Experiences with NFS over IB and iWARP RDMA

OpenFabric Workshop, Sonoma, CA

May 1, 2007

Helen Chen, Noah Fischer, Matt Leininger, and Mitch Williams
Sandia National Laboratories
SAND 2007 – 2137C

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

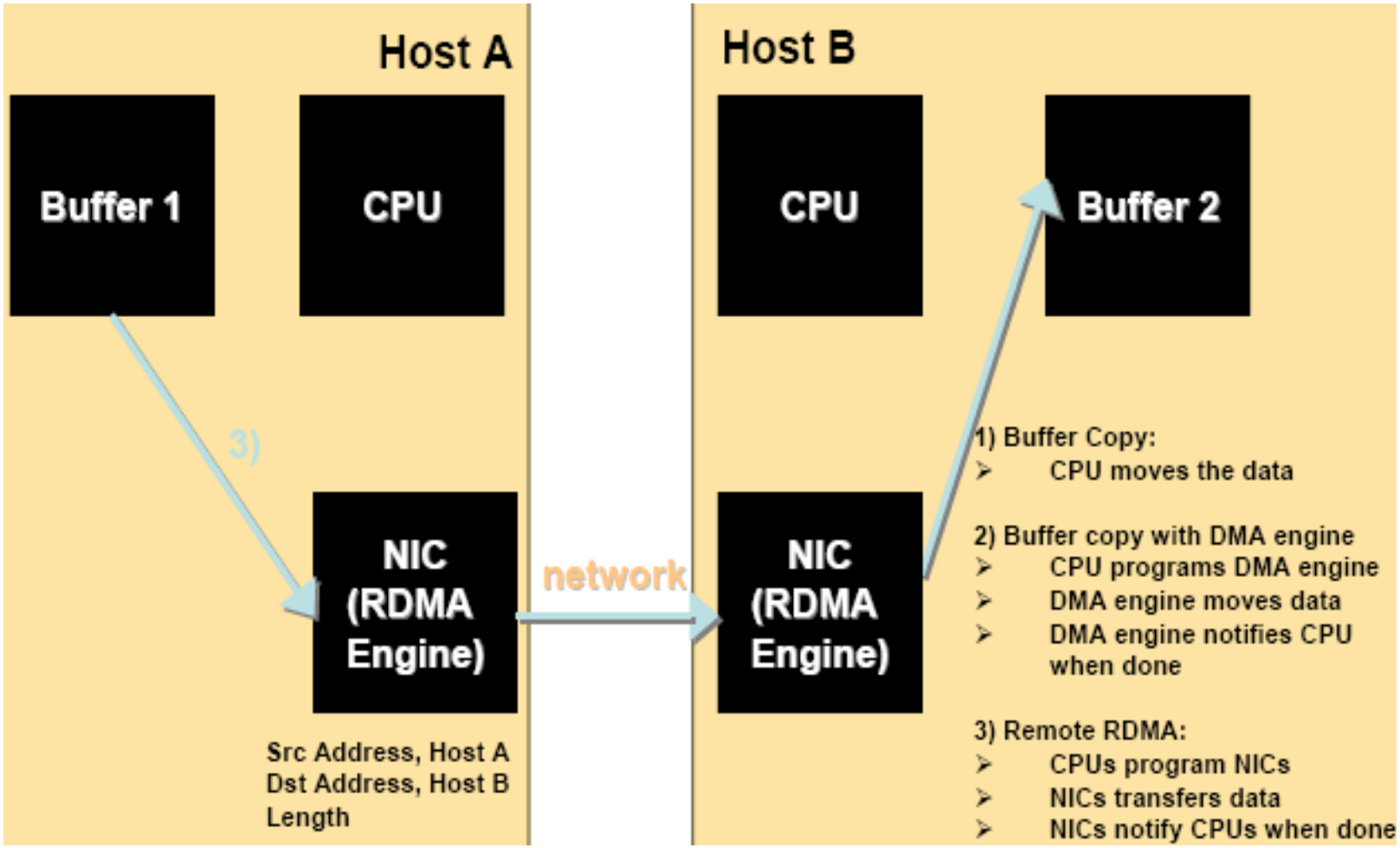
Outline

- Motivation
- Previous Study – NFS over RDMA (SDR IB)
- This Study – extends the previous study to include DDR IB and 10 GbE iWARP
  - The Testbed
  - The Benchmark
  - Results and Analysis
- Summary and Future Plans
  - The parallel NFS research collaboration with Open Grid Computing

Motivation

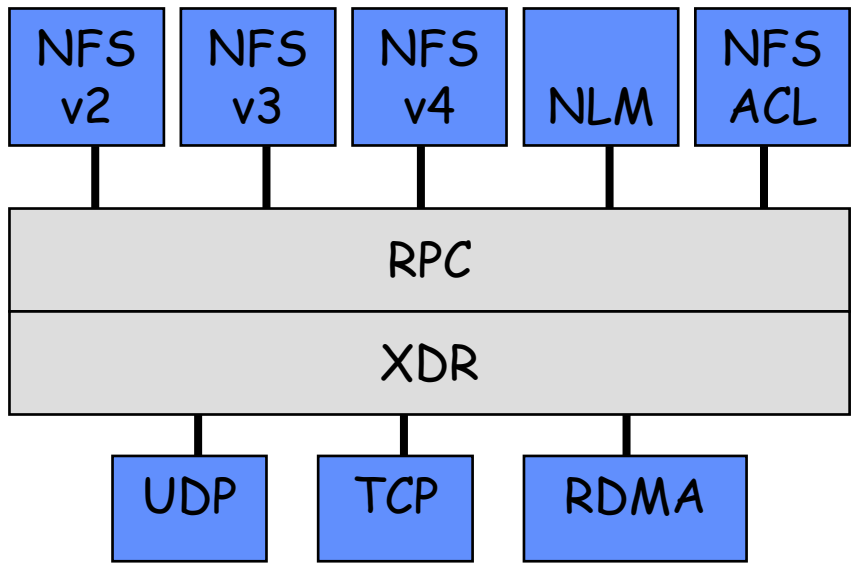
- **Scaling I/O for Commodity Clusters**
  - **While multi-core processor technology speeds ahead, filesystem capability is falling far behind.**
  - **Panasas, Lustre, and GPFS are being developed outside of the Linux main stream, and they are complex to administer**
  - **The Linux mainstream distributed filesystem, NFS, is slowly being improved in functionality (NFSv4, NFS-over-RDMA, parallel-NFS),**

How RDMA Works



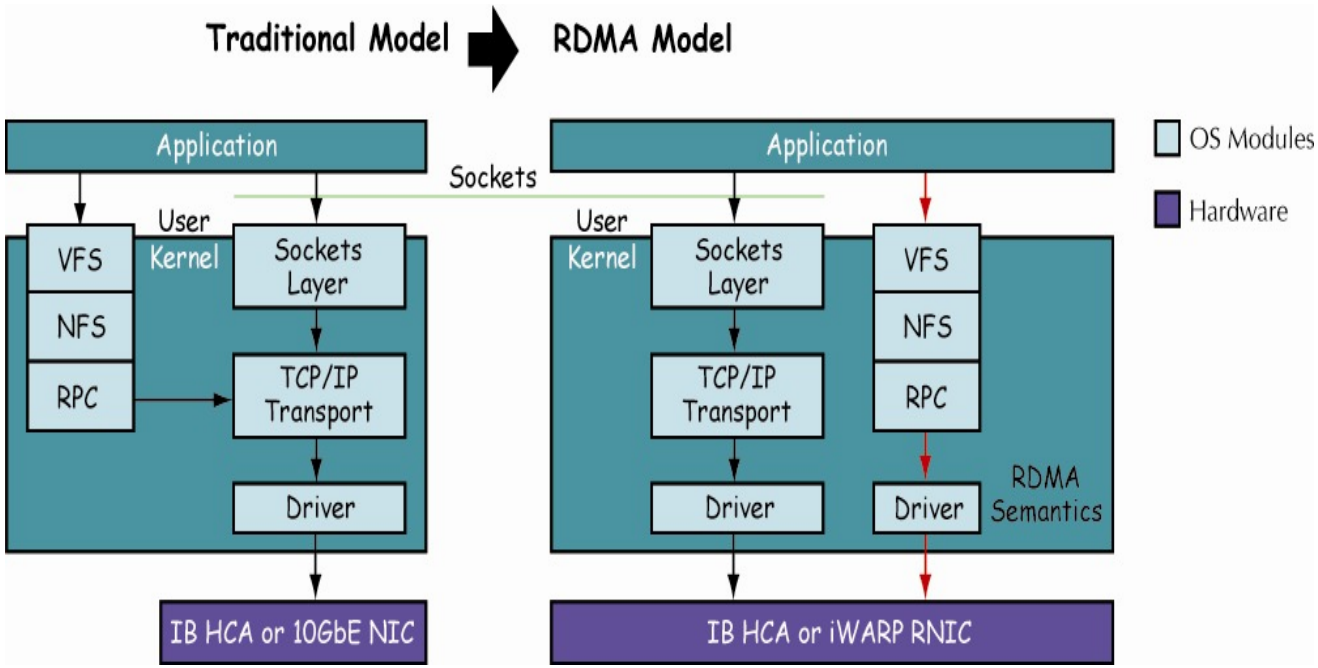
The NFS RDMA Architecture

- **NFS is a family of protocol layered over RPC**
- **XDR encodes RPC requests and results onto RPC transports**
- **NFS RDMA is implemented as a new RPC transport mechanism**
- **Selection of transport is an NFS mount option**



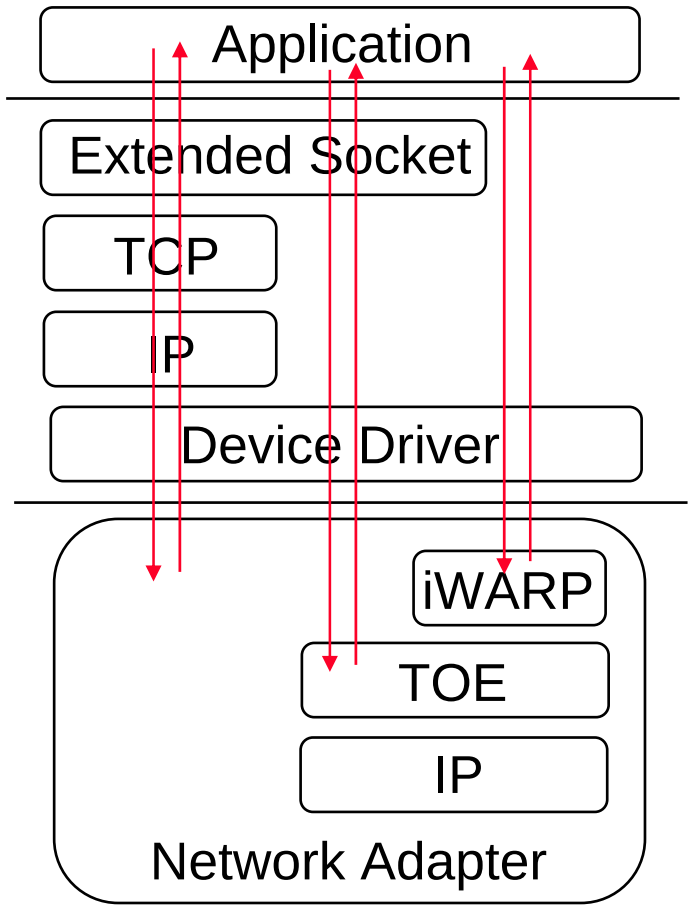
Brent Callaghan, Theresa Lingutla-Raj, Alex Chiu, Peter Staubach, Omer Asad, "NFS over RDMA", ACM SIGCOMM 2003 Workshops, August 25-27, 2003

The NFS Protocol Stack



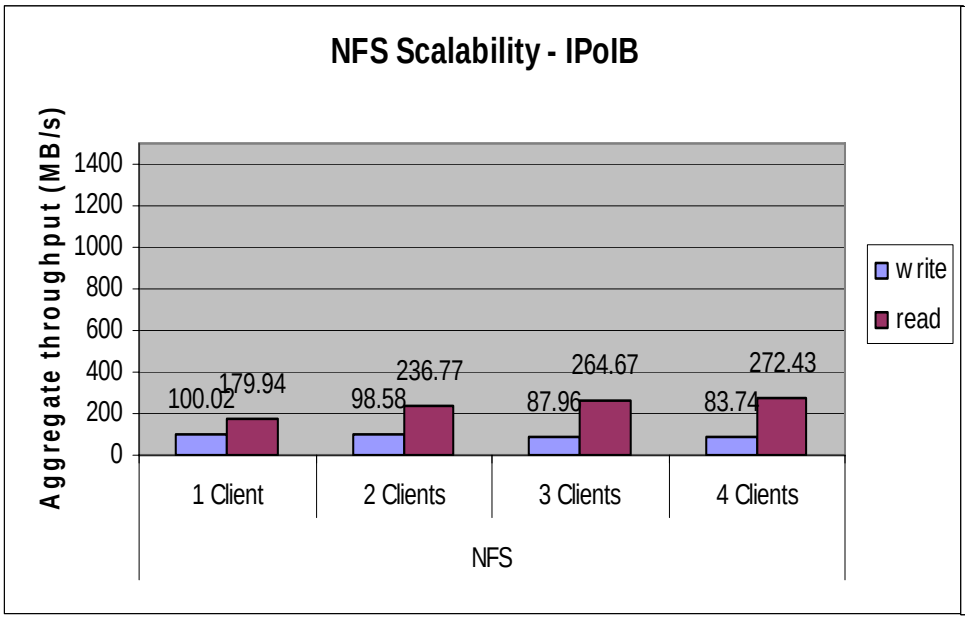
iWARP - RDMA protocol for TCP/IP

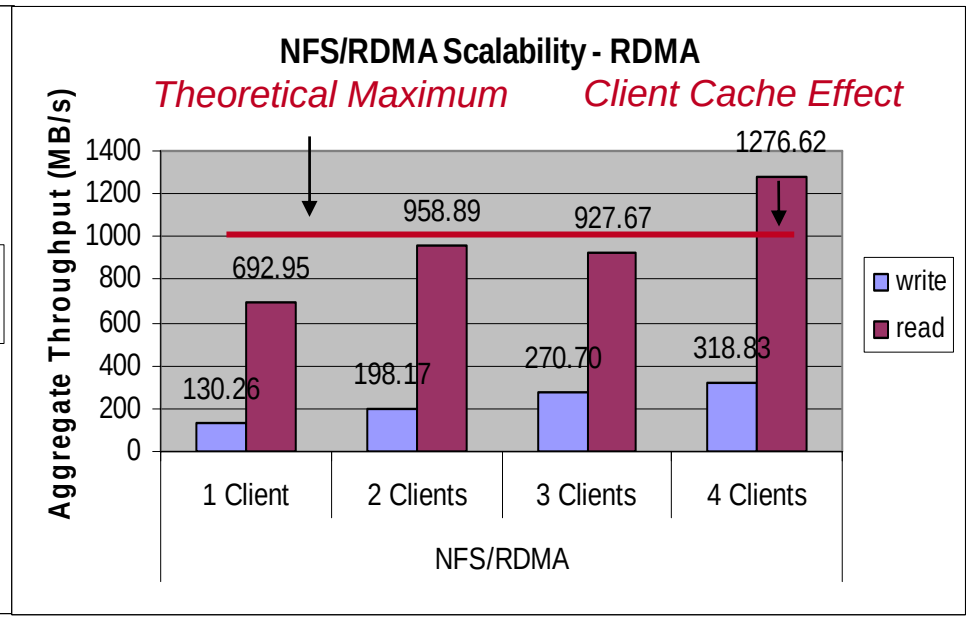
- iWARP is the set of RDMA protocols for TCP/IP
- RNIC is a RDMA capable NIC with offloaded iWARP as well as TCP/IP (TOE)
- RNIC typically exposes NIC, TOE and iWARP interfaces to upper layer applications



Previous Study – NFS over IB RDMA vs. TCP (IPoIB)

- **NFS over RDMA can easily fill the 10 Gigabit (1GB) pipe**
  Iozone -i 0 –I 1 –r 64k –s 2g





http://www.openfabrics.org/archives/sep2006devcon.htm

H

h

Sandia National Laboratories

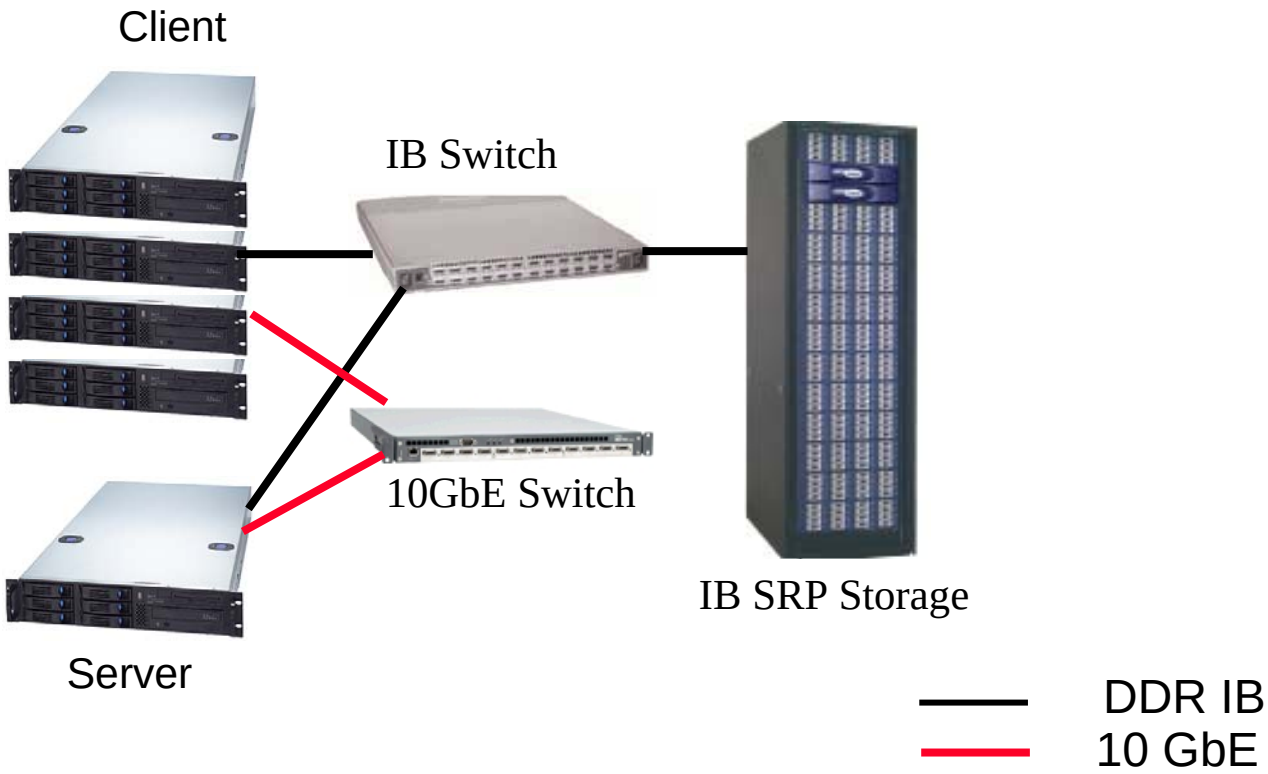
Write Performance Issues

- Linux's NFS client implementation lack concurrency
  - Pdflush activated when dirty page cache reached 34%
  - Application I/O's blocked while cached data being flushed
  - Most visible with RDMA due to huge bandwidth capacity and CPU efficiency
- Being addressed by the Linux kernel community (Talpey, Tucker, et. al.)
  - Multi-threading support to pipeline application I/O

This study evaluates NFS RDMA transport vs. TCP, using Iozone reading from server cache

Sandia National Laboratories

This Study –DDR IB and iWARP RPC Transport with SRP IB Storage



Key Testbed Hardware

- **Mainboard: iWILL DK8ES**
  - **Dual Core Dual Socket 2.4 Ghz AMD Opteron**
  - **Dual Channel 400 Registered memory**
    - **4 GB on server**
    - **2 GB on client**
- **DDR IB Switch: Mellanox InfiniScale III 24-port switch**
- **DDR IB HCA: PCI-E Mellanox MT25204 InfiniHost III Lx**
- **10 GbE Switch: Fujitsu XG700 CX4**
- **10 GbE RNIC: PCI-X and PCI-E Chelsio Terminator 3**
- **SRP SDP IB Storage: DDN S2A 9550**

Key Testbed Software

- RedHat Enterprise AS Release 4 Update 4
- Kernel: Linux 2.6.18.8
  - http://kernel.org
- NFS/RDMA update 7
  - http://sourceforge.net/projects/nfs-rdma/
- oneSIS used to boot all the nodes
  - http://www.oneSIS.org
- OpenFabric Enterprise Distribution 1.2 Beta
  - IB, iWARP, SRP, etc.
  - http://www.openfabrics.org

Sandia National Laboratories

NFS Test Configuration

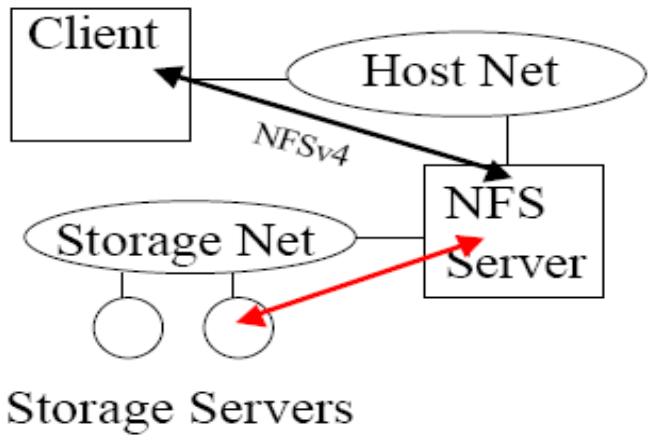
- One NFS server and one to four clients
- Ext2 filesystem built on IB SRP Storage at SDR
- TCP/IPoIB (MTU 2048), TCP/IPoIB-CM (MTU 65520), and IB RDMA transport at DDR
- Host TCP/IP and RNIC (iWARP) transport at 10GbE rate (MTU 9000)
- Clients ran IOZONE reading 128KB records
- Write and read 2GB file on all clients to avoid client-side cache effect and server-side disk I/O

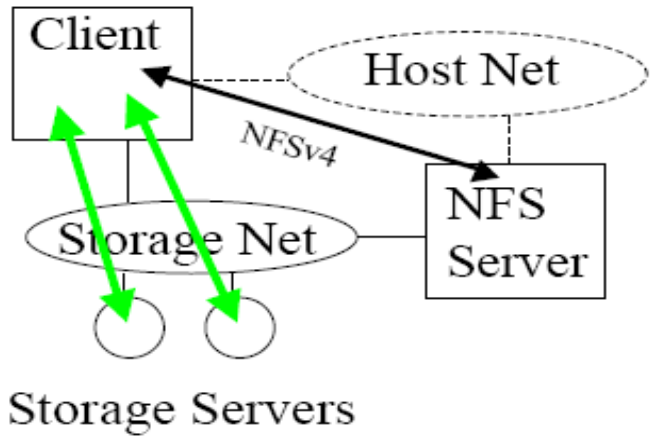
To allow evaluation of NFS RDMA transport

- System resources monitored using “vmstat” at one second intervals
- All tests repeated 10 times

Sandia National Laboratories

The pNFS Architecture



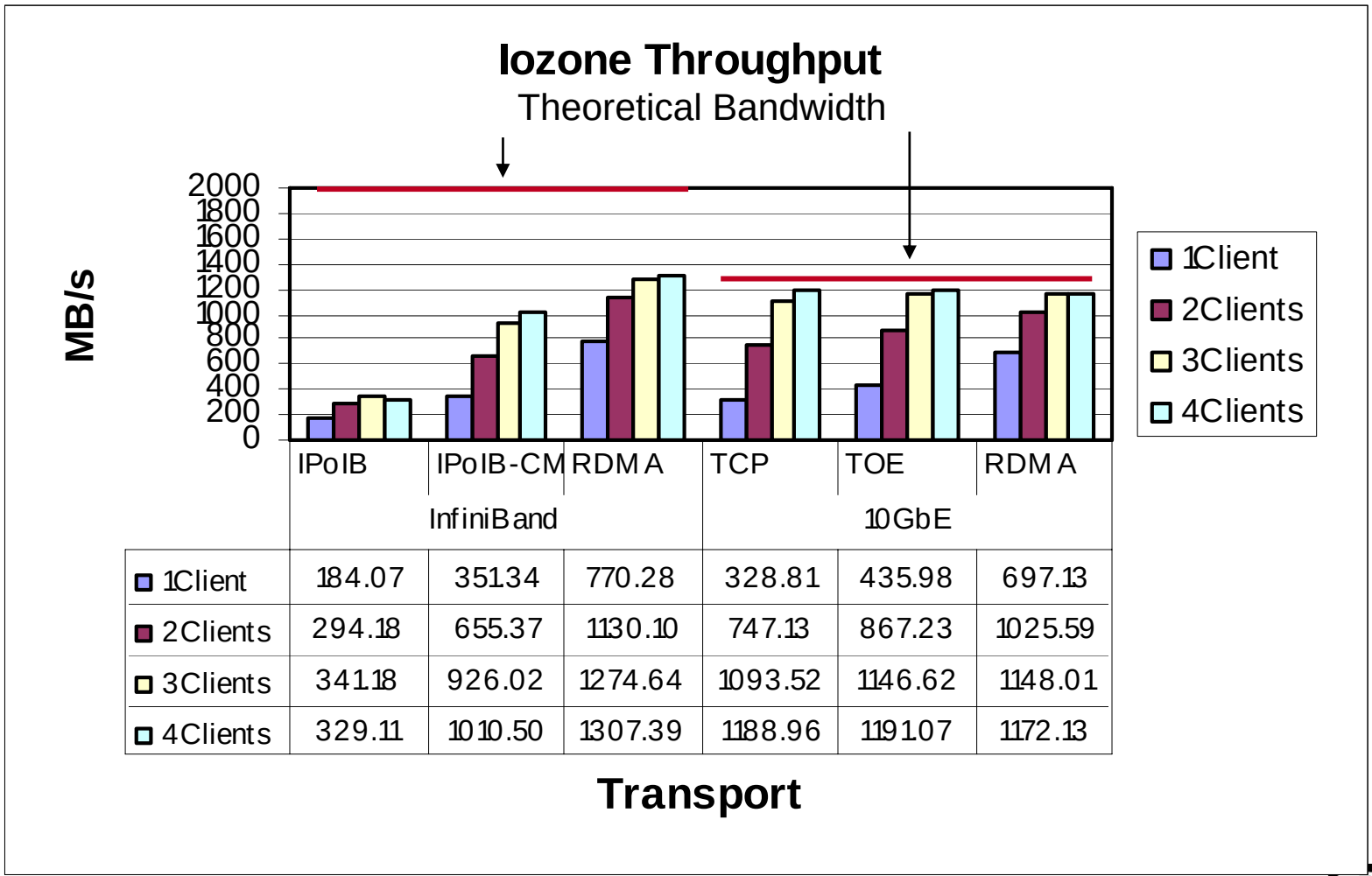


- **pNFS extends NFSv4**
  - **To allow out-of-band I/O**
  - **A Standards-based scalable I/O solution**
- **Asymmetric, Out-of-band solutions offer scalability**
  - **Control path (open/close) different from Data Path (read/write)**

http://www3.ietf.org/proceedings/04nov/slides/nfsv4-8/pnfs-reqs-ietf61.ppt

Sandia National Laboratories

NFS Throughput



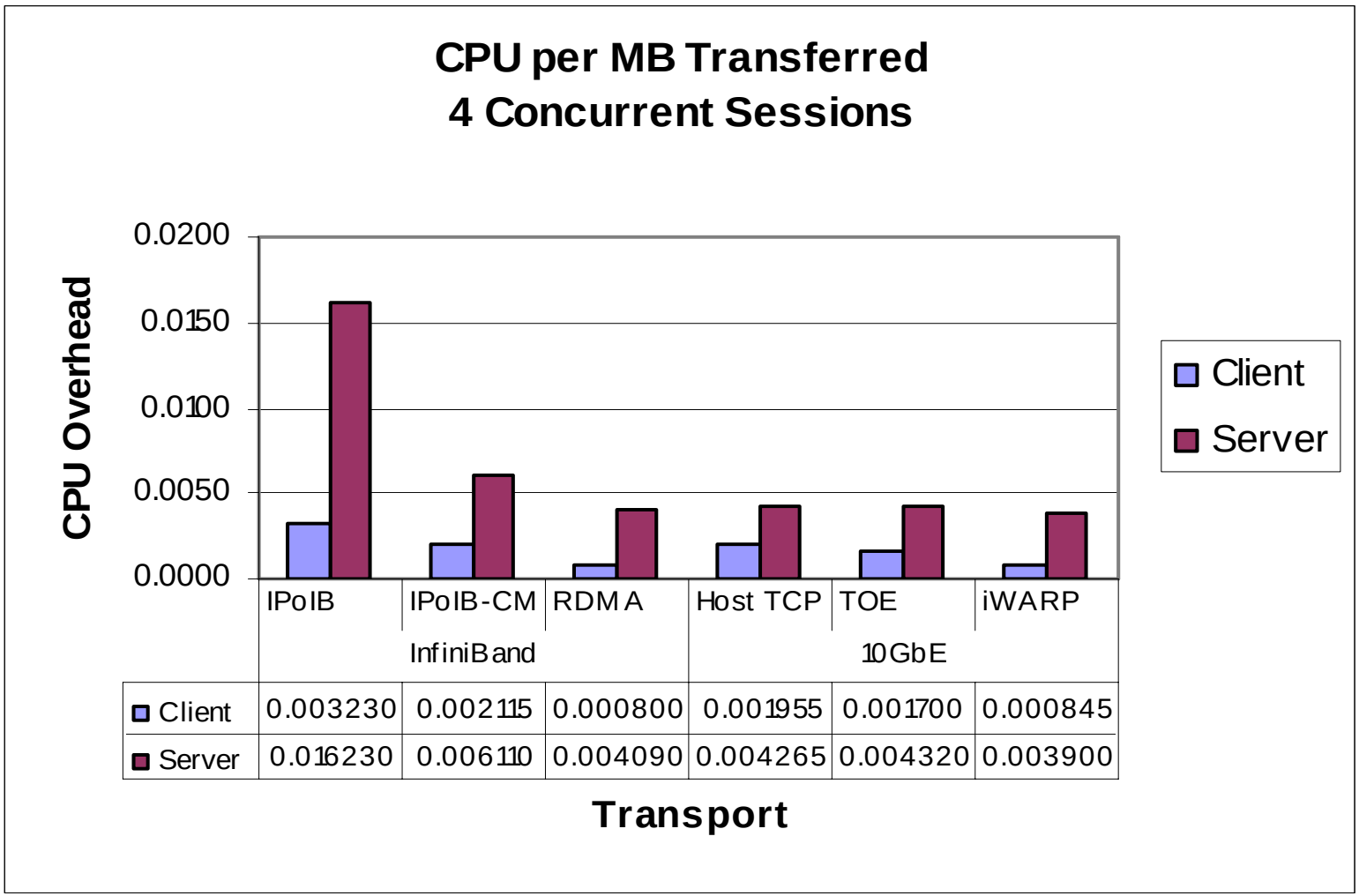
	IPoIB	IPoIB-CM	RDMA	TCP	TOE	RDMA
1Client	184.07	351.34	770.28	328.81	435.98	697.13
2Clients	294.18	655.37	1130.10	747.13	867.23	1025.59
3Clients	341.18	926.02	1274.64	1093.52	1146.62	1148.01
4Clients	329.11	1010.50	1307.39	1188.96	1191.07	1172.13

Sandia National Laboratories

NFS Throughput Summary

- NFS over RDMA
  - Can take advantage of the IB DDR pipe (theoretical maximum bandwidth 2.0GB)
  - Throughput is limited by the 10GbE rate (theoretical maximum bandwidth 1.25GB)
  - Both RDMA transport out performed their TCP counter part, most noticeably in the case of IB
- NFS over TCP
  - IPoIB-CM significantly better than IPoIB-UD
    - 65520 MTU concern – fragmentation at IB-GE gateway
  - 10GbE TCP, both RNIC's TOE and host stack, performed surprisingly well
  - Can easily filled the 10GbE pipe as well
  - A great all-in-one adapter

NFS CPU Efficiency



	IPoIB	IPoIB-CM	RDMA	Host TCP	TOE	iWARP
Client	0.003230	0.002115	0.000800	0.001955	0.001700	0.000845
Server	0.016230	0.006110	0.004090	0.004265	0.004320	0.003900

Sandia National Laboratories

NFS CPU Efficiency Summary

- **Host Efficiency is based on CPU per MB transferred**

  $$\sum \%cpu / 100 / file\text{-}size$$

- **IB RDMA and 10GbE iWARP delivered comparable CPU efficiency**
- **RDMA demonstrated better CPU performance than TCP**
  - **Most significant in IB**
  - **Both TOE and host TCP performed extremely well, with TOE better than host stack**

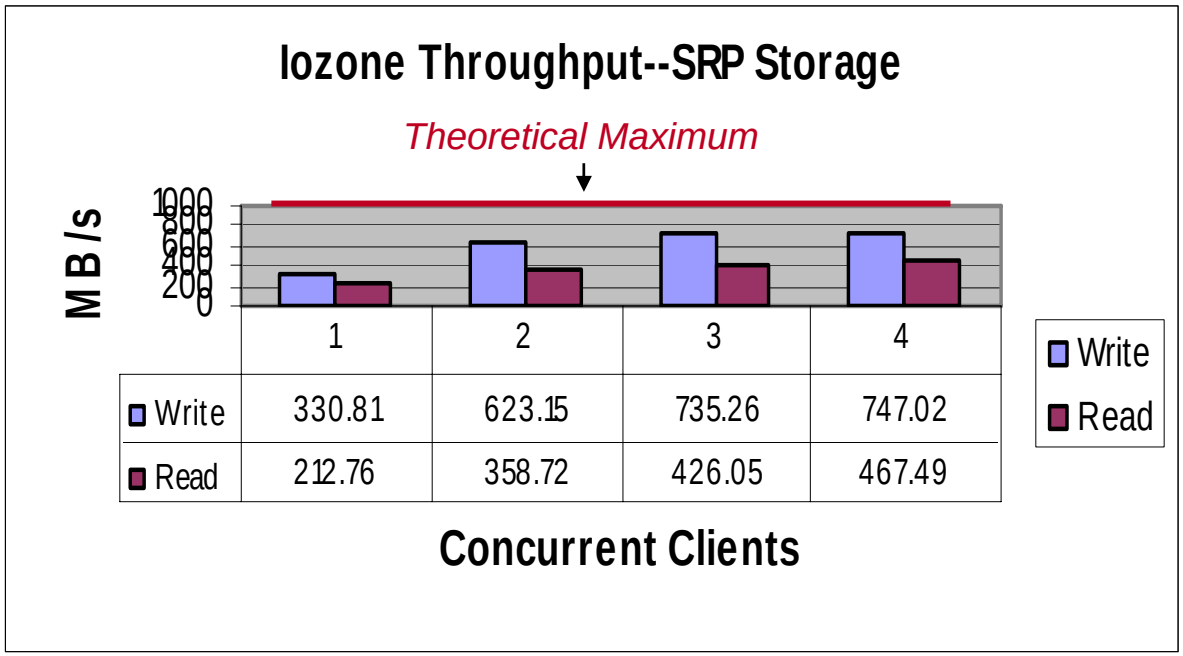
SRP Target and Initiator Configuration

- Target – DDN S2A9550
  - 1 Controller, 1 SDR IB link
  - 4 Power LUN's each stripped across 4 Tiers (8 plus 1 Parity of 250GB SATAII disks)
  - Block size = 4096
- Initiator – OFED 1.2 beta
  - Increased maximum number of gather/scatter entries per I/O
    - modprobe ib_srp srp_sg_tablesize (scatter and gather) =64
  - Increased Filesystem read-ahead sector count to 1024
    - hdparm –a 1024

Preliminary SRP Performance

- 1 to 4 concurrent sessions from 1 to 4 Initiators
  "Iozone –i 0 –r 128k –s 8g –f /mnt/srp1/test"

8g > Initiator memory; measurement involved Disk I/O



Sandia National Laboratories

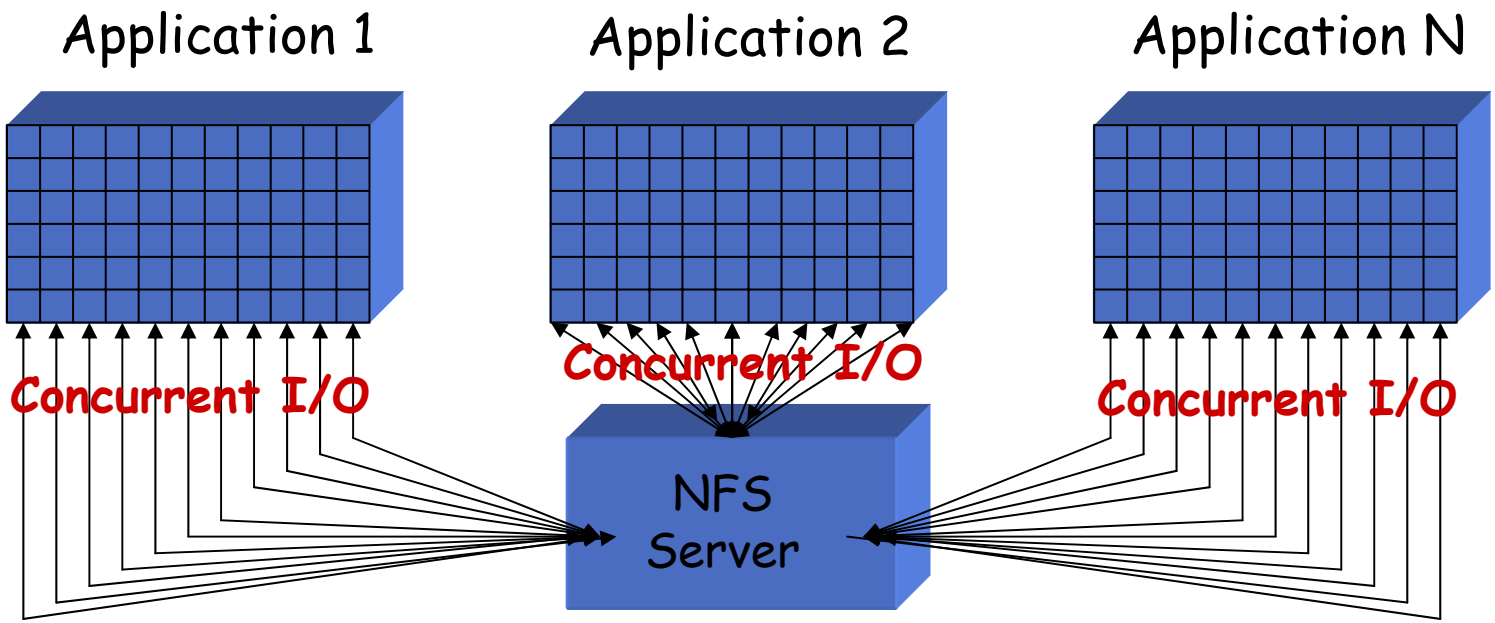
Preliminary SRP Performance

- Good but still room for tuning
  - Adjust maximum outstanding SCSI requests per LUN
  - Increase maximum SCSI command payload size
  - Evaluate Linux I/O Scheduling Algorithms
  - etc…

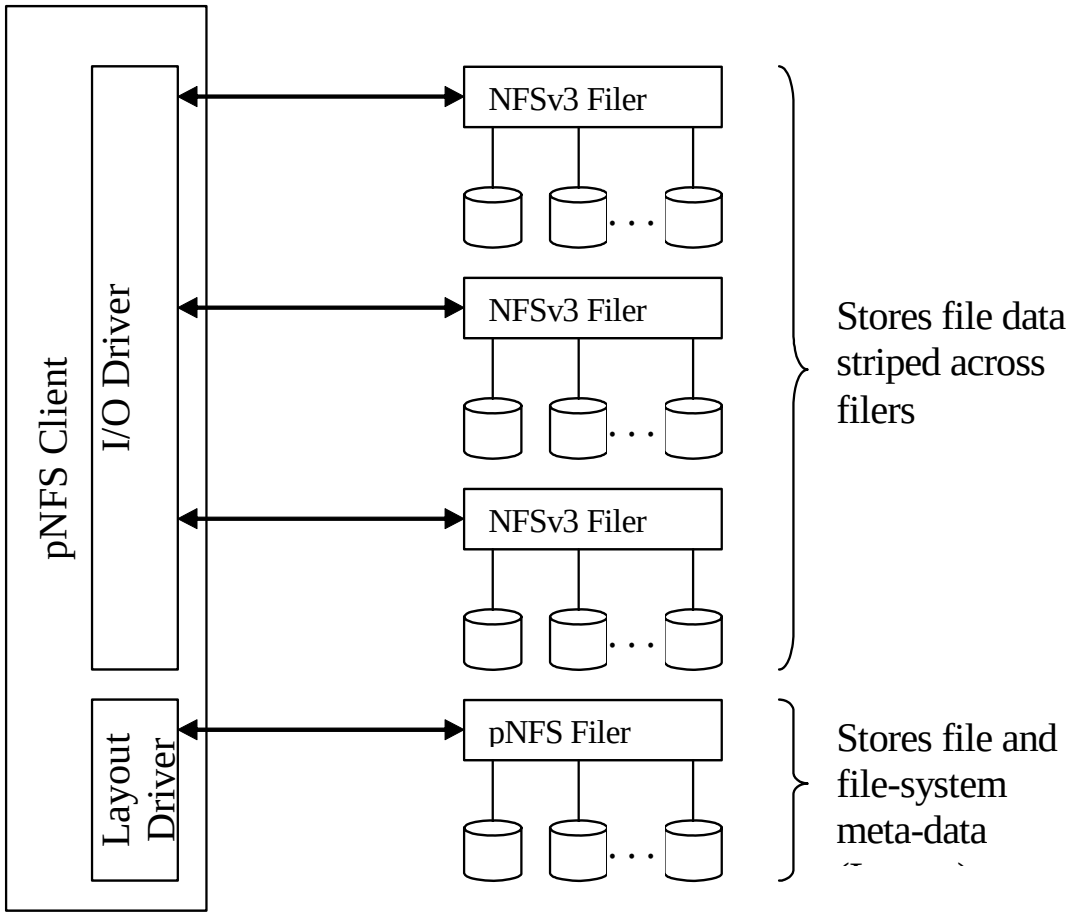
Sandia National Laboratories

Future Plans: The Need for pNFS

- Large number of concurrent requests from parallel applications
- Require parallelism in addition to RDMA



The Sandia - Open Grid Computing Research Collaboration

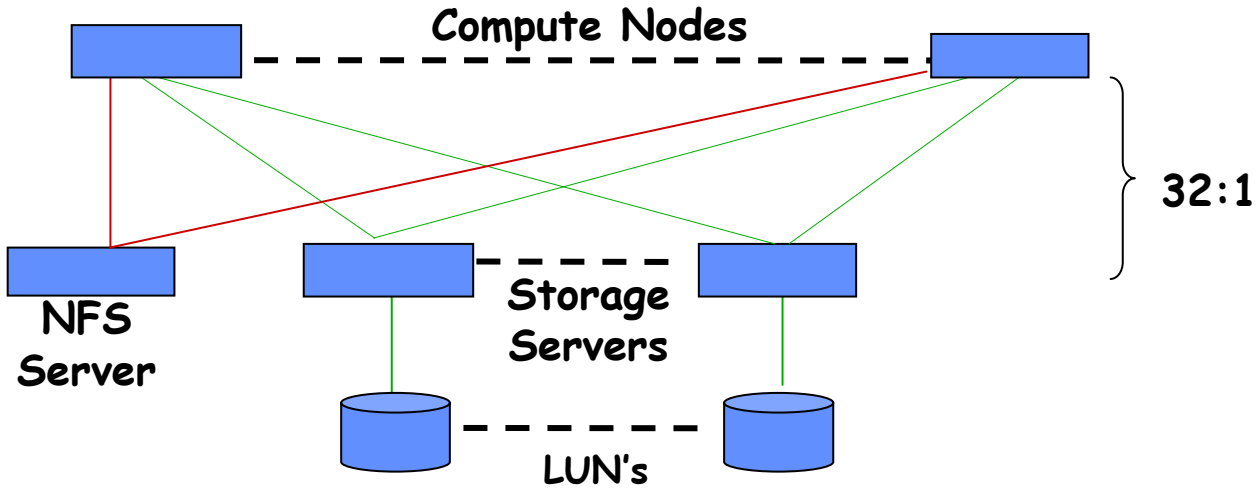


- **Based on the CITI implementation at UMICH**
- **Modified to stripe pNFS file data across RDMA enabled Linux NFSv3 Filers**
- **Open Source Linux environment**

Sandia National Laboratories

Future SRP Study

- Each Storage Server has to handle multiple independent large sequential writes and/or reads
  - Concurrent sequential I/O turned random
  - A challenge for Parallel Filesystem and Storage vendors



Acknowledgment

- Tom Tucker from Open Grid Computing and Tom Talpey from Network Appliance for their in depth technical support for NFS/RDMA
- Chas Williams from NRL, Randy Kreiser from DDN, and Dror Goldenberg from Mellanox for their assistance in IB SRP
- Felix Marti from Chelsio and Steve Wise from Open Grid Computing for their expertise in iWARP
- Jim Brandt from Sandia for his technical input and review

Sandia National Laboratories