2016 Acquisition Status

April 2016 All Hands Meeting

Chip Watson
Jefferson Lab

Outline

Current Status of Acquisition

More KNL Details

Award Options

2016 LQCD Machine

FY16 & partial FY17 funds now planned for JLab yield a two phase procurement of ~$1.3M, aiming at operations by October 1 and January 1 respectively (we are trying for a month earlier for each).

Process:

The timeline & process is the same as previous years (alternatives analysis, acquisition planning, best value procurement process). The goal is also the same:

Optimize the portfolio of machines to get the most science on the portfolio of applications.

Current Status of the Procurement

The Request For Procurement (RFP) process is underway at JLab, with a published Statement of Work, to which vendors will respond within a few weeks.

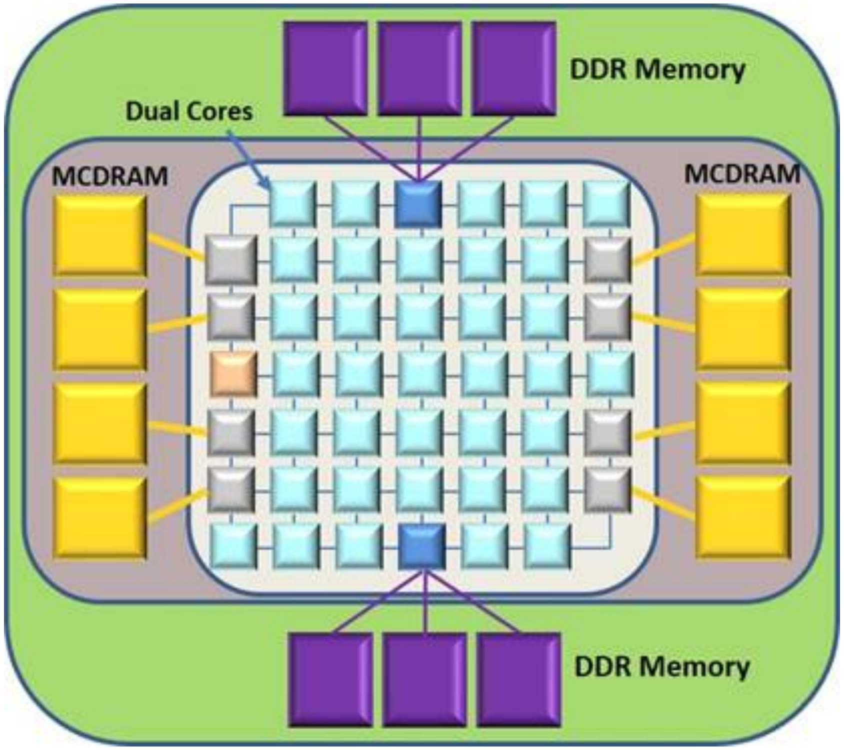
The Statement of Work (SOW) is identical to anticipated in the Alternatives Analysis and Procurement Plan, with 2 alternatives being evaluated for best value:

1. Xeon Phi Knights Landing (homogeneous)
2. Mixed 50:50 (by dollars) GPU & conventional x86

While this procurement heavily weights benchmark performance, "Best Value" allows us to take into account many other facets, including software readiness. Post evaluation, award is not restricted to 50:50.

Knights Landing

General: single socket x86 processor, with memory and PCI bus to get to network adapters, disk, etc.



- ✓ Binary compatible ISA
- ✓ AVX-512 (twice Xeon width)
- ✓ Out-of-order execution
- ✓ Advanced branch prediction
- ✓ Scatter gather, unaligned R/W
- ✓ 16 GB on package MCDRAM
- ✓ 6 DDR4 ports "up to 384 GB"
- ✓ 1 MB L2 cache per 2 core tile

(figure shows up to 72 cores if all are real & operational)

software.intel.com/en-us/articles/what-disclosures-has-intel-made-about-knights-landing

KNL Performance Details

3+ Teraflops double, 6+ Teraflops single

- SPECint*_rate_base2006: at least >~0.6x perf of 2-socket Intel® Xeon® processor E5-2697v3 (Haswell, 14 cores, 2.6 GHz)
- SPECfp*_rate_base2006: at least >~0.8x perf of 2-socket Intel® Xeon® processor E5-2697v3

Implication: for running lots of small integer or floating point intensive serial processes, this chip will perform like a dual socket system of roughly comparable price (the chips above are very expensive chips).

Running on early access silicon confirms that the performance of KNL is good for LQCD. Details are regrettably NDA.

KNL Additional Details

Memory: 6 DDR4 channels, 8 high bandwidth on package

- 90 GB/s to main memory, “sustained”
- “over 400 GB/s streams” (confirmed)
- Configurations: 96 GB, 192 GB, 384 GB

Actual cores will be 64+

- Can be divided into 4 “quadrants” of 16+
- Can be treated as 4 NUMA zones (like 4 socket server)

(we have run 4 openMP processes in this mode)

PCIe gen 3 x 16

- Can support either 100g Infiniband, or 100g Intel Omnipath fabric (SOW specifies this speed as a requirement)

GPU Info

The new Pascal GPU appears to be too late for this procurement, therefore k80 is the most likely GPU option:

- Quad k80 => 8 GPU chips / node
- Paired with 16 core Broadwell x86, preserves 4 cores / GPU

(Pascal is a powerful GPU, and from public information is more powerful per socket than KNL. It also appears to be more expensive per socket. The only data point I have is that an 8 GPU server is being offered for ~ $128K, $16K/socket, standard discounts expected.)

We will take into account USQCD's ability to absorb more GPUs, and will take into account anticipated BNL k80 resources. (This still leads to a total reduction in GPU resources if KNL is selected.)

Procurement Benchmarks

Based upon anticipated usage, performance will not be solely LQCD solvers, as analysis is becoming a larger and larger factor in USQCD resources. SOW specifies the following benchmarks:

- Memory bandwidth, and transfer speeds between memories
- Dslash performance (as a proxy for all inverters)
- Contraction / distillation: execution speed of an analysis code trace with a memory footprint for a set of 384x384 matrices that is much larger than 16 GB; key is batched zgemm performance

  code does application paging (least recently used ejection and restore on demand); analysis will assume overlap of serial code, transfer to/from fast memory, and zgemm execution time
- Network bandwidth

Award Considerations

Host memory size

- KNL: 96 GB or 192 GB (15% upgrade cost/node, just a guess)
- GPU host: 256 GB or 512 GB (maybe 10% cost/node)

Compute vs. Disk vs. Memory

- Highest priority is to reach each multiple of 32 sockets with FY16 + FY17 funds
- Next priority is total disk while retiring old servers; current plan is to add 256 TB this summer and another 256 TB in the Fall, but either could be doubled for ~$30K, losing a few KNL or GPUs
- Large memory footprint is another goal; if we buy 128+ KNL nodes, base would be 14 TB (128 nodes x 96+16 GB/node); largest today is pi0 256 nodes x 128 GB/node = 32 TB

Please help the project in making the best selection
by providing your input!