Updating The NFS RDMA Standards

Chuck Lever, Oracle
Tom Talpey, Microsoft

Today's Purpose

- Confirm the NFS/RDMA specifications require attention
- Agree on a framework for updating these specifications
- Defer deep technical discussion

Useful Definitions

- A **bulk payload** is an RPC argument or result that is not immediately processed by the receiver, and is conveyed separately
  - The data portion of such an argument or result is eligible for RDMA transfer

- An **upper layer binding** is a set of rules that determine:
  - Which upper layer operations MAY send or receive bulk payload
  - Which RPC arguments or results MAY be considered bulk payload

Useful Definitions

- An **RPC/RDMA message** consists of
  - An RPC/RDMA header
  - An RPC header
  - An Upper Layer Protocol message

- The **inline** portion of an RPC/RDMA message is conveyed with RDMA SEND

- RPC/RDMA header represents bulk payload with **chunks**, which are conveyed separately

Useful Definitions

- With an **RDMA_MSG** proc, only chunks are moved via RDMA transfer; everything else is inline and moved via RDMA SEND

- With an **RDMA_NOMSG** proc, the RPC header, arguments or results, and ULP message are all moved via RDMA transfer

Existing Documents

- RFC 5666: “Remote Direct Memory Access Transport for Remote Procedure Call” (2010)

- RFC 5667: “Network File System (NFS) Direct Data Placement” (2010)

- Clustered with RFCs 5661 - 5665

RFC 5666: Needed Clarifications

- Restrictions on RDMA_NOMSG
- Padding requirements when there is inline content following a read chunk
- How to handle multiple bulk payloads in a single RPC

RFC 5666: Potential Enhancements

- In-band receive buffer size negotiation
- Bi-directional RPC
- Remote Memory Region invalidation
- Message chaining

RFC 5666: Strategies

- Clarifications only:
  1. Use errata process, and/or
  2. Create a normative “Updates” document

- Any extension or incompatible protocol change requires:
  - An RPC/RDMA protocol version bump
  - A new normative document

RFC 5667: Proposed Updates

- Section 5 (NFSv4.0/NFSv4.1) is inadequate
  - Is an operation in an NFSv4 COMPOUND an RPC argument?
    - No SYMLINK operation in NFSv4
- No upper layer binding is provided for pNFS operations
- NFSv4.2 introduces operations that could be eligible for RDMA (*e.g.* READ_PLUS)

RFC 5667: Strategies

1. Repair RFC 5667:
   - Leave NFSv2 and NFSv3 upper layer binding as-is
   - Move NFSv4.0 and NFSv4.1 upper layer binding to a normative “Updates” document

2. Replace RFC 5667:
   - Create a normative “Obsoletes” document that copies RFC 5667 with corrections and replaces section 5 outright

Areas Not Covered by Existing Documents

RPC/RDMA with RPCSEC GSS

- Likely no issues with strong authentication
- How are bulk payloads handled when using integrity checking or encryption?
  - Current Solaris implementation uses RDMA_NOMSG

NFSv4.2 Upper Layer Binding

- At least READ_PLUS operation needs discussion
- One of the following could be used:
  1. Cover NFSv4.2 upper layer binding in the new NFSv4.0 and NFSv4.1 document
  2. Add NFSv4.2 upper layer binding to existing NFSv4.2 draft specification
  3. Cover NFSv4.2 upper layer binding in separate new document

Existing pNFS Layouts

- pNFS DS operations include:
  - READ and WRITE operations
  - Covered elsewhere (NFSv4.1, SCSI, *etc.*)

- pNFS MDS operations include:
  - Callbacks and layout-related operations
  - Large MDS operations need an upper layer binding

Additional Layout Types

- Transitional block-over-RDMA technologies
  - iSER
  - SRP
- Persistent memory technologies
  - NVMe on Fabrics
  - RDMA targeting byte-addressable persistent memory

Discussion and Hum