Cori - A System to Support Data-Intensive Computing

Tina Declerck, Katie Antypas, Deborah Bard, Wahid Bhimji, Shane Canon, Shreyas Cholia, Helen (Yun) He, Douglas Jacobsen, Prabhat, Nicholas J. Wright
NERSC
Lawrence Berkeley National Laboratory
Berkeley, CA USA
tmdeclerck@lbl.gov, kantypas@lbl.gov, djbard@lbl.gov, wbhimji@lbl.gov, scanon@lbl.gov, scholia@lbl.gov, yhe@lbl.gov, dmjacobsen@lbl.gov, prabhat@lbl.gov, njwright@lbl.gov

***Abstract*— The first phase of Cori, NERSC's next generation supercomputer, a Cray XC40, has been configured to specifically support data intensive computing. With increasing dataset sizes coming from experimental and observational facilities, including telescopes, sensors, detectors, microscopes, sequencers, and, supercomputers, scientific users from the Department of Energy, Office of Science are increasingly relying on NERSC for extreme scale data analytics. This paper will discuss the Cori Phase 1 architecture, and installation into the new and energy efficient CRT facility, and explains how the system will be combined with the larger Cori Phase 2 system based on the Intel Knights Landing processor. In addition, the paper will describe the unique features and configuration of the Cori system that allow it to support data-intensive science.**

Keywords-Cori;Data-intensive;RSIP;SDN;KNL;

I. INTRODUCTION

The first phase of Cori, NERSC's next generation supercomputer, a Cray XC40, has been configured to specifically support data intensive computing. With increasing dataset sizes coming from experimental and observational facilities, including telescopes, sensors, detectors, microscopes, sequencers, and, supercomputers, scientific users from the Department of Energy, Office of Science are increasingly relying on NERSC for extreme scale data analytics. To support these requirements, NERSC is also working to improve both hardware and software system capabilities to support this new and growing workload. This paper will discuss the Cori Phase 1 architecture, and describe the unique features and configuration of the Cori system that allow it to support data-intensive science. In addition, the paper will explain how the system will be combined with the larger Cori Phase 2 system based on the Intel Knights Landing processor and support both data intensive and HPC workloads.

What does data intensive computing mean to NERSC? To understand some of the choices NERSC has made in configuring the system, it's important to understand how NERSC defines data intensive computing. In the broadest sense, we take the term to mean applications or workflows that require a significant amount of data movement. Some applications will need to read or write large datasets from disk, and perform a significant amount of I/O. Other projects many needs to transfer large amounts of data from an experimental facility, share large amounts of data through gateways, or need to store or search through large amounts of data in databases. Still others may needs to analyze a large amount of data, but require only a modest amount of computing. Our goal is to be able to work support the broad NERSC user requirements for data-intensive computing within a single system.

II. CORI PHASE I CONFIGURATION

Cori Phase 1, a Cray XC40, has approximately 1600 dual-socket Haswell compute nodes. Each compute node contains two 16-core Haswell processors per node running at 2.3 GHz with 128GB of DDR4 memory running at 2133 MHz, providing data intensive application with a large amount of memory per node. Cori Phase 1 uses the Cray Aries interconnect with the dragonfly topology. In addition, Cori Phase 1 contains 144 DataWarp nodes, configured to form an NVRAM 'burst buffer' provide accelerated read and write performance. The system also has a 30 PB Lustre file system based on the Cray Sonexion 2000 with 124 scalable storage units (SSUs) and 2 additional DNE units (ADUs) provide both fast data paths and large storage capacity, and the 4 additional metadata servers to help spread the load of the MDS increasing metadata performance.

III. II I/O IMPROVEMENT CRAY DATAWARP

One of the top improvements NERSC users consistently request in requirement reviews and feedback is more storage and better I/O performance. On the Cori Phase 1 system, NERSC is providing new capabilities to specifically address some of the I/O concerns of users. First, it includes a Burst Buffer, based on the Cray DataWarp technology. To meet the needs of NERSC users, in particular those with data-intensive applications, NERSC has invested in the Cray DataWarp, an intermediate layer of non-volatile storage that sits between the fast on-node DRAM and the slower (but higher capacity) parallel file system (PFS). This "Burst

Buffer" provides users with a configurable layer of fast I/O that can improve application IO in several ways:

- Improved IO bandwidth for reads/writes, for example, for checkpoint-restart applications.
- Improved performance for complex IO patterns, for example high IOPS (IO operations per second).
- Improved capability for complex workflows, for example combining simulation, analysis and visualization codes.

A. Hardware

The DataWarp SSDs sit on specialized nodes that bridge the internal Aries interconnect of the compute system and the SAN fabric of the PFS, through the I/O nodes.

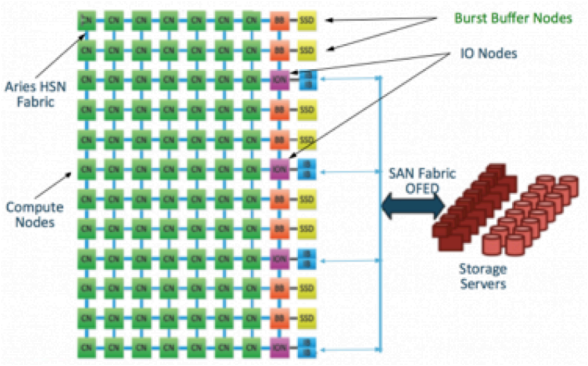

Figure 2. Cori DataWarp Architecture

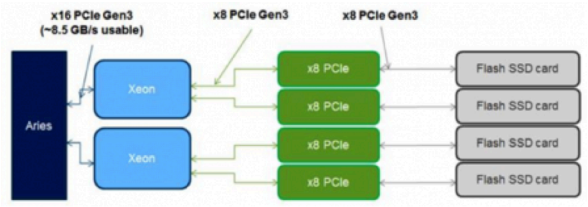

Figure 3. DataWarp Node Architecture

The flash memory is attached to Burst Buffer nodes that are packaged two nodes to a blade. Each Burst Buffer node contains an Intel Xeon processor with 64 GB of DDR3 memory, and two 3.2 TB NAND flash SSD modules attached over two PCIe gen3 x8 interfaces. Each Burst Buffer node is attached to a Cray Aries network interconnect over a PCIe gen3 x16 interface. Each Burst Buffer node provides approximately 6.4 TB of usable capacity and a peak of approximately 5.7 GB/sec of sequential read and write bandwidth.

B. Software

NERSC has engaged with Cray in Non-Recurring Engineering (NRE) of the DataWarp software. This consists of several layers of software:

- A Logical Volume Manager (LVM) groups the 4 SSDs on a device into one logical block device
- XFS is used to create a designated Burst Buffer allocation
- The DataWarp File System (DWFS) coordinates data between the DataWarp nodes
- The Cray Virtualization System (DVS) is used to control IO exchange between the DataWarp nodes and the PFS.

Access to the Burst Buffer resource is integrated with the SLURM scheduler. When a user submits a job requesting a Burst Buffer allocation, an XFS is created to mount that allocation so that the user sees a single namespace, even though data might be striped over several DataWarp nodes.

Future software capability will include transparent caching where a user does not need to explicitly move data between the PFS and the Burst Buffer, and in-transit analysis capabilities, which will use the on-package Xeon chip to process data directly on the SSD.

C. Performance

The Burst Buffer has performed very well on standard benchmarks, like IOR (ref) and above the required contractual performance for almost all benchmarks with the exception of MPI-IO shared file writes. The burst buffer on Cori Phase 1 also outperforms the configured Lustre file system on Cori. It is important to note that the final configuration of the system with Cori Phase 2 installed will include double the amount of DataWarp hardware, but the same Lustre file system.

NERSC has launched a Burst Buffer Early User Program assist real applications in getting performance out of the burst buffer and to assess the real application performance. And while still in the early stages of testing and development, DataWarp is providing improved I/O capabilities to a applications. A more detailed look at the Burst Buffer use cases and early performance results is

TABLE I. CORI DATAWARP PERFORMANCE

	All Burst Buffer Nodes : 1120 Compute Nodes ; 4 MPI Ranks per node					
	Posix FPP		MPIO Shared File		IOPS	
	Read	Write	Read	Write	Read	Write
Required (GB/s) or IOPS	820	820	656	656	7200000	7200000
Best Measured	905	873	803	351	12591978	12527427

included in the CUG 2016 paper, *Bhimji, W., Accelerating Science with the NERSC Burst Buffer Early User Program (ref)*

IV. III I/O Improvement Parallel File System

In addition to the DataWarp burst buffer, the Cori Phase 1 system offers users a Lustre file system of increased bandwidth and capacity, providing over 700GB/sec of peak bandwidth and 30PB of disk capacity. NERSC is in the process of testing some new system capabilities that will allow better support of the file system. To ensure file system updates are not affecting performance, the Sonexion 2000 file system has been configured with two file systems: a small test file system that can be used for performance and regression testing and a large production file system for user applications. This configuration is not typical and NERSC worked closely with Cray to enable this capability. Although the two file systems can not be mounted at the same time, having the option to do performance testing on a clean, small file system allows NERSC to better assess the performance and impact of an upgrade. The primary issue with measuring file system performance on the large production file system is that performance changes depending on how full the Lustre file system is and where the data is being written to disk since some sectors of the disk provide faster access than other sectors, making performance variable and making it difficult to get a performance baseline. The small test file system is comprised of a small portion of every SSU on the Sonexion that allows a full test of all parts of the file system, but with a partition that can be cleared of data, which means the performance should be reasonably consistent. With this configuration, a problem on a specific SSU or an upgrade that has an impact on performance should be identified on the test file system. This provides a way to identify issues after an upgrade or modification more quickly and easily.

A. Peak Cori scratch Lustre IO performance

The high performance Cori Lustre file system has 30 PB disk space and >700 GB/sec IO bandwidth. The plot below shows the required contractual rates and best-measured Read and Write I/O rates. Left 2 columns are Posix: File Per Process, and Right 2 columns are MPI-IO: Single Shared File.

V. Software Defined Networking

Detectors and imaging facilities coming online in the next 3-5 years are expected to produce data at a rate of approximately 1Tb/second. These scientific data-intensive workloads require systems that have the ability to ingest and process data from scientific instruments and sensor networks. For example, light sources like the LCLS and next-generation electron microscopes are expected to generate data at over 1 Tb/s within the next five years. These users need to have the flexibility to stream data to Cori compute nodes so it can be either processed directly in memory or be staged to the burst buffer for processing after that data is collected. Benchmarks have illustrated that the current realm-specific IP (RSIP) solution is not capable of addressing many of these new use cases of users who need to

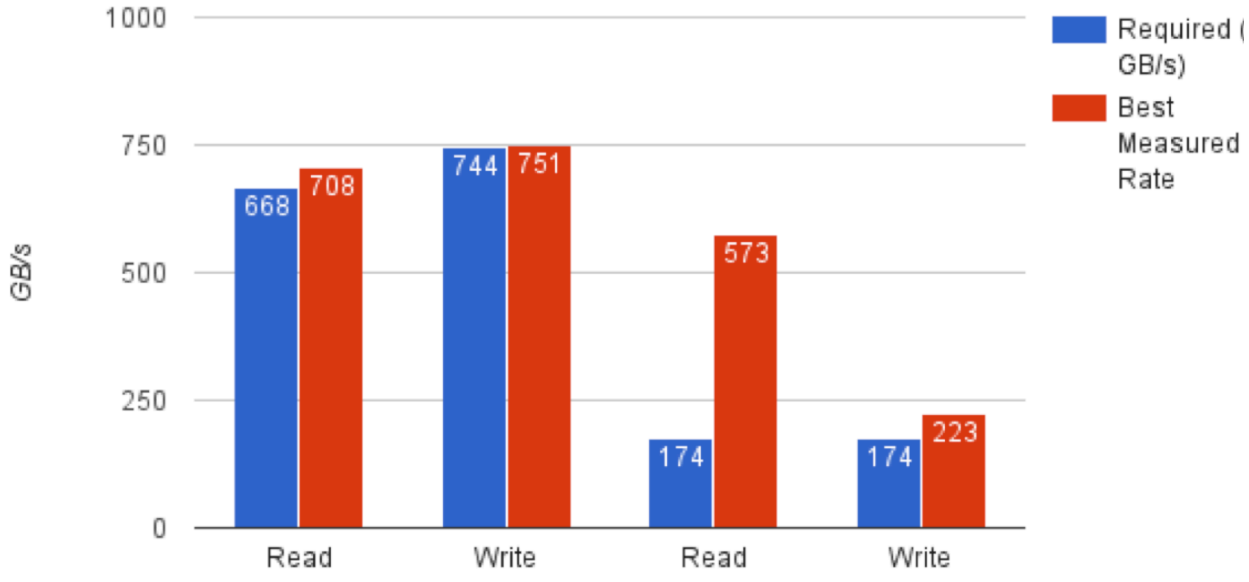

Figure 4. Cori Scratch File System Performance

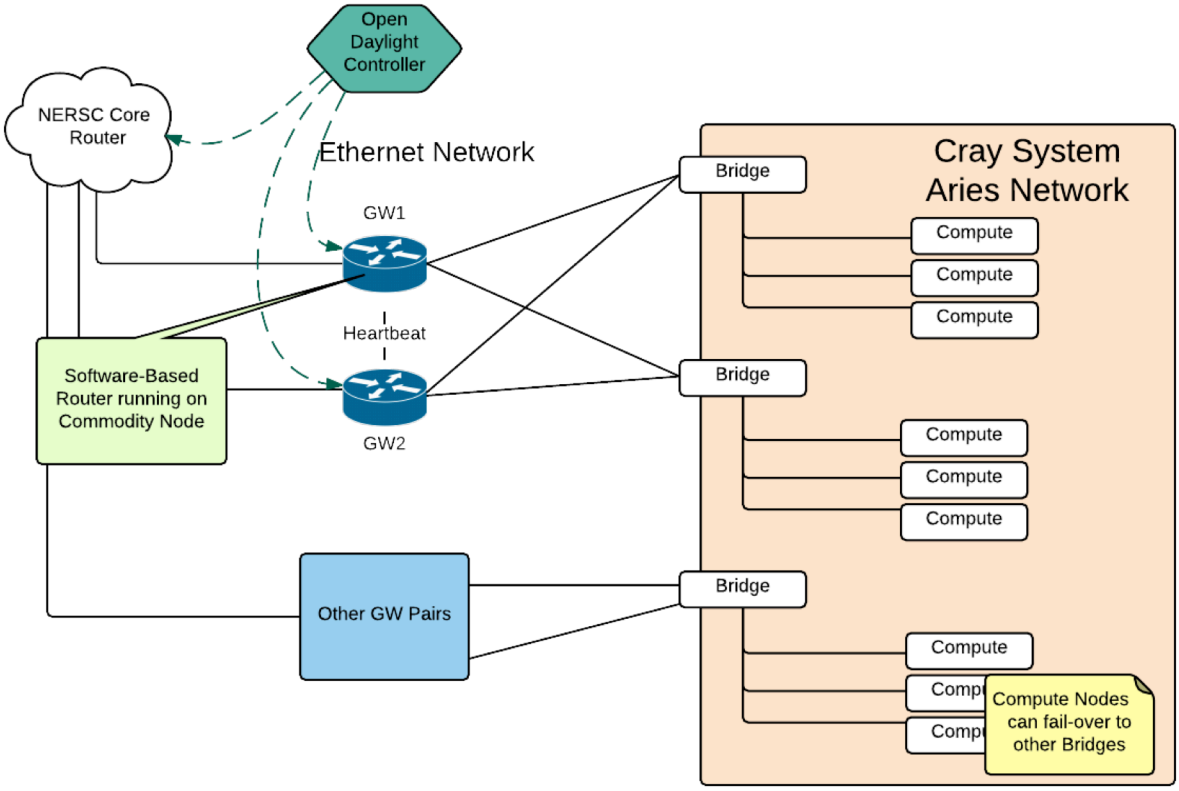

Figure 5. Software Defined Network Configuration on Cori

move data in and out of the system at high data rates. One primary drawback is the RSIP only allows for connections that are initiated from compute nodes in the Cray system. One primary drawback is that RSIP only allows for connections that are initiated from compute nodes in the Cray system. Use cases for connections initiated outside the system such as streaming data directly from experimental facilities for initial processing have lead to testing technologies such as software defined networking (SDN). In partnership with Cray, NERSC is exploring new techniques to efficiently move data into and out of the Cori. In addition, we are looking at ways to integrate emerging Software Defined Networking capabilities. The ultimate goal is to allow a beamline scientist at the LCLS and other experimental facilities to co-schedule networking bandwidth, Cori compute resources and burst buffer bandwidth so that they could effectively act as an extension of the beamline instruments. To realize this vision NERSC needs to be able to provide dynamic scheduling and provisioning of networking end-points that map to compute resources in Cori. This problem will be addressed in multiple phases with the initial phases focused on basic capabilities to do high-bandwidth Network Address Translation (NAT) from the private address space used inside Cori to routable public addresses. The existing RSIP nodes will be re-provisioned to act as IP bridges between the Aries network and a local Ethernet network. Software-based routers running Brocades vRouter will be connected to this Ethernet network and will be responsible for routing IP traffic to external networks. These routers are also associated with a SDN controller based on the OpenDaylight Controller. Eventually this controller will be integrated with the SLURM scheduler to handle dynamic provisioning.

VI. WORKFLOWS

NERSC's mission has been expanding into closer interactions with experimental and observational facilities, whose users often have different requirements than traditional HPC modeling and simulation users. As a result, how we configure and manage our systems must be adapted to support these new use cases. Often users analyzing large data sets from an experimental facility have more complex workflows. A workflow might include filtering data, moving data, running multiple pipelined analysis codes on the data, in addition to post-processing. To support more complex workflows, the Cori Phase 1 system includes a much larger number of login nodes than NERSC has deployed in the past, some of which are configured as workflow nodes. These workflow nodes can be reserved for specific users or experiments if necessary. The Cori system has been designed with data-intensive

workflows in mind. We define a workflow as a set of coordinated tasks that need to be run in an automated fashion. Workflow systems can facilitate the execution of various workflow elements. Additionally, they may also perform data movement in and out of the system, along with managing intermediate data products across different stages of the workflow. Another aspect common to many workflow systems is the ability to support very large numbers of independent but coordinated tasks.

Workflow systems have been developed to address evolving needs of scientific computing as data-intensive jobs become increasingly important. We consider 3 key features on Cori that can augment both scientific workflows:

A. Network Connectivity to External Nodes

Specifically many workflow systems are managed by a control node or database that manages the different tasks and stages in a workflow. The compute nodes on Cori must be able to talk to a persistent service that lives outside the Cori network. This is also needed to facilitate movement of data in and out of the compute nodes, in the case where individual workflow tasks need to pull down units of work or publish results to/from an data source service.

B. NVRAM filesystem for in-situ workflows

Multi-Stage Workflows will generate data between each step in the workflow. For example, in a cosmological image analysis pipeline, the first stage of the workflow may involve some kind of noise filtering algorithm, which must be fed into the next step that involves performing a series of subtractions on the images. As the size of the data grows it, it becomes increasingly impractical to move the data in and out of the file system from a performance standpoint. The NVRAM layer provides a very convenient intermediate staging area for this data. Applications can write out intermediate data into a fast NVRAM layer, which is then fed to the next step in the workflow, without a performance bottleneck

C. Real-time Queues for time sensitive analyses

Cori supports a real-time queue for time sensitive analyses. Users can request a small number of on-demand nodes if their jobs have special needs that cannot be accommodated through the regular batch system. The real-time queue enables immediate access to a set of nodes, for jobs that are under the real-time wallclock limit (currently 12 hours). Typically this is used for real time processing linked to some experiment or event. Use cases for the real-time queue include:

1) ALS

Real-time Analysis of light source data from the Advanced Light Source must occur in conjunction with the data-taking phase at the beamline. Since scientists have time reserved on the beamline, it is highly desirable to enable prompt analysis of data being taken on the beamline, to be able to provide prompt user feedback on the results. The ALS uses the SPOT Portal to transfer data from the ALS to NERSC, analyze it over the real-time queue, and provide interactive visualizations over the web based on results streaming from the real-time partition.

2) PTF

This experiment runs a real-time survey at Palomar Observatory in which the data is processed minutes after the shutter closes on NERSC resources with follow-up directed by the discoveries made at NERSC.

3) DIII-D

The DIII-D National Fusion Facility is a DOE/SC User Facility. When operating, fusion plasmas of approximately 10 second duration are made every 20 minutes or so. Those 20 minutes are used for intense physics analysis. There is a long chain of analysis codes that are run locally, but as codes expand, the ability to run a few larger codes at a site like NERSC to support "between shot data analysis" can be very attractive. For this to be valuable, the wall clock turn around is very important (minutes). This code will read data from DIII-D's MDSplus repository and then write it results back to the same repository.

D. Scheduler and Queue improvements to support data-intensive Computing

In addition, the scheduler and scheduling policies we had traditionally used to support our modeling and simulation workloads proved inflexible for dealing with the more dynamic needs of data-intensive users. One of the key changes we made to support this new workload was a change to Native SLURM. The scheduler's closer ties to the compute nodes provide easier diagnostics, cleaner access to the data, and faster startup. Some of the capabilities NERSC is utilizing allow for users to access a part of the system on a real-time basis. This allows for immediate access to compute cycles for a class of users that have shown a need for this type of use such as data streaming from experimental sites for initial processing. There are also some downsides to using native SLURM vs. ALPS (Application Level Placement Scheduler). ALPS is intended to provide an interface for external Workload Managers (WLM) to place, launch, and manage jobs while they are running on Cray compute nodes. Some of the capabilities provided by ALPS don't provide the ability to support native schedulers. However, some of the new features introduced by Cray, specifically DataWarp, provide capabilities that can be used by all applications both data-intensive and simulation and modeling and have been integrated into the SLURM scheduler. Users from experimental facilities also often come with expectations of a specific operating system or require a complex set of installed libraries. For these users, NERSC is allowing users to bring their own images to Cori, by way of a new capability called Shifter, which is based on Docker containers.

Our implementation of the real-time queue is drawing a number of capabilities of the SLURM scheduler, particular possible due its integration as the resource manager on the system ("native" SLURM). Users granted access to real-time are provided a per-project Quality-of-Service (QoS) access to the schedule that inserts jobs into

the batch queue at higher priority than is achievable by all non-real-time jobs, but also restricts how many resources that particular project can use concurrently (called "Grp" limits). Once a project exceeds its allocated resource limits, those jobs queue up and are no longer real-time. Furthermore, users are encouraged to "share" nodes either with their own jobs or with other real-time users (at the user's option), however the user can also gain exclusive node access. This capability is drawing on SLURM's use of the "native" WLM interface provide by Cray, wherein we can run up to 32 individual jobs per Cray compute node if that is desirable for the workload at hand. Therefore, since real-time users can very finely request exactly the resources needed, sharing resources when possible, we can deliver more real-time access with fewer dedicated resources. To ensure that real-time jobs get immediate service, real-time jobs are granted access to the entire system, with a small number of nodes dedicated to just real-time work - 8 are used in the current configuration. Another SLURM feature allows us to weight those 8 dedicated nodes so that real-time work will always attempt to schedule on those before using other resources.

VII. Summary of System Features for Data Intensive Science

A summary of the features NERSC considers as part of it's data solution include:

- Burst Buffer for high bandwidth, low latency I/O
- Large amount of memory per node (128 GB/node) as well as high-mem nodes (775GB/node) accessed via a separate high-memory queue.
- Large number of login/interactive nodes to support applications with advanced workflows
    - Used for spark, ipython, fireworks and experiment specific workflows? (e.g. ATLAS LHC experiment)
- Immediate access (realtime) queues for jobs requiring real-time data ingestion or analysis
- High throughput and serial (shared) queues can handle a large number of jobs
- Improved outbound Internet connections to communicate with the outside world. (e.g. to access a database in another center.) via RSIP
- R&D for high bandwidth transfers in and out of the compute node with Software Defined Networking
- User-defined images/Shifter
- High-performance Lustre Filesystem

VIII. Early User Experiences

A. System Usage Info

Early users were enabled on the Cori Phase 1 system in multiple phases from October through November 2015 and all users were enabled by November 12th, 2015.

Figure 6 below shows the Cori usage information from October 2015 to April 2016. During the early user period, users are able to run free of charge, until the new allocation year began on January 12, 2016.

The top panel shows the jobs size breakdown and the second panel shows the system utilization information. 162 Million NERSC hours were used from Oct 29, 2015 to Jan 11, 2016. The plot below shows the usage from different science areas before and after the beginning of the allocation year. The Cori Phase 1 system is particularly popular with the Materials Sciences community.

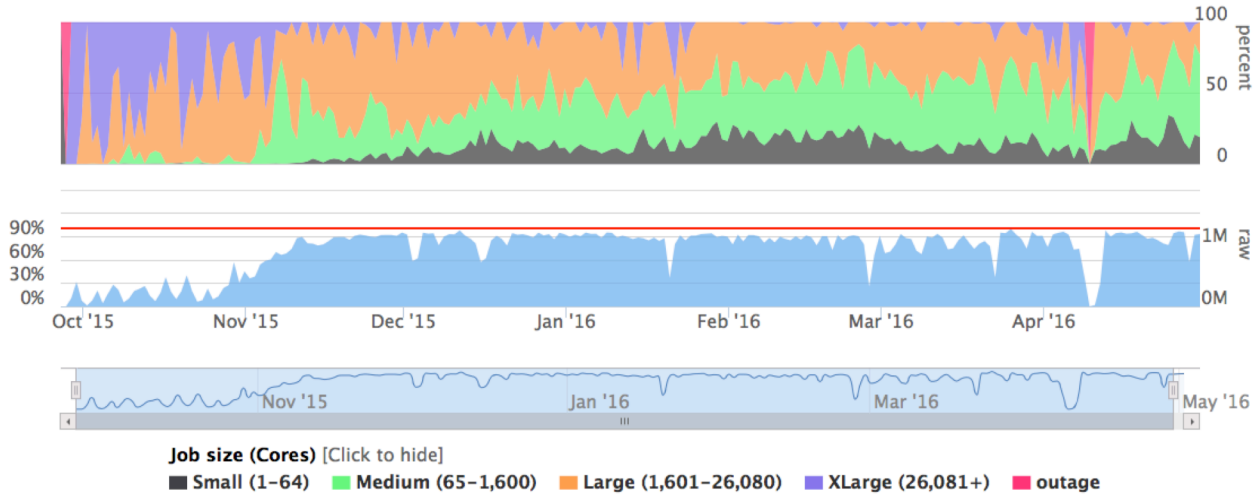

Figure 6. Job Size Breakdown and System Utilization

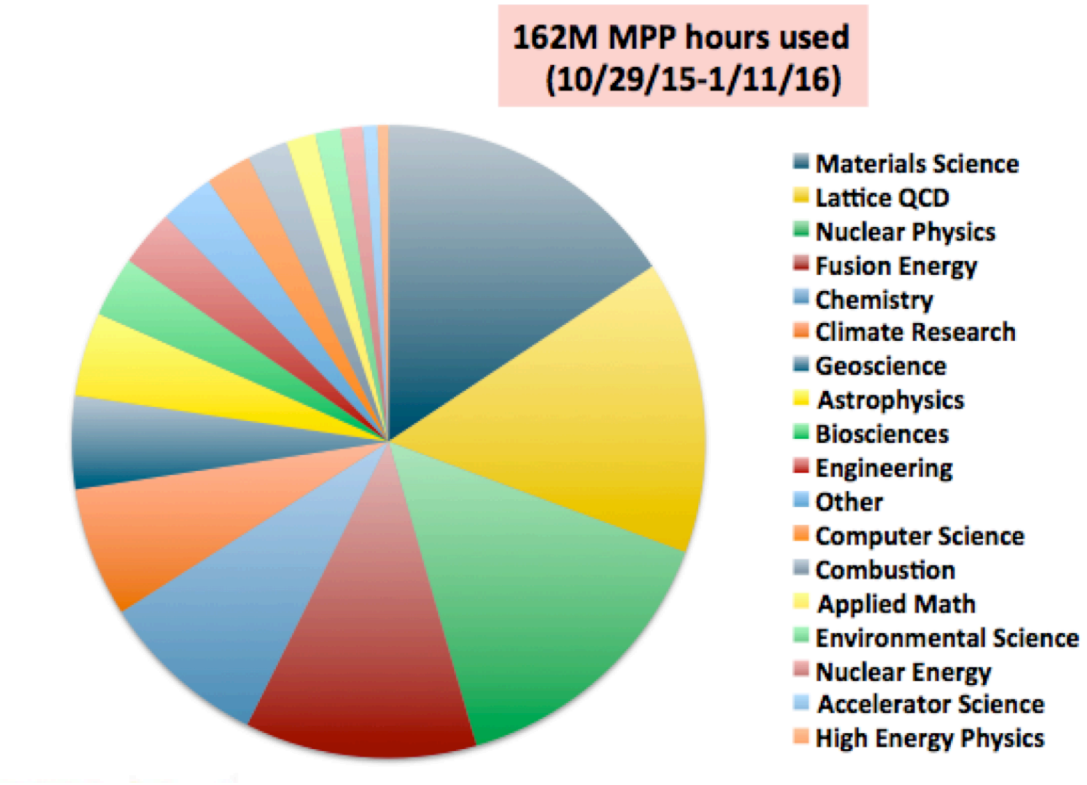

Figure 7. Cori Usage Before Allocation Year End

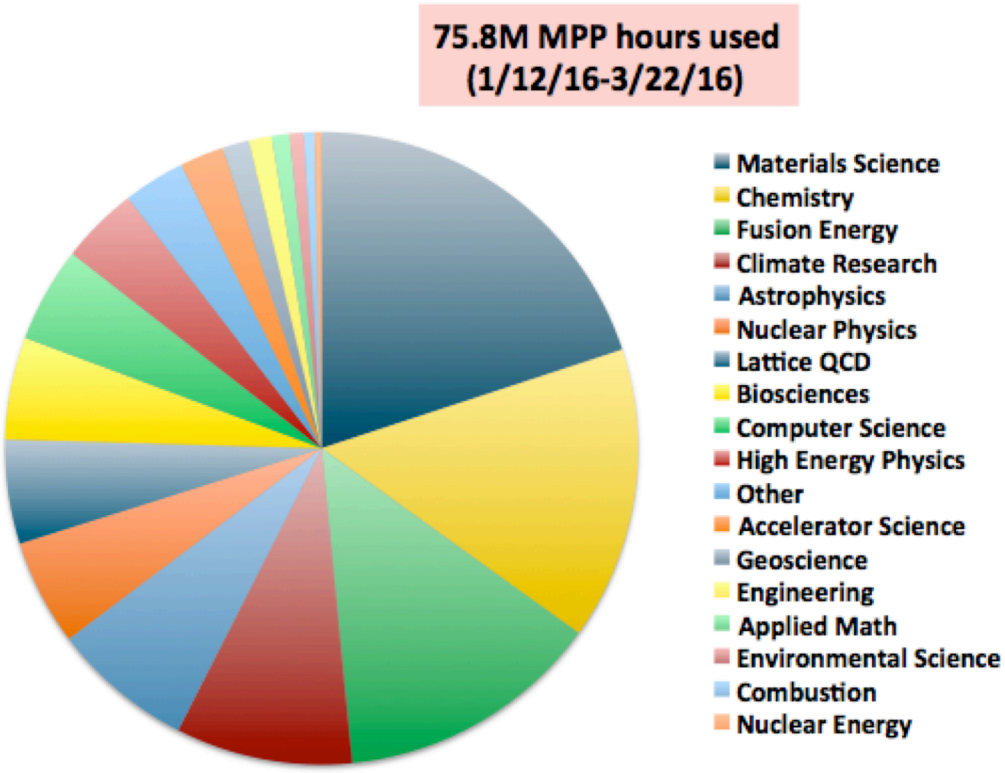

Figure 8. Cori Usage Beginning of the New Allocation Year

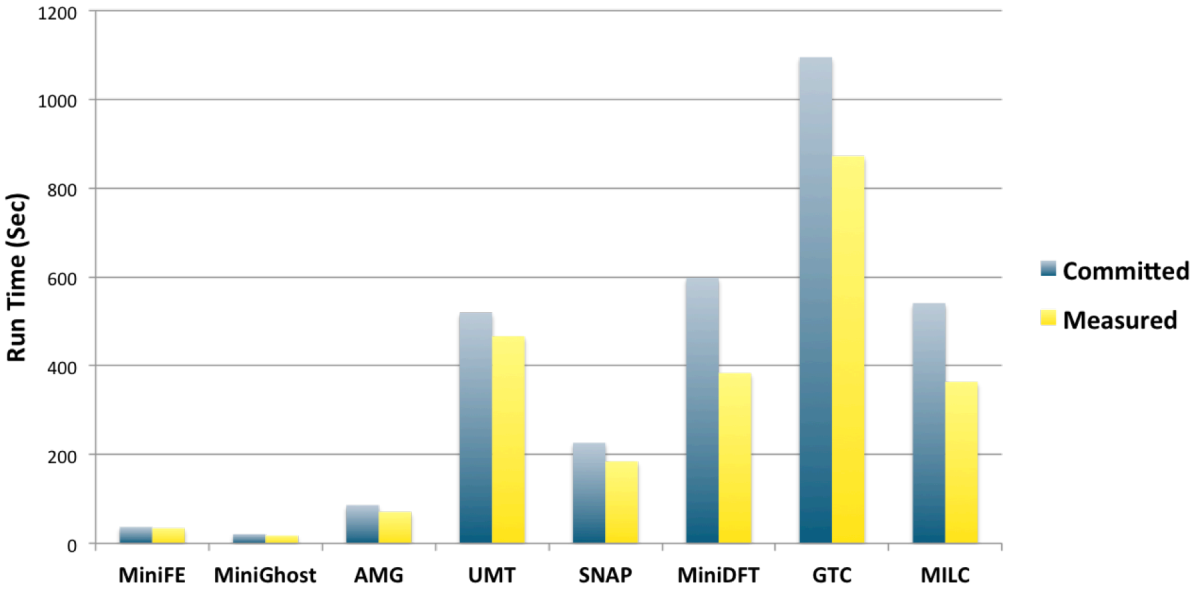

Figure 1. Cori SSP Performance

B. *Programming and Running Jobs*

The programming environment on Cori Phase 1 is very similar to NERSC's previous Cray systems Hopper/Edison and thus porting applications to Cori Phase 1 was straightforward for most.

The aspect that users need to adjust the most is the transition of the batch scheduler from Torque/Moab to SLURM. In order to help users, we provided detailed documentations on SLURM transition guide, example batch scripts, and tutorials. Also we worked with some specific applications and users for the porting. CESM is one such example.

In general users are happy with the Cori Phase 1 system and are using the system productively. The biggest complaint from users is the long wait times which is a testament to the popularity of the system.

C. *Benchmarking Performance*

The Sustained System Performance (SSP) (ref) is an aggregate, unweighted measure of computational capability relevant to achievable scientific work. It is used as a way of measuring, reporting, and projecting the performance of a given system using a set of benchmark programs that represent a workload. The plot below shows the Cori Phase 1 SSP performance. The committed SSP value from Cray is 68.2 TF/sec, while we are achieving 83.0 TF/sec.

IX. Conclusion

In summary, in order to provide the best support for both our traditional modeling and simulation, as well as data intensive workloads, we are looking at new technologies, configurations and ways of managing our HPC systems. On Cori Phase 1 we have introduced a number of new capabilities to support data-intensive science.

In the fall of 2016, the larger Phase 2 Cori system, featuring the Intel Knights Landing (KNL) processors will be delivered to NERSC and integrated into the Phase 1 system. Providing a large-scale system for simulations as well as data analysis.

X. References

[1] IOR Benchmark, https://github.com/LLNL/ior, https://asc.llnl.gov/CORAL-benchmarks/Summaries/IOR_Summary_v1.0.pdf

[2] Kramer, William, John Shalf, and Erich Strohmaier. "The NERSC sustained system performance (SSP) metric." *Lawrence Berkeley National Laboratory* (2005).

[3] Bhimji, W., Accelerating Science with the NERSC Burst Buffer Early User Program, Cray Users Group, London, 2016.