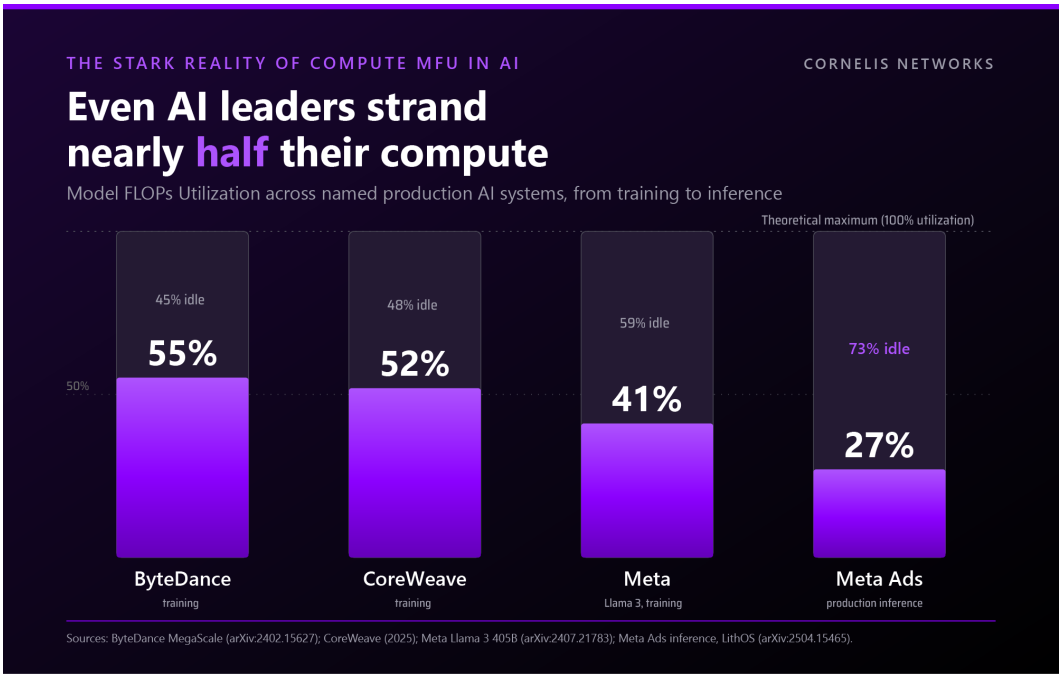
FASTER PIPES, STALLED CLUSTERS: THE NETWORK IS THE CONSTRAINT



The Contradiction In The Data Center

Raw compute is scaling at a rate the industry has never seen to meet AI requirements. Worldwide data center capital spending jumped 57% in 2025, with the four largest US cloud providers alone raising their outlays 76%, and Dell'Oro Group forecasts data center capex will surpass $1 trillion in 2026.[1]

Cluster performance tells a different story. Inside many of the most advanced AI data centers running right now, GPUs that cost $30,000 to $40,000 apiece sit idle a large share of the time, stalled and waiting for data the network cannot move fast enough.

The operators' own reports paint a clear picture. Model FLOPs Utilization (MFU), the share of a cluster's theoretical compute that real workloads actually use, runs from a high of 55% at ByteDance[2] to 52% at CoreWeave,[3] and 41% at Meta on its Llama 3 405B training run.[4] In production inference the picture is starker still: a Meta Ads serving system averaged just 27%.[5] Even the best-run systems turn barely half of the compute they paid for into useful work, and many turn far less. That gap has many causes,[6] from kernel efficiency and memory bandwidth to pipeline bubbles and hardware failures; studies of real deep-learning jobs find low utilization is chronic and rarely traces to a single culprit. But a large and addressable share of it is the network, in the form of exposed communication: the time processors spend waiting on data that cannot be hidden behind computation. It is the largest source of waste an operator can attack by changing the fabric rather than the silicon.

This is not only a training problem. Classic HPC pushed hard, predictable traffic between processors. Modern AI training does the same at a far larger scale, and inference, running in its own clusters, brings a different but equally network-bound pattern: many small, latency-sensitive transfers as a single request moves across pools of hardware. Training and inference reach the same conclusion by different routes. In both, the network increasingly decides how much of the hardware actually works.

So what is wrong with the network? Today's dominant fabrics are passive. They move packets from one point to another and understand nothing about the work riding on top. When tens of thousands of processors have to synchronize every few milliseconds, a passive fabric cannot see the collective operation they are all waiting on, so congestion builds, dropped packets get resent, and exposed communication climbs. Expensive hardware then sits idle as the network catches up. That is the problem the rest of this paper takes on.

What The Market Is Doing About It

Confronted with stalled clusters, operators respond by buying bigger pipes: higher-bandwidth InfiniBand or Ethernet, faster NICs, and denser racks. All these choices attack the symptom, and all carry a very high price tag. The logic is to push more bits per second and trust that the stalls will clear themselves.

Unfortunately, actual deployment analysis proves that more bandwidth does not cure congestion, packet loss, or synchronization stalls. It just leaves the same gap with a higher line rate. The two fabrics carrying nearly all of this traffic, InfiniBand and Ethernet, were never designed for the job in front of them. Both predate modern AI by decades, and both are inherently passive, shuttling data while understanding nothing about the workload riding on top. A passive fabric has no idea that 10,000 GPUs are blocked on the same collective operation, so it can do nothing to protect that moment from a single slow path. Recent products bolt a few smarter features onto the edges, but the core posture has not changed. Run a passive network faster and you still have a passive network. The structural flaw simply travels at a higher speed.

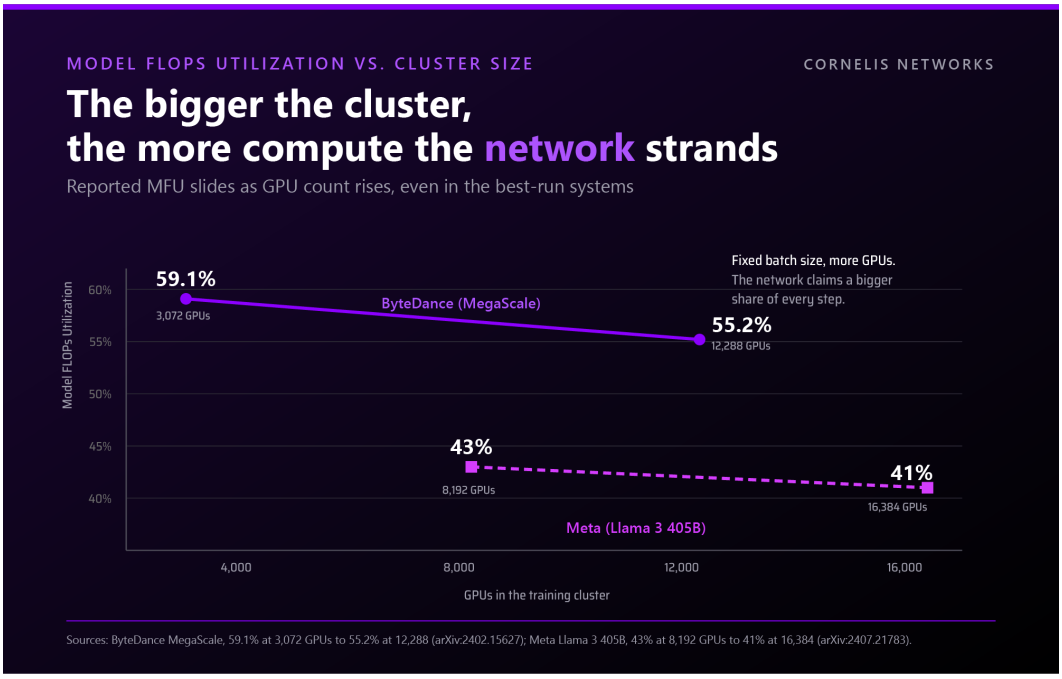
The constraint is not the speed of the pipe. It is everything the pipe fails to manage once tens of thousands of processors depend on it at once. High utilization is something an operator today has to engineer the network for; the hardware does not deliver it alone.

This is the trap the market keeps walking into: every generation of the same architecture gets sold on throughput, and every generation leaves the real problem, idle compute waiting on the fabric, untouched.

At Scale, The Performance Gap Gets Worse

Put a dollar figure on this waste, and it stops looking abstract. Half of the most expensive hardware in the building sitting at idle is not a rounding error in any organization's budget. On a 10,000-GPU cluster, a single percentage point of utilization is worth roughly $3.5M a year, about 100 GPUs running flat out at a typical cloud rate. Moving utilization a few points in either direction is enough to fund or sink entire IT programs. At current prices and current scale, this difference decides whether a build justifies its cost.

You can also read the gap from how operators behave, without anyone admitting the network is the problem. Some overprovision, buying more accelerators than the workload needs so the job still finishes on time even at low utilization. Others brute-force their way past the stalls by standing up a larger cluster to hit a target a well-tuned fabric would have reached with fewer nodes. Both moves are rational, and both are expensive.



Unfortunately, this "solution" only compounds the problem: the waste grows with scale. ByteDance watched utilization slide from 59.1% to 55.2% as it added GPUs, for the plain reason that the ratio of computation to communication falls as a cluster grows.[2] Meta reported the same pattern training Llama 3 405B, where utilization dropped from 43% on 8,000 GPUs to 41% on 16,000.[4] The trend is consistent across operators and architectures. The more you scale, the larger the share the network takes off the top.

A New Scorecard For Your AI Infrastructure

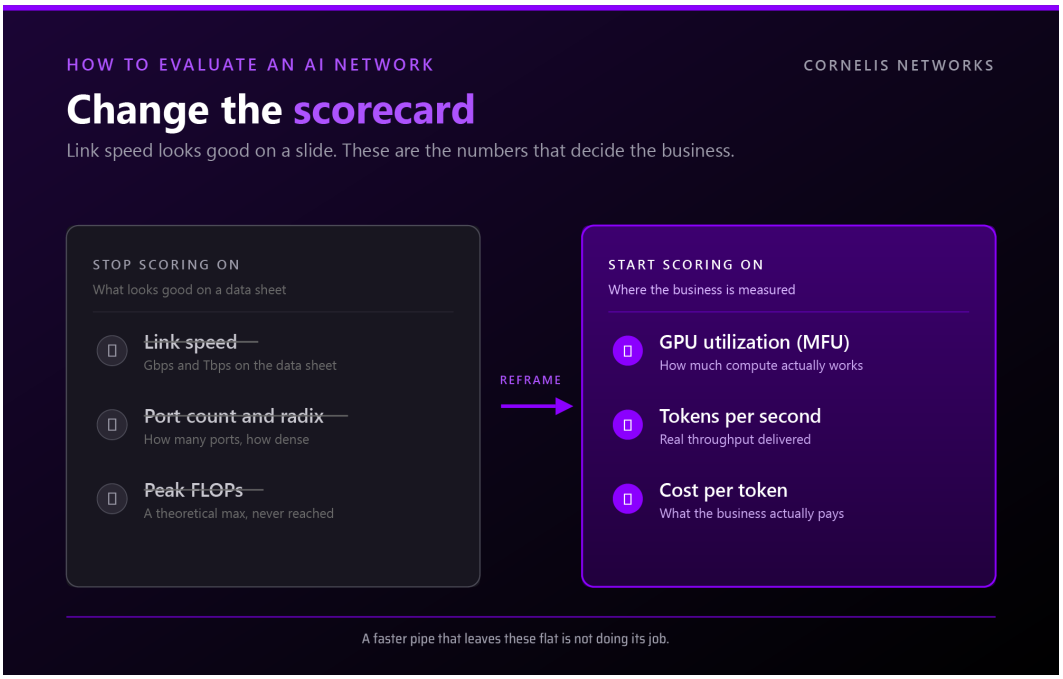
This is where the conversation has to turn from diagnosis to a standards-based path ahead. If the network is the constraint, buyers need to shop for alternatives with urgency backed by economic impact. A short list of what to expect from suppliers includes:

- Lossless delivery at scale, so a single dropped packet never triggers a retransmit that stalls a synchronization step across the whole cluster.
- Congestion managed in real time, as it forms, instead of mopped up after it has already idled the hardware.
- Programmable collective operations that run inside the network itself. Fixed-function offloads already exist on some fabrics, but the network should handle the full, evolving set of AI and HPC collectives, not a handful of hard-wired ones, and without stealing cycles from the processors that should be computing.
- A fabric purpose-built for the communication patterns of AI and HPC, not one retrofitted from an era that never imagined them. It must also adapt as those patterns evolve, from mixture-of-experts inference to AI agents.
- Openness you can hold a vendor to, built on open standards, so the fabric never decides which accelerator you are allowed to run and never locks you to a single vendor for the life of the deployment.

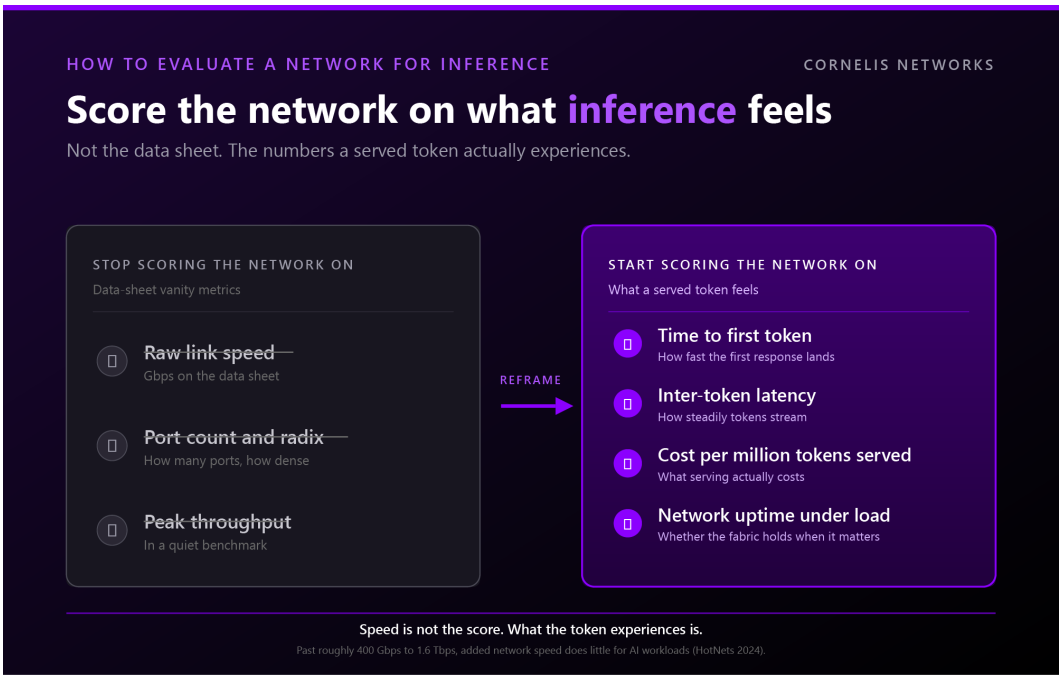
That last expectation grows more important as the accelerator landscape diversifies. The data center of the future will run a mix of architectures, not one.

It matters for a second reason, too. The compute fabric has two jobs, not one. Scale-out connects systems across the data center. Scale-up connects accelerators inside a single system so they work as one machine. Today, those two layers usually come from different vendors running different fabrics, which puts a seam at the busiest point in the system and hands an operator two lock-in decisions instead of one. Carrying one active architecture across both layers is what closes that seam.

No part of this problem is theoretical. But no part of it is straightforward, either. The operators at the top of the utilization range, ByteDance and Meta among them, are already custom-engineering capabilities onto passive fabrics as stop-gaps, waiting for the industry to catch up with off-the-shelf solutions.[4] ByteDance shows both the promise and the limit. It lifted utilization from a 47.7% baseline to 55.2% on 12,288 GPUs, and only by re-engineering how computation and communication overlap.[2] Even with that effort, nearly half the cluster sits stranded. Custom engineering can narrow the gap. It cannot close it. Out of this struggle between ineffective passive networks and costly custom approaches, a third path can emerge: a network that ships standard with the long-standing capabilities the incumbents never solved, and the new demands of AI and HPC they were never built to meet.



That change also demands a new scorecard for the network itself. Stop grading it on the data-sheet numbers that flatter a spec sheet but say little about delivered work: raw link speed, port count, and peak throughput on a quiet benchmark. To select infrastructure that supports the pace of AI innovation, grade the network on what a served token actually experiences: time to first token, inter-token latency, cost per million tokens served, and uptime under real load. A network that moves those numbers is bridging infrastructure investment to AI outcomes. One that posts a fast link speed but leaves them flat is not.



The Reframe

While everyone is focused on GPU advancement, the foundational truth is that the next real performance unlock in AI infrastructure will not come from more compute alone. A large part of it will come from the network that ties the compute together. Compute optimization has carried the industry an extraordinary distance, and is now hitting a limit that more silicon cannot fix. As clusters scale, the bottleneck shifts from the processors to the movement between them.

This is the problem Cornelis is built to solve. For now, the idea worth holding onto is the reframe itself. In a building packed with the most powerful processors ever made, the constraint has shifted to the fabric running between them. The answer is a network that evolves from passive to active, keeping pace with AI software and model innovation. That is where the next gain is hiding, and it will not come from buying more of the same.

References

1. Dell'Oro Group — opener capex "Data Center Capex Surges 57 Percent in 2025 as AI Deployments Accelerate," March 17, 2026.
2. ByteDance / Peking University, MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs. 55.2% MFU on 12,288 GPUs; optimization breakdown; MFU 59.1% to 55.2% with scale. arXiv:2402.15627
3. CoreWeave, CoreWeave Leads the Charge in AI Infrastructure Efficiency (March 2025). >50% MFU on NVIDIA Hopper vs. 35 to 45% public benchmarks; general-purpose clouds not built for AI. coreweave.com
4. Meta AI, The Llama 3 Herd of Models. 38 to 43% BF16 MFU; 43% to 41% from 8K to 16K GPUs; 466 interruptions over 54 days, ~78% hardware-related. arXiv:2407.21783
5. LithOS (Carnegie Mellon University and Meta, 2025) — the Meta Ads production inference figure of ~27% (Figure 1) LithOS: An Operating System for Efficient Machine Learning on GPUs. arXiv:2504.15465
6. Microsoft Research — Gao et al., An Empirical Study on Low GPU Utilization of Deep Learning Jobs (ICSE 2024).

Learn more about industry leading AI and HPC scale-out network at cornelis.com

CORNELIS™

Other names and brands may be claimed as the property of others. All information provided here is subject to change without notice. Contact your Cornelis Networks representative to obtain the latest Cornelis Networks product specifications and roadmaps. Cornelis Networks technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Copyright © 2026, Cornelis Networks. All rights reserved.