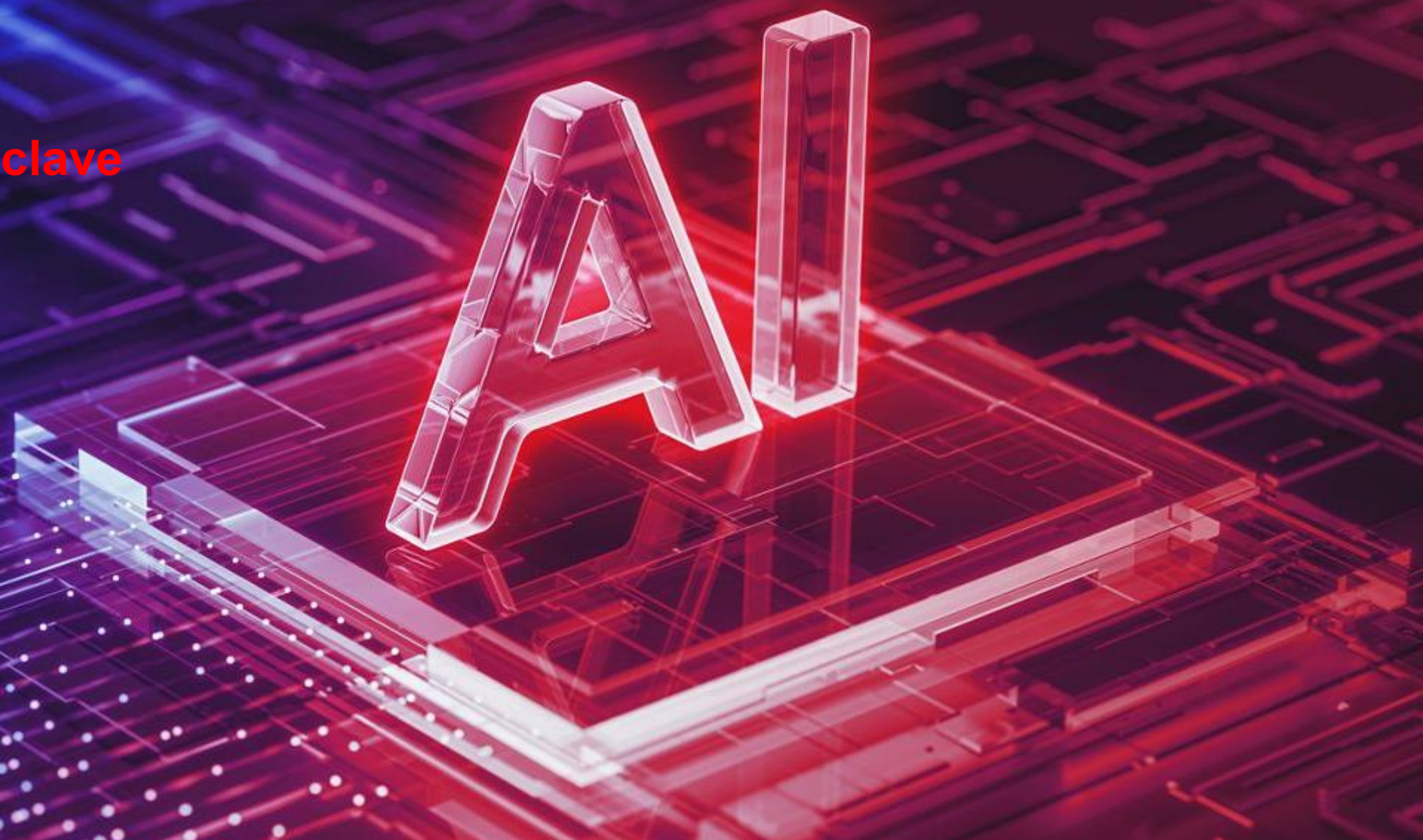
Beyond Connectivity | AI-Led Evolution for the Future

AI Thought Leadership Conclave

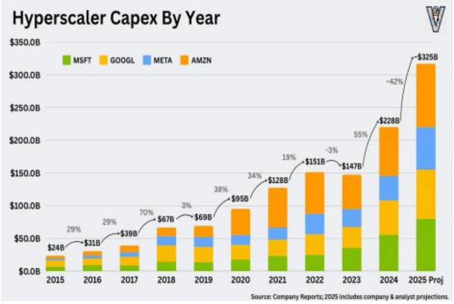


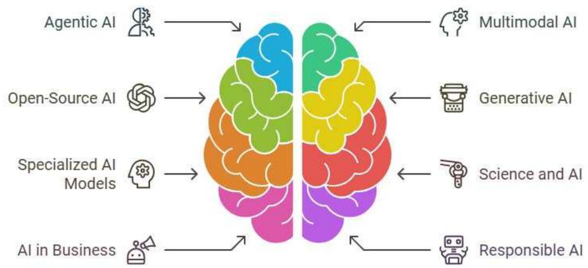
Mombasawala Mohmedsaeed
Dec 9, 2025
Bengaluru, India

KEYSIGHT

A Robust and Expanding Industry

AI Market Update





Social Impact

- AI to automate ~50% of current work within 2 decades.
- Personalized AI assistants will augment daily tasks and decision-making.

Economic Growth

- AI market to exceed $2T by 2030, with rapid, cross-sector adoption.
- Hyperscaler investment level aims to sustain the global AI growth.

Models

- Foundational models now exceed 1T parameters.
- Agentic AI and Reasoning Models require 100x compute, 20-50x tokens.
- Small Language Models, multi-agent, and multimodal drive edge efficiency

Infrastructure

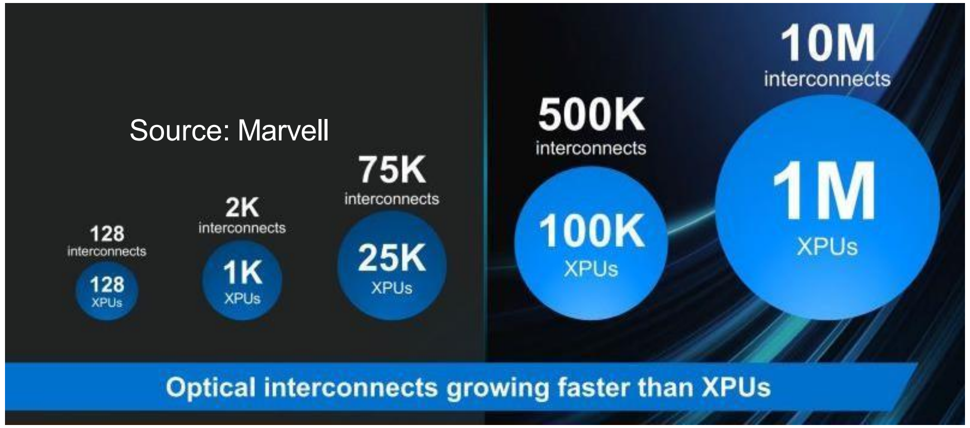
- Interconnects and the network as now the critical bottleneck, require 5-10x more per GPU.
- Energy demands necessitate chiplets, Si-Ph, and novel power-optimized innovations.

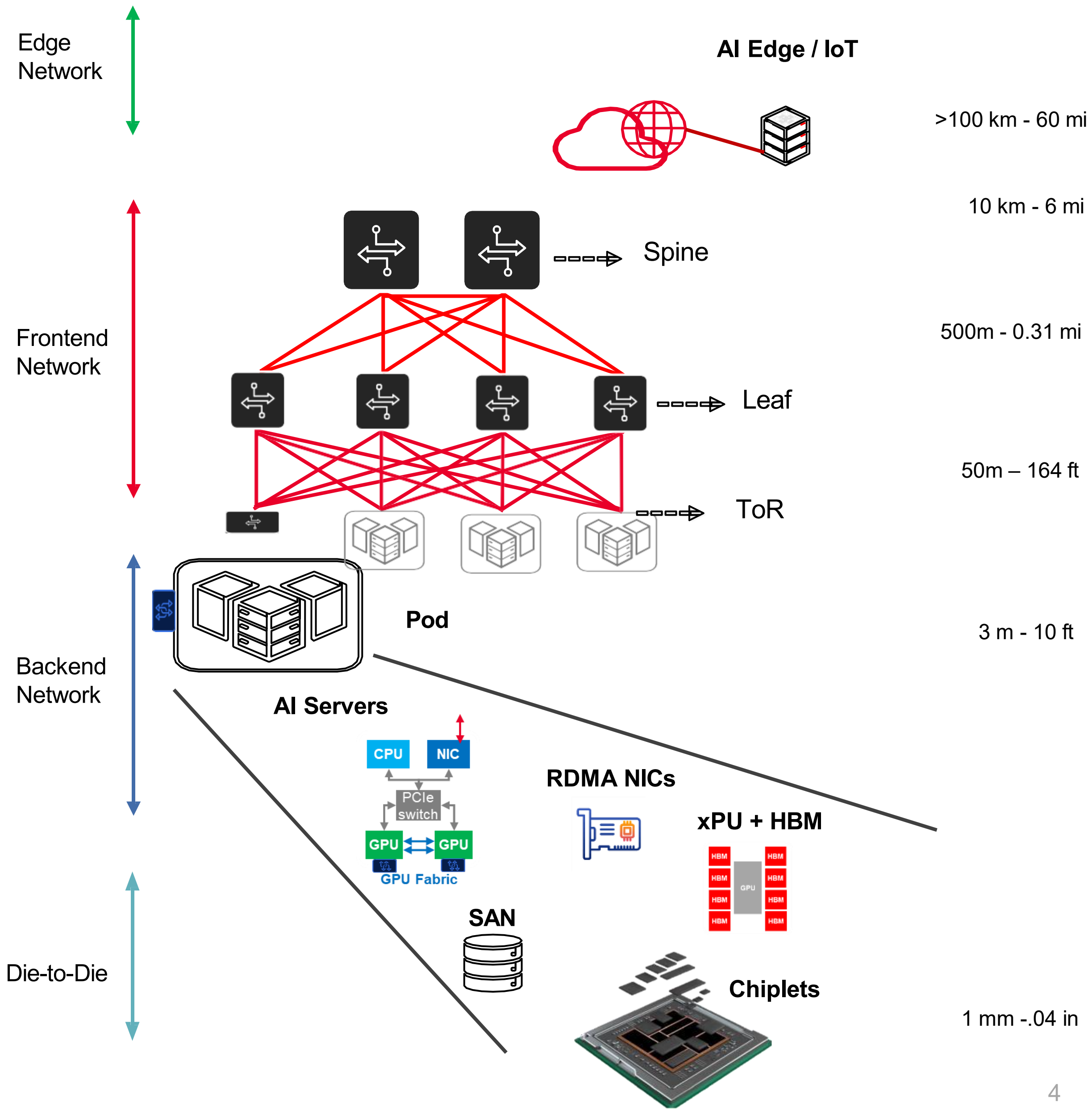
Today	Next
400/800G	1.6/3.2T
DDR5 8.4 GT/s	DDR6/HBM3 12.8 GT/s
100 Gb/s	224/448 Gb/s
PCIe 5 32 GT/s	PCIe 7 128 GT/s
5G 10 Gbit/s	6G 100+ Gbit/s

AI Infrastructure

Adapting Hyperscale DC to Edge AI

- Training Clusters: 100k+ GPUs in 2024 and **path to 600k**
- 800G/1.6T links, 112/224G lanes and path to 448G
- Power need **100+MW, 160% increase by 2030**
- New protocols for transport and congestion management

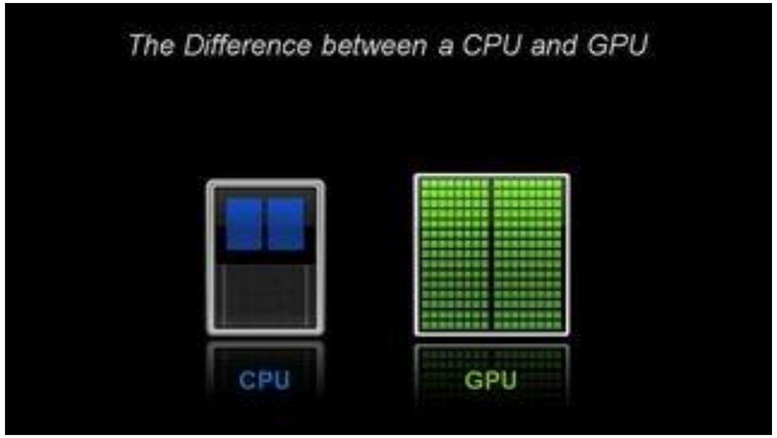




Graphics Processing Unit (GPU)

- GPUs are the de facto engines of AI computing
- Originally developed to handle complex graphics tasks in video games and computer graphics applications
- Evolved to become essential components in a wide range of computing tasks, such as Deep Learning
- GPUs perform much more work for every unit of energy than CPUs

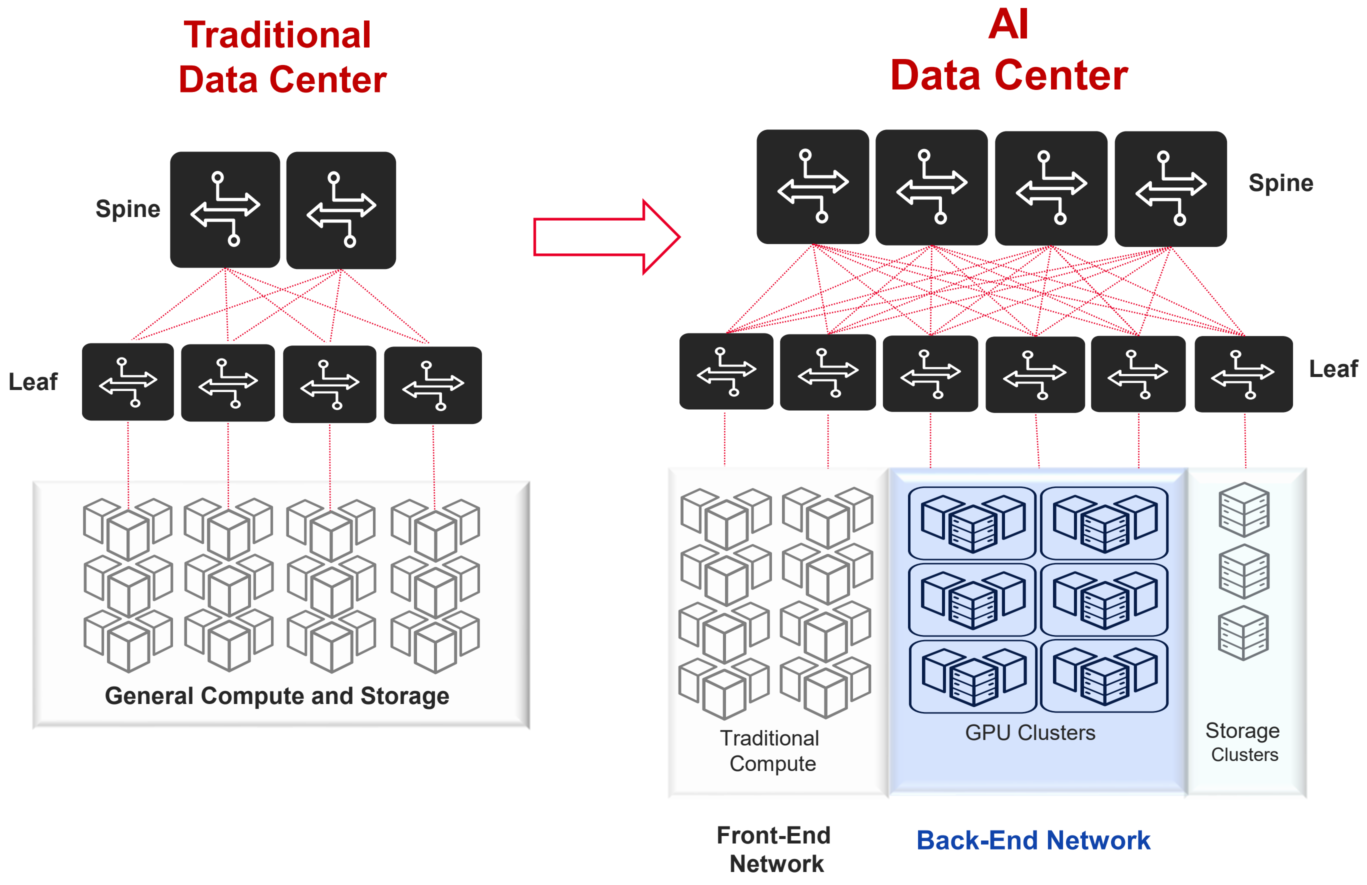
CPU	GPU / NPU
10s of cores	100s of cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Handful of operations in parallel	Thousands of operations in parallel



Evolving Data Center Architecture

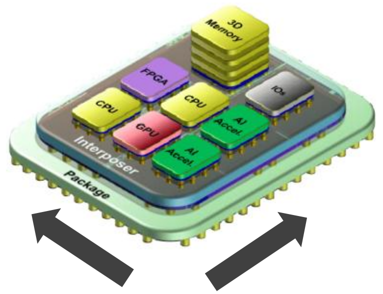
Front-End and Back-End Networks

- **AI/ML workloads**: pushing data centers to evolve their network architecture
- **AI-Specific Networking:** a dedicated Back-End network for AI workloads to isolate them from other data center traffic and ensure low-latency communication.
- **Back-End AI/ML clusters:** consists of hundreds to thousands of AI/ML accelerators, CPUs, storage devices, Switches, and Network Interface Cards (NICs) connected to GPUs
- **Frontend** is traditional cloud services and internet



AI Infrastructure Scaling – Use cases an challenges

Each network type has different Interconnects requirements



Scale In

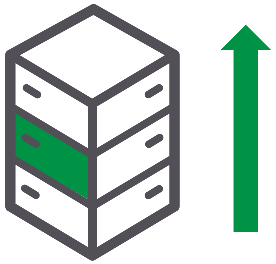
Die to Die - Network

Scales In the Chip

Bandwidth, Density, Reticle Limit, Yield

10 -100’s of Chiplets
1-50 TB/s per chip
2D - 3D Advanced Packaging
Co-packaged Copper and Optics

Standards: UCIe, BoW



Scale Up

GPU to GPU - Network

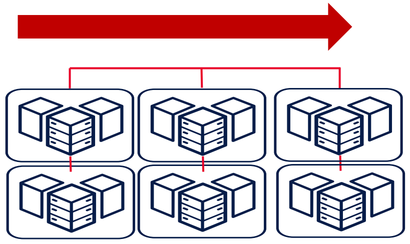
Scales Up the Rack (1-2m)

Bandwidth, Latency, Power

xPU: 10s ~ 100s

~25-100Tb/s per rack
Latency: ~ <1 us
SerDes: 200G - 400G
Shifting Copper to Optics

Standards: Nvlink, UALink



Scale Out

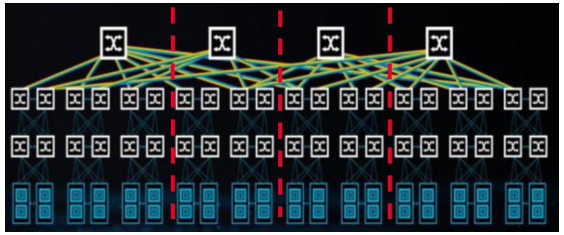
Cluster to Cluster - Network

Scales Out across DC (2km)

Resilient, Bandwidth, Scale

10k ~ 100k xPU
~100Tb/s to 10 Petab/s
Latency: 1 ~ 100s of us
SerDes: 100 – 200 Gb/s
Energy efficient optics

Standards: Ultra Ethernet, InfiniBand, Coherent-Lite



Multi-Data Center

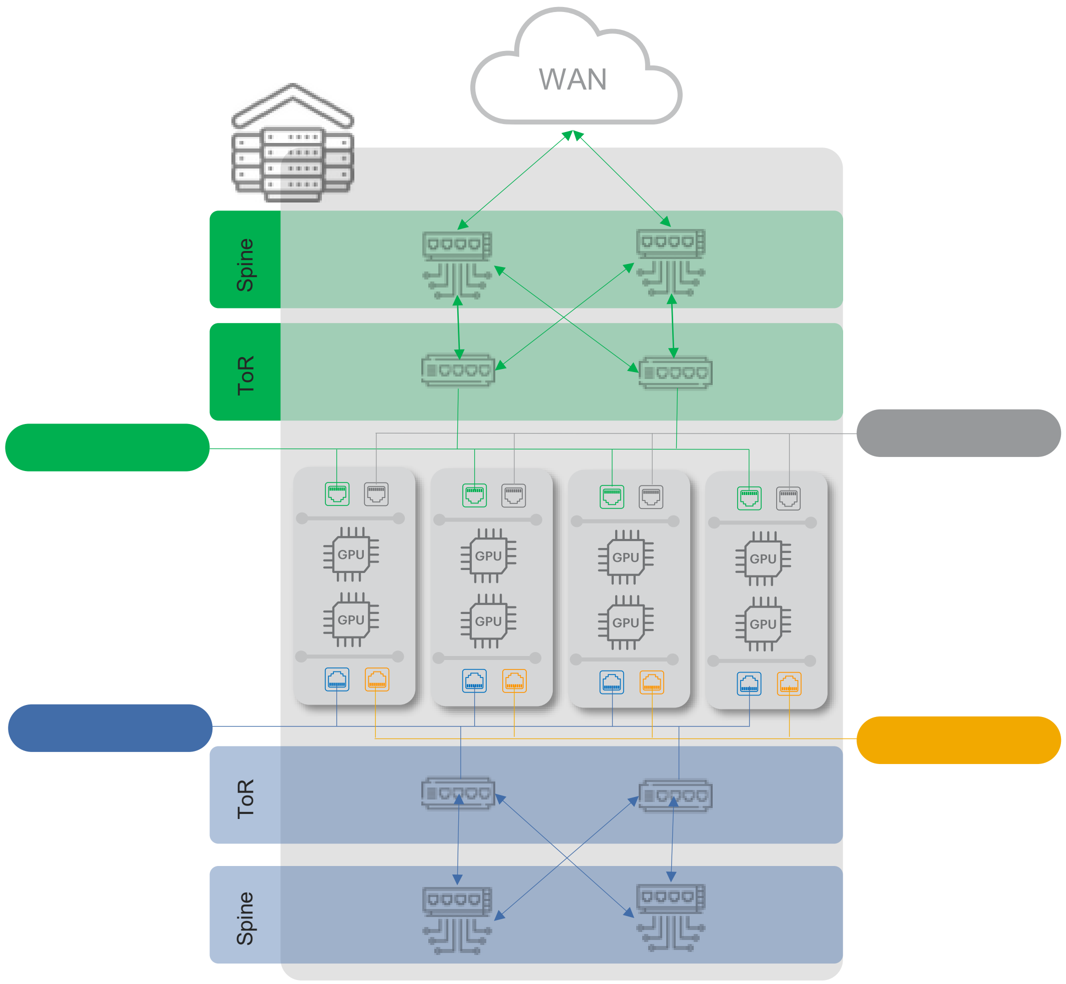
DC to DC - Network

Scales across a region (100km)

Power Capacity, Space Limitations

> 100k xPU
10s of Petab/s
Latency: 100s of us
Coherent Optics

Standards: 800ZR, 1600ZR



Network Topology drawn for simplicity over accuracy

	Front End Network	Back End Network
Network's Purpose	Connecting Servers (x86, ARM, etc. Connecting Servers to Internet	Connecting Specialized End-Points (GPUs, Storage, etc.)
Network Bandwidth Drivers	Migration to the Cloud	Connecting Specialized End-Points (GPUs, Storage, etc.)
Meeting the Bandwidth Needs	YoY CPU and Network Increase Scale-out within a generation	HPC & Storage Explosion of AI/ML Step Function increase in Bandwidth
Interconnect	Ethernet (Massive Investment)	Proprietary (Limited Investment) → Ethernet (Massive Investment)

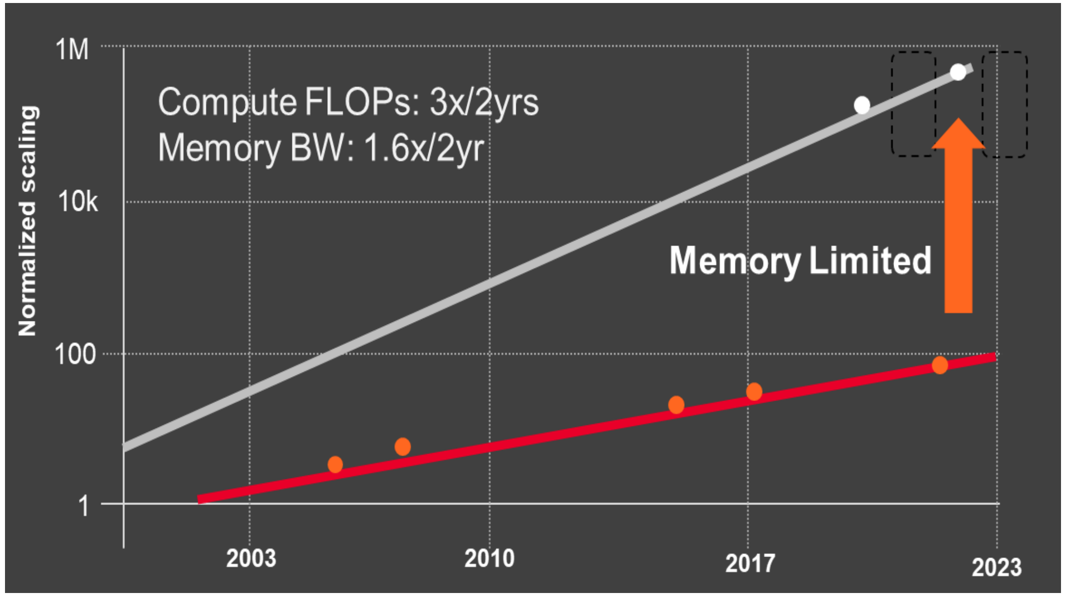
Bottleneck - Compute-Memory Gap

Memory bandwidth has not kept up with compute

- Different Memory technologies
- HBM – Training
- LPDDR6 – Inference

Breaking the memory wall:

- More efficient Models
- Efficient training
- Optics and Disaggregation

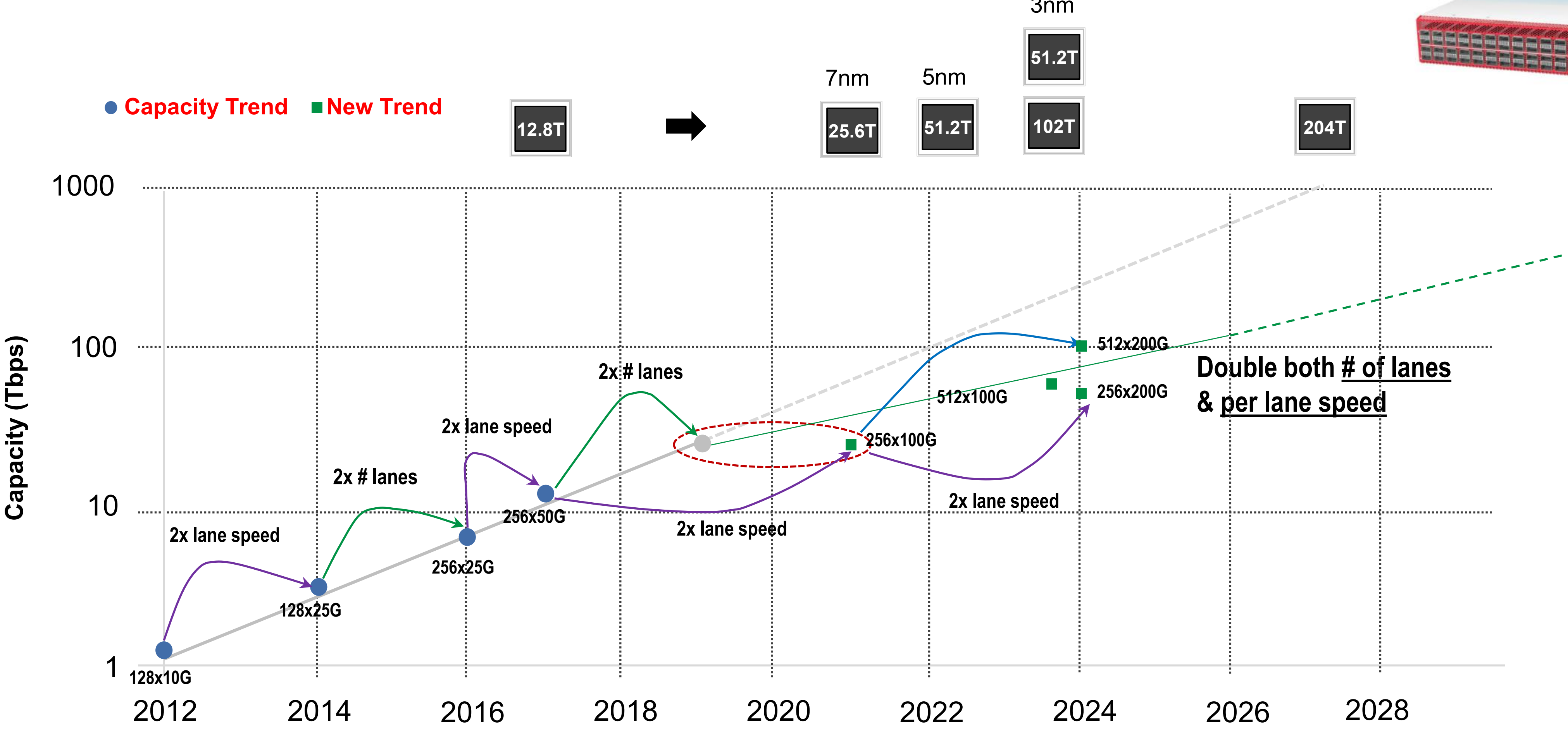


Memory AI Use Case

	HBM3E	GDDR7	LPDDR6	DDR6
Target Devices	High-performance GPUs, AI accelerators	High-performance GPUs, gaming PCs	Mobile devices, laptops, edge computing	Desktops, workstations, servers
Bandwidth per pin (GB/s)	9.2	32	16.8	16
Power Consumption	Higher	Higher	Lower	Lower
Cost	Highest	High	Moderate	Moderate
Typical Use Cases	AI Deep learning, scientific computing, high-performance graphics	High-end gaming PCs, AI inference engine	AI inference on Devices Smartphones, tablets, embedded systems	Desktops, workstations, servers, network equipment

Bottleneck – Network –Compute

Ethernet Switching ASIC drive the Network Evolution



Network/Interconnect is now the bottleneck as it falls behind compute

Challenges: Design Verification is Only Getting Harder

Keysight helps you solve these challenges

- 112 to 224 to 448G
- PAM4 or PAM6 or PAM8
- Shrinking Design Cycles

- Small dimensions
- Chips stacked on chips
- Chips packaged inside a bigger package

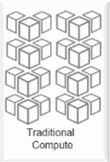
- CPO, LPO
- SiPho
- Chiplets

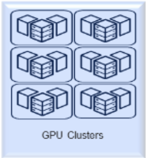
- Jitter budgets in the 10s of fs
- Closed RT Eyes
- New error correction schemes

- Congestion
- Crosstalk

- Must meet tough standards
- Need to work with many other vendors
- Who’s problem is it if it fails?

AI DC Transformation – Optimized Networks for Performance





	Front-End	AI/ML Back-End
Rack Bandwidth	3.2T – 12.8 T	>>100 T
Rack Power	~10 kW	100 kW +
Reliability	Low Importance	Critical Importance
Latency	Low Importance	High Importance

- **FLOPS**: measures the computational workload of a model in terms of floating-point operations.
- **Tokens**: AI models use tokens to process and generate text, enabling features like prediction and generation. It is the compute output
- The relationship between FLOPs and tokens is that the cost of processing text (in terms of FLOPs) is related to the number of tokens processed.

$$Performance = \frac{Tokens}{(Watt + \$)}$$

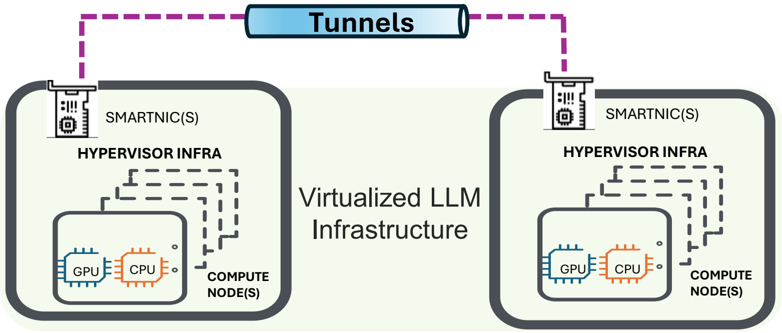
Speed | Power | Thermal | Latency | Density | Reliability

Design Priorities for AI/ML Interconnects

The Operation of AI ML Network Infrastructure

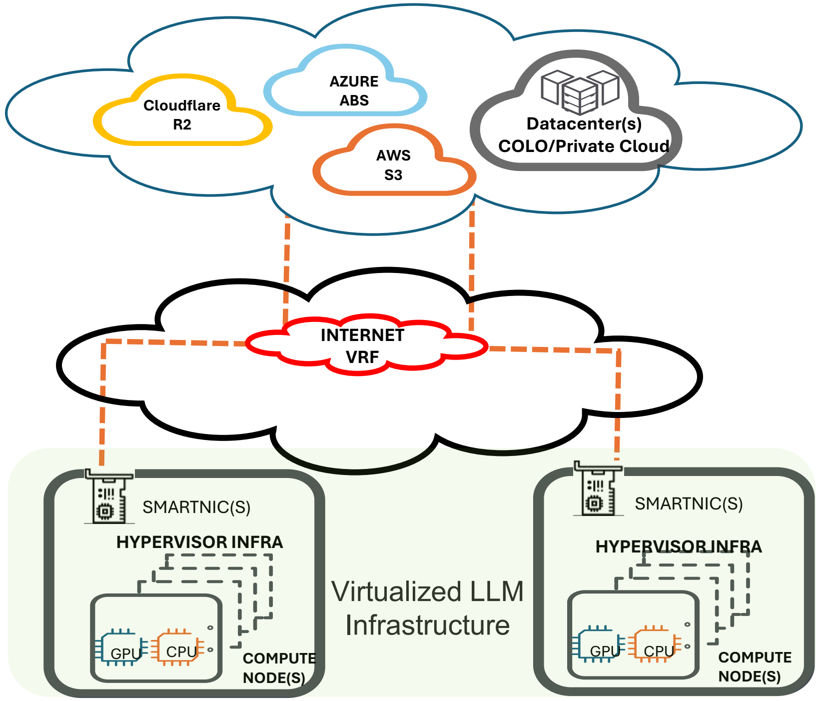
Backend Data Center for AI Models Training
East-West Traffic Test Demands -

- Distributed GPU/CPU architectures
- Collective communications & parallel processing among GPU nodes
- Hyper-virtualized infrastructures for multi tenancy
- Immense performance needs for lossless connectivity and minimum tail-end latency



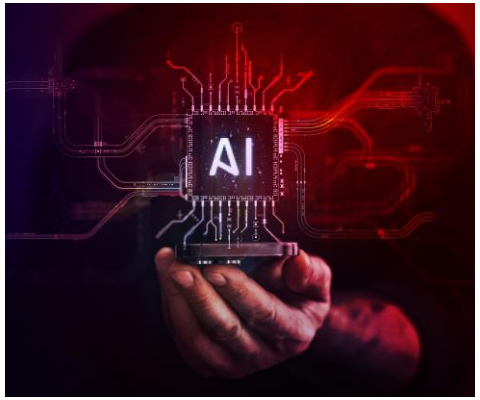
Front-end Data Center for Inference Workloads
North-South Network Traffic Test Demands -

- GPUs need high-speed access to block/remote storages
- Provisions to secure data in motion
- Ultra-low latency demands



AI | Model Training | Inference

- **AI (Artificial Intelligence)** is a **tool** that helps machines learn from data and act "smart" and behave like how humans think or make decisions - it's still just code and math under the hood!



- **Model training** is the process of teaching a machine learning (ML) model to make accurate predictions by learning patterns from data

- During **training**, the model learns what a dog looks like

- **Inference** is when a trained AI model is used to make predictions or decisions based on new, unseen data.

- During **inference**, you give it a new photo, and it says: “This is a dog!”

AI Model Training

3 Step Process

Step 1: Data preparation

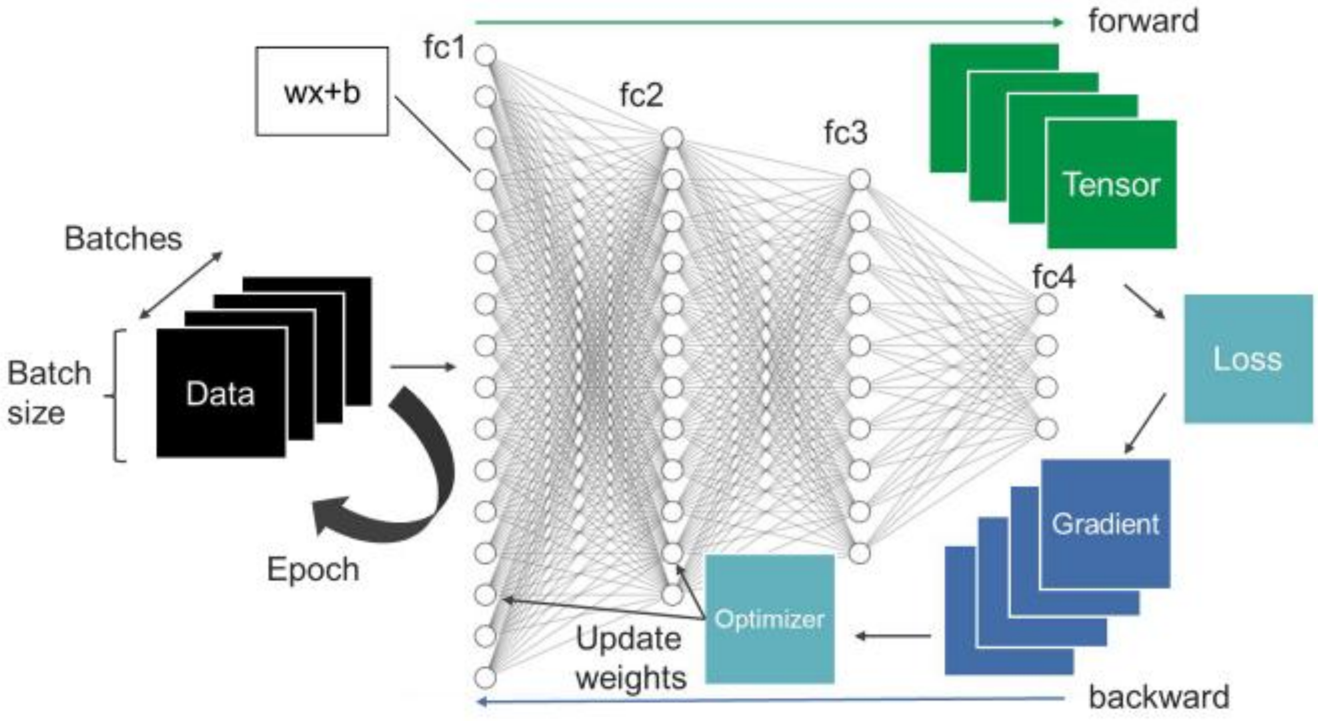
- Collect and preprocess large datasets (for example, text files, images, and audio).
- Tokenize and normalize data to ensure consistency and efficiency.
- Split data into training, validation, and testing sets.

Step 2: Model definition

- Define the architecture of the AI model (for example, neural network and decision tree).
- Specify hyperparameters (for example, learning rate, batch size, and number of layers).

Step 3: Model training

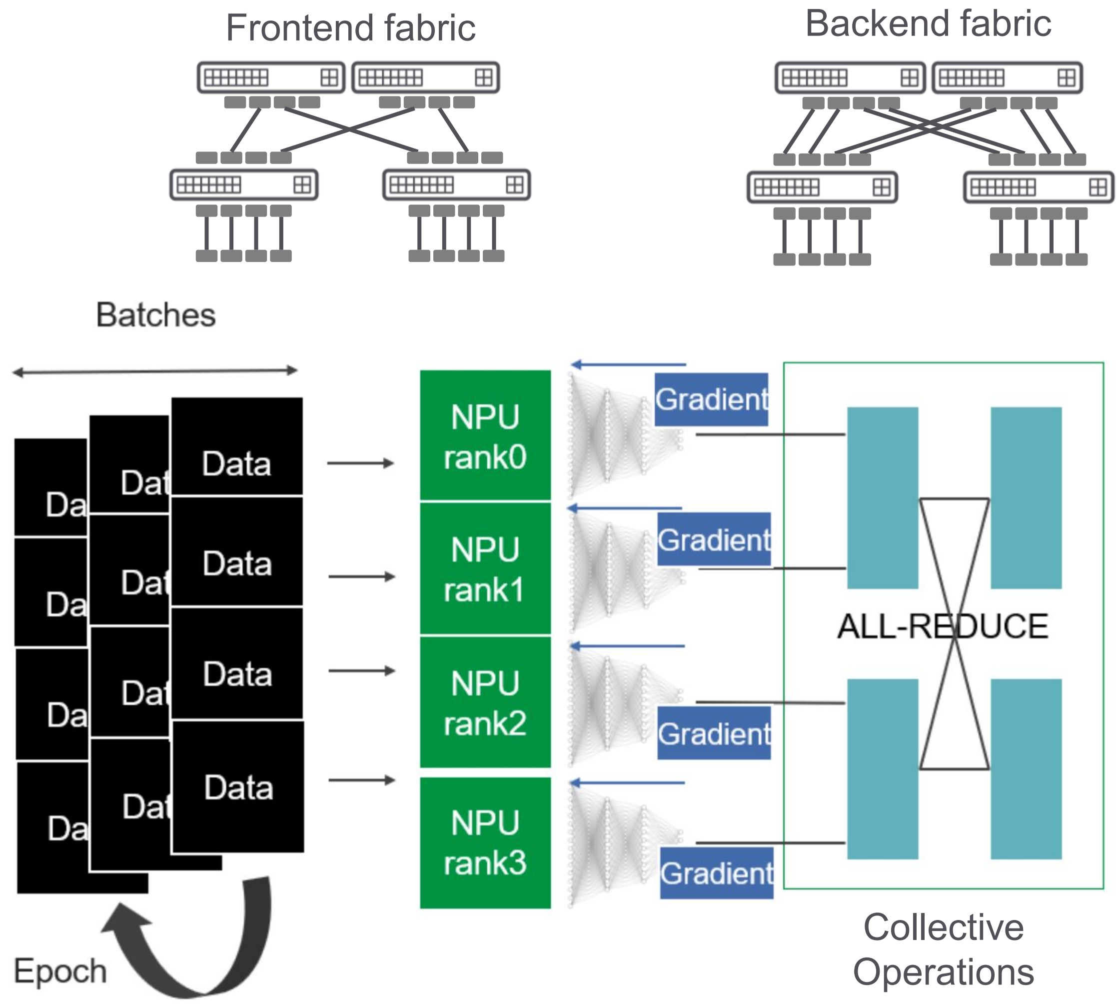
- Initialize the model's weights and biases.
- Feedforward pass: Compute outputs for each sample in the training set.
- Backpropagation: Calculate gradients and update model parameters by using an optimization algorithm (for example, Stochastic Gradient Descent and Adam).
- Repeat the preceding steps until convergence or a stopping criterion is reached.



Network role in AI clusters

Scaling up systems, scaling out clusters

- Accelerate model training with **Data Parallelism**
- Split large models across GPUs with **Tensor** and **Pipeline Parallelism**
- Subdivide complex problems among several models with **Mixture of Experts**

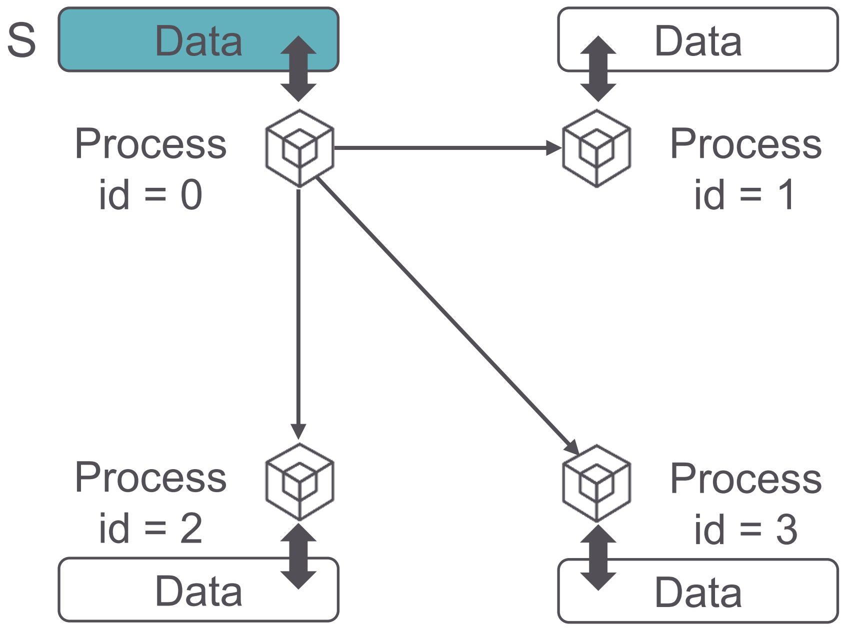


Data Parallel Architecture

Collective Operations Terminology

Broadcast as an Example

- World = group of processes
- Rank = id of the process
- Collective size (n) = number of ranks
- Data Size (S) = size of memory buffer
  - Per rank
  - Commonly, the same among all ranks
- Rank 0 = root rank
  - Broadcast
  - Gather

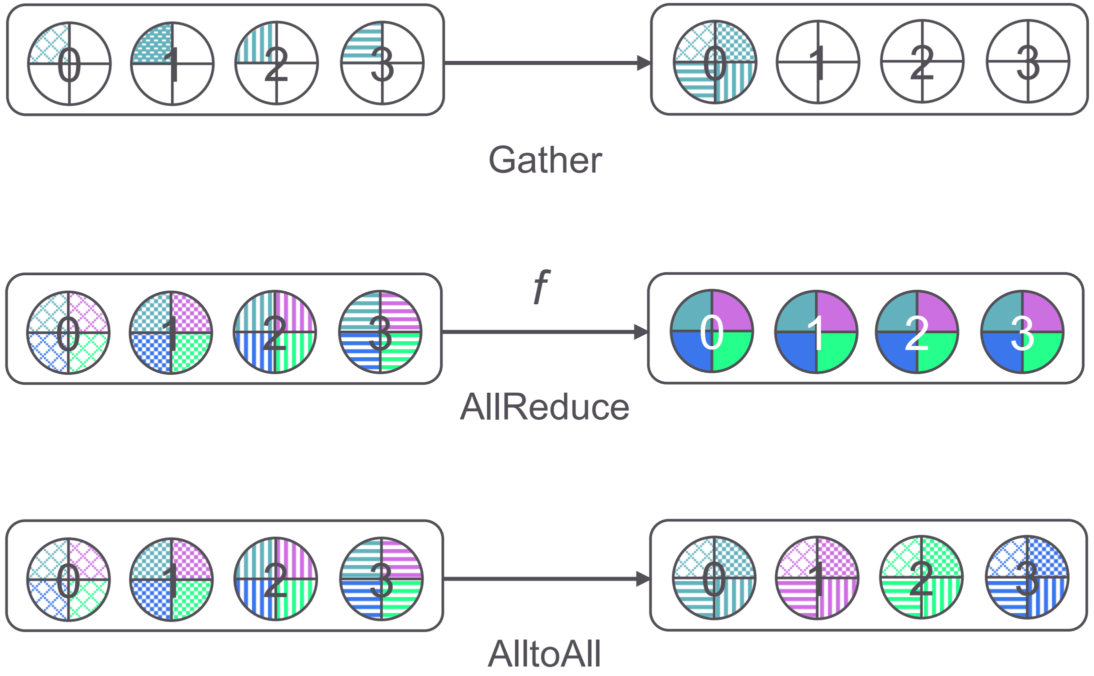


Broadcast over collective size 4

Broadcast initial and final data placement

Types of Collective Operations

- Common types for AI workloads:
  - Broadcast
  - Gather
  - AllReduce
  - AlltoAll
  - ReduceScatter
  - AllGather
- Reduce implies math with data (*f*)
- *All* or *Scatter* – symmetry

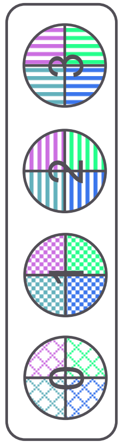


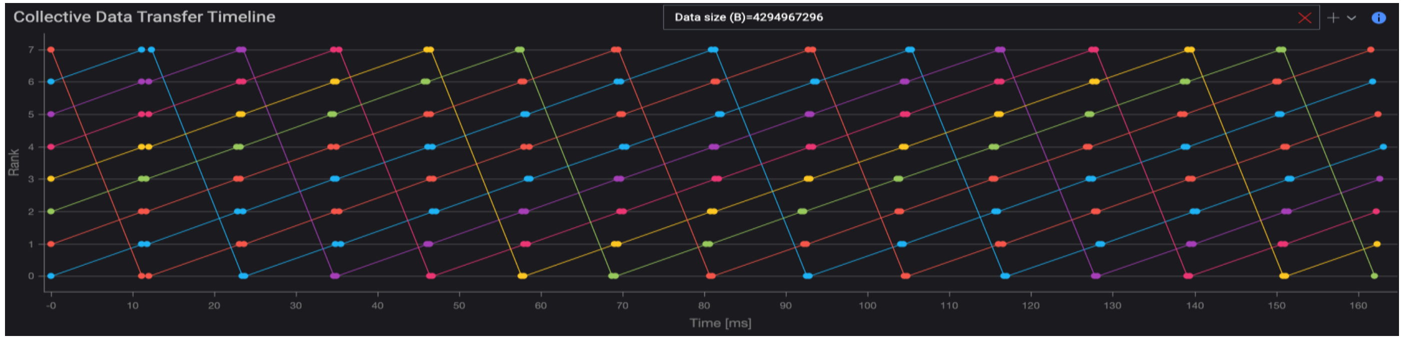
Collective Algorithms

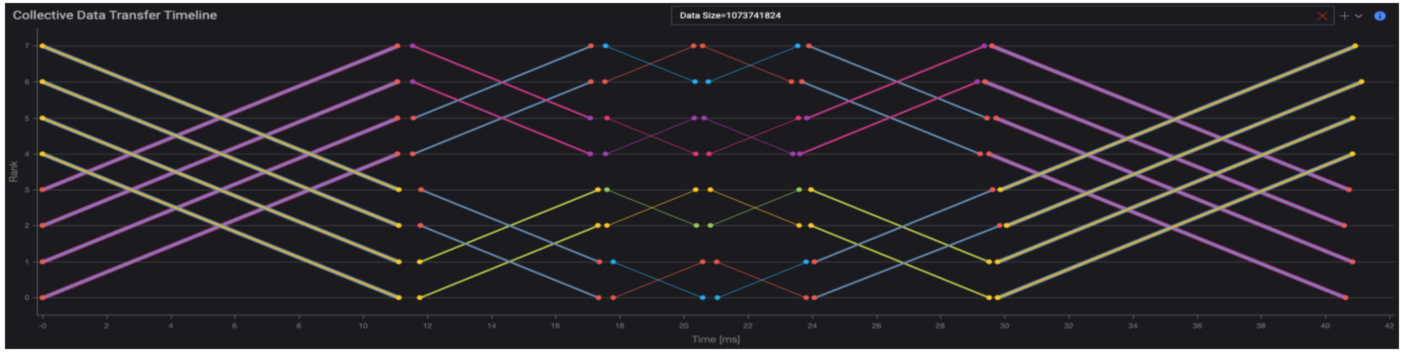
Implement Collective Operations

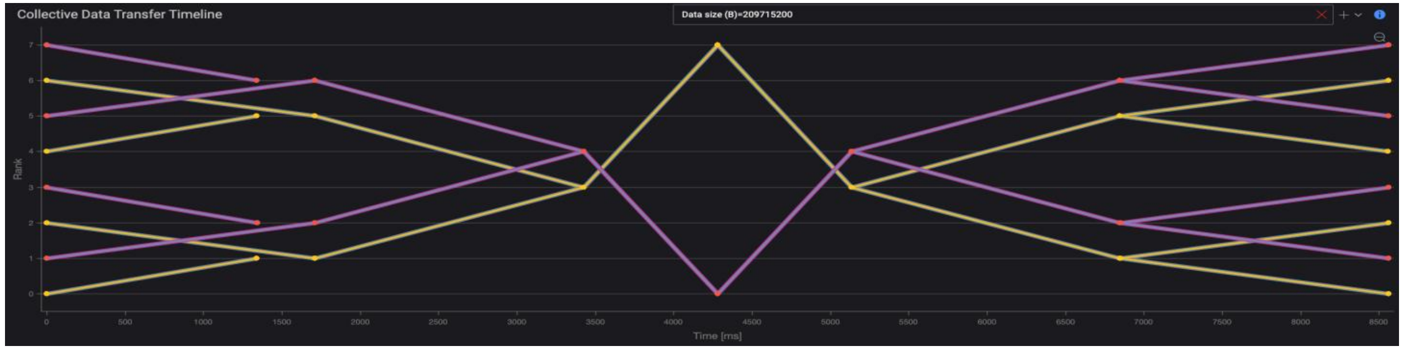
Example: AllReduce

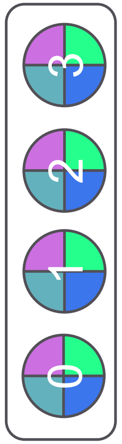
- Ring
- Halving-Doubling
- Double Binary Tree
- …



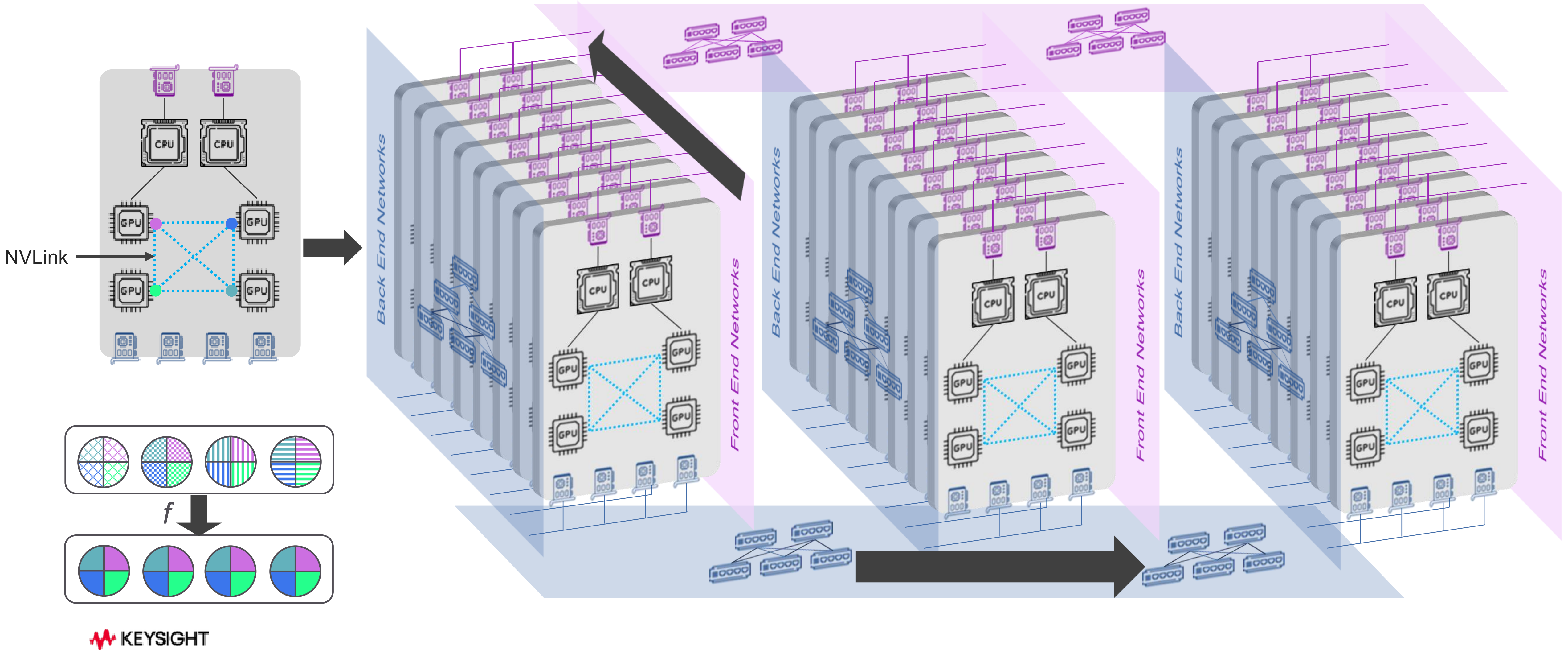


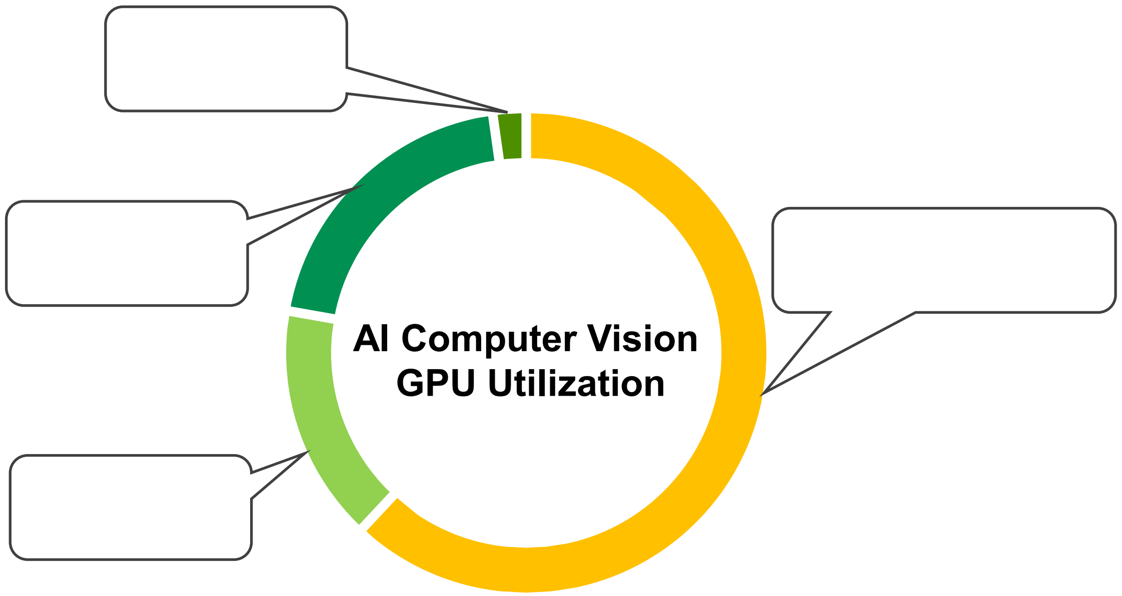






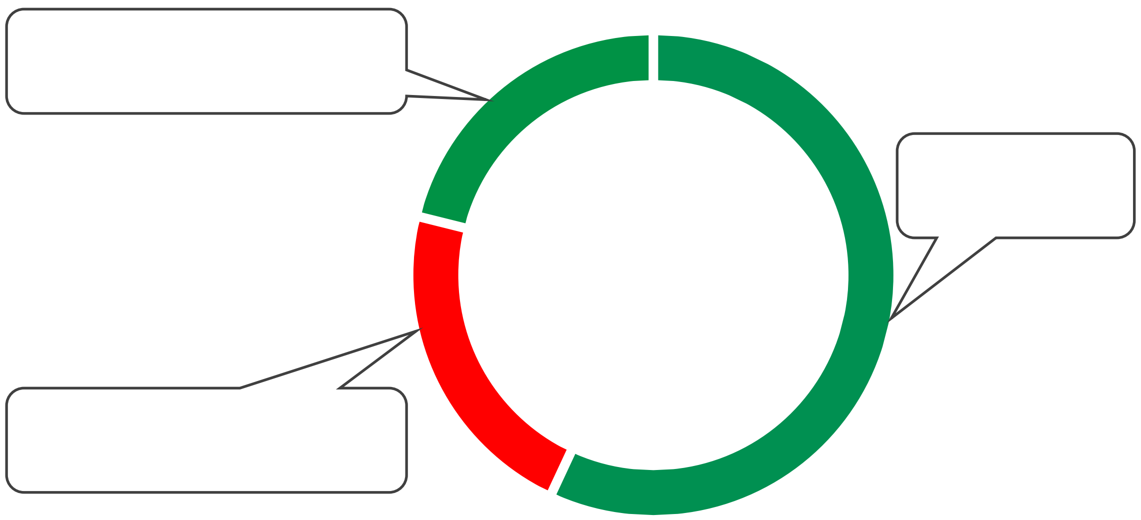
- Software libraries
  - NCCL
  - RCCL
  - MSCCL
  - …


NVLink
CPU
GPU
Back End Networks
Front End Networks
f
KEYSIGHT



GPUs waiting on data
>50%

Vision transformer (ViT) example. Source: https://github.com/facebookresearch/HolisticTraceAnalysis/

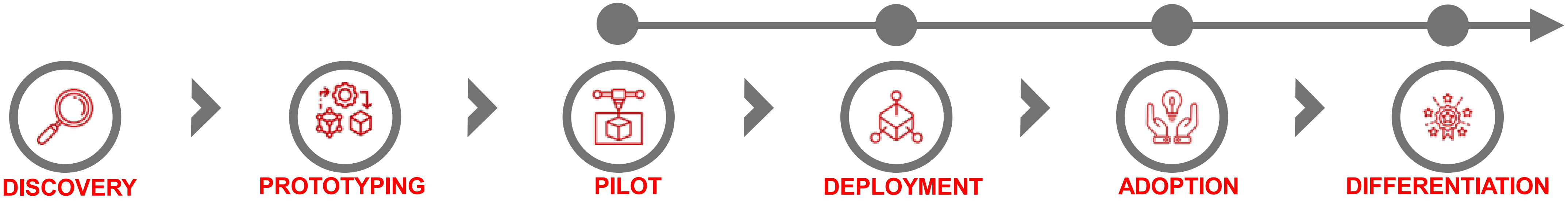


Training task failures
>43%

Source: Unicron: Economizing Self-Healing LLM Training at Scale, Tao He[1], Xue Li[1], Zhibin Wang[1,2], Kun Qian[1], Jingbo Xu[1], Wenyuan Yu[1], Jingren Zhou[1]
[1]Alibaba Group, [2]Nanjing University

KEYSIGHT

Infrastructure Scale Challenges



Scale Challenges

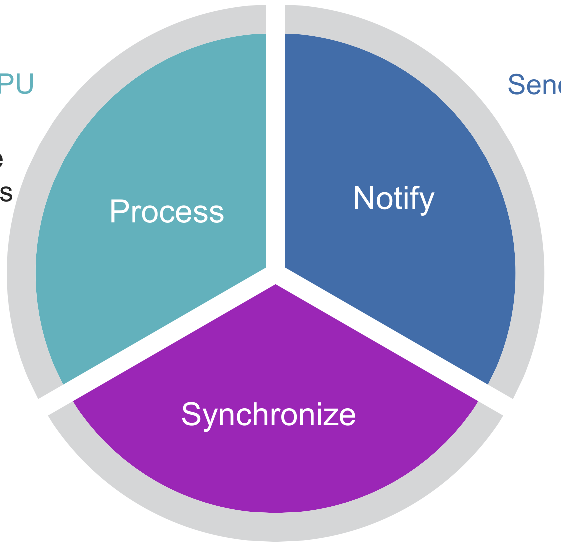
	PILOT	DEPLOYMENT	ADOPTION	DIFFERENTIATION
Bottle Necks	Performance	Performance and Cost	Cost	Performance
Trigger	Inference Latency GPU access	Trade-offs Reliability Data I/O	Usage Growth	Demand for Next-Gen AI Models
Root Cause	Ad-hoc cloud performance Inadequate bandwidth	Over- or under-provisioning Legacy infrastructure	Lack of optimization Poor workload placement	Lack GPU elasticity Inadequate interconnect

Execute Instructions on GPU

High Bandwidth Compute can saturate Network Links

Send Results of Computation

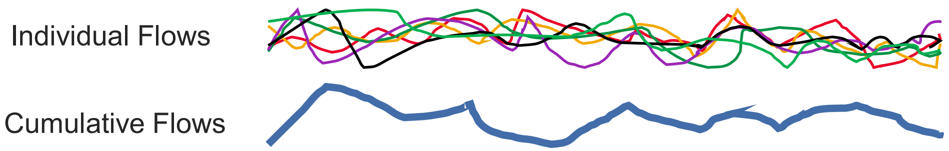
Different Collective Operations



Wait for Everyone to Compute

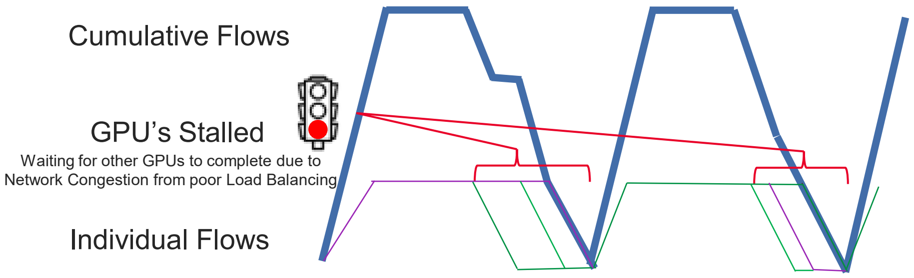
Creates Synchronization between GPUs
Computation stalls waiting for the slowest path
Job Completion Time (JCT) is based on worst case tail-latency

Traditional Data Center Traffic Pattern



Many asynchronous small bandwidth flows
Chaotic pattern averages out to consistent load

AI (All-to-All Collective) Traffic Pattern



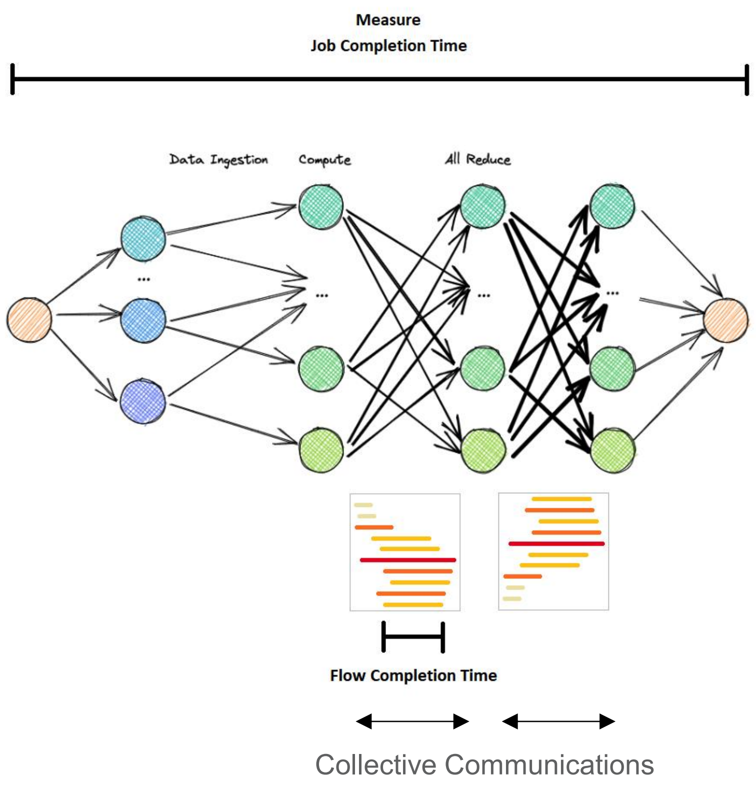
Few **synchronous** high bandwidth flows
Synchronization magnifies the long tail latency & bad load balancing decisions

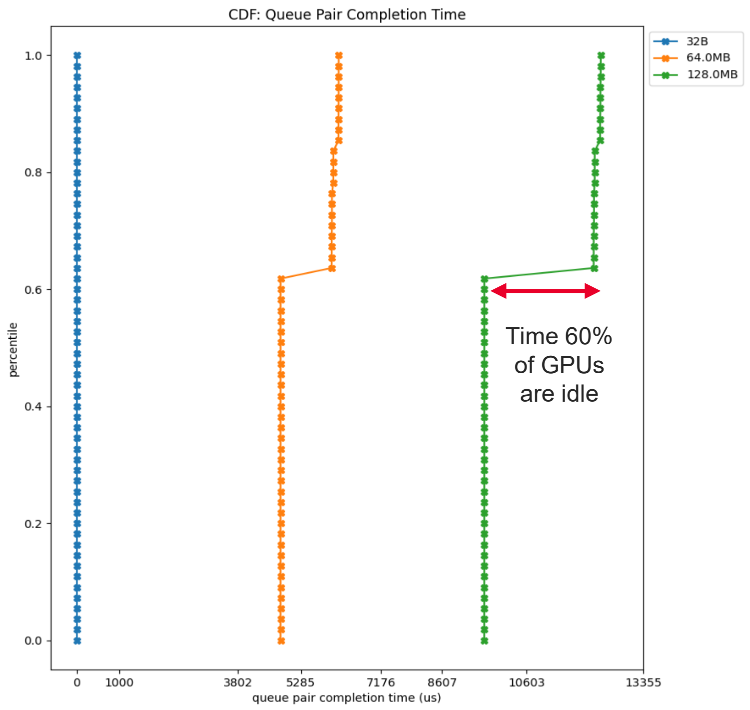
Network is the bottleneck in AI model training

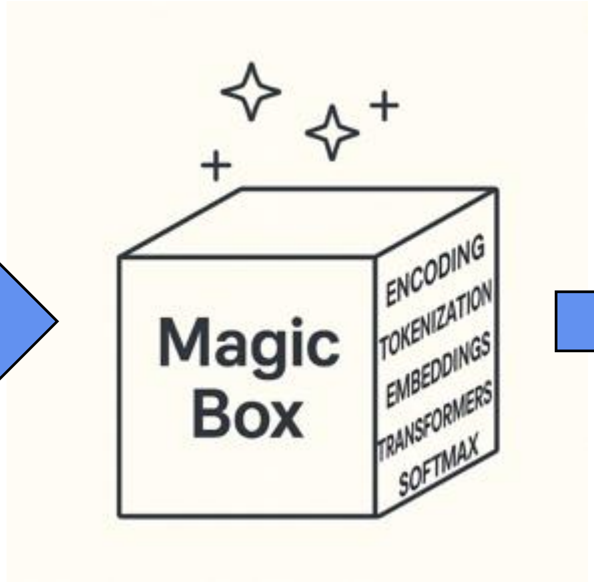
Job Completion Time Factors
- Data Ingestion
- Computation
- Collective Communications

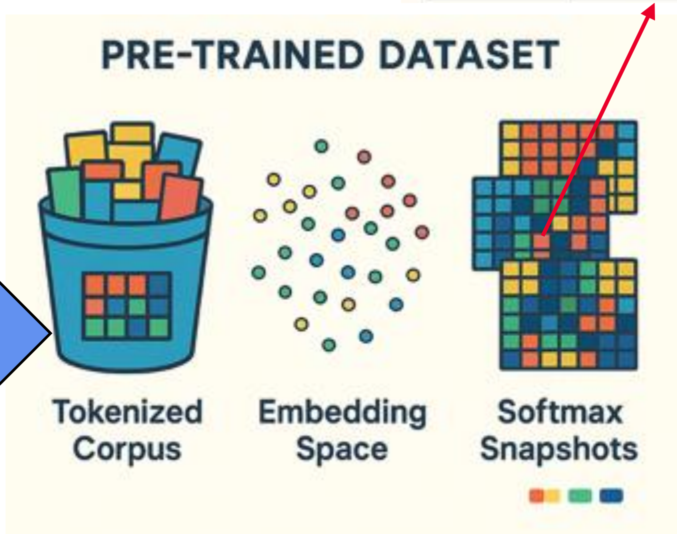
Network tail latency
- Defines wasted GPU time

Contributors
- Data exchange algorithm
- Software stack
- System I/O
- DPU (NIC)
- Network fabric

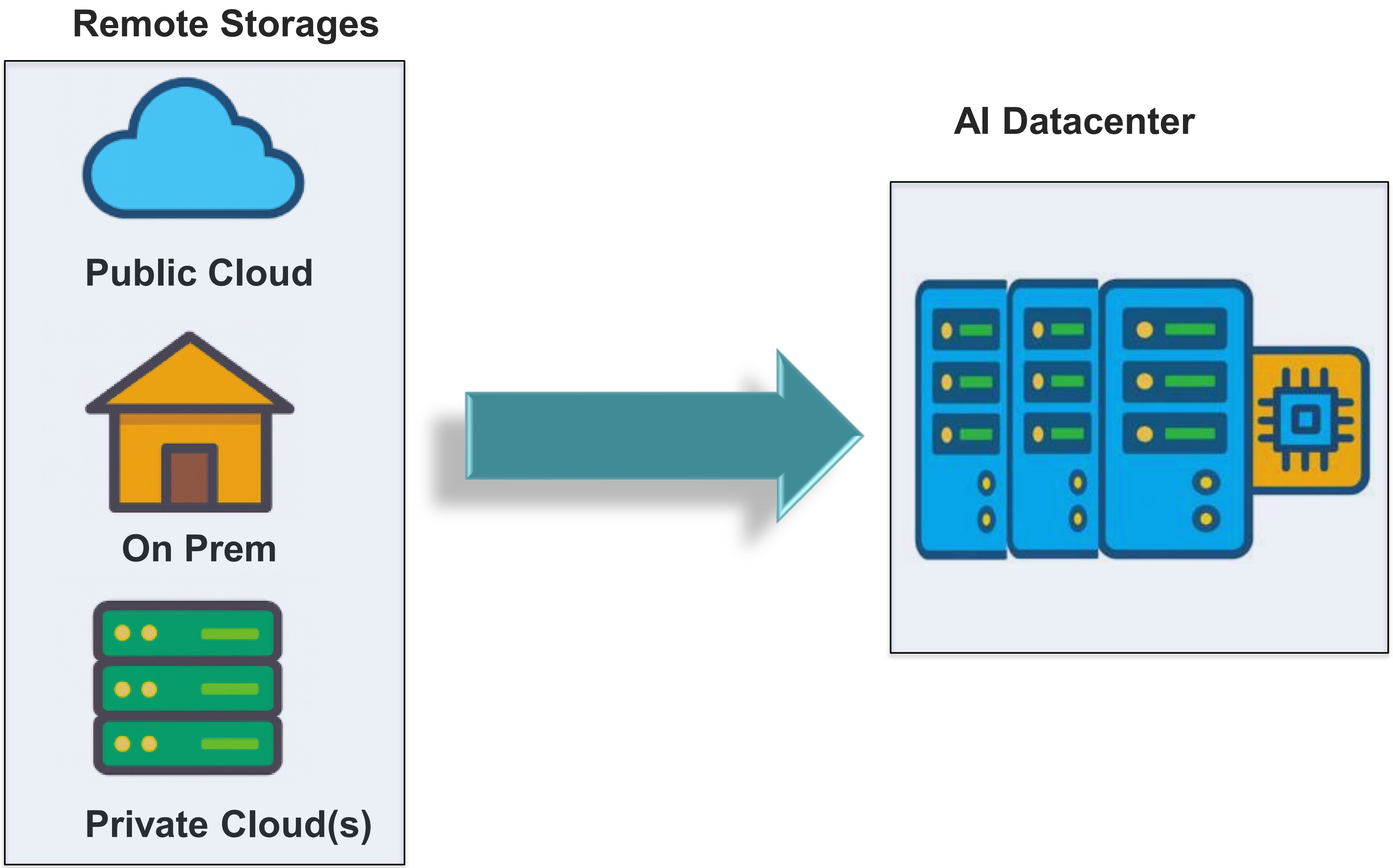






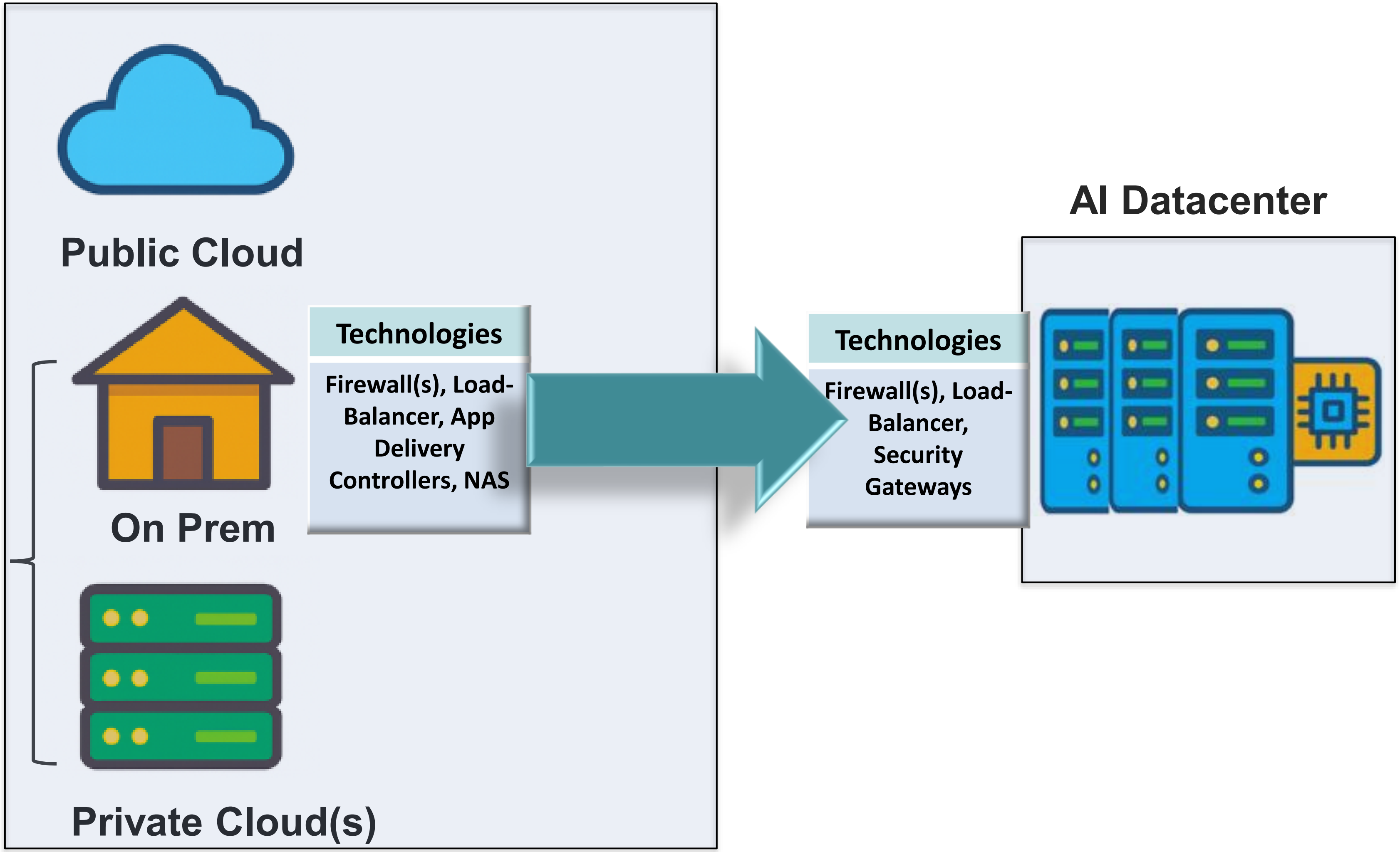


Learn	something	new
0,0295	0,004	0,0037
0,0631	0,004	0,0092
0,3867	0,084	0,0035
0,0725	0,098	0,0019
0,0221	0,034	0,0005
0,1733	0,005	0,3129
0,0431	0,129	0,0344
0,0133	0,536	0,2245
0,0867	0,457	0,5545


Remote Storages
Public Cloud
On Prem
Private Cloud(s)
AI Datacenter

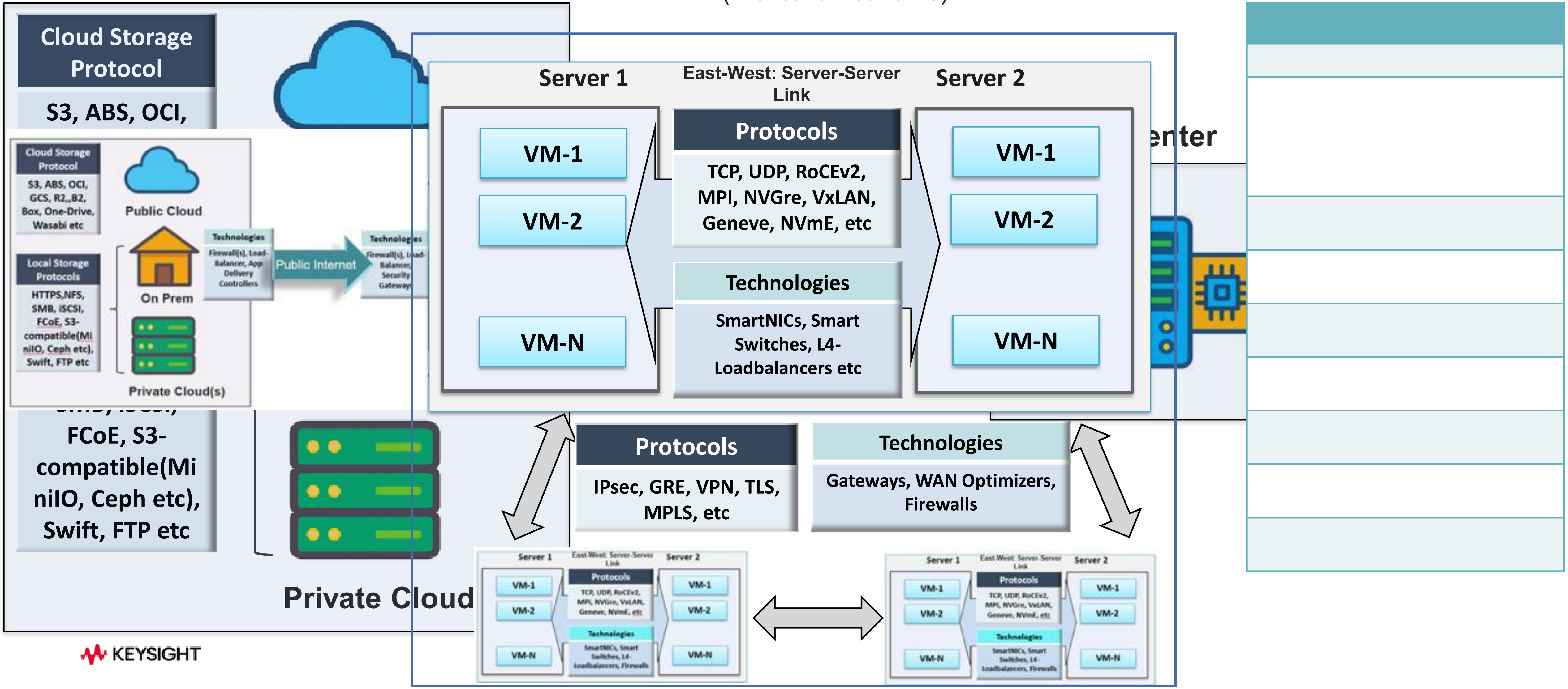
Cloud Storage Protocol
S3, ABS, OCI, GCS, R2,B2, Box, One-Drive, Wasabi etc

Local Storage Protocols
HTTPS,NFS, SMB, iSCSI, FCoE, S3-compatible(MinilO, Ceph etc), Swift, FTP etc



KEYSIGHT

AI Datacenter
(Frontend Networks)



KEYSIGHT

Protocols

Cloud:S3, ABS, OCI, GCS, R2,B2, Box, One-Drive, Wasabi etc

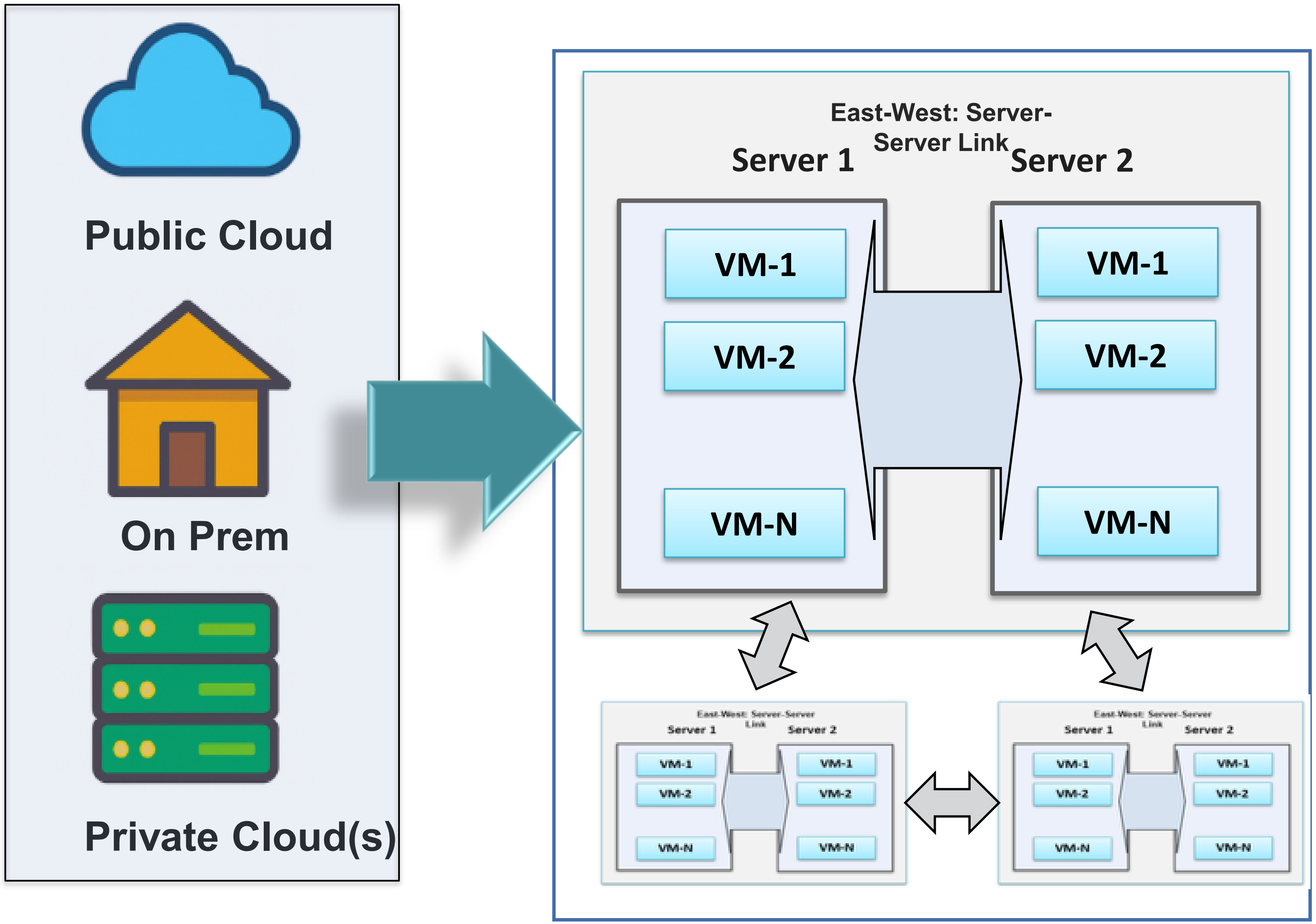
Local: HTTPS,NFS, SMB, iSCSI, FCoE, S3-compatible(MiniIO, Ceph etc), Swift, FTP etc

Technologies

Firewal, Load-Balancer, Security Gateways, NAS Eg: F5, PAN, Cloud Services, A10 etc

Remote Storages

AI Datacenter
(Frontend Networks)



Protocols

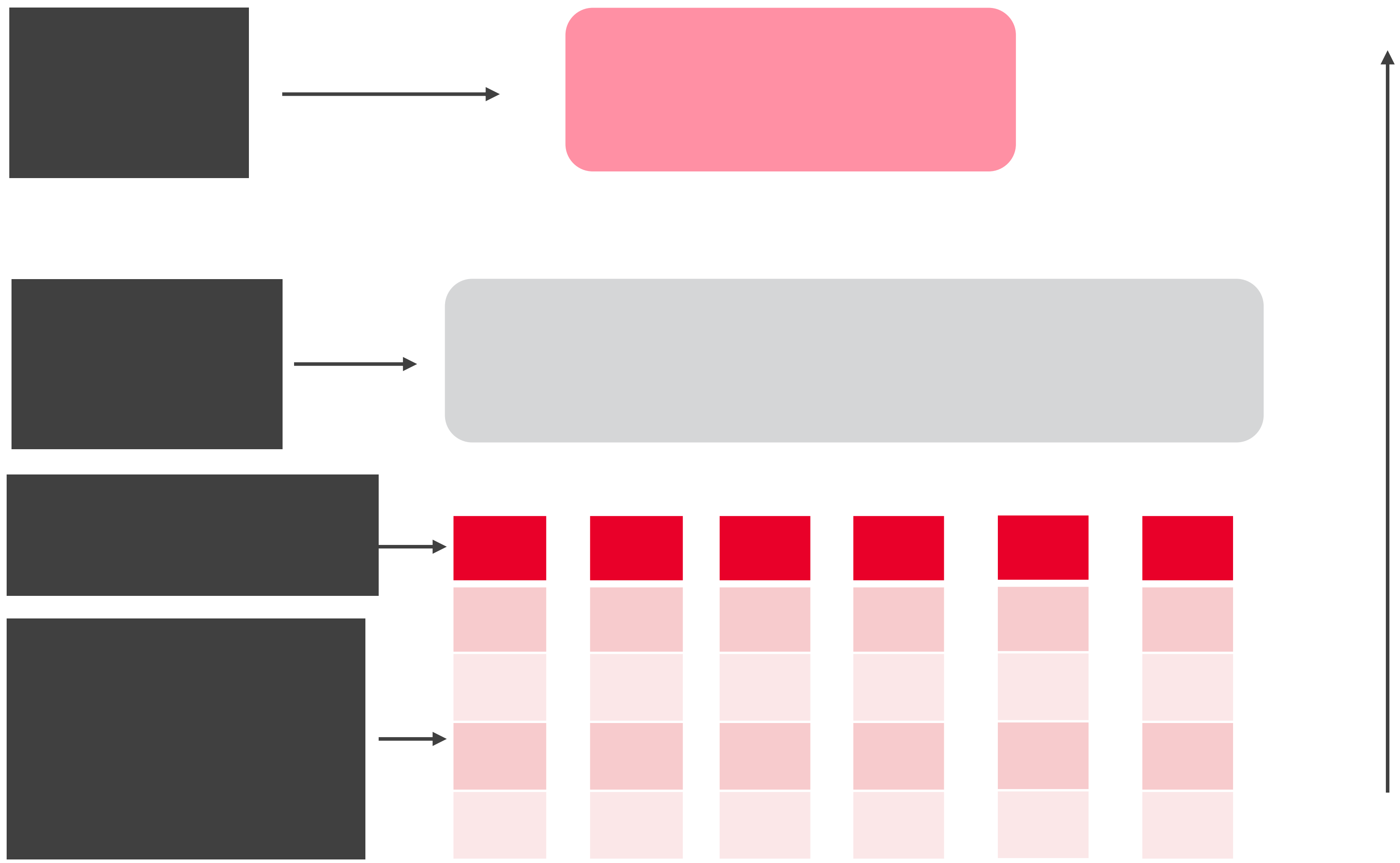
Local DC:TCP, UDP, RoCEv2, MPI, NVGre, VxLAN, Geneve, NVmE, etc

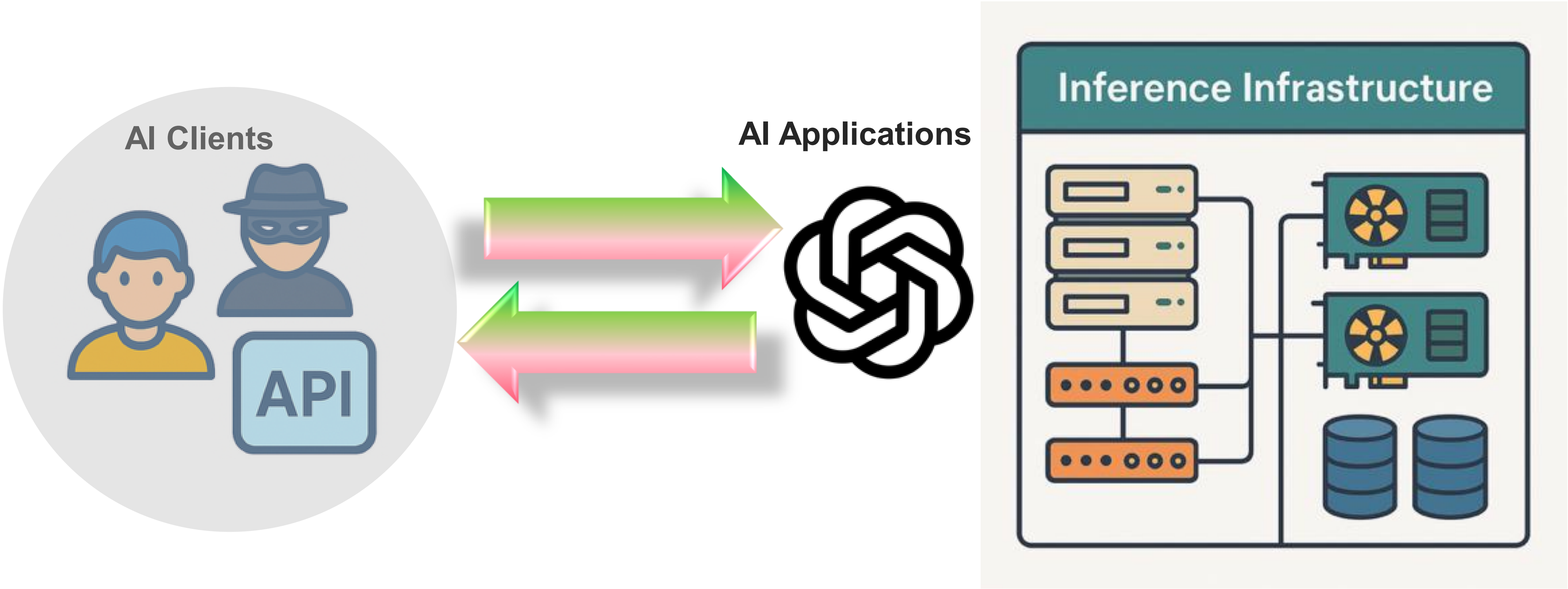
Remote DC: IPsec, GRE, VPN, TLS, MPLS, etc

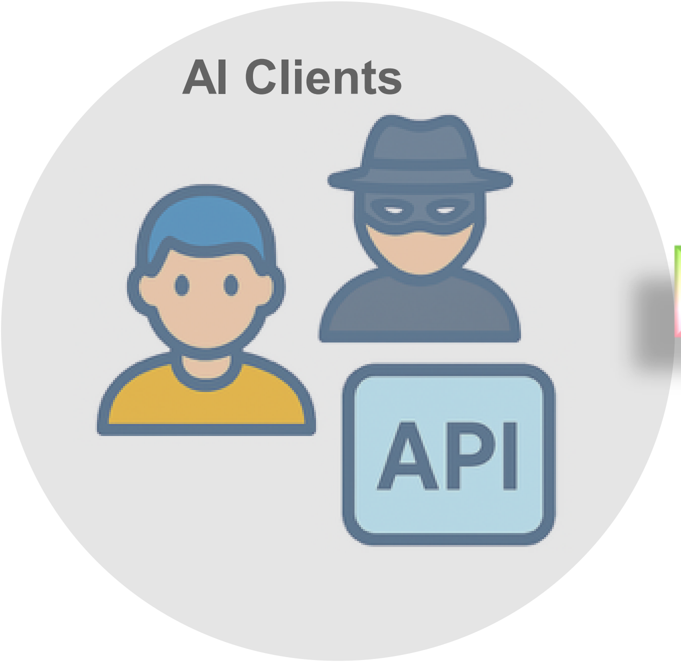
Technologies

SmartNICs, Smart Switches, L4-Loadbalancers, Gateways, Firewalls, WAN Optimizers

KEYSIGHT



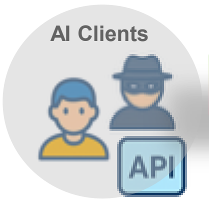

AI Clients
API
AI Applications
Inference Infrastructure



Application Frontend Technologies and Features

- Auth Gateway
- Load Balancer
- TLS termination/Re-Initiation
- Web App Firewalling
- Rate limits/Shaping

Usecases (Testing Load Balancers and WAFs)
HTTP/HTTPS KPI performances
TLS KPIs -1.2/1.3 -PQC -Various Cipher perf -Bulk throughput perf -Private Key exchanges
WAF Security Efficacy



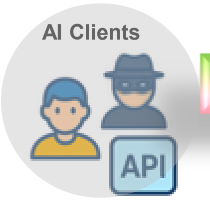
Application Frontend Technologies and Features

- Auth Gateway
- Load Balancer
- TLS termination/Re-Initiation
- Web Firewalling
- Rate limits/Shaping

AI Security Gateways/ Model Guardrails

- Prompt injection filtering
- PII redaction, enforcement
- Toxicity/Bias checks
- Hallucination checks

Usecases (Testing AI Security Request/Response)
Prompt Injection efficacy
Response blocking/Masking capabilities
Evasion detection capabilities
App KPI impacts of Security Policies



Application Frontend Technologies and Features

- Auth Gateway
- Load Balancer
- TLS termination/Re-Initiation
- Web Firewalling
- Rate limits/Shaping

AI Security Gateways/ Model Guardrails

- Prompt injection filtering
- PII redaction, enforcement
- Toxicity/Bias checks
- Hallucination checks

Inference Load-balancing(DPU)

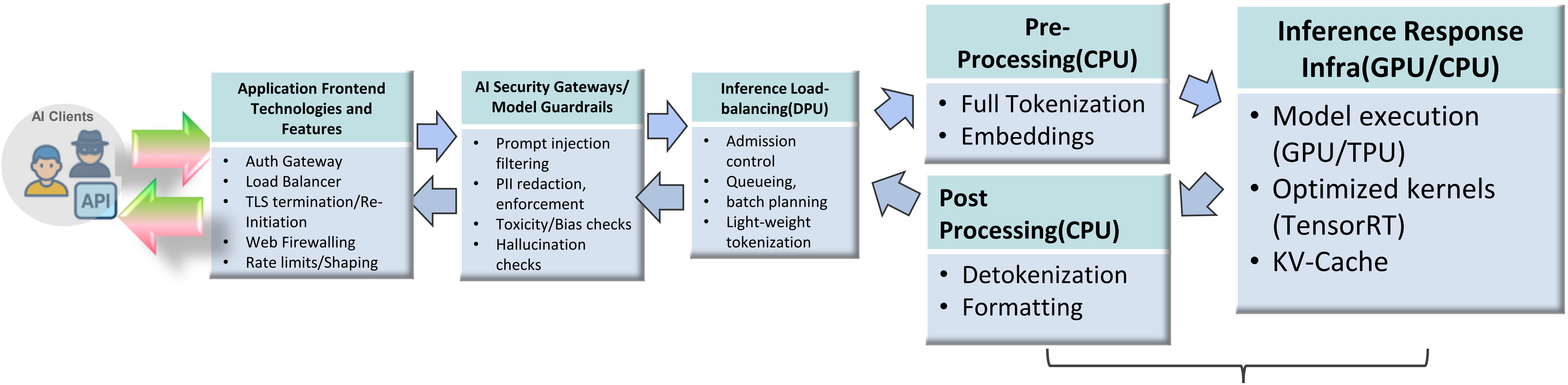
- Admission control
- Queueing
- Batch Planning
- Light-weight tokenization

Testing Pre-Inference Infrastructure(DPU)

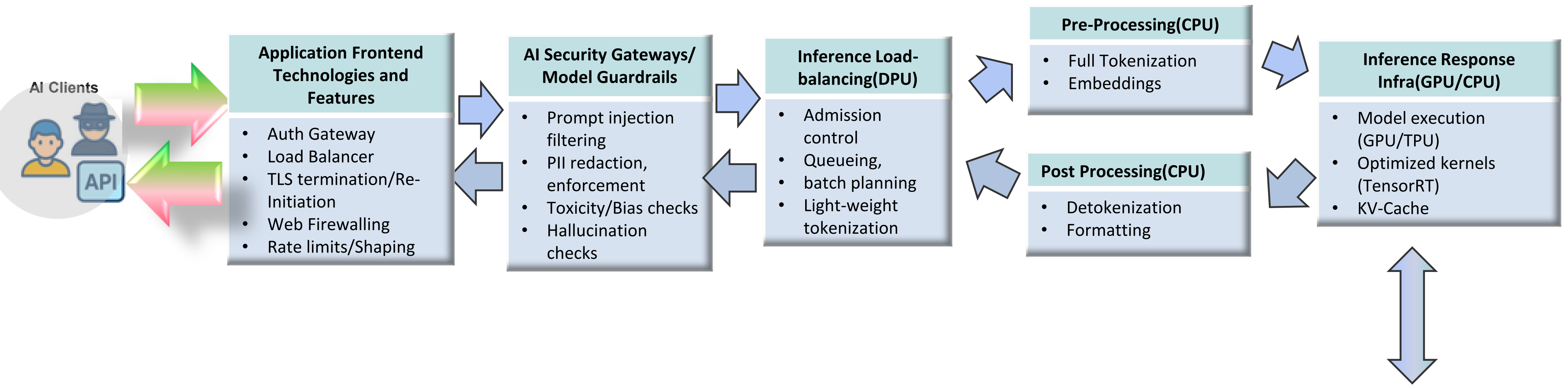
HTTP/GRPC traffic at Scale validation

Prompt Handling at scale

Authentications



Usecases at Processing and Inference Response layers	
Various Prompt Handling capacity at scale - Simple fetch prompts - Complex philosophical prompts - Multi-modal prompts - Mix of prompts - Cold/Warm/Hot Prompts for Cache testing - MCP/A2A/RAG executing prompts	Token Performance and Latency -Time to First Token, Inter-Token time, Token Latencies
	Token Rate, Max Tokens, Temperature
	GPU / CPU Infrastructure benchmark

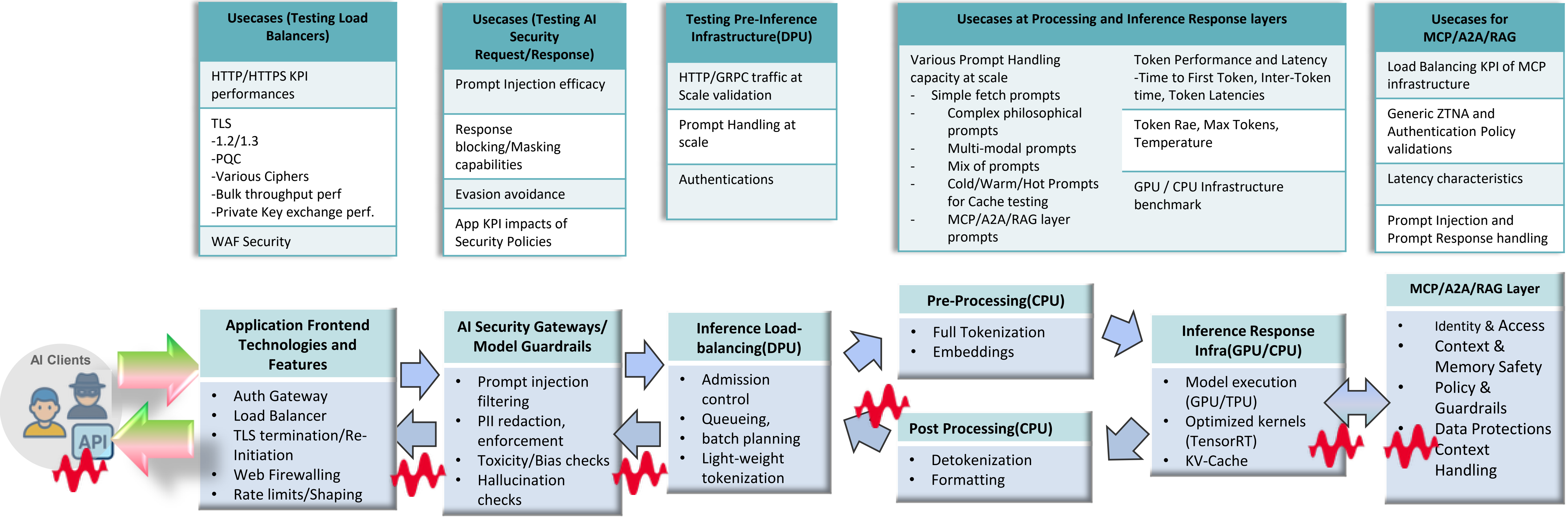


Usecases for MCP/A2A/RAG

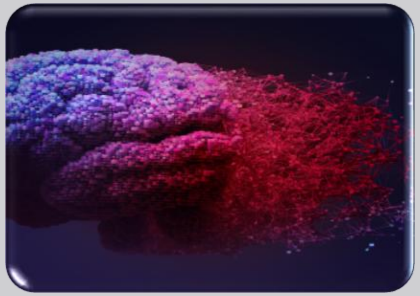
Usecases for MCP/A2A/RAG
Load Balancing KPI of MCP infrastructure(CC,CPS, Tput, TPS)
Generic ZTNA and Authentication Policy validations
Latency characteristics
Prompt Injection and Prompt Response handling

MCP/A2A/RAG Layer

- Identity & Access
- Context & Memory Safety
- Policy & Guardrails
- Data Protections
- Context Handling


Usecases (Testing Load Balancers)
HTTP/HTTPS KPI performances
TLS
-1.2/1.3
-PQC
-Various Ciphers
-Bulk throughput perf
-Private Key exchange perf.
WAF Security
Usecases (Testing AI Security Request/Response)
Prompt Injection efficacy
Response blocking/Masking capabilities
Evasion avoidance
App KPI impacts of Security Policies
Testing Pre-Inference Infrastructure(DPU)
HTTP/GRPC traffic at Scale validation
Prompt Handling at scale
Authentications
Usecases at Processing and Inference Response layers
Various Prompt Handling capacity at scale
- Simple fetch prompts
- Complex philosophical prompts
- Multi-modal prompts
- Mix of prompts
- Cold/Warm/Hot Prompts for Cache testing
- MCP/A2A/RAG layer prompts
Token Performance and Latency -Time to First Token, Inter-Token time, Token Latencies
Token Rae, Max Tokens, Temperature
GPU / CPU Infrastructure benchmark
Usecases for MCP/A2A/RAG
Load Balancing KPI of MCP infrastructure
Generic ZTNA and Authentication Policy validations
Latency characteristics
Prompt Injection and Prompt Response handling
AI Clients
API
Application Frontend Technologies and Features
• Auth Gateway
• Load Balancer
• TLS termination/Re-Initiation
• Web Firewalling
• Rate limits/Shaping
AI Security Gateways/ Model Guardrails
• Prompt injection filtering
• PII redaction, enforcement
• Toxicity/Bias checks
• Hallucination checks
Inference Load-balancing(DPU)
• Admission control
• Queueing,
• batch planning
• Light-weight tokenization
Pre-Processing(CPU)
• Full Tokenization
• Embeddings
Post Processing(CPU)
• Detokenization
• Formatting
Inference Response Infra(GPU/CPU)
• Model execution (GPU/TPU)
• Optimized kernels (TensorRT)
• KV-Cache
MCP/A2A/RAG Layer
• Identity & Access
• Context & Memory Safety
• Policy & Guardrails
• Data Protections
• Context Handling

AI Development Phases



AI Factory Infrastructure
- GPU
- Servers
- Interconnects
- Data Center

Edge AI
- AI Devices
- AI Smartphones
- AI Sensors

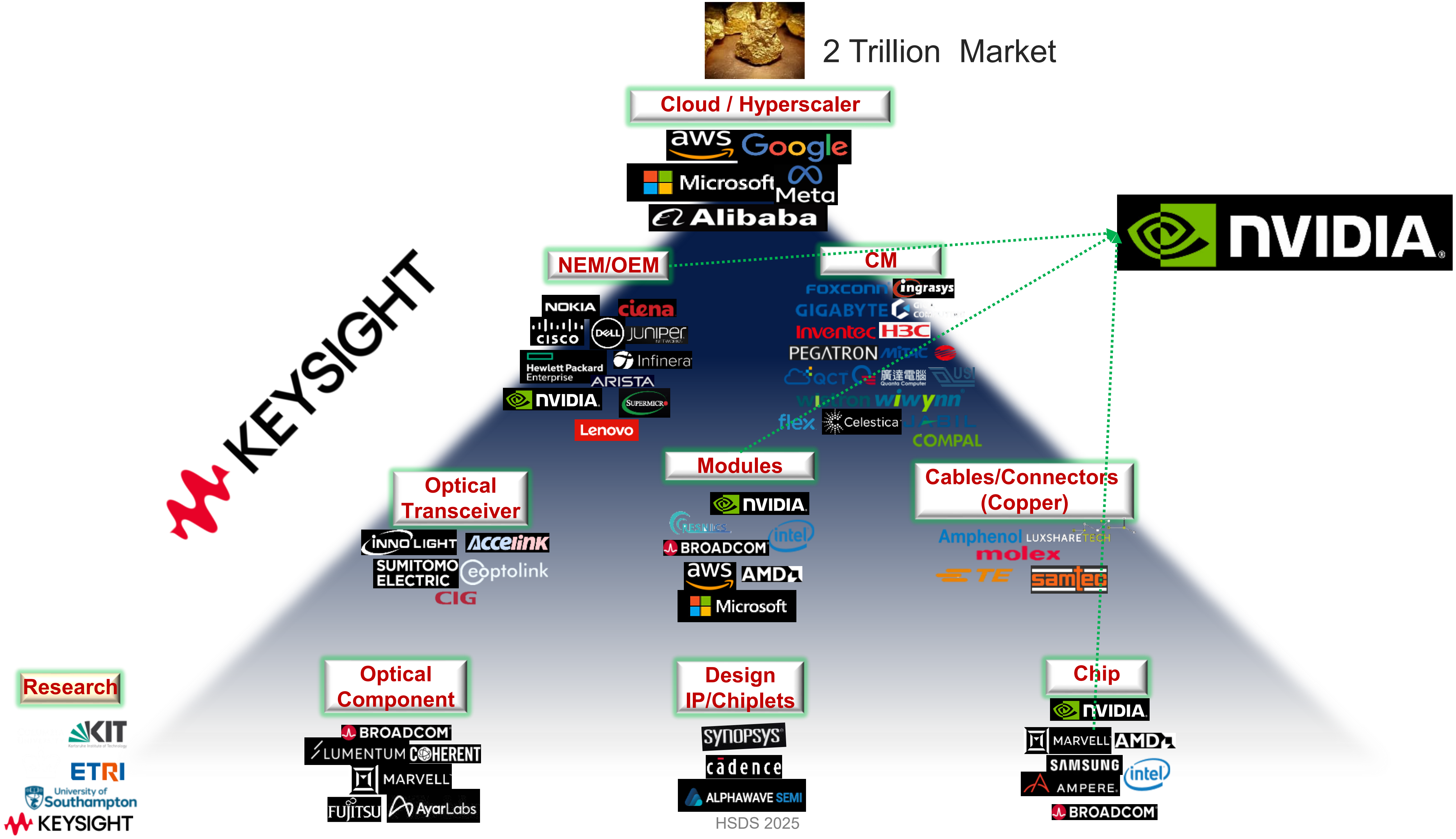
AI Applications
- Agents
- AI Personal Assistants

Physical AI
- Self Driving Cars
- Humanoid Robot

Integrated AI Systems
- Defense systems
- Smart Grid
- Intelligent Transport
- Space Exploration

Foundation Models

Keysight Value: Enabling the AI Industry across the supply chain



Keysight AI Infrastructure Solution Components

Keysight solutions cover entire workflow

Simulation | Emulation | Validation | Testing | Manufacturing

Applications

Complementary adjacencies aligned to customers' expanding design & test needs

+

Protocol

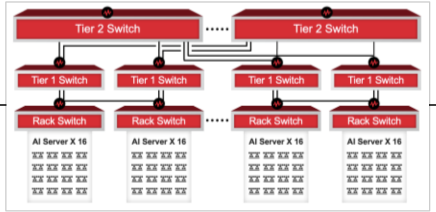
Protocol test and emulation of real-world systems in the lab

+

Physical

High-performance measurement solutions across analog and digital domains

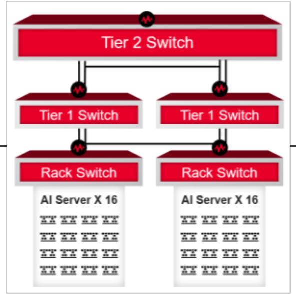
Network Fabric, Systems and AI Models



AI Data Center Solutions

Workload Emulation for Network Fabric and System Benchmarking

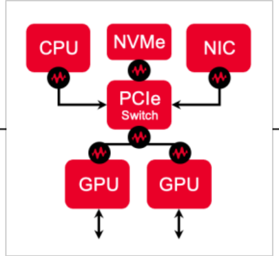
Servers, Switches, Routers, Devices



High-Speed Networking, Computing & Security

Comprehensive offerings for components and system interoperability

Chips, Interconnects and Modules



High-Speed Design & Test

State-of-the-art AI infrastructure characterization for switches, interfaces, CPUs, GPUs, accelerators, and transceiver modules

AI as it appears to Evolve

AI Stages	**Artificial Narrow Intelligence (ANI)** Execute specific focused tasks, without the ability to self-expand functionality	**Artificial General Intelligence (AGI)** Perform broad tasks, reason, and improve capabilities comparable to humans	**Artificial Super Intelligence (ASI)** Demonstrate intelligence beyond human capabilities
Timing	**Today**	**About 2030?**	**Soon after AGI**
Implications	Outperform people in specific repetitive functions, such as driving, medical diagnosis, and financial advice	Compete with humans across all endeavors, such as earning university degrees and convincing people that it is human	Outperform people, helping to achieve societal objectives, or threatening human race
	Jobs enhanced	Jobs at risk	Humanity at risk?

GPT defined
NLU defined

KEYSIGHT

Source: Lawtomated

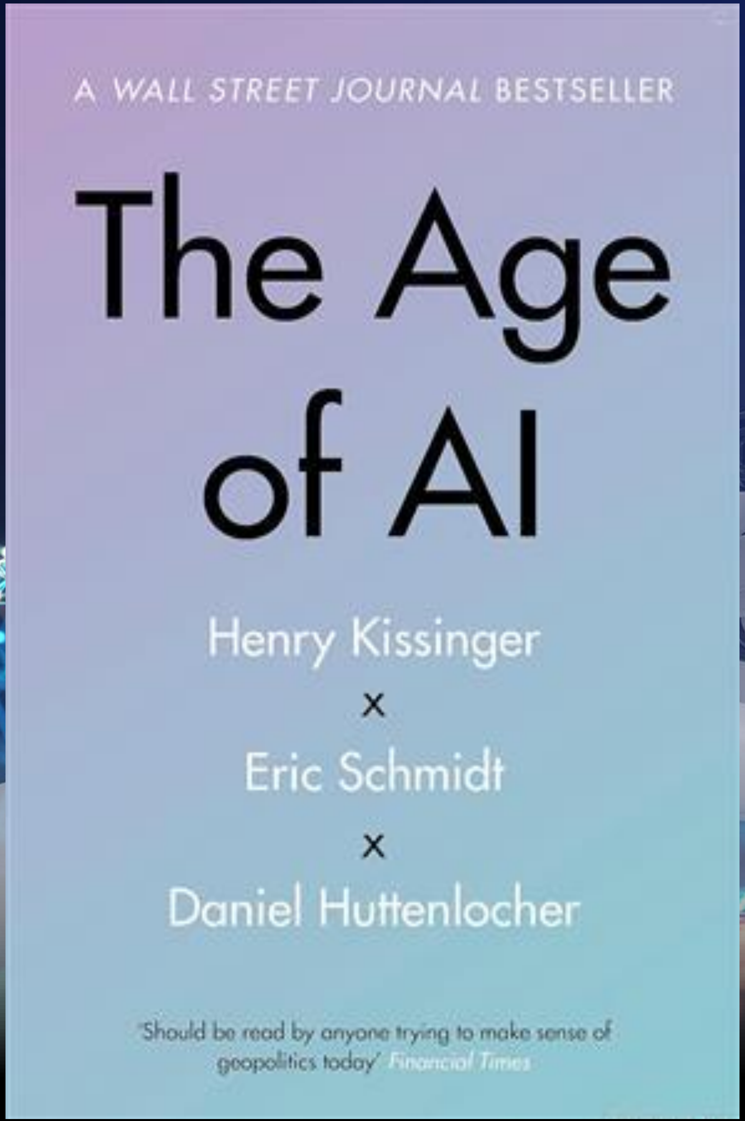
AI is Already Here, Pervading…..

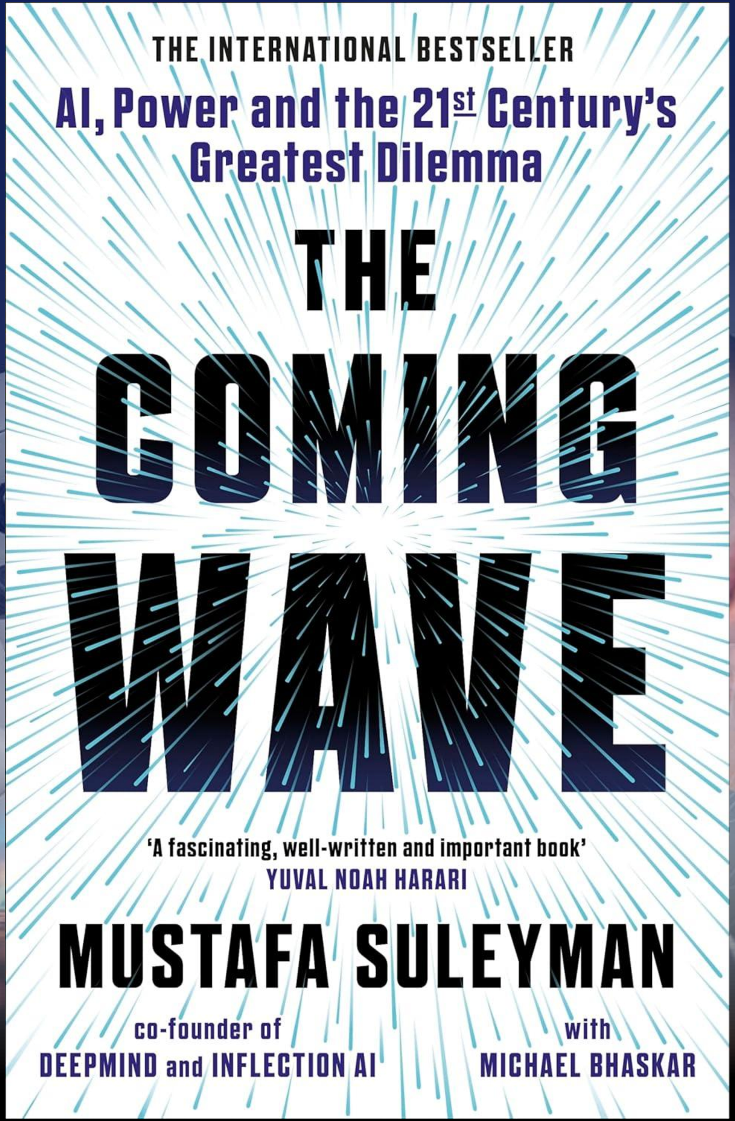
Race has
already
begun

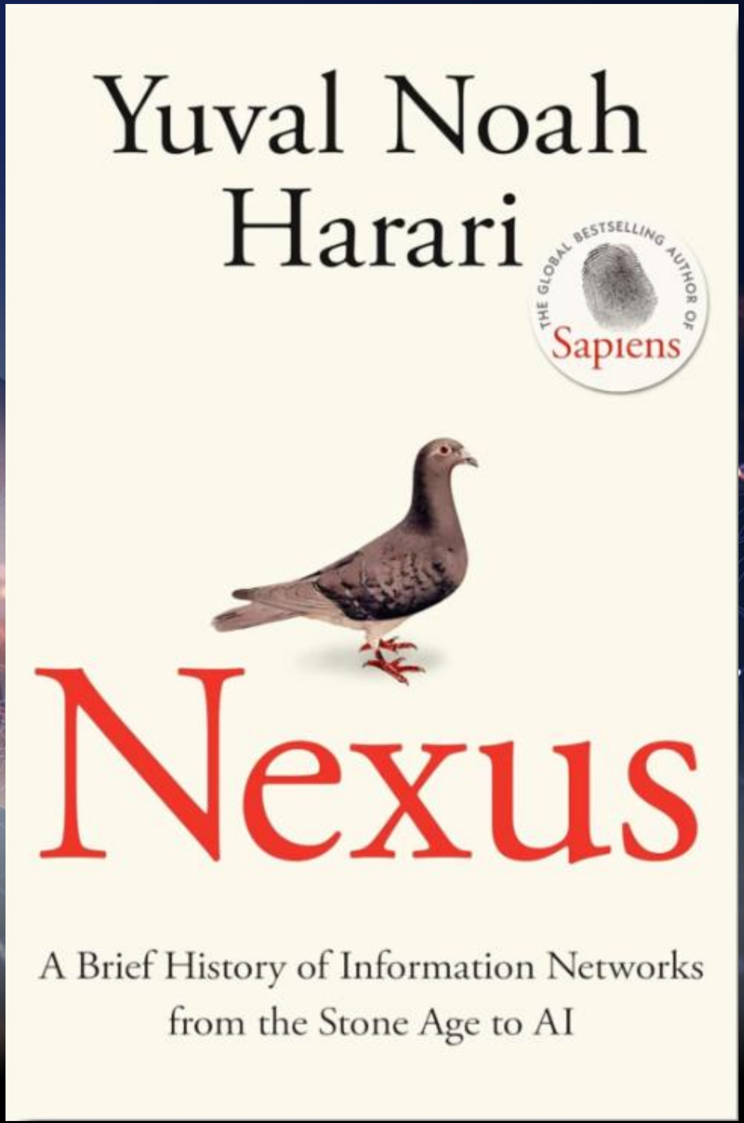
Competition is
fierce
and vehement

Upskill to outperform in AI world

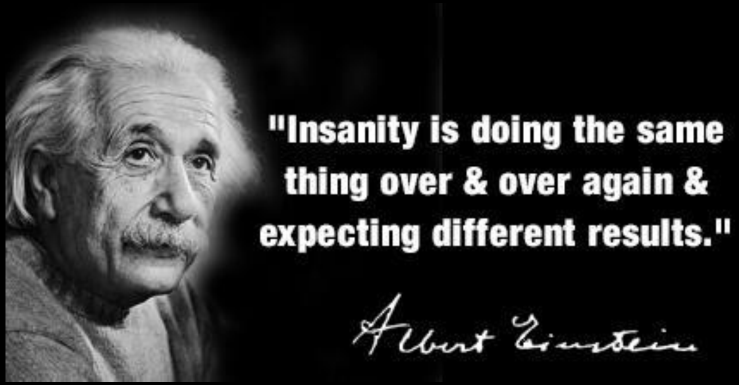
Suggested Reading………

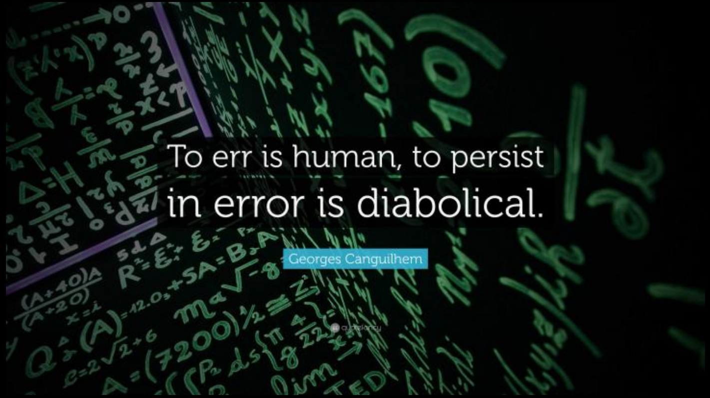


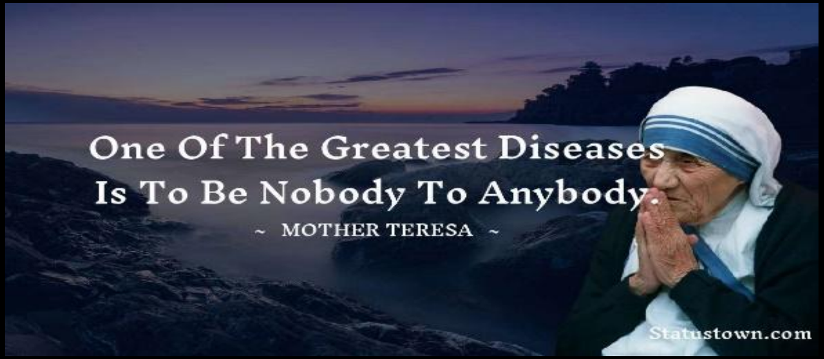


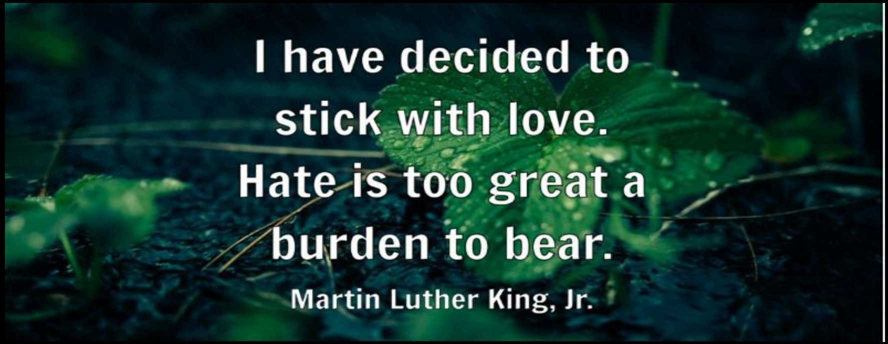


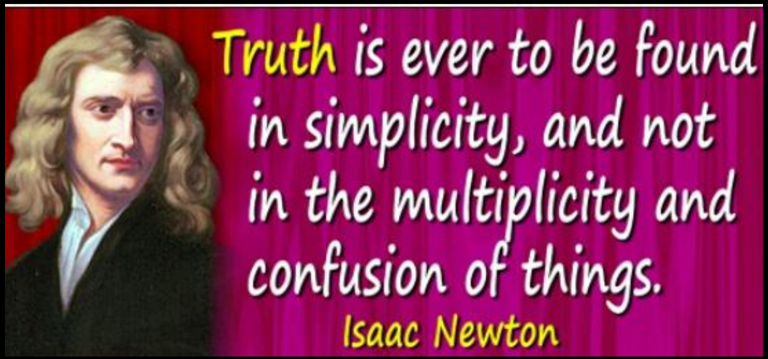
Some of my favorite quotes











"The good thing about science is that it's true whether you believe in it or not"

Neil deGrasse Tyson

KEYSIGHT