CAPACITY PLANNER · SUMMARY REPORT

Support copilot — Qwen3-32B on H200 — 10 racks (320 GPUs)

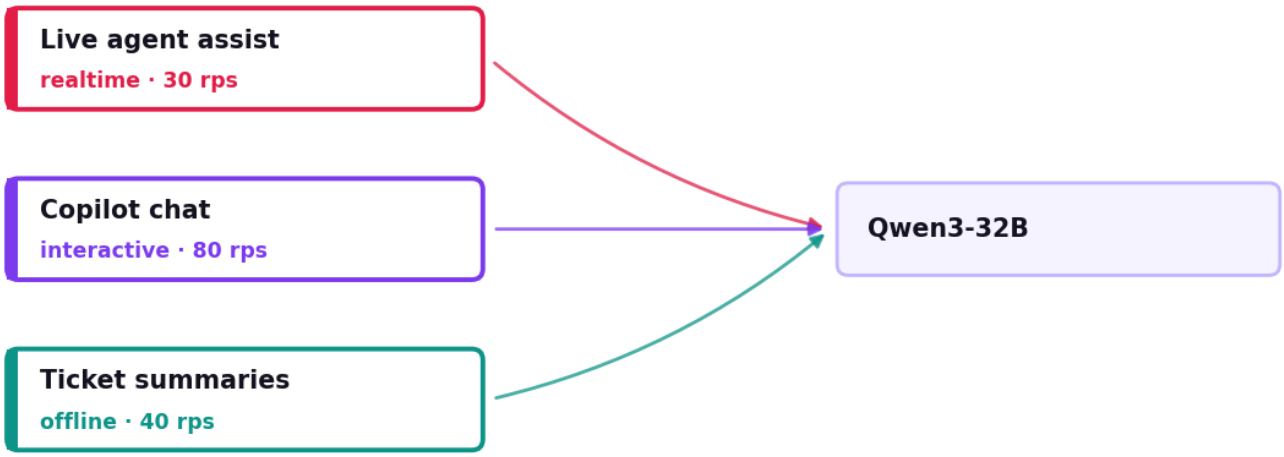
Generated 2026-10-03 05:14 UTC · digital-twin sizing estimate

■ **Stale. The workload or topology has been edited since this result was computed. Re-run the cell for a current answer.**

16	0.5	MET
GPUs provisioned	racks	SLO verdict

Workload definition

Support copilot — Qwen3-32B — 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
Qwen3-32B	qwen3-32b	32.8B	float8	float8	64	40,960

Traffic patterns

Pattern	Models touched, in order	rps	SLO tier	pNN	Prompt tok	Output tok
Live agent assist	Qwen3-32B	30	realtime	p95	700	150
Copilot chat	Qwen3-32B	80	interactive	p95	1,500	300
Ticket summaries	Qwen3-32B	40	offline	p95	2,500	600

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

© 2026 AlphaTango AI, Inc.

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 — 10 racks (320 GPUs) — 1 rack group, 320 GPUs in total.



Rack group	Hardware	Racks	Boxes/rack	GPUs/box	GPUs
H200 racks	h200_sxm	10	4	8	320

Availability

Strategy	Failure domain	Spare %
none	—	—

Between groups

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

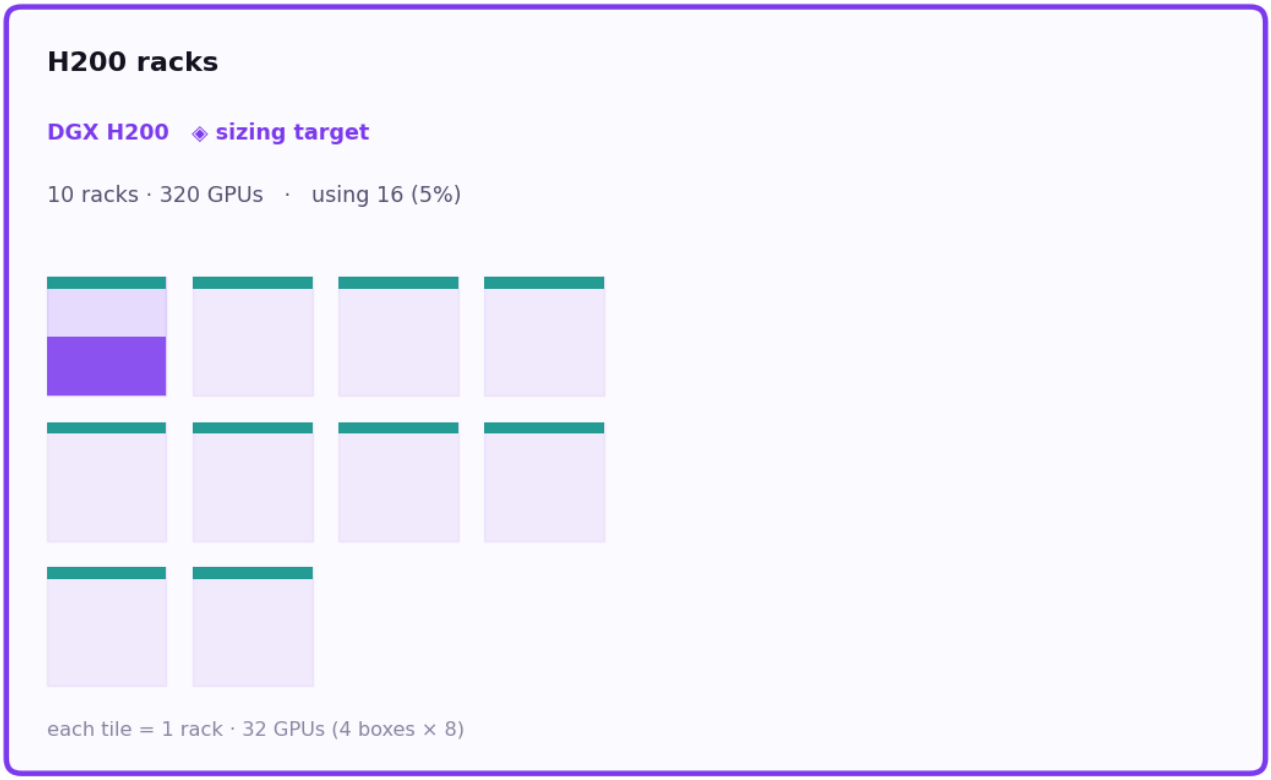
RECOMMENDED CLUSTER

0.5 racks

16 GPUs provisioned

16	11.6 kW	$0.0391
GPUs provisioned	facility power (base)	per 1M tokens

SLO-meeting base: **16** GPUs · **0.5** 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
Qwen3-32B	H200 racks	TP4	4× replicas	16	0.5	150 rps	Live agent assist, Copilot chat, Ticket 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
✓ qwen3-32b	H200 racks	16	0.5	$28,800	0.0114	MET

Per-pattern SLO attainment

© 2026 AlphaTango AI, Inc.

Pattern	SLO tier	rps	Judged at	TTFT	E2E	Result
Live agent assist	realtime	30	p95	0.376s	4.818s	MET
Copilot chat	interactive	80	p95	0.467s	8.766s	MET
Ticket summaries	offline	40	p95	0.669s	12.891s	MET

Economics: **$0.0391** / 1M tokens · **0.0114** kWh / 1M · **$28,800** per month.

per-model placement summed across the fleet

Capacity sensitivity — footprint vs traffic

Traffic	GPUs	Racks	SLO
×1	16	0.5	MET
×1.5	24	0.75	MET
×2	36	1.12	MET

Live summary — final

1 model placed.

Model	Rack group	State	GPUs	Serves
Qwen3-32B	H200 racks	placed	16	Live agent assist, Copilot chat, Ticket 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.