Scaling Inference Prefill with High-Radix Photonic Interconnects

Arulselvan Madhavan (Presenter)
Peter Carson, Taylor Groves, Thomas Graham

Hot Interconnects | Virtual Conference | August 19, 2026

Agenda

Part 1 — The Inference and ML Systems Landscape

Part 2 — What can Photonic Interconnects Offer?

Part 3 — Modeling Inference Prefill Impact of Photonic Interconnects

Part 1

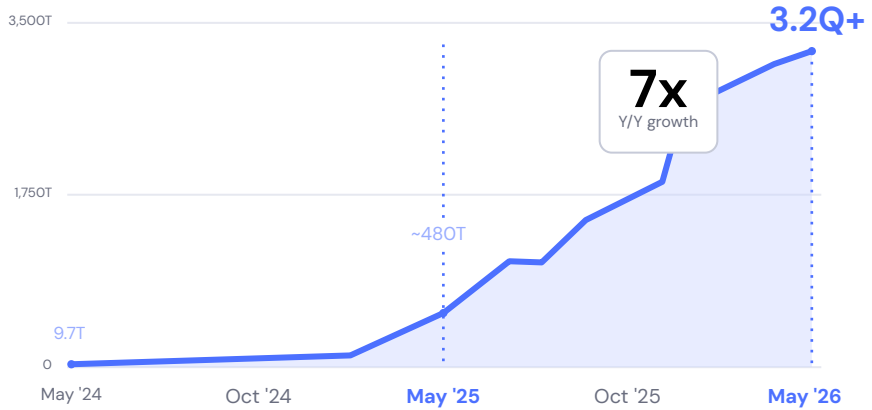
The Inference & ML Systems Landscape

Inference is the dominant ML workload

Human generated stock of training data : 300T tokens by 2026[1]
Training data usage : ~30T tokens over 3 months
Inference data : >100T tokens in a day[2]

Model	Release	Training tokens	Days
Inkling	July 2026	45 Trillion	~100 - 120 days
Llama 4 Scout	April 2025	~40 Trillion	~90 days
Qwen3	May 2025	36 Trillion	~75 - 90 days
DeepSeek-V4 Pro	April 2026	33 Trillion	~85 days
MAI-Thinking-1	June 2026	30 Trillion	Undisclosed

Google I/O 2026
Monthly Tokens Processed[2]



1 Villalobos et al. (2024) "Will we run out of data? Limits of LLM scaling based on human-generated data"
2 Google (2026) "Google I/O 2026: Sundar pichai's opening keynote"

Specialized hardware for prefill and decode

Prefill (throughput): AMD Helios, AWS Trainium3
Decode (latency): Cerebras WSE (CS-3)

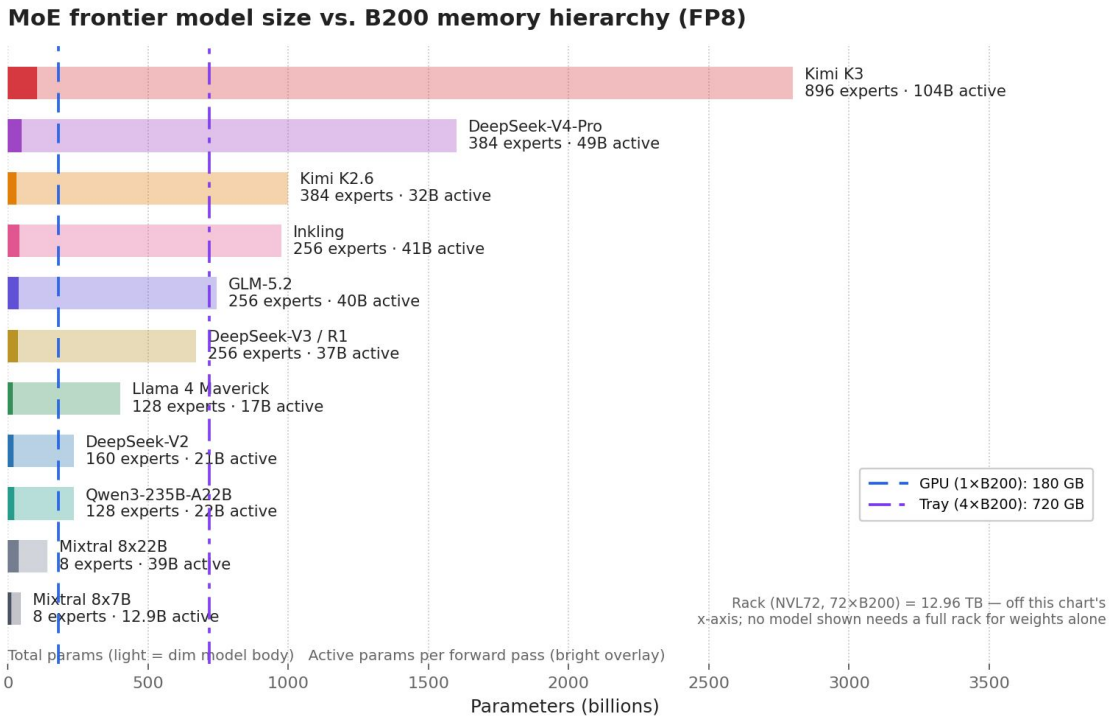
Prefill

Decode

Source: AMD and Cerebras Systems (2026) "AMD and Cerebras announce industry-leading ultra-low latency and high-throughput AI inference"

Models must be sharded: weights

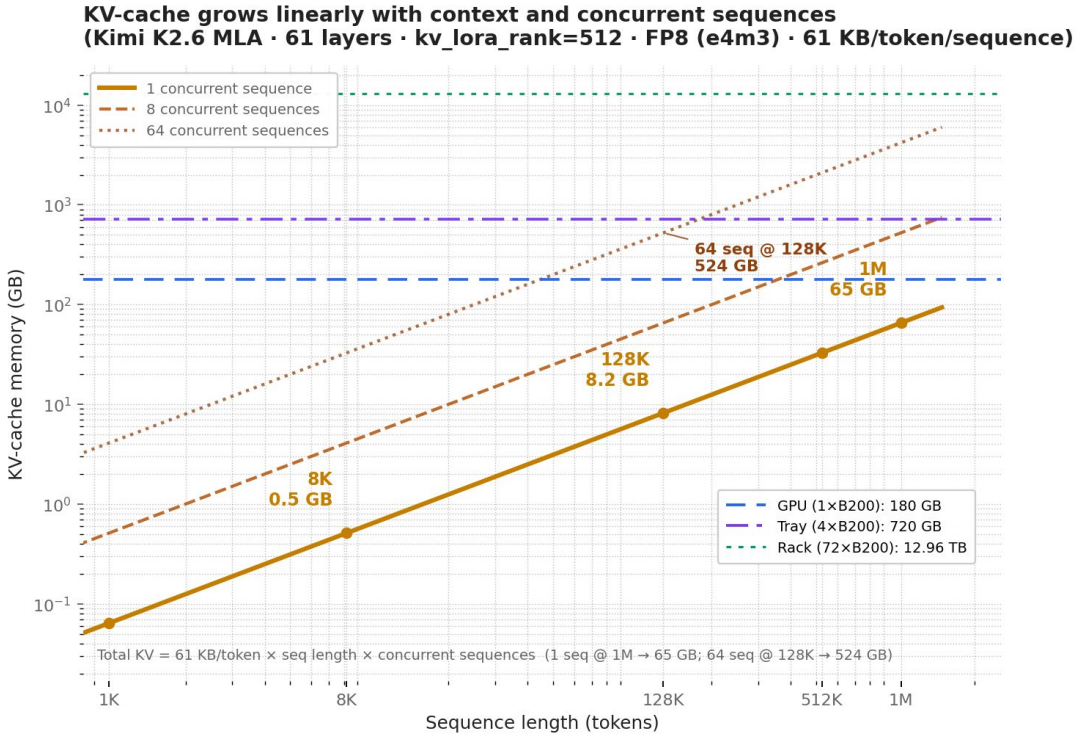
Weights exceed single-GPU HBM even at FP8



1 DeepSeek-AI (2024) "DeepSeek-V3 technical report" arXiv preprint arXiv:2412.19437; Kimi Team, Moonshot AI (2026) "Kimi K3: Open frontier intelligence"; DeepSeek-AI (2026) "DeepSeek-V4: Towards highly efficient million-token context intelligence," (2024) "DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model"; Moonshot AI (2026) "Kimi K2.6 model card"; Zhipu AI / Z.ai (2026) "GLM-5.2 model card"; Thinking Machines Lab (2026) "Inkling: Our open-weights model"; Mistral AI (2023) "Mixtral of experts," (2024) "Cheaper, better, faster, stronger"; Meta AI (2025) "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation"; Qwen Team (2025) "Qwen3 technical report"

Models must be sharded: KV-cache

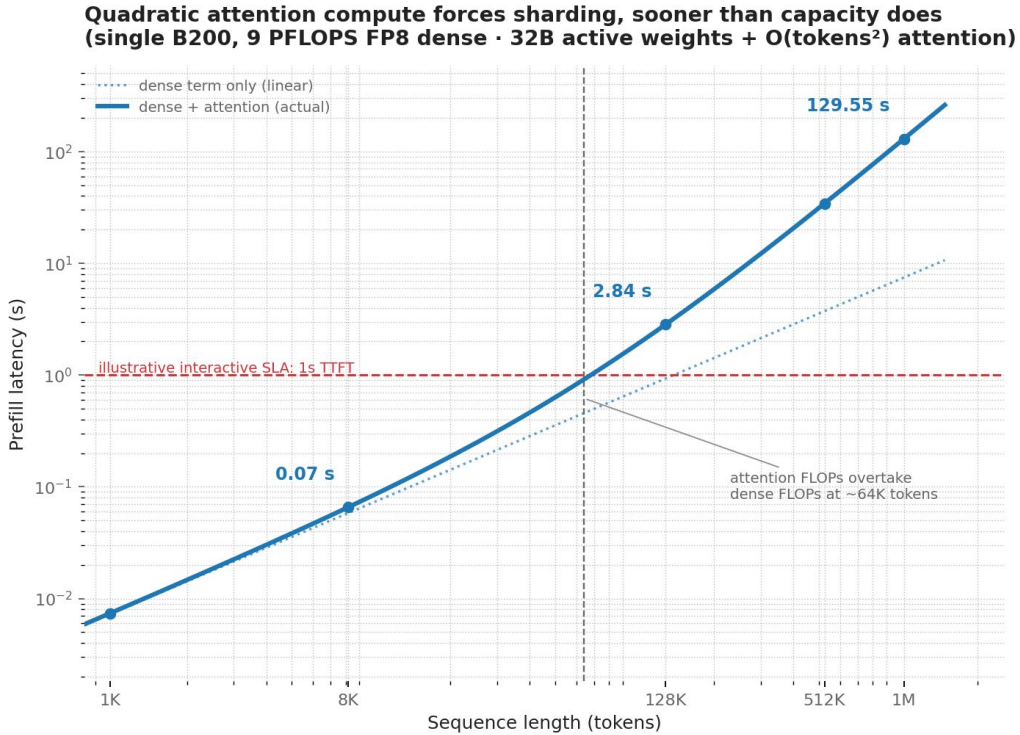
KV-cache adds memory pressure at long context, even at FP8



1 Moonshot AI (2026) "Kimi K2.6 model card"
2 vLLM Project (2026) "The state of FP8 KV-cache and attention quantization in vLLM"

Models must be sharded: prefill compute

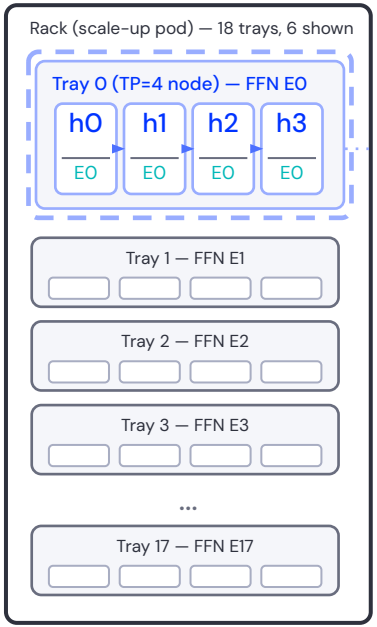
Prefill compute forces sharding even sooner than either memory limit



1 Fernando (2026) "The bandwidth wall: A roofline for local LLMs"
2 Moonshot AI (2026) "Kimi K2.6 model card"

Rack: tensor parallel is the innermost shard

Tensor parallel splits the tensors across compute units and introduce All-Reduce



TP — Tensor Parallel

TP degree is bounded by # attention heads and embedding size

Current MoE models: 64–128 attention heads · 4,096–7,168 embedding dim

Range: [1, node-level]

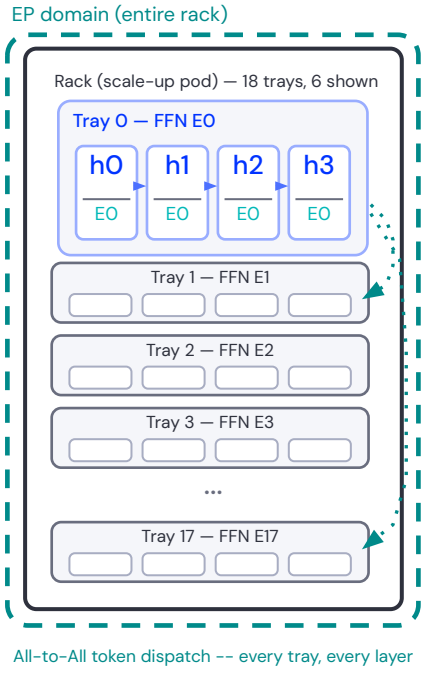
TP incurs All-Reduce communication every layer

Each tray = 1 expert's FFN – that's EP, next slide

Source: Li et al. (2024) "Flash communication: Reducing tensor parallelization bottleneck for fast large language model inference"

Rack: expert parallel fills the whole pod

Expert Parallel places experts across sets of compute units. Introduces All2All communication



TP=4 — still runs inside every tray

Bounded by # attention heads and embedding size · incurs All-Reduce every layer

EP — Expert Parallel

EP degree is bounded by # of routed experts

Current MoE models: 64–896 routed experts · top-K per token (e.g. top-8 to top-16)

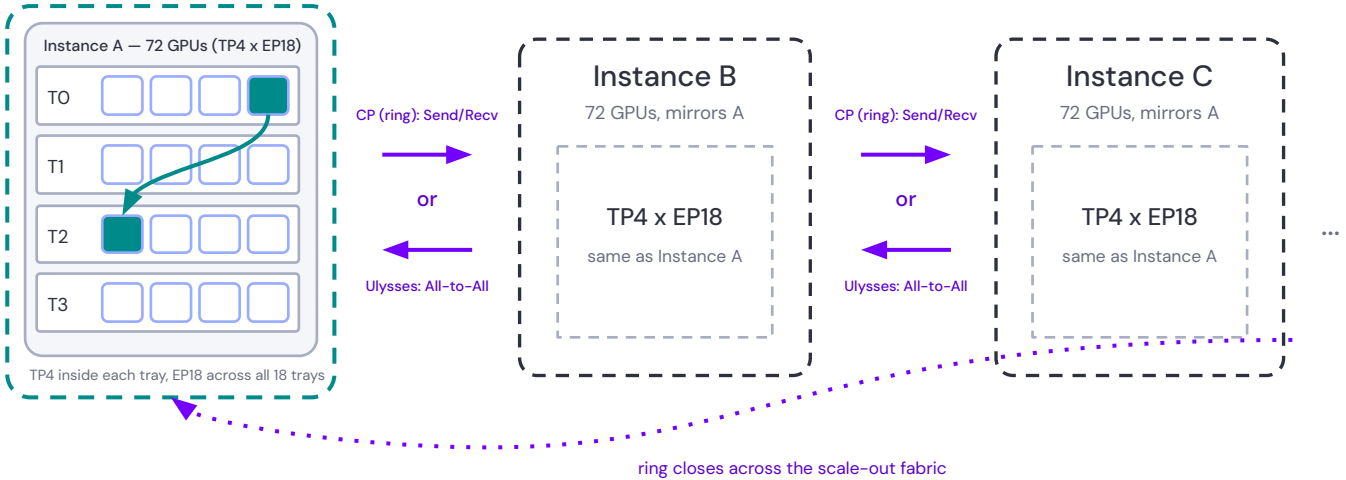
Range: [1, rack-level or pod-level]

EP incurs All-to-All communication every layer

Source: DeepSeek-AI (2024) "DeepSeek-V3 technical report" arXiv preprint arXiv:2412.19437; Kimi Team, Moonshot AI (2026) "Kimi K3: Open frontier intelligence"

Multi-Rack: context parallel spans the fabric

Each rack is one full model instance.; CP spans the sequence across instances



CP / SP — Context · Sequence Parallel

Splits sequences across model instances · requires inter-instance comms
Size set by context-length regime, not model architecture · range: 1K–1M tokens

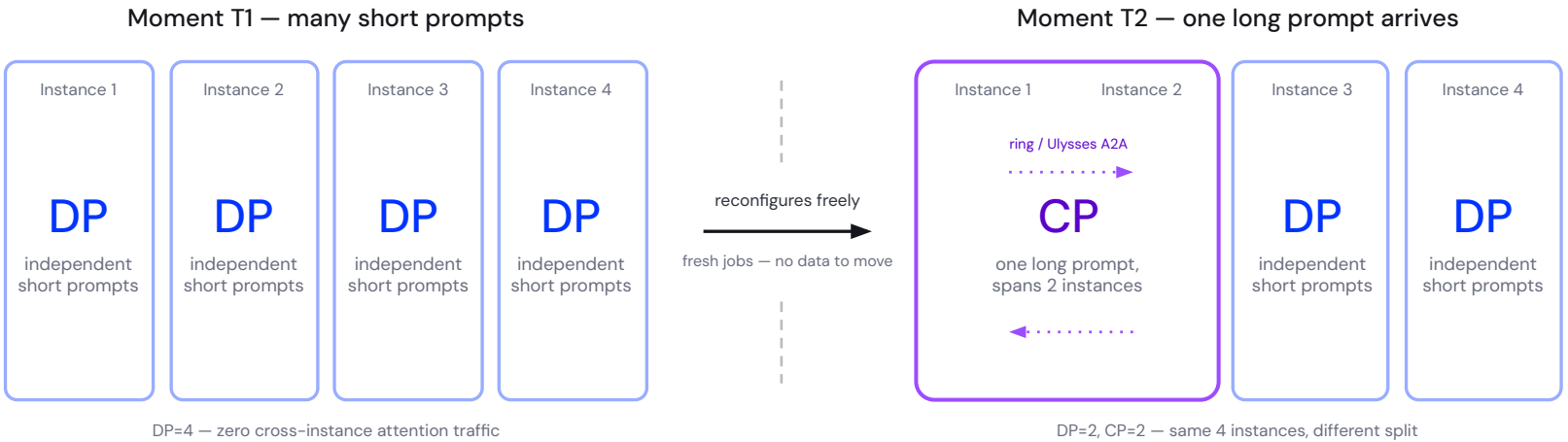
SP: All-Gather
CP (ring): Send/Recv · Ulysses: All-to-All

Source: Yang et al. (2024) “Context parallelism for scalable million-token inference”

Data parallel and context parallel

Flexible serving: the same instances, repartitioned by request mix.

Each instance = 1 rack = 72 GPUs (TP4 x EP18); fresh prefill jobs pick DP or CP at admission



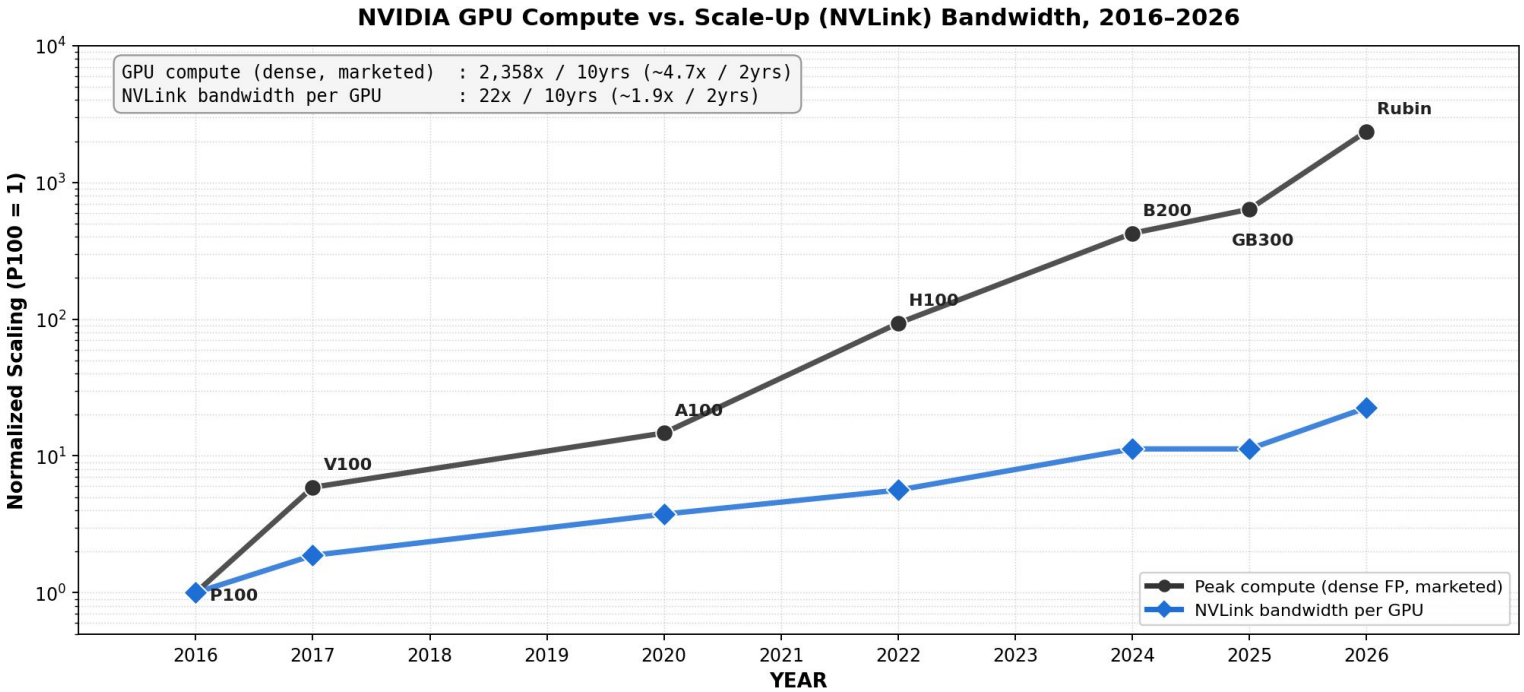
DP <-> CP: the same pod, reassigned per request

Long sequence -> more instances join one CP group. Many short sequences -> more independent DP jobs.
Both draw from the same fixed pool of 72-GPU instances; only the split between them changes.

Source: Wu et al. (2024) "LoongServe: Efficiently serving long-context large language models with elastic sequence parallelism" Proceedings of the ACM SIGOPS 30th symposium on operating systems principles (SOSP '24)

Scale-Up is Not Keeping Up

GPU compute : 2,358× over 10 yrs (~4.7×/2 yrs)
NVLink BW : 22× over 10 yrs (~1.9×/2 yrs)



Source: Patel et al. (2025) "NVIDIA GTC 2025 – built for reasoning, vera rubin, kyber, CPO, dynamo inference, jensen math, feynman"

Part 2

What Can Photonic Interconnects Offer?

Optics vs Copper

Reach

1,000×

Optical links span 1 km
vs ~1 m for copper —
rack-scale becomes row-scale.

Optics — 1 km

Copper — 1 m

Port density

8×

One bidirectional fiber vs. four
30 AWG wires (2 differential
pairs).

1 bidirectional fiber

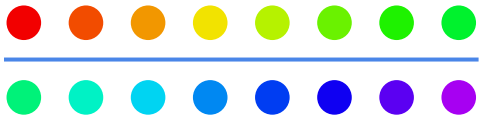
4×30 AWG wires (2 diff pairs)

Bandwidth density

16 λ

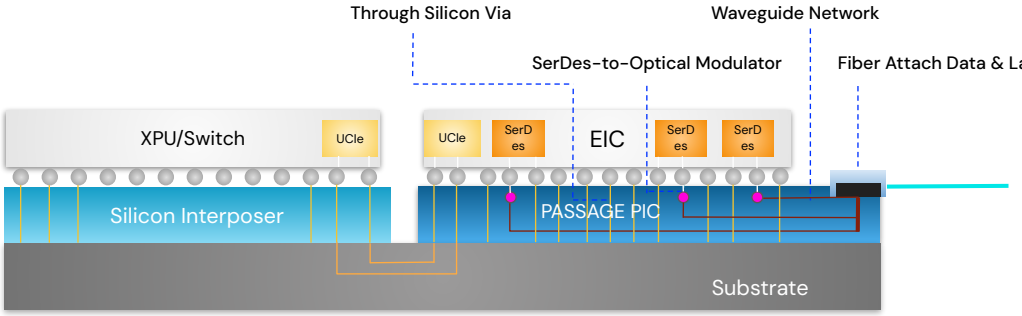
16 lambdas multiplexed
on a single fiber —
demonstrated today.

16 wavelengths, one fiber

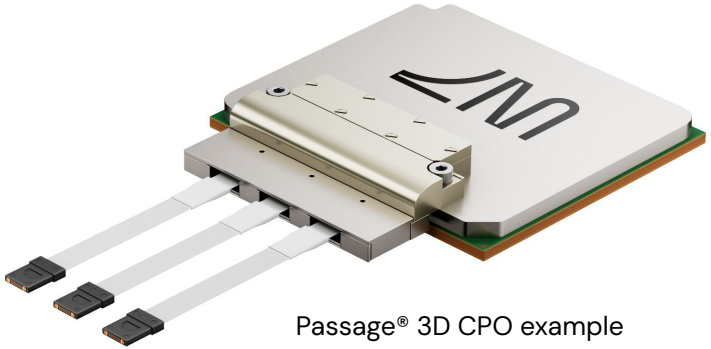


Example Implementation of Photonic Interconnects

3D Co-Packaged Optics (CPO)



- Advanced node Electrical Integrated Circuit (EIC) 3D packaged with a Photonic Integrated Circuit (PIC)
- UCIe interface connect 3D CPO to XPU/Switch ASICs
- 112 Gbps PAM4 SerDes, 16λ WDM for 1.6 Tbps/fiber
- 64 fibers → 64 Tbps per 3D CPO
- 2× 3D CPO on XPU → 112 Tbps BiDi I/O per XPU



Passage® 3D CPO example

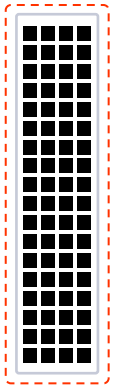
Reach & Port Density Set the Scale-Up Domain

Electrical: NVL72, copper NVLink, in-rack (72 GPUs) · Optical: CPO, rack-to-rack (1,152 GPUs)

1,000× reach × 8× port density → 16× larger scale-up domain

Electrical
(copper NVLink)

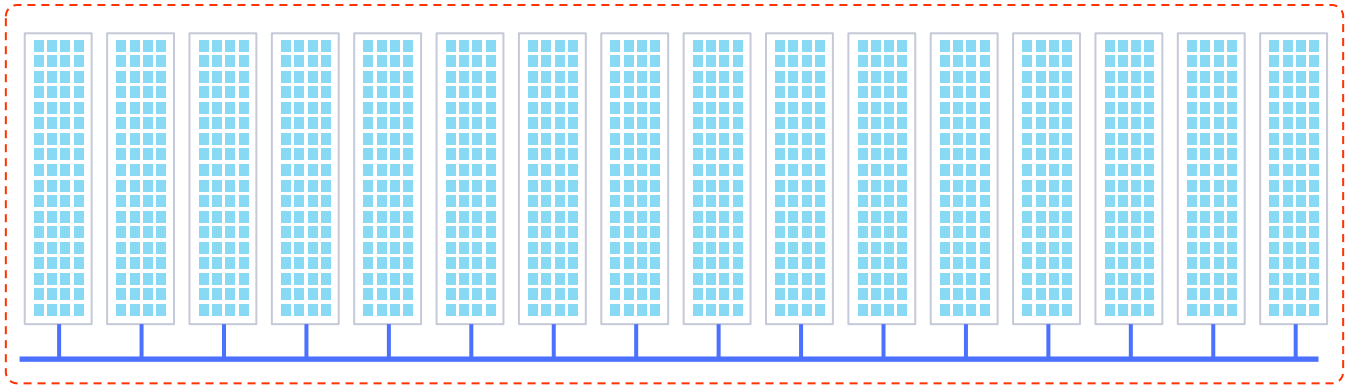
Scale-up domain



NVL72 (copper NVLink, in-rack)
72 GPUs · 1 rack · ~1 m reach

Optical
(CPO fabric)

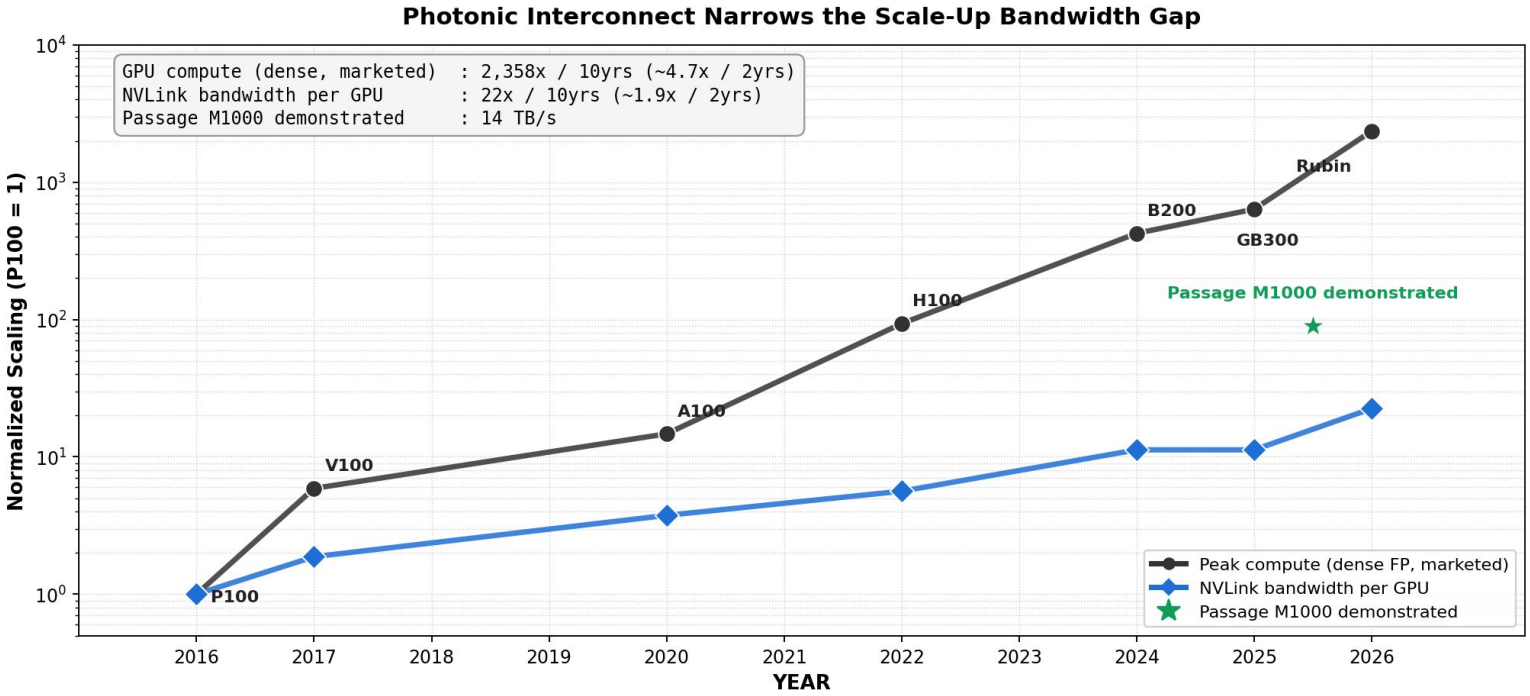
Scale-up domain



CPO, rack-to-rack — optical CPO fabric spine
1152 GPUs · 16 racks · ~1 km reach

Source: SemiAnalysis (2026) “Nvidia: The inference kingdom expands”

Photonics shortens the gap



Source: Lightmatter (2026) “Passage: 3D photonics for AI applications,” (2025) “Passage M1000 EVK – 3D photonic interconnect for AI infrastructure”

Part 3

Modeling Inference Prefill Impact of Photonic Interconnects

Modeling Methodology

Production partitionings → XLA MLIR cost walk → overlapped prefill latency (compute + collectives)
Compare: electrical baseline vs optical twin — same compute/HBM, 4× scale-up BW, pod to 1,152 GPUs
Sweeps: device count (~8M input tokens per context) and batch size (fixed cluster per regime)

Production partitionings -> MLIR -> compute + collective cost walk

Scenarios We Modeled

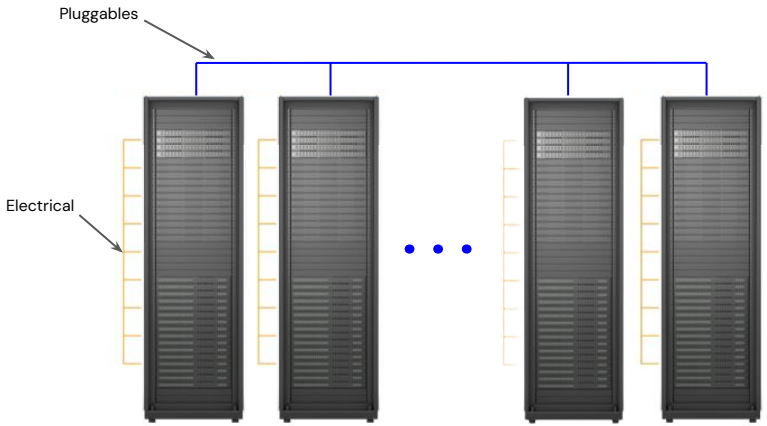
- Frontier MoE model: DeepSeek R1 variant, 42B active / 840B total parameters, FP4
- Three context regimes: low (8K), medium (128K), large (1M)
- Current and futuristic hardware specs: electrical baseline vs. 4× bandwidth optical scale-up
- Serving Capacity: 8M tokens

Modeling network setup

Networks

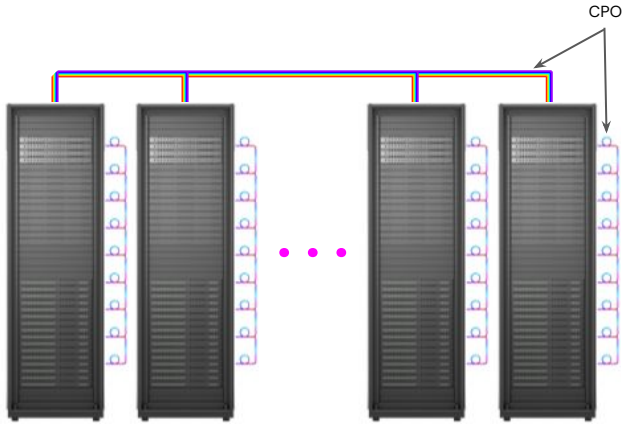
Electrical+Optical Pluggables

Electrical 14.4-28.8 Tbps per GPU in rack
Pluggable optics across racks



Passage 3D CPO

Optical = 4X electrical bandwidth per GPU in rack and across racks



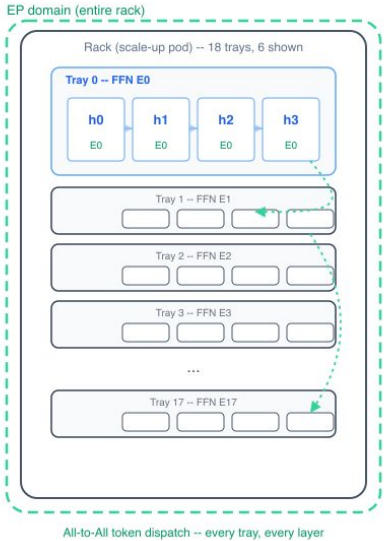
All bandwidth numbers are bi-directional

Low Context Regime (8K)

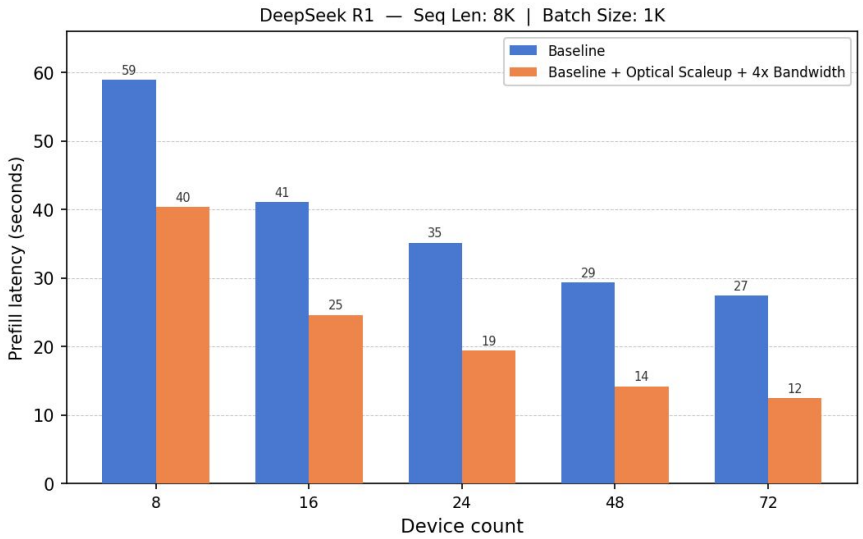
(1024 users x 8K context = 8M tokens)

Up to 2.2x reduction in latency

Model deployed in a rack — TP + EP



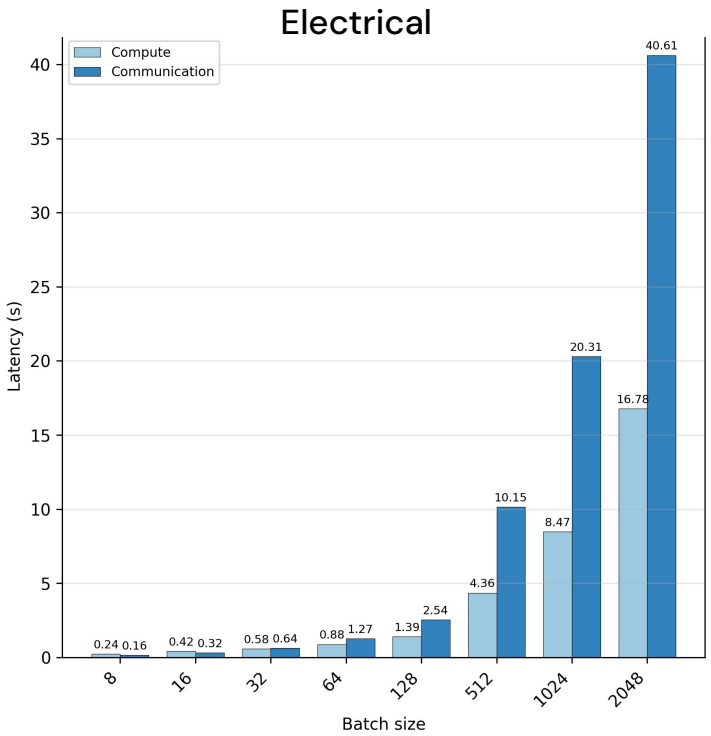
Prefill Latency vs. Device Count

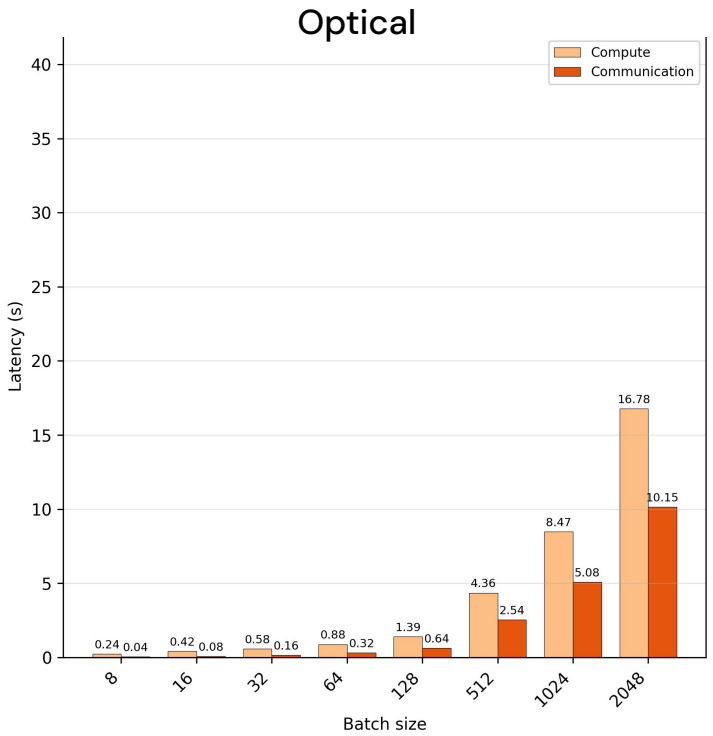


Model	: DeepSeek R1 variant, 42B active / 840B total parameters, MoE, FP4
Electrical baseline	: 14 PFLOPS/GPU, HBM 7.9 TB/s, SU BW 900 GB/s, scale-up pod 72 GPUs (NVIDIA B300)
Optical baseline	: 4× SU BW, 1152-GPU pod; all else same as electrical

Handle more users at lower latency

Compute vs. Comm at 8K prefill · 72 devices

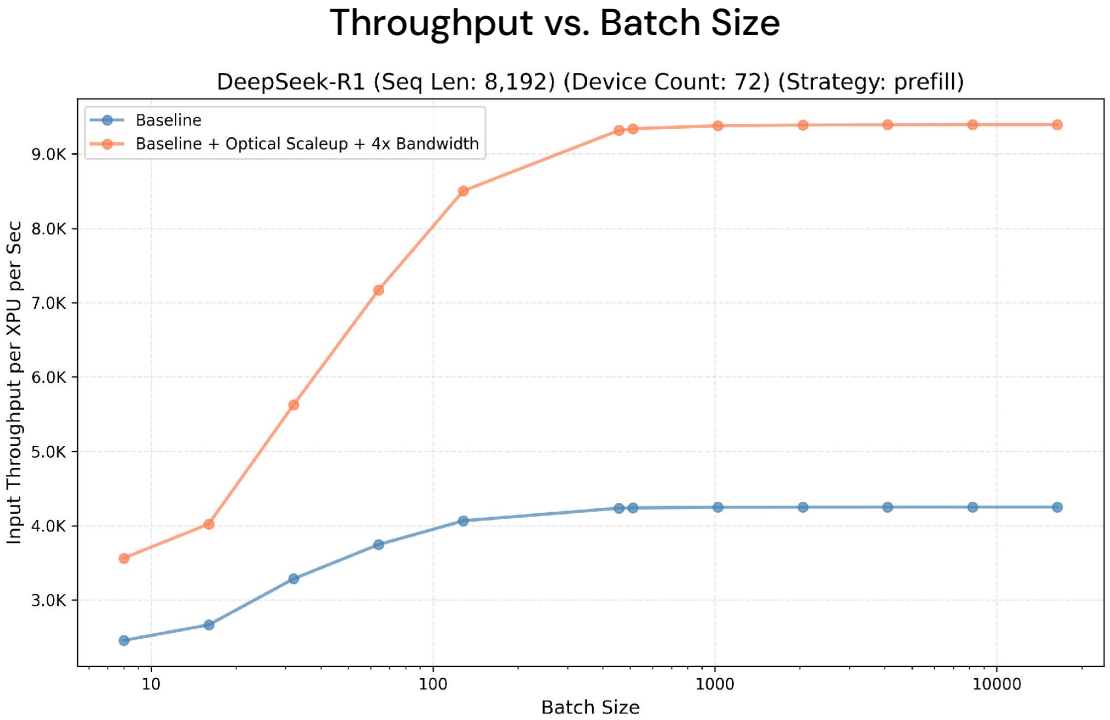




Low Context — Throughput and Latency

8K prefill · 72 devices.

At 1024 batch size, throughput increases from 4K to 9K tokens per second per XPU

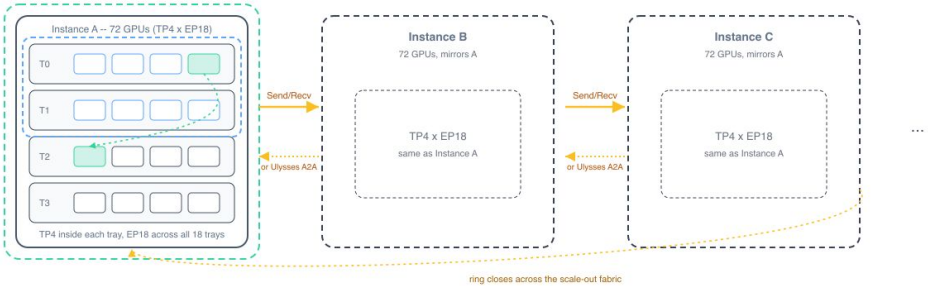


Medium Context Regime (128K)

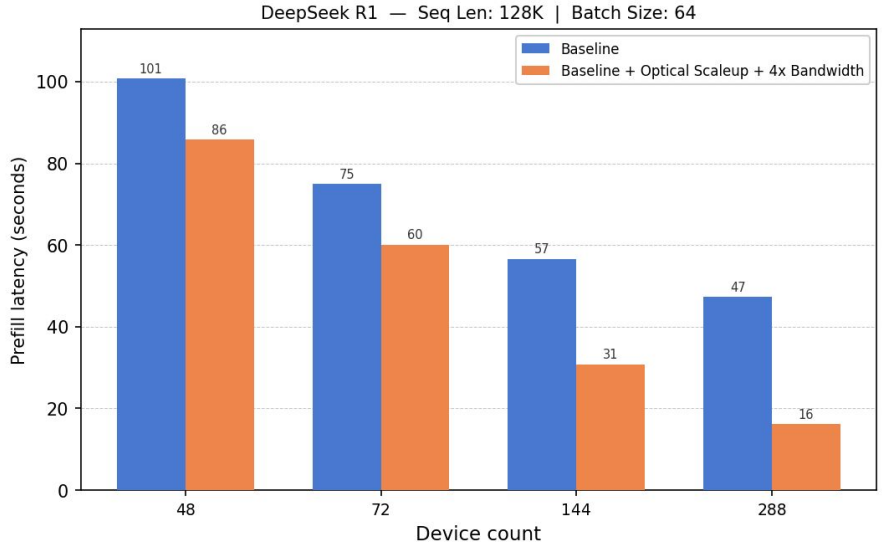
(64 users x 128K context = 8M tokens)

Up to 2.9x reduction in latency

Compute can afford to spill beyond a rack



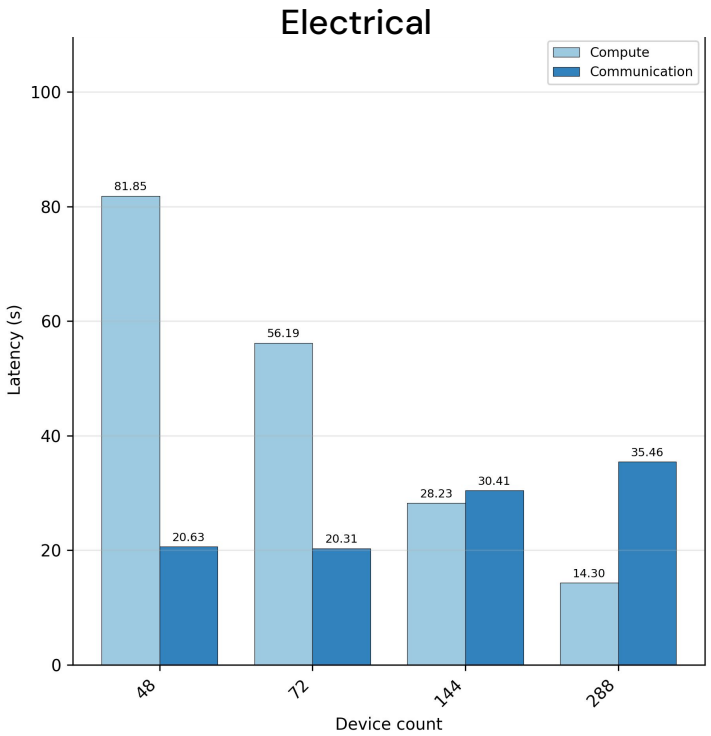
Prefill Latency vs. Device Count

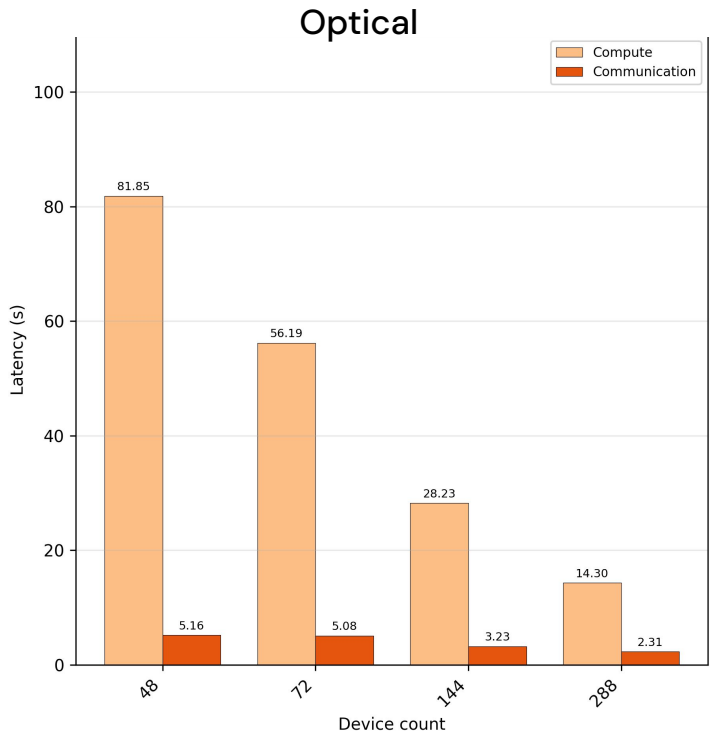


Model	: DeepSeek R1 variant, 42B active / 840B total parameters, MoE, FP4
Electrical baseline	: 14 PFLOPS/GPU, HBM 7.9 TB/s, SU BW 900 GB/s, scale-up pod 72 GPUs (NVIDIA B300)
Optical baseline	: 4× SU BW, 1152-GPU pod; all else same as electrical

Compute Spill to Scale-Out

128K prefill · batch 64; Ratio of Scale-up to Scale-out is 900GBps to 100GBps



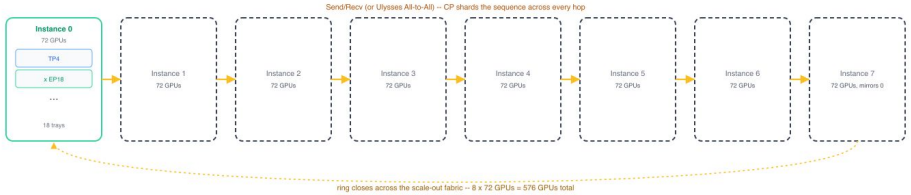


Large Context Regime (1M)

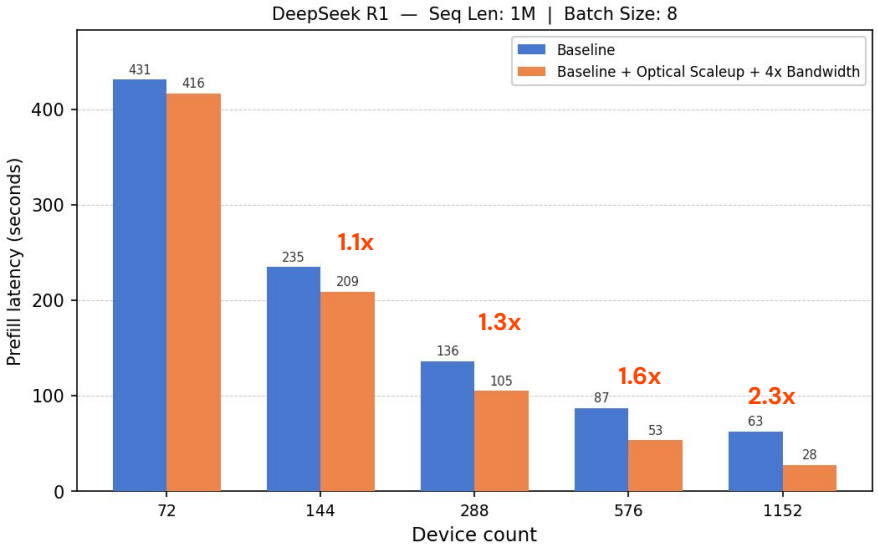
(8 users x 1M context = 8M tokens)

Up to 2.3x reduction in latency

Compute scaling to a full row



Prefill Latency vs. Device Count

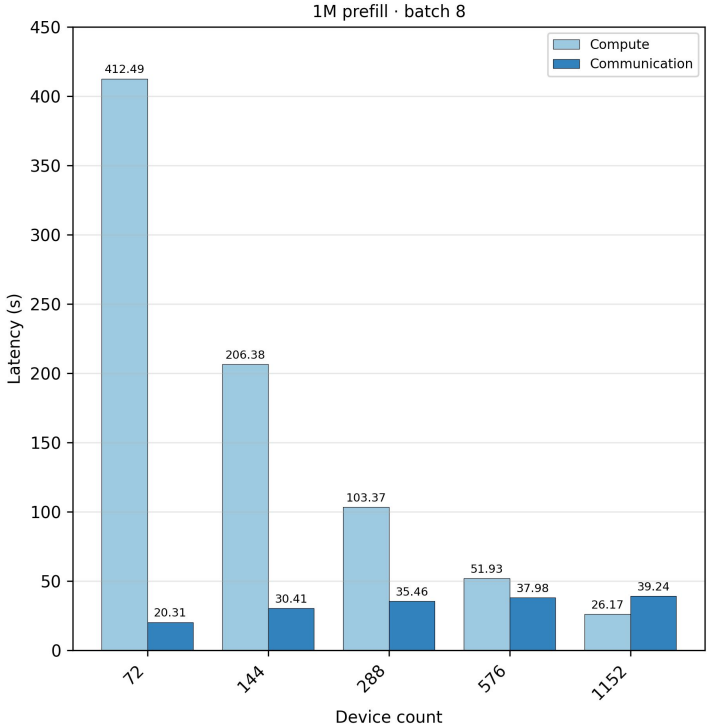


Model	: DeepSeek R1 variant, 42B active / 840B total parameters, MoE, FP4
Electrical baseline	: 14 PFLOPS/GPU, HBM 7.9 TB/s, SU BW 900 GB/s, scale-up pod 72 GPUs (NVIDIA B300)
Optical baseline	: 4× SU BW, 1152-GPU pod; all else same as electrical

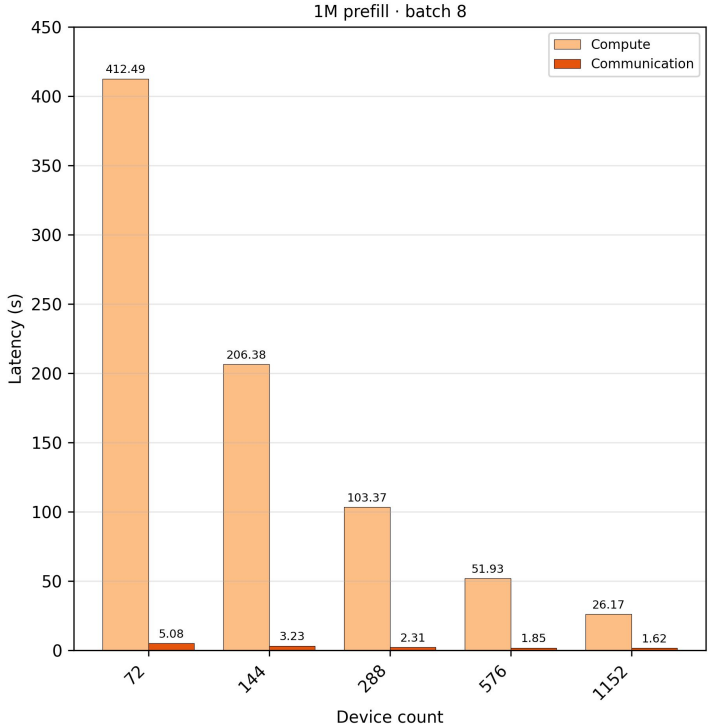
Scale-Out Costs Communication

1M prefill · batch 8 — same device sweep as prior slide; up to 1,152 GPUs

Electrical

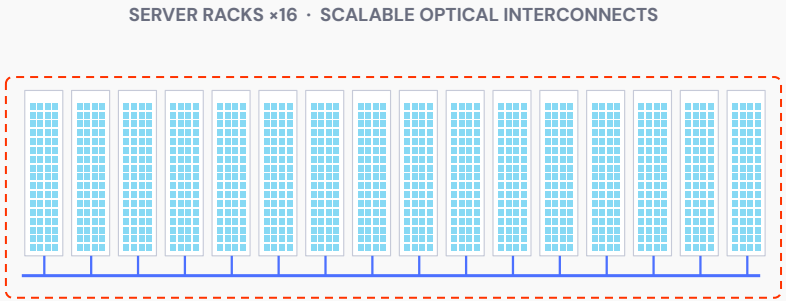


Optical



Conclusion

Conclusion: Optical Scale-Up Latency Reduction for Prefill



4x Bandwidth Expansion

Across a 16-rack, 1,152-GPU pod

Context	Current HW (14 PFLOPS · 7.9 TB/s HBM · 900 GB/s SU)	Future HW (25 PFLOPS · 22 TB/s HBM · 1800 GB/s SU)
Low (8K)	2.2×	2.4×
Medium (128K)	2.9×	5.7×
Large (1M)	2.3×	4.5×

Summary of Cross-Regime Performance

- **Optical Baseline:** 4× Scale-Up Bandwidth, 1152-GPU pod structure
- **Sweep Scope:** Multi-config analysis evaluating current and future hardware platforms across 4 distinct context lengths (1K–1M)
- **Performance:** Delivers up to 5.7× latency reduction over equivalent electrical configurations


HOT
INTERCONNECTS

References

References (1/5)

1. Google. 2026. “Google I/O 2026: Sundar pichai’s opening keynote.” https://blog.google/innovation-and-ai/sundar-pichai-io-2026/.

2. P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn. 2024. “Will we run out of data? Limits of LLM scaling based on human-generated data.” https://arxiv.org/abs/2211.04325.

3. Q. Zhang, F. Lyu, Z. Sun, L. Wang, W. Zhang, W. Hua, et al. 2025. “A survey on test-time scaling in large language models: What, how, where, and how well?” https://arxiv.org/abs/2503.24235.

4. S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. 2023. “Tree of thoughts: Deliberate problem solving with large language models.” https://arxiv.org/abs/2305.10601.

5. C. Lin, H. Jiang, J. Xu, Y. Yang, and L. Qiu. 2025. “Efficient multi-round LLM inference over disaggregated serving.” https://arxiv.org/abs/2602.14516.

6. SemiAnalysis. 2026. “WekaTrace: Claude code execution telemetry profile.” https://huggingface.co/datasets/semianalysisai/cc-traces-weka-062126.

References (2/5)

7. G. Steele, and NVIDIA. 2026. “How the NVIDIA vera rubin platform is solving agentic AI’s scale-up problem.” https://developer.nvidia.com/blog/how-the-nvidia-vera-rubin-platform-is-solving-agentic-ais-scale-up-problem/.

8. DeepSeek-AI. 2024. “DeepSeek-V3 technical report.” arXiv preprint arXiv:2412.19437. https://arxiv.org/abs/2412.19437.

9. Kimi Team, Moonshot AI. 2026. “Kimi K3: Open frontier intelligence.” https://arxiv.org/abs/2607.24653.

10. DeepSeek-AI. 2026. “DeepSeek-V4: Towards highly efficient million-token context intelligence.” https://arxiv.org/abs/2606.19348.

11. Moonshot AI. 2026. “Kimi K2.6 model card.” https://huggingface.co/moonshotai/Kimi-K2.6.

12. Zhipu AI / Z.ai. 2026. “GLM-5.2 model card.” https://huggingface.co/zai-org/GLM-5.2.

References (3/5)

13. Thinking Machines Lab. 2026. “Inkling: Our open-weights model.” https://thinkingmachines.ai/model-card/inkling/.

14. Mistral AI. 2023. “Mixtral of experts.” https://mistral.ai/news/mixtral-of-experts/.

15. Mistral AI. 2024. “Cheaper, better, faster, stronger.” https://mistral.ai/news/mixtral-8x22b/.

16. DeepSeek-AI. 2024. “DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model.” https://arxiv.org/abs/2405.04434.

17. Meta AI. 2025. “The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.” https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

18. Qwen Team. 2025. “Qwen3 technical report.” https://arxiv.org/abs/2505.09388.

References (4/5)

19. vLLM Project. 2026. “The state of FP8 KV-cache and attention quantization in vLLM.” https://vllm.ai/blog/2026-04-22-fp8-kvcache.

20. T. Fernando. 2026. “The bandwidth wall: A roofline for local LLMs.” https://tensorfoundry.io/blog/roofline-llm-apple-silicon.

21. Q. Li, B. Zhang, L. Ye, Y. Zhang, W. Wu, Y. Sun, et al. 2024. “Flash communication: Reducing tensor parallelization bottleneck for fast large language model inference.” https://arxiv.org/abs/2412.04964.

22. A. (Jie) Yang, J. Yang, A. Ibrahim, X. Xie, B. Tang, G. Sizov, et al. 2024. “Context parallelism for scalable million-token inference.” https://arxiv.org/abs/2411.01783.

23. B. Wu, S. Liu, Y. Zhong, P. Sun, X. Liu, and X. Jin. 2024. “LoongServe: Efficiently serving long-context large language models with elastic sequence parallelism.” Proceedings of the ACM SIGOPS 30th symposium on operating systems principles (SOSP ’24). https://arxiv.org/abs/2404.09526.

24. D. Patel, M. Xie, D. Nishball, and SemiAnalysis. 2025. “NVIDIA GTC 2025 – built for reasoning, vera rubin, kyber, CPO, dynamo inference, jensen math, feynman.” https://newsletter.semianalysis.com/p/nvidia-gtc-2025-built-for-reasoning-vera-rubin-kyber-cpo-dynamo-inference-jensen-math-feynman.

References (5/5)

25. Lightmatter. 2025. “Passage L200/L200X 3D co-packaged optics product brief.”

26. SemiAnalysis. 2026. “Nvidia: The inference kingdom expands.” https://newsletter.semianalysis.com/p/nvidia-the-inference-kingdom-expands.

27. Lightmatter. 2026. “Passage: 3D photonics for AI applications.” https://lightmatter.co/products/passage/.

28. Lightmatter. 2025. “Passage M1000 EVK – 3D photonic interconnect for AI infrastructure.” https://lightmatter.co/products/passage-m1000-evk/.