SAND2006-7127C

Scaling NFS through RDMA for Cluster Computing

SuperComputing 2006

Storage Challenge

Tampa, Florida November 14, 2006

Dov Cohen, Helen Y. Chen, Jackie Chen, Ramanan Sankaran*, Noah Fisher, Jeff Decker

Sandia National Laboratories, CA

*** Oak Ridge National Laboratories, TN**

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
- RDMA technologies
- NFS over RDMA
- Testbed hardware and software
- Preliminary results and analysis
- Conclusion
- Ongoing work ad Future Plans

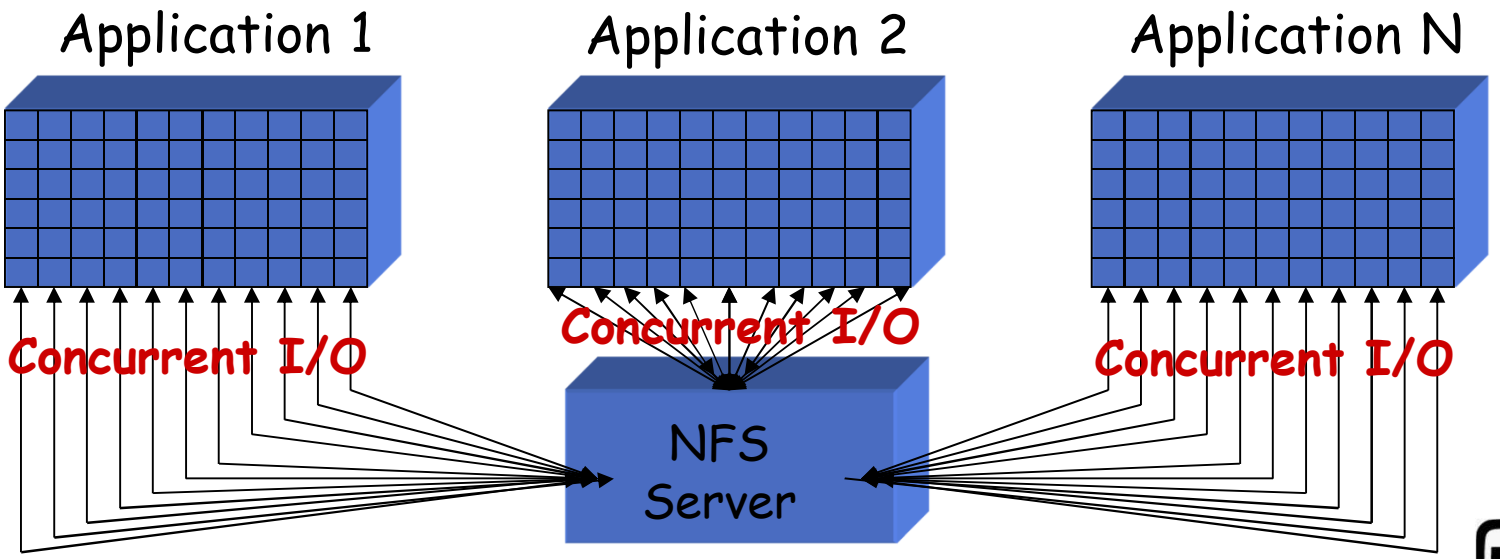
Sandia National Laboratories

Network File System (NFS)

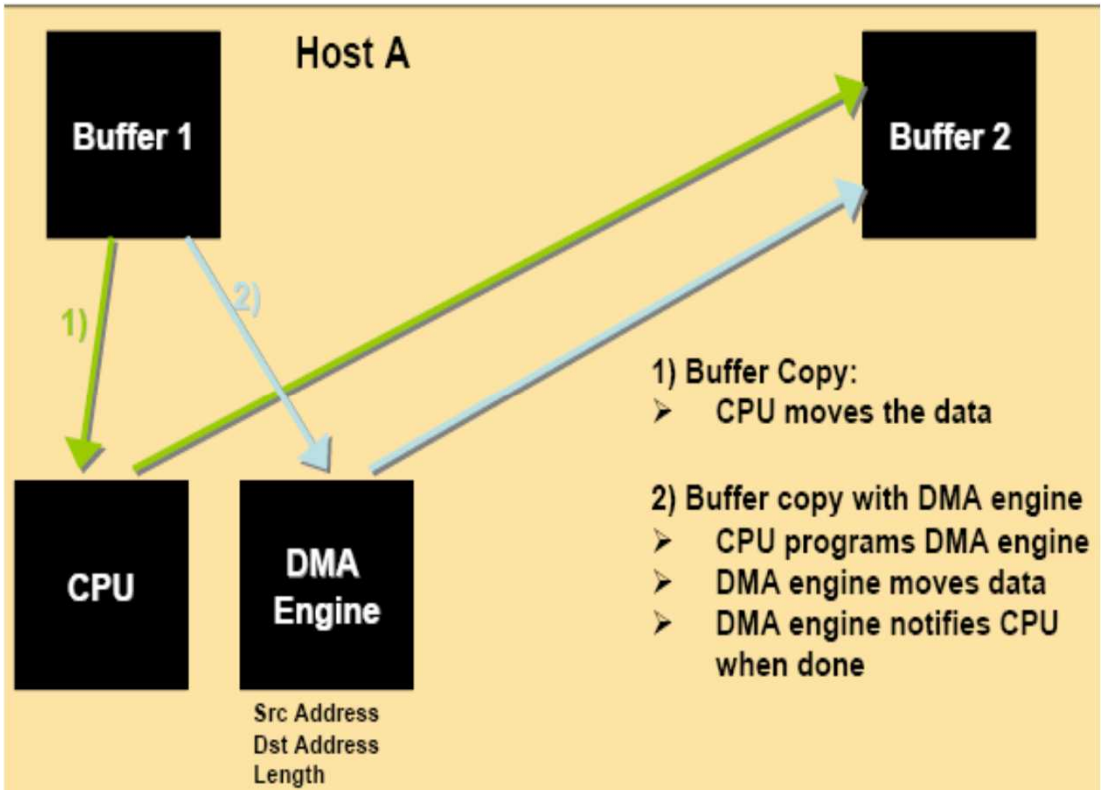
- A network attached storage file access protocol layered on RPC, typically carried over UDP/TCP over IP
- Allow files to be shared among multiple clients across LAN and WAN
- Standard, stable and mature protocol adopted for cluster platform

Scalability Limitations of NFS in Cluster Computing

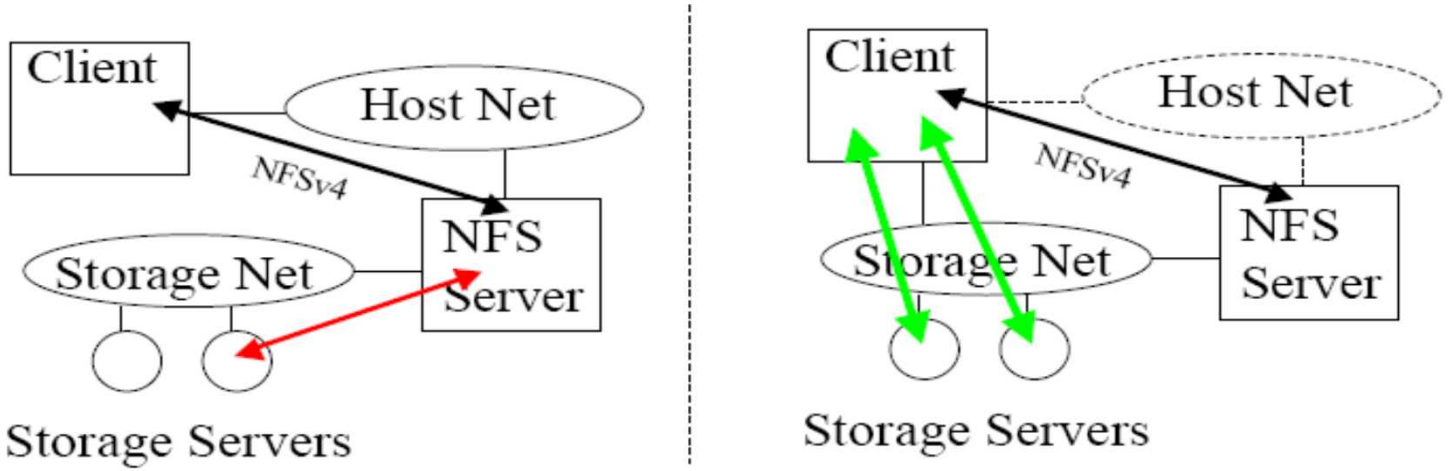
- Large number of concurrent requests from parallel applications
- Parallel I/O requests serialized by NFS to a large extend
- Need RDMA and pNFS



Direct Memory Access



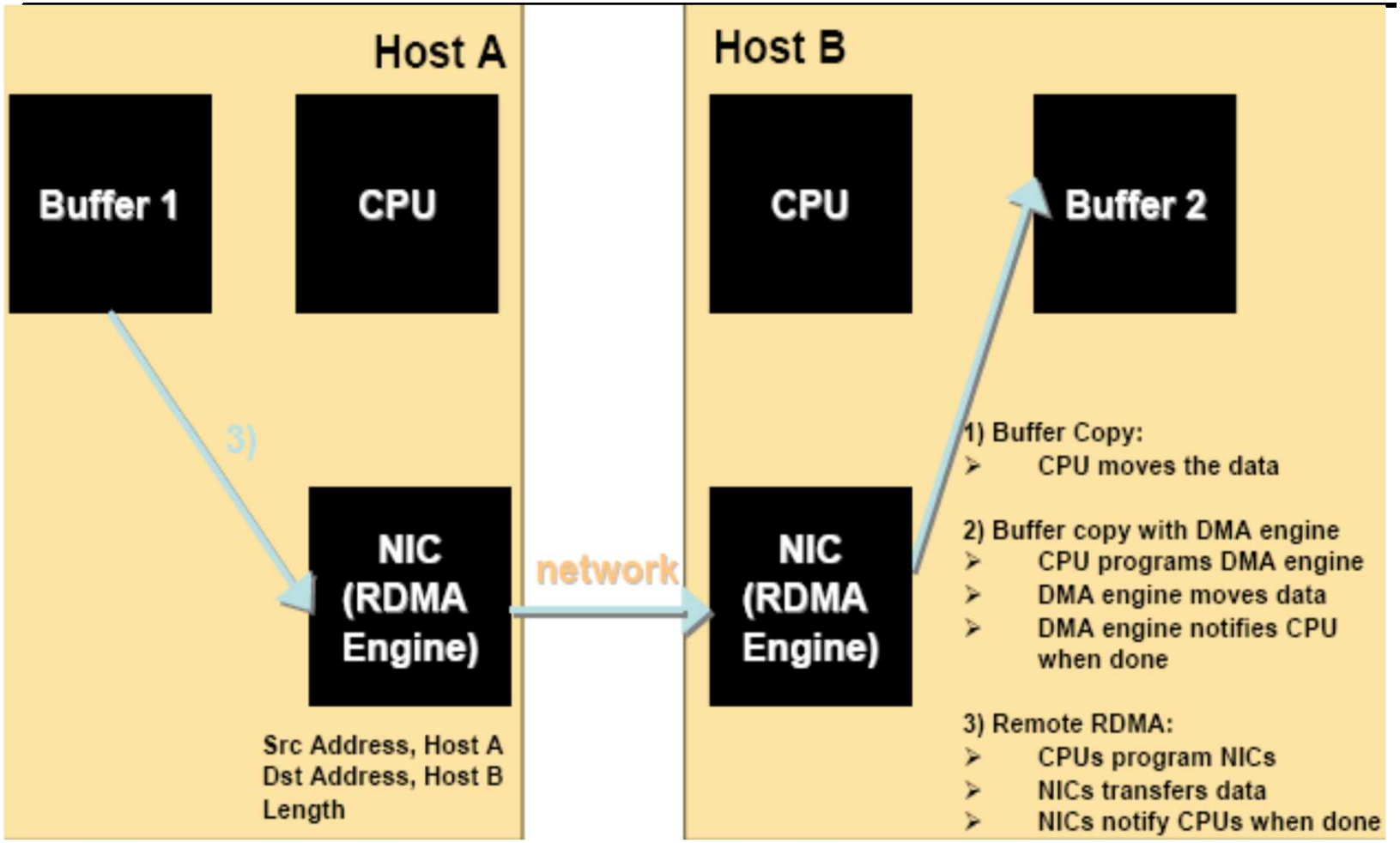
Parallel Network File System (PNFS)



- pNFS extends NFSv4
  - Minimum extension to allow out-of-band I/O
  - 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

Remote Memory Direct Access



NFS over RDMA

NFS moves big chunks of data incurring many copies with each RPC

Cluster Computing

- High bandwidth and low latency

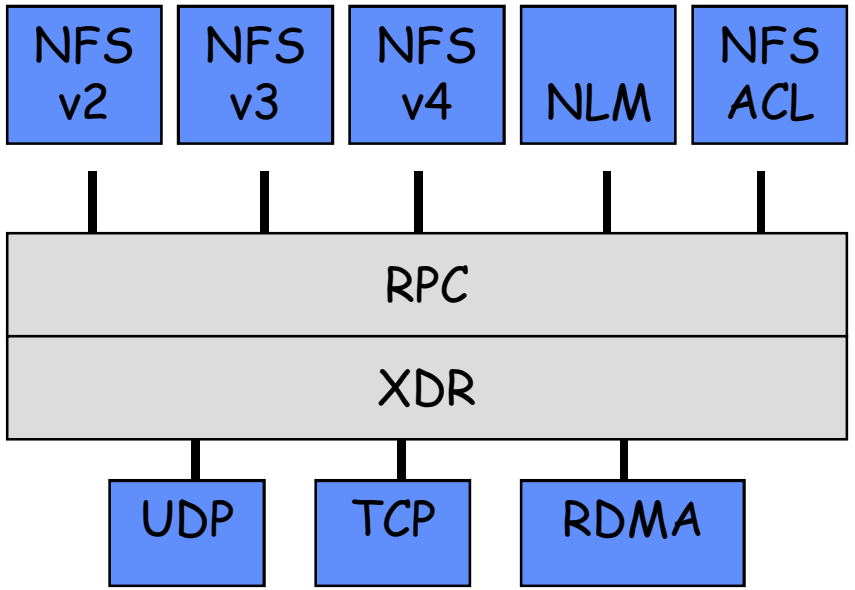
RDMA

- Offload protocol processing
- Offload host memory I/O bus
- Essential Component for 10/20 Gbps networks

http://www.ietf.org/internet-drafts/
draft-ietf-nfsv4-nfs-rdma-problem-statement-04.txt

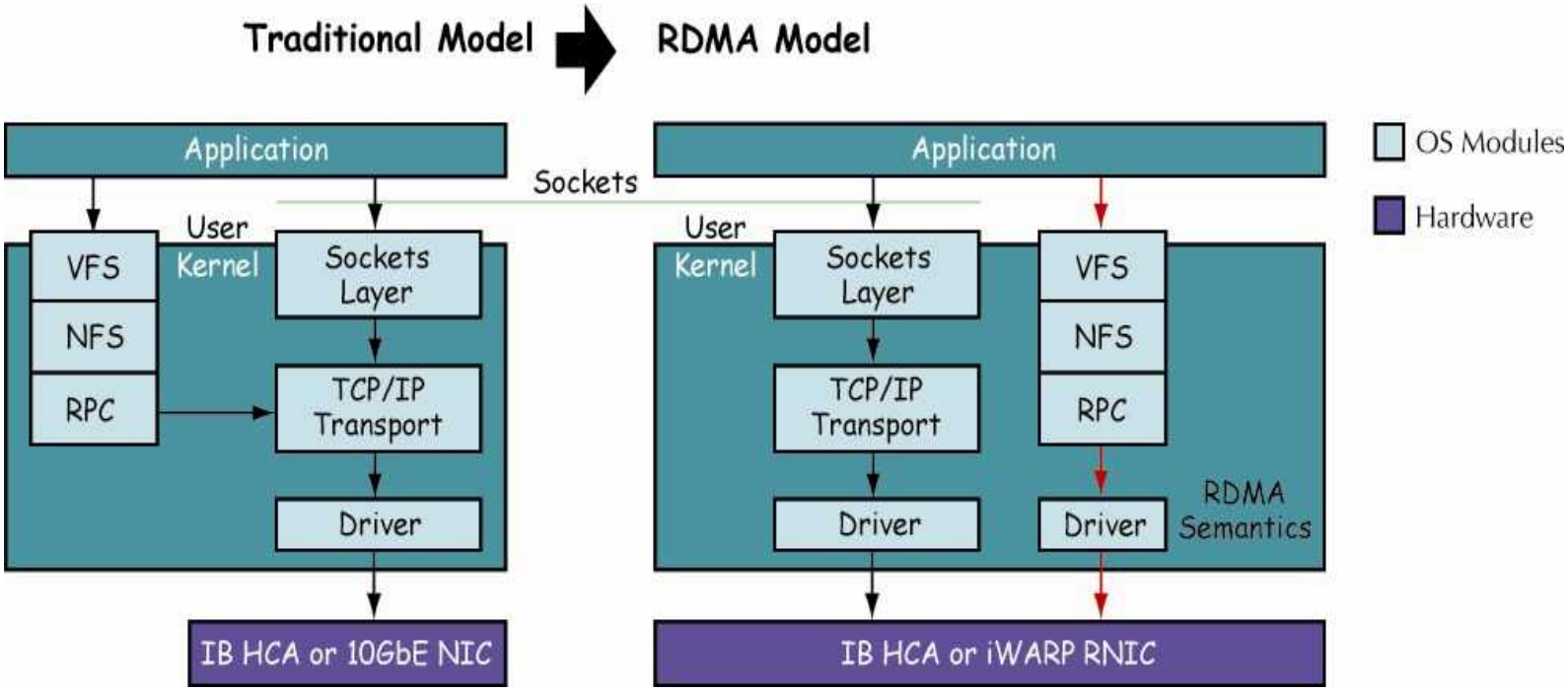
The NFS RDMA Architecture

- 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



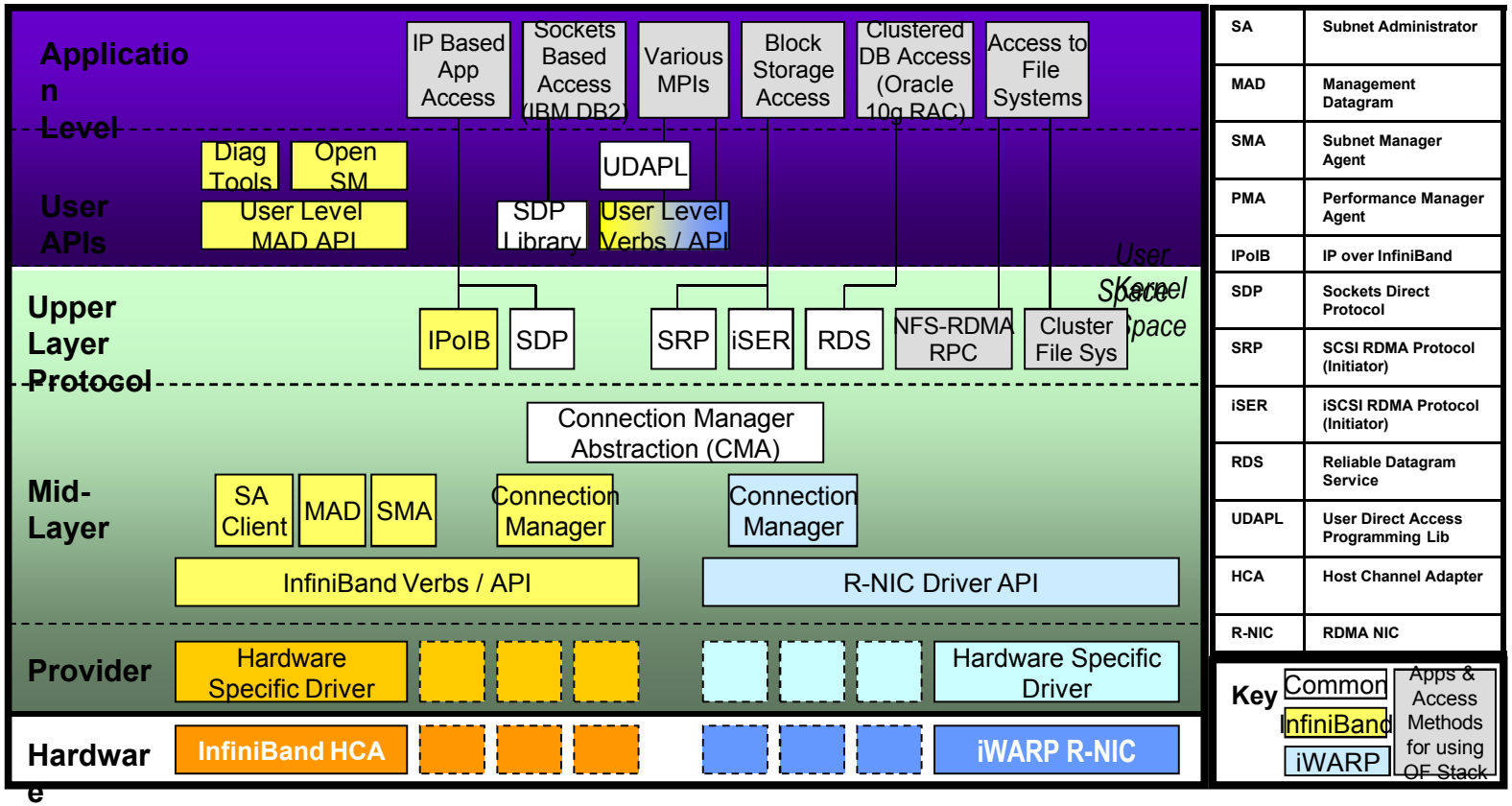
Brent Callaghan, Theresa Lingutla-Raj, Alex Chiu, Peter Staubach, Omer Asad, "NFS over RDMA", ACM SIGCOMM 2003 Workshops, August 25-27, 2003

RDMA Model



Sandia National Laboratories

The OpenFabric Stack



OPENFABRICS ALLIANCE

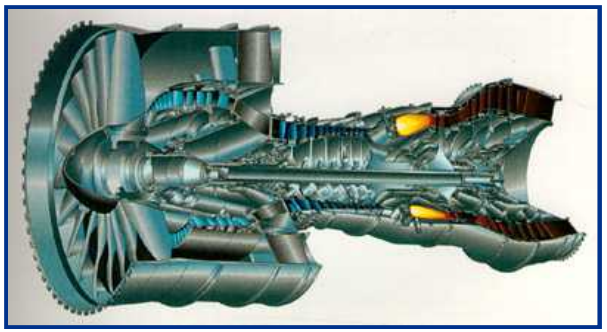
Offers a common, open source, and open development RDMA application programming interface

http://openfabrics.org

Sandia National Laboratories

Application Requirements: S3D
Direct Numerical Simulation (DNS) of Turbulent Combustion

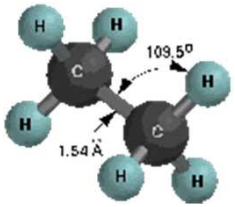
Turbulent Combustion is a Grand Challenge



Combustor size~1m

•Turbulent Combustion involves coupled phenomena at a wide range of scales.
•$O(10^4)$ continuum scales.

Molecular reactions ~ 1nm



DNS Approach and Role

•Fully resolve all continuum scales without using sub-grid models

•Only a limited range of scales is computationally feasible.
  •Terascale computing = DNS with $O(10^3)$ scales for cold flow.

•DNS is limited to small domains. Device-scale simulations are out of reach.

CRF

Sandia National Laboratories

S3D I/O Requirements

- Jet Simulation (Re=10,000) on X1E (20TF)
  - At the rate of one data dump every hour, **I/O rate is 64GB/hour**
- On a Petaflop system, required **I/O rate is 3.2 TB/hour**
- To achieve **5% maximum overhead**, I/O has to occur at **64TB/hour** or 17 GB/s
- It will be useful to dump data more often than once an hour if higher I/O rates are available

S3D Simulations	Grid points	Platform	Size per dump
Jet, Re=3,000	150M	XT3 (NCCS)	19GB
Jet, Re=5,000	350M	SP (NERSC)	45GB
Jet, Re=10,000	500M	X1E (NCCS)	64GB
Bunsen, u'/Sl=3	52M	X1E (NCCS)	8GB
Bunsen, u'/Sl=6	88M	XT3 (NCCS)	13GB
Bunsen, u'/Sl=10	200M	XT3(NCCS)	29GB

CRF

Sandia National Laboratories

Testbed System

- **Network File System**
  - NFS/RDMA release candidate 4
    - http://sourceforge.net/projects/nfs-rdma
  - Kernel: Linux 2.6.16.5 with deadline I/O scheduler
- **IB Fabric**
  - Switch: Flextronics InfiniScale III 24-port switch
  - HCA: Mellanox MT25208 InfiniHost III Ex
  - Software Stack: OpenFabric IB stack svn 7442
- **Server and Clients**
  - CPU: Dual 2.2 Ghz AMD Opterons
  - RAM: 8 GB (server) and 2 GB (clients)
- **Storage : Software RAID 0**

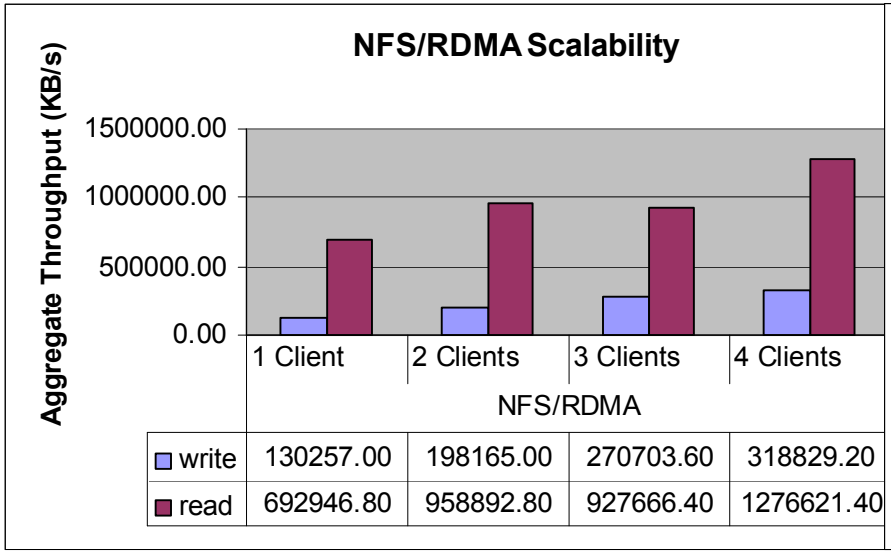
Sandia National Laboratories

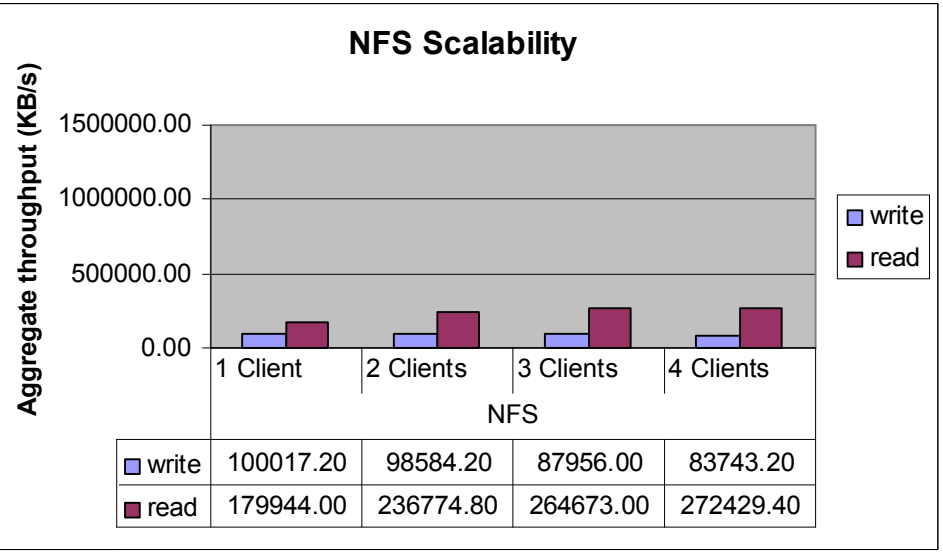
System Performance Benchmarks

I/O Rates (MB/Sec)	Local	NFS (IPoIB)	NFS/RDMA
Write	266.11	100.02	130.26
Read	1518.20	179.94	692.94

- Reads are from server cache reflecting
  - TCP RPC transport achieved ~180 MB/s (1.4 Gb/s) of throughput
  - RDMA RPC transport was capable of delivering ~700MB/s (5.6Gb/s) throughput

Scalability Tests - Throughput

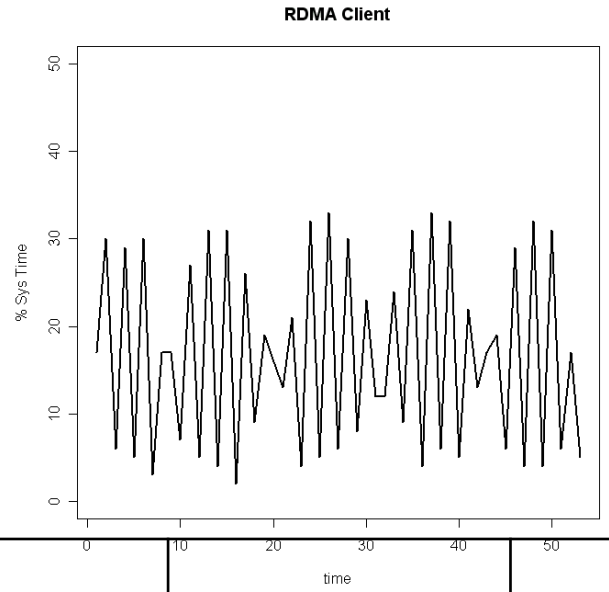


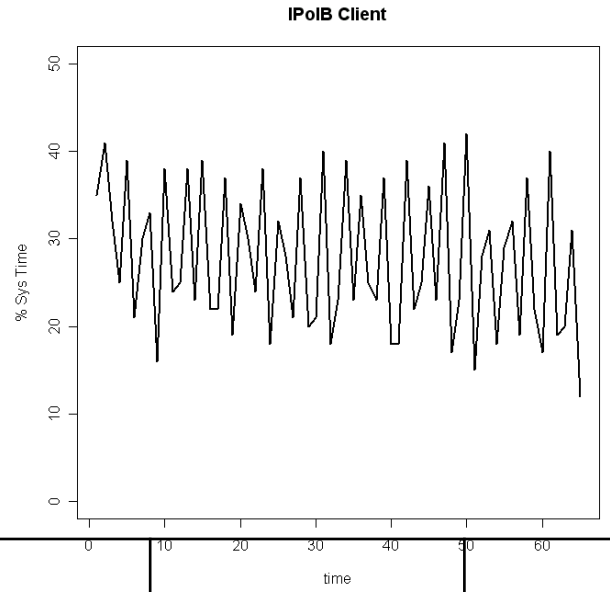


- To minimize the impact of disk I/O
  - One 5GB, two 2.5GB, three 1.67GB, four 1.25GB
- Ignored rewrite and reread due to client-side cache effect

Sandia National Laboratories

CPU Utilization Client Side

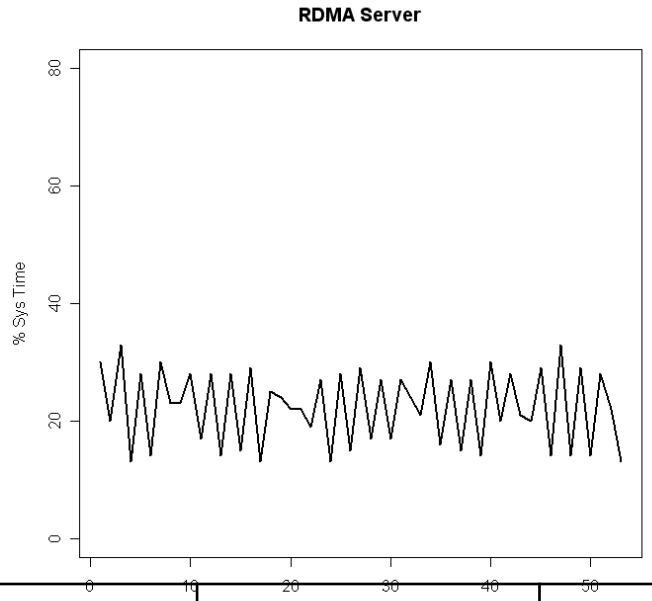


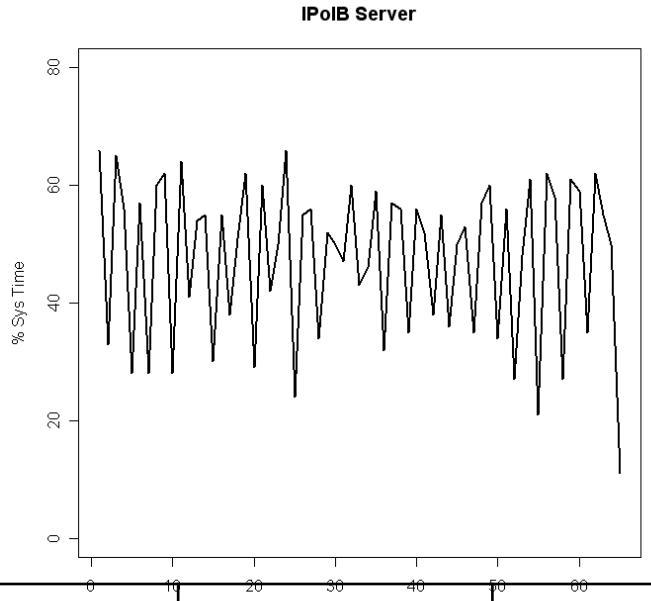


Mode	BW (MB/sec)	Context Switch	Interrupt	% Sys
RDMA	94	3614	6376	16.58
IPoIB	76	5215	5215	27.67

Sandia National Laboratories

Server CPU Utilization – 1 client

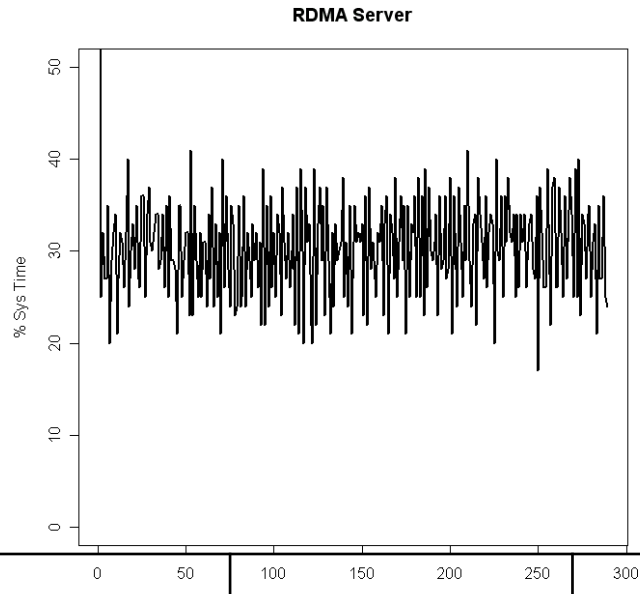


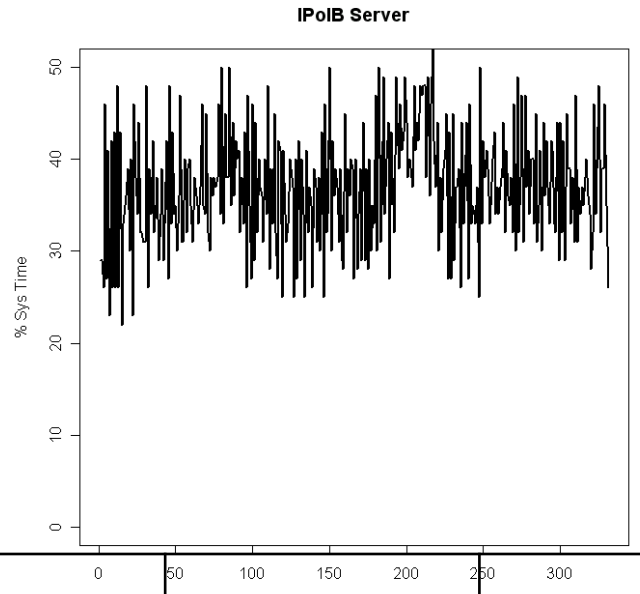


Mode	BW (MB/sec)	Context Switch	Interrupt	% Sys
RDMA	94	20084	10021	22
IPoIB	76	22227	6323	47

Sandia National Laboratories

Server CPU Utilization – 6 clients

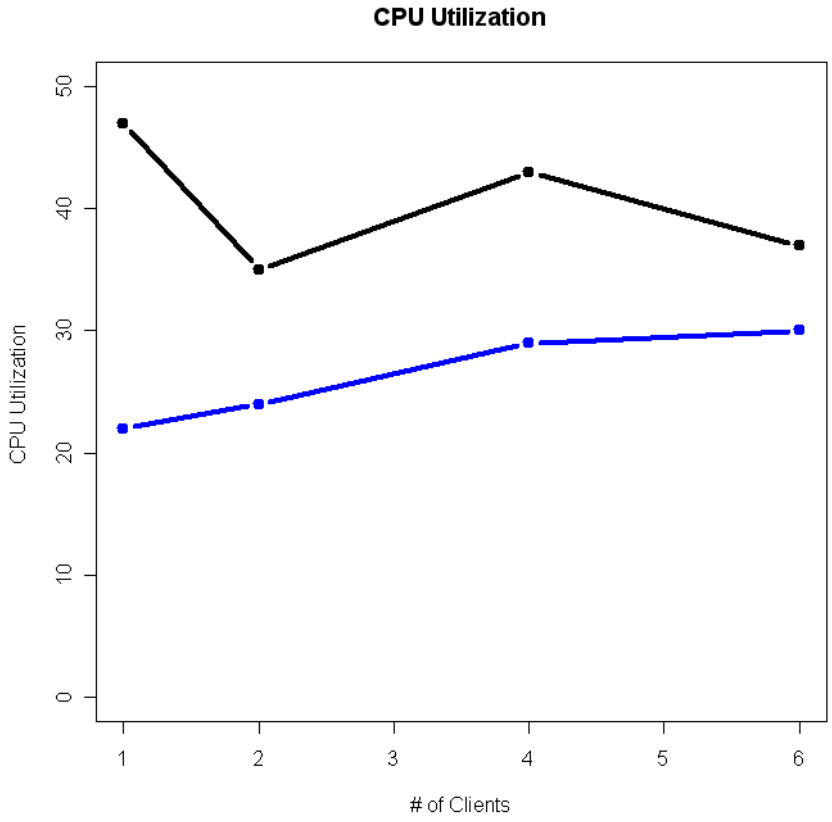


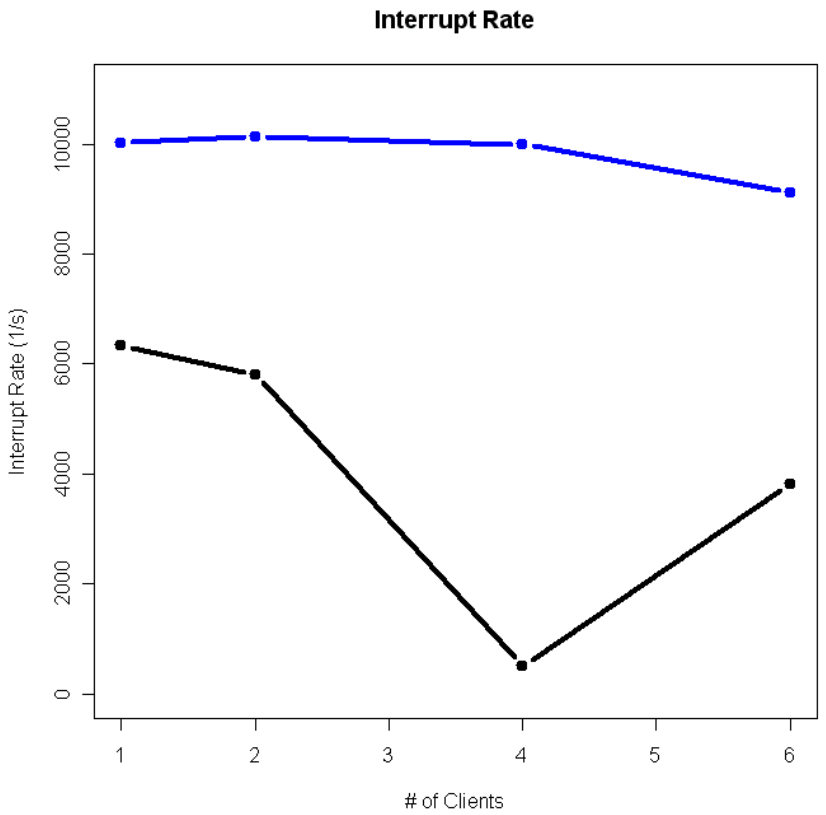


Mode	BW (MB/sec)	Context Switch	Interrupt	% Sys
RDMA	120	12844	9115	30
IPoIB	90	12021	3810	37

Sandia National Laboratories

Server Performance vs # of Clients





Sandia National Laboratories

S3D Run Server Performance

Mode	Total Time (hours)	Context Switch	Interrupt	% Sys
RDMA	13.9	12957	2506	7.7 +/- 10.2
IPoIB	14.5	15729	2135	13.9 +/- 4.8

- Overall Computation Time reduced by 5 %
- Fraction of the time server spends in Sys state is reduced by 44 %

S3D Run Client Performance

Mode	Apparent BW (MB/sec)	Context Switch	Interrupt	% Run	% Sys
RDMA	415 +/- 35	10512	1147	17.6 +/- 18.8	4.9 +/- 4.4
IPoIB	132 +/- 19	10126	1122	16.49 +/- 18.6	4.5 +/- 5.2

- Apparent Bandwidth increased by factor > 2x
  - Caching effect in RDMA system
- Clients spend a larger fraction of their time in the run state

Conclusions

- **NFS/RDMA demonstrated:**
  - **More efficient server CPU utilization**
  - **Enhanced caching performance, can take advantage of superior memory to memory bandwidth**
- ***NFS/RDMA has the potential to improve application and system level performance!***
- ***NFS/RDMA can easily take advantage of the bandwidth in 10/20 Gigabit network for large file accesses***