A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

Benchmarking SSD-Based Lustre File System Configurations

Rick Mohr and Paul Peltz Jr.

National Institute for Computational Sciences
University of Tennessee

XSEDE14, July 13-18, 2014

Introduction

- Application Acceleration Center of Excellence (AACE) is a partnership with NICS, Cray, and Intel established in 2011
- AACE's Beacon Project
  - Funded by NSF and the State of Tennessee
  - Acquired two Intel Xeon Phi clusters for the exploration of MIC technology
  - #1 on the Green 500 (Nov 2012)
  - Nodes with SSDs for I/O experimentation
- How should the SSDs be used?
  - Allocate nodes to batch jobs?
  - Layer storage technology on top of SSDs?

ZFS/Lustre on SSDs

- Lustre is a good choice for users
  - Users are familiar with using it
  - Can use all I/O nodes to boost performance
  - Easy to share among multiple users
- ZFS is a good choice for SSDs
  - Copy-on-write provides wear leveling benefits
  - Potentially use compression to increase performance/capacity
- Lustre 2.4 adds ZFS support
  - Opportunity to compare ZFS vs. mdraid and ZFS vs. ldiskfs

Beacon Hardware

- 48x Compute Nodes
  - 2x 8-core Intel Xeon processors
  - 4x Intel Xeon Phi coprocessors
  - 256 GB RAM
  - 960 GB of local SSD
- 6x I/O Nodes
  - 2x 8-core Intel Xeon processors
  - 256 GB RAM
  - 16x Intel 710 SSDs (300 GB each)
  - 4x LSI SAS9211-4i RAID cards (4 disks each)
- FDR Infiniband Fabric
  - Bandwidth of 56 Gb/s

Beacon I/O Node Software

- CentOS 6.2
- Kernel 2.6.32-358.23.2.el6_lustre
  - Standard patched kernel supplied by Intel
- Lustre-2.4.3 (server)
- Lustre-1.8.9 (client)
- zfs-0.6.1
- e2fsprogs-1.42.7.wc2-7
- mdadm-3.2.2-9

Benchmarking Methodology

- Goals
  - Test hardware/software performance
  - Verify vendor claims
  - Gauge real-world performance
  - Identify bottlenecks or misconfigurations
- Need a systematic approach
  - Bottom-up testing (disk → Lustre)
  - Test individual components before testing combinations
  - Build up layer-by-layer

Benchmarking tools

- xdd-6.5
  - SSD and RAID benchmarking
  - Use multiple threads to saturate targets
- IOzone-3.420
  - ext4 and ZFS benchmarking
- ib_write_bw-2.6 (perftest-1.3.0-2.el6 rpm)
  - Benchmark RDMA over Infiniband
- lnet_selftest (Lustre 2.4.3)
  - Benchmark LNet performance
- IOR-3.0.1
  - Lustre benchmarking

SSD Benchmarks

- **Test individual drives**
  - Verify vendor claims and consistent performance
- **Test multiple drives**
  - Check scaling behavior and potential bottlenecks
- **Sequential I/O xdd flags:** -timelimit 60 -blocksize 512 -reqsize 2048 -passes 3 -dio -queuedepth 3 -seek sequential
- **Random I/O xdd flags:** -timelimit 60 -blocksize 4096 -reqsize 1 -passes 1 -dio -queuedepth 32 -seek random -seek seed $TIME
    - Before writes, issue PURGE command to reset drive
    - Add appropriate “-seek range” option base on SSD size
    - Run write command ~2hrs to precondition drive

SSD Benchmarks (Single Drive)

Test Type	Intel Specs	Benchmark Results
Sequential Read	270 MB/s	281 MB/s
Sequential Write	210 MB/s	219 MB/s
Random Read	38,500 IOPS	39,287 IOPS
Random Write	2000 IOPS	2260 IOPS

- **Drives perform slightly better than vendor specs**

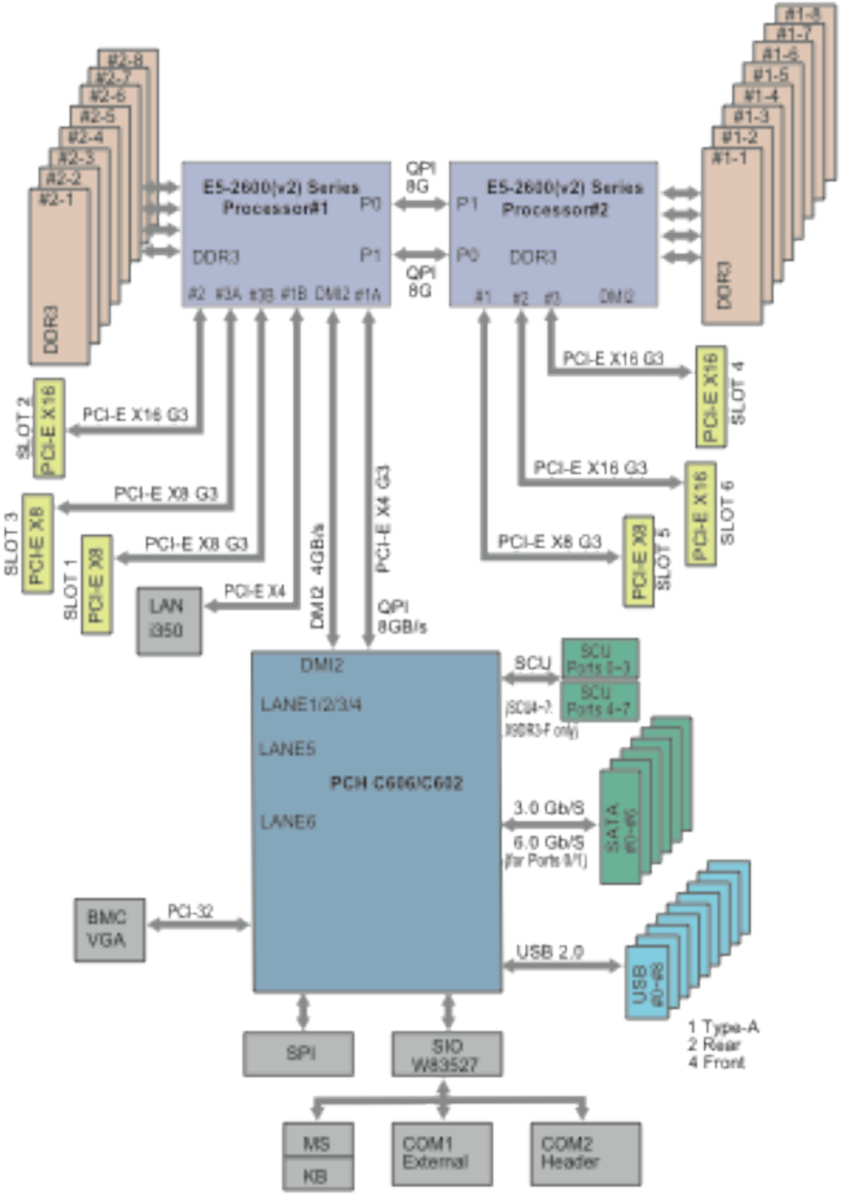
SSD Benchmarks (Multiple Drives)

- Scaling per RAID card
  - Test groups of 4 drives simultaneously
  - Results show good scaling
    - Single host example: 879 MB/s, 878 MB/s, 874 MB/s, 878 MB/s
- Scaling across all RAID cards
  - Test all 16 drives simultaneously
  - Write tests: 3542 MB/s ≈ 16 x 219 MB/s
  - Read tests: 3633 MB/s ≠ 16 x 281 MB/s (4496 MB/s)
  - Read speeds per drive vary from 134 MB/s to 285 MB/s

Sample xdd Results (16 SSDs)

T	Q	Bytes	Ops	Time	Rate	IOPS	Latency	%CPU	OP_Type	ReqSize
0	3	10148118528	9678	60.013	169.098	161.26	0.0062	0.01	read	1048576
1	3	10152312832	9682	60.016	169.160	161.32	0.0062	0.01	read	1048576
2	3	10149167104	9679	60.015	169.111	161.28	0.0062	0.01	read	1048576
3	3	11326717952	10802	60.016	188.729	179.99	0.0056	0.01	read	1048576
4	3	16871587840	16090	60.009	281.152	268.13	0.0037	0.01	read	1048576
5	3	16871587840	16090	60.009	281.153	268.13	0.0037	0.01	read	1048576
6	3	16885219328	16103	60.010	281.373	268.34	0.0037	0.01	read	1048576
7	3	16885219328	16103	60.009	281.378	268.34	0.0037	0.01	read	1048576
8	3	16873684992	16092	60.008	281.193	268.17	0.0037	0.01	read	1048576
9	3	16870539264	16089	60.011	281.126	268.10	0.0037	0.01	read	1048576
10	3	16883122176	16101	60.007	281.351	268.32	0.0037	0.01	read	1048576
11	3	16864247808	16083	60.010	281.022	268.00	0.0037	0.01	read	1048576
12	3	13420724224	12799	60.011	223.638	213.28	0.0047	0.01	read	1048576
13	3	13417578496	12796	60.012	223.583	213.23	0.0047	0.01	read	1048576
14	3	13407092736	12786	60.012	223.406	213.06	0.0047	0.01	read	1048576
15	3	13425967104	12804	60.010	223.729	213.36	0.0047	0.01	read	1048576

I/O Node Block Diagram



- "Slow" drives match the x16 PCI slots
- Tried moving IB and RAID cards
- Solution: Change BIOS settings
  - Configure x16 slot as 2-x8 slot
  - This works, but not sure why

RAID & File System Testing

- For different RAID levels, compare:
  - Standard Linux mdraid (RAID-0/5/6)
  - mdraid with ext4 file system
  - Equivalent ZFS configuration (zpool / raidz / raidz2)
- Focus on sequential read/write speeds
  - xdd for mdraid tests
    - Same command used for SSD testing except that queuedepth is 6 for writes and 10 for reads
  - IOzone for ext4/zfs tests
    - `iozone —ec -t8 -r1M -s100g -+n -i0 -i1`
- All RAID devices composed of 8 SSDs
  - Chosen to allow uniformity of OSTs and MDT

RAID & File System Results

	RAID-0	RAID-0 / ext4	zpool
Seq. Write	1701 MB/s	1406 MB/s	1466 MB/s
Seq. Read	2159 MB/s	1962 MB/s	1859 MB/s

	RAID-5	RAID-5 / ext4	raidz
Seq. Write	400 MB/s	338 MB/s	1236 MB/s
Seq. Read	1786 MB/s	1581 MB/s	1568 MB/s

	RAID-6	RAID-6 / ext4	raidz2
Seq. Write	319 MB/s	243 MB/s	1059 MB/s
Seq. Read	1773 MB/s	1532 MB/s	1401 MB/s

- **Based on these results, raidz was selected.**

Infiniband Testing

- Before testing Lustre, need to make sure interconnect is working as expected
- Use ib_write_bw to test IB RDMA speed
  - 5.9 GB/s compared to 6.8 GB/s (theoretical)
- Use lnet_selftest to check LNet performance
  - 5.4 GB/s (single client to single server)
  - Possibly higher with other Lustre tuning

Lustre Testing

- One MDS server, four OSS servers
  - MDS server has one MDT and one OST
  - OSS servers have two OSTs
  - Each OST is 8-disk raidz setup
  - MDT has 8 drives configured as a mirrored zpool
- Use IOR to test speeds
  - POSIX, file-per-process, 1 MB requests, 60 secs duration, stripe_count=1
  - Best performance: 9 clients, 3 processes per client, files evenly distributed over OSTs
- Results: 12.2 GB/s (writes), 12.1 GB/s (reads)

Future Work

- Metadata testing
- Random I/O benchmarking
- Compare Lustre/ZFS with Lustre/ldiskfs
- RAID-10 vs mirrored zpool
- ZFS compression/deduplication
- Investigate optimal tuning for mdraid
- System load monitoring
- Re-run tests with Lustre 2.5 clients

Conclusions

- Proper benchmarking requires:
  - Systematic approach
  - Time
- ZFS is the best choice…..for this case
  - Provides reliability with less performance loss
  - Hardware drives the software choice
- ZFS flexibility and features make it promising for Lustre deployments
- More work to be done

Questions?

This material is based upon work supported by the National Science Foundation under Grant Number 1137097 and by the University of Tennessee through the Beacon Project.