HYPERION RESEARCH

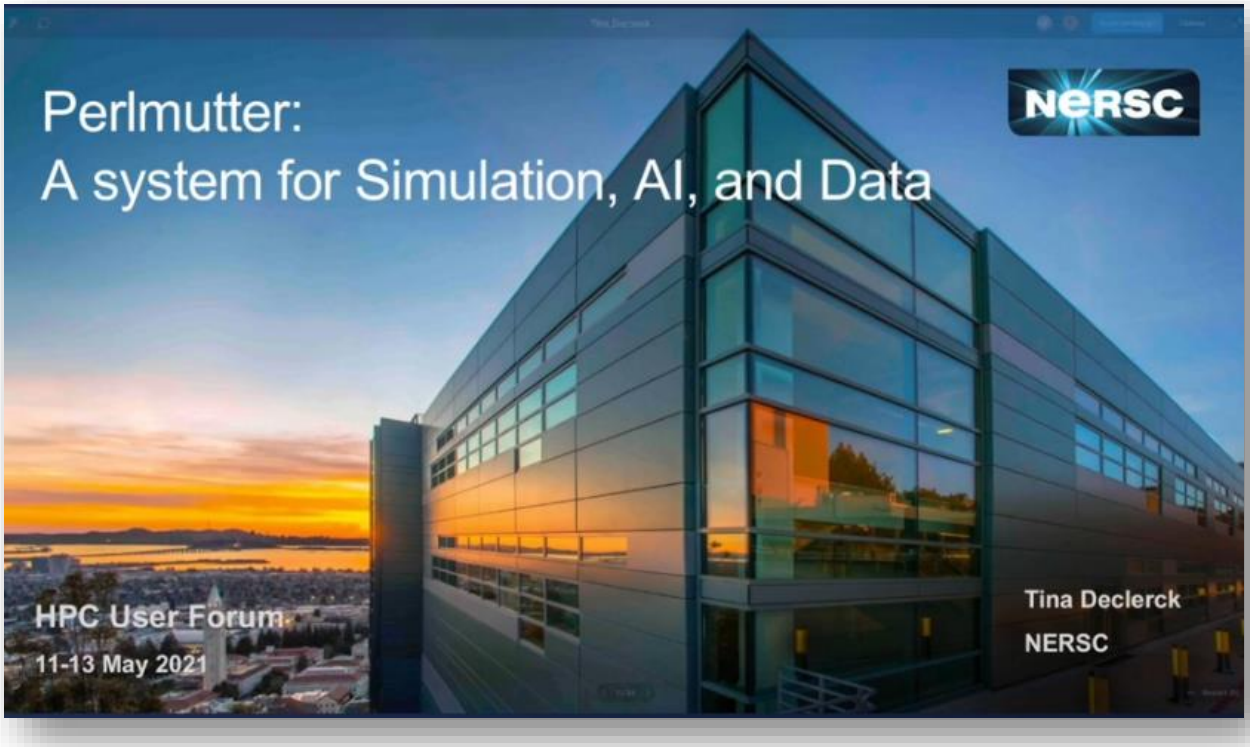
HPC User Forum Update

Perlmutter: A System for Simulation, AI, and Data

Thomas Sorensen and Earl Joseph
July 2021

IN THIS UPDATE

The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In May 2021, the 76th HPC User Forum took place virtually. This update summarizes a presentation from that virtual conference given by Tina Declerck, System Department Head at the National Energy Research Scientific Computing Center (NERSC), titled *Perlmutter: A System for Simulation, AI, and Data*. In addition to updates on NERSC and its role as the Mission HPC facility for the DOE Office of Science, Declerck provided information on the developments of the new Perlmutter system, its capabilities, structure, and future.



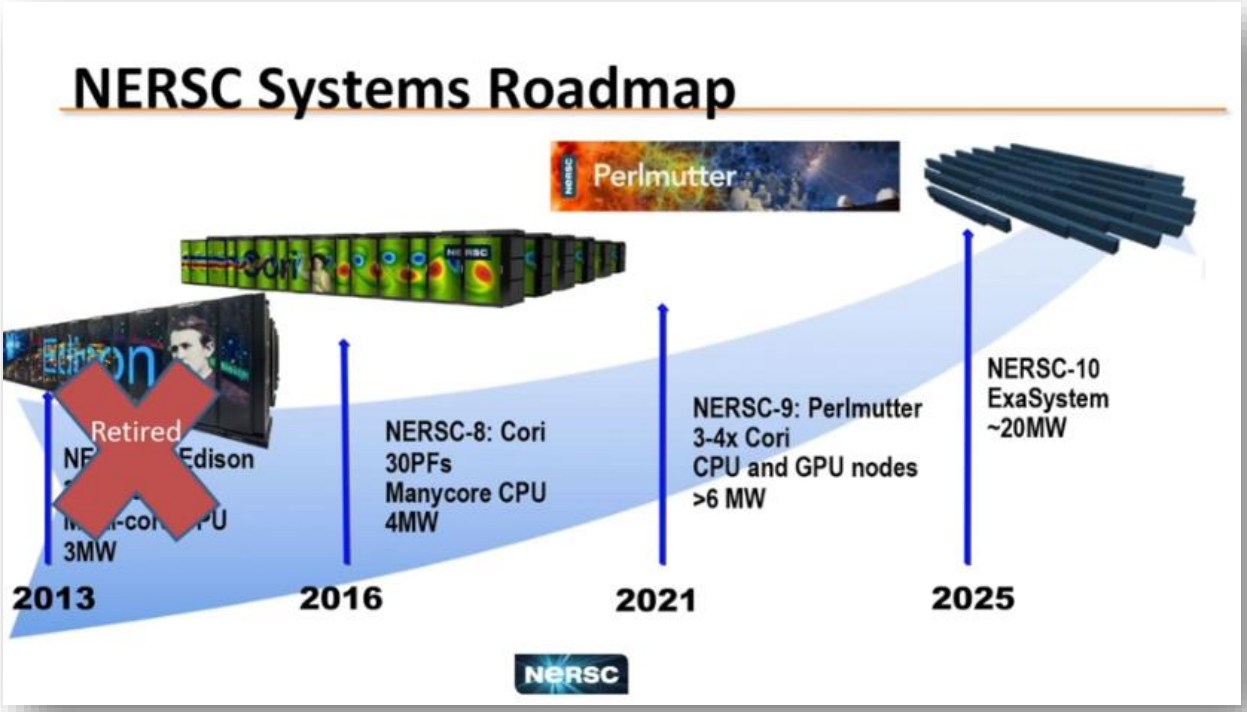
Source: NERSC and Hyperion Research, 2021

PRESENTATION: PERLMUTTER: A SYSTEM FOR SIMULATION, AI, AND DATA GIVEN BY TINA DECLERCK, SYSTEM DEPARTMENT HEAD AT NERSC

Declerk, who leads the System Department that is composed of the Computational Systems, Security and Networking, Operations Technology, and Building Infrastructure groups, began by highlighting the reach and volume of scientific research conducted at the site: over 7,000 users with 800 projects, 700 codes, in all 50 states and 40 other countries. Additionally, the science performed on NESRC systems is responsible for roughly 2,000 publications per year.

NERSC has a dual mission: to advance science and maintain the state-of-the-art in supercomputing. To meet this high standard, they collaborate with vendors years in advance before a system's delivery, deliberating on hardware components that are yet to be in production and that are often among the first sites to get them. The site is tightly coupled with DOE workflows and those of their experimental and observational facilities. 90% (nearly 8 billion core hours) of their cycles are assigned to the DOE for distribution among their different experiments. The other 10% is split evenly among the ASCR Leadership Computing Challenge (ALCC) and the Directors Discretionary Reserve, which can be used for various purposes, most recently in aiding Coronavirus research.

FIGURE 1

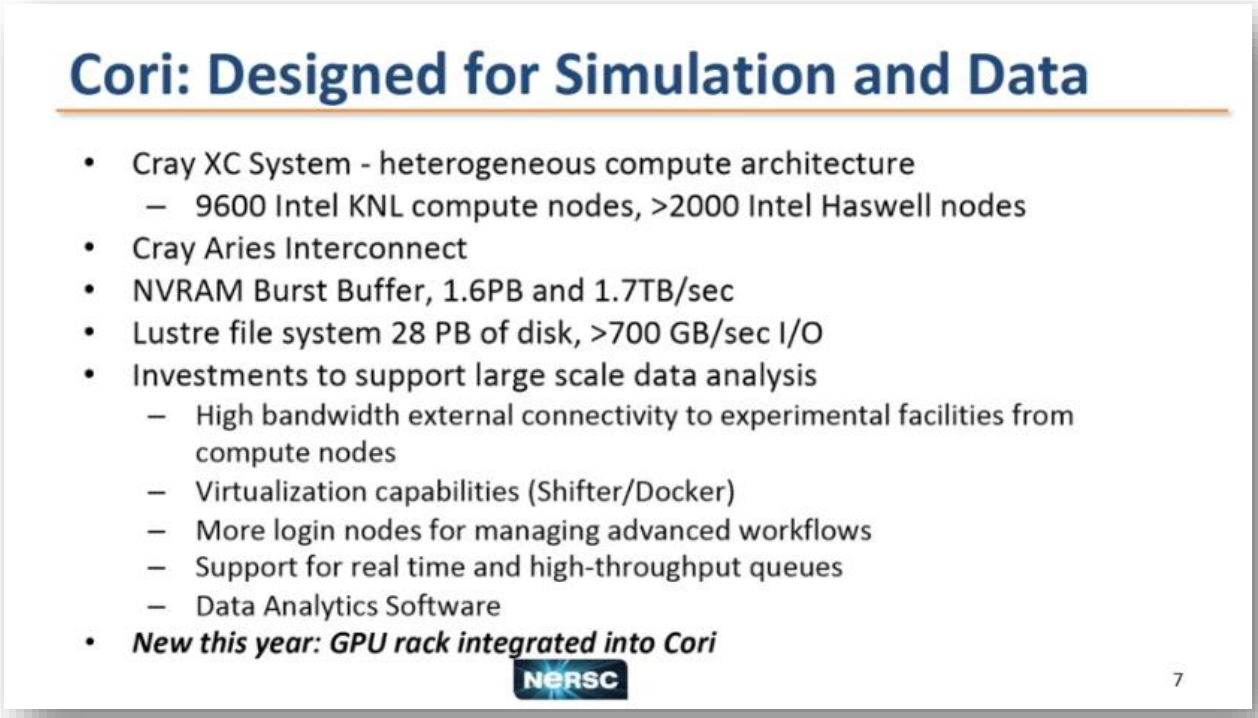


Source: NERSC and Hyperion Research, 2021

Declerck contextualized Perlmutter by providing a brief background on NERSC's modern systems roadmap. The Cray XC30 Edison machine (NERSC-7), built in 2013, was retired in 2019 to prepare for Perlmutter (NERSC-9), leaving the 30PF Cori (NERSC-8) as the only functioning NERSC supercomputer during the interim. Cori is a Cray XC40 system with a heterogenous compute architecture. This was the first system at NERSC aimed at supporting large scale data analysis.

Recent additions to the system were improved bandwidth to external networks for their experimental facilities, and more virtualization capabilities such as Shifter and Docker, allowing for users who have codes that are validated and tested to continue to run those in a container and not have to necessarily use NERSC's latest software stack to keep up to date with fixes, patches, and security updates. Recently, they have added GPU racks into Cori, primarily to allow users a smoother transition to the Perlmutter system. Of the near 8 billion core hours provided in 2020, 39.7% were allocated to jobs which use more than 1,024 nodes. This is all to support the large-scale science that goes on at the site, like simulations of particle in cell plasma, stellar mergers, and quantum circuits. As Perlmutter is still being stood up, Declerck and her team have already started planning for NERSC-10, a ~20MW exascale system slated for 2025.

FIGURE 2



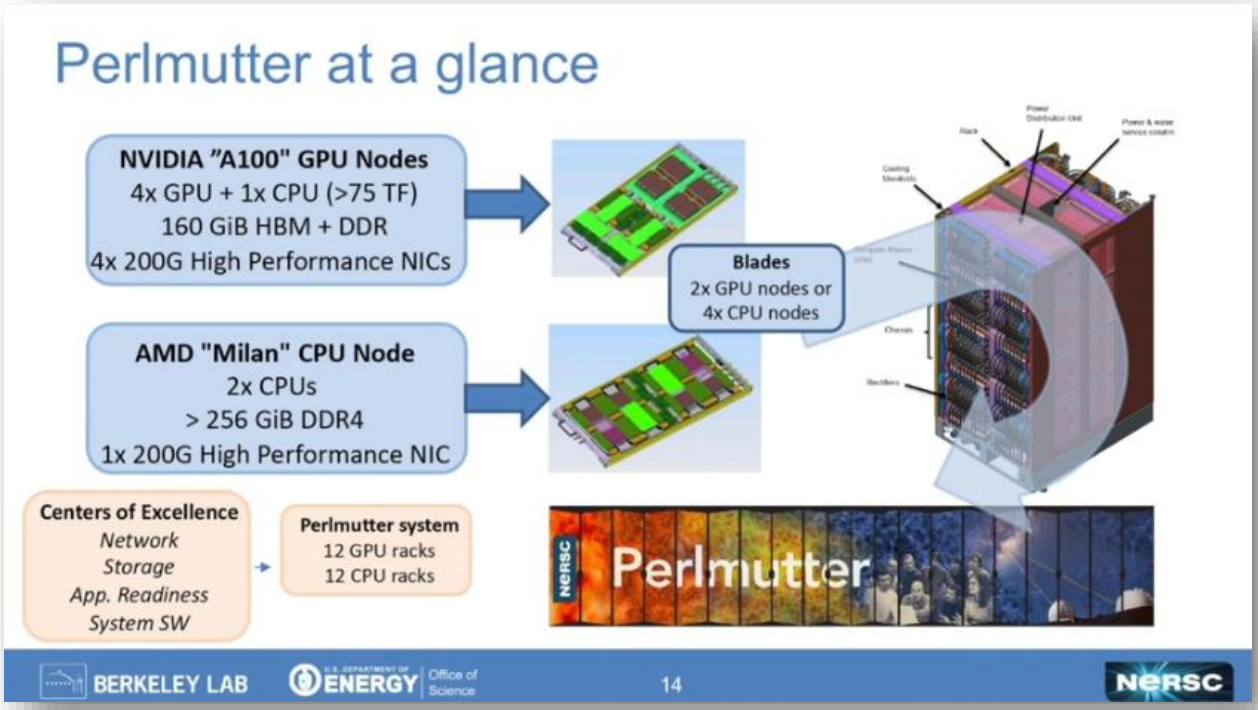
Source: NERSC and Hyperion Research, 2020

Next, Declerck introduced the newest addition to the NERSC HPC lineage: Perlmutter (NERSC-9). Named after 2011 Nobel Prize winner Saul Perlmutter, this Cray Shasta system provides 3-4x the capability of Cori inside its CPU-only partition comprised of 2 AMD Milan CPUs with 256GiB DDR4. It also has over 6,000 GPU nodes with 4 NVIDIA A100 GPUs each with Tensor Cores, NVLink-3, and

High-BW memory, along with a single AMD Milan CPU. The high-performance, scalable, low latency ethernet compatible network is enabled by Cray "Slingshot" connectivity. In the past, the high-speed network has only connected the compute nodes. In this case, the storage and the login nodes are all going to be connected with the high-speed network. Because of this, they are expecting the overall interaction of the system to be much improved. With 6x the bandwidth of Cori, the file system will be the largest all-flash system supported by Lustre yet.

Declerck explains that since its inception, Perlmutter (NERSC-9) had the requirements of simulation and data users in mind. With its all-flash file system, optimized network for data ingestion from experimental facilities, and real-time scheduling capabilities, the Perlmutter machine will empower researchers and computational scientists in a broad range of localities. The modernized, cloud-based system software supports rolling upgrades for improved resilience and additional user flexibility. As one of the first sites to get access to much of this technology, Declerck and her team worked closely with the engineers that developed the software and hardware components to address the challenges that come with erecting such a large and powerful machine.

FIGURE 3



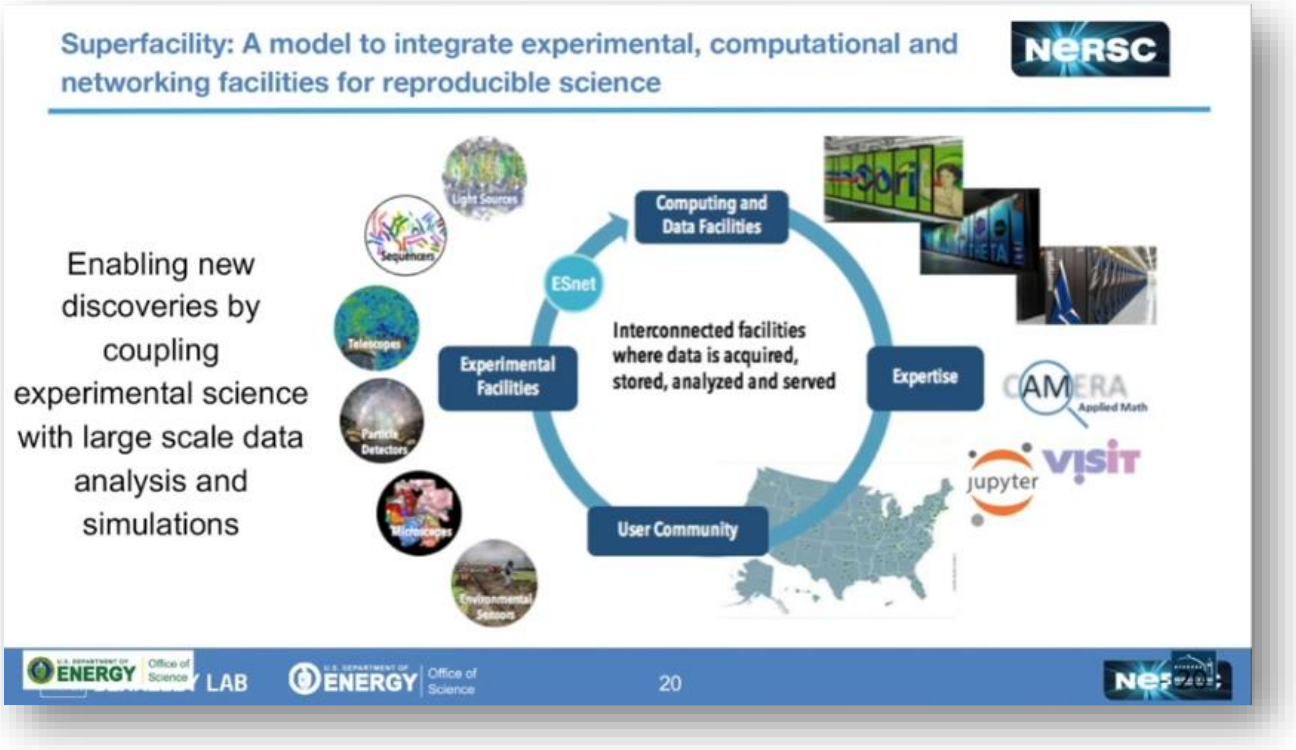
Source: NERSC and Hyperion Research, 2021

To help realize the power of their system, NERSC has a robust application readiness program used to prepare some workloads for the new system. In this program, called NESAP (NERSC Exascale Science Application Program), they partner with application development teams and vendors to port and optimize key applications to DOE's Office of Science. Application teams work with experts at NERSC to improve parallelism and prepare code for use. This process of sharing resources and

working collaboratively to identify and fix areas needing improvement in code result in marked improvements for all measured application types.

NERSC has a continued dedication to offering diverse and unique capabilities to their users, including their community file system, which, while more distant from the machine, has a large capacity. Currently at 75 PB and scheduled to be 150-300 PB by the time of Perlmutter's deployment, Declerck places this nearline file system in the middle of a spectrum between the long-term, HPSS storage, and the Lustre all-flash memory used when running jobs. For some users who need data bases or API portals to do their science, NERSC has enabled certain edge services, now offering over 90.

FIGURE 4



Source: NERSC and Hyperion Research, 2021

A key method of enabling the work done in data intensive scientific research is what NERSC refers to as the Superfacility model. This model, designed to integrate experimental, computational, and networking facilities for reproducible science, allows facilities to analyze data at a much faster pace. This system of allowing data to stream directly from an observational experiment onto a large-scale system for analysis would allow for more efficient, accurate, timely, and cost-effective laboratory use. There are some challenges to implementing this system, including the need for a reliable way to reserve bandwidth and computational resources efficiently, as well as address the diverse needs of scientific facilities that produce data of different kinds.

NERSC's mission is to accelerate scientific discovery at the DOE Office of Science through high performance computing and data analysis. For Declerck, everything they do is to advance science. The Perlmutter machine is the result of years of collaborative effort. Its design, components, and features are meant to reflect the needs of scientists and their research. NERSC continues to set a high standard not only with a powerful flagship computer, but with a holistic approach to community collaboration, networking, and the lifecycle of modern scientific research.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
