

NFSv4 Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 20, 2014

W. Adamson  
NetApp  
N. Williams  
Cryptonector  
October 17, 2013

NFSv4 Multi-Domain FedFS Requirements  
draft-adamson-nfsv4-multi-domain-federated-fs-reqs-03

Abstract

This document describes constraints to the NFSv4.0 and NFSv4.1 protocols as well as the use of multi-domain capable file systems, name resolution services, and security services required to fully enable a multi-domain NFSv4 federated file system.

## Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 20, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1.	Introduction . . . . .	4
2.	Terminology . . . . .	5
3.	NFSv4 Server Identity Mapping . . . . .	7
4.	Multi-domain Constraints to the NFSv4 Protocol . . . . .	8
4.1.	Name@domain Constraints . . . . .	8
4.2.	RPC Security Constraints . . . . .	8
5.	Resolving Multi-domain Authorization Information . . . . .	10
5.1.	GSS-API Authorization Payload . . . . .	11
6.	Setting and Retrieving NFSv4 Multi-domain ACLs . . . . .	12
7.	Security Considerations . . . . .	13
8.	References . . . . .	14
8.1.	Normative References . . . . .	14
8.2.	Informative References . . . . .	15
	Authors' Addresses . . . . .	16

## 1. Introduction

This document describes constraints to the NFSv4.0 and NFSv4.1 protocols as well as the use of multi-domain capable file systems, name resolution services, and security services required to fully enable an NFSv4 multi-domain federated file system.

The definition of an NFSv4 multi-domain federated file system combines these concepts:

1. NFSv4 Domain, Pseudo file system and referrals: The NFSv4.0 [RFC3530] and NFSv4.1 [RFC5661] (hereafter referred to as NFSv4) protocols enable the construction of a distributed file system which can join NFSv4 servers from multiple NFSv4 domains, each potentially using separate name resolution services and separate security services, into a common multi-domain name space.
2. The Federated File System (FedFS): [RFC5716] describes the requirements and administrative tools to construct a uniform file server based namespace that is capable of spanning a whole enterprise and that is easy to manage.
3. Multi-domain capable filesystem: A local filesystem that uses a local ID form that can represent identities from both local and remote domains. For example, an SSID based local ID form where the SSID contains both a domain and a user or group component. We note that many file systems exported by NFSv4 use 32 bit POSIX UID and GIDs as a local ID form and are therefore not domain aware and not able to participate in an NFSv4 multi-domain federated file system. There are ways to overcome this deficiency, but these practices are beyond the scope of this document.

An NFSv4 multi-domain federated file system uses the FedFS to join multiple NFSv4 domains each consisting of NFSv4 servers that export multi-domain capable filesystems, into a uniform NFSv4 server-based name space capable of spanning multiple enterprises.

## 2. Terminology

**Name Service:** provides the mapping between {NFSv4 domain, group or use name} and {NFSv4 domain, local ID} via lookups. Can be applied to local or remote domains. Often provided by a Directory Service such as LDAP.

**Domain:** This term is used in multiple contexts where it has different meanings. Here we provide specific definitions used in this document.

**DNS domain:** a set of computers, services, or any internet resource identified by an DNS domain name [RFC1034].

**Security realm or domain:** a set of configured security providers, users, groups, security roles, and security policies running a single security protocol and administered by a single entity. E.g. a Kerberos Realm. While a typical configuration is to use the uppercase DNS domain name as the Kerberos realm name they are independent.

**NFSv4 domain:** a set of users, groups and computers running NFSv4 protocols running a single name service, and identified by a unique NFSv4 domain name. See [RFC5661] Section 5.9 "Interpreting owner and owner\_group". An NFSv4 domain can include multiple DNS domains and multiple security realms but only one name service.

**Multi-domain:** In this document this always refers to multiple NFSv4 domains.

**FedFS domain:** A file name space that can cross multiple shares on multiple file servers using file-access protocols such as NFSv4 or CIFS [CIFS]. A FedFS domain is typically a single administrative entity, and has a name that is similar to a DNS domain name. Also known as a Federation.

**Administrative domain:** a set of users, groups, computers and services administered by a single entity. Can include multiple DNS domains, NFSv4 domains, security domains, and FedFS domains.

**Local representation of identity:** an object such as a uidNumber (UID) or gidNumber (GID) [RFC2307], or a Windows Security Identifier (SID), or other such representation of a user or a group of users on-disk in a file system.

Principal: an RPCSEC\_GSS authentication identity. Usually, but not always, a user; rarely, if ever, a group; sometimes a host.

Authorization Context: A collection of information about a principal such as username, userID, group membership, etcetera used in authorization decisions.

### 3. NFSv4 Server Identity Mapping

NFSv4 deals with two kinds of identities: authentication identities (referred to here as "principals") and authorization identities ("users" and "groups" of users). NFSv4 supports multiple authentication methods, each authenticating an "initiator principal" (typically representing a user) to an "acceptor principal" (always corresponding to the NFSv4 server). NFSv4 does not prescribe how to represent authorization identities on file systems. All file access decisions constitute "authorization" and are made by NFSv4 servers using authorization context information and file metadata related to authorization, such as a file's access control list (ACL).

NFSv4 servers therefore must perform two kinds of mappings:

1. A mapping between the authentication identity and the authorization context information.
2. A mapping between the on-the-wire authorization identity representation and the on-disk authorization identity representation.

Many aspects of these mappings are entirely implementation specific, but some require multi-domain capable name resolution and security services. In order to interoperate in a multi-domain NFSv4 FedFS file system, NFSv4 servers must use such services in compatible ways.

#### 4. Multi-domain Constraints to the NFSv4 Protocol

In order to service as many environments as possible, the NFSv4 protocol is designed to allow administrators freedom to configure their NFSv4 domains as they please. Joining NFSv4 domains under a single file namespace imposes slightly on this freedom. Here we describe the required constraints.

##### 4.1. Name@domain Constraints

NFSv4 uses a syntax of the form "name@domain" as the on wire representation of the "who" field of an NFSv4 access control entry (ACE) for users and groups. This design provides a level of indirection that allows NFSv4 clients and servers with different internal representations of authorization identity to interoperate even when referring to authorization identities from different NFSv4 domains.

NFSv4 multi-domain capable sites need to meet the following requirements in order to ensure that NFSv4 clients and servers can map between name@domain and internal representations reliably:

- o The NFSv4 domain portion of name@domain MUST be unique within the FedFS NFSv4 multi-domain namespace. See [RFC3530] section 5.9 "Interpreting owner and owner\_group" for a discussion on NFSv4 domain configuration.
- o The name portion of name@domain MUST be unique within the specified NFSv4 domain.
- o Every local representation of a user and of a group MUST have a canonical name@domain, and it must be possible to return the canonical name@domain for any identity stored on disk, at least when required infrastructure servers (such as name services) are online.

##### 4.2. RPC Security Constraints

As described in [RFC5661] section 2.2.1.1 "RPC Security Flavors":

NFSv4.1 clients and servers MUST implement RPCSEC\_GSS.  
(This requirement to implement is not a requirement to use.) Other flavors, such as AUTH\_NONE, and AUTH\_SYS, MAY be implemented as well.

The underlying RPCSEC\_GSS security mechanism used in a multi-domain NFSv4 FedFS is REQUIRED to employ a method of cross NFSv4 domain trust so that a principal from a security service in one NFSv4 domain



can be authenticated in another NFSv4 domain that uses a security service with the same security mechanism. Kerberos, and PKU2U [I-D.zhu-pku2u] are examples of such security services.

The AUTH\_NONE security flavor can be useful in a multi-domain NFSv4 FedFS to grant universal access to public data without any credentials.

The AUTH\_SYS security flavor uses a host-based authentication model where the weakly authenticated host (the NFSv4 client) asserts the user's authorization identities using small integers, uidNumber and gidNumber [RFC2307], as user and group identity representations. Because this authorization ID representation has no DNS domain component, AUTH\_SYS can only be used in a name space where all NFSv4 clients and servers share an [RFC2307] name service. A shared name service is required because uidNumbers and gidNumbers are passed in the RPC credential; there is no negotiation of namespace in AUTH\_SYS. Collisions can occur if multiple name services are used. AUTH\_SYS can not be used in an NFSv4 multi-domain federated file system.

## 5. Resolving Multi-domain Authorization Information

When an RPCSEC\_GSS principal is seeking access to files on an NFSv4 server, after authenticating the principal, the server must obtain in a secure manner the principal's authorization context information from an authoritative source: e.g. the name service in the principal's NFSv4 domain.

In the local NFSv4 domain case where the principal is seeking access to files on an NFSv4 server in the principal's NFSv4 home domain, the server administrator has knowledge of the local policies and methods for obtaining the principal's authorization information and the mappings to local representation of identity. E.g. the administrator can configure secure access to the local NFSv4 domain name service.

In the multi-domain case where a principal from a remote NFSv4 domain is seeking access to files on an NFSv4 server not in the principal's domain, there is no assumption of:

- o remote name service configuration knowledge
- o the form of the remote authorization context information presented to the NFSv4 server by the remote name service for mapping to a local representation.

There are several methods the NFSv4 server can use to obtain the NFSv4 domain authoritative authorization information for a remote principal, listed in order of preference:

1. A mechanism specific GSS-API authorization payload containing credential authorization data described in the following section. This is the preferred method as it is part of the GSS-API authentication and avoids requiring any knowledge of a remote NFSv4 domain name service configuration, and has a set form of authorization context information to allow mapping to a local representation.
2. When there is a security agreement between the local and remote NFSv4 domain name services plus regular update data feeds, the NFSv4 server local NFSv4 domain name service can be authoritative for principal's in a remote NFSv4 domain. In this case, the NFSv4 server makes a query to it's local NFSv4 domain name service just as it does when servicing a local domain principal. While this requires detailed knowledge of the remote NFSv4 domains name service, the authorization context information presented to the NFSv4 server is in the same form as a query for a local principal.

3. An authenticated direct query from the NFSv4 server to the principal's NFSv4 domain authoritative name service. This requires the NFSv4 server to have detailed knowledge of the remote NFSv4 domain's authoritative name service and detailed knowledge of the form of the resultant authorization context information.

Once the authorization context data for the remote principal has been obtained, the remote information must be mapped into local representations suitable for use in file system ACLs. This is the first mapping described in Section 3.

#### 5.1. GSS-API Authorization Payload

To avoid requiring detailed knowledge of remote NFSv4 domain name services, authorization context information SHOULD be obtained from the credentials authenticating a principal; the GSS-API represents such information as attributes of the initiator principal name.

For example:

- o Kerberos 5 [RFC4120] has a method for conveying "authorization data", both client-asserted as well as KDC-authenticated authorization data. Some KDC implementation, notably Windows KDCs, use this feature to convey a "privilege attribute certificate" [PAC] listing the principal's user and group global IDs as "security identifiers" (SIDs).
- o Some KDCs (will) issue Kerberos Tickets with the General PAD [I-D.sorce-krbwg-general-pac] (PAD) as Kerberos authorization data listing user and group names along with their uidNumber and gidNumber [RFC2307], the name of the DNS domain [ANDROS: review General PAD RFC to ensure this can be the NFSv4 domain] along with a unique DNS domain identifier and other information. The General PAD authorization data MUST be authenticated in the sense that its contents must come from an authenticated, trusted source, such as a name server or the issuer of the principal's credential.
- o PKIX [RFC5280] certificates allow for extensions that could be used similarly.

## 6. Setting and Retrieving NFSv4 Multi-domain ACLs

When servicing a set acl request, the NFSv4 server must be able to map the name@domain in the ACE who field to a local representation of ID. When servicing a get acl request, the NFSv4 server must be able to map the local representation of ID in the file system ACL to the name@domain form. This mapping between name@domain and local representation of ID must [ANDROS: MUST?] be done against an authoritative source. This is the second mapping described in Section 3.

The local name-service is authoritative for these mappings for remote users and groups when one of the first two methods in (Section 5) is used to keep the local name-service updated with remote information.

## 7. Security Considerations

Some considerations to come

## 8. References

### 8.1. Normative References

- [RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2307] Howard, L., "An Approach for Using LDAP as a Network Information Service", RFC 2307, March 1998.
- [RFC3530] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.
- [RFC4120] Neuman, C., Yu, T., Hartman, S., and K. Raeburn, "The Kerberos Network Authentication Service (V5)", RFC 4120, July 2005.
- [RFC5661] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.
- [RFC5716] Lentini, J., Everhart, C., Ellard, D., Tewari, R., and M. Naik, "Requirements for Federated File Systems", RFC 5716, January 2010.
- [I-D.zhu-pku2u]  
Zhu, L., Altman, J., and N. Williams, "Public Key Cryptography Based User-to-User Authentication - (PKU2U)", draft-zhu-pku2u-09 (work in progress), November 2008.
- [I-D.sorce-krbwg-general-pac]  
Sorce, S., Yu, T., and T. Hardjono, "A Generalized PAC for Kerberos V5", draft-sorce-krbwg-general-pac-01 (work in progress), December 2010.
- [PAC] Brezak, J., "Utilizing the Windows 2000 Authorization Data in Kerberos Tickets for Access Control to Resources", October 2002.
- [CIFS] Microsoft Corporation, "[MS-CIFS] -- v20130118 Common Internet File System (CIFS) Protocol", January 2013.

## 8.2. Informative References

- [RFC5280] Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., and W. Polk, "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile", RFC 5280, May 2008.

Authors' Addresses

William A. (Andy) Adamson  
NetApp

Email: andros@netapp.com

Nicolas Williams  
Cryptonector

Email: nico@cryptonector.com





NFSv4  
Internet-Draft  
Intended status: Standards Track  
Expires: April 23, 2014

B. Halevy  
Primary Data  
October 20, 2013

Parallel NFS (pNFS) Flexible Files Layout  
draft-bhalevy-nfsv4-flex-files-01

Abstract

Parallel NFS (pNFS) extends Network File System version 4 (NFSv4) to allow clients to directly access file data on the storage used by the NFSv4 server. This ability to bypass the server for data access can increase both performance and parallelism, but requires additional client functionality for data access, some of which is dependent on the class of storage used, a.k.a. the Layout Type. The main pNFS operations and data types in NFSv4 Minor version 1 specify a layout-type-independent layer; layout-type-specific information is conveyed using opaque data structures whose internal structure is further defined by the particular layout type specification. This document specifies the NFSv4.1 Flexible Files pNFS Layout as a companion to the main NFSv4 Minor version 1 specification for use of pNFS with Data Servers over NFSv4 or higher minor versions using flexible, per-file striping topology.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 23, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	4
1.1. Requirements Language . . . . .	4
2. Method of Operation . . . . .	4
2.1. Security models . . . . .	5
2.2. State and Locking Models . . . . .	5
3. XDR Description of the Flexible Files Layout Protocol . . . . .	6
3.1. Code Components Licensing Notice . . . . .	6
4. Device Addressing and Discovery . . . . .	8
4.1. pnfs_ff_device_addr . . . . .	8
4.2. Data Server Multipathing . . . . .	9
5. Flexible Files Layout . . . . .	10
5.1. pnfs_ff_layout . . . . .	11
5.2. Striping Topologies . . . . .	14
5.2.1. PFSP_SPARSE_STRIPING . . . . .	14
5.2.2. PFSP_DENSE_STRIPING . . . . .	15
5.2.3. PFSP_RAID_4 . . . . .	16
5.2.4. PFSP_RAID_5 . . . . .	16
5.2.5. PFSP_RAID_PQ . . . . .	17
5.2.6. RAID Usage and Implementation Notes . . . . .	18
5.3. Mirroring . . . . .	18
6. Recovering from Client I/O Errors . . . . .	18
7. Flexible Files Layout Return . . . . .	19
7.1. pflr_errno . . . . .	20
7.2. pnfs_ff_ioerr . . . . .	21
7.3. pnfs_ff_iostats . . . . .	22
7.4. pnfs_ff_layoutreturn . . . . .	23
8. Flexible Files Creation Layout Hint . . . . .	23
8.1. pnfs_ff_layouthint . . . . .	24
9. Recalling Layouts . . . . .	25
9.1. CB_RECALL_ANY . . . . .	25
10. Client Fencing . . . . .	26
11. Security Considerations . . . . .	26
12. Striping Topologies Extensibility . . . . .	27
13. IANA Considerations . . . . .	27
14. Normative References . . . . .	27
Appendix A. Acknowledgments . . . . .	28
Author's Address . . . . .	29

## 1. Introduction

In pNFS, the file server returns typed layout structures that describe where file data is located. There are different layouts for different storage systems and methods of arranging data on storage devices. This document defines the layout used with file-based data servers that are accessed using the Network File System (NFS) Protocol: NFSv3 (RFC1813 [1]), NFSv4 (RFC3530 [2]) and its newer minor version - NFSv4.1 (RFC5661 [3]).

In contrast to the LAYOUT4\_NFSV4\_1\_FILES layout type (RFC5661 [3]) that also uses NFSv4.1 to access the data server, the Flexible Files layout defines a model of device metadata and striping patterns that is inspired by the object layout (RFC5664 [4]) that provide flexible, per-file striping patterns and simple device information suitable aggregating standalone NFS servers into a centrally managed pNFS cluster.

To provide a global state model equivalent to that of the files layout a back-end control protocol may be implemented between the metadata server (MDS) and NFSv4.1 data servers (DSs). It is out of scope for this document to specify the wire protocol of such a protocol, yet the requirements for the protocol are specified in RFC5661 [3]. The actual protocol definition of a standard back-end control protocol conforming to these requirements is encouraged to be specified within the IETF as a separate RFC.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [5].

## 2. Method of Operation

This section describes the semantics and format of flexible file-based layouts for pNFS. Flexible file-based layouts use the LAYOUT4\_FLEX\_FILES layout type. The LAYOUT4\_FLEX\_FILES type defines striping data across multiple NFS Data Servers.

For the purpose of this discussion, we will distinguish between user files served by the metadata server, to be referred to as User Files; vs. user files served by Data Servers, to be referred to as Component Objects.

Component Objects are addressable by their NFS filehandle. Each Component Object may store a whole User File or parts of it, in case

the User File is striped across multiple Component Objects. The striping pattern is provided by `pfl_striping_pattern` as defined below.

Data Servers may be accessed using different versions of the NFS protocol. It is required that the server **MUST** use Data Servers of the same NFS version and minor version for striping data within each layout. The NFS version and minor version define the respective security, state, and locking models to be used, as described below.

### 2.1. Security models

With NFSv3 Data Servers, the Metadata Server uses synthetic uids and gids for the Component Objects, where the uid owner of the Component Objects is allowed read/write access and the gid owner is allowed read only access. As part of the layout, the client is provided with the rpc credentials to be used (`XREF pfcf_auth`) to access the Object. Fencing off clients is achieved by using `SETATTR` by the server to change the uid and/or gid owners of the Component Objects to implicitly revoke the outstanding rpc credentials. Note: it is recommended to implement common access control methods at the Data Server filesystem exports level to allow only the Metadata Server root (super user) access to the Data Server, and to set the owner of all directories holding Component Objects to the root user. This security method, when using weak auth flavors such as `AUTH_SYS`, provides a practical model to enforce access control and fence off cooperative clients, but it can not protect against malicious clients; hence it provides a level of security equivalent to NFSv3.

With NFSv4.x Data Servers, the Metadata Server sets the user and group owners, mode bits, and ACL of the Component Objects to be the same as the User File. And the client must authenticate with the Data Server and go through the same authorization process it would go through via the Metadata Server.

### 2.2. State and Locking Models

User File `OPEN`, `LOCK`, and `DELEGATION` operations are always executed only against the Metadata Server.

With NFSv4 Data Servers, the Metadata Server, in response to the state changing operation, executes them against the respective Component Objects on the Data Server(s). It then sends the Data Server open stateid as part of the layout (`XREF pfcf_stateid`) and it is then used by the client for executing `READ/WRITE` operations against the Data Server.

Standalone NFSv4.1 Data Servers that do not return the

EXCHGID4\_FLAG\_USE\_PNFS\_DS flag to EXCHANGE\_ID are used the same way as NFSv4 Data Servers.

NFSv4.1 Clustered Data Servers that do identify themselves with the EXCHGID4\_FLAG\_USE\_PNFS\_DS flag to EXCHANGE\_ID use a back-end control protocol as described in RFC5661 [3] to implement a global stateid model as defined there.

### 3. XDR Description of the Flexible Files Layout Protocol

This document contains the external data representation (XDR [6]) description of the NFSv4.1 flexible files layout protocol. The XDR description is embedded in this document in a way that makes it simple for the reader to extract into a ready-to-compile form. The reader can feed this document into the following shell script to produce the machine readable XDR description of the NFSv4.1 objects layout protocol:

```
#!/bin/sh
grep '^ *///' $* | sed 's?^ */// ??' | sed 's?^ *///$??'
```

That is, if the above script is stored in a file called "extract.sh", and this document is in a file called "spec.txt", then the reader can do:

```
sh extract.sh < spec.txt > pnfs_flex_files_prot.x
```

The effect of the script is to remove leading white space from each line, plus a sentinel sequence of "///".

The embedded XDR file header follows. Subsequent XDR descriptions, with the sentinel sequence are embedded throughout the document.

Note that the XDR code contained in this document depends on types from the NFSv4.1 nfs4\_prot.x file ([7]). This includes both nfs types that end with a 4, such as offset4, length4, etc., as well as more generic types such as uint32\_t and uint64\_t.

#### 3.1. Code Components Licensing Notice

The XDR description, marked with lines beginning with the sequence "///", as well as scripts for extracting the XDR description are Code Components as described in Section 4 of "Legal Provisions Relating to IETF Documents" [8]. These Code Components are licensed according to the terms of Section 4 of "Legal Provisions Relating to IETF Documents".

```
/// /*
/// * Copyright (c) 2012 IETF Trust and the persons identified
/// * as authors of the code. All rights reserved.
/// *
/// * Redistribution and use in source and binary forms, with
/// * or without modification, are permitted provided that the
/// * following conditions are met:
/// *
/// * o Redistributions of source code must retain the above
/// *   copyright notice, this list of conditions and the
/// *   following disclaimer.
/// *
/// * o Redistributions in binary form must reproduce the above
/// *   copyright notice, this list of conditions and the
/// *   following disclaimer in the documentation and/or other
/// *   materials provided with the distribution.
/// *
/// * o Neither the name of Internet Society, IETF or IETF
/// *   Trust, nor the names of specific contributors, may be
/// *   used to endorse or promote products derived from this
/// *   software without specific prior written permission.
/// *
/// * THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS
/// * AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED
/// * WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
/// * IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS
/// * FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO
/// * EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE
/// * LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL,
/// * EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT
/// * NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR
/// * SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS
/// * INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF
/// * LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY,
/// * OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING
/// * IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF
/// * ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
/// *
/// * This code was derived from draft-bhalevy-nfsv4-flex-files-01.
[[RFC Editor: please insert RFC number if needed]]
/// * Please reproduce this note if possible.
/// */
///
/// /*
/// * pnfs_flex_files_prot.x
/// */
///
/// /*
```



```
/// * The following include statements are for example only.
/// * The actual XDR definition files are generated separately
/// * and independently and are likely to have a different name.
/// */
/// %#include <nfs4_prot.x>
/// %#include <rpc_prot.x>
///
```

#### 4. Device Addressing and Discovery

Data operations to a data server require the client to know the network address of the data server. The GETDEVICEINFO NFSv4.1 operation is used by the client to retrieve that information.

##### 4.1. pnfs\_ff\_device\_addr

The pnfs\_ff\_device\_addr data structure is returned by the server as the storage-protocol-specific opaque field da\_addr\_body in the device\_addr4 structure by a successful GETDEVICEINFO operation [3].

```
/// struct pnfs_ff_device_addr {
///     multipath_list4      pfda_netaddrs;
///     uint32_t             pfda_version;
///     uint32_t             pfda_minorversion;
///     pathname4            pfda_path;
/// };
///
```

The pfda\_netaddrs field is used to locate the data server. It MUST be set by the server to a list holding one or more of the device network addresses.

pfda\_version and pfda\_minorversion represent the NFS protocol to be used to access the data server. This layout specification defines the semantics for pfda\_versions 3 and 4. If pfda\_version equals 3 then server MUST set pfda\_minorversion to 0 and the client MUST access the data server using the NFSv3 protocol (RFC1813 [1]). If pfda\_version equals 4 then the server MUST set pfda\_minorversion to either 0 or 1 and the client MUST access the data server using NFSv4 (RFC3530 [2]) or NFSv4.1 (RFC5661 [3]), respectively.

pfda\_path MAY be set by the server to an exported path on the data server for device identification. If provided, the path MUST exist and be accessible to the client. If the path does not exist, the client MUST ignore this device information and any layouts referring to the respective deviceid until valid device information is acquired.

#### 4.2. Data Server Multipathing

The flexible file layout supports multipathing to multiple data server addresses. Data-server-level multipathing is used for bandwidth scaling via trunking and for higher availability of use in the case of a data-server failure. Multipathing allows the client to switch to another data server address which may be that of another data server that is exporting the same data stripe unit, without having to contact the metadata server for a new layout.

To support data server multipathing, `pfda_netaddrs` contains an array of one more data server network addresses. This array (data type `multipath_list4`) represents a list of data servers (each identified by a network address), with the possibility that some data servers will appear in the list multiple times.

The client is free to use any of the network addresses as a destination to send data server requests. If some network addresses are less optimal paths to the data than others, then the MDS SHOULD NOT include those network addresses in `pfda_netaddrs`. If less optimal network addresses exist to provide failover, the RECOMMENDED method to offer the addresses is to provide them in a replacement device-ID-to-device-address mapping, or a replacement device ID. When a client finds no response from the data server using all addresses available in `pfda_netaddrs`, it SHOULD send a `GETDEVICEINFO` to attempt to replace the existing device-ID-to-device-address mappings. If the MDS detects that all network paths represented by `pfda_netaddrs` are unavailable, the MDS SHOULD send a `CB_NOTIFY_DEVICEID` (if the client has indicated it wants device ID notifications for changed device IDs) to change the device-ID-to-device-address mappings to the available addresses. If the device ID itself will be replaced, the MDS SHOULD recall all layouts with the device ID, and thus force the client to get new layouts and device ID mappings via `LAYOUTGET` and `GETDEVICEINFO`.

Generally, if two network addresses appear in `pfda_netaddrs`, they will designate the same data server. When the data server is accessed over NFSv4.1 or higher minor version the two data server addresses will support the implementation of client ID or session trunking (the latter is RECOMMENDED) as defined in RFC5661 [3]. The two data server addresses will share the same server owner or major ID of the server owner. It is not always necessary for the two data server addresses to designate the same server with trunking being used. For example, the data could be read-only, and the data consist of exact replicas.

## 5. Flexible Files Layout

The layout4 type is defined in the [3] protocol as follows:

```
/// enum layouttype4 {
///     LAYOUT4_NFSV4_1_FILES    = 1,
///     LAYOUT4_OSD2_OBJECTS     = 2,
///     LAYOUT4_BLOCK_VOLUME     = 3,
///     LAYOUT4_FLEX_FILES       = 4
/// [[RFC Editor: please insert layouttype assigned by IANA]]
/// };
///
/// struct layout_content4 {
///     layouttype4      loc_type;
///     opaque           loc_body<>;
/// };
///
/// struct layout4 {
///     offset4          lo_offset;
///     length4          lo_length;
///     layoutiomode4     lo_iomode;
///     layout_content4  lo_content;
/// };
```

This document defines structure associated with the layouttype4 value LAYOUT4\_FLEX\_FILES. NFSv4.1 RFC5661 [3] specifies the loc\_body structure as an XDR type "opaque". The opaque layout is uninterpreted by the generic pNFS client layers, but obviously must be interpreted by the flexible files layout driver. This section defines the structure of this opaque value, pnfs\_ff\_layout4.

## 5.1. pnfs\_ff\_layout

```

    /// enum pnfs_ff_stripping_pattern {
    ///     PFSP_SPARSE_STRIPING = 1,
    ///     PFSP_DENSE_STRIPING  = 2,
    ///     PFSP_RAID_4           = 4,
    ///     PFSP_RAID_5           = 5,
    ///     PFSP_RAID_PQ          = 6
    /// };
    ///
    /// enum pnfs_ff_comp_type {
    ///     PNFS_FF_COMP_MISSING = 0,
    ///     PNFS_FF_COMP_PACKED  = 1,
    ///     PNFS_FF_COMP_FULL    = 2
    /// };
    ///
    /// struct pnfs_ff_comp_full {
    ///     deviceid4      pfcf_deviceid;
    ///     nfs_fh4         pfcf_fhandle;
    ///     stateid4        pfcf_stateid;
    ///     opaque_auth     pfcf_auth;
    ///     uint32_t         pfcf_metric;
    /// };
    ///
    /// union pnfs_ff_comp switch (pnfs_ff_comp_type pfc_type) {
    ///     case PNFS_FF_COMP_MISSING:
    ///         void;
    ///
    ///     case PNFS_FF_COMP_PACKED:
    ///         deviceid4      pfcf_deviceid;
    ///
    ///     case PNFS_FF_COMP_FULL:
    ///         pnfs_ff_comp_full pfcf_full;
    /// };
    ///
    /// struct pnfs_ff_layout {
    ///     pnfs_ff_stripping_pattern pfl_stripping_pattern;
    ///     uint32_t                  pfl_num_comps;
    ///     uint32_t                  pfl_mirror_cnt;
    ///     length4                   pfl_stripe_unit;
    ///     nfs_fh4                   pfl_global_fh;
    ///     uint32_t                  pfl_comps_index;
    ///     pnfs_ff_comp              pfl_comps<>;
    /// };
    ///

```

The `pnfs_ff_layout` structure specifies a layout over a set of Component Objects. The layout parameterizes the algorithm that maps

the file's contents within the returned byte range, as represented by `lo_offset` and `lo_length`, over the Component Objects.

It is possible that the file is concatenated from more than one layout segment. Each layout segment MAY represent different striping parameters, applying respectively only to the layout segment byte range.

This section provides a brief introduction to the layout parameters. See Section 5.2 for a more detailed description of the different striping schemes and the respective interpretation of the layout parameters for each striping scheme.

In addition to mapping data using simple striping schemes where loss of a single component object results in data loss, the layout parameters support mirroring and more advanced redundancy schemes that protect against loss of component objects. `pfl_striping_pattern` represents the algorithm to be used for mapping byte offsets in the file address space to corresponding component objects in the returned layout and byte offsets in the component's address space. `pfl_striping_pattern` also represents methods for storing and retrieving redundant data that can be used to recover from failure or loss of component objects.

`pfl_num_comps` is the total number of component objects the file is striped over within the returned byte range, not counting mirrored components (See `pfl_mirror_cnt` below). Note that the server MAY grow the file by adding more components to the stripe while clients hold valid layouts until the file has reached its final stripe width.

`pfl_mirror_cnt` represents the number of mirrors each component in the stripe has. If there is no mirroring then `pfl_mirror_cnt` MUST be 0. Otherwise, the number of entries listed in `pfl_comps` MUST be a multiple of (`pfl_mirror_cnt+1`).

`pfl_stripe_unit` is the number of bytes placed on one component before advancing to the next one in the list of components. When the file is striped over a single component object (`pfl_num_comps` equals to 1), the stripe unit has no use and the server SHOULD set it to the server default value or to zero; otherwise, `pfl_stripe_unit` MUST NOT be set to zero.

The `pfl_comps` field represents an array of component objects. The data placement algorithm that maps file data onto component objects assumes that each component object occurs exactly once in the array of components. Therefore, component objects MUST appear in the `pfl_comps` array only once. The components array may represent all objects comprising the file, in which case `pfl_comps_index` is set to

zero and the number of entries in the `pfl_comps` array is equal to `pfl_num_comps * (pfl_mirror_cnt + 1)`. The server MAY return fewer components than `pfl_num_comps`, provided that the returned byte range represented by `lo_offset` and `lo_count` maps in whole into the set of returned component objects. In this case, `pfl_comps_index` represents the logical position of the returned components array, `pfl_comps`, within the full array of components that comprise the file. `pfl_comps_index` MUST be a multiple of `(pfl_mirror_cnt + 1)`.

Each component object in the `pfl_comps` array is described by the `pnfs_ff_comp` type.

When a component object is unavailable `pfc_type` is set to `PNFS_FF_COMP_MISSING` and no other information for this component is returned. When a data redundancy scheme is being used, as represented by `pfl_striping_pattern`, the client MAY use a respective data recovery algorithm to reconstruct data that is logically stored on the missing component using user data and redundant data stored on the available components in the containing stripe.

The server MUST set the same `pfc_type` for all available components to either `PNFS_FF_COMP_PACKED` or `PNFS_FF_COMP_FULL`.

When NFSv4.1 Clustered Data Servers are used, the metadata server implements the global state model where all data servers share the same stateid and filehandle for the file. In such case, the client MUST use the open, delegation, or lock stateid returned by the metadata server for the file for accessing the Data Servers for READ and WRITE; the global filehandle to be used by the client is provided by `pfl_global_fh`. If the metadata server filehandle for the file is being used by all data servers then `pfl_global_fh` MAY be set to an empty filehandle.

`pfc_p_deviceid` or `pfc_f_deviceid` provide the deviceid of the data server holding the Component Object.

When standalone data servers are used, either over NFSv4 or NFSv4.1, `pfl_global_fh` SHOULD be set to an empty filehandle and it MUST be ignored by the client and `pfc_f_handle` provides the filehandle of the Data Server file holding the Component Object, and `pfc_f_stateid` provides the stateid to be used by the client to access the file.

For NFSv3 Data Servers, `pfc_auth` provides the rpc credentials to be used by the client to access the Component Objects. For NFSv4.x Data Servers, the server SHOULD use the `AUTH_NONE` flavor and a zero length opaque body to minimize the returned structure length. The client MUST ignore `pfx_auth` in this case.

When `pfl_mirror_cnt` is not zero `pfcf_metric` indicates the distance to the client the distance of the respective component object, otherwise the server MUST set `pfcf_metric` to zero. When reading data, the client the client is advised to read from components with the lowest `pfcf_metric`. When there are several components with the same `pfcf_metric` client implementations may implement a load distribution algorithm to evenly distribute the read load across several devices and by so provide larger bandwidth.

## 5.2. Striping Topologies

This section describes the different data mapping schemes in detail.

`pnfs_ff_striping_pattern` determines the algorithm and placement of redundant data. This section defines the different redundancy algorithms. Note: The term "RAID" (Redundant Array of Independent Disks) is used in this document to represent an array of Component Objects that store data for an individual User File. The objects are stored on independent Data Servers. User File data is encoded and striped across the array of Component Objects using algorithms developed for block-based RAID systems.

### 5.2.1. PFSP\_SPARSE\_STRIPING

The mapping from the logical offset within a file ( $L$ ) to the Component Object  $C$  and object-specific offset  $O$  is direct and straight forward as defined by the following equations:

$L$ : logical offset into the file

$W$ : stripe width  
 $W = \text{pfl\_num\_comps}$

$S$ : number of bytes in a stripe  
 $S = W * \text{pfl\_stripe\_unit}$

$N$ : stripe number  
 $N = L / S$

$C$ : component index corresponding to  $L$   
 $C = (L \% S) / \text{pfl\_stripe\_unit}$

$O$ : The component offset corresponding to  $L$   
 $O = L$

Note that this computation does not accommodate the same object appearing in the `pfl_comps` array multiple times. Therefore the server may not return layouts with the same object appearing multiple

times. If needed the server can return multiple layout segments each covering a single instance of the object.

PFSP\_SPARSE\_STRIPING means there is no parity data, so all bytes in the component objects are data bytes located by the above equations for C and O. If a component object is marked as PNFS\_FF\_COMP\_MISSING, the pNFS client MUST either return an I/O error if this component is attempted to be read or, alternatively, it can retry the READ against the pNFS server.

#### 5.2.2. PFSP\_DENSE\_STRIPING

The mapping from the logical offset within a file (L) to the component object C and object-specific offset O is defined by the following equations:

L: logical offset into the file

W: stripe width  
 $W = \text{pfl\_num\_comps}$

S: number of bytes in a stripe  
 $S = W * \text{pfl\_stripe\_unit}$

N: stripe number  
 $N = L / S$

C: component index corresponding to L  
 $C = (L \% S) / \text{pfl\_stripe\_unit}$

O: The component offset corresponding to L  
 $O = (N * \text{pfl\_stripe\_unit}) + (L \% \text{pfl\_stripe\_unit})$

Note that this computation does not accommodate the same object appearing in the pfl\_comps array multiple times. Therefore the server may not return layouts with the same object appearing multiple times. If needed the server can return multiple layout segments each covering a single instance of the object.

PFSP\_DENSE\_STRIPING means there is no parity data, so all bytes in the component objects are data bytes located by the above equations for C and O. If a component object is marked as PNFS\_FF\_COMP\_MISSING, the pNFS client MUST either return an I/O error if this component is attempted to be read or, alternatively, it can retry the READ against the pNFS server.

Note that the layout depends on the file size, which the client learns from the generic return parameters of LAYOUTGET, by doing



GETATTR commands to the Metadata Server. The client uses the file size to decide if it should fill holes with zeros or return a short read. Striping patterns can cause cases where Component Objects are shorter than other components because a hole happens to correspond to the last part of the Component Object.

#### 5.2.3. PFSP\_RAID\_4

PFSP\_RAID\_4 means that the last component object in the stripe contains parity information computed over the rest of the stripe with an XOR operation. If a Component Object is unavailable, the client can read the rest of the stripe units in the damaged stripe and recompute the missing stripe unit by XORing the other stripe units in the stripe. Or the client can replay the READ against the pNFS server that will presumably perform the reconstructed read on the client's behalf.

When parity is present in the file, then the number of parity devices is taken into account in the above equations when calculating (D), the number of data devices in a stripe, as follows:

P: number of parity devices in each stripe  
 $P = 1$

D: number of data devices in a stripe  
 $D = W - P$

I: parity device index  
 $I = D$

#### 5.2.4. PFSP\_RAID\_5

PNFS\_OBJ\_RAID\_5 means that the position of the parity data is rotated on each stripe. In the first stripe, the last component holds the parity. In the second stripe, the next-to-last component holds the parity, and so on. In this scheme, all stripe units are rotated so that I/O is evenly spread across objects as the file is read sequentially. The rotated parity layout is illustrated here, with hexadecimal numbers indicating the stripe unit.

```
0 1 2 P
4 5 P 3
8 P 6 7
P 9 a b
```

Note that the math for RAID\_5 is similar to RAID\_4 only that the device indices for each stripe are rotated backwards. So start with the equations above for RAID\_4, then compute the rotation as

described below.

P: number of parity devices in each stripe  
 $P = 1$

PC: Parity Cycle  
 $PC = W$

R: The parity rotation index  
 (N is as computed in above equations for RAID-4)  
 $R = N \% PC$

I: parity device index  
 $I = (W + W - (R + 1) * P) \% W$

Cr: The rotated device index  
 (C is as computed in the above equations for RAID-4)  
 $Cr = (W + C - (R * P)) \% W$

Note: W is added above to avoid negative numbers modulo math.

#### 5.2.5. PFSP\_RAID\_PQ

PFSP\_RAID\_PQ is a double-parity scheme that uses the Reed-Solomon P+Q encoding scheme [9]. In this layout, the last two component objects hold the P and Q data, respectively. P is parity computed with XOR. The Q computation is described in detail by Anvin [10]. The same polynomial " $x^8+x^4+x^3+x^2+1$ " and Galois field size of  $2^8$  are used here. Clients may simply choose to read data through the metadata server if two or more components are missing or damaged.

The equations given above for embedded parity can be used to map a file offset to the correct component object by setting the number of parity components (P) to 2 instead of 1 for RAID-5 and computing the Parity Cycle length as the Lowest Common Multiple [11] of `pfl_num_comps` and P, divided by P, as described below. Note: This algorithm can be used also for RAID-5 where  $P=1$ .

P: number of parity devices  
 $P = 2$

PC: Parity cycle:  
 $PC = LCM(W, P) / P$

Q: The device index holding the Q component  
 (I is as computed in the above equations for RAID-5)  
 $Qdev = (I + 1) \% W$

#### 5.2.6. RAID Usage and Implementation Notes

RAID layouts with redundant data in their stripes require additional serialization of updates to ensure correct operation. Otherwise, if two clients simultaneously write to the same logical range of an object, the result could include different data in the same ranges of mirrored tuples, or corrupt parity information. It is the responsibility of the metadata server to enforce serialization requirements such as this. For example, the metadata server may do so by not granting overlapping write layouts within mirrored objects.

Many alternative encoding schemes exist for  $P \geq 2$  [12]. These involve  $P$  or  $Q$  equations different than those used in PFSP\_RAID\_PQ. Thus, if one of these schemes is to be used in the future, a distinct value must be added to `pnfs_ff_striping_pattern` for it. While Reed-Solomon codes are well understood, recently discovered schemes such as Liberation codes are more computationally efficient for small `group_widths`, and Cauchy Reed-Solomon codes are more computationally efficient for higher values of  $P$ .

#### 5.3. Mirroring

The `pfl_mirror_cnt` is used to replicate a file by replicating its Component Objects. If there is no mirroring, then `pfs_mirror_cnt` MUST be 0. If `pfl_mirror_cnt` is greater than zero, then the size of the `pfl_comps` array MUST be a multiple of  $(\text{pfl\_mirror\_cnt} + 1)$ . Thus, for a classic mirror on two objects, `pfl_mirror_cnt` is one. Note that mirroring can be defined over any striping pattern.

Replicas are adjacent in the `olo_components` array, and the value  $C$  produced by the above equations is not a direct index into the `pfl_comps` array. Instead, the following equations determine the replica component index  $RC_i$ , where  $i$  ranges from 0 to `pfl_mirror_cnt`.

$FW = \text{size of pfl\_comps array} / (\text{pfl\_mirror\_cnt} + 1)$

$C = \text{component index for striping or two-level striping}$   
as calculated using above equations

$i$  ranges from 0 to `pfl_mirror_cnt`, inclusive  
 $RC_i = C * (\text{pfl\_mirror\_cnt} + 1) + i$

#### 6. Recovering from Client I/O Errors

The pNFS client may encounter errors when directly accessing the Data Servers. However, it is the responsibility of the Metadata Server to recover from the I/O errors. When the LAYOUT4\_FLEX\_FILES layout type

is used, the client MUST report the I/O errors to the server at LAYOUTRETURN time using the `pflr_ioerr4` structure (see Section 7.1).

The metadata server analyzes the error and determines the required recovery operations such as repairing any parity inconsistencies, recovering media failures, or reconstructing missing objects.

The metadata server SHOULD recall any outstanding layouts to allow it exclusive write access to the stripes being recovered and to prevent other clients from hitting the same error condition. In these cases, the server MUST complete recovery before handing out any new layouts to the affected byte ranges.

Although it MAY be acceptable for the client to propagate a corresponding error to the application that initiated the I/O operation and drop any unwritten data, the client SHOULD attempt to retry the original I/O operation by requesting a new layout using LAYOUTGET and retry the I/O operation(s) using the new layout, or the client MAY just retry the I/O operation(s) using regular NFS READ or WRITE operations via the metadata server. The client SHOULD attempt to retrieve a new layout and retry the I/O operation using the Data Server first and only if the error persists, retry the I/O operation via the metadata server.

## 7. Flexible Files Layout Return

`layoutreturn_file4` is used in the LAYOUTRETURN operation to convey layout-type specific information to the server. It is defined in the NFSv4.1 [3] as follows:

```

struct layoutreturn_file4 {
    offset4      lrf_offset;
    length4      lrf_length;
    stateid4     lrf_stateid;
    /* layouttype4 specific data */
    opaque       lrf_body<>;
};

union layoutreturn4 switch(layoutreturn_type4 lr_returntype) {
    case LAYOUTRETURN4_FILE:
        layoutreturn_file4      lr_layout;
    default:
        void;
};

struct LAYOUTRETURN4args {
    /* CURRENT_FH: file */
    bool                lora_reclaim;
    layoutreturn_stateid lora_recallstateid;
    layouttype4         lora_layout_type;
    layoutiomode4       lora_iomode;
    layoutreturn4       lora_layoutreturn;
};

```

If the `lora_layout_type` layout type is `LAYOUT4_FLEX_FILES`, then the `lrf_body` opaque value is defined by the `pnfs_ff_layoutreturn4` type.

The `pnfs_ff_layoutreturn4` type allows the client to report I/O error information or layout usage statistics back to the metadata server as defined below.

#### 7.1. pflr\_errno

```

/// enum pflr_errno {
///     PNFS_FF_ERR_EIO                = 1,
///     PNFS_FF_ERR_NOT_FOUND          = 2,
///     PNFS_FF_ERR_NO_SPACE           = 3,
///     PNFS_FF_ERR_BAD_STATEID        = 4,
///     PNFS_FF_ERR_NO_ACCESS          = 5,
///     PNFS_FF_ERR_UNREACHABLE        = 6,
///     PNFS_FF_ERR_RESOURCE           = 7
/// };
///

```

`pflr_errno4` is used to represent error types when read/write errors are reported to the metadata server. The error codes serve as hints to the metadata server that may help it in diagnosing the exact

reason for the error and in repairing it.

- o PNFS\_FF\_ERR\_EIO indicates the operation failed because the Data Server experienced a failure trying to access the object. The most common source of these errors is media errors, but other internal errors might cause this as well. In this case, the metadata server should go examine the broken object more closely; hence, it should be used as the default error code.
- o PNFS\_FF\_ERR\_NOT\_FOUND indicates the object ID specifies a Component Object that does not exist on the Data Server.
- o PNFS\_FF\_ERR\_NO\_SPACE indicates the operation failed because the Data Server ran out of free capacity during the operation.
- o PNFS\_FF\_ERR\_BAD\_STATEID indicates the stateid is not valid.
- o PNFS\_FF\_ERR\_NO\_ACCESS indicates the rpc credentials do not allow the requested operation. This may happen when the client is fenced off. The client will need to return the layout and get a new one with fresh credentials.
- o PNFS\_FF\_ERR\_UNREACHABLE indicates the client did not complete the I/O operation at the Data Server due to a communication failure. Whether or not the I/O operation was executed by the Data Server is undetermined.
- o PNFS\_FF\_ERR\_RESOURCE indicates the client did not issue the I/O operation due to a local problem on the initiator (i.e., client) side, e.g., when running out of memory. The client MUST guarantee that the Data Server WRITE operation was never sent.

## 7.2. pnfs\_ff\_ioerr

```
/// struct pnfs_ff_ioerr {  
///     deviceid4      ioe_deviceid;  
///     nfs_fh4         ioe_fhandle;  
///     offset4         ioe_comp_offset;  
///     length4         ioe_comp_length;  
///     bool            ioe_iswrite;  
///     pnfs_ff_errno   ioe_errno;  
/// };  
///
```

The `pnfs_ff_ioerr4` structure is used to return error indications for Component Objects that generated errors during data transfers. These are hints to the metadata server that there are problems with that object. For each error, "ioe\_deviceid", "ioe\_fhandle",

"ioe\_comp\_offset", and "ioe\_comp\_length" represent the Component Object and byte range within the object in which the error occurred; "ioe\_iswrite" is set to "true" if the failed Data Server operation was data modifying, and "ioe\_errno" represents the type of error.

Component byte ranges in the optional `pnfs_ff_ioerr4` structure are used for recovering the object and MUST be set by the client to cover all failed I/O operations to the component.

### 7.3. `pnfs_ff_iostats`

```
/// struct pnfs_ff_iostats {  
///     offset4      ios_offset;  
///     length4      ios_length;  
///     uint32_t      ios_duration;  
///     uint32_t      ios_rd_count;  
///     uint64_t      ios_rd_bytes;  
///     uint32_t      ios_wr_count;  
///     uint64_t      ios_wr_bytes;  
/// };  
///
```

With pNFS, the data transfers are performed directly between the pNFS client and the data servers. Therefore, the metadata server has no visibility to the I/O stream and cannot use any statistical information about client I/O to optimize data storage location. `pnfs_ff_iostats4` MAY be used by the client to report I/O statistics back to the metadata server upon returning the layout. Since it is infeasible for the client to report every I/O that used the layout, the client MAY identify "hot" byte ranges for which to report I/O statistics. The definition and/or configuration mechanism of what is considered "hot" and the size of the reported byte range is out of the scope of this document. It is suggested for client implementation to provide reasonable default values and an optional run-time management interface to control these parameters. For example, a client can define the default byte range resolution to be 1 MB in size and the thresholds for reporting to be 1 MB/second or 10 I/O operations per second. For each byte range, `ios_offset` and `ios_length` represent the starting offset of the range and the range length in bytes. `ios_duration` represents the number of seconds the reported burst of I/O lasted. `ios_rd_count`, `ios_rd_bytes`, `ios_wr_count`, and `ios_wr_bytes` represent, respectively, the number of contiguous read and write I/Os and the respective aggregate number of bytes transferred within the reported byte range.

#### 7.4. pnfs\_ff\_layoutreturn

```
/// struct pnfs_ff_layoutreturn {  
///     pnfs_ff_ioerr          pflr_ioerr_report<>;  
///     pnfs_ff_iostats        pflr_iostats_report<>;  
/// };  
///
```

When object I/O operations failed, "pflr\_ioerr\_report<>" is used to report these errors to the metadata server as an array of elements of type `pnfs_ff_ioerr4`. Each element in the array represents an error that occurred on the Component Object identified by `<ioe_deviceid, ioe_fhandle>`. If no errors are to be reported, the size of the `pflr_ioerr_report<>` array is set to zero. The client MAY also use "pflr\_iostats\_report<>" to report a list of I/O statistics as an array of elements of type `pnfs_ff_iostats4`. Each element in the array represents statistics for a particular byte range. Byte ranges are not guaranteed to be disjoint and MAY repeat or intersect.

#### 8. Flexible Files Creation Layout Hint

The `layouthint4` type is defined in the NFSv4.1 [3] as follows:

```
struct layouthint4 {  
    layouttype4          loh_type;  
    opaque                loh_body<>;  
};
```

The `layouthint4` structure is used by the client to pass a hint about the type of layout it would like created for a particular file. If the `loh_type` layout type is `LAYOUT4_FLEX_FILES`, then the `loh_body` opaque value is defined by the `pnfs_ff_layouthint` type.



## 8.1. pnfs\_ff\_layouthint

```

    /// union pnfs_ff_max_comps_hint switch (bool pfmv_valid) {
    ///     case TRUE:
    ///         uint32_t                omx_max_comps;
    ///     case FALSE:
    ///         void;
    /// };
    ///
    /// union pnfs_ff_stripe_unit_hint switch (bool pfsu_valid) {
    ///     case TRUE:
    ///         length4                osu_stripe_unit;
    ///     case FALSE:
    ///         void;
    /// };
    ///
    /// union pnfs_ff_mirror_cnt_hint switch (bool pfmc_valid) {
    ///     case TRUE:
    ///         uint32_t                omc_mirror_cnt;
    ///     case FALSE:
    ///         void;
    /// };
    ///
    /// union pnfs_ff_striping_pattern_hint switch (bool pfsp_valid) {
    ///     case TRUE:
    ///         pnfs_ff_striping_pattern    pfsp_striping_pattern;
    ///     case FALSE:
    ///         void;
    /// };
    ///
    /// struct pnfs_ff_layouthint {
    ///     pnfs_ff_max_comps_hint        pflh_max_comps_hint;
    ///     pnfs_ff_stripe_unit_hint      pflh_stripe_unit_hint;
    ///     pnfs_ff_mirror_cnt_hint       pflh_mirror_cnt_hint;
    ///     pnfs_ff_striping_pattern_hint pflh_striping_pattern_hint;
    /// };
    ///

```

This type conveys hints for the desired data map. All parameters are optional so the client can give values for only the parameters it cares about, e.g. it can provide a hint for the desired number of mirrored components, regardless of the striping pattern selected for the file. The server should make an attempt to honor the hints, but it can ignore any or all of them at its own discretion and without failing the respective CREATE operation.

## 9. Recalling Layouts

The Flexible Files metadata server should recall outstanding layouts in the following cases:

- o When the file's security policy changes, i.e., Access Control Lists (ACLs) or permission mode bits are set.
- o When the file's layout changes, rendering outstanding layouts invalid.
- o When there are sharing conflicts. For example, the server will issue stripe-aligned layout segments for RAID-5 objects. To prevent corruption of the file's parity, multiple clients must not hold valid write layouts for the same stripes. An outstanding READ/WRITE (RW) layout should be recalled when a conflicting LAYOUTGET is received from a different client for LAYOUTIOMODE4\_RW and for a byte range overlapping with the outstanding layout segment.

### 9.1. CB\_RECALL\_ANY

The metadata server can use the CB\_RECALL\_ANY callback operation to notify the client to return some or all of its layouts. The NFSv4.1 [3] defines the following types:

```
const RCA4_TYPE_MASK_FF_LAYOUT_MIN      = -2;
const RCA4_TYPE_MASK_FF_LAYOUT_MAX      = -1;
[[RFC Editor: please insert assigned constants]]
```

```
struct CB_RECALL_ANY4args {
    uint32_t      craa_objects_to_keep;
    bitmap4       craa_type_mask;
};
```

Typically, CB\_RECALL\_ANY will be used to recall client state when the server needs to reclaim resources. The `craa_type_mask` bitmap specifies the type of resources that are recalled and the `craa_objects_to_keep` value specifies how many of the recalled objects the client is allowed to keep. The Flexible Files layout type mask flags are defined as follows. They represent the iomode of the recalled layouts. In response, the client SHOULD return layouts of the recalled iomode that it needs the least, keeping at most `craa_objects_to_keep` object-based layouts.

```
/// enum pnfs_ff_cb_recall_any_mask {  
///     PNFS_FF_RCA4_TYPE_MASK_READ = -2,  
///     PNFS_FF_RCA4_TYPE_MASK_RW   = -1  
[[RFC Editor: please insert assigned constants]]  
/// };  
///
```

The PNFS\_FF\_RCA4\_TYPE\_MASK\_READ flag notifies the client to return layouts of iomode LAYOUTIOMODE4\_READ. Similarly, the PNFS\_FF\_RCA4\_TYPE\_MASK\_RW flag notifies the client to return layouts of iomode LAYOUTIOMODE4\_RW. When both mask flags are set, the client is notified to return layouts of either iomode.

## 10. Client Fencing

In cases where clients are uncommunicative and their lease has expired or when clients fail to return recalled layouts within a lease period at the least (see "Recalling a Layout"[3]), the server MAY revoke client layouts and/or device address mappings and reassign these resources to other clients. To avoid data corruption, the metadata server MUST fence off the revoked clients from the respective objects as described in Section 2.1.

## 11. Security Considerations

The pNFS extension partitions the NFSv4 file system protocol into two parts, the control path and the data path (storage protocol). The control path contains all the new operations described by this extension; all existing NFSv4 security mechanisms and features apply to the control path. The combination of components in a pNFS system is required to preserve the security properties of NFSv4 with respect to an entity accessing data via a client, including security countermeasures to defend against threats that NFSv4 provides defenses for in environments where these threats are considered significant.

The metadata server enforces the file access-control policy at LAYOUTGET time. The client should use suitable authorization credentials for getting the layout for the requested iomode (READ or RW) and the server verifies the permissions and ACL for these credentials, possibly returning NFS4ERR\_ACCESS if the client is not allowed the requested iomode. If the LAYOUTGET operation succeeds the client receives, as part of the layout, a set of credentials allowing it I/O access to the specified objects corresponding to the requested iomode. When the client acts on I/O operations on behalf of its local users, it MUST authenticate and authorize the user by

issuing respective OPEN and ACCESS calls to the metadata server, similar to having NFSv4 data delegations. If access is allowed, the client uses the corresponding (READ or RW) credentials to perform the I/O operations at the object storage devices. When the metadata server receives a request to change a file's permissions or ACL, it SHOULD recall all layouts for that file and it MUST fence off the clients holding outstanding layouts for the respective file by implicitly invalidating the outstanding credentials on all Component Objects comprising before committing to the new permissions and ACL. Doing this will ensure that clients re-authorize their layouts according to the modified permissions and ACL by requesting new layouts. Recalling the layouts in this case is courtesy of the server intended to prevent clients from getting an error on I/Os done after the client was fenced off.

## 12. Striping Topologies Extensibility

New striping topologies that are not specified in this document may be specified using @@@. These must be documented in the IETF by submitting an RFC augmenting this protocol provided that:

- o New striping topologies MUST be wire-protocol compatible with the Flexible Files Layout protocol as specified in this document.
- o Some members of the data structures specified here may be declared as optional or mandatory-not-to-be-used.
- o Upon acceptance by the IETF as a RFC, new striping topology constants MUST be registered with IANA (Section 13).

## 13. IANA Considerations

As described in NFSv4.1 [3], new layout type numbers have been assigned by IANA. This document defines the protocol associated with the existing layout type number, LAYOUT4\_FLEX\_FILES.

A new IANA registry should be assigned to register new data map striping topologies described by the enumerated type: @@@.

## 14. Normative References

- [1] IETF, "NFS Version 3 Protocol Specification", RFC 1813, June 1995.
- [2] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.

- [3] Shepler, S., Ed., Eisler, M., Ed., and D. Noveck, Ed., "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.
- [4] Halevy, B., Ed., Welch, B., Ed., and J. Zelenka, Ed., "Object-Based Parallel NFS (pNFS) Operations", RFC 5664, January 2010.
- [5] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [6] Eisler, M., "XDR: External Data Representation Standard", STD 67, RFC 4506, May 2006.
- [7] Shepler, S., Ed., Eisler, M., Ed., and D. Noveck, Ed., "Network File System (NFS) Version 4 Minor Version 1 External Data Representation Standard (XDR) Description", RFC 5662, January 2010.
- [8] IETF Trust, "Legal Provisions Relating to IETF Documents", November 2008,  
<<http://trustee.ietf.org/docs/IETF-Trust-License-Policy.pdf>>.
- [9] MacWilliams, F. and N. Sloane, "The Theory of Error-Correcting Codes, Part I", 1977.
- [10] Anvin, H., "The Mathematics of RAID-6", May 2009,  
<<http://kernel.org/pub/linux/kernel/people/hpa/raid6.pdf>>.
- [11] The free encyclopedia, Wikipedia., "Least common multiple", April 2011,  
<[http://en.wikipedia.org/wiki/Least\\_common\\_multiple](http://en.wikipedia.org/wiki/Least_common_multiple)>.
- [12] Plank, James S., and Luo, Jianqiang and Schuman, Catherine D. and Xu, Lihao and Wilcox-O'Hearn, Zooko, "A Performance Evaluation and Examination of Open-source Erasure Coding Libraries for Storage", 2007.

#### Appendix A. Acknowledgments

The pNFS Objects Layout was authored and revised by Brent Welch, Jim Zelenka, Benny Halevy, and Boaz Harrosh.

Those who provided miscellaneous comments to early drafts of this document include: Matt W. Benjamin, Adam Emerson, Tom Haynes, J. Bruce Fields, and Lev Solomonov.

Author's Address

Benny Halevy  
PrimaryData, Inc.

Email: [bhalevy@primarydata.com](mailto:bhalevy@primarydata.com)

URI: <http://www.primarydata.com>



NFSv4  
Internet-Draft  
Intended status: Informational  
Expires: March 13, 2014

D. Noveck  
EMC  
September 09, 2013

NFS Protocol Extension: Retrospect and Prospect  
draft-dnoveck-nfs-extension-00

Abstract

This document surveys the processes by which the NFS protocol has been extended in the past and considers how the mechanisms by which the protocol is extended might best be modified in the future.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on March 13, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.



## Table of Contents

1. Introduction . . . . .	2
2. Protocol Extension . . . . .	3
3. Protocol Extension Mechanisms . . . . .	3
3.1. Specific Protocol Mechanisms Designed for Extension . . . . .	4
3.2. Protocol Extension by XDR Replacement . . . . .	4
3.3. Protocol Extension by XDR Extension . . . . .	5
3.4. Combination of Protocol Extension Mechanisms . . . . .	6
4. Pre-IETF NFS Versioning . . . . .	6
4.1. The Pre-IETF NFS Environment . . . . .	6
4.2. Transition from NFSv2 to NFSv3 . . . . .	7
5. NFS Versioning (so far) Within IETF . . . . .	7
5.1. Transition from NFSv3 to NFSv4 . . . . .	7
5.2. Initial Minor Versioning Model for NFSv4 . . . . .	8
5.3. Transition from NFSv4.0 to NFSv4.1 . . . . .	10
5.4. Transition from NFSv4.1 to NFSv4.2 . . . . .	11
5.5. Evolution of Minor Versioning Model within NFSv4 . . . . .	12
5.6. Current Minor Versioning Model for NFSv4 . . . . .	13
5.7. Review of NFSv4 Versioning so far . . . . .	14
6. NFSv4 Versioning Now . . . . .	15
6.1. Current NFS Versioning Practices . . . . .	15
6.2. Problems with Current NFS Versioning Approach . . . . .	15
7. Going Forward with a New NFSv4 Extension Approach . . . . .	17
7.1. Extension Mechanisms used for Protocol Updates . . . . .	17
7.2. Requirements for a New NFSv4 Extension Approach . . . . .	18
7.3. Principles upon which to Base a New NFSv4 Extension Approach . . . . .	19
7.4. Work Going Forward in Creating a New NFSv4 Extension Approach . . . . .	20
8. Security Considerations . . . . .	21
9. IANA Considerations . . . . .	21
10. Acknowledgements . . . . .	21
11. References . . . . .	21
11.1. Normative References . . . . .	21
11.2. Informative References . . . . .	22
Author's Address . . . . .	23

## 1. Introduction

This document examines the subject of protocol extension within the NFS family of protocols. In order to better understand the issues that exist going forward with NFSv4, we examine the history of protocol extension throughout the development of NFS including both the pre-IETF period and the development of successive NFSv4 minor versions.

We then use this history as a basis upon which to explore the issues involved in providing a modified extension paradigm that builds on the work already done, but is more flexible.

## 2. Protocol Extension

Often, protocols require means by which they can be extended. Such extension may be needed to meet new requirements, to correct protocol weaknesses exposed by experience, or even to correct protocol bugs (as can happen when protocols are published as RFC's without fully fleshed-out implementations).

We need to distinguish here between protocol "extension" and "versioning". Versioning is a form of protocol extension but not every form of protocol extension can be accommodated within a versioning paradigm.

When a versioning paradigm is in place, groups of extensions are conceived of as ordered, allowing extensions in subsequent versions to build upon those in previous versions. When multiple extensions are combined into a single version, each of the extensions may be built assuming that the others will be present as well. In such cases, there can be the opportunity to make design changes in the protocol, allowing elements of the protocol to be restructured, sometimes in major ways.

When a versioning paradigm is in effect and extensions are optional, extensions cannot build upon one another, since the presence of any particular extension cannot be assumed. In such cases, the ability to restructure the protocol is reduced, but smaller changes may be introduced more easily.

In this latter case, it is not clear that the word "versioning" is appropriate. Nevertheless, in this document, we will, as in the phrase "NFSv4 minor versioning" use the existing terminology without necessarily subscribing to the view that "versioning" is the appropriate description.

## 3. Protocol Extension Mechanisms

Some factors that are often relevant in deciding on the means by which a protocol will be extended.

- o Whether extensions are to be individually selectable (i.e. optional) or assumed to be always present, allowing one to build upon another earlier one.
- o The size and scope of extensions that will be made.

- o Compatibility issues with existing implementations.
- o Issues that relate to ensuring that when individual extensions, separately arrived at, are each optionally allowed, that the ones used are compatible and can be used effectively together.
- o The overall implementation framework. For example, RPC-based protocols may do extension by means of the RPC version mechanism.

While it is possible to use different sorts of extension mechanisms for different sorts of extensions, protocols typically do not take advantage of that flexibility.

On the other hand, protocols do, as NFS has done, change their preferred extension mechanisms in response to long-term changes in requirements. However, once having done so, they rarely switch back. Changing extension mechanisms is a big step, both conceptually and in implementation terms, and is not frequently repeated.

### 3.1. Specific Protocol Mechanisms Designed for Extension

Often, protocols will be designed with specific mechanisms, designed to allow protocol extension. An example is the provision for TCP options (see [RFC0793] and [RFC2780].) Most often, such mechanisms are designed to allow individual extensions to be designed and implemented independently, with any dependency relations between extensions specified separately and not enforced by the extension mechanism itself.

### 3.2. Protocol Extension by XDR Replacement

RPC-based protocols may, and often do, provide for protocol extension by replacing the XDR for one version with that for another and using the RPC versioning mechanism to manage selection of the proper protocol variant. The use of the RPC versioning mechanism enforces a versioning paradigm of this sort on protocols using this extension mechanism.

This extension mechanism allows very extensive protocol changes, up to and including the replacement of one protocol by an entirely different one. For some kinds of protocol extensions, this seems the only way to effect the change.

### 3.3. Protocol Extension by XDR Extension

It is possible to replace an XDR definition by one which is an extension in the sense that

- o The set of messages described by the second definition is a superset of that described by the first.
- o Each message within the set of messages described by the first definition is recognized as having exact the same structure/interpretation by the second definition.
- o Each message within the set of messages described by the second definition but not the first definition must be recognized as part of an unsupported extension.

Within an XDR/RPC framework, extensions can be arrived at by:

- o Addition of previously unspecified RPC requests.
- o Addition of new, previously unused, values to existing enums.
- o Addition of previously unassigned bit values to a flag word.
- o Addition of new cases to existing switches, provided that the existing switch did not contain a default case.

Such an extension relation between XDR descriptions is reflexive and transitive and thus defines a partial order. In practice, provisions have to be made to make sure that two extensions of the same description are compatible (i.e. either one is an extension of the other, or there is a another description that is a valid extension of both).

To put things in concrete terms, such compatibility can be assured if measures are taken to ensure:

- o That the same request number is not used for two different requests.
- o That the same enum value is not assigned two different meanings.
- o That the same bit flag value is not assigned two different meanings.
- o That whenever the same case value is added to the same switch in two different extensions, the content assigned to the two matching added cases is the same.

- o That default cases are never added to existing switches.

### 3.4. Combination of Protocol Extension Mechanisms

It is possible to use multiple of these means of protocol extension simultaneously. When this is done, the result is a composite extension mechanism. For example, if the XDR replacement or XDR extension mechanism is adopted, a protocol-specific mechanism can be added to it, if the protocol-specific mechanism is built on objects whose XDR definition is sufficiently generic. (e.g. opaque arrays or feature bitmasks).

It can be argued that the NFSv4 attribute model provides such an embedded protocol-specific extension mechanism, since sets of attribute values are specified as XDR opaque arrays and attribute sets are specified as variable-length arrays of 32-bit integers allowing new attribute bits to be defined outside of the XDR definition framework.

Note that there exists specification text that suggests that attributes are part of the XDR specification, making it hard to reach a firm conclusion on the matter. However, the resolution of this question does not affect the other matters discussed below, since, in either case, we have an extension mechanism that allows optional extensions.

## 4. Pre-IETF NFS Versioning

### 4.1. The Pre-IETF NFS Environment

NFSv2 and NFSv3 were described by the informational RFC's [RFC1094] and [RFC1813]. These documents each described existing interoperating client and server implementations. Thus they started with running code. If there was a rough consensus in effect, it was that these were useful protocols to use and thus that someone building a client or server had to interoperate with the existing implementations.

The following characteristics of protocol development during that period are noteworthy.

- o Most client implementations were implemented on very similar systems, in terms of the API's supported and many specifics of the local filesystems exported by servers.
- o Often, the important client and server implementations were done by the same organization.

As a result of these commonalities, specifications tended to avoid a lot of detail that would have been required in a more diverse environment. New features were thought of in terms of generally understood client and server implementation frameworks and it was generally clear which of those could be implemented without markedly changing that framework.

#### 4.2. Transition from NFSv2 to NFSv3

There were a number of significant changes involved, but only the first two were of major importance.

- o Converting file sizes and offsets from 32-bit to 64-bit unsigned integers.
- o Support for uncommitted WRITES and the COMMIT request.
- o Provision for WRITE to return atomic pre- and post-write file attributes, in order to allow a client to determine whether another client was writing the file.

Interestingly enough, this feature was not carried over into NFSv4.

- o The REaddirPLUS request.
- o The addition of NFS3ERR\_JUKEBOX (the precursor of NFS4ERR\_DELAY).

Of these only the first actually needed something as drastic as the XDR replacement model. The others could have been handled simply by adding new RPC requests and an enum value to an existing NFSv2 XDR. Since, NFS's extension mechanism was then XDR replacement, such choices were not available.

### 5. NFS Versioning (so far) Within IETF

#### 5.1. Transition from NFSv3 to NFSv4

NFSv4 was the first NFS version published as a Standards track document. Although an initial [RFC3010], entitled "NFS version 4 protocol" was published as a Proposed Standard, it was never implemented due to issues with the design of locking.

Subsequently, [RFC3530], entitled "Network File System (NFS) version 4 Protocol" was published as a Proposed Standard, obsoleting [RFC3010]. Currently, there are bis documents, [RFC3530bis] and [RFC3530bis-dotx], nearing publication.

The set of changes made to create NFSv4 was larger by far than that for NFSv3. A partial list follows.

- o Conversion to a stateful protocol, including support for locking. Locking included support for OPENS (with share reservations), byte-range locking (optionally including mandatory byte-range locks) and delegation.
- o The COMPOUND operation.
- o Conversion to a bi-directional protocol, by the addition of (optional) callbacks.
- o Internationalization.
- o Support for filesystems doing case-insensitive name matching.
- o A new, extensible file attribute model, including support for acls, and conversion of user and group to a string model, opening the way for multi-domain support.
- o Support for named attributes.
- o Merger of ancillary protocols (e.g. MOUNT, NSM, NLM) into the NFS Protocol proper.
- o Support for crossing of filesystem boundaries within a server's file name space (originally done for the incorporation of MOUNT functionality).
- o Support for such multi-server operations as migration, replication, and referral.

Referrals were not explicitly mentioned in [RFC3530] and are explained in [RFC3530bis].

- o Creation of a minor versioning model (to be discussed in Section 5.2) to allow further protocol extension.

These features/extensions were implemented via the XDR replacement model. This was the only realistic alternative, not only because of the size of the list, but also because some of the changes undercut some central design elements of the pre-IETF NFS protocol.

## 5.2. Initial Minor Versioning Model for NFSv4

The minor versioning model for NFSv4 is an XDR extension model. It was presented within a versioning paradigm but the fact that all the

added features were to be (at least initially) optional indicated that features were intended to be built independently, and that clients were expected to deal with their presence or absence. Note that the term "features" is not explicitly defined. We assume that the definition includes operations within COMPOUND or CB\_COMPOUND, attributes, flag bits enum values, and new cases of XDR switch definitions.

Now let's look at some specifics of the minor version rules established for NFSv4 in [RFC3530]. Note that some of these were significantly modified by [RFC5661] and [NFSv42], as discussed in Section 5.6.

- o No RPC requests may be added. Thus COMPOUND (and NULL) are to be the only requests within all NFSv4 minor versions.

Similarly for callbacks, CB\_COMPOUND and CB\_NULL are the only requests within callback program.

- o The arguments for COMPOUND and CB\_COMPOUND contain a 32-bit minorversion field.

Although this field is part of the minor versioning paradigm, it is not clear how useful it is, as long as all extensions are optional. For a more detailed discussion of this issue, see Section 5.6.

- o Features may be defined as optional, recommended, or mandatory.

These designations apply to implementation by the server. For clients, no operations are mandatory, although it is hard to imagine an NFSv4.0 client that does not use PUTFH or SETCLIENTID, for example.

- o Features may be upgraded or downgraded along the optional/recommended/mandatory scale.
- o Features may be declared "mandatory to not implement". This allows the deletion of a feature while retaining as reserved the value previously assigned.
- o Clients and servers that support a particular minor version must support all previous minor versions as well.
- o New features may not be designated as mandatory in the minor version in which they are introduced.



- o Clients are not allowed to use stateids, filehandles, or similar returned objects from the COMPOUND procedure with one minor within another COMPOUND procedure with a different value of the minorversion field.

This model was subsequently modified in [RFC5661] and in [NFSv42]. See Section 5.3 and Section 5.4 for details.

Many of the events anticipated in the model presented above have never been realized and it may be that they never will be realized. See Section 5.5 for some details. Examples are:

- o There have never been recommended operations.
- o There have never been optional attributes.
- o Features have never been upgraded or downgraded in the transition between minor versions.

### 5.3. Transition from NFSv4.0 to NFSv4.1

NFSv4.1 made a major change to NFSv4.0. It was able to do so using an XDR extension model although it did not follow the rules laid out in Section 5.2. Specifically, some features were declared "infrastructural" and thus mandatory upon introduction.

Note that at the same time, the requirement that clients and servers support previous minor versions changed from a "must" to a "SHOULD". Presumably, this change reflects the fact that a minor version with substantial infrastructural changes is essentially a new protocol, making the "must" seem dubious. Whether the "SHOULD" here meets the requirements of [RFC2119] needs to be explored.

NFSv4.1 was described in [RFC5661] and [RFC5662], each of which was published as a Proposed Standard.

The following features were added as infrastructural features.

- o Support for a sessions model including support for EOS (exactly-once semantics).

Note that COMPOUND was taken advantage of to avoid adding slot and sequence information to the request header. Instead this information is packaged in a SEQUENCE or CB\_SEQUENCE operation at the start of the COMPOUND or CB\_COMPOUND.

- o A new set of operations were added which enable the client and server to identify themselves to one another.

Although these are often thought of as part of the sessions model, in fact they are logically distinct.

- o RECLAIM\_COMPLETE to allow better server sequencing of lock reclaim operations.

There also a number of optional features.

- o Parallel NFS
- o WANT\_DELEGATION to allow delegations to be obtained apart from opens.
- o Directory delegations and notifications.
- o The FS\_LOCATIONS\_INFO and FS\_STATUS attributes.

Note that there has been little implementation work on the last two of these.

Parallel NFS created an alternate protocol extension mechanism for NFS. New pNFS mapping types could be added. Existing mapping types might have their own extension mechanisms. There also exists the possibility that features might be added within the NFSv4 protocol proper, designed to, or capable of, interacting with particular mapping types. This document will not address these issues but eventually, the NFSv4 Protocol will have to deal with them.

#### 5.4. Transition from NFSv4.1 to NFSv4.2

While NFSv4.2 has not been defined in an RFC, it is fairly close to completion. The descriptions in [NFSv42] and [NFSv42-dotx] can serve as useful references.

The following features (all optional) are provided for in NFSv4.2:

- o Support for labeled NFS.
- o Server-side copy.
- o An operation fence option on EXCHANGE\_ID.
- o Application data holes (formerly application data blocks).
- o Disk-space reservation (nominally "recommended" since it is implemented by an attribute and attributes have never been declared "optional").

- o Hole-punching operations.
- o READ\_PLUS

Note that there are two piece of infrastructure that are used by multiple features above. These are not "infrastructural" in the sense mentioned in Section 5.3 (i.e. they are not mandatory), but they do serve an infrastructural role in that are required to be present if one of the optional features that use them are supported.

- o WRITE\_PLUS used to implement (ordinary) hole punching and application data holes.
- o OFFLOAD operations/callbacks used to support WRITE\_PLUS and server-side copy.

#### 5.5. Evolution of Minor Versioning Model within NFSv4

As noted above, there have been changes made by [RFC5661] and [NFSv42] in the NFSV4 minor versioning model.

- o NFSv4.1 (in [RFC5661] introduced the concept of "infrastructural" features (i.e. those defined as mandatory at initial introduction).
- o NFSv4.2 (in [NFSv42] added the concept of feature obsolescence allowing implementers to get early notice of the expectation that some features are on the path to becoming mandatory-to-not-implement.

With these changes, we can classify potential minor versions, starting with those that currently exist.

- o Minor version zero which introduced a new (major) version of the NFS protocol. All of the operations within it are new and a subset are effectively mandatory.
- o Minor versions which introduce a new operation and make it mandatory (based on its being infrastructural).

Currently, the only such version is minor version one, although there may be others in the future.

- o All other minor versions. These add only optional/recommended features, each present or absent on the server with clients needing to be able to deal individually with their presence or absence.

Currently, the only such version is minor version two. It is likely there will be others in the future and it may be that all future minor versions will be of this character.

We term versions in the first two categories "infrastructure-level versions". Such versions form an ascending sequence in which the difference between, for example, NFSv4.0 and NFSv4.1, is very similar to the difference between NFSv2 and NFSv3. Clients may be designed for one or the other, or for both but a client capable of interacting with both is really choosing between two different protocols.

We term versions in the last category "optional-feature-only versions".

Note that although the concept of optional features being upgraded to mandatory status remains, it is likely that it will not be used very much, if at all, in the future. The situation is similar for the case of features being downgraded to mandatory-to-not-implement.

Given the diversity of NFS clients and servers, it is highly unlikely that a new non-infrastructure feature will be so broadly necessary/desirable that a consensus to make it mandatory would be likely to arise. Such a decision would prevent servers not implementing such a feature from incorporating other later-developed features. It is only when a feature is judged so useful by users that people will not use servers without it, that adoption will become universal. At this point, a decision to make it mandatory would merely ratify what had already happened on its own.

Except in the case of a universally recognized mistake, any downgrading to mandatory-to-not-implement, would only happen when a replacement becomes mandatory so the considerations above make that situation equally unlikely to occur.

Still, it is possible that versions making such feature status changes will be created in the future. We will call any such "mandatory-feature-change" versions.

#### 5.6. Current Minor Versioning Model for NFSv4

Minor versions which are infrastructure-level or which are, mandatory-feature-change versions form an ascending sequence in which we also have a versioning paradigm, implemented using XDR extension.

Optional-feature-only versions are fundamentally different. Each NFSv4.2 server implements the same protocol as NFSv4.1 with a particular set of optional features beyond those that are mandatory. This set may range from the null set all the way to all of the

optional features. Here, it appears that the versioning paradigm is not appropriate to the reality of the extension mechanism.

As a way of illustrating the basic point here, let us consider two servers each of which only supports operations within NFSv4.1:

- o The first server "supports" NFSv4.2 but none of the optional features added in [NFSv4.2]. In this case, any attempt by a client to use one of those features will result in an NFS4ERR\_OPNOTSUPP being returned.
- o Let us say that the second server does not support NFSv4.2 and supports precisely the same set of features. In this case, a request will be rejected (with error NFS4ERR\_....) if its COMPOUND minorversion field is two and if the field is one, any unsupported NFSv4.2 operation will be rejected with NFS4ERR\_OP\_INVALID.

Although this obeys the rules as they stand, there is no real value for the client, the server, or the protocol in making these artificial distinctions. Optional-feature-only minor versions such as NFSv4.2 are not minor versions in the same sense that NFSv4.0 and NFSv4.1 are. In this case the minorversion field is not providing any information, while the set of operations supported is the important thing that the server implementer chooses and the client needs to know.

In later sections we will discuss how this mismatch might be best addressed as NFSv4 development proceeds.

## 5.7. Review of NFSv4 Versioning so far

To summarize protocol extension as it applies to the NFSV4 protocols:

- o NFSV4.0 was implemented using the XDR replacement approach inherited from NFSv2 and NFSv3. As was to be expected given the nature and scope of the changes, its development took considerable time.

It defined a protocol extension approach based on the XDR extension mechanism which was designed to enable future development of minor versions. However, this mechanism was not used as part of the implementation of NFSv4.0.

- o NFSV4.1 was implemented using the XDR extension mechanism. To implement sessions, it was forced to modify the extension approach in the only way that was viable in the circumstances. As a result, the specification process took a long time, since it made significant structural changes to the protocol and also because it

had to specify the entire protocol, and not just a set of extensions.

- o NFSv4.2 returned to the original XDR extension mechanism and was intended to be a small incremental update with a one-hundred page (or less) specification. The fact that this turned out to be a multi-year effort has occasioned concern and we will attempt to see how the process can be streamlined.

## 6. NFSv4 Versioning Now

### 6.1. Current NFS Versioning Practices

The following pattern was followed for NFSv4.2, and, unless changes are made, seems likely to persist.

- o Various features are sketched out in individual drafts
- o The working group reaches a decision (i.e. by rough consensus) as to the extensions/features to be included in the minor version.
- o The existing individual drafts are combined into a draft of a working group document intended to eventually become the RFC describing the new minor version.
- o This document goes through further refinement and cycles of working group document review. At some point a companion -dot-x document is prepared and reviewed by the working group as well.
- o The two documents go through working group last call, IESG review, and RFC publication.

This pattern of development is not a good fit for the kind of minor version that NFSv4.2 is and many future such minor versions will be. Such versions consist of a set of mostly unrelated features, each individually selectable or not by implementers, artificially yoked together. In essence, we have a "feature batch" rather than a minor version.

### 6.2. Problems with Current NFS Versioning Approach

A number of issues have been noted with the current process for NFSv4.2, leading to the conclusion that the process needs to be revised in some way for future minor versions, of the same sort.

- o It takes too long to get a minor version drafted and through working group IESG review. Despite the fact that NFSv4.2 was intended to be a fairly minimal minor version, describable in a

one-hundred-page spec, it looks like the pace of development is such that there will be about a four-year gap between the time that NFSv4.2 was started and an equivalent point for NFSv4.3, if that pattern is maintained.

- o We still do not have significant active implementations in which proposed last-minute protocol changes can be tested for validity. As an example of the problem, consider the decision to pass source stateids to the COPY op. If there were an implementation of inter-server server-side copy, the problems that this created (since stateids are tied to clientids in NFSv4.1 and beyond) would have quickly become manifest.
- o Many features within NFSv4.2 have not received the kind of searching review appropriate to this stage of specification development. Some examples are discussed below.

Some instances of problems/issues ascribable to a lack of searching document review:

- o The state of the IO hints feature is most unsatisfactory. It is not clear how, or even if, it is possible to specify in a way that interoperable clients and servers can be written.
- o It was the general understanding within the group that labeled NFS required use of RPCSEC\_GSSv3, while RPCSEC\_GSSv3 was not being worked on, and had little chance of being worked on.
- o The security for inter-server copy was specified to be dependent on RPCSEC\_GSSv3, yet, when it was found that RPCSEC\_GSSv3 was not on the horizon, it turned out that a simple alternative was available, and, the functionality needed for inter-server copy was not really anticipated for RPCSEC\_GSSv3 either.

If we look at the problems above, we can understand better how such problems can arise. In short, the decision as to what features to include within a minor version, is not a good use of the rough consensus model and in proceeding on that basis, the group created a set of perverse incentives that undercut the process. Also, as the process goes on for a long time, as is likely, these perverse incentives are intensified. Consider the following points:

- o It is not clear exactly what the consensus as to proposed minor version contents actually means. Working group members might interpret it as meaning "These features are worth pursuing and they should be pursued". However, if they thought the definition was more like, "each of these features is so important that, if it is not ready, any other feature, including the one I'm interested

in, should be delayed also", then it is hard to imagine any such rough consensus existing. Note that, given the minor versioning implementation laid out above (in Section 6.1), the latter definition is, for functional purposes, the effective meaning, of the minor version content consensus.

- o Given that many features are linked together, any delay in one feature, once it is accepted as part of the feature batch, delays all of the features, making it hard for people to comment forthrightly on any significant specification inadequacies. Not only will it delay your preferred feature, but if the problems are not fixed, the only recourse is an extreme penalty. As a result, it often seems not worth pursuing these sorts of issues.
- o As the version turnaround cycle is so long, it is very difficult to remove a feature from a minor version feature batch. Given that these are all features that have enough interest to be in the minor version, it is hard to kick the feature into the next minor version, given that will certainly mean a multi-year delay, even if the feature could be guaranteed admission to the next feature batch.
- o Given that responsibility for a minor version is transferred to the editor of minor version definition documents at an early stage, we have a process in which it is not clear who has responsibility to follow up on the work necessary other than the minor version editor who may not have the required time and expertise in all cases. There is not a designated feature owner with responsibility to make the feature happen.

## 7. Going Forward with a New NFSv4 Extension Approach

As we work to correct the issues noted above, and fill out the details of a modified extension paradigm, we will have to take note of the design considerations put forth in [RFC6709].

### 7.1. Extension Mechanisms used for Protocol Updates

It is generally held to be the case, that a document updating a minor version RFC, is not allowed to extend the XDR. Sometimes typedefs are added to make it clearer how particular fields are used. Also comments have been added and minimal reformatting done, but even addition of new error codes, as opposed to adding existing error codes to the list of those allowed to be returned by a given operation, has not been allowed.



Given that we have an XDR extension paradigm in place, it does not make sense to prohibit XDR extensions to be made in these update documents.

It should be noted that the prohibition of such XDR changes is not explicitly mentioned anywhere, to my knowledge. Rather, it seems to be a piece of NFSv4 versioning folklore which needs to be either justified or discarded.

Acknowledging this change would allow us to do the following sorts of things (in addition to the specification bug fixes now allowed) in bis documents and other documents which update minor version definition RFC's. While it would be theoretically possible to add entirely new features, working group and IESG review should keep additions limited to the following two sorts of items.

- o Adding new operations to correct protocol bugs, subject to the proviso that a server implementing a replacement operation must also support the operation replaced. In this case, uniqueness for values such as operation and attribute numbers would assured by defining them in a later minor version's XDR definition document, or some update thereof. If the bug fix does not apply to that later minor version, it can be treated there as mandatory-to-not-implement, to match what it replaces.
- o Backporting of features whose usefulness has been proven in a subsequent minor version and can be easily made available in an earlier minor version. In this case, values such as operation and attribute numbers are already assured of uniqueness, due to their assignment as unique in the later minor version.

Doing things this way would address the issues that have given rise to the perceived need for "micro-versioning". Note that the sorts of changes we would be making would not require any change in the minorversion field, at all.

## 7.2. Requirements for a New NFSv4 Extension Approach

The following requirements will govern construction of a possible new protocol extension approach for NFSv4.

- o That individual extensions not be documented together (i.e. in the same document) unless there are good reasons to do so (e.g. the extensions use common facilities not documented anywhere else, there is a dependency relationship among the extensions allowing one to be used only when others are available). This would allow for shorter documents and a less trying review process.

- o The process should allow for protocol bugs to be fixed, even if the problem is found after RFC publication. Just as we have the errata process to fix spec issues, we should be able to fix bugs/oversights in the XDR, as long as compatibility issues with existing implementations are addressed.
- o That the process gives the working group an appropriate opportunity to review extensions and reject those that are architecturally inappropriate.
- o That the process gives the working group and IESG an appropriate opportunity to review extension specifications and get them fixed, without adversely affecting other unrelated features.
- o That the process appropriately assigns the responsibility for making proposed extensions real on those proposing them. This should include, at an appropriate time, work to get "running code" (i.e. client and server implementations) to demonstrate implementation feasibility.

### 7.3. Principles upon which to Base a New NFSv4 Extension Approach

The following principles seem the best way to meet these requirements without a disruptive major-version-scale shift in the NFSv4 definition (i.e. something as big as the shift from NFSv3 to NFSv4 or from NFSv4.0 to NFSv4.1)

- o That the versioning paradigm not be used where it cannot actually be taken advantage of (i.e. where all extensions are optional).
- o That the extension mechanism be usable to correct protocol bugs in his documents and other RFC's updating existing RFC's. See Section 7.1 for details.
- o That a new, more flexible workflow be designed for ongoing protocol extension development. This should include some recognition of the role of the development of implementations in the maturing of a feature.
- o That appropriate version negotiation/discovery features be added to allow clients to discover what facilities a server supports, without having to attempt to use them all.

#### 7.4. Work Going Forward in Creating a New NFSv4 Extension Approach

The following set of steps are necessary in many possible ways of proceeding along the way to a new extension approach for NFSv4. Details and actual sequencing will reflect choices that the working group makes.

- o Creation of a new standards-track document defining how protocol extension and versioning are to work in NFSv4.

Since it would supersede existing treatments of the issues, it should be recorded as updating specifications for NFSv4.0 ([RFC3530] or the RFC arising from [RFC3530bis], when approved), for NFSv4.1 ([RFC5661]), and for NFSv4.2 (the RFC arising from [NFSv42] when approved).

- o Establishment of a framework for extension discovery (and negotiation if that is judged necessary) to replace testing of operations for an NFS3ERR\_OPNOTSUPP response.

The operations and attributes for this framework might be documented in the document defining the protocol extension framework, in a free-standing standards-track document, or as an infrastructural feature in a new minor version. In any of those cases, they could be incorporated as optional in earlier minor versions by documents updating the minor version specification RFC's.

- o Design of processes to review proposed extensions for architectural suitability, to reserve necessary operation codes, attribute numbers, and enum and flag values, and to decide when and if minor versions should be created to upgrade or downgrade features.

Such processes might well be documented in a working group informational document, but any such documentation should probably be done only after the working group is satisfied that the processes are working well.

As far as the value reservation issue, we will have to decide whether an IANA-based approach, as recommended by [RFC6709], is required or whether simpler procedures, implemented within the working group, are adequate.

As far as the document specifying the NFSv4 extension and versioning framework, the following are important elements:

- o Clearly separate the concepts of protocol extension and versioning to allow the basic XDR extension approach to be used in contexts other than creation of a minor version.
- o Clearly specify rules and expectations for updates to already defined minor versions.
- o Discourage any future incorporation of the definition of protocol extensions in minor version definition documents, except where the extension is infrastructural. Thus the basic function of minor version definition documents would be to specify what features already defined are included and their status (experimental, optional, recommended or mandatory) in that minor version.
- o Resolve and clarify cases where the original minor versioning rules don't match the extension/versioning model as it has evolved (e.g. how far the obligation of a client to support earlier versions extends, across what sorts of version changes does it make sense to require non-use of filehandles, stateids, etc.)

## 8. Security Considerations

Since no substantive protocol changes are proposed here, no security considerations apply.

As extensions are designed and specified, their security issues will be addressed and each extension will receive the appropriate security review from the NFSv4 working group and IESG.

## 9. IANA Considerations

The current document does not require any actions by IANA.

Depending on decisions that the working group makes about how to address the issues raised here, future documents may require actions by IANA.

## 10. Acknowledgements

The author wishes to thank Chuck Lever of Oracle for his helpful document review and many important suggestions.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

## 11.2. Informative References

## [NFSv42-dotx]

Haynes, T., Ed., "NFS Version 4 Minor Version 2 External Data Representation Standard (XDR) Description ", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-minorversion2-dot-x-19.txt>>.

Work in progress.

[NFSv42] Haynes, T., Ed., "NFS Version 4 Minor Version 2 ", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-minorversion2-20.txt>>.

Work in progress.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[RFC1094] Nowicki, B., "NFS: Network File System Protocol specification", RFC 1094, March 1989.

[RFC1813] Callaghan, B., Pawlowski, B., and P. Staubach, "NFS Version 3 Protocol Specification", RFC 1813, June 1995.

[RFC2780] Bradner, S. and V. Paxson, "IANA Allocation Guidelines For Values In the Internet Protocol and Related Headers", BCP 37, RFC 2780, March 2000.

[RFC3010] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "NFS version 4 Protocol", RFC 3010, December 2000.

[RFC3530] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.

## [RFC3530bis-dotx]

Haynes, T., Ed. and D. Noveck, Ed., "Network File System (NFS) Version 4 Protocol External Data Representation Standard (XDR) Description ", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-rfc3530bis-dot-x-18.txt>>.

Work in progress.

## [RFC3530bis]

Haynes, T., Ed. and D. Noveck, Ed., "Network File System (NFS) Version 4 Protocol ", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-rfc3530bis-27.txt>>.

Work in progress.

[RFC5661] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.

[RFC5662] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 External Data Representation Standard (XDR) Description", RFC 5662, January 2010.

[RFC6709] Carpenter, B., Aboba, B., and S. Cheshire, "Design Considerations for Protocol Extensions", RFC 6709, September 2012.

#### Author's Address

David Noveck  
EMC Corporation  
228 South Street  
Hopkinton, MA 01748  
US

Phone: +1 508 249 5748  
Email: [david.noveck@emc.com](mailto:david.noveck@emc.com)

NFSv4  
Internet-Draft  
Intended status: Informational  
Expires: March 20, 2014

D. Noveck, Ed.  
EMC  
P. Shivam  
C. Lever  
B. Baker  
ORACLE  
September 16, 2013

NFSv4 migration: Implementation experience and spec issues to resolve  
draft-ietf-nfsv4-migration-issues-04

## Abstract

The migration feature of NFSv4 provides for moving responsibility for a single filesystem from one server to another, without disruption to clients. Recent implementation experience has shown problems in the existing specification for this feature. This document discusses the issues which have arisen, explores the options available for curing the issues, and explains the choices made in updating the NFSv4.0 and NFSv4.1 specifications, to address migration.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on March 20, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Conventions . . . . .	3
3. NFSv4.0 Implementation Experience . . . . .	4
3.1. Implementation issues . . . . .	4
3.1.1. Failure to free migrated state on client reboot . . . .	4
3.1.2. Server reboots resulting in a confused lease situation . . . . .	5
3.1.3. Client complexity issues . . . . .	6
3.2. Sources of Protocol difficulties . . . . .	8
3.2.1. Issues with nfs_client_id4 generation and use . . . .	8
3.2.2. Issues with lease proliferation . . . . .	10
4. Issues to be resolved in NFSv4.0 . . . . .	10
4.1. Possible changes to nfs_client_id4 client-string . . . .	10
4.2. Possible changes to handle differing nfs_client_id4 string values . . . . .	11
4.3. Possible changes to add a new operation . . . . .	12
4.4. Other issues within migration-state sections . . . . .	12
4.5. Issues within other sections . . . . .	13
5. Proposed resolution of NFSv4.0 protocol difficulties . . . .	13
5.1. Proposed changes: nfs_client_id4 client-string . . . .	13
5.2. Proposed changes: merged (vs. synchronized) leases . . .	14
5.3. Other proposed changes to migration-state sections . . .	16
5.3.1. Proposed changes: Client ID migration . . . . .	16
5.3.2. Proposed changes: Callback re-establishment . . . .	16
5.3.3. Proposed changes: NFS4ERR_LEASE_MOVED rework . . . .	17
5.4. Proposed changes to other sections . . . . .	17
5.4.1. Proposed changes: callback update . . . . .	17
5.4.2. Proposed changes: clientid4 handling . . . . .	18
5.4.3. Proposed changes: NFS4ERR_CLID_INUSE . . . . .	19
6. Results of proposed changes for NFSv4.0 . . . . .	20
6.1. Results: Failure to free migrated state on client reboot	20
6.2. Results: Server reboots resulting in confused lease situation . . . . .	21
6.3. Results: Client complexity issues . . . . .	22
6.4. Result summary . . . . .	23
7. Issues for NFSv4.1 . . . . .	23
7.1. Addressing state merger in NFSv4.1 . . . . .	24
7.2. Addressing pNFS relationship with migration . . . . .	25
7.3. Addressing server owner changes in NFSv4.1 . . . . .	25
8. Security Considerations . . . . .	26



9. IANA Considerations . . . . .	27
10. Acknowledgements . . . . .	27
11. References . . . . .	27
11.1. Normative References . . . . .	27
11.2. Informative References . . . . .	27
Authors' Addresses . . . . .	28

## 1. Introduction

This document is in the informational category, and while the facts it reports may have normative implications, any such normative significance reflects the readers' preferences. For example, we may report that the reboot of a client with migrated state results in state not being promptly cleared and that this will prevent granting of conflicting lock requests at least for the lease time, which is a fact. While it is to be expected that client and server implementers will judge this to be a situation that is best avoided, the judgment as to how pressing this issue should be considered is a judgment for the reader, and eventually the nfsv4 working group to make.

We do explore possible ways in which such issues can be avoided, with minimal negative effects, given that the working group has decided to address these issues, but the choice of exactly how to address these is best given effect in one or more standards-track documents and/or errata.

This document focuses on NFSv4.0, since that is where the majority of implementation experience has been. Nevertheless, there is discussion of the implications of the NFSv4.0 experience for migration in NFSv4.1, as well as discussion of other issues with regard to the treatment of migration in NFSv4.1.

## 2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

In the context of this informational document, these normative keywords will always occur in the context of a quotation, most often direct but sometimes indirect. The context will make it clear whether the quotation is from:

- o The current definitive definition of the NFSv4.0 protocol, whether that is the original NFSv4.0 specification [RFC3530], or its expected successor [RFC3530bis].

As the identity of that document may change during the lifetime of this document, we will often refer to the current or pending definition of NFSv4.0 and quote from portions of the documents that are identical among all existing drafts. Given that RFC3530 and all RFC3530bis drafts agree as to the issues under discussion, this should not cause undue difficulty. Note that to simplify document maintenance, section names rather than section numbers are used when referring to sections in existing documents so that only minimal changes will be necessary as the identity of the document defining NFSv4.0 changes.

- o The current definitive definition of the NFSv4.1 protocol [RFC5661].
- o A proposed or possible text to serve as a replacement for the current definitive document text. Sometimes, a number of possible alternative texts may be listed and benefits and detriments of each examined in turn.

### 3. NFSv4.0 Implementation Experience

#### 3.1. Implementation issues

Note that the examples below reflect current experience which arises from clients implementing the recommendation to use different `nfs_client_id4` id strings for different server addresses, i.e. using what is later referred to herein as the "non-uniform client-string approach."

This is simply because that is the experience implementers have had. The reader should not assume that in all cases, this practice is the source of the difficulty. It may be so in some cases but clearly it is not in all cases.

##### 3.1.1. Failure to free migrated state on client reboot

The following sort of situation has proved troublesome:

- o A client C establishes a `clientid4` C1 with server ABC specifying an `nfs_client_id4` with id string value "C-ABC" and boot verifier 0x111.
- o The client begins to access files in filesystem F on server ABC, resulting in generating stateids S1, S2, etc. under the lease for `clientid` C1. It may also access files on other filesystems on the same server.

- o The filesystem is migrated from server ABC to server XYZ. When transparent state migration is in effect, stateids S1 and S2 and clientid4 C1 are now available for use by client C at server XYZ.
- o Client C reboots and attempts to access data on server XYZ, whether in filesystem F or another. It does a SETCLIENTID with an nfs\_client\_id4 with id string value "C-XYZ" and boot verifier 0x112. There is thus no occasion to free stateids S1 and S2 since they are associated with a different client name and so lease expiration is the only way that they can be gotten rid of.

Note here that while it seems clear to us in this example that C-XYZ and C-ABC are from the same client, the server has no way to determine the structure of the "opaque" id string. In the protocol, it really is treated as opaque. Only the client knows which nfs\_client\_id4 values designate the same client on a different server.

### 3.1.2. Server reboots resulting in a confused lease situation

Further problems arise from scenarios like the following.

- o Client C talks to server ABC using an nfs\_client\_id4 id string such as "C-ABC" and a boot verifier v1. As a result, a lease with clientid4 c.i is established: {v1, "C-ABC", c.i}.
- o fs\_a1 migrates from server ABC to server XYZ along with its state. Now server XYZ also has a lease: {v1, "C-ABC", c.i}.
- o Server ABC reboots.
- o Client C talks to server ABC using an nfs\_client\_id4 id string such as "C-ABC" and a boot verifier v1. As a result, a lease with clientid4 c.j is established: {v1, "C-ABC", c.j}.
- o fs\_a2 migrates from server ABC to server XYZ. Now server XYZ also has a lease: {v1, "C-ABC", c.j}.
- o Now server XYZ has two leases that match {v1, "C-ABC", \*}, when the protocol clearly assumes there can be only one.

Note that if the client used "C" (rather than "C-ABC") as the nfs\_client\_id4 id string, the exact same situation would arise.

One of the first cases in which this sort of situation has resulted in difficulties is in connection with doing a SETCLIENTID for callback update.

The SETCLIENTID for callback update only includes the `nfs_client_id4`, assuming there can only be one such with a given `nfs_client_id4` value. If there were multiple, confirmed client records with identical `nfs_client_id4` id string values, there would be no way to map the callback update request to the correct client record. Apart from the migration handling specified in [RFC3530] and [RFC3530bis], such a situation cannot arise.

One possible accommodation for this particular issue that has been used is to add a RENEW operation along with SETCLIENTID (on a callback update) to disambiguate the client.

When the client updates the callback info to the destination, the client would, by convention, send a compound like this:

```
{ RENEW clientid4, SETCLIENTID nfs_client_id4,verf,cb }
```

The presence of the `clientid4` in the compound would allow the server to differentiate among the various leases that it knows of, all with the same `nfs_client_id4` value.

While this would be a reasonable patch for an isolated protocol weakness, interoperable clients and servers would require that the protocol truly be updated to allow such a situation, specifically that of multiple `clientid4`'s with the same `nfs_client_id4` value. The protocol is currently designed and implemented assuming this cannot happen. We need to either prevent the situation from happening, or fully adapt to the possibilities which can arise. See Section 4 for a discussion of such issues.

### 3.1.3. Client complexity issues

Consider the following situation:

- o There are a set of clients  $C_1$  through  $C_n$  accessing servers  $S_1$  through  $S_m$ . Each server manages some significant number of filesystems with the filesystem count  $L$  being significantly greater than  $m$ .
- o Each client  $C_x$  will access a subset of the servers and so will have up to  $m$  clientids, which we will call  $C_{xy}$  for server  $S_y$ .
- o Now assume that for load-balancing or other operational reasons, numbers of filesystems are migrated among the servers. As a result, each client-server pair will have up to  $m$  clientids and each client will have up to  $m^2$  clientids. If we add the possibility of server reboot, the only bound on a client's clientid count is  $L$ .

Now, instead of a `clientid4` identifying a client-server pair, we have many more entities for the client to deal with. In addition, it isn't clear how new state is to be incorporated in this structure.

The limitations of the migrated state (inability to be freed on reboot) would argue against adding more such state but trying to avoid that would run into its own difficulties. For example, a single lockowner string presented under two different `clientids` would appear as two different entities.

Thus we have to choose between:

- o indefinite prolongation of foreign `clientids` even after all transferred state is gone.
- o having multiple requests for the same lockowner-string-named entity carried on in parallel by separate identically named lockowners under different `clientid4`'s
- o Adding serialization at the lock-owner string level, in addition to that at the lockowner level.

In any case, we have gone (in adding migration as it was described) from a situation in which

- o Each client has a single `clientid4/lease` for each server it talks to.
- o Each client has a single `nfs_client_id4` for each server it talks to.
- o Every state id can be mapped to an associated lease based on the server it was obtained from.

To one in which

- o Each client may have multiple `clientid4`'s for a single server.
- o For each stateid, the client must separately record the `clientid4` that it is assigned to, or it must manage separate "state blobs" for each fsid and map those to `clientid4`'s.
- o Before doing an operation that can result in a stateid, the client must either find a "state blob" based on fsid or create a new one, possibly with a new `clientid4`.
- o There may be multiple `clientid4`'s all connected to the same server and using the same `nfs_clientid4`.

This sort of additional client complexity is troublesome and needs to be eliminated.

### 3.2. Sources of Protocol difficulties

#### 3.2.1. Issues with nfs\_client\_id4 generation and use

The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Client ID" says:

The second field, id is a variable length string that uniquely defines the client.

There are two possible interpretations of the phrase "uniquely defines" in the above:

- o The relation between strings and clients is a function from such strings to clients so that each string designates a single client.
- o The relation between strings and clients is a bijection between such strings and clients so that each string designates a single client and each client is named by a single string.

The first interpretation would make these client-strings like phone numbers (a single person can have several) while the second would make them like social security numbers.

Endless debate about the true meaning of "uniquely defines" in this context is quite possible but not very helpful. The following points should be noted though:

- o The second interpretation is more consistent with the way "uniquely defines" is used elsewhere in the spec.
- o The spec as now written intends the first interpretation (or is internally inconsistent). In fact, it recommends, although it doesn't "RECOMMEND" that a single client have at least as many client-strings as server addresses that it interacts with. It says, in the third bullet point regarding construction of the string (which we shall henceforth refer to as client-string-BP3):

The string should be different for each server network address that the client accesses, rather than common to all server network addresses.

- o If internode interactions are limited to those between a client and its servers, there is no occasion for servers to be concerned with the question of whether two client-strings designate the same

client, so that there is no occasion for the difference in interpretation to matter.

- o When transparent migration of client state occurs between two servers, it becomes important to determine when state on two different servers is for the same client or not, and this distinction becomes very important.

Given the need for the server to be aware of client identity with regard to migrated state, either client-string construction rules will have to change or there will be a need to get around current issues, or perhaps a combination of these two will be required. Later sections will examine the options and propose a solution.

One consideration that may indicate that this cannot remain exactly as it is today has to do with the fact that the current explanation for this behavior is not correct. The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Client ID" says:

The reason is that it may not be possible for the client to tell if the same server is listening on multiple network addresses. If the client issues SETCLIENTID with the same id string to each network address of such a server, the server will think it is the same client, and each successive SETCLIENTID will cause the server to begin the process of removing the client's previous leased state.

In point of fact, a "SETCLIENTID with the same id string" sent to multiple network addresses will be treated as all from the same client but will not "cause the server to begin the process of removing the client's previous leased state" unless the server believes it is a different instance of the same client, i.e. if the id string is the same and there is a different boot verifier. If the client does not reboot, the verifier should not change. If it does reboot, the verifier will change, and the server should "begin the process of removing the client's previous leased state."

The situation of multiple SETCLIENTID requests received by a server on multiple network addresses is exactly the same, from the protocol design point of view, as when multiple (i.e. duplicate) SETCLIENTID requests are received by the server on a single network address. The same protocol mechanisms that prevent erroneous state deletion in the latter case prevent it in the former case. There is no reason for special handling of the multiple-network-appearance case, in this regard.

### 3.2.2. Issues with lease proliferation

It is often felt that this is a consequence of the client-string construction issues, and it is certainly the case that the two are closely connected in that non-uniform client-strings make it impossible for the server to appropriately combine leases from the same client.

However, even where the server could combine leases from the same client, it needs to be clear how and when it will do so, so that the client will be prepared. These issues will have to be addressed at various places in the spec.

This could be enough only if we are prepared to do away with the "should" recommending non-uniform client-strings and replace it with a "should not" or even a "SHOULD NOT". Current client implementation patterns make this an unpalatable choice for use as a general solution, but it is reasonable to "RECOMMEND" this choice for a well-defined subset of clients. One alternative would be to create a way for the server to infer from client behavior which leases are held by the same client and use this information to do appropriate lease mergers. Prototyping and detailed specification work has shown that this could be done but the resulting complexity is such that a better choice is to "RECOMMEND" use of the uniform client-string approach for clients supporting the migration feature.

Because of the discussion of client-string construction in [RFC3530] and [RFC3530bis], most existing clients implement the non-uniform client-string approach. As a result, existing servers may not have been tested with clients implementing uniform client-strings. As a consequence, care must be taken to preserve interoperability between UCS-capable clients and servers that don't tolerate uniform client strings for one reason or another.

## 4. Issues to be resolved in NFSv4.0

### 4.1. Possible changes to nfs\_client\_id4 client-string

The fact that the reason given in client-string-BP3 is not valid makes the existing "should" insupportable. We can't either

- o Keep a reason we know is invalid.
- o Keep saying "should" without giving a reason.

What are often presented as reasons that motivate use of the non-uniform approach always turn out to be cases in which, if the uniform approach were used, the server will treat a client which accesses



that server via two different IP addresses as part of a single client, as it in fact is. This may be disconcerting to a client unaware that the two IP addresses connect to the same server. This is not a reason to use the non-uniform approach but is better thought of as an illustration of the fact that those using the uniform approach need to be aware of the possibility of server trunking and its effect on server behavior.

If it is possible to reliably infer the existence of trunking of server IP addresses from observed server behavior, use of the uniform approach would be more desirable, although compatibility issues would have to be dealt with.

An alternative to having the client infer the existence of trunking of IP server addresses, is to make this information available to the client directly. See Section 4.3 for details.

It is always possible that a valid new reason will be found, but so far none has been proposed. Given the history, the burden of proof should be on those asserting the validity of a proposed new reason.

So we will assume for now that the "should" will have to go. The question is what to replace it with.

- o We can't say "MUST NOT", despite the problems this raises for migration since this is pretty late in the day for such a change. Many currently operating clients obey the existing "should". Similar considerations would apply for "SHOULD NOT" or "should not".
- o Dropping client-string-BP3 entirely is a possibility but, given the context and history, it would just be a confusing version of "SHOULD NOT".
- o Using "MAY" would clearly specify that both ways of doing this are valid choices for clients and that servers will have to deal with clients that make either choice.
- o This might be modified by a "SHOULD" (or even a "MUST") for particular groups of clients.
- o There will have to be some text explaining why a client might make either choice but, except for the particular cases referred to above, we will have to make sure that it is truly descriptive, and not slanted in either direction.

#### 4.2. Possible changes to handle differing nfs\_client\_id4 string values

Given the difficulties caused by having different `nfs_client_id4` client-string values for the same client, we have two choices:

- o Deprecate the existing treatment and basically say the client is on its own doing migration, if it follows it.
- o Introduce a way of having the client provide client identity information to the server, if it can be done compatibly while staying within the bounds of v4.0.

#### 4.3. Possible changes to add a new operation

It might be possible to return server-identity information to the client, just as is done in NFSv4.1 by the response to the `EXCHANGE_ID` operation. This could be done by a `SETCLIENTID_PLUS` optional operation, which acts like `SETCLIENTID`, except that it returns server identity information. Such information could be used by clients, making it possible for them to be aware of server trunking relationships, rather than having to infer them from server behavior.

It had always been thought that protocol extensions such as this were not appropriate to bis documents and other documents updating NFSv4 protocol definition RFC's. However, it is argued in [NFS-ext] that protocol extensions, similar to those allowed between minor versions, should be acceptable to correct mistakes within a minor version.

A decision to adopt this approach will depend on nfsv4 working group discussion and would probably best be effected by means of a standards-track document laying out a modified NFSv4 extension/versioning model for all minor versions.

In view of the time to effect such changes, this approach is not likely to be adopted in an RFC updating [RFC3530] or [RFC3530bis], such as [migr-v4.0-update]. Still, it is worth keeping in mind, if implementers have difficulties inferring trunking relationships using the techniques discussed there.

#### 4.4. Other issues within migration-state sections

There are a number of issues where the existing text is unclear and/or wrong and needs to be fixed in some way.

- o Lack of clarity in the discussion of moving clientids (as well as stateids) as part of moving state for migration.
- o The discussion of synchronized leases is wrong in that there is no way to determine (in the current spec) when leases are for the same client and also wrong in suggesting a benefit from leases

synchronized at the point of transfer. What is needed is merger of leases, which is necessary to keep client complexity requirements from getting out of hand.

- o Lack of clarity in the discussion of LEASE\_MOVED handling, including failure to fully address situations in which transparent state migration did not occur.

#### 4.5. Issues within other sections

There are a number of cases in which certain sections, not specifically related to migration, require additional clarification. This is generally because text that is clear in a context in which leases and clientids are created in one place and live there forever may need further refinement in the more dynamic environment that arises as part of migration.

Some examples:

- o Some people are under the impression that updating callback endpoint information for an existing client, as used during migration, may cause the destination server to free existing state. There need to be additions to clarify the situation.
- o The handling of the sets of clientid4's maintained by each server needs to be clarified. In particular, the issue of how the client adapts to the presumably independent and uncoordinated clientid4 sets needs to be clearly addressed
- o Statements regarding handling of invalid clientid4's need to be clarified and/or refined in light of the possibilities that arise due to lease motion and merger.
- o Confusion and lack of clarity about NFS4ERR\_CLID\_INUSE.

#### 5. Proposed resolution of NFSv4.0 protocol difficulties

This section lists the changes which we believe are necessary to resolve the difficulties mentioned above. Such change, along with other clarifications found to be desirable during drafting and review are contained in [migr-v4.0-update].

##### 5.1. Proposed changes: nfs\_client\_id4 client-string

We propose replacing client-string-BP3 with the following text and adding the following proposed to provide implementation guidance.

The string MAY be different for each server network address that the client accesses, rather than common to all server network addresses.

In addition, given the importance of the issue of client identity and the fact that both client string-approaches are to be considered valid, a greatly expanded treatment of client identity desirable. It should have the following major elements.

- o It should fully describe the consequences of making the string different for each network address (the non-uniform client-string approach) and of making it the same for all network addresses (the uniform client string approach).
- o It should give helpful guidance about the factors that might affect client implementation choice between these approaches.
- o It should describe the compatibility issues that might cause servers to be incompatible with the uniform approach and give guidance about dealing with these.
- o It should describe how a client using the uniform approach might use server behavior to determine server address trunking patterns.
- o It should present a clearer and more complete set of recommendations to guide client string construction.

## 5.2. Proposed changes: merged (vs. synchronized) leases

The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Migration and State" says:

As part of the transfer of information between servers, leases would be transferred as well. The leases being transferred to the new server will typically have a different expiration time from those for the same client, previously on the old server. To maintain the property that all leases on a given server for a given client expire at the same time, the server should advance the expiration time to the later of the leases being transferred or the leases already present. This allows the client to maintain lease renewal of both classes without special effort:

There are a number of problems with this and any resolution of our difficulties must address them somehow.

- o The current v4.0 spec recommends that the client make it essentially impossible to determine when two leases are from "the same client".
- o It is not appropriate to speak of "maintain[ing] the property that all leases on a given server for a given client expire at the same time", since this is not a property that holds even in the absence of migration. A server listening on multiple network addresses may have the same client appear as multiple clients with no way to recognize the client as the same.
- o Even if the client identity issue could be resolved, advancing the lease time at the point of migration would not maintain the desired synchronization property. The leases would be synchronized until one of them was renewed, after which they would be unsynchronized again.

To avoid client complexity, we need to have no more than one lease between a single client and a single server. This requires merger of leases since there is no real help from synchronizing them at a single instant.

For the uniform approach, the destination server would simply merge leases as part of state transfer, since two leases with the same `nfs_client_id4` values must be for the same client.

We have made the following decisions as far as proposed normative statements regarding for state merger. They reflect the facts that we want to support fully migration support in the simplest way possible and that we can't say MUST since we have older clients and servers to deal with.

- o Clients SHOULD use the uniform client-string approach in order to get good migration support.
- o Servers SHOULD provide automatic lease merger during state migration so that clients using the uniform id approach get the support automatically.

If the clients and the servers obey the SHOULD's, having more than a single lease for a given client-server pair will be a transient situation, cleaned up as part of adapting to use of migrated state.

Since clients and servers will be a mixture of old and new and because nothing is a MUST we have to ensure that no combination will show worse behavior than is exhibited by current (i.e. old) clients and servers.

### 5.3. Other proposed changes to migration-state sections

#### 5.3.1. Proposed changes: Client ID migration

The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Migration and State" says:

In the case of migration, the servers involved in the migration of a filesystem SHOULD transfer all server state from the original to the new server. This must be done in a way that is transparent to the client. This state transfer will ease the client's transition when a filesystem migration occurs. If the servers are successful in transferring all state, the client will continue to use stateids assigned by the original server. Therefore the new server must recognize these stateids as valid. This holds true for the client ID as well. Since responsibility for an entire filesystem is transferred with a migration event, there is no possibility that conflicts will arise on the new server as a result of the transfer of locks.

This poses some difficulties, mostly because the part about "client ID" is not clear:

- o It isn't clear what part of the paragraph the "this" in the statement "this holds true ..." is meant to signify.
- o The phrase "the client ID" is ambiguous, possibly indicating the clientid4 and possibly indicating the nfs\_client\_id4.
- o If the text means to suggest that the same clientid4 must be used, the logic is not clear since the issue is not the same as for stateids of which there might be many. Adapting to the change of a single clientid, as might happen as a part of lease migration, is relatively easy for the client.

We have decided that it is best to address this issue as follows:

- o Make it clear that both clientid4 and nfs\_client\_id4 (including both id string and boot verifier) are to be transferred.
- o Indicate that the initial transfer will result in the same clientid4 after transfer but this is not guaranteed since there may conflict with an existing clientid4 on the destination server and because lease merger can result in a change of the clientid4.

#### 5.3.2. Proposed changes: Callback re-establishment

The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Migration and State" says:

A client SHOULD re-establish new callback information with the new server as soon as possible, according to sequences described in sections "Operation 35: SETCLIENTID - Negotiate Client ID" and "Operation 36: SETCLIENTID\_CONFIRM - Confirm Client ID". This ensures that server operations are not blocked by the inability to recall delegations.

The above will need to be fixed to reflect the possibility of merging of leases,

#### 5.3.3. Proposed changes: NFS4ERR\_LEASE\_MOVED rework

The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Notification of Migrated Lease" says:

Upon receiving the NFS4ERR\_LEASE\_MOVED error, a client that supports filesystem migration MUST probe all filesystems from that server on which it holds open state. Once the client has successfully probed all those filesystems which are migrated, the server MUST resume normal handling of stateful requests from that client.

There is a lack of clarity that is prompted by ambiguity about what exactly probing is and what the interlock between client and server must be. This has led to some worry about the scalability of the probing process, and although the time required does scale linearly with the number of filesystems that the client may have state for with respect to a given server, the actual process can be done efficiently.

To address these issues we propose rewriting the above to be more clear and to give suggestions about how to do the required scanning efficiently.

#### 5.4. Proposed changes to other sections

##### 5.4.1. Proposed changes: callback update

Some changes are necessary to reduce confusion about the process of callback information update and in particular to make it clear that no state is freed as a result:

- o Make it clear that after migration there are confirmed entries for transferred clientid4/nfs\_client\_id4 pairs.
- o Be explicit in the sections headed "otherwise," in the descriptions of SETCLIENTID and SETCLIENTID\_CONFIRM, that these don't apply in the cases we are concerned about.

#### 5.4.2. Proposed changes: clientid4 handling

To address both of the clientid4-related issues mentioned in Section 4.5, we propose replacing the last three paragraphs of the section entitled "Client ID" with the following:

Once a SETCLIENTID and SETCLIENTID\_CONFIRM sequence has successfully completed, the client uses the shorthand client identifier, of type clientid4, instead of the longer and less compact nfs\_client\_id4 structure. This shorthand client identifier (a client ID) is assigned by the server and should be chosen so that it will not conflict with a client ID previously assigned by same server. This applies across server restarts or reboots.

Distinct servers MAY assign clientid4's independently, and will generally do so. Therefore, a client has to be prepared to deal with multiple instances of the same clientid4 value received on distinct IP addresses, denoting separate entities. When trunking of server IP addresses is not a consideration, a client should keep track of (IP-address, clientid4) pairs, so that each pair is distinct. In the face of possible trunking of server IP addresses, the client will use the receipt of the same clientid4 from multiple IP-addresses, as an indication that the two IP-addresses may be trunked and proceed to determine, from the observed server behavior whether the two addresses are in fact trunked.

When a clientid4 is presented to a server and that clientid4 is not recognized, the server will reject the request with the error NFS4ERR\_STALE\_CLIENTID. This can occur for a number of reasons:

- \* A server reboot causing loss of the server's knowledge of the client
- \* Client error sending an incorrect clientid4 or a valid clientid4 to the wrong server.
- \* Loss of lease state due to lease expiration.



- \* Client or server error causing the server to believe that the client has rebooted (i.e. receiving a SETCLIENTID with an nfs\_client\_id4 which has a matching id string and a non-matching boot verifier).
- \* Migration of all state under the associated lease causes its non-existence to be recognized on the source server.
- \* Merger of state under the associated lease with another lease under a different clientid causes the clientid4 serving as the source of the merge to cease being recognized on its server.

In the event of a server reboot, or loss of lease state due to lease expiration, the client must obtain a new clientid4 by use of the SETCLIENTID operation and then proceed to any other necessary recovery for the server reboot case (See the section entitled "Server Failure and Recovery"). In cases of server or client error resulting in this error, use of SETCLIENTID to establish a new lease is desirable as well.

In the last two cases, different recovery procedures are required. Note that in cases in which there is any uncertainty about which sort of handling is applicable, the distinguishing characteristic is that in reboot-like cases, the clientid4 and all associated stateids cease to exist while in migration-related cases, the clientid4 ceases to exist while the stateids are still valid.

The client must also employ the SETCLIENTID operation when it receives a NFS4ERR\_STALE\_STATEID error using a stateid derived from its current clientid4, since this indicates a situation, such as server reboot which has invalidated the existing clientid4 and associated stateids (see the section entitled "lock-owner" for details).

See the detailed descriptions of SETCLIENTID and SETCLIENTID\_CONFIRM for a complete specification of the operations.

#### 5.4.3. Proposed changes: NFS4ERR\_CLID\_INUSE

It appears to be the intention that only a single principal be used for client establishment between any client-server pair. However:

- o There is no explicit statement to this effect.
- o The error that indicates a principal conflict has a name which does not clarify this issue: NFS4ERR\_CLID\_INUSE.

- o The definition of the error is also not very helpful: "The SETCLIENTID operation has found that a client id is already in use by another client".

As a result, servers exist which reject a SETCLIENTID simply because there already exists a clientid for the same client, established using a different IP address. Although this is generally understood to be erroneous, such servers still exist and the spec should make the correct behavior clear.

Although the error name cannot be changed, the following changes should be made to avoid confusion:

- o The definition of the error should be changed to read as follows:

The SETCLIENTID operation has found that the specified nfs\_client\_id4 was previously presented with a different principal and that client instance currently holds an active lease. A server MAY return this error if the same principal is used but a change in authentication flavor gives good reason to reject the new SETCLIENTID operation as not bona fide.

- o In the description of SETCLIENTID, the phrase "then the server returns a NFS4ERR\_CLID\_INUSE error" should be expanded to read "then the server returns a NFS4ERR\_CLID\_INUSE error, since use of a single client with multiple principals is not allowed."

## 6. Results of proposed changes for NFSv4.0

The purpose of this section is to examine the troubling results reported in Section 3.1. We will look at the scenarios as they would be handled within the proposal.

Because the choice of uniform vs. non-uniform nfs\_client\_id4 id strings is a "SHOULD" in these cases, we will designate clients that follow this recommendation by SHOULD-UF-CID.

We will also have to take account of any merger-related "SHOULD" clauses to better understand how they have addressed the issues seen. We abbreviate as follows:

- o SHOULD-SVR-AM refers to the server obeying the SHOULD which RECOMMENDS that they merge leases with identical nfs\_client\_id4 id strings and boot verifiers.

### 6.1. Results: Failure to free migrated state on client reboot

Let's look at the troublesome situation cited in Section 3.1.1. We have already seen what happens when SHOULD-UF-CID does not hold. Now let's look at the situation in which SHOULD-UF-CID holds, whether SHOULD-SVR-AM is in effect or not.

- o A client C establishes a clientid4 C1 with server ABC specifying an nfs\_client\_id4 with id string value "C" and boot verifier 0x111.
- o The client begins to access files in filesystem F on server ABC, resulting in generating stateids S1, S2, etc. under the lease for clientid C1. It may also access files on other filesystems on the same server.
- o The filesystem is migrated from ABC to server XYZ. When transparent state migration is in effect, stateids S1 and S2 and lease {0x111, "C", C1} are now available for use by client C at server XYZ.
- o Client C reboots and attempts to access data on server XYZ, whether in filesystem F or another. It does a SETCLIENTID with an nfs\_client\_id4 with id string value "C" and boot verifier 0x112. The state associated with lease {0x111, "C", C1} is deleted as part of creating {0x112, "C", C2}. No problem.

The correctness signature for this issue is

SHOULD-UF-CID

so if you have clients and servers that obey the SHOULD clauses, the problem is gone regardless of the choice on the MAY.

## 6.2. Results: Server reboots resulting in confused lease situation

Now let's consider the scenario given in Section 3.1.2. We have already seen what happens when SHOULD-UF-CID does not hold. Now let's look at the situation in which SHOULD-UF-CID holds and SHOULD-SVR-AM holds as well.

- o Client C talks to server ABC using an nfs\_client\_id4 id string such as "C-ABC" and boot verifier v1. As a result a lease with clientid4 c.i established: {v1, "C-ABC", c.i}.
- o Filesystem fs\_a1 migrates from server ABC to server XYZ along with its state. Now server XYZ also has a lease: {v1, "C-ABC", c.i}
- o Server ABC reboots.

- o Client C talks to server ABC using an `nfs_client_id4` id string such as "C-ABC" and boot verifier `v1`. As a result a lease with `clientid4 c.j` established: `{v1, "C-ABC", c.j}`.
- o `fs_a2` migrates from server ABC to server XYZ. As part of migration the incoming lease is seen to denote same `nfs_client_id4` and so is merged with `{v1, "C-ABC", c.i}`.
- o Now server XYZ has only one lease that matches `{v1, "C-ABC", *}`, so the problem is solved

Now let's consider the same scenario in the situation in which SHOULD-UF-CID holds and SHOULD-SVR-AM holds as well.

- o Client C talks to server ABC using an `nfs_client_id4` id string "C" and boot verifier `v1`. As a result a lease with `clientid4 c.i` is established: `{v1, "C", c.i}`.
- o `fs_a1` migrates from server ABC to server XYZ along with its state. Now XYZ also has a lease: `{v1, "C", c.i}`
- o Server ABC reboots.
- o Client C talks to server ABC using an `nfs_client_id4` id string "C" and boot verifier `v1`. As a result a lease with `clientid4 c.j` is established: `{v1, "C", c.j}`.
- o `fs_a2` migrates from server ABC to server XYZ. As part of migration the incoming lease is seen to denote the same `nfs_client_id4` and so is merged with `{v1, "C", c.i}`.
- o Now server XYZ has only one lease that matches `{v1, "C", *}`, so the problem is solved

The correctness signature for this issue is

SHOULD-SVR-AM

so if you have clients and servers that obey the SHOULD clauses, the problem is gone regardless of the choice on the MAY.

### 6.3. Results: Client complexity issues

Consider the following situation:

- o There are a set of clients  $C_1$  through  $C_n$  accessing servers  $S_1$  through  $S_m$ . Each server manages some significant number of filesystems with the filesystem count  $L$  being significantly greater than  $m$ .
- o Each client  $C_x$  will access a subset of the servers and so will have up to  $m$  clientids, which we will call  $C_{xy}$  for server  $S_y$ .
- o Now assume that for load-balancing or other operational reasons, numbers of filesystems are migrated among the servers. As a result, depending on how this handled, the number of clientids may explode. See below.

Now look what will happen under various scenarios:

- o We have previously (in Section 3.1.3) looked at this in case of client following the non-uniform client-string approach. In that case, each client-server pair could have up to  $m$  clientids and each client will have up to  $m^2$  clientids. If we add the possibility of server reboot, the only bound on a client's clientid count is  $L$ .
- o If we look at this in the SHOULD-UF-CID case in which the SHOULD-SVR-AM condition holds, the situation is no different. Although the server has the client identity information that could enable same-client-same-server leases to be combined, it does not do so. We still have up to  $L$  clientids per client.
- o On the other hand, if we look at the SHOULD-UF-CID case in which SHOULD-SVR-AM holds, the problem is gone. There can be no more than  $m$  clientids per client, and  $n$  clientids per server.

The correctness signature for this issue is

(SHOULD-UF-CID & SHOULD-SVR-AM)

so if you have clients and servers that obey the SHOULD clauses, the problem is gone regardless of the choice on the MAY.

#### 6.4. Result summary

We have seen that (SHOULD-SVR-AM & SHOULD-UF-CID) are sufficient to solve the problems people have experienced.

#### 7. Issues for NFSv4.1

Because NFSv4.1 embraces the uniform client-string approach, addressing migration issues is simpler. In the terms of Section 6,

we already have SHOULD-UF-CID, for NFSv4.1, as advised by section 2.4 of [RFC5661], simplifying the work to be done.

Nevertheless, there are some issues that will have to be addressed. Some examples:

- o The other necessary part of addressing migration issues, which we call above SHOULD-SVR-AM, is not currently addressed by NFSv4.1 and changes need to be made to make it clear that state needs to be appropriately merged as part of migration, to avoid multiple clientids between a client-server pair.
- o There needs to be some clarification of how migration, and particularly transparent state migration, should interact with pNFS layouts.
- o The current discussion (in [RFC5661]), of the possibility of server\_owner changes is incomplete and confusing.

Discussion of how to resolve these issues will appear in the sections below.

#### 7.1. Addressing state merger in NFSv4.1

The existing treatment of state transfer in [RFC5661], has similar problems to that in [RFC3530] and [RFC3530bis] in that it assumes that the state for multiple filesystems on different servers will not be merged so that it appears under a single common clientid. We've already seen the reasons that this is a problem, with regard to NFSv4.0.

Although we don't have the problems stemming from the non-uniform client-string approach, there are a number of complexities in the existing treatment of state management in the section entitled "Lock State and File System Transitions" in [RFC5661] that make this non-trivial to address:

- o Migration is currently treated together with other sorts of filesystem transitions including transitioning between replicas without any NFS4ERR\_MOVED errors.
- o There is separate handling and discussion of the cases of matching and non-matching server scopes.
- o In the case of matching server scopes, the text calls for an impossible degree of transparency.

- o In the case of non-matching server scopes, the text does not mention transparent state migration at all, resulting in a functional regression from NFSV4.0

## 7.2. Addressing pNFS relationship with migration

This is made difficult because, within the PNFS framework, migration might mean any of several things:

- o Transfer of the MDS, leaving DS's alone.

This would be minimally disruptive to those using layouts but would require the pNFS control protocol to support the DS being directed to a new MDS.

- o Transfer of a DS, leaving everything else in place.

Such a transfer can be handled without using migration at all. The server can recall/revoke layouts, as appropriate.

- o Transfer of the filesystem to a new filesystem with both MDS and DS's moving.

In such a transfer, an entirely different set of DS's will be at the target location. There may even be no pNFS support on the destination filesystem at all.

Migration needs to support both the first and last of these models.

## 7.3. Addressing server owner changes in NFSv4.1

Section 2.10.5 of [RFC5661] states the following.

The client should be prepared for the possibility that `eir_server_owner` values may be different on subsequent `EXCHANGE_ID` requests made to the same network address, as a result of various sorts of reconfiguration events. When this happens and the changes result in the invalidation of previously valid forms of trunking, the client should cease to use those forms, either by dropping connections or by adding sessions. For a discussion of lock reclaim as it relates to such reconfiguration events, see Section 8.4.2.1.

While this paragraph is literally true in that such reconfiguration events can happen and clients have to deal with them, it is confusing in that it can be read as suggesting that clients have to deal with them without disruption, which in general is impossible.

A clearer alternative would be:

It is always possible that, as a result of various sorts of reconfiguration events, `eir_server_scope` and `eir_server_owner` values may be different on subsequent `EXCHANGE_ID` requests made to the same network address.

In most cases such reconfiguration events will be disruptive and indicate that an IP address formerly connected to one server is now connected to an entirely different one.

Some guidelines on client handling of such situations follow:

- \* When `eir_server_scope` changes, the client has no assurance that any id's it obtained previously (e.g. file handles) can be validly used on the new server, and, even if the new server accepts them, there is no assurance that this is not due to accident. Thus it is best to treat all such state as lost/stale although a client may assume that the probability of inadvertent acceptance is low and treat this situation as within the next case.
- \* When `eir_server_scope` remains the same and `eir_server_owner.so_major_id` changes, the client can use filehandles it has and attempt reclaims. It may find that these are now stale but if `NFS4ERR_STALE` is not received, he can proceed to reclaim his opens.
- \* When `eir_server_scope` and `eir_server_owner.so_major_id` remain the same, the client has to use the now-current values of `eir_server-owner.so_minor_id` in deciding on appropriate forms of trunking.

## 8. Security Considerations

The current definitive definitions of the NFSv4.0 protocol, [RFC3530] and [RFC3530bis] both agree. The section entitled "Security Considerations" encourages that clients protect the integrity of the `SECINFO` operation, any `GETATTR` operation for the `fs_locations` attribute, and the operations `SETCLIENTID/SETCLIENTID_CONFIRM`. A migration recovery event can use any or all of these operations. We do not recommend any change here.



## 9. IANA Considerations

This document does not require actions by IANA.

## 10. Acknowledgements

The editor and authors of this document gratefully acknowledge the contributions of Trond Myklebust of NetApp and Robert Thurlow of Oracle. We also thank Tom Haynes of NetApp and Spencer Shepler of Microsoft for their guidance and suggestions.

Special thanks go to members of the Oracle Solaris NFS team, especially Rick Mesta and James Wahlig, for their work implementing an NFSv4.0 migration prototype and identifying many of the issues documented here.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3530] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.
- [RFC3530bis]  
Haynes, T., Ed. and D. Noveck, Ed., "Network File System (NFS) Version 4 Protocol", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-rfc3530bis-27.txt>>.
- Work in progress.
- [RFC5661] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.

### 11.2. Informative References

- [NFS-ext] Noveck, D., "NFS Protocol Extension: Retrospect and Prospect", 2013, <<http://www.ietf.org/id/draft-dnoveck-nfs-extension-00.txt>>.
- Work in progress.
- [migr-v4.0-update]

Noveck, D., Ed., Shivam, P., Lever, C., and B. Baker,  
"NFSv4.0 migration: Specification Update", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-rfc3530-migration-update-02.txt>>.

Work in progress.

#### Authors' Addresses

David Noveck (editor)  
EMC Corporation  
228 South Street  
Hopkinton, MA 01748  
US

Phone: +1 508 249 5748  
Email: david.noveck@emc.com

Piyush Shivam  
Oracle Corporation  
5300 Riata Park Ct.  
Austin, TX 78727  
US

Phone: +1 512 401 1019  
Email: piyush.shivam@oracle.com

Charles Lever  
Oracle Corporation  
1015 Granger Avenue  
Ann Arbor, MI 48104  
US

Phone: +1 248 614 5091  
Email: chuck.lever@oracle.com

Bill Baker  
Oracle Corporation  
5300 Riata Park Ct.  
Austin, TX 78727  
US

Phone: +1 512 401 1081  
Email: bill.baker@oracle.com

NFSv4  
Internet-Draft  
Updates: 3530 (if approved)  
Intended status: Standards Track  
Expires: April 15, 2014

D. Noveck, Ed.  
EMC  
P. Shivam  
C. Lever  
B. Baker  
ORACLE  
October 12, 2013

NFSv4.0 migration: Specification Update  
draft-ietf-nfsv4-rfc3530-migration-update-03

## Abstract

The migration feature of NFSv4 allows for responsibility for a single filesystem to move from one server to another, without disruption to clients. Recent implementation experience has shown problems in the existing specification for this feature in NFSv4.0. This document clarifies and corrects the NFSv4.0 specification (RFC3530 and possible successors) to address these problems.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 15, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Conventions . . . . .	3
3. Background . . . . .	3
4. Client Identity Definition . . . . .	5
4.1. Differences from Replaced Sections . . . . .	5
4.2. Client Identity data items . . . . .	5
4.3. Server Release of Client ID . . . . .	10
4.4. client id string Approaches . . . . .	10
4.5. Non-Uniform client id string Approach . . . . .	12
4.6. Uniform client id string Approach . . . . .	13
4.7. Mixing client id string Approaches . . . . .	14
4.8. Trunking Determination when using Uniform client id strings . . . . .	16
4.9. Client id string construction details . . . . .	21
5. Locking and Multi-Server Namespace . . . . .	22
5.1. Changes from Replaced Sections . . . . .	22
5.2. Lock State and Filesystem Transitions . . . . .	23
5.3. Migration and State . . . . .	23
5.3.1. Migration and clientid's . . . . .	24
5.3.2. Migration and state owner information . . . . .	26
5.4. Replication and State . . . . .	29
5.5. Notification of Migrated Lease . . . . .	30
5.6. Migration and the Lease_time Attribute . . . . .	33
6. Server Implementation Considerations . . . . .	33
6.1. Relation of Locking State Transfer to Other Aspects of Filesystem Motion . . . . .	33
6.2. Preventing Locking State Modification During Transfer . .	35
7. Additional Changes . . . . .	38
7.1. Summary of Additional Changes from Previous Documents . .	38
7.2. NFS4ERR_CLID_INUSE definition . . . . .	39
7.3. NFS4ERR_DELAY return from RELEASE_LOCKOWNER . . . . .	39
7.4. Operation 35: SETCLIENTID - Negotiate Client ID . . . . .	40
7.5. Security Considerations revision . . . . .	44
8. Security Considerations . . . . .	44
9. IANA Considerations . . . . .	44
10. Acknowledgements . . . . .	44
11. References . . . . .	45
11.1. Normative References . . . . .	45
11.2. Informative References . . . . .	45
Authors' Addresses . . . . .	45

## 1. Introduction

This document is a standards track document which corrects the existing definitive specification of the NFSv4.0 protocol, in [RFC3530] and the one expected to become definitive (now in [cur-rfc3530-bis]). Given this fact, one should take the current document into account when learning about NFSv4.0, particularly if one is concerned with issues that relate to:

- o Filesystem migration, particularly when it involves transparent state migration.
- o The construction and interpretation of the `nfs_clientid4` structure and particularly the requirements on the id string within it, referred to below as a "client id string".

## 2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

## 3. Background

Implementation experience with transparent state migration has exposed a number of problems with the existing specification of this feature, in [RFC3530] and in RFC3530bis (see the draft at [cur-rfc3530-bis]). The symptoms were:

- o After migration of a filesystem, a reboot of the associated client was not appropriately dealt with, in that the state associated with the rebooting client was not promptly freed.
- o Situations can arise whereby a given server has multiple leases with the same `nfs_client_id4` (id and verifier), when the protocol clearly assumes there can be only one.
- o Excessive client implementation complexity since clients have to deal with situations in which a single client can wind up with its locking state with a given server divided among multiple leases each with its own `clientid4`.

An analysis of these symptoms leads to the conclusion that existing specifications have erred. They assume that locking state, including both state ids and `clientid4`'s, should be transferred as part of transparent state migration. The troubling symptoms arise from the failure to describe how migrating state is to be integrated with existing client definition structures on the destination server.

Specification of requirements for the server to appropriately merge stateids associated with a common client boot instance encounters a difficult problem. The issue is that the common client practice with regard to the presentation of unique strings specifying client identity makes it essentially impossible for the client to determine whether or not two stateids, originally generated on different servers are referable to the same client. This practice is allowed and endorsed, although not "RECOMMENDED", by existing NFSv4.0 specifications ([RFC3530] and RFC3530bis, whose current draft is at [cur-rfc3530-bis]).

To further complicate matters, upon prototyping of clients implementing an alternative approach, it has been found that there exist servers which do not work well with these new clients. It appears that current circumstances, in which a particular client implementation pattern had been adopted universally, has resulted in servers not being able to interoperate against alternate client implementation patterns. As a result, we have a situation which requires careful attention to compatibility issues to untangle.

This document updates the existing NFSv4.0 specifications ([RFC3530] and RFC3530bis, whose current draft is at [cur-rfc3530-bis]) as follows:

- o It makes clear that NFSv4.0 supports multiple approaches to the construction of client id strings, including that formerly endorsed by existing NFSV4.0 specifications, and currently widely deployed.
- o It addresses the potential compatibility issues that might arise for clients adopting a previously non-favored client id construction approach including the existence of servers which have problems with the new approach.
- o It gives some guidance regarding the factors that might govern clients' choice of a client id construction approach and RECOMMENDS that clients construct client id strings in manner that supports lease merger if they intend to support transparent state migration.
- o It specifies how state is to be transparently migrated, including defining how state that arrives at a new server as part of migration is to be merged into existing leases for clients connected to the target server.

- o It makes further clarifications and corrections to address cases where the specification text does not take proper account of the issues raised by state migration or where it has been found that the existing text is insufficiently clear.

For a more complete explanation of the choices made in addressing these issues, see [info-migr]).

#### 4. Client Identity Definition

This chapter is a replacement for sections 8.1.1 and 8.1.2 in [RFC3530] and for sections 9.1.1 and 9.1.2 in RFC3530bis (see the draft at [cur-rfc3530-bis]). The replaced sections are named "client ID" and "Server Release of Clientid."

It supersedes the replaced sections.

##### 4.1. Differences from Replaced Sections

Because of the need for greater attention to and careful description of this area, this chapter is much larger than the sections it replaces. The principal changes/additions made by this chapter are:

- o It corrects inconsistencies regarding the possible role or non-role of client IP address in construction of client id strings.
- o It clearly addresses the need to save client id strings or any changeable values that are used in their construction.
- o It provides a more complete description of circumstances leading to clientid4 invalidity and the appropriate recovery actions.
- o It presents, as valid alternatives, two approaches to client id string construction (named "uniform" and "non-uniform") and gives some implementation guidance to help implementers choose one or the other of these.
- o It adds a discussion of issues involved for clients in interacting with servers whose behavior is not consistent with use of uniform client id strings
- o It adds a description of how server behavior might be used by the client to determine server address trunking patterns.

##### 4.2. Client Identity data items

The NFSv4 protocol contains a number of protocol entities to identify clients and client-based entities, for locking-related purposes:

- o The `nfs_client_id4` structure which uniquely identifies a specific client boot instance. That identification is presented to the server by doing a `SETCLIENTID` operation.
- o The `clientid4` which is returned by the server upon completion of a successful `SETCLIENTID` operation. This id is used by the client to identify itself when doing subsequent locking-related operations. A `clientid4` is associated with a particular lease whereby a client instance holds state on a server instance and may become invalid due to client reboot, server reboot, or other circumstances.
- o Opaque arrays which are used together with the `clientid4` to designate within-client entities (e.g. processes) as the owners of opens (open-owners) and owners of byte-range locks (lock-owners).

The basis of the client identification infrastructure is encapsulated in the following data structure:

```
struct nfs_client_id4 {  
    verifier4      verifier;  
    opaque         id<NFS4_OPAQUE_LIMIT>;  
};
```

The `nfs_client_id4` structure uniquely defines a client boot instance as follows:

- o The `id` field is a variable-length string which uniquely identifies a specific client. Although, we describe it as a string and it is often referred to as a "client string," it should be understood that the protocol defines this as opaque data. In particular, those receiving such an id should not assume that it will be in the UTF-8 encoding. Servers MUST NOT reject an `nfs_client_id4` simply because the id string does not follow the rules of UTF-8 encoding.

The string MAY be different for each server network address that the client accesses, rather than common to all server network addresses.

- o The `verifier` is a client incarnation identifier that is used by the server to detect client reboots. Only if the `verifier` is different from that which the server has previously recorded in connection with the client (as identified by the `id` field) does the server cancel the client's leased state, once it receives confirmation of the new `nfs_clientd4` via `SETCLIENTID_CONFIRM`.



As a security measure, the server MUST NOT cancel a client's leased state if the principal that established the state for a given id string is not the same as the principal issuing the SETCLIENTID.

There are several considerations for how the client generates the id string:

- o The string should be unique so that multiple clients do not present the same string. The consequences of two clients presenting the same string range from one client getting an error to one client having its leased state abruptly and unexpectedly canceled.
- o The string should be selected so that subsequent incarnations (e.g., reboots) of the same client cause the client to present the same string. The implementer is cautioned against an approach that requires the string to be recorded in a local file because this precludes the use of the implementation in an environment where there is no local disk and all file access is from an NFSv4 server.
- o The string MAY be different for each server network address that the client accesses, rather than common to all server network addresses.

The considerations that might influence a client to use different strings for different network server addresses are explained in Section 4.4.

- o The algorithm for generating the string should not assume that the client's network address is forever fixed. Changes might occur between client incarnations and even while the client is still running in its current incarnation.

Having the client id string change simply because of a network address change would mean that successive SETCLIENTID operations for the same client would appear as from different clients, interfering with the use of the `nfs_client_id4` verifier to cancel state associated with previous boot instances of the same client.

The difficulty is more severe if the client address is the only client-based information in the client id string. In such a case, there is a real risk that, after the client gives up the network address, another client, using a similar algorithm for generating the id string, will generate a conflicting id string.

Once a SETCLIENTID and SETCLIENTID\_CONFIRM sequence has successfully completed, the client uses the shorthand client identifier, of type `clientid4`, instead of the longer and less compact `nfs_client_id4` structure. This shorthand client identifier (a client ID) is assigned by the server and should be chosen so that it will not conflict with a client ID previously assigned by same server. This applies across server restarts or reboots.

Note that the SETCLIENTID and SETCLIENTID\_CONFIRM operations have a secondary purpose of establishing the information the server needs to make callbacks to the client for the purpose of supporting delegations. The client is able to change this information via SETCLIENTID and SETCLIENTID\_CONFIRM within the same incarnation of the client without causing removal of the client's leased state.

Distinct servers MAY assign `clientid4`'s independently, and will generally do so. Therefore, a client has to be prepared to deal with multiple instances of the same `clientid4` value received on distinct IP addresses, denoting separate entities. When trunking of server IP addresses is not a consideration, a client should keep track of (IP-address, `clientid4`) pairs, so that each pair is distinct. For a discussion of how to address the issue in the face of possible trunking of server IP addresses, see Section 4.4.

Owners of opens and owners of byte-range locks are separate entities and remain separate even if the same opaque arrays are used to designate owners of each. The protocol distinguishes between open-owners (represented by `open_owner4` structures) and lock-owners (represented by `lock_owner4` structures).

Both sorts of owners consist of a `clientid4` and an opaque owner string. For each client, the set of distinct owner values used with that client constitutes the set of owners of that type, for the given client.

Each open is associated with a specific open-owner while each byte-range lock is associated with a lock-owner and an open-owner, the latter being the open-owner associated with the open file under which the LOCK operation was done.

When a `clientid4` is presented to a server and that `clientid4` is not valid, the server will reject the request with the an error that depends on the reason for `clientid4` invalidity. The error `NFS4ERR_ADMIN_REVOKED` is returned when the invalidation is the result of administrative action, When the `clientid4` is unrecognizable, the error `NFS4ERR_STALE_CLIENTID` or `NFS4ERR_EXPIRED` may be returned. An unrecognizable `clientid4` can occur for a number of reasons:

- o A server reboot causing loss of the server's knowledge of the client. (Always returns NFS4ERR\_STALE\_CLIENTID)
- o Client error sending an incorrect clientid4 or a valid clientid4 to the wrong server. (May return either error).
- o Loss of lease state due to lease expiration. (Always returns NFS4ERR\_EXPIRED)
- o Client or server error causing the server to believe that the client has rebooted (i.e. receiving a SETCLIENTID with an nfs\_client\_id4 which has a matching id string and a non-matching boot verifier). (May return either error).
- o Migration of all state under the associated lease causes its non-existence to be recognized on the source server. (Always returns NFS4ERR\_STALE\_CLIENTID)
- o Merger of state under the associated lease with another lease under a different clientid causes the clientid4 serving as the source of the merge to cease being recognized on its server. (Always returns NFS4ERR\_STALE\_CLIENTID)

In the event of a server reboot, loss of lease state due to lease expiration, or administrative revocation of a clientid4, the client must obtain a new clientid4 by use of the SETCLIENTID operation and then proceed to any other necessary recovery for the server reboot case (See the section entitled "Server Failure and Recovery"). In cases of server or client error resulting in this error, use of SETCLIENTID to establish a new lease is desirable as well.

In the last two cases, different recovery procedures are required. See Section 5.3 for details. Note that in cases in which there is any uncertainty about which sort of handling is applicable, the distinguishing characteristic is that in reboot-like cases, the clientid4 and all associated stateids cease to exist while in migration-related cases, the clientid4 ceases to exist while the stateids are still valid.

The client must also employ the SETCLIENTID operation when it receives a NFS4ERR\_STALE\_STATEID error using a stateid derived from its current clientid4, since this indicates a situation, such as server reboot which has invalidated the existing clientid4 and associated stateids (see the section entitled "lock-owner" for details).

See the detailed descriptions of SETCLIENTID and SETCLIENTID\_CONFIRM for a complete specification of these operations.

#### 4.3. Server Release of Client ID

If the server determines that the client holds no associated state for its `clientid4`, the server may choose to release that `clientid4`. The server may make this choice for an inactive client so that resources are not consumed by those intermittently active clients. If the client contacts the server after this release, the server must ensure the client receives the appropriate error so that it will use the `SETCLIENTID/SETCLIENTID_CONFIRM` sequence to establish a new identity. It should be clear that the server must be very hesitant to release a client ID since the resulting work on the client to recover from such an event will be the same burden as if the server had failed and restarted. Typically a server would not release a client ID unless there had been no activity from that client for many minutes.

Note that if the id string in a `SETCLIENTID` request is properly constructed, and if the client takes care to use the same principal for each successive use of `SETCLIENTID`, then, barring an active denial of service attack, `NFS4ERR_CLID_INUSE` should never be returned.

However, client bugs, server bugs, or perhaps a deliberate change of the principal owner of the id string (such as may occur in the case in which a client changes security flavors, and under the new flavor, there is no mapping to the previous owner) will in rare cases result in `NFS4ERR_CLID_INUSE`.

In that event, when the server gets a `SETCLIENTID` specifying a client id string for which the server has a `clientid4` that currently has no state, or for which it has state, but where the lease has expired, the server **MUST** allow the `SETCLIENTID`, rather than returning `NFS4ERR_CLID_INUSE`. The server **MUST** then confirm the new client ID if followed by the appropriate `SETCLIENTID_CONFIRM`.

#### 4.4. client id string Approaches

One particular aspect of the construction of the `nfs_client_id4` string has proved recurrently troublesome. The client has a choice of:

- o Presenting the same id string to multiple server addresses. This is referred to as the "uniform client id string approach" and is discussed in Section 4.6.
- o Presenting different id strings to multiple server addresses. This is referred to as the "non-uniform client id string approach" and is discussed in Section 4.5.

Note that implementation considerations, including compatibility with existing servers, may make it desirable for a client to use both approaches, based on configuration information, such as mount options. This issue will be discussed in Section 4.7.

Construction of the client id string has arisen as a difficult issue because of the way in which the NFS protocols have evolved.

- o NFSv3 as a stateless protocol had no need to identify the state shared by a particular client-server pair. (See [RFC1813]). Thus there was no occasion to consider the question of whether a set of requests come from the same client, or whether two server IP addresses are connected to the same server. As the environment was one in which the user supplied the target server IP address as part of incorporating the remote filesystem in the client's file name space, there was no occasion to take note of server trunking. Within a stateless protocol, the situation was symmetrical. The client has no server identity information and the server has no client identity information.
- o NFSv4.1 is a stateful protocol with full support for client and server identity determination (See [RFC5661]). This enables the server to be aware when two requests come from the same client (they are on sessions sharing a clientid4) and the client to be aware when two server IP addresses are connected to the same server (they return the same server name in responding to an EXCHANGE\_ID).

NFSv4.0 is unfortunately halfway between these two. The two client id string approaches have arisen in attempts to deal with the changing requirements of the protocol as implementation has proceeded and features that were not very substantial in early implementations of [RFC3530], became more substantial as implementation proceeded.

- o In the absence of any implementation of the fs\_locations-related features (replication, referral, and migration), the situation is very similar to that of NFSv3, with the addition of state but with no concern to provide accurate client and server identity determination. This is the situation that gave rise to the non-uniform client id string approach.
- o In the presence of replication and referrals, the client may have occasion to take advantage of knowledge of server trunking information. Even more important, transparent state migration, by transferring state among servers, causes difficulties for the non-uniform client id string approach, in that the two different client id strings sent to different IP addresses may wind up on the same IP address, adding confusion.

- o A further consideration is that client implementations typically provide NFSv4.1 by augmenting their existing NFSv4.0 implementation, not by providing two separate implementations. Thus the more NFSv4.0 and NFSv4.1 can work alike, the less complex are clients. This is a key reason why those implementing NFSv4.0 clients might prefer using the uniform client string model, even if they have chosen not to provide fs\_locations-related features in their NFSv4.0 client.

Both approaches have to deal with the asymmetry in client and server identity information between client and server. Each seeks to make the client's and the server's views match. In the process, each encounters some combination of inelegant protocol features and/or implementation difficulties. The choice of which to use is up to the client implementer and the sections below try to give some useful guidance.

#### 4.5. Non-Uniform client id string Approach

The non-uniform client id string approach is an attempt to handle these matters in NFSv4.0 client implementations in as NFSv3-like a way as possible.

For a client using the non-uniform approach, all internal recording of clientid4 values is to include, whether explicitly or implicitly, the server IP address so that one always has an (IP-address, clientid4) pair. Two such pairs from different servers are always distinct even when the clientid4 values are the same, as they may occasionally be. In this approach, such equality is always treated as simple happenstance.

Making the client id string different on different server IP addresses results in a situation in which a server has no way of tying together information from the same client, when the client accesses multiple server IP addresses. As a result, it will treat a single client as multiple clients with separate leases for each server network address. Since there is no way in the protocol for the client to determine if two network addresses are connected to the same server, the resulting lack of knowledge is symmetrical and can result in simpler client implementations in which there is a single clientid/lease per server network addresses.

Support for migration, particularly with transparent state migration, is more complex in the case of non-uniform client id strings. For example, migration of a lease can result in multiple leases for the same client accessing the same server addresses, vitiating many of the advantages of this approach. Therefore, client implementations that support migration with transparent state migration SHOULD NOT

use the non-uniform client id string approach, except where it is necessary for compatibility with existing server implementations (For details of arranging use of multiple client id string approaches, see Section 4.7).

#### 4.6. Uniform client id string Approach

When the client id string is kept uniform, the server has the basis to have a single clientid4/lease for each distinct client. The problem that has to be addressed is the lack of explicit server identity information, which was made available in NFSv4.1.

When the same client id string is given to multiple IP addresses, the client can determine whether two IP addresses correspond to a single server, based on the server's behavior. This is the inverse of the strategy adopted for the non-uniform approach in which different server IP addresses are told about different clients, simply to prevent a server from manifesting behavior that is inconsistent with there being a single server for each IP address, in line with the traditions of NFS. So, to compare:

- o In the non-uniform approach, servers are told about different clients because, if the server were to use accurate information as to client identity, two IP addresses on the same server would behave as if they were talking to the same client, which might prove disconcerting to a client not expecting such behavior.
- o In the uniform approach, the servers are told about there being a single client, which is, after all, the truth. Then, when the server uses this information, two IP addresses on the same server will behave as if they are talking to the same client, and this difference in behavior allows the client to infer the server IP address trunking configuration, even though NFSv4.0 does not explicitly provide this information.

The approach given in the section below shows one example of how this might be done.

The uniform client id string approach makes it necessary to exercise more care in the definition of the `nfs_client_id4` boot verifier:

- o In [RFC3530], the client is told to change the boot verifier when reboot occurs, but there is no explicit statement as to the converse, so that any requirement to keep the verifier constant unless rebooting is only present by implication.
- o Many existing clients change the boot verifier every time they destroy and recreate the data structure that tracks an <IP-

address, clientid4> pair. This might happen if the last mount of a particular server is removed, and then a fresh mount is created. Also, note that this might result in each <IP-address, clientid4> pair having its own boot verifier that is independent of the others.

- o Within the uniform client id string approach, an `nfs_client_id4` designates a globally known client instance, so that the boot verifier should change if and only if a new client instance is created, typically as a result of a reboot.

The following are advantages for the implementation of using the uniform client id string approach:

- o Clients can take advantage of server trunking (and clustering with single-server-equivalent semantics) to increase bandwidth or reliability.
- o There are advantages in state management so that, for example, we never have a delegation under one clientid revoked because of a reference to the same file from the same client under a different clientid.
- o The uniform client id string approach allows the server to do any necessary automatic lease merger in connection with transparent state migration, without requiring any client involvement. This consideration is of sufficient weight to cause us to RECOMMEND use of the uniform client id string approach for clients supporting transparent state migration.

The following implementation considerations might cause issues for client implementations.

- o This approach is considerably different from the non-uniform approach, which most client implementations have been following. Until substantial implementation experience is obtained with this approach, reluctance to embrace something so new is to be expected.
- o Mapping between server network addresses and leases is more complicated in that it is no longer a one-to-one mapping.

How to balance these considerations depends on implementation goals.

#### 4.7. Mixing client id string Approaches

As noted above, a client which needs to use the uniform client id string approach (e.g. to support migration), may also need to support



existing servers with implementations that do not work properly in this case.

Some examples of such server issues include:

- o Some existing NFSv4.0 server implementations of IP address failover depend on clients' use of a non-uniform client id string approach. In particular, when a server supports both its own IP address and one failed over from a partner server, it may have separate sets of state applicable to the two IP addresses, owned by different servers but residing on a single one.

In this situation, some servers have relied on clients' use of the non-uniform client id string approach, as suggested but not mandated by [RFC3530], to keep these sets of state separate, and will have problems in handling clients using the uniform client id string approach, in that such clients will see changes in trunking relationships whenever server failover and giveback occur.

- o Some existing servers incorrectly return NFS4ERR\_CLID\_INUSE simply because there already exists a clientid for the same client, established using a different IP address. This causes difficulty for a multi-homed client using the uniform client id string approach.

Although this behavior is not correct, such servers still exist and the spec should give clients guidance about dealing with the situation, as well as making the correct behavior clear.

In order to support use of these sorts of servers, the client can use different client id string approaches for different mounts, as long as:

- o The uniform client id string approach is used when accessing servers that may return NFS4ERR\_MOVED and the client wishes to enable transparent state migration."
- o The non-uniform client id string approach is used when accessing servers whose implementations make them incompatible with the uniform client id string approach

One effective way for clients to handle this is to support the uniform client id string approach as the default, but allow a mount option to specify use of the non-uniform client id string approach for particular mount points, as long as such mount points are not used when migration is to be supported.

In the case in which the same server has multiple mounts, and both approaches are specified for the same server, the client could have multiple clientid's corresponding to the same server, one for each approach and would then have to keep these separate.

#### 4.8. Trunking Determination when using Uniform client id strings

This section provides an example of how trunking determination could be done by a client following the uniform client id string approach (whether this is used for all mounts or not). Clients need not follow this procedure but implementers should make sure that the issues dealt with by this procedure are all properly addressed.

We need to clarify the various possible purposes of trunking determination and the corresponding requirements as to server behavior. The following points should be noted:

- o The primary purpose of the trunking determination algorithm is to make sure that, if the server treats client requests on two IP addresses as part of the same client, the client will not be blind-sided and encounter disconcerting server behavior, as mentioned in Section 4.6. Such behavior could occur if the client were unaware that all of its client requests for the two IP addresses were being handled as part of a single client talking to a single server.
- o A second purpose is to be able to use knowledge of trunking relationships for better performance, etc.
- o If a server were to give out distinct clientid's in response to receiving the same nfs\_client\_id4 on different network addresses, and acted as if these were separate clients, the primary purpose of trunking determination would be met, as long as the server did not treat them as part of the same client. In this case, the server would be acting, with regard to that client, as if it were two distinct servers. This would interfere with the secondary purpose of trunking determination but there is nothing the client can do about that.
- o Suppose a server were to give such a client two different clientid's but act as if they were one. That is the only way that the server could behave in a way that would defeat the primary purpose of the trunking determination algorithm.

Servers MUST NOT do that.

For a client using the uniform approach, clientid4 values are treated as important information in determining server trunking patterns.

For two different IP addresses to return the same clientid4 value is a necessary, though not a sufficient condition for them to be considered as connected to the same server. As a result, when two different IP addresses return the same clientid4, the client needs to determine, using the procedure given below or otherwise, whether the IP addresses are connected to the same server. For such clients, all internal recording of clientid4 values needs to include, whether explicitly or implicitly, identification of the server from which the clientid4 was received so that one always has a (server, clientid4) pair. Two such pairs from different servers are always considered distinct even when the clientid4 values are the same, as they may occasionally be.

In order to make this approach work, the client must have accessible, for each nfs\_client\_id4 used by the uniform approach (only one in general) a list of all server IP addresses, together with the associated clientid4 values, SETCLIENTID principals and authentication flavors. As a part of the associated data structures, there should be the ability to mark a server IP structure as having the same server as another and to mark an IP address as currently unresolved. One way to do this is to allow each such entry to point to another with the pointer value being one of:

- o A pointer to another entry for an IP address associated with the same server, where that IP address is the first one referenced to access that server.
- o A pointer to the current entry if there is no earlier IP address associated with the same server, i.e. where the current IP address is the first one referenced to access that server. We'll refer to such an IP address as the lead IP address for a given server.
- o The value NULL if the address's server identity is currently unresolved.

In order to keep the above information current, in the interests of the most effective trunking determination, RENEWS should be periodically done on each server. However, even if this is not done, the primary purpose of the trunking determination algorithm, to prevent confusion due to trunking hidden from the client, will be achieved.

Given this apparatus, when a SETCLIENTID is done and a clientid4 returned, the data structure can be searched for a matching clientid4 and if such is found, further processing can be done to determine whether the clientid4 match is accidental, or the result of trunking.

In this algorithm, when SETCLIENTID is done it will use the common `nfs_client_id4` and specify the current target IP address as part of the callback parameters. We call the `clientid4` and SETCLIENTID verifier returned by this operation XC and XV.

Note that when the client has done previous SETCLIENTID's, to any IP addresses, with more than one principal or authentication flavor, we have the possibility of receiving NFS4ERR\_CLID\_INUSE, since we do not yet know which of our connections with existing IP addresses might be trunked with our current one. In the event that the SETCLIENTID fails with NFS4ERR\_CLID\_INUSE, one must try all other combinations of principals and authentication flavors currently in use and eventually one will be correct and not return NFS4ERR\_CLID\_INUSE.

Note that at this point, no SETCLIENTID\_CONFIRM has yet been done. This is because our SETCLIENTID has either established a new `clientid4` on a previously unknown server or changed the callback parameters on a `clientid4` associated with some already known server. Given that we don't want to confirm something that we are not sure we want to happen, what is to be done next depends on information about existing `clientid4`'s.

- o If no matching `clientid4` is found, the IP address X and `clientid4` XC are added to the list and considered as having no existing known IP addresses trunked with it. The IP address is marked as a lead IP address for a new server. A SETCLIENTID\_CONFIRM is done using XC and XV.
- o If a matching `clientid4` is found which is marked unresolved, processing on the new IP address is suspended. In order to simplify processing, there can only be one unresolved IP address for any given `clientid4`.
- o If one or more matching `clientid4`'s is found, none of which is marked unresolved, the new IP address is entered and marked unresolved. After applying the steps below to each of the lead IP addresses with a matching `clientid4`, the address will have been resolved: It may be determined to be part of an already known server as a new IP address to be added to an existing set of IP addresses for that server. Otherwise, it will be recognized as a new server. At the point at which this determination is made, the unresolved indication is cleared and any suspended SETCLIENTID processing is restarted

So for each lead IP address IPn with a `clientid4` matching XC, the following steps are done.

- o If the principal for IPn does not match that for X, the IP address is skipped, since it is impossible for IPn and X to be trunked in these circumstances. If the principal does match but the authentication flavor does not, the authentication flavor already used should be used for address X as well. This will avoid any possibility that NFS4ERR\_CLID\_INUSE will be returned for the SETCLIENTID and SETCLIENTID\_CONFIRM to be done below, as long as the server(s) at IP addresses IPn and X are correctly implemented.
- o A SETCLIENTID is done to update the callback parameters to reflect the possibility that X will be marked as associated with the server whose lead IP address is IPn. The specific callback parameters chosen, in terms of cb\_client4 and callback\_ident, are up to the client and should reflect its preferences as to callback handling for the common clientid, in the event that X and IPn are trunked together. So assume that we do that SETCLIENTID on IP address IPn and get back a setclientid\_confirm value (in the form of a verifier4) SCn.

Note that the NFSv4.0 specification requires the server to make sure that such verifiers are very unlikely to be regenerated. Given that it is already highly unlikely that the clientid XC is duplicated by distinct servers, the probability that Sc is duplicated as well has to be considered vanishingly small. Note also that the callback update procedure can be repeated multiple times to reduce the probability of spurious matches further.

- o Note that we don't want this to happen if address X is not associated with this server. So we do a SETCLIENTID\_CONFIRM on address X using the setclientid\_confirm value SCn.
- o If the setclientid\_confirm value generated on X is accepted on IPn, then X and IPn are recognized as connected to the same server and the entry for X is marked as associated with IPn. The entry is now resolved and processing can be restarted for IP addresses whose clientid4 matched XC but whose resolution had been deferred.
- o If the confirm value generated on IPn is not accepted on X, then X and IPn are distinct and the callback update will not be confirmed. So we go on to the next IPn, until we run out of them. If it happens that we run out of potential matches, then we can treat X as connected to a distinct server and then update and confirm its callback parameters on that basis.

Note here that we may set a number of possible values for the callback parameters to be used for XC, one for the possibility that X is untrunked, and others for each potential match with an existing IPn. Although there are multiple such updates at most one will be

confirmed and, if X is untrunked, its original callback parameters will be put in effect by its SETCLIENTID\_CONFIRM.

The procedure described above must be performed so as to exclude the possibility that multiple SETCLIENTID's, done to different server IP addresses and returning the same clientid4 might "race" in such a fashion that there is no explicit determination of whether they correspond to the same server. The following possibilities for serialization are all valid and implementers may choose among them based on a tradeoff between performance and complexity. They are listed in order of increasing parallelism:

- o An NFSv4.0 client might serialize all instances of SETCLIENTID/SETCLIENTID\_CONFIRM processing, either directly or by serializing mount operations involving use of NFSv4.0. While doing so will prevent the races mentioned above, this degree of serialization can cause performance issues when there is a high volume of mount operations.
- o One might instead serialize the period of processing that begins when the clientid4 received from the server is processed and ends when all trunking determination for that server is completed. This prevents the races mentioned above, without adding to delay except when trunking determination is common.
- o One might avoid much of the serialization implied above, by allowing trunking determination for distinct clientid4 values to happen in parallel, with serialization of trunking determination happening independently for each distinct clientid4 value.

The procedure above has made no explicit mention of the possibility that server reboot can occur at any time. To address this possibility the client should periodically use the clientid4 XC in RENEW operations, directed to both the IP address X and the current lead IP address that is currently being tested for identity.

- o When XC becomes invalid on X, the resolution process should be terminated, subject to being redone later. Before redoing the resolution, XC should be checked on all the lead IP addresses on which it was valid. Once a new clientid4 is established on any servers on which XC became invalid, a new clientid4 can be established on X and the resolution process for X can be restarted.
- o When XC does not become invalid on X, but becomes invalid on the current IPn being tested, it should be concluded that X and IPn do not match and that it is time to advance to the next IPn, if any.

- o In the event of a reboot detected on any server lead IP, the set of IP addresses associated with the server should not change and state should be re-established for the lease as a whole, using all available connected server IP addresses. It is prudent to verify connectivity by doing a RENEW using the new clientid4 on each such server address before using it, however.

If we have run out of IPn's without finding a matching server, X is considered as having no existing known IP addresses trunked with it. The IP address is marked as a lead IP address for a new server. A SETCLIENTID\_CONFIRM is done using XC and XV.

#### 4.9. Client id string construction details

This section gives more detailed guidance on client id construction. Note that among the items suggested for inclusion, there are many that may conceivably change. In order for the client id string to remain valid in such circumstances, the client should either:

- o Use a saved copy of such value, rather than the changeable value itself.
- o Save the constructed client id string, rather than constructing it anew at SETCLIENTID time, based on unchangeable parameters and saved copies of changeable data items.

A file is not always a valid choice to store such information, given the existence of diskless clients. In such situations, whatever facilities exist for a client to store configuration information such as boot arguments should be used.

Given the considerations listed in Section 4.2, an example of a well generated id string is one that includes:

- o The client's network address, or more safely, an address that has previously been used in that capacity.
- o For a user level NFSv4.0 client, it should contain additional information to distinguish the client from other user level clients running on the same host, such as a universally unique identifier (UUID).
- o Additional information that tends to be unique, such as one or more of:
  - \* The client machine's serial number (for privacy reasons, it is best to perform some one way function on the serial number).

- \* A MAC address. Note that this can cause difficulties when there are configuration changes or when a client has multiple network adapters.
- \* The timestamp of when the NFSv4 software was first installed on the client (though this is subject to the previously mentioned caution about using information that is stored in a file, because the file might only be accessible over NFSv4).
- \* A true random number, generally established once and saved.

## 5. Locking and Multi-Server Namespace

This chapter is a replacement for section 7.7.6, "Lock State and File System transitions", in RFC3530bis (see the draft at [cur-rfc3530-bis]).

With respect to [RFC3530], it serves as a replacement for section 8.14, "Migration, Replication, and State".

It supersedes the replaced sections.

### 5.1. Changes from Replaced Sections

These changes can be briefly summarized as follows:

- o Adding text to address the case of stateid conflict on migration.
- o Specifying that when leases are moved, as a result of filesystem migration, they are to be merged with leases on the destination server that are connected to the same client.
- o Adding text that deals with the case of a clientid4 being changed on state transfer as a result of conflict with an existing clientid4.
- o Adding a section describing how information associated with openowners and lockowners is to be managed with regard to migration.
- o The description of handling of the NFS4ERR\_LEASE\_MOVED has been rewritten for greater clarity.



## 5.2. Lock State and Filesystem Transitions

When responsibility for handling a given filesystem is transferred to a new server (migration) or the client chooses to use an alternate server (e.g., in response to server unresponsiveness) in the context of filesystem replication, the appropriate handling of state shared between the client and server (i.e., locks, leases, stateids, and client IDs) is as described below. The handling differs between migration and replication.

If a server replica or a server immigrating a filesystem agrees to, or is expected to, accept opaque values from the client that originated from another server, then it is a wise implementation practice for the servers to encode the "opaque" values in network byte order. When doing so, servers acting as replicas or immigrating filesystems will be able to parse values like stateids, directory cookies, filehandles, etc. even if their native byte order is different from that of other servers cooperating in the replication and migration of the filesystem.

## 5.3. Migration and State

In the case of migration, the servers involved in the migration of a filesystem SHOULD transfer all server state associated with the migrating filesystem from source to the destination server. This must be done in a way that is transparent to the client. This state transfer will ease the client's transition when a filesystem migration occurs. If the servers are successful in transferring all state, the client will continue to use stateids assigned by the original server. Therefore the new server must recognize these stateids as valid and treat them as representing the same locks as they did on the source server.

In this context, the phrase "the same locks" means:

- o That they are associated with the same file
- o That they represent the same types of locks, whether opens, delegations, advisory byte-range locks, or mandatory byte-range locks.
- o That they have the same lock particulars, including such things as access modes, deny modes, and byte ranges.
- o That they are associated with the same owner string(s).

If transferring stateids from server to server would result in a conflict for an existing stateid for the destination server with the

existing client, transparent state migration MUST NOT happen for that client. Servers participating in using transparent state migration should co-ordinate their stateid assignment policies to make this situation unlikely or impossible. The means by which this might be done, like all of the inter-server interactions for migration, are not specified by the NFS version 4.0 protocol.

A client may determine the disposition of migrated state by using a stateid associated with the migrated state on the new server.

- o If the stateid is not valid and an error NFS4ERR\_BAD\_STATEID is received, either transparent state migration has not occurred or the state was purged due to boot verifier mismatch.
- o If the stateid is valid, transparent state migration has occurred.

Since responsibility for an entire filesystem is transferred with a migration event, there is no possibility that conflicts will arise on the destination server as a result of the transfer of locks.

The servers may choose not to transfer the state information upon migration. However, this choice is discouraged, except where specific issues such as stateid conflicts make it necessary. When a server implements migration and it does not transfer state information, it SHOULD provide a filesystem-specific grace period, to allow clients to reclaim locks associated with files in the migrated filesystem. If it did not do so, clients would have to re-obtain locks, with no assurance that a conflicting lock was not granted after the filesystem was migrated and before the lock was re-obtained.

In the case of migration without state transfer, when the client presents state information from the original server (e.g. in a RENEW op or a READ op of zero length), the client must be prepared to receive either NFS4ERR\_STALE\_CLIENTID or NFS4ERR\_BAD\_STATEID from the new server. The client should then recover its state information as it normally would in response to a server failure. The new server must take care to allow for the recovery of state information as it would in the event of server restart.

In those situations in which state has not been transferred, as shown by a return of NFS4ERR\_BAD\_STATEID, the client may attempt to reclaim locks in order to take advantage of cases in which the destination server has set up a file-system-specific grace period in support of the migration.

#### 5.3.1. Migration and clientid's

Handling of clientid values is similar to that for stateids. However, there are some differences that derive from the fact that a clientid is an object which spans multiple filesystems while a stateid is inherently limited to a single filesystem.

The clientid4 and nfs\_client\_id4 information (id string and boot verifier) will be transferred with the rest of the state information and the destination server should use that information to determine appropriate clientid4 handling. Although the destination server may make state stored under an existing lease available under the clientid4 used on the source server, the client should not assume that this is always so. In particular,

- o If there is an existing lease with an nfs\_client\_id4 that matches a migrated lease (same id string and boot verifier), the server SHOULD merge the two, making the union of the sets of stateids available under the clientid4 for the existing lease. As part of the lease merger, the expiration time of the lease will reflect renewal done within either of the ancestor leases (and so will reflect the latest of the renewals).
- o If there is an existing lease with an nfs\_client\_id4 that partially matches a migrated lease (same id string and a different boot verifier), the server MUST eliminate one of the two, possibly invalidating one of the ancestor clientid4's. Since boot verifiers are not ordered, the later lease renewal time will prevail.
- o If the destination server already has the transferred clientid4 in use for another purpose, it is free to substitute a different clientid4 and associate that with the transferred nfs\_client\_id4.

When leases are not merged, the transfer of state should result in creation of a confirmed client record with empty callback information but matching the {v, x, c} with v and x derived from the transferred client information and c chosen by the destination server.

In such cases, the client SHOULD re-establish new callback information with the new server as soon as possible, according to sequences described in sections "Operation 35: SETCLIENTID - Negotiate Client ID" and "Operation 36: SETCLIENTID\_CONFIRM - Confirm Client ID". This ensures that server operations are not delayed due to an inability to recall delegations. The client can determine the new clientid (the value c) from the response to SETCLIENTID.

The client can use its own information about leases with the destination server to see if lease merger should have happened. When there is any ambiguity, the client MAY use the above procedure to set

the proper callback information and find out, as part of the process, the correct value of its clientid with respect to the server in question.

#### 5.3.2. Migration and state owner information

In addition to stateids, the locks they represent, and clientid information, servers also need to transfer information related to the current status of openowners and lockowners.

This information includes:

- o The sequence number of the last operation associated with the particular owner.
- o Information regarding the results of the last operation, sufficient to allow reissued operations to be correctly responded to.

When clients are implemented to isolate each openowner and lockowner to a particular filesystem, the server SHOULD transfer this information together with the lock state. The owner ceases to exist on the source server and is reconstituted on the destination server.

Note that when servers take this approach for all owners whose state is limited to the particular filesystem being migrated, doing so will not cause difficulties for clients not adhering to an approach in which owners are isolated to particular filesystems. As long as the client recognizes the loss of transferred state, the protocol allows the owner in question to disappear and the client may have to deal with an owner confirmation request that would not have occurred in the absence of the migration.

When migration occurs and the source server discovers an owner whose state includes the migrated filesystem but other filesystems as well, it cannot transfer the associated owner state. Instead, the existing owner state stays in place but propagation of owner state is done as specified below

- o When the current seqid for an owner represents an operation associated with the filesystem being migrated, owner status SHOULD be propagated to the destination filesystem.
- o When the current seqid for an owner does not represent an operation associated with the filesystem being migrated, owner status MAY be propagated to the destination filesystem.

- o When the owner in question has never been used for an operation involving the migrated filesystem, the owner information SHOULD NOT be propagated to the destination filesystem.

Note that a server may obey all of the conditions above without the overhead of keeping track of set of filesystems that any particular owner has been associated with. Consider a situation in which the source server has decided to keep lock-related state associated with a filesystem fixed, preparatory to propagating it to the destination filesystem. If a client is free to create new locks associated with existing owners on other filesystems, the owner information may be propagated to the destination filesystem, even though, at the time the filesystem migration is recognized by the client to have occurred, the last operation associated with the owner may not be associated with the migrating filesystem.

When source server propagates owner-related state associated with owners that span multiple filesystems, it will propagate the owner sequence value to the destination server, while retaining it on the source server, as long as there exists state associated with the owner. When owner information is propagated in this way, source and destination servers start with the same owner sequence value which is then updated independently, as the client makes owner-related requests to the servers. Note that each server will have some period in which the associated sequence value for an owner is identical to the one transferred as part of migration. At those times, when a server receives a request with a matching owner sequence value, it MUST NOT respond with the associated stored response if the associated filesystem is not, when the reissued request is received, part of the set of filesystems handled by that server.

One sort of case may require more complex handling. When multiple filesystem are migrated, in sequence, to a specific destination server, an owner may be migrated to a destination server, on which it was already present, leading to the issue of how the resident owner information and that being newly migrated are to be reconciled.

If filesystem migration encounters a situation where owner information needs to be merged, it MAY decline to transfer such state, even if it chooses to handle other cases in which locks for a given owner are spread among multiple filesystems.

As a way of understanding the situations which need to be addressed when owner information needs to be merged, consider the following scenario:

- o There is client C and two servers X and Y. There are two clientid's designating C, which we refer to as CX and CY.

- o Initially server X supports filesystems F1, F2, F3, and F4. These will be migrated, one-at-a-time, to server Y.
- o While these migrations are proceeding, the client makes locking requests for filesystem F1 through F4 on behalf of owner O (either a lockowner or an openowner), with each request going to X or Y depending on where the relevant filesystem is being supported at the time the request is made.
- o Once the first migration event occurs, client C will maintain two instances for owner O, one for each server.
- o It is always possible that C may make a request of server X relating to owner O, and before receiving a response, find the target filesystem has moved to Y, and need to re-issue the request to server Y.
- o At the same time, C may make a request of server Y relating to owner O, and this too may encounter a lost-response situation.

As a result of such situations, the server needs to provide support for dealing with retransmission of owner-sequenced requests that diverges from the typical model in which there is support for retransmission of replies only for a request whose sequence value exactly matches the last one sent. Such support only needs to be provided for requests issued before the migration event whose status as the last by sequence is invalidated by the migration event.

When servers do support such merger of owner information on the destination server, the following rules are to be adhered to:

- o When an owner sequence value is propagated to a destination server where it already exists, the resulting sequence value is to be the greater of the one present on the destination server and the one being propagated as part of migration.
- o In the event that an owner sequence value on a server represents a request applying to a filesystem currently present on the server, it is not to be rendered invalid simply because that sequence value is changed as a result of owner information propagation as part of filesystem migration. Instead, it is retained until it can be deduced that the client in question has received the reply.

As a result of the operation of these rules, there are three ways in which we can have more reply data than what is typically present, i.e. data for a single request per owner whose sequence is the last one received, where the next sequence to be used is one beyond that.

- o When the owner sequence value for a migrating filesystem is greater than the corresponding value on the destination server, the last request for the owner in effect at the destination server needs to be retained, even though it is no longer one less the next sequence to be received.
- o When the owner sequence value for a migrating filesystem is less than the corresponding value on the destination server the last request for the owner in effect on the migrating filesystem needs to be retained, even though it is no longer one less the next sequence to be received.
- o When the owner sequence value for a migrating filesystem is equal to the corresponding value on the destination server, one has two different "last" requests which both must be retained. The next sequence value to be used is one beyond the sequence value shared by these two requests.

Here are some guidelines as to when servers can drop such additional reply data which is created as part of owner information migration.

- o The server SHOULD NOT drop this information simply because it receives a new sequence value for the owner in question, since that request may have been issued before the client was aware of the migration event.
- o The server SHOULD drop this information if it receives a new sequence value for the owner in question and the request relates to the same filesystem.
- o The server SHOULD drop the part of this information that relates to non-migrated filesystems, if it receives a new sequence value for the owner in question and the request relates to a non-migrated filesystem.
- o The server MAY drop this information when it receives a new sequence value for the owner in question a considerable period of time (more than one or two lease periods) after the migration occurs.

#### 5.4. Replication and State

Since client switch-over in the case of replication is not under server control, the handling of state is different. In this case, leases, stateids and client IDs do not have validity across a transition from one server to another. The client must re-establish its locks on the new server. This can be compared to the re-establishment of locks by means of reclaim-type requests after a

server reboot. The difference is that the server has no provision to distinguish requests reclaiming locks from those obtaining new locks or to defer the latter. Thus, a client re-establishing a lock on the new server (by means of a LOCK or OPEN request), may have the requests denied due to a conflicting lock. Since replication is intended for read-only use of filesystems, such denial of locks should not pose large difficulties in practice. When an attempt to re-establish a lock on a new server is denied, the client should treat the situation as if its original lock had been revoked.

#### 5.5. Notification of Migrated Lease

A filesystem can be migrated to another server while a client that has state related to that filesystem is not actively submitting requests to it. In this case, the migration is reported to the client during lease renewal. Lease renewal can occur either explicitly via a RENEW operation, or implicitly when the client performs a lease-renewing operation on another filesystem on that server.

In order for the client to schedule renewal of leases that may have been relocated to the new server, the client must find out about lease relocation before those leases expire. Similarly, when migration occurs but there has not been transparent state migration, the client needs to find out about the change soon enough to be able to reclaim the lock within the destination server's grace period. To accomplish this, all operations which implicitly renew leases for a client (such as OPEN, CLOSE, READ, WRITE, RENEW, LOCK, and others), will return the error NFS4ERR\_LEASE\_MOVED if responsibility for any of the leases to be renewed has been transferred to a new server. Note that when the transfer of responsibility leaves remaining state for that lease on the source server, the lease is renewed just as it would have been in the NFS4ERR\_OK case, despite returning the error. The transfer of responsibility happens when the server receives a GETATTR(fs\_locations) from the client for each filesystem for which a lease has been moved to a new server. Normally it does this after receiving an NFS4ERR\_MOVED for an access to the filesystem but the server is not required to verify that this happens in order to terminate the return of NFS4ERR\_LEASE\_MOVED. By convention, the compounds containing GETATTR(fs\_locations) SHOULD include an appended RENEW operation to permit the server to identify the client getting the information.

Note that the NFS4ERR\_LEASE\_MOVED error is only required when responsibility for at least one stateid has been affected. In the case of a null lease, where the only associated state is a clientid, an NFS4ERR\_LEASE\_MOVED error SHOULD NOT be generated.



Upon receiving the NFS4ERR\_LEASE\_MOVED error, a client that supports filesystem migration MUST perform the necessary GETATTR operation for each of the filesystems containing state that have been migrated and so give the server evidence that it is aware of the migration of the filesystem. Once the client has done this for all migrated filesystems on which the client holds state, the server MUST resume normal handling of stateful requests from that client.

One way in which clients can do this efficiently in the presence of large numbers of filesystems is described below. This approach divides the process into two phases, one devoted to finding the migrated filesystems and the second devoted to doing the necessary GETATTRs.

The client can find the migrated filesystems by building and issuing one or more COMPOUND requests, each consisting of a set of PUTFH/GETFH pairs, each pair using an fh in one of the filesystems in question. All such COMPOUND requests can be done in parallel. The successful completion of such a request indicates that none of the filesystems interrogated have been migrated while termination with NFS4ERR\_MOVED indicates that the filesystem getting the error has migrated while those interrogated before it in the same COMPOUND have not. Those whose interrogation follows the error remain in an uncertain state and can be interrogated by restarting the requests from after the point at which NFS4ERR\_MOVED was returned or by issuing a new set of COMPOUND requests for the filesystems which remain in an uncertain state.

Once the migrated filesystems have been found, all that is needed is for the client to give evidence to the server that it is aware of the migrated status of filesystems found by this process, by interrogating the fs\_locations attribute for an fh within each of the migrated filesystems. The client can do this by building and issuing one or more COMPOUND requests, each of which consists of a set of PUTFH operations, each followed by a GETATTR of the fs\_locations attribute. A RENEW is necessary to enable the operations to be associated with the lease returning NFS4ERR\_LEASE\_MOVED. Once the client has done this for all migrated filesystems on which the client holds state, the server will resume normal handling of stateful requests from that client.

In order to support legacy clients that do not handle the NFS4ERR\_LEASE\_MOVED error correctly, the server SHOULD time out after a wait of at least two lease periods, at which time it will resume normal handling of stateful requests from all clients. If a client attempts to access the migrated files, the server MUST reply NFS4ERR\_MOVED. In this situation, it is likely that the client would find its lease expired although a server may use "courtesy" locks to mitigate the issue.

When the client receives an NFS4ERR\_MOVED error, the client can follow the normal process to obtain the destination server information (through the fs\_locations attribute) and perform renewal of those leases on the new server. If the server has not had state transferred to it transparently, the client will receive either NFS4ERR\_STALE\_CLIENTID or NFS4ERR\_STALE\_STATEID from the new server, as described above. The client can then recover state information as it does in the event of server failure.

Aside from recovering from a migration, there are other reasons a client may wish to retrieve fs\_locations information from a server. When a server becomes unresponsive, for example, a client may use cached fs\_locations data to discover an alternate server hosting the same filesystem data. A client may periodically request fs\_locations data from a server in order to keep its cache of fs\_locations data fresh.

Since a GETATTR(fs\_locations) operation would be used for refreshing cached fs\_locations data, a server could mistake such a request as indicating recognition of an NFS4ERR\_LEASE\_MOVED condition. Therefore a compound which is not intended to signal that a client has recognized a migrated lease SHOULD be prefixed with a guard operation which fails with NFS4ERR\_MOVED if the file handle being queried is no longer present on the server. The guard can be as simple as a GETFH operation.

Though unlikely, it is possible that the target of such a compound could be migrated in the time after the guard operation is executed on the server but before the GETATTR(fs\_locations) operation is encountered. When a client issues a GETATTR(fs\_locations) operation as part of a compound not intended to signal recognition of a migrated lease, it SHOULD be prepared to process fs\_locations data in the reply that shows the current location of the filesystem is gone.

## 5.6. Migration and the Lease\_time Attribute

In order that the client may appropriately manage its leases in the case of migration, the destination server must establish proper values for the lease\_time attribute.

When state is transferred transparently, that state should include the correct value of the lease\_time attribute. The lease\_time attribute on the destination server must never be less than that on the source since this would result in premature expiration of leases granted by the source server. Upon migration in which state is transferred transparently, the client is under no obligation to re-fetch the lease\_time attribute and may continue to use the value previously fetched (on the source server).

In the case in which lease merger occurs as part of state transfer, the lease\_time attribute of the destination lease remains in effect. The client can simply renew that lease with its existing lease\_time attribute. State in the source lease is renewed at the time of transfer so that it cannot expire, as long as the destination lease is appropriately renewed.

If state has not been transferred transparently (i.e., the client needs to reclaim or re-obtain its locks), the client should fetch the value of lease\_time on the new (i.e., destination) server, and use it for subsequent locking requests. However the server must respect a grace period at least as long as the lease\_time on the source server, in order to ensure that clients have ample time to reclaim their locks before potentially conflicting non-reclaimed locks are granted. The means by which the new server obtains the value of lease\_time on the old server is left to the server implementations. It is not specified by the NFS version 4.0 protocol.

## 6. Server Implementation Considerations

This chapter provides suggestions to help server implementers deal with issues involved in the transparent transfer of filesystem-related data between servers. Servers are not obliged to follow these suggestions, but should be sure that their approach to the issues handle all the potential problems addressed below.

### 6.1. Relation of Locking State Transfer to Other Aspects of Filesystem Motion

In many cases, state transfer will be part of a larger function wherein the contents of a filesystem are transferred from server to server. Although specifics will vary with the implementation, the relation between the transfer of persistent file data and metadata

and the transfer of state will typically be described by one of the cases below.

- o In some implementations, access to the on-disk contents of a filesystem can be transferred from server to server by making the storage devices on which the filesystem resides physically accessible from multiple servers, and transferring the right and responsibility for handling that filesystem from server to server.

In such implementations, the transfer of locking state happens on its own, as described in Section 6.2. The transfer of physical access to the filesystem happens after the locking state is transferred and before any subsequent access to the filesystem. In cases where such transfer is not instantaneous, there will be a period in which all operations on the filesystem are held off, either by having the operations themselves return NFS4ERR\_DELAY, or, where this is not allowed, by using the techniques described below in Section 6.2.

- o In other implementations, filesystem data and metadata must be copied from the server where it has existed to the destination server. Because of the typical amounts of data involved, it is generally not practical to hold off access to the filesystem while this transfer is going on. Normal access to the filesystem, including modifying operations, will generally happen while the transfer is going on.

Eventually the filesystem copying process will complete. At this point, there will be two valid copies of the filesystem, one on each of the source and destination servers. Servers may maintain that state of affairs by making sure that each modification to filesystem data is done on both the source and destination servers.

Although the transfer of locking state can begin before the above state of affairs is reached, servers will often wait until it is arrived at to begin transfer of locking state. Once the transfer of locking state is completed, as described in the section below, clients may be notified of the migration event and access the destination filesystem on the destination server.

- o Another case in which filesystem data and metadata must be copied from server to server involves a variant of the pattern above. In cases in which a single filesystem moves between or among a small set of servers, it will transition to a server on which a previous instantiation of that same filesystem existed before. In such cases, it is often more efficient to update the previous filesystem instance to reflect changes made while the active

filesystem was residing elsewhere, rather than copying the filesystem data anew.

In such cases, the copying of filesystem data and metadata is replaced by a process which validates each visible filesystem object, copying new objects and updating those that have changed since the filesystem was last present on the destination server. Although this process is generally shorter than a complete copy, it is generally long enough that it is not practical to hold off access to the filesystem while this update is going on.

Eventually the filesystem updating process will complete. At this point, there will be two valid copies of the filesystem, one on each of the source and destination servers. Servers may maintain that state of affairs just as is done in the previous case. Similarly, the transfer of locking state, once it is complete, allows the clients to be notified of the migration event and access the destination filesystem on the destination server.

## 6.2. Preventing Locking State Modification During Transfer

When transferring locking state from the source to a destination server, there will be occasions when the source server will need to prevent operations that modify the state being transferred. For example, if the locking state at time T is sent to the destination server, any state change that occurs on the source server after that time but before the filesystem transfer is made effective will mean that the state on the destination server will differ from that on the source server, which matches what the client would expect to see.

In general, a server can prevent some set of server-maintained data from changing by returning NFS4ERR\_DELAY on operations which attempt to change that data. In the case of locking state for NFSv4.0, there are two specific issues that might interfere:

- o Returning NFS4ERR\_DELAY will not prevent state from changing in that owner-based sequence values will still change, even though NFS4ERR\_DELAY is returned. For example OPEN and LOCK will change state (in the form of owner seqid values) even when they return NFS4ERR\_DELAY.
- o Some operations which modify locking state are not allowed to return NFS4ERR\_DELAY.

Note that the first problem and many instances of the second can be addressed by returning NFS4ERR\_DELAY on the operations that establish a filehandle within the target as one of the filehandles associated with the request, i.e. as either the current or saved filehandle.

This would require returning NFS4ERR\_DELAY under the following circumstances:

- o On a PUTFH that specifies a filehandle within the target filesystem.
- o On a LOOKUP or LOOKUPP that crosses into the target filesystem.

Note that if the server establishes and maintains a situation in which no request has, as either the current or saved filehandle, a filehandle within the target filesystem, no special handling of SAVEFH or RESTOREFH is required. Thus the fact that these operations cannot return NFS4ERR\_DELAY is not a problem since neither will establish a filehandle in the target filesystem as the current filehandle.

If the server is to establish the situation described above, it may have to take special note of long-running requests which started before state migration. Part of any solution to this issue will involve distinguishing two separate points in time at which handling for the target filesystem will change. Let us distinguish;

- o A time T after which the previously mentioned operations will return NFS4ERR\_DELAY.
- o A later time T' at which the server can consider filesystem locking state fixed, making it possible for it to be sent to the destination server.

For a server to decide on T', it must ensure that requests started before T, cannot change target filesystem locking state, given that all those started after T are dealt with by returning NFS4ERR\_DELAY upon setting filehandles within the target filesystem. Among the ways of doing this are:

- o Keeping track of the earliest request started which is still in execution (for example, by keeping a list of active requests ordered by request start time). The server can then define T' to be the first time at which the earliest-started active request started after time T.
- o Keeping track of the count of requests, started before time T which have a filehandle within the target filesystem as either the current or saved filehandle. The server can then define T' to be the first time after T at which the count is zero.

The set of operations that change locking state include two that cannot be dealt with by the above approach, because they are not

filesystem-specific and do not use a current filehandle as an implicit parameter.

- o RENEW can be dealt with by applying the renewal to state for non-transitioning filesystems. The effect of renewal for the transitioning filesystem can be ignored, as long as the servers make sure that the lease on the destination server has an expiration time that is no earlier than the latest renewal done on the source server. This can be easily accomplished by making the lease expiration on the destination server equal to the time the state transfer was completed plus the lease period.
- o RELEASE\_LOCKOWNER can be handled by propagating the fact of the lockowner deletion (e.g. by using an RPC) to the destination server. Such a propagation RPC can be done as part of the operation or the existence of the deletion can be recorded locally and propagation of owner deletions to the destination server done as a batch later. In either case, the actual deletions on the destination server have to be delayed until all of the other state information has been transferred.

Alternatively, RELEASE\_LOCKOWNER can be dealt with by returning NFS4ERR\_DELAY. In order to avoid compatibility issues for clients not prepared to accept NFS4ERR\_DELAY in response to RELEASE\_LOCKOWNER, care must be exercised. (See Section 7.3 for details.)

The approach outlined above, wherein NFS\$ERR\_DELAY is returned based primarily on the use of current and saved filehandles in the filesystem, prevents all reference to the transitioning filesystem, rather than limiting the delayed operations to those that change locking state on the transitioning filesystem. Because of this, servers may choose to limit the time during which this broad approach is used by adopting a layered approach to the issue.

- o During the preparatory phase, operations that change, create, or destroy locks or modify the valid set of stateids will return NFS4ERR\_DELAY. During this phase, owner-associated seqids may change, and the identity of the filesystem associated with the last request for a given owner may change as well. Also, RELEASE\_LOCKOWNER operations may be processed without returning NFS4ERR\_DELAY as long as the fact of the lockowner deletion is recorded locally for later transmission.
- o During the restrictive phase, operations that change locking state for the filesystem in transition are prevented by returning NFS4ERR\_DELAY on any attempt to make a filehandle within that filesystem either the current or saved filehandle for a request.

RELEASE\_LOCKOWNER operations may return NFS4ERR\_DELAY, but if they are processed, the lockowner deletion needs to be communicated immediately to the destination server.

A possible sequence would be the following.

- o The server enters the preparatory phase for the transitioning filesystem.
- o At this point locking state, including stateids, locks, owner strings are transferred to the destination server. The seqids associated with owners are either not transferred, or transferred on a provisional basis, subject to later change.
- o After the above has been transferred, the server may enter the restrictive phase for the filesystem.
- o At this point, the updated seqid values may be sent to the destination server.

Reporting regarding pending owner deletions (as a result of RELEASE\_LOCKOWNER operations) can be communicated at the same time.

- o Once it is known that all of this information has been transferred to the destination server, and there are no pending RELEASE\_LOCKOWNER notifications outstanding, the source server may treat the filesystem transition as having occurred and return NFS4ERR\_MOVED when an attempt is made to access it.

## 7. Additional Changes

This chapter contains a number of items which relate to the changes in the chapters above, but which, for one reason or another, exist in different portions of the specification to be updated.

### 7.1. Summary of Additional Changes from Previous Documents

We summarize here all the remaining changes, not included in the two main chapters.

- o New definition of the CLID\_INUSE error.
- o A revised description of SETCLIENTID, which brings the description into sync with the rest of the spec regarding CLID\_INUSE.
- o A revision to the Security Considerations section, indicating why integrity protection is needed for the SETCLIENTID operation.



- o A revision of the error definitions chapter to allow `RELEASE_LOCKOWNER` to return `NFS4ERR_DELAY`, with appropriate constraints to assure interoperability with clients not expecting this error to be returned.

## 7.2. `NFS4ERR_CLID_INUSE` definition

The definition of this error is now as follows

The `SETCLIENTID` operation has found that the id string within the specified `nfs_client_id4` was previously presented with a different principal and that client instance currently holds an active lease. A server MAY return this error if the same principal is used but a change in authentication flavor gives good reason to reject the new `SETCLIENTID` operation as not bona fide.

## 7.3. `NFS4ERR_DELAY` return from `RELEASE_LOCKOWNER`

The existing error tables should be considered modified to allow `NFS4ERR_DELAY` to be returned by `RELEASE_LOCKOWNER`. However, the scope of this addition is limited and is not to be considered as making this error return generally acceptable.

It needs to be made clear that servers may not return this error to clients not prepared to support filesystem migration. Such clients may be following the error specifications in [RFC3530] and [cur-rfc3530-bis] and so might not expect `NFS4ERR_DELAY` to be returned on `RELEASE_LOCKOWNER`.

The following constraint applies to this additional error return, as if it were a note appearing together with the newly allowed error code:

In order to make server state fixed for a filesystem being migrated, a server MAY return `NFS4ERR_DELAY` in response to a `RELEASE_LOCKOWNER` that will affect locking state being propagated to a destination server. The source server MUST NOT do so unless it is likely that it will later return `NFS4ERR_MOVED` for the filesystem in question.

In the context of lockowner release, the set of filesystems such that server state being made fixed can result in `NFS4ERR_DELAY` must include the filesystem on which the operation associated with the current lockowner seqid was performed.

In addition, this set may include other filesystems on which an operation associated with an earlier seqid for the current lockowner seqid was performed, since servers will have to deal

with the issue of an owner being used in succession for multiple filesystems.

Thus, a client that is prepared to receive NFS4ERR\_MOVED after creating state associated with a given filesystem, it also needs to be prepared to receive NFS4ERR\_DELAY in response to RELEASE\_LOCKOWNER, if it has used that owner in connection with a file on that filesystem.

#### 7.4. Operation 35: SETCLIENTID - Negotiate Client ID

##### 7.4.1. SYNOPSIS

```
client, callback, callback_ident -> clientid, setclientid_confirm
```

##### 7.4.2. ARGUMENT

```
struct SETCLIENTID4args {
    nfs_client_id4  client;
    cb_client4      callback;
    uint32_t        callback_ident;
};
```

##### 7.4.3. RESULT

```
struct SETCLIENTID4resok {
    clientid4      clientid;
    verifier4      setclientid_confirm;
};

union SETCLIENTID4res switch (nfsstat4 status) {
    case NFS4_OK:
        SETCLIENTID4resok      resok4;
    case NFS4ERR_CLID_INUSE:
        clientaddr4      client_using;
    default:
        void;
};
```

##### 7.4.4. DESCRIPTION

The client uses the SETCLIENTID operation to notify the server of its intention to use a particular client identifier, callback, and callback\_ident for subsequent requests that entail creating lock, share reservation, and delegation state on the server. Upon

successful completion the server will return a shorthand client ID which, if confirmed via a separate step, will be used in subsequent file locking and file open requests. Confirmation of the client ID must be done via the SETCLIENTID\_CONFIRM operation to return the client ID and setclientid\_confirm values, as verifiers, to the server. The reason why two verifiers are necessary is that it is possible to use SETCLIENTID and SETCLIENTID\_CONFIRM to modify the callback and callback\_ident information but not the shorthand client ID. In that event, the setclientid\_confirm value is effectively the only verifier.

The callback information provided in this operation will be used if the client is provided an open delegation at a future point. Therefore, the client must correctly reflect the program and port numbers for the callback program at the time SETCLIENTID is used.

The callback\_ident value is used by the server on the callback. The client can leverage the callback\_ident to eliminate the need for more than one callback RPC program number, while still being able to determine which server is initiating the callback.

#### 7.4.5. IMPLEMENTATION

To understand how to implement SETCLIENTID, make the following notations. Let:

- x be the value of the client.id subfield of the SETCLIENTID4args structure.
  - v be the value of the client.verifier subfield of the SETCLIENTID4args structure.
  - c be the value of the client ID field returned in the SETCLIENTID4resok structure.
  - k represent the value combination of the fields callback and callback\_ident fields of the SETCLIENTID4args structure.
  - s be the setclientid\_confirm value returned in the SETCLIENTID4resok structure.
- { v, x, c, k, s } be a quintuple for a client record. A client record is confirmed if there has been a SETCLIENTID\_CONFIRM operation to confirm it. Otherwise it is unconfirmed. An unconfirmed record is established by a SETCLIENTID call.

##### 7.4.5.1. IMPLEMENTATION (preparatory phase)

Since SETCLIENTID is a non-idempotent operation, let us assume that the server is implementing the duplicate request cache (DRC).

When the server gets a SETCLIENTID { v, x, k } request, it first does a number of preliminary checks as listed below before proceeding to the main part of SETCLIENTID processing.

- o It first looks up the request in the DRC. If there is a hit, it returns the result cached in the DRC. The server does NOT remove client state (locks, shares, delegations) nor does it modify any recorded callback and callback\_ident information for client { x }.
- o Otherwise (i.e. in the case of any DRC miss), the server takes the client id string x, and searches for confirmed client records for x that the server may have recorded from previous SETCLIENTID calls. If there are no such, or if all such records have a recorded principal which matches that of the current request's principal, then
- o If there is a confirmed client record with a matching client id string and a non-matching principal, the server checks the current state of the associated lease. If there is no associated state for the lease, or the lease has expired, the server proceeds to the main part of SETCLIENTID
- o Otherwise, the server is being asked to do a SETCLIENTID for a client by a non-matching principal while there is active state and the server rejects the SETCLIENTID request returning an NFS4ERR\_CLID\_INUSE error, since use of a single client with multiple principals is not allowed. Note that even though the previously used clientaddr is returned with this error, the use of the same id string with multiple clientaddr's is not prohibited, while its use with multiple principals is prohibited.

#### 7.4.5.2. IMPLEMENTATION (main phase)

If the SETCLIENTID has not been dealt with by DRC processing, and has not been rejected with an NFS4ERR\_CLID\_INUSE error, then the main part of SETCLIENTID processing proceeds, as described below.

- o The server checks if it has recorded a confirmed record for { v, x, c, l, s }, where l may or may not equal k. If so, and since the id verifier v of the request matches that which is confirmed and recorded, the server treats this as a probable callback information update and records an unconfirmed { v, x, c, k, t } and leaves the confirmed { v, x, c, l, s } in place, such that t != s. It does not matter if k equals l or not. Any pre-existing unconfirmed { v, x, c, \*, \* } is removed.

The server returns { c, t }. It is indeed returning the old clientid4 value c, because the client apparently only wants to update callback value k to value l. It's possible this request is one from the Byzantine router that has stale callback information, but this is not a problem. The callback information update is only confirmed if followed up by a SETCLIENTID\_CONFIRM { c, t }.

The server awaits confirmation of k via SETCLIENTID\_CONFIRM { c, t }.

The server does NOT remove client (lock/share/delegation) state for x.

- o The server has previously recorded a confirmed { u, x, c, l, s } record such that v != u, l may or may not equal k, and has not recorded any unconfirmed { \*, x, \*, \*, \* } record for x. The server records an unconfirmed { v, x, d, k, t } (d != c, t != s).

The server returns { d, t }.

The server awaits confirmation of { d, k } via SETCLIENTID\_CONFIRM { d, t }.

The server does NOT remove client (lock/share/delegation) state for x.

- o The server has previously recorded a confirmed { u, x, c, l, s } record such that v != u, l may or may not equal k, and recorded an unconfirmed { w, x, d, m, t } record such that c != d, t != s, m may or may not equal k, m may or may not equal l, and k may or may not equal l. Whether w == v or w != v makes no difference. The server simply removes the unconfirmed { w, x, d, m, t } record and replaces it with an unconfirmed { v, x, e, k, r } record, such that e != d, e != c, r != t, r != s.

The server returns { e, r }.

The server awaits confirmation of { e, k } via SETCLIENTID\_CONFIRM { e, r }.

The server does NOT remove client (lock/share/delegation) state for x.

- o The server has no confirmed { \*, x, \*, \*, \* } for x. It may or may not have recorded an unconfirmed { u, x, c, l, s }, where l may or may not equal k, and u may or may not equal v. Any unconfirmed record { u, x, c, l, \* }, regardless whether u == v or l == k, is replaced with an unconfirmed record { v, x, d, k, t } where d != c, t != s.

The server returns { d, t }.

The server awaits confirmation of { d, k } via SETCLIENTID\_CONFIRM { d, t }. The server does NOT remove client (lock/share/delegation) state for x.

The server generates the clientid and setclientid\_confirm values and must take care to ensure that these values are extremely unlikely to ever be regenerated.

#### 7.5. Security Considerations revision

The last paragraph of the "Security Considerations" section should be revised to read as follows:

Because the operations SETCLIENTID/SETCLIENTID\_CONFIRM are responsible for the release of client state, it is imperative that the principal used for these operations is checked against and match the previous use of these operations. In addition, use of integrity protection is desirable on the SETCLIENTID operation, to prevent an attack whereby a change in the boot verifier forces an undesired loss of client state. See the section "Client Identity Definition" for further discussion.

#### 8. Security Considerations

Is modified as specified in Section 7.5.

#### 9. IANA Considerations

This document does not require actions by IANA.

#### 10. Acknowledgements

The editor and authors of this document gratefully acknowledge the contributions of Trond Myklebust of NetApp and Robert Thurlow of Oracle. We also thank Tom Haynes of NetApp and Spencer Shepler of Microsoft for their guidance and suggestions.

Special thanks go to members of the Oracle Solaris NFS team, especially Rick Mesta and James Wahlig, for their work implementing

an NFSv4.0 migration prototype and identifying many of the issues addressed here.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3530] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.

### 11.2. Informative References

- [RFC1813] Callaghan, B., Pawlowski, B., and P. Staubach, "NFS Version 3 Protocol Specification", RFC 1813, June 1995.
- [RFC5661] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.
- [cur-rfc3530-bis] Haynes, T., Ed. and D. Noveck, Ed., "Network File System (NFS) Version 4 Protocol", 2013, <<http://www.ietf.org/id/draft-ietf-nfsv4-rfc3530bis-26.txt>>. Work in progress.
- [info-migr] Noveck, D., Ed., Shivam, P., Lever, C., and B. Baker, "NFSv4 migration: Implementation experience and spec issues to resolve ", 2012, <<http://www.ietf.org/id/draft-ietf-nfsv4-migration-issues-03.txt>>. Work in progress.

## Authors' Addresses

David Noveck (editor)  
EMC Corporation  
228 South Street  
Hopkinton, MA 01748  
US

Phone: +1 508 249 5748  
Email: david.noveck@emc.com

Piyush Shivam  
Oracle Corporation  
5300 Riata Park Ct.  
Austin, TX 78727  
US

Phone: +1 512 401 1019  
Email: piyush.shivam@oracle.com

Charles Lever  
Oracle Corporation  
1015 Granger Avenue  
Ann Arbor, MI 48104  
US

Phone: +1 734 274 2396  
Email: chuck.lever@oracle.com

Bill Baker  
Oracle Corporation  
5300 Riata Park Ct.  
Austin, TX 78727  
US

Phone: +1 512 401 1081  
Email: bill.baker@oracle.com



NFSv4  
Internet-Draft  
Obsoletes: 3530 (if approved)  
Intended status: Standards Track  
Expires: April 22, 2014

T. Haynes, Ed.  
NetApp  
D. Noveck, Ed.  
EMC  
October 19, 2013

Network File System (NFS) Version 4 Protocol  
draft-ietf-nfsv4-rfc3530bis-28.txt

## Abstract

The Network File System (NFS) version 4 is a distributed file system protocol which builds on the heritage of NFS protocol version 2, RFC 1094, and version 3, RFC 1813. Unlike earlier versions, the NFS version 4 protocol supports traditional file access while integrating support for file locking and the mount protocol. In addition, support for strong security (and its negotiation), compound operations, client caching, and internationalization have been added. Of course, attention has been applied to making NFS version 4 operate well in an Internet environment.

This document, together with the companion XDR description document, RFCNFSv4XDR, obsoletes RFC 3530 as the definition of the NFS version 4 protocol.

## Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 22, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

## Table of Contents

1.	Introduction . . . . .	9
1.1.	NFS Version 4 Goals . . . . .	9
1.2.	Definitions in the companion document NFS Version 4 Protocol are Authoritative . . . . .	9
1.3.	Overview of NFSv4 Features . . . . .	10
1.3.1.	RPC and Security . . . . .	10
1.3.2.	Procedure and Operation Structure . . . . .	10
1.3.3.	Filesystem Model . . . . .	11
1.3.4.	OPEN and CLOSE . . . . .	13
1.3.5.	File Locking . . . . .	13
1.3.6.	Client Caching and Delegation . . . . .	13
1.4.	General Definitions . . . . .	14
1.5.	Changes since RFC 3530 . . . . .	16
1.6.	Changes since RFC 3010 . . . . .	17
2.	Protocol Data Types . . . . .	18
2.1.	Basic Data Types . . . . .	18
2.2.	Structured Data Types . . . . .	20
3.	RPC and Security Flavor . . . . .	24
3.1.	Ports and Transports . . . . .	24
3.1.1.	Client Retransmission Behavior . . . . .	25
3.2.	Security Flavors . . . . .	26
3.2.1.	Security mechanisms for NFSv4 . . . . .	26
3.3.	Security Negotiation . . . . .	27
3.3.1.	SECINFO . . . . .	28
3.3.2.	Security Error . . . . .	28
3.3.3.	Callback RPC Authentication . . . . .	28
4.	Filehandles . . . . .	29
4.1.	Obtaining the First Filehandle . . . . .	29
4.1.1.	Root Filehandle . . . . .	30
4.1.2.	Public Filehandle . . . . .	30
4.2.	Filehandle Types . . . . .	31
4.2.1.	General Properties of a Filehandle . . . . .	31
4.2.2.	Persistent Filehandle . . . . .	32
4.2.3.	Volatile Filehandle . . . . .	32
4.2.4.	One Method of Constructing a Volatile Filehandle . . . . .	33
4.3.	Client Recovery from Filehandle Expiration . . . . .	34
5.	Attributes . . . . .	35
5.1.	REQUIRED Attributes . . . . .	36
5.2.	RECOMMENDED Attributes . . . . .	36
5.3.	Named Attributes . . . . .	37
5.4.	Classification of Attributes . . . . .	38
5.5.	Set-Only and Get-Only Attributes . . . . .	39
5.6.	REQUIRED Attributes - List and Definition References . . . . .	39
5.7.	RECOMMENDED Attributes - List and Definition References . . . . .	40
5.8.	Attribute Definitions . . . . .	41

5.8.1.	Definitions of REQUIRED Attributes . . . . .	41
5.8.2.	Definitions of Uncategorized RECOMMENDED Attributes . . . . .	43
5.9.	Interpreting owner and owner_group . . . . .	49
5.10.	Character Case Attributes . . . . .	52
6.	Access Control Attributes . . . . .	52
6.1.	Goals . . . . .	52
6.2.	File Attributes Discussion . . . . .	53
6.2.1.	Attribute 12: acl . . . . .	53
6.2.2.	Attribute 33: mode . . . . .	68
6.3.	Common Methods . . . . .	68
6.3.1.	Interpreting an ACL . . . . .	68
6.3.2.	Computing a Mode Attribute from an ACL . . . . .	69
6.4.	Requirements . . . . .	70
6.4.1.	Setting the mode and/or ACL Attributes . . . . .	71
6.4.2.	Retrieving the mode and/or ACL Attributes . . . . .	72
6.4.3.	Creating New Objects . . . . .	72
7.	NFS Server Name Space . . . . .	74
7.1.	Server Exports . . . . .	74
7.2.	Browsing Exports . . . . .	74
7.3.	Server Pseudo Filesystem . . . . .	75
7.4.	Multiple Roots . . . . .	75
7.5.	Filehandle Volatility . . . . .	76
7.6.	Exported Root . . . . .	76
7.7.	Mount Point Crossing . . . . .	76
7.8.	Security Policy and Name Space Presentation . . . . .	77
8.	Multi-Server Namespace . . . . .	77
8.1.	Location Attributes . . . . .	78
8.2.	File System Presence or Absence . . . . .	78
8.3.	Getting Attributes for an Absent File System . . . . .	79
8.3.1.	GETATTR Within an Absent File System . . . . .	79
8.3.2.	READDIR and Absent File Systems . . . . .	80
8.4.	Uses of Location Information . . . . .	81
8.4.1.	File System Replication . . . . .	82
8.4.2.	File System Migration . . . . .	82
8.4.3.	Referrals . . . . .	83
8.5.	Location Entries and Server Identity . . . . .	84
8.6.	Additional Client-Side Considerations . . . . .	84
8.7.	Effecting File System Referrals . . . . .	85
8.7.1.	Referral Example (LOOKUP) . . . . .	86
8.7.2.	Referral Example (READDIR) . . . . .	90
8.8.	The Attribute fs_locations . . . . .	92
8.8.1.	Inferring Transition Modes . . . . .	94
9.	File Locking and Share Reservations . . . . .	95
9.1.	Opens and Byte-Range Locks . . . . .	96
9.1.1.	Client ID . . . . .	97
9.1.2.	Server Release of Client ID . . . . .	100
9.1.3.	Stateid Definition . . . . .	100

9.1.4.	lock-owner . . . . .	106
9.1.5.	Use of the Stateid and Locking . . . . .	107
9.1.6.	Sequencing of Lock Requests . . . . .	109
9.1.7.	Recovery from Replayed Requests . . . . .	110
9.1.8.	Interactions of multiple sequence values . . . . .	110
9.1.9.	Releasing state-owner State . . . . .	111
9.1.10.	Use of Open Confirmation . . . . .	112
9.2.	Lock Ranges . . . . .	113
9.3.	Upgrading and Downgrading Locks . . . . .	114
9.4.	Blocking Locks . . . . .	114
9.5.	Lease Renewal . . . . .	115
9.6.	Crash Recovery . . . . .	116
9.6.1.	Client Failure and Recovery . . . . .	116
9.6.2.	Server Failure and Recovery . . . . .	117
9.6.3.	Network Partitions and Recovery . . . . .	118
9.7.	Recovery from a Lock Request Timeout or Abort . . . . .	126
9.8.	Server Revocation of Locks . . . . .	126
9.9.	Share Reservations . . . . .	128
9.10.	OPEN/CLOSE Operations . . . . .	128
9.10.1.	Close and Retention of State Information . . . . .	129
9.11.	Open Upgrade and Downgrade . . . . .	130
9.12.	Short and Long Leases . . . . .	131
9.13.	Clocks, Propagation Delay, and Calculating Lease Expiration . . . . .	131
9.14.	Migration, Replication and State . . . . .	132
9.14.1.	Migration and State . . . . .	132
9.14.2.	Replication and State . . . . .	133
9.14.3.	Notification of Migrated Lease . . . . .	133
9.14.4.	Migration and the Lease_time Attribute . . . . .	134
10.	Client-Side Caching . . . . .	135
10.1.	Performance Challenges for Client-Side Caching . . . . .	135
10.2.	Delegation and Callbacks . . . . .	136
10.2.1.	Delegation Recovery . . . . .	138
10.3.	Data Caching . . . . .	142
10.3.1.	Data Caching and OPENS . . . . .	143
10.3.2.	Data Caching and File Locking . . . . .	144
10.3.3.	Data Caching and Mandatory File Locking . . . . .	145
10.3.4.	Data Caching and File Identity . . . . .	146
10.4.	Open Delegation . . . . .	147
10.4.1.	Open Delegation and Data Caching . . . . .	149
10.4.2.	Open Delegation and File Locks . . . . .	150
10.4.3.	Handling of CB_GETATTR . . . . .	151
10.4.4.	Recall of Open Delegation . . . . .	154
10.4.5.	OPEN Delegation Race with CB_RECALL . . . . .	156
10.4.6.	Clients that Fail to Honor Delegation Recalls . . . . .	156
10.4.7.	Delegation Revocation . . . . .	157
10.5.	Data Caching and Revocation . . . . .	158
10.5.1.	Revocation Recovery for Write Open Delegation . . . . .	158

10.6.	Attribute Caching . . . . .	159
10.7.	Data and Metadata Caching and Memory Mapped Files . . . . .	161
10.8.	Name Caching . . . . .	163
10.9.	Directory Caching . . . . .	164
11.	Minor Versioning . . . . .	165
12.	Internationalization . . . . .	168
12.1.	Introduction . . . . .	168
12.2.	String Encoding . . . . .	169
12.3.	Normalization . . . . .	169
12.4.	Types with Processing Defined by Other Internet Areas . . . . .	170
12.5.	UTF-8 Related Errors . . . . .	171
12.6.	Handling of component names that are not valid UTF-8 strings . . . . .	172
13.	Error Values . . . . .	173
13.1.	Error Definitions . . . . .	173
13.1.1.	General Errors . . . . .	175
13.1.2.	Filehandle Errors . . . . .	176
13.1.3.	Compound Structure Errors . . . . .	177
13.1.4.	File System Errors . . . . .	178
13.1.5.	State Management Errors . . . . .	180
13.1.6.	Security Errors . . . . .	181
13.1.7.	Name Errors . . . . .	182
13.1.8.	Locking Errors . . . . .	182
13.1.9.	Reclaim Errors . . . . .	184
13.1.10.	Client Management Errors . . . . .	184
13.1.11.	Attribute Handling Errors . . . . .	185
13.2.	Operations and their valid errors . . . . .	185
13.3.	Callback operations and their valid errors . . . . .	192
13.4.	Errors and the operations that use them . . . . .	193
14.	NFSv4 Requests . . . . .	197
14.1.	Compound Procedure . . . . .	198
14.2.	Evaluation of a Compound Request . . . . .	198
14.3.	Synchronous Modifying Operations . . . . .	199
14.4.	Operation Values . . . . .	199
15.	NFSv4 Procedures . . . . .	200
15.1.	Procedure 0: NULL - No Operation . . . . .	200
15.2.	Procedure 1: COMPOUND - Compound Operations . . . . .	200
15.3.	Operation 3: ACCESS - Check Access Rights . . . . .	204
15.4.	Operation 4: CLOSE - Close File . . . . .	207
15.5.	Operation 5: COMMIT - Commit Cached Data . . . . .	208
15.6.	Operation 6: CREATE - Create a Non-Regular File Object . . . . .	210
15.7.	Operation 7: DELEGPURGE - Purge Delegations Awaiting Recovery . . . . .	213
15.8.	Operation 8: DELEGRETURN - Return Delegation . . . . .	214
15.9.	Operation 9: GETATTR - Get Attributes . . . . .	215
15.10.	Operation 10: GETFH - Get Current Filehandle . . . . .	217
15.11.	Operation 11: LINK - Create Link to a File . . . . .	218
15.12.	Operation 12: LOCK - Create Lock . . . . .	219

15.13.	Operation 13: LOCKT - Test For Lock . . . . .	223
15.14.	Operation 14: LOCKU - Unlock File . . . . .	225
15.15.	Operation 15: LOOKUP - Lookup Filename . . . . .	226
15.16.	Operation 16: LOOKUPP - Lookup Parent Directory . . . . .	228
15.17.	Operation 17: NVERIFY - Verify Difference in Attributes . . . . .	229
15.18.	Operation 18: OPEN - Open a Regular File . . . . .	230
15.19.	Operation 19: OPENATTR - Open Named Attribute Directory . . . . .	240
15.20.	Operation 20: OPEN_CONFIRM - Confirm Open . . . . .	241
15.21.	Operation 21: OPEN_DOWNGRADE - Reduce Open File Access .	243
15.22.	Operation 22: PUTFH - Set Current Filehandle . . . . .	244
15.23.	Operation 23: PUTPUBFH - Set Public Filehandle . . . . .	245
15.24.	Operation 24: PUTROOTFH - Set Root Filehandle . . . . .	246
15.25.	Operation 25: READ - Read from File . . . . .	247
15.26.	Operation 26: READDIR - Read Directory . . . . .	249
15.27.	Operation 27: READLINK - Read Symbolic Link . . . . .	253
15.28.	Operation 28: REMOVE - Remove Filesystem Object . . . . .	254
15.29.	Operation 29: RENAME - Rename Directory Entry . . . . .	256
15.30.	Operation 30: RENEW - Renew a Lease . . . . .	258
15.31.	Operation 31: RESTOREFH - Restore Saved Filehandle . . . .	259
15.32.	Operation 32: SAVEFH - Save Current Filehandle . . . . .	260
15.33.	Operation 33: SECINFO - Obtain Available Security . . . . .	261
15.34.	Operation 34: SETATTR - Set Attributes . . . . .	265
15.35.	Operation 35: SETCLIENTID - Negotiate Client ID . . . . .	267
15.36.	Operation 36: SETCLIENTID_CONFIRM - Confirm Client ID .	271
15.37.	Operation 37: VERIFY - Verify Same Attributes . . . . .	274
15.38.	Operation 38: WRITE - Write to File . . . . .	276
15.39.	Operation 39: RELEASE_LOCKOWNER - Release Lockowner State . . . . .	280
15.40.	Operation 10044: ILLEGAL - Illegal operation . . . . .	281
16.	NFSv4 Callback Procedures . . . . .	282
16.1.	Procedure 0: CB_NULL - No Operation . . . . .	282
16.2.	Procedure 1: CB_COMPOUND - Compound Operations . . . . .	282
16.2.6.	Operation 3: CB_GETATTR - Get Attributes . . . . .	284
16.2.7.	Operation 4: CB_RECALL - Recall an Open Delegation .	285
16.2.8.	Operation 10044: CB_ILLEGAL - Illegal Callback Operation . . . . .	286
17.	Security Considerations . . . . .	287
18.	IANA Considerations . . . . .	289
18.1.	Named Attribute Definitions . . . . .	289
18.1.1.	Initial Registry . . . . .	290
18.1.2.	Updating Registrations . . . . .	290
19.	References . . . . .	290
19.1.	Normative References . . . . .	290
19.2.	Informative References . . . . .	291
Appendix A.	Acknowledgments . . . . .	294
Appendix B.	RFC Editor Notes . . . . .	295

Authors' Addresses . . . . .	295
------------------------------	-----



## 1. Introduction

### 1.1. NFS Version 4 Goals

The Network Filesystem version 4 (NFSv4) protocol is a further revision of the NFS protocol defined already by versions 2 [RFC1094] and 3 [RFC1813]. It retains the essential characteristics of previous versions: design for easy recovery, independent of transport protocols, operating systems and file systems, simplicity, and good performance. The NFSv4 revision has the following goals:

- o Improved access and good performance on the Internet.

The protocol is designed to transit firewalls easily, perform well where latency is high and bandwidth is low, and scale to very large numbers of clients per server.

- o Strong security with negotiation built into the protocol.

The protocol builds on the work of the Open Network Computing (ONC) Remote Procedure Call (RPC) working group in supporting the RPCSEC\_GSS protocol (see both [RFC2203] and [RFC5403]). Additionally, the NFS version 4 protocol provides a mechanism to allow clients and servers the ability to negotiate security and require clients and servers to support a minimal set of security schemes.

- o Good cross-platform interoperability.

The protocol features a file system model that provides a useful, common set of features that does not unduly favor one file system or operating system over another.

- o Designed for protocol extensions.

The protocol is designed to accept standard extensions that do not compromise backward compatibility.

This document, together with the companion XDR description document [I-D.ietf-nfsv4-rfc3530bis-dot-x], obsoletes RFC 3530 [RFC3530] as the authoritative document describing NFSv4. It does not introduce any over-the-wire protocol changes, in the sense that previously valid requests remain valid.

### 1.2. Definitions in the companion document NFS Version 4 Protocol are Authoritative

[I-D.ietf-nfsv4-rfc3530bis-dot-x], NFS Version 4 Protocol, contains

the definitions in XDR description language of the constructs used by the protocol. Inside this document, several of the constructs are reproduced for purposes of explanation. The reader is warned of the possibility of errors in the reproduced constructs outside of [I-D.ietf-nfsv4-rfc3530bis-dot-x]. For any part of the document that is inconsistent with [I-D.ietf-nfsv4-rfc3530bis-dot-x], [I-D.ietf-nfsv4-rfc3530bis-dot-x] is to be considered authoritative.

### 1.3. Overview of NFSv4 Features

To provide a reasonable context for the reader, the major features of NFSv4 protocol will be reviewed in brief. This will be done to provide an appropriate context for both the reader who is familiar with the previous versions of the NFS protocol and the reader who is new to the NFS protocols. For the reader new to the NFS protocols, some fundamental knowledge is still expected. The reader should be familiar with the XDR and RPC protocols as described in [RFC5531] and [RFC4506]. A basic knowledge of file systems and distributed file systems is expected as well.

#### 1.3.1. RPC and Security

As with previous versions of NFS, the External Data Representation (XDR) and RPC mechanisms used for the NFSv4 protocol are those defined in [RFC5531] and [RFC4506]. To meet end to end security requirements, the RPCSEC\_GSS framework (both version 1 in [RFC2203] and version 2 in [RFC5403]) will be used to extend the basic RPC security. With the use of RPCSEC\_GSS, various mechanisms can be provided to offer authentication, integrity, and privacy to the NFS version 4 protocol. Kerberos V5 will be used as described in [RFC4121] to provide one security framework. With the use of RPCSEC\_GSS, other mechanisms may also be specified and used for NFS version 4 security.

To enable in-band security negotiation, the NFSv4 protocol has added a new operation which provides the client with a method of querying the server about its policies regarding which security mechanisms must be used for access to the server's file system resources. With this, the client can securely match the security mechanism that meets the policies specified at both the client and server.

#### 1.3.2. Procedure and Operation Structure

A significant departure from the previous versions of the NFS protocol is the introduction of the COMPOUND procedure. For the NFSv4 protocol, there are two RPC procedures, NULL and COMPOUND. The COMPOUND procedure is defined in terms of operations and these operations correspond more closely to the traditional NFS procedures.

With the use of the COMPOUND procedure, the client is able to build simple or complex requests. These COMPOUND requests allow for a reduction in the number of RPCs needed for logical file system operations. For example, without previous contact with a server a client will be able to read data from a file in one request by combining LOOKUP, OPEN, and READ operations in a single COMPOUND RPC. With previous versions of the NFS protocol, this type of single request was not possible.

The model used for COMPOUND is very simple. There is no logical OR or ANDing of operations. The operations combined within a COMPOUND request are evaluated in order by the server. Once an operation returns a failing result, the evaluation ends and the results of all evaluated operations are returned to the client.

The NFSv4 protocol continues to have the client refer to a file or directory at the server by a "filehandle". The COMPOUND procedure has a method of passing a filehandle from one operation to another within the sequence of operations. There is a concept of a "current filehandle" and "saved filehandle". Most operations use the "current filehandle" as the file system object to operate upon. The "saved filehandle" is used as temporary filehandle storage within a COMPOUND procedure as well as an additional operand for certain operations.

#### 1.3.3. Filesystem Model

The general file system model used for the NFSv4 protocol is the same as previous versions. The server file system is hierarchical with the regular files contained within being treated as opaque byte streams. In a slight departure, file and directory names are encoded with UTF-8 to deal with the basics of internationalization.

The NFSv4 protocol does not require a separate protocol to provide for the initial mapping between path name and filehandle. Instead of using the older MOUNT protocol for this mapping, the server provides a ROOT filehandle that represents the logical root or top of the file system tree provided by the server. The server provides multiple file systems by gluing them together with pseudo file systems. These pseudo file systems provide for potential gaps in the path names between real file systems.

##### 1.3.3.1. Filehandle Types

In previous versions of the NFS protocol, the filehandle provided by the server was guaranteed to be valid or persistent for the lifetime of the file system object to which it referred. For some server implementations, this persistence requirement has been difficult to meet. For the NFSv4 protocol, this requirement has been relaxed by

introducing another type of filehandle, volatile. With persistent and volatile filehandle types, the server implementation can match the abilities of the file system at the server along with the operating environment. The client will have knowledge of the type of filehandle being provided by the server and can be prepared to deal with the semantics of each.

#### 1.3.3.2. Attribute Types

The NFSv4 protocol has a rich and extensible file object attribute structure, which is divided into REQUIRED, RECOMMENDED, and named attributes (see Section 5).

Several (but not all) of the REQUIRED attributes are derived from the attributes of NFSv3 (see definition of the `fattnr3` data type in [RFC1813]). An example of a REQUIRED attribute is the file object's type (Section 5.8.1.2) so that regular files can be distinguished from directories (also known as folders in some operating environments) and other types of objects. REQUIRED attributes are discussed in Section 5.1.

An example of the RECOMMENDED attributes is an `acl` (Section 6.2.1). This attribute defines an Access Control List (ACL) on a file object. An ACL provides file access control beyond the model used in NFSv3. The ACL definition allows for specification of specific sets of permissions for individual users and groups. In addition, ACL inheritance allows propagation of access permissions and restriction down a directory tree as file system objects are created. RECOMMENDED attributes are discussed in Section 5.2.

A named attribute is an opaque byte stream that is associated with a directory or file and referred to by a string name. Named attributes are meant to be used by client applications as a method to associate application-specific data with a regular file or directory. NFSv4.1 modifies named attributes relative to NFSv4.0 by tightening the allowed operations in order to prevent the development of non-interoperable implementations. Named attributes are discussed in Section 5.3.

#### 1.3.3.3. Multi-server Namespace

NFSv4 contains a number of features to allow implementation of namespaces that cross server boundaries and that allow and facilitate a non-disruptive transfer of support for individual file systems between servers. They are all based upon attributes that allow one file system to specify alternate or new locations for that file system.

These attributes may be used together with the concept of absent file systems, which provide specifications for additional locations but no actual file system content. This allows a number of important facilities:

- o Location attributes may be used with absent file systems to implement referrals whereby one server may direct the client to a file system provided by another server. This allows extensive multi-server namespaces to be constructed.
- o Location attributes may be provided for present file systems to provide the locations of alternate file system instances or replicas to be used in the event that the current file system instance becomes unavailable.
- o Location attributes may be provided when a previously present file system becomes absent. This allows non-disruptive migration of file systems to alternate servers.

#### 1.3.4. OPEN and CLOSE

The NFSv4 protocol introduces OPEN and CLOSE operations. The OPEN operation provides a single point where file lookup, creation, and share semantics can be combined. The CLOSE operation also provides for the release of state accumulated by OPEN.

#### 1.3.5. File Locking

With the NFSv4 protocol, the support for byte range file locking is part of the NFS protocol. The file locking support is structured so that an RPC callback mechanism is not required. This is a departure from the previous versions of the NFS file locking protocol, Network Lock Manager (NLM). The state associated with file locks is maintained at the server under a lease-based model. The server defines a single lease period for all state held by a NFS client. If the client does not renew its lease within the defined period, all state associated with the client's lease may be released by the server. The client may renew its lease with use of the RENEW operation or implicitly by use of other operations (primarily READ).

#### 1.3.6. Client Caching and Delegation

The file, attribute, and directory caching for the NFSv4 protocol is similar to previous versions. Attributes and directory information are cached for a duration determined by the client. At the end of a predefined timeout, the client will query the server to see if the related file system object has been updated.

For file data, the client checks its cache validity when the file is opened. A query is sent to the server to determine if the file has been changed. Based on this information, the client determines if the data cache for the file should be kept or released. Also, when the file is closed, any modified data is written to the server.

If an application wants to serialize access to file data, file locking of the file data ranges in question should be used.

The major addition to NFSv4 in the area of caching is the ability of the server to delegate certain responsibilities to the client. When the server grants a delegation for a file to a client, the client is guaranteed certain semantics with respect to the sharing of that file with other clients. At OPEN, the server may provide the client either a `OPEN_DELEGATE_READ` or `OPEN_DELEGATE_WRITE` delegation for the file. If the client is granted a `OPEN_DELEGATE_READ` delegation, it is assured that no other client has the ability to write to the file for the duration of the delegation. If the client is granted a `OPEN_DELEGATE_WRITE` delegation, the client is assured that no other client has read or write access to the file.

Delegations can be recalled by the server. If another client requests access to the file in such a way that the access conflicts with the granted delegation, the server is able to notify the initial client and recall the delegation. This requires that a callback path exist between the server and client. If this callback path does not exist, then delegations cannot be granted. The essence of a delegation is that it allows the client to locally service operations such as OPEN, CLOSE, LOCK, LOCKU, READ, or WRITE without immediate interaction with the server.

#### 1.4. General Definitions

The following definitions are provided for the purpose of providing an appropriate context for the reader.

**Absent File System:** A file system is "absent" when a namespace component does not have a backing file system.

**Byte:** In this document, a byte is an octet, i.e., a datum exactly 8 bits in length.

**Client:** The client is the entity that accesses the NFS server's resources. The client may be an application that contains the logic to access the NFS server directly. The client may also be the traditional operating system client that provides remote file system services for a set of applications.

With reference to byte-range locking, the client is also the entity that maintains a set of locks on behalf of one or more applications. This client is responsible for crash or failure recovery for those locks it manages.

Note that multiple clients may share the same transport and connection and multiple clients may exist on the same network node.

**Client ID:** A 64-bit quantity used as a unique, short-hand reference to a client supplied Verifier and ID. The server is responsible for supplying the Client ID.

**File System:** The file system is the collection of objects on a server that share the same fsid attribute (see Section 5.8.1.9).

**Lease:** An interval of time defined by the server for which the client is irrevocably granted a lock. At the end of a lease period the lock may be revoked if the lease has not been extended. The lock must be revoked if a conflicting lock has been granted after the lease interval.

All leases granted by a server have the same fixed interval. Note that the fixed interval was chosen to alleviate the expense a server would have in maintaining state about variable length leases across server failures.

**Lock:** The term "lock" is used to refer to both record (byte-range) locks as well as share reservations unless specifically stated otherwise.

**Server:** The "Server" is the entity responsible for coordinating client access to a set of file systems.

**Stable Storage:** NFSv4 servers must be able to recover without data loss from multiple power failures (including cascading power failures, that is, several power failures in quick succession), operating system failures, and hardware failure of components other than the storage medium itself (for example, disk, nonvolatile RAM).

Some examples of stable storage that are allowable for an NFS server include:

- (1) Media commit of data, that is, the modified data has been successfully written to the disk media, for example, the disk platter.

- (2) An immediate reply disk drive with battery-backed on-drive intermediate storage or uninterruptible power system (UPS).
- (3) Server commit of data with battery-backed intermediate storage and recovery software.
- (4) Cache commit with uninterruptible power system (UPS) and recovery software.

**Stateid:** A stateid is a 128-bit quantity returned by a server that uniquely identifies the open and locking states provided by the server for a specific open-owner or lock-owner/open-owner pair for a specific file and type of lock.

**Verifier:** A 64-bit quantity generated by the client that the server can use to determine if the client has restarted and lost all previous lock state.

#### 1.5. Changes since RFC 3530

The main changes from RFC 3530 [RFC3530] are:

- o The XDR definition has been moved to a companion document [I-D.ietf-nfsv4-rfc3530bis-dot-x]
- o Updates for the latest IETF intellectual property statements
- o There is a restructured and more complete explanation of multi-server namespace features.
- o Updating handling of domain names to reflect Internationalized Domain Names in Applications (IDNA) [RFC5891].
- o The previously required LIPKEY and SPKM-3 security mechanisms have been removed.
- o Some clarification on a client re-establishing callback information to the new server if state has been migrated.
- o A third edge case was added for Courtesy locks and network partitions.
- o The definition of stateid was strengthened.



### 1.6. Changes since RFC 3010

This definition of the NFSv4 protocol replaces or obsoletes the definition present in [RFC3010]. While portions of the two documents have remained the same, there have been substantive changes in others. The changes made between [RFC3010] and this document represent implementation experience and further review of the protocol. While some modifications were made for ease of implementation or clarification, most updates represent errors or situations where the [RFC3010] definition were untenable.

The following list is not all inclusive of all changes but presents some of the most notable changes or additions made:

- o The state model has added an `open_owner4` identifier. This was done to accommodate Posix based clients and the model they use for file locking. For Posix clients, an `open_owner4` would correspond to a file descriptor potentially shared amongst a set of processes and the `lock_owner4` identifier would correspond to a process that is locking a file.
- o Clarifications and error conditions were added for the handling of the owner and group attributes. Since these attributes are string based (as opposed to the numeric uid/gid of previous versions of NFS), translations may not be available and hence the changes made.
- o Clarifications for the ACL and mode attributes to address evaluation and partial support.
- o For identifiers that are defined as XDR opaque, limits were set on their size.
- o Added the `mounted_on_fileid` attribute to allow Posix clients to correctly construct local mounts.
- o Modified the `SETCLIENTID/SETCLIENTID_CONFIRM` operations to deal correctly with confirmation details along with adding the ability to specify new client callback information. Also added clarification of the callback information itself.
- o Added a new operation `RELEASE_LOCKOWNER` to enable notifying the server that a `lock_owner4` will no longer be used by the client.
- o `RENEW` operation changes to identify the client correctly and allow for additional error returns.

- o Verify error return possibilities for all operations.
- o Remove use of the `pathname4` data type from LOOKUP and OPEN in favor of having the client construct a sequence of LOOKUP operations to achieve the same effect.

## 2. Protocol Data Types

The syntax and semantics to describe the data types of the NFS version 4 protocol are defined in the XDR [RFC4506] and RPC [RFC5531] documents. The next sections build upon the XDR data types to define types and structures specific to this protocol.

### 2.1. Basic Data Types

These are the base NFSv4 data types.

Data Type	Definition
<code>int32_t</code>	<code>typedef int int32_t;</code>
<code>uint32_t</code>	<code>typedef unsigned int uint32_t;</code>
<code>int64_t</code>	<code>typedef hyper int64_t;</code>
<code>uint64_t</code>	<code>typedef unsigned hyper uint64_t;</code>
<code>attrlist4</code>	<code>typedef opaque attrlist4&lt;&gt;;</code> Used for file/directory attributes.
<code>bitmap4</code>	<code>typedef uint32_t bitmap4&lt;&gt;;</code> Used in attribute array encoding.
<code>changeid4</code>	<code>typedef uint64_t changeid4;</code> Used in the definition of <code>change_info4</code> .
<code>clientid4</code>	<code>typedef uint64_t clientid4;</code> Shorthand reference to client identification.
<code>count4</code>	<code>typedef uint32_t count4;</code> Various count parameters (READ, WRITE, COMMIT).
<code>length4</code>	<code>typedef uint64_t length4;</code> Describes LOCK lengths.
<code>mode4</code>	<code>typedef uint32_t mode4;</code> Mode attribute data type.
<code>nfs_cookie4</code>	<code>typedef uint64_t nfs_cookie4;</code> Opaque cookie value for READDIR.
<code>nfs_fh4</code>	<code>typedef opaque nfs_fh4&lt;NFS4_FHSIZE&gt;;</code> Filehandle definition.
<code>nfs_ftype4</code>	<code>enum nfs_ftype4;</code> Various defined file types.
<code>nfsstat4</code>	<code>enum nfsstat4;</code> Return value for operations.
<code>offset4</code>	<code>typedef uint64_t offset4;</code>

	Various offset designations (READ, WRITE, LOCK, COMMIT).
qop4	typedef uint32_t qop4;
sec_oid4	Quality of protection designation in SECINFO. typedef opaque sec_oid4<>; Security Object Identifier. The sec_oid4 data type is not really opaque. Instead it contains an ASN.1 OBJECT IDENTIFIER as used by GSS-API in the mech_type argument to GSS_Init_sec_context. See [RFC2743] for details.
seqid4	typedef uint32_t seqid4;
utf8string	Sequence identifier used for file locking. typedef opaque utf8string<>; UTF-8 encoding for strings.
utf8str_cis	typedef utf8string utf8str_cis;
utf8str_cs	Case-insensitive UTF-8 string. typedef utf8string utf8str_cs;
utf8str_mixed	Case-sensitive UTF-8 string. typedef utf8string utf8str_mixed;
component4	UTF-8 strings with a case-sensitive prefix and a case-insensitive suffix. typedef utf8str_cs component4;
linktext4	Represents pathname components. typedef opaque linktext4;
ascii_REQUIRED4	Symbolic link contents ("symbolic link" is defined in an Open Group [openg_symlink] standard). typedef utf8string ascii_REQUIRED4;
pathname4	String MUST be sent as ASCII and thus is automatically UTF-8. typedef component4 pathname4<>; Represents path name for fs_locations.
nfs_lockid4	typedef uint64_t nfs_lockid4;
verifier4	typedef opaque verifier4[NFS4_VERIFIER_SIZE]; Verifier used for various operations (COMMIT, CREATE, OPEN, READDIR, WRITE) NFS4_VERIFIER_SIZE is defined as 8.

End of Base Data Types

Table 1

## 2.2. Structured Data Types

### 2.2.1. nfstime4

```
struct nfstime4 {  
    int64_t      seconds;  
    uint32_t     nseconds;  
};
```

The `nfstime4` structure gives the number of seconds and nanoseconds since midnight or 0 hour January 1, 1970 Coordinated Universal Time (UTC). Values greater than zero for the seconds field denote dates after the 0 hour January 1, 1970. Values less than zero for the seconds field denote dates before the 0 hour January 1, 1970. In both cases, the `nseconds` field is to be added to the seconds field for the final time representation. For example, if the time to be represented is one-half second before 0 hour January 1, 1970, the seconds field would have a value of negative one (-1) and the `nseconds` field would have a value of one-half second (500000000). Values greater than 999,999,999 for `nseconds` are considered invalid.

This data type is used to pass time and date information. A server converts to and from its local representation of time when processing time values, preserving as much accuracy as possible. If the precision of timestamps stored for a file system object is less than defined, loss of precision can occur. An adjunct time maintenance protocol is recommended to reduce client and server time skew.

### 2.2.2. time\_how4

```
enum time_how4 {  
    SET_TO_SERVER_TIME4 = 0,  
    SET_TO_CLIENT_TIME4 = 1  
};
```

### 2.2.3. setttime4

```
union setttime4 switch (time_how4 set_it) {  
    case SET_TO_CLIENT_TIME4:  
        nfstime4      time;  
    default:  
        void;  
};
```

The above definitions are used as the attribute definitions to set time values. If `set_it` is `SET_TO_SERVER_TIME4`, then the server uses its local representation of time for the time value.

#### 2.2.4. specdata4

```
struct specdata4 {
    uint32_t specdata1; /* major device number */
    uint32_t specdata2; /* minor device number */
};
```

This data type represents additional information for the device file types NF4CHR and NF4BLK.

#### 2.2.5. fsid4

```
struct fsid4 {
    uint64_t      major;
    uint64_t      minor;
};
```

This type is the file system identifier that is used as a mandatory attribute.

#### 2.2.6. fs\_location4

```
struct fs_location4 {
    utf8str_cis      server<>;
    pathname4        rootpath;
};
```

#### 2.2.7. fs\_locations4

```
struct fs_locations4 {
    pathname4        fs_root;
    fs_location4     locations<>;
};
```

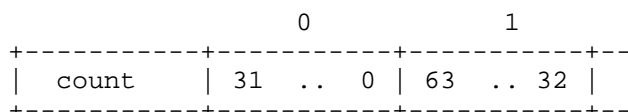
The fs\_location4 and fs\_locations4 data types are used for the fs\_locations recommended attribute which is used for migration and replication support.

#### 2.2.8. fattr4

```
struct fattr4 {
    bitmap4          attrmask;
    attrlist4        attr_vals;
};
```

The fattr4 structure is used to represent file and directory attributes.

The bitmap is a counted array of 32 bit integers used to contain bit values. The position of the integer in the array that contains bit  $n$  can be computed from the expression  $(n / 32)$  and its bit within that integer is  $(n \bmod 32)$ .



#### 2.2.9. change\_info4

```
struct change_info4 {
    bool                atomic;
    changeid4           before;
    changeid4           after;
};
```

This structure is used with the CREATE, LINK, REMOVE, RENAME operations to let the client know the value of the change attribute for the directory in which the target file system object resides.

#### 2.2.10. clientaddr4

```
struct clientaddr4 {
    /* see struct rpcb in RFC 1833 */
    string r_netid<>;    /* network id */
    string r_addr<>;     /* universal address */
};
```

The clientaddr4 structure is used as part of the SETCLIENTID operation to either specify the address of the client that is using a client ID or as part of the callback registration. The r\_netid and r\_addr fields respectively contain a netid and uaddr. The netid and uaddr concepts are defined in [RFC5665]. The netid and uaddr formats for TCP over IPv4 and TCP over IPv6 are defined in [RFC5665], specifically Tables 2 and 3 and Sections 5.2.3.3 and 5.2.3.4.

#### 2.2.11. cb\_client4

```
struct cb_client4 {
    unsigned int        cb_program;
    clientaddr4         cb_location;
};
```

This structure is used by the client to inform the server of its call back address; includes the program number and client address.

## 2.2.12. nfs\_client\_id4

```
struct nfs_client_id4 {  
    verifier4    verifier;  
    opaque       id<NFS4_OPAQUE_LIMIT>;  
};
```

This structure is part of the arguments to the SETCLIENTID operation.

## 2.2.13. open\_owner4

```
struct open_owner4 {  
    clientid4    clientid;  
    opaque       owner<NFS4_OPAQUE_LIMIT>;  
};
```

This structure is used to identify the owner of open state.

## 2.2.14. lock\_owner4

```
struct lock_owner4 {  
    clientid4    clientid;  
    opaque       owner<NFS4_OPAQUE_LIMIT>;  
};
```

This structure is used to identify the owner of file locking state.

## 2.2.15. open\_to\_lock\_owner4

```
struct open_to_lock_owner4 {  
    seqid4       open_seqid;  
    stateid4     open_stateid;  
    seqid4       lock_seqid;  
    lock_owner4  lock_owner;  
};
```

This structure is used for the first LOCK operation done for an open\_owner4. It provides both the open\_stateid and lock\_owner such that the transition is made from a valid open\_stateid sequence to that of the new lock\_stateid sequence. Using this mechanism avoids the confirmation of the lock\_owner/lock\_seqid pair since it is tied to established state in the form of the open\_stateid/open\_seqid.

#### 2.2.16. stateid4

```
struct stateid4 {  
    uint32_t      seqid;  
    opaque        other[NFS4_OTHER_SIZE];  
};
```

This structure is used for the various state sharing mechanisms between the client and server. For the client, this data structure is read-only. The server is required to increment the seqid field monotonically at each transition of the stateid. This is important since the client will inspect the seqid in OPEN stateids to determine the order of OPEN processing done by the server.

### 3. RPC and Security Flavor

The NFSv4 protocol is a RPC application that uses RPC version 2 and the XDR as defined in [RFC5531] and [RFC4506]. The RPCSEC\_GSS security flavors as defined in version 1 ([RFC2203]) and version 2 ([RFC5403]) MUST be implemented as the mechanism to deliver stronger security for the NFSv4 protocol. However, deployment of RPCSEC\_GSS is optional.

#### 3.1. Ports and Transports

Historically, NFSv2 and NFSv3 servers have resided on port 2049. The registered port 2049 [RFC3232] for the NFS protocol SHOULD be the default configuration. Using the registered port for NFS services means the NFS client will not need to use the RPC binding protocols as described in [RFC1833]; this will allow NFS to transit firewalls.

Where an NFSv4 implementation supports operation over the IP network protocol, the supported transport layer between NFS and IP MUST be an IETF standardised transport protocol that is specified to avoid network congestion; such transports include TCP and SCTP. To enhance the possibilities for interoperability, an NFSv4 implementation MUST support operation over the TCP transport protocol, at least until such time as a standards track RFC revises this requirement to use a different IETF standardised transport protocol with appropriate congestion control.

If TCP is used as the transport, the client and server SHOULD use persistent connections. This will prevent the weakening of TCP's congestion control via short lived connections and will improve performance for the Wide Area Network (WAN) environment by eliminating the need for SYN handshakes.



To date, all NFSv4 implementations are TCP based, i.e., there are none for SCTP nor UDP. UDP by itself is not sufficient as a transport for NFSv4, neither is UDP in combination with some other mechanism (e.g., DCCP [RFC4340], NORM [RFC5740]).

As noted in Section 17, the authentication model for NFSv4 has moved from machine-based to principal-based. However, this modification of the authentication model does not imply a technical requirement to move the TCP connection management model from whole machine-based to one based on a per user model. In particular, NFS over TCP client implementations have traditionally multiplexed traffic for multiple users over a common TCP connection between an NFS client and server. This has been true, regardless of whether the NFS client is using AUTH\_SYS, AUTH\_DH, RPCSEC\_GSS or any other flavor. Similarly, NFS over TCP server implementations have assumed such a model and thus scale the implementation of TCP connection management in proportion to the number of expected client machines. It is intended that NFSv4 will not modify this connection management model. NFSv4 clients that violate this assumption can expect scaling issues on the server and hence reduced service.

#### 3.1.1. Client Retransmission Behavior

When processing a NFSv4 request received over a reliable transport such as TCP, the NFSv4 server MUST NOT silently drop the request, except if the established transport connection has been broken. Given such a contract between NFSv4 clients and servers, clients MUST NOT retry a request unless one or both of the following are true:

- o The transport connection has been broken
- o The procedure being retried is the NULL procedure

Since reliable transports, such as TCP, do not always synchronously inform a peer when the other peer has broken the connection (for example, when an NFS server reboots), the NFSv4 client may want to actively "probe" the connection to see if has been broken. Use of the NULL procedure is one recommended way to do so. So, when a client experiences a remote procedure call timeout (of some arbitrary implementation specific amount), rather than retrying the remote procedure call, it could instead issue a NULL procedure call to the server. If the server has died, the transport connection break will eventually be indicated to the NFSv4 client. The client can then reconnect, and then retry the original request. If the NULL procedure call gets a response, the connection has not broken. The client can decide to wait longer for the original request's response, or it can break the transport connection and reconnect before re-sending the original request.

For callbacks from the server to the client, the same rules apply, but the server doing the callback becomes the client, and the client receiving the callback becomes the server.

### 3.2. Security Flavors

Traditional RPC implementations have included AUTH\_NONE, AUTH\_SYS, AUTH\_DH, and AUTH\_KRB4 as security flavors. With [RFC2203] an additional security flavor of RPCSEC\_GSS has been introduced which uses the functionality of GSS-API [RFC2743]. This allows for the use of various security mechanisms by the RPC layer without the additional implementation overhead of adding RPC security flavors. For NFSv4, the RPCSEC\_GSS security flavor MUST be used to enable the mandatory security mechanism. Other flavors, such as, AUTH\_NONE, AUTH\_SYS, and AUTH\_DH MAY be implemented as well.

#### 3.2.1. Security mechanisms for NFSv4

RPCSEC\_GSS, via GSS-API, normalizes access to mechanisms that provide security services. Therefore, NFSv4 clients and servers MUST support the Kerberos V5 security mechanism.

The use of RPCSEC\_GSS requires selection of mechanism, quality of protection (QOP), and service (authentication, integrity, privacy). For the mandated security mechanisms, NFSv4 specifies that a QOP of zero is used, leaving it up to the mechanism or the mechanism's configuration to map QOP zero to an appropriate level of protection. Each mandated mechanism specifies a minimum set of cryptographic algorithms for implementing integrity and privacy. NFSv4 clients and servers MUST be implemented on operating environments that comply with the REQUIRED cryptographic algorithms of each REQUIRED mechanism.

##### 3.2.1.1. Kerberos V5 as a Security Triple

The Kerberos V5 GSS-API mechanism as described in [RFC4121] MUST be implemented with the RPCSEC\_GSS services as specified in the following table:

column descriptions:

1 == number of pseudo flavor  
 2 == name of pseudo flavor  
 3 == mechanism's OID  
 4 == RPCSEC\_GSS service  
 5 == NFSv4 clients MUST support  
 6 == NFSv4 servers MUST support

1	2	3	4	5	6
390003	krb5	1.2.840.113554.1.2.2	rpc_gss_svc_none	yes	yes
390004	krb5i	1.2.840.113554.1.2.2	rpc_gss_svc_integrity	yes	yes
390005	krb5p	1.2.840.113554.1.2.2	rpc_gss_svc_privacy	no	yes

Note that the pseudo flavor is presented here as a mapping aid to the implementor. Because this NFS protocol includes a method to negotiate security and it understands the GSS-API mechanism, the pseudo flavor is not needed. The pseudo flavor is needed for NFSv3 since the security negotiation is done via the MOUNT protocol as described in [RFC2623].

At the time this document was specified, the Advanced Encryption Standard (AES) with HMAC-SHA1 was a REQUIRED algorithm set for Kerberos V5. In contrast, when NFSv4.0 was first specified in [RFC3530], weaker algorithm sets were REQUIRED for Kerberos V5, and were REQUIRED in the NFSv4.0 specification, because the Kerberos V5 specification at the time did not specify stronger algorithms. The NFSv4 specification does not specify REQUIRED algorithms for Kerberos V5, and instead, the implementor is expected to track the evolution of the Kerberos V5 standard if and when stronger algorithms are specified.

#### 3.2.1.1.1. Security Considerations for Cryptographic Algorithms in Kerberos V5

When deploying NFSv4, the strength of the security achieved depends on the existing Kerberos V5 infrastructure. The algorithms of Kerberos V5 are not directly exposed to or selectable by the client or server, so there is some due diligence required by the user of NFSv4 to ensure that security is acceptable where needed. Guidance is provided in [RFC6649] as to why weak algorithms should be disabled by default.

### 3.3. Security Negotiation

With the NFSv4 server potentially offering multiple security mechanisms, the client needs a method to determine or negotiate which mechanism is to be used for its communication with the server. The

NFS server may have multiple points within its file system name space that are available for use by NFS clients. In turn the NFS server may be configured such that each of these entry points may have different or multiple security mechanisms in use.

The security negotiation between client and server SHOULD be done with a secure channel to eliminate the possibility of a third party intercepting the negotiation sequence and forcing the client and server to choose a lower level of security than required or desired. See Section 17 for further discussion.

#### 3.3.1. SECINFO

The new SECINFO operation will allow the client to determine, on a per filehandle basis, what security triple (see [RFC2743]) is to be used for server access. In general, the client will not have to use the SECINFO operation except during initial communication with the server or when the client crosses policy boundaries at the server. It is possible that the server's policies change during the client's interaction therefore forcing the client to negotiate a new security triple.

#### 3.3.2. Security Error

Based on the assumption that each NFSv4 client and server MUST support a minimum set of security (i.e., Kerberos-V5 under RPCSEC\_GSS), the NFS client will start its communication with the server with one of the minimal security triples. During communication with the server, the client may receive an NFS error of NFS4ERR\_WRONGSEC. This error allows the server to notify the client that the security triple currently being used is not appropriate for access to the server's file system resources. The client is then responsible for determining what security triples are available at the server and choose one which is appropriate for the client. See Section 15.33 for further discussion of how the client will respond to the NFS4ERR\_WRONGSEC error and use SECINFO.

#### 3.3.3. Callback RPC Authentication

Except as noted elsewhere in this section, the callback RPC (described later) MUST mutually authenticate the NFS server to the principal that acquired the client ID (also described later), using the security flavor the original SETCLIENTID operation used.

For AUTH\_NONE, there are no principals, so this is a non-issue.

AUTH\_SYS has no notions of mutual authentication or a server principal, so the callback from the server simply uses the AUTH\_SYS

credential that the user used when he set up the delegation.

For AUTH\_DH, one commonly used convention is that the server uses the credential corresponding to this AUTH\_DH principal:

```
unix.host@domain
```

where host and domain are variables corresponding to the name of server host and directory services domain in which it lives such as a Network Information System domain or a DNS domain.

Regardless of what security mechanism under RPCSEC\_GSS is being used, the NFS server MUST identify itself in GSS-API via a GSS\_C\_NT\_HOSTBASED\_SERVICE name type. GSS\_C\_NT\_HOSTBASED\_SERVICE names are of the form:

```
service@hostname
```

For NFS, the "service" element is

```
nfs
```

Implementations of security mechanisms will convert nfs@hostname to various different forms. For Kerberos V5, the following form is RECOMMENDED:

```
nfs/hostname
```

For Kerberos V5, nfs/hostname would be a server principal in the Kerberos Key Distribution Center database. This is the same principal the client acquired a GSS-API context for when it issued the SETCLIENTID operation, therefore, the realm name for the server principal must be the same for the callback as it was for the SETCLIENTID.

#### 4. Filehandles

The filehandle in the NFS protocol is a per server unique identifier for a file system object. The contents of the filehandle are opaque to the client. Therefore, the server is responsible for translating the filehandle to an internal representation of the file system object.

##### 4.1. Obtaining the First Filehandle

The operations of the NFS protocol are defined in terms of one or more filehandles. Therefore, the client needs a filehandle to

initiate communication with the server. With the NFSv2 protocol [RFC1094] and the NFSv3 protocol [RFC1813], there exists an ancillary protocol to obtain this first filehandle. The MOUNT protocol, RPC program number 100005, provides the mechanism of translating a string based file system path name to a filehandle which can then be used by the NFS protocols.

The MOUNT protocol has deficiencies in the area of security and use via firewalls. This is one reason that the use of the public filehandle was introduced in [RFC2054] and [RFC2055]. With the use of the public filehandle in combination with the LOOKUP operation in the NFSv2 and NFSv3 protocols, it has been demonstrated that the MOUNT protocol is unnecessary for viable interaction between NFS client and server.

Therefore, the NFSv4 protocol will not use an ancillary protocol for translation from string based path names to a filehandle. Two special filehandles will be used as starting points for the NFS client.

#### 4.1.1. Root Filehandle

The first of the special filehandles is the ROOT filehandle. The ROOT filehandle is the "conceptual" root of the file system name space at the NFS server. The client uses or starts with the ROOT filehandle by employing the PUTROOTFH operation. The PUTROOTFH operation instructs the server to set the "current" filehandle to the ROOT of the server's file tree. Once this PUTROOTFH operation is used, the client can then traverse the entirety of the server's file tree with the LOOKUP operation. A complete discussion of the server name space is in Section 7.

#### 4.1.2. Public Filehandle

The second special filehandle is the PUBLIC filehandle. Unlike the ROOT filehandle, the PUBLIC filehandle may be bound or represent an arbitrary file system object at the server. The server is responsible for this binding. It may be that the PUBLIC filehandle and the ROOT filehandle refer to the same file system object. However, it is up to the administrative software at the server and the policies of the server administrator to define the binding of the PUBLIC filehandle and server file system object. The client may not make any assumptions about this binding. The client uses the PUBLIC filehandle via the PUTPUBFH operation.

#### 4.2. Filehandle Types

In the NFSv2 and NFSv3 protocols, there was one type of filehandle with a single set of semantics. This type of filehandle is termed "persistent" in NFS Version 4. The semantics of a persistent filehandle remain the same as before. A new type of filehandle introduced in NFS Version 4 is the "volatile" filehandle, which attempts to accommodate certain server environments.

The volatile filehandle type was introduced to address server functionality or implementation issues which make correct implementation of a persistent filehandle infeasible. Some server environments do not provide a file system level invariant that can be used to construct a persistent filehandle. The underlying server file system may not provide the invariant or the server's file system programming interfaces may not provide access to the needed invariant. Volatile filehandles may ease the implementation of server functionality such as hierarchical storage management or file system reorganization or migration. However, the volatile filehandle increases the implementation burden for the client.

Since the client will need to handle persistent and volatile filehandles differently, a file attribute is defined which may be used by the client to determine the filehandle types being returned by the server.

##### 4.2.1. General Properties of a Filehandle

The filehandle contains all the information the server needs to distinguish an individual file. To the client, the filehandle is opaque. The client stores filehandles for use in a later request and can compare two filehandles from the same server for equality by doing a byte-by-byte comparison. However, the client **MUST NOT** otherwise interpret the contents of filehandles. If two filehandles from the same server are equal, they **MUST** refer to the same file system object. Servers **SHOULD** try to maintain a one-to-one correspondence between filehandles and file system objects but this is not required. Clients **MUST** use filehandle comparisons only to improve performance, not for correct behavior. All clients need to be prepared for situations in which it cannot be determined whether two filehandles denote the same object and in such cases, avoid making invalid assumptions which might cause incorrect behavior. Further discussion of filehandle and attribute comparison in the context of data caching is presented in Section 10.3.4.

As an example, in the case that two different path names when traversed at the server terminate at the same file system object, the server **SHOULD** return the same filehandle for each path. This can

occur if a hard link is used to create two file names which refer to the same underlying file object and associated data. For example, if paths /a/b/c and /a/d/c refer to the same file, the server SHOULD return the same filehandle for both path names traversals.

#### 4.2.2. Persistent Filehandle

A persistent filehandle is defined as having a fixed value for the lifetime of the file system object to which it refers. Once the server creates the filehandle for a file system object, the server MUST accept the same filehandle for the object for the lifetime of the object. If the server restarts or reboots the NFS server must honor the same filehandle value as it did in the server's previous instantiation. Similarly, if the file system is migrated, the new NFS server must honor the same filehandle as the old NFS server.

The persistent filehandle will become stale or invalid when the file system object is removed. When the server is presented with a persistent filehandle that refers to a deleted object, it MUST return an error of NFS4ERR\_STALE. A filehandle may become stale when the file system containing the object is no longer available. The file system may become unavailable if it exists on removable media and the media is no longer available at the server or the file system in whole has been destroyed or the file system has simply been removed from the server's name space (i.e., unmounted in a UNIX environment).

#### 4.2.3. Volatile Filehandle

A volatile filehandle does not share the same longevity characteristics of a persistent filehandle. The server may determine that a volatile filehandle is no longer valid at many different points in time. If the server can definitively determine that a volatile filehandle refers to an object that has been removed, the server should return NFS4ERR\_STALE to the client (as is the case for persistent filehandles). In all other cases where the server determines that a volatile filehandle can no longer be used, it should return an error of NFS4ERR\_FHEXPIRED.

The mandatory attribute "fh\_expire\_type" is used by the client to determine what type of filehandle the server is providing for a particular file system. This attribute is a bitmask with the following values:

**FH4\_PERSISTENT:** The value of FH4\_PERSISTENT is used to indicate a persistent filehandle, which is valid until the object is removed from the file system. The server will not return NFS4ERR\_FHEXPIRED for this filehandle. FH4\_PERSISTENT is defined as a value in which none of the bits specified below are set.



**FH4\_VOLATILE\_ANY:** The filehandle may expire at any time, except as specifically excluded (i.e., **FH4\_NOEXPIRE\_WITH\_OPEN**).

**FH4\_NOEXPIRE\_WITH\_OPEN:** May only be set when **FH4\_VOLATILE\_ANY** is set. If this bit is set, then the meaning of **FH4\_VOLATILE\_ANY** is qualified to exclude any expiration of the filehandle when it is open.

**FH4\_VOL\_MIGRATION:** The filehandle will expire as a result of migration. If **FH4\_VOLATILE\_ANY** is set, **FH4\_VOL\_MIGRATION** is redundant.

**FH4\_VOL\_RENAME:** The filehandle will expire during rename. This includes a rename by the requesting client or a rename by any other client. If **FH4\_VOLATILE\_ANY** is set, **FH4\_VOL\_RENAME** is redundant.

Servers which provide volatile filehandles that may expire while open (i.e., if **FH4\_VOL\_MIGRATION** or **FH4\_VOL\_RENAME** is set or if **FH4\_VOLATILE\_ANY** is set and **FH4\_NOEXPIRE\_WITH\_OPEN** not set), should deny a **RENAME** or **REMOVE** that would affect an **OPEN** file of any of the components leading to the **OPEN** file. In addition, the server should deny all **RENAME** or **REMOVE** requests during the grace period upon server restart.

Note that the bits **FH4\_VOL\_MIGRATION** and **FH4\_VOL\_RENAME** allow the client to determine that expiration has occurred whenever a specific event occurs, without an explicit filehandle expiration error from the server. **FH4\_VOLATILE\_ANY** does not provide this form of information. In situations where the server will expire many, but not all filehandles upon migration (e.g., all but those that are open), **FH4\_VOLATILE\_ANY** (in this case with **FH4\_NOEXPIRE\_WITH\_OPEN**) is a better choice since the client may not assume that all filehandles will expire when migration occurs, and it is likely that additional expirations will occur (as a result of file **CLOSE**) that are separated in time from the migration event itself.

#### 4.2.4. One Method of Constructing a Volatile Filehandle

A volatile filehandle, while opaque to the client could contain:

[volatile bit = 1 | server boot time | slot | generation number]

- o slot is an index in the server volatile filehandle table
- o generation number is the generation number for the table entry/  
slot

When the client presents a volatile filehandle, the server makes the following checks, which assume that the check for the volatile bit has passed. If the server boot time is less than the current server boot time, return NFS4ERR\_FHEXPIRED. If slot is out of range, return NFS4ERR\_BADHANDLE. If the generation number does not match, return NFS4ERR\_FHEXPIRED.

When the server reboots, the table is gone (it is volatile).

If volatile bit is 0, then it is a persistent filehandle with a different structure following it.

#### 4.3. Client Recovery from Filehandle Expiration

If possible, the client should recover from the receipt of an NFS4ERR\_FHEXPIRED error. The client must take on additional responsibility so that it may prepare itself to recover from the expiration of a volatile filehandle. If the server returns persistent filehandles, the client does not need these additional steps.

For volatile filehandles, most commonly the client will need to store the component names leading up to and including the file system object in question. With these names, the client should be able to recover by finding a filehandle in the name space that is still available or by starting at the root of the server's file system name space.

If the expired filehandle refers to an object that has been removed from the file system, obviously the client will not be able to recover from the expired filehandle.

It is also possible that the expired filehandle refers to a file that has been renamed. If the file was renamed by another client, again it is possible that the original client will not be able to recover. However, in the case that the client itself is renaming the file and the file is open, it is possible that the client may be able to recover. The client can determine the new path name based on the processing of the rename request. The client can then regenerate the new filehandle based on the new path name. The client could also use the compound operation mechanism to construct a set of operations like:

```
RENAME A B
LOOKUP B
GETFH
```

Note that the COMPOUND procedure does not provide atomicity. This

example only reduces the overhead of recovering from an expired filehandle.

## 5. Attributes

To meet the requirements of extensibility and increased interoperability with non-UNIX platforms, attributes need to be handled in a flexible manner. The NFSv3 `fatattr3` structure contains a fixed list of attributes that not all clients and servers are able to support or care about. The `fatattr3` structure cannot be extended as new needs arise and it provides no way to indicate non-support. With the NFSv4.0 protocol, the client is able to query what attributes the server supports and construct requests with only those supported attributes (or a subset thereof).

To this end, attributes are divided into three groups: REQUIRED, RECOMMENDED, and named. Both REQUIRED and RECOMMENDED attributes are supported in the NFSv4.0 protocol by a specific and well-defined encoding and are identified by number. They are requested by setting a bit in the bit vector sent in the GETATTR request; the server response includes a bit vector to list what attributes were returned in the response. New REQUIRED or RECOMMENDED attributes may be added to the NFSv4 protocol as part of a new minor version by publishing a Standards Track RFC which allocates a new attribute number value and defines the encoding for the attribute. See Section 11 for further discussion.

Named attributes are accessed by the OPENATTR operation, which accesses a hidden directory of attributes associated with a file system object. OPENATTR takes a filehandle for the object and returns the filehandle for the attribute hierarchy. The filehandle for the named attributes is a directory object accessible by LOOKUP or REaddir and contains files whose names represent the named attributes and whose data bytes are the value of the attribute. For example:

LOOKUP	"foo"	; look up file
GETATTR	attrbits	
OPENATTR		; access foo's named attributes
LOOKUP	"xllicon"	; look up specific attribute
READ	0,4096	; read stream of bytes

Named attributes are intended for data needed by applications rather than by an NFS client implementation. NFS implementors are strongly encouraged to define their new attributes as RECOMMENDED attributes

by bringing them to the IETF Standards Track process.

The set of attributes that are classified as REQUIRED is deliberately small since servers need to do whatever it takes to support them. A server should support as many of the RECOMMENDED attributes as possible but, by their definition, the server is not required to support all of them. Attributes are deemed REQUIRED if the data is both needed by a large number of clients and is not otherwise reasonably computable by the client when support is not provided on the server.

Note that the hidden directory returned by OPENATTR is a convenience for protocol processing. The client should not make any assumptions about the server's implementation of named attributes and whether or not the underlying file system at the server has a named attribute directory. Therefore, operations such as SETATTR and GETATTR on the named attribute directory are undefined.

#### 5.1. REQUIRED Attributes

These MUST be supported by every NFSv4.0 client and server in order to ensure a minimum level of interoperability. The server MUST store and return these attributes, and the client MUST be able to function with an attribute set limited to these attributes. With just the REQUIRED attributes some client functionality may be impaired or limited in some ways. A client may ask for any of these attributes to be returned by setting a bit in the GETATTR request, and the server MUST return their value.

#### 5.2. RECOMMENDED Attributes

These attributes are understood well enough to warrant support in the NFSv4.0 protocol. However, they may not be supported on all clients and servers. A client MAY ask for any of these attributes to be returned by setting a bit in the GETATTR request but must handle the case where the server does not return them. A client MAY ask for the set of attributes the server supports and SHOULD NOT request attributes the server does not support. A server should be tolerant of requests for unsupported attributes and simply not return them rather than considering the request an error. It is expected that servers will support all attributes they comfortably can and only fail to support attributes that are difficult to support in their operating environments. A server should provide attributes whenever they don't have to "tell lies" to the client. For example, a file modification time should be either an accurate time or should not be supported by the server. At times this will be difficult for clients, but a client is better positioned to decide whether and how to fabricate or construct an attribute or whether to do without the

attribute.

### 5.3. Named Attributes

These attributes are not supported by direct encoding in the NFSv4 protocol but are accessed by string names rather than numbers and correspond to an uninterpreted stream of bytes that are stored with the file system object. The name space for these attributes may be accessed by using the OPENATTR operation. The OPENATTR operation returns a filehandle for a virtual "named attribute directory", and further perusal and modification of the name space may be done using operations that work on more typical directories. In particular, READDIR may be used to get a list of such named attributes, and LOOKUP and OPEN may select a particular attribute. Creation of a new named attribute may be the result of an OPEN specifying file creation.

Once an OPEN is done, named attributes may be examined and changed by normal READ and WRITE operations using the filehandles and stateids returned by OPEN.

Named attributes and the named attribute directory may have their own (non-named) attributes. Each of these objects must have all of the REQUIRED attributes and may have additional RECOMMENDED attributes. However, the set of attributes for named attributes and the named attribute directory need not be, and typically will not be, as large as that for other objects in that file system.

Named attributes might be the target of delegations. However, since granting of delegations is at the server's discretion, a server need not support delegations on named attributes.

It is RECOMMENDED that servers support arbitrary named attributes. A client should not depend on the ability to store any named attributes in the server's file system. If a server does support named attributes, a client that is also able to handle them should be able to copy a file's data and metadata with complete transparency from one location to another; this would imply that names allowed for regular directory entries are valid for named attribute names as well.

In NFSv4.0, the structure of named attribute directories is restricted in a number of ways, in order to prevent the development of non-interoperable implementations in which some servers support a fully general hierarchical directory structure for named attributes while others support a limited but adequate structure for named attributes. In such an environment, clients or applications might come to depend on non-portable extensions. The restrictions are:

- o CREATE is not allowed in a named attribute directory. Thus, such objects as symbolic links and special files are not allowed to be named attributes. Further, directories may not be created in a named attribute directory, so no hierarchical structure of named attributes for a single object is allowed.
- o If OPENATTR is done on a named attribute directory or on a named attribute, the server MUST return an error.
- o Doing a RENAME of a named attribute to a different named attribute directory or to an ordinary (i.e., non-named-attribute) directory is not allowed.
- o Creating hard links between named attribute directories or between named attribute directories and ordinary directories is not allowed.

Names of attributes will not be controlled by this document or other IETF Standards Track documents. See Section 18 for further discussion.

#### 5.4. Classification of Attributes

Each of the REQUIRED and RECOMMENDED attributes can be classified in one of three categories: per server (i.e., the value of the attribute will be the same for all file objects that share the same server), per file system (i.e., the value of the attribute will be the same for some or all file objects that share the same fsid attribute (Section 5.8.1.9) and server owner), or per file system object. Note that it is possible that some per file system attributes may vary within the file system. Note that it is possible that some per file system attributes may vary within the file system, depending on the value of the "homogeneous" (Section 5.8.2.16) attribute. Note that the attributes `time_access_set` and `time_modify_set` are not listed in this section because they are write-only attributes corresponding to `time_access` and `time_modify`, and are used in a special instance of SETATTR.

- o The per-server attribute is:

`lease_time`

- o The per-file system attributes are:

`supported_attrs`, `fh_expire_type`, `link_support`, `symlink_support`,  
`unique_handles`, `aclsupport`, `cansettime`, `case_insensitive`,  
`case_preserving`, `chown_restricted`, `files_avail`, `files_free`,  
`files_total`, `fs_locations`, `homogeneous`, `maxfilesize`, `maxname`,

maxread, maxwrite, no\_trunc, space\_avail, space\_free,  
space\_total, time\_delta,

- o The per-file system object attributes are:

type, change, size, named\_attr, fsid, rdattrib\_error, filehandle,  
acl, archive, fileid, hidden, maxlink, mimetype, mode,  
numlinks, owner, owner\_group, rawdev, space\_used, system,  
time\_access, time\_backup, time\_create, time\_metadata,  
time\_modify, mounted\_on\_fileid

For quota\_avail\_hard, quota\_avail\_soft, and quota\_used, see their definitions below for the appropriate classification.

#### 5.5. Set-Only and Get-Only Attributes

Some REQUIRED and RECOMMENDED attributes are set-only; i.e., they can be set via SETATTR but not retrieved via GETATTR. Similarly, some REQUIRED and RECOMMENDED attributes are get-only; i.e., they can be retrieved via GETATTR but not set via SETATTR. If a client attempts to set a get-only attribute or get a set-only attribute, the server MUST return NFS4ERR\_INVALID.

#### 5.6. REQUIRED Attributes - List and Definition References

The list of REQUIRED attributes appears in Table 2. The meaning of the columns of the table are:

- o Name: The name of attribute
- o Id: The number assigned to the attribute. In the event of conflicts between the assigned number and [I-D.ietf-nfsv4-rfc3530bis-dot-x], the latter is authoritative, but in such an event, it should be resolved with Errata to this document and/or [I-D.ietf-nfsv4-rfc3530bis-dot-x]. See [ISEG\_errata] for the Errata process.
- o Data Type: The XDR data type of the attribute.
- o Acc: Access allowed to the attribute. R means read-only (GETATTR may retrieve, SETATTR may not set). W means write-only (SETATTR may set, GETATTR may not retrieve). R W means read/write (GETATTR may retrieve, SETATTR may set).
- o Defined in: The section of this specification that describes the attribute.

Name	Id	Data Type	Acc	Defined in:
supported_attrs	0	bitmap4	R	Section 5.8.1.1
type	1	nfs_ftype4	R	Section 5.8.1.2
fh_expire_type	2	uint32_t	R	Section 5.8.1.3
change	3	uint64_t	R	Section 5.8.1.4
size	4	uint64_t	R W	Section 5.8.1.5
link_support	5	bool	R	Section 5.8.1.6
symlink_support	6	bool	R	Section 5.8.1.7
named_attr	7	bool	R	Section 5.8.1.8
fsid	8	fsid4	R	Section 5.8.1.9
unique_handles	9	bool	R	Section 5.8.1.10
lease_time	10	nfs_lease4	R	Section 5.8.1.11
rdattr_error	11	nfsstat4	R	Section 5.8.1.12
filehandle	19	nfs_fh4	R	Section 5.8.1.13

Table 2

#### 5.7. RECOMMENDED Attributes - List and Definition References

The RECOMMENDED attributes are defined in Table 3. The meanings of the column headers are the same as Table 2; see Section 5.6 for the meanings.

Name	Id	Data Type	Acc	Defined in:
acl	12	nfsace4<>	R W	Section 6.2.1
aclsupport	13	uint32_t	R	Section 6.2.1.2
archive	14	bool	R W	Section 5.8.2.1
cansettime	15	bool	R	Section 5.8.2.2
case_insensitive	16	bool	R	Section 5.8.2.3
case_preserving	17	bool	R	Section 5.8.2.4
chown_restricted	18	bool	R	Section 5.8.2.5
fileid	20	uint64_t	R	Section 5.8.2.6
files_avail	21	uint64_t	R	Section 5.8.2.7
files_free	22	uint64_t	R	Section 5.8.2.8
files_total	23	uint64_t	R	Section 5.8.2.9
fs_locations	24	fs_locations	R	Section 5.8.2.10
hidden	25	bool	R W	Section 5.8.2.11
homogeneous	26	bool	R	Section 5.8.2.12
maxfilesize	27	uint64_t	R	Section 5.8.2.13
maxlink	28	uint32_t	R	Section 5.8.2.14
maxname	29	uint32_t	R	Section 5.8.2.15
maxread	30	uint64_t	R	Section 5.8.2.16
maxwrite	31	uint64_t	R	Section 5.8.2.17



mimetype	32	ascii_ REQUIRED4<>	R W	Section 5.8.2.18
mode	33	mode4	R W	Section 6.2.2
mounted_on_fileid	55	uint64_t	R	Section 5.8.2.19
no_trunc	34	bool	R	Section 5.8.2.20
numlinks	35	uint32_t	R	Section 5.8.2.21
owner	36	utf8<>	R W	Section 5.8.2.22
owner_group	37	utf8<>	R W	Section 5.8.2.23
quota_avail_hard	38	uint64_t	R	Section 5.8.2.24
quota_avail_soft	39	uint64_t	R	Section 5.8.2.25
quota_used	40	uint64_t	R	Section 5.8.2.26
rawdev	41	specdata4	R	Section 5.8.2.27
space_avail	42	uint64_t	R	Section 5.8.2.28
space_free	43	uint64_t	R	Section 5.8.2.29
space_total	44	uint64_t	R	Section 5.8.2.30
space_used	45	uint64_t	R	Section 5.8.2.31
system	46	bool	R W	Section 5.8.2.32
time_access	47	nfstime4	R	Section 5.8.2.33
time_access_set	48	settime4	W	Section 5.8.2.34
time_backup	49	nfstime4	R W	Section 5.8.2.35
time_create	50	nfstime4	R W	Section 5.8.2.36
time_delta	51	nfstime4	R	Section 5.8.2.37
time_metadata	52	nfstime4	R	Section 5.8.2.38
time_modify	53	nfstime4	R	Section 5.8.2.39
time_modify_set	54	settime4	W	Section 5.8.2.40

Table 3

## 5.8. Attribute Definitions

### 5.8.1. Definitions of REQUIRED Attributes

#### 5.8.1.1. Attribute 0: supported\_attrs

The bit vector that would retrieve all REQUIRED and RECOMMENDED attributes that are supported for this object. The scope of this attribute applies to all objects with a matching fsid.

#### 5.8.1.2. Attribute 1: type

Designates the type of an object in terms of one of a number of special constants:

- o NF4REG designates a regular file.
- o NF4DIR designates a directory.

- o NF4BLK designates a block device special file.
- o NF4CHR designates a character device special file.
- o NF4LNK designates a symbolic link.
- o NF4SOCK designates a named socket special file.
- o NF4FIFO designates a fifo special file.
- o NF4ATTRDIR designates a named attribute directory.
- o NF4NAMEDATTR designates a named attribute.

Within the explanatory text and operation descriptions, the following phrases will be used with the meanings given below:

- o The phrase "is a directory" means that the object's type attribute is NF4DIR or NF4ATTRDIR.
- o The phrase "is a special file" means that the object's type attribute is NF4BLK, NF4CHR, NF4SOCK, or NF4FIFO.
- o The phrase "is an regular file" means that the object's type attribute is NF4REG or NF4NAMEDATTR.

#### 5.8.1.3. Attribute 2: fh\_expire\_type

Server uses this to specify filehandle expiration behavior to the client. See Section 4 for additional description.

#### 5.8.1.4. Attribute 3: change

A value created by the server that the client can use to determine if file data, directory contents, or attributes of the object have been modified. The server MAY return the object's time\_metadata attribute for this attribute's value but only if the file system object cannot be updated more frequently than the resolution of time\_metadata.

#### 5.8.1.5. Attribute 4: size

The size of the object in bytes.

#### 5.8.1.6. Attribute 5: link\_support

TRUE, if the object's file system supports hard links.

#### 5.8.1.7. Attribute 6: symlink\_support

TRUE, if the object's file system supports symbolic links.

#### 5.8.1.8. Attribute 7: named\_attr

TRUE, if this object has named attributes. In other words, object has a non-empty named attribute directory.

#### 5.8.1.9. Attribute 8: fsid

Unique file system identifier for the file system holding this object. The fsid attribute has major and minor components, each of which are of data type uint64\_t.

#### 5.8.1.10. Attribute 9: unique\_handles

TRUE, if two distinct filehandles are guaranteed to refer to two different file system objects.

#### 5.8.1.11. Attribute 10: lease\_time

Duration of the lease at server in seconds.

#### 5.8.1.12. Attribute 11: rdattrib\_error

Error returned from an attempt to retrieve attributes during a READDIR operation.

#### 5.8.1.13. Attribute 19: filehandle

The filehandle of this object (primarily for READDIR requests).

### 5.8.2. Definitions of Uncategorized RECOMMENDED Attributes

The definitions of most of the RECOMMENDED attributes follow. Collections that share a common category are defined in other sections.

#### 5.8.2.1. Attribute 14: archive

TRUE, if this file has been archived since the time of last modification (deprecated in favor of time\_backup).

#### 5.8.2.2. Attribute 15: cansettime

TRUE, if the server is able to change the times for a file system object as specified in a SETATTR operation.

#### 5.8.2.3. Attribute 16: case\_insensitive

TRUE, if file name comparisons on this file system are case insensitive.

#### 5.8.2.4. Attribute 17: case\_preserving

TRUE, if file name case on this file system is preserved.

#### 5.8.2.5. Attribute 18: chown\_restricted

If TRUE, the server will reject any request to change either the owner or the group associated with a file if the caller is not a privileged user (for example, "root" in UNIX operating environments or in Windows 2000, the "Take Ownership" privilege).

#### 5.8.2.6. Attribute 20: fileid

A number uniquely identifying the file within the file system.

#### 5.8.2.7. Attribute 21: files\_avail

File slots available to this user on the file system containing this object -- this should be the smallest relevant limit.

#### 5.8.2.8. Attribute 22: files\_free

Free file slots on the file system containing this object - this should be the smallest relevant limit.

#### 5.8.2.9. Attribute 23: files\_total

Total file slots on the file system containing this object.

#### 5.8.2.10. Attribute 24: fs\_locations

Locations where this file system may be found. If the server returns NFS4ERR\_MOVED as an error, this attribute MUST be supported.

The server specifies the root path for a given server by returning a path consisting of zero path components.

#### 5.8.2.11. Attribute 25: hidden

TRUE, if the file is considered hidden with respect to the Windows API.

## 5.8.2.12. Attribute 26: homogeneous

TRUE, if this object's file system is homogeneous, i.e., all objects in the file system (all objects on the server with the same fsid) have common values for all per-file-system attributes.

## 5.8.2.13. Attribute 27: maxfilesize

Maximum supported file size for the file system of this object.

## 5.8.2.14. Attribute 28: maxlink

Maximum number of links for this object.

## 5.8.2.15. Attribute 29: maxname

Maximum file name size supported for this object.

## 5.8.2.16. Attribute 30: maxread

Maximum amount of data the READ operation will return for this object.

## 5.8.2.17. Attribute 31: maxwrite

Maximum amount of data the WRITE operation will accept for this object. This attribute SHOULD be supported if the file is writable. Lack of this attribute can lead to the client either wasting bandwidth or not receiving the best performance.

## 5.8.2.18. Attribute 32: mimetype

MIME body type/subtype of this object.

## 5.8.2.19. Attribute 55: mounted\_on\_fileid

Like fileid, but if the target filehandle is the root of a file system, this attribute represents the fileid of the underlying directory.

UNIX-based operating environments connect a file system into the namespace by connecting (mounting) the file system onto the existing file object (the mount point, usually a directory) of an existing file system. When the mount point's parent directory is read via an API like `readdir()`, the return results are directory entries, each with a component name and a fileid. The fileid of the mount point's directory entry will be different from the fileid that the `stat()` system call returns. The `stat()` system call is returning the fileid

of the root of the mounted file system, whereas `readdir()` is returning the fileid that `stat()` would have returned before any file systems were mounted on the mount point.

Unlike NFSv3, NFSv4.0 allows a client's LOOKUP request to cross other file systems. The client detects the file system crossing whenever the filehandle argument of LOOKUP has an fsid attribute different from that of the filehandle returned by LOOKUP. A UNIX-based client will consider this a "mount point crossing". UNIX has a legacy scheme for allowing a process to determine its current working directory. This relies on `readdir()` of a mount point's parent and `stat()` of the mount point returning fileids as previously described. The `mounted_on_fileid` attribute corresponds to the fileid that `readdir()` would have returned as described previously.

While the NFSv4.0 client could simply fabricate a fileid corresponding to what `mounted_on_fileid` provides (and if the server does not support `mounted_on_fileid`, the client has no choice), there is a risk that the client will generate a fileid that conflicts with one that is already assigned to another object in the file system. Instead, if the server can provide the `mounted_on_fileid`, the potential for client operational problems in this area is eliminated.

If the server detects that there is no mounted point at the target file object, then the value for `mounted_on_fileid` that it returns is the same as that of the fileid attribute.

The `mounted_on_fileid` attribute is RECOMMENDED, so the server SHOULD provide it if possible, and for a UNIX-based server, this is straightforward. Usually, `mounted_on_fileid` will be requested during a READDIR operation, in which case it is trivial (at least for UNIX-based servers) to return `mounted_on_fileid` since it is equal to the fileid of a directory entry returned by `readdir()`. If `mounted_on_fileid` is requested in a GETATTR operation, the server should obey an invariant that has it returning a value that is equal to the file object's entry in the object's parent directory, i.e., what `readdir()` would have returned. Some operating environments allow a series of two or more file systems to be mounted onto a single mount point. In this case, for the server to obey the aforementioned invariant, it will need to find the base mount point, and not the intermediate mount points.

#### 5.8.2.20. Attribute 34: `no_trunc`

If this attribute is TRUE, then if the client uses a file name longer than `name_max`, an error will be returned instead of the name being truncated.

## 5.8.2.21. Attribute 35: numlinks

Number of hard links to this object.

## 5.8.2.22. Attribute 36: owner

The string name of the owner of this object.

## 5.8.2.23. Attribute 37: owner\_group

The string name of the group ownership of this object.

## 5.8.2.24. Attribute 38: quota\_avail\_hard

The value in bytes that represents the amount of additional disk space beyond the current allocation that can be allocated to this file or directory before further allocations will be refused. It is understood that this space may be consumed by allocations to other files or directories.

## 5.8.2.25. Attribute 39: quota\_avail\_soft

The value in bytes that represents the amount of additional disk space that can be allocated to this file or directory before the user may reasonably be warned. It is understood that this space may be consumed by allocations to other files or directories though there may exist server side rules as to which other files or directories.

## 5.8.2.26. Attribute 40: quota\_used

The value in bytes that represents the amount of disk space used by this file or directory and possibly a number of other similar files or directories, where the set of "similar" meets at least the criterion that allocating space to any file or directory in the set will reduce the "quota\_avail\_hard" of every other file or directory in the set.

Note that there may be a number of distinct but overlapping sets of files or directories for which a quota\_used value is maintained, e.g., "all files with a given owner", "all files with a given group owner", etc. The server is at liberty to choose any of those sets when providing the content of the quota\_used attribute, but should do so in a repeatable way. The rule may be configured per file system or may be "choose the set with the smallest quota".

## 5.8.2.27. Attribute 41: rawdev

Raw device number of file of type NF4BLK or NF4CHR. The device number is split into major and minor numbers. If the file's type attribute is not NF4BLK or NF4CHR, the value returned SHOULD NOT be considered useful.

## 5.8.2.28. Attribute 42: space\_avail

Disk space in bytes available to this user on the file system containing this object -- this should be the smallest relevant limit.

## 5.8.2.29. Attribute 43: space\_free

Free disk space in bytes on the file system containing this object -- this should be the smallest relevant limit.

## 5.8.2.30. Attribute 44: space\_total

Total disk space in bytes on the file system containing this object.

## 5.8.2.31. Attribute 45: space\_used

Number of file system bytes allocated to this object.

## 5.8.2.32. Attribute 46: system

This attribute is TRUE if this file is a "system" file with respect to the Windows operating environment.

## 5.8.2.33. Attribute 47: time\_access

The time\_access attribute represents the time of last access to the object by a READ operation sent to the server. The notion of what is an "access" depends on the server's operating environment and/or the server's file system semantics. For example, for servers obeying Portable Operating System Interface (POSIX) semantics, time\_access would be updated only by the READ and REaddir operations and not any of the operations that modify the content of the object [16], [17], [read\_apil], [readdir\_apil], [write\_apil]. Of course, setting the corresponding time\_access\_set attribute is another way to modify the time\_access attribute.

Whenever the file object resides on a writable file system, the server should make its best efforts to record time\_access into stable storage. However, to mitigate the performance effects of doing so, and most especially whenever the server is satisfying the read of the object's content from its cache, the server MAY cache access time



updates and lazily write them to stable storage. It is also acceptable to give administrators of the server the option to disable time\_access updates.

5.8.2.34. Attribute 48: time\_access\_set

Sets the time of last access to the object. SETATTR use only.

5.8.2.35. Attribute 49: time\_backup

The time of last backup of the object.

5.8.2.36. Attribute 50: time\_create

The time of creation of the object. This attribute does not have any relation to the traditional UNIX file attribute "ctime" or "change time".

5.8.2.37. Attribute 51: time\_delta

Smallest useful server time granularity.

5.8.2.38. Attribute 52: time\_metadata

The time of last metadata modification of the object.

5.8.2.39. Attribute 53: time\_modify

The time of last modification to the object.

5.8.2.40. Attribute 54: time\_modify\_set

Sets the time of last modification to the object. SETATTR use only.

5.9. Interpreting owner and owner\_group

The RECOMMENDED attributes "owner" and "owner\_group" (and also users and groups within the "acl" attribute) are represented in terms of a UTF-8 string. To avoid a representation that is tied to a particular underlying implementation at the client or server, the use of the UTF-8 string has been chosen. Note that section 6.1 of RFC 2624 [RFC2624] provides additional rationale. It is expected that the client and server will have their own local representation of owner and owner\_group that is used for local storage or presentation to the end user. Therefore, it is expected that when these attributes are transferred between the client and server, the local representation is translated to a syntax of the form "user@dns\_domain". This will allow for a client and server that do not use the same local

representation the ability to translate to a common syntax that can be interpreted by both.

Similarly, security principals may be represented in different ways by different security mechanisms. Servers normally translate these representations into a common format, generally that used by local storage, to serve as a means of identifying the users corresponding to these security principals. When these local identifiers are translated to the form of the owner attribute, associated with files created by such principals, they identify, in a common format, the users associated with each corresponding set of security principals.

The translation used to interpret owner and group strings is not specified as part of the protocol. This allows various solutions to be employed. For example, a local translation table may be consulted that maps a numeric identifier to the `user@dns_domain` syntax. A name service may also be used to accomplish the translation. A server may provide a more general service, not limited by any particular translation (which would only translate a limited set of possible strings) by storing the owner and owner\_group attributes in local storage without any translation or it may augment a translation method by storing the entire string for attributes for which no translation is available while using the local representation for those cases in which a translation is available.

Servers that do not provide support for all possible values of the owner and owner\_group attributes SHOULD return an error (`NFS4ERR_BADOWNER`) when a string is presented that has no translation, as the value to be set for a SETATTR of the owner, owner\_group, or acl attributes. When a server does accept an owner or owner\_group value as valid on a SETATTR (and similarly for the owner and group strings in an acl), it is promising to return that same string (for which see below) when a corresponding GETATTR is done. For some internationalization-related exceptions where this is not possible, see below. Configuration changes (including changes from the mapping of the string to the local representation) and ill-constructed name translations (those that contain aliasing) may make that promise impossible to honor. Servers should make appropriate efforts to avoid a situation in which these attributes have their values changed when no real change to ownership has occurred.

The "dns\_domain" portion of the owner string is meant to be a DNS domain name. For example, `user@example.org`. Servers should accept as valid a set of users for at least one domain. A server may treat other domains as having no valid translations. A more general service is provided when a server is capable of accepting users for multiple domains, or for all domains, subject to security constraints.

As an implementation guide, both clients and servers may provide a means to configure the "dns\_domain" portion of the owner string. For example, the DNS domain name might be "lab.example.org", but the user names are defined in "example.org". In the absence of such a configuration, or as a default, the current DNS domain name of the server should be the value used for the "dns\_domain".

As mentioned above, it is desirable that a server when accepting a string of the form user@domain or group@domain in an attribute, return this same string when that corresponding attribute is fetched. Internationalization issues (for a general discussion of which see Section 12) may make this impossible and the client needs to take note of the following situations:

- o The string representing the domain may be converted to equivalent U-label (see [RFC5890]), if presented using a form other than a U-label. See Section 12.4 for details.
- o The user or group may be returned in a different form, due to normalization issues, although it will always be a canonically equivalent string.

In the case where there is no translation available to the client or server, the attribute value will be constructed without the "@". Therefore, the absence of the "@" from the owner or owner\_group attribute signifies that no translation was available at the sender and that the receiver of the attribute should not use that string as a basis for translation into its own internal format. Even though the attribute value cannot be translated, it may still be useful. In the case of a client, the attribute string may be used for local display of ownership.

To provide a greater degree of compatibility with NFSv3, which identified users and groups by 32-bit unsigned user identifiers and group identifiers, owner and group strings that consist of ASCII-encoded decimal numeric values with no leading zeros can be given a special interpretation by clients and servers that choose to provide such support. The receiver may treat such a user or group string as representing the same user as would be represented by an NFSv3 uid or gid having the corresponding numeric value.

A server SHOULD reject such a numeric value if the security mechanism is kerberized. I.e., in such a scenario, the client will already need to form "user@domain" strings. For any other security mechanism, the server SHOULD accept such numeric values. As an implementation note, the server could make such an acceptance be configurable. If the server does not support numeric values or if it is configured off, then it MUST return an NFS4ERR\_BADOWNER error. If

the security mechanism is kerberized and the client attempts to use the special form, then the server SHOULD return an NFS4ERR\_BADOWNER error when there is a valid translation for the user or owner designated in this way. In that case, the client must use the appropriate user@domain string and not the special form for compatibility.

The client MUST always accept numeric values if the security mechanism is not RPCSEC\_GSS. A client can determine if a server supports numeric identifiers by first attempting to provide a numeric identifier. If this attempt rejected with an NFS4ERR\_BADOWNER error, the the client should only use named identifiers of the form "user@dns\_domain".

The owner string "nobody" may be used to designate an anonymous user, which will be associated with a file created by a security principal that cannot be mapped through normal means to the owner attribute.

#### 5.10. Character Case Attributes

With respect to the case\_insensitive and case\_preserving attributes, each Universal Multiple-octet coded Character Set-4 (UCS-4) [ISO.10646-1.1993] character (which UTF-8 encodes) has a "long descriptive name" RFC1345 [RFC1345] which may or may not include the word "CAPITAL" or "SMALL". The presence of SMALL or CAPITAL allows an NFS server to implement unambiguous and efficient table driven mappings for case insensitive comparisons, and non-case-preserving storage, although there are variations that occur additional characters with a name including "SMALL" or "CAPITAL" are added in a subsequent version of Unicode.

### 6. Access Control Attributes

Access Control Lists (ACLs) are file attributes that specify fine grained access control. This chapter covers the "acl", "aclsupport", "mode", file attributes, and their interactions. Note that file attributes may apply to any file system object.

#### 6.1. Goals

ACLs and modes represent two well established models for specifying permissions. This chapter specifies requirements that attempt to meet the following goals:

- o If a server supports the mode attribute, it should provide reasonable semantics to clients that only set and retrieve the mode attribute.

- o If a server supports ACL attributes, it should provide reasonable semantics to clients that only set and retrieve those attributes.
- o On servers that support the mode attribute, if ACL attributes have never been set on an object, via inheritance or explicitly, the behavior should be traditional UNIX-like behavior.
- o On servers that support the mode attribute, if the ACL attributes have been previously set on an object, either explicitly or via inheritance:
  - \* Setting only the mode attribute should effectively control the traditional UNIX-like permissions of read, write, and execute on owner, owner\_group, and other.
  - \* Setting only the mode attribute should provide reasonable security. For example, setting a mode of 000 should be enough to ensure that future opens for read or write by any principal fail, regardless of a previously existing or inherited ACL.
- o When a mode attribute is set on an object, the ACL attributes may need to be modified so as to not conflict with the new mode. In such cases, it is desirable that the ACL keep as much information as possible. This includes information about inheritance, AUDIT and ALARM ACEs, and permissions granted and denied that do not conflict with the new mode.

## 6.2. File Attributes Discussion

### 6.2.1. Attribute 12: acl

The NFSv4.0 ACL attribute contains an array of access control entries (ACEs) that are associated with the file system object. Although the client can read and write the acl attribute, the server is responsible for using the ACL to perform access control. The client can use the OPEN or ACCESS operations to check access without modifying or reading data or metadata.

The NFS ACE structure is defined as follows:

```
typedef uint32_t      acetype4;
```

```
typedef uint32_t      aceflag4;
```

```
typedef uint32_t      acemask4;
```

```
struct nfsace4 {  
    acetype4          type;  
    aceflag4          flag;  
    acemask4          access_mask;  
    utf8str_mixed     who;  
};
```

To determine if a request succeeds, the server processes each `nfsace4` entry in order. Only ACEs which have a "who" that matches the requester are considered. Each ACE is processed until all of the bits of the requester's access have been ALLOWED. Once a bit (see below) has been ALLOWED by an `ACCESS_ALLOWED_ACE`, it is no longer considered in the processing of later ACEs. If an `ACCESS_DENIED_ACE` is encountered where the requester's access still has unALLOWED bits in common with the "access\_mask" of the ACE, the request is denied. When the ACL is fully processed, if there are bits in the requester's mask that have not been ALLOWED or DENIED, access is denied.

Unlike the ALLOW and DENY ACE types, the ALARM and AUDIT ACE types do not affect a requester's access, and instead are for triggering events as a result of a requester's access attempt. Therefore, AUDIT and ALARM ACEs are processed only after processing ALLOW and DENY ACEs.

The NFSv4.0 ACL model is quite rich. Some server platforms may provide access control functionality that goes beyond the UNIX-style mode attribute, but which is not as rich as the NFS ACL model. So that users can take advantage of this more limited functionality, the server may support the `acl` attributes by mapping between its ACL model and the NFSv4.0 ACL model. Servers must ensure that the ACL they actually store or enforce is at least as strict as the NFSv4 ACL that was set. It is tempting to accomplish this by rejecting any ACL that falls outside the small set that can be represented accurately. However, such an approach can render ACLs unusable without special client-side knowledge of the server's mapping, which defeats the purpose of having a common NFSv4 ACL protocol. Therefore servers should accept every ACL that they can without compromising security. To help accomplish this, servers may make a special exception, in the case of unsupported permission bits, to the rule that bits not ALLOWED or DENIED by an ACL must be denied. For example, a UNIX-style server might choose to silently allow read attribute permissions even though an ACL does not explicitly allow those permissions. (An ACL that explicitly denies permission to read attributes should still be rejected.)

The situation is complicated by the fact that a server may have multiple modules that enforce ACLs. For example, the enforcement for NFSv4.0 access may be different from, but not weaker than, the

enforcement for local access, and both may be different from the enforcement for access through other protocols such as Server Message Block (SMB). So it may be useful for a server to accept an ACL even if not all of its modules are able to support it.

The guiding principle with regard to NFSv4 access is that the server must not accept ACLs that appear to make access to the file more restrictive than it really is.

#### 6.2.1.1. ACE Type

The constants used for the type field (acetype4) are as follows:

```
const ACE4_ACCESS_ALLOWED_ACE_TYPE      = 0x00000000;
const ACE4_ACCESS_DENIED_ACE_TYPE       = 0x00000001;
const ACE4_SYSTEM_AUDIT_ACE_TYPE        = 0x00000002;
const ACE4_SYSTEM_ALARM_ACE_TYPE        = 0x00000003;
```

All four bit types are permitted in the acl attribute.

Value	Abbreviation	Description
ACE4_ACCESS_ALLOWED_ACE_TYPE	ALLOW	Explicitly grants the access defined in acemask4 to the file or directory.
ACE4_ACCESS_DENIED_ACE_TYPE	DENY	Explicitly denies the access defined in acemask4 to the file or directory.
ACE4_SYSTEM_AUDIT_ACE_TYPE	AUDIT	LOG (in a system dependent way) any access attempt to a file or directory which uses any of the access methods specified in acemask4.

ACE4_SYSTEM_ALARM_ACE_TYPE	ALARM	Generate a system ALARM (system dependent) when any access attempt is made to a file or directory for the access methods specified in <code>acemask4</code> .
----------------------------	-------	---

The "Abbreviation" column denotes how the types will be referred to throughout the rest of this chapter.

#### 6.2.1.2. Attribute 13: `aclsupport`

A server need not support all of the above ACE types. This attribute indicates which ACE types are supported for the current file system. The bitmask constants used to represent the above definitions within the `aclsupport` attribute are as follows:

```
const ACL4_SUPPORT_ALLOW_ACL      = 0x00000001;
const ACL4_SUPPORT_DENY_ACL      = 0x00000002;
const ACL4_SUPPORT_AUDIT_ACL     = 0x00000004;
const ACL4_SUPPORT_ALARM_ACL     = 0x00000008;
```

Servers which support either the ALLOW or DENY ACE type SHOULD support both ALLOW and DENY ACE types.

Clients should not attempt to set an ACE unless the server claims support for that ACE type. If the server receives a request to set an ACE that it cannot store, it MUST reject the request with `NFS4ERR_ATTRNOTSUPP`. If the server receives a request to set an ACE that it can store but cannot enforce, the server SHOULD reject the request with `NFS4ERR_ATTRNOTSUPP`.

Support for any of the ACL attributes is optional (albeit, RECOMMENDED).

#### 6.2.1.3. ACE Access Mask

The bitmask constants used for the access mask field are as follows:



```
const ACE4_READ_DATA           = 0x00000001;
const ACE4_LIST_DIRECTORY      = 0x00000001;
const ACE4_WRITE_DATA          = 0x00000002;
const ACE4_ADD_FILE            = 0x00000002;
const ACE4_APPEND_DATA         = 0x00000004;
const ACE4_ADD_SUBDIRECTORY    = 0x00000004;
const ACE4_READ_NAMED_ATTRS    = 0x00000008;
const ACE4_WRITE_NAMED_ATTRS   = 0x00000010;
const ACE4_EXECUTE             = 0x00000020;
const ACE4_DELETE_CHILD        = 0x00000040;
const ACE4_READ_ATTRIBUTES     = 0x00000080;
const ACE4_WRITE_ATTRIBUTES    = 0x00000100;

const ACE4_DELETE              = 0x00010000;
const ACE4_READ_ACL            = 0x00020000;
const ACE4_WRITE_ACL           = 0x00040000;
const ACE4_WRITE_OWNER         = 0x00080000;
const ACE4_SYNCHRONIZE          = 0x00100000;
```

Note that some masks have coincident values, for example, ACE4\_READ\_DATA and ACE4\_LIST\_DIRECTORY. The mask entries ACE4\_LIST\_DIRECTORY, ACE4\_ADD\_FILE, and ACE4\_ADD\_SUBDIRECTORY are intended to be used with directory objects, while ACE4\_READ\_DATA, ACE4\_WRITE\_DATA, and ACE4\_APPEND\_DATA are intended to be used with non-directory objects.

#### 6.2.1.3.1. Discussion of Mask Attributes

##### ACE4\_READ\_DATA

Operation(s) affected:

READ

OPEN

Discussion:

Permission to read the data of the file.

Servers SHOULD allow a user the ability to read the data of the file when only the ACE4\_EXECUTE access mask bit is allowed.

##### ACE4\_LIST\_DIRECTORY

Operation(s) affected:

READDIR

Discussion:

Permission to list the contents of a directory.

#### ACE4\_WRITE\_DATA

Operation(s) affected:

WRITE

OPEN

SETATTR of size

Discussion:

Permission to modify a file's data.

#### ACE4\_ADD\_FILE

Operation(s) affected:

CREATE

LINK

OPEN

RENAME

Discussion:

Permission to add a new file in a directory. The CREATE operation is affected when `nfs_ftype4` is `NF4LNK`, `NF4BLK`, `NF4CHR`, `NF4SOCK`, or `NF4FIFO`. (`NF4DIR` is not listed because it is covered by `ACE4_ADD_SUBDIRECTORY`.) OPEN is affected when used to create a regular file. LINK and RENAME are always affected.

#### ACE4\_APPEND\_DATA

Operation(s) affected:

WRITE

OPEN

SETATTR of size

Discussion:

The ability to modify a file's data, but only starting at EOF. This allows for the notion of append-only files, by allowing ACE4\_APPEND\_DATA and denying ACE4\_WRITE\_DATA to the same user or group. If a file has an ACL such as the one described above and a WRITE request is made for somewhere other than EOF, the server SHOULD return NFS4ERR\_ACCESS.

ACE4\_ADD\_SUBDIRECTORY

Operation(s) affected:

CREATE

RENAME

Discussion:

Permission to create a subdirectory in a directory. The CREATE operation is affected when nfs\_ftype4 is NF4DIR. The RENAME operation is always affected.

ACE4\_READ\_NAMED\_ATTRS

Operation(s) affected:

OPENATTR

Discussion:

Permission to read the named attributes of a file or to lookup the named attributes directory. OPENATTR is affected when it is not used to create a named attribute directory. This is when 1.) createdir is TRUE, but a named attribute directory already exists, or 2.) createdir is FALSE.

## ACE4\_WRITE\_NAMED\_ATTRS

Operation(s) affected:

OPENATTR

Discussion:

Permission to write the named attributes of a file or to create a named attribute directory. OPENATTR is affected when it is used to create a named attribute directory. This is when createdir is TRUE and no named attribute directory exists. The ability to check whether or not a named attribute directory exists depends on the ability to look it up, therefore, users also need the ACE4\_READ\_NAMED\_ATTRS permission in order to create a named attribute directory.

## ACE4\_EXECUTE

Operation(s) affected:

READ

Discussion:

Permission to execute a file.

Servers SHOULD allow a user the ability to read the data of the file when only the ACE4\_EXECUTE access mask bit is allowed. This is because there is no way to execute a file without reading the contents. Though a server may treat ACE4\_EXECUTE and ACE4\_READ\_DATA bits identically when deciding to permit a READ operation, it SHOULD still allow the two bits to be set independently in ACLs, and MUST distinguish between them when replying to ACCESS operations. In particular, servers SHOULD NOT silently turn on one of the two bits when the other is set, as that would make it impossible for the client to correctly enforce the distinction between read and execute permissions.

As an example, following a SETATTR of the following ACL:

nfsuser:ACE4\_EXECUTE:ALLOW

A subsequent GETATTR of ACL for that file SHOULD return:

nfsuser:ACE4\_EXECUTE:ALLOW

Rather than:

nfsuser:ACE4\_EXECUTE/ACE4\_READ\_DATA:ALLOW

ACE4\_EXECUTE

Operation(s) affected:

LOOKUP

OPEN

REMOVE

RENAME

LINK

CREATE

Discussion:

Permission to traverse/search a directory.

ACE4\_DELETE\_CHILD

Operation(s) affected:

REMOVE

RENAME

Discussion:

Permission to delete a file or directory within a directory.  
See Section 6.2.1.3.2 for information on how ACE4\_DELETE and  
ACE4\_DELETE\_CHILD interact.

ACE4\_READ\_ATTRIBUTES

Operation(s) affected:

GETATTR of file system object attributes

VERIFY

NVERIFY

READDIR

Discussion:

The ability to read basic attributes (non-ACLs) of a file. On a UNIX system, basic attributes can be thought of as the stat level attributes. Allowing this access mask bit would mean the entity can execute "ls -l" and stat. If a READDIR operation requests attributes, this mask must be allowed for the READDIR to succeed.

ACE4\_WRITE\_ATTRIBUTES

Operation(s) affected:

SETATTR of time\_access\_set, time\_backup,  
time\_create, time\_modify\_set, mimetype, hidden, system

Discussion:

Permission to change the times associated with a file or directory to an arbitrary value. Also permission to change the mimetype, hidden and system attributes. A user having ACE4\_WRITE\_DATA or ACE4\_WRITE\_ATTRIBUTES will be allowed to set the times associated with a file to the current server time.

ACE4\_DELETE

Operation(s) affected:

REMOVE

Discussion:

Permission to delete the file or directory. See Section 6.2.1.3.2 for information on ACE4\_DELETE and ACE4\_DELETE\_CHILD interact.

ACE4\_READ\_ACL

Operation(s) affected:

GETATTR of acl

NVERIFY

VERIFY

Discussion:

Permission to read the ACL.

ACE4\_WRITE\_ACL

Operation(s) affected:

SETATTR of acl and mode

Discussion:

Permission to write the acl and mode attributes.

ACE4\_WRITE\_OWNER

Operation(s) affected:

SETATTR of owner and owner\_group

Discussion:

Permission to write the owner and owner\_group attributes. On UNIX systems, this is the ability to execute chown() and chgrp().

ACE4\_SYNCHRONIZE

Operation(s) affected:

NONE

Discussion:

Permission to use the file object as a synchronization primitive for interprocess communication. This permission is not enforced or interpreted by the NFSv4.0 server on behalf of the client.

Typically, the ACE4\_SYNCHRONIZE permission is only meaningful on local file systems, i.e., file systems not accessed via NFSv4.0. The reason that the permission bit exists is that

some operating environments, such as Windows, use ACE4\_SYNCHRONIZE.

For example, if a client copies a file that has ACE4\_SYNCHRONIZE set from a local file system to an NFSv4.0 server, and then later copies the file from the NFSv4.0 server to a local file system, it is likely that if ACE4\_SYNCHRONIZE was set in the original file, the client will want it set in the second copy. The first copy will not have the permission set unless the NFSv4.0 server has the means to set the ACE4\_SYNCHRONIZE bit. The second copy will not have the permission set unless the NFSv4.0 server has the means to retrieve the ACE4\_SYNCHRONIZE bit.

Server implementations need not provide the granularity of control that is implied by this list of masks. For example, POSIX-based systems might not distinguish ACE4\_APPEND\_DATA (the ability to append to a file) from ACE4\_WRITE\_DATA (the ability to modify existing contents); both masks would be tied to a single "write" permission. When such a server returns attributes to the client, it would show both ACE4\_APPEND\_DATA and ACE4\_WRITE\_DATA if and only if the write permission is enabled.

If a server receives a SETATTR request that it cannot accurately implement, it should err in the direction of more restricted access, except in the previously discussed cases of execute and read. For example, suppose a server cannot distinguish overwriting data from appending new data, as described in the previous paragraph. If a client submits an ALLOW ACE where ACE4\_APPEND\_DATA is set but ACE4\_WRITE\_DATA is not (or vice versa), the server should either turn off ACE4\_APPEND\_DATA or reject the request with NFS4ERR\_ATTRNOTSUPP.

#### 6.2.1.3.2. ACE4\_DELETE vs. ACE4\_DELETE\_CHILD

Two access mask bits govern the ability to delete a directory entry: ACE4\_DELETE on the object itself (the "target"), and ACE4\_DELETE\_CHILD on the containing directory (the "parent").

Many systems also take the "sticky bit" (MODE4\_SVTX) on a directory to allow unlink only to a user that owns either the target or the parent; on some such systems the decision also depends on whether the target is writable.

Servers SHOULD allow unlink if either ACE4\_DELETE is permitted on the target, or ACE4\_DELETE\_CHILD is permitted on the parent. (Note that this is true even if the parent or target explicitly denies one of these permissions.)



If the ACLs in question neither explicitly ALLOW nor DENY either of the above, and if MODE4\_SVTX is not set on the parent, then the server SHOULD allow the removal if and only if ACE4\_ADD\_FILE is permitted. In the case where MODE4\_SVTX is set, the server may also require the remover to own either the parent or the target, or may require the target to be writable.

This allows servers to support something close to traditional UNIX-like semantics, with ACE4\_ADD\_FILE taking the place of the write bit.

#### 6.2.1.4. ACE flag

The bitmask constants used for the flag field are as follows:

```
const ACE4_FILE_INHERIT_ACE           = 0x000000001;
const ACE4_DIRECTORY_INHERIT_ACE      = 0x000000002;
const ACE4_NO_PROPAGATE_INHERIT_ACE   = 0x000000004;
const ACE4_INHERIT_ONLY_ACE           = 0x000000008;
const ACE4_SUCCESSFUL_ACCESS_ACE_FLAG = 0x000000010;
const ACE4_FAILED_ACCESS_ACE_FLAG     = 0x000000020;
const ACE4_IDENTIFIER_GROUP           = 0x000000040;
```

A server need not support any of these flags. If the server supports flags that are similar to, but not exactly the same as, these flags, the implementation may define a mapping between the protocol-defined flags and the implementation-defined flags.

For example, suppose a client tries to set an ACE with ACE4\_FILE\_INHERIT\_ACE set but not ACE4\_DIRECTORY\_INHERIT\_ACE. If the server does not support any form of ACL inheritance, the server should reject the request with NFS4ERR\_ATTRNOTSUPP. If the server supports a single "inherit ACE" flag that applies to both files and directories, the server may reject the request (i.e., requiring the client to set both the file and directory inheritance flags). The server may also accept the request and silently turn on the ACE4\_DIRECTORY\_INHERIT\_ACE flag.

##### 6.2.1.4.1. Discussion of Flag Bits

###### ACE4\_FILE\_INHERIT\_ACE

Any non-directory file in any sub-directory will get this ACE inherited.

###### ACE4\_DIRECTORY\_INHERIT\_ACE

Can be placed on a directory and indicates that this ACE should be added to each new directory created.

If this flag is set in an ACE in an ACL attribute to be set on a non-directory file system object, the operation attempting to set

the ACL SHOULD fail with NFS4ERR\_ATTRNOTSUPP.

#### ACE4\_INHERIT\_ONLY\_ACE

Can be placed on a directory but does not apply to the directory; ALLOW and DENY ACEs with this bit set do not affect access to the directory, and AUDIT and ALARM ACEs with this bit set do not trigger log or alarm events. Such ACEs only take effect once they are applied (with this bit cleared) to newly created files and directories as specified by the above two flags.

If this flag is present on an ACE, but neither ACE4\_DIRECTORY\_INHERIT\_ACE nor ACE4\_FILE\_INHERIT\_ACE is present, then an operation attempting to set such an attribute SHOULD fail with NFS4ERR\_ATTRNOTSUPP.

#### ACE4\_NO\_PROPAGATE\_INHERIT\_ACE

Can be placed on a directory. This flag tells the server that inheritance of this ACE should stop at newly created child directories.

#### ACE4\_SUCCESSFUL\_ACCESS\_ACE\_FLAG

#### ACE4\_FAILED\_ACCESS\_ACE\_FLAG

The ACE4\_SUCCESSFUL\_ACCESS\_ACE\_FLAG (SUCCESS) and ACE4\_FAILED\_ACCESS\_ACE\_FLAG (FAILED) flag bits may be set only on ACE4\_SYSTEM\_AUDIT\_ACE\_TYPE (AUDIT) and ACE4\_SYSTEM\_ALARM\_ACE\_TYPE (ALARM) ACE types. If during the processing of the file's ACL, the server encounters an AUDIT or ALARM ACE that matches the principal attempting the OPEN, the server notes that fact, and the presence, if any, of the SUCCESS and FAILED flags encountered in the AUDIT or ALARM ACE. Once the server completes the ACL processing, it then notes if the operation succeeded or failed. If the operation succeeded, and if the SUCCESS flag was set for a matching AUDIT or ALARM ACE, then the appropriate AUDIT or ALARM event occurs. If the operation failed, and if the FAILED flag was set for the matching AUDIT or ALARM ACE, then the appropriate AUDIT or ALARM event occurs. Either or both of the SUCCESS or FAILED can be set, but if neither is set, the AUDIT or ALARM ACE is not useful.

The previously described processing applies to ACCESS operations even when they return NFS4\_OK. For the purposes of AUDIT and ALARM, we consider an ACCESS operation to be a "failure" if it fails to return a bit that was requested and supported.

#### ACE4\_IDENTIFIER\_GROUP

Indicates that the "who" refers to a GROUP as defined under UNIX or a GROUP ACCOUNT as defined under Windows. Clients and servers MUST ignore the ACE4\_IDENTIFIER\_GROUP flag on ACEs with a who

value equal to one of the special identifiers outlined in Section 6.2.1.5.

#### 6.2.1.5. ACE Who

The "who" field of an ACE is an identifier that specifies the principal or principals to whom the ACE applies. It may refer to a user or a group, with the flag bit `ACE4_IDENTIFIER_GROUP` specifying which.

There are several special identifiers which need to be understood universally, rather than in the context of a particular DNS domain. Some of these identifiers cannot be understood when an NFS client accesses the server, but have meaning when a local process accesses the file. The ability to display and modify these permissions is permitted over NFS, even if none of the access methods on the server understands the identifiers.

Who	Description
OWNER	The owner of the file
GROUP	The group associated with the file.
EVERYONE	The world, including the owner and owning group.
INTERACTIVE	Accessed from an interactive terminal.
NETWORK	Accessed via the network.
DIALUP	Accessed as a dialup user to the server.
BATCH	Accessed from a batch job.
ANONYMOUS	Accessed without any authentication.
AUTHENTICATED	Any authenticated user (opposite of ANONYMOUS)
SERVICE	Access from a system service.

Table 4

To avoid conflict, these special identifiers are distinguished by an appended "@" and should appear in the form "xxxx@" (with no domain name after the "@"). For example: ANONYMOUS@.

The `ACE4_IDENTIFIER_GROUP` flag MUST be ignored on entries with these special identifiers. When encoding entries with these special identifiers, the `ACE4_IDENTIFIER_GROUP` flag SHOULD be set to zero.

##### 6.2.1.5.1. Discussion of EVERYONE@

It is important to note that "EVERYONE@" is not equivalent to the UNIX "other" entity. This is because, by definition, UNIX "other" does not include the owner or owning group of a file. "EVERYONE@"

means literally everyone, including the owner or owning group.

#### 6.2.2. Attribute 33: mode

The NFSv4.0 mode attribute is based on the UNIX mode bits. The following bits are defined:

```
const MODE4_SUID = 0x800; /* set user id on execution */
const MODE4_SGID = 0x400; /* set group id on execution */
const MODE4_SVTX = 0x200; /* save text even after use */
const MODE4_RUSR = 0x100; /* read permission: owner */
const MODE4_WUSR = 0x080; /* write permission: owner */
const MODE4_XUSR = 0x040; /* execute permission: owner */
const MODE4_RGRP = 0x020; /* read permission: group */
const MODE4_WGRP = 0x010; /* write permission: group */
const MODE4_XGRP = 0x008; /* execute permission: group */
const MODE4_OTH = 0x004; /* read permission: other */
const MODE4_WOTH = 0x002; /* write permission: other */
const MODE4_XOTH = 0x001; /* execute permission: other */
```

Bits MODE4\_RUSR, MODE4\_WUSR, and MODE4\_XUSR apply to the principal identified in the owner attribute. Bits MODE4\_RGRP, MODE4\_WGRP, and MODE4\_XGRP apply to principals identified in the owner\_group attribute but who are not identified in the owner attribute. Bits MODE4\_OTH, MODE4\_WOTH, MODE4\_XOTH apply to any principal that does not match that in the owner attribute, and does not have a group matching that of the owner\_group attribute.

Bits within the mode other than those specified above are not defined by this protocol. A server MUST NOT return bits other than those defined above in a GETATTR or REaddir operation, and it MUST return NFS4ERR\_INVAL if bits other than those defined above are set in a SETATTR, CREATE, OPEN, VERIFY or NVERIFY operation.

#### 6.3. Common Methods

The requirements in this section will be referred to in future sections, especially Section 6.4.

##### 6.3.1. Interpreting an ACL

###### 6.3.1.1. Server Considerations

The server uses the algorithm described in Section 6.2.1 to determine whether an ACL allows access to an object. However, the ACL may not be the sole determiner of access. For example:

- o In the case of a file system exported as read-only, the server may deny write permissions even though an object's ACL grants it.
- o Server implementations MAY grant ACE4\_WRITE\_ACL and ACE4\_READ\_ACL permissions to prevent a situation from arising in which there is no valid way to ever modify the ACL.
- o All servers will allow a user the ability to read the data of the file when only the execute permission is granted (i.e., If the ACL denies the user the ACE4\_READ\_DATA access and allows the user ACE4\_EXECUTE, the server will allow the user to read the data of the file).
- o Many servers have the notion of owner-override in which the owner of the object is allowed to override accesses that are denied by the ACL. This may be helpful, for example, to allow users continued access to open files on which the permissions have changed.
- o Many servers have the notion of a "superuser" that has privileges beyond an ordinary user. The superuser may be able to read or write data or metadata in ways that would not be permitted by the ACL.

#### 6.3.1.2. Client Considerations

Clients SHOULD NOT do their own access checks based on their interpretation the ACL, but rather use the OPEN and ACCESS operations to do access checks. This allows the client to act on the results of having the server determine whether or not access should be granted based on its interpretation of the ACL.

Clients must be aware of situations in which an object's ACL will define a certain access even though the server will not enforce it. In general, but especially in these situations, the client needs to do its part in the enforcement of access as defined by the ACL. To do this, the client MAY send the appropriate ACCESS operation prior to servicing the request of the user or application in order to determine whether the user or application should be granted the access requested. For examples in which the ACL may define accesses that the server doesn't enforce see Section 6.3.1.1.

#### 6.3.2. Computing a Mode Attribute from an ACL

The following method can be used to calculate the MODE4\_R\*, MODE4\_W\* and MODE4\_X\* bits of a mode attribute, based upon an ACL.

First, for each of the special identifiers OWNER@, GROUP@, and

EVERYONE@, evaluate the ACL in order, considering only ALLOW and DENY ACEs for the identifier EVERYONE@ and for the identifier under consideration. The result of the evaluation will be an NFSv4 ACL mask showing exactly which bits are permitted to that identifier.

Then translate the calculated mask for OWNER@, GROUP@, and EVERYONE@ into mode bits for, respectively, the user, group, and other, as follows:

1. Set the read bit (MODE4\_RUSR, MODE4\_RGRP, or MODE4\_OTH) if and only if ACE4\_READ\_DATA is set in the corresponding mask.
2. Set the write bit (MODE4\_WUSR, MODE4\_WGRP, or MODE4\_WOTH) if and only if ACE4\_WRITE\_DATA and ACE4\_APPEND\_DATA are both set in the corresponding mask.
3. Set the execute bit (MODE4\_XUSR, MODE4\_XGRP, or MODE4\_XOTH), if and only if ACE4\_EXECUTE is set in the corresponding mask.

#### 6.3.2.1. Discussion

Some server implementations also add bits permitted to named users and groups to the group bits (MODE4\_RGRP, MODE4\_WGRP, and MODE4\_XGRP).

Implementations are discouraged from doing this, because it has been found to cause confusion for users who see members of a file's group denied access that the mode bits appear to allow. (The presence of DENY ACEs may also lead to such behavior, but DENY ACEs are expected to be more rarely used.)

The same user confusion seen when fetching the mode also results if setting the mode does not effectively control permissions for the owner, group, and other users; this motivates some of the requirements that follow.

#### 6.4. Requirements

The server that supports both mode and ACL must take care to synchronize the MODE4\_\*USR, MODE4\_\*GRP, and MODE4\_\*OTH bits with the ACEs which have respective who fields of "OWNER@", "GROUP@", and "EVERYONE@" so that the client can see semantically equivalent access permissions exist whether the client asks for owner, owner\_group and mode attributes, or for just the ACL.

Many requirements refer to Section 6.3.2, but note that the methods have behaviors specified with "SHOULD". This is intentional, to avoid invalidating existing implementations that compute the mode

according to the withdrawn POSIX ACL draft ([P1003.1e]), rather than by actual permissions on owner, group, and other.

#### 6.4.1. Setting the mode and/or ACL Attributes

##### 6.4.1.1. Setting mode and not ACL

When any of the nine low-order mode bits are changed because the mode attribute was set, and no ACL attribute is explicitly set, the `acl` attribute must be modified in accordance with the updated value of those bits. This must happen even if the value of the low-order bits is the same after the mode is set as before.

Note that any `AUDIT` or `ALARM` ACEs are unaffected by changes to the mode.

In cases in which the permissions bits are subject to change, the `acl` attribute **MUST** be modified such that the mode computed via the method in Section 6.3.2 yields the low-order nine bits (`MODE4_R*`, `MODE4_W*`, `MODE4_X*`) of the mode attribute as modified by the attribute change. The ACL attributes **SHOULD** also be modified such that:

1. If `MODE4_RGRP` is not set, entities explicitly listed in the ACL other than `OWNER@` and `EVERYONE@` **SHOULD NOT** be granted `ACE4_READ_DATA`.
2. If `MODE4_WGRP` is not set, entities explicitly listed in the ACL other than `OWNER@` and `EVERYONE@` **SHOULD NOT** be granted `ACE4_WRITE_DATA` or `ACE4_APPEND_DATA`.
3. If `MODE4_XGRP` is not set, entities explicitly listed in the ACL other than `OWNER@` and `EVERYONE@` **SHOULD NOT** be granted `ACE4_EXECUTE`.

Access mask bits other than those listed above, appearing in `ALLOW` ACEs, **MAY** also be disabled.

Note that ACEs with the flag `ACE4_INHERIT_ONLY_ACE` set do not affect the permissions of the ACL itself, nor do ACEs of the type `AUDIT` and `ALARM`. As such, it is desirable to leave these ACEs unmodified when modifying the ACL attributes.

Also note that the requirement may be met by discarding the `acl` in favor of an ACL that represents the mode and only the mode. This is permitted, but it is preferable for a server to preserve as much of the ACL as possible without violating the above requirements. Discarding the ACL makes it effectively impossible for a file created with a mode attribute to inherit an ACL (see Section 6.4.3).

#### 6.4.1.2. Setting ACL and not mode

When setting the acl and not setting the mode attribute, the permission bits of the mode need to be derived from the ACL. In this case, the ACL attribute SHOULD be set as given. The nine low-order bits of the mode attribute (MODE4\_R\*, MODE4\_W\*, MODE4\_X\*) MUST be modified to match the result of the method Section 6.3.2. The three high-order bits of the mode (MODE4\_SUID, MODE4\_SGID, MODE4\_SVTX) SHOULD remain unchanged.

#### 6.4.1.3. Setting both ACL and mode

When setting both the mode and the acl attribute in the same operation, the attributes MUST be applied in this order: mode, then ACL. The mode-related attribute is set as given, then the ACL attribute is set as given, possibly changing the final mode, as described above in Section 6.4.1.2.

#### 6.4.2. Retrieving the mode and/or ACL Attributes

This section applies only to servers that support both the mode and ACL attributes.

Some server implementations may have a concept of "objects without ACLs", meaning that all permissions are granted and denied according to the mode attribute, and that no ACL attribute is stored for that object. If an ACL attribute is requested of such a server, the server SHOULD return an ACL that does not conflict with the mode; that is to say, the ACL returned SHOULD represent the nine low-order bits of the mode attribute (MODE4\_R\*, MODE4\_W\*, MODE4\_X\*) as described in Section 6.3.2.

For other server implementations, the ACL attribute is always present for every object. Such servers SHOULD store at least the three high-order bits of the mode attribute (MODE4\_SUID, MODE4\_SGID, MODE4\_SVTX). The server SHOULD return a mode attribute if one is requested, and the low-order nine bits of the mode (MODE4\_R\*, MODE4\_W\*, MODE4\_X\*) MUST match the result of applying the method in Section 6.3.2 to the ACL attribute.

#### 6.4.3. Creating New Objects

If a server supports any ACL attributes, it may use the ACL attributes on the parent directory to compute an initial ACL attribute for a newly created object. This will be referred to as the inherited ACL within this section. The act of adding one or more ACEs to the inherited ACL that are based upon ACEs in the parent directory's ACL will be referred to as inheriting an ACE within this



section.

In the presence or absence of the mode and ACL attributes, the behavior of CREATE and OPEN SHOULD be:

1. If just the mode is given in the call:

In this case, inheritance SHOULD take place, but the mode MUST be applied to the inherited ACL as described in Section 6.4.1.1, thereby modifying the ACL.

2. If just the ACL is given in the call:

In this case, inheritance SHOULD NOT take place, and the ACL as defined in the CREATE or OPEN will be set without modification, and the mode modified as in Section 6.4.1.2

3. If both mode and ACL are given in the call:

In this case, inheritance SHOULD NOT take place, and both attributes will be set as described in Section 6.4.1.3.

4. If neither mode nor ACL are given in the call:

In the case where an object is being created without any initial attributes at all, e.g., an OPEN operation with an opentype4 of OPEN4\_CREATE and a createmode4 of EXCLUSIVE4, inheritance SHOULD NOT take place. Instead, the server SHOULD set permissions to deny all access to the newly created object. It is expected that the appropriate client will set the desired attributes in a subsequent SETATTR operation, and the server SHOULD allow that operation to succeed, regardless of what permissions the object is created with. For example, an empty ACL denies all permissions, but the server should allow the owner's SETATTR to succeed even though WRITE\_ACL is implicitly denied.

In other cases, inheritance SHOULD take place, and no modifications to the ACL will happen. The mode attribute, if supported, MUST be as computed in Section 6.3.2, with the MODE4\_SUID, MODE4\_SGID and MODE4\_SVTX bits clear. If no inheritable ACEs exist on the parent directory, the rules for creating acl attributes are implementation defined.

#### 6.4.3.1. The Inherited ACL

If the object being created is not a directory, the inherited ACL SHOULD NOT inherit ACEs from the parent directory ACL unless the ACE4\_FILE\_INHERIT\_FLAG is set.

If the object being created is a directory, the inherited ACL should inherit all inheritable ACEs from the parent directory, those that have ACE4\_FILE\_INHERIT\_ACE or ACE4\_DIRECTORY\_INHERIT\_ACE flag set. If the inheritable ACE has ACE4\_FILE\_INHERIT\_ACE set, but ACE4\_DIRECTORY\_INHERIT\_ACE is clear, the inherited ACE on the newly created directory MUST have the ACE4\_INHERIT\_ONLY\_ACE flag set to prevent the directory from being affected by ACEs meant for non-directories.

When a new directory is created, the server MAY split any inherited ACE which is both inheritable and effective (in other words, which has neither ACE4\_INHERIT\_ONLY\_ACE nor ACE4\_NO\_PROPAGATE\_INHERIT\_ACE set), into two ACEs, one with no inheritance flags, and one with ACE4\_INHERIT\_ONLY\_ACE set. This makes it simpler to modify the effective permissions on the directory without modifying the ACE which is to be inherited to the new directory's children.

### 7. NFS Server Name Space

#### 7.1. Server Exports

On a UNIX server the name space describes all the files reachable by pathnames under the root directory or "/". On a Windows NT server the name space constitutes all the files on disks named by mapped disk letters. NFS server administrators rarely make the entire server's file system name space available to NFS clients. More often portions of the name space are made available via an "export" feature. In previous versions of the NFS protocol, the root filehandle for each export is obtained through the MOUNT protocol; the client sends a string that identifies an object in the exported name space and the server returns the root filehandle for it. The MOUNT protocol supports an EXPORTS procedure that will enumerate the server's exports.

#### 7.2. Browsing Exports

The NFSv4 protocol provides a root filehandle that clients can use to obtain filehandles for these exports via a multi-component LOOKUP. A common user experience is to use a graphical user interface (perhaps a file "Open" dialog window) to find a file via progressive browsing through a directory tree. The client must be able to move from one

export to another export via single-component, progressive LOOKUP operations.

This style of browsing is not well supported by the NFSv2 and NFSv3 protocols. The client expects all LOOKUP operations to remain within a single server file system. For example, the device attribute will not change. This prevents a client from taking name space paths that span exports.

An automounter on the client can obtain a snapshot of the server's name space using the EXPORTS procedure of the MOUNT protocol. If it understands the server's pathname syntax, it can create an image of the server's name space on the client. The parts of the name space that are not exported by the server are filled in with a "pseudo file system" that allows the user to browse from one mounted file system to another. There is a drawback to this representation of the server's name space on the client: it is static. If the server administrator adds a new export the client will be unaware of it.

### 7.3. Server Pseudo Filesystem

NFSv4 servers avoid this name space inconsistency by presenting all the exports within the framework of a single server name space. An NFSv4 client uses LOOKUP and REaddir operations to browse seamlessly from one export to another. Portions of the server name space that are not exported are bridged via a "pseudo file system" that provides a view of exported directories only. A pseudo file system has a unique fsid and behaves like a normal, read only file system.

Based on the construction of the server's name space, it is possible that multiple pseudo file systems may exist. For example,

```
/a      pseudo file system
/a/b    real file system
/a/b/c  pseudo file system
/a/b/c/d real file system
```

Each of the pseudo file systems are considered separate entities and therefore will have a unique fsid.

### 7.4. Multiple Roots

The DOS and Windows operating environments are sometimes described as having "multiple roots". Filesystems are commonly represented as disk letters. MacOS represents file systems as top level names. NFSv4 servers for these platforms can construct a pseudo file system above these root names so that disk letters or volume names are simply directory names in the pseudo root.

### 7.5. Filehandle Volatility

The nature of the server's pseudo file system is that it is a logical representation of file system(s) available from the server. Therefore, the pseudo file system is most likely constructed dynamically when the server is first instantiated. It is expected that the pseudo file system may not have an on disk counterpart from which persistent filehandles could be constructed. Even though it is preferable that the server provide persistent filehandles for the pseudo file system, the NFS client should expect that pseudo file system filehandles are volatile. This can be confirmed by checking the associated "fh\_expire\_type" attribute for those filehandles in question. If the filehandles are volatile, the NFS client must be prepared to recover a filehandle value (e.g., with a multi-component LOOKUP) when receiving an error of NFS4ERR\_FHEXPIRED.

### 7.6. Exported Root

If the server's root file system is exported, one might conclude that a pseudo file system is not needed. This would be wrong. Assume the following file systems on a server:

```
/          disk1  (exported)
/a         disk2  (not exported)
/a/b       disk3  (exported)
```

Because disk2 is not exported, disk3 cannot be reached with simple LOOKUPS. The server must bridge the gap with a pseudo file system.

### 7.7. Mount Point Crossing

The server file system environment may be constructed in such a way that one file system contains a directory which is 'covered' or mounted upon by a second file system. For example:

```
/a/b          (file system 1)
/a/b/c/d      (file system 2)
```

The pseudo file system for this server may be constructed to look like:

```
/              (place holder/not exported)
/a/b           (file system 1)
/a/b/c/d       (file system 2)
```

It is the server's responsibility to present the pseudo file system that is complete to the client. If the client sends a lookup request for the path "/a/b/c/d", the server's response is the filehandle of

the file system `"/a/b/c/d"`. In previous versions of the NFS protocol, the server would respond with the filehandle of directory `"/a/b/c/d"` within the file system `"/a/b"`.

The NFS client will be able to determine if it crosses a server mount point by a change in the value of the `"fsid"` attribute.

#### 7.8. Security Policy and Name Space Presentation

The application of the server's security policy needs to be carefully considered by the implementor. One may choose to limit the viewability of portions of the pseudo file system based on the server's perception of the client's ability to authenticate itself properly. However, with the support of multiple security mechanisms and the ability to negotiate the appropriate use of these mechanisms, the server is unable to properly determine if a client will be able to authenticate itself. If, based on its policies, the server chooses to limit the contents of the pseudo file system, the server may effectively hide file systems from a client that may otherwise have legitimate access.

As suggested practice, the server should apply the security policy of a shared resource in the server's namespace to the components of the resource's ancestors. For example:

```
/
/a/b
/a/b/c
```

The `/a/b/c` directory is a real file system and is the shared resource. The security policy for `/a/b/c` is Kerberos with integrity. The server should apply the same security policy to `/`, `/a`, and `/a/b`. This allows for the extension of the protection of the server's namespace to the ancestors of the real shared resource.

For the case of the use of multiple, disjoint security mechanisms in the server's resources, the security for a particular object in the server's namespace should be the union of all security mechanisms of all direct descendants.

#### 8. Multi-Server Namespace

NFSv4 supports attributes that allow a namespace to extend beyond the boundaries of a single server. It is RECOMMENDED that clients and servers support construction of such multi-server namespaces. Use of such multi-server namespaces is OPTIONAL, however, and for many purposes, single-server namespaces are perfectly acceptable. Use of

multi-server namespaces can provide many advantages, however, by separating a file system's logical position in a namespace from the (possibly changing) logistical and administrative considerations that result in particular file systems being located on particular servers.

### 8.1. Location Attributes

NFSv4 contains RECOMMENDED attributes that allow file systems on one server to be associated with one or more instances of that file system on other servers. These attributes specify such file system instances by specifying a server address target (either as a DNS name representing one or more IP addresses or as a literal IP address) together with the path of that file system within the associated single-server namespace.

The `fs_locations` RECOMMENDED attribute allows specification of the file system locations where the data corresponding to a given file system may be found.

### 8.2. File System Presence or Absence

A given location in an NFSv4 namespace (typically but not necessarily a multi-server namespace) can have a number of file system instance locations associated with it via the `fs_locations` attribute. There may also be an actual current file system at that location, accessible via normal namespace operations (e.g., LOOKUP). In this case, the file system is said to be "present" at that position in the namespace, and clients will typically use it, reserving use of additional locations specified via the location-related attributes to situations in which the principal location is no longer available.

When there is no actual file system at the namespace location in question, the file system is said to be "absent". An absent file system contains no files or directories other than the root. Any reference to it, except to access a small set of attributes useful in determining alternate locations, will result in an error, `NFS4ERR_MOVED`. Note that if the server ever returns the error `NFS4ERR_MOVED`, it MUST support the `fs_locations` attribute.

While the error name suggests that we have a case of a file system that once was present, and has only become absent later, this is only one possibility. A position in the namespace may be permanently absent with the set of file system(s) designated by the location attributes being the only realization. The name `NFS4ERR_MOVED` reflects an earlier, more limited conception of its function, but this error will be returned whenever the referenced file system is absent, whether it has moved or not.

Except in the case of GETATTR-type operations (to be discussed later), when the current filehandle at the start of an operation is within an absent file system, that operation is not performed and the error NFS4ERR\_MOVED is returned, to indicate that the file system is absent on the current server.

Because a GETFH cannot succeed if the current filehandle is within an absent file system, filehandles within an absent file system cannot be transferred to the client. When a client does have filehandles within an absent file system, it is the result of obtaining them when the file system was present, and having the file system become absent subsequently.

It should be noted that because the check for the current filehandle being within an absent file system happens at the start of every operation, operations that change the current filehandle so that it is within an absent file system will not result in an error. This allows such combinations as PUTFH-GETATTR and LOOKUP-GETATTR to be used to get attribute information, particularly location attribute information, as discussed below.

### 8.3. Getting Attributes for an Absent File System

When a file system is absent, most attributes are not available, but it is necessary to allow the client access to the small set of attributes that are available, and most particularly that which gives information about the correct current locations for this file system, `fs_locations`.

#### 8.3.1. GETATTR Within an Absent File System

As mentioned above, an exception is made for GETATTR in that attributes may be obtained for a filehandle within an absent file system. This exception only applies if the attribute mask contains at least the `fs_locations` attribute bit, which indicates the client is interested in a result regarding an absent file system. If it is not requested, GETATTR will result in an NFS4ERR\_MOVED error.

When a GETATTR is done on an absent file system, the set of supported attributes is very limited. Many attributes, including those that are normally REQUIRED, will not be available on an absent file system. In addition to the `fs_locations` attribute, the following attributes SHOULD be available on absent file systems. In the case of RECOMMENDED attributes, they should be available at least to the same degree that they are available on present file systems.

**fsid:** This attribute should be provided so that the client can determine file system boundaries, including, in particular, the boundary between present and absent file systems. This value must be different from any other fsid on the current server and need have no particular relationship to fsids on any particular destination to which the client might be directed.

**mounted\_on\_fileid:** For objects at the top of an absent file system, this attribute needs to be available. Since the fileid is within the present parent file system, there should be no need to reference the absent file system to provide this information.

Other attributes SHOULD NOT be made available for absent file systems, even when it is possible to provide them. The server should not assume that more information is always better and should avoid gratuitously providing additional information.

When a GETATTR operation includes a bit mask for the attribute `fs_locations`, but where the bit mask includes attributes that are not supported, GETATTR will not return an error, but will return the mask of the actual attributes supported with the results.

Handling of VERIFY/NVERIFY is similar to GETATTR in that if the attribute mask does not include `fs_locations` the error `NFS4ERR_MOVED` will result. It differs in that any appearance in the attribute mask of an attribute not supported for an absent file system (and note that this will include some normally REQUIRED attributes) will also cause an `NFS4ERR_MOVED` result.

### 8.3.2. READDIR and Absent File Systems

A READDIR performed when the current filehandle is within an absent file system will result in an `NFS4ERR_MOVED` error, since, unlike the case of GETATTR, no such exception is made for READDIR.

Attributes for an absent file system may be fetched via a READDIR for a directory in a present file system, when that directory contains the root directories of one or more absent file systems. In this case, the handling is as follows:

- o If the attribute set requested includes `fs_locations`, then fetching of attributes proceeds normally and no `NFS4ERR_MOVED` indication is returned, even when the `rdattr_error` attribute is requested.
- o If the attribute set requested does not include `fs_locations`, then if the `rdattr_error` attribute is requested, each directory entry for the root of an absent file system will report `NFS4ERR_MOVED` as



the value of the `rdattr_error` attribute.

- o If the attribute set requested does not include either of the attributes `fs_locations` or `rdattr_error` then the occurrence of the root of an absent file system within the directory will result in the `REaddir` failing with an `NFS4ERR_MOVED` error.
- o The unavailability of an attribute because of a file system's absence, even one that is ordinarily `REQUIRED`, does not result in any error indication. The set of attributes returned for the root directory of the absent file system in that case is simply restricted to those actually available.

#### 8.4. Uses of Location Information

The location-bearing attribute of `fs_locations` provides, together with the possibility of absent file systems, a number of important facilities in providing reliable, manageable, and scalable data access.

When a file system is present, these attributes can provide alternative locations, to be used to access the same data, in the event of server failures, communications problems, or other difficulties that make continued access to the current file system impossible or otherwise impractical. Under some circumstances, multiple alternative locations may be used simultaneously to provide higher-performance access to the file system in question. Provision of such alternate locations is referred to as "replication" although there are cases in which replicated sets of data are not in fact present, and the replicas are instead different paths to the same data.

When a file system is present and becomes absent, clients can be given the opportunity to have continued access to their data, at an alternate location. In this case, a continued attempt to use the data in the now-absent file system will result in an `NFS4ERR_MOVED` error and, at that point, the successor locations (typically only one although multiple choices are possible) can be fetched and used to continue access. Transfer of the file system contents to the new location is referred to as "migration", but it should be kept in mind that there are cases in which this term can be used, like "replication", when there is no actual data migration per se.

Where a file system was not previously present, specification of file system location provides a means by which file systems located on one server can be associated with a namespace defined by another server, thus allowing a general multi-server namespace facility. A designation of such a location, in place of an absent file system, is

called a "referral".

Because client support for location-related attributes is OPTIONAL, a server may (but is not required to) take action to hide migration and referral events from such clients, by acting as a proxy, for example.

#### 8.4.1. File System Replication

The `fs_locations` attribute provides alternative locations, to be used to access data in place of or in addition to the current file system instance. On first access to a file system, the client should obtain the value of the set of alternate locations by interrogating the `fs_locations` attribute.

In the event that server failures, communications problems, or other difficulties make continued access to the current file system impossible or otherwise impractical, the client can use the alternate locations as a way to get continued access to its data. Multiple locations may be used simultaneously, to provide higher performance through the exploitation of multiple paths between client and target file system.

The alternate locations may be physical replicas of the (typically read-only) file system data, or they may reflect alternate paths to the same server or provide for the use of various forms of server clustering in which multiple servers provide alternate ways of accessing the same physical file system. How these different modes of file system transition are represented within the `fs_locations` attribute and how the client deals with file system transition issues will be discussed in detail below.

Multiple server addresses, whether they are derived from a single entry with a DNS name representing a set of IP addresses or from multiple entries each with its own server address, may correspond to the same actual server.

#### 8.4.2. File System Migration

When a file system is present and becomes absent, clients can be given the opportunity to have continued access to their data, at an alternate location, as specified by the `fs_locations` attribute. Typically, a client will be accessing the file system in question, get an `NFS4ERR_MOVED` error, and then use the `fs_locations` attribute to determine the new location of the data.

Such migration can be helpful in providing load balancing or general resource reallocation. The protocol does not specify how the file system will be moved between servers. It is anticipated that a

number of different server-to-server transfer mechanisms might be used with the choice left to the server implementor. The NFSv4 protocol specifies the method used to communicate the migration event between client and server.

The new location may be an alternate communication path to the same server or, in the case of various forms of server clustering, another server providing access to the same physical file system. The client's responsibilities in dealing with this transition depend on the specific nature of the new access path as well as how and whether data was in fact migrated. These issues will be discussed in detail below.

When an alternate location is designated as the target for migration, it must designate the same data. Where file systems are writable, a change made on the original file system must be visible on all migration targets. Where a file system is not writable but represents a read-only copy (possibly periodically updated) of a writable file system, similar requirements apply to the propagation of updates. Any change visible in the original file system must already be effected on all migration targets, to avoid any possibility that a client, in effecting a transition to the migration target, will see any reversion in file system state.

#### 8.4.3. Referrals

Referrals provide a way of placing a file system in a location within the namespace essentially without respect to its physical location on a given server. This allows a single server or a set of servers to present a multi-server namespace that encompasses file systems located on multiple servers. Some likely uses of this include establishment of site-wide or organization-wide namespaces, or even knitting such together into a truly global namespace.

Referrals occur when a client determines, upon first referencing a position in the current namespace, that it is part of a new file system and that the file system is absent. When this occurs, typically by receiving the error NFS4ERR\_MOVED, the actual location or locations of the file system can be determined by fetching the `fs_locations` attribute.

The locations-related attribute may designate a single file system location or multiple file system locations, to be selected based on the needs of the client.

Use of multi-server namespaces is enabled by NFSv4 but is not required. The use of multi-server namespaces and their scope will depend on the applications used and system administration

preferences.

Multi-server namespaces can be established by a single server providing a large set of referrals to all of the included file systems. Alternatively, a single multi-server namespace may be administratively segmented with separate referral file systems (on separate servers) for each separately administered portion of the namespace. The top-level referral file system or any segment may use replicated referral file systems for higher availability.

Generally, multi-server namespaces are for the most part uniform, in that the same data made available to one client at a given location in the namespace is made available to all clients at that location.

#### 8.5. Location Entries and Server Identity

As mentioned above, a single location entry may have a server address target in the form of a DNS name that may represent multiple IP addresses, while multiple location entries may have their own server address targets that reference the same server.

When multiple addresses for the same server exist, the client may assume that for each file system in the namespace of a given server network address, there exist file systems at corresponding namespace locations for each of the other server network addresses. It may do this even in the absence of explicit listing in `fs_locations`. Such corresponding file system locations can be used as alternate locations, just as those explicitly specified via the `fs_locations` attribute.

If a single location entry designates multiple server IP addresses, the client cannot assume that these addresses are multiple paths to the same server. In most cases, they will be, but the client **MUST** verify that before acting on that assumption. When two server addresses are designated by a single location entry and they correspond to different servers, this normally indicates some sort of misconfiguration, and so the client should avoid using such location entries when alternatives are available. When they are not, clients should pick one of IP addresses and use it, without using others that are not directed to the same server.

#### 8.6. Additional Client-Side Considerations

When clients make use of servers that implement referrals, replication, and migration, care should be taken that a user who mounts a given file system that includes a referral or a relocated file system continues to see a coherent picture of that user-side file system despite the fact that it contains a number of server-side

file systems that may be on different servers.

One important issue is upward navigation from the root of a server-side file system to its parent (specified as ".." in UNIX), in the case in which it transitions to that file system as a result of referral, migration, or a transition as a result of replication. When the client is at such a point, and it needs to ascend to the parent, it must go back to the parent as seen within the multi-server namespace rather than sending a LOOKUPP operation to the server, which would result in the parent within that server's single-server namespace. In order to do this, the client needs to remember the filehandles that represent such file system roots and use these instead of issuing a LOOKUPP operation to the current server. This will allow the client to present to applications a consistent namespace, where upward navigation and downward navigation are consistent.

Another issue concerns refresh of referral locations. When referrals are used extensively, they may change as server configurations change. It is expected that clients will cache information related to traversing referrals so that future client-side requests are resolved locally without server communication. This is usually rooted in client-side name look up caching. Clients should periodically purge this data for referral points in order to detect changes in location information.

A potential problem exists if a client were to allow an open owner to have state on multiple file systems on server, in that it is unclear how the sequence numbers associated with open owners are to be dealt with, in the event of transparent state migration. A client can avoid such a situation, if it ensures that any use of an open owner is confined to a single file system.

A server MAY decline to migrate state associated with open owners that span multiple file systems. In cases in which the server chooses not to migrate such state, the server MUST return NFS4ERR\_BAD\_STATEID when the client uses those stateids on the new server.

The server MUST return NFS4ERR\_STALE\_STATEID when the client uses those stateids on the old server, regardless of whether migration has occurred or not.

#### 8.7. Effecting File System Referrals

Referrals are effected when an absent file system is encountered, and one or more alternate locations are made available by the fs\_locations attribute. The client will typically get an

NFS4ERR\_MOVED error, fetch the appropriate location information, and proceed to access the file system on a different server, even though it retains its logical position within the original namespace. Referrals differ from migration events in that they happen only when the client has not previously referenced the file system in question (so there is nothing to transition). Referrals can only come into effect when an absent file system is encountered at its root.

The examples given in the sections below are somewhat artificial in that an actual client will not typically do a multi-component look up, but will have cached information regarding the upper levels of the name hierarchy. However, these examples are chosen to make the required behavior clear and easy to put within the scope of a small number of requests, without getting unduly into details of how specific clients might choose to cache things.

#### 8.7.1. Referral Example (LOOKUP)

Let us suppose that the following COMPOUND is sent in an environment in which /this/is/the/path is absent from the target server. This may be for a number of reasons. It may be the case that the file system has moved, or it may be the case that the target server is functioning mainly, or solely, to refer clients to the servers on which various file systems are located.

- o PUTROOTFH
- o LOOKUP "this"
- o LOOKUP "is"
- o LOOKUP "the"
- o LOOKUP "path"
- o GETFH
- o GETATTR(fsid,fileid,size,time\_modify)

Under the given circumstances, the following will be the result.

- o PUTROOTFH --> NFS\_OK. The current fh is now the root of the pseudo-fs.
- o LOOKUP "this" --> NFS\_OK. The current fh is for /this and is within the pseudo-fs.

- o LOOKUP "is" --> NFS\_OK. The current fh is for /this/is and is within the pseudo-fs.
- o LOOKUP "the" --> NFS\_OK. The current fh is for /this/is/the and is within the pseudo-fs.
- o LOOKUP "path" --> NFS\_OK. The current fh is for /this/is/the/path and is within a new, absent file system, but ... the client will never see the value of that fh.
- o GETFH --> NFS4ERR\_MOVED. Fails because current fh is in an absent file system at the start of the operation, and the specification makes no exception for GETFH.
- o GETATTR(fsid,fileid,size,time\_modify) Not executed because the failure of the GETFH stops processing of the COMPOUND.

Given the failure of the GETFH, the client has the job of determining the root of the absent file system and where to find that file system, i.e., the server and path relative to that server's root fh. Note here that in this example, the client did not obtain filehandles and attribute information (e.g., fsid) for the intermediate directories, so that it would not be sure where the absent file system starts. It could be the case, for example, that /this/is/the is the root of the moved file system and that the reason that the look up of "path" succeeded is that the file system was not absent on that operation but was moved between the last LOOKUP and the GETFH (since COMPOUND is not atomic). Even if we had the fsids for all of the intermediate directories, we could have no way of knowing that /this/is/the/path was the root of a new file system, since we don't yet have its fsid.

In order to get the necessary information, let us re-send the chain of LOOKUPS with GETFHs and GETATTRs to at least get the fsids so we can be sure where the appropriate file system boundaries are. The client could choose to get fs\_locations at the same time but in most cases the client will have a good guess as to where file system boundaries are (because of where NFS4ERR\_MOVED was, and was not, received) making fetching of fs\_locations unnecessary.

OP01: PUTROOTFH --> NFS\_OK

- Current fh is root of pseudo-fs.

OP02: GETATTR(fsid) --> NFS\_OK

- Just for completeness. Normally, clients will know the fsid of the pseudo-fs as soon as they establish communication with a server.

OP03: LOOKUP "this" --> NFS\_OK

OP04: GETATTR(fsid) --> NFS\_OK

- Get current fsid to see where file system boundaries are. The fsid will be that for the pseudo-fs in this example, so no boundary.

OP05: GETFH --> NFS\_OK

- Current fh is for /this and is within pseudo-fs.

OP06: LOOKUP "is" --> NFS\_OK

- Current fh is for /this/is and is within pseudo-fs.

OP07: GETATTR(fsid) --> NFS\_OK

- Get current fsid to see where file system boundaries are. The fsid will be that for the pseudo-fs in this example, so no boundary.

OP08: GETFH --> NFS\_OK

- Current fh is for /this/is and is within pseudo-fs.

OP09: LOOKUP "the" --> NFS\_OK

- Current fh is for /this/is/the and is within pseudo-fs.

OP10: GETATTR(fsid) --> NFS\_OK

- Get current fsid to see where file system boundaries are. The fsid will be that for the pseudo-fs in this example, so no boundary.

OP11: GETFH --> NFS\_OK

- Current fh is for /this/is/the and is within pseudo-fs.

OP12: LOOKUP "path" --> NFS\_OK



- Current fh is for /this/is/the/path and is within a new, absent file system, but ...
- The client will never see the value of that fh.

OP13: GETATTR(fsid, fs\_locations) --> NFS\_OK

- We are getting the fsid to know where the file system boundaries are. In this operation, the fsid will be different than that of the parent directory (which in turn was retrieved in OP10). Note that the fsid we are given will not necessarily be preserved at the new location. That fsid might be different, and in fact the fsid we have for this file system might be a valid fsid of a different file system on that new server.
- In this particular case, we are pretty sure anyway that what has moved is /this/is/the/path rather than /this/is/the since we have the fsid of the latter and it is that of the pseudo-fs, which presumably cannot move. However, in other examples, we might not have this kind of information to rely on (e.g., /this/is/the might be a non-pseudo file system separate from /this/is/the/path), so we need to have other reliable source information on the boundary of the file system that is moved. If, for example, the file system /this/is had moved, we would have a case of migration rather than referral, and once the boundaries of the migrated file system was clear we could fetch fs\_locations.
- We are fetching fs\_locations because the fact that we got an NFS4ERR\_MOVED at this point means that it is most likely that this is a referral and we need the destination. Even if it is the case that /this/is/the is a file system that has migrated, we will still need the location information for that file system.

OP14: GETFH --> NFS4ERR\_MOVED

- Fails because current fh is in an absent file system at the start of the operation, and the specification makes no exception for GETFH. Note that this means the server will never send the client a filehandle from within an absent file system.

Given the above, the client knows where the root of the absent file system is (/this/is/the/path) by noting where the change of fsid occurred (between "the" and "path"). The fs\_locations attribute also gives the client the actual location of the absent file system, so that the referral can proceed. The server gives the client the bare minimum of information about the absent file system so that there will be very little scope for problems of conflict between information sent by the referring server and information of the file

system's home. No filehandles and very few attributes are present on the referring server, and the client can treat those it receives as transient information with the function of enabling the referral.

#### 8.7.2. Referral Example (READDIR)

Another context in which a client may encounter referrals is when it does a READDIR on a directory in which some of the sub-directories are the roots of absent file systems.

Suppose such a directory is read as follows:

- o PUTROOTFH
- o LOOKUP "this"
- o LOOKUP "is"
- o LOOKUP "the"
- o READDIR (fsid, size, time\_modify, mounted\_on\_fileid)

In this case, because rdattn\_error is not requested, fs\_locations is not requested, and some of the attributes cannot be provided, the result will be an NFS4ERR\_MOVED error on the READDIR, with the detailed results as follows:

- o PUTROOTFH --> NFS\_OK. The current fh is at the root of the pseudo-fs.
- o LOOKUP "this" --> NFS\_OK. The current fh is for /this and is within the pseudo-fs.
- o LOOKUP "is" --> NFS\_OK. The current fh is for /this/is and is within the pseudo-fs.
- o LOOKUP "the" --> NFS\_OK. The current fh is for /this/is/the and is within the pseudo-fs.
- o READDIR (fsid, size, time\_modify, mounted\_on\_fileid) --> NFS4ERR\_MOVED. Note that the same error would have been returned if /this/is/the had migrated, but it is returned because the directory contains the root of an absent file system.

So now suppose that we re-send with rdattn\_error:

- o PUTROOTFH
- o LOOKUP "this"
- o LOOKUP "is"
- o LOOKUP "the"
- o READDIR (rdattr\_error, fsid, size, time\_modify, mounted\_on\_fileid)

The results will be:

- o PUTROOTFH --> NFS\_OK. The current fh is at the root of the pseudo-fs.
- o LOOKUP "this" --> NFS\_OK. The current fh is for /this and is within the pseudo-fs.
- o LOOKUP "is" --> NFS\_OK. The current fh is for /this/is and is within the pseudo-fs.
- o LOOKUP "the" --> NFS\_OK. The current fh is for /this/is/the and is within the pseudo-fs.
- o READDIR (rdattr\_error, fsid, size, time\_modify, mounted\_on\_fileid) --> NFS\_OK. The attributes for directory entry with the component named "path" will only contain rdattr\_error with the value NFS4ERR\_MOVED, together with an fsid value and a value for mounted\_on\_fileid.

So suppose we do another READDIR to get fs\_locations (although we could have used a GETATTR directly, as in Section 8.7.1).

- o PUTROOTFH
- o LOOKUP "this"
- o LOOKUP "is"
- o LOOKUP "the"
- o READDIR (rdattr\_error, fs\_locations, mounted\_on\_fileid, fsid, size, time\_modify)

The results would be:

- o PUTROOTFH --> NFS\_OK. The current fh is at the root of the pseudo-fs.
- o LOOKUP "this" --> NFS\_OK. The current fh is for /this and is within the pseudo-fs.
- o LOOKUP "is" --> NFS\_OK. The current fh is for /this/is and is within the pseudo-fs.
- o LOOKUP "the" --> NFS\_OK. The current fh is for /this/is/the and is within the pseudo-fs.
- o READDIR (rdattr\_error, fs\_locations, mounted\_on\_fileid, fsid, size, time\_modify) --> NFS\_OK. The attributes will be as shown below.

The attributes for the directory entry with the component named "path" will only contain:

- o rdattr\_error (value: NFS\_OK)
- o fs\_locations
- o mounted\_on\_fileid (value: unique fileid within referring file system)
- o fsid (value: unique value within referring server)

The attributes for entry "path" will not contain size or time\_modify because these attributes are not available within an absent file system.

#### 8.8. The Attribute fs\_locations

The fs\_locations attribute is structured in the following way:

```
struct fs_location4 {
    utf8str_cis          server<>;
    pathname4            rootpath;
};

struct fs_locations4 {
    pathname4            fs_root;
    fs_location4         locations<>;
};
```

The `fs_location4` data type is used to represent the location of a file system by providing a server name and the path to the root of the file system within that server's namespace. When a set of servers have corresponding file systems at the same path within their namespaces, an array of server names may be provided. An entry in the server array is a UTF-8 string and represents one of a traditional DNS host name, IPv4 address, IPv6 address, or a zero-length string. A zero-length string SHOULD be used to indicate the current address being used for the RPC call. It is not a requirement that all servers that share the same rootpath be listed in one `fs_location4` instance. The array of server names is provided for convenience. Servers that share the same rootpath may also be listed in separate `fs_location4` entries in the `fs_locations` attribute.

The `fs_locations4` data type and `fs_locations` attribute contain an array of such locations. Since the namespace of each server may be constructed differently, the "fs\_root" field is provided. The path represented by `fs_root` represents the location of the file system in the current server's namespace, i.e., that of the server from which the `fs_locations` attribute was obtained. The `fs_root` path is meant to aid the client by clearly referencing the root of the file system whose locations are being reported, no matter what object within the current file system the current filehandle designates. The `fs_root` is simply the pathname the client used to reach the object on the current server (i.e., the object to which the `fs_locations` attribute applies).

When the `fs_locations` attribute is interrogated and there are no alternate file system locations, the server SHOULD return a zero-length array of `fs_location4` structures, together with a valid `fs_root`.

As an example, suppose there is a replicated file system located at two servers (`servA` and `servB`). At `servA`, the file system is located at path `/a/b/c`. At `servB` the file system is located at path `/x/y/z`. If the client were to obtain the `fs_locations` value for the directory at `/a/b/c/d`, it might not necessarily know that the file system's root is located in `servA`'s namespace at `/a/b/c`. When the client switches to `servB`, it will need to determine that the directory it first referenced at `servA` is now represented by the path `/x/y/z/d` on `servB`. To facilitate this, the `fs_locations` attribute provided by `servA` would have an `fs_root` value of `/a/b/c` and two entries in `fs_locations`. One entry in `fs_locations` will be for itself (`servA`) and the other will be for `servB` with a path of `/x/y/z`. With this information, the client is able to substitute `/x/y/z` for the `/a/b/c` at the beginning of its access path and construct `/x/y/z/d` to use for the new server.

Note that: there is no requirement that the number of components in each rootpath be the same; there is no relation between the number of components in rootpath or fs\_root, and none of the components in each rootpath and fs\_root have to be the same. In the above example, we could have had a third element in the locations array, with server equal to "servC", and rootpath equal to "/I/II", and a fourth element in locations with server equal to "servD" and rootpath equal to "/aleph/beth/gimel/dalet/he".

The relationship between fs\_root to a rootpath is that the client replaces the pathname indicated in fs\_root for the current server for the substitute indicated in rootpath for the new server.

For an example of a referred or migrated file system, suppose there is a file system located at serv1. At serv1, the file system is located at /az/buky/vedi/glagoli. The client finds that object at glagoli has migrated (or is a referral). The client gets the fs\_locations attribute, which contains an fs\_root of /az/buky/vedi/glagoli, and one element in the locations array, with server equal to serv2, and rootpath equal to /izhitsa/fita. The client replaces /az/buky/vedi/glagoli with /izhitsa/fita, and uses the latter pathname on serv2.

Thus, the server MUST return an fs\_root that is equal to the path the client used to reach the object to which the fs\_locations attribute applies. Otherwise, the client cannot determine the new path to use on the new server.

#### 8.8.1. Inferring Transition Modes

When fs\_locations is used, information about the specific locations should be assumed based on the following rules.

The following rules are general and apply irrespective of the context.

- o All listed file system instances should be considered as of the same handle class if and only if the current fh\_expire\_type attribute does not include the FH4\_VOL\_MIGRATION bit. Note that in the case of referral, filehandle issues do not apply since there can be no filehandles known within the current file system nor is there any access to the fh\_expire\_type attribute on the referring (absent) file system.
- o All listed file system instances should be considered as of the same fileid class if and only if the fh\_expire\_type attribute indicates persistent filehandles and does not include the FH4\_VOL\_MIGRATION bit. Note that in the case of referral, fileid

issues do not apply since there can be no fileids known within the referring (absent) file system nor is there any access to the `fh_expire_type` attribute.

- o All file system instances servers should be considered as of different change classes.
- o All file system instances servers should be considered as of different readdir classes.

For other class assignments, handling of file system transitions depends on the reasons for the transition:

- o When the transition is due to migration, that is, the client was directed to a new file system after receiving an `NFS4ERR_MOVED` error, the target should be treated as being of the same write-verifier class as the source.
- o When the transition is due to failover to another replica, that is, the client selected another replica without receiving and `NFS4ERR_MOVED` error, the target should be treated as being of a different write-verifier class from the source.

The specific choices reflect typical implementation patterns for failover and controlled migration, respectively.

See Section 17 for a discussion on the recommendations for the security flavor to be used by any `GETATTR` operation that requests the `"fs_locations"` attribute.

## 9. File Locking and Share Reservations

Integrating locking into the NFS protocol necessarily causes it to be stateful. With the inclusion of share reservations the protocol becomes substantially more dependent on state than the traditional combination of NFS and NLM (Network Lock Manager) [xnfs]. There are three components to making this state manageable:

- o clear division between client and server
- o ability to reliably detect inconsistency in state between client and server
- o simple and robust recovery mechanisms

In this model, the server owns the state information. The client requests changes in locks and the server responds with the changes

made. Non-client-initiated changes in locking state are infrequent. The client receives prompt notification of such changes and can adjust its view of the locking state to reflect the server's changes.

Individual pieces of state created by the server and passed to the client at its request are represented by 128-bit stateids. These stateids may represent a particular open file, a set of byte-range locks held by a particular owner, or a recallable delegation of privileges to access a file in particular ways or at a particular location.

In all cases, there is a transition from the most general information that represents a client as a whole to the eventual lightweight stateid used for most client and server locking interactions. The details of this transition will vary with the type of object but it always starts with a client ID.

To support Win32 share reservations it is necessary to atomically OPEN or CREATE files and apply the appropriate locks in the same operation. Having a separate share/unshare operation would not allow correct implementation of the Win32 OpenFile API. In order to correctly implement share semantics, the previous NFS protocol mechanisms used when a file is opened or created (LOOKUP, CREATE, ACCESS) need to be replaced. The NFSv4 protocol has an OPEN operation that subsumes the NFSv3 methodology of LOOKUP, CREATE, and ACCESS. However, because many operations require a filehandle, the traditional LOOKUP is preserved to map a file name to filehandle without establishing state on the server. The policy of granting access or modifying files is managed by the server based on the client's state. These mechanisms can implement policy ranging from advisory only locking to full mandatory locking.

#### 9.1. Opens and Byte-Range Locks

It is assumed that manipulating a byte-range lock is rare when compared to READ and WRITE operations. It is also assumed that server restarts and network partitions are relatively rare. Therefore it is important that the READ and WRITE operations have a lightweight mechanism to indicate if they possess a held lock. A byte-range lock request contains the heavyweight information required to establish a lock and uniquely define the owner of the lock.

The following sections describe the transition from the heavy weight information to the eventual stateid used for most client and server locking and lease interactions.



#### 9.1.1.1. Client ID

For each LOCK request, the client must identify itself to the server. This is done in such a way as to allow for correct lock identification and crash recovery. A sequence of a SETCLIENTID operation followed by a SETCLIENTID\_CONFIRM operation is required to establish the identification onto the server. Establishment of identification by a new incarnation of the client also has the effect of immediately breaking any leased state that a previous incarnation of the client might have had on the server, as opposed to forcing the new client incarnation to wait for the leases to expire. Breaking the lease state amounts to the server removing all lock, share reservation, and, where the server is not supporting the CLAIM\_DELEGATE\_PREV claim type, all delegation state associated with same client with the same identity. For discussion of delegation state recovery, see Section 10.2.1.

Owners of opens and owners of byte-range locks are separate entities and remain separate even if the same opaque arrays are used to designate owners of each. The protocol distinguishes between open-owners (represented by open\_owner4 structures) and lock-owners (represented by lock\_owner4 structures).

Both sorts of owners consist of a clientid and an opaque owner string. For each client, the set of distinct owner values used with that client constitutes the set of owners of that type, for the given client.

Each open is associated with a specific open-owner while each byte-range lock is associated with a lock-owner and an open-owner, the latter being the open-owner associated with the open file under which the LOCK operation was done.

Client identification is encapsulated in the following structure:

```
struct nfs_client_id4 {  
    verifier4    verifier;  
    opaque       id<NFS4_OPAQUE_LIMIT>;  
};
```

The first field, verifier is a client incarnation verifier that is used to detect client reboots. Only if the verifier is different from that which the server has previously recorded for the client (as identified by the second field of the structure, id) does the server start the process of canceling the client's leased state.

The second field, id is a variable length string that uniquely defines the client.

There are several considerations for how the client generates the id string:

- o The string should be unique so that multiple clients do not present the same string. The consequences of two clients presenting the same string range from one client getting an error to one client having its leased state abruptly and unexpectedly canceled.
- o The string should be selected so the subsequent incarnations (e.g., reboots) of the same client cause the client to present the same string. The implementor is cautioned against an approach that requires the string to be recorded in a local file because this precludes the use of the implementation in an environment where there is no local disk and all file access is from an NFSv4 server.
- o The string should be different for each server network address that the client accesses, rather than common to all server network addresses. The reason is that it may not be possible for the client to tell if the same server is listening on multiple network addresses. If the client issues SETCLIENTID with the same id string to each network address of such a server, the server will think it is the same client, and each successive SETCLIENTID will cause the server to begin the process of removing the client's previous leased state.
- o The algorithm for generating the string should not assume that the client's network address won't change. This includes changes between client incarnations and even changes while the client is still running in its current incarnation. This means that if the client includes just the client's and server's network address in the id string, there is a real risk, after the client gives up the network address, that another client, using a similar algorithm for generating the id string, will generate a conflicting id string.

Given the above considerations, an example of a well generated id string is one that includes:

- o The server's network address.
- o The client's network address.
- o For a user level NFSv4 client, it should contain additional information to distinguish the client from other user level clients running on the same host, such as an universally unique identifier (UUID).

- o Additional information that tends to be unique, such as one or more of:
  - \* The client machine's serial number (for privacy reasons, it is best to perform some one way function on the serial number).
  - \* A MAC address.
  - \* The timestamp of when the NFSv4 software was first installed on the client (though this is subject to the previously mentioned caution about using information that is stored in a file, because the file might only be accessible over NFSv4).
  - \* A true random number. However since this number ought to be the same between client incarnations, this shares the same problem as that of the using the timestamp of the software installation.

As a security measure, the server MUST NOT cancel a client's leased state if the principal that established the state for a given id string is not the same as the principal issuing the SETCLIENTID.

Note that SETCLIENTID and SETCLIENTID\_CONFIRM has a secondary purpose of establishing the information the server needs to make callbacks to the client for purpose of supporting delegations. It is permitted to change this information via SETCLIENTID and SETCLIENTID\_CONFIRM within the same incarnation of the client without removing the client's leased state.

Once a SETCLIENTID and SETCLIENTID\_CONFIRM sequence has successfully completed, the client uses the shorthand client identifier, of type clientid4, instead of the longer and less compact nfs\_client\_id4 structure. This shorthand client identifier (a client ID) is assigned by the server and should be chosen so that it will not conflict with a client ID previously assigned by the server. This applies across server restarts or reboots. When a client ID is presented to a server and that client ID is not recognized, as would happen after a server reboot, the server will reject the request with the error NFS4ERR\_STALE\_CLIENTID. When this happens, the client must obtain a new client ID by use of the SETCLIENTID operation and then proceed to any other necessary recovery for the server reboot case (See Section 9.6.2).

The client must also employ the SETCLIENTID operation when it receives a NFS4ERR\_STALE\_STATEID error using a stateid derived from its current client ID, since this also indicates a server reboot which has invalidated the existing client ID (see Section 9.6.2 for details).

See the detailed descriptions of SETCLIENTID and SETCLIENTID\_CONFIRM for a complete specification of the operations.

#### 9.1.2. Server Release of Client ID

If the server determines that the client holds no associated state for its client ID, the server may choose to release the client ID. The server may make this choice for an inactive client so that resources are not consumed by those intermittently active clients. If the client contacts the server after this release, the server must ensure the client receives the appropriate error so that it will use the SETCLIENTID/SETCLIENTID\_CONFIRM sequence to establish a new identity. It should be clear that the server must be very hesitant to release a client ID since the resulting work on the client to recover from such an event will be the same burden as if the server had failed and restarted. Typically a server would not release a client ID unless there had been no activity from that client for many minutes.

Note that if the id string in a SETCLIENTID request is properly constructed, and if the client takes care to use the same principal for each successive use of SETCLIENTID, then, barring an active denial of service attack, NFS4ERR\_CLID\_INUSE should never be returned.

However, client bugs, server bugs, or perhaps a deliberate change of the principal owner of the id string (such as the case of a client that changes security flavors, and under the new flavor, there is no mapping to the previous owner) will in rare cases result in NFS4ERR\_CLID\_INUSE.

In that event, when the server gets a SETCLIENTID for a client ID that currently has no state, or it has state, but the lease has expired, rather than returning NFS4ERR\_CLID\_INUSE, the server MUST allow the SETCLIENTID, and confirm the new client ID if followed by the appropriate SETCLIENTID\_CONFIRM.

#### 9.1.3. Stateid Definition

When the server grants a lock of any type (including opens, byte-range locks, and delegations), it responds with a unique stateid that represents a set of locks (often a single lock) for the same file, of the same type, and sharing the same ownership characteristics. Thus, opens of the same file by different open-owners each have an identifying stateid. Similarly, each set of byte-range locks on a file owned by a specific lock-owner has its own identifying stateid. Delegations also have associated stateids by which they may be referenced. The stateid is used as a shorthand reference to a lock

or set of locks, and given a stateid, the server can determine the associated state-owner or state-owners (in the case of an open-owner/lock-owner pair) and the associated filehandle. When stateids are used, the current filehandle must be the one associated with that stateid.

All stateids associated with a given client ID are associated with a common lease that represents the claim of those stateids and the objects they represent to be maintained by the server. See Section 9.5 for a discussion of the lease.

Each stateid must be unique to the server. Many operations take a stateid as an argument but not a clientid, so the server must be able to infer the client from the stateid.

#### 9.1.3.1. Stateid Types

With the exception of special stateids (see Section 9.1.3.3), each stateid represents locking objects of one of a set of types defined by the NFSv4 protocol. Note that in all these cases, where we speak of guarantee, it is understood there are situations such as a client restart, or lock revocation, that allow the guarantee to be voided.

- o Stateids may represent opens of files.

Each stateid in this case represents the OPEN state for a given client ID/open-owner/filehandle triple. Such stateids are subject to change (with consequent incrementing of the stateid's seqid) in response to OPENS that result in upgrade and OPEN\_DOWNGRADE operations.

- o Stateids may represent sets of byte-range locks.

All locks held on a particular file by a particular owner and all gotten under the aegis of a particular open file are associated with a single stateid with the seqid being incremented whenever LOCK and LOCKU operations affect that set of locks.

- o Stateids may represent file delegations, which are recallable guarantees by the server to the client, that other clients will not reference, or will not modify a particular file, until the delegation is returned.

A stateid represents a single delegation held by a client for a particular filehandle.

#### 9.1.3.2. Stateid Structure

Stateids are divided into two fields, a 96-bit "other" field identifying the specific set of locks and a 32-bit "seqid" sequence value. Except in the case of special stateids (see Section 9.1.3.3), a particular value of the "other" field denotes a set of locks of the same type (for example, byte-range locks, opens, or delegations), for a specific file or directory, and sharing the same ownership characteristics. The seqid designates a specific instance of such a set of locks, and is incremented to indicate changes in such a set of locks, either by the addition or deletion of locks from the set, a change in the byte-range they apply to, or an upgrade or downgrade in the type of one or more locks.

When such a set of locks is first created, the server SHOULD return a stateid with seqid value of one. On subsequent operations that modify the set of locks, the server is required to increment the "seqid" field by one whenever it returns a stateid for the same state-owner/file/type combination and there is some change in the set of locks actually designated. In this case, the server will return a stateid with an "other" field the same as previously used for that state-owner/file/type combination, with an incremented "seqid" field. This pattern continues until the seqid is incremented past NFS4\_UINT32\_MAX, and one (not zero) SHOULD be the next seqid value. The purpose of the incrementing of the seqid is to allow the server to communicate to the client the order in which operations that modified locking state associated with a stateid have been processed.

In making comparisons between seqids, both by the client in determining the order of operations and by the server in determining whether the NFS4ERR\_OLD\_STATEID is to be returned, the possibility of the seqid being swapped around past the NFS4\_UINT32\_MAX value needs to be taken into account.

#### 9.1.3.3. Special Stateids

Stateid values whose "other" field is either all zeros or all ones are reserved. They may not be assigned by the server but have special meanings defined by the protocol. The particular meaning depends on whether the "other" field is all zeros or all ones and the specific value of the "seqid" field.

The following combinations of "other" and "seqid" are defined in NFSv4:

- o When "other" and "seqid" are both zero, the stateid is treated as a special anonymous stateid, which can be used in READ, WRITE, and SETATTR requests to indicate the absence of any open state

associated with the request. When an anonymous stateid value is used, and an existing open denies the form of access requested, then access will be denied to the request.

- o When "other" and "seqid" are both all ones, the stateid is a special READ bypass stateid. When this value is used in WRITE or SETATTR, it is treated like the anonymous value. When used in READ, the server MAY grant access, even if access would normally be denied to READ requests.

If a stateid value is used which has all zero or all ones in the "other" field, but does not match one of the cases above, the server MUST return the error NFS4ERR\_BAD\_STATEID.

Special stateids, unlike other stateids, are not associated with individual client IDs or filehandles and can be used with all valid client IDs and filehandles.

#### 9.1.3.4. Stateid Lifetime and Validation

Stateids must remain valid until either a client restart or a server restart or until the client returns all of the locks associated with the stateid by means of an operation such as CLOSE or DELEGRETURN. If the locks are lost due to revocation as long as the client ID is valid, the stateid remains a valid designation of that revoked state. Stateids associated with byte-range locks are an exception. They remain valid even if a LOCKU frees all remaining locks, so long as the open file with which they are associated remains open.

It should be noted that there are situations in which the client's locks become invalid, without the client requesting they be returned. These include lease expiration and a number of forms of lock revocation within the lease period. It is important to note that in these situations, the stateid remains valid and the client can use it to determine the disposition of the associated lost locks.

An "other" value must never be reused for a different purpose (i.e. different filehandle, owner, or type of locks) within the context of a single client ID. A server may retain the "other" value for the same purpose beyond the point where it may otherwise be freed but if it does so, it must maintain "seqid" continuity with previous values.

One mechanism that may be used to satisfy the requirement that the server recognize invalid and out-of-date stateids is for the server to divide the "other" field of the stateid into two fields.

- o An index into a table of locking-state structures.

- o A generation number which is incremented on each allocation of a table entry for a particular use.

And then store in each table entry,

- o The client ID with which the stateid is associated.
- o The current generation number for the (at most one) valid stateid sharing this index value.
- o The filehandle of the file on which the locks are taken.
- o An indication of the type of stateid (open, byte-range lock, file delegation).
- o The last "seqid" value returned corresponding to the current "other" value.
- o An indication of the current status of the locks associated with this stateid. In particular, whether these have been revoked and if so, for what reason.

With this information, an incoming stateid can be validated and the appropriate error returned when necessary. Special and non-special stateids are handled separately. (See Section 9.1.3.3 for a discussion of special stateids.)

When a stateid is being tested, and the "other" field is all zeros or all ones, a check that the "other" and "seqid" fields match a defined combination for a special stateid is done and the results determined as follows:

- o If the "other" and "seqid" fields do not match a defined combination associated with a special stateid, the error NFS4ERR\_BAD\_STATEID is returned.
- o If the combination is valid in general but is not appropriate to the context in which the stateid is used (e.g., an all-zero stateid is used when an open stateid is required in a LOCK operation), the error NFS4ERR\_BAD\_STATEID is also returned.
- o Otherwise, the check is completed and the special stateid is accepted as valid.

When a stateid is being tested, and the "other" field is neither all zeros or all ones, the following procedure could be used to validate an incoming stateid and return an appropriate error, when necessary, assuming that the "other" field would be divided into a table index



and an entry generation.

- o If the table index field is outside the range of the associated table, return NFS4ERR\_BAD\_STATEID.
- o If the selected table entry is of a different generation than that specified in the incoming stateid, return NFS4ERR\_BAD\_STATEID.
- o If the selected table entry does not match the current filehandle, return NFS4ERR\_BAD\_STATEID.
- o If the stateid represents revoked state or state lost as a result of lease expiration, then return NFS4ERR\_EXPIRED, NFS4ERR\_BAD\_STATEID, or NFS4ERR\_ADMIN\_REVOKED, as appropriate.
- o If the stateid type is not valid for the context in which the stateid appears, return NFS4ERR\_BAD\_STATEID. Note that a stateid may be valid in general, but be invalid for a particular operation, as, for example, when a stateid which doesn't represent byte-range locks is passed to the non-from\_open case of LOCK or to LOCKU, or when a stateid which does not represent an open is passed to CLOSE or OPEN\_DOWNGRADE. In such cases, the server MUST return NFS4ERR\_BAD\_STATEID.
- o If the "seqid" field is not zero, and it is greater than the current sequence value corresponding the current "other" field, return NFS4ERR\_BAD\_STATEID.
- o If the "seqid" field is less than the current sequence value corresponding the current "other" field, return NFS4ERR\_OLD\_STATEID.
- o Otherwise, the stateid is valid and the table entry should contain any additional information about the type of stateid and information associated with that particular type of stateid, such as the associated set of locks, such as open-owner and lock-owner information, as well as information on the specific locks, such as open modes and byte ranges.

#### 9.1.3.5. Stateid Use for I/O Operations

Clients performing Input/Output (I/O) operations need to select an appropriate stateid based on the locks (including opens and delegations) held by the client and the various types of state-owners sending the I/O requests. SETATTR operations that change the file size are treated like I/O operations in this regard.

The following rules, applied in order of decreasing priority, govern

the selection of the appropriate stateid. In following these rules, the client will only consider locks of which it has actually received notification by an appropriate operation response or callback.

- o If the client holds a delegation for the file in question, the delegation stateid SHOULD be used.
- o Otherwise, if the entity corresponding to the lock-owner (e.g., a process) sending the I/O has a byte-range lock stateid for the associated open file, then the byte-range lock stateid for that lock-owner and open file SHOULD be used.
- o If there is no byte-range lock stateid, then the OPEN stateid for the current open-owner, and that OPEN stateid for the open file in question SHOULD be used.
- o Finally, if none of the above apply, then a special stateid SHOULD be used.

Ignoring these rules may result in situations in which the server does not have information necessary to properly process the request. For example, when mandatory byte-range locks are in effect, if the stateid does not indicate the proper lock-owner, via a lock stateid, a request might be avoidably rejected.

The server however should not try to enforce these ordering rules and should use whatever information is available to properly process I/O requests. In particular, when a client has a delegation for a given file, it SHOULD take note of this fact in processing a request, even if it is sent with a special stateid.

#### 9.1.3.6. Stateid Use for SETATTR Operations

In the case of SETATTR operations, a stateid is present. In cases other than those that set the file size, the client may send either a special stateid or, when a delegation is held for the file in question, a delegation stateid. While the server SHOULD validate the stateid and may use the stateid to optimize the determination as to whether a delegation is held, it SHOULD note the presence of a delegation even when a special stateid is sent, and MUST accept a valid delegation stateid when sent.

#### 9.1.4. lock-owner

When requesting a lock, the client must present to the server the client ID and an identifier for the owner of the requested lock. These two fields are referred to as the lock-owner and the definition of those fields are:

- o A client ID returned by the server as part of the client's use of the SETCLIENTID operation.
- o A variable length opaque array used to uniquely define the owner of a lock managed by the client.

This may be a thread id, process id, or other unique value.

When the server grants the lock, it responds with a unique stateid. The stateid is used as a shorthand reference to the lock-owner, since the server will be maintaining the correspondence between them.

#### 9.1.1.5. Use of the Stateid and Locking

All READ, WRITE and SETATTR operations contain a stateid. For the purposes of this section, SETATTR operations which change the size attribute of a file are treated as if they are writing the area between the old and new size (i.e., the range truncated or added to the file by means of the SETATTR), even where SETATTR is not explicitly mentioned in the text. The stateid passed to one of these operations must be one that represents an OPEN (e.g., via the open-owner), a set of byte-range locks, or a delegation, or it may be a special stateid representing anonymous access or the special bypass stateid.

If the state-owner performs a READ or WRITE in a situation in which it has established a lock or share reservation on the server (any OPEN constitutes a share reservation) the stateid (previously returned by the server) must be used to indicate what locks, including both byte-range locks and share reservations, are held by the state-owner. If no state is established by the client, either byte-range lock or share reservation, a stateid of all bits 0 is used. Regardless whether a stateid of all bits 0, or a stateid returned by the server is used, if there is a conflicting share reservation or mandatory byte-range lock held on the file, the server MUST refuse to service the READ or WRITE operation.

Share reservations are established by OPEN operations and by their nature are mandatory in that when the OPEN denies READ or WRITE operations, that denial results in such operations being rejected with error NFS4ERR\_LOCKED. Byte-range locks may be implemented by the server as either mandatory or advisory, or the choice of mandatory or advisory behavior may be determined by the server on the basis of the file being accessed (for example, some UNIX-based servers support a "mandatory lock bit" on the mode attribute such that if set, byte-range locks are required on the file before I/O is possible). When byte-range locks are advisory, they only prevent the granting of conflicting lock requests and have no effect on READs or

WRITES. Mandatory byte-range locks, however, prevent conflicting I/O operations. When they are attempted, they are rejected with NFS4ERR\_LOCKED. When the client gets NFS4ERR\_LOCKED on a file it knows it has the proper share reservation for, it will need to issue a LOCK request on the region of the file that includes the region the I/O was to be performed on, with an appropriate locktype (i.e., READ\*\_LT for a READ operation, WRITE\*\_LT for a WRITE operation).

With NFSv3, there was no notion of a stateid so there was no way to tell if the application process of the client sending the READ or WRITE operation had also acquired the appropriate byte-range lock on the file. Thus there was no way to implement mandatory locking. With the stateid construct, this barrier has been removed.

Note that for UNIX environments that support mandatory file locking, the distinction between advisory and mandatory locking is subtle. In fact, advisory and mandatory byte-range locks are exactly the same in so far as the APIs and requirements on implementation. If the mandatory lock attribute is set on the file, the server checks to see if the lock-owner has an appropriate shared (read) or exclusive (write) byte-range lock on the region it wishes to read or write to. If there is no appropriate lock, the server checks if there is a conflicting lock (which can be done by attempting to acquire the conflicting lock on the behalf of the lock-owner, and if successful, release the lock after the READ or WRITE is done), and if there is, the server returns NFS4ERR\_LOCKED.

For Windows environments, there are no advisory byte-range locks, so the server always checks for byte-range locks during I/O requests.

Thus, the NFSv4 LOCK operation does not need to distinguish between advisory and mandatory byte-range locks. It is the NFS version 4 server's processing of the READ and WRITE operations that introduces the distinction.

Every stateid other than the special stateid values noted in this section, whether returned by an OPEN-type operation (i.e., OPEN, OPEN\_DOWNGRADE), or by a LOCK-type operation (i.e., LOCK or LOCKU), defines an access mode for the file (i.e., READ, WRITE, or READ-WRITE) as established by the original OPEN which began the stateid sequence, and as modified by subsequent OPENS and OPEN\_DOWNGRADES within that stateid sequence. When a READ, WRITE, or SETATTR which specifies the size attribute, is done, the operation is subject to checking against the access mode to verify that the operation is appropriate given the OPEN with which the operation is associated.

In the case of WRITE-type operations (i.e., WRITES and SETATTRs which set size), the server must verify that the access mode allows writing

and return an NFS4ERR\_OPENMODE error if it does not. In the case, of READ, the server may perform the corresponding check on the access mode, or it may choose to allow READ on opens for WRITE only, to accommodate clients whose write implementation may unavoidably do reads (e.g., due to buffer cache constraints). However, even if READs are allowed in these circumstances, the server MUST still check for locks that conflict with the READ (e.g., another open specifying denial of READs). Note that a server which does enforce the access mode check on READs need not explicitly check for conflicting share reservations since the existence of OPEN for read access guarantees that no conflicting share reservation can exist.

A stateid of all bits 1 (one) MAY allow READ operations to bypass locking checks at the server. However, WRITE operations with a stateid with bits all 1 (one) MUST NOT bypass locking checks and are treated exactly the same as if a stateid of all bits 0 were used.

A lock may not be granted while a READ or WRITE operation using one of the special stateids is being performed and the range of the lock request conflicts with the range of the READ or WRITE operation. For the purposes of this paragraph, a conflict occurs when a shared lock is requested and a WRITE operation is being performed, or an exclusive lock is requested and either a READ or a WRITE operation is being performed. A SETATTR that sets size is treated similarly to a WRITE as discussed above.

#### 9.1.6. Sequencing of Lock Requests

Locking is different than most NFS operations as it requires "at-most-one" semantics that are not provided by ONC RPC. ONC RPC over a reliable transport is not sufficient because a sequence of locking requests may span multiple TCP connections. In the face of retransmission or reordering, lock or unlock requests must have a well defined and consistent behavior. To accomplish this, each lock request contains a sequence number that is a consecutively increasing integer. Different state-owners have different sequences. The server maintains the last sequence number (L) received and the response that was returned. The server SHOULD assign a seqid value of one for the first request issued for any given state-owner.

Note that for requests that contain a sequence number, for each state-owner, there should be no more than one outstanding request.

If a request (r) with a previous sequence number ( $r < L$ ) is received, it is rejected with the return of error NFS4ERR\_BAD\_SEQID. Given a properly-functioning client, the response to (r) must have been received before the last request (L) was sent. If a duplicate of last request ( $r == L$ ) is received, the stored response is returned.

If a request beyond the next sequence ( $r == L + 2$ ) is received, it is rejected with the return of error NFS4ERR\_BAD\_SEQID. Sequence history is reinitialized whenever the SETCLIENTID/SETCLIENTID\_CONFIRM sequence changes the client verifier.

Since the sequence number is represented with an unsigned 32-bit integer, the arithmetic involved with the sequence number is mod  $2^{32}$ . Note that when the seqid wraps, it SHOULD bypass zero and use one as the next seqid value. For an example of modulo arithmetic involving sequence numbers see [RFC0793].

It is critical the server maintain the last response sent to the client to provide a more reliable cache of duplicate non-idempotent requests than that of the traditional cache described in [Chet]. The traditional duplicate request cache uses a least recently used algorithm for removing unneeded requests. However, the last lock request and response on a given state-owner must be cached as long as the lock state exists on the server.

The client MUST monotonically increment the sequence number for the CLOSE, LOCK, LOCKU, OPEN, OPEN\_CONFIRM, and OPEN\_DOWNGRADE operations. This is true even in the event that the previous operation that used the sequence number received an error. The only exception to this rule is if the previous operation received one of the following errors: NFS4ERR\_STALE\_CLIENTID, NFS4ERR\_STALE\_STATEID, NFS4ERR\_BAD\_STATEID, NFS4ERR\_BAD\_SEQID, NFS4ERR\_BADXDR, NFS4ERR\_RESOURCE, NFS4ERR\_NOFILEHANDLE, or NFS4ERR\_MOVED.

#### 9.1.7. Recovery from Replayed Requests

As described above, the sequence number is per state-owner. As long as the server maintains the last sequence number received and follows the methods described above, there are no risks of a Byzantine router re-sending old requests. The server need only maintain the (state-owner, sequence number) state as long as there are open files or closed files with locks outstanding.

LOCK, LOCKU, OPEN, OPEN\_DOWNGRADE, and CLOSE each contain a sequence number and therefore the risk of the replay of these operations resulting in undesired effects is non-existent while the server maintains the state-owner state.

#### 9.1.8. Interactions of multiple sequence values

Some Operations may have multiple sources of data for request sequence checking and retransmission determination. Some Operations have multiple sequence values associated with multiple types of state-owners. In addition, such Operations may also have a stateid

with its own seqid value, that will be checked for validity.

As noted above, there may be multiple sequence values to check. The following rules should be followed by the server in processing these multiple sequence values within a single operation.

- o When a sequence value associated with a state-owner is unavailable for checking because the state-owner is unknown to the server, it takes no part in the comparison.
- o When any of the state-owner sequence values are invalid, NFS4ERR\_BAD\_SEQID is returned. When a stateid sequence is checked, NFS4ERR\_BAD\_STATEID, or NFS4ERR\_OLD\_STATEID is returned as appropriate, but NFS4ERR\_BAD\_SEQID has priority.
- o When any one of the sequence values matches a previous request, for a state-owner, it is treated as a retransmission and not re-executed. When the type of the operation does not match that originally used, NFS4ERR\_BAD\_SEQID is returned. When the server can determine that the request differs from the original it may return NFS4ERR\_BAD\_SEQID.
- o When multiple of the sequence values match previous operations, but the operations are not the same, NFS4ERR\_BAD\_SEQID is returned.
- o When there are no available sequence values available for comparison and the operation is an OPEN, the server indicates to the client that an OPEN\_CONFIRM is required, unless it can conclusively determine that confirmation is not required (e.g., by knowing that no open-owner state has ever been released for the current clientid).

#### 9.1.9. Releasing state-owner State

When a particular state-owner no longer holds open or file locking state at the server, the server may choose to release the sequence number state associated with the state-owner. The server may make this choice based on lease expiration, for the reclamation of server memory, or other implementation specific details. Note that when this is done, a retransmitted request, normally identified by a matching state-owner sequence may not be correctly recognized, so that the client will not receive the original response that it would have if the state-owner state was not released.

If the server were able to be sure that a given state-owner would never again be used by a client, such an issue could not arise. Even when the state-owner state is released and the client subsequently

uses that state-owner, retransmitted requests will be detected as invalid and the request not executed, although the client may have a recovery path that is more complicated than simply getting the original response back transparently.

In any event, the server is able to safely release state-owner state (in the sense that retransmitted requests will not be erroneously acted upon) when the state-owner is not currently being utilized by the client (i.e., there are no open files associated with an open-owner and no lock stateids associated with a lock-owner). The server may choose to hold the state-owner state in order to simplify the recovery path, in the case in which retransmissions of currently active requests are received. However, the period it chooses to hold this state is implementation specific.

In the case that a LOCK, LOCKU, OPEN\_DOWNGRADE, or CLOSE is retransmitted after the server has previously released the state-owner state, the server will find that the state-owner has no files open and an error will be returned to the client. If the state-owner does have a file open, the stateid will not match and again an error is returned to the client.

#### 9.1.10. Use of Open Confirmation

In the case that an OPEN is retransmitted and the open-owner is being used for the first time or the open-owner state has been previously released by the server, the use of the OPEN\_CONFIRM operation will prevent incorrect behavior. When the server observes the use of the open-owner for the first time, it will direct the client to perform the OPEN\_CONFIRM for the corresponding OPEN. This sequence establishes the use of an open-owner and associated sequence number. Since the OPEN\_CONFIRM sequence connects a new open-owner on the server with an existing open-owner on a client, the sequence number may have any value. The OPEN\_CONFIRM step assures the server that the value received is the correct one. (see Section 15.20 for further details.)

There are a number of situations in which the requirement to confirm an OPEN would pose difficulties for the client and server, in that they would be prevented from acting in a timely fashion on information received, because that information would be provisional, subject to deletion upon non-confirmation. Fortunately, these are situations in which the server can avoid the need for confirmation when responding to open requests. The two constraints are:

- o The server must not bestow a delegation for any open which would require confirmation.



- o The server MUST NOT require confirmation on a reclaim-type open (i.e., one specifying claim type CLAIM\_PREVIOUS or CLAIM\_DELEGATE\_PREV).

These constraints are related in that reclaim-type opens are the only ones in which the server may be required to send a delegation. For CLAIM\_NULL, sending the delegation is optional while for CLAIM\_DELEGATE\_CUR, no delegation is sent.

Delegations being sent with an open requiring confirmation are troublesome because recovering from non-confirmation adds undue complexity to the protocol while requiring confirmation on reclaim-type opens poses difficulties in that the inability to resolve the status of the reclaim until lease expiration may make it difficult to have timely determination of the set of locks being reclaimed (since the grace period may expire).

Requiring open confirmation on reclaim-type opens is avoidable because of the nature of the environments in which such opens are done. For CLAIM\_PREVIOUS opens, this is immediately after server reboot, so there should be no time for open-owners to be created, found to be unused, and recycled. For CLAIM\_DELEGATE\_PREV opens, we are dealing with either a client reboot situation or a network partition resulting in deletion of lease state (and returning NFS4ERR\_EXPIRED). A server which supports delegations can be sure that no open-owners for that client have been recycled since client initialization or deletion of lease state and thus can ensure that confirmation will not be required.

## 9.2. Lock Ranges

The protocol allows a lock owner to request a lock with a byte range and then either upgrade or unlock a sub-range of the initial lock. It is expected that this will be an uncommon type of request. In any case, servers or server file systems may not be able to support sub-range lock semantics. In the event that a server receives a locking request that represents a sub-range of current locking state for the lock owner, the server is allowed to return the error NFS4ERR\_LOCK\_RANGE to signify that it does not support sub-range lock operations. Therefore, the client should be prepared to receive this error and, if appropriate, report the error to the requesting application.

The client is discouraged from combining multiple independent locking ranges that happen to be adjacent into a single request since the server may not support sub-range requests and for reasons related to the recovery of file locking state in the event of server failure. As discussed in the Section 9.6.2 below, the server may employ

certain optimizations during recovery that work effectively only when the client's behavior during lock recovery is similar to the client's locking behavior prior to server failure.

### 9.3. Upgrading and Downgrading Locks

If a client has a write lock on a record, it can request an atomic downgrade of the lock to a read lock via the LOCK request, by setting the type to READ\_LT. If the server supports atomic downgrade, the request will succeed. If not, it will return NFS4ERR\_LOCK\_NOTSUPP. The client should be prepared to receive this error, and if appropriate, report the error to the requesting application.

If a client has a read lock on a record, it can request an atomic upgrade of the lock to a write lock via the LOCK request by setting the type to WRITE\_LT or WRITEW\_LT. If the server does not support atomic upgrade, it will return NFS4ERR\_LOCK\_NOTSUPP. If the upgrade can be achieved without an existing conflict, the request will succeed. Otherwise, the server will return either NFS4ERR\_DENIED or NFS4ERR\_DEADLOCK. The error NFS4ERR\_DEADLOCK is returned if the client issued the LOCK request with the type set to WRITEW\_LT and the server has detected a deadlock. The client should be prepared to receive such errors and if appropriate, report the error to the requesting application.

### 9.4. Blocking Locks

Some clients require the support of blocking locks. The NFS version 4 protocol must not rely on a callback mechanism and therefore is unable to notify a client when a previously denied lock has been granted. Clients have no choice but to continually poll for the lock. This presents a fairness problem. Two new lock types are added, READW and WRITEW, and are used to indicate to the server that the client is requesting a blocking lock. The server should maintain an ordered list of pending blocking locks. When the conflicting lock is released, the server may wait the lease period for the first waiting client to re-request the lock. After the lease period expires the next waiting client request is allowed the lock. Clients are required to poll at an interval sufficiently small that it is likely to acquire the lock in a timely manner. The server is not required to maintain a list of pending blocked locks as it is not used to provide correct operation but only to increase fairness. Because of the unordered nature of crash recovery, storing of lock state to stable storage would be required to guarantee ordered granting of blocking locks.

Servers may also note the lock types and delay returning denial of the request to allow extra time for a conflicting lock to be

released, allowing a successful return. In this way, clients can avoid the burden of needlessly frequent polling for blocking locks. The server should take care in the length of delay in the event the client retransmits the request.

If a server receives a blocking lock request, denies it, and then later receives a nonblocking request for the same lock, which is also denied, then it should remove the lock in question from its list of pending blocking locks. Clients should use such a nonblocking request to indicate to the server that this is the last time they intend to poll for the lock, as may happen when the process requesting the lock is interrupted. This is a courtesy to the server, to prevent it from unnecessarily waiting a lease period before granting other lock requests. However, clients are not required to perform this courtesy, and servers must not depend on them doing so. Also, clients must be prepared for the possibility that this final locking request will be accepted.

#### 9.5. Lease Renewal

The purpose of a lease is to allow a server to remove stale locks that are held by a client that has crashed or is otherwise unreachable. It is not a mechanism for cache consistency and lease renewals may not be denied if the lease interval has not expired.

The client can implicitly provide a positive indication that it is still active and that the associated state held at the server, for the client, is still valid. Any operation made with a valid clientid (DELEGPURGE, LOCK, LOCKT, OPEN, RELEASE\_LOCKOWNER, or RENEW) or a valid stateid (CLOSE, DELEGRETURN, LOCK, LOCKU, OPEN, OPEN\_CONFIRM, OPEN\_DOWNGRADE, READ, SETATTR, or WRITE) informs the server to renew all of the leases for that client (i.e., all those sharing a given client ID). In the latter case, the stateid must not be one of the special stateids consisting of all bits 0 or all bits 1.

Note that if the client had restarted or rebooted, the client would not be making these requests without issuing the SETCLIENTID/SETCLIENTID\_CONFIRM sequence. The use of the SETCLIENTID/SETCLIENTID\_CONFIRM sequence (one that changes the client verifier) notifies the server to drop the locking state associated with the client. SETCLIENTID/SETCLIENTID\_CONFIRM never renews a lease.

If the server has rebooted, the stateids (NFS4ERR\_STALE\_STATEID error) or the client ID (NFS4ERR\_STALE\_CLIENTID error) will not be valid hence preventing spurious renewals.

This approach allows for low overhead lease renewal which scales well. In the typical case no extra RPC calls are required for lease

renewal and in the worst case one RPC is required every lease period (i.e., a RENEW operation). The number of locks held by the client is not a factor since all state for the client is involved with the lease renewal action.

Since all operations that create a new lease also renew existing leases, the server must maintain a common lease expiration time for all valid leases for a given client. This lease time can then be easily updated upon implicit lease renewal actions.

## 9.6. Crash Recovery

The important requirement in crash recovery is that both the client and the server know when the other has failed. Additionally, it is required that a client sees a consistent view of data across server restarts or reboots. All READ and WRITE operations that may have been queued within the client or network buffers must wait until the client has successfully recovered the locks protecting the READ and WRITE operations.

### 9.6.1. Client Failure and Recovery

In the event that a client fails, the server may recover the client's locks when the associated leases have expired. Conflicting locks from another client may only be granted after this lease expiration. If the client is able to restart or reinitialize within the lease period the client may be forced to wait the remainder of the lease period before obtaining new locks.

To minimize client delay upon restart, open and lock requests are associated with an instance of the client by a client supplied verifier. This verifier is part of the initial SETCLIENTID call made by the client. The server returns a client ID as a result of the SETCLIENTID operation. The client then confirms the use of the client ID with SETCLIENTID\_CONFIRM. The client ID in combination with an opaque owner field is then used by the client to identify the open owner for OPEN. This chain of associations is then used to identify all locks for a particular client.

Since the verifier will be changed by the client upon each initialization, the server can compare a new verifier to the verifier associated with currently held locks and determine that they do not match. This signifies the client's new instantiation and subsequent loss of locking state. As a result, the server is free to release all locks held which are associated with the old client ID which was derived from the old verifier.

Note that the verifier must have the same uniqueness properties of

the verifier for the COMMIT operation.

#### 9.6.2. Server Failure and Recovery

If the server loses locking state (usually as a result of a restart or reboot), it must allow clients time to discover this fact and re-establish the lost locking state. The client must be able to re-establish the locking state without having the server deny valid requests because the server has granted conflicting access to another client. Likewise, if there is the possibility that clients have not yet re-established their locking state for a file, the server must disallow READ and WRITE operations for that file. The duration of this recovery period is equal to the duration of the lease period.

A client can determine that server failure (and thus loss of locking state) has occurred, when it receives one of two errors. The NFS4ERR\_STALE\_STATEID error indicates a stateid invalidated by a reboot or restart. The NFS4ERR\_STALE\_CLIENTID error indicates a client ID invalidated by reboot or restart. When either of these are received, the client must establish a new client ID (see Section 9.1.1) and re-establish the locking state as discussed below.

The period of special handling of locking and READs and WRITEs, equal in duration to the lease period, is referred to as the "grace period". During the grace period, clients recover locks and the associated state by reclaim-type locking requests (i.e., LOCK requests with reclaim set to true and OPEN operations with a claim type of either CLAIM\_PREVIOUS or CLAIM\_DELEGATE\_PREV). During the grace period, the server must reject READ and WRITE operations and non-reclaim locking requests (i.e., other LOCK and OPEN operations) with an error of NFS4ERR\_GRACE.

If the server can reliably determine that granting a non-reclaim request will not conflict with reclamation of locks by other clients, the NFS4ERR\_GRACE error does not have to be returned and the non-reclaim client request can be serviced. For the server to be able to service READ and WRITE operations during the grace period, it must again be able to guarantee that no possible conflict could arise between an impending reclaim locking request and the READ or WRITE operation. If the server is unable to offer that guarantee, the NFS4ERR\_GRACE error must be returned to the client.

For a server to provide simple, valid handling during the grace period, the easiest method is to simply reject all non-reclaim locking requests and READ and WRITE operations by returning the NFS4ERR\_GRACE error. However, a server may keep information about granted locks in stable storage. With this information, the server could determine if a regular lock or READ or WRITE operation can be

safely processed.

For example, if a count of locks on a given file is available in stable storage, the server can track reclaimed locks for the file and when all reclaims have been processed, non-reclaim locking requests may be processed. This way the server can ensure that non-reclaim locking requests will not conflict with potential reclaim requests. With respect to I/O requests, if the server is able to determine that there are no outstanding reclaim requests for a file by information from stable storage or another similar mechanism, the processing of I/O requests could proceed normally for the file.

To reiterate, for a server that allows non-reclaim lock and I/O requests to be processed during the grace period, it **MUST** determine that no lock subsequently reclaimed will be rejected and that no lock subsequently reclaimed would have prevented any I/O operation processed during the grace period.

Clients should be prepared for the return of NFS4ERR\_GRACE errors for non-reclaim lock and I/O requests. In this case the client should employ a retry mechanism for the request. A delay (on the order of several seconds) between retries should be used to avoid overwhelming the server. Further discussion of the general issue is included in [Floyd]. The client must account for the server that is able to perform I/O and non-reclaim locking requests within the grace period as well as those that cannot do so.

A reclaim-type locking request outside the server's grace period can only succeed if the server can guarantee that no conflicting lock or I/O request has been granted since reboot or restart.

A server may, upon restart, establish a new value for the lease period. Therefore, clients should, once a new client ID is established, refetch the lease\_time attribute and use it as the basis for lease renewal for the lease associated with that server. However, the server must establish, for this restart event, a grace period at least as long as the lease period for the previous server instantiation. This allows the client state obtained during the previous server instance to be reliably re-established.

#### 9.6.3. Network Partitions and Recovery

If the duration of a network partition is greater than the lease period provided by the server, the server will have not received a lease renewal from the client. If this occurs, the server may cancel the lease and free all locks held for the client. As a result, all stateids held by the client will become invalid or stale. Once the client is able to reach the server after such a network partition,

all I/O submitted by the client with the now invalid stateids will fail with the server returning the error NFS4ERR\_EXPIRED. Once this error is received, the client will suitably notify the application that held the lock.

#### 9.6.3.1. Courtesy Locks

As a courtesy to the client or as an optimization, the server may continue to hold locks, including delegations, on behalf of a client for which recent communication has extended beyond the lease period, delaying the cancellation of the lease. If the server receives a lock or I/O request that conflicts with one of these courtesy locks or if it runs out of resources, the server MAY cause lease cancellation to occur at that time and henceforth return NFS4ERR\_EXPIRED when any of the stateids associated with the freed locks is used. If lease cancellation has not occurred and the server receives a lock or I/O request that conflicts with one of the courtesy locks, the requirements are as follows:

- o In the case of a courtesy lock which is not a delegation, it MUST free the courtesy lock and grant the new request.
- o In the case of lock or IO request which conflicts with a delegation which is being held as courtesy lock, the server MAY delay resolution of request but MUST NOT reject the request and MUST free the delegation and grant the new request eventually.
- o In the case of a requests for a delegation which conflicts with a delegation which is being held as courtesy lock, the server MAY grant the new request or not as it chooses, but if it grants the conflicting request, the delegation held as courtesy lock MUST be freed.

If the server does not reboot or cancel the lease before the network partition is healed, when the original client tries to access a courtesy lock which was freed, the server SHOULD send back a NFS4ERR\_BAD\_STATEID to the client. If the client tries to access a courtesy lock which was not freed, then the server SHOULD mark all of the courtesy locks as implicitly being renewed.

#### 9.6.3.2. Lease Cancellation

As a result of lease expiration, leases may be cancelled, either immediately upon expiration or subsequently, depending on the occurrence of a conflicting lock or extension of the period of partition beyond what the server will tolerate.

When a lease is cancelled, all locking state associated with it is

freed and use of any the associated stateids will result in NFS4ERR\_EXPIRED being returned. Similarly, use of the associated clientid will result in NFS4ERR\_EXPIRED being returned.

The client should recover from this situation by using SETCLIENTID followed by SETCLIENTID\_CONFIRM, in order to establish a new clientid. Once a lock is obtained using this clientid, a lease will be established.

#### 9.6.3.3. Client's Reaction to a Freed Lock

There is no way for a client to predetermine how a given server is going to behave during a network partition. When the partition heals, either the client still has all of its locks, it has some of its locks, or it has none of them. The client will be able to examine the various error return values to determine its response.

##### NFS4ERR\_EXPIRED:

All locks have been freed as a result of a lease cancellation which occurred during the partition. The client should use a SETCLIENTID to recover.

##### NFS4ERR\_ADMIN\_REVOKED:

The current lock has been revoked before, during, or after the partition. The client SHOULD handle this error as it normally would.

##### NFS4ERR\_BAD\_STATEID:

The current lock has been revoked/released during the partition and the server did not reboot. Other locks MAY still be renewed. The client need not do a SETCLIENTID and instead SHOULD probe via a RENEW call.

##### NFS4ERR\_RECLAIM\_BAD:

The current lock has been revoked during the partition and the server rebooted. The server might have no information on the other locks. They may still be renewable.

##### NFS4ERR\_NO\_GRACE:

The client's locks have been revoked during the partition and the server rebooted. None of the client's locks will be renewable.



**NFS4ERR\_OLD\_STATEID:**

The server has not rebooted. The client SHOULD handle this error as it normally would.

**9.6.3.4. Edge Conditions**

When a network partition is combined with a server reboot, then both the server and client have responsibilities to ensure that the client does not reclaim a lock which it should no longer be able to access. Briefly those are:

- o Client's responsibility: A client MUST NOT attempt to reclaim any locks which it did not hold at the end of its most recent successfully established client lease.
- o Server's responsibility: A server MUST NOT allow a client to reclaim a lock unless it knows that it could not have since granted a conflicting lock. However, in deciding whether a conflicting lock could have been granted, it is permitted to assume its clients are responsible, as above.

A server may consider a client's lease "successfully established" once it has received an open operation from that client.

The above are directed to CLAIM\_PREVIOUS reclaims and not to CLAIM\_DELEGATE\_PREV reclaims, which generally do not involve a server reboot. However, when a server persistently stores delegation information to support CLAIM\_DELEGATE\_PREV across a period in which both client and server are down at the same time, similar strictures apply.

The next sections give examples showing what can go wrong if these responsibilities are neglected, and provides examples of server implementation strategies that could meet a server's responsibilities.

**9.6.3.4.1. First Server Edge Condition**

The first edge condition has the following scenario:

1. Client A acquires a lock.
2. Client A and server experience mutual network partition, such that client A is unable to renew its lease.
3. Client A's lease expires, so server releases lock.

4. Client B acquires a lock that would have conflicted with that of Client A.
5. Client B releases the lock
6. Server reboots
7. Network partition between client A and server heals.
8. Client A issues a RENEW operation, and gets back a NFS4ERR\_STALE\_CLIENTID.
9. Client A reclaims its lock within the server's grace period.

Thus, at the final step, the server has erroneously granted client A's lock reclaim. If client B modified the object the lock was protecting, client A will experience object corruption.

#### 9.6.3.4.2. Second Server Edge Condition

The second known edge condition follows:

1. Client A acquires a lock.
2. Server reboots.
3. Client A and server experience mutual network partition, such that client A is unable to reclaim its lock within the grace period.
4. Server's reclaim grace period ends. Client A has no locks recorded on server.
5. Client B acquires a lock that would have conflicted with that of Client A.
6. Client B releases the lock.
7. Server reboots a second time.
8. Network partition between client A and server heals.
9. Client A issues a RENEW operation, and gets back a NFS4ERR\_STALE\_CLIENTID.
10. Client A reclaims its lock within the server's grace period.

As with the first edge condition, the final step of the scenario of

the second edge condition has the server erroneously granting client A's lock reclaim.

#### 9.6.3.4.3. Handling Server Edge Conditions

In both of the above examples, the client attempts reclaim of a lock that it held at the end of its most recent successfully established lease; thus, it has fulfilled its responsibility.

The server, however, has failed, by granting a reclaim, despite having granted a conflicting lock since the reclaimed lock was last held.

Solving these edge conditions requires that the server either assume after it reboots that edge condition occurs, and thus return NFS4ERR\_NO\_GRACE for all reclaim attempts, or that the server record some information in stable storage. The amount of information the server records in stable storage is in inverse proportion to how harsh the server wants to be whenever the edge conditions occur. The server that is completely tolerant of all edge conditions will record in stable storage every lock that is acquired, removing the lock record from stable storage only when the lock is unlocked by the client and the lock's owner advances the sequence number such that the lock release is not the last stateful event for the owner's sequence. For the two aforementioned edge conditions, the harshest a server can be, and still support a grace period for reclaims, requires that the server record in stable storage information some minimal information. For example, a server implementation could, for each client, save in stable storage a record containing:

- o the client's id string
- o a boolean that indicates if the client's lease expired or if there was administrative intervention (see Section 9.8) to revoke a byte-range lock, share reservation, or delegation
- o a timestamp that is updated the first time after a server boot or reboot the client acquires byte-range locking, share reservation, or delegation state on the server. The timestamp need not be updated on subsequent lock requests until the server reboots.

The server implementation would also record in the stable storage the timestamps from the two most recent server reboots.

Assuming the above record keeping, for the first edge condition, after the server reboots, the record that client A's lease expired means that another client could have acquired a conflicting record lock, share reservation, or delegation. Hence the server must reject

a reclaim from client A with the error NFS4ERR\_NO\_GRACE or NFS4ERR\_RECLAIM\_BAD.

For the second edge condition, after the server reboots for a second time, the record that the client had an unexpired record lock, share reservation, or delegation established before the server's previous incarnation means that the server must reject a reclaim from client A with the error NFS4ERR\_NO\_GRACE or NFS4ERR\_RECLAIM\_BAD.

Regardless of the level and approach to record keeping, the server MUST implement one of the following strategies (which apply to reclaims of share reservations, byte-range locks, and delegations):

1. Reject all reclaims with NFS4ERR\_NO\_GRACE. This is super harsh, but necessary if the server does not want to record lock state in stable storage.
2. Record sufficient state in stable storage to meet its responsibilities. In doubt, the server should err on the side of being harsh.

In the event that, after a server reboot, the server determines that there is unrecoverable damage or corruption to the the stable storage, then for all clients and/or locks affected, the server MUST return NFS4ERR\_NO\_GRACE.

#### 9.6.3.4.4. Client Edge Condition

A third edge condition effects the client and not the server. If the server reboots in the middle of the client reclaiming some locks and then a network partition is established, the client might be in the situation of having reclaimed some, but not all locks. In that case, a conservative client would assume that the non-reclaimed locks were revoked.

The third known edge condition follows:

1. Client A acquires a lock 1.
2. Client A acquires a lock 2.
3. Server reboots.
4. Client A issues a RENEW operation, and gets back a NFS4ERR\_STALE\_CLIENTID.
5. Client A reclaims its lock 1 within the server's grace period.

6. Client A and server experience mutual network partition, such that client A is unable to reclaim its remaining locks within the grace period.
7. Server's reclaim grace period ends.
8. Client B acquires a lock that would have conflicted with Client A's lock 2.
9. Client B releases the lock.
10. Server reboots a second time.
11. Network partition between client A and server heals.
12. Client A issues a RENEW operation, and gets back a NFS4ERR\_STALE\_CLIENTID.
13. Client A reclaims both lock 1 and lock 2 within the server's grace period.

At the last step, the client reclaims lock 2 as if it had held that lock continuously, when in fact a conflicting lock was granted to client B.

This occurs because the client failed its responsibility, by attempting to reclaim lock 2 even though it had not held that lock at the end of the lease that was established by the SETCLIENTID after the first server reboot. (The client did hold lock 2 on a previous lease. But it is only the most recent lease that matters.)

A server could avoid this situation by rejecting the reclaim of lock 2. However, to do so accurately it would have to ensure that additional information about individual locks held survives reboot. Server implementations are not required to do that, so the client must not assume that the server will.

Instead, a client MUST reclaim only those locks which it successfully acquired from the previous server instance, omitting any that it failed to reclaim before a new reboot. Thus, in the last step above, client A should reclaim only lock 1.

#### 9.6.3.4.5. Client's Handling of Reclaim Errors

A mandate for the client's handling of the NFS4ERR\_NO\_GRACE and NFS4ERR\_RECLAIM\_BAD errors is outside the scope of this specification, since the strategies for such handling are very dependent on the client's operating environment. However, one

potential approach is described below.

When the client's reclaim fails, it could examine the change attribute of the objects the client is trying to reclaim state for, and use that to determine whether to re-establish the state via normal OPEN or LOCK requests. This is acceptable provided the client's operating environment allows it. In other words, the client implementor is advised to document for his users the behavior. The client could also inform the application that its byte-range lock or share reservations (whether they were delegated or not) have been lost, such as via a UNIX signal, a GUI pop-up window, etc. See Section 10.5, for a discussion of what the client should do for dealing with unreclaimed delegations on client state.

For further discussion of revocation of locks see Section 9.8.

#### 9.7. Recovery from a Lock Request Timeout or Abort

In the event a lock request times out, a client may decide to not retry the request. The client may also abort the request when the process for which it was issued is terminated (e.g., in UNIX due to a signal). It is possible though that the server received the request and acted upon it. This would change the state on the server without the client being aware of the change. It is paramount that the client re-synchronize state with server before it attempts any other operation that takes a seqid and/or a stateid with the same state-owner. This is straightforward to do without a special re-synchronize operation.

Since the server maintains the last lock request and response received on the state-owner, for each state-owner, the client should cache the last lock request it sent such that the lock request did not receive a response. From this, the next time the client does a lock operation for the state-owner, it can send the cached request, if there is one, and if the request was one that established state (e.g., a LOCK or OPEN operation), the server will return the cached result or if never saw the request, perform it. The client can follow up with a request to remove the state (e.g., a LOCKU or CLOSE operation). With this approach, the sequencing and stateid information on the client and server for the given state-owner will re-synchronize and in turn the lock state will re-synchronize.

#### 9.8. Server Revocation of Locks

At any point, the server can revoke locks held by a client and the client must be prepared for this event. When the client detects that its locks have been or may have been revoked, the client is responsible for validating the state information between itself and

the server. Validating locking state for the client means that it must verify or reclaim state for each lock currently held.

The first instance of lock revocation is upon server reboot or re-initialization. In this instance the client will receive an error (NFS4ERR\_STALE\_STATEID or NFS4ERR\_STALE\_CLIENTID) and the client will proceed with normal crash recovery as described in the previous section.

The second lock revocation event is the inability to renew the lease before expiration. While this is considered a rare or unusual event, the client must be prepared to recover. Both the server and client will be able to detect the failure to renew the lease and are capable of recovering without data corruption. For the server, it tracks the last renewal event serviced for the client and knows when the lease will expire. Similarly, the client must track operations which will renew the lease period. Using the time that each such request was sent and the time that the corresponding reply was received, the client should bound the time that the corresponding renewal could have occurred on the server and thus determine if it is possible that a lease period expiration could have occurred.

The third lock revocation event can occur as a result of administrative intervention within the lease period. While this is considered a rare event, it is possible that the server's administrator has decided to release or revoke a particular lock held by the client. As a result of revocation, the client will receive an error of NFS4ERR\_ADMIN\_REVOKED. In this instance the client may assume that only the state-owner's locks have been lost. The client notifies the lock holder appropriately. The client MUST NOT assume the lease period has been renewed as a result of a failed operation.

When the client determines the lease period may have expired, the client must mark all locks held for the associated lease as "unvalidated". This means the client has been unable to re-establish or confirm the appropriate lock state with the server. As described in Section 9.6, there are scenarios in which the server may grant conflicting locks after the lease period has expired for a client. When it is possible that the lease period has expired, the client must validate each lock currently held to ensure that a conflicting lock has not been granted. The client may accomplish this task by issuing an I/O request, either a pending I/O or a zero-length read, specifying the stateid associated with the lock in question. If the response to the request is success, the client has validated all of the locks governed by that stateid and re-established the appropriate state between itself and the server.

If the I/O request is not successful, then one or more of the locks

associated with the stateid was revoked by the server and the client must notify the owner.

### 9.9. Share Reservations

A share reservation is a mechanism to control access to a file. It is a separate and independent mechanism from byte-range locking. When a client opens a file, it issues an OPEN operation to the server specifying the type of access required (READ, WRITE, or BOTH) and the type of access to deny others (OPEN4\_SHARE\_DENY\_NONE, OPEN4\_SHARE\_DENY\_READ, OPEN4\_SHARE\_DENY\_WRITE, or OPEN4\_SHARE\_DENY\_BOTH). If the OPEN fails the client will fail the application's open request.

Pseudo-code definition of the semantics:

```
if (request.access == 0)
    return (NFS4ERR_INVALID)
else if ((request.access & file_state.deny) ||
        (request.deny & file_state.access))
    return (NFS4ERR_DENIED)
```

This checking of share reservations on OPEN is done with no exception for an existing OPEN for the same open-owner.

The constants used for the OPEN and OPEN\_DOWNGRADE operations for the access and deny fields are as follows:

```
const OPEN4_SHARE_ACCESS_READ    = 0x00000001;
const OPEN4_SHARE_ACCESS_WRITE   = 0x00000002;
const OPEN4_SHARE_ACCESS_BOTH    = 0x00000003;

const OPEN4_SHARE_DENY_NONE      = 0x00000000;
const OPEN4_SHARE_DENY_READ      = 0x00000001;
const OPEN4_SHARE_DENY_WRITE     = 0x00000002;
const OPEN4_SHARE_DENY_BOTH      = 0x00000003;
```

### 9.10. OPEN/CLOSE Operations

To provide correct share semantics, a client MUST use the OPEN operation to obtain the initial filehandle and indicate the desired access and what access, if any, to deny. Even if the client intends to use a stateid of all 0's or all 1's, it must still obtain the filehandle for the regular file with the OPEN operation so the appropriate share semantics can be applied. Clients that do not have a deny mode built into their programming interfaces for opening a file should request a deny mode of OPEN4\_SHARE\_DENY\_NONE.



The OPEN operation with the CREATE flag, also subsumes the CREATE operation for regular files as used in previous versions of the NFS protocol. This allows a create with a share to be done atomically.

The CLOSE operation removes all share reservations held by the open-owner on that file. If byte-range locks are held, the client SHOULD release all locks before issuing a CLOSE. The server MAY free all outstanding locks on CLOSE but some servers may not support the CLOSE of a file that still has byte-range locks held. The server MUST return failure, NFS4ERR\_LOCKS\_HELD, if any locks would exist after the CLOSE.

The LOOKUP operation will return a filehandle without establishing any lock state on the server. Without a valid stateid, the server will assume the client has the least access. For example, if one client opened a file with OPEN4\_SHARE\_DENY\_BOTH and another client accesses the file via a filehandle obtained through LOOKUP, the second client could only read the file using the special read bypass stateid. The second client could not WRITE the file at all because it would not have a valid stateid from OPEN and the special anonymous stateid would not be allowed access.

#### 9.10.1. Close and Retention of State Information

Since a CLOSE operation requests deallocation of a stateid, dealing with retransmission of the CLOSE, may pose special difficulties, since the state information, which normally would be used to determine the state of the open file being designated, might be deallocated, resulting in an NFS4ERR\_BAD\_STATEID error.

Servers may deal with this problem in a number of ways. To provide the greatest degree assurance that the protocol is being used properly, a server should, rather than deallocate the stateid, mark it as close-pending, and retain the stateid with this status, until later deallocation. In this way, a retransmitted CLOSE can be recognized since the stateid points to state information with this distinctive status, so that it can be handled without error.

When adopting this strategy, a server should retain the state information until the earliest of:

- o Another validly sequenced request for the same open-owner, that is not a retransmission.
- o The time that an open-owner is freed by the server due to period with no activity.

- o All locks for the client are freed as a result of a SETCLIENTID.

Servers may avoid this complexity, at the cost of less complete protocol error checking, by simply responding NFS4\_OK in the event of a CLOSE for a deallocated stateid, on the assumption that this case must be caused by a retransmitted close. When adopting this approach, it is desirable to at least log an error when returning a no-error indication in this situation. If the server maintains a reply-cache mechanism, it can verify the CLOSE is indeed a retransmission and avoid error logging in most cases.

#### 9.11. Open Upgrade and Downgrade

When an OPEN is done for a file and the open-owner for which the open is being done already has the file open, the result is to upgrade the open file status maintained on the server to include the access and deny bits specified by the new OPEN as well as those for the existing OPEN. The result is that there is one open file, as far as the protocol is concerned, and it includes the union of the access and deny bits for all of the OPEN requests completed. Only a single CLOSE will be done to reset the effects of both OPENS. Note that the client, when issuing the OPEN, may not know that the same file is in fact being opened. The above only applies if both OPENS result in the OPENed object being designated by the same filehandle.

When the server chooses to export multiple filehandles corresponding to the same file object and returns different filehandles on two different OPENS of the same file object, the server MUST NOT "OR" together the access and deny bits and coalesce the two open files. Instead the server must maintain separate OPENS with separate stateids and will require separate CLOSEs to free them.

When multiple open files on the client are merged into a single open file object on the server, the close of one of the open files (on the client) may necessitate change of the access and deny status of the open file on the server. This is because the union of the access and deny bits for the remaining opens may be smaller (i.e., a proper subset) than previously. The OPEN\_DOWNGRADE operation is used to make the necessary change and the client should use it to update the server so that share reservation requests by other clients are handled properly. The stateid returned has the same "other" field as that passed to the server. The "seqid" value in the returned stateid MUST be incremented, even in situations in which there is no change to the access and deny bits for the file.

### 9.12. Short and Long Leases

When determining the time period for the server lease, the usual lease tradeoffs apply. Short leases are good for fast server recovery at a cost of increased RENEW or READ (with zero length) requests. Longer leases are certainly kinder and gentler to servers trying to handle very large numbers of clients. The number of RENEW requests drop in proportion to the lease time. The disadvantages of long leases are slower recovery after server failure (the server must wait for the leases to expire and the grace period to elapse before granting new lock requests) and increased file contention (if client fails to transmit an unlock request then server must wait for lease expiration before granting new locks).

Long leases are usable if the server is able to store lease state in non-volatile memory. Upon recovery, the server can reconstruct the lease state from its non-volatile memory and continue operation with its clients and therefore long leases would not be an issue.

### 9.13. Clocks, Propagation Delay, and Calculating Lease Expiration

To avoid the need for synchronized clocks, lease times are granted by the server as a time delta. However, there is a requirement that the client and server clocks do not drift excessively over the duration of the lock. There is also the issue of propagation delay across the network which could easily be several hundred milliseconds as well as the possibility that requests will be lost and need to be retransmitted.

To take propagation delay into account, the client should subtract it from lease times (e.g., if the client estimates the one-way propagation delay as 200 msec, then it can assume that the lease is already 200 msec old when it gets it). In addition, it will take another 200 msec to get a response back to the server. So the client must send a lock renewal or write data back to the server 400 msec before the lease would expire.

The server's lease period configuration should take into account the network distance of the clients that will be accessing the server's resources. It is expected that the lease period will take into account the network propagation delays and other network delay factors for the client population. Since the protocol does not allow for an automatic method to determine an appropriate lease period, the server's administrator may have to tune the lease period.

#### 9.14. Migration, Replication and State

When responsibility for handling a given file system is transferred to a new server (migration) or the client chooses to use an alternate server (e.g., in response to server unresponsiveness) in the context of file system replication, the appropriate handling of state shared between the client and server (i.e., locks, leases, stateids, and client IDs) is as described below. The handling differs between migration and replication. For related discussion of file server state and recover of such see the sections under Section 9.6.

If a server replica or a server immigrating a file system agrees to, or is expected to, accept opaque values from the client that originated from another server, then servers SHOULD encode the "opaque" values in network byte order. This way, servers acting as replicas or immigrating file systems will be able to parse values like stateids, directory cookies, filehandles, etc. even if their native byte order is different from other servers cooperating in the replication and migration of the file system.

##### 9.14.1. Migration and State

In the case of migration, the servers involved in the migration of a file system SHOULD transfer all server state from the original to the new server. This must be done in a way that is transparent to the client. This state transfer will ease the client's transition when a file system migration occurs. If the servers are successful in transferring all state, the client will continue to use stateids assigned by the original server. Therefore the new server must recognize these stateids as valid. This holds true for the client ID as well. Since responsibility for an entire file system is transferred with a migration event, there is no possibility that conflicts will arise on the new server as a result of the transfer of locks.

As part of the transfer of information between servers, leases would be transferred as well. The leases being transferred to the new server will typically have a different expiration time from those for the same client, previously on the old server. To maintain the property that all leases on a given server for a given client expire at the same time, the server should advance the expiration time to the later of the leases being transferred or the leases already present. This allows the client to maintain lease renewal of both classes without special effort.

The servers may choose not to transfer the state information upon migration. However, this choice is discouraged. In this case, when the client presents state information from the original server (e.g.,

in a RENEW op or a READ op of zero length), the client must be prepared to receive either NFS4ERR\_STALE\_CLIENTID or NFS4ERR\_STALE\_STATEID from the new server. The client should then recover its state information as it normally would in response to a server failure. The new server must take care to allow for the recovery of state information as it would in the event of server restart.

A client SHOULD re-establish new callback information with the new server as soon as possible, according to sequences described in Section 15.35 and Section 15.36. This ensures that server operations are not blocked by the inability to recall delegations.

#### 9.14.2. Replication and State

Since client switch-over in the case of replication is not under server control, the handling of state is different. In this case, leases, stateids and client IDs do not have validity across a transition from one server to another. The client must re-establish its locks on the new server. This can be compared to the re-establishment of locks by means of reclaim-type requests after a server reboot. The difference is that the server has no provision to distinguish requests reclaiming locks from those obtaining new locks or to defer the latter. Thus, a client re-establishing a lock on the new server (by means of a LOCK or OPEN request), may have the requests denied due to a conflicting lock. Since replication is intended for read-only use of file systems, such denial of locks should not pose large difficulties in practice. When an attempt to re-establish a lock on a new server is denied, the client should treat the situation as if his original lock had been revoked.

#### 9.14.3. Notification of Migrated Lease

In the case of lease renewal, the client may not be submitting requests for a file system that has been migrated to another server. This can occur because of the implicit lease renewal mechanism. The client renews leases for all file systems when submitting a request to any one file system at the server.

In order for the client to schedule renewal of leases that may have been relocated to the new server, the client must find out about lease relocation before those leases expire. To accomplish this, all operations which implicitly renew leases for a client (such as OPEN, CLOSE, READ, WRITE, RENEW, LOCK, and others), will return the error NFS4ERR\_LEASE\_MOVED if responsibility for any of the leases to be renewed has been transferred to a new server. This condition will continue until the client receives an NFS4ERR\_MOVED error and the server receives the subsequent GETATTR(fs\_locations) for an access to

each file system for which a lease has been moved to a new server. By convention, the compound including the GETATTR(fs\_locations) SHOULD append a RENEW operation to permit the server to identify the client doing the access.

Upon receiving the NFS4ERR\_LEASE\_MOVED error, a client that supports file system migration MUST probe all file systems from that server on which it holds open state. Once the client has successfully probed all those file systems which are migrated, the server MUST resume normal handling of stateful requests from that client.

In order to support legacy clients that do not handle the NFS4ERR\_LEASE\_MOVED error correctly, the server SHOULD time out after a wait of at least two lease periods, at which time it will resume normal handling of stateful requests from all clients. If a client attempts to access the migrated files, the server MUST reply NFS4ERR\_MOVED.

When the client receives an NFS4ERR\_MOVED error, the client can follow the normal process to obtain the new server information (through the fs\_locations attribute) and perform renewal of those leases on the new server. If the server has not had state transferred to it transparently, the client will receive either NFS4ERR\_STALE\_CLIENTID or NFS4ERR\_STALE\_STATEID from the new server, as described above. The client can then recover state information as it does in the event of server failure.

#### 9.14.4. Migration and the Lease\_time Attribute

In order that the client may appropriately manage its leases in the case of migration, the destination server must establish proper values for the lease\_time attribute.

When state is transferred transparently, that state should include the correct value of the lease\_time attribute. The lease\_time attribute on the destination server must never be less than that on the source since this would result in premature expiration of leases granted by the source server. Upon migration in which state is transferred transparently, the client is under no obligation to re-fetch the lease\_time attribute and may continue to use the value previously fetched (on the source server).

If state has not been transferred transparently (i.e., the client sees a real or simulated server reboot), the client should fetch the value of lease\_time on the new (i.e., destination) server, and use it for subsequent locking requests. However the server must respect a grace period at least as long as the lease\_time on the source server, in order to ensure that clients have ample time to reclaim their

locks before potentially conflicting non-reclaimed locks are granted. The means by which the new server obtains the value of `lease_time` on the old server is left to the server implementations. It is not specified by the NFS version 4 protocol.

## 10. Client-Side Caching

Client-side caching of data, of file attributes, and of file names is essential to providing good performance with the NFS protocol. Providing distributed cache coherence is a difficult problem and previous versions of the NFS protocol have not attempted it. Instead, several NFS client implementation techniques have been used to reduce the problems that a lack of coherence poses for users. These techniques have not been clearly defined by earlier protocol specifications and it is often unclear what is valid or invalid client behavior.

The NFSv4 protocol uses many techniques similar to those that have been used in previous protocol versions. The NFSv4 protocol does not provide distributed cache coherence. However, it defines a more limited set of caching guarantees to allow locks and share reservations to be used without destructive interference from client side caching.

In addition, the NFSv4 protocol introduces a delegation mechanism which allows many decisions normally made by the server to be made locally by clients. This mechanism provides efficient support of the common cases where sharing is infrequent or where sharing is read-only.

### 10.1. Performance Challenges for Client-Side Caching

Caching techniques used in previous versions of the NFS protocol have been successful in providing good performance. However, several scalability challenges can arise when those techniques are used with very large numbers of clients. This is particularly true when clients are geographically distributed which classically increases the latency for cache re-validation requests.

The previous versions of the NFS protocol repeat their file data cache validation requests at the time the file is opened. This behavior can have serious performance drawbacks. A common case is one in which a file is only accessed by a single client. Therefore, sharing is infrequent.

In this case, repeated reference to the server to find that no conflicts exist is expensive. A better option with regards to

performance is to allow a client that repeatedly opens a file to do so without reference to the server. This is done until potentially conflicting operations from another client actually occur.

A similar situation arises in connection with file locking. Sending file lock and unlock requests to the server as well as the read and write requests necessary to make data caching consistent with the locking semantics (see Section 10.3.2) can severely limit performance. When locking is used to provide protection against infrequent conflicts, a large penalty is incurred. This penalty may discourage the use of file locking by applications.

The NFSv4 protocol provides more aggressive caching strategies with the following design goals:

- o Compatibility with a large range of server semantics.
- o Provide the same caching benefits as previous versions of the NFS protocol when unable to provide the more aggressive model.
- o Requirements for aggressive caching are organized so that a large portion of the benefit can be obtained even when not all of the requirements can be met.

The appropriate requirements for the server are discussed in later sections in which specific forms of caching are covered (see Section 10.4).

## 10.2. Delegation and Callbacks

Recallable delegation of server responsibilities for a file to a client improves performance by avoiding repeated requests to the server in the absence of inter-client conflict. With the use of a "callback" RPC from server to client, a server recalls delegated responsibilities when another client engages in sharing of a delegated file.

A delegation is passed from the server to the client, specifying the object of the delegation and the type of delegation. There are different types of delegations but each type contains a stateid to be used to represent the delegation when performing operations that depend on the delegation. This stateid is similar to those associated with locks and share reservations but differs in that the stateid for a delegation is associated with a client ID and may be used on behalf of all the open-owners for the given client. A delegation is made to the client as a whole and not to any specific process or thread of control within it.



Because callback RPCs may not work in all environments (due to firewalls, for example), correct protocol operation does not depend on them. Preliminary testing of callback functionality by means of a CB\_NULL procedure determines whether callbacks can be supported. The CB\_NULL procedure checks the continuity of the callback path. A server makes a preliminary assessment of callback availability to a given client and avoids delegating responsibilities until it has determined that callbacks are supported. Because the granting of a delegation is always conditional upon the absence of conflicting access, clients must not assume that a delegation will be granted and they must always be prepared for OPENS to be processed without any delegations being granted.

Once granted, a delegation behaves in most ways like a lock. There is an associated lease that is subject to renewal together with all of the other leases held by that client.

Unlike locks, an operation by a second client to a delegated file will cause the server to recall a delegation through a callback.

On recall, the client holding the delegation must flush modified state (such as modified data) to the server and return the delegation. The conflicting request will not be acted on until the recall is complete. The recall is considered complete when the client returns the delegation or the server times out its wait for the delegation to be returned and revokes the delegation as a result of the timeout. In the interim, the server will either delay responding to conflicting requests or respond to them with NFS4ERR\_DELAY. Following the resolution of the recall, the server has the information necessary to grant or deny the second client's request.

At the time the client receives a delegation recall, it may have substantial state that needs to be flushed to the server. Therefore, the server should allow sufficient time for the delegation to be returned since it may involve numerous RPCs to the server. If the server is able to determine that the client is diligently flushing state to the server as a result of the recall, the server MAY extend the usual time allowed for a recall. However, the time allowed for recall completion SHOULD NOT be unbounded.

An example of this is when responsibility to mediate opens on a given file is delegated to a client (see Section 10.4). The server will not know what opens are in effect on the client. Without this knowledge the server will be unable to determine if the access and deny state for the file allows any particular open until the delegation for the file has been returned.

A client failure or a network partition can result in failure to respond to a recall callback. In this case, the server will revoke the delegation which in turn will render useless any modified state still on the client.

Clients need to be aware that server implementors may enforce practical limitations on the number of delegations issued. Further, as there is no way to determine which delegations to revoke, the server is allowed to revoke any. If the server is implemented to revoke another delegation held by that client, then the client may be able to determine that a limit has been reached because each new delegation request results in a revoke. The client could then determine which delegations it may not need and preemptively release them.

#### 10.2.1. Delegation Recovery

There are three situations that delegation recovery must deal with:

- o Client reboot or restart
- o Server reboot or restart (see Section 9.6.3.1)
- o Network partition (full or callback-only)

In the event the client reboots or restarts, the confirmation of a SETCLIENTID done with an `nfs_client_id4` with a new verifier4 value will result in the release of byte-range locks and share reservations. Delegations, however, may be treated a bit differently.

There will be situations in which delegations will need to be reestablished after a client reboots or restarts. The reason for this is the client may have file data stored locally and this data was associated with the previously held delegations. The client will need to reestablish the appropriate file state on the server.

To allow for this type of client recovery, the server MAY allow delegations to be retained after other sort of locks are released. This implies that requests from other clients that conflict with these delegations will need to wait. Because the normal recall process may require significant time for the client to flush changed state to the server, other clients need to be prepared for delays that occur because of a conflicting delegation. In order to give clients a chance to get through the reboot process during which leases will not be renewed, the server MAY extend the period for delegation recovery beyond the typical lease expiration period. For open delegations, such delegations that are not released are

reclaimed using OPEN with a claim type of CLAIM\_DELEGATE\_PREV. (See Section 10.5 and Section 15.18 for discussion of open delegation and the details of OPEN respectively).

A server MAY support a claim type of CLAIM\_DELEGATE\_PREV, but if it does, it MUST NOT remove delegations upon SETCLIENTID\_CONFIRM and instead MUST make them available for client reclaim using CLAIM\_DELEGATE\_PREV. The server MUST NOT remove the delegations until either the client does a DELEGPURGE, or one lease period has elapsed from the time the later of the SETCLIENTID\_CONFIRM or the last successful CLAIM\_DELEGATE\_PREV reclaim.

Note that the requirement stated above is not meant to imply that when the client is no longer obliged, as required above, to retain delegation information, that it should necessarily dispose of it. Some specific cases are:

- o When the period is terminated by the occurrence of DELEGPURGE, deletion of unreclaimed delegations is appropriate and desirable.
- o When the period is terminated by a lease period elapsing without a successful CLAIM\_DELEGATE\_PREV reclaim, and that situation appears to be the result of a network partition (i.e., lease expiration has occurred), a server's lease expiration approach, possibly including the use of courtesy locks would normally provide for the retention of unreclaimed delegations. Even in the event that lease cancellation occurs, such delegation should be reclaimed using CLAIM\_DELEGATE\_PREV as part of network partition recovery.
- o When the period of non-communicating is followed by a client reboot, unreclaimed delegations, should also be reclaimable by use of CLAIM\_DELEGATE\_PREV as part of client reboot recovery.
- o When the period is terminated by a lease period elapsing without a successful CLAIM\_DELEGATE\_PREV reclaim, and lease renewal is occurring, the server may well conclude that unreclaimed delegations have been abandoned, and consider the situation as one in which an implied DELEGPURGE should be assumed.

A server that supports a claim type of CLAIM\_DELEGATE\_PREV MUST support the DELEGPURGE operation, and similarly a server that supports DELEGPURGE MUST support CLAIM\_DELEGATE\_PREV. A server which does not support CLAIM\_DELEGATE\_PREV MUST return NFS4ERR\_NOTSUPP if the client attempts to use that feature or performs a DELEGPURGE operation.

Support for a claim type of CLAIM\_DELEGATE\_PREV, is often referred to as providing for "client-persistent delegations" in that they allow

use of client persistent storage on the client to store data written by the client, even across a client restart. It should be noted that, with the optional exception noted below, this feature requires persistent storage to be used on the client and does not add to persistent storage requirements on the server.

One good way to think about client-persistent delegations is that for the most part, they function like "courtesy locks", with special semantic adjustments to allow them to be retained across a client restart, which cause all other sorts of locks to be freed. Such locks are generally not retained across a server restart. The one exception is the case of simultaneous failure of the client and server and is discussed below.

When the server indicates support of CLAIM\_DELEGATE\_PREV (implicitly) by returning NFS\_OK to DELEGPURGE, a client with a write delegation, can use write-back caching for data to be written to the server, deferring the write-back, until such time as the delegation is recalled, possibly after intervening client restarts. Similarly, when the server indicates support of CLAIM\_DELEGATE\_PREV, a client with a read delegation and an open-for-write subordinate to that delegation, may be sure of the integrity of its persistently cached copy of the file after a client restart without specific verification of the change attribute.

When the server reboots or restarts, delegations are reclaimed (using the OPEN operation with CLAIM\_PREVIOUS) in a similar fashion to byte-range locks and share reservations. However, there is a slight semantic difference. In the normal case, if the server decides that a delegation should not be granted, it performs the requested action (e.g., OPEN) without granting any delegation. For reclaim, the server grants the delegation but a special designation is applied so that the client treats the delegation as having been granted but recalled by the server. Because of this, the client has the duty to write all modified state to the server and then return the delegation. This process of handling delegation reclaim reconciles three principles of the NFSv4 protocol:

- o Upon reclaim, a client claiming resources assigned to it by an earlier server instance must be granted those resources.
- o The server has unquestionable authority to determine whether delegations are to be granted and, once granted, whether they are to be continued.
- o The use of callbacks is not to be depended upon until the client has proven its ability to receive them.

When a client has more than a single open associated with a delegation, state for those additional opens can be established using OPEN operations of type CLAIM\_DELEGATE\_CUR. When these are used to establish opens associated with reclaimed delegations, the server MUST allow them when made within the grace period.

Situations in which there is a series of client and server restarts where there is no restart of both at the same time, are dealt with via a combination of CLAIM\_DELEGATE\_PREV and CLAIM\_PREVIOUS reclaim cycles. Persistent storage is needed only on the client. For each server failure, a CLAIM\_PREVIOUS reclaim cycle is done, while for each client restart, a CLAIM\_DELEGATE\_PREV reclaim cycle is done.

To deal with the possibility of simultaneous failure of client and server (e.g., a data center power outage), the server MAY persistently store delegation information so that it can respond to a CLAIM\_DELEGATE\_PREV reclaim request which it receives from a restarting client. This is the one case in which persistent delegation state can be retained across a server restart. A server is not required to store this information, but if it does do so, it should do so for write delegations and for read delegations, during the pendency of which (across multiple client and/or server instances), some open-for-write was done as part of delegation. When the space to persistently record such information is limited, the server should recall delegations in this class in preference to keeping them active without persistent storage recording.

When a network partition occurs, delegations are subject to freeing by the server when the lease renewal period expires. This is similar to the behavior for locks and share reservations, and, as for locks and share reservations it may be modified by support for "courtesy locks" in which locks are not freed in the absence of a conflicting lock request. Whereas, for locks and share reservations, freeing of locks will occur immediately upon the appearance of a conflicting request, for delegations, the server may institute period during which conflicting requests are held off. Eventually the occurrence of a conflicting request from another client will cause revocation of the delegation.

A loss of the callback path (e.g., by later network configuration change) will have a similar effect in that it can also result in revocation of a delegation. A recall request will fail and revocation of the delegation will result.

A client normally finds out about revocation of a delegation when it uses a stateid associated with a delegation and receives one of the errors NFS4ERR\_EXPIRED, NFS4ERR\_BAD\_STATEID, or NFS4ERR\_ADMIN\_REVOKED (NFS4ERR\_EXPIRED indicates that all lock state associated with the

client has been lost). It also may find out about delegation revocation after a client reboot when it attempts to reclaim a delegation and receives NFS4ERR\_EXPIRED. Note that in the case of a revoked OPEN\_DELEGATE\_WRITE delegation, there are issues because data may have been modified by the client whose delegation is revoked and separately by other clients. See Section 10.5.1 for a discussion of such issues. Note also that when delegations are revoked, information about the revoked delegation will be written by the server to stable storage (as described in Section 9.6). This is done to deal with the case in which a server reboots after revoking a delegation but before the client holding the revoked delegation is notified about the revocation.

Note that when there is a loss of a delegation, due to a network partition in which all locks associated with the lease are lost, the client will also receive the error NFS4ERR\_EXPIRED. This case can be distinguished from other situations in which delegations are revoked by seeing that the associated clientid becomes invalid so that NFS4ERR\_STALE\_CLIENTID is returned when it is used.

When NFS4ERR\_EXPIRED is returned, the server MAY retain information about the delegations held by the client, deleting those that are invalidated by a conflicting request. Retaining such information will allow the client to recover all non-invalidated delegations using the claim type CLAIM\_DELEGATE\_PREV, once the SETCLIENTID\_CONFIRM is done to recover. Attempted recovery of a delegation that the client has no record of, typically because they were invalidated by conflicting requests, will get the error NFS4ERR\_BAD\_RECLAIM. Once a reclaim is attempted for all delegations that the client held, it SHOULD do a DELEGPURGE to allow any remaining server delegation information to be freed.

### 10.3. Data Caching

When applications share access to a set of files, they need to be implemented so as to take account of the possibility of conflicting access by another application. This is true whether the applications in question execute on different clients or reside on the same client.

Share reservations and byte-range locks are the facilities the NFS version 4 protocol provides to allow applications to coordinate access by providing mutual exclusion facilities. The NFSv4 protocol's data caching must be implemented such that it does not invalidate the assumptions that those using these facilities depend upon.

### 10.3.1. Data Caching and OPENS

In order to avoid invalidating the sharing assumptions that applications rely on, NFSv4 clients should not provide cached data to applications or modify it on behalf of an application when it would not be valid to obtain or modify that same data via a READ or WRITE operation.

Furthermore, in the absence of open delegation (see Section 10.4) two additional rules apply. Note that these rules are obeyed in practice by many NFSv2 and NFSv3 clients.

- o First, cached data present on a client must be revalidated after doing an OPEN. Revalidating means that the client fetches the change attribute from the server, compares it with the cached change attribute, and if different, declares the cached data (as well as the cached attributes) as invalid. This is to ensure that the data for the OPENed file is still correctly reflected in the client's cache. This validation must be done at least when the client's OPEN operation includes DENY=WRITE or BOTH thus terminating a period in which other clients may have had the opportunity to open the file with WRITE access. Clients may choose to do the revalidation more often (i.e., at OPENS specifying DENY=NONE) to parallel the NFSv3 protocol's practice for the benefit of users assuming this degree of cache revalidation. Since the change attribute is updated for data and metadata modifications, some client implementors may be tempted to use the time\_modify attribute and not the change attribute to validate cached data, so that metadata changes do not spuriously invalidate clean data. The implementor is cautioned in this approach. The change attribute is guaranteed to change for each update to the file, whereas time\_modify is guaranteed to change only at the granularity of the time\_delta attribute. Use by the client's data cache validation logic of time\_modify and not the change attribute runs the risk of the client incorrectly marking stale data as valid.
- o Second, modified data must be flushed to the server before closing a file OPENed for write. This is complementary to the first rule. If the data is not flushed at CLOSE, the revalidation done after the client OPENS a file is unable to achieve its purpose. The other aspect to flushing the data before close is that the data must be committed to stable storage, at the server, before the CLOSE operation is requested by the client. In the case of a server reboot or restart and a CLOSEd file, it may not be possible to retransmit the data to be written to the file. Hence, this requirement.

### 10.3.2. Data Caching and File Locking

For those applications that choose to use file locking instead of share reservations to exclude inconsistent file access, there is an analogous set of constraints that apply to client side data caching. These rules are effective only if the file locking is used in a way that matches in an equivalent way the actual READ and WRITE operations executed. This is as opposed to file locking that is based on pure convention. For example, it is possible to manipulate a two-megabyte file by dividing the file into two one-megabyte regions and protecting access to the two regions by file locks on bytes zero and one. A lock for write on byte zero of the file would represent the right to do READ and WRITE operations on the first region. A lock for write on byte one of the file would represent the right to do READ and WRITE operations on the second region. As long as all applications manipulating the file obey this convention, they will work on a local file system. However, they may not work with the NFSv4 protocol unless clients refrain from data caching.

The rules for data caching in the file locking environment are:

- o First, when a client obtains a file lock for a particular region, the data cache corresponding to that region (if any cached data exists) must be revalidated. If the change attribute indicates that the file may have been updated since the cached data was obtained, the client must flush or invalidate the cached data for the newly locked region. A client might choose to invalidate all of non-modified cached data that it has for the file but the only requirement for correct operation is to invalidate all of the data in the newly locked region.
- o Second, before releasing a write lock for a region, all modified data for that region must be flushed to the server. The modified data must also be written to stable storage.

Note that flushing data to the server and the invalidation of cached data must reflect the actual byte ranges locked or unlocked. Rounding these up or down to reflect client cache block boundaries will cause problems if not carefully done. For example, writing a modified block when only half of that block is within an area being unlocked may cause invalid modification to the region outside the unlocked area. This, in turn, may be part of a region locked by another client. Clients can avoid this situation by synchronously performing portions of write operations that overlap that portion (initial or final) that is not a full block. Similarly, invalidating a locked area which is not an integral number of full buffer blocks would require the client to read one or two partial blocks from the server if the revalidation procedure shows that the data which the



client possesses may not be valid.

The data that is written to the server as a prerequisite to the unlocking of a region must be written, at the server, to stable storage. The client may accomplish this either with synchronous writes or by following asynchronous writes with a COMMIT operation. This is required because retransmission of the modified data after a server reboot might conflict with a lock held by another client.

A client implementation may choose to accommodate applications which use byte-range locking in non-standard ways (e.g., using a byte-range lock as a global semaphore) by flushing to the server more data upon a LOCKU than is covered by the locked range. This may include modified data within files other than the one for which the unlocks are being done. In such cases, the client must not interfere with applications whose READs and WRITEs are being done only within the bounds of record locks which the application holds. For example, an application locks a single byte of a file and proceeds to write that single byte. A client that chose to handle a LOCKU by flushing all modified data to the server could validly write that single byte in response to an unrelated unlock. However, it would not be valid to write the entire block in which that single written byte was located since it includes an area that is not locked and might be locked by another client. Client implementations can avoid this problem by dividing files with modified data into those for which all modifications are done to areas covered by an appropriate byte-range lock and those for which there are modifications not covered by a byte-range lock. Any writes done for the former class of files must not include areas not locked and thus not modified on the client.

#### 10.3.3. Data Caching and Mandatory File Locking

Client side data caching needs to respect mandatory file locking when it is in effect. The presence of mandatory file locking for a given file is indicated when the client gets back NFS4ERR\_LOCKED from a READ or WRITE on a file it has an appropriate share reservation for. When mandatory locking is in effect for a file, the client must check for an appropriate file lock for data being read or written. If a lock exists for the range being read or written, the client may satisfy the request using the client's validated cache. If an appropriate file lock is not held for the range of the read or write, the read or write request must not be satisfied by the client's cache and the request must be sent to the server for processing. When a read or write request partially overlaps a locked region, the request should be subdivided into multiple pieces with each region (locked or not) treated appropriately.

#### 10.3.4. Data Caching and File Identity

When clients cache data, the file data needs to be organized according to the file system object to which the data belongs. For NFSv3 clients, the typical practice has been to assume for the purpose of caching that distinct filehandles represent distinct file system objects. The client then has the choice to organize and maintain the data cache on this basis.

In the NFSv4 protocol, there is now the possibility to have significant deviations from a "one filehandle per object" model because a filehandle may be constructed on the basis of the object's pathname. Therefore, clients need a reliable method to determine if two filehandles designate the same file system object. If clients were simply to assume that all distinct filehandles denote distinct objects and proceed to do data caching on this basis, caching inconsistencies would arise between the distinct client side objects which mapped to the same server side object.

By providing a method to differentiate filehandles, the NFSv4 protocol alleviates a potential functional regression in comparison with the NFSv3 protocol. Without this method, caching inconsistencies within the same client could occur and this has not been present in previous versions of the NFS protocol. Note that it is possible to have such inconsistencies with applications executing on multiple clients but that is not the issue being addressed here.

For the purposes of data caching, the following steps allow an NFSv4 client to determine whether two distinct filehandles denote the same server side object:

- o If GETATTR directed to two filehandles returns different values of the fsid attribute, then the filehandles represent distinct objects.
- o If GETATTR for any file with an fsid that matches the fsid of the two filehandles in question returns a unique\_handles attribute with a value of TRUE, then the two objects are distinct.
- o If GETATTR directed to the two filehandles does not return the fileid attribute for both of the handles, then it cannot be determined whether the two objects are the same. Therefore, operations which depend on that knowledge (e.g., client side data caching) cannot be done reliably. Note that if GETATTR does not return the fileid attribute for both filehandles, it will return it for neither of the filehandles, since the fsid for both filehandles is the same.

- o If GETATTR directed to the two filehandles returns different values for the fileid attribute, then they are distinct objects.
- o Otherwise they are the same object.

#### 10.4. Open Delegation

When a file is being OPENed, the server may delegate further handling of opens and closes for that file to the opening client. Any such delegation is recallable, since the circumstances that allowed for the delegation are subject to change. In particular, the server may receive a conflicting OPEN from another client, the server must recall the delegation before deciding whether the OPEN from the other client may be granted. Making a delegation is up to the server and clients should not assume that any particular OPEN either will or will not result in an open delegation. The following is a typical set of conditions that servers might use in deciding whether OPEN should be delegated:

- o The client must be able to respond to the server's callback requests. The server will use the CB\_NULL procedure for a test of callback ability.
- o The client must have responded properly to previous recalls.
- o There must be no current open conflicting with the requested delegation.
- o There should be no current delegation that conflicts with the delegation being requested.
- o The probability of future conflicting open requests should be low based on the recent history of the file.
- o The existence of any server-specific semantics of OPEN/CLOSE that would make the required handling incompatible with the prescribed handling that the delegated client would apply (see below).

There are two types of open delegations, OPEN\_DELEGATE\_READ and OPEN\_DELEGATE\_WRITE. A OPEN\_DELEGATE\_READ delegation allows a client to handle, on its own, requests to open a file for reading that do not deny read access to others. It MUST, however, continue to send all requests to open a file for writing to the server. Multiple OPEN\_DELEGATE\_READ delegations may be outstanding simultaneously and do not conflict. A OPEN\_DELEGATE\_WRITE delegation allows the client to handle, on its own, all opens. Only one OPEN\_DELEGATE\_WRITE delegation may exist for a given file at a given time and it is inconsistent with any OPEN\_DELEGATE\_READ delegations.

When a single client holds a `OPEN_DELEGATE_READ` delegation, it is assured that no other client may modify the contents or attributes of the file. If more than one client holds an `OPEN_DELEGATE_READ` delegation, then the contents and attributes of that file are not allowed to change. When a client has an `OPEN_DELEGATE_WRITE` delegation, it may modify the file data since no other client will be accessing the file's data. The client holding a `OPEN_DELEGATE_WRITE` delegation may only affect file attributes which are intimately connected with the file data: `size`, `time_modify`, `change`.

When a client has an open delegation, it does not send `OPENS` or `CLOSES` to the server but updates the appropriate status internally. For a `OPEN_DELEGATE_READ` delegation, opens that cannot be handled locally (opens for write or that deny read access) must be sent to the server.

When an open delegation is made, the response to the `OPEN` contains an open delegation structure which specifies the following:

- o the type of delegation (read or write)
- o space limitation information to control flushing of data on close (`OPEN_DELEGATE_WRITE` delegation only, see Section 10.4.1)
- o an `nfsace4` specifying read and write permissions
- o a `stateid` to represent the delegation for `READ` and `WRITE`

The delegation `stateid` is separate and distinct from the `stateid` for the `OPEN` proper. The standard `stateid`, unlike the delegation `stateid`, is associated with a particular open-owner and will continue to be valid after the delegation is recalled and the file remains open.

When a request internal to the client is made to open a file and open delegation is in effect, it will be accepted or rejected solely on the basis of the following conditions. Any requirement for other checks to be made by the delegate should result in open delegation being denied so that the checks can be made by the server itself.

- o The access and deny bits for the request and the file as described in Section 9.9.
- o The read and write permissions as determined below.

The `nfsace4` passed with delegation can be used to avoid frequent `ACCESS` calls. The permission check should be as follows:

- o If the nfsace4 indicates that the open may be done, then it should be granted without reference to the server.
- o If the nfsace4 indicates that the open may not be done, then an ACCESS request must be sent to the server to obtain the definitive answer.

The server may return an nfsace4 that is more restrictive than the actual ACL of the file. This includes an nfsace4 that specifies denial of all access. Note that some common practices such as mapping the traditional user "root" to the user "nobody" may make it incorrect to return the actual ACL of the file in the delegation response.

The use of delegation together with various other forms of caching creates the possibility that no server authentication will ever be performed for a given user since all of the user's requests might be satisfied locally. Where the client is depending on the server for authentication, the client should be sure authentication occurs for each user by use of the ACCESS operation. This should be the case even if an ACCESS operation would not be required otherwise. As mentioned before, the server may enforce frequent authentication by returning an nfsace4 denying all access with every open delegation.

#### 10.4.1. Open Delegation and Data Caching

OPEN delegation allows much of the message overhead associated with the opening and closing files to be eliminated. An open when an open delegation is in effect does not require that a validation message be sent to the server unless there exists a potential for conflict with the requested share mode. The continued endurance of the "OPEN\_DELEGATE\_READ delegation" provides a guarantee that no OPEN for write and thus no write has occurred that did not originate from this client. Similarly, when closing a file opened for write and if OPEN\_DELEGATE\_WRITE delegation is in effect, the data written does not have to be flushed to the server until the open delegation is recalled. The continued endurance of the open delegation provides a guarantee that no open and thus no read or write has been done by another client.

For the purposes of open delegation, READs and WRITEs done without an OPEN are treated as the functional equivalents of a corresponding type of OPEN. This refers to the READs and WRITEs that use the special stateids consisting of all zero bits or all one bits. Therefore, READs or WRITEs with a special stateid done by another client will force the server to recall a OPEN\_DELEGATE\_WRITE delegation. A WRITE with a special stateid done by another client will force a recall of OPEN\_DELEGATE\_READ delegations.

With delegations, a client is able to avoid writing data to the server when the CLOSE of a file is serviced. The file close system call is the usual point at which the client is notified of a lack of stable storage for the modified file data generated by the application. At the close, file data is written to the server and through normal accounting the server is able to determine if the available file system space for the data has been exceeded (i.e., server returns NFS4ERR\_NOSPC or NFS4ERR\_DQUOT). This accounting includes quotas. The introduction of delegations requires that a alternative method be in place for the same type of communication to occur between client and server.

In the delegation response, the server provides either the limit of the size of the file or the number of modified blocks and associated block size. The server must ensure that the client will be able to flush data to the server of a size equal to that provided in the original delegation. The server must make this assurance for all outstanding delegations. Therefore, the server must be careful in its management of available space for new or modified data taking into account available file system space and any applicable quotas. The server can recall delegations as a result of managing the available file system space. The client should abide by the server's state space limits for delegations. If the client exceeds the stated limits for the delegation, the server's behavior is undefined.

Based on server conditions, quotas or available file system space, the server may grant OPEN\_DELEGATE\_WRITE delegations with very restrictive space limitations. The limitations may be defined in a way that will always force modified data to be flushed to the server on close.

With respect to authentication, flushing modified data to the server after a CLOSE has occurred may be problematic. For example, the user of the application may have logged off the client and unexpired authentication credentials may not be present. In this case, the client may need to take special care to ensure that local unexpired credentials will in fact be available. This may be accomplished by tracking the expiration time of credentials and flushing data well in advance of their expiration or by making private copies of credentials to assure their availability when needed.

#### 10.4.2. Open Delegation and File Locks

When a client holds a OPEN\_DELEGATE\_WRITE delegation, lock operations may be performed locally. This includes those required for mandatory file locking. This can be done since the delegation implies that there can be no conflicting locks. Similarly, all of the revalidations that would normally be associated with obtaining locks

and the flushing of data associated with the releasing of locks need not be done.

When a client holds a `OPEN_DELEGATE_READ` delegation, lock operations are not performed locally. All lock operations, including those requesting non-exclusive locks, are sent to the server for resolution.

#### 10.4.3. Handling of `CB_GETATTR`

The server needs to employ special handling for a `GETATTR` where the target is a file that has a `OPEN_DELEGATE_WRITE` delegation in effect. The reason for this is that the client holding the `OPEN_DELEGATE_WRITE` delegation may have modified the data and the server needs to reflect this change to the second client that submitted the `GETATTR`. Therefore, the client holding the `OPEN_DELEGATE_WRITE` delegation needs to be interrogated. The server will use the `CB_GETATTR` operation. The only attributes that the server can reliably query via `CB_GETATTR` are size and change.

Since `CB_GETATTR` is being used to satisfy another client's `GETATTR` request, the server only needs to know if the client holding the delegation has a modified version of the file. If the client's copy of the delegated file is not modified (data or size), the server can satisfy the second client's `GETATTR` request from the attributes stored locally at the server. If the file is modified, the server only needs to know about this modified state. If the server determines that the file is currently modified, it will respond to the second client's `GETATTR` as if the file had been modified locally at the server.

Since the form of the change attribute is determined by the server and is opaque to the client, the client and server need to agree on a method of communicating the modified state of the file. For the size attribute, the client will report its current view of the file size. For the change attribute, the handling is more involved.

For the client, the following steps will be taken when receiving a `OPEN_DELEGATE_WRITE` delegation:

- o The value of the change attribute will be obtained from the server and cached. Let this value be represented by `c`.
- o The client will create a value greater than `c` that will be used for communicating modified data is held at the client. Let this value be represented by `d`.

- o When the client is queried via CB\_GETATTR for the change attribute, it checks to see if it holds modified data. If the file is modified, the value d is returned for the change attribute value. If this file is not currently modified, the client returns the value c for the change attribute.

For simplicity of implementation, the client MAY for each CB\_GETATTR return the same value d. This is true even if, between successive CB\_GETATTR operations, the client again modifies in the file's data or metadata in its cache. The client can return the same value because the only requirement is that the client be able to indicate to the server that the client holds modified data. Therefore, the value of d may always be  $c + 1$ .

While the change attribute is opaque to the client in the sense that it has no idea what units of time, if any, the server is counting change with, it is not opaque in that the client has to treat it as an unsigned integer, and the server has to be able to see the results of the client's changes to that integer. Therefore, the server MUST encode the change attribute in network order when sending it to the client. The client MUST decode it from network order to its native order when receiving it and the client MUST encode it network order when sending it to the server. For this reason, the change attribute is defined as an unsigned integer rather than an opaque array of bytes.

For the server, the following steps will be taken when providing a OPEN\_DELEGATE\_WRITE delegation:

- o Upon providing a OPEN\_DELEGATE\_WRITE delegation, the server will cache a copy of the change attribute in the data structure it uses to record the delegation. Let this value be represented by sc.
- o When a second client sends a GETATTR operation on the same file to the server, the server obtains the change attribute from the first client. Let this value be cc.
- o If the value cc is equal to sc, the file is not modified and the server returns the current values for change, time\_metadata, and time\_modify (for example) to the second client.
- o If the value cc is NOT equal to sc, the file is currently modified at the first client and most likely will be modified at the server at a future time. The server then uses its current time to construct attribute values for time\_metadata and time\_modify. A new value of sc, which we will call nsc, is computed by the server, such that  $nsc \geq sc + 1$ . The server then returns the constructed time\_metadata, time\_modify, and nsc values to the



requester. The server replaces `sc` in the delegation record with `nsc`. To prevent the possibility of `time_modify`, `time_metadata`, and `change` from appearing to go backward (which would happen if the client holding the delegation fails to write its modified data to the server before the delegation is revoked or returned), the server SHOULD update the file's metadata record with the constructed attribute values. For reasons of reasonable performance, committing the constructed attribute values to stable storage is OPTIONAL.

As discussed earlier in this section, the client MAY return the same `cc` value on subsequent `CB_GETATTR` calls, even if the file was modified in the client's cache yet again between successive `CB_GETATTR` calls. Therefore, the server must assume that the file has been modified yet again, and MUST take care to ensure that the new `nsc` it constructs and returns is greater than the previous `nsc` it returned. An example implementation's delegation record would satisfy this mandate by including a boolean field (let us call it "modified") that is set to `FALSE` when the delegation is granted, and an `sc` value set at the time of grant to the `change` attribute value. The modified field would be set to `TRUE` the first time `cc != sc`, and would stay `TRUE` until the delegation is returned or revoked. The processing for constructing `nsc`, `time_modify`, and `time_metadata` would use this pseudo code:

```
if (!modified) {
    do CB_GETATTR for change and size;

    if (cc != sc)
        modified = TRUE;
} else {
    do CB_GETATTR for size;
}

if (modified) {
    sc = sc + 1;
    time_modify = time_metadata = current_time;
    update sc, time_modify, time_metadata into file's metadata;
}
```

This would return to the client (that sent `GETATTR`) the attributes it requested, but make sure `size` comes from what `CB_GETATTR` returned. The server would not update the file's metadata with the client's modified size.

In the case that the file attribute `size` is different than the server's current value, the server treats this as a modification regardless of the value of the `change` attribute retrieved via

CB\_GETATTR and responds to the second client as in the last step.

This methodology resolves issues of clock differences between client and server and other scenarios where the use of CB\_GETATTR break down.

It should be noted that the server is under no obligation to use CB\_GETATTR and therefore the server MAY simply recall the delegation to avoid its use.

#### 10.4.4. Recall of Open Delegation

The following events necessitate recall of an open delegation:

- o Potentially conflicting OPEN request (or READ/WRITE done with "special" stateid)
- o SETATTR issued by another client
- o REMOVE request for the file
- o RENAME request for the file as either source or target of the RENAME

Whether a RENAME of a directory in the path leading to the file results in recall of an open delegation depends on the semantics of the server file system. If that file system denies such RENAMES when a file is open, the recall must be performed to determine whether the file in question is, in fact, open.

In addition to the situations above, the server may choose to recall open delegations at any time if resource constraints make it advisable to do so. Clients should always be prepared for the possibility of recall.

When a client receives a recall for an open delegation, it needs to update state on the server before returning the delegation. These same updates must be done whenever a client chooses to return a delegation voluntarily. The following items of state need to be dealt with:

- o If the file associated with the delegation is no longer open and no previous CLOSE operation has been sent to the server, a CLOSE operation must be sent to the server.
- o If a file has other open references at the client, then OPEN operations must be sent to the server. The appropriate stateids will be provided by the server for subsequent use by the client

since the delegation stateid will not longer be valid. These OPEN requests are done with the claim type of CLAIM\_DELEGATE\_CUR. This will allow the presentation of the delegation stateid so that the client can establish the appropriate rights to perform the OPEN. (see Section 15.18 for details.)

- o If there are granted file locks, the corresponding LOCK operations need to be performed. This applies to the OPEN\_DELEGATE\_WRITE delegation case only.
- o For a OPEN\_DELEGATE\_WRITE delegation, if at the time of recall the file is not open for write, all modified data for the file must be flushed to the server. If the delegation had not existed, the client would have done this data flush before the CLOSE operation.
- o For a OPEN\_DELEGATE\_WRITE delegation when a file is still open at the time of recall, any modified data for the file needs to be flushed to the server.
- o With the OPEN\_DELEGATE\_WRITE delegation in place, it is possible that the file was truncated during the duration of the delegation. For example, the truncation could have occurred as a result of an OPEN UNCHECKED4 with a size attribute value of zero. Therefore, if a truncation of the file has occurred and this operation has not been propagated to the server, the truncation must occur before any modified data is written to the server.

In the case of OPEN\_DELEGATE\_WRITE delegation, file locking imposes some additional requirements. To precisely maintain the associated invariant, it is required to flush any modified data in any region for which a write lock was released while the OPEN\_DELEGATE\_WRITE delegation was in effect. However, because the OPEN\_DELEGATE\_WRITE delegation implies no other locking by other clients, a simpler implementation is to flush all modified data for the file (as described just above) if any write lock has been released while the OPEN\_DELEGATE\_WRITE delegation was in effect.

An implementation need not wait until delegation recall (or deciding to voluntarily return a delegation) to perform any of the above actions, if implementation considerations (e.g., resource availability constraints) make that desirable. Generally, however, the fact that the actual open state of the file may continue to change makes it not worthwhile to send information about opens and closes to the server, except as part of delegation return. Only in the case of closing the open that resulted in obtaining the delegation would clients be likely to do this early, since, in that case, the close once done will not be undone. Regardless of the client's choices on scheduling these actions, all must be performed

before the delegation is returned, including (when applicable) the close that corresponds to the open that resulted in the delegation. These actions can be performed either in previous requests or in previous operations in the same COMPOUND request.

#### 10.4.5. OPEN Delegation Race with CB\_RECALL

The server informs the client of recall via a CB\_RECALL. A race case which may develop is when the delegation is immediately recalled before the COMPOUND which established the delegation is returned to the client. As the CB\_RECALL provides both a stateid and a filehandle for which the client has no mapping, it cannot honor the recall attempt. At this point, the client has two choices, either do not respond or respond with NFS4ERR\_BADHANDLE. If it does not respond, then it runs the risk of the server deciding to not grant it further delegations.

If instead it does reply with NFS4ERR\_BADHANDLE, then both the client and the server might be able to detect that a race condition is occurring. The client can keep a list of pending delegations. When it receives a CB\_RECALL for an unknown delegation, it can cache the stateid and filehandle on a list of pending recalls. When it is provided with a delegation, it would only use it if it was not on the pending recall list. Upon the next CB\_RECALL, it could immediately return the delegation.

In turn, the server can keep track of when it issues a delegation and assume that if a client responds to the CB\_RECALL with a NFS4ERR\_BADHANDLE, then the client has yet to receive the delegation. The server SHOULD give the client a reasonable time both to get this delegation and to return it before revoking the delegation. Unlike a failed callback path, the server should periodically probe the client with CB\_RECALL to see if it has received the delegation and is ready to return it.

When the server finally determines that enough time has lapsed, it SHOULD revoke the delegation and it SHOULD NOT revoke the lease. During this extended recall process, the server SHOULD be renewing the client lease. The intent here is that the client not pay too onerous a burden for a condition caused by the server.

#### 10.4.6. Clients that Fail to Honor Delegation Recalls

A client may fail to respond to a recall for various reasons, such as a failure of the callback path from server to the client. The client may be unaware of a failure in the callback path. This lack of awareness could result in the client finding out long after the failure that its delegation has been revoked, and another client has

modified the data for which the client had a delegation. This is especially a problem for the client that held a `OPEN_DELEGATE_WRITE` delegation.

The server also has a dilemma in that the client that fails to respond to the recall might also be sending other NFS requests, including those that renew the lease before the lease expires. Without returning an error for those lease renewing operations, the server leads the client to believe that the delegation it has is in force.

This difficulty is solved by the following rules:

- o When the callback path is down, the server MUST NOT revoke the delegation if one of the following occurs:
  - \* The client has issued a `RENEW` operation and the server has returned an `NFS4ERR_CB_PATH_DOWN` error. The server MUST renew the lease for any byte-range locks and share reservations the client has that the server has known about (as opposed to those locks and share reservations the client has established but not yet sent to the server, due to the delegation). The server SHOULD give the client a reasonable time to return its delegations to the server before revoking the client's delegations.
  - \* The client has not issued a `RENEW` operation for some period of time after the server attempted to recall the delegation. This period of time MUST NOT be less than the value of the `lease_time` attribute.
- o When the client holds a delegation, it cannot rely on operations, except for `RENEW`, that take a `stateid`, to renew delegation leases across callback path failures. The client that wants to keep delegations in force across callback path failures must use `RENEW` to do so.

#### 10.4.7. Delegation Revocation

At the point a delegation is revoked, if there are associated opens on the client, the applications holding these opens need to be notified. This notification usually occurs by returning errors for `READ/WRITE` operations or when a `close` is attempted for the open file.

If no opens exist for the file at the point the delegation is revoked, then notification of the revocation is unnecessary. However, if there is modified data present at the client for the file, the user of the application should be notified. Unfortunately,

it may not be possible to notify the user since active applications may not be present at the client. See Section 10.5.1 for additional details.

#### 10.5. Data Caching and Revocation

When locks and delegations are revoked, the assumptions upon which successful caching depend are no longer guaranteed. For any locks or share reservations that have been revoked, the corresponding owner needs to be notified. This notification includes applications with a file open that has a corresponding delegation which has been revoked. Cached data associated with the revocation must be removed from the client. In the case of modified data existing in the client's cache, that data must be removed from the client without it being written to the server. As mentioned, the assumptions made by the client are no longer valid at the point when a lock or delegation has been revoked. For example, another client may have been granted a conflicting lock after the revocation of the lock at the first client. Therefore, the data within the lock range may have been modified by the other client. Obviously, the first client is unable to guarantee to the application what has occurred to the file in the case of revocation.

Notification to a lock owner will in many cases consist of simply returning an error on the next and all subsequent READS/Writes to the open file or on the close. Where the methods available to a client make such notification impossible because errors for certain operations may not be returned, more drastic action such as signals or process termination may be appropriate. The justification for this is that an invariant for which an application depends on may be violated. Depending on how errors are typically treated for the client operating environment, further levels of notification including logging, console messages, and GUI pop-ups may be appropriate.

##### 10.5.1. Revocation Recovery for Write Open Delegation

Revocation recovery for a `OPEN_DELEGATE_WRITE` delegation poses the special issue of modified data in the client cache while the file is not open. In this situation, any client which does not flush modified data to the server on each close must ensure that the user receives appropriate notification of the failure as a result of the revocation. Since such situations may require human action to correct problems, notification schemes in which the appropriate user or administrator is notified may be necessary. Logging and console messages are typical examples.

If there is modified data on the client, it must not be flushed normally to the server. A client may attempt to provide a copy of

the file data as modified during the delegation under a different name in the file system name space to ease recovery. Note that when the client can determine that the file has not been modified by any other client, or when the client has a complete cached copy of the file in question, such a saved copy of the client's view of the file may be of particular value for recovery. In other cases, recovery using a copy of the file based partially on the client's cached data and partially on the server copy as modified by other clients, will be anything but straightforward, so clients may avoid saving file contents in these situations or mark the results specially to warn users of possible problems.

Saving of such modified data in delegation revocation situations may be limited to files of a certain size or might be used only when sufficient disk space is available within the target file system. Such saving may also be restricted to situations when the client has sufficient buffering resources to keep the cached copy available until it is properly stored to the target file system.

#### 10.6. Attribute Caching

The attributes discussed in this section do not include named attributes. Individual named attributes are analogous to files and caching of the data for these needs to be handled just as data caching is for regular files. Similarly, LOOKUP results from an OPENATTR directory are to be cached on the same basis as any other pathnames and similarly for directory contents.

Clients may cache file attributes obtained from the server and use them to avoid subsequent GETATTR requests. Such caching is write through in that modification to file attributes is always done by means of requests to the server and should not be done locally and cached. The exception to this are modifications to attributes that are intimately connected with data caching. Therefore, extending a file by writing data to the local data cache is reflected immediately in the size as seen on the client without this change being immediately reflected on the server. Normally such changes are not propagated directly to the server but when the modified data is flushed to the server, analogous attribute changes are made on the server. When open delegation is in effect, the modified attributes may be returned to the server in the response to a CB\_GETATTR call.

The result of local caching of attributes is that the attribute caches maintained on individual clients will not be coherent. Changes made in one order on the server may be seen in a different order on one client and in a third order on a different client.

The typical file system application programming interfaces do not

provide means to atomically modify or interrogate attributes for multiple files at the same time. The following rules provide an environment where the potential incoherency mentioned above can be reasonably managed. These rules are derived from the practice of previous NFS protocols.

- o All attributes for a given file (per-fsid attributes excepted) are cached as a unit at the client so that no non-serializability can arise within the context of a single file.
- o An upper time boundary is maintained on how long a client cache entry can be kept without being refreshed from the server.
- o When operations are performed that modify attributes at the server, the updated attribute set is requested as part of the containing RPC. This includes directory operations that update attributes indirectly. This is accomplished by following the modifying operation with a GETATTR operation and then using the results of the GETATTR to update the client's cached attributes.

Note that if the full set of attributes to be cached is requested by REaddir, the results can be cached by the client on the same basis as attributes obtained via GETATTR.

A client may validate its cached version of attributes for a file by fetching just both the change and time\_access attributes and assuming that if the change attribute has the same value as it did when the attributes were cached, then no attributes other than time\_access have changed. The reason why time\_access is also fetched is because many servers operate in environments where the operation that updates change does not update time\_access. For example, POSIX file semantics do not update access time when a file is modified by the write system call. Therefore, the client that wants a current time\_access value should fetch it with change during the attribute cache validation processing and update its cached time\_access.

The client may maintain a cache of modified attributes for those attributes intimately connected with data of modified regular files (size, time\_modify, and change). Other than those three attributes, the client MUST NOT maintain a cache of modified attributes. Instead, attribute changes are immediately sent to the server.

In some operating environments, the equivalent to time\_access is expected to be implicitly updated by each read of the content of the file object. If an NFS client is caching the content of a file object, whether it is a regular file, directory, or symbolic link, the client SHOULD NOT update the time\_access attribute (via SETATTR or a small READ or REaddir request) on the server with each read that



is satisfied from cache. The reason is that this can defeat the performance benefits of caching content, especially since an explicit SETATTR of `time_access` may alter the change attribute on the server. If the change attribute changes, clients that are caching the content will think the content has changed, and will re-read unmodified data from the server. Nor is the client encouraged to maintain a modified version of `time_access` in its cache, since this would mean that the client will either eventually have to write the access time to the server with bad performance effects, or it would never update the server's `time_access`, thereby resulting in a situation where an application that caches access time between a close and open of the same file observes the access time oscillating between the past and present. The `time_access` attribute always means the time of last access to a file by a read that was satisfied by the server. This way clients will tend to see only `time_access` changes that go forward in time.

#### 10.7. Data and Metadata Caching and Memory Mapped Files

Some operating environments include the capability for an application to map a file's content into the application's address space. Each time the application accesses a memory location that corresponds to a block that has not been loaded into the address space, a page fault occurs and the file is read (or if the block does not exist in the file, the block is allocated and then instantiated in the application's address space).

As long as each memory mapped access to the file requires a page fault, the relevant attributes of the file that are used to detect access and modification (`time_access`, `time_metadata`, `time_modify`, and `change`) will be updated. However, in many operating environments, when page faults are not required these attributes will not be updated on reads or updates to the file via memory access (regardless of whether the file is a local file or is being accessed remotely). A client or server MAY fail to update attributes of a file that is being accessed via memory mapped I/O. This has several implications:

- o If there is an application on the server that has memory mapped a file that a client is also accessing, the client may not be able to get a consistent value of the change attribute to determine whether its cache is stale or not. A server that knows that the file is memory mapped could always pessimistically return updated values for change so as to force the application to always get the most up to date data and metadata for the file. However, due to the negative performance implications of this, such behavior is OPTIONAL.

- o If the memory mapped file is not being modified on the server, and instead is just being read by an application via the memory mapped interface, the client will not see an updated `time_access` attribute. However, in many operating environments, neither will any process running on the server. Thus NFS clients are at no disadvantage with respect to local processes.
- o If there is another client that is memory mapping the file, and if that client is holding a `OPEN_DELEGATE_WRITE` delegation, the same set of issues as discussed in the previous two bullet items apply. So, when a server does a `CB_GETATTR` to a file that the client has modified in its cache, the response from `CB_GETATTR` will not necessarily be accurate. As discussed earlier, the client's obligation is to report that the file has been modified since the delegation was granted, not whether it has been modified again between successive `CB_GETATTR` calls, and the server **MUST** assume that any file the client has modified in cache has been modified again between successive `CB_GETATTR` calls. Depending on the nature of the client's memory management system, this weak obligation may not be possible. A client **MAY** return stale information in `CB_GETATTR` whenever the file is memory mapped.
- o The mixture of memory mapping and file locking on the same file is problematic. Consider the following scenario, where the page size on each client is 8192 bytes.
  - \* Client A memory maps first page (8192 bytes) of file X
  - \* Client B memory maps first page (8192 bytes) of file X
  - \* Client A write locks first 4096 bytes
  - \* Client B write locks second 4096 bytes
  - \* Client A, via a `STORE` instruction modifies part of its locked region.
  - \* Simultaneous to client A, client B issues a `STORE` on part of its locked region.

Here the challenge is for each client to resynchronize to get a correct view of the first page. In many operating environments, the virtual memory management systems on each client only know a page is modified, not that a subset of the page corresponding to the respective lock regions has been modified. So it is not possible for each client to do the right thing, which is to only write to the server that portion of the page that is locked. For example, if client A simply writes out the page, and then client B writes out the

page, client A's data is lost.

Moreover, if mandatory locking is enabled on the file, then we have a different problem. When clients A and B issue the STORE instructions, the resulting page faults require a byte-range lock on the entire page. Each client then tries to extend their locked range to the entire page, which results in a deadlock.

Communicating the NFS4ERR\_DEADLOCK error to a STORE instruction is difficult at best.

If a client is locking the entire memory mapped file, there is no problem with advisory or mandatory byte-range locking, at least until the client unlocks a region in the middle of the file.

Given the above issues the following are permitted:

- o Clients and servers MAY deny memory mapping a file they know there are byte-range locks for.
- o Clients and servers MAY deny a byte-range lock on a file they know is memory mapped.
- o A client MAY deny memory mapping a file that it knows requires mandatory locking for I/O. If mandatory locking is enabled after the file is opened and mapped, the client MAY deny the application further access to its mapped file.

#### 10.8. Name Caching

The results of LOOKUP and REaddir operations may be cached to avoid the cost of subsequent LOOKUP operations. Just as in the case of attribute caching, inconsistencies may arise among the various client caches. To mitigate the effects of these inconsistencies and given the context of typical file system APIs, an upper time boundary is maintained on how long a client name cache entry can be kept without verifying that the entry has not been made invalid by a directory change operation performed by another client.

When a client is not making changes to a directory for which there exist name cache entries, the client needs to periodically fetch attributes for that directory to ensure that it is not being modified. After determining that no modification has occurred, the expiration time for the associated name cache entries may be updated to be the current time plus the name cache staleness bound.

When a client is making changes to a given directory, it needs to determine whether there have been changes made to the directory by

other clients. It does this by using the change attribute as reported before and after the directory operation in the associated change\_info4 value returned for the operation. The server is able to communicate to the client whether the change\_info4 data is provided atomically with respect to the directory operation. If the change values are provided atomically, the client is then able to compare the pre-operation change value with the change value in the client's name cache. If the comparison indicates that the directory was updated by another client, the name cache associated with the modified directory is purged from the client. If the comparison indicates no modification, the name cache can be updated on the client to reflect the directory operation and the associated timeout extended. The post-operation change value needs to be saved as the basis for future change\_info4 comparisons.

As demonstrated by the scenario above, name caching requires that the client revalidate name cache data by inspecting the change attribute of a directory at the point when the name cache item was cached. This requires that the server update the change attribute for directories when the contents of the corresponding directory is modified. For a client to use the change\_info4 information appropriately and correctly, the server must report the pre and post operation change attribute values atomically. When the server is unable to report the before and after values atomically with respect to the directory operation, the server must indicate that fact in the change\_info4 return value. When the information is not atomically reported, the client should not assume that other clients have not changed the directory.

#### 10.9. Directory Caching

The results of READDIR operations may be used to avoid subsequent READDIR operations. Just as in the cases of attribute and name caching, inconsistencies may arise among the various client caches. To mitigate the effects of these inconsistencies, and given the context of typical file system APIs, the following rules should be followed:

- o Cached READDIR information for a directory which is not obtained in a single READDIR operation must always be a consistent snapshot of directory contents. This is determined by using a GETATTR before the first READDIR and after the last of READDIR that contributes to the cache.
- o An upper time boundary is maintained to indicate the length of time a directory cache entry is considered valid before the client must revalidate the cached information.

The revalidation technique parallels that discussed in the case of name caching. When the client is not changing the directory in question, checking the change attribute of the directory with GETATTR is adequate. The lifetime of the cache entry can be extended at these checkpoints. When a client is modifying the directory, the client needs to use the change\_info4 data to determine whether there are other clients modifying the directory. If it is determined that no other client modifications are occurring, the client may update its directory cache to reflect its own changes.

As demonstrated previously, directory caching requires that the client revalidate directory cache data by inspecting the change attribute of a directory at the point when the directory was cached. This requires that the server update the change attribute for directories when the contents of the corresponding directory is modified. For a client to use the change\_info4 information appropriately and correctly, the server must report the pre and post operation change attribute values atomically. When the server is unable to report the before and after values atomically with respect to the directory operation, the server must indicate that fact in the change\_info4 return value. When the information is not atomically reported, the client should not assume that other clients have not changed the directory.

## 11. Minor Versioning

To address the requirement of an NFS protocol that can evolve as the need arises, the NFSv4 protocol contains the rules and framework to allow for future minor changes or versioning.

The base assumption with respect to minor versioning is that any future accepted minor version must follow the IETF process and be documented in a standards track RFC. Therefore, each minor version number will correspond to an RFC. Minor version 0 of the NFS version 4 protocol is represented by this RFC. The COMPOUND and CB\_COMPOUND procedures support the encoding of the minor version being requested by the client.

The following items represent the basic rules for the development of minor versions. Note that a future minor version may decide to modify or add to the following rules as part of the minor version definition.

### 1. Procedures are not added or deleted

To maintain the general RPC model, NFSv4 minor versions will not add to or delete procedures from the NFS program.

2. Minor versions may add operations to the COMPOUND and CB\_COMPOUND procedures.

The addition of operations to the COMPOUND and CB\_COMPOUND procedures does not affect the RPC model.

1. Minor versions may append attributes to the bitmap4 that represents sets of attributes and to the fatattr4 that represents sets of attribute values.

This allows for the expansion of the attribute model to allow for future growth or adaptation.

2. Minor version X must append any new attributes after the last documented attribute.

Since attribute results are specified as an opaque array of per-attribute XDR encoded results, the complexity of adding new attributes in the midst of the current definitions would be too burdensome.

3. Minor versions must not modify the structure of an existing operation's arguments or results.

Again, the complexity of handling multiple structure definitions for a single operation is too burdensome. New operations should be added instead of modifying existing structures for a minor version.

This rule does not preclude the following adaptations in a minor version.

- \* adding bits to flag fields, such as new attributes to GETATTR's bitmap4 data type, and providing corresponding variants of opaque arrays, such as a notify4 used together with such bitmaps
- \* adding bits to existing attributes like ACLs that have flag words
- \* extending enumerated types (including NFS4ERR\_\*) with new values

4. Minor versions must not modify the structure of existing attributes.
5. Minor versions must not delete operations.

This prevents the potential reuse of a particular operation "slot" in a future minor version.

6. Minor versions must not delete attributes.
7. Minor versions must not delete flag bits or enumeration values.
8. Minor versions may declare an operation MUST NOT be implemented.

Specifying that an operation MUST NOT be implemented is equivalent to obsoleting an operation. For the client, it means that the operation MUST NOT be sent to the server. For the server, an NFS error can be returned as opposed to "dropping" the request as an XDR decode error. This approach allows for the obsolescence of an operation while maintaining its structure so that a future minor version can reintroduce the operation.

1. Minor versions may declare that an attribute MUST NOT be implemented.
  2. Minor versions may declare that a flag bit or enumeration value MUST NOT be implemented.
9. Minor versions may downgrade features from REQUIRED to RECOMMENDED, or RECOMMENDED to OPTIONAL.
  10. Minor versions may upgrade features from OPTIONAL to RECOMMENDED or RECOMMENDED to REQUIRED.
  11. A client and server that support minor version X SHOULD support minor versions 0 through X-1 as well.
  12. Except for infrastructural changes, no new features may be introduced as REQUIRED in a minor version.

This rule allows for the introduction of new functionality and forces the use of implementation experience before designating a feature as REQUIRED. On the other hand, some classes of features are infrastructural and have broad effects. Allowing infrastructural features to be RECOMMENDED or OPTIONAL complicates implementation of the minor version.

13. A client MUST NOT attempt to use a stateid, filehandle, or similar returned object from the COMPOUND procedure with minor version X for another COMPOUND procedure with minor version Y, where  $X \neq Y$ .

## 12. Internationalization

### 12.1. Introduction

This section uses NFSv4.0 internationalization, as implemented by existing clients and servers, as the basis upon which NFSv4.0 clients may implement internationalization support. This procedure, while necessary, may result in confusion if we do not clearly understand that we are mixing prescription and description, and why, in this particular case, this is a valid thing to do.

Note that in this chapter, the keywords "MUST", "SHOULD", and "MAY", retain their normal meanings. However, in deriving this specification from implementation patterns, we document below how the normative terms used derive from the behavior of existing implementations.

- o Behavior implemented by all existing clients or servers is described using "MUST", since new implementations need to follow existing ones to be assured of interoperability. While it is possible that different behavior might be workable, we have found no case where this seems reasonable.
- o Behavior implemented by no existing clients or servers is described using "MUST NOT", if such behavior poses interoperability problems.
- o Behavior implemented by most existing clients or servers, where that behavior is more desirable than any alternative is described using "SHOULD", since new implementations need to follow that existing practice unless there are strong reasons to do otherwise.

This also holds for "SHOULD NOT".

- o Behavior implemented by some not all existing clients or servers, is described using "MAY", indicating that new implementations have a choice as to whether they will behave in that way. Thus, new implementations will have the same flexibility that existing ones do.
- o Behavior implemented by all existing clients or servers, so far as is known, but where there remains some uncertainty as to details is described using "should". Such cases primarily concern details of error returns. New implementations should follow existing practice even though such situations generally do not affect interoperability.

In the case of a MAY, SHOULD, or SHOULD NOT that applies to servers,



clients need to be aware that there are servers which may or may not take the specified action, and they need to be prepared for either eventuality.

## 12.2. String Encoding

Strings that potentially contain non-ASCII Characters are represented in NFSv4 using the UTF-8 encoding of Unicode. See [RFC2279] for precise encoding and decoding rules.

Some details of the protocol treatment depend on the type of string:

- o For strings that are component names, any non-ASCII characters SHOULD be represented using the UTF-8 encoding of Unicode.

In many cases, clients have no knowledge of the encoding being used, with the encoding done at user-level under control of a per-process locale specification. As a result, it may be impossible for the NFSv4 client to enforce use of UTF-8. Use of non-UTF-8 encodings can be problematic since it may interfere with access to files stored using encoding and because normalization-related processing (see Section 12.3) would result in name aliasing in the case of a non-UTF-8 encoding resulted characters strings that have multiple equivalent Unicode encodings.

- o For strings whose form is defined by other internet standards, non-ASCII characters MUST be represented using the UTF-8 encoding of Unicode. In addition other sorts of restrictions defined by those standards need to be addressed. See Section 12.4 for details.
- o The contents of symbolic links (of type linktext4 in the XDR) MUST be treated as opaque data by NFSv4 servers. Although UTF-8 encoding is often used, it need not be. In this respect, the contents of symbolic links are like the contents of regular files in that their encoding is not within the scope of this specification.
- o For other sorts of strings, any non-ASCII characters SHOULD be represented using the UTF-8 encoding of Unicode.

## 12.3. Normalization

The client and server operating environments may differ in their policies and operational methods with respect to character normalization (See [Unicode1] for a discussion of normalization forms). This difference may also exist between applications on the same client. This adds to the difficulty of providing a single

normalization policy for the protocol that allows for maximal interoperability. This issue is similar to the character case issues where the server may or may not support case insensitive file name matching and may or may not preserve the character case when storing file names. The protocol does not mandate a particular behavior but allows for a range of useful behaviors.

The NFS version 4 protocol does not mandate the use of a particular normalization form at this time. A subsequent minor version of the NFSv4 protocol might specify a particular normalization form. Therefore, the server and client can expect that they may receive unnormalized characters within protocol requests and responses. If the operating environment requires normalization, then the implementation must normalize the various UTF-8 encoded strings within the protocol before presenting the information to an application (at the client) or local file system (at the server).

Server implementations MAY normalize file names to conform to a particular normalization form before using the resulting string when looking up or creating a file. Servers MAY also perform normalization-insensitive string comparisons without modifying name to match a particular normalization form. Except in cases in which component names are excluded from normalization-related handling because they are not valid UTF-8 strings, a server MUST make the same choice (as to whether to normalize or not, the target form of normalization and whether to do normalization-insensitive string comparisons) in the same way for all accesses to a particular filesystem. Servers MUST NOT reject a file name because it doesn't conform to a particular normalization form.

#### 12.4. Types with Processing Defined by Other Internet Areas

There are two types of strings which NFSv4 deals with whose processing is defined by other Internet standards, and where issues related to different handling choices by server operating systems or server file systems do not apply.

These are as follows:

- o Server names as they appear in the `fs_locations` attribute. Note that for most purposes, such server names will only be sent by the server to the client. The exception is use of the `fs_locations` attribute in a `VERIFY` or `NVERIFY` operation.
- o Principal suffixes which are used to denote sets of users and groups, and are in the form of domain names.

The general rules for handling all of these domain-related strings

are similar and independent of the role of the sender or receiver as client or server although the consequences of failure to obey these rules may be different for client or server. The server can report errors when it is sent invalid strings, whereas the client will simply ignore invalid string or use a default value in their place.

The string sent SHOULD be in the form of a U-label although it MAY be in the form of an A-label or a UTF-8 string that would not map to itself when canonicalized by applying `ToUnicode(ToASCII(...))`. The receiver needs to be able to accept domain and server names in any of the formats allowed. The server MUST reject, using the error `NFS4ERR_INVALID`, a string which is not valid UTF-8 or which begins with "xn--" and violates the rules for a valid A-label.

When a domain string is part of `id@domain` or `group@domain`, the server SHOULD map domain strings which are A-labels (see [RFC5890]) or are UTF-8 domain names which are not U-labels, to the corresponding U-label, using `ToUnicode(domain)` or `ToUnicode(ToASCII(domain))`. As a result, the domain name returned within a `userid` on a `GETATTR` may not match that sent when the `userid` is set using `SETATTR`, although when this happens, the domain will be in the form of a U-label. When the server does not map domain strings which are not U-labels into a U-label, which it MAY do, it MUST NOT modify the domain, and the domain returned on a `GETATTR` of the `userid` MUST be the same as that used when setting the `userid` by the `SETATTR`.

The server MAY implement `VERIFY` and `NVERIFY` without translating internal state to a string form, so that, for example, a user principal which represents a specific numeric user id, will match a different principal string which represents the same numeric user id.

#### 12.5. UTF-8 Related Errors

Where the client sends an invalid UTF-8 string, the server MAY return an `NFS4ERR_INVALID` error. This includes cases in which inappropriate prefixes are detected and where the count includes trailing bytes that do not constitute a full UCS character.

Requirements for server handling of component names which are not valid UTF-8, when a server does not return `NFS4ERR_INVALID` in response to receiving these are described in Section 12.6.

Where the client supplied string is not rejected with `NFS4ERR_INVALID` but contains characters that are not supported by the server as a value for that string (e.g., names containing slashes or characters that have more than two octets on a filesystem that supports Unicode characters only), the server should return an `NFS4ERR_BADCHAR` error.

Where a UTF-8 string is used as a file name, and the file system, while supporting all of the characters within the name, does not allow that particular name to be used, the error should return the error NFS4ERR\_BADNAME. This includes such situations as file system prohibitions of "." and ".." as file names for certain operations, and similar constraints

#### 12.6. Handling of component names that are not valid UTF-8 strings

In cases in which the server receives a component name that is not a valid UTF-8 string, the required handling depends on whether a file object is being created or looked up. Object creation happens for the component name in LINK and CREATE, and the second component name in RENAME. Object lookup happens for the component name in LOOKUP and REMOVE, and the first component name in RENAME. The component name in OPEN will result in object lookup and also object creation if the lookup fails and the other OPEN parameters allow file creation.

With regard to normalization-related processing, it is generally inhibited, both in the case of object creation and object lookup:

- o Characters which are not valid UTF-8, have no canonically equivalent Unicode string so normalization-related processing cannot happen.
- o In cases in which a string has valid UTF-8 character strings that do have canonically equivalent Unicode strings, but if the component name is not valid UTF-8, the server MAY perform normalization-related processing on valid UTF-8 substrings within it that do have canonical equivalents.

When creation of a component name which is not valid UTF-8 occurs, and is allowed by the server:

- o A subsequent lookup of the same component name in the same directory MUST result in finding the created file object.
- o A READDIR of the directory in which the name is created MUST result in an entry containing the component name used for the creation.
- o A subsequent lookup using any other string as the component name SHOULD NOT find the originally created object.
- o A subsequent lookup using any valid UTF-8 string MUST NOT find the originally created object.

When a lookup of a component name which is not valid UTF-8 occurs,

and is allowed by the server:

- o The result of the lookup MUST NOT be any directory entry created with a valid UTF-8 component name.
- o The result of the lookup MUST be a directory entry created with the identical invalid UTF-8 string, if one exists in the directory.

### 13. Error Values

NFS error numbers are assigned to failed operations within a Compound (COMPOUND or CB\_COMPOUND) request. A Compound request contains a number of NFS operations that have their results encoded in sequence in a Compound reply. The results of successful operations will consist of an NFS4\_OK status followed by the encoded results of the operation. If an NFS operation fails, an error status will be entered in the reply and the Compound request will be terminated.

#### 13.1. Error Definitions

Protocol Error Definitions

Error	Number	Description
NFS4_OK	0	Section 13.1.3.1
NFS4ERR_ACCESS	13	Section 13.1.6.1
NFS4ERR_ATTRNOTSUPP	10032	Section 13.1.11.1
NFS4ERR_ADMIN_REVOKED	10047	Section 13.1.5.1
NFS4ERR_BADCHAR	10040	Section 13.1.7.1
NFS4ERR_BADHANDLE	10001	Section 13.1.2.1
NFS4ERR_BADNAME	10041	Section 13.1.7.2
NFS4ERR_BADOWNER	10039	Section 13.1.11.2
NFS4ERR_BADTYPE	10007	Section 13.1.4.1
NFS4ERR_BADXDR	10036	Section 13.1.1.1
NFS4ERR_BAD_COOKIE	10003	Section 13.1.1.2
NFS4ERR_BAD_RANGE	10042	Section 13.1.8.1
NFS4ERR_BAD_SEQID	10026	Section 13.1.8.2
NFS4ERR_BAD_STATEID	10025	Section 13.1.5.2
NFS4ERR_CLID_INUSE	10017	Section 13.1.10.1
NFS4ERR_DEADLOCK	10045	Section 13.1.8.3
NFS4ERR_DELAY	10008	Section 13.1.1.3
NFS4ERR_DENIED	10010	Section 13.1.8.4
NFS4ERR_DQUOT	69	Section 13.1.4.2
NFS4ERR_EXIST	17	Section 13.1.4.3
NFS4ERR_EXPIRED	10011	Section 13.1.5.3

NFS4ERR_FBIG	27	Section 13.1.4.4
NFS4ERR_FHEXPIRED	10014	Section 13.1.2.2
NFS4ERR_FILE_OPEN	10046	Section 13.1.4.5
NFS4ERR_GRACE	10013	Section 13.1.9.1
NFS4ERR_INVAL	22	Section 13.1.1.4
NFS4ERR_IO	5	Section 13.1.4.6
NFS4ERR_ISDIR	21	Section 13.1.2.3
NFS4ERR_LEASE_MOVED	10031	Section 13.1.5.4
NFS4ERR_LOCKED	10012	Section 13.1.8.5
NFS4ERR_LOCKS_HELD	10037	Section 13.1.8.6
NFS4ERR_LOCK_NOTSUPP	10043	Section 13.1.8.7
NFS4ERR_LOCK_RANGE	10028	Section 13.1.8.8
NFS4ERR_MINOR_VERS_MISMATCH	10021	Section 13.1.3.2
NFS4ERR_MLINK	31	Section 13.1.4.7
NFS4ERR_MOVED	10019	Section 13.1.2.4
NFS4ERR_NAMETOOLONG	63	Section 13.1.7.3
NFS4ERR_NOENT	2	Section 13.1.4.8
NFS4ERR_NOFILEHANDLE	10020	Section 13.1.2.5
NFS4ERR_NOSPC	28	Section 13.1.4.9
NFS4ERR_NOTDIR	20	Section 13.1.2.6
NFS4ERR_NOTEMPTY	66	Section 13.1.4.10
NFS4ERR_NOTSUPP	10004	Section 13.1.1.5
NFS4ERR_NOT_SAME	10027	Section 13.1.11.3
NFS4ERR_NO_GRACE	10033	Section 13.1.9.2
NFS4ERR_NXIO	6	Section 13.1.4.11
NFS4ERR_OLD_STATEID	10024	Section 13.1.5.5
NFS4ERR_OPENMODE	10038	Section 13.1.8.9
NFS4ERR_OP_ILLEGAL	10044	Section 13.1.3.3
NFS4ERR_PERM	1	Section 13.1.6.2
NFS4ERR_RECLAIM_BAD	10034	Section 13.1.9.3
NFS4ERR_RECLAIM_CONFLICT	10035	Section 13.1.9.4
NFS4ERR_RESOURCE	10018	Section 13.1.3.4
NFS4ERR_RESTOREFH	10030	Section 13.1.4.12
NFS4ERR_ROFS	30	Section 13.1.4.13
NFS4ERR_SAME	10009	Section 13.1.11.4
NFS4ERR_SERVERFAULT	10006	Section 13.1.1.6
NFS4ERR_STALE	70	Section 13.1.2.7
NFS4ERR_STALE_CLIENTID	10022	Section 13.1.10.2
NFS4ERR_STALE_STATEID	10023	Section 13.1.5.6
NFS4ERR_SYMLINK	10029	Section 13.1.2.8
NFS4ERR_TOOSMALL	10005	Section 13.1.1.7
NFS4ERR_WRONGSEC	10016	Section 13.1.6.3
NFS4ERR_XDEV	18	Section 13.1.4.14

Table 5

### 13.1.1. General Errors

This section deals with errors that are applicable to a broad set of different purposes.

#### 13.1.1.1. NFS4ERR\_BADXDR (Error Code 10036)

The arguments for this operation do not match those specified in the XDR definition. This includes situations in which the request ends before all the arguments have been seen. Note that this error applies when fixed enumerations (these include booleans) have a value within the input stream which is not valid for the enum. A replier may pre-parse all operations for a Compound procedure before doing any operation execution and return RPC-level XDR errors in that case.

#### 13.1.1.2. NFS4ERR\_BAD\_COOKIE (Error Code 10003)

Used for operations that provide a set of information indexed by some quantity provided by the client or cookie sent by the server for an earlier invocation. Where the value cannot be used for its intended purpose, this error results.

#### 13.1.1.3. NFS4ERR\_DELAY (Error Code 10008)

For any of a number of reasons, the replier could not process this operation in what was deemed a reasonable time. The client should wait and then try the request with a new RPC transaction ID.

Some example of situations that might lead to this situation:

- o A server that supports hierarchical storage receives a request to process a file that had been migrated.
- o An operation requires a delegation recall to proceed and waiting for this delegation recall makes processing this request in a timely fashion impossible.

#### 13.1.1.4. NFS4ERR\_INVALID (Error Code 22)

The arguments for this operation are not valid for some reason, even though they do match those specified in the XDR definition for the request.

#### 13.1.1.5. NFS4ERR\_NOTSUPP (Error Code 10004)

Operation not supported, either because the operation is an OPTIONAL one and is not supported by this server or because the operation MUST NOT be implemented in the current minor version.

#### 13.1.1.6. NFS4ERR\_SERVERFAULT (Error Code 10006)

An error occurred on the server which does not map to any of the specific legal NFSv4 protocol error values. The client should translate this into an appropriate error. UNIX clients may choose to translate this to EIO.

#### 13.1.1.7. NFS4ERR\_TOOSMALL (Error Code 10005)

Used where an operation returns a variable amount of data, with a limit specified by the client. Where the data returned cannot be fitted within the limit specified by the client, this error results.

#### 13.1.2. Filehandle Errors

These errors deal with the situation in which the current or saved filehandle, or the filehandle passed to PUTFH intended to become the current filehandle, is invalid in some way. This includes situations in which the filehandle is a valid filehandle in general but is not of the appropriate object type for the current operation.

Where the error description indicates a problem with the current or saved filehandle, it is to be understood that filehandles are only checked for the condition if they are implicit arguments of the operation in question.

##### 13.1.2.1. NFS4ERR\_BADHANDLE (Error Code 10001)

Illegal NFS filehandle for the current server. The current filehandle failed internal consistency checks. Once accepted as valid (by PUTFH), no subsequent status change can cause the filehandle to generate this error.

##### 13.1.2.2. NFS4ERR\_FHEXPIRED (Error Code 10014)

A current or saved filehandle which is an argument to the current operation is volatile and has expired at the server.

##### 13.1.2.3. NFS4ERR\_ISDIR (Error Code 21)

The current or saved filehandle designates a directory when the current operation does not allow a directory to be accepted as the target of this operation.

##### 13.1.2.4. NFS4ERR\_MOVED (Error Code 10019)

The file system which contains the current filehandle object is not present at the server. It may have been relocated, migrated to



another server or may have never been present. The client may obtain the new file system location by obtaining the "fs\_locations" or attribute for the current filehandle. For further discussion, refer to Section 8.

#### 13.1.1.2.5. NFS4ERR\_NOFILEHANDLE (Error Code 10020)

The logical current or saved filehandle value is required by the current operation and is not set. This may be a result of a malformed COMPOUND operation (i.e., no PUTFH or PUTROOTFH before an operation that requires the current filehandle be set).

#### 13.1.1.2.6. NFS4ERR\_NOTDIR (Error Code 20)

The current (or saved) filehandle designates an object which is not a directory for an operation in which a directory is required.

#### 13.1.1.2.7. NFS4ERR\_STALE (Error Code 70)

The current or saved filehandle value designating an argument to the current operation is invalid. The file system object referred to by that filehandle no longer exists or access to it has been revoked.

#### 13.1.1.2.8. NFS4ERR\_SYMLINK (Error Code 10029)

The current filehandle designates a symbolic link when the current operation does not allow a symbolic link as the target.

### 13.1.1.3. Compound Structure Errors

This section deals with errors that relate to overall structure of a Compound request (by which we mean to include both COMPOUND and CB\_COMPOUND), rather than to particular operations.

There are a number of basic constraints on the operations that may appear in a Compound request.

#### 13.1.1.3.1. NFS\_OK (Error code 0)

Indicates the operation completed successfully, in that all of the constituent operations completed without error.

#### 13.1.1.3.2. NFS4ERR\_MINOR\_VERS\_MISMATCH (Error code 10021)

The minor version specified is not one that the current listener supports. This value is returned in the overall status for the Compound but is not associated with a specific operation since the results must specify a result count of zero.

#### 13.1.3.3. NFS4ERR\_OP\_ILLEGAL (Error Code 10044)

The operation code is not a valid one for the current Compound procedure. The opcode in the result stream matched with this error is the ILLEGAL value, although the value that appears in the request stream may be different. Where an illegal value appears and the replier pre-parses all operations for a Compound procedure before doing any operation execution, an RPC-level XDR error may be returned in this case.

#### 13.1.3.4. NFS4ERR\_RESOURCE (Error Code 10018)

For the processing of the Compound procedure, the server may exhaust available resources and cannot continue processing operations within the Compound procedure. This error will be returned from the server in those instances of resource exhaustion related to the processing of the Compound procedure.

#### 13.1.4. File System Errors

These errors describe situations which occurred in the underlying file system implementation rather than in the protocol or any NFSv4.x feature.

##### 13.1.4.1. NFS4ERR\_BADTYPE (Error Code 10007)

An attempt was made to create an object with an inappropriate type specified to CREATE. This may be because the type is undefined, because it is a type not supported by the server, or because it is a type for which create is not intended such as a regular file or named attribute, for which OPEN is used to do the file creation.

##### 13.1.4.2. NFS4ERR\_DQUOT (Error Code 69)

Resource (quota) hard limit exceeded. The user's resource limit on the server has been exceeded.

##### 13.1.4.3. NFS4ERR\_EXIST (Error Code 17)

A file system object of the specified target name (when creating, renaming or linking) already exists.

##### 13.1.4.4. NFS4ERR\_FBIG (Error Code 27)

Filesystem object too large. The operation would have caused a file system object to grow beyond the server's limit.

#### 13.1.4.5. NFS4ERR\_FILE\_OPEN (Error Code 10046)

The operation is not allowed because a file system object involved in the operation is currently open. Servers may, but are not required to disallow linking-to, removing, or renaming open file system objects.

#### 13.1.4.6. NFS4ERR\_IO (Error Code 5)

Indicates that an I/O error occurred for which the file system was unable to provide recovery.

#### 13.1.4.7. NFS4ERR\_MLINK (Error Code 31)

The request would have caused the server's limit for the number of hard links a file system object may have to be exceeded.

#### 13.1.4.8. NFS4ERR\_NOENT (Error Code 2)

Indicates no such file or directory. The file system object referenced by the name specified does not exist.

#### 13.1.4.9. NFS4ERR\_NOSPC (Error Code 28)

Indicates no space left on device. The operation would have caused the server's file system to exceed its limit.

#### 13.1.4.10. NFS4ERR\_NOTEMPTY (Error Code 66)

An attempt was made to remove a directory that was not empty.

#### 13.1.4.11. NFS4ERR\_NXIO (Error Code 6)

I/O error. No such device or address.

#### 13.1.4.12. NFS4ERR\_RESTOREFH (Error Code 10030)

The RESTOREFH operation does not have a saved filehandle (identified by SAVEFH) to operate upon.

#### 13.1.4.13. NFS4ERR\_ROFS (Error Code 30)

Indicates a read-only file system. A modifying operation was attempted on a read-only file system.

#### 13.1.4.14. NFS4ERR\_XDEV (Error Code 18)

Indicates an attempt to do an operation, such as linking, that inappropriately crosses a boundary. This may be due to such boundaries as:

- o That between file systems (where the fsids are different).
- o That between different named attribute directories or between a named attribute directory and an ordinary directory.
- o That between regions of a file system that the file system implementation treats as separate (for example for space accounting purposes), and where cross-connection between the regions are not allowed.

#### 13.1.5. State Management Errors

These errors indicate problems with the stateid (or one of the stateids) passed to a given operation. This includes situations in which the stateid is invalid as well as situations in which the stateid is valid but designates revoked locking state. Depending on the operation, the stateid when valid may designate opens, byte-range locks, or file delegations.

##### 13.1.5.1. NFS4ERR\_ADMIN\_REVOKED (Error Code 10047)

A stateid designates locking state of any type that has been revoked due to administrative interaction, possibly while the lease is valid, or because a delegation was revoked because of failure to return it, while the lease was valid.

##### 13.1.5.2. NFS4ERR\_BAD\_STATEID (Error Code 10025)

A stateid generated by the current server instance was used which either:

- o Does not designate any locking state (either current or superseded) for a current (state-owner, file) pair.
- o Designates locking state that was freed after lease expiration but without any lease cancelation, as may happen in the handling of "courtesy locks".

#### 13.1.5.3. NFS4ERR\_EXPIRED (Error Code 10011)

A stateid or clientid designates locking state of any type that has been revoked or released due to cancellation of the client's lease, either immediately upon lease expiration, or following a later request for a conflicting lock.

#### 13.1.5.4. NFS4ERR\_LEASE\_MOVED (Error Code 10031)

A lease being renewed is associated with a file system that has been migrated to a new server.

#### 13.1.5.5. NFS4ERR\_OLD\_STATEID (Error Code 10024)

A stateid is provided with a seqid value that is not the most current.

#### 13.1.5.6. NFS4ERR\_STALE\_STATEID (Error Code 10023)

A stateid generated by an earlier server instance was used.

#### 13.1.6. Security Errors

These are the various permission-related errors in NFSv4.

##### 13.1.6.1. NFS4ERR\_ACCESS (Error Code 13)

Indicates permission denied. The caller does not have the correct permission to perform the requested operation. Contrast this with NFS4ERR\_PERM (Section 13.1.6.2), which restricts itself to owner or privileged user permission failures.

##### 13.1.6.2. NFS4ERR\_PERM (Error Code 1)

Indicates requester is not the owner. The operation was not allowed because the caller is neither a privileged user (root) nor the owner of the target of the operation.

##### 13.1.6.3. NFS4ERR\_WRONGSEC (Error Code 10016)

Indicates that the security mechanism being used by the client for the operation does not match the server's security policy. The client should change the security mechanism being used and re-send the operation. SECINFO can be used to determine the appropriate mechanism.

#### 13.1.7. Name Errors

Names in NFSv4 are UTF-8 strings. When the strings are not of length zero, the error NFS4ERR\_INVAL results. When they are not valid UTF-8 the error NFS4ERR\_INVAL also results, but servers may accommodate file systems with different character formats and not return this error. Besides this, there are a number of other errors to indicate specific problems with names.

##### 13.1.7.1. NFS4ERR\_BADCHAR (Error Code 10040)

A UTF-8 string contains a character which is not supported by the server in the context in which it being used.

##### 13.1.7.2. NFS4ERR\_BADNAME (Error Code 10041)

A name string in a request consisted of valid UTF-8 characters supported by the server but the name is not supported by the server as a valid name for current operation. An example might be creating a file or directory named ".." on a server whose file system uses that name for links to parent directories.

This error should not be returned due a normalization issue in a string. When a file system keeps names in a particular normalization form, it is the server's responsibility to do the appropriate normalization, rather than rejecting the name.

##### 13.1.7.3. NFS4ERR\_NAMETOOLONG (Error Code 63)

Returned when the filename in an operation exceeds the server's implementation limit.

#### 13.1.8. Locking Errors

This section deal with errors related to locking, both as to share reservations and byte-range locking. It does not deal with errors specific to the process of reclaiming locks. Those are dealt with in the next section.

##### 13.1.8.1. NFS4ERR\_BAD\_RANGE (Error Code 10042)

The range for a LOCK, LOCKT, or LOCKU operation is not appropriate to the allowable range of offsets for the server. E.g., this error results when a server which only supports 32-bit ranges receives a range that cannot be handled by that server. (See Section 15.12.4).

#### 13.1.8.2. NFS4ERR\_BAD\_SEQID (Error Code 10026)

The sequence number (seqid) in a locking request is neither the next expected number or the last number processed.

#### 13.1.8.3. NFS4ERR\_DEADLOCK (Error Code 10045)

The server has been able to determine a file locking deadlock condition for a blocking lock request.

#### 13.1.8.4. NFS4ERR\_DENIED (Error Code 10010)

An attempt to lock a file is denied. Since this may be a temporary condition, the client is encouraged to re-send the lock request until the lock is accepted. See Section 9.4 for a discussion of the re-send.

#### 13.1.8.5. NFS4ERR\_LOCKED (Error Code 10012)

A read or write operation was attempted on a file where there was a conflict between the I/O and an existing lock:

- o There is a share reservation inconsistent with the I/O being done.
- o The range to be read or written intersects an existing mandatory byte range lock.

#### 13.1.8.6. NFS4ERR\_LOCKS\_HELD (Error Code 10037)

An operation was prevented by the unexpected presence of locks.

#### 13.1.8.7. NFS4ERR\_LOCK\_NOTSUPP (Error Code 10043)

A locking request was attempted which would require the upgrade or downgrade of a lock range already held by the owner when the server does not support atomic upgrade or downgrade of locks.

#### 13.1.8.8. NFS4ERR\_LOCK\_RANGE (Error Code 10028)

A lock request is operating on a range that overlaps in part a currently held lock for the current lock owner and does not precisely match a single such lock where the server does not support this type of request, and thus does not implement POSIX locking semantics [fcntl]. See Section 15.12.5, Section 15.13.5, and Section 15.14.5 for a discussion of how this applies to LOCK, LOCKT, and LOCKU respectively.

#### 13.1.8.9. NFS4ERR\_OPENMODE (Error Code 10038)

The client attempted a READ, WRITE, LOCK or other operation not sanctioned by the stateid passed (e.g., writing to a file opened only for read).

#### 13.1.9. Reclaim Errors

These errors relate to the process of reclaiming locks after a server restart.

##### 13.1.9.1. NFS4ERR\_GRACE (Error Code 10013)

The server is in its recovery or grace period which should at least match the lease period of the server. A locking request other than a reclaim could not be granted during that period.

##### 13.1.9.2. NFS4ERR\_NO\_GRACE (Error Code 10033)

The server cannot guarantee that it has not granted state to another client which may conflict with this client's state. No further reclaims from this client will succeed.

##### 13.1.9.3. NFS4ERR\_RECLAIM\_BAD (Error Code 10034)

The server cannot guarantee that it has not granted state to another client which may conflict with the requested state. However, this applies only to the state requested in this call; further reclaims may succeed.

Unlike NFS4ERR\_RECLAIM\_CONFLICT, this can occur between correctly functioning clients and servers: the "edge condition" scenarios described in Section 9.6.3.1 leave only the server knowing whether the client's locks are still valid, and NFS4ERR\_RECLAIM\_BAD is the server's way of informing the client that they are not.

##### 13.1.9.4. NFS4ERR\_RECLAIM\_CONFLICT (Error Code 10035)

The reclaim attempted by the client conflicts with a lock already held by another client. Unlike NFS4ERR\_RECLAIM\_BAD, this can only occur if one of the clients misbehaved.

#### 13.1.10. Client Management Errors

This sections deals with errors associated with requests used to create and manage client IDs.



#### 13.1.10.1. NFS4ERR\_CLID\_INUSE (Error Code 10017)

The SETCLIENTID operation has found that a client id is already in use by another client.

#### 13.1.10.2. NFS4ERR\_STALE\_CLIENTID (Error Code 10022)

A client ID not recognized by the server was used in a locking or SETCLIENTID\_CONFIRM request.

#### 13.1.11. Attribute Handling Errors

This section deals with errors specific to attribute handling within NFSv4.

##### 13.1.11.1. NFS4ERR\_ATTRNOTSUPP (Error Code 10032)

An attribute specified is not supported by the server. This error MUST NOT be returned by the GETATTR operation.

##### 13.1.11.2. NFS4ERR\_BADOWNER (Error Code 10039)

Returned when an owner or owner\_group attribute value or the who field of an ace within an ACL attribute value cannot be translated to a local representation.

##### 13.1.11.3. NFS4ERR\_NOT\_SAME (Error Code 10027)

This error is returned by the VERIFY operation to signify that the attributes compared were not the same as those provided in the client's request.

##### 13.1.11.4. NFS4ERR\_SAME (Error Code 10009)

This error is returned by the NVERIFY operation to signify that the attributes compared were the same as those provided in the client's request.

#### 13.2. Operations and their valid errors

This section contains a table which gives the valid error returns for each protocol operation. The error code NFS4\_OK (indicating no error) is not listed but should be understood to be returnable by all operations except ILLEGAL.

Valid error returns for each protocol operation

Operation	Errors
ACCESS	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
CLOSE	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADHANDLE, NFS4ERR_BAD_SEQID, NFS4ERR_BAD_STATEID, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCKS_HELD, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_OLD_STATEID, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID
COMMIT	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK
CREATE	NFS4ERR_ACCESS, NFS4ERR_ATTRNOTSUPP, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADOWNER, NFS4ERR_BADTYPE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_EXIST, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NAMETOOLONG, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTDIR, NFS4ERR_PERM, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
DELEGPURGE	NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_NOTSUPP, NFS4ERR_LEASE_MOVED, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE_CLIENTID

DELEGRETURN	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BAD_STATEID, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_INVAL, NFS4ERR_LEASE_MOVED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID
GETATTR	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
GETFH	NFS4ERR_BADHANDLE, NFS4ERR_FHEXPIRED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
ILLEGAL LINK	NFS4ERR_BADXDR, NFS4ERR_OP_ILLEGAL NFS4ERR_ACCESS, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_EXIST, NFS4ERR_FHEXPIRED, NFS4ERR_FILE_OPEN, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MLINK, NFS4ERR_MOVED, NFS4ERR_NAMETOOLONG, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTDIR, NFS4ERR_NOTSUPP, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_WRONGSEC, NFS4ERR_XDEV
LOCK	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADHANDLE, NFS4ERR_BAD_RANGE, NFS4ERR_BAD_SEQID, NFS4ERR_BAD_STATEID, NFS4ERR_BADXDR, NFS4ERR_DEADLOCK, NFS4ERR_DELAY, NFS4ERR_DENIED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCK_NOTSUPP, NFS4ERR_LOCK_RANGE, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NO_GRACE, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_RECLAIM_BAD, NFS4ERR_RECLAIM_CONFLICT, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_CLIENTID, NFS4ERR_STALE_STATEID

LOCKT	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_BAD_RANGE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_DENIED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVALID, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCK_RANGE, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_CLIENTID
LOCKU	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADHANDLE, NFS4ERR_BAD_RANGE, NFS4ERR_BAD_SEQID, NFS4ERR_BAD_STATEID, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVALID, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCK_RANGE, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_OLD_STATEID, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID
LOOKUP	NFS4ERR_ACCESS, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_INVALID, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NAMETOOLONG, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTDIR, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_WRONGSEC
LOOKUPP	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTDIR, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_WRONGSEC
NVERIFY	NFS4ERR_ACCESS, NFS4ERR_ATTRNOTSUPP, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVALID, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_SAME, NFS4ERR_SERVERFAULT, NFS4ERR_STALE

OPEN	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_ATTRNOTSUPP, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADOWNER, NFS4ERR_BADXDR, NFS4ERR_BAD_SEQID, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_EXIST, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVALID, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NAMETOOLONG, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTDIR, NFS4ERR_NOTSUP, NFS4ERR_NO_GRACE, NFS4ERR_OLD_STATEID, NFS4ERR_PERM, NFS4ERR_RECLAIM_BAD, NFS4ERR_RECLAIM_CONFLICT, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_SHARE_DENIED, NFS4ERR_STALE, NFS4ERR_STALE_CLIENTID, NFS4ERR_SYMLINK, NFS4ERR_WRONGSEC
OPENATTR	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_FHEXPIRED, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTSUPP, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
OPEN_CONFIRM	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADHANDLE, NFS4ERR_BAD_SEQID, NFS4ERR_BAD_STATEID, NFS4ERR_BADXDR, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_INVALID, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_OLD_STATEID, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID
OPEN_DOWNGRADE	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_BAD_SEQID, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_INVALID, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCKS_HELD, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_OLD_STATEID, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID

PUTFH	NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_MOVED, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_WRONGSEC
PUTPUBFH	NFS4ERR_DELAY, NFS4ERR_SERVERFAULT, NFS4ERR_WRONGSEC
PUTROOTFH	NFS4ERR_DELAY, NFS4ERR_SERVERFAULT, NFS4ERR_WRONGSEC
READ	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID, NFS4ERR_SYMLINK
READDIR	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_BAD_COOKIE, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTDIR, NFS4ERR_NOT_SAME, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_TOOSMALL
READLINK	NFS4ERR_ACCESS, NFS4ERR_BADHANDLE, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOTSUP, NFS4ERR_RESOURCE, NFS4ERR_NOFILEHANDLE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
RELEASE_LOCKOWNER	NFS4ERR_BADXDR, NFS4ERR_EXPIRED, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCKS_HELD, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE_CLIENTID
REMOVE	NFS4ERR_ACCESS, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_FILE_OPEN, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NAME_TOO_LONG, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTDIR, NFS4ERR_NOTEMPTY, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE

RENAME	NFS4ERR_ACCESS, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_EXIST, NFS4ERR_FHEXPIRED, NFS4ERR_FILE_OPEN, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NAMETOOLONG, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTDIR, NFS4ERR_NOTEMPTY, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_WRONGSEC, NFS4ERR_XDEV
RENEW	NFS4ERR_ACCESS, NFS4ERR_BADXDR, NFS4ERR_CB_PATH_DOWN, NFS4ERR_EXPIRED, NFS4ERR_LEASE_MOVED, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE_CLIENTID
RESTOREFH	NFS4ERR_BADHANDLE, NFS4ERR_FHEXPIRED, NFS4ERR_MOVED, NFS4ERR_RESOURCE, NFS4ERR_RESTOREFH, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_WRONGSEC
SAVEFH	NFS4ERR_BADHANDLE, NFS4ERR_FHEXPIRED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
SECINFO	NFS4ERR_ACCESS, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADNAME, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_INVAL, NFS4ERR_MOVED, NFS4ERR_NAMETOOLONG, NFS4ERR_NOENT, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTDIR, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
SETATTR	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_ATTRNOTSUPP, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADOWNER, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_PERM, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID

SETCLIENTID	NFS4ERR_BADXDR, NFS4ERR_CLID_INUSE, NFS4ERR_DELAY, NFS4ERR_INVALID, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT
SETCLIENTID_CONFIRM	NFS4ERR_BADXDR, NFS4ERR_CLID_INUSE, NFS4ERR_DELAY, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE_CLIENTID
VERIFY	NFS4ERR_ACCESS, NFS4ERR_ATTRNOTSUPP, NFS4ERR_BADCHAR, NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVALID, NFS4ERR_IO, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOT_SAME, NFS4ERR_RESOURCE, NFS4ERR_SERVERFAULT, NFS4ERR_STALE
WRITE	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BADHANDLE, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_DQUOT, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVALID, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LEASE_MOVED, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NXIO, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_RESOURCE, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_STALE_STATEID, NFS4ERR_SYMLINK

Table 6

### 13.3. Callback operations and their valid errors

This section contains a table which gives the valid error returns for each callback operation. The error code NFS4\_OK (indicating no error) is not listed but should be understood to be returnable by all callback operations with the exception of CB\_ILLEGAL.



Valid error returns for each protocol callback operation

Callback Operation	Errors
CB_GETATTR	NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_DELAY, NFS4ERR_INVAL, NFS4ERR_SERVERFAULT
CB_ILLEGAL	NFS4ERR_BADXDR, NFS4ERR_OP_ILLEGAL
CB_RECALL	NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_SERVERFAULT

Table 7

#### 13.4. Errors and the operations that use them

Error	Operations
NFS4ERR_ACCESS	ACCESS, COMMIT, CREATE, GETATTR, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, READ, READDIR, READLINK, REMOVE, RENAME, RENEW, SECINFO, SETATTR, VERIFY, WRITE
NFS4ERR_ADMIN_REVOKED	CLOSE, DELEGRETURN, LOCK, LOCKU, OPEN, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, SETATTR, WRITE
NFS4ERR_ATTRNOTSUPP	CREATE, NVERIFY, OPEN, SETATTR, VERIFY
NFS4ERR_BADCHAR	CREATE, LINK, LOOKUP, NVERIFY, OPEN, REMOVE, RENAME, SECINFO, SETATTR, VERIFY
NFS4ERR_BADHANDLE	ACCESS, CB_GETATTR, CB_RECALL, CLOSE, COMMIT, CREATE, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, PUTFH, READ, READDIR, READLINK, REMOVE, RENAME, RESTOREFH, SAVEFH, SECINFO, SETATTR, VERIFY, WRITE
NFS4ERR_BADNAME	CREATE, LINK, LOOKUP, OPEN, REMOVE, RENAME, SECINFO
NFS4ERR_BADOWNER	CREATE, OPEN, SETATTR
NFS4ERR_BADTYPE	CREATE

NFS4ERR_BADXDR	ACCESS, CB_GETATTR, CB_ILLEGAL, CB_RECALL, CLOSE, COMMIT, CREATE, DELEGPURGE, DELEGRETURN, GETATTR, ILLEGAL, LINK, LOCK, LOCKT, LOCKU, LOOKUP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, PUTFH, READ, READDIR, RELEASE_LOCKOWNER, REMOVE, RENAME, RENEW, SECINFO, SETATTR, SETCLIENTID, SETCLIENTID_CONFIRM, VERIFY, WRITE
NFS4ERR_BAD_COOKIE	READDIR
NFS4ERR_BAD_RANGE	LOCK, LOCKT, LOCKU
NFS4ERR_BAD_SEQID	CLOSE, LOCK, LOCKU, OPEN, OPEN_CONFIRM, OPEN_DOWNGRADE
NFS4ERR_BAD_STATEID	CB_RECALL, CLOSE, DELEGRETURN, LOCK, LOCKU, OPEN, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, SETATTR, WRITE
NFS4ERR_CB_PATH_DOWN	RENEW
NFS4ERR_CLID_INUSE	SETCLIENTID, SETCLIENTID_CONFIRM
NFS4ERR_DEADLOCK	LOCK
NFS4ERR_DELAY	ACCESS, CB_GETATTR, CB_RECALL, CLOSE, COMMIT, CREATE, DELEGPURGE, DELEGRETURN, GETATTR, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_DOWNGRADE, PUTFH, PUTPUBFH, PUTROOTFH, READ, READDIR, READLINK, REMOVE, RENAME, SECINFO, SETATTR, SETCLIENTID, SETCLIENTID_CONFIRM, VERIFY, WRITE
NFS4ERR_DENIED	LOCK, LOCKT
NFS4ERR_DQUOT	CREATE, LINK, OPEN, OPENATTR, RENAME, SETATTR, WRITE
NFS4ERR_EXIST	CREATE, LINK, OPEN, RENAME
NFS4ERR_EXPIRED	CLOSE, DELEGRETURN, LOCK, LOCKT, LOCKU, OPEN, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, RELEASE_LOCKOWNER, RENEW, SETATTR, WRITE
NFS4ERR_FBIG	OPEN, SETATTR, WRITE
NFS4ERR_FHEXPIRED	ACCESS, CLOSE, COMMIT, CREATE, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, PUTFH, READ, READDIR, READLINK, REMOVE, RENAME, RESTOREFH, SAVEFH, SECINFO, SETATTR, VERIFY, WRITE

NFS4ERR_FILE_OPEN	LINK, REMOVE, RENAME
NFS4ERR_GRACE	GETATTR, LOCK, LOCKT, LOCKU, NVERIFY, OPEN, READ, REMOVE, RENAME, SETATTR, VERIFY, WRITE
NFS4ERR_INVAL	ACCESS, CB_GETATTR, CLOSE, COMMIT, CREATE, DELEGRETURN, GETATTR, LINK, LOCK, LOCKT, LOCKU, LOOKUP, NVERIFY, OPEN, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, READDIR, READLINK, REMOVE, RENAME, SECINFO, SETATTR, SETCLIENTID, VERIFY, WRITE
NFS4ERR_IO	ACCESS, COMMIT, CREATE, GETATTR, LINK, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, READ, READDIR, READLINK, REMOVE, RENAME, SETATTR, VERIFY, WRITE
NFS4ERR_ISDIR	CLOSE, COMMIT, LINK, LOCK, LOCKT, LOCKU, OPEN, OPEN_CONFIRM, READ, READLINK, SETATTR, WRITE
NFS4ERR_LEASE_MOVED	CLOSE, DELEGPURGE, DELEGRETURN, LOCK, LOCKT, LOCKU, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, RELEASE_LOCKOWNER, RENEW, SETATTR, WRITE
NFS4ERR_LOCKED	READ, SETATTR, WRITE
NFS4ERR_LOCKS_HELD	CLOSE, OPEN_DOWNGRADE, RELEASE_LOCKOWNER
NFS4ERR_LOCK_NOTSUPP	LOCK
NFS4ERR_LOCK_RANGE	LOCK, LOCKT, LOCKU
NFS4ERR_MLINK	LINK
NFS4ERR_MOVED	ACCESS, CLOSE, COMMIT, CREATE, DELEGRETURN, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, PUTFH, READ, READDIR, READLINK, REMOVE, RENAME, RESTOREFH, SAVEFH, SECINFO, SETATTR, VERIFY, WRITE
NFS4ERR_NAMETOOLONG	CREATE, LINK, LOOKUP, OPEN, REMOVE, RENAME, SECINFO
NFS4ERR_NOENT	LINK, LOOKUP, LOOKUPP, OPEN, OPENATTR, REMOVE, RENAME, SECINFO
NFS4ERR_NOFILEHANDLE	ACCESS, CLOSE, COMMIT, CREATE, DELEGRETURN, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, READDIR, READLINK, REMOVE, RENAME, SAVEFH, SECINFO, SETATTR, VERIFY, WRITE

NFS4ERR_NOSPC	CREATE, LINK, OPEN, OPENATTR, RENAME, SETATTR, WRITE
NFS4ERR_NOTDIR	CREATE, LINK, LOOKUP, LOOKUPP, OPEN, READDIR, REMOVE, RENAME, SECINFO
NFS4ERR_NOTEMPTY	REMOVE, RENAME
NFS4ERR_NOTSUP	OPEN, READLINK
NFS4ERR_NOTSUPP	DELEGPURGE, DELEGRETURN, LINK, OPENATTR
NFS4ERR_NOT_SAME	READDIR, VERIFY
NFS4ERR_NO_GRACE	LOCK, OPEN
NFS4ERR_NXIO	WRITE
NFS4ERR_OLD_STATEID	CLOSE, DELEGRETURN, LOCK, LOCKU, OPEN, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, SETATTR, WRITE
NFS4ERR_OPENMODE	LOCK, READ, SETATTR, WRITE
NFS4ERR_OP_ILLEGAL	CB_ILLEGAL, ILLEGAL
NFS4ERR_PERM	CREATE, OPEN, SETATTR
NFS4ERR_RECLAIM_BAD	LOCK, OPEN
NFS4ERR_RECLAIM_CONFLICT	LOCK, OPEN
NFS4ERR_RESOURCE	ACCESS, CLOSE, COMMIT, CREATE, DELEGPURGE, DELEGRETURN, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, READDIR, READLINK, RELEASE_LOCKOWNER, REMOVE, RENAME, RENEW, RESTOREFH, SAVEFH, SECINFO, SETATTR, SETCLIENTID, SETCLIENTID_CONFIRM, VERIFY, WRITE
NFS4ERR_RESTOREFH	RESTOREFH
NFS4ERR_ROFS	COMMIT, CREATE, LINK, OPEN, OPENATTR, OPEN_DOWNGRADE, REMOVE, RENAME, SETATTR, WRITE
NFS4ERR_SAME	NVERIFY
NFS4ERR_SERVERFAULT	ACCESS, CB_GETATTR, CB_RECALL, CLOSE, COMMIT, CREATE, DELEGPURGE, DELEGRETURN, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, PUTFH, PUTPUBFH, PUTROOTFH, READ, READDIR, READLINK, RELEASE_LOCKOWNER, REMOVE, RENAME, RENEW, RESTOREFH, SAVEFH, SECINFO, SETATTR, SETCLIENTID, SETCLIENTID_CONFIRM, VERIFY, WRITE
NFS4ERR_SHARE_DENIED	OPEN

NFS4ERR_STALE	ACCESS, CLOSE, COMMIT, CREATE, DELEGRETURN, GETATTR, GETFH, LINK, LOCK, LOCKT, LOCKU, LOOKUP, LOOKUPP, NVERIFY, OPEN, OPENATTR, OPEN_CONFIRM, OPEN_DOWNGRADE, PUTFH, READ, READDIR, READLINK, REMOVE, RENAME, RESTOREFH, SAVEFH, SECINFO, SETATTR, VERIFY, WRITE
NFS4ERR_STALE_CLIENTID	DELEGPURGE, LOCK, LOCKT, OPEN, RELEASE_LOCKOWNER, RENEW, SETCLIENTID_CONFIRM
NFS4ERR_STALE_STATEID	CLOSE, DELEGRETURN, LOCK, LOCKU, OPEN_CONFIRM, OPEN_DOWNGRADE, READ, SETATTR, WRITE
NFS4ERR_SYMLINK	COMMIT, LOOKUP, LOOKUPP, OPEN, READ, WRITE
NFS4ERR_TOOSMALL	READDIR
NFS4ERR_WRONGSEC	LINK, LOOKUP, LOOKUPP, OPEN, PUTFH, PUTPUBFH, PUTROOTFH, RENAME, RESTOREFH
NFS4ERR_XDEV	LINK, RENAME

Table 8

#### 14. NFSv4 Requests

For the NFSv4 RPC program, there are two traditional RPC procedures: NULL and COMPOUND. All other functionality is defined as a set of operations and these operations are defined in normal XDR/RPC syntax and semantics. However, these operations are encapsulated within the COMPOUND procedure. This requires that the client combine one or more of the NFSv4 operations into a single request.

The NFS4\_CALLBACK program is used to provide server to client signaling and is constructed in a similar fashion as the NFSv4 program. The procedures CB\_NULL and CB\_COMPOUND are defined in the same way as NULL and COMPOUND are within the NFS program. The CB\_COMPOUND request also encapsulates the remaining operations of the NFS4\_CALLBACK program. There is no predefined RPC program number for the NFS4\_CALLBACK program. It is up to the client to specify a program number in the "transient" program range. The program and port number of the NFS4\_CALLBACK program are provided by the client as part of the SETCLIENTID/SETCLIENTID\_CONFIRM sequence. The program and port can be changed by another SETCLIENTID/SETCLIENTID\_CONFIRM sequence, and it is possible to use the sequence to change them within a client incarnation without removing relevant leased client state.

### 14.1. Compound Procedure

The COMPOUND procedure provides the opportunity for better performance within high latency networks. The client can avoid cumulative latency of multiple RPCs by combining multiple dependent operations into a single COMPOUND procedure. A compound operation may provide for protocol simplification by allowing the client to combine basic procedures into a single request that is customized for the client's environment.

The CB\_COMPOUND procedure precisely parallels the features of COMPOUND as described above.

The basic structure of the COMPOUND procedure is:

```
+-----+-----+-----+-----+-----+-----+
| tag | minorversion | numops | op + args | op + args | op + args |
+-----+-----+-----+-----+-----+-----+
```

and the reply's structure is:

```
+-----+-----+-----+-----+
| last status | tag | numres | status + op + results |
+-----+-----+-----+-----+
```

The numops and numres fields, used in the depiction above, represent the count for the counted array encoding use to signify the number of arguments or results encoded in the request and response. As per the XDR encoding, these counts must match exactly the number of operation arguments or results encoded.

### 14.2. Evaluation of a Compound Request

The server will process the COMPOUND procedure by evaluating each of the operations within the COMPOUND procedure in order. Each component operation consists of a 32 bit operation code, followed by the argument of length determined by the type of operation. The results of each operation are encoded in sequence into a reply buffer. The results of each operation are preceded by the opcode and a status code (normally zero). If an operation results in a non-zero status code, the status will be encoded and evaluation of the compound sequence will halt and the reply will be returned. Note that evaluation stops even in the event of "non error" conditions such as NFS4ERR\_SAME.

There are no atomicity requirements for the operations contained within the COMPOUND procedure. The operations being evaluated as part of a COMPOUND request may be evaluated simultaneously with other

COMPOUND requests that the server receives.

A COMPOUND is not a transaction and it is the client's responsibility for recovering from any partially completed COMPOUND procedure. These may occur at any point due to errors such as NFS4ERR\_RESOURCE and NFS4ERR\_DELAY. Note that these errors can occur in an otherwise valid operation string. Further, a server reboot which occurs in the middle of processing a COMPOUND procedure may leave the client with the difficult task of determining how far COMPOUND processing has proceeded. Therefore, the client should avoid overly complex COMPOUND procedures in the event of the failure of an operation within the procedure.

Each operation assumes a "current" and "saved" filehandle that is available as part of the execution context of the compound request. Operations may set, change, or return the current filehandle. The "saved" filehandle is used for temporary storage of a filehandle value and as operands for the RENAME and LINK operations.

#### 14.3. Synchronous Modifying Operations

NFSv4 operations that modify the file system are synchronous. When an operation is successfully completed at the server, the client can depend that any data associated with the request is now on stable storage (the one exception is in the case of the file data in a WRITE operation with the UNSTABLE4 option specified).

This implies that any previous operations within the same compound request are also reflected in stable storage. This behavior enables the client's ability to recover from a partially executed compound request which may resulted from the failure of the server. For example, if a compound request contains operations A and B and the server is unable to send a response to the client, depending on the progress the server made in servicing the request the result of both operations may be reflected in stable storage or just operation A may be reflected. The server must not have just the results of operation B in stable storage.

#### 14.4. Operation Values

The operations encoded in the COMPOUND procedure are identified by operation values. To avoid overlap with the RPC procedure numbers, operations 0 (zero) and 1 are not defined. Operation 2 is not defined but reserved for future use with minor versioning.

## 15. NFSv4 Procedures

### 15.1. Procedure 0: NULL - No Operation

#### 15.1.1. SYNOPSIS

<null>

#### 15.1.2. ARGUMENT

void;

#### 15.1.3. RESULT

void;

#### 15.1.4. DESCRIPTION

Standard NULL procedure. Void argument, void response. This procedure has no functionality associated with it. Because of this it is sometimes used to measure the overhead of processing a service request. Therefore, the server should ensure that no unnecessary work is done in servicing this procedure.

### 15.2. Procedure 1: COMPOUND - Compound Operations

#### 15.2.1. SYNOPSIS

compoundargs -> compoundres

#### 15.2.2. ARGUMENT

```
union nfs_argop4 switch (nfs_opnum4 argop) {
    case <OPCODE>: <argument>;
    ...
};

struct COMPOUND4args {
    utf8str_cs    tag;
    uint32_t      minorversion;
    nfs_argop4    argarray<>;
};
```



## 15.2.3. RESULT

```
union nfs_resop4 switch (nfs_opnum4 resop) {
    case <OPCODE>: <argument>;
    ...
};

struct COMPOUND4res {
    nfsstat4      status;
    utf8str_cs    tag;
    nfs_resop4    resarray<>;
};
```

## 15.2.4. DESCRIPTION

The COMPOUND procedure is used to combine one or more of the NFS operations into a single RPC request. The main NFS RPC program has two main procedures: NULL and COMPOUND. All other operations use the COMPOUND procedure as a wrapper.

The COMPOUND procedure is used to combine individual operations into a single RPC request. The server interprets each of the operations in turn. If an operation is executed by the server and the status of that operation is NFS4\_OK, then the next operation in the COMPOUND procedure is executed. The server continues this process until there are no more operations to be executed or one of the operations has a status value other than NFS4\_OK.

In the processing of the COMPOUND procedure, the server may find that it does not have the available resources to execute any or all of the operations within the COMPOUND sequence. In this case, the error NFS4ERR\_RESOURCE will be returned for the particular operation within the COMPOUND procedure where the resource exhaustion occurred. This assumes that all previous operations within the COMPOUND sequence have been evaluated successfully. The results for all of the evaluated operations must be returned to the client.

The server will generally choose between two methods of decoding the client's request. The first would be the traditional one-pass XDR decode, in which decoding of the entire COMPOUND precedes execution of any operation within it. If there is an XDR decoding error in this case, an RPC XDR decode error would be returned. The second method would be to make an initial pass to decode the basic COMPOUND request and then to XDR decode each of the individual operations, as the server is ready to execute it. In this case, the server may encounter an XDR decode error during such an operation decode, after previous operations within the COMPOUND have been executed. In this

case, the server would return the error NFS4ERR\_BADXDR to signify the decode error.

The COMPOUND arguments contain a "minorversion" field. The initial and default value for this field is 0 (zero). This field will be used by future minor versions such that the client can communicate to the server what minor version is being requested. If the server receives a COMPOUND procedure with a minorversion field value that it does not support, the server MUST return an error of NFS4ERR\_MINOR\_VERS\_MISMATCH and a zero length resultdata array.

Contained within the COMPOUND results is a "status" field. If the results array length is non-zero, this status must be equivalent to the status of the last operation that was executed within the COMPOUND procedure. Therefore, if an operation incurred an error then the "status" value will be the same error value as is being returned for the operation that failed.

Note that operations, 0 (zero) and 1 (one) are not defined for the COMPOUND procedure. Operation 2 is not defined but reserved for future definition and use with minor versioning. If the server receives a operation array that contains operation 2 and the minorversion field has a value of 0 (zero), an error of NFS4ERR\_OP\_ILLEGAL, as described in the next paragraph, is returned to the client. If an operation array contains an operation 2 and the minorversion field is non-zero and the server does not support the minor version, the server returns an error of NFS4ERR\_MINOR\_VERS\_MISMATCH. Therefore, the NFS4ERR\_MINOR\_VERS\_MISMATCH error takes precedence over all other errors.

It is possible that the server receives a request that contains an operation that is less than the first legal operation (OP\_ACCESS) or greater than the last legal operation (OP\_RELEASE\_LOCKOWNER). In this case, the server's response will encode the opcode OP\_ILLEGAL rather than the illegal opcode of the request. The status field in the ILLEGAL return results will set to NFS4ERR\_OP\_ILLEGAL. The COMPOUND procedure's return results will also be NFS4ERR\_OP\_ILLEGAL.

The definition of the "tag" in the request is left to the implementor. It may be used to summarize the content of the compound request for the benefit of packet sniffers and engineers debugging implementations. However, the value of "tag" in the response SHOULD be the same value as provided in the request. This applies to the tag field of the CB\_COMPOUND procedure as well.

## 15.2.4.1. Current Filehandle

The current and saved filehandle are used throughout the protocol. Most operations implicitly use the current filehandle as a argument and many set the current filehandle as part of the results. The combination of client specified sequences of operations and current and saved filehandle arguments and results allows for greater protocol flexibility. The best or easiest example of current filehandle usage is a sequence like the following:

PUTFH fh1	{fh1}
LOOKUP "compA"	{fh2}
GETATTR	{fh2}
LOOKUP "compB"	{fh3}
GETATTR	{fh3}
LOOKUP "compC"	{fh4}
GETATTR	{fh4}
GETFH	

Figure 1

In this example, the PUTFH (Section 15.22) operation explicitly sets the current filehandle value while the result of each LOOKUP operation sets the current filehandle value to the resultant file system object. Also, the client is able to insert GETATTR operations using the current filehandle as an argument.

The PUTROOTFH (Section 15.24) and PUTPUBFH (Section 15.24) operations also set the current filehandle. The above example would replace "PUTFH fh1" with PUTROOTFH or PUTPUBFH with no filehandle argument in order to achieve the same effect (on the assumption that "compA" is directly below the root of the namespace).

Along with the current filehandle, there is a saved filehandle. While the current filehandle is set as the result of operations like LOOKUP, the saved filehandle must be set directly with the use of the SAVEFH operation. The SAVEFH operation copies the current filehandle value to the saved value. The saved filehandle value is used in combination with the current filehandle value for the LINK and RENAME operations. The RESTOREFH operation will copy the saved filehandle value to the current filehandle value; as a result, the saved filehandle value may be used a sort of "scratch" area for the client's series of operations.

#### 15.2.5. IMPLEMENTATION

Since an error of any type may occur after only a portion of the operations have been evaluated, the client must be prepared to recover from any failure. If the source of an NFS4ERR\_RESOURCE error was a complex or lengthy set of operations, it is likely that if the number of operations were reduced the server would be able to evaluate them successfully. Therefore, the client is responsible for dealing with this type of complexity in recovery.

The client SHOULD NOT construct a COMPOUND which mixes operations for different client IDs.

#### 15.3. Operation 3: ACCESS - Check Access Rights

##### 15.3.1. SYNOPSIS

(cfh), accessreq -> supported, accessrights

##### 15.3.2. ARGUMENT

```
const ACCESS4_READ      = 0x00000001;
const ACCESS4_LOOKUP    = 0x00000002;
const ACCESS4_MODIFY     = 0x00000004;
const ACCESS4_EXTEND     = 0x00000008;
const ACCESS4_DELETE     = 0x00000010;
const ACCESS4_EXECUTE    = 0x00000020;
```

```
struct ACCESS4args {
    /* CURRENT_FH: object */
    uint32_t      access;
};
```

## 15.3.3. RESULT

```
struct ACCESS4resok {
    uint32_t      supported;
    uint32_t      access;
};

union ACCESS4res switch (nfsstat4 status) {
    case NFS4_OK:
        ACCESS4resok   resok4;
    default:
        void;
};
```

## 15.3.4. DESCRIPTION

ACCESS determines the access rights that a user, as identified by the credentials in the RPC request, has with respect to the file system object specified by the current filehandle. The client encodes the set of access rights that are to be checked in the bit mask "access". The server checks the permissions encoded in the bit mask. If a status of NFS4\_OK is returned, two bit masks are included in the response. The first, "supported", represents the access rights for which the server can verify reliably. The second, "access", represents the access rights available to the user for the filehandle provided. On success, the current filehandle retains its value.

Note that the supported field will contain only as many values as were originally sent in the arguments. For example, if the client sends an ACCESS operation with only the ACCESS4\_READ value set and the server supports this value, the server will return only ACCESS4\_READ even if it could have reliably checked other values.

The results of this operation are necessarily advisory in nature. A return status of NFS4\_OK and the appropriate bit set in the bit mask does not imply that such access will be allowed to the file system object in the future. This is because access rights can be revoked by the server at any time.

The following access permissions may be requested:

ACCESS4\_READ: Read data from file or read a directory.

ACCESS4\_LOOKUP: Look up a name in a directory (no meaning for non-directory objects).

ACCESS4\_MODIFY: Rewrite existing file data or modify existing directory entries.

ACCESS4\_EXTEND: Write new data or add directory entries.

ACCESS4\_DELETE: Delete an existing directory entry.

ACCESS4\_EXECUTE: Execute file (no meaning for a directory).

On success, the current filehandle retains its value.

#### 15.3.5. IMPLEMENTATION

In general, it is not sufficient for the client to attempt to deduce access permissions by inspecting the uid, gid, and mode fields in the file attributes or by attempting to interpret the contents of the ACL attribute. This is because the server may perform uid or gid mapping or enforce additional access control restrictions. It is also possible that the server may not be in the same ID space as the client. In these cases (and perhaps others), the client cannot reliably perform an access check with only current file attributes.

In the NFSv2 protocol, the only reliable way to determine whether an operation was allowed was to try it and see if it succeeded or failed. Using the ACCESS operation in the NFSv4 protocol, the client can ask the server to indicate whether or not one or more classes of operations are permitted. The ACCESS operation is provided to allow clients to check before doing a series of operations which might result in an access failure. The OPEN operation provides a point where the server can verify access to the file object and method to return that information to the client. The ACCESS operation is still useful for directory operations or for use in the case the UNIX API "access" is used on the client.

The information returned by the server in response to an ACCESS call is not permanent. It was correct at the exact time that the server performed the checks, but not necessarily afterward. The server can revoke access permission at any time.

The client should use the effective credentials of the user to build the authentication information in the ACCESS request used to determine access rights. It is the effective user and group credentials that are used in subsequent read and write operations.

Many implementations do not directly support the ACCESS4\_DELETE permission. Operating systems like UNIX will ignore the ACCESS4\_DELETE bit if set on an access request on a non-directory object. In these systems, delete permission on a file is determined

by the access permissions on the directory in which the file resides, instead of being determined by the permissions of the file itself. Therefore, the mask returned enumerating which access rights can be determined will have the ACCESS4\_DELETE value set to 0. This indicates to the client that the server was unable to check that particular access right. The ACCESS4\_DELETE bit in the access mask returned will then be ignored by the client.

#### 15.4. Operation 4: CLOSE - Close File

##### 15.4.1. SYNOPSIS

```
(cfh), seqid, open_stateid -> open_stateid
```

##### 15.4.2. ARGUMENT

```
struct CLOSE4args {  
    /* CURRENT_FH: object */  
    seqid4      seqid;  
    stateid4     open_stateid;  
};
```

##### 15.4.3. RESULT

```
union CLOSE4res switch (nfsstat4 status) {  
    case NFS4_OK:  
        stateid4      open_stateid;  
    default:  
        void;  
};
```

##### 15.4.4. DESCRIPTION

The CLOSE operation releases share reservations for the regular or named attribute file as specified by the current filehandle. The share reservations and other state information released at the server as a result of this CLOSE is only associated with the supplied stateid. The sequence id provides for the correct ordering. State associated with other OPENS is not affected.

If byte-range locks are held, the client SHOULD release all locks before issuing a CLOSE. The server MAY free all outstanding locks on CLOSE but some servers may not support the CLOSE of a file that still has byte-range locks held. The server MUST return failure if any locks would exist after the CLOSE.

On success, the current filehandle retains its value.

#### 15.4.5. IMPLEMENTATION

Even though CLOSE returns a stateid, this stateid is not useful to the client and should be treated as deprecated. CLOSE "shuts down" the state associated with all OPENS for the file by a single open-owner. As noted above, CLOSE will either release all file locking state or return an error. Therefore, the stateid returned by CLOSE is not useful for operations that follow.

### 15.5. Operation 5: COMMIT - Commit Cached Data

#### 15.5.1. SYNOPSIS

(cfh), offset, count -> verifier

#### 15.5.2. ARGUMENT

```
struct COMMIT4args {  
    /* CURRENT_FH: file */  
    offset4      offset;  
    count4      count;  
};
```

#### 15.5.3. RESULT

```
struct COMMIT4resok {  
    verifier4      writeverf;  
};  
  
union COMMIT4res switch (nfsstat4 status) {  
    case NFS4_OK:  
        COMMIT4resok    resok4;  
    default:  
        void;  
};
```

#### 15.5.4. DESCRIPTION

The COMMIT operation forces or flushes data to stable storage for the file specified by the current filehandle. The flushed data is that which was previously written with a WRITE operation which had the stable field set to UNSTABLE4.

The offset specifies the position within the file where the flush is



to begin. An offset value of 0 (zero) means to flush data starting at the beginning of the file. The count specifies the number of bytes of data to flush. If count is 0 (zero), a flush from offset to the end of the file is done.

The server returns a write verifier upon successful completion of the COMMIT. The write verifier is used by the client to determine if the server has restarted or rebooted between the initial WRITE(s) and the COMMIT. The client does this by comparing the write verifier returned from the initial writes and the verifier returned by the COMMIT operation. The server must vary the value of the write verifier at each server event or instantiation that may lead to a loss of uncommitted data. Most commonly this occurs when the server is rebooted; however, other events at the server may result in uncommitted data loss as well.

On success, the current filehandle retains its value.

#### 15.5.5. IMPLEMENTATION

The COMMIT operation is similar in operation and semantics to the POSIX `fsync()` [`fsync`] system call that synchronizes a file's state with the disk (file data and metadata is flushed to disk or stable storage). COMMIT performs the same operation for a client, flushing any unsynchronized data and metadata on the server to the server's disk or stable storage for the specified file. Like `fsync()`, it may be that there is some modified data or no modified data to synchronize. The data may have been synchronized by the server's normal periodic buffer synchronization activity. COMMIT should return `NFS4_OK`, unless there has been an unexpected error.

COMMIT differs from `fsync()` in that it is possible for the client to flush a range of the file (most likely triggered by a buffer-reclamation scheme on the client before file has been completely written).

The server implementation of COMMIT is reasonably simple. If the server receives a full file COMMIT request, that is starting at offset 0 and count 0, it should do the equivalent of `fsync()`'ing the file. Otherwise, it should arrange to have the cached data in the range specified by offset and count to be flushed to stable storage. In both cases, any metadata associated with the file must be flushed to stable storage before returning. It is not an error for there to be nothing to flush on the server. This means that the data and metadata that needed to be flushed have already been flushed or lost during the last server failure.

The client implementation of COMMIT is a little more complex. There

are two reasons for wanting to commit a client buffer to stable storage. The first is that the client wants to reuse a buffer. In this case, the offset and count of the buffer are sent to the server in the COMMIT request. The server then flushes any cached data based on the offset and count, and flushes any metadata associated with the file. It then returns the status of the flush and the write verifier. The other reason for the client to generate a COMMIT is for a full file flush, such as may be done at close. In this case, the client would gather all of the buffers for this file that contain uncommitted data, do the COMMIT operation with an offset of 0 and count of 0, and then free all of those buffers. Any other dirty buffers would be sent to the server in the normal fashion.

After a buffer is written by the client with the stable parameter set to UNSTABLE4, the buffer must be considered as modified by the client until the buffer has either been flushed via a COMMIT operation or written via a WRITE operation with stable parameter set to FILE\_SYNC4 or DATA\_SYNC4. This is done to prevent the buffer from being freed and reused before the data can be flushed to stable storage on the server.

When a response is returned from either a WRITE or a COMMIT operation and it contains a write verifier that is different than previously returned by the server, the client will need to retransmit all of the buffers containing uncommitted cached data to the server. How this is to be done is up to the implementor. If there is only one buffer of interest, then it should probably be sent back over in a WRITE request with the appropriate stable parameter. If there is more than one buffer, it might be worthwhile retransmitting all of the buffers in WRITE requests with the stable parameter set to UNSTABLE4 and then retransmitting the COMMIT operation to flush all of the data on the server to stable storage. The timing of these retransmissions is left to the implementor.

The above description applies to page-cache-based systems as well as buffer-cache-based systems. In those systems, the virtual memory system will need to be modified instead of the buffer cache.

## 15.6. Operation 6: CREATE - Create a Non-Regular File Object

### 15.6.1. SYNOPSIS

```
(cfh), name, type, attrs -> (cfh), cinfo, attrset
```

## 15.6.2. ARGUMENT

```
union createtype4 switch (nfs_ftype4 type) {
    case NF4LNK:
        linktext4 linkdata;
    case NF4BLK:
    case NF4CHR:
        specdata4 devdata;
    case NF4SOCK:
    case NF4FIFO:
    case NF4DIR:
        void;
    default:
        void; /* server should return NFS4ERR_BADTYPE */
};

struct CREATE4args {
    /* CURRENT_FH: directory for creation */
    createtype4    objtype;
    component4     objname;
    fattr4         createattrs;
};
```

## 15.6.3. RESULT

```
struct CREATE4resok {
    change_info4    cinfo;
    bitmap4         attrset; /* attributes set */
};

union CREATE4res switch (nfsstat4 status) {
    case NFS4_OK:
        CREATE4resok resok4;
    default:
        void;
};
```

## 15.6.4. DESCRIPTION

The CREATE operation creates a non-regular file object in a directory with a given name. The OPEN operation MUST be used to create a regular file.

The objname specifies the name for the new object. The objtype determines the type of object to be created: directory, symlink, etc.

If an object of the same name already exists in the directory, the server will return the error NFS4ERR\_EXIST.

For the directory where the new file object was created, the server returns change\_info4 information in cinfo. With the atomic field of the change\_info4 struct, the server will indicate if the before and after change attributes were obtained atomically with respect to the file object creation.

If the objname is of zero length, NFS4ERR\_INVALID will be returned. The objname is also subject to the normal UTF-8, character support, and name checks. See Section 12.5 for further discussion.

If the objname has a length of 0 (zero), or if objname does not obey the UTF-8 definition, the error NFS4ERR\_INVALID will be returned.

The current filehandle is replaced by that of the new object.

The createattrs specifies the initial set of attributes for the object. The set of attributes may include any writable attribute valid for the object type. When the operation is successful, the server will return to the client an attribute mask signifying which attributes were successfully set for the object.

If createattrs includes neither the owner attribute nor an ACL with an ACE for the owner, and if the server's file system both supports and requires an owner attribute (or an owner ACE) then the server MUST derive the owner (or the owner ACE). This would typically be from the principal indicated in the RPC credentials of the call, but the server's operating environment or file system semantics may dictate other methods of derivation. Similarly, if createattrs includes neither the group attribute nor a group ACE, and if the server's file system both supports and requires the notion of a group attribute (or group ACE), the server MUST derive the group attribute (or the corresponding owner ACE) for the file. This could be from the RPC call's credentials, such as the group principal if the credentials include it (such as with AUTH\_SYS), from the group identifier associated with the principal in the credentials (e.g., POSIX systems have a user database [getpwnam] that has the group identifier for every user identifier), inherited from directory the object is created in, or whatever else the server's operating environment or file system semantics dictate. This applies to the OPEN operation too.

Conversely, it is possible the client will specify in createattrs an owner attribute or group attribute or ACL that the principal indicated the RPC call's credentials does not have permissions to create files for. The error to be returned in this instance is

NFS4ERR\_PERM. This applies to the OPEN operation too.

#### 15.6.5. IMPLEMENTATION

If the client desires to set attribute values after the create, a SETATTR operation can be added to the COMPOUND request so that the appropriate attributes will be set.

### 15.7. Operation 7: DELEGPURGE - Purge Delegations Awaiting Recovery

#### 15.7.1. SYNOPSIS

clientid ->

#### 15.7.2. ARGUMENT

```
struct DELEGPURGE4args {  
    clientid4      clientid;  
};
```

#### 15.7.3. RESULT

```
struct DELEGPURGE4res {  
    nfsstat4      status;  
};
```

#### 15.7.4. DESCRIPTION

Purges all of the delegations awaiting recovery for a given client. This is useful for clients which do not commit delegation information to stable storage to indicate that conflicting requests need not be delayed by the server awaiting recovery of delegation information.

This operation is provided to support clients that record delegation information on stable storage on the client. In this case, DELEGPURGE should be issued immediately after doing delegation recovery (using CLAIM\_DELEGATE\_PREV) on all delegations known to the client. Doing so will notify the server that no additional delegations for the client will be recovered allowing it to free resources, and avoid delaying other clients who make requests that conflict with the unrecovered delegations. All client SHOULD use DELEGPURGE as part of recovery once it is known that no further CLAIM\_DELEGATE\_PREV recovery will be done. This includes clients that do not record delegation information on stable storage, who would then do a DELEGPURGE immediately after SETCLIENTID\_CONFIRM.

The set of delegations known to the server and the client may be different. The reasons for this include:

- o A client may fail after making a request which resulted in delegation but before it received the results and committed them to the client's stable storage.
- o A client may fail after deleting its indication that a delegation exists but before the delegation return is fully processed by the server.
- o In the case in which the server and the client restart, the server may have limited persistent recording of delegation to a subset of those in existence.
- o A client may have only persistently recorded information about a subset of delegations.

The server MAY support DELEGPURGE, but its support or non-support should match that of CLAIM\_DELEGATE\_PREV:

- o A server may support both DELEGPURGE and CLAIM\_DELEGATE\_PREV.
- o A server may support neither DELEGPURGE nor CLAIM\_DELEGATE\_PREV.

This fact allows a client starting up to determine if the server is prepared to support persistent storage of delegation information and thus whether it may use write-back caching to local persistent storage, relying on CLAIM\_DELEGATE\_PREV recovery to allow such changed data to be flushed safely to the server in the event of client restart.

## 15.8. Operation 8: DELEGRETURN - Return Delegation

### 15.8.1. SYNOPSIS

(cfh), stateid ->

### 15.8.2. ARGUMENT

```
struct DELEGRETURN4args {  
    /* CURRENT_FH: delegated file */  
    stateid4      deleg_stateid;  
};
```

## 15.8.3. RESULT

```
struct DELEGRETURN4res {  
    nfsstat4      status;  
};
```

## 15.8.4. DESCRIPTION

Returns the delegation represented by the current filehandle and stateid.

Delegations may be returned when recalled or voluntarily (i.e., before the server has recalled them). In either case the client must properly propagate state changed under the context of the delegation to the server before returning the delegation.

## 15.9. Operation 9: GETATTR - Get Attributes

## 15.9.1. SYNOPSIS

(cfh), attrbits -> attrbits, attrvals

## 15.9.2. ARGUMENT

```
struct GETATTR4args {  
    /* CURRENT_FH: directory or file */  
    bitmap4      attr_request;  
};
```

## 15.9.3. RESULT

```
struct GETATTR4resok {  
    fattr4      obj_attributes;  
};  
  
union GETATTR4res switch (nfsstat4 status) {  
    case NFS4_OK:  
        GETATTR4resok  resok4;  
    default:  
        void;  
};
```

#### 15.9.4. DESCRIPTION

The GETATTR operation will obtain attributes for the file system object specified by the current filehandle. The client sets a bit in the bitmap argument for each attribute value that it would like the server to return. The server returns an attribute bitmap that indicates the attribute values for which it was able to return values, followed by the attribute values ordered lowest attribute number first.

The server MUST return a value for each attribute that the client requests if the attribute is supported by the server. If the server does not support an attribute or cannot approximate a useful value then it MUST NOT return the attribute value and MUST NOT set the attribute bit in the result bitmap. The server MUST return an error if it supports an attribute on the target but cannot obtain its value. In that case no attribute values will be returned.

File systems which are absent should be treated as having support for a very small set of attributes as described in GETATTR Within an Absent File System (Section 8.3.1), even if previously, when the file system was present, more attributes were supported.

All servers MUST support the REQUIRED attributes as specified in the section File Attributes (Section 5), for all file systems, with the exception of absent file systems.

On success, the current filehandle retains its value.

#### 15.9.5. IMPLEMENTATION

Suppose there is a OPEN\_DELEGATE\_WRITE delegation held by another client for file in question and size and/or change are among the set of attributes being interrogated. The server has two choices. First, the server can obtain the actual current value of these attributes from the client holding the delegation by using the CB\_GETATTR callback. Second, the server, particularly when the delegated client is unresponsive, can recall the delegation in question. The GETATTR MUST NOT proceed until one of the following occurs:

- o The requested attribute values are returned in the response to CB\_GETATTR.
- o The OPEN\_DELEGATE\_WRITE delegation is returned.
- o The OPEN\_DELEGATE\_WRITE delegation is revoked.



Unless one of the above happens very quickly, one or more NFS4ERR\_DELAY errors will be returned while a delegation is outstanding.

#### 15.10. Operation 10: GETFH - Get Current Filehandle

##### 15.10.1. SYNOPSIS

```
(cfh) -> filehandle
```

##### 15.10.2. ARGUMENT

```
/* CURRENT_FH: */  
void;
```

##### 15.10.3. RESULT

```
struct GETFH4resok {  
    nfs_fh4      object;  
};  
  
union GETFH4res switch (nfsstat4 status) {  
    case NFS4_OK:  
        GETFH4resok      resok4;  
    default:  
        void;  
};
```

##### 15.10.4. DESCRIPTION

This operation returns the current filehandle value.

On success, the current filehandle retains its value.

##### 15.10.5. IMPLEMENTATION

Operations that change the current filehandle like LOOKUP or CREATE do not automatically return the new filehandle as a result. For instance, if a client needs to lookup a directory entry and obtain its filehandle then the following request is needed.

```
PUTFH (directory filehandle)  
LOOKUP (entry name)  
GETFH
```

## 15.11. Operation 11: LINK - Create Link to a File

## 15.11.1. SYNOPSIS

```
(sfh), (cfh), newname -> (cfh), cinfo
```

## 15.11.2. ARGUMENT

```
struct LINK4args {  
    /* SAVED_FH: source object */  
    /* CURRENT_FH: target directory */  
    component4      newname;  
};
```

## 15.11.3. RESULT

```
struct LINK4resok {  
    change_info4      cinfo;  
};  
  
union LINK4res switch (nfsstat4 status) {  
    case NFS4_OK:  
        LINK4resok resok4;  
    default:  
        void;  
};
```

## 15.11.4. DESCRIPTION

The LINK operation creates an additional newname for the file represented by the saved filehandle, as set by the SAVEFH operation, in the directory represented by the current filehandle. The existing file and the target directory must reside within the same file system on the server. On success, the current filehandle will continue to be the target directory. If an object exists in the target directory with the same name as newname, the server must return NFS4ERR\_EXIST.

For the target directory, the server returns change\_info4 information in cinfo. With the atomic field of the change\_info4 struct, the server will indicate if the before and after change attributes were obtained atomically with respect to the link creation.

If the newname has a length of 0 (zero), or if newname does not obey the UTF-8 definition, the error NFS4ERR\_INVALID will be returned.

## 15.11.5. IMPLEMENTATION

Changes to any property of the "hard" linked files are reflected in all of the linked files. When a link is made to a file, the attributes for the file should have a value for numlinks that is one greater than the value before the LINK operation.

The statement "file and the target directory must reside within the same file system on the server" means that the fsid fields in the attributes for the objects are the same. If they reside on different file systems, the error, NFS4ERR\_XDEV, is returned. On some servers, the filenames, "." and "..", are illegal as newname.

In the case that newname is already linked to the file represented by the saved filehandle, the server will return NFS4ERR\_EXIST.

Note that symbolic links are created with the CREATE operation.

## 15.12. Operation 12: LOCK - Create Lock

## 15.12.1. SYNOPSIS

```
(cfh) locktype, reclaim, offset, length, locker -> stateid
```

## 15.12.2. ARGUMENT

```
enum nfs_lock_type4 {  
    READ_LT          = 1,  
    WRITE_LT         = 2,  
    READW_LT         = 3,    /* blocking read */  
    WRITEW_LT        = 4     /* blocking write */  
};
```

```
/*
 * For LOCK, transition from open_owner to new lock_owner
 */
struct open_to_lock_owner4 {
    seqid4      open_seqid;
    stateid4     open_stateid;
    seqid4      lock_seqid;
    lock_owner4  lock_owner;
};

/*
 * For LOCK, existing lock_owner continues to request file locks
 */
struct exist_lock_owner4 {
    stateid4     lock_stateid;
    seqid4      lock_seqid;
};

union locker4 switch (bool new_lock_owner) {
    case TRUE:
        open_to_lock_owner4    open_owner;
    case FALSE:
        exist_lock_owner4      lock_owner;
};

/*
 * LOCK/LOCKT/LOCKU: Record lock management
 */
struct LOCK4args {
    /* CURRENT_FH: file */
    nfs_lock_type4  locktype;
    bool            reclaim;
    offset4         offset;
    length4         length;
    locker4         locker;
};
```

## 15.12.3. RESULT

```
struct LOCK4denied {
    offset4      offset;
    length4      length;
    nfs_lock_type4 locktype;
    lock_owner4  owner;
};

struct LOCK4resok {
    stateid4      lock_stateid;
};

union LOCK4res switch (nfsstat4 status) {
    case NFS4_OK:
        LOCK4resok      resok4;
    case NFS4ERR_DENIED:
        LOCK4denied      denied;
    default:
        void;
};
```

## 15.12.4. DESCRIPTION

The LOCK operation requests a byte-range lock for the byte range specified by the offset and length parameters. The lock type is also specified to be one of the `nfs_lock_type4s`. If this is a reclaim request, the reclaim parameter will be TRUE;

Bytes in a file may be locked even if those bytes are not currently allocated to the file. To lock the file from a specific offset through the end-of-file (no matter how long the file actually is) use a length field with all bits set to 1 (one). If the length is zero, or if a length which is not all bits set to one is specified, and length when added to the offset exceeds the maximum 64-bit unsigned integer value, the error `NFS4ERR_INVALID` will result.

Some servers may only support locking for byte offsets that fit within 32 bits. If the client specifies a range that includes a byte beyond the last byte offset of the 32-bit range, but does not include the last byte offset of the 32-bit and all of the byte offsets beyond it, up to the end of the valid 64-bit range, such a 32-bit server MUST return the error `NFS4ERR_BAD_RANGE`.

In the case that the lock is denied, the owner, offset, and length of a conflicting lock are returned.

On success, the current filehandle retains its value.

#### 15.12.5. IMPLEMENTATION

If the server is unable to determine the exact offset and length of the conflicting lock, the same offset and length that were provided in the arguments should be returned in the denied results. Section 9 contains a full description of this and the other file locking operations.

LOCK operations are subject to permission checks and to checks against the access type of the associated file. However, the specific right and modes required for various type of locks, reflect the semantics of the server-exported file system, and are not specified by the protocol. For example, Windows 2000 allows a write lock of a file open for READ, while a POSIX-compliant system does not.

When the client makes a lock request that corresponds to a range that the lock-owner has locked already (with the same or different lock type), or to a sub-region of such a range, or to a region which includes multiple locks already granted to that lock-owner, in whole or in part, and the server does not support such locking operations (i.e., does not support POSIX locking semantics), the server will return the error NFS4ERR\_LOCK\_RANGE. In that case, the client may return an error, or it may emulate the required operations, using only LOCK for ranges that do not include any bytes already locked by that lock-owner and LOCKU of locks held by that lock-owner (specifying an exactly-matching range and type). Similarly, when the client makes a lock request that amounts to upgrading (changing from a read lock to a write lock) or downgrading (changing from write lock to a read lock) an existing record lock, and the server does not support such a lock, the server will return NFS4ERR\_LOCK\_NOTSUPP. Such operations may not perfectly reflect the required semantics in the face of conflicting lock requests from other clients.

When a client holds an OPEN\_DELEGATE\_WRITE delegation, the client holding that delegation is assured that there are no opens by other clients. Thus, there can be no conflicting LOCK operations from such clients. Therefore, the client may be handling locking requests locally, without doing LOCK operations on the server. If it does that, it must be prepared to update the lock status on the server, by sending appropriate LOCK and LOCKU operations before returning the delegation.

When one or more clients hold OPEN\_DELEGATE\_READ delegations, any LOCK operation where the server is implementing mandatory locking semantics MUST result in the recall of all such delegations. The

LOCK operation may not be granted until all such delegations are returned or revoked. Except where this happens very quickly, one or more NFS4ERR\_DELAY errors will be returned to requests made while the delegation remains outstanding.

The locker argument specifies the lock-owner that is associated with the LOCK request. The locker4 structure is a switched union that indicates whether the client has already created byte-range locking state associated with the current open file and lock-owner. There are multiple cases to be considered, corresponding to possible combinations of whether locking state has been created for the current open file and lock-owner, and whether the boolean new\_lock\_owner is set. In all of the cases, there is a lock\_seqid specified, whether the lock-owner is specified explicitly or implicitly. This seqid value is used for checking lock-owner sequencing/replay issues. When the given lock-owner is not known to the server, this establishes an initial sequence value for the new lock-owner.

- o In the case in which the state has been created and the boolean is false, the only part of the argument other than lock\_seqid is just a stateid representing the set of locks associated with that open file and lock-owner.
- o In the case in which the state has been created and the boolean is true, the server rejects the request with the error NFS4ERR\_BAD\_SEQID. The only exception is where there is a retransmission of a previous request in which the boolean was true. In this case, the lock\_seqid will match the original request and the response will reflect the final case, below.
- o In the case where no byte-range locking state has been established and the boolean is true, the argument contains an open\_to\_lock\_owner structure which specifies the stateid of the open file and the lock-owner to be used for the lock. Note that although the open-owner is not given explicitly, the open\_seqid associated with it is used to check for open-owner sequencing issues. This case provides a method to use the established state of the open\_stateid to transition to the use of a lock stateid.

### 15.13. Operation 13: LOCKT - Test For Lock

#### 15.13.1. SYNOPSIS

```
(cfh) locktype, offset, length, owner -> {void, NFS4ERR_DENIED ->
owner}
```

## 15.13.2. ARGUMENT

```
struct LOCKT4args {
    /* CURRENT_FH: file */
    nfs_lock_type4  locktype;
    offset4         offset;
    length4         length;
    lock_owner4     owner;
};
```

## 15.13.3. RESULT

```
union LOCKT4res switch (nfsstat4 status) {
    case NFS4ERR_DENIED:
        LOCK4denied    denied;
    case NFS4_OK:
        void;
    default:
        void;
};
```

## 15.13.4. DESCRIPTION

The LOCKT operation tests the lock as specified in the arguments. If a conflicting lock exists, the owner, offset, length, and type of the conflicting lock are returned; if no lock is held, nothing other than NFS4\_OK is returned. Lock types READ\_LT and READW\_LT are processed in the same way in that a conflicting lock test is done without regard to blocking or non-blocking. The same is true for WRITE\_LT and WRITEW\_LT.

The ranges are specified as for LOCK. The NFS4ERR\_INVALID and NFS4ERR\_BAD\_RANGE errors are returned under the same circumstances as for LOCK.

On success, the current filehandle retains its value.

## 15.13.5. IMPLEMENTATION

If the server is unable to determine the exact offset and length of the conflicting lock, the same offset and length that were provided in the arguments should be returned in the denied results. Section 9 contains further discussion of the file locking mechanisms.

LOCKT uses a lock\_owner4 rather a stateid4, as is used in LOCK to identify the owner. This is because the client does not have to open



the file to test for the existence of a lock, so a stateid may not be available.

The test for conflicting locks SHOULD exclude locks for the current lock-owner. Note that since such locks are not examined the possible existence of overlapping ranges may not affect the results of LOCKT. If the server does examine locks that match the lock-owner for the purpose of range checking, NFS4ERR\_LOCK\_RANGE may be returned. In the event that it returns NFS4\_OK, clients may do a LOCK and receive NFS4ERR\_LOCK\_RANGE on the LOCK request because of the flexibility provided to the server.

When a client holds an OPEN\_DELEGATE\_WRITE delegation, it may choose (see Section 15.12.5)) to handle LOCK requests locally. In such a case, LOCKT requests will similarly be handled locally.

#### 15.14. Operation 14: LOCKU - Unlock File

##### 15.14.1. SYNOPSIS

(cfh) type, seqid, stateid, offset, length -> stateid

##### 15.14.2. ARGUMENT

```
struct LOCKU4args {
    /* CURRENT_FH: file */
    nfs_lock_type4  locktype;
    seqid4          seqid;
    stateid4        lock_stateid;
    offset4         offset;
    length4         length;
};
```

##### 15.14.3. RESULT

```
union LOCKU4res switch (nfsstat4 status) {
    case NFS4_OK:
        stateid4        lock_stateid;
    default:
        void;
};
```

#### 15.14.4. DESCRIPTION

The LOCKU operation unlocks the byte-range lock specified by the parameters. The client may set the locktype field to any value that is legal for the nfs\_lock\_type4 enumerated type, and the server MUST accept any legal value for locktype. Any legal value for locktype has no effect on the success or failure of the LOCKU operation.

The ranges are specified as for LOCK. The NFS4ERR\_INVALID and NFS4ERR\_BAD\_RANGE errors are returned under the same circumstances as for LOCK.

On success, the current filehandle retains its value.

#### 15.14.5. IMPLEMENTATION

If the area to be unlocked does not correspond exactly to a lock actually held by the lock-owner the server may return the error NFS4ERR\_LOCK\_RANGE. This includes the case in which the area is not locked, where the area is a sub-range of the area locked, where it overlaps the area locked without matching exactly or the area specified includes multiple locks held by the lock-owner. In all of these cases, allowed by POSIX locking [fcntl] semantics, a client receiving this error, should if it desires support for such operations, simulate the operation using LOCKU on ranges corresponding to locks it actually holds, possibly followed by LOCK requests for the sub-ranges not being unlocked.

When a client holds an OPEN\_DELEGATE\_WRITE delegation, it may choose (see Section 15.12.5)) to handle LOCK requests locally. In such a case, LOCKU requests will similarly be handled locally.

### 15.15. Operation 15: LOOKUP - Lookup Filename

#### 15.15.1. SYNOPSIS

(cfh), component -> (cfh)

#### 15.15.2. ARGUMENT

```
struct LOOKUP4args {  
    /* CURRENT_FH: directory */  
    component4      objname;  
};
```

## 15.15.3. RESULT

```
struct LOOKUP4res {  
    /* CURRENT_FH: object */  
    nfsstat4      status;  
};
```

## 15.15.4. DESCRIPTION

This operation LOOKUPS or finds a file system object using the directory specified by the current filehandle. LOOKUP evaluates the component and if the object exists the current filehandle is replaced with the component's filehandle.

If the component cannot be evaluated either because it does not exist or because the client does not have permission to evaluate the component, then an error will be returned and the current filehandle will be unchanged.

If the component is of zero length, NFS4ERR\_INVALID will be returned. The component is also subject to the normal UTF-8, character support, and name checks. See Section 12.5 for further discussion.

## 15.15.5. IMPLEMENTATION

If the client wants to achieve the effect of a multi-component lookup, it may construct a COMPOUND request such as (and obtain each filehandle):

```
PUTFH (directory filehandle)  
LOOKUP "pub"  
GETFH  
LOOKUP "foo"  
GETFH  
LOOKUP "bar"  
GETFH
```

NFSv4 servers depart from the semantics of previous NFS versions in allowing LOOKUP requests to cross mount points on the server. The client can detect a mount point crossing by comparing the fsid attribute of the directory with the fsid attribute of the directory looked up. If the fsids are different then the new directory is a server mount point. UNIX clients that detect a mount point crossing will need to mount the server's file system. This needs to be done to maintain the file object identity checking mechanisms common to UNIX clients.

Servers that limit NFS access to "shares" or "exported" file systems should provide a pseudo-file system into which the exported file systems can be integrated, so that clients can browse the server's name space. The clients' view of a pseudo file system will be limited to paths that lead to exported file systems.

Note: previous versions of the protocol assigned special semantics to the names "." and "..". NFSv4 assigns no special semantics to these names. The LOOKUPP operator must be used to lookup a parent directory.

Note that this operation does not follow symbolic links. The client is responsible for all parsing of filenames including filenames that are modified by symbolic links encountered during the lookup process.

If the current filehandle supplied is not a directory but a symbolic link, the error NFS4ERR\_SYMLINK is returned as the error. For all other non-directory file types, the error NFS4ERR\_NOTDIR is returned.

#### 15.16. Operation 16: LOOKUPP - Lookup Parent Directory

##### 15.16.1. SYNOPSIS

```
(cfh) -> (cfh)
```

##### 15.16.2. ARGUMENT

```
/* CURRENT_FH: object */  
void;
```

##### 15.16.3. RESULT

```
struct LOOKUPP4res {  
    /* CURRENT_FH: directory */  
    nfsstat4      status;  
};
```

##### 15.16.4. DESCRIPTION

The current filehandle is assumed to refer to a regular directory or a named attribute directory. LOOKUPP assigns the filehandle for its parent directory to be the current filehandle. If there is no parent directory an NFS4ERR\_NOENT error must be returned. Therefore, NFS4ERR\_NOENT will be returned by the server when the current filehandle is at the root or top of the server's file tree.

## 15.16.5. IMPLEMENTATION

As for LOOKUP, LOOKUPP will also cross mount points.

If the current filehandle is not a directory or named attribute directory, the error NFS4ERR\_NOTDIR is returned.

## 15.17. Operation 17: NVERIFY - Verify Difference in Attributes

## 15.17.1. SYNOPSIS

(cfh), fattr -> -

## 15.17.2. ARGUMENT

```
struct NVERIFY4args {  
    /* CURRENT_FH: object */  
    fattr4          obj_attributes;  
};
```

## 15.17.3. RESULT

```
struct NVERIFY4res {  
    nfsstat4          status;  
};
```

## 15.17.4. DESCRIPTION

This operation is used to prefix a sequence of operations to be performed if one or more attributes have changed on some file system object. If all the attributes match then the error NFS4ERR\_SAME must be returned.

On success, the current filehandle retains its value.

## 15.17.5. IMPLEMENTATION

This operation is useful as a cache validation operator. If the object to which the attributes belong has changed then the following operations may obtain new data associated with that object. For instance, to check if a file has been changed and obtain new data if it has:

```
PUTFH (public)  
LOOKUP "foobar"  
NVERIFY attrbits attrs
```

READ 0 32767

In the case that a recommended attribute is specified in the NVERIFY operation and the server does not support that attribute for the file system object, the error NFS4ERR\_ATTRNOTSUPP is returned to the client.

When the attribute rdattnr\_error or any write-only attribute (e.g., time\_modify\_set) is specified, the error NFS4ERR\_INVALID is returned to the client.

## 15.18. Operation 18: OPEN - Open a Regular File

### 15.18.1. SYNOPSIS

(cfh), seqid, share\_access, share\_deny, owner, openhow, claim ->  
(cfh), stateid, cinfo, rflags, attrset, delegation

### 15.18.2. ARGUMENT

```
/*
 * Various definitions for OPEN
 */
enum createmode4 {
    UNCHECKED4      = 0,
    GUARDED4        = 1,
    EXCLUSIVE4      = 2
};

union createhow4 switch (createmode4 mode) {
    case UNCHECKED4:
    case GUARDED4:
        fattnr4      createattnr;
    case EXCLUSIVE4:
        verifier4    createverf;
};

enum opentype4 {
    OPEN4_NOCREATE  = 0,
    OPEN4_CREATE    = 1
};

union openflag4 switch (opentype4 opentype) {
    case OPEN4_CREATE:
        createhow4    how;
    default:
        void;
};
```

```
/* Next definitions used for OPEN delegation */
enum limit_by4 {
    NFS_LIMIT_SIZE          = 1,
    NFS_LIMIT_BLOCKS        = 2
    /* others as needed */
};

struct nfs_modified_limit4 {
    uint32_t      num_blocks;
    uint32_t      bytes_per_block;
};

union nfs_space_limit4 switch (limit_by4 limitby) {
    /* limit specified as file size */
    case NFS_LIMIT_SIZE:
        uint64_t      filesize;
    /* limit specified by number of blocks */
    case NFS_LIMIT_BLOCKS:
        nfs_modified_limit4      mod_blocks;
} ;

enum open_delegation_type4 {
    OPEN_DELEGATE_NONE      = 0,
    OPEN_DELEGATE_READ      = 1,
    OPEN_DELEGATE_WRITE     = 2
};

enum open_claim_type4 {
    CLAIM_NULL              = 0,
    CLAIM_PREVIOUS          = 1,
    CLAIM_DELEGATE_CUR      = 2,
    CLAIM_DELEGATE_PREV     = 3
};

struct open_claim_delegate_cur4 {
    stateid4      delegate_stateid;
    component4    file;
};

union open_claim4 switch (open_claim_type4 claim) {
    /*
     * No special rights to file.
     * Ordinary OPEN of the specified file.
     */
    case CLAIM_NULL:
        /* CURRENT_FH: directory */
        component4      file;
    /*
```

```

    * Right to the file established by an
    * open previous to server reboot. File
    * identified by filehandle obtained at
    * that time rather than by name.
    */
case CLAIM_PREVIOUS:
    /* CURRENT_FH: file being reclaimed */
    open_delegation_type4    delegate_type;

/*
 * Right to file based on a delegation
 * granted by the server. File is
 * specified by name.
 */
case CLAIM_DELEGATE_CUR:
    /* CURRENT_FH: directory */
    open_claim_delegate_cur4    delegate_cur_info;

/*
 * Right to file based on a delegation
 * granted to a previous boot instance
 * of the client. File is specified by name.
 */
case CLAIM_DELEGATE_PREV:
    /* CURRENT_FH: directory */
    component4    file_delegate_prev;
};

/*
 * OPEN: Open a file, potentially receiving an open delegation
 */
struct OPEN4args {
    seqid4        seqid;
    uint32_t      share_access;
    uint32_t      share_deny;
    open_owner4   owner;
    openflag4     openhow;
    open_claim4   claim;
};

```

### 15.18.3. RESULT

```

struct open_read_delegation4 {
    stateid4 stateid;    /* Stateid for delegation*/
    bool      recall;    /* Pre-recalled flag for
                        delegations obtained
                        by reclaim (CLAIM_PREVIOUS) */
};

```



```
nfsace4 permissions; /* Defines users who don't
                      need an ACCESS call to
                      open for read */
};

struct open_write_delegation4 {
    stateid4 stateid; /* Stateid for delegation */
    bool      recall; /* Pre-recalled flag for
                      delegations obtained
                      by reclaim
                      (CLAIM_PREVIOUS) */

    nfs_space_limit4
        space_limit; /* Defines condition that
                      the client must check to
                      determine whether the
                      file needs to be flushed
                      to the server on close. */

    nfsace4 permissions; /* Defines users who don't
                          need an ACCESS call as
                          part of a delegated
                          open. */
};

union open_delegation4
switch (open_delegation_type4 delegation_type) {
    case OPEN_DELEGATE_NONE:
        void;
    case OPEN_DELEGATE_READ:
        open_read_delegation4 read;
    case OPEN_DELEGATE_WRITE:
        open_write_delegation4 write;
};

/*
 * Result flags
 */

/* Client must confirm open */
const OPEN4_RESULT_CONFIRM = 0x00000002;
/* Type of file locking behavior at the server */
const OPEN4_RESULT_LOCKTYPE_POSIX = 0x00000004;

struct OPEN4resok {
    stateid4 stateid; /* Stateid for open */
    change_info4 cinfo; /* Directory Change Info */
    uint32_t rflags; /* Result flags */
};
```

```
    bitmap4      attrset;      /* attribute set for create*/
    open_delegation4 delegation; /* Info on any open
                                delegation */
};

union OPEN4res switch (nfsstat4 status) {
    case NFS4_OK:
        /* CURRENT_FH: opened file */
        OPEN4resok      resok4;
    default:
        void;
};
```

#### 15.18.4. Warning to Client Implementors

OPEN resembles LOOKUP in that it generates a filehandle for the client to use. Unlike LOOKUP though, OPEN creates server state on the filehandle. In normal circumstances, the client can only release this state with a CLOSE operation. CLOSE uses the current filehandle to determine which file to close. Therefore, the client MUST follow every OPEN operation with a GETFH operation in the same COMPOUND procedure. This will supply the client with the filehandle such that CLOSE can be used appropriately.

Simply waiting for the lease on the file to expire is insufficient because the server may maintain the state indefinitely as long as another client does not attempt to make a conflicting access to the same file.

#### 15.18.5. DESCRIPTION

The OPEN operation creates and/or opens a regular file in a directory with the provided name. If the file does not exist at the server and creation is desired, specification of the method of creation is provided by the openhow parameter. The client has the choice of three creation methods: UNCHECKED4, GUARDED4, or EXCLUSIVE4.

If the current filehandle is a named attribute directory, OPEN will then create or open a named attribute file. Note that exclusive create of a named attribute is not supported. If the createmode is EXCLUSIVE4 and the current filehandle is a named attribute directory, the server will return EINVAL.

UNCHECKED4 means that the file should be created if a file of that name does not exist and encountering an existing regular file of that name is not an error. For this type of create, createattrs specifies the initial set of attributes for the file. The set of attributes

may include any writable attribute valid for regular files. When an UNCHECKED4 create encounters an existing file, the attributes specified by createattrs are not used, except that when an size of zero is specified, the existing file is truncated. If GUARDED4 is specified, the server checks for the presence of a duplicate object by name before performing the create. If a duplicate exists, an error of NFS4ERR\_EXIST is returned as the status. If the object does not exist, the request is performed as described for UNCHECKED4. For each of these cases (UNCHECKED4 and GUARDED4) where the operation is successful, the server will return to the client an attribute mask signifying which attributes were successfully set for the object.

EXCLUSIVE4 specifies that the server is to follow exclusive creation semantics, using the verifier to ensure exclusive creation of the target. The server should check for the presence of a duplicate object by name. If the object does not exist, the server creates the object and stores the verifier with the object. If the object does exist and the stored verifier matches the client provided verifier, the server uses the existing object as the newly created object. If the stored verifier does not match, then an error of NFS4ERR\_EXIST is returned. No attributes may be provided in this case, since the server may use an attribute of the target object to store the verifier. If the server uses an attribute to store the exclusive create verifier, it will signify which attribute by setting the appropriate bit in the attribute mask that is returned in the results.

For the target directory, the server returns change\_info4 information in cinfo. With the atomic field of the change\_info4 struct, the server will indicate if the before and after change attributes were obtained atomically with respect to the link creation.

Upon successful creation, the current filehandle is replaced by that of the new object.

The OPEN operation provides for Windows share reservation capability with the use of the share\_access and share\_deny fields of the OPEN arguments. The client specifies at OPEN the required share\_access and share\_deny modes. For clients that do not directly support SHARES (i.e., UNIX), the expected deny value is DENY\_NONE. In the case that there is a existing SHARE reservation that conflicts with the OPEN request, the server returns the error NFS4ERR\_SHARE\_DENIED. For a complete SHARE request, the client must provide values for the owner and seqid fields for the OPEN argument. For additional discussion of SHARE semantics see Section 9.9.

In the case that the client is recovering state from a server failure, the claim field of the OPEN argument is used to signify that

the request is meant to reclaim state previously held.

The "claim" field of the OPEN argument is used to specify the file to be opened and the state information which the client claims to possess. There are four basic claim types which cover the various situations for an OPEN. They are as follows:

CLAIM\_NULL: For the client, this is a new OPEN request and there is no previous state associate with the file for the client.

CLAIM\_PREVIOUS: The client is claiming basic OPEN state for a file that was held previous to a server reboot. Generally used when a server is returning persistent filehandles; the client may not have the file name to reclaim the OPEN.

CLAIM\_DELEGATE\_CUR: The client is claiming a delegation for OPEN as granted by the server. Generally this is done as part of recalling a delegation.

CLAIM\_DELEGATE\_PREV: The client is claiming a delegation granted to a previous client instance. This claim type is for use after a SETCLIENTID\_CONFIRM and before the corresponding DELEGPURGE in two situations: after a client reboot and after a lease expiration that resulted in loss of all lock state. The server MAY support CLAIM\_DELEGATE\_PREV. If it does support CLAIM\_DELEGATE\_PREV, SETCLIENTID\_CONFIRM MUST NOT remove the client's delegation state, and the server MUST support the DELEGPURGE operation.

The following errors apply to use of the CLAIM\_DELEGATE\_PREV claim type:

- o NFS4ERR\_NOTSUPP is returned if the server does not support this claim type.
- o NFS4ERR\_INVALID is returned if the reclaim is done at an inappropriate time, e.g., after DELEGPURGE has been done.
- o NFS4ERR\_BAD\_RECLAIM is returned if the other error conditions do not apply and the server has no record of the delegation whose reclaim is being attempted.

For OPEN requests whose claim type is other than CLAIM\_PREVIOUS (i.e., requests other than those devoted to reclaiming opens after a server reboot) that reach the server during its grace or lease expiration period, the server returns an error of NFS4ERR\_GRACE.

For any OPEN request, the server may return an open delegation, which allows further opens and closes to be handled locally on the client

as described in Section 10.4. Note that delegation is up to the server to decide. The client should never assume that delegation will or will not be granted in a particular instance. It should always be prepared for either case. A partial exception is the reclaim (CLAIM\_PREVIOUS) case, in which a delegation type is claimed. In this case, delegation will always be granted, although the server may specify an immediate recall in the delegation structure.

The rflags returned by a successful OPEN allow the server to return information governing how the open file is to be handled.

OPEN4\_RESULT\_CONFIRM indicates that the client MUST execute an OPEN\_CONFIRM operation before using the open file.  
OPEN4\_RESULT\_LOCKTYPE\_POSIX indicates the server's file locking behavior supports the complete set of Posix locking techniques [fcntl]. From this the client can choose to manage file locking state in a way to handle a mis-match of file locking management.

If the component is of zero length, NFS4ERR\_INVALID will be returned. The component is also subject to the normal UTF-8, character support, and name checks. See Section 12.5 for further discussion.

When an OPEN is done and the specified open-owner already has the resulting filehandle open, the result is to "OR" together the new share and deny status together with the existing status. In this case, only a single CLOSE need be done, even though multiple OPENs were completed. When such an OPEN is done, checking of share reservations for the new OPEN proceeds normally, with no exception for the existing OPEN held by the same owner. In this case, the stateid returned as an "other" field that matches that of the previous open while the "seqid" field is incremented to reflect the change status due to the new open.

If the underlying file system at the server is only accessible in a read-only mode and the OPEN request has specified ACCESS\_WRITE or ACCESS\_BOTH, the server will return NFS4ERR\_ROFS to indicate a read-only file system.

As with the CREATE operation, the server MUST derive the owner, owner ACE, group, or group ACE if any of the four attributes are required and supported by the server's file system. For an OPEN with the EXCLUSIVE4 createmode, the server has no choice, since such OPEN calls do not include the createattrs field. Conversely, if createattrs is specified, and includes owner or group (or corresponding ACEs) that the principal in the RPC call's credentials does not have authorization to create files for, then the server may return NFS4ERR\_PERM.

In the case of a OPEN which specifies a size of zero (e.g., truncation) and the file has named attributes, the named attributes are left as is. They are not removed.

#### 15.18.6. IMPLEMENTATION

The OPEN operation contains support for EXCLUSIVE4 create. The mechanism is similar to the support in NFSv3 [RFC1813]. As in NFSv3, this mechanism provides reliable exclusive creation. Exclusive create is invoked when the how parameter is EXCLUSIVE4. In this case, the client provides a verifier that can reasonably be expected to be unique. A combination of a client identifier, perhaps the client network address, and a unique number generated by the client, perhaps the RPC transaction identifier, may be appropriate.

If the object does not exist, the server creates the object and stores the verifier in stable storage. For file systems that do not provide a mechanism for the storage of arbitrary file attributes, the server may use one or more elements of the object meta-data to store the verifier. The verifier must be stored in stable storage to prevent erroneous failure on retransmission of the request. It is assumed that an exclusive create is being performed because exclusive semantics are critical to the application. Because of the expected usage, exclusive CREATE does not rely solely on the normally volatile duplicate request cache for storage of the verifier. The duplicate request cache in volatile storage does not survive a crash and may actually flush on a long network partition, opening failure windows. In the UNIX local file system environment, the expected storage location for the verifier on creation is the meta-data (time stamps) of the object. For this reason, an exclusive object create may not include initial attributes because the server would have nowhere to store the verifier.

If the server cannot support these exclusive create semantics, possibly because of the requirement to commit the verifier to stable storage, it should fail the OPEN request with the error, NFS4ERR\_NOTSUPP.

During an exclusive CREATE request, if the object already exists, the server reconstructs the object's verifier and compares it with the verifier in the request. If they match, the server treats the request as a success. The request is presumed to be a duplicate of an earlier, successful request for which the reply was lost and that the server duplicate request cache mechanism did not detect. If the verifiers do not match, the request is rejected with the status, NFS4ERR\_EXIST.

Once the client has performed a successful exclusive create, it must

issue a SETATTR to set the correct object attributes. Until it does so, it should not rely upon any of the object attributes, since the server implementation may need to overload object meta-data to store the verifier. The subsequent SETATTR must not occur in the same COMPOUND request as the OPEN. This separation will guarantee that the exclusive create mechanism will continue to function properly in the face of retransmission of the request.

Use of the GUARDED4 attribute does not provide exactly-once semantics. In particular, if a reply is lost and the server does not detect the retransmission of the request, the operation can fail with NFS4ERR\_EXIST, even though the create was performed successfully. The client would use this behavior in the case that the application has not requested an exclusive create but has asked to have the file truncated when the file is opened. In the case of the client timing out and retransmitting the create request, the client can use GUARDED4 to prevent against a sequence like: create, write, create (retransmitted) from occurring.

For SHARE reservations, the client must specify a value for share\_access that is one of READ, WRITE, or BOTH. For share\_deny, the client must specify one of NONE, READ, WRITE, or BOTH. If the client fails to do this, the server must return NFS4ERR\_INVALID.

Based on the share\_access value (READ, WRITE, or BOTH) the client should check that the requester has the proper access rights to perform the specified operation. This would generally be the results of applying the ACL access rules to the file for the current requester. However, just as with the ACCESS operation, the client should not attempt to second-guess the server's decisions, as access rights may change and may be subject to server administrative controls outside the ACL framework. If the requester is not authorized to READ or WRITE (depending on the share\_access value), the server must return NFS4ERR\_ACCESS. Note that since the NFS version 4 protocol does not impose any requirement that READs and WRITES issued for an open file have the same credentials as the OPEN itself, the server still must do appropriate access checking on the READs and WRITES themselves.

If the component provided to OPEN resolves to something other than a regular file, an error will be returned to the client. If it is a directory, NFS4ERR\_ISDIR is returned; otherwise, NFS4ERR\_SYMLINK is returned. Note that NFS4ERR\_SYMLINK is returned for both symlinks and for special files of other types; NFS4ERR\_INVALID would be inappropriate, since the arguments provided by the client were correct, and the client cannot necessarily know at the time it sent the OPEN that the component would resolve to a non-regular file.

If the current filehandle is not a directory, the error NFS4ERR\_NOTDIR will be returned.

If a COMPOUND contains an OPEN which establishes a OPEN\_DELEGATE\_WRITE delegation, then a subsequent GETATTR inside that COMPOUND SHOULD NOT result in a CB\_GETATTR to the client. The server SHOULD understand the GETATTR to be for the same client ID and avoid querying the client, which will not be able to respond. This sequence of OPEN, GETATTR SHOULD be understood as an atomic retrieval of the initial size and change attribute. Further, the client SHOULD NOT construct a COMPOUND which mixes operations for different client IDs.

#### 15.19. Operation 19: OPENATTR - Open Named Attribute Directory

##### 15.19.1. SYNOPSIS

(cfh) createdir -> (cfh)

##### 15.19.2. ARGUMENT

```
struct OPENATTR4args {  
    /* CURRENT_FH: object */  
    bool    createdir;  
};
```

##### 15.19.3. RESULT

```
struct OPENATTR4res {  
    /* CURRENT_FH: named attr directory */  
    nfsstat4    status;  
};
```

##### 15.19.4. DESCRIPTION

The OPENATTR operation is used to obtain the filehandle of the named attribute directory associated with the current filehandle. The result of the OPENATTR will be a filehandle to an object of type NF4ATTRDIR. From this filehandle, READDIR and LOOKUP operations can be used to obtain filehandles for the various named attributes associated with the original file system object. Filehandles returned within the named attribute directory will have a type of NF4NAMEDATTR.

The createdir argument allows the client to signify if a named attribute directory should be created as a result of the OPENATTR



operation. Some clients may use the OPENATTR operation with a value of FALSE for createdir to determine if any named attributes exist for the object. If none exist, then NFS4ERR\_NOENT will be returned. If createdir has a value of TRUE and no named attribute directory exists, one is created. The creation of a named attribute directory assumes that the server has implemented named attribute support in this fashion and is not required to do so by this definition.

#### 15.19.5. IMPLEMENTATION

If the server does not support named attributes for the current filehandle, an error of NFS4ERR\_NOTSUPP will be returned to the client.

#### 15.20. Operation 20: OPEN\_CONFIRM - Confirm Open

##### 15.20.1. SYNOPSIS

(cfh), seqid, stateid -> stateid

##### 15.20.2. ARGUMENT

```
struct OPEN_CONFIRM4args {
    /* CURRENT_FH: opened file */
    stateid4      open_stateid;
    seqid4        seqid;
};
```

##### 15.20.3. RESULT

```
struct OPEN_CONFIRM4resok {
    stateid4      open_stateid;
};

union OPEN_CONFIRM4res switch (nfsstat4 status) {
    case NFS4_OK:
        OPEN_CONFIRM4resok      resok4;
    default:
        void;
};
```

##### 15.20.4. DESCRIPTION

This operation is used to confirm the sequence id usage for the first time that a open-owner is used by a client. The stateid returned from the OPEN operation is used as the argument for this operation

along with the next sequence id for the open-owner. The sequence id passed to the OPEN\_CONFIRM must be 1 (one) greater than the seqid passed to the OPEN operation. If the server receives an unexpected sequence id with respect to the original open, then the server assumes that the client will not confirm the original OPEN and all state associated with the original OPEN is released by the server.

On success, the current filehandle retains its value.

#### 15.20.5. IMPLEMENTATION

A given client might generate many open\_owner4 data structures for a given client ID. The client will periodically either dispose of its open\_owner4s or stop using them for indefinite periods of time. The latter situation is why the NFSv4 protocol does not have an explicit operation to exit an open\_owner4: such an operation is of no use in that situation. Instead, to avoid unbounded memory use, the server needs to implement a strategy for disposing of open\_owner4s that have no current open state for any files and have not been used recently. The time period used to determine when to dispose of open\_owner4s is an implementation choice. The time period should certainly be no less than the lease time plus any grace period the server wishes to implement beyond a lease time. The OPEN\_CONFIRM operation allows the server to safely dispose of unused open\_owner4 data structures.

In the case that a client issues an OPEN operation and the server no longer has a record of the open\_owner4, the server needs to ensure that this is a new OPEN and not a replay or retransmission.

Servers MUST NOT require confirmation on OPENs that grant delegations or are doing reclaim operations. See Section 9.1.10 for details. The server can easily avoid this by noting whether it has disposed of one open\_owner4 for the given client ID. If the server does not support delegation, it might simply maintain a single bit that notes whether any open\_owner4 (for any client) has been disposed of.

The server must hold unconfirmed OPEN state until one of three events occur. First, the client sends an OPEN\_CONFIRM request with the appropriate sequence id and stateid within the lease period. In this case, the OPEN state on the server goes to confirmed, and the open\_owner4 on the server is fully established.

Second, the client sends another OPEN request with a sequence id that is incorrect for the open\_owner4 (out of sequence). In this case, the server assumes the second OPEN request is valid and the first one is a replay. The server cancels the OPEN state of the first OPEN request, establishes an unconfirmed OPEN state for the second OPEN request, and responds to the second OPEN request with an indication

that an OPEN\_CONFIRM is needed. The process then repeats itself. While there is a potential for a denial of service attack on the client, it is mitigated if the client and server require the use of a security flavor based on Kerberos V5 or some other flavor that uses cryptography.

What if the server is in the unconfirmed OPEN state for a given open\_owner4, and it receives an operation on the open\_owner4 that has a stateid but the operation is not OPEN, or it is OPEN\_CONFIRM but with the wrong stateid? Then, even if the seqid is correct, the server returns NFS4ERR\_BAD\_STATEID, because the server assumes the operation is a replay: if the server has no established OPEN state, then there is no way, for example, a LOCK operation could be valid.

Third, neither of the two aforementioned events occur for the open\_owner4 within the lease period. In this case, the OPEN state is canceled and disposal of the open\_owner4 can occur.

#### 15.21. Operation 21: OPEN\_DOWNGRADE - Reduce Open File Access

##### 15.21.1. SYNOPSIS

(cfh), stateid, seqid, access, deny -> stateid

##### 15.21.2. ARGUMENT

```
struct OPEN_DOWNGRADE4args {
    /* CURRENT_FH: opened file */
    stateid4      open_stateid;
    seqid4        seqid;
    uint32_t      share_access;
    uint32_t      share_deny;
};
```

##### 15.21.3. RESULT

```
struct OPEN_DOWNGRADE4resok {
    stateid4      open_stateid;
};

union OPEN_DOWNGRADE4res switch(nfsstat4 status) {
    case NFS4_OK:
        OPEN_DOWNGRADE4resok      resok4;
    default:
        void;
};
```

## 15.21.4. DESCRIPTION

This operation is used to adjust the `share_access` and `share_deny` bits for a given open. This is necessary when a given open-owner opens the same file multiple times with different `share_access` and `share_deny` flags. In this situation, a close of one of the opens may change the appropriate `share_access` and `share_deny` flags to remove bits associated with opens no longer in effect.

The `share_access` and `share_deny` bits specified in this operation replace the current ones for the specified open file. The `share_access` and `share_deny` bits specified must be exactly equal to the union of the `share_access` and `share_deny` bits specified for some subset of the OPENs in effect for current open-owner on the current file. If that constraint is not respected, the error `NFS4ERR_INVALID` should be returned. Since `share_access` and `share_deny` bits are subsets of those already granted, it is not possible for this request to be denied because of conflicting share reservations.

As the `OPEN_DOWNGRADE` may change a file to be not-open-for-write and a write byte-range lock might be held, the server may have to reject the `OPEN_DOWNGRADE` with a `NFS4ERR_LOCKS_HELD`.

On success, the current filehandle retains its value.

## 15.22. Operation 22: PUTFH - Set Current Filehandle

## 15.22.1. SYNOPSIS

```
filehandle -> (cfh)
```

## 15.22.2. ARGUMENT

```
struct PUTFH4args {  
    nfs_fh4      object;  
};
```

## 15.22.3. RESULT

```
struct PUTFH4res {  
    /* CURRENT_FH: */  
    nfsstat4      status;  
};
```

#### 15.22.4. DESCRIPTION

Replaces the current filehandle with the filehandle provided as an argument.

If the security mechanism used by the requester does not meet the requirements of the filehandle provided to this operation, the server MUST return NFS4ERR\_WRONGSEC.

See Section 15.2.4.1 for more details on the current filehandle.

#### 15.22.5. IMPLEMENTATION

Commonly used as the first operator in an NFS request to set the context for following operations.

#### 15.23. Operation 23: PUTPUBFH - Set Public Filehandle

##### 15.23.1. SYNOPSIS

- -> (cfh)

##### 15.23.2. ARGUMENT

void;

##### 15.23.3. RESULT

```
struct PUTPUBFH4res {  
    /* CURRENT_FH: public fh */  
    nfsstat4      status;  
};
```

##### 15.23.4. DESCRIPTION

Replaces the current filehandle with the filehandle that represents the public filehandle of the server's name space. This filehandle may be different from the "root" filehandle which may be associated with some other directory on the server.

The public filehandle represents the concepts embodied in [RFC2054], [RFC2055], [RFC2224]. The intent for NFSv4 is that the public filehandle (represented by the PUTPUBFH operation) be used as a method of providing WebNFS server compatibility with NFSv2 and NFSv3.

The public filehandle and the root filehandle (represented by the PUTROOTFH operation) should be equivalent. If the public and root

filehandles are not equivalent, then the public filehandle MUST be a descendant of the root filehandle.

#### 15.23.5. IMPLEMENTATION

Used as the first operator in an NFS request to set the context for following operations.

With the NFSv2 and 3 public filehandle, the client is able to specify whether the path name provided in the LOOKUP should be evaluated as either an absolute path relative to the server's root or relative to the public filehandle. [RFC2224] contains further discussion of the functionality. With NFSv4, that type of specification is not directly available in the LOOKUP operation. The reason for this is because the component separators needed to specify absolute vs. relative are not allowed in NFSv4. Therefore, the client is responsible for constructing its request such that the use of either PUTROOTFH or PUTPUBFH are used to signify absolute or relative evaluation of an NFS URL respectively.

Note that there are warnings mentioned in [RFC2224] with respect to the use of absolute evaluation and the restrictions the server may place on that evaluation with respect to how much of its namespace has been made available. These same warnings apply to NFSv4. It is likely, therefore that because of server implementation details, an NFSv3 absolute public filehandle lookup may behave differently than an NFSv4 absolute resolution.

There is a form of security negotiation as described in [RFC2755] that uses the public filehandle a method of employing Simple and Protected GSSAPI Negotiation Mechanism (SNEGO) [RFC4178]. This method is not available with NFSv4 as filehandles are not overloaded with special meaning and therefore do not provide the same framework as NFSv2 and NFSv3. Clients should therefore use the security negotiation mechanisms described in this RFC.

#### 15.24. Operation 24: PUTROOTFH - Set Root Filehandle

##### 15.24.1. SYNOPSIS

- -> (cfh)

##### 15.24.2. ARGUMENT

void;

## 15.24.3. RESULT

```
struct PUTROOTFH4res {  
    /* CURRENT_FH: root fh */  
    nfsstat4      status;  
};
```

## 15.24.4. DESCRIPTION

Replaces the current filehandle with the filehandle that represents the root of the server's name space. From this filehandle a LOOKUP operation can locate any other filehandle on the server. This filehandle may be different from the "public" filehandle which may be associated with some other directory on the server.

See Section 15.2.4.1 for more details on the current filehandle.

## 15.24.5. IMPLEMENTATION

Commonly used as the first operator in an NFS request to set the context for following operations.

## 15.25. Operation 25: READ - Read from File

## 15.25.1. SYNOPSIS

(cfh), stateid, offset, count -> eof, data

## 15.25.2. ARGUMENT

```
struct READ4args {  
    /* CURRENT_FH: file */  
    stateid4      stateid;  
    offset4       offset;  
    count4        count;  
};
```

## 15.25.3. RESULT

```
struct READ4resok {
    bool        eof;
    opaque      data<>;
};

union READ4res switch (nfsstat4 status) {
    case NFS4_OK:
        READ4resok    resok4;
    default:
        void;
};
```

## 15.25.4. DESCRIPTION

The READ operation reads data from the regular file identified by the current filehandle.

The client provides an offset of where the READ is to start and a count of how many bytes are to be read. An offset of 0 (zero) means to read data starting at the beginning of the file. If offset is greater than or equal to the size of the file, the status, NFS4\_OK, is returned with a data length set to 0 (zero) and eof is set to TRUE. The READ is subject to access permissions checking.

If the client specifies a count value of 0 (zero), the READ succeeds and returns 0 (zero) bytes of data again subject to access permissions checking. The server may choose to return fewer bytes than specified by the client. The client needs to check for this condition and handle the condition appropriately.

The stateid value for a READ request represents a value returned from a previous byte-range lock or share reservation request or the stateid associated with a delegation. The stateid is used by the server to verify that the associated share reservation and any byte-range locks are still valid and to update lease timeouts for the client.

If the read ended at the end-of-file (formally, in a correctly formed READ request, if offset + count is equal to the size of the file), or the read request extends beyond the size of the file (if offset + count is greater than the size of the file), eof is returned as TRUE; otherwise it is FALSE. A successful READ of an empty file will always return eof as TRUE.

If the current filehandle is not a regular file, an error will be



returned to the client. In the case the current filehandle represents a directory, NFS4ERR\_ISDIR is returned; otherwise, NFS4ERR\_INVALID is returned.

For a READ with a stateid value of all bits 0, the server MAY allow the READ to be serviced subject to mandatory file locks or the current share deny modes for the file. For a READ with a stateid value of all bits 1, the server MAY allow READ operations to bypass locking checks at the server.

On success, the current filehandle retains its value.

#### 15.25.5. IMPLEMENTATION

If the server returns a "short read" (i.e., fewer data than requested and eof is set to FALSE), the client should send another READ to get the remaining data. A server may return less data than requested under several circumstances. The file may have been truncated by another client or perhaps on the server itself, changing the file size from what the requesting client believes to be the case. This would reduce the actual amount of data available to the client. It is possible that the server reduces the transfer size and so returns a short read result. Server resource exhaustion may also result in a short read.

If mandatory byte-range locking is in effect for the file, and if the byte-range corresponding to the data to be read from the file is WRITE\_LT locked by an owner not associated with the stateid, the server will return the NFS4ERR\_LOCKED error. The client should try to get the appropriate READ\_LT via the LOCK operation before reattempting the READ. When the READ completes, the client should release the byte-range lock via LOCKU.

If another client has an OPEN\_DELEGATE\_WRITE delegation for the file being read, the delegation must be recalled, and the operation cannot proceed until that delegation is returned or revoked. Except where this happens very quickly, one or more NFS4ERR\_DELAY errors will be returned to requests made while the delegation remains outstanding. Normally, delegations will not be recalled as a result of a READ operation since the recall will occur as a result of an earlier OPEN. However, since it is possible for a READ to be done with a special stateid, the server needs to check for this case even though the client should have done an OPEN previously.

#### 15.26. Operation 26: READDIR - Read Directory

## 15.26.1. SYNOPSIS

```
(cfh), cookie, cookieverf, dircount, maxcount, attr_request ->
cookieverf { cookie, name, attrs }
```

## 15.26.2. ARGUMENT

```
struct READDIR4args {
    /* CURRENT_FH: directory */
    nfs_cookie4      cookie;
    verifier4        cookieverf;
    count4           dircount;
    count4           maxcount;
    bitmap4          attr_request;
};
```

## 15.26.3. RESULT

```
struct entry4 {
    nfs_cookie4      cookie;
    component4       name;
    fattr4           attrs;
    entry4           *nextentry;
};

struct dirlist4 {
    entry4           *entries;
    bool             eof;
};

struct READDIR4resok {
    verifier4        cookieverf;
    dirlist4         reply;
};

union READDIR4res switch (nfsstat4 status) {
    case NFS4_OK:
        READDIR4resok  resok4;
    default:
        void;
};
```

## 15.26.4. DESCRIPTION

The READDIR operation retrieves a variable number of entries from a file system directory and returns client requested attributes for each entry along with information to allow the client to request additional directory entries in a subsequent READDIR.

The arguments contain a cookie value that represents where the READDIR should start within the directory. A value of 0 (zero) for the cookie is used to start reading at the beginning of the directory. For subsequent READDIR requests, the client specifies a cookie value that is provided by the server on a previous READDIR request.

The cookieverf value should be set to 0 (zero) when the cookie value is 0 (zero) (first directory read). On subsequent requests, it should be a cookieverf as returned by the server. The cookieverf must match that returned by the READDIR in which the cookie was acquired. If the server determines that the cookieverf is no longer valid for the directory, the error NFS4ERR\_NOT\_SAME must be returned.

The dircount portion of the argument is a hint of the maximum number of bytes of directory information that should be returned. This value represents the length of the names of the directory entries and the cookie value for these entries. This length represents the XDR encoding of the data (names and cookies) and not the length in the native format of the server.

The maxcount value of the argument is the maximum number of bytes for the result. This maximum size represents all of the data being returned within the READDIR4resok structure and includes the XDR overhead. The server may return less data. If the server is unable to return a single directory entry within the maxcount limit, the error NFS4ERR\_TOOSMALL will be returned to the client.

Finally, attr\_request represents the list of attributes to be returned for each directory entry supplied by the server.

On successful return, the server's response will provide a list of directory entries. Each of these entries contains the name of the directory entry, a cookie value for that entry, and the associated attributes as requested. The "eof" flag has a value of TRUE if there are no more entries in the directory.

The cookie value is only meaningful to the server and is used as a "bookmark" for the directory entry. As mentioned, this cookie is used by the client for subsequent READDIR operations so that it may continue reading a directory. The cookie is similar in concept to a

READ offset but should not be interpreted as such by the client. Ideally, the cookie value should not change if the directory is modified since the client may be caching these values.

In some cases, the server may encounter an error while obtaining the attributes for a directory entry. Instead of returning an error for the entire READDIR operation, the server can instead return the attribute 'fattr4\_rdattrib\_error'. With this, the server is able to communicate the failure to the client and not fail the entire operation in the instance of what might be a transient failure. Obviously, the client must request the fattr4\_rdattrib\_error attribute for this method to work properly. If the client does not request the attribute, the server has no choice but to return failure for the entire READDIR operation.

For some file system environments, the directory entries "." and ".." have special meaning and in other environments, they may not. If the server supports these special entries within a directory, they should not be returned to the client as part of the READDIR response. To enable some client environments, the cookie values of 0, 1, and 2 are to be considered reserved. Note that the UNIX client will use these values when combining the server's response and local representations to enable a fully formed UNIX directory presentation to the application.

For READDIR arguments, cookie values of 1 and 2 SHOULD NOT be used and for READDIR results cookie values of 0, 1, and 2 MUST NOT be returned.

On success, the current filehandle retains its value.

#### 15.26.5. IMPLEMENTATION

The server's file system directory representations can differ greatly. A client's programming interfaces may also be bound to the local operating environment in a way that does not translate well into the NFS protocol. Therefore the use of the dircount and maxcount fields are provided to allow the client the ability to provide guidelines to the server. If the client is aggressive about attribute collection during a READDIR, the server has an idea of how to limit the encoded response. The dircount field provides a hint on the number of entries based solely on the names of the directory entries. Since it is a hint, it may be possible that a dircount value is zero. In this case, the server is free to ignore the dircount value and return directory information based on the specified maxcount value.

The cookieverf may be used by the server to help manage cookie values

that may become stale. It should be a rare occurrence that a server is unable to continue properly reading a directory with the provided cookie/cookieverf pair. The server should make every effort to avoid this condition since the application at the client may not be able to properly handle this type of failure.

The use of the cookieverf will also protect the client from using READDIR cookie values that may be stale. For example, if the file system has been migrated, the server may or may not be able to use the same cookie values to service READDIR as the previous server used. With the client providing the cookieverf, the server is able to provide the appropriate response to the client. This prevents the case where the server may accept a cookie value but the underlying directory has changed and the response is invalid from the client's context of its previous READDIR.

Since some servers will not be returning "." and ".." entries as has been done with previous versions of the NFS protocol, the client that requires these entries be present in READDIR responses must fabricate them.

## 15.27. Operation 27: READLINK - Read Symbolic Link

### 15.27.1. SYNOPSIS

```
(cfh) -> linktext
```

### 15.27.2. ARGUMENT

```
/* CURRENT_FH: symlink */  
void;
```

### 15.27.3. RESULT

```
struct READLINK4resok {  
    linktext4    link;  
};  
  
union READLINK4res switch (nfsstat4 status) {  
    case NFS4_OK:  
        READLINK4resok resok4;  
    default:  
        void;  
};
```

## 15.27.4. DESCRIPTION

READLINK reads the data associated with a symbolic link. The data is a UTF-8 string that is opaque to the server. That is, whether created by an NFS client or created locally on the server, the data in a symbolic link is not interpreted when created, but is simply stored.

On success, the current filehandle retains its value.

## 15.27.5. IMPLEMENTATION

A symbolic link is nominally a pointer to another file. The data is not necessarily interpreted by the server, just stored in the file. It is possible for a client implementation to store a path name that is not meaningful to the server operating system in a symbolic link. A READLINK operation returns the data to the client for interpretation. If different implementations want to share access to symbolic links, then they must agree on the interpretation of the data in the symbolic link.

The READLINK operation is only allowed on objects of type NF4LNK. The server should return the error, NFS4ERR\_INVALID, if the object is not of type, NF4LNK.

## 15.28. Operation 28: REMOVE - Remove Filesystem Object

## 15.28.1. SYNOPSIS

(cfh), filename -> change\_info

## 15.28.2. ARGUMENT

```
struct REMOVE4args {  
    /* CURRENT_FH: directory */  
    component4      target;  
};
```

## 15.28.3. RESULT

```
struct REMOVE4resok {
    change_info4    cinfo;
};

union REMOVE4res switch (nfsstat4 status) {
    case NFS4_OK:
        REMOVE4resok    resok4;
    default:
        void;
};
```

## 15.28.4. DESCRIPTION

The REMOVE operation removes (deletes) a directory entry M named by filename from the directory corresponding to the current filehandle. If the entry in the directory was the last reference to the corresponding file system object, the object may be destroyed.

For the directory where the filename was removed, the server returns change\_info4 information in cinfo. With the atomic field of the change\_info4 struct, the server will indicate if the before and after change attributes were obtained atomically with respect to the removal.

If the target is of zero length, NFS4ERR\_INVALID will be returned. The target is also subject to the normal UTF-8, character support, and name checks. See Section 12.5 for further discussion.

On success, the current filehandle retains its value.

## 15.28.5. IMPLEMENTATION

NFSv3 required a different operator RMDIR for directory removal and REMOVE for non-directory removal. This allowed clients to skip checking the file type when being passed a non-directory delete system call (e.g., unlink() [unlink] in POSIX) to remove a directory, as well as the converse (e.g., a rmdir() on a non-directory) because they knew the server would check the file type. NFSv4 REMOVE can be used to delete any directory entry independent of its file type. The implementor of an NFSv4 client's entry points from the unlink() and rmdir() system calls should first check the file type against the types the system call is allowed to remove before issuing a REMOVE. Alternatively, the implementor can produce a COMPOUND call that includes a LOOKUP/VERIFY sequence to verify the file type before a REMOVE operation in the same COMPOUND call.

The concept of last reference is server specific. However, if the `numlinks` field in the previous attributes of the object had the value 1, the client should not rely on referring to the object via a filehandle. Likewise, the client should not rely on the resources (disk space, directory entry, and so on) formerly associated with the object becoming immediately available. Thus, if a client needs to be able to continue to access a file after using `REMOVE` to remove it, the client should take steps to make sure that the file will still be accessible. The usual mechanism used is to `RENAME` the file from its old name to a new hidden name.

If the server finds that the file is still open when the `REMOVE` arrives:

- o The server **SHOULD NOT** delete the file's directory entry if the file was opened with `OPEN4_SHARE_DENY_WRITE` or `OPEN4_SHARE_DENY_BOTH`.
- o If the file was not opened with `OPEN4_SHARE_DENY_WRITE` or `OPEN4_SHARE_DENY_BOTH`, the server **SHOULD** delete the file's directory entry. However, until last `CLOSE` of the file, the server **MAY** continue to allow access to the file via its filehandle.

## 15.29. Operation 29: `RENAME` - Rename Directory Entry

### 15.29.1. SYNOPSIS

```
(sfh), oldname, (cfh), newname -> source_cinfo, target_cinfo
```

### 15.29.2. ARGUMENT

```
struct RENAME4args {  
    /* SAVED_FH: source directory */  
    component4      oldname;  
    /* CURRENT_FH: target directory */  
    component4      newname;  
};
```



## 15.29.3. RESULT

```
struct RENAME4resok {
    change_info4    source_cinfo;
    change_info4    target_cinfo;
};

union RENAME4res switch (nfsstat4 status) {
    case NFS4_OK:
        RENAME4resok    resok4;
    default:
        void;
};
```

## 15.29.4. DESCRIPTION

The RENAME operation renames the object identified by oldname in the source directory corresponding to the saved filehandle, as set by the SAVEFH operation, to newname in the target directory corresponding to the current filehandle. The operation is required to be atomic to the client. Source and target directories must reside on the same file system on the server. On success, the current filehandle will continue to be the target directory.

If the target directory already contains an entry with the name, newname, the source object must be compatible with the target: either both are non-directories or both are directories and the target must be empty. If compatible, the existing target is removed before the rename occurs (See Section 15.28 for client and server actions whenever a target is removed). If they are not compatible or if the target is a directory but not empty, the server will return the error, NFS4ERR\_EXIST.

If oldname and newname both refer to the same file (they might be hard links of each other), then RENAME should perform no action and return success.

For both directories involved in the RENAME, the server returns change\_info4 information. With the atomic field of the change\_info4 struct, the server will indicate if the before and after change attributes were obtained atomically with respect to the rename.

If the oldname refers to a named attribute and the saved and current filehandles refer to different file system objects, the server will return NFS4ERR\_XDEV just as if the saved and current filehandles represented directories on different file systems.

If the oldname or newname is of zero length, NFS4ERR\_INVALID will be returned. The oldname and newname are also subject to the normal UTF-8, character support, and name checks. See Section 12.5 for further discussion.

#### 15.29.5. IMPLEMENTATION

The RENAME operation must be atomic to the client. The statement "source and target directories must reside on the same file system on the server" means that the fsid fields in the attributes for the directories are the same. If they reside on different file systems, the error, NFS4ERR\_XDEV, is returned.

Based on the value of the fh\_expire\_type attribute for the object, the filehandle may or may not expire on a RENAME. However, server implementors are strongly encouraged to attempt to keep filehandles from expiring in this fashion.

On some servers, the file names "." and ".." are illegal as either oldname or newname, and will result in the error NFS4ERR\_BADNAME. In addition, on many servers the case of oldname or newname being an alias for the source directory will be checked for. Such servers will return the error NFS4ERR\_INVALID in these cases.

If either of the source or target filehandles are not directories, the server will return NFS4ERR\_NOTDIR.

### 15.30. Operation 30: RENEW - Renew a Lease

#### 15.30.1. SYNOPSIS

```
clientid -> ()
```

#### 15.30.2. ARGUMENT

```
struct RENEW4args {  
    clientid4      clientid;  
};
```

#### 15.30.3. RESULT

```
struct RENEW4res {  
    nfsstat4      status;  
};
```

#### 15.30.4. DESCRIPTION

The RENEW operation is used by the client to renew leases which it currently holds at a server. In processing the RENEW request, the server renews all leases associated with the client. The associated leases are determined by the clientid provided via the SETCLIENTID operation.

#### 15.30.5. IMPLEMENTATION

When the client holds delegations, it needs to use RENEW to detect when the server has determined that the callback path is down. When the server has made such a determination, only the RENEW operation will renew the lease on delegations. If the server determines the callback path is down, it returns NFS4ERR\_CB\_PATH\_DOWN. Even though it returns NFS4ERR\_CB\_PATH\_DOWN, the server MUST renew the lease on the byte-range locks and share reservations that the client has established on the server. If for some reason the lock and share reservation lease cannot be renewed, then the server MUST return an error other than NFS4ERR\_CB\_PATH\_DOWN, even if the callback path is also down. In the event that the server has conditions such that it could return either NFS4ERR\_CB\_PATH\_DOWN or NFS4ERR\_LEASE\_MOVED, NFS4ERR\_LEASE\_MOVED MUST be handled first.

The client that issues RENEW MUST choose the principal, RPC security flavor, and if applicable, GSS-API mechanism and service via one of the following algorithms:

- o The client uses the same principal, RPC security flavor -- and if the flavor was RPCSEC\_GSS -- the same mechanism and service that was used when the client id was established via SETCLIENTID\_CONFIRM.
- o The client uses any principal, RPC security flavor mechanism and service combination that currently has an OPEN file on the server. I.e., the same principal had a successful OPEN operation, the file is still open by that principal, and the flavor, mechanism, and service of RENEW match that of the previous OPEN.

The server MUST reject a RENEW that does not use one the aforementioned algorithms, with the error NFS4ERR\_ACCESS.

#### 15.31. Operation 31: RESTOREFH - Restore Saved Filehandle

##### 15.31.1. SYNOPSIS

(sfh) -> (cfh)

## 15.31.2. ARGUMENT

```
/* SAVED_FH: */  
void;
```

## 15.31.3. RESULT

```
struct RESTOREFH4res {  
    /* CURRENT_FH: value of saved fh */  
    nfsstat4      status;  
};
```

## 15.31.4. DESCRIPTION

Set the current filehandle to the value in the saved filehandle. If there is no saved filehandle then return the error NFS4ERR\_RESTOREFH.

## 15.31.5. IMPLEMENTATION

Operations like OPEN and LOOKUP use the current filehandle to represent a directory and replace it with a new filehandle. Assuming the previous filehandle was saved with a SAVEFH operator, the previous filehandle can be restored as the current filehandle. This is commonly used to obtain post-operation attributes for the directory, e.g.,

```
PUTFH (directory filehandle)  
SAVEFH  
GETATTR attrbits      (pre-op dir attrs)  
CREATE optbits "foo" attrs  
GETATTR attrbits      (file attributes)  
RESTOREFH  
GETATTR attrbits      (post-op dir attrs)
```

## 15.32. Operation 32: SAVEFH - Save Current Filehandle

## 15.32.1. SYNOPSIS

```
(cfh) -> (sfh)
```

## 15.32.2. ARGUMENT

```
/* CURRENT_FH: */  
void;
```

## 15.32.3. RESULT

```
struct SAVEFH4res {  
    /* SAVED_FH: value of current fh */  
    nfsstat4      status;  
};
```

## 15.32.4. DESCRIPTION

Save the current filehandle. If a previous filehandle was saved then it is no longer accessible. The saved filehandle can be restored as the current filehandle with the RESTOREFH operator.

On success, the current filehandle retains its value.

## 15.32.5. IMPLEMENTATION

## 15.33. Operation 33: SECINFO - Obtain Available Security

## 15.33.1. SYNOPSIS

```
(cfh), name -> { secinfo }
```

## 15.33.2. ARGUMENT

```
struct SECINFO4args {  
    /* CURRENT_FH: directory */  
    component4      name;  
};
```

## 15.33.3. RESULT

```

/*
 * From RFC 2203
 */
enum rpc_gss_svc_t {
    RPC_GSS_SVC_NONE          = 1,
    RPC_GSS_SVC_INTEGRITY     = 2,
    RPC_GSS_SVC_PRIVACY       = 3
};

struct rpcsec_gss_info {
    sec_oid4      oid;
    qop4          qop;
    rpc_gss_svc_t service;
};

/* RPCSEC_GSS has a value of '6' - See RFC 2203 */
union secinfo4 switch (uint32_t flavor) {
    case RPCSEC_GSS:
        rpcsec_gss_info      flavor_info;
    default:
        void;
};

typedef secinfo4 SECINFO4resok<>;

union SECINFO4res switch (nfsstat4 status) {
    case NFS4_OK:
        SECINFO4resok resok4;
    default:
        void;
};

```

## 15.33.4. DESCRIPTION

The SECINFO operation is used by the client to obtain a list of valid RPC authentication flavors for a specific directory filehandle, file name pair. SECINFO should apply the same access methodology used for LOOKUP when evaluating the name. Therefore, if the requester does not have the appropriate access to LOOKUP the name then SECINFO must behave the same way and return NFS4ERR\_ACCESS.

The result will contain an array which represents the security mechanisms available, with an order corresponding to server's preferences, the most preferred being first in the array. The client is free to pick whatever security mechanism it both desires and

supports, or to pick in the server's preference order the first one it supports. The array entries are represented by the `secinfo4` structure. The field 'flavor' will contain a value of `AUTH_NONE`, `AUTH_SYS` (as defined in [RFC5531]), or `RPCSEC_GSS` (as defined in [RFC2203]).

For the flavors `AUTH_NONE` and `AUTH_SYS`, no additional security information is returned. For a return value of `RPCSEC_GSS`, a security triple is returned that contains the mechanism object id (as defined in [RFC2743]), the quality of protection (as defined in [RFC2743]) and the service type (as defined in [RFC2203]). It is possible for `SECINFO` to return multiple entries with flavor equal to `RPCSEC_GSS` with different security triple values.

On success, the current filehandle retains its value.

If the name has a length of 0 (zero), or if name does not obey the UTF-8 definition, the error `NFS4ERR_INVALID` will be returned.

#### 15.33.5. IMPLEMENTATION

The `SECINFO` operation is expected to be used by the NFS client when the error value of `NFS4ERR_WRONGSEC` is returned from another NFS operation. This signifies to the client that the server's security policy is different from what the client is currently using. At this point, the client is expected to obtain a list of possible security flavors and choose what best suits its policies.

As mentioned, the server's security policies will determine when a client request receives `NFS4ERR_WRONGSEC`. The operations which may receive this error are: `LINK`, `LOOKUP`, `LOOKUPP`, `OPEN`, `PUTFH`, `PUTPUBFH`, `PUTROOTFH`, `RENAME`, `RESTOREFH`, and indirectly `READDIR`. `LINK` and `RENAME` will only receive this error if the security used for the operation is inappropriate for saved filehandle. With the exception of `READDIR`, these operations represent the point at which the client can instantiate a filehandle into the "current filehandle" at the server. The filehandle is either provided by the client (`PUTFH`, `PUTPUBFH`, `PUTROOTFH`) or generated as a result of a name to filehandle translation (`LOOKUP` and `OPEN`). `RESTOREFH` is different because the filehandle is a result of a previous `SAVEFH`. Even though the filehandle, for `RESTOREFH`, might have previously passed the server's inspection for a security match, the server will check it again on `RESTOREFH` to ensure that the security policy has not changed.

If the client wants to resolve an error return of `NFS4ERR_WRONGSEC`, the following will occur:

- o For LOOKUP and OPEN, the client will use SECINFO with the same current filehandle and name as provided in the original LOOKUP or OPEN to enumerate the available security triples.
- o For LINK, PUTFH, RENAME, and RESTOREFH, the client will use SECINFO and provide the parent directory filehandle and object name which corresponds to the filehandle originally provided by the PUTFH RESTOREFH, or for LINK and RENAME, the SAVEFH.
- o For LOOKUPP, PUTROOTFH and PUTPUBFH, the client will be unable to use the SECINFO operation since SECINFO requires a current filehandle and none exist for these two operations. Therefore, the client must iterate through the security triples available at the client and reattempt the PUTROOTFH or PUTPUBFH operation. In the unfortunate event none of the MANDATORY security triples are supported by the client and server, the client SHOULD try using others that support integrity. Failing that, the client can try using AUTH\_NONE, but because such forms lack integrity checks, this puts the client at risk. Nonetheless, the server SHOULD allow the client to use whatever security form the client requests and the server supports, since the risks of doing so are on the client.

The READDIR operation will not directly return the NFS4ERR\_WRONGSEC error. However, if the READDIR request included a request for attributes, it is possible that the READDIR request's security triple does not match that of a directory entry. If this is the case and the client has requested the rdatatr\_error attribute, the server will return the NFS4ERR\_WRONGSEC error in rdatatr\_error for the entry.

Note that a server MAY use the AUTH\_NONE flavor to signify that the client is allowed to attempt to use authentication flavors that are not explicitly listed in the SECINFO results. Instead of using a listed flavor, the client might then, for instance opt to use an otherwise unlisted RPCSEC\_GSS mechanism instead of AUTH\_NONE. It may wish to do so in order to meet an application requirement for data integrity or privacy. In choosing to use an unlisted flavor, the client SHOULD always be prepared to handle a failure by falling back to using AUTH\_NONE or another listed flavor. It MUST NOT assume that identity mapping is supported, and should be prepared for the fact that its identity is squashed.

See Section 17 for a discussion on the recommendations for security flavor used by SECINFO.



## 15.34. Operation 34: SETATTR - Set Attributes

## 15.34.1. SYNOPSIS

```
(cfh), stateid, attrmask, attr_vals -> attrset
```

## 15.34.2. ARGUMENT

```
struct SETATTR4args {  
    /* CURRENT_FH: target object */  
    stateid4      stateid;  
    fattr4        obj_attributes;  
};
```

## 15.34.3. RESULT

```
struct SETATTR4res {  
    nfsstat4      status;  
    bitmap4       attrset;  
};
```

## 15.34.4. DESCRIPTION

The SETATTR operation changes one or more of the attributes of a file system object. The new attributes are specified with a bitmap and the attributes that follow the bitmap in bit order.

The stateid argument for SETATTR is used to provide byte-range locking context that is necessary for SETATTR requests that set the size attribute. Since setting the size attribute modifies the file's data, it has the same locking requirements as a corresponding WRITE. Any SETATTR that sets the size attribute is incompatible with a share reservation that specifies OPEN4\_SHARE\_DENY\_WRITE. The area between the old end-of-file and the new end-of-file is considered to be modified just as would have been the case had the area in question been specified as the target of WRITE, for the purpose of checking conflicts with byte-range locks, for those cases in which a server is implementing mandatory byte-range locking behavior. A valid stateid SHOULD always be specified. When the file size attribute is not set, the special stateid consisting of all bits zero MAY be passed.

On either success or failure of the operation, the server will return the attrset bitmask to represent what (if any) attributes were successfully set. The attrset in the response is a subset of the bitmap4 that is part of the obj\_attributes in the argument.

On success, the current filehandle retains its value.

#### 15.34.5. IMPLEMENTATION

If the request specifies the owner attribute to be set, the server SHOULD allow the operation to succeed if the current owner of the object matches the value specified in the request. Some servers may be implemented in a way as to prohibit the setting of the owner attribute unless the requester has privilege to do so. If the server is lenient in this one case of matching owner values, the client implementation may be simplified in cases of creation of an object (e.g., an exclusive create via OPEN) followed by a SETATTR.

The file size attribute is used to request changes to the size of a file. A value of zero causes the file to be truncated, a value less than the current size of the file causes data from new size to the end of the file to be discarded, and a size greater than the current size of the file causes logically zeroed data bytes to be added to the end of the file. Servers are free to implement this using holes or actual zero data bytes. Clients should not make any assumptions regarding a server's implementation of this feature, beyond that the bytes returned will be zeroed. Servers MUST support extending the file size via SETATTR.

SETATTR is not guaranteed atomic. A failed SETATTR may partially change a file's attributes, hence the reason why the reply always includes the status and the list of attributes that were set.

If the object whose attributes are being changed has a file delegation that is held by a client other than the one doing the SETATTR, the delegation(s) must be recalled, and the operation cannot proceed to actually change an attribute until each such delegation is returned or revoked. In all cases in which delegations are recalled, the server is likely to return one or more NFS4ERR\_DELAY errors while the delegation(s) remains outstanding, although it might not do that if the delegations are returned quickly.

Changing the size of a file with SETATTR indirectly changes the time\_modify and change attributes. A client must account for this as size changes can result in data deletion.

The attributes time\_access\_set and time\_modify\_set are write-only attributes constructed as a switched union so the client can direct the server in setting the time values. If the switched union specifies SET\_TO\_CLIENT\_TIME4, the client has provided an nfstime4 to be used for the operation. If the switch union does not specify SET\_TO\_CLIENT\_TIME4, the server is to use its current time for the SETATTR operation.

If server and client times differ, programs that compare client time to file times can break. A time maintenance protocol should be used to limit client/server time skew.

Use of a COMPOUND containing a VERIFY operation specifying only the change attribute, immediately followed by a SETATTR, provides a means whereby a client may specify a request that emulates the functionality of the SETATTR guard mechanism of NFSv3. Since the function of the guard mechanism is to avoid changes to the file attributes based on stale information, delays between checking of the guard condition and the setting of the attributes have the potential to compromise this function, as would the corresponding delay in the NFSv4 emulation. Therefore, NFSv4 servers should take care to avoid such delays, to the degree possible, when executing such a request.

If the server does not support an attribute as requested by the client, the server should return NFS4ERR\_ATTRNOTSUPP.

A mask of the attributes actually set is returned by SETATTR in all cases. That mask MUST NOT include attribute bits not requested to be set by the client. If the attribute masks in the request and reply are equal, the status field in the reply MUST be NFS4\_OK.

#### 15.35. Operation 35: SETCLIENTID - Negotiate Client ID

##### 15.35.1. SYNOPSIS

client, callback, callback\_ident -> clientid, setclientid\_confirm

##### 15.35.2. ARGUMENT

```
struct SETCLIENTID4args {
    nfs_client_id4  client;
    cb_client4      callback;
    uint32_t        callback_ident;
};
```

## 15.35.3. RESULT

```
struct SETCLIENTID4resok {
    clientid4      clientid;
    verifier4      setclientid_confirm;
};

union SETCLIENTID4res switch (nfsstat4 status) {
    case NFS4_OK:
        SETCLIENTID4resok      resok4;
    case NFS4ERR_CLID_INUSE:
        clientaddr4      client_using;
    default:
        void;
};
```

## 15.35.4. DESCRIPTION

The client uses the SETCLIENTID operation to notify the server of its intention to use a particular client identifier, callback, and callback\_ident for subsequent requests that entail creating lock, share reservation, and delegation state on the server. Upon successful completion the server will return a shorthand client ID which, if confirmed via a separate step, will be used in subsequent file locking and file open requests. Confirmation of the client ID must be done via the SETCLIENTID\_CONFIRM operation to return the client ID and setclientid\_confirm values, as verifiers, to the server. The reason why two verifiers are necessary is that it is possible to use SETCLIENTID and SETCLIENTID\_CONFIRM to modify the callback and callback\_ident information but not the shorthand client ID. In that event, the setclientid\_confirm value is effectively the only verifier.

The callback information provided in this operation will be used if the client is provided an open delegation at a future point. Therefore, the client must correctly reflect the program and port numbers for the callback program at the time SETCLIENTID is used.

The callback\_ident value is used by the server on the callback. The client can leverage the callback\_ident to eliminate the need for more than one callback RPC program number, while still being able to determine which server is initiating the callback.

## 15.35.5. IMPLEMENTATION

To understand how to implement SETCLIENTID, make the following notations. Let:

- x be the value of the client.id subfield of the SETCLIENTID4args structure.
  - v be the value of the client.verifier subfield of the SETCLIENTID4args structure.
  - c be the value of the client ID field returned in the SETCLIENTID4resok structure.
  - k represent the value combination of the fields callback and callback\_ident fields of the SETCLIENTID4args structure.
  - s be the setclientid\_confirm value returned in the SETCLIENTID4resok structure.
- { v, x, c, k, s } be a quintuple for a client record. A client record is confirmed if there has been a SETCLIENTID\_CONFIRM operation to confirm it. Otherwise it is unconfirmed. An unconfirmed record is established by a SETCLIENTID call.

Since SETCLIENTID is a non-idempotent operation, let us assume that the server is implementing the duplicate request cache (DRC).

When the server gets a SETCLIENTID { v, x, k } request, it processes it in the following manner.

- o It first looks up the request in the DRC. If there is a hit, it returns the result cached in the DRC. The server does NOT remove client state (locks, shares, delegations) nor does it modify any recorded callback and callback\_ident information for client { x }.

For any DRC miss, the server takes the client id string x, and searches for client records for x that the server may have recorded from previous SETCLIENTID calls. For any confirmed record with the same id string x, if the recorded principal does not match that of SETCLIENTID call, then the server returns a NFS4ERR\_CLID\_INUSE error.

For brevity of discussion, the remaining description of the processing assumes that there was a DRC miss, and that where the server has previously recorded a confirmed record for client x, the aforementioned principal check has successfully passed.

- o The server checks if it has recorded a confirmed record for { v, x, c, l, s }, where l may or may not equal k. If so, and since the id verifier v of the request matches that which is confirmed and recorded, the server treats this as a probable callback information update and records an unconfirmed { v, x, c, k, t }

and leaves the confirmed { v, x, c, l, s } in place, such that t != s. It does not matter if k equals l or not. Any pre-existing unconfirmed { v, x, c, \*, \* } is removed.

The server returns { c, t }. It is indeed returning the old clientid4 value c, because the client apparently only wants to update callback value k to value l. It's possible this request is one from the Byzantine router that has stale callback information, but this is not a problem. The callback information update is only confirmed if followed up by a SETCLIENTID\_CONFIRM { c, t }.

The server awaits confirmation of k via SETCLIENTID\_CONFIRM { c, t }.

The server does NOT remove client (lock/share/delegation) state for x.

- o The server has previously recorded a confirmed { u, x, c, l, s } record such that v != u, l may or may not equal k, and has not recorded any unconfirmed { \*, x, \*, \*, \* } record for x. The server records an unconfirmed { v, x, d, k, t } (d != c, t != s).

The server returns { d, t }.

The server awaits confirmation of { d, k } via SETCLIENTID\_CONFIRM { d, t }.

The server does NOT remove client (lock/share/delegation) state for x.

- o The server has previously recorded a confirmed { u, x, c, l, s } record such that v != u, l may or may not equal k, and recorded an unconfirmed { w, x, d, m, t } record such that c != d, t != s, m may or may not equal k, m may or may not equal l, and k may or may not equal l. Whether w == v or w != v makes no difference. The server simply removes the unconfirmed { w, x, d, m, t } record and replaces it with an unconfirmed { v, x, e, k, r } record, such that e != d, e != c, r != t, r != s.

The server returns { e, r }.

The server awaits confirmation of { e, k } via SETCLIENTID\_CONFIRM { e, r }.

The server does NOT remove client (lock/share/delegation) state for x.

- o The server has no confirmed { \*, x, \*, \*, \* } for x. It may or may not have recorded an unconfirmed { u, x, c, l, s }, where l may or may not equal k, and u may or may not equal v. Any unconfirmed record { u, x, c, l, \* }, regardless whether u == v or l == k, is replaced with an unconfirmed record { v, x, d, k, t } where d != c, t != s.

The server returns { d, t }.

The server awaits confirmation of { d, k } via SETCLIENTID\_CONFIRM { d, t }. The server does NOT remove client (lock/share/delegation) state for x.

The server generates the clientid and setclientid\_confirm values and must take care to ensure that these values are extremely unlikely to ever be regenerated.

#### 15.36. Operation 36: SETCLIENTID\_CONFIRM - Confirm Client ID

##### 15.36.1. SYNOPSIS

clientid, setclientid\_confirm -> -

##### 15.36.2. ARGUMENT

```
struct SETCLIENTID_CONFIRM4args {
    clientid4      clientid;
    verifier4      setclientid_confirm;
};
```

##### 15.36.3. RESULT

```
struct SETCLIENTID_CONFIRM4res {
    nfsstat4      status;
};
```

##### 15.36.4. DESCRIPTION

This operation is used by the client to confirm the results from a previous call to SETCLIENTID. The client provides the server supplied (from a SETCLIENTID response) client ID. The server responds with a simple status of success or failure.

## 15.36.5. IMPLEMENTATION

The client must use the SETCLIENTID\_CONFIRM operation to confirm the following two distinct cases:

- o The client's use of a new shorthand client identifier (as returned from the server in the response to SETCLIENTID), a new callback value (as specified in the arguments to SETCLIENTID) and a new callback\_ident (as specified in the arguments to SETCLIENTID) value. The client's use of SETCLIENTID\_CONFIRM in this case also confirms the removal of any of the client's previous relevant leased state. Relevant leased client state includes byte-range locks, share reservations, and where the server does not support the CLAIM\_DELEGATE\_PREV claim type, delegations. If the server supports CLAIM\_DELEGATE\_PREV, then SETCLIENTID\_CONFIRM MUST NOT remove delegations for this client; relevant leased client state would then just include byte-range locks and share reservations.
- o The client's re-use of an old, previously confirmed, shorthand client identifier, a new callback value, and a new callback\_ident value. The client's use of SETCLIENTID\_CONFIRM in this case MUST NOT result in the removal of any previous leased state (locks, share reservations, and delegations)

We use the same notation and definitions for v, x, c, k, s, and unconfirmed and confirmed client records as introduced in the description of the SETCLIENTID operation. The arguments to SETCLIENTID\_CONFIRM are indicated by the notation { c, s }, where c is a value of type clientid4, and s is a value of type verifier4 corresponding to the setclientid\_confirm field.

As with SETCLIENTID, SETCLIENTID\_CONFIRM is a non-idempotent operation, and we assume that the server is implementing the duplicate request cache (DRC).

When the server gets a SETCLIENTID\_CONFIRM { c, s } request, it processes it in the following manner.

- o It first looks up the request in the DRC. If there is a hit, it returns the result cached in the DRC. The server does not remove any relevant leased client state nor does it modify any recorded callback and callback\_ident information for client { x } as represented by the shorthand value c.

For a DRC miss, the server checks for client records that match the shorthand value c. The processing cases are as follows:



- o The server has recorded an unconfirmed { v, x, c, k, s } record and a confirmed { v, x, c, l, t } record, such that s != t. If the principals of the records do not match that of the SETCLIENTID\_CONFIRM, the server returns NFS4ERR\_CLID\_INUSE, and no relevant leased client state is removed and no recorded callback and callback\_ident information for client { x } is changed. Otherwise, the confirmed { v, x, c, l, t } record is removed and the unconfirmed { v, x, c, k, s } is marked as confirmed, thereby modifying recorded and confirmed callback and callback\_ident information for client { x }.

The server does not remove any relevant leased client state.

The server returns NFS4\_OK.

- o The server has not recorded an unconfirmed { v, x, c, \*, \* } and has recorded a confirmed { v, x, c, \*, s }. If the principals of the record and of SETCLIENTID\_CONFIRM do not match, the server returns NFS4ERR\_CLID\_INUSE without removing any relevant leased client state and without changing recorded callback and callback\_ident values for client { x }.

If the principals match, then what has likely happened is that the client never got the response from the SETCLIENTID\_CONFIRM, and the DRC entry has been purged. Whatever the scenario, since the principals match, as well as { c, s } matching a confirmed record, the server leaves client x's relevant leased client state intact, leaves its callback and callback\_ident values unmodified, and returns NFS4\_OK.

- o The server has not recorded a confirmed { \*, \*, c, \*, \* }, and has recorded an unconfirmed { \*, x, c, k, s }. Even if this is a retry from client, nonetheless the client's first SETCLIENTID\_CONFIRM attempt was not received by the server. Retry or not, the server doesn't know, but it processes it as if were a first try. If the principal of the unconfirmed { \*, x, c, k, s } record mismatches that of the SETCLIENTID\_CONFIRM request the server returns NFS4ERR\_CLID\_INUSE without removing any relevant leased client state.

Otherwise, the server records a confirmed { \*, x, c, k, s }. If there is also a confirmed { \*, x, d, \*, t }, the server MUST remove the client x's relevant leased client state, and overwrite the callback state with k. The confirmed record { \*, x, d, \*, t } is removed.

Server returns NFS4\_OK.

- o The server has no record of a confirmed or unconfirmed { \*, \*, c, \*, s }. The server returns NFS4ERR\_STALE\_CLIENTID. The server does not remove any relevant leased client state, nor does it modify any recorded callback and callback\_ident information for any client.

The server needs to cache unconfirmed { v, x, c, k, s } client records and await for some time their confirmation. As should be clear from the record processing discussions for SETCLIENTID and SETCLIENTID\_CONFIRM, there are cases where the server does not deterministically remove unconfirmed client records. To avoid running out of resources, the server is not required to hold unconfirmed records indefinitely. One strategy the server might use is to set a limit on how many unconfirmed client records it will maintain, and then when the limit would be exceeded, remove the oldest record. Another strategy might be to remove an unconfirmed record when some amount of time has elapsed. The choice of the amount of time is fairly arbitrary but it is surely no higher than the server's lease time period. Consider that leases need to be renewed before the lease time expires via an operation from the client. If the client cannot issue a SETCLIENTID\_CONFIRM after a SETCLIENTID before a period of time equal to that of a lease expires, then the client is unlikely to be able maintain state on the server during steady state operation.

If the client does send a SETCLIENTID\_CONFIRM for an unconfirmed record that the server has already deleted, the client will get NFS4ERR\_STALE\_CLIENTID back. If so, the client should then start over, and send SETCLIENTID to reestablish an unconfirmed client record and get back an unconfirmed client ID and setclientid\_confirm verifier. The client should then send the SETCLIENTID\_CONFIRM to confirm the client ID.

SETCLIENTID\_CONFIRM does not establish or renew a lease. However, if SETCLIENTID\_CONFIRM removes relevant leased client state, and that state does not include existing delegations, the server MUST allow the client a period of time no less than the value of lease\_time attribute, to reclaim, (via the CLAIM\_DELEGATE\_PREV claim type of the OPEN operation) its delegations before removing unreclaimed delegations.

#### 15.37. Operation 37: VERIFY - Verify Same Attributes

##### 15.37.1. SYNOPSIS

(cfh), fattr -> -

## 15.37.2. ARGUMENT

```
struct VERIFY4args {  
    /* CURRENT_FH: object */  
    fattr4          obj_attributes;  
};
```

## 15.37.3. RESULT

```
struct VERIFY4res {  
    nfsstat4        status;  
};
```

## 15.37.4. DESCRIPTION

The VERIFY operation is used to verify that attributes have a value assumed by the client before proceeding with following operations in the compound request. If any of the attributes do not match then the error NFS4ERR\_NOT\_SAME must be returned. The current filehandle retains its value after successful completion of the operation.

## 15.37.5. IMPLEMENTATION

One possible use of the VERIFY operation is the following compound sequence. With this the client is attempting to verify that the file being removed will match what the client expects to be removed. This sequence can help prevent the unintended deletion of a file.

```
PUTFH (directory filehandle)  
LOOKUP (file name)  
VERIFY (filehandle == fh)  
PUTFH (directory filehandle)  
REMOVE (file name)
```

This sequence does not prevent a second client from removing and creating a new file in the middle of this sequence but it does help avoid the unintended result.

In the case that a recommended attribute is specified in the VERIFY operation and the server does not support that attribute for the file system object, the error NFS4ERR\_ATTRNOTSUPP is returned to the client.

When the attribute rdattr\_error or any write-only attribute (e.g., time\_modify\_set) is specified, the error NFS4ERR\_INVALID is returned to the client.

## 15.38. Operation 38: WRITE - Write to File

## 15.38.1. SYNOPSIS

(cfh), stateid, offset, stable, data -> count, committed, writeverf

## 15.38.2. ARGUMENT

```
enum stable_how4 {
    UNSTABLE4      = 0,
    DATA_SYNC4    = 1,
    FILE_SYNC4     = 2
};

struct WRITE4args {
    /* CURRENT_FH: file */
    stateid4      stateid;
    offset4       offset;
    stable_how4   stable;
    opaque        data<>;
};
```

## 15.38.3. RESULT

```
struct WRITE4resok {
    count4      count;
    stable_how4 committed;
    verifier4   writeverf;
};

union WRITE4res switch (nfsstat4 status) {
    case NFS4_OK:
        WRITE4resok    resok4;
    default:
        void;
};
```

## 15.38.4. DESCRIPTION

The WRITE operation is used to write data to a regular file. The target file is specified by the current filehandle. The offset specifies the offset where the data should be written. An offset of 0 (zero) specifies that the write should start at the beginning of the file. The count, as encoded as part of the opaque data parameter, represents the number of bytes of data that are to be written. If the count is 0 (zero), the WRITE will succeed and return

a count of 0 (zero) subject to permissions checking. The server may choose to write fewer bytes than requested by the client.

Part of the write request is a specification of how the write is to be performed. The client specifies with the `stable` parameter the method of how the data is to be processed by the server. If `stable` is `FILE_SYNC4`, the server must commit the data written plus all file system metadata to stable storage before returning results. This corresponds to the NFS version 2 protocol semantics. Any other behavior constitutes a protocol violation. If `stable` is `DATA_SYNC4`, then the server must commit all of the data to stable storage and enough of the metadata to retrieve the data before returning. The server implementor is free to implement `DATA_SYNC4` in the same fashion as `FILE_SYNC4`, but with a possible performance drop. If `stable` is `UNSTABLE4`, the server is free to commit any part of the data and the metadata to stable storage, including all or none, before returning a reply to the client. There is no guarantee whether or when any uncommitted data will subsequently be committed to stable storage. The only guarantees made by the server are that it will not destroy any data without changing the value of `verf` and that it will not commit the data and metadata at a level less than that requested by the client.

The `stateid` value for a `WRITE` request represents a value returned from a previous byte-range lock or share reservation request or the `stateid` associated with a delegation. The `stateid` is used by the server to verify that the associated share reservation and any byte-range locks are still valid and to update lease timeouts for the client.

Upon successful completion, the following results are returned. The `count` result is the number of bytes of data written to the file. The server may write fewer bytes than requested. If so, the actual number of bytes written starting at `location`, `offset`, is returned.

The server also returns an indication of the level of commitment of the data and metadata via `committed`. If the server committed all data and metadata to stable storage, `committed` should be set to `FILE_SYNC4`. If the level of commitment was at least as strong as `DATA_SYNC4`, then `committed` should be set to `DATA_SYNC4`. Otherwise, `committed` must be returned as `UNSTABLE4`. If `stable` was `FILE4_SYNC`, then `committed` must also be `FILE_SYNC4`: anything else constitutes a protocol violation. If `stable` was `DATA_SYNC4`, then `committed` may be `FILE_SYNC4` or `DATA_SYNC4`: anything else constitutes a protocol violation. If `stable` was `UNSTABLE4`, then `committed` may be either `FILE_SYNC4`, `DATA_SYNC4`, or `UNSTABLE4`.

The final portion of the result is the write verifier. The write

verifier is a cookie that the client can use to determine whether the server has changed instance (boot) state between a call to WRITE and a subsequent call to either WRITE or COMMIT. This cookie must be consistent during a single instance of the NFSv4 protocol service and must be unique between instances of the NFSv4 protocol server, where uncommitted data may be lost.

If a client writes data to the server with the stable argument set to UNSTABLE4 and the reply yields a committed response of DATA\_SYNC4 or UNSTABLE4, the client will follow up some time in the future with a COMMIT operation to synchronize outstanding asynchronous data and metadata with the server's stable storage, barring client error. It is possible that due to client crash or other error that a subsequent COMMIT will not be received by the server.

For a WRITE with a stateid value of all bits 0, the server MAY allow the WRITE to be serviced subject to mandatory file locks or the current share deny modes for the file. For a WRITE with a stateid value of all bits 1, the server MUST NOT allow the WRITE operation to bypass locking checks at the server and are treated exactly the same as if a stateid of all bits 0 were used.

On success, the current filehandle retains its value.

#### 15.38.5. IMPLEMENTATION

It is possible for the server to write fewer bytes of data than requested by the client. In this case, the server should not return an error unless no data was written at all. If the server writes less than the number of bytes specified, the client should issue another WRITE to write the remaining data.

It is assumed that the act of writing data to a file will cause the time\_modified of the file to be updated. However, the time\_modified of the file should not be changed unless the contents of the file are changed. Thus, a WRITE request with count set to 0 should not cause the time\_modified of the file to be updated.

The definition of stable storage has been historically a point of contention. The following expected properties of stable storage may help in resolving design issues in the implementation. Stable storage is persistent storage that survives:

1. Repeated power failures.
2. Hardware failures (of any board, power supply, etc.).

### 3. Repeated software crashes, including reboot cycle.

This definition does not address failure of the stable storage module itself.

The verifier is defined to allow a client to detect different instances of an NFSv4 protocol server over which cached, uncommitted data may be lost. In the most likely case, the verifier allows the client to detect server reboots. This information is required so that the client can safely determine whether the server could have lost cached data. If the server fails unexpectedly and the client has uncommitted data from previous WRITE requests (done with the stable argument set to UNSTABLE4 and in which the result committed was returned as UNSTABLE4 as well) it may not have flushed cached data to stable storage. The burden of recovery is on the client and the client will need to retransmit the data to the server.

A suggested verifier would be to use the time that the server was booted or the time the server was last started (if restarting the server without a reboot results in lost buffers).

The committed field in the results allows the client to do more effective caching. If the server is committing all WRITE requests to stable storage, then it should return with committed set to FILE\_SYNC4, regardless of the value of the stable field in the arguments. A server that uses an NVRAM accelerator may choose to implement this policy. The client can use this to increase the effectiveness of the cache by discarding cached data that has already been committed on the server.

Some implementations may return NFS4ERR\_NOSPC instead of NFS4ERR\_DQUOT when a user's quota is exceeded. In the case that the current filehandle is a directory, the server will return NFS4ERR\_ISDIR. If the current filehandle is not a regular file or a directory, the server will return NFS4ERR\_INVALID.

If mandatory file locking is on for the file, and corresponding record of the data to be written file is read or write locked by an owner that is not associated with the stateid, the server will return NFS4ERR\_LOCKED. If so, the client must check if the owner corresponding to the stateid used with the WRITE operation has a conflicting read lock that overlaps with the region that was to be written. If the stateid's owner has no conflicting read lock, then the client should try to get the appropriate write byte-range lock via the LOCK operation before re-attempting the WRITE. When the WRITE completes, the client should release the byte-range lock via LOCKU.

If the stateid's owner had a conflicting read lock, then the client has no choice but to return an error to the application that attempted the WRITE. The reason is that since the stateid's owner had a read lock, the server either attempted to temporarily effectively upgrade this read lock to a write lock, or the server has no upgrade capability. If the server attempted to upgrade the read lock and failed, it is pointless for the client to re-attempt the upgrade via the LOCK operation, because there might be another client also trying to upgrade. If two clients are blocked trying upgrade the same lock, the clients deadlock. If the server has no upgrade capability, then it is pointless to try a LOCK operation to upgrade.

### 15.39. Operation 39: RELEASE\_LOCKOWNER - Release Lockowner State

#### 15.39.1. SYNOPSIS

```
lock-owner -> ()
```

#### 15.39.2. ARGUMENT

```
struct RELEASE_LOCKOWNER4args {  
    lock_owner4    lock_owner;  
};
```

#### 15.39.3. RESULT

```
struct RELEASE_LOCKOWNER4res {  
    nfsstat4    status;  
};
```

#### 15.39.4. DESCRIPTION

This operation is used to notify the server that the lock\_owner is no longer in use by the client and that future client requests will not reference this lock\_owner. This allows the server to release cached state related to the specified lock\_owner. If file locks, associated with the lock\_owner, are held at the server, the error NFS4ERR\_LOCKS\_HELD will be returned and no further action will be taken.

#### 15.39.5. IMPLEMENTATION

The client may choose to use this operation to ease the amount of server state that is held. Information that can be released when a RELEASE\_LOCKOWNER is done includes the specified lock-owner string, the seqid associated with the lock-owner, any saved reply for the



lock-owner, and any lock stateids associated with that lock-owner.

Depending on the behavior of applications at the client, it may be important for the client to use this operation since the server has certain obligations with respect to holding a reference to lock-owner-associated state as long as an associated file is open. Therefore, if the client knows for certain that the lock\_owner will no longer be used, either to reference existing lock stateids associated with the lock-owner to create new ones, it should use `RELEASE_LOCKOWNER`.

#### 15.40. Operation 10044: ILLEGAL - Illegal operation

##### 15.40.1. SYNOPSIS

```
<null> -> ()
```

##### 15.40.2. ARGUMENT

```
void;
```

##### 15.40.3. RESULT

```
struct ILLEGAL4res {  
    nfsstat4      status;  
};
```

##### 15.40.4. DESCRIPTION

This operation is a place holder for encoding a result to handle the case of the client sending an operation code within COMPOUND that is not supported. See Section 15.2.4 for more details.

The status field of ILLEGAL4res MUST be set to NFS4ERR\_OP\_ILLEGAL.

##### 15.40.5. IMPLEMENTATION

A client will probably not send an operation with code `OP_ILLEGAL` but if it does, the response will be ILLEGAL4res just as it would be with any other invalid operation code. Note that if the server gets an illegal operation code that is not `OP_ILLEGAL`, and if the server checks for legal operation codes during the XDR decode phase, then the ILLEGAL4res would not be returned.

## 16. NFSv4 Callback Procedures

The procedures used for callbacks are defined in the following sections. In the interest of clarity, the terms "client" and "server" refer to NFS clients and servers, despite the fact that for an individual callback RPC, the sense of these terms would be precisely the opposite.

### 16.1. Procedure 0: CB\_NULL - No Operation

#### 16.1.1. SYNOPSIS

<null>

#### 16.1.2. ARGUMENT

void;

#### 16.1.3. RESULT

void;

#### 16.1.4. DESCRIPTION

Standard NULL procedure. Void argument, void response. Even though there is no direct functionality associated with this procedure, the server will use CB\_NULL to confirm the existence of a path for RPCs from server to client.

### 16.2. Procedure 1: CB\_COMPOUND - Compound Operations

#### 16.2.1. SYNOPSIS

compoundargs -> compoundres

#### 16.2.2. ARGUMENT

```
enum nfs_cb_opnum4 {
    OP_CB_GETATTR      = 3,
    OP_CB_RECALL       = 4,
    OP_CB_ILLEGAL       = 10044
};
```

```

union nfs_cb_argop4 switch (unsigned argop) {
    case OP_CB_GETATTR:
        CB_GETATTR4args          opcbgetattr;
    case OP_CB_RECALL:
        CB_RECALL4args           opcbrecall;
    case OP_CB_ILLEGAL:
        void;
};

```

```

struct CB_COMPOUND4args {
    utf8str_cs      tag;
    uint32_t        minorversion;
    uint32_t        callback_ident;
    nfs_cb_argop4   argarray<>;
};

```

#### 16.2.3. RESULT

```

union nfs_cb_resop4 switch (unsigned resop) {
    case OP_CB_GETATTR:      CB_GETATTR4res  opcbgetattr;
    case OP_CB_RECALL:      CB_RECALL4res   opcbrecall;
    case OP_CB_ILLEGAL:      CB_ILLEGAL4res  opcbillegal;
};

```

```

struct CB_COMPOUND4res {
    nfsstat4        status;
    utf8str_cs      tag;
    nfs_cb_resop4   resarray<>;
};

```

#### 16.2.4. DESCRIPTION

The CB\_COMPOUND procedure is used to combine one or more of the callback procedures into a single RPC request. The main callback RPC program has two main procedures: CB\_NULL and CB\_COMPOUND. All other operations use the CB\_COMPOUND procedure as a wrapper.

In the processing of the CB\_COMPOUND procedure, the client may find that it does not have the available resources to execute any or all of the operations within the CB\_COMPOUND sequence. In this case, the error NFS4ERR\_RESOURCE will be returned for the particular operation within the CB\_COMPOUND procedure where the resource exhaustion occurred. This assumes that all previous operations within the CB\_COMPOUND sequence have been evaluated successfully.

Contained within the CB\_COMPOUND results is a 'status' field. This status must be equivalent to the status of the last operation that

was executed within the CB\_COMPOUND procedure. Therefore, if an operation incurred an error then the 'status' value will be the same error value as is being returned for the operation that failed.

For the definition of the "tag" field, see Section 15.2.

The value of `callback_ident` is supplied by the client during SETCLIENTID. The server must use the client supplied `callback_ident` during the CB\_COMPOUND to allow the client to properly identify the server.

Illegal operation codes are handled in the same way as they are handled for the COMPOUND procedure.

#### 16.2.5. IMPLEMENTATION

The CB\_COMPOUND procedure is used to combine individual operations into a single RPC request. The client interprets each of the operations in turn. If an operation is executed by the client and the status of that operation is NFS4\_OK, then the next operation in the CB\_COMPOUND procedure is executed. The client continues this process until there are no more operations to be executed or one of the operations has a status value other than NFS4\_OK.

#### 16.2.6. Operation 3: CB\_GETATTR - Get Attributes

##### 16.2.6.1. SYNOPSIS

```
fh, attr_request -> attrmask, attr_vals
```

##### 16.2.6.2. ARGUMENT

```
struct CB_GETATTR4args {  
    nfs_fh4 fh;  
    bitmap4 attr_request;  
};
```

## 16.2.6.3. RESULT

```
struct CB_GETATTR4resok {
    fattr4  obj_attributes;
};

union CB_GETATTR4res switch (nfsstat4 status) {
    case NFS4_OK:
        CB_GETATTR4resok      resok4;
    default:
        void;
};
```

## 16.2.6.4. DESCRIPTION

The CB\_GETATTR operation is used by the server to obtain the current modified state of a file that has been OPEN\_DELEGATE\_WRITE delegated. The attributes size and change are the only ones guaranteed to be serviced by the client. See Section 10.4.3 for a full description of how the client and server are to interact with the use of CB\_GETATTR.

If the filehandle specified is not one for which the client holds a OPEN\_DELEGATE\_WRITE delegation, an NFS4ERR\_BADHANDLE error is returned.

## 16.2.6.5. IMPLEMENTATION

The client returns attrmask bits and the associated attribute values only for the change attribute, and attributes that it may change (time\_modify, and size).

## 16.2.7. Operation 4: CB\_RECALL - Recall an Open Delegation

## 16.2.7.1. SYNOPSIS

```
stateid, truncate, fh -> ()
```

## 16.2.7.2. ARGUMENT

```
struct CB_RECALL4args {
    stateid4      stateid;
    bool          truncate;
    nfs_fh4       fh;
};
```

## 16.2.7.3. RESULT

```
struct CB_RECALL4res {  
    nfsstat4      status;  
};
```

## 16.2.7.4. DESCRIPTION

The CB\_RECALL operation is used to begin the process of recalling an open delegation and returning it to the server.

The truncate flag is used to optimize recall for a file which is about to be truncated to zero. When it is set, the client is freed of obligation to propagate modified data for the file to the server, since this data is irrelevant.

If the handle specified is not one for which the client holds an open delegation, an NFS4ERR\_BADHANDLE error is returned.

If the stateid specified is not one corresponding to an open delegation for the file specified by the filehandle, an NFS4ERR\_BAD\_STATEID is returned.

## 16.2.7.5. IMPLEMENTATION

The client should reply to the callback immediately. Replying does not complete the recall except when an error was returned. The recall is not complete until the delegation is returned using a DELEGRETURN.

## 16.2.8. Operation 10044: CB\_ILLEGAL - Illegal Callback Operation

## 16.2.8.1. SYNOPSIS

```
<null> -> ()
```

## 16.2.8.2. ARGUMENT

```
void;
```

## 16.2.8.3. RESULT

```
/*
 * CB_ILLEGAL: Response for illegal operation numbers
 */
struct CB_ILLEGAL4res {
    nfsstat4      status;
};
```

## 16.2.8.4. DESCRIPTION

This operation is a place-holder for encoding a result to handle the case of the client sending an operation code within COMPOUND that is not supported. See Section 15.2.4 for more details.

The status field of CB\_ILLEGAL4res MUST be set to NFS4ERR\_OP\_ILLEGAL.

## 16.2.8.5. IMPLEMENTATION

A server will probably not send an operation with code OP\_CB\_ILLEGAL but if it does, the response will be CB\_ILLEGAL4res just as it would be with any other invalid operation code. Note that if the client gets an illegal operation code that is not OP\_ILLEGAL, and if the client checks for legal operation codes during the XDR decode phase, then the CB\_ILLEGAL4res would not be returned.

## 17. Security Considerations

NFS has historically used a model where, from an authentication perspective, the client was the entire machine, or at least the source IP address of the machine. The NFS server relied on the NFS client to make the proper authentication of the end-user. The NFS server in turn shared its files only to specific clients, as identified by the client's source IP address. Given this model, the AUTH\_SYS RPC security flavor simply identified the end-user using the client to the NFS server. When processing NFS responses, the client ensured that the responses came from the same IP address and port number that the request was sent to. While such a model is easy to implement and simple to deploy and use, it is certainly not a safe model. Thus, NFSv4 mandates that implementations support a security model that uses end to end authentication, where an end-user on a client mutually authenticates (via cryptographic schemes that do not expose passwords or keys in the clear on the network) to a principal on an NFS server. Consideration should also be given to the integrity and privacy of NFS requests and responses. The issues of end to end mutual authentication, integrity, and privacy are

discussed as part of Section 3.

When an NFSv4 mandated security model is used and a security principal or an NFSv4 name in `user@dns_domain` form needs to be translated to or from a local representation as described in Section 5.9, the translation SHOULD be done in a secure manner that preserves the integrity of the translation. For communication with a name service such as LDAP ([RFC4511]), this means employing a security service that uses authentication and data integrity. Kerberos and Transport Layer Security (TLS) ([RFC5246]) are examples of such a security service.

Note that being REQUIRED to implement does not mean REQUIRED to use; AUTH\_SYS can be used by NFSv4 clients and servers. However, AUTH\_SYS is merely an OPTIONAL security flavor in NFSv4, and so interoperability via AUTH\_SYS is not assured.

For reasons of reduced administration overhead, better performance and/or reduction of CPU utilization, users of NFSv4 implementations may choose to not use security mechanisms that enable integrity protection on each remote procedure call and response. The use of mechanisms without integrity leaves the customer vulnerable to an attacker in between the NFS client and server that modifies the RPC request and/or the response. While implementations are free to provide the option to use weaker security mechanisms, there are two operations in particular that warrant the implementation overriding user choices.

The first such operation is SECINFO. It is recommended that the client issue the SECINFO call such that it is protected with a security flavor that has integrity protection, such as RPCSEC\_GSS with a security triple that uses either `rpc_gss_svc_integrity` or `rpc_gss_svc_privacy` (`rpc_gss_svc_privacy` includes integrity protection) service. Without integrity protection encapsulating SECINFO and therefore its results, an attacker in the middle could modify results such that the client might select a weaker algorithm in the set allowed by server, making the client and/or server vulnerable to further attacks.

The second operation that SHOULD use integrity protection is any GETATTR for the `fs_locations` attribute. The attack has two steps. First the attacker modifies the unprotected results of some operation to return NFS4ERR\_MOVED. Second, when the client follows up with a GETATTR for the `fs_locations` attribute, the attacker modifies the results to cause the client migrate its traffic to a server controlled by the attacker.

Because the operations SETCLIENTID/SETCLIENTID\_CONFIRM are



responsible for the release of client state, it is imperative that the principal used for these operations is checked against and match the previous use of these operations. See Section 9.1.1 for further discussion.

## 18. IANA Considerations

This section uses terms that are defined in [RFC5226].

### 18.1. Named Attribute Definitions

IANA has created a registry called the "NFSv4 Named Attribute Definitions Registry" for [RFC3530] and [RFC5661]. This section introduces no new changes, but it does recap the intent.

The NFSv4 protocol supports the association of a file with zero or more named attributes. The name space identifiers for these attributes are defined as string names. The protocol does not define the specific assignment of the name space for these file attributes. The IANA registry promotes interoperability where common interests exist. While application developers are allowed to define and use attributes as needed, they are encouraged to register the attributes with IANA.

Such registered named attributes are presumed to apply to all minor versions of NFSv4, including those defined subsequently to the registration. Where the named attribute is intended to be limited with regard to the minor versions for which they are not be used, the assignment in registry will clearly state the applicable limits.

All assignments to the registry are made on a First Come First Served basis, per section 4.1 of [RFC5226]. The policy for each assignment is Specification Required, per section 4.1 of [RFC5226].

Under the NFSv4 specification, the name of a named attribute can in theory be up to  $2^{32} - 1$  bytes in length, but in practice NFSv4 clients and servers will be unable to handle a string that long. IANA should reject any assignment request with a named attribute that exceeds 128 UTF-8 characters. To give IESG the flexibility to set up bases of assignment of Experimental Use and Standards Action, the prefixes of "EXPE" and "STDS" are Reserved. The zero length named attribute name is Reserved.

The prefix "PRIV" is allocated for Private Use. A site that wants to make use of unregistered named attributes without risk of conflicting with an assignment in IANA's registry should use the prefix "PRIV" in all of its named attributes.

Because some NFSv4 clients and servers have case insensitive semantics, the fifteen additional lower case and mixed case permutations of each of "EXPE", "PRIV", and "STDS", are Reserved (e.g. "expe", "expE", "exPe", etc. are Reserved). Similarly, IANA must not allow two assignments that would conflict if both named attributes were converted to a common case.

The registry of named attributes is a list of assignments, each containing three fields for each assignment.

1. A US-ASCII string name that is the actual name of the attribute. This name must be unique. This string name can be 1 to 128 UTF-8 characters long.
2. A reference to the specification of the named attribute. The reference can consume up to 256 bytes (or more if IANA permits).
3. The point of contact of the registrant. The point of contact can consume up to 256 bytes (or more if IANA permits).

#### 18.1.1. Initial Registry

There is no initial registry.

#### 18.1.2. Updating Registrations

The registrant is always permitted to update the point of contact field. To make any other change will require Expert Review or IESG Approval.

### 19. References

#### 19.1. Normative References

- [I-D.ietf-nfsv4-rfc3530bis-dot-x]  
Haynes, T. and D. Noveck, "NFSv4 Version 0 XDR Description", draft-ietf-nfsv4-rfc3530bis-dot-x-18 (work in progress), Aug 2013.
- [ISO.10646-1.1993]  
International Organization for Standardization,  
"Information Technology - Universal Multiple-octet coded  
Character Set (UCS) - Part 1: Architecture and Basic  
Multilingual Plane", ISO Standard 10646-1, May 1993.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", March 1997.

- [RFC2203] Eisler, M., Chiu, A., and L. Ling, "RPCSEC\_GSS Protocol Specification", RFC 2203, September 1997.
- [RFC2277] Alvestrand, H., "IETF Policy on Character Sets and Languages", BCP 18, RFC 2277, January 1998.
- [RFC2743] Linn, J., "Generic Security Service Application Program Interface Version 2, Update 1", RFC 2743, January 2000.
- [RFC5403] Eisler, M., "RPCSEC\_GSS Version 2", RFC 5403, February 2009.
- [RFC5531] Thurlow, R., "RPC: Remote Procedure Call Protocol Specification Version 2", RFC 5531, May 2009.
- [RFC5665] Eisler, M., Ed., "IANA Considerations for Remote Procedure Call (RPC) Network Identifiers and Universal Address Formats", RFC 5665, January 2010.
- [RFC5890] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Definitions and Document Framework", RFC 5890, August 2010.
- [RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, August 2010.
- [RFC6649] Astrand, L. and T. Yu, "Deprecate DES, RC4-HMAC-EXP, and Other Weak Cryptographic Algorithms in Kerberos", RFC 6649, July 2012.
- [Unicode1] The Unicode Consortium, "The Unicode Standard, Version 3.0, ISBN 0-201-61633-5".
- [openg\_symlink] The Open Group, "Section 3.372 of Chapter 3 of Base Definitions of The Open Group Base Specifications Issue 6, IEEE Std 1003.1, 2004 Edition, HTML Version ([www.opengroup.org](http://www.opengroup.org)), ISBN 1931624232", 2004.

## 19.2. Informative References

- [Chet] Juszczak, C., "Improving the Performance and Correctness of an NFS Server", USENIX Conference Proceedings , June 1990.
- [Floyd] Floyd, S. and V. Jacobson, "The Synchronization of Periodic Routing Messages", IEEE/ACM Transactions on

Networking 2(2), pp. 122-136, April 1994.

[ISEG\_errata]

IESG, "IESG Processing of RFC Errata for the IETF Stream", July 2008.

[P1003.1e]

Institute of Electrical and Electronics Engineers, Inc., "IEEE Draft P1003.1e", 1997.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[RFC1094] Nowicki, B., "NFS: Network File System Protocol specification", RFC 1094, March 1989.

[RFC1345] Simonsen, K., "Character Mnemonics and Character Sets", RFC 1345, June 1992.

[RFC1813] Callaghan, B., Pawlowski, B., and P. Staubach, "NFS Version 3 Protocol Specification", RFC 1813, June 1995.

[RFC1833] Srinivasan, R., "Binding Protocols for ONC RPC Version 2", RFC 1833, August 1995.

[RFC2054] Callaghan, B., "WebNFS Client Specification", RFC 2054, October 1996.

[RFC2055] Callaghan, B., "WebNFS Server Specification", RFC 2055, October 1996.

[RFC2224] Callaghan, B., "NFS URL Scheme", RFC 2224, October 1997.

[RFC2623] Eisler, M., "NFS Version 2 and Version 3 Security Issues and the NFS Protocol's Use of RPCSEC\_GSS and Kerberos V5", RFC 2623, June 1999.

[RFC2624] Shepler, S., "NFS Version 4 Design Considerations", RFC 2624, June 1999.

[RFC2755] Chiu, A., Eisler, M., and B. Callaghan, "Security Negotiation for WebNFS", RFC 2755, January 2000.

[RFC3010] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3010, December 2000.

[RFC3232] Reynolds, J., "Assigned Numbers: RFC 1700 is Replaced by

an On-line Database", RFC 3232, January 2002.

- [RFC3454] Hoffman, P. and M. Blanchet, "Preparation of Internationalized Strings ("stringprep")", RFC 3454, December 2002.
- [RFC3530] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.
- [RFC4121] Zhu, L., Jaganathan, K., and S. Hartman, "The Kerberos Version 5 Generic Security Service Application Program Interface (GSS-API) Mechanism: Version 2", RFC 4121, July 2005.
- [RFC4178] Zhu, L., Leach, P., Jaganathan, K., and W. Ingersoll, "The Simple and Protected Generic Security Service Application Program Interface (GSS-API) Negotiation Mechanism", RFC 4178, October 2005.
- [RFC4340] Kohler, E., Handley, M., and S. Floyd, "Datagram Congestion Control Protocol (DCCP)", RFC 4340, March 2006.
- [RFC4506] Eisler, M., "XDR: External Data Representation Standard", RFC 4506, May 2006.
- [RFC4511] Sermersheim, J., "Lightweight Directory Access Protocol (LDAP): The Protocol", RFC 4511, June 2006.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.
- [RFC5661] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.
- [RFC5740] Adamson, B., Bormann, C., Handley, M., and J. Macker, "Negative-acknowledgment (NACK)-Oriented Reliable Multicast (NORM) Transport Protocol", RFC 5740, November 2009.
- [fcntl] The Open Group, "Section 'fcntl()' of System Interfaces of The Open Group Base Specifications Issue 6 IEEE Std 1003.1, 2004 Edition, HTML Version ([www.opengroup.org](http://www.opengroup.org))",

ISBN 1931624232", 2004.

[fsync] The Open Group, "Section 'fsync()' of System Interfaces of The Open Group Base Specifications Issue 6 IEEE Std 1003.1, 2004 Edition, HTML Version ([www.opengroup.org](http://www.opengroup.org)), ISBN 1931624232", 2004.

[getpwnam] The Open Group, "Section 'getpwnam()' of System Interfaces of The Open Group Base Specifications Issue 6 IEEE Std 1003.1, 2004 Edition, HTML Version ([www.opengroup.org](http://www.opengroup.org)), ISBN 1931624232", 2004.

[read\_api] The Open Group, "Section 'read()' of System Interfaces of The Open Group Base Specifications Issue 6, IEEE Std 1003.1, 2004 Edition", 2004.

[readdir\_api] The Open Group, "Section 'readdir()' of System Interfaces of The Open Group Base Specifications Issue 6, IEEE Std 1003.1, 2004 Edition", 2004.

[unlink] The Open Group, "Section 'unlink()' of System Interfaces of The Open Group Base Specifications Issue 6 IEEE Std 1003.1, 2004 Edition, HTML Version ([www.opengroup.org](http://www.opengroup.org)), ISBN 1931624232", 2004.

[write\_api] The Open Group, "Section 'write()' of System Interfaces of The Open Group Base Specifications Issue 6, IEEE Std 1003.1, 2004 Edition", 2004.

[xnfs] The Open Group, "Protocols for Interworking: XNFS, Version 3W, ISBN 1-85912-184-5", February 1998.

## Appendix A. Acknowledgments

A bis is certainly built on the shoulders of the first attempt. Spencer Shepler, Brent Callaghan, David Robinson, Robert Thurlow, Carl Beame, Mike Eisler, and David Noveck are responsible for a great deal of the effort in this work.

Rob Thurlow clarified how a client should contact a new server if a migration has occurred.

David Black, Nico Williams, Mike Eisler, Trond Myklebust, James

Lentini, and Mike Kupfer read many drafts of Section 12 and contributed numerous useful suggestions, without which the necessary revision of that section for this document would not have been possible.

Peter Staubach read almost all of the drafts of Section 12 leading to the published result and his numerous comments were always useful and contributed substantially to improving the quality of the final result.

James Lentini graciously read the rewrite of Section 8 and his comments were vital in improving the quality of that effort.

Rob Thurlow, Sorin Faibish, James Lentini, Bruce Fields, and Trond Myklebust were faithful attendants of the biweekly triage meeting and accepted many an action item.

Bruce Fields was a good sounding board for both the Third Edge Condition and Courtesy Locks in general. He was also the leading advocate of stamping out backport issues from [RFC5661].

Marcel Telka was a champion of straightening out the difference between a lock-owner and an open-owner. He has also been diligent in reviewing the final document.

Benjamin Kaduk reminded us that DES is dead and Nico Williams helped us close the lid on the coffin.

#### Appendix B. RFC Editor Notes

[RFC Editor: please remove this section prior to publishing this document as an RFC]

[RFC Editor: prior to publishing this document as an RFC, please replace all occurrences of RFCNFSv4XDR with RFCxxxx where xxxx is the RFC number assigned to the XDR document.]

[RFC Editor: Please note that there is also a reference entry that needs to be modified for the companion document.]

Authors' Addresses

Thomas Haynes (editor)  
NetApp  
495 E Java Dr  
Sunnyvale, CA 95054  
USA

Phone: +1 408 419 3018  
Email: thomas@netapp.com

David Noveck (editor)  
EMC Corporation  
228 South Street  
Hopkinton, MA 01748  
US

Phone: +1 508 249 5748  
Email: david.noveck@emc.com





NFSv4  
Internet-Draft  
Intended status: Standards Track  
Expires: April 22, 2014

T. Haynes, Ed.  
NetApp  
D. Noveck, Ed.  
EMC  
October 19, 2013

Network File System (NFS) Version 4  
External Data Representation Standard (XDR) Description  
draft-ietf-nfsv4-rfc3530bis-dot-x-19.txt

## Abstract

The Network File System (NFS) version 4 is a distributed filesystem protocol which owes its heritage to NFS protocol version 2, RFC 1094, and version 3, RFC 1813. Unlike earlier versions, the NFS version 4 protocol supports traditional file access, while integrating support for file locking and the mount protocol. In addition, support for strong security (and its negotiation), compound operations, client caching, and internationalization have been added. Of course, attention has been applied to making NFS version 4 operate well in an Internet environment.

RFC3530bis formally obsoleting RFC 3530. This document, together with RFC3530bis replaces RFC 3530 as the definition of the NFS version 4 protocol.

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 22, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

## Table of Contents

1. XDR Description of NFSv4.0 . . . . .	4
2. Security Considerations . . . . .	37
3. IANA Considerations . . . . .	37
4. Normative References . . . . .	37
Appendix A. Acknowledgments . . . . .	37
Appendix B. RFC Editor Notes . . . . .	37
Authors' Addresses . . . . .	38

## 1. XDR Description of NFSv4.0

This document contains the XDR ([RFC4506]) description of NFSv4.0 protocol ([I-D.ietf-nfsv4-rfc3530bis]).

The XDR description is provided in this document in a way that makes it simple for the reader to extract it into ready to compile form. The reader can feed this document in the following shell script to produce the machine readable XDR description of NFSv4.0:

```
#!/bin/sh
grep "^ *///" | sed 's?^ */// ??' | sed 's?^ *///$??'
```

I.e. if the above script is stored in a file called "extract.sh", and this document is in a file called "spec.txt", then the reader can do:

```
sh extract.sh < spec.txt > nfs4_prot.x
```

The effect of the script is to remove leading white space from each line, plus a sentinel sequence of "///".

The XDR description, with the sentinel sequence follows:

```
/// /*
///  * This file was machine generated for
///  * draft-ietf-nfsv4-rfc3530bis-28
///  * Last updated Sat Oct 19 11:28:52 PDT 2013
/// */
/// /*
///  * Copyright (C) The IETF Trust (2009-2011)
///  * All Rights Reserved.
///  *
///  * Copyright (C) The Internet Society (1998-2011).
///  * All Rights Reserved.
/// */
///
/// /*
///  *      nfs4_prot.x
///  *
/// */
/// /*
///  * Basic typedefs for RFC 1832 data type definitions
/// */
/// /*
///  * typedef int          int32_t;
///  * typedef unsigned int uint32_t;
```

```

/// * typedef hyper          int64_t;
/// * typedef unsigned hyper  uint64_t;
/// */
///
/// /*
/// * Sizes
/// */
/// const NFS4_FH_SIZE          = 128;
/// const NFS4_VERIFIER_SIZE    = 8;
/// const NFS4_OTHER_SIZE       = 12;
/// const NFS4_OPAQUE_LIMIT     = 1024;
///
/// const NFS4_INT64_MAX        = 0x7fffffffffffffff;
/// const NFS4_UINT64_MAX       = 0xffffffffffffffff;
/// const NFS4_INT32_MAX        = 0x7fffffff;
/// const NFS4_UINT32_MAX       = 0xffffffff;
///
/// /*
/// * File types
/// */
/// enum nfs_ftype4 {
///     NF4REG = 1,      /* Regular File */
///     NF4DIR = 2,      /* Directory */
///     NF4BLK = 3,      /* Special File - block device */
///     NF4CHR = 4,      /* Special File - character device */
///     NF4LNK = 5,      /* Symbolic Link */
///     NF4SOCK = 6,     /* Special File - socket */
///     NF4FIFO = 7,     /* Special File - fifo */
///     NF4ATTRDIR
///         = 8,         /* Attribute Directory */
///     NF4NAMEDATTR
///         = 9          /* Named Attribute */
/// };
///
/// /*
/// * Error status
/// */
/// enum nfsstat4 {
///     NFS4_OK          = 0,      /* everything is okay */
///     NFS4ERR_PERM      = 1,      /* caller not privileged */
///     NFS4ERR_NOENT     = 2,      /* no such file/directory */
///     NFS4ERR_IO        = 5,      /* hard I/O error */
///     NFS4ERR_NXIO      = 6,      /* no such device */
///     NFS4ERR_ACCESS    = 13,     /* access denied */
///     NFS4ERR_EXIST     = 17,     /* file already exists */
///     NFS4ERR_XDEV      = 18,     /* different filesystems */
///     /* Unused/reserved 19 */

```

```
/// NFS4ERR_NOTDIR          = 20, /* should be a directory */
/// NFS4ERR_ISDIR           = 21, /* should not be directory */
/// NFS4ERR_INVAL           = 22, /* invalid argument */
/// NFS4ERR_FBIG            = 27, /* file exceeds server max */
/// NFS4ERR_NOSPC           = 28, /* no space on filesystem */
/// NFS4ERR_ROFS            = 30, /* read-only filesystem */
/// NFS4ERR_MLINK           = 31, /* too many hard links */
/// NFS4ERR_NAMETOOLONG      = 63, /* name exceeds server max */
/// NFS4ERR_NOTEMPTY        = 66, /* directory not empty */
/// NFS4ERR_DQUOT           = 69, /* hard quota limit reached*/
/// NFS4ERR_STALE           = 70, /* file no longer exists */
/// NFS4ERR_BADHANDLE       = 10001,/* Illegal filehandle */
/// NFS4ERR_BAD_COOKIE      = 10003,/* READDIR cookie is stale */
/// NFS4ERR_NOTSUPP         = 10004,/* operation not supported */
/// NFS4ERR_TOOSMALL        = 10005,/* response limit exceeded */
/// NFS4ERR_SERVERFAULT     = 10006,/* undefined server error */
/// NFS4ERR_BADTYPE         = 10007,/* type invalid for CREATE */
/// NFS4ERR_DELAY           = 10008,/* file "busy" - retry */
/// NFS4ERR_SAME            = 10009,/* nverify says attrs same */
/// NFS4ERR_DENIED          = 10010,/* lock unavailable */
/// NFS4ERR_EXPIRED         = 10011,/* lock lease expired */
/// NFS4ERR_LOCKED          = 10012,/* I/O failed due to lock */
/// NFS4ERR_GRACE           = 10013,/* in grace period */
/// NFS4ERR_FHEXPIRED       = 10014,/* filehandle expired */
/// NFS4ERR_SHARE_DENIED    = 10015,/* share reserve denied */
/// NFS4ERR_WRONGSEC        = 10016,/* wrong security flavor */
/// NFS4ERR_CLID_INUSE      = 10017,/* clientid in use */
/// NFS4ERR_RESOURCE        = 10018,/* resource exhaustion */
/// NFS4ERR_MOVED           = 10019,/* filesystem relocated */
/// NFS4ERR_NOFILEHANDLE    = 10020,/* current FH is not set */
/// NFS4ERR_MINOR_VERS_MISMATCH = 10021,/* minor vers not supp */
/// NFS4ERR_STALE_CLIENTID  = 10022,/* server has rebooted */
/// NFS4ERR_STALE_STATEID   = 10023,/* server has rebooted */
/// NFS4ERR_OLD_STATEID     = 10024,/* state is out of sync */
/// NFS4ERR_BAD_STATEID     = 10025,/* incorrect stateid */
/// NFS4ERR_BAD_SEQID       = 10026,/* request is out of seq. */
/// NFS4ERR_NOT_SAME        = 10027,/* verify - attrs not same */
/// NFS4ERR_LOCK_RANGE      = 10028,/* lock range not supported*/
/// NFS4ERR_SYMLINK         = 10029,/* should be file/directory*/
/// NFS4ERR_RESTOREFH       = 10030,/* no saved filehandle */
/// NFS4ERR_LEASE_MOVED     = 10031,/* some filesystem moved */
/// NFS4ERR_ATTRNOTSUPP     = 10032,/* recommended attr not sup*/
/// NFS4ERR_NO_GRACE        = 10033,/* reclaim outside of grace*/
/// NFS4ERR_RECLAIM_BAD     = 10034,/* reclaim error at server */
/// NFS4ERR_RECLAIM_CONFLICT = 10035,/* conflict on reclaim */
/// NFS4ERR_BADXDR          = 10036,/* XDR decode failed */
/// NFS4ERR_LOCKS_HELD      = 10037,/* file locks held at CLOSE*/
/// NFS4ERR_OPENMODE        = 10038,/* conflict in OPEN and I/O*/
```

```

/// NFS4ERR_BADOWNER      = 10039,/* owner translation bad */
/// NFS4ERR_BADCHAR       = 10040,/* utf-8 char not supported*/
/// NFS4ERR_BADNAME       = 10041,/* name not supported */
/// NFS4ERR_BAD_RANGE     = 10042,/* lock range not supported*/
/// NFS4ERR_LOCK_NOTSUPP   = 10043,/* no atomic up/downgrade */
/// NFS4ERR_OP_ILLEGAL     = 10044,/* undefined operation */
/// NFS4ERR_DEADLOCK      = 10045,/* file locking deadlock */
/// NFS4ERR_FILE_OPEN     = 10046,/* open file blocks op. */
/// NFS4ERR_ADMIN_REVOKED  = 10047,/* lockowner state revoked */
/// NFS4ERR_CB_PATH_DOWN   = 10048 /* callback path down */
/// };
///
/// /*
///  * Basic data types
///  */
/// typedef opaque          attrlist4<>;
/// typedef uint32_t        bitmap4<>;
/// typedef uint64_t        changeid4;
/// typedef uint64_t        clientid4;
/// typedef uint32_t        count4;
/// typedef uint64_t        length4;
/// typedef uint32_t        mode4;
/// typedef uint64_t        nfs_cookie4;
/// typedef opaque          nfs_fh4<NFS4_FHSIZE>;
/// typedef uint64_t        offset4;
/// typedef uint32_t        qop4;
/// typedef opaque          sec_oid4<>;
/// typedef uint32_t        seqid4;
/// typedef opaque          utf8string<>;
/// typedef utf8string      utf8str_cis;
/// typedef utf8string      utf8str_cs;
/// typedef utf8string      utf8str_mixed;
/// typedef utf8str_cs      component4;
/// typedef opaque          linktext4;
/// typedef utf8string      ascii_REQUIRED4;
/// typedef component4      pathname4<>;
/// typedef uint64_t        nfs_lockid4;
/// typedef opaque          verifier4[NFS4_VERIFIER_SIZE];
///
///
/// /*
///  * Timeval
///  */
/// struct nfstime4 {
///     int64_t        seconds;
///     uint32_t       nseconds;
/// };
///

```



```
/// enum time_how4 {
///     SET_TO_SERVER_TIME4 = 0,
///     SET_TO_CLIENT_TIME4 = 1
/// };
///
/// union settime4 switch (time_how4 set_it) {
///     case SET_TO_CLIENT_TIME4:
///         nfstime4      time;
///     default:
///         void;
/// };
///
///
///
/// /*
///  * File attribute definitions
///  */
///
/// /*
///  * FSID structure for major/minor
///  */
/// struct fsid4 {
///     uint64_t      major;
///     uint64_t      minor;
/// };
///
///
/// /*
///  * Filesystem locations attribute for relocation/migration
///  */
/// struct fs_location4 {
///     utf8str_cis      server<>;
///     pathname4        rootpath;
/// };
///
/// struct fs_locations4 {
///     pathname4        fs_root;
///     fs_location4     locations<>;
/// };
///
///
/// /*
///  * Various Access Control Entry definitions
///  */
///
/// /*
///  * Mask that indicates which Access Control Entries
///  * are supported. Values for the fattr4_aclsupport attribute.
```

```
/// */
/// const ACL4_SUPPORT_ALLOW_ACL      = 0x00000001;
/// const ACL4_SUPPORT_DENY_ACL       = 0x00000002;
/// const ACL4_SUPPORT_AUDIT_ACL      = 0x00000004;
/// const ACL4_SUPPORT_ALARM_ACL      = 0x00000008;
///
///
/// typedef uint32_t          acetype4;
///
///
/// /*
///  * acetype4 values, others can be added as needed.
///  */
/// const ACE4_ACCESS_ALLOWED_ACE_TYPE      = 0x00000000;
/// const ACE4_ACCESS_DENIED_ACE_TYPE      = 0x00000001;
/// const ACE4_SYSTEM_AUDIT_ACE_TYPE      = 0x00000002;
/// const ACE4_SYSTEM_ALARM_ACE_TYPE      = 0x00000003;
///
///
///
/// /*
///  * ACE flag
///  */
/// typedef uint32_t          aceflag4;
///
///
///
/// /*
///  * ACE flag values
///  */
/// const ACE4_FILE_INHERIT_ACE            = 0x00000001;
/// const ACE4_DIRECTORY_INHERIT_ACE      = 0x00000002;
/// const ACE4_NO_PROPAGATE_INHERIT_ACE    = 0x00000004;
/// const ACE4_INHERIT_ONLY_ACE           = 0x00000008;
/// const ACE4_SUCCESSFUL_ACCESS_ACE_FLAG  = 0x00000010;
/// const ACE4_FAILED_ACCESS_ACE_FLAG     = 0x00000020;
/// const ACE4_IDENTIFIER_GROUP           = 0x00000040;
///
///
///
/// /*
///  * ACE mask
///  */
/// typedef uint32_t          acemask4;
///
///
///
/// /*
///  * ACE mask values
///  */
```

```
/// const ACE4_READ_DATA          = 0x00000001;
/// const ACE4_LIST_DIRECTORY      = 0x00000001;
/// const ACE4_WRITE_DATA          = 0x00000002;
/// const ACE4_ADD_FILE             = 0x00000002;
/// const ACE4_APPEND_DATA         = 0x00000004;
/// const ACE4_ADD_SUBDIRECTORY    = 0x00000004;
/// const ACE4_READ_NAMED_ATTRS    = 0x00000008;
/// const ACE4_WRITE_NAMED_ATTRS   = 0x00000010;
/// const ACE4_EXECUTE             = 0x00000020;
/// const ACE4_DELETE_CHILD        = 0x00000040;
/// const ACE4_READ_ATTRIBUTES     = 0x00000080;
/// const ACE4_WRITE_ATTRIBUTES    = 0x00000100;
///
/// const ACE4_DELETE              = 0x00010000;
/// const ACE4_READ_ACL            = 0x00020000;
/// const ACE4_WRITE_ACL           = 0x00040000;
/// const ACE4_WRITE_OWNER         = 0x00080000;
/// const ACE4_SYNCHRONIZE         = 0x00100000;
///
///
/// /*
///  * ACE4_GENERIC_READ -- defined as combination of
///  *     ACE4_READ_ACL |
///  *     ACE4_READ_DATA |
///  *     ACE4_READ_ATTRIBUTES |
///  *     ACE4_SYNCHRONIZE
///  */
///
/// const ACE4_GENERIC_READ = 0x00120081;
///
/// /*
///  * ACE4_GENERIC_WRITE -- defined as combination of
///  *     ACE4_READ_ACL |
///  *     ACE4_WRITE_DATA |
///  *     ACE4_WRITE_ATTRIBUTES |
///  *     ACE4_WRITE_ACL |
///  *     ACE4_APPEND_DATA |
///  *     ACE4_SYNCHRONIZE
///  */
///
/// const ACE4_GENERIC_WRITE = 0x00160106;
///
///
/// /*
///  * ACE4_GENERIC_EXECUTE -- defined as combination of
///  *     ACE4_READ_ACL
///  *     ACE4_READ_ATTRIBUTES
///  *     ACE4_EXECUTE
///  *     ACE4_SYNCHRONIZE
```

```
/// */
/// const ACE4_GENERIC_EXECUTE = 0x001200A0;
///
///
/// /*
///  * Access Control Entry definition
///  */
/// struct nfsace4 {
///     acetype4          type;
///     aceflag4          flag;
///     acemask4          access_mask;
///     utf8str_mixed     who;
/// };
///
///
/// /*
///  * Field definitions for the fattr4_mode attribute
///  */
/// const MODE4_SUID = 0x800; /* set user id on execution */
/// const MODE4_SGID = 0x400; /* set group id on execution */
/// const MODE4_SVTX = 0x200; /* save text even after use */
/// const MODE4_RUSR = 0x100; /* read permission: owner */
/// const MODE4_WUSR = 0x080; /* write permission: owner */
/// const MODE4_XUSR = 0x040; /* execute permission: owner */
/// const MODE4_RGRP = 0x020; /* read permission: group */
/// const MODE4_WGRP = 0x010; /* write permission: group */
/// const MODE4_XGRP = 0x008; /* execute permission: group */
/// const MODE4_OTH = 0x004; /* read permission: other */
/// const MODE4_WOTH = 0x002; /* write permission: other */
/// const MODE4_XOTH = 0x001; /* execute permission: other */
///
///
/// /*
///  * Special data/attribute associated with
///  * file types NF4BLK and NF4CHR.
///  */
/// struct specdata4 {
///     uint32_t specdata1; /* major device number */
///     uint32_t specdata2; /* minor device number */
/// };
///
///
/// /*
///  * Values for fattr4_fh_expire_type
///  */
/// const FH4_PERSISTENT = 0x00000000;
/// const FH4_NOEXPIRE_WITH_OPEN = 0x00000001;
/// const FH4_VOLATILE_ANY = 0x00000002;
```

```
/// const    FH4_VOL_MIGRATION          = 0x00000004;
/// const    FH4_VOL_RENAME              = 0x00000008;
///
///
/// typedef bitmap4                      fatattr4_supported_attrs;
/// typedef nfs_ftype4                  fatattr4_type;
/// typedef uint32_t                    fatattr4_fh_expire_type;
/// typedef changeid4                  fatattr4_change;
/// typedef uint64_t                    fatattr4_size;
/// typedef bool                        fatattr4_link_support;
/// typedef bool                        fatattr4_symlink_support;
/// typedef bool                        fatattr4_named_attr;
/// typedef fsid4                      fatattr4_fsid;
/// typedef bool                        fatattr4_unique_handles;
/// typedef uint32_t                    fatattr4_lease_time;
/// typedef nfsstat4                    fatattr4_rdatattr_error;
///
/// typedef nfsace4                     fatattr4_acl<>;
/// typedef uint32_t                    fatattr4_aclsupport;
/// typedef bool                        fatattr4_archive;
/// typedef bool                        fatattr4_cansettime;
/// typedef bool                        fatattr4_case_insensitive;
/// typedef bool                        fatattr4_case_preserving;
/// typedef bool                        fatattr4_chown_restricted;
/// typedef uint64_t                    fatattr4_fileid;
/// typedef uint64_t                    fatattr4_files_avail;
/// typedef nfs_fh4                     fatattr4_filehandle;
/// typedef uint64_t                    fatattr4_files_free;
/// typedef uint64_t                    fatattr4_files_total;
/// typedef fs_locations4               fatattr4_fs_locations;
/// typedef bool                        fatattr4_hidden;
/// typedef bool                        fatattr4_homogeneous;
/// typedef uint64_t                    fatattr4_maxfilesize;
/// typedef uint32_t                    fatattr4_maxlink;
/// typedef uint32_t                    fatattr4_maxname;
/// typedef uint64_t                    fatattr4_maxread;
/// typedef uint64_t                    fatattr4_maxwrite;
/// typedef ascii_REQUIRED4              fatattr4_mimetype;
/// typedef mode4                       fatattr4_mode;
/// typedef uint64_t                    fatattr4_mounted_on_fileid;
/// typedef bool                        fatattr4_no_trunc;
/// typedef uint32_t                    fatattr4_numlinks;
/// typedef utf8str_mixed                fatattr4_owner;
/// typedef utf8str_mixed                fatattr4_owner_group;
/// typedef uint64_t                    fatattr4_quota_avail_hard;
/// typedef uint64_t                    fatattr4_quota_avail_soft;
/// typedef uint64_t                    fatattr4_quota_used;
/// typedef specdata4                   fatattr4_rawdev;
```

```
/// typedef uint64_t          fattr4_space_avail;
/// typedef uint64_t          fattr4_space_free;
/// typedef uint64_t          fattr4_space_total;
/// typedef uint64_t          fattr4_space_used;
/// typedef bool              fattr4_system;
/// typedef nfstime4          fattr4_time_access;
/// typedef settime4          fattr4_time_access_set;
/// typedef nfstime4          fattr4_time_backup;
/// typedef nfstime4          fattr4_time_create;
/// typedef nfstime4          fattr4_time_delta;
/// typedef nfstime4          fattr4_time_metadata;
/// typedef nfstime4          fattr4_time_modify;
/// typedef settime4          fattr4_time_modify_set;
///
///
/// /*
///  * Mandatory Attributes
///  */
/// const FATTR4_SUPPORTED_ATTRS = 0;
/// const FATTR4_TYPE = 1;
/// const FATTR4_FH_EXPIRE_TYPE = 2;
/// const FATTR4_CHANGE = 3;
/// const FATTR4_SIZE = 4;
/// const FATTR4_LINK_SUPPORT = 5;
/// const FATTR4_SYMLINK_SUPPORT = 6;
/// const FATTR4_NAMED_ATTR = 7;
/// const FATTR4_FSID = 8;
/// const FATTR4_UNIQUE_HANDLES = 9;
/// const FATTR4_LEASE_TIME = 10;
/// const FATTR4_RDATTR_ERROR = 11;
/// const FATTR4_FILEHANDLE = 19;
///
/// /*
///  * Recommended Attributes
///  */
/// const FATTR4_ACL = 12;
/// const FATTR4_ACLSUPPORT = 13;
/// const FATTR4_ARCHIVE = 14;
/// const FATTR4_CANSETTIME = 15;
/// const FATTR4_CASE_INSENSITIVE = 16;
/// const FATTR4_CASE_PRESERVING = 17;
/// const FATTR4_CHOWN_RESTRICTED = 18;
/// const FATTR4_FILEID = 20;
/// const FATTR4_FILES_AVAIL = 21;
/// const FATTR4_FILES_FREE = 22;
/// const FATTR4_FILES_TOTAL = 23;
/// const FATTR4_FS_LOCATIONS = 24;
/// const FATTR4_HIDDEN = 25;
```

```
/// const FATTR4_HOMOGENEOUS      = 26;
/// const FATTR4_MAXFILESIZE      = 27;
/// const FATTR4_MAXLINK          = 28;
/// const FATTR4_MAXNAME          = 29;
/// const FATTR4_MAXREAD          = 30;
/// const FATTR4_MAXWRITE         = 31;
/// const FATTR4_MIMETYPE         = 32;
/// const FATTR4_MODE              = 33;
/// const FATTR4_NO_TRUNC          = 34;
/// const FATTR4_NUMLINKS          = 35;
/// const FATTR4_OWNER             = 36;
/// const FATTR4_OWNER_GROUP       = 37;
/// const FATTR4_QUOTA_AVAIL_HARD  = 38;
/// const FATTR4_QUOTA_AVAIL_SOFT  = 39;
/// const FATTR4_QUOTA_USED        = 40;
/// const FATTR4_RAWDEV            = 41;
/// const FATTR4_SPACE_AVAIL       = 42;
/// const FATTR4_SPACE_FREE        = 43;
/// const FATTR4_SPACE_TOTAL       = 44;
/// const FATTR4_SPACE_USED        = 45;
/// const FATTR4_SYSTEM            = 46;
/// const FATTR4_TIME_ACCESS       = 47;
/// const FATTR4_TIME_ACCESS_SET   = 48;
/// const FATTR4_TIME_BACKUP       = 49;
/// const FATTR4_TIME_CREATE       = 50;
/// const FATTR4_TIME_DELTA        = 51;
/// const FATTR4_TIME_METADATA     = 52;
/// const FATTR4_TIME_MODIFY       = 53;
/// const FATTR4_TIME_MODIFY_SET   = 54;
/// const FATTR4_MOUNTED_ON_FILEID = 55;
///
/// /*
///  * File attribute container
///  */
/// struct fattr4 {
///     bitmap4      attrmask;
///     attrlist4     attr_vals;
/// };
///
/// /*
///  * Change info for the client
///  */
/// struct change_info4 {
///     bool         atomic;
///     changeid4    before;
///     changeid4    after;
/// };
```

```
///
///
/// struct clientaddr4 {
///     /* see struct rpcb in RFC 1833 */
///     string r_netid<>;      /* network id */
///     string r_addr<>;      /* universal address */
/// };
///
///
/// /*
///  * Callback program info as provided by the client
///  */
/// struct cb_client4 {
///     unsigned int    cb_program;
///     clientaddr4     cb_location;
/// };
///
///
/// /*
///  * Stateid
///  */
/// struct stateid4 {
///     uint32_t        seqid;
///     opaque          other[NFS4_OTHER_SIZE];
/// };
///
///
/// /*
///  * Client ID
///  */
/// struct nfs_client_id4 {
///     verifier4       verifier;
///     opaque          id<NFS4_OPAQUE_LIMIT>;
/// };
///
///
/// struct open_owner4 {
///     clientid4       clientid;
///     opaque          owner<NFS4_OPAQUE_LIMIT>;
/// };
///
///
/// struct lock_owner4 {
///     clientid4       clientid;
///     opaque          owner<NFS4_OPAQUE_LIMIT>;
/// };
///
///
/// enum nfs_lock_type4 {
```



```

///      READ_LT          = 1,
///      WRITE_LT         = 2,
///      READW_LT         = 3,      /* blocking read */
///      WRITEW_LT        = 4      /* blocking write */
/// };
///
///
/// const ACCESS4_READ      = 0x00000001;
/// const ACCESS4_LOOKUP    = 0x00000002;
/// const ACCESS4_MODIFY    = 0x00000004;
/// const ACCESS4_EXTEND    = 0x00000008;
/// const ACCESS4_DELETE    = 0x00000010;
/// const ACCESS4_EXECUTE   = 0x00000020;
///
/// struct ACCESS4args {
///      /* CURRENT_FH: object */
///      uint32_t      access;
/// };
///
/// struct ACCESS4resok {
///      uint32_t      supported;
///      uint32_t      access;
/// };
///
/// union ACCESS4res switch (nfsstat4 status) {
/// case NFS4_OK:
///      ACCESS4resok      resok4;
/// default:
///      void;
/// };
///
/// struct CLOSE4args {
///      /* CURRENT_FH: object */
///      seqid4          seqid;
///      stateid4         open_stateid;
/// };
///
/// union CLOSE4res switch (nfsstat4 status) {
/// case NFS4_OK:
///      stateid4         open_stateid;
/// default:
///      void;
/// };
///
/// struct COMMIT4args {
///      /* CURRENT_FH: file */
///      offset4          offset;
///      count4           count;

```

```
/// };
///
/// struct COMMIT4resok {
///     verifier4      writeverf;
/// };
///
/// union COMMIT4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         COMMIT4resok   resok4;
///     default:
///         void;
/// };
///
/// union createtype4 switch (nfs_ftype4 type) {
///     case NF4LNK:
///         linktext4 linkdata;
///     case NF4BLK:
///     case NF4CHR:
///         specdata4 devdata;
///     case NF4SOCK:
///     case NF4FIFO:
///     case NF4DIR:
///         void;
///     default:
///         void; /* server should return NFS4ERR_BADTYPE */
/// };
///
/// struct CREATE4args {
///     /* CURRENT_FH: directory for creation */
///     createtype4   objtype;
///     component4    objname;
///     fattr4        createattrs;
/// };
///
/// struct CREATE4resok {
///     change_info4   cinfo;
///     bitmap4        attrset; /* attributes set */
/// };
///
/// union CREATE4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         CREATE4resok   resok4;
///     default:
///         void;
/// };
///
/// struct DELEGPURGE4args {
///     clientid4      clientid;
```

```
/// };
///
/// struct DELEGPURGE4res {
///     nfsstat4      status;
/// };
///
/// struct DELEGRETURN4args {
///     /* CURRENT_FH: delegated file */
///     stateid4      deleg_stateid;
/// };
///
/// struct DELEGRETURN4res {
///     nfsstat4      status;
/// };
///
/// struct GETATTR4args {
///     /* CURRENT_FH: directory or file */
///     bitmap4       attr_request;
/// };
///
/// struct GETATTR4resok {
///     fattr4        obj_attributes;
/// };
///
/// union GETATTR4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         GETATTR4resok  resok4;
///     default:
///         void;
/// };
///
/// struct GETFH4resok {
///     nfs_fh4       object;
/// };
///
/// union GETFH4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         GETFH4resok    resok4;
///     default:
///         void;
/// };
///
/// struct LINK4args {
///     /* SAVED_FH: source object */
///     /* CURRENT_FH: target directory */
///     component4    newname;
/// };
///
```

```

    /// struct LINK4resok {
    ///     change_info4    cinfo;
    /// };
    ///
    /// union LINK4res switch (nfsstat4 status) {
    ///     case NFS4_OK:
    ///         LINK4resok resok4;
    ///     default:
    ///         void;
    /// };
    ///
    /// /*
    ///  * For LOCK, transition from open_owner to new lock_owner
    ///  */
    /// struct open_to_lock_owner4 {
    ///     seqid4            open_seqid;
    ///     stateid4          open_stateid;
    ///     seqid4            lock_seqid;
    ///     lock_owner4       lock_owner;
    /// };
    ///
    /// /*
    ///  * For LOCK, existing lock_owner continues to request file locks
    ///  */
    /// struct exist_lock_owner4 {
    ///     stateid4          lock_stateid;
    ///     seqid4            lock_seqid;
    /// };
    ///
    /// union locker4 switch (bool new_lock_owner) {
    ///     case TRUE:
    ///         open_to_lock_owner4    open_owner;
    ///     case FALSE:
    ///         exist_lock_owner4      lock_owner;
    /// };
    ///
    /// /*
    ///  * LOCK/LOCKT/LOCKU: Record lock management
    ///  */
    /// struct LOCK4args {
    ///     /* CURRENT_FH: file */
    ///     nfs_lock_type4 locktype;
    ///     bool            reclaim;
    ///     offset4         offset;
    ///     length4         length;
    ///     locker4         locker;
    /// };
    ///

```

```
/// struct LOCK4denied {
///     offset4      offset;
///     length4      length;
///     nfs_lock_type4 locktype;
///     lock_owner4   owner;
/// };
///
/// struct LOCK4resok {
///     stateid4      lock_stateid;
/// };
///
/// union LOCK4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         LOCK4resok      resok4;
///     case NFS4ERR_DENIED:
///         LOCK4denied     denied;
///     default:
///         void;
/// };
///
/// struct LOCKT4args {
///     /* CURRENT_FH: file */
///     nfs_lock_type4 locktype;
///     offset4        offset;
///     length4        length;
///     lock_owner4     owner;
/// };
///
/// union LOCKT4res switch (nfsstat4 status) {
///     case NFS4ERR_DENIED:
///         LOCK4denied     denied;
///     case NFS4_OK:
///         void;
///     default:
///         void;
/// };
///
/// struct LOCKU4args {
///     /* CURRENT_FH: file */
///     nfs_lock_type4 locktype;
///     seqid4         seqid;
///     stateid4        lock_stateid;
///     offset4         offset;
///     length4         length;
/// };
///
/// union LOCKU4res switch (nfsstat4 status) {
///     case NFS4_OK:
```

```
///      stateid4      lock_stateid;
///  default:
///      void;
/// };
///
/// struct LOOKUP4args {
///      /* CURRENT_FH: directory */
///      component4      objname;
/// };
///
/// struct LOOKUP4res {
///      /* CURRENT_FH: object */
///      nfsstat4      status;
/// };
///
/// struct LOOKUP4res {
///      /* CURRENT_FH: directory */
///      nfsstat4      status;
/// };
///
/// struct NVERIFY4args {
///      /* CURRENT_FH: object */
///      fattr4      obj_attributes;
/// };
///
/// struct NVERIFY4res {
///      nfsstat4      status;
/// };
///
/// const OPEN4_SHARE_ACCESS_READ    = 0x00000001;
/// const OPEN4_SHARE_ACCESS_WRITE   = 0x00000002;
/// const OPEN4_SHARE_ACCESS_BOTH    = 0x00000003;
///
/// const OPEN4_SHARE_DENY_NONE      = 0x00000000;
/// const OPEN4_SHARE_DENY_READ      = 0x00000001;
/// const OPEN4_SHARE_DENY_WRITE     = 0x00000002;
/// const OPEN4_SHARE_DENY_BOTH      = 0x00000003;
/// /*
///  * Various definitions for OPEN
///  */
/// enum createmode4 {
///      UNCHECKED4      = 0,
///      GUARDED4        = 1,
///      EXCLUSIVE4      = 2
/// };
///
/// union createhow4 switch (createmode4 mode) {
///  case UNCHECKED4:
```

```
/// case GUARDED4:
///     fattr4          createattrs;
/// case EXCLUSIVE4:
///     verifier4       createverf;
/// };
///
/// enum opentype4 {
///     OPEN4_NOCREATE   = 0,
///     OPEN4_CREATE     = 1
/// };
///
/// union openflag4 switch (opentype4 opentype) {
///     case OPEN4_CREATE:
///         createhow4     how;
///     default:
///         void;
/// };
///
/// /* Next definitions used for OPEN delegation */
/// enum limit_by4 {
///     NFS_LIMIT_SIZE      = 1,
///     NFS_LIMIT_BLOCKS    = 2
///     /* others as needed */
/// };
///
/// struct nfs_modified_limit4 {
///     uint32_t            num_blocks;
///     uint32_t            bytes_per_block;
/// };
///
/// union nfs_space_limit4 switch (limit_by4 limitby) {
///     /* limit specified as file size */
///     case NFS_LIMIT_SIZE:
///         uint64_t         filesize;
///     /* limit specified by number of blocks */
///     case NFS_LIMIT_BLOCKS:
///         nfs_modified_limit4    mod_blocks;
/// } ;
///
/// enum open_delegation_type4 {
///     OPEN_DELEGATE_NONE    = 0,
///     OPEN_DELEGATE_READ    = 1,
///     OPEN_DELEGATE_WRITE   = 2
/// };
///
/// enum open_claim_type4 {
///     CLAIM_NULL            = 0,
///     CLAIM_PREVIOUS        = 1,
```

```
///          CLAIM_DELEGATE_CUR          = 2,
///          CLAIM_DELEGATE_PREV         = 3
/// };
///
/// struct open_claim_delegate_cur4 {
///     stateid4      delegate_stateid;
///     component4    file;
/// };
///
/// union open_claim4 switch (open_claim_type4 claim) {
///     /*
///     * No special rights to file.
///     * Ordinary OPEN of the specified file.
///     */
///     case CLAIM_NULL:
///         /* CURRENT_FH: directory */
///         component4    file;
///     /*
///     * Right to the file established by an
///     * open previous to server reboot. File
///     * identified by filehandle obtained at
///     * that time rather than by name.
///     */
///     case CLAIM_PREVIOUS:
///         /* CURRENT_FH: file being reclaimed */
///         open_delegation_type4    delegate_type;
///     /*
///     * Right to file based on a delegation
///     * granted by the server. File is
///     * specified by name.
///     */
///     case CLAIM_DELEGATE_CUR:
///         /* CURRENT_FH: directory */
///         open_claim_delegate_cur4    delegate_cur_info;
///     /*
///     * Right to file based on a delegation
///     * granted to a previous boot instance
///     * of the client. File is specified by name.
///     */
///     case CLAIM_DELEGATE_PREV:
///         /* CURRENT_FH: directory */
///         component4    file_delegate_prev;
///     };
///
///     /*
///     * OPEN: Open a file, potentially receiving an open delegation
```



```
/// */
/// struct OPEN4args {
///     seqid4          seqid;
///     uint32_t         share_access;
///     uint32_t         share_deny;
///     open_owner4      owner;
///     openflag4        openhow;
///     open_claim4      claim;
/// };
///
/// struct open_read_delegation4 {
///     stateid4 stateid; /* Stateid for delegation*/
///     bool      recall; /* Pre-recalled flag for
///                       delegations obtained
///                       by reclaim (CLAIM_PREVIOUS) */
///     nfsace4 permissions; /* Defines users who don't
///                           need an ACCESS call to
///                           open for read */
/// };
///
/// struct open_write_delegation4 {
///     stateid4 stateid; /* Stateid for delegation */
///     bool      recall; /* Pre-recalled flag for
///                       delegations obtained
///                       by reclaim
///                       (CLAIM_PREVIOUS) */
///     nfs_space_limit4
///     space_limit; /* Defines condition that
///                  the client must check to
///                  determine whether the
///                  file needs to be flushed
///                  to the server on close. */
///     nfsace4 permissions; /* Defines users who don't
///                           need an ACCESS call as
///                           part of a delegated
///                           open. */
/// };
///
/// union open_delegation4
/// switch (open_delegation_type4 delegation_type) {
///     case OPEN_DELEGATE_NONE:
///         void;
///     case OPEN_DELEGATE_READ:
///         open_read_delegation4 read;
///     case OPEN_DELEGATE_WRITE:
```

```

    ///          open_write_delegation4 write;
    /// };
    ///
    /// /*
    ///  * Result flags
    ///  */
    ///
    /// /* Client must confirm open */
    /// const OPEN4_RESULT_CONFIRM      = 0x000000002;
    /// /* Type of file locking behavior at the server */
    /// const OPEN4_RESULT_LOCKTYPE_POSIX = 0x000000004;
    ///
    /// struct OPEN4resok {
    ///     stateid4      stateid;      /* Stateid for open */
    ///     change_info4   cinfo;        /* Directory Change Info */
    ///     uint32_t        rflags;      /* Result flags */
    ///     bitmap4        attrset;      /* attribute set for create */
    ///     open_delegation4 delegation; /* Info on any open
    ///                                     delegation */
    /// };
    ///
    /// union OPEN4res switch (nfsstat4 status) {
    ///     case NFS4_OK:
    ///         /* CURRENT_FH: opened file */
    ///         OPEN4resok      resok4;
    ///     default:
    ///         void;
    /// };
    ///
    /// struct OPENATTR4args {
    ///     /* CURRENT_FH: object */
    ///     bool      createdir;
    /// };
    ///
    /// struct OPENATTR4res {
    ///     /* CURRENT_FH: named attr directory */
    ///     nfsstat4      status;
    /// };
    ///
    /// struct OPEN_CONFIRM4args {
    ///     /* CURRENT_FH: opened file */
    ///     stateid4      open_stateid;
    ///     seqid4        seqid;
    /// };
    ///
    /// struct OPEN_CONFIRM4resok {
    ///     stateid4      open_stateid;
    /// };

```

```
///
/// union OPEN_CONFIRM4res switch (nfsstat4 status) {
///   case NFS4_OK:
///       OPEN_CONFIRM4resok      resok4;
///   default:
///       void;
/// };
///
/// struct OPEN_DOWNGRADE4args {
///     /* CURRENT_FH: opened file */
///     stateid4      open_stateid;
///     seqid4        seqid;
///     uint32_t      share_access;
///     uint32_t      share_deny;
/// };
///
/// struct OPEN_DOWNGRADE4resok {
///     stateid4      open_stateid;
/// };
///
/// union OPEN_DOWNGRADE4res switch(nfsstat4 status) {
///   case NFS4_OK:
///       OPEN_DOWNGRADE4resok      resok4;
///   default:
///       void;
/// };
///
/// struct PUTFH4args {
///     nfs_fh4      object;
/// };
///
/// struct PUTFH4res {
///     /* CURRENT_FH: */
///     nfsstat4      status;
/// };
///
/// struct PUTPUBFH4res {
///     /* CURRENT_FH: public fh */
///     nfsstat4      status;
/// };
///
/// struct PUTROOTFH4res {
///     /* CURRENT_FH: root fh */
///     nfsstat4      status;
/// };
///
/// struct READ4args {
///     /* CURRENT_FH: file */
```

```

    ///      stateid4      stateid;
    ///      offset4      offset;
    ///      count4       count;
    /// };
    ///
    /// struct READ4resok {
    ///      bool           eof;
    ///      opaque         data<>;
    /// };
    ///
    /// union READ4res switch (nfsstat4 status) {
    ///      case NFS4_OK:
    ///          READ4resok      resok4;
    ///      default:
    ///          void;
    /// };
    ///
    /// struct READDIR4args {
    ///      /* CURRENT_FH: directory */
    ///      nfs_cookie4      cookie;
    ///      verifier4        cookieverf;
    ///      count4           dircount;
    ///      count4           maxcount;
    ///      bitmap4          attr_request;
    /// };
    ///
    /// struct entry4 {
    ///      nfs_cookie4      cookie;
    ///      component4       name;
    ///      fattr4           attrs;
    ///      entry4           *nextentry;
    /// };
    ///
    /// struct dirlist4 {
    ///      entry4           *entries;
    ///      bool             eof;
    /// };
    ///
    /// struct READDIR4resok {
    ///      verifier4        cookieverf;
    ///      dirlist4         reply;
    /// };
    ///
    /// union READDIR4res switch (nfsstat4 status) {
    ///      case NFS4_OK:
    ///          READDIR4resok  resok4;
    ///      default:

```

```
///          void;
/// };
///
///
/// struct READLINK4resok {
///     linktext4      link;
/// };
///
/// union READLINK4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         READLINK4resok resok4;
///     default:
///         void;
/// };
///
/// struct REMOVE4args {
///     /* CURRENT_FH: directory */
///     component4      target;
/// };
///
/// struct REMOVE4resok {
///     change_info4     cinfo;
/// };
///
/// union REMOVE4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         REMOVE4resok   resok4;
///     default:
///         void;
/// };
///
/// struct RENAME4args {
///     /* SAVED_FH: source directory */
///     component4      oldname;
///     /* CURRENT_FH: target directory */
///     component4      newname;
/// };
///
/// struct RENAME4resok {
///     change_info4     source_cinfo;
///     change_info4     target_cinfo;
/// };
///
/// union RENAME4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         RENAME4resok   resok4;
///     default:
///         void;
```

```

    /// };
    ///
    /// struct RENEW4args {
    ///     clientid4      clientid;
    /// };
    ///
    /// struct RENEW4res {
    ///     nfsstat4      status;
    /// };
    ///
    /// struct RESTOREFH4res {
    ///     /* CURRENT_FH: value of saved fh */
    ///     nfsstat4      status;
    /// };
    ///
    /// struct SAVEFH4res {
    ///     /* SAVED_FH: value of current fh */
    ///     nfsstat4      status;
    /// };
    ///
    /// struct SECINFO4args {
    ///     /* CURRENT_FH: directory */
    ///     component4     name;
    /// };
    ///
    /// /*
    ///  * From RFC 2203
    ///  */
    /// enum rpc_gss_svc_t {
    ///     RPC_GSS_SVC_NONE      = 1,
    ///     RPC_GSS_SVC_INTEGRITY = 2,
    ///     RPC_GSS_SVC_PRIVACY   = 3
    /// };
    ///
    /// struct rpcsec_gss_info {
    ///     sec_oid4      oid;
    ///     qop4          qop;
    ///     rpc_gss_svc_t service;
    /// };
    ///
    /// /* RPCSEC_GSS has a value of '6' - See RFC 2203 */
    /// union secinfo4 switch (uint32_t flavor) {
    ///     case RPCSEC_GSS:
    ///         rpcsec_gss_info      flavor_info;
    ///     default:
    ///         void;
    /// };
    ///
    ///

```

```
/// typedef secinfo4 SECINFO4resok<>;
///
/// union SECINFO4res switch (nfsstat4 status) {
///   case NFS4_OK:
///     SECINFO4resok resok4;
///   default:
///     void;
/// };
///
/// struct SETATTR4args {
///   /* CURRENT_FH: target object */
///   stateid4      stateid;
///   fattr4        obj_attributes;
/// };
///
/// struct SETATTR4res {
///   nfsstat4      status;
///   bitmap4       attrset;
/// };
///
/// struct SETCLIENTID4args {
///   nfs_client_id4 client;
///   cb_client4     callback;
///   uint32_t       callback_ident;
/// };
///
/// struct SETCLIENTID4resok {
///   clientid4      clientid;
///   verifier4      setclientid_confirm;
/// };
///
/// union SETCLIENTID4res switch (nfsstat4 status) {
///   case NFS4_OK:
///     SETCLIENTID4resok      resok4;
///   case NFS4ERR_CLID_INUSE:
///     clientaddr4      client_using;
///   default:
///     void;
/// };
///
/// struct SETCLIENTID_CONFIRM4args {
///   clientid4      clientid;
///   verifier4      setclientid_confirm;
/// };
///
/// struct SETCLIENTID_CONFIRM4res {
///   nfsstat4      status;
/// };
///
```

```
///
/// struct VERIFY4args {
///     /* CURRENT_FH: object */
///     fattr4          obj_attributes;
/// };
///
/// struct VERIFY4res {
///     nfsstat4        status;
/// };
///
/// enum stable_how4 {
///     UNSTABLE4        = 0,
///     DATA_SYNC4      = 1,
///     FILE_SYNC4       = 2
/// };
///
/// struct WRITE4args {
///     /* CURRENT_FH: file */
///     stateid4         stateid;
///     offset4          offset;
///     stable_how4      stable;
///     opaque            data<>;
/// };
///
/// struct WRITE4resok {
///     count4           count;
///     stable_how4      committed;
///     verifier4        writeverf;
/// };
///
/// union WRITE4res switch (nfsstat4 status) {
///     case NFS4_OK:
///         WRITE4resok    resok4;
///     default:
///         void;
/// };
///
/// struct RELEASE_LOCKOWNER4args {
///     lock_owner4      lock_owner;
/// };
///
/// struct RELEASE_LOCKOWNER4res {
///     nfsstat4        status;
/// };
///
/// struct ILLEGAL4res {
///     nfsstat4        status;
/// };
///
```



```
///
/// /*
///  * Operation arrays
///  */
///
/// enum nfs_opnum4 {
///     OP_ACCESS           = 3,
///     OP_CLOSE            = 4,
///     OP_COMMIT           = 5,
///     OP_CREATE           = 6,
///     OP_DELEGPURGE       = 7,
///     OP_DELEGRETURN      = 8,
///     OP_GETATTR          = 9,
///     OP_GETFH            = 10,
///     OP_LINK             = 11,
///     OP_LOCK             = 12,
///     OP_LOCKT            = 13,
///     OP_LOCKU            = 14,
///     OP_LOOKUP           = 15,
///     OP_LOOKUPP          = 16,
///     OP_NVERIFY          = 17,
///     OP_OPEN             = 18,
///     OP_OPENATTR         = 19,
///     OP_OPEN_CONFIRM     = 20,
///     OP_OPEN_DOWNGRADE   = 21,
///     OP_PUTFH            = 22,
///     OP_PUTPUBFH         = 23,
///     OP_PUTROOTFH        = 24,
///     OP_READ             = 25,
///     OP_READDIR          = 26,
///     OP_READLINK         = 27,
///     OP_REMOVE           = 28,
///     OP_RENAME           = 29,
///     OP_RENEW            = 30,
///     OP_RESTOREFH        = 31,
///     OP_SAVEFH           = 32,
///     OP_SECINFO          = 33,
///     OP_SETATTR          = 34,
///     OP_SETCLIENTID      = 35,
///     OP_SETCLIENTID_CONFIRM = 36,
///     OP_VERIFY           = 37,
///     OP_WRITE            = 38,
///     OP_RELEASE_LOCKOWNER = 39,
///     OP_ILLEGAL          = 10044
/// };
///
/// union nfs_argop4 switch (nfs_opnum4 argop) {
///     case OP_ACCESS:      ACCESS4args opaccess;
```

```
/// case OP_CLOSE:          CLOSE4args opclose;
/// case OP_COMMIT:         COMMIT4args opcommit;
/// case OP_CREATE:         CREATE4args opcreate;
/// case OP_DELEGPURGE:      DELEGPURGE4args opdelegpurge;
/// case OP_DELEGRETURN:     DELEGRETURN4args opdelegreturn;
/// case OP_GETATTR:        GETATTR4args opgetattr;
/// case OP_GETFH:          void;
/// case OP_LINK:           LINK4args oplink;
/// case OP_LOCK:           LOCK4args oplock;
/// case OP_LOCKT:          LOCKT4args oplockt;
/// case OP_LOCKU:          LOCKU4args oplocku;
/// case OP_LOOKUP:         LOOKUP4args oplookup;
/// case OP_LOOKUPP:        void;
/// case OP_NVERIFY:        NVERIFY4args opnverify;
/// case OP_OPEN:           OPEN4args opopen;
/// case OP_OPENATTR:       OPENATTR4args opopenattr;
/// case OP_OPEN_CONFIRM:   OPEN_CONFIRM4args opopen_confirm;
/// case OP_OPEN_DOWNGRADE: OPEN_DOWNGRADE4args opopen_downgrade;
///
/// case OP_PUTFH:          PUTFH4args opputfh;
/// case OP_PUTPUBFH:       void;
/// case OP_PUTROOTFH:      void;
/// case OP_READ:           READ4args opread;
/// case OP_READDIR:        READDIR4args opreaddir;
/// case OP_READLINK:       void;
/// case OP_REMOVE:         REMOVE4args opremove;
/// case OP_RENAME:         RENAME4args oprename;
/// case OP_RENEW:          RENEW4args oprenew;
/// case OP_RESTOREFH:      void;
/// case OP_SAVEFH:         void;
/// case OP_SECINFO:        SECINFO4args opsecinfo;
/// case OP_SETATTR:        SETATTR4args opsetattr;
/// case OP_SETCLIENTID:    SETCLIENTID4args opsetclientid;
/// case OP_SETCLIENTID_CONFIRM: SETCLIENTID_CONFIRM4args
///                               opsetclientid_confirm;
/// case OP_VERIFY:         VERIFY4args opverify;
/// case OP_WRITE:          WRITE4args opwrite;
/// case OP_RELEASE_LOCKOWNER:
///
///                               RELEASE_LOCKOWNER4args
///                               oprelease_lockowner;
/// case OP_ILLEGAL:        void;
/// };
///
/// union nfs_resop4 switch (nfs_opnum4 resop) {
/// case OP_ACCESS:         ACCESS4res opaccess;
/// case OP_CLOSE:         CLOSE4res opclose;
/// case OP_COMMIT:        COMMIT4res opcommit;
/// case OP_CREATE:        CREATE4res opcreate;
```

```

/// case OP_DELEGPURGE:      DELEGPURGE4res opdelegpurge;
/// case OP_DELEGRETURN:     DELEGRETURN4res opdelegreturn;
/// case OP_GETATTR:         GETATTR4res opgetattr;
/// case OP_GETFH:           GETFH4res opgetfh;
/// case OP_LINK:             LINK4res oplink;
/// case OP_LOCK:             LOCK4res oplock;
/// case OP_LOCKT:            LOCKT4res oplockt;
/// case OP_LOCKU:            LOCKU4res oplocku;
/// case OP_LOOKUP:           LOOKUP4res oplookup;
/// case OP_LOOKUPP:          LOOKUPP4res oplookupp;
/// case OP_NVERIFY:          NVERIFY4res opnverify;
/// case OP_OPEN:             OPEN4res opopen;
/// case OP_OPENATTR:         OPENATTR4res opopenattr;
/// case OP_OPEN_CONFIRM:     OPEN_CONFIRM4res opopen_confirm;
/// case OP_OPEN_DOWNGRADE:   OPEN_DOWNGRADE4res
///                             opopen_downgrade;
/// case OP_PUTFH:            PUTFH4res opputfh;
/// case OP_PUTPUBFH:         PUTPUBFH4res opputpubfh;
/// case OP_PUTROOTFH:        PUTROOTFH4res opputrootfh;
/// case OP_READ:             READ4res opread;
/// case OP_READDIR:          READDIR4res opreaddir;
/// case OP_READLINK:         READLINK4res opreadlink;
/// case OP_REMOVE:           REMOVE4res opremove;
/// case OP_RENAME:           RENAME4res oprename;
/// case OP_RENEW:            RENEW4res oprenew;
/// case OP_RESTOREFH:        RESTOREFH4res oprestorefh;
/// case OP_SAVEFH:           SAVEFH4res opsavefh;
/// case OP_SECINFO:          SECINFO4res opsecinfo;
/// case OP_SETATTR:          SETATTR4res opsetattr;
/// case OP_SETCLIENTID:      SETCLIENTID4res opsetclientid;
/// case OP_SETCLIENTID_CONFIRM:
///                             SETCLIENTID_CONFIRM4res
///                             opsetclientid_confirm;
/// case OP_VERIFY:           VERIFY4res opverify;
/// case OP_WRITE:            WRITE4res opwrite;
/// case OP_RELEASE_LOCKOWNER:
///                             RELEASE_LOCKOWNER4res
///                             oprelease_lockowner;
/// case OP_ILLEGAL:          ILLEGAL4res opillegal;
/// };
///
/// struct COMPOUND4args {
///     utf8str_cs      tag;
///     uint32_t         minorversion;
///     nfs_argop4       argarray<>;
/// };
///

```

```

    /// struct COMPOUND4res {
    ///     nfsstat4      status;
    ///     utf8str_cs     tag;
    ///     nfs_resop4     resarray<>;
    /// };
    ///
    /// /*
    ///  * Remote file service routines
    ///  */
    /// program NFS4_PROGRAM {
    ///     version NFS_V4 {
    ///         void
    ///             NFSPROC4_NULL(void) = 0;
    ///
    ///             COMPOUND4res
    ///             NFSPROC4_COMPOUND(COMPOUND4args) = 1;
    ///
    ///     } = 4;
    /// } = 100003;
    ///
    /// /*
    ///  * NFS4 Callback Procedure Definitions and Program
    ///  */
    /// struct CB_GETATTR4args {
    ///     nfs_fh4 fh;
    ///     bitmap4 attr_request;
    /// };
    ///
    /// struct CB_GETATTR4resok {
    ///     fattr4 obj_attributes;
    /// };
    ///
    /// union CB_GETATTR4res switch (nfsstat4 status) {
    ///     case NFS4_OK:
    ///         CB_GETATTR4resok      resok4;
    ///     default:
    ///         void;
    /// };
    ///
    /// struct CB_RECALL4args {
    ///     stateid4      stateid;
    ///     bool          truncate;
    ///     nfs_fh4       fh;
    /// };
    ///
    /// struct CB_RECALL4res {
    ///     nfsstat4      status;

```

```

    /// };
    ///
    /// /*
    ///  * CB_ILLEGAL: Response for illegal operation numbers
    ///  */
    /// struct CB_ILLEGAL4res {
    ///     nfsstat4      status;
    /// };
    ///
    /// /*
    ///  * Various definitions for CB_COMPOUND
    ///  */
    /// %
    /// enum nfs_cb_opnum4 {
    ///     OP_CB_GETATTR          = 3,
    ///     OP_CB_RECALL           = 4,
    ///     OP_CB_ILLEGAL          = 10044
    /// };
    ///
    /// union nfs_cb_argop4 switch (unsigned argop) {
    ///     case OP_CB_GETATTR:
    ///         CB_GETATTR4args      opcbgetattr;
    ///     case OP_CB_RECALL:
    ///         CB_RECALL4args       opcbrecall;
    ///     case OP_CB_ILLEGAL:
    ///         void;
    /// };
    ///
    /// union nfs_cb_resop4 switch (unsigned resop) {
    ///     case OP_CB_GETATTR:      CB_GETATTR4res  opcbgetattr;
    ///     case OP_CB_RECALL:      CB_RECALL4res   opcbrecall;
    ///     case OP_CB_ILLEGAL:      CB_ILLEGAL4res  opcbillegal;
    /// };
    ///
    ///
    /// struct CB_COMPOUND4args {
    ///     utf8str_cs      tag;
    ///     uint32_t         minorversion;
    ///     uint32_t         callback_ident;
    ///     nfs_cb_argop4    argarray<>;
    /// };
    ///
    /// struct CB_COMPOUND4res {
    ///     nfsstat4      status;
    ///     utf8str_cs    tag;
    ///     nfs_cb_resop4 resarray<>;
    /// };
    ///
    ///

```

```
///
/// /*
///  * Program number is in the transient range since the client
///  * will assign the exact transient program number and provide
///  * that to the server via the SETCLIENTID operation.
///  */
/// program NFS4_CALLBACK {
///     version NFS_CB {
///         void
///         CB_NULL(void) = 0;
///         CB_COMPOUND4res
///         CB_COMPOUND(CB_COMPOUND4args) = 1;
///     } = 1;
/// } = 0x40000000;
```

## 2. Security Considerations

See the Security Considerations section of [I-D.ietf-nfsv4-rfc3530bis].

## 3. IANA Considerations

This document does not have any IANA considerations.

## 4. Normative References

- [I-D.ietf-nfsv4-rfc3530bis]  
Haynes, T. and D. Noveck, "NFS Version 4 Protocol",  
draft-ietf-nfsv4-rfc3530bis-27 (work in progress),  
Aug 2013.
- [RFC4506] Eisler, M., "XDR: External Data Representation Standard",  
STD 67, RFC 4506, May 2006.

## Appendix A. Acknowledgments

David Quigley tested the extraction of the .x file from this document and corrected the two resulting errors.

## Appendix B. RFC Editor Notes

[RFC Editor: please remove this section prior to publishing this document as an RFC]

[RFC3530bis should be replaced by the RFC number of draft-ietf-nfsv4-rfc3530bis in this draft.]

[RFC Editor: Please note that there is also a reference entry that needs to be modified for the companion document.]

#### Authors' Addresses

Thomas Haynes (editor)  
NetApp  
495 E Java Dr  
Sunnyvale, CA 95054  
USA

Phone: +1 408 419 3018  
Email: thomas@netapp.com

David Noveck (editor)  
EMC Corporation  
32 Coslin Drive  
Southborough, MA 01772  
US

Phone: +1 508 305 8404  
Email: novecd@emc.com





NFSv4  
Internet-Draft  
Intended status: Standards Track  
Expires: April 20, 2014

W. Adamson  
NetApp  
N. Williams  
Cryptonector  
October 17, 2013

Remote Procedure Call (RPC) Security Version 3  
draft-ietf-nfsv4-rpcsec-gssv3-05.txt

## Abstract

This document specifies version 3 of the Remote Procedure Call (RPC) security protocol (RPCSEC\_GSS). This protocol provides for compound authentication of client hosts and users to server (constructed by generic composition), security label assertions for multi-level and type enforcement, structured privilege assertions, and channel bindings.

## Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [1].

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 20, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents  
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Applications of RPCSEC_GSSv3 . . . . .	4
2. The RPCSEC_GSSv3 protocol . . . . .	5
2.1. New auth_stat values . . . . .	9
2.2. RPC message credential and verifier . . . . .	10
2.3. Control Messages . . . . .	10
2.3.1. Create request . . . . .	11
2.3.2. Destruction request . . . . .	15
2.3.3. List request . . . . .	16
2.3.4. Extensibility . . . . .	16
2.4. Data Messages . . . . .	17
3. Security Considerations . . . . .	17
4. IANA Considerations . . . . .	18
5. References . . . . .	19
5.1. Normative References . . . . .	19
5.2. Informative References . . . . .	19
Appendix A. Acknowledgments . . . . .	20
Appendix B. RFC Editor Notes . . . . .	20
Authors' Addresses . . . . .	20

## 1. Introduction

The original RPCSEC\_GSS protocol [2] provided for authentication of RPC clients and servers to each other using the Generic Security Services Application Programming Interface (GSS-API) [3]. The second version of RPCSEC\_GSS [4] added support for channel bindings [5].

We find that GSS-API mechanisms are insufficient for communicating certain aspects of a client's authority to a server. The GSS-API and its mechanisms certainly could be extended to address this shortcoming, but it seems be far simpler to address it at the application layer, namely, in this case, RPCSEC\_GSS.

The motivation for RPCSEC\_GSSv3 is to add support for labeled security and server-side copy for NFSv4 (see [6] and [9]). Both of these features require assertions of authority from the client.

Assertions need to be verified. One party that can verify an assertion is the client host, which can authenticate to the server using its own credentials. We can also require users to verify an assertion as well. This calls for compound authentication.

Because the design of RPCSEC\_GSSv3 relies on either RPCSEC\_GSS version 1 (though version 2 can be used) to do the actual GSS-API security context establishment, we add support for channel binding so that implementors who have implemented RPCSEC\_GSS version 1 but not version 2 can provide a (simplified) channel binding implementation using RPCSEC\_GSSv3.

We therefore describe a new version of RPCSEC\_GSS that allows for the following:

- o Client-side assertions of authority:
  - \* Security labels for multi-level, type enforcement, and other labeled security models. See [10], [11], [12], [6] and [9].
  - \* Application-specific structured privileges. For an example see server-side copy [6].
  - \* Compound authentication of the client host and user to the server done by binding two RPCSEC\_GSS handles.
  - \* Simplified channel binding.

Assertions of labels and privileges are evaluated by the server, which may then map the asserted values to other values, all according to server-side policy.

We also add an option for enumerating active server-side privileges and supported label format specifiers (LFS). The LFS and Label Format Registry are described in detail in [13].

RPCSEC\_GSSv3 is patterned as follows:

- o A client uses an existing RPCSEC\_GSSv1 (or RPCSEC\_GSSv2) context handle to protect RPCSEC\_GSSv3 exchanges (this will be termed the "parent" handle)
- o The server issues a "child" RPCSEC\_GSSv3 handle, but the underlying GSS-API security context for the parent handle is used in all subsequent exchanges using the child handle. This works because the RPCSEC\_GSS handle is included in the integrity protected RPCSEC\_GSS auth/verifier header for all versions of RPCSEC\_GSS. The child context, however, has its own sequence number space and window, distinct from that of the parent.

[[Comment.1: RFC22203 states that when data integrity is used, the seq\_num in the rpc\_gss\_data\_t must be the same as in the credential. This means that using data integrity with GSS3 context's can not simply construct it using the parent context as the seq\_num must be from the GSS3 context. --AA]]

This means that RPCSEC\_GSSv3 depends on RPCSEC\_GSS versions 1 and/or 2 for actual GSS-API security context establishment. This keeps the specification of RPCSEC\_GSSv3 simple by avoiding the need to duplicate the core functionality of RPCSEC\_GSS version 1.

#### 1.1. Applications of RPCSEC\_GSSv3

The common uses of RPCSEC\_GSSv3, particularly for NFSv4 [6], are expected to be:

- a. labeled security: client-side process label assertion [+ privilege assertion] + compound client host & user authentication;
- b. compound client host & user authentication [+ critical structured privilege assertions] used in inter-server server-side copy;

Labeled NFS (see Section 8 of [6]) uses the subject label provided by the client via the RPCSEC\_GSSv3 layer to enforce MAC access to objects owned by the server to enable server guest mode or full mode labeled NFS.

[[Comment.2: check that this language states what NFSv4.2 labeled NFS problem we are really solving. (setting labels on the server) --AA]]

A traditional inter-server file copy entails the user gaining access to a file on the source, reading it, and writing it to a file on the destination. In secure NFSv4 inter-server server-side copy (see Section 3.4.1 of [6]), the user first secures access to both source and destination files, and then uses RPCSEC\_GSSv3 compound authentication and structured privileges to authorize the destination to copy the file from the source on behalf of the user.

## 2. The RPCSEC\_GSSv3 protocol

This document contains the External Data Representation (XDR) ([7]) definitions for the RPCSEC\_GSSv3 protocol.

The XDR description is provided in this document in a way that makes it simple for the reader to extract into ready to compile form. The reader can feed this document in the following shell script to produce the machine readable XDR description of RPCSEC\_GSSv3:

```
#!/bin/sh
grep "^ *///" | sed 's?^ */// ??' | sed 's?^ *///$??'
```

I.e. if the above script is stored in a file called "extract.sh", and this document is in a file called "spec.txt", then the reader can do:

```
sh extract.sh < spec.txt > rpcsec_gss_v3.x
```

The effect of the script is to remove leading white space from each line, plus a sentinel sequence of "///".

The XDR description, with the sentinel sequence follows:

```
/// /*
///  * Copyright (c) 2013 IETF Trust and the persons
///  * identified as the document authors. All rights
///  * reserved.
///  *
///  * The document authors are identified in [RFC2203],
///  * [RFC5403], and [RFCxxxx].
///  *
///  * Redistribution and use in source and binary forms,
///  * with or without modification, are permitted
///  * provided that the following conditions are met:
///  *
///  * o Redistributions of source code must retain the above
///  *   copyright notice, this list of conditions and the
///  *   following disclaimer.
///  *
```

```
/// * o Redistributions in binary form must reproduce the
/// *   above copyright notice, this list of
/// *   conditions and the following disclaimer in
/// *   the documentation and/or other materials
/// *   provided with the distribution.
/// *
/// * o Neither the name of Internet Society, IETF or IETF
/// *   Trust, nor the names of specific contributors, may be
/// *   used to endorse or promote products derived from this
/// *   software without specific prior written permission.
/// *
/// *   THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS
/// *   AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED
/// *   WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
/// *   IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS
/// *   FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO
/// *   EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE
/// *   LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL,
/// *   EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT
/// *   NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR
/// *   SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS
/// *   INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF
/// *   LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY,
/// *   OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING
/// *   IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF
/// *   ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
/// */
///
/// /*
/// * This code was derived from [RFC2203]. Please
/// * reproduce this note if possible.
/// */
///
/// /*
/// * rpcsec_gss_v3.x
/// */
///
/// enum rpc_gss_service_t {
///     /* Note: the enumerated value for 0 is reserved. */
///     rpc_gss_svc_none           = 1,
///     rpc_gss_svc_integrity      = 2,
///     rpc_gss_svc_privacy        = 3,
///     rpc_gss_svc_channel_prot   = 4
/// };
///
/// enum rpc_gss_proc_t {
///     RPCSEC_GSS_DATA            = 0,
///     RPCSEC_GSS_INIT            = 1,
```

```
///          RPCSEC_GSS_CONTINUE_INIT = 2,
///          RPCSEC_GSS_DESTROY        = 3,
///          RPCSEC_GSS_BIND_CHANNEL   = 4
/// };
///
/// struct rpc_gss_cred_vers_1_t {
///     rpc_gss_proc_t      gss_proc; /* control procedure */
///     unsigned int        seq_num;  /* sequence number */
///     rpc_gss_service_t   service;  /* service used */
///     opaque              handle<>; /* context handle */
/// };
///
/// enum rpc_gss3_proc_t {
///     RPCSEC_GSS3_DATA = 0,
///     RPCSEC_GSS3_LIST = 5,
///     RPCSEC_GSS3_CREATE = 6,
///     RPCSEC_GSS3_DESTROY = 7
/// };
///
/// struct rpc_gss_cred_vers_3_t {
///     rpc_gss3_proc_t      gss_proc;
///     unsigned int        seq_num;
///     rpc_gss_service_t   service;
///     opaque              handle<>;
/// };
///
/// const RPCSEC_GSS_VERS_1 = 1;
/// const RPCSEC_GSS_VERS_2 = 2;
/// const RPCSEC_GSS_VERS_3 = 3; /* new */
///
/// union rpc_gss_cred_t switch (unsigned int rgc_version) {
///     case RPCSEC_GSS_VERS_1:
///     case RPCSEC_GSS_VERS_2:
///         rpc_gss_cred_vers_1_t rgc_cred_v1;
///     case RPCSEC_GSS_VERS_3: /* new */
///         rpc_gss_cred_vers_3_t rgc_cred_v3;
/// };
///
/// const MAXSEQ = 0x80000000;
///
/// struct rpc_gss3_extension {
///     int      type;
///     bool     critical;
///     opaque   data<>;
/// };
///
/// struct rpc_gss3_gss_binding {
///     unsigned int    vers;
```

```
///      opaque      handle<>;
///      opaque      nonce<>;
///      opaque      mic<>;
/// };
///
/// typedef opaque rpc_gss3_chan_binding<>;
///
/// struct rpc_gss3_lfs {
///     unsigned int lfs_id;
///     unsigned int pi_id;
/// };
///
/// struct rpc_gss3_label {
///     rpc_gss3_lfs    lfs;
///     opaque          label<>;
/// };
///
/// typedef string rpc_gss3_list_name<>;
/// struct rpc_gss3_privs {
///     rpc_gss3_list_name    listname;
///     opaque                privilege<>;
/// };
///
/// enum rpc_gss3_assertion_type {
///     LABEL = 0,
///     PRIVS = 1
/// };
///
/// union rpc_gss3_assertion_u
///     switch (rpc_gss3_assertion_type atype) {
/// case LABEL:
///     rpc_gss3_label    label;
/// case PRIVILEGES:
///     rpc_gss3_privs    privs;
/// default:
///     opaque            ext<>;
/// };
///
/// struct rpc_gss3_assertion {
///     bool                critical;
///     rpc_gss3_assertion_u    assertion;
/// };
///
/// struct rpc_gss3_create_args {
///     rpc_gss3_gss_binding    *compound_binding;
///     rpc_gss3_chan_binding    *chan_binding_mic;
///     rpc_gss3_assertion        assertions<>;
///     rpc_gss3_extension        extensions<>;
```



```

    /// };
    ///
    /// struct rpc_gss3_create_res {
    ///     opaque                handle<>;
    ///     rpc_gss3_chan_binding *chan_binding_mic;
    ///     rpc_gss3_assertion    granted_assertions<>;
    ///     rpc_gss3_extension    granted_extensions<>;
    /// };
    ///
    /// enum rpc_gss3_list_item {
    ///     LABEL = 0,
    ///     PRIV = 1,
    /// };
    ///
    /// struct rpc_gss3_list_args {
    ///     rpc_gss3_list_item    list_what<>;
    /// };
    ///
    /// union rpc_gss3_list_item_u
    ///     switch (rpc_gss3_list_item itype) {
    /// case LABEL:
    ///     rpc_gss3_lable        labels<>;
    /// case PRIV:
    ///     rpc_gss3_list_name    privs<>;
    /// default:
    ///     opaque                ext<>;
    /// };
    ///
    /// typedef rpc_gss3_list_item_u rpc_gss3_list_res<>;

```

## 2.1. New auth\_stat values

RPCSEC\_GSSv3 requires the addition of several values to the auth\_stat enumerated type definition:

```

enum auth_stat {
    ...
    /*
     * RPCSEC_GSS errors
     */
    RPCSEC_GSS3_COMPOUND_PROBLEM = <>,
    RPCSEC_GSS3_LABEL_PROBLEM = <>,
    RPCSEC_GSS3_UNKNOWN_ASSERTION = <>
    RPCSEC_GSS3_UNKNOWN_EXTENSION = <>
    RPCSEC_GSS3_UNKNOWN_MESSAGE = <>
};

```

[[Comment.3: fix above into YYY. All the entries are TBD... --NW]]

## 2.2. RPC message credential and verifier

The `rpc_gss_cred_vers_3_t` type is used in much the same way that `rpc_gss_cred_vers_1_t` is used in `RPCSEC_GSSv1`, that is: as the arm of the `rpc_gss_cred_t` discriminated union in the RPC message header `opaque_auth` structure corresponding to version 3 (`RPCSEC_GSS_VERS_3`). It differs from `rpc_gss_cred_vers_1_t` in that:

- a. the values for `gss_proc` corresponding to control messages are different.
- b. the `handle` field is the `RPCSEC_GSSv3` (child) handle, except for the `RPCSEC_GSS3_CREATE` control message where it is set to the parent context handle.

For all `RPCSEC_GSSv3` data and control messages, the `verifier` field in the RPC message header is constructed in the `RPCSEC_GSSv1` manner using the parent GSS-API security context.

## 2.3. Control Messages

There are three `RPCSEC_GSSv3` control messages: `RPCSEC_GSS3_CREATE`, `RPCSEC_GSS3_DESTROY`, and `RPCSEC_GSS3_LIST`.

`RPCSEC_GSSv3` control messages are similar to the `RPCSEC_GSSv1` `RPCSEC_GSS_DESTROY` control message (see section 5.4 [2]) in that the sequence number in the request must be valid, and the header checksum in the verifier must be valid. In other words, they look a lot like an `RPCSEC_GSSv3` data message with the header procedure set to `NULLPROC`.

As in `RPCSEC_GSSv1`, the `RPCSEC_GSSv3` control messages may contain information following the verifier in the body of the `NULLPROC` procedure.

The client **MUST** use one of the following security services to protect any `RPCSEC_GSSv3` control message:

- o `rpc_gss_svc_channel_prot` (see `RPCSEC_GSSv2`)
- o `rpc_gss_svc_integrity`
- o `rpc_gss_svc_privacy`

Specifically the client **MUST NOT** use `rpc_gss_svc_none`.

For `RPCSEC_GSSv3` control messages the `rpc_gss_cred_vers_3_t` in the RPC message `opaque_auth` structure is encoded as follows:

1. the union `rpc_gss_cred_t` version is set to 3 with the value being of type `rpc_gss_cred_vers_3_t` instead of `rpc_gss_cred_vers_1_t`.
2. the `gss_proc` is set to one of `RPCSEC_GSS3_CREATE`, `RPCSEC_GSS3_DESTROY`, or `RPCSEC_GSS3_LIST`.
3. the `seq_num` is a valid sequence number for the context in the handle field.
4. the `rpc_gss_service_t` is one of `rpc_gss_svc_integrity`, `rpc_gss_svc_privacy`, or `rpc_gss_svc_channel_prot`.
5. the `rpc_gss_cred_vers_3_t` handle field is either set to the parent context handle for `RPCSEC_GSS3_CREATE`, or to the GSS3 child handle for `RPCSEC_GSS3_LIST` and `RPCSEC_GSS3_DESTROY`.

#### 2.3.1. Create request

As noted in the introduction, `RPCSEC_GSSv3` relies on the `RPCSEC_GSS` version 1 parent context (though version 2 can be used) secure connection to do the actual GSS-API GSS3 security context establishment. As such, the `rpc_gss_cred_vers_3_t` fields in the RPC Call `opaque_auth` use the parent context handle and `seq_num` stream.

The `RPCSEC_GSS3_CREATE` call message binds one or more items of several kinds into a new `RPCSEC_GSSv3` context handle:

- o another `RPCSEC_GSS` (version 1, 2, or 3) context handle (compound authentication)
- o a channel binding
- o authorization assertions (labels, privileges)
- o extensions (see Section 2.3.4 )

The reply to this message consists of either an error or an `rpc_gss3_create_res` structure which includes a new `RPCSEC_GSSv3` handle, termed the "child" which is used for subsequent control and data messages.

Upon successful `RPCSEC_GSS3_CREATE`, both the client and the server should associate the resultant GSSv3 child context handle with the parent context handle in their GSS context caches so as to be able to reference the parent context given the child context handle.

[[Comment.4: Destruction of the parent context => first destroy child handle. IOW fail the `RPCSEC_GSS_DESTROY` of parent with new

RPCSEC\_GSS3\_CONTEXT\_EXISTS error code: What about the lifetime of the GSS3 context. Is this meant to be long lived?? --AA]]

Server policies should take into account the identity of the client and/or user as authenticated via the GSS-API. Server implementation and policy MAY result in labels, privileges, and identities being mapped to concepts and values that are local to the server.

#### 2.3.1.1. Compound authentication

RPCSEC\_GSSv3 allows for compound authentication of client hosts and users to servers. As in non-compound authentication, there is a parent handle used to protect the RPCSEC\_GSS3\_CREATE call message, and a resultant RPCSEC\_GSSv3 child handle. In addition to the parent handle, the compound authentication create control message has a handle referenced via the `compound_binding` field of the RPCSEC\_GSS3\_CREATE arguments structure (`rpc_gss3_create_args`) termed the "inner" handle, as well as a nonce and a MIC of that nonce created using the GSS-API security context associated with the "inner" handle.

All uses of a child context handle that is bound to an inner context MUST be treated as speaking for the initiator principal (as modified by any assertions in the RPCSEC\_GSS3\_CREATE message) of the inner context handle's GSS-API security context.

This feature is needed, for example, when a client wishes to use authority assertions that the server may only grant if a user and a client are authenticated together to the server. Thus a server may refuse to grant requested authority to a user acting alone (e.g., via an unprivileged user-space program), or to a client acting alone (e.g. when a client is acting on behalf of a user) but may grant requested authority to a client acting on behalf of a user if the server identifies the user and trusts the client.

It is assumed that an unprivileged user-space program would not have access to client host credentials needed to establish a GSS-API security context authenticating the client to the server, therefore an unprivileged user-space program could not create an RPCSEC\_GSSv3 RPCSEC\_GSS3\_CREATE message that successfully binds a client and a user security context.

Clients using RPCSEC\_GSS context binding MUST use, as the parent context handle, an RPCSEC\_GSS context handle that corresponds to a GSS-API security context that authenticates the client host, and for the inner context handle it SHOULD use a context handle to authenticate a user. The reverse (parent handle authenticates user, inner authenticates client) MUST NOT be used. Other compounds might

eventually make sense.

An `RPCSEC_GSSv3` context handle that is bound to another `RPCSEC_GSS` context **MUST** be treated by servers as authenticating the GSS-API initiator principal authenticated by the inner context handle's GSS-API security context. This principal may be mapped to a server-side notion of user or principal as modified by any identity assertions by the client in the same `RPCSEC_GSS3_CREATE` request that the server accepts.

#### 2.3.1.2. Channel binding

`RPCSEC_GSSv3` provides a different way to do channel binding than `RPCSEC_GSSv2`. Specifically:

- a. `RPCSEC_GSSv3` builds on `RPCSEC_GSSv1` by reusing existing, established context handles rather than providing a different RPC security flavor for establishing context handles,
- b. channel bindings data are not hashed because the community now agrees that it is the secure channel's responsibility to produce channel bindings data of manageable size.

(a) is useful in keeping `RPCSEC_GSSv3` simple in general, not just for channel binding. (b) is useful in keeping `RPCSEC_GSSv3` simple specifically for channel binding.

Channel binding is accomplished as follows. The client prefixes the channel bindings data octet string with the channel type as described in [5], then the client calls `GSS_GetMIC()` to get a MIC of resulting octet string, using the parent `RPCSEC_GSS` context handle's GSS-API security context. The MIC is then placed in the `chan_binding_mic` field of `RPCSEC_GSS3_CREATE` arguments (`rpc_gss3_create_args`).

If the `chan_binding_mic` field of the arguments of a `RPCSEC_GSS3_CREATE` control message is set, then the server **MUST** verify the client's channel binding MIC if the server supports this feature. If channel binding verification succeeds then the server **MUST** generate a new MIC of the same channel bindings and place it in the `chan_binding_mic` field of the `RPCSEC_GSS3_CREATE` results. If channel binding verification fails or the server doesn't support channel binding then the server **MUST** indicate this in its reply by not including a `chan_binding_mic` value (`chan_binding_mic` is an optional field).

The client **MUST** verify the result's `chan_binding_mic` value, if the server included it, by calling `GSS_VerifyMIC()` with the given MIC and the channel bindings data (including the channel type prefix). If

client-side channel binding verification fails then the client MUST call `RPCSEC_GSS3_DESTROY`. If the client requested channel binding but the server did not include a `chan_binding_mic` field in the results, then the client MAY continue to use the resulting context handle as though channel binding had never been requested, otherwise (if the client really wanted channel binding) it MUST call `RPCSEC_GSS3_DESTROY`.

As per-RPCSEC\_GSSv2 [4]:

"Once a successful [channel binding] procedure has been performed on an [RPCSEC\_GSSv3] context handle, the initiator's implementation may map application requests for `rpc_gss_svc_none` and `rpc_gss_svc_integrity` to `rpc_gss_svc_channel_prot` credentials. And if the secure channel has privacy enabled, requests for `rpc_gss_svc_privacy` can also be mapped to `rpc_gss_svc_channel_prot`."

Any `RPCSEC_GSSv3` context handle that has been bound to a secure channel in this way SHOULD be used only with the `rpc_gss_svc_channel_prot`, and SHOULD NOT be used with `rpc_gss_svc_none` nor `rpc_gss_svc_integrity` -- if the secure channel does not provide privacy protection then the client MAY use `rpc_gss_svc_privacy` where privacy protection is needed or desired.

#### 2.3.1.3. Label assertions

`RPCSEC_GSSv3` clients MAY assert a security label in some LSF by binding this assertion into an `RPCSEC_GSSv3` context handle. This is done by including an assertion of type `rpc_gss3_label` in the 'assertions' field (discriminant: 'LABEL') of the `RPCSEC_GSS3_CREATE` arguments to the desired LSF and label.

Label encoding is specified to mirror the NFSv4 `sec_label` attribute described in Section 12.2.2 of [6]. The label format specifier (LFS) is an identifier used by the client to establish the syntactic format of the security label and the semantic meaning of its components. The policy identifier (PI) is an optional part of the definition of an LFS which allows for clients and server to identify specific security policies. The opaque label field of `rpc_gss3_label` is dependent on the MAC model to interpret and enforce.

[[Comment.5: Check that this Label definition provides all the required pieces to enable full mode when combined with NFSv4.2 LNFS. Specifically, how does the client find out and respond if a server has changed a label. --AA]]

If a label itself requires privacy protection (i.e., that the user

can assert that label is a secret) then the client MUST use the `rpc_gss_svc_privacy` protection service for the `RPCSEC_GSS3_CREATE` request or, if the parent handle is bound to a secure channel that provides privacy protection, `rpc_gss_svc_channel_prot`.

If a client wants to ensure that the server understands the asserted label then it MUST set the 'critical' field of the label assertion to TRUE, otherwise it MUST set it to FALSE.

Servers that do not support labeling MUST ignore non-critical label assertions. Servers that do not support the requested LFS MUST either ignore non-critical label assertions or map them to a suitable label in a supported LFS. Servers that do not support labeling or do not support the requested LFS MUST return an error if the label request is critical. Servers that support labeling in the requested LFS MAY map the requested label to different label as a result of server-side policy evaluation.

#### 2.3.1.4. Structured privilege assertions

A structured privilege is an RPC application defined structure that is opaque, and is encoded in the `rpc_gss3_privs` privilege field. Encoding, server verification and any server policies for structured privileges are described by the RPC application definition. The `listname` field of `rpc_gss3_privs` is a description string used to list the privilege.

A successful structured privilege assertion `RPCSEC_GSS3_CREATE` call must return all accepted privileges in the `rpc_gss3_privs` `granted_assertions` field.

Section 3.4.1.2. "Inter-Server Copy with `RPCSEC_GSSv3`" of [6] shows an example of structured privilege definition and use.

#### 2.3.2. Destruction request

The `RPCSEC_GSS3_DESTROY` control message is the same as the `RPCSEC_GSSv1` `RPCSEC_GSS_DESTROY` control message, but with the version 3 header. Specifically, the `rpc_gss_cred_vers_3_t` fields in the RPC Call `opaque_auth` use the GSS3 context handle and `seq_num` stream. As with all `RPCSEC_GSSv3` messages, the header checksum uses the parent context, and needs to be valid.

The server sends a response as it would to a data request. The client and server must then destroy the context for the session.

### 2.3.3. List request

The `RPCSEC_GSS3_LIST` control message is similar to `RPCSEC_GSS3_DESTROY` message. Specifically, the `rpc_gss_cred_vers_3_t` fields in the RPC Call `opaque_auth` use the GSS3 context handle and `seq_num` stream. As with all `RPCSEC_GSSv3` messages, the header checksum uses the parent context, and needs to be valid.

The `RPCSEC_GSS3_LIST` control message consists of a single integer indicating what should be listed, and the reply consists of an error or the requested list. The client may list LFSs or structured privilege listnames.

The result is an opaque octet string containing a list of LFSs [encoding TBD] or a list of active structured privileges [encoding TBD].

### 2.3.4. Extensibility

New fields may be added through the 'extensions' typed hole. All such extensions have a 'critical' flag.

[[Comment.6: Should we keep the extensions types hole? I think not... --AA]]

Assertion types may be added in the future by adding arms to the 'rpc\_gss3\_assertion\_u' union. Every assertion has a 'critical' flag that can be used to indicate criticality. Other assertion types are described elsewhere and include:

- o Client-side assertions of identity:
  - \* Primary client/user identity
  - \* Supplementary group memberships of the client/user, including support for specifying deltas to the membership list as seen on the server.

New control message types may be added.

Servers receiving unknown critical client assertions or unknown `RPCSEC_GSS_v3` extensions MUST return an error.

There is no IANA or other registry for `RPCSEC_GSSv3` extensions. All extensions MUST be done by IETF Protocol Action.



## 2.4. Data Messages

RPCSEC\_GSS3\_DATA messages differ from from RPCSEC\_GSSv1 data messages in that the version number used MUST be '3' instead of '1'. As noted in Section 2.2 the RPCSEC\_GSSv3 context handle is used along with it's sequence number stream.

For RPCSEC\_GSSv3 data messages the `rpc_gss_cred_vers_3_t` in the RPC message `opaque_auth` structure is encoded as follows:

1. the union `rpc_gss_cred_t` version is set to 3 with the value being of type `rpc_gss_cred_vers_3_t` instead of `rpc_gss_cred_vers_1_t`.
2. the `gss_proc` is set to `RPCSEC_GSS3_DATA`
3. the `seq_num` is a valid GSS3 context (child context) sequence number.
4. just as in `RPCSEC_GSSv1`, the `rpc_gss_service_t` is one of `rpc_gss_svc_none`, `rpc_gss_svc_integrity`, `rpc_gss_svc_privacy`, or `rpc_gss_svc_channel_prot`.
5. the handle field is set to the (child) `RPCSEC_GSSv3` context handle

## 3. Security Considerations

This entire document deals with security issues.

The `RPCSEC_GSSv3` protocol allows for client-side assertions of data that is relevant to server-side authorization decisions. These assertions must be evaludated by the server in the context of whether the client and/or user are authenticated, whether compound authentication was used, whether the client is trusted, what ranges of assertions are allowed for the client and the user (separately or together), and any relevant server-side policy.

The security semantics of assertions carried by `RPCSEC_GSSv3` are application protocol-specific.

`RPCSEC_GSSv3` supports a notion of critical assertions (and extensions), but there's no need for peers to tell each other what assertions were granted, or what they were mapped to.

Note that `RPSEC_GSSv3` is not a complete solution for labeling: it conveys the labels of actors, but not the labels of objects. RPC application protocols may require extending in order to carry object

label information.

There may be interactions with NFSv4's callback security scheme and NFSv4.1's GSS-API "SSV" mechanisms. Specifically, the NFSv4 callback scheme requires that the server initiate GSS-API security contexts, which does not work well in practice, and in the context of client-side processes running as the same user but with different privileges and security labels the NFSv4 callback security scheme seems particularly unlikely to work well. NFSv4.1 has the server use an existing, client-initiated RPCSEC\_GSS context handle to protect server-initiated callback RPCs. The NFSv4.1 callback security scheme lacks all the problems of the NFSv4 scheme, however, it is important that the server pick an appropriate RPCSEC\_GSS context handle to protect any callbacks. Specifically, it is important that the server use RPCSEC\_GSS context handles which authenticate the client to protect any callbacks relating to server state initiated by RPCs protected by RPCSEC\_GSSv3 contexts.

[[Comment.7: [Add text about interaction with GSS-SSV...] --NW]]

[[Comment.8: I see no reason to use RPCSEC\_GSSv3 contexts for NFSv4.x back channel. --AA]]

[[Comment.9: Since GSS3 requires an RPCSEC\_GSSv1 or v2 context handle to establish a GSS3 context, SSV can not be used as this draft is written.]]

[[Comment.10: AFAICS the reason to use SSV is to avoid using a client machine credential which means compound authentication can not be used. Since GSS3 requires an RPCSEC\_GSSv1 or v2 context handle to establish a GSS3 context, SSV can not be used as the parent context for GSSv3. --AA]]

#### 4. IANA Considerations

This section uses terms that are defined in [8].

There are no IANA considerations in this document. TBDs in this document will be assigned by the ONC RPC registrar (which is not IANA, XXX: verify).

#### 5. References

### 5.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", March 1997.
- [2] Eisler, M., Chiu, A., and L. Ling, "RPCSEC\_GSS Protocol Specification", RFC 2203, September 1997.
- [3] Linn, J., "Generic Security Service Application Program Interface Version 2, Update 1", RFC 2743, January 2000.
- [4] Srinivasan, R., "RPC: Remote Procedure Call Protocol Specification Version 2", RFC 1831, August 1995.
- [5] Williams, N., "On the Use of Channel Bindings to Secure Channels", RFC 5056, November 2007.
- [6] Haynes, T., "NFS Version 4 Minor Version 2", draft-ietf-nfsv4-minorversion2-19 (Work In Progress), March 2013.
- [7] Eisler, M., "XDR: External Data Representation Standard", RFC 4506, May 2006.
- [8] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

### 5.2. Informative References

- [9] Haynes, T., "Requirements for Labeled NFS", draft-ietf-nfsv4-labreqs-03 (work in progress).
- [10] "Section 46.6. Multi-Level Security (MLS) of Deployment Guide: Deployment, configuration and administration of Red Hat Enterprise Linux 5, Edition 6", 2011.
- [11] Smalley, S., "The Distributed Trusted Operating System (DTOS) Home Page", <<http://www.cs.utah.edu/flux/fluke/html/dtos/HTML/dtos.html>>.
- [12] Carter, J., "Implementing SELinux Support for NFS", <[http://www.nsa.gov/research/\\_files/selinux/papers/nfsv3.pdf](http://www.nsa.gov/research/_files/selinux/papers/nfsv3.pdf)>.
- [13] Quigley, D. and J. Lu, "Registry Specification for MAC Security Label Formats", draft-quigley-label-format-registry (work in progress), 2011.

Appendix A. Acknowledgments

Appendix B. RFC Editor Notes

[RFC Editor: please remove this section prior to publishing this document as an RFC]

[RFC Editor: prior to publishing this document as an RFC, please replace all occurrences of RFCTBD10 with RFCxxxx where xxxx is the RFC number of this document]

Authors' Addresses

William A. (Andy) Adamson  
NetApp  
3629 Wagner Ridge Ctt  
Ann Arbor, MI 48103  
USA

Phone: +1 734 665 1204  
Email: andros@netapp.com

Nico Williams  
cryptonector.com  
13115 Tamayo Dr  
Austin, TX 78729  
USA

Email: nico@cryptonector.com

