Cori: A Cray XC Pre-Exascale System for NERSC

Katie Antypas, *KAntypas@lbl.gov*
Nicholas Wright, *NJWright@lbl.gov*
Nicholas P. Cardo, *NPCardo@lbl.gov*
Allison Andrews, *MNAndrews@lbl.gov*
Matthew Cordery, *MJCordery@lbl.gov*
National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
Berkeley, CA 94720 USA

***Abstract*—The next supercomputer for the National Energy Research Scientific Computing Center (NERSC) will be a next-generation Cray XC system. The system, named "Cori" after Nobel Laureate Gerty Cori, will bring together technological advances in processors, memory, and storage to enable the solution of the worlds most challenging scientific problems. This new, next-generation Cray XC supercomputer will utilize Intel's next-generation Intel® Xeon Phi™ processor -- code-named "Knights Landing" -- a self-hosted, manycore processor with on-package high bandwidth memory and delivering more than 3 teraFLOPS of double-precision peak performance per single socket node. Scheduled for delivery in mid-2016, the new system will deliver 10x the sustained computing capability of NERSC's Hopper system, a Cray XE6 supercomputer. With the excitement of bringing new technology to bear on world-class scientific problems also come the many challenges in application development and system management. Strategies to overcome these challenges are key to successful deployment of this system.**

Keywords-NERSC-8; CORI; NERSC; LBNL;

I. INTRODUCTION

A. Mission Need

The DOE Office of Science (SC) is the lead federal agency supporting basic and applied research programs that accomplish DOE missions in efficient energy use, reliable energy sources, improved environmental quality, and fundamental understanding of matter and energy. One of two principal thrusts within SC is the direct support of the development, construction, and operation of unique, open-access High Performance Computing (HPC) scientific user facilities.

The National Energy Research Scientific Computing (NERSC) Center is the primary computing facility for the DOE Office of Science. NERSC supports the entire spectrum of SC research, and its mission is to accelerate the pace of scientific discovery through high performance computing and data analysis. An Office of Science user facility, NERSC annually serves over 5,000 scientists annually throughout the United States and the world, encompassing approximately 700 projects utilizing over 600 discrete applications. These researchers, working remotely from Department of Energy laboratories, other Federal agencies, industry, and universities, use NERSC resources and services to further the mission of SC. Computational science conducted at NERSC covers the entire range of scientific disciplines, but is focused on research that supports DOE's missions and scientific goals. The results of the scientific use of NERSC are documented in over 1,500 peer reviewed scientific papers per year as well as in NERSC annual reports and other materials.

The NERSC-8 project was launched to provide additional HPC resources to the Department of Energy Office of Science research community. The NERSC-8 system was scoped to provide a significant increase in resources, at least 10 times the sustained performance of the Hopper system on a set of representative DOE benchmarks. The Hopper system is a Cray XE6 system deployed at NERSC for the NERSC-6 project in 2010 and accepted in 2011. The NERSC-8 system requirements also stated that the system needed to be a platform that begins to transition the broad user community to more energy-efficient manycore architectures.

B. Collaboration with ACES

NERSC is collaborating with ACES (Alliance for Computing at the Extreme Scale, a partnership between Los Alamos National Laboratory and Sandia National Laboratory) on the procurement of NERSC-8 and ACES next generation system called Trinity. Both organizations planned to procure HPC systems in the 2015/2016 timeframe and there were numerous advantages that motivated the decision to work together on the NERSC-8 and Trinity projects. Principally, the collaboration supports the strategy of each of the program's headquarters, SC and NNSA, who seek to leverage each other's investments, provide a unified message to vendors during this time of rapidly changing technology in the HPC market, and build a broader coalition to overcome the challenges transitioning the user communities to more energy efficient architectures. The Office of Science and the National Nuclear Security Administration have been working together for many years and most recently the two programs are partnering on the Design Forward and Fast Forward projects, with the goal of

developing the critical technologies that will be needed for extreme-scale computing.

NERSC and ACES have been working together since the spring of 2012 creating joint technical requirements, developing a common set of application benchmarks and jointly conducting vendor market surveys. In the fall of 2013, the teams released a joint Request for Proposal (RFP) for two independent systems to be delivered in the 2015/2016 timeframe, one for NERSC-8 and one for Trinity. While the two teams are collaborating, each project has its own mission drivers, project management requirements, and will choose the vendor and architecture which provides the best value to the organization.

C. System Configuration

Cori is the next-generation Cray XC system and will be liquid cooled. With 50 cabinets and over 9,300 compute nodes, Cori will have an overall performance of more than 27 Petaflops.

The compute portion of the system will take advantage of the next-generation Intel Xeon Phi processor, code-named "Knights Landing". With over 60 cores per processor with multiple hardware threads, the latest Intel Xeon Phi processor will be capable of delivering 3 Teraflops of performance. One feature of this processor is that it is not an accelerator, but rather a self hosted processor and will take advantage of the Message Passing Interface (MPI) plus OpenMP programming model to deliver high performance to scientific applications. The Knights Landing processor also features high bandwidth on package memory.

Each compute node will have a single self-hosted Intel Xeon Phi processor with between 64 and 128 Gigabytes of memory. The system will utilize the Aries high-speed interconnect to deliver high performance communications between compute nodes.

Accessibility to the system will be gained through 14 external login nodes.

Data Direct Networks latest storage technology will provide 28 Petabytes of storage for the scratch Lustre File System by Cray. This file system will provide 432 GB/s of performance to the compute nodes.

II. Architecture

Cori will be a 50 cabinet Cray XC system that is designed provide ten times the sustained performance of the Hopper Cray XE6 system on a set of representative science benchmarks. The Cori system will use the Knights Landing (KNL) Intel processor, a self-hosted manycore processor that will support the MPI+OpenMP programming model and will include high bandwidth on-package memory. The NERSC-8 system will utilize the Aries interconnect and will come with a Lustre file system for high-bandwidth access to storage.

A. Knights Landing

Each node of Cori will contain a next-generation Intel Xeon-Phi "Knights Landing" (KNL) processor. This is a self-hosted manycore processor, with greater than 60 cores. Each core will have a 512b vector unit, utilizing the AVX 512 instruction set. Each core will have improved single-thread performance over current generation Xeon-Phi co-processor as well was having higher performance per watt.

Each node will also contain high bandwidth on-package memory of similar capacity or greater than in current Xeon-Phi co-processor local memory (> 8GB). This memory can be configured in as a cache or as an extension of the memory space (a scratch pad). We expect that one early question that our users will explore will be to gain an understanding how their applications will optimally use this resource.

B. Aries interconnect

The system will utilize the Cray Aries interconnect, which has been described extensively previously [Faanes, G., Bataineh, A., Roweth, D., Froese, E., Alverson, B., Johnson, T., Kopnick, J., Higgins M., & Reinhard, J. (2012, November). Cray cascade: a scalable HPC system based on a Dragonfly network. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (p. 103). IEEE Computer Society Press.]. In this case the only significant difference from previously installed XC machines is that the nodes are single socket, and thus there are four single sockets connected to each Aries chip.

III. I/O Subsystem

A. I/O Features

Cori will include several features designed to enable users to store, access, and transfer data at high performance. Cori will have local Lustre "scratch" storage capable of 432GB/s, and with a capacity of 28.5 petabytes. Additionally, there will be DVS gateways to provide access to NERSC shared filesystems (NGF) which hold user home directories, and static scientific data sets("project" storage.) Cori will also support 40 gigabit Ethernet connectivity for access to archival storage, as well as offsite internet resources.

B. Lustre Scratch

Cori's scratch storage will consist of 12 DDN block storage arrays, and 96 Lustre OSS servers providing access to 9600 disk drives. Metadata services will be provided by 4 metadata servers, which along with software support in the form of Lustre DNE will support scaling metadata operation rates beyond the capabilities of a single server which has been a limiting bottleneck in previous systems.

C. Burst Buffer

As systems continue on a path of exponential performance growth, it becomes increasingly difficult (expensive) to build storage systems that can keep up. In order to fill the gap in performance between CPU/memory

and disk, the NERSC-8 contract includes an option for a "Burst Buffer," a layer of Non-Volatile Random-Access Memory (NVRAM) designed to accelerate I/O performance for some programs. NVRAM is the technology typically used in "Flash" memory. The precise architecture of the NERSC-8 Burst Buffer and the way in which users would access it, should the contract option be exercised, is still being worked out.

IV. Programming Environment

Bringing together an effective set of tools for application development is only part of the overall picture for successful deployment. Having a robust programming environment along with a strategy for moving applications forward are key elements for overall success.

A. Programming Model Strategy

The necessary characteristics for broad adoption of a new programming model are:

- Performance: 10x-50x performance improvement
- Portability: Code performs well on multiple platforms
- Durability: Solution must be good for a decade or more
- Availability/Ubiquity: Cannot be a proprietary solution

A near-term strategy was developed in order to move to a new programming model that includes the following guidelines:

- Smooth progression to exascale from a user's point of view
- Support for legacy code, albeit at less than optimal performance
- Support for a variety of programming models
- Support optimized libraries

B. Applications on Cori

Due to architectural similarities with current NERSC platforms, applications will likely run successfully on Cori without any changes. However, in order to gain performance applications need to take advantage of the Knights Landing architecture. In order to achieve performance, applications must:

- Exploit more parallelism
- Express thread-level parallelism
- Exploit data level parallelism
- Manage data placement and movement
- Accommodate less memory per process space
- Exploit high bandwidth on-package memory

V. Integration and System Management

A. Cray XC30 Similarities

The liquid cooled Cray XC has many similarities to the current Cray XC30 liquid cooled system. These similarities include:

- Rack configuration
- Cooling system
- Water temperature and flow rate controls
- Power requirements
- Physical cabinet dimensions
- Cabinet weight
- Cabling requirements

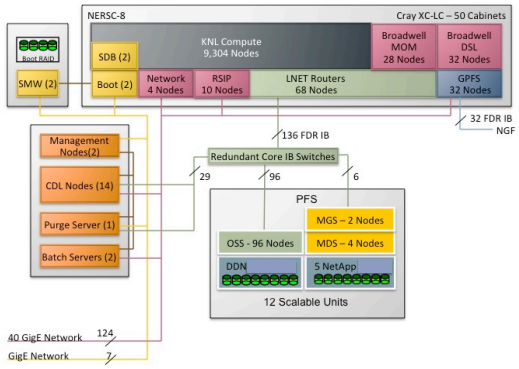
NERSC has gained vast amounts of knowledge and experience with running a Cray XC30, which is the architecture of the currently installed Edison system. This knowledge and expertise was key to understanding the facility impacts of Cori's installation. NERSC was able to leverage this experience to build upon the lessons learned in order to design a facility integration solution.

B. Integrated System

There are four distinct parts to the physical system:

1. System Management
2. Mainframe
3. External Nodes
4. I/O Subsystem

These four parts of the system need to work in concert with each other in order to provide optimal performance to the users and their applications.



A combination of networking technologies will be used to bind these parts of the system together providing high performance integration and to make the system accessible. Data transfer between the storage subsystem and the compute and login nodes will be possible through FDR InfiniBand connections. A combination of 1 GigE and 40 GigE network connections will provide the integrated communications as well as performance connectivity for accessing the system.

C. Seismic Isolation

Cori will be installed in NERSC's new computer facility located in the Computational Research and Theory Facility (CRT) at Lawrence Berkeley National Laboratory (LBNL).

In its current facility, the Oakland Scientific Facility (OSF), NERSC has been using seismic isolation platforms between the computer room raised floor and the computer cabinets in order to minimize the effects of seismic events. There are a number of challenges with this solution including the need for custom platforms for custom cabinets as well as placement on the raised floor. The CRT computer room will be outfitted with a seismic isolation floor, eliminating the need for custom platforms under each cabinet.

D. Floor Loading

As cabinet sizes grow and the density of the components in the cabinets increases, so do the effects on floor loading. The Cray XC cabinets are expected to weigh over 3,400 lbs yielding a floor loading of over 220 lbs/sf. The seismic isolation floor will have a load rating of 500 lbs/sf, easily accommodating these large heavy cabinets.

E. Reliability, Availability, Serviceability

A number of key features of the Cray XC and Cray Linux Environment (CLE) will be utilized to provide improvements in reliability, availability, and serviceability. Where possible, redundancy has been introduced to mitigate the risk of single points of failure. High availability configurations will be used for the System Management Workstations (SMWs), boot nodes, System DataBase nodes (SDB), Cray Image Management servers (CIMS) and batch system. High availability configurations require redundant hardware as well as software to detect failures within the high availability configuration. The Lustre scratch file system will also be configured for high availability. Using dual-pathed servers and dual InfiBand switches mitigates loss of access to all parts of the file system due to a single point of failure.

NERSC runs production level computing resources for the United States Department of Energy's Office of Science. As such, there is an expectation of high availability of the computational resources to the scientific community. As previously mentioned, Cori will be configured for high availability where possible. Minimizing the need to remove the system from service will also play an important role in maximizing the ability to produce scientific results. By utilizing the warmboot capability of the system to perform node level service, the length of time required to take the system out of service will be reduced.

VI. Application Readiness

NERSC has a broad and diverse workload so the success of NERSC-8 depends critically on being able to have a significant fraction of the workload perform well on new many-core technologies. However, these new technologies will make it challenging to meet this goal. Achieving maximum performance on these new chips will require that applications expose more parallelism and use it efficiently. Additionally, longer vector units and high bandwidth on-package memories may require a reformulation of application data structures. Indeed, some applications may benefit from substantial changes to their core algorithms. Unlike previous systems, then, NERSC users will likely require substantial assistance in porting and optimizing their codes. To this end, we developed an Application Readiness Plan, which is a combination of resources from NERSC, individual application code teams, the prime vendor (Cray), and their CPU OEM (Intel).

The success of the application readiness effort will depend on all stakeholders being actively involved. This active involvement requires a plan that deploys resources and initiates action to their best effect. To achieve this, NERSC's application readiness plan is a combination of broad education and targeted engagement. Below, we describe how these two elements will be deployed. Overall, it is NERSC's philosophy that its application readiness effort be as technologically agnostic as possible, with all stakeholders targeting optimizations that can be useful to many-core architectures in general and both minimizing and compartmentalizing any technology-specific optimizations.

A. Engagement:

Given the breadth of NERSC's workload, it is not feasible for NERSC staff (nor the vendors) to be directly involved in porting and optimizing every code in the workload to the NERSC-8 architecture. To address this limitation, NERSC staff have identified of order 20 codes that comprise approximately two-thirds of NERSC's current workload. These codes will receive direct assistance from NERSC in their porting and optimization efforts, including some fraction of the vendor resources and will be assigned to one of three code teams:

1. *Target Application:* These teams are generally comprised of code developers that have a strong LBNL presence but would significantly benefit from expert optimization advice.

2. *Advanced User:* These teams are generally comprised of code developers that have significant experience with optimizing their codes for various architectures but would still benefit from expert advice from Cray and Intel on directly targeting many-core techniques in general and their Xeon Phi processor specifically.

3. *Third-Party Developers:* These teams target community-based applications and libraries used and possibly developed by LBNL researchers but, generally, with more a more geographically dispersed developer base.

Rather than begin all code team activity after contract signing, NERSC will delay the deployment of most code teams until approximately the last half of 2015. This timing ensures that there is not a large lag time between code team deployment and the availability of many-core hardware and should ensure that teams maintain their motivation until the arrival of the main NERSC-8 system.

Before the deployment of the majority of the code teams, however, NERSC will initiate two prototype code teams in order to determine the best practices in coordinating the resources between the various stakeholders. The initiation of these two code teams will begin in the second half of 2014.

After the prototype teams have completed their initial work, in the second half of 2015, we will commence the bulk of the application readiness effort by instantiating the remaining code teams. At present, NERSC envisions code team developers as the primary drivers of code optimizations, with Cray providing expertise not only on optimizing for the many-core nodes, but also for the interconnect and parallel file system and Intel providing expert advice on compiler usage, deep information on exploiting the Knights Landing processor and memory system, and advice on many core programming. NERSC Application Readiness Team members will be responsible for coordinating the teams with the vendor resources and ensuring that timely progress is made and deliverables are met.

For the third party codes in particular, NERSC plans on hosting several 'developer conferences' to bring together active developers of the codes to focus and direct community optimization efforts for many-core as well as to provide an opportunity for the community to benefit from 'lessons learned' by NERSC groups and the wider DOE HPC community. The plan is to use these workshops to identify issues and opportunities and to try to drive the adoption of specific software engineering action items by the community.

B. Education/Training

Because NERSC is only able to directly target a limited set of the entire NERSC workload, providing training to all NERSC users is of paramount importance for the success of NERSC-8. To this end, NERSC will provide a series of quarterly training sessions, in addition to the standard yearly course on parallel computing, addressing various issues of importance to the success of optimizing for many-core architectures. These courses will begin in the second half of 2014 and will begin targeting the performance of applications on NERSC's Cray XC30 (Edison) as a proxy for Knights Landing, thus giving users ample to prepare without having to wait for NERSC-8 hardware to arrive. A series of NERSC provided training courses is envisioned as follows:

Phase 1: Shared memory programming with OpenMP
Phase 2: Vectorization and Memory Locality
Phase 3: Intel compiler directives/intrinsics for Xeon Phi and use of on-package memory

Despite having concerns about its own workload, NERSC is keenly aware that the issues that it faces are also shared by its colleagues other DOE laboratories and the broader HPC community. To this end, NERSC intends to proactively share its experiences. Our current plans are as follows:

a) Sharing of training materials.
b) Inviting representatives from other laboratories to collaborative efforts and developer meetings.
c) Attend collaborative efforts at other DOE laboratories.
d) Hosting multi-laboratory conferences where technical staff can share their experiences/results with many-core programming and discuss new tools and algorithms and hacks.
e) Engaging with current code optimization efforts by Cray, Intel, and/or other DOE labs where these efforts overlap with the NERSC workload.
f) Presenting at conferences and publishing white papers on our experiences porting and optimizing for many core architectures in general and the Knights Landing processor specifically.

ACKNOWLEDGMENT

This work was supported by the Director, Office of Science, Office of Advance Scientific Computing Research of the U.S. Department of Energy under contract No. DEAC02-05CH11231.