Can analog or photonic hardware replace LLM inference silicon? AIMC for compute, optics for decode interconnect

W. Schulz

Jul 30, 2026

Abstract

About 54 pre-registered experiments found no general replacement. Analog MoE routing adds no accuracy, and the FFN it accelerates is under 19% of decode energy. Optical interconnect's energy win collapses under honest laser power, leaving bandwidth density at extreme scale, mostly in decode.

Consolidates the whole campaign: ~54 pre-registered experiments across 5 run-once rounds + a 106-agent deep-research literature cross-check + the B6 physics decider. Question: *can neuromorphic photonic/analog hardware replace inference hardware for foundational LLMs?* Method throughout: pre-register predictions + kill-conditions, run once, report failures, no strawmen (methodology-honest-evaluation).

Substrate for algorithm rounds: frozen ViT-B/16 CIFAR-100 + TinyImageNet + synthetic. Cost rounds built on calibrated `pim_baseline.AIMC_RERAM` constants. Diagnostic testbed — read relative movements, not absolute accuracies.

What each round settled

Round 1 — where's the bottleneck? (E1–E10)

Substrate works: selection law `ACC=A_sel·p_within` transfers to analog exactly (E1); analog experts feasible at 3-bit/ENOB~4 (E2); MoE recreates the batch-1 weight-resident regime at E≈128 (E5); PISA-style dependency-aware dispatch mapping cuts stalls 92% (E10). Bottleneck = routing ε≈0.5, and only routing (E9).

Round 2 — the router (R1–R10)

Corrected round 1's own headline: "write-time structured codes" were an artifact of a badly-fit baseline — structured codes == trivial ridge-to-one-hot exactly (R1/R2/R3/R8). Real lever = a plain discriminative ridge linear router (CIFAR 0.42→0.73). With it, ACC rises with expert count (R5) and survives 4-bit (R7). The reframe that matters: a monolithic classifier (0.760) BEATS the best MoE (0.741) — routing buys NO accuracy, only conditional compute. Soft routing, gating, metric keys all failed.

Round 3 — sparse-analog compute economics (C1–C10)

C3 is decisive (Amdahl): the FFN that analog-MoE accelerates is 18.8% of decode energy dense, 2.8% sparse; attention+KV dominate 81–99%. Conditional compute touches <3% of decode. C6: 9-pt accuracy tax, unfixed by grouping. C9: on language-like overlapping data the routing lever evaporates. C10: drift not a blocker. (C2 leakage model inconclusive; C5 "476×" was a retired weight-restreaming strawman.)

Round 4 — optical interconnect for KV movement (I1–I13)

Attacks the 81%. Optical KV movement = modest win in the *right* place. I2 break-even 3.2 cm; I6 dispatch is tiny vs KV. I11 (sensitivity): the energy win is NOT robust — flips at opt floor ~2.84 pJ/b vs base electrical (2.4× margin) and ~1.68 pJ/b vs advanced SerDes (1.4× margin); honest laser wall-plug (~3 pJ/b) puts optics *behind* advanced electrical. I12/I13 (GQA/quant realism): every threshold shifts 8–64× — pin-wall L* moves 24k→381k→1.5M context; on-package HBM holds batch 3–165 at 32k, so off-package is forced only by ~1M context or high batch, not ordinary serving.

Round 5 — stress the interconnect win (S1–S10)

The board-level energy win is DEAD under honest laser power: utilization cliff at U=0.42 (S2); at realistic U=0.3 optics is 3.4× worse (S10); end-to-end −32% (S9). Culprit = laser static × low utilization (thermal negligible, S4). Survives only at long reach (rack/disaggregated, S5/S8) and as a power-independent bandwidth-density necessity (S7) — which I12 shows only bites at extreme scale.

B6 — the physics decider (deep-research-identified)

REFUTED as a blocker. Analog accumulation SNR is width-independent to d=8192 for ternary/2/4-bit, for both tiled and untiled, as long as the ADC is range-calibrated ($\sqrt{d}$ signal / $\sqrt{d}$ noise cancel). Collapse appears only under a naive fixed-range ADC — a design mistake real AIMC avoids. Binding constraint reverts to the known B5 ceiling (~4-bit + hardware-aware retrain + drift). Residual risk = *correlated* column drift (a B5/retention problem), not width.

The final verdict

Can photonic/analog hardware replace foundational-LLM inference hardware? No — but there is one narrow, honest niche, and it is an interconnect story, not a compute story.

1. Analog compute is feasible but pointless for this goal. It works (B6, E2), but (a) MoE routing buys zero accuracy over a monolithic classifier (R-round), and (b) the FFN it accelerates is <19% of decode energy (C3). You'd be optimizing the wrong 19% with a structure that adds a routing tax.
2. The only durable lever is optical interconnect for data movement — and even it is bounded on two sides:
    - Energy win is fragile. It flips within the plausible CPO parameter range once laser wall-plug power and realistic (~30%) link utilization are charged (S-round, I11). It survives only at long reach (rack/disaggregated).

- The robust part — bandwidth-density necessity — only bites at extreme scale. With the GQA + KV-quantization every 2026 model uses, electrical suffices to ~381k context; optics becomes *physically required* only at ~1M-context / high-batch pooled-KV serving (I12/I13).

3. So the defensible north star is narrow and specific: *optical interconnect for KV/activation movement in extreme-scale serving (long-context, high-batch, disaggregated-memory) — deployed for bandwidth density you cannot get electrically, NOT for an energy win, and NOT to do the model's math.* It competes against advancing electrical SerDes and mature programmable-switch ASICs.

This reconverges — from four independent directions — on the program's standing conclusion (project-photonic-falsification-machine, project-validate-and-ship): the durable optical win is interconnect, not compute; and a strong simple baseline beats the clever structure. The co-design is not refuted as physics (B6), but it is bounded to a narrow, interconnect-only, extreme-scale niche as an *engineering* proposition.

Honest bounds on this whole campaign

- Algorithm rounds use frozen-ViT/prototype experts — directional, not exact vs LLM works that train experts end-to-end (deep-research caveat A_testbed).
- Cost/link constants are literature-scale ASSUMPTIONS; I11 shows the energy verdicts are sensitive to them (the density verdict is sturdier).
- Not yet built: B7 (photonic-vs-AIMC for transformers, externally unresolved — LightML tension).

B8 — the decisive system build (DONE): confirms the terminus

End-to-end decode cost, every finding baked in (`results/SYNTHESIS-B8.md`, `b8_results.json`): - B8a (moderate, on-package KV): analog-MoE (10,108 μJ/tok) ties *digital* expert-parallel (10,258, ±1.5%) while paying a 1.9-pt accuracy tax. The conditional-compute win is an MoE-sparsity win digital hardware also captures — analog adds nothing at the system level and costs accuracy. - B8b (extreme, off-package KV): optical KV at honest 4.6 pJ/b wall-plug = 1.6× *worse* energy than board-electrical. Optics is an energy-LOSER. - B8c: optics' only durable rationale = bandwidth-density feasibility (electrical needs metres of impossible beachfront at scale). - B8d verdict: NO general replacement. The surviving niche (optical pooled-KV fabric at extreme-scale serving, deployed for feasibility not energy) = already Celestial AI / Lightmatter's shipping thesis → the differential contribution is quantification + negatives, not a new direction.

FINAL TERMINUS (co-design): the analog-MoE co-design is feasible physics (B6) but not a worthwhile engineering bet — MoE routing adds no accuracy and the FFN is the wrong slice. Fifth reconvergence on interconnect-not-compute + strong-baseline-wins.

A-series — "But what about AIMC?" (REVISES the terminus)

The MoE framing hid AIMC (the program's established right competitor). Running full-model AIMC directly (`results/SYNTHESIS-aimc.md`, `aimc_results.json`): - A4: AIMC crushes static-weight matmul — 30.6× vs digital, 4.2× vs optics (worse-PIM re-confirmed: if analog, AIMC not photonics). - A2/A6: full-model AIMC decode win is real but context-dependent — 3.86× at L=2k → 1.2× at L=32k → 1.05× at L=131k (dilutes as KV movement dominates). - A3: AIMC's real niche is PREFILL (compute-bound, no KV re-read): 19× at P=512, 9.3× at P=2048, diluting only when P^2 attention takes over. - A5: ADC dominates AIMC energy (60%), not the 2 fJ MAC — an interface story like optics. - A7: feasible at 4-bit + hardware-aware retrain + drift-managed (B6: not a width wall).

REVISION: "analog compute is a system-level no-op" was too harsh — an artifact of the MoE-long-context-decode framing. AIMC is a REAL efficiency for the COMPUTE-BOUND regime (prefill 9–19×, short-context decode ~4×), the strongest analog substrate, beating optics. It does NOT transform long-context autoregressive decode (~1.2×, KV-bound — the same wall optics/MoE hit). Clean separation of the program's two truths: AIMC owns the compute half of LLM inference (prefill/short-context/batched); the KV/interconnect problem owns the other half. "Replace inference HW" is still No as a whole, but AIMC is a defensible accelerator for the (large, growing) compute-bound share — a complement, not a replacement.

UPDATE 2026-07-30 — the interconnect claim was on the WRONG PHASE (deep-research sweep)

A hostile verification sweep against the 2026 literature, prompted by a YC kernel/chip talk that asserted "*networking can consume up to 50% of total runtime for Llama MoE prefill.*"

The 50%-of-prefill number is REFUTED as stated — it is a blend of three real numbers

Real source	What it actually measures
MixNet, LLaMA-MoE 42–58%	training, 128×H800, 400 Gb/s
UBEP ~50% (arXiv 2607.06202, SIGCOMM'26)	decode, Huawei CloudMatrix384
TokenWeave's "~50% DeepSeek EP"	traces to an SGLang blog that never says it

Every measured inference-*prefill* number is far lower: 9.6% comms + 12.3% sync on GB200 (NVIDIA DWDP); 8–11% dense TP=8 on A100 NVLink (FLUX, arXiv 2406.06858); TokenWeave measures 9–23% end-to-end across Llama-3.3-70B / Qwen2.5-72B / Mixtral-8x22B. And UBEP, the one genuine ~50% *inference* figure, debunks itself: *"MoE communication accounts for roughly 50% of per-token latency … but only about 20% of actual hardware execution time, suggesting substantial dependency stalls."* Half the "networking cost" is stall, not network work.

The premise is INVERTED, and this relocates our own claim

arXiv 2507.14392 (OSU, Jul 2025): *"The decode stage dominates communication volume … generating 127× more operations than prefill."* TP=8 across two nodes *improves* TTFT (90→30 ms) while wrecking TPOT. Prefill is compute-bound and tolerates communication; decode goes communication-bound off-node.

This program's interconnect analysis (rounds 4–5, I1–I13 / S1–S10) framed the optical-KV story around movement in general and leaned on prefill-side intuitions. The correct phase is decode.

And the relocation makes the case STRONGER, first-party

DeepSeek ISCA'25 (arXiv 2505.09343) — in the dual-microbatch-overlap regime communication *is* the whole critical path: 120.96 μs/layer at 50 GB/s InfiniBand → 14.76 ms TPOT floor (67 tok/s); at 900 GB/s NVLink → 0.82 ms (~1200 tok/s). 18× from interconnect bandwidth alone. And FLUX shows the *interconnect*, not the workload, sets the number: 8–11% on A100 NVLink vs 40–75% on the same A100s over PCIe.

Interconnect bandwidth binds hardest in decode — the same phase where data residency wins and where AIMC does *not* help (A-series: long-context decode ~1.2×, KV-bound). The two halves of this program's verdict interlock rather than compete: AIMC owns the compute half; the decode half is an interconnect problem, and that is where optics has its case.

Citation hygiene flagged by the sweep — applies to our own drafts

- Lina's all-to-all percentage differs across v1 / v2 / ATC'23 (34.1% vs 37.4%). Cite with a version number or not at all. Its Table 1 also splits inference 26.2–32.5% vs training 35.4–39.5% — inference runs ~8 points *lower*, the direction that undercuts the headline.
- T3 (ASPLOS'24) 34–46% (→75% if GEMMs double) is SIMULATION, not measurement.
- FlashMoE's "68%" cites Janus and Lancet; neither says it (Lancet says 40%, training).

Net effect on the campaign's four standing verdicts: unchanged. The optical-interconnect survivor keeps its rack-reach qualifier and gains a phase correction and a much better first-party number. Nothing here rehabilitates optical *compute*.

The four standing verdicts, re-gated against 2026 sources

#	Claim	Verdict
(a)	AIMC owns the COMPUTE half (prefill 9–19×, short decode ~4×), not long-context decode (~1.2×, KV-bound)	SPLIT. Decode half STILL VALID — confirmed 4-for-4 by IBM primary sources that keep KV+attention digital by design. Compute half NEEDS REVISION: digital CIM now wins the like-for-like INT8 metric, and our 9–19× / 4× / 1.2× have no external support
(b)	Optical compute dead; optical interconnect survives	STILL VALID, two corrections: reach qualifier moves to *beyond* rack (copper won the NVL72 scale-up domain outright), and the phase is decode, not prefill
(c)	The mechanism is data residency, not cheap multiplies	STILL VALID — the strongest surviving claim. SN40L ~3.6–4× B200 accuracy-per-joule; GroqRack 4× faster decode while 41× slower at prefill — residency exactly as characterised
(d)	The win decays as d_model grows	NEEDS REVISION — right direction, wrong variable. Restate against total params vs a frozen ~0.5 GB/reticle ceiling, with MoE *accelerating* the decay

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

> Only ~1.5×/yr of the digital baseline's improvement is silicon. The 18–33×-per-16-months headline numbers are precision + architecture + software. So the 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 (IBM's own NeurIPS 2025 result).

That is the whole moat argument inverted. 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. Any future analog claim must clear a native-4-bit digital baseline, not an INT8 one.

Claim-drift audit — inputs that were wrong, including ones this repo relayed

Input	Truth
"NVIDIA for prefill + Cerebras for decode" (deployed split)	does not exist
"~3× intelligence-per-watt over 2 yr"	5.3× (3.1× was the model-only term)
"imperfect routers capture 50–70%"	80% of oracle / 64.3% absolute
"EnCharge ~150 TOPS/W"	~24, precision undisclosed
"NVIDIA CPO 3.5×"	5×
"88.7% of queries routable"	correct
"18× intelligence-per-joule in 16 months"	correct

Unclosed gap, flagged: copper reach in metres at 224G/lane, and the copper-vs-optical pJ/bit crossover — five attempts, no primary source. Any crossover-reach constant in **`photonics/`** cost models is therefore unsourced and must be labelled as such.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/analog-photonic-codesign-synthesis