Using the pNFS SCSI/NVMe Layout over Fabrics
draft-faibish-nfsv4-scsi-nvme-layout-over-fabrics-00

Sorin Faibish <faibish.sorin@dell.com>

David Black <david.black@dell.com>

Christoph Hellwig <hch@lst.de>

Motivation: Add support for new NVMe-oF transport protocol to pNFS

- NVMe-oF: NVMe over Fabrics
  - Fabric extension of NVMe (Non Volatile Memory Express) SSD interface
- New storage systems support NVMe-oF transports to access NVMe-based storage
  - The NVMe devices are faster than older SCSI devices
  - But connecting them using old transports and switches introduces inefficiencies – New faster transports are needed
  - Memory of storage servers may use persistent memory (e.g., Intel Optane)
  - New hosts can access remote persistent memory using RDMA

Motivation: extending pNFS to NVMe

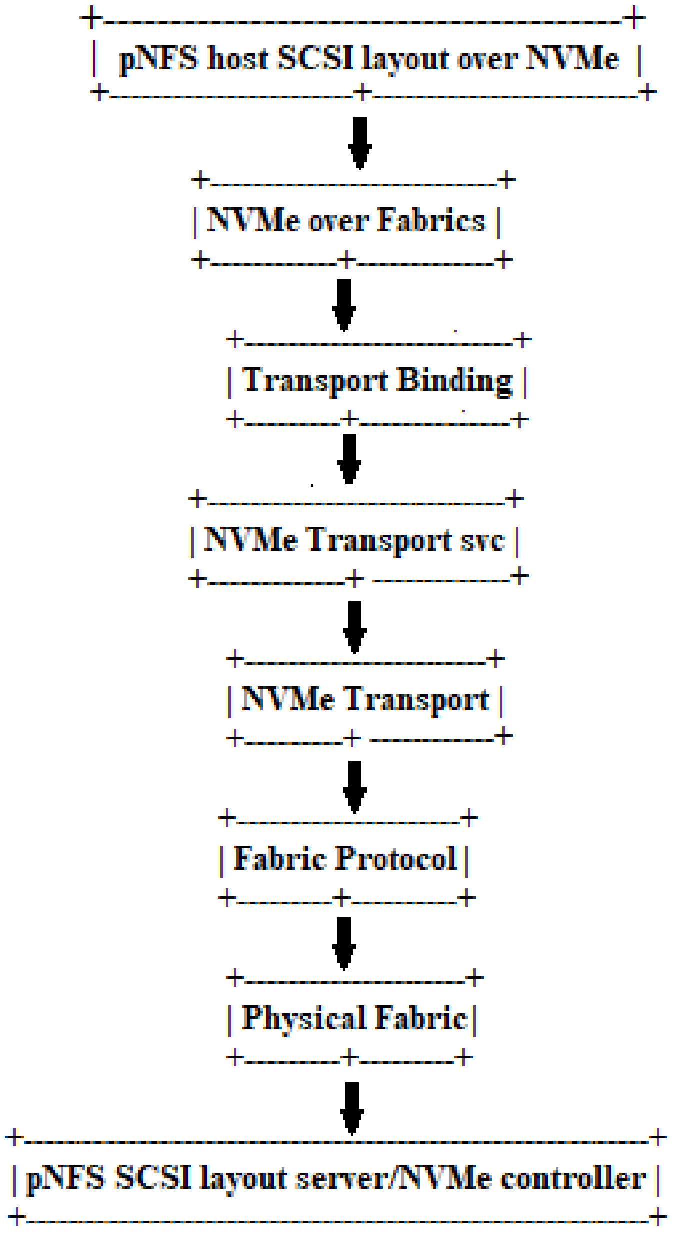
- How to extend current pNFS from SCSI to NVMe
  - New transports used to connect to NVMe-oF based servers (instead of SCSI)
  - NVMe-oF transports: Fibre Channel, RDMA (e.g., iWARP) and native TCP
  - pNFS needs NVMe layout support for these new NVMe-oF transports
- Start from pNFS SCSI layout (RFC 8154), extend to support NVMe
  - Draft introduces NVMe details for pNFS servers & clients

Expanding pNFS to support NVMe

- pNFS SCSI layout [RFC8154] allows pNFS clients to directly perform I/O to block storage devices bypassing the MDS (MetaData Server).
- This draft adapts the pNFS SCSI layout to enable use of NVMe-oF
  - Provides FC, RDMA or TCP access for devices using NVMe-oF
  - Enable implementers to start from the pNFS SCSI layout and the NVMe standards (currently NVMe-oF 1.1 and NVMe 1.4) to implement the pNFS NVMe layout
  - References the NVMe-oF Transport specifications for FC, RDMA and TCP

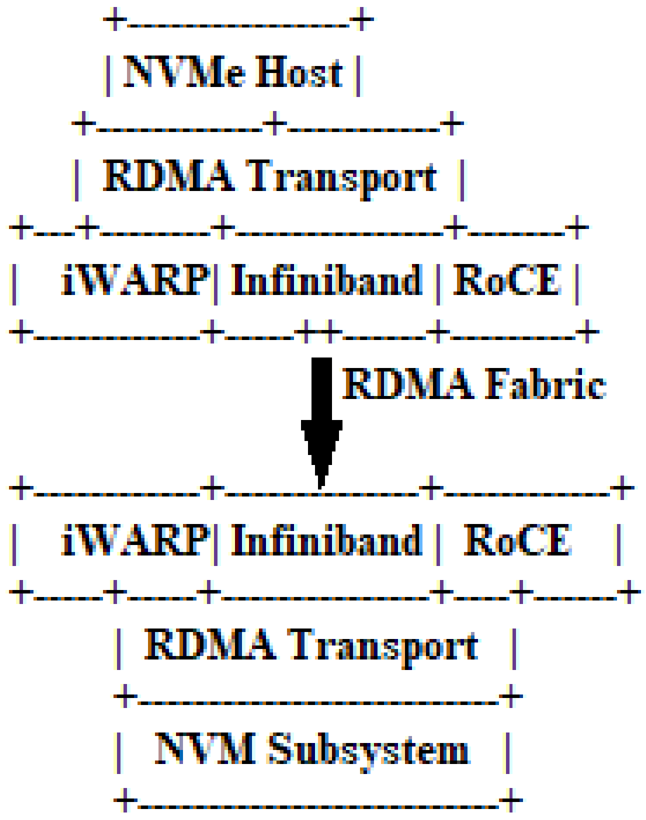
What is needed from pNFS server?

- Requires the pNFS storage devices to support the underlying NVMe-oF Transport to provide reliable NVMe command and data delivery
- NVMe over Fabrics architecture and commands used by pNFS clients to access pNFS storage devices.
- Layers shown in diagram



RDMA Transfer Protocol for pNFS/NFSv4.2

- NVMe port (PortID) supports multiple NVMe-oF Transports if more than one Transport is supported by the underlying network/fabric:
  - iWARP, RoCE or both
- The diagram illustrates the layering of the RDMA Transport and common RDMA providers (iWARP, InfiniBand™, and RoCE[v2]) within the host and NVM subsystem
- Separate TCP and FC transports not shown



Transfer Protocol for pNFS/NFSv4.2 (cont.)

- NVMe-oF allows multiple pNFS clients to connect to different controllers on the same subsystem (pNFS storage device)
- An association is established between a host and a controller when the host connects to a controller's Admin Queue
- The pNFS client also acts as a NVMe host and NVMe controllers are used as the pNFS storage devices.
- pNFS clients MAY connect to pNFS storage devices using different network protocols and different NVMe-oF transports.
- The NVM subsystem may require a host to use fabric secure channel, NVMe in-band authentication, or both.

Volume identification

- pNFS SCSI layout uses SCSI Device Identification VPD page to identify the devices used by a layout.
- NVMe-oF storage devices need to provide analogous unique identifiers based on EUI-64 and/or NGUID identifiers. Details to be worked out.
- UUID identification could be added but MUST use a large enough enum value to avoid conflict with possible future SCSI changes.

Client Fencing

- SCSI layout uses Persistent Reservations (PR) to provide client fencing.
- Both the MDS and the pNFS Clients have to register a key with the storage device, and the MDS has to create a reservation on the storage device.
- To allow fencing individual systems, each system MUST use a unique persistent reservation key.
- The MDS MUST generate a key for itself and a key for each pNFS client that accesses SCSI layout volumes before exporting a volume.
- The reservation key applies to all access by an individual pNFS client

Client Fencing – NVMe-oF

- NVMe Reservations: Similar to SCSI Persistent Reservations.
  - MDS Registration and Reservation
  - pNFS client registration
  - Multi-host reservation used: “Exclusive Access – All Registrants”
- Fencing actions:
  - MDS preempt a client’s registration to fence client (Reservation Acquire command)
  - Registration preemption removes client from reservation, hence denies access.
- Client Recovery after Fencing:
  - MUST commit all layouts
  - Future GETDEVICEINFO calls MAY require new pNFS client registration

Volatile Write Caches

- Carry SCSI layout volatile write cache support forward to NVMe
  - pNFS server required to commit cache to stable storage on Layout Commit
- NVMe: Flush command analogous to SCSI SYNCHRONIZE CACHE command
- Unrelated: (new) RDMA Flush extension at transport layer
  - Reason: RDMA operations are reversed from NVMe-oF commands
    - E.g., NVMe-oF Write command: RDMA Read pulls data from pNFS client.
  - Don’t need to flush NVMe-oF Read data (pushed to client via RDMA Write) to stable storage

Asks from NFSv4 WG

- Existing WG milestone: “use of NVMe in accessing a pNFS SCSI Layout”
  - Current date: August 2020
- Initial (-00) draft will be submitted sometime after this meeting
  - Would like that draft reviewed by WG members
  - Expect to ask that subsequent draft version be adopted for that WG milestone
- Will want to adjust date for that milestone
  - August 2020 completion appears unlikely
  - Suggest end of year, e.g., January 2021