The last hardware flank: analog prefill survives only as an energy-bound niche

W. Schulz

Jul 30, 2026

Abstract

Five pre-registered kill tests on analog in-memory prefill with optically disaggregated decode. Shipping the KV cache costs under 0.1% of prefill energy, demoting the optical claim; the cache survives 5 bits but collapses at 4. Against measured chips, digital wins throughput per area, leaving an energy-bound, short-prompt niche.

By the end of July 2026 this project's hardware program had closed almost everything. Optical decode-compute was dead. Analog in-memory computing survived in one honest lane — small edge models under the window-policy rule — and optical interconnect survived only as a bandwidth-and-reach claim about training and prefill. One flank was still open: prefill itself, the compute-bound pass that turns a prompt into a key-value cache before generation begins.

It was worth closing properly, because it is the one system where both surviving hardware claims would be load-bearing at once: analog compute doing the prefill, optical links shipping the produced cache to a separate decode fleet. Prefill/decode disaggregation is already an established serving design (DistServe, for example, disaggregates the two phases onto separate GPUs), so this is a product hypothesis rather than a thought experiment.

Five kill tests were pre-registered with expected outcomes written down first. All five are now closed, and the composite verdict is narrower and more useful than the hypothesis that started it.

What the tests found

Long-prompt dilution behaves exactly as predicted. The energy win against a digital baseline is 19.1× at 512 tokens, 9.3× at 2048, 5.9× at 4096, 3.6× at 8192, and 1.7× at 32k — because causal attention is dynamic, grows quadratically, and cannot be put on a crossbar. The niche is short-to-mid prompts, and the kill threshold was never approached.

The disaggregation tax was refuted, and the pre-registration was wrong by about three orders of magnitude. Shipping the produced cache to the decode fleet costs 0.43 mJ against 459 mJ of prefill compute at 2048 tokens — under a tenth of a percent at every prompt length. Prefill/decode disaggregation costs essentially nothing in energy. The consequence lands on this project's own romantic version of the story: the composition claim is demoted. Optical links save 0.3 mJ of a 460 mJ job, which is energy-irrelevant. Optical interconnect's relevance to prefill survives only at the bandwidth, reach, and concurrency level — its original claim, and no more.

The quality gate cleared, positively, on a real checkpoint. The deferred question was whether a key-value cache produced on quantized analog weights is usable by a clean digital decoder. At every bit-width

where the quantized model works at all, prefill-only quantization matches fully-quantized behaviour — 0.722 against 0.722 at eight bits, 0.778 against 0.722 at six, 1.000 against 1.000 at five. The cache is portable, and the binding constraint on the product is not cache corruption. What binds instead is a sharp precision cliff between five and four bits, where accuracy falls from 1.000 to 0.111 under naive quantization — which independently locates, on this project's own substrate, exactly the boundary that hardware-aware retraining exists to move.

That result also required withdrawing an earlier verdict. A first attempt pre-registered four bits, got zeros everywhere, and its kill fired — but the fully-analog control was itself destroyed, and a physically implausible ordering appeared in the numbers. There was no working analog baseline, so the run could not answer the question, and its kill is withdrawn and logged as an instrument failure rather than a result.

And the throughput test, which had been withdrawn as uncalibrated, is now closed against fabricated silicon. The original version swung entirely on a guessed tile area: with the guess, analog won by about 20×; with a published per-core figure, digital won by about 2×. A verdict that inverts on an unmeasured constant is not a verdict, so no claim was made at the time.

Closing it against measured chips settles it. The best fabricated multi-core analog part we found, IBM's 64-core phase-change chip, delivers 0.40 TOPS/mm² in its high-precision (four-phase) read mode and 1.55 in its low-precision (one-phase) mode; its single core, without fabric, 1.59; an older-node analog flash part about 0.07; and a matched-node digital estimate about 1.0, against this cost model's array constant of 4.0. At the high-precision operating point, digital wins throughput per unit area by roughly 2.5× at matched granularity, up to 10× comparing chip against array, and about 58× against the older-node analog part. At the chip's low-precision mode the comparison would roughly level against the digital estimate, so the area verdict is conditional on the precision the analog part is run at. The guessed tile had implied 41 TOPS/mm², which is about 26× above the best fabricated core ever measured — that guess, not physics, produced the earlier inversion.

The verdict

Structurally, nothing changed when the guess was replaced by measurements, which is the reassuring part. The analog case remains a joules-per-token case only, and the underlying dichotomy stands with measured constants attached: digital wins silicon-bound fleets by roughly 22× on weight density, analog wins power-bound fleets by roughly 9× at the mid-length operating point.

> The prefill flank narrows but survives: an energy-bound, short-to-mid-prompt, five-bit-or-retrained, disaggregated-prefill niche that exists if and only if the deployment is capped by megawatts rather than by wafer supply.

That is the last open hardware flank, and it is now bounded on all five axes. It is also a smaller claim than the one the flank was opened to test — the romantic version, in which two surviving hardware advantages compose into a system story, was killed by its own energy arithmetic. In this program that is usually the sign that what remains is real.

The companion design paper was revised the same day with every one of these inputs closed: the prompt-length bound, the calibrated silicon numbers, and a falsifier that had threatened the window rule and is now defused. It is ship-ready as a defensive publication, with only the venue decision open.

Sources

The five kill tests, the energy figures and the cost-model constants are this project's own run-once results (`codesign-analog-moe/run_prefill_flank.py` and `docs/DIRECTION-prefill-flank.md` in the Quantum-Intelligence repository). The fabricated-chip constants come from:

1. Le Gallo, M., Khaddam-Aljameh, R., Stanisavljevic, M., et al. (2023). "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference." *Nature Electronics* 6: 680–693. doi:10.1038/s41928-023-01010-1 · arXiv:2212.02872 — MVM area efficiency 1.55 TOPS/mm² (one-phase) and 0.40 TOPS/mm² (four-phase), 8-bit input/output.
2. Khaddam-Aljameh, R., Stanisavljevic, M., Fornt Mas, J., et al. (2022). "HERMES-Core—A 1.59-TOPS/mm² PCM on 14-nm CMOS In-Memory Compute Core Using 300-ps/LSB Linearized CCO-Based ADCs." *IEEE Journal of Solid-State Circuits* 57: 1027–1038. doi:10.1109/JSSC.2022.3140414
3. Mythic M1076 analog matrix processor: 25 TOPS on a 40 nm process (vendor launch materials, 2021); die area of about 360 mm² as reported by *All About Circuits*. The ~0.07 TOPS/mm² figure is derived from these and is approximate.
4. Zhong, Y., Liu, S., Chen, J., et al. (2024). "DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving." *OSDI 2024*. arXiv:2401.09670

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