intel Look Inside.™

齐芯协力 驭算兴业

2014高性能计算合作伙伴峰会

2014 HPC Partner Summit

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.

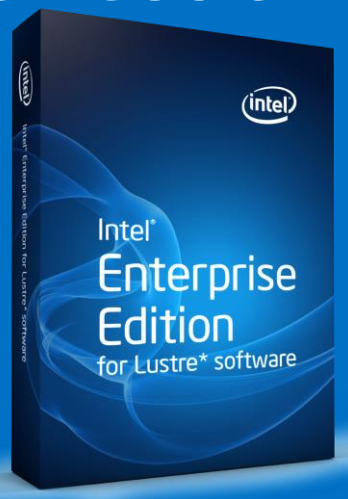
Intel, the Intel logo, Xeon, Intel Core, Pentium, and Intel Xeon Phi are trademarks of Intel Corporation in the U.S. and/or other countries.*Other names and brands may be claimed as the property of others.

Benchmarking a High Performance Lustre System

He, Wanqing

Michael Hebenstreit

October 14, 2014



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)

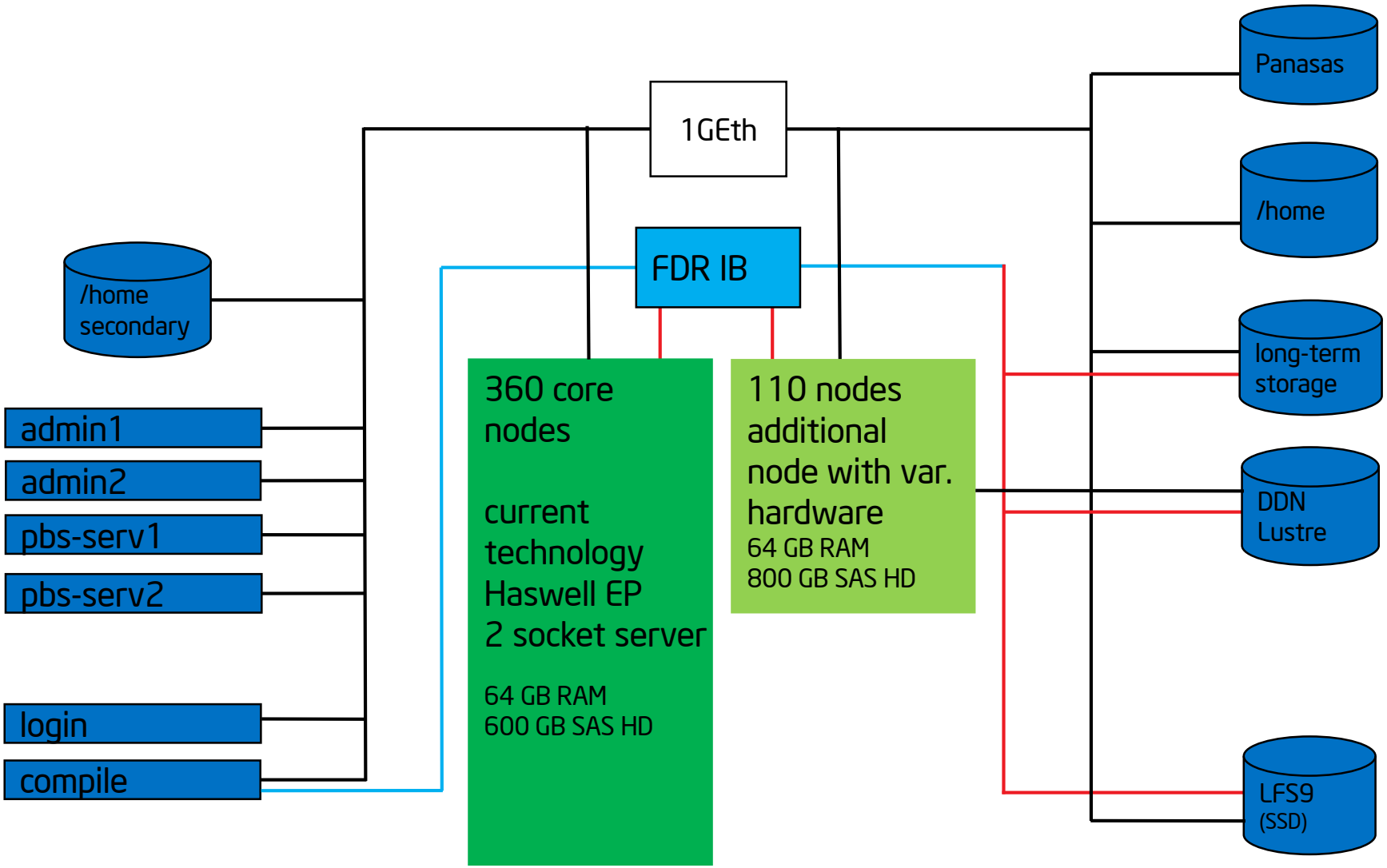
The primary mission of the Intel CRT Datacenter (CRT-DC), located in Rio Rancho, New Mexico, is to support benchmarking in the HPC market segment

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



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

Lustre system lfs09 installed Q4 2013

1 Meta Data Server (MDS)
Intel® Server System R2224GZ4GC4
2 x Intel®Xeon®CPU E5-2680 @ 2.70GHz

64 GB memory

3 Raid controllersLSI Logic / Symbios Logic MegaRAID SAS 2208 [Thunderbolt] (rev 05)
6 OST (Targets) targets per server
Each target is 4 SSDs "Intel DC S3500, 600GB "

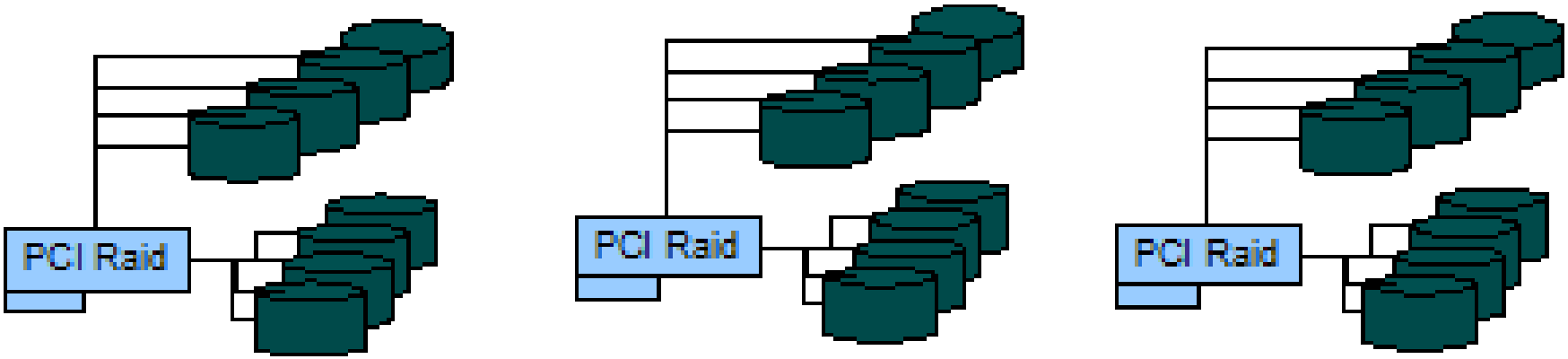
Mellanox ConnectX-3 FDR InfiniBand

8 OSS (Storage Server)
Intel® Server System R2224GZ4GC4
2 x Intel® Xeon®CPU E5-2680 @ 2.70GHz
64 GB memory

3 Raid controllersLSI Logic / Symbios Logic MegaRAID SAS 2208 [Thunderbolt] (rev 05)
6 OST (Targets) targets per server
Each target is 4 SSDs "Intel DC S3500, 600GB "

Mellanox ConnectX-3 FDR InfiniBand

Raid setup in storage nodes



Test sizing

8 OSSs, each connected to backbone via FDR InfiniBand

=> Maximum OSS bandwidth is 8 * 6GB/s == 48 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

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

Good test programs

dd

- Simple to use
- No cluster model
- huge synchronization problems

Iozone

- 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 Iozone 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

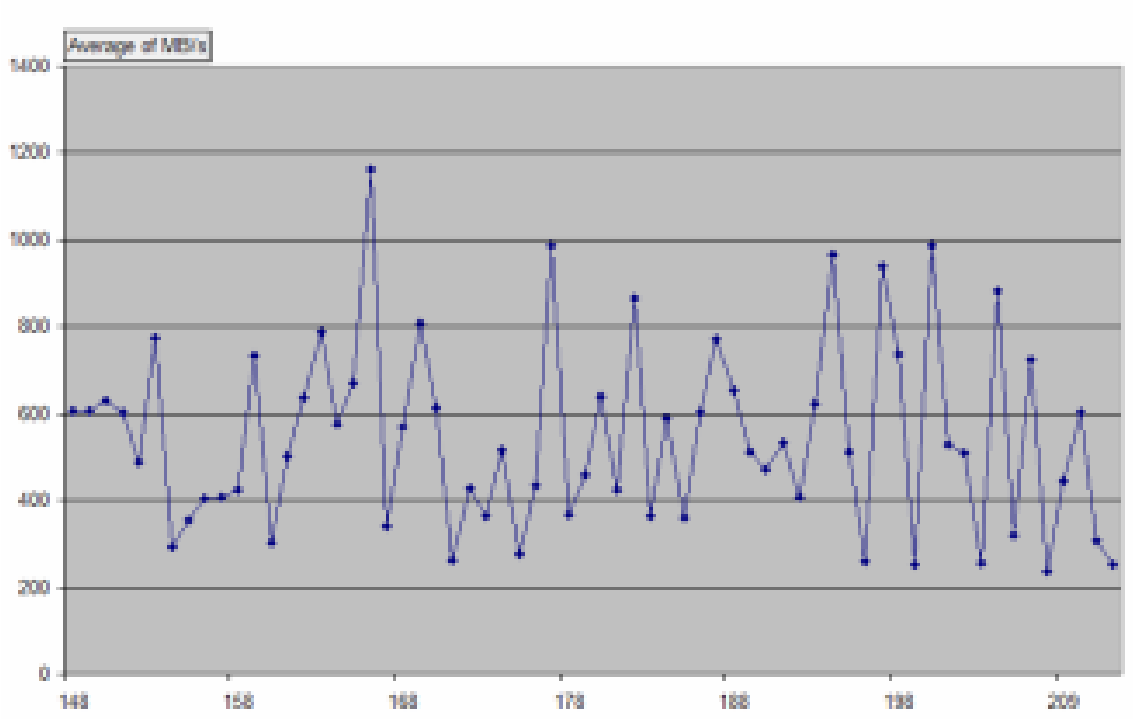
Note: 128 nodes, 100 GB per node – complete test uses 12 TB

Common pitfall: synchronisation

Multiple clients compete for resources

Slight delays in startup create huge differences in results

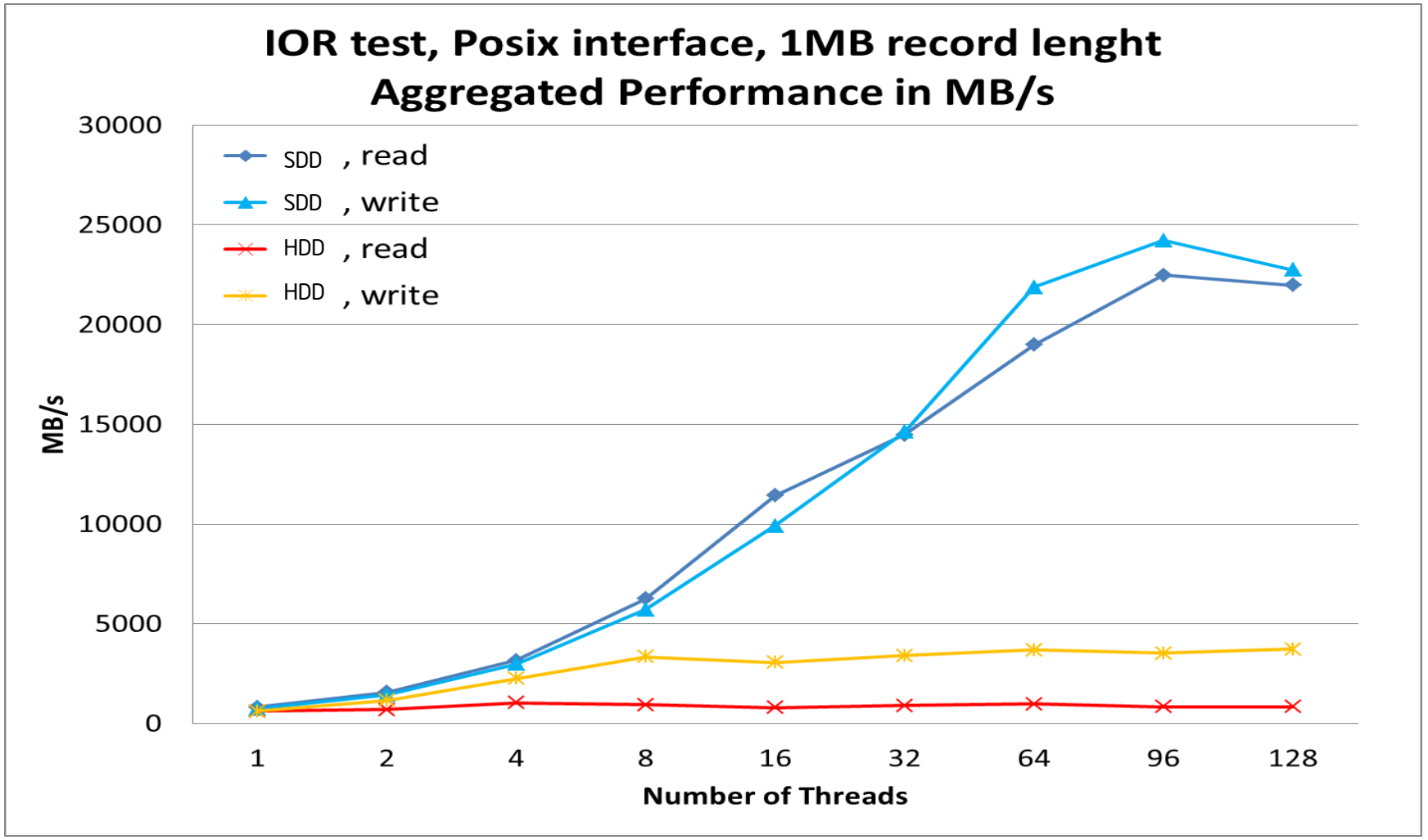
Unless all client report similar results the aggregated performance is over-estimated!



Selected Results from Iozone test (compared SSD system against HDD solution)

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

Selected Results from IOR test (compared SSD system against HDD solution)



Findings

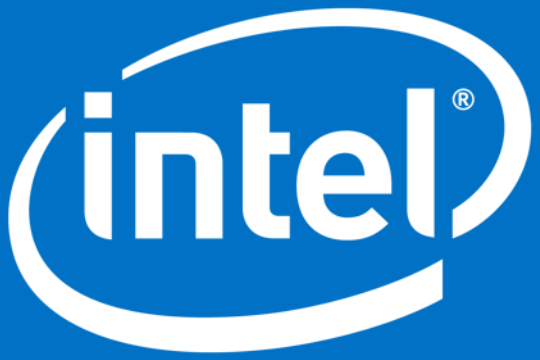
HDD and SDD solutions provide similar results for SINGLE clients

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


intel

Benchmark activites

http://software.intel.com/en-us/articles/performance-characterization-lustre-file-system-based-upon-intel-solid-state-disks/

http://software.intel.com/en-us/articles/performance-comparison-of-the-cluster-file-systems-at-the-intel-crt-dc/?wapkw=(performance+comparison)

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

http://www.intel.com/content/dam/www/public/us/en/documents/performance-briefs/lustre-cluster-file-system-performance-evaluation.pdf