# pNFS/PVFS2 over InfiniBand: Early Experiences

Lei Chai Xiangyong Ouyang **Ranjit Noronha**
Dhabaleswar K. Panda

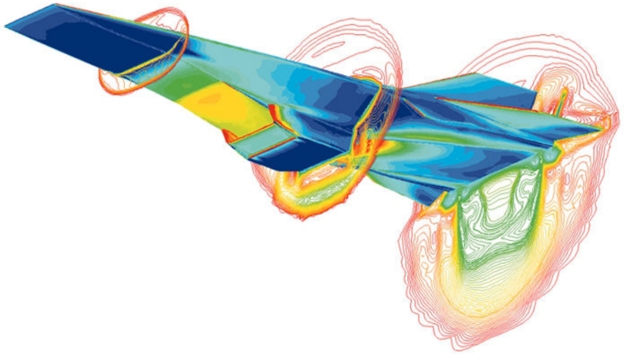
Department of Computer Science and Engineering
Ohio State University

# Outline of the talk

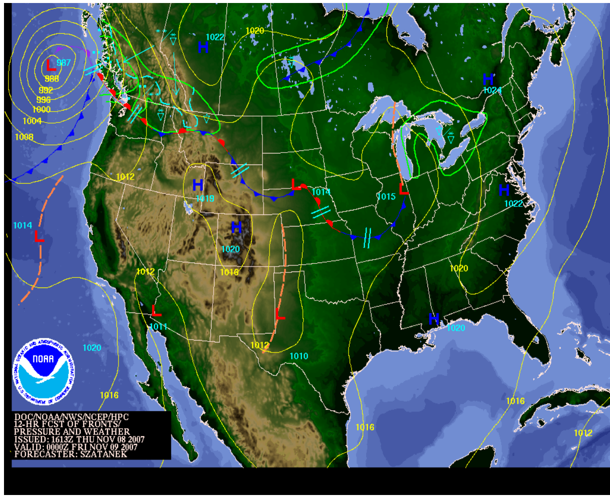
- Introduction and Background
- Problem statement
- Design of experiments
- Results
- Conclusions and future work

# Introduction

- Petascale Environments
  - Requires high-performance I/O systems to provide data in a sustained high throughput manner
- NFS
  - Widely deployed
  - Single server bottleneck
- Parallel file systems
  - PVFS2, Lustre, GPFS, etc
  - Good parallel performance
- Can pNFS bridge the gap between NFS and parallel file systems and be the solution for petascale file systems?

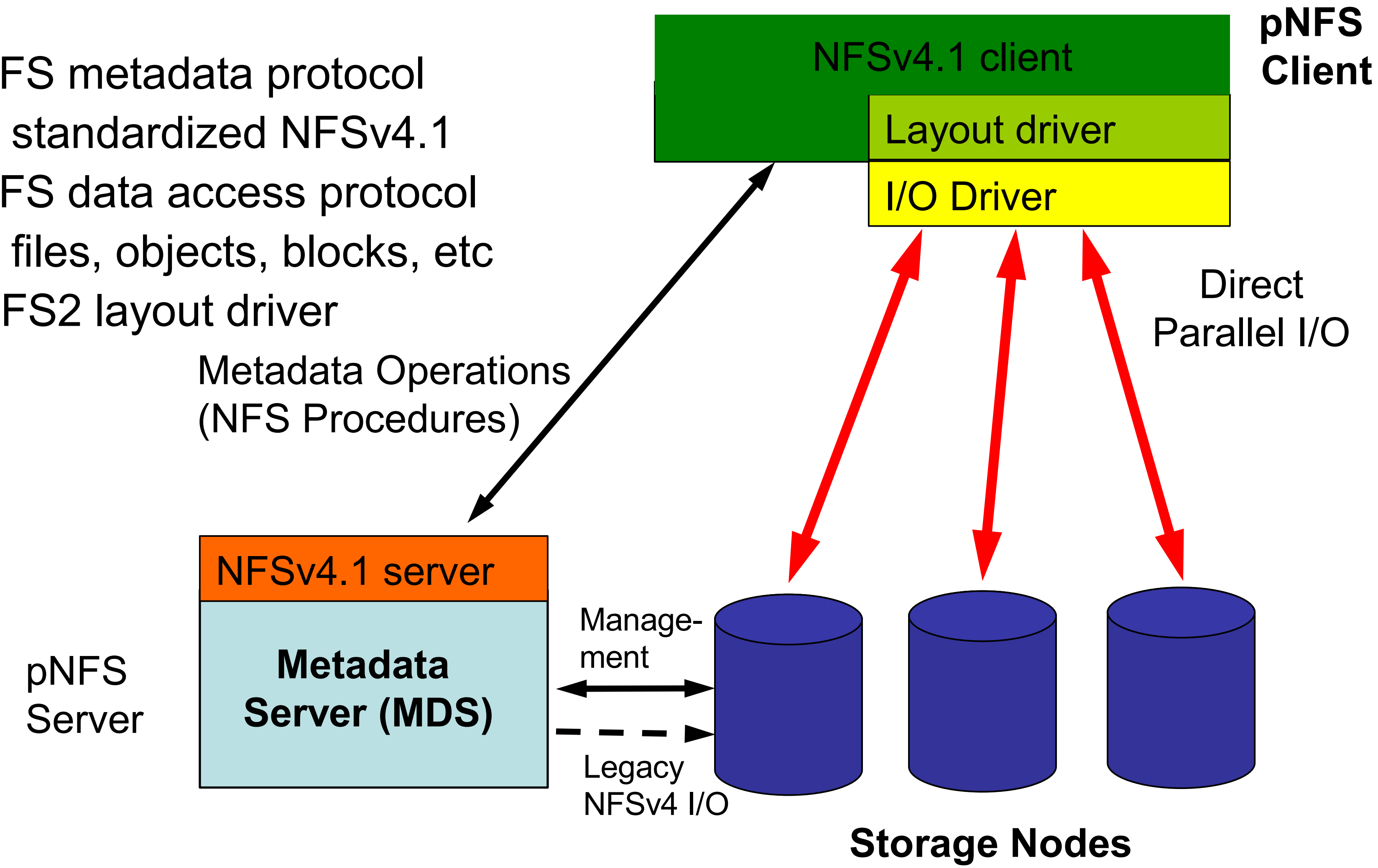


Dryden Flight Research Center ED97 43968-01
HYPER-X AT MACH 7: This computational fluid dynamic (CFD) image is of the Hyper-X vehicle at the Mach 7 test condition with the engine operating.



# Background – pNFS Architecture

- pNFS metadata protocol
  - standardized NFSv4.1
- pNFS data access protocol
  - files, objects, blocks, etc
- PVFS2 layout driver



# Background - InfiniBand

- Commodity High Performance Interconnect
- Communication semantics
  - Send/Recv
  - Remote Direct Memory Access (RDMA)
  - Communication Offload
- Performance characteristics
  - Low latency (< 2 μs)
  - High Bandwidth
  - Low CPU utilization
- InfiniBand standard supports
  - Single data rate (SDR) – 10Gbps
  - Double data rate (DDR) – 20Gbps
  - Quad data rate (QDR) – 40Gbps
- Widely deployed in clusters

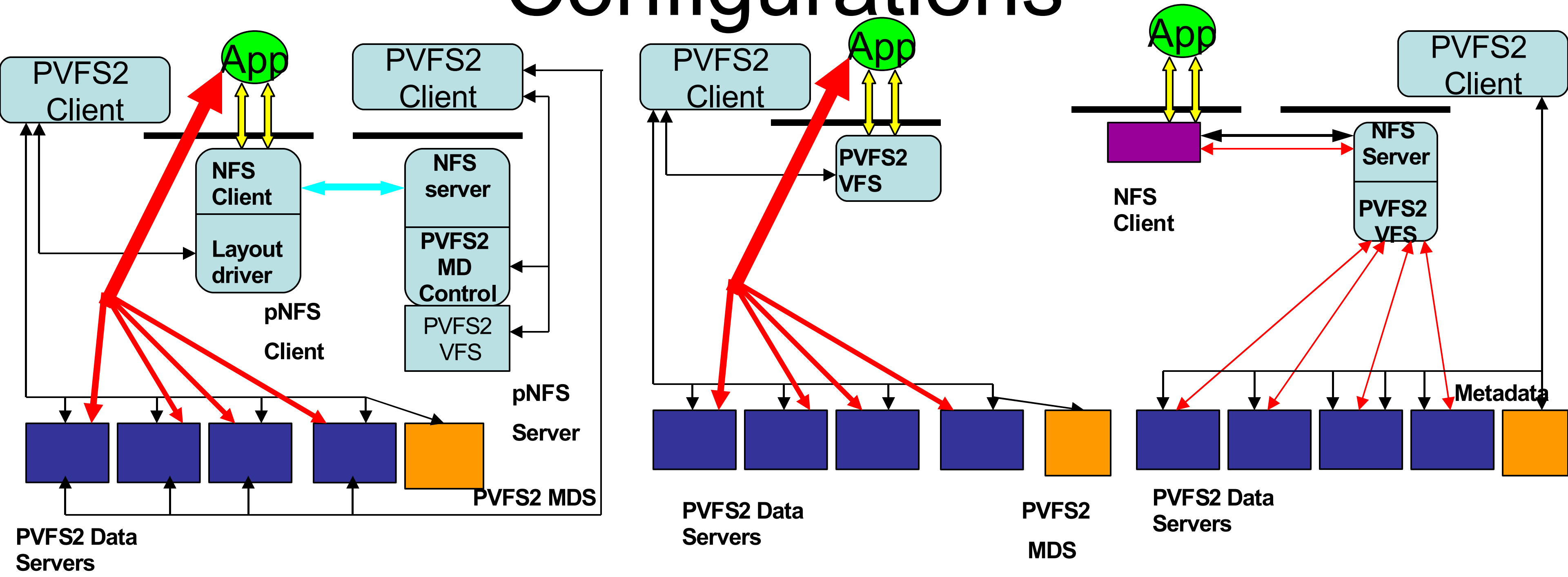
# Problem Statement

- What are the advantages of InfiniBand over Gigabit Ethernet in a parallel file system environment?
- How much is the performance gain of using pNFS instead of the traditional single server NFS?
- Any potential overhead introduced by the pNFS PVFS2 layout driver compared with native PVFS2?
- How does pNFS scale with an increasing number of I/O servers?

# Outline of the talk

- Introduction and Background
- Problem statement
- Design of experiments
  - File System Configuration
  - Network Transports
  - Node Setup/Benchmarks
- Results
- Conclusions and future work

# Experimental File System Configurations



(a) pNFS/PVFS2

(b) PVFS2

(c) NFS/PVFS2

↔ ***NFS Metadata***

→ ***Data Paths***

→ ***Control Paths***

↔ ***I/O Requests***

- pNFS with a PVFS2 layout driver (pNFS/PVFS2)
- PVFS2 with a VFS mount (PVFS2)
- NFSv4 server using a PVFS2 file system as the backend (NFSv4/PVFS2)

# Experimental Setup - Network Transports

- Either InfiniBand or GigE is used as the transport
  - Native IB -OpenIB Gen2 (IB)
  - IP over InfiniBand (IPoIB)
  - TCP over Ethernet (GigE)

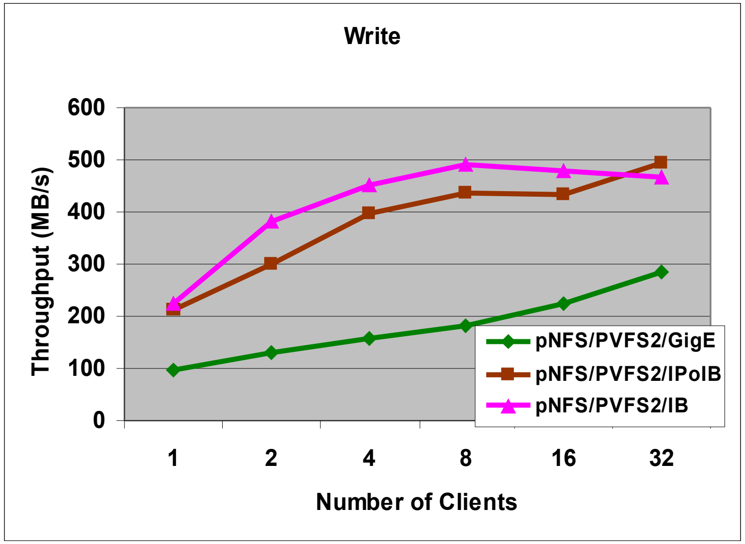
# Experimental Setup-Node Setup

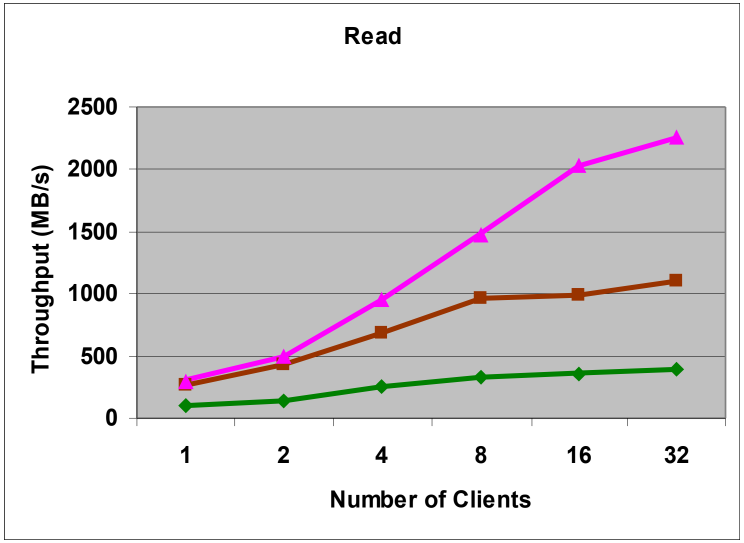
- Hardware
  - Intel Clovertown cluster with 32 compute nodes and 8 storage nodes
  - Each node is equipped with a  2.33 GHz
  - 6GB main memory, PCI-Express bus
  - Connected by both Gigabit Ethernet and Mellanox InfiniBand DDR cards
  - Each storage node is equipped with 3ware RAID controller, 16 disks in RAID-0 configuration
- Benchmark
  - IOzone multi-thread Write/Read throughput tests
    - File size 256MB
    - Record size 2MB
    - 1 process per client

# Outline of the talk

- Introduction and Background
- Problem statement
- Design of experiments
- Results
  - Network and Protocol Impact
  - Setup Comparison (Native IB)
  - Setup Comparison (IPoIB)
  - Scalability with varying I/O servers
  - Alternate Techniques (NFS/RDMA)
- Conclusions and future work

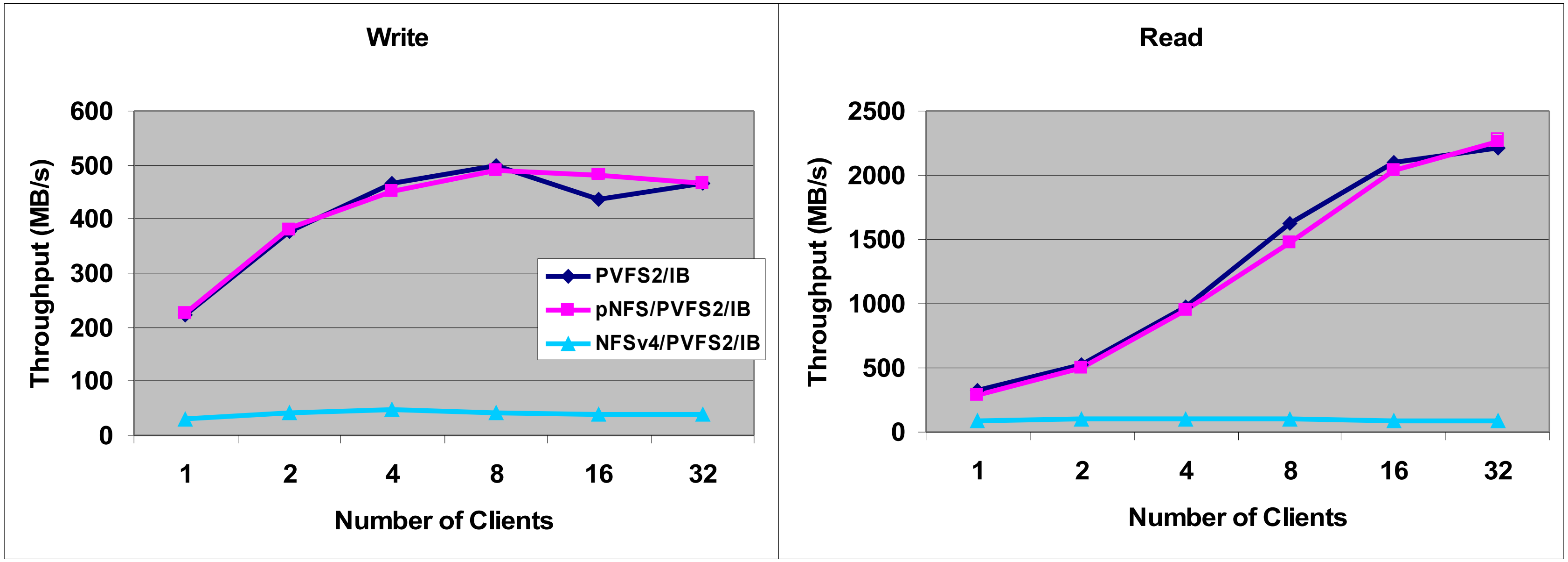
# Network and Protocol Impact





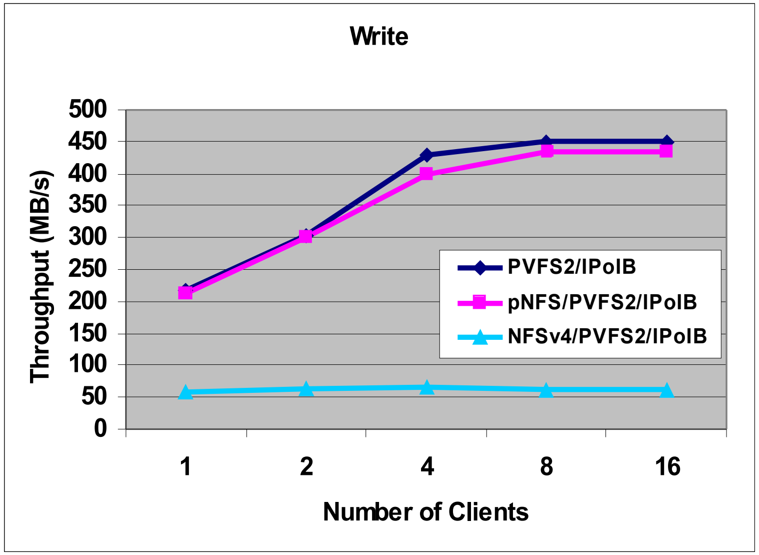
- Results with 4 I/O servers
- Compared with GigE, IPoIB improves throughput by up to
  - Write 150%
  - Read 200%
- Compared with GigE, Native IB improves throughput by up to
  - Write 190%
  - Read 480%

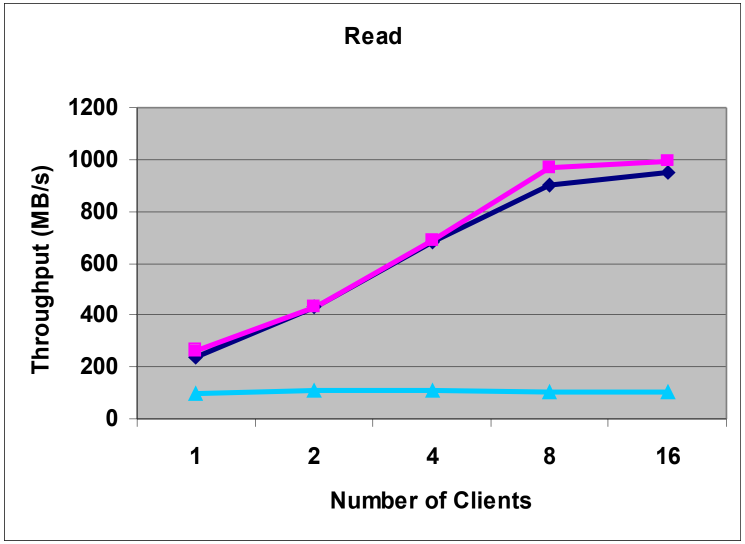
# Setup Comparison (Native IB)



- pNFS/PVFS2 peak throughput:
  - Write 490MB/s
  - Read 2256MB/s
- pNFS/PVFS2 performs comparably with native PVFS2
- pNFS/PVFS2 improves performance significantly compared with NFSv4/PVFS2

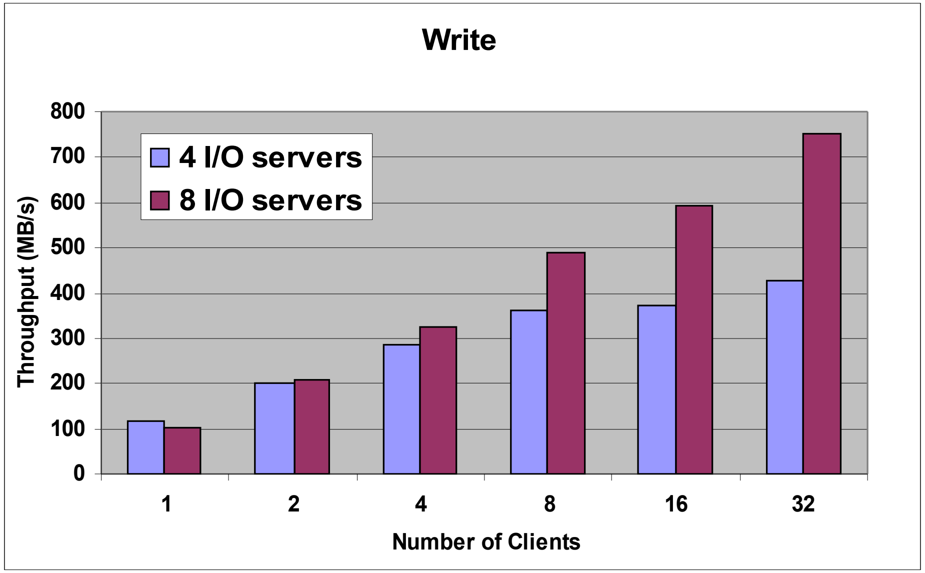
# Setup Comparison (IPoIB)

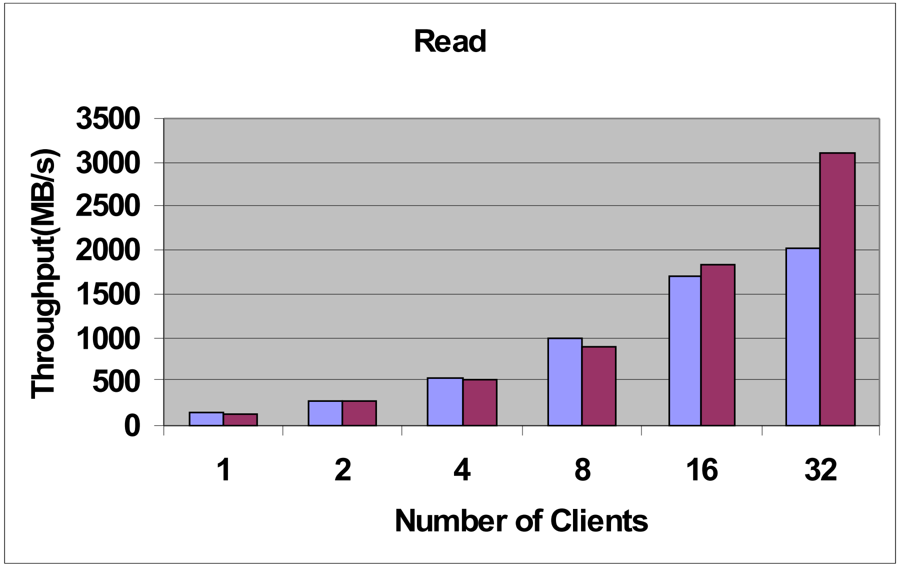




- pNFS/PVFS2 peak throughput:
  - Write 435MB/s
  - Read 1107MB/s
- Same trend
  - pNFS/PVFS2 performs comparably with native PVFS2
  - pNFS/PVFS2 improves performance significantly compared with NFSv4/PVFS2

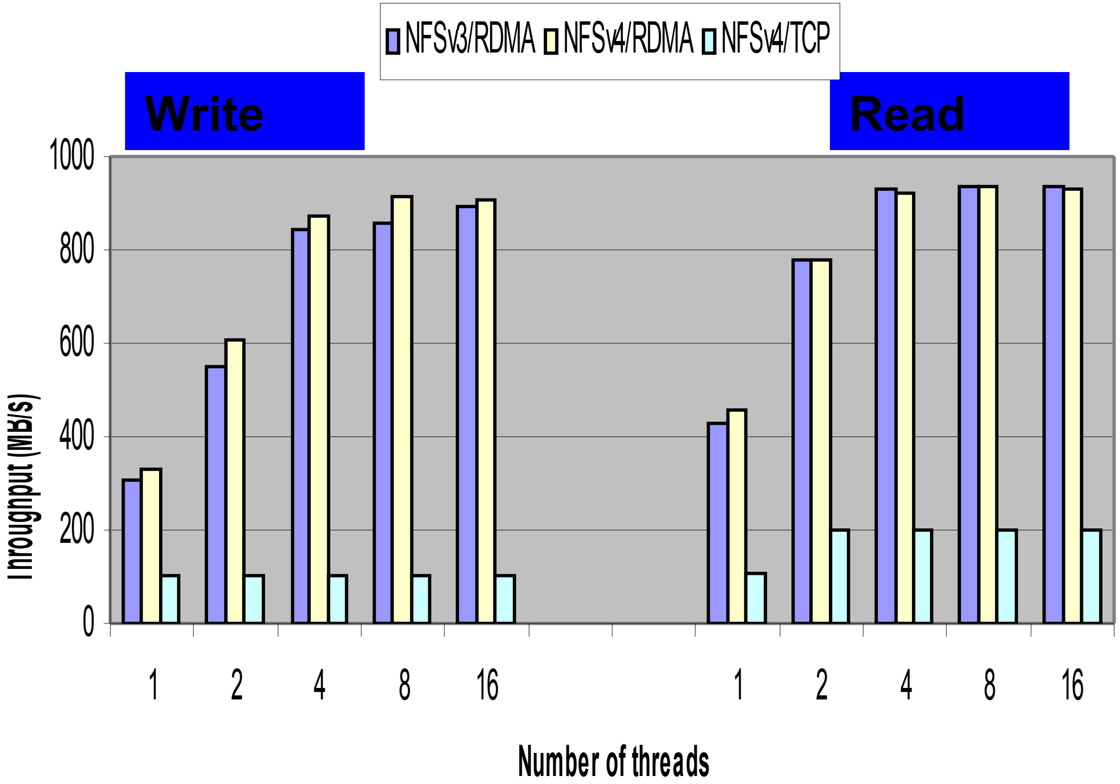
# pNFS Scalability with I/O Servers





- pNFS/PVFS2/IB (native IB)
- Peak READ throughput
  - 3099 MB/s (8 I/O servers)
- Peak WRITE throughput
  - 754 MB/s (8 I/O servers)

# Alternate Techniques (NFS/RDMA)



- OpenSolaris NFS over RDMA Project
  - Collaboration with Sun and NetApp
  - Improved performance compared to TCP/IP (IPoIB)
  - To be incorporated into OpenSolaris kernel

***NFSv4 READ bandwidth is 933 MB/s***

***NFSv4 WRITE bandwidth is 917 MB/s***

http://nowlab.cse.ohio-state.edu/projects/nfsrdma/index.html

# Conclusions

- What are the advantages of using InfiniBand over Gigabit Ethernet in a parallel file system environment?
  - InfiniBand significantly improves pNFS/PVFS2 performance
    - Write throughput 490MB/s
    - Read throughput 2256MB/s
    - Up to 480% improvement compared with using GigE
- How much is the performance gain of using pNFS instead of the traditional single server NFS?
  - pNFS/PVFS2 provides significantly higher throughput and shows better scalability than NFS/PVFS2
    - Write up to11 times improvement
    - Read up to 24 times improvement

# Conclusions (Cont'd)

- Any potential overhead introduced by the pNFS PVFS2 layout driver compared with native PVFS2?
  - Very little overhead
    - pNFS/PVFS2 achieves the same performance as the native PVFS2
- How does pNFS scale with an increasing number of I/O servers?
  - 754 MB/s (aggregate Write)
  - 3099 MB/s (aggregate Read)
- To conclude
  - Performance evaluation of pNFS/PVFS2 on an InfiniBand cluster
  - pNFS is promising as the file system solution for clusters

# Future Work

- File based layout, e.g. NFS/RDMA
- Larger scale experiments with more I/O servers and clients
- Application level evaluation
- Using 10 GigE/iWARP as the underlying transport

# Acknowledgements

Our research is supported by the following organizations

Office of Science
U.S. DEPARTMENT OF ENERGY

NSF

Cisco Systems
Empowering the Internet Generation

Sun microsystems

NetApp

# Thank you

{chail, ouyangx, noronha, panda}@cse.ohio-state.edu

Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

Project Web Page

http://nowlab.cse.ohio-state.edu/projects/nfsrdma/index.html