Cori (2016) and Beyond – Ensuring NERSC Users Stay Productive

Nicholas J. Wright
Advanced Technologies Group Lead

Heterogeneous Multi-Core 4 Workshop
17 September 2014
NERSC Systems Today

**Edison: 2.39PF, 333 TB RAM**
Cray XC30 5,192 nodes, 125K Cores

**Hopper: 1.3PF, 212 TB RAM**
Cray XE6 6,384 nodes 150K Cores

**Data-Intensive Systems**
Carver, PDSF, JGI, KBASE, HEP
14x QDR

**Vis & Analytics**
Data Transfer Nodes
Adv. Arch. Testbeds
Science Gateways

**Ethernet & IB Fabric**
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

**Global Scratch**
3.6 PB
5 x SFA12KE

**/project**
5 KB
DDN9900 & NexSAN

**/home**
250 TB
NetApp 5460

**HPSS**
50 PB stored, 240 PB capacity, 20 years of community data

**Cray XE6**
6,384 nodes 150K Cores

**Cray XC30**
5,192 nodes, 125K Cores

**7.6 PB Local Scratch 163 GB/s**
16 x FDR IB
80 GB/s

**2.2 PB Local Scratch 70 GB/s**
16 x QDR IB
50 GB/s

**16 x FDR IB**
12 GB/s

**2 x 10 Gb**
1 x 100 Gb

**Software Defined Networking**

**3.6 PB**
5 x SFA12KE

**5 PB**
DDN9900 & NexSAN

**250 TB**
NetApp 5460

**50 PB stored, 240 PB capacity, 20 years of community data**

**Hopper: 1.3PF, 212 TB RAM**
Cray XE6 6,384 nodes 150K Cores

**Data-Intensive Systems**
Carver, PDSF, JGI, KBASE, HEP
14x QDR

**Vis & Analytics**
Data Transfer Nodes
Adv. Arch. Testbeds
Science Gateways

**Ethernet & IB Fabric**
Science Friendly Security
Production Monitoring
Power Efficiency
WAN
NERSC-8 (Cori) Mission Need

The Department of Energy Office of Science requires an HPC system to support the rapidly increasing computational demands of the entire spectrum of DOE SC computational research.

- Provide a significant increase in computational capabilities, at least 10 times the sustained performance of the Hopper system on a set of representative DOE benchmarks.
- Delivery in the 2015/2016 time frame.
- Provide high bandwidth access to existing data stored by continuing research projects.
- Platform needs to begin to transition users to more energy-efficient many-core architectures.
Many Core Architectures: Why Now?

To sustain historic performance growth, the DOE community must prepare for new architectures

Hennessy, Patterson. Computer Architecture, a Quantitative Approach. 5th Ed. 2011

“Moore’s Law” will continue.

- Memory wall
- Power wall
- ILP wall

Exponential performance increase; do not rewrite software, just buy a new machine!

hmmm...

oh no!
Why can’t we keep doing what we’ve been doing?

Optimization target for hardware has evolved to new direction (but pmodels have not kept up)

Old Constraints

• Peak clock frequency as primary limiter for performance improvement
• Cost: FLOPs are biggest cost for system: optimize for compute
• Concurrency: Modest growth of parallelism by adding nodes
• Memory scaling: maintain byte per flop capacity and bandwidth
• Locality: MPI+X model (uniform costs within node & between nodes)
• Uniformity: Assume uniform system performance
• Reliability: It’s the hardware’s problem

New Constraints

• Power is primary design constraint for future HPC system design
• Cost: Data movement dominates: optimize to minimize data movement
• Concurrency: Exponential growth of parallelism within chips
• Memory Scaling: Compute growing 2x faster than capacity or bandwidth
• Locality: must reason about data locality and possibly topology
• Heterogeneity: Architectural and performance non-uniformity increase
• Reliability: Cannot count on hardware protection alone

Fundamentally breaks our current programming paradigm and computing ecosystem
Cori Configuration

• **64 Cabinets of Cray XC System**
  -- Over 9,300 ‘Knights Landing’ compute nodes
    • Self-hosted (not an accelerator)
    • Greater than 60 cores per node with four hardware threads each
    • 64-128 GB memory per node
    • High bandwidth on-package memory
  -- Over 1,900 ‘Haswell’ compute nodes
    • Data partition
  -- 14 external login nodes
  -- Aries Interconnect (same as on Edison)
  -- 10x Hopper sustained performance using NERSC SSP metric

• **Lustre File system**
  -- 28 PB capacity, 432 GB/sec peak performance

• **NVRAM “Burst Buffer” for I/O acceleration**
• **Significant Intel and Cray application transition support**
• **Delivery in mid-2016; installation in new LBNL CRT**
Stand-Alone Processor vs. CoProcessor

**Older Knights Corner Co-Processor**

Next-Generation Knights Landing *stand-alone* processor:
No “offloading” bottlenecks
Intel “Knights Landing” Processor

• Next generation Xeon-Phi, >3TF peak
• Single socket processor - Self-hosted, not a co-processor, not an accelerator
• Greater than 60 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
• Intel® "Silvermont" architecture enhanced for high performance computing
• 512b vector units (32 flops/clock – AVX 512)
• 3X single-thread performance over current generation Xeon-Phi co-processor
• High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory
• Higher performance per watt
Knights Landing Integrated On-Package Memory

**Cache Model**
Let the hardware automatically manage the integrated on-package memory as an “L3” cache between KNL CPU and external DDR

**Flat Model**
Manually manage how your application uses the integrated on-package memory and external DDR for peak performance

**Hybrid Model**
Harness the benefits of both cache and flat models by segmenting the integrated on-package memory

*Maximum performance through higher memory bandwidth and flexibility*
Programming Model Considerations

• Knight’s Landing is a self-hosted part
  — Users can focus on adding parallelism to their applications without concerning themselves with PCI-bus transfers

• MPI + OpenMP preferred programming model
  — Should enable NERSC users to make robust code changes

• MPI-only will work – performance may not be optimal

• On package MCDRAM
  — How to optimally use?
    • Explicitly or implicitly??
Beyond Cori
DOE Fast Forward and Design Forward Programs
http://www.exascaleinitiative.org/

- **Objective:** *Accelerate movement of innovative ideas from computer architecture research into products through public/private partnerships*
- Evaluate advanced research concepts and develop quantitative evidence of their benefit for DOE applications (using DOE Codesign Center apps)
- Engage DOE applications teams to understand technology trends constraints (how it impacts their code development)
- Understand how to *program* these new features
- Develop strong evidence to promote adoption by product teams

- **Fast Forward Technology Areas**
  - **Processor:** AMD, Intel, NVIDIA
  - **Memory:** IBM, AMD
  - **Storage:** WhamCloud

- **Design Forward**
  - Interconnects Architecture (Review Next week at LBNL / July 22-24)
  - System Architecture (under review)
Abstract Machine Model for Exascale

3D Stacked Memory

Thin Cores / Accelerators

Fat Core

Fat Core

Integrated NIC for Off-Chip Communication

Coherence Domain

DRAM

NVRAM
Summary

• NERSC’s goal is to provide *usable* Exascale computing.

• Cori is first step
  – Exciting technology!
    • Processor and Burst Buffer represent first steps along the path to exascale
  – Extensive vendor and community support

• Disruptive technology changes are coming
  – Understand how they will effect you!
  – Note that Exascale era architectures are starting to crystalize