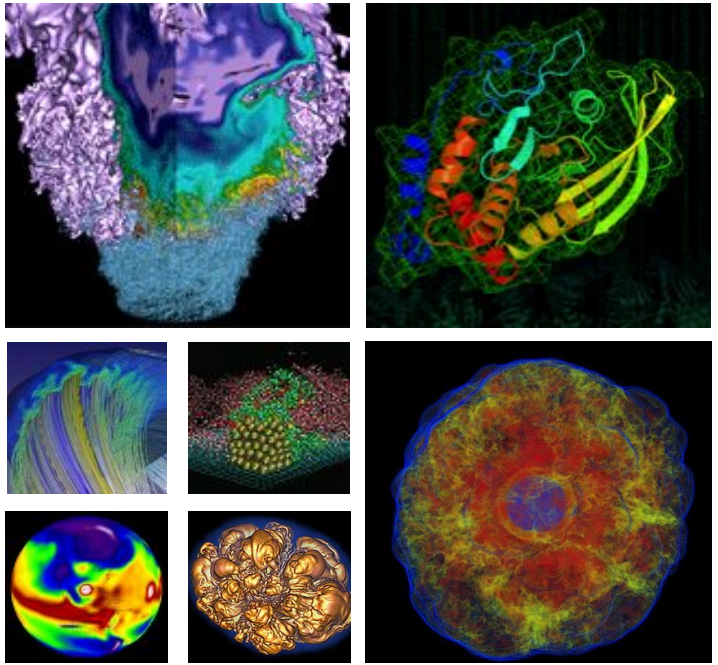
NUG Monthly Meeting



NERSC

16 March, 2023

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Today's plan

- Interactive - please participate!
  - Raise hand or just speak up
  - **NERSC User Slack** (link in chat), **#webinars** channel

- Agenda:
  - Win-of-the-month
  - Today-I-learned
  - User Community Survey
  - Announcements/CFPs
  - Topic of the day: **Science Highlights**
  - Coming meetings: topic suggestions/requests?

Win of the month

Show off an achievement, or shout out someone else's achievement, e.g.:

- Had a paper accepted
- Solved a bug
- A scientific achievement (maybe candidate for Science Highlight, or **High Impact Scientific Achievement award**)
- An **Innovative Use of High Performance Computing** (also a candidate for an award) (https://www.nersc.gov/science/nersc-hpc-achievement-awards/)

Please let us know of award-worthy work from you or your colleagues - tell us what you did, and what was the key insight?

Today I learned

What surprised you that might benefit other users to hear about?

(and might help NERSC identify documentation improvements!)

Eg:

- Something you got stuck on, hit a dead end, or turned out to be wrong about
  - Give others the benefit of your experience!
  - Opportunity to improve NERSC documentation
- A tip for using NERSC
- Something you learned that might benefit other NERSC users

"If we knew what it was we were doing, it would not be called research, would it?" - Einstein

Announcements and CFPs

Perlmutter charging holiday (See email yesterday)

- (Yesterday) we temporarily disabled a SS11 feature for performant GPU-RDMA to mitigate a critical issue leading to node failures. This is expected to substantially affect performance of applications using GPU-RDMA capabilities for inter-node communication (such as CUDA-Aware MPI or GASNet), but will allow jobs that were previously crashing to run. **We expect to be able to remove this mitigation** during the next scheduled maintenance on **March 22**, 2023.

- For the **two weeks** starting (yesterday) and ending at 11:59:59 pm (Pacific time) on March 29, **all CPU and GPU jobs on Perlmutter will run free of charge** against your allocation to encourage you to use the system. Please let us know if you see problems via a ticket at https://help.nersc.gov .

Announcements and CFPs

NERSC Appropriate Use Policy and Code of Conduct

- Many users have still not read and agreed to these!
- To read and sign it - just log into Iris.nersc.gov - if you haven't signed it yet, a dialog will pop up with it
- Users who have (had) not signed were sent an email about it this week

Announcements and CFPs

Cori Retirement end of April

- Migrating from Cori to Perlmutter - office hours March 31
- Training last week - materials and recordings are available via https://www.nersc.gov/users/training/events/migrating-from-cori-to-perlmutter-training-march2023/

Announcements and CFPs

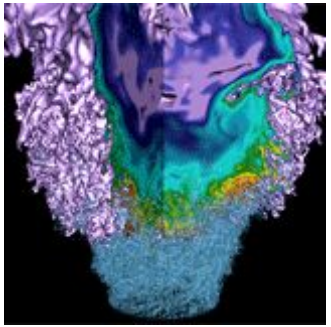
See weekly email for these upcoming events:

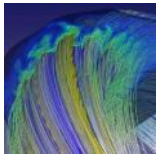
- Attention Students: NERSC Summer Internships Available!
  - https://www.nersc.gov/research-and-development/internships/
- US Research Software Engineer Association Conference - submissions open
  - Oct 16-18, Chicago IL - https://us-rse.org/usrse23/
- DOE Cross-facility Workflows Workshop April 12, 2023
- NERSC is hiring! Several open positions, see https://lbl.referrals.selectminds.com/page/nersc-careers-85

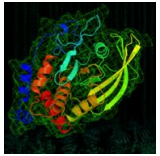
Announcements and CFPs

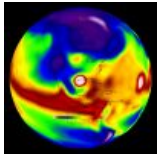
Others?

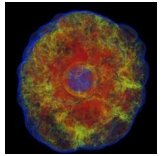
NERSC Science Highlights

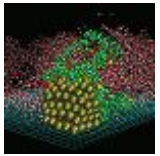






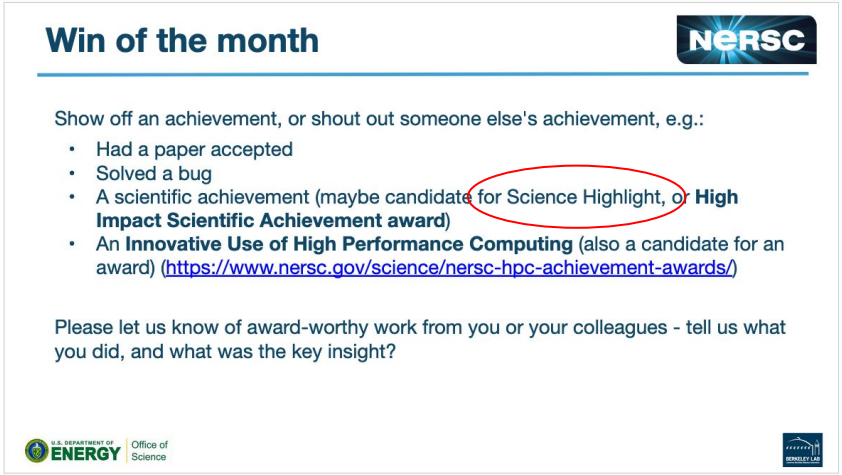






NERSC

What *are* these "Science Highlights"?



- What does one look like?
- Why does it matter?
- What sort of thing gets highlighted?
- Where to find them?
- How do I get my work into that collection?

What does one look like?

Can take the form of:

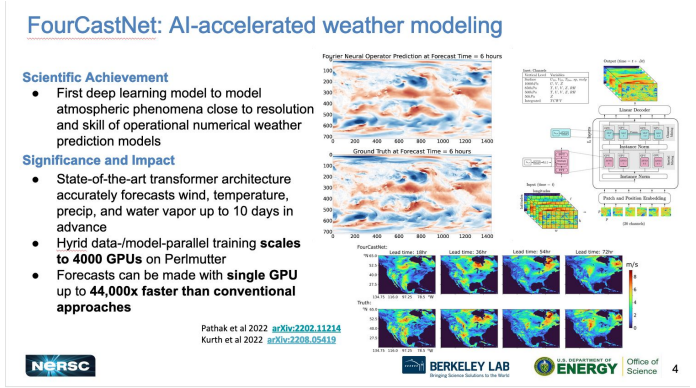
- A short summary slide
- A longer news article

Describes, in non-domain-expert terms:

- A scientific achievement
- The significance and impact of that achievement

For work that used NERSC resources

- Compute, storage or other systems
- Collaborations with NERSC staff



Why does it matter?

NERSC MISSION

The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate scientific discovery at the DOE Office of Science through high performance computing and data analysis.

It's what we're here for!

- Paying attention to the science being done helps keep focus on our mission

Presented to DOE in regular reports, and used in our annual report

- Visibility for NERSC
- Visibility for NERSC users

Some recent examples

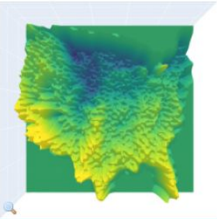
Hot off the press:

- NERSC Users in Mathematics for Experimental Data, and Earth and Environmental Science, devised an approach to Gaussian Processes that distributes a large-but-sparse covariance matrix over many Perlmutter GPU nodes, computes the locally-small submatrices on the GPUs, then reassembles the result
- Applied the method to a large climate dataset, and demonstrated an ability to accommodate the huge volumes of data collected in climate science
- Breaking the matrix down into submatrices small enough to fit in a GPU enabled a ~25x speedup over Cori
- Published in Nature Scientific Reports (March 2023)

NEW MATH METHODS AND PERLMUTTER HPC COMBINE TO DELIVER RECORD-BREAKING ML ALGORITHM

MARCH 13, 2023
By Elizabeth Ball
Contact: cscomms@lbl.gov

Using the *Perlmutter* supercomputer at the National Energy Research Scientific Computing Center (NERSC), researchers at Lawrence Berkeley National Laboratory (Berkeley Lab) have devised a new mathematical method for analyzing extremely large datasets – and, in the process, demonstrated proof of principle on a record-breaking dataset of more than five million points. Based on a mathematical framework known as Gaussian processes and harnessing the power of high performance computing, the achievement introduces a new tool for machine learning and other fields that deal with massive datasets.



The record-breaking calculation was performed on a dataset composed of daily maximum temperatures (°C) across the United States between 1990 and 2019. (Credit: Marcus Noack, Berkeley Lab)

Gaussian processes are part of a family of probabilistic problem-solving methods; they are powerful, flexible, and naturally include a quantification of the degree of uncertainty in their findings, a key element of their predictive power. Exact Gaussian processes bypass sampling and approximation techniques in favor of using complete datasets to make predictions. As a result, they are so computationally expensive and their output so cumbersome to store that they have only been used for small and medium-sized datasets.

A team led by Marcus Noack, a research scientist in the Mathematics for Experimental Data Group at Berkeley Lab, has proposed a new approach to Gaussian processes that takes advantage of datasets' natural sparsity, making them scalable up to millions of data points while remaining exact. The team's record-breaking calculation was **published in Nature Scientific** Reports in March 2023.

Discovering Sparsity

Gaussian processes require the computation, storage, and processing of a large matrix that assigns each pair of points a numerical label describing the relationship between them. However, for most modern datasets, many of those numbers are likely to be zero, meaning there is no relationship between them at all. This new algorithm allows the flexibility to identify those zero entries, ignore them, and only attend to the non-zero numbers – the ones with actual relational value. The resulting matrix of relational values is called the covariance matrix, and identifying and stripping out all zeroes makes it far more efficient to calculate, store, and work with in subsequent computations. This sparser version of the matrix also offers much-improved scalability, and no theoretical limitations to further scaling up have been discovered.

Some recent examples

Investigating energy (heat) transfer through 2 layers of 2D semiconductor materials

- Experimentally discovered that directing light at one layer increased heat dissipation >100x
- Ran ab-initio simulations on Cori to understand the mechanism
- May enable new, smaller semiconductors
- Published in Nature Nanotechnology

SHINING A LIGHT ON ELECTRONS' ROLE IN ENERGY TRANSFER AMONG 2D MATERIALS

Simulations run at NERSC yield surprising findings

MARCH 6, 2023
By Patrick Riley
Contact: cscomms@lbl.gov

Backed by supercomputing prowess at the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory (Berkeley Lab), a group of scientists examined heat and energy movements between certain two-dimensional (2D) materials, spurring discoveries that could pave the way for a new generation of transistors.

Artistic depiction of electron transfer driven by an ultrashort laser pulse, across an interface between two atomically-thin materials. This transfer is facilitated by an interlayer 'bridge' state that electrons are able to access due to lattice vibrations in both materials. (Credit: Gregory M. Stewart/SLAC)

Marrying theory and praxis, the researchers shined a light on the surprising role electrons play in the energy transfer between layers of tungsten diselenide (WSe_2) and tungsten disulfide (WS_2), a pair of 2D semiconductor materials. In their study the scientists found that even though the layers aren't tightly bound to each other, electrons provide a connection between them that dramatically speeds up heat transfer.

This class of atomically thin materials could lead to advances in new electronic devices, but one of the challenges with those materials has been that the heat dissipation is "very slow, to the point of being a bottleneck for many practical devices," said Felipe da Jornada, an assistant professor of materials science and engineering at Stanford University, a principal investigator at the SLAC National Accelerator Laboratory, and a co-author of the paper, which was recently published in Nature Nanotechnology.

However, the researchers found that by directing light at the WSe_2 and exciting electrons – as opposed to simply heating the heterostructure using a metal contact layer – they were able to transfer heat very quickly to the WS2 layer. At least two orders of magnitude faster, da Jornada said.

"It's really much faster and, of course, we were quite excited about that, for the broad application of how this sort of new mechanism can be used for managing thermal challenges in new materials and new devices," he said.

Sophisticated calculations highlight electrons' journey

From the observations made by researchers conducting the experiments – studying devices consisting of stacked monolayers of WSe_2 and WS_2 – the scientists then leaned on theoretical calculations to better understand what was happening.

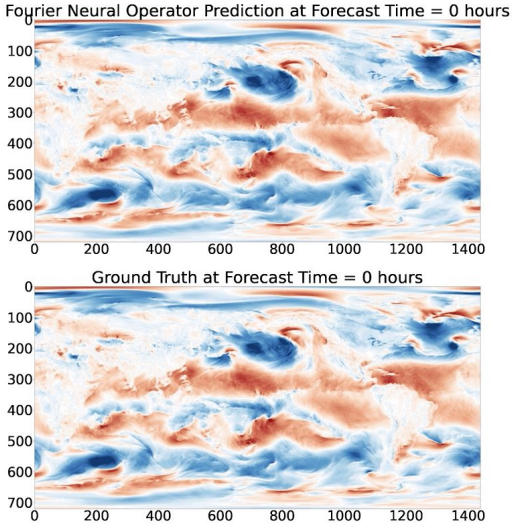
FourCastNet: AI-accelerated weather modeling

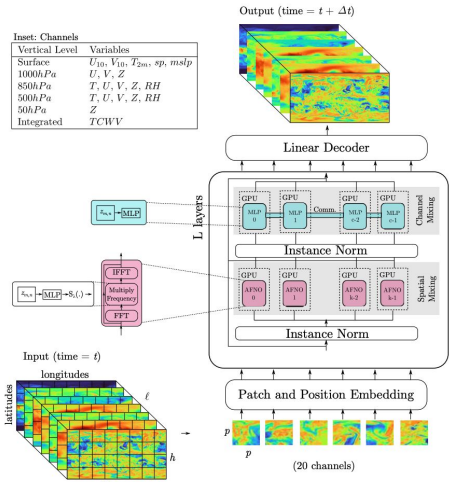
Scientific Achievement

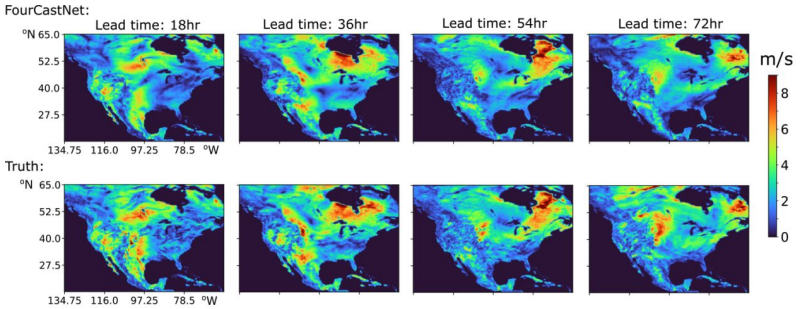
- First deep learning model to model atmospheric phenomena close to resolution and skill of operational numerical weather prediction models

Significance and Impact

- State-of-the-art transformer architecture accurately forecasts wind, temperature, precip, and water vapor up to 10 days in advance
- Hybrid data-/model-parallel training **scales to 4000 GPUs** on Perlmutter
- Forecasts can be made with **single GPU** up to **44,000x faster than conventional approaches**







Pathak et al 2022 arXiv:2202.11214
Kurth et al 2022 arXiv:2208.05419

Some recent examples

Plasma accelerators are a promising development for shrinking the size and cost of particle accelerators.

- "WarpX enables computational explorations of key physics questions in the transport and acceleration of particle beams in long chains of plasma channels, which could yield significant savings in the design and characterization of plasma-based colliders before they are built"
- ECP Project, built on AMReX, and Adaptive Mesh Refinement library
- Ran successfully at scale on Frontier, Fugaku, Summit, and Perlmutter - which earned it the 2022 Gordon Bell Prize

WARPX CODE SHINES AT THE EXASCALE LEVEL

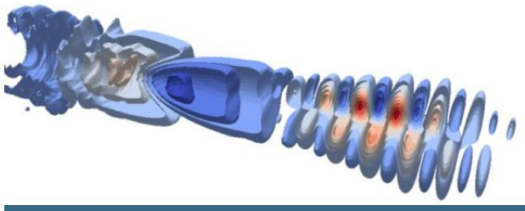
ECP-funded project takes plasma accelerator modeling to new heights

FEBRUARY 2, 2023
By Kathy Kincade
Contact: cscomms@lbl.gov

This article originally appeared on the Exascale Computing Project website.

Long valued for their role in scientific discovery and in a variety of medical and industrial applications, particle accelerators have been used in many areas of fundamental research and credited with enabling Nobel Prize–winning research in physics and chemistry. But these high-end instruments also occupy a lot of space and carry hefty price tags. Even smaller accelerators, such as those used in medical centers for proton therapy, need large spaces to accommodate their hardware, power supplies, and radiation shielding.

Fortunately, over the last several years physicists, engineers, and computational scientists have been working to create more affordable and accessible particle accelerators by shrinking both the size and the cost while increasing the capability. One of the most exciting developments in these efforts is the plasma accelerator, which uses lasers or particle beams rather than radio-frequency waves to generate the accelerating field, allowing these devices to support accelerating-electric fields many orders of magnitude greater than conventional accelerators with a much smaller footprint – even able to fit on a tabletop.



WarpX: longitudinal electric field in a laser-plasma accelerator rendered with the ECP software libraries Ascent & VTK-m as the simulation was running. Credit: Axel Huebl (Berkeley Lab)

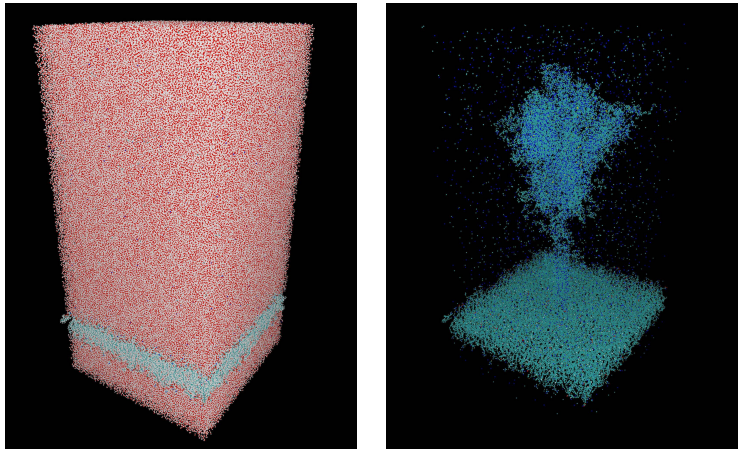
The reduced size of plasma accelerators, however, presents challenges in controlling the intricate ultrafast processes at play, which are often on the picosecond and micrometer scale. Thus, realizing their compact designs requires novel mathematical and software capabilities to enable high-performance, high-fidelity modeling that can capture the full complexity of acceleration processes over a large range of space and timescales – simulations that are often computationally intensive. To address this issue, the WarpX project – a Lawrence Berkeley National Laboratory (Berkeley Lab)-led effort that also drew the attention and support of DOE's Exascale Computing Project (ECP) – has spent the last six years creating a novel, highly parallel and highly optimized

First Exascale *Ab-Initio* MD Calculation

Breaking the Exaop Barrier for the Electronic Structure Problem in *Ab-Initio* Molecular Dynamics

(mixed FP16/FP32)

- Team showed that for certain problems, mixed precision arithmetic can achieve scientifically valid results in electronic structure calculations.
- The non-orthogonal local submatrix method applied to electronic-structure based molecular dynamics simulations exceeds 1.1 EOP/s in FP16/FP32 mixed floating-point arithmetic
- Used 4,400 NVIDIA A100 GPUs on Perlmutter
- The method achieves a sustained fraction of peak performance of about 80%.
- Example calculations are performed for SARS-CoV-2 spike proteins with up to 83 million atoms.



SARS-CoV-2 spike protein in aqueous solution: full cell (left) and without hydrogen and oxygen atoms (right).

Robert Schade, Tobias Kenter, Hossam Elgabarty, Michael Lass, Thomas D. Kühne, Christian Plessl, Paderborn University

arXiv:2205.12182v1 24 May 2022

Some recent examples

NERSC Users simulating the SARS-CoV-2 spike protein used a submatrix method to distribute a large sparse-matrix problem as a parallel set of small dense-matrix problems.

And demonstrated using mixed-precision methods (FP16 for some parts of the problem, FP32 for other parts), can give valid results alongside the ability to use the tensor cores in Perlmutter's GPUs.

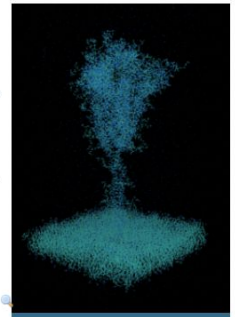
With these methods, they achieved ExaOps (10^{18} operations/second) across 4400 GPUs (1100 nodes) on Perlmutter

Home » News » News » Science News » Materials Science Simulation Achieves Extreme Performance at NERSC

MATERIALS SCIENCE SIMULATION ACHIEVES EXTREME PERFORMANCE AT NERSC

SEPTEMBER 7, 2022
By Elizabeth Ball
Contact: cscomms@lbl.gov

Using the new *Perlmutter* system at the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory (Berkeley Lab), a team of researchers led by Paderborn University scientists Thomas D. Kühne and Christian Plessl used a new mixed-precision method to conduct the first electronic structure simulation that executed more than quintillion (10^{18}) operations per second (exaops). The team's mixed-precision method is well-suited to running on Perlmutter's thousands of GPU processors.



Simulation graphic of the COVID-19 spike protein simulated in aqueous solution, with the hydrogen and oxygen atoms removed.

Of the quintillion-operations milestone, Plessl said: "The dimension of this number becomes clearer when you consider that the universe is about 10^{18} seconds old. That means that if a human had performed a calculation every second since the time of the Big Bang, this calculation does the same work in a single second."

Scientific simulations typically use "64-bit" arithmetic to achieve the high-precision results needed to represent physical systems and processes. The Paderborn team was able to show that some real-world problems of interest can use lower-precision arithmetic for some operations using their new method, a method that takes great advantage of the "tensor" cores on Perlmutter's NVIDIA A100 GPU accelerators.

The calculation used 4,400 GPUs on Perlmutter to perform a simulation of the SARS-CoV-2 spike protein. Kühne and Plessl used the submatrix method they introduced in 2020 for the approximate calculations. In this method, complex chemical calculations are broken down into independent pieces performed on small dense matrices. Because it uses many nodes working on smaller problems at once — what computing scientists call parallelism — the method lends itself to efficiency and scaling up and down for differently sized uses.

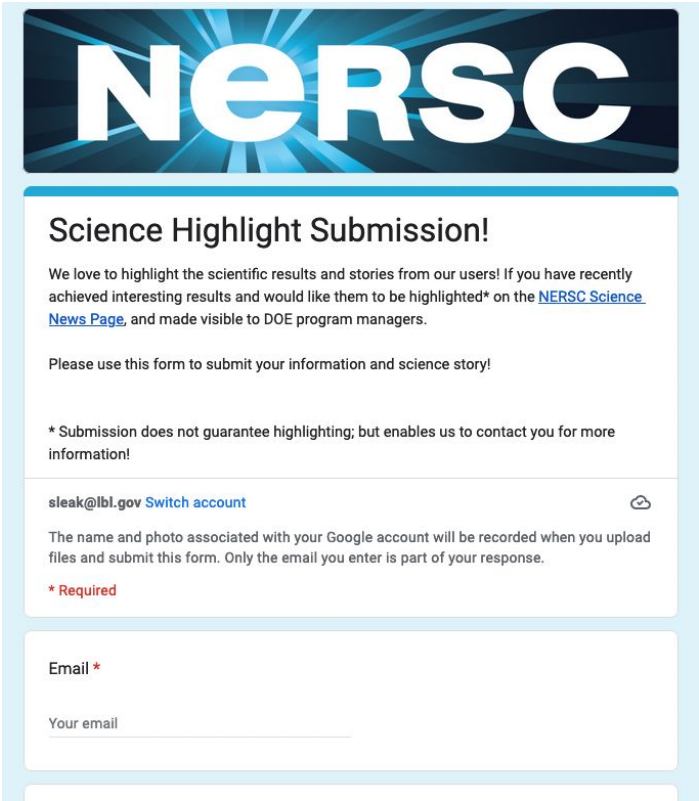
"What's neat about it is that it's a method that's inherently extremely parallel, so it's extremely scalable," said Plessl. "And that's the reason we're able to target the largest supercomputers in the world using this method. The other benefit of the method is that it's very suitable for GPUs because it kind of converts a problem that is a sparse-matrix problem that is hard to solve on a CPU to a very parallel implementation where you can work on much smaller dense matrices. From a computer science perspective, I think it's quite exciting."

Where to find Science Highlights

https://www.nersc.gov/science/science-news/

How to get my work written up here?

https://docs.google.com/forms/d/e/1FAIpQLScP4bRCtcde43nqUx4Z_sz780G9HsXtpecQ_qlPKvGafDVVKQ/viewform



Coming up

Upcoming topics:

- April: Julia at NERSC
- May: JupyterHub

We'd love to hear more lightning talks **from NERSC users** about the research you use NERSC for!

Nominate a topic at https://forms.gle/WjYx7zV7SAz2CaYz7

Thank You