CAPACITY PLANNER · SUMMARY REPORT

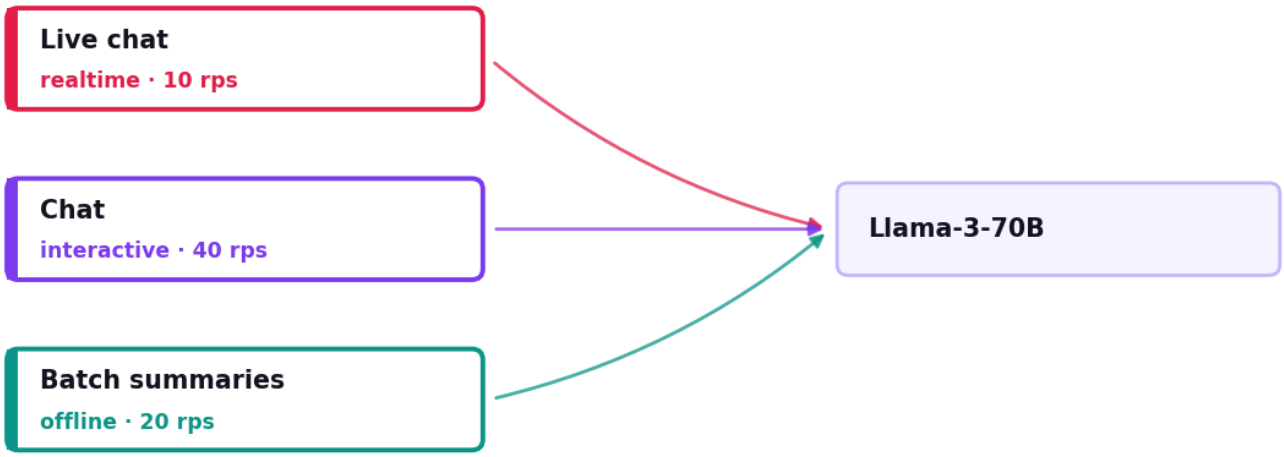
Chat assistant — Llama-3-70B on Vera Rubin NVL144 — 3 racks (216 GPUs)

Generated 2026-10-03 07:18 UTC · digital-twin sizing estimate

12	0.17	MET
GPUs provisioned	racks	SLO verdict

Workload definition

Chat assistant — Llama-3-70B — 1 model, 3 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
Llama-3-70B	llama3-70b	70B	float16	float16	80	8,192

Traffic patterns

Pattern	Models touched, in order	rps	SLO tier	pNN	Prompt tok	Output tok
Live chat	Llama-3-70B	10	realtime	p95	600	120
Chat	Llama-3-70B	40	interactive	p95	1,200	250
Batch summaries	Llama-3-70B	20	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

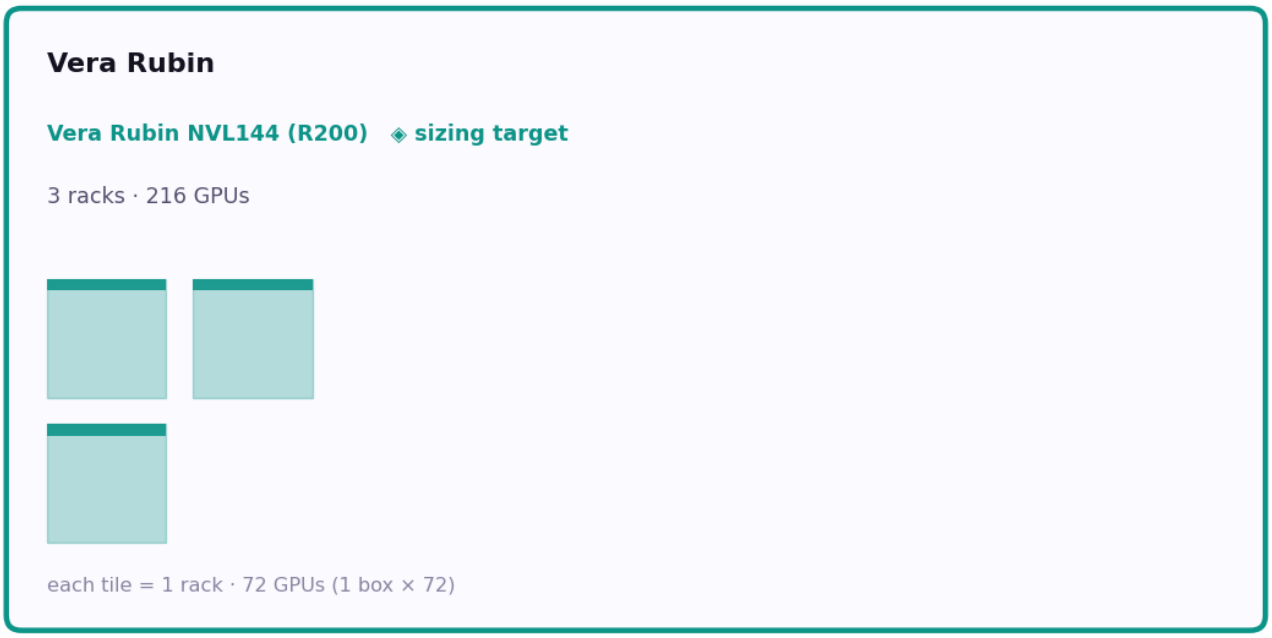
Tier	Source	Judged at	Max TTFT	Max TPOT	Max E2E	Priority
interactive	catalog	p95	2s	75ms	10s	high

© 2026 AlphaTango AI, Inc.

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

Topology

Vera Rubin NVL144 — 3 racks (216 GPUs) — 1 rack group, 216 GPUs in total.



Rack group	Hardware	Racks	Boxes/rack	GPUs/box	GPUs
Vera Rubin	rubin_sxm	3	1	72	216

Availability

Strategy	Failure domain	Spare %
none	—	—

Between groups

Fabric	Links	Per-link Gb/s
RoCEv2 fabric	8	—

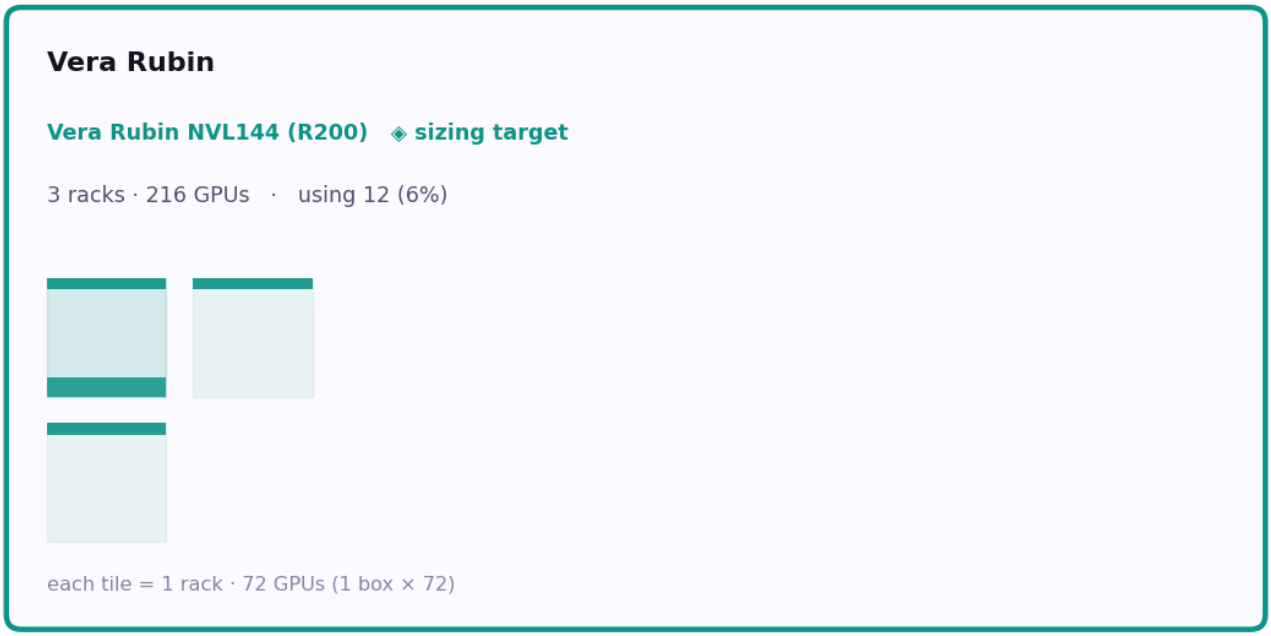
RECOMMENDED CLUSTER

0.17 racks

12 GPUs provisioned

12	19 kW	$0.2287
GPUs provisioned	facility power (base)	per 1M tokens

SLO-meeting base: **12** GPUs · **0.17** 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
Llama-3-70B	Vera Rubin	TP2	6× replicas	12	0.17	70 rps	Live chat, Chat, Batch summaries

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
✓ llama3-70b	Vera Rubin	12	0.17	$69,120	0.0453	MET

Per-pattern SLO attainment

Pattern	SLO tier	rps	Judged at	TTFT	E2E	Result
Live chat	realtime	10	p95	0.297s	4.06s	MET
Chat	interactive	40	p95	0.537s	6.836s	MET

© 2026 AlphaTango AI, Inc.

Pattern	SLO tier	rps	Judged at	TTFT	E2E	Result
Batch summaries	offline	20	p95	0.758s	11.19s	MET

Economics: **$0.2287** / 1M tokens · **0.0453** kWh / 1M · **$69,120** per month.

per-model placement summed across the fleet; monthly cost and $/1M include HA spare GPUs at their group's rate; power excludes spares

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
Llama-3-70B	Vera Rubin	placed	12	Live chat, Chat, Batch summaries

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.