RPC-over-RDMA version two

2022 Progress Report

Chuck Lever <chuck.lever@oracle.com>

Linux Prototype

- Based on the existing v1 implementation
- Client and server
- Limited: only v1 credit accounting; no transport properties, peer authentication, or new error codes
- In other words, no change since 2020

Flow Control Updates

- rpcrdma-version-two-07 Section 4.2.1 has been rewritten with help from Jana Igeyar:
  - A classic credit-based flow control mechanism is described that provides full-duplex management of peer Receive buffers
  - This enables message chains of unlimited size, asynchronous credit grants, and in-band control messages, and it enables rarely-used RPC mechanisms such as Call-only transactions

RPC-over-RDMA transport layer security

- We anticipate that the IESG might make transport layer security a requirement for a new version of the RPC-over-RDMA protocol
- A mechanism to exchange authentication material has been proposed as a transport property, but I would prefer:
  - The use of a well-established Internet building block such as TLSv1.3
  - An RDMA transport level solution analogous to TLS with TCP

RPC-over-RDMA transport layer security

- The IETF has no purview over RoCE, InfiniBand, and others, but does have authority over MPA/DDP, formerly known as iWARP
- Section 5.4.2 of RFC 5042 considers the use of TLS under DDP/RDMAP and rejects it. Section 5.4.3 of RFC 5042 proposes the use of IPsec or DTLS as a transport below a TCP-like layer which would then convey RDDP on top of that. The reasoning for this complexity is unconvincing.
- Perhaps it is time to consider a simple specification of DDP/RDMAP on QUIC.

NFSv4.2 READ_PLUS

- Because an NFS client cannot predict the content of the returned segment list, it must register a Reply chunk and parse the returned list. This guarantees that direct data placement cannot be used.
- The NFS/RDMA Upper Layer Bindings therefore do not allow any READ_PLUS result data item to be eligible for DDP.
- However, READ_PLUS is required to handle large sparse files efficiently: they avoid transmission of large ranges of zero bytes, and help server filesystems avoid hole instantiation on read.
- Possible action: a brief document extending either RFC 8166 or 8267

WG Bureaucratic Actions

- The performance benefits of v2 are met (in practice) with RFC 8797 and the upcoming pNFS-NVMe layout type
- Evaluate the priority of work on rpcrdma-version-two based on:
  - Current number of RPC/RDMA v2 prototypes
  - Other projects in front of the WG (*i.e.*, rfc5661bis, TLS, etc)
  - Available prototyping, authorship, review, and stewardship resources

WG Bureaucratic Actions

- Remove the milestone for delivery of rpcrdma-version-two and nfs-ulb-v2

Supplemental Material

Bibliography

- RFC 8166 - RPC over an RDMA Transport
- RFC 8797 - Remote Direct Memory Access - Connection Manager (RDMA-CM) Private Data for RPC-over-RDMA Version 1
- https://datatracker.ietf.org/doc/draft-ietf-nfsv4-rpcrdma-version-two
- https://datatracker.ietf.org/doc/draft-ietf-nfsv4-nfs-ulb-v2