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

MEASURED on the 2× RTX PRO 6000 we rented, 11 Sep 2026

qwen36-35b-a3b
NVFP4 · FP8 KV (fleet config)
2× RTX PRO 6000

$5.80–$6.69

qwen38-27b
FP8 KV (fleet config)
2× RTX PRO 6000

$18–$21

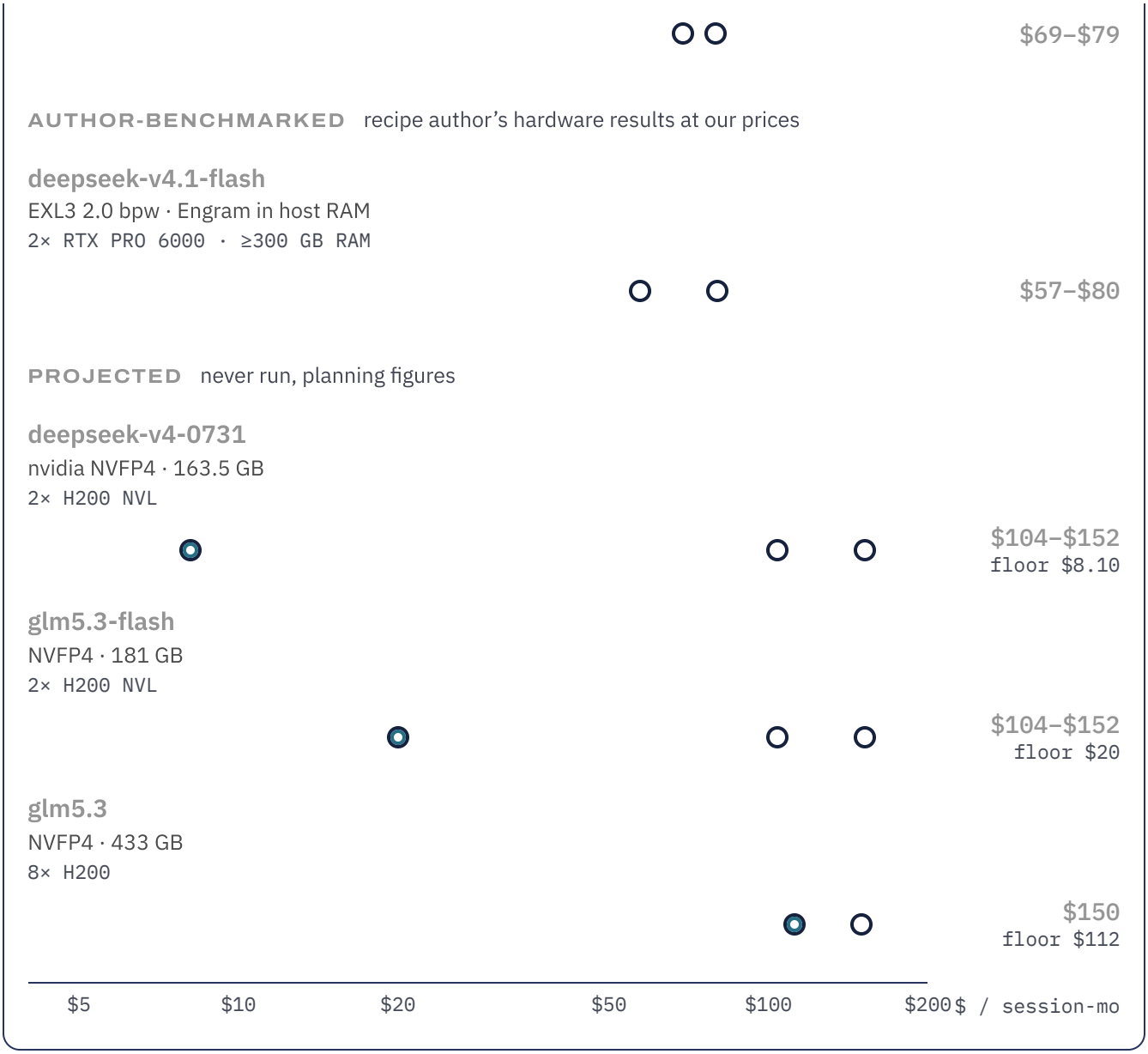
qwen38-27b
BF16 KV
2× RTX PRO 6000

$36–$41

qwen38-flash-next
NVFP4 · expert-parallel
2× RTX PRO 6000

$53–$61

qwen38-flash-next
NVFP4 · expert-parallel + MTP
2× RTX PRO 6000



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
Measured	**qwen36-35b-a3b** NVFP4 · FP8 KV (fleet config)	2× RTX PRO 6000	$1.97–$2.27
Measured	**qwen38-27b** FP8 KV (fleet config)	2× RTX PRO 6000	$1.97–$2.27
Measured	**qwen38-27b** BF16 KV	2× RTX PRO 6000	$1.97–$2.27
Measured	**qwen38-flash-next** NVFP4 · expert parallel	2× RTX PRO 6000	$1.97–$2.27

	NVFP4 · expert-parallel		
Measured	**qwen38-flash-next** NVFP4 · expert-parallel + MTP	2× RTX PRO 6000	$1.97–$2.27
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
Projected	**deepseek-v4-0731** nvidia NVFP4 · 163.5 GB	2× H200 NVL	$5.21–$7.58
Projected	**glm5.3-flash** NVFP4 · 181 GB	2× H200 NVL	$5.21–$7.58
Projected	**glm5.3** NVFP4 · 433 GB	8× H200	$30.04

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.