DATA Intensive Computing:

Sun's HPC I/O Strategy

Presented at

Lustre User Group

04/23/07

Larry McIntosh
Global Advanced Computing Solutions
Sun Microsystems, Inc.

Agenda

- Look at Customer's Data Access Requirements
- Look at Customer's Data Sizing Requirements
- Look at Sun's Thumper -- Real Life
- Look at Thumper Target Markets
- Examine where Sun is deploying Thumper with Lustre
- Discuss Sun's strategy for HPC I/O
- Discuss Sun's Three Tier HPC Storage Architecture

Customer Observations and Needs

Customer Data Observations and Needs

- Multi-Node large-scale deployments are ramping up
- Compute and Data opportunities are coupled and beg for a balanced solution
- Scalable Object Based File Systems have matured
- Customers seek answers from established vendors

Customer's are expanding their I/O requirements which encompass

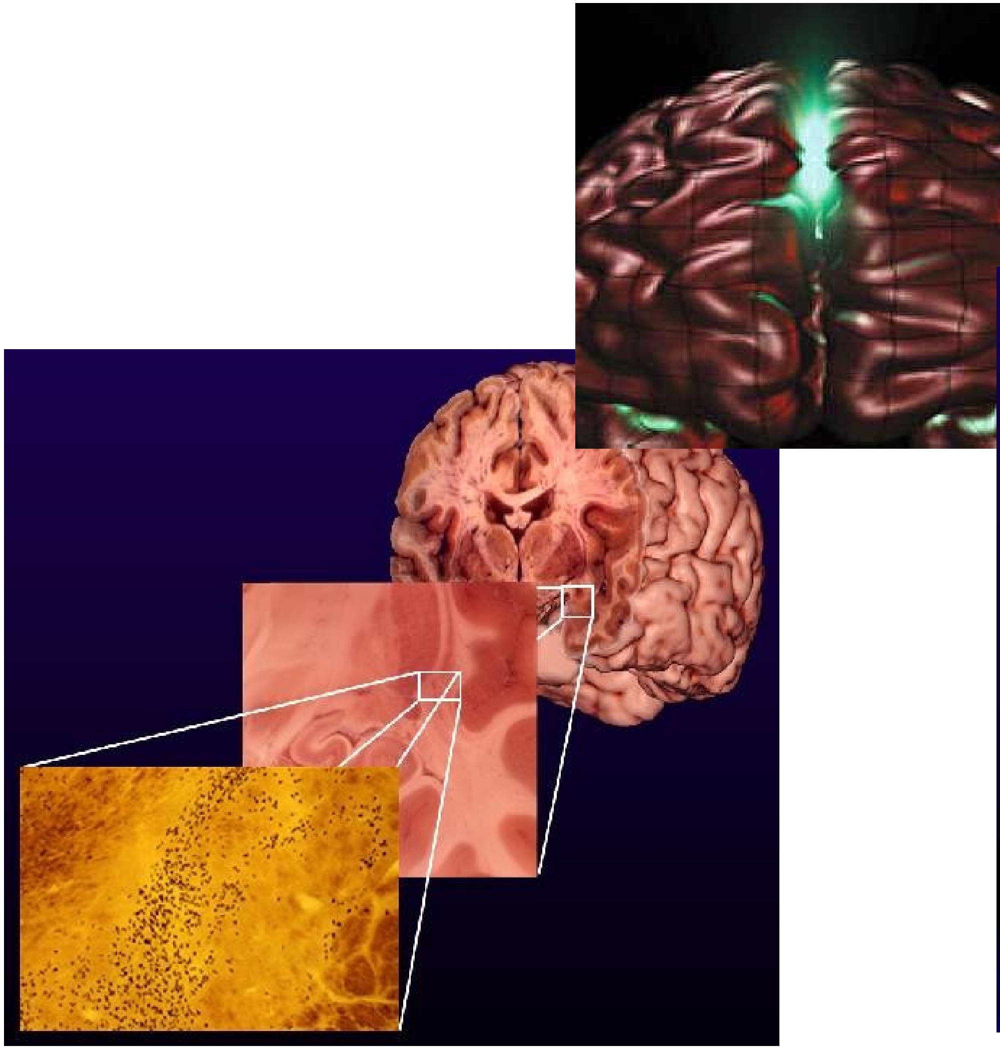
- Storage Access directly across Interconnects – Ethernet & Infiniband
- Increased Parallel Client Access to data
- Need for High Performance boost over NFS/NAS
- Requirement to process Even Larger File Sizes
- Need for Simplistic View of the FS Space

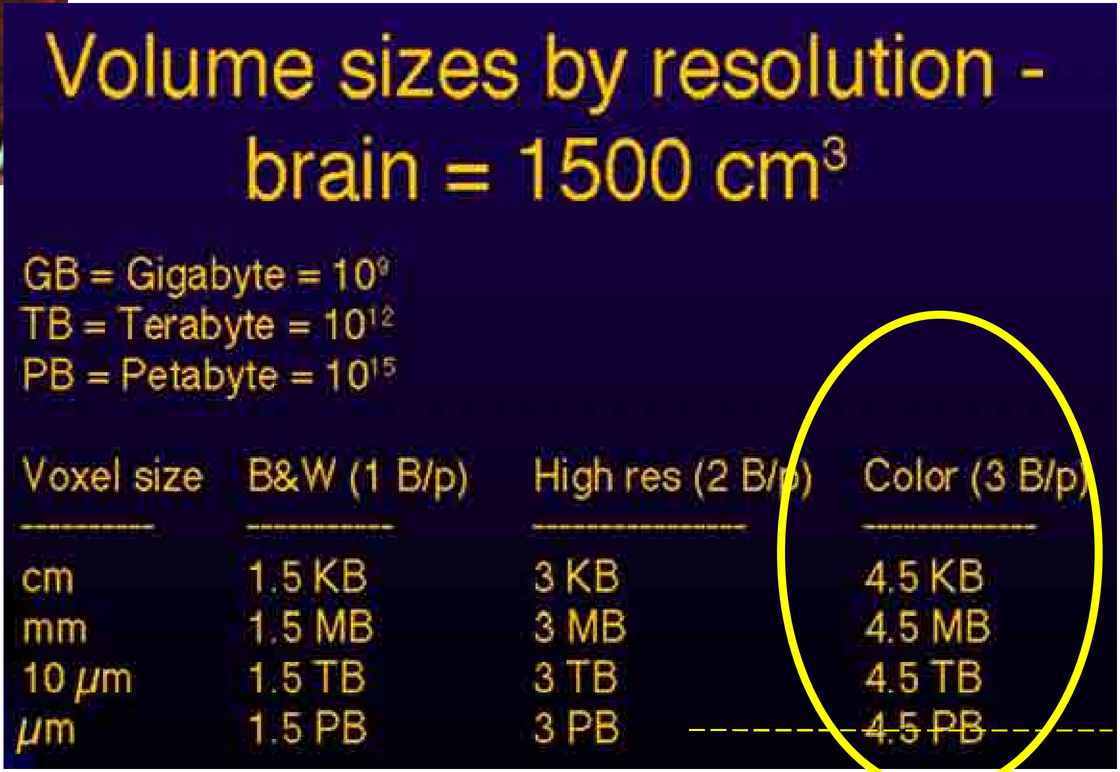
Why are Petabytes of FS Sizing needed?

BIRN's Challenges of Large Distributed Data – Human Brain

Dr. Art Toga (UCLA) was one of the first to articulate the magnitude of the challenge of human brain data - and address it!

Each Brain is Big Data and Comparisons Must be Made Between Many!



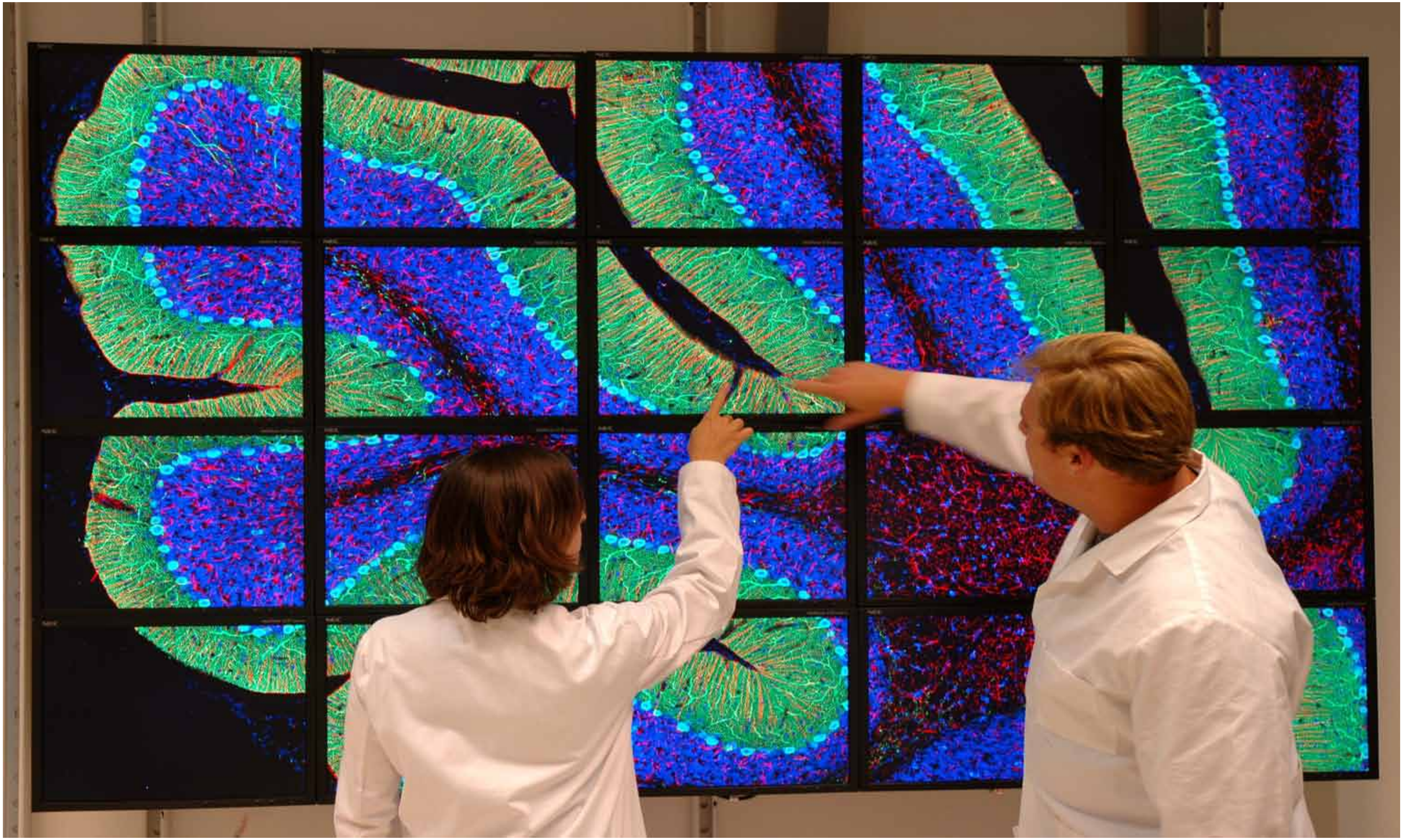


Volume sizes by resolution - brain = 1500 cm^3

GB = Gigabyte = 10^9
TB = Terabyte = 10^{12}
PB = Petabyte = 10^{15}

Voxel size	B&W (1 B/p)	High res (2 B/p)	Color (3 B/p)
cm	1.5 KB	3 KB	4.5 KB
mm	1.5 MB	3 MB	4.5 MB
10 μm	1.5 TB	3 TB	4.5 TB
μm	1.5 PB	3 PB	4.5 PB

High Resolution assists with scientific discovery but data challenges grow



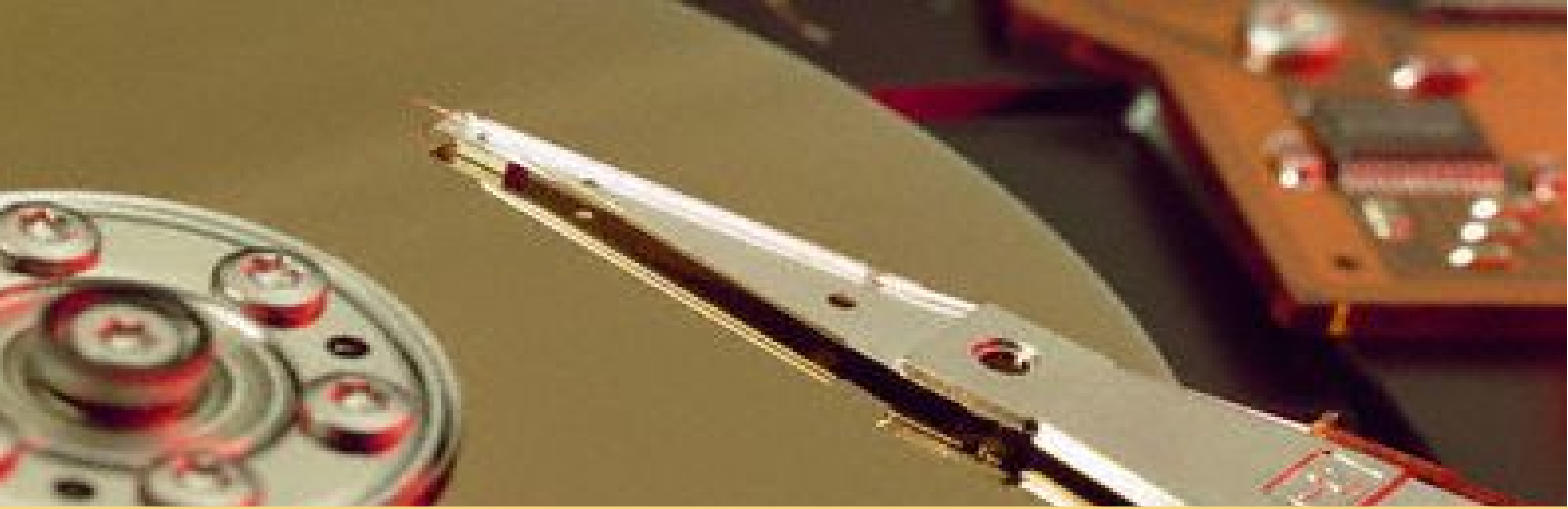
What is Sun's Thumper Offering?

High Performance Server

Thumper

- **A High-Performance 4 way Server**
  - Dual Opteron Dual-Core processors
  - Up to 16GB Memory (2GB Dimms)
- **With on Board High Density SATA**
  - 48 direct attached hot-plug SATA II drives
  - 24TB in 4 RU
- **And Enterprise Class Server RAS**
  - ILOM
  - Fans
  - Power Supplies
  - Hard Disk

n+x



Storage

Thumper SATA Disk

- SATA technology adoption is the fastest growth area in disk storage
- SATA disk for Thumper
  - > Enterprise Grade Drive with 1M hour MTBF
  - > 7,200 RPM
  - > Platter speed of 57MB/s
  - > 2GB/s throughput serial read
- Hot swap Disk
- Individual LED light for each drive
- SW RAID provides the flexibility of RAID configs and performance -- RAID 0+1, 5, 6, RAID Z, RAID Z2
  - > Not Your Father's SW RAID...

Thumper

CPU:

4way Opteron
8-16GB Mem
10 PCI-X bridges

Storage Capacity:

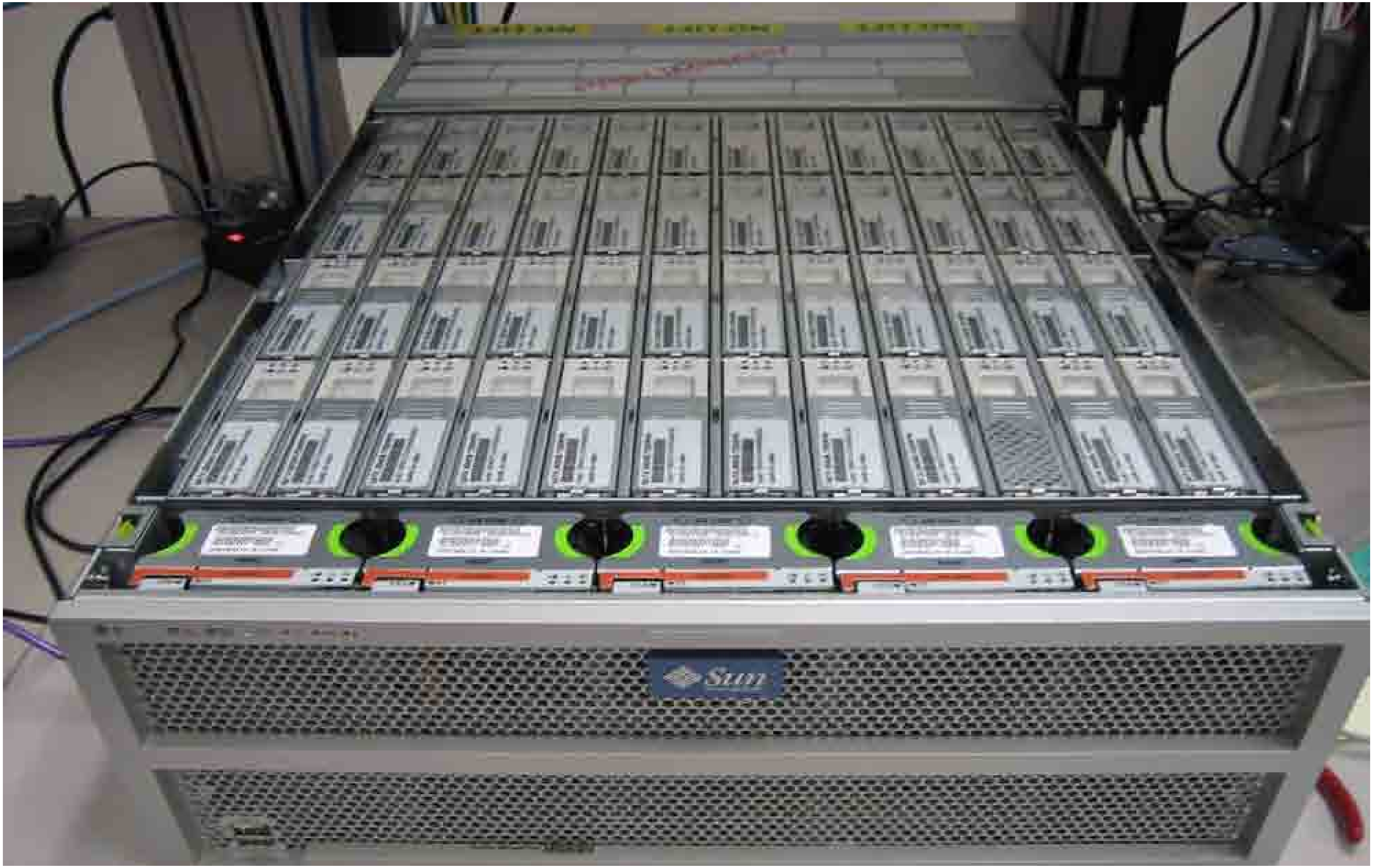
48 disks@500GB

24 TB per 4U
240 TB per rack

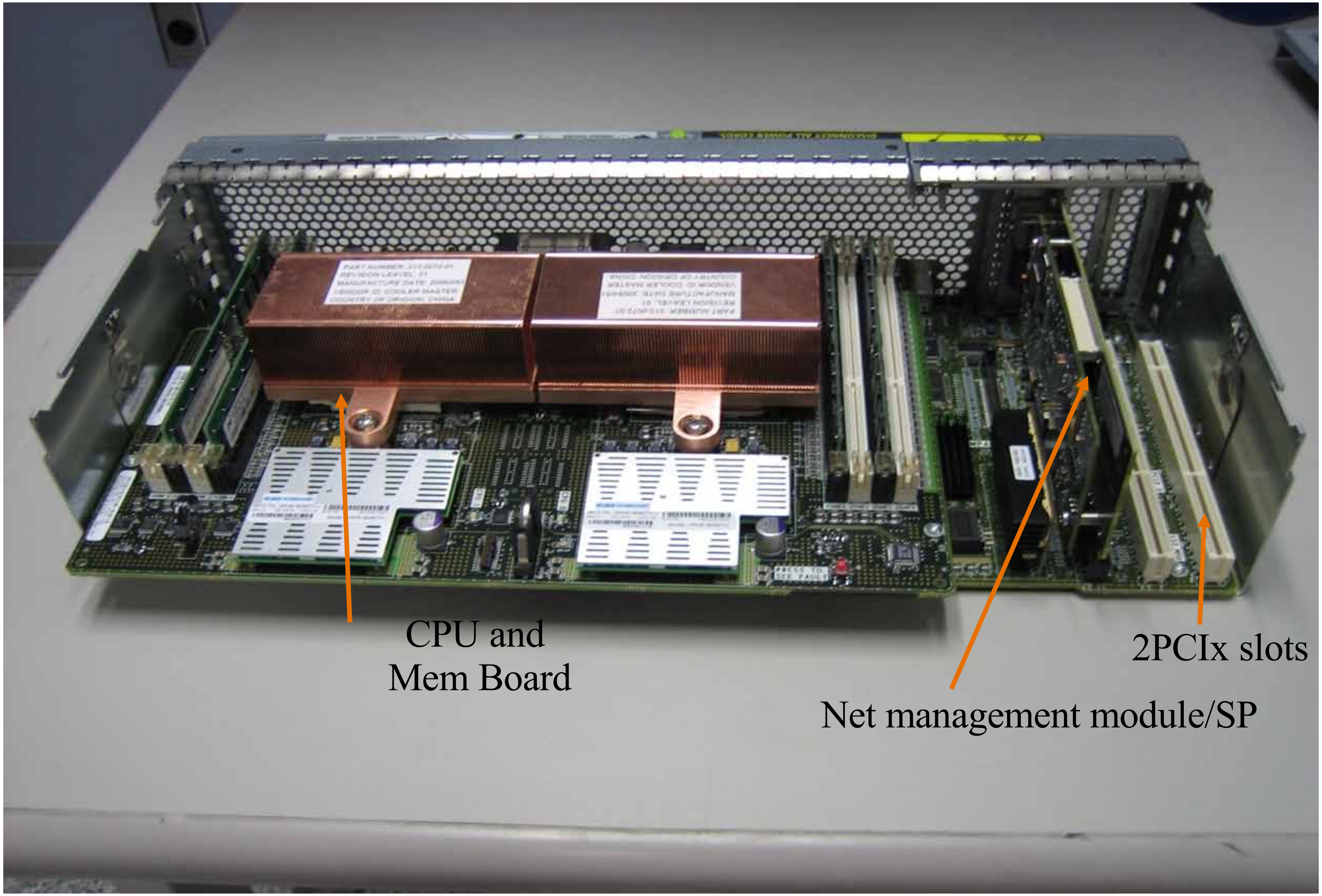
Throughput:

10 Gbps per 4U
100 Gbps per rack

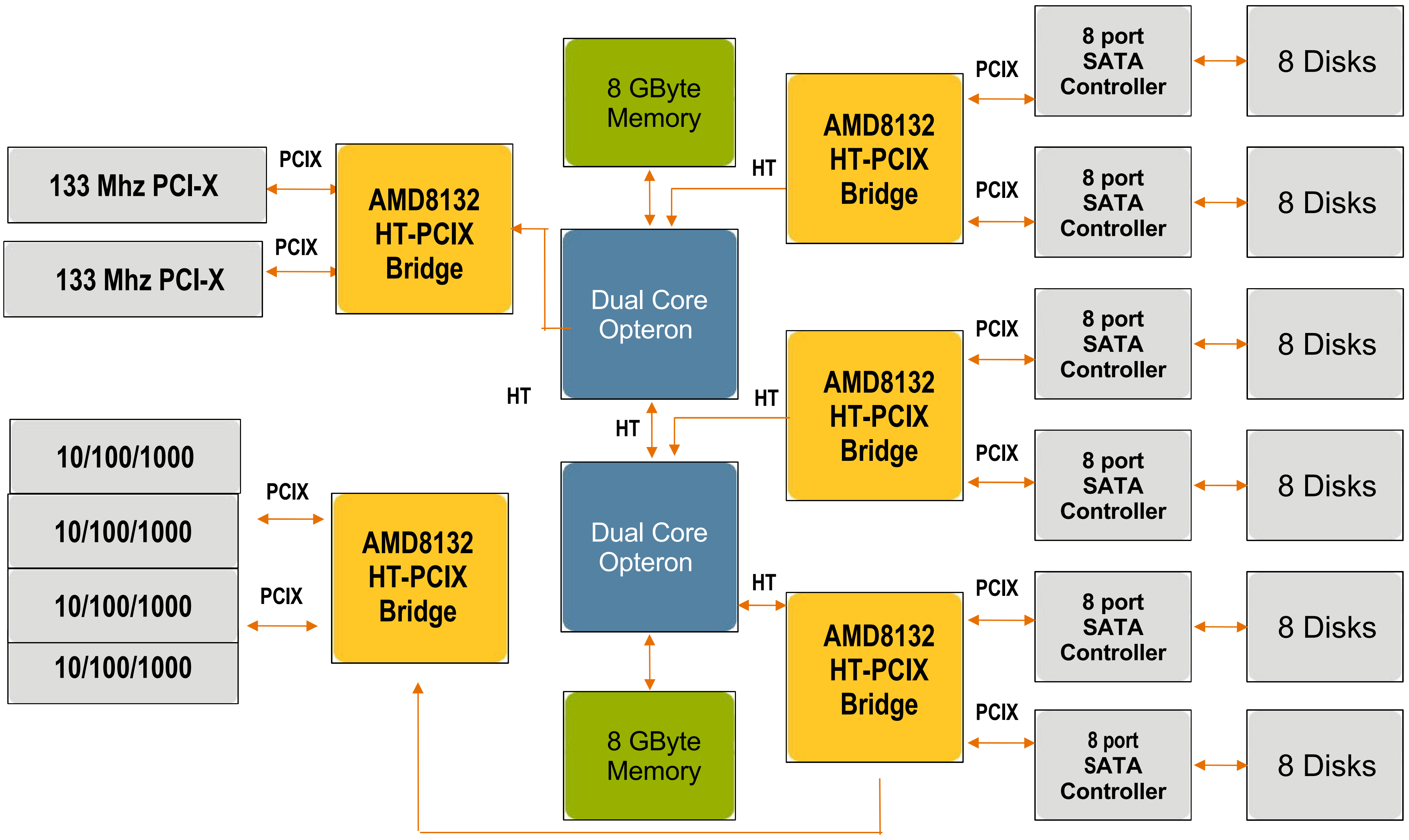
Top View of Thumper



Thumper Internals



Thumper Block Diagram



Dramatically Higher Throughput

Measured peak throughput of 2.5GB/s with ZFS.
This is peak, averaged over one second.
Over a longer period, measured 2.1GB/s.

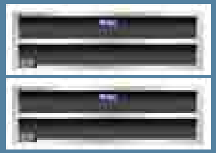
The disks have a platter speed of 57MB/s each,
for a theoretical max of 2.7GB/s
without a file system

Dramatically Higher Storage Density

48 Terabytes/Rack
1/5th density

Traditional Storage

240 Terabytes/Rack
5x the density
10x 4-way servers



Sun Thumper

Thumper Storage Capacity Roadmap

Disk Date	Disk Capacity	Raw Capacity	Net Capacity
Q3CY 05	250GB	12TByte	9.6TByte
Q3CY 05	500 GB	24 TByte	20 TByte
Q3CY 07	750 GB	36TByte	29TByte
Q4CY 07	1000 GB	48 TByte	40 TByte

Note: Net Density Assumes 80% Efficiency

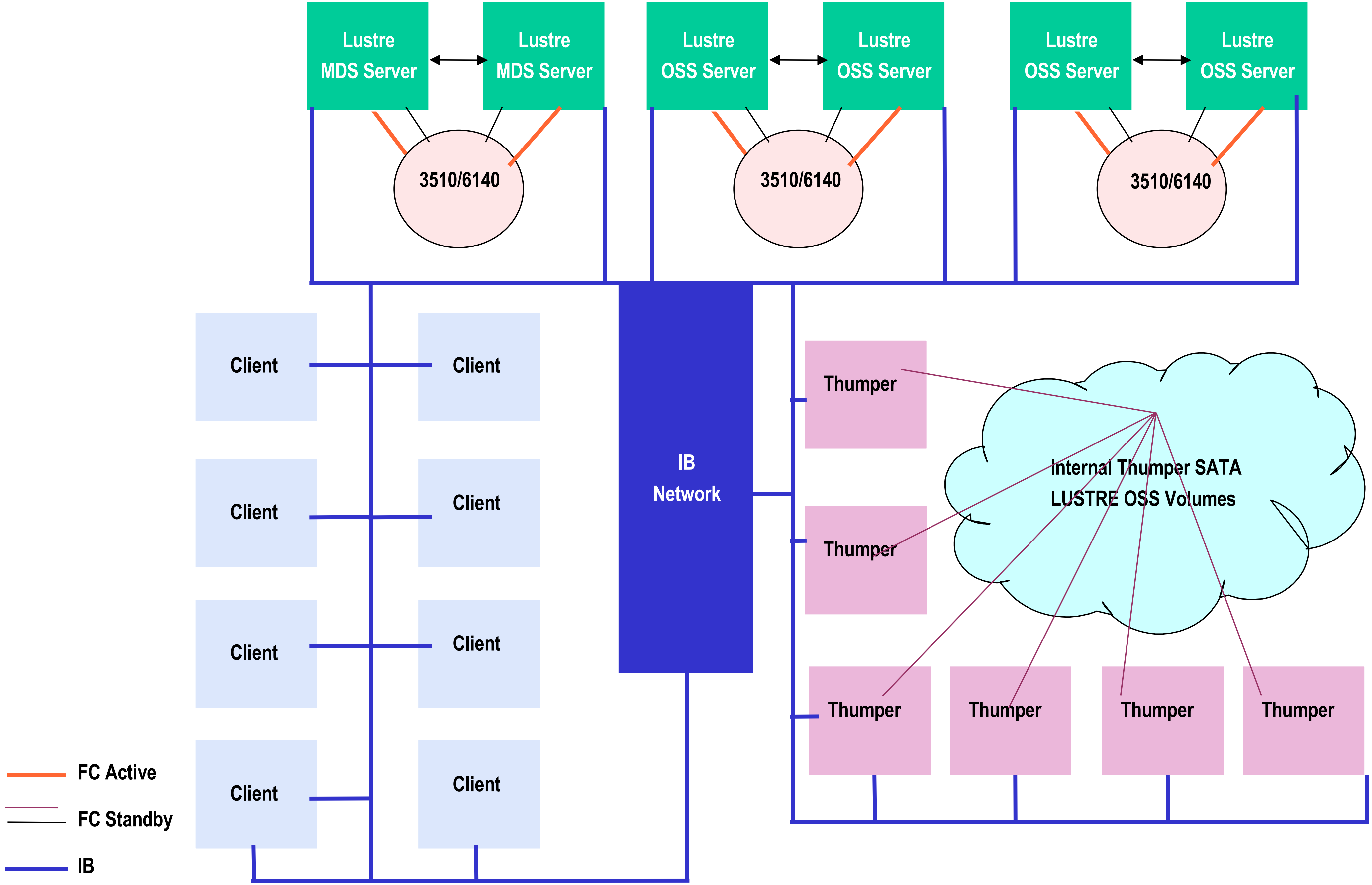
Target Market and Workloads

- HPC/Grid Computing Data Server/storage
  - > Thumper has the highest density and capacity storage for Grid storage node at very low cost, coupled with lustre scalable FS, provides one of the highest data throughputs in Grid environments
- Streaming server/storage
  - > Thumper's high bandwidth IO provides the large network connection throughput for streaming needs at extremely low cost
- Data Warehouse applications
  - > Thumper's large memory bandwidth and disk throughput coupled with ZFS makes it an ideal solution for data storing, searching and mining in a 24 TB system and scales up
- Archiving/online back up
  - > Thumper's high density, high capacity disk storage at low cost per GB provides the most economically viable solution for large archives of data online

Target Application: High Performance and Grid Computing

- Challenge
  - Data requirements for analysis and visualization have scaled beyond the capabilities of current network attached storage.
  - Cost of disk storage skyrockets as capacities grow rapidly
- Solution
  - Linux: Lustre parallel file-system provides scalability to create clusters of Thumpers
  - Thumper is one of the best OSSs in a Lustre Storage Cluster environment with enhancements being made for further features and improvements
  - Thumper provides unprecedented disk storage density and low cost.

Lustre Architecture using Sun Thumper & STK



Thumper Lustre Storage Pool

Where has Sun Deployed This Architecture?

Some Great Things happening at TITECH!

TITECH, Sun, and CFS Implemented Thumper with Lustre

Titech Deployment – 655 16way-Galaxy4, 42 Thumpers, 1PB

Thumpers Racked at TITECH

TITECH Thumper

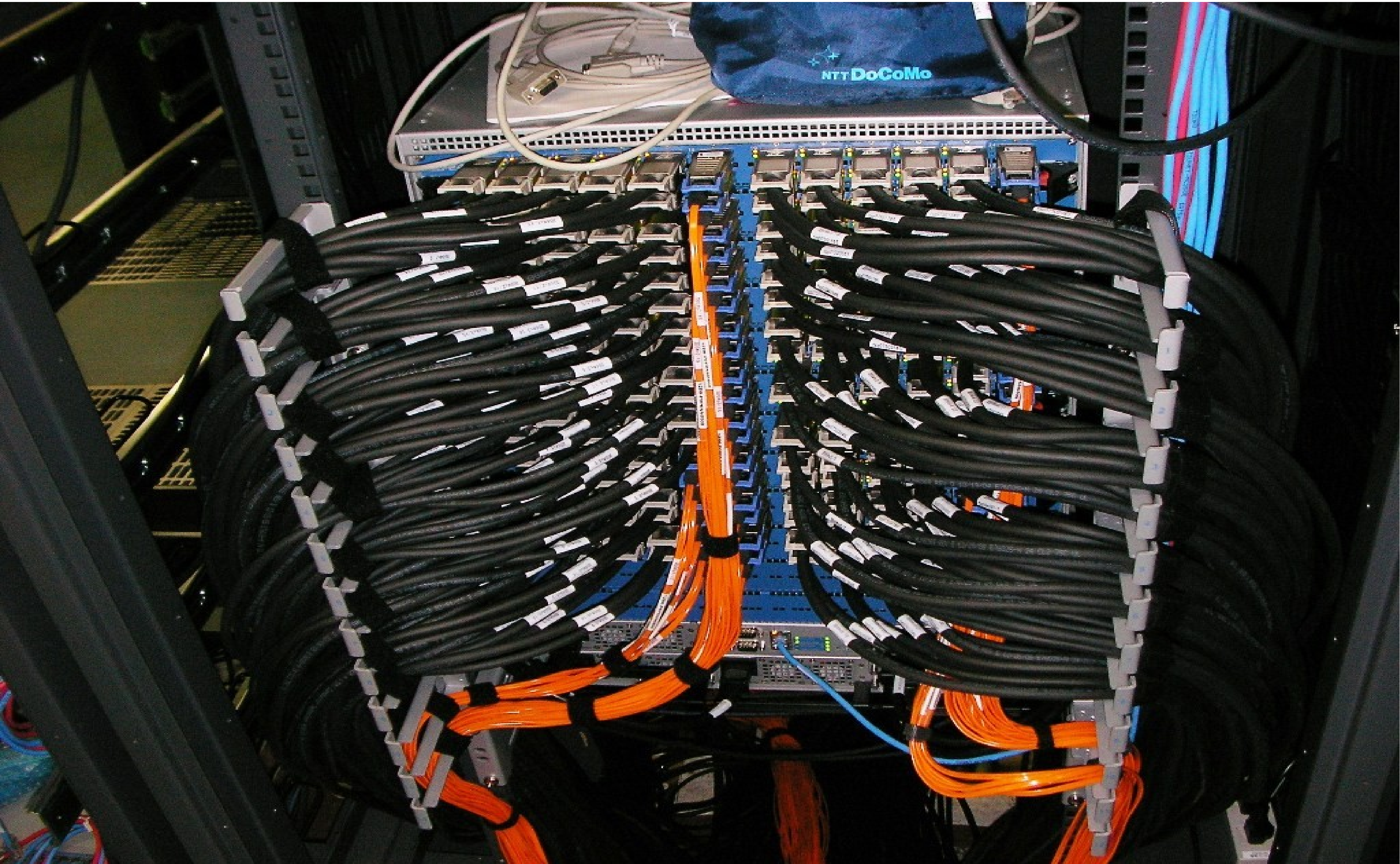
42 Systems

Accessed via IB

Flexible Storage Pool

Flexible Number of File Systems for specific File I/O

Single IB Network – Servers and Thumpers



Recent Lustre and Thumper Performance Results

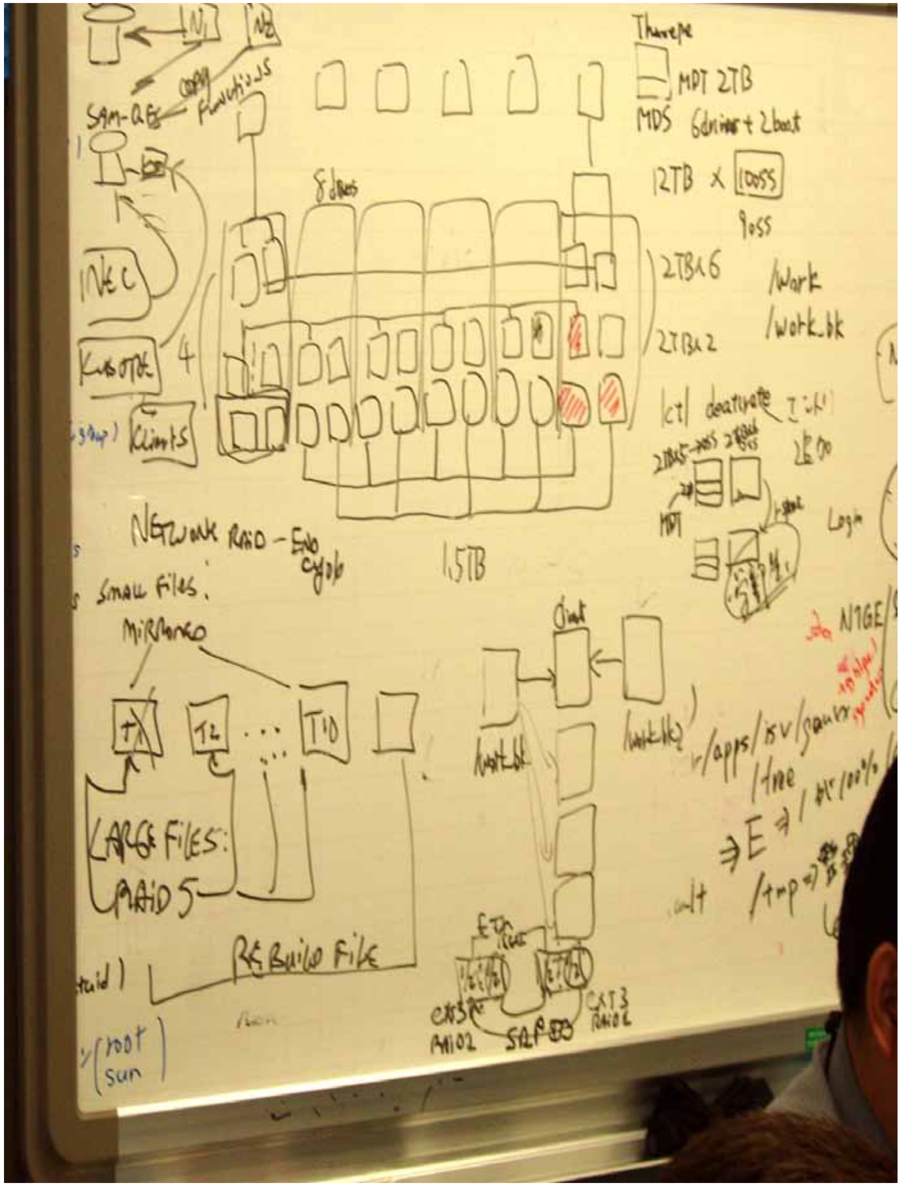
Titech and CFS testing on Thumper thus far reveals:

- 2GB/s read 1.25GB/s write throughput on 42 Drives
  - > Further Work – in - progress
- Close to 1GB/s out of Thumper

  with IB on Raid 5 Lustre FS
- 10GB with 16 Galaxy 4 clients and 10 Thumpers

Future WIP on RAID 6 and OFED IB

- > Further performance specs will come



Where else is Sun Deploying This Architecture?

Sun has deployed this architecture and is focusing on deploying further at:

- ARSC
- Brazil
- DKRZ
- KISTI
- TACC – Karl will discuss further
- Many Other Future Bids..

Thumper Real Life Results at ARSC

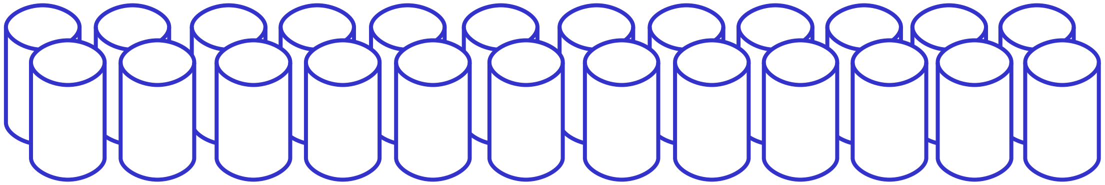
- Customer utilized 8 processes on 15 x4600 clients for a total of 120 writers and 6 Thumper OSS
- Test covered **writes**
- IB Infrastructure – Voltaire Driver with support for one HCA
- IOZONE report from ARSC showed BW throughput on data writes of 4683420.05 KB/sec as shown on next page

Thumper Real Life Results at ARSC--IOZONE

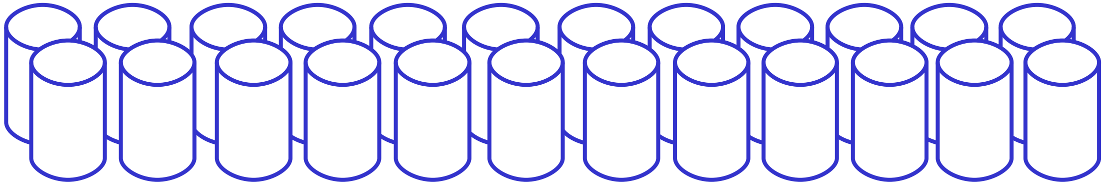
```
Command line used: /wrkdir/mitchell/I.A.3.a/iozone -w -e -M -t 120 -s 1g -r 512K -i0 -+m
    /wrkdir/mitchell/I.A.3.a/client.list

Output is in Kbytes/sec

Time Resolution = 0.000001 seconds.

Processor cache size set to 1024 Kbytes.

Processor cache line size set to 32 bytes.

File stride size set to 17 * record size.

Throughput test with 120 processes

Each process writes a 1048576 Kbyte file in 512 Kbyte records


Test running:

Children see throughput for 120 initial writers    = 4683420.05 KB/sec

Min throughput per process                         =   25979.76 KB/sec

Max throughput per process                         =   59258.05 KB/sec

Avg throughput per process                         =   39028.50 KB/sec

Min xfer                                   =  460800.00 KB
```

Sun's Three Tier Storage Architecture

Sun's HPC Three Tier Storage Architecture

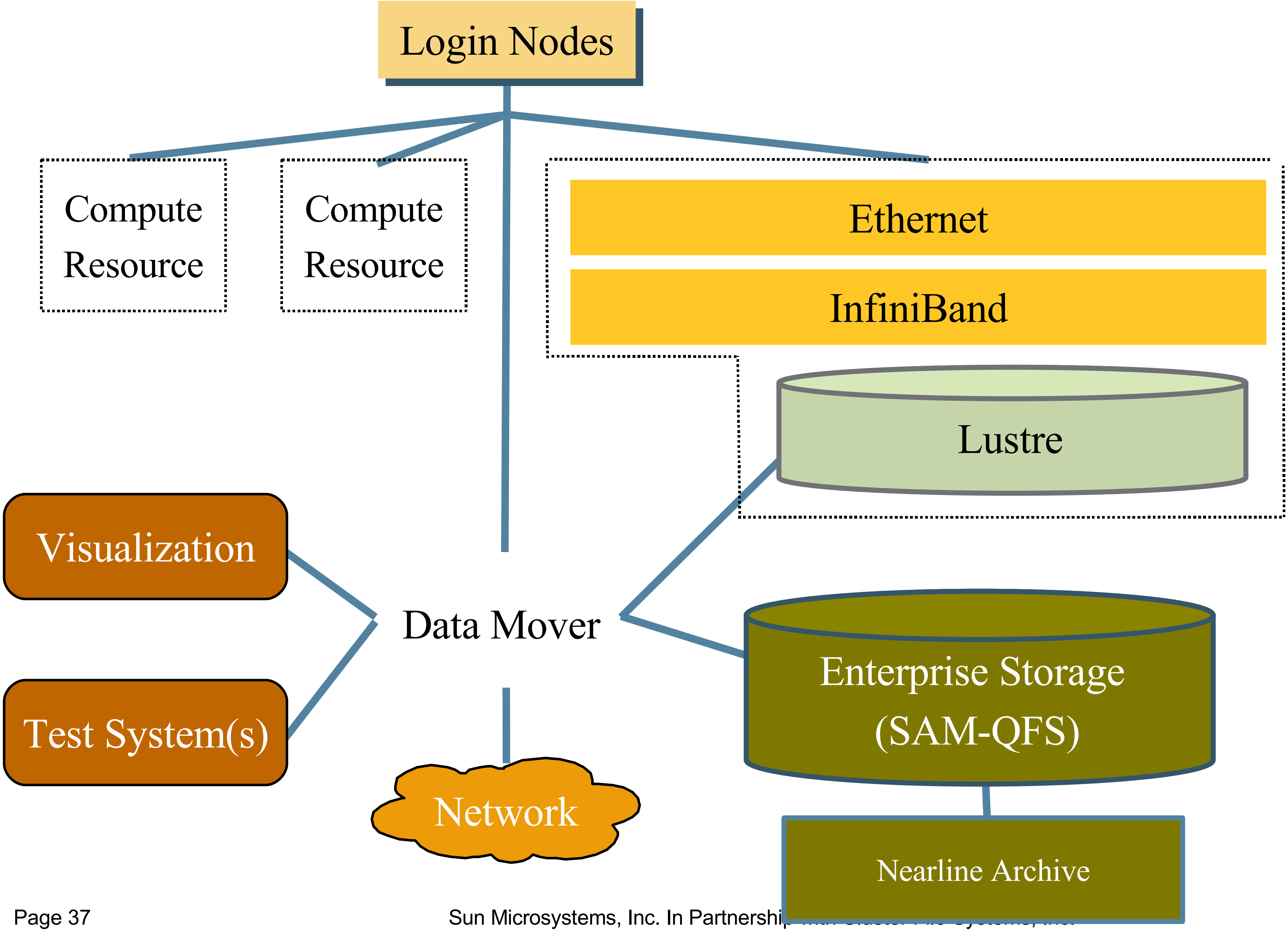


High Speed, High I/O
Computational Facing
PFS



Medium Speed Parking Space
For Post Processing

Low Speed Archival Facility
For Data Life Cycle Management



Login Nodes

Compute Resource

Compute Resource

Ethernet

InfiniBand

Lustre

Visualization

Test System(s)

Data Mover

Network

Enterprise Storage (SAM-QFS)

Nearline Archive

Observations and Direction

- Customers are re-architecting workloads for horizontal scale to enable lower cost deployments
- Customers want direct access of their data across Interconnects such as Ethernet & Infiniband
- Use of Lustre coupled with with Sun Storage Offerings provides a winning combination from a competitive perspective
- Sun and CFS are working jointly on running Lustre on Solaris/ZFS
- Sun, CFS, and TACC are working jointly on further enhancements of SW RAID
- Sun, CFS, Mellanox, and TACC are working jointly on IB OFED improvements with Thumper

DATA Intensive Computing:

Sun's HPC I/O Strategy

Presented at

Lustre User Group

04/23/07

Larry McIntosh

Thanks..

Sun Microsystems, Inc.