Argonne Innovations in Computing

From AVIDAC to Aurora … and everything in between

Argonne has always been a leader in computing innovation.

Complementing theoretical and experimental research, **computation** is a third pillar of science— revolutionizing how scientists learn, experiment, and theorize.

Argonne
NATIONAL LABORATORY

August 11, 2025

Welcome to the World of HPC

Michael E. Papka
Senior Scientist / Argonne Distinguished Fellow and Division Director, Argonne Leadership Computing Facility
Deputy Associate Laboratory Director, Computing, Environment and Life Sciences, Argonne National Laboratory

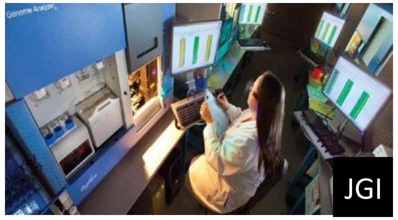
Warren S. McCulloch Professor of Computer Science, University of Illinois Chicago

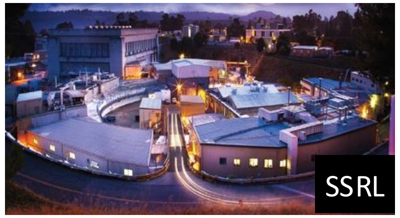
Department of Energy National Laboratories

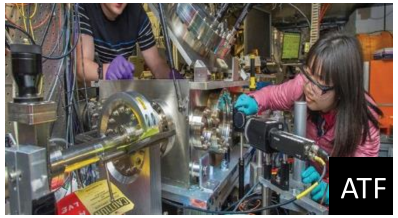
U.S. DEPARTMENT OF ENERGY

Department of Energy User Facilities

FY 2024
28 scientific
user facilities
>39,500 users







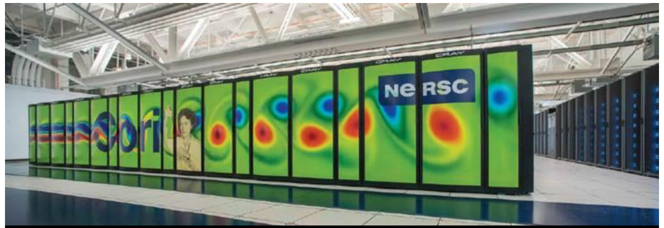
U.S. DEPARTMENT OF ENERGY | Office of Science

Department of Energy ASCR Facilities

- **Enable World-Class Scientific Discovery** - Provide cutting-edge supercomputing, data, and networking (via ESnet) that empower breakthrough research across disciplines, from materials science to climate modeling.
- **Advance Computational Science and Algorithms** - Drive the development of next-generation computational methods, algorithms, and software to solve complex scientific and engineering problems.
- **Deliver Capability and Capacity for National Priorities** - Balance leadership-class facilities (LCFs) for the largest, most complex simulations (capability) with centers like NERSC supporting broad, high-throughput workloads (capacity), all underpinning U.S. national security, energy, and innovation.
- **Support and Grow Diverse Research Communities** - Serve thousands of users across national labs, universities, and industry, accelerating discovery through open access, partnerships, and community engagement.
- **Push the Frontiers of Exascale, AI, and Quantum** - Lead the exploration of exascale computing, artificial intelligence, quantum computing, and ultra-fast networking to shape the future of scientific computing.

Argonne Leadership Computing Facility

Oak Ridge Leadership Computing Facility



National Energy Research Scientific Computing Center

Energy Sciences Network

A Brief History of Supercomputing

The history of HPC is among the most dramatic instances of human achievement through scientific discoveries and engineering innovations.

— Thomas Sterling et. al. **High Performance Computing: Modern Systems and Practices**

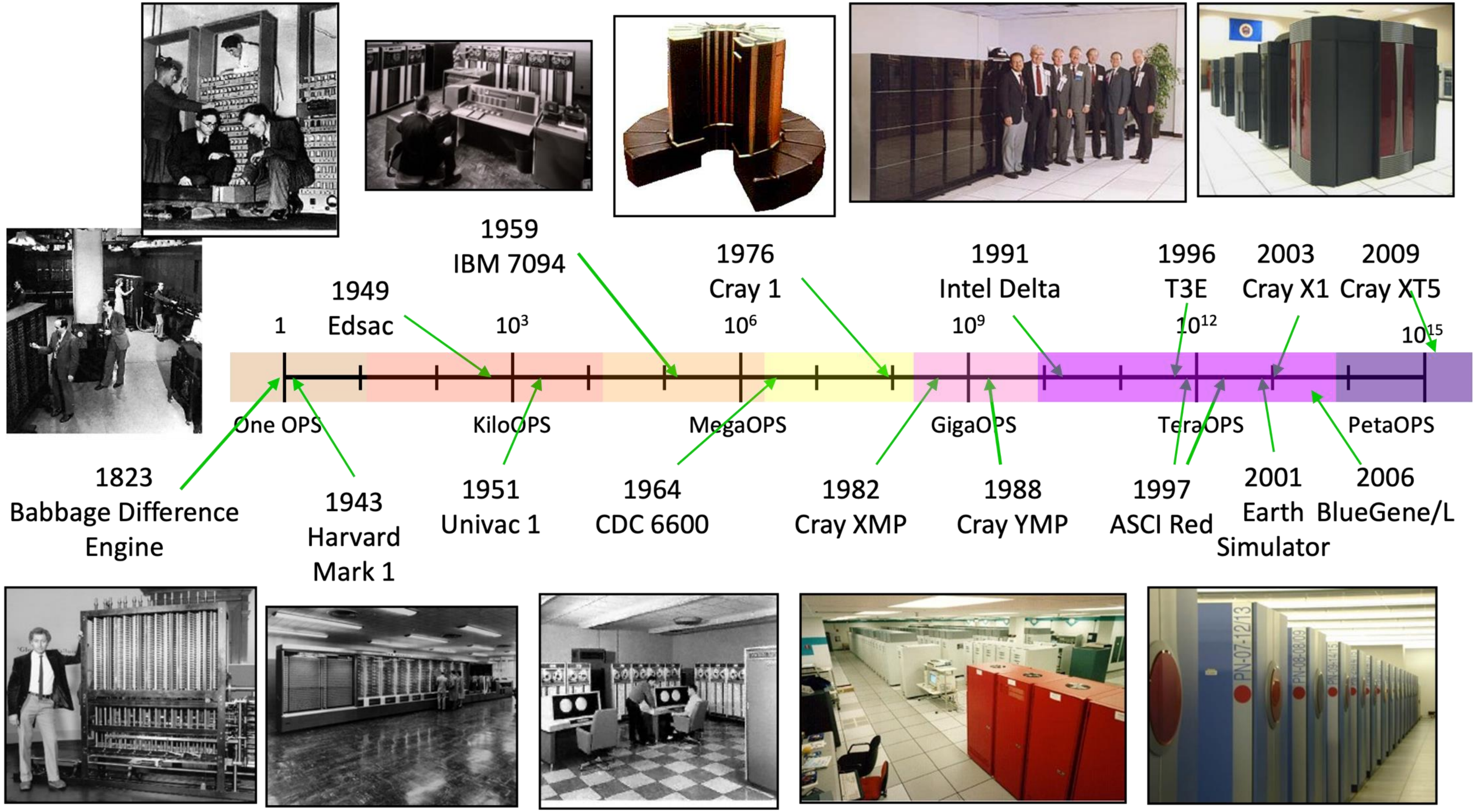
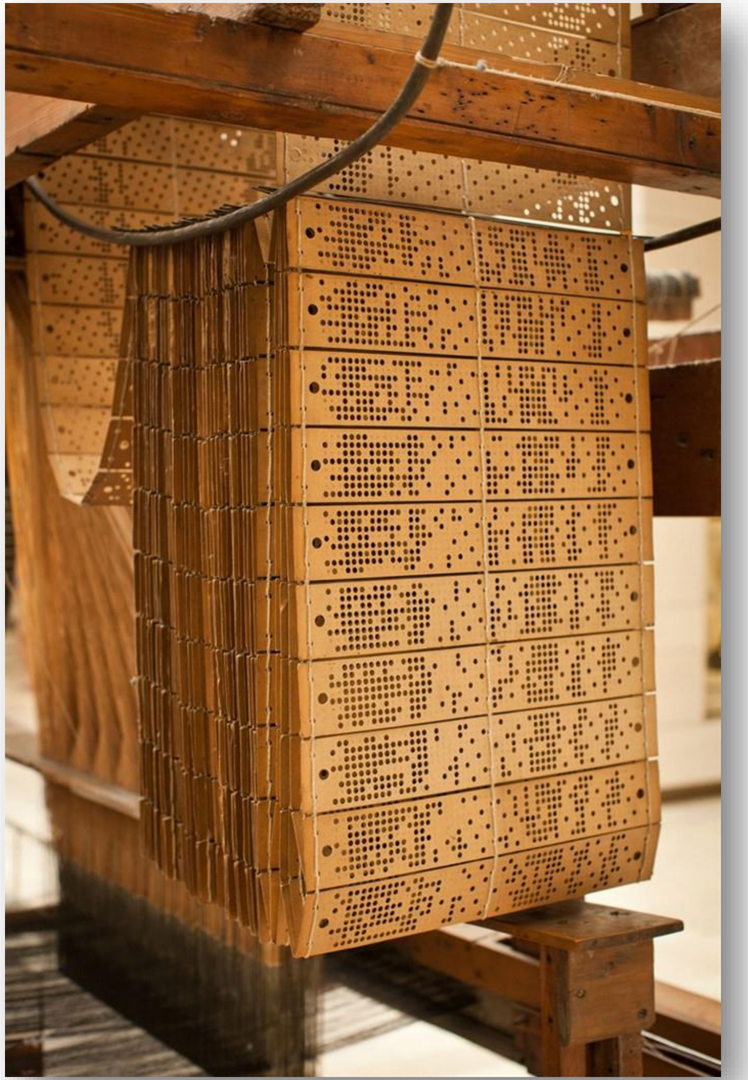


Image: Timeline created by Thomas Sterling

A Brief History of Supercomputing

- In a single lifetime, the capability of supercomputers has gained a growth factor of more than 10 trillion (10,000,000,000,000)
- HPC has seen performance gains at the rate of 200x each decade
- Changes in architecture enabled by new emergent technologies (vacuum tubes to transistors, spinning disk to solid state drives)
- Nine major periods mark HPC history
- Automated calculators through mechanical technologies
    - von Neumann architecture in vacuum tubes
    - Instruction-level parallelism
    - Vector processing and integration
    - Single-instruction multiple data array
    - Communicating sequential processors and very large-scale integration
    - Multicore petaflops
    - Heterogeneous exaflops
    - Quantum

Automated Calculators through Mechanical Technologies

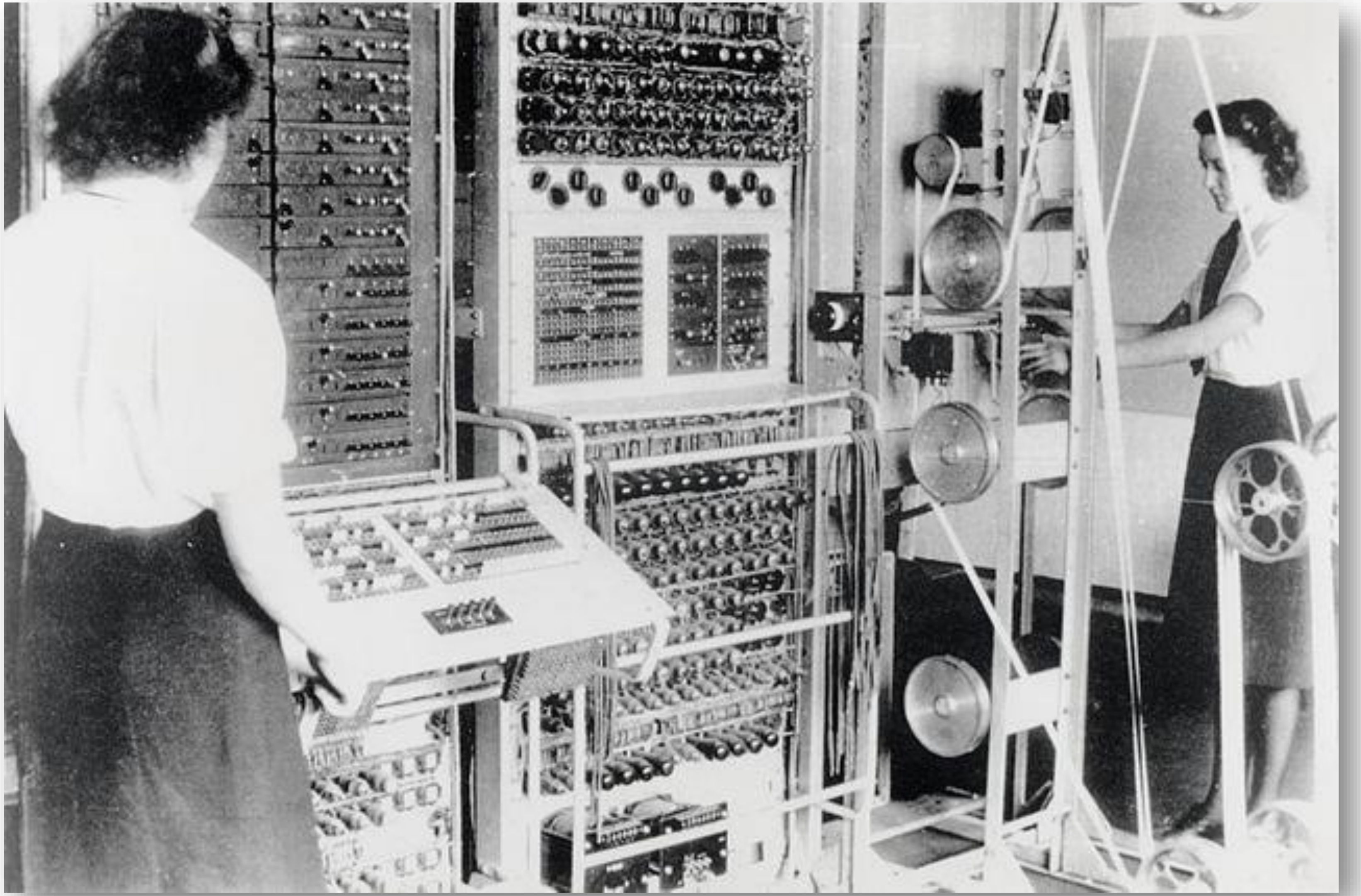


Images: Left image is the Jacquard loom in the *Making It* gallery in the National Museum of Scotland and the right image a collection of punch cards from that loom.

Automated Calculators through Mechanical Technologies

Image: Harvard Mark I, 50 feet long and containing some 750,000 components, was used for calculations during World War II

von Neumann Architecture in Vacuum Tubes



Wikipedia: Colossus Mark II code-breaking computer being operated by Dorothy Du Boisson (left) and Elsie Booker (right) in 1943

von Neumann Architecture in Vacuum Tubes

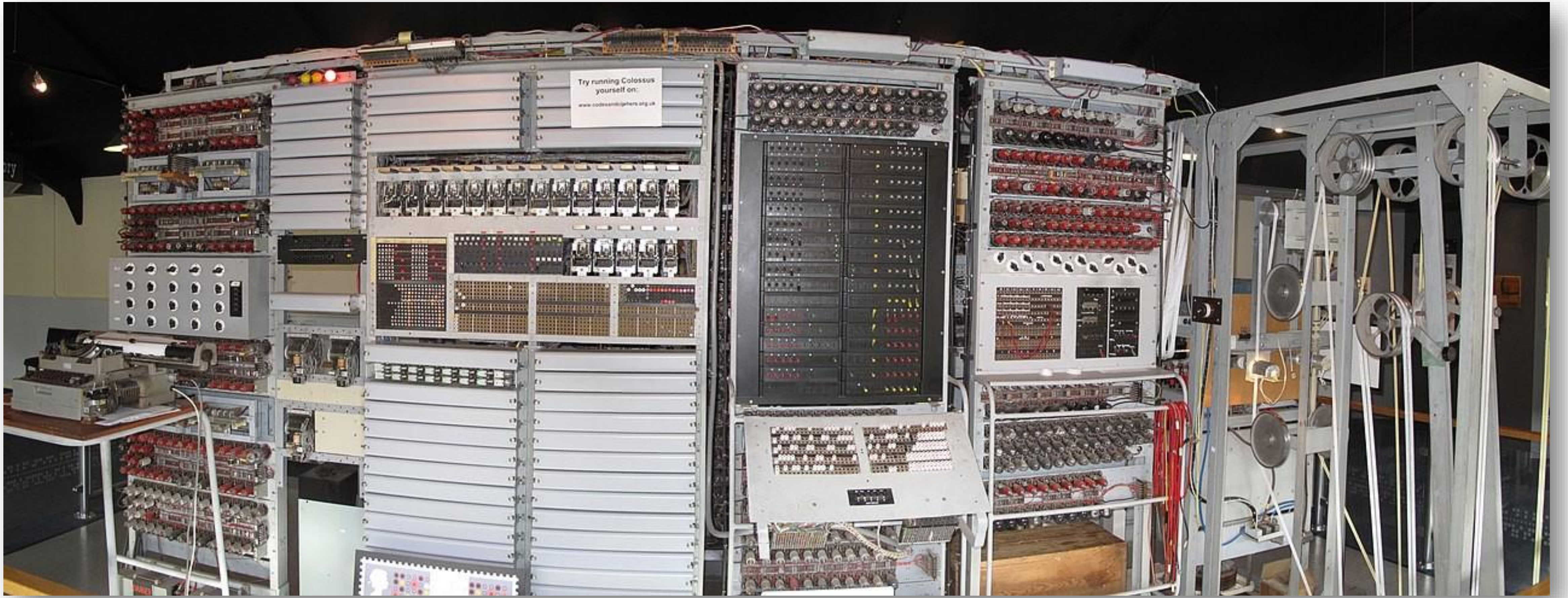
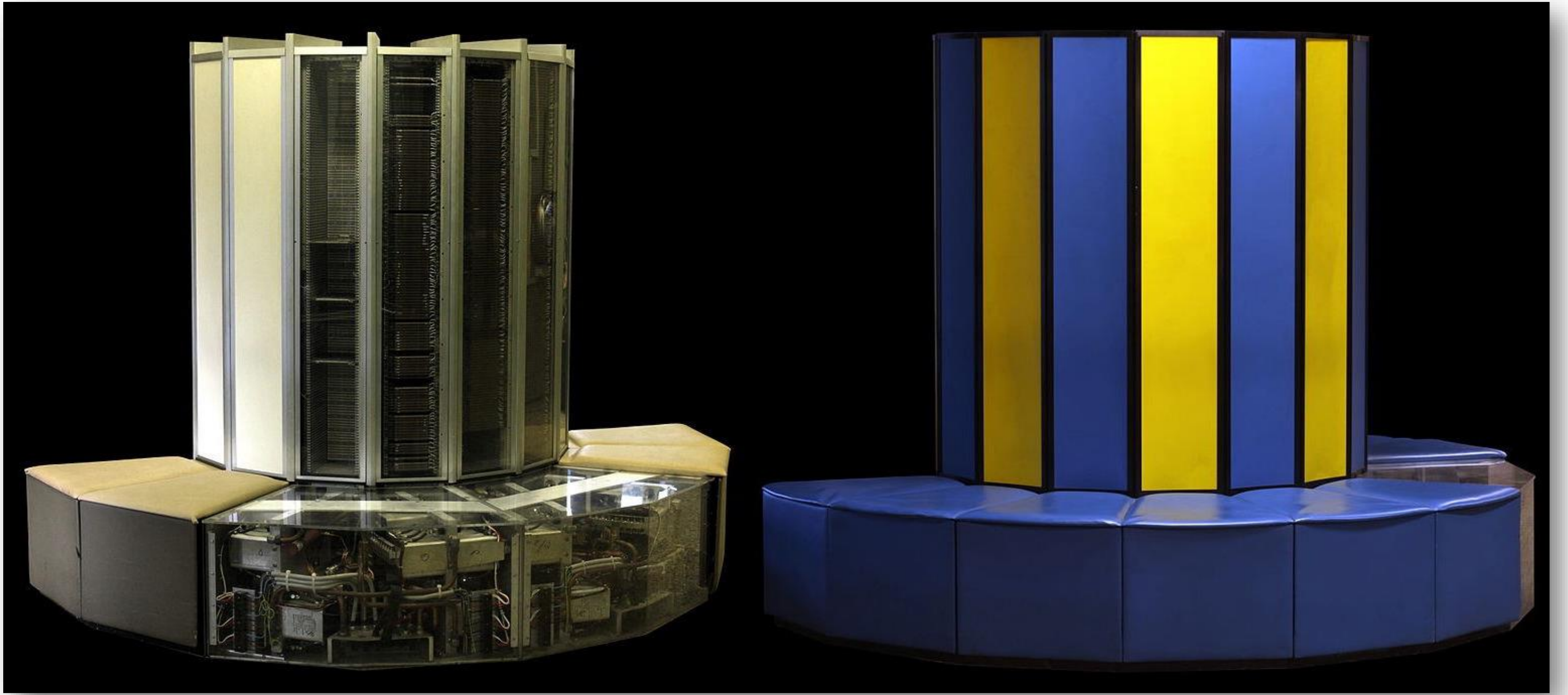


Image: Colossus Front view of the Colossus rebuild by Tony Sale between 1993 and 2008

Instruction-level parallelism

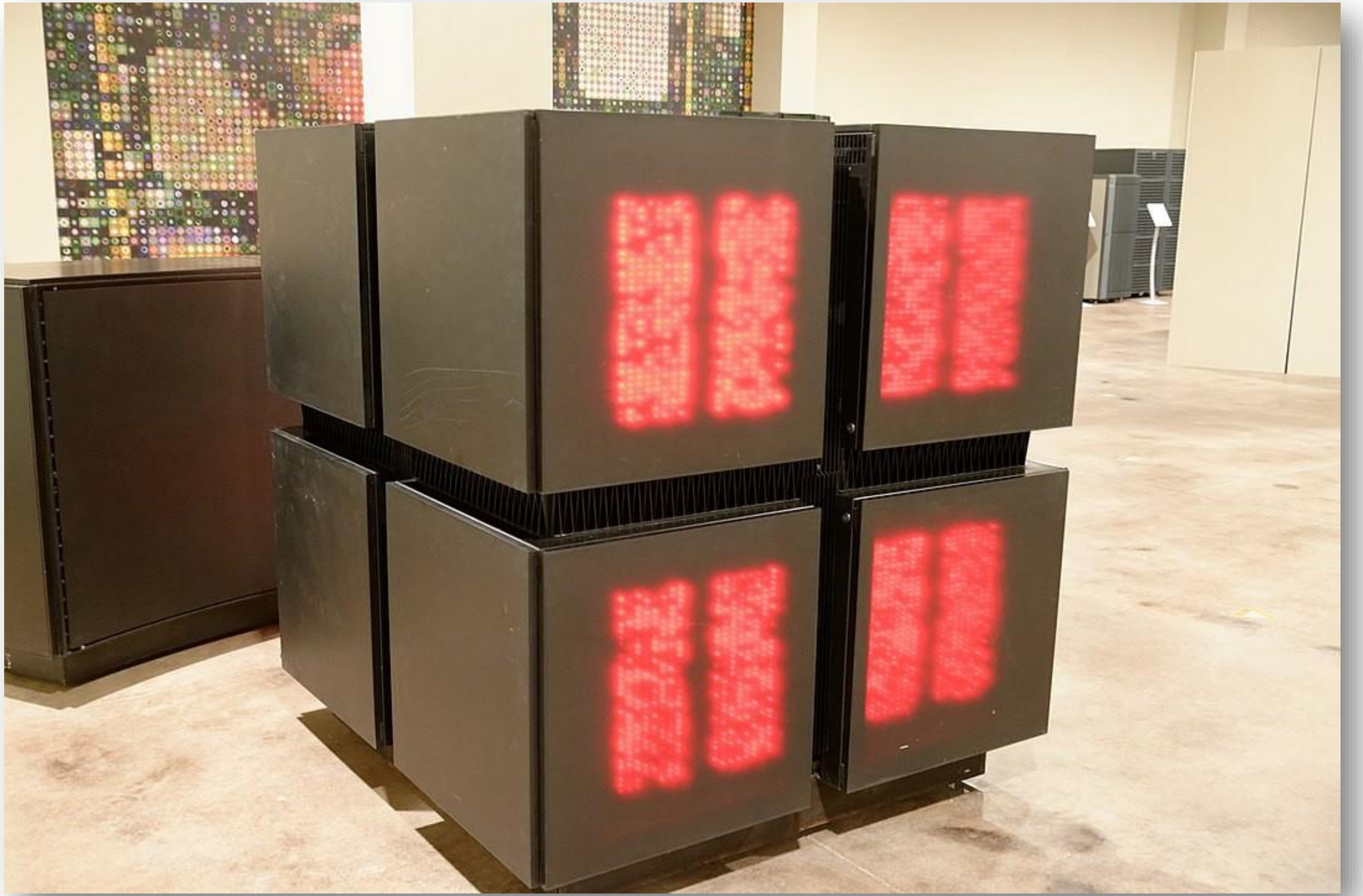
Wikipedia: CDC6600 computer with operator console in foreground

Vector Processing



Wikipedia: Cray 1 (left) and Cray X-MP (right)

Massively Parallel



Wikipedia: CM-2 built by Thinking Machines Corporation

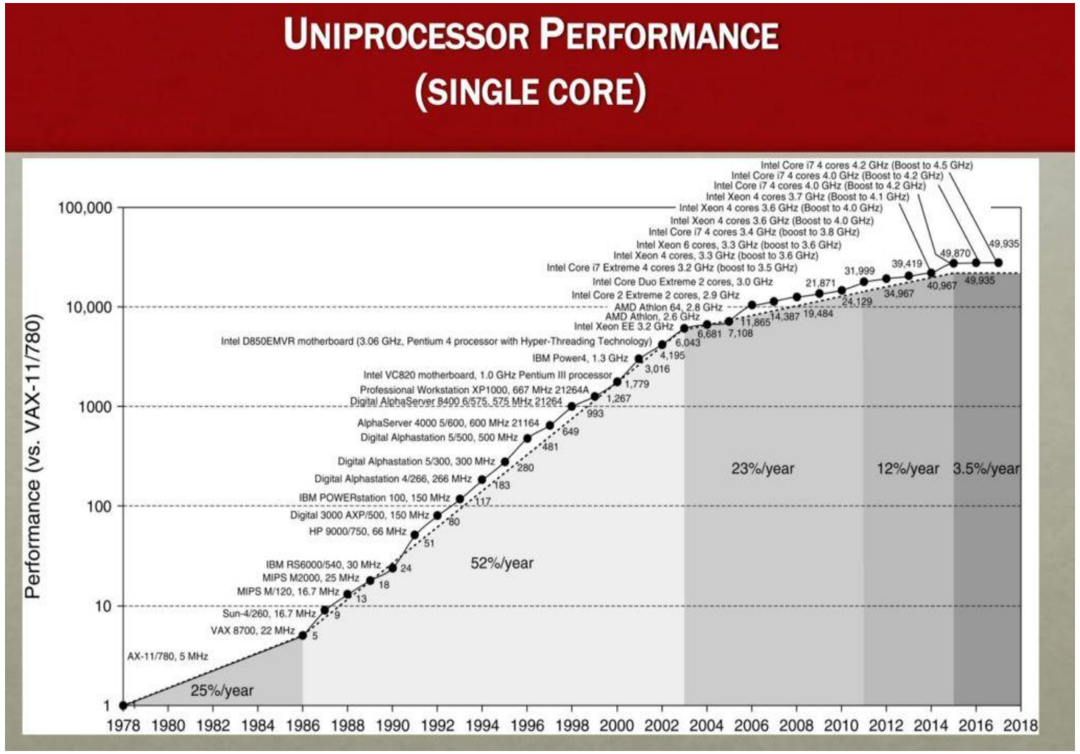
Communicating Sequential Processors and VLSI

Images: The Intel Paragon at Computer History Museum (left) and Thomas Sterling in front of NASA Beowulf system (right)

Exascale

- **Enabling Technology**
  - 3-D die stacking
  - Optical inter/intra socket networking
  - End of Moore's Law
  - Processor-in-Memory (PIM)
- **Fundamental Concepts**
  - Innovative execution models
  - Dynamic adaptive resource management
  - Message-driven computation
  - Multi-threading
- **Accomplishments**
  - Billion-way parallelism
  - Energy <20pJ/op
  - Runtime system software

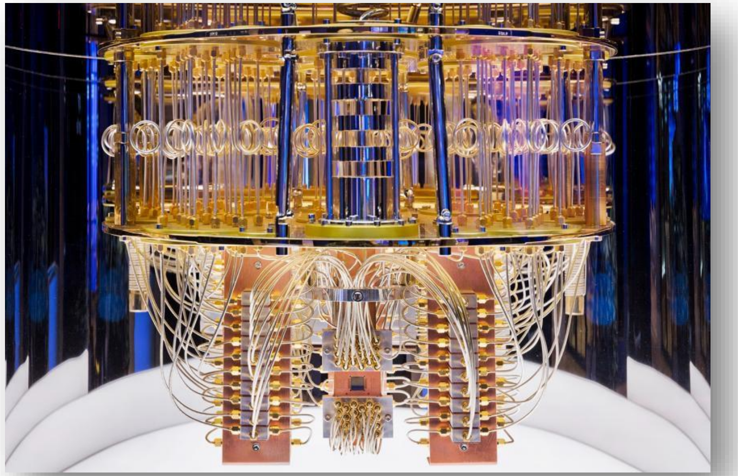
3D die



End of Moore's Law

Quantum

- **Enabling Technology**
  - Cryogenics
  - Josephson Junctions
  - Graphene and NCT
- **Fundamental Concepts**
  - Quantum Mechanics
- **Accomplishments**
  - A few qubits
  - Adiabatic switching
  - Algorithms
  - Witnessing entanglement



IBM quantum computer

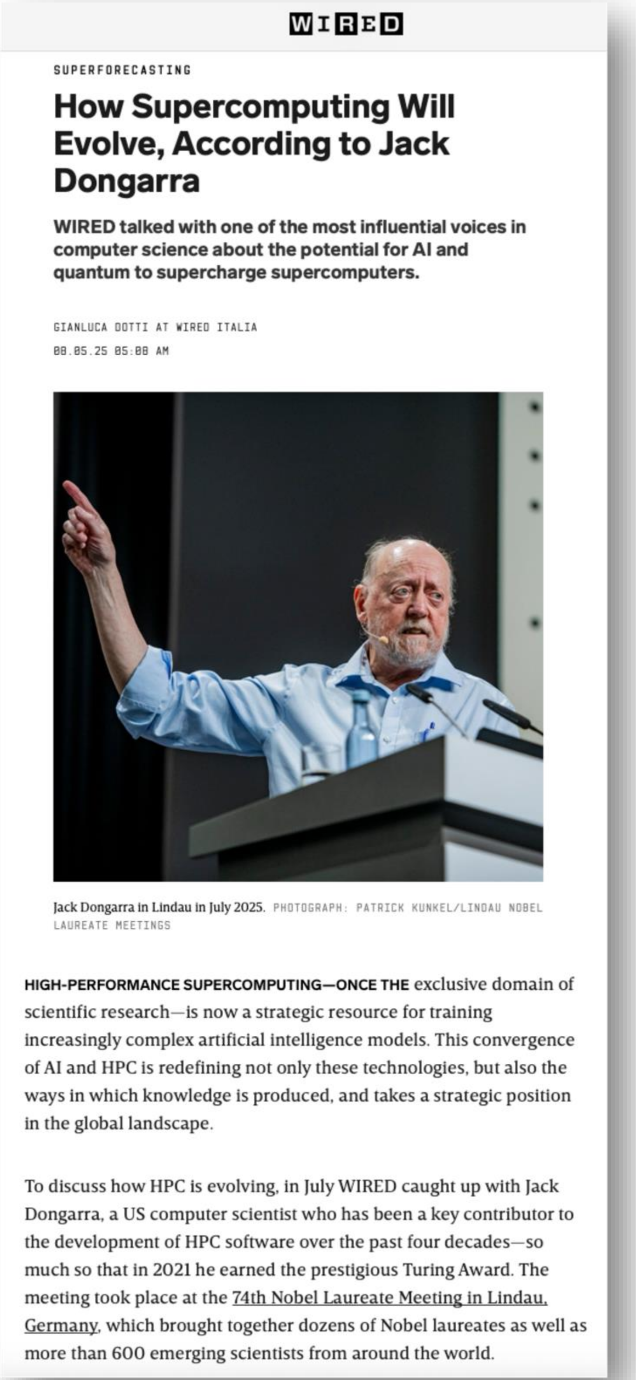
IBM quantum computer in cryostat

NOTE: Chicago home to lots of quantum work; Argonne's Q-Next, UIC part of Co-design Center for Quantum Advantage, UChicago's EPiQC: Enabling Practical-scale Quantum Computation - efforts at most major universities and both Argonne and Fermi National Laboratory'

Future of HPC (according to Jack)

1. **AI Becoming Central in HPC**
   AI is rapidly reshaping scientific research, moving from a purely computational tool to a powerful method for generating approximations that are then refined with traditional simulations.
2. **Multi-Accelerator Architectures**
   Supercomputing centers now rely heavily on GPUs, but the future may introduce even more specialized accelerators: quantum units, neuromorphic chips (brain-inspired architectures), and optical computing, performing operations at near-light speed.
3. **Quantum Computing "Winter"**
   While the quantum realm is fascinating, current systems remain rudimentary. Quantum computers deliver probability distributions rather than clear-cut answers, requiring many runs and offering potential, not certainty.
4. **Geopolitical-Driven Innovation**
   Fabrication remains a critical dependency: while U.S. designs often use Taiwan Semiconductor Manufacturing Company (TSMC), Chinese systems are said to be produced domestically—though potentially still relying on Taiwan.
5. **AI-Assisted Programming**
   AI is making software development more accessible: developers can prompt AI to draft code, then optimize it—perhaps even using natural language alone in the future.

Source: *How Supercomputing Will Evolve, According to Jack Dongarra* by Gianluca Dotti at WIRED Italia


WIRED

SUPERFORECASTING

How Supercomputing Will Evolve, According to Jack Dongarra

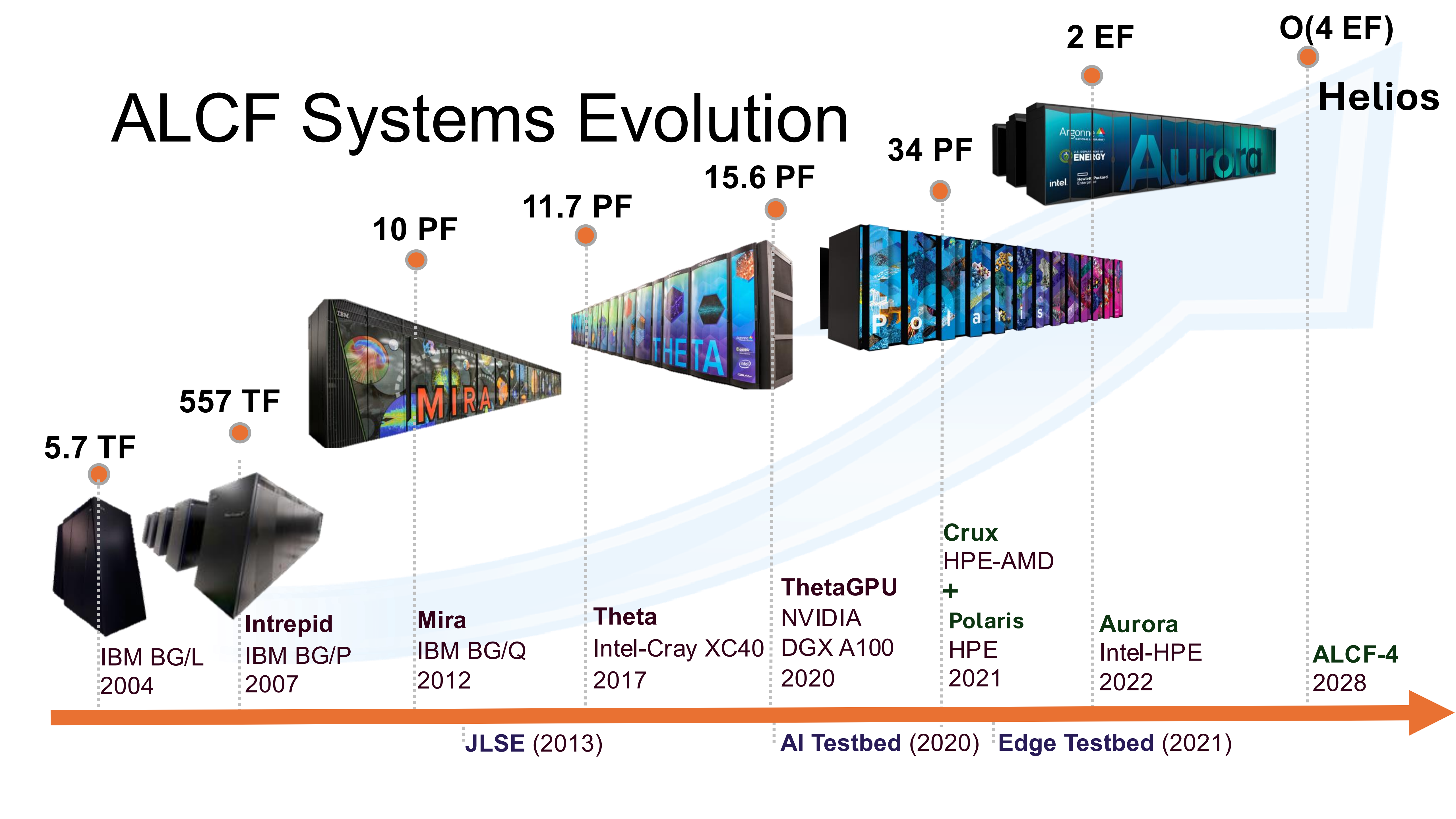
WIRED talked with one of the most influential voices in computer science about the potential for AI and quantum to supercharge supercomputers.

GIANLUCA DOTTI AT WIRED ITALIA
08.05.25 05:00 AM

Jack Dongarra in Lindau in July 2025. PHOTOGRAPH: PATRICK KUNKEL/LINDAU NOBEL LAUREATE MEETINGS

HIGH-PERFORMANCE SUPERCOMPUTING—ONCE THE exclusive domain of scientific research—is now a strategic resource for training increasingly complex artificial intelligence models. This convergence of AI and HPC is redefining not only these technologies, but also the ways in which knowledge is produced, and takes a strategic position in the global landscape.

To discuss how HPC is evolving, in July WIRED caught up with Jack Dongarra, a US computer scientist who has been a key contributor to the development of HPC software over the past four decades—so much so that in 2021 he earned the prestigious Turing Award. The meeting took place at the 74th Nobel Laureate Meeting in Lindau, Germany, which brought together dozens of Nobel laureates as well as more than 600 emerging scientists from around the world.


ALCF Systems Evolution
5.7 TF
557 TF
10 PF
11.7 PF
15.6 PF
34 PF
2 EF
O(4 EF)
Helios
IBM BG/L
2004
Intrepid
IBM BG/P
2007
Mira
IBM BG/Q
2012
Theta
Intel-Cray XC40
2017
ThetaGPU
NVIDIA
DGX A100
2020
Crux
HPE-AMD
+
Polaris
HPE
2021
Aurora
Intel-HPE
2022
ALCF-4
2028
JLSE (2013)
AI Testbed (2020)
Edge Testbed (2021)

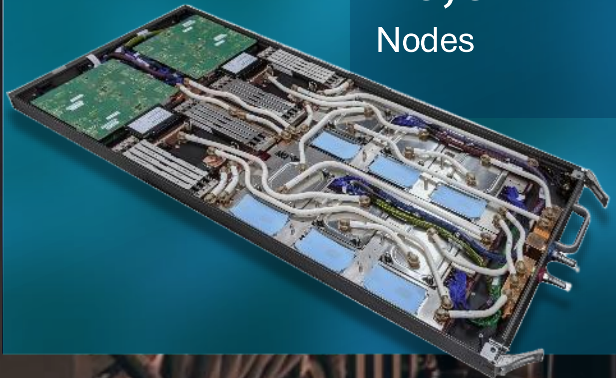
Aurora Specifications

Compute

21,248
CPUs

63,744
GPUs

10,624
Nodes



Fabric

Peak Injection Bandwidth
2.12 PB/s

Peak Bisection Bandwidth
0.69 PB/s

Dragonfly Topology

Memory

10.9PB DDR Capacity	1.36PB HBM CPU Capacity	8.16PB HBM GPU Capacity
5.95PB/s Peak DDR BW	30.5PB/s Peak HBM BW CPU	208.9PB/s Peak HBM BW GPU

Storage

230PB DAOS Capacity	31TB/s DAOS Bandwidth	1024 DAOS Node #

ALCF AI Testbed

Cutting-edge AI accelerators for science

GroqRack (Available for Allocation Requests)

GroqRack Inference

System Size: 72 Accelerators (9 nodes x 8 Accelerators per node)

Compute Units per Accelerator: 5120 vector ALUs

Performance of a single accelerator (TFlops): >188 (FP16) >750 (INT8)

Software Stack Support: GroqWare SDK, ONNX

Interconnect: RealScale TM

Cerebras CS-2 (Available for Allocation Requests)

Cerebras CS-2 Wafer-Scale Cluster WSE-2

System Size: 2 Nodes (each with a Wafer scale engine) including Memory-X and Swarm-X

Compute Units per Accelerator: 850,000 Cores

Performance of a single accelerator (TFlops): >5780 (FP16)

Software Stack Support: Cerebras SDK, Tensorflow, Pytorch

Interconnect: Ethernet-based

SambaNova Dataflow (Available for Allocation Requests)

SambaNova DataScale SN30

System Size: 64 Accelerators (8 nodes and 8 accelerators per node)

Compute Units per Accelerator: 1280 Programmable compute units

Performance of a single accelerator (TFlops): >660 (BF16)

Software Stack Support: SambaFlow, Pytorch

Interconnect: Ethernet-based

Graphcore Bow Pod64 (Available for Allocation Requests)

Graphcore Intelligent Processing Unit (IPU)

System Size: 64 Accelerators (4 nodes x 16 Accelerators per node)

Compute Units per Accelerator: 1472 independent processing units

Performance of a single accelerator (TFlops): >250 (FP16)

Software Stack Support: PopArt, Tensorflow, Pytorch, ONNX

Interconnect: IPU Link

Habana Gaudi-1

Habana Gaudi Tensor Processing Cores

System Size: 16 Accelerators (2 nodes x 8 Accelerators per node)

Compute Units per Accelerator: 8 TPC + GEMM engine

Performance of a single accelerator (TFlops): >150 (FP16)

Software Stack Support: Synapse AI, TensorFlow and PyTorch

Interconnect: Ethernet-based

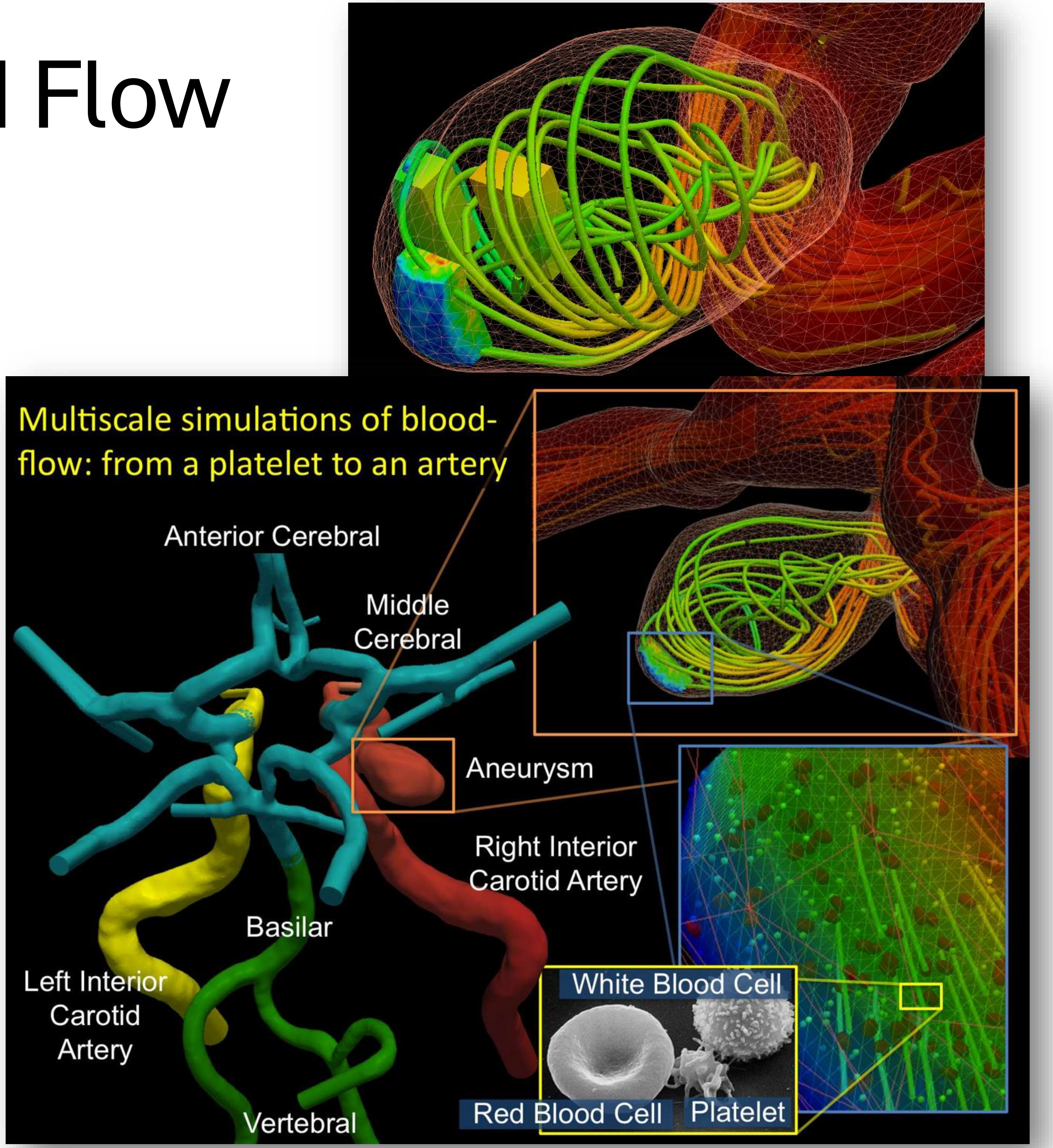
Community Data Sharing with Eagle

- A global filesystem deployed to bring larger and more capable production-level file sharing to facility users
- A space for broader distribution of reassembled data acquired from various experiments
  - Data originating at the ALCF
  - Greater scientific community
- Science community can access uploaded data, and ALCF users are able to directly access the data for analysis
- Designed to foster experimentation
  - Analysts are able to write new algorithms to attempt analyses that have never been performed

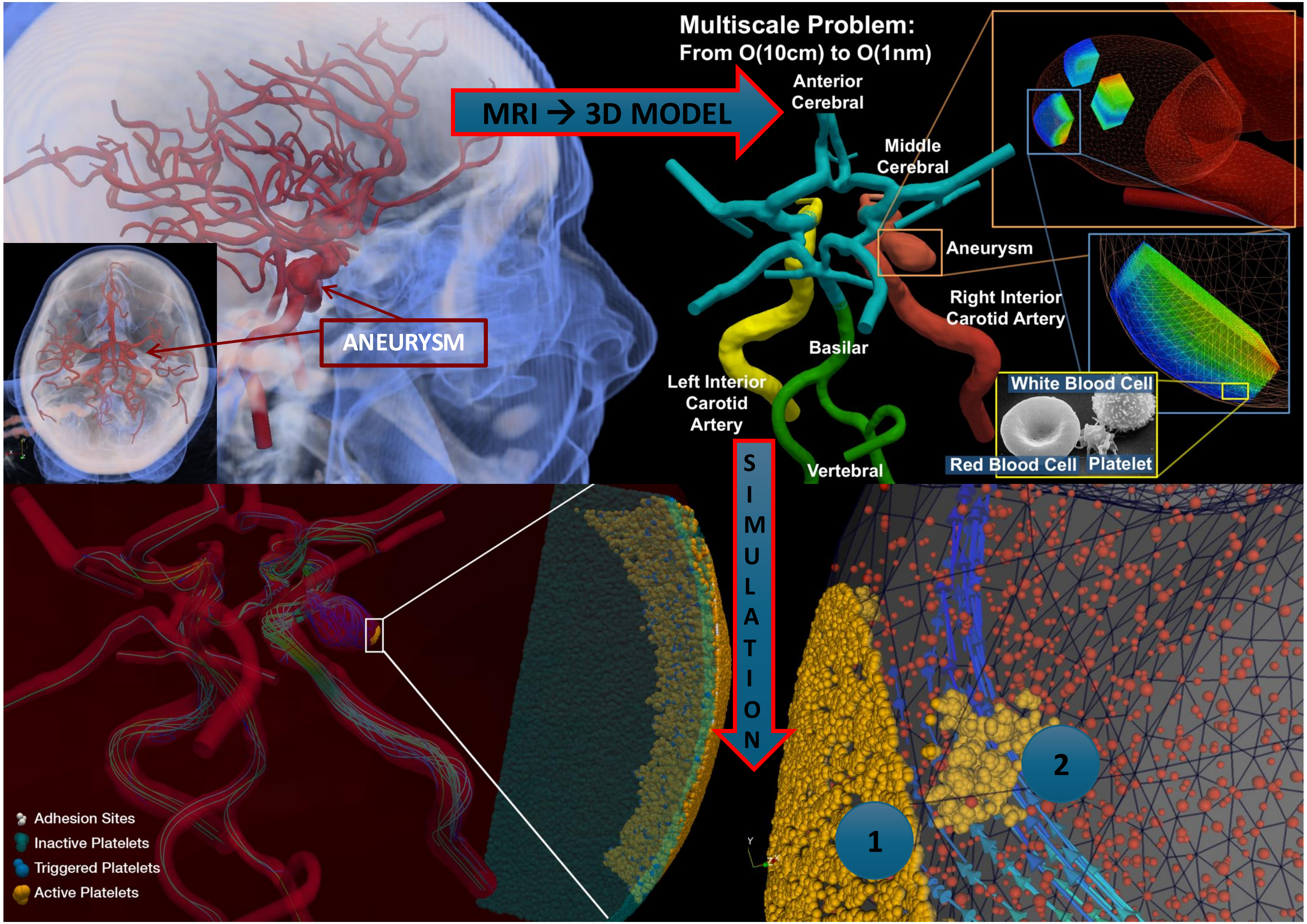
- **HPE ClusterStor E1000**
- **100 petabytes of usable capacity**
- **8,480 disk drives**
- **Lustre filesystem**
- **160 Object Storage Targets**
- **40 Metadata Targets**
- **HDR InfiniBand network**
- **650 GB/s rate on data transfers**

Application to Brain Blood Flow

- Multi-scale modeling of arterial blood flow can shed light on the interaction between events happening at micro and mesoscales (adhesion of red blood cells to the arterial wall, clot formation) and at macro-scales (change in flow patterns due to the clot)
- Coupled numerical simulations of such multiscale flow require state-of-the-art computers and algorithms, along with techniques for multi-scale visualizations
- Computer-aided design represents a huge speedup over traditional experiments



Science team: G. Karniadakis, L. Grinberg, V. Morozov, D. Fedosov, J. Insley, M. E. Papka

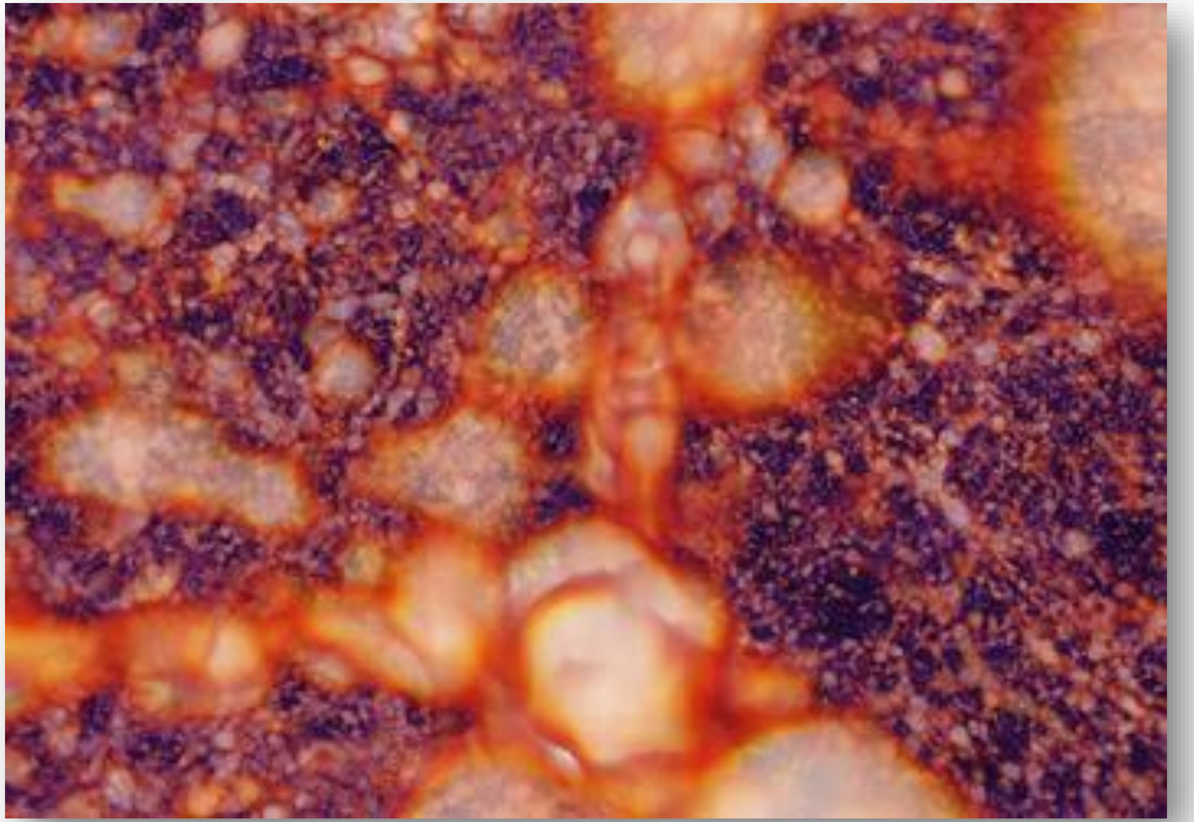


ANEURYSM
MRI → 3D MODEL
Multiscale Problem:
From O(10cm) to O(1nm)
Anterior Cerebral
Middle Cerebral
Aneurysm
Right Interior Carotid Artery
Basilar
Left Interior Carotid Artery
Vertebral
White Blood Cell
Red Blood Cell
Platelet
SIMULATION
Adhesion Sites
Inactive Platelets
Triggered Platelets
Active Platelets
1
2

Red Blood Cell Animation

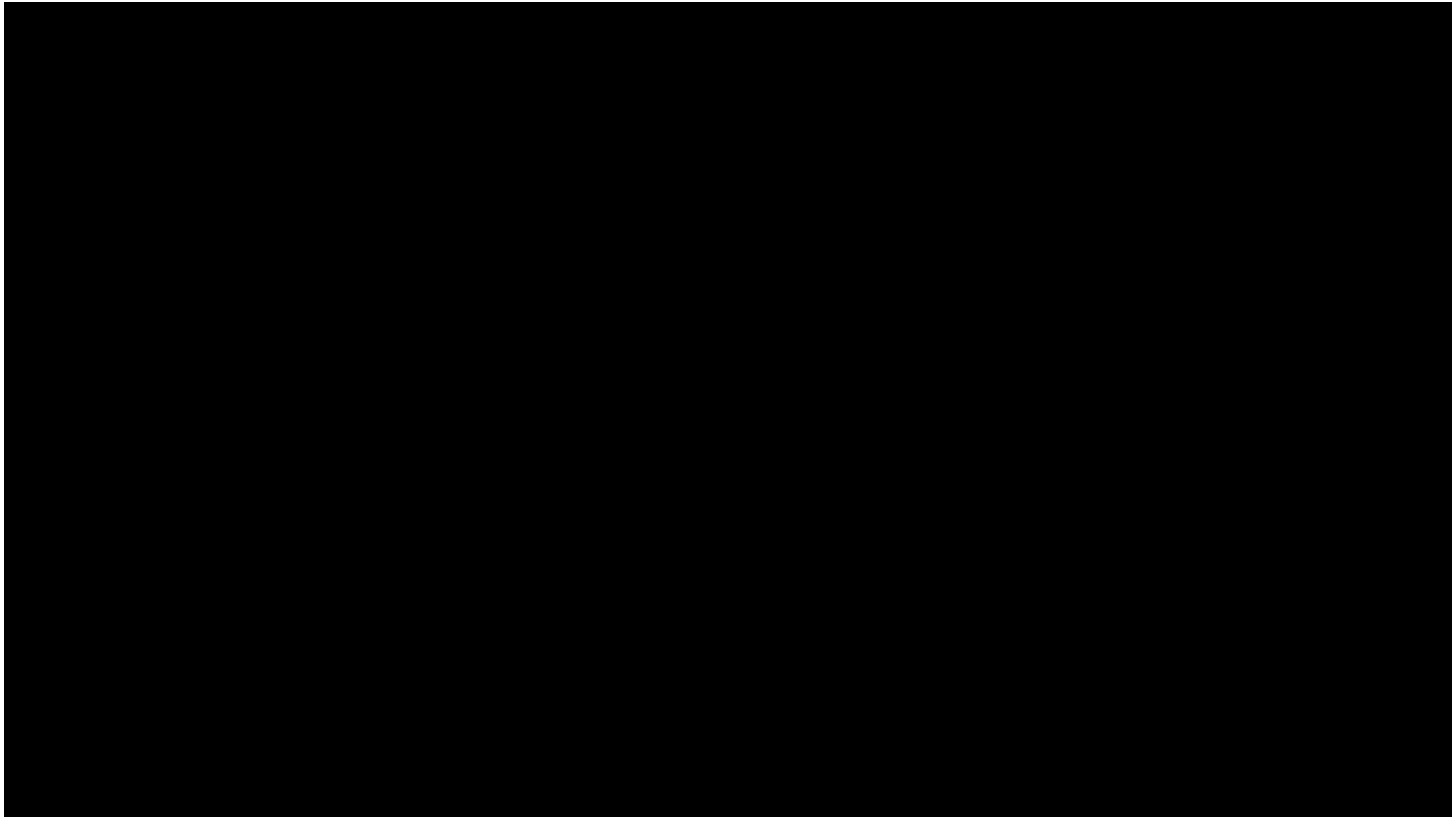
Computing the Dark Universe

- Understanding the nature of dark energy and dark matter is the foremost challenge in cosmology today.
- Simulations on Mira allow cosmologists to make theoretical predictions that can be tested against data gathered using powerful telescopes and space probes.
- HACC (Hardware/Hybrid Accelerated Cosmology Code) represents new generation of code: built from scratch, very efficient to port between systems, gets more "science per core."

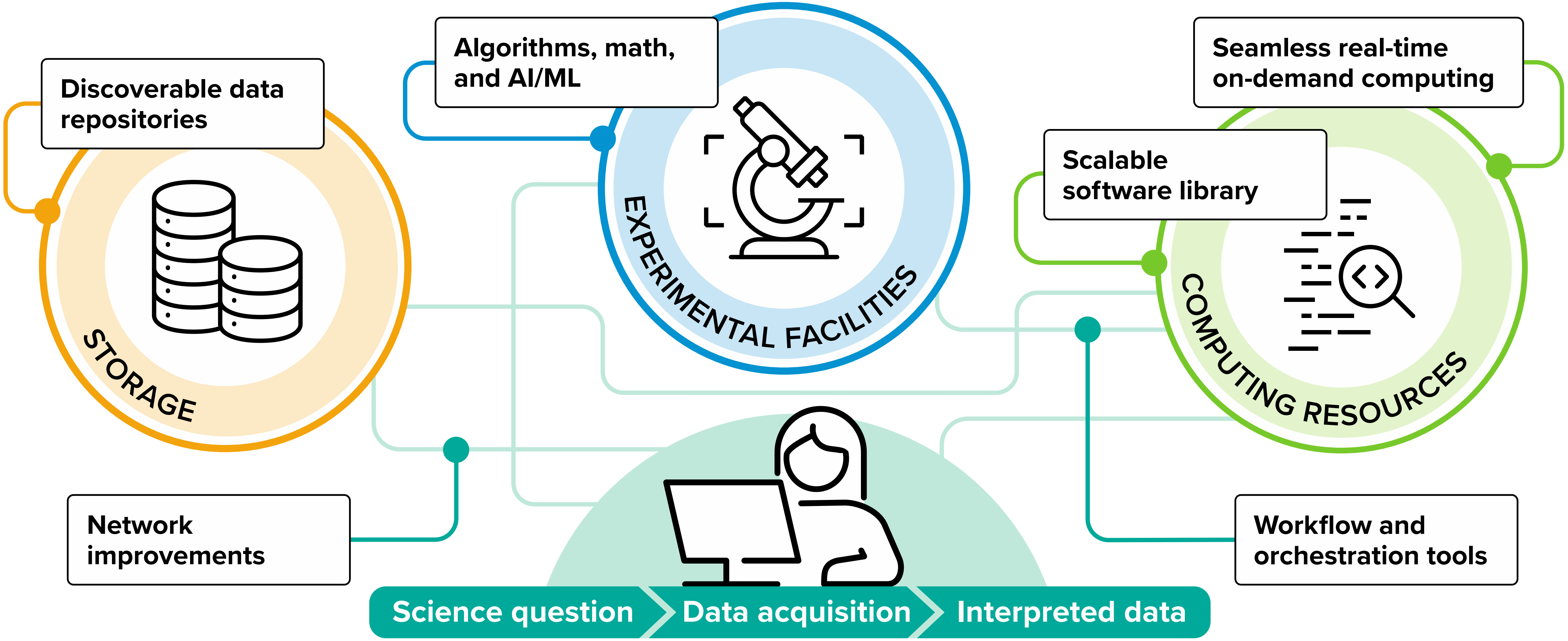


Science team: S. Habib, V. Morozov, N. Frontiere, H. Finkel, A. Pope, K. Heitmann, K. Kumaran, V. Vishwanath, T. Peterka, J. A. Insley, D. Daniel, P. Fasel, Z. Lukic

Probing the Cosmic Structure of the Dark Universe



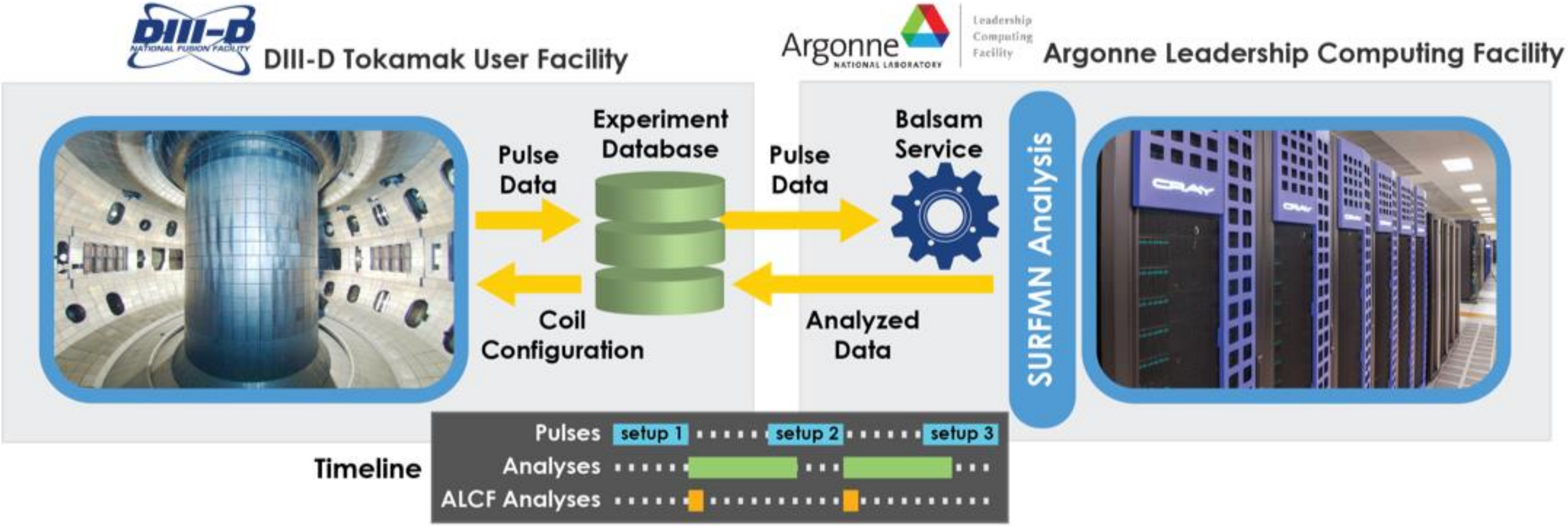
Integrated Research Infrastructure



Argonne National Laboratory's **Nexus** Effort (https://www.anl.gov/nexus-connect)

~2018

Experiments Integrating Research Infrastructure

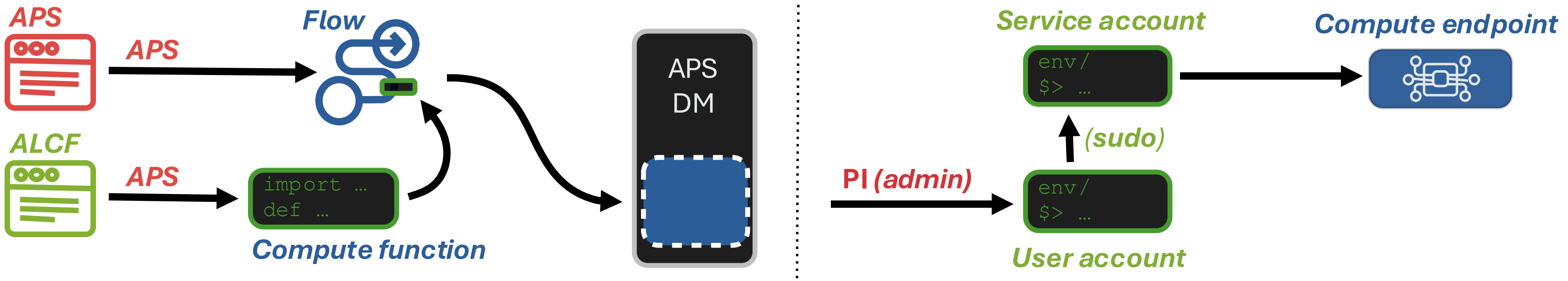


M. Kostuk, T. D. Uram, T. Evans, D. M. Orlov, M. E. Papka and D. Schissel, *Automatic Between-Pulse Analysis of DIII-D Experimental Data Performed Remotely on a Supercomputer at Argonne Leadership Computing Facility*, **Fusion Science and Technology**, 74(2), 2018.

Putting the Research Infrastructure Together (Nexus)

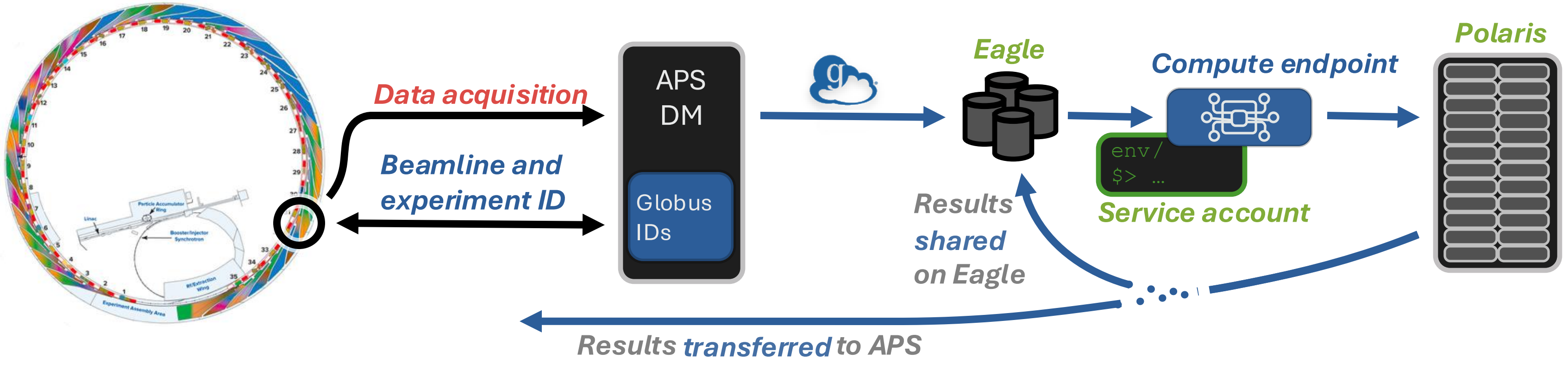
One-time configuration per beamline

Globus	APS		ALCF		
Globus creates client ***identities***	*APS registers* ***Flow templates***	*APS adds credentials to* ***Data Management*** *(DM) system*	*ALCF adds* ***service account***	*APS PI configures* ***environment***	*APS installs Globus* ***Compute endpoint***



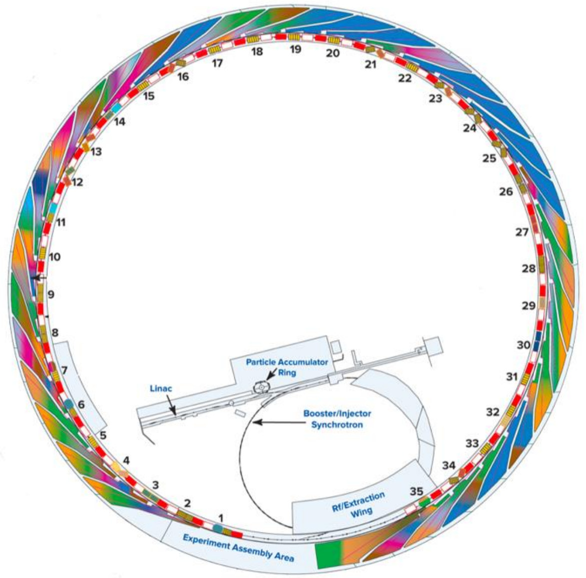
Automated workflow during experiments

APS		ALCF	
APS *links experiment to* ***beamline identity***	*DM system triggers* ***Globus Flow*** *during or following data acquisition*	*Globus* ***transfers*** *APS data to Eagle*	*Globus Compute launches* ***analysis*** *on behalf of beamline*

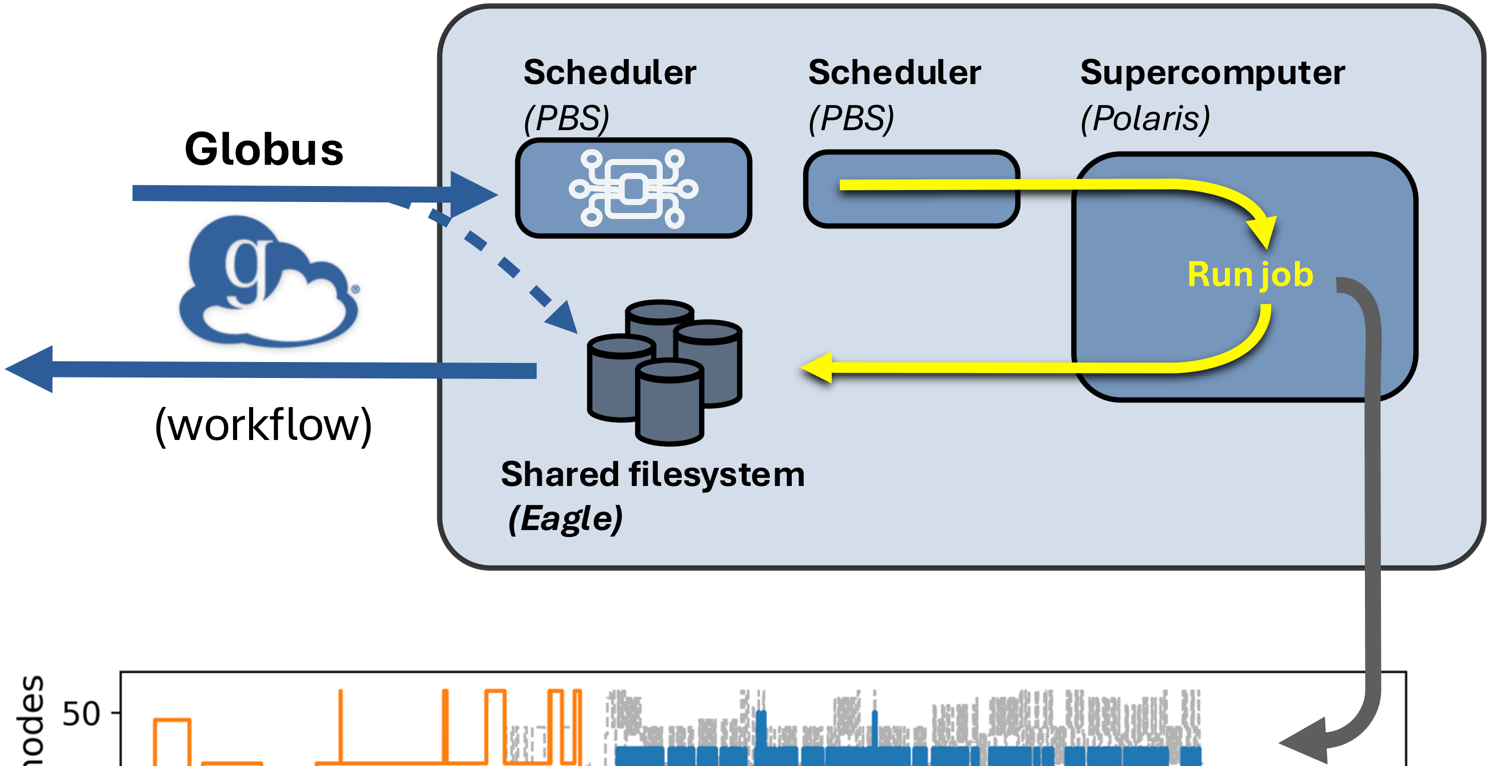


No Human in the Loop Experiment

APS (data acquisition)



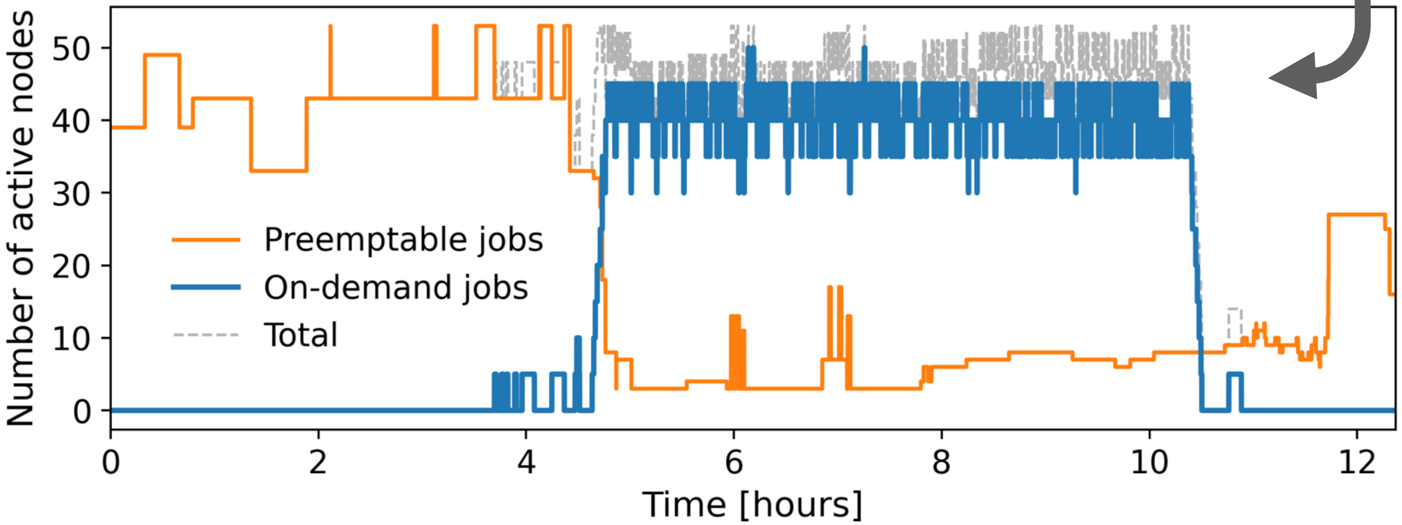
ALCF (on-demand data analysis)



Computation example

(Part of Laue experiment)

April 1st, 2023



Thank You