Advanced Research Computing

New Storage Technologies - First Impressions: SanDisk IF150 & Intel Omni-Path

Brian Marshall

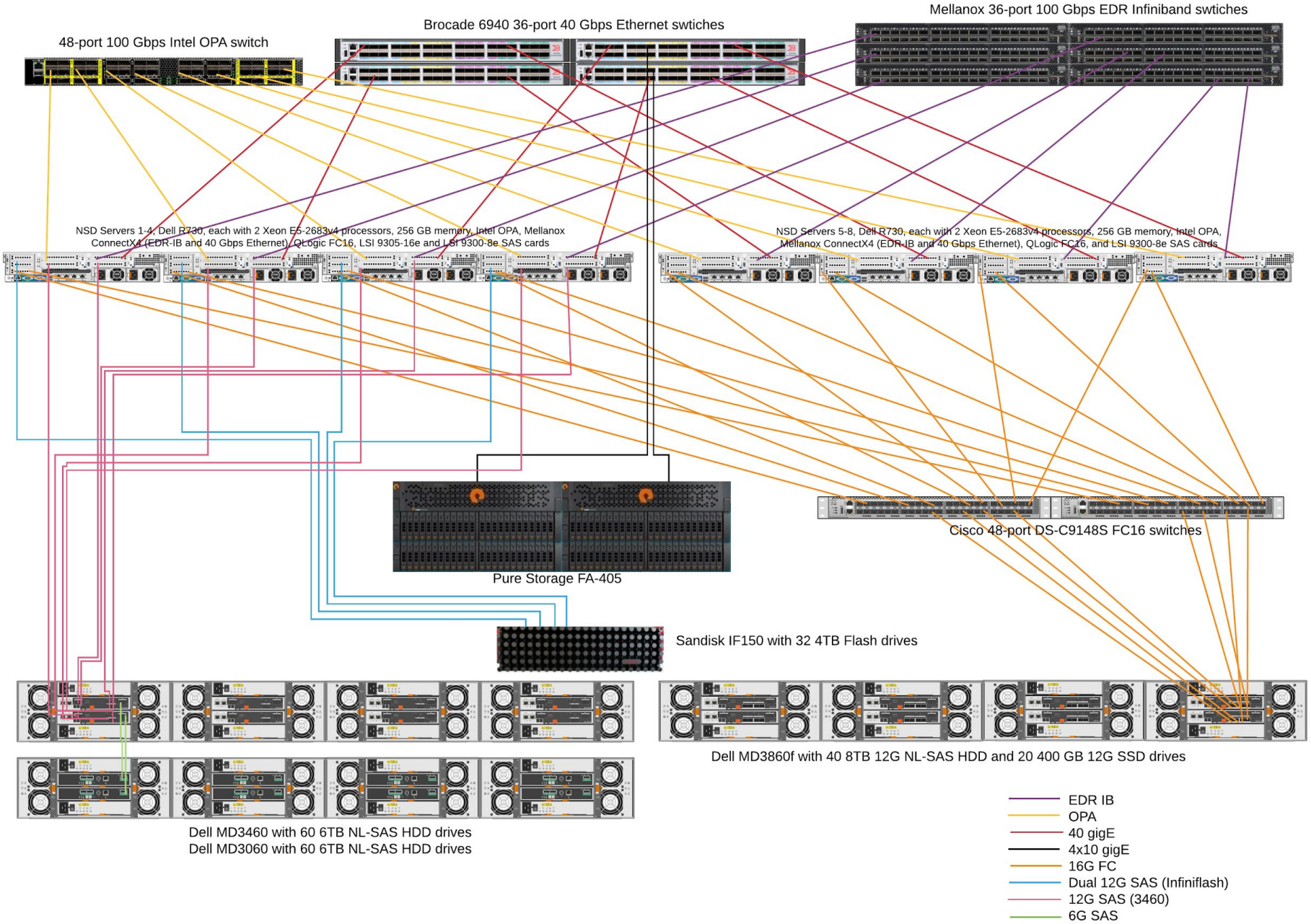
GPFS UG - SC16

November 13, 2016

Presenter Background

- Brian Marshall
- Computational Scientist at Virginia Tech - Advanced Research Computing
- Highly skilled in the compute side of HPC (accelerators, parallel programming, etc.)
- New to GPFS (and storage in general) but put some time in this Summer due to organizational needs

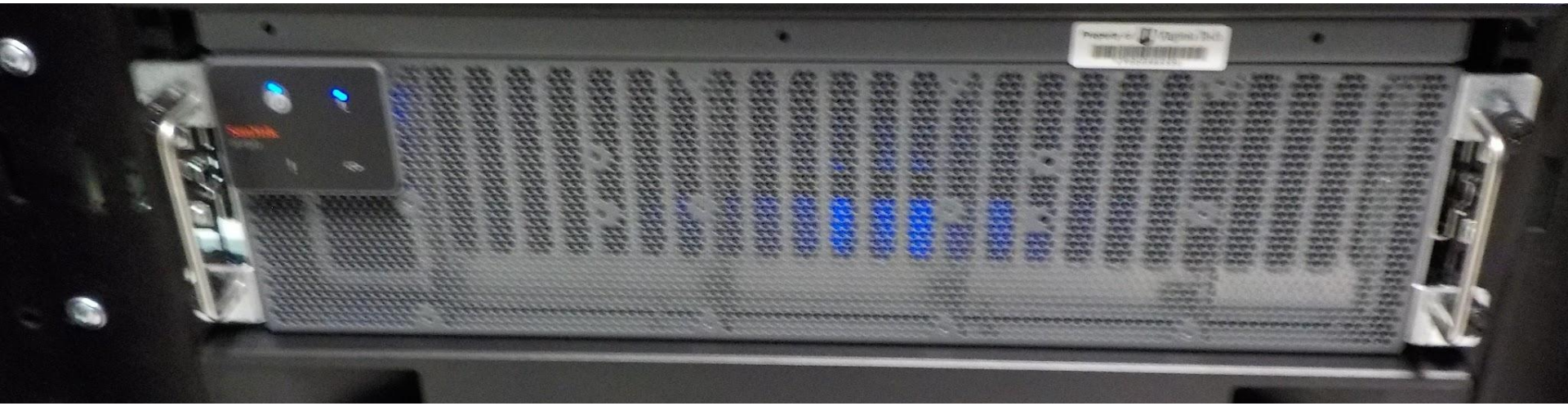
ARC Storage - ClaytorLake FY15 and FY16



Virginia Tech
Advanced Research Computing

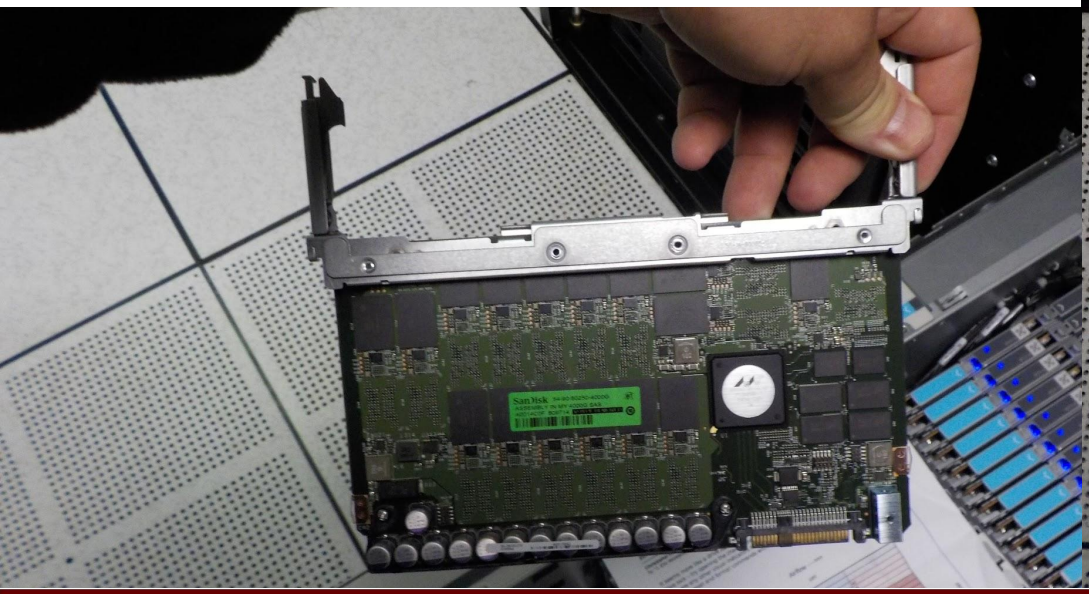
SanDisk IF150

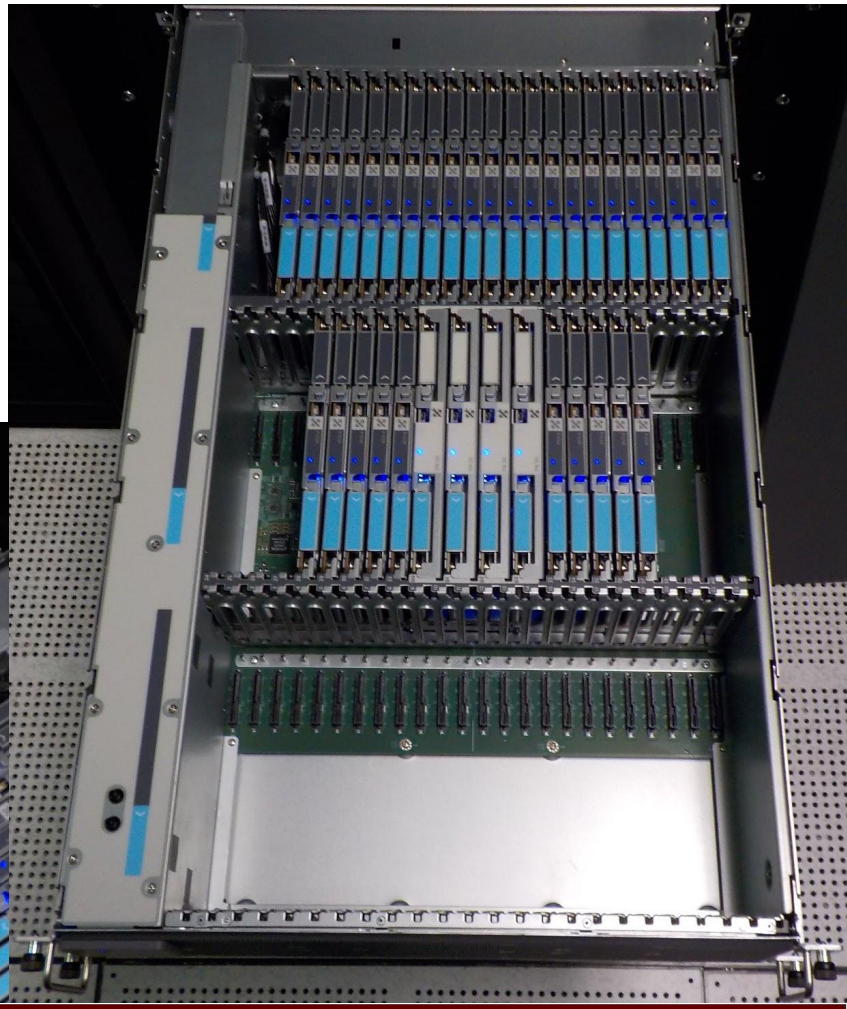
- Same hardware as used in the DeepFlash product
- JBOF - Just a Bunch Of Flash
- Tray can hold up to 64 8TB or 4TB SSD
- 8 x 12 GBps SAS connections



More Pictures

ARC has 32 4 TB drives => 128





Theoretical Numbers

- Each SSD has 2 read connections, each can do 250 MBps
- VT - 32 SSD * 2 connect * 250 MBps = 16 GBps
- drive side expansion board (DSEB) limits to 13 GBps
- To engage both SSD connections you need to SAS cables out of the box.
- BUG - A Marvell firmware bug prevents to different PCs from talking to a SSD simultaneously.
- VT - We did dual SAS connections to 4 NSD Servers instead of single connection to 8 NSD servers. Alternative is to populate all slots for max bandwidth.

Filesystem Configuration

- SSDs have a 8KB page size. SanDisk recommends doing at least a 4KB low level blocksize (smallest chunk of I/O). Make GPFS subblock equal to pagesize. SanDisk has not seen improvement in throughput beyond 256KB GPFS blocksize

  32 * 8 KB = 256KB (or larger) blocksize.

- SanDisk recommends using data replication over RAID and is investigating GPFS Native RAID (GNR) for the future.

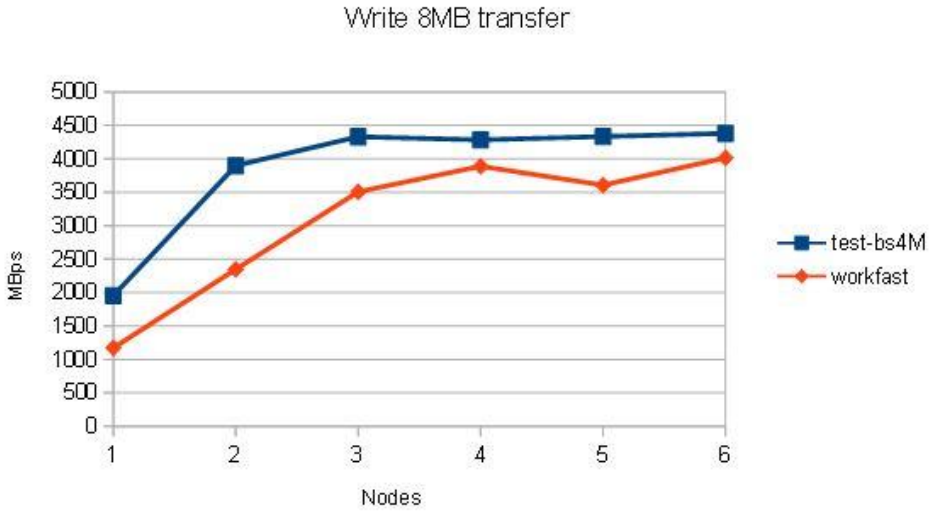
The Installation

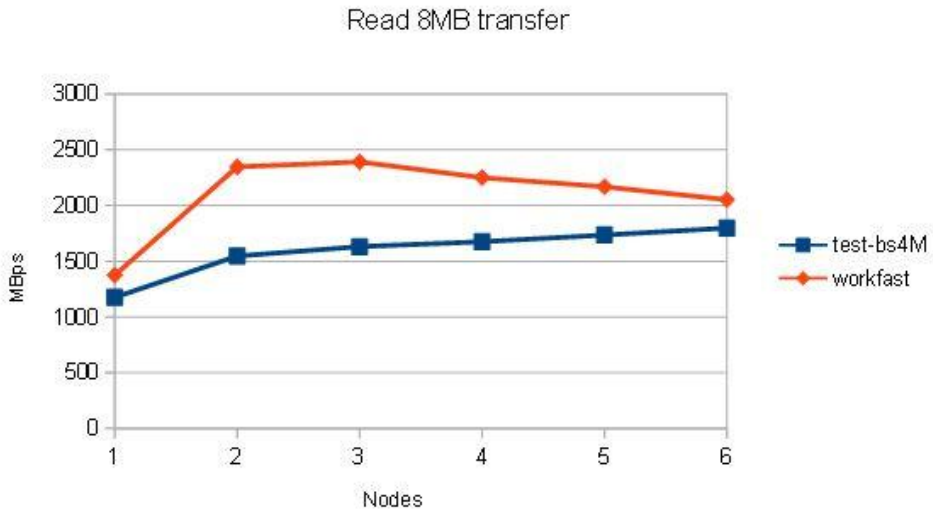
- Mostly painless
- SanDisk provides utilities to setup most OS settings
- Make sure you update firmware on SAS cards
- Up and running after about 6 hours

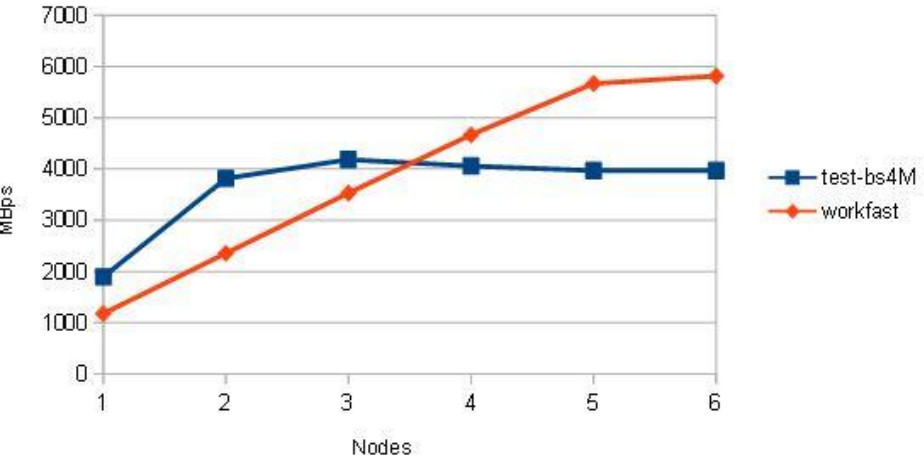
Advanced Research Computing

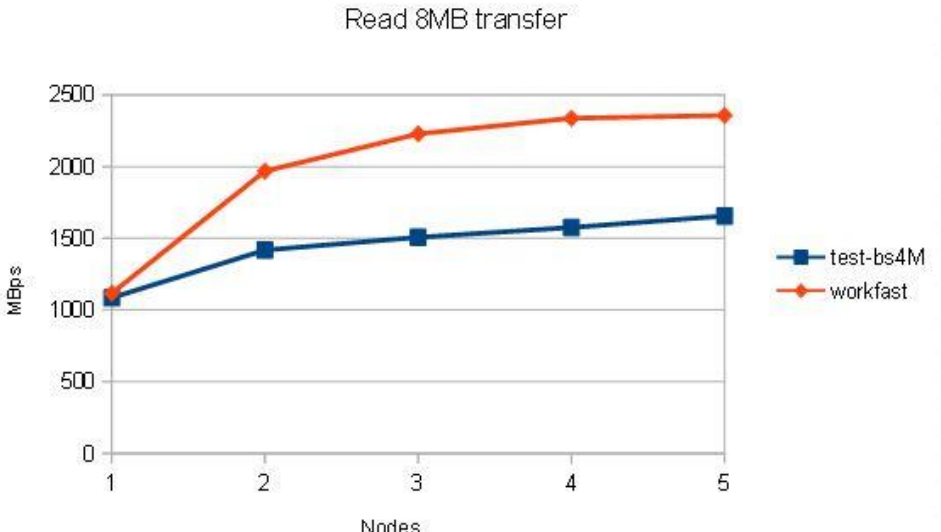
Benchmark - Dual 10gigE

Only 3 NSD Servers & network read issues. Bottom NSD Servers on EDR









Benchmark - EDR IB

Using 2 Nodes connected via EDR IB

- Max Write 6800 MBps
- Max Read 11300 MBps

mdtest Results

Metadata is stored on Pure Storage FA-405, so these numbers are not that interesting but included for completeness

```
mdtest-1.9.4-rc was launched with 3 total task(s) on 3 node(s)
Command line used: /home/mimarsh2/newriver/ior-benchmarks/mdtest/mdtest -n 1000 -i
5 -d /gpfs/workfast/mdtest-runs/mdtest-runs
Path: /gpfs/workfast/mdtest-runs
FS: 111.8 TiB   Used FS: 9.4%   Inodes: 83.8 Mi   Used Inodes: 0.1%

3 tasks, 3000 files/directories

SUMMARY: (of 5 iterations)
   Operation                     Max            Min           Mean        Std Dev
   ---------                     ---            ---           ----        -------
   Directory creation:         116.175        110.609        113.068          1.910
   Directory stat    :      150731.465     124751.269     141876.893       9400.443
   Directory removal :         119.422        109.833        114.539          3.041
   File creation     :        7789.348       5450.310       6684.189        743.539
   File stat         :     1005587.149     970828.794     995085.346      12854.880
   File read         :      418899.794      34695.019     258393.663     150006.606
   File removal      :       68477.687      21224.695      53210.109      17316.715
   Tree creation     :        9404.269       7002.177       8690.938        922.063
   Tree removal      :          66.242         25.183         49.289         17.728
```

Real-World Application Results

To Be Continued....

When moving from dev to production, don’t forget to flash the firmware on ALL SAS cards in NSD servers

Intel Omni-Path

- Intel's next generation interconnect
- 100 Gbps throughput; low latency
- Competitor to EDR Infiniband
- Go to the Intel booth for more info

OPA Storage Connection

You just bought a new OPA connected compute cluster. How do you connect it to storage?

Options

1. 10 gig Ethernet only
2. Buy a new storage contain with OPA
3. Build a Omni-Path Storage Router
4. Install OPA cards in current NSD Servers

ARC chose #4

ARC NSD Servers now have a single port OPA card and a dual port EDR IB card (1 port used as 40 gig Ethernet)

The 40gig Ethernet network is connected to all clusters; EDR and OPA serve as fast networks to some clusters

OPA + EDR NSD Server

1. Use the Intel Storage Router Design as a guide for deploying Mellanox EDR IB and Intel OPA in the same system. Don't use Mellanox MXM
2. GPFS does not support 2 RDMA fabrics; we stayed with EDR for RDMA
3. Keep the daemon and admin networks on 10 gig Ethernet because OPA is new
4. Use subnets to specify the IP over Fabric networks to use OPA for data

Benchmark

2 Broadwell nodes writing to HDD scratch filesystem

- Max Write 7710 MBps
- Max Read 3094 MBps

Acknowledgements

Thank you to all that supported these efforts:
Eric Wonderley ←In the audience
Valdis Kletnieks
Chris Snapp
Josh Akers
Brandon Sawyers
Chris Konger
Gary Hess
Mike Moyer
Vijay Agarwala
Wanda Baber
Tim Rhodes
Umar Kalim
Christopher Howard - SanDisk
Ali Ahmed - Intel
Lindsay Todd - IBM

Questions

???

ARC Wants You!!

If you thought this was fun and interesting, Virginia Tech ARC has 2 systems engineer positions open