The SLO Frontier per GPU Generation

At interactive chat a current flagship GPU serves 2 to 3× the tokens of a previous-generation one, and an older inference GPU serves none

Pranav Varshney
Nodus Compute

Viswa Kotra
Nodus Compute

Abstract

Fleets buy and plan GPUs by the hour and compare generations by peak throughput, but what a GPU-hour delivers to users depends on the latency target the work must meet. Nodus, the infrastructure execution layer for AI workloads, measured it directly. On an older inference GPU, a previous-generation training GPU and a current flagship GPU, each serving Qwen3-8B with the same stock vLLM 0.30.0 configuration on production chat and long-prompt traces, we found the highest load at which each GPU meets each of three SLO tiers, over 3 seeds with an A/A on every GPU. On chat at the interactive tier, a p95 TTFT of 0.5 s and a p95 TPOT of 30 ms, the flagship GPU served 2357 SLO-met output tokens per GPU-second against 852 for the previous-generation GPU, 2.77×, or 2.09× by the highest load met, and 2.18× when only completion counts. An independent verifier re-measured the previous-generation GPU on 4 fresh seeds and put the token ratio between 2.16× and 4.15× at 95% confidence, so the flagship GPU serves roughly 2 to 3 times the previous-generation GPU's interactive chat work. The older GPU served none at the interactive tier, because its TPOT with one request at a time is already 33.6 ms, and neither the older nor the previous-generation GPU served any long-prompt tokens at that tier. At every SLO knee the previous-generation and flagship GPUs read 95.1% to 99.5% busy in nvidia-smi, while DCGM showed 18.0% to 28.6% of the flagship GPU's SM time idle, and holding the SLO gave up 9.3% to 38.8% of the throughput the same flagship GPU reached when saturated. Qwen3-1.7B on the older GPU served as many interactive chat tokens per GPU-hour as Qwen3-8B on the previous-generation GPU.

1 Introduction

A GPU-hour is the unit in which fleets buy and plan capacity, and GPU generations are usually compared by peak throughput or by benchmarks with one fixed latency constraint [1] [2]. Users do not consume peak throughput. They consume tokens that arrive within a latency target, and the target depends on the product: an interactive assistant needs a fast first token and a fast stream, while a batch summary needs only to finish. Prior work showed that request size, rate and SLO decide the cheapest GPU type for a workload [3] and that offered load sets the cost per token [4].

We measure the frontier directly on three GPU generations that are still deployed in volume, with the same engine, the same model and the same production traces, at three SLO tiers. The result is a table of SLO-met tokens per GPU-hour that a fleet can plan with, the price ratio at which an older GPU breaks even at each tier, and a measurement of how much of a paid GPU-hour stays idle at the best load that still meets the SLO.

The metric is SLO-met output tokens per GPU-second at the highest load at which every SLO term holds, the goodput of [5], multiplied by 3600 for tokens per GPU-hour.

2 Experimental setup

System. One whole GPU per host: an older inference GPU, a previous-generation training GPU and a current flagship GPU, all on the same driver. Every GPU runs stock vLLM 0.30.0 [6] with the same flags, a maximum length of 9216, a memory fraction of 0.90, at most 256 sequences and a step budget of 2048 tokens, with prefix caching, chunked prefill and CUDA graphs at their defaults. The model is Qwen3-8B [7] in BF16 on all three GPUs and in FP8 on the flagship GPU, with Qwen3-1.7B in BF16 on the older and previous-generation GPUs as a second point.

Workloads. Chat replays the public 2023 LLM inference conversation trace released with Splitwise [8], with prompts capped at 4096 tokens and answers at 512, a mean of 1056 tokens in and 251 out. Long-prompt RAG replays the Mooncake conversation trace [9] with its shared 512-token prefix blocks, capped at 8192 and 512, a mean of 5847 in and 321 out. Arrivals keep the traces' bursts, answers have exactly the trace's length, and prompt tokens are salted per run so that a repeat cannot reuse an earlier run's prefix cache.

SLO tiers. The interactive tier needs a p95 TTFT of at most 0.5 s and a p95 TPOT of at most 30 ms, and the standard tier 1.0 s and 50 ms. Both need at least 99% of requests completed. The relaxed tier counts completion only, and its throughput is read from 256 closed-loop clients.

Load search. For each GPU, model, workload and seed, an open-loop ladder raises the load until the standard tier fails, bisects each tier on a grid of 10.7% steps and rechecks the best load with a fresh window. Seeds 1 to 3 run in every cell, and seed 1 is repeated with a new salt on every GPU and workload as an A/A. A floor probe sends one request at a time to show whether a tier can be met at any load. nvidia-smi is sampled on every GPU, and DCGM SM-active and tensor-active on the flagship GPU.

3 Results

Table 1: SLO-met output tokens per GPU-hour, in millions, for Qwen3-8B, mean of 3 seeds, at the interactive (int), standard (std) and relaxed (rlx) tiers. Flagship, Previous generation and Older are the current flagship GPU, the previous-generation training GPU and the older inference GPU.

GPU	Chat int	Chat std	Chat rlx	RAG int	RAG std	RAG rlx
Flagship, FP8	12.03	12.98	14.17	2.64	2.82	3.61
Flagship, BF16	8.48	11.69	12.89	1.92	2.64	3.14
Previous generation, BF16	3.07	3.25	5.92	0	1.03	1.43
Older, BF16	0	0.10	0.93	0	0	0.24

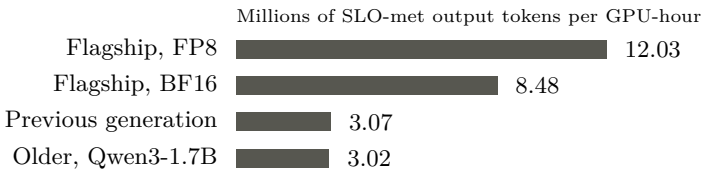


Figure 1: Interactive chat tokens served within SLO per GPU-hour, mean of 3 seeds. Qwen3-8B on the flagship and previous-generation GPUs, Qwen3-1.7B on the older GPU, which serves none with Qwen3-8B.

The interactive tier separates the generations. At the interactive tier on chat the flagship GPU served 2357 SLO-met output tokens per GPU-second against 852 for the previous-generation GPU, 2.77× by ratio of means and 2.51 to 3.68 per seed. When only completion

counts the gap is 2.18×, 3581 against 1645. The metric counts the tokens of whichever slice of the bursty trace the passing window drew, so it is noisy: on seed 3 the previous-generation GPU passed 5.06 requests per second on windows that received 2.87 at 168 tokens per request and served 481 tokens per GPU-second, against 1072 at its passing load of 3.74, which gives that seed its 3.68×. By the highest load met the ratio is 2.09×, and by the tokens of the best passing load 2.25×. The same effect puts the standard tier at 3.59×. Per GPU-hour, the flagship GPU serves 8.48 million interactive chat tokens, the previous-generation GPU 3.07 million and the older GPU none.

Some zeros hold at any load. With one request at a time, the older GPU's TPOT for Qwen3-8B was 33.6 ms at p50, above the interactive tier's 30 ms, so no load meets that tier, and its p95 TTFT on long prompts was 2.0834 s, above the standard tier's 1 s. The previous-generation GPU's p95 TTFT on long prompts was 0.6046 s with one request at a time, so it serves nothing at the interactive tier on RAG, where the flagship GPU serves 533 tokens per GPU-second. These zeros come from latency floors on an idle GPU, not from contention, but they are floors of this stack rather than of the hardware. The older GPU would need about 91% of its peak memory bandwidth to reach 30 ms per token, while the previous-generation GPU's long-prompt prefill ran at about 70% of its compute peak, against a full-peak floor of about 0.43 s, so a different engine or step budget could move that zero.

Model size moves the frontier as much as the generation. Qwen3-1.7B on the older GPU served 3.02 million interactive chat tokens per GPU-hour, as many as Qwen3-8B on the previous-generation GPU at 3.07 million. For a fleet of older GPUs, the size of the model it serves decides whether it earns anything at a latency tier.

Busy by nvidia-smi, idle by DCGM. At every SLO knee the previous-generation and flagship GPUs read 95.1% to 99.5% utilization in nvidia-smi. On the flagship GPU, DCGM showed 18.0% to 28.6% of SM time idle at the interactive and standard knees, against 6.1% to 8.0% when saturated. Holding the SLO gave up 9.3% to 38.8% of the throughput the same flagship GPU reached when saturated, and 28.3% to 48.4% on the previous-generation GPU. Utilization dashboards report this capacity as used, and it is the capacity that sharing a GPU with lower-priority work can target.

FP8 helps most where latency binds. On the flagship GPU, FP8 weights raised interactive chat to 3342 tokens per GPU-second, 1.42× BF16, and interactive RAG by 1.37×, against 1.10× and 1.15× at the relaxed tier. Output quality was not measured, because prompts are random tokens with forced lengths.

Break-even prices. Dividing tokens per GPU-hour gives the price ratio at which a GPU matches the flagship GPU's cost per SLO-met token. For chat on the previous-generation GPU it is 0.36 at the interactive tier, 0.28 at standard and 0.46 at relaxed, and 0.26, 0.25 and 0.42 against the flagship GPU with FP8. In every tier and workload measured, an hour of the previous-generation GPU is worth less than half an hour of the flagship GPU, and much less once a latency SLO applies.

Noise and the limits of the comparison. The A/A reproduced the best load exactly in all 14 SLO-tier A/A cells, and relaxed readings repeated within 0.2% to 4.4%, and 13.2% on RAG on the older GPU. The A/A replays the same arrivals and lengths with new prompt tokens, so it does not measure the dominant noise, which is seed to seed: 1767 to 2724 tokens per GPU-second for the flagship GPU on interactive chat, and a coefficient of variation of 35% on tokens and 23% on load over 7 seeds of the previous-generation GPU. Two flagship-GPU cells on seed 3 rest on one window instead of a recheck. Every GPU ran one engine configuration. On one seed, a step budget of 8192 tokens raised the previous-generation GPU's standard-tier chat load from 5.60 to 8.40 requests per second, so the standard-tier ratios hold for this configuration, not for each GPU's best one.

Pre-registration. The frontier gate, 3 seeds in every core cell and an A/A on every GPU, passed. The hypothesis that the generation ratio is larger at the interactive tier than at the relaxed tier held on the token metric on every seed, but it is fragile on chat: on the best load alone the ordering flips on 2 of 3 seeds, and by the tokens of the best passing load seed 3 gives 1.66× against 2.08×. Its robust part is the zeros. The hypothesis that nvidia-smi

reads below 95% at the standard knee failed. FP8 beat BF16 beyond the A/A spread on every seed at the interactive and relaxed tiers.

4 Independent verification

A separate verifier recomputed the headline cells, ratios and floors from the raw files and matched them, and confirmed that every GPU received identical traffic at the same seed and load. Its slot was the previous-generation GPU's host, so it re-measured the previous-generation GPU on fresh seeds 11 to 14 and could not re-measure the flagship GPU. The previous-generation GPU served 1061, 546, 471 and 903 interactive chat tokens per GPU-second, a mean of 745, inside the researcher's range, and its A/A reproduced the best load exactly. Against the researcher's flagship-GPU seeds, the token ratio was 2.98× over all 7 seeds of the previous-generation GPU with a bootstrap 95% CI of 2.16× to 4.15×, and the load ratio 2.43× with a CI of 1.94× to 3.04×. On a fresh long-prompt seed the previous-generation GPU again served nothing at the interactive tier, with a p95 TTFT of 0.6055 s for one request at a time. The verifier classed the study as a fresh measurement of a known effect [3], new in its production traces, p95 tiers, three generations including the older GPU and current stack, and recommended quoting the gap as roughly 2 to 3× rather than as a precise figure.

5 Conclusion

Measured the same way on three generations, what a GPU-hour serves depends on the latency tier as much as on the GPU. The flagship GPU served roughly 2 to 3× the interactive chat work of the previous-generation GPU, 2.77× by the registered token metric and 2.09× by load, and 2.18× when only completion counts, the older GPU served none at the interactive tier with Qwen3-8B, and every GPU looked fully busy in nvidia-smi while 18.0% to 28.6% of the flagship GPU's SM time stayed idle at its SLO knee. The study covers one GPU per generation, one engine configuration, one model family and random-token prompts. For a fleet, planning in SLO-met tokens per GPU-hour per tier shows which GPUs to keep on latency-bound work, which to move to batch work, and which to give smaller models.

References

[1] V. J. Reddi et al. MLPerf Inference Benchmark. ISCA 2020.

[2] SemiAnalysis. InferenceMAX: open-source continuous inference benchmarking. 2025.

[3] T. Griggs et al. Mélange: Cost Efficient Large Language Model Serving by Exploiting GPU Heterogeneity. arXiv preprint, 2024.

[4] C. Patil. Beyond Per-Token Pricing: A Concurrency-Aware Methodology for LLM Infrastructure Cost Estimation. arXiv preprint, 2026.

[5] Y. Zhong et al. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving. OSDI 2024.

[6] W. Kwon et al. Efficient Memory Management for Large Language Model Serving with PagedAttention. SOSP 2023.

[7] A. Yang et al. Qwen3 Technical Report. arXiv preprint, 2025.

[8] P. Patel et al. Splitwise: Efficient Generative LLM Inference Using Phase Splitting. ISCA 2024.

[9] R. Qin et al. Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving. FAST 2025.