



Hewlett Packard  
Enterprise

# Private Cloud AI



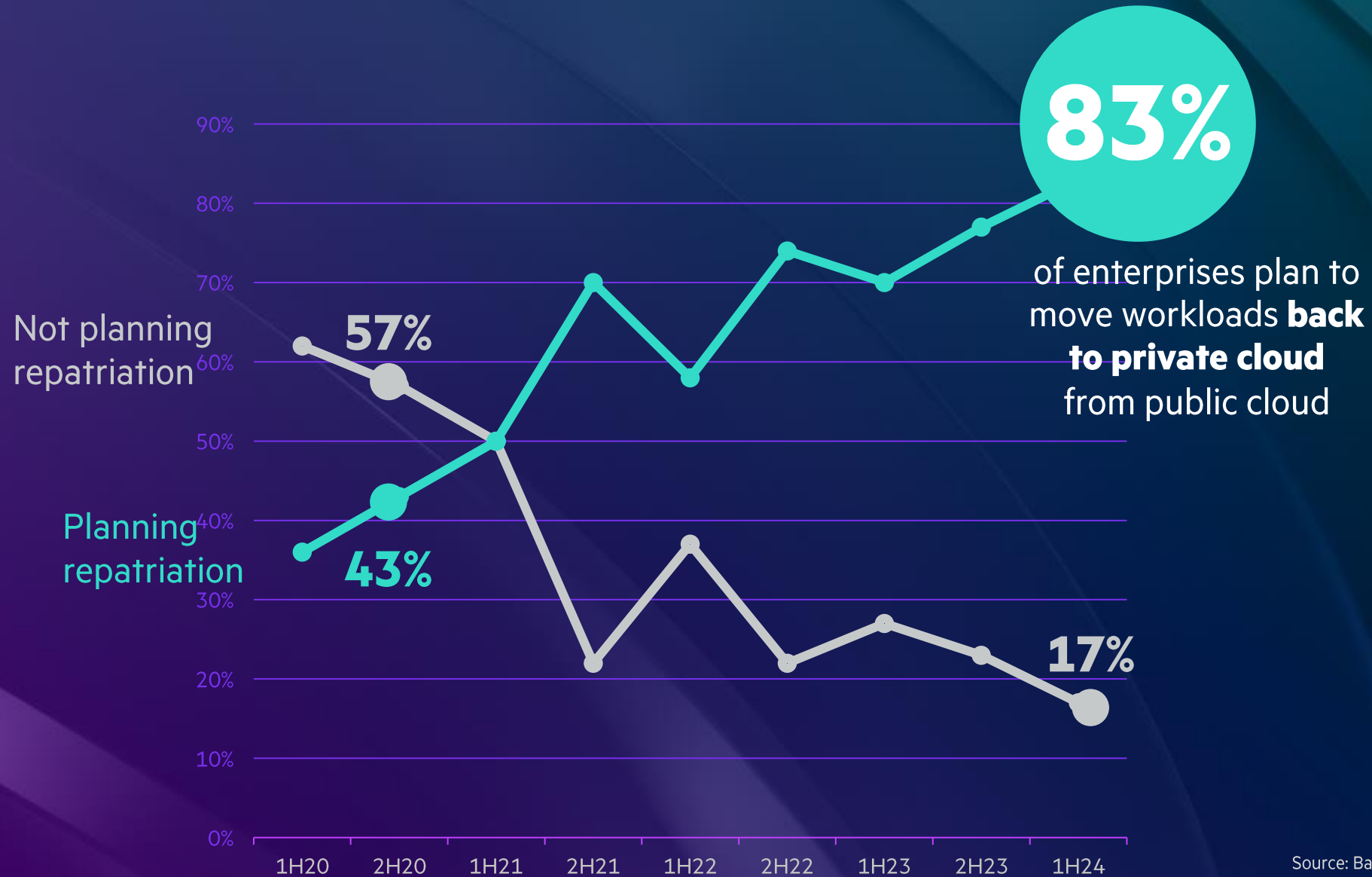
**Guido Trampe**

Sales Manager AI driven solutions

[guido.trampe@hpe.com](mailto:guido.trampe@hpe.com)

+31 624130576

# Gen AI is driving the shift to private cloud



# Top 5 Gen AI use cases



## Code generation

“Co-pilot” for SW engineers to generate code faster—on prem with own codebase and coding standards



## Document creation

Automate form-filling—insurance, prescriptions and more; create marketing briefs, parts manuals and more



## Search engine

Enhance search results with internal documents and information



## Q & A chat

Create an internal support chatbot with your own data for an enhanced knowledge base and faster resolutions



## Customer service

Boost productivity for customer service teams and enhance customer experiences with optimized call centers



# AI is transforming every industry, but getting started is difficult ...

## Enterprise AI adoption reality



Only  
**33%**  
have enterprise strategy

Only  
**~20%**  
of AI projects go to production

**7**  
months from pilot to  
operationalization

## Common enterprise challenges

**Time to  
productivity**

**Data privacy  
and control**

**Data  
accessibility**

**Rate of AI  
innovation/scalability**

# NVIDIA AI Computing by HPE

Co-developed solutions to simplify enterprise AI

Turnkey Private Cloud

AI Services & Training

AI Optimized Systems

Enterprise-grade Control

PEOPLE

TECHNOLOGY

ECONOMICS

Inference, Tuning & Training

Virtual Assistants

Process Automation

Content & Product Creation

# HPE Private Cloud AI

A full-stack, turnkey private cloud for production AI

AI models

AI software

AI infrastructure

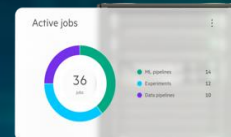
AI services

Instant AI productivity

Secure and unified data access

Enterprise-grade control

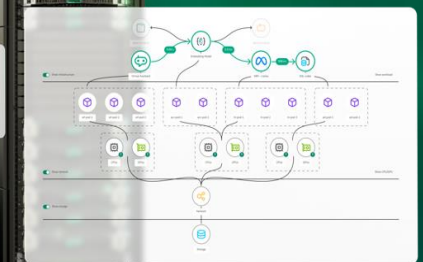
Cloud experience everywhere



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

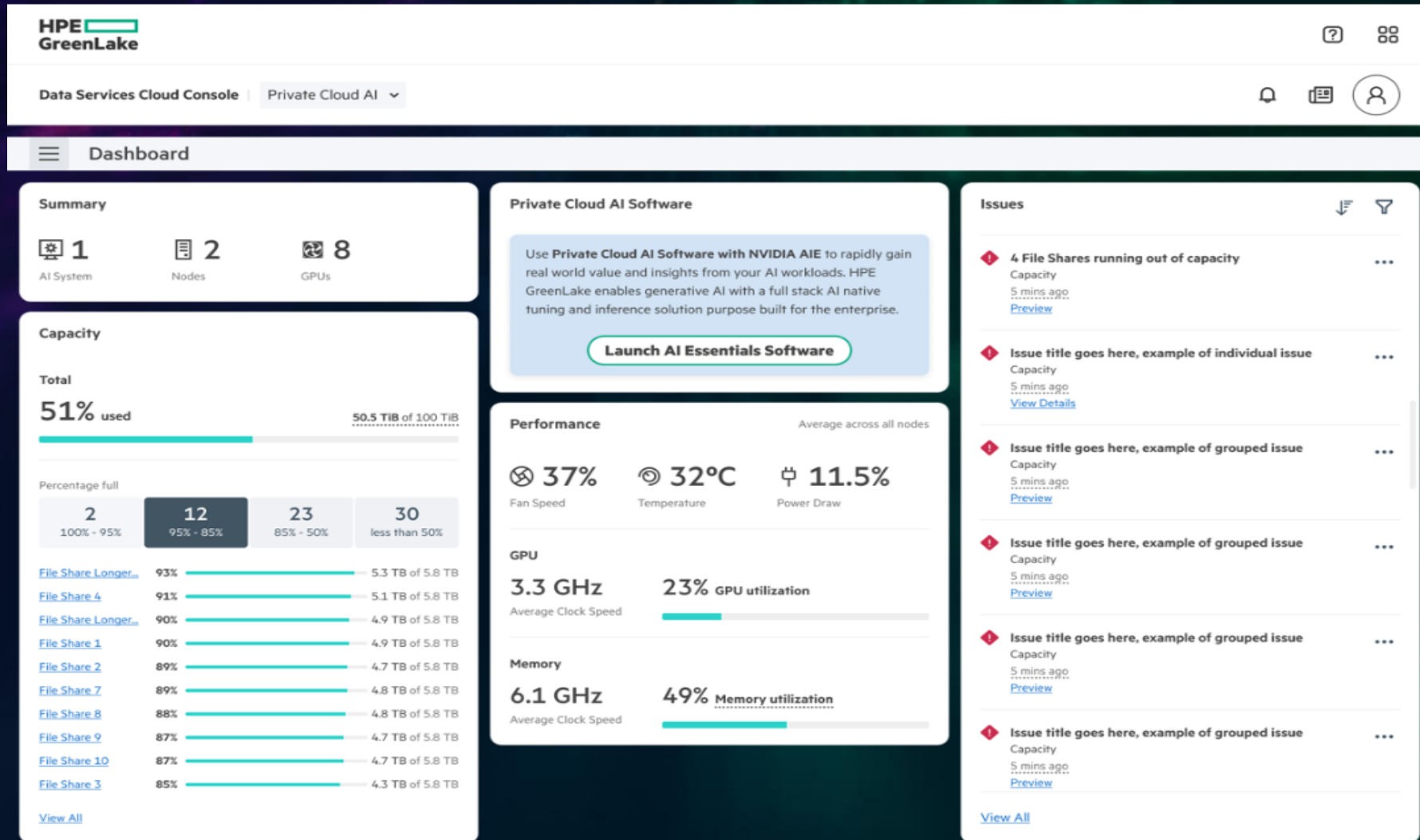


← Inference

RAG

→ Fine-tuning

# PCAI Controlplane Dashboard



# PCAI Controlplane

## User management

The screenshot displays the HPE GreenLake user management interface. At the top, the HPE GreenLake logo and navigation menu (Home, Services, Devices) are visible. A notification banner at the top reads "Automation - Platform Notification Test [click here](#)". The main section is titled "Identity & Access" and "Users". It features summary cards for "Total users" (171), "Active users" (80), and "Unverified users" (0). A search bar labeled "Search Users" is present. Below the search bar, a table lists 171 users. The table has columns for Name, Email, Role, and Last login. A modal dialog titled "Invite User" is open in the foreground, prompting the user to enter an email address and assign an initial HPE GreenLake role. The role is currently set to "Workspace Operator". The modal also includes a "Contact Email Address (Optional)" field and "Cancel" and "Send Invite" buttons.

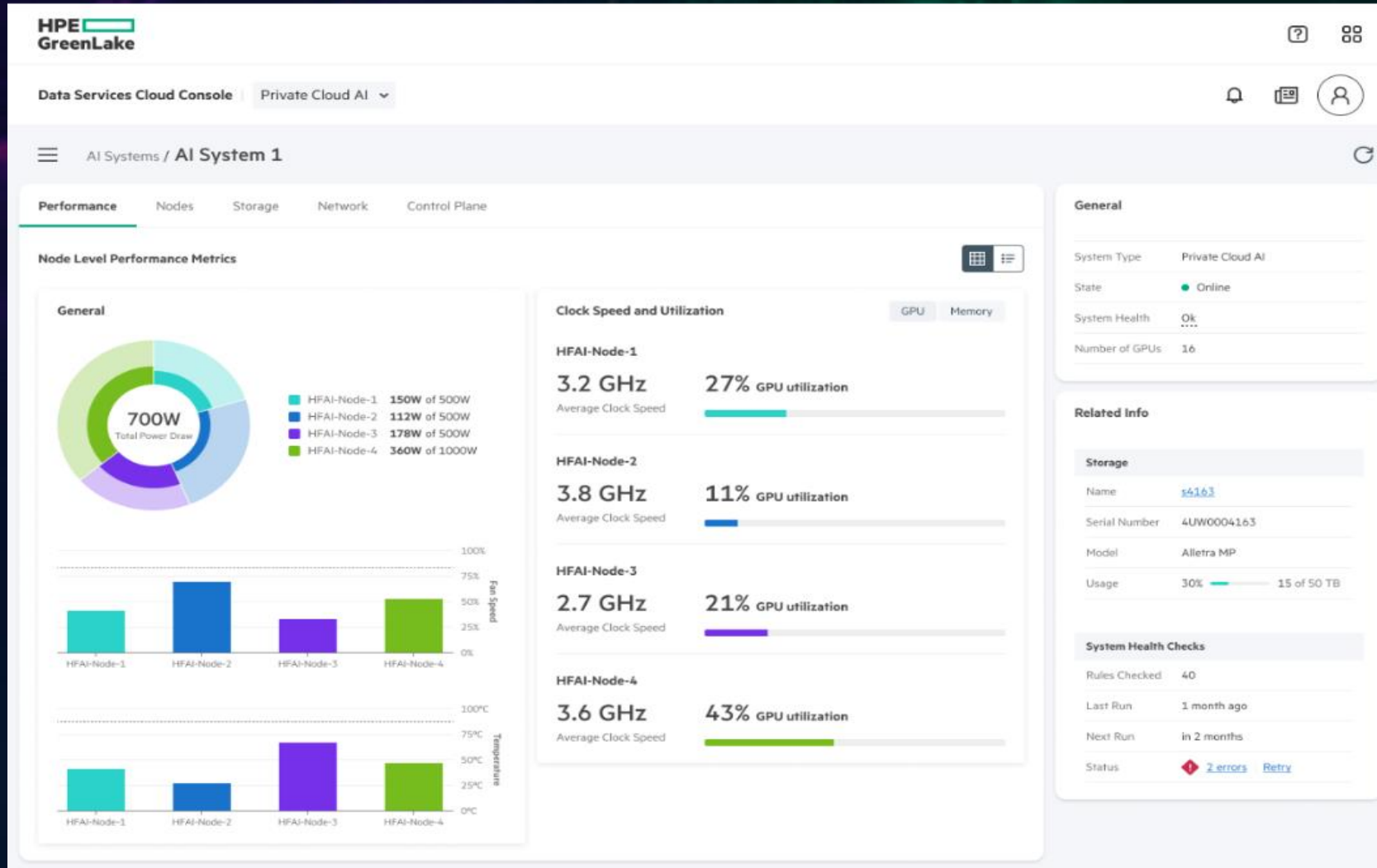
Name	Email	Role	Last login
AM AAFAK MOHAMMAD	aafak-n	Administrator	Apr 1, 2024, 05:36:56
AC Abdullah Chattha	abdulla	Administrator	Apr 24, 2024, 00:12:20
AM Abhijeet Maharana	abhijeet.maharana@hpe.com	VERIFIED Workspace Administrator	Apr 10, 2024, 05:44:33
AN Abhinav Nayak	abhinav.nayak@hpe.com	VERIFIED Workspace Administrator	Apr 23, 2024, 10:57:33

- **Add users and assign them roles.**
  - Manage cluster access, their role in a cluster, by team or by individual user.
  - Once invited, they are ready to log into their HPE Private Cloud AI environment.





# PCAI Controlplane Performance Metrics



# AI essentials

## Pre-configured Development Environment

The screenshot displays the HPE AI Essentials web interface. At the top, it shows the user's name 'Pavo (itg) GreenLake' and the environment 'AI Essentials'. Below this, the 'HPE AI Essentials' header is visible. The main content area is titled 'Notebook Servers' and includes a 'New Notebook Server' button. A search bar is present, and a table lists two items:

Name	Type	Status
group-c19	jupyter	Running
op	jupyter	Running

Below the table, a file launcher window is open, showing a file named 'smart-retail.zip' selected. The launcher provides various options for opening files, including 'Notebook', 'Console', and 'Other' (Terminal, Text File, Markdown File, Python File, Show Contextual Help).

- **End-to-end Development**

- The complete environment for every stage of development, from experimentation to production.
- Connect to any Data Source as if it were local to the Notebook, so data is available directly to the specialist, no matter where it resides.

- **Pre-configured kernels.**

- Data and AI/ML professionals can immediately get started with Notebooks pre-loaded with libraries for their use case, including Spark, Tensorflow and PyTorch.

# Curated set of AI solution accelerators with HPE Private Cloud AI

The image displays two overlapping screenshots of the HPE Private Cloud AI dashboard. The background screenshot shows a grid of six AI solution accelerators, each with a title, description, supported frameworks, and a 'Deploy' button. The foreground screenshot shows a detailed view of the 'NeMo Curator' accelerator.

**Tools & Frameworks**

**NVIDIA AI Enterprise** Data Engineering

**NeMo Curator**  
Version 1.0.0 | Ready

NeMo Curator for GPU-accelerated data curation of high 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 proprietary 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...

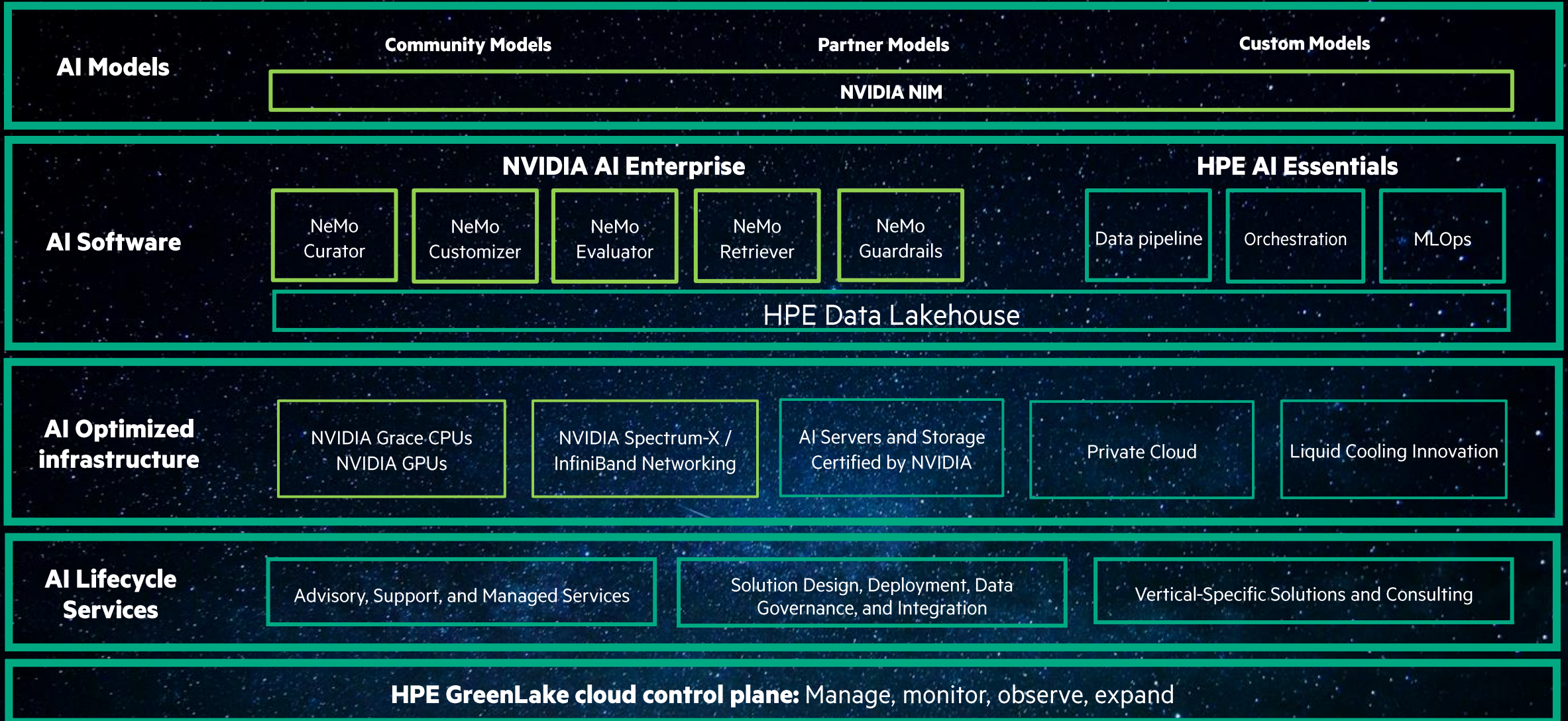
Feast

[Deploy](#)



# HPE Private Cloud AI

Co-developed solutions to simplify enterprise AI



# AI optimized turnkey solutions

Best for

Inferencing

Inferencing  
+ RAG

Inferencing + RAG  
+ Fine-tuning

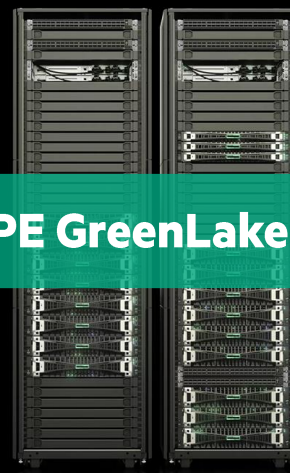
Inferencing + RAG  
+ Fine-tuning



Small



Medium



Large



Extra Large\*

Unified experience through HPE GreenLake cloud

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 250TB

217 TB to 529 TB

670 TB to 1088 TB

670 TB to 1088 TB

Networking

100GbE NVIDIA Networking

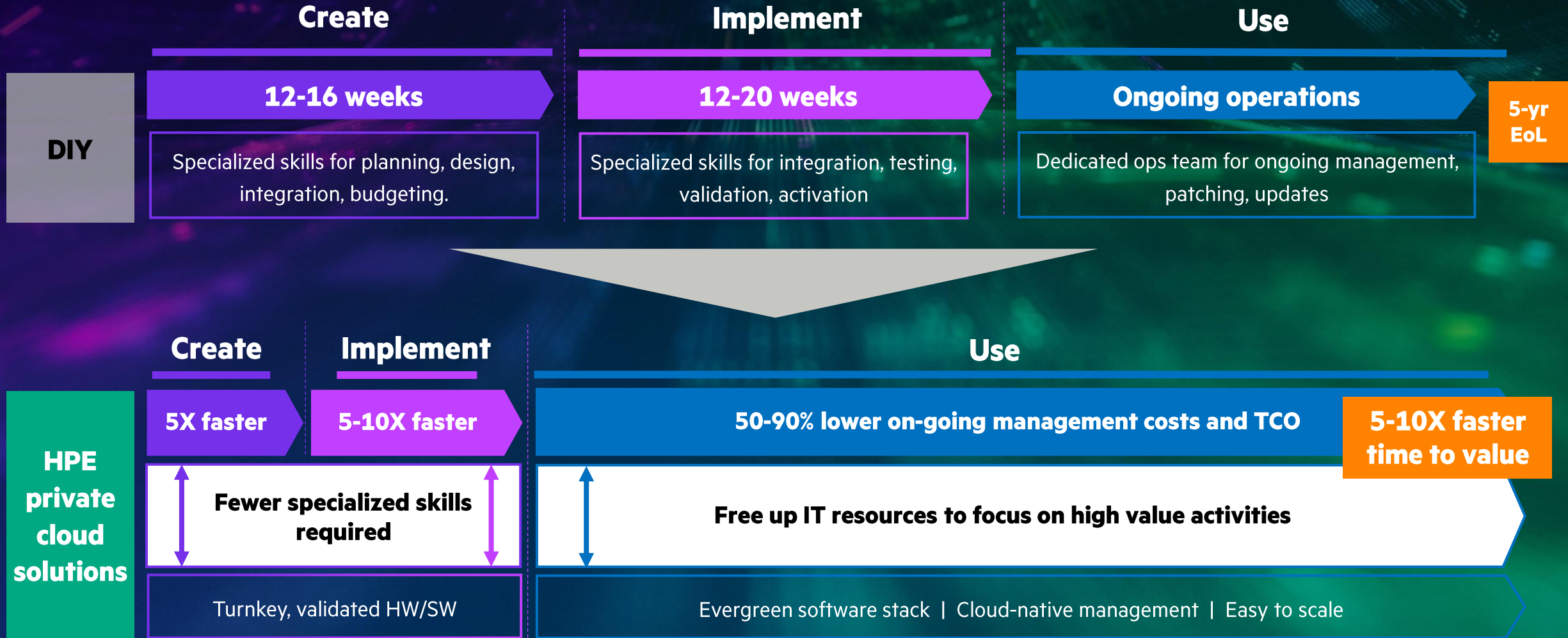
200GbE NVIDIA Networking

400GbE NVIDIA Networking

800GbE NVIDIA Networking

\*XL coming in H1 FY25

# Engineered systems drive speed and efficiency



5-yr EoL

5-10X faster time to value



**Within the next 3 years,  
anything not connected  
to AI will be considered  
obsolete or ineffective.**

McKinsey & Company



# What to remember about HPE Private Cloud AI

A full-stack, turnkey private cloud for production AI

## Ready to run out of the box

Fully managed, pre-integrated NVIDIA accelerated computing, networking, and software with HPE compute, storage, and software

## Scale as you grow

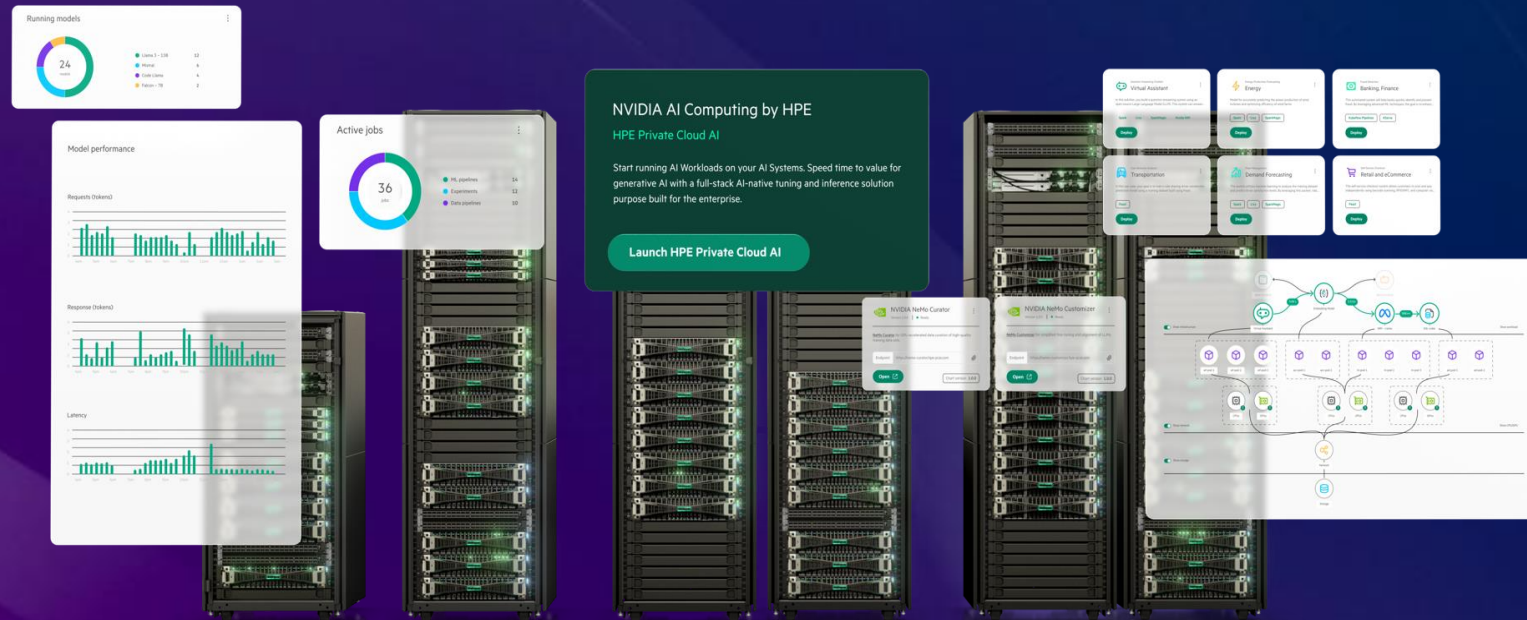
Available in modular configurations sized for inference, RAG-based applications and model customization and fine tuning

## Data privacy and control

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

## AI lifecycle management

Keep up with AI innovation with access to the latest AI models, development software and pre-configured solution accelerators







**Thank you!**

