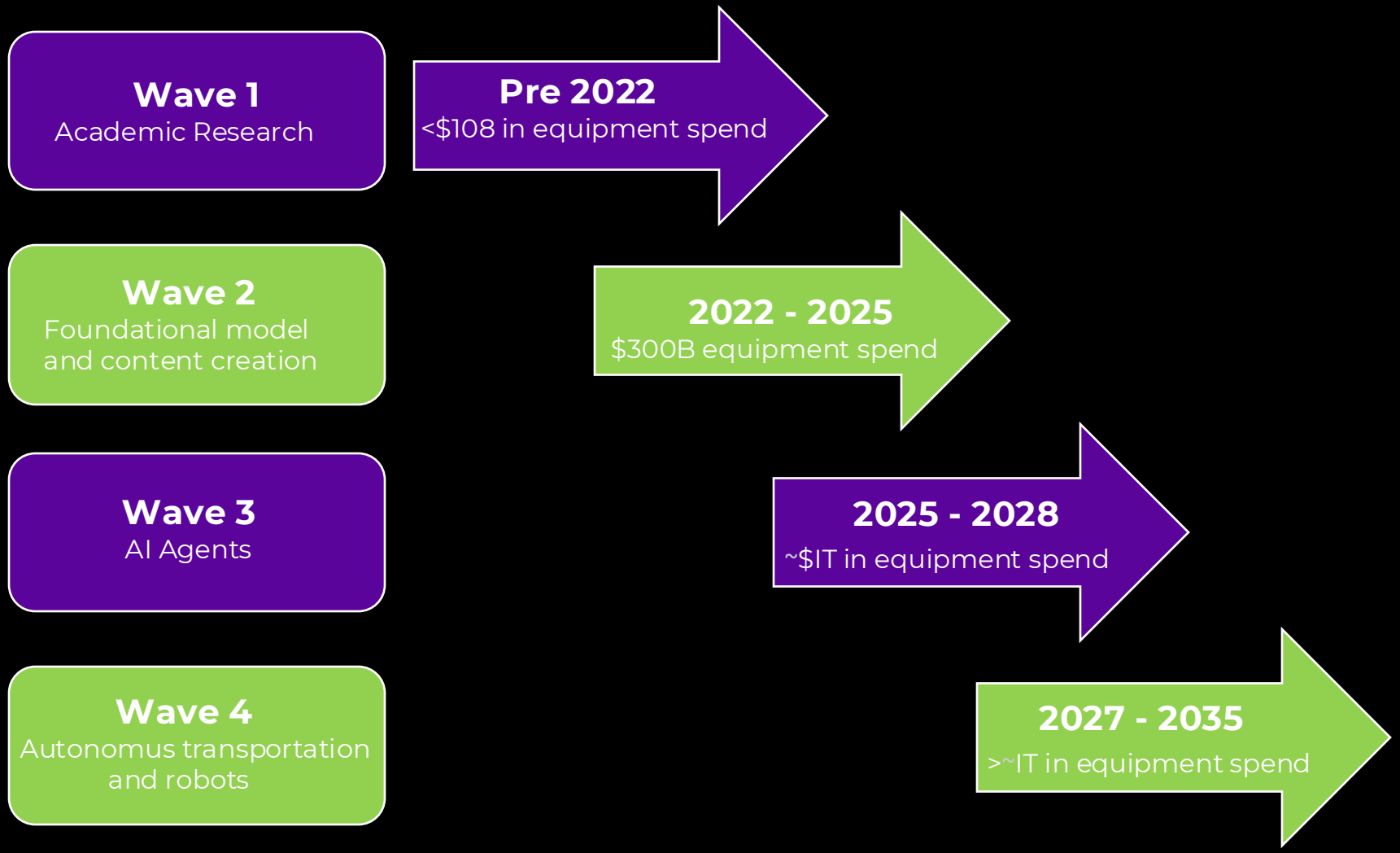
The Myth around AI networks & AI for networks

Mikael Holmberg
Distinguished Engineer – Member of the Office of the CTO

AI Waves 2022 - 2035



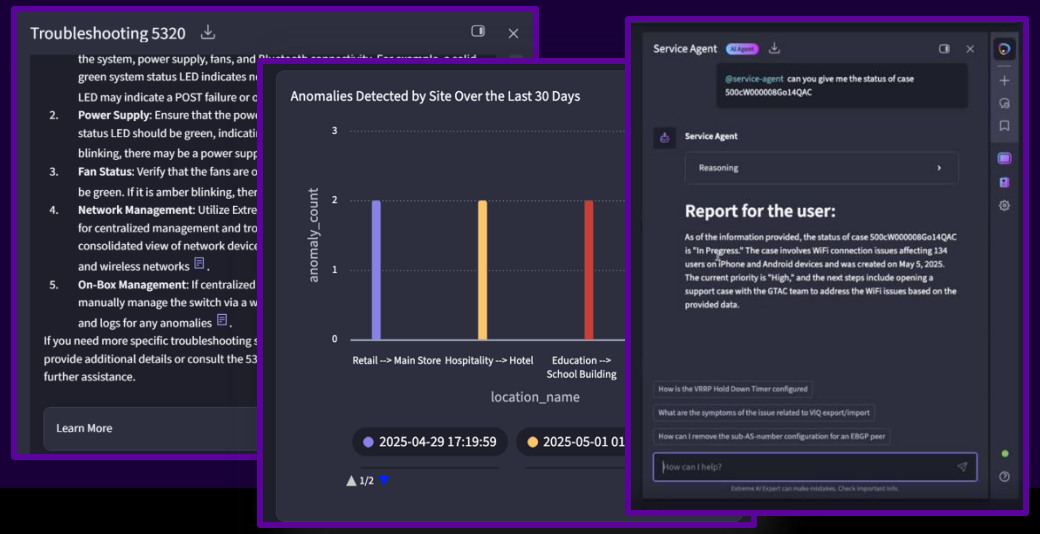
Soure: 650 Group

Modern Networks Need More than AIOps

AI Expert
Conversational

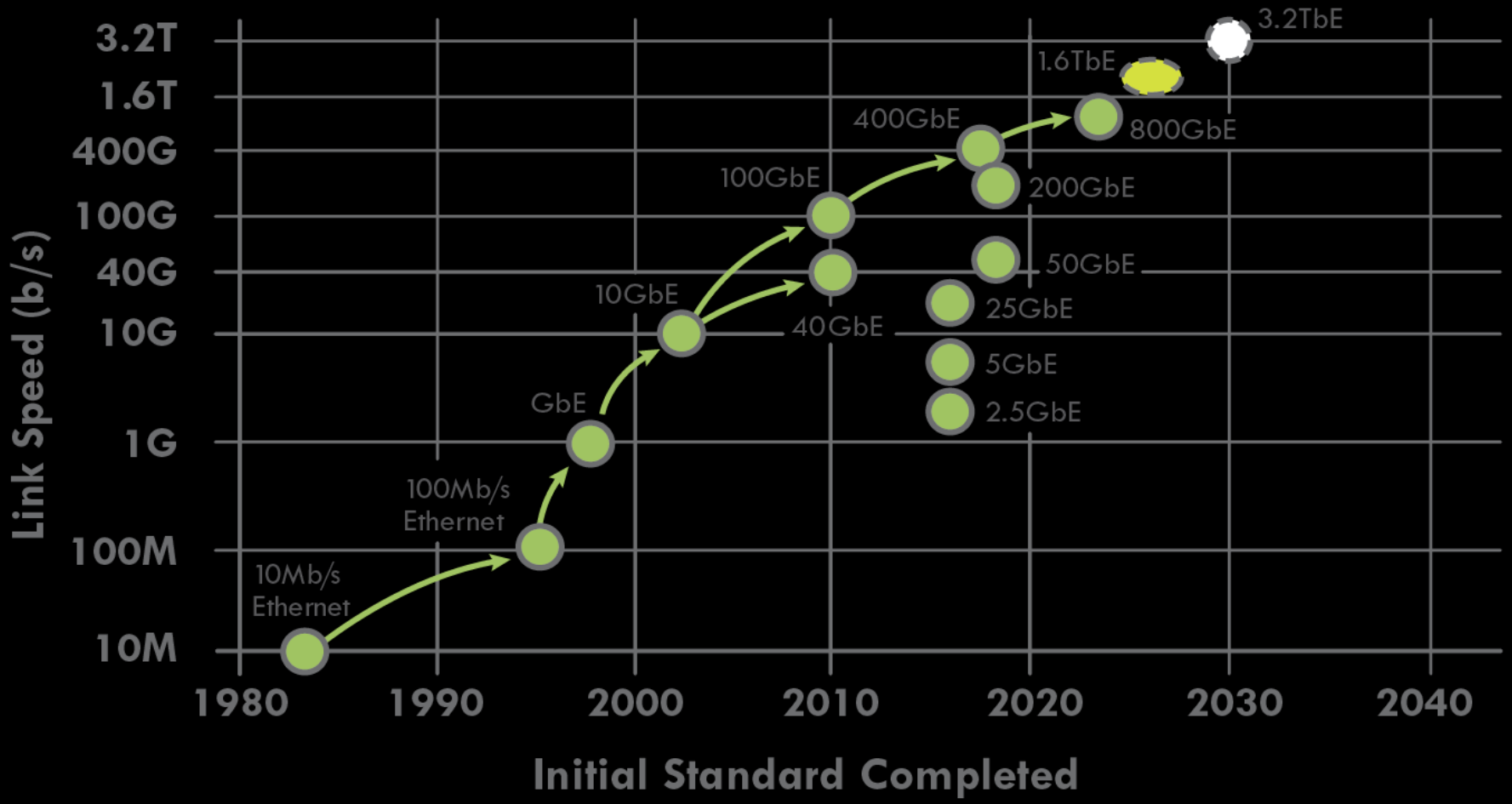
AI Canvas
Interactive, Multi-Modal

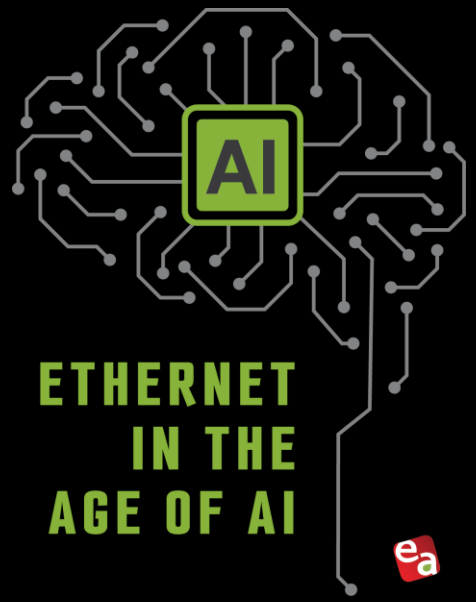
AI Agents
Autonomous with HITL



Conceptual

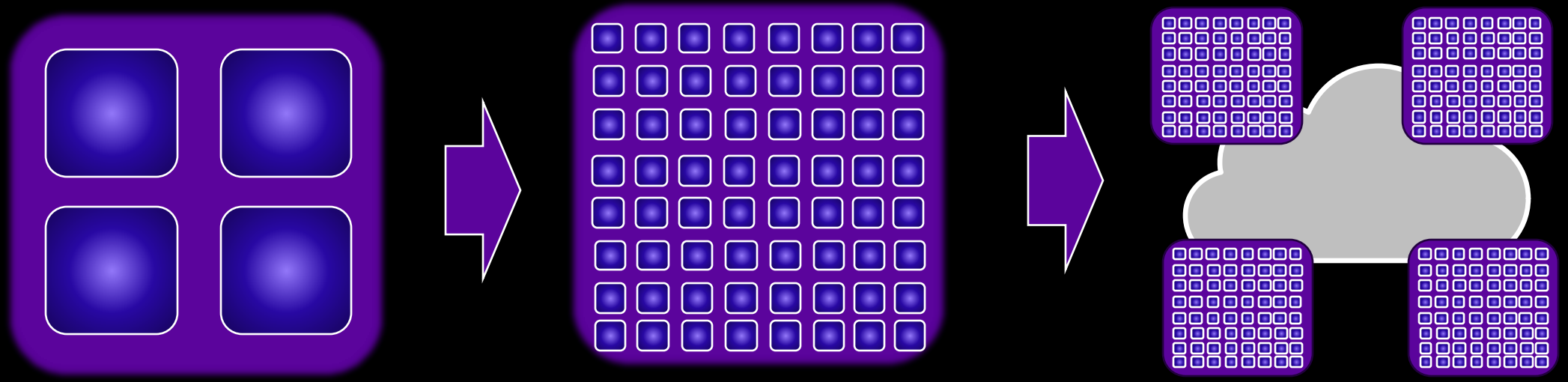
Ethernet Speeds Evolution





Source: Ethernet Alliance

Acceleration of Computes for AI Workloads



CPU

Optimized for serial tasks

GPU

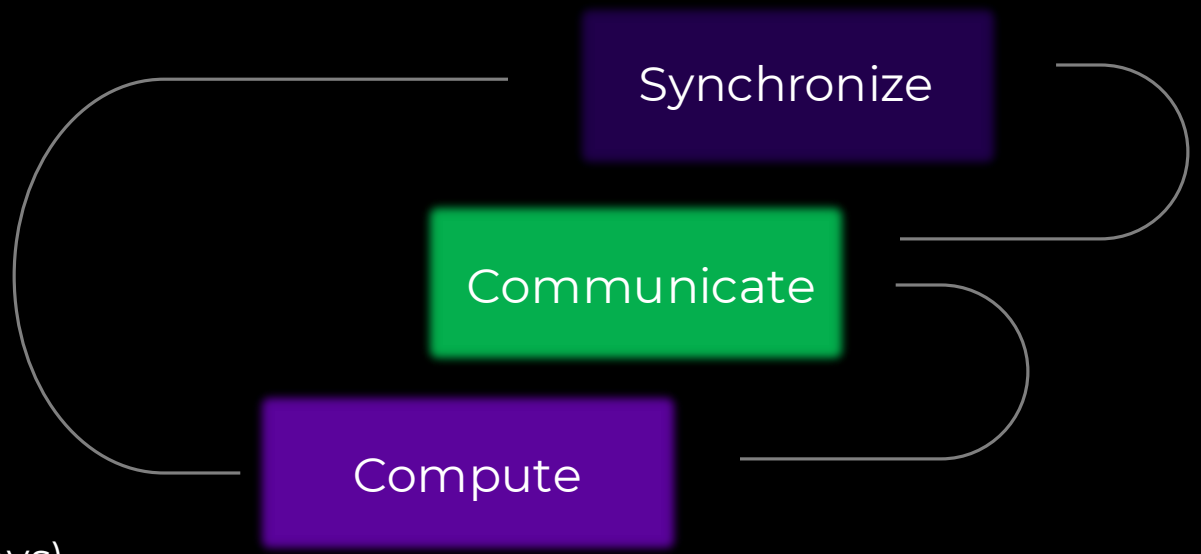
Optimized for parallel tasks

GPU Clusters

Scale-up network for AI workloads

AI Workload - Networking Uniqueness

- ✓ Fewer flows (low entropy)
- ✓ High bandwidth flows
- ✓ Synchronized and bursty traffic
- ✓ Links are saturated in micro-seconds (<<RTT)
- ✓ Training jobs run for long periods of time (hours, days)
- ✓ Tail latency impacts job completion time significantly

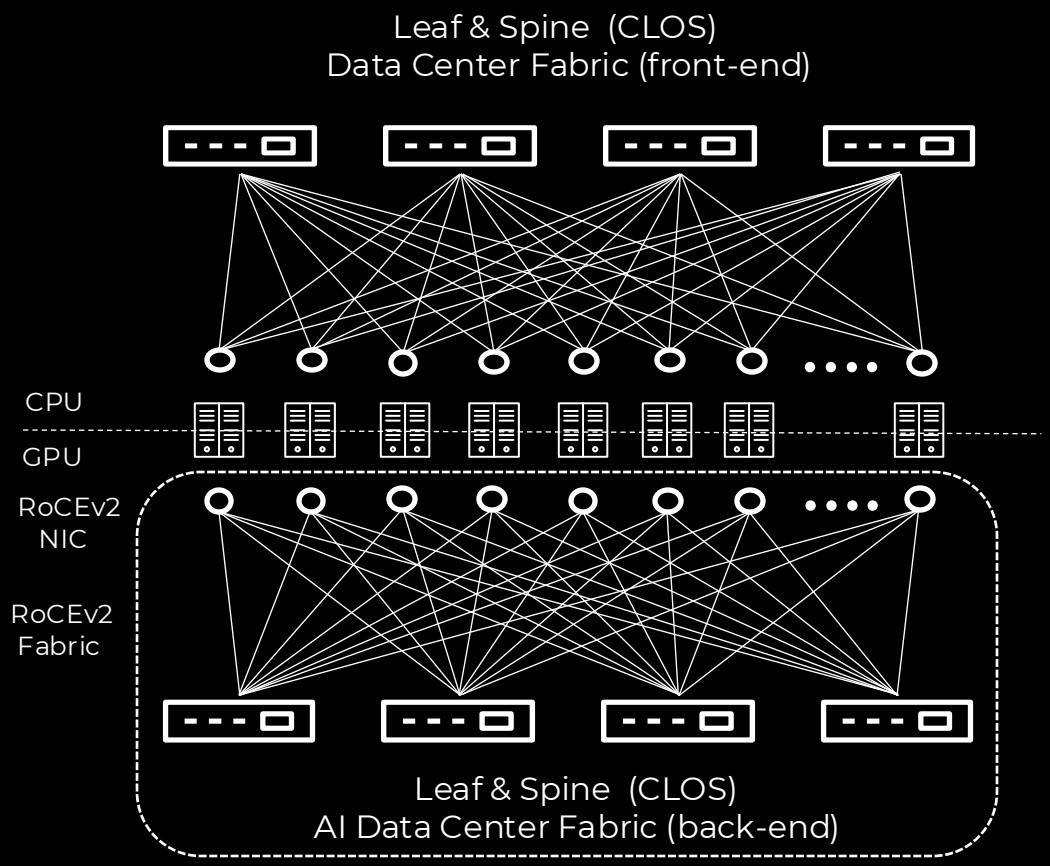


JCT (Job Completion Time)

AI Data Center Fabric Networks

AI workloads present new challenges to current Ethernet networks

- Need for higher scale,
- Higher bandwidth density
- Multi-pathing
- Fast reaction to congestion, and inter-dependency on the progress of individual flows
- AI workloads within compute clusters (GPU – GPU)



AI Workload Challenges

INCAST (IN)

Incast traffic patterns happens when multiple sources target the same destination.

OBLIVIOUS BULK SYNCHRONIZATION (OBS)

Computation steps are interleaved with global communication steps that often synchronize processes – Three dimensional parallellism in AI deep-learning 1. Number of processes 2. Duration of computation 3. Size of communication (end point).

LATENCY SENSITIVE (LS)

Some workloads are latency sensitive, could fall into the OBS category – complex, data-dependent, message chains forming critical performance paths in application.

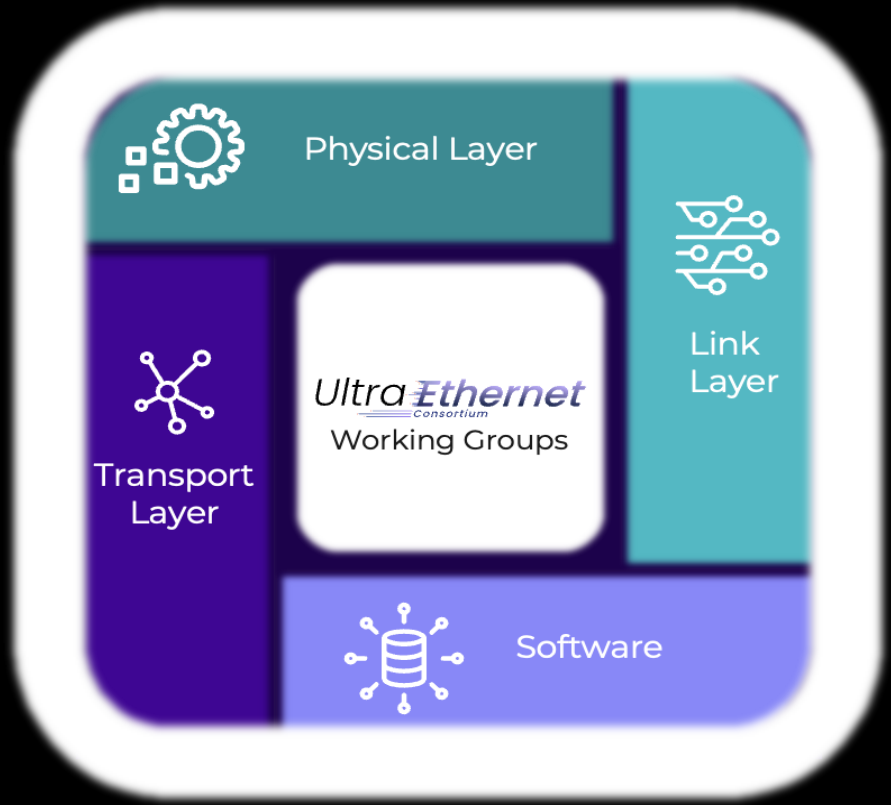
DEPLOYMENT CHARASTERISTICS

Confidential compute – all traffic must be encrypted on the wire.
Multitenancy - managing thousends of from single host.
Cost effective low-diameter topologies requires advanced load balancing and routing for high bandwidth requirements.

Ultra Ethernet Consortium – Working Groups

A group of vendors and operators have teamed up to form the Ultra Ethernet Consortium (UEC), as there are concerns that today's traditional network interconnects cannot provide the required performance, scale, and bandwidth to keep up with AI demands,

The consortium **aims to address these concerns by adding new capabilitie**s to the known and proven Ethernet technology specification, adding number of new features and capabilities.



Goal to develop specifications, APIs and source code to define protocols, interfaces and data structure.

AI Cluster Components

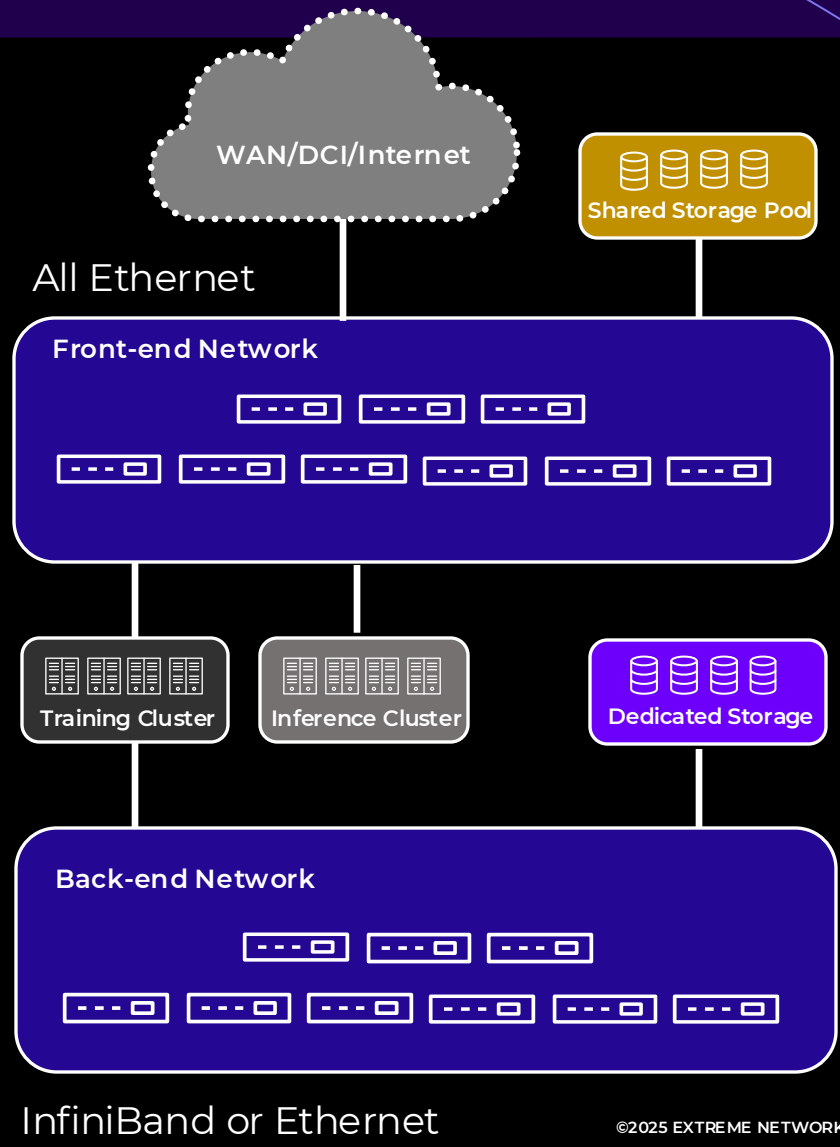
AI Data Center Cluster Networks

Front-end:

- Inference clusters use this network
- Shared storage pools
- Management network for training

Back-end:

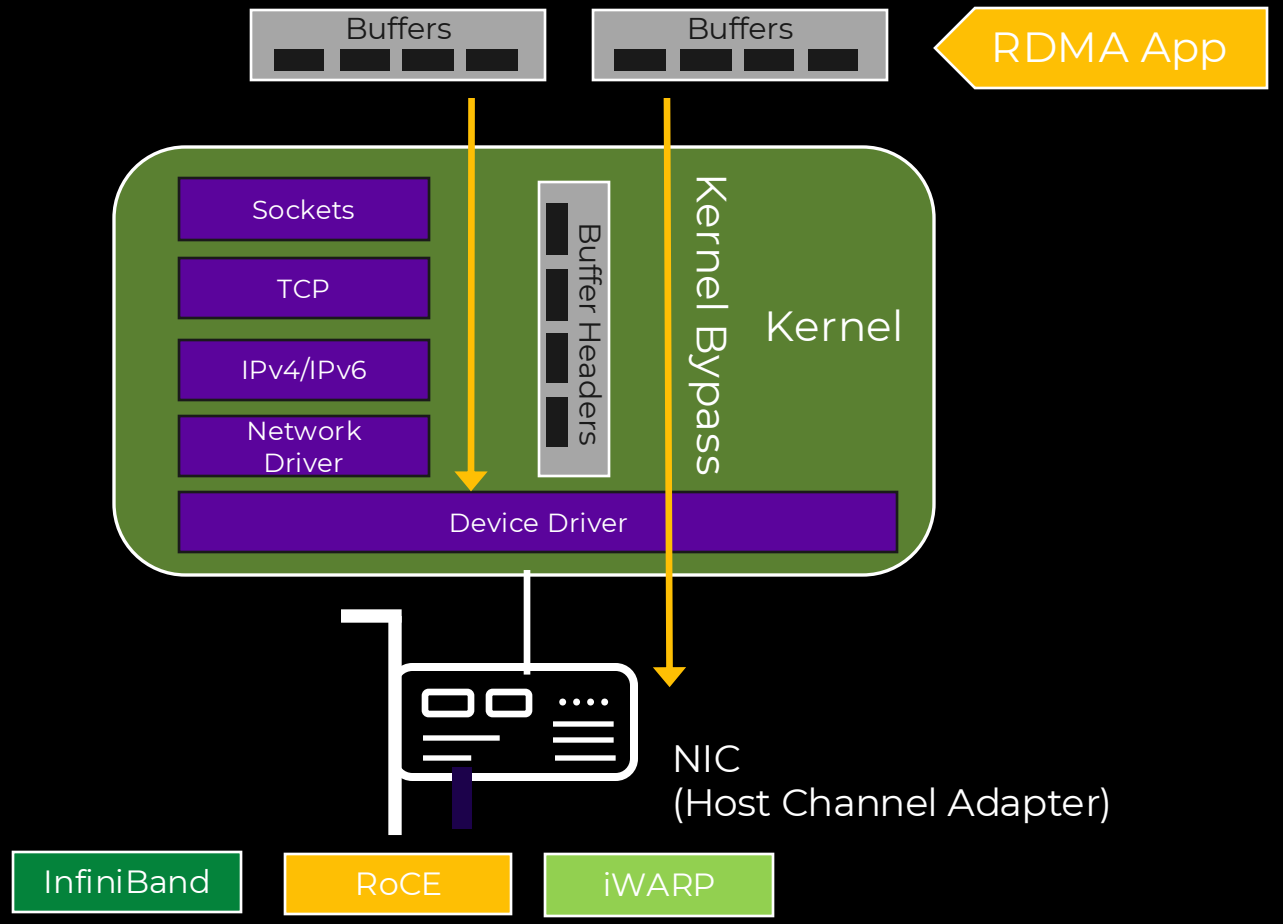
- GPU Compute Fabric
- Dedicated Storage Fabric
  - Could be converged with compute



RDMA (Remote Direct Memory Access)

Remote Direct Memory Access (RDMA) is an ultra-high-speed network memory access technology

- Allows programs to access the memory of remote compute extremely fast.
- With RDMA network access does not need to go through the OS kernel (sockets, TCP/IP, etc.),
- It would consume CPU time with kernel operations.
- RDMA bypasses OS kernel overheads and enables direct memory access to the Network Interface Card (NIC).

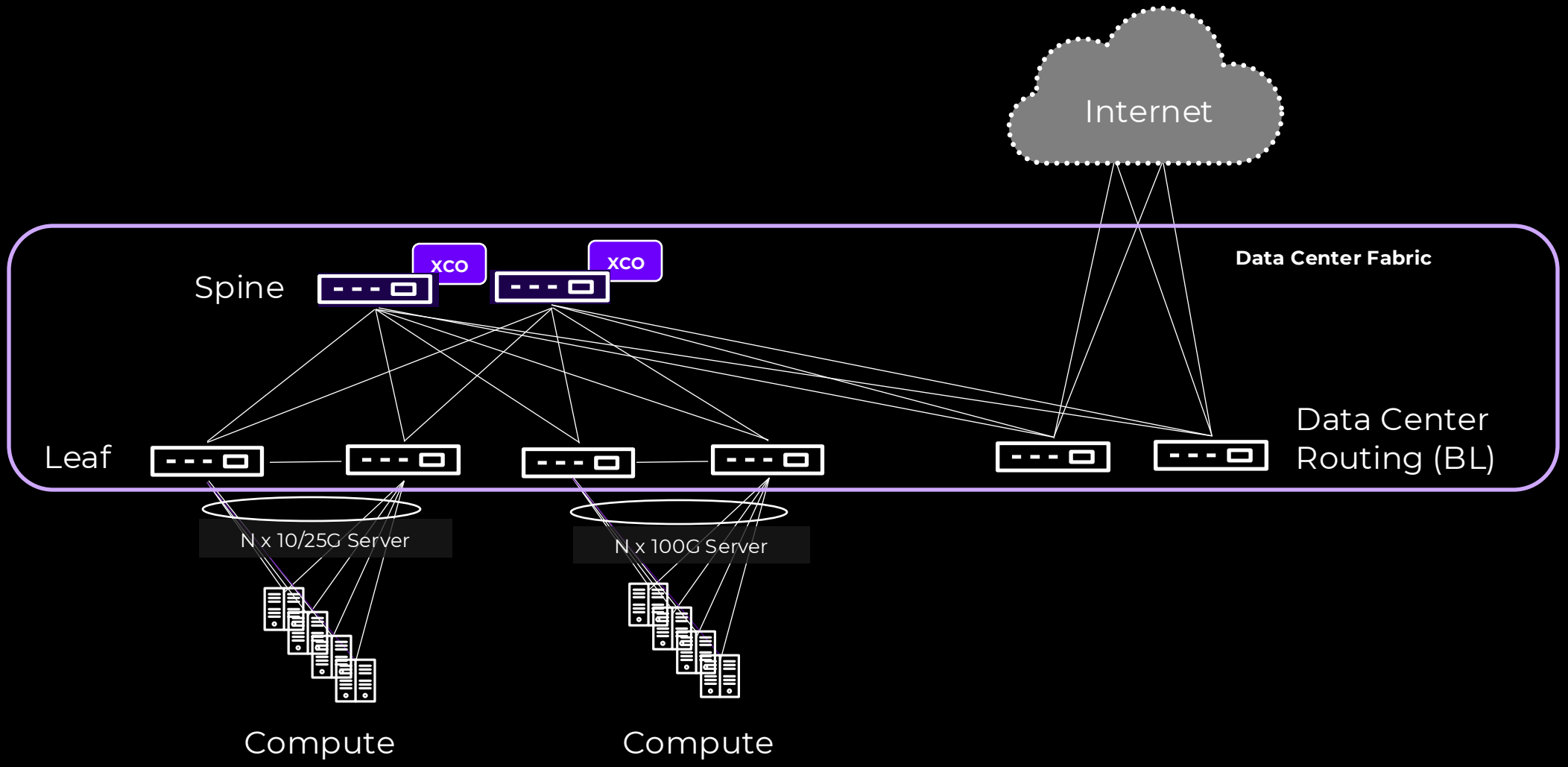


AI Data Center Technology Reference

- Compared to expensive network deployments like InfiniBand, RoCE is a relatively cheaper option, although it still cannot be considered inexpensive.

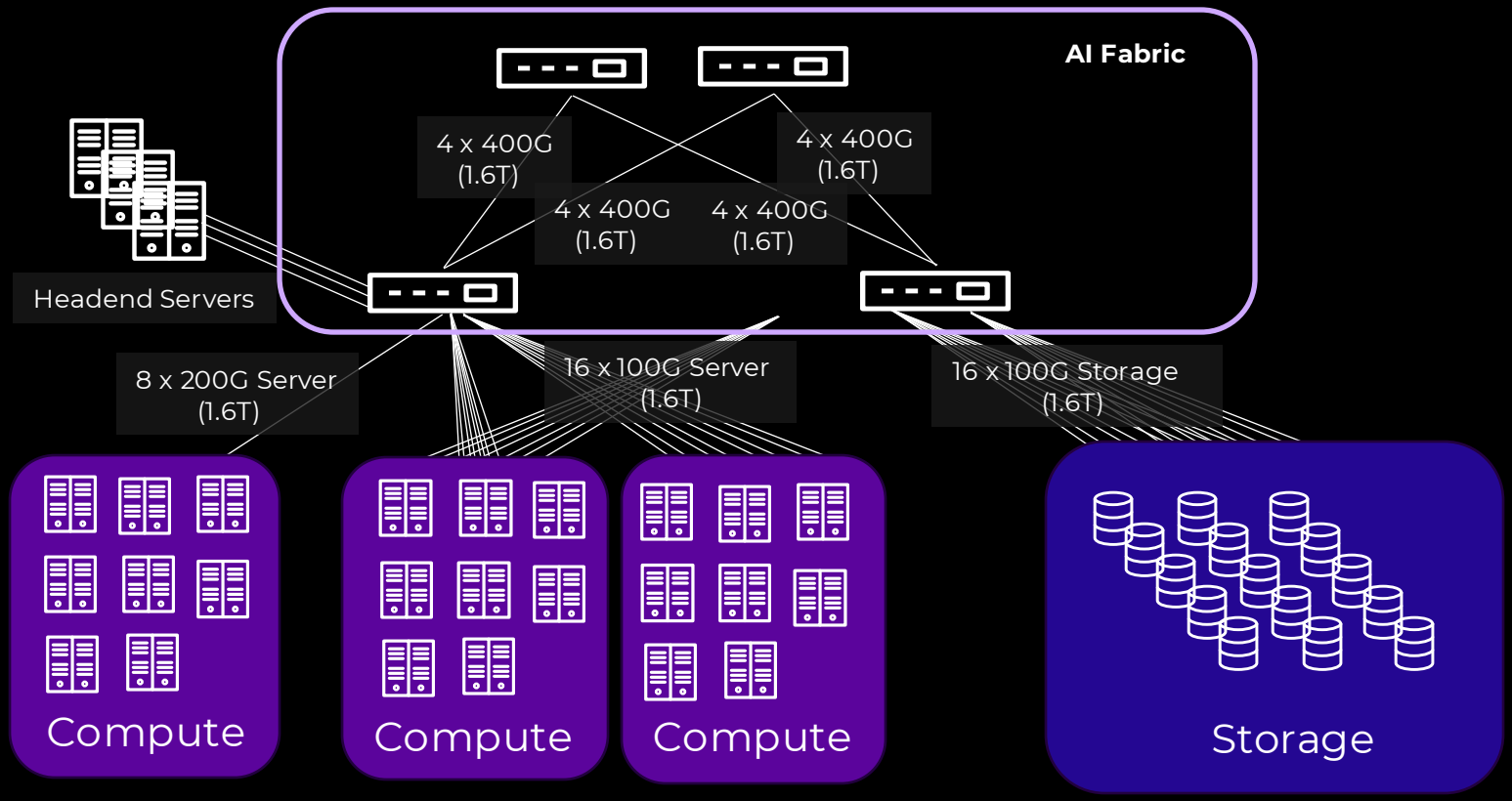
Protocol	InfiniBand (IB)	RoCEv2	Ultra Ethernet (UE)
Hardware Requirements	Both NICs and switches	Both NICs and standard Ethernet switches	Both NICs and standard Ethernet switches with additional optional features
Losless Mechanism	Hop-by-hop credit at L2, rate based	Hop-by-hop PFC at L2. Go Back N recovery, rate based	Adaptive Sender congestion window. Receiver controlled credit
Congestion Management	FECN and BECN	ECN marking and CNP, WRED for buffer control, QoS	PFC and ECN, QoS, optional packet trimming in switches
Load Balancing	Per packet adaptive routing	ECMP	Packet spraying, ECMP
Latency	Very low	Low	Info not available *
Control Plane	Centralized Subnet Manager	Distributed	Distributed

IP Fabric Data Center Scale-Out Design for CSP 5G Workloads



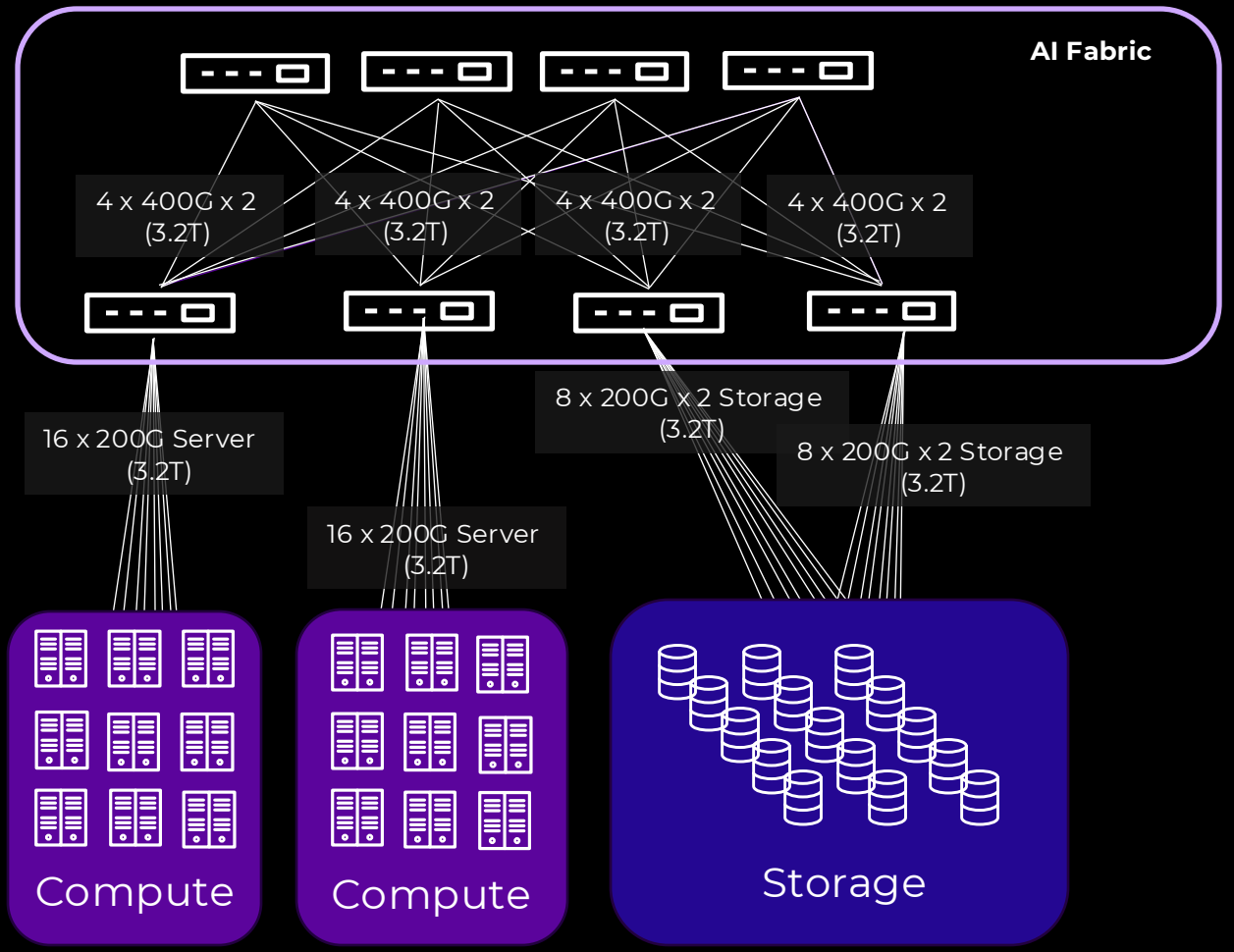
Front-End AI Data Center Reference Design

- **CPU – Storage Traffic**
  - Front-end to back-end
  - 25G/100G
- **CPU – NS Traffic**
  - N to S Traffic
  - 100G/200G

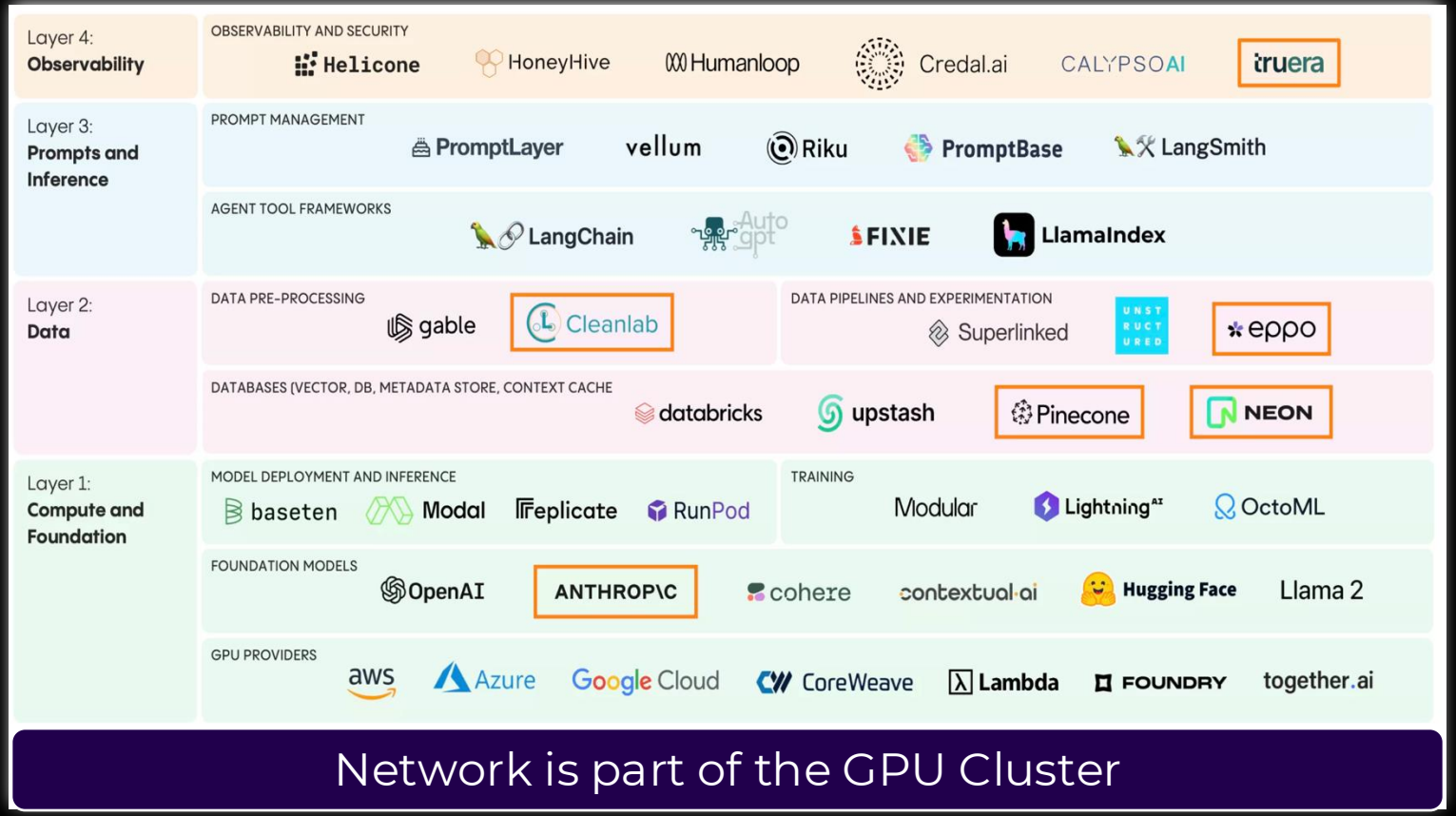


Back-End AI Data Center Reference Design

- **CPU – Storage Traffic**
  - Front-end to back-end
  - 25G/100G
- **GPU – GPU Traffic**
  - Back-end Traffic
  - 100G/200G/400G
- **GPU – Storage Traffic**
  - Back-end Traffic
  - 100G/200G
- **CPU – NS Traffic**
  - N to S Traffic
  - 100G/200G



The AI Stack



Source:Menlo Ventures

Extreme®
networks