turbalance

BERKELEY LAB

Breaking Through the Network Wall

Petabit-Class Switching for Single-Hop MoE Training at Scale

P. Berdesinski, L. Montigny · Turbalance

J. Shalf, G. Michelogiannakis · Lawrence Berkeley National Laboratory

IEEE Hot Interconnects 33 (HotI 2026) · August 2026

The Study in Three Parts

1

The model

A frontier-scale Mixture-of-Experts, scaled to thousands of experts and 20T parameters.

2

The fabrics

Two ways to connect it: a conventional multi-tier network vs a flat, single-hop using petabit switches.

3

The tests

Two simulated training runs: a pure network stress test and a realistic parallelism mix.

We investigate the impact of petabit switches on training performance for a large Mixture-of-Experts model.

The Network Wall

Compute has outscaled the network

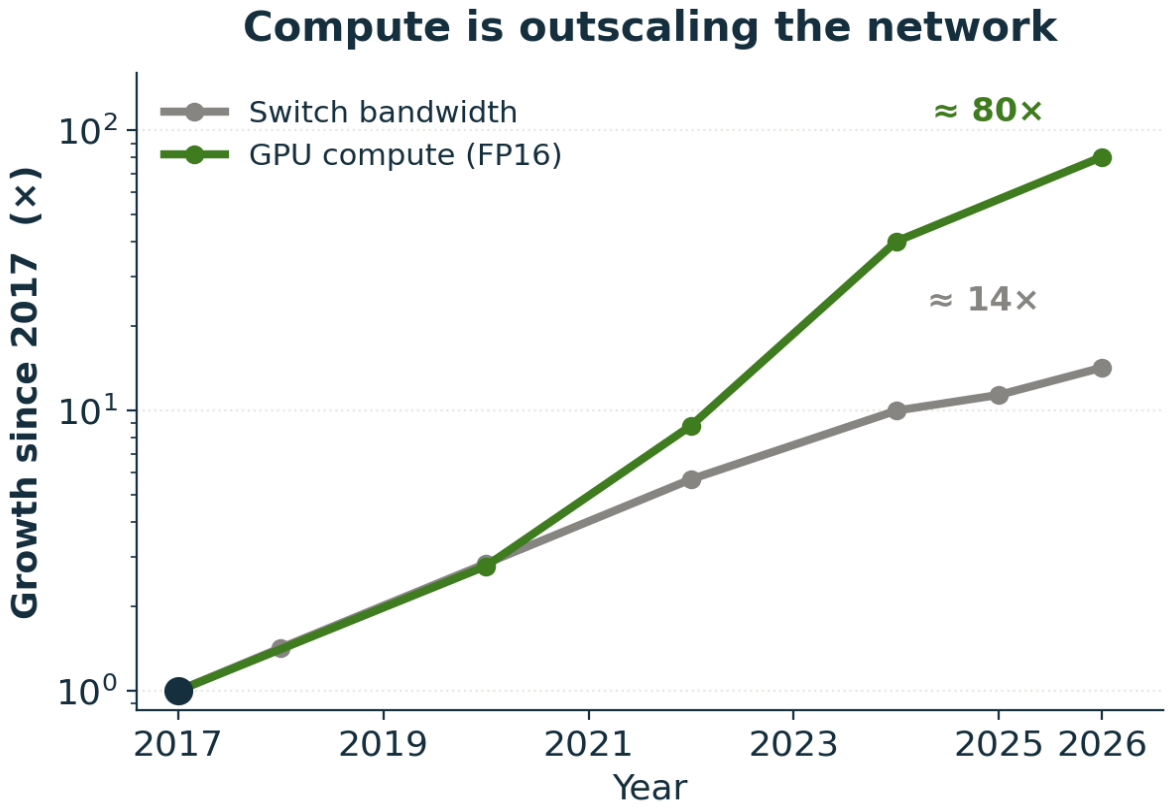
the interconnect, not the GPU, now limits large-model training.

Mixture-of-Experts sharpens the problem

expert routing turns every training step into an All-to-All.

Today, the network limits how fast large models train.

Can a switch with 10× the radix take the network off the critical path by collapsing the fabric?



Source: Broadcom Tomahawk and NVIDIA FP16 specifications; 2026 points projected.

Today's Switches Hit a Radix Wall

The leading edge & beyond

State of the Art: 512 × 200G lanes

Current (shipping from 2H2026) ASICs top out at 102.4 Tbps — Tomahawk 6, Silicon One G300, Spectrum-6.

Multi-ASIC box doesn’t change radix

NVIDIA's Spectrum-6 SN6800 reaches 409.6 Tbps by bundling four switch chips in a box with an internal shuffle — but the fabric still sees the same 512 lane per-chip radix.

Next Generation: only ~2× radix expected

Trend suggests next-generation ASICs will land at 204.8 or 230.4 Tbps.

Effects

Limits the size of single-hop domains

A radix of 512 caps a single-hop domain at a few hundred xPUs.

Multi-tier fabrics leads to poor scaling

Scaling beyond those limits requires non-blocking CLOS trees, meaning more switches, cabling, and power, and higher tail latency.

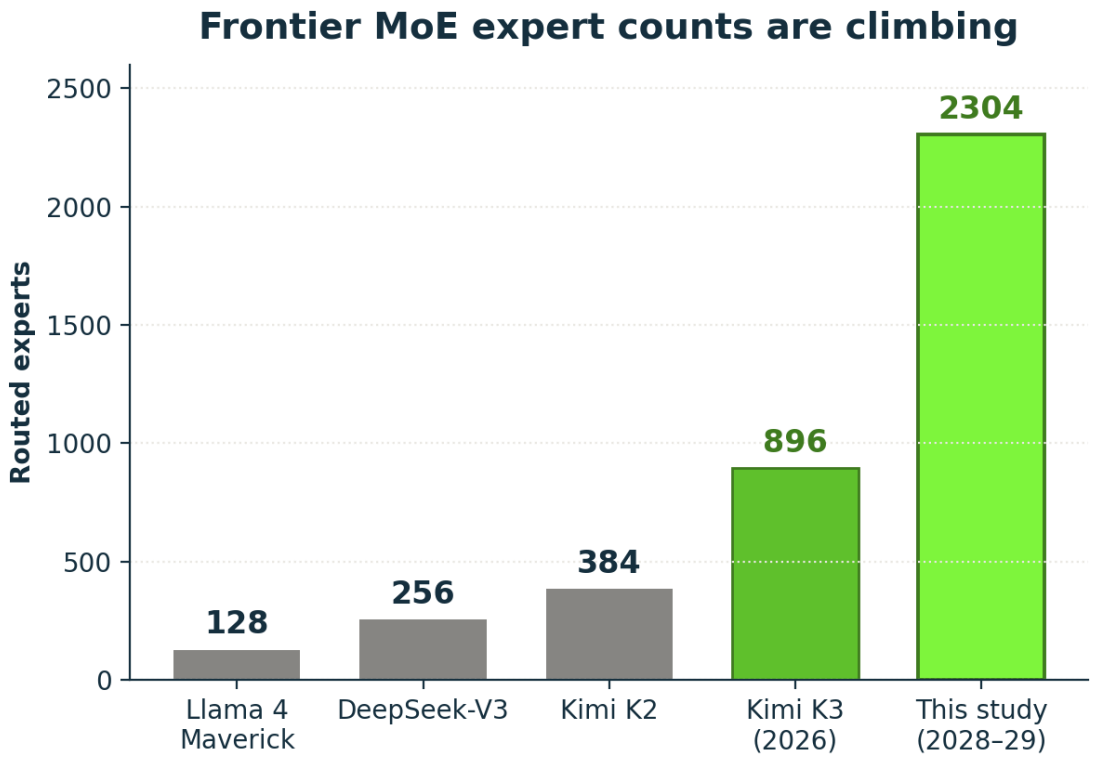
We evaluate at 10× the radix: A 1-Pbps switch

5,120 × 200G lanes on one ASIC — 10× today's radix at XDR / Ultra Ethernet lane rates — deliver **≈ 1 Pbps** of capacity: 5,120 GPU packages in a single tier, every pair **1 hop** apart with full bisection bandwidth and uniform latency.

Methodology — TurbaSim

- **Discrete-event simulator —** for large-scale AI-training interconnects; its flow-level backend models point-to-point traffic for every collective.
- **Trace-driven —** it ingests Chakra execution traces and replays Collectives and send/receive operations.
- **Multiple backends —** packet, flow or analytical — trading fidelity for speed, and built to model novel high-radix topologies at scale.
- **Apples-to-apples —** it replays one trace on each fabric, so every measured difference is the network.
- **Hardware- and topology-agnostic —** the same engine models NVIDIA, AMD or custom accelerators across scale-up, scale-out and scale-across fabrics.

The Workload — a 20T 2,304-Expert MoE



Expert counts are climbing

Kimi K3 (2026) already routes 16 of 896 experts — and SemiAnalysis reports Moonshot recommends serving it on a scale-up world of at least 64 accelerators. As models add experts, the interconnects matter more.

So we test the projected frontier

We deliberately pick 2,304 experts (16 active) — a ~20T / ~165B-active model derived from Llama 4 Maverick — projecting to 2028–29.

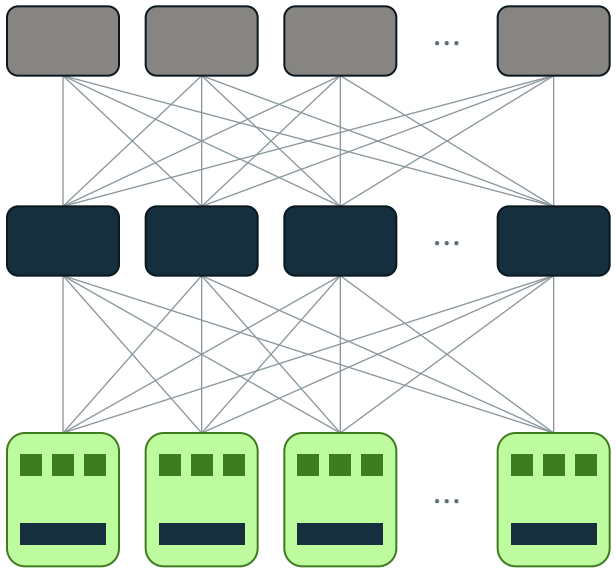
Platform: 2,304 NVIDIA Rubin Ultra packages · 16 racks × 144.

Expert counts: Llama 4, DeepSeek-V3, Kimi K2/K3 as reported; SemiAnalysis on K3 serving.

Two Fabrics, Same 28.8 Tbps Scale-Up

Baseline — scale-up + scale-out

Per-rack 28.8Tbps NVLink + 6.4Tbps 2-tier CLOS with 230.4T switches



96 spine switches

6 × 16 planes · 230.4T

256 leaf switches

64 × 4 pods · 230.4T

16 × 400G per package = 6.4 Tbps scale-out

16 × NVL144 racks

144 pkgs + 36 NVSwitches each

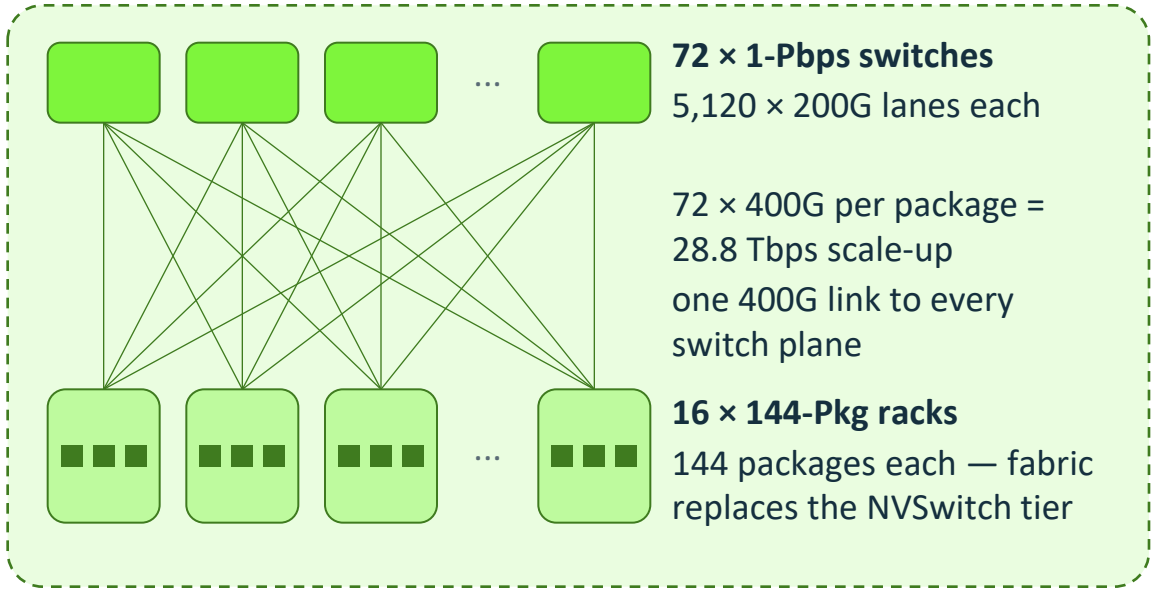
28.8 Tbps NVLink scale-up, in-rack only

576 scale-up switches + 352 scale-out switches · 3 hop diameter

352 × 230.4T switches (256 leaf + 96 spine) · 4-rail, 16-plane CLOS

28.8T scale-up + 6.4T scale-out per pkg

Petabit — one flat, single-hop domain



72 × 1-Pbps switches

5,120 × 200G lanes each

72 × 400G per package = 28.8 Tbps scale-up

one 400G link to every switch plane

16 × 144-Pkg racks

144 packages each — fabric replaces the NVSwitch tier

72 switches total (>80% fewer units)· 1 hop diameter

72 × 1-Pbps switches (5,120×200G) · flat 72-plane scale-up

28.8T per pkg (72×400G) · single hop, diameter 1 · >80% fewer switches.

Two Workload Scenarios

Scenario A · Maximum expert parallelism

EP 2304 — TP1 · PP1 · DP1

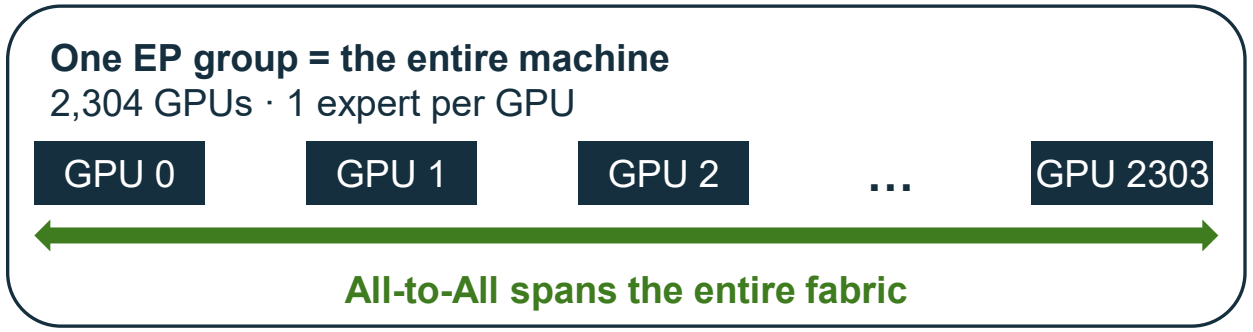
- Eliminates TP collectives and pipeline bubbles (TP = PP = 1)
- Minimizes per-GPU expert-weight residency in HBM (one expert per GPU)
- Exposes peak All-to-All load: dispatch/combine traverse the full fabric

Scenario B · Mixed parallelism

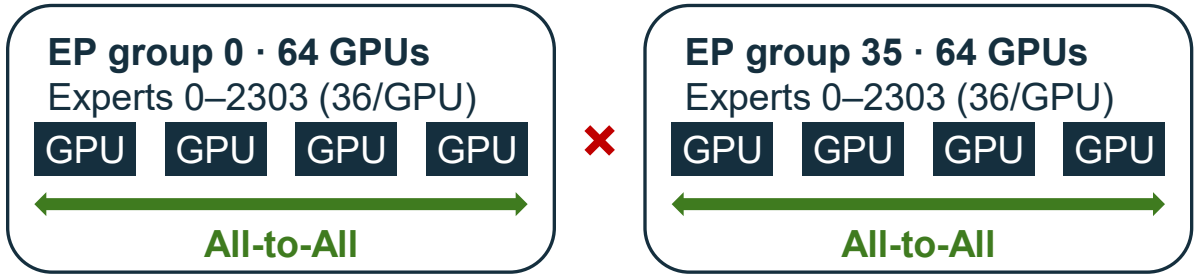
EP 64 — TP4 · PP3 · DP3

- Confines All-to-All to 64-GPU groups, bounding path length and contention
- TP/PP/DP dimensions provide memory headroom for activations and batch
- Preserves communication locality under system-level scaling

A — EP 2304: one All-to-All domain



B — EP 64: All-to-All contained per group



… 36 EP groups total · no All-to-All between groups

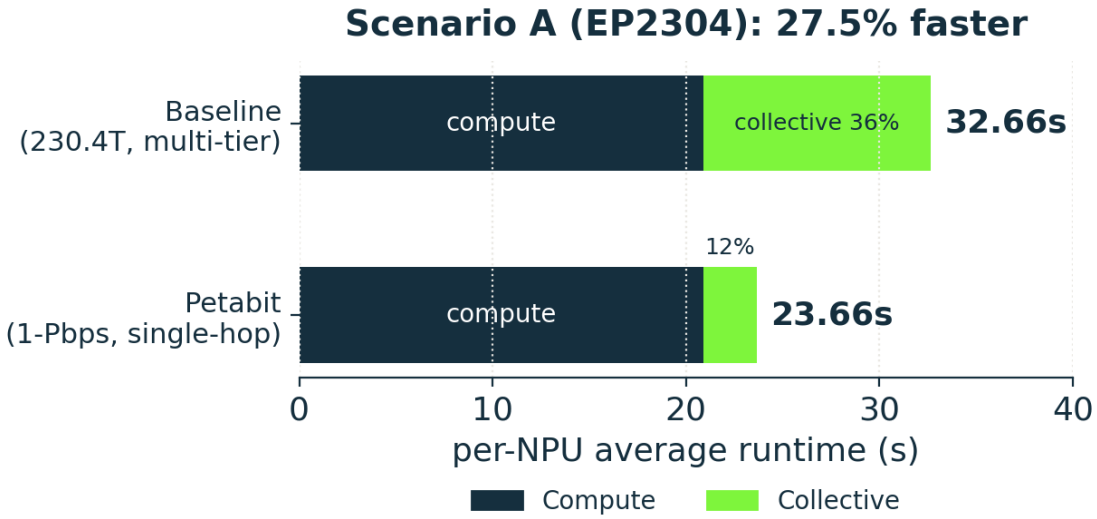
Result A — Maximum Expert Parallelism

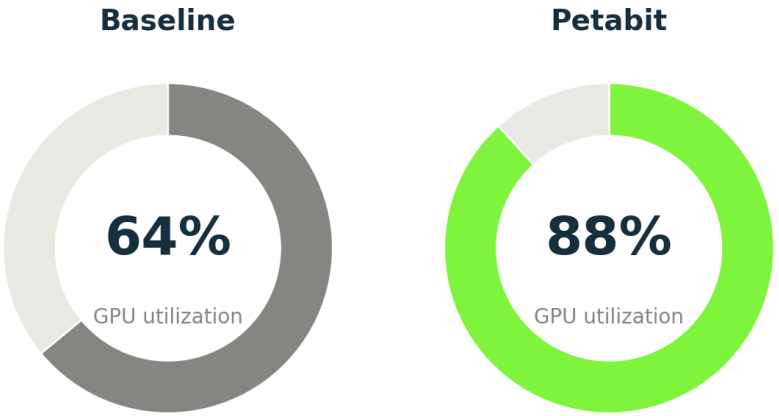
27.5%
faster job completion time

64 % → 88%
GPU utilization (+24 pts)

4.50×
All-to-All collective speedup

36% → 12%
of the iteration spent in collectives





EP2304 · TP1 · PP1 · DP1

72 switches vs 352 · single hop vs 3 tiers

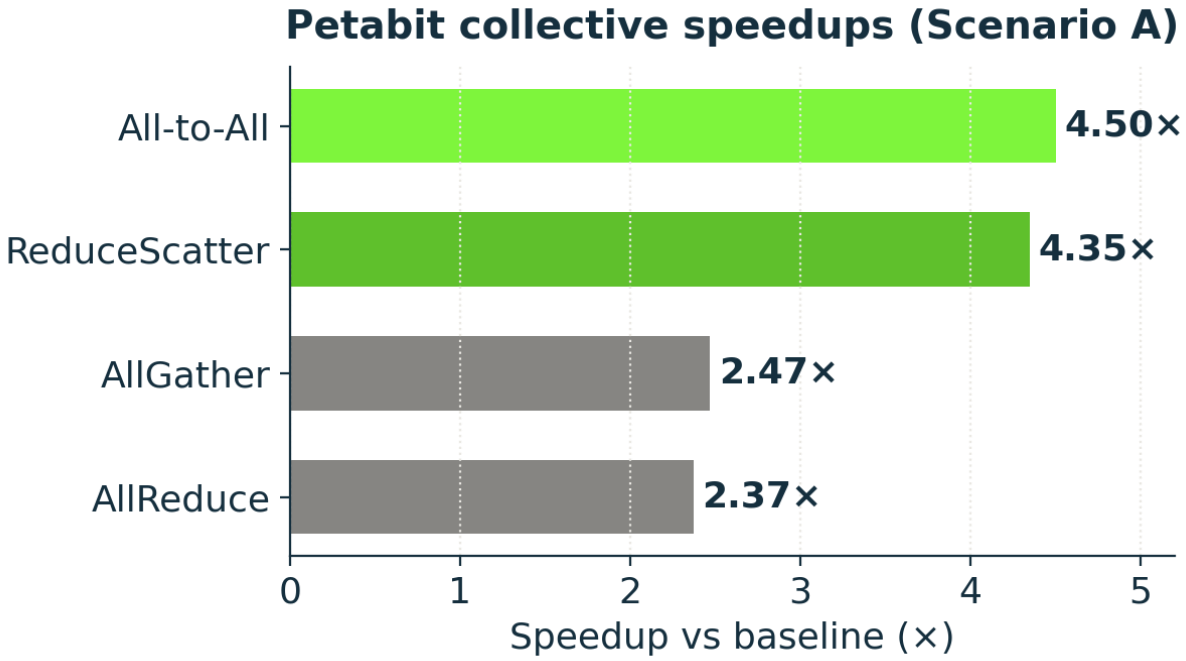
Where the Time Goes: Collective Speedups

All-to-All dominates the gap.

On the single-hop fabric, All-to-All completes 4.50× faster and ReduceScatter 4.35×.

Removing multi-tier congestion collapses All-to-All from ~27% of the iteration to under 10%.

Compute time is identical — the entire improvement is network-driven.



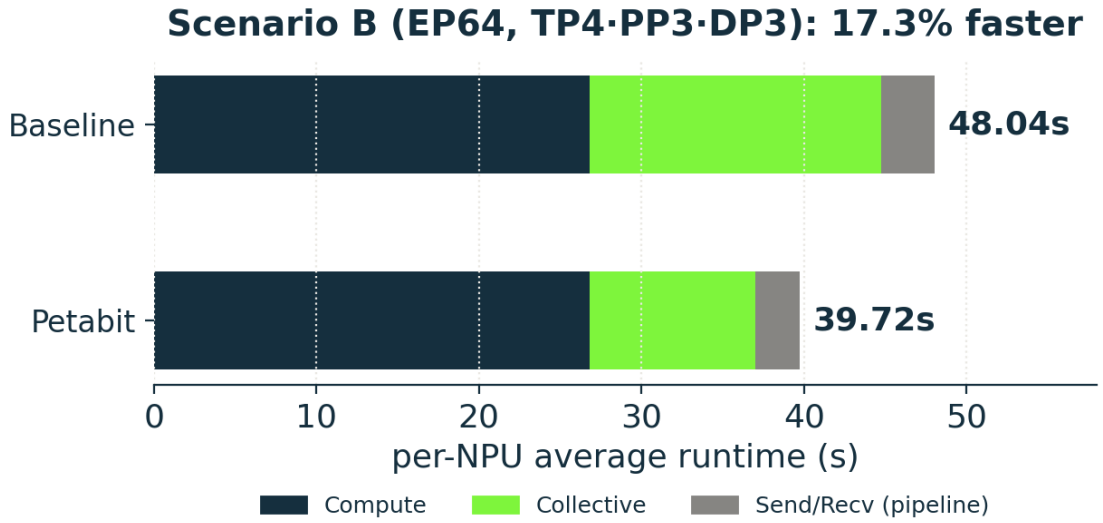
Result B — Mixed Parallelism

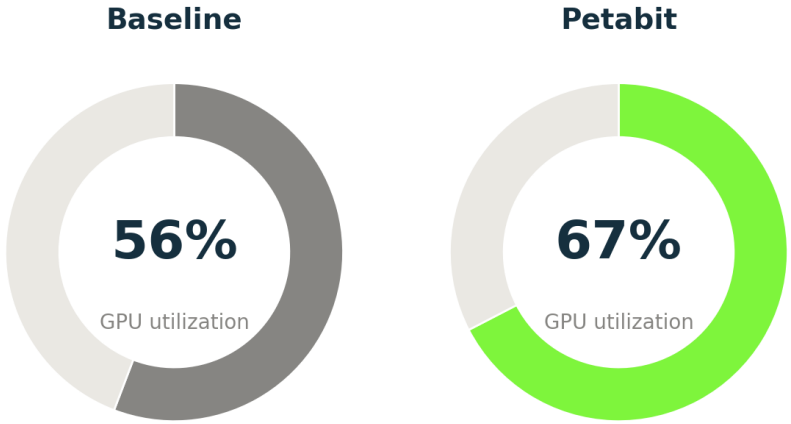
17.3%
faster job completion time

56 % → 67%
GPU utilization (+12 pts)

4.50×
All-to-All speedup sustained

↓ stalls
shorter pipeline receive idle





EP64 · TP4 · PP3 · DP3 — a mixed parallelism deployment
TP stays on NVLink · EP / PP / DP cross the racks

Impact: Why It Matters

64 → 88%

GPU utilization

The GPUs spend far less time waiting on the fabric.

27.5% faster

job completion

Same model, same GPUs — the difference is the network.

Better economics

Higher utilization means more useful tokens per GPU-hour — directly lowering cost per token, dollar and watt.

At Colossus 1 scale (~200,000 GPUs), 64 → 88% is worth **≈ $1.3B per year**

Eases the HBM squeeze

High expert parallelism cuts HBM per GPU, freeing HBM for more accelerator units.

A path to frontier-scale MoE

As models add experts and parameters, the margin of a flat fabric over a multi-tier one should widen.

Fewer switches

72 instead of 352 — 80% fewer — cutting power, cabling and congestion at the same scale-up bandwidth.

The payoff is utilization and economics — fewer switches is the bonus.

Future Work

- **Scale to the full domain**

  Extend the simulation to the switch's full 5,040-package single-tier reach.

- **Revisit 4D parallelism**

  Rebalance TP / EP / PP / DP now that a flat fabric makes every package equidistant.

- **Disaggregated inference**

  Apply the same fabric comparison to prefill / decode inference workloads.

- **Switch-level co-design**

  Explore buffer sizing, congestion control and plane count in TurbaSim.

Thank you — Questions?

Breaking Through the Network Wall: Petabit-Class Switching for Single-Hop MoE Training at Scale

IEEE Hot Interconnects 33 (HotI 2026) · in collaboration with LBNL

Philipp Berdesinski	philipp.berdesinski@turbalance.com
Laurent Montigny	laurent.montigny@turbalance.com
John Shalf	jshalf@lbl.gov
George Michelogiannakis	mihelog@lbl.gov