CAPACITY PLANNER · SUMMARY REPORT

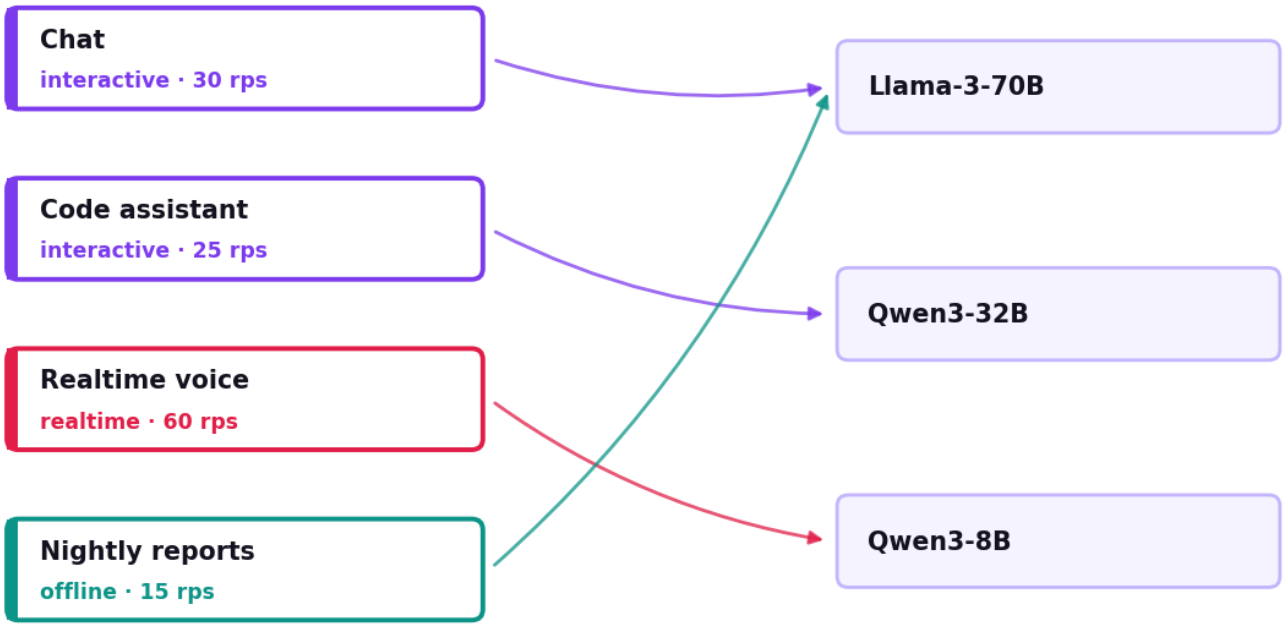
Model portfolio — three models on H200 + Vera Rubin — two clusters

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

41	1.28	MET
GPUs provisioned	racks	SLO verdict

Workload definition

Model portfolio — three models — 3 models, 4 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
Qwen3-32B	qwen3-32b	32.8B	float8	float8	64	40,960
Qwen3-8B	qwen3-8b	8.2B	float16	float16	36	40,960

Traffic patterns

Pattern	Models touched, in order	rps	SLO tier	pNN	Prompt tok	Output tok
Chat	Llama-3-70B	30	interactive	p95	1,200	250
Code assistant	Qwen3-32B	25	interactive	p95	6,000	300

© 2026 AlphaTango AI, Inc.

Pattern	Models touched, in order	rps	SLO tier	pNN	Prompt tok	Output tok
Realtime voice	Qwen3-8B	60	realtime	p95	500	100
Nightly reports	Llama-3-70B	15	offline	p95	2,500	400

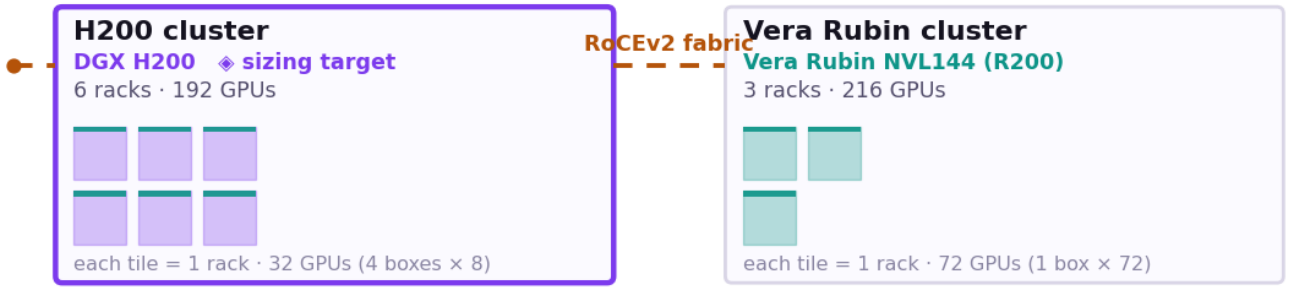
Arrivals are Poisson at the stated mean rate. Each pattern must meet all three ceilings of its SLO tier at that tier's percentile.

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

H200 + Vera Rubin — two clusters — 2 rack groups, 408 GPUs in total.



Rack group	Hardware	Racks	Boxes/rack	GPUs/box	GPUs
H200 cluster	h200_sxm	6	4	8	192
Vera Rubin cluster	rubin_sxm	3	1	72	216

Availability

Strategy	Failure domain	Spare %
none	—	—

Between groups

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

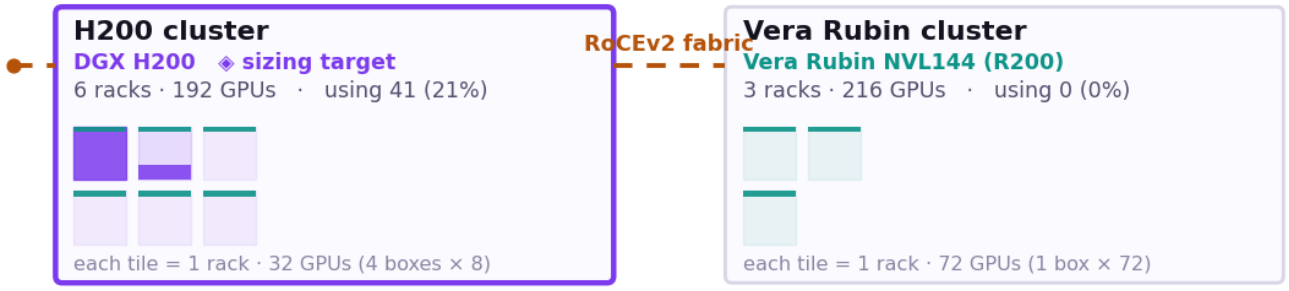
RECOMMENDED CLUSTER

1.28 racks

41 GPUs provisioned

41	23.6 kW	$0.1006
GPUs provisioned	facility power (base)	per 1M tokens

SLO-meeting base: **41** GPUs · **1.28** 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	H200 cluster	TP8	3× replicas	24	0.75	45 rps	Chat, Nightly reports
Qwen3-32B	H200 cluster	TP2	8× replicas	16	0.5	25 rps	Code assistant
Qwen3-8B	H200 cluster	TP1	1× replicas	1	0.03	60 rps	Realtime voice

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	H200 cluster	24	0.75	$43,200	0.0466	MET
✓ qwen3-32b	H200 cluster	16	0.5	$28,800	0.0143	MET
✓ qwen3-8b	H200 cluster	1	0.03	$1,800	0.005	MET

Showing the chosen placement only; the full report lists all 6 measured candidates.

Per-pattern SLO attainment

Pattern	SLO tier	rps	Judged at	TTFT	E2E	Result
Chat	interactive	30	p95	0.361s	6.174s	MET
Code assistant	interactive	25	p95	1.623s	9.025s	MET
Realtime voice	realtime	60	p95	0.258s	3.59s	MET
Nightly reports	offline	15	p95	0.67s	8.57s	MET

Economics: **$0.1006** / 1M tokens · **0.0231** kWh / 1M · **$73,800** 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

© 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

3 models placed.

Model	Rack group	State	GPUs	Serves
Llama-3-70B	H200 cluster	placed	24	Chat, Nightly reports
Qwen3-32B	H200 cluster	placed	16	Code assistant
Qwen3-8B	H200 cluster	placed	1	Realtime voice

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.