TENSOR™ NETWORKS

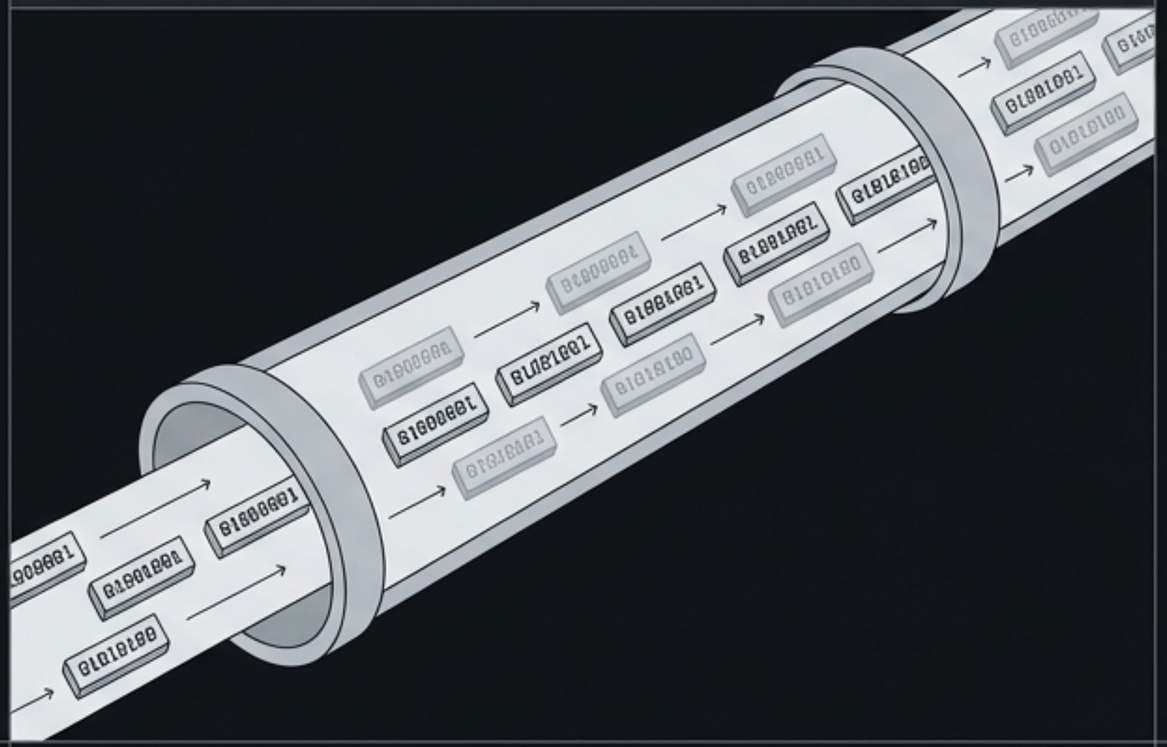
Predictive Determinism for AI Storage Fabrics

Eradicating the Incast Bottleneck with PTCP

The Storage Determinism Mandate

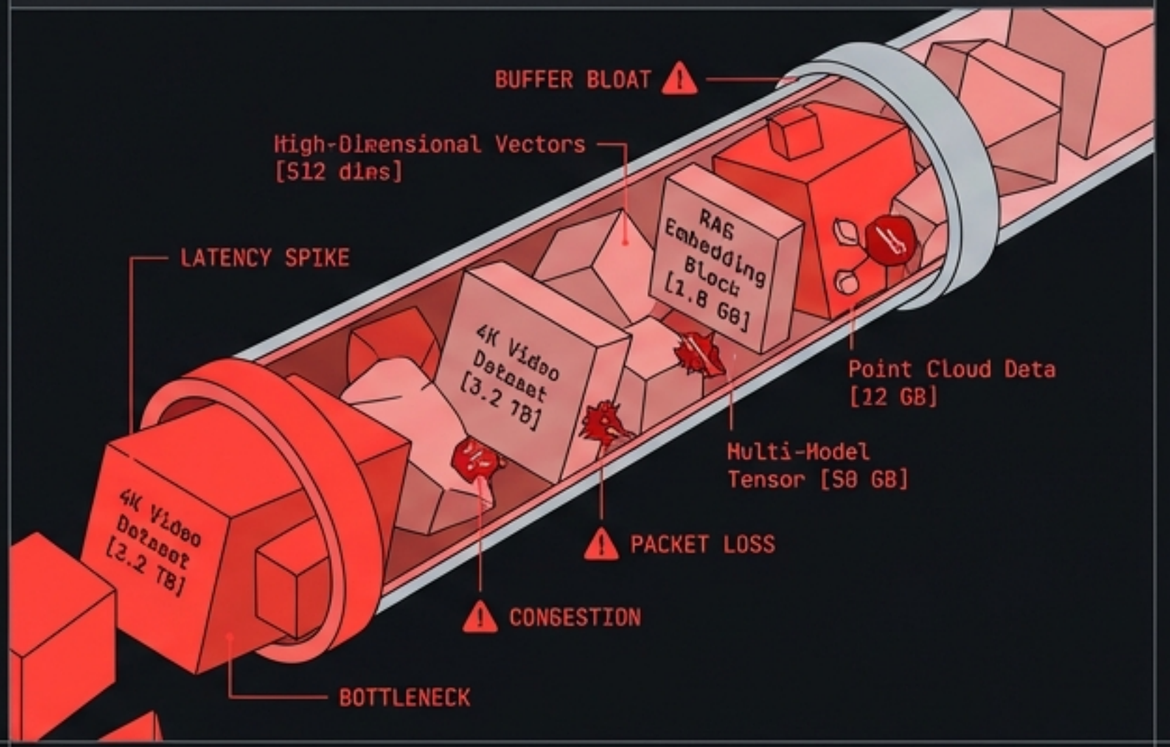
Multimodal & RAG AI: Massive video and high-dimensional vector datasets have shifted the definitive performance bottleneck directly to the storage network. Best-effort is no longer viable.

Legacy AI



Processing text tokens relied on "best-effort" reactive networking.

Multimodal & RAG AI

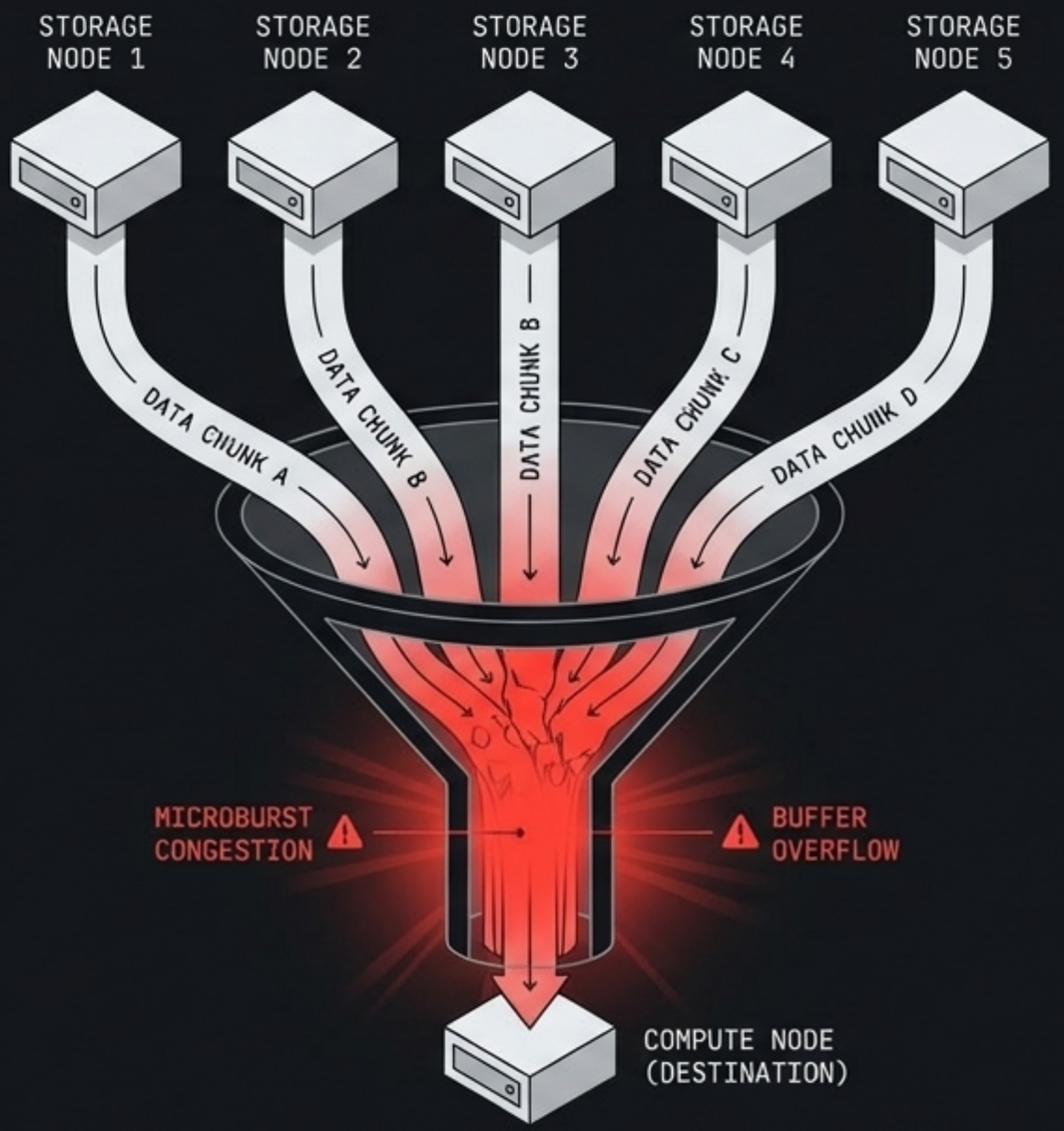


Multimodal & RAG AI: Massive video and high-dimensional vector datasets have shifted the definitive performance bottleneck directly to the storage network. Best-effort is no longer viable.

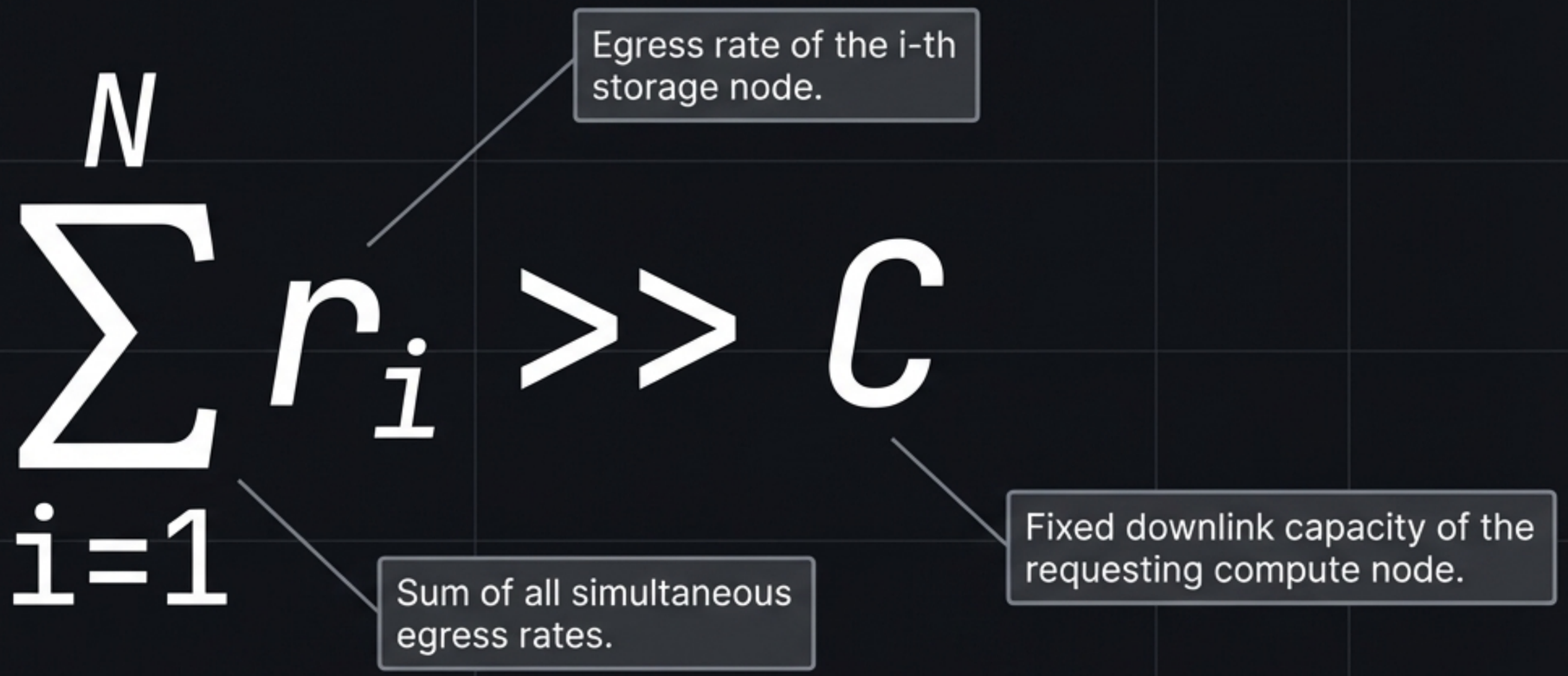
The Definitive Bottleneck: Many-to-One Incast

When a GPU cluster initiates a parallel read—such as a sharded RAG vector search—multiple storage nodes respond simultaneously.

This creates an instantaneous microburst at the Top-of-Rack (ToR) switch ingress, overwhelming the destination link.



The Physics of Storage Congestion



Conclusion: Microbursts are a mathematical certainty in distributed AI storage, fundamentally breaking traditional network protocols.

The Infrastructure Failure Chain

Stochastic collapse migrates from hardware buffers directly into the kernel's transport layer.

1. Shallow ToR Buffers Overflow

2. AIMD "Sawtooth" Collapse or PFC Storms

3. 50ms to 200ms Retransmission Penalty

RETRANSMISSION PENALTY:

50ms - 200ms

PACKET LOSS CONFIRMED
CONNECTION STALLED

The "Lossless" Illusion

TCP/IP

- Drops packets.
- Triggers retransmission timeouts.
- Stalls AI prompt generation.

RoCEv2

- Generates PFC PAUSE frames.
- Triggers Head-of-Line (HoL) blocking.
- Catastrophic "PFC storms" freeze fabric.

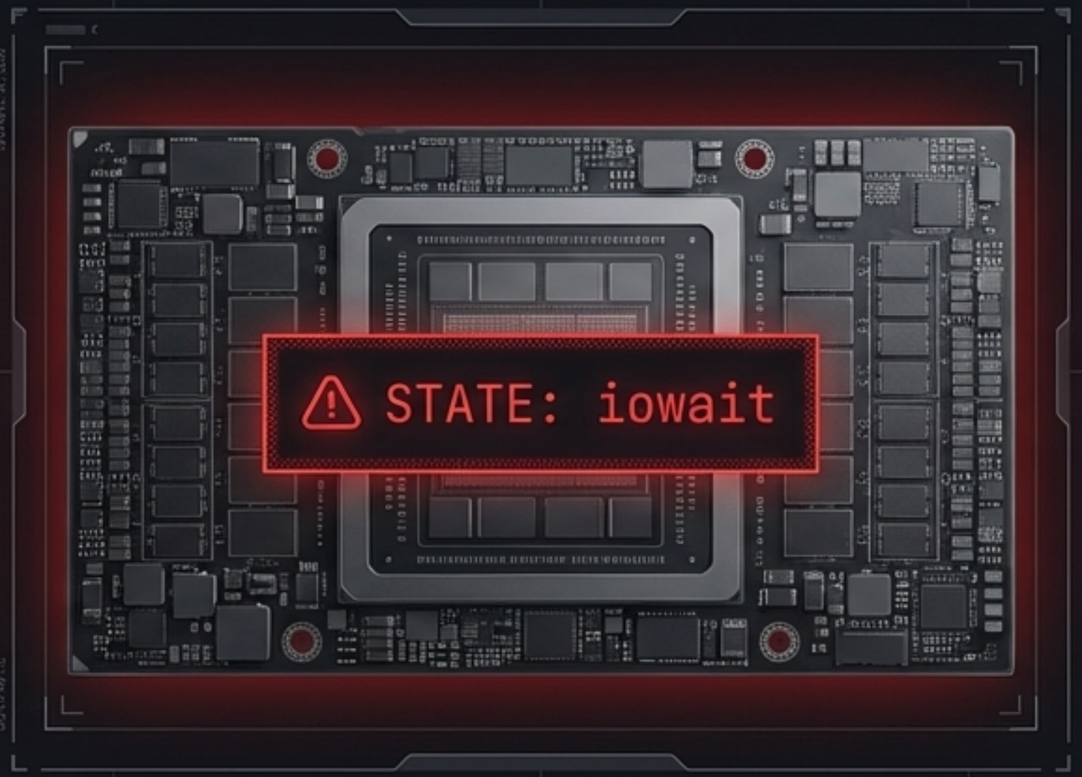
The SR-IOV Paradox

Using Single Root I/O Virtualization (SR-IOV) accelerates storage throughput, which paradoxically exacerbates Incast by hitting ToR switches with microbursts even faster.

The True Cost: Idle Silicon & Lost ROI

200ms Timeouts

Massive spikes in Time To First Token (TTFT) due to TCP retransmissions.

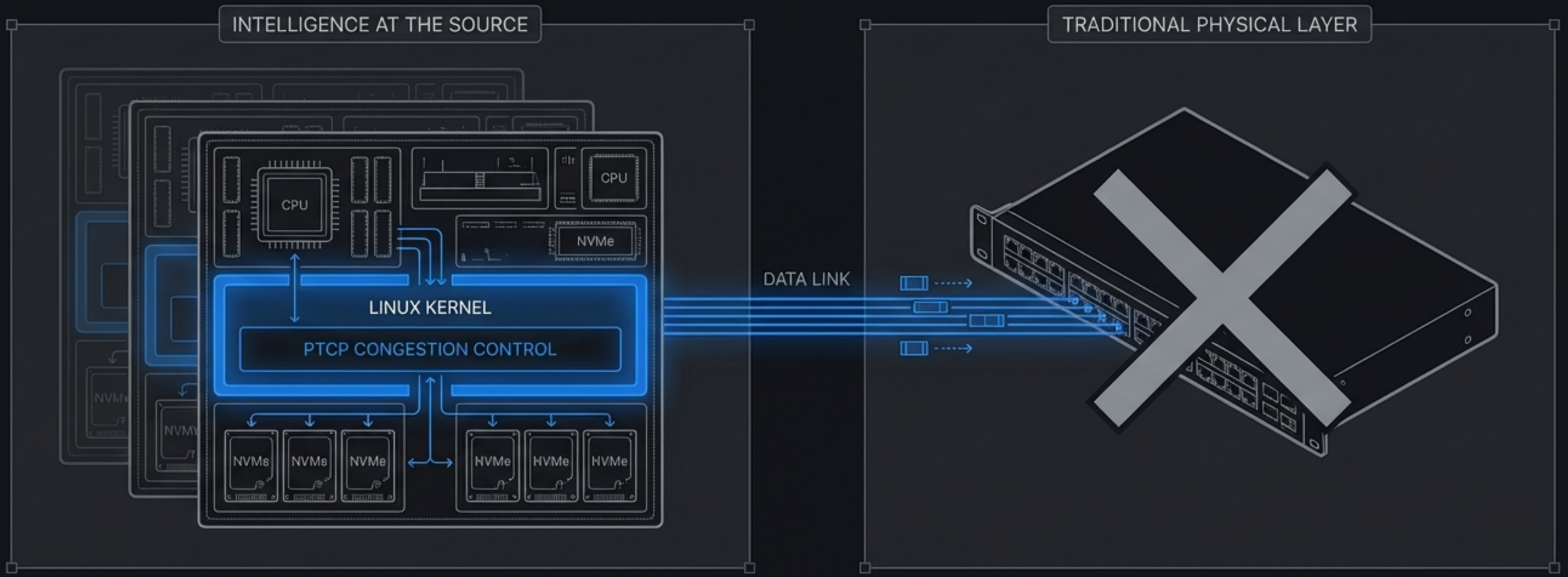


$$ Millions Lost

Idle compute silicon trapped waiting on avoidable "network weather".

Predictive Tensor Control Plane (PTCP)

Moving intelligence to the source.



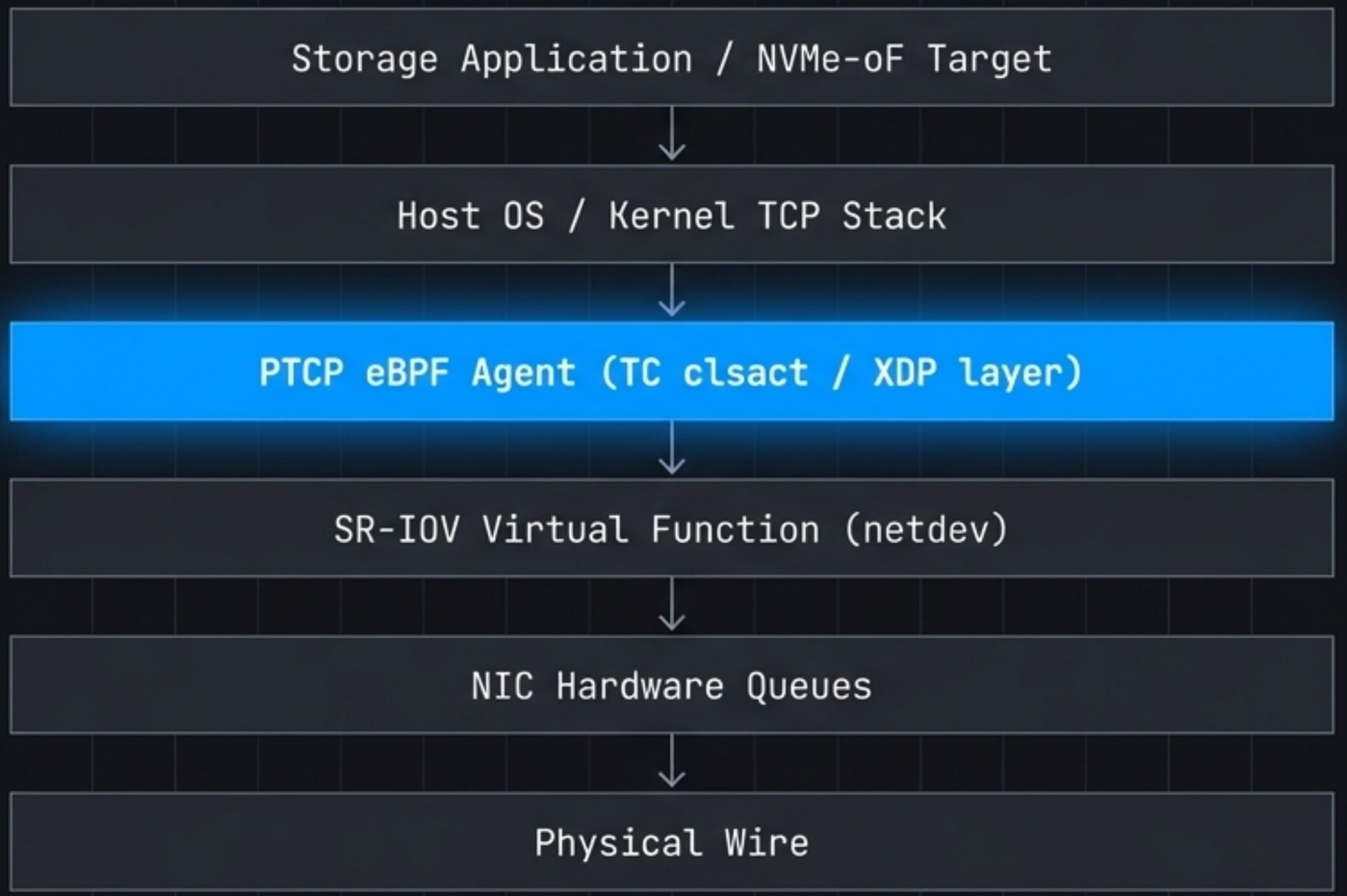
Transforming standard Ethernet into a deterministic, lossless fabric by moving congestion control out of the physical switch and into the storage kernel.

Matrix: Reactive vs. Predictive Control

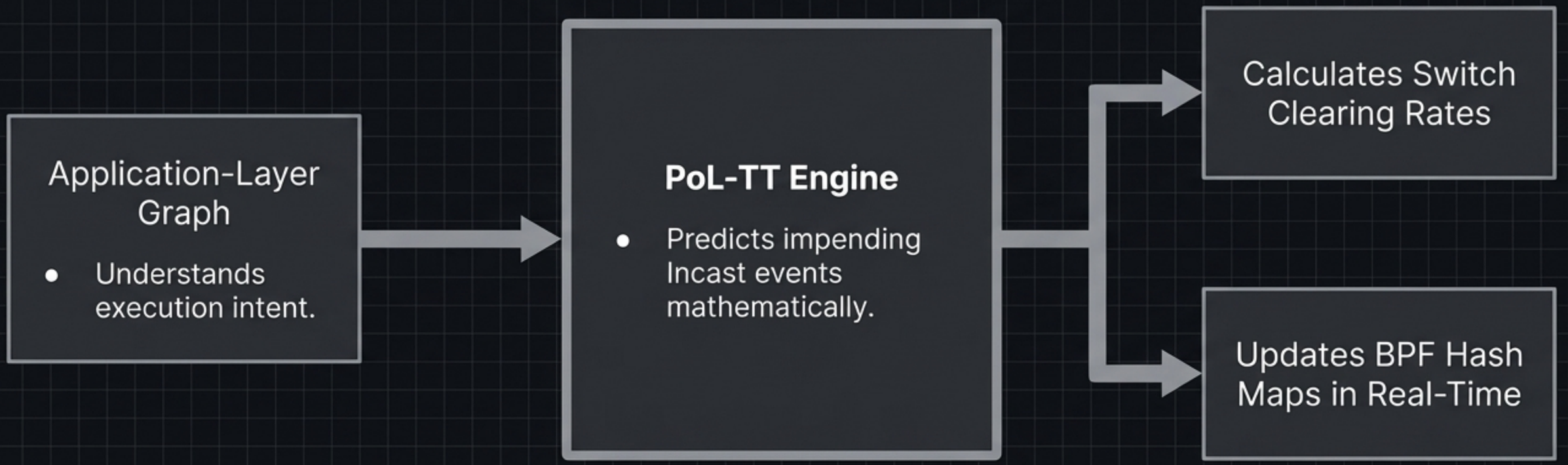
Feature	Legacy TCP / RoCEv2 (Reactive)	PTCP (Predictive)
Congestion	Responds after drops or PAUSE frames	O(1) pacing at the source
Buffer Use	High (Buffer bloat & HoL blocking)	Near-Zero (Mathematical interleaving)
Latency	Jitter-heavy (50ms+ spikes)	Deterministic (Sub-millisecond bounds)
Behavior	AIMD Sawtooth or PFC Storms	Linear, constant-rate delivery
Hardware	Deep-buffer switches or InfiniBand	Commodity Lossless Ethernet

Core Architecture: eBPF & SR-IOV

PTCP paces data after the host OS stack but before the NIC hardware queues, maintaining hardware acceleration while enforcing determinism.



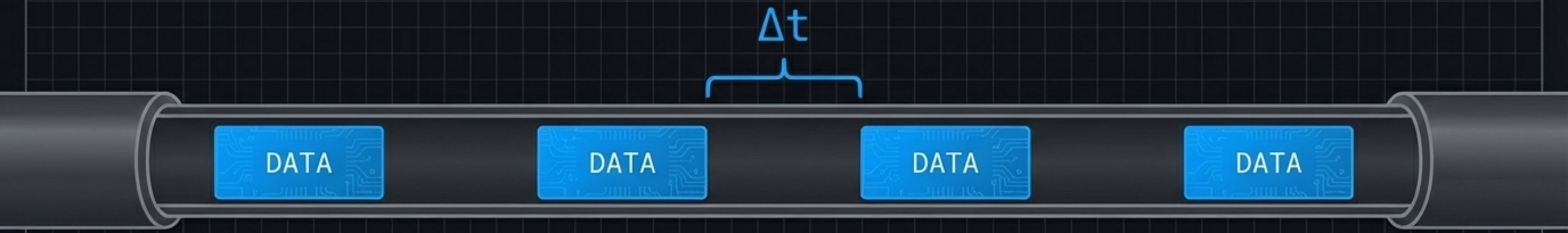
The Brain: Pattern-of-Life Tensor Train (PoL-TT)



The Mechanism: Earliest Departure Time (EDT) Pacing

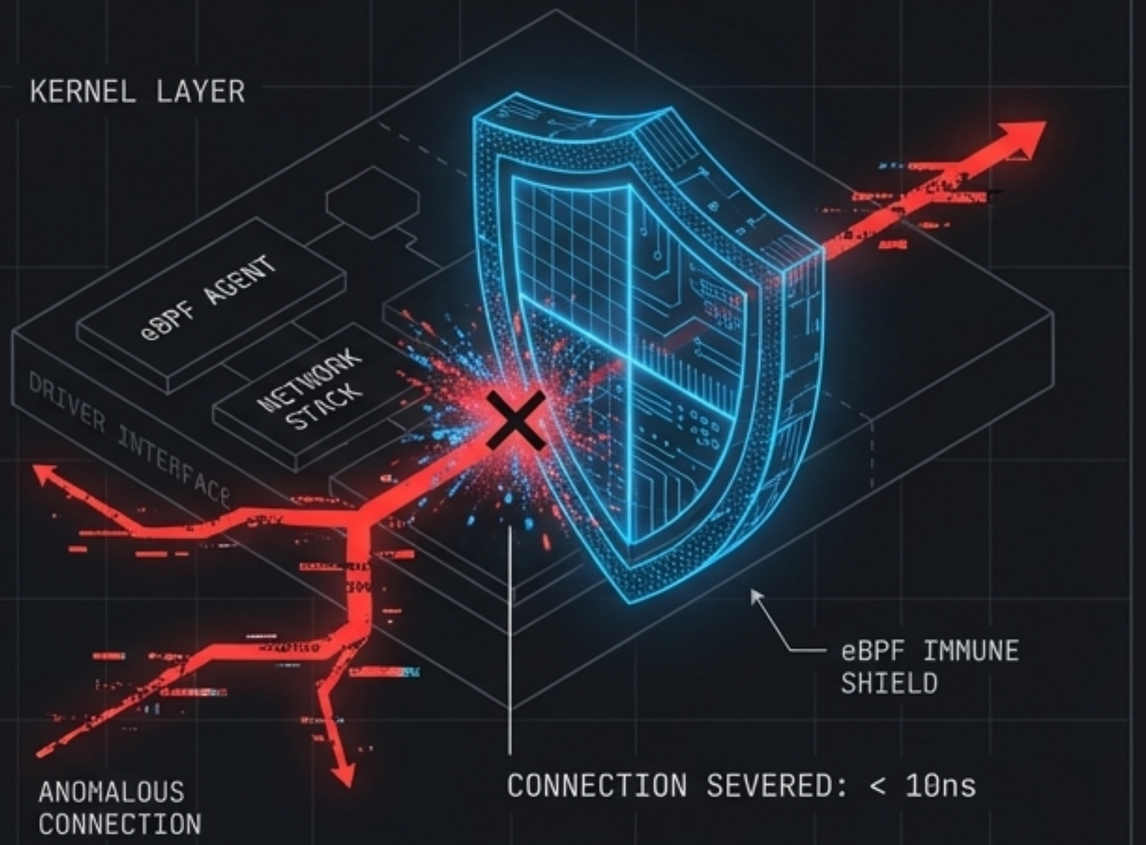
```
tc filter add dev ens1np0v0 egress
   bpf direct-action
   direct-action obj ptcp_tc.o sec tc
```

- The eBPF agent manipulates `skb->tstamp` with O(1) efficiency.
- Inserts nanosecond-precise inter-packet gaps (\Delta Δt).
- Ensures data hits the NIC hardware queues already
- Ensures data hits the NIC hardware queues already mathematically interleaved for the ToR switch.



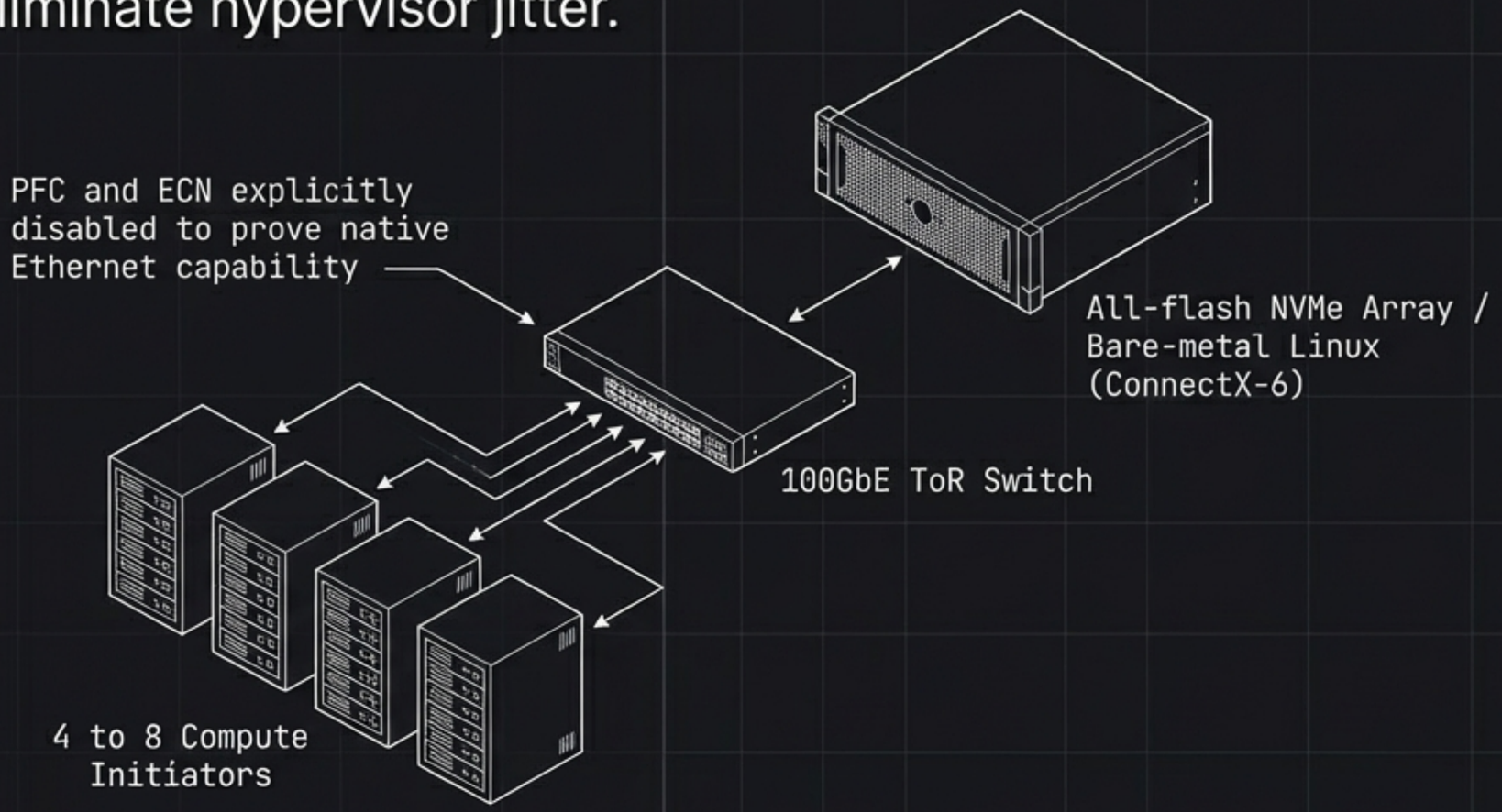
Autonomous Zero-Trust Security

- Acts as a kernel-level immune system.
- Monitors for anomalous mass encryption or lateral data dumps.
- If a node violates its PoL-TT baseline, the eBPF agent severs the connection at the kernel layer in nanoseconds—providing security without the latency of traditional firewalls.

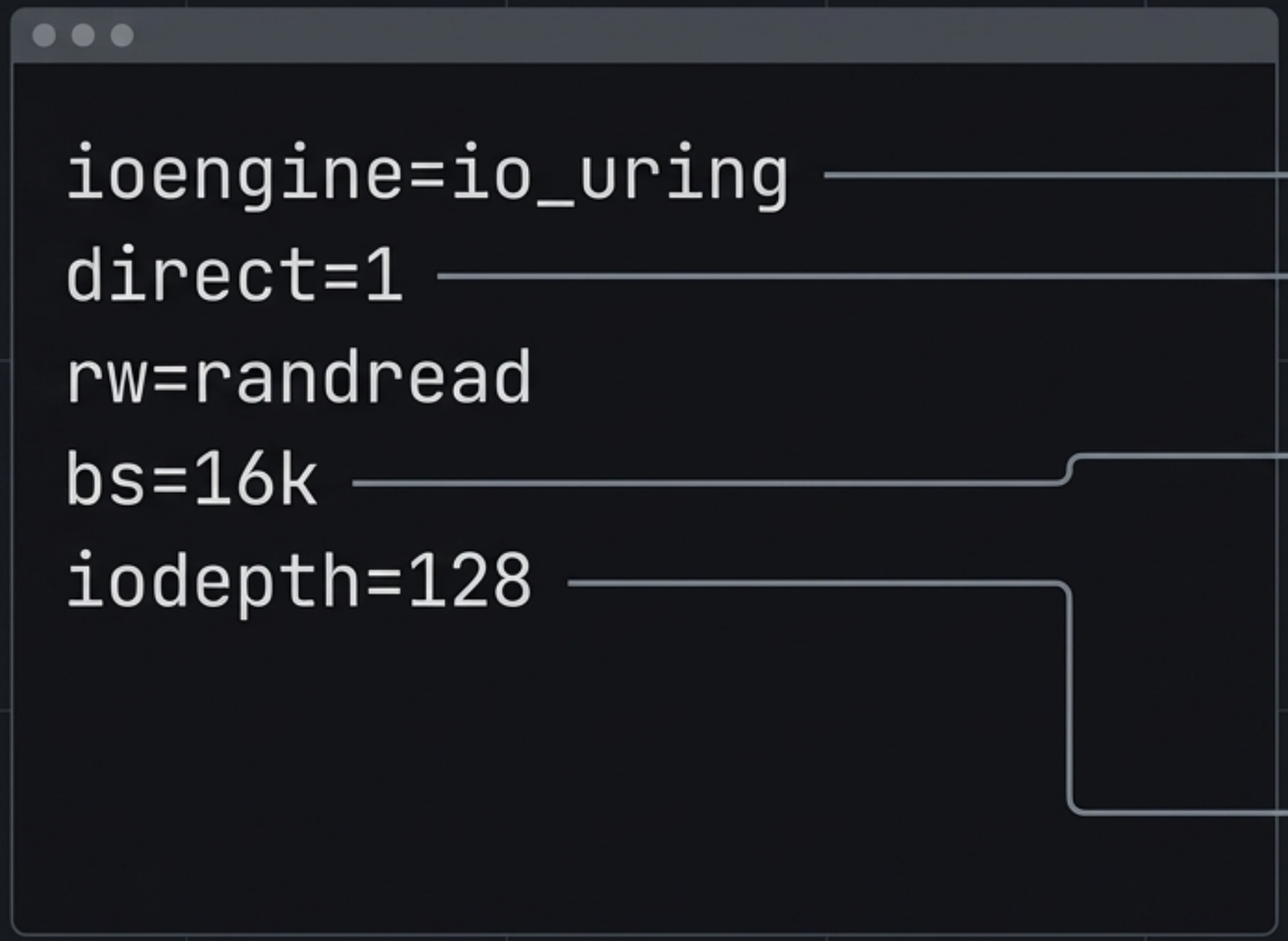


Empirical Validation: The Enterprise Storage Lab

Hardware-isolated Virtual Functions (VFs) presented directly to the storage subsystem to eliminate hypervisor jitter.



Workload Generation: The Stress Test



Bypasses kernel syscall overhead and page cache.

Simulates typical RAG vector retrieval block size.

Forces massive concurrency to trigger Many-to-One Incast.

Phase I: The Signature of Failure (Baseline)

Traffic Control (TC)

Rapidly increasing 'dropped' counts.

```
root@lab-node-01:~# tc -s qdisc show dev eth0
qdisc mq 1: root handle 1: drop_ovl rate 100Gbit limit 1000
    Sent 1593028492 bytes 1061989 pkt  (dropped 2340516б, overlimits 0 requeues 0)
qdisc mq 2: root handle 2: drop_ovl rate 1086bit limit 1000
    Sent 1600129485 bytes 1066859 pkt  (dropped 23498215, overlimits 0 requeues 0)
qdisc mq 3: root handle 3: drop_ovl rate 100Gbit limit 1000
    Sent 1614039281 bytes 1076195 pkt  (dropped 2358930б, overlimits 0 requeues 0)
qdisc mq 4: root handle 4: drop_ovl rate 100Gbit limit 1000
    Sent 1628930128 bytes 1086021 pkt  (dropped 25691450, overlimits 0 requeues 0)
```

PCAP Analysis

A storm of duplicate ACKs, 'fast retransmissions,' and massive tail latency spikes.

```
root@lab-monitor:~# tcpdump -nnvvS -i eth0 tcp port 443
09:30:14.123456 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.]
09:30:14.123460 IP 10.0.0.3.35214 > 10.0.0.2.443: Flags [.], ack 138492302, win 65535, options [nop,nop,TS val 3320038491 ecr 3320038491], length 0
09:30:14.123462 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.]
09:30:14.123464 IP 10.0.0.3.35214 > 10.0.0.2.443: Flags [.], ack 138492302, win 65535, options [nop,nop,TS val 3320038492 ecr 3320038491], length 0
[DUP ACK 1#] (Alert Red)
09:30:14.123467 IP 10.0.0.3.35214 > 10.0.0.2.443: Flags [.], ack 138492302, win 65535, options [nop,nop,TS val 3320038493 ecr 3320038491], length 0
[DUP ACK 2#] (Alert Red)
09:30:14.123469 IP 10.0.0.3.35214 > 10.0.0.2.443: Flags [.], ack 138492302, win 65535, options [nop,nop,TS val 3320038494 ecr 3320038491], length 0
[DUP ACK 3#] (Alert Red)
09:30:14.123475 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [P.], seq 138492302:138493750, ack 394201940, win 65535, options [nop,nop,TS val 3320038495 ecr
3320038494], length 1448 [Retransmission] (Alert Red)
```

THE TOR BUFFERS HAVE OVERFLOWED

Phase II: PTCP-Enforced Fabric

Traffic Control (TC)

High 'delayed' metrics (proving EDT pacing is active), but strictly zero drops.

```
root@lab-node-01:~# tc -s qdisc show dev eth0
qdisc mq 1: root handle 1: drop_ovl rate 100Gbit limit 1000
    Sent 1593028492 bytes 1061989 pkt  (dropped 0, overlimits 0 requeues 0 delayed 45012)
qdisc mq 2: root handle 2: drop_ovl rate 100Gbit limit 1000
    Sent 1600129485 bytes 1066859 pkt  (dropped 0, overlimits 0 requeues 0 delayed 51235)
qdisc mq 3: root handle 3: drop_ovl rate 100Gbit limit 1000
    Sent 1614039281 bytes 1076195 pkt  (dropped 0, overlimits 0 requeues 0 delayed 39012)
qdisc mq 4: root handle 4: drop_ovl rate 100Gbit limit 1000
    Sent 1628930128 bytes 1086021 pkt  (dropped 0, overlimits 0 requeues 0 delayed 48219)
```

PCAP Analysis

Clean, uninterrupted TCP streams. Zero retransmissions.

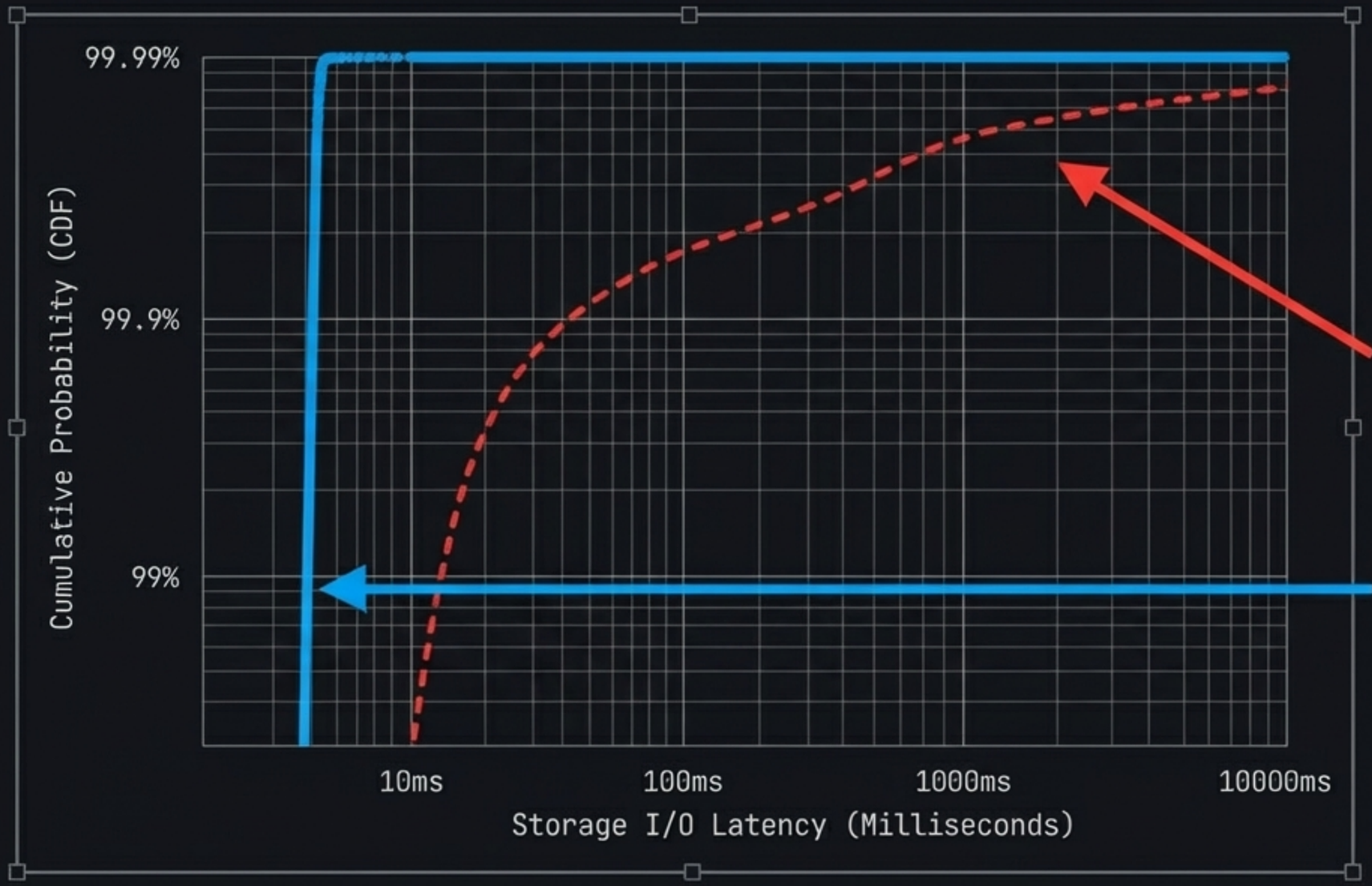
```
root@lab-monitor:~# tcpdump -nnvvS -i eth0 tcp port 443
09:30:14.123456 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.]
09:30:14.123458 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.], seq 138492302:138493750, ack 394201940, win 65535, options  [nop,nop,TS val 3320038495 ecr
3320038491], length 1448
09:30:14.123458 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.], seq 138493750:138495198, ack 394201940, win 65535, options  [nop,nop,TS val 3320038496 ecr
3320038491], length 1448
09:30:14.123460 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.], seq 138495198:138496646, ack 394201940, win 65535, options  [nop,nop,TS val 3320038497 ecr
3320038491], length 1448
09:30:14.123462 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.], seq 138496646:138498094, ack 394201940, win 65535, options  [nop,nop,TS val 3320038498 ecr
3320038491], length 1448
09:30:14.123464 IP 10.0.0.2.443 > 10.0.0.3.35214: Flags [.],
3320038491], length 1448
```

THE MICROBURST HAS BEEN MATHEMATICALLY INTERLEAVED

Measuring the Tail: Cumulative Distribution Function

Decoding the CDF: The ‘ROI Vector’

Medians hide network congestion. The 99.99th percentile dictates AI cluster efficiency.



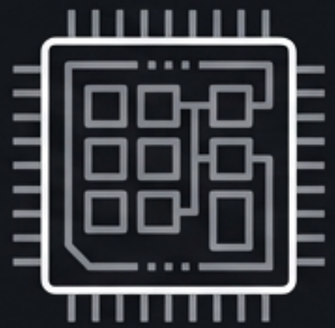
Baseline (Alert Red (#ff3333):
Buffer Bloat / TCP Retransmission Tail. The 'ROI Vector' where multi-million dollar GPUs sit idle.

PTCP (Precision Blue (#00aaff):
Zero-Jitter Mathematical Bound (< 2ms). Eradication of network weather.

Infrastructure Justification & ROI

Pillar 1

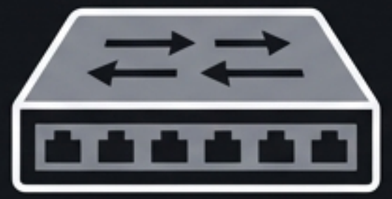
GPU Compute Yield Reclamation



Eliminates `iowait states`; keeps expensive clusters **100% saturated**.

Pillar 2

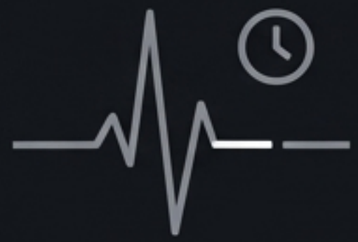
Network CapEx Deferral



`InfiniBand Avoidance`; reduces CapEx by up to **40%**.

Pillar 3

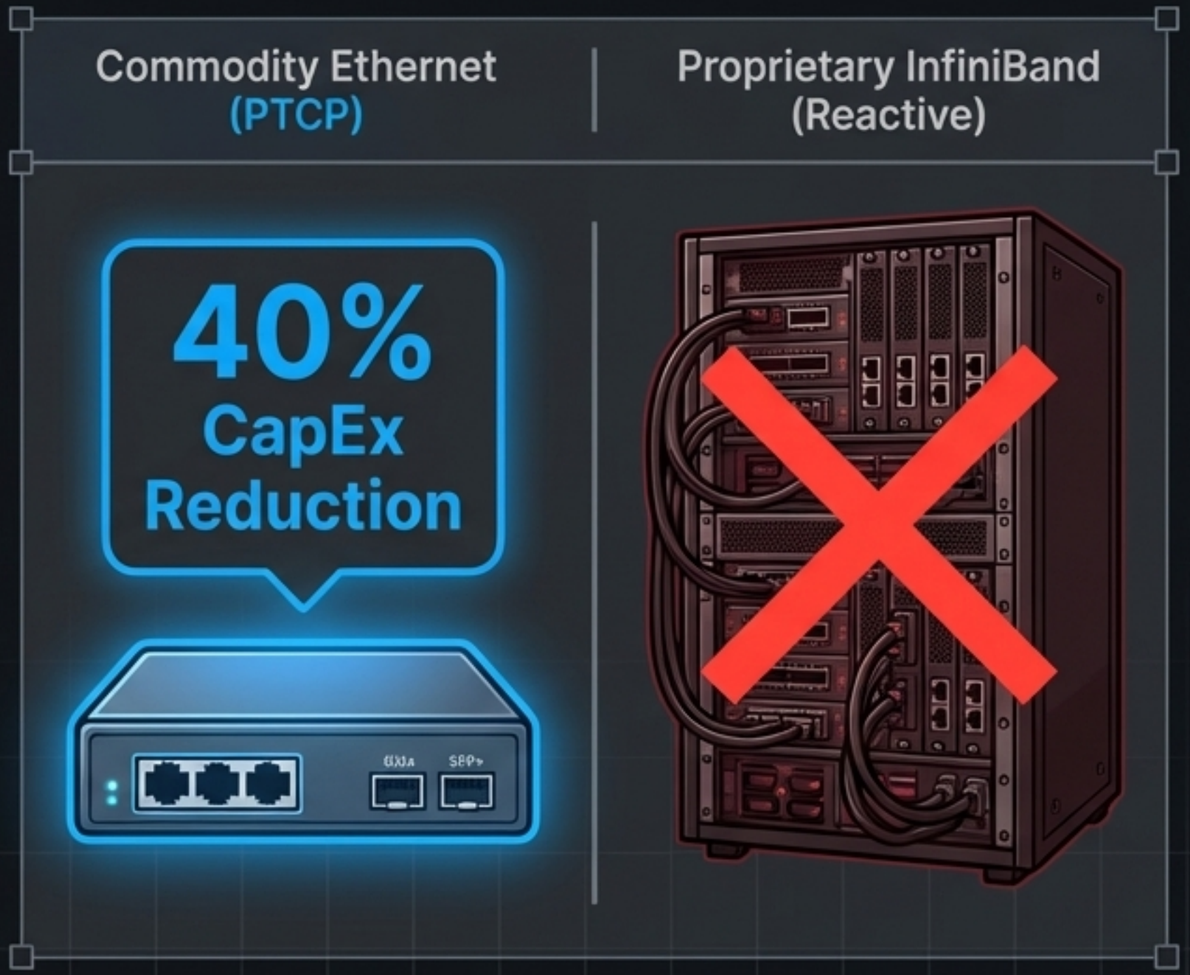
Application Responsiveness



Bounds `P99.99` latency to **sub-millisecond maximums**; accelerates TTFT for synchronous AI interactions.

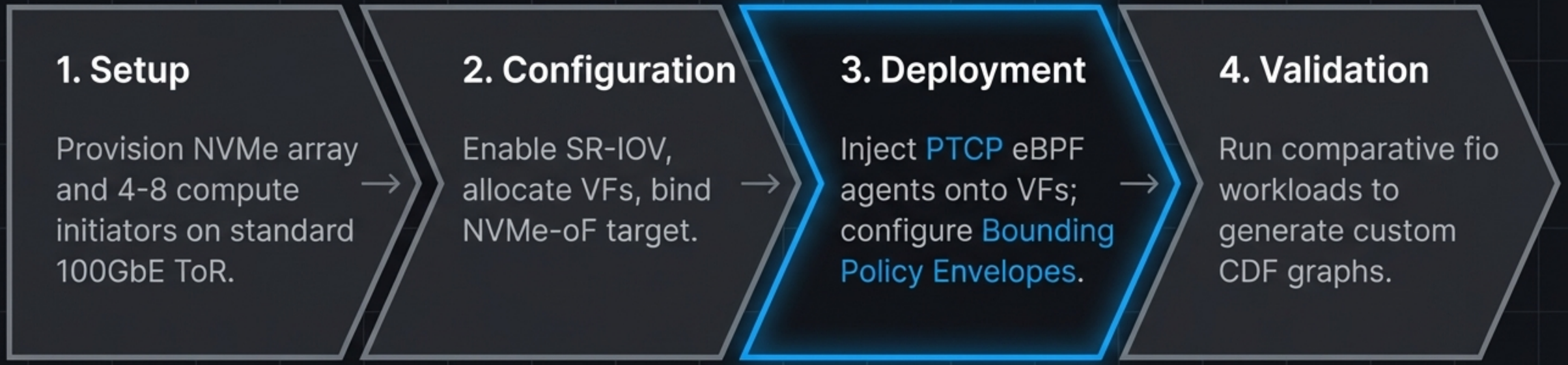
The CapEx Advantage: InfiniBand Avoidance

- **PTCP** provides **lossless, deterministic storage** routing over commoditized Ethernet.
- Eliminates the forced migration to expensive NVIDIA **InfiniBand** hardware or ultra-deep-buffer **ASICs**.
- Achieves high-performance **AI fabric determinism** while reducing network CapEx by **40%**.



The 45-Day Pilot Roadmap

Ring-fenced Enterprise Storage Lab deployment



Pilot Deliverables & Success Criteria

Category	Success Metric Lock-up
Packet Loss Rate	**Zero drops** recorded at the ToR switch under 100% load.
Tail Latency Stability	**99.99th percentile < 2ms** (Eradication of 50ms+ spikes).
Hardware Efficiency	**Negligible CPU overhead** on Storage Target (< 2% increase).
Data Integrity	**Zero TCP retransmissions** in heavy incast scenarios.

The Deterministic Fabric for Modern AI

PTCP reclaims lost GPU yield and defers massive CapEx by enforcing mathematical determinism at the kernel level.

```
Action: Approve Statement of Work (SOW) to commence Phase 1 Pilot.
```