Parallel File System Testing for the Lunatic Fringe: the care and feeding of restless I/O Power Users

Richard Hedges, Bill Loewe, Chris Morrone, Tyce McLarty

Scalable I/O Project: Lawrence Livermore National Laboratory

UCRL-PRES-211216

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

ASC Requirements Today

Advanced **S**imulation and **C**omputing Program

- 100 teraflops
- 50 terabytes memory
- 2 petabytes RAID Disk
- 100 GB/sec I/O to a single parallel app

System capabilities have increased by factor of over 100 since the inception of ASC

ASC Programming Model

- 10,000's compute nodes
- Distributed memory (MPI coordination)
- C, C++, FORTRAN
- Parallel I/O from all nodes to shared files or to sets of files (file per process) on single mount point
- Visualization cluster may mount file system for post processing (no copy across network)

Simulations may run for weeks on thousands of processors and write a terabyte per day

Multiple Platform Technology Tracks

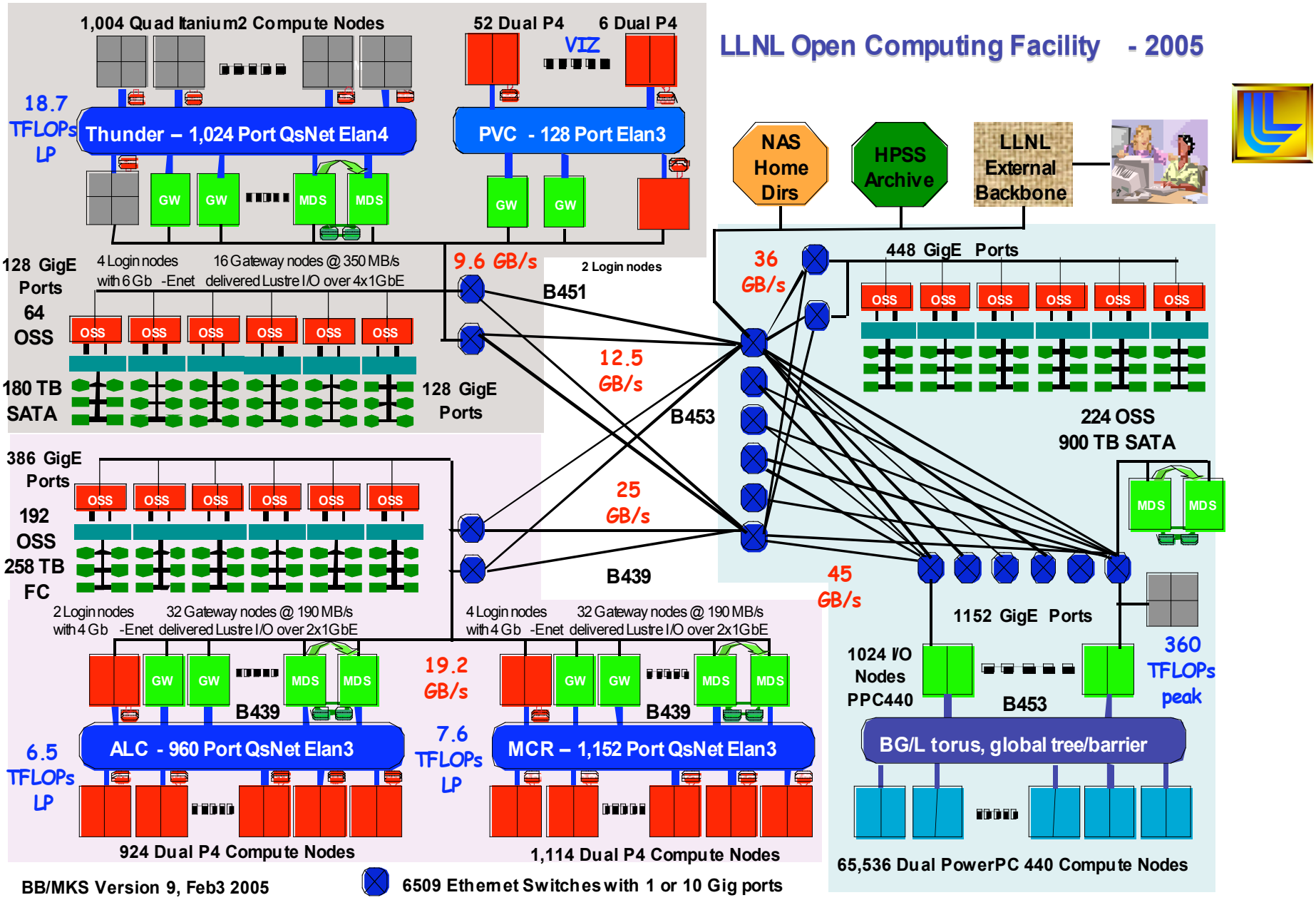
	ASC	**Commodity**	**Cellular (BlueGene)**
Integrator	IBM	LLNL	IBM
Processor	Power[2-5]	Intel IA32 IA64, AMD*	PowerPC 440
OS	AIX	Linux	Linux and “Linux like”
Switch	Federated	Quadrics, InfiniBand*	Custom torus/tree
Filesystem	GPFS	Lustre	Lustre for us

Lustre Filesystem

- Our chosen file system solution for commodity track (and beyond)
- Collaboration of 3 DOE Labs, HP, Intel, and Cluster File Systems, Inc.
- Active development started ~3 years ago

- Object Based parallel file system
- Open Source
- Designed for high performance and scalability
- Share data between multiple computers and clusters

Lustre is running in production while still under early and active development


LLNL Open Computing Facility - 2005
1,004 Quad Itanium2 Compute Nodes
52 Dual P4
6 Dual P4
VIZ
18.7 TFLOPs LP
Thunder – 1,024 Port QsNet Elan4
PVC - 128 Port Elan3
GW
MDS
NAS Home Dirs
HPSS Archive
LLNL External Backbone
128 GigE Ports
64 OSS
180 TB SATA
4 Login nodes with 6 Gb -Enet
16 Gateway nodes @ 350 MB/s delivered Lustre I/O over 4x1GbE
9.6 GB/s
2 Login nodes
B451
128 GigE Ports
36 GB/s
448 GigE Ports
OSS
12.5 GB/s
B453
224 OSS
900 TB SATA
386 GigE Ports
192 OSS
258 TB FC
25 GB/s
B439
45 GB/s
1152 GigE Ports
MDS
2 Login nodes with 4 Gb -Enet
32 Gateway nodes @ 190 MB/s delivered Lustre I/O over 2x1GbE
4 Login nodes with 4 Gb -Enet
32 Gateway nodes @ 190 MB/s delivered Lustre I/O over 2x1GbE
19.2 GB/s
1024 I/O Nodes PPC440
360 TFLOPs peak
6.5 TFLOPs LP
ALC - 960 Port QsNet Elan3
7.6 TFLOPs LP
MCR – 1,152 Port QsNet Elan3
BG/L torus, global tree/barrier
924 Dual P4 Compute Nodes
1,114 Dual P4 Compute Nodes
65,536 Dual PowerPC 440 Compute Nodes
BB/MKS Version 9, Feb3 2005
6509 Ethernet Switches with 1 or 10 Gig ports

Lustre Testing Overview

- 450 nodes dedicated to continuous automated testing
- Throughput and metadata, performance and stress tests
- Customer discovered bugs receive top priority
- CFS developer (Andreas Dilger) supporting us full time
- Customers crucial to resolution of subtle problems

- Huge resource commitment
- Close collaboration with customers and developer
- Local development of tests

All necessary for our success with Lustre

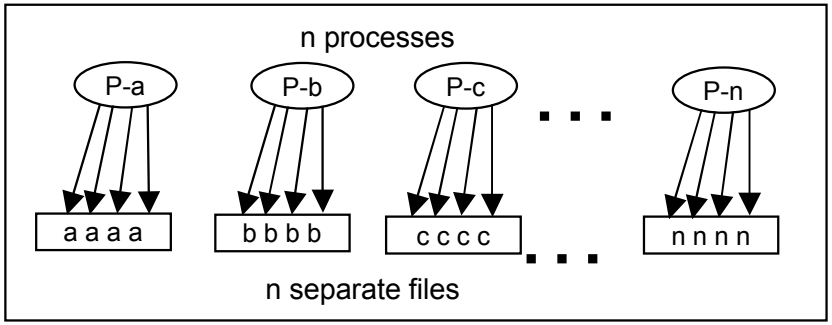
IOR (the workhorse)

Parallel file system bandwidth test and more

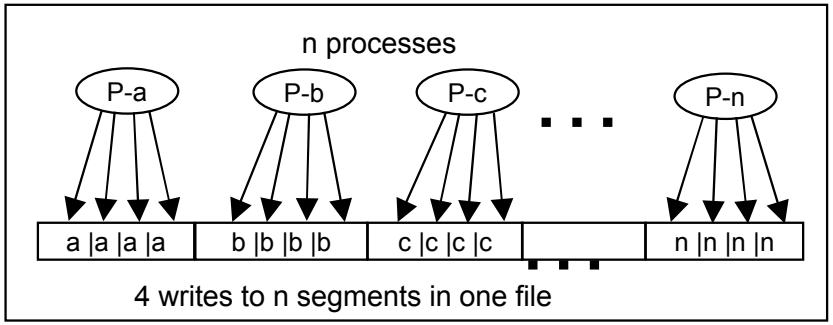
- MPI synchronized (one test on 32k processors)
- Shared or File per process operation
- Supports several interfaces:
  - POSIX, MPI-IO, HDF5, NetCDF, FORTRAN next?
- Size and number of transfers variable

Concisely models ASC applications usage patterns

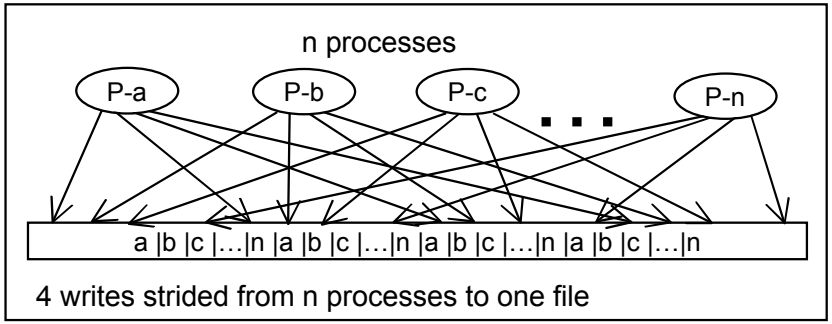
ASC File Access Modes



File Per Process Model



Shared File: Segmented Access Model



Shared File: Strided Access Model

Metadata tests

mdtest

- performance test
- File and directory creation and deletion rates
- Run parameters:
  - Shared directory or per task
  - # of files in each directory
  - Size of each file

simul

- stress test
- simultaneous operations
  - Shared or individual
  - Files or directories
- Operations: open, close, file stat, lseek, read, write, chdir, dir stat, readdir, mkdir, rmdir, unlink, rename, creat, truncate,symlink, readlink, link

These reproduce issues discovered in more complex situations and provide relevant performance measures.

Daily testing report

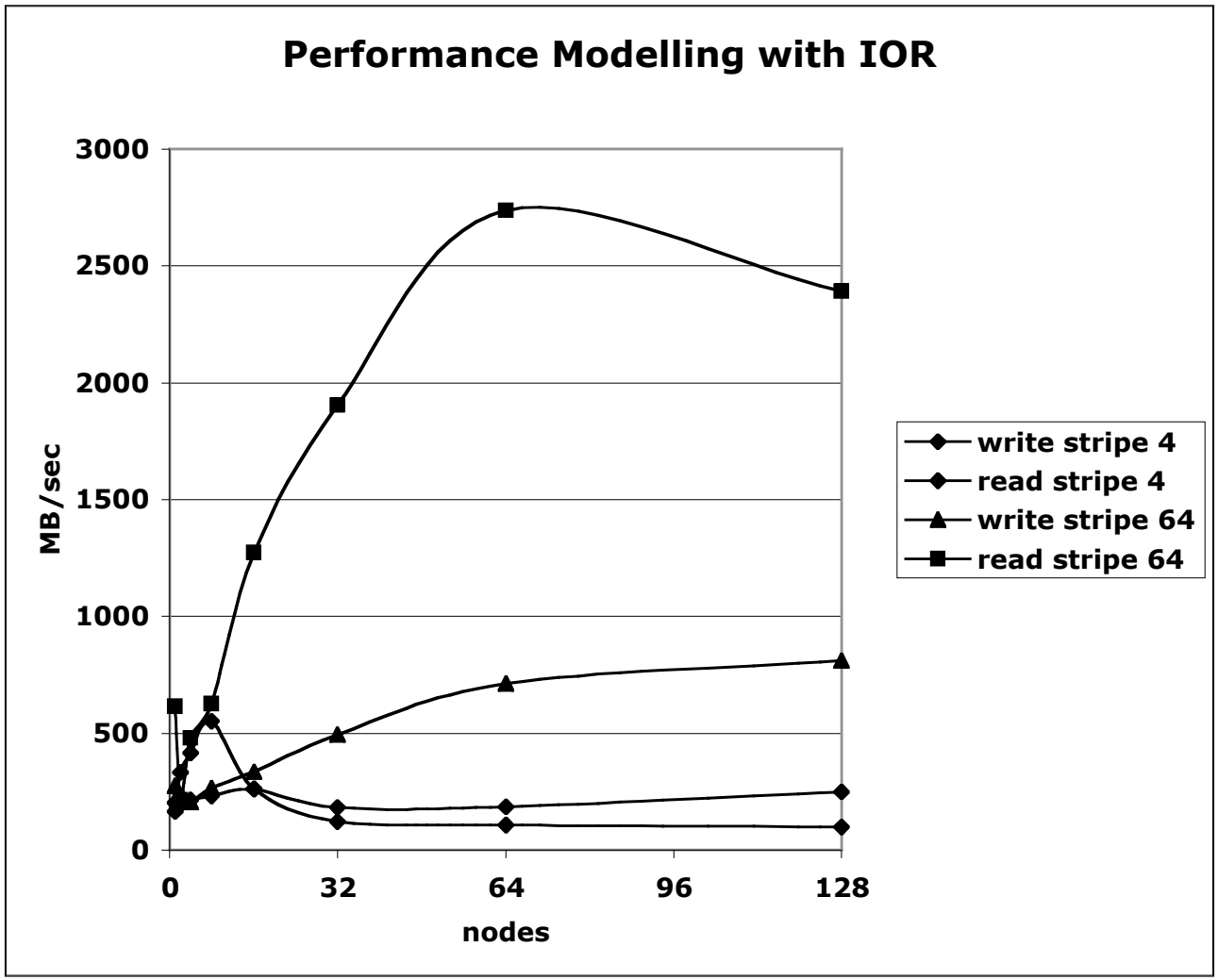
lustre-lite-1.2.8.6-2.4.21_p4smp_79chaos_200503091123
lustre-modules-1.2.8.6-2.4.21_p4smp_79chaos_200503091123
N.B. This is the bz#3389 code. We have not yet activated the groups upcall, though.

General results:

---

atomicity : No tests run
bonnie : No tests run
cabot : Passed 3 of 3 100%
dbench : Passed 3 of 3 100%
fsx : No tests run
ior : Passed 166 of 168 99%
test 22852: transfer failed
test 22879: transfer failed
iozone : No tests run
mdtest : Passed 2 of 2 100%
prodcon : No tests run
simul : Passed 2 of 2 100%

IOR: effect of Lustre Striping



The Cabot Bug

- User observes miswritten data (2k bytes of 56Gbytes!)
- Code is FORTRAN, I/O is file per process
- Sample test coded in F77, refined in F90 by user
- “strace”ing sample code associates error w insertion of control words
- Error rate now 1-2 per day on 200 nodes
- Lustre developer associates error w write of an incomplete page
- IOR parameters selected for shared file w 1000 byte interleaved transfers: now all writes are incomplete page
- Errors now observed in minutes on 5-10 nodes
- Problem resolved and fixed
- Fixes staged on test systems and into production ~3 weeks
- Tests added to regression suite

Productive collaboration of customer, developer and test team all crucial to resolution of this subtle error

Thriving on the Lunatic Fringe

- Our extreme requirements compel aggressive pursuit of technology (and not just the file system, trust me!).
- We actively develop tests, try to find the problems before our customers, but …..
- When we don't, we count on great collaboration from customer, developers, etc., in debugging.
- Extensive testing is essential for development and production to have a chance to coexist.

- Codes available for download:

http://www.llnl.gov/icc/lc/siop/downloads/download.html