UT-BATTELLE

A Center Wide File System Using Lustre and Recent I/O Benchmark Results on XT3

Shane Canon
Oak Ridge National Laboratory

CUG 2006
Lugano, Switzerland

May 11, 2006

National Center for Computational Sciences

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

Outline

- Overview of NLCF
- Motivation for a Center Wide File System
- Initial Plan for Spider
- Experience to Date
- Recent Results on the XT3
- Future Plans

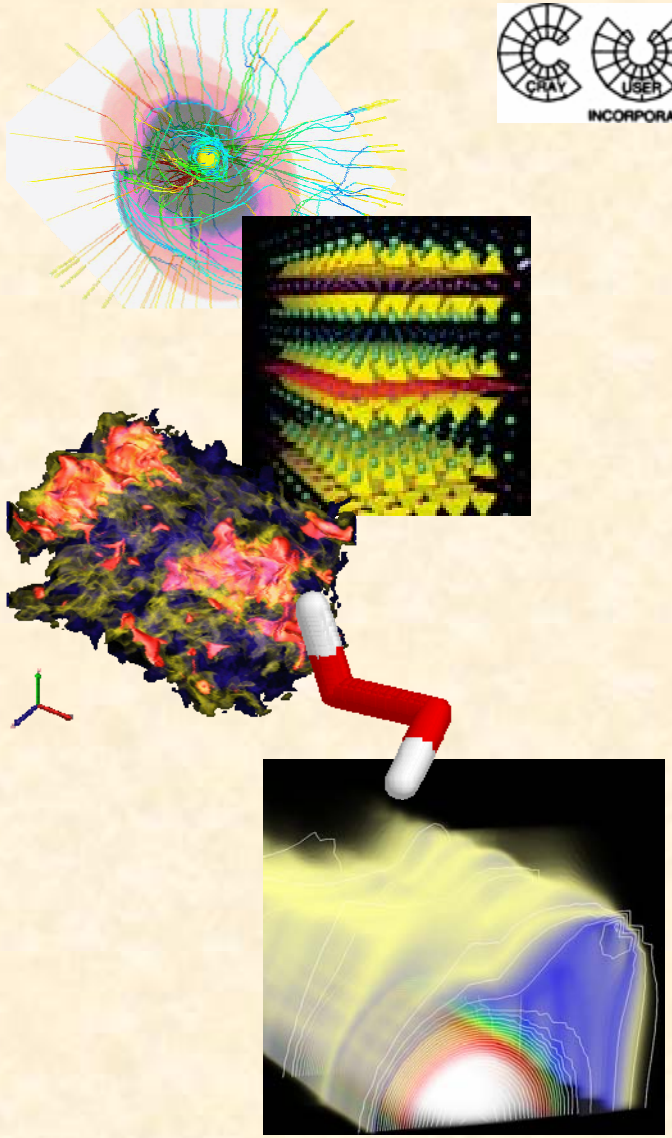
Mission of the NLCF

Provide Leadership Class Computing to enable breakthrough science

The goal of leadership systems is to provide computational capability that is atleast 100 times greater than what is currently available.

Users of the NLCF

- Diverse set of disciplines from the Office of Science
  - Astronomy
  - Chemistry
  - Climate
  - Combustin
  - Fusion
  - Material Science
- INICTE Program includes science and industry



NLCF Resources

Computing

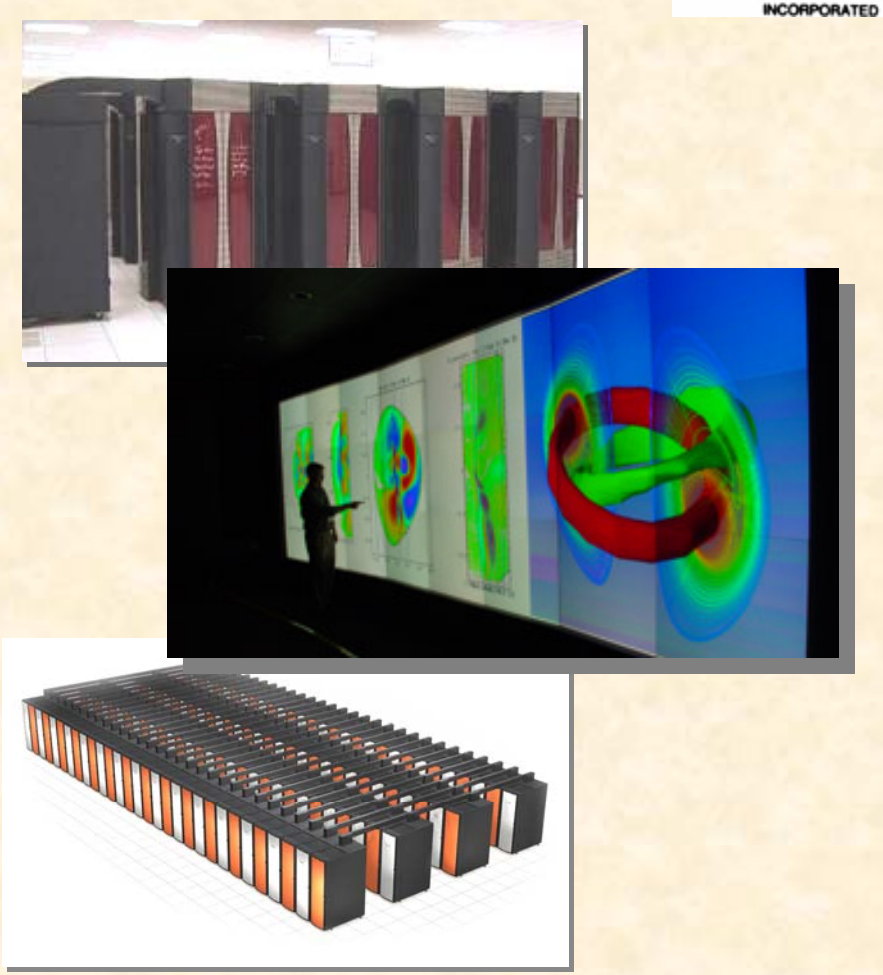
- Jaguar – XT3
- Phoenix – X1E

Data Analytics

- Ewok – End-to-end
- Ram – SGI Altix
- Everest – PowerWall
- Hawk – Viz Cluster

Storage

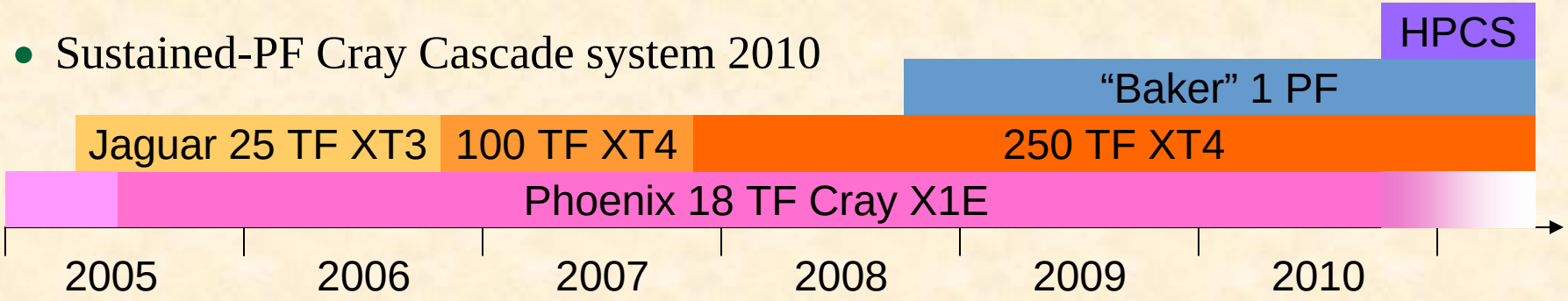
- HPSS
- Spider



ORNL Milestones:
Deliver 1PF system in 2008
Deliver 250 TF by 2007

Roadmap

- Upgrade existing 25 TF XT3 to dual-core 100 TF system in 2006
- Upgrade 100 TF to 250 TF in late-2007
- Deploy 1 PF Cray “Baker” late 2008
- Sustained-PF Cray Cascade system 2010



- 18 TF Cray Phoenix and 25 TF Cray Jaguar currently in production

Motivation for Central File System

- Common file system
  - less file movement
  - Users are already struggling with this
- Leverage hardware
  - Bandwidth is expensive
  - Even more critical with PF system where 100s-1000s of GB/s are needed

User Quote

From a recent User Survey…

What are the current time-intensive bottlenecks for your work flow process? What might this process be for you in 5 years (e.g,. with >1 PF)?

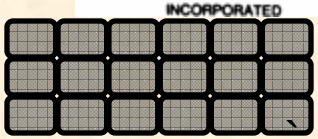
Managing and analyzing data is still (and likely to remain) the bottleneck. The proposal to have a ***common parallel file system*** *and a commodity cluster that can be used for interactive analysis and visualization* ***could have a major impact on this bottleneck****.*

Architecture for a Common File System

Phoenix
Cray X1E

NFS Servers

Grow central storage; disk and bandwidth



Data Analysis
& Visualization

Jaguar
Cray XT3

Increase HPSS bandwidth

HPSS

Increase WAN bandwidth

ESnet, USN, TeraGrid, Internet2, NLR

Late 2006
- 100 TB
- 10 GB/s (aggregate)

2008
- 1-10 PB
- 300-750 GB/s (aggregate)

Baker

Initial Spider System

- 20 OSSs and 1 MDS
  - Dual Dual-core Opteron
  - 8 GB RAM
  - Dual Port 2Gb Fibre-Channel
  - 10Gb Ethernet (PCI-X)
- Connected to Force10 E1200
- 2 DDN 8500 Couplets

Initial Configuration and Testing

- Initial role out was smooth
- Mounted on Hawk Viz Cluster
- Cluster connected by 2 10Gb links connected to aggregator switches
- Viz nodes connected by 1Gb links
- Testing demonstrated saturation of both the 1Gb links and effectively use the 2 10 Gb links
- Used this for testing up to 40 way

Performance on SGI Altix

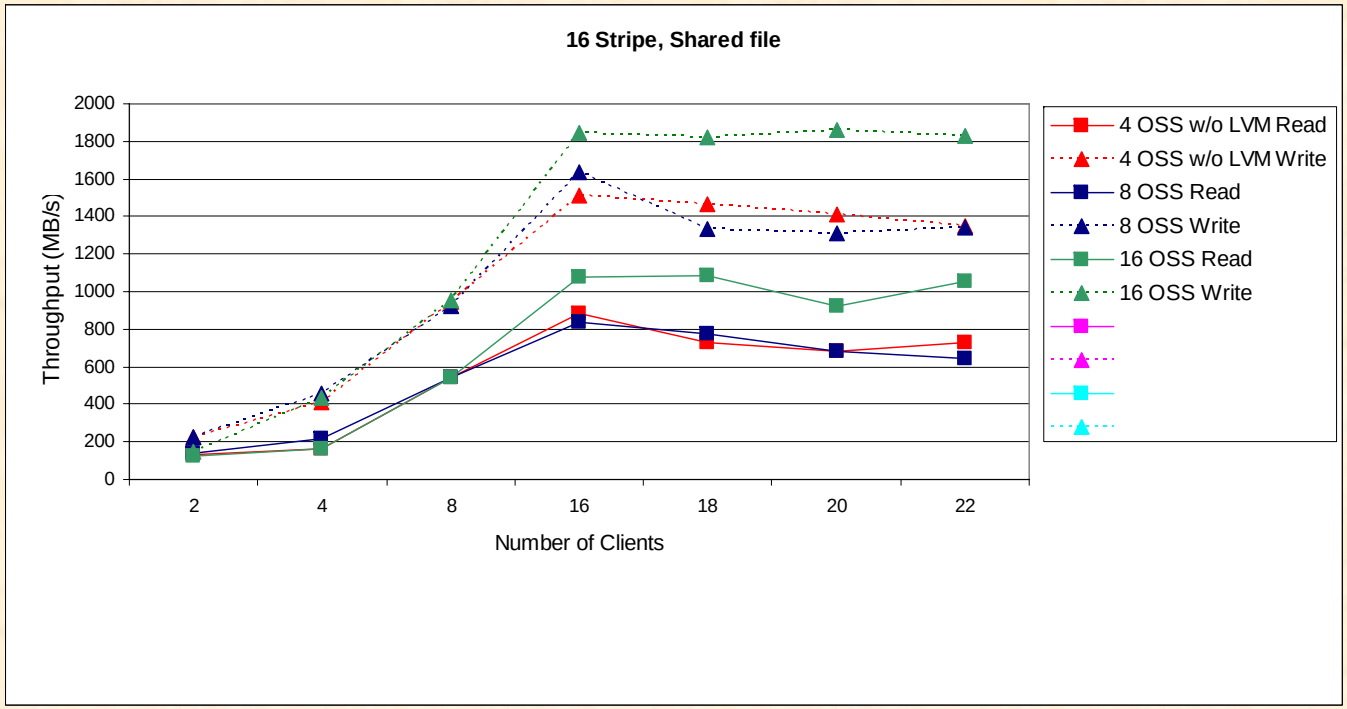
- Single interface bandwidth is constrained (roughly 6.5 GB/s on a PCI-X based 10 Gb card)
- Need to effectively stripe traffic across multiple network adapters (open question)
- Single kernel means that lustre client must be well optimized for large SMP-like system. This is in sharp contrast to the MPP and clusters systems that have seen the most investment for Lustre.
- Still unclear how well Lustre can scale on this type of system and how much we will invest in it.

LNET – A New Networking Layer for Lustre

- CFS developed a routing capability for Lustre under a contract with ORNL
- Developed LNET
- Standard in 1.4.6 now
- Allows routing from multiple networkings, including the XT3 SeaStar network

Study of OSS/OST balance

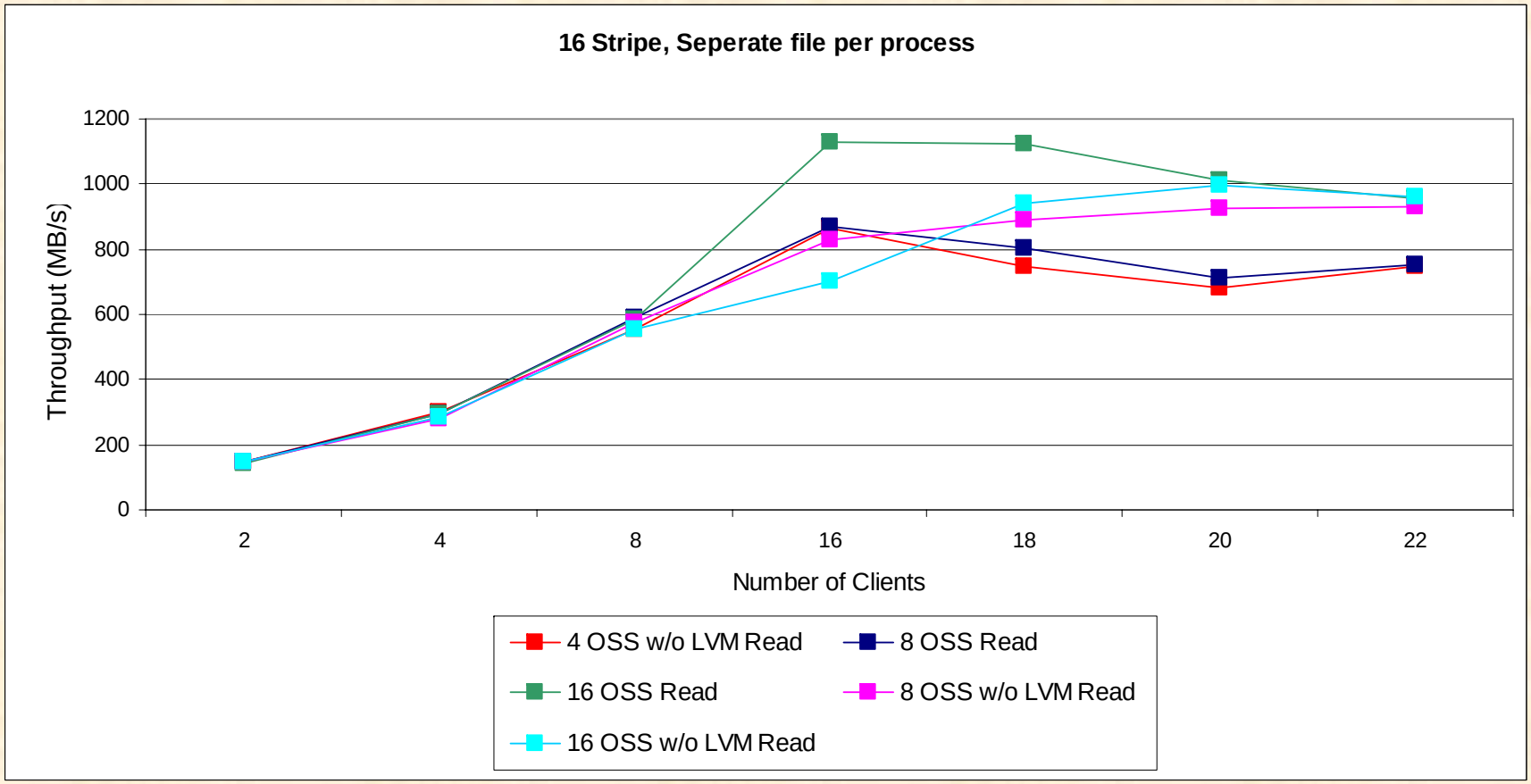
- Question: What is a reasonable number of targets to host on a single OSS?
- A: For spider nodes, small difference until clients approached the number of targets.



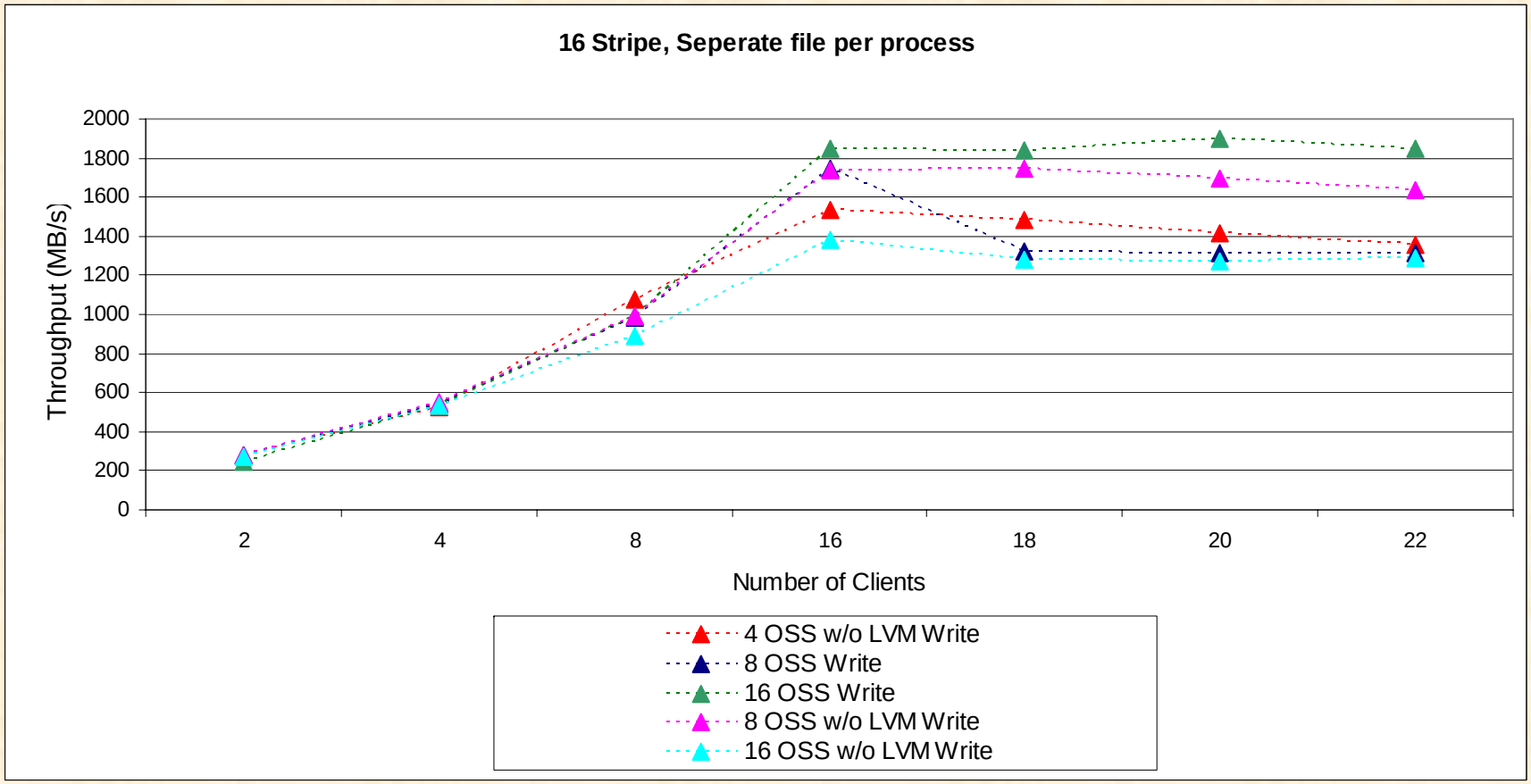
Impact of LVM on performance

- Question: Does running LVM on top of target devices impact performance?
- Yes….
- but the impact can be both positive or negative
- Preliminary results indicate LVM imposes a noticeable negative impact on performance (~33%) and scaling at 2 OSSs per OST. But had a small positive impact with one OST per OSS.
- Also, encountered stability issues with high number of OSSs/OSTs.

Plots of LVM impact (read)



Plot of LVM impact (write)



Future Plans

- IB Testing
  - DDN 9500 using SRP
  - IB network between systems
- Commodity storage testing
- Portals function shipping project
- HPSS Integration

UT-BATTELLE

Recent I/O Benchmark Results on XT3

National Center for Computational Sciences

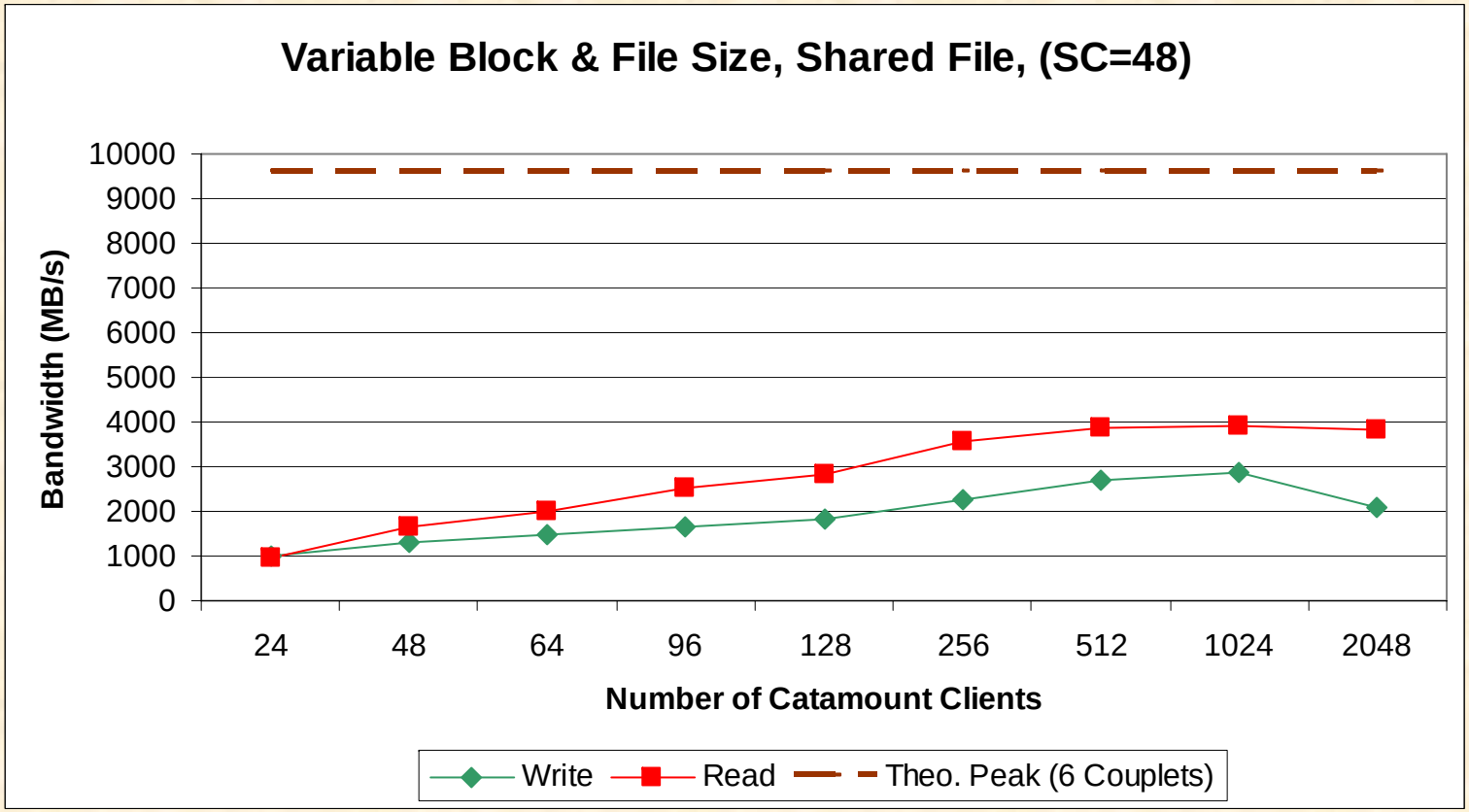
OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

Recent Results on Jaguar

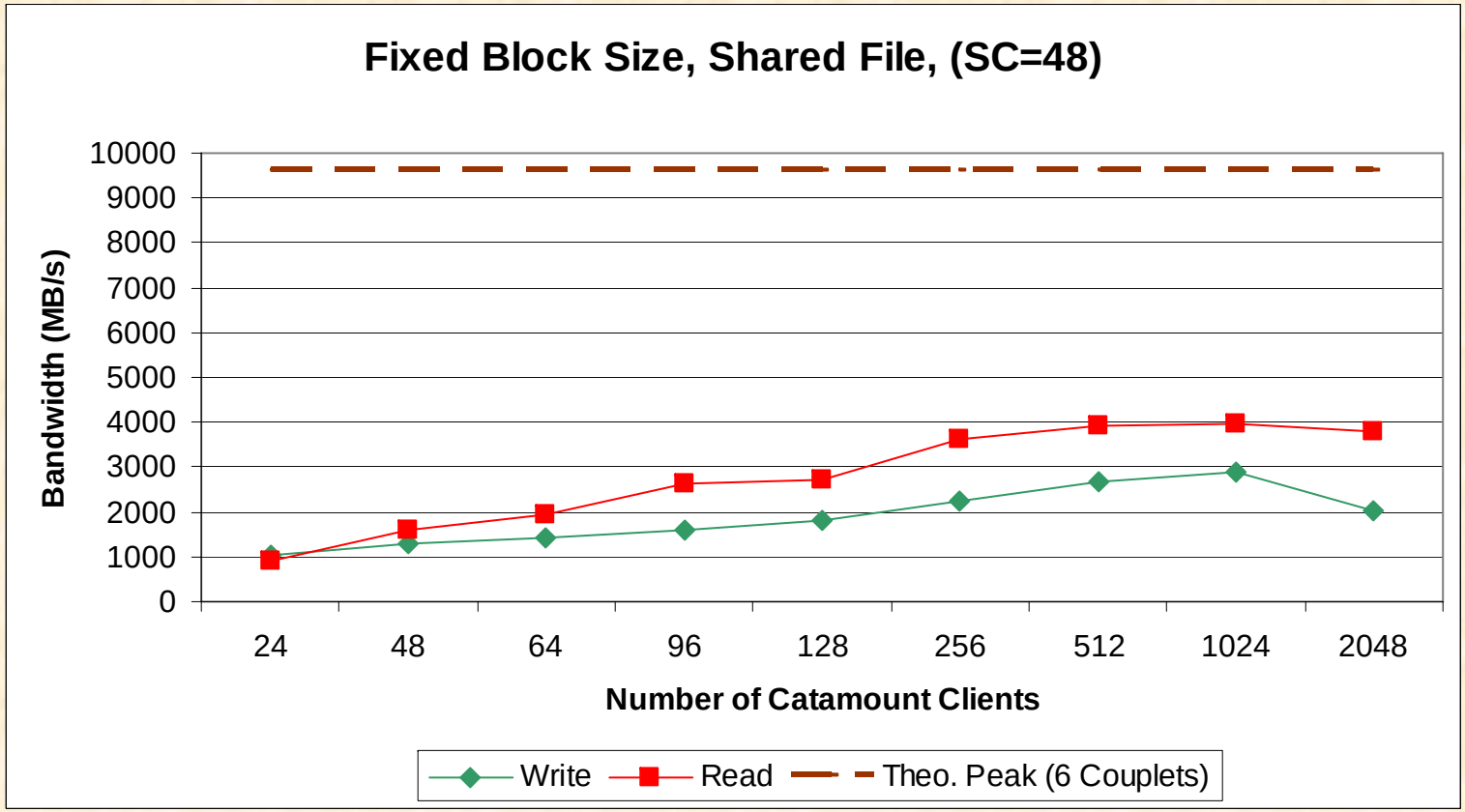
- All tests are against a
  - 24 OSS/48 OST file system
  - DDN 8500 with Fibre Channel disks

- Test were done using IOR
  - Combination of single and multiple files
  - Varying block sizes
  - Varying stripe count

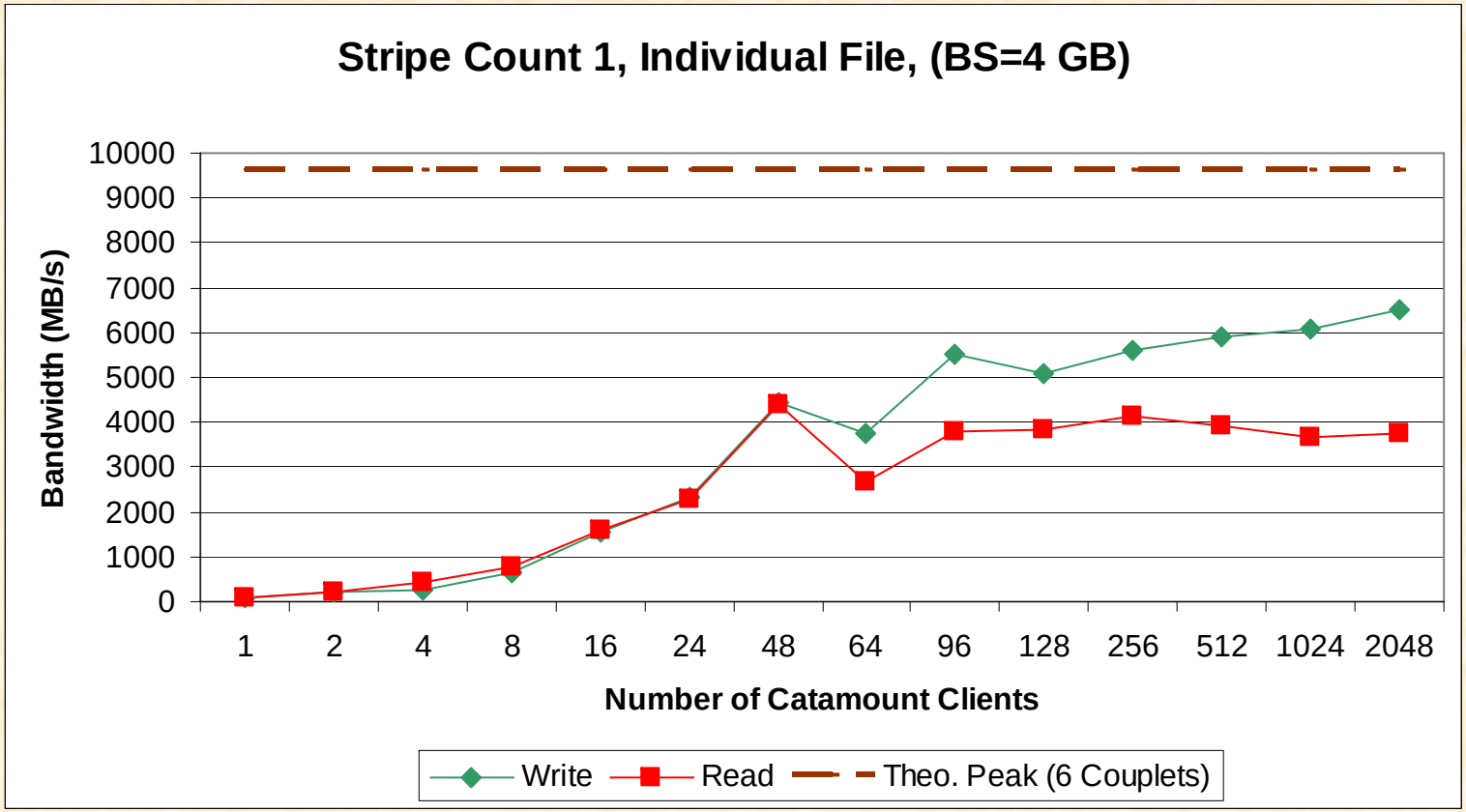
XT3 Benchmark



XT3 Benchmark



XT3 Benchmark



Acknowledgements

Oak Ridge National Laboratory

- Josh Lothian
- Don Maxwell
- Sarp Oral
- Sergey Shpanskiy
- David Vasil

Cluster File Systems, Inc.

- Peter Bojanic
- Michael MacDonald