

Lawrence Livermore National Laboratory

Thinking about I/O on BG/P Systems

January 21, 2010



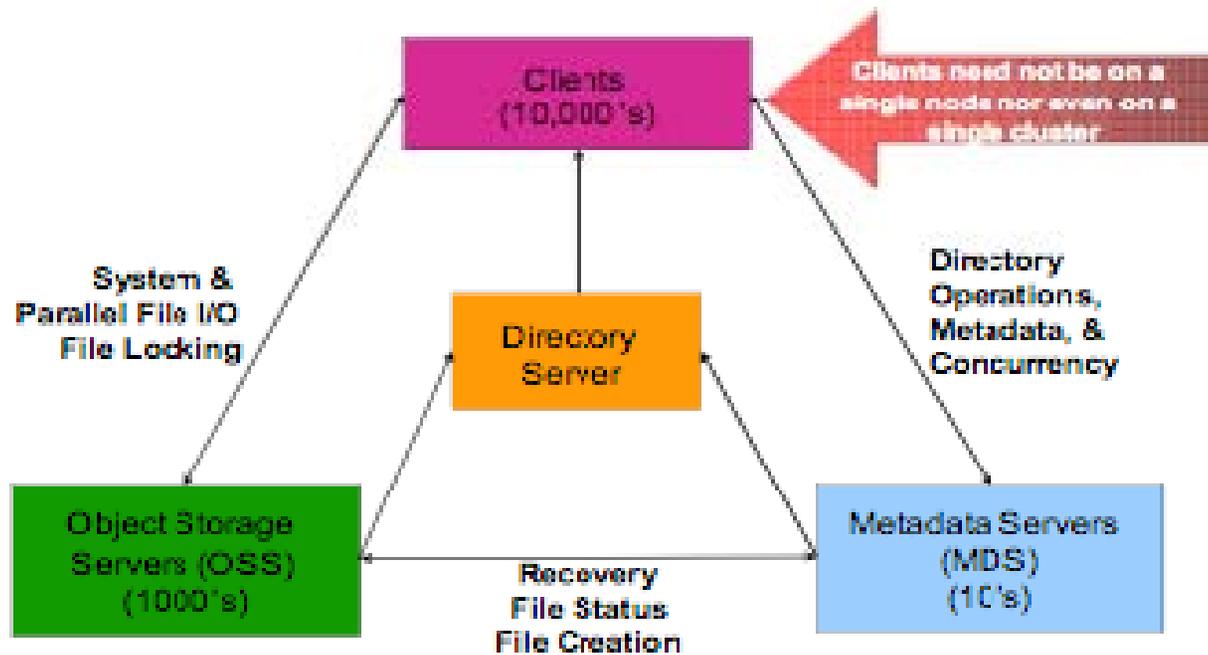
Richard Hedges

Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, CA 94551
This work performed under the auspices of the U.S. Department of Energy by
Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

LLNL-PRES-422764

Lustre Overview

Lustre Architecture



Lustre Overview continued

- Good data patterns
 - large contiguous transfers (broken into 1 Mbyte chunks)
- Bad data patterns
 - small transfers to random offsets
- Caching
 - Aggregate small contiguous transfer of monotonically increasing offsets into large transfers
- O_DIRECT
 - disable caching: all I/O call synchronous to storage



File I/O on BG/P

- I/O calls function shipped to ION
- 128 nodes share 1 ION (16 for Dawndev)
- ION has 4 Gbytes memory

- Caching of $4 \times [128 \text{ or even } 16]$ data streams in 4 Gbytes memory problematic



Local example

- From Jeff Grandy
- Example generates and write meshes to disk using the Silo interface (to HDF5)
 - Test representative of creating inputs to Overlink
 - 145 small meshes: 100 to 6000 zones
- run on Dawndev on 20 processors, 1 ION
 - 189 Mbytes, 950 files



BG/P strace observations

- Facility to trace function shipped system calls now supported
- Standard driver
 - Dominated by small I/O, with a few chunks in the 30-60 kbyte range
- Core driver
 - Entire file is written to memory, and appears to be flushed in nice 1 Mbyte chunks



Local Results

O_DIRECT=on			
	Standard driver	Core driver	
	7:30	0:22	Fairly consistent
O_DIRECT=off			
	Standard driver	Core driver	
	2:17	5:00	Best of 6
	4:30	6:33	Worst of 6



Discussion

- Need to figure out how to support user switchable O_DIRECT in production
- Investigating split virtual file driver in HDF5 layer

