



IP Storage

A Performance and Security Study

June 11, 2003

Helen Y. Chen, Neal Bierbaum
High Performance Computing and Networking

Jamie Van Randwyk, Frank Bielecki
Information Security



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.





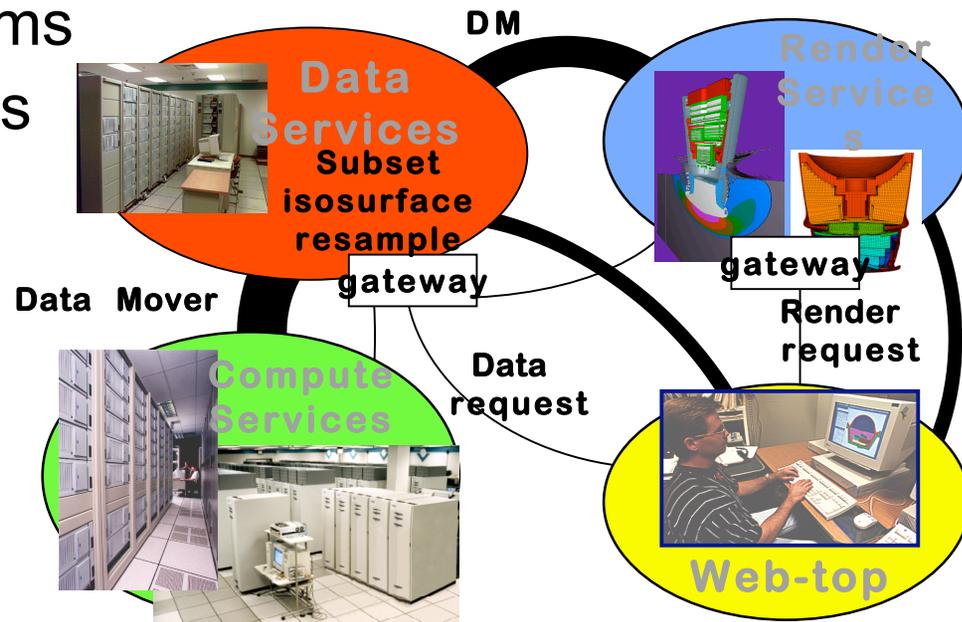
Outline

- **The IP Storage Proposal**
- **FY03 Accomplishment**
 - **Background studies**
 - **Benchmark methodology**
 - **Industry partners and testbed**
 - **Preliminary result and analysis**
- **Conclusion**
- **FY04 Plans**



The Big Science Challenge

- Large scientific calculations generate hundreds of terabytes of data
- Compute resources are geographically distributed
 - Compute platforms
 - Graphics engines
 - Storage
 - Scientists
- Ad Hoc solutions lack performance and robustness

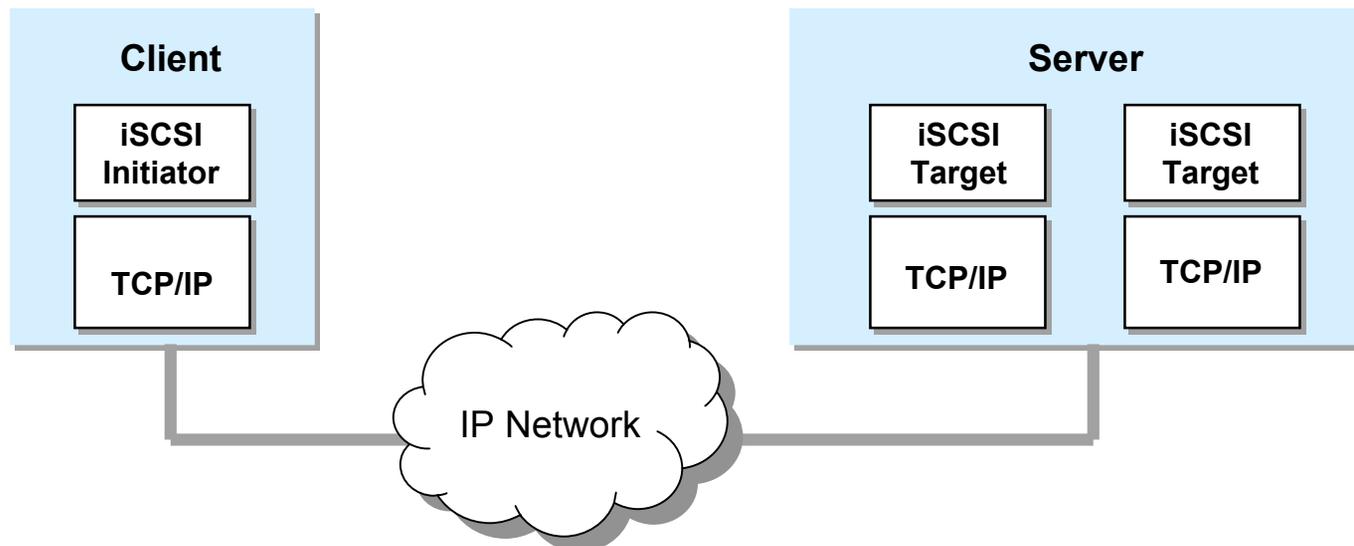




Research Objective

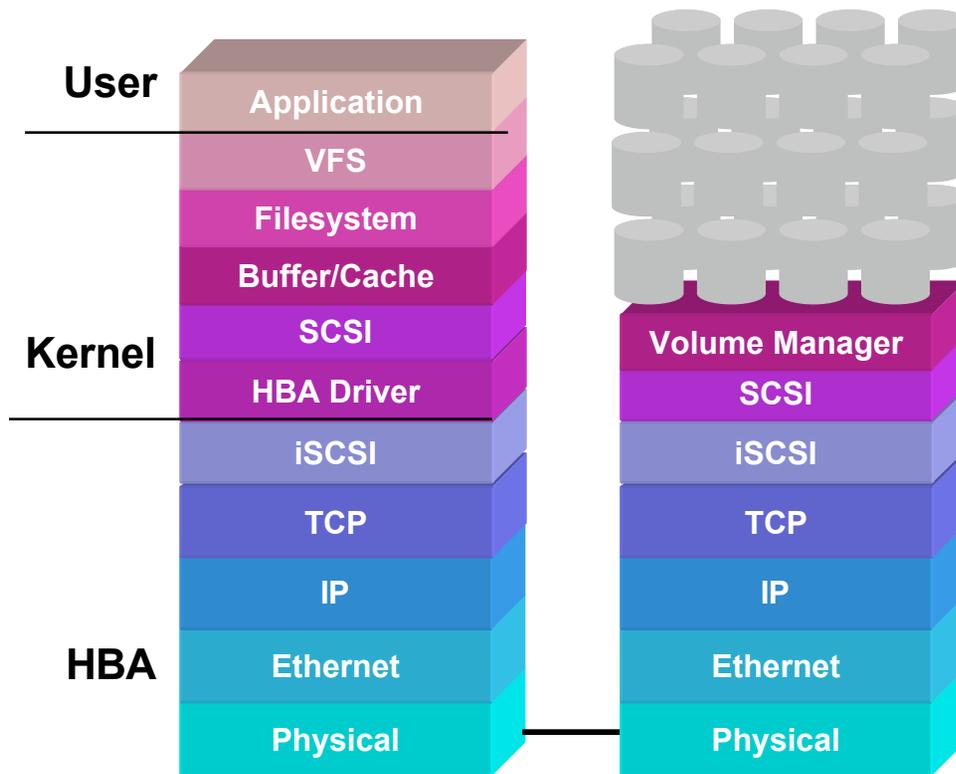
Transparently deliver high throughput, low latency network and storage performance to large scale scientific applications

- **Use standard based IP storage and leverage on**
 - **Raid, SCSI and IP level parallel technique and failover mechanisms**
 - **Ubiquitous deployment of TCP/IP and Ethernet**
 - **Block I/O over distance**





FY03 Accomplishment



Background Studies

- Traffic profiling of scientific applications
- Linux Kernel
 - Filesystem
 - Buffer/Cache
 - SCSI Class Driver
 - Device Driver
- iSCSI transport
 - Resource discovery
 - Login and authentication
 - Security negotiation
 - SCSI operations and performance tuning
 - Error detection and recovery



FY03 Accomplishment (continued)

- **Benchmark Tools**

- ***Xdd*** is a tool for measuring and characterizing file as well as storage I/O on single or cluster of systems
- ***IOzone*** is a filesystem benchmark tool to generate and measure performance of file level operations
- ***Performance Co Pilot (PCP)*** is a system-level performance monitoring tool to correlate application performance with platform activity
- ***Ethereal*** is a public domain network protocol analyzer that can parse iSCSI and NFS in addition to TCP/IP



Fy03 Accomplishment (continued)

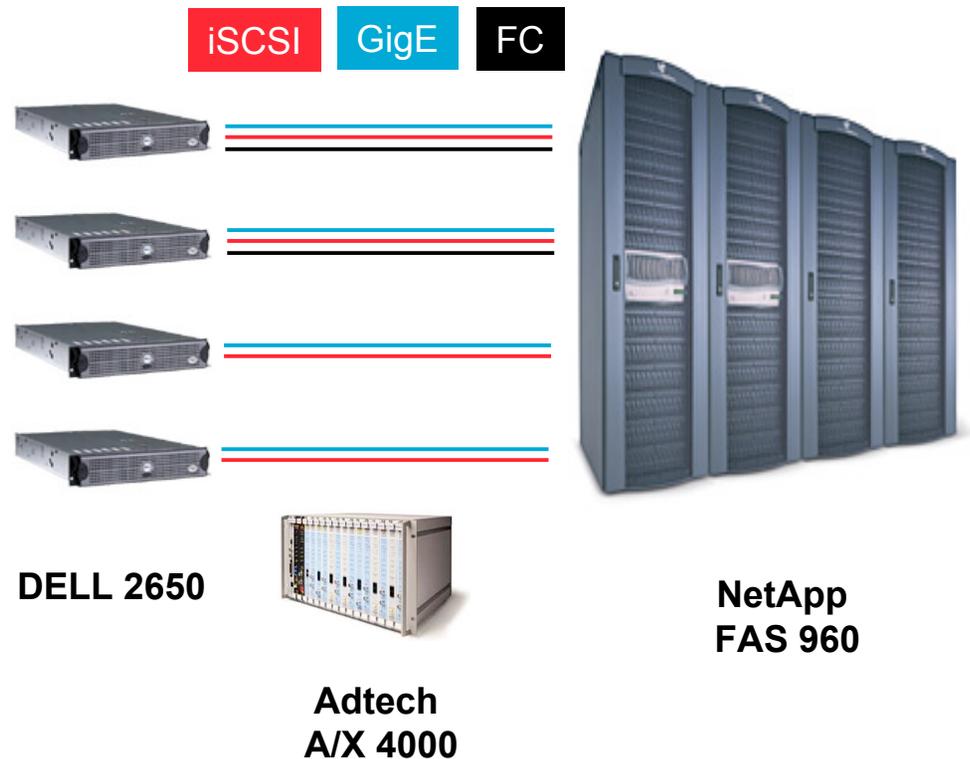
Benchmark methodology

- **Developed a test control program that integrates test definition, test execution, as well as organization of test results and related information**
 - XML definition files define the test environment, the test program parameters, and the scheduling of simultaneous runs across multiple hosts
 - Results of each run are reported in a series of XML and HTML files to allow consistent, unambiguous search and processing of results from thousands of test runs
 - The test program can support a wide variety of testing projects and will be released for general Sandia use at the end of this project
- **Written a post processing tool that correlates benchmark and monitoring statistics of a test, and generate a spreadsheet entry for analysis and plotting**



FY03 Accomplishment (continued)

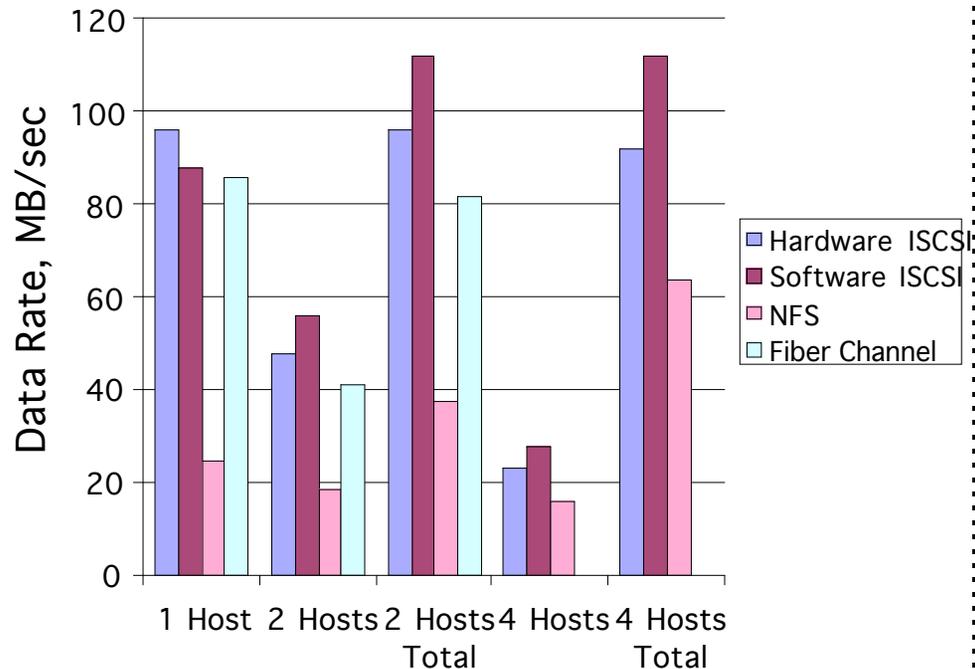
- **Dell 2650 Server**
 - Dual 2.4 Ghz Intel Xeon, 2GB memory, RH 2.4.18 Kernel
- **Adaptec iSCSI HBA**
 - iSCSI firmware and TCP off load engine
 - 512 MB onboard memory
 - Gigabit Ethernet, 1500 B MTU
- **NetApp Appliance**
 - Dual 2.2 Ghz P4, 6GB memory, 6TB disks
 - NFS
 - Software iSCSI over kernel TCP/IP
 - 4 Gigabit Ethernet, 1500 and 9000 B MTU
 - 2 Fibre Channel, 2048 byte frame
- **Cisco Systems**
 - Software iSCSI driver for Linux
- **SysKonnnect**
 - Gigabit Ethernet NIC, 1500 & 9000 B MTU
 - Used for software iSCSI and NFS tests
- **Adtech delay simulator**





Preliminary Results and Analysis

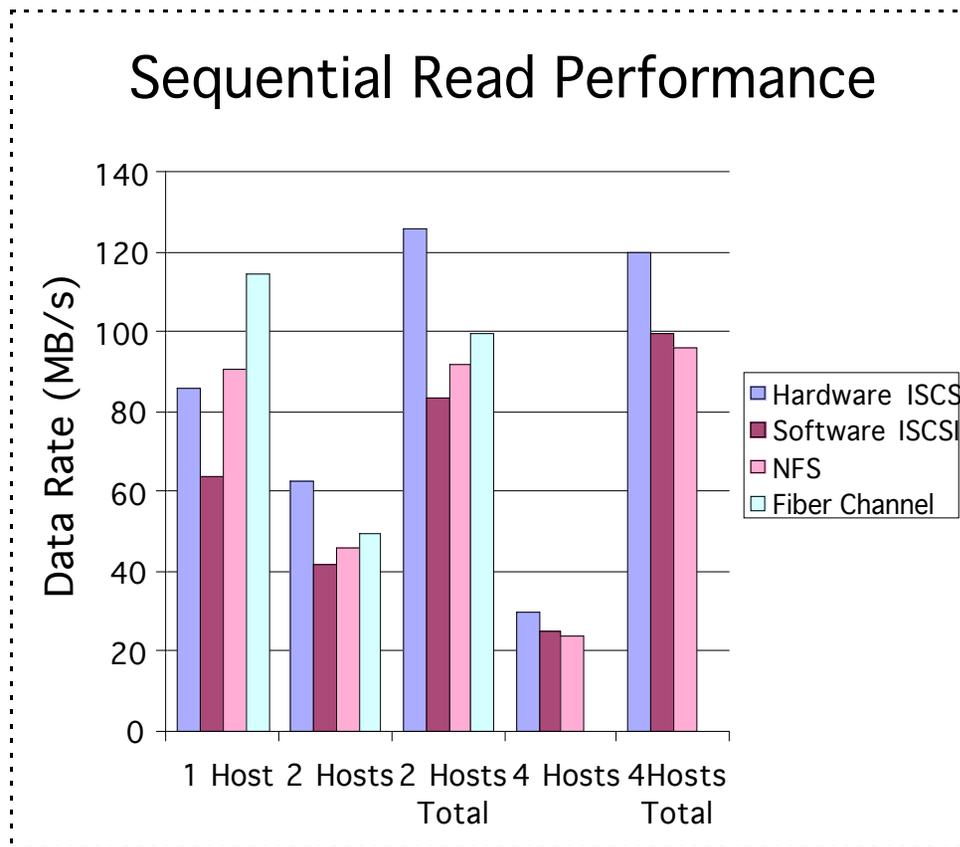
Sequential Write Performance



- Hardware iSCSI outperformed Fibre Channel in single and multi host tests
- Hardware iSCSI used 1024 Byte Protocol Data Unit, causing server CPU to bottleneck in multi-host tests
- Software iSCSI achieved better aggregate throughput with Jumbo Frame support
- NFS protocol had limited its per session performance
- Server bottleneck prevented linear speed up in multi-host runs



Preliminary Results and Analysis (continued)

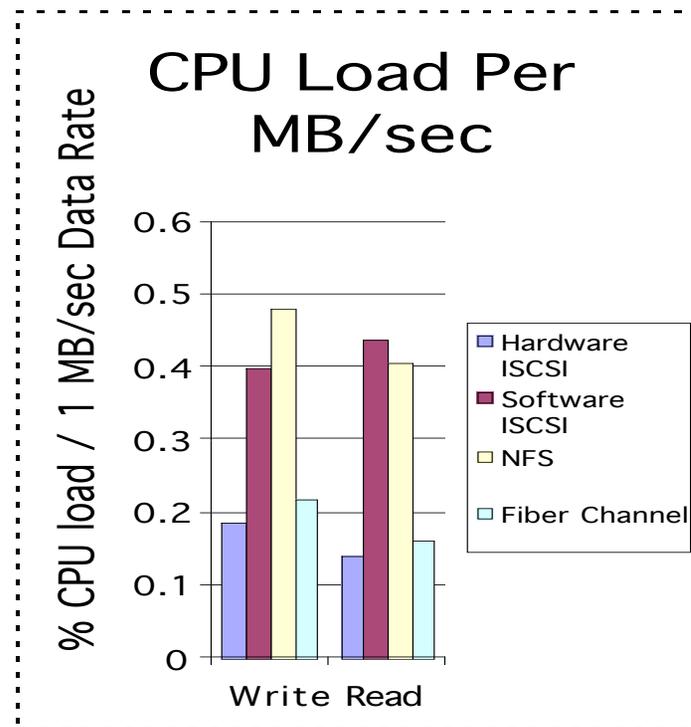
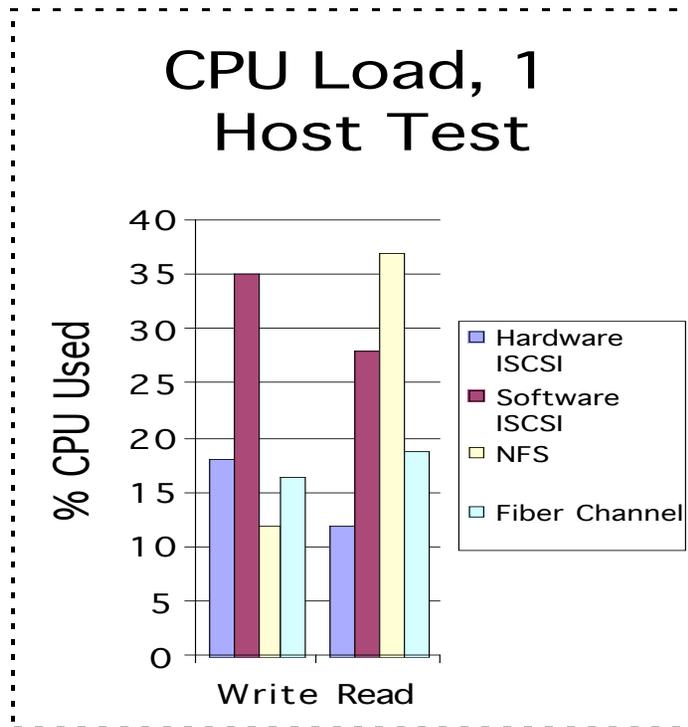


- Fibre channel delivered best throughput in single host test
- Hardware iSCSI achieved best aggregate throughput, reaching the upper bound of server's resource
- NFS achieved similar throughput among single and multi host tests, due to server limitation
- High protocol processing overhead limited NFS and software iSCSI performance



Preliminary Results and Analysis (continued)

Hardware iSCSI exhibited most efficient CPU utilization





Conclusion

- **iSCSI transport can deliver FC like performance in SAN environments**
- **Using Jumbo Frame with Software iSCSI, 1.5 Ghz of Intel Xeon is used to process iSCSI and TCP/IP protocol to deliver 1 Gbps**
- **With IPsec, hardware HBA would certainly be necessary to deliver Gigabit per second line rate**
- **Hardware iSCSI and TOE support is more critical at server end**
- **Computer architecture and TOE will be issues at 10 Gbps**
 - RDMA, PCI-Express, and TOE design

iSCSI can offer better cost/performance ratio than FC SAN!

- Hardware iSCSI HBA \$600, Gigabit Ethernet Switch \$800/port
- Fibre Channel HBA \$1200, Fibre Channel Switch \$1200/port



Hardware iSCSI Support for WAN

- **Support Jumbo Frame**
 - Achievable BW $\leq (MSS * C) / (RTT * P^{1/2})$
 - Require RDMA to decrease onboard memory requirement
- **Expose tunable iSCSI TCP parameters**
 - TCP window
 - Timestamp option
 - SACK, etc.
- **Implement parallel TCP connections per iSCSI session**
 - Failover mechanism
 - Higher aggregate throughput
- **Allow iSCSI and SCSI level tuning**
 - Tagged/linked SCSI commands
 - MaxBurstLength, MaxRecvDataSegmentLength, FirstBurstLength
 - MaxCmdSN, etc.



Future Plans

- **Measure WAN performance**
 - Throughput
 - Congestion
 - QoS requirements
 - Failover mechanisms
- **Conduct security study**
 - IPsec and IKE
 - Authentication, data integrity and privacy protection
 - Performance impact
- **HPC application**
 - Global parallel filesystem
 - Cluster filesystem
 - Remote replication / mirroring
- **Publish in refereed journal or conference**

Implementation and Deployment ?