LLM SaaS — 24-region Copilot

Client: North American AI Copilot · **Region:** Global

Workload

70B MoE, streaming completions, 180M req/day

Deployment topology

3 Core IDC + 41 Hotel Edge sites, KV-cache tiering

Measured results

Metric	Value	Note
TTFT	-62%	median
Serving cost	-38%	per token
Regions live	24	in 9 weeks

Methodology

Figures anonymised at customer request. Latencies measured at the client SDK boundary over rolling 14-day windows. Cost deltas reference the customer's prior production stack, normalised per token-equivalent.

© Aether Compute Group. Contact: partners@aether-compute.com