

# Pure Storage Solve 5 AI Infrastructure Challenges

Andrew Ho (何與暉)

Manager, Systems Engineering, Greater China

# 業界首創 專為儲存設計的 生成式AI Copilot

## DEMO



# Pure AI Copilot

My oracle workload hosted on host-oradb-3 is experiencing slowness. Can you identify the root-cause and suggest remediation measures?



I see that **host-oradb-3** is connected to **prg1-labrack4-fax50-13**.

**prg1-labrack4-fax50-13** is showing high average read (**5.2ms**) and write (**2.3ms**) latency, which is higher than the average latency for the last 6 months.

A likely cause of the higher latency is the **high busyness (95%)**. Average busyness for the last week and last 3 months is 85% and 80% respectively.

Here are the details on the **3 most active volumes** on the array prg1-labrack4-fax50-13:

Volume Name	Host Name	Avg Read IOPS	Avg Read BW (MB/s)	Avg Write IOPS	Avg Write BW (MB/s)
vol1	host-oradb-3	2,010	90	10,243	312
vol2	host-oradb-3	4	4	12,385	1,463
vol3	host-oradb-3	4	4	12,385	1,463



Your organization's name is **The Pure Platform**.

Your organization's security strategy is ranked as follows:

- Overall Percentile: 27% out of over 12K organizations
- Industry Vertical: Tech HW & SW
- Vertical Percentile: 38%

Would you like to understand how your security strategy



To improve your organization's security, here are the key areas where improvement is needed:

## Feature

Alert Watcher Monitor

SMTP Channel Security

Syslog Channel Security

SNMP Channel Security

NTP

Support Channel Security

Directory Services Channel Security

SAML Channel Security

Offload Channel Security

Array Connection Encryption

忙啥都行  
別忙Pure

Monday, June 17, 2024.

Slowness:

High latency, High busyness

- **Latency:** Average read latency is 5.2ms, and peak read latency is 2.3ms.

- **Busyness:** Average array busyness is 95%, and peak busyness is 100%

FileStore-2

High Space Utilization

- **Capacity Usage:** 85% of the available capacity.

- **Growth Rate:** 15% in the last 6 months

Would you like to proceed with further analysis for the unhealthy arrays?

# Why AI? And Why Not?

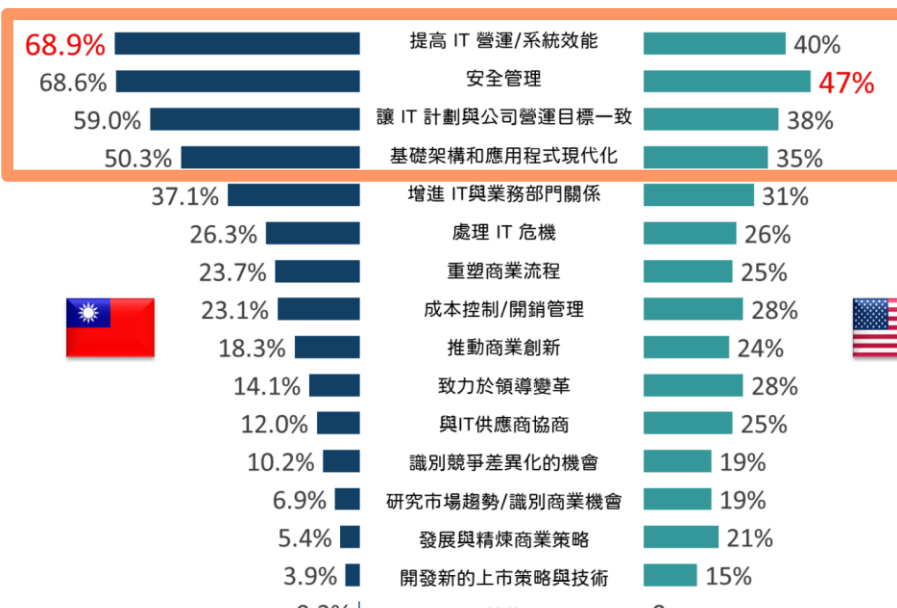


## 一個IT人服務多少企業員工?

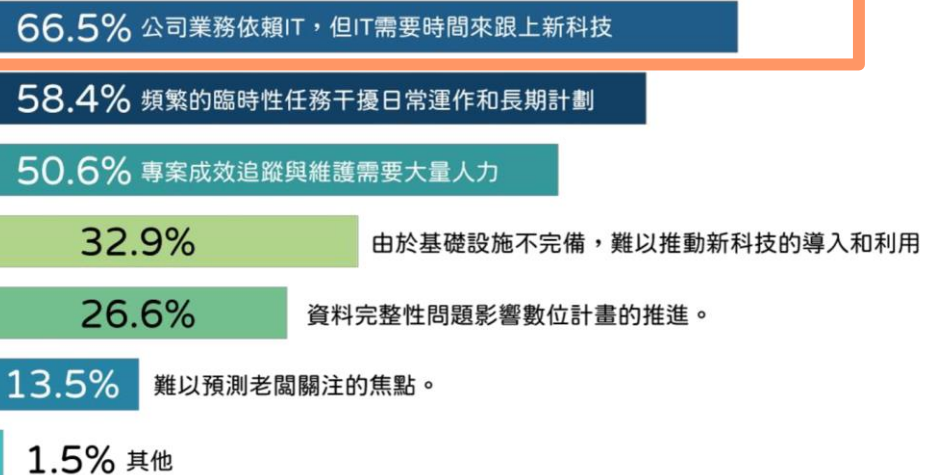


CIO Taiwan 55

## 目前您聚焦的重點項目?



## 科技導入，最具挑戰或最艱難的任務?



# AI Storage Infrastructure Challenges

Performance

Operational  
Efficiencies

High  
Costs

Reliability,  
Resilience

Future Growth  
Uncertainty

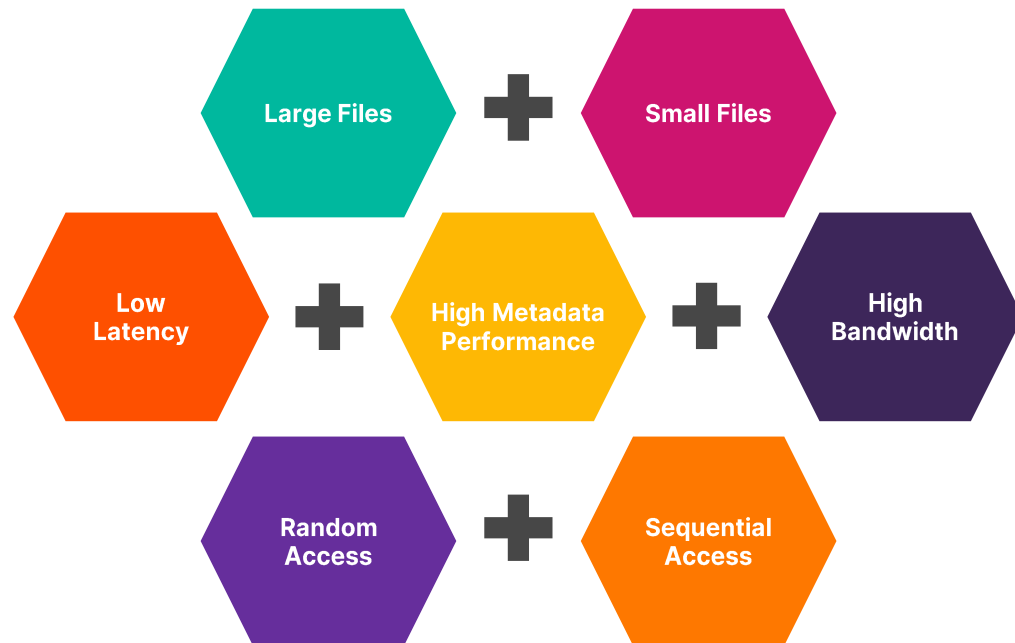


# Deliver performance beyond throughput and IOPS

*Train your model in days instead of months*

any job | any protocol | any size | any object count | any processing type

Ingestion | Persistence | Processing | Training | Inference



AI workloads require multi-dimensional performance

## Predictable expansion

Scale performance granularly on demand

## Performance without complexity

No tuning needed

## Performance without deep expertise

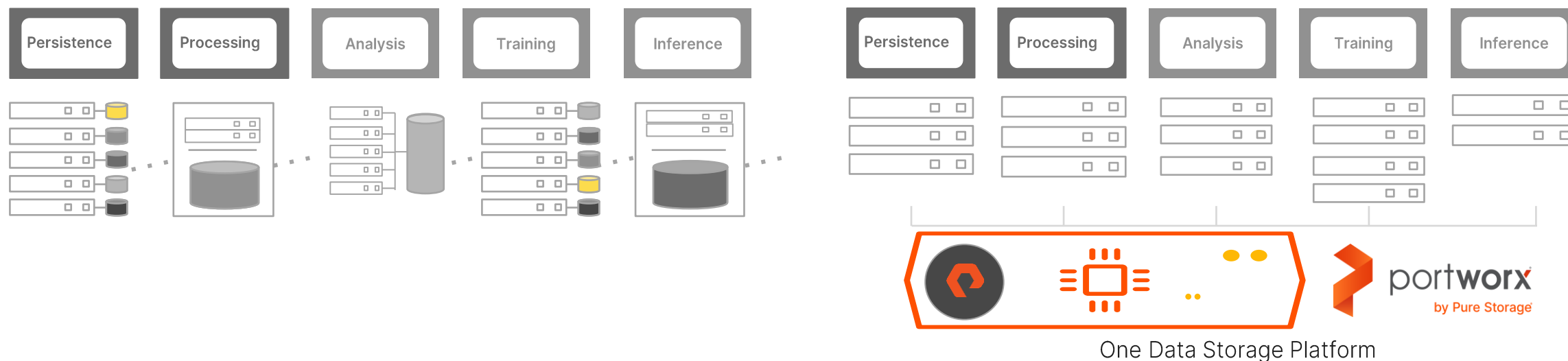
Intuitive and easy to use interface

## Industry leading Watts/IOPS & Watts/GB/Sec

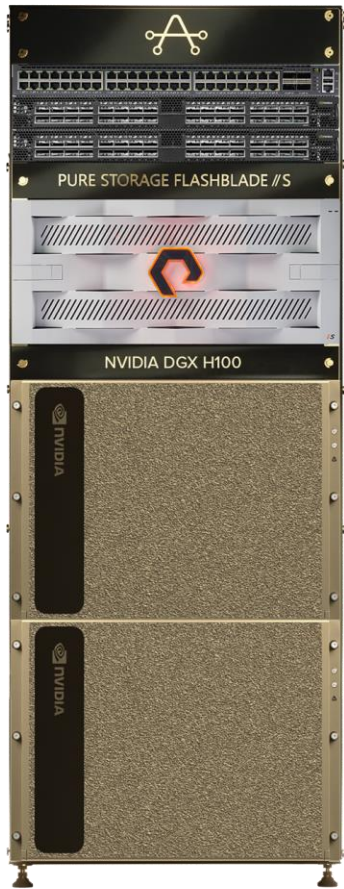
85% less energy consumption

100s of organizations power their AI training with Pure  
1,000s run high throughput applications on Pure

# Deliver efficiency for data curation and AI steps on one data storage platform



# AI Infrastructure Re-imagined, Optimized, and Ready for Enterprise AI-at-Scale



**Industry's first to simplify AI-at-scale**

**Aligned with NVIDIA**

DGX BasePOD Reference Architecture

Support for GPU Direct Storage

**Pure Storage is in use for AI by over 100 organizations**





# NVIDIA DGX SuperPod Certification\*

Futureproof AI performance  
with unmatched simplicity

Fully validated design with performance  
guarantees

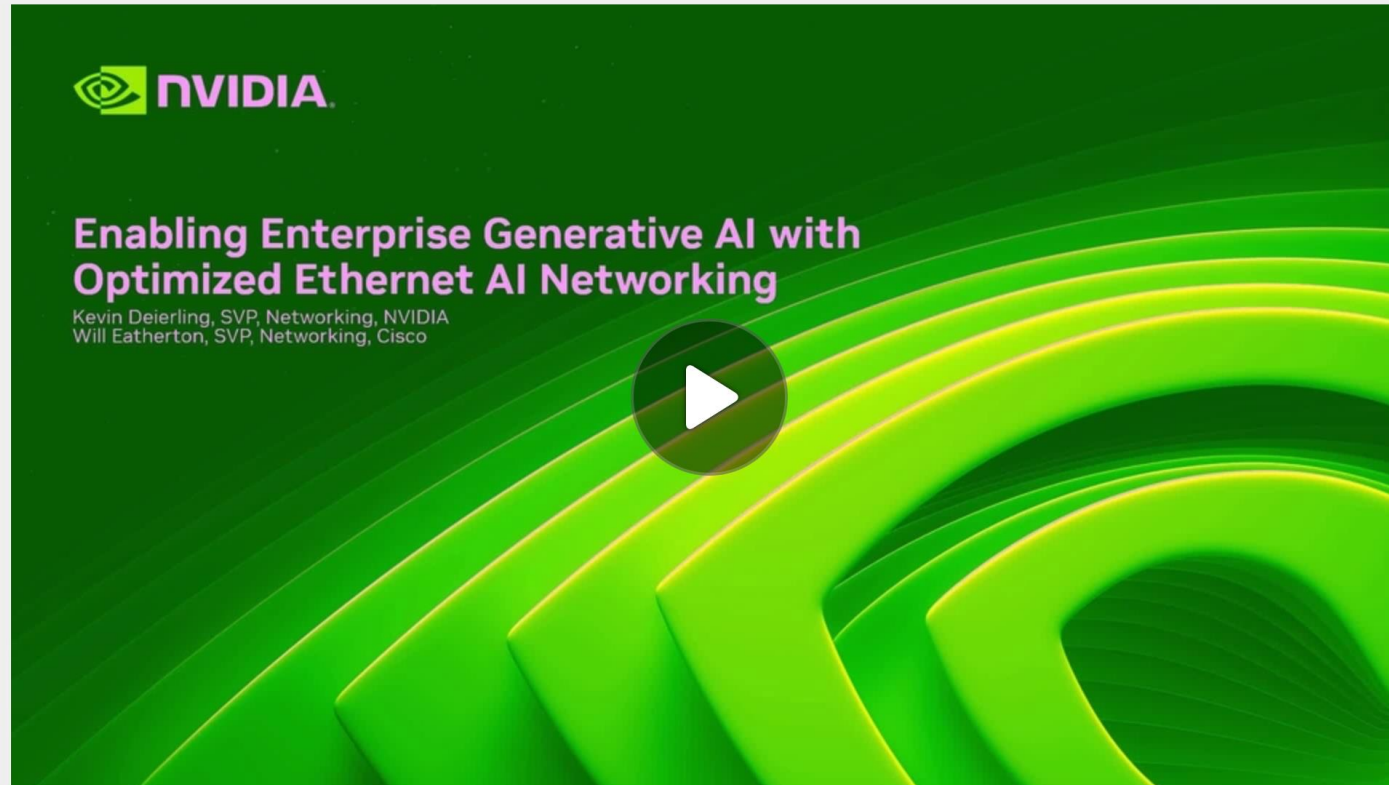
Easily scale in both size and performance for  
maximum AI performance at the right price

Seamless, non-disruptive upgrades increase  
performance without downtime

Ethernet-based storage greatly simplifies  
enterprise integration for large scale AI  
training and inference

\*NVIDIA DGX SuperPOD certification expected H2 CY2024. While Pure Storage is committed to pursuing these certifications, it should be understood that any forward-looking statements about certifications are based on current expectations and are not promises or guarantees.





## Enabling Enterprise Generative AI with Optimized Ethernet AI Networking

**Kevin Deierling**, SVP, Networking, NVIDIA

**Will Eatherton**, Senior Vice President, Networking Engineering, Cisco

Rate Now

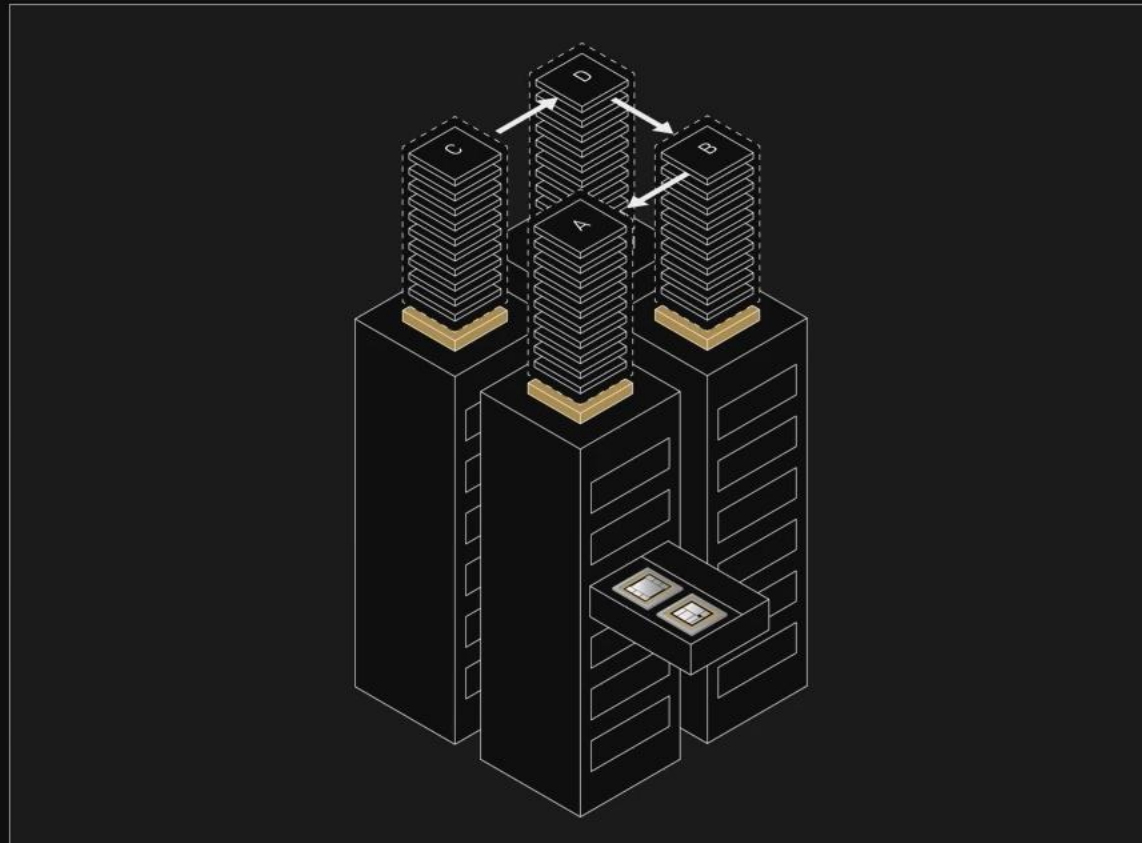
Share

Favorite

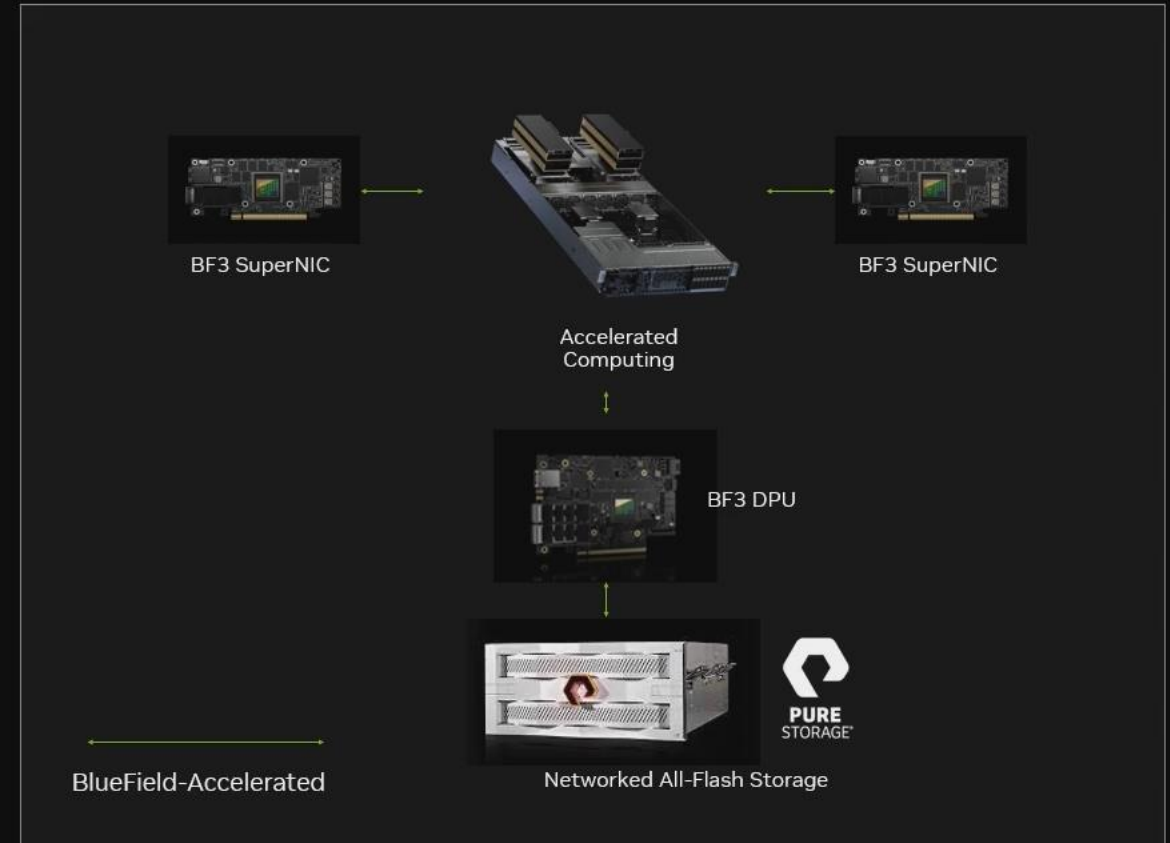
Add to list

# Deploying RAG at Enterprise Scale

Distributed, accelerated RAG workflows across 100's of enterprise data sources and servicing 1,000s of users



**Modern Enterprise Data Center**  
Disaggregated, Micro-Services, Scaled Out

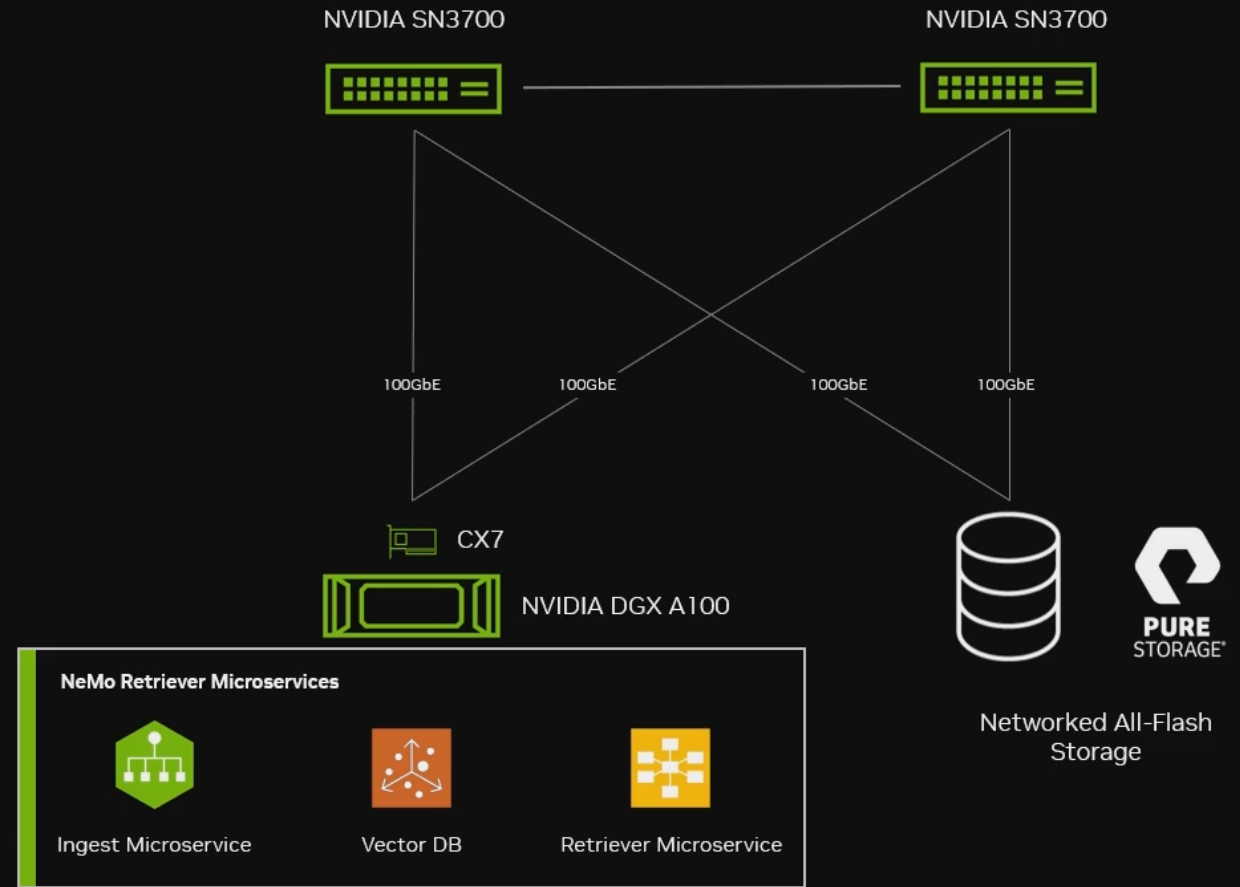
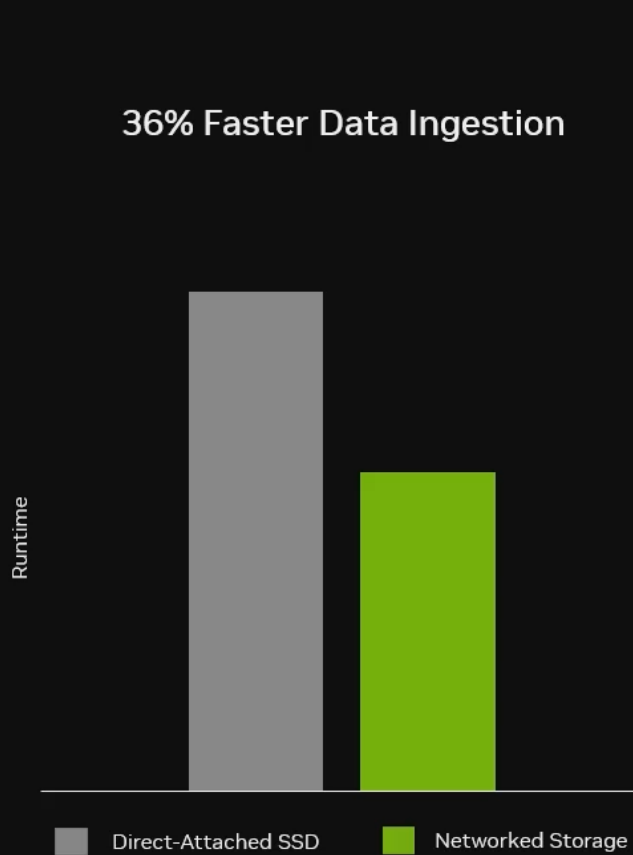


**Accelerated Ethernet Networking for AI**  
Powered by NVIDIA BlueField

# Networked Storage Improves Data Ingestion Performance

NeMo Retriever Microservices with All the Benefits of Networked Storage

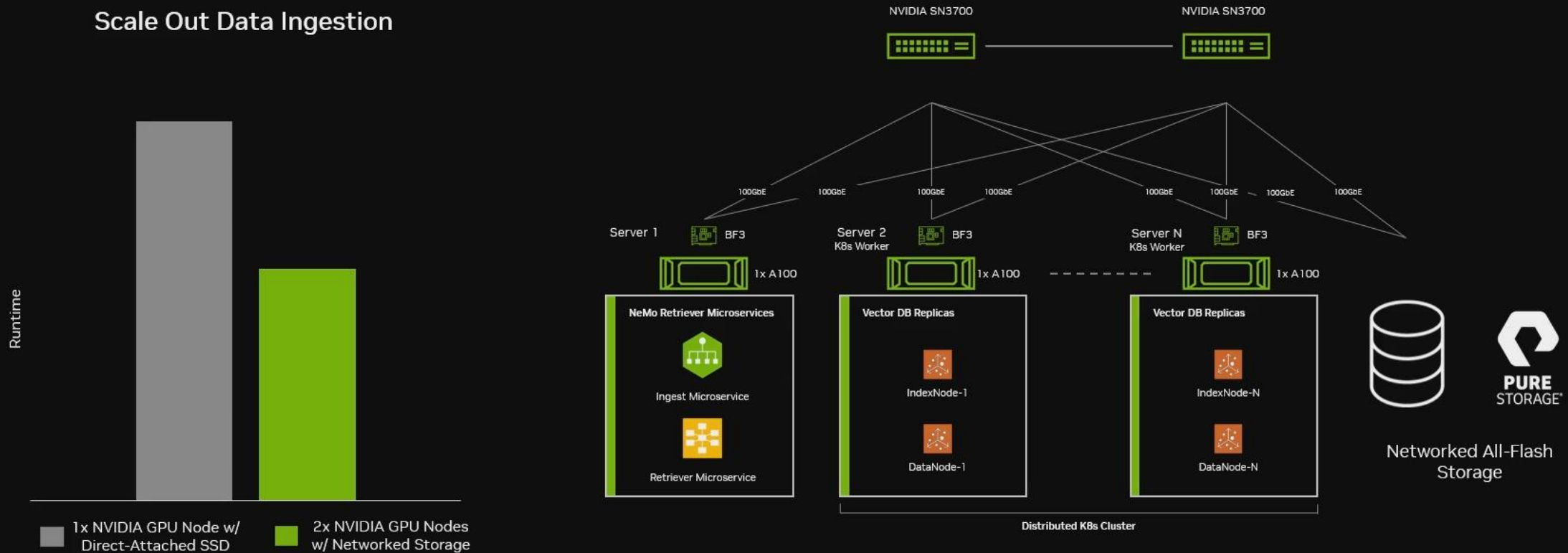
36% Faster Data Ingestion



# Multi-Node Data Ingest Scale Out

Optimized Networking & Storage to Scale Out Embedding and Indexing Performance

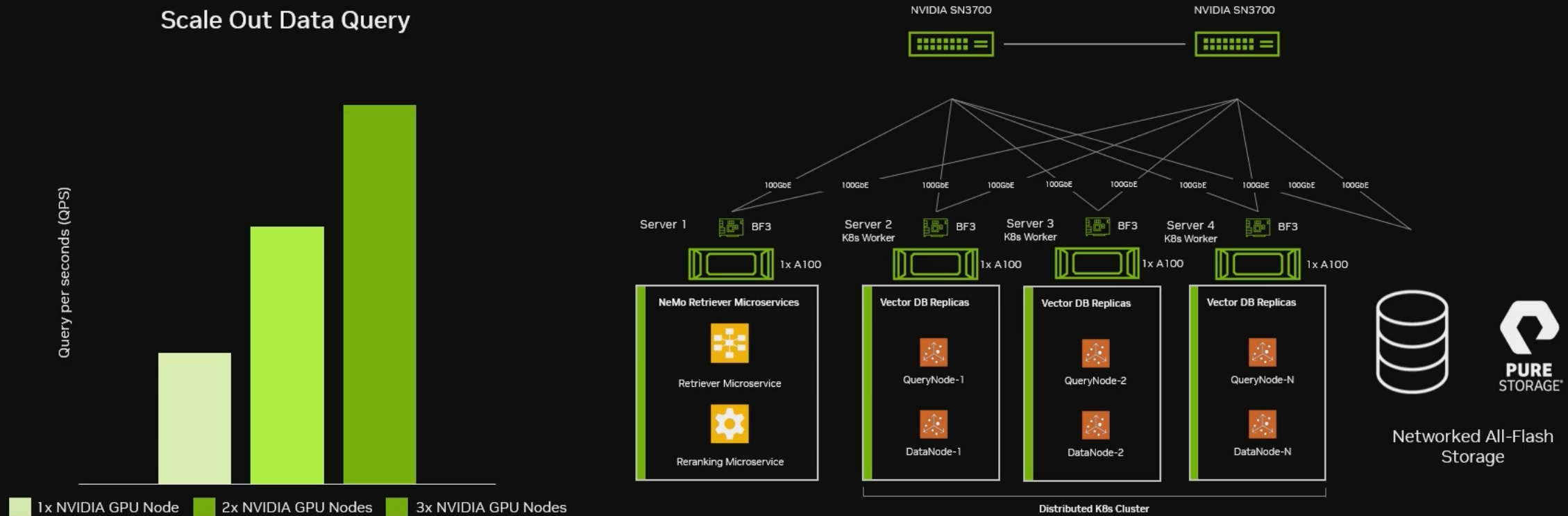
## Scale Out Data Ingestion



# Multi-Node Data Query Scale Out

Optimized Networking & Storage to Scale Out Query Performance

## Scale Out Data Query



# Networked Storage is the Optimal Data Platform for Generative AI



## Linear Scaling

Deliver high throughput, low latency  
Scale performance and capacity  
Support dozens of AI servers



## Peak Utilization

Eliminate stranded local storage  
Share data across servers and GPUs  
Support multiple stages of the AI workflow



## Data Protection

RAID and hot spare drives  
Backup and disaster recovery  
Data encryption



## Composable Storage

Rapid and flexible provisioning  
File and/or object storage  
Multi-tenant and cloud support

[Read the Blog: Scaling Enterprise RAG with All Flash Networked Storage](#)

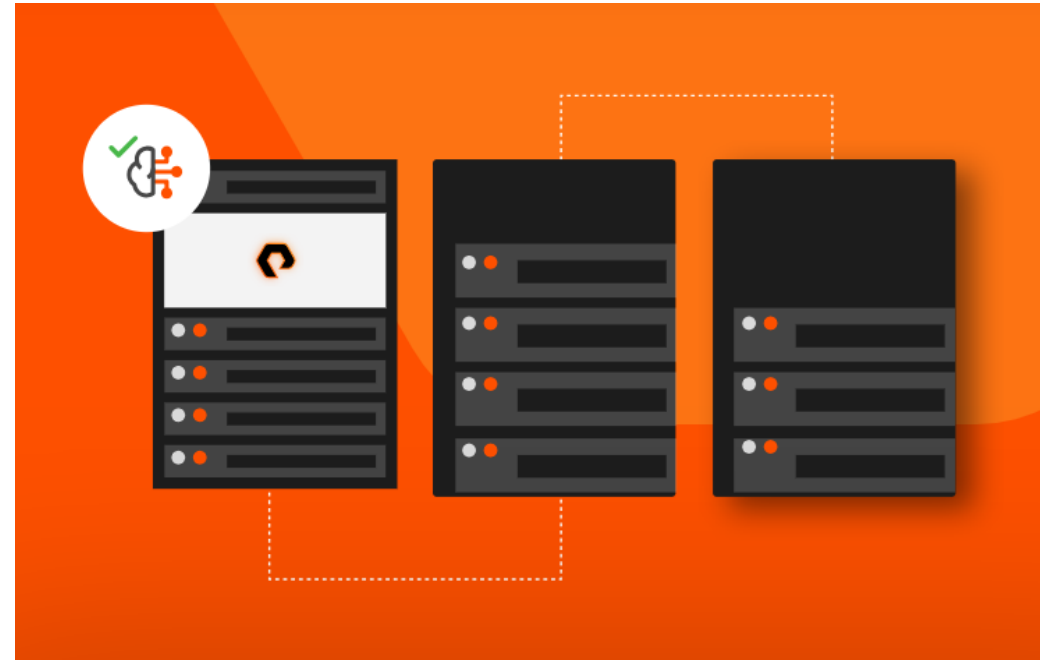


Pure Storage FlashBlade

## Fast Track AI Adoption with Pure Storage

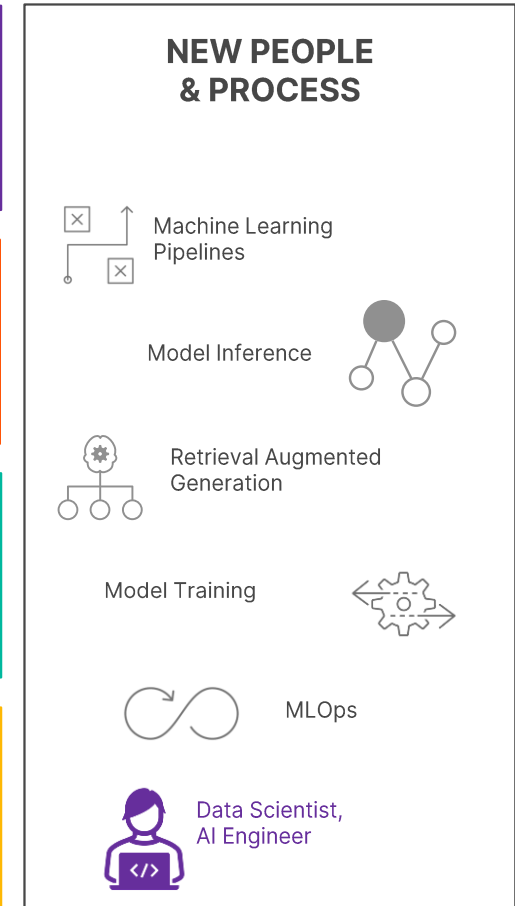
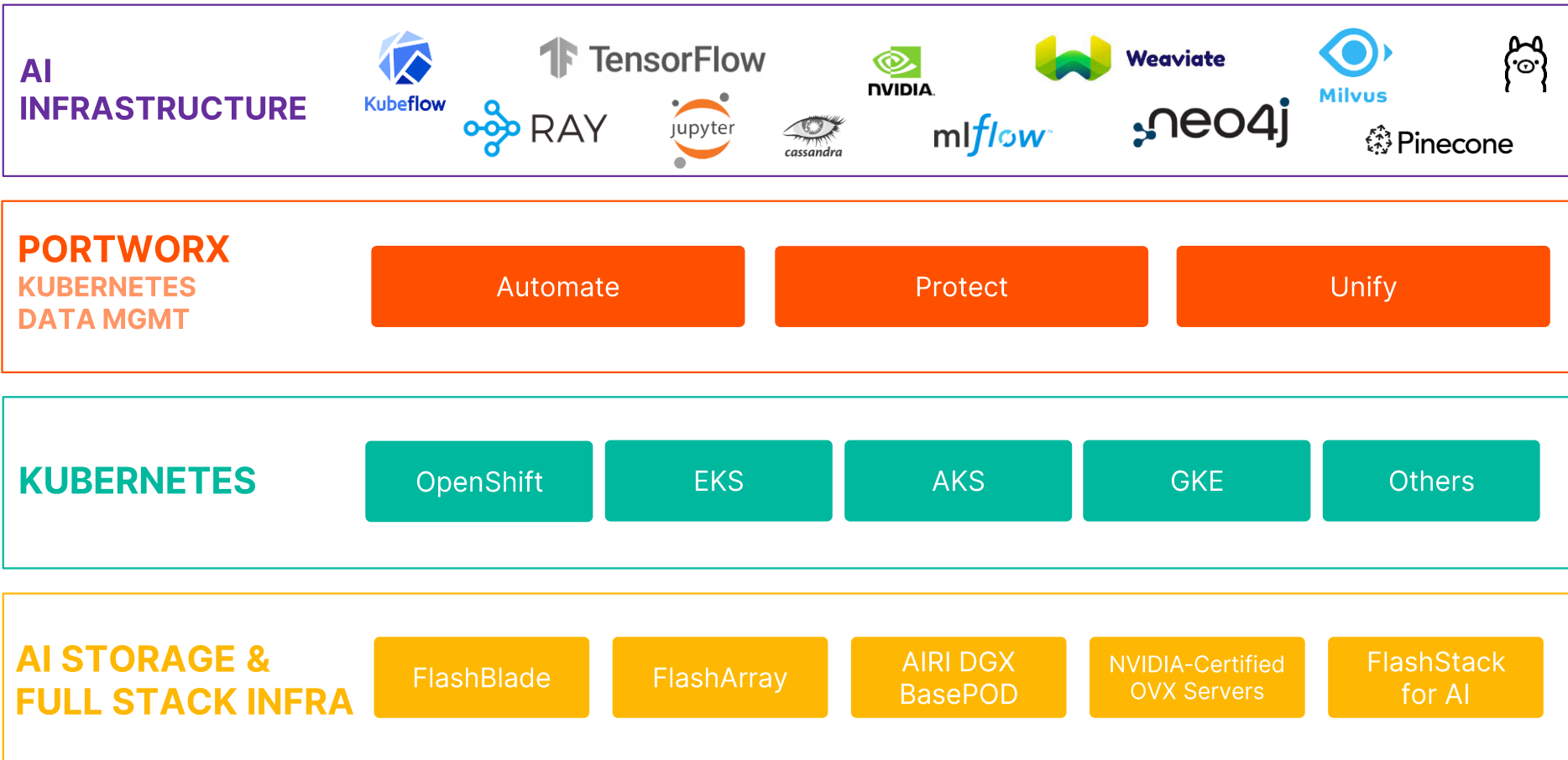
### Validated NVIDIA OVX Server Storage Reference Architecture

- Accelerate Deployment
- Simplify AI Infrastructure
- Expand compute server choices
- Powered by FlashBlade//S and NVIDIA OVX Servers with L40S GPUs





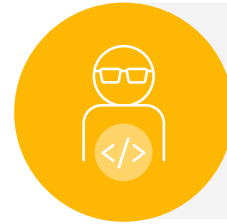
# Simplify Model Serving with Portworx by Pure Storage



# Accelerate Model Training and Serving using Portworx



Train your models in the cloud and serve them on-prem using Portworx



Database Platform As a Service for self-service deployments



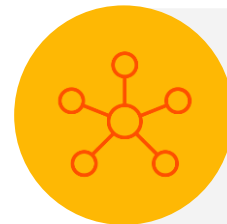
Avoiding wasted GPU resources by running Data on Kubernetes



Protection from node failures, zone failures, cluster failures



High Performance storage for Vector databases running on Kubernetes



Run AI models and applications on the same unified stack



# Integrate Mission-Critical Data with AI Clusters

Pure Secure Application Workspaces make storage transparent to application owners with automated access to AI innovation



Secure Multi-tenancy

Policy Governance Tools

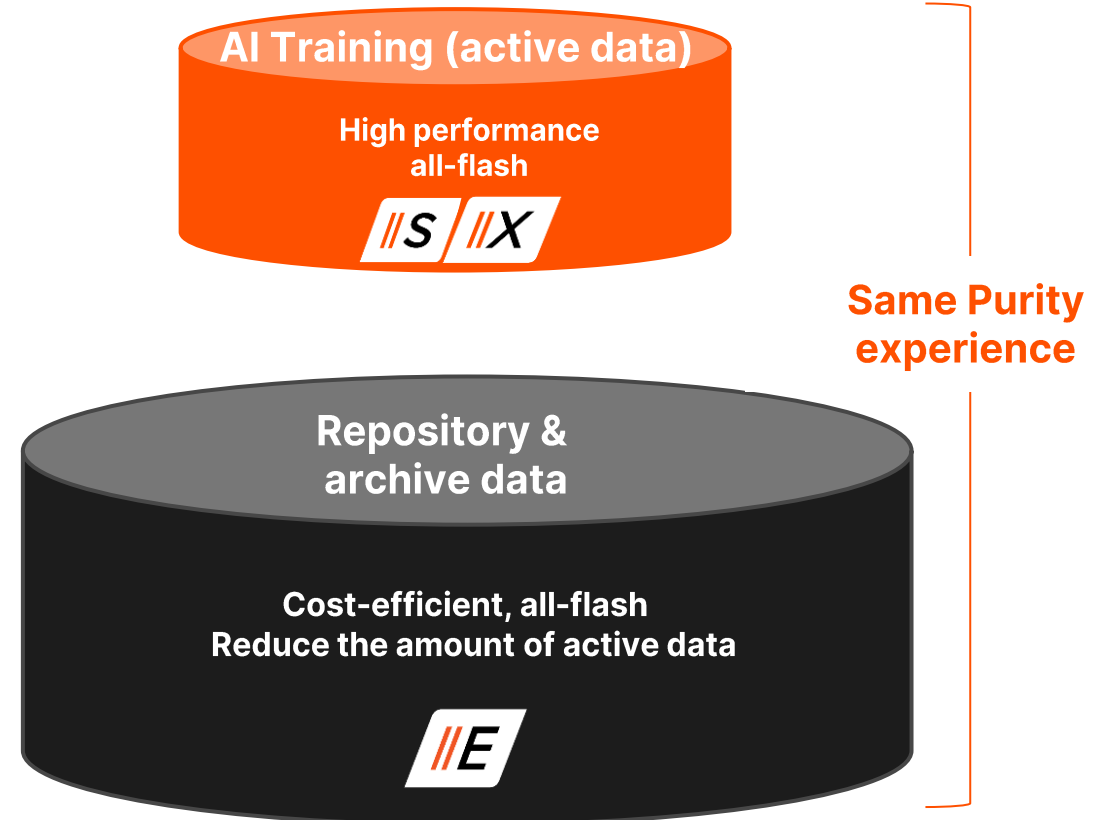
Kubernetes Container Management

# Lower TCO for both active and repository data

All data is always accessible

## Data trends with AI

- Training data sets reaching PBs
- Data retention period growing
- Need to store inference results for further model training
- Use of historical data for AI training increasing
- Need to lower access time for repository data
- SLAs for repository data, same cost



2-5x less space & power | 10x-20x the reliability | 50% lower TCO | 85% less e-waste

Faster time to results irrespective of where the data lives  
Modern, all-flash storage for AI repository data at disk economics

# Evergreen//One for AI

Solve for AI unpredictability with Storage-as-a-Service

Focus efforts on advancing AI initiatives versus managing storage

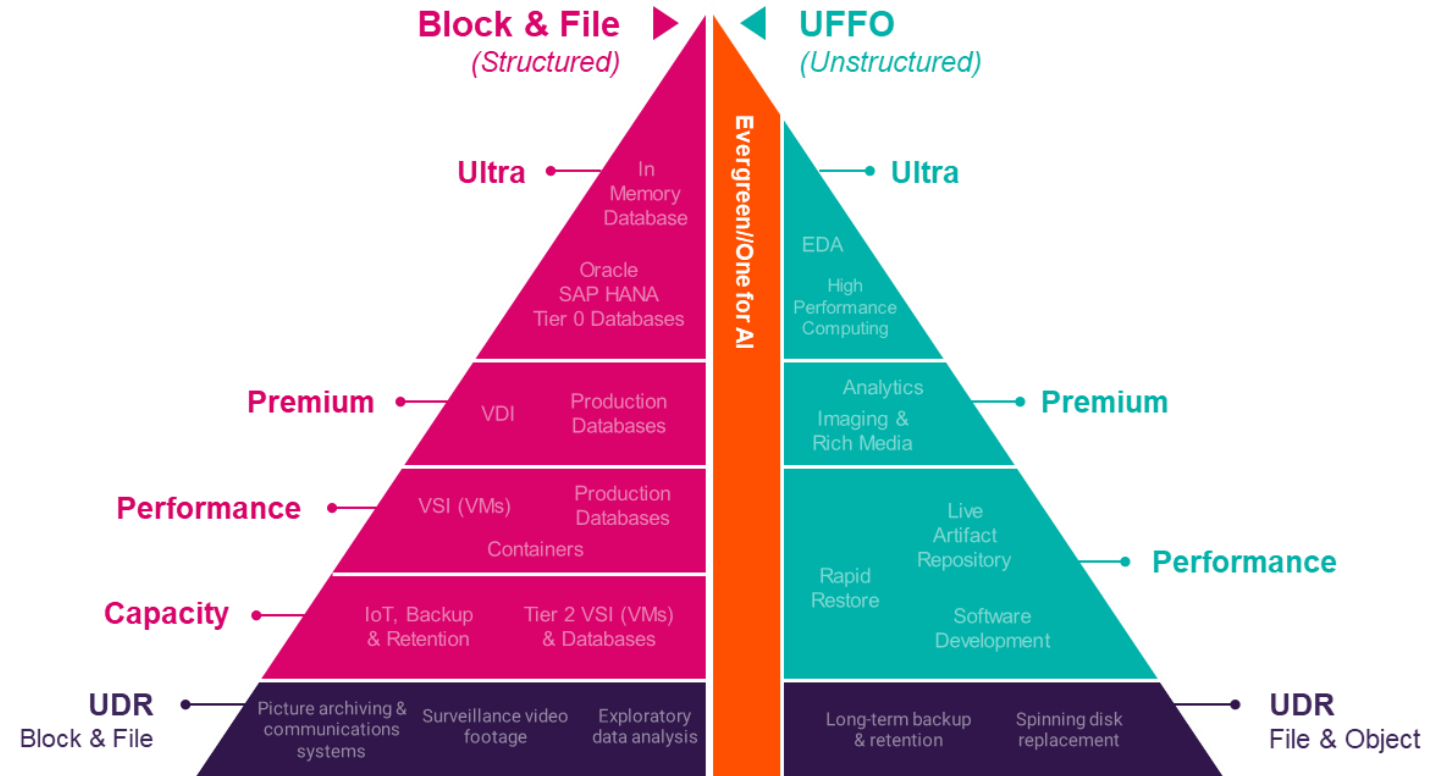
Receive guaranteed performance based on maximum bandwidth requirements for GPUs

Proven for AI with NVIDIA OVX server, DGX BasePOD, and DGX SuperPOD\* certification





Get a cost-effective base rate per GB/s of bandwidth and a low marginal rate for actual data stored

\*NVIDIA DGX SuperPOD certification expected H2 CY2024. While Pure Storage is committed to pursuing these certifications, it should be understood that any forward-looking statements about certifications are based on current expectations and are not promises or guarantees.

## Evergreen//One Service Catalog



### Designed for any AI workload

 Inferencing
  RAG
  Fine-tuning
  Training



# Data Resiliency: **Solution Areas**

Pure Storage Data Resiliency Portfolio

**Snapshot  
Recovery &  
Archiving**

**Backup & Rapid  
Recovery**

**Replication &  
Disaster  
Recovery**

**Container  
Protection &  
Security**

**Security  
Analytics &  
Ransomware  
Protection**

**Data Protection  
Software  
Integration**

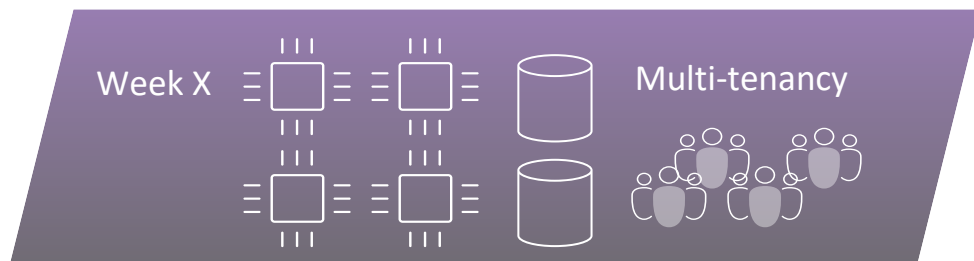
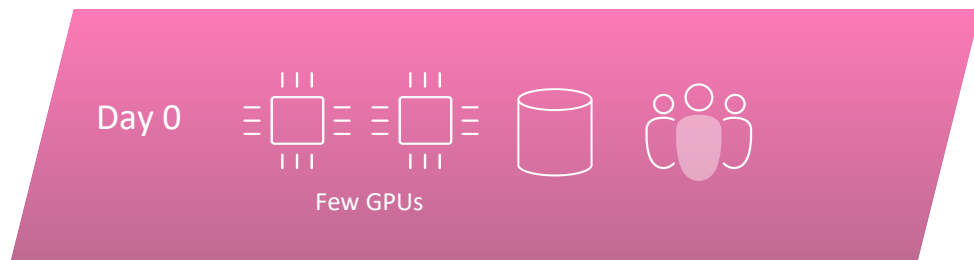
**Active/Active  
Business  
Continuity**

**Hybrid Cloud  
Protection &  
Resiliency**

# Future-Proof AI Storage

## As you add GPUs, effortlessly adjust and upgrade storage

As the adoption of AI grows in your organization, you add more GPUs and storage



Tuning and upgrading storage everytime you add GPUs need not be long and tedious



**Deploy once, No rebuys**



**Never migrate your AI environment**



**Always expandable**



**Always up to date**



**Always on AI environment, Full performance**



# What customers are realizing



**2X**  
increase in storage data  
processing for faster  
AI performance

**> 2.5X**  
increase in GPU usage, from  
30% to 80%

**MEDIAZEN**

Reduced time to market  
for new AI services from  
6 - 12 months to  
**2 weeks**

Accelerates AI-powered voice  
recognition modeling cycle by  
**96%**

## Industry leading storage technology

**>10x more reliable**

>10-30x Fewer Service Visits

**2-5x\* less power and space**

10x vs Existing Hard Disk Systems

**50%+ lower TCO**

**Most consistent product line**

1 Purity, 2 Architectures, 1 Management

**5-10x less labor to operate**

**Products are never-obsolete**

Always-improving, Non-disruptively





Uncomplicate Data Storage, Forever