齐芯协力 驭算兴业
2014高性能计算合作伙伴峰会
2014 HPC Partner Summit

*Other names and brands may be claimed as the property of others.
Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. EXCEPT AS PROVIDED IN INTEL’S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT.

UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

This document contains information on products in the design phase of development. The information here is subject to change without notice. Do not finalize a design with this information. Contact your local Intel sales office or your distributor to obtain the latest specification before placing your product order.

All products, dates, and figures are preliminary for planning purposes and are subject to change without notice.

The code names Westmere, Sandy Bridge, Ivy Bridge, Knights Ferry and Knights Corner are presented in this document are only for use by Intel to identify products, technologies, or services in development, that have not been made commercially available to the public, i.e., announced, launched or shipped. They are not “commercial” names for products or services and are not intended to function as trademarks.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document, or other Intel literature may be obtained by calling 1-800-548-4725 or by visiting Intel’s website at http://www.intel.com.

Intel, the Intel logo, Xeon, Intel Core, Pentium, and Intel Xeon Phi are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

Copyright © 2012, Intel Corporation. All Rights Reserved.
Benchmarking a High Performance Lustre System

He, Wanqing
Michael Hebenstreit
October 14, 2014

*Other names and brands may be claimed as the property of others.*
what is CRT-DC

In an ongoing effort to provide customers with world class solutions in High Performance Computing (HPC), Intel Corporation has established the CRT Datacenter (CRT-DC)

The primary mission of the Intel CRT Datacenter (CRT-DC), located in Rio Rancho, New Mexico, is to support benchmarking in the HPC market segment

The cluster is known as Endeavour and is upgraded on a regular basis with the latest hardware and software.

As part of this role, the CRT Datacenter also supports benchmarking on pre-production hardware by both OEMs and end customers.

A secondary mission of the CRT Datacenter is to support HPC ISVs in testing their HPC applications.
The Endeavour Benchmarking Cluster

- /home secondary
- admin1
- admin2
- pbs-serv1
- pbs-serv2
- login
- compile

- 360 core nodes
- current technology
- Haswell EP
- 2 socket server
- 64 GB RAM
- 600 GB SAS HD

- FDR IB
- 1GEth

- 110 nodes
- additional node with var. hardware
- 64 GB RAM
- 800 GB SAS HD

- Panasas
- /home
- long-term storage
- DDN Lustre
- LFS9 (SSD)

*Other names and brands may be claimed as the property of others.*
CRT-DC & Lustre – a 6 year history

Started out with small systems using Intel storage boxes (12*3.5” SAS drives; 6 boxes each 1 OST; 1 MDT)

Used self build as well as commercial systems from DDN and Terrascala

Various benchmarking activities over time
<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Meta Data Server (MDS)</td>
<td>Intel® Server System R2224GZ4GC4</td>
</tr>
<tr>
<td></td>
<td>2 x Intel® Xeon® CPU E5-2680 @ 2.70GHz</td>
</tr>
<tr>
<td></td>
<td>64 GB memory</td>
</tr>
<tr>
<td></td>
<td>3 Raid controllers LSI Logic / Symbios Logic MegaRAID SAS 2208 [Thunderbolt] (rev 05)</td>
</tr>
<tr>
<td></td>
<td>6 OST (Targets) targets per server</td>
</tr>
<tr>
<td></td>
<td>Each target is 4 SSDs “Intel DC S3500, 600GB”</td>
</tr>
<tr>
<td></td>
<td>Mellanox ConnectX-3 FDR InfiniBand</td>
</tr>
<tr>
<td>8 OSS (Storage Server)</td>
<td>Intel® Server System R2224GZ4GC4</td>
</tr>
<tr>
<td></td>
<td>2 x Intel® Xeon® CPU E5-2680 @ 2.70GHz</td>
</tr>
<tr>
<td></td>
<td>64 GB memory</td>
</tr>
<tr>
<td></td>
<td>3 Raid controllers LSI Logic / Symbios Logic MegaRAID SAS 2208 [Thunderbolt] (rev 05)</td>
</tr>
<tr>
<td></td>
<td>6 OST (Targets) targets per server</td>
</tr>
<tr>
<td></td>
<td>Each target is 4 SSDs “Intel DC S3500, 600GB”</td>
</tr>
<tr>
<td></td>
<td>Mellanox ConnectX-3 FDR InfiniBand</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Raid setup in storage nodes

*Other names and brands may be claimed as the property of others.*
Test sizing

8 OSSs, each connected to backbone via FDR InfiniBand

=> Maximum OSS bandwidth is $8 \times 6\text{GB/s} = 48\text{ GB/s}$ (later achieved in tests over 40 GB/s I/O)

Single Client at maximum does 6GB/s (FDR speed)

=> So you need to test on at least 8 nodes in parallel

Depending on test single I/O thread does 50 to 900 MB/s

=> You need up to $48/0.05 = 960$ threads (or cores)

Each node has 24 HW cores => $960/24$ at minimum 40 nodes

Practical limits on single node performance => calculate to use 128 nodes

*Other names and brands may be claimed as the property of others.*
Test parameters

You need to test over a wide range of use cases

- 1-24 I/O threads per node
- 1-128 nodes
- LSF stripe size can vary (typically 1-4)
- Record size varies from 1kb to 4MB
- Iozone uses 8 different tests

Repeat each test at least 3 times to detect screw-ups

That’s a lot of testing; try to cut down the number of different tests with a good selection out of the possible tests

*Other names and brands may be claimed as the property of others.*
Good test programs

**dd**
- Simple to use
- No cluster model
- Huge synchronization problems

**lozone**
- Wide range of tests
- Even in cluster mode still some synchronization problems

**IOR**
- Only a few I/O models
Common pitfalls: caching issues

64 GB cache per node screws results (read results can be off by x10)

Use 100GB output file

- Works with fasts streaming tests
- Does not work well char based lozone tests (too slow)

Alternative - use a program to BLOCK memory

- Program calls malloc, memset and memlock
- Memory is blocked from use as cache
- Blocking 90% of 64 GB leaves 6GB caches, so file size can be 10GB
Common pitfalls: file size

On multi thread tests each node should always read/write the same amount.

Example

- File size 100 GB
- Single thread writes 100 GB
- Distributing between 10 I/O threads - each thread should do 10GB

Note: 128 nodes, 100 GB per node - complete test uses 12 TB
Common pitfall: synchronisation

Multiple clients compete for resources

Slight delays in startup create huge differences in results

Unless all client report similar results the aggregated performance is over-estimated!
## Selected Results from lozone test (compared SSD system against HDD solution)

<table>
<thead>
<tr>
<th></th>
<th>Test/Number of Nodes =&gt;</th>
<th>1</th>
<th>4</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SDD</strong></td>
<td><strong>initial_writers</strong></td>
<td>232</td>
<td>473</td>
<td>2523</td>
<td>4953</td>
<td>9644</td>
<td>20649</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>initial_writers</strong></td>
<td>105</td>
<td>404</td>
<td>2077</td>
<td>3839</td>
<td>3790</td>
<td>3768</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>readers</strong></td>
<td>663</td>
<td>2255</td>
<td>8518</td>
<td>16824</td>
<td>31129</td>
<td>44087</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>readers</strong></td>
<td>513</td>
<td>1733</td>
<td>6580</td>
<td>8454</td>
<td>2406</td>
<td>1977</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>reverse_readers</strong></td>
<td>534</td>
<td>1810</td>
<td>6323</td>
<td>9793</td>
<td>15569</td>
<td>28691</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>reverse_readers</strong></td>
<td>408</td>
<td>1556</td>
<td>5829</td>
<td>2564</td>
<td>1427</td>
<td>1560</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>stride_readers</strong></td>
<td>620</td>
<td>1910</td>
<td>7098</td>
<td>12541</td>
<td>20338</td>
<td>32631</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>stride_readers</strong></td>
<td>391</td>
<td>1604</td>
<td>5898</td>
<td>5780</td>
<td>2385</td>
<td>1654</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>random_readers</strong></td>
<td>384</td>
<td>1552</td>
<td>5636</td>
<td>4740</td>
<td>1927</td>
<td>1459</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>random_writers</strong></td>
<td>330</td>
<td>1100</td>
<td>3618</td>
<td>7592</td>
<td>14337</td>
<td>29007</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>random_writers</strong></td>
<td>146</td>
<td>666</td>
<td>2975</td>
<td>3719</td>
<td>3836</td>
<td>3611</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>mixed_workload</strong></td>
<td>484</td>
<td>1309</td>
<td>5404</td>
<td>9497</td>
<td>15840</td>
<td>27907</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>mixed_workload</strong></td>
<td>392</td>
<td>1063</td>
<td>4315</td>
<td>5047</td>
<td>4107</td>
<td>2171</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>fwriters</strong></td>
<td>363</td>
<td>961</td>
<td>3533</td>
<td>7738</td>
<td>13401</td>
<td>29276</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>fwriters</strong></td>
<td>98</td>
<td>440</td>
<td>2746</td>
<td>4165</td>
<td>4018</td>
<td>3894</td>
</tr>
<tr>
<td><strong>SDD</strong></td>
<td><strong>freaders</strong></td>
<td>713</td>
<td>2276</td>
<td>8616</td>
<td>16231</td>
<td>30582</td>
<td>44847</td>
</tr>
<tr>
<td><strong>HDD</strong></td>
<td><strong>freaders</strong></td>
<td>442</td>
<td>1769</td>
<td>6720</td>
<td>10225</td>
<td>2553</td>
<td>2231</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Selected Results from IOR test
(compared SSD system against HDD solution)

IOR test, Posix interface, 1MB record length
Aggregated Performance in MB/s

- SDD, read
- SDD, write
- HDD, read
- HDD, write

*Other names and brands may be claimed as the property of others.*
Findings

HDD and SDD solutions provide similar results for SINGLE clients.

Differences come out in SCALING.

SSDs do not suffer from access patterns like reverse directional I/O.

SDD are not as prone to performance losses due to concurrent accesses – aka the performance flattens out at some point, but does not diminish as much as HDD based solutions do.

Advise to datacenters – use SSDs for High Performance small size scratch systems, use HDD for large size storage solution.
Benchmark activities


https://communities.intel.com/docs/DOC-19265


*Other names and brands may be claimed as the property of others.