Enabling AI for Science at Scale on the Perlmutter Supercomputer

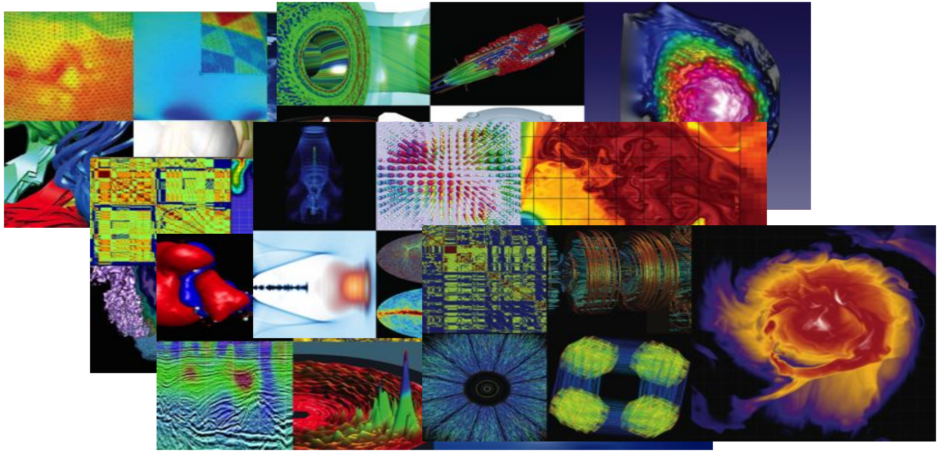
SOS26
March 14, 2024

Steven Farrell
Data & AI Services, NERSC
Lawrence Berkeley National Lab

Outline

- NERSC AI strategy
- Enabling NCCL on Slingshot 11
- MLPerf HPC and Perlmutter
- Current and future directions

NERSC: Mission HPC for the Dept. of Energy Office of Science



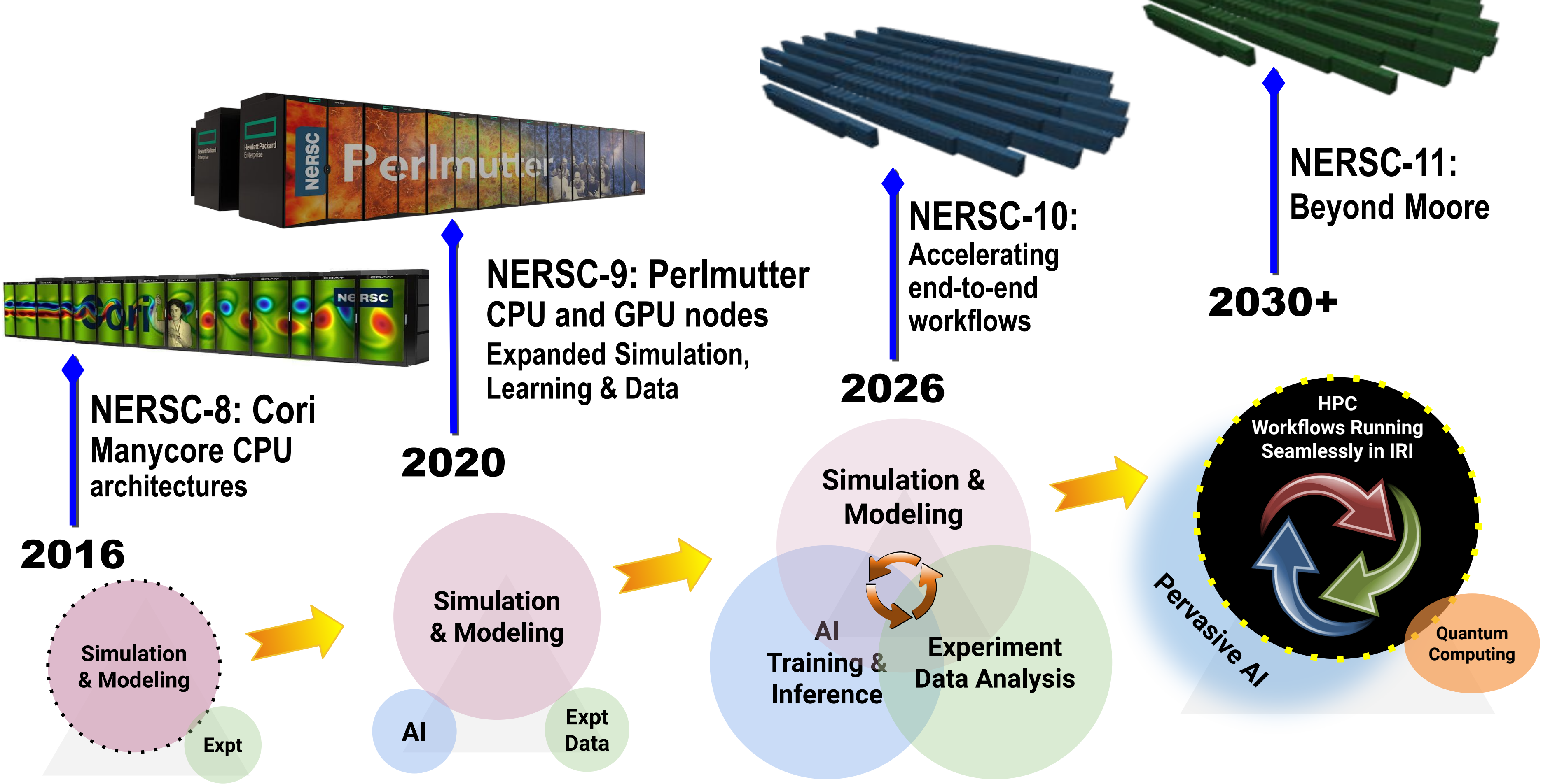
Large compute and data systems

- Perlmutter: ~7k A100 GPUs
- 128PB Community Filesystem ....

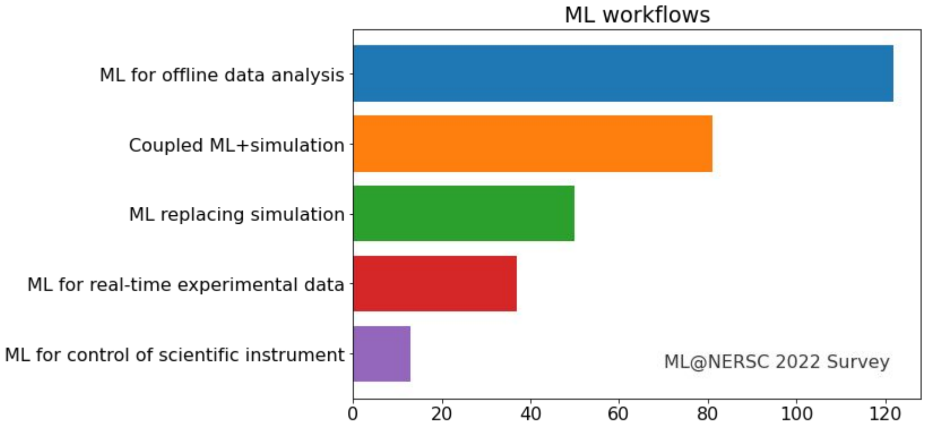
Broad science user base

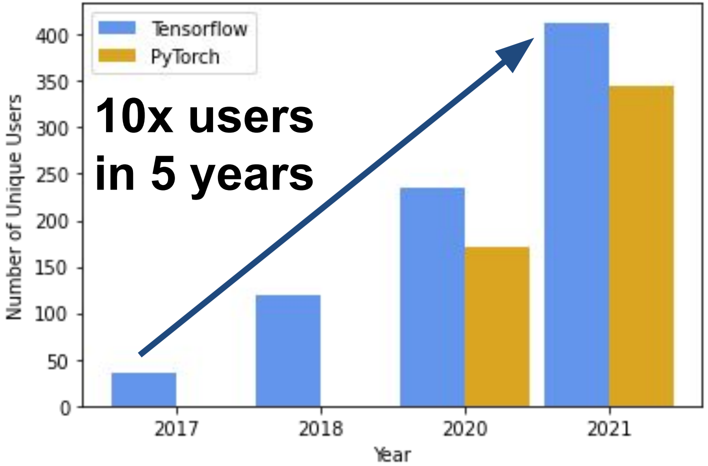
- ~10,000 users,
- 1,000 projects,

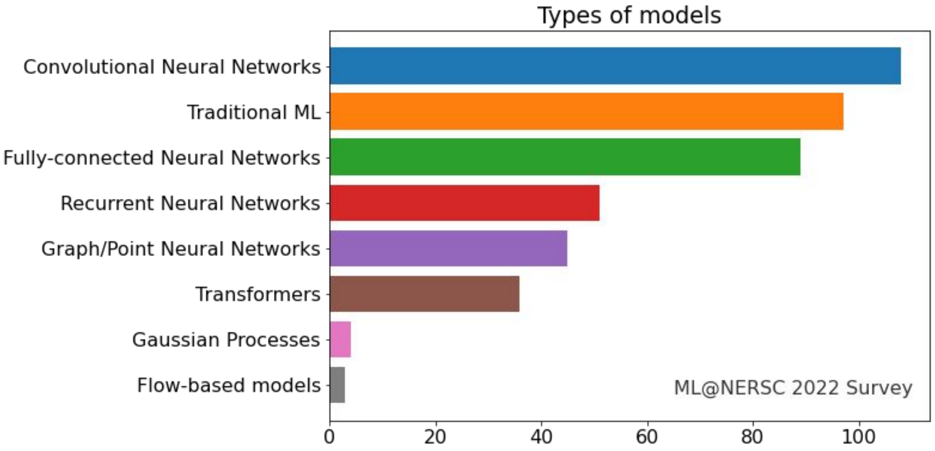
NERSC roadmap

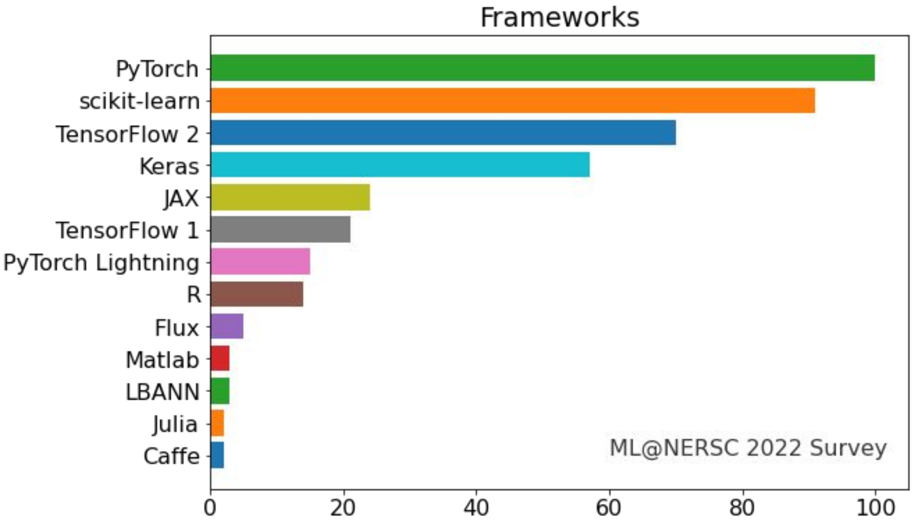


A growing scientific AI workload

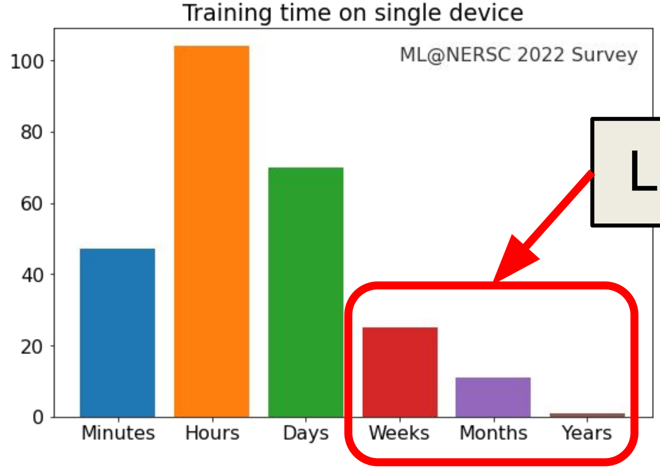


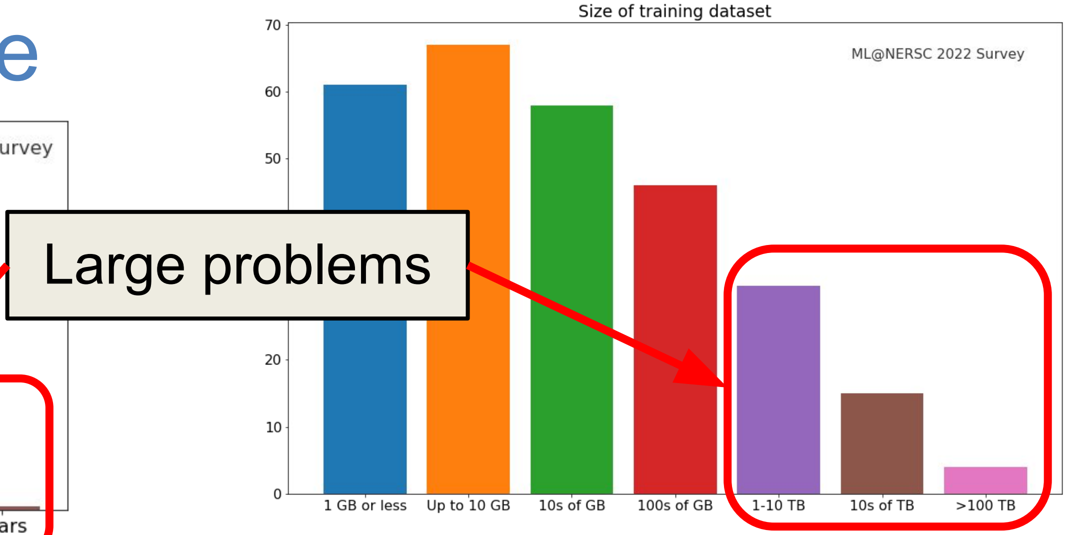


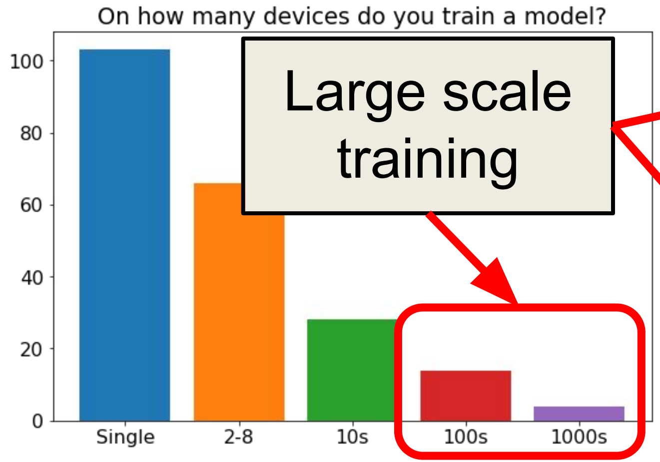


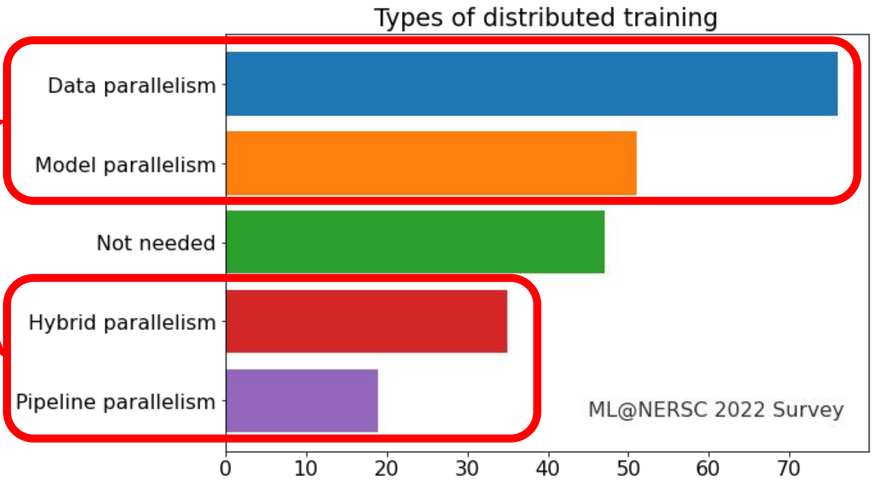


Need for AI at scale



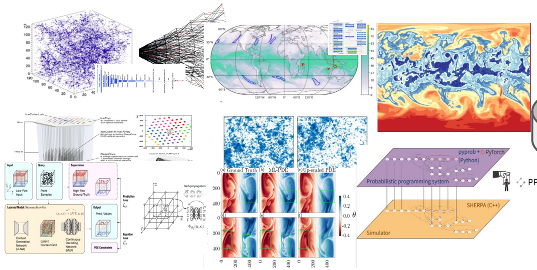






NERSC AI Strategy

Methods and Applications



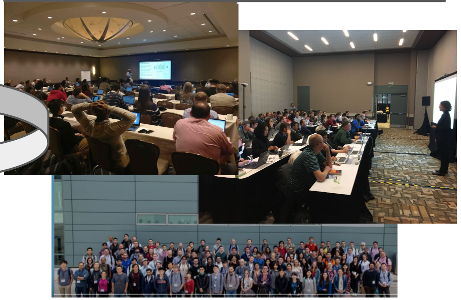
Deployment

Automation | **Interactivity**

Software Frameworks and Libraries

Systems w/ Accelerators

Empowerment



- ***Deploy*** optimized hardware and software systems
  - Work with vendors for optimized AI software
- ***Apply*** AI for science using cutting-edge techniques
  - “NESAP” and strategic projects - leverage lessons learned for scalable impact
- ***Empower*** and develop workforce through seminars, training and schools as well as staff, student intern and postdoctoral programs
  - Over 20 DL@Scale tutorials (e.g. SC18-23), 1000s of total participants

Perlmutter: A Scientific AI Supercomputer

HPE/Cray Shasta system

Phase 1 (Dedicated May `21):

- 12 GPU cabinets with 4x NVIDIA A100 GPU nodes; Total >6000 GPUs
- 35 PB of All-Flash storage

Phase 2 (Integrated in 2022):

- 12 AMD CPU-only cabinets
- HPE/Cray Slingshot high performance ethernet-based network

Optimized software stack for AI

Application readiness program (NESAP)

NVIDIA

HOME AI NETWORKING DRIVING GAMING PRO GRAPHICS AUTONOMOUS MACHINES HEALTHCA

Need for Speed: Researchers Switch on World's Fastest AI Supercomputer

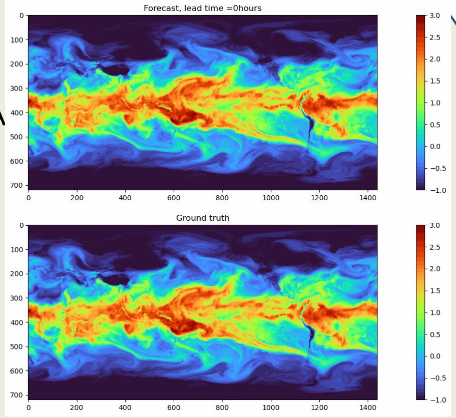
NVIDIA blog May 2021

NESAP and Perlmutter are Enabling Adoption of Large-scale and Groundbreaking AI

FourCastNet

Pathak et al. 2022 **arXiv:2202.11214**
Collab with Nvidia, Caltech, … (+ now LBL EESA)

- Forecasts global weather at high-resolution.
- Prediction skill of numerical model; 10000s times faster



Jaideep Pathak
former NERSC
Postdoc now NVIDIA

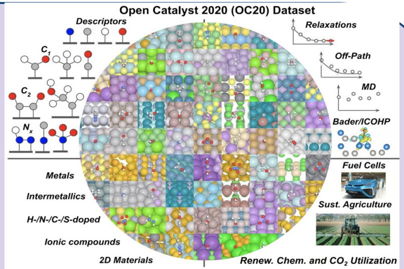
Shashank
Subramanian
Former NERSC
Postdoc now Staff

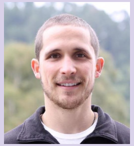
Jared Willard
NERSC Postdoc

CatalysisDL

Chanussot et al. 2021
Collab with CMU, MetaAI, …
arXiv:2010.09990

- NeurIPS 2021-23 Competitions
- Pre-trained models now used with DFT - e.g. FineTuna; AdsorbML





Brandon Wood
former NERSC
Postdoc now Meta AI

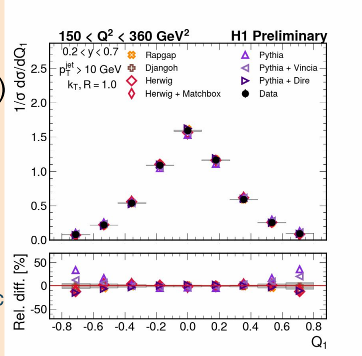
Wenbin Xu
NERSC postdoc

HEP-ML

Collab with LBL Physics division (and H1 Collaboration)

- AI "Unfolding" extracts new physics insights from data
  - Requires Perlmutter for 1000s of UQ runs

Vinicius Mikuni
NERSC Postdoc



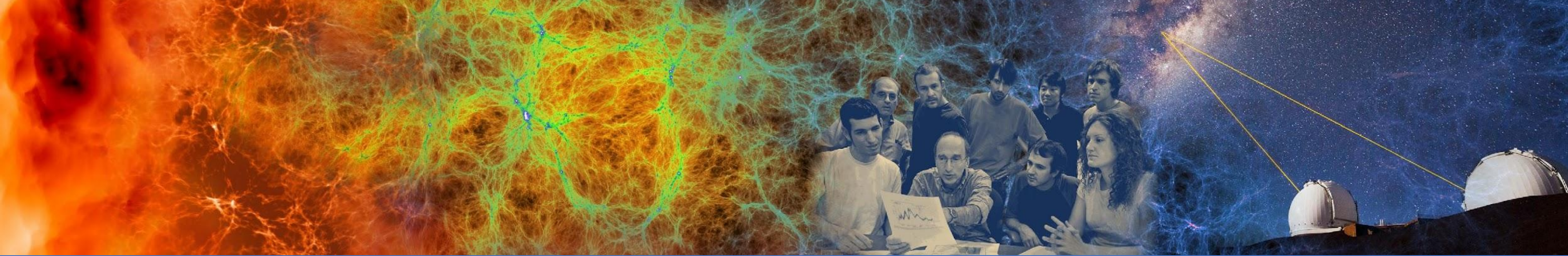
Empowering the science communities for Deep Learning

The Deep Learning for Science School at Berkeley Lab https://dl4sci-school.lbl.gov/

- Lectures, demos, hands-on sessions, posters: 2019 in person (videos, slides, code)
- 2020 summer webinar series ***focussed on science and computing***.
  Recorded talks: https://dl4sci-school.lbl.gov/agenda

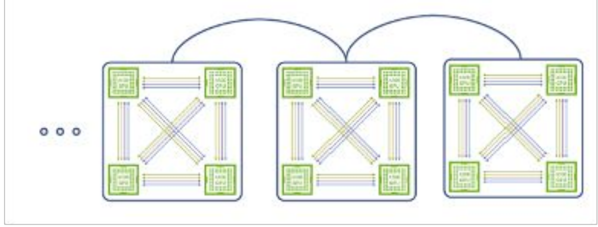
The *Deep Learning at Scale* Tutorial

- Run since 2018 with Cray, Intel, OLCF and NVIDIA
- Powered by Perlmutter since 2021 with **hands-on material for distributed training**
- Featuring sophisticated ViT science example with content on GPU optimization, data + model parallelism
- SC23 material: https://github.com/NERSC/sc23-dl-tutorial



Enable NCCL performance on Perlmutter

NCCL on Perlmutter background



NVIDIA Collective Communications Library (NCCL)

- Critical for high-performance distributed training in deep learning frameworks
- Need high-bandwidth, low-latency P2P all-reduce between GPUs
- NCCL uses NVLINK on-node, interconnect across-node

Perlmutter deployment

- Phase 1, Slingshot 10, 2 NICs (100Gb), RoCE
- Phase 2, Slingshot 11, 4 NICs (200Gb), requires libfabric

Initial performance

- TCP for inter-node communication => *2-3x perf drop!!*

Benchmarks	Phase I, SS10	Phase II, SS11
NCCL-Tests AllReduce (32 MB) 2 Node (GB/s) (**higher is better**)	26	9.5
Tensorflow 2 + Horovod (ResNet/ImageNet) (w/Shifter) 2 Nodes (samples/second) (**higher is better**)	4700	3900
DeepCam-4k 8 Node Runtime (min) (**lower is better**)	5.2	7.0

NERSC

NCCL plugin for Slingshot 11

Leverage AWS open-source libfabric NCCL plugin for their EFA network

- Initial efforts led by Josh Romero, Jim Dinan (NVIDIA)

AWS OFI NCCL

AWS OFI NCCL is a plug-in which enables EC2 developers to use libfabric as a network provider while running NVIDIA's NCCL based applications.

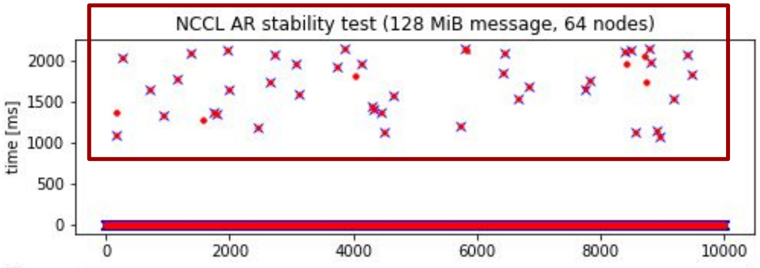
https://github.com/aws/aws-ofi-nccl

Early days (2022 Q3-4): focus on getting something functional at mid-intermediate scale, then improve performance

- Deployed NCCL builds with plugin as baremetal and shifter modules
- Multiple iterations of issues, debugging, adapting as SS11 was hardened

NCCL remaining issues

64 nodes

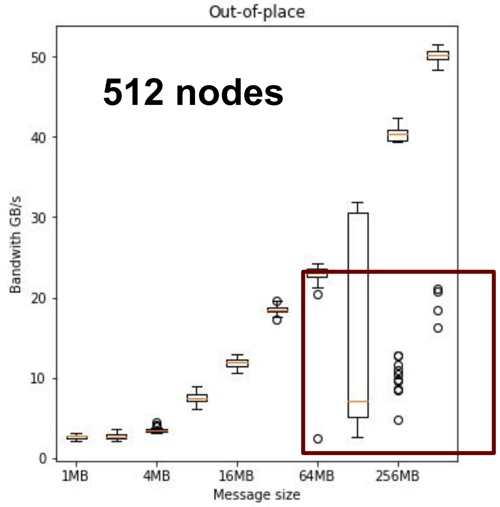


Jobs hanging

- Multiple types of hangs caused by different parts of the hardware/software stack
- Some intermittent/only emergent at very large scale; we struggled with these for months!
- Found and removed some “bad nodes”

Spurious performance drops

- Saw intermittent substantial drops in NCCL bandwidth (>10-100x reduction); worst at large scale
- Root-caused to the protocol used in Slingshot for queueing messages
- NVIDIA devs worked with Igor Gorodetsky (HPE) to resolve
- Fixed in Slingshot Host Software (SHS) 2.1.0 Q2 2023

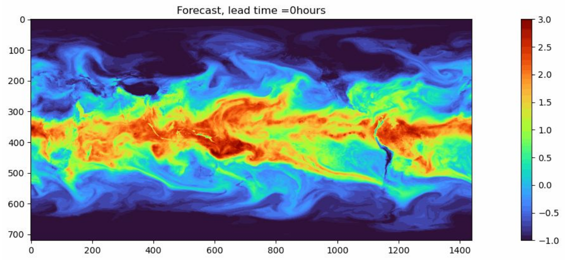


NCCL current status

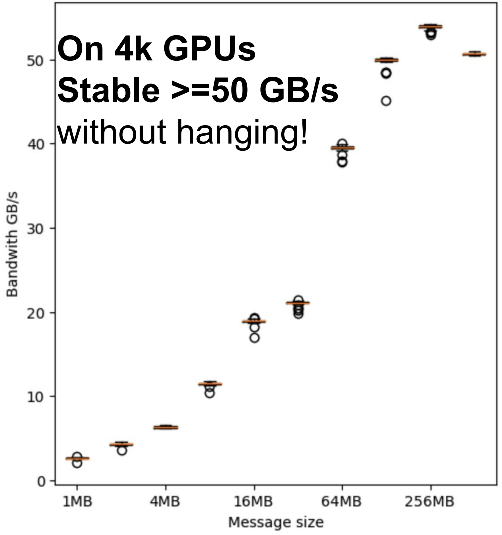
Hangs and biggest performance issues seem resolved!

Impact on real workloads: FourCastNet++ ([PASC 2023]) hybrid data-model parallel DL weather model training

- Performance on 4k GPUs now sees 60% end-to-end speedup from SS10

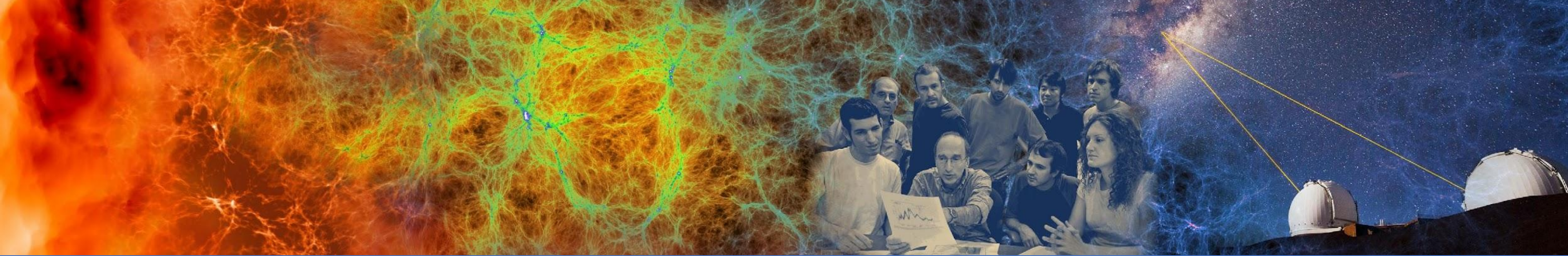


Mission Accomplished 🎉



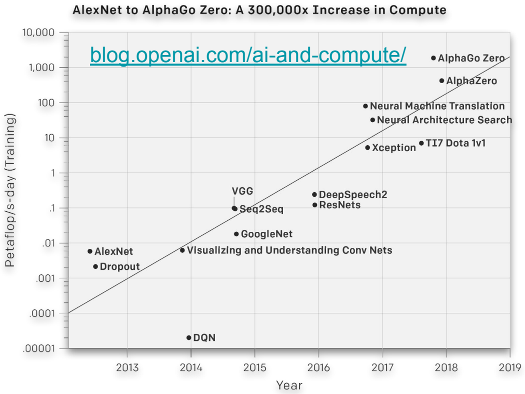
Next steps

- Further optimizations
- Improved integration into container runtimes (e.g. podman)
- Solidify long-term support plan across orgs as NCCL, Slingshot evolve



MLPerf HPC on Perlmutter

The need for HPC ML benchmarks



- **We see a growing ML workloads at HPC centers**
- **We need good benchmarks for this emerging, evolving workload**
- **We therefore formed an HPC Working Group in MLCommons**
  - to build a community focused on issues related to scientific AI on HPC
  - to develop the MLPerf HPC benchmark suite

MLPerf HPC

A benchmark suite for ML training workloads on HPC systems

- Large-scale scientific problems and datasets
- Modeled after MLPerf Training rules

Two measurement types

- Time-to-train (strong-scaling):
  - Traditional measurement from MLPerf Training
  - Measures time to train to target quality
- Throughput (weak-scaling):
  - Submitter trains many models concurrently to target quality
  - Measures models trained per unit time (higher is better)
  - Captures aggregate ML capabilities of HPC system of any scale

Submission Process

Submitters optimize benchmarks and measure time to train to convergence

- Handle stochasticity by running multiple times and taking average

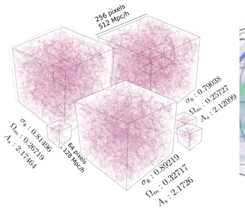
Divisions offer balance between comparability and innovation

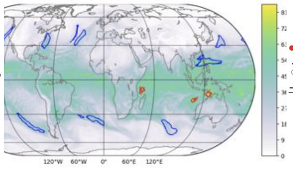
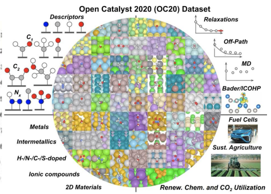
- *Closed division*: submissions must be “consistent” with reference model, training procedure, but allow some hyperparameter tuning
- *Open division*: can change the model and training procedure

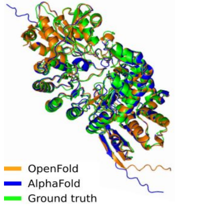
Submission period followed by *peer review process*

- Submitters review each other’s code and results
- Can also “borrow” hyperparameters to rerun in this phase

MLPerf HPC benchmarks





Benchmark	Task	Dataset	Reference Model	Quality Target
DeepCAM	Climate segmentation	CAM5+TECA simulation, image size (768x1152x16), 8.8 TB	DeepCAM 2D CNN based on DeepLabv3+	0.82 IOU
CosmoFlow	Cosmology parameter prediction	CosmoFlow N-body simulation, 3D cubes of size 128^3, 10.2 TB	CosmoFlow 3D CNN	0.124 MAE
OpenCatalyst	Quantum molecular modeling	Open Catalyst 2020 (OC20) S2EF, 300GB	DimeNet++ GraphNN	0.036 Forces MAE
OpenFold (*NEW*)	Protein Folding	OpenProteinSet and Protein Data Bank (May 2022 snapshot)	AlphaFold2 (PyTorch)	0.8 Local Distance Difference Test (IDDT)

MLPerf HPC results highlights

Four successful submission rounds since 2020

- ~90 total results
- 12 total submitting orgs, 15 total HPC systems
  - from DOE, NSF, Europe, Asia, vendors
- Time-to-train results scaled up to 2,048 GPUs
- Throughput results scaled up to
  5,120 GPUs (Perlmutter),
  82,944 CPUs (Fugaku)
- Impressive speedups year after year
  - DeepCAM 14x speedup from v0.7 -> v3.0
  - OpenCatalyst 10x speedup from v1.0 -> v3.0

Systems
ANL: ThetaGPU
Clemson: Palmetto
CSCS: Piz Daint
Dell: 32XE8545-4xA100-SXM4-40GB
Fujitsu: ABCI
Helmholtz AI: HoreKa, JUWELS
RIKEN+Fujitsu: Fugaku
LBNL (+HPE): Cori, Cori-GPU, Perlmutter
NCSA: HAL
NVIDIA: Selene
TACC: Frontera-RTX, Longhorn

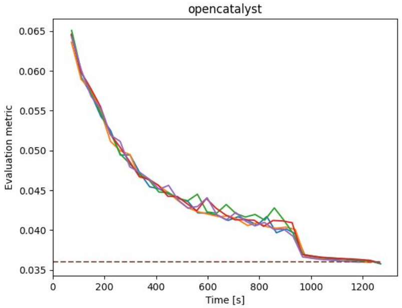
Perlmutter and MLPerf HPC v3.0

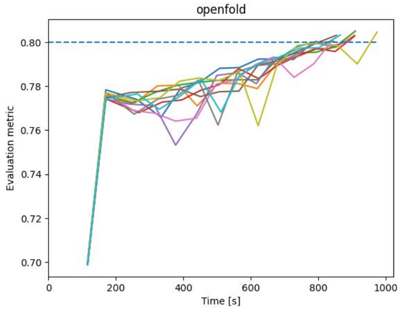
NERSC partnered with HPE and NVIDIA to submit results using Perlmutter

- Previously submitted with Phase 1 system (and slingshot 10)

We utilized various optimization techniques to get good performance

- NGC containers in shifter
- PyTorch JIT compilation and CUDA graphs
- Optimized data movement from Lustre
- NVIDIA DALI for data loading



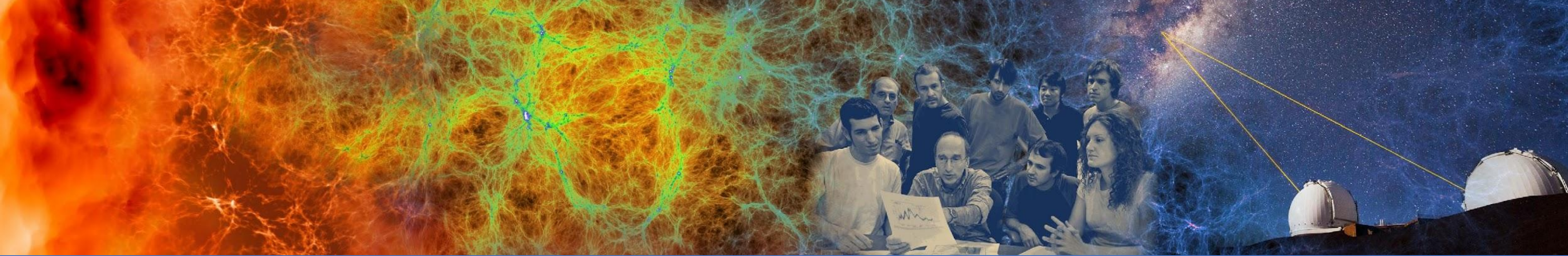


We achieved highly competitive results

- Excellent improvements over our previous results
- Overall a very valuable experience for us

Nodes	GPUs	CosmoFlow	DeepCAM	OpenCatalyst	OpenFold
128	512	4.73		21.04	
224	896				16.11
512	2048		1.81		

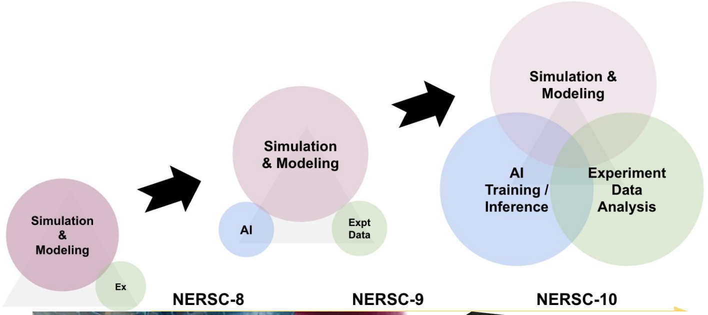
Full results: https://mlcommons.org/benchmarks/training-hpc/
Press release: https://mlcommons.org/2023/11/mlperf-training-v3-1-hpc-v3-0-results/
Code and log files: https://github.com/mlcommons/hpc_results_v3.0/

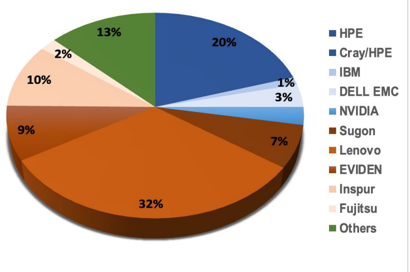


Current and future directions

Drivers for the future

- Rapidly-evolving AI methods, software
- Expanding AI user base, use cases, computational needs
- Foundational models for science
- Maturing of AI applications
- Growing need for inference
- IRI workflows (streaming data, APIs, realtime)
- More complex workflows (AI+sim, active learning, etc.)
- New capabilities in AI for HPC operations
- Larger, complex HPC systems
- Evolving HPC vendor landscape

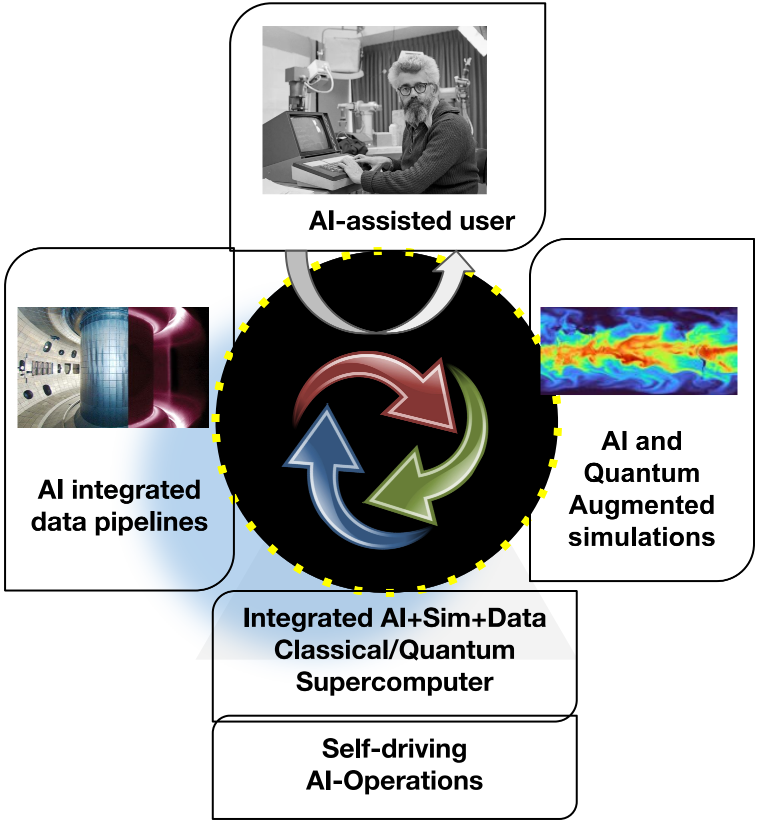




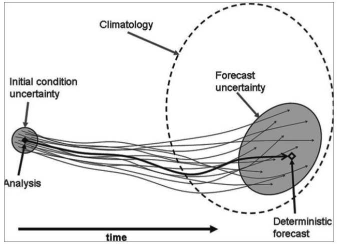
127 Top 500 US Systems

Future NERSC Strategy for Pervasive AI Ecosystem

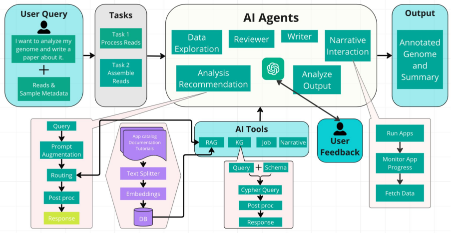
- **System** hardware and software that liberates scientists in application of large AI models
  - Integrated AI+Data+Sim accelerators, workflow and data management reflected in the architecture of NERSC-10 and beyond
  - Highly-instrumented,“Self-driving” systems
- **Service** platform for seamless experimentation and integration of AI with simulation and data
  - Host foundational AI models and datasets
  - Intelligent AI-driven interfaces to compute
- **Applications** for science with large-scale, science-informed, robust, transferable models
- **Ecosystem** to empower scientists to use pervasive AI with human and AI-driven expertise



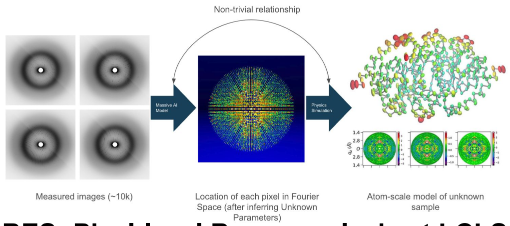
As AI Becomes Pervasive Science will be Transformed



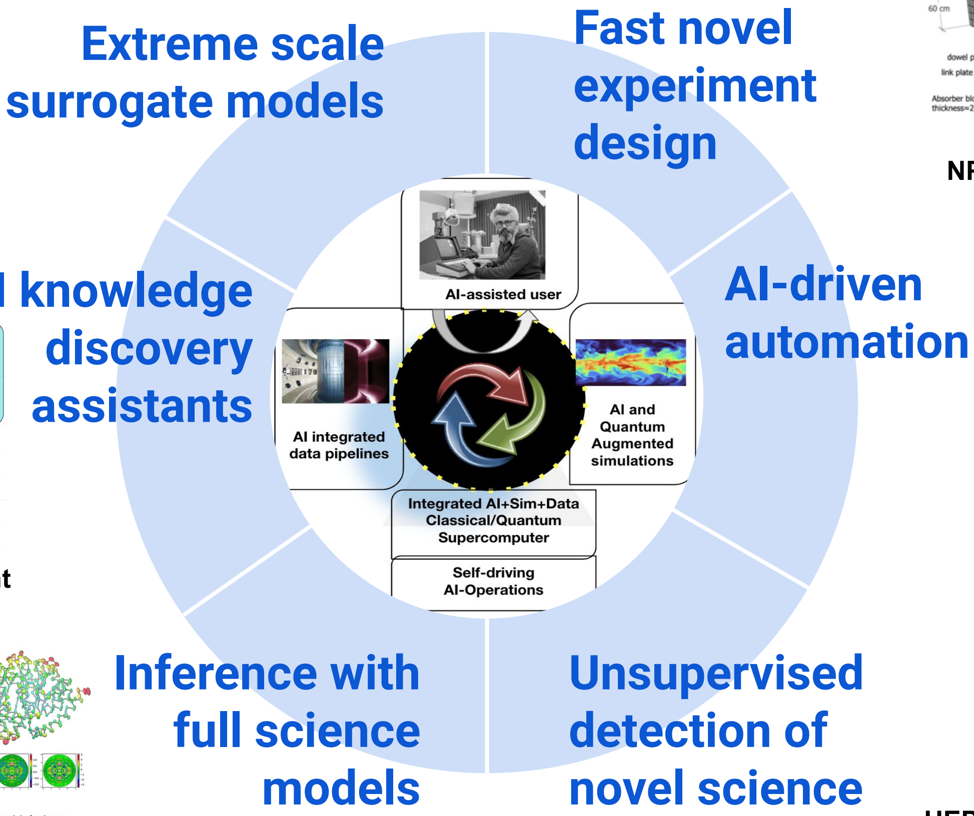
BER: Climate hindcasting with large ensembles

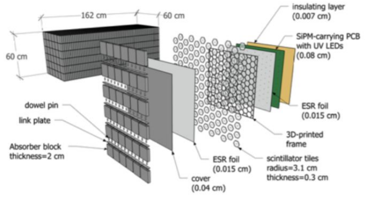


BER: KBase Knowledge Assistant



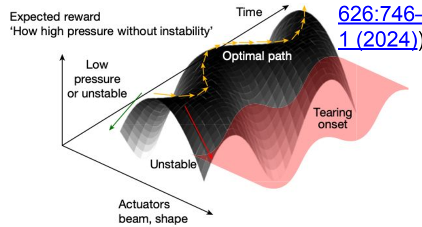
BES: Pixel-level Bragg analysis at LCLS-II



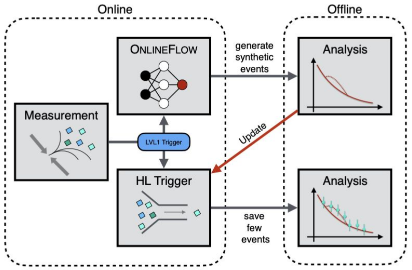


NP: EIC Calorimeter design

(ref: Nature 626:746–751 (2024))

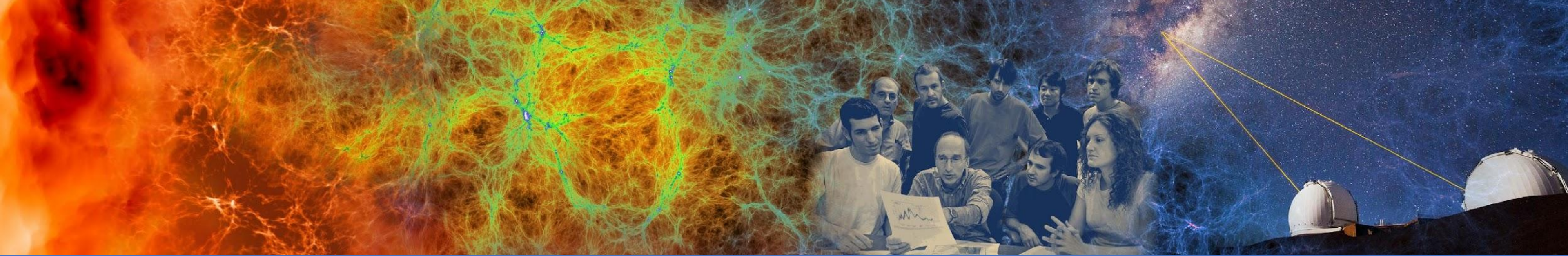


FES: Instability avoidance for ITER



HEP: "Anomaly" detection for HL-LHC

(ref: SciPost Phys. 13, 087 (2022))



Thank you for listening!

For inquiries: SFarrell@lbl.gov