National Energy Research Scientific Computing Center

2021 ANNUAL REPORT

U.S. DEPARTMENT OF ENERGY | Office of Science

Table of Contents

The National Energy Research Scientific Computing Center is the mission high-performance computing facility for the U.S. Department of Energy's Office of Science. NERSC is managed by Lawrence Berkeley National Laboratory and funded by the U.S. Department of Energy's Office of Science Advanced Scientific Computing Research Office. NERSC's mission is to accelerate scientific discovery at the U.S. Department of Energy Office of Science through HPC and extreme data analysis.

Director’s Note

In 2021, NERSC had over 9,000 scientists working on more than 900 research projects spanning the range of the U.S. Department of Energy (DOE) Office of Science scientific disciplines. More than 2,000 refereed publications cited NERSC, a demonstration of the reach and impact NERSC continues to have on the science community.

2021 also marked an exciting new era at NERSC: the Phase I installation of *Perlmutter*, the NERSC-9 supercomputer. Named for longtime NERSC user and Nobelist Saul Perlmutter, the new system represents NERSC and Berkeley Lab’s deep commitment to and key role in team science. *Perlmutter* is the first NERSC system designed from the beginning for simulation, modeling, and data analysis. Phase I features 12 GPU cabinets with 6,144 NVIDIA A100 Tensor Core GPUs. In 2022, we will take delivery of Phase II, consisting of 12 CPU cabinets with over 3,000 AMD EPYC Milan processors. The system’s very fast HPE Cray Slingshot network will allow users to stream data seamlessly from remote experimental and observational facilities. The system also features a 35-PB all-flash file system and a software stack for data analysis and AI.

A select and growing group of early users worked with NERSC staff in 2021 for the dual purpose of doing their research and stabilizing *Perlmutter*, helping us prepare for a typical NERSC workload. Here are some highlights:

- **Over 250 science projects used *Perlmutter* Phase 1 successfully in the early science period in 2021.**
- **On a broad set of applications, NESAP teams achieved an average speedup of 40x compared with the earlier Edison system.**
- A 20-billion-atom record-scale molecular dynamics simulation with LAMMPS was a Gordon Bell finalist at SC21.
- **HPL results placed *Perlmutter* at #5 on Top500 (65 PF) and #6 in the Green 500 (25.5 GF/Watt) — the only Top10 system in both the Top500 and Green500.**
- The FourCastNet AI global weather prediction application was run at scale on *Perlmutter* and featured in NVIDIA’s keynote at GTC 2021, as well as being a Gordon Bell run submission.
- *Perlmutter’s* status as a world-class system for artificial intelligence was demonstrated through an impressive set of MLPerf HPC 1.0 results. *Perlmutter* achieved leading results for several benchmarks, as well as the largest-scale GPU submission on 5,120 A100s.
- We overcame numerous COVID-19 challenges to successfully complete a 12.5-MW power and cooling upgrade in Wang Hall in preparation for the delivery of *Perlmutter*.

Constructing the Superfacility

During 2021, NERSC continued its work on the Superfacility effort with the goal of seamlessly integrating compute centers with experimental and observational facilities using ESnet. The

Superfacility project achieved its goal, with five of the eight science engagements demonstrating automated pipelines that analyze data from remote facilities at large scale, without routine human intervention, using these capabilities:

- Near-real-time computing support
- Dynamic, high-performance networking
- Data management and movement tools, including Globus
- API-driven automation
- HPC-scale notebooks via Jupyter
- Authentication using Federated Identity
- Container-based edge services supported via Spin.

All eight teams made significant advances in their workflows as a result of work achieved by the project team. In four cases, we have gone beyond demonstrations and can now provide production-level services for their experiment teams:

- Dark Energy Spectroscopic Instrument (DESI): Automated nightly data movement from telescope to NERSC and deadline-driven data analysis
- Linac Coherent Light Source (LCLS): Automated data movement and analysis using Cori and *Perlmutter*
- Lux-Zeplin experiment (LZ): Automated 24/7 data analysis from the dark matter detector
- National Center for Electron Microscopy (NCEM): Automated workflow pulling data from the 4D STEM camera to Cori for near-real-time data processing.

The Superfacility project also launched the federated identity pilot, enabling users from Berkeley Lab and other DOE labs to log into their NERSC accounts using credentials from their home institutions. This was extended to logins to NERSC's Jupyter hub. Superfacility technical projects have been integrated with NERSC-9, including Jupyter, Globus, real-time QOS and CVMFS, and all Superfacility science teams can now use *Perlmutter*.

Striving for Energy Efficiency

We are proud of the fact that NERSC is one of the most energy-efficient supercomputing centers in the world. We maintained a 12-month power usage effectiveness (PUE) value of 1.07, placing NERSC in the upper echelon of supercomputing and data centers in terms of energy efficiency. PUE measures the percentage of power that is delivered to systems; a value of 1.0 is ideal. Many supercomputer and data centers have a PUE between 1.3 and 1.8. NERSC has been a key contributor to a Lab-wide 26% reduction in energy use in recent years, for which the Lab was awarded ISO50001 status by the International Organization for Standardization (ISO) and a DOE Sustainability Award in the Innovative Approach to Sustainability category. The cooling-system changes made as part of that process increased our previous non-compute energy savings from 37% to 42% — and we continue to press forward with new solutions to meet our aggressive goals in this area.

A Range of Innovations

Our staff continues to support users by answering tickets and holding training events — in 2021, NERSC HPC consultants and account support staff answered over 7,000 tickets while maintaining a three-day SLA metric of nearly 93%. When surveyed, 90% of users reported being "moderately satisfied" or "very satisfied" with NERSC consulting, and 96% of user ticket survey respondents were satisfied with the ticket resolution. NERSC offered 21 distinct training events to NERSC users on a variety of topics ranging from the basics of using NERSC resources to advanced, specialized trainings. All trainings were held online, and recordings were posted to the NERSC YouTube channel, with professional captioning for accessibility. NERSC's best practices for online training sessions were formalized and published in a paper in the *Journal of Computational Science Education* in April 2022.

Other operational highlights include:

- We made investments in the Community Filesystem, the HPSS tape library, and improved networking to better support our users.
- We made improvements to the Slurm scheduler and the implementation of a flex queue, which allowed the use of ~414M compute hours that would otherwise have been wasted.
- NERSC leveraged the continuous integration (CI) infrastructure developed in collaboration with the Exascale Computing Project to install and test software.
- We upgraded our network to 400 Gbps in preparation for the launch of ESnet6, the new generation of our scientific network that supports scientific partnerships and further integrates the nation's scientific user facilities and high performance compute resources.
- We continue to work with the developers of the DMTCP checkpoint-restart library to ensure that the MANA library (a plugin enabling DMTCP to work independent of which MPI library or network underlies a machine's software stack) is bug-free and performant. We have begun a three-way collaboration with Northeastern University (home of DMTCP) and a company called MemVerge, to this end. We held the Second International Symposium on Checkpointing in Supercomputing as an SC21 workshop, and have begun collaborating with academics and national lab and industry contacts on developing a checkpoint/restart standard (similar to the MPI standard or OpenMP standard process), for which the first version is targeted to be released at SC22.
- We kept NERSC systems and storage secure despite increasing attempts to compromise HPC systems worldwide.

As we look to the future, the NERSC-10 mission need was signed by DOE, formally kicking off the project to follow *Perlmutter* in 2026. NERSC staff have developed an exciting vision for this system, which we will share with you all in the coming year. We are also investing in technology explorations for 2030 that range from compute in networking and storage to quantum computing.

As I look back on the year, I am amazed by how much our staff accomplished. It is a true testament to their dedication and hard work. I am looking forward to 2022 and seeing all the great science that will be produced by our users.

— **Sudip Dosanjh**
NERSC Division Director

NERSC by the Numbers

2021 NERSC USERS ACROSS US AND WORLD

50 States + Washington D.C. and Puerto Rico

46 Countries

~9,000 ANNUAL USERS FROM ~800 Institutions + National Labs

27% Graduate Students

17% Postdoctoral Fellows

14% Staff Scientists

11% University Faculty

7% Undergraduate Students

6% Professional Staff

59% Universities

29% DOE Labs

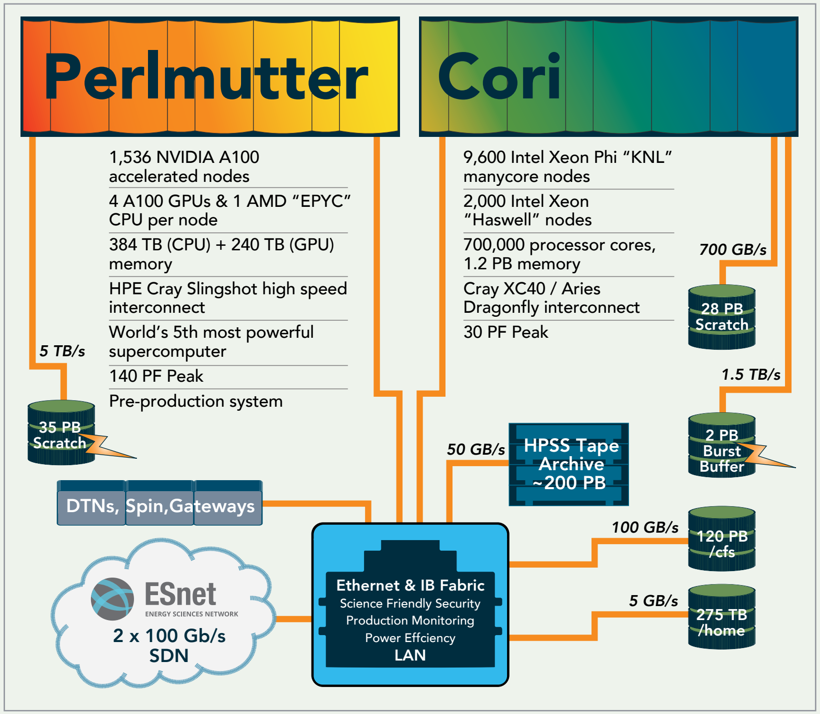
5% Other Government Labs

3% Industry

1% Small Businesses

<1% Private Labs

NERSC Systems 2021



8.25 BILLION

COMPUTE HOURS USED IN 2021

>2,000

Refereed Publications Cited NERSC

Data Stored

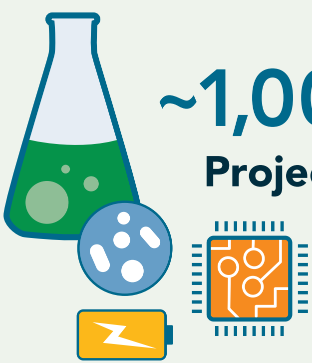
200

Petabytes

Top Science Disciplines

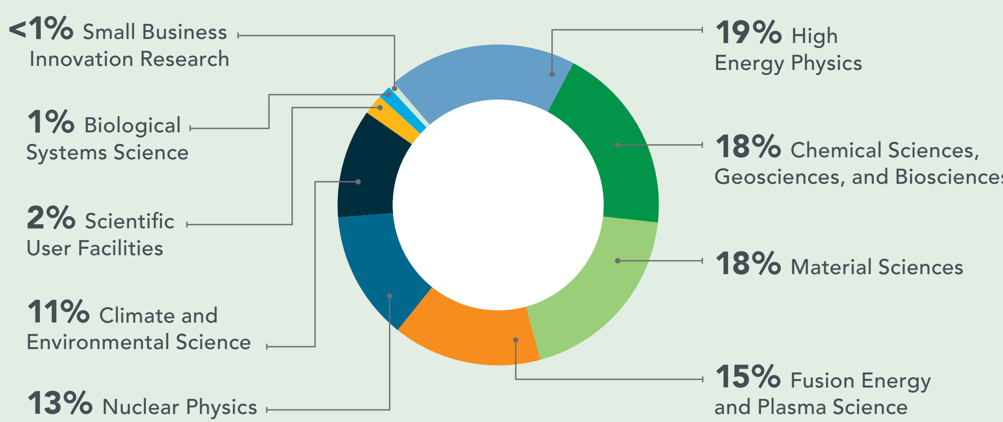
(By computational hours used)

Nuclear Physics
Scientific User Facilities
Chemical Sciences
Material Sciences
High Energy Physics
Geosciences
Biological Systems Science
Fusion Energy
Small Business Innovation Research
Biosciences
Plasma Science
Climate and Environmental Science



~1,000 Projects

Breakdown of Compute Used by DOE Program



2021 Milestones

Bringing *Perlmutter* to Life

2021 brought the fruition of a much anticipated project: initial installation and early science use of *Perlmutter*, the next-generation supercomputer named for Berkeley Lab physicist and Nobelist Saul Perlmutter.

Perlmutter is an HPE Cray EX supercomputer that features both GPU-accelerated and CPU-only nodes and has a projected performance of three to four times that of NERSC's Cori supercomputer. Debuting in the Top500 ranking in June 2021 as the fifth-fastest supercomputer in the world, the system is designed for the future of high performance computing (HPC), particularly the machine learning that researchers are increasingly turning to for results that would otherwise be nearly impossible. Eventually serving more than 8,000 researchers around the world, *Perlmutter* will expand what's possible for simulation and near-real-time data analysis, with a degree of energy efficiency that's among the best in the world.

Figure 1. Installation of the decorative panels on the exterior of the computer cabinets for *Perlmutter*.

Perlmutter is being installed in two phases. Phase 1, which includes the system's GPU-accelerated nodes and scratch file system, became available for early science campaigns starting in summer of 2021; Phase 2 will add CPU-only nodes in 2022. In addition to installing the components of *Perlmutter* in 2021, NERSC staff made a series of custom enhancements to the system, including:

- Integrating NVIDIA system software on NVIDIA GPUs with existing software infrastructure provided by HPE on both user access and compute nodes.
- Applying an extensive and expanding automated testing suite in ReFrame to all software deployed on *Perlmutter* to ensure a stable environment for users. HPE has now internally adopted the tests NERSC shared with them during the testing process.
- Integrating *Perlmutter* with NERSC's global storage and network infrastructure, which required routing between *Perlmutter's* internal fabric and the fabric in NERSC's global filesystems. Use of virtual/floating IP addresses and the virtual router redundancy protocol (VRRP) solved these problems.
- Noting the lack of an interface for checking network health or monitoring trends in the Slingshot Fabric Manager, and creating a Fabric Manager Monitor application to run on the Kubernetes platform, allowing for easy monitoring of the Slingshot network.
- Instituting a new, more resilient deployment of Slurm Workload Manager to better integrate with the Kubernetes platform, including separate deployments for different management tasks to ensure redundancy.
- Collaborating with HPE to improve the reliability and security of the Shasta system management networks on which *Perlmutter* is based.
- Adding an additional out-of-band management network to *Perlmutter* to enable remote access during the COVID-19 pandemic. This out-of-band access allowed us to perform major system installation and reconfiguration tasks entirely remotely. We estimate that the accessibility provided by this network saved four to six weeks of calendar time in making *Perlmutter* operational, and we expect it to continue to be a valuable aid during the Phase 2 installation, system updates, diagnosing system problems, and keeping *Perlmutter* a reliable system over its lifetime.

Early User Onboarding and User Integration Team

During acceptance of *Perlmutter*, it was vital that the machine be tested with a workload representative of the normal workload at NERSC. To do that, we allowed strategic groups of users early access to the system. Users were chosen from our NERSC Exascale Science Applications Program and Exascale Computing Project engagements, with an eye toward users who would run a wide range of applications and be understanding of the type of disruptions that routinely occur when stabilizing a new system.

Before the first users were allowed on the system, the *Perlmutter* User Integration Team began meeting weekly, with between-meetings communication taking place via a dedicated Slack channel and long-term internal issues tracked via GitLab. A subset of the User Integration Team attended daily meetings with the Systems team and the vendors to track user-facing issues and communicate updates to the rest of the team, while others had direct engagements with the science teams running on *Perlmutter* or the vendors providing the software. The entire team met weekly to review outstanding issues and prioritize fixes or changes. Members of the User Integration Team wrote 135 tests in the ReFrame testing framework that were used for validating changes to the environment before exposing them to users.

Initially we added about a dozen users to the machine, and slowly increased the number of users over time as system stability improved. Because the system was changing so rapidly early on and the pool of users was so small then, we provided user support via live interactions — a setup beneficial to both users and staff because it provided rapid, actionable feedback that was useful for debugging, stabilizing, and improving the user environment. When the number of active users exceeded 150, we transitioned to a new *Perlmutter*-focused Slack channel and then to the traditional NERSC ticketing system (ServiceNow).

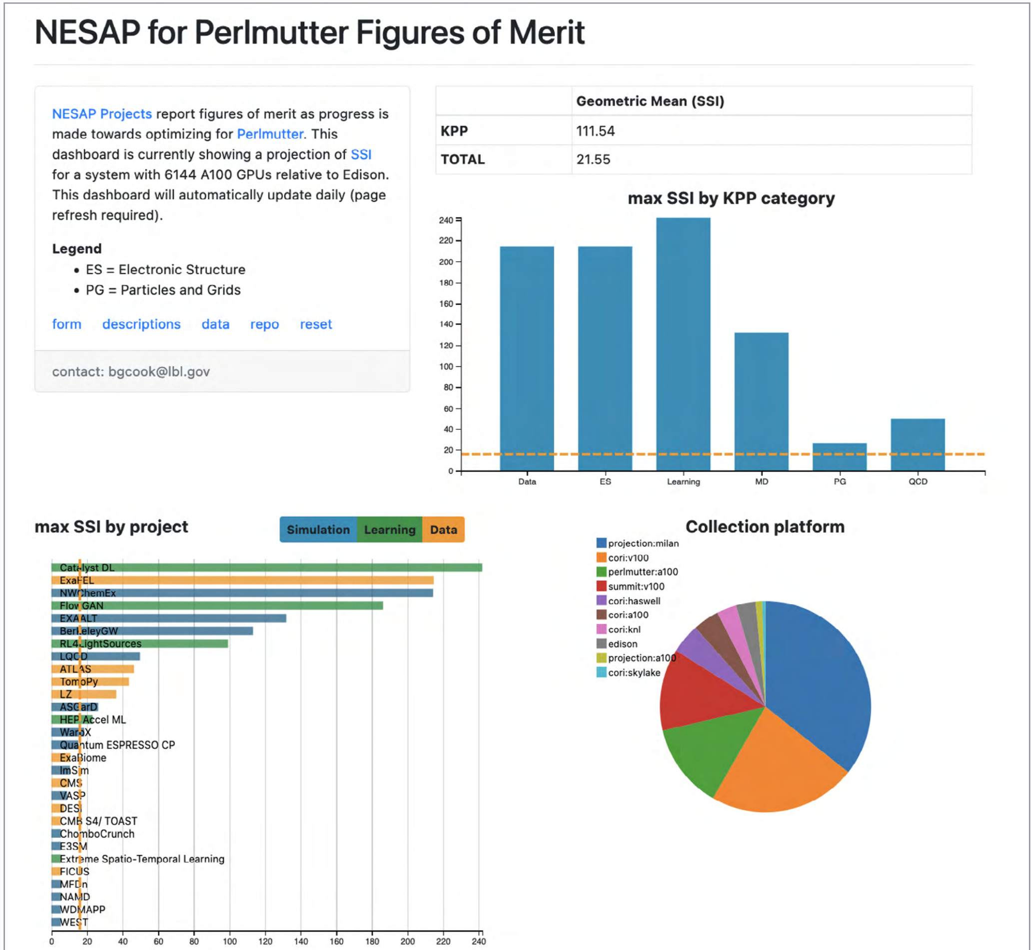
Despite the considerable churn necessary for acceptance, NERSC staff were able to deliver a stable and productive system for users in 2021.

NESAP for *Perlmutter*: Preparing the NERSC Workload

The NERSC Exascale Science Applications Program (NESAP) is a collaborative effort in which NERSC partners with code teams, vendors, and library and tools developers to prepare for advanced architectures and new systems. NESAP began in late 2014 to help users prepare for the Cori manycore Knights Landing/Xeon Phi architecture; today it targets *Perlmutter*, NERSC's first production GPU-based system.

NESAP aims to improve science outcomes with upcoming systems — a broader goal than simply optimizing legacy codebases. At present, there are 59 NESAP projects in place that are designed to support research on the *Perlmutter* system, spanning three focus areas:

- **NESAP for Simulations:** Cutting-edge simulations of complex physical phenomena
- **NESAP for Data:** Data analysis science pipelines that process massive datasets from experimental and observational science facilities
- **NESAP for Learning:** Machine learning and deep learning solutions to improve scientific discovery potential on experimental or simulation data, or improve HPC applications by replacing parts of the software stack or



algorithms with machine learning/deep learning solutions.

NESAP made substantial progress in 2021, with several projects showing figure-of-merit improvements through the combined efforts of NERSC postdocs, NERSC staff liaisons, vendor partners, and project developers. Progress is being tracked by a NERSC-developed custom dashboard that maintains a live view of the current status of each NESAP project and a projection of the total scientific throughput, as measured by the scalable system improvement (SSI) metric, by project and by category. These metrics increased noticeably in 2021; for example, the geometric mean SSIs are both about 25% higher than at the end of 2020 (see Figure 2).

Figure 2. NESAP-developed custom online dashboard for tracking applications progress towards *Perlmutter* performance goals, as measured by the scalable system improvement (SSI) metric.[1]

[1] The SSI is not a simple speedup factor but aims to provide a single metric to measure performance improvement for a wide variety of application and platform characteristics. A detailed discussion can be found at **https://www.nersc.gov/research-and-development/benchmarking-and-workload-characterization/ssi/**

Current NESAP projects running on *Perlmutter* include:

NESAP Efforts on EXAALT Lead to Gordon Bell Finalist Honors

EXAALT, an ECP project that joined the NESAP program in 2019, was honored as a Gordon Bell Award finalist at SC21. Since 2019, the performance of Spectral Neighborhood Analysis Potential (SNAP) in LAMMPS on a single NVIDIA V100 GPU has seen a 26x speedup compared to the pre-2019 implementation on the same architecture.

Researchers from University of South Florida, along with scientists from Los Alamos, NVIDIA, and NERSC, used the optimized LAMMPS-SNAP implementation to simulate the behavior of 20 billion atoms of BC8 carbon at extreme pressure (12Mbpar) and temperature (5000K) over 1 nanosecond of physical time. This simulation took 24 hours on 4,650 nodes (27,900 NVIDIA V100 GPUs) of Summit and was made possible by the efficient implementation of LAMMPS-SNAP using the CUDA backend of the Kokkos framework.

FourCastNet: Scaling AI to Unprecedented Resolution in Climate and Weather Forecasting

Forecasting global weather phenomena at high resolution is of the utmost importance to human efforts in understanding and adapting to changes in the climate.

Doing so also consumes substantial resources at HPC centers around the world, as modeling atmospheric dynamics with traditional solvers is computationally heavy. In 2021 the NeuWCast NESAP for Learning project made important headway in reducing this computational burden by introducing a fully data-driven forecasting model, powered by recent advances in AI algorithms, that can accurately forecast important variables such as surface winds, temperature, and precipitation at a fraction of the computational cost.

NESAP for Learning postdocs at NERSC were central to the conception, development, scaling, and optimization of this model, and heavily contributed to the writing process as well.[2] The model can be trained in a distributed fashion using GPUs from *Perlmutter* in a matter of hours, and can produce forecasts using a single GPU in seconds. The speed of inference enables substantial speedups (in the tens of thousands) over traditional numerical solvers for ensemble forecasting.

ExaFEL: Workflow and Performance Portability for the Next Generation of XFEL Data Analysis

ExaFEL performs comprehensive analysis of X-ray diffraction patterns from protein crystals to extract a higher level of atomic detail and thus molecular function. X-ray pulses from SLAC's Linac Coherent Light Source can capture molecular movies of enzymes reacting in femtoseconds. We are developing unique algorithms to interpret the experimental diffraction images to pixel-level resolution.

[2] https ://arxiv.org/abs/2202.11214

XFEL data analysis must be done in real time to enable good decision-making during ongoing experiments. Data-collection rates are outpacing computational resources at the experimental sites, requiring a Superfacility approach.[3] This approach has proven fruitful for enabling high-impact research and was published in *Nature*[4] and *Nature Scientific Reports*.[5]

Figure 3. An inside view of *Perlmutter* reveals that the compute cabinets are liquid-cooled. Redundancy in cooling units is provided as needed. All non-compute nodes housed in service cabinets are air-cooled.

[3] https://arxiv.org/abs/2106.11469

[4] https://doi.org/10.1038/s41586-021-04218-3

[5] https://doi.org/10.1038/s41598-021-00236-3

NESAP Postdoc Anticipates *Perlmutter* for Deep Learning

Deep learning is becoming more prevalent in scientific domains because datasets continue to increase in size. Brandon Wood, a NESAP postdoc at NERSC, spent 2021 helping to create a large catalysis dataset called Open Catalyst 2020, or OC20, and an associated community challenge with Zack Ulissi's group at Carnegie Mellon University and researchers at Facebook AI Research (FAIR).

According to Wood, one thing has become obvious: The way chemists and materials scientists think about and utilize GPU resources is rapidly changing.

Going beyond chemistry and materials science have typically been data-poor domains compared to other fields such as image processing or natural language processing. At the outset of Wood's work with Ulissi's group, they had a dataset of around 50,000 examples, and models such as crystal graph convolutional neural network (CGCNN) with 1000s of parameters could be trained to convergence on one GPU in a day. Conversely, with the largest OC20 training dataset (~130M examples) and the largest version of the DimeNet++ model reported in their paper (~10M parameters), it took 256 GPUs the better part of a week to train (1000s of GPU days). DimeNet++, a GNN that incorporates angular information through directional message passing, is memory intensive, which limits the local batch size and results in using many GPUs to achieve a sufficiently large global batch size. This massive change in compute requirements quickly outpaced their resources.

This is where *Perlmutter*, with its 6,000+ NVIDIA A100 GPUs, each with 40GB of memory, comes in. One goal of this NESAP project (CatalysisDL) was to gain an understanding of the limits of deep learning on a large materials dataset. The pre-*Perlmutter* paradigm was one of improved model performance when the number of trainable parameters was increased and the data was not yet overfitted. The GPU resources on *Perlmutter* allow researchers to explore questions about model size that have been prohibitively expensive for academic researchers in the past. Going beyond training a number of large models, moving forward, the *Perlmutter* ecosystem will enable other necessary aspects of modern deep learning workflows such as model iteration, retraining of models with new data, training state-of-the-art models, hyperparameter optimization, putting models into production, and incorporating other data science tools such as JupyterHub and Dask — an increasingly valuable resource for the academic community interested in deep learning for science.

This piece was adapted from ***a NERSC staff blog post*** *by NESAP postdoc researcher Brandon Wood.*

COVID-19 Response: Thriving Remotely

As COVID vaccines became available for most Americans and the COVID pandemic entered a new phase, NERSC maintained a mixed approach. User-facing staff worked remotely, with a single member of the Operations Technology Group (OTG) always on site to monitor the machine room. Staff who directly worked on the facility electrical, cooling, water, and other operational facets made regular, solo visits to the machine room to assess its state and fix any problems.

Groups with some need to work on site entered the site only when necessary. The network team worked about 30% on site, and the facility team about 25%. The storage team was on site 5–10% of the time and relied more on contractors, vendors, and NERSC Operations to assist with hardware installation, problem resolution, and de-installation. Onsite visits were sometimes performed on weekends as staff tried to juggle going on site with needs at home.

From the user perspective, little changed at NERSC. Pre-pandemic, users did not often come to NERSC in person except for training events or the annual user meeting. We have continued to hone our process for providing purely online training events and shared our best practices for remote training through a paper at the Eighth SC Workshop on Best Practices for HPC Training and Education at SC21. Throughout the year, we continued to hold many user training sessions, and users expressed positive feedback.

Supporting COVID research

NERSC also continued to support COVID-19-related research in 2021, helping to identify and grant awards to researchers who could leverage NERSC's capabilities in research teams that were supported in 2021. Projects were provided with resources on Cori, including its GPU-accelerated nodes and large-memory nodes that were procured with funding from the Coronavirus Aid, Relief, and Economic Security (CARES) Act. In addition, each research project was assigned an HPC technical expert from the NERSC staff to serve as a liaison. The liaisons helped teams navigate NERSC policies and procedures, provided optimized applications, set up edge services, implemented expedited job throughput when needed, and provided general consulting support.

NERSC is a member of the national COVID-19 High Performance Computing Consortium as a resource provider and also contributed to the proposal review process. This unique private-public effort brings together federal government, industry, and academic leaders to volunteer free computing time and resources on their world-class machines.

Building Out the Superfacility Concept

The Superfacility concept was first proposed almost 10 years ago as a model to leverage HPC for experimental science. In its simplest form, it can be considered a model of connected facilities: an experimental facility connected to a high-performance network (usually ESnet), connected to a supercomputing center. But to make this a productive environment for a scientist, we also need to consider the full ecosystem of infrastructure, software, tools, and expertise needed to make connected facilities easy to use.

A key principle behind the Superfacility Project has been to ensure that the solutions developed to enable new science scale across multiple user communities. We wanted to avoid unsustainable one-off solutions that place large support burdens on both science teams and facility staff, and instead design an ecosystem of interconnected technologies that would fulfill the needs of the majority of the NERSC user base. To this end, in late 2021 we selected a set of seven (later eight) science engagements that would drive our requirements and provide us with highly engaged beta testers. We chose these engagements carefully, based on:

- Science area: Representing experiment teams across the DOE SC
- Workflow complexity: Representing a range of complicated end-to-end workflows that have data sources within Berkeley Lab, elsewhere in the US, and internationally.
- Scale: From relatively small teams (a few scientists) to the largest teams at NERSC (a few hundred scientists), with compute needs ranging from a few million to a few hundred million hours a year, and with data rates in the TB/year to PB/year.
- Timeframe: From experiments currently taking data to more forward-looking projects preparing for future large data rates.

Introducing Real-time Scheduling

One unprecedented feature of the workloads we aim to support with the Superfacility Project is near-real-time computing. Several of our science engagement teams require HPC-scale computing at NERSC to analyze their data in the moment, monitoring and controlling their experiments as they operate (such as seeing results of an experiment within minutes to decide how to adjust an experiment for the next X-ray shot). Others require short turnaround data analysis to ensure timely decision making (such as analyzing nightly telescope data by morning). This is a new paradigm in scheduling busy HPC centers, which typically rely on a batch system to get through the constant queue of compute jobs waiting to run.

For NERSC, the solution to this problem included both policy decisions and engineering work.

NERSC identified two solutions, both of which are in use on NERSC systems today:

1. The "real-time" queue

Since 2017, NERSC has supported small-scale compute needs that require very short turnaround via a "real-time" queue. Access to this queue is tightly constrained to NERSC users who have genuine need for it (including ALS, DESI, and LZ). A small set of 20 compute nodes is held in reserve to serve this queue. If all of these nodes are occupied, incoming near-real-time jobs are assigned a high priority in the Slurm scheduler, which allocates them to the next available compute nodes. In a system as busy as NERSC's Cori (which typically has more than 2,000 jobs running simultaneously), the wait for resources is typically minimal.

2. Reservations + preemption

The near-real-time needs of experiment facilities like LCLS are projected to rise rapidly in the next five years; when LCLS reaches full capacity, researchers there will regularly require hundreds of petaflops of compute to monitor and guide their experiments. This scale is very difficult to supply via the real-time queue, so in collaboration with SchedMD, we developed a system of reservations to provide the required near-real-time compute availability while maintaining high utilization of our systems. In the case of LCLS, the need is not random or unpredictable — LCLS needs this level of compute capacity for specific experiments that happen at scheduled times in the year. With advance warning, we can reserve compute nodes for the duration of the LCLS experiment so they can surge to NERSC when they need them.

The challenge in this scenario is that experiments don't always run smoothly. During a shift at an experiment, equipment may fail, or samples may be hard to adjust, so for much of the time the reserved compute nodes may be sitting idle. To avoid this and maintain good use of our compute resources for science, we enabled preemptible jobs to run in reservations. This means that any flagged compute job can run on nodes set aside for a reservation — with the expectation that the jobs will be canceled within a given time frame (generally a couple of minutes) so that the nodes can be given to the experiment team when they have compute jobs flowing to NERSC (see Figure 4). This capability was funded by NERSC via a non-recurring engineering agreement with SchedMD and was deployed in 2021.

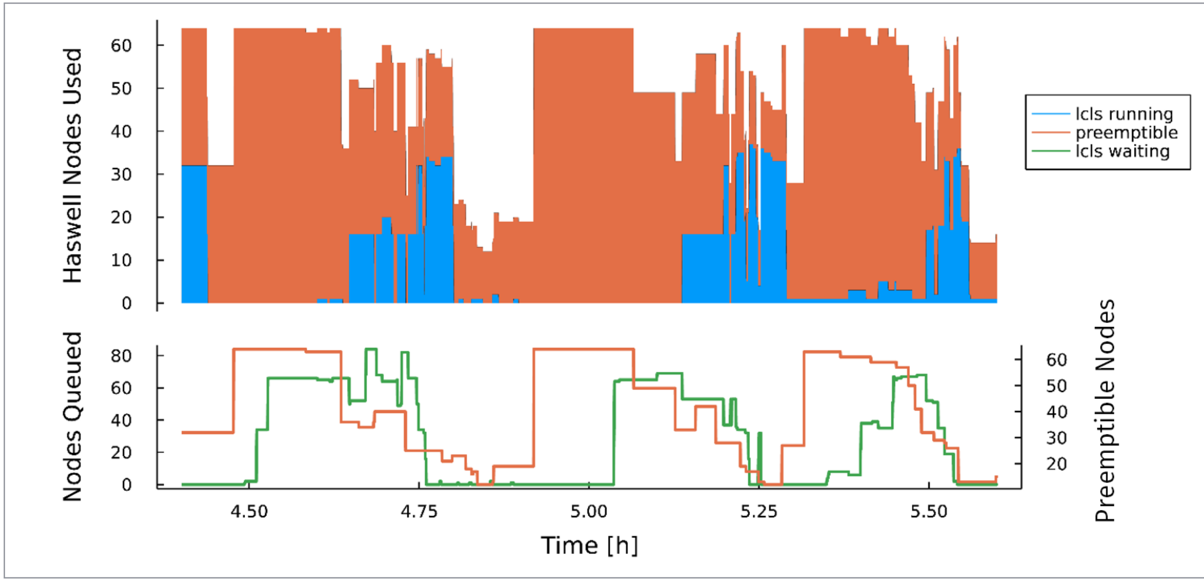


Figure 4. Example of compute usage of a reservation for a running experiment, and how near-real-time data analysis jobs (blue) can co-exist with preemptible jobs (orange). 64 nodes (2,048 cores) were reserved, to which near-real-time jobs have priority access. Near-real-time data analysis is bursty by nature; data analysis requires a large amount of computational resources immediately after measurement runs (approximately every 30 minutes). While data is being collected, however, reserved nodes would sit idle. This usage pattern is ripe for preemptible jobs as these can "sit between" near-real-time data analysis bursts.

FedID Takes NERSC User Security to the Next Level

Every discussion of distributed workflows eventually touches on the issue of authentication and the complexity and inconvenience brought on by having multiple passwords and accounts across instrument and compute facilities. The work in this area was intended as a first step toward addressing these problems by embracing modern approaches to identity management and leveraging modern distributed authentication protocols — OpenID Connect (OIDC) and the Security Assertion Markup Language (SAML) — and applying them to science use cases.

Federated identity (FedID) is the concept of establishing a trustworthy association between an individual's digital identities; federated authentication is the mechanism that uses these associations to authenticate to systems and services. NERSC is piloting FedID with Berkeley Lab, allowing users with Berkeley Lab accounts to link them to their NERSC accounts, then use them to log into core NERSC systems (see Figure 5). During the pilot, the systems at NERSC that support federated authentication are Iris (account and allocation management), ServiceNow (support tickets), the NERSC web site, and Spin. In the coming months, the FedID capability will be expanded to hundreds more institutions and additional services at NERSC.

The distributed nature of FedID requires that trust be placed in additional technology and other organizations and their security practices. NERSC underwent a full study comparing its conventional, self-contained authentication versus the federated approach, which found that security requirements would be satisfied by:

- Scoping its implementation to DOE-managed facilities and members of the InCommon Federation, an established trust network of higher education and research institutions created to enable federated identity in the academic community.
- Restricting any non-DOE participating institutions to those that comply with SIRTFI (a set of best practices for organizational security) and the REFEDS MFA Profile (a protocol extension that signals when users perform multi-factor authentication).

The initial pilot of FedID for the subset of NERSC users that are affiliated with Berkeley Lab has shown that we can provide an enrollment process that is easy to navigate, that the impact of additional support tickets is minimal, and that, if made available, many users will take advantage of it (more than 300 Berkeley Lab users signed up in the first six weeks of availability). In the meantime, work remains to make this more broadly useful:

- Expand eligible users to members of other DOE labs and eventually to all members of the InCommon identity providers that meet our security requirements.

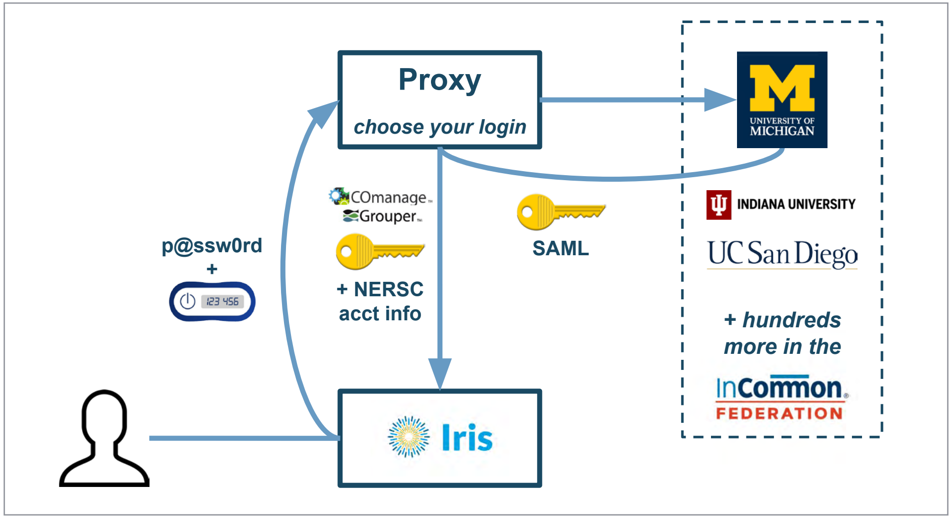


Figure 5. Federated authentication login flow at NERSC.

- Expand the applications that can utilize our federating identity framework by enhancing our login proxy to work with the OpenID protocol in addition to the SAML protocol.
- Develop a web application that can provide SSH certificates via a federated login.
- Improve the user experience during the multi-factor authentication (MFA) step-up process. Currently, if the user's login provider doesn't signal to NERSC that a multi- factor authentication was used, we "step up" the authentication by issuing an MFA challenge for a NERSC token. The current software doesn't limit the number of MFA challenges a user will receive, and adding the concept of an MFA session to the proxy would improve the user experience by reducing the number of MFA challenges to the one per day they have grown accustomed to in our legacy login process.
- Leverage capabilities made available through the Office of Science Distributed Computing and Data Ecosystem project. Developments out of this project could streamline the process of managing the members of science projects that work across DOE facilities, as well as streamline the process of enrolling a federated identity at NERSC.

NERSC API: A Modern, Secure Interface for Science

As automation becomes increasingly required for complex workflows and science gateways become an increasingly popular digital interface for user communities, it is important for HPC centers to provide a modern and secure means for software to interface with batch workload managers, storage systems, and other supercomputing facility resources. For these reasons, the NERSC API debuted in late 2021.

The NERSC API is a modular application programming interface that allows science teams and experimental facilities to securely log into NERSC systems and perform their work. The API was developed with input from several science teams currently using NERSC resources via mostly home-grown software. This allowed us to identify common abstractions that then became API components.

- The following API components are available to our users: Account: Get accounting data about the user's projects, roles, groups, and compute and storage allocations
- Compute: Run batch jobs and query job and queue statuses on NERSC compute resources
- Status: Query the status of NERSC component systems
- Storage: Transfer files between Globus endpoints
- Utilities: Traverse the file system, upload and download small files, and execute commands
- Tasks: Query the status and results of asynchronous operations (most endpoints are asynchronous, and will run in the background until complete)
- Reservations: Make and amend requests for NERSC compute resources ahead of time (currently under implementation)

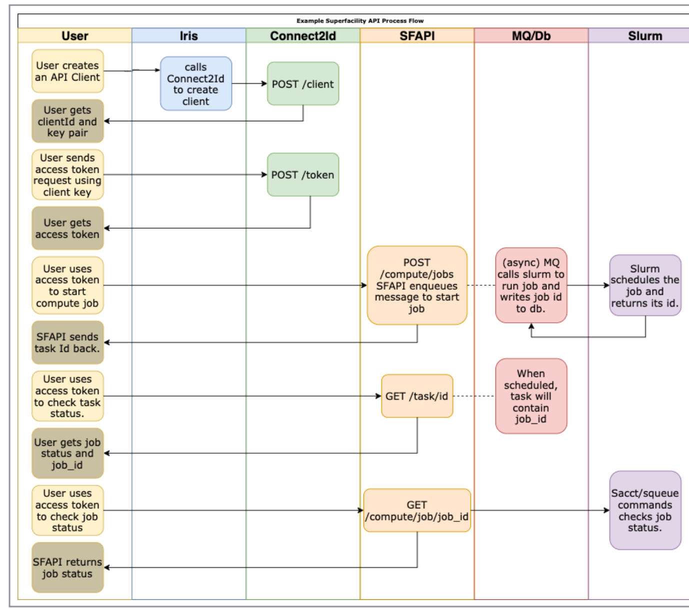


Figure 6. A typical scientific workflow using the NERSC API.

Figure 6 is the call graph of a typical scientific workflow using the NERSC API.

The NERSC API is designed for high-throughput, concurrent access. For this reason, it is deployed on our in-house cloud, Spin. It runs as a set of scalable, containerized microservices so the maintenance team can monitor its performance and easily scale up each component in case of unforeseen demand.

It is essential that the API be secure as well as easy to use. We undertook a full security review of the API throughout the course of its development and deployed it to users only once we had established a well-tested mechanism to grant access. Users need to be granted an MFA-validated token to access API endpoints, which is tied to a specific IP address. This mitigates the risk of a token being stolen and used elsewhere.

The NERSC API soft-launched in late 2021 and has already been adopted by more than 20 users. These projects use the API for automated data management and movement, job submission, software continuous integration, and monitoring NERSC status. For example, LCLS relies on the API as a key part of an automated data analysis pipeline used for real-time data processing during a running experiment. Detailed demonstrations have already been given at various venues, and focused efforts toward user training are ongoing, including tutorials at major conferences and dedicated training sessions for staff and users.

The ability for users to programmatically request reservations in the NERSC API is still in progress. In the near future, we envision a semi-automated workflow for users to request, manage, and cancel reservations. The system should be able to handle overlapping reservation intervals, cancellations, and notifications to all parties involved.

Besides reservations, we hope to support better logging and monitoring around the API. Once better monitoring is in place, usage statistics should give us insight into ways to improve the API, add higher-level functionality, and more.

Innovations in Job, System, and Facility Monitoring

LDMS Offers Improved Management and Insights

NERSC has continued to enhance and leverage Lightweight Distributed Metric Services (LDMS) to enable NERSC staff and users to gain new insights into the NERSC workload. We deployed a scalable pipeline to merge LDMS data with Slurm accounting logs to enable Slurm job and job step level analysis.

This, combined with the Superfacility API, has allowed us to integrate with the Iris web interface to provide on-demand plots of metrics such as memory usage versus time for user jobs (Figure 7). This can help with diagnosing common user issues such as running out of memory.

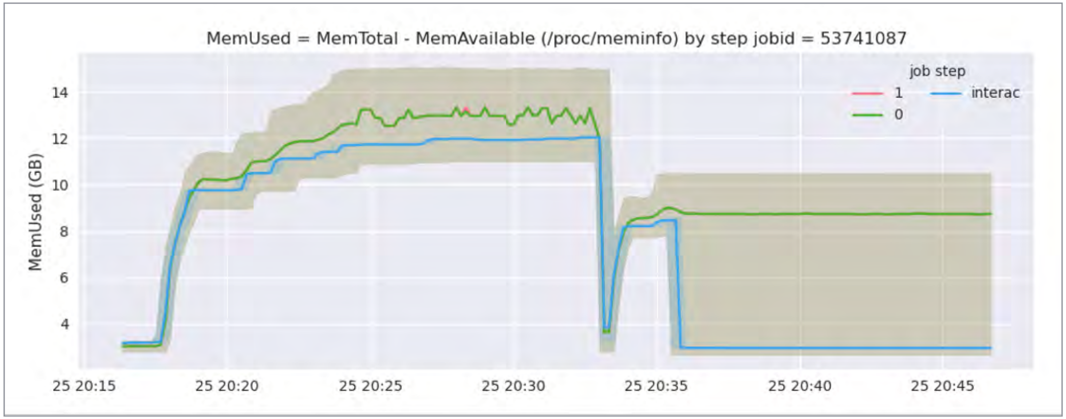


Figure 7. An on-demand plot of memory usage versus time for user jobs.

In addition to these improvements, we have leveraged the new Slurm + LDMS data pipeline with summer student researchers to prototype methods to "fingerprint" jobs. This opens new avenues to aggregating and understanding the NERSC workload. We also used this functionality to explore and understand the available metrics for GPUs and how they might be visualized by users. These enhancements and the insights we can provide to our users and NERSC staff demonstrate how the LDMS framework serves as a foundation for increasingly sophisticated monitoring, job analysis, and gathering business intelligence.

Enhanced *Perlmutter* Monitoring

Perlmutter provides significantly more diagnostic and monitoring data than any previous system at NERSC. This additional data opens the door to real-time system monitoring and helps the NERSC Operations team respond even more quickly to issues by helping identify root causes as they occur and saving logs for later analysis. Because the tools provided by HPE were insufficient to support real-time monitoring, NERSC staff collaborated to create their own scalable solutions.

The first step was to upgrade NERSC's Operations Monitoring and Network Infrastructure (OMNI) data-collection warehouse. OMNI was designed from the beginning to accommodate large data streams, but none as big as the monitoring data flowing out of *Perlmutter*. Modernizing OMNI to be more scalable and robust was vital. NERSC upgraded the cluster to utilize containerization deployed using lightweight Kubernetes (k3s). Today, OMNI consists of multiple k3s clusters, each of which is self-contained with master and worker nodes and its own control plane. When we need to monitor data from new sources, we can rapidly deploy k3s resources without interrupting any of the currently running services or infrastructure.

NERSC then deployed scalable tools on top of OMNI, such as VictoriaMetrics, an open-source, high-performance, time-series database in which Perlmutter metrics are stored and queried for analysis. Visualizations of data can proceed on multiple servers at once due to the performance characteristics of VictoriaMetrics databases. NERSC also deployed and integrated Loki, an open-source log aggregation tool, to stream logs into OMNI. Previously, logs and events were not stored together; now, they can be accessed from the same source and directly compared and correlated.

NERSC uses Grafana to visualize Perlmutter monitoring data, transforming the data into visualizations enabling critical insight into system health and performance (see Figure 8). The value of the system was demonstrated when we were able to troubleshoot issues with *Perlmutter's* cooling system and gain insights that HPE then used to further diagnose the underlying cause. Another example is shown in Figure 9, which provides a dashboard element to monitor the status of Slurm.

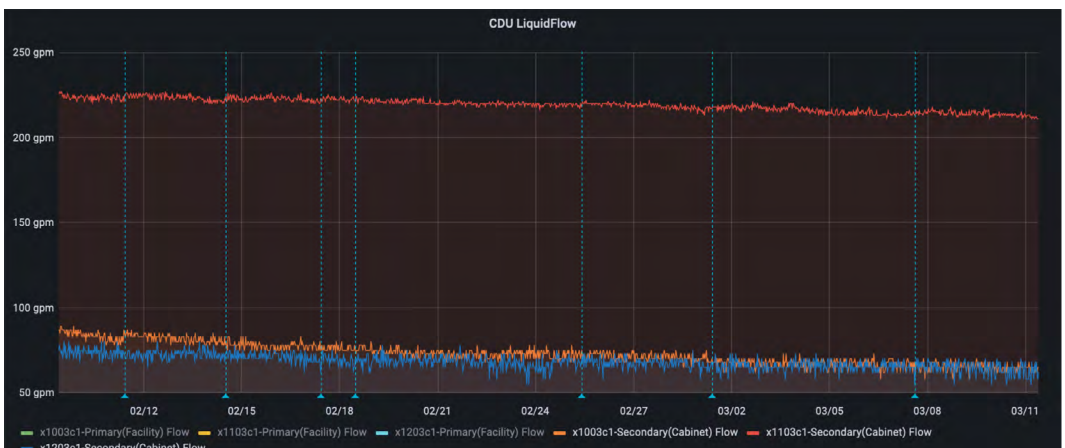


Figure 8. A NERSC Grafana dashboard visualizing coolant flow data from *Perlmutter.*

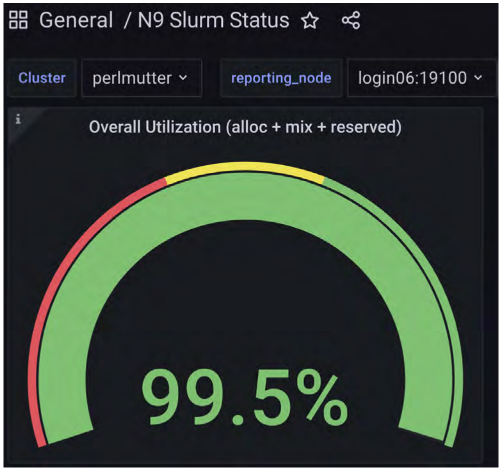


Figure 9. NERSC's Slurm status monitor provides information about node allocation, partition state, and system utilization.

Figure 10. Sample Slack message about a *Perlmutter* critical alert.

In addition to visualizations, NERSC uses the monitoring system to send automated updates over Slack to on-call staff. Like text messages, Slack messages can be accessed from a mobile device, but Slack is preferred because more information can be included in a single Slack message compared to a single text message (Figure 10), giving the staff member key information on the event to streamline corrective actions.

Superfacility Model Brings COVID Research into Real Time

Responding to the coronavirus pandemic has required a global response and the efforts of thousands of people. Among them are scientists working at NERSC and the Linac Coherent Light Source (LCLS) at SLAC National Accelerator Laboratory, who are contributing to the work through research partnerships leveraging the superfacility model for real-time data analysis.

In one project, researchers working at LCLS used X-ray crystallography to capture detailed images of the structure of the SARS-CoV-2 virus, focusing on two of the viral proteases — enzymes that make the virus's life cycle possible — and how to keep them from functioning. Identifying proteins that bind to them and interfering with their role in viral reproduction is a key step toward finding treatments for COVID-19.

The LCLS data analytics team used NERSC's Cori supercomputer to process data and provide results in real time, allowing the researchers to monitor experiments, begin analysis, and make changes nimbly. Their work generated large datasets, which traveled to NERSC via the Energy Sciences Network and which the computing resources at NERSC are well prepared to handle.

"Previously, you'd have to go home and crunch the numbers after the fact. This way, you can see any bad data while the experiment is still running and find out whether there's a difference [in results] between the proteins, so you can move on to the next sample and use your precious beam time well," said Johannes Blaschke, who worked on the project.

Collaborating with NERSC and ESnet has been invaluable, added research scientist Aaron Brewster, a project scientist in Berkeley Lab's Molecular Biophysics and Integrated Bioimaging division. "The data rates LCLS can produce can only be handled using advanced supercomputing facilities. They allow us to match the processing rate to the data production rate, letting us study small differences in atomic structure in near real time."

This is the promise of the superfacility concept: connecting facilities so they can be used by scientists seamlessly as one. As experimental data volumes continue to skyrocket, the power of supercomputers like those at NERSC becomes an increasingly crucial tool enabling scientists to spot patterns that might be difficult to see because of the size and complexity of the datasets.

"This is an emerging need, being able to move data between research sites and computing centers," said Debbie Bard, who leads NERSC's Data Science Engagement group. "This

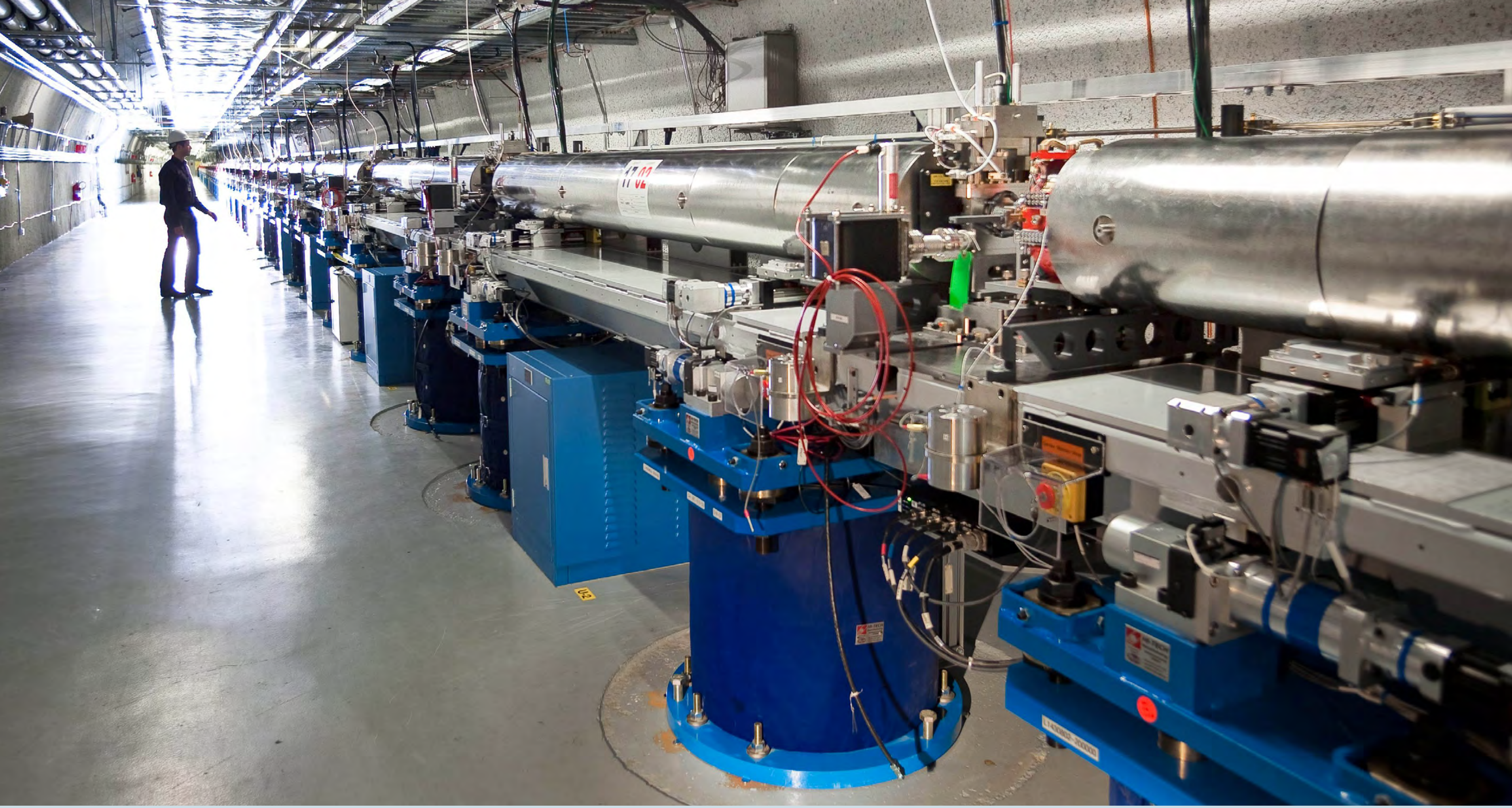


Figure 11. Researchers from Turkey working at the Linac Coherent Light Source have used X-ray crystallography to capture detailed images of the structure of the SARS-CoV-2 virus, which were then transported to NERSC in near-real time via ESnet.

particular work with LCLS is one of many experiments that need our resources. But part of the idea is that every experiment should have this capability."

Collaboration between LCLS and NERSC on COVID is forward-facing in more ways than one. Though it's just one part of the research ecosystem, Bard noted, "this work is urgent in two senses: we need to analyze the data coming from the experiment urgently, and the science of COVID research is very urgent."

Stepping into the Future with Quantum

Computing Sciences at Berkeley Lab has always been reaching into the future of what's possible in computing — and in the coming years, what's possible in computing will change dramatically as quantum computing becomes a reality. The possibilities of quantum are wide-ranging and inspiring; in the not-too-distant future, quantum information systems are expected to revolutionize how we study disease, design new materials, analyze molecular processes, solve massive data challenges, and much more.

Berkeley Lab is already partnering with industry and academia to fabricate and test quantum-based devices, develop software and algorithms, build a prototype quantum computer and network, and apply these innovations for breakthroughs in a broad range of science areas.

As part of this collaborative effort, in November 2021 NERSC accepted proposals for the QIS@*Perlmutter* program and later awarded more than 250,000 *Perlmutter* GPU node hours to 16 quantum information science (QIS) projects across the U.S. and beyond. These pioneering research efforts include simulating defects in materials for QIS, applying quantum deep learning algorithms to high-energy physics data analysis, and developing surrogate models for variational quantum algorithms, among others.

NERSC sees its role in the budding QIS field as a centralized resource for users who want to bridge the gap between classical computing and quantum computing for applications in chemistry, physics, materials science, drug discovery, and more. Many of the science problems NERSC users are currently focused on are quantum mechanical in nature; by combining classical and quantum resources, NERSC is looking to enhance and expand these research efforts both in the near term and beyond.

Innovations in Equitable Hiring

New Resources

As part of an ongoing effort, in 2021 a team of NERSC staff worked to collect and produce a trove of material to support equitable hiring practices. The outcome of this effort was a collection of curated hiring resources shared center-wide via Google Drive. The information included is aimed at hiring managers but is available and may be useful to anyone participating in the hiring process.

To develop these materials, the team met with a senior Berkeley Lab technical recruiter to learn about best practices. They surveyed several University of California-produced videos about best practices for hiring, with an emphasis on equitable hiring and reducing bias. The major themes that emerged from these efforts were:

- Hiring managers should write and advertise positions using neutral language to avoid discouraging a subset of applicants.
- To avoid bias, it is important to define key position requirements in advance.
- It is important to score candidates in a uniform way against the requirements that were determined before any evaluation began.

These documented practices serve as a living collection of information within NERSC to help improve the center's hiring practices, primarily by making our hiring processes more consistent and predictable, reducing aspects of the hiring process that are unnecessarily difficult for hiring managers, and supporting equity and diversity in hiring.

Those participating in the hiring process at NERSC now have access to:

- A README that guides readers through the hiring process with best practices, including the recommendation that all hiring committee members should take bias training, and information about UC-provided targeted bias training for hiring managers
- The official Berkeley Lab talent acquisition guide
- Candidate communications templates
- Interview and resume-scoring templates
- A curated list of interview questions, down-selected from a larger list of questions provided by hiring managers at Berkeley Lab.

In 2021, these practices were implemented for various job requisitions and feedback was generally positive, including constructive feedback that has been further incorporated into the materials.

New Policies for NESAP Engineers

Over the lifetime of the NERSC NESAP program, this program has been a significant hiring pipeline for junior staff into NERSC. However, we found that limiting this pool of potential staff to only postdocs is missing out on a larger pool of talented individuals who may not meet the relatively strict requirements for working as a postdoc at Berkeley Lab. These requirements prohibit anyone who has five years of previous postdoc experience or equivalent experience from being hired into a postdoc role. Furthermore, by definition, a postdoc requires a Ph.D., which further narrows the applicant pool.

In 2020, several NERSC staff made a push to open the NESAP hiring program to candidates without a Ph.D. The hope was that by widening the NESAP pool, the diversity of staff would expand. Toward this end, in late 2020 NERSC added a new role: NESAP engineer. This position has duties similar to those associated with current NESAP postdoc roles, but with additional consulting, software development, and support responsibilities.

In 2021, four new staff were hired into this role, including one who did not have a Ph.D.; they are working in the Application Performance Group and the Data and Analytics Services Group with code teams, on system integration, and on providing system software and infrastructure. Since the addition of this role is still relatively new, we will continue to assess this new position to determine whether it is helping NERSC reach its goal of strengthening and diversifying its workforce.

Innovations in Sustainability

Berkeley Lab has a wide range of aggressive sustainability goals set by DOE, the California state government, and the University of California. These Lab-wide goals include targets to:

- Keep the overhead cooling energy consumption of its HPC facility (NERSC) below 10% of the consumption used by the computers (<1.1 power usage effectiveness [PUE])
- Increase water consumption efficiency 36% by 2025 from a 2007 baseline
- Improve energy intensity (use effectiveness) by 2% year over year.

To help achieve these goals, the Sustainable Berkeley Lab team collaborates with partners site-wide to reduce the Lab's climate, waste, and water footprint. Through these efforts, the Lab has implemented ISO50001 energy management programs and achieved third-party certification for both energy and water performance improvements. Since NERSC is the largest single divisional user of energy at the Lab (~40% of the total energy usage), it has played an important role in this process and is a formally designated Significant Energy User (SEU) and a member of the Lab's core ISO50001 team. This process leads to operational cost savings and also aligns with the Lab's and DOE's mission and broad commitment to sustainability, stewardship, and leading by example.

Figure 12. The first phase of *Perlmutter* was installed and came on line in 2021.

While pursuing ISO 50001 certification was voluntary, managing energy and water performance aligns with the Lab's business strategy. For this reason, the Lab has implemented the ISO50001 for both energy and water performance improvement. While NERSC has achieved energy-efficiency recognition for its HPC center, the ISO50001 process focuses on continuous improvement through multidisciplinary teamwork within the Berkeley Lab pool of subject matter expertise.

As a part of the ISO50001 process, NERSC has identified a number of areas of improvement and optimization. A summary of the changes and their impacts are shown in Table 1. Some of the technical details of these improvements have been highlighted in this year's and previous Operational Assessment Reports. A key to many of these improvements is an ongoing enhancement of instrumentation and monitoring systems, working towards increased facility coverage of real-time measurement capabilities (Figure 13). Examples of these are highlighted in Figure 14.

Table 1. A Summary of Areas Where NERSC Has Improved and Optimized Its Energy Efficiency Since Becoming Involved in the ISO50001 Process

	MEASUREMENT	ENERGY SAVINGS (KWH)	WATER SAVINGS (GAL)	COST SAVINGS
1	Install Firmware to Enable ESS Mode for UPSs	350,000	140,000	$26,690
2	Implement Tower Water Supply Temperature Reset and Reduced Tower Water Pump Speed	260,000	0	$18,460
3	Reset Cooling Water Temperature Setpoint and Enable Cray Dynamic Fan Control	275,000	110,000	$20,970
4	Install New Heat Exchanger	760,000	300,000	$57,900
5	Install Bypass Valves	35,000	10,000	$2,620
6	Reset Cray Air Temperature Setpoint	82,000	30,000	$6,220
7	Balancing Adjustment to Alleviate Pump Check Valve Issues	150,052	0	$10,650
8	Adjusting AHU Balancing Dampers	140,000	0	$9,940
9	Install Booster Pump	Pending verification	Pending verification	Pending verification
	TOTAL	**2,052,052**	**590,000**	**153,450**
	Total Non-IT Energy 4,866,000 kWh			
	Savings as a % of Non-IT Energy 42%			

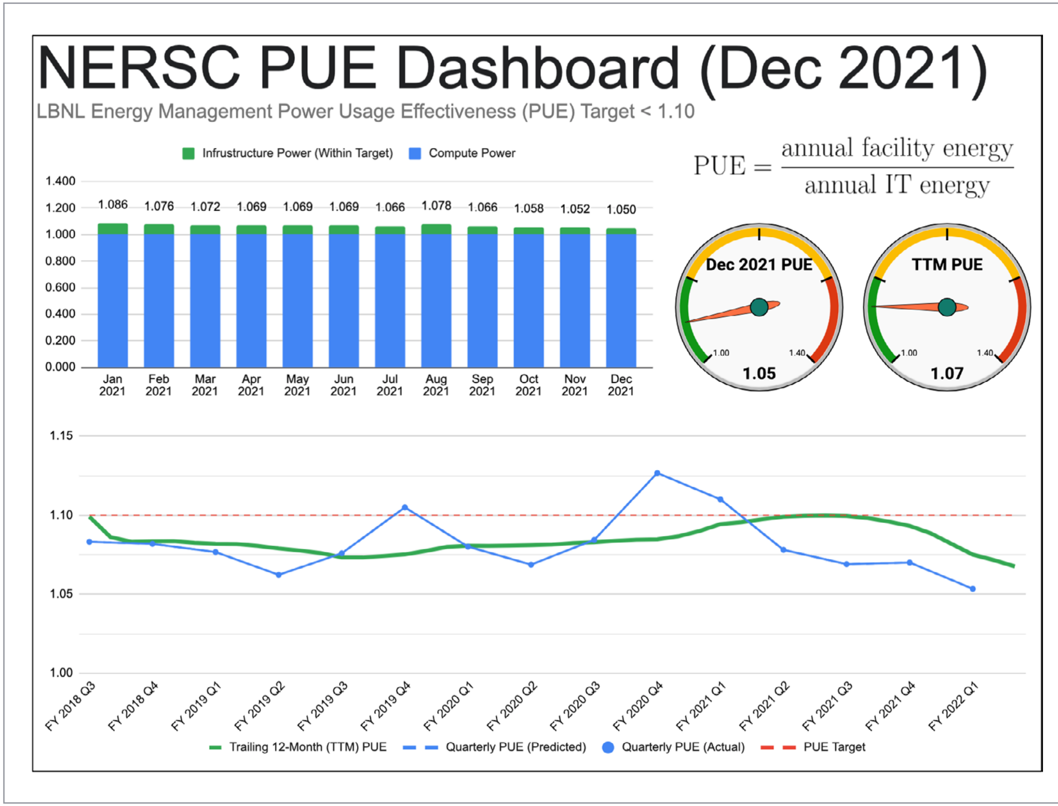


Figure 13. As part of its energy-efficiency optimization efforts, NERSC continuously enhances its instrumentation and monitoring systems, working toward increased facility coverage of real-time measurement capabilities.

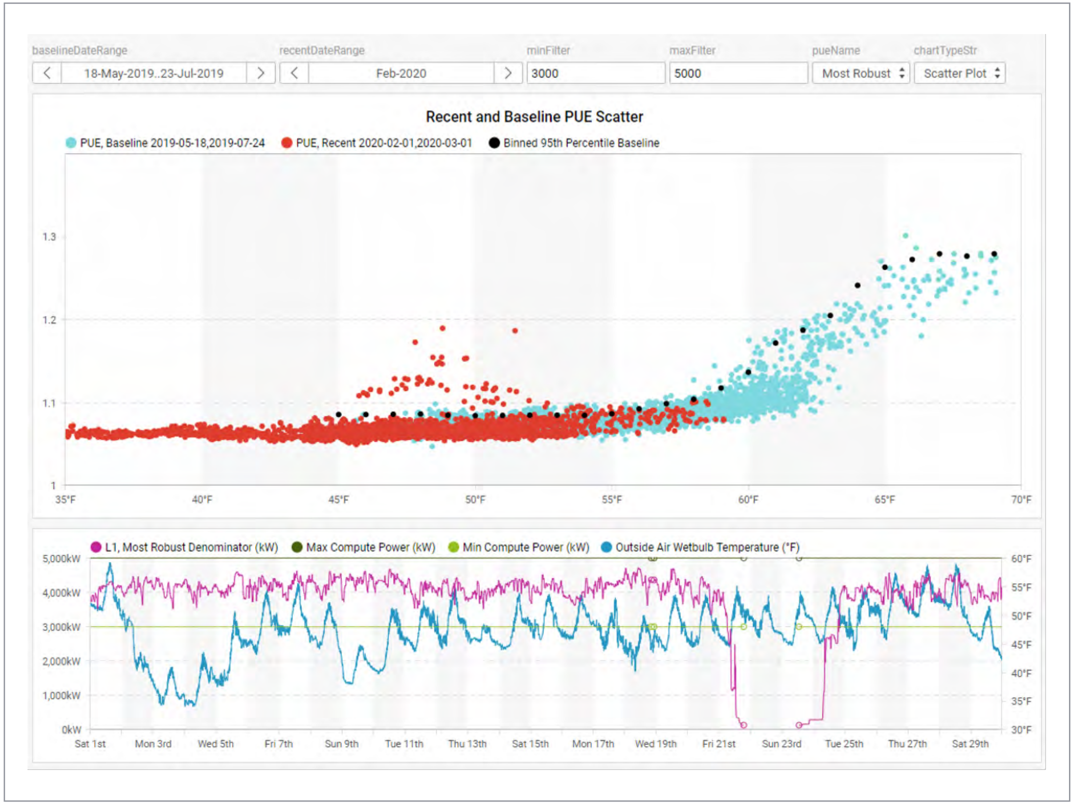


Figure 14. As part of its energy-efficiency optimization efforts, NERSC continuously enhances its instrumentation and monitoring systems, working toward increased facility coverage of real-time measurement capabilities.

Berkeley Lab Recognized for Sustainability in 2021

In 2021, the Lab received award recognition for its ISO50001 work. In September, the implementation team received a 2021 U.S. Department of Energy (DOE) Sustainability Award in the Innovative Approach to Sustainability category. The award singled out Berkeley Lab's ISO50001 as "best in class" due to its use of cross-divisional teams, application of energy management principles to incorporate water management, promoting a culture of continual improvement, and maintaining an ecosystem of online guidance to streamline implementation and ensure consistency. The award also recognized the publicly available Energy and Water Management System (EWMS), a clear resource benefit for other organizations to conceptualize and implement their own ISO 50001 programs.

In addition, the Clean Energy Ministerial (CEM), an international forum that promotes policies and programs to advance clean energy, awarded Berkeley Lab with a 2021 Energy Management Insight Award for helping to build international awareness of ISO50001 and contributing to global knowledge of energy management implementation. The award, announced in December, especially recognizes the EWMS and is described in a detailed case study on the CEM website that highlights many insights, benefits, and accomplishments as well as NERSC's specific contributions to these efforts.

Since the EWMS program began, Lab-wide infrastructure has improved the energy efficiency of its operations by 26%, saved $2.17 million in utility costs, and avoided 4,400 metric tons of greenhouse gas. The NERSC focus on savings is principally on non-compute infrastructure, which averaged 8% of compute load during 2021. During 2021, the NERSC collaboration in ISO50001 maintained the previously attained 37% of non-compute savings and increased it further to 42% using control interconnections between the Cori XC-Series blower cooling systems and the Building 59 cooling water plant systems.

Another area of significant effort during 2021 was the development of Building Management Systems (BMS) control code to transition NERSC from 24/7 cooling tower operation to staging towers on HPC cooling demand. Prior to the NERSC-9 Facility Upgrade project, NERSC had four cooling towers always in operation. With an expansion to seven cooling towers in 2020, we also deployed tower water isolation valves to cycle off towers when needed. During 2021, we developed complicated BMS code that controlled three directly interactive but independent control loops: the number of cooling tower cells in operation, cooling tower fan speeds, and tower water flow versus cooling water temperature setpoint. After many iterations of the BMS code and iterative feedback from our monitoring tools, we attained stable and energy-efficient cooling water plant operation. We are now gathering long-term performance data for further optimization in 2022.

The ongoing NERSC engagement in the EMWS Core Team and Berkeley Lab campus operations has shown that NERSC energy and water efficiency is not limited to our local systems and hardware. Programmatic and regularized coordination with other Lab divisions and multidisciplinary expertise are critical to identifying resource savings opportunities. Since the ISO 50001 process is about continuous improvements, this is arguably the best outcome of the EMWS program.

The EMWS Core Team continues to identify new operational improvements, including facility machine-learning strategies, improved cooling and outside air-sensing controls hardware, and expanded outside air-quality monitoring for increasingly regular California wildfire seasons. Another area of opportunity is to increase water-use efficiency during the coming years. The EMWS team is exploring wastewater reuse options from both Lab and UC Berkeley effluent sources. Expected water consumption from NERSC, with *Perlmutter* fully online, will be approximately 37 million gallons/year. Preliminary study estimates show a range of options delivering from 10% to as high as 70% of reduced water consumption depending on capital investment levels. Another area being explored is waste heat recovery from NERSC into a campus District Hot Water loop. This district loop would extract waste heat ahead of the NERSC cooling towers, thereby providing water savings as well as delivering building space air conditioning energy use savings across the Lab campus.

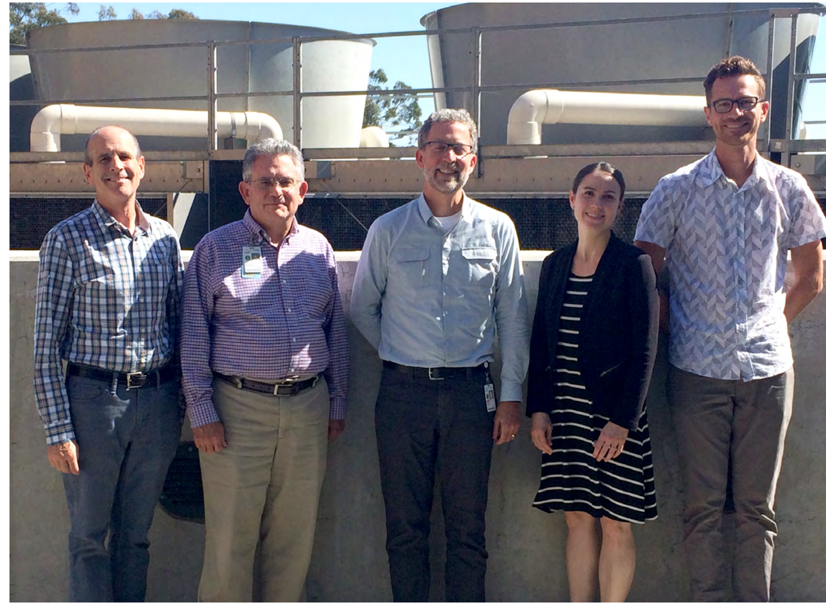


Figure 15. DOE honors NERSC efficiency optimization team (left to right): Norm Bourassa and Jeff Broughton of NERSC; John Elliott and Deidre Carter of Sustainable Berkeley Lab; and Walker Johnson of kW Engineering. Not pictured: Brent Draney, Jeff Grounds, and Ernest Jew of NERSC; Mark Friedrich, Berkeley Lab.

In Their Own Words: Summer Students Learn and Contribute

Josh Geden, a student at Duke University, participated in the 2021 Computing Sciences Summer Program as an intern in the NERSC Data & Analytics Services Group and the Computational Research Division's Data Science and Technology department. Here, Geden shares his account of working to make Jupyter, a program popular with NERSC scientists, even more user-friendly.

How I Spent My Summer: Making Jupyter Tools — and Groundbreaking Science — Even Better

Each summer, Berkeley Lab hosts dozens of college students through the Computing Sciences Summer Program, and this year I had the opportunity to work with NERSC and the Computational Research Division (CRD) on a variety of projects related to Jupyter. This effort was part of a collaboration between NERSC and CRD to enhance Jupyter tools at NERSC.

The core feature of Jupyter is the notebook, an interactive computational environment that combines text, code, data analytics, and visualization into a single document that can be easily reproduced and shared. These notebooks can be used to provide a more user-friendly way to interact with the Cori supercomputer, and are very popular among scientists who utilize NERSC resources. Recent metrics indicate that more than 30% of user interaction with Cori now goes through NERSC's Jupyter deployment.

The beauty of Jupyter, from a developer's point of view, is its modular design and browser-based user interface (UI). Frontend extensions can be developed as simple typescript objects and integration with the main Jupyter installation is handled through the included Jupyter package manager. For extensions that require a backend, developers follow the same steps to create the frontend, and then they can add a simple web application with the Tornado Python library. And because Jupyter runs in the browser, anyone with an understanding of simple HTML and CSS can develop user-facing components with ease. Jupyter also uses Bootstrap in the frontend, which allows developers to reuse common class names to maintain a cohesive UI. This helps extensions blend in and feel like a true part of the Jupyter extension.

At the beginning of my internship, I focused on developing two JupyterLab extensions. JupyterLab is the most recent notebook interface, and if you log onto https://jupyter.nersc.gov, that will be the default option for how notebooks are displayed. The first extension I developed adds a new tab to the help menu that provides links to external documentation. These links can be configured by an admin by including a simple JSON entry in a Jupyter configuration file, which will set the links for every user. The second extension I developed adds the ability to have announcements within JupyterLab. This extension will periodically poll an external API for any announcements, and if the API returns a non-empty JSON object then it will display a button on the JupyterLab status bar. Users can then click on this button to open a window with the announcement text.

The next project I focused on was something we call the JupyterHub Entrypoint Service. My first two projects were a good warmup for the Entrypoint Service, as it proved to be my largest project and the main focus of my internship. A common issue that users at NERSC were running into was ensuring that package dependencies and versioning were consistent across multiple notebooks and user profiles. This is particularly important for certain multi-user projects such as the Advanced Light Source and the Rubin Observatory Legacy Survey of Space and Time Dark Energy Science Collaboration, both of which have very specific package requirements that every user has to use for their analysis. The good news was that this problem is easily solved by using virtual package environments such as conda environments or even container images. The bad news was that there wasn't an easy way for users to launch their Jupyter notebooks in these environments through JupyterHub.

JupyterHub is a part of the Jupyter environment that enables spawning Jupyter notebooks on remote servers. When you go to https://jupyter.nersc.gov, the page where you land is the hub. At NERSC we use the hub to control which type of node on Cori a notebook gets launched or to modify the settings of that launch. My Entrypoint Service is an extension of JupyterHub that adds a registry of custom environments that users wish to launch their notebooks in. Users can view a simple UI that allows them to add custom environments such as conda environments and custom startup scripts by path, and it works with NERSC's Shifter API to provide users with a list of available Shifter images. Users can then select one of these environments to be their selected custom entry point. These settings are then available to the hub through an API.

The emphasis of my internship on programming may be a little surprising for other interns. When attending intern events and the concluding poster session, every other intern I met had research-type projects. But to support these researchers, NERSC and Berkeley Lab have to have a dedicated team of developers that are there to implement the features necessary for scientific study.

"The biggest takeaway from my internship: that science is a team sport. While I may not be as gifted at physics or chemistry, I was still able to support the researchers in these fields by developing tools that improve the usability of NERSC's supercomputing resources, which in turn makes their jobs of conducting groundbreaking research a little bit easier."

Science Highlights

NERSC's strategic planning and engagement activities have helped enable yet another highly productive year of science for thousands of users. In this section, we document the methods and practices we use to monitor our contributions to the DOE mission via science highlights and accomplishments; describe the quality of our engagements with strategic stakeholders, including other DOE programs and partner institutions; and outline how we build understanding of the HPC landscape to enable facility and DOE mission success.

Self-Supervised ML Adds Depth, Breadth & Speed to Sky Surveys (HEP)

THE SCIENCE. Researchers from Berkeley Lab developed a self-supervised learning approach that overcomes shortcomings of existing methods for extracting meaningful information from massive sky survey datasets. The method produces representations that can be used to outperform supervised learning methods trained only on labeled data. An early application of this technique was used to discover more than 1,000 previously unknown gravitational lenses in DES data.

THE IMPACT. Sky surveys like the DES, DESI, and LSST-DESC are the largest data generators in astronomy, currently imaging tens of millions to billions of galaxies over the lifetime of a single survey, making automated analysis tools a necessity. This new technique allows for efficient feature extraction, which simplifies and accelerates analysis tasks like classification of galaxies, distance estimates, similarity searches, and outlier detection.

ADDITIONAL DETAILS. The researchers used NERSC's Cori and *Perlmutter* supercomputers to demonstrate their approach. Data preprocessing and preparation was done with Cori's Intel Xeon Phi CPUs, while the bulk of training was conducted using PyTorch on the GPU partitions of Cori and *Perlmutter*. Additional analysis was conducted on Cori's Intel Haswell processors, and the full datasets and source code were made available through NERSC's public data portal.

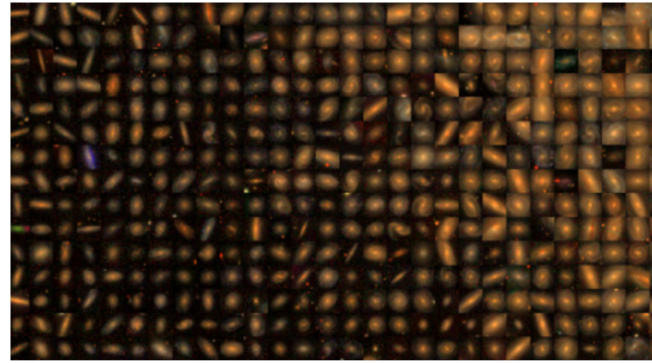


Figure 16. Sky surveys have become increasingly labor-intensive when it comes to sifting through the gathered datasets to find the most relevant information. In recent years, machine learning techniques have added a welcome advance to the process. *Image: Peter Harrington, Berkeley Lab*

NERSC PI: Zarija Lukić, Lawrence Berkeley National Laboratory

PROJECT FUNDING AND ALLOCATION AWARD: DOE Office of Science, Office of High Energy Physics (HEP)

PUBLICATION: Stein, George; Blaum, Jacqueline; Harrington, Peter; Medan, Tomislav and Lukic, Zarija, "Mining for strong gravitational lenses with self-supervised learning," arXiv:2110.00023. Hayat, Md Abul; Stein, George; Harrington, Peter; Lukic, Zarija and Mustafa, Mustafa, "Self-supervised Representation Learning for Astronomical Images"; Astrophysical Journal Letters, 911 2021 APR, 10.3847/2041-8213/abf2c7

NERSC Helps Shed New Light on Matter/Antimatter Behavior (NP)

THE SCIENCE. Scientists have produced definitive evidence for a phenomenon predicted more than 80 years ago: the creation of electron/positron pairs from the interaction of two real photons of light. The results are based on a detailed analysis of data from the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory.

THE IMPACT. The finding that pairs of electrons and positrons — particles of matter and antimatter — can be created directly from two interacting real photons had never been observed. The team's new analysis, backed by better statistics and updated theoretical calculations, shows this effect arising from photon interactions associated with close approaches of energetic gold ions produced at RHIC.

ADDITIONAL DETAILS. The results were derived from a detailed analysis of more than 6,000 pairs of electrons and positrons produced in glancing particle collisions. These measurements utilized NERSC resources to establish data corrections extracted by embedding simulated data into real data, a process that STAR runs routinely at NERSC.

NERSC PI: Jeff Porter, Lawrence Berkeley National Laboratory

PROJECT FUNDING AND ALLOCATION AWARD: DOE Office of Science, Office of Nuclear Physics (NP)

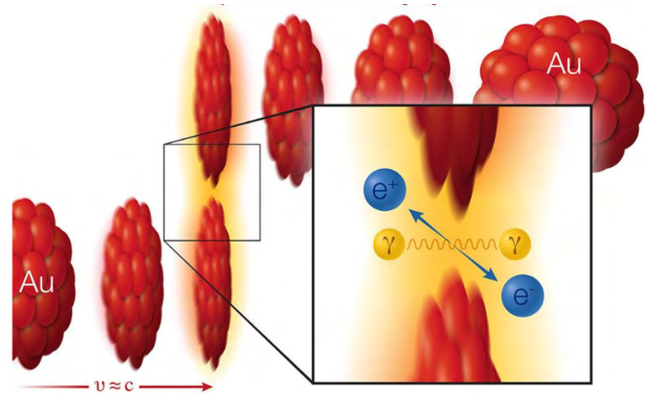


Figure 17. Making matter from light: Two gold (Au) ions (red) move in opposite directions at 99.995% of the speed of light (v, for velocity = approximately c, the speed of light). As the ions pass one another without colliding, two photons (γ) from the electromagnetic cloud surrounding the ions can interact with each other to create a matter-antimatter pair: an electron (e-) and positron (e+). *Image: Brookhaven National Laboratory*

PUBLICATIONS: Adam, J. et al; STAR Collaboration, "Measurement of e(+)e(-) Momentum and Angular Distributions from Linearly Polarized Photon Collisions"; Physical Review Letters, 127 2021 Jul 27, 10.1103/PhysRevLett.127.052302

NERSC Simulations Help Scientists Propose Source of Unexplained Solar Jets (FES)

THE SCIENCE. Using NERSC's Cori supercomputer to simulate the solar corona, an aura of plasma that surrounds the Sun, researchers found that swirls and flashes of X-ray light, together known as coronal jets, could be caused by globs of plasma emerging from the sun in ball shapes that resemble magnetic shapes known as spheromaks.

THE IMPACT. Large coronal jets contribute to outpourings of particles in the solar wind and can affect communications satellites and power grids on Earth. Recent studies have shown that the electric grid and satellites may be more vulnerable to extreme solar eruptions than previously believed, so insights into how the jets form are important to helping scientists predict their occurrence and prepare Earth for their impact.

ADDITIONAL DETAILS. The research was led by a team from the Princeton Plasma Physics Laboratory and utilized the HYM code, which was developed to study plasma stability in fusion energy reactors. The simulations indicate that a dome-shaped magnetic structure forms on the sun's surface prior to the coronal jets. When the structure detaches from the solar surface, it starts tilting. As it does so, the dome's magnetic field lines interact with the surrounding lines in a process known as magnetic reconnection, which releases the stored magnetic energy in the form of plasma jets.

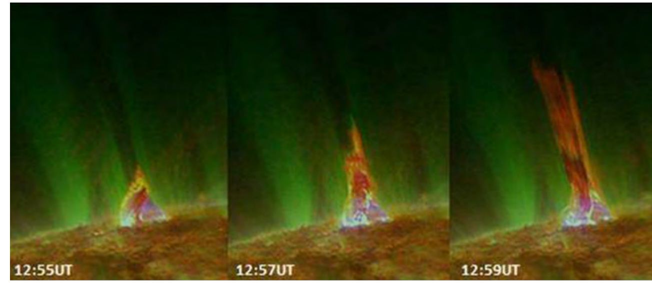


Figure 18. The eruption of a coronal jet, as captured in an extreme ultraviolet composite by the Solar Dynamics Observatory's Atmospheric Imaging Assembly. *Image: SDO/AIA/Solar Influences Data Analysis Center*

NERSC PI: Elena Belova, Princeton Plasma Physics Laboratory

PROJECT FUNDING AND ALLOCATION AWARD: DOE Office of Science, Office of Fusion Energy Science (FES)

PUBLICATION: Latham, J.; Belova, E. V.; Yamada, M.,"Numerical study of coronal plasma jet formation"; Physics of Plasmas, 28:012901; 2021 Jan 6, 10.1063/5.0025136

Simulations Run at NERSC Help Confirm a Global Climate 'Teleconnection' (BER)

THE SCIENCE. Research from Pacific Northwest National Laboratory (PNNL) describes how climate conditions in one part of the world can, over time, influence climate outcomes thousands of kilometers away. Simulations run on NERSC'S Cori supercomputer helped confirm this link. Specifically, the researchers found the Arctic region and the Western U.S. are connected by this relationship, and that regional land and sea surface warming caused by sea ice loss distantly triggers hotter and drier conditions in the American West later in the year.

THE IMPACT. Simulations run at NERSC revealed that, as Arctic sea ice melts and the surrounding land and sea surfaces warm, a vortex forms in the atmosphere above the heated area. If a vortex has already formed, as is sometimes the case, the added heat from ice loss strengthens it. This vortex, spinning counterclockwise like a cyclone, is spawned by differences in air pressure. As a result, the vortex constantly pushes the polar jet stream out of its typical pattern, diverting moist air away from the Western U.S. This, in turn, causes a second vortex, spinning clockwise, to form under the ridge of the polar jet above the Western U.S., bringing with it clear skies, dry conditions, and other fire-favorable weather. This dynamics-driven connection warms and dries out the western U.S. region, the researchers noted. By uncovering the mechanism behind that teleconnection, they hope those in charge of managing forests and preparing for wildfire will be more informed.

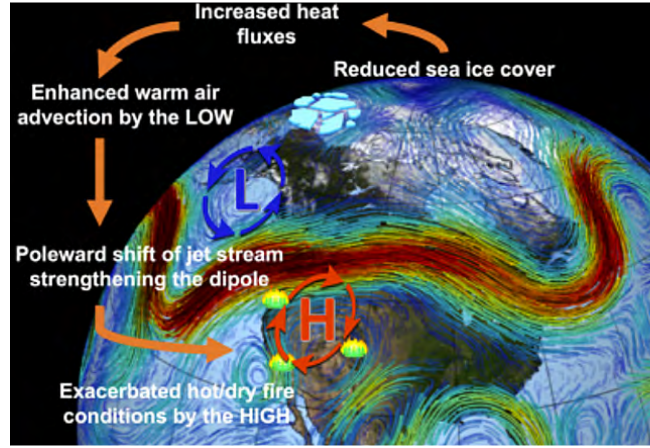


Figure 19. A schematic diagram of the teleconnection between Arctic sea-ice loss and increasing fire hazards over the western U.S. *Background image: NASA/Goddard Space Flight Center Scientific Visualization Studio*

ADDITIONAL DETAILS. The study authors used four decades of recorded sea ice data to run a series of model experiments on Cori, using about 10 million compute hours over two years. This approach utilized RESFire, a region-specific ecosystem-feedback fire model developed as part of NCAR's Community Earth System Model to simulate observed burn patterns and trends and to understand fire-related climatic processes. They also introduced some improvements to the RESFire model to optimize model parameters and settings.

In addition to the simulations and data analyses, the researchers also used NERSC as a hub for some of the CMIP6 simulations and reanalysis data and as part of a climate model comparison hackathon, giving DOE scientists and collaborators access to the data so they didn't have to duplicate or download the common datasets to their own clusters.farming — were included in the simulations. In addition to harnessing the computational power of NERSC's Cori supercomputer, the team also relied on a special allocation of 50 TB of data storage on Cori's high-speed file system.

NERSC PI: Hailong Wang, PNNL

PROJECT FUNDING AND ALLOCATION AWARD: DOE Office of Science, Office of Biological and Environmental Research (BER)

PUBLICATION: Zou, Yufei; Rasch, Philip J.; Wang, Hailong; Xie, Zuowei; Zhang, Rudong, "Increasing large wildfires over the western United States linked to diminishing sea ice in the Arctic," Nature Communications, 12:6048, Oct 26, 2021 doi.org/10.1038/s41467-021-26232-9

Though distantly spaced, the Arctic and the western United States are connected through a relationship between sea ice and wildfire weather. As Arctic sea ice melts, regional surface warming can shift atmospheric circulation patterns that ultimately bring about drier conditions in western states like Washington, California, and Oregon, according to a recent study led by PNNL researchers.

NERSC Resources Power Advances in Solar Cell Efficiency (BES)

THE SCIENCE. Research based on calculations run on NERSC's Cori supercomputer has shown that, counter to common belief, all-inorganic materials known as perovskites have the potential for greater energy generation efficiencies in solar cells compared to materials currently being used. The University of California at Santa Barbara research team's results were featured on the cover of the October 20, 2021 journal *Cell Reports Physical Science*.

THE IMPACT. Hybrid organic-inorganic perovskites have already demonstrated very high photovoltaic efficiencies of greater than 25%. The prevailing wisdom in the field is that the organic (carbon- and hydrogen-containing) molecules in the material are crucial to achieving this impressive performance because they are believed to suppress energy-losing "nonradiative recombination" events. The new research has shown that this assumption is incorrect and all-inorganic materials have the potential to outperform hybrid perovskites.

ADDITIONAL DETAILS. The researchers ran comprehensive simulations of inorganic and hybrid materials, examining how the recombination events occurred. The simulations showed that organic molecules in the hybrid perovskite can break up, which strongly decreases efficiency. Inorganic perovskites are

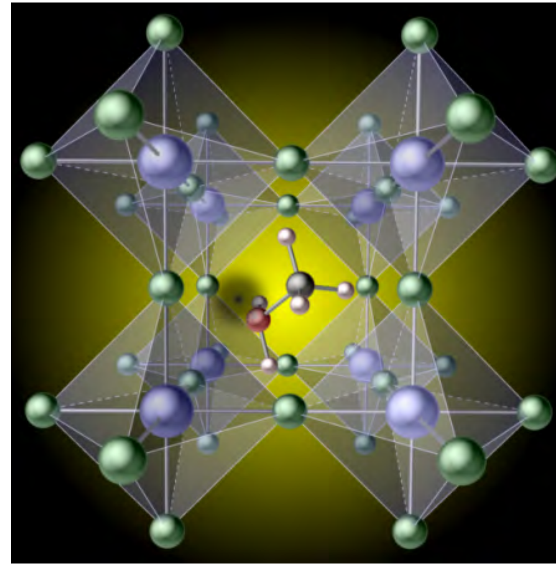


Figure 20. A hydrogen vacancy (the black spot left of center) created by removing hydrogen from a methylammonium molecule traps carriers in the prototypical hybrid perovskite, methylammonium lead iodide. *Image: UC Santa Barbara*

more difficult to create experimentally, but recent research has shown that their production is feasible.

NERSC PI: Chris G. Van de Walle, UC Santa Barbara

PROJECT FUNDING AND ALLOCATION AWARD: DOE Office of Science, Office of Basic Energy Sciences (BES)

PUBLICATION: Zhang, Xie; Turiansky, Mark E.; Van de Walle, Chris G., "All-inorganic halide perovskites as candidates for efficient solar cells"; Cell Reports Physical Science, 2:100604; 2021 Oct 12, 10.1016/j.xcrp.2021.100604

Acronyms and Abbreviations

ALS
Advanced Light Source, Lawrence Berkeley National Laboratory

AMR
Adaptive Mesh Refinement

ANL
Argonne National Laboratory

API
Application Programming Interface

ASCR
Office of Advanced Scientific Computing Research

BER
Office of Biological and Environmental Research

BES
Office of Basic Energy Sciences

BNL
Brookhaven National Laboratory

CERN
European Organization for Nuclear Research

CMB
Cosmic Microwave Background

CPU
Central Processing Unit

CSCS
Swiss National Supercomputing Centre

DESI
Dark Energy Spectroscopic Instrument

DFT
Density Functional Theory

DTN
Data Transfer Node

ECP
Exascale Computing Project

FES
Office of Fusion Energy Sciences

GB
Gigabytes

Gbps
Gigabits Per Second

GPCNet
Global Performance and Congestion Network

GPU
Graphics Processing Unit

HDF5
Hierarchical Data Format 5

HEP
Office of High Energy Physics

HPC4Mfg
High Performance Computing for Manufacturing

JGI
Joint Genome Institute

KNL
Knights Landing Processors

LANL
Los Alamos National Laboratory

LCLS
Linac Coherent Light Source

LLNL
Lawrence Livermore National Laboratory

LZ Dark Matter Experiment
LUX-Zeplin Dark Matter Experiment

MFA
Multi-Factor Authentication

MHD
Magnetohydrodynamic

NCEM
National Center for Electron Microscopy

NESAP
NERSC Exascale Scientific Application Program

NIM
NERSC Information Management

NOAA
National Oceanic and Atmospheric Administration

NP
Office of Nuclear Physics

OLCF
Oak Ridge Leadership Computing Facility

OpenMP
Open Multi-Processing

OpenMSI
Open Mass Spectrometry Imaging

PB
Petabytes

PNNL
Pacific Northwest National Laboratory

PPPL
Princeton Plasma Physics Laboratory

PUE
Power Usage Effectiveness

SENSE
Software-defined Network for End-to-End Networked Science at Exascale

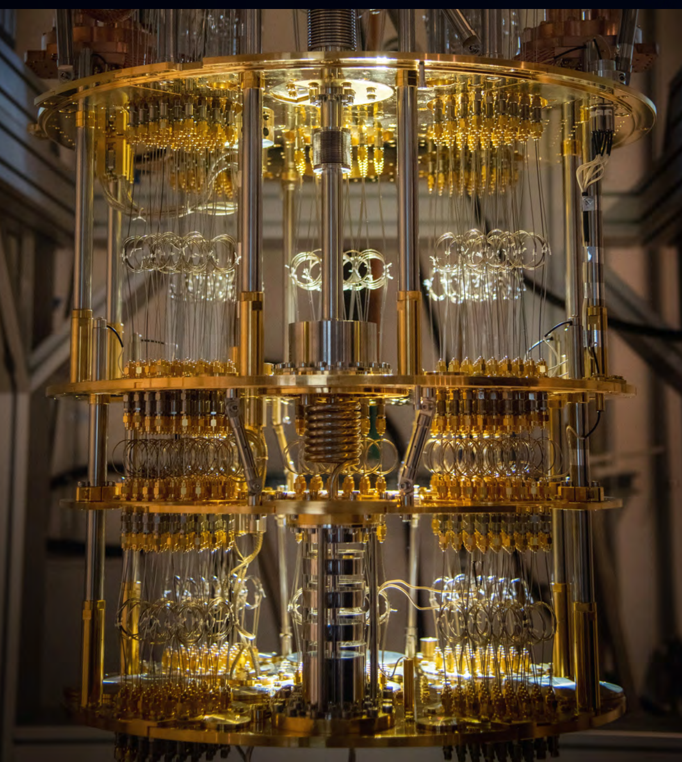
SciDAC
Scientific Discovery Through Advanced Computing

SDN
Software-defined Networking

SLURM
Simple Linux Utility for Resource Management

TAP
Trusted Access Platform

TB
Terabytes



FOR MORE INFORMATION ABOUT NERSC, CONTACT:

NERSC Communications
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720-8148
Email: CSComms@lbl.gov

NERSC's Web Site
www.nersc.gov

NERSC Annual Report Editor
Elizabeth Ball

Design
Design, layout, illustration, photography, and printing coordination:
Berkeley Lab Creative Services

Cover Image
iStock by Getty Images

22-CSAO-9045