HPCS I/O Scenarios Update

John Carrier
17 April 2013

Introduction

- The Defense Advanced Research Projects Agency (DARPA) provided its High Productivity Computing Systems (HPCS) vendors with a set of 14 representative I/O workloads, which they called the “HPCS File I/O Scenarios”
- The HPCS Scenarios focus on the scalability of the proposed HPCS file system and storage hardware, not on absolute performance
- At LUG 2012, Cray announced the availability of its implementation of the HPCS I/O Scenarios and presented results from its initial evaluation of the Scenarios on ORNL's Cray XT4
- This presentation describes updates to our tests and the results from our successful demonstration to DARPA of Cray’s next generation XC30 supercomputer

HPCS I/O Scenarios

- **Streaming I/O**
  1. Single stream with large data blocks operating in half duplex mode
  2. Single stream with large data blocks operating in full duplex mode
  3. Multiple streams with large data blocks operating in full duplex mode
- **Parallel I/O**
  5. Checkpoint/restart with large I/O requests
  6. Checkpoint/restart with small I/O requests
  7. Checkpoint/restart large file count per directory - large I/Os
  8. Checkpoint/restart large file count per directory - small I/Os
  13. Small block random I/O to multiple files
  14. Small block random I/O to a single file
- **Metadata I/O**
  4. Extreme file creation rates
  9. Walking through directory trees
  10. Parallel walking through directory trees
  11. Random stat() system call to files in the file system – one (1) process
  12. Random stat() system call to files in the file system - multiple processes

Updates to Cray's Scenarios Tests

- Time-based completion
  Added a command line option to exit tests after a specified time since test completions based on transfer sizes as defined in the Scenarios are indeterminate
- Improved command line syntax
  Updated the syntax and format of the command line interface to simplify deployment of the tests within Cray's automated test harness
- Decoupled reads and writes
  Added a command line switch to allow the writes and reads to be completed independently to remove risk of accessing any cached data during reads following writes

All updates have been uploaded to Cray's HPCS I/O SourceForge repository (http://hpcs-io.cray.com/)

Introducing the Cray XC30 Supercomputer

Adaptive Supercomputing

- Flexible Processor Options & Upgrades
- Hybrid Systems
- Adaptive Network Routing
- Advanced Adaptive Programming Tools

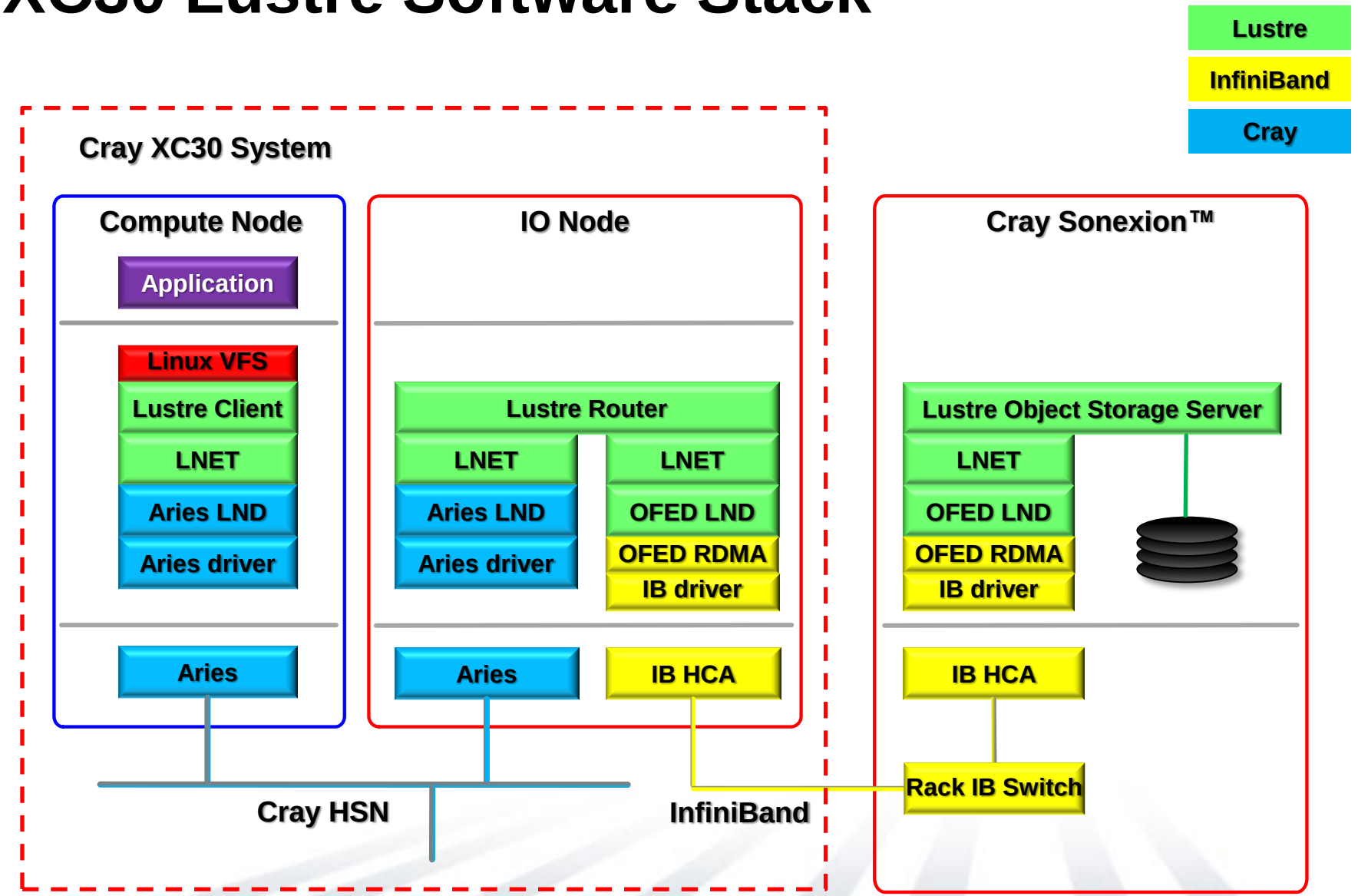
Scalable Performance

- Enhanced Aries Interconnect
- Global Network Bandwidth
- HPC Development Tools
- Cray Linux

Comprehensive HPC Integration

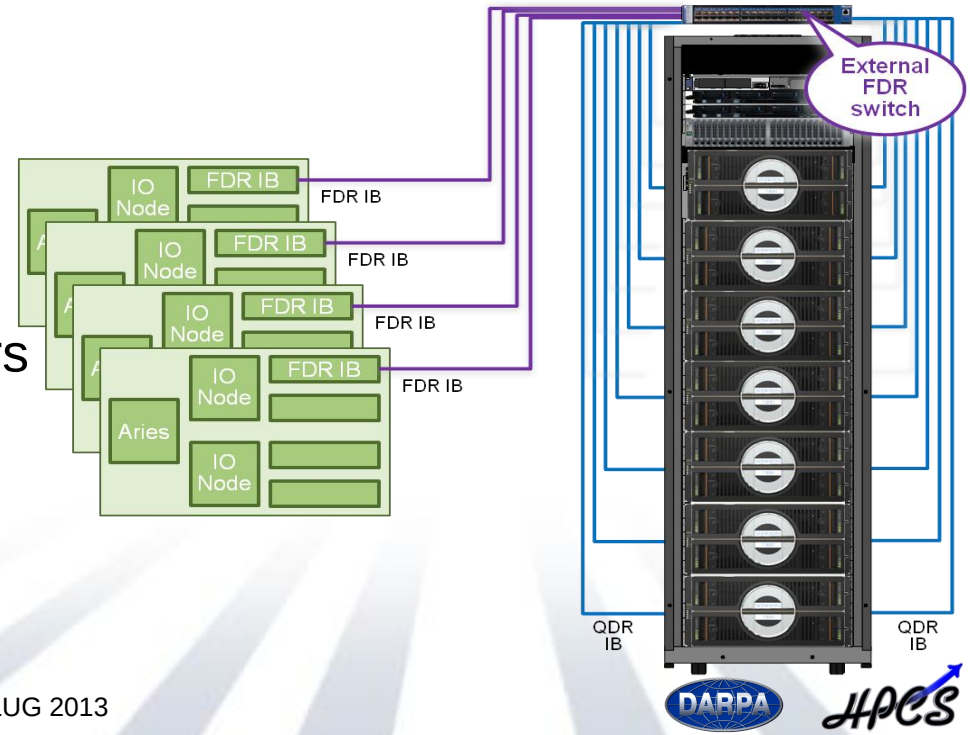
- HW & Networking
- SW Environment & Partner Ecosystem
- Storage
- Reliability & Resiliency

XC30 Lustre Software Stack



DARPA Demo System Configuration

- 4 total cabinets (compute + I/O)
  - 188 compute blades
    = 752 compute nodes
    =1504 Intel SandyBridge Sockets
  - 2 cabinet groups in Dragonfly network topology
  - Peak performance of 249 TF
- 4 I/O blades
  - 8 IO nodes total with 2 PCIe3 x8 slots per node
  - 4 IO nodes configured as Lustre LNET routers with one FDR IB HCA per node to give >20 GB/s aggregate I/O from computer
- 7 Cray Sonexion 1300 Controllers
  - 1300s have QDR IB HCAs and only used with XC30 for the DARPA Demo
  - XC30 ships with Sonexion 1600 and FDR IB

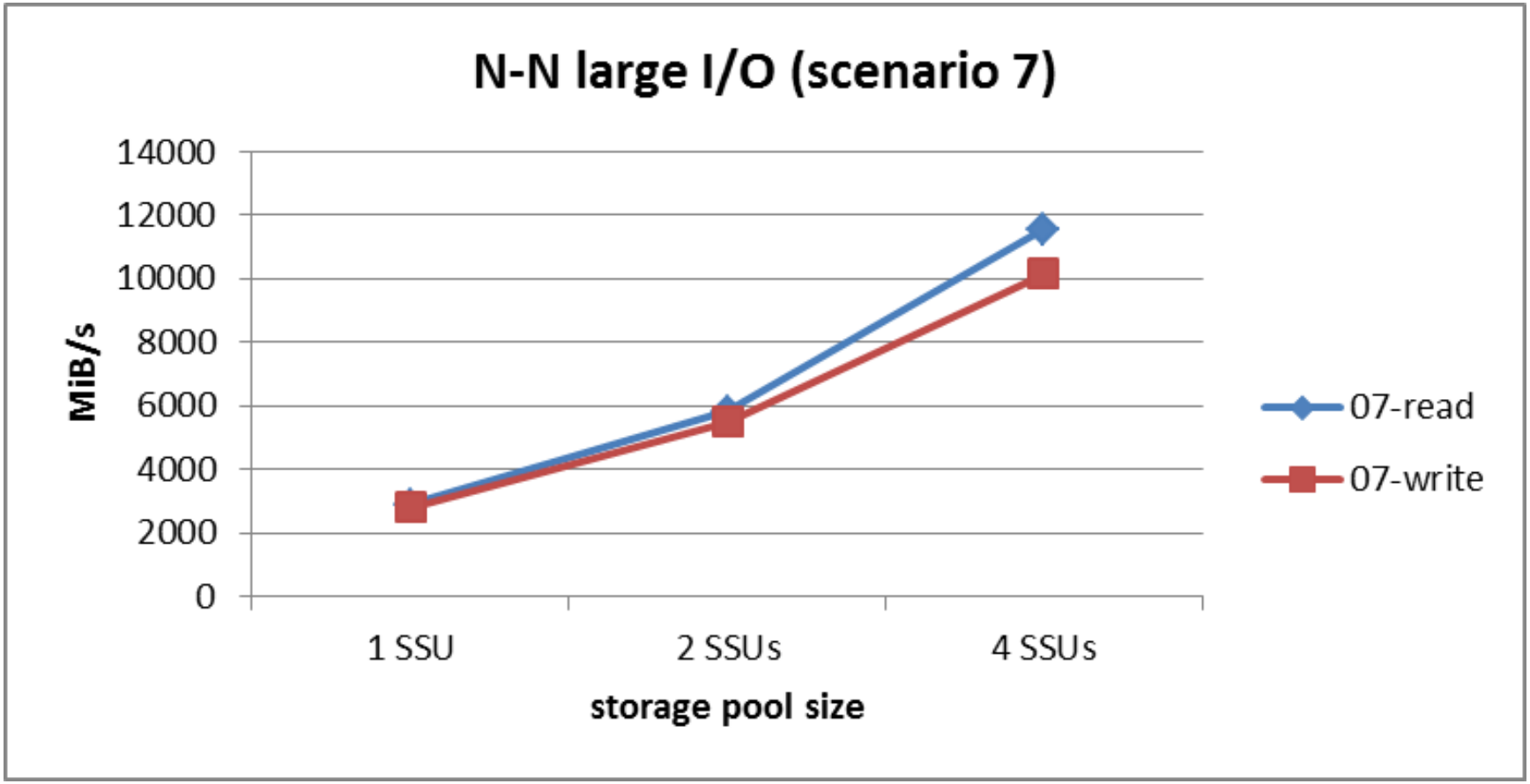


Storage Pool Definitions

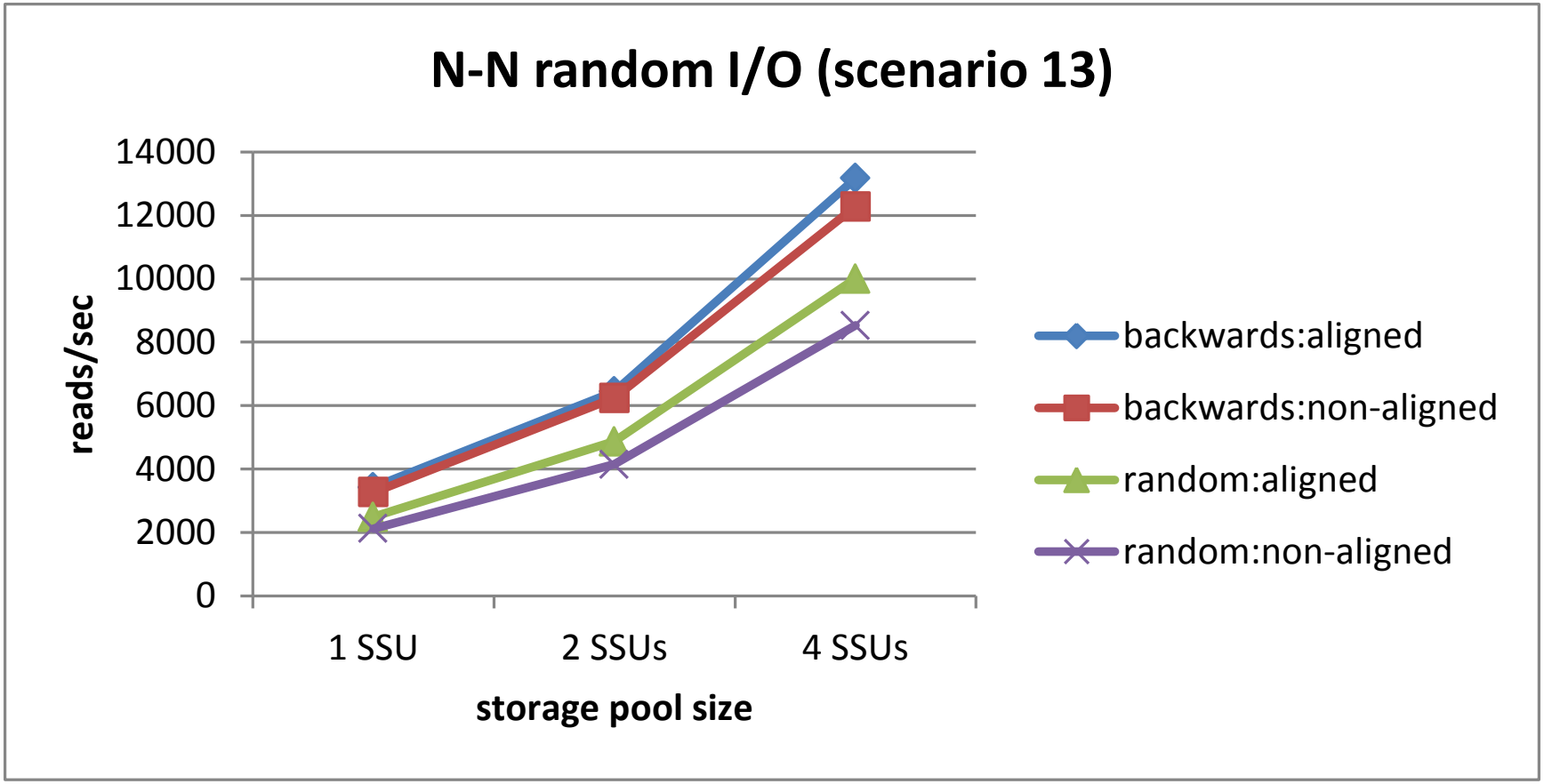
- The HPCS demonstration system had a single file system composed of seven (7) Sonexion 1300 SSUs
  - Each SSU had 2 OSS nodes; each OSS had 4 OSTs
  - Theoretical performance per SSU was ~3 GB/s
- Created three different sized storage configurations using Lustre pools that span 1, 2, or 4 SSUs from the single file system
  - Avoided reconfiguring the file system between scenario runs
- Used 4 client nodes per OST while running the parallel I/O tests to provide consistent workloads between storage pools

Pool	# of SSUs	# of OSSs	# of OSTs	# of clients	Expected Raw BW
FS1	1	2	8	32	3 GB/s
FS2	2	4	16	64	6 GB/s
FS3	4	8	32	128	12 GB/s

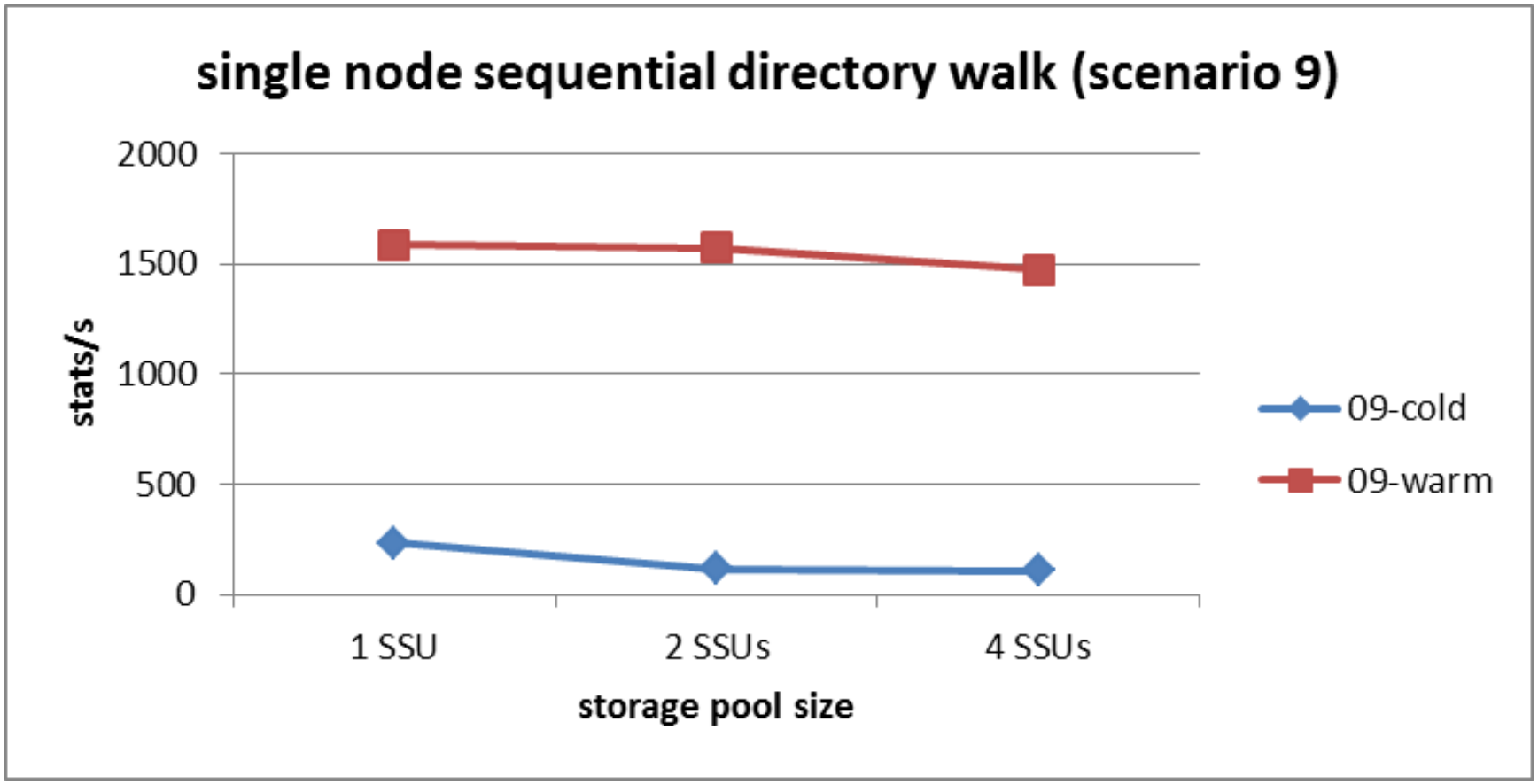
Large I/O with Scenario 7 (N-N)



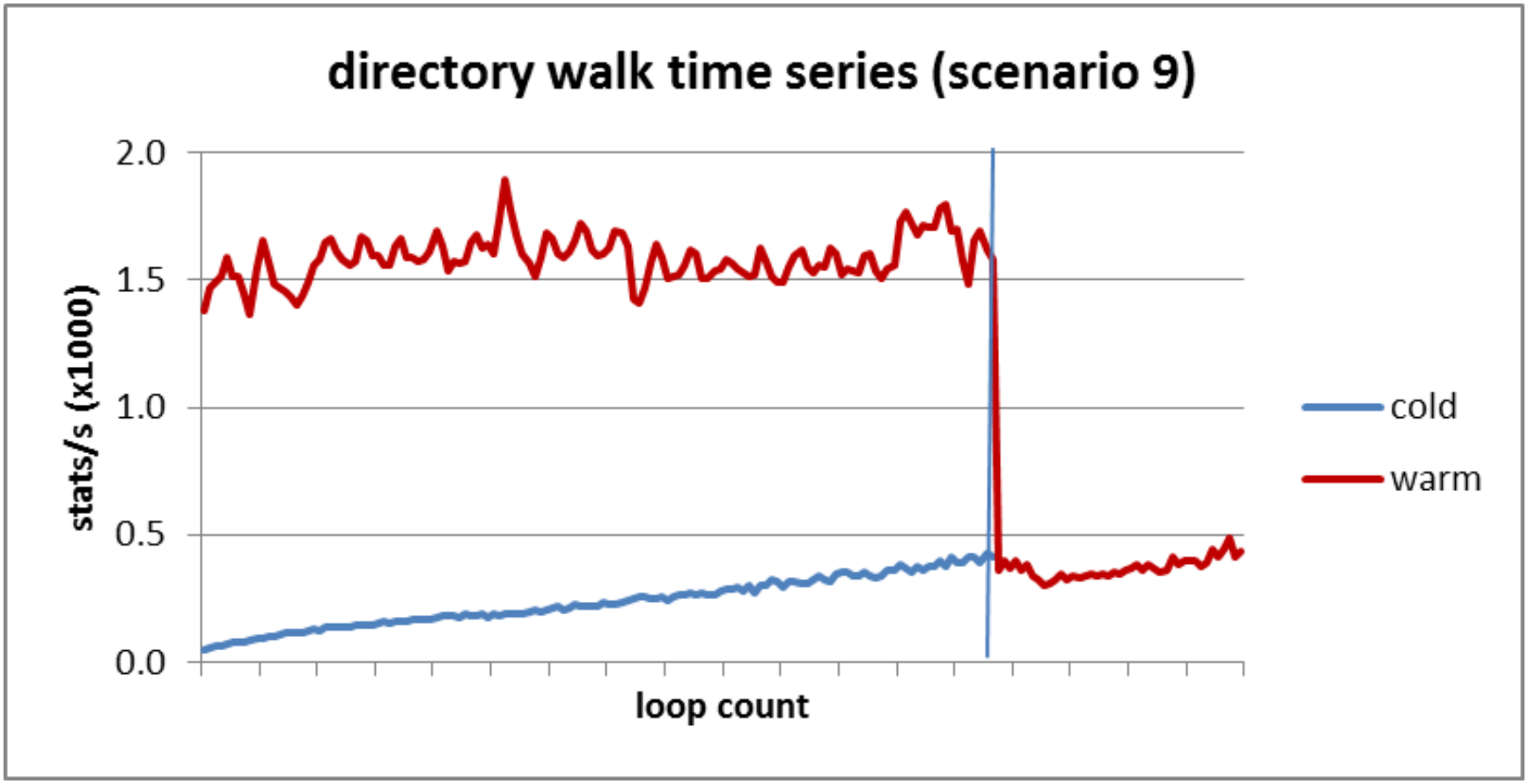
Random I/O with Scenario 13 (N-N)



Metadata performance with Scenario 9



Unexpected metadata results



- Initial warm rates were lower than expected because they overran the number of files scanned during the cold runs
- Inspected the intermediate results to find the warm cache rates

Conclusions

- Cray has demonstrated an update to its implementation of the HPCS I/O Scenarios
- Lustre shown to be a very scalable file system (again)
- HPCS Scenarios are about the scalability of the file system and storage hardware, not the absolute performance
  - DARPA defined workloads important to their HPCS mission partners
  - Other workload definitions are possible

- References
  - http://hpcs-io.cray.com/
  - http://sourceforge.net/projects/hpcs-io/files/DARPA.HPCS.IO.Scenarios.2011.pdf
  - http://www.opensfs.org/wp-content/uploads/2011/11/Carrier_LUG12_HPCS-Scenarios.pdf

Thank You

- This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.
- Our tests were performed at Cray's Chippewa Falls computer facility. The author gratefully acknowledges the assistance of Glen Overby, Brad Stevens, Steve East, Jeff Garlough, and Linda Finnegan of Cray's testing group; Mark Swan and Dick Sandness of Cray's benchmarking group.