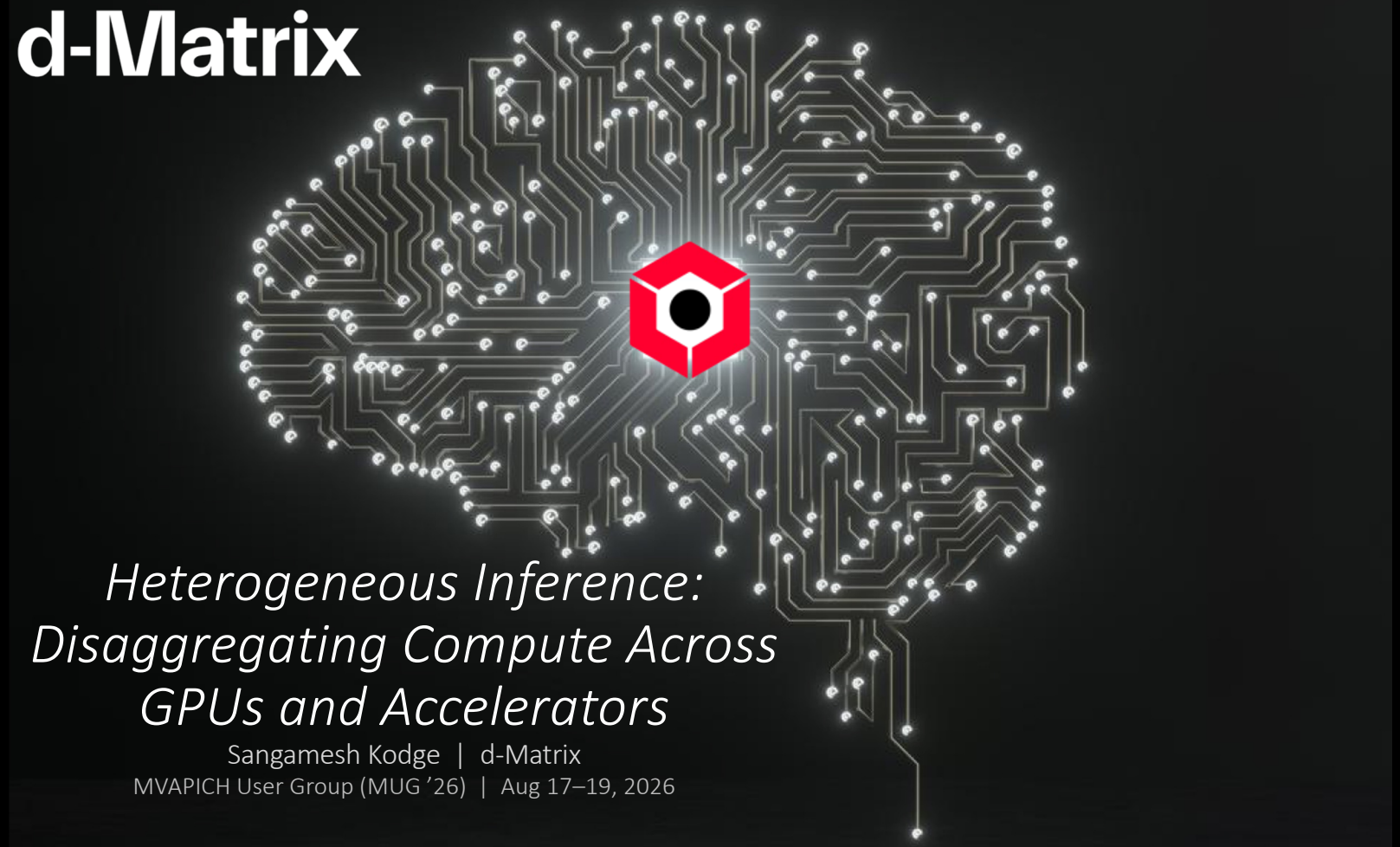
d-Matrix



Heterogeneous Inference: Disaggregating Compute Across GPUs and Accelerators

Sangamesh Kodge | d-Matrix

MVAPICH User Group (MUG '26) | Aug 17–19, 2026

AI: Skepticism to Reality

THE NEW YORK TIMES — MAY 2023

'The Godfather of A.I.' leaves Google and warns of danger ahead

"I thought it was 30 to 50 years or even longer away. Obviously, I no longer think that."

— Geoffrey Hinton

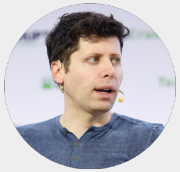
CNBC — APR 2025

Nadella: as much as 30% of Microsoft code is written by AI

"...maybe 20-30% of the code inside our repos today... written by software."

— Satya Nadella, LlamaCon

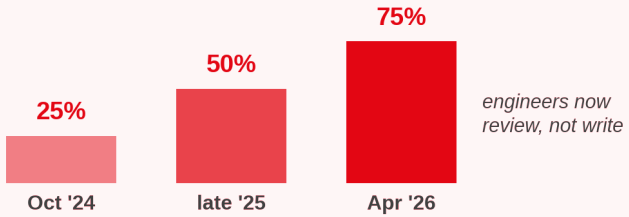
BLOG.SAMALTMAN.COM — JAN 2025



"We are now confident we know how to build AGI as we have traditionally understood it."

— Sam Altman, OpenAI CEO

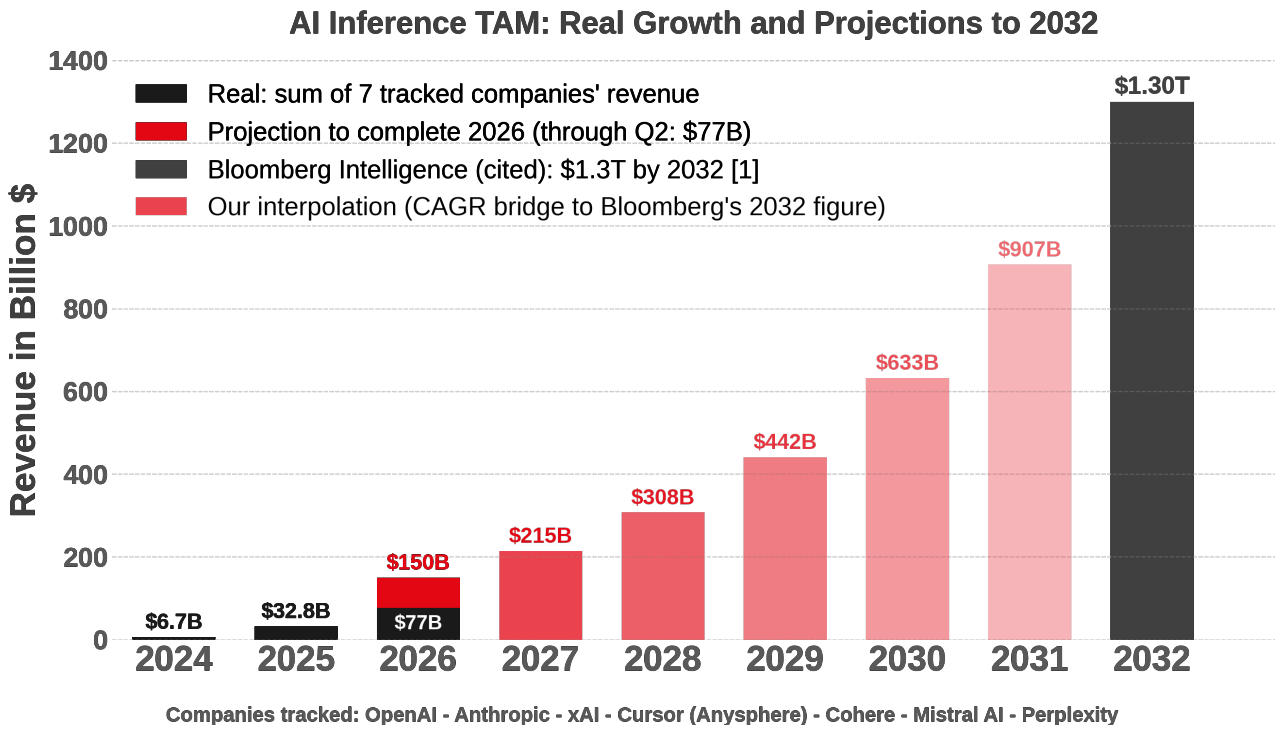
GOOGLE — SHARE OF NEW CODE WRITTEN BY AI



Sources: NYT (May 2023) - blog.samaltman.com (Jan 2025) - CNBC (Apr 2025) - Alphabet Q3'24 earnings & Cloud Next '26 (Pichai). Photos: Wikimedia Commons, CC BY-SA 4.0 / CC BY 2.0.

It happened faster than anyone thought

LLM Inference Market



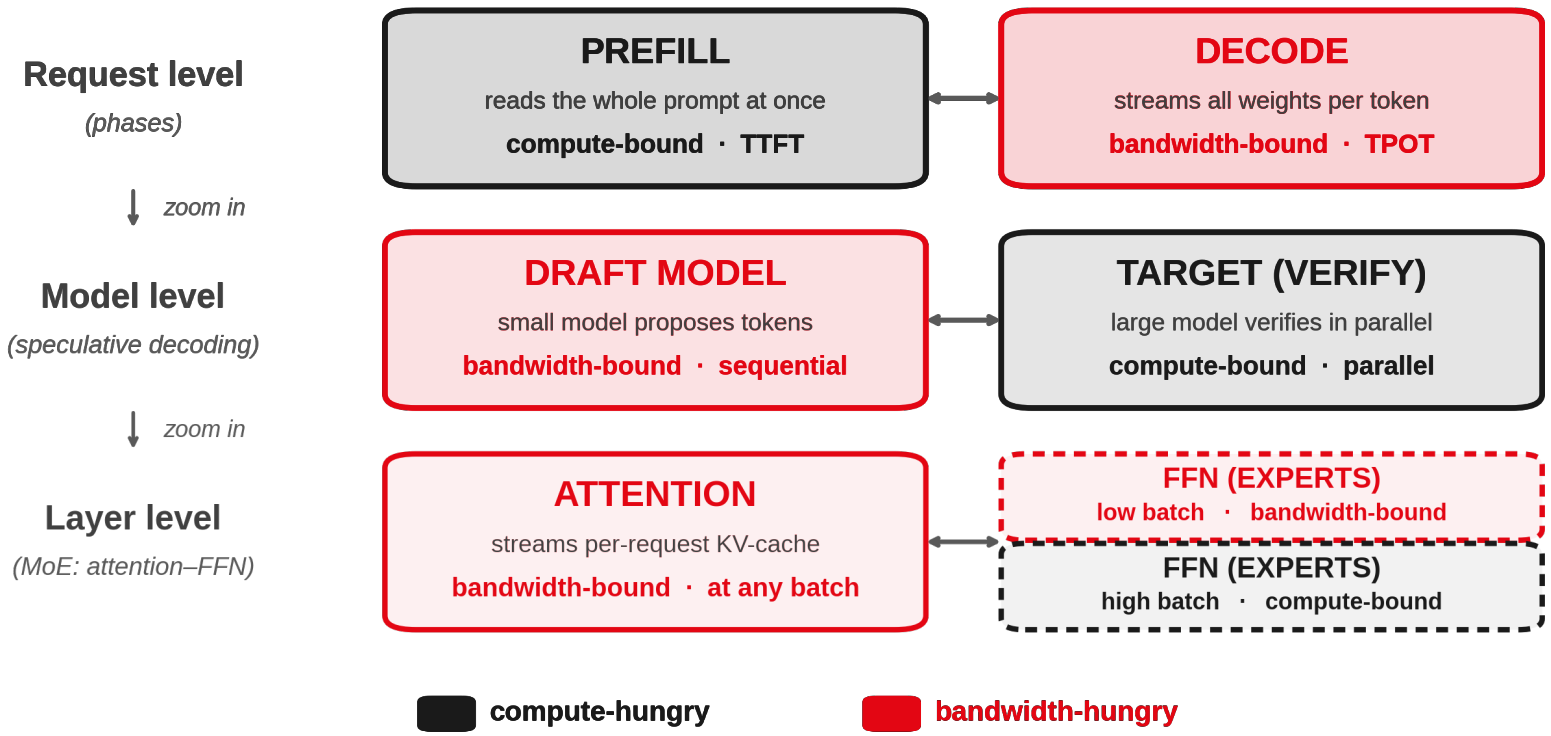
- LLM Inference revenue growth > 3x for 2 consecutive years.
- Bloomberg Intelligence : Generative AI inference market should grow beyond 1 Trillion by 2032 (32% CAGR).

Capturing this market isn't just about model quality — it's about serving economics.

At $1.3T scale, the winners will be whoever can serve tokens **cheaper** and **faster** than the rest.

[1] Bloomberg forecast of Inference market

Inference is not monolithic

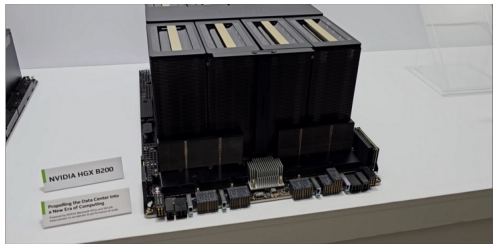


Inference is not one computation — at several levels we see opposite hardware demands.

Different Hardware designs

NVIDIA B200

compute-optimal GPU + HBM



MEMORY CAPACITY
192 GB (HBM3e)

MEMORY BANDWIDTH
8 TB/s

COMPUTE (DENSE)
4.5 PFLOPS

d-Matrix Corsair

bandwidth-optimal, digital in-memory compute

MEMORY CAPACITY
2 GB (on-chip SRAM)

MEMORY BANDWIDTH
150 TB/s

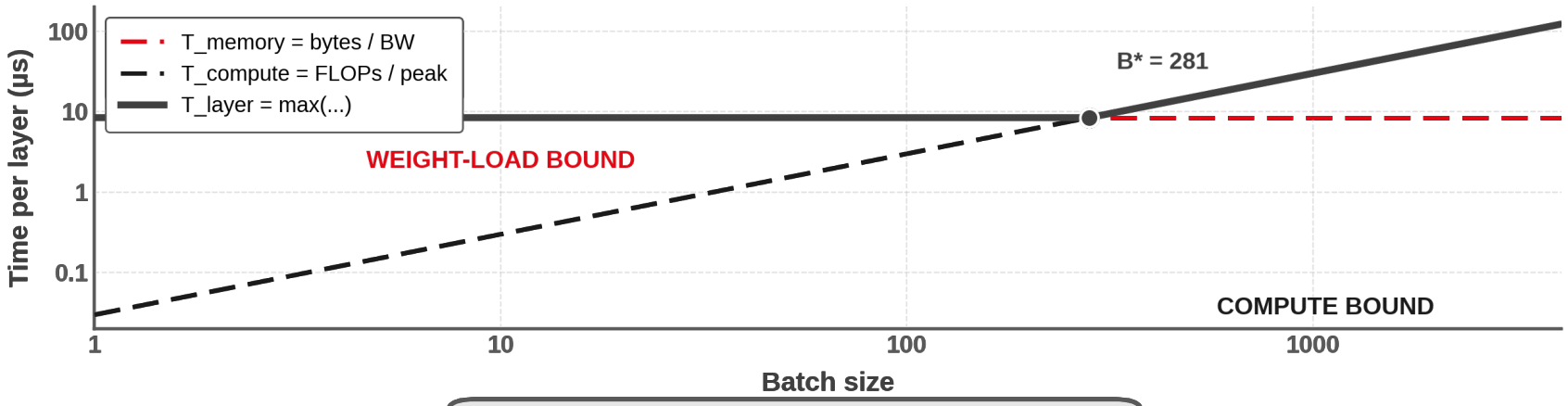
COMPUTE (DENSE)
2.4 PFLOPS

Opposite strengths — neither chip dominates.

Images: B200 image: Wikimedia Commons, Geekerwan (CC BY 3.0).

The Latency Equations

T_layer = max(T_memory , T_compute)



LAYER (MLP) — on B200

Weight params = 8192 × 8192 ≈ 67.1M

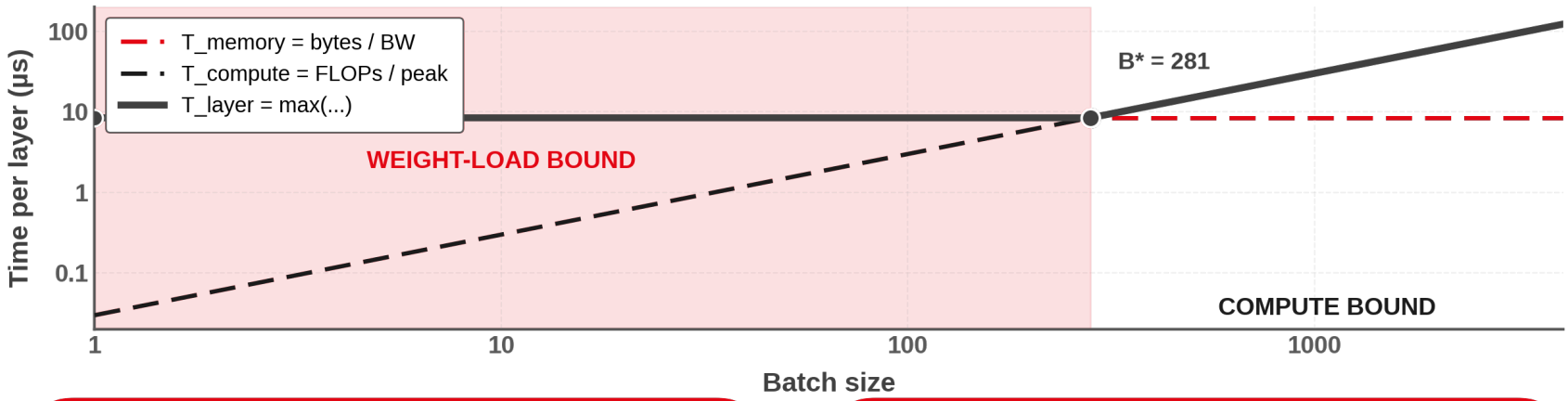
Precision = FP8 (1 B/param)

Weight bytes = 64 MiB

FLOPs/token = 2 × params ≈ 134M

The Latency Equations

T_layer = max(T_memory , T_compute)



BATCH = 1

T_memory = 8.4 µs

T_compute = 30 ns

T_layer = 8.4 µs
MEMORY-BOUND

BATCH = 2

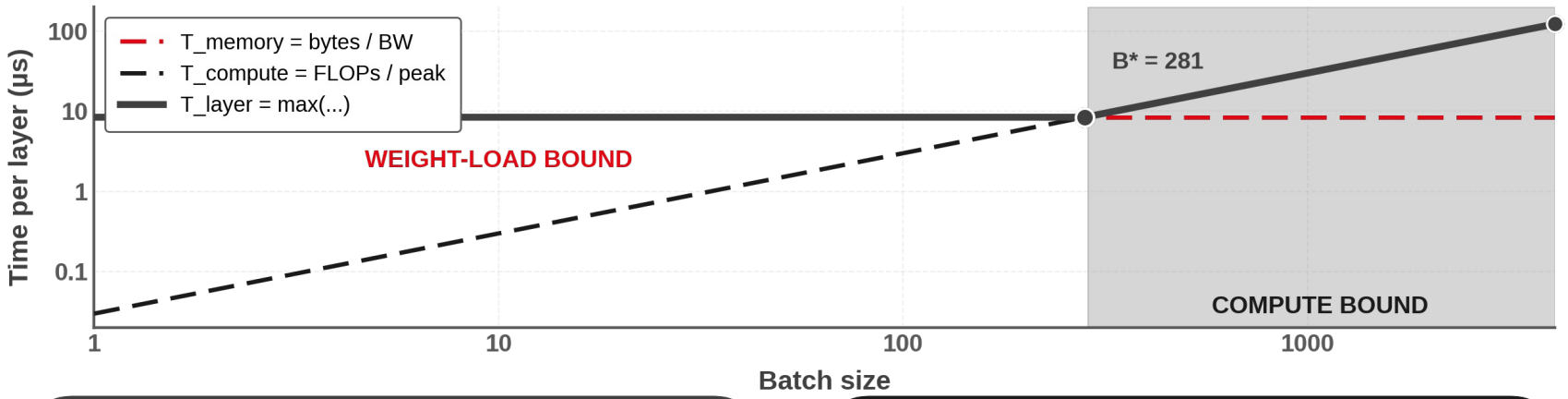
T_memory = 8.4 µs

T_compute = 60 ns

T_layer = 8.4 µs
MEMORY-BOUND

The Latency Equations

T_layer = max(T_memory , T_compute)



BATCH = 281 (crossover)

T_memory = 8.4 µs

T_compute = 8.4 µs

T_layer = 8.4 µs
TIE

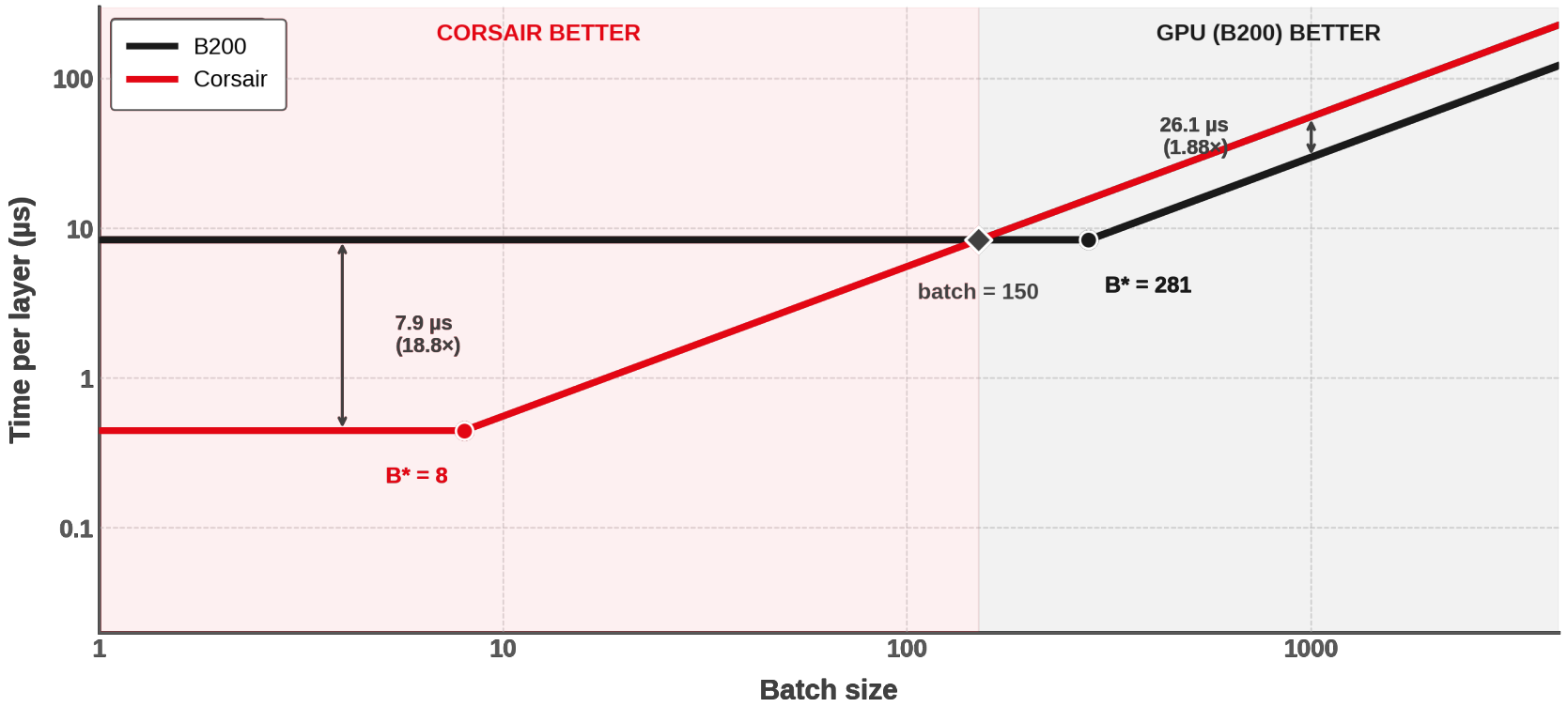
BATCH = 4096

T_memory = 8.4 µs

T_compute = 122.2 µs

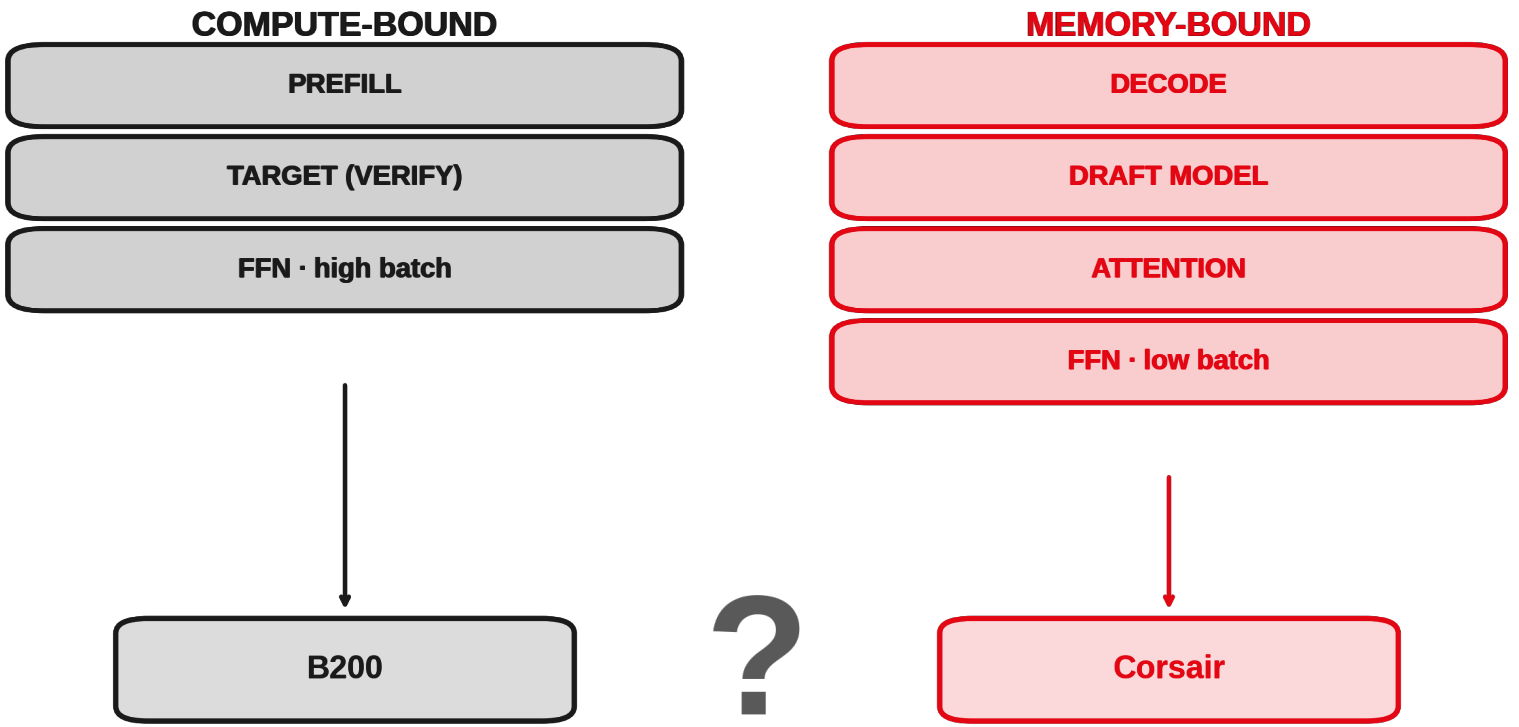
T_layer = 122.2 µs
COMPUTE-BOUND

The Latency Equations



Below batch 150, Corsair wins on bandwidth. Above it, B200's raw compute takes over.

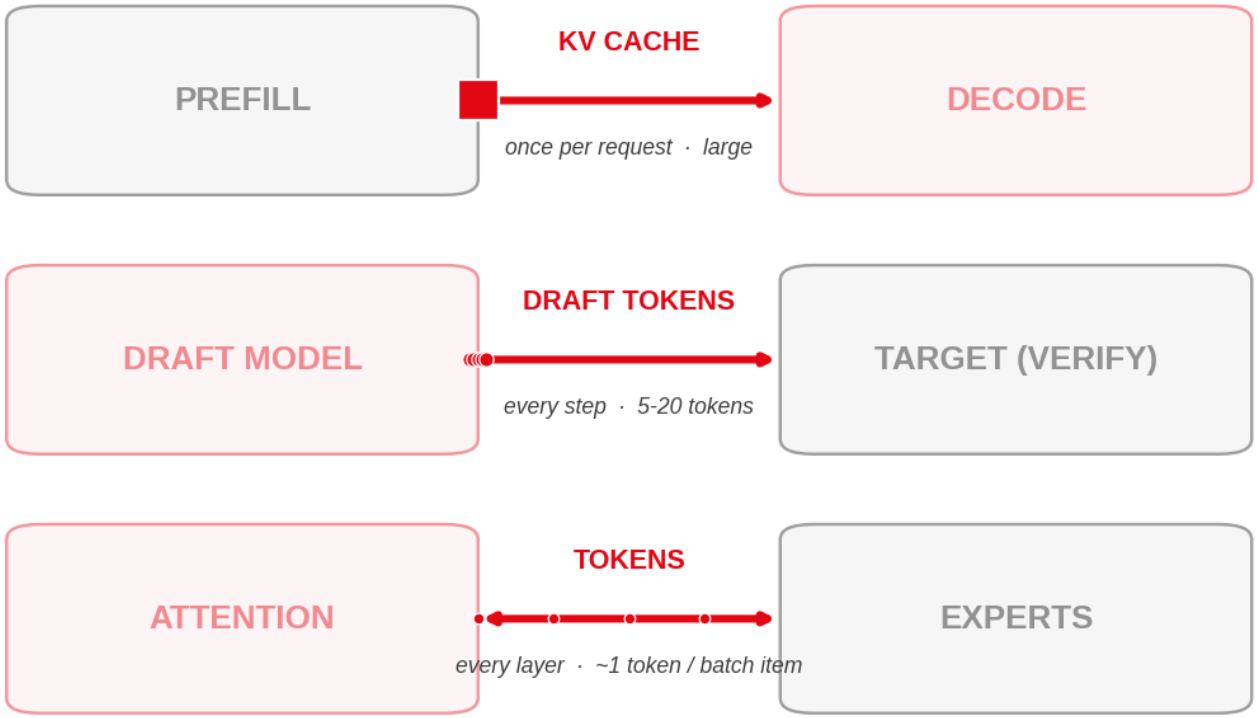
Disaggregated Inference



We just gave every phase, every model, every layer its perfect chip. Did we miss anything?

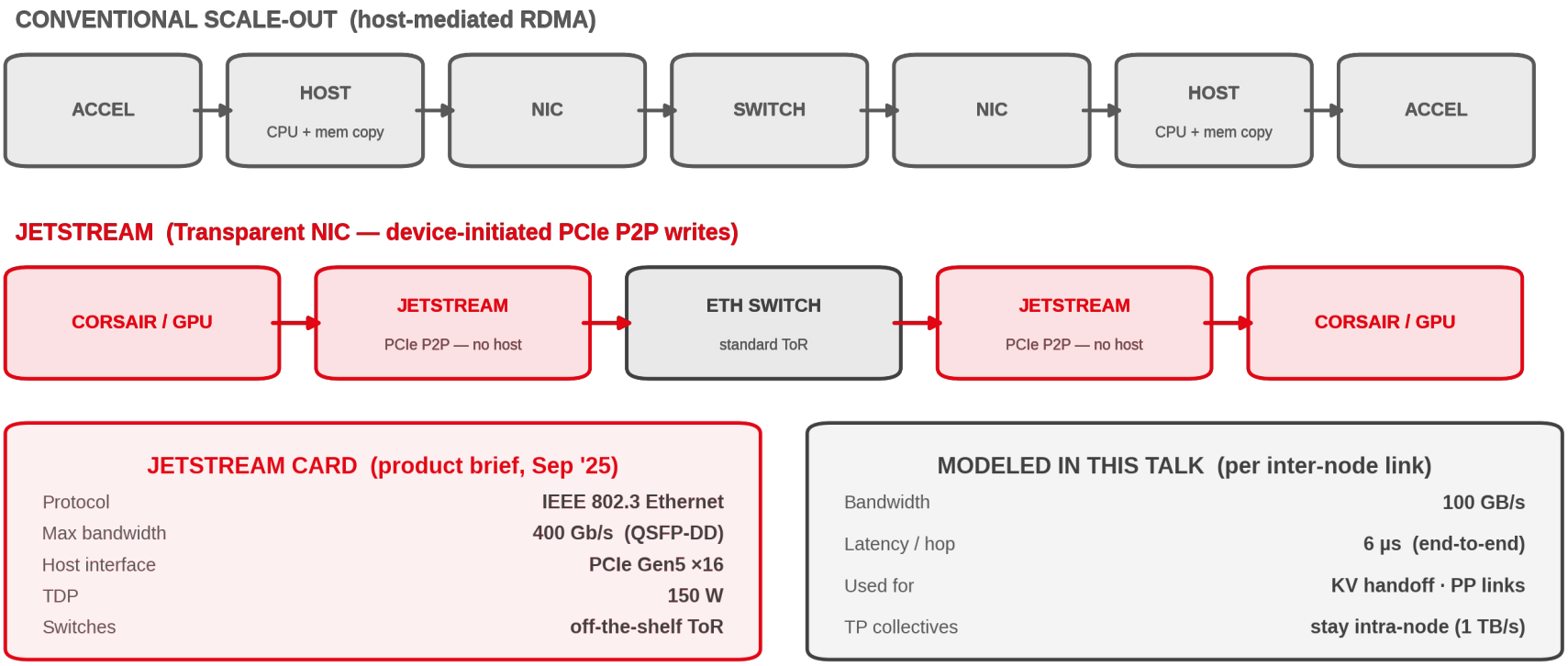
Disaggregated Inference Overheads

- Prefill → Decode: the entire KV cache moves, once per request.
- Draft → Target: 5-20 draft tokens per batch cross every single decoding step.
- Attention ↔ Experts: only ~1 token per batch item — but at every layer, every step.



Disaggregation isn't free — every split has associated communication overheads.

Communication for Disaggregation



JETSTREAM CARD (product brief, Sep '25)

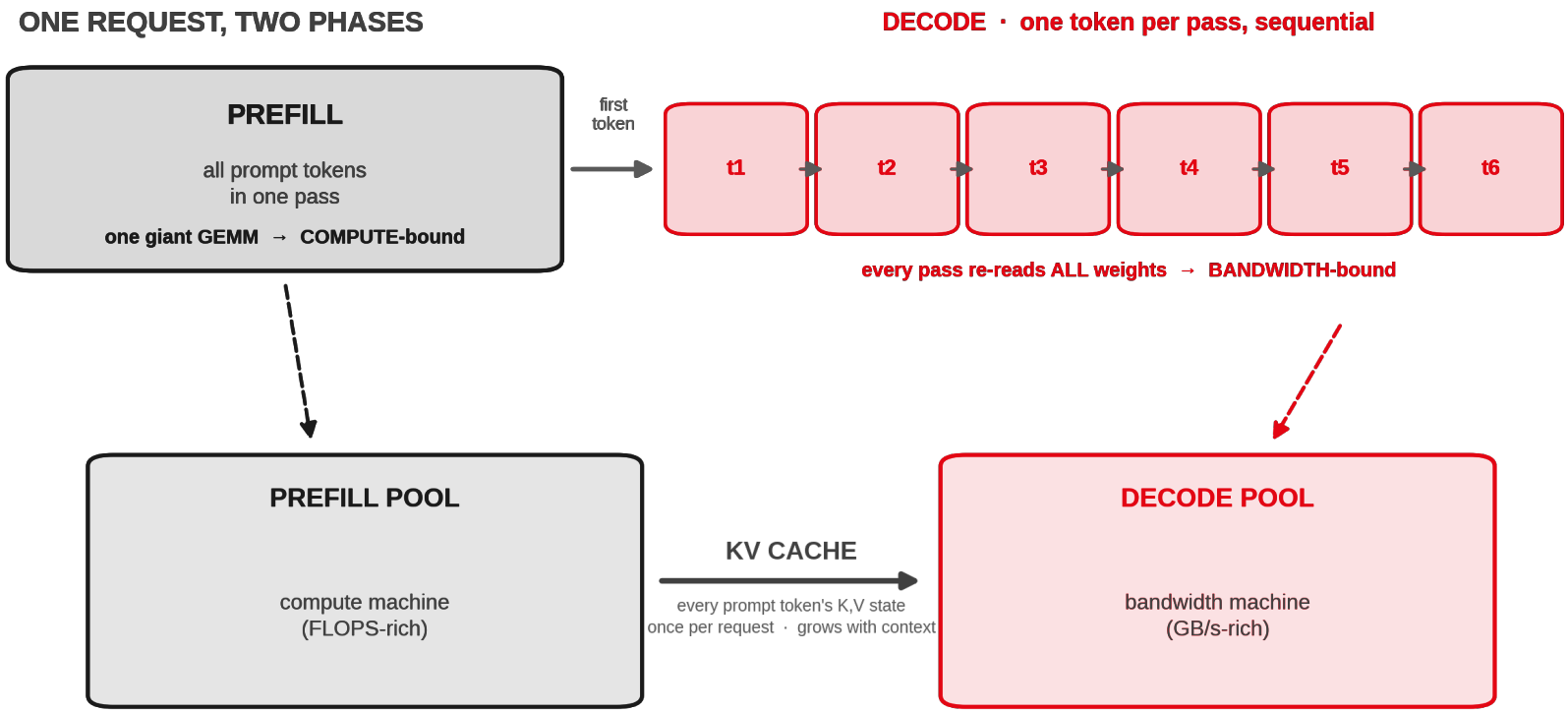
Protocol	IEEE 802.3 Ethernet
Max bandwidth	400 Gb/s (QSFP-DD)
Host interface	PCIe Gen5 ×16
TDP	150 W
Switches	off-the-shelf ToR

MODELED IN THIS TALK (per inter-node link)

Bandwidth	100 GB/s
Latency / hop	6 µs (end-to-end)
Used for	KV handoff · PP links
TP collectives	stay intra-node (1 TB/s)

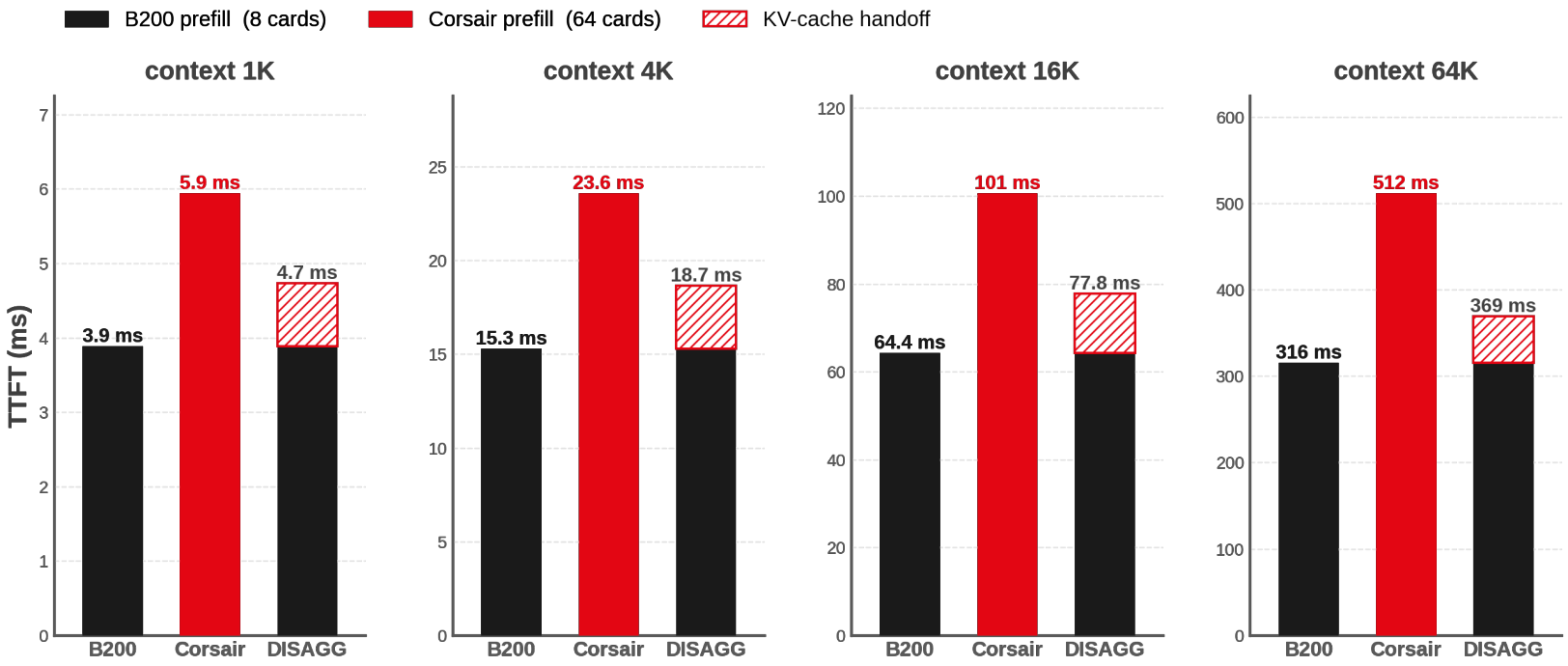
Interconnect Bandwidth and latency can become significant for some workloads, making disaggregation infeasible.

Prefill-Decode Disaggregation



KV Cache transfer required for Prefill-Decode Disaggregation.

Prefill-Decode Disaggregation : Results

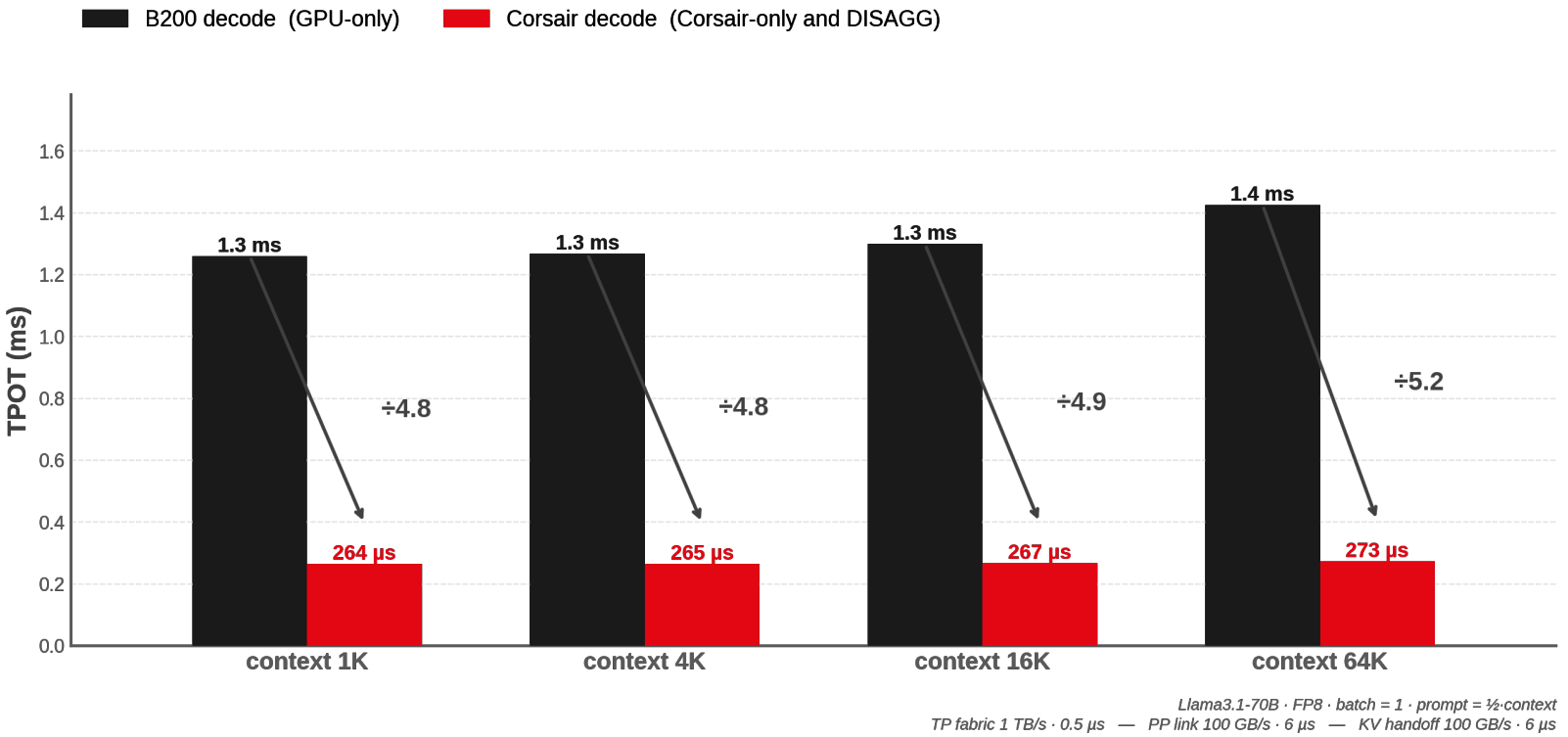


- Prefill being compute bound is >1.5x better on GPU.
- Disaggregation requires KV communication.
- B200 Prefill + KV cache communication faster than Corsair prefill

Llama3.1-70B · FP8 · batch = 1 · prompt = ½·context
TP fabric 1 TB/s · 0.5 µs — PP link 100 GB/s · 6 µs — KV handoff 100 GB/s · 6 µs

KV cache handoff is cheaper than processing prefill on Corsair

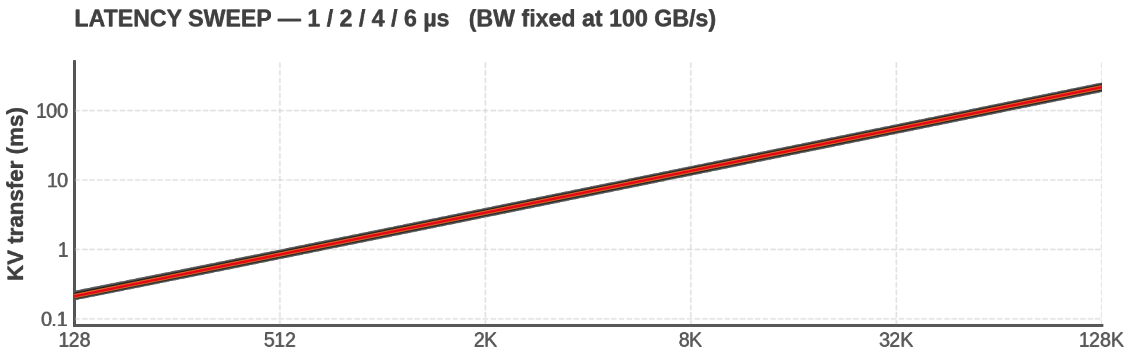
Prefill-Decode Disaggregation: Results



- Decode being memory bound, achieve much superior performance on Corsair.
- We see >4.8x better TPOT. TPOT improves with increase in context as the KV cache size increases.

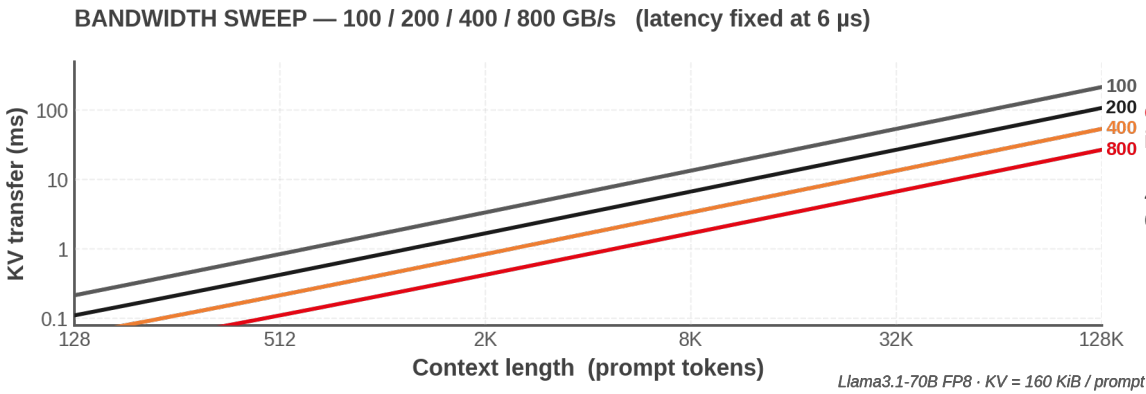
Prefill-Decode disaggregation achieves TTFT of GPU and TPOT of Corsair at marginal KV cache communication.

Prefill-Decode Disaggregation: Results



all four curves lie on top of each other

6 µs → 1 µs changes the transfer by < 3% at any context



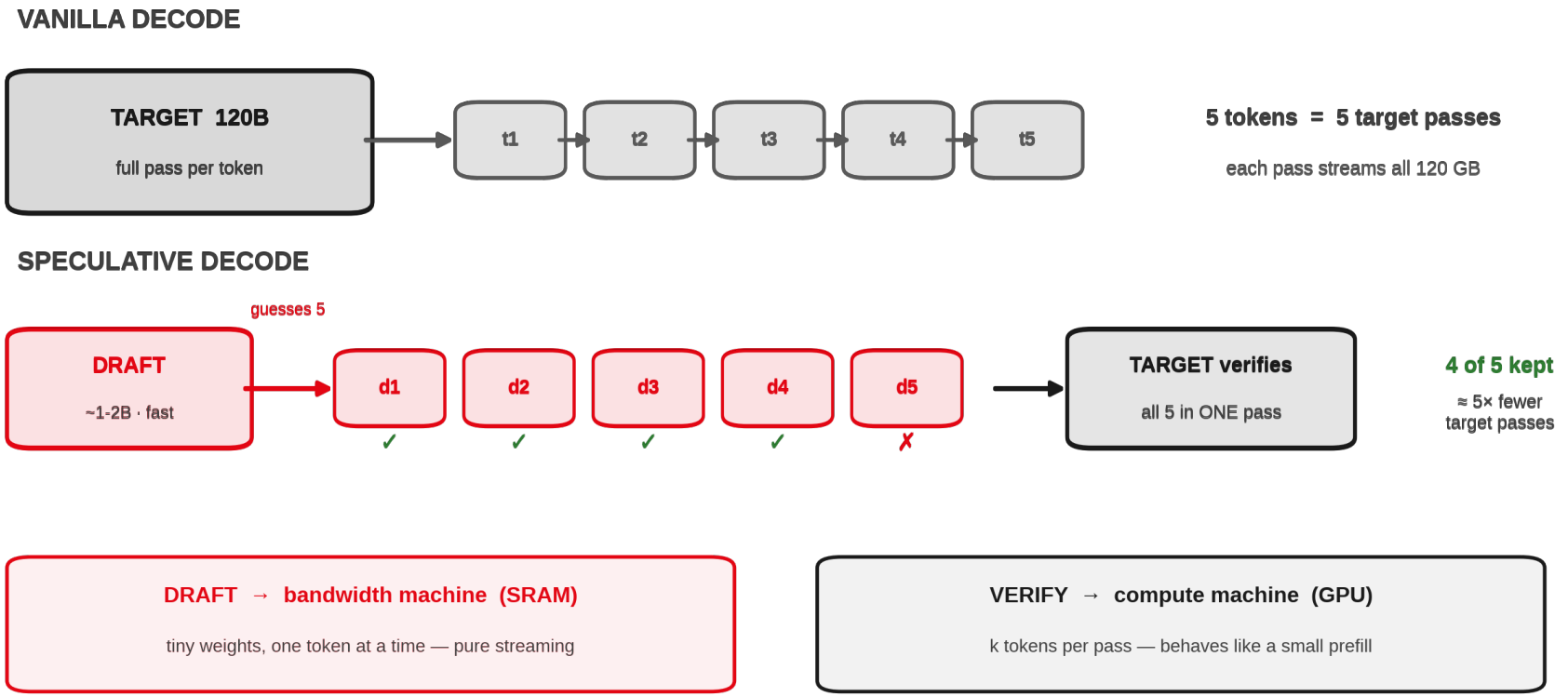
every 2× in bandwidth is 2× off the handoff

4K context: 0.67 GB moves in 6.7 ms @100 → 0.8 ms @800 GB/s

Llama3.1-70B FP8 · KV = 160 KiB / prompt token · transfer = KV bytes / BW + latency · single link, no overlap

KV Cache moved per request – Network Bandwidth sensitive.

Speculative Decoding

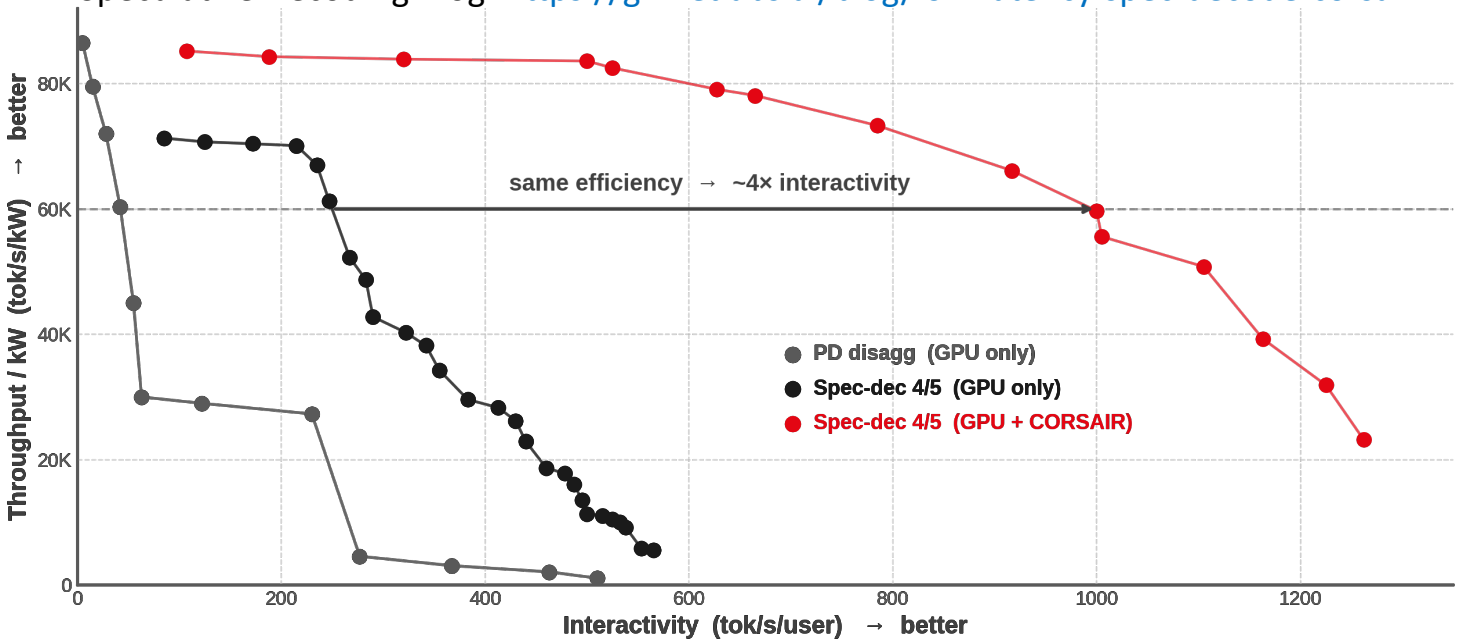


Speculative Decoding Blog: https://gimletlabs.ai/blog/low-latency-spec-decode-corsair

Speculative Decoding: Result

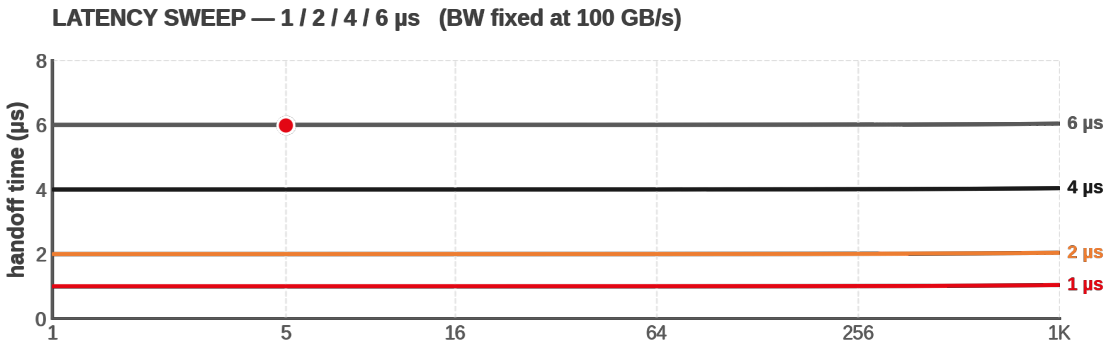
- PD baseline : **Increasing TPS/KW sharply drops Interactivity.**
- Spec-dec GPU only **improves the curve** but **Interactivity limited to 550 Tok/s/User.**
- Spec-dec with Corsair improve Interactivity to 1250 Tok/s/User while maintaining high TPS/KW for wide range of interactivity.
- At 60K TPS/KW Corsair spec-dec has 4x Superior interactivity over GPU spec-dec

Speculative Decoding Blog: https://gimletlabs.ai/blog/low-latency-spec-decode-corsair



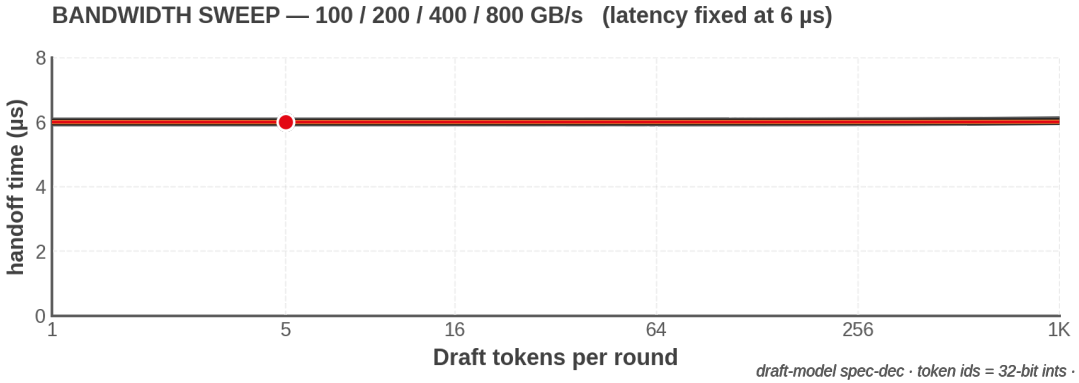
gpt-oss-120B target · 1.6B draft on 2 Corsair cards · 8K in / 1K out · 4-of-5 tokens accepted
Recreated from Gimlet Labs — gimletlabs.ai/blog/low-latency-spec-decode-corsair (points digitized)

Speculative Decoding: Result



handoff time IS the latency — 6 µs → 1 µs is a 6× win

● operating point: 5 tokens = 20 B
wire time @100 GB/s: 0.2 ns



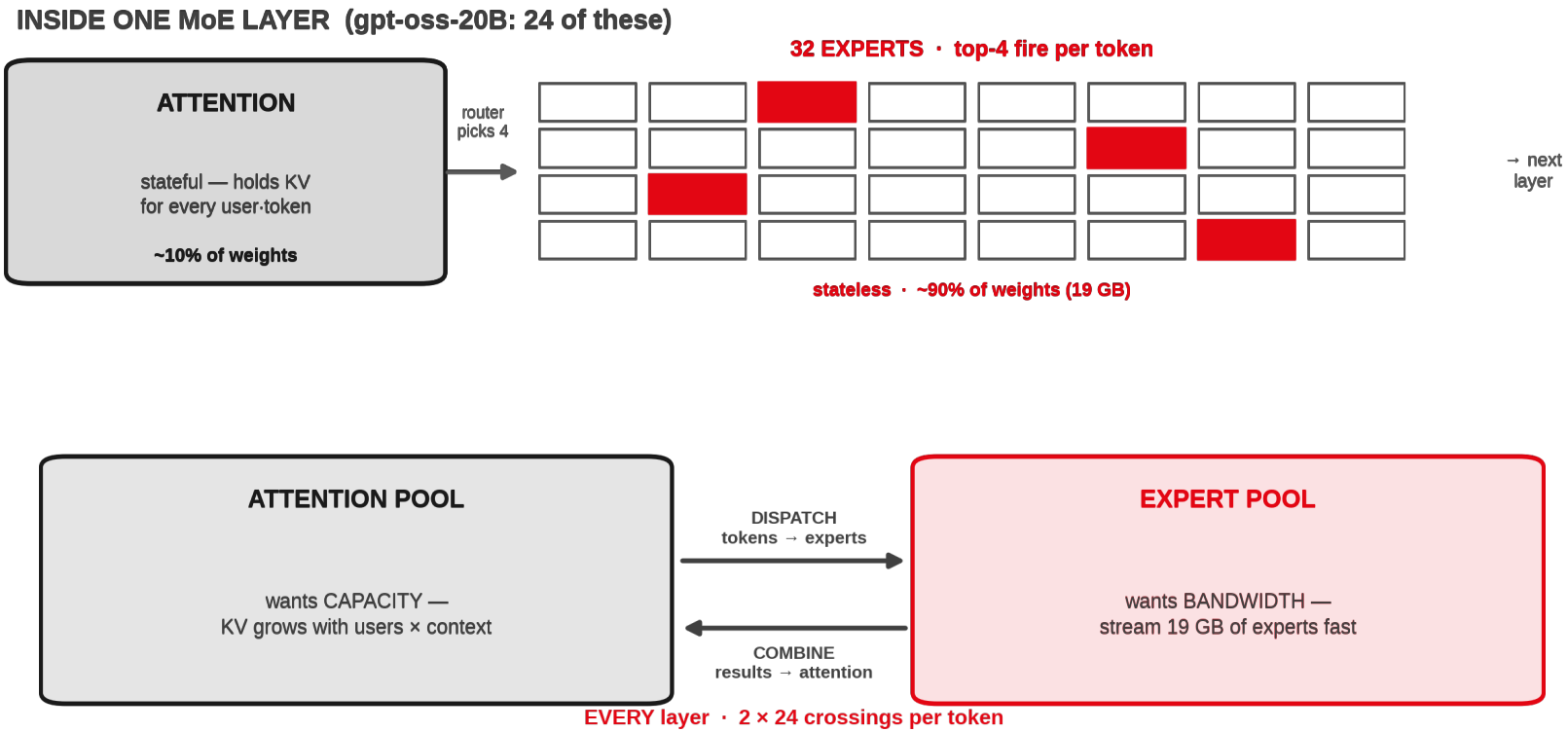
all four curves lie on top of each other — 8× the BW buys nothing

even 1K token ids = 4 KB:
wire time 40 ns vs 6 µs latency

draft-model spec-dec · token ids = 32-bit ints · handoff = tokens × 4 B / BW + latency · one round trip per draft round

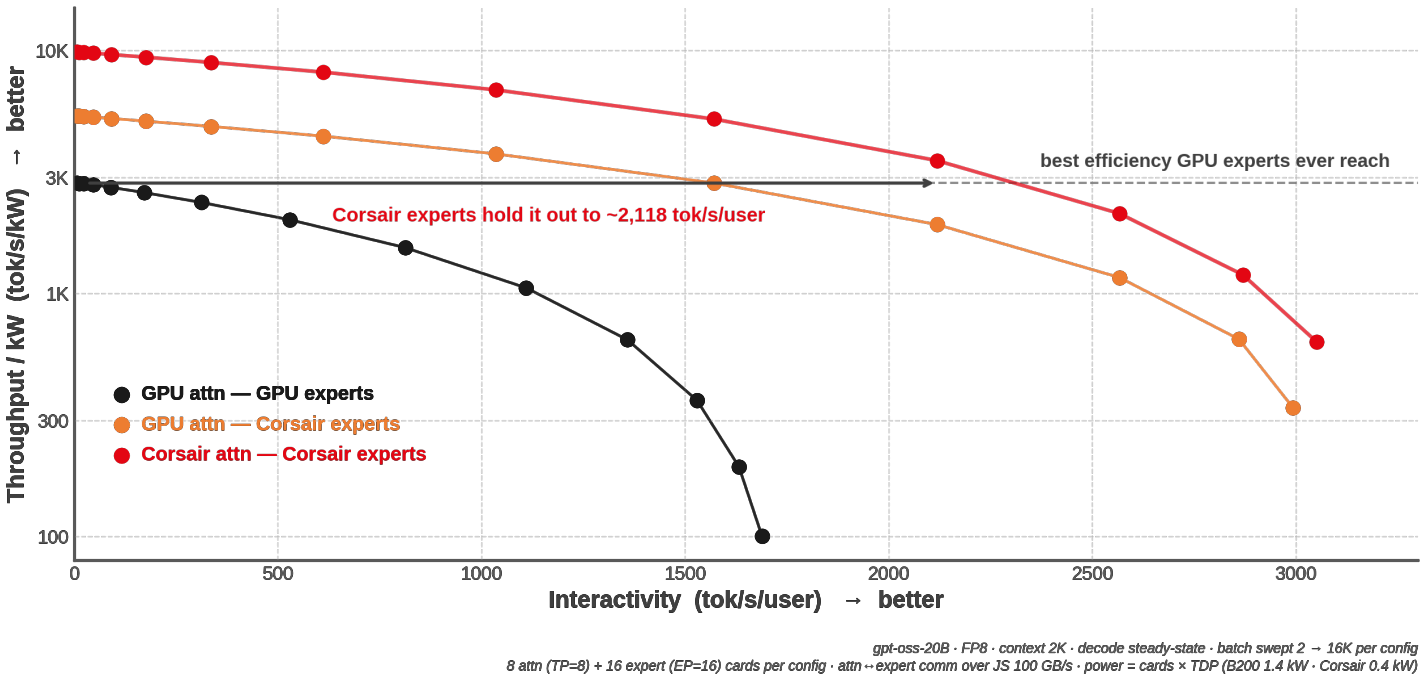
Draft token ids moved per speculation step– Network Latency sensitive.

AFD for MoEs



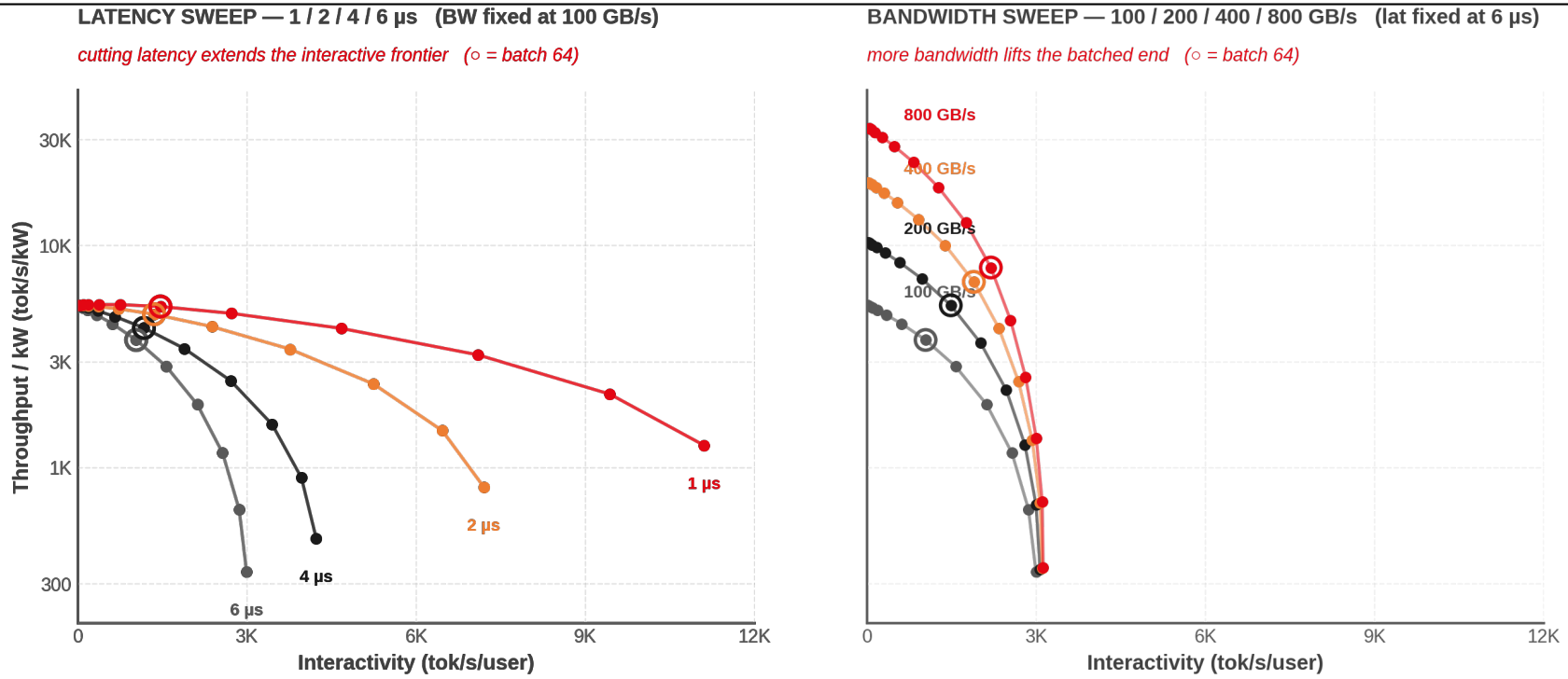
ISCA paper: https://ieeexplore.ieee.org/abstract/document/11617716

AFD Results



- GPU-only AFD hits an efficiency ceiling near **2.9K tok/s/kW**, and interactivity collapses past ~1,500 tok/s/user.
- Moving **only the experts** to Corsair lifts the whole frontier — experts stream 19 GB per token-step; they are the bandwidth problem.
- All-Corsair tops the chart: up to **~10K tok/s/kW** and **~3,050 tok/s/user**, on **9.6 kW vs 33.6 kW** for the all-GPU config.
- At the GPU's best-ever efficiency, Corsair experts hold that same efficiency out to **~2,100 tok/s/user** — interactivity the GPU config never reaches at any batch.

AFD Results



GPU attn — Corsair experts · gpt-oss-20B FP8 · decode steady-state · batch 2 → 16K · power = cards × TDP (17.6 kW)

Improved Network Latency improves Interactivity and Network Bandwithd improves TPS/KW

Workload dependent Network requirements

WORKLOAD	WHAT CROSSES	HOW OFTEN	VOLUME	NETWORK MUST DELIVER
PREFILL–DECODE	KV cache (prompt K,V state)	once per request	160 KiB / prompt tok 0.31 GB @ 4K ctx	**BANDWIDTH** latency forgiving — one bulk transfer, overlappable
SPEC-DEC	draft tokens + accept verdicts	every draft round (~5 tokens)	KBs per round	**LATENCY** round trip is the cost — long draft windows shift it to BW
AFD	activations to experts and results back	every MoE layer 2 × 24 per token	~0.8 MB per generated token	**LATENCY and BANDWIDTH** latency sets the interactive frontier; bandwidth sets the batched ceiling

PD: Llama3.1-70B FP8, batch 1 · Spec-dec: draft-model scheme · AFD: gpt-oss-20B, FP8 dispatch + FP16 combine, 24 MoE layers

The Industry moving towards Disaggregation

ORIGINS — ACADEMIA '24

DistServe: 7.4× goodput from PD split
Splitwise: different GPUs per phase

OSDI '24 · ISCA '24

SPEC-DEC — ON CORSAIR

Gimlet Labs: draft on Corsair, verify on GPU
→ 2–10× interactivity at the same energy

Mar 11 2026 · gimletlabs.ai

FRAMEWORKS & PRODUCTION

NVIDIA Dynamo: 30× DeepSeek-R1
DeepSeek serves 608B tokens/day disagg

GTC '25 · GA Mar '26 · DeepSeek

HETEROGENEOUS — d-MATRIX × PARASAIL

NVIDIA prefill + Corsair decode:
up to 10× tokens/s · 3× energy efficiency

Jul 7 2026 · d-matrix.ai · PRNewswire

GOING LIVE — SILICON VALLEY DATA CENTER

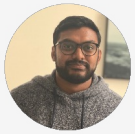
“The heterogenous insanity begins! d-Matrix going live in our data center”
— Mike Henry, CEO, Parasail

Aug 2026 · LinkedIn

Thank You

Sangamesh Kodge

skodge@d-matrix.ai



Nithesh Kurella

nitheshk@d-matrix.ai

Sudeep Bhoja

CTO

sbhoja@d-matrix.ai

Sai Rahul Chalamalasetti

sairahul@d-matrix.ai

David Karlov

dkarlov@d-matrix.ai

Questions welcome — come find us, or write to any of the above


d-Matrix