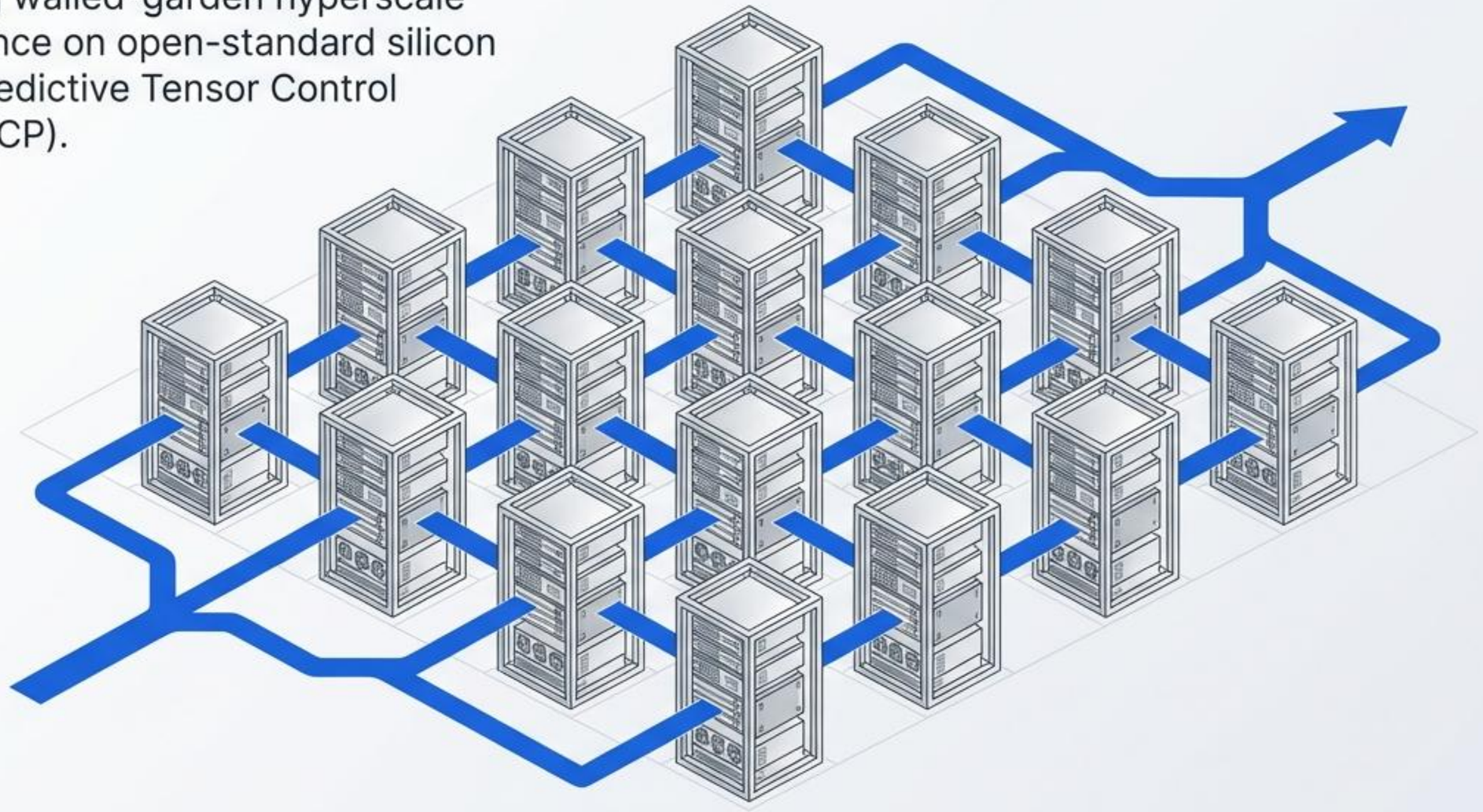
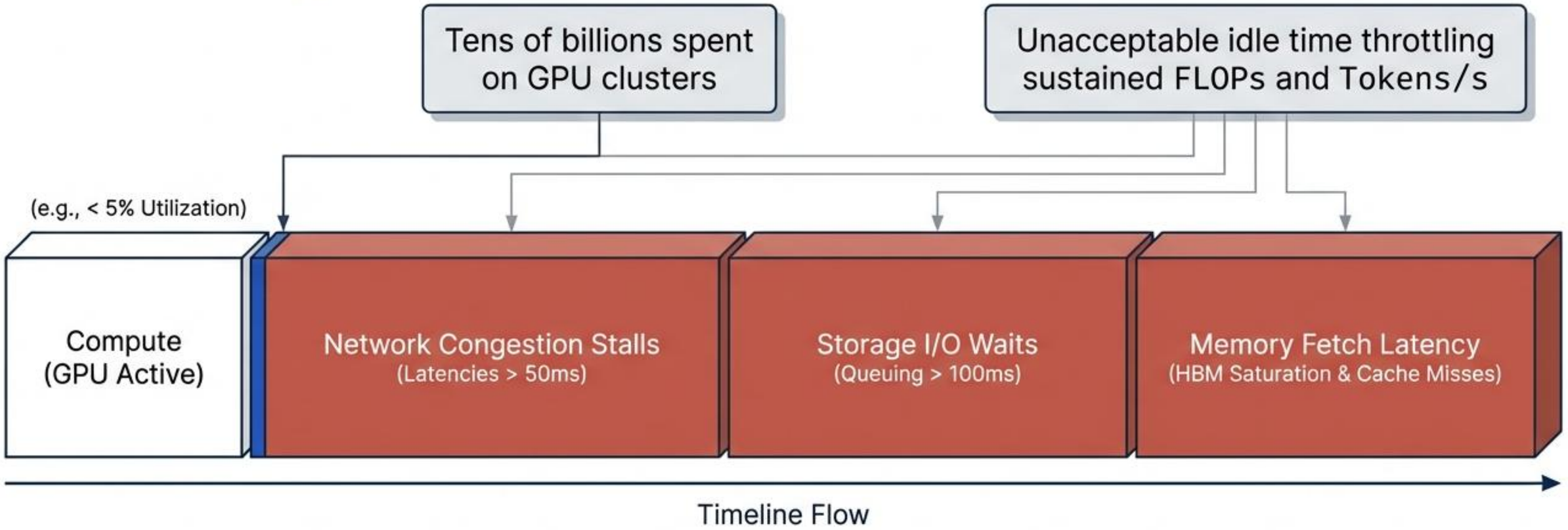
The Synchronized Engine

Unlocking walled-garden hyperscale performance on open-standard silicon via the Predictive Tensor Control Plane (PTCP).



Disclaimer: Forward-Looking Statements This presentation may contain forward-looking statements regarding future product plans, expected features, or anticipated release dates.
These statements outline our general product direction and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for our products remain at our sole discretion and are subject to change without notice. These statements do not represent a commitment, promise, or legal obligation to deliver any specific material, code, or functionality

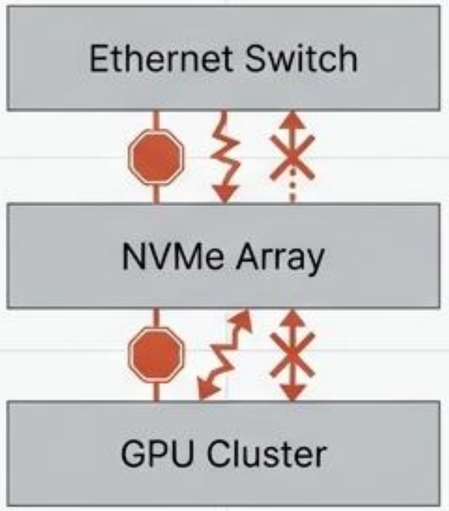
The multi-billion-dollar compute engine is starving



Compute accelerators are scaling **exponentially**. The network and storage fabrics feeding them remain **strictly reactive**. Data gravity is **paralyzing** the AI factory.

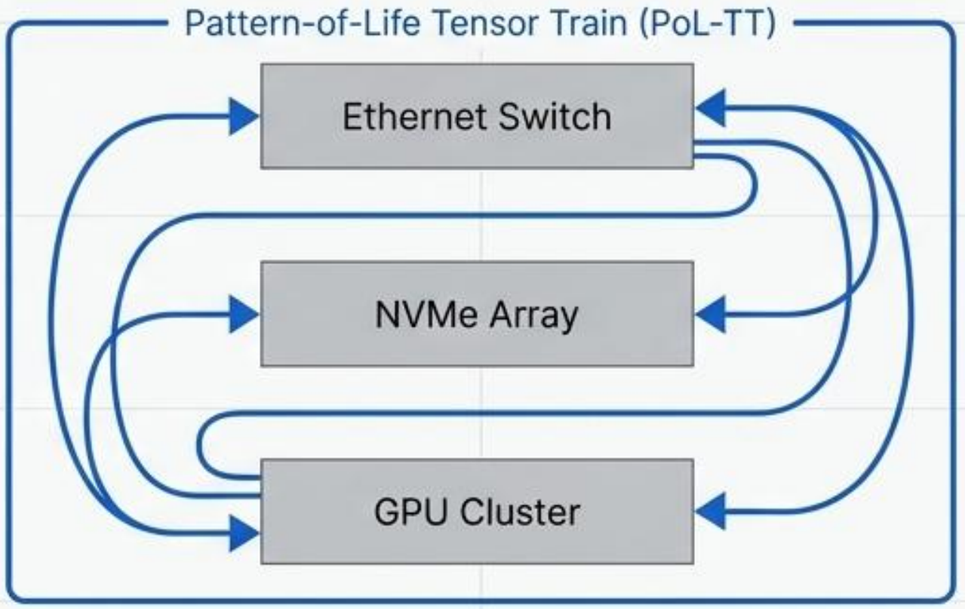
Shifting from reactive hardware to a mathematically predictive pipeline

The Status Quo



Components wait for explicit commands or packet drops.
High latency, massive microbursts.

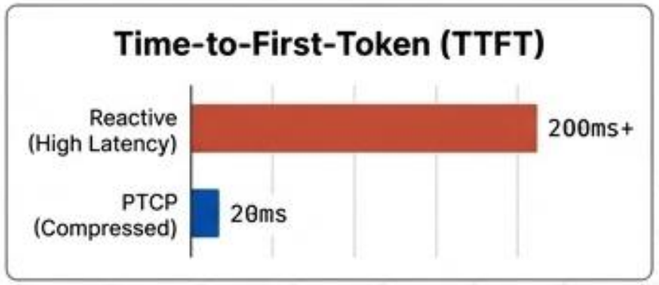
The PTCP Architecture



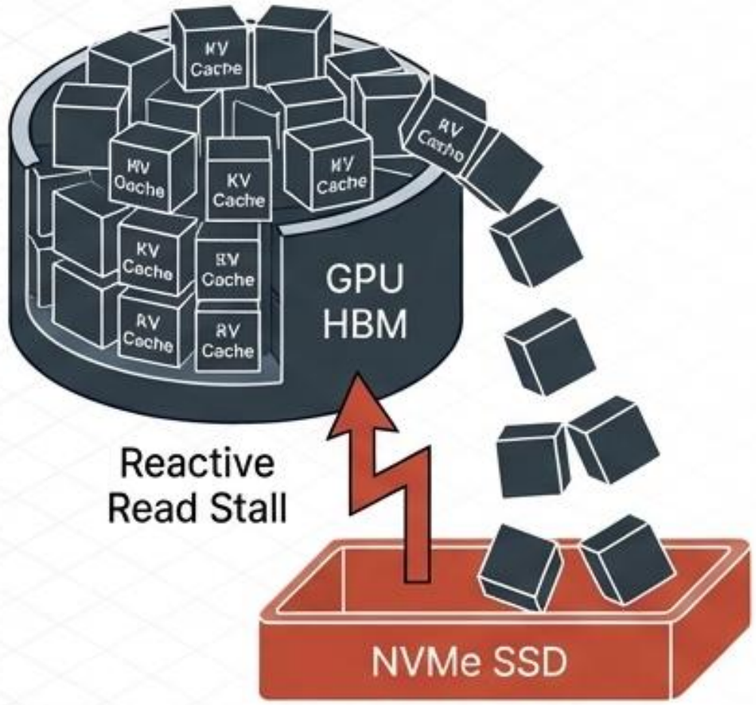
Data moves before it is requested.
Continuous, synchronized pacing.

PTCP operates as a predictive semantic layer, forecasting workload phases to pre-position data and pace traffic across open-standard COTS hardware.

Defeating the Memory Wall during LLM inference

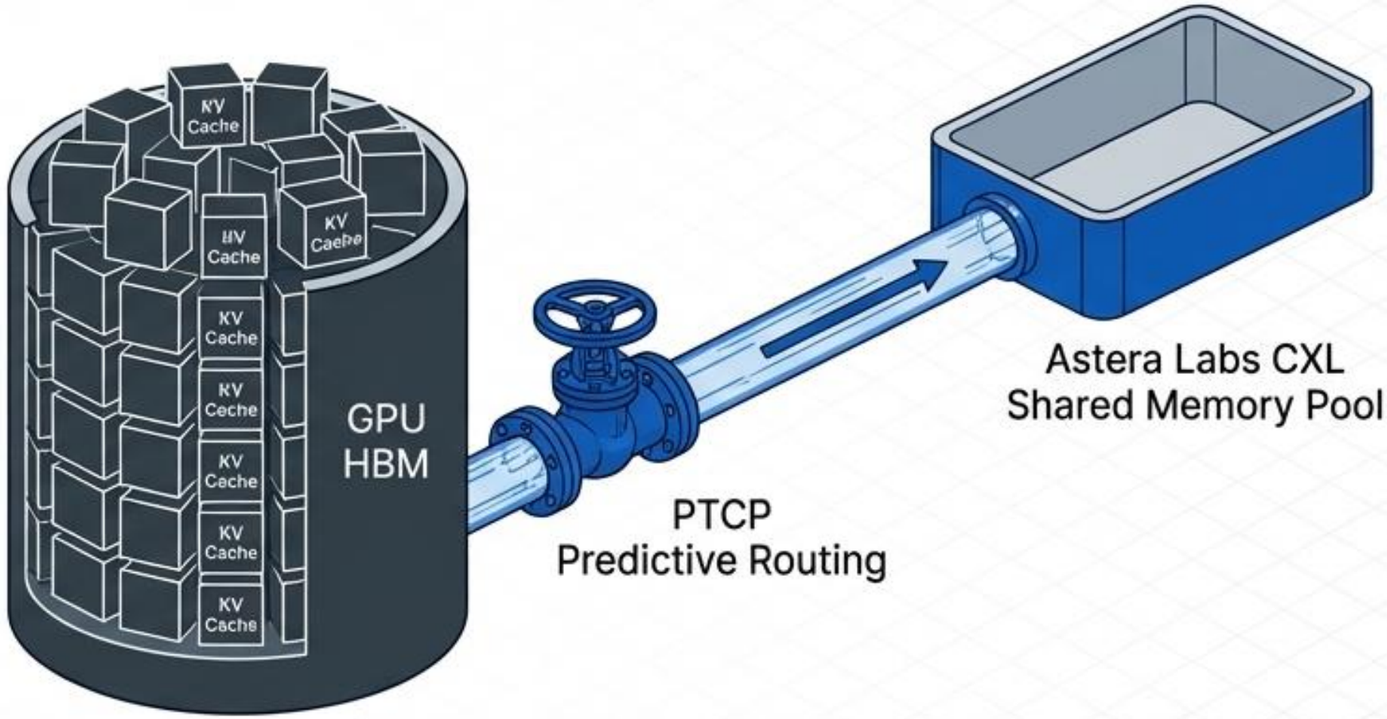


The Status Quo



Components wait for explicit commands or packet drops. High latency, massive microbursts.

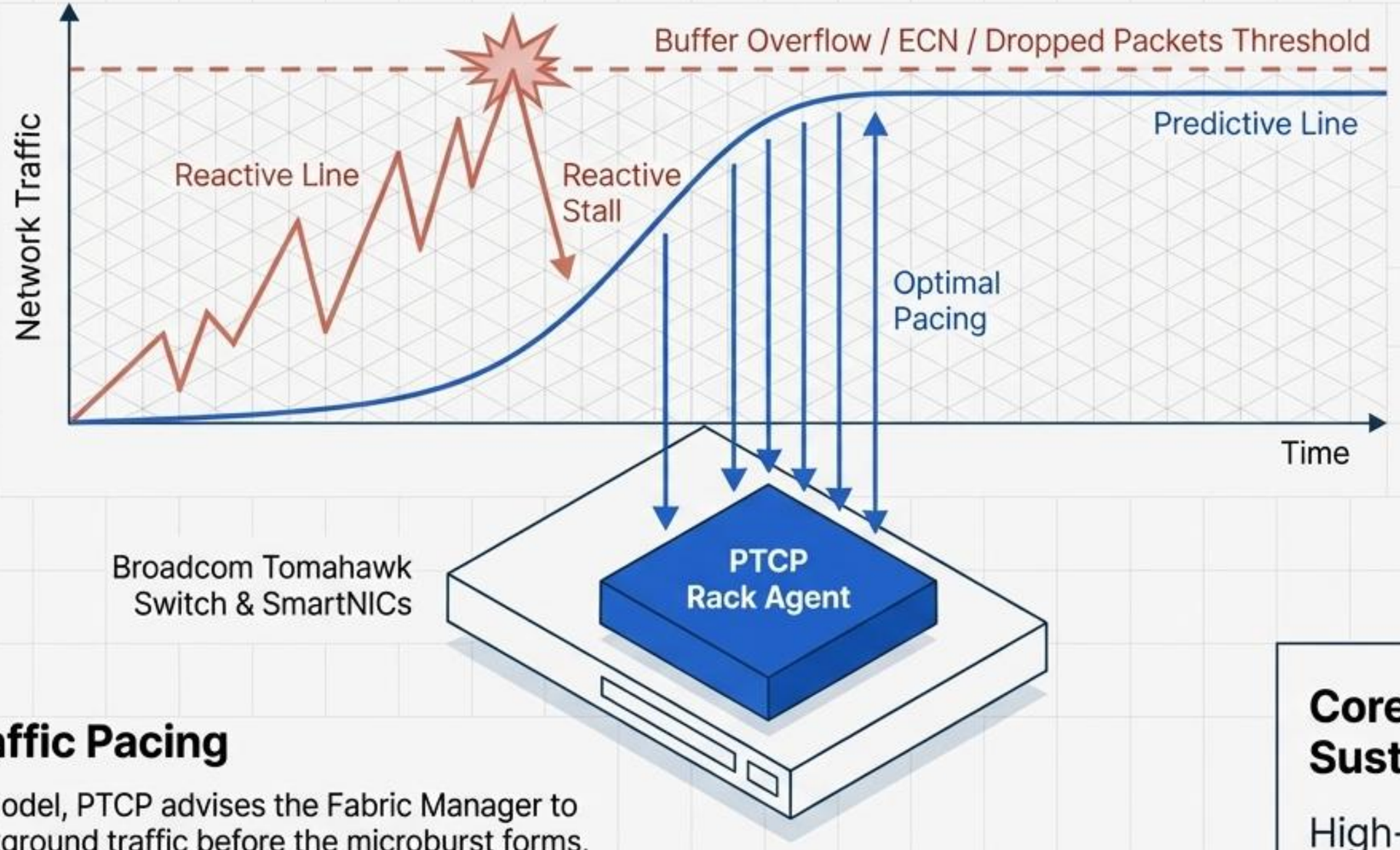
The PTCP Advantage



Data moves before it is requested. Continuous, synchronized pacing.

Predictive Cache Tiering: PTCP forecasts prefix reuse before the prompt is fully processed, moving hot KV caches into high-speed memory tiers to eliminate TTFT-crushing disk reads.

Taming the all-reduce storm in distributed training



Pre-emptive Traffic Pacing

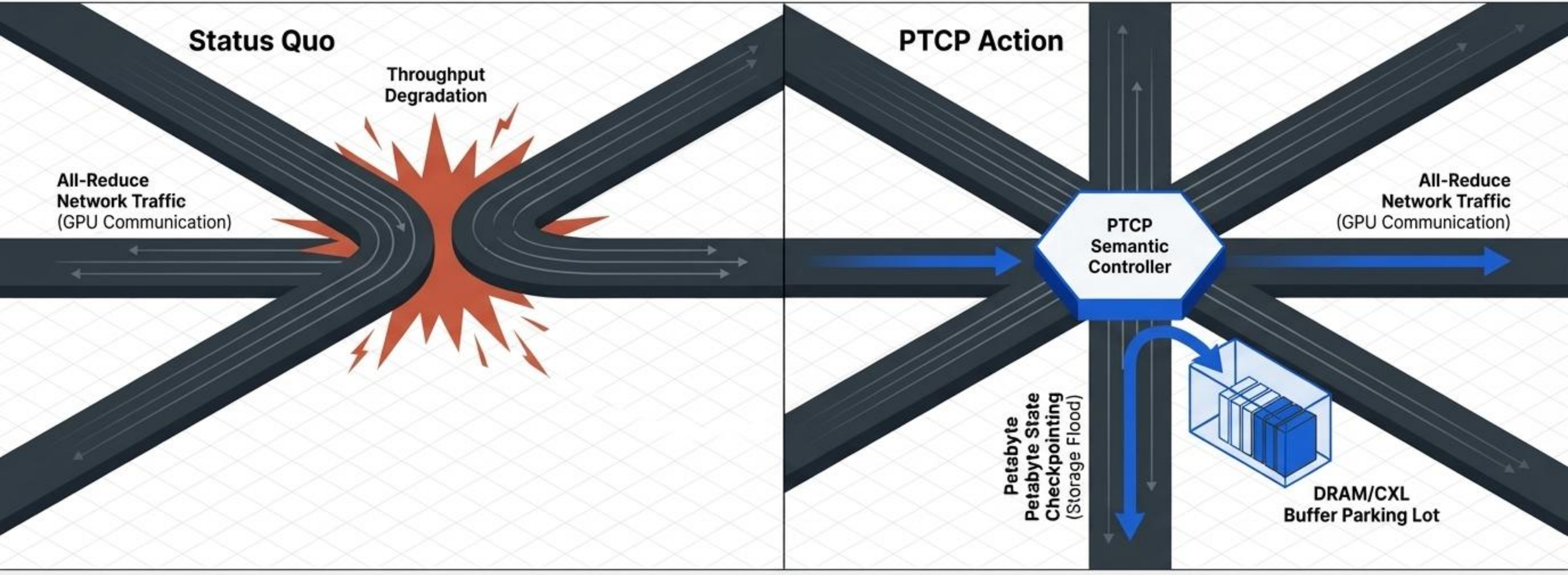
By evaluating the PoL model, PTCP advises the Fabric Manager to shift paths or pace background traffic before the microburst forms, matching proprietary NVLink performance on open Ethernet.

Core Metric: Sustained FLOPs

High-value [...]

Cross-layer synchronization for resilient checkpointing

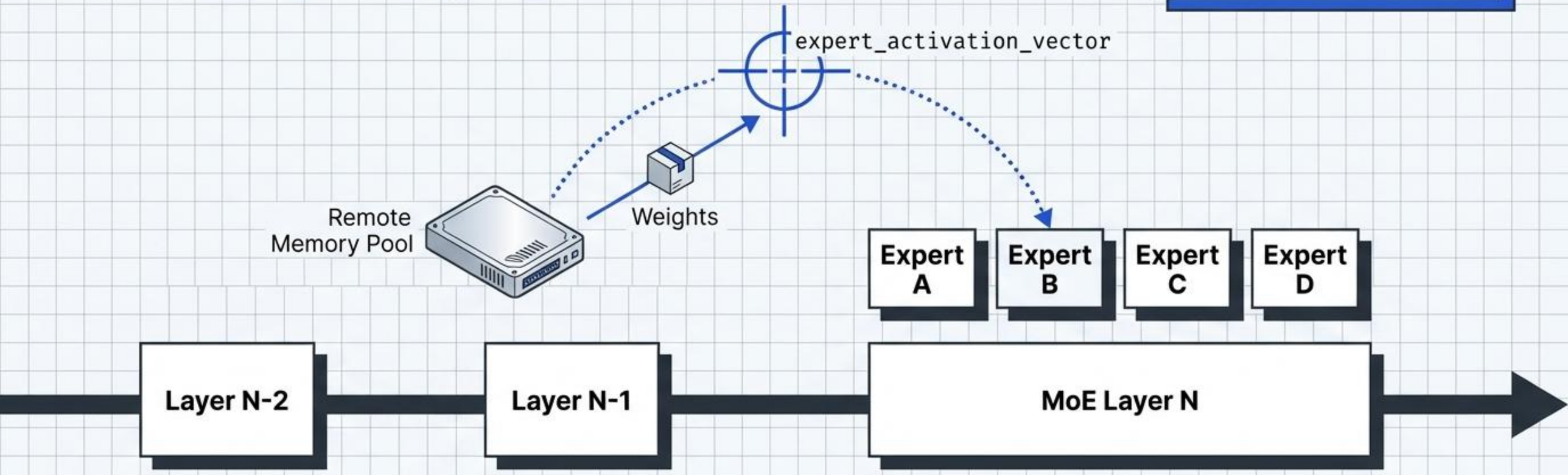
Core Metric: Cluster Uptime & Throughput	
Metric Code	Fira Code
Check-wermase	10% ~ 357 MB/s



Converged Orchestration: The same math governing the DPUs informs the storage controller. Checkpoints are dynamically buffered until the network's collective communication window closes.

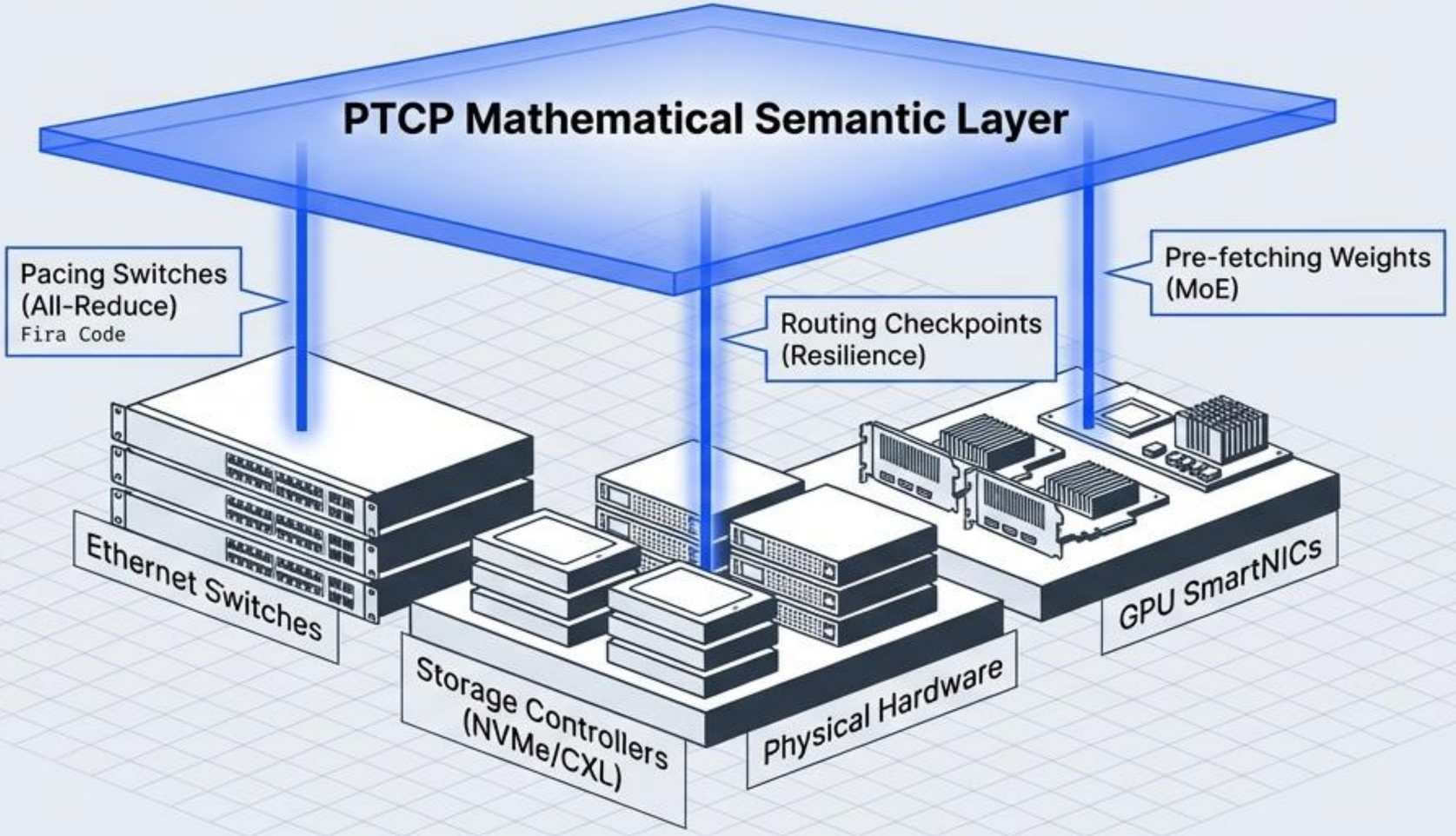
Probabilistic weight fetching for Mixture of Experts

Core Metric: Tokens/s (Inference Saturation)



By tracking the expert activation vector in real-time, PTCP probabilistically pre-fetches the required weights just-in-time, keeping the inference pipeline saturated without latency spikes.

The Converged Orchestration Layer

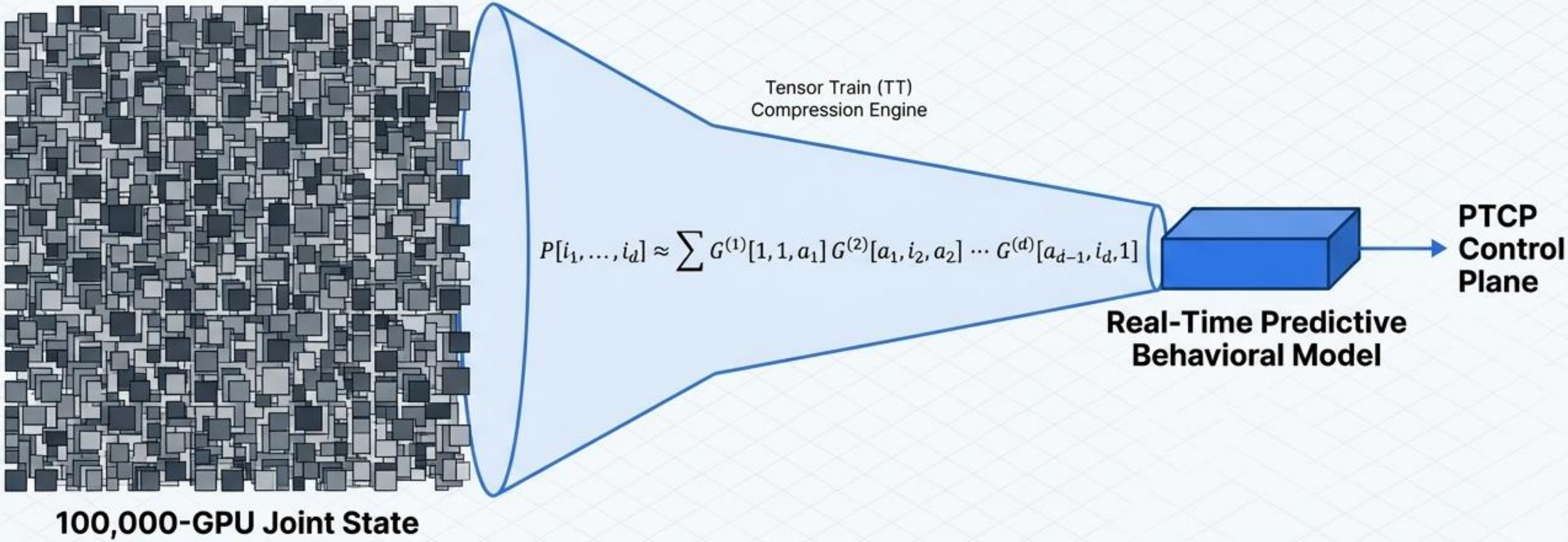


The insight: The exact same mathematical model simultaneously orchestrates network, storage, and memory. The factory is no longer a collection of disjointed parts, but a single, mathematically tractable pipeline.

The Hyperscale Paradigm Matrix

Dimension	Reactive Ecosystem	Predictive AI Factory (PTCP)
Congestion Handling	React & Drop (ECN)	Pace & Shift (Pre-emptive)
Storage I/O	Wait & Stall (Explicit Read)	Pre-fetch & Tier (CXL pools)
Hardware Strategy	Proprietary Walled Garden (NVLink)	Open Standard (Ethernet + CXL/NVMe)
Dominant Constraint	Unpredictable Memory Wall	Saturated Math Engine

Overcoming the curse of dimensionality

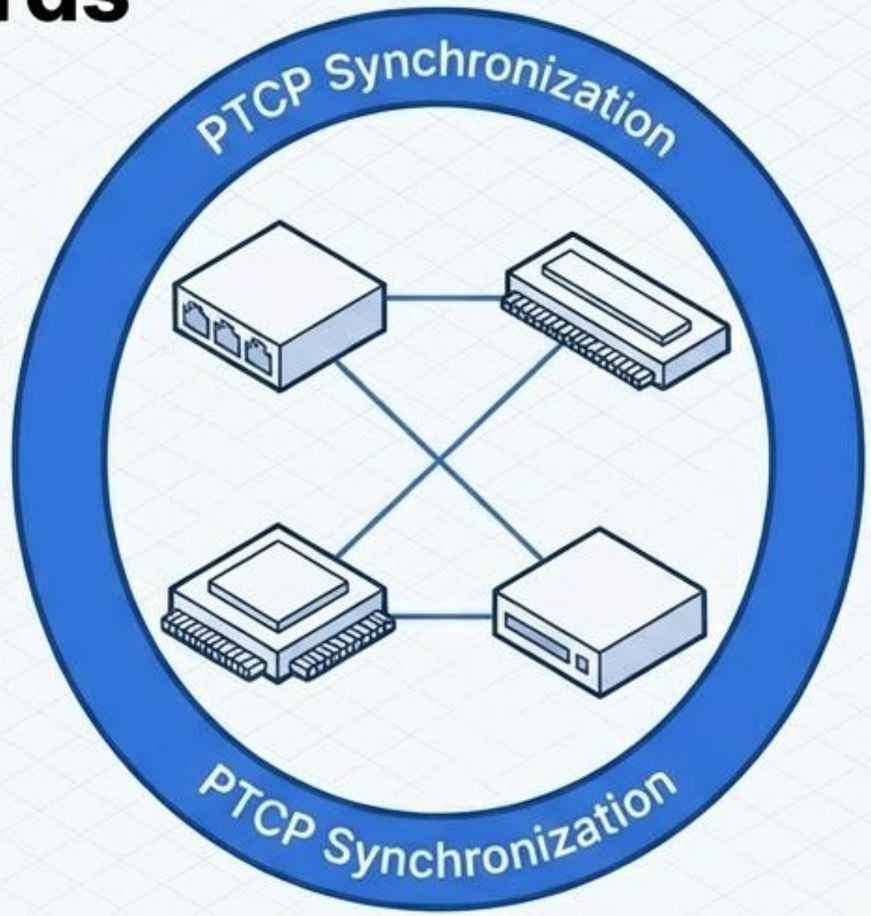


Tracking the state of an entire AI factory creates an unmanageable dense tensor. PTCP utilizes Tensor Train compression to distill this state into a real-time predictive engine.

The strategic moat: Walled-garden performance on open standards

Proprietary Ecosystems
(e.g., NVLink)

- Vendor Lock-in
- Supply Chain Vulnerability



Because COTS hardware is highly optimized for matrix math, PTCP rides directly on top of open protocols. Hyperscalers achieve **tightly coupled** synchronization without dictatorial pricing or supply chain lock-in.

Continuous Token Generation. Saturated FLOPs.

The Predictive Tensor Control Plane transforms the AI factory from a reactive bottleneck into a synchronized engine. Math replaces lock-in. Compute never starves.

Thank You for Your Partnership We look forward to building the future of networking together. Please don't hesitate to reach out with any questions.

Tensor Networks, Inc. Contact Us:

information@tensornetworks.net