SDC

STORAGE DEVELOPER CONFERENCE

SNIA ■ SANTA CLARA, 2015

pNFS/RDMA: Possibilities

Chuck Lever
Oracle Corporation

The opinions expressed in this presentation are the presenter's own, and do not represent the views of Oracle or anyone else.

What If . . . ?

- Given these storage trends:
  - Throughput of networks is increasing
  - Latency of persistent storage is dropping exponentially
  - Capacity is off the charts

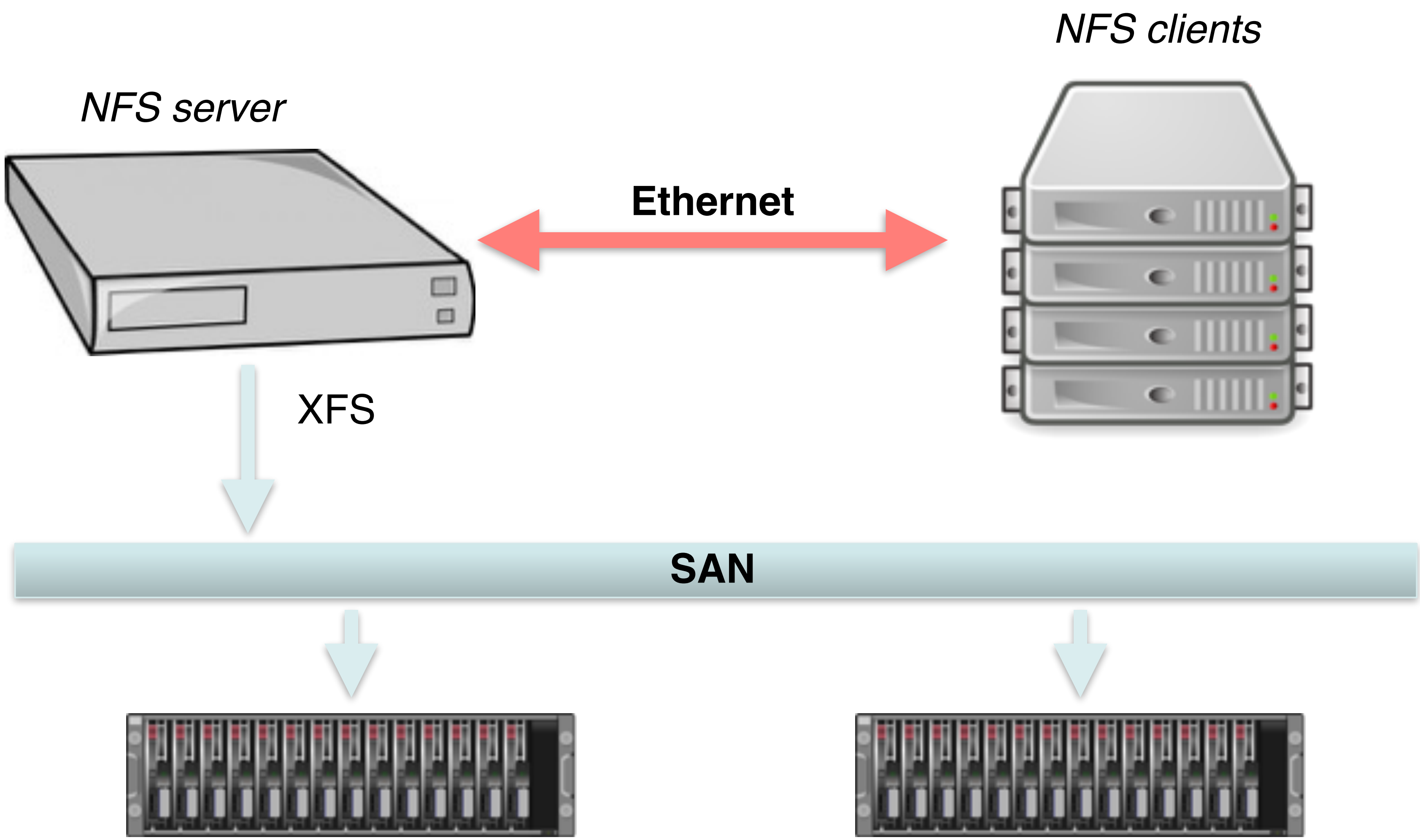
- How can NFS make good use of our new Persistent Memory overlords?

Traditional NFS

Traditional NFS Operation

- Each NFS file resides on one server
- Applications locate files via a POSIX directory structure
- Clients access data via NFS READ and WRITE operations

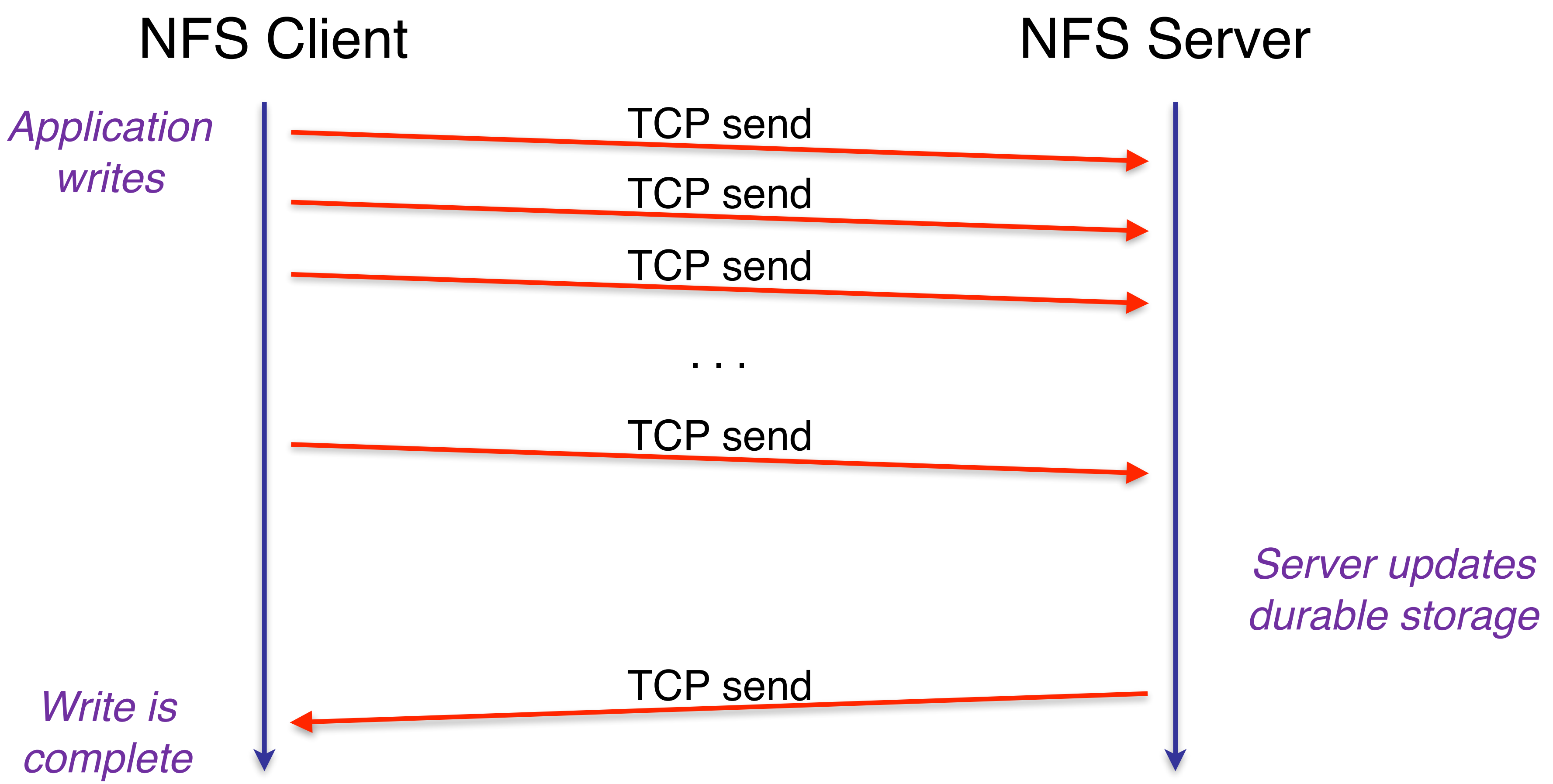
Traditional NFS Server Storage Topology



Traditional NFS Weaknesses

- One RPC issued at a time per TCP socket
- Typically one or a few TCP sockets are shared across a server's shares
- Data throughput is constrained by the server

Traditional NFS FILE_SYNC WRITE



Two-phase Commit

- To avoid waiting for durable storage on every WRITE, NFSv3 introduced unstable WRITE plus COMMIT
  - Client flushes data to server asynchronously
  - Client sends COMMIT
  - Server makes written data durable
- Transport bottlenecks remained

What Is pNFS?

Data / Metadata Separation

- NFS protocol manages metadata
  - Directory structure
  - File open and lock state
  - File data layout information
  - Fall-back I/O mechanism

- Separate protocol and transports handle I/O

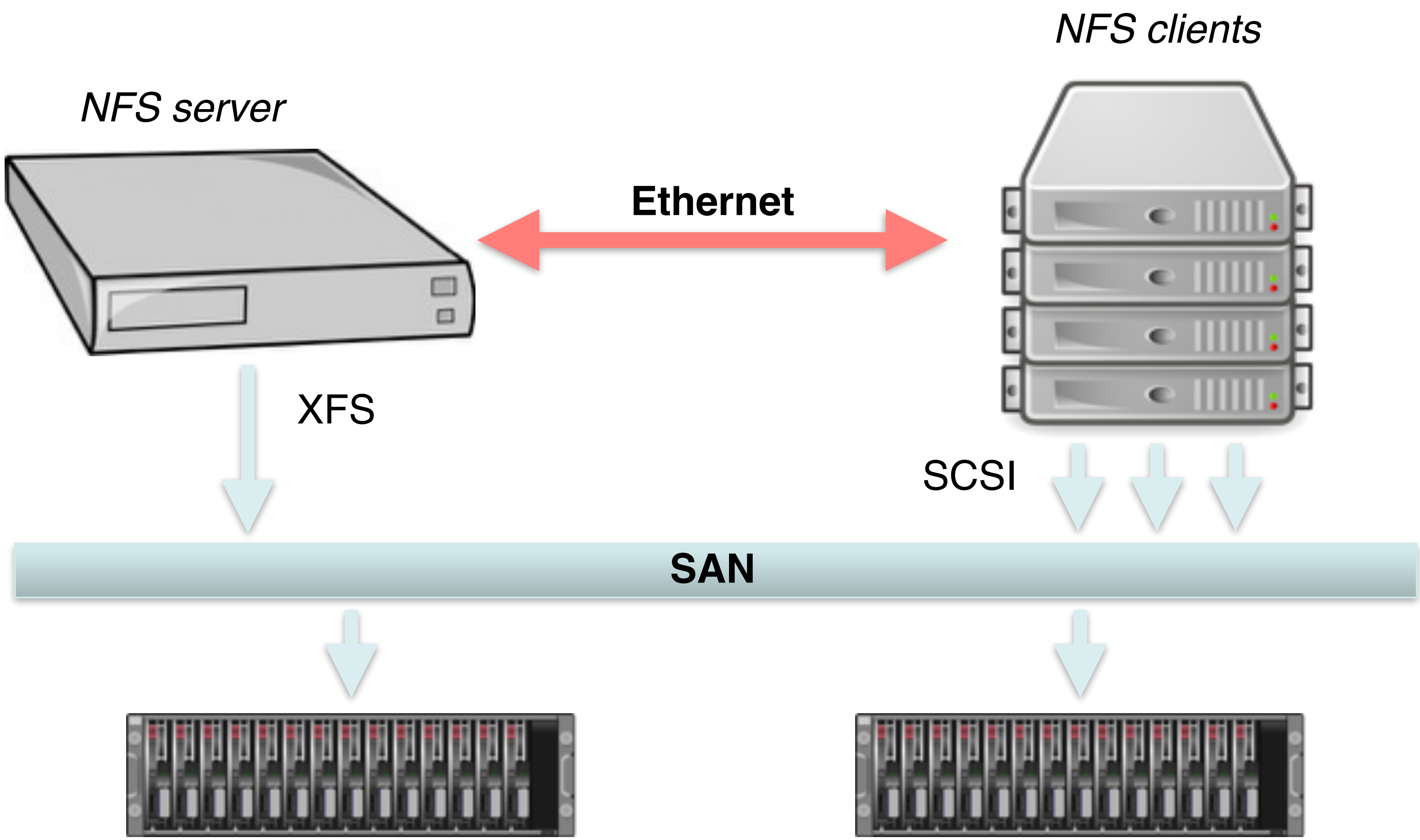
pNFS Layout Types

- A layout type:
  - Specifies which transport protocol to use
  - How to locate file data
  - Specified separately from NFS protocol
- A layout instance tells where a file's data resides
  - Which NFS server and file, or
  - Which SCSI LUN at which LBA

Parallel NFS In A Nutshell

- Applications retain single-server view of files
- NFS server manages data layout
- Each NFS client can stripe file I/O across multiple storage services
- Data and metadata operations run concurrently
- Clients and servers share a storage fabric
  - SCSI, iSCSI, iSER, SRP
  - Object-based storage
  - NFS

pNFS Server Storage Topology



Example Usage Scenarios

- High Performance Computing
  - Parallel I/O
  - Greater file capacity
- Deployments where storage clients and servers share a storage fabric
  - Each client can be directed to a particular server
  - Each file can be placed on a particular server

What Is NFS/RDMA?

What Is Remote Direct Memory Access?

- I/O-like access of the physical memory on another host
  - Strong ordering of operations
  - Asynchronous: completion fires when an operation finishes
  - Datagram channel: SEND and RECV
  - Data transfer: READ and WRITE

RDMA Ready For 100Gbps Fabrics

- Zero-copy is possible on both send *and* receive
  - No CPU cache footprint until app accesses data
- Transport resources are pre-allocated
  - No resource allocation in data path
  - Reduced opportunity for deadlock
- Data transfer is concurrent with other transport operations

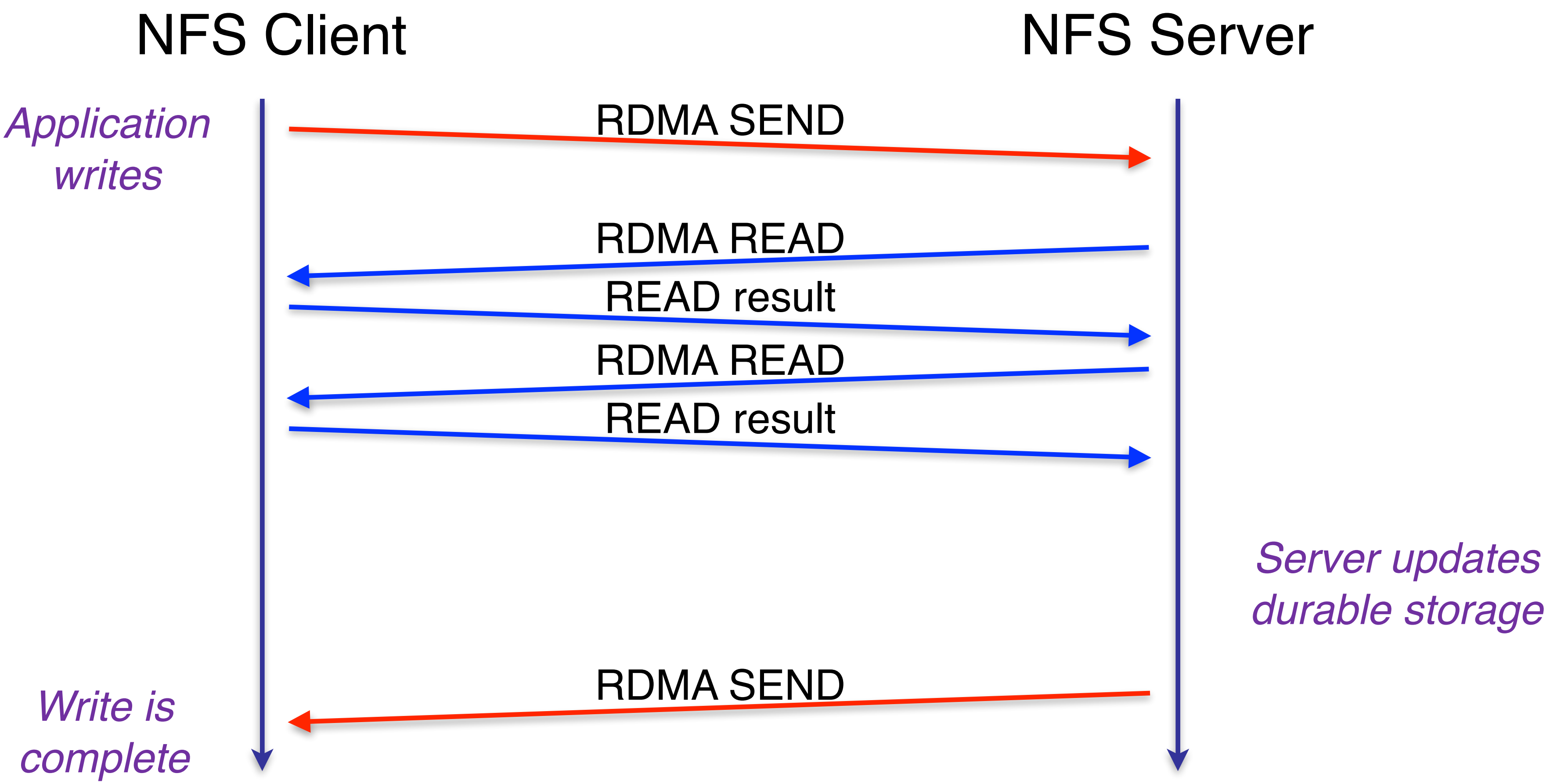
NFS/RDMA Concepts

- Each RPC is conveyed by RDMA operations
  - Ultra-low round-trip latency
- RNICs handle bulk data transfer
  - Low CPU overhead
  - High bandwidth

Data / Metadata Separation

- Non-I/O operations conveyed via RDMA SEND
  - GETATTR, LOOKUP, and so on
- Data operations (*i.e.* NFS READ and WRITE) utilize RDMA READ and WRITE
  - Server initiates all RDMA transfer
  - After that, neither host CPU is involved

NFS/RDMA FILE_SYNC WRITE



Example Usage Scenarios

- Use NFS/RDMA instead of NFS/TCP on IPoIB
  - See “RDMA On 100Gbps Fabrics”
- Latency-sensitive SLAs
- CPU-intensive client workloads
- One-time bulk-data movement (*e.g.* backup)

pNFS and NFS/RDMA

Why pNFS/RDMA?

- Client gets **direct** access to durable storage
  - *E.g.* ultra-low latency Persistent Memory
  - No protocol translation overhead
  - Data not even read into server DRAM

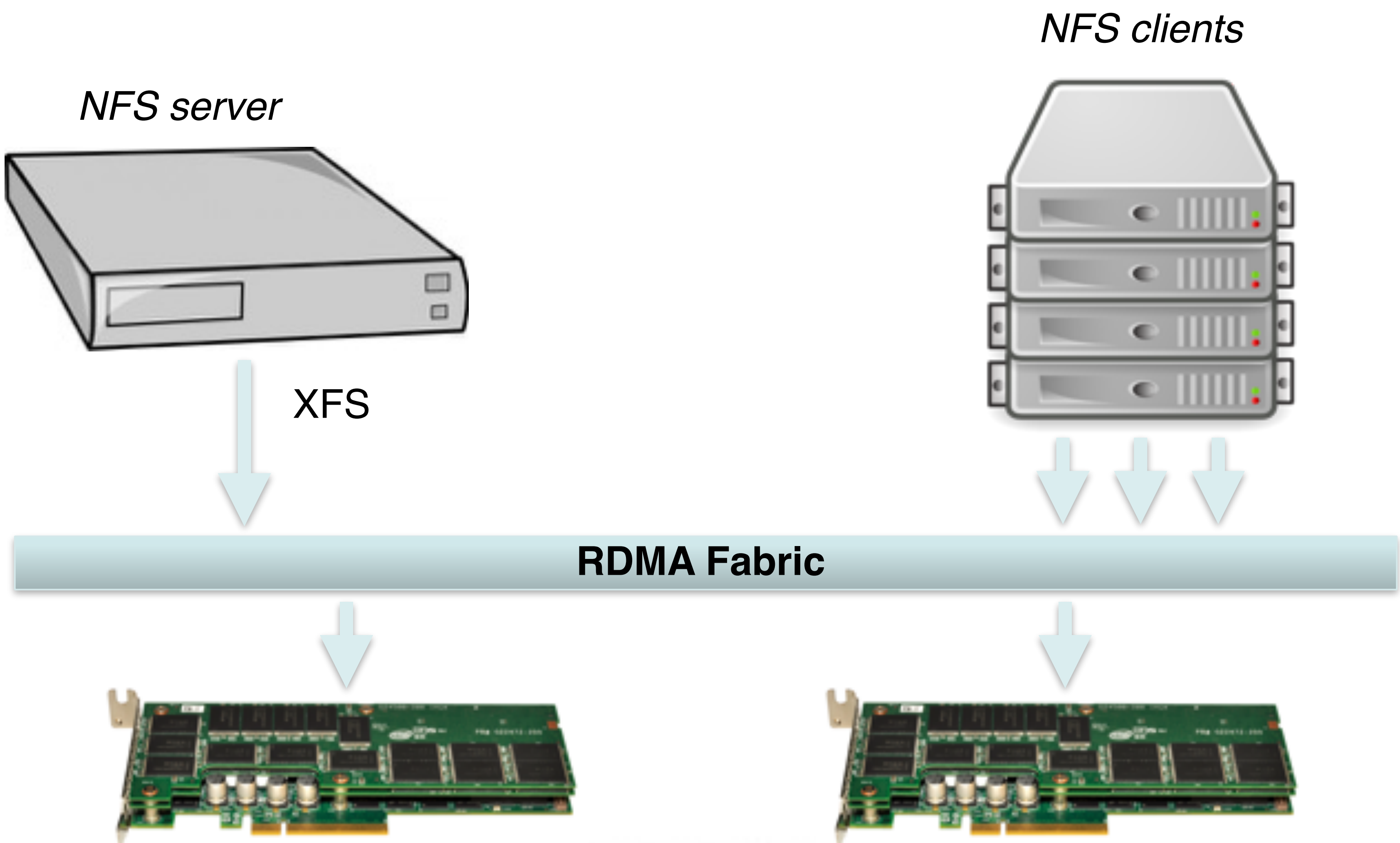
Why pNFS/RDMA?

- Multiple transport connections per client mount point
  - Multiple QPs
  - Multiple RNICs

Why pNFS/RDMA?

- Single converged fabric shared between pNFS clients and servers
  - Rather than “pNFS/TCP with SCSI”
  - Instead use “pNFS/RDMA with SRP”

pNFS/RDMA Server Storage Topology



Next Steps

What's Needed For NFS/RDMA

- NFSv4.1 on RDMA is a pre-requisite
  - Bi-directional RPC-over-RDMA
  - Lots of backchannel session slots
  - NFSv4.1 Upper Layer Binding to RPC-over-RDMA

What's Needed For pNFS

- A new pNFS layout type is not required for operation with SRP or iSER

- Proposal: a new pNFS layout type for accessing remote Persistent Memory devices directly
  - Device naming
  - Ensuring data durability
  - Error handling and fencing
  - Authentication, data privacy

Questions / Discussion

Appendix

NFS Reference Material

- pNFS Standards
  - NFSv4.1: RFC 5661
  - pNFS layouts: RFCs 5662 - 5665
- NFS/RDMA Standards
  - RPC-over-RDMA: RFC 5666
  - NFS/RDMA ULB: RFC 5667