Hewlett Packard Enterprise

Operated by Selectium

Global and regional leader

HPE’s solutions for HPC and AI

Đorđe Ristić, Milan Avramović

May 2025

Providing the brightest minds of humankind – our customers – supercomputing systems that are the **most efficient, most powerful** and **most complete,** so that they can **achieve their mission faster**

National Centers & Academics

Oil & Gas, Energy

Defense

Weather & Climate

Sovereign AI

Hewlett Packard
Enterprise
Operated by Selectium

The World’s Only Three Verified Exascale Supercomputers

All exascale supercomputers are built on HPE Cray Supercomputing EX – 100% Liquid cooled

#1

LLNL’s El Capitan supercomputer is #1 on the TOP500

1.742 exaflops of FP64 performance[1]

10 952 liquid-cooled APUs

#2

ORNL’s Frontier supercomputer is #2 on the TOP500

1.353 exaflops of FP64 performance[1]

10.2 exaflops of HPL-MxP

9 604 liquid-cooled CPUs

38 416 liquid-cooled GPUs

#3

ANL’s Aurora supercomputer is #3 on the TOP500

1.012 exaflops of FP64 performance[1]

10.6 exaflops of HPL-MxP

21 248 liquid-cooled CPUs

63 744 liquid-cooled GPUs

HPE Cray Supercomputing EX broke the exascale barrier for the THIRD time!

- Source:Top500.org, November 2024 list

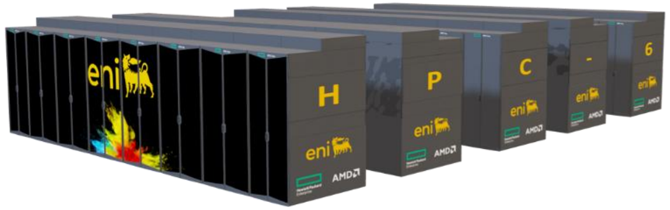
The European fastest Supercomputers

First three supercomputers in Europe are built on HPE Cray Supercomputing EX

#1

ENI's HPC6 supercomputer is #5 on the TOP500

477 petaflops of FP64 performance[1]



3 330 liquid-cooled CPUs

13 320 liquid-cooled GPUs

#2

CSCS's Alps supercomputer is #7 on the TOP500

434 petaflops of FP64 performance[1]

10 400 liquid-cooled CPUs

10 400 liquid-cooled GPUs

#3

CSC's LUMI supercomputer is #8 on the TOP500

379 petaflops of FP64 performance[1]

2.35 exaflops of HPL-MxP

2 919 liquid-cooled CPUs

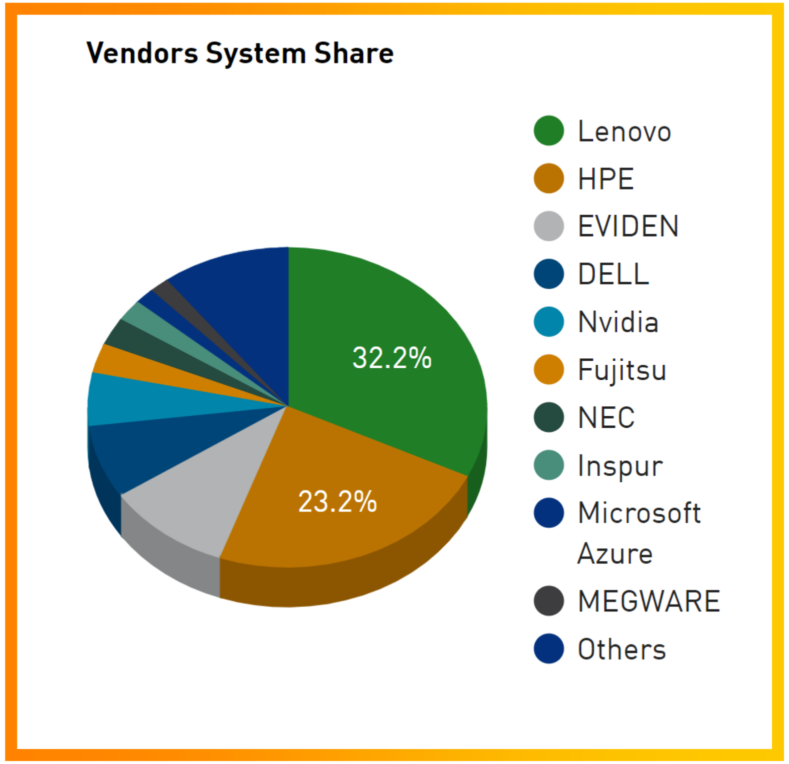
11 664 liquid-cooled GPUs

HPE Cray Supercomputing EX deliver 1.5 exaflops of aggregate performance in Europe

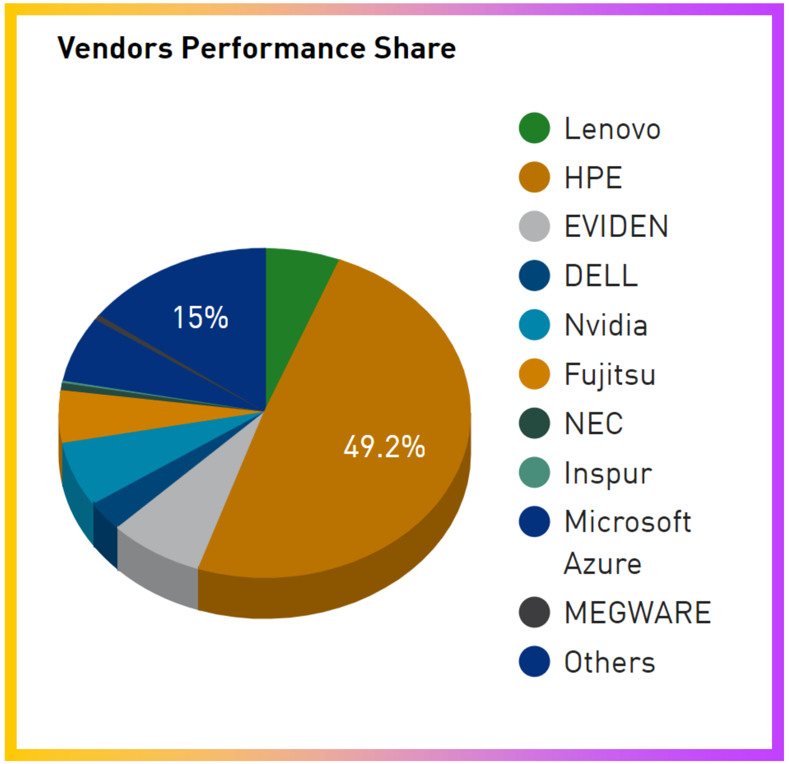
[1] Source:Top500.org, November 2024 list

Global Leader in Supercomputing

Providing our customers with performance they require



Vendors System Share



Vendors Performance Share

HPE Cray Supercomputing EX systems delivers outstanding performance

Source:Top500.org, November 2024 list

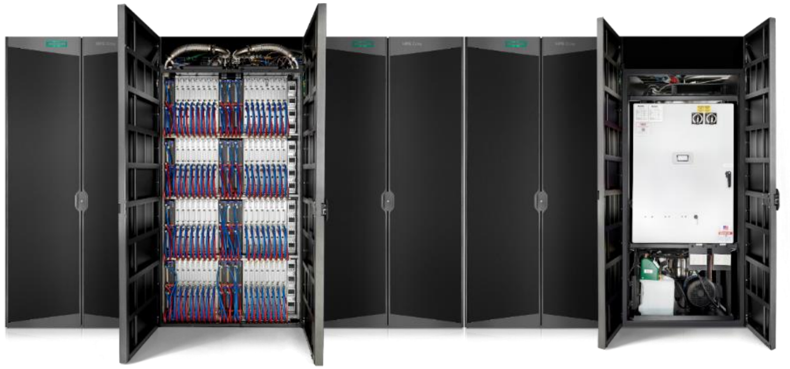
Why not just go to the **Public Cloud** for training, fine-tuning and serving generative AI?

Because supercomputing-type workloads like generative AI that run 24x7 can be **4 to 5 times more expensive** to run in the public cloud than in a private environment.*

Because the public clouds were built on a cloud-native architecture. They are **not built on AI-native** infrastructure that has proven its scalability and cost-effectiveness for many years.

*Source: Earl Joseph, CEO, Hyperion Research, May 2023

Three solutions to cover key use-cases for AI



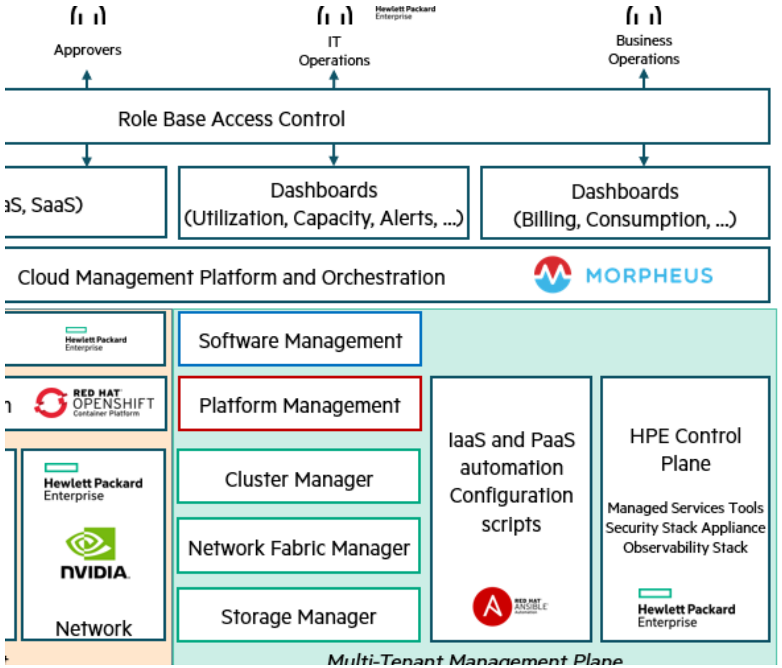
Supercomputing Solution for National Level AI Training

- Highly efficient solution for national AI. Best for LLM training. 1000s of projects.
- The fastest and the most efficient AI infrastructure in the world.

NVIDIA AI Computing by HPE
HPE Private Cloud AI
Start running AI Workloads on your AI Systems. Speed time to value for generative AI with a full-stack AI-native tuning and inference solution purpose built for the enterprise.
Launch HPE Private Cloud AI
Active jobs
36

Private Cloud AI for Enterprise

- Turn-key solution, with full stack of the infrastructure hardware and software for enterprise
- Solution accelerators, full stack of a validated software, RAG & finetuning



Approvers

IT Operations

Business Operations

Role Base Access Control

aS, SaaS)

Dashboards (Utilization, Capacity, Alerts, ...)

Dashboards (Billing, Consumption, ...)

Cloud Management Platform and Orchestration

MORPHEUS

Hewlett Packard Enterprise

RED HAT OPENSHIFT Container Platform

Hewlett Packard Enterprise

NVIDIA

Network

Software Management

Platform Management

Cluster Manager

Network Fabric Manager

Storage Manager

IaaS and PaaS automation Configuration scripts

RED HAT ANSIBLE

HPE Control Plane

Managed Services Tools
Security Stack Appliance
Observability Stack

Hewlett Packard Enterprise

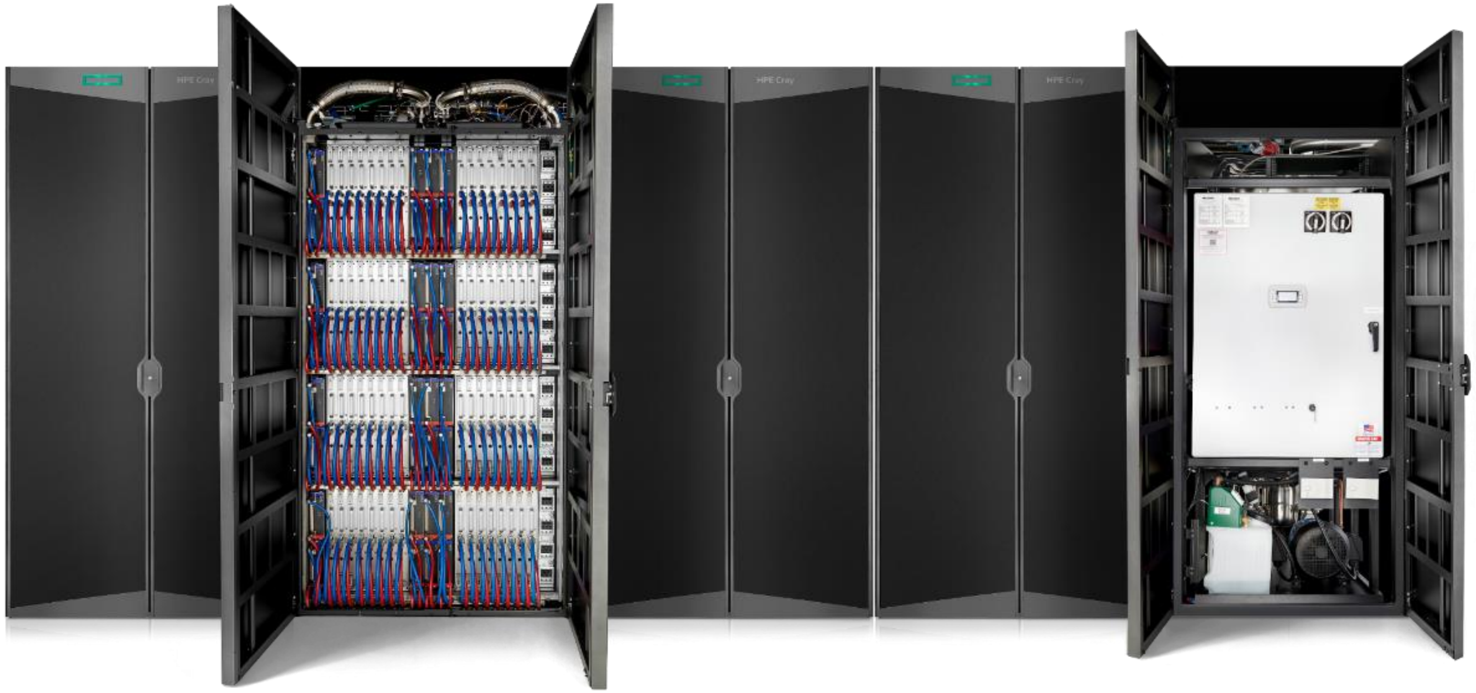
Multi-Tenant Management Plane

AI factory for Cloud Solution Providers

- Cloud Solution Provider class multitenant platform for multi-national level deployments
- Easy integration with multiple clouds, unlimited customization capabilities, groundbreaking security

Supercomputing Solution for National Level AI Training

- Highly efficient solution for national AI. Best for LLM training. 1000s of projects.
- The fastest and the most efficient AI infrastructure in the world.



Private Cloud AI for Enterprise

- Turn-key solution, with full stack of the infrastructure hardware and software for enterprise
- Solution accelerators, full stack of a validated software, RAG & finetuning

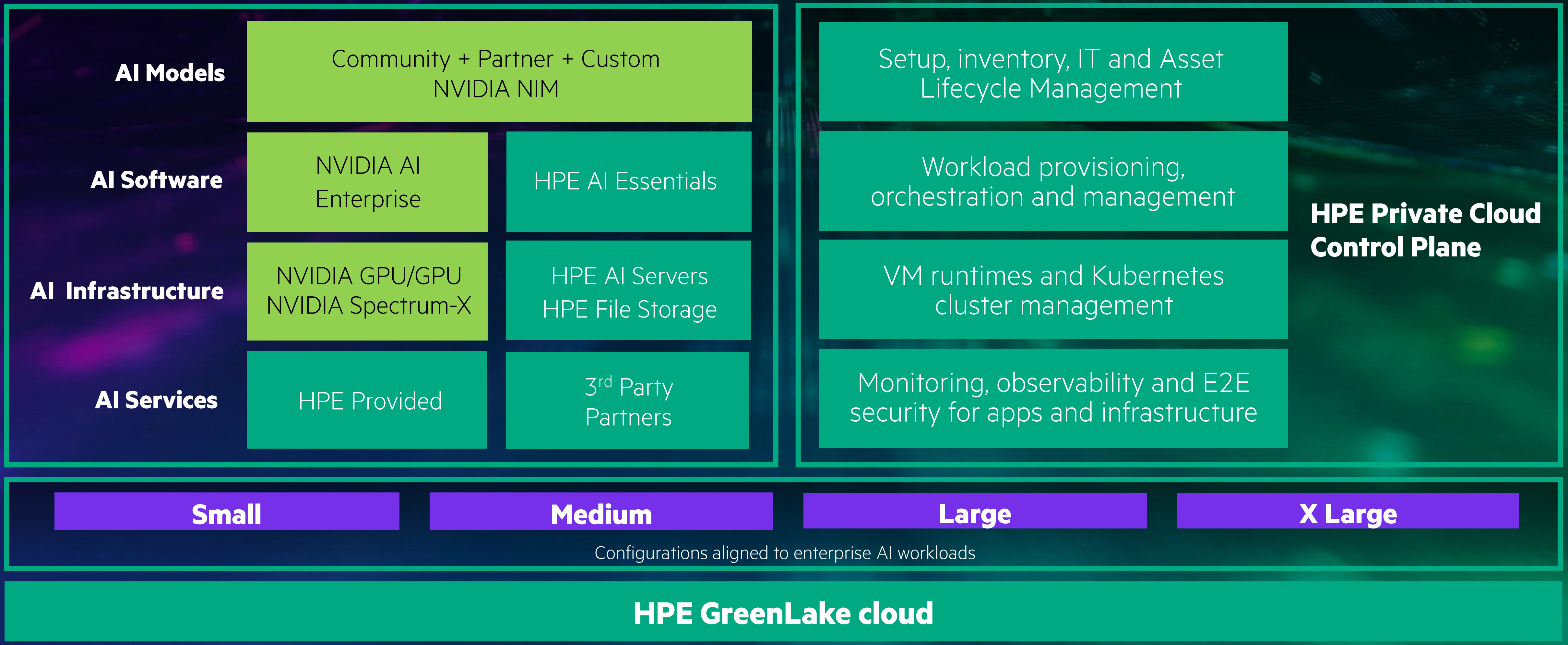
NVIDIA AI Computing by HPE

HPE Private Cloud AI

Start running AI Workloads on your AI Systems. Speed time to value for generative AI with a full-stack AI-native tuning and inference solution purpose built for the enterprise.

NVIDIA AI Computing by HPE

HPE Private Cloud AI



AI Models

Community + Partner + Custom
NVIDIA NIM

AI Software

NVIDIA AI Enterprise

HPE AI Essentials

AI Infrastructure

NVIDIA GPU/GPU
NVIDIA Spectrum-X

HPE AI Servers
HPE File Storage

AI Services

HPE Provided

3rd Party Partners

Setup, inventory, IT and Asset Lifecycle Management

Workload provisioning, orchestration and management

VM runtimes and Kubernetes cluster management

Monitoring, observability and E2E security for apps and infrastructure

HPE Private Cloud Control Plane

Small | **Medium** | **Large** | **X Large**

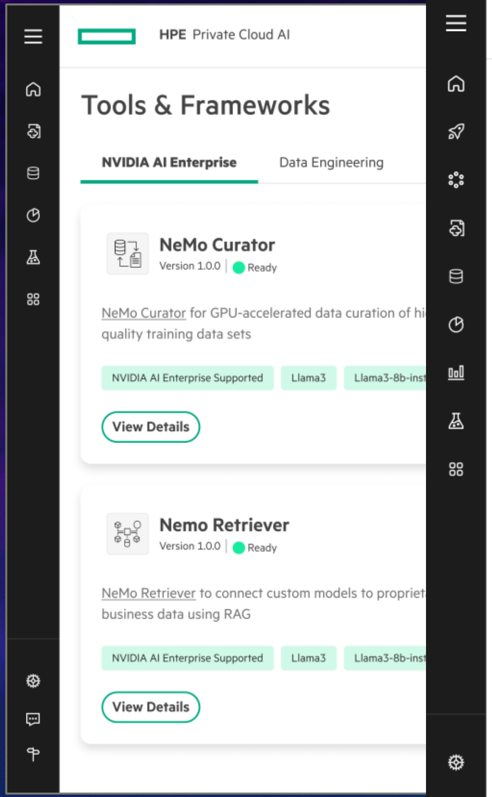
Configurations aligned to enterprise AI workloads

HPE GreenLake cloud

Hewlett Packard Enterprise | NVIDIA

Curated and supported set of AI tools

End-to-end software platform



HPE Private Cloud AI

Tools & Frameworks

NVIDIA AI Enterprise | Data Engineering

NeMo Curator
Version 1.0.0 | Ready

NeMo Curator for GPU-accelerated data curation of hi
quality training data sets

NVIDIA AI Enterprise Supported · Llama3 · Llama3-8b-inst

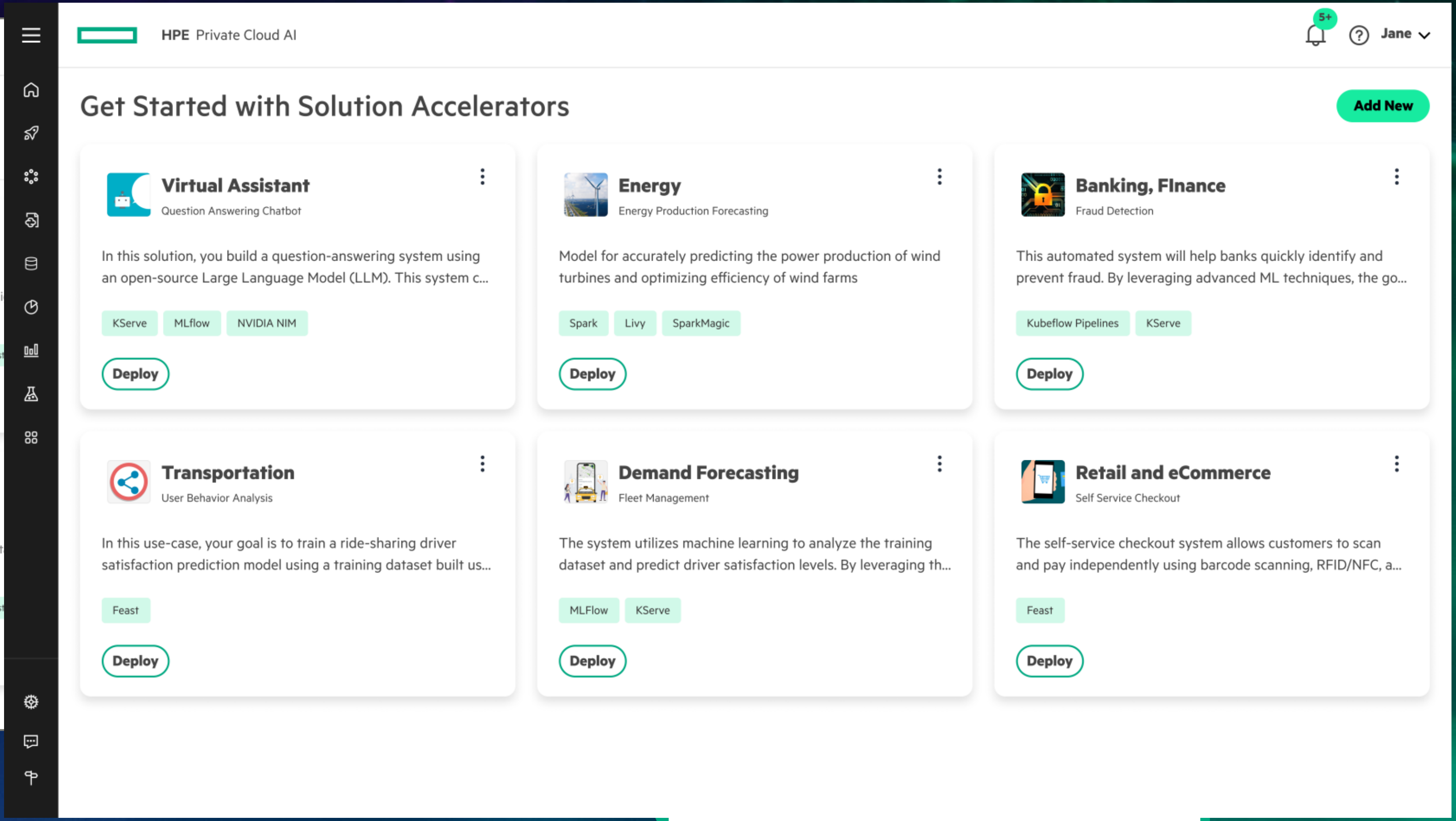
View Details

Nemo Retriever
Version 1.0.0 | Ready

NeMo Retriever to connect custom models to propriet
business data using RAG

NVIDIA AI Enterprise Supported · Llama3 · Llama3-8b-inst

View Details



HPE Private Cloud AI

Get Started with Solution Accelerators

Add New

Virtual Assistant
Question Answering Chatbot

In this solution, you build a question-answering system using an open-source Large Language Model (LLM). This system c...

KServe · MLflow · NVIDIA NIM

Deploy

Energy
Energy Production Forecasting

Model for accurately predicting the power production of wind turbines and optimizing efficiency of wind farms

Spark · Livy · SparkMagic

Deploy

Banking, FInance
Fraud Detection

This automated system will help banks quickly identify and prevent fraud. By leveraging advanced ML techniques, the go...

Kubeflow Pipelines · KServe

Deploy

Transportation
User Behavior Analysis

In this use-case, your goal is to train a ride-sharing driver satisfaction prediction model using a training dataset built us...

Feast

Deploy

Demand Forecasting
Fleet Management

The system utilizes machine learning to analyze the training dataset and predict driver satisfaction levels. By leveraging th...

MLFlow · KServe

Deploy

Retail and eCommerce
Self Service Checkout

The self-service checkout system allows customers to scan and pay independently using barcode scanning, RFID/NFC, a...

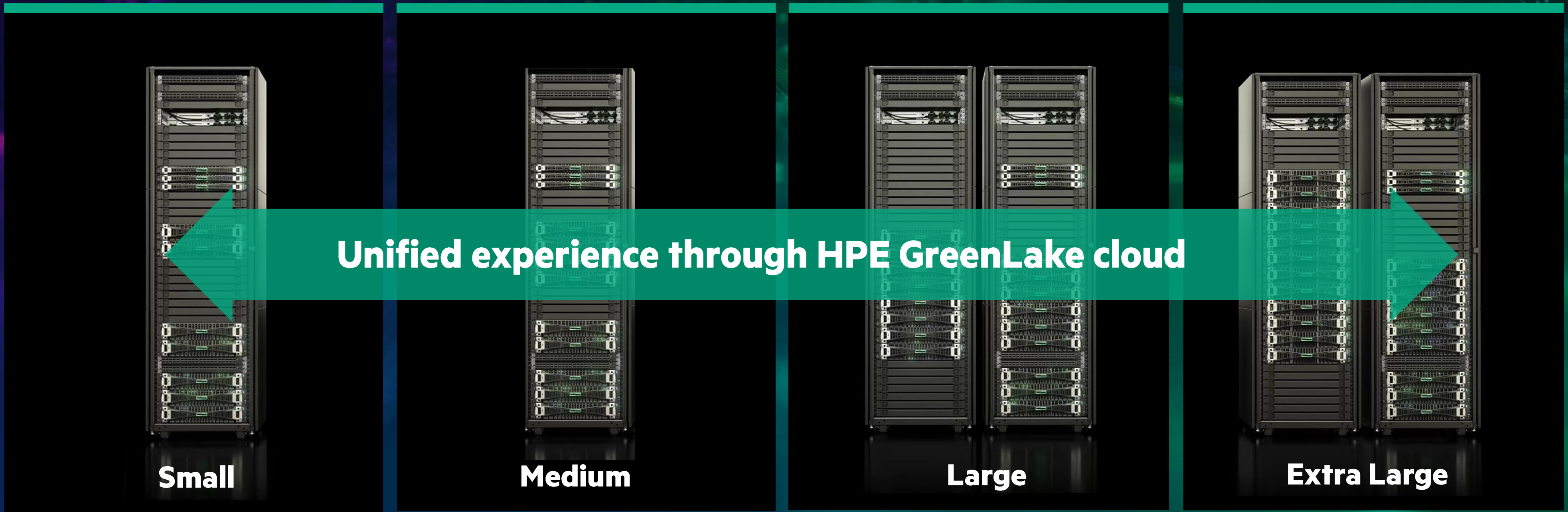
Feast

Deploy

Hewlett Packard Enterprise | NVIDIA

AI optimized and scalable hardware

For your Inferencing, RAG, and Fine-tuning workloads



Unified experience through HPE GreenLake cloud

	Small	Medium	Large	Extra Large
Compute	4 or 8 NVIDIA L40S GPUs	8 or 16 NVIDIA L40S GPUs	16 or 32 NVIDIA H100 NVL GPUs	16 or 32 NVIDIA GH200 NVL2
***Storage**	109 TB to 500TB	217 TB to 500 TB	670 TB to 1088 TB	670 TB to 1088 TB
Networking	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking	800GbE NVIDIA Networking
Power	up to 8 kW rack	up to 17.7 kW	up to 16.5 kW x 2	up to 16.5 kW x 2

* Storage Includes 10+ TB Data Fabric Lakehouse capacity in addition to File storage.
Upgraded configurations shown

Hewlett Packard
Enterprise

Fueling the future with improved seismic imaging

ExxonMobil improves decision-making and doubles its chances of discovering oil and gas with advanced imaging technology

Hewlett Packard Enterprise | AMD

Sustainable cultivation that's good for biodiversity and the economy

University of Zagreb researchers leverage supercomputing to speed up crop gene analysis by more than 30x

Multi-purpose HPC cluster

Institute of Physics Belgrade hosts a supercomputer for University-wide projects

Dozens of projects for researchers from all over the country and region

Regional pioneer, in HPC since 1997, moving to HPE in 2013, expanded in 2024

With assistance from local HPE team

IPB HPC clusters from HPE

2013

105 nodes
Compute HPE SL250
Management HPE DL380 Gen8
Infiniband interconnect
Lustre cluster-wide filesystem
Storage HPE MSA 2000 G3
IO performance ~100MB/s per client
Room neutral with HPE ARCS

2024

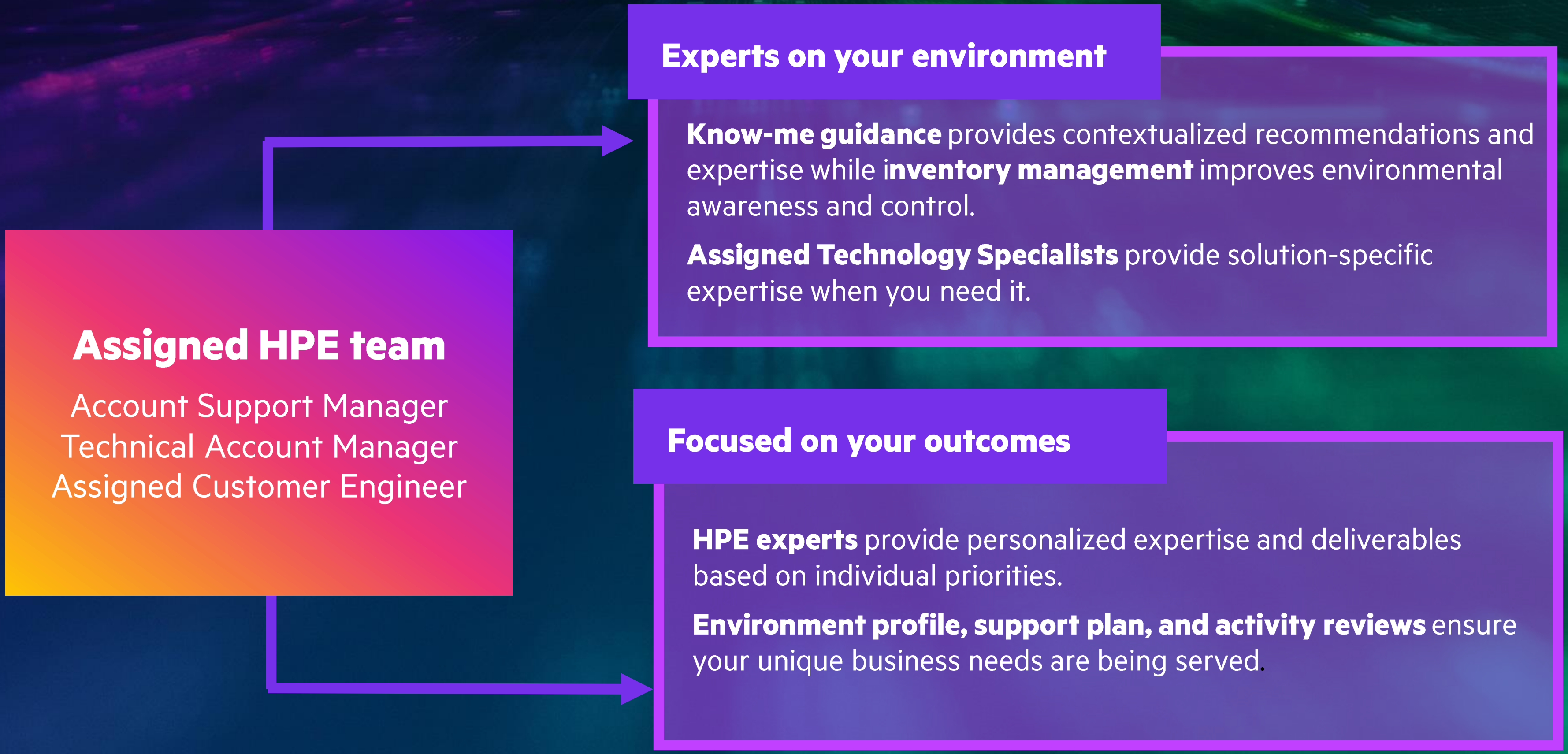
24 nodes
Compute HPE XL220
Management HPE DL380 Gen10
Infiniband EDR interconnect
Lustre cluster-wide filesystem
Storage HPE MSA 2060
IO performance ~200MB/s per client
Room neutral with HPE ARCS

Future

Considering merging clusters
Considering expansion
Software updates for interoperability
Lustre PFS updates

Including a worry-free operational experience

Maintain the value of your investment with HPE Tech Care and Complete Care Services



Assigned HPE team

Account Support Manager
Technical Account Manager
Assigned Customer Engineer

Experts on your environment

Know-me guidance provides contextualized recommendations and expertise while i**nventory management** improves environmental awareness and control.

Assigned Technology Specialists provide solution-specific expertise when you need it.

Focused on your outcomes

HPE experts provide personalized expertise and deliverables based on individual priorities.

Environment profile, support plan, and activity reviews ensure your unique business needs are being served.

[1] NVIDIA Grace Hopper Superchip Architecture Whitepaper

HPC / AI Products Portfolio 2025

HPC and AI systems portfolio

Leadership-class supercomputing

100% fanless direct liquid cooling

The next frontier of supercomputing systems redesigned for HPC, AI, and converged workloads

HPE Cray Supercomputing EX4000

HPE Cray Supercomputing EX2500

HPE Slingshot combines the performance of a supercomputing interconnect with the cost-effectiveness of Ethernet

Accelerated AI

70% direct liquid cooling, liquid to air cooling

Purpose-built, 8-way AI servers for AI model training, tuning and inference

HPE ProLiant Compute XD680

HPE ProLiant Compute XD685

HPE Cray XD670

HPE Cray XD675

HPE ProLiant Compute accelerating AI applications for Enterprises

HPE ProLiant Compute DL380a Gen12

HPE ProLiant Compute DL384 Gen12

Mainstream HPC/AI

Density-optimized, scale-out compute for HPC and AI workloads

HPE Cray XD2000

HPE Cray XD665

Integrated HPC and AI software portfolio, including application and software development ecosystem, system management suite, orchestration tools, enhanced compute environment & more

HPE Services Experts available globally to accelerate your strategic HPC and AI initiatives

Purpose-built storage

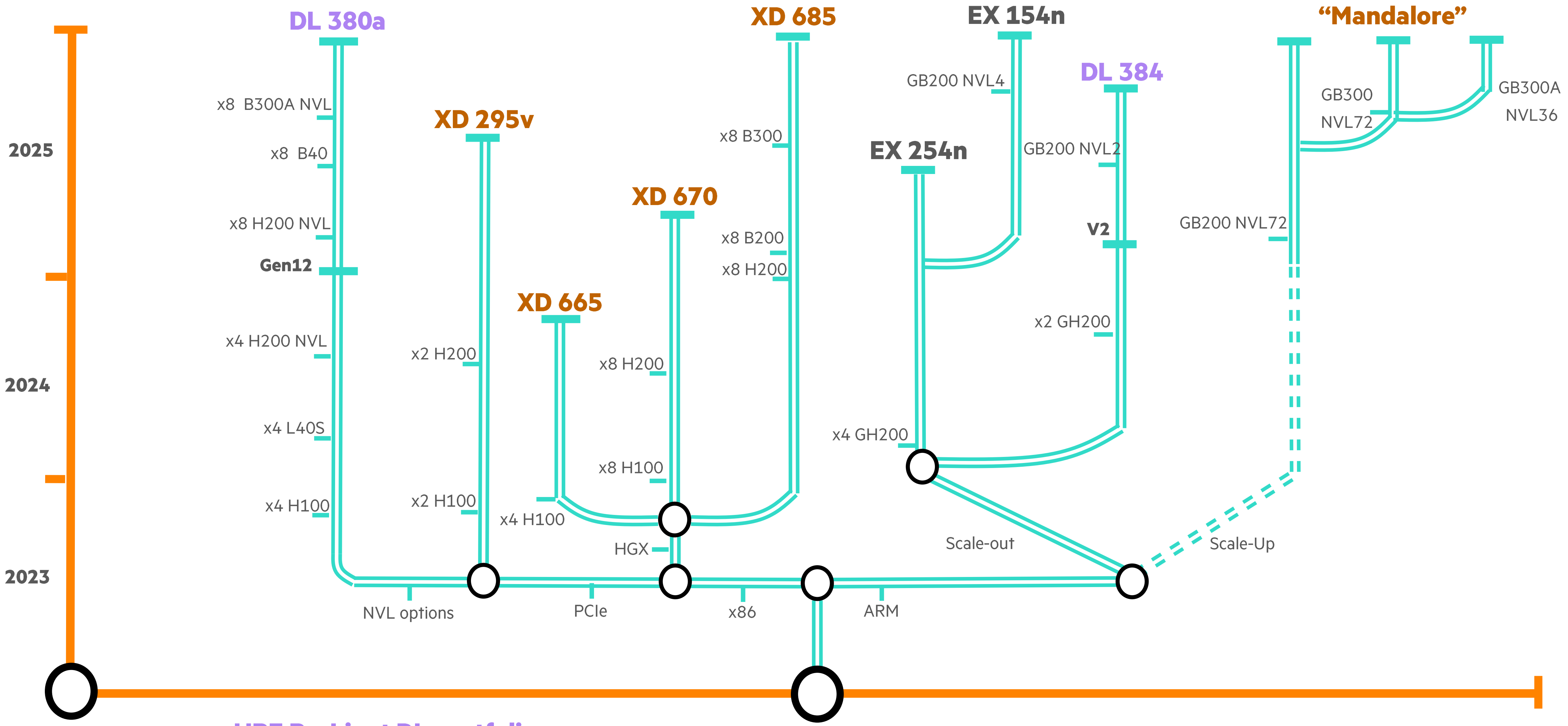
Unprecedented data storage price/performance for HPC, AI, and converged workloads

HPE Cray Supercomputing Storage Systems E2000

Cray ClusterStor E1000 Storage Systems

HPE Cray Storage Systems C500

HPE – Nvidia GPU Servers - Subway MAP



HPE ProLiant DL portfolio

HPE ProLiant/Cray XD portfolio

HPE Cray EX portfolio

Hewlett Packard Enterprise
Operated by Selectium

HPE ProLiant Compute XD

Optimized for natural language processing, large language model training and multimodal training

New HPE-designed modular 5U chassis

- Faster go-to-market
- Choice of GPUs, CPUs, boards and cooling methods
- Optimized pod deployment leveraging a compact 8-nodes-per-rack arrangement

DLC Capable

- Energy-efficiency, enablement of >1KW GPUs. Optional HPE ARCS for room-neutrality

Adaptable GPUs

- Nvidia B200 (Air | DLC)
- AMD MI300x (Air | DLC)
- AMD MI325x (Air | DLC)
- NVIDIA H200 (Air | DLC)
- Intel Gaudi 3 (Air | DLC)
- Future GPUs

HPE ProLiant Compute XD680

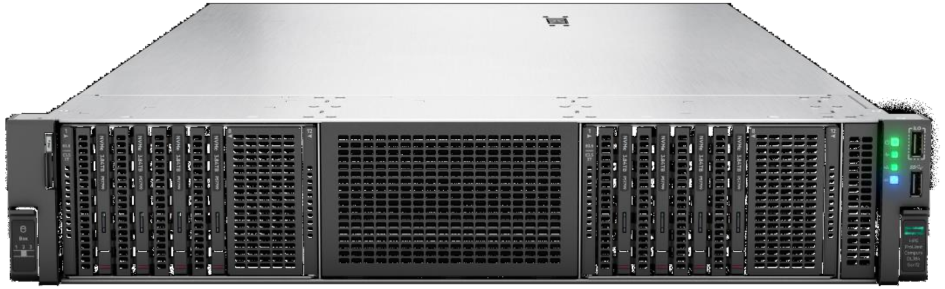
HPE ProLiant Compute XD685

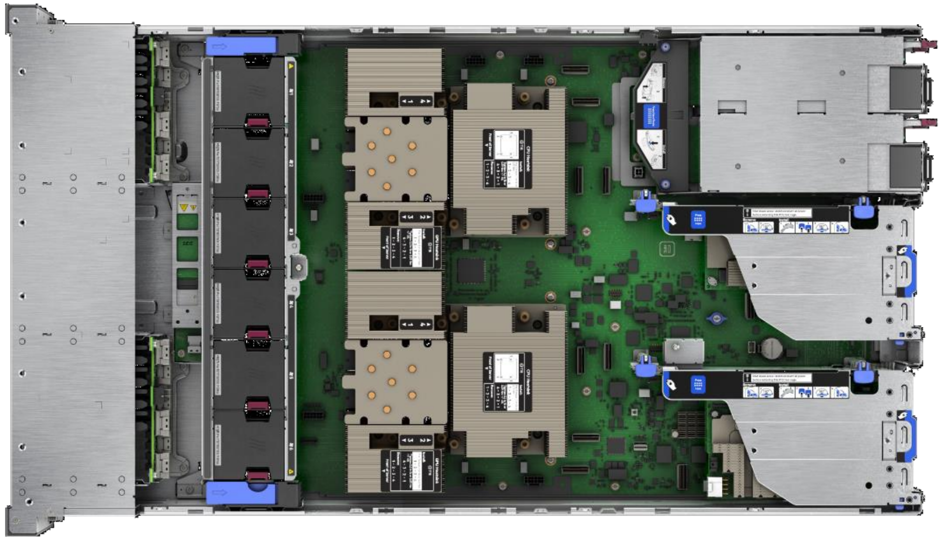
Hewlett Packard Enterprise
Operated by Selectium

Grace Hopper in the HPE ProLiant Compute DL384 Gen12

- Cutting-edge NVIDIA GH200 Superchip
  - NVIDIA GH200 NVL2 for up to 1248GB coherent memory for the largest models
  - Enterprise high-end AI Inference workloads including generative AI models
  - AI Training and HPC

- Enterprise Features with the Latest Air-cooled Superchip:
  - Maximum inferences/watt for generative AI inference
  - Maximum performance/GPU and performance/$ in our air-cooled portfolio
  - Enterprise management (iLO-based) for enterprise AI
  - Broad operating system, options, and services available globally





Rack-level comparison between parallel and NFS storage

For a job with 5 hours run time checkpointing every 15 minutes 3 TB of checkpoint data

Every 15 minutes your compute nodes wait for

4 seconds

for the checkpoint files to be written

Every 15 minutes your compute nodes wait for

One minute and 7 seconds

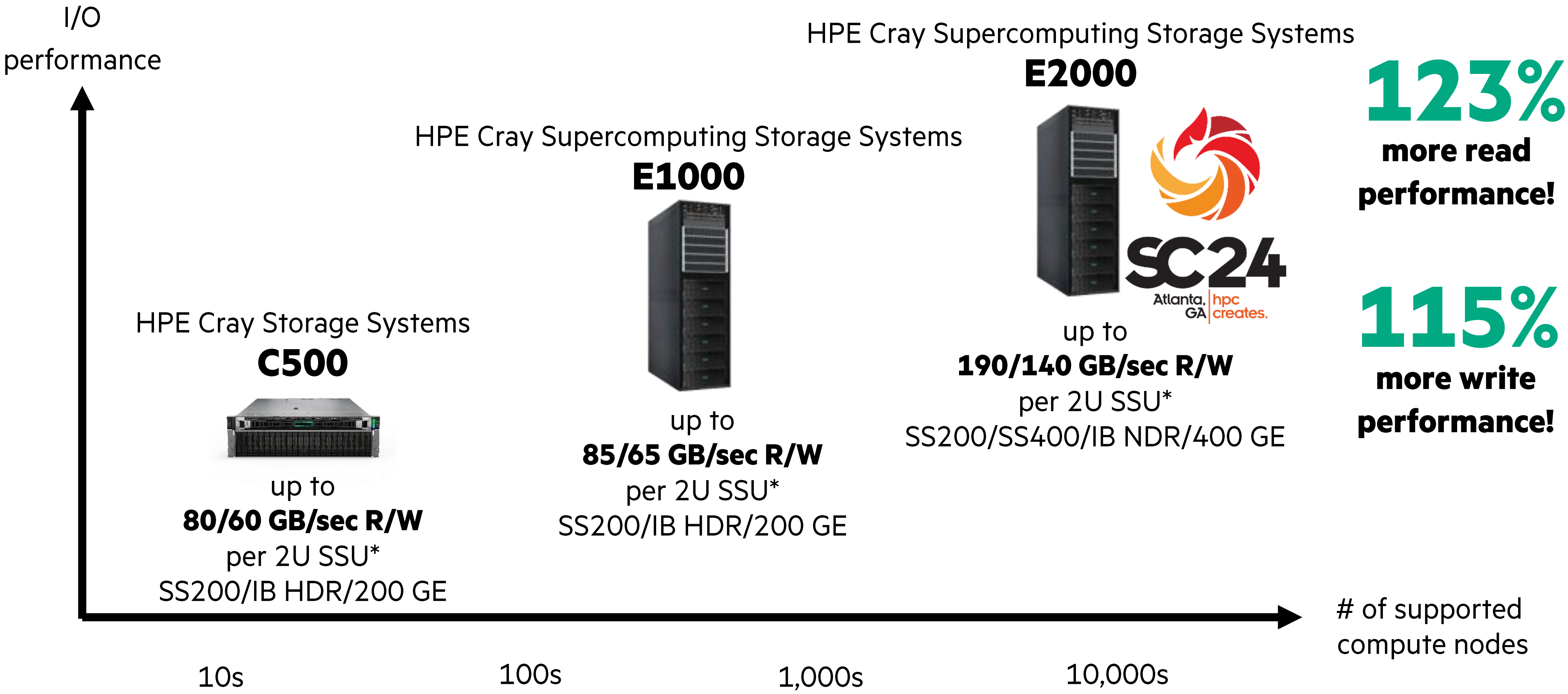
for the checkpoint files to be written

Rack of Cray ClusterStor E1000 with
960 GB/sec write throughput

Rack of NFS storage with
45 GB/sec write throughput

HPC Storage portfolio

Storage solutions for any performance requirements and compute cluster sizes



I/O performance

HPE Cray Storage Systems

C500

up to

80/60 GB/sec R/W

per 2U SSU*

SS200/IB HDR/200 GE

HPE Cray Supercomputing Storage Systems

E1000

up to

85/65 GB/sec R/W

per 2U SSU*

SS200/IB HDR/200 GE

HPE Cray Supercomputing Storage Systems

E2000

SC24 Atlanta, GA | hpc creates.

up to

190/140 GB/sec R/W

per 2U SSU*

SS200/SS400/IB NDR/400 GE

# of supported compute nodes

10s

100s

1,000s

10,000s

123%

more read performance!

115%

more write performance!

*Scalable Storage Unit

Thank you!

d.ristic@selectium.com
m.avramovic@selectium.com