Accepted at the 2027 Design, Automation and Test in Europe Conference (DATE 2027)
This is the authors' version submitted for peer review. The final version will appear in the DATE 2027 proceedings.

Glass-Photonic Flattened-Butterfly: A Switchless Optical Scale-Up Fabric for MoE LLM Training

Seongwon Yoon and Shimeng Yu
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
gabrielyoon@gatech.edu, shimeng.yu@ece.gatech.edu

***Abstract*—Mixture-of-Experts (MoE) training is dominated by the expert all-to-all (A2A). An electrical scale-up domain serves it only within 8–72 GPUs, and an expert-parallel (EP) group that exceeds the domain spills onto scale-out links at a fraction of the bandwidth and several times the energy per byte. We present a glass-photonic Flattened-Butterfly (FB) scale-up node of sixteen GPUs on a glass substrate. Adjacent hops use the electrical redistribution layer, longer hops use laser-written glass waveguides at depths set by distance, and one MTP-16 connector per GPU carries 800GB/s between panels, which tiles panels into a switchless optical fabric. In packet-level simulation against calibrated NVLink domains, Glass-FB trains at the speed of an NVL72-class domain for every expert group the domain can hold, at 2.4–3.1× less interconnect energy per iteration and 2.8× less static interconnect power. It trains 1.7–3.6× faster than an 8-GPU HGX server. Beyond the domain boundary it is 1.24× faster than NVL72 at 3.4× less energy, and 1.9× faster than HGX-8.**

***Index Terms*—optical fabric, glass substrate, glass photonics, co-packaged optics, flattened butterfly, GPU scale-up domain, MoE LLM, all-to-all**

I. Introduction

Frontier language models increasingly adopt the Mixture-of-Experts (MoE) architecture, and training them with expert parallelism (EP) makes the token-routing all-to-all (A2A) the dominant collective, which production measurements put at 33–55% of a training iteration [1] (Fig. 1). Today it runs on an electrical scale-up domain, such as the eight NVLink-connected GPUs of an HGX server or the 72 GPUs of a GB200 NVL72 rack. An expert group larger than the domain must send its A2A through scale-out NICs, at a fraction of the bandwidth and several times the energy per byte. Keeping larger expert groups on scale-up links calls for a topology that reaches many GPUs in few hops without a switch tier.

Kim and Dally at Stanford, with Abts at Cray, proposed the Flattened Butterfly (FB) in 2007 [2], and the topology was carried to on-chip networks the same year [3]. Every node links directly to its whole row and column, so no two nodes are more than two hops apart, and on balanced traffic FB costs about half of a Clos network of the same performance [2]. Its cost sits in the long row and column links, and in 2007 optical cables were too expensive to carry them, so that work priced the network with electrical cables and repeaters. Packaging has since changed those terms. Laser-written waveguides at multiple depths [4], [5], embedded photonic ICs [6] and co-packaged edge connectors [7], [8] are reported in the literature or already shipping, and on a glass substrate they carry a long link at optical bandwidth across a panel. We propose an FB scale-up node of 16 GPUs built from these parts. Inside the panel the fabric is a diameter-2 switchless butterfly, and across panels each GPU has one MTP-16 optical port at 800GB/s. Against an NVL72-class domain simulated at the same packet level, Glass-FB trains at the domain's speed for every EP it can hold, at 2.4–3.1× less interconnect energy per iteration, because a laser and ring-tuning budget of 2.8× less power replaces the switch tier.

Our contributions are (i) the glass-photonic FB node design with its waveguide, power and thermal budgets (Sec. III); (ii) EP-aware placement and cabling that keep the expert A2A inside a panel where it fits (Sec. IV); and (iii) a packet-level evaluation against HGX-8 and NVL72-class domains with a calibrated NVSwitch model and a per-tier energy accounting (Sec. V).

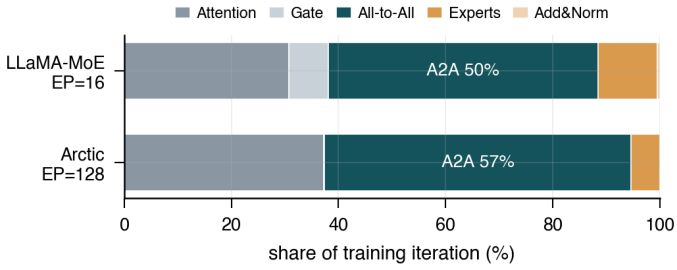
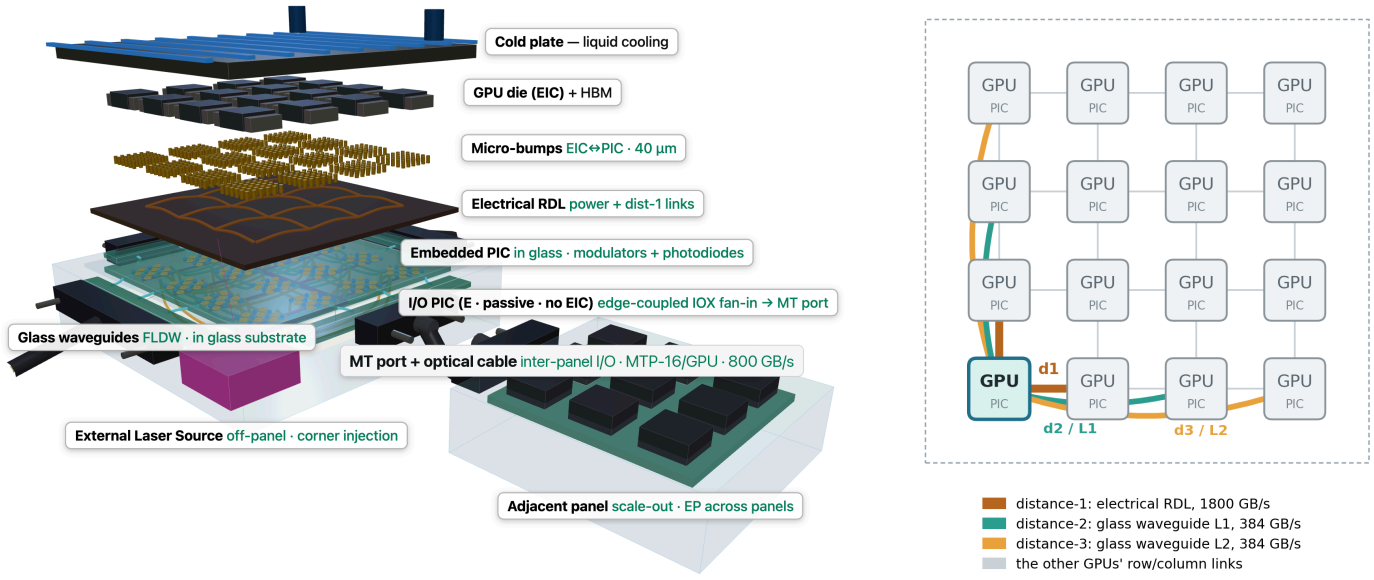


Fig. 1: MoE-layer phase breakdown of the training iteration. Expert all-to-all takes 50–57% of the iteration in our HGX-8 runs.

II. Background and Motivation

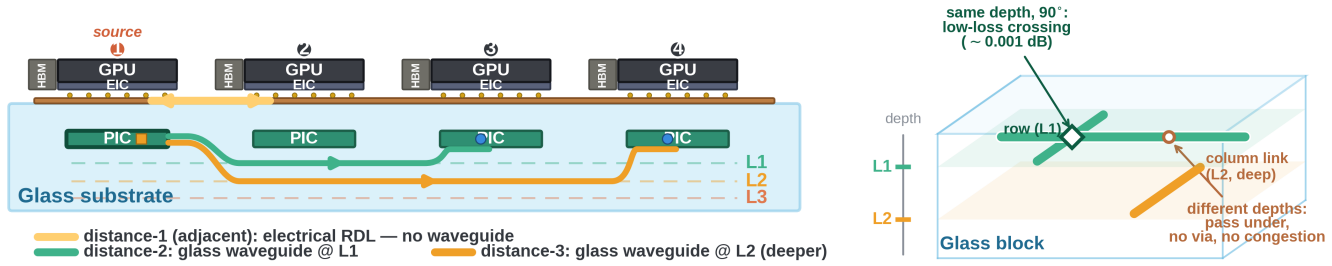
Energy consumption of AI data centers. Data centers used about 415 TWh of electricity in 2024, 1.5% of global consumption, and their use is projected to double to about 945 TWh by 2030 [9]. Most of that power goes to the accelerators rather than to the fabric between them, and yet the fabric governs a large share of what a training run costs. The expert A2A occupies 33–55% of a training iteration [1], and for all of that time the accelerators wait while the whole machine keeps drawing power, so time saved in the fabric is energy saved across every GPU in the job. The fabric's own draw is not small at rack scale either, since driving the NVLink spine of a rack with pluggable optics would take 20 kW of a 120 kW rack budget [10]. An



(a) Glass-photonic FB node.

(b) 4×4 FB topology (top view).

Fig. 2: Glass-photonic FB node system architecture. (a) A GPU tile on a glass substrate. Beneath the die an embedded PIC couples to FLDW waveguides routed at depths set by hop distance, fed by an external laser, with switchless edge ports for inter-panel I/O. The electrical stack is drawn exploded upward. (b) Every GPU links directly to its whole row and column (diameter 2), with distance-1 links electrical and distance-≥ 2 links optical.



(a) Distance-layered paths across four tiles (side view).

(b) Distance-layered 3D routing.

Fig. 3: Distance-layered glass waveguides and coupling in the glass-photonic FB node. (a) Adjacent tiles use the surface RDL. Distance-2 hops edge-couple into a shallow waveguide (L1) and distance-3 hops into a deeper one (L2), and both couple evanescently up into the destination PIC. (b) A row link (L1) and a column link (L2) cross in-plane at different depths, with no via.

interconnect that carries the same collective sooner and at fewer picojoules per bit therefore lowers both terms.

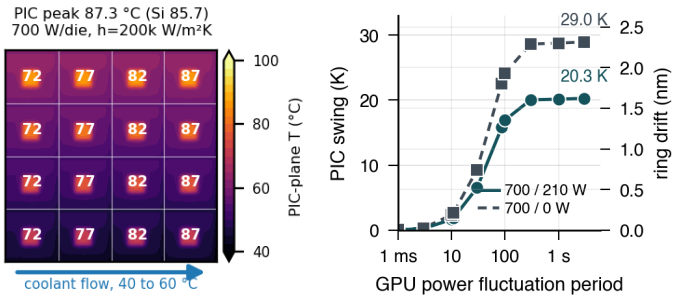
Scale-up bandwidth. An MoE layer sends each token to its top-k experts and back through two A2As, so every activated token crosses the EP group twice per layer and needs scale-up bandwidth. A scale-up domain serves more than 12.8 Tb/s per GPU at 100–250 ns of latency, while a scale-out network serves about 1.6 Tb/s per GPU at microsecond latency [10]. NVLink 5 gives each GPU of a GB200 NVL72 rack 1.8 TB/s within a 72-GPU domain [11], and widening such a domain adds switch trays and electrical links whose power grows with it.

Glass substrates and panel-scale nodes. Glass is emerging as the substrate for packages that outgrow organic laminates, and glass core substrates have been announced for next-generation data-center packages [12], [13]. Glass also guides light. Femtosecond-laser direct writing (FLDW) makes low-loss 3D waveguides [4], [5] (0.04 dB/cm at 1550 nm [14]) at multiple depths, panel-scale photonic interconnects embed photonic ICs in the glass itself [6], and photonic interconnects in the glass core are being developed for AI data-center packages [15]. A glass-core package that carries XPU and HBM chiplets beside co-packaged optics has been co-optimized for warpage, through-glass power delivery and XPU-to-optics signaling by the same substrate maker [13], [16]. A glass panel can therefore carry a whole scale-up node, the GPUs and their optical fabric together, and on such a panel FB's link lengths map onto waveguide depth, with short hops shallow and long hops deep.

III. Glass-Photonic FB Node Design

A. Physical stack and topology

Fig. 2 shows the node. A GPU compute die and its HBM sit on the glass substrate under a top-side liquid cold plate.



(a) PIC-plane heatmap. (b) PIC temperature swing.

Fig. 4: Thermal feasibility of the glass-photonic FB node (Ansys thermal). (a) PIC-plane temperature heatmap at the operating point of Table I. The PIC tracks the die within 0.13°C, and peaks rise along the coolant axis to 87.3°C. (b) Peak-to-peak PIC temperature swing against the period of a 50%-duty GPU power square wave.

The die carries the electronic IC (EIC) with its SerDes, drivers and TIAs, and it couples down through the redistribution layer (RDL) and ~40 μm micro-bumps into a photonic IC (PIC) of modulators and photodiodes embedded in the glass, a thin slab inset so it faces the FLDW waveguides just below (Fig. 3a). The RDL also carries power and the distance-1 links. Embedding only the PIC keeps the temperature-sensitive micro-rings off the hot EIC, whose hybrid bonding cuts ring-heater efficiency by 42–62% [17]. The panel is a 4×4 grid of 16 GPUs. Each GPU has a direct FB link to every peer in its row and column, so the panel has a two-hop diameter and holds a 16-way expert group on two optical layers.

B. Distance-layered waveguides and coupling

Distance-1 neighbors use the electrical RDL (Fig. 3), and longer hops use FLDW waveguides at a depth set by distance, L1 (~50 μm) for distance-2 and L2 (~300 μm) for distance-3. Light couples into the receiver PIC evanescently from below (0.16–0.4 dB measured [7], [18]), and the worst intra-panel path, distance-3 at about 9 cm, stays within 0.8 dB. Each optical hop costs about 40 ns over a wide, low-rate WDM link [19] with no FEC or retiming, against about 150 ns for retimed copper [20].

C. External laser, edge I/O, and budgets

An external laser injected at a panel corner supplies the carriers. Keeping the laser off the panel holds the heat of the least reliable component away from the glass and makes that component replaceable on its own. Each panel edge carries a switchless I/O PIC, an ion-exchange fan-out and four MTP-16 ports. One MTP-16 (8 TX + 8 RX fibers) carries 800GB/s per direction with four 200G wavelengths per fiber, the 800GBASE-FR4 class of IEEE 802.3dj [21] shipping in Tomahawk 6 [22]. Every GPU hosts one port on its nearest edge, so a panel presents 16 ports and 800GB/s of switchless egress per GPU. The 16 ports are cabled according to a port map. The map connects each panel to the panels its job communicates with, and it is generated from the same (dp, pp, ep) layout that drives

TABLE I: Thermal model of the glass panel node.

Layer (bottom up)	t (mm)	k (W/mK)	ρ (kg/m^3)	c_p (J/kgK)
Board	0.50	1.0	1900	1200
Glass substrate	0.70	1.2	2500	800
PIC (embedded)	0.05	80	2330	700
RDL	0.10	200 / 60‡	8900	385
GPU die	0.75	150	2330	700

Item	Value
Panel	4×4 tiles, 28.3 mm at 30.3 mm pitch
GPU die	700 W, half in a central hotspot of 1/9 of the area
PIC	15.5 / 37.3 / 59.1 W at interior / edge / corner tiles
Cold plate	die back, h=200,000 W/m^{2}K, 40 to 60 °C

‡in-plane / through-plane.

placement, so every cross-panel path is one hop. A fixed four-edge mesh of the panel grid is the job-independent alternative.

Waveguide budget. One waveguide carries $32\,\lambda \times 32Gb/s = 128GB/s$ per direction, and a tile at panel-scale bump density [6] fields about 60 waveguides per GPU in both directions. Only peers at distance ≥ 2 use them, so a corner GPU has optical degree 4. With n waveguides per direction per optical link and 6.25 for the port, $2 \cdot 4n + 2 \cdot 6.25 \leq 60$ gives $n \leq 5$, and we provision n=3 (384GB/s). A corner GPU then lights 36.5 of its 60 waveguides, while egress concentrated on four gateway GPUs would need 50 each, so one port per GPU is required.

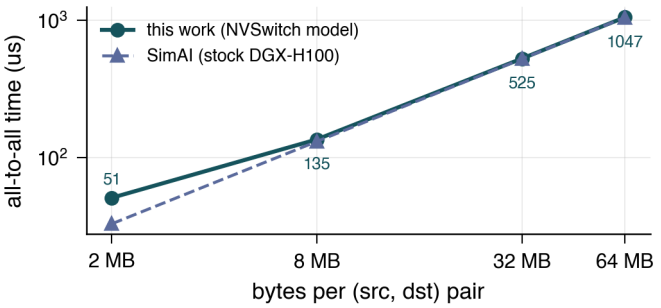
D. Thermal feasibility

We solve a full-package Ansys thermal model of the panel (Table I) at the cold-plate heat-transfer coefficient at which the GPU die meets its 85–90°C junction specification. Across all 16 tiles, the peak temperature of the embedded PIC is within 0.13°C of the die above it at every heat-transfer coefficient solved (Fig. 4a), so the PIC adds no thermal constraint beyond the one the GPU already sets. For the optical link, the relevant quantity is how far the PIC temperature swings as the GPU load changes, because microring resonances drift with temperature (Fig. 4b). A step from idle to 700 W raises the PIC by 28.8 K with a 44 ms rise time. A training job applies a periodic load instead, and the package stack filters it. With the GPU alternating between 700 and 210 W at 50% duty, one microbatch period (10.8 ms) swings the PIC by 1.8 K and one iteration period (86.8 ms) by 15.8 K. That swing moves an untuned silicon ring by about 1.3 nm at 12 Hz, three orders of magnitude slower than a kHz tuning loop [23].

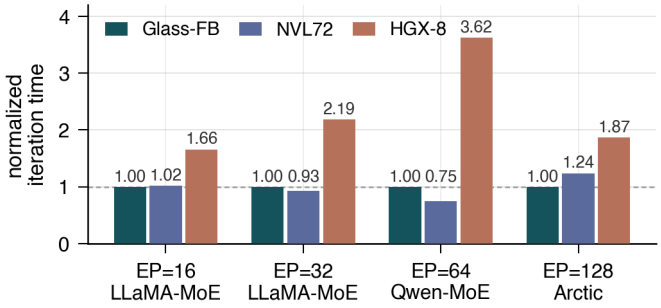
IV. EP-Aware Panel Placement and Routing

A job numbers its GPUs by logical rank, and the default order interleaves the parallel groups. Placed on panels in that order, one expert group can be spread over several panels, and its A2A then runs on the inter-panel ports. EP-aware placement renumbers the ranks so that the members of each expert group are consecutive and fill panels in order. With t and e the tensor- and expert-parallel degrees, a logical rank is $n = \mathrm{hi} \cdot (e\,t) + \mathrm{ep} \cdot t + \mathrm{tp}$, where hi indexes the data- and pipeline-parallel copy. The placement ϕ moves the expert index to the innermost position,

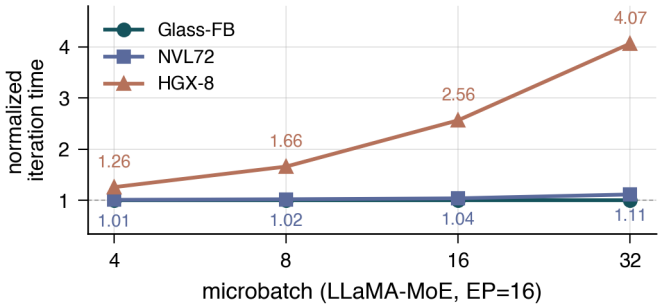
$$\phi(n) = \big(\underbrace{\lfloor n/(e\,t)\rfloor}_{\text{hi: DP/PP}} \cdot t + \underbrace{n \bmod t}_{\text{tp: TP}}\big)\, e + \underbrace{\lfloor n/t\rfloor \bmod e}_{\text{ep: EP}},$$



(a) NVSwitch model validation.



(b) Training iteration time of the three fabrics at each EP.



(c) Iteration time against microbatch size.

Fig. 5: (a) The NVSwitch model running an 8-GPU HGX H100 under a synthetic all-to-all of 2–64 MB per pair, against SimAI's stock DGX-H100 model [24]. (b) Training iteration time at EP 16, 32, 64 and 128, normalized to Glass-FB within each EP, where it takes 87.7, 75.3, 39.9 and 194.6 ms. (c) Iteration time against microbatch size at EP=16 (LLaMA-MoE), normalized to Glass-FB at each microbatch.

so consecutive physical ranks belong to one expert group, the group's A2A runs on intra-panel links, and with $\mathrm{EP} \cdot \mathrm{TP} = 16$ each copy fills exactly one panel.

Routing. A flow to another panel leaves through the GPU that holds a port to that panel. A flow between two GPUs of one panel that share no row or column is relayed through one intermediate GPU, along the row first and then along the column.

V. Evaluation

A. Methodology

Simulation environment. We combine packet-level simulation with real-GPU profiling to evaluate Glass-FB. Operator times are profiled on H100 GPUs, and every fabric is simulated in htsim [25], a packet-level, congestion-aware simulator. As is standard for new scale-up fabrics [1], [26], [27], we extend htsim with a Glass-FB class carrying the panel model, its link rates, placement and routing.

TABLE II: Training configurations (sequence length 1024, microbatch swept 4–32, four MoE layers per task graph). GPUs = DP × TP × PP × EP.

Model	Hidden	GPUs	DP	TP	PP	EP	top-k
LLaMA-MoE-3.5B	4096	128	2	1	4	16	2
LLaMA-MoE-3.5B	4096	256	2	1	4	32	2
Qwen1.5-MoE-A2.7B	2048	512	2	1	4	64	4
Arctic-480B	7168	1024	2	1	4	128	2

TABLE III: Fabrics compared (bandwidths per direction). An NVL72 rack hosts power-of-two parallelism groups, so 64 of its 72 GPUs form the domain.

Fabric	Domain (GPUs)	In-domain link (GB/s)	Cross-domain (GB/s per GPU)
Glass-FB (ours)	16	1800 elec., 384 opt.	800 (MTP-16)†
HGX-8	8	450 (18×25, NVLink4)	50 (400G NIC)
NVL72	64 of 72	900 (18×50, NVLink5)	100 (800G NIC)

†One MTP-16 port per GPU at 200G/lane.

Baselines and calibration. Table III defines the fabrics, which run the same task graphs on the same GPUs and differ only in the interconnect. NVL72 and HGX-8 are simulated as packet-level NVSwitch domains, each a single-stage Clos of 18 or 4 switch chips with one queued link per GPU and chip, 250 ns per hop, and every transfer striped across all chips. We validate the model on an 8-GPU HGX H100 under a synthetic all-to-all, where it agrees with SimAI's stock DGX-H100 model [24] within 3.5% above 8 MB (Fig. 5a).

Workloads and power. Task graphs come from FlexFlow [28] profiles of the models in Table II, with a Zipf expert-popularity skew. Every fabric pays a static term plus the hop-bytes it moves on each tier times that tier's pJ/bit, and a byte that crosses two links counts twice. Glass-FB's tiers are charged the 32 Gb/s-per-lane link energies of [6, Table 1], 1.15 for the electrical RDL tier (an XSR SerDes PHY) and 1.15 for the optical tiers (a 32-wavelength photonic link). NVLink is charged 5 and an 800G NIC 16, the 224G SerDes and scale-out figures of [10]. The static term for a panel is 42.9 W of laser and ring tuning, which covers the 7,808 carriers a panel lights when all of its ports carry traffic, at 1 mW per carrier behind a 25% wall-plug laser and 0.75 mW per ring. NVSwitch draws 7.5 W per GPU on NVL72 and 12.5 W per GPU on HGX-8, about 27 and 25 W per switch chip. These are deliberately low charges for a switch tier of this bandwidth, and a heavier one widens every energy gap we report.

B. Training iteration time across fabrics and EP sizes

Fig. 5b compares the training iteration time of the three fabrics at EP 16, 32, 64 and 128, normalized to Glass-FB at each EP. Up to EP 64 the expert group fits inside one NVL72 domain. On Glass-FB the same group fills one panel at EP 16 and spans two panels at EP 32 and four at EP 64.

HGX-8 is the slowest fabric at every EP. Its domain holds eight GPUs, so the expert A2A already leaves the server at

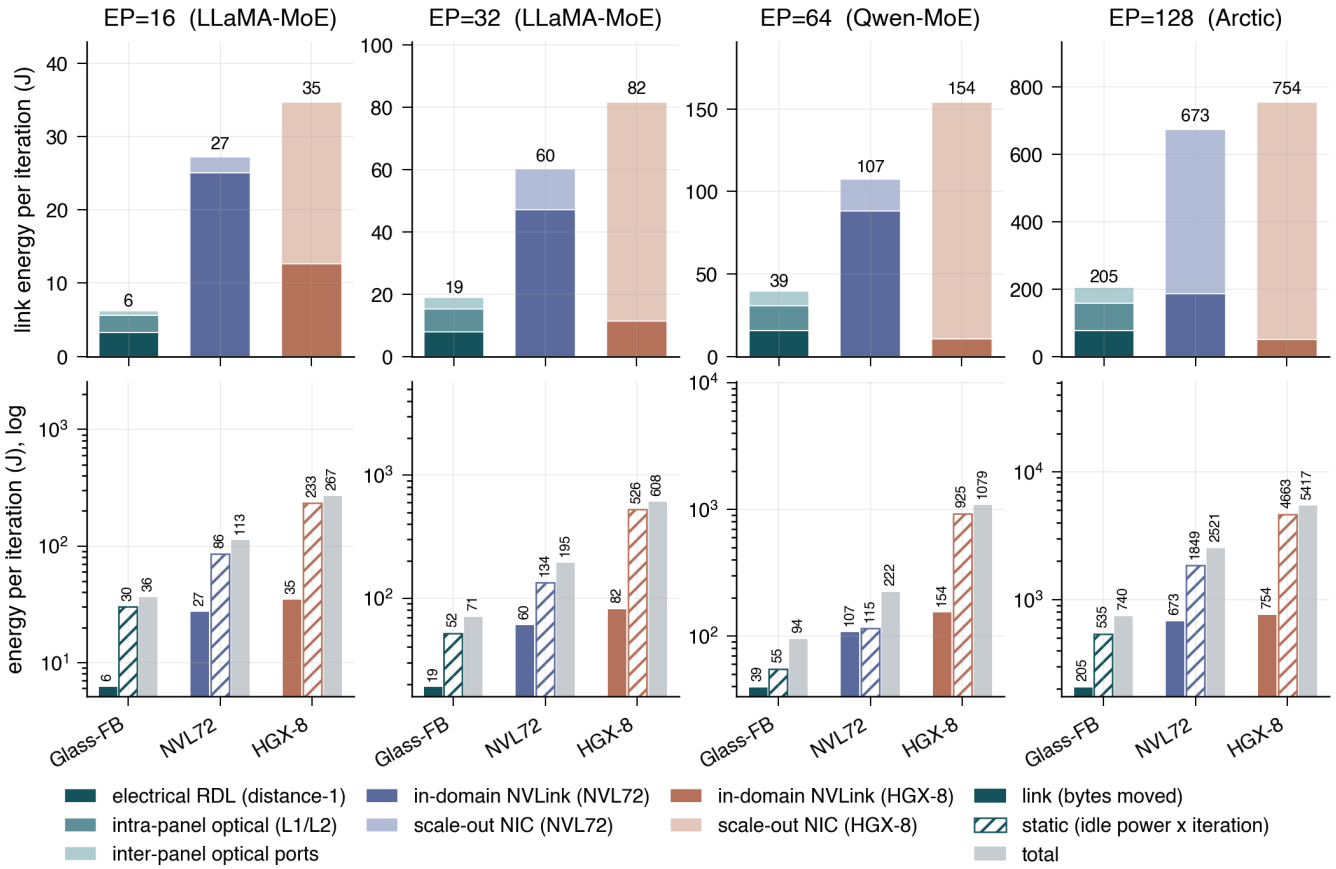


Fig. 6: Breakdown of interconnect energy per iteration. The top row is link energy on bytes moved, by tier. The bottom row, on a log scale, sets that link energy beside the static term (idle power over the iteration) and their total.

EP 16 and runs over 50 GB/s NICs. Glass-FB finishes 1.7, 2.2, 3.6 and 1.9× sooner at the four EPs.

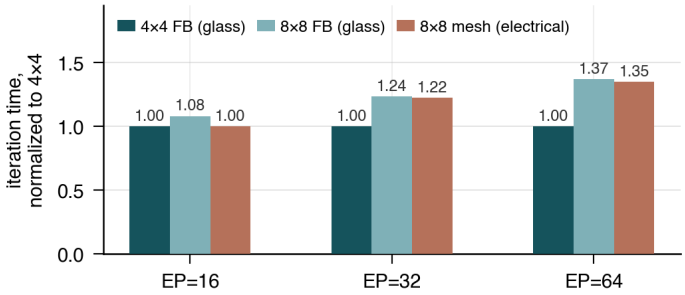
Up to EP 64 Glass-FB and NVL72 finish an iteration within 1.33× of each other. Glass-FB is 1.02× faster at EP 16. NVL72 is 1.08× faster at EP 32 and 1.33× faster at EP 64, and it spends 2.4–3.1× more interconnect energy to do so (Sec. V-D). What separates them is the A2A traffic that crosses panel boundaries. NVL72 keeps the whole group inside one domain, and its A2A takes about 8 ms per iteration at every EP. On Glass-FB the A2A grows from 5.6 ms at EP 16 to 17.4 ms at EP 64, as the share of its traffic that leaves a panel grows from 10% to 22%.

At EP 128 the expert group is larger than the scale-up domain of every fabric. Part of each A2A then crosses a domain boundary, and what decides the iteration is how much traffic crosses and at what bandwidth. HGX-8 sends 81% of its traffic across 50 GB/s NICs, and NVL72 sends 45% across 100 GB/s NICs. Glass-FB sends only 24% of its traffic off a panel, and that traffic leaves through 800 GB/s ports, eight times the per-GPU bandwidth of an NVL72 NIC. Glass-FB completes the iteration in 194.6 ms, against 240.7 ms on NVL72 and 364.3 ms on HGX-8.

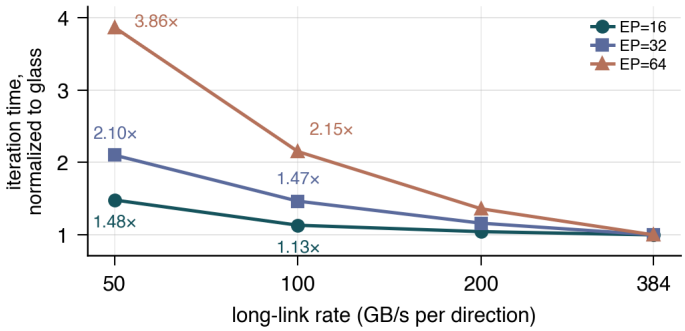
C. Panel size, substrate and microbatch size

Two limits set how many GPUs belong on one panel. One is the size of glass that can be processed today, and the other is the waveguide budget each GPU shares among its optical peers. Glass core substrates with through-glass vias are being scaled to 510×515 mm panels [13], and package substrates as large as 120×150 mm are shown today [29]. A 4×4 panel is 121 mm on a side, close to that package size, while an 8×8 panel would be 242 mm. We ran two 64-GPU panels on the same task graphs to see whether a larger panel would pay off in training time if it were built (Fig. 7a). An 8×8 flattened butterfly gives each GPU three times as many optical peers, and the same waveguide budget then leaves each optical link 128 GB/s instead of 384. An 8×8 electrical mesh keeps its 1800 GB/s links but reaches only the four nearest neighbors, so traffic to any other GPU is relayed hop by hop. Both panels are 22–37% slower than the 4×4 panel at EP 32 and 64. At EP 16 the mesh matches the 4×4 panel and the 8×8 butterfly is 8% slower. We did not simulate a 6×6 panel, because 36 GPUs do not divide the power-of-two parallelism groups and an expert group could not fill whole panels under the placement of Sec. IV.

A switchless wafer-scale system carries its non-adjacent links as long-reach electrical links [20]. We compare the 4×4 butterfly on the panel-scale glass substrate with the same butterfly whose distance-≥ 2 links are long-reach electrical instead. The electrical butterfly is 1.13, 1.47 and 2.15× slower at EP 16, 32 and 64 at the same buffer size (Fig. 7b). The penalty grows with EP because a larger group puts more of its A2A on the long links. The penalty comes from the bandwidth of those links rather than their latency, which alone costs 0.1–



(a) 64-GPU panels against the 16-GPU panel.



(b) Panel-scale glass substrate against wafer-scale integration.

Fig. 7: Effect of panel size and substrate on training iteration time. (a) An 8×8 glass flattened butterfly and an 8×8 electrical mesh, normalized to the 4×4 Glass-FB panel at each EP. (b) Iteration time against the rate of the distance-≥ 2 links, carried at the long-reach electrical latency and normalized to the glass design point at each EP.

2%. With the long links charged 16 pJ/bit [10] and no static term, the electrical butterfly uses 1.0, 1.6 and 2.5× Glass-FB's interconnect energy per iteration.

A larger microbatch sends more tokens through every A2A and loads the links harder. From microbatch 4 to 32 at EP=16, Glass-FB's iteration grows 1.16×, NVL72's 1.28× and HGX-8's 3.8× (Fig. 5c). Glass-FB therefore stays 1.01–1.11× ahead of NVL72, and its lead over HGX-8 grows from 1.26× to 4.1×. Cabling matters once an expert group spans panels. At EP=32, Glass-FB cabled with the port map of Sec. III-C is 2.6× faster than with a fixed four-edge mesh.

D. Interconnect energy per iteration across fabrics

Fig. 6 splits each fabric's interconnect energy per iteration into link energy on the bytes it moves and static energy from idle power. At EP 16, 32, 64 and 128 Glass-FB spends 36, 71, 94 and 740 J per iteration, against 113, 195, 222 and 2521 J on NVL72 and 267, 608, 1079 and 5417 J on HGX-8.

Glass-FB is ahead on the bytes it moves. At EP 16 to 64 NVL72's link energy is 2.7–4.4× Glass-FB's, because every Glass-FB tier costs 1.15 pJ/bit while NVL72 pays 5 pJ/bit on each of the two NVLink hops of a flow. Within Glass-FB the panel ports take a growing share of that link energy, from 10% at EP 16 to 24% at EP 128, as the expert group spans more panels. HGX-8 moves most of its bytes over NICs at 16 pJ/bit, and those NICs carry 64–93% of its link energy.

Idle power decides how wide the total gap becomes. It is 52–76% of NVL72's interconnect energy, 86–87% of HGX-8's and 58–83% of Glass-FB's. Per 16 GPUs a panel draws 42.9 W of laser and ring tuning, against 120 W of NVL72 switches and 200 W of HGX-8 switches, 2.8× and 4.7× less. Static energy is that power times the iteration time, so the ratio also follows which fabric finishes first. It narrows to 2.1× at EP 64, where NVL72 finishes 1.33× sooner, and widens to 3.5× at EP 128, where Glass-FB finishes 1.24× sooner. The total gap to NVL72 is therefore narrowest at EP 64, at 2.4×, and 3.1× and 2.8× at EP 16 and 32. At EP 128 NVL72 also pays 488 J on its NICs, 72% of its link energy, and the gap is 3.4×. Against HGX-8 the gap is 7.3–11.5×.

The widest per-byte gap is at the panel boundary. Glass-FB's 800GB/s port crosses it at 8.5 W, including the laser and ring tuning of its carriers, while NVL72's 100GB/s NIC costs 12.8 W. The port carries 8× the bandwidth at 12× the bandwidth per watt. Per token at LLaMA-MoE with EP=32, Glass-FB spends 0.54 mJ, NVL72 1.5 mJ and HGX-8 4.6 mJ.

VI. Related Work

The panel-scale photonic interconnect of Hsueh et al. [6] is the closest design to ours. It places chiplets and memory on a glass panel with embedded photonics and reconfigures them with optical crossbar switches, as data-center fabrics do with a switch tier [1], while we keep the embedded PIC and use a fixed switchless flattened butterfly. Each building block of our node has been reported on its own, including evanescent coupling, glass waveguides and multilayer packaging [7], [30]–[33], thermo-optic tuning and interposer thermals [34], [35], and laser-written 3D photonics [4], [5]. Inter-chiplet topologies for glass substrates [36] keep substrate links electrical. Wafer-scale systems [20], [26], [27], [37], expert placement on switched chiplets [38], [39] and Lightmatter's photonic interposer [10] are the closest precedents to our evaluation.

VII. Conclusion

We presented a glass-photonic Flattened-Butterfly scale-up node for MoE training. Sixteen GPUs sit on one glass panel, where distance-layered waveguides carry the long links, the electrical RDL carries the adjacent hops, and one MTP-16 connector per GPU tiles panels into a switchless optical fabric. In packet-level simulation Glass-FB stays within 1.33× of an NVL72-class domain for every expert group that domain can hold and uses 2.4–3.1× less interconnect energy per iteration. Beyond the boundary it is 1.24× faster than NVL72 and 1.9× faster than HGX-8, at 3.4× less energy. The saving comes from the structure rather than from a faster link. A panel replaces the switch tier with a laser and ring-tuning budget of 2.8× less power, and it crosses its boundary with 8× the bandwidth per GPU at 12× the bandwidth per watt. The 4×4 panel also trains faster than the 64-GPU panels we measured, and under the GPU's own cold plate the embedded PIC stays within 0.13°C of the die and swings by under 2 K per microbatch.

References

[1] X. Liao, Y. Sun, H. Tian, X. Wan, Y. Jin, Z. Wang *et al.*, "MixNet: A runtime reconfigurable optical-electrical fabric for distributed mixture-of-experts training," in *Proceedings of the ACM SIGCOMM 2025 Conference*, 2025, pp. 554–574.

[2] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: A cost-efficient topology for high-radix networks," in *Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA)*, 2007, pp. 126–137.

[3] J. Kim, J. Balfour, and W. J. Dally, "Flattened butterfly topology for on-chip networks," in *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2007, pp. 172–182.

[4] X. Han, Y. Wang, J. Hu, L. Zhong, and J. Qiu, "Laser printing of large-area conformal 3D photonic circuits in glass," *Laser & Photonics Reviews*, vol. 18, no. 9, p. 2400060, 2024.

[5] L. Zhong, Y. Wang, X. Han, J. Hu, and J. Qiu, "Laser direct lithography of large-area three-dimensional integrated photonics: Technological challenges and advances," *Chip*, vol. 4, no. 4, p. 100157, 2025.

[6] T.-C. Hsueh, B. Lin, Z. Chen, and Y. Fainman, "Panel-scale reconfigurable photonic interconnects for scalable AI computation," *IEEE Open J. Solid-State Circuits Soc.*, vol. 5, pp. 437–453, 2025.

[7] L. Brusberg, J. A. Holguín-Lerma, J. R. Rodrigues, B. J. Johnson, J. R. Grenier, M. Granados-Baez *et al.*, "High-density evanescent chip coupling with detachable fiber connector for co-packaged optics," in *Optical Fiber Communication Conference (OFC)*, 2025, p. Th3H.1.

[8] B. Fleury and A. Ramaswamy, "Optimized fiber connections and routing for co-packaged optics (CPO)," Market Focus presentation, European Conference on Optical Communication (ECOC), Sep. 2024.

[9] International Energy Agency, "Energy and AI," World Energy Outlook Special Report, Apr. 2025.

[10] M. Bernadskiy, P. Carson, T. Graham, T. Groves, H. J. Lee, and E. Yeh, "Accelerating frontier MoE training with 3D integrated optics," in *IEEE Symposium on High-Performance Interconnects (HOTI)*, 2025, pp. 25–35.

[11] NVIDIA, "GB200 NVL72," Product page, accessed Sep. 2026.

[12] Intel Corporation, "Intel unveils industry-leading glass substrates to meet demand for more powerful compute," Press release, Sep. 2023.

[13] Dai Nippon Printing, "DNP develops TGV glass core substrate for semiconductor packages," Press release, Mar. 2023.

[14] T. T. Fernandez, A. Ross-Adams, M. Bakovic, J. U. Thomas, O. Hochrein, H. Mahmodi *et al.*, "High fidelity waveguides through vitreous media recomposition for next generation photonic devices," *Advanced Optical Materials*, vol. 13, no. 20, 2025.

[15] M. R. H. Mojumder, Z. Guo, N. Li, and M. Swaminathan, "Photonic interconnects in glass core for AI data center applications," in *IEEE 76th Electronic Components and Technology Conference (ECTC)*, 2026, pp. 856–861.

[16] S. Kuramochi, T. Tai, K. Fujimoto, M. Tanaka, and T. Takano, "Optimization of glass core substrate design for energy-efficient high-density AI server packages," *SID Symposium Digest of Technical Papers*, vol. 57, pp. 661–664, 2026.

[17] D. Coenen, M. Kim, H. Oprins, X. Chang, F. Ferraro, Y. Ban, and J. Van Campenhout, "Thermal challenges for resonant Si photonic modulators in system-on-wafer applications," *IEEE J. Sel. Top. Quantum Electron.*, vol. 32, no. 2, pp. 1–8, 2026.

[18] P. Xu, J. He, K. Kennes, A. Patil, A. Dvoretskii, A. Podpod *et al.*, "Collective die-to-wafer bonding enabling low-loss evanescent coupling for optically interconnected system-on-wafer," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 16, no. 1, pp. 186–197, 2026.

[19] S. Daudlin, A. Rizzo, S. Lee, D. Khilwani, C. Ou, S. Wang *et al.*, "Three-dimensional photonic integration for ultra-low-energy, high-bandwidth interchip data links," *Nature Photonics*, vol. 19, no. 5, pp. 502–509, 2025.

[20] Y. Feng and K. Ma, "Switch-less dragonfly on wafers: A scalable interconnection architecture based on wafer-scale integration," in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2024, pp. 1–17.

[21] IEEE 802.3 Ethernet Working Group, "IEEE P802.3dj 200 Gb/s, 400 Gb/s, 800 Gb/s, and 1.6 Tb/s Ethernet task force," Draft standard, 2025.

[22] Broadcom Inc., "Broadcom announces Tomahawk 6 – Davisson, the industry's first 102.4-Tbps Ethernet switch with co-packaged optics," Press release, Oct. 2025.

[23] K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," *Nanophotonics*, vol. 3, no. 4–5, pp. 269–281, 2014.

[24] X. Wang, Q. Li, Y. Xu, G. Lu, D. Li, L. Chen *et al.*, "SimAI: Unifying architecture design and performance tuning for large-scale large language model training with scalability and precision," in *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2025, pp. 541–558.

[25] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley, "Improving datacenter performance and robustness with multipath TCP," in *Proceedings of the ACM SIGCOMM 2011 Conference*, 2011, pp. 266–277.

[26] S. Rashidi, W. Won, S. Srinivasan, P. Gupta, and T. Krishna, "FRED: A wafer-scale fabric for 3D parallel DNN training," in *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA)*, 2025, pp. 34–48.

[27] H. Wang, Z. Wang, H. Wang, J. Hou, T. Wei, C. Li *et al.*, "WATOS: Efficient LLM training strategies and architecture co-exploration for wafer-scale chip," in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2026, pp. 1–19.

[28] Z. Jia, M. Zaharia, and A. Aiken, "Beyond data and model parallelism for deep neural networks," in *Proceedings of Machine Learning and Systems (MLSys)*, 2019.

[29] TOPPAN Holdings, "TOPPAN and Tekscend Photomask participate in SEMICON Japan 2025," Press release, Dec. 2025.

[30] L. W. Yeary, L. Brusberg, and J. R. Grenier, "Integrated optical and electrical connectivity in advanced packaging: The role of glass substrates," in *IEEE Optical Interconnects and Packaging Conference (OIP)*, 2025, pp. 1–2.

[31] M. Letz, J.-U. Thomas, D. Hochfilzer, M. L. Haffner, J. Schwietering, and H. Schröder, "Glass panel based packaging: Optical waveguides for data transport in large packages," in *Optical Interconnects and Packaging 2026, Proc. SPIE 13903*, 2026.

[32] T. Amemiya, "Multilayer glass structure for advancing packaging and substrate technologies," *SID Symposium Digest of Technical Papers*, vol. 56, no. 1, pp. 1301–1303, 2025.

[33] X. Chen, B. Gao, Z. Sheng, X. Wu, H. He, and H. Xue, "Review of glass substrates for co-packaged optics: Fabrication and performance of integrated electro-optical structures," *Advanced Photonics Nexus*, vol. 5, no. 3, p. 034003, 2026.

[34] Z. Yang, S. B. N. Gourikutty, H. K. Cheemalamarri, A. Sundaram, Y. Sun, W. Luo *et al.*, "Low-power thermo-optic switching enabled by spiral waveguides and glass–silicon hybrid bonding," *IEEE Access*, vol. 13, pp. 209 829–209 837, 2025.

[35] P. Gupta, A. K. Mallik, K. Krohnert, S. Latkowski, C. Okonkwo, S. Alomari *et al.*, "Thermal and electrical study of glass interposers in co-packaged electronic–photonic systems," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 15, no. 8, pp. 1625–1635, 2025.

[36] P. Iff, M. Besta, and T. Hoefler, "FoldedHexaTorus: An inter-chiplet interconnect topology for chiplet-based systems using organic and glass substrates," arXiv:2504.19878, 2025.

[37] T. Jiang, S. Qi, Z. Wang, N. Kang, F. Yang, Y. Wang *et al.*, "Credit-guided congestion control on wafer-scale on-chip networks for molecular dynamics," in *Proceedings of the ACM SIGCOMM 2026 Conference*, 2026, pp. 1561–1575.

[38] X. Tang, J. Hou, D. Jiang, T. Wei, J. Liu, J. Deng *et al.*, "MoEntwine: Unleashing the potential of wafer-scale chips for large-scale expert parallel inference," in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2026, pp. 1–15.

[39] S. Luo, Y. Han, P. Li, J. Qin, J. Peng, Y. Zhao *et al.*, "Mozart: Modularized and efficient MoE training on 3.5D wafer-scale chiplet architectures," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 38, 2025, pp. 125 742–125 766.