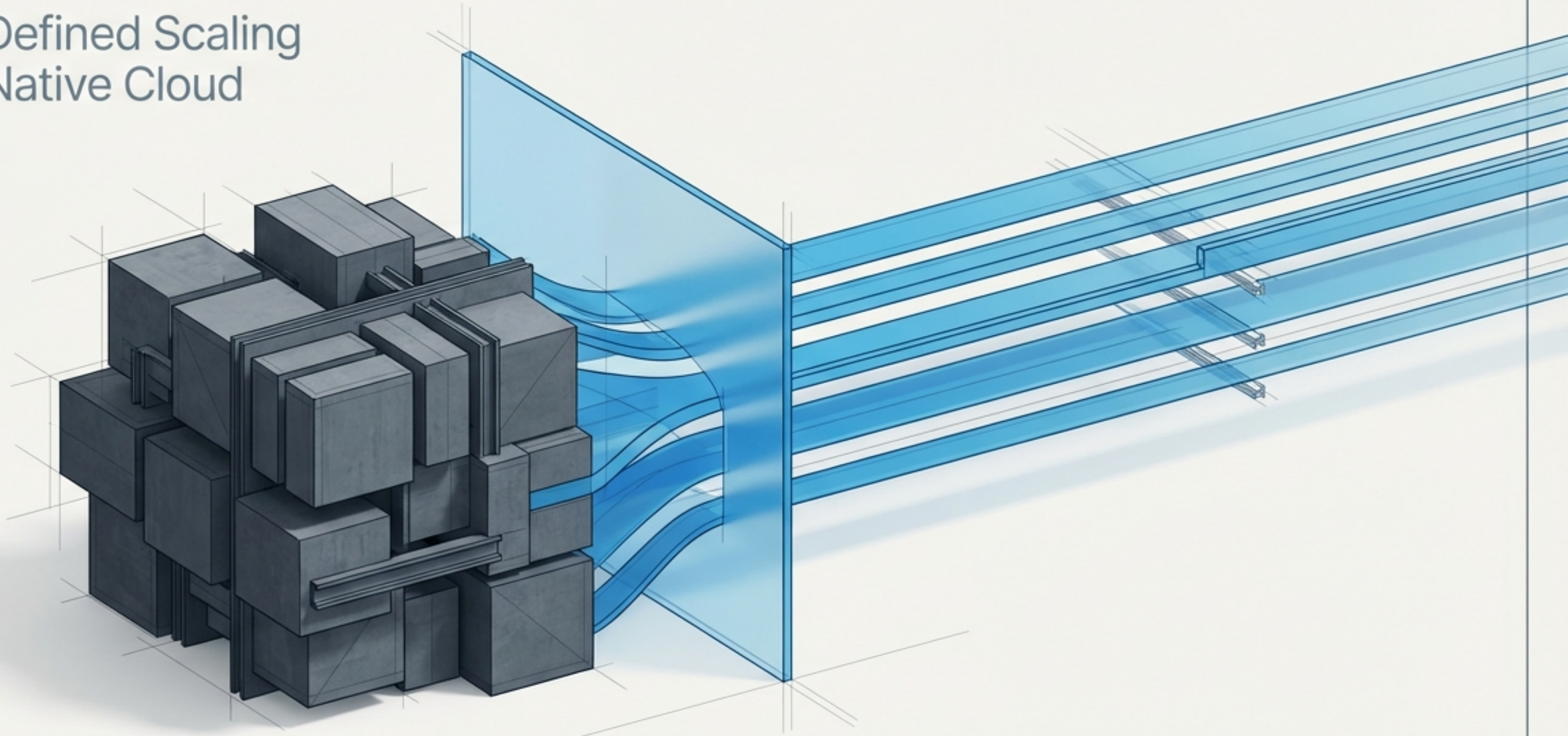
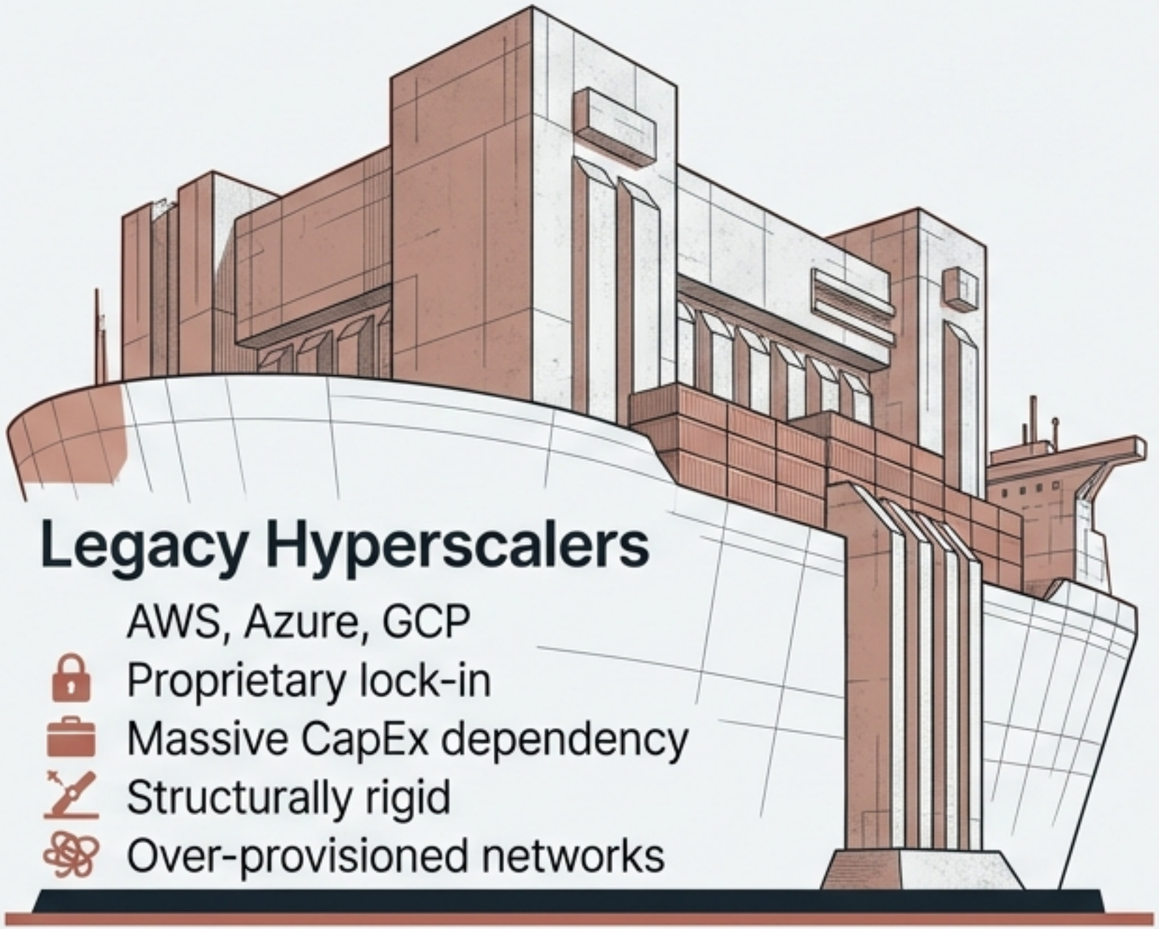
PTCP: The Predictive Tensor Control Plane

Software-Defined Scaling for the AI-Native Cloud

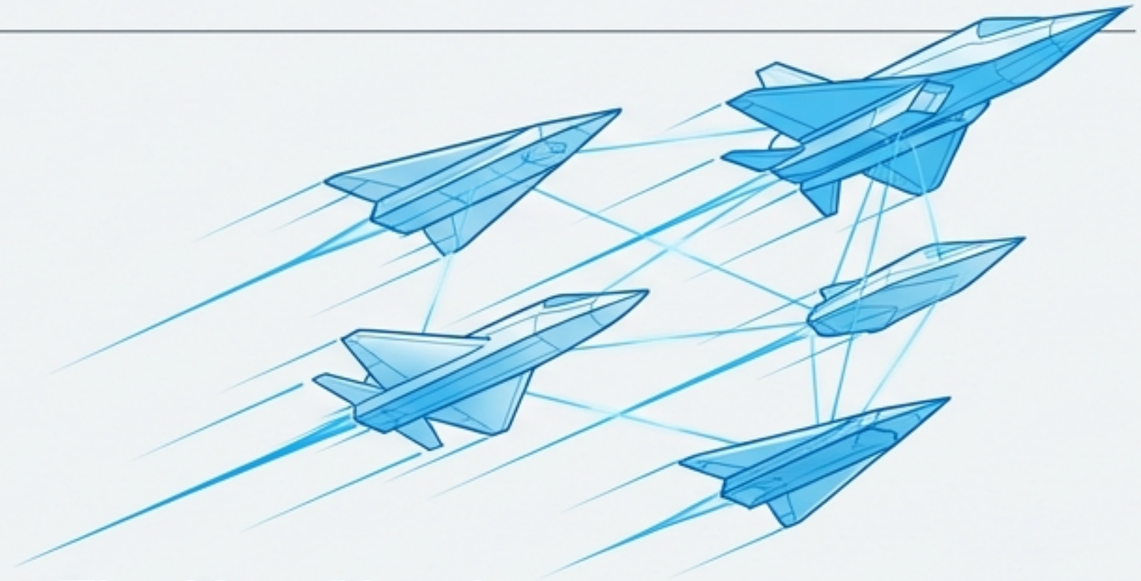


The AI compute market is splitting into two distinct paradigms



Legacy Hyperscalers

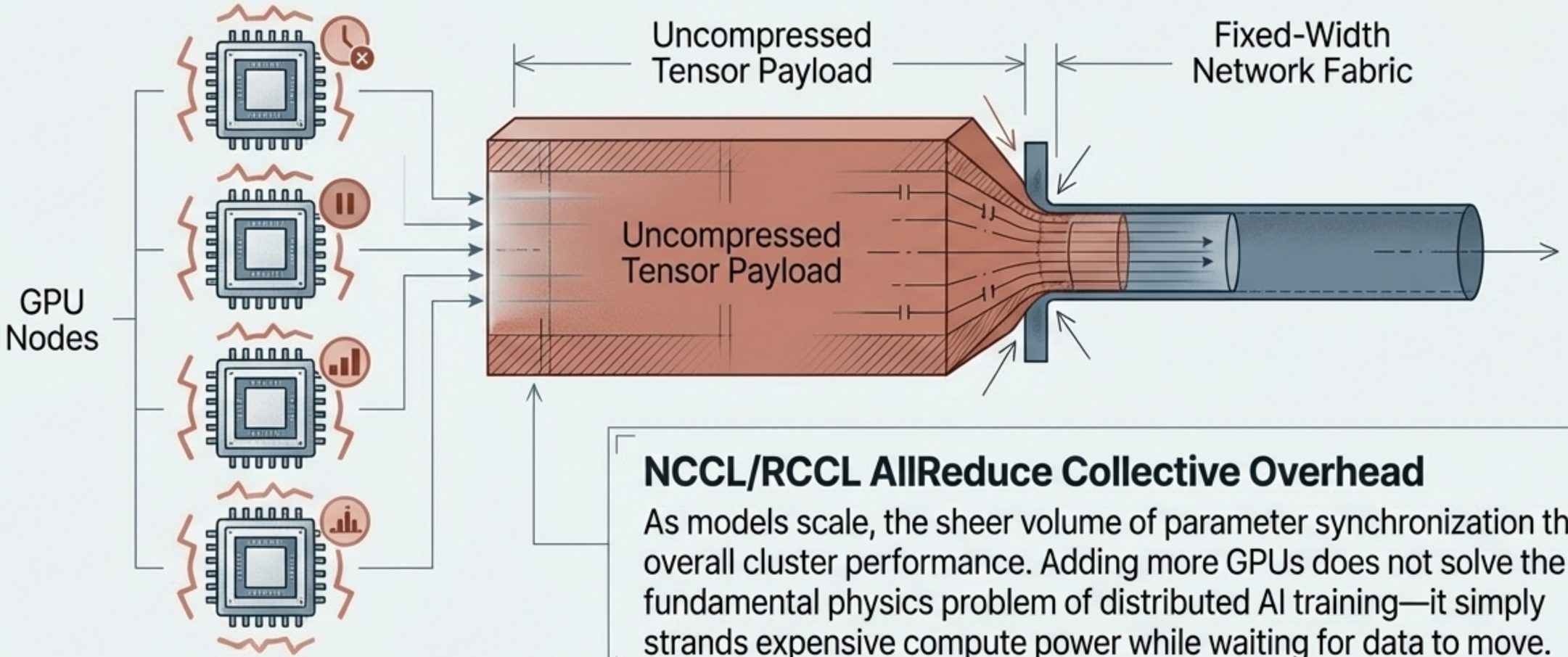
- AWS, Azure, GCP
- Proprietary lock-in
- Massive CapEx dependency
- Structurally rigid
- Over-provisioned networks



The Neo Clouds

- TensorWave, VultR, CoreWeave, Lambda
- Hardware-diverse (AMD Instinct, Custom)
- Cost-efficient & flexible
- Aggressively agile
- Optimized networking

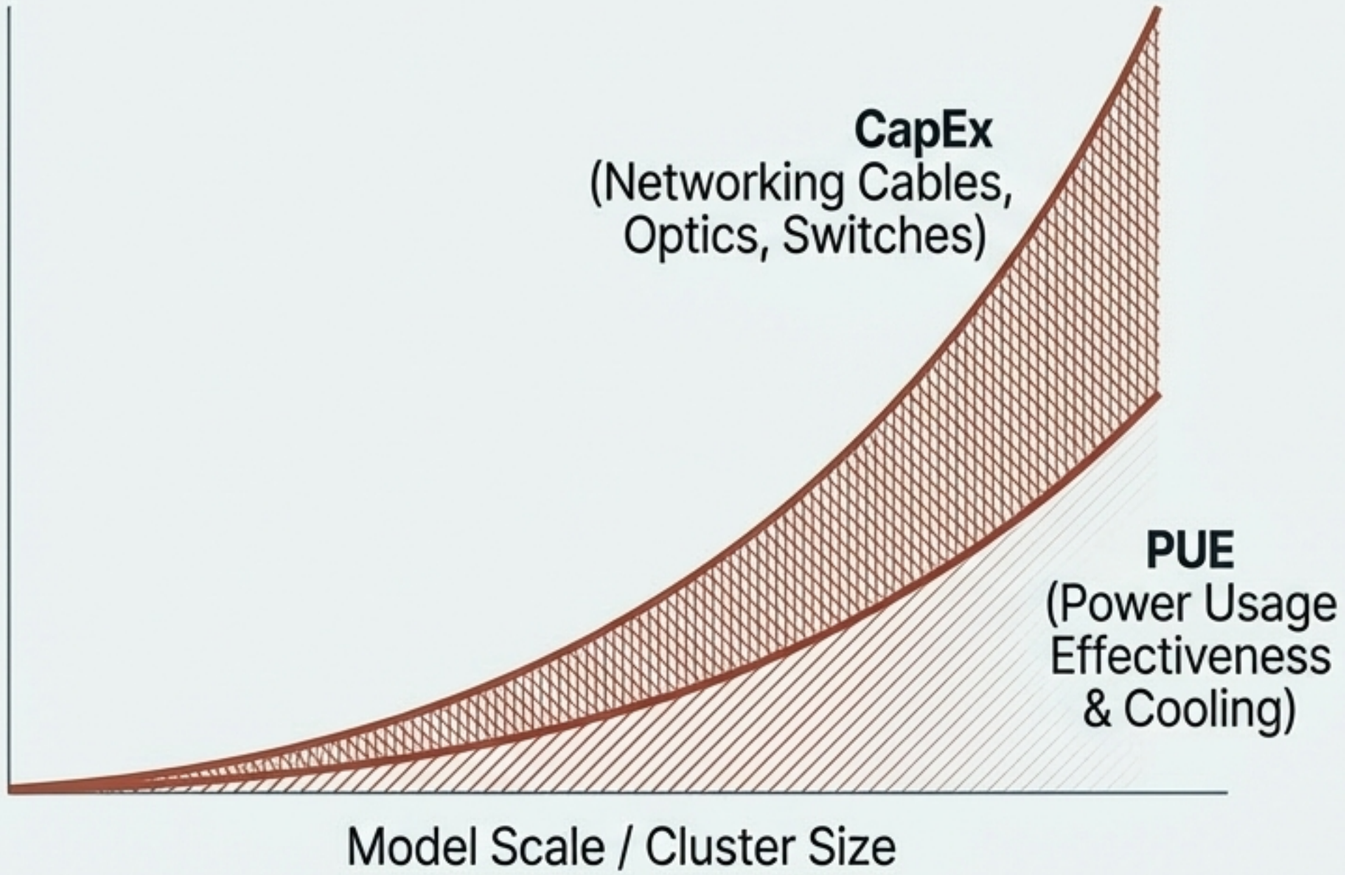
The Communication Wall throttles scale and bleeds efficiency



NCCL/RCCL AllReduce Collective Overhead

As models scale, the sheer volume of parameter synchronization throttles overall cluster performance. Adding more GPUs does not solve the fundamental physics problem of distributed AI training—it simply strands expensive compute power while waiting for data to move.

Winning a brute-force hardware war against hyperscalers is mathematically impossible



Hyperscalers have the balance sheets to over-provision hardware and widen the pipes.

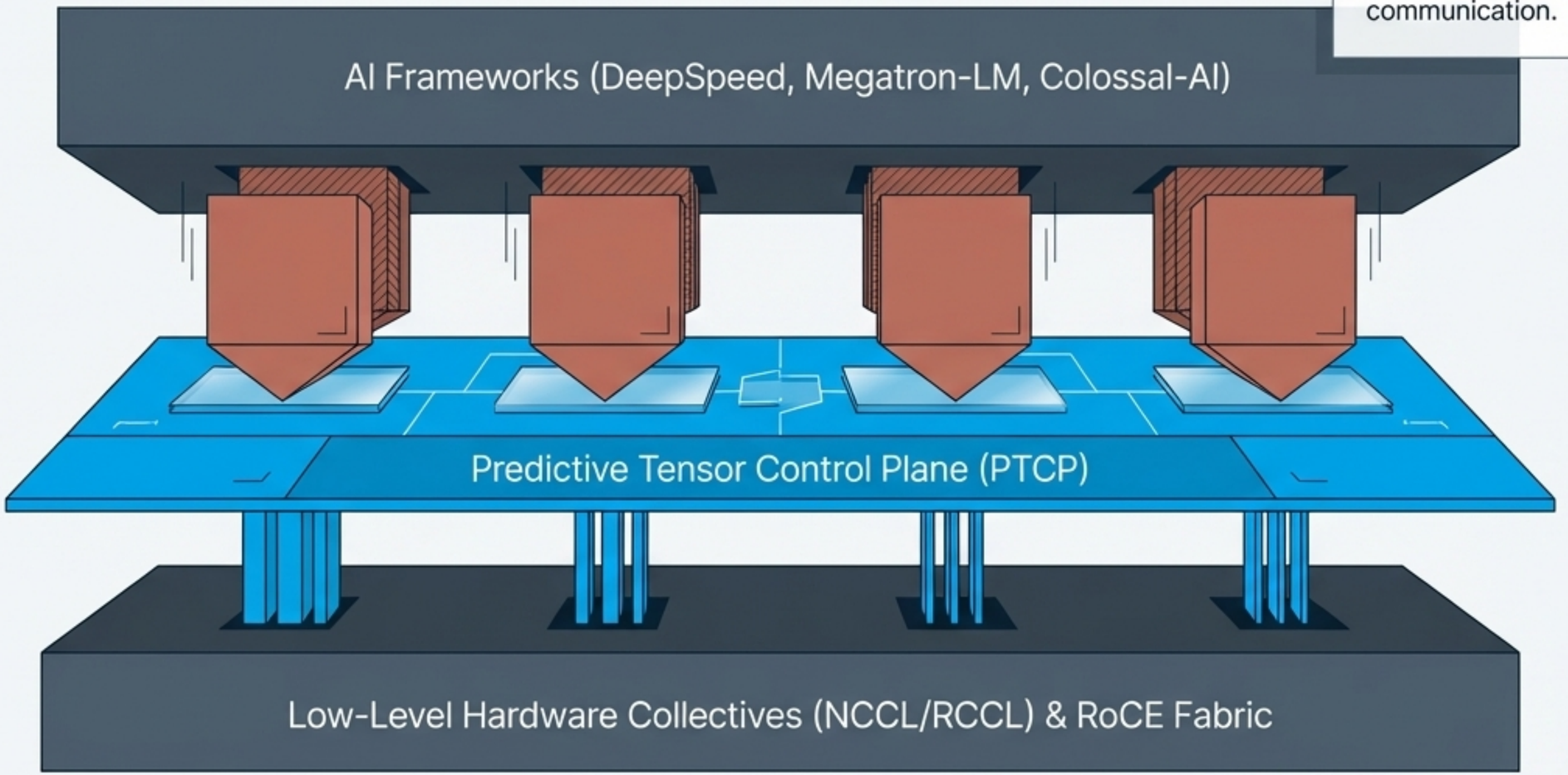
Neo clouds cannot outspend a trillion-dollar monolith on infrastructure alone.

Pure hardware parity is a death sentence for margins.

PTCP shrinks the physical footprint of the workload at the tensor level

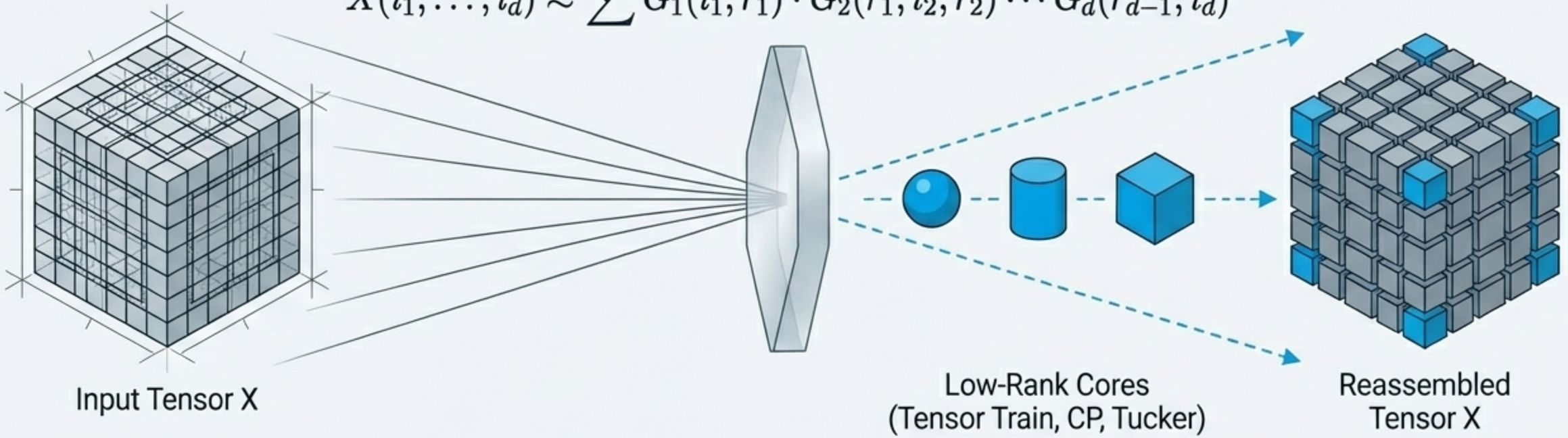
The Breakthrough

A production-grade software overlay that intelligently compresses high-order tensors before communication.



Adaptive tensor-network decompositions rewrite the rules of data gravity

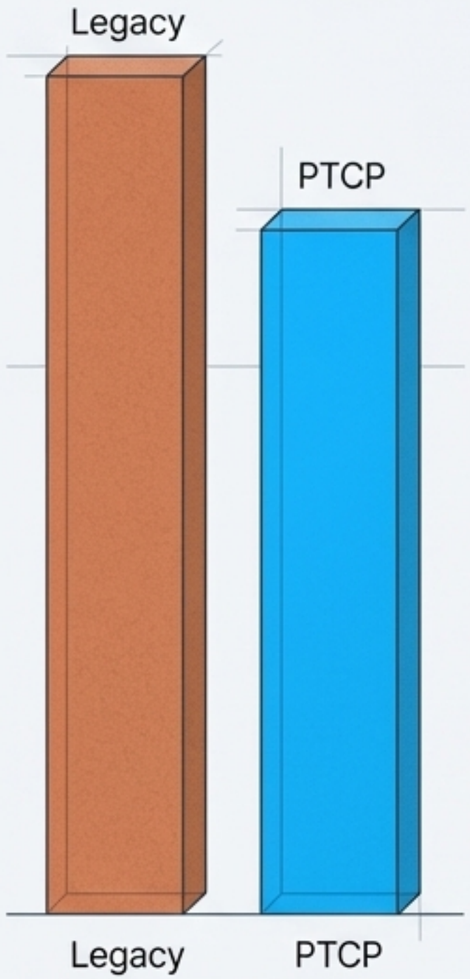
$$X(i_1, \ldots, i_d) \approx \sum G_1(i_1, r_1) \cdot G_2(r_1, i_2, r_2) \cdots G_d(r_{d-1}, i_d)$$



PTCP's proprietary adaptive rank selection algorithm dynamically chooses the optimal decomposition type and rank per tensor based on shape, layer type, and workload statistics.

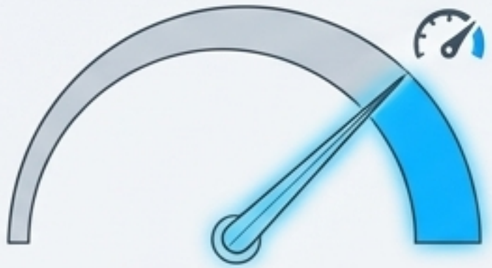
Production-grade compression without compromising model integrity

Communication Volume



Drops by 15–40%+

Effective Throughput



Increases by 15–40%+

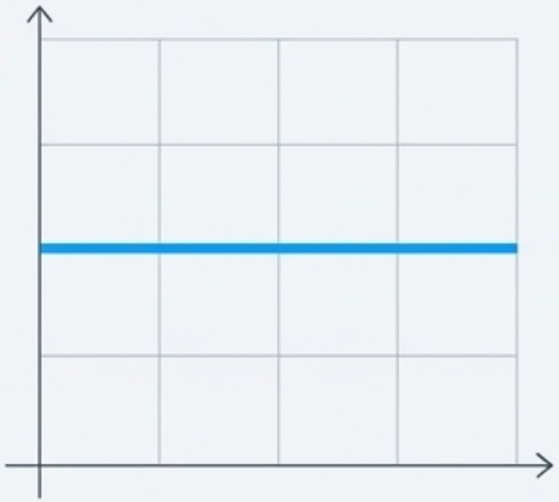
(tokens/sec or training steps)

System Overhead

Only 3–8%

of total step time

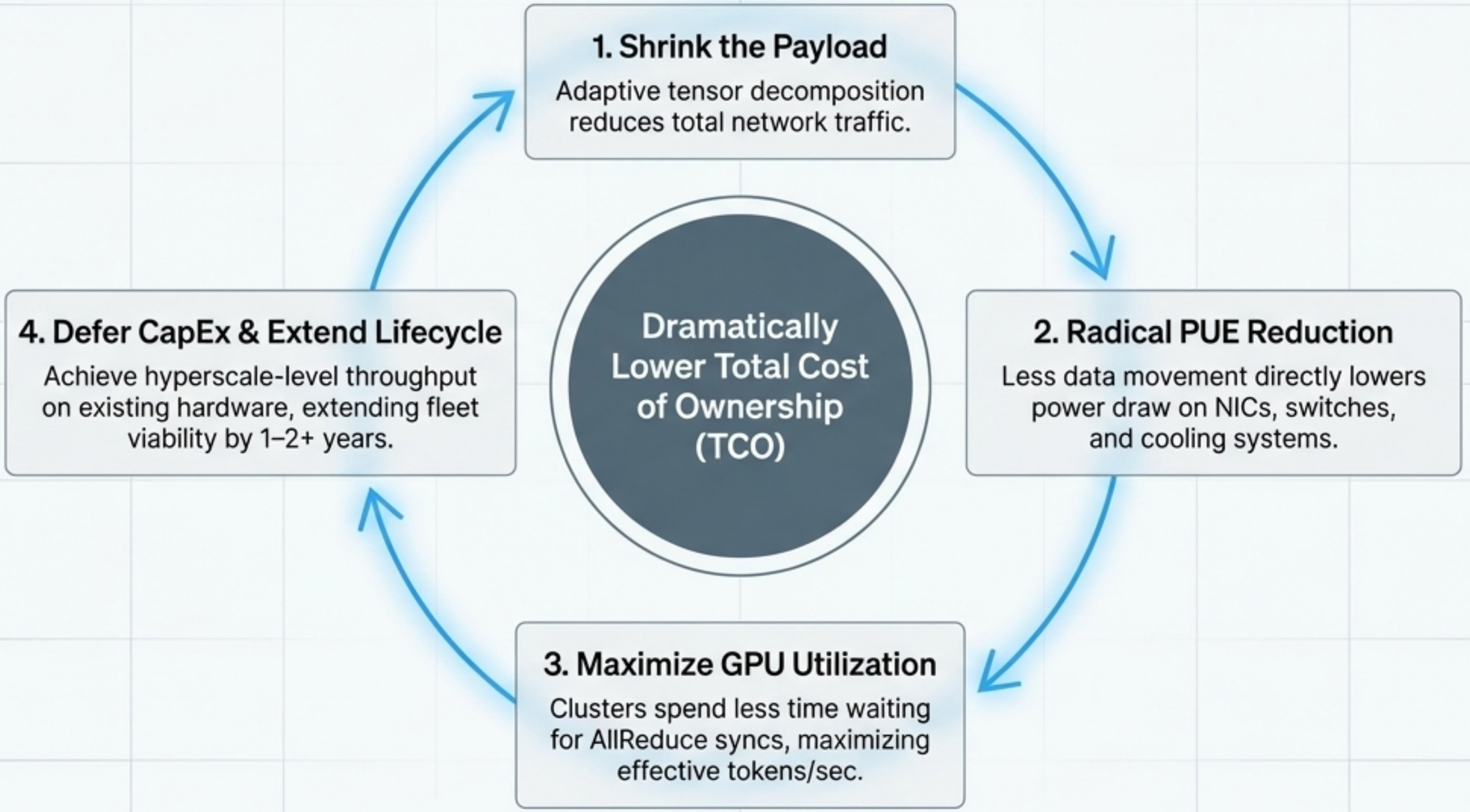
Precision Tracking



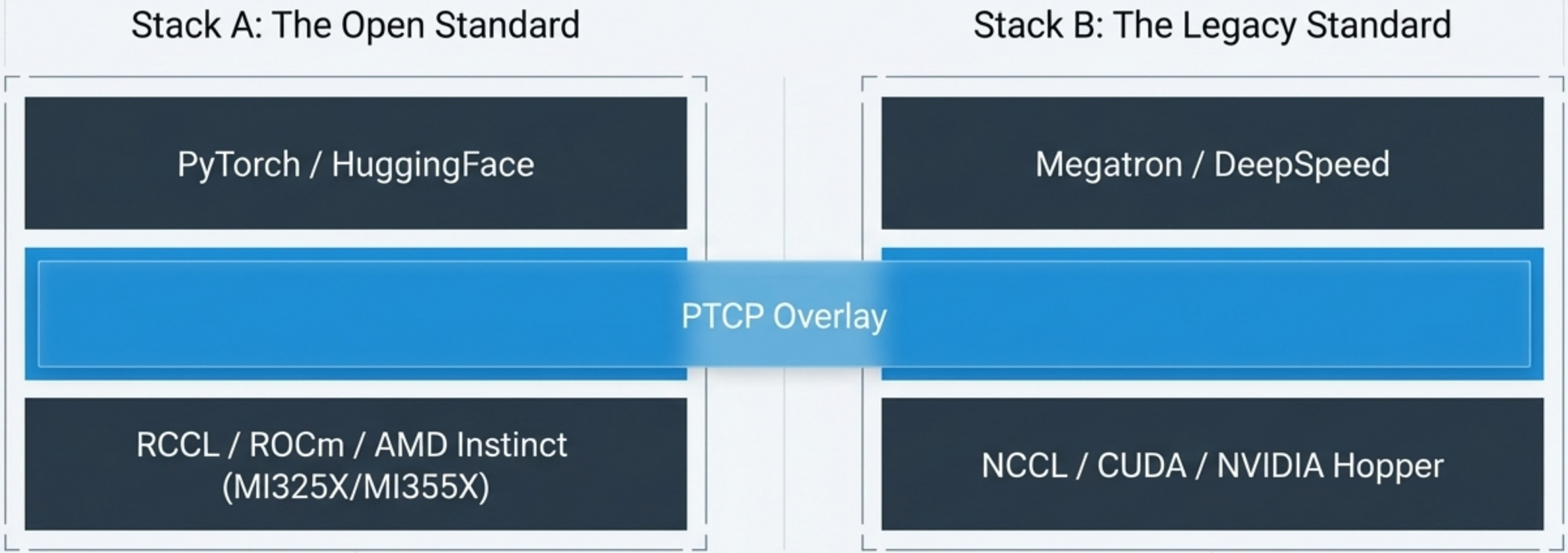
< 1% Degradation

Tunable accuracy impact maintained strictly within tolerance ε.

The Neo Cloud Efficiency Flywheel transforms network reduction into compounding ROI



Seamless deployment across diverse hardware and framework architectures

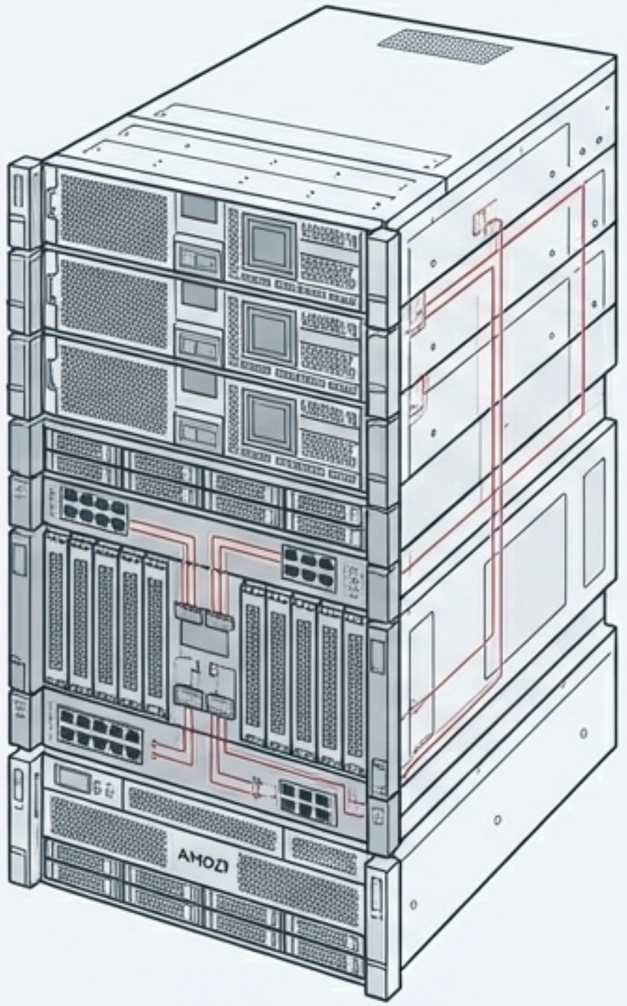


PTCP is entirely complementary to quantization, sparsity, or new physical interconnects. It specifically targets payload volume, delivering gains that scale precisely as your infrastructure scales, without forcing a massive re-architecture.

Strategic divergence: Building the Software-Defined Moat

	Legacy Hyperscalers	PTCP-Optimized Neo Clouds
Scaling Strategy	Brute-force hardware over-provisioning	**Algorithmic volume reduction**
Hardware Dependency	Absolute lock-in to proprietary networks / NVIDIA	**Hardware-agnostic, maximizing ROCm/AMD efficiency**
Cost Structure	Crippling upfront CapEx	**Deferred CapEx / 1–2 year lifecycle extension**
Network Architecture	Fat-tree congestion management	**Hyper-efficient, payload-compressed smart fabrics**

Case Study: Amplifying the AMD Instinct stack for dominant price-performance



Target Profile: TensorWave (High-density, AMD-focused compute)

The Challenge: AMD Instinct + Pollara-based clusters are inherently **communication-heavy.** Standard RCCL AllReduce protocols bottleneck raw compute capability.

The PTCP Intervention: Transparently integrates into the open ROCm stack to compress payloads before the fabric layer, acting as a force multiplier for the hardware.

The Business Outcome: Unlocks 20–35% better price-performance on MI325X and MI355X platforms. **Enables higher customer density per rack, reinforcing positioning as the ultimate cost-effective alternative to monolithic NVIDIA clouds.**

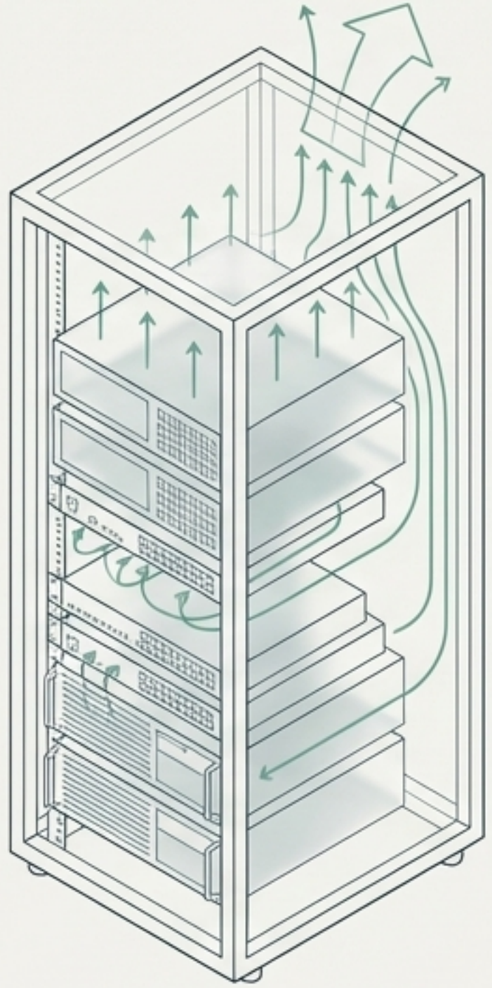
Case Study: Driving the sustainable cloud through radical PUE reduction

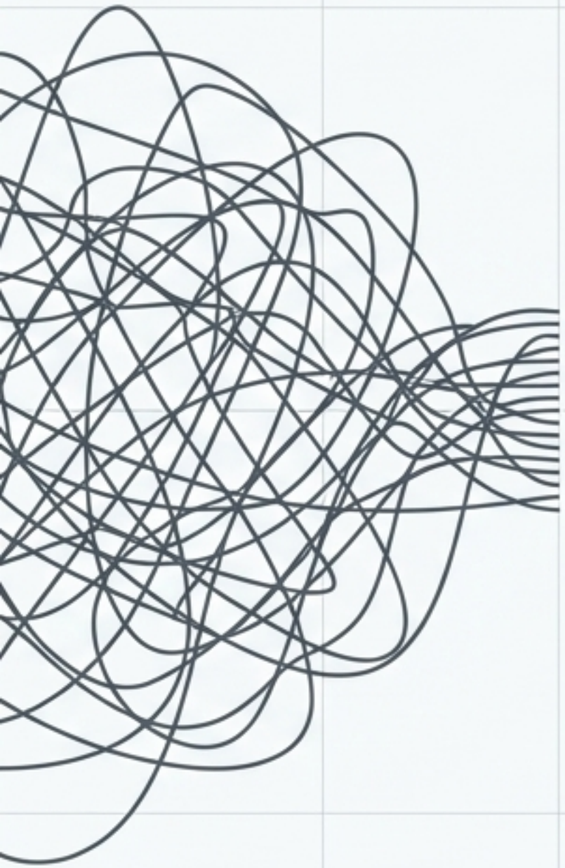
Target Profile: VultR (Sustainable, flexible, high-density infrastructure)

The Challenge: AI networking hardware (NICs, optics, switches) consumes massive amounts of power, driving up Power Usage Effectiveness (PUE) and limiting rack density due to strict thermal ceilings.

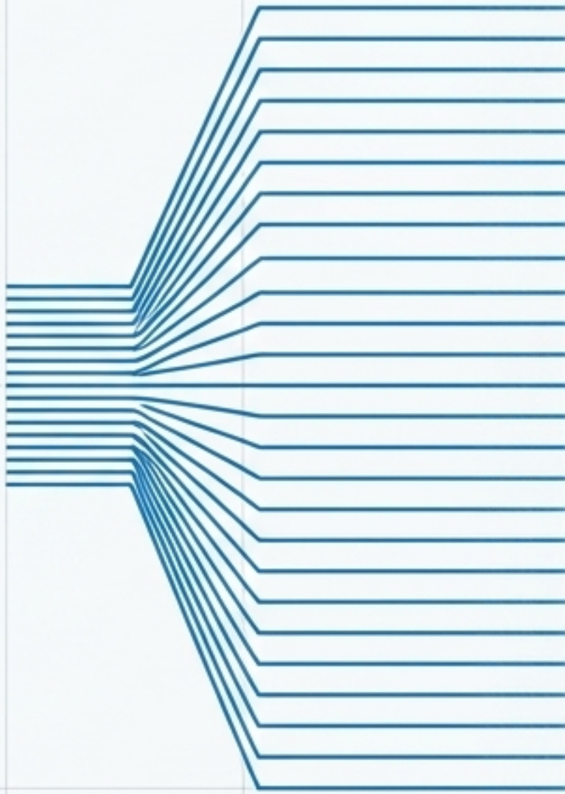
The PTCP Intervention: A 40% reduction in network volume translates to a direct, proportional drop in switch active-state power draw and data-movement heat generation.

The Business Outcome: Enables VultR to bring differentiated, Enables VultR to bring differentiated, communication-optimized, and highly sustainable clusters to market. Captures price-sensitive AI startups and research teams focused on ESG mandates.





Hardware parity is a myth.
For Neo Clouds,
outspending hyperscalers
is impossible.
The ultimate competitive
moat is software-defined
efficiency at the tensor level.



A rapid, 6-to-9 month path to a proprietary, co-branded offering

1. The Proof

Rapid Proof-of-Concept (PoC) on a representative client workload (e.g., Llama-3 70B training). Tuned for strict error bounds (ε).

2. The Integration

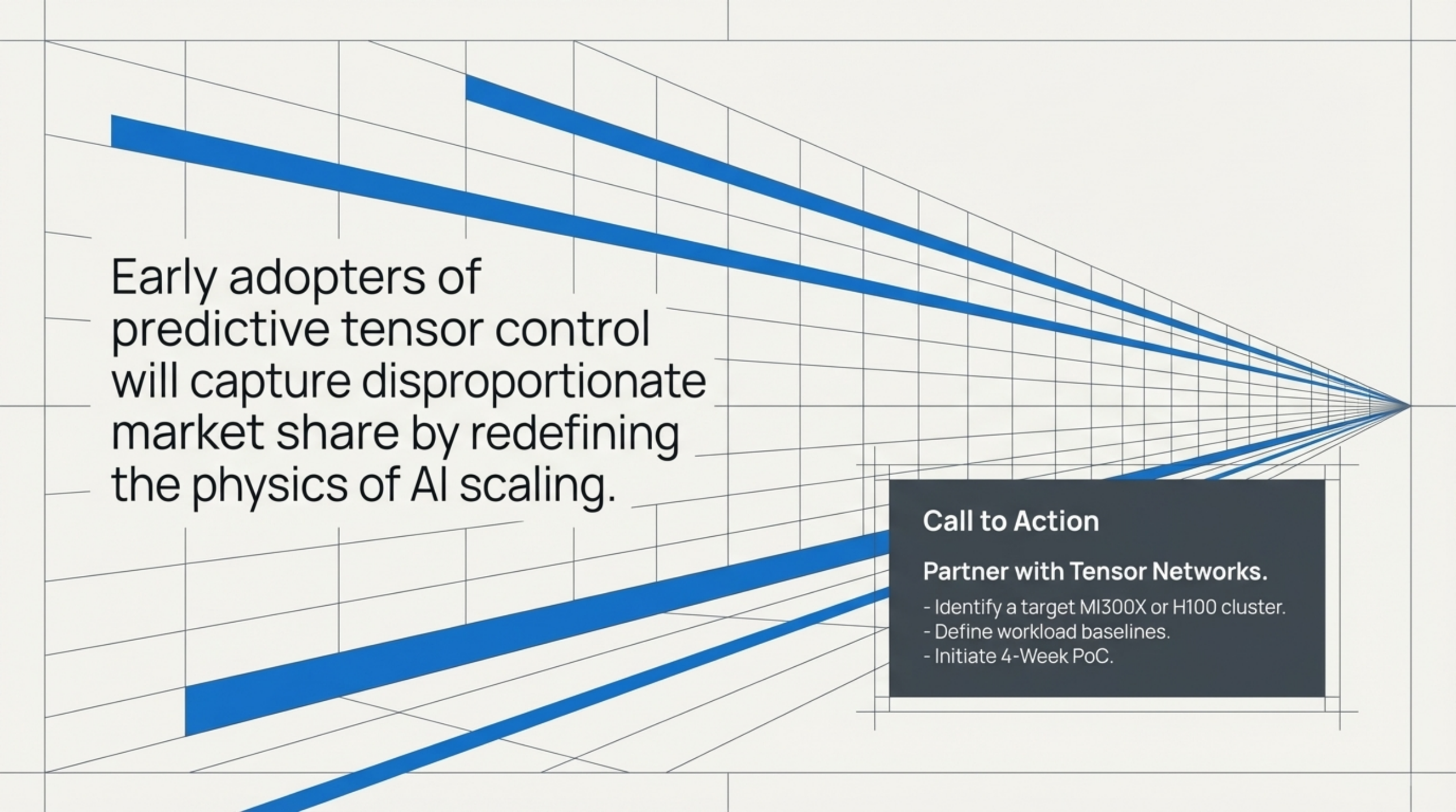
Joint engineering optimization to perfectly map PTCP to your existing RoCE and NCCL/RCCL networking stack.

3. The Rollout

Staged deployment, beginning with your highest-value, most communication-constrained customer clusters.

4. Market Execution

Co-marketing launch of the new tier: “PTCP-Optimized AI Infrastructure,” directly targeting hyperscaler defectors.


Early adopters of predictive tensor control will capture disproportionate market share by redefining the physics of AI scaling.
Call to Action
Partner with Tensor Networks.
- Identify a target MI300X or H100 cluster.
- Define workload baselines.
- Initiate 4-Week PoC.