Workgroup: NFSv4
Internet-Draft:
draft-hellwig-nfsv4-scsi-layout-nvme-02
Published: 2 June 2022
Intended Status: Informational
Expires: 4 December 2022
Authors: C. Hellwig

Using the Parallel NFS (pNFS) SCSI Layout with NVMe

Abstract

This document explains how to use the Parallel Network File System (pNFS) SCSI Layout Type with transports using the NVMe or NVMe over Fabrics protocol.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The pNFS Small Computer System Interface (SCSI) layout [RFC8154] is a layout type that allows NFS clients to directly perform I/O to block storage devices while bypassing the MDS. It is specified by using concepts from the SCSI protocol family for the data path to the storage devices. This documents explains how to access PCI Express, RDMA or Fibre Channel devices using the NVM Express protocol [NVME] using the SCSI layout type by leveraging the SCSI Translation Reference ([NVME-STLR]). This document does not amend the pNFS SCSI layout document in any way, instead of explains how to map the SCSI constructs used in the pNFS SCSI layout document to NVMe concepts using the NVMe SCSI translation reference.

1.1. Conventions Used in This Document

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].

1.2. General Definitions

The following definitions are provided for the purpose of providing an appropriate context for the reader.

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.

Server The "server" is the entity responsible for coordinating client access to a set of file systems and is identified by a server owner.

2. SCSI Layout mapping to NVMe

The SCSI layout definition [RFC8154] only references few SCSI specific concepts directly. This document uses the NVMe SCSI Translation Reference document ([NVME-STLR]) to provide mappings from these SCSI concepts to NVM Express ([NVME]) concepts that SHOULD be used when using the pNFS SCSI layout with NVMe devices.

The NVMe SCSI Translation Reference is used to define the NVMe command and concepts that SHOULD be used to implement the pNFS SCSI layout. Implementations MAY or MAY not use an actual SCSI to NVMe translation layer.

2.1. Volume Identification

The pNFS SCSI layout uses the Device Identification VPD page (page code 0x83) from [SPC4] to identify the devices used by a layout. Section 6.1.4 of [NVME-STLR] lists ways to build SCSI Device Identification descriptors from NVMe Identify data. To be used as storage devices for the pNFS SCSI layout, NVMe devices MUST support either the EUI64 or NGUID value in the Identify Namespace data, as the methods based on the Serial Number for legacy devices might not be suitable for unique addressing needs and thus MUST NOT be used. If possible NVMe devices uses as storage devices for the pNFS SCSI layout SHOULD support the NGUID value as it is the larger identifier.

2.2. Client Fencing

The SCSI layout uses Persistent Reservations to provide client fencing. For this both the MDS and the Clients have to register a key with the storage device, and the MDS has to create a reservation on the storage device. Section 6.7 of [NVME-STLR] contains a full mapping of the required PERSISTENT RESERVE IN and PERSISTENT RESERVE OUT SCSI command to NVMe commands which SHOULD be used when using NVMe devices as storage devices for the pNFS SCSI layout. One important difference between SCSI and NVMe Persistent Reservations is that NVMe reservation keys always apply to all controllers used by a host (as indicated by the NVMe HOSTID). This behavior is somewhat similar to setting the ALL_TG_PT bit when registering a SCSI Reservation key, but actually guaranteed to work reliably.

2.3. Volatile write caches

The equivalent of the WCE bit in the Caching Mode Page in [SBC3] is the Write Cache Enable field in the NVMe Get Features command, see Section 6.3.3.2 of [NVME-STLR]. If a write cache is enable on a NVMe device used as a storage device for the pNFS SCSI layout, the MDS must ensure to use the NVMe FLUSH command to flush the volatile write cache.

3. Security Considerations

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

4. IANA Considerations

The document does not require any actions by IANA.

5. Normative References

[NVME] NVM Express, Inc., "NVM Express Revision 1.3", May 2017.

[NVME-STLR] NVM Express, Inc., "NVM Express: SCSI Translation Reference Revision 1.5", June 2015.

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

[RFC8154] Hellwig, C., "Parallel NFS (pNFS) Small Computer System Interface (SCSI) Layout", May 2017.

[SBC3] INCITS Technical Committee T10, "SCSI Block Commands-3", ANSI INCITS INCITS 514-2014, ISO/IEC 14776-323, 2014.

[SPC4] INCITS Technical Committee T10, "SCSI Primary Commands-4", ANSI INCITS 513-2015, 2015.

Author's Address

Christoph Hellwig

Email: hch@lst.de