project_flotilla · rented GPU economics · 12 Sep 2026

Flotilla Session Rate Card

What one concurrent session with a 262K-token context costs per month at 160 rented hours, for every model in scope. Each range spans the vast.ai offers we actually saw for a box that can run the model. Measured rows come from our own rented hardware; projected rows lean on a concurrency assumption.

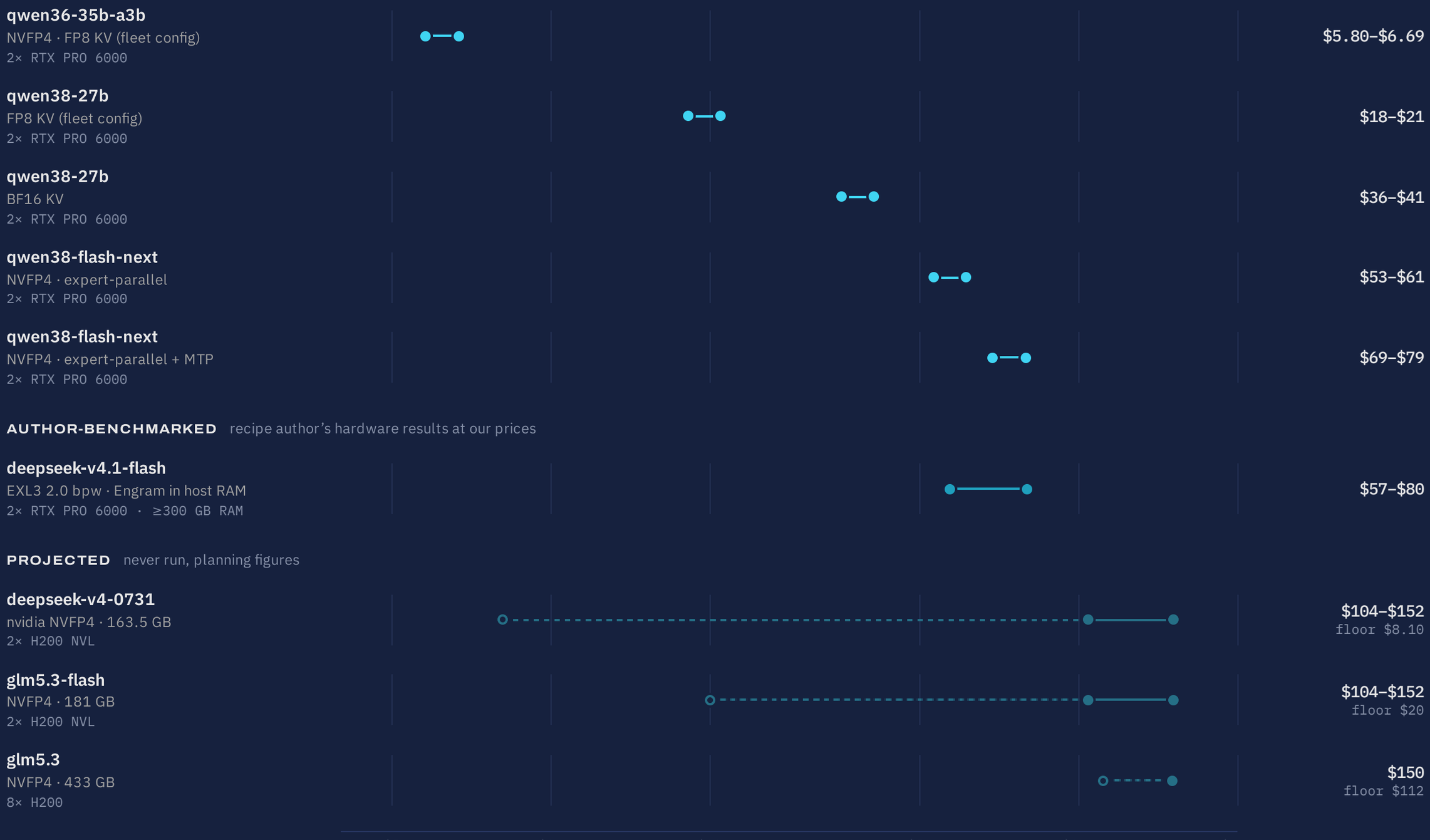
box $/h × 160 h ÷ concurrent 262K sessions per box = $/session-month

Concurrent sessions are the lower of two limits: what fits in the KV cache, and what the engine will run at once.

Cost per concurrent session, per month

USD · log scale · 160 h/month

Measured on our rented box · Recipe author's benchmark, our prices · Projected, never run · Floor if the KV cache were the only limit



Inputs behind every row

Hourly prices are the range of real offers seen on 11–12 Sep 2026 for a box that meets each model's requirements. Box cost is the hourly price times 160 hours.

Evidence	Model and configuration	Box	$/h	Box, 160 h	Concurrent sessions	$/session-month	Engine	Basis
Measured	**qwen36-35b-a3b** NVFP4 · FP8 KV (fleet config)	2× RTX PRO 6000	$1.97–$2.27	$315–$363	54.27	$5.80–$6.69	vLLM nightly e7edf17c	Measured KV ledger on the rented box; the cache fills first
Measured	**qwen38-27b** FP8 KV (fleet config)	2× RTX PRO 6000	$1.97–$2.27	$315–$363	17.28	$18–$21	vLLM nightly e7edf17c	Measured KV ledger; the cache fills first
Measured	**qwen38-27b** BF16 KV	2× RTX PRO 6000	$1.97–$2.27	$315–$363	8.86	$36–$41	vLLM nightly e7edf17c	Measured KV ledger; the cache fills first
Measured	**qwen38-flash-next** NVFP4 · expert-parallel	2× RTX PRO 6000	$1.97–$2.27	$315–$363	5.92	$53–$61	vLLM nightly e7edf17c	Measured KV ledger; its attention forces a BF16 cache
Measured	**qwen38-flash-next** NVFP4 · expert-parallel + MTP	2× RTX PRO 6000	$1.97–$2.27	$315–$363	4.58	$69–$79	vLLM nightly e7edf17c	Measured KV ledger; the MTP draft uses cache too

Evidence	Model and configuration	Box	$/h	Box, 160 h	Concurrent sessions	$/session-month	Engine	Basis
Author-benchmarked	**deepseek-v4.1-flash** EXL3 2.0 bpw · Engram in host RAM	2× RTX PRO 6000 · ≥300 GB RAM	$2.86–$4.00	$457–$640	8 cache alone ≈ 14.5	$57–$80	vLLM day-0 image + EXL3	The recipe runs at most 8 sequences; its author benchmarked up to 4 streams
Projected	**deepseek-v4-0731** nvidia NVFP4 · 163.5 GB	2× H200 NVL	$5.21–$7.58	$834–$1,212	8 planned cache alone ≈ 41–103	$104–$152 floor $8.10	vLLM or SGLang	Assumes 8 per 2 GPUs, borrowed from V4.1-Flash; cache size per token inferred from the V4.1 card
Projected	**glm5.3-flash** NVFP4 · 181 GB	2× H200 NVL	$5.21–$7.58	$834–$1,212	8 planned cache alone ≈ 42	$104–$152 floor $20	vLLM per-model image	Assumes 8 per 2 GPUs; only 11 of its 45 layers hold KV cache
Projected	**glm5.3** NVFP4 · 433 GB	8× H200	$30.04	$4,806	32 planned cache alone ≈ 43	$150 floor $112	vLLM or SGLang	Assumes 8 per 2 GPUs; cache size per token derived from its config

Runtime behind every row

Measured on 2× RTX PRO 6000 Max-Q with vLLM nightly e7edf17c at TP=2 and a 262K window, 11 Sep 2026. Decode is aggregate throughput across concurrent streams. DeepSeek-V4.1-Flash's figures are its recipe author's, on the same card at 300 W with CUDA graphs and DSpark speculative decoding. Projected rows have never run: context is the model's configured maximum and the KV pool is our projection.

Evidence	Model and configuration	Decode tok/s	Cold prefill tok/s	Max context	KV pool	Concurrency
Measured	**qwen36-35b-a3b** NVFP4 · FP8 KV (fleet config)	197 → 6,396 1 → 128 streams	30.1K → 15.9K 8K → 132K prompt	262K served window	14.2M tokens measured ledger	256 max seqs; 64 concurrent 27K agent sessions at ≤0.37 s p95 TTFT
Measured	**qwen38-27b** FP8 KV (fleet config)	92.6 → 2,908 1 → 64 streams	10.0K → 5.9K 8K → 132K prompt	262K served window	4.53M tokens measured ledger	256 max seqs
Measured	**qwen38-27b** BF16 KV	96.7 → 2,873 1 → 64 streams	10.1K → 6.5K 8K → 132K prompt	262K served window	2.32M tokens measured ledger	256 max seqs
Measured	**qwen38-flash-next** NVFP4 · expert-parallel	97.5 → 1,846 1 → 64 streams	13.7K → 13.8K 8K → 132K, flat	262K served window	1.55M tokens measured ledger	256 max seqs; its attention forces a BF16 cache
Measured	**qwen38-flash-next** NVFP4 · expert-parallel + MTP	125.9 → 2,074 1 → 64 streams	12.6K → 13.3K 8K → 132K prompt	262K served window	1.20M tokens measured ledger	256 max seqs; MTP adds 12–29% decode at every level up to 64
Author-benchmarked	**deepseek-v4.1-flash** EXL3 2.0 bpw · Engram in host RAM	113.3 → 277.2 1 → 4 streams @2K; 116.6 @46K	4,044 → 8,085 46K prompt, 1 → 4 streams	512K configured; model max 1M	≈3.8M tokens 8 GiB cache	8 max seqs; benchmarked to 4 streams; agentic run held 240–254 tok/s over 3–4 streams with 87% prefix-cache hits
Projected	**deepseek-v4-0731** nvidia NVFP4 · 163.5 GB	not measured	not measured	1M vendor recipe uses 393K	≈10.6–27.0M tokens projected, 2× H200 NVL	8 planned; vendor tested TP=8 on B200
Projected	**glm5.3-flash** NVFP4 · 181 GB	not measured	not measured	1M configured maximum	≈10.9M tokens projected, 2× H200 NVL	8 planned; only 11 of 45 layers hold KV cache
Projected	**glm5.3** NVFP4 · 433 GB	not measured	not measured	1M configured maximum	≈11.3M tokens projected, 8× H200	32 planned

Reading it

Two different limits fill a box.

The qwen rows run out of KV cache first, so their session counts come straight from the KV ledger we measured. DeepSeek-V4.1-Flash runs out of compute first: its recipe runs at most 8 sequences, though its cache could hold about 14.5. Its author benchmarked only up to 4 streams, and at 4 a session would cost $114–$160. For contrast, qwen36-35b-a3b held 64 concurrent 27K-token agent sessions at 0.37 s p95 time-to-first-token or better, which is $4.92–$5.68 each.

Projected rows are planning figures, not quotes.

deepseek-v4-0731, glm5.3-flash and glm5.3 have never run for us. They assume 8 concurrent sessions per 2 GPUs, borrowed from V4.1-Flash's recipe. The dashed line runs down to what they would cost if the KV cache were the only limit, which is where the lower figures quoted earlier came from. Quote the solid range and treat the dashed end as a best case.

Prices moved all day, and some costs are left out.

2× RTX PRO 6000 went for $1.97–$2.27/h. A V4.1-Flash host with at least 300 GB of RAM and CUDA 13 ran $2.86–$4.00/h, and the cheapest of those fails our 0.95 reliability floor. 2× H200 NVL ran $5.21–$7.58/h, and 8× H200 $30.04/h. Bandwidth, disk and cold-start time are not included.

Sources: `FLOTILLA-GUIDANCE.md` §0 and §1.10b, `flotilla/runs/dsv41.sh`, and lane commits `a88aed6a` and `78cf47fa`, plus the vast.ai market on 11–12 Sep 2026.

Every performance figure here is self-reported by hardware we do not control. DeepSeek-V4.1-Flash's numbers belong to its recipe author until our own run reproduces them.