Deepak Bansal

Corporate Vice President – Microsoft

- **Leads Azure SDN platform** and Secure Networks pillar.
- Expert in **cloud, AI infrastructure, and distributed systems**; 50+ patents.
- **Opensource leader**: Initiated DASH and ReTiNA projects.
- **Founding member of Azure**; ex-VP Azure Networking; former ONF board member.
- **Education**: MS, MIT; BS, IIT Delhi.

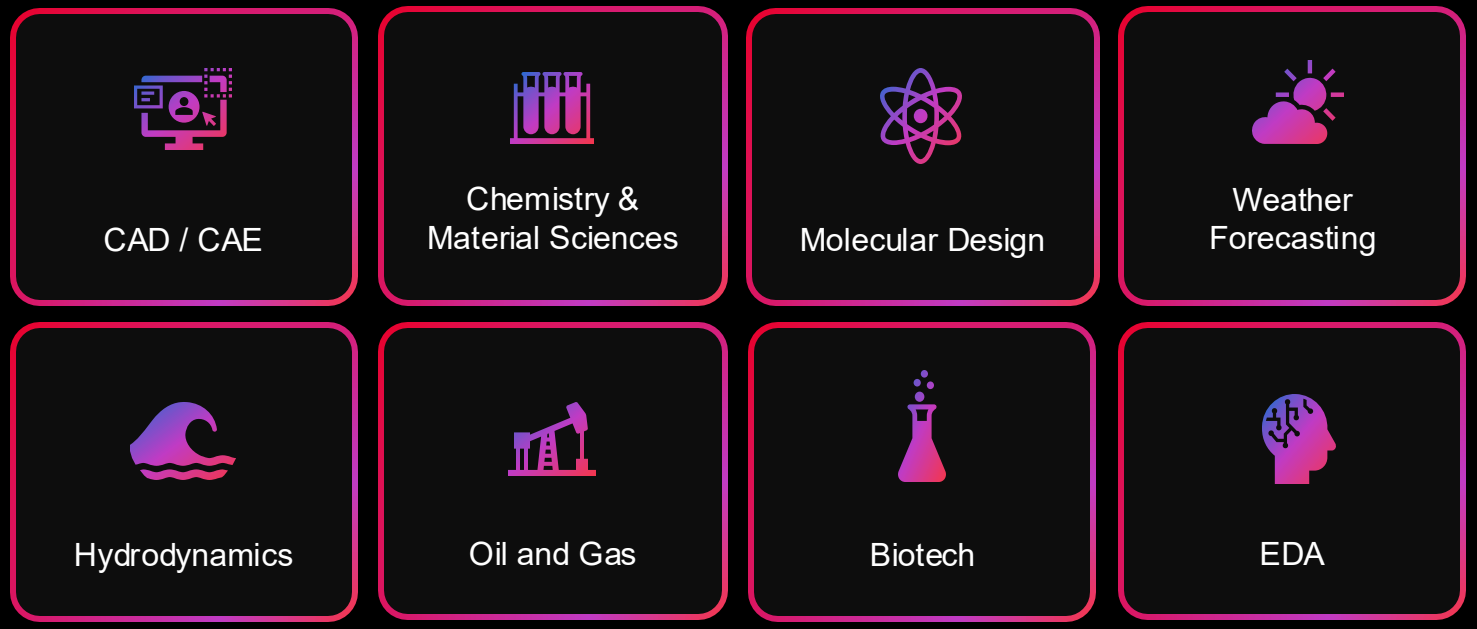
KEYSIGHT

AI DATA CENTER SUMMIT

Interconnect Fabrics & SDN: AI-Centric Evolution

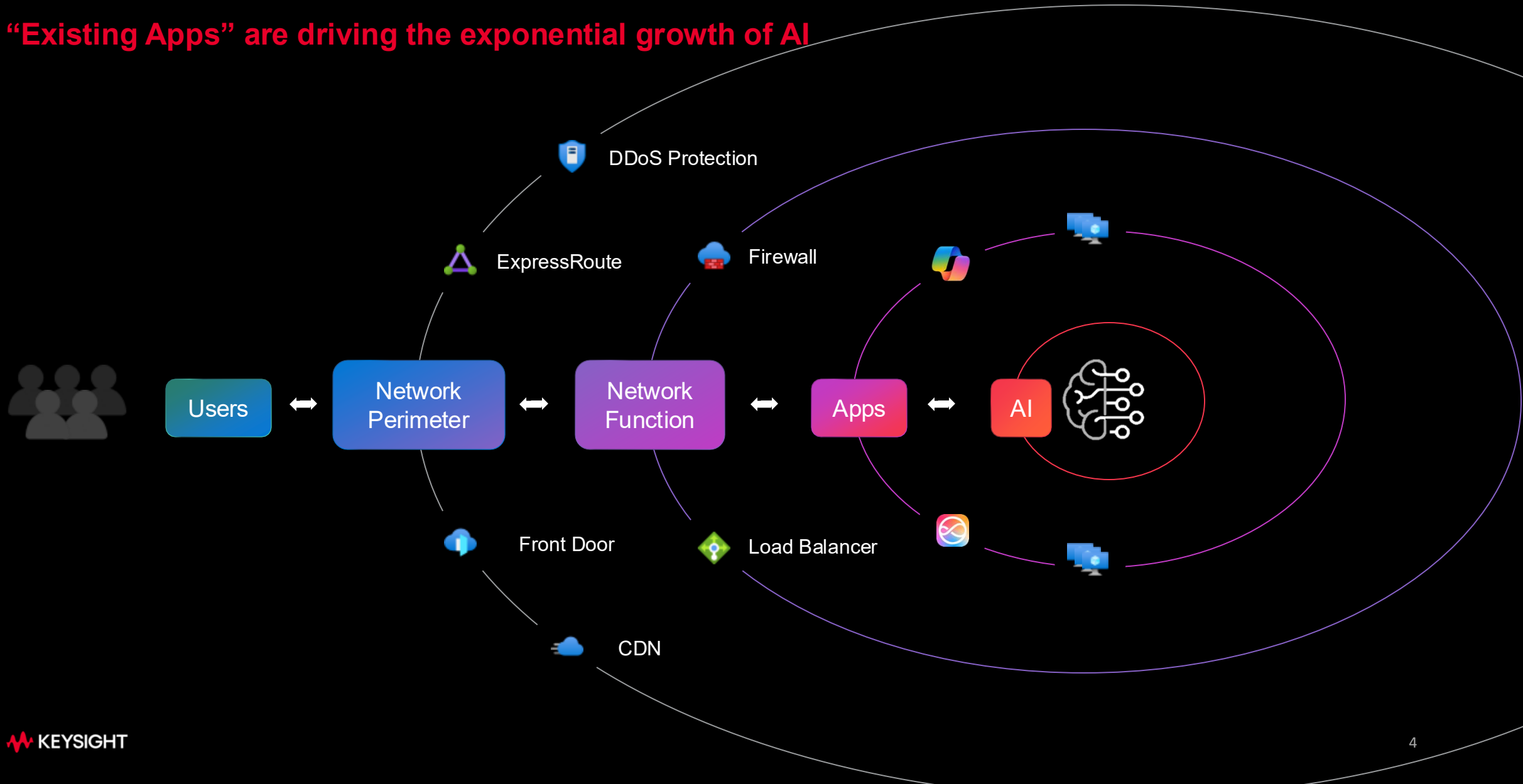
Deepak Bansal
Corporate Vice President – Microsoft

AI Is a Game Changer for More and More Applications



AI Is the Heart of a Larger Cloud System

"Existing Apps" are driving the exponential growth of AI



AI Cloud Providers Must Provide

- **Private, secure, and reliable infrastructure** to connect users to AI-enabled apps.
- **Flexibility** to match the right AI technologies to already established or newly created apps.
- **Ability to scale** with insatiable demand — regionally and globally.

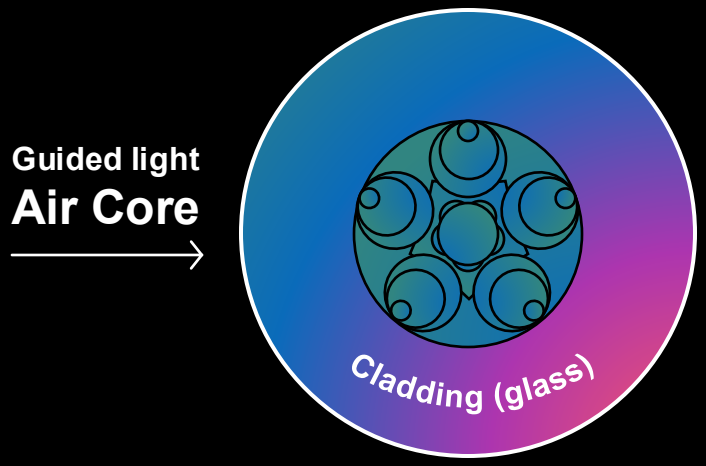
Requirements

- **High bandwidth** and **reliable networks.**
- **Low latency** to connect users to AI-enabled apps.
- **Security** to safeguard data and identity.

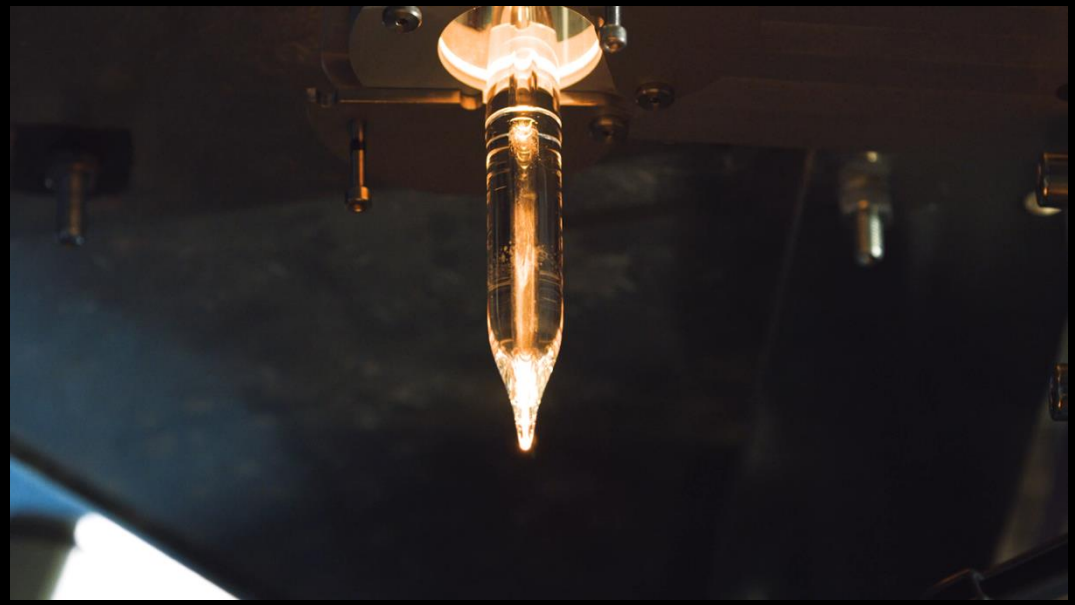
Hollow Core Fiber (HFC)

47% improvement in speed

Innovative Hollow Core Fiber (HCF)

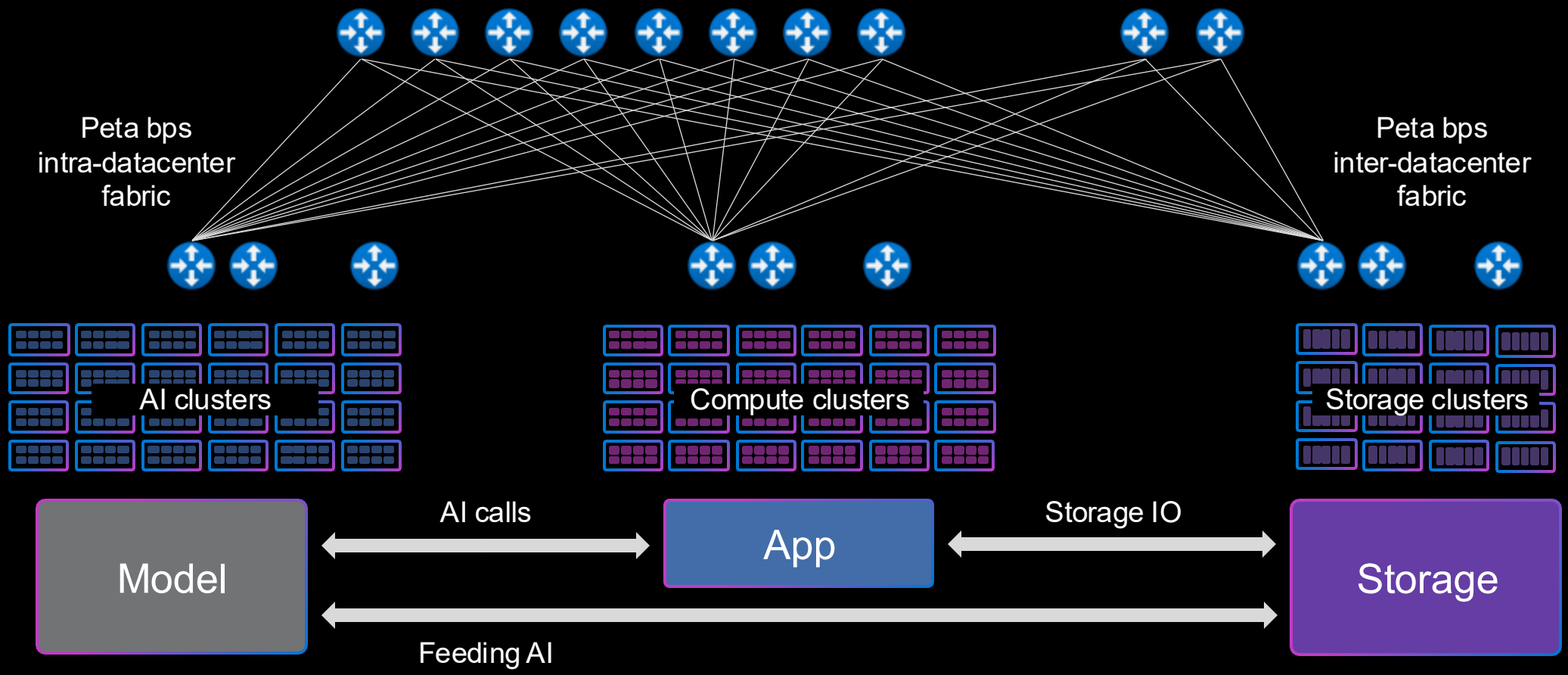


1.5X Faster



HCF in fab facility

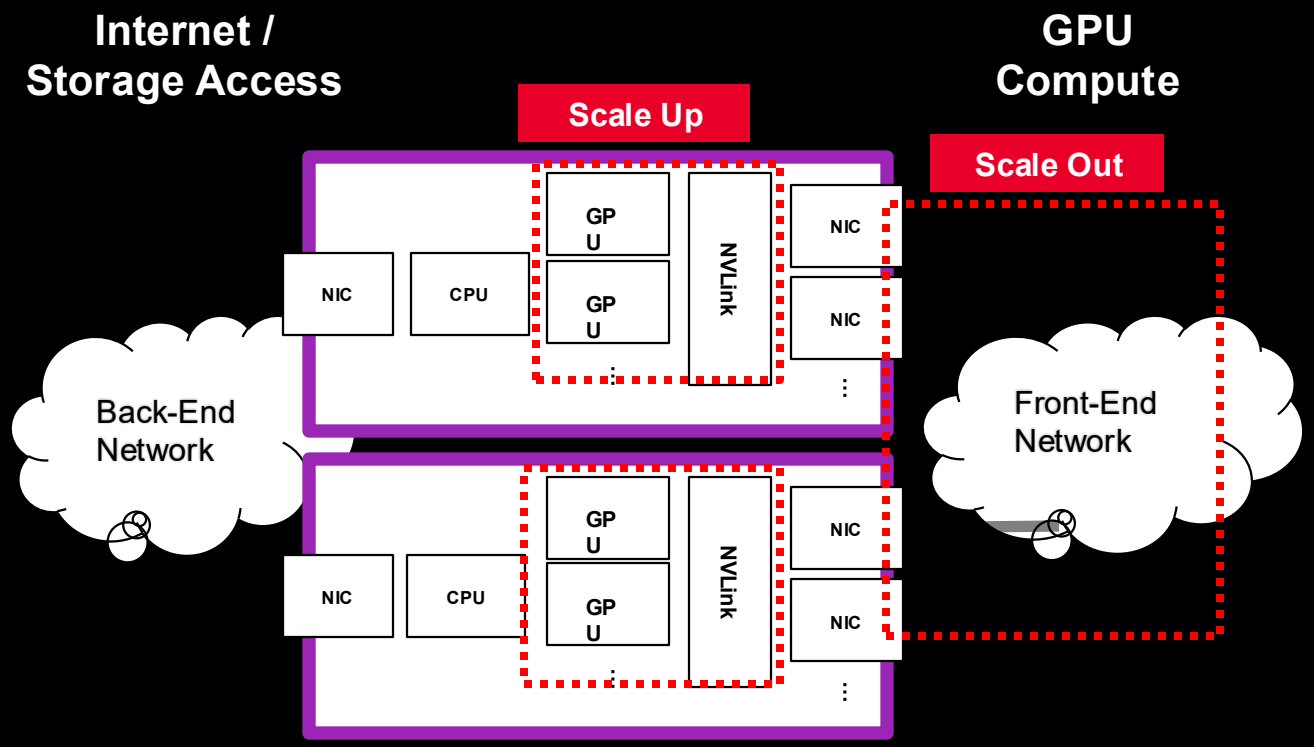
Massive Bandwidth Required to Interconnect Storage, Apps, and AI



AI Back-End Network in the Cloud

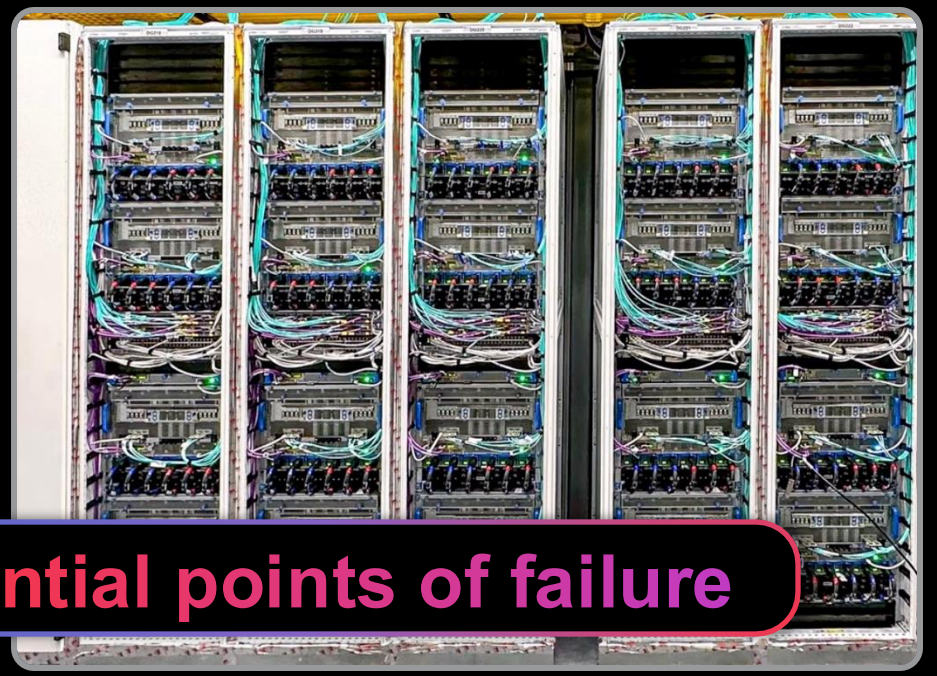
Raising the bar for hyperscale data center networks

- Long-lasting and large flows of training data
- Large bursts of data sent synchronously
- Long training time demands reliable networks
- Retries of failed jobs increased costs → efficient traffic management, monitoring, and visibility
- AI applications need fast processing and response → lossless traffic with low latency
- Traffic using RoCEv2 has low entropy for ECMP → traffic engineering technology for AI back-end network



Some Clusters Now Contain More than 100k GPUs

Reliable infrastructure is key to avoid re-runs



100s of 1000s of potential points of failure

Instability from inadequate cooling addressed with OCP chassis and rack design

Defect-free builds achieved with high levels of automation before commissioning

Ethernet / InfiniBand interconnect addressed with defect detection and streaming telemetry

Maintaining Network Performance and Uptime

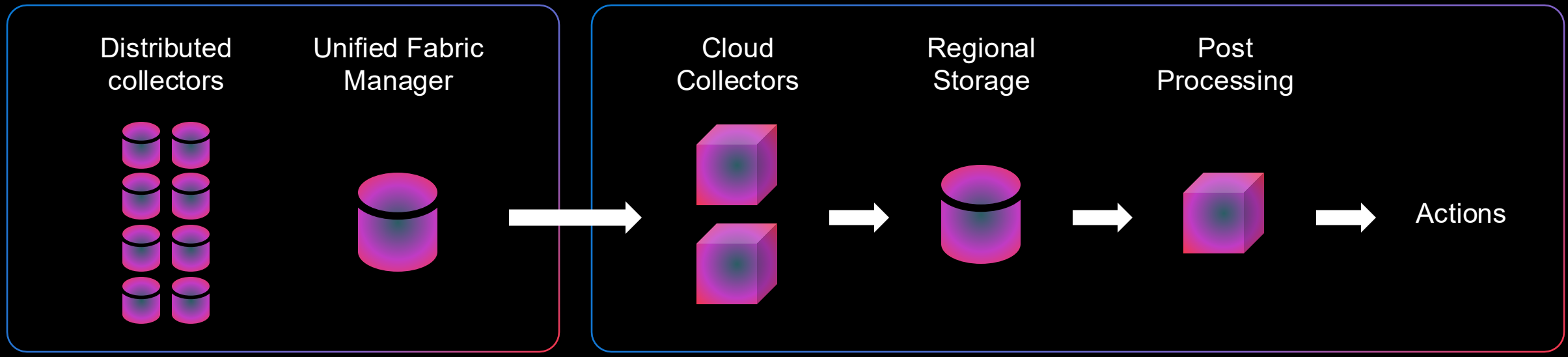
Huge deployments require automated health monitoring, ticketing, and mitigation

Back-End Networking

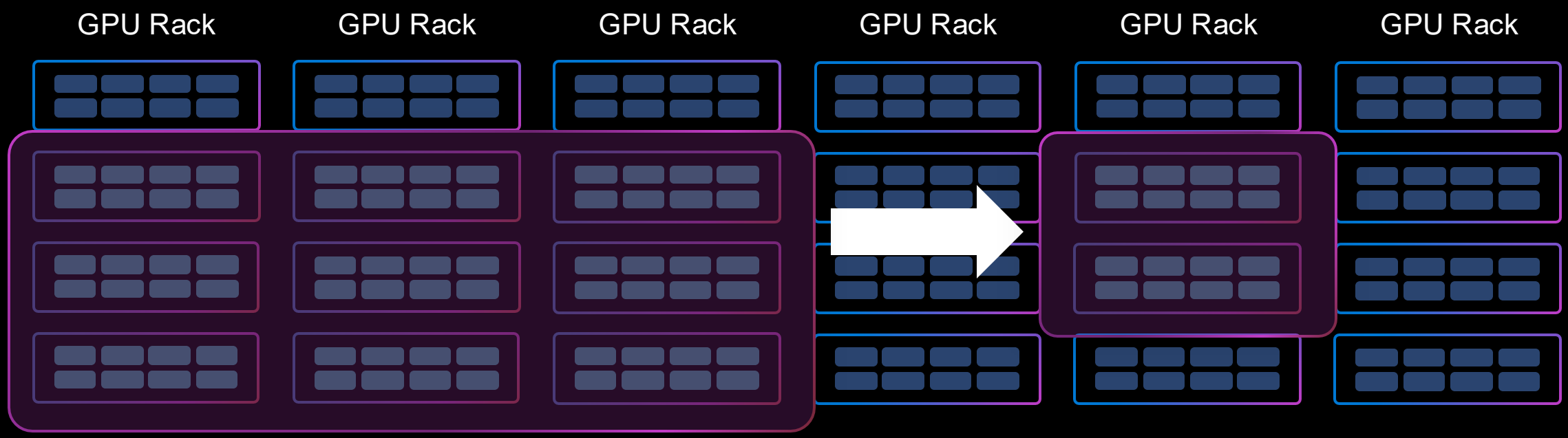
- Performance and defect streaming

Cloud

- AI-based monitoring and defect analysis
- Automated mitigation, ticketing, and resolution



Training and Inference Have Different Performance Requirements



Scale-out arbitrary amounts of AI GPU VMs to train your model

Batch, share, and overlay jobs to optimize cost vs. performance

Scale-in AI tech for the lighter needs of inferencing

Regional or Global Presence Is Key for Some Customers

60+ AI regions

275k+ Miles of fiber

2Pbps+ WAN capacity

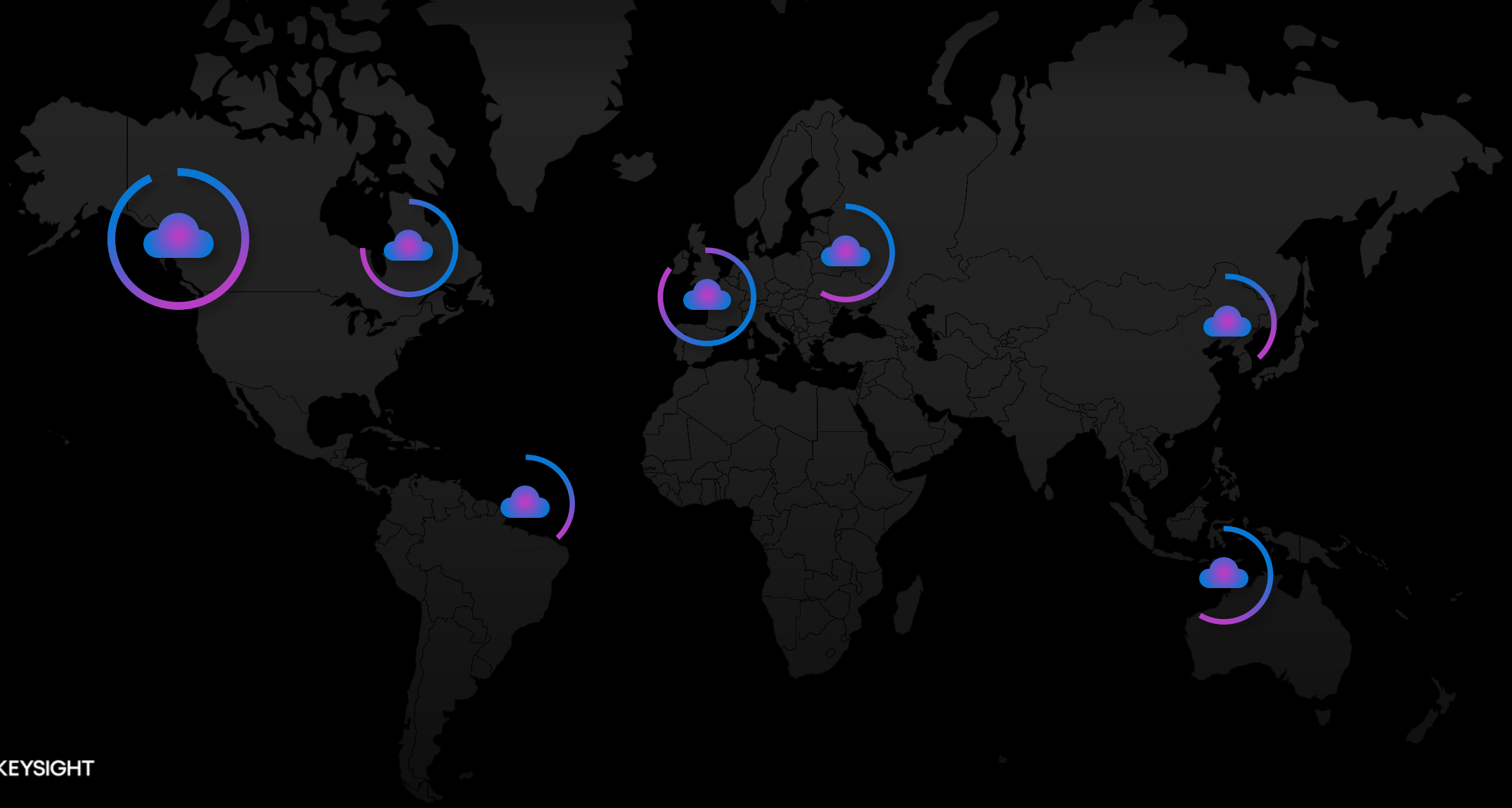
200T Peering capacity

40k+ Peering connections

KEYSIGHT

Global Schedulers Offer AI Everywhere You Need It

Transparent pre-empt, scale up, scale down, run



AI Appliance: Integrating Distributed GPUs to Azure

Use case: Secure 1.6 Tbps+ to storage over WAN

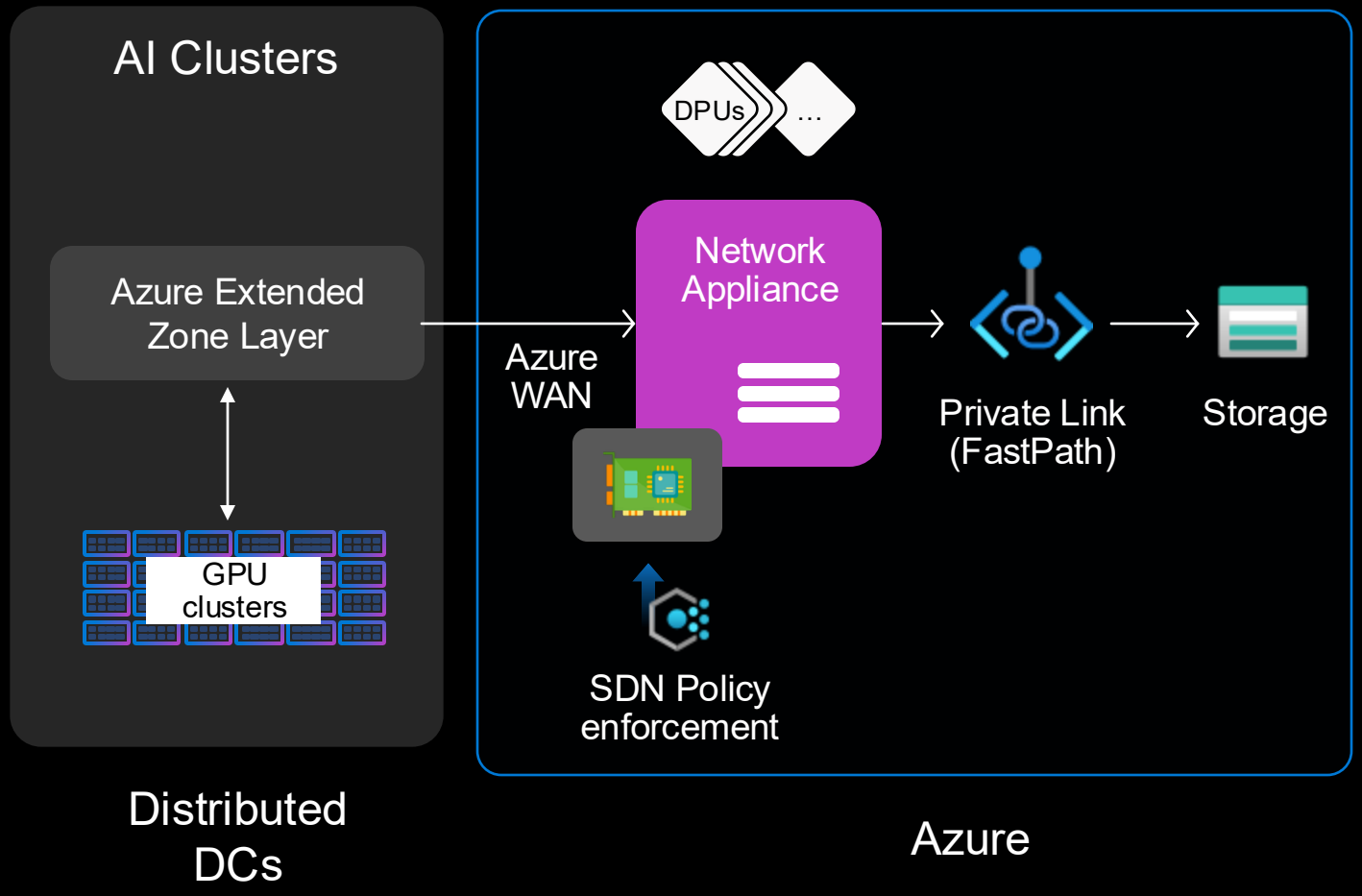
Secure access to Azure storage
- FastPath private link to Azure storage over WAN

Extreme network scale / performance
- 12.8 Tbps+

Horizontally scale DPUs
- 12.8 Tbps+ via ECMP

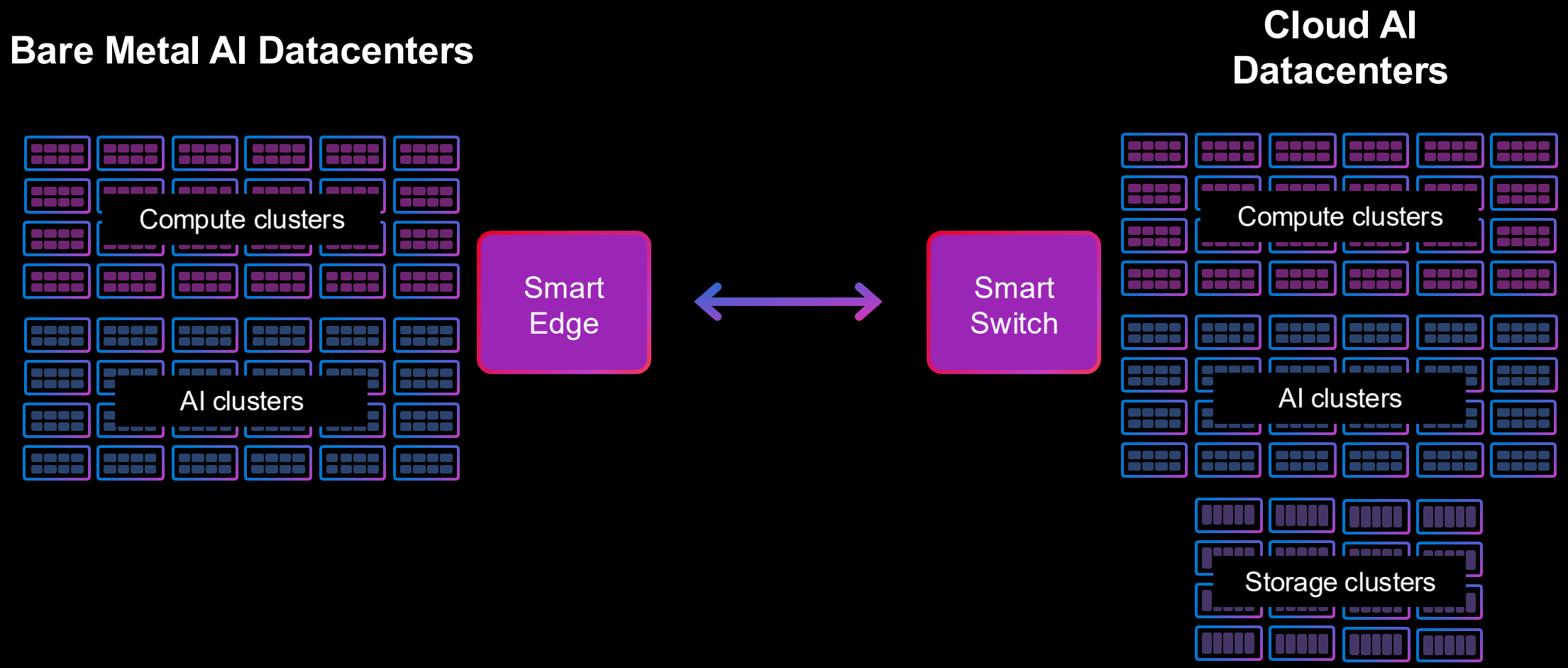
Lower training time
- Model load times significantly lowered training checkpoint time reductions

Increased efficiencies
- Run Azure services efficiently



Partners Are Essential to Keep Up with Demand

DASH-based SmartSwitch allows for bare-metal virtualization

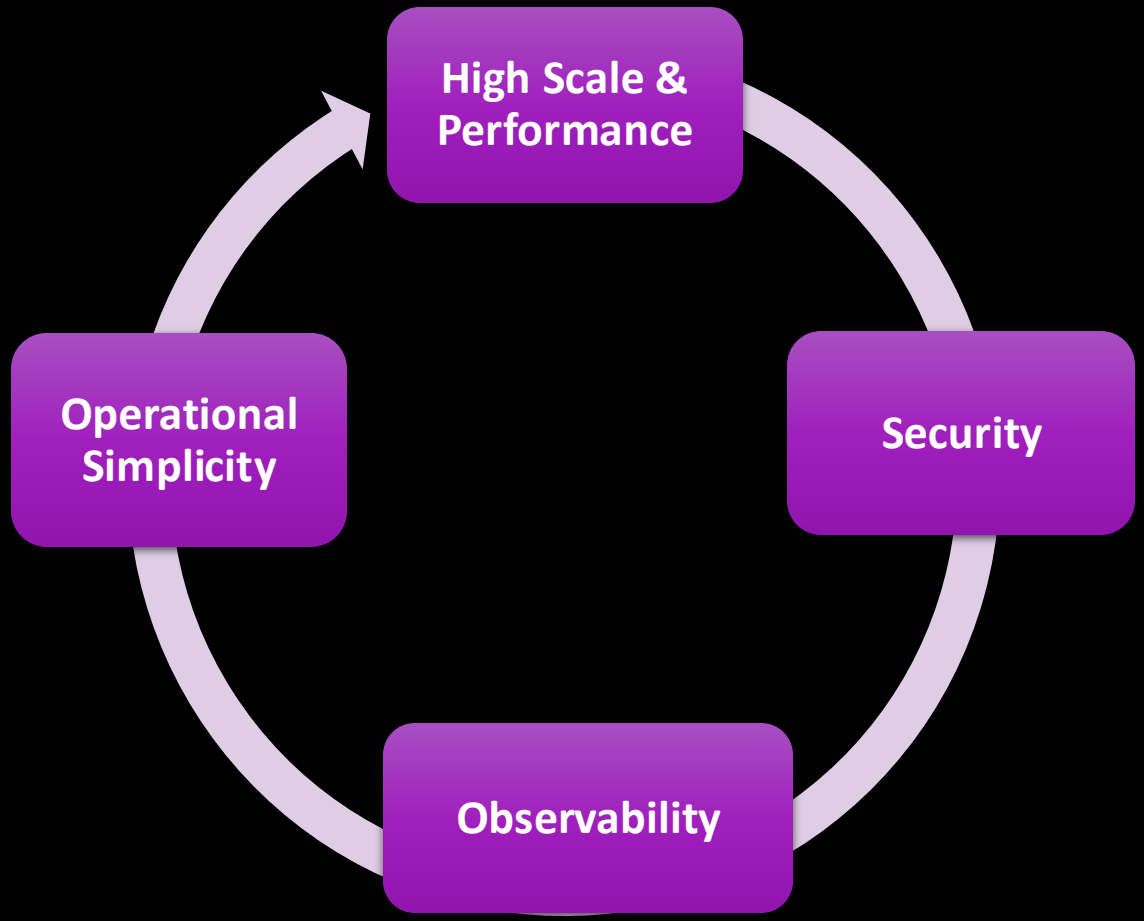


Application Networking for AI

AI Workloads

- Training
- Inference
- AI Apps & Edge AI
- Data Pipelines

Networking Demands



Azure Container Networking Solutions

High Scale & Performance:
Azure-managed container networking interface (CNI)

Security:
Advanced L3-L7 Policies and encryption

Observability:
Advanced Network analysis at workload level

Operational Simplicity:
Multi-cluster growth and Azure support

AI Appliance: Integrating Distributed GPUs to Azure

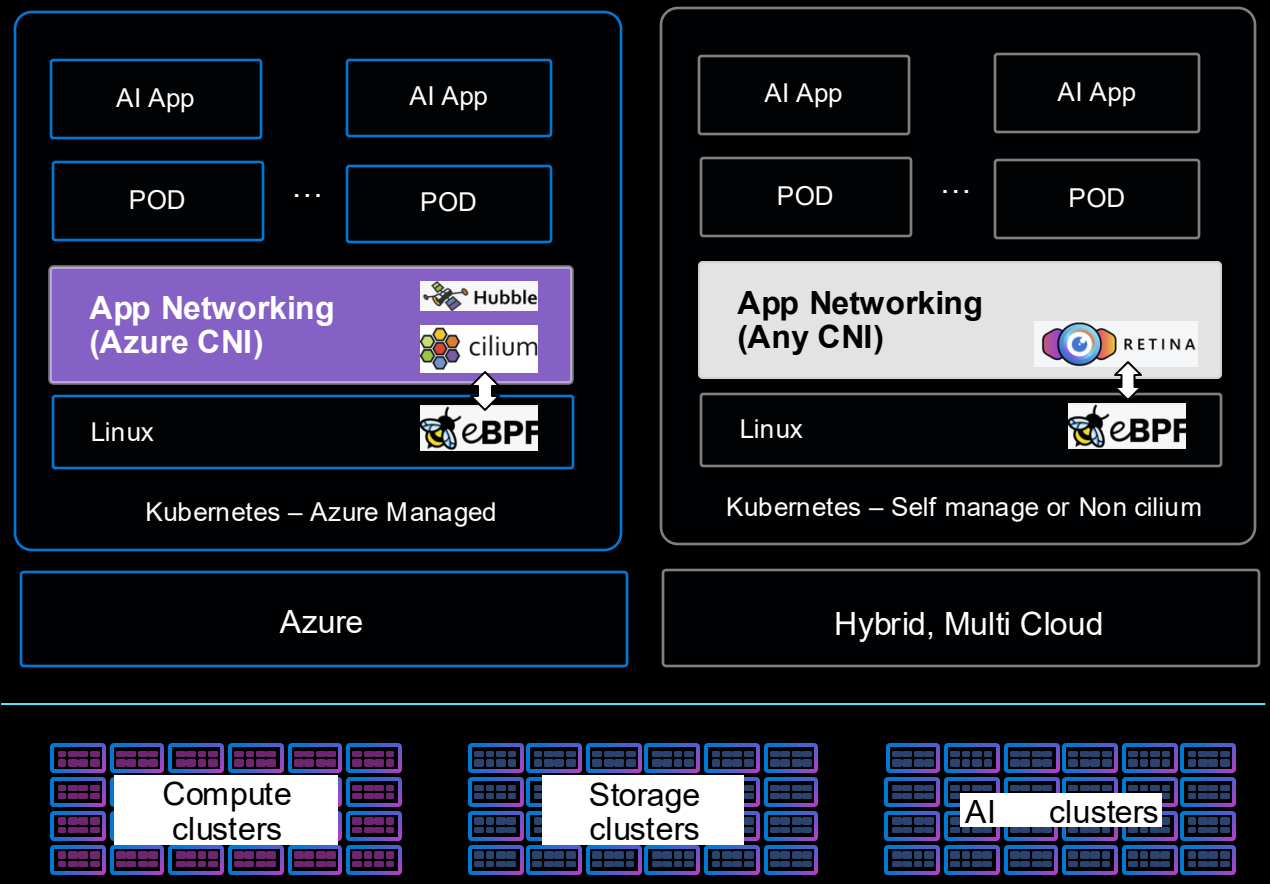
Use case: Secure 1.6 Tbps+ to storage over WAN

Azure CNI:

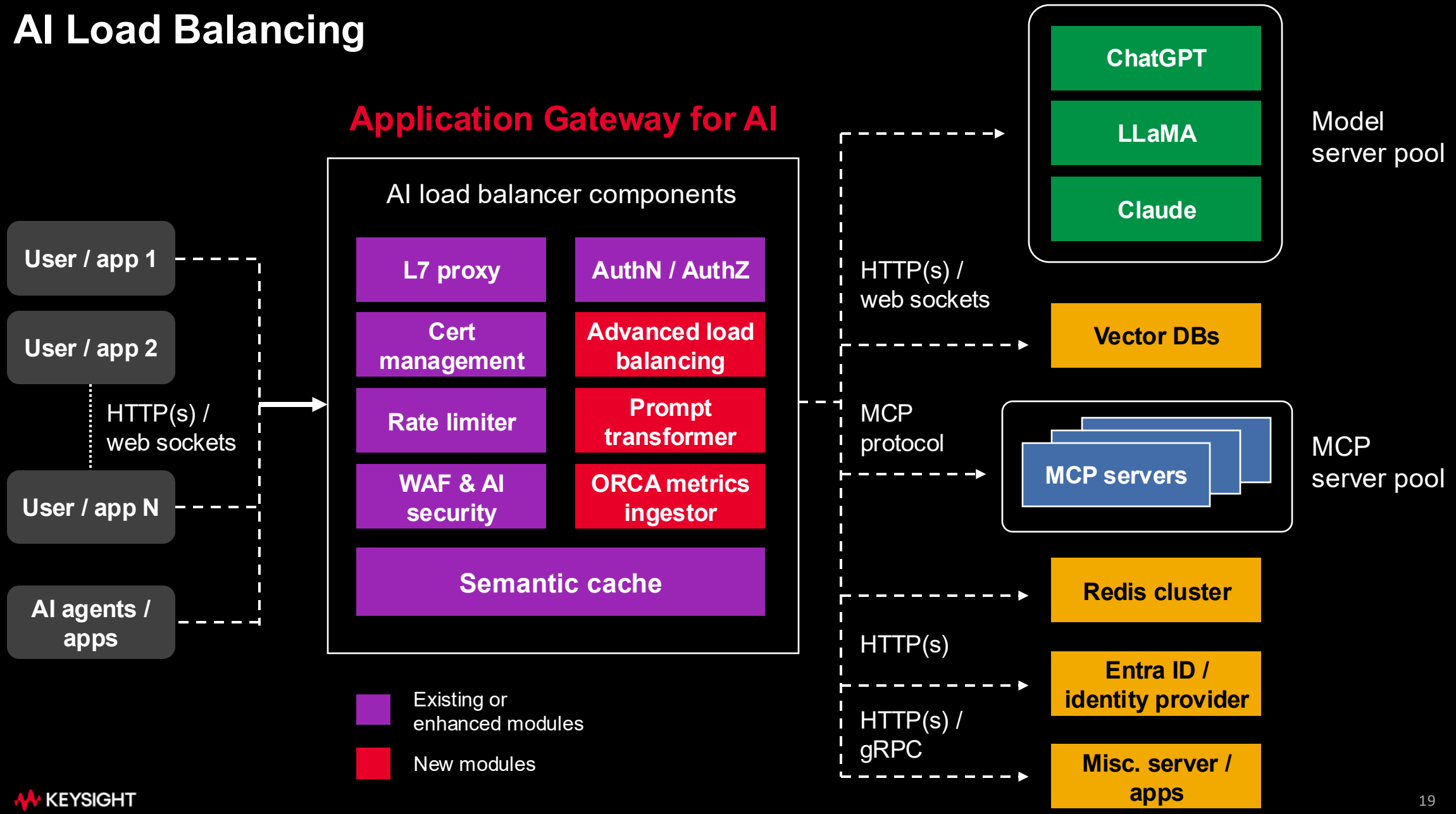
- Azure CNI powered by Cilium (built on eBPF Linux kernel)
- High-performance, scalable data plane for AI workloads
- Azure Integrations for Flexible IPAM and VNET routing options scaling to 1M+ IPs
- Granular network security policies L4-L7, Filtering on FQDNs, Encryption
- Observability and AI insights with Pod and protocol metrics, flow logs

Any CNI:

- Retina (opensource) supports observability for many scenarios: customer self-managed, hybrid and multi cloud.

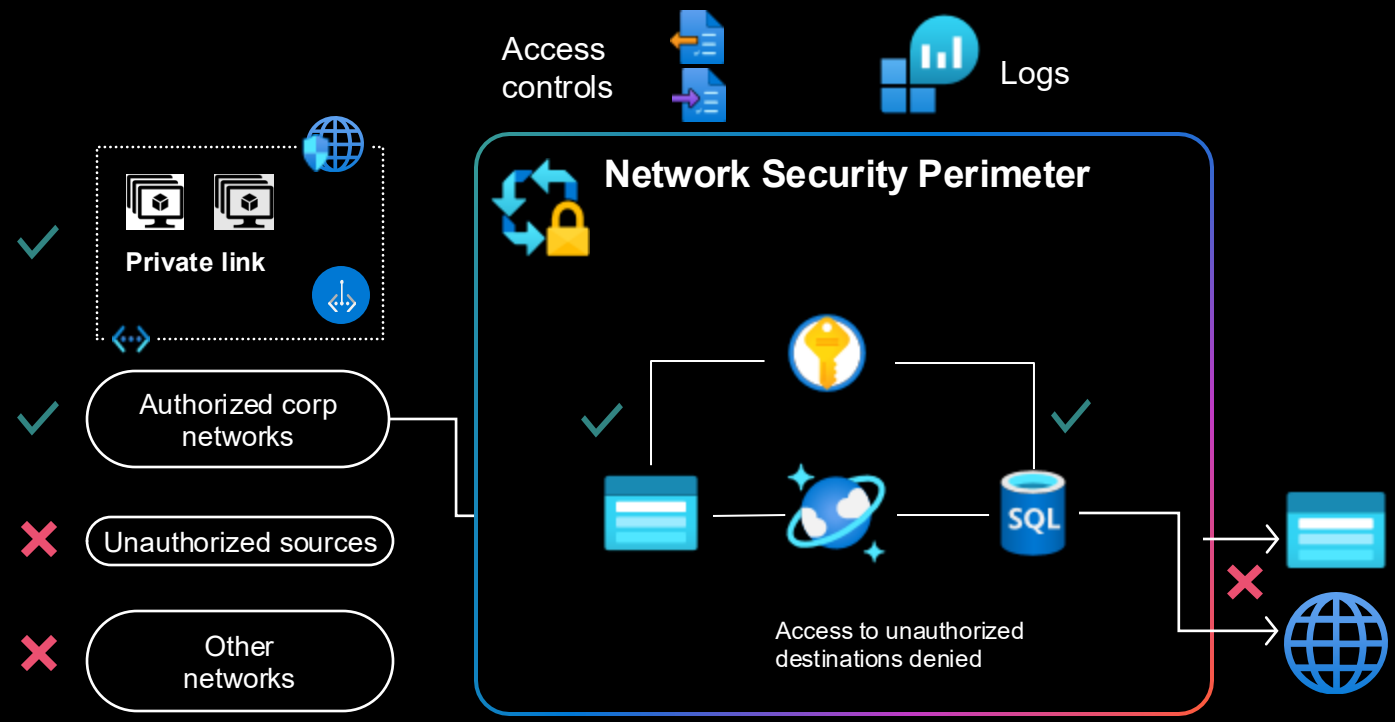


AI Load Balancing



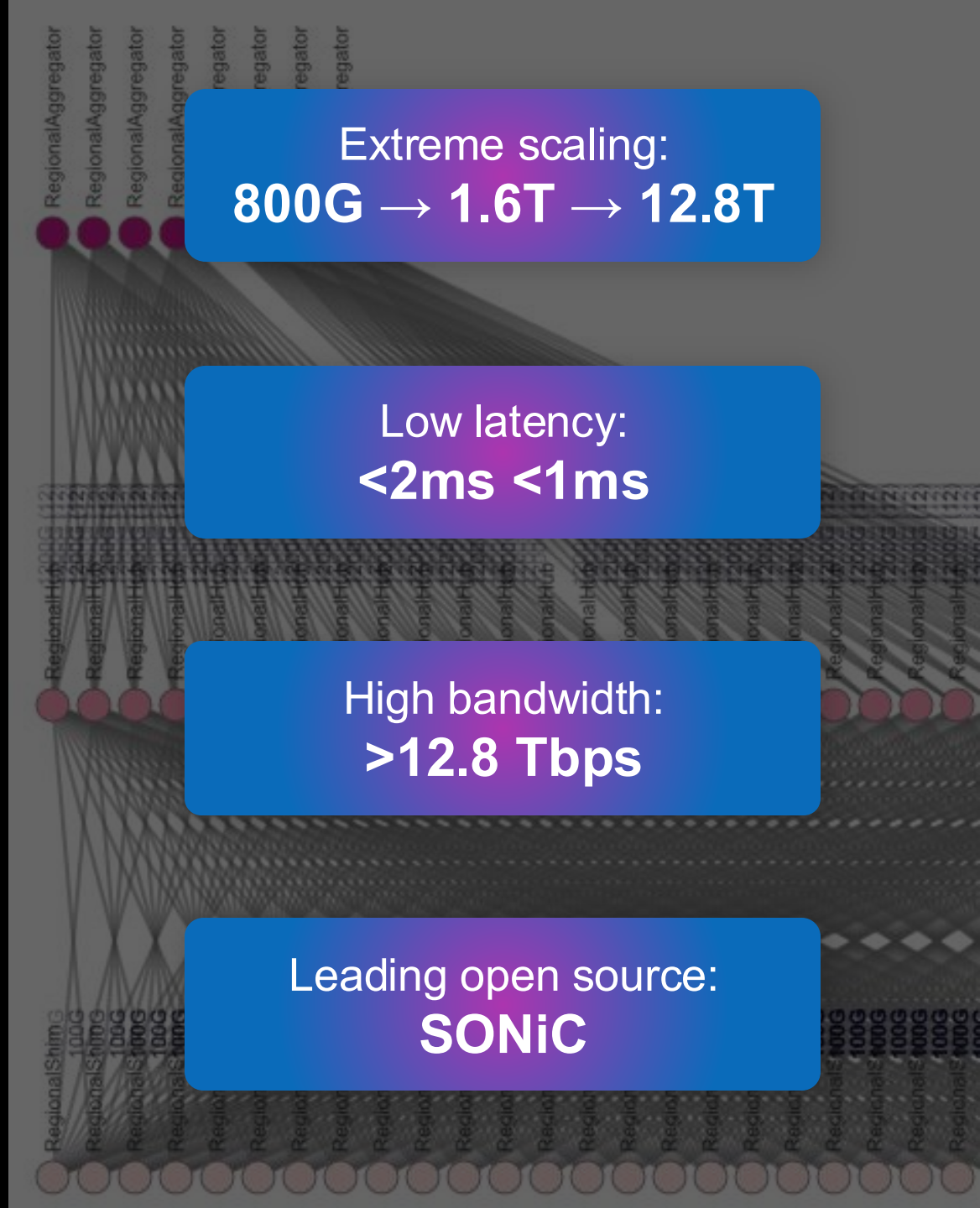
Security Perimeter

- Ensures data and models are isolated
- Unified firewall for multi-tenant resources
- Granular access control for PaaS to PaaS
- Egress controls for data exfil protection



Networking for AI

- **High-performance / security** network for AI
- For all types of accelerators: Nvidia, AMD, Maia
- Dedicated, **low-latency** WAN for AI
- First large-scale co-operative transport (host + network) for **fast failover and load balance** via SRv6 on SONiC
- Millisecond-level, **high-frequency telemetry** for troubleshooting
- Bare metal host with **SDN policies**
- Secure and high-bandwidth connectivity via **private link** to storage for training



KEYSIGHT

100B+ near-term
AI infrastructure investments
enabling customers to

Build the future of their business on AI

This information is subject to change without notice. © Keysight Technologies, 2025, Published in USA, September 30, 2025, 7125-1079.EN