Red Storm IO Performance Analysis

IEEE Cluster 2007

17 Sept

James H. Laros III

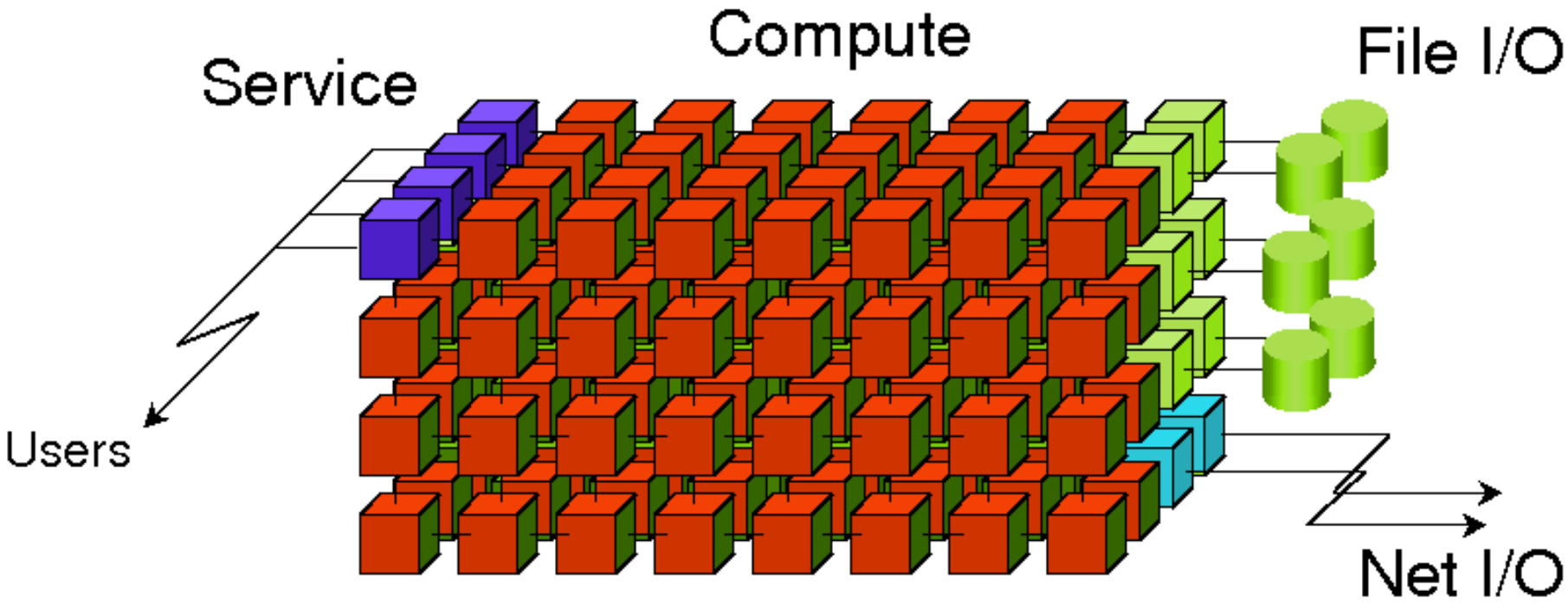
Sandia National Labs

jhlaros@sandia.gov

http://www.cs.sandia.gov/RSIOPA

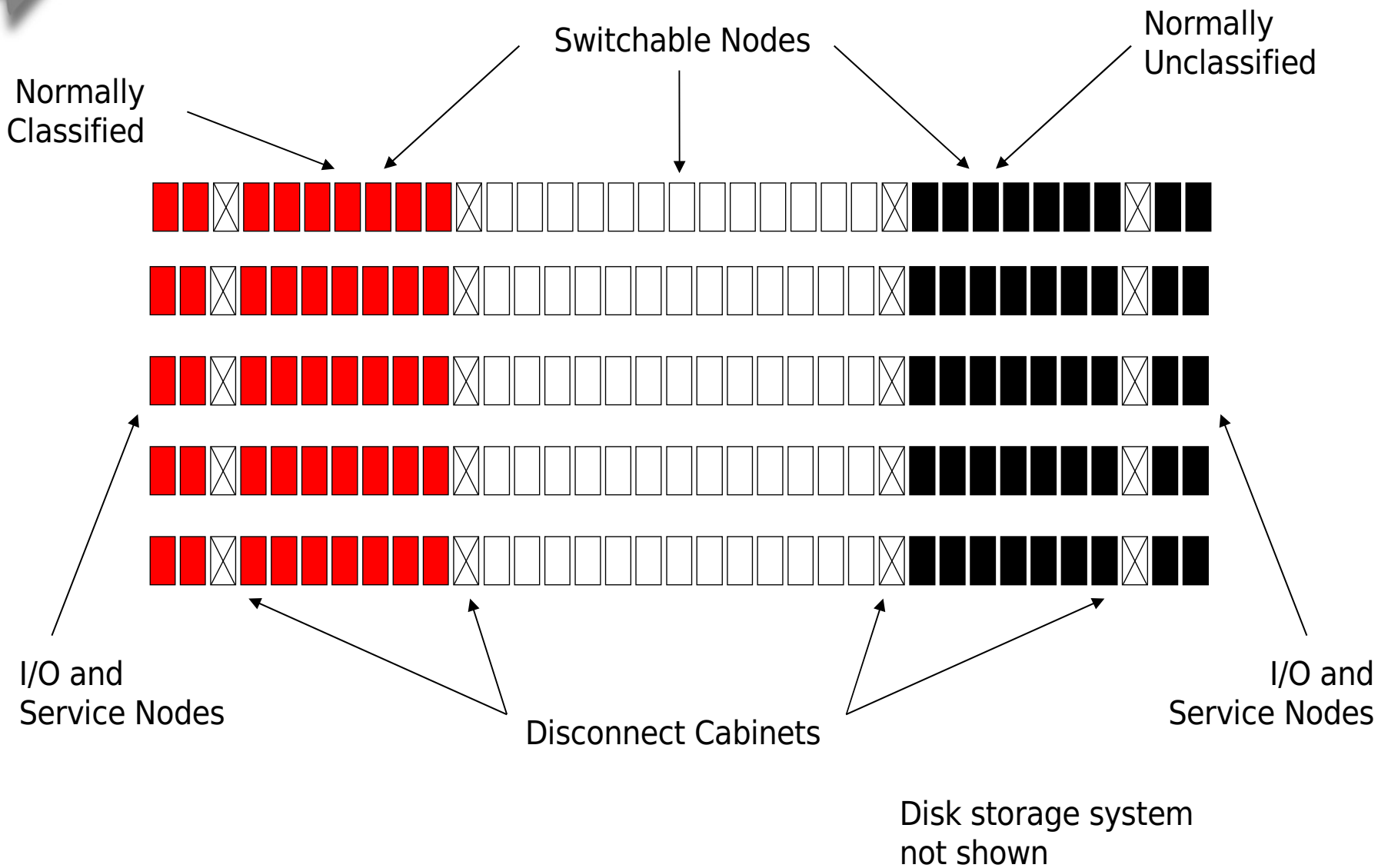
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Red Storm Architecture (Logical View)



Sandia National Laboratories

View From the Cheap Seats



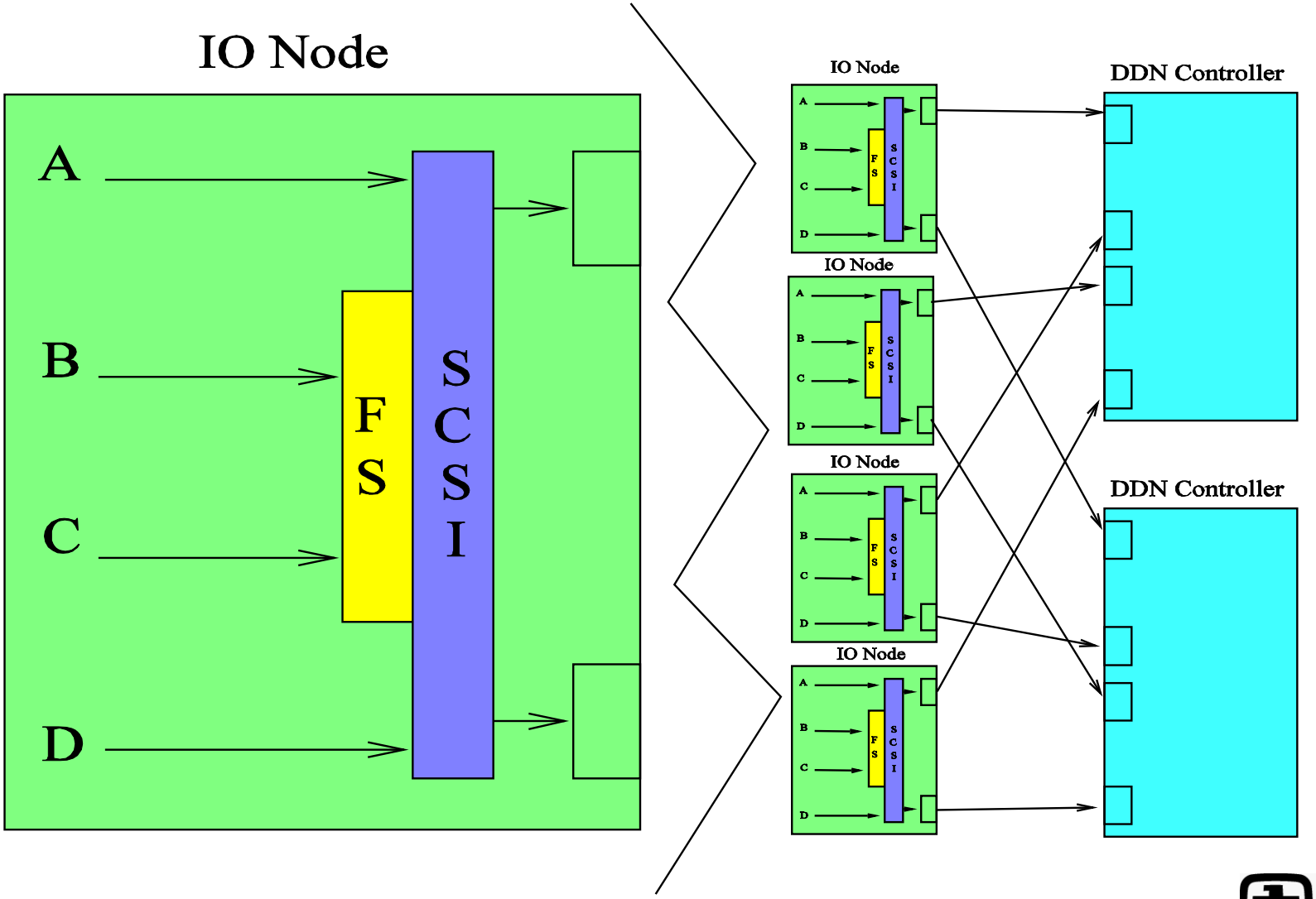
Sandia National Laboratories

Some Important Numbers

- 12960 nodes (2.4Ghz Opteron dual core, 25920p)
- 3D Mesh Topology
- 3.6 TB/sec Bisection Bandwidth
- 2.1 GB/sec Individual Link Speed (unidirectional)
- Light Weight Kernel (Catamount) – Compute
- Cray Modified SUSE Linux – IO
- Qlogics 2300 2Gb dual port HBA
- Data Direct Networks S2A 8500 controllers
  - 4, 2Gb ports per controller

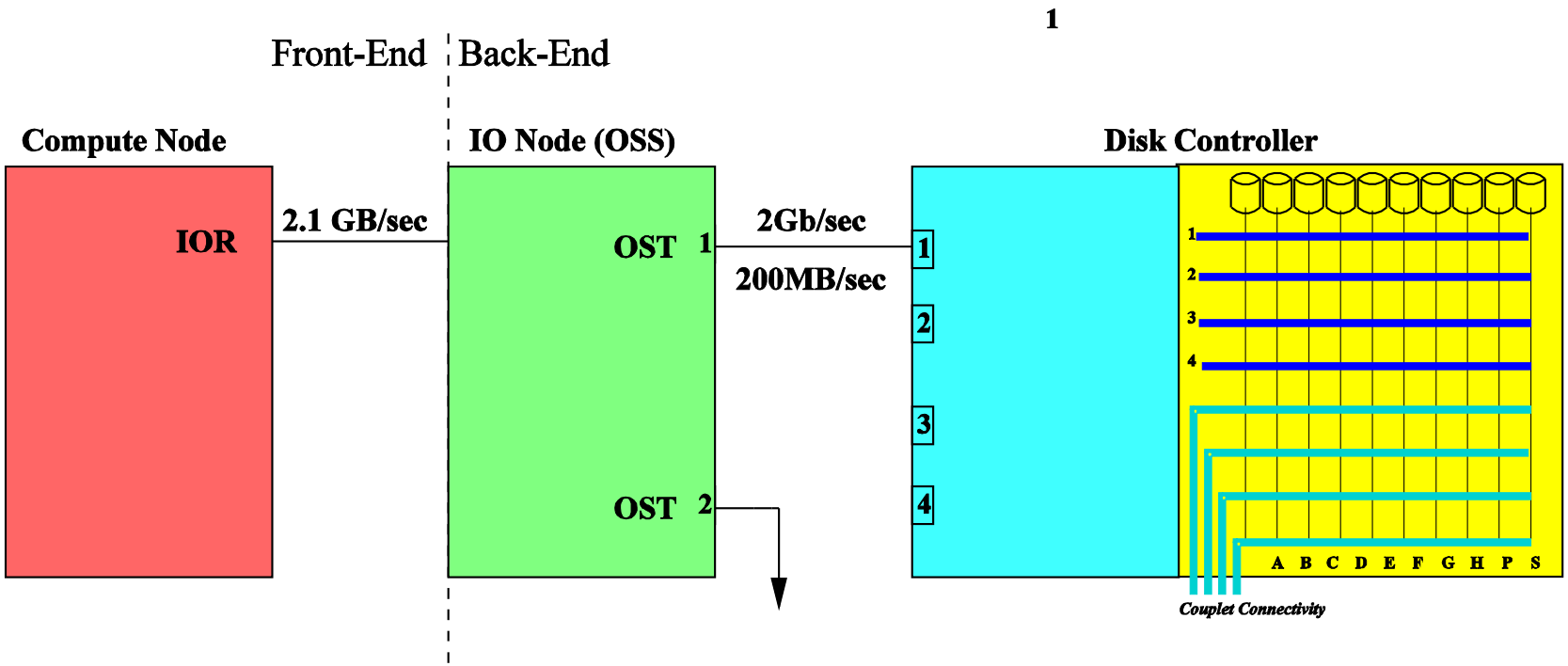
Sandia National Laboratories

IO Node/Controller Configuration



Sandia National Laboratories

End to End IO Path

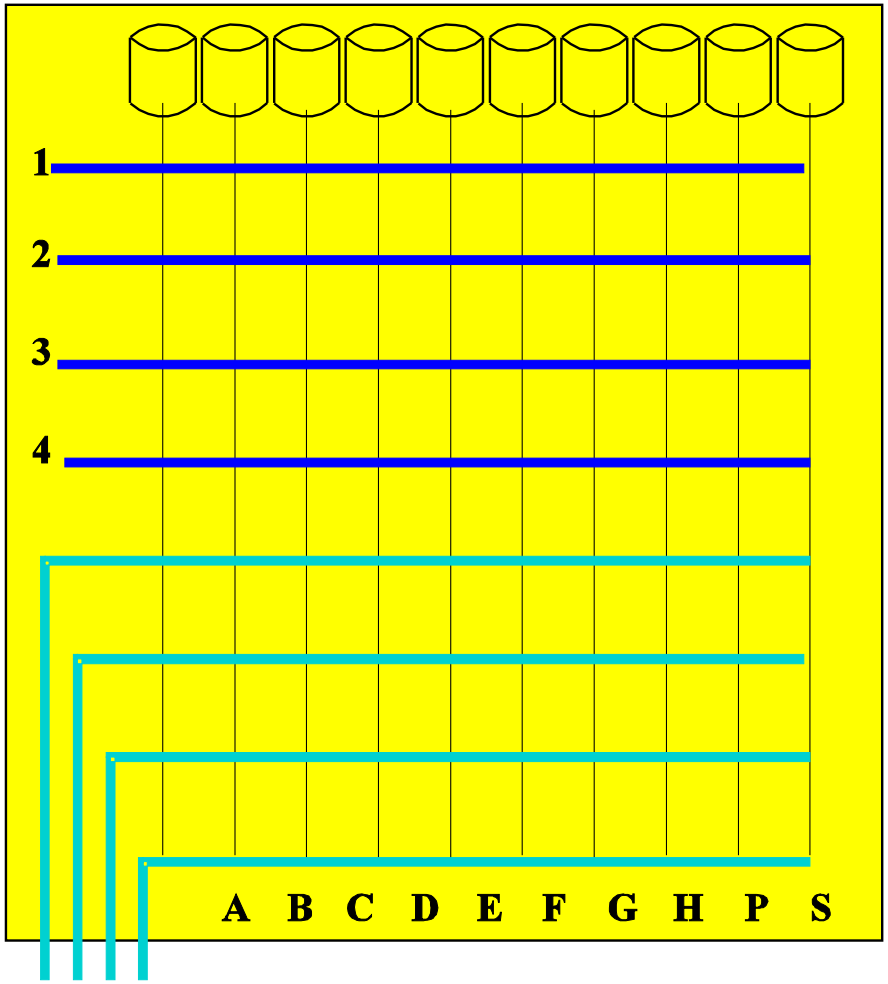


Sandia National Laboratories

Goal?

- 50000MB/sec from application to parallel file-system
  - Read or write
  - File-per-process or Shared-file
- Lustre is the parallel file-system
- File-system configuration
  - 160 OSSs (IO nodes)
  - 2 OSTs per OSS (320 OSTs)

Controller Internals



Couplet Connectivity

Sandia National Laboratories

Back End Testing
(Single path Theoretical)

- What is the limiting factor using a single port of controller?
- Internel disk channels (A-H,P,S) 1Gb/sec (100MB/sec)
- Disk 43MB/sec (min) – 78MB/sec (max)
- Controller port 2Gb/sec (200MB/sec)
- HBA 2Gb/sec (200MB/sec)

8 DDN data channels (A-H) → 800MB/sec

43MB/sec × 1 disk/channel × 8 channels = 344MB/sec (min)

78MB/sec × 1 disk/channel × 8 channels = 624MB/sec (max)

- Inside the controller disk limits the rate - BUT
- Still limited by 200MB/sec controller port/HBA port

Sandia National Laboratories

Back End Testing (Aggregate Theoretical)

- What is the limiting factor when using all four ports of controller?

8 DDN data channels (A-H) → 800MB/sec
800MB/sec ÷ 4 ports/controller = 200MB/sec/port/controller

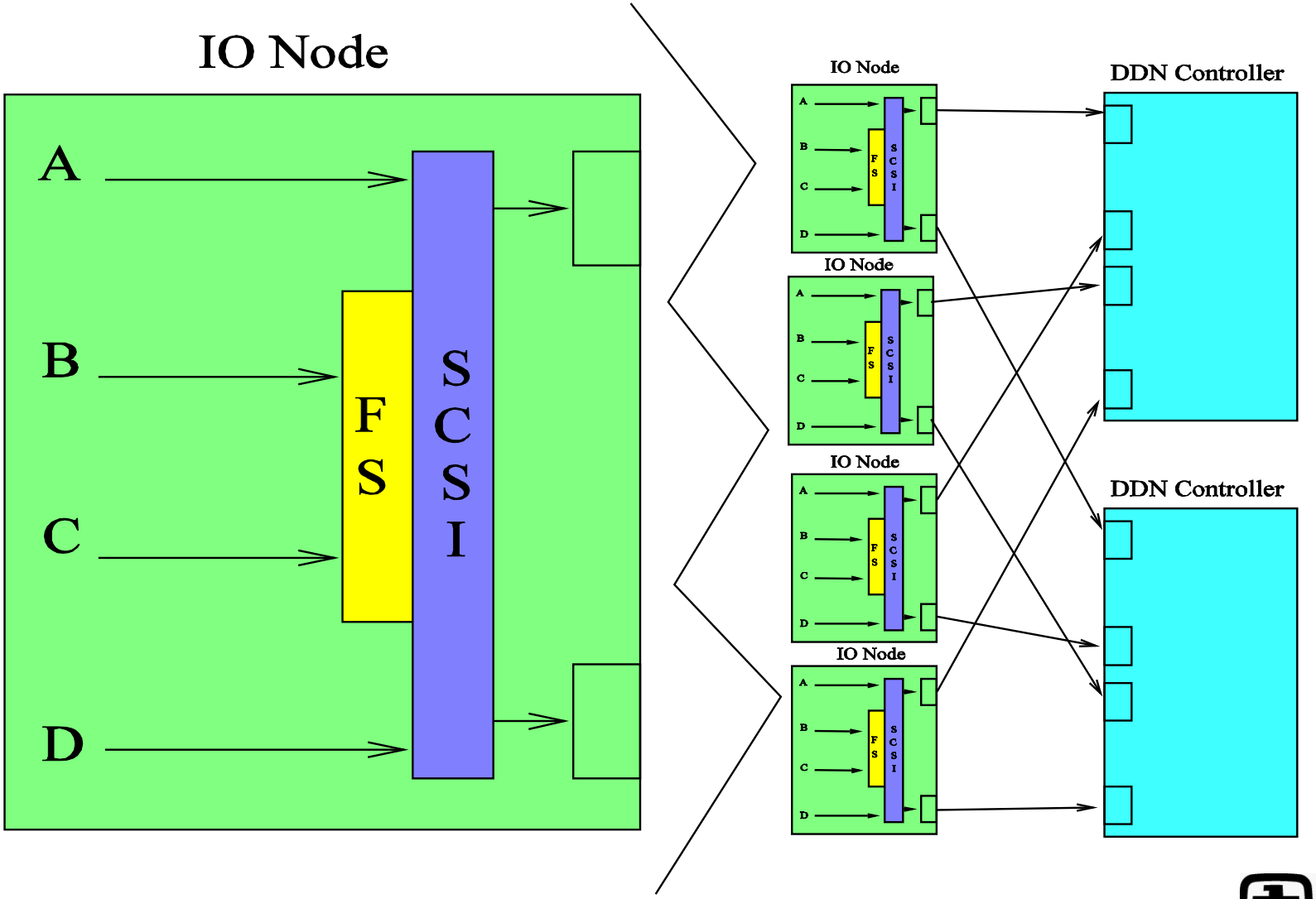
- Each port on controller gets ¼ of each data channel

100MB/sec/channel ÷ 4 ports = 25MB/sec/channel/port

- Minimum per disk rate is 43MB/sec
  - Exceeds shared per data channel rate
- Still limited by controller port/HBA/data channel rate (200MB/sec)

Sandia National Laboratories

IO Node/Controller Configuration (AGAIN)



Sandia National Laboratories

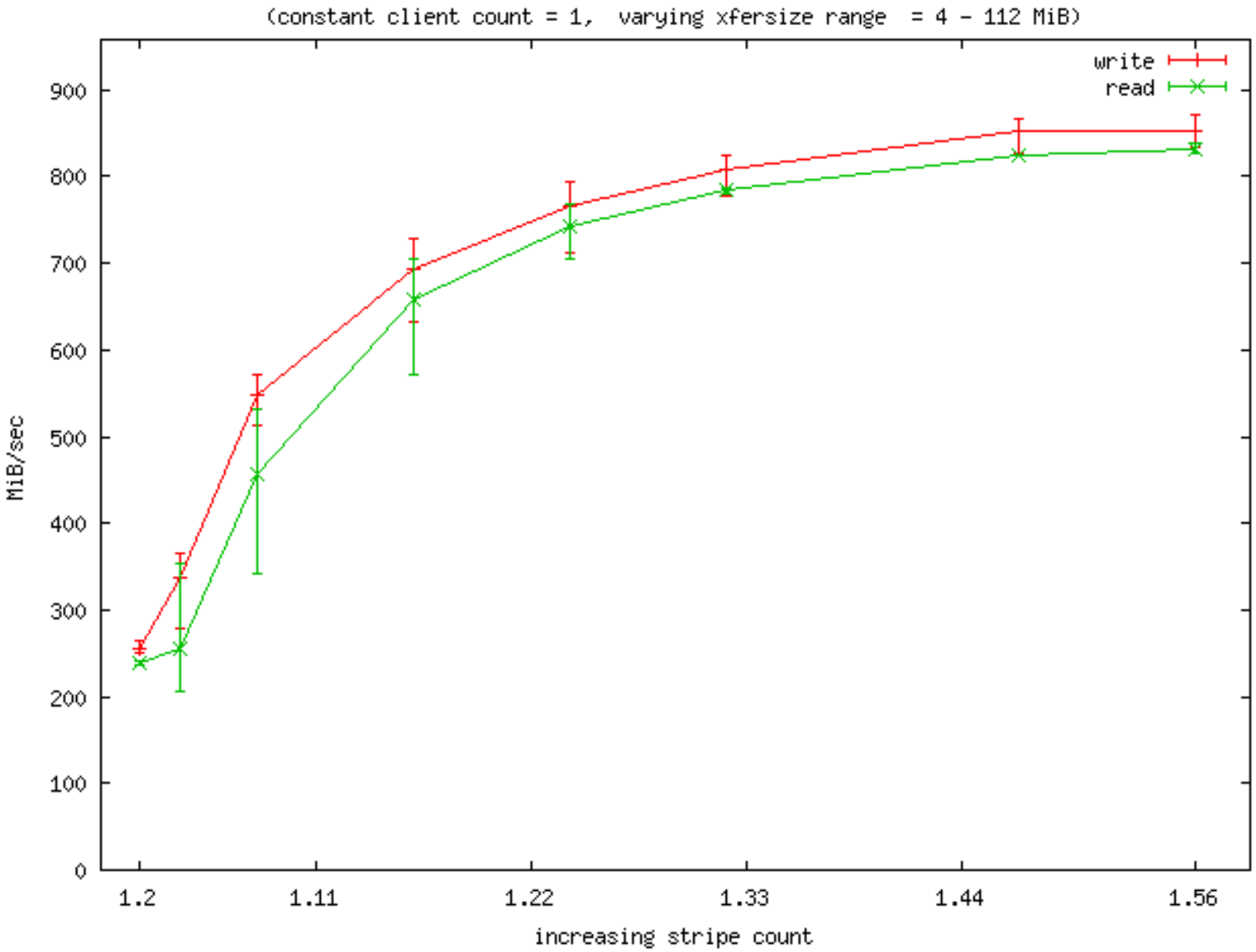
Back End Testing (Demonstrable)

- Single port on IO node, Single port Controller
  - SCSI layer 196.23MB/sec (A)
  - File-system layer 179.84MB/sec (B)
- Both ports on IO node, one port each on separate controllers
  - SCSI layer same (196MB/sec) (A and D)
  - File-system layer 102.35MB/sec (B and C)
- One port each on four IO nodes, all four ports on a single controller
  - SCSI layer 195.13MB/sec (Ax4)
  - File-system layer 140.82MB/sec (Bx4)

Parallel File-System Tests

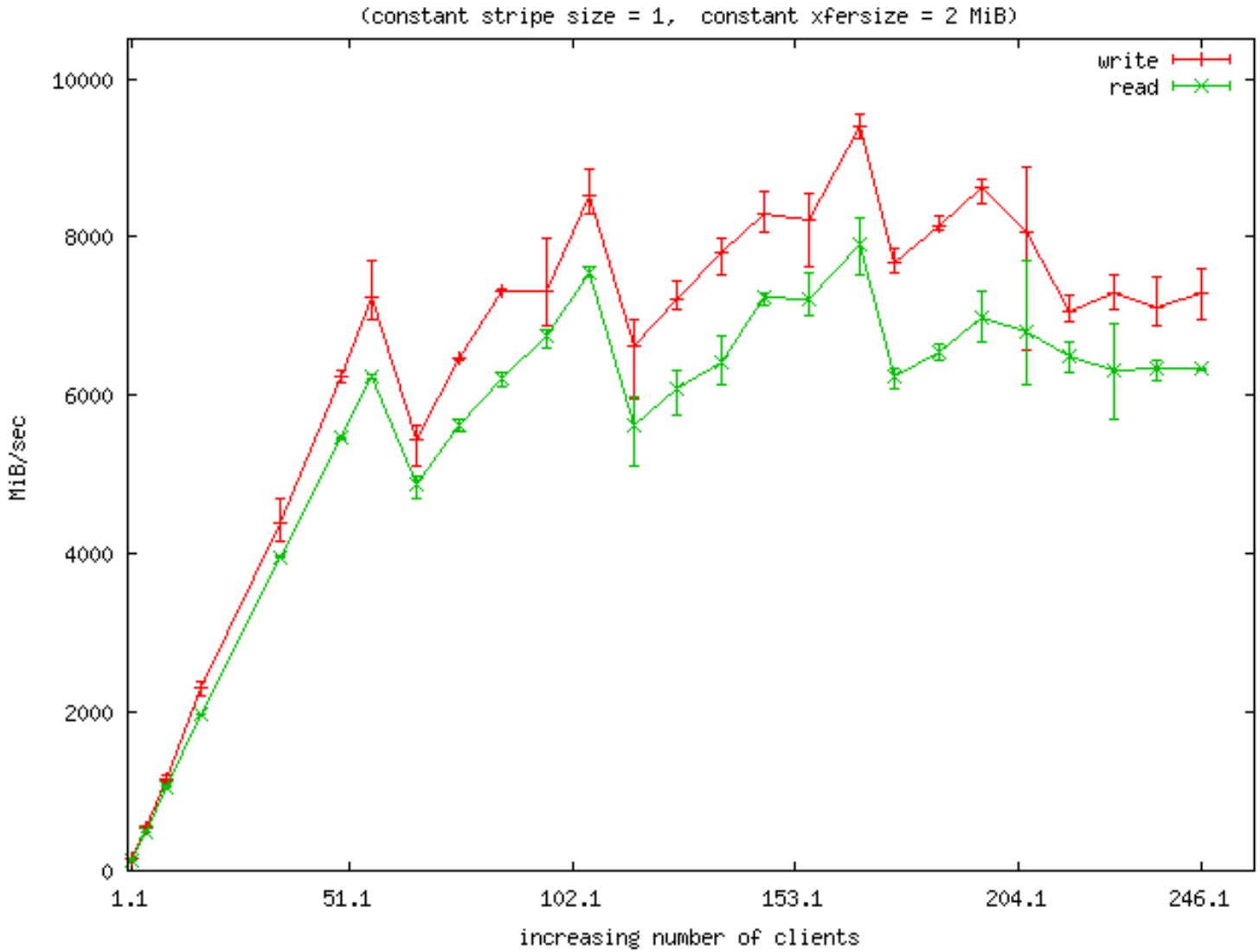
- IOR – parallel application
  - File-per-process
  - Shared-file
- Lustre
  - OSS/OST assignment carefully controlled
- Testing Oversubscription
  - 60:1, ratio of compute to IO nodes in initial configuration
  - Achieve this by limiting OSS/OSTs used

Front End Test



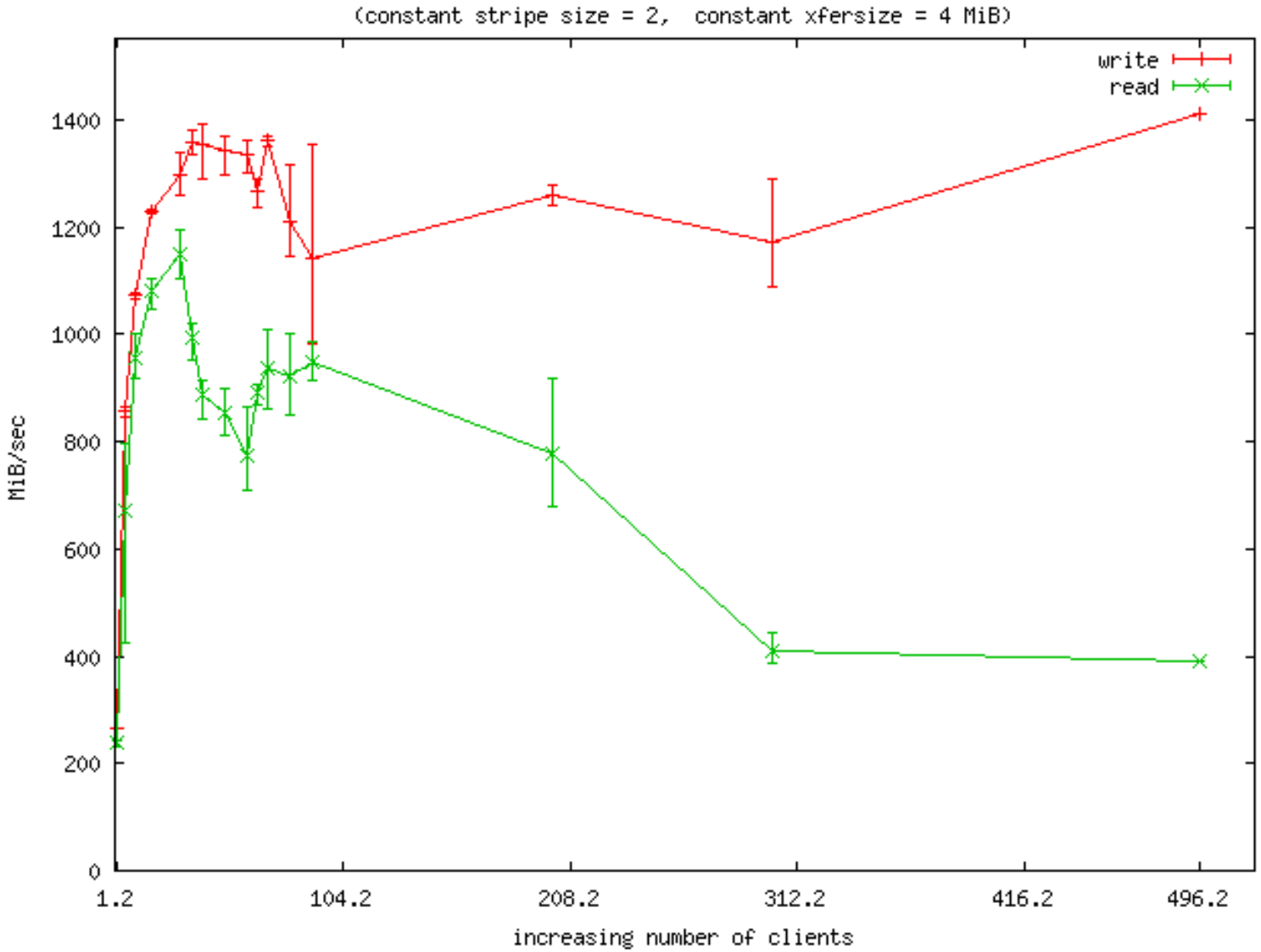
Sandia National Laboratories

File-Per-Process



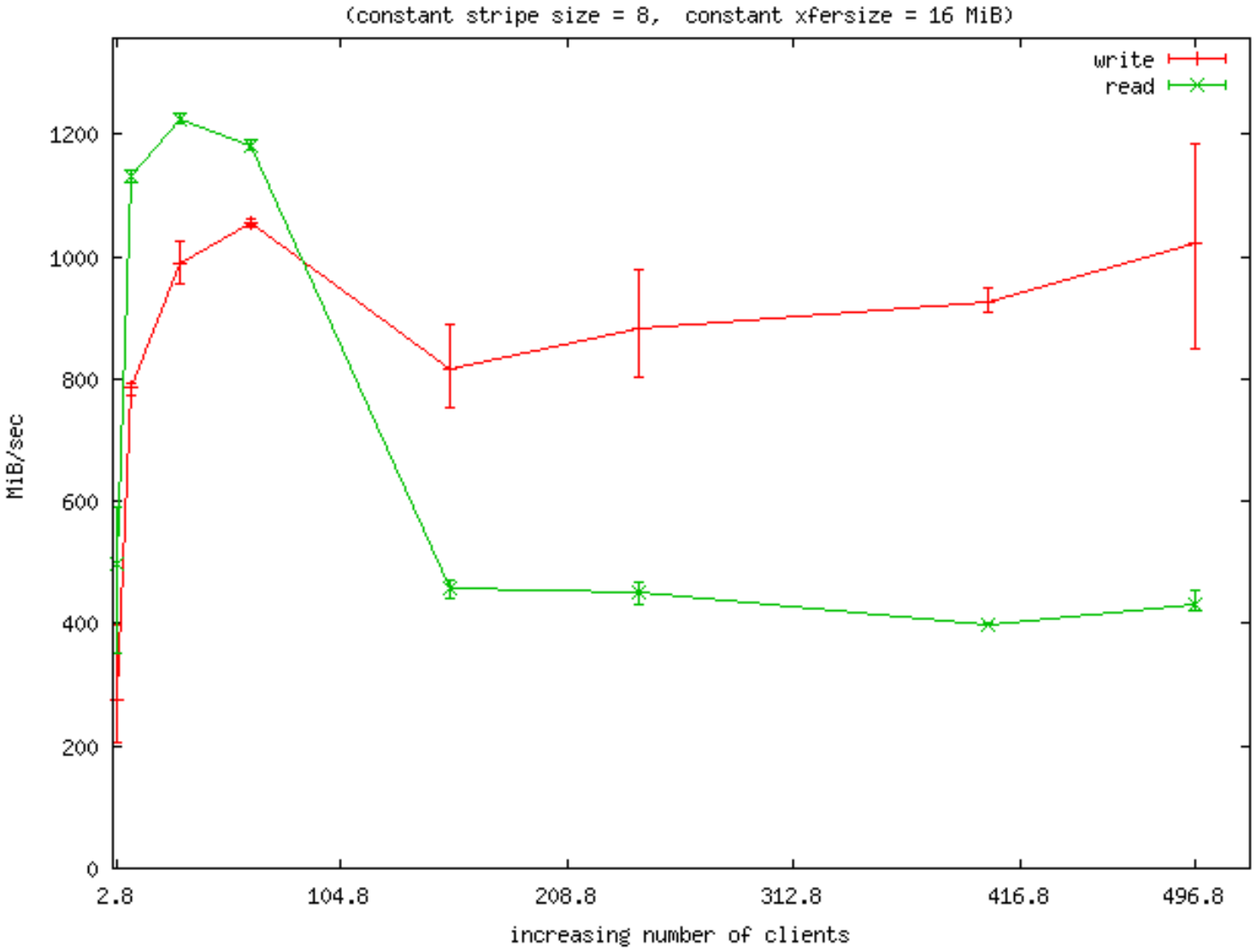
Sandia National Laboratories

File-Per-Process (Oversubscribed)



Sandia National Laboratories

Shared-File (Oversubscribed)



Sandia National Laboratories

Conclusions

- Physical configuration sufficient to achieve goals

195MB/sec per port using 320 ports yields 62400MB/sec

- Initial non-Lustre file-system testing not good
- Lustre results more promising
  - 154.15MB/sec (avg) × 320 OSTs = 49280MB/sec (from file-per-process oversubscribed)
  - 176.31MB/sec (max) × 320 OSTs = 56419.2MB/sec (from file-per-process oversubscribed)
  - Unfortunately only for writing
  - Only for file-per-process
- Read performance suffered in all cases
- Write performance for shared-file also insufficient
- Results shared with Cray and CFS

Future

- Testing will continue after software or hardware upgrades
- Larger scale testing
  - Demonstrated 54104MB/sec (writing)
    - 86% of theoretical!
    - 320 OSTs
    - 640 clients
    - File-per-process
    - Not reliably repeatable ☹
  - Demonstrated > 50000MB/sec (writing)
    - Client counts up to 3200
    - Indicates rates can be maintained at larger scale
    - Not reliably repeatable ☹
  - Little good to say about read rates

http://www.cs.sandia.gov/RSIOPA

Acknowledgements

- **Team participants**

Lee Ward

Ruth Klundt

Sue Kelly

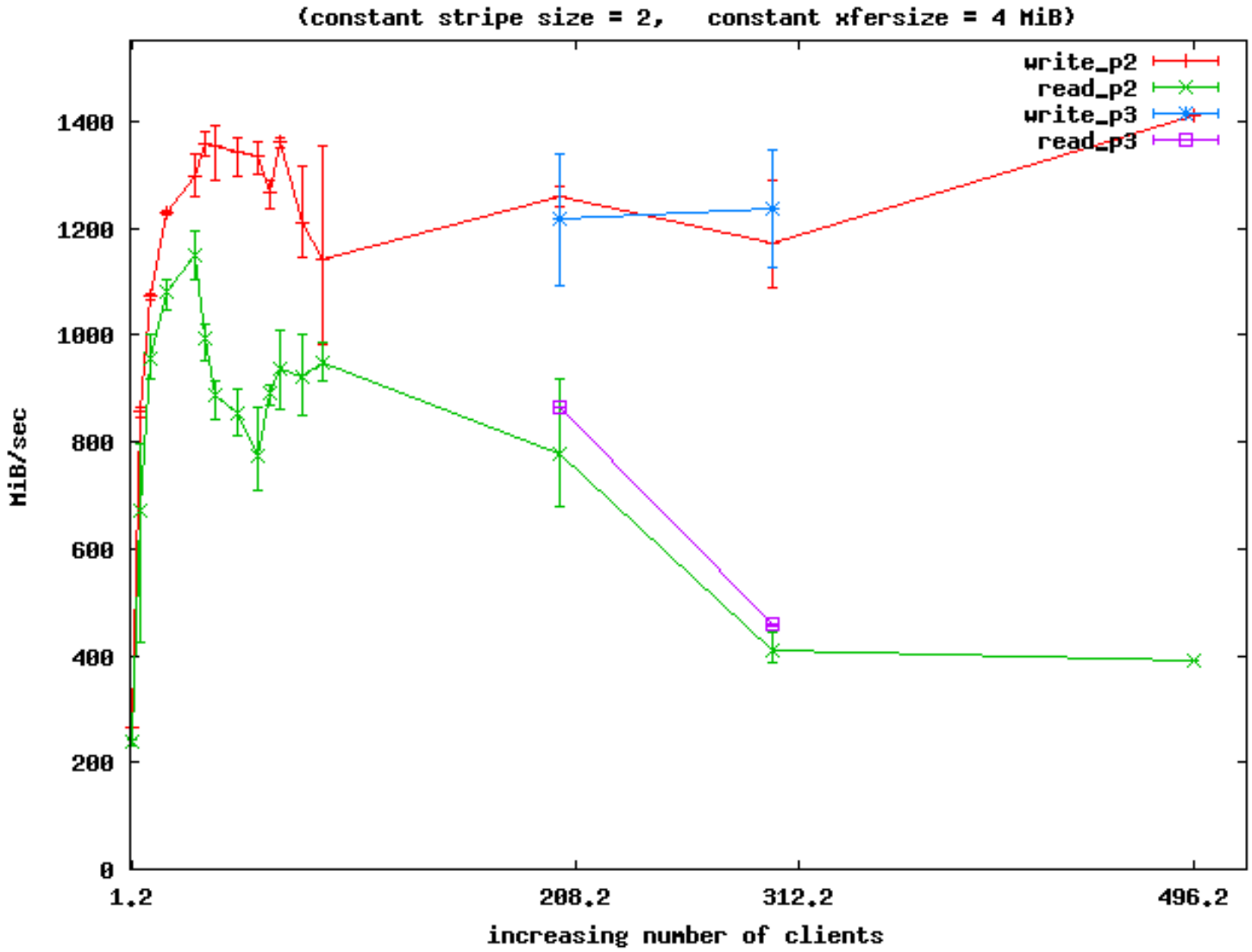
Jim Tomkins

Brian Kellogg

Thanks also to Bob Ballance and John Noe for accommodating the long system time periods required, and the Cray admin staff for being responsive 24 hours a day.

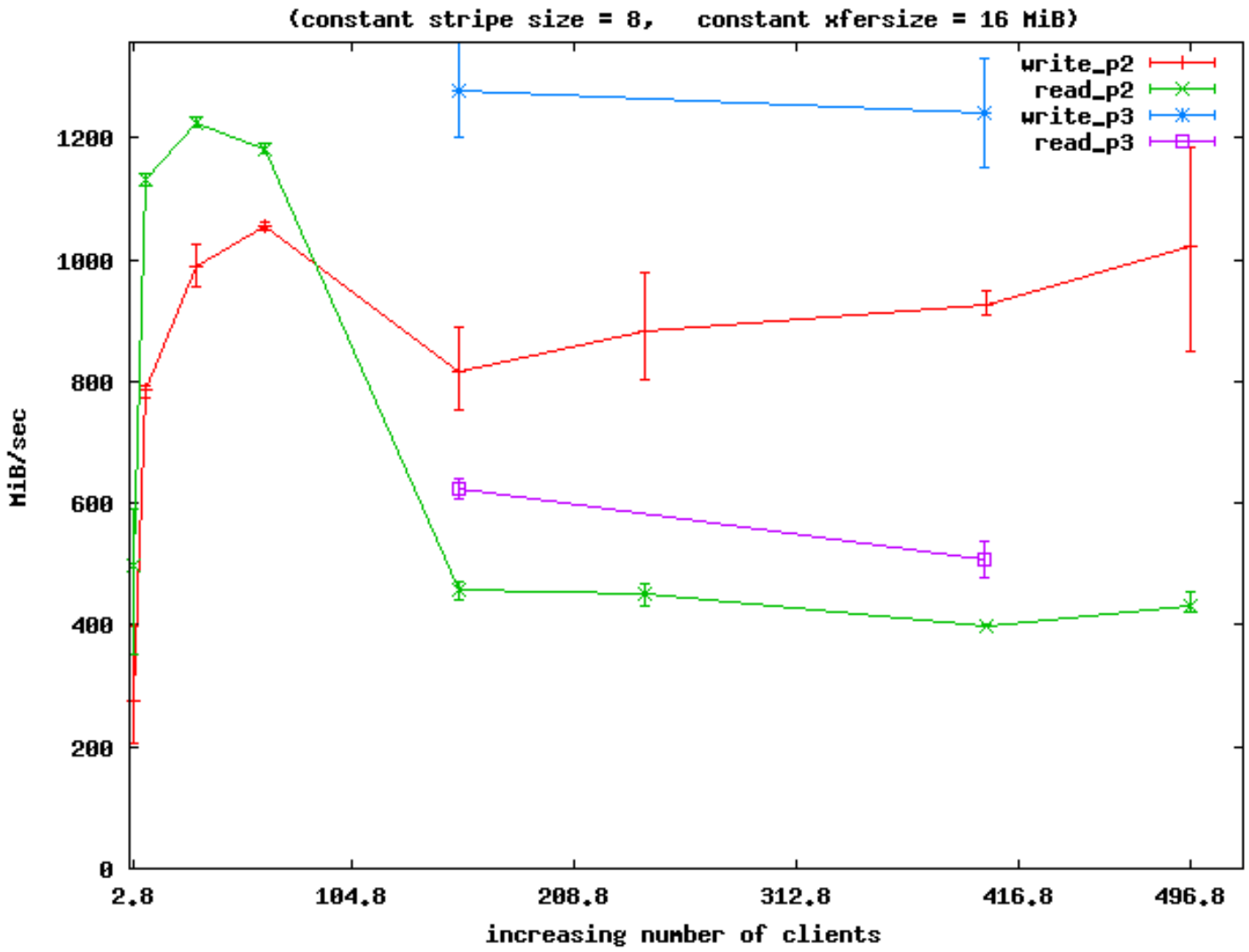
Sandia National Laboratories

Revisit File-Per-Process



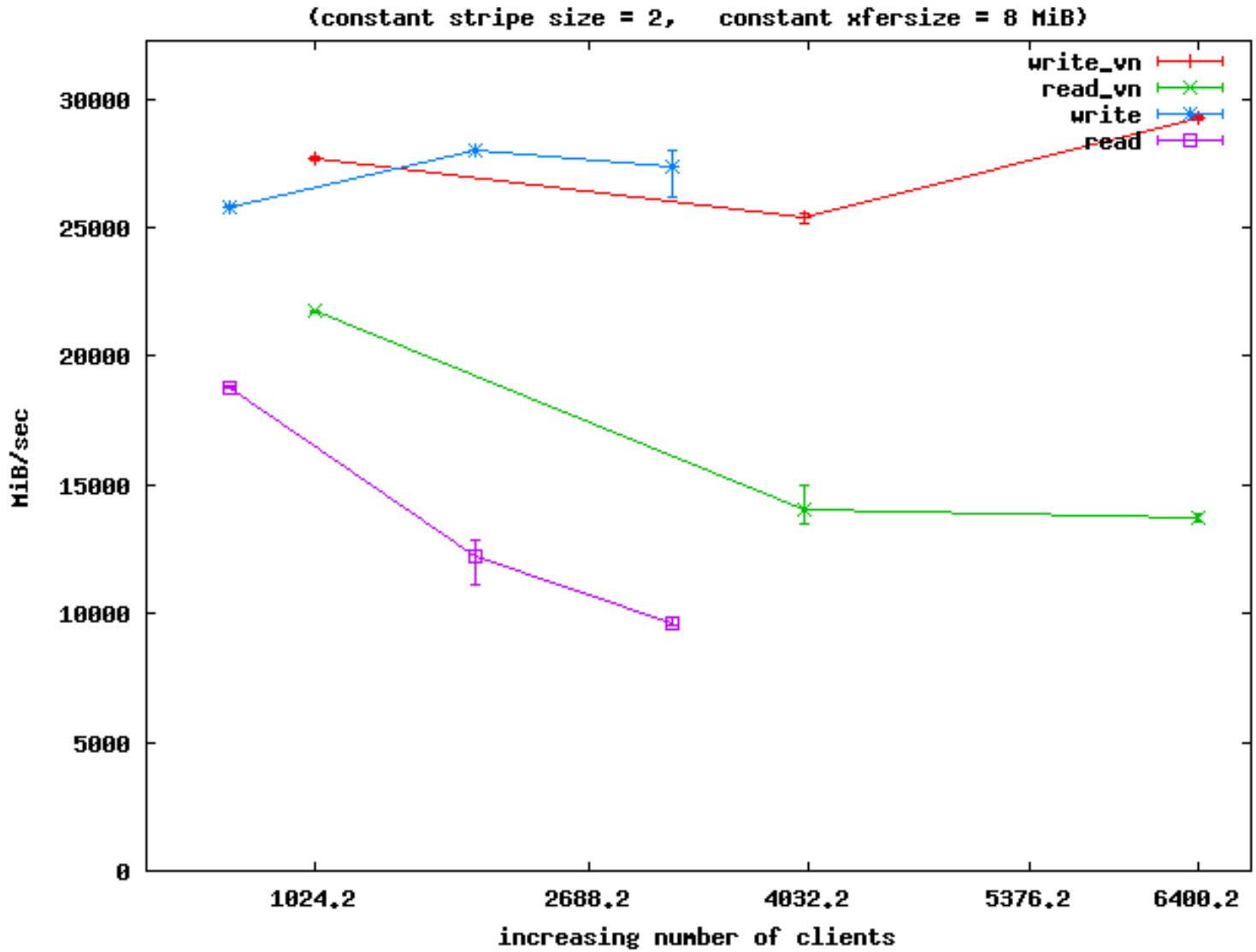
Sandia National Laboratories

Revisit Shared-File



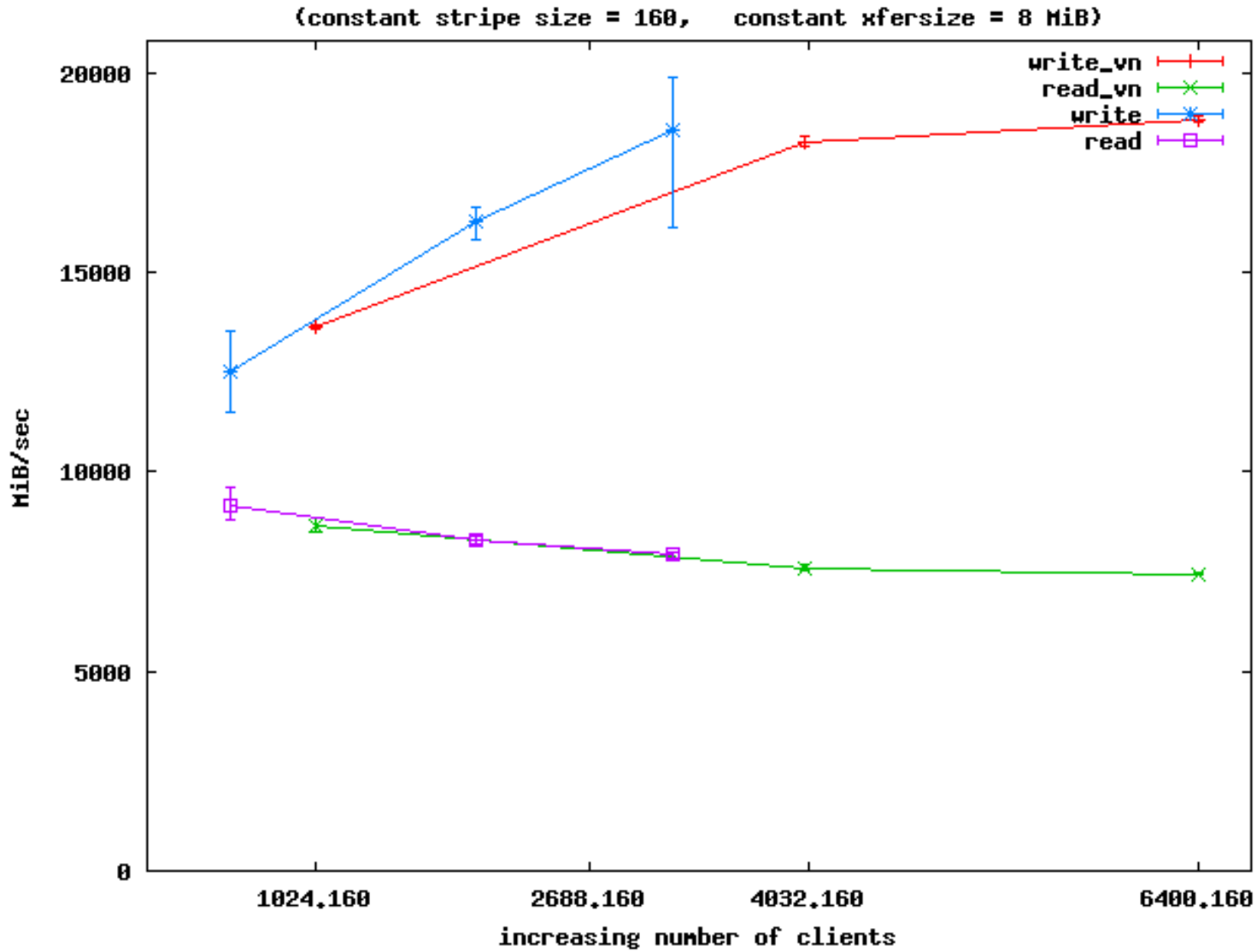
Sandia National Laboratories

File-Per-Process
(Larger client counts)



Sandia National Laboratories

Shared-File
(Larger client counts)



Sandia National Laboratories