Where optics-native inference actually pays: a cost-model study

W. Schulz

Jun 27, 2026

Abstract

Speed-of-light inference is real but narrow and conditional: an optics-native state-space model beats a GPU on batch-1 decode latency only inside a tight box — WDM microrings, small-to-mid width, ~5-bit precision, error-tolerant workloads. The model reproduces a published 2025 chip's throughput exactly; the defensible contribution is the cost model itself.

A digital-twin / co-design study scoring architectures for end-to-end optical inference latency, energy, and precision, with the four structural bottlenecks as first-class terms. Synthesizes `EXPERIMENT_LOG.md` (Phase 1) and `EXPERIMENT_LOG_2.md` (Phase 2). Model = `photonic/` package (174 tests). All numbers reproducible from `experiments/`.

Abstract

"Speed-of-light inference" is real but narrow and conditional. An optics-native state-space model (S4-SSM) beats a GPU on batch-1 decode latency, but only inside a tight box: WDM microring hardware, small-to-mid width, ~5-bit precision, error-tolerant (classification) workloads, against a fused GPU. Outside it — coherent meshes at today's sizes, prefill, high precision, transformers, large width — optics loses. The model reproduces a published 2025 chip's throughput exactly and explains its device-vs-system energy gap, so the bounds are trustworthy. The defensible contribution is the cost model itself.

The four structural walls (the falsification checklist)

Any optics-as-compute claim must survive all four:

1. Tiling — an L×L mesh tiling an N×N weight pays $(N/L)^2$ electronic accumulation passes. This cancels the O(N) "speed of light" advantage, leaving a constant factor set by N/L.
2. Precision — analog dot products carry a $\sqrt{D}$ dimension penalty; real meshes deliver ~4–5 effective bits. Every demand $\geq$6 bits is below the analog cliff.
3. Cascadability — optical signal degrades per layer; depth before forced regeneration is precision-dependent ($\approx$3–15 layers at 10% tolerance, 0 at the 1e-4 cliff).
4. Nonlinearity / accumulation — staying optical needs both a cascadable optical nonlinearity (unsolved) and coherent optical accumulation (capped by phase-locking).

Method

A per-op cost model (`matmul`, `nonlinear`, `softmax`, `conv`, `reduce`) over a `HardwareConfig`, returning latency, energy, O-E-O crossings, and a graceful precision verdict. Architectures (transformer / SSM / linear-attention, prefill + decode) are scored against a roofline GPU baseline. Metric: batch-1 per-token latency (the "speed of light" north star).

Key results (Phase-1 headline → Phase-2 corrected)

Claim	Phase-1	Phase-2 correction	Final
SSM decode speedup	15–56×	fused GPU removes ~5× launch overhead	single-digit×
Best hardware path	coherent MZI	WDM helps interface, but small-mesh weight-reload dominates (residency)	coherent L≥64 (weights resident)
Coherent accumulation lever	23×	rode a hidden free-fan-out assumption	~2×
WDM contribution	(uncredited bug)	`tiling.cost` ignored `n_wavelengths`	~`w`× , bit-affordable
Full 130M model vs current GPU	(per-block only)	end-to-end, fused B200, residency-correct	single-digit×, residency-gated
Niche viability	asserted	tested: 5-bit argmax intact	real for classification

Validation (credibility anchor)

Against *Complex-valued MVM using a scalable coherent photonic processor* (Science Advances 2025, 16×16 mesh, 40 Gbaud, 6-bit): - Throughput: model $2 \cdot L \cdot \text{baud}$ = 1.28 TOPS = reported 1.28 TOPS (1.0×). - Energy: measured 340 fJ/MAC system-level = 170× the 2 fJ device-only constant — independently confirming that "interface-tax" energy claims are device-vs-system fragile.

The fence (where optics-as-compute pays)

IN: batch-1 decode · a mesh large enough to hold the weights resident (the binding requirement — see below) · width d≤~1024 · ~5-bit · error-tolerant argmax (RF modulation ID, keyword spotting) · well-thermally-stabilized mesh · vs a fused current GPU → single-digit× latency, not 50×. OUT: prefill/batched · a mesh too small to hold the weights (per-token thermal reload tax) · high precision/regression · transformers · large width · poorly-stabilized mesh → GPU wins.

The binding requirement: weight residency (corrected)

An early read favoured WDM microrings. The residency-correct model reverses this: a small mesh (e.g. L=32) cannot hold a real weight matrix, so it reprograms weights per token — a thermal phase-shifter reload (ms-scale) that dwarfs every interface optimisation, WDM included. WDM correctly relieves the converter wall (a real latency credit, now in `tiling.cost`), but it does nothing for weight reload. So the binding lever is a mesh big enough (L≥~64) to keep weights resident; the coherent L≥64 path wins (≈55× at d=256 vs a per-op GPU; single-digit× vs a fused GPU). WDM is a *secondary* interface optimisation, not the headline.

Why the money went to interconnect

Optical interconnect captures a comparable batch-1 latency gain at full precision, with none of the four walls (it moves data, doesn't transform it). Per unit of risk it dominates optics-as-compute. Our model

agrees with the market for model-internal reasons.

Honest open gaps

- Single-shot latency unvalidated (public chips report throughput).
- WDM credit and fan-out gains assume crosstalk-free lanes; modeled as a bit-budget discount, not from device crosstalk data.
- GPU energy efficiency held at H100 values (conservative for the GPU).
- Drift tax assumes mK-class stabilization; fast-drift meshes pay 1.4–42× throughput.

Conclusion

The contribution is this falsification machine, not a chip. It converts "speed of light inference" from a slogan into a bounded, validated, falsifiable claim: a ~2×, ~5-bit, classification-only, WDM-microring corner — useful, real, and far smaller than the headlines.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/optics-native-inference