



# From Exascale to AI Factories

Bill Mannel, Sales Director, SLED and CTO Office, North America

September 3, 2025

# The world's three fastest, verified supercomputers



ranked  
**SUPERCOMPUTER**  
in the world.  
at 1.742 exaflops.



ranked  
**SUPERCOMPUTER**  
in the world.  
at 1.353 exaflops.



ranked  
**SUPERCOMPUTER**  
in the world.  
at 1.012 exaflops.

HPE delivers leadership-class

# SUPERCOMPUTING

innovation



# **1**

# **2**

# **3**

**60%**

**63%**

HPE has built the  
**THREE FASTEST VERIFIED SUPERCOMPUTERS**  
in the world

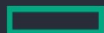
HPE has built  
**60% of the TOP10**  
performing  
supercomputers in  
the world

HPE has built  
**63% of the TOP30**  
most energy  
efficient  
supercomputers  
in the world

Sources: June 2025 Top500



# HPE powers 6 of the world's Top 10 supercomputers



Sources: June 2025 Top500

# Enabling large-scaling AI workloads around the globe



**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



**BriCS**



**21 EFLOPS**

single-precision AI performance  
with NVIDIA GH200 superchips

# Infrastructure DOES Matter

In a new [research paper](#), scientists from the famed Oak Ridge National Laboratory trained a one trillion parameter model using just a few thousand GPUs in their Frontier supercomputer, the most powerful non-distributed supercomputer in the world and one of only two exascale systems globally.

They used just 3,072 GPUs to train the giant large language model out of 37,888 AMD GPUs housed in Frontier. That means the researchers trained a model comparable to ChatGPT's rumored size of a trillion parameters on just 8% of Frontier's computing power.

*AI Business, January 8, 2024, Ben Wojecki, Jr.*



# Exascale can solve the world's most challenging problems

- Predicting extreme weather events
- Discovering life-saving precision medicine
- Clean energy futures
- Accelerating Oil & Gas exploration
- High-fidelity modeling for product development



# HPE Platform Solutions

Servers for every ambition



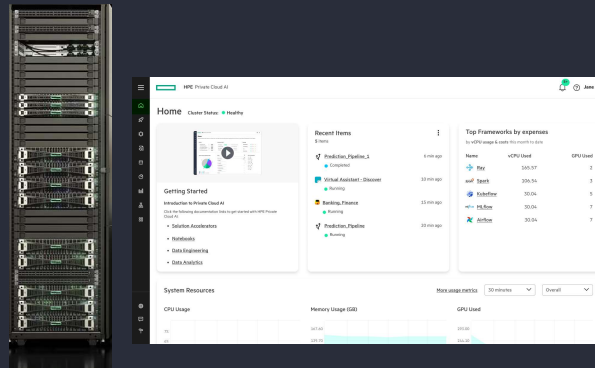
# Journey to HPC/AI

## Enterprise Grade AI



AI Optimized  
Compute and Storage

## Engineered Systems



Turnkey Private Cloud AI Solutions

## Model Building : Research : Service Providers



HPC and Super Computing

## AI Services

## AI Software Platform

## Data Management

## Hybrid Cloud Platform



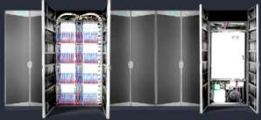
# HPC and AI systems portfolio

## 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

### Hybrid air and direct liquid cooling, liquid to air cooling

Rack-scale solution for AI models > 1 trillion parameters



#### NVIDIA GB200 NVL 72 by HPE



#### NVIDIA GB300 NVL 72 by HPE

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

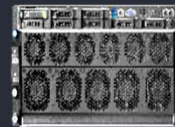
#### HPE Cray XD670



#### HPE ProLiant XD685



#### HPE ProLiant XD690



HPE ProLiant Compute accelerating AI applications for Enterprises

#### HPE ProLiant Compute DL380a Gen12



#### HPE ProLiant Compute DL384 Gen12



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

## Mainstream HPC

Density-optimized, scale-out compute for HPC

#### HPE ProLiant Compute XD230



#### HPE Cray XD2000



## 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 AI Factory

GPU accelerated HPC platforms designed to perform



# AI factories provide a blueprint for AI success

From pilots to profits and purpose-driven outcomes

## AI today

High failure rate

Scattered across silos and teams

No clear plan for operational scale

Fragmented operating model

## AI factory



Measurable and predictable



Centralized workflows and reusable assets



Speed and predictability of deployment



Unified environment

## Outcomes

Faster time to AI ROI

Repeatable AI success

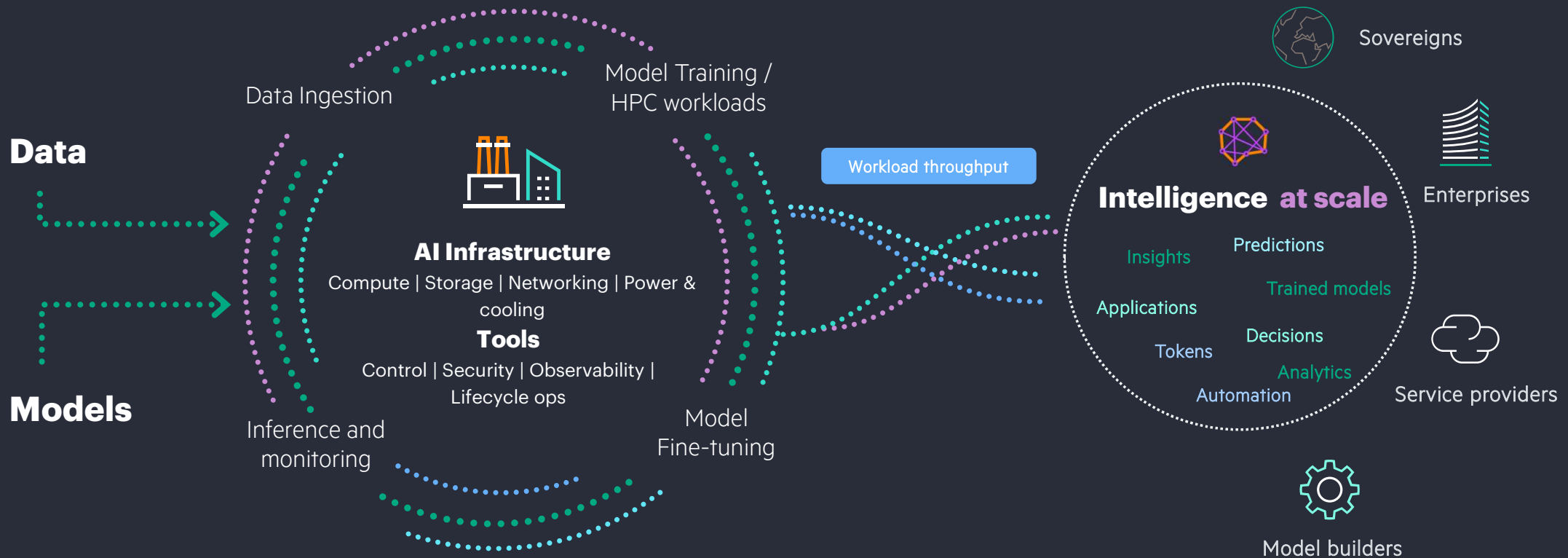
AI innovation at scale

Secure governance & control



# AI factories shaping the future of AI

Industrializing intelligence at unprecedented scale



The **New Engine** of Innovation and Transformation



# Unlock autonomy with your AI factory

Own your data, intelligence, and outcomes



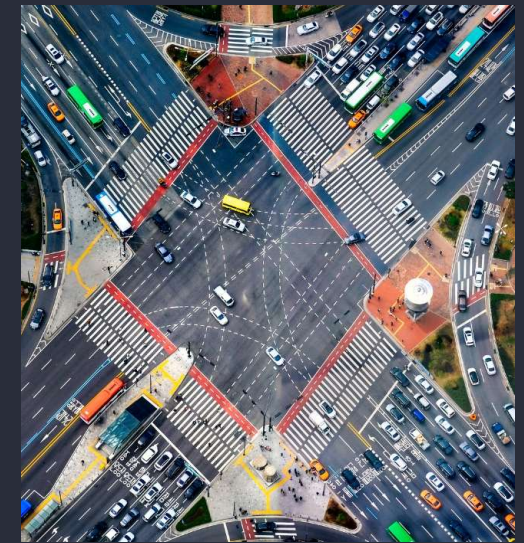
**Data sovereignty  
and security**



**Governance and  
Compliance**



**Flexibility and  
Cost control**

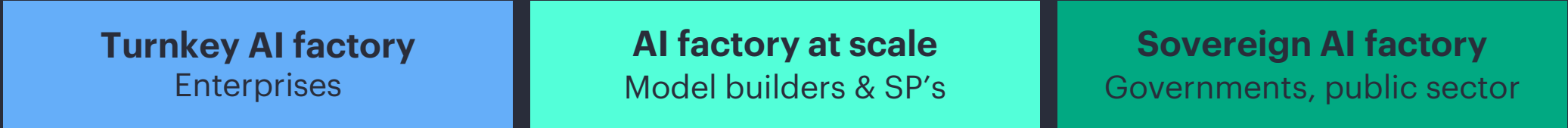


**Performance  
'at scale'**

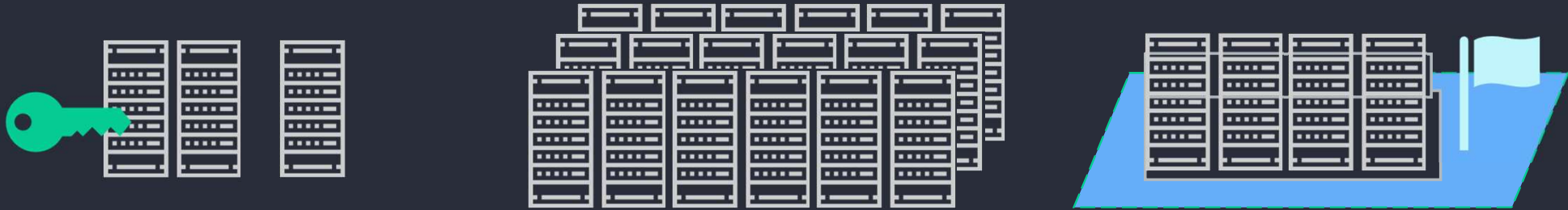


# Expanded AI factory portfolio from HPE

for every AI ambition, across clouds, cores and countries



Common control plane: HPE Morpheus and HPE OpsRamp



Turnkey, engineered systems

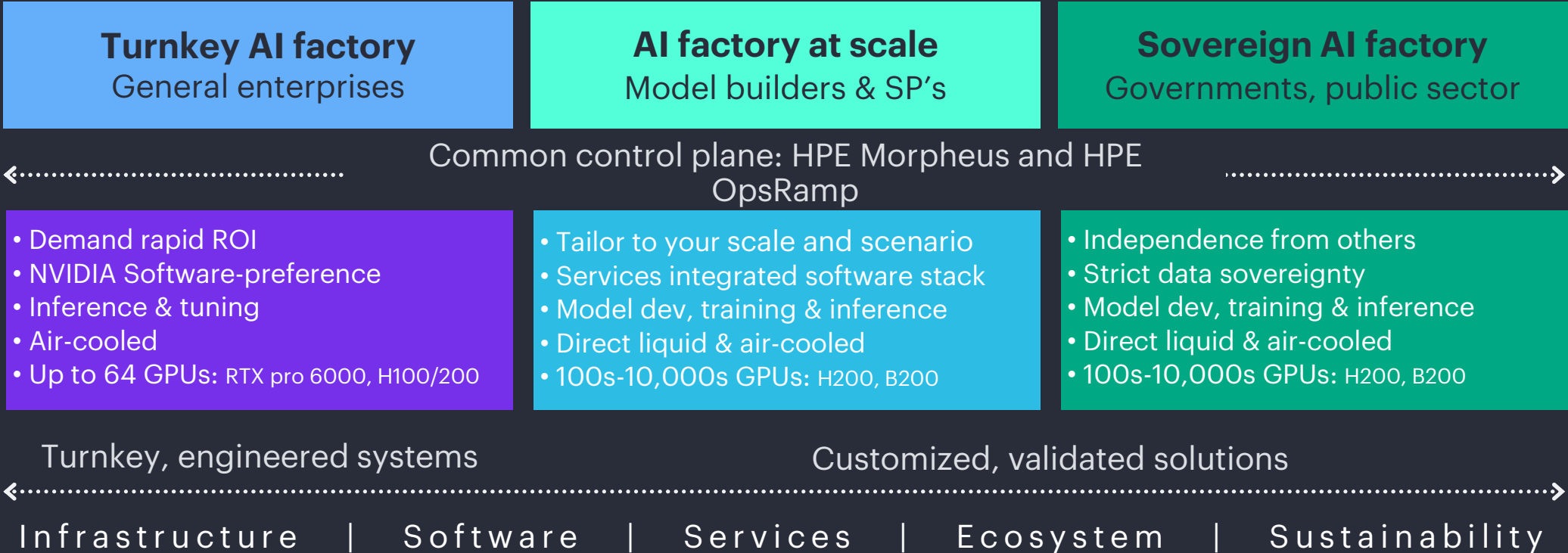
Customized, validated solutions

Infrastructure | Software | Services | Ecosystem | Sustainability



# Expanded AI factory portfolio from HPE

for every AI ambition, across clouds, cores and countries



# HPE Private Cloud AI

Turnkey AI factory

AI factory at scale

Sovereign AI factory

A full-stack, turnkey private cloud for production AI

## Ready to run out of the box

Full stack, HPE servers, NVIDIA accelerated computing, AI software suite, network, and storage

## Scale as you grow

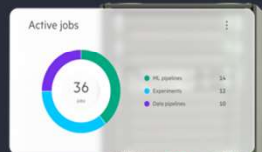
Available in modular configurations sized for inference, RAG-based applications and and fine tuning

## Data privacy and control

Eliminate data silos with one global namespace for seamless access to different data types, anywhere

## AI lifecycle management

Cloud control plane and observability and access to the latest AI models, and development tools



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

NVIDIA NeMo Curator

Open AI

NVIDIA NeMo Customizer

Open AI

Virtual Assistant

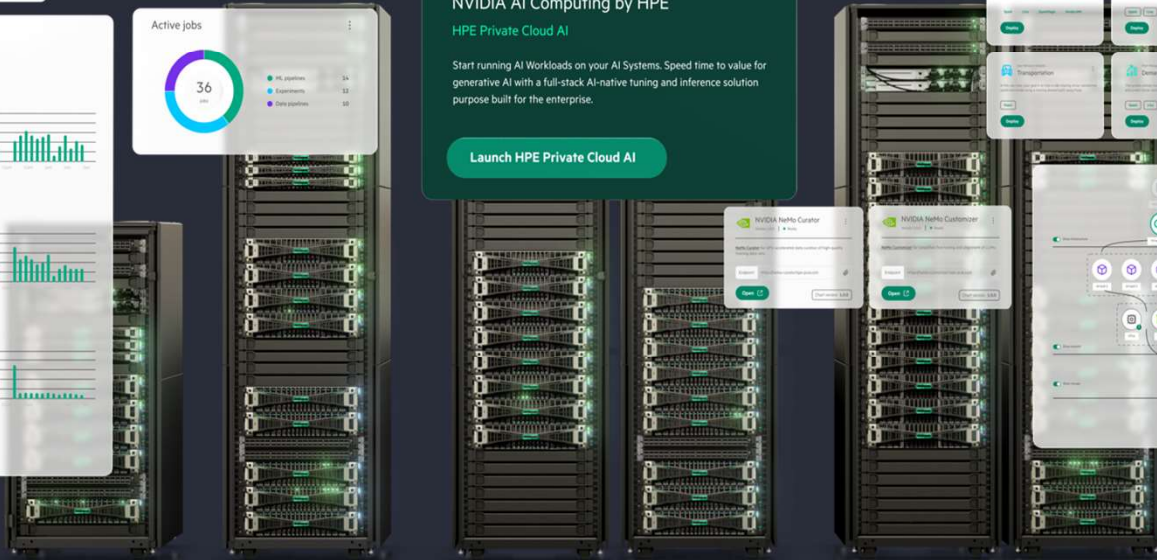
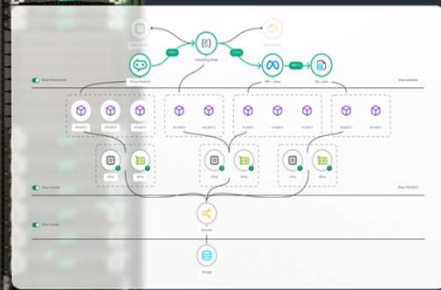
Energy

Banking Finance

Transportation

Demand Forecasting

Retail and Commerce



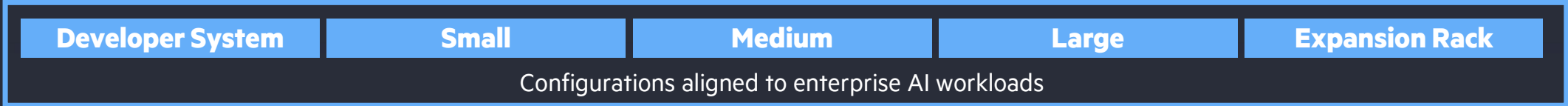
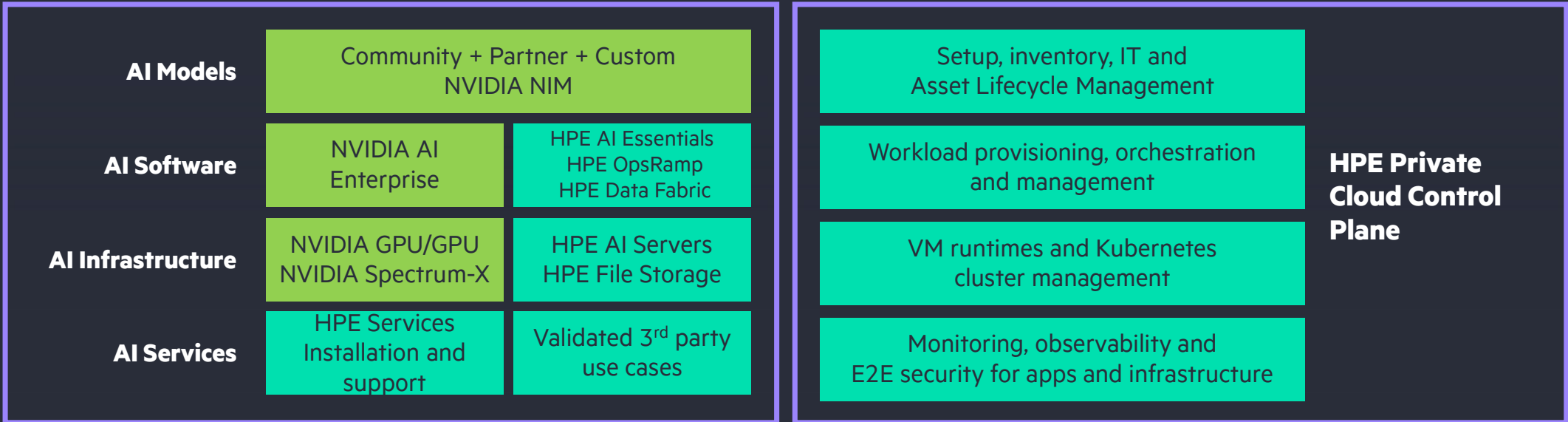
# Turnkey AI factory

HPE Private Cloud AI

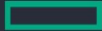
Turnkey AI factory

AI factory at scale

Sovereign AI factory



**HPE GreenLake cloud**



# HPE Private Cloud AI

Enterprise AI infrastructure simplified

Turnkey AI factory

AI factory at scale

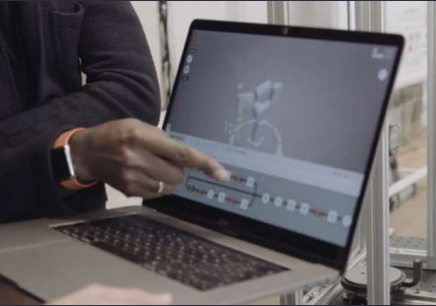
Sovereign AI factory

Instant AI productivity

Secure and unified data access

End-to-end AI software platform

Use cases



Development



Computer vision



Agentic AI



Physical AI



RAG



# Gen AI-optimized and scalable hardware

Turnkey AI factory

AI factory at scale

Sovereign AI factory

Start fast

Scale fast



Developer system



Small



Medium



Large



Expansion rack

2 NVIDIA H100 NVL GPU's  
32 TB Integrated  
Customer Network  
Up to 2.2 kW

8 NVIDIA RTX Pro 6000  
GPU's  
109 TB  
400GbE NVIDIA Networking  
up to 12 kW rack

8 NVIDIA H200 GPUs  
109 TB  
400GbE NVIDIA Networking  
up to 13 kW

16 NVIDIA H200 GPUs  
217 TB  
400GbE NVIDIA Networking  
up to 17.4 kW

16 NVIDIA H200 GPUs  
400GbE NVIDIA Networking  
Up to 12 kw

←..... HPE GreenLake .....→

# AI Factory at Scale Characteristics

Turnkey AI factory

AI factory at scale

Sovereign AI  
factory

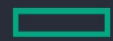
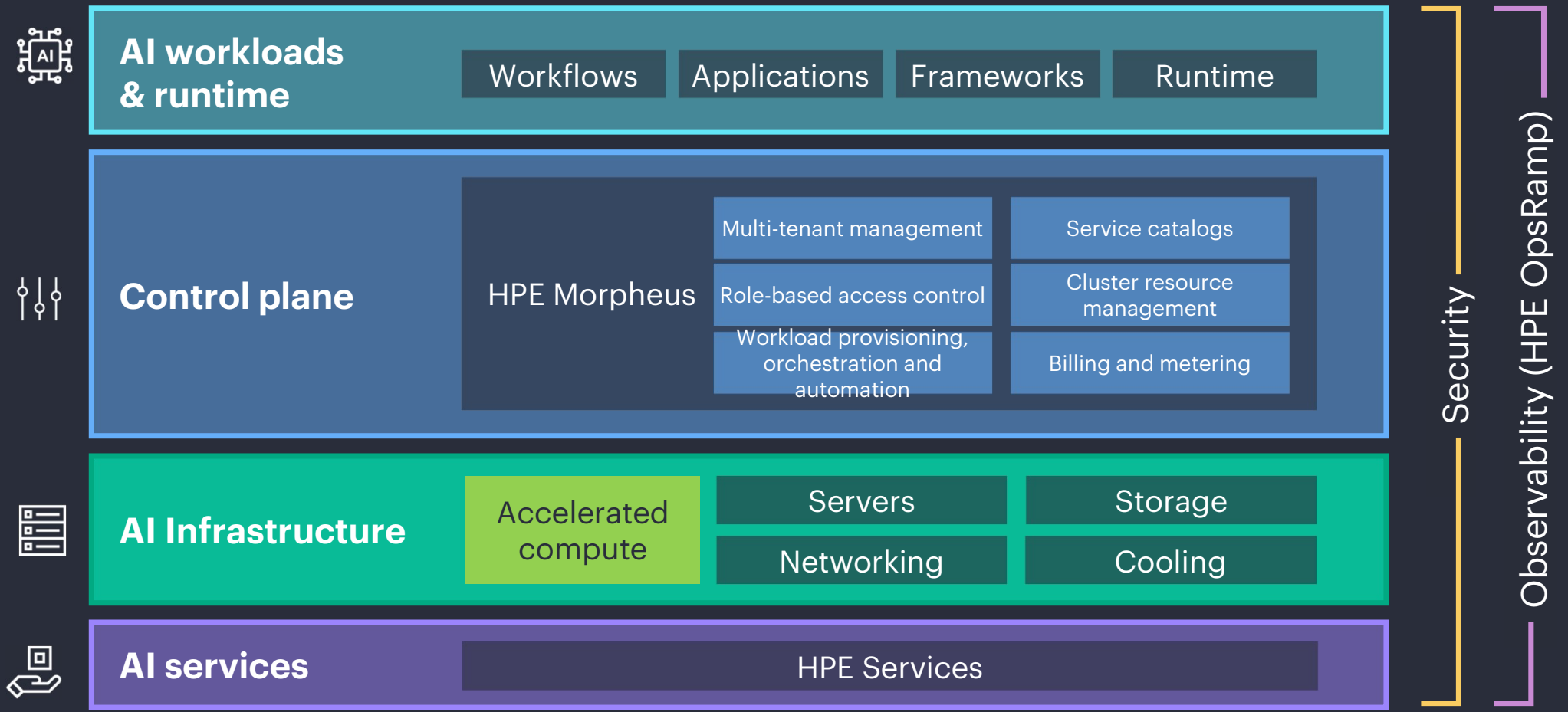


## Flexibility and Scale

- Host **multiple tenants** securely
- **Dynamically provide** GPU-backed tenant resources (GPUaaS)
- Extensive automation & **self-service capabilities for tenants**
- Automated resource monitoring & **billing**
- **Central control** & observability
- Centrally **provide value-added services** to tenants, such as specialized app blueprints

# At scale and sovereign AI factory solutions

Turnkey AI factory    **AI factory at scale**    Sovereign AI factory



# HPE Morpheus Enterprise software

Private Cloud service provider control plane

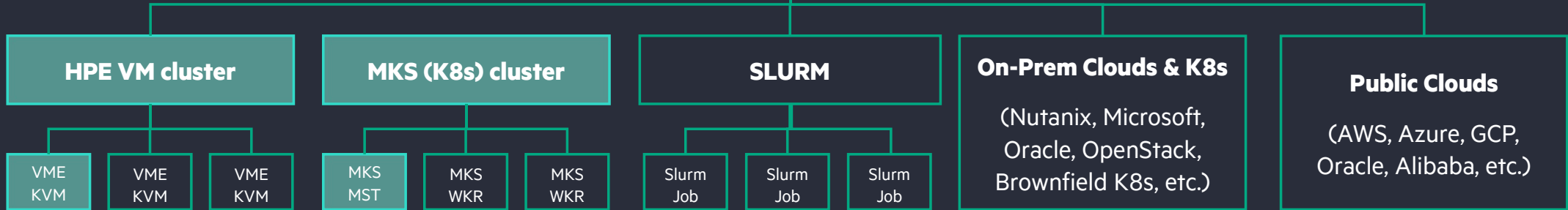
Turnkey AI factory

AI factory at scale

Sovereign AI factory



## Morpheus appliance



# Technologies Powering AI Factories at scale

Turnkey AI factory

AI factory at scale

Sovereign AI factory



Compute: CPU



Compute: GPU



Interconnect



Storage



Development



Management



# Providing GPU Choice

AMD Instinct MI355X GPUs

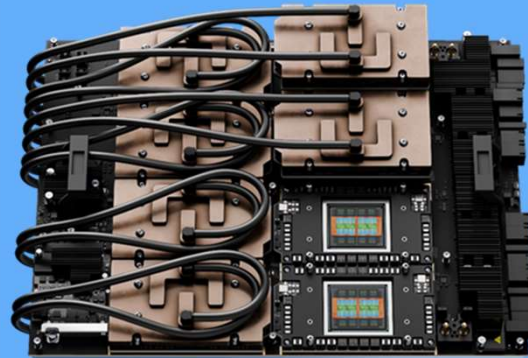
Turnkey AI factory

AI factory at scale

Sovereign AI  
factory



**HPE ProLiant XD685**  
Direct Liquid Cooled (DLC)



**AMD Instinct MI355X GPUs**

## AMD Instinct MI355X GPU

- 288 GB HBM3E GPU memory
- 2.3 TB total GPU memory (8 GPU)
- 520 billion parameters on a single GPU<sup>1</sup>
- 8 TB/s of memory bandwidth
- 34 TFLOPS of FP64 performance

**Thank you for your time!**  
**Questions?**



# Thank You

