

AI INFRASTRUCTURE

AI Has Reopened The Capacity Planning Problem

Why cloud elasticity stopped working

Cloud Didn't Eliminate Capacity Planning

It outsourced it — to three companies who got very good at absorbing the forecasting risk invisibly.

It Outsourced It

Hyperscalers ran the forecast, the procurement, and the lead-time risk. You consumed the output as an API call and a monthly invoice.

Elasticity wasn't infinite capacity. It was somebody else's capacity plan.

THE CONSTRAINT

GPUs Don't Behave Like General Cloud

QUEUES

Waitlists by instance family and region — sometimes quarters, not minutes.

RESERVATIONS

Multi-year volume commitments replace spot availability as the default.

LEAD TIMES

Months to over a year, commitment to delivery, on high-end accelerators.

Purchased $\neq$ Usable

Capacity Illusion Index

The fraction of purchased GPU capacity that actually produces useful work after scheduling overhead, fragmentation, and idle time.

The invoice number and the workload number are different numbers.

THREE GENERATIONS

The Planning Muscle Nobody Rebuilt

PRE-CLOUD

Discipline present

Forecast, procure, wait, manage utilization manually.

ELASTIC CLOUD

Discipline invisible

Scale up, scale down, pay invoice. No muscle required.

AI INFRASTRUCTURE

Discipline needed again

Reservations, allocation windows, queue modeling — not yet rebuilt.

DIAGNOSTIC QUESTION

"If your primary AI workload doubled tomorrow, could your organization estimate when the required capacity would actually be available — not just when the budget would be approved?"

**Elasticity wasn't infinite capacity.
It was somebody else's capacity plan.**

The bill for not noticing that has now come due.