Network File System Version 4
Internet-Draft
Intended status: Experimental
Expires: April 2, 2017

C. Lever
Oracle

September 29, 2016

RDMA Connection Manager Private Messages For RPC-Over-RDMA Version One
draft-cel-nfsv4-rpcrdma-cm-pvt-msg-00

Abstract

This document specifies the format of RDMA-CM Private Data exchanged between RPC-over-RDMA Version One peers. Such messages indicate peer support for Remote Invalidation and larger-than-default inline thresholds, but can be extended. The Private Data message format defined in this document is experimental only.

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 2, 2017.

Copyright Notice

Table of Contents

1. Introduction

RPC-over-RDMA Version One, specified in [I-D.ietf-nfsv4-rfc5666bis], enables the use of direct data placement for upper layer protocols based on RPC [RFC5531]. However, there are some recognized shortcomings of the RPC-over-RDMA Version One protocol. The two most immediate shortcomings are:

- Setting up an explicit RDMA operation (RDMA Read or Write) can be costly. The small default size of inline thresholds requires the use of explicit RDMA operations even for relatively small messages and data payloads.

- Unlike most other contemporary RDMA-enabled storage protocols, there is no facility in RPC-over-RDMA Version One that enables the use of Remote Invalidation [RFC5042].

The original specification of RPC-over-RDMA Version One provided an out-of-band protocol for passing inline threshold settings between connected peers. However, [I-D.ietf-nfsv4-rfc5666bis] deprecates this protocol because it was not fully specified and thus it was never implemented.

Work on [I-D.ietf-nfsv4-rfc5666bis] has demonstrated that the RPC-over-RDMA Version One protocol as it stands is challenging to extend while maintaining interoperability. Therefore, another out-of-band mechanism is required to help relieve these limitations for RPC-over-RDMA Version One implementations.

This document specifies a simple, non-XDR-based message format designed to pass between RPC-over-RDMA Version One peers when an RDMA transport connection is first established. The purpose of this message format is to enable experimentation with parameters of the base transport layer over which RPC-over-RDMA runs. Future versions of RPC-over-RDMA may make use of these experimental results, providing similar information exchange as part of the XDR-defined base transport protocol.

2. Advertised Transport Capabilities

2.1. Inline Threshold Size

Section 4.3.2 of [I-D.ietf-nfsv4-rfc5666bis] defines the term "inline threshold." There are a pair of inline thresholds per transport connection, one for each direction of message flow, which limit the size of messages conveyed using RDMA Send. If an incoming message exceeds the size of a receiver's inline threshold, the receive operation fails and the connection is typically terminated. To send a message larger than a receiver's inline threshold, an NFS client uses explicit RDMA operations, which are typically more costly than RDMA Send.

The default value of this threshold for RPC-over-RDMA Version One connections is 1024 bytes (see Section 4.3.3 of [I-D.ietf-nfsv4-rfc5666bis]). This is adequate for nearly all NFS Version 3 procedures. NFS Version 4 COMPOUNDs are larger, on average, forcing clients to use explicit RDMA operations for frequently-issued requests such as LOOKUP and GETATTR.

If a sender and receiver can agree on a larger inline threshold, a greater portion of frequently-issued NFS Version 4 operations can avoid the use of explicit RDMA operations. Explicit RDMA can be avoided for smaller I/O requests as well.

Thus each peer advertises the largest message size it can send and the largest size it can receive. The requester MUST use the smaller of its maximum send size and the responder's maximum receive size as the requester-to-responder inline threshold. The responder MUST use the smaller of its maximum send size and the requester's maximum receive size as the responder-to-requester inline threshold.

2.2. Support for Remote Invalidation

A description of Remote Invalidation and a full discussion of the design issues can be found in [I-D.cel-nfsv4-reminv-design].

Without altering the XDR definition of RPC-over-RDMA Version One messages that carry chunk lists, it's not possible to provide fully generic support for Remote Invalidation. However, it is possible to provide a simple signaling mechanism for a requester to indicate it can deal with Responder's Choice (see Section 2.3 of [I-D.cel-nfsv4-reminv-design]). In this case, the responder is allowed to invalidate any STag in an RPC-over-RDMA request.

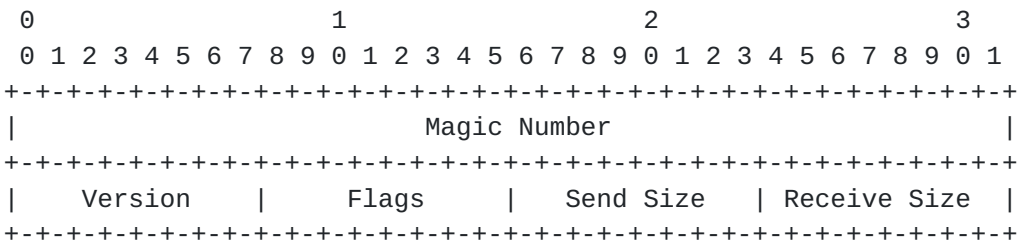
Thus each peer advertises its ability to support Responder's Choice Remote Invalidation. If both peers support it, then the responder MAY use RDMA Send With Invalidate rather than RDMA Send to convey RPC-over-RDMA reply messages.

3. Private Data Message Format

When an RPC-over-RDMA Version One transport connection is established, a requester and responder MAY populate the CM Private Data field exchanged as part of CM connection establishment (refer to Section 12.7.35 of [IBTA-IB]). For RPC-over-RDMA Version One, the CM Private Data field is formatted as described in this section. Requesters and responders use the same format.

3.1. Fixed Mandatory Fields

The first 8 octets of the CM Private Data field MUST be formatted as follows:



Magic Number
: This field contains a fixed 32-bit value that identifies the content of the Private Data field as an RPC-over-RDMA Version One CM Private Data message. The value of this field MUST be 0xf6ab0e18, in big-endian order.

Version
: This 8-bit field contains a message format version number. The value "1" in this field means only the first eight octets are present, they appear in the order described in this section, and they each have the meaning defined in this section.

Flags
: This 8-bit field contains eight boolean flags that indicate the support status of optional features, such as Remote Invalidation. The meaning of these flags is defined in Section 3.1.1.

Send Size
: This 8-bit field contains an encoded value corresponding to the largest message size this peer can send using RDMA Send. The value is encoded as described in Section 3.1.2.

Receive Size
: This 8-bit field contains an encoded value corresponding to the largest message size this peer can receive via posted receive buffers. The value is encoded as described in Section 3.1.2.

3.1.1. Feature Support Flags

The bits in the Flags field are labeled from bit 8 to bit 15, as shown in the diagram above. When the Version field contains the value "1", the bits in the Flags field have the following meaning:

Bit 15
: When this bit is asserted (one), the sender supports the use of Remote Invalidation, as described in [I-D.cel-nfsv4-reminv-design]. When this bit is clear (zero), the sender does not support Remote Invalidation.

Bits 14 - 8
: These bits are reserved and must be clear (zero).

3.1.2. Inline Threshold Encoding

Inline threshold sizes from 1KB to 256KB can be represented in the Send Size and Receive Size fields. A sender computes the encoded value by dividing the actual value by 1024 and subtracting one from the result. A receiver decodes this value by performing complementary operations.

3.2. Extending The Private Message Format

The Private Data format described above can be extended to add additional optional fields which follow the first eight octets or to make use of one of the reserved bits in the Flags fields. To introduce such changes while preserving interoperability, a new Version number is allocated, and new fields and bit flags are defined. A description of how receivers should behave if they do not recognize the new format must also be provided. If this document is

still a personal draft in the Experiemental category, it must be updated to document the new Private Data message format as above.

4. Interoperability Considerations

This extension is intended to interoperate with other RPC-over-RDMA Version One implementations which do not support the exchange of CM Private Data. When a peer does not receive a CM Private Data message which conforms to Section 3, it MUST assume the remote peer supports only the default RPC-over-RDMA Version One settings as defined in [I-D.ietf-nfsv4-rfc5666bis]. In other words, the peer behaves as if a Private Data message was received in which bit 8 of the Flags field is clear (zero), and both Size fields contain the value zero.

5. IANA Considerations

There are no IANA considerations for this document.

6. Security Considerations

RDMA-CM Private Data typically traverses the link layer in the clear. The same considerations apply here that are described in the Security Considerations section of [I-D.ietf-nfsv4-rfc5666bis].

7. References

7.1. Normative References

[I-D.ietf-nfsv4-rfc5666bis]
Lever, C., Simpson, W., and T. Talpey, "Remote Direct Memory Access Transport for Remote Procedure Call, Version One", draft-ietf-nfsv4-rfc5666bis-07 (work in progress), May 2016.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC5042] Pinkerton, J. and E. Deleganes, "Direct Data Placement Protocol (DDP) / Remote Direct Memory Access Protocol (RDMAP) Security", RFC 5042, DOI 10.17487/RFC5042, October 2007, <http://www.rfc-editor.org/info/rfc5042>.

7.2. Informative References

[I-D.cel-nfsv4-reminv-design]
Lever, C., "Using Remote Invalidation With RPC-Over-RDMA Transports", draft-cel-nfsv4-reminv-design-04 (work in progress), September 2016.

[IBTA-IB] InfiniBand Trade Association, "InfiniBand(TM) Architecture Specification Volume 1 Release 1.2", November 2007, <http://www.infinibandta.org>.

[RFC5531] Thurlow, R., "RPC: Remote Procedure Call Protocol Specification Version 2", RFC 5531, DOI 10.17487/RFC5531, May 2009, <http://www.rfc-editor.org/info/rfc5531>.

Appendix A. Acknowledgments

Thanks to Christoph Hellwig of HGST and Devesh Sharma of Broadcom for suggesting this approach.

Special thanks go to Transport Area Director Spencer Dawkins, nfsv4 Working Group Chair Spencer Shepler, and nfsv4 Working Group Secretary Thomas Haynes for their support.

Author's Address

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

Phone: +1 734 274 2396
Email: chuck.lever@oracle.com