

JLab “SciPhi-XVI” KNL Cluster



Thomas Jefferson National Accelerator Facility

Sandy Philpott

JLab High Performance Computing

<https://lqcd.jlab.org>

HEPiX Wigner April 26, 2017

Recall from last talk...

The USQCD project's FY16 procurement was installed at JLab in August:

3 racks of 64 7230 KNL nodes, ~ \$ 1M, 49TF USQCD benchmarks

In FY17 (October 3), exercised the upgrade option in the award, for a 4th rack, ~ \$ 0.25M, installed late October

Competitive procurement...

- Intel Xeon Phi / Knights Landing
- NVIDIA Pascal GPU, CUDA
- Intel Broadwell CPU server

Consideration factors

- hardware availability timeline
- high speed network – 100 Gbps price/performance; Omni-Path or Infiniband
- available configurations
- **reflective benchmarks of USQCD codes**

Strategy: optimize portfolio of machines to get the most science on the portfolio of applications

KNL proposals achieved highest aggregate score on all performance metrics and sub-metrics

Updates since Fall16 HEPiX

4th rack of 64 nodes installed, total of 264 nodes

- and more temp location power circuits

Top500 #397! Green500 #10!

- Relocation into Hot Aisle Containment System in December
- MCDRAM flat/cache reboots (sometimes hangs?!)
 - daemon to switch not yet automated...
- Multi-node job insight
 - Reboot/power cycle between every job?
 - That's hard on the disks ...
- Early-ship KNL silicon ... 26 nodes have been replaced
- User benchmarks ... are sometimes hard to get ...
- Testbed – into production July 1 for project year start
 - CentOS 7.3, Xppsl 1.5.1, Lustre 2.9 client, OPA 10.3
 - New production BIOS soon: F01.02.0049 ... after processor screenings

Top500 #397!

SciPhi XVI - KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path

Site:	Thomas Jefferson National Accelerator Facility
System URL:	http://qed.jlab.org/
Manufacturer:	Koi Computers
Cores:	16,896
Linpack Performance (Rmax)	425.868 TFlop/s
Theoretical Peak (Rpeak)	702.874 TFlop/s
Nmax	2,457,600
Power:	111.00 kW (Submitted)
Memory:	54,912 GB
Processor:	Intel Xeon Phi 7230 64C 1.3GHz
Interconnect:	Intel Omni-Path
Operating System:	Linux
Compiler:	Intel Parallel Studio 2017
Math Library:	Intel MKL
MPI:	Intel MPI 5.1.3.210

RANKING

List	Rank	System	Vendor	Total Cores	Rmax (TFlops)	Rpeak (TFlops)	Power (kW)
11/2016	397	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	Koi Computers	16,896	425.9	702.9	111.00

Green500 #10 !

Green500 List for November 2016

Listed below are the November 2016 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site	System	Total Power(kW)
1	9462.1	NVIDIA Corporation	NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100	349.5
2	7453.5	Swiss National Supercomputing Centre (CSCS)	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100	1312
3	6673.8	Advanced Center for Computing and Communication, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	150.0
4	6051.3	National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	15371
5	5806.3	Fujitsu Technology Solutions GmbH	PRIMERGY CX1640 M1, Intel Xeon Phi 7210 64C 1.3GHz, Intel Omni-Path	77
6	4985.7	Joint Center for Advanced High Performance Computing	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	2718.7
7	4688.0	DOE/SC/Argonne National Laboratory	Cray XC40, Intel Xeon Phi 7230 64C 1.3GHz, Aries interconnect	1087
8	4112.1	Stanford Research Computing Center	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190
9	4086.8	Academic Center for Computing and Media Studies (ACCMS), Kyoto University	Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect	748.1
10	3836.6	Thomas Jefferson National Accelerator Facility	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	111

KNL Cluster Hardware

264 Intel Xeon Phi 7230

64 cores, 1.3GHz, 256 threads, 192GB RAM, 16GB MCDRAM

1 TB Seagate disk

Intel Omni-Path 100gbps

Total: 16,896 cores, >50 TB memory

Server Board:

Intel S7200AP

BIOS: S72C610.86B.01.01.0190.080520162104 (08/05/2016)

KNL Cluster Software

CentOS 7.2:

3.10.0-327.22.2.el7.xppsl_1.4.1.3272.x86_64

openMPI:

- openmpi-1.10.0-10.el7.x86_64
- openmpi_intel_hfi-1.10.2-8.x86_64
- openmpi_gcc-1.10.2-8.x86_64
- openmpi_gcc_hfi-1.10.2-8.x86_64
- openmpi_pgi_hfi-1.10.2-8.x86_64

Mvapich2

- mvapich2_gcc_hfi-2.1-4.x86_64
- mvapich2_pgi_hfi-2.1-4.x86_64
- mvapich2_intel_hfi-2.1-4.x86_64
- mvapich2_gcc-2.1-4.x86_64

Intel Parallel Studio

- parallel_studio_xe_2016.3.067
- parallel_studio_xe_2017

Memory options

Changing nodes with reboot to update BIOS to set MCDRAM to Flat or Cache mode per user job tag...

latest version of syscfg

(hwloc_dump_hwdata showed different settings?!)

works ... but nodes would often randomly hang during the process, remaining down until a required a power cycle ...

Turns out, Intel requires a CPU reset (IPMI “power cycle”) anyway to guarantee the memory setting

New solution in the works ...

Update the BIOS, check the chip set; if “old” – update ...

But this takes several hours of manual intervention on each node! Final arrangements await ...

Omni-Path

Spine : 4 64 port switches; Leaf switch uplinks in bundles of 6,6,4
for 256 nodes, 2 core switches consumed, 3rd supports 4 links to the file system plus 8 additional nodes with a little room to spare

32 nodes connect to a 48 port leaf switch

- nominally 2:1 oversubscribed, but effectively full bandwidth per node

Need more than a single core to drive OPA to full bandwidth

/dev/hfi1 driver – use the correct one, or

hfi_userinit: PSM2 and driver version mismatch

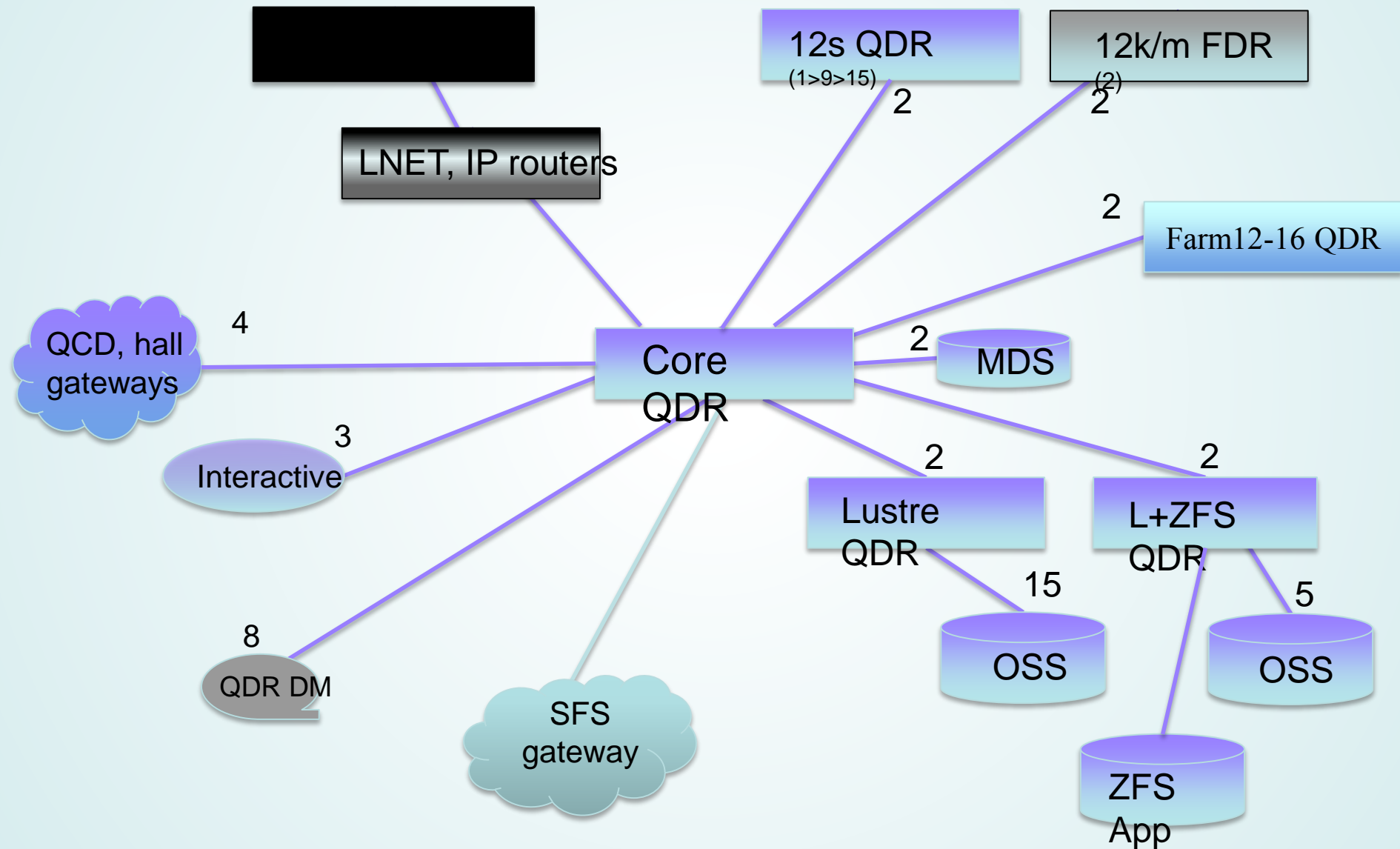
Driver initialization failure on /dev/hfi1 (err=23)

[3] MPI startup(): tmi fabric is not available and fallback fabric is not enabled

Managed OPA switch subnet manager stopped running several times;
recommended for <100 nodes

so, installed subnet manager on 2 hosts

Scientific Computing Network Fabrics



OPA to IB File Systems

Omni-Path 16p nodes need to reach Lustre and NFS file services on the Infiniband QDR fabric...

Lustre: /cache, /volatile

Use 4 LNet routers between OPA and IB

NFS : /home, /dist, /work

Use 2 JLab hosts with IB and OPA connections, with IP routing

Still To Do...

- Benchmark users' USQCD codes relative to conventional cores
 - Currently assuming 3 Jpsi core hour, the reference for USQCD
- Bin nodes into fast/medium/slow, with 80% in medium, so jobs land on consistently performant nodes (we've seen >10% spread) ?
- Automate cache/flat memory reboots based on job demands
 - PBS/Torque/Maui
- Testbed into production by end of June, software and BIOS updates
- Investigate Zonesort kernel module, Intel Cluster Checker tool
- Continue to share experiences with other KNL sites

Fall16 HEPiX HPC BOF

To share experiences between large sites and smaller sites

- Procurements
- Allocations
- Benchmarks
- Software
- Porting
- Lessons learned

Follow-on meetings as attendance permits ...