September 13, 2024

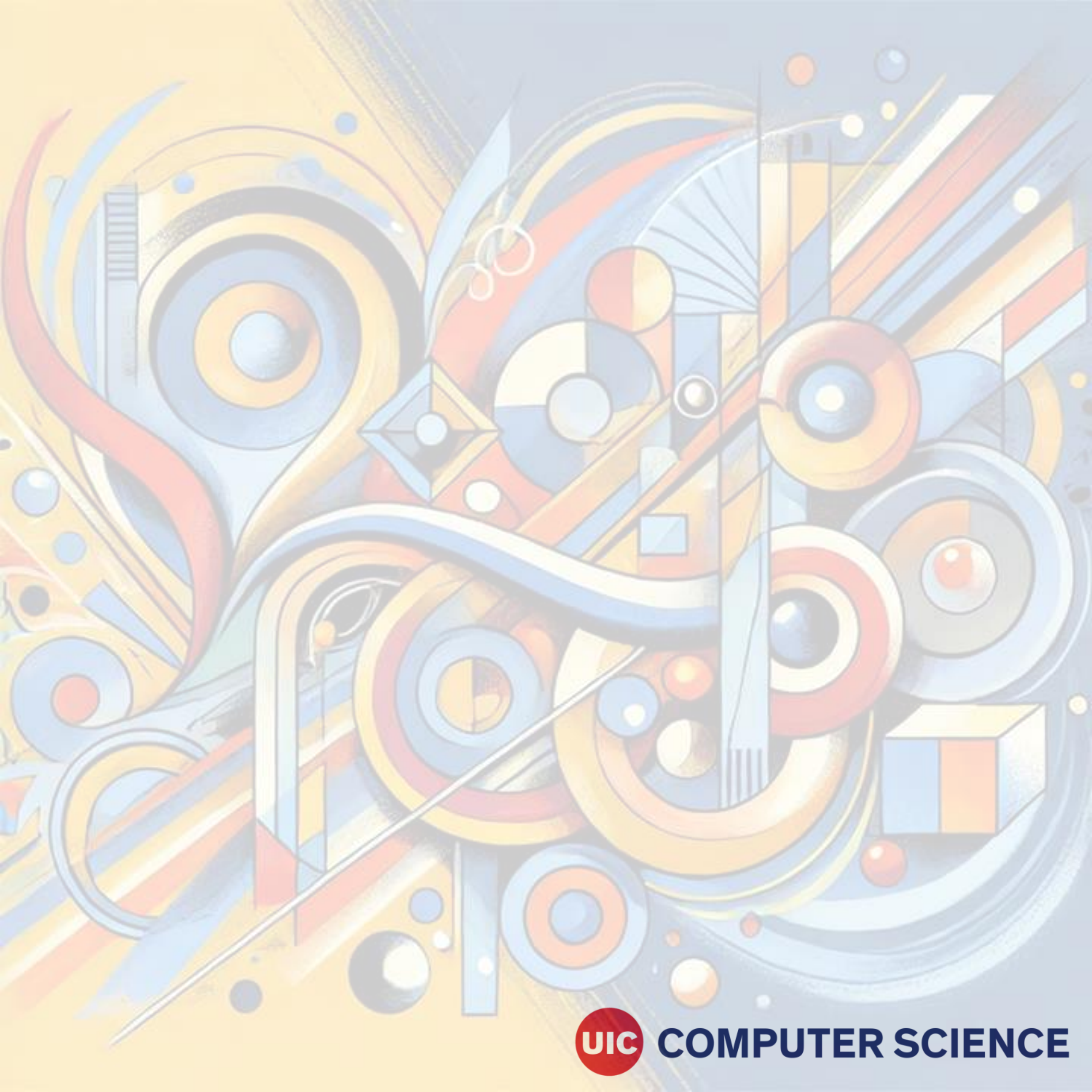
Intersection of AI and HPC

Michael E. Papka
Professor, Computer Science, University of Illinois Chicago
Senior Scientist, Computing, Environment and Life Sciences, Argonne National Laboratory

UIC COMPUTER SCIENCE

Overview

- **Evolution of HPC with AI**
- Challenges
- Opportunities
- Future Directions

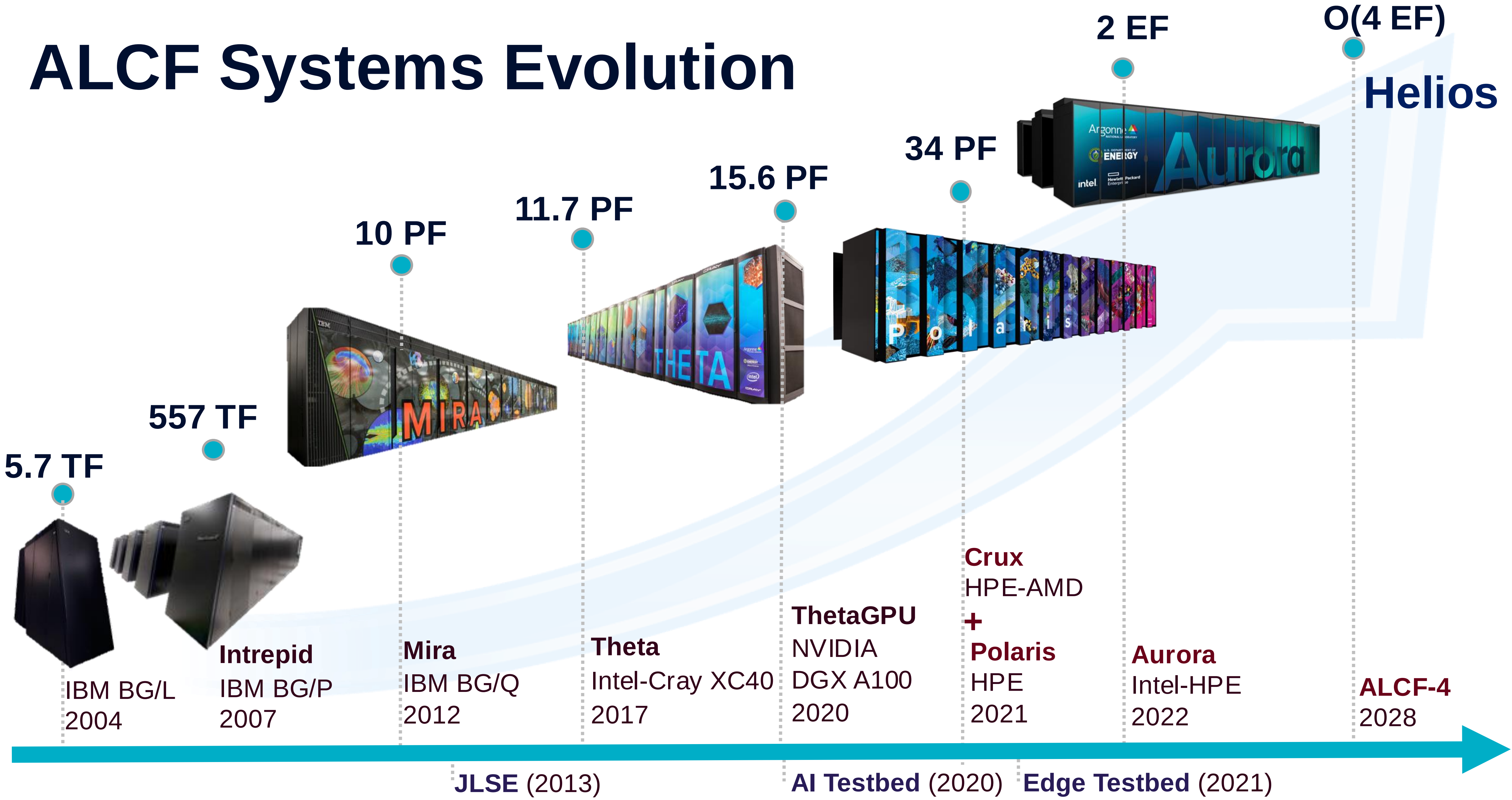


UIC COMPUTER SCIENCE

GPU Integration into Data Centers for Science

- **2006–2008: Early Adoption of GPGPU -** NVIDIA launches CUDA, enabling GPUs for general-purpose computing (molecular dynamics, astrophysics)
- **2012: Breakthrough at Scale -** Titan (OLCF) supercomputer pioneer's hybrid CPU-GPU architecture (climate, materials science)
- **2015–2017: AI and Deep Learning Revolution -** GPUs become central to AI and machine learning. NVIDIA's Volta GPUs (V100) drive AI-accelerated research (genomics, climate modeling)
- **2018–2020: Widespread GPU Adoption** - Summit (OLCF) and other top systems use GPUs for AI and traditional HPC tasks (healthcare, energy, and materials science)
- **2023–2024: Exascale Era and Democratization of AI** - Systems like Aurora (ALCF) and Frontier (ORNL) leverage GPUs for exascale computing, supporting large-scale simulations, AI, and data-driven research

ALCF Systems Evolution



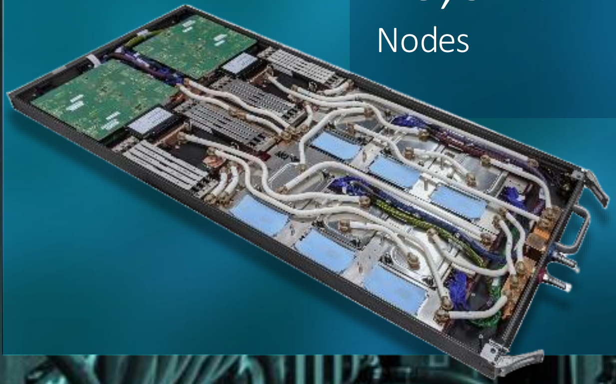
Aurora Specifications

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 #

TOP500 Supercomputers

- Fastest machines in the world, according to HPL

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	OpenAI
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
6	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	1,305,600	270.00	353.75	5,194
7	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	241.20	306.31	7,494
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 32C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR, EVIDEN EuroHPC/BSC Spain	663,040	175.30	249.44	4,159
9	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
10	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia NVIDIA Corporation United States	485,888	121.40	188.65	NVIDIA

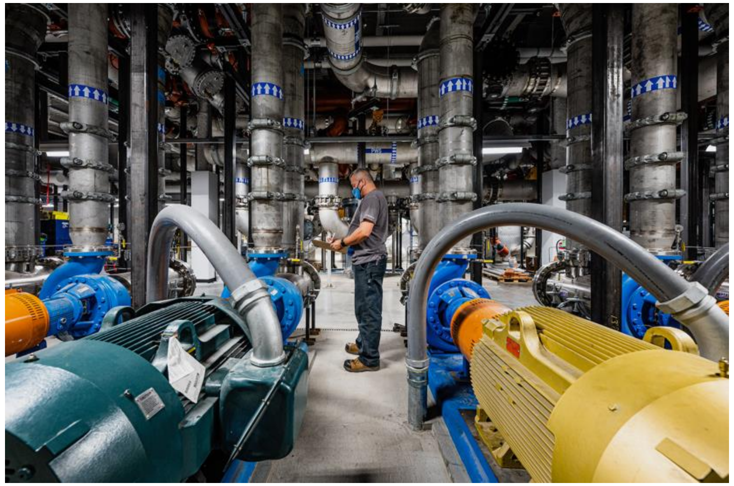
June 2024

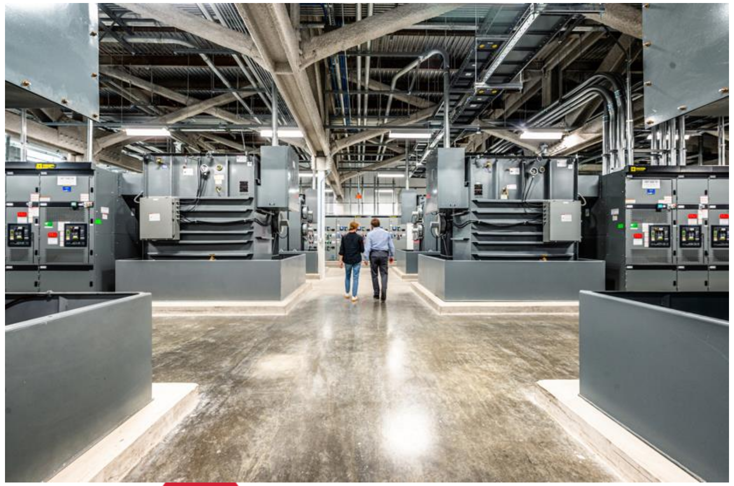
Rank	Site	Computer	Cores	HPL-MxP (Eflop/s)	TOP500 Rank	HPL Rmax (Eflop/s)	Speedup
1	DOE/SC/ANL	Aurora	9,264,128	10.600	2	1.0120	10.5
2	DOE/SC/ORNL	Frontier	8,699,904	10.200	1	1.2060	8.5
3	EuroHPC/CSC	LUMI	2,752,704	2.350	5	0.3797	6.2
4	RIKEN	Fugaku	7,630,848	2.000	4	0.4420	4.5
5	EuroHPC/CINECA	Leonardo	1,824,768	1.842	7	0.2412	7.6
6	DOE/SC/ORNL	Summit	2,414,592	1.411	9	0.1486	9.5
7	NVIDIA	Selene	555,520	0.630	15	0.0635	9.9
8	DOE/SC/LBNL	Perlmutter	888,832	0.590	14	0.0792	7.4
9	FZJ	JUWELS BM	449,280	0.470	21	0.0441	10.7
10	GENCI-CINES	Adastra	319,072	0.303	20	0.0461	6.6

- Fastest machines in the world, according to HPL-MxP

Role of HPC Facilities in Advancing AI

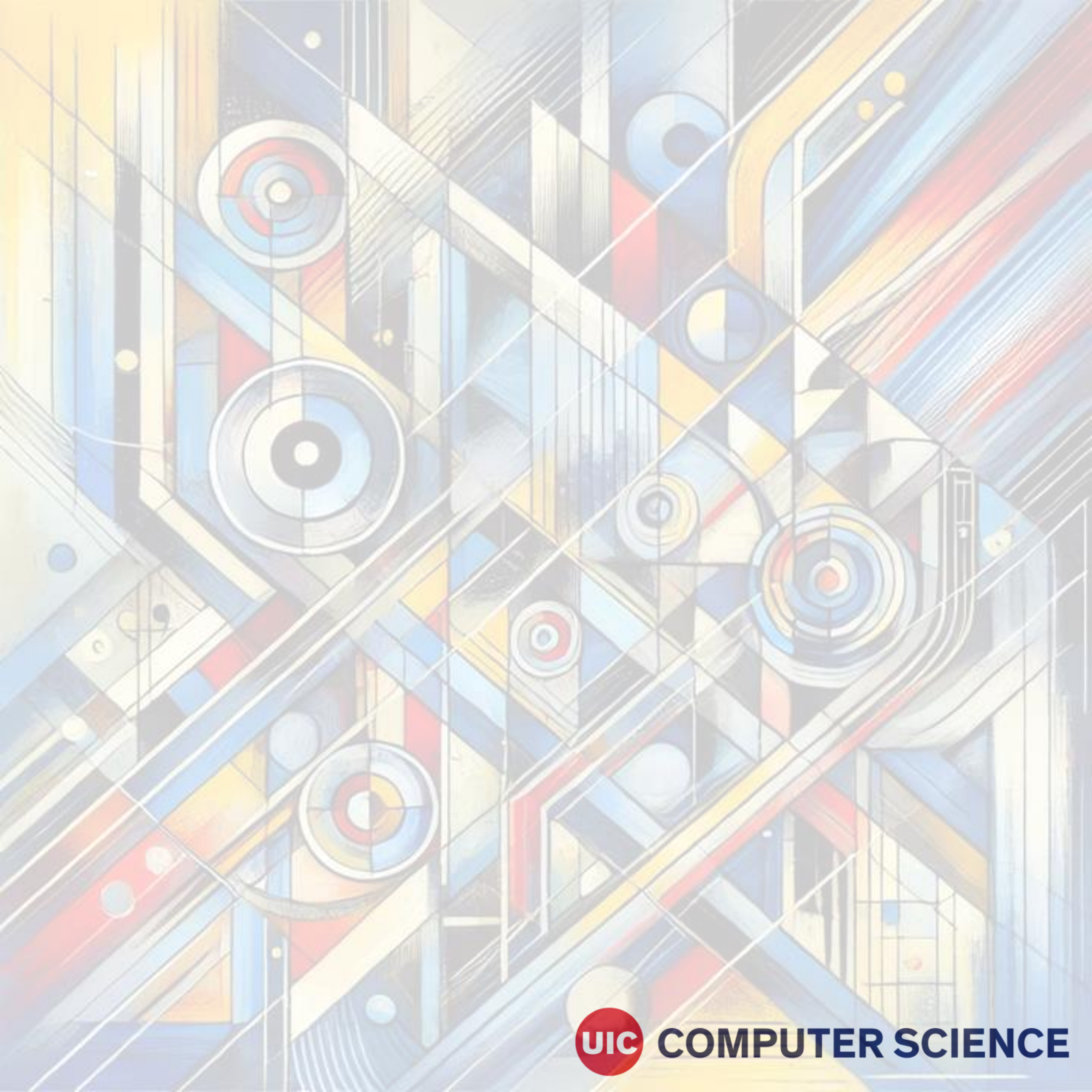
- **AI Model Scaling:** HPC enables the training of larger, more complex AI models that would not be feasible on traditional systems
- **Infrastructure Support:** Specialized hardware (like GPUs) and high-speed networks at scale tailored for optimizing AI workflows
- **Collaboration and Accessibility:** Open up AI research by democratizing access to resources for diverse and underfunded research communities





Overview

- Evolution of HPC with AI
- **Challenges**
- Opportunities
- Future Directions



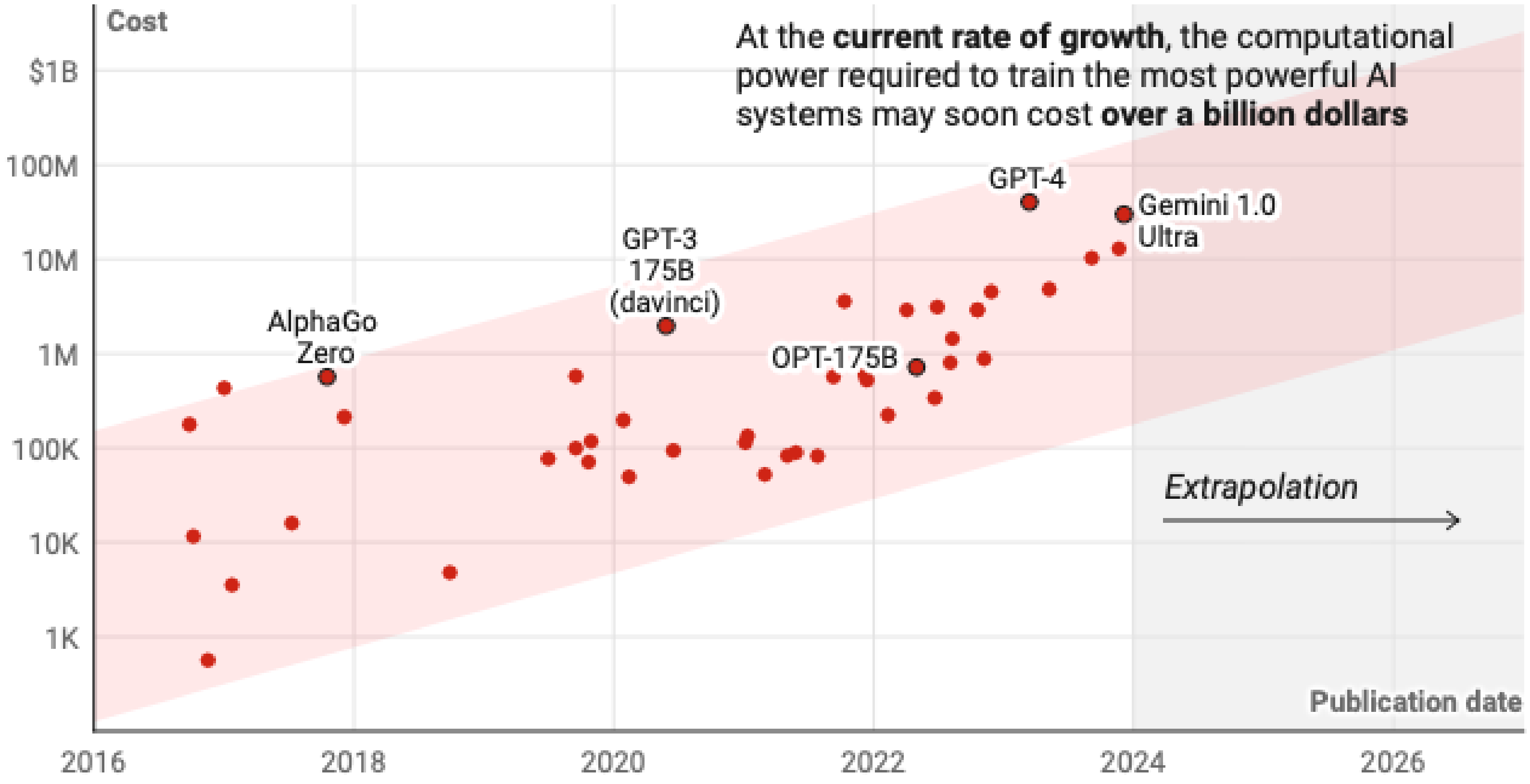
Frontier versus Foundation

- **Foundation models** are broad, versatile models pre-trained on large datasets, which can be adapted (fine-tuned) for specific tasks [GPT-4 (OpenAI), PaLM 2 (Google), Claude (Anthropic), Gemini (Google DeepMind), LLaMA 3 (Meta), Mistral, Falcon]
- **Frontier models** push the cutting edge of technology and AI capabilities, often built on new architectures or techniques, such as exascale computing systems

A frontier model can be a foundation model if it's at the cutting edge!

Cost* of Compute Power to Train Frontier AI

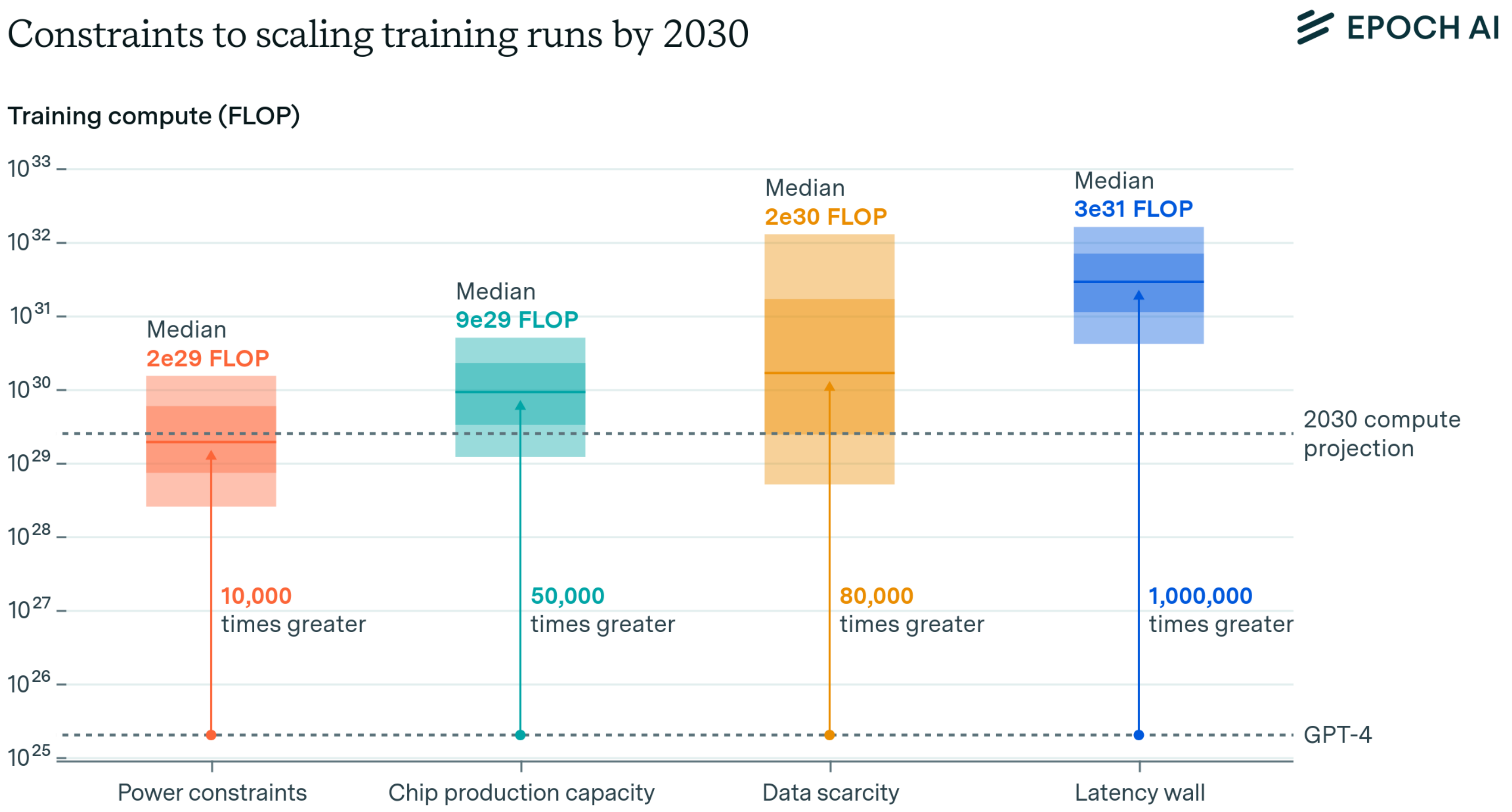
**Cost includes amortized hardware acquisition and energy consumption*



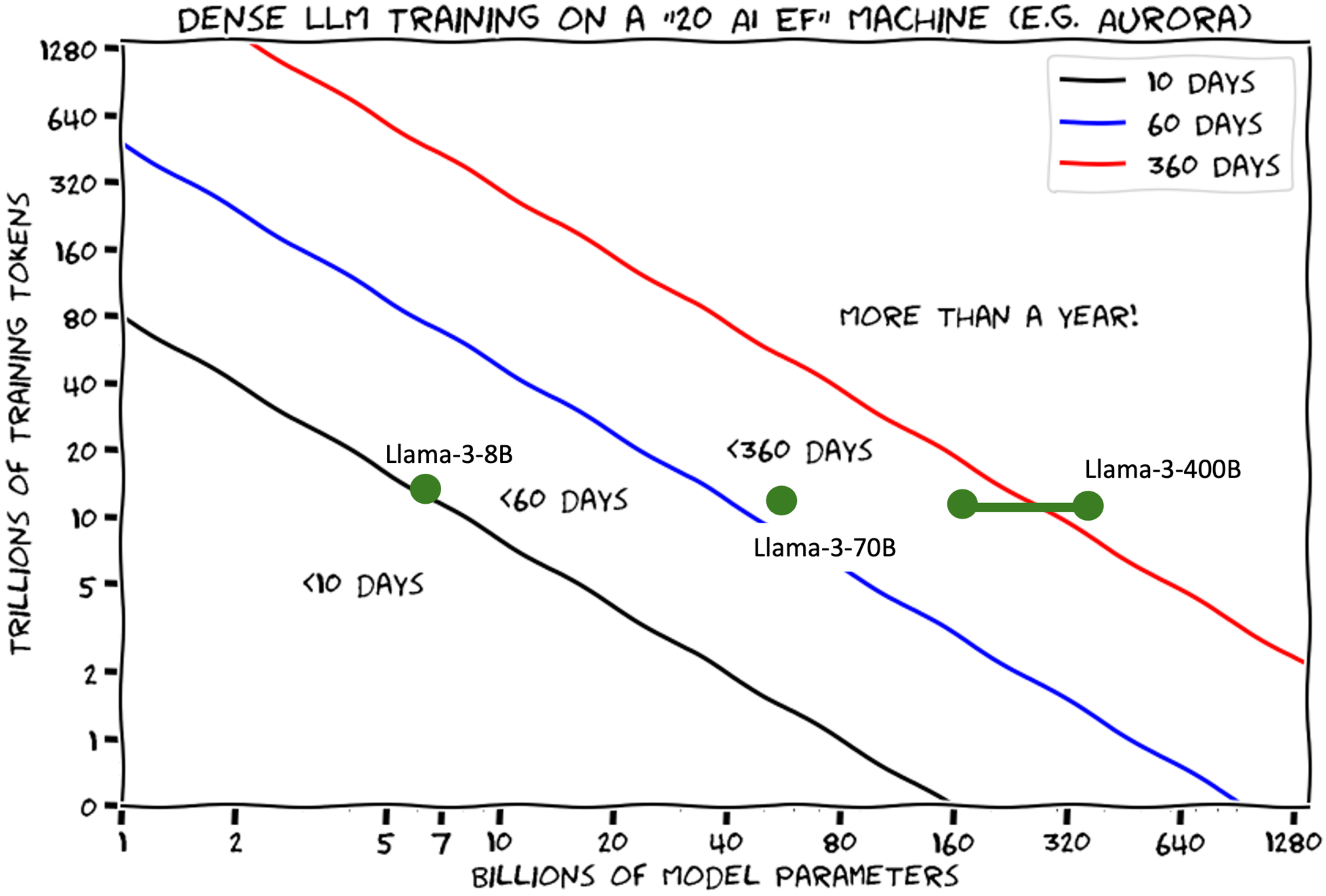
The cost of the computational power required to train the most powerful AI systems has doubled every nine months

Source: Will Henshall for Time, **Data Source:** Epoch AI

Constraints to Scaling Training Runs by 2030



Source: Can AI Scaling Continue Through 2030? **Data Source:** EPOCH AI



Source: Rick Stevens – Argonne National Laboratory

Real World Experience

- “A day-long run without a system failure would be outstanding,” ... “Our goal is still hours but longer than Frontier’s current failure rate”, ... “we’re not super far off our goal. The issues span lots of different categories, the GPUs are just one.” - Dan Swinhoe, *Frontier supercomputer suffering ‘daily hardware failures’ during testing* in **Data Centre Dynamics**, October 10, 2022
- *Faulty Nvidia H100 GPUs and HBM3 memory caused half of failures during LLama 3 training — one failure every three hours for Meta's 16,384 GPU training cluster* - Anton Shilov, **tom's Hardware**, July 27, 2024

Llama 3 405B Interruptions (54 days)

419 unexpected interruptions:

- 148 (30.1%) various GPU failures (including NVLink failures)
- 72 (17.2%) were caused by HBM3 memory failures
- 2 CPUs failed

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-trainin
f unexpected interruptions were attributed to confirmed or suspected hardware issues.

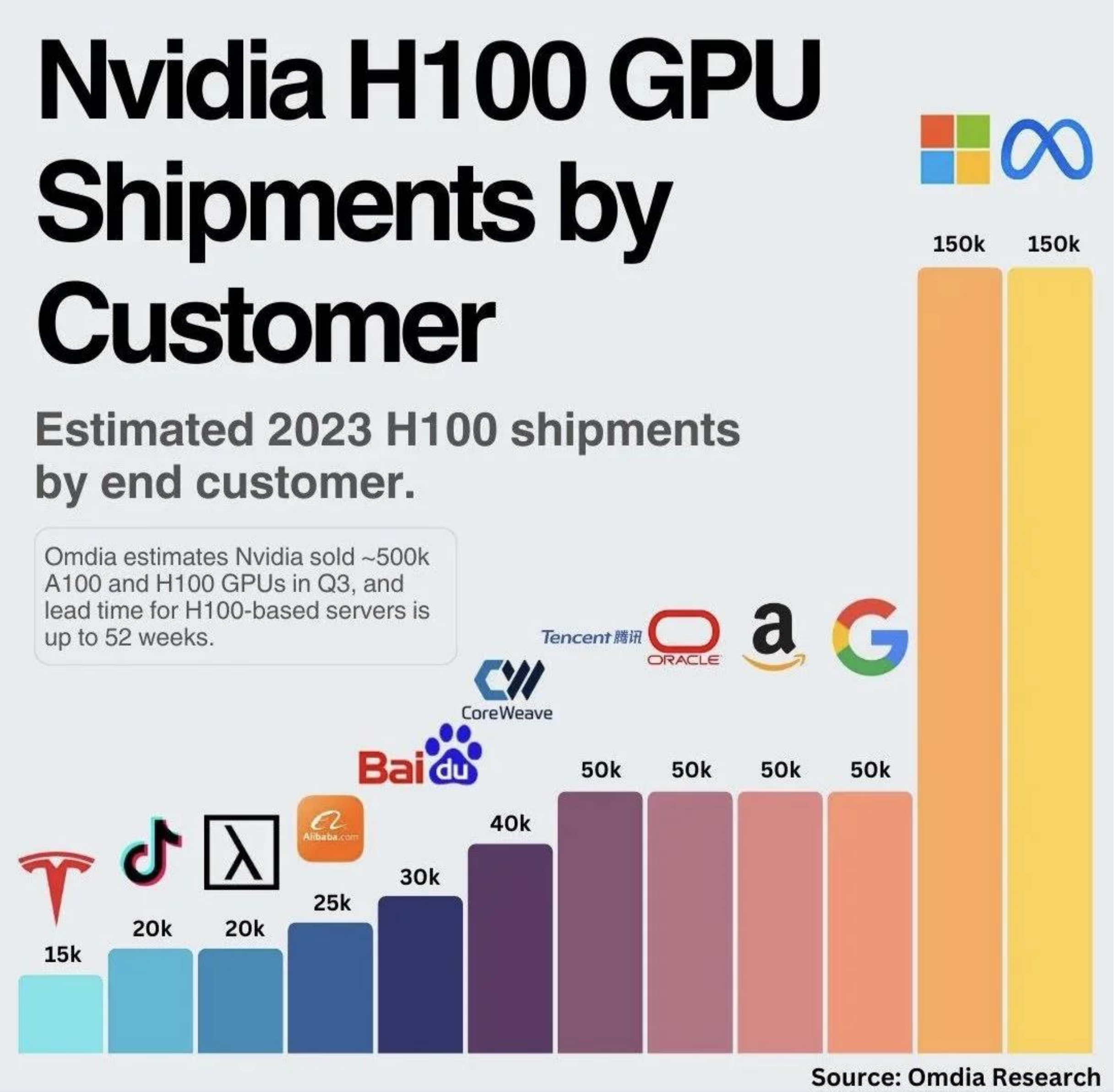
Tom's HARDWARE, July 27, 2024

Scaling of Systems

Headlines of 2024

- Tesla bought 50K, planning 100K upgraded 300K GPU systems
- Meta bought 350K GPUs

Number of GPUs	MTBF (hours)
50,000	13.58
100,000	7.05
500,000	1.87
1,000,000	1.28



Aurora Mean Time Between Failures (exercise)

Problem: Aurora has 63,744 GPUs. We need to determine the system-wide Mean Time Between Failures (MTBF) based on the number of GPUs and the given failure probability.
Approach: Calculate the system-wide MTBF by accounting for the number of GPUs and their individual failure probabilities.
Given: Number of GPUs ($N = 63,744$), Failure probability per GPU ($P_{\text{GPU}} = 10^{-4}$)
Formula:

$$P_{\text{system failure}} \approx N \cdot P_{\text{GPU}}$$

$$\text{MTBF(system)} = \frac{1}{N \cdot P_{\text{GPU}}}$$

Calculation:

$$P_{\text{system failure}} = 63,744 \times 10^{-4} = 6.3744$$

$$\text{MTBF(system)} = \frac{1}{6.3744} \approx 10.76 \text{ hours}$$

Conclusion: The system-wide MTBF is around 10.76 hours, which indicates that minimizing the per-GPU failure rate is crucial to improving system reliability.

Feasibility of Training Models on Aurora/Polaris

AuroraGPT set of models (1.5B, 7B, 13B, 70B, 200B, 1T, …)

Aurora BFP16 HGEMM ~ 180 TF per tile x (127,488 tiles) ⟹ 22.9 EF/s

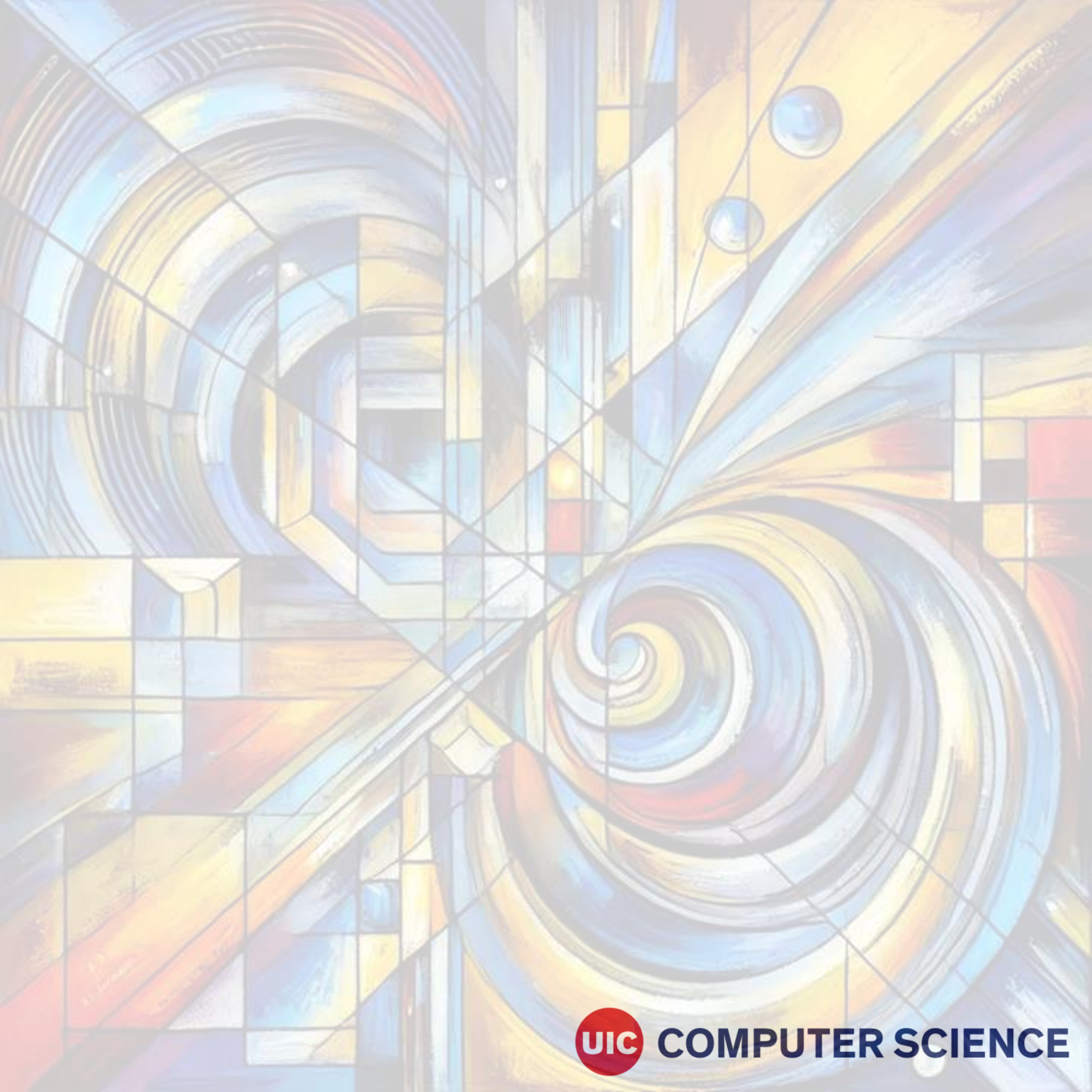
Model Size (# of Parameters in Billions)	Training Tokens (Trillions)	Training F/P/T	Total Training Compute (Flops in BF16)	Total Training Compute (EF-days)	Aurora Time (Days)	Aurora Time (Hours)	Polaris Time (Days)	Polaris Time (Hours)	Cloud Cost ($3 GPU/hr)
1.5	1	6	9E+21	0.10	0.01	0.25	1	36	$46,871
1.5	2	6	1.8E+22	0.21	0.02	0.49	3	71	$93,741
1.5	3	6	2.7E+22	0.31	0.03	0.74	4	107	$140,612
7	1	6	4.2E+22	0.49	0.05	1.14	7	167	$218,729
7	2	6	8.4E+22	0.97	0.10	2.29	14	333	$437,459
7	3	6	1.26E+23	1.46	0.14	3.43	21	500	$656,188
70	2	6	8.4E+23	9.72	0.95	22.88	139	3,333	$4,374,588
70	3	6	1.26E+24	14.58	1.43	34.31	208	5,000	$6,561,882
70	4	6	1.68E+24	19.44	1.91	45.75	278	6,667	$8,749,176
200	6	6	7.2E+24	83.33	8.17	196.08	1,190	28,571	$37,496,471
200	10	6	1.2E+25	138.89	13.62	326.80	1,984	47,619	$62,494,118
200	15	6	1.8E+25	208.33	20.42	490.20	2,976	71,429	$93,741,176
1000	10	6	6E+25	694.44	68.08	1633.99	9,921	238,095	$312,470,588
1000	20	6	1.2E+26	1388.89	136.17	3267.97	19,841	476,190	$624,941,176
1000	30	6	1.8E+26	2083.33	204.25	4901.96	29,762	714,286	$937,411,765

We are assuming about 40% efficiency for LLM BFP16 flops utilization relative to HGEMM measurements

Will every domain build its own model? Need their own system?

Overview

- Evolution of HPC with AI
- Challenges
- **Opportunities**
- Future Directions



UIC COMPUTER SCIENCE

Data

Scientific & Engineering Datasets

- Mathematics
- Biology
- Materials
- Chemistry
- Particle Physics
- Nuclear Physics
- Computer Science
- Climate
- Medicine
- Cosmology
- Fusion Energy
- Accelerators
- Reactors
- Energy Systems
- Manufacturing

Text and Code Corpora

- General Text
- Media
- News
- Humanities
- History
- Law
- Digital Libraries
- OSTI Archive
- Scientific Journals
- arXiv
- Code repositories
- Data.gov
- PubMed
- Agency Archives

DOE and NNSA Exascale Systems
FASST Common AI Software Frameworks
FASST Responsible AI Techniques

Compute

Training

Training

Models

Open Science Foundation Models

National Security Foundation Models

Integrated Research Infrastructure
Online Experimental Facilities
Strategic Partnerships

Applications

Tuned and Adapted Downstream Models

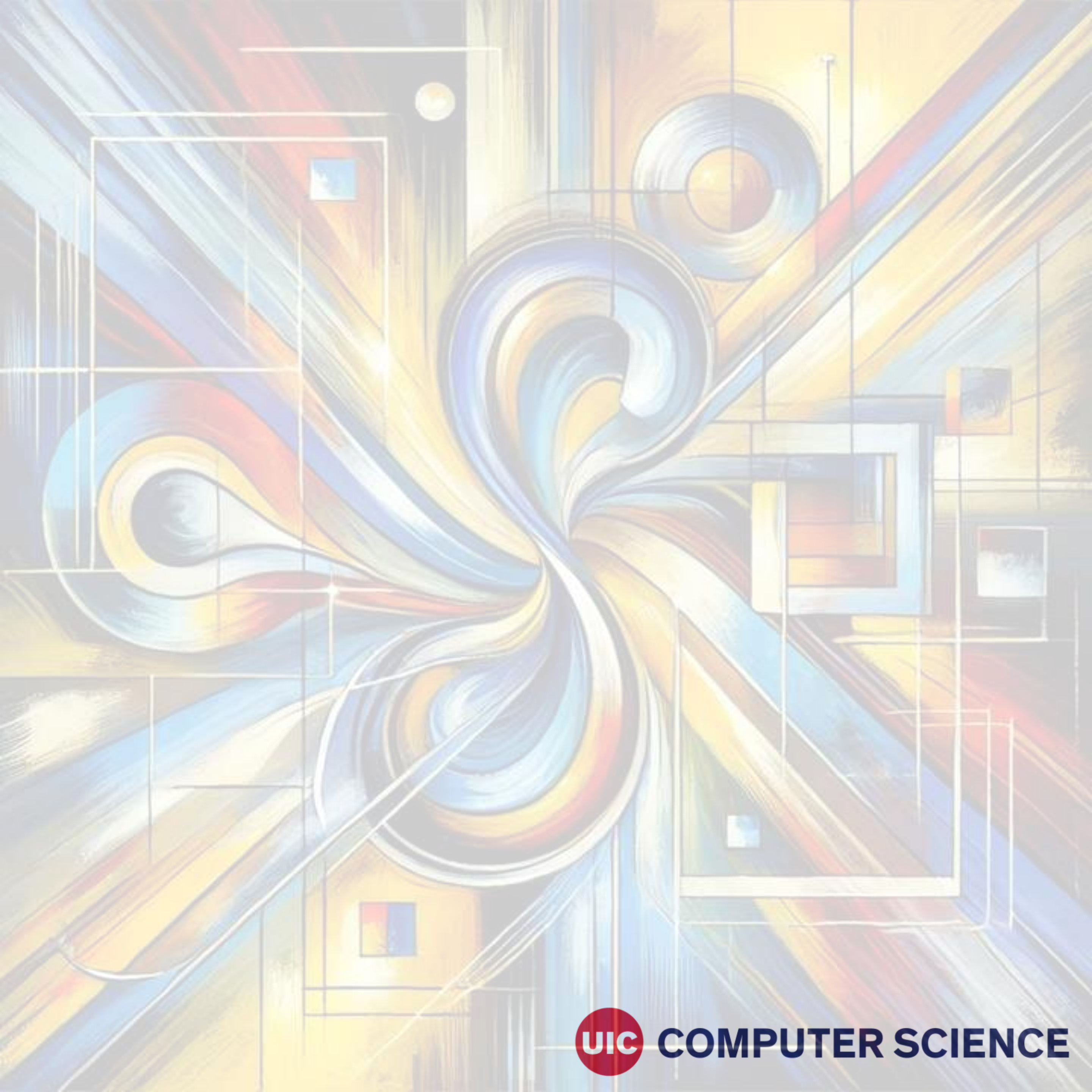
Exemplar DOE Mission Tasks

- Scientific Discovery
- Digital Twins
- Inverse Design
- Code Optimization
- Accelerated Simulations
- Autonomous Experiments
- Secure Data Infrastructure
- Co-Design

UIC COMPUTER SCIENCE

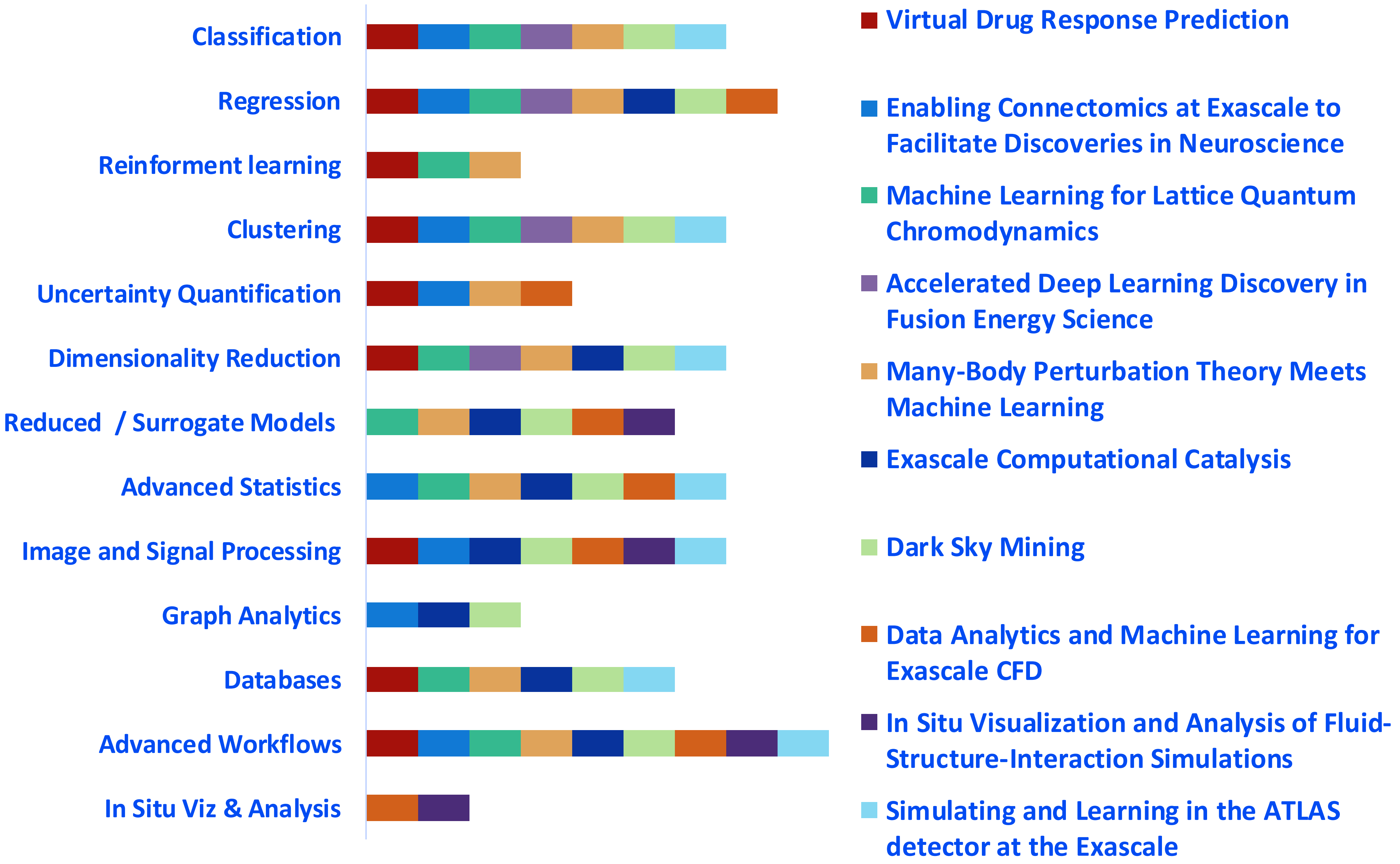
Overview

- Evolution of HPC with AI
- Challenges
- Opportunities
- **Future Directions**



UIC COMPUTER SCIENCE

AURORA ESP Data and Learning Projects and Methods



ALCF AI Testbeds

Cerebras (CS-2)

SambaNova

Graphcore

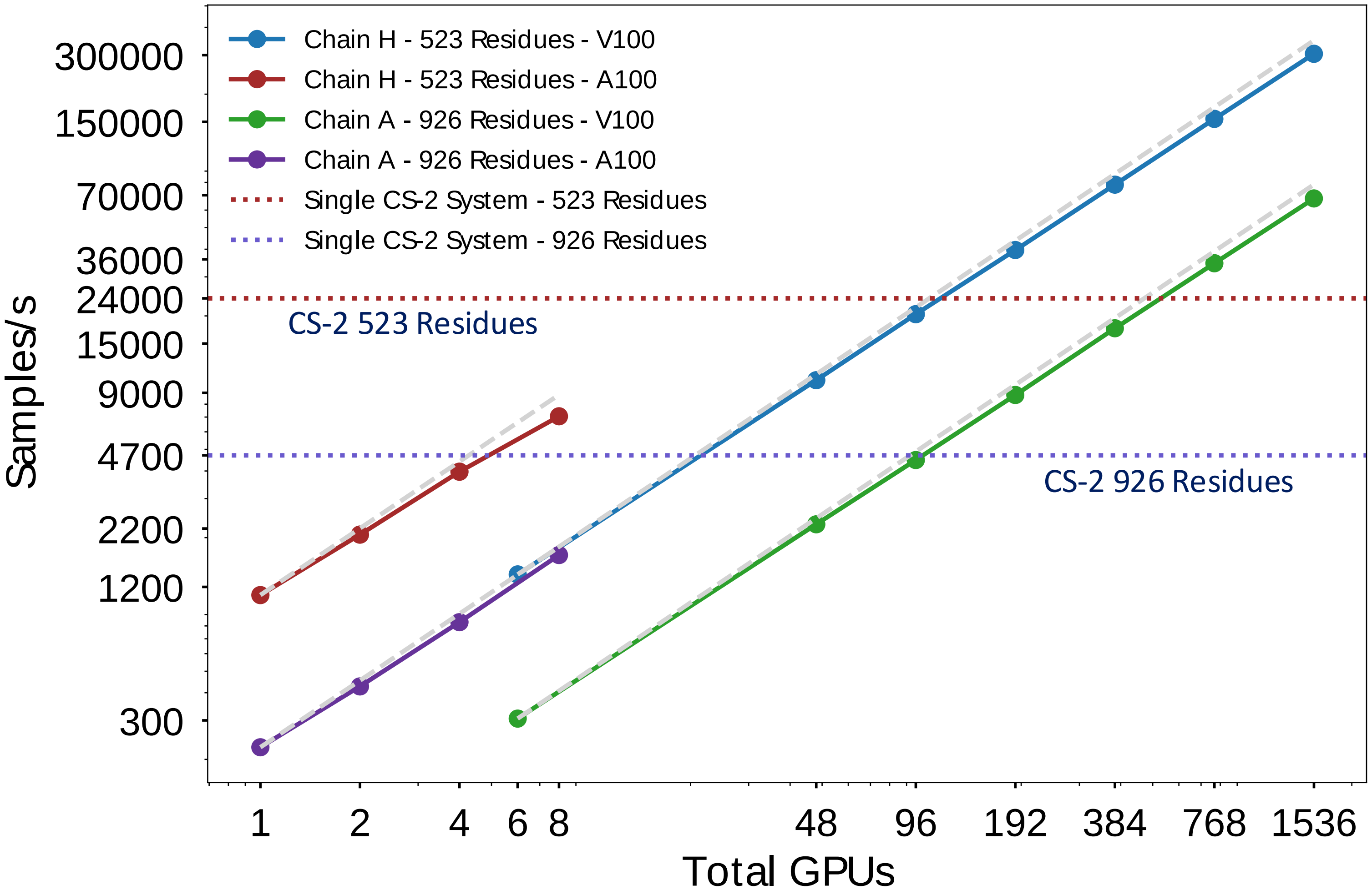
Habana

Groq

- Infrastructure of next-generation machines with hardware accelerators customized for artificial intelligence (AI) applications.
- Provide a platform to evaluate usability and performance of machine learning based HPC applications running on these accelerators.
- The goal is to better understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

	Cerebras CS2	SambaNova Cardinal SN10	Groq GroqCard	GraphCore GC200 IPU	Habana Gaudi1	NVIDIA A100
Compute Units	850,000 Cores	640 PCUs	5120 vector ALUs	1472 IPUs	8 TPC + GEMM engine	6912 CUDA Cores
On-Chip Memory	40 GB	>300MB	230MB	900MB	24 MB	192KB L1 40MB L2
Process	7nm	7nm	14nm	7nm	7nm	7nm
System Size	2 Nodes	2 nodes (8 cards per node)	4 nodes (8 cards per node)	1 node (8 cards per node)	2 nodes (8 cards per node)	1 card
Estimated Performance of a card (TFlops)	>5780 (FP16)	>300 (BF16)	>188 (FP16)	>250 (FP16)	>150 (FP16)	312 (FP16), 156 (FP32)
Software Stack Support	Tensorflow, Pytorch	SambaFlow, Pytorch	GroqAPI, ONNX	Tensorflow, Pytorch, PopArt	Synapse AI, TensorFlow and PyTorch	Tensorflow, Pytorch, etc
Interconnect	Ethernet-based	Infiniband	RealScale ™	IPU Link	Ethernet-based	NVLink

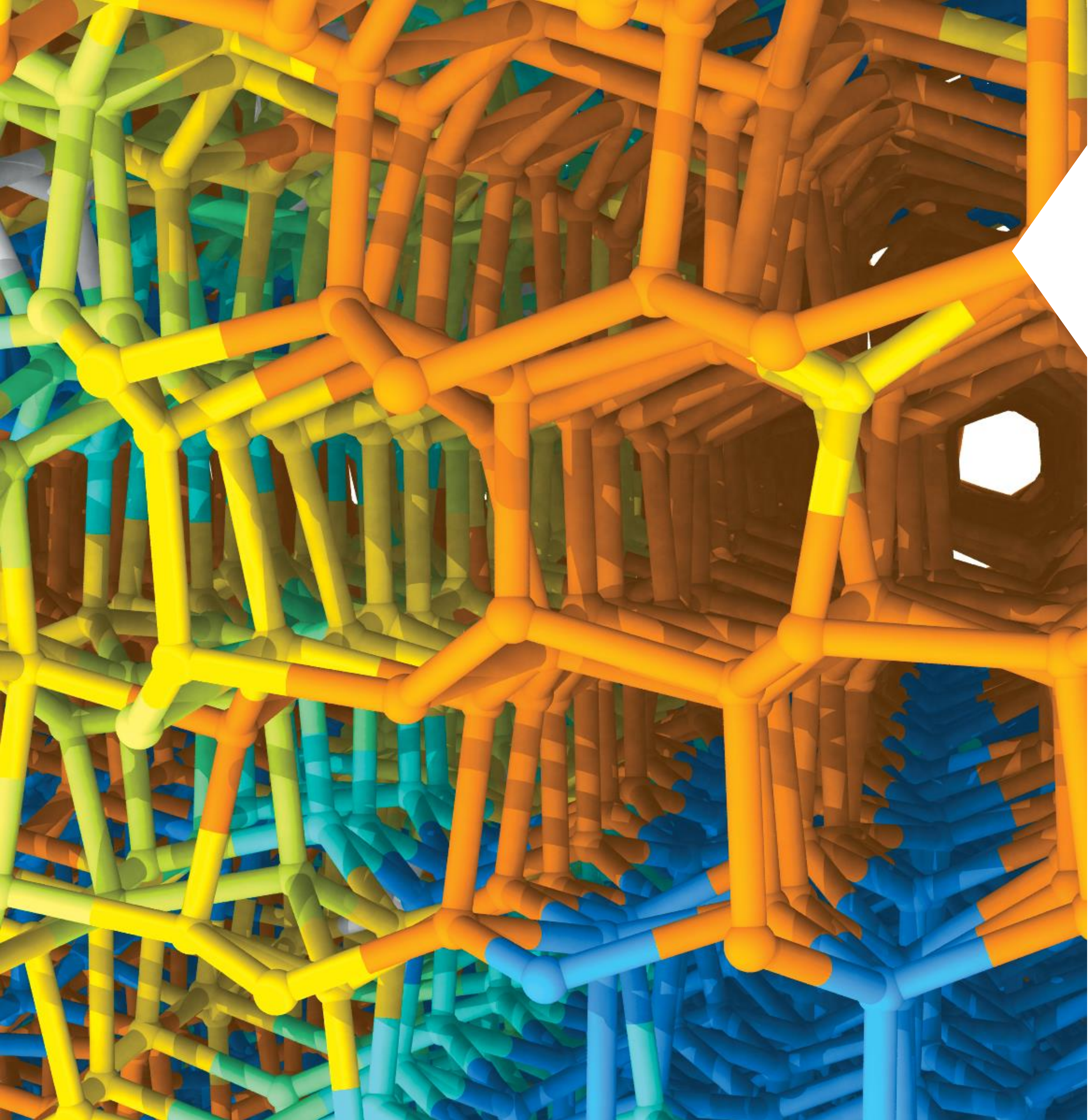
COVID-19 CVAE Training on Summit and Cerebras CS-2



- Single CS-2 delivers performance of over 100 GPUs on CVAE
- Results are for **out-of-the-box performance** based on model config not optimized for CS-2.

Performance	523 X 523	926 X 926
Throughput (samples/sec)		
1x CS-2 System	24,000	4700
1x V100 GPU	228	51
1x A100 GPU	~1100	~150
Speedup (CS2 vs. GPU)		
1 x V100 GPU	113x	101x
1 x A100 GPU	~22X	~32X

Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action, SC21 COVID19 Gordon Bell Finalist, To appear in IJHPCA 2022 https://www.biorxiv.org/content/10.1101/2021.10.09.463779v1.full.pdf



Getting Started on ALCF AI Testbed

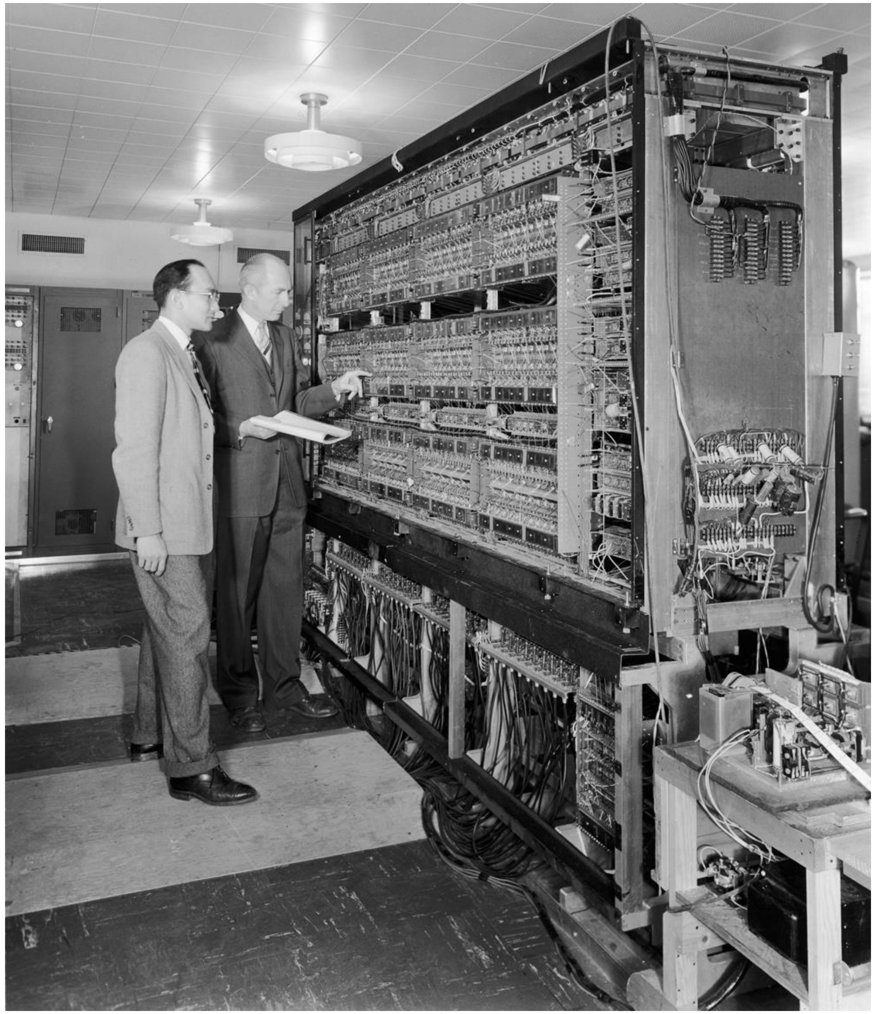
Director’s Discretionary (DD) awards support various project objectives from scaling code to preparing for future computing competition to production scientific computing in support of strategic partnerships.

Cerebras CS-2 and SambaNova Datascale systems are available for allocations!

- Allocation Request Form
- AI Testbed User Guide

UIC COMPUTER SCIENCE

First and Latest Argonne Computer



Thank You

COMPUTER SCIENCE
COLLEGE OF ENGINEERING