Tensor Networks, Inc.

PTCP AI Factory
Fast eBPF / Slow Path Control for Next-Generation AI Data Centers

Clean-layout edition: SONiC top-of-rack switch and BlueField DPU architecture for NVIDIA, AMD, Intel, and Broadcom AI infrastructure.

GPU Flow | DPU Fast Path | SONiC Slow Path | RAG Storage | Token Metrics

Postdoctoral-style non-fiction whitepaper and technical capabilities analysis
Prepared for executive, technical, investor, licensee, and infrastructure-operator review
Source-aligned with Tensor Networks Combined Platform collateral
July 2026

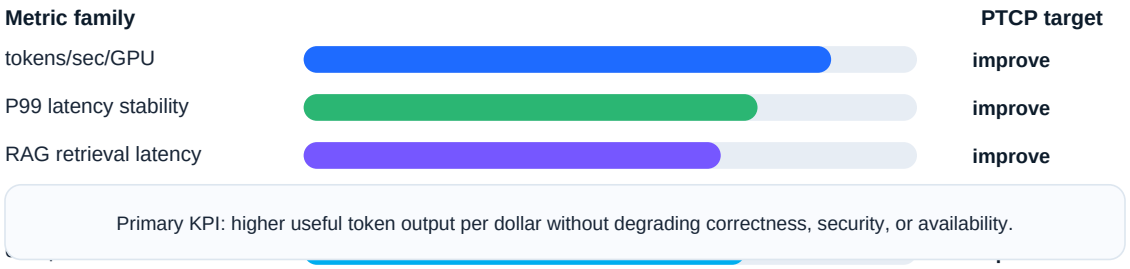
Executive Summary

PTCP AI Factory is framed here as a cleanly deployable data-center control plane that improves the productivity of AI infrastructure by coupling telemetry, topology modeling, fast-path enforcement, and slow-path policy orchestration. The objective is practical: increase useful token output per dollar on existing accelerator, DPU, NIC, ToR, and storage infrastructure.

Strategic thesis

The next bottleneck in large AI infrastructure is not only accelerator supply; it is the coordination of compute, network, storage, scheduling, security, and retrieval paths at cluster scale.

Business value scorecard: token productivity, tail latency, RAG retrieval, and GPU idle time

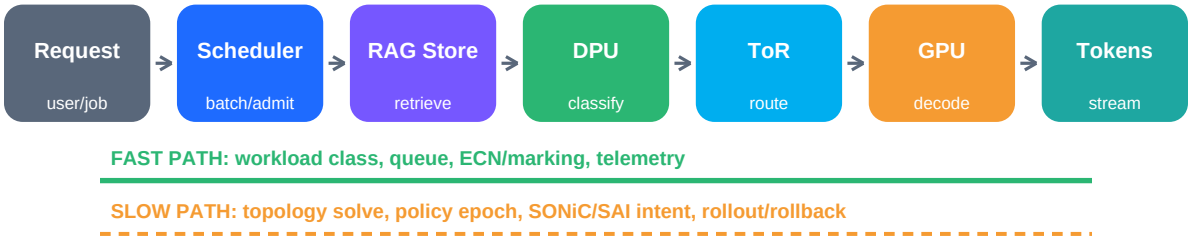


- **Compute:** keep high-cost GPUs fed with data, retrieval context, and collective traffic without collapsing into east-west congestion.
- **Network:** distinguish latency-critical token, RAG, control, and checkpoint paths; then route or queue them deterministically.
- **Storage:** treat vector search, object retrieval, and KV-cache movement as first-class token productivity paths.
- **Operations:** produce evidence that optimization changes improved throughput without degrading security, reliability, or policy compliance.

1. Why AI Data Centers Need a Control Plane

At 10,000-GPU scale, the bottleneck is often an interaction of scheduler choices, storage locality, network incast, retransmits, queue starvation, and uneven batch flow. PTCP AI Factory addresses that interaction as a topology problem instead of a collection of disconnected point problems.

End-to-end token flow: client request to streamed tokens

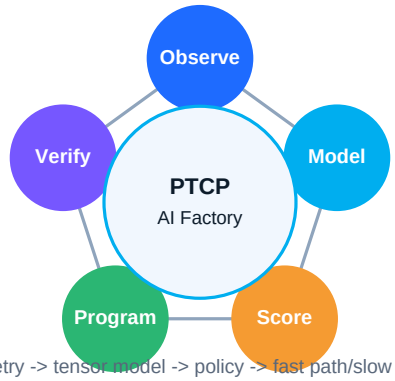


- Tokens/sec is a system metric. It is shaped by GPU kernels, memory bandwidth, storage retrieval, network jitter, request batching, and service placement.
- RAG moves the bottleneck beyond raw GPU math: vector lookup, document retrieval, embedding stores, cache misses, and object-store bursts can stall inference.
- PTCP AI Factory creates a continuous control loop around those dependencies, then programs safe controls at either the BlueField DPU/IPU side or the SONiC top-of-rack control layer.

2. PTCP AI Factory as a Deterministic Overlay

The supplied Tensor Networks source describes AI Factory as a topographic, deterministic software overlay for massive AI workloads. This clean-layout edition expresses that claim as a concrete software architecture: observe, model, score, program, verify.

Closed-loop PTCP AI Factory operating model



Closed loop: topology telemetry -> tensor model -> policy -> fast path/slow path -> verified token metrics

Source-aligned framing

The Tensor Networks source states that AI Factory maps the geometric topology of the cluster and routes data with precision so GPU assets are not waiting idly for payloads. In this document, that thesis is expanded into a deployment model for eBPF, DPU, and SONiC environments.

- **Observe:** collect workload, flow, queue, storage, and accelerator telemetry.
- **Model:** compress the cluster into a high-dimensional tensor/topology representation.
- **Program:** deploy bounded fast-path and slow-path controls with rollback.
- **Verify:** track tokens/sec, tail latency, GPU idle time, and policy evidence.

3. Deployment Placement: DPU and ToR

PTCP AI Factory can be expressed through two complementary insertion patterns. In a DPU-centered pattern, flow classification and telemetry occur near the server. In a ToR-centered pattern, policy orchestration and ASIC-safe routing intent occur through SONiC and SAI abstractions.

PTCP placement across GPU, DPU, ToR, storage, and operations

Two legal insertion points

1) BlueField / IPU / DPU fast path: classify, mark, meter, and steer high-value AI flows near the host.

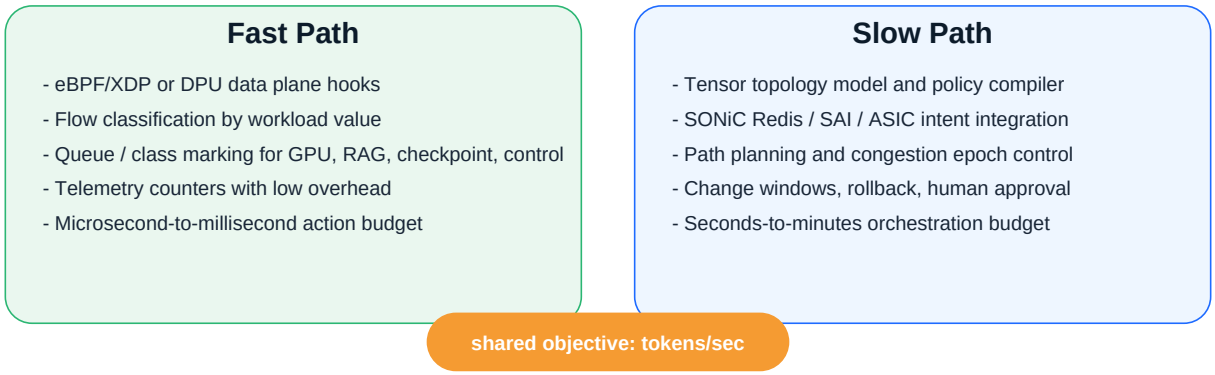
2) SONiC ToR slow path: compile policy intent into queue, route, ECN, telemetry, and ASIC control abstractions.

- **DPU/BlueField/IPU placement:** suited to per-host flow marking, storage/security offload, service isolation, and data-path telemetry.
- **SONiC ToR placement:** suited to switch control-plane intent, queue profiles, route preference, ECN/PFC monitoring, telemetry export, and ASIC-programming boundaries.
- **Hybrid placement:** the most powerful model: host/DPU fast path for high-frequency actions and ToR/controller slow path for policy epochs and topology solves.

4. Fast Path and Slow Path Responsibilities

The fast path should do only bounded, low-latency work. The slow path should perform heavier reasoning, policy compilation, and cross-cluster coordination. This separation prevents the optimization layer from becoming another source of production risk.

Clean separation of low-latency enforcement and policy orchestration

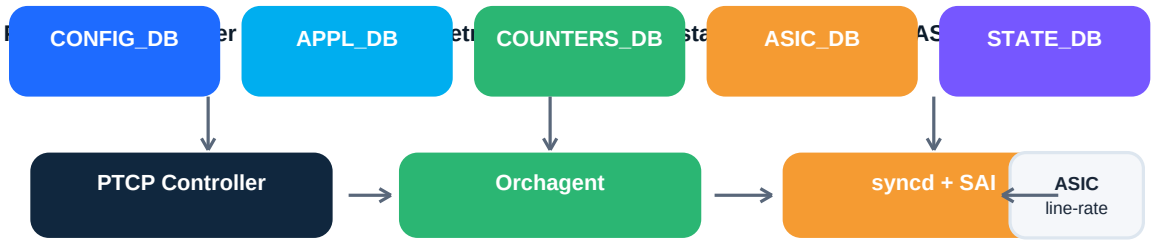


Layer	Example controls	Why it matters for token metrics
Fast eBPF / DPU path	Classify GPU, RAG, checkpoint, and control flows; mark queues; expose counters; block invalid tool paths.	Protects token-critical flows from noisy neighbors and large background movements.
Slow controller path	Solve topology; compute geodesic path choices; update SONiC intent; schedule canary rollout; stage rollback.	Improves sustained cluster throughput while keeping changes auditable and reversible.
Verification path	Compare pre/post tokens/sec, p99 latency, retrieval latency, and GPU idle time.	Prevents marketing claims from replacing measurable proof.

5. SONiC Top-of-Rack Pattern

SONiC is relevant because many AI data centers use open switch operating systems, Redis-backed state databases, containerized switch services, and the Switch Abstraction Interface to program ASICs in a vendor-neutral way. PTCP should use those boundaries rather than bypass them.

SONiC ToR slow-path integration pattern

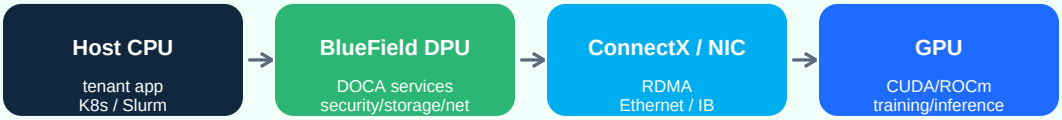


- **Read:** collect link, port, queue, route, ECN/PFC, and counter data from SONiC databases and telemetry collectors.
- **Decide:** compute topology-aware policy epochs outside the line-rate ASIC path.
- **Write:** publish safe intent through sanctioned SONiC/SAI control points so syncd and the vendor SDK remain responsible for ASIC programming.
- **Rollback:** keep known-good control-plane state and signed change evidence.

6. BlueField DPU / Infrastructure Processor Pattern

A DPU or IPU can host infrastructure services outside the tenant workload domain. In a PTCP AI Factory design, the DPU is a strong fast-path location for AI-flow classification, storage offload, security separation, and telemetry collection close to the server.

DPU-resident fast path and infrastructure service isolation



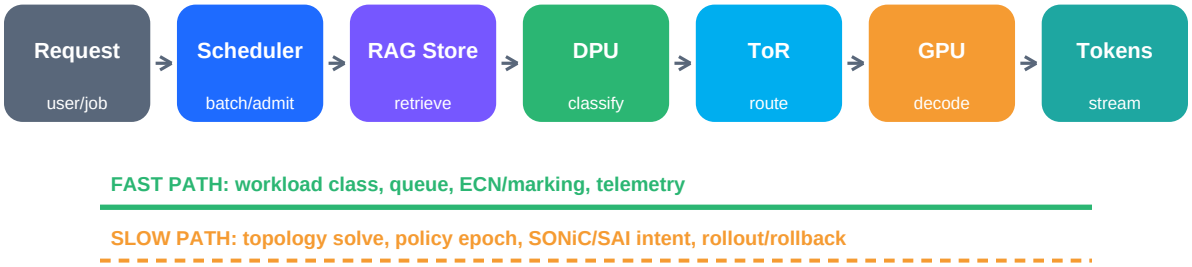
- PTCP fast path agents can execute close to packet ingress/egress.
- DOCA-style services allow infrastructure isolation, offload, storage and security acceleration.
- Policy remains auditable: signed bundles record what was changed and why.

- **Near-host visibility:** identify model serving, all-reduce, RAG, checkpoint, and control-plane flows before they stress the fabric.
- **Isolation:** separate infrastructure services from the application host, reducing CPU interference and limiting blast radius.
- **Storage and RDMA integration:** support high-value path decisions for vector store, object store, NVMe, and RDMA/UCX-like movement.
- **Security posture:** DPU/IPU placement can enforce allowed paths without reading payloads or modifying model logic.

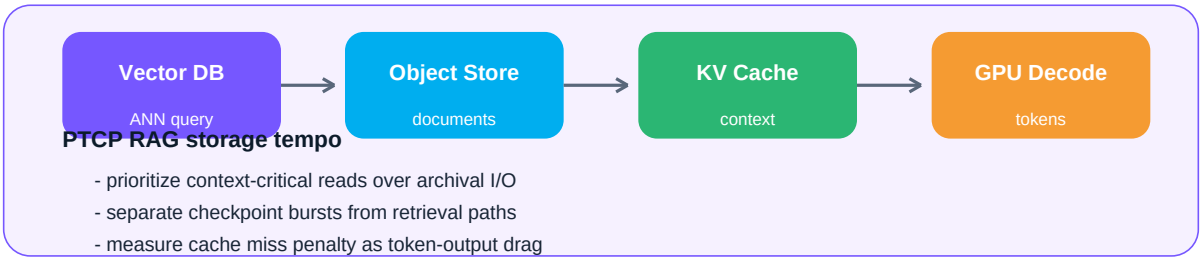
7. GPU -> DPU -> ToR -> Storage Flow

The core data-center flow for PTCP AI Factory is not a single packet path. It is a system graph linking request admission, model placement, context retrieval, NIC/DPU classification, ToR fabric behavior, storage locality, and GPU decode work.

Fast-path and slow-path overlay across the token generation pipeline



RAG and KV-cache storage as first-class token paths



PTCP RAG storage tempo

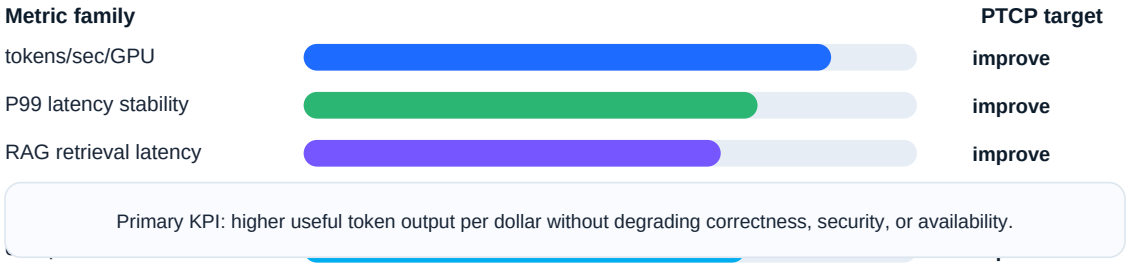
- prioritize context-critical reads over archival I/O
- separate checkpoint bursts from retrieval paths
- measure cache miss penalty as token-output drag
- feed ACISO evidence with storage and identity context

- The fast path protects what directly affects user-visible tokens: context retrieval, KV-cache, token streaming, and control messages.
- The slow path prevents pathological behavior: queue collapse, checkpoint storms, congestion spread, and topology choices that amplify storage latency.

8. Token Metrics Improved by PTCP AI Factory

A serious AI Factory claim must be expressed through token economics, not generic network performance. The performance question is: how much more useful token output does the same infrastructure produce at the same or lower risk?

Token productivity scorecard



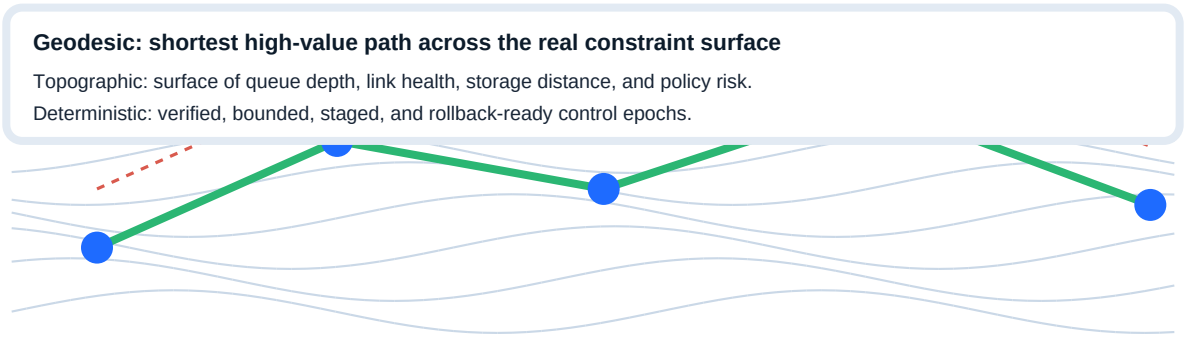
Primary KPI: higher useful token output per dollar without degrading correctness, security, or availability.

Metric	Definition	PTCP improvement mechanism
Cluster tokens/sec	Total useful generated tokens per second across model replicas.	Reduces stalls from storage, network, scheduling, and queue contention.
Tokens/sec/GPU	Useful generated tokens per GPU, adjusted for model and batch profile.	Recovers stranded GPU cycles and prevents uneven feeding of accelerators.
Time to first token	Request arrival to first streamed token.	Prioritizes RAG retrieval, context setup, and admission timing.
P95/P99 inter-token latency	Tail jitter between token emissions.	Separates critical flows from checkpoints, bulk replication, and background IO.
Cost per million tokens	All-in cost divided by useful token output.	Improves utilization before buying incremental GPU capacity.

9. Geodesic, Topographic, and Deterministic

These terms are useful only if they map to measurable controls. In PTCP AI Factory, geodesic means path selection over a real constraint surface; topographic means representing queues, links, storage, and policy as a navigable state space; deterministic means bounded, verifiable, rollback-ready actions.

From common topology to geodesic, topographic, deterministic routing



Term	Operational definition	Data-center implication
Geodesic	Preferred path across a cost surface, not simply hop count.	Routes around queue depth, failing links, hot storage, or policy risk.
Topographic	Cluster state represented as a surface of link, queue, storage, GPU, and identity constraints.	Operators can reason about why one flow should avoid another.
Deterministic	Controls are compiled, bounded, staged, and verifiable.	Prevents self-inflicted outages from optimization experiments.

10. NVIDIA Infrastructure Relevance

For NVIDIA-centered environments, PTCP AI Factory should complement the CUDA, NCCL, NVLink/NVSwitch, ConnectX, BlueField, Spectrum-X, InfiniBand, and Ethernet stack rather than replacing it. The strategic value is overlay control: measuring where token productivity is constrained and placing bounded controls near the DPU, NIC, or ToR layer.

Vendor-neutral overlay position

PTCP is an overlay: it complements the accelerator, NIC/DPU, switch ASIC, storage, and orchestration stack.

It should not claim vendor certification or line-rate ASIC replacement without customer-specific validation.

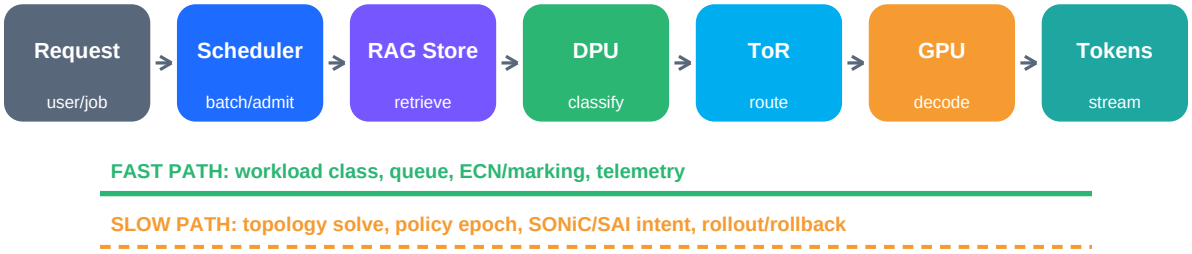
- **BlueField / DOCA:** a natural PTCP fast-path location for infrastructure isolation, networking, storage, and security services.
- **ConnectX / RDMA / GPUDirect:** PTCP telemetry can classify and prioritize movement that directly affects GPU training or inference flow.
- **Spectrum-X / Ethernet / InfiniBand fabrics:** PTCP should observe congestion, ECN/PFC behavior, route stability, and tail latency at fabric boundaries.
- **Token impact:** the goal is not merely faster links; it is fewer GPU stalls, better RAG path predictability, and more consistent streamed-token output.

11. AMD Infrastructure Relevance

For AMD-centered environments, PTCP AI Factory can align with Instinct accelerators, ROCm software, EPYC hosts, and Pensando-style infrastructure acceleration. The emphasis is vendor-neutral Ethernet and storage/RAG flow discipline around high-memory accelerators.

- **Instinct MI300-class accelerators:** high HBM capacity and bandwidth make storage and context-path discipline strategically important for inference and RAG-heavy use cases.
- **ROCm:** PTCP does not replace model software; it surrounds the cluster with telemetry, control, and policy enforcement.
- **Pensando/DPU-like control points:** classify, isolate, and observe infrastructure services near the server where host CPU contention and tenant isolation matter.
- **Ethernet fabrics:** PTCP can map flow criticality across open switching, RoCE-like paths, and storage neighborhoods.

AMD-style flow emphasis: HBM, ROCm jobs, Ethernet, and RAG storage

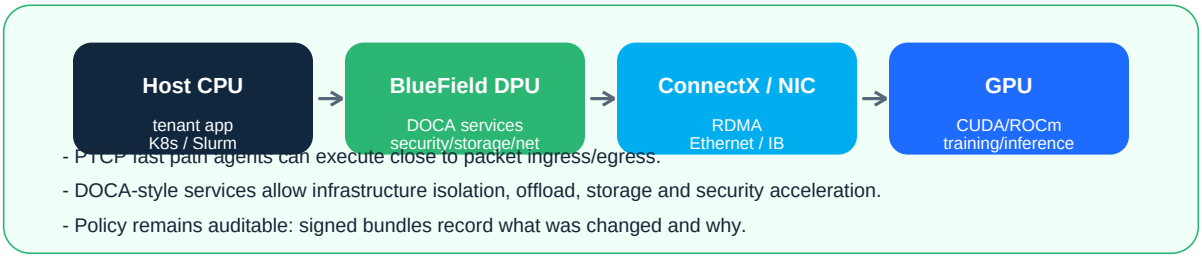


12. Intel Infrastructure Relevance

For Intel-centered environments, PTCP AI Factory is especially relevant to Ethernet-scale AI fabrics, Gaudi-style deployments, Xeon hosts, IPU-based offload, and RAG workloads that benefit from open networking and infrastructure separation.

- **Gaudi 3:** publicly positioned around standard Ethernet infrastructure and enterprise RAG relevance, making topology-aware Ethernet control commercially important.
- **Intel IPU:** provides infrastructure isolation and task offload opportunities for storage, networking, and security services.
- **Xeon and host software:** PTCP can reduce host overhead by moving selected infrastructure enforcement into DPU/IPU or switch-control contexts.
- **Token impact:** the same PTCP scorecard applies: higher tokens/sec, lower tail latency, faster RAG retrieval, and less idle accelerator time.

Infrastructure processing pattern also maps to Intel IPU-style offload



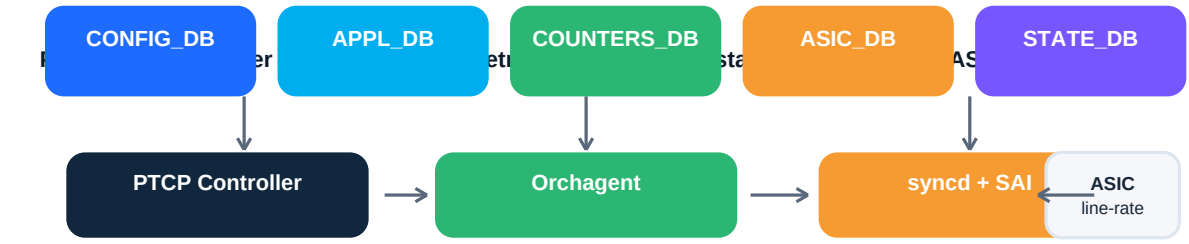
- PTCP fast path agents can execute close to packet ingress/egress.
- DOCA-style services allow infrastructure isolation, offload, storage and security acceleration.
- Policy remains auditable: signed bundles record what was changed and why.

13. Broadcom and Merchant Silicon Relevance

Broadcom switch silicon is central to many Ethernet data-center fabrics. PTCP AI Factory should treat the ASIC as the line-rate truth and use SONiC/SAI-compatible slow-path orchestration for safe policy changes.

- **Tomahawk-class switching:** PTCP should not claim to replace the ASIC. It should exploit the ASIC more effectively by controlling queue intent, topology policy, and flow classes.
- **Thor / NIC / adapter ecosystems:** PTCP can use host and adapter telemetry to understand token-path health across RoCE, Ethernet, and storage flows.
- **SONiC/SAI:** the right abstraction boundary lets PTCP remain portable across vendor ASICs while still capturing precise data-center topology state.
- **Strategic value:** merchant silicon makes open, multi-vendor AI networking possible; PTCP supplies the policy intelligence that common topologies lack by default.

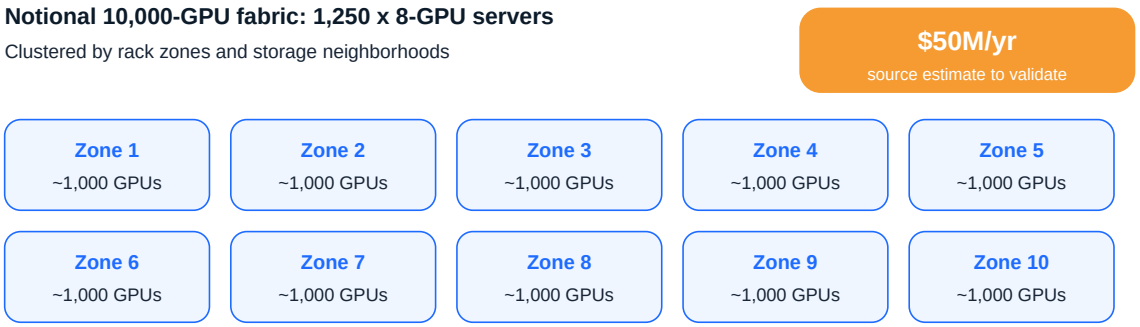
Broadcom-style ToR ASIC integration through SONiC and SAI boundaries



14. 10,000-GPU Scale Economics

At 10,000 GPUs, small utilization improvements create large financial leverage. The Tensor Networks source includes a modeled $50M/year savings figure for a 10K-GPU datacenter. This document treats that figure as a modeling anchor requiring proof-of-value validation in a real customer environment.

Notional 10,000-GPU productivity surface

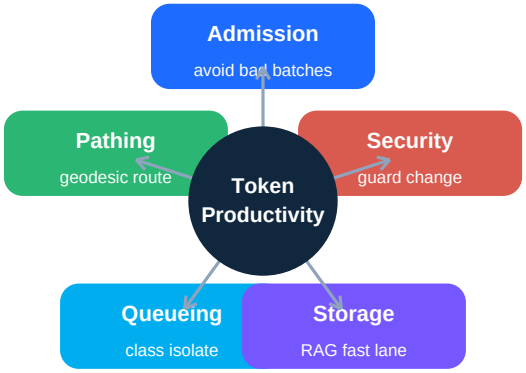


Lever	Example measurable effect	Business interpretation
GPU idle recovery	Reduce time GPUs wait for data, storage, or network flow.	More tokens from existing capex.
RAG path optimization	Lower p95 retrieval latency and cache miss penalty.	Better user experience and lower cost per answer.
Tail-latency stability	Reduce p99 token jitter under mixed workloads.	More reliable service-level agreements.
Checkpoint isolation	Prevent training/checkpoint bursts from starving inference.	Higher blended infrastructure utilization.

15. Specific Performance Mechanisms

PTCP AI Factory improves data-center performance through multiple mechanisms that compound. None should be marketed as magic; each should be proven with explicit measurements.

Mechanism stack that improves token productivity



- **Admission control:** prevents pathological batches or retrieval storms from entering the wrong service path.
- **Path selection:** steers high-value flows over the least-risk topology surface rather than a default shortest path.
- **Queue isolation:** separates token-critical, RAG-critical, checkpoint, replication, telemetry, and background flows.
- **Storage locality:** places retrieval traffic near useful storage and cache neighborhoods when possible.
- **Change safety:** stages controls so optimization cannot become an outage.

16. eBPF and Kernel-Safe Enforcement

eBPF is attractive because it can add networking, security, and observability logic inside the Linux kernel at runtime without changing kernel source code or loading a kernel module. In a PTCP design, eBPF should be used only for bounded, verifier-approved logic such as classification, marking, counting, or enforcement at supported hooks.

Fast-path discipline

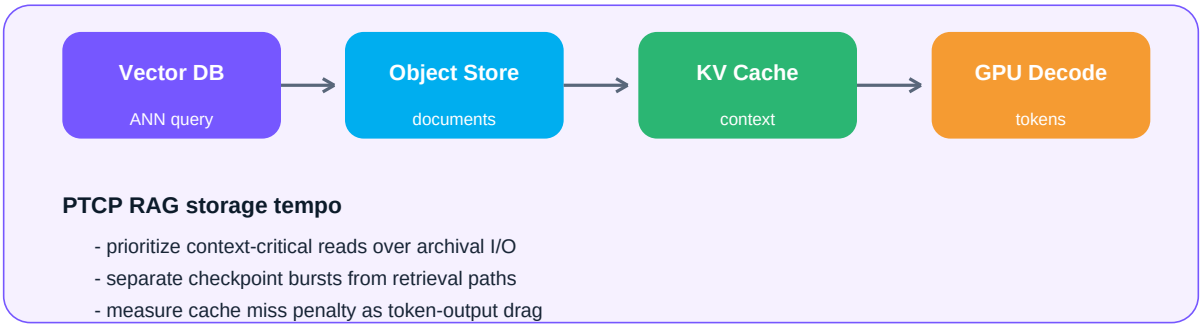
PTCP eBPF programs should be small, auditable, signed, and rollback-ready. They should not become a hidden second network operating system.

eBPF function	PTCP use	Guardrail
Flow classification	Identify GPU, RAG, checkpoint, control, tenant, or storage flow classes.	Use metadata and policy, not invasive payload inspection.
Queue marking	Apply DSCP/TC/priority intent where supported.	Do not override site policy or hardware safety controls.
Telemetry counters	Export per-class counters for topology modeling.	Keep overhead measurable and bounded.
Containment hooks	Block disallowed path classes when integrated with Zero-Day policy.	Require signed policy and rollback.

17. RAG, KV Cache, and Storage Path Controls

Modern inference workloads increasingly depend on retrieval, context extension, and cache management. A PTCP AI Factory implementation should measure these paths as part of the token pipeline rather than leaving storage as a black box.

RAG / KV-cache path controls and measurement points



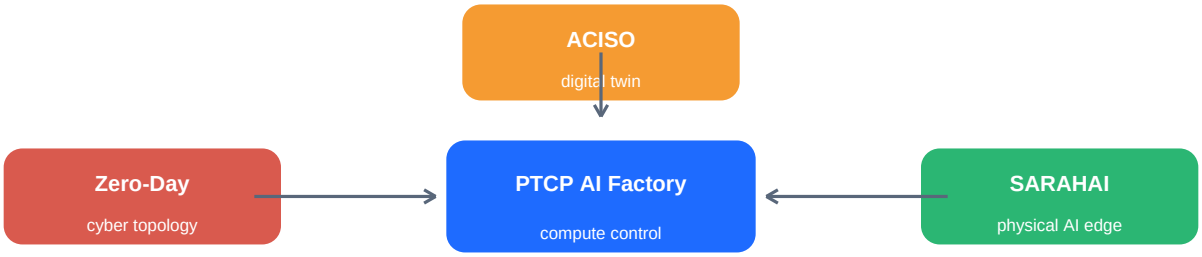
- Measure request-to-vector-query, vector-query-to-document, document-to-context-build, and context-build-to-GPU timing.
- Prioritize retrieval for active inference sessions over background indexing, cold archival reads, and checkpoint bursts.
- Detect storage-neighborhood hot spots and push slow-path policy changes that route around them.
- Track cache miss penalty as a direct contributor to lower tokens/sec and higher time-to-first-token.

18. Security, Reliability, and Change Control

Optimization software inside the data-center fabric must be governed like critical infrastructure. PTCP AI Factory should be deployed with strict role separation, staged rollout, signed artifacts, and immutable evidence.

Risk	Mitigation
Misconfigured fast path causes packet loss.	Verifier-approved programs; canary deployment; automatic rollback; human approval for high-impact actions.
Slow path overwrites operator intent.	Policy diffing; explicit ownership boundaries; SONiC/SAI-compatible integration; signed change records.
Vendor-specific feature drift.	Adapter layer per infrastructure family; do not depend on undocumented ASIC behavior.
Benchmark overclaiming.	Use shadow mode, A/B tests, repeatable workloads, and customer-controlled baselines.
Security exposure from optimization agents.	Least privilege, signed binaries, ACISO evidence, and integration with Zero-Day topology monitoring.

PTCP suite adjacency around AI Factory

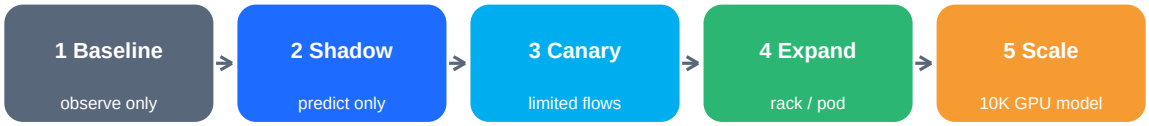


AI Factory can be the performance core; Zero-Day, SARAHAI, and ACISO turn optimization into governed resilience.

19. Proof-of-Value Protocol

A credible PTCP AI Factory deployment should begin as measurement, then move to shadow prediction, then canary control, and only then scale. The objective is to prove improved token metrics without requiring a disruptive data-center redesign.

Validation sequence for customer proof-of-value



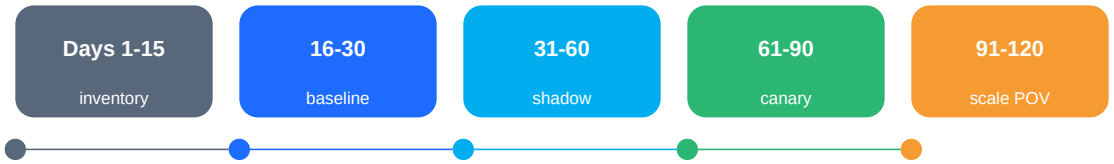
Validation principle: never ship an optimization claim without pre/post metrics, rollback evidence, and workload repeatability.
Required metrics: tokens/sec/GPU, p95/p99 latency, GPU idle, RDMA retransmit, RAG latency, queue depth, cost/token.

Test family	Workload examples	Success metric
LLM inference	batch serving, streaming chat, mixture of batch sizes	tokens/sec, TTFT, p99 inter-token latency
RAG pipeline	vector search, document fetch, context assembly	p95 retrieval, cache misses, answer latency
Training / fine-tune	collectives, checkpoint, gradient sync	MFU proxy, stall time, retransmit, queue depth
Mixed workload	inference + RAG + checkpoint + indexing	SLA preservation under contention
Failure scenario	link degradation, hot rack, storage slow zone	route stability and rollback evidence

20. 120-Day Deployment Roadmap

This roadmap is designed to create a defensible performance dossier rather than a speculative benchmark. It keeps production risk low while still proving the value of topology-aware control.

120-day PTCP AI Factory deployment roadmap



Deliverable

A signed performance dossier showing where token productivity improved, where risk was reduced, and what controls were changed.

- **Days 1-15:** inventory fabric, accelerators, DPUs/IPUs, storage, RAG services, model serving, and telemetry sources.
- **Days 16-30:** capture baseline token metrics and topology state without control.
- **Days 31-60:** run PTCP in shadow mode; compare predicted bottlenecks to actual token stalls.
- **Days 61-90:** deploy limited fast-path classification and slow-path policy canaries.
- **Days 91-120:** scale to selected pods or zones; deliver signed performance and risk-reduction dossier.

21. Strategic Conclusion

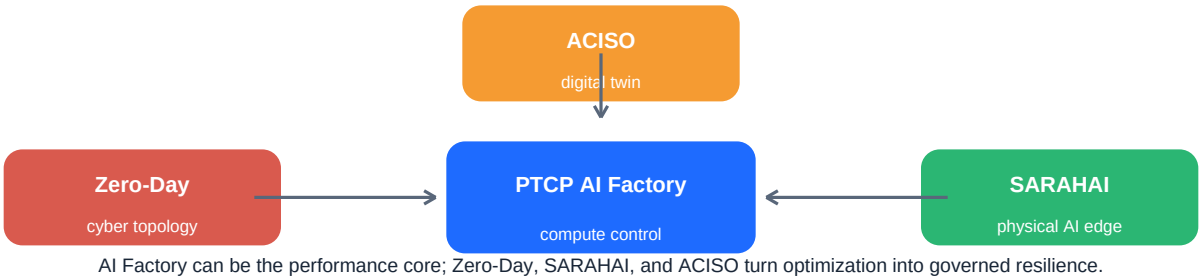
PTCP AI Factory is strategically relevant because AI data centers are becoming cyber-physical production systems. The winning metric is not just raw FLOPs; it is useful, safe, reliable, governed token productivity at scale.

The point

A data center at 10,000 GPUs is a living topology. PTCP AI Factory makes that topology measurable, programmable, and economically accountable, using fast-path enforcement and slow-path intelligence on existing infrastructure.

- For operators, the immediate value is more productive use of existing infrastructure before purchasing additional accelerator capacity.
- For platform teams, the value is one control-plane language across GPU, DPU, ToR, storage, and RAG paths.
- For risk leaders, the value is auditable, rollback-ready optimization rather than opaque automation.
- For licensees, the value is an overlay that can be tailored to NVIDIA, AMD, Intel, Broadcom, SONiC, and DPU/IPU environments without asserting replacement of the underlying hardware stack.

Suite-level advantage: optimization plus cybersecurity plus physical AI plus governance



AI Factory can be the performance core; Zero-Day, SARAHAI, and ACISO turn optimization into governed resilience.

22. Source Notes and Claim Boundaries

This paper is a non-fiction technical and strategic analysis prepared for marketing, investor, licensee, and proof-of-value conversations. It is not a benchmark report, deployment certification, or vendor endorsement statement.

Reference area	Source basis used in this document
Tensor Networks source collateral	Combined Platform whitepaper: AI infrastructure, cybersecurity, cyber-physical operations, PTCP AI Factory, SARAHAI, Zero-Day, ACISO, and 10K-GPU savings model.
eBPF	eBPF.io: kernel-origin technology for safely and efficiently extending OS capabilities at runtime; event hooks, verifier, JIT, maps, networking/security/observability use cases.
SONiC	SONiC architecture wiki: Redis-based state infrastructure, containerized services, APPL_DB/ASIC_DB/COUNTERS_DB, Orchagent, syncd, SAI, and vendor SDK boundary.
NVIDIA	NVIDIA DOCA/BlueField public documentation: BlueField/ConnectX services for offload, acceleration, isolation, networking, storage, security, management, DPDK/SPDK/Linux Netlink, RDMA/GPUDirect.
AMD	AMD MI300X and ROCm public materials: Instinct accelerators for AI/HPC, HBM capacity/bandwidth, ROCm developer ecosystem, Pensando-adjacent infrastructure acceleration.
Intel	Intel Gaudi and IPU public materials: Ethernet-oriented AI accelerator positioning and IPU infrastructure isolation/offload.
Broadcom	Public reporting on Tomahawk/Tomahawk Ultra/Tomahawk 6 Ethernet AI networking and merchant-silicon relevance to large AI clusters.

- **Claim boundary:** PTCP performance benefits must be validated with customer workloads, vendor-qualified environments, and controlled baselines.
- **No endorsement implied:** NVIDIA, AMD, Intel, Broadcom, SONiC, OCP, Linux Foundation, and other names are used for infrastructure compatibility analysis only.
- **Safety boundary:** Any eBPF, DPU, or ToR integration should be delivered under least privilege, signed policy, canary deployment, and rollback controls.