CPB-2026-002

GPU memory and bandwidth for LLM inference 2026: H100, H200 and L40S across nine open models

A derived analysis of model fit, KV-cache capacity and decode bandwidth ceilings for nine Qwen, Llama, DeepSeek and MiniMax models on H100, H200 and L40S GPUs.

VERSION
1.0

PUBLISHED
6 November 2026

UPDATED
6 November 2026

ONLINE VERSION
https://www.cloudproviderbenchmarks.com/reports/gpu-memory-bandwidth-llm-inference-2026/

Contents

Executive summary

1. Purpose and method

2. GPUs

3. Model fit and KV budget

4. Concurrency by context length

5. Decode ceilings

6. Availability and prices

7. Interpretation

8. Next version

Limitations

Version history

Sources

DIRECT ANSWER

At FP8, 7B–8B models fit on any of the three GPUs; 32B dense models fit on one H100 or H200 but leave little KV room on an L40S; 70B–72B models need one H200 or two H100s; DeepSeek-V3 needs a full 8-GPU H200 node. Decode ceilings scale with bandwidth: H200 ≈ 1.43× H100 SXM ≈ 5.6× L40S.

EXECUTIVE SUMMARY

This report applies a published memory model to nine open-weight LLMs — Qwen2.5 7B and 72B, Qwen3 32B and 30B-A3B, Llama 3.1 8B and 3.3 70B, DeepSeek-R1-Distill-Qwen-32B, DeepSeek-V3 and MiniMax-M1 — on NVIDIA H100, H200 and L40S. It reports the minimum GPU count for each model and precision, the KV-cache budget, the maximum concurrent sequences per context length and the bandwidth-bound decode ceiling. All values are derived from specifications and labelled as such.

KEY FINDINGS

01 Llama 3.3 70B at FP8 fits on one H200 with about 49 GB of KV budget (about 18 concurrent 8K-token sequences) and on two H100s with about 65 GB.

02 Qwen3 32B at FP8 on one H100 holds about 16 concurrent 8K sequences; on one H200, about 40.

03 Qwen3 30B-A3B's FP8 decode ceiling is about 1,015 tokens/s on H100 SXM because only ~3.3B parameters are active per token.

04 On L40S, a 32B FP8 model leaves about 8 GB of KV budget — three 8K sequences — which limits its use to small models or multi-card setups.

05 Derived values only; measured results are reported on the companion benchmark pages.

1. Purpose and method

Choosing a GPU configuration for LLM inference starts with memory: do the weights fit, and how much KV cache remains for concurrent requests? This report answers that for nine models and three GPUs using the GPU and LLM inference methodology:

- weights = parameters × bytes per parameter;
- KV per token = 2 × layers × KV heads × head dimension × 2 bytes;
- usable memory = 85% of device memory (90% utilisation − 5% reserve);

- decode ceiling = memory bandwidth ÷ active weight bytes.

2. GPUs

GPU specifications (NVIDIA datasheets)

GPU	ARCHITECTURE	MEMORY (GB)	MEMORY TYPE	BANDWIDTH (GB/S)	FP8 (TFLOPS)	BF16 (TFLOPS)	FP32 (TFLOPS
H100 SXM5 SXM5	Hopper	80	HBM3	3,350	1,979	989	6
H200 SXM SXM5	Hopper	141	HBM3e	4,800	1,979	989	6
L40S PCIe dual-slot	Ada Lovelace	48	GDDR6 (ECC)	864	733	362	91.

Tensor values are dense (non-sparse) peaks; NVIDIA publishes sparse figures at twice these values.

3. Model fit and KV budget

Model memory requirements and minimum GPU count (FP8 weights)

MODEL	PARAMS (B)	BF16 (GB)	FP8 (GB)	INT4 (GB)	KV / TOKEN (KIB)	KV @32K (GB)	MIN H100 SXM5	M
Llama 3.1 8B Instruct Dense · GQA	8	16	8	4	128	4.3	1×	
Llama 3.3 70B Instruct Dense · GQA	70.6	141	71	35	320	10.7	2×	
Qwen2.5 7B Instruct Dense · GQA	7.6	15	8	4	56	1.9	1×	
Qwen3 32B Dense · GQA	32.8	66	33	16	256	8.6	1×	
Qwen3 30B-A3B MoE · 3.3B active · GQA	30.5	61	31	15	96	3.2	1×	
Qwen2.5 72B Instruct Dense · GQA	72.7	145	73	36	320	10.7	2×	
DeepSeek-R1-Distill-Qwen-32B Dense · GQA	32.8	66	33	16	256	8.6	1×	
DeepSeek-V3 MoE · 37B active · MLA	671	1,342	671	336	69	2.3	>8	
MiniMax-M1-80k MoE · 45.9B active · HYBRID-LINEAR	456	912	456	228	–	–	8×	

Weight size = parameters × bytes per parameter. KV = 2 × layers × KV heads × head dim × 2 bytes (latent size for MLA). Minimum GPUs: 85% usable memory and at least 4 GB of KV headroom per GPU. A capacity model, not a measurement.

4. Concurrency by context length

The following matrices show which concurrency levels fit at each context length for representative configurations.

Llama 3.1 8B Instruct · FP8 · 1× L40S

Executable scenario matrix (KV-cache capacity model) · KV budget ≈ 32.8 GB

CONTEXT	1	5	10	25	50	100	MAX CONC.
8K	✓	✓	✓	✓	–	–	30
16K	✓	✓	✓	–	–	–	15
32K	✓	✓	–	–	–	–	7
64K	✓	–	–	–	–	–	3
128K	✓	–	–	–	–	–	1

✓ scenario fits the memory model · — KV cache exceeded (not executed) · ✕ weights do not fit · n/a context above model maximum. Model: 90% memory utilisation, 5% runtime reserve, 16-bit KV cache.

Qwen3 32B · FP8 · 1× H100 SXM5

Executable scenario matrix (KV-cache capacity model) · KV budget ≈ 35.2 GB

CONTEXT	1	5	10	25	50	100	MAX CONC.
8K	✓	✓	✓	–	–	–	16
16K	✓	✓	–	–	–	–	8
32K	✓	–	–	–	–	–	4
64K	✓	–	–	–	–	–	2
128K	✓	–	–	–	–	–	1

✓ scenario fits the memory model · — KV cache exceeded (not executed) · ✕ weights do not fit · n/a context above model maximum. Model: 90% memory utilisation, 5% runtime reserve, 16-bit KV cache.

Qwen3 32B · FP8 · 1× H200 SXM

Executable scenario matrix (KV-cache capacity model) · KV budget ≈ 87.0 GB

CONTEXT	1	5	10	25	50	100	MAX CONC.
8K	✓	✓	✓	✓	–	–	40
16K	✓	✓	✓	–	–	–	20
32K	✓	✓	✓	–	–	–	10
64K	✓	✓	–	–	–	–	5
128K	✓	–	–	–	–	–	2

✓ scenario fits the memory model · — KV cache exceeded (not executed) · ✕ weights do not fit · n/a context above model maximum. Model: 90% memory utilisation, 5% runtime reserve, 16-bit KV cache.

Llama 3.3 70B Instruct · FP8 · 1× H200 SXM

Executable scenario matrix (KV-cache capacity model) · KV budget ≈ 49.3 GB

CONTEXT	1	5	10	25	50	100	MAX CONC.
8K	✓	✓	✓	–	–	–	18
16K	✓	✓	–	–	–	–	9
32K	✓	–	–	–	–	–	4
64K	✓	–	–	–	–	–	2
128K	✓	–	–	–	–	–	1

✓ scenario fits the memory model · — KV cache exceeded (not executed) · ✕ weights do not fit · n/a context above model maximum. Model: 90% memory utilisation, 5% runtime reserve, 16-bit KV cache.

Llama 3.3 70B Instruct · FP8 · 2× H100 SXM5

Executable scenario matrix (KV-cache capacity model) · KV budget ≈ 65.4 GB

CONTEXT	1	5	10	25	50	100	MAX CONC.
8K	✓	✓	✓	–	–	–	24
16K	✓	✓	✓	–	–	–	12
32K	✓	✓	–	–	–	–	6
64K	✓	–	–	–	–	–	3
128K	✓	–	–	–	–	–	1

✓ scenario fits the memory model · — KV cache exceeded (not executed) · ✕ weights do not fit · n/a context above model maximum. Model: 90% memory utilisation, 5% runtime reserve, 16-bit KV cache.

5. Decode ceilings

Memory-bandwidth ceiling for single-stream decode (FP8, tok/s)

Theoretical upper bound — not a measurement

MODEL	H100 SXM5	H200 SXM	L40S
Llama 3.1 8B Instruct	417	598	108
Llama 3.3 70B Instruct	95 2× TP	68	24 2× TP
Qwen2.5 7B Instruct	440	631	114
Qwen3 32B	102	146	26
Qwen3 30B-A3B	1,015	1,455	262
Qwen2.5 72B Instruct	92 2× TP	66	24 2× TP
DeepSeek-R1-Distill-Qwen-32B	102	146	26
DeepSeek-V3	does not fit	1,038 8× TP	does not fit

Ceiling = aggregate memory bandwidth ÷ active weight bytes read per token (batch 1, excluding KV reads and communication, assuming ideal tensor parallelism). Real engines reach a fraction of this bound.

Memory-bandwidth ceiling for single-stream decode (BF16, tok/s)

Theoretical upper bound — not a measurement

MODEL	H100 SXM5	H200 SXM	L40S
Llama 3.1 8B Instruct	209	299	54
Llama 3.3 70B Instruct	95 4× TP	68 2× TP	24 4× TP
Qwen2.5 7B Instruct	220	315	57
Qwen3 32B	102 2× TP	73	26 2× TP
Qwen3 30B-A3B	508	727	262 2× TP
Qwen2.5 72B Instruct	92 4× TP	66 2× TP	24 4× TP
DeepSeek-R1-Distill-Qwen-32B	102 2× TP	73	26 2× TP
DeepSeek-V3	does not fit	does not fit	does not fit

Ceiling = aggregate memory bandwidth ÷ active weight bytes read per token (batch 1, excluding KV reads and communication, assuming ideal tensor parallelism). Real engines reach a fraction of this bound.

6. Availability and prices

GPU availability and published per-GPU-hour list prices by provider

PROVIDER	NVIDIA H100	NVIDIA H200	NVIDIA L40S	TÜRKIYE IN
PortvMind Cloud	Generally available gpu.h100s / gpu.h100m profiles TRY 195.835 / GPU-h 1, 2, 3 or 4 GPUs per server; 80 GB (81,920 MB) VRAM per GPU; 26 vCPU and 496 GB RAM per GPU; local NVMe disk included.	Generally available GPU / AI Compute (H200 option) Listed as an available GPU option; configurations and prices are not included in the public calculator table and are provided via the console or sales.	Generally available GPU / AI Compute (L40S option) Listed as an available GPU option; configurations and prices are not included in the public calculator table and are provided via the console or sales.	Yes
Amazon Web Services	Generally available EC2 P5 Single-GPU p5.4xlarge and 8-GPU p5.48xlarge sizes; availability varies by region.	Generally available EC2 P5e / P5en 8-GPU instances; availability varies by region.	Generally available EC2 G6e 1 to 8 L40S GPUs per instance.	Yes
Microsoft Azure	Generally available NC H100 v5 / ND H100 v5 NC H100 v5 with one or two H100 NVL GPUs; ND H100 v5 with eight H100 GPUs.	Generally available ND H200 v5 Eight H200 GPUs per VM.	Not offered L40S SKUs are documented for Azure Local / Azure Stack Hub (customer-site hardware).	No
Google Cloud	Generally available A3 High / A3 Mega Up to 8 H100 GPUs per VM; zone availability varies.	Generally available A3 Ultra 8 H200 GPUs per VM; zone availability varies.	Not offered	Announce
DigitalOcean	Generally available GPU Droplet — NVIDIA HGX H100 $4.410 / GPU-h 1× or 8× H100 80 GB; 1× plan with 20 vCPU and 240 GiB RAM.	Generally available GPU Droplet — NVIDIA HGX H200 $4.470 / GPU-h 1× or 8× H200 141 GB; 1× plan with 24 vCPU and 240 GiB RAM.	Generally available GPU Droplet — NVIDIA L40S $1.570 / GPU-h 1× L40S 48 GB with 8 vCPU and 64 GiB RAM.	No
Vultr	Generally available Cloud GPU — NVIDIA HGX H100 $2.300 / GPU-h 8× H100 80 GB SXM (640 GB GPU RAM, 216 vCPU, 1,914 GB RAM).	Unverified Not listed on the public pricing page H200 was not listed on the Vultr pricing page at the last review date.	Generally available Cloud GPU — NVIDIA L40S $1.671 / GPU-h 1, 2, 4 or 8× L40S 48 GB; 1× plan with 16 vCPU and 180 GB RAM.	No
Hetzner	Not offered	Not offered	Not offered	No

Published per-GPU-hour list prices

PROVIDER	GPU	PER GPU-HOUR	PER GB-HOUR	BASIS
PortvMind	NVIDIA H100	TRY 195.835	TRY 2.4479	gpu.h100s.xlarge (1× H100) hourly list price, tr-ist-01, excluding tax. Multi-GPU profiles price from ₺141.12 per GPU-hour (3× and 4× profiles). · 6 November 2026
DigitalOcean	NVIDIA H100	$4.410	$0.0551	On-demand, 1× H100 GPU Droplet. · 6 November 2026
DigitalOcean	NVIDIA H200	$4.470	$0.0317	On-demand, 1× H200 GPU Droplet. · 6 November 2026
DigitalOcean	NVIDIA L40S	$1.570	$0.0327	On-demand, 1× L40S GPU Droplet. · 6 November 2026
Vultr	NVIDIA H100	$2.300	$0.0287	On-demand "starting at" per-GPU rate, 8-GPU HGX configuration. · 6 November 2026
Vultr	NVIDIA L40S	$1.671	$0.0348	On-demand per-GPU rate, 1× L40S. · 6 November 2026

7. Interpretation

- **Memory determines configuration; bandwidth determines speed.** The H200's advantage over the H100 is entirely in these two numbers.
- **MoE models change the trade-off.** Qwen3 30B-A3B needs the memory of a 30B model but decodes like a 3B model.
- **Long contexts are expensive.** At 128K tokens, even an 8B model on an H100 supports only three concurrent sequences with a 16-bit KV cache.
- **The L40S suits small models.** Above ~14B parameters at FP8, its 48 GB constrains concurrency severely.

8. Next version

Version 2.0 of this report will add measured TTFT, TPOT, throughput and cost per million tokens from verified vLLM runs on each tracked provider where GPUs are available.

Limitations

- All figures are derived from specifications using the stated formulas; they are planning bounds, not measurements.
- A 16-bit KV cache without prefix sharing is assumed.
- MiniMax-M1's hybrid attention is not covered by the KV formula.
- SXM variants are modelled for H100 and H200; PCIe and NVL variants differ.

Version history

1.0	6 November 2026	First edition. Derived analysis only; measured results to follow in a later version.

Sources

[1] NVIDIA H100 Tensor Core GPU — NVIDIA Vendor datasheet
accessed: 6 November 2026

[2] NVIDIA H200 Tensor Core GPU — NVIDIA Vendor datasheet
accessed: 6 November 2026

[3] NVIDIA L40S GPU — NVIDIA Vendor datasheet
accessed: 6 November 2026

[4] Llama 3.3 70B Instruct — Meta Official documentation
accessed: 6 November 2026

[5] Llama 3.1 8B Instruct — Meta Official documentation
accessed: 6 November 2026

[6] Qwen3 32B — Qwen Official documentation
accessed: 6 November 2026

[7] Qwen3 30B-A3B — Qwen Official documentation
accessed: 6 November 2026

[8] Qwen2.5 72B Instruct — Qwen Official documentation
accessed: 6 November 2026

[9] Qwen2.5 7B Instruct — Qwen Official documentation
accessed: 6 November 2026

[10] DeepSeek-V3 — DeepSeek Official documentation
accessed: 6 November 2026

[11] DeepSeek-R1-Distill-Qwen-32B — DeepSeek Official documentation
accessed: 6 November 2026

[12] MiniMax-M1-80k — MiniMax Official documentation
accessed: 6 November 2026

[13] vLLM documentation — vLLM project Official documentation
accessed: 6 November 2026