THIS AI INFRASTRUCTURE CYCLE ALREADY HAPPENED ONCE

Virtualization ran the same five stages, 2005–2015.

EVERY EXPANSION CYCLE RUNS THE SAME FIVE STAGES

VIRTUALIZATION RAN THIS CYCLE
2005 – 2015

Cheap hardware. Any team could provision a VM in an afternoon.

No capacity review. No ownership model. No one asking why.

AI INFRASTRUCTURE IS AT THE SAME POINT NOW

~5%

average enterprise GPU fleet utilization
against $401B in new 2026 AI infrastructure spend

HYPERVISOR COMMODITIZATION DIDN'T END THE CYCLE

Operating the environment became harder than building it.

That's what actually forced the reckoning.

PHYSICAL CONSTRAINTS ARE BECOMING VISIBLE AGAIN

Power

Cooling

Accelerator lead times

Upstream memory supply supply

COORDINATION OVERHEAD IS THE OTHER HALF

- Reservation systems
- Allocation policy
- Chargeback / showback
- Approval workflows

Nobody builds a VM approval board during a growth phase.

TWO CEILINGS, NOT ONE

#106

DENSITY CEILING

Limits what the hardware can do.

#132

COORDINATION DENSITY

Limits what the organization can coordinate.

Consolidation happens when both ceilings are visible at once.

SIGNALS YOU'RE ALREADY IN THE CONSOLIDATION PHASE

GPU reservation systems replacing on-demand provisioning

An internal AI platform team with its own headcount

Formal allocation reviews before new capacity

Recurring utilization governance, not a one-time audit

Shared inference clusters instead of per-team capacity

Chargeback or showback attributing GPU cost

CONSOLIDATION BEGINS WHEN OPERATING GETS HARDER THAN EXPANDING

rack2cloud.com | Follow The Architect for more