SNIA DEVELOPER CONFERENCE

SDC 25

By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Assessing AI storage communication performance at scale

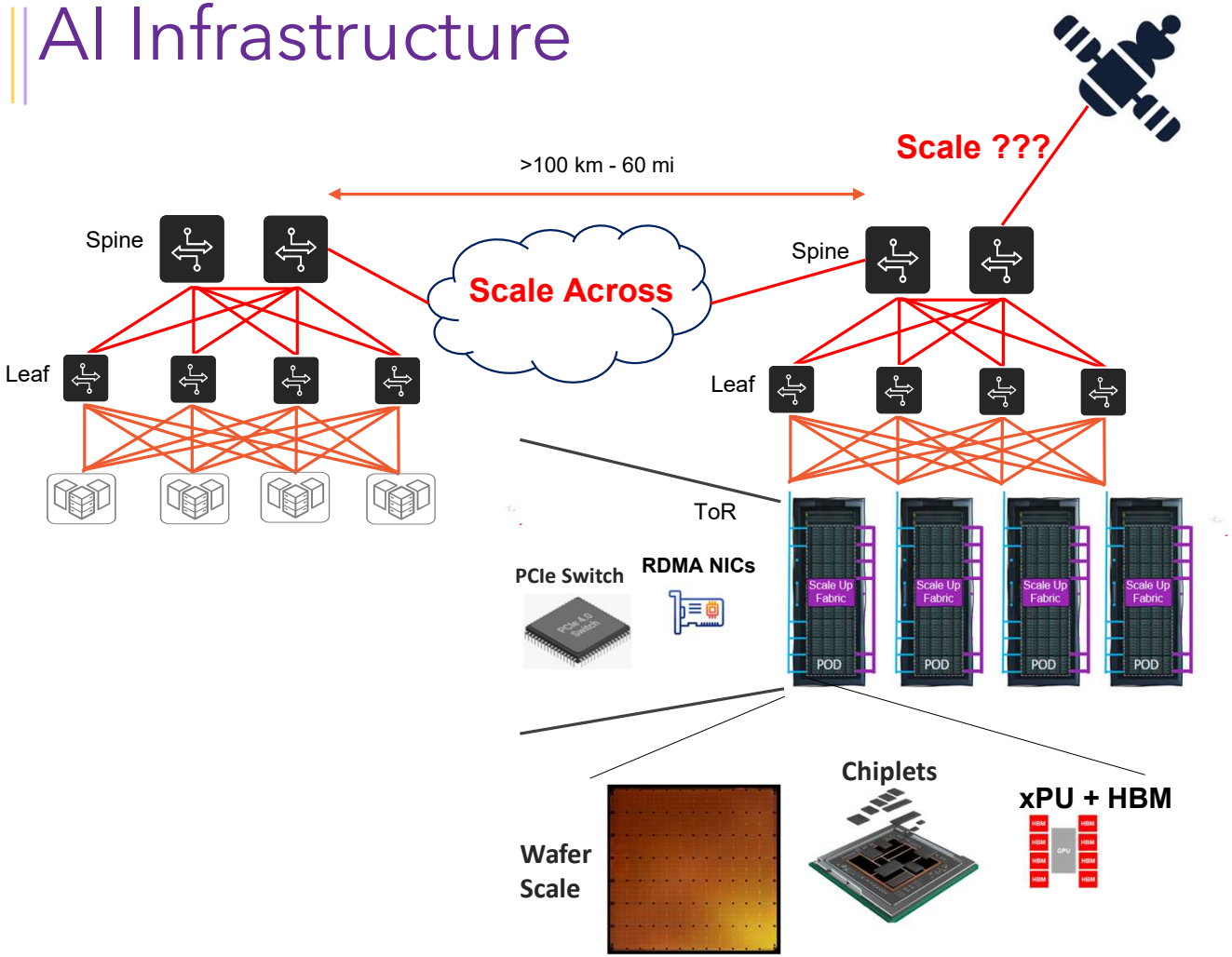
Venkat Pullela, CTO, Networking, Keysight
Cliff Tavares, Sr. Dir Engineering, Keysight

KEYSIGHT www.sniadeveloper.org

Agenda

- Scale in/up/out/across
- On-prem, cloud, remote storage
- Training, Inference, Ingestion, Checkpointing
- AI – Unique, Stateful, reliable transport
- Testing at all levels

AI Infrastructure



Scale Out
InfiniBand, Ultra Ethernet
Congestion Protocols

500m - 0.31 mi

50m – 164 ft

Scale Up
Nvlink, Infiniband,
UALink, SUE …

3 m - 10 ft

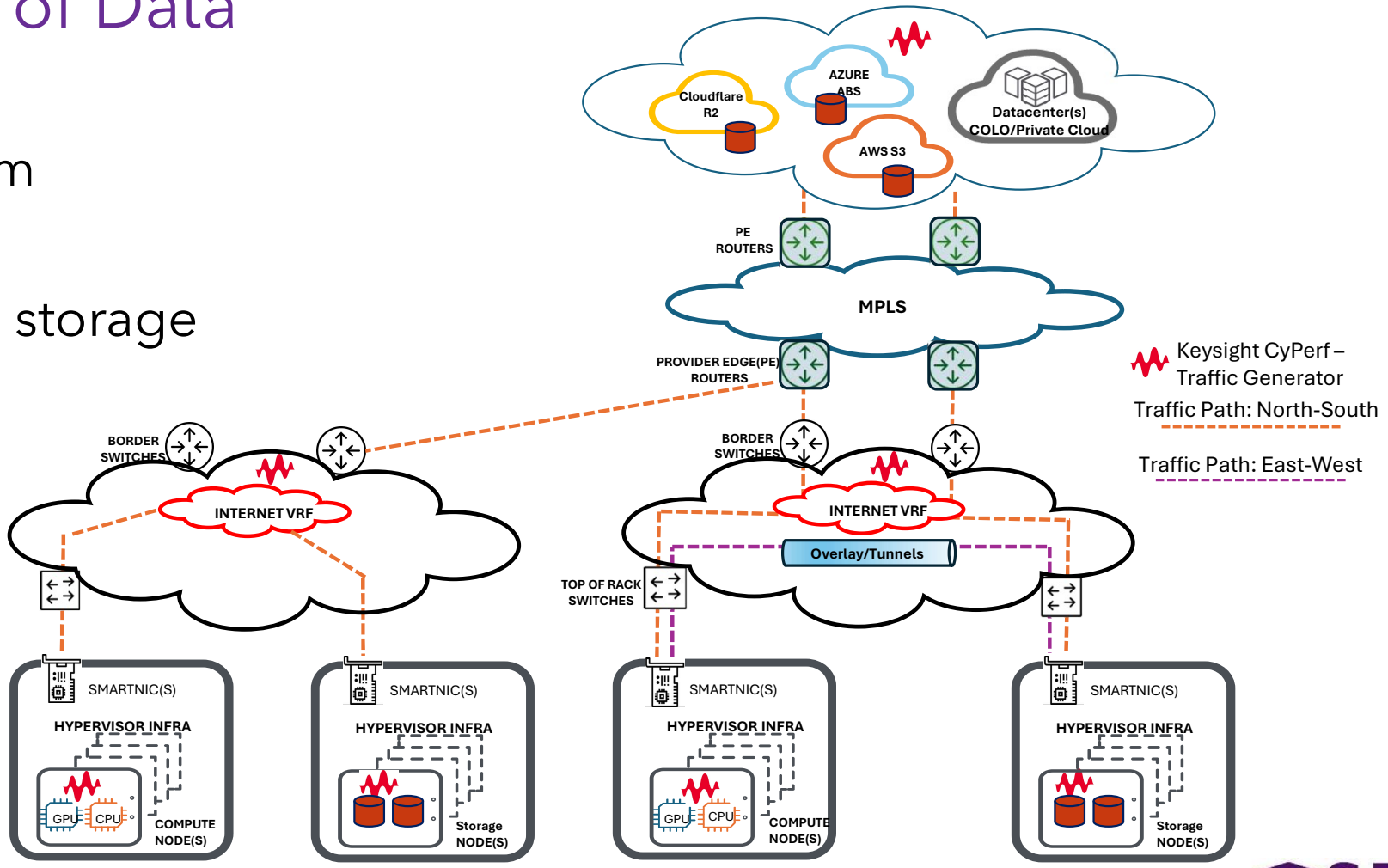
Scale In
NoC - D2D, AXI/CHI
UCIe/BoW

30 cm -.12 in

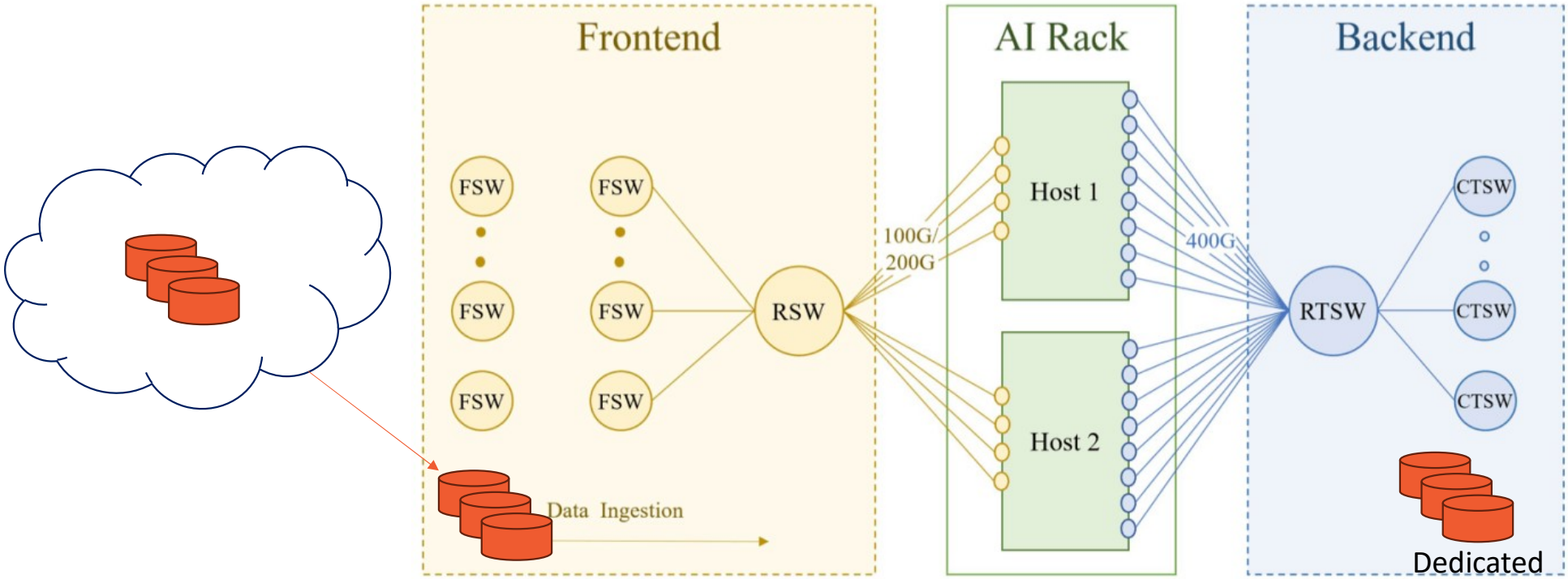
1 mm -.04 in

Source of Data

- On-prem
- Cloud
- Remote storage



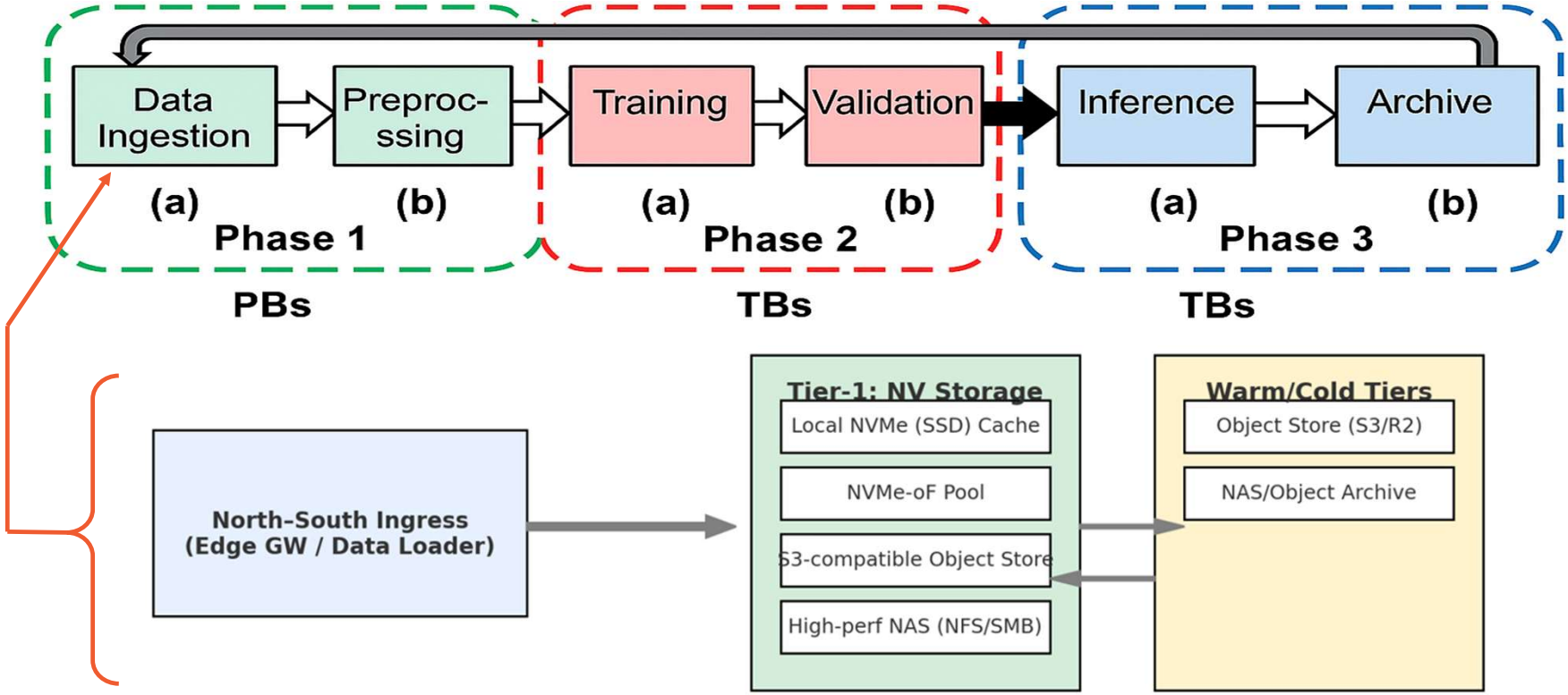
Where is the storage



Picture Credit : @Scale Meta

Data Movement

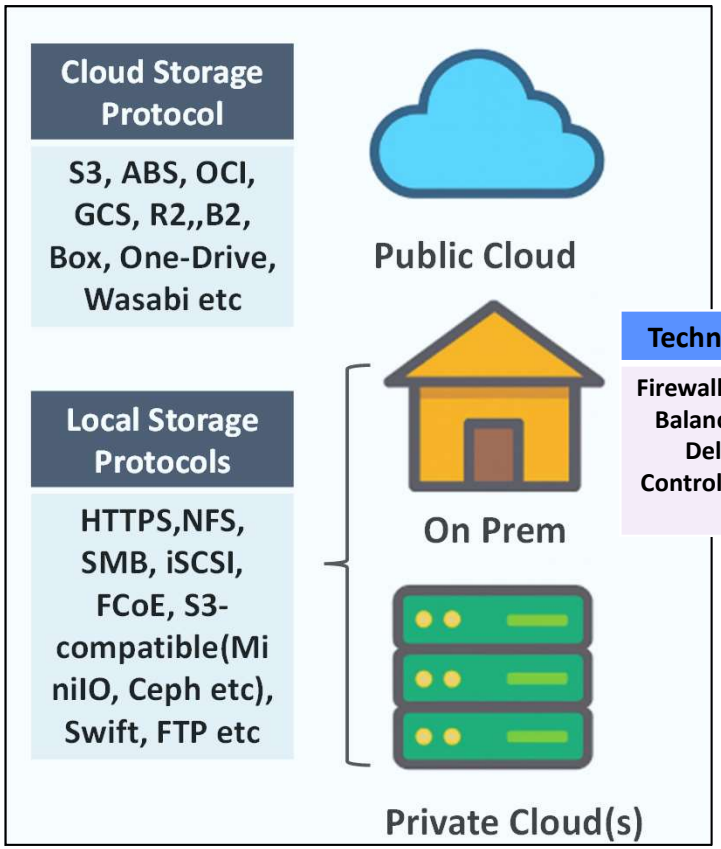
Picture Credit : Pratik Mishra, AMD

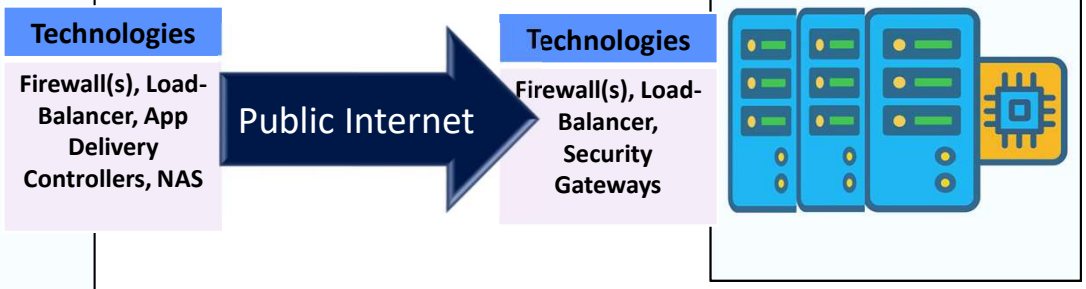


Data Lands in AI DC NV Storage Tier

Remote Storage to Datacenter

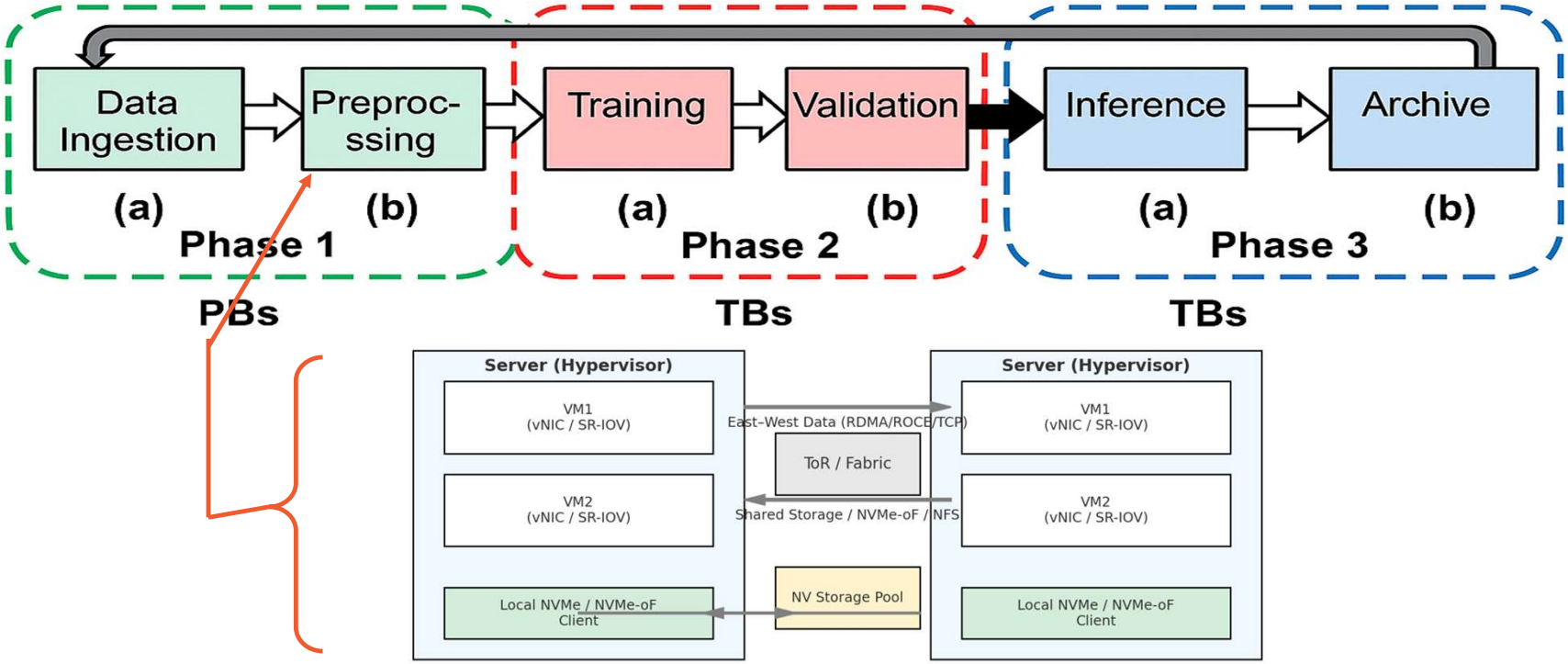
Remote Storage to Datacenter Validation Use Cases
Raw Stateful Performance KPIs
Storage Protocol Scalability
Small App Transaction Scale
Max App Bandwidth Scale
Max and Median Latency
KPI variations Over Time
Encryption/decryption Perf
Packet Drops and recovery
Choking Points
Security Efficacy
Elephant Flow Performance





Data Movement

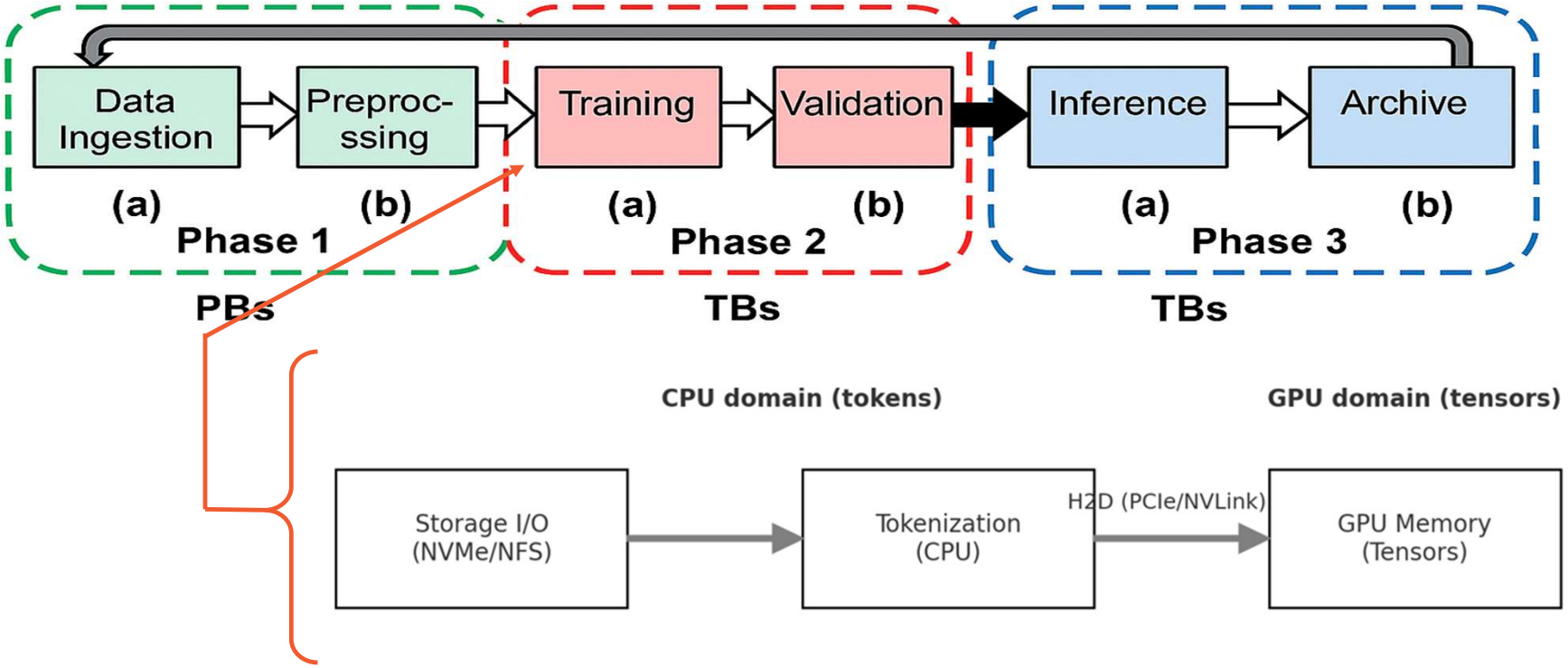
Picture Credit : Pratik Mishra, AMD



Improve data quality, remove errors, noise, inconsistencies to ensure accuracy
Text extraction in images, metadata/context extraction, chunking, RAG

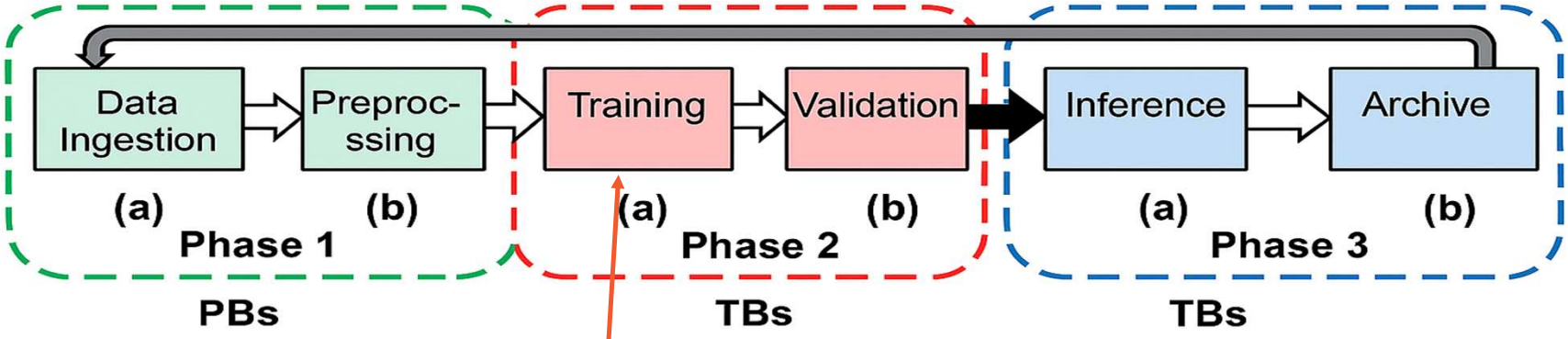
Data Movement

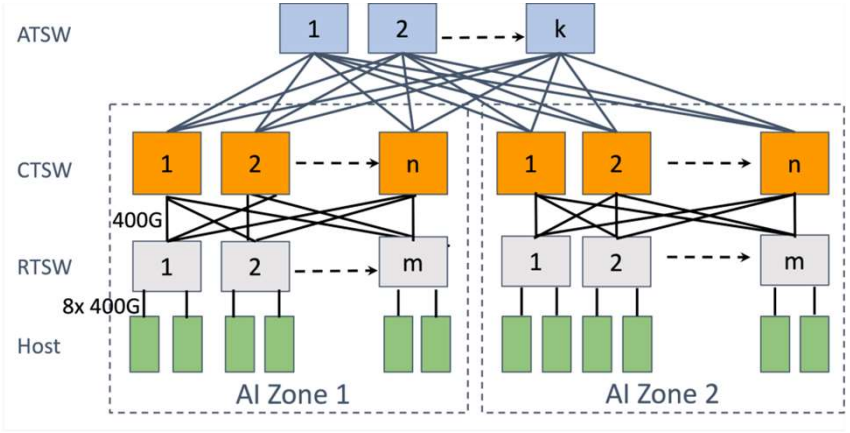
Picture Credit : Pratik Mishra, AMD



Training

Picture Credit : Pratik Mishra, AMD

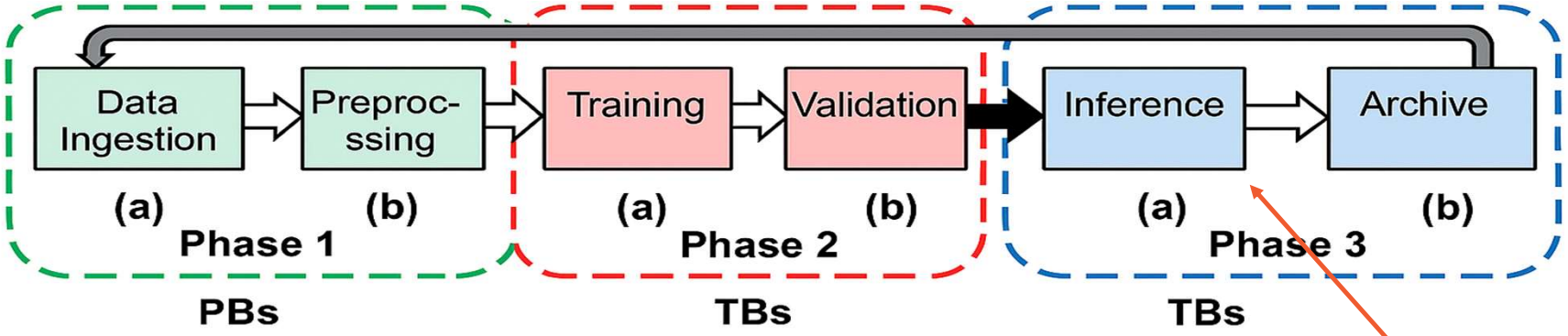


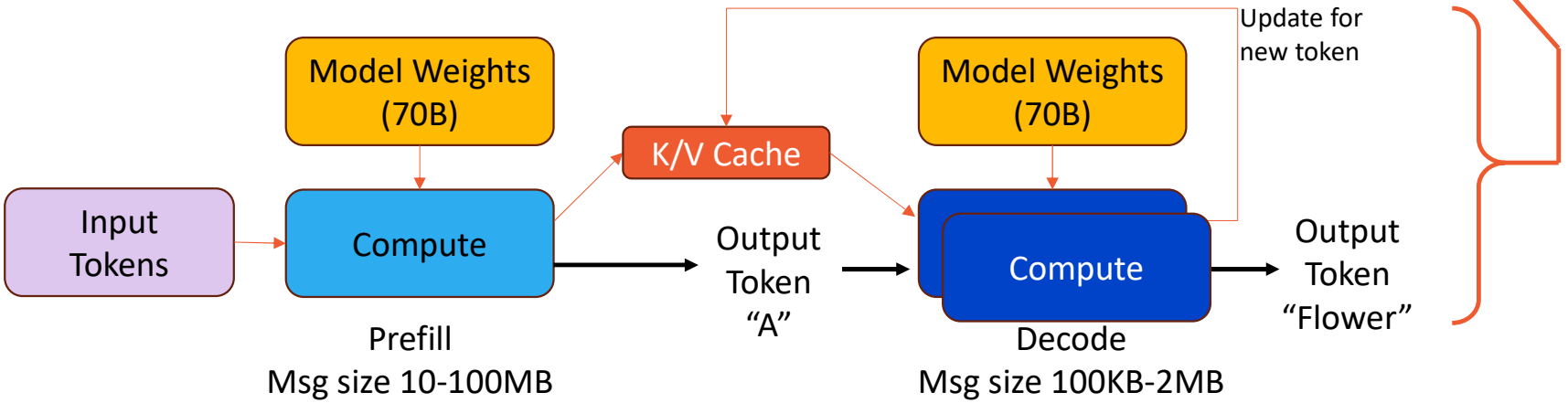


Picture Credit : @Scale Meta

Inference

Picture Credit : Pratik Mishra, AMD





How is AI unique

- AI is a system design problem
- Even components needs to be tested at Scale
  - Context, Events, Triggers
  - How does it perform at scale?
- Bringing System Test to pre-silicon – Shift left
- Design to Deployment Continuum
  - Pre-Silicon to Post-Silicon is just the first step

Testing at all layers

Applications - Capture and replay application patterns

Collectives – Visualize Cluster level communications and compute

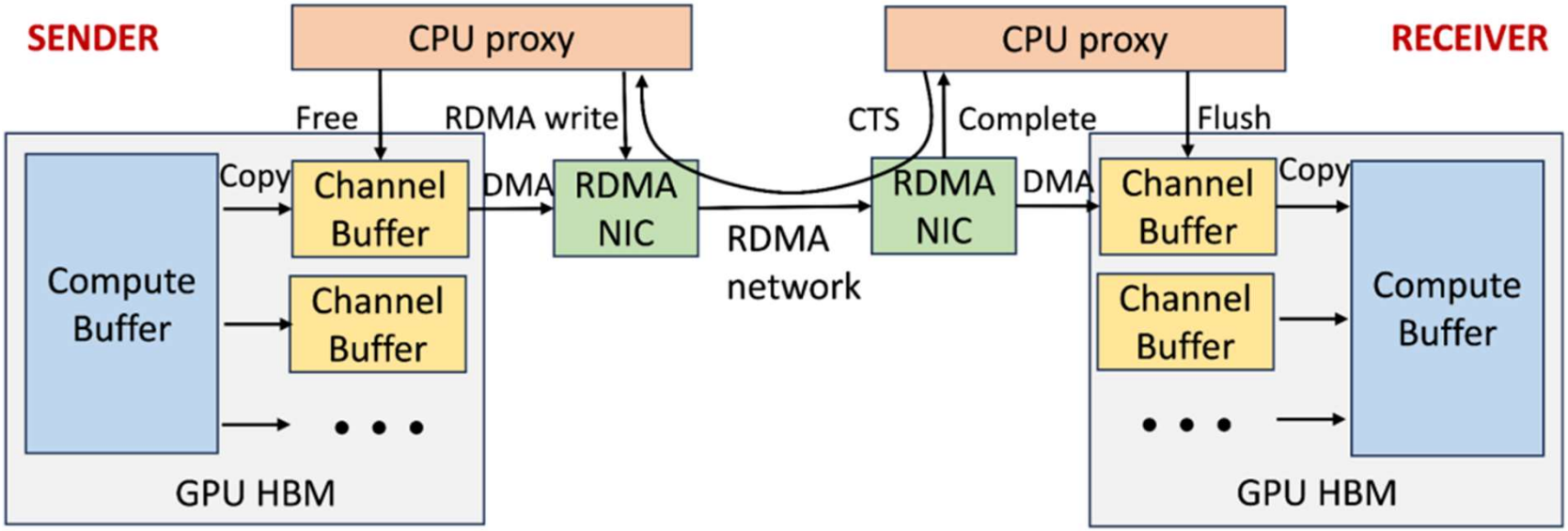
Transport – Stateful, Reliable, Congestion control RoCEv2, UEC

Packet – Load balancing and Performance

Link – Error detection and correction, LLR, CBFC

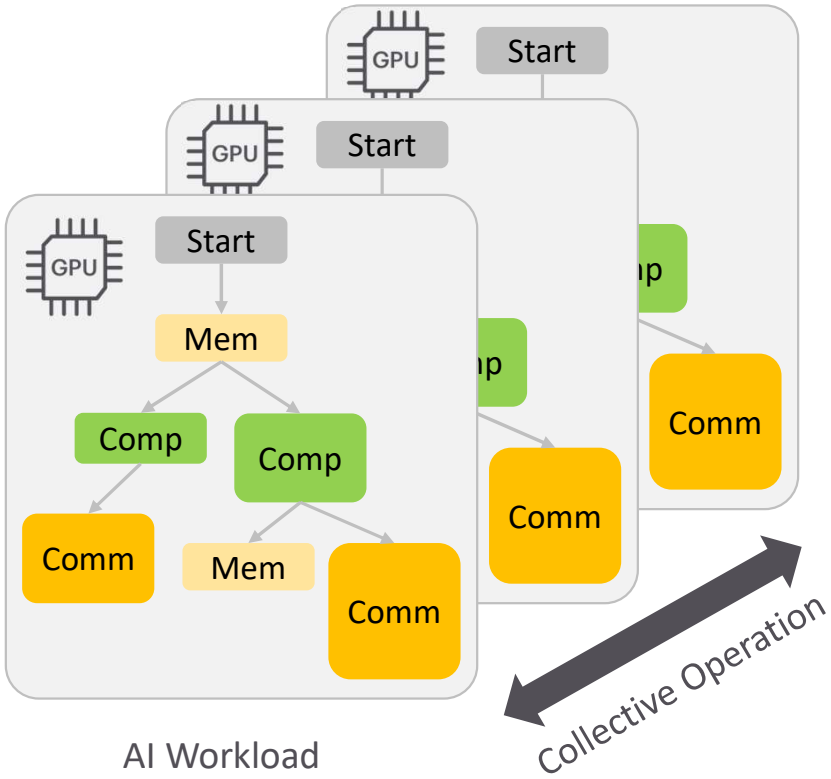
KEYSIGHT SDC 25

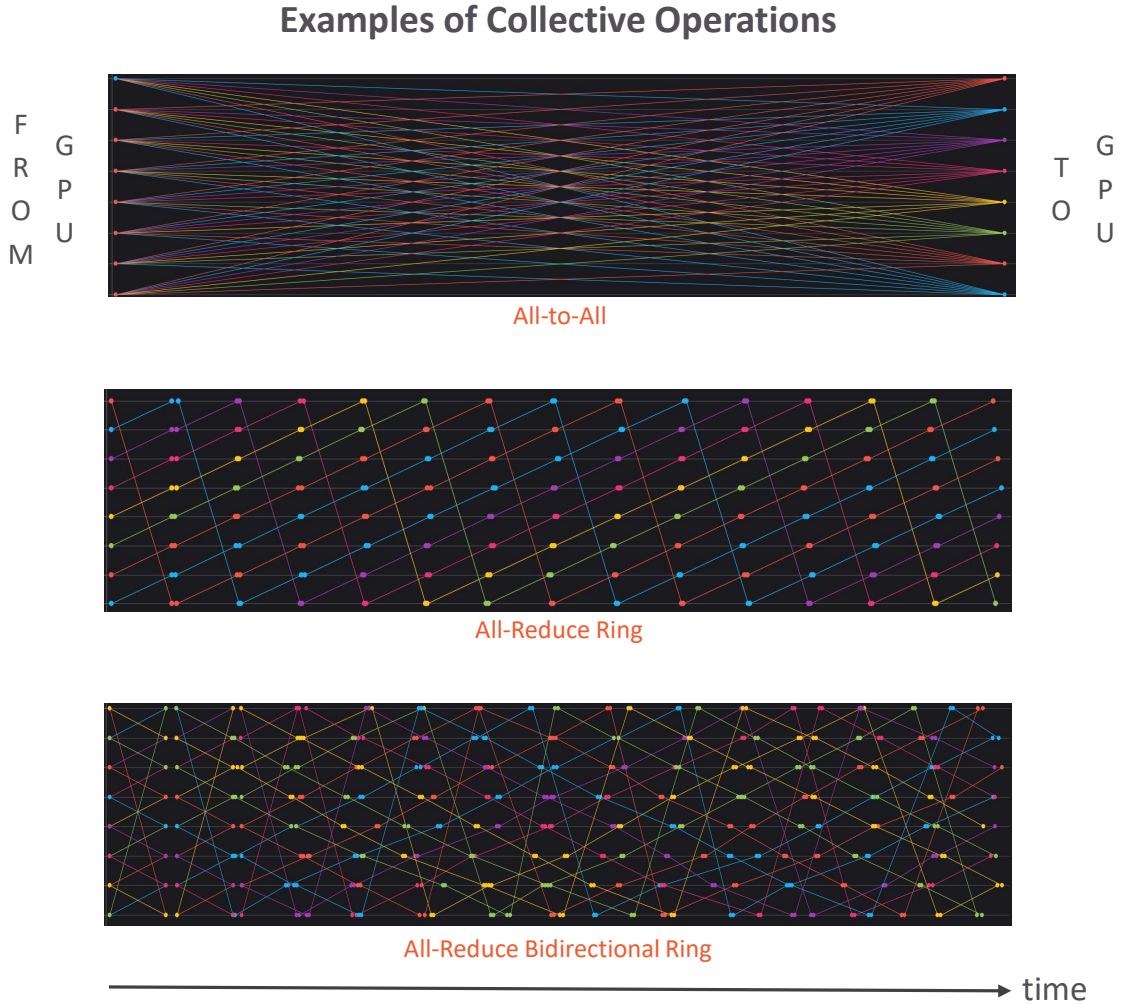
Transport



Picture Credit : @Scale Meta

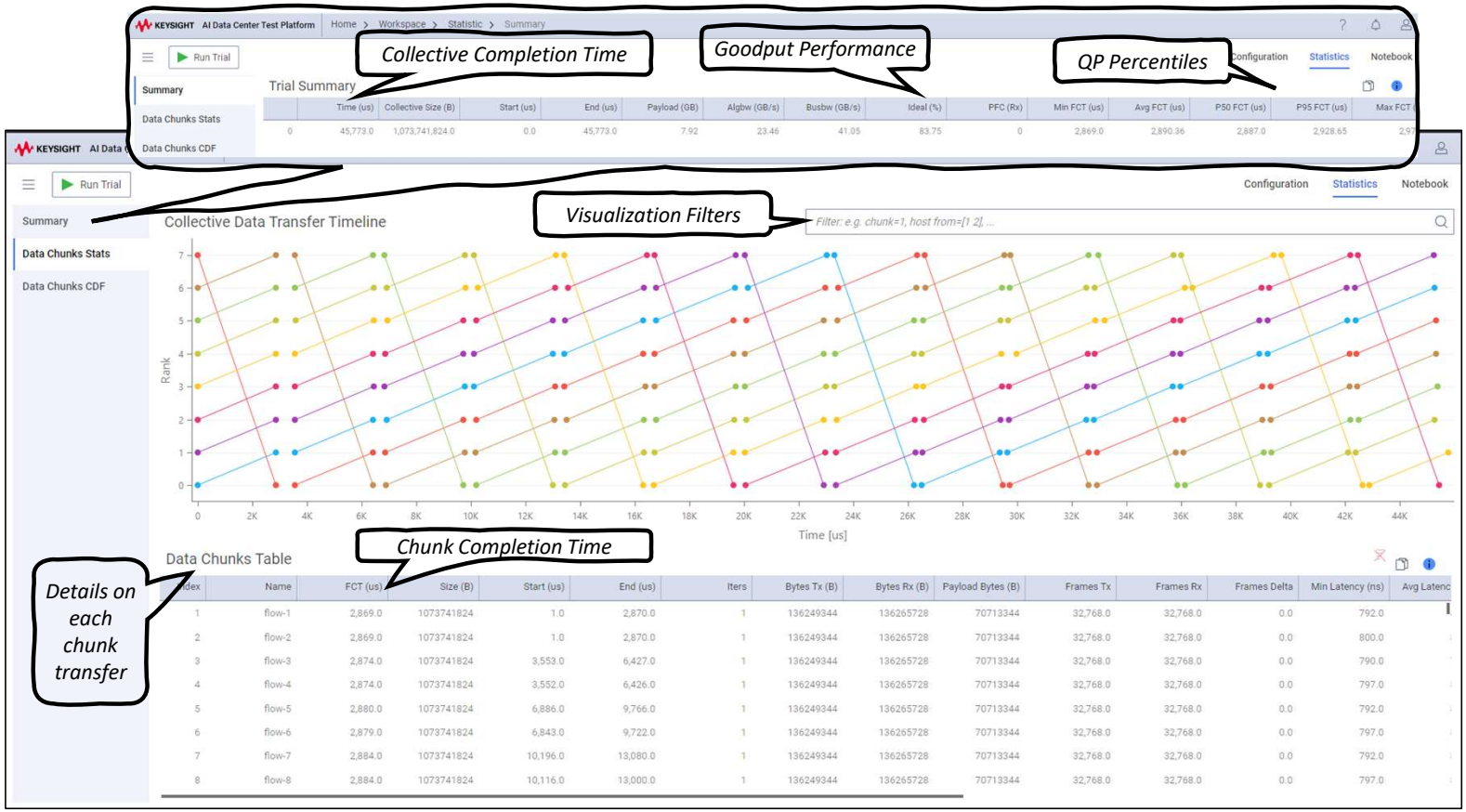
Collectives



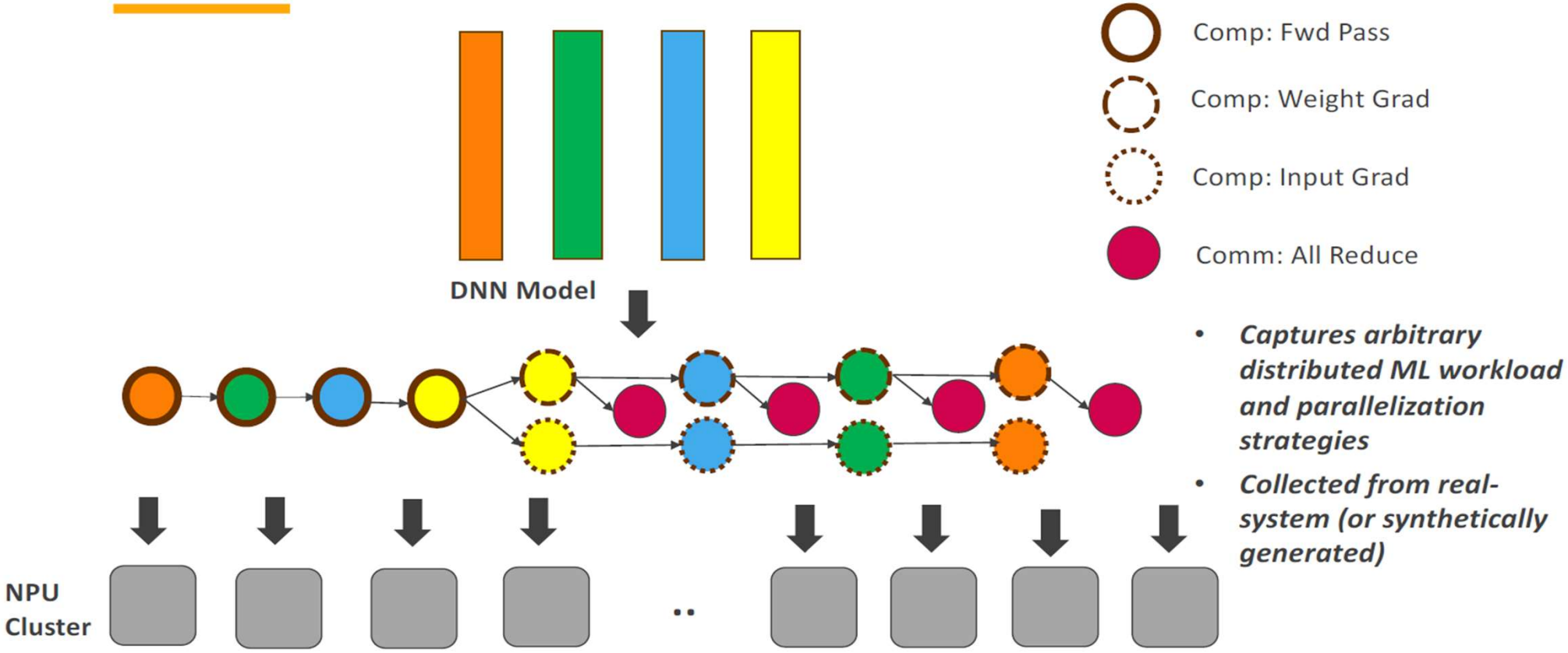


Collective Communications

Measurements



AI Application



Chakra

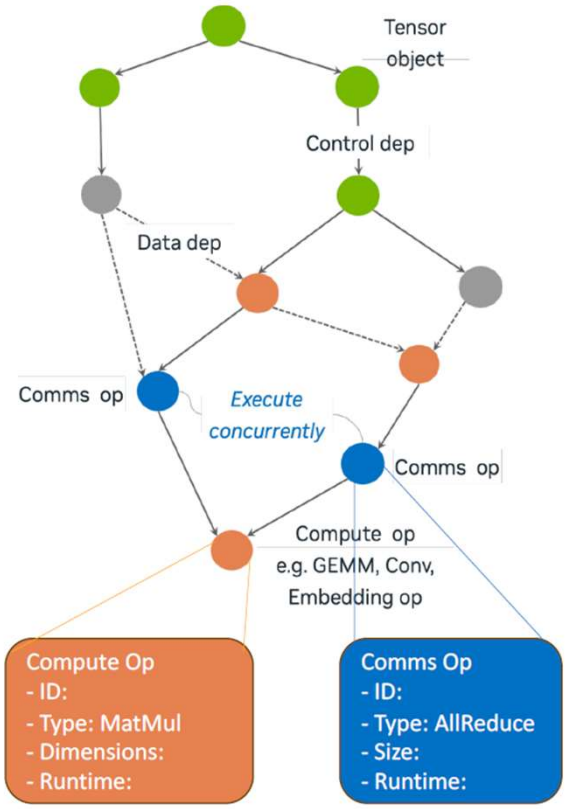
ML Commons

Extensible and standardized graph format to represent AI workloads

- **Nodes**: primitive operators and tensor objects with attributes and timing
- **Edges**: data and control dependency

Benefits

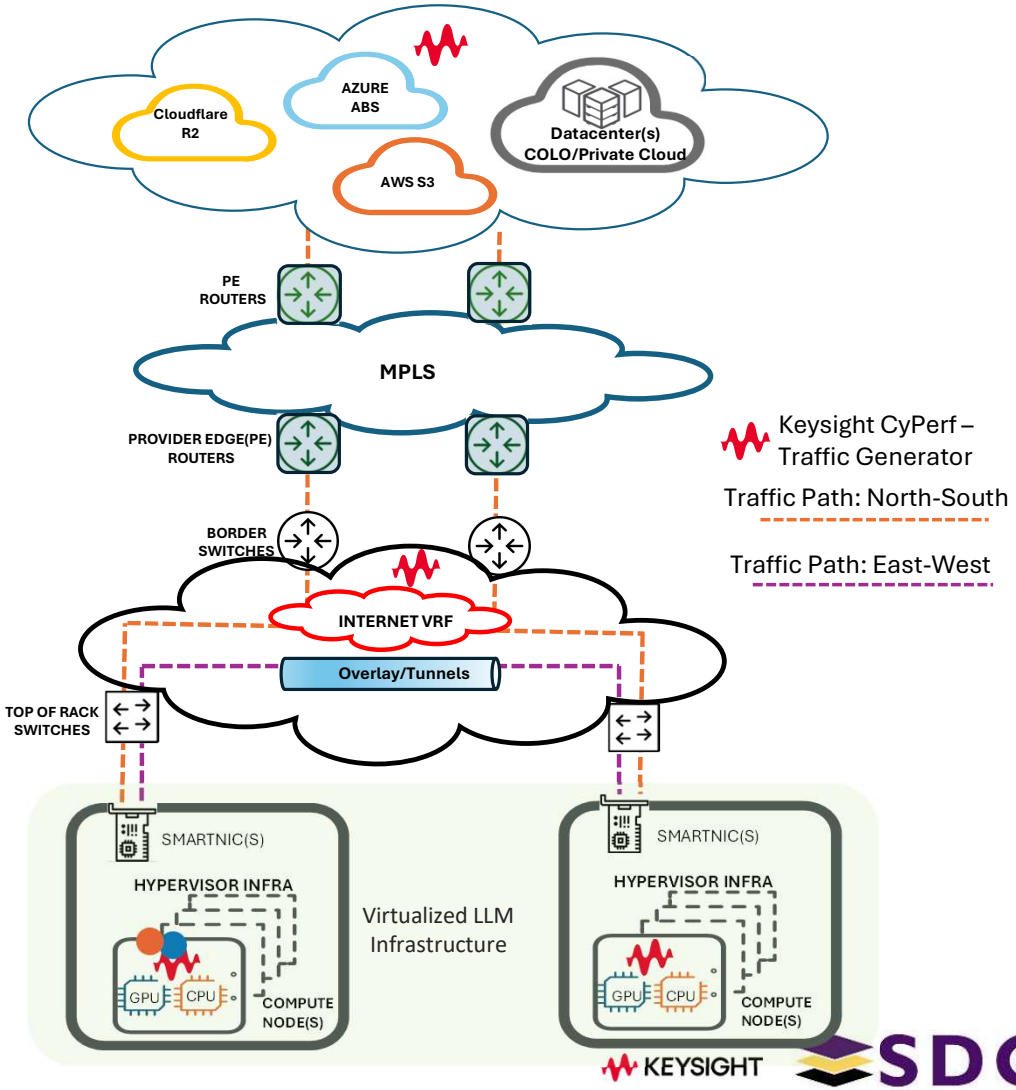
- Isolate comms and compute operators
- Operator, dependencies, and timing for replay, simulation, and analysis
- Flexible to represent both workloads and collective implementations
- Graph transformations to obscure sensitive IP



KEYSIGHT SDC 25

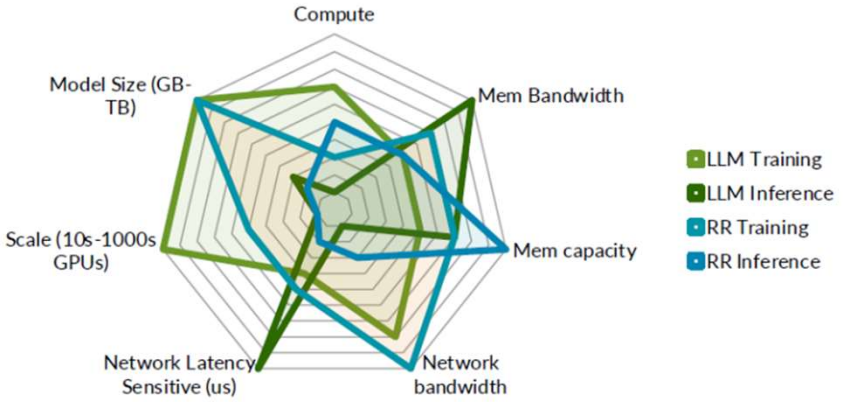
Validating AI Infra

- Validate AI rack, modular DC
- Integration of AI infra with hybrid cloud
- Capacity and Capability to host Multiple workloads at the same time
- Multi-tenancy - Impact of noisy neighbor, adversarial workloads
- Orchestration - Optimization based on Power, JCT, custom SLAs, Internal/Presidential vs. external
- LLM Agent capacity and security

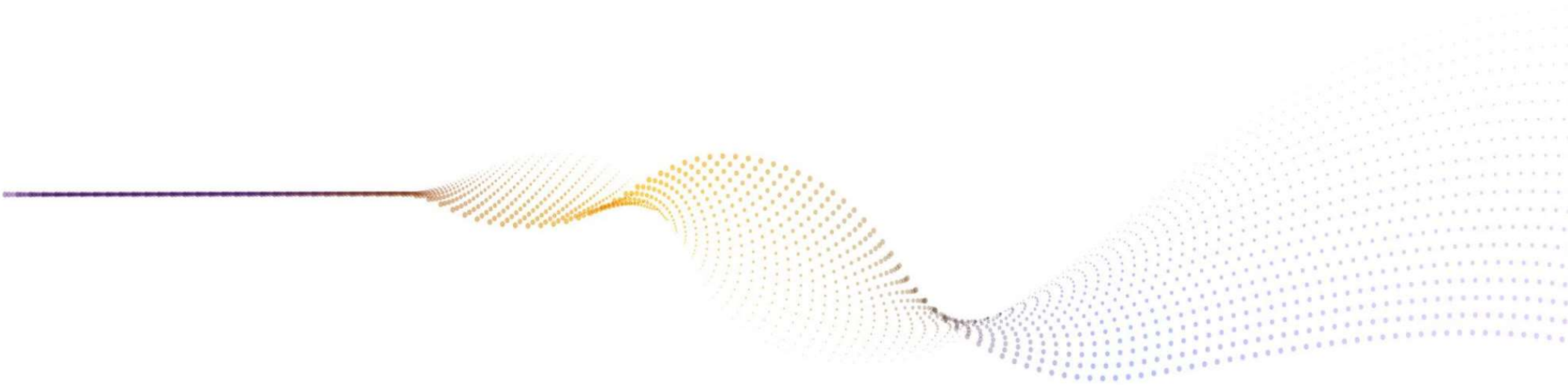


Validating AI Infra

- East-West Traffic Testing (For AI Training Efficiency)
  - Performance - Transactions/s, Connections/s, Throughput, IP Scale. Time to first token
  - Latency Characteristics, Encapsulations, RFC 2544
- North-South Traffic Testing (For Real-Time AI Inferencing)
  - NAT / Proxy, Encryptions, Resiliency
  - Applications - HTTP, Database, S3, R2, ABS, Storage, Video/Voice
  - Apps+ Attacks
- Test real-world cyber threats to remain operational under stress
  - GPU/CPU, memory, network utilization



Different workloads stress Infrastructure in different dimensions



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

