The Confluence of HPC and Experimental and Observational Science

NERSC

Sudip Dosanjh
NERSC Director

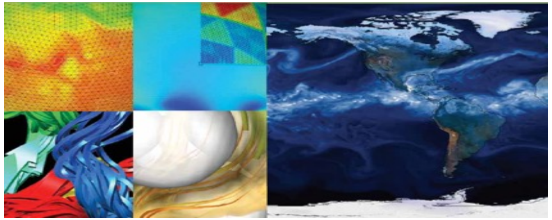
CLSAC
October 6, 2022

Acknowledgements: Debbie Bard, Jack Deslippe, Katie Antypas, Wahid Bhimji, et al.

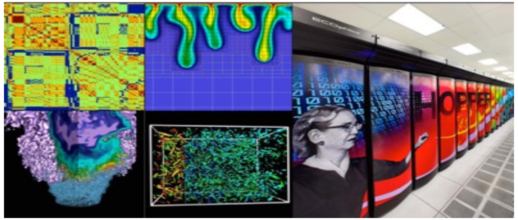
NERSC: Mission HPC for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

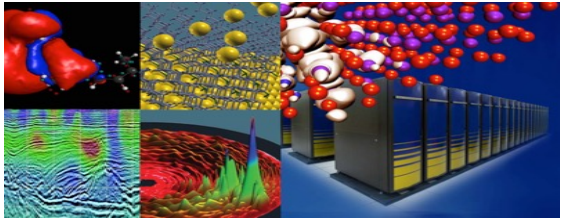
Largest funder of physical science research in the U.S.



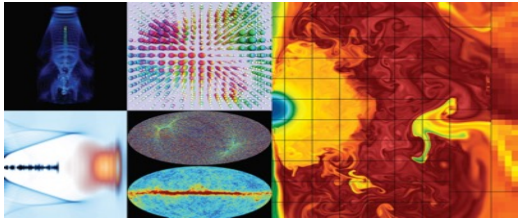
Biological and Environmental Research



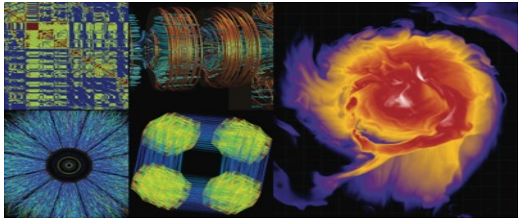
Computing



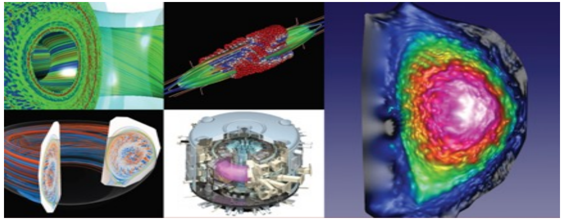
Basic Energy Sciences



High Energy Physics

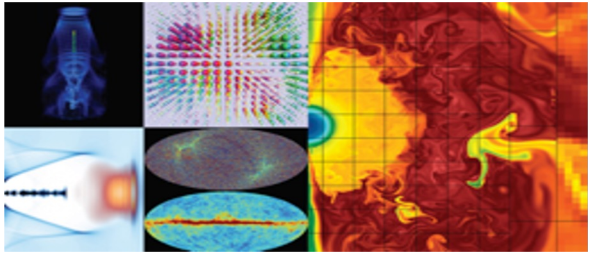


Nuclear Physics

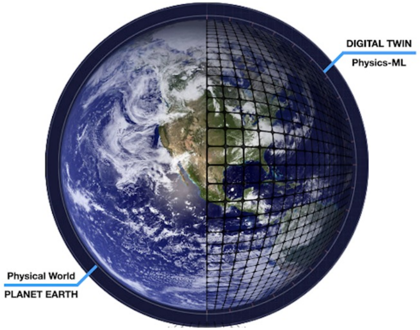


Fusion Energy, Plasma Physics

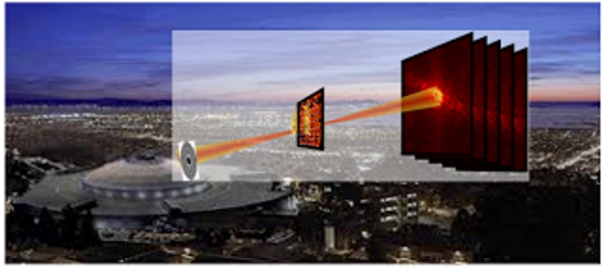
We accelerate scientific discovery for thousands of Office of Science users with 3 advanced and novel capability thrusts:



Large scale applications for simulation, modeling and data analysis

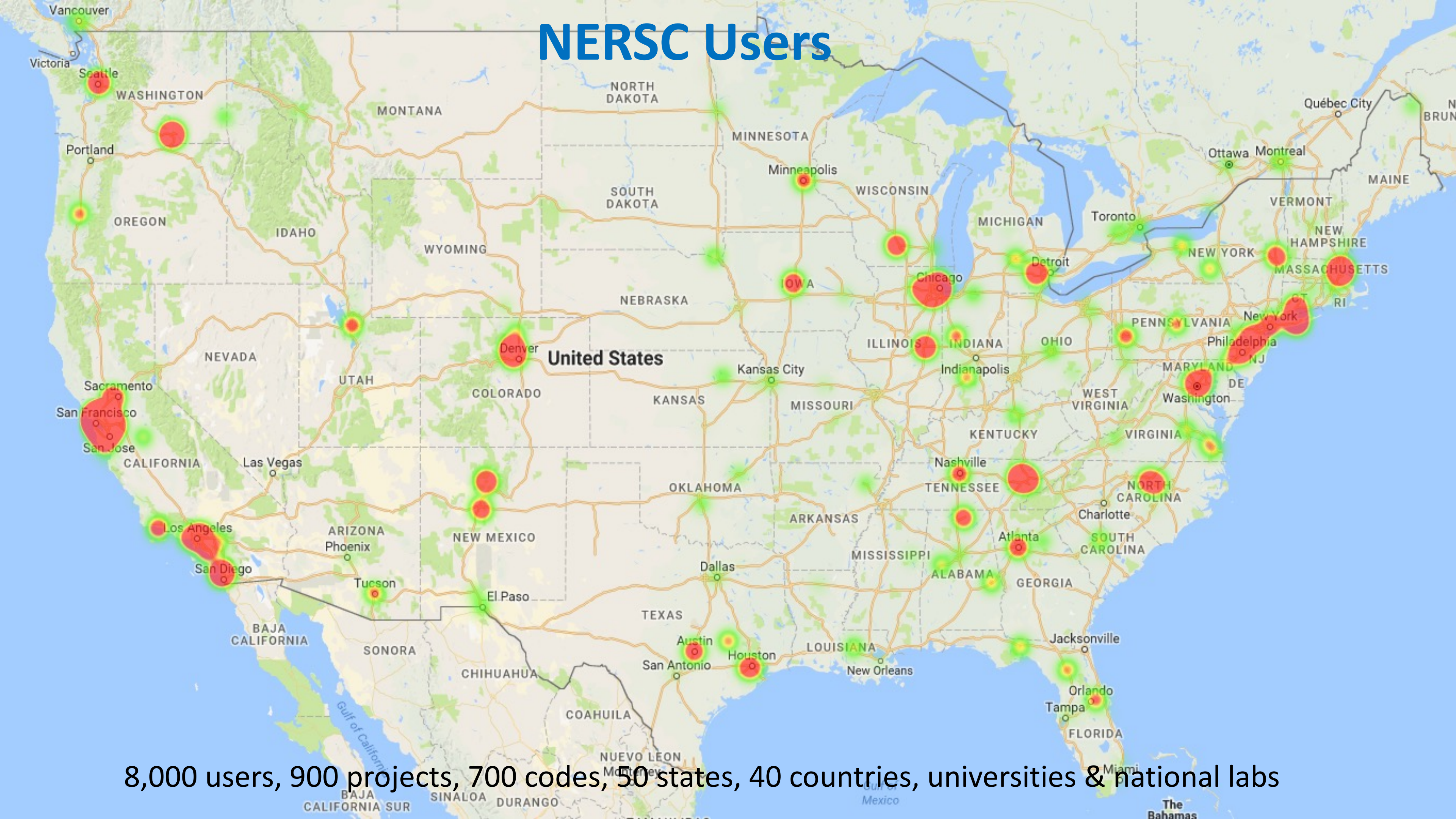


Complex experimental and AI driven workflows



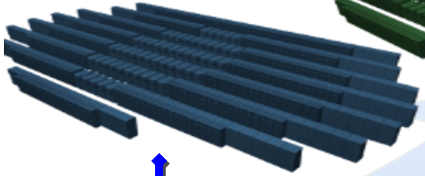
Urgent and interactive computing

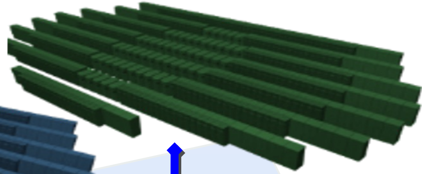
NERSC Users



8,000 users, 900 projects, 700 codes, 50 states, 40 countries, universities & national labs

NERSC Systems Roadmap





2013

NERSC-7:
Edison
Multicore CPU

2016

NERSC-8: Cori
Manycore CPU
NESAP Launched:
transition applications to advanced architectures

2020

NERSC-9:
CPU and GPU nodes
Continued transition of applications and support for complex workflows

2026

NERSC-10:
Exa system

2030+

NERSC-11:
Beyond Moore

NESAP Applications Cover the Broad Workload

Electronic Structure	
Quantum ESPRESSO	BES
NWChemEX	BES
VASP	BES
MFDn	NP
WEST	BES
BerkeleyGW	BES

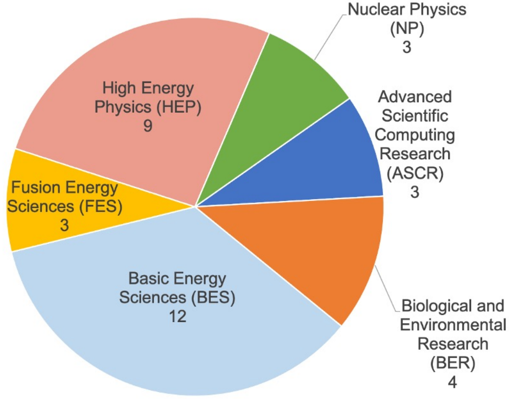
Molecular Dynamics	
EXAALT	FES, NP, BES
NAMD	BES, BER

Data	
DESI	HEP
TomoPy	BES
ATLAS	HEP
ExaFel	BES, ECP
CMS	HEP
ExaBiome	BER, ECP
TOAST	HEP
JGI WorkFlows	BER
LZ	HEP

Learning	
ExaRL	BES
HEP Accel ML	HEP
Catalyst ML	BES
Extreme Spatio-Temporal ML	ASCR
FlowGAN	ASCR

Particles & Grids	
ASGarD	FES, ASCR
WarpX	HEP, ECP
ImSim	HEP
ChomboCrunch	BES, ECP
E3SM	BER, ECP
WDMAPP	FES, ECP

LQCD	
LQCD Consortium	HEP, NP



Tier 1 NESAP Teams

+29 Tier 2 NESAP teams

58 Total NESAP Teams

Broad impact and enablement

Perlmutter Supports all Viable GPU Programming models and languages

SYCL

OpenMP

OpenACC

Fortran

HIP

kokkos

Community Codes

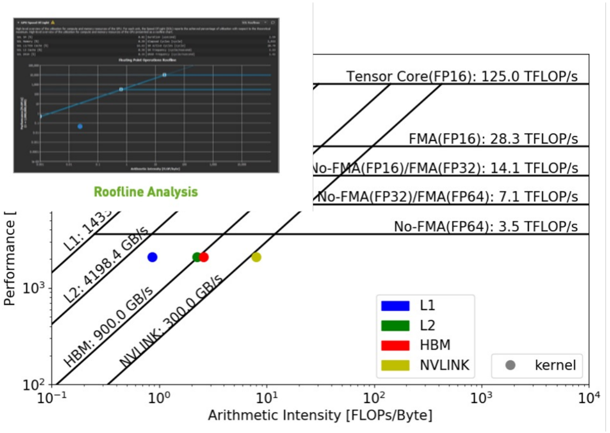
Community Resources

NERSC Documentation

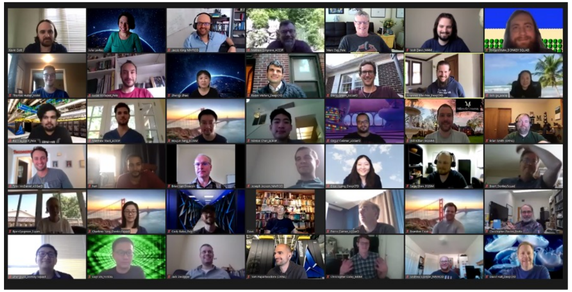
NERSC TRAINING EVENTS

SC20

Vendor tools



gpuhackathons.org

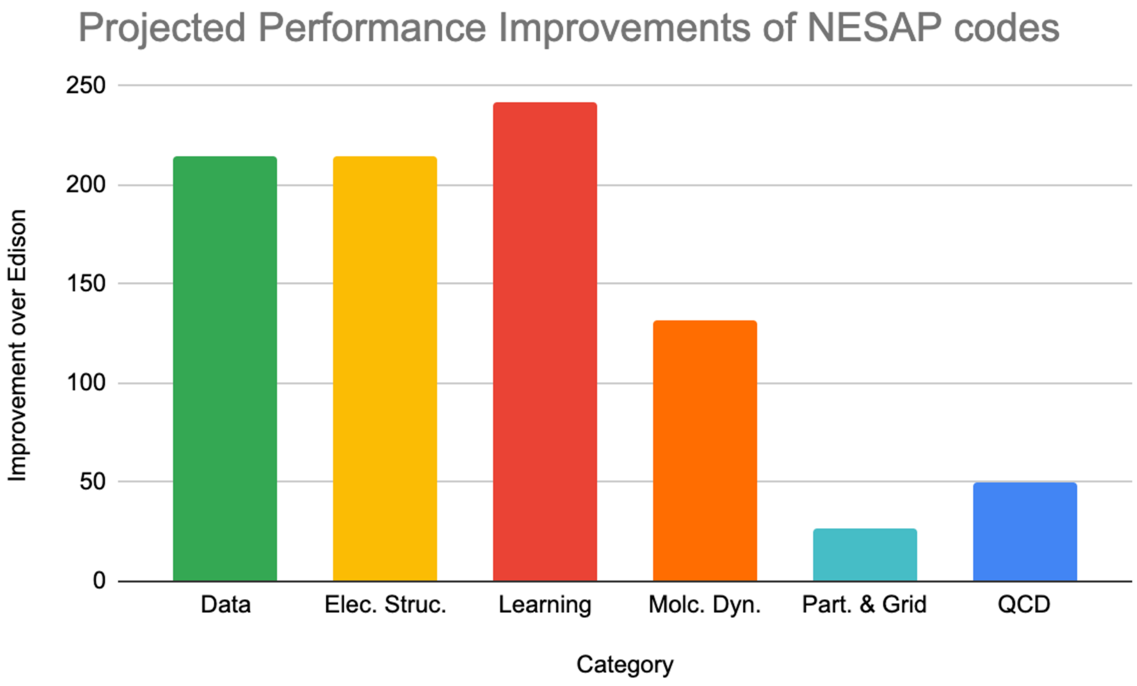


NESAP has helped make Perlmutter a high-performing and productive system

For almost 2 years, the NESAP effort has worked with a wide range of code teams and helped prepare them to use the Perlmutter system effectively

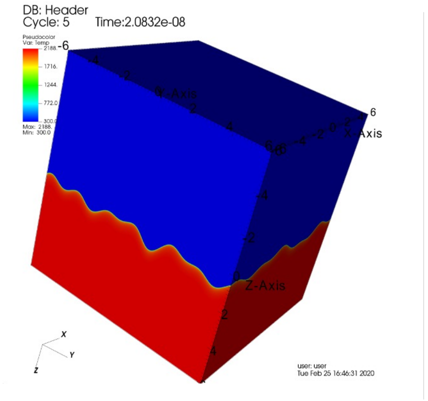
- 56 Applications targeted
- Six broad categories
  - Covers Data, Simulation and Learning applications
  - Fully representative of the current and future workload at NERSC
- Multiple levels of engagement with post-docs, hack-a-thons, training
- Excellent improvements in every category

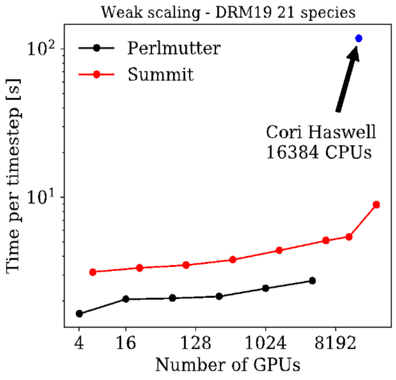
Continuing to work with the code teams in person and remotely as Perlmutter is put into production

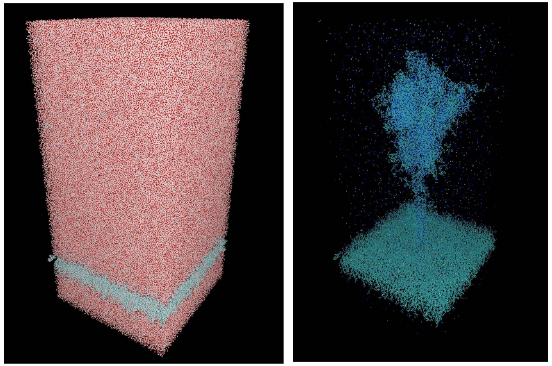


Perlmutter Early Science

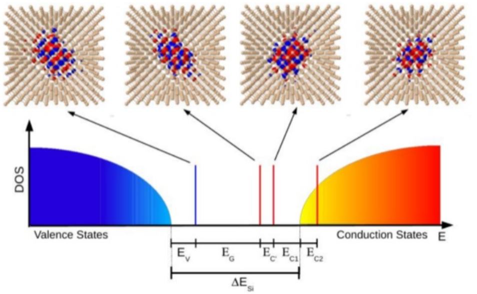
Combustion Modelling (Pele). Methane flame simulations.



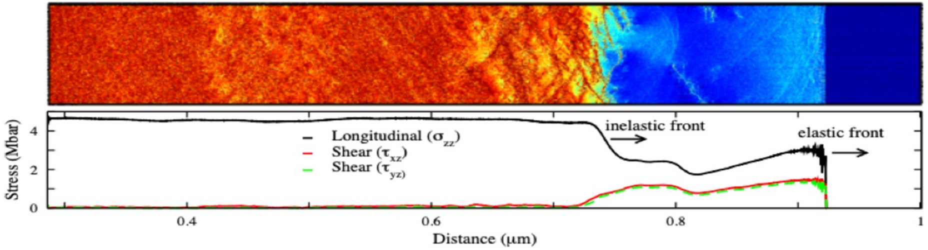




Exascale (mixed-precision) Ab-Initio MD simulations of SARS-CoV-2 spike protein in aqueous solution: full cell (left) and without hydrogen and oxygen atoms.



QuBit Design:
First Principle Excited State Calculations on localized electronic states near complex material defects



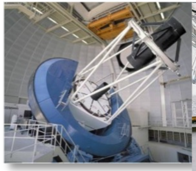
Gordon Bell Finalist: 20 Billion atom simulation of shock wave propagation in diamond in astrophysical environments.

Early Science Teams

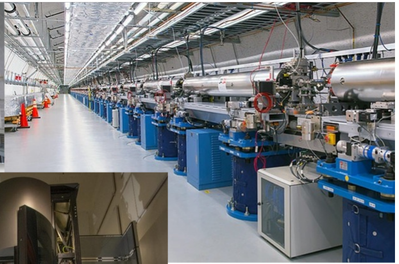
Over 230 Projects Have Run on
Perlmutter During Early Science

A changing computing landscape challenges us to think differently about supporting the Office of Science workload

Growth of experimental and observational data and the need for interactive feedback through real-time data analysis and simulation and modeling



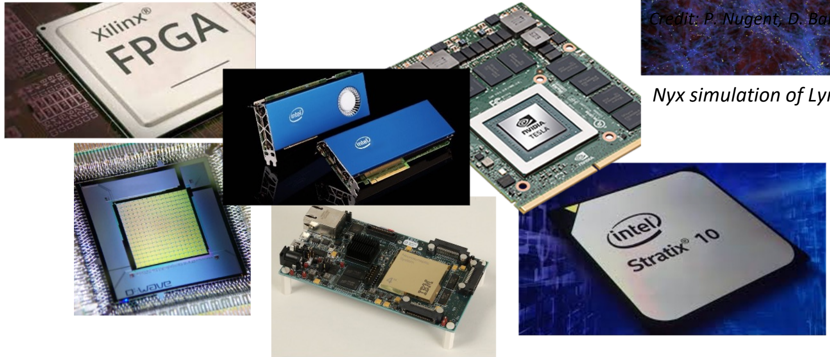
DESI



LCLS-II

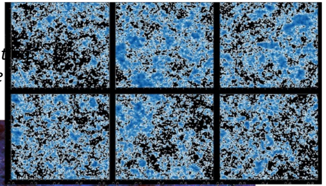
NCEM

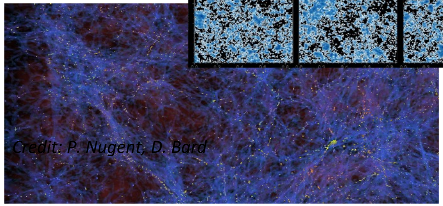
The proliferation of accelerators and new technologies



Use of advanced data analytics and AI in simulations as well as for integration of multimodal data sets

GANs used to construct surrogate models for creating weak lensing convergence maps



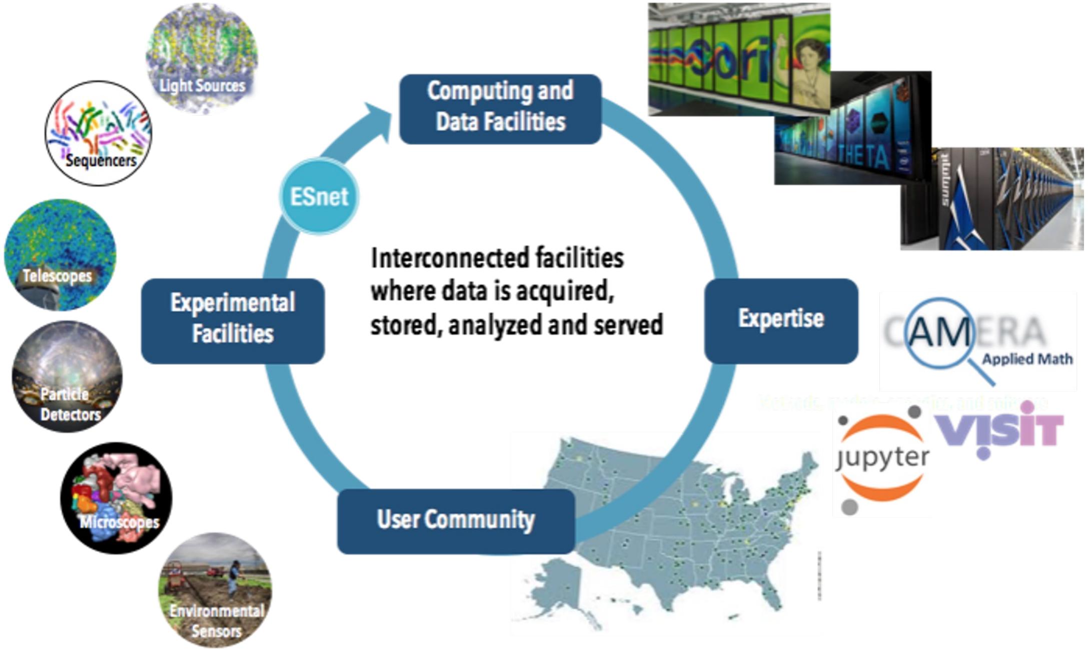


Nyx simulation of Lyman alpha forest

The Superfacility Model: an ecosystem of connected facilities, software and expertise to enable new modes of discovery

Superfacility: Computing Facilities, ESnet and research working together to support experimental science

- A model to integrate experimental, computational and networking facilities for reproducible science
- Enabling new discoveries by coupling experimental science with large scale data analysis and simulations

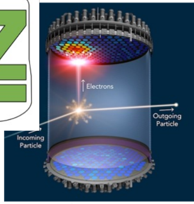


The Superfacility ‘project’ coordinated work to demonstrate automated pipelines

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

ALS

LZ



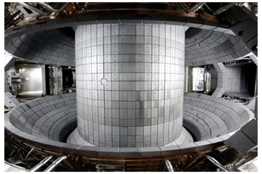
FICUS
Facilities Integrating Collaborations for User Science
JGI-EMSL Collaborative Science Call

LSST DESC
Dark Energy Science Collaboration

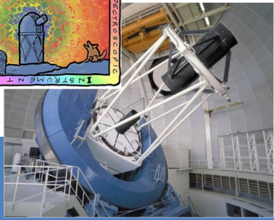
MOLECULAR FOUNDRY

LCLS

KFE



DARK ENERGY SPECTROSCOPIC INSTRUMENT



DESI uses NERSC for nightly data processing

near Tucson, AZ

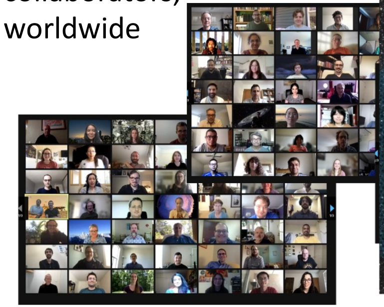
raw data (realtime)

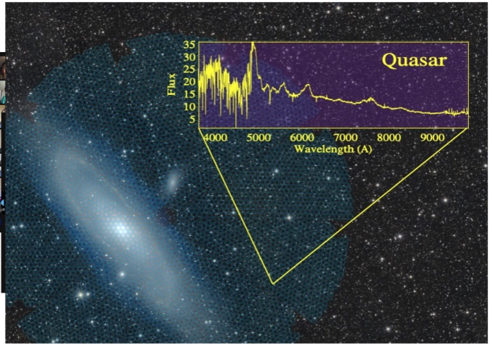
updated survey plan (daily)

realtime queue during night; regular queue for big reruns

Direct login, Jupyter, Spin, Globus

hundreds of collaborators, worldwide



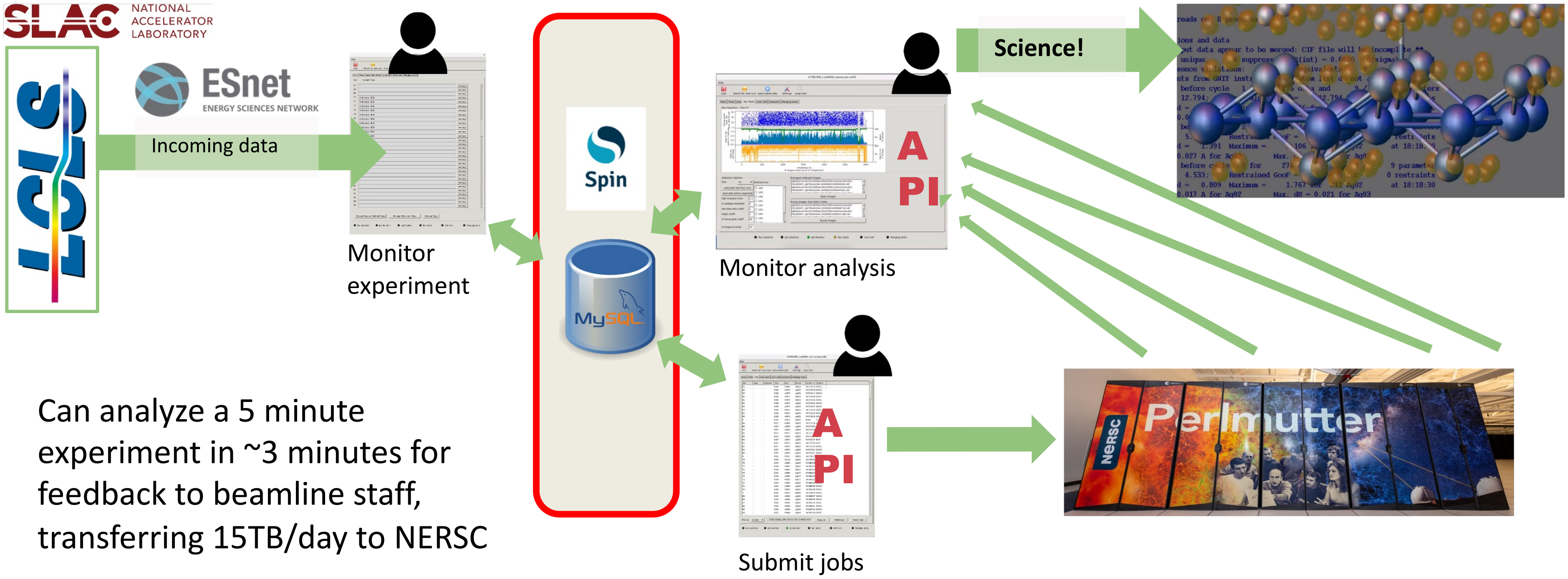


- NESAP for code optimization:
  - 2.5x improvement in per-node throughput using Perlmutter A100 compared to Cori V100 GPU (x25 compared to Edison).
- Realtime/advanced scheduling for nightly data processing
  - need to process up to 100 GB/night before breakfast to guide telescope operations
- Spin used to monitor data quality and analysis

Biggest remaining challenge: Robustness / Resilience, especially “soft” outages, e.g. transient I/O or slurm failures

Maximizing science is different than maximizing FLOPS or CPU-hours delivered

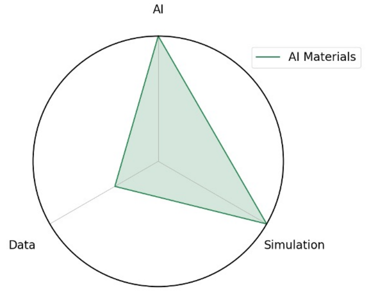
LCLS is using NERSC for realtime collaborative distributed data analysis

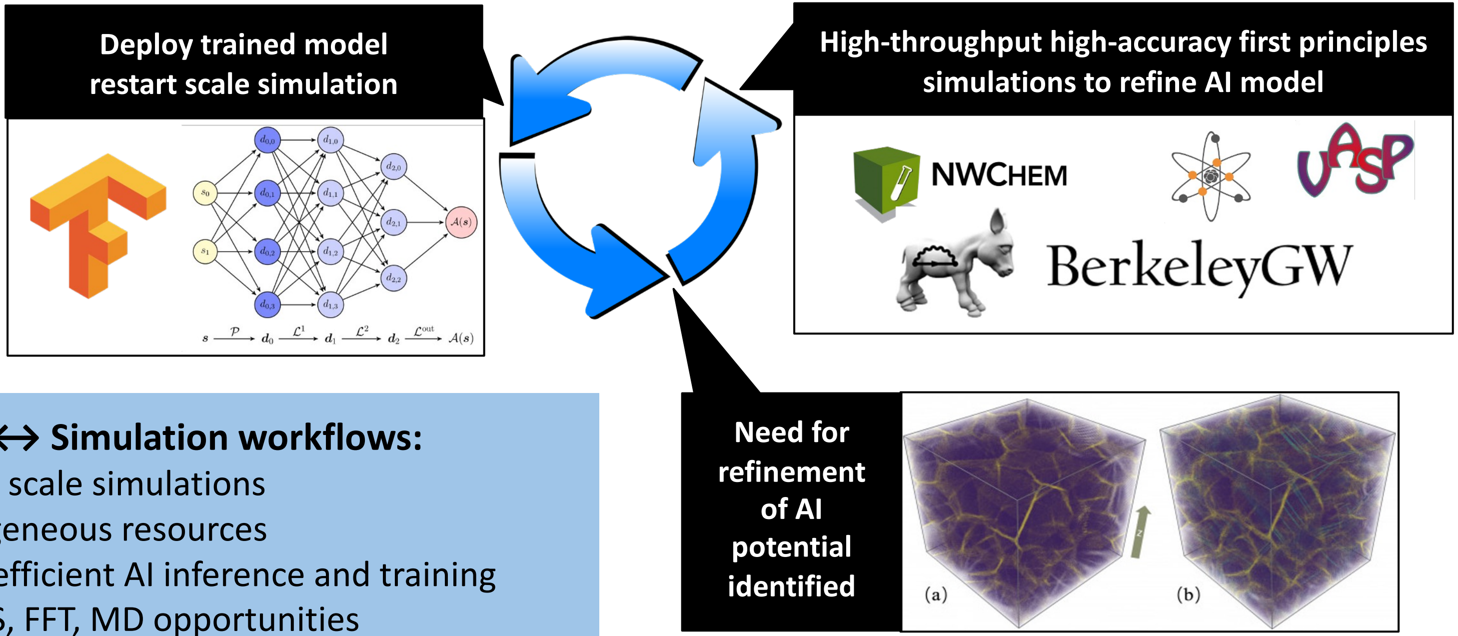


Can analyze a 5 minute experiment in ~3 minutes for feedback to beamline staff, transferring 15TB/day to NERSC

BES: AI-Enhanced Materials Design Workflow

AI accelerated "traditional" simulation at scale = direct access to experimental regime



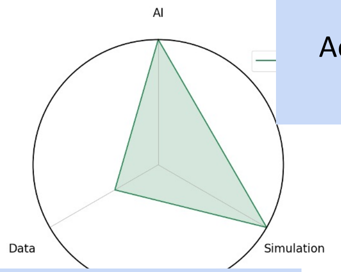


Requirements for AI ↔ Simulation workflows:

- Support for large scale simulations
- Dynamic heterogeneous resources
- Accelerators for efficient AI inference and training
- Accelerated BLAS, FFT, MD opportunities
- Advanced scheduling and data management tools

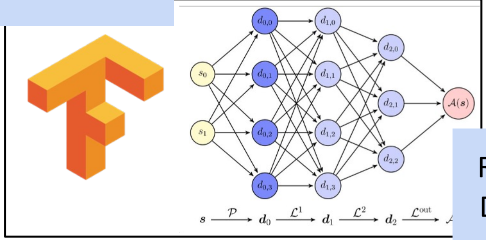
BES: AI-Enhanced Materials Design Workflow

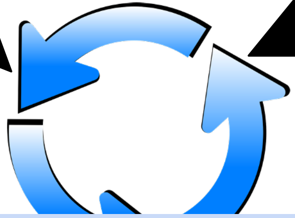
AI accelerated “traditional” simulation at scale = direct access to experimental regime



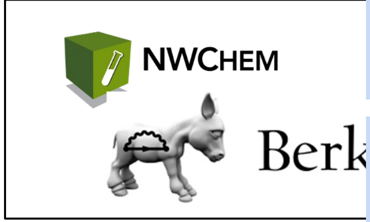
Benefits from Dedicated AI Accelerators for Inference

y trained model
scale simulation





High-throughput high-acc
simulations to ref



Requires Near Full-System (“exascale”) Tightly Coupled Runs on Complex Material Geometries

Benefits from Accelerators for BLAS, FFT, MD

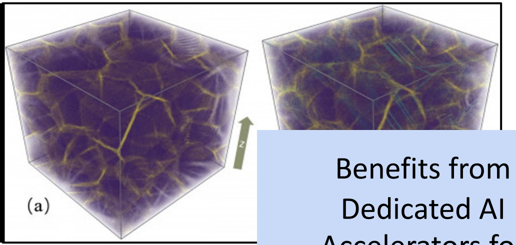
Requires Capability to Dynamically Expand / Reduce Running Job Size - Frequency of Minutes

May be Connected to Experiment with Real-Time Need: X-Ray Scattering, Photo-emission, Absorption, Tomography

or AI ↔ Simulation workflows:

- large scale simulations
- eterogeneous resources
- rs for efficient AI inference and training
- Accelerated BLAS, FFT, MD opportunities
- Advanced scheduling and data management tools

d for
ement
AI
potential
identified



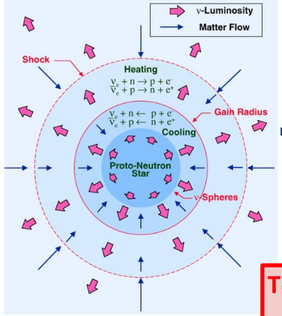
Benefits from Dedicated AI Accelerators for Training

HEP: DUNE Supernova Neutrino Trigger

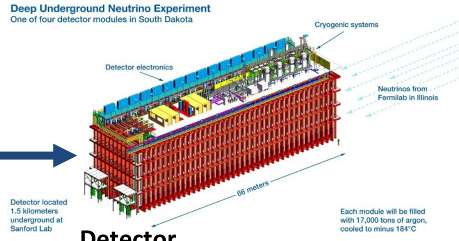
DUNE: Neutrino beam from Fermilab toward detector at Sanford Underground Research Facility

- **Neutrino oscillations & interactions with matter**
- **⇨ Core collapse supernovae ⇦**

"Nearby" core collapse supernova



neutrinos



Detector

Full detector readout
Up to 184 TB compressed
4h over 100Gbs

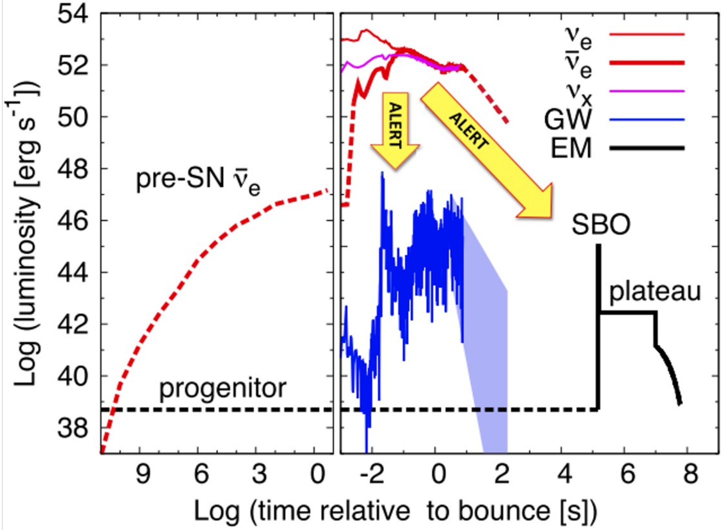
ESnet
ENERGY SCIENCES NETWORK

Telescopes need to observe the supernova within hours

Telescopes

early warning

NERSC-10



response window

Requirements for disruptive (~1 trigger/month):

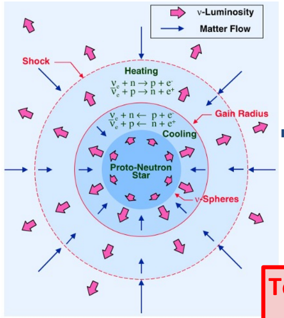
- Compute nodes available on demand, no queue wait
- High-bandwidth ingress to compute nodes on demand
- No phone calls, emails, tickets; just scripted API calls

HEP: DUNE Supernova Neutrino Trigger

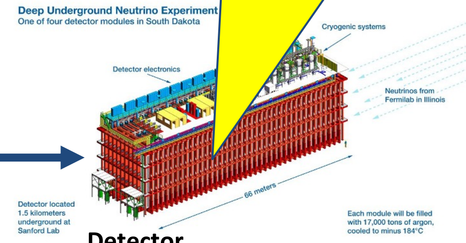
DUNE: Neutrino beam from Fermilab toward detector at Sanford Underground Research Facility

- **Neutrino oscillations & interactions**
- **⇨ Core collapse supernovae ⇦**

"Nearby" core collapse supernova



neutrinos

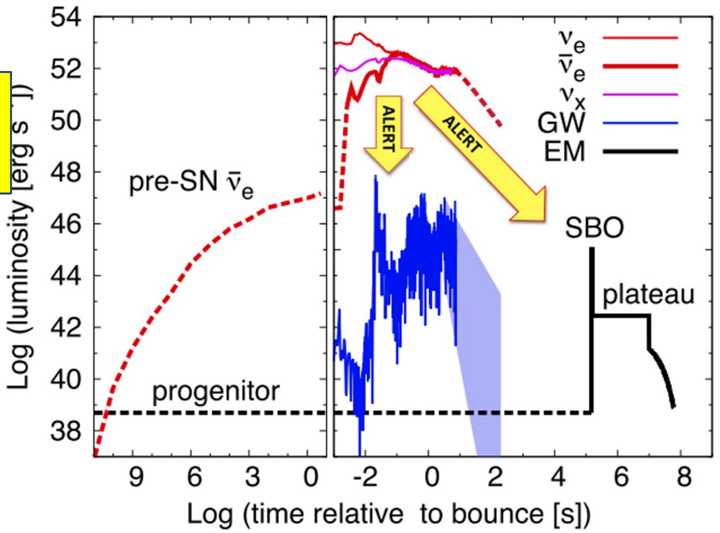


Detector

Large data volume collected in minutes, egress limited by network BW

Full detector readout
Up to 184 TB compressed
4h over 100Gbs

ESnet
ENERGY SCIENCES NETWORK



response window

NERSC-10

Immediate launch of compute jobs.130k CPU hours to process full dataset.

disruptive (~1 trigger/month):

- vailable on demand, no queue wait
- High-bandwidth ingress to compute nodes on demand
- No phone calls, emails, tickets; just scripted API calls

Telescopes need to observe the supernova within hours

early warning

Telescopes

Results shared to huge astronomy community via web portal

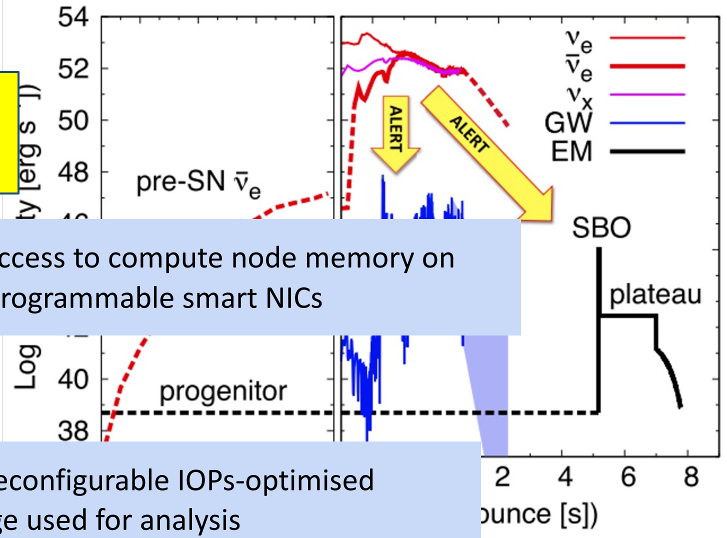
HEP: DUNE Supernova Neutrino Trigger

DUNE: Neutrino beam from Fermilab toward detector at Sanford Underground Research Facility

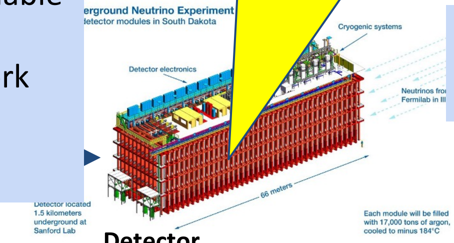
- **Neutrino oscillations & interactions**
- **⇨ Core collapse supernovae ⇦**

Large data volume collected in minutes, egress limited by network BW

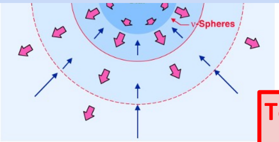
Needs automation - programmable interfaces for entire workflow, from data movement to network config to job launch and monitoring



High-BW access to compute node memory on demand: programmable smart NICs



Detector



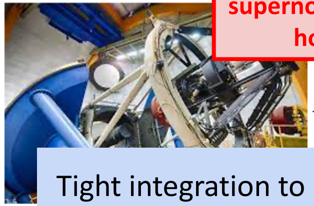
Full detector readout
Up to 184 TB compress
4h over 100Gbs

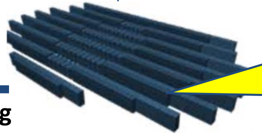
ESnet
ENERGY SCIENCES NETWORK

Fast reconfigurable IOPs-optimised storage used for analysis

Telescopes need to observe the supernova within hours

response window





early warning

NERSC-10

Immediate launch of compute jobs.130k CPU hours to process full dataset.

disruptive (≈1 trigger/month):
queue wait
s on demand
mails, tickets; just scripted API calls

May benefit from data reduction on smartNIC

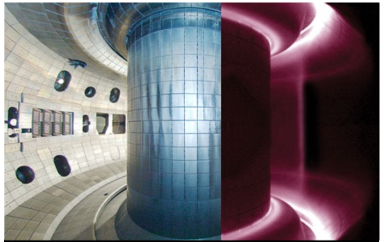
May benefit from AI-specialised hardware

Tight integration to secure, scalable Spin services from N10 file systems

ts shared to huge
nomy community
via web portal

FES: DIII-D workflow with real-time feedback

DIII-D National Fusion Facility (San Diego, CA)



Event:
Anomaly detected at DIII-D experiment

Scientists and operators in control room have ~20 minutes to decide what to do before the next discharge

Fast data transfer, real-time compute

++

Quickly process data, use AI at scale to locate and analyze anomalies, compare against simulations, and use AI to recommend action.

Real-time HPC capabilities in N10 enable:

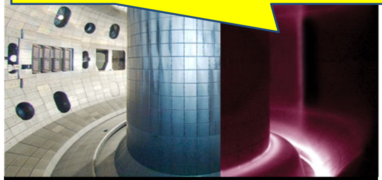
- Agile experimental decision making tools
- Optimized use of valuable experimental time
- Enhanced scientific productivity

Real-time results from N10 workflow can help answer:

- What caused the anomaly?
- Adjust experimental parameters?
- Safe to continue?

FES: DIII-D workflow with real-time feedback

End data stored at local sites



Event:
Anomaly detected at
DIII-D experiment

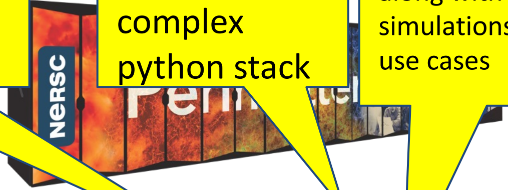
Scientists and operators in control room have ~20 minutes to decide what to do before the next discharge

Need to launch different dependent simulation types simultaneously across node types

Need containers for complex python stack

Need some compute nodes, constant during the experiment along with large scale simulations for some use cases

real-ti...
compute



Quickly process data, use AI at scale to locate and analyze anomalies, compare against simulations, and use AI to recommend action.

Real-time HPC capabilities in N10 enable:

- Agile experimental decision making tools
- Opti...ized use of ...luable experimental time
- ...entific ...luctivity

Real-time results from N10 workflow can help answer:

- What c... d t...
- ...djust expe...

Time allocated to science team on instrument - know in advance when will need NERSC

~30 shots/day. Need NERSC resources available for full experiment

Small network needs: Image data ~10GB/shot

Need to stream data to compute node memory - direct secure path (ADIOS)

Need to trigger compute when data arrives, via API or controller in Spin, tightly coupled to filesystems

FES: DIII-D workflow with real-tin

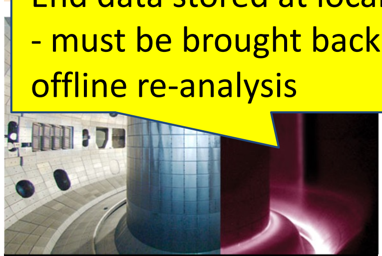
Simulations can scale up to ~2000 nodes.

Need some compute nodes, constant during the experiment along with large scale simulations for some use cases

Need to launch different dependent simulation types simultaneously across node types

Need containers for complex python stack

End data stored at local sites (security) - must be brought back to NERSC for offline re-analysis



Event:
Anomaly detected at DIII-D experiment

Scientists and operators in control room have ~20 minutes to decide what to do before the next discharge

real-ti
compute

May benefit from AI accelerator

Quickly process data, use AI at sc le to locate and analyze anomalies, compare against simulations, and use AI to recommend action.

Real-time HPC capabilities in N10 enable:

- Agile experimental decision making tools
- Opti ized use of luable experimental time
- entific luctivity

Real-time workflo

- What
- dius

smart NICs allow many workflows like this to exist within N10, with secure access to all file systems (today Spin only has access to CFS), tighter coupling between persistent service and compute on an integrated network.

Time allocated to science team on instrument - know in advance when will need NERSC

~30 shots/day.
Need NERSC

Small network needs: Image data

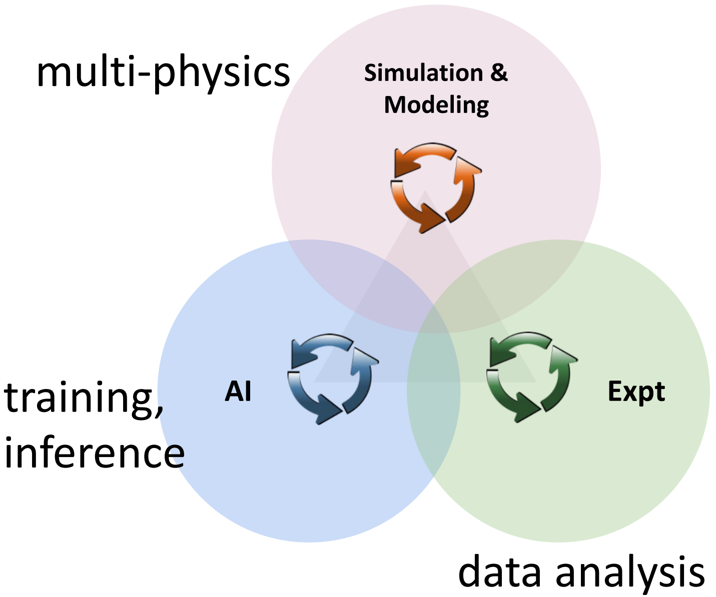
Need to str to compute node direct h (ADIOS)

Spin, tightly coupled to filesystems

Needs automation - programmable interfaces for entire workflow, from data movement to network config to job launch and monitoring

What is 'an HPC' workflow?

Interconnected computational/dataflow tasks (> 1) with data products;
Task Coupling and/or Data Movement between tasks (control flow & data flow)

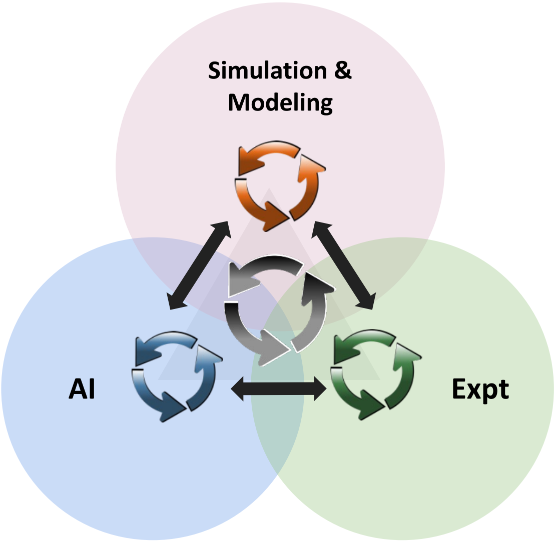


Coupled tasks to enable HPC Science Campaigns

- Experiment to high-performance data analytics Workflow
- High-performance Modeling & Simulation Workflow
- High-performance AI Workflow

- **Coupled workflows within, between or across all three modes**

Growing Opportunity and Complexity



NERSC users require a paradigm shift in the way we design, configure and operate HPC systems

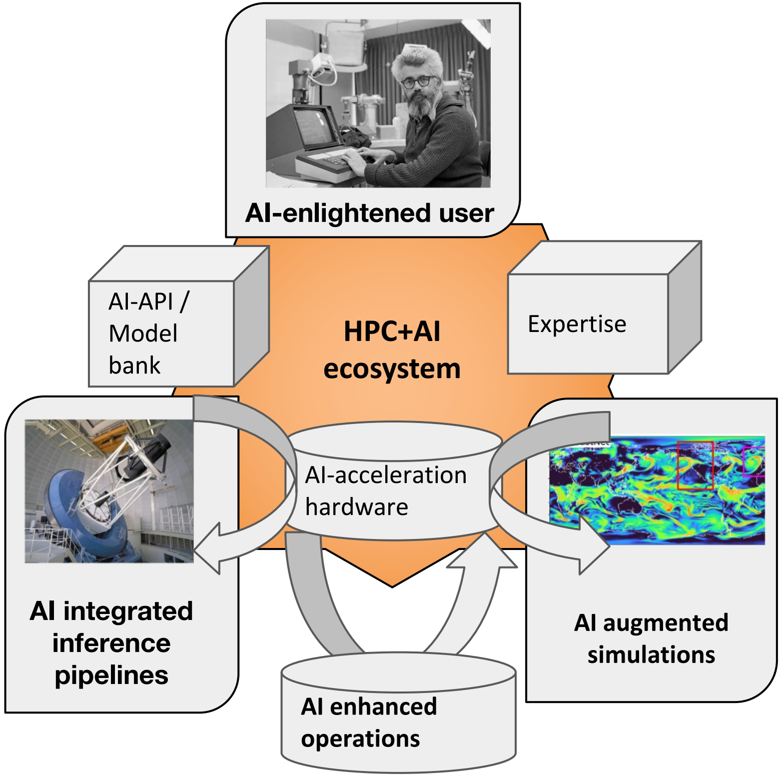
Users require an ***integrated ecosystem*** that supports new paradigms for ***data analysis*** *with* ***real-time interactive feedback*** *between experiments and simulations. Users need the ability to* ***search, analyze, reuse,*** *and* ***combine data*** *from* ***different sources*** *into* ***large scale simulations and AI models.***

NERSC-10 Mission Need Statement: *The NERSC-10 system will accelerate end-to-end DOE SC workflows and enable new modes of scientific discovery through the integration of experiment, data analysis, and simulation.*

NERSC-10 AI Vision

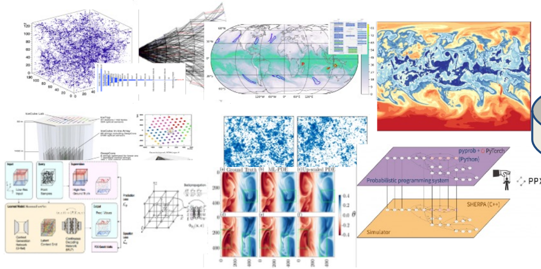
World-leading *open-science AI ecosystem* with

- ***Underlying AI-acceleration system hardware and software*** that are best-in-class, including compute, workflow and data management
- ***Service platform*** *that offers*:
  - Interactivity for large-scale model exploration
  - “Foundation” model hosting and retraining
    - Incorporating novel AI4Sci techniques
    - Accessible to AI novices and experts
  - Coupled AI, simulation and data pipelines
- ***Applications*** for *science* with new approaches to achieve large, science-informed, uncertainty-aware and transferable models
- ***Expertise, consulting and education***

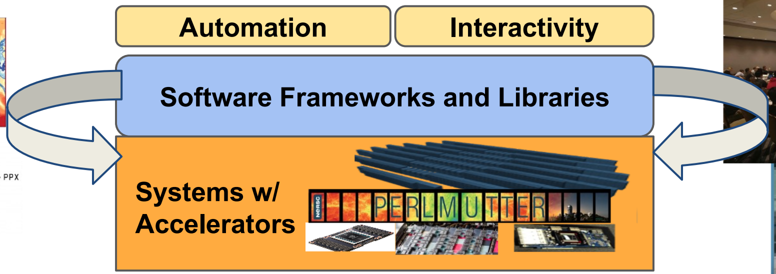


NERSC AI Strategy

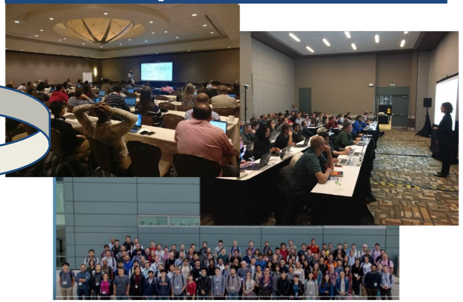
Methods and Applications



Deployment



Empowerment



- ***Deploy*** optimized hardware and software systems
  - Currently Perlmutter >6000 A100 GPUs; Work with vendors for optimized AI software
  - >6x increase in usage of DL frameworks since 2018
  - Improve performance, e.g through benchmarking (e.g. MLPerf HPC))
- ***Apply*** *ML* for science using cutting-edge methods
  - “NESAP for Learning” application readiness program with postdocs, early access etc.
  - Other targeted engagements that push model development, scale and performance
  - Leverage lessons learned for all users
- ***Empower*** through seminars, training and schools
  - E.g. Deep Learning at Scale tutorial at Supercomputing (SC21 material here)

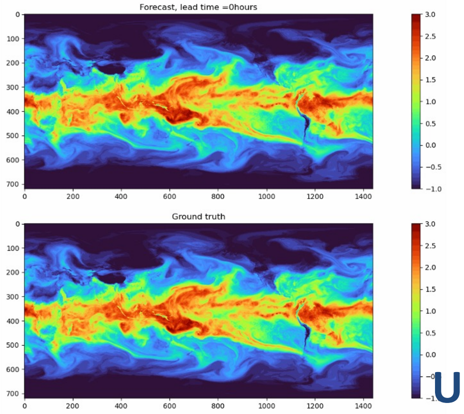
Transformative AI for new science

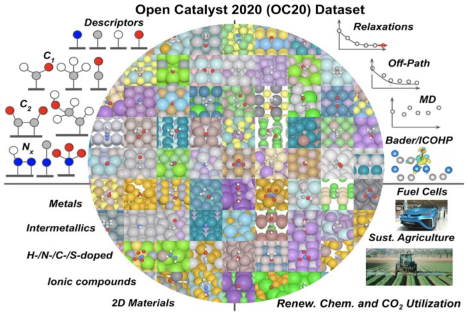
FourCastNet

Pathak et al. 2022 [arXiv:2202.11214]

Forecasts global weather at high-resolution. Hybrid data/ model parallel @ 4000 GPUs

First deep-learning with skill of numerical weather prediction





CatalysisDL

Chanussot et al. 2021

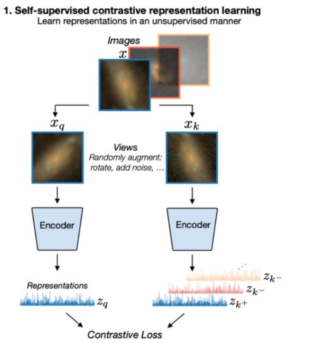
[arXiv:2010.09990]

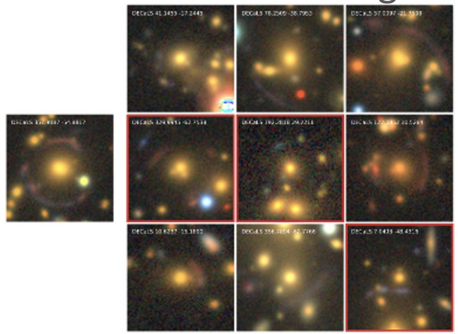
Co-developed largest catalysis dataset ([OC20]); [Graph-parallel NN approaches] and [NeurIPS 2021 Competition]

Self-supervised sky surveys

Stein et. al. (2021) [arXiv:2110.00023]

Uncovered thousands of undiscovered strong-lenses



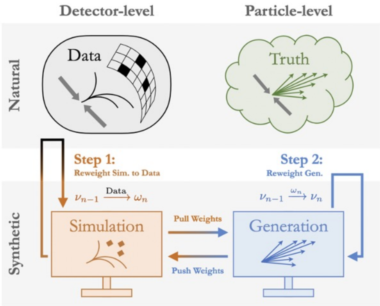


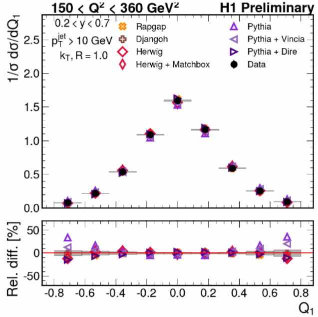
Unfolding for particle physics

H1 Collaboration: announced [at DIS2022]

New ML approach extracts new physics insights from old data.

Requires Perlmutter for multiple distributed training runs



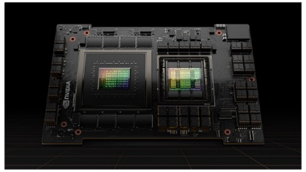


AI Accelerator Strategy and Pathfinding

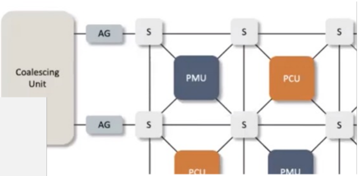
Exploring *Novel AI Acceleration* options for NERSC-10 system

- Formed a cross-LBL AI hardware working group in 2020 to evaluate the potential of AI-focussed accelerators for science
  - Defining benchmarks and metrics for scientific ML
  - Collating science experiences on AI-specific hardware
- Leading development of scientific AI benchmarks on HPC systems: e.g. MLPerf HPC
- Deepening understanding of performance bottlenecks on current and emerging hardware - e.g. HPC DL Architectural requirements (PMBS 2021); MLPerf HPC analysis (MLHPC 21); Scientific ML on Graphcore (PMBS 2022) ...
- Fully exploiting Perlmutter, current state-of-the-art AI accelerator system, with novel and at-scale AI applications:
  - Refreshing NESAP for learning applications
  - Will help find bottlenecks and evaluate needs for future
- Also prototyping, evaluating and developing *AI service platforms*

Nvidia Grace Hopper

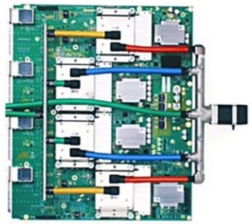


SambaNova

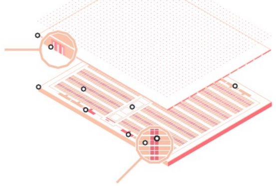


Cerebras

Google TPUv4

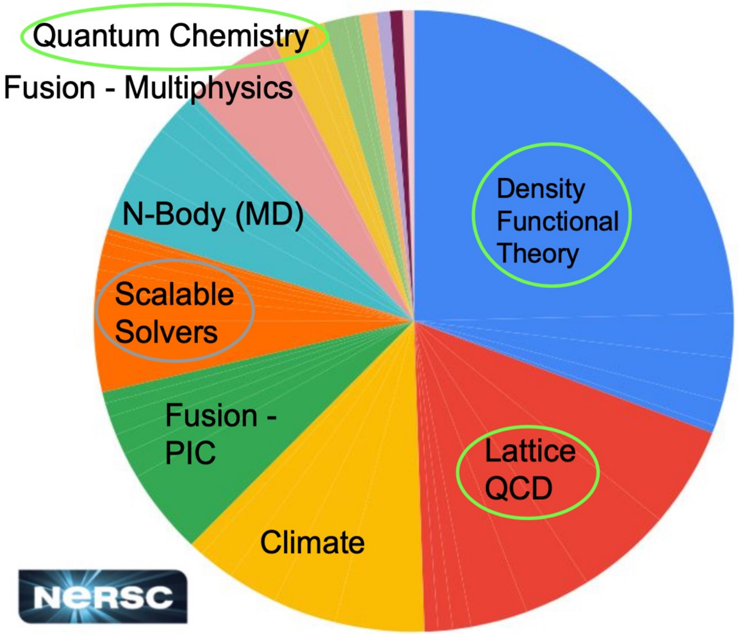


Graphcore



Quantum Computing applies to more than 50% of the NERSC workload

Top algorithms among NERSC codes allocation year 2018



	Logical Qubits	Note
Quantum Chemistry	∝ active orbitals 10^1-10^2	Possible NISQ "killer app" - NAS
Density Functional Theory	∝ bands 10^3 - 10^5	Algorithm published. Like ab initio, but larger systems.
Lattice QCD	∝ lattice sites. 10^6-10^9	Algorithm published.
Machine Learning	???	Framework published. Tensor-Flow Quantum
Scalable Solvers	???	Kernels published. (Ax=b, FFT)

NERSC Quantum Roadmap

2022	2022-2024	2024-2028	2028-203?
• Ramp up engagement with QIS community • Director's Discretionary Reserve Call for quantum information science (QIS) on Perlmutter	• Engage with quantum hardware companies • Enable user access to quantum hardware • Development and testing of classical-quantum hybrid algorithms	• Integration of near-term (NISQ) quantum hardware becoming standard - Users requesting both classical and quantum resources	• Fault-tolerant quantum hardware becoming available • Full integration with traditional HPC • Users routinely solve problems using quantum hardware !

Optimal integration of classical and quantum processors is an open area of research

Summary

- The evolving NERSC workload requires a shift in how to design future systems
- NERSC-11 is likely to be even a greater technological change from previous systems