Hewlett Packard Enterprise

Technologies for Future HPC and AI Systems

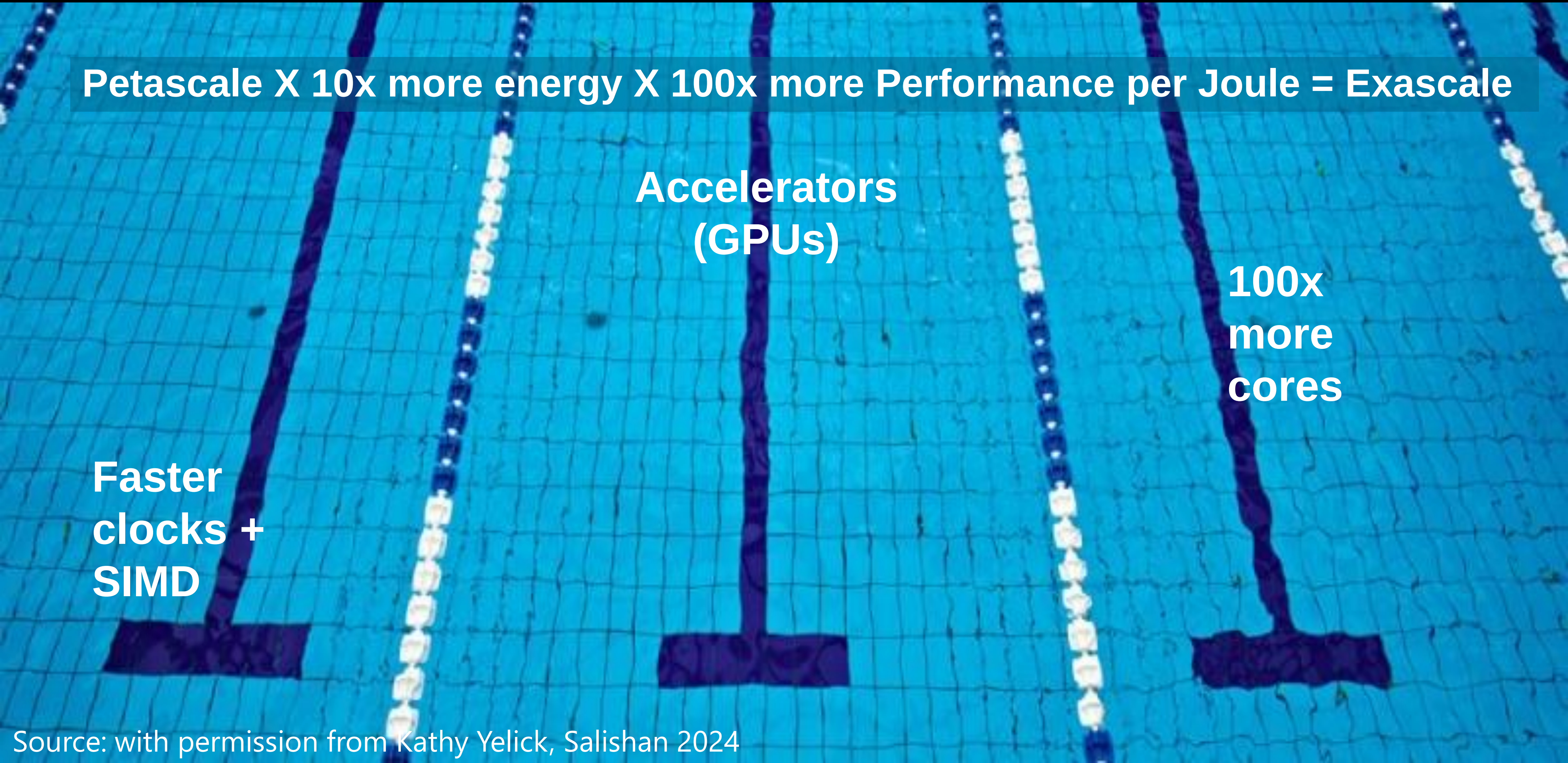
Dr. Robert W. Wisniewski
HPE Fellow
Chief Architect HPC and AI Solutions

March 18, 2025

Agenda

- Quick review of how we arrived at exascale
- Technologies to move us forward

Exascale Architecture Plans (2008)



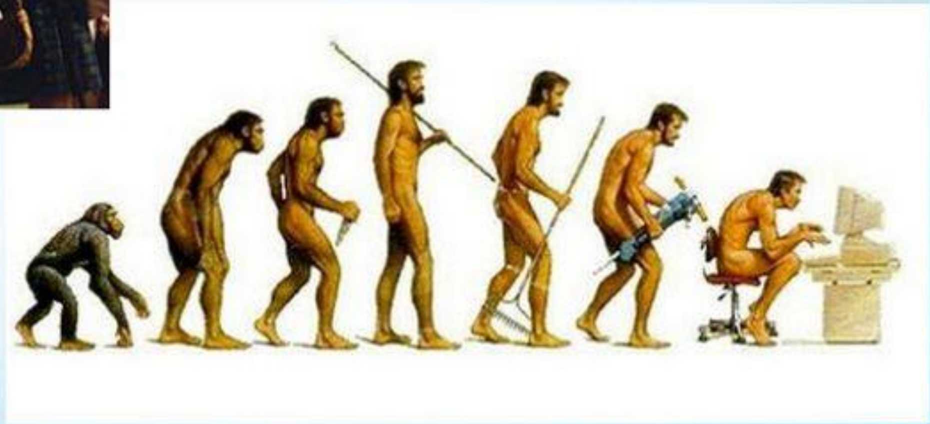
Source: with permission from Kathy Yelick, Salishan 2024

The Swim Lanes

Source: Wisniewski Salishan 2011

How to Get There

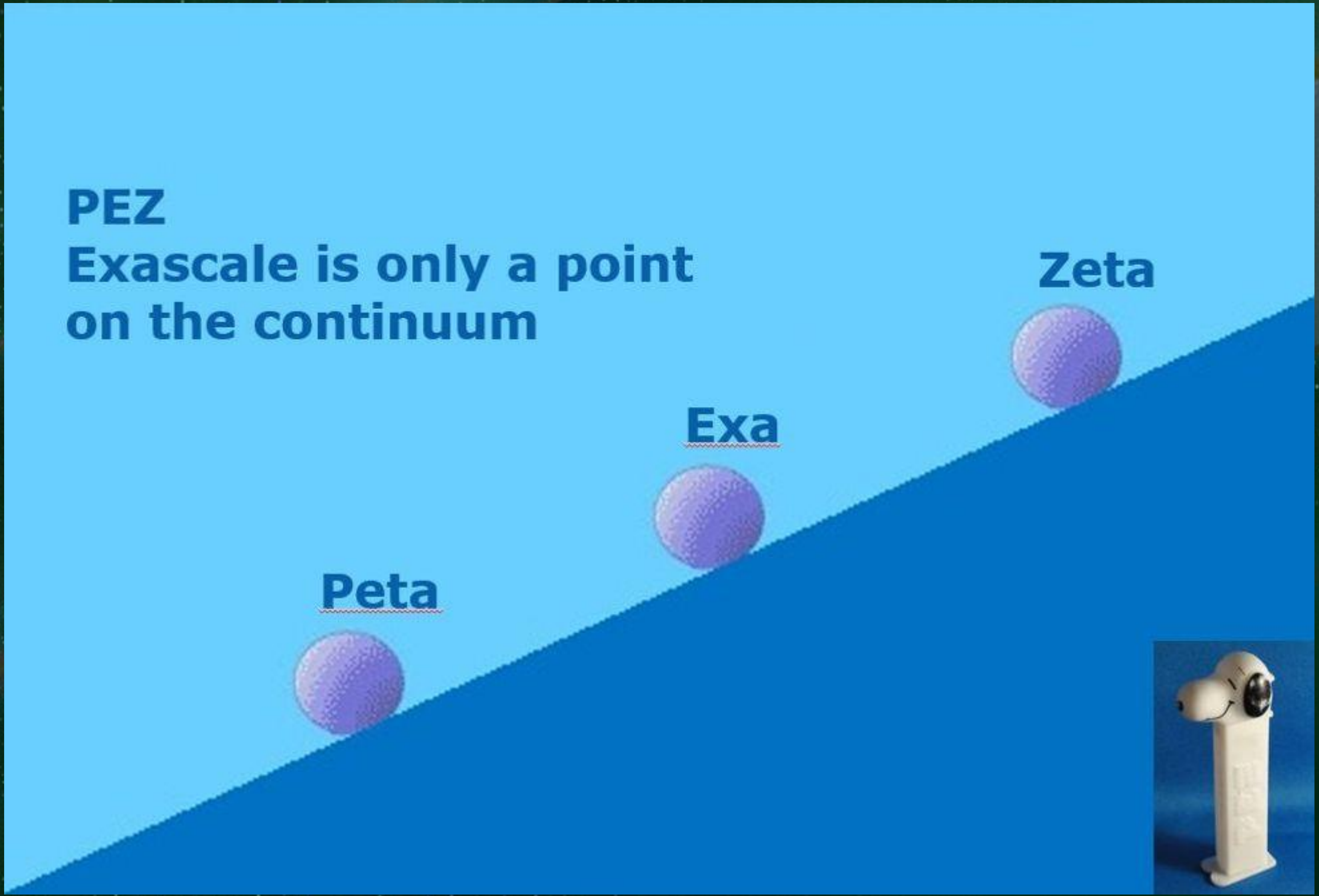
Revolutionary versus Evolutionary



- Which one ?

Source: Wisniewski SOS 2014

PEZ – A Continuum



Source: Wisniewski SOS 2014

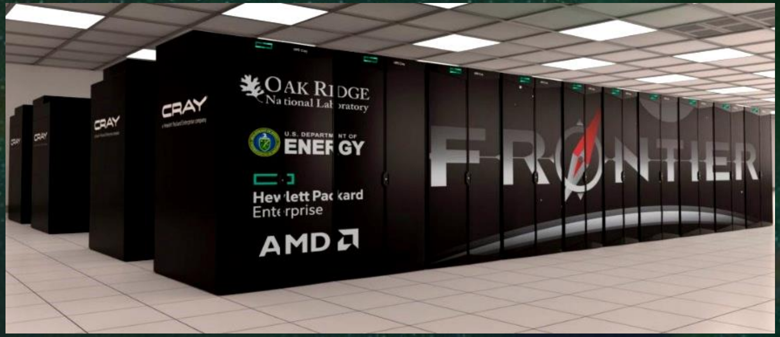
HPE Large-Scale HPC and AI Machines

Helping organizations tackle the grand challenges of humankind

37,632
GPUs

63,744
GPUs

44,544
APUs



100% liquid-cooled HPE Cray EX supercomputer

High performance GPU accelerated blades

HPE Slingshot exascale interconnect

Cray ClusterStor file systems

Enabling Large-Scaling AI Workloads Around the Globe

Los Alamos
NATIONAL LABORATORY
EST. 1943

10 EFLOPS

single-precision AI Performance
with NVIDIA GH200 superchips

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

20 EFLOPS

single-precision AI Performance
with NVIDIA GH200 superchips

NATIONAL COMPOSITES CENTRE

21 EFLOPS

single-precision AI Performance
with NVIDIA GH200 superchips

Power – A Little or A Lot

- Frontier 22.7 MW (https://en.wikipedia.org/wiki/Frontier_(supercomputer))
- Aurora 38.7 MW (https://en.wikipedia.org/wiki/Aurora_(supercomputer))

- Combined powers a small city (40K people)

Source: AI generated

Power – A Little or A Lot

- GPT-4 training used over 50 gigawatt-hours
  - 0.02% of the electricity California generates in a year
  - 2200 hours or 92 days on Frontier
  - 10T mode estimate 5000 gigawatt-hours
- LHC 200 MW

https://sites.uci.edu/energyobserver/2012/11/28/introduction-to-the-large-hadron-collider-at-cern-2/

Power – A Little or A Lot

- The Gigawatt Data Center Campus is Coming
- Amazon Web Services recently bought a data center co-located with a nuclear power facility, where it hopes to gradually deploy up to 960 megawatts
- https://www.datacenterfrontier.com/hyperscale/article/55021675/the-gigawatt-data-center-campus-is-coming

Source: AI generated

Parallelism and Fat versus Thin Nodes

- Sequoia 20PF circa 2012, had 96 racks *1024 nodes/rack +16 cores/node == 1,572,864 *4 threads/core == 6,291,456 threads
- Concern was we would need 50x that number of threads

https://en.wikipedia.org/wiki/Sequoia_(supercomputer)

Parallelism and Fat versus Thin Nodes

image source:
Oak Ridge National Lab

- Frontier is 74 cabinets, 128 nodes per cabinet
  - 1 AMD Epyc 7713 "Trento" CPU and 4 AMD Instinct MI250X GPUs per node
- Frontier has 9,472 CPUs * 64 cores/CPU == 606,208 cores * 2 threads/core == 1,212,416 threads
- Frontier has 37,888 GPUs each GPU has 2 GCD (Graphic Compute Dies) with 110 CU (Compute Units) per die == 8,335,360 cores with 64 threads (a wavefront) == 533,463,040 threads

Parallelism and Fat versus Thin Nodes

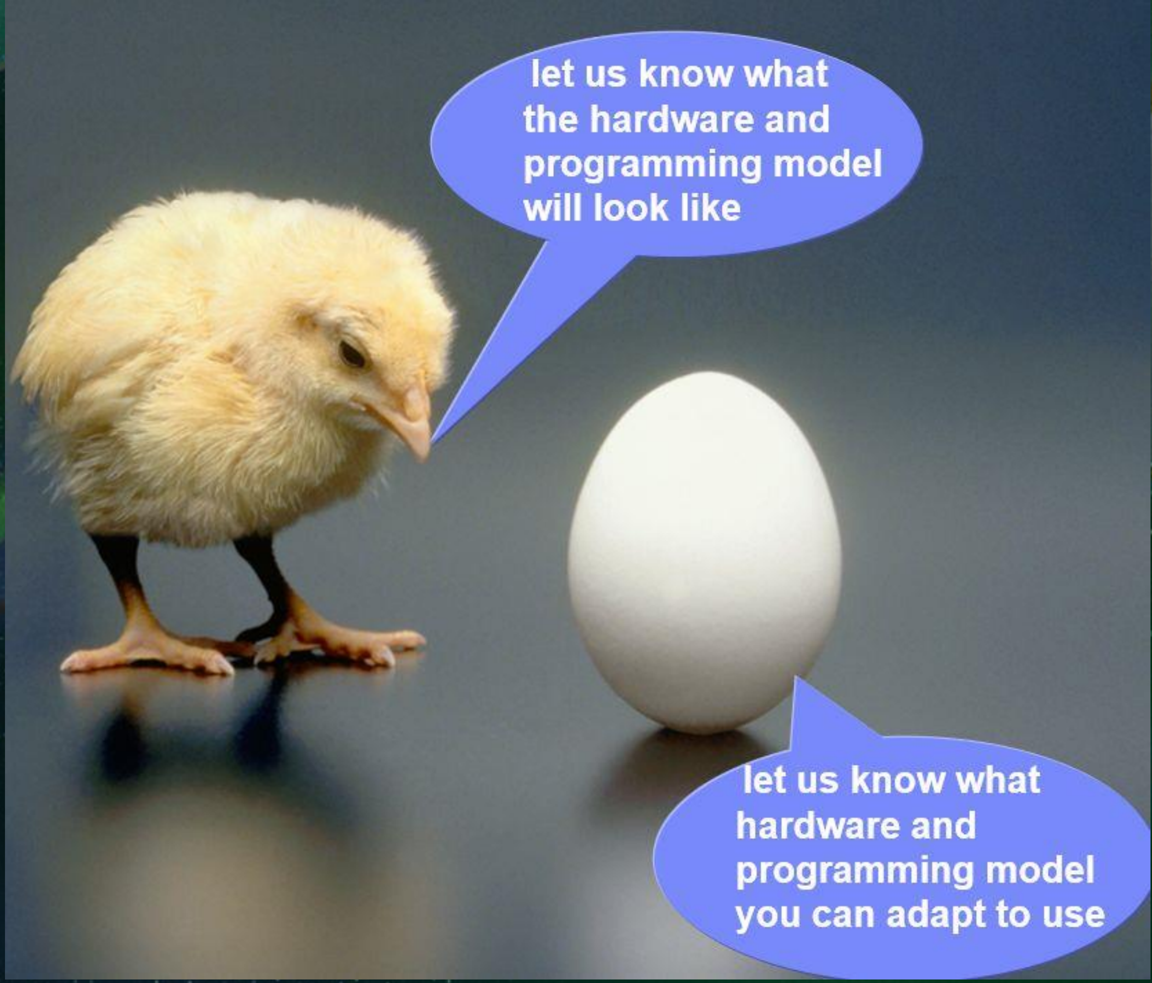
- Sequoia 6,291,456 threads
- Frontier
  - 1,212,416 CPU threads
  - 533,463,040 GPU threads

- The number of GPU threads exceeds what we thought thread count would be
  - The number of CPU threads is meaningfully less than predicted
  - GPU hardware and software help hide that high degree of parallelism

- Fat nodes help significantly
  - Lower surface to volume ratios reduces global communication
  - Fewer OSes put less pressure on the reliability of each instance
  - Fewer nodes ease the burden of providing scalable and reliable system management software

Software and Programming Model

- Programming model did not substantially change
  - Did not need all new language/runtime and model
  - MPI + X still here
  - Kokkos and Raja emerged and their usage broadened
  - Kokkos also helping drive C++ standards
- Moving to GPUs was a massive effort, but primarily due to accelerator model and parallelism rather than the GPU itself
- Fat nodes relieve some of the software scalability challenges
  - Helped with reliability challenges due to absolute number of instance of software stack running
    - Has not solved hardware MTBF



Source: Wisniewski Salishan 2011

Where We are Going: Taking Stock

- Thought we were going to do it in 20MW
  - Many people did not think so, but that was the target
- Thought it was going to take a new programming model and rewrite of all codes
  - There was a massive effort to restructure codes for GPUs
    - –Will the work that was done, at least for the codes that utilized Kokkos or Raja, carry forward
- Thought parallelism was going to swamp us
  - It grew, but we managed to [mostly] hide it with a hierarchical layer
- Thought reliability was going to require fault tolerant computing
  - We managed to eke this one out, but MTBF for capability jobs is counted in hours now
- New theme: mixed precision playing an increasing important role
- New theme: AI

- HPC has become like an aircraft carrier

Source: AI generated

Source: AI generated

Where We Are Going: Technical Themes

- Hide complexity behind a layer
  - Threading, parallelism: small and large, programming model
- Improve performance through tighter coupling
  - Compute to memory, compute to compute, compute to communication
- Macro heterogeneity
  - Quantum common example, but perhaps more : AI training, AI inference, HPC
- Handle reliability
  - Enhance approach to fault tolerance, tolerate failures in the small at least
- Complex workflows
  - Spanning machines and sites
  - Spanning edge to supercomputer to cloud
  - Containing massive and secured data
- Sustainability and power

Arkouda

An open-source Python package providing interactive data analytics at supercomputing scale.

>>> Transform the way you work with big data

EASY TO USE

Provides an API data scientists are familiar with based on Pandas/NumPy

FAST & SCALABLE

Outperforms NumPy on a single Node and has scaled up to 8,000+ Nodes

POWERED BY CHAPEL

Powered by a parallel distributed server written in Chapel

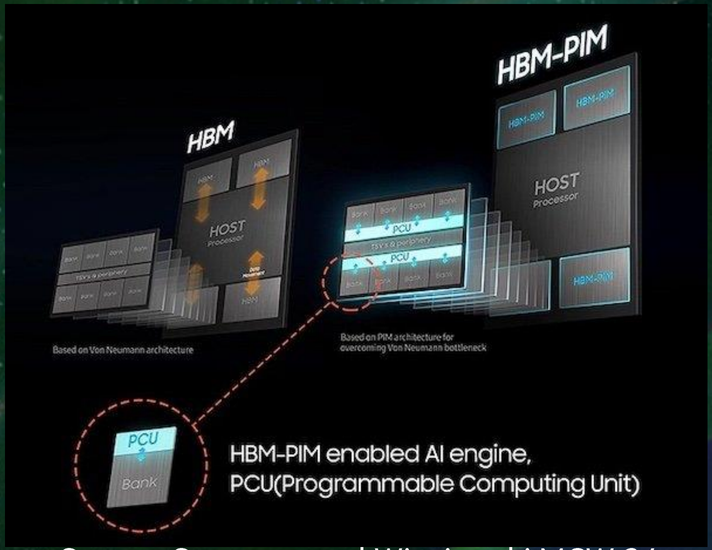
EXTENSIBLE & CUSTOMIZABLE

Extend Arkouda's capabilities by creating specialized functions

Tight Coupling

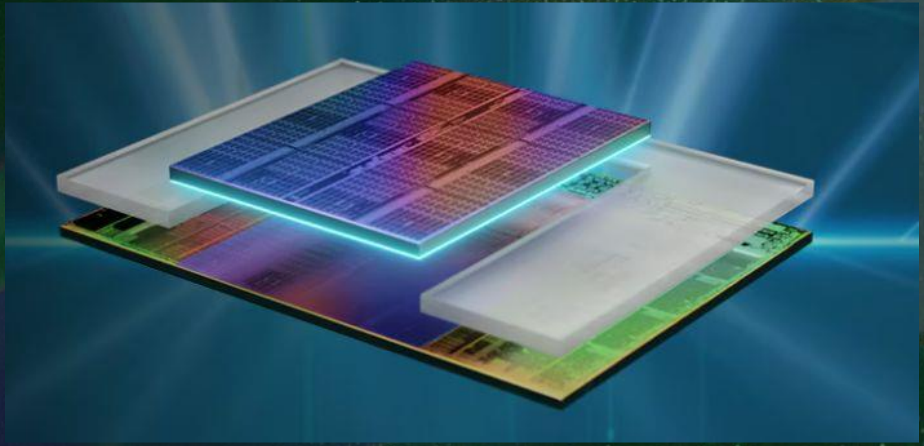
Samsung AXDIMM

Source: Samsung and Wisniewski MCW 24



Source: Samsung and Wisniewski MCW 24

AMD V-Cache

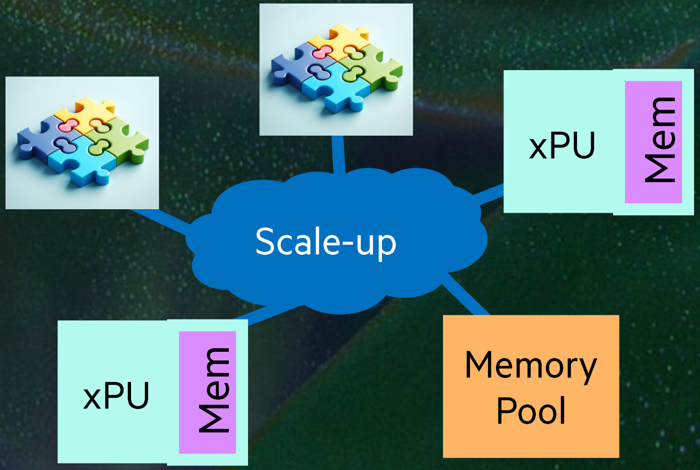


https://www.amd.com/en/products/processors/technologies/3d-v-cache.html

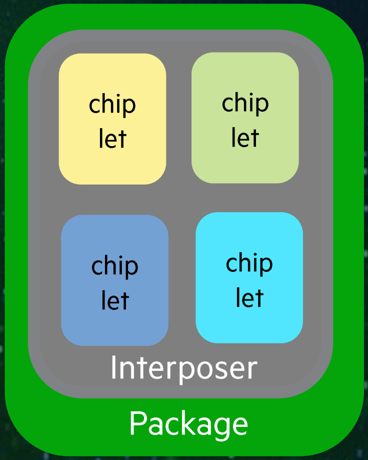
Compute to memory

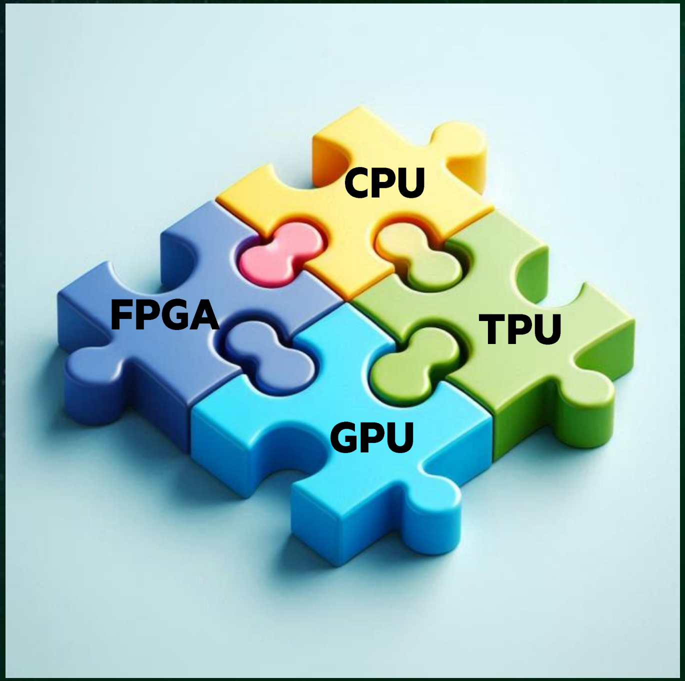
Compute to compute

Compute to communication



Source: AI generated





Source: AI generated

Quantum Computing Integration at HPE

Integrating classical and quantum systems

to harness diverse accelerators that maximize run-time, efficiency, sustainability, and security

Unified workflow environment

Simplify the end user experience

Software framework to harness accelerators most suitable for each segment of a workflow

Large-scale quantum simulation

Toward industrial scale

HPC systems used to simulate and test quantum advancements

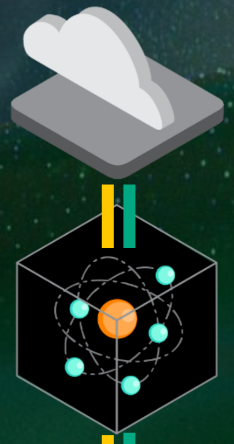
Quantum-inspired accelerators

Solve intractable problems

Non-conventional acceleration of algorithms explored by the quantum computing community

Hewlett Packard **Labs** + **HPE HPC & AI Business Group** + **Innovation partners** (academic, industrial, government)

Integration of quantum accelerators



Accelerators

FPGAs

GPUs

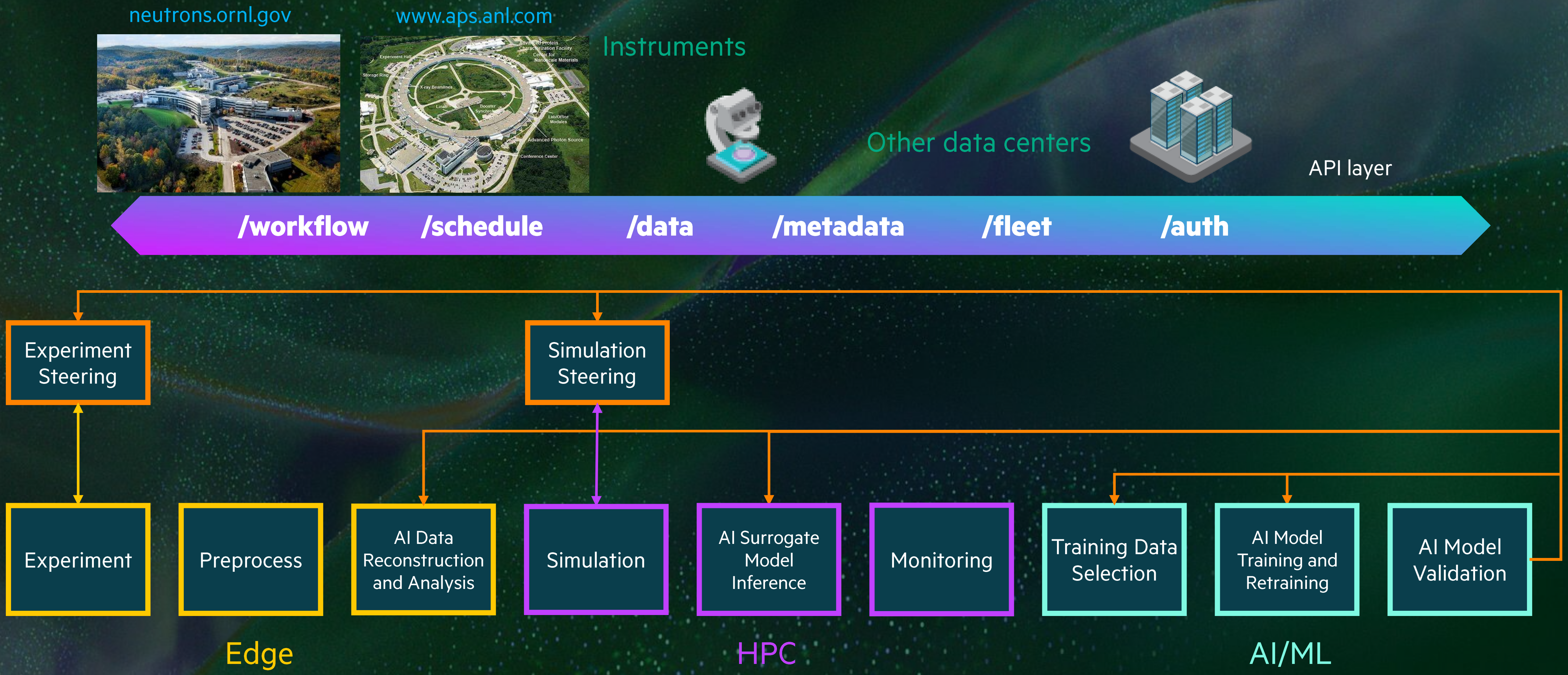
CPUs

Heterogenous computing development

Quantum computing development

Common Federation Framework: Workflow Deployment SDK

Enables Federated Hybrid Workflows on Data from Edge to Extreme Scale to Cloud

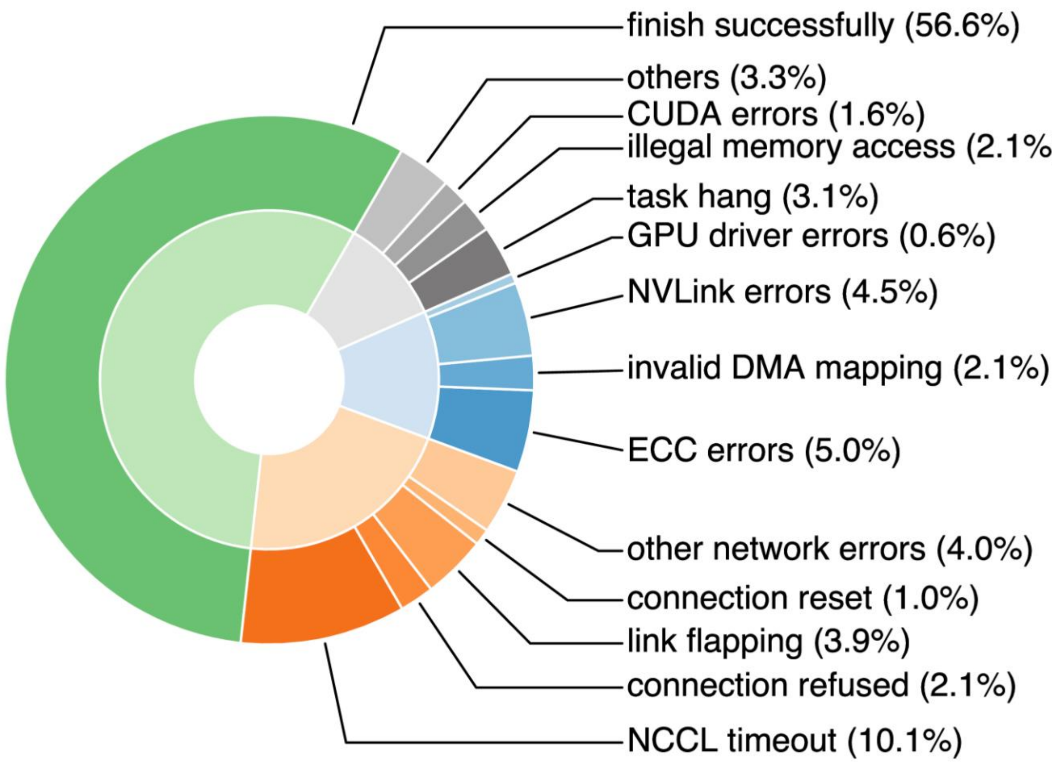


AI Strategy

- HPE delivers the most HPC computing on the top 500
  - HPE sells over 2x the amount of dedicated AI computing as HPC computing
- Driving to make common AI frameworks work out of the box on our systems
  - Working to address networking, compiler, development, etc. issues
- We will leverage our expertise to augment and enhance AI systems
  - Provide tools and capability to scale AI and get it to be reliable
  - Provide frameworks to connect HPC to AI
  - Provide tools to build and deploy federated AI workflows

What Happens at Scale

- As leadership-class AI workloads have grown, concerns about reliability have increased



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%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

https://arxiv.org/pdf/2401.00134

https://arxiv.org/pdf/2407.21783

Coupling AI and HPC

Dragon is a composable distributed runtime that enables users to create sophisticated, scalable, resilient, and high-performance AI/HPC applications, workflows, and services through standard Python interfaces.

Distributed Python and interactive Jupyter

Resilient training and inference

Easy-to-use and HPC-optimized data exchange between applications

Observability for AI/HPC workflows

Multiprocessing | AI | Data | Telemetry

Dragon Core

- 2 – 100X faster data processing than Ray
- Scalable to over 1000 nodes
- Multi-system features offer a hybrid experience, spanning from laptop to supercomputers
- Open-source or HPE-optimized packages
- Well-documented with numerous cookbook examples and easy setup

https://developer.hpe.com/platform/dragonhpc/home/

https://github.com/DragonHPC/dragon

Holistic Power and energy Management (HPM)

Concept: System Administrator and/or User define optimization policy

Minimum resources ⟷ **Maximum performance**

Holistic power and energy management tools

HPE's automatic workload control delivers desired outcome

Possible constraints
Power
Cooling
OPEX
Sustainability

Monitoring
- Data
- Events
- Decisions

- Dynamically balance between available power/cooling, optimized resource usage, and workload performance
- Balance facility efficiency and system operation with minimal performance impact

* Potentially up to 50+% power and TCO savings

A TCO Savings Example

Estimates for 8 theoretical racks (each 200kW IT nameplate power)	No Management	Uniform Static	Strategy 1	Combine (Strategy 1&2)	Combine (Strategy 1&2)
Application Performance	100%	>90%	>99.1%	>95%	>90%
IT compute power (MW)	1.6	1.2	1.2	0.9	0.7
Facility Power procured (MW)	2.3	1.7	1.7	1.2	1.0
OPEX 5 years (Million US)	>=8.4	8.4	8.4	6.0	5.0
CAPEX savings (Million US)	0.0	**4.3**	**4.3**	**7.5**	**9.3**
OPEX savings over 5 years(Million US)	0.0	0.0	0.0	**2.4**	**3.4**
Potential annual OPEX savings (Million US)	0.0	0.0	0.0	**0.5**	**0.7**
Perf/procured Watt efficiency (relative)	**1.00**	**1.23**	**1.35**	**1.79**	**2.14**

Racks

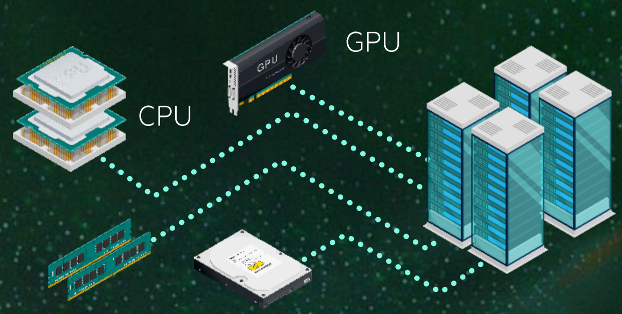
- Publicly vendors have stated chip powers through 1200W
  - https://www.theregister.com/2024/03/18/nvidia_turns_up_the_ai/
  - Likely to increase 2x
  - Keeping current density drives significant rack power and cooling challenges

Cooling efficiency and capacity (kW/rack) increases from left to right

- **Leadership Class Performance**
  - The fastest and most capable HPC/AI solutions are ready for the future, with cutting-edge chip technology, advanced workload software and the latest in high-speed fabric
- **Open Standards**
  - An open rack framework with industry standard OCP motherboards decrease time to market while being adaptable with rapidly changing HPC and emerging AI-focused architectures
- **Revolutionary Cooling**
  - Innovative power management and cooling infrastructure enables customers to match workload needs and sustainability goals with warm facility water

The Future of Sustainable Data Centers

Configuration



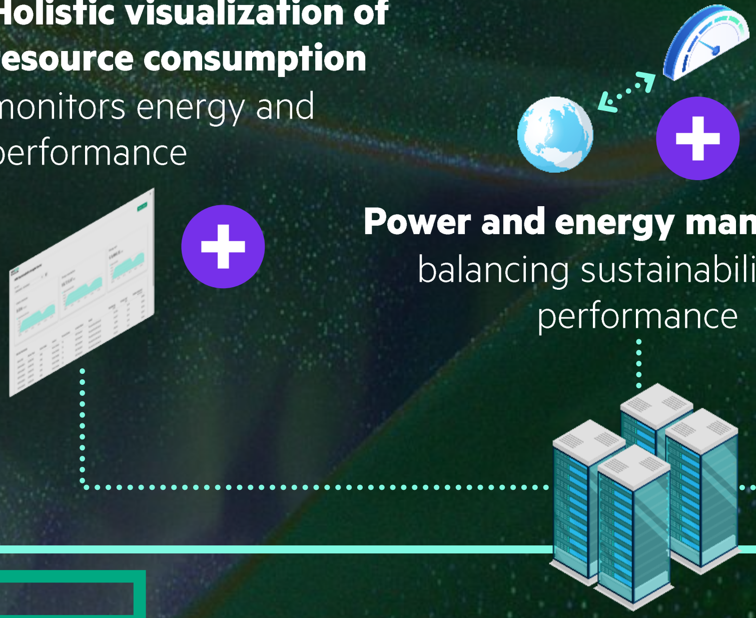
Performance-Energy system configuration tool
Configure hardware (virtual + physical) based on workloads

Operations

Holistic visualization of resource consumption monitors energy and performance

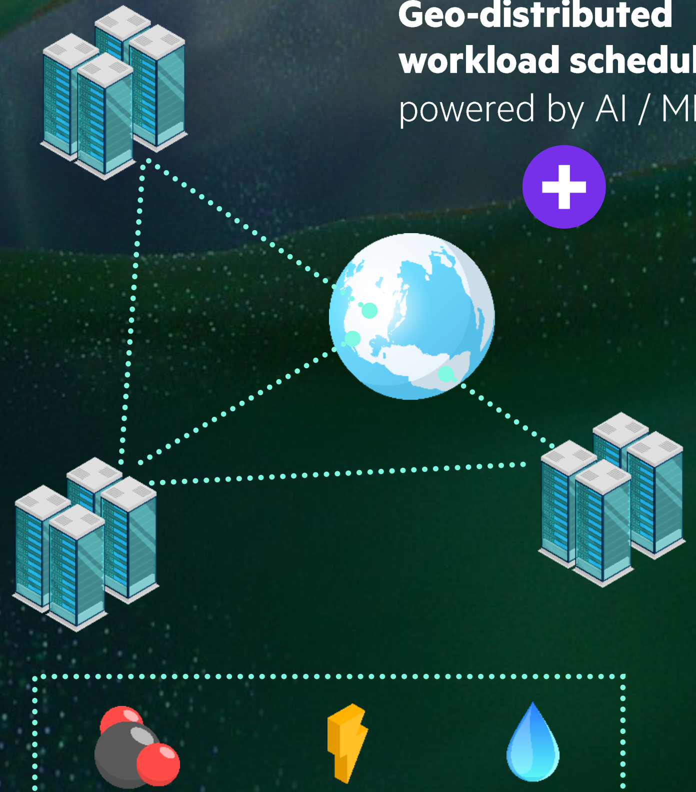
Power and energy management balancing sustainability and performance

Data center digital twin powered by AI/ML views and controls the data center



Scaling

Geo-distributed workload scheduling powered by AI / ML



Carbon
Energy
Water

Thank You