PUBLIC-RELEASE POSTDOCTORAL WHITE PAPER

THE CONTROL LAYER THE AI FACTORY IS MISSING

Why the Predictive Tensor Control Plane (PTCP) Belongs in the Top-Ten Class of AI Data-Center Technologies

Cross-domain optimization across compute, network, storage, and cybersecurity

COMPUTE	NETWORK	STORAGE	CYBER
Accepted work per accelerator-dollar	Lower collective and KV tail latency	State locality and data-path control	Topology-native bounded containment

Tensor Networks, Inc.

440 N Wolfe Rd, MS #03, Sunnyvale, California 94085 | www.ptcp.ai

CAGE 8FD88 | UEI J87LWHL7NNJ3 | Evidence cutoff: July 25, 2026

Top-ten is an analytic technology-class judgment made in this paper, not an external league-table ranking. This non-enabling edition describes capabilities and validation boundaries but intentionally withholds proprietary algorithms, equations, feature engineering, model structures, parameters, thresholds, source code, and implementation recipes.

Executive Abstract

The next constraint on artificial intelligence is not only model quality or accelerator peak FLOPS. It is the ability to turn an expensive, heterogeneous, failure-prone AI data center into a coherent controlled system.

A modern AI factory is a coupled cyber-physical system. Accelerators wait on collective communication; inference engines wait on KV-state transfer; retrieval pipelines wait on storage and network locality; schedulers trade throughput against tail latency; security controls change reachability and resource availability; power and thermal limits bound all of them. Local optimizers can be excellent within a layer and still produce a globally inferior outcome because they optimize different objectives, operate at different timescales, and observe only fragments of system state.

The Predictive Tensor Control Plane (PTCP) is designed for the missing coordination function. It uses proprietary cross-domain analytics to develop a governed operational view, identify consequential change, evaluate risk-sensitive alternatives, and authorize only bounded actions. A deliberative analytic path supports learning, replay, review, and authorization; a constrained enforcement path applies approved, time-limited intent through existing infrastructure interfaces. The proprietary state model, scoring composition, optimization objective, parameters, thresholds, training process, and compilation logic are intentionally withheld. PTCP remains an overlay rather than a demand to replace the installed accelerator, network, storage, or cyber stack.

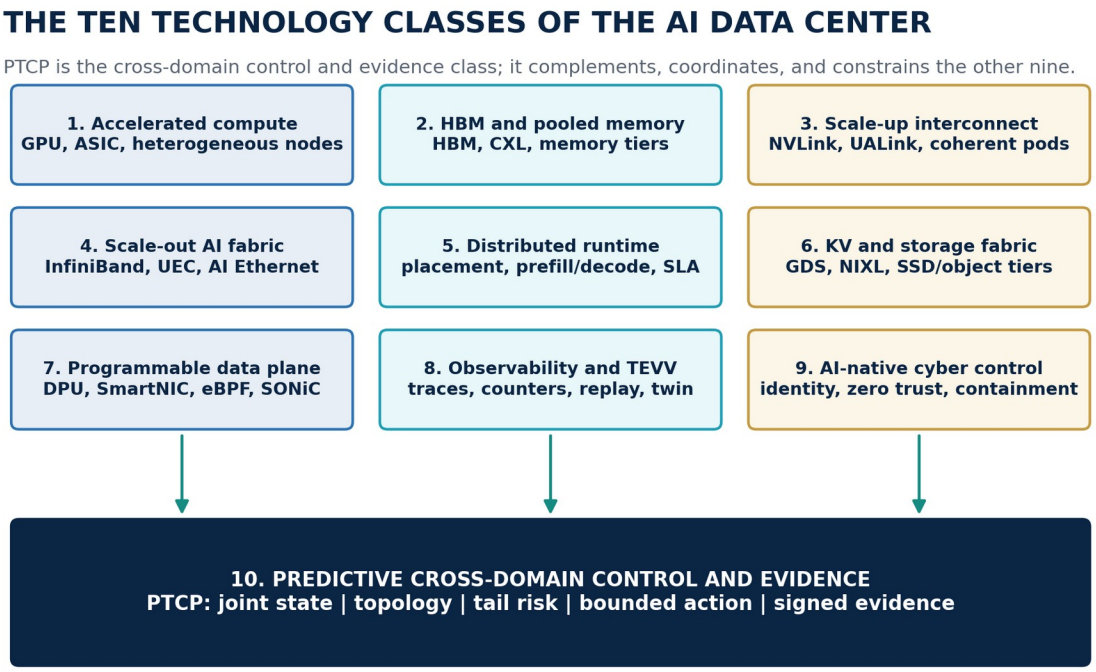

Figure 1. A functional taxonomy of the ten technology classes that determine AI-factory outcomes.

Author taxonomy based on accelerator, memory, interconnect, runtime, data-movement, observability, and security dependencies.

TOP-TEN FINDING PTCP scores 43 of 50 on the paper's disclosed qualification rubric. Its defining strengths are cross-domain reach, bottleneck leverage, auditability, and performance-security convergence. Its material weakness is independent field maturity, scored 2 of 5. The top-ten conclusion therefore supports evaluation and strategic priority, not unrestricted production deployment.

The market need is structural. Gartner forecasts $787.99 billion in worldwide data-center systems spending for 2026, up 55.8 percent, and $2.59 trillion in worldwide AI spending. The International Energy Agency reports that data centers consumed about 415 TWh in 2024 and projects about 945 TWh by 2030; it also estimates roughly 20 percent of planned data-center projects could face delay risk unless grid constraints are addressed. These figures do not prove demand for PTCP, but they prove that small percentage improvements in accepted work, latency, resilience, or containment act on a very large and increasingly constrained capital base. [19]-[22]

Domain	PTCP control objective	Primary measurable outcome	Safety guardrail
Compute	Place, resize, pace, or protect work using coupled state	Accepted output/$, MFU, job completion, p99 latency	No priority inversion or protected-job starvation
Network	Predict congestion and route or pace within policy	Collective time, queue tail, retransmit, path stability	No oscillation, unsafe reroute, or fabric non-regression
Storage	Preserve KV and data locality across memory and storage tiers	TTFT, cache hit, stall time, bytes moved, SSD write load	No corruption, privacy breach, or pathological cache churn
Cyber	Detect topology deformation and apply bounded containment	Detection lead time, false-positive cost, containment time	No unauthorized isolation; signed expiry and rollback

Executive decision rule

Decision	Rationale
Prioritize evaluation	The structural leverage and cross-domain reach justify top-tier technical and commercial diligence.
Withhold broad authority	Independent production superiority is not established; begin in shadow and advisory modes.
Test a coupled wedge	Select a use case where strong native compute, fabric, storage, and security controls leave measurable residual loss.
Scale only with evidence	Expand one target and action class at a time after causal benefit, guardrails, rollback, and operating economics pass.

Contents

The editable edition uses real Word Heading styles; the publication PDF preserves the same hierarchy.

Document Status, Method, and Claim Discipline

CORE CONCLUSION PTCP belongs in the top-ten class of AI data-center enabling technologies because it addresses a missing systems function: predictive, cross-domain, safety-bounded control over compute, network, storage, and cyber state. This is a reasoned technology-class conclusion, not a claim that PTCP has been independently ranked tenth or has already outperformed every incumbent in production.

The paper uses four methods. First, it constructs a technology taxonomy from the functional dependencies of a modern AI data center. Second, it evaluates PTCP against ten disclosed qualification criteria. Third, it maps PTCP capability areas to causal bottlenecks and measurable outcomes in four domains. Fourth, it separates scientific validity, implementation evidence, vendor-controlled testing, and independent field proof. A conclusion is permitted only at the evidence level actually reached.

Claim class	Status at cutoff	Permissible statement
External ranking	Not found	Do not say an analyst, standards body, or independent survey ranks PTCP in the top ten.
Technology-class thesis	Supported by the disclosed framework	PTCP can be called top-ten-class if the reader accepts the taxonomy and qualification criteria.
Scientific foundations	Established components	The cited mathematical, statistical, graph, and programmable-infrastructure foundations are legitimate; PTCP's proprietary selection, combination, and tuning are withheld.
Implemented system	Substantial Tensor-supplied evidence	PTCP is more than a concept; target-specific compatibility and line-rate behavior still require attestation.
Performance outcomes	Internal or simulated at cutoff	Use internal results as hypotheses and engineering priors, not customer or third-party production claims.
Production superiority	Not publicly established	Require a pre-registered independent proof-of-value against strong combined baselines.

Market figures are used to establish urgency, not to calculate an addressable market by simple addition. Gartner's AI, data-center systems, and information-security categories overlap. Vendor specifications are cited as evidence of design direction and interface reality, not as independent proof of vendor performance. Company-specific implementation and internal test statements are identified as Tensor-supplied evidence.

PUBLIC DISCLOSURE BOUNDARY This edition establishes what PTCP is intended to accomplish, how it fits into an AI data center, what evidence supports its evaluation, and how buyers can validate it. It does not disclose proprietary algorithms, equations, feature definitions, model structures, training or update procedures, scoring composition, optimization objectives, parameters, thresholds, solver logic, source code, policy schemas, or target-specific implementation recipes.

1. The Top-Ten Thesis

PTCP is not top-ten because it is a fashionable algorithm. It is top-ten-class because cross-domain predictive control is a first-order dependency of useful, safe, and economical AI at data-center scale.

A technology class deserves strategic priority when three conditions hold. First, failure or underperformance in that class materially constrains the mission outcome. Second, the class creates leverage over multiple downstream cost or risk centers. Third, no neighboring layer can fully substitute for its function. Accelerators satisfy these conditions because no amount of orchestration creates absent compute. AI fabrics satisfy them because distributed jobs cannot outrun communication. Storage and KV tiers satisfy them because inference cannot reuse state it cannot access. Cyber controls satisfy them because unsafe autonomy is not commercially usable autonomy.

Cross-domain predictive control satisfies the same test. An accelerator scheduler generally cannot infer a malicious reachability change. A switch cannot know whether a short queue is delaying a mission-critical decode stage or harmless batch work. A storage tier cannot know that preserving one KV object will reduce GPU idling more than preserving another. A SIEM can correlate alerts yet rarely prices the performance cost of quarantine. PTCP is intended to supply the shared state, risk objective, action constraint, and evidence chain needed to coordinate those specialized systems.

- Top-ten means a technology class, not a vendor leaderboard.
- PTCP is the proposed implementation of the cross-domain predictive control and evidence class.
- The classification is independent of whether PTCP ultimately wins a given procurement.
- Production superiority remains a hypothesis until independent testing establishes causal benefit.
- The correct strategic response is a gated evaluation with strong baselines and kill criteria.

THESIS IN ONE SENTENCE As AI factories become more expensive, heterogeneous, distributed, and autonomous, the layer that can safely optimize the coupled system becomes as strategically important as the components it coordinates.

2. The Ten Technology Classes

The taxonomy is functional: each class supplies a non-substitutable systems capability, while products may span more than one class.

#	Technology class	Why it is indispensable	Representative interfaces or ecosystems
1	Accelerated and heterogeneous compute	Provides dense tensor, vector, and inference execution	NVIDIA Blackwell, AMD Instinct, AI ASICs, CUDA/ROCm

#	Technology class	Why it is indispensable	Representative interfaces or ecosystems
2	High-bandwidth and pooled memory	Feeds accelerators and enlarges the economical state space	HBM, CXL memory expansion/pooling
3	Scale-up accelerator interconnect	Makes many accelerators act as a coherent pod	NVLink/NVSwitch, UALink
4	Scale-out AI fabric	Carries collectives, KV state, checkpoints, and storage traffic across pods	InfiniBand, Spectrum-X, UEC Ethernet
5	Distributed runtime and scheduling	Places model phases and workers against SLA and resource constraints	Kubernetes, Slurm, Dynamo, vLLM, SGLang
6	KV, data-movement, and storage fabric	Moves and retains model, retrieval, checkpoint, and KV state	GDS, NIXL, KVBM, SSD/object/distributed storage
7	Programmable infrastructure and offload	Creates enforceable, low-latency telemetry and action surfaces	DPU, SmartNIC, eBPF/XDP/TC, SONiC
8	Observability, replay, and TEVV	Makes distributed behavior measurable, reproducible, and governable	OpenTelemetry, DCGM/DOCA telemetry, digital twin
9	AI-native cybersecurity and zero trust	Constrains identities, tools, topology, and autonomous actions	NIST CSF/AI RMF, zero trust, EDR/XDR, SIEM/SOAR
10	Cross-domain predictive control and evidence	Optimizes the coupled objective and proves bounded action	PTCP slow path, safe envelope, fast path, evidence pack

The classes are intentionally not ordered by market capitalization or technical glamour. They are arranged from physical execution toward system coordination. A single rack-scale system already binds these functions tightly: NVIDIA's public GB200 NVL72 specification describes 72 Blackwell GPUs, 13.4 TB of HBM3E, 576 TB/s aggregate HBM bandwidth, and 130 TB/s of NVLink bandwidth in one rack. UALink defines an open scale-up fabric for up to 1,024 accelerators, while the Ultra Ethernet Consortium's current public specification addresses interoperable AI and HPC scale-out communication. CXL 4.0 extends the coherent-memory control surface with 128 GT/s signaling, bundled ports, and expanded memory reliability and error visibility. These are not merely faster parts; they create new control surfaces and new failure couplings. [6]-[10], [31]

2.1 Why PTCP is a distinct class

PTCP is not an inference engine, scheduler, network controller, storage system, or SIEM. Its proposed function is to provide cross-domain operational awareness, policy-bounded recommendations, and evidence linking authorized decisions to outcomes. Its internal state representation, analytical transformations, candidate generation, ranking, and tuning are withheld. A mature runtime may optimize prefill and decode; PTCP can add network, storage, identity, and risk context. A mature fabric may optimize paths; PTCP can tell it which workload externalities and cyber constraints should shape the route. Distinctness arises from the joint objective, not from duplicating every actuator.

3. Why Optimization Must Cross Domains

The AI data center is governed by bottleneck propagation: the slowest coupled stage determines useful output, while rare tail events determine service credibility.

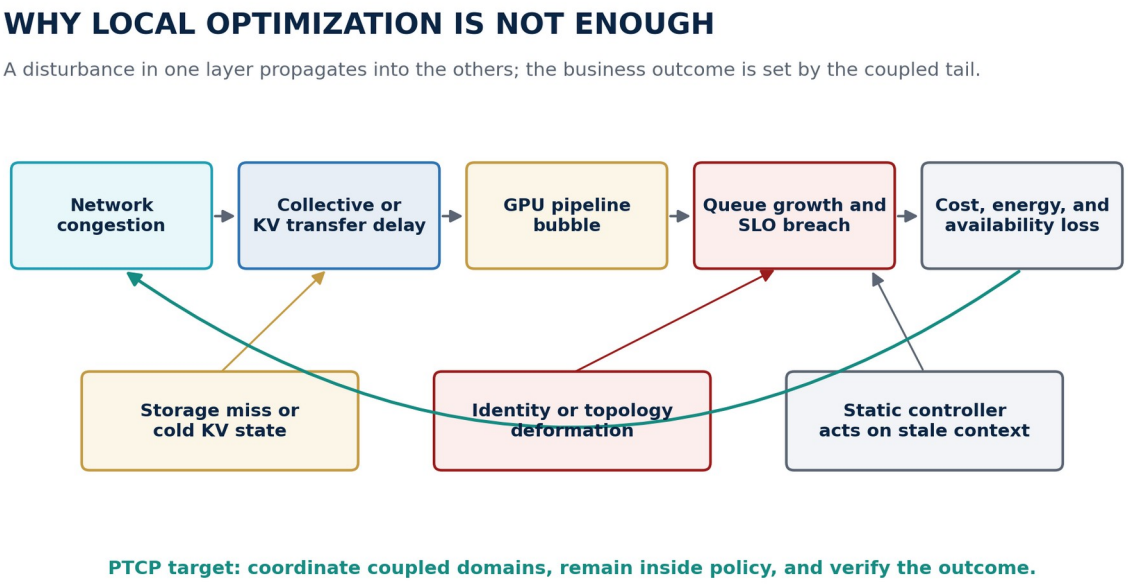

Figure 2. Cross-layer disturbances turn local inefficiency into system-level tail loss.

Author synthesis; the chain is conceptual and should be estimated empirically for each workload.

Distributed training and inference convert communication variance into compute waste. A congested link, slow rank, failing RNIC, asymmetric route, or topology mismatch can delay a collective or KV transfer; downstream accelerators remain allocated but idle; queues accumulate; SLOs fail; energy and capital continue to be consumed. The same propagation can begin in storage with a cold object or in cyber with a control that abruptly removes reachability. The system cost is nonlinear because one straggler can stall many dependent workers and one unsafe response can expand blast radius.

NVIDIA Dynamo's current documentation makes the coupling explicit: disaggregated serving separates prefill and decode but KV transfer becomes a critical path. If RDMA silently falls back to TCP, time-to-first-token and throughput may be dominated by state movement rather than model compute. Its KV Block Manager spans GPU memory, host memory, SSD, and remote storage, creating a placement problem that crosses accelerator, network, and storage domains. This is exactly the type of coupled state for which a higher-order controller can be useful, provided it adds value beyond Dynamo's own router and planner. [11]-[14]

Peer-reviewed production systems reinforce the causal argument. MegaScale co-designed model execution, computation-communication overlap, networking, fault tolerance, and observability, reporting 55.2 percent MFU for a 175-billion-parameter model on 12,288 GPUs. PIPEMORPH shows how congestion, RNIC or switch defects, and topological asymmetry create cascading pipeline bubbles and GPU head-of-line blocking; its reported gains belong to PIPEMORPH, not PTCP. Together they establish

that topology-aware cross-layer coordination is a real systems problem while leaving PTCP-specific advantage open. [33], [34]

Disturbance	Propagation path	Local controller blind spot	PTCP hypothesis
Collective congestion	queue -> straggler -> pipeline bubble -> job delay	Switch sees packets, not mission value	Route or pace using workload criticality and tail risk
Cold or misplaced KV	storage/network fetch -> TTFT -> decode underfill	Cache sees reuse, not global resource pressure	Coordinate KV tier, path, and worker placement
GPU thermal/power cap	lower clocks -> queue growth -> SLO pressure	Scheduler may lack fabric and cyber state	Rebalance or degrade gracefully inside policy
Compromised or rogue agent	identity abuse -> topology change -> resource exhaustion	Security action may ignore performance blast radius	Propose a policy-compliant bounded response with expiry and rollback
Stale telemetry	incorrect state -> harmful control -> oscillation	Each plane may trust its own clock and schema	Reject stale policy; enter local safe mode

3.1 The objective is accepted work, not isolated utilization

Maximizing GPU utilization without quality, latency, safety, or fairness constraints can simply run more low-value work. Minimizing network latency without job context can accelerate the wrong flows. Maximizing cache hit rate can increase SSD write amplification or retain sensitive state too long. PTCP should therefore optimize accepted work: output that meets quality, service-level, security, policy, and mission-priority requirements. This makes the objective harder, but also more economically truthful.

3.2 Observability prerequisites for causal control

Cross-domain correlation is not enough to justify actuation. PTCP must preserve the conditions under which an observed relationship could support a causal decision: synchronized clocks, traceable provenance, explicit missingness, known sampling and aggregation, and a record of which controller owned each actuator. Replay and bounded intervention then test whether the candidate action changes the intended outcome without shifting cost into another domain.

Prerequisite	Required evidence before active control
Time integrity	Clock source, skew budget, event-time uncertainty, and ordering policy
Data provenance	Sensor identity, schema and firmware version, collection path, and signature
Missingness	Drop rate, sampling gaps, stale-state age, and confidence adjustment
Control ownership	Current controller, action priority, conflict rule, dwell time, and rollback owner
Causal test	Counterfactual baseline, replay or intervention design, outcome, and non-regression measures

4. PTCP Capability Architecture and Disclosure Boundary

PTCP combines proprietary cross-domain analytics, governed decision support, bounded enforcement, and evidence capture without exposing its internal algorithms.

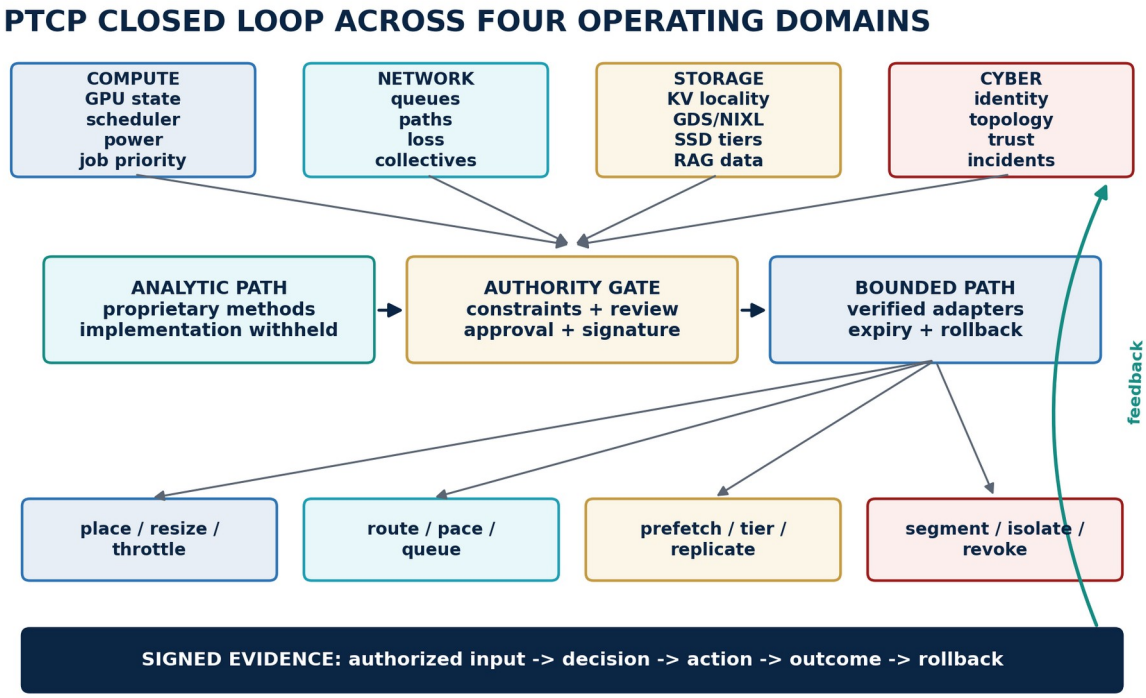

Figure 3. Public capability view of the PTCP reference architecture.

Non-enabling architecture view; internal algorithms, data structures, parameters, and control logic are withheld.

4.1 Cross-domain operational awareness

PTCP is intended to correlate authorized operational signals from compute, network, storage, identity, security, power, and service-management domains into a time-aware operational view. That view supports detection of cross-domain dependencies and conditions that isolated controllers may not recognize. The implementation is proprietary. This paper does not disclose the internal data model, dimensionalization, feature engineering, representation structure, model-update process, treatment of missing data, training procedure, or decision thresholds.

The public claim is functional and testable: PTCP should provide more useful cross-domain awareness than the customer's current combined observability and control baseline, at acceptable latency, cost, and false-positive burden. Independent evaluation can test that outcome without receiving the proprietary implementation.

4.2 Topology-risk diagnostics

PTCP includes topology-aware diagnostics intended to identify material changes in reachability, trust, workload flow, resource relationships, and infrastructure behavior. These diagnostics are used to prioritize investigation and support bounded response; they are not presented as a universal physical

constant or a substitute for identity, endpoint, content, or application security. The score composition, transformations, weighting, calibration, windows, baselines, uncertainty treatment, and classification logic are Tensor Networks trade secrets and are not disclosed.

4.3 Risk-aware decision support

PTCP is designed to evaluate alternatives against performance, resilience, cost, safety, policy, and mission constraints while giving explicit attention to adverse-tail outcomes. The decision layer operates only within authority established by the system owner and relevant domain authorities. This paper does not disclose the proprietary objective function, candidate-generation process, ranking logic, solver, penalties, hyperparameters, confidence rules, escalation thresholds, or control-update law.

4.4 Two timescales and authority separation

A deliberative path performs analysis, replay, review, and authorization. A separate bounded path applies only approved intent through verified target interfaces. The enforcement layer does not invent a new global objective. It must reject, degrade, or revert when authority, integrity, freshness, compatibility, or required evidence is absent. Exact message formats, policy schemas, state machines, verification logic, and target-specific mappings are intentionally withheld.

TRADE-SECRET BOUNDARY Names of public scientific methods may be cited as background, but this paper does not state which methods PTCP uses in a given release, how they are combined, how models are trained or updated, how scores are calculated, how alternatives are generated or ranked, or how target-specific policy is compiled. Those details require controlled diligence under appropriate confidentiality.

5. Compute Optimization

PTCP improves compute economics indirectly and directly: it reduces waiting, aligns placement with coupled resource state, and protects high-value work from cross-layer interference.

Peak accelerator throughput is not realized throughput. Even within a rack-scale NVLink domain, useful output depends on model parallelism, memory pressure, collective timing, worker health, thermal state, batch composition, and phase-specific behavior. Across racks, the scheduling problem expands to network paths, storage locality, and tenant policy. Heterogeneous fleets add different memory capacities, numerical formats, software stacks, and failure modes. The controller must reason about the resource bundle required by a stage, not a GPU in isolation.

PTCP compute mechanism	Decision surface	Expected benefit	Required comparison
Criticality-aware placement	scheduler or runtime adapter	Lower waiting and fewer straggler-amplified bubbles	Incumbent scheduler with network-aware placement
Prefill/decode coordination	Dynamo/vLLM/SGLang planner interface	Better resource shape and queue balance	Native runtime planner and aggregated serving
Risk-aware admission	queue and SLA policy	Protect p99 and mission-priority work during overload	Static priority, token bucket, and autoscaling
Failure-domain avoidance	node/rack/path health context	Fewer retries and cascading stalls	Health-aware scheduler alone
Power/thermal response	clock, placement, or degradation policy	More accepted work under facility constraints	DCGM/thermal controller plus scheduler

5.1 Accepted output per accelerator-dollar

A useful primary metric is accepted output per accelerator-dollar-hour, not gross tokens per second. For training, accepted output can mean validated progress per unit time at a required loss and reliability level. For inference, it can mean quality-qualified tokens or completed requests meeting latency and safety SLOs. The commercial test is straightforward even without disclosing an algorithm: independently measured improvement across compute, network, storage, resilience, and risk must exceed the full incremental cost of licensing, integration, operations, and change management under an agreed counterfactual.

5.2 Why PTCP should remain accelerator-neutral

NVIDIA, AMD, and custom AI accelerators expose different telemetry and control surfaces. A strategic PTCP layer should normalize semantics without erasing vendor-specific detail. Its core contract should describe workload state, capacity, health, memory, path dependency, and allowable action, with adapters translating to DCGM, ROCm, Kubernetes, Slurm, or a proprietary runtime. Portability is part of the top-ten case because buyers increasingly need leverage across architectures; it is also an evidence risk because abstraction can hide material hardware differences. Kubernetes Dynamic Resource Allocation provides a credible slow-path interface for accelerator and attached-device intent, including prioritized requests, partitions, and health-related controls. PTCP should emit constrained resource intent through such ownership models rather than bypass them. [32]

COMPUTE NON-CLAIM PTCP does not create accelerator FLOPS, replace CUDA/ROCm, or make a poorly parallelized model efficient. It can improve the probability that installed compute is fed, scheduled, protected, and used for accepted work.

6. Network Optimization

The network is the synchronization substrate of distributed AI; PTCP's opportunity is to add workload, topology, storage, and cyber context to mature fabric controls.

AI networks are becoming specialized because collective communication and KV movement are sensitive to bandwidth, tail latency, path asymmetry, congestion, and loss. NVIDIA Spectrum-X advertises topology-aware adaptive routing and telemetry-based congestion control. UEC 1.0 and its current 1.0.3 publication define an interoperable Ethernet-based stack for AI and HPC. These developments reduce the space left for a naive overlay: PTCP must integrate with, and outperform the combined outcome of, advanced native controls rather than compare itself with shortest-path routing alone. NCCL already performs topology-aware collectives across PCIe, NVLink, InfiniBand, and IP sockets; PTCP's defensible scope is higher-order cross-job and cross-domain context, not replacement of collective algorithms. [8]-[10], [15], [35]

Network problem	Evidence category	Bounded action	Primary KPI
Incipient collective congestion	authorized workload and fabric evidence	pace, mark, or select approved alternate path	collective completion, p99 queue, loss
Elephant/mice interference	authorized flow, priority, and service evidence	queue intent or rate envelope	protected-flow tail without starvation
Path or device anomaly	proprietary topology-risk and device-health evidence	avoid, drain, or quarantine limited scope	recovery time and false reroute cost
KV transfer critical path	authorized cache, transfer, and service evidence	co-select worker, path, and transfer tier	TTFT and decode idle time
Control instability	authorized controller-health and timing evidence	freeze, damp, expire, or revert	oscillation, route churn, rollback RTO

6.1 Topology is both a performance and security object

A communication graph encodes which ranks, services, storage endpoints, agents, tools, and identities can reach one another. Performance changes when edge capacity or latency changes; security changes when an edge appears, disappears, or violates authorized reachability. PTCP can therefore use one graph representation for two distinct questions: which topology minimizes risk-sensitive completion cost, and which topology remains permitted. The objectives must not be collapsed. Security constraints define the feasible set; performance optimization selects within it.

6.2 Where PTCP could add unique value

Native congestion control reacts quickly to local signals. PTCP's potential advantage is slower, higher-order context: scheduled stage transitions, KV placement, maintenance windows, identity trust, cross-fabric state, and counterfactual cost. The fast path should delegate microsecond response to the fabric and inject only stable, validated intent. If PTCP attempts to micromanage every packet from a remote controller, it will likely lose on latency and stability.

7. Storage and Data-Movement Optimization

In AI inference, storage is no longer a background system: model weights, retrieval context, checkpoints, and KV state form an active hierarchy from HBM to remote object storage.

NVIDIA GPUDirect Storage provides a direct DMA path between storage and GPU memory, avoiding a CPU bounce buffer and potentially reducing CPU load, latency, and bandwidth bottlenecks. NIXL abstracts transfers across CPU memory, GPU memory, file, block, and object storage. Dynamo's KVBM presents a write-through hierarchy spanning GPU, host, SSD, and remote storage. These technologies make data movement faster and more flexible, but they also create a placement and lifecycle control problem: what state should reside where, for how long, on which path, for which tenant, at what write cost, and under which security constraints? NIXL's documented telemetry includes transfer bytes, request counts, transfer time, errors, and dropped-event counts; dropped telemetry should lower PTCP confidence rather than disappear inside an average. [12]-[14], [16], [37]

Mooncake, the KV-cache-centric serving platform used by Kimi, is a particularly important benchmark architecture. The FAST 2025 Best Paper reports joint use of underused CPU, DRAM, SSD, and NIC resources across thousands of nodes and more than 100 billion tokens daily. The authors report 59-498 percent higher effective request capacity than selected baselines in real-trace tests and more requests than their earlier internal systems on A800 and H800 clusters. These are author-reported Mooncake results, not PTCP evidence. They demonstrate that cross-resource cache scheduling is already a competitive systems capability; PTCP must add portability, security-aware control, and independent evidence beyond that bar. [36]

Storage objective	Evidence category	PTCP decision	Failure mode to test
KV reuse	authorized workload, memory, and tenant-policy evidence	retain, evict, replicate, or prefetch	churn makes offload slower than recompute
GDS path selection	authorized device, path, locality, and compatibility evidence	choose approved direct path or safe fallback	silent bounce path or unsupported topology
Checkpoint flow	authorized job, transfer, and recovery-policy evidence	schedule, pace, tier, or replicate	checkpoint burst harms collectives
RAG data locality	authorized demand, placement, and service-level evidence	co-place worker, index shard, and path	hot shard creates correlated tail
SSD endurance	authorized device-health and lifecycle evidence	filter disk offload and adjust tier budget	write amplification shortens device life

7.1 Storage optimization is a control problem, not a cache-size problem

A larger cache can reduce misses while increasing capital, power, privacy exposure, and eviction complexity. A lower-latency path can be undesirable if it crosses an untrusted boundary. Keeping all KV state in HBM improves access but competes with model and batch memory. PTCP should treat storage state as part of the risk-sensitive resource allocation problem and make the trade visible. Every placement action should be reversible or reconstructible, tenant-scoped, and evidenced.

7.2 Data semantics and privacy boundaries

PTCP's strongest deployment posture is payload-blind by default: use metadata, counters, locality identifiers, hashes, lifetimes, sensitivity labels, and policy without ingesting model prompts or customer content. Some optimization may benefit from semantic signals, but this expands privacy, security, and regulatory scope. Semantic access should be a separately authorized module, never an implicit prerequisite of the core control plane.

STORAGE NON-CLAIM PTCP does not replace a distributed file system, object store, vector database, GDS, NIXL, or KVBM. Its claimed role is to coordinate their placement and transfer decisions with compute, network, and cyber state.

8. Cybersecurity Optimization

AI infrastructure requires security controls that understand dynamic topology, autonomous identities, tool use, and resource exhaustion without turning every anomaly into a data-center outage.

The April 2026 multinational guidance on agentic AI warns that autonomy, interconnected components, increased complexity, and evolving capability expand attack surface and risk. It recommends incremental deployment, continuous assessment, explicit accountability, rigorous monitoring, human oversight, and established cybersecurity paradigms. NIST CSF 2.0 adds Govern to the core functions; the AI RMF organizes Govern, Map, Measure, and Manage. PTCP's safe envelope, evidence pack, shadow mode, bounded action, and rollback are compatible with those governance outcomes, but compatibility is not certification. [23]-[27]

Zero trust also prevents a topology-only interpretation of security. NIST SP 800-207A removes implicit trust based on network location and emphasizes application and service identity with granular runtime enforcement. NSA-led AI data-security guidance recommends signatures, provenance, trusted infrastructure, and lifecycle protection. PTCP should therefore bind every topology decision to authenticated identity and signed data lineage; graph anomaly cannot substitute for semantic poisoning, data-loss prevention, or endpoint defense. [38], [39]

The same multinational agentic-AI guidance states that localized component analysis and logging can be insufficient for risks emerging from interactions among models, humans, datasets, tools, hardware, and guardrails, and it points to system-theoretic methods such as STPA, STPA-Sec, and CAST. That is unusually direct independent support for the problem class PTCP addresses: system-level state,

bounded action, and cross-domain evidence. It is not evidence that PTCP's implementation detects or contains those risks better than alternatives. [23]

Threat or failure	Evidence category	Possible bounded response	Control safeguard
Rogue or over-privileged agent	authorized identity, tool-use, and reachability evidence	revoke capability, isolate session, preserve evidence	per-action identity; human escalation
Resource-exhaustion attack	authorized resource, service, and identity evidence	rate limit or quarantine limited scope	mission-priority floor; time-limited scope
Lateral movement	proprietary topology-risk and access-control evidence	segment path or redirect to inspection	do not rely on payload blindness alone
Compromised control adapter	authorized integrity and compatibility evidence	reject policy and enter safe mode	hardware root of trust and key rotation
Telemetry poisoning	authorized sensor-integrity and provenance evidence	down-weight, freeze, or request attestation	independent sources and replay audit

8.1 Security-performance convergence

Conventional security tools often decide whether an event is suspicious; infrastructure controllers decide whether a path or resource is efficient. PTCP can connect the decisions without merging accountability. For example, a suspected compromised agent may be isolated by cutting the smallest permitted set of communication edges while preserving critical inference. The security authority determines what must be blocked; the performance optimizer finds the least disruptive compliant action; the evidence plane records the basis and outcome.

8.2 Payload-blind is risk-reducing, not magic

Topology and behavior can expose attacks that content inspection misses, and they can support encrypted or privacy-sensitive environments. They cannot determine every malicious semantic intent. PTCP should overlay EDR/XDR, SIEM/SOAR, zero trust, microsegmentation, identity, secrets management, and model-specific safeguards. Claims of universal zero-day detection would be scientifically indefensible; the testable claim is improved detection or containment for specified topology and behavior classes at an acceptable false-positive cost.

9. Cross-Domain Operating Scenarios

PTCP's strategic value appears when one decision requires facts and actuators from more than one domain.

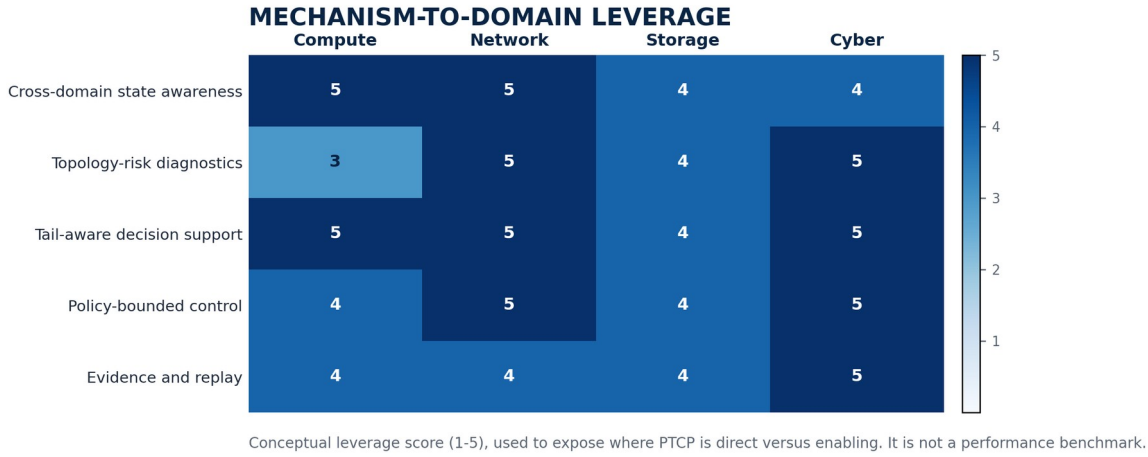

Figure 4. PTCP capability areas have different leverage in each domain.

Conceptual author scoring, not a benchmark. Scores expose test priorities.

9.1 Scenario A: retrieval-heavy inference surge

A burst of long-context requests raises prefill demand, creates hot retrieval shards, and increases KV transfer. The native runtime can add prefill workers, but the best placement depends on GPU memory, index locality, RDMA path state, storage tier, and tenant priority. PTCP uses proprietary cross-domain analytics to recommend a bounded configuration through existing runtime and policy interfaces. Its internal representation, candidate generation, and ranking logic remain withheld. Success is measured by TTFT and accepted throughput without SSD churn, network regression, or privacy-boundary violation.

9.2 Scenario B: collective straggler with ambiguous cause

One training stage develops periodic bubbles. GPU telemetry suggests underutilization, switch counters show moderate queueing, and the scheduler sees no node failure. PTCP correlates authorized multi-domain evidence and tests proprietary recommendations in replay before any bounded action. Exact correlation, diagnosis, and prioritization logic are withheld. The falsifier is simple: if a single-domain health rule finds and fixes the issue equally well, the higher-order controller did not add sufficient value.

9.3 Scenario C: suspected agent compromise during overload

An agent begins calling an unusual tool while the cluster is under heavy load. A blunt quarantine could interrupt a critical workflow; a performance-only policy might preserve the malicious path. Security policy marks prohibited edges and required evidence. PTCP proposes a policy-compliant bounded response intended to preserve unaffected work, limits its duration, and escalates for human review. Exact selection and optimization logic are withheld. Evaluation must count both containment effectiveness and business disruption.

9.4 Scenario D: facility power constraint

A grid or thermal event reduces the facility power envelope. PTCP receives the constraint as a hard safe-set update, then reevaluates workload admission, placement, network paths, and storage transfers. It may defer low-priority checkpoint replication, shift inference to more efficient nodes, protect latency-sensitive tenants, and preserve cyber monitoring. The goal is maximum accepted work inside the facility constraint, not maximum utilization of every component.

Scenario	Domains jointly required	Primary benefit	Key non-regression test
RAG surge	compute + network + storage + policy	TTFT and throughput	privacy, SSD endurance, protected-flow tail
Training straggler	compute + network + storage	job completion and accelerator efficiency	fabric stability and scheduler fairness
Agent compromise	cyber + network + compute	containment with lower disruption	no unauthorized cut or evidence loss
Power constraint	compute + network + storage + cyber	accepted work under cap	critical monitoring and mission floor

10. The Top-Ten Qualification Scorecard

A disclosed rubric is more credible than an unsupported superlative. PTCP is scored on strategic importance, systems leverage, deployability, governance, and evidence maturity.

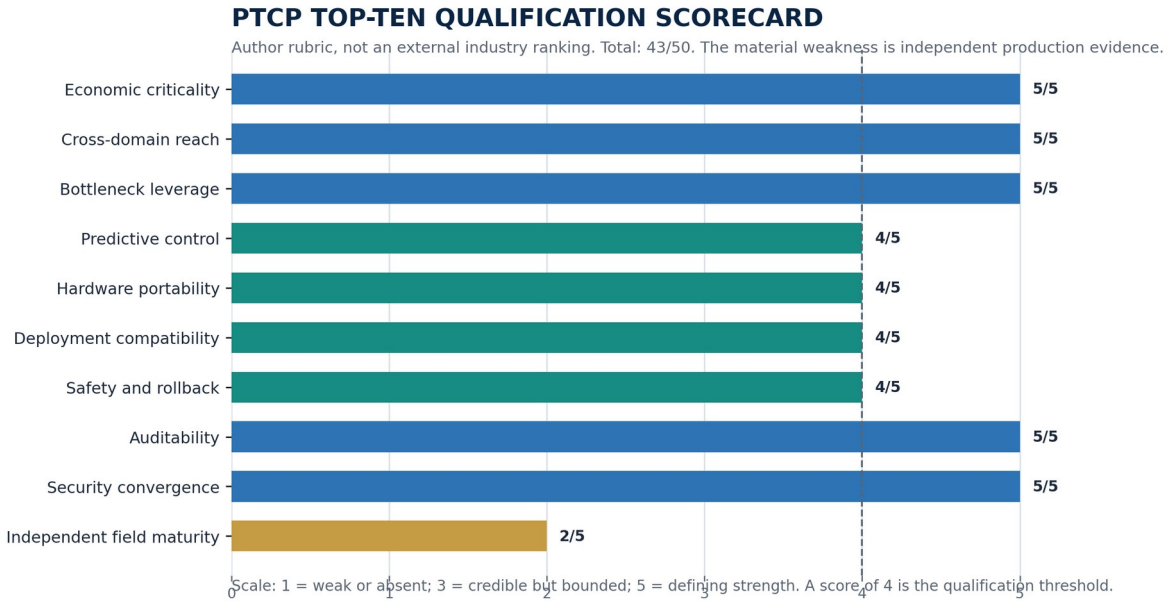

Figure 5. PTCP qualification scorecard: 43 of 50.

Author judgment. The score is reproducible only to the extent readers accept the disclosed criteria and evidence.

Criterion	Score	Reason	What would change the score
Economic criticality	5	Acts on scarce accelerator, fabric, storage, energy, and risk budgets	Demonstrate that buyers do not value or fund the outcome

10.1 Interpretation

The 43-point total exceeds the paper's 40-point qualification threshold, and nine criteria are at least 4. The result says PTCP has sufficient structural importance and design leverage to be treated as a top-ten strategic technology class. It does not erase the low maturity score. A product can be strategically top-ten and still require disciplined validation before broad authority. In fact, the gap between structural importance and field maturity is the central management fact.

GOVERNANCE CONSEQUENCE Give PTCP top-tier evaluation priority, not top-tier unbounded authority. Strategic importance determines the rigor and speed of the test; field evidence determines the scope of deployment.

11. Complementarity and Competitive Position

PTCP wins only if it makes strong incumbent technologies more effective together; a strategy based on pretending those technologies do not exist will fail.

Existing class or product	What it already does well	PTCP complement	Overlap risk
Accelerator management	device health, clocks, partitions, utilization	mission, network, storage, and cyber context	vendor platform expands upward
Scheduler/runtime	worker placement, scaling, request routing, KV awareness	tail risk, topology, policy, and cross-runtime evidence	native planner absorbs features
AI fabric control	congestion control, adaptive routing, telemetry	workload value, storage locality, identity constraints	fabric has enough global context
Storage/KV manager	tiering, offload, transfer, reuse	co-optimization with compute path and security	cache optimizer achieves same result
SIEM/SOAR/XDR	alerting, investigation, endpoint and workflow response	topology-aware policy-bounded infrastructure response	security platform adds infrastructure control
Observability platform	collection, correlation, dashboards, alerts	counterfactual decision and bounded actuation	AIOps adds safe closed-loop control

11.1 The defensible wedge

The strongest entry is not a claim to replace every controller. It is a high-value coordination problem with measurable cross-domain residual loss: for example, KV-aware inference across a congested multi-tier environment, or topology-native containment where blunt segmentation would cause excessive operational harm. PTCP begins in observation and recommendation, exports evidence into the customer's existing workflow, and earns authority one action class at a time.

11.2 Moat composition

- PoL and physical-AI issued patents provide provenance and potentially useful claim scope, subject to counsel diligence.
- A normalized joint-state schema and time-synchronized evidence corpus become difficult to recreate once deployed.
- Target adapters, compatibility tests, and rollback transcripts create deployment know-how.
- Customer-accepted baselines and causal evidence are more defensible than unpublished benchmark claims.
- Safety certification and operator trust can compound across sites if the product remains disciplined.

Patents do not prove performance, exclusivity, freedom to operate, or coverage of every later PTCP mechanism. The moat thesis should be supported by a counsel-reviewed claim map, license-scope verification, protected implementation IP, and customer evidence - not by the existence of an issued patent alone. [1], [5]

12. Public Reference Architecture and Trust Boundaries

A production PTCP should be modular, target-aware, partition-safe, and capable of proving every consequential decision.

Plane	Core services	Trust boundary	Failure behavior
Collection	authorized telemetry ingress and provenance services	authenticated sources, integrity, clock and schema controls	reject or quarantine invalid input; never invent freshness
Analytics	proprietary cross-domain state and risk services; implementation withheld	tenant separation, least privilege, encrypted state	degrade to observation or an approved safe posture
Decision	proprietary recommendation and bounded-decision services; implementation withheld	independent policy and authority verification	reject unauthorized or insufficiently supported output
Governance	review, authorization, signature, scope, compatibility, and expiry	separate approval and deployment duties	deny unsigned, stale, incompatible, or out-of-scope intent
Enforcement	verified interfaces to approved infrastructure controllers	local capability limits and target attestation	local safe mode and bounded rollback
Evidence	authorized input reference, decision, approval, action, outcome, and recovery record	tamper-evident record and access control	withhold high-impact action if evidence is unavailable

12.1 Public control boundary

Every consequential action must be bound to an authenticated target, authorized scope, compatible implementation, limited lifetime, accountable approver, tested recovery path, and tamper-evident record. The enforcement layer must reject intent that fails those governance conditions. Exact schemas, tokens, field definitions, serialization, signing workflow, compatibility rules, and verification state machines are proprietary and are not disclosed in this public edition.

12.2 Observability contract

OpenTelemetry provides vendor-neutral traces, metrics, logs, and context propagation, while GPU, NIC, DPU, switch, and storage stacks expose specialized counters. PTCP should preserve raw provenance and normalize only into versioned semantic fields. Clock uncertainty, sampling period, missingness, aggregation, and sensor authority must be explicit. Without those properties, a sophisticated model can make a precisely wrong decision. [17], [18]

12.3 Hardware and line-rate boundary

Tensor-supplied v14.2 materials describe a functioning control-plane lineage, APIs, session and identity support, fast-path fleet interfaces, and deterministic userspace tests. They do not publicly establish live NIC/DPU line-rate performance across all target variants. Production qualification therefore requires per-target compilation, verifier output, throughput and latency measurement, reboot and upgrade behavior, failover, partition, and rollback tests.

13. Evidence Maturity and Validation Status

PTCP's strongest evidence is cumulative: legitimate scientific components, issued-IP provenance, implemented software, and an explicit testing architecture. Its weakest evidence is independent production efficacy.

Evidence level	Available support	Strength	Limitation
Scientific	Public mathematical, statistical, graph, and programmable-infrastructure foundations cited in the references	Foundations are independently established	PTCP's proprietary combination and operational tuning are not thereby proven
Patent/public record	US 11,308,384 B1 PoL patent; US 9,696,404 B1 physical tracking patent; recorded Tensor notices	Issued provenance and described empirical examples	Counsel must verify scope, license, exclusivity, and FTO
Implementation	Tensor code lineages, APIs, manuals, fast/slow-path design, safe-envelope and proof-pack tests	System exists beyond a concept paper	Environment-specific compatibility and scale remain open
Internal quantitative	v14.2 simulation reports 13-point MFU uplift, 37.8% RAG latency reduction, 65.9% tail-risk reduction	Useful engineering priors and reproducible hypotheses	Vendor-controlled simulation; not a customer field claim
Independent production	No public multi-site causal evidence identified	Clear, testable next milestone	Cannot claim general superiority or realized savings

13.1 What validated means here

A mathematical component is validated when its properties and implementation can be reproduced. A software service is validated when it builds, passes deterministic tests, enforces interfaces, and fails as specified. A target integration is validated when compatibility, latency, capacity, safety, and rollback are attested on that target. A performance claim is validated only when a pre-registered comparison demonstrates causal benefit inside a stated envelope. A business claim is validated only when customers deploy, pay, realize value, and renew. PTCP currently occupies different rungs for different claims.

13.2 Internal results as priors, not conclusions

Internal simulation can narrow the experimental space, expose instrument errors, and establish plausible effect sizes. It can also encode favorable traces, baselines, assumptions, or economic weights. The cited v14.2 results should therefore determine what to test, not what to market. Independent evaluators should receive the frozen configuration, workload corpus, baseline implementations, analysis plan, and all negative results.

EVIDENCE RULE No result may be generalized beyond the tested workload, topology, hardware, software version, traffic mix, failure model, and control authority. Every external claim should name its evidence envelope.

14. Risks, Objections, and Falsifiers

The top-ten thesis is useful only if it remains falsifiable. PTCP should lose authority or scope when simpler or native controls perform equally well.

Objection	Why it matters	Test	Decision rule
Native platforms already solve the problem	PTCP may add cost and instability	Compare against combined scheduler + fabric + cache + security baseline	Narrow or stop if no incremental value
Proprietary methods add unnecessary complexity	Complexity may exceed benefit	Compare against simpler combined baselines without access to proprietary internals	Use the simplest system meeting outcome and safety requirements
Topology diagnostic is fragile	A proprietary diagnostic may vary across environments	Test repeatability, drift, and robustness across representative conditions	Reject if prioritization is unstable
Cross-domain controller is too slow	Decision arrives after the event	Measure detection-to-action against event timescale	Delegate local response or remove use case
Controller creates blast radius	One policy can harm many workloads	Fault injection, malicious update, partition, stale clock, rollback	No active scale until all safety gates pass
False positives overwhelm operators	Cyber value can be erased by disruption	Incident corpus with business-impact-weighted precision/recall	Fail if expected disruption exceeds avoided loss
Integration destroys economics	Adapters become bespoke services	Track engineering hours, time-to-value, upgrade burden	Stop target without repeatable connector profile
Patent moat is overstated	Commercial defense may be weak	Independent claim chart, license and FTO diligence	Market on evidence, not unverified exclusivity

14.1 Control-theoretic risk

Cross-domain feedback can oscillate when sensors are delayed, multiple controllers act on the same resource, or policy changes faster than the system settles. PTCP must identify controller ownership, timescale, gain, and conflict for each actuator. Appropriate stability safeguards are mandatory; their exact design and tuning are withheld. The safe default is advisory mode until stability is demonstrated under representative delay and failure.

14.2 Organizational risk

Compute, network, storage, and cyber teams often have separate budgets, SLOs, data models, and change boards. A technically correct optimizer can fail because no owner authorizes cross-domain action. The deployment must specify decision rights: who sets mission weights, who defines prohibited edges, who approves canaries, who owns rollback, and who accepts the evidence. Governance integration is part of the product, not a post-sale detail.

KILL CRITERIA Do not expand active control if PTCP fails the pre-registered primary benefit, violates any safety guardrail, increases total operating burden, or cannot reproduce the result under an independent evaluator.

15. 120-Day Independent Proof-of-Value

The largest uncertainty - incremental production value over strong baselines - can be converted into evidence through a staged, reversible, pre-registered program.

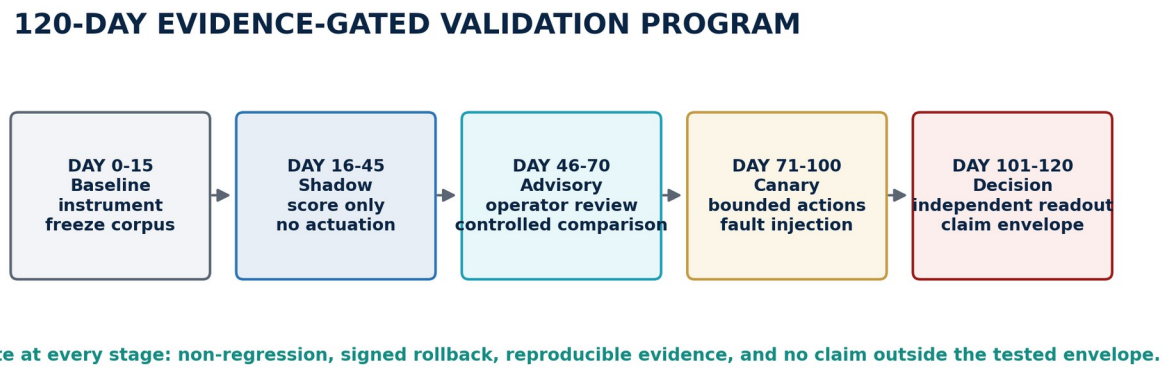

Figure 6. Evidence-gated progression from observation to bounded authority.

Author validation design.

1. Name one primary outcome, strong combined baselines, minimum effect, guardrails, workload strata, exclusions, and stop rules.
2. Place all implementation artifacts, configurations, test fixtures, and analysis materials under independent custody; the public report exposes versions and hashes, not proprietary source or parameters.
3. Synchronize compute, fabric, storage, identity, power, and SLO telemetry with known uncertainty.
4. Run representative steady state, bursts, failures, maintenance, and security incidents without PTCP.

5. Generate proprietary recommendations without changing production; measure timeliness, consistency, and operator usefulness.
6. Compare the complete PTCP service against strong combined baselines without disclosing proprietary internals.
7. Have operators accept, reject, and annotate recommendations; measure decision utility and burden.
8. Grant one low-blast-radius action class with signature, magnitude limit, expiry, and tested rollback.
9. Exercise stale telemetry, partition, target mismatch, malicious policy, overload, sensor poisoning, and rollback.
10. Report positive and negative results, uncertainty, tested versions, and permitted external language.

15.1 Minimum scorecard

Domain	Primary outcome	Guardrail	Evidence artifact
Compute	accepted output/$, MFU, job or request completion	no starvation, priority inversion, or quality loss	trace-to-job causal report
Network	collective time, p95/p99 queue, retransmit/loss	no instability, unsafe reroute, or unfairness	path and policy replay
Storage	TTFT, cache hit, stall time, bytes moved, SSD writes	no corruption, privacy breach, or churn regression	object-tier and transfer ledger
Cyber	lead time, recall, false-positive cost, containment time	no unauthorized isolation or evidence gap	incident timeline and signed action
Safety	policy rejection, expiry, rollback RTO, partition behavior	100% rejection of unsigned/stale/mismatched policy	fault-injection transcript
Economics	realized benefit minus full incremental cost	positive conservative base case	customer-accepted value model

15.2 Statistical discipline

Randomize or alternate treatment windows where operationally safe; stratify by workload and load regime; report confidence intervals and distribution tails; correct for repeated experiments; preserve raw event order; and separate engineering tuning from the confirmatory evaluation. For rare incidents, supplement live testing with replay and fault injection, but do not label simulated incident results as field prevention. The evaluator should be independent of the feature-development team.

16. Strategic Recommendations

Treat PTCP as a strategic systems technology whose authority must be earned through evidence, one target and action class at a time.

1. Use 'cross-domain predictive control and evidence plane' and disclose that top-ten is the paper's technology-class thesis.
2. Begin where compute, network, storage, and cyber context creates measurable residual loss that native controls do not already eliminate.
3. Deliver observable, reviewable recommendations and evidence before requesting actuation.
4. Publish exact hardware/software versions, telemetry contracts, control limits, latency, failure behavior, and rollback evidence.
5. Compare against modern scheduler, fabric, KV, storage, and security controls operating together.
6. Let domain owners define hard constraints; let PTCP optimize only inside the signed feasible set.
7. Prioritize independent causal proof, customer acceptance, and repeatability over a broad list of unsupported capabilities.
8. Every marketing statement must inherit the tested envelope and evidence level.

16.1 Board and buyer decision

Decision	Minimum evidence	Authority allowed
Fund evaluation	implemented core, security review, frozen claims, named target	lab and offline replay
Enter customer shadow	telemetry contract, privacy review, baseline plan, no-actuation guarantee	observation and recommendation
Run canary	shadow calibration, target attestation, signed rollback, operator approval	one bounded action class
Expand domain/site	causal primary benefit, all guardrails, independent readout, acceptable operations cost	specified tested envelope only
Make broad market claim	repeatable multi-site evidence and reviewed claim language	no broader than aggregate evidence

Conclusion

PTCP belongs in the top-ten class of AI data-center technologies because it addresses the missing coordination problem created by the success of the other nine.

Accelerators, HBM, scale-up links, scale-out fabrics, distributed runtimes, KV and storage systems, programmable infrastructure, observability, and AI-native security are all necessary. Their interaction

determines accepted output, tail latency, resilience, and risk. A cross-domain controller can create disproportionate value by reducing waiting and propagation rather than by making any one component incrementally faster.

PTCP offers a technically coherent implementation of that class: proprietary cross-domain analytics, topology-aware risk assessment, governed decision support, bounded enforcement, and an intrinsic evidence chain. Its architecture aligns with the industry's movement toward rack-scale compute, open accelerator interconnect, AI-specific Ethernet, distributed inference, tiered KV storage, programmable DPUs, and governed agentic systems.

The conclusion remains deliberately bounded. PTCP has validated foundations and substantial implementation evidence, but no public independent record yet establishes general production superiority. That gap lowers its maturity score; it does not erase the strategic category. The rational course is to give PTCP top-tier evaluation priority, run a 120-day pre-registered proof against strong combined baselines, and grant authority only to configurations and claims that survive.

Glossary

Term	Definition
Accepted output	Work that meets quality, latency, security, policy, and mission requirements, not merely raw tokens or utilization.
AI fabric	The scale-up and scale-out communication substrate carrying collectives, KV state, checkpoints, and storage traffic.
CVaR	Conditional Value-at-Risk, a public risk-analysis concept cited as background; PTCP's internal risk methods are not disclosed.
D_topo	Tensor Networks' name for a proprietary topology-risk diagnostic; its composition, parameters, and thresholds are withheld.
Evidence envelope	The workload, topology, hardware, software, traffic, failure, and authority conditions for which a result is valid.
Evidence pack	A signed, versioned record of state, candidate policy, authorization, action, outcome, and rollback.
Fast path	Constrained local enforcement applying only approved, signed, target-valid, time-limited policy.
GDS	GPUDirect Storage; a direct DMA path between storage and GPU memory.
HBM	High-bandwidth memory placed close to accelerators.
KV cache	Key-value attention state reused during transformer inference.
KVBM	Dynamo KV Block Manager; a multi-tier KV cache and memory layer.
MFU	Model FLOPs Utilization; realized model computation relative to theoretical accelerator capacity.
NIXL	NVIDIA Inference Xfer Library; a transfer abstraction across heterogeneous memory and storage.
Pattern of Life	A model of normal multidimensional operational behavior.
PoL-TT	Pattern-of-Life Tensor Train, a Tensor Networks capability name; its representation, training, update, and scoring methods are withheld.
PTCP	Predictive Tensor Control Plane; a model-agnostic analytic, policy, actuation, and evidence overlay.
Safe envelope	An authorized boundary limiting where and how PTCP may act; exact policy representation and enforcement logic are withheld.
Shadow mode	Observation and decision generation without active enforcement.
Slow path	A deliberative analytic and governance layer; its internal models, objectives, and decision logic are withheld.
TEVV	Testing, evaluation, verification, and validation.
Top-ten-class	A strategic technology-class judgment under the disclosed rubric, not an independent vendor ranking.
TTFT	Time to first token.
UEC	Ultra Ethernet Consortium and its AI/HPC Ethernet communication specification.
UALink	An open accelerator scale-up interconnect standard.

References

[1] Tensor Networks, Inc., PTCP public platform overview, accessed July 25, 2026. Official source

[2] J. Ochoa, 'Predictive Tensor Control Plane: Holographic Geodesic Routing and Topology-Native Security in Information Networks,' Tensor Networks, 2026. Official source

[3] I. V. Oseledets, 'Tensor-Train Decomposition,' SIAM Journal on Scientific Computing 33(5), 2011. Official source

[4] R. T. Rockafellar and S. Uryasev, 'Optimization of Conditional Value-at-Risk,' Journal of Risk 2(3), 2000. Official source

[5] C.-C. Ni et al., 'Ricci curvature of the Internet topology,' IEEE INFOCOM, 2015. Official source

[6] NVIDIA, GB200 NVL72 rack-scale architecture and specifications, accessed July 25, 2026. Official source

[7] AMD, Instinct MI350X accelerator specifications, launched June 12, 2025. Official source

[8] UALink Consortium, UALink mission and 1.0 scale-up architecture. Official source

[9] UALink Consortium, UALink specifications, including 200G 1.0. Official source

[10] Ultra Ethernet Consortium, specification history; v1.0.3 published July 16, 2026. Official source

[11] NVIDIA, Dynamo disaggregated serving documentation. Official source

[12] NVIDIA, Dynamo user guide: KV transfer as a critical path in disaggregated serving. Official source

[13] NVIDIA, Dynamo KV Block Manager documentation. Official source

[14] NVIDIA, Dynamo KV cache offloading documentation. Official source

[15] NVIDIA, Spectrum-X Ethernet networking platform. Official source

[16] NVIDIA, GPUDirect Storage Overview Guide. Official source

[17] Linux Kernel Documentation, BPF and XDP. Official source

[18] OpenTelemetry, vendor-neutral observability framework and documentation. Official source

[19] Gartner, 'Worldwide IT Spending to Grow 13.5% in 2026, Totaling $6.31 Trillion,' Apr. 22, 2026. Official source

[20] Gartner, 'Worldwide AI Spending to Grow 47% in 2026,' May 19, 2026. Official source

[21] Gartner, 'Worldwide End-User Spending on Information Security to Total $213 Billion in 2025,' July 29, 2025; includes 2026 forecast. Official source

[22] International Energy Agency, Energy and AI, executive summary, 2025. Official source

[23] NSA, CISA, ASD and international partners, 'Careful Adoption of Agentic AI Services,' Apr. 30, 2026. Official source

[24] NIST, Artificial Intelligence Risk Management Framework 1.0, NIST AI 100-1. Official source

[25] NIST, Cybersecurity Framework 2.0, NIST CSWP 29. Official source

[26] NIST, Concept Note: AI RMF Profile on Trustworthy AI in Critical Infrastructure, Apr. 2026. Official source

[27] MITRE, Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS), accessed July 25, 2026. Official source

[28] U.S. Patent No. 11,308,384 B1, 'Method and framework for pattern of life analysis,' Apr. 19, 2022. Official source

[29] U.S. Patent No. 9,696,404 B1, 'Real-time camera tracking system using optical flow feature points,' July 4, 2017. Official source

[30] Tensor Networks, Inc., PTCP AI Factory v14.2 internal proof-pack materials, June 2026; results are internal/simulated pending hardware and customer attestation. Official source

[31] Compute Express Link Consortium, CXL 4.0 architecture overview; specification released Nov. 18, 2025. Official source

[32] Kubernetes, Dynamic Resource Allocation updates in Kubernetes v1.36, May 7, 2026. Official source

[33] Z. Jiang et al., 'MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs,' USENIX NSDI, 2024. Official source

[34] T. Wu et al., 'PIPEMORPH: Mitigating Communication Stragglers in Pipeline-Parallel Training,' USENIX NSDI, 2026. Official source

[35] NVIDIA, NCCL User Guide: topology-aware collective communication, accessed July 25, 2026. Official source

[36] Y. Qin et al., 'Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving,' USENIX FAST, 2025. Official source

[37] NVIDIA, NIXL architecture and telemetry documentation, accessed July 25, 2026. Official source

[38] NIST, SP 800-207A, 'A Zero Trust Architecture Model for Access Control in Cloud-Native Applications in Multi-Cloud Environments,' Sept. 2023. Official source

[39] NSA and partner agencies, 'AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems,' May 22, 2025. Official source

Appendix A. Cross-Domain KPI Dictionary

The proof-of-value requires a shared metric contract so improvements in one domain are not purchased through hidden regression in another.

KPI	Definition	Unit	Direction	Required strata
Accepted output/$	quality- and SLO-qualified work divided by full accelerator cost	work/$	higher	model, tenant, load, hardware
MFU	model FLOPs executed divided by theoretical available FLOPs	%	higher	job stage and accelerator type
Job completion time	submit-to-complete or stage completion	s/min/hr	lower	job class, scale, priority
TTFT	request arrival to first generated token	ms/s	lower	prompt length, model, cache state
Inter-token latency	time between generated tokens	ms	lower	concurrency and decode length
Collective tail	p95/p99 duration of selected collective operation	ms	lower	message size, topology, rank count
Queue tail	p95/p99 occupancy or sojourn time	bytes/ms	lower	link/queue class and load
KV cache hit	requests or blocks served from target cache tier	%	higher	tenant, prefix, tier
Storage stall	accelerator or worker waiting on data movement	ms/request	lower	object/tier/path
SSD write load	bytes or writes attributable to offload and checkpoint policy	TB/day	bounded	device and object class
Detection lead time	time between first observable malicious condition and alert	s/min	lower	attack or failure class
False-positive cost	business and operator impact of incorrect response	$/hours	lower	response class and tenant
Containment time	verified time to stop prohibited behavior or reachability	s/min	lower	incident class and scope
Rollback RTO	time from rollback trigger to verified safe state	s/min	lower	target and action type
Policy churn	control changes per resource per unit time	count/hr	bounded	target and workload phase

A.1 Economic reconciliation

Domain benefits must be reconciled to avoid double counting. Reduced collective time may already appear as higher accepted compute output; lower TTFT may already appear as greater completed-request value. The economic model should map each operational delta to one customer value line,

subtract incremental hardware and operating cost, apply a realization factor, and report conservative, base, and upside cases. Market-spend figures are context, not revenue forecasts.

Appendix B. Acceptance and Fault-Injection Matrix

No high-impact control path should receive production authority until its normal, degraded, adversarial, and recovery behavior is demonstrated.

Test	Injected condition	Expected behavior	Pass evidence
Unsigned policy	valid action body with absent/invalid signature	target rejects with no side effect	adapter log and unchanged state
Expired policy	replay after expiry	target rejects and records reason	time and decision transcript
Version mismatch	policy compiled for different target build	target rejects before action	attestation and error code
Stale telemetry	delay or freeze one sensor family	confidence falls; action freezes or degrades	state uncertainty and no unsafe actuation
Clock skew	offset event sources beyond budget	events quarantined or corrected	synchronization alarm and provenance
Controller partition	disconnect slow path from target	local safe policy continues then expires	availability and expiry transcript
Fast-path crash	kill adapter or DPU process	failover or safe default without loop	recovery time and state integrity
Fabric congestion	incast, elephant interference, or path fault	bounded response improves target KPI	before/after trace and stability
KV path fallback	remove RDMA or force TCP fallback	detect path change and avoid harmful disaggregation	TTFT/throughput and route evidence
Storage churn	small mis-sized lower tier	controller avoids repeated harmful offload	hit/miss, writes, and latency
Malicious update	authorized identity submits out-of-envelope magnitude	verifier rejects and alerts	policy diff and security event
Telemetry poisoning	correlated false counters	cross-source inconsistency detected	sensor trust change and frozen action
Quarantine false positive	benign topology deviation	advisory or minimal reversible response	business impact and rollback
Rollback under load	trigger rollback at peak demand	returns to verified baseline within RTO	state hash, SLO, and operator confirmation
Upgrade	target or adapter rolling upgrade	compatibility gate and no authority gap	version matrix and canary transcript

B.1 Acceptance authority

The customer system owner accepts performance results; the security authority accepts containment and identity controls; the platform owner accepts integration and rollback; and an evaluation lead independent of the PTCP feature team accepts the analysis. No single party should silently waive another domain's hard guardrail.

