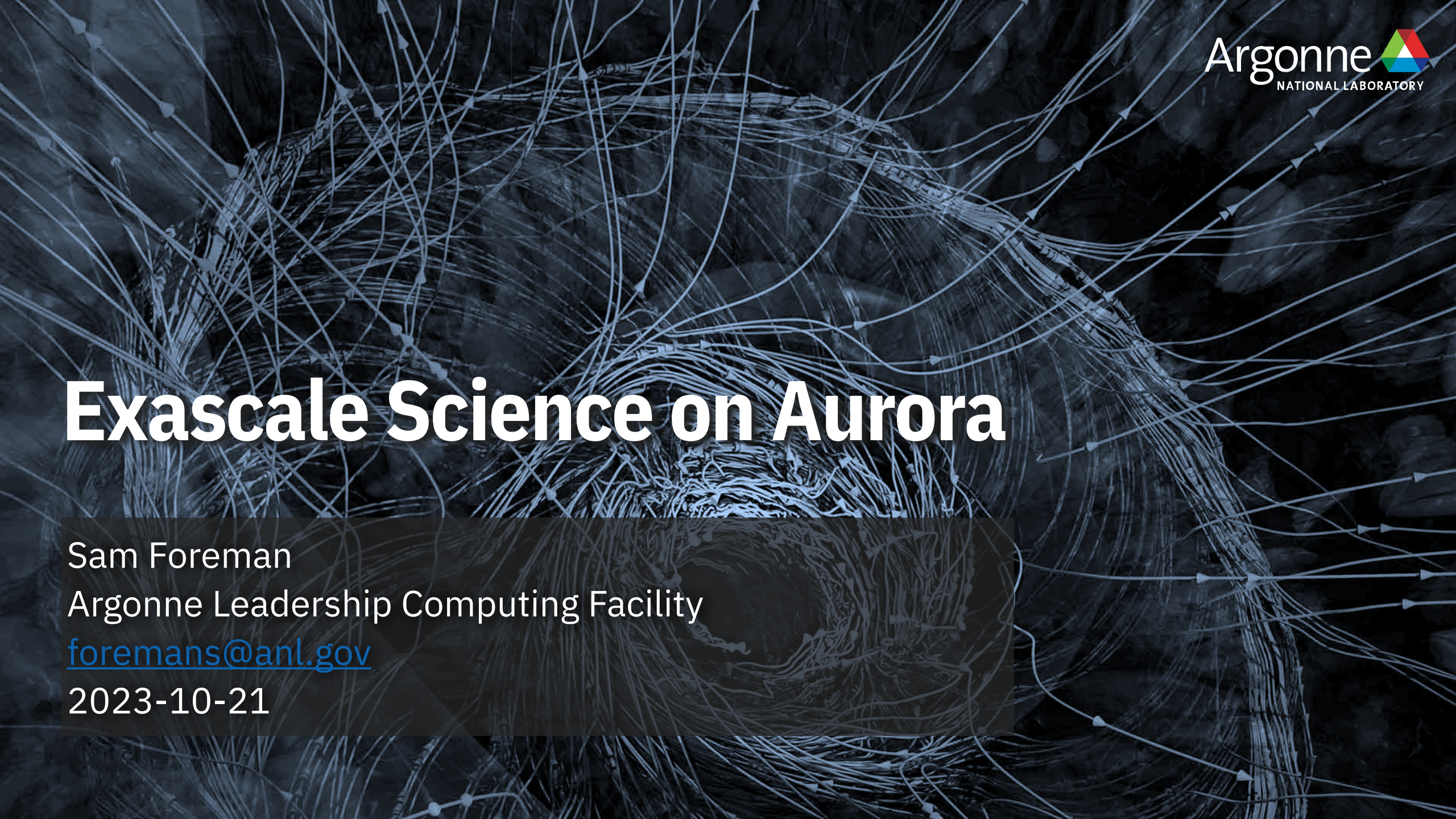


Exascale Science on Aurora

Sam Foreman
Argonne Leadership Computing Facility
foremans@anl.gov
2023-10-21

About Me

- 🏡 samforeman.me
  - Data Science @ ALCF
- Undergrad (2010 — 2015):
  - UIUC:
    - Engineering Physics
    - Applied Mathematics
- Grad School (2015 — 2019):
  - University of Iowa
    - PhD. Physics
    - "A Machine Learning Approach to Lattice Gauge Theory"
- Postdoc (2019-2022) @ ALCF

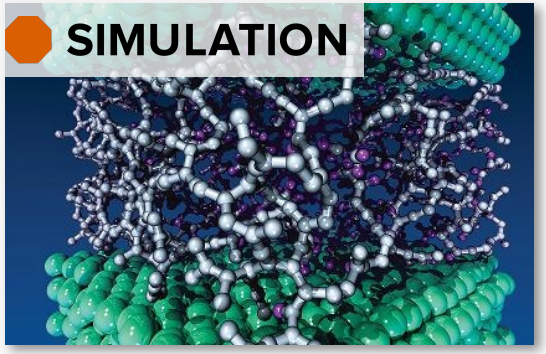
- **Current Research:**
  - AI + Science:
    - *GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics**
    - Building better sampling methods for Lattice QCD
    - Foundation models for long term climate forecasting
  - Scaling Large Language Models
    - Optimizing distibuted training across thousands of GPUs
    - Building new parallelism techniques for efficient scaling
- You can get a live view of some of my recent talks *here*

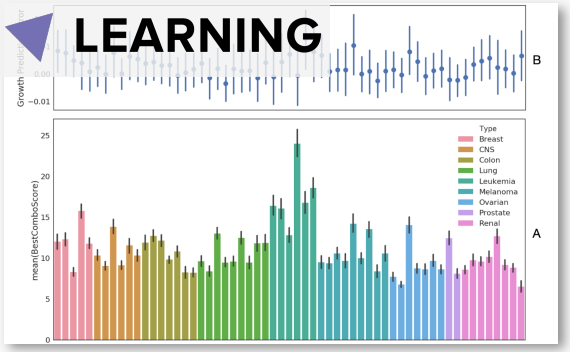
**ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research*

Argonne Leadership Computing Facility (ALCF)

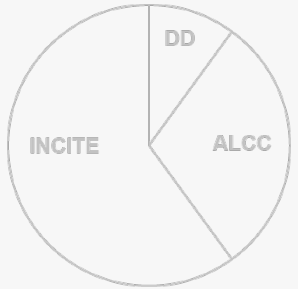
The ALCF provides world-class computing resources to the scientific community.

- Users pursue scientific challenges
- In-house experts help maximize results
- Resources **fully dedicated to open science**





ALCF offers different pipelines based on your computational readiness.



(Apply to the allocation program that fits your needs)

Architecture supports three types of computing:

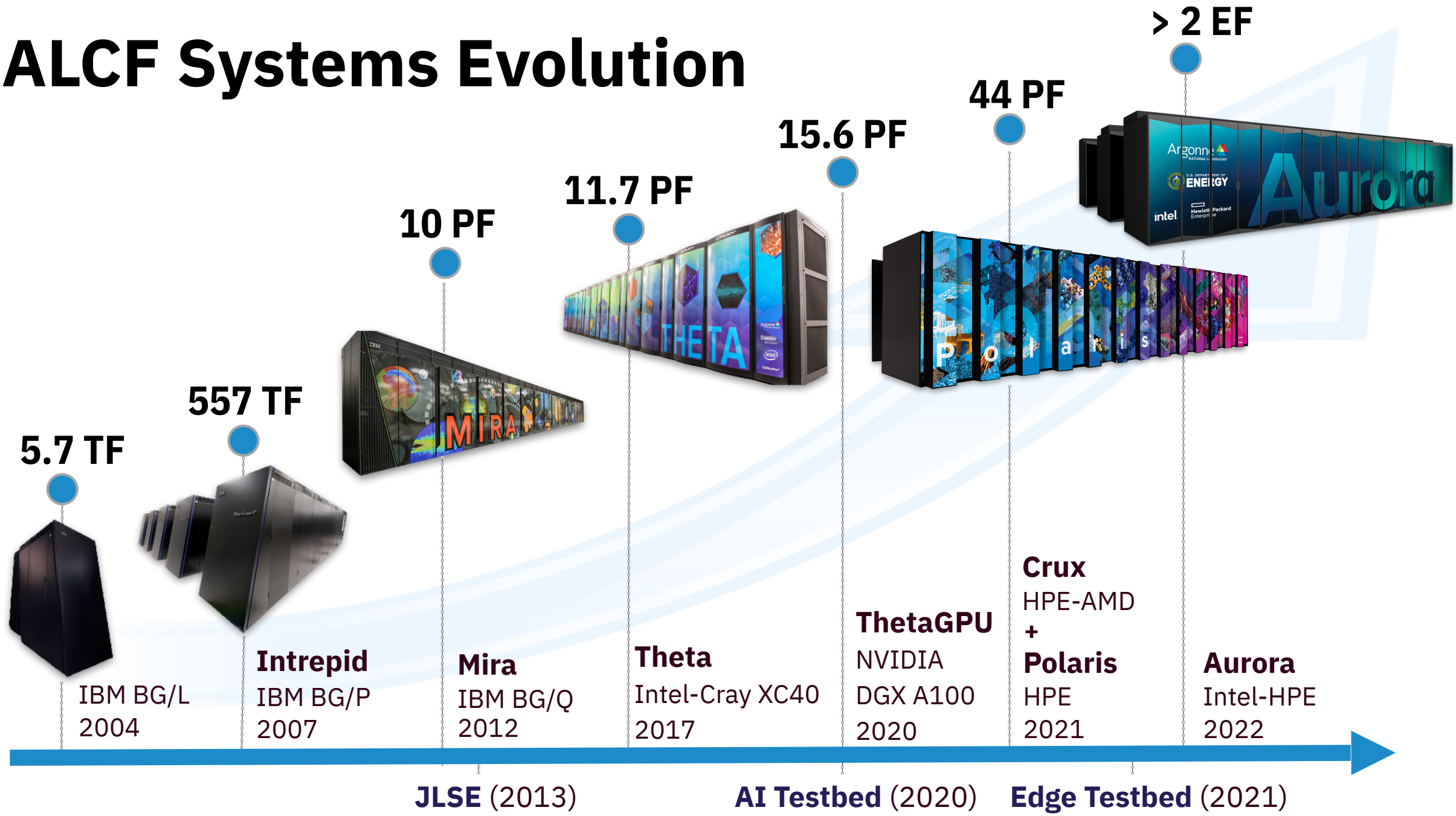
1. **Large-scale Simulation**
   - PDEs, traditional HPC
2. **Data Intensive Applications**
3. **Deep Learning and Emerging Science AI**
   - Training + inference
   - Scalable pipelines (for science)

Argonne NATIONAL LABORATORY

ALCF

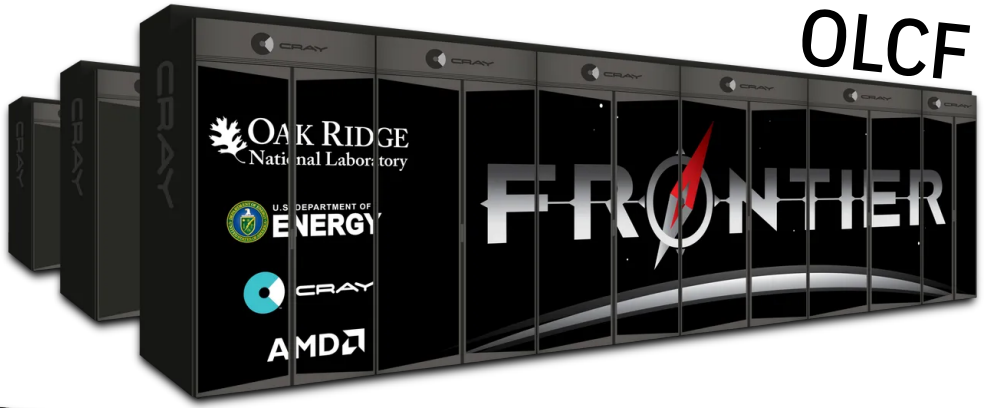
Full Group @ ALCF (+ summer students !!)

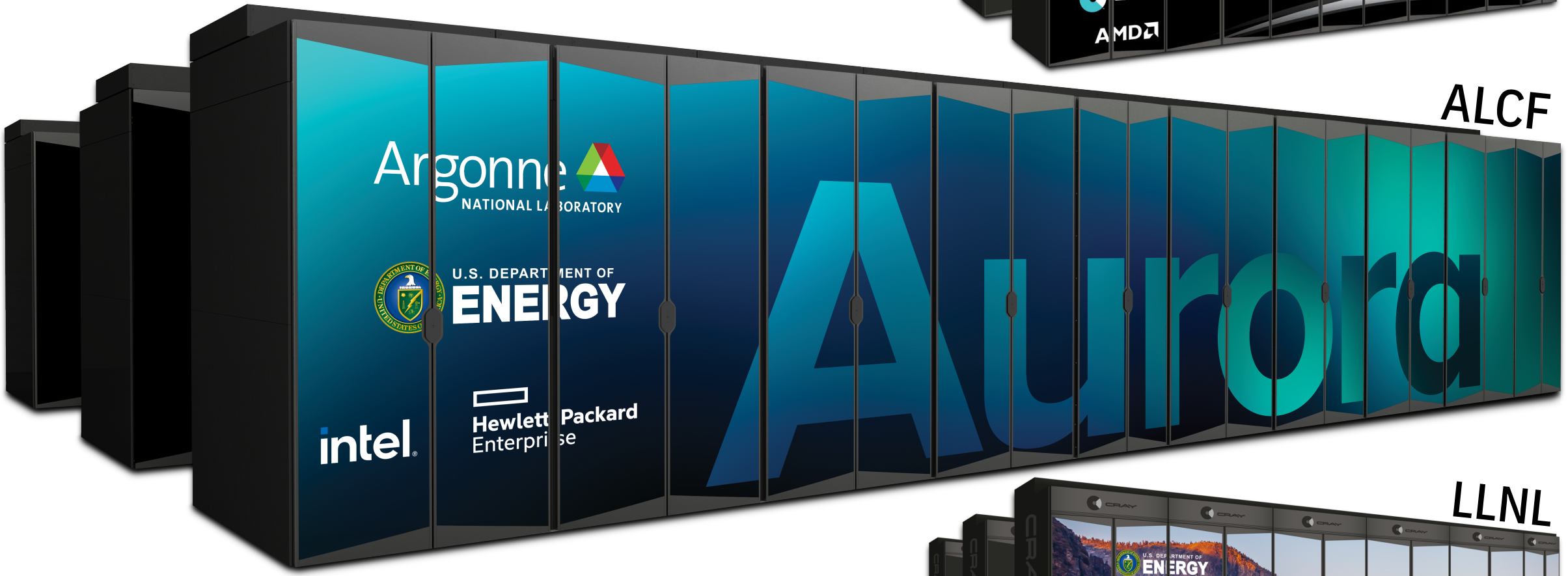
ALCF Systems Evolution



US Exascale Platforms

Intel CPU/GPUs, AMD CPU/GPUs

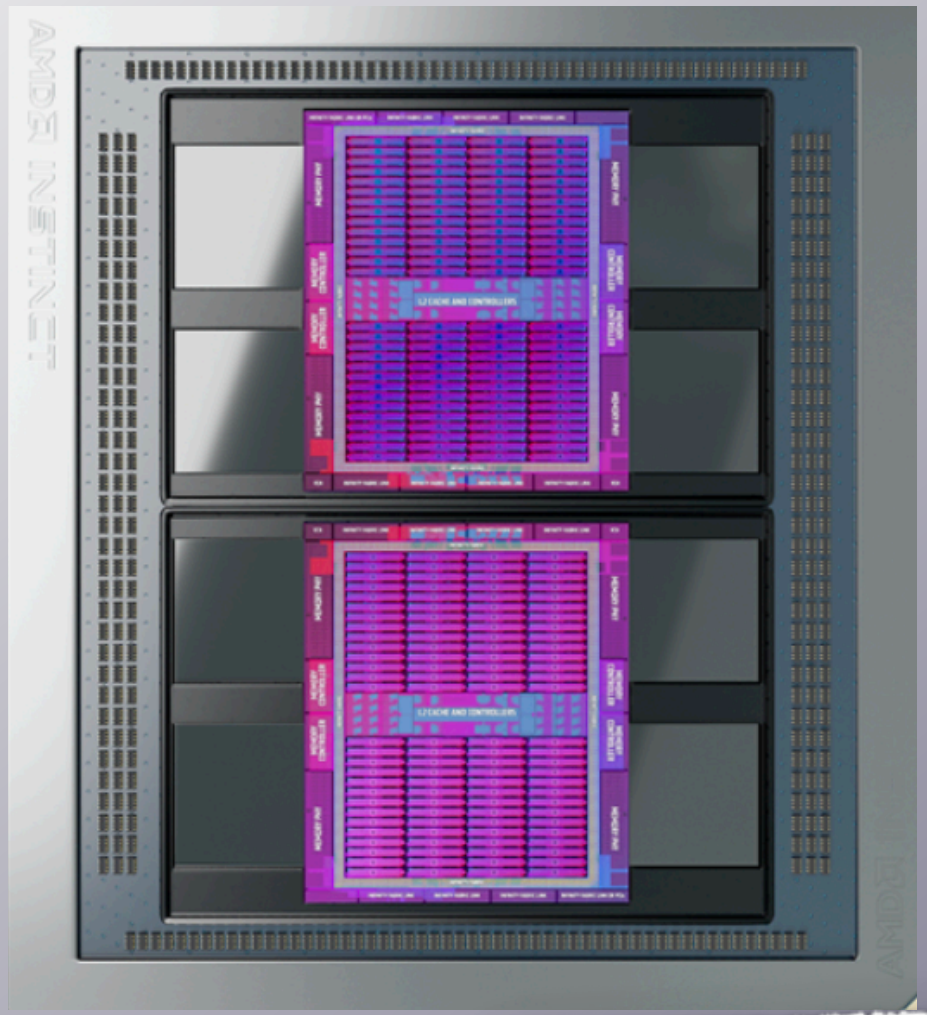


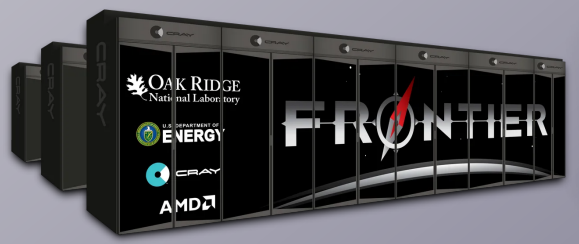


These architectures will excel at **AI + HPC**

>2.3B Transistors

58B transistors on TN6



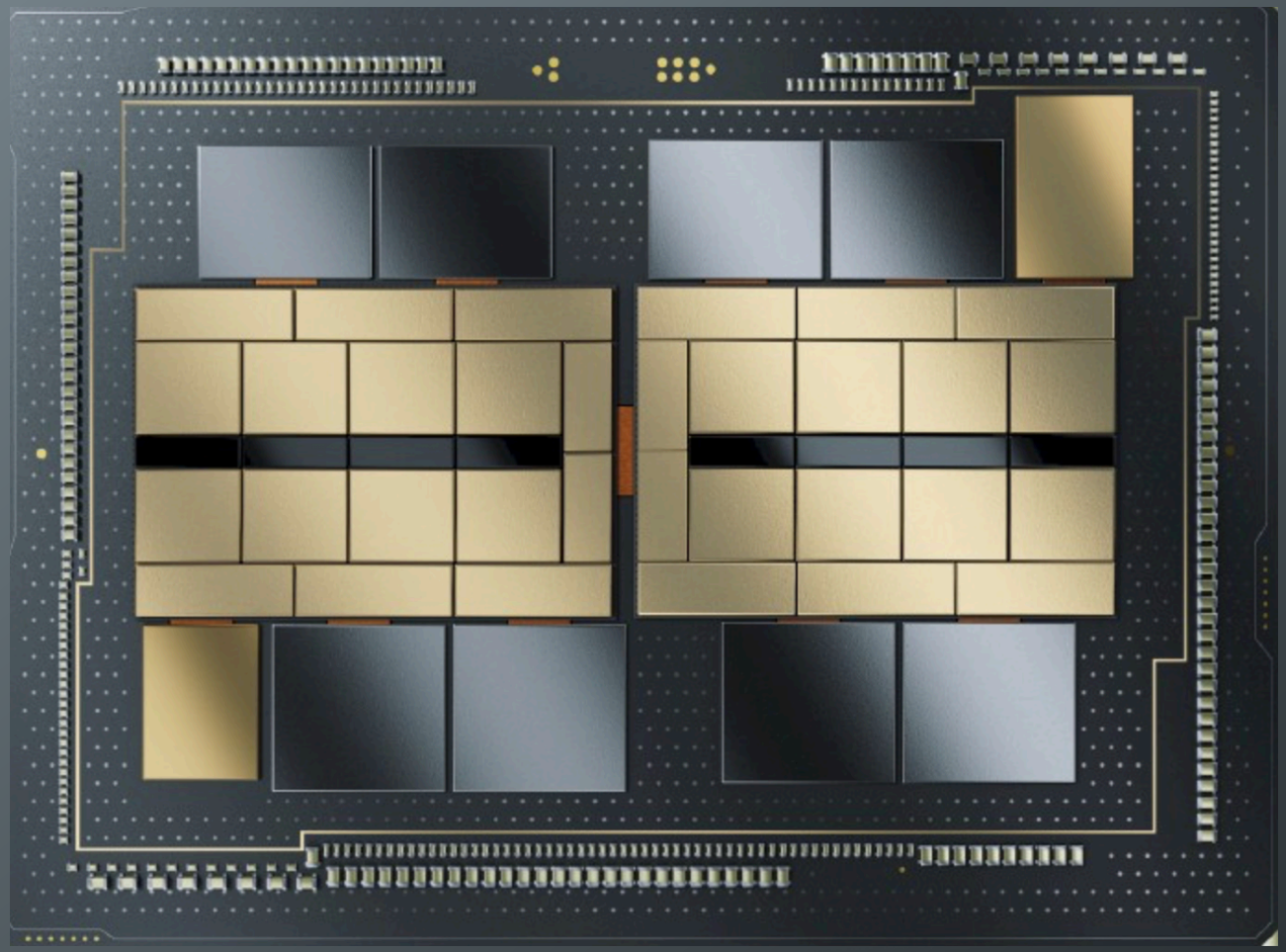


Frontier ≈ 4×10^4

MI-250X (AMD)

>6B Transistors

>100B transistors on TN5, (IN7, IN10)



Aurora ≈ 6×10^4

Xe (PVC) (Intel)

Aurora[1]

Leadership Computing Facility
Exascale Supercomputer

PEAK PERFORMANCE
≧ **2 Exaflops DP**

Intel GPU
Ponte Vecchio

Intel Xeon PROCESSOR
Sapphire Rapids + HBM

PLATFORM
HPE Cray-Ex

Compute Node
2 SPR+HBM processor;
6 PVC; Unified Memory Architecture;
8 fabric endpoints;

GPU Architecture
Xe arch-based “Ponte Vecchio” GPU
Tile-based chiplets
HBM stack
Foveros 3D integration

System Interconnect
HPE Slingshot 11; Dragonfly topology with adaptive routing

Network Switch
25.6 Tb/s per switch, from 64–200 Gb/s ports (25 GB/s per direction)

Node Performance
>130 TF

System Size
>9,000 nodes

Aggregate System Memory
>10 PB aggregate System Memory

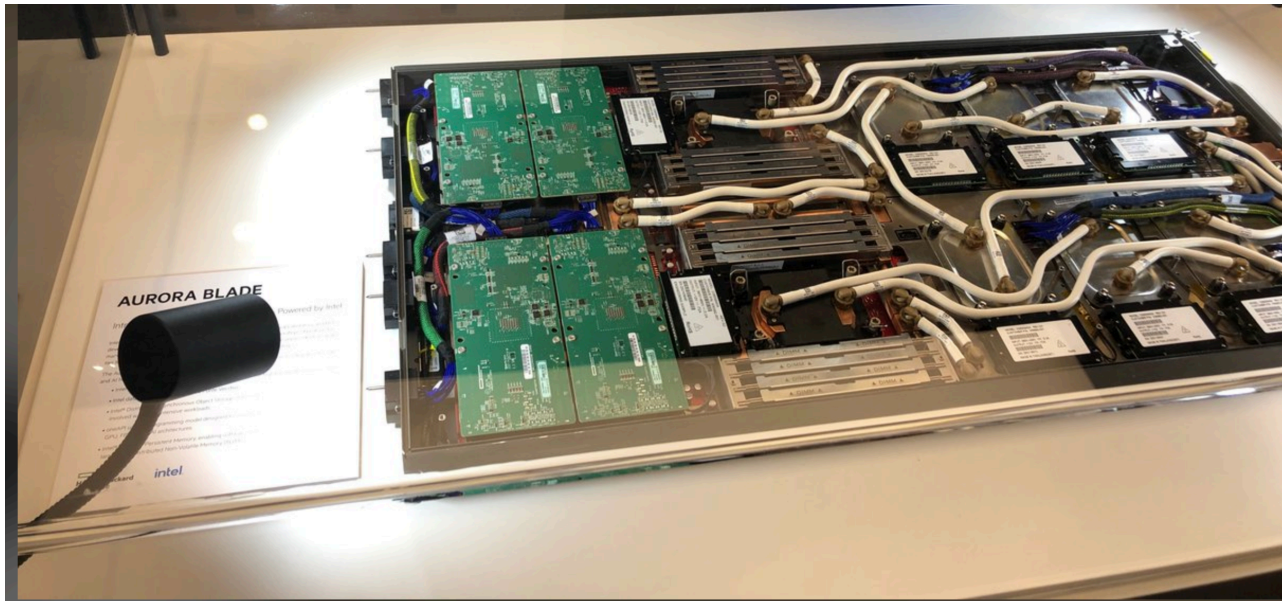
High-Performance Storage
220 PB @ EC16+2, ≧**25 TB/s**
DAOS

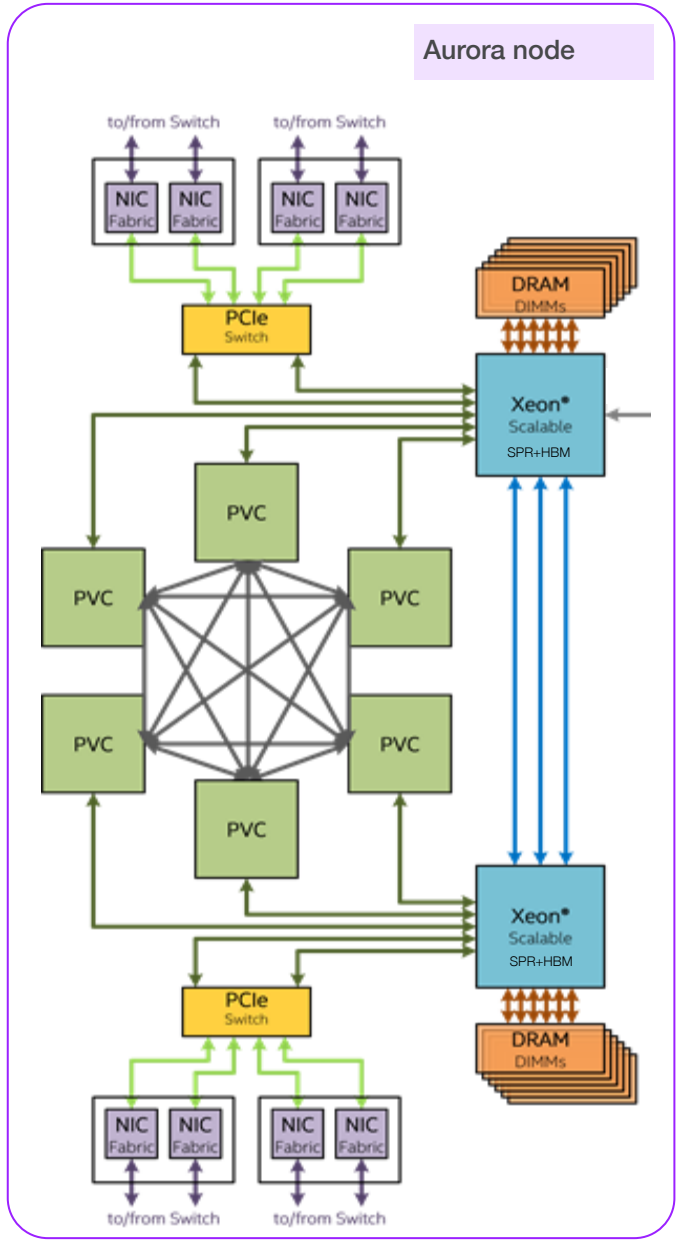
Programming Models
oneAPI, MPI, OpenMP, C/C++, Fortran, SYCL/DPC++
Python-based environments
ML + Deep learning frameworks

1: *The Computer That WIll Change Everything*

Aurora Compute Node

- 6 Xe Architecture based GPUs (Ponte Vecchio)
  - All to all connection
- 2 Intel Xeon (Sapphire Rapids) processors
- Unified Memory Architecture across CPUs and GPUs
- 8 Slingshot Fabric endpoints





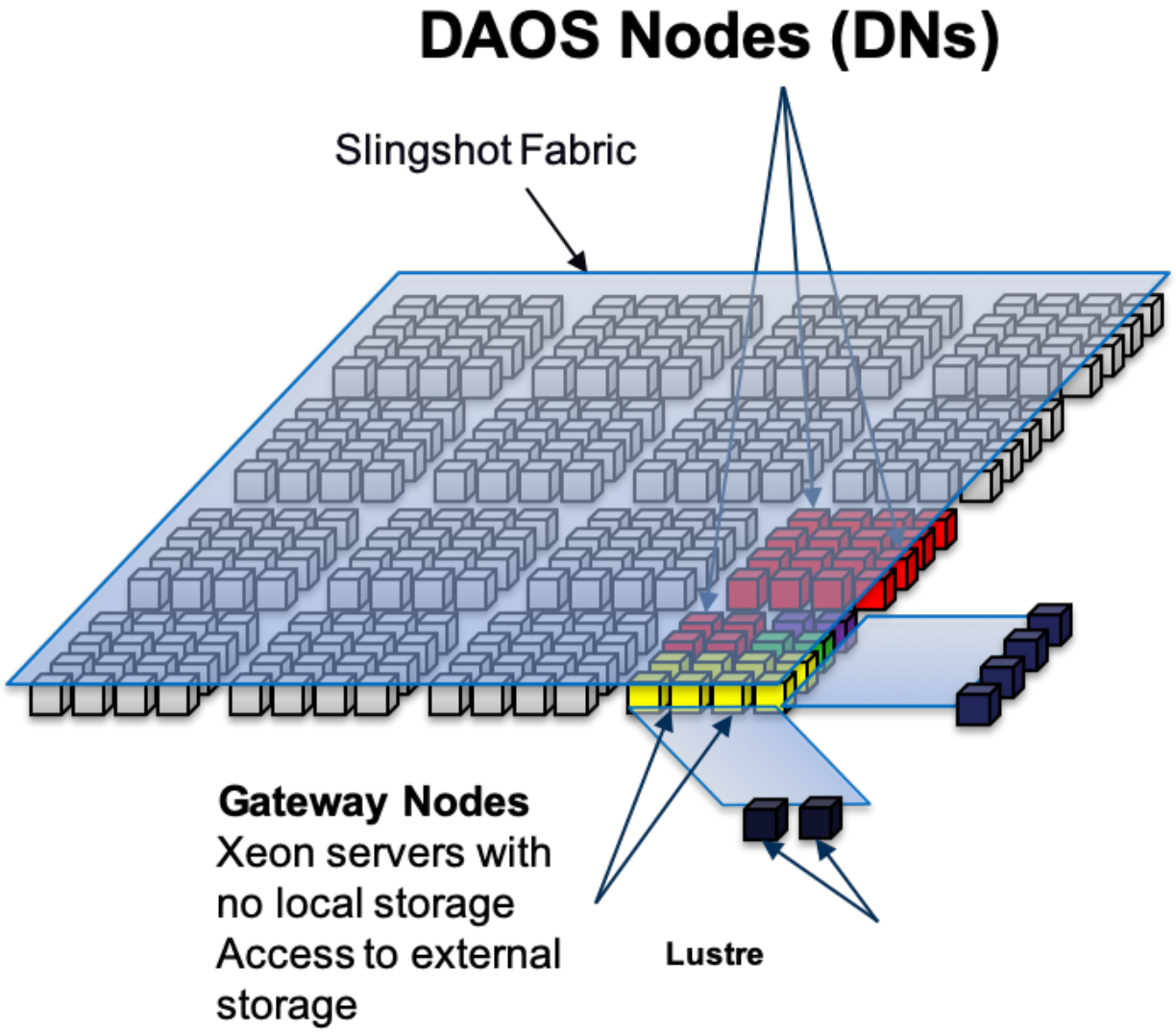
Aurora Cabinets Installed at Argonne

The Aurora Supercomputer at ALCF


Aurora
> 60,000 Intel GPUs
Science Starts in 2023

Distributed Asynchronous Object Store (DAOS)

- Primary storage system for Aurora
- Offers high performance in bandwidth and IO operations
  - **230 PB capacity**
  - **≥ 25 TB/s**
- Provides a flexible storage API that enables new I/O paradigms
- Provides compatibility with existing I/O models such as POSIX, MPI-IO and HDF5
- Open source storage solution



oneAPI

- Industry specification from Intel
  - Language and libraries to target programming across diverse architectures (DPC++, APIs, low level interface)
- Intel oneAPI products and toolkits
  - **Languages**
    - Fortran (w/ OpenMP 5+)
    - C/C++ (w/ OpenMP 5+)
    - DPC++
    - Python
  - **Libraries**
    - oneAPI MKL (oneMKL)
    - oneAPI Deep Neural Network Library (oneDNN)
    - oneAPI Data Analytics Library (oneDAL)
    - MPI
  - **Tools**
    - Intel Advisor
    - Intel VTune
    - Intel Inspector

https://software.intel.com/oneapi

Available Aurora Programming Models

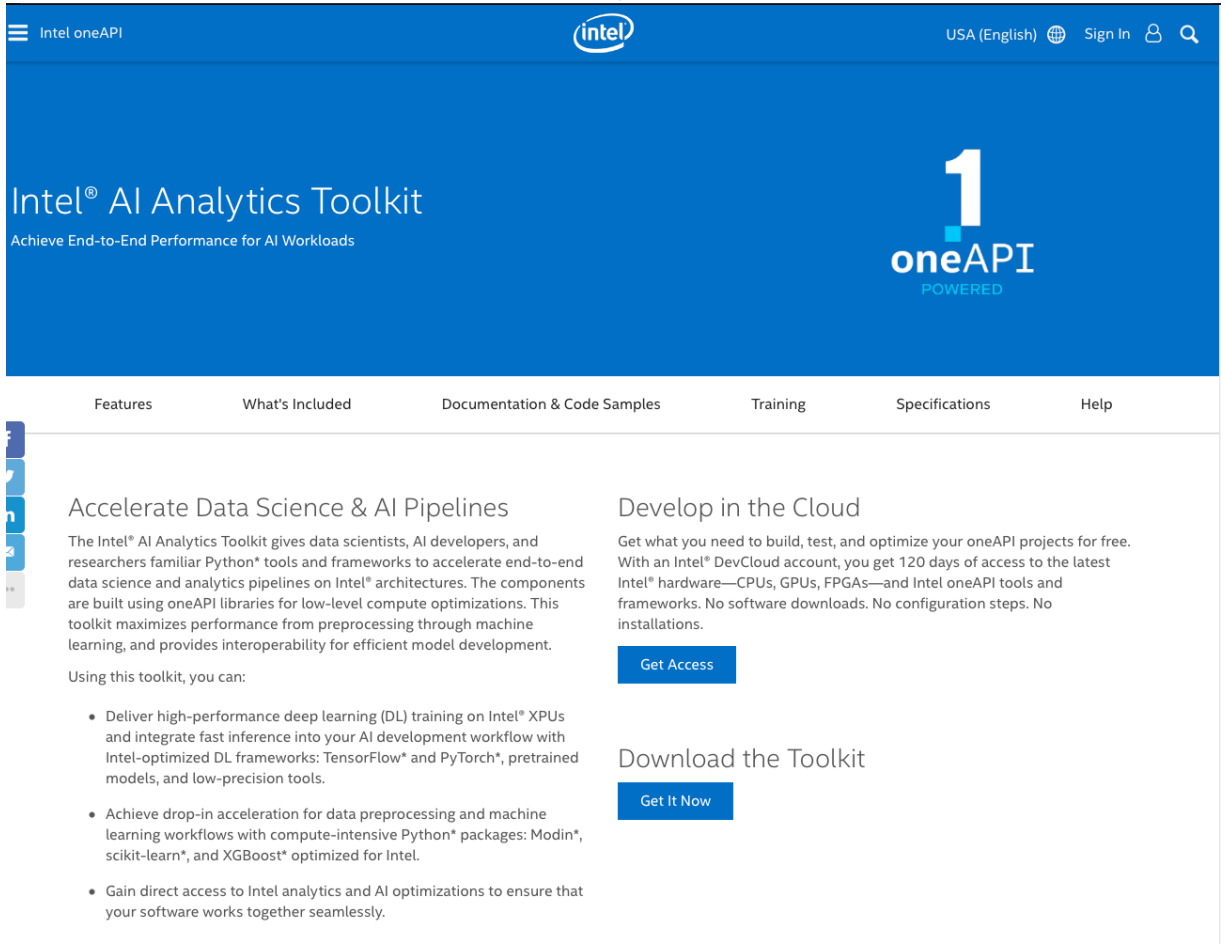
- Aurora applications may use:
  - DPC++/SYCL
  - OpenMP
  - Kokkos
  - Raja
  - OpenCL
- Experimental
  - HIP
- Not available on Aurora:
  - ~~CUDA~~
  - ~~OpenACC~~

Data Science and Learning on Aurora

Aurora Will Provide for a Familiar, Productive and Performant HPC and AI Software Stack

- Python Ecosystem:
  - `Numba, NumPy, etc.`
- **ML + Deep Learning Frameworks**:
  - OneDAL, scikit-learn, XGBoost, etc.
  - `PyTorch`, `TensorFlow`, `Horovod`, `DDP`, `DeepSpeed`, 🤗 `Accelerate`
- *Optimized + scalable communication* with OneCCL
- Spark BigData Analytics
- DAOS for fast I/O + workflows
- Profiling and debugging tools

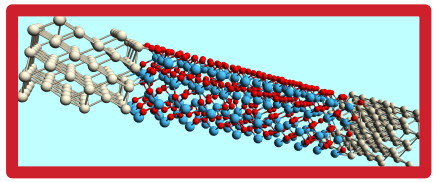
Intel AI Analytics Toolkit

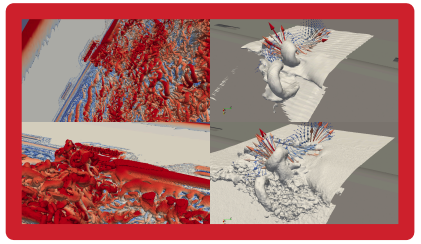


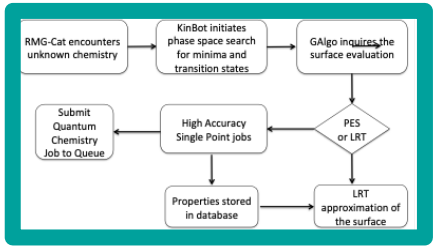
https://software.intel.com/content/www/us/en/develop/tools/oneapi/ai-analytics-toolkit.html

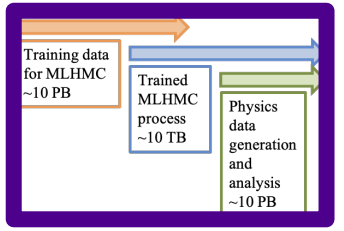
Aurora ESP Projects

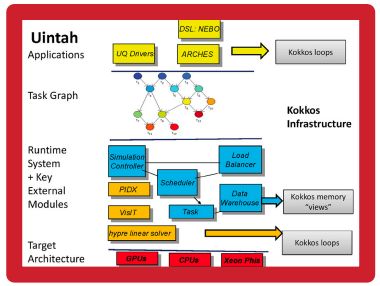
S D L

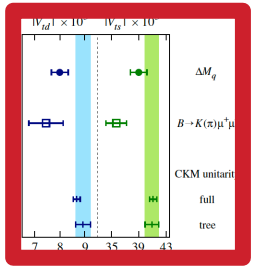


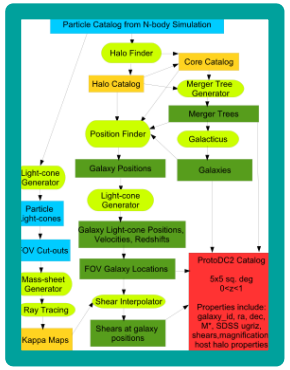


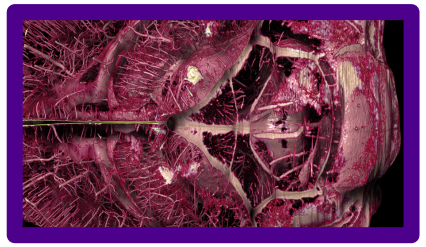


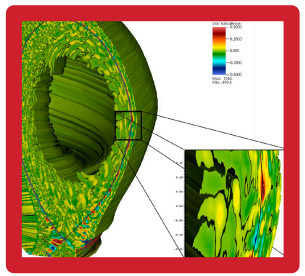


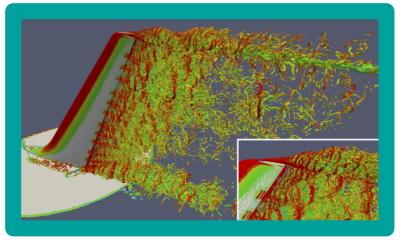


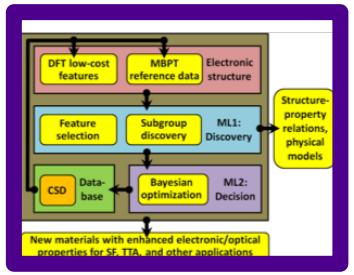


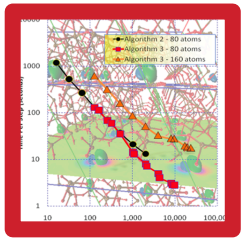


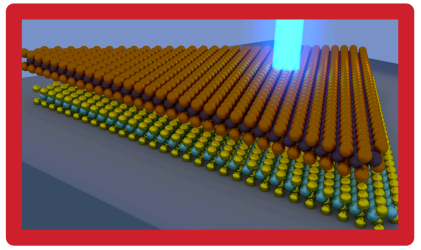


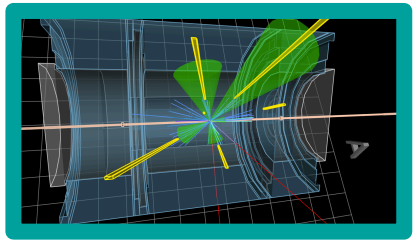


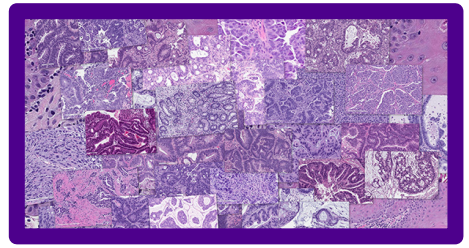


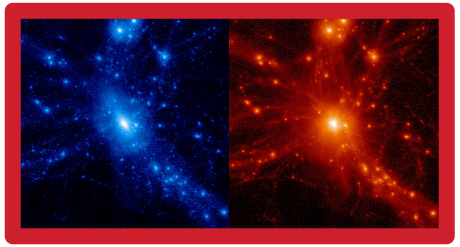


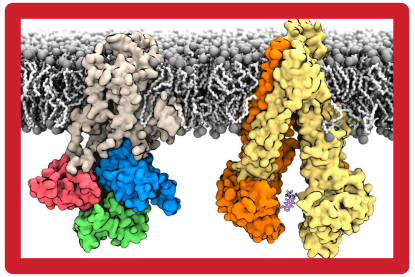


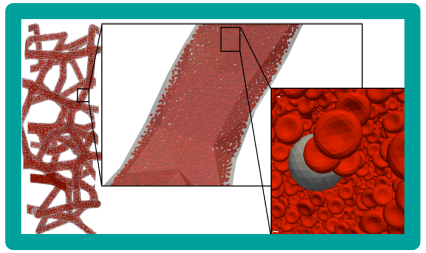


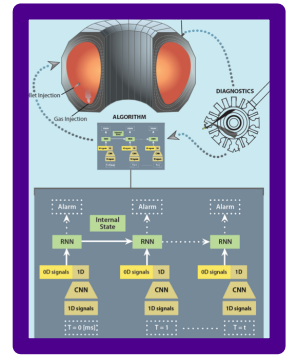












Argonne NATIONAL LABORATORY

Aurora ESP Projects

S D L

Anouar Benali (ANL)

Extending Moore's Law computing with Quantum Monte Carlo

Martin Berzins (U. Utah)

Design & evaluation of high-efficiency boilers for energy production using a hierarchical V/UQ approach

CS Chang (PPPL)

High fidelity simulation of fusion reactor boundary plasmas

Theresa Windus (Ames)

NWChemEx: Tackling Chemical, Materials & Biochemical Challenges in the Exascale Era

Katrin Heitmann (ANL)

Extreme-Scale Cosmological Hydrodynamics

Ken Jansen (U. Colorado)

Extreme Scale Unstructured Adaptive CFD: From Multiphase Flow to Aerodynamic Flow Control

Norman Christ (Columbia)

Lattice Quantum Chromodynamics Calculations for Particle and Nuclear Physics

Aiichiro Nakano (USC)

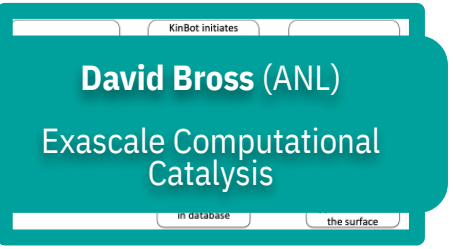
Metascalable Layered Materials Genome

Benoit Roux (U. Chicago)

Free Energy Landscapes of Membrane Transport Proteins

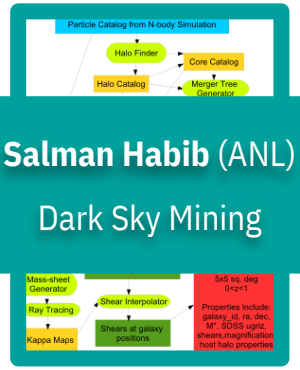
David Bross (ANL)

Exascale Computational Catalysis



Salman Habib (ANL)

Dark Sky Mining



Ken Jansen (U. Colorado)

Data Analytics and Machine Learning for Exascale CFD

Walter Hopkins(ANL)

Simulating and Learning in the ATLAS detector at the Exascale

Amanda Randles (Duke U.)

Extreme-scale In Situ Visualization and Analysis of Fluid-Structure-Interaction Simulations

Will Detmold (MIT)

Machine Learning for Lattice Quantum Chromodynamics

Nicola Ferrier (ANL)

Enabling Connectomics at Exascale to Facilitate Discoveries in Neuroscience

Noa Marom (CMU)

Many-Body Perturbation Theory Meets Machine Learning to Discover Singlet Fission Materials

Rick Stevens (ANL)

Virtual Drug Response Prediction

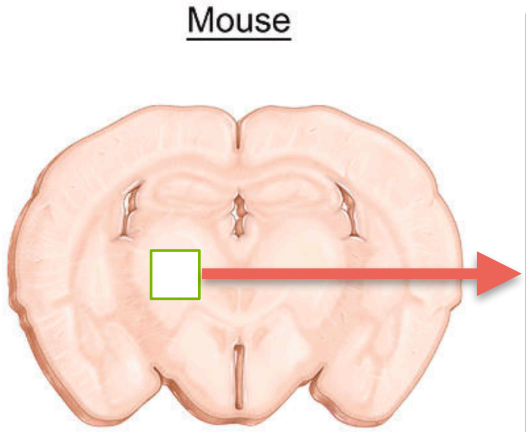
Bill Tang (Princeton)

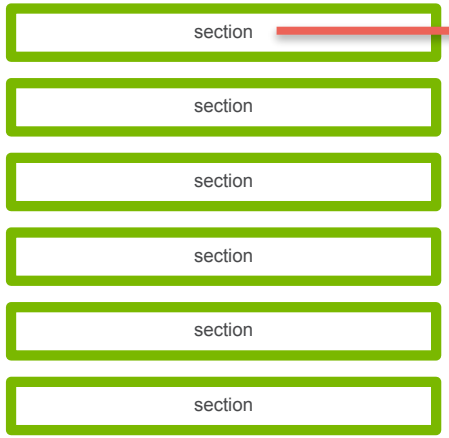
Accelerated Deep Learning Discovery in Fusion Energy Science

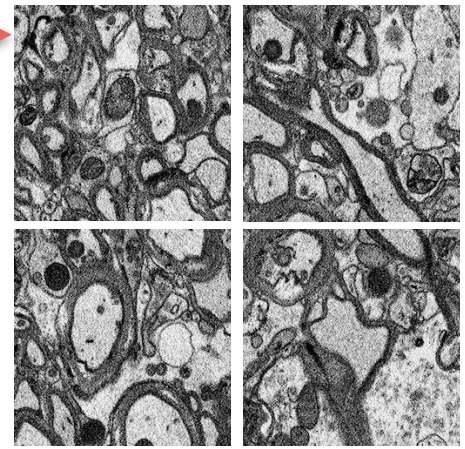
Argonne NATIONAL LABORATORY

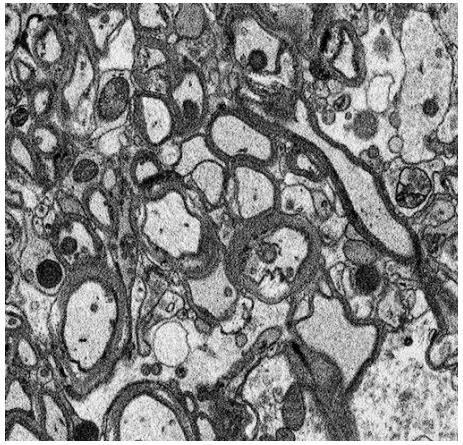
Data from Gregg Wildenberg, Kasthuri Lab, UChicago

Connectomics Data-Driven Models









~1cm^3

Mouse brain: 70M neurons

~1mm^3

25000
40nm sections
1mm x 1mm
(6nm resolution)

Each section
imaged with EM as
N tiles (8 bit)
80K x 40K pixels

Sections
stitched
together

How much image data is 1mm³ ? 1×10^{15} voxels --> ~1 PB

Data Challenges in Connectomics

Mouse brain: 70M neurons

~1cm³

How much image data is 1cm³ ? **~1EB**

Human brain: 80B neurons

~1000cm³

How much image data is 1000cm³ ? **~1000 EB** (6nm x 6nm x 40nm)

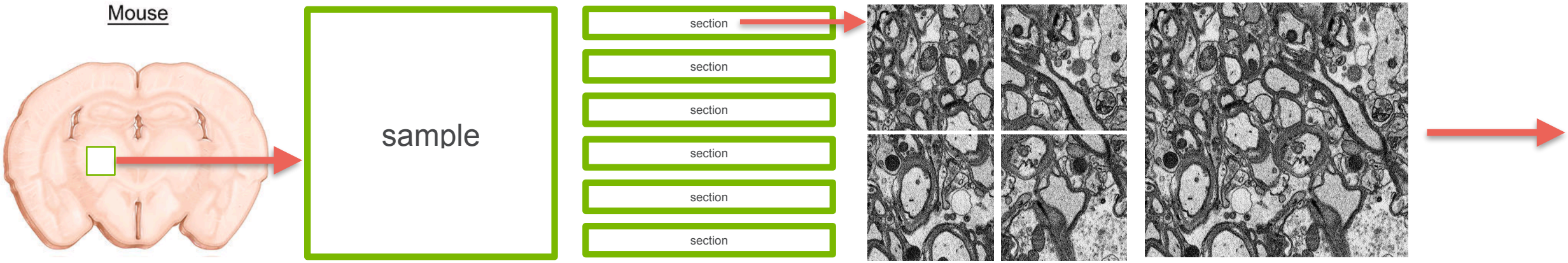
Reconstructed data will be much larger:

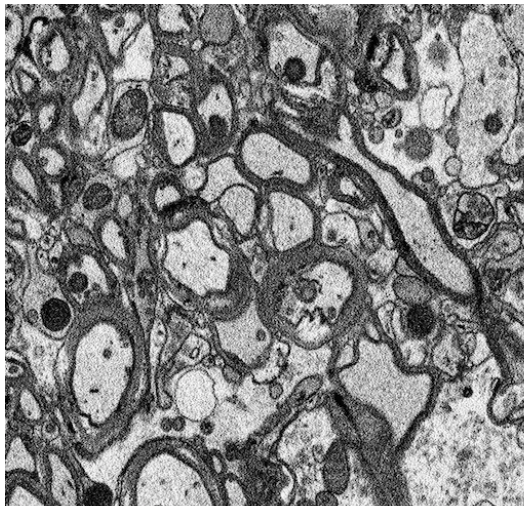
- Segmentation labels for each voxel
  - 4x voxel data
- 3D Mesh
- Skeleton

The structures are expected to be used to seed simulations to study flow in neurotransmitters, in better modeling the brain, among others.

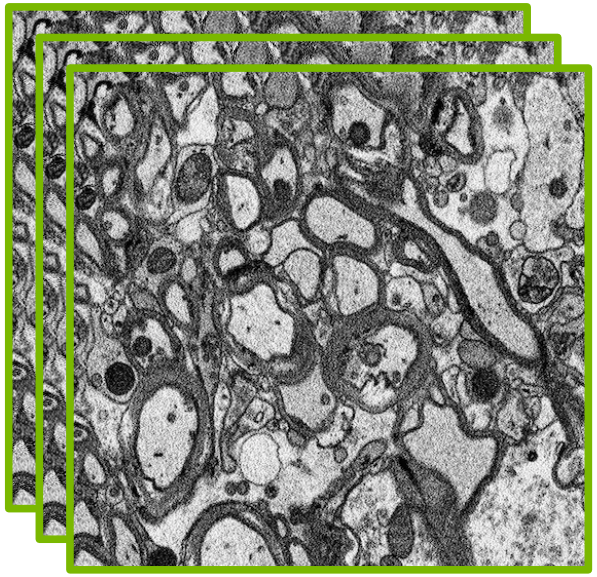
Connectomics Processing

Data from Gregg Wildenberg, Kasthuri Lab, UChicago

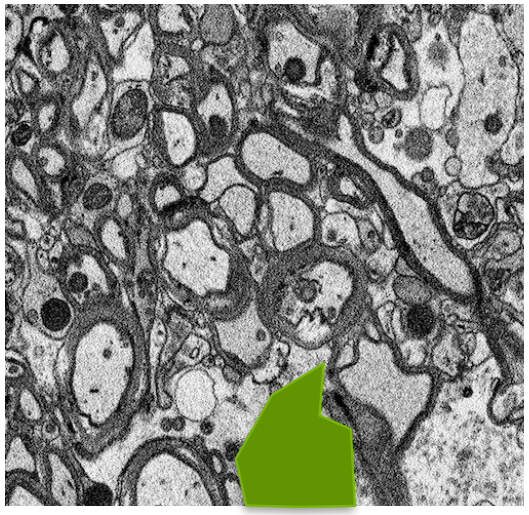




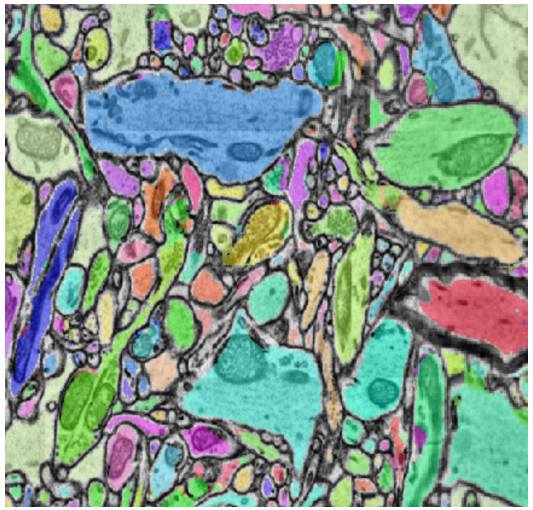
Sections stitched together



Align sections



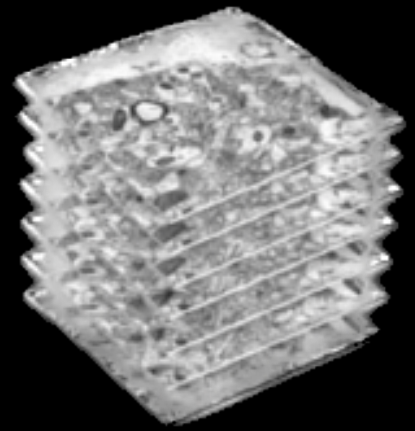
Mask out non-target objects



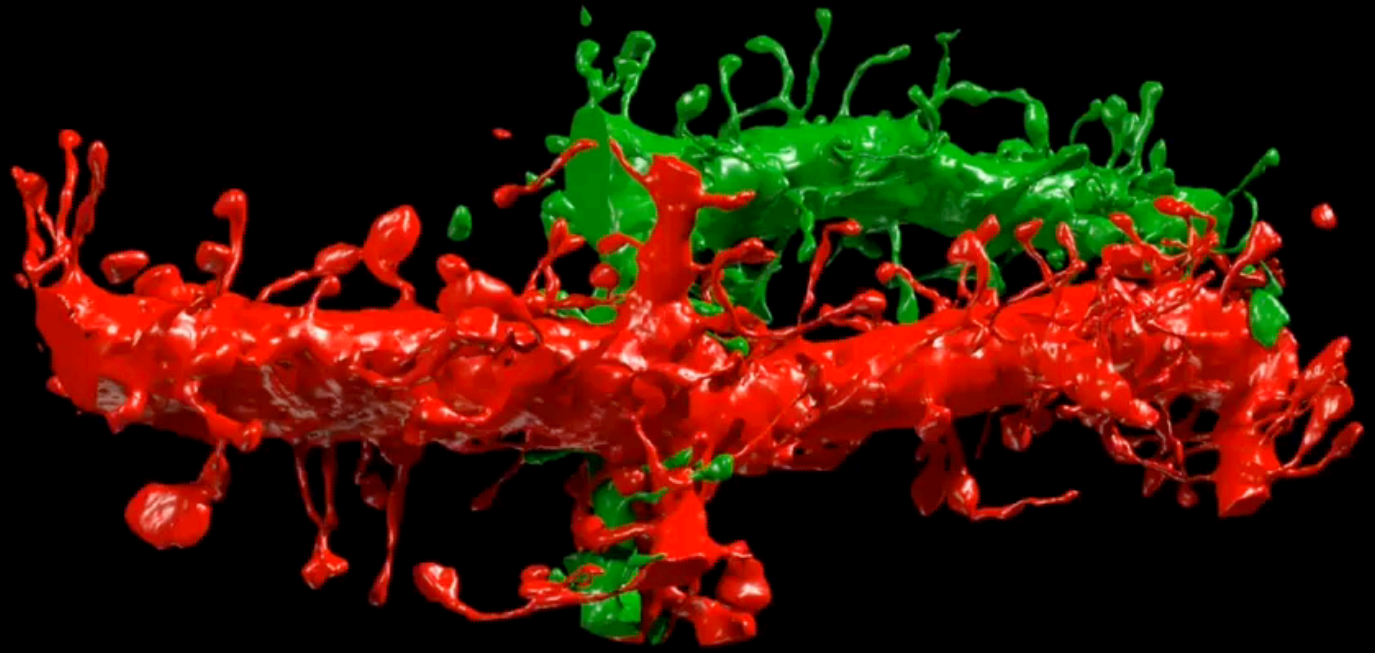
Segment target objects

Reconstructing the Brain Connectivity

Kasthuri et al, Cell 2015

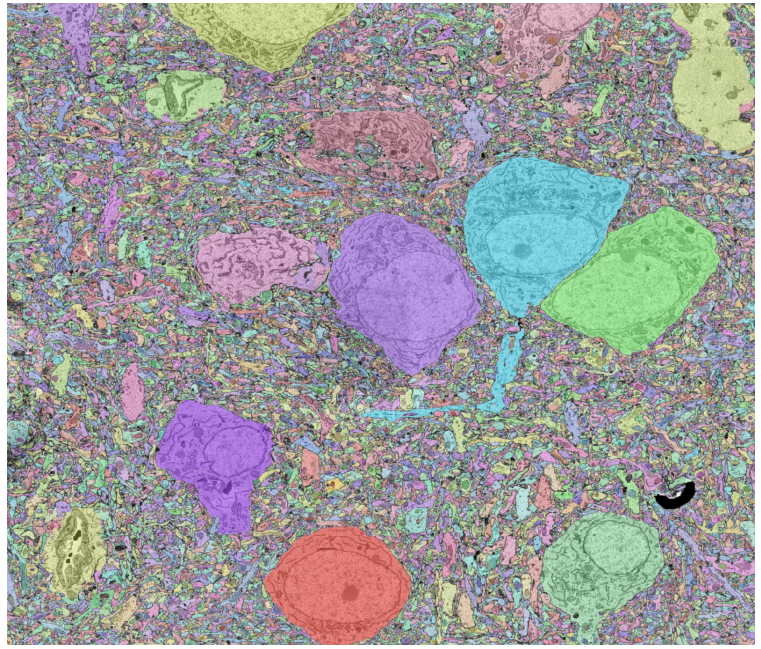


EM Images



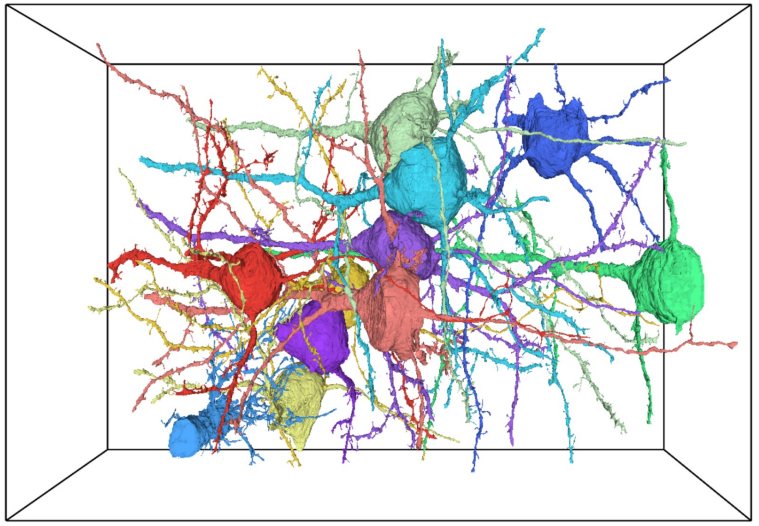
Large-Scale Reconstruction

- Inference (and training) has scaled on CPU-based and GPU-based supercomputers (parallel granularity: overlapping subvolumes)
  - –Achieved million-way concurrency on Theta supercomputer
- Image stitching and alignment components are being scaled as well to ensure a scalable end-to-end pipeline



Exascale Inference Problem:

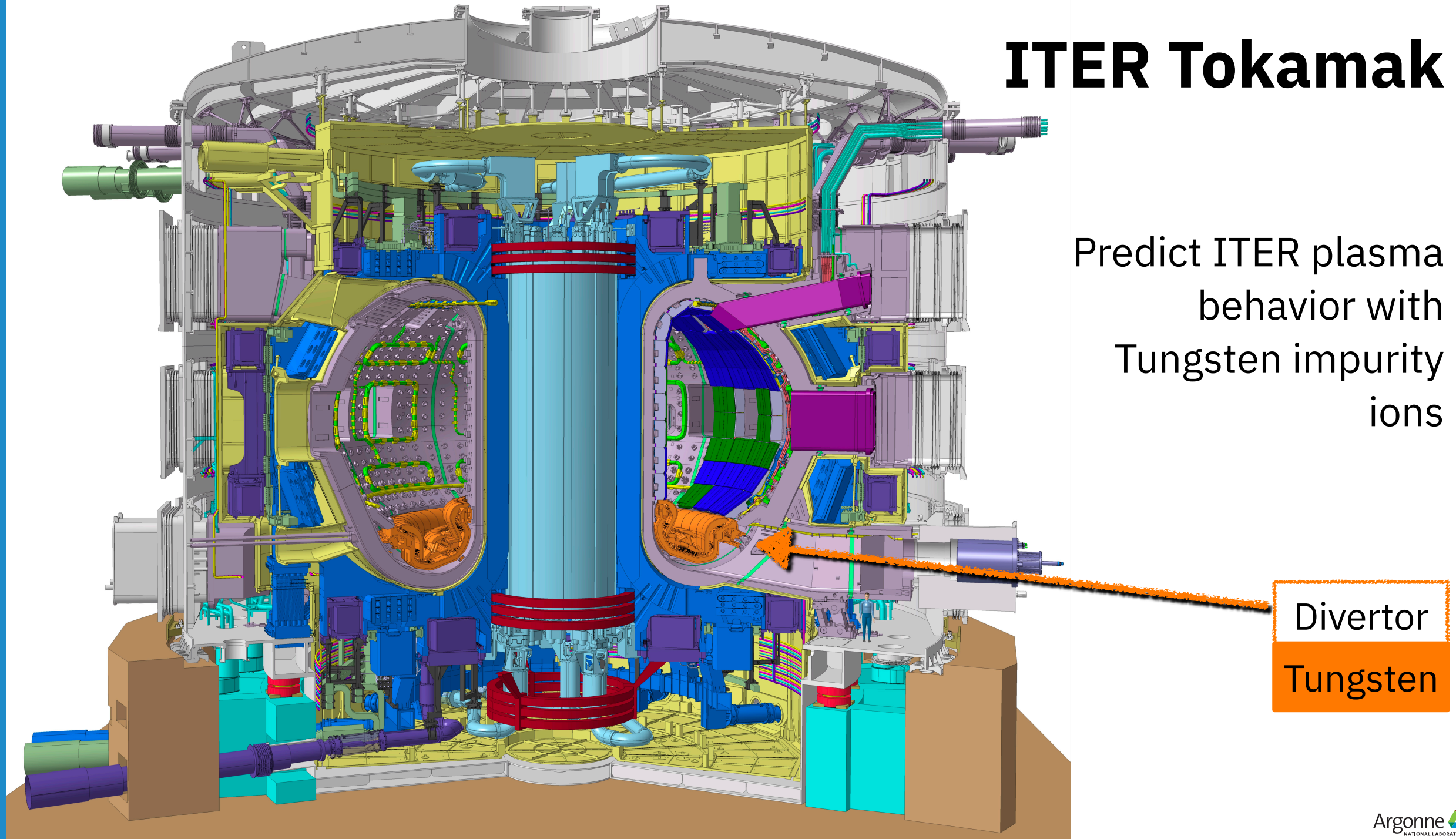
- On a single GPU (A100), we achieve ~80 MegaVoxels/hour using 32-bit (There is still room for improvement here)
- In reduced precision (8-16 bits), we expect ~1 GigaVoxel/hour per GPU
- 1 PetaVoxel (1mm^3) will take ~1M GPU node hours
- Approximately, **24 hours on a system with 50K GPUs** (considering overlapping sub-volumes)
- For a mouse brain (1cm^3), 1 ExaVoxel, we would need **~3 years on an exascale system**



Dong, et al, “Scaling Distributed Training of Flood-Filling Networks on HPC Infrastructure for Brain Mapping”, 2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS) at SC19

Vescovi, et al, “Toward an Automated HPC Pipeline for Processing Large Scale Electron Microscopy Data”, 2020 IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP) at SC19

ITER Tokamak



Predict ITER plasma behavior with Tungsten impurity ions

Showcase

ExaSMR: NekRS Performance on Ponte Vecchio

Ponte Vecchio with Intel OneAPI DPC++ implementation

1.5x performance lead

ExaSMR: *Small modular reactors (SMRs) and advanced reactor concepts (ARCs) will deliver clean, flexible, reliable, and affordable electricity while avoiding the traditional limitations of large nuclear reactor designs,*
https://www.exascaleproject.org/research-project/exasmr/

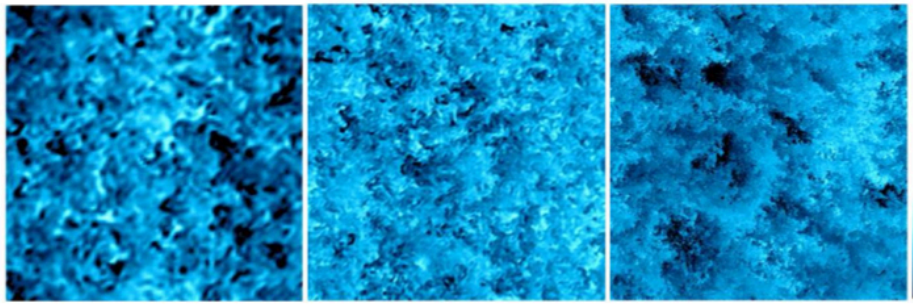
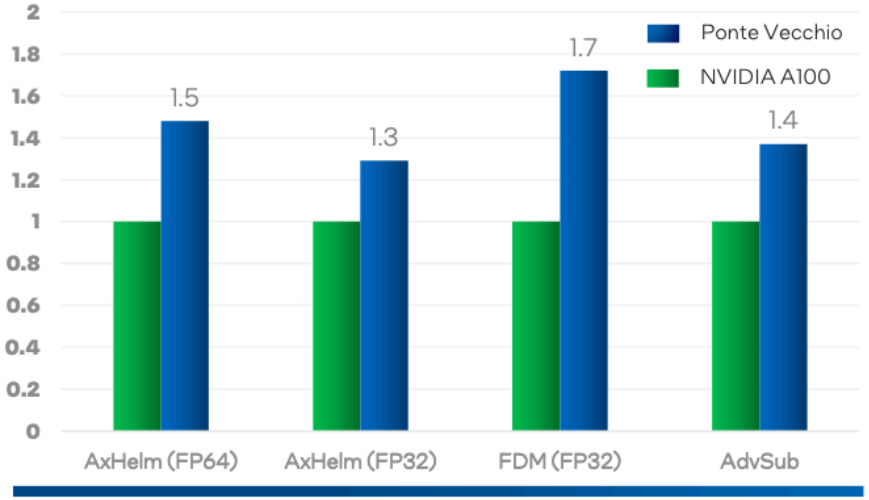


Figure 10: NekRS: potential temperature distributions in [K] at time 6h and z=100m on different resolutions of Δx= 3.12m (left), 1.56m (center), and 0.78m (right) corresponding to the number of grid points, n=128^3, 256^3, and 512^3, respectively. Δx represents the average grid-spacing for the spectral elements, $E = 16^3$, 32^3 and 64^3 and the polynomial order $N = 8$ on the domain 400m×400m×400m.

https://ceed.exascaleproject.org/docs/ceed-ms38-report.pdf

Relative Performance of NekRS Benchmarks w/ problem size of 8196 (Averaged throughput, higher is better)



Application Summary:

NekRS is an open-source Navier Stokes solver based on the spectral element method targeting classical processors and accelerators like GPUs. The code started as a fork of libParanumal in 2019. For API portable programming OCCA is used.
https://github.com/argonne-lcf/nekRS/

OCCA is an open-source library which aims to make it easy to program different types of devices (e.g. CPU, GPU, FPGA). It provides a unified API for interacting with backend device APIs (e.g. OpenMP, CUDA, OpenCL), uses just-in-time compilation to build backend kernel, and provide a kernel language, a minor extension to C, to abstract programming for each backend.
https://libocca.org

HOT CHIPS

- See backup for workloads and configurations. Results may vary.
- Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

intel.

H. Jiang, "Intel's Ponte Vecchio GPU : Architecture, Systems & Software," *2022 IEEE Hot Chips 34 Symposium (HCS)*, 2022, pp. 1-29, doi: 10.1109/HCS55958.2022.9895631.

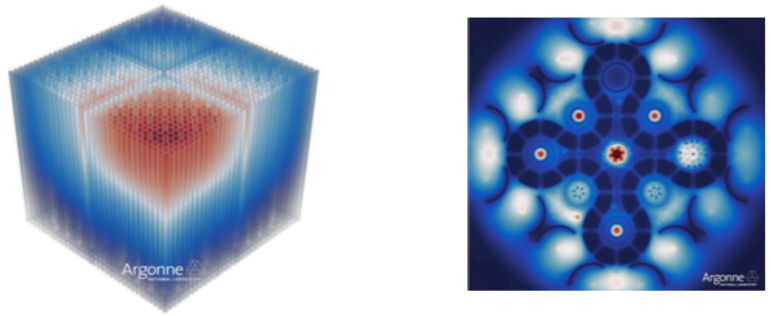
Showcase

ExaSMR: OpenMC Performance on Ponte Vecchio

Monte Carlo particle transport code for exascale computations

Ponte Vecchio with OpenMP Target offload

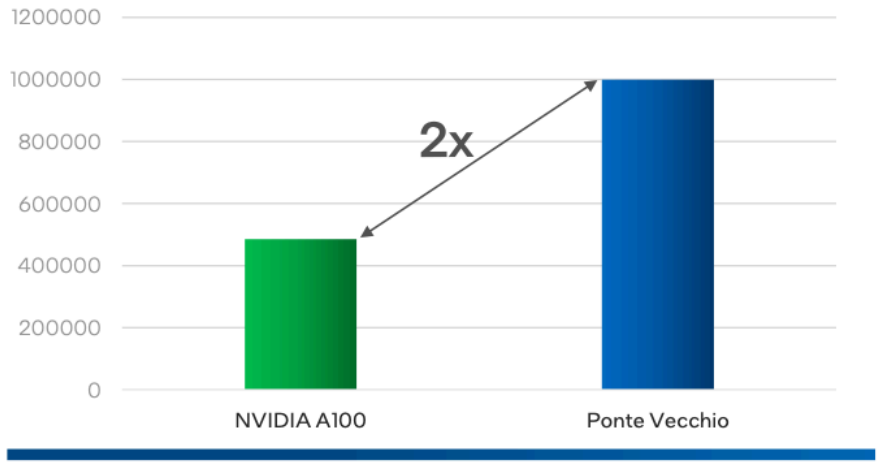
2x performance lead



Exascale Compute Project Annual Meeting 2022 presentation:
https://www.alcf.anl.gov/events/2022-ecp-annual-meeting
International Conference on Physics of Reactors 2022 presentation:
https://www.ans.org/meetings/physor2022/session/view-976/

https://docs.openmc.org

OpenMC Depleted Fuel Inactive Batch Performance on HM-Large Reactor with 40M particles (particles/second, Higher is better)



Application Summary: OpenMC is a Monte Carlo particle transport application that has recently been ported to the OpenMP target offloading programming model for use on GPU-based systems. The Monte Carlo method employed by OpenMC is considered the "gold standard" for high-fidelity simulation while also having the advantage of being a general-purpose method able to simulate nearly any geometry or material without the need for domain-specific assumptions. However, despite the extreme advantages in ease of use and accuracy, Monte Carlo methods like those in OpenMC often suffer from a very high computational cost. The extreme performance gains OpenMC has achieved on GPUs, as compared to traditional CPU architectures, is finally bringing within reach a much larger class of problems that historically were deemed too expensive to simulate using Monte Carlo methods. The leap in performance that GPUs are now offering carries with it the potential to disrupt a number of engineering technology stacks that have traditionally been dominated by non-general deterministic methods. For instance, faster MC applications may greatly expand the design space and simplify the regulation process for new nuclear reactor designs -- potentially improving the economics of nuclear energy and therefore helping to solve the world's climate crisis.

HOT CHIPS

- See backup for workloads and configurations. Results may vary.
- Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

intel.

H. Jiang, "Intel's Ponte Vecchio GPU : Architecture, Systems & Software," *2022 IEEE Hot Chips 34 Symposium (HCS)*, 2022, pp. 1-29, doi: 10.1109/HCS55958.2022.9895631.

Student Opportunities

- DOE's: **Science Undergraduate Laboratory Internship (SULI)**
  - **Application**:
    - How to Apply
    - open **NOW**
    - **due January 9, 2024**

- Argonne's: **Student Research Participation Program (SRP)**
  - Applications for Spring 2024 **NOW OPEN !!**
  - Deadline: **Friday, October 27, 2023**

- **Reach out!**

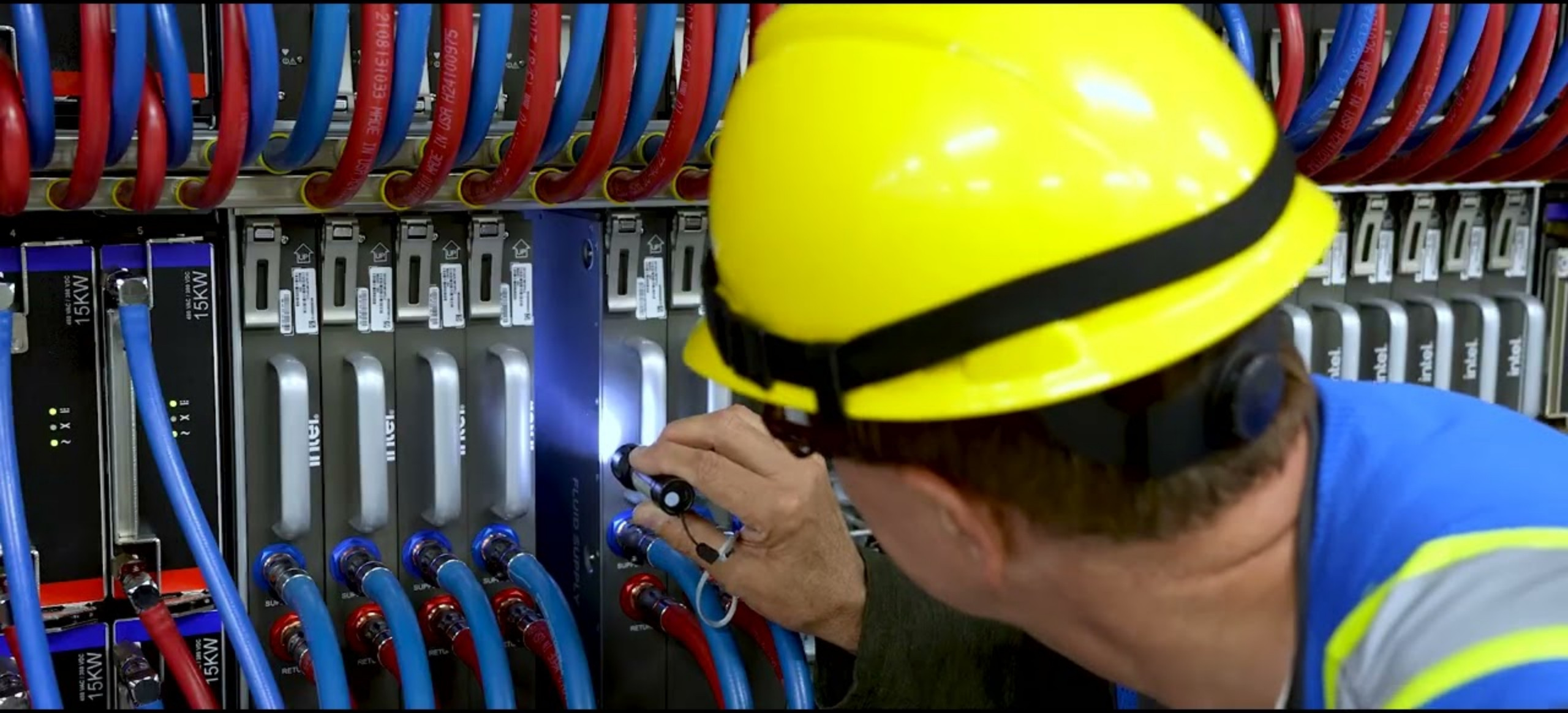
APPLY NOW!!

Undergraduate Programs

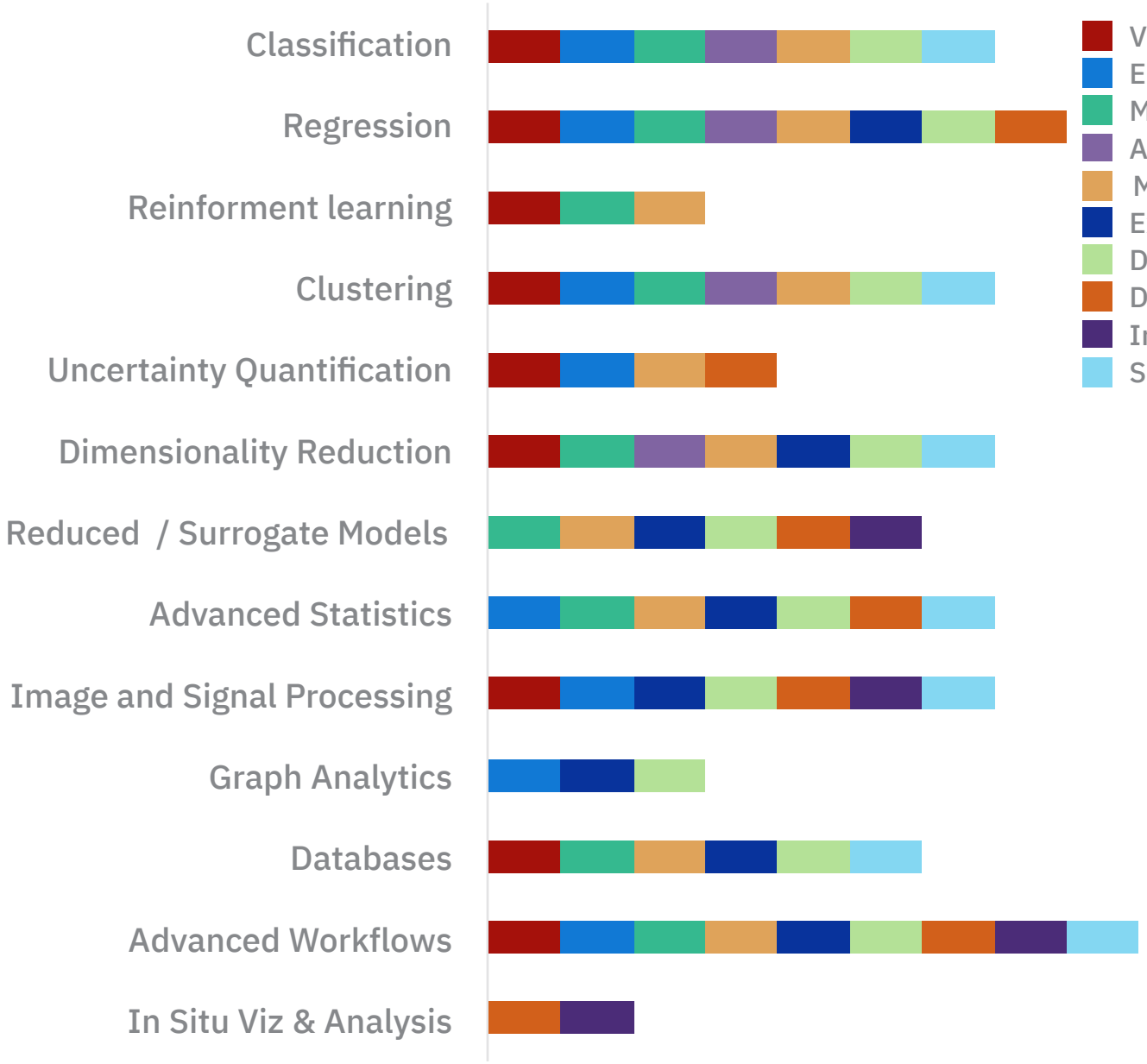
- Undergraduate Research Aide Program (more info)
- Community College Internship
- Professional Career Internship Program
- Minority Serving Institutions Partnership Program
- Lee Teng Undergrauate Fellowship in Accelerator Science
- Visiting Student Program for Undergraduate Students
- Seasonal Internship Program
- Sustainable Research Pathways (SRP)

Thank You!

foremans@anl.gov @saforem2 samforeman.me


15KW
15KW
15KW
intel

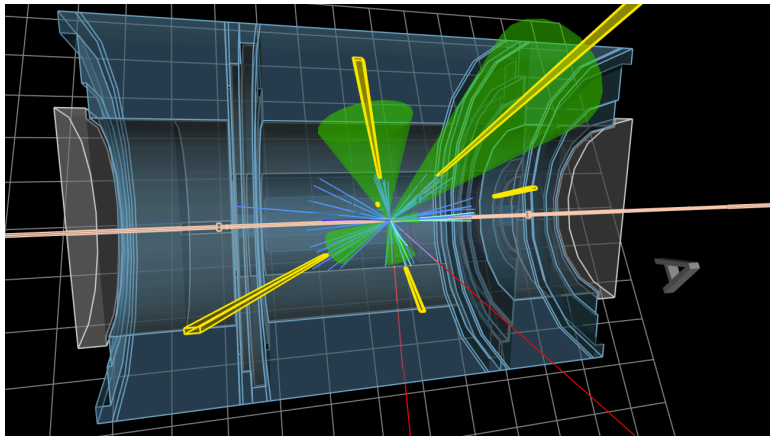
AURORA ESP Data and Learning Projects and Methods



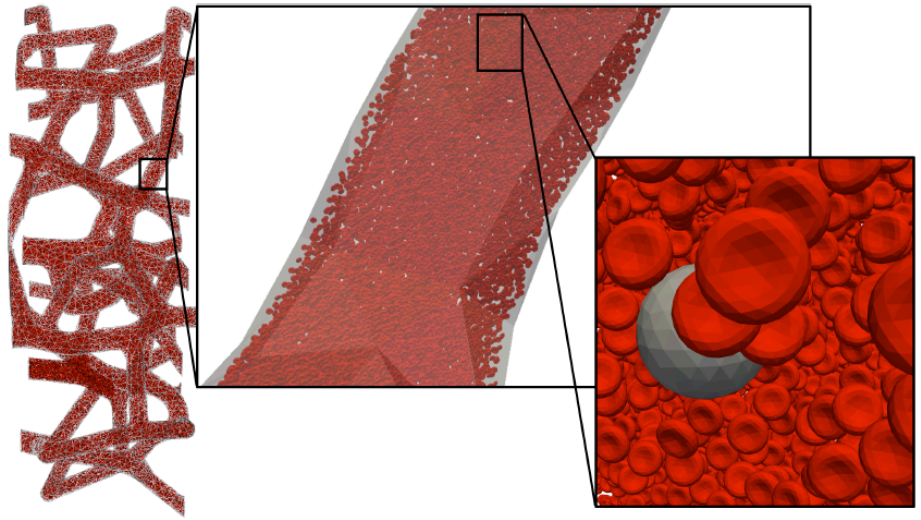
- Virtual Drug Response Prediction
- Enabling Connectomics at Exascale to Facilitate Discoveries in Neuroscience
- Machine Learning for Lattice Quantum Chromodynamics
- Accelerated Deep Learning Discovery in Fusion Energy Science
- Many-Body Perturbation Theory Meets Machine Learning
- Exascale Computational Catalysis
- Dark Sky Mining
- Data Analytics and Machine Learning for Exascale CFD
- In Situ Visualization and Analysis of Fluid-Structure-Interaction Simulations
- Simulating and Learning in the ATLAS detector at the Exascale

Learning **Data**

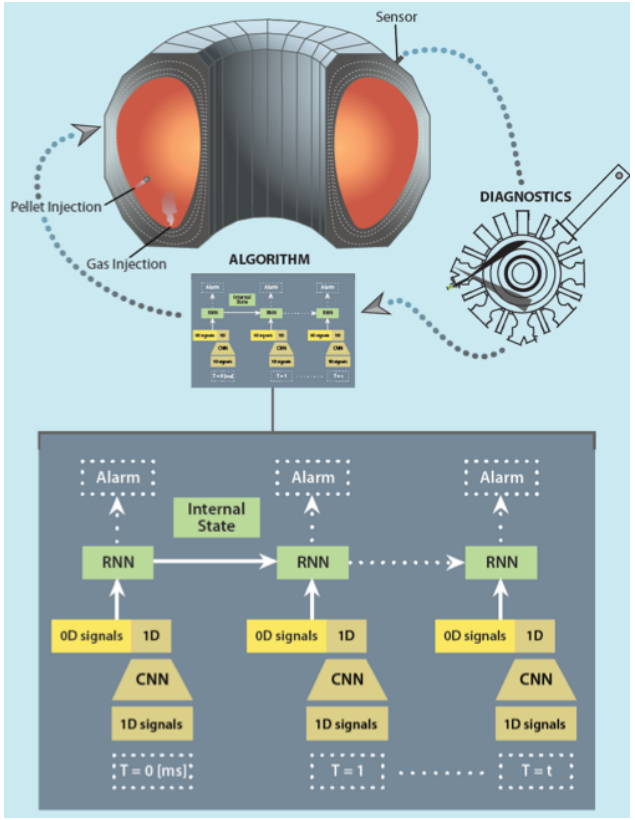
Exascale Simulation, Data and Learning



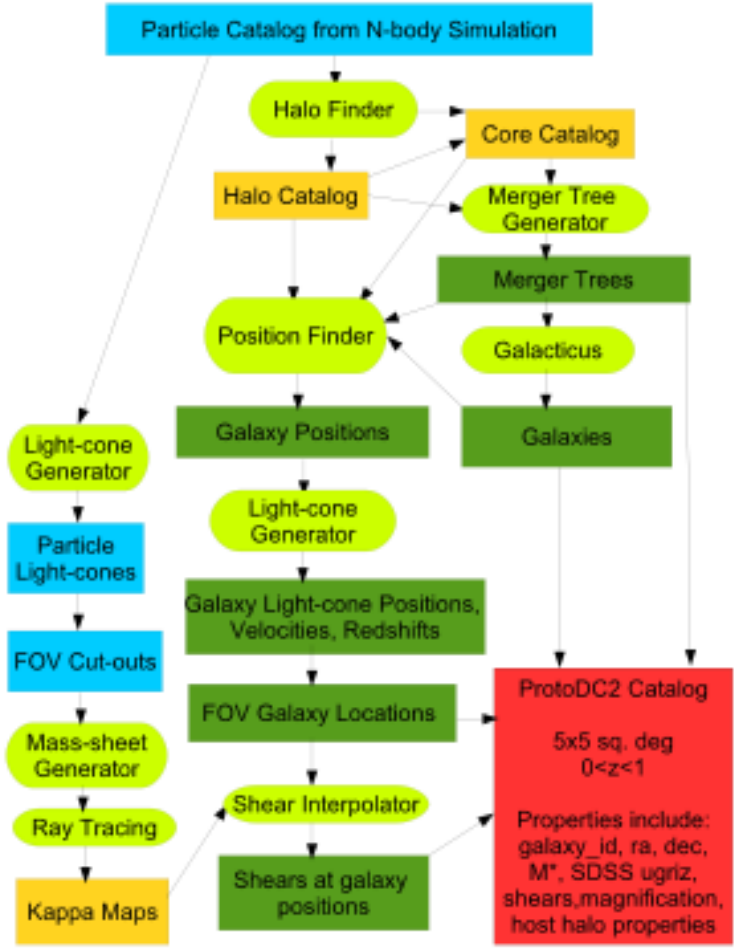
High Energy Physics



Bioscience, CFD



Fusion



Cosmology