CAPACITY PLANNER · SUMMARY REPORT

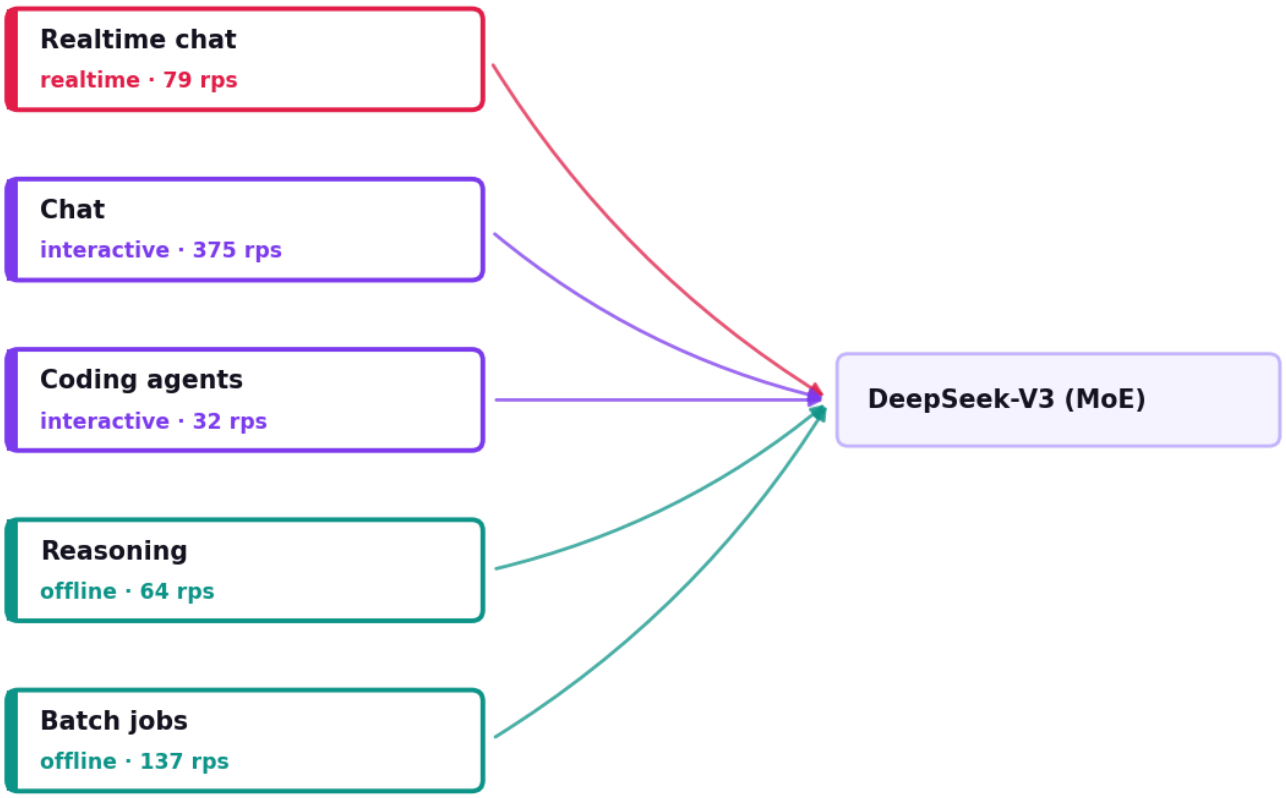
DeepSeek-V3 API — busiest hour on Two H200 clusters — 384 GPUs

Generated 2026-10-03 04:59 UTC · digital-twin sizing estimate

264	8.25	MET
GPUs provisioned	racks	SLO verdict

Workload definition

DeepSeek-V3 API — busiest hour — 1 model, 5 traffic patterns.



Each traffic pattern is one interaction flowing across one or more models in order; numbered arrows show the sequence. Pattern colour is its SLO tier.

Models

Model	Catalog id	Params	Weights	KV	Layers	Context
DeepSeek-V3 (MoE)	deepseek-v3	671B	float8	float8	61	131,072

Traffic patterns

© 2026 AlphaTango AI, Inc.

Pattern	Models touched, in order	rps	SLO tier	pNN	Prompt tok	Output tok
Realtime chat	DeepSeek-V3 (MoE)	79	realtime	p95	800	120
Chat	DeepSeek-V3 (MoE)	375	interactive	p95	1,633	105
Coding agents	DeepSeek-V3 (MoE)	32	interactive	p95	8,596	182
Reasoning	DeepSeek-V3 (MoE)	64	offline	p95	1,500	4,000
Batch jobs	DeepSeek-V3 (MoE)	137	offline	p95	2,000	500

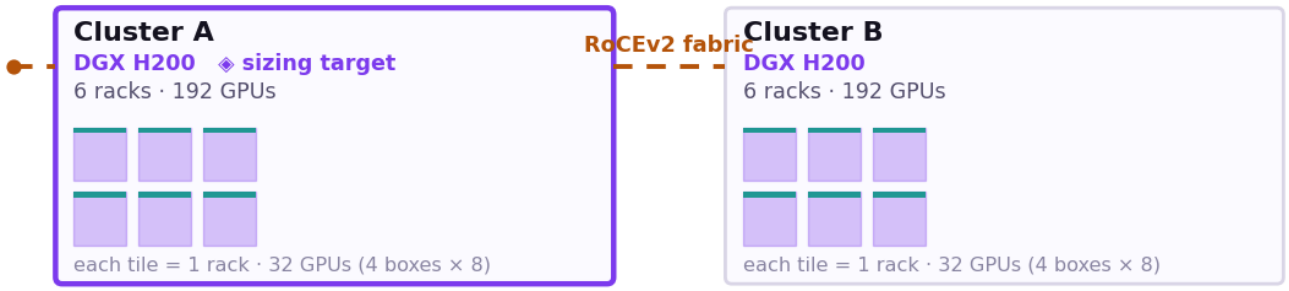
Arrivals are Poisson at the stated mean rate. Each pattern must meet all three ceilings of its SLO tier at that tier's percentile.

SLO tiers in use

Tier	Source	Judged at	Max TTFT	Max TPOT	Max E2E	Priority
interactive	catalog	p95	2s	75ms	10s	high
offline	catalog	p95	4h	100,000ms	4h	low
realtime	catalog	p95	0.5s	50ms	5s	critical

Topology

Two H200 clusters — 384 GPUs — 2 rack groups, 384 GPUs in total.



Rack group	Hardware	Racks	Boxes/rack	GPUs/box	GPUs
Cluster A	h200_sxm	6	4	8	192
Cluster B	h200_sxm	6	4	8	192

Availability

Strategy	Failure domain	Spare %
none	—	—

Between groups

Fabric	Links	Per-link Gb/s
RoCEv2 fabric	4	—

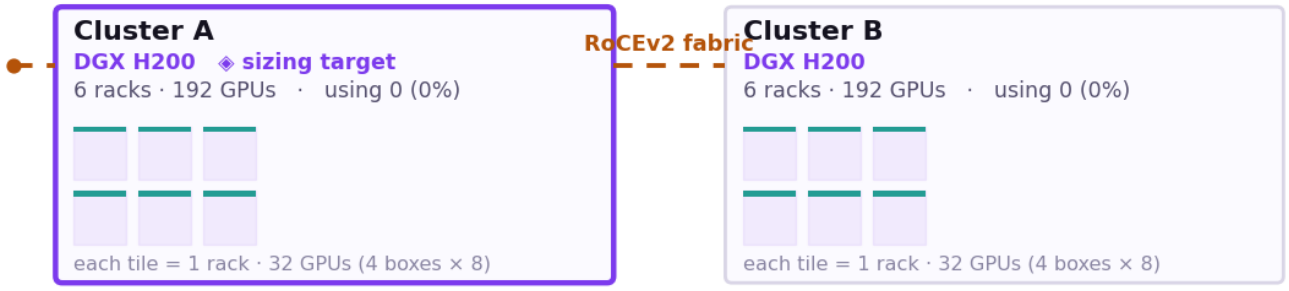
RECOMMENDED CLUSTER

8.25 racks

264 GPUs provisioned

264	152.7 kW	$0.1115
GPUs provisioned	facility power (base)	per 1M tokens

SLO-meeting base: **264** GPUs · **8.25** racks · HA ×**1** (none)



Shaded racks are those the recommendation occupies, at the provisioned (post-HA) count.

Placement — what runs where (auto-placed)

Model	Rack group	TP	Config	GPUs	Racks	Serves	Patterns
DeepSeek-V3	Cluster A + Cluster B	TP8	33× replicas	264	8.25	687 rps	Realtime chat, Chat, Coding agents, Reasoning, Batch jobs

Per-model placement across rack groups; each model sized on its group's hardware, then summed.

Validated on each hardware

Every model sized on every rack group in the topology, ranked by **cost**. The chosen row is marked ✓.

Model	Rack group	GPUs	Racks	$/month	kWh/1M	SLO
✓ deepseek-v3	Cluster A	264	8.25	$475,200	0.0258	MET

Showing the chosen placement only; the full report lists all 2 measured candidates.

Per-pattern SLO attainment

Pattern	SLO tier	rps	Judged at	TTFT	E2E	Result
Realtime chat	realtime	79	p95	0.188s	4.902s	MET
Chat	interactive	375	p95	0.378s	4.823s	MET
Coding agents	interactive	32	p95	0.378s	4.823s	MET
Reasoning	offline	64	p95	0.312s	38.214s	MET
Batch jobs	offline	137	p95	0.312s	38.214s	MET

Economics: **$0.1115** / 1M tokens · **0.0258** kWh / 1M · **$475,200** per month.

per-model placement summed across the fleet

© 2026 AlphaTango AI, Inc.

Capacity sensitivity — footprint vs traffic

Not computed for this run. Tick **compute sensitivity** on the Run tab and re-size to see how the footprint grows with traffic.

Live summary — final

1 model placed.

Model	Rack group	State	GPUs	Serves
DeepSeek-V3	Cluster A + Cluster B	placed	264	Realtime chat, Chat, Coding agents, Reasoning, Batch jobs

Estimate only · no production access · no customer data leaves your hands. AlphaTango's optimizer can often meet the same SLO on fewer racks and a lower $/token than a static sizing — hello@alphatango.ai

© 2026 AlphaTango AI, Inc.