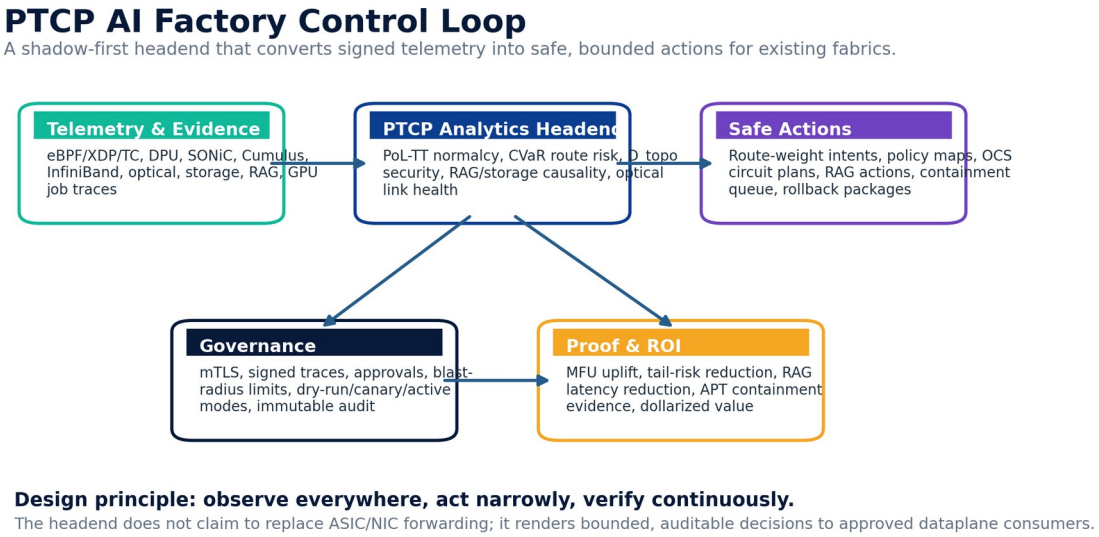
PTCP AI Factory

Topographical, Geodesic, and Deterministic Control for High-Yield AI Datacenters

A post-doctoral professional whitepaper for prospective datacenter, AI factory, MSSP, OEM, Gov/SLED, and infrastructure buyers



PTCP AI Factory converts signed telemetry into safe, evidence-backed control-plane decisions.

Claim discipline

This whitepaper distinguishes implemented capabilities, deterministic internal proof-pack results, and hardware-attested external claims. PTCP AI Factory is a control-plane decision and evidence layer; live dataplane enforcement remains with approved switches, DPUs, controllers, storage systems, or optical devices.

Executive Abstract

AI datacenters have crossed a threshold where the network, storage, and security control planes directly determine usable AI output. Large training jobs create synchronized RoCEv2 elephant flows; ECMP hash collisions and microbursts trigger queue pressure and PFC/ECN events; tail latency forces synchronous GPU jobs to wait for the slowest participant; and storage/RAG latency causes context retrieval to become a first-order performance limiter. At the same time, advanced persistent threats, compromised DPUs/hosts, route abuse, agentic AI misuse, and lateral movement increasingly appear as topology deformation before they appear as simple payload signatures.

PTCP AI Factory addresses these problems by functioning as a topology-native control-plane headend. It ingests signed telemetry from fabrics, DPUs, eBPF/XDP/TC fastpath agents, SONiC/Cumulus switches, InfiniBand telemetry, optical components, storage controllers, RAG pipelines, and AI workload systems. The headend then applies Pattern-of-Life Tensor Train reasoning, CVaR-aware geodesic route selection, D_topo topological security scoring, RAG/storage causal analysis, and deterministic safe-envelope governance to recommend and stage bounded actions.

The business value is the conversion of noisy infrastructure telemetry into provable operational decisions: lower tail risk, better GPU yield, faster RAG retrieval, faster APT containment, safer change control, and dollarized evidence for operators and executives. PTCP AI Factory does not ask operators to replace their existing fabric. It augments SONiC, Cumulus, DPU/eBPF agents, optical systems, storage controllers, SIEM/SOAR, microsegmentation, and AI workload schedulers with a vendor-neutral decision and proof layer.

The latest diligence-closure architecture strengthens the product posture by requiring signed traces, explicit claim boundaries, evidence-required proof packs, visual eBPF fastpath fleet management, and a decoupled frontend bundle. Deterministic internal proof-pack results show material simulated/internal improvements, including +13 percentage points of MFU uplift, 37.8% RAG latency reduction, and 65.9% tail-risk reduction in the v14.2 self-test. These are not live hardware claims; they are repeatable internal proof-pack results and should be converted into external claims only with named, hardware-attested traces.

Buyer takeaway

PTCP AI Factory is valuable because it turns AI factory topology into an executable management object: observe the fabric, model the deformation, stage the safest action, verify the result, and convert the outcome into evidence and ROI.

Table of Contents

Glossary of Key Terms

Term	Meaning for Operators
AI Factory	A datacenter or distributed infrastructure environment purpose-built to train, infer, retrieve, and serve AI workloads at scale.
APT	Advanced Persistent Threat: a stealthy attacker or campaign that maintains presence, moves laterally, and exploits trust paths over time.
CVaR	Conditional Value-at-Risk: a tail-risk metric that penalizes rare but severe network or workload outcomes.
D_topo	PTCP topology defect score combining anomaly likelihood, graph-curvature gradient, and cut-capacity deviation.
eBPF/XDP/TC	Linux kernel programmable datapath technologies used for bounded fastpath observation, marking, and policy enforcement when attached to supported hosts, DPUs, or switches.
Geodesic Routing	In PTCP, the lowest-risk route under learned information-geometric link metrics; not a literal spacetime geodesic.
MFU	Model FLOPs Utilization: the fraction of theoretical model compute that is realized as useful work.
OCS	Optical Circuit Switch: an optical switching element used to create direct optical paths in AI and cloud networks.
PoL-TT	Pattern-of-Life Tensor Train: a bounded-rank telemetry normalcy model for high-dimensional infrastructure behavior.
RAG	Retrieval-Augmented Generation: an AI architecture that combines external retrieval with LLM generation to ground responses.
Safe Envelope	A policy boundary for actions, including approval gates, blast-radius limits, rollout stages, rollback, and post-action proof.
Topological Security	Security analysis based on changes in graph structure, trust, cuts, routes, curvature, and behavior rather than payload content alone.

1. The AI Factory Problem Landscape

AI workloads expose failure modes that traditional datacenter designs were not optimized to control. Training clusters rely on synchronized communication across GPUs. When heavy flows collide on a small set of paths, switch buffers fill, PFC pause frames propagate, and GPU work stalls behind the slowest collective operation. The Geodesic Datacenter paper frames this as the RoCEv2 elephant-flow, ECMP microburst, PFC storm, tail-latency, and MFU problem that AI factories must solve.

Modern AI operators also face a storage/RAG performance problem. A model can only generate quickly if context retrieval, vector search, object fetch, metadata access, and cache placement are aligned with the fabric topology. RAG latency increasingly becomes a performance limiter for inference, agentic workflows, and enterprise knowledge applications.

Security has become similarly topological. APTs and agentic AI risks frequently manifest as lateral movement, privilege drift, route abuse, stealthy data movement, compromised DPU/host behavior, or unusual service dependencies. These patterns may be visible as graph deformation even when payloads are encrypted, compressed, or impractical to inspect at scale.

Problem	Operational consequence
RoCEv2 elephant-flow collisions	Heavy synchronized GPU flows hash onto the same physical links, producing microbursts.
PFC / ECN pressure	Queue pressure and lossless Ethernet control feedback create tail-latency and cluster-wide stalls.
GPU underutilization	MFU drops when all-reduce or data retrieval stalls make accelerators wait.
RAG/storage hotspots	Vector search, object retrieval, metadata, and cache locality cause inference delays.
APT and AI-era security risk	Stealthy lateral movement or agentic misuse deforms trust and cut structure before obvious signature alerts.
Multivendor complexity	Operators manage SONiC, Cumulus, DPUs, optics, storage, SIEM/SOAR, and schedulers as disconnected systems.

Table 1. Top-of-mind AI datacenter problems addressed by PTCP AI Factory.

2. What PTCP AI Factory Is

PTCP AI Factory is a topology-native control-plane and evidence headend. It does not replace line-rate packet forwarding. Instead, it computes safe decisions and publishes route-weight intents, policy-map outputs, circuit plans, containment recommendations, and proof artifacts to the underlying systems that already own dataplane authority.

The product combines five functions: a signed telemetry ingest plane; a mathematical inference plane based on PoL-TT, CVaR, D_topo, and optical/storage scoring; a safe action rendering plane; a visual operator experience; and a proof/ROI plane. This makes it a control-plane product rather than a single observability dashboard.

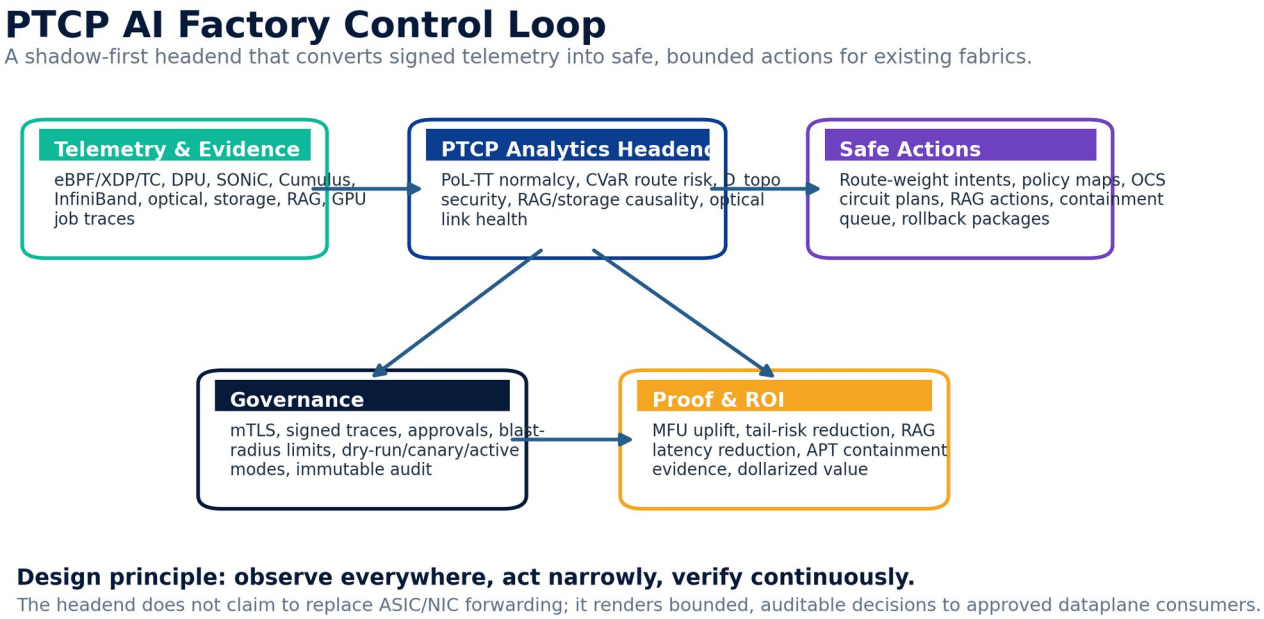


Figure 2. PTCP AI Factory control loop: signed evidence, mathematical inference, safe actions, governance, and proof.

3. Architecture and Data Flow

PTCP AI Factory uses a split architecture. The slow path is the headend: it performs analytics, proof, policy generation, UI/UX, and governance. The fast path is the eBPF/XDP/TC and DPU client layer: it observes flows, exports telemetry, enforces signed maps only when approved, and returns kernel-level state to the headend. Optical transceivers and OCS systems are treated as telemetry/control endpoints rather than generic Linux eBPF runtimes.

The architecture is deliberately vendor-neutral. It can ingest telemetry from SONiC, Cumulus Linux, DPUs, InfiniBand, optical components, storage controllers, Prometheus, OpenTelemetry, and AI workload metadata. It then stages actions through existing fabric controllers, NOS paths, DPU policy maps, OCS control interfaces, storage/RAG controls, SIEM/SOAR, or ITSM systems.

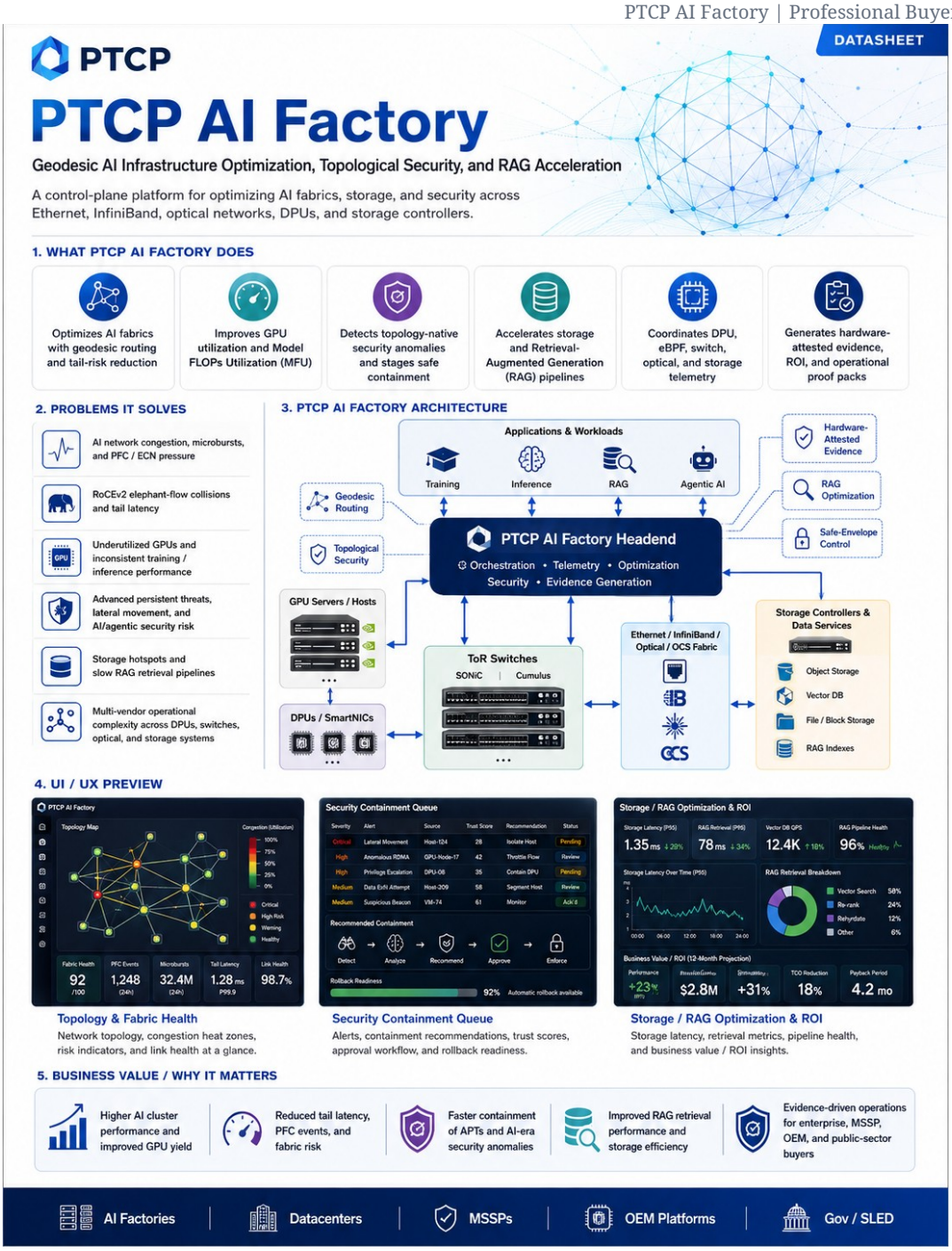


Figure 3. PTCP AI Factory datasheet overview: what it does, problems solved, architecture, UI/UX previews, and buyer value.

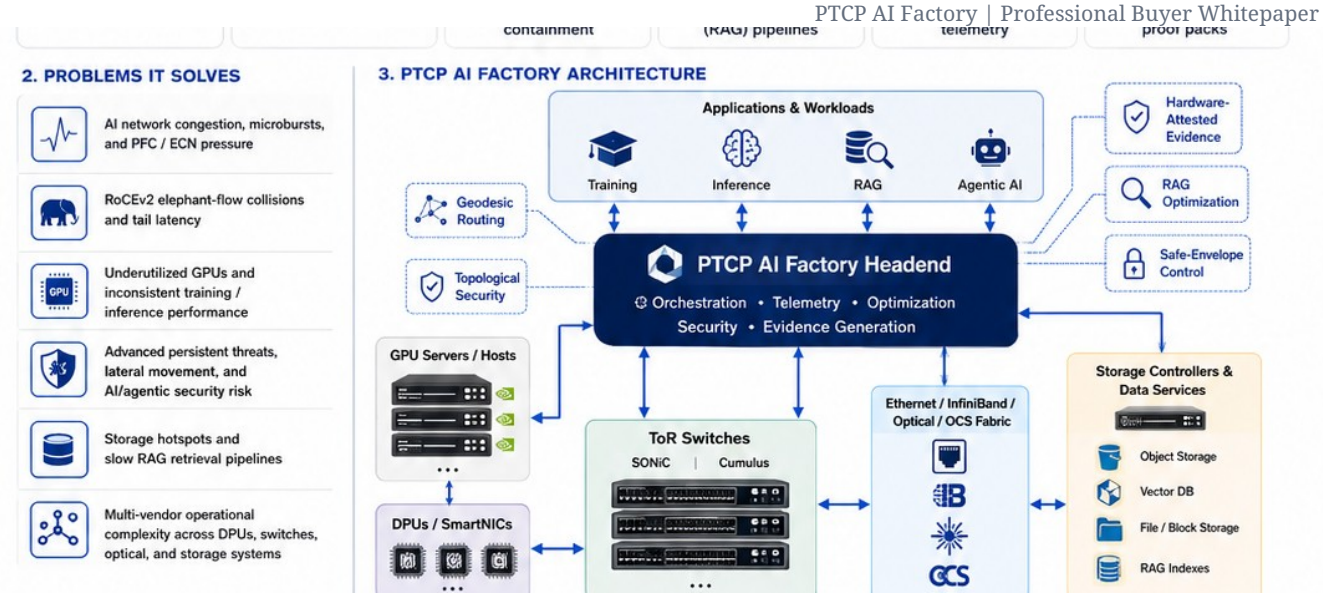


Figure 4. Datasheet architecture view: PTCP AI Factory headend coordinating workloads, DPUs, switches, fabric, optical/OCS, storage, RAG, and evidence.

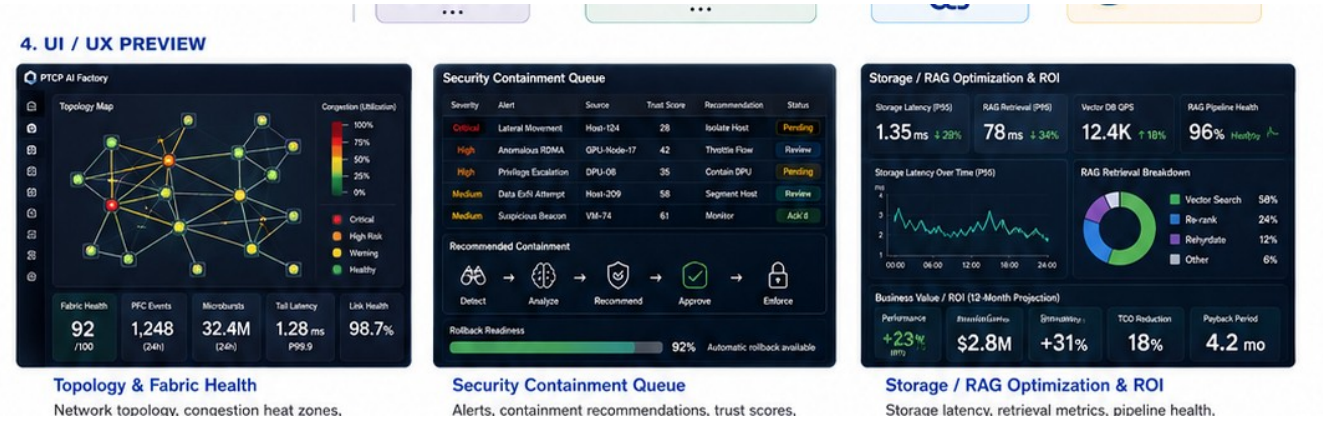


Figure 5. Operator UI/UX previews: topology/fabric health, security containment queue, and storage/RAG optimization/ROI.

4. How PTCP AI Factory Increases Datacenter Performance

Mechanism	Performance value
Predictive fabric scoring	PTCP converts queue, loss, jitter, trust, optical health, PFC/ECN, and AI job context into normalized topology state rather than treating each counter independently.
Tail-risk routing	Candidate routes are evaluated not only by mean cost but by CVaR tail risk, reducing exposure to rare but expensive congestion states.
DPU/eBPF fastpath observation	XDP/TC clients export fastpath evidence and can enforce signed, bounded maps under approval. The Python headend remains a slowpath controller.
RAG/storage locality	PTCP correlates vector search, object retrieval, metadata latency, cache hit rate, and GPU idle to identify end-to-end

	retrieval bottlenecks.
Optical geodesic planning	OCS and optical telemetry allow the headend to recommend circuit schedules and path changes that align optical capacity with AI communication phases.
Proof-before-claims	Signed traces and hardware profiles allow operators to convert internal recommendations into measurable before/after evidence.

Table 2. PTCP AI Factory performance mechanisms.

5. How PTCP AI Factory Addresses Advanced Persistent Threats

APTs are rarely just single malicious packets. They are campaigns that reshape trust paths, dependencies, lateral movement options, and data movement over time. PTCP AI Factory treats this as a topology problem: a cyberattack is a deformation of telemetry distribution and graph geometry. D_topo captures the combination of anomaly evidence, graph-curvature changes, and cut-capacity perturbation.

This is especially relevant in encrypted, AI-driven, and high-throughput environments. Payload inspection is expensive and incomplete; PTCP instead asks whether the topology itself is becoming unsafe. The containment outcome is not automatically block-everything. The system stages the smallest safe containment or drain action, ties it to a rationale, requires approval when blast radius is material, and verifies the outcome with signed traces and audit evidence.

For agentic AI risk, the same model applies. Prompt injection, tool misuse, autonomous scanning, policy poisoning, shadow workloads, and high-bandwidth exfiltration appear as changes in trust, path, rate, graph cut, or RAG/storage access behavior. PTCP can elevate these manifestations into a security containment queue without making unverifiable assumptions about the content of every packet.

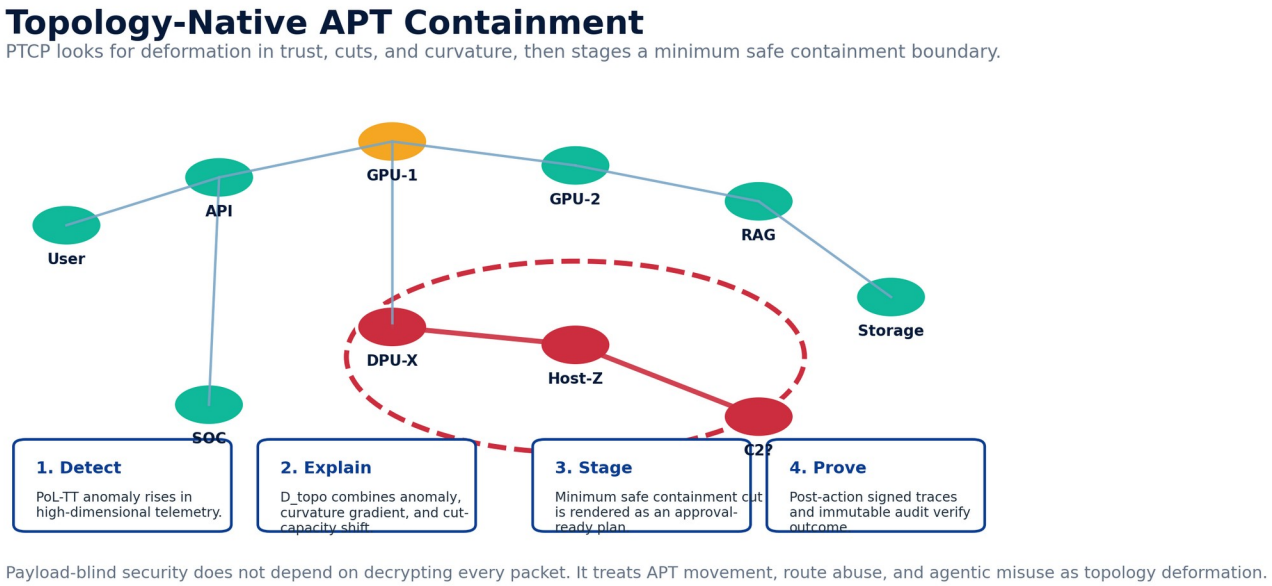


Figure 6. APT and AI-era threat containment through topology-native evidence and safe-envelope response.

6. Topographical, Geodesic, and Deterministic Differentiation

Topographical means PTCP evaluates the network, storage, optical, and security environment as a graph with capacity, distance, cuts, trust, and curvature. This differs from treating the AI factory as a collection of isolated counters.

Geodesic means PTCP searches for the minimum-risk path or action under learned information-geometric link metrics. In the PTCP paper, geodesic routing is explicitly a control-plane metric; it is not a physical spacetime claim. TNQG contributes a capacity/cut/distance vocabulary as an operational modeling layer, not as a completed proof of quantum gravity.

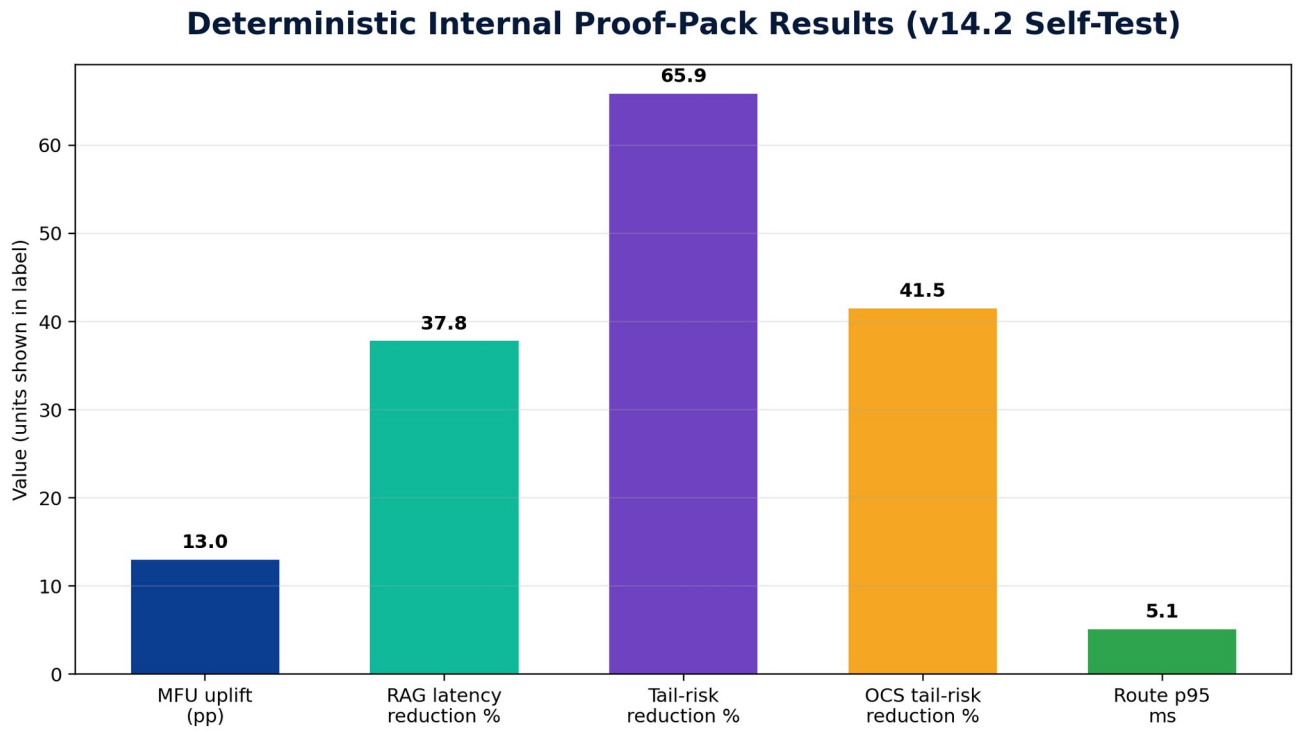
Deterministic means recommendations are reproducible, hashable, and governed. Signed traces, policy-map hashes, mTLS identities, route/action plan hashes, and immutable evidence packs allow operators to verify how a recommendation was produced. This is the opposite of opaque auto-remediation.

Commercial differentiation

PTCP AI Factory is not just another AI fabric dashboard. It is a vendor-neutral decision layer that connects performance, RAG/storage, optical, eBPF/DPU telemetry, and topological security through the same graph-based evidence model.

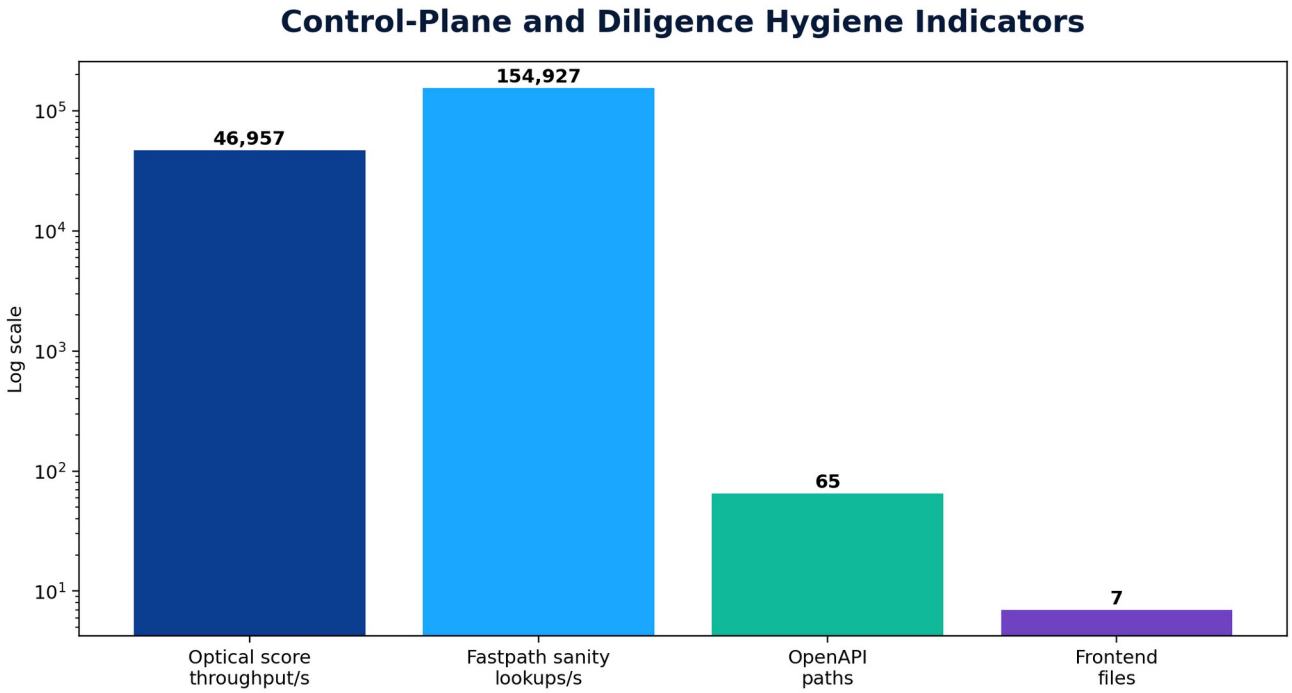
7. Benchmark Results and Evidence Boundary

The following values are deterministic internal proof-pack results from the v14.2 self-test and performance-test harness. They are useful for repeatability and engineering validation, but they are not live customer hardware claims. External claims should be made only from named hardware-attested traces.



Boundary: deterministic internal validation only. External performance claims require named hardware-attested traces.

Figure 7. v14.2 deterministic internal proof-pack results: selected performance and risk indicators.



Fastpath lookups/s are userspace sanity checks, not line-rate NIC/DPU claims. v14.2 also reports line_rate_claim=false.

Figure 8. Control-plane and diligence hygiene indicators. Log scale; no line-rate claim is made.

Metric	Internal result	Interpretation
Trace status	ATTESTED	Signed trace/evidence model accepted.
Proof status	FIELD_MEASURED	Proof-pack path executed with supplied trace data.
MFU uplift	+13.0 percentage points	Modeled internal before/after AI-job improvement.
RAG latency reduction	37.8%	Storage/RAG causal-chain improvement in self-test.
Tail-risk reduction	65.9%	Control-plane risk reduction in deterministic replay.
OCS tail-risk reduction	41.5%	Optical circuit scoring and advisory plan result.
v14 PoL-TT status	TRAINED	TT-SVD density path trained in diligence closure layer.
v14 routing engine	bounded custom router, networkx=false	Diligence-closing route path avoids NetworkX for v14 decision surface.
v14.1 fastpath static status	PASS	eBPF/XDP/TC source and manifest checks passed.

v14.2 UI status	PASS	Enterprise UI/UX and fastpath fleet management checks passed.
Line-rate claim	False	Explicitly false until verified on target hardware.

Table 3. Internal deterministic proof-pack results and claim boundaries.

8. Enterprise UI/UX and eBPF Fastpath Management

The most recent diligence-driven release corrects a major commercialization gap: the UI must not be a raw JSON sandbox. A production buyer needs to see agent lifecycle, fastpath attach state, policy drift, health, trust, telemetry, and proof status in one guided console. PTCP AI Factory now exposes a decoupled frontend bundle, session login, OIDC/SAML metadata, and visual fleet state APIs for eBPF fastpath clients.

The UI objective is not to replace CLI or APIs. It is to reduce operational friction by giving network engineers a guided workflow: enroll agents, confirm mTLS, verify XDP/TC state, stage signed policy maps, watch ring-buffer telemetry, run proof packs, review containment/optimization recommendations, and export ROI/evidence.

Figure 9. eBPF fastpath fleet management UI: attach state, policy sync, trust, and telemetry.

9. Market Comparables and Status-Quo Alternatives

Peer category	What it does well	PTCP AI Factory differentiation
NVIDIA Spectrum-X / AI Ethernet fabrics	Purpose-built Ethernet networking for AI workloads with adaptive routing, congestion control, telemetry, and AI performance claims.	PTCP is complementary: a vendor-neutral control/evidence layer that can reason across multiple fabrics, storage/RAG, security, and optics rather than owning the switch stack.
SONiC / Cumulus / open NOS	Switch operating systems and automation surfaces for white-box or vendor-disaggregated networks.	PTCP renders route/QoS/config diffs, telemetry collectors, and rollback evidence without replacing the NOS.
Cilium / eBPF cloud networking	eBPF dataplane, observability, and	PTCP uses eBPF as a fastpath agent

	security for cloud-native/Kubernetes environments.	substrate, but adds AI-factory topology optimization, RAG/storage causality, optical planning, and proof packs.
Optical circuit switching / OCP OCS	Optical switching projects and hardware provide low-latency, energy-efficient circuit fabrics and open control interfaces.	PTCP schedules, validates, and proves circuit recommendations based on AI workload, optical health, tail risk, and rollback feasibility.
Storage/RAG platforms	Platforms such as WEKA, VAST, NetApp, vector DBs, and object/file systems optimize data and retrieval infrastructure.	PTCP correlates storage behavior with RAG latency, GPU idle, topology, and network path risk to recommend cross-domain actions.
SIEM/SOAR/microsegmentation	Alerting, orchestration, ticketing, and segmentation tools handle security response workflows.	PTCP supplies topology-native risk, minimum safe containment, and signed proof rather than generic alerts or static segments.

Table 4. PTCP AI Factory positioning against adjacent market categories.

10. Adoption Model for Operators

1. Shadow assessment: Deploy collectors only; ingest signed traces; no actuation.

2. Hardware profile certification: Bind traces to SONiC, Cumulus, DPU, InfiniBand, optical, storage, or RAG profiles.

3. Proof-pack generation: Run before/after or replay evidence for fabric, security, storage/RAG, and optical paths.

4. Advisory action staging: Render route intents, policy maps, OCS plans, RAG actions, and containment plans for review.

5. Bounded canary: Apply only to preapproved small scopes after proof, certification, approval, blast-radius, rollback, and audit gates pass.

6. Production operation: Run recurring evidence packs, UI dashboards, API integrations, reporting, and executive ROI.

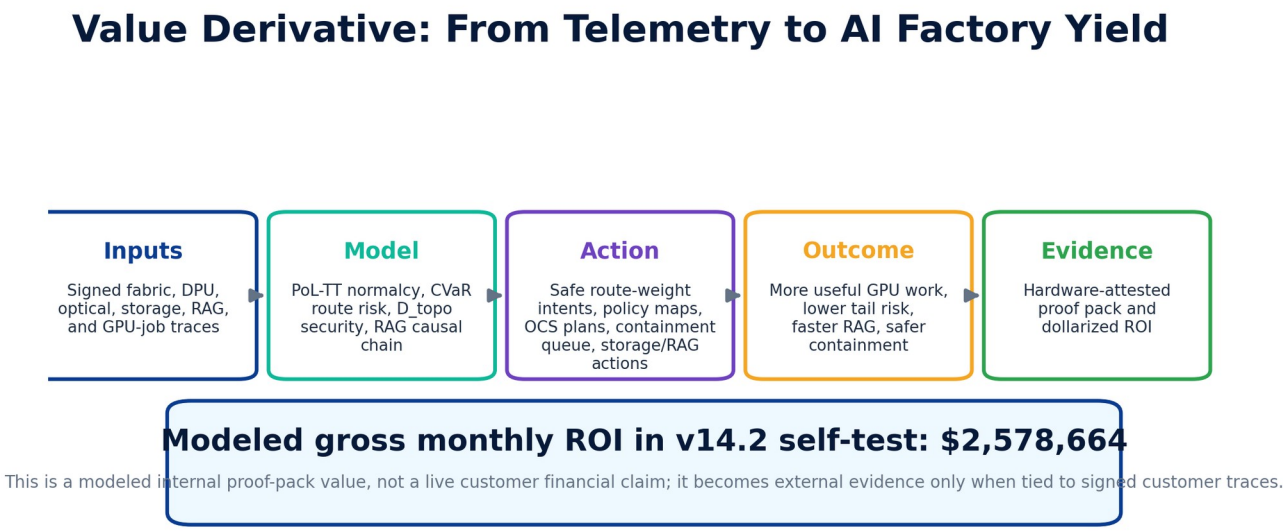


Figure 10. Value derivative: how PTCP converts telemetry into safe actions, evidence, and buyer-facing ROI.

11. Buyer Evaluation Checklist

- Can PTCP ingest signed traces from the buyer's actual fabric, DPU, optical, storage, RAG, and AI workload systems?
- Does the proof pack show named hardware profiles, immutable trace hashes, clock-skew checks, and exact version metadata?
- Are internal results labeled separately from hardware-attested external claims?
- Do fastpath agents show XDP/TC attach state, policy hash, drift, trust score, and ring-buffer telemetry in the UI?
- Can recommendations be exported as route-weight intents, policy maps, config diffs, tickets, or safe OCS/RAG/storage plans?
- Does every action have a safe envelope, approval requirement, blast-radius estimate, rollback plan, and post-action proof?
- Does the ROI report translate technical improvements into recovered GPU-hours, reduced RAG latency, avoided downtime, and analyst-time savings?

Appendix A. Benchmark Methodology and Internal Proof-Pack Results

The benchmark results in this whitepaper are derived from deterministic internal self-test and performance-test harnesses executed against the PTCP AI Factory v14.2 codebase. The harness is designed to validate control-plane behavior, proof-pack generation, UI/UX health, fastpath management APIs, and claim-boundary enforcement.

External claims must be generated from named hardware profiles and signed trace packs. The product intentionally reports line_rate_claim=false in control-plane tests. This protects both operators and Tensor Networks from conflating a headend proof harness with live ASIC, NIC, DPU, or optical device performance.


```
{
 "mfu_uplift_pct_points": 13.0,
 "rag_latency_reduction_pct": 37.804878,
 "tail_risk_reduction_pct": 65.853659,
 "v142_diligence_status": "DILIGENCE_READY",
 "line_rate_claim": false
}
```


Appendix B. API and Integration Surface

API group	Representative routes
Telemetry and evidence	/api/v4/fastpath/telemetry/ingest, /api/v4/validation/field-proof/run, /api/v3/optical/telemetry/ingest
Fastpath fleet	/api/v4/fastpath/agents/register, /api/v4/fastpath/agents/heartbeat, /api/v4/fastpath/fleet/state, /api/v4/fastpath/telemetry/timeseries
Policy and actions	/api/v4/fastpath/policies/compile, /api/v4/fastpath/policies/deploy, /api/v3/optical/circuits/render, /api/v2/fabric/v13/geodesic-route/recommend
Security	/api/v4/security/dtopo/evaluate, /api/v2/security/v13/topological-risk/evaluate
Storage/RAG	/api/v2/storage/v13/telemetry/ingest, /api/v2/rag/v13/causal-chain, /api/v2/rag/v13/action-plan/render
Commercial evidence	/api/v2/evidence/v13/procurement-pack, /api/v2/roi/v13/calculate, /api/v2/licensing/v13/evaluate

Table 5. Representative PTCP AI Factory integration surfaces.

Appendix C. Claim Boundary and Safety Model

PTCP AI Factory is best described as a control-plane decision and evidence layer. The headend computes route-weight intents, risk scores, policy maps, action plans, and proof packs. Dataplane systems - ASICs, NICs, DPUs, optical devices, storage controllers, and fabric controllers - remain the authorities for live enforcement.

Safe operation follows the principle: observe everywhere, act narrowly, verify continuously. Live action is gated by proof-pack pass, hardware profile certification, signed agent identity, mTLS, operator approval, preflight simulation, bounded blast radius, rollback verification, post-action proof, and immutable audit.

TNQG language is used as an operational graph-geometry vocabulary: capacity, cut, distance, curvature, and geodesic-like paths. It is not marketed here as a completed proof of quantum gravity.

Appendix D. References

[1] Tensor Networks, Inc., Predictive Tensor Control Plane (PTCP): Tensor-Train Telemetry, Risk-Aware Geodesic Routing, and Topology-Native Security. Uploaded technical paper.

[2] Tensor Networks, Inc., Tensor-Network Quantum Gravity as an Operational Reconstruction Program. Uploaded technical paper.

[3] Tensor Networks, Inc., The Geodesic Datacenter: Accelerating GPU Ethernet Fabrics and Compute Utilization with PTCP + TNQG. Uploaded whitepaper.

[4] PTCP AI Factory v13.1 / v14.x codebase and diligence-closure artifacts. Uploaded source files and deterministic self-test outputs.

[5] NVIDIA Spectrum-X official product materials: AI Ethernet networking platform, adaptive routing, congestion control, and telemetry.

[6] SONiC Foundation official project materials: open-source Linux-based NOS for multi-vendor switches and ASICs.

[7] Open Compute Project Optical Circuit Switching project materials: OCS interfaces and OpenConfig/gNMI/gNOI/gNSI direction.

[8] NVIDIA DOCA/BlueField official materials: DPU platform for networking, storage, security, and management services.

[9] Lumentum Optical Circuit Switch official materials: MEMS-based optical switching for AI and cloud networks.

[10] Cilium official materials: cloud-native eBPF networking, security, and observability.

[11] WEKA/VAST/Google RAG public materials: RAG as retrieval plus generation, and the need for high-performance retrieval/storage pipelines.