The wrong phase: optical interconnect's case is in decode, not prefill

W. Schulz

Jul 30, 2026

Abstract

A source-checked review found the 'networking is 50% of MoE prefill' figure blends training, decode, and an unsupported claim; measured prefill communication is 9–23%. Decode dominates communication, so that is where optical interconnect's case lives. Digital's move to native 4-bit also erodes the precision edge analog designs relied on.

This project's hardware track has one surviving positive claim. Optical *compute* is dead — killed in these pages across three campaigns, most decisively by an analog in-memory competitor that, in this project's cost model, beats it by roughly an order of magnitude on both latency and energy at equal precision. What survived is optical interconnect: the idea that the durable advantage of light is moving data, not multiplying with it.

A hostile verification sweep against the 2026 literature has now found that the claim was attached to the wrong half of the workload. The claim survives. Its phase does not.

The number that started it

The occasion was a talk asserting that *networking can consume up to 50% of total runtime for Llama MoE prefill.* If true, that is a large, well-lit target, and it is where an interconnect story would want to live.

It is not true as stated. The figure is a blend of three real numbers from three different settings:

Real source	What it actually measures
MixNet, LLaMA-MoE, 42–58% [1]	training, 128 accelerators, 400 Gb/s
UBEP, ~50% (SIGCOMM 2026) [2]	decode, on a specific 384-chip supernode
TokenWeave's "~50% for DeepSeek expert parallelism" [3]	traces to an SGLang team blog post that never says it

Every measured *inference-prefill* number is far lower. NVIDIA's own DWDP work reports 9.6% communication plus 12.3% synchronization in the context (prefill) phase on GB200 [5]; TokenWeave measures 9–23% of end-to-end inference latency across three large models on an eight-GPU H100 node [3]. (FLUX's often-quoted 8–11% for dense tensor parallelism on NVLink-connected A100s is a *training* figure [4].) And UBEP — the one genuine ~50% *inference* figure — debunks itself in its own text: MoE communication accounts for roughly half of per-token latency but *only about 20% of actual hardware execution time*, which the authors attribute to dependency stalls [2]. Half of the "networking cost" is not network work.

The premise is inverted

The deeper finding is that the phase assignment is backwards. Communication volume in autoregressive inference is dominated by decode, which generates on the order of 127× more communication operations than prefill (for 128 generated tokens, in one profiling study) [6]. Spreading tensor parallelism across two nodes *improves* time-to-first-token — 90 ms down to 30 ms — while wrecking time-per-output-token [6].

The mechanism is not subtle. Prefill is compute-bound and therefore tolerates communication; it has arithmetic to hide behind. Decode has almost none, so once it goes off-node it goes communication-bound.

This project's own interconnect analysis framed the optical-KV story around data movement in general, and leaned on prefill-side intuitions in doing so. The correct phase is decode, and two chapters already published in this collection — the window-law analysis and the closure of the prefill flank — carry the older framing.

The relocation makes the case stronger

This is the unusual case where a correction improves the claim it corrects, because the decode literature supplies a first-party number the prefill literature never did.

DeepSeek's own systems paper works through the regime in which communication *is* the entire critical path: at 50 GB/s of InfiniBand bandwidth, 120.96 μs per expert-parallel communication step (two steps per layer, 61 layers) implies a floor of 14.76 ms per output token — about 67 tokens per second. At 900 GB/s of NVLink (a GB200 NVL72, which is not the hardware DeepSeek-V3 was served on), the same arithmetic gives 0.82 ms, or roughly 1200 tokens per second. That is 18× from interconnect bandwidth alone — a theoretical upper-bound calculation by the model's authors, who state that real deployments will fall well short of it, not a measurement [7].

And FLUX shows that it is the *interconnect*, not the workload, that sets the number: 8–11% exposed communication in training on NVLink-connected A100s becomes 40–75% in training and prefill on A100s connected over PCIe [4]. The fabric is the variable.

Interconnect bandwidth binds hardest in decode — which is the same phase where data residency wins, and the same phase where analog in-memory compute does *not* help, because long-context decode is bound by the key-value cache rather than by multiplication. So the two halves of this program's hardware verdict interlock rather than compete. Analog compute owns the compute half of inference. The decode half is an interconnect problem, and that is where optics has its case.

The four standing verdicts, re-gated

Claim	Verdict
Analog in-memory owns the compute half (prefill, short-context decode), not long-context decode	Split. The decode half is confirmed four ways over, by primary sources that keep the cache and attention digital by design. The compute half needs revision: digital compute-in-memory now wins the like-for-like integer metric, and our own speedup figures have no external support.

Claim	Verdict
Optical compute dead; optical interconnect survives	Still valid, with two corrections: the reach qualifier moves to *beyond* the rack — copper won the current scale-up domain outright — and the phase is decode, not prefill.
The mechanism is data residency, not cheap multiplies	Still valid, and the strongest surviving claim. One weight-resident architecture posts roughly 3.6–4× the accuracy-per-joule of a current flagship GPU (a ratio derived from published per-device measurements [8]); another delivers 4× lower per-token decode latency than an A100 while delivering 41× lower prefill throughput [9]. Residency behaves exactly as characterized.
The win decays as model width grows	Needs revision — right direction, wrong variable. It should be restated against total parameters versus a frozen per-reticle memory ceiling, with mixture-of-experts *accelerating* the decay.

The finding to act on first — our cost models are stale for the opposite reason

This is the part that inverts an argument rather than relocating it.

> Only about 1.5× per year of the digital baseline's improvement is silicon. The headline figures of 18–33× in sixteen months are precision, architecture and software. [verification pending] So this project's cost models are not stale because GPUs got faster. They are stale because digital moved to native 4-bit — which is exactly analog's accuracy ceiling.

That is the whole moat argument turned around. This program priced analog's advantage as *tolerance of low precision*. Digital has now taken the low-precision territory outright, at the precision where analog tops out, and the analog vendor's own research arm published the ceiling [10]. Any future analog claim in this project must clear a native-4-bit digital baseline, not an 8-bit one.

Citation hygiene, applied to our own drafts

The sweep produced a set of rules that the project's own papers now have to follow, because the failures it found are the kind we could commit.

A widely cited expert-parallelism percentage differs across versions of the same paper — 34.1% in one, 37.4% in another [13]. Cite it with a version number or not at all. The same paper's table splits inference at 26.2–32.5% against training at 35.4–39.5%: inference runs about eight points *lower*, which is the direction that undercuts the headline it is usually cited to support. A frequently quoted 34–46% figure is simulation, not measurement. And a 68% claim cites two papers, neither of which says it — one says 40%, for training.

The sweep also audited inputs this repository itself relayed. A widely repeated "NVIDIA for prefill, Cerebras for decode" deployment split does not exist. A ~3× intelligence-per-watt figure over two years is actually 5.3×, of which 3.1× was the model-only term [8]. "Imperfect routers capture 50–70%" is really 80% of oracle, 64.3% absolute [8]. A ~150 TOPS/W analog figure is an 8-bit test-chip claim; the shipping card works out to about 24 (200 TOPS peak in an 8.25 W card, precision of the peak figure not

stated; the vendor quotes up to 40 TOPS/W at 8-bit for the NPU cores alone) [11]. A co-packaged-optics power-efficiency figure quoted at 3.5× was the vendor's 2025 launch number; its current figure is 5× [12]. Two relayed numbers checked out unchanged.

One gap is flagged as unclosed: copper reach in metres at the current per-lane signalling rate, and the copper-versus-optical energy-per-bit crossover. Five attempts produced no primary source. Any crossover-reach constant in this project's photonic cost models is therefore unsourced and must be labelled as such — which is a live caveat on published work, not a note for the future.

What this changes, and what it does not

Nothing here rehabilitates optical compute. The interconnect survivor keeps its reach qualifier and gains a phase correction plus a much better first-party number, so the net effect on the campaign's four standing verdicts is smaller than the effort that produced it — which is the ordinary outcome of a gate that works.

What it changes is where to look. The claim was pointed at the phase with the friendlier headline and the weaker case. Moved to the phase where the constraint actually binds, it is supported by the model authors' own bandwidth arithmetic rather than by a blend of three misquoted percentages.

Sources

1. Liao, X., Sun, Y., Tian, H., Wan, X., et al. (2025). "MixNet: A Runtime Reconfigurable Optical-Electrical Fabric for Distributed Mixture-of-Experts Training." *ACM SIGCOMM 2025*. arXiv:2501.03905 — the two all-to-all phases take 42–58% of LLaMA-MoE *training* iteration time.
2. Liu, Y., Liu, C., Shen, S., et al. (2026). "UBEP: Re-architecting Expert Parallelism Communication Library for Production Superpods." *ACM SIGCOMM 2026*. arXiv:2607.06202 — MoE communication is roughly 50% of per-token decode latency but about 20% of hardware execution time.
3. Gond, R., Kwatra, N., & Ramjee, R. (2025). "TokenWeave: Efficient Compute-Communication Overlap for Distributed LLM Inference." arXiv:2505.11329 — 9–23% of end-to-end inference latency on 8×H100; its "~50%" DeepSeek figure is attributed to an SGLang team post.
4. Chang, L.-W., Bao, W., Hou, Q., et al. (2024). "FLUX: Fast Software-based Communication Overlap On GPUs Through Kernel Fusion." arXiv:2406.06858 — about 8–11% exposed tensor-parallel communication in A100 NVLink training; 40–75% in A100 PCIe training and prefill.
5. Li, W., Peng, J., Jing, Z., Zhang, T., et al. (NVIDIA) (2026). "DWDP: Distributed Weight Data Parallelism for High-Performance LLM Inference on NVL72." Technical report. arXiv:2604.01621 — context-phase kernel breakdown on GB200: communication 9.60%, synchronization 12.26% of iteration latency.
6. Xu, L., Kandadi Suresh, K., Anthony, Q., et al. (2025). "Characterizing Communication Patterns in Distributed Large Language Model Inference." arXiv:2507.14392 — decode generates 127× more communication operations than prefill for 128 output tokens; TP=8 across two nodes improves TTFT from 90 to 30 ms while TPOT rises to 11.56 ms.
7. Zhao, C., Deng, C., Ruan, C., et al. (2025). "Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures." *ISCA '25*. arXiv:2505.09343 — the 14.76 ms (50 GB/s IB) and 0.82 ms (900 GB/s NVL72) TPOT figures, both stated as theoretical upper limits.
8. Saad-Falcon, J., Narayan, A., Akengin, H. O., Griffin, J. W., et al. (2025). "Intelligence per Watt:

Measuring Intelligence Efficiency of Local AI." arXiv:2511.07885 — 5.3× intelligence-per-watt gain 2023–2025 (3.1× model, 1.7× accelerator); a router at 80% accuracy captures 80% of the oracle's 80.4% energy saving; per-device intelligence-per-joule for SN40L and B200.

9. Usami, S., Vishwanath, V., & Bethel, E. W. (2026). "Prefill/Decode-Aware Evaluation of LLM Inference on Emerging AI Accelerators." arXiv:2606.17104 — on Llama-2-7B, A100 GPUs reach up to 41× higher prefill throughput; GroqRack shows 4× lower per-token decode latency.
10. Büchel, J., Chalas, I., Acampa, G., Chen, A., et al. (IBM) (2025). "Analog Foundation Models." *NeurIPS 2025.* arXiv:2505.09663 — off-the-shelf LLMs do not reach 4-bit-level performance on AIMC hardware without hardware-aware training.
11. Mann, T. (2025, Feb 17). "The future of AI is … analog? Upstart bags $100M to push GPU-like brains on less juice." *The Register.* theregister.com — EnCharge's 150 TOPS/W claim at 8-bit, for test chips. EN100 product specification (M.2 card: 200+ TOPS peak, card power up to 8.25 W): EnCharge AI, en100.enchargeai.com.
12. NVIDIA, "Silicon Photonics" product page, nvidia.com ("5x better power efficiency"); launch figure "3.5× more power efficiency" as quoted in *Optics & Photonics News*, March 2025, opticaopn.org.
13. Li, J., Jiang, Y., Zhu, Y., et al. (2023). "Accelerating Distributed MoE Training and Inference with Lina." *USENIX ATC '23.* arXiv:2210.17223 — all-to-all share is 37.4% in v1 and 34.1% in v2; v2's table gives inference 26.2–32.5% vs training 35.4–39.5%.

The sweep's working notes, including the claim-drift audit table, are in `codesign-analog-moe/MASTER-SYNTHESIS.md` *in the Quantum-Intelligence repository.*

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/optical-interconnect-decode