

Latency Debt

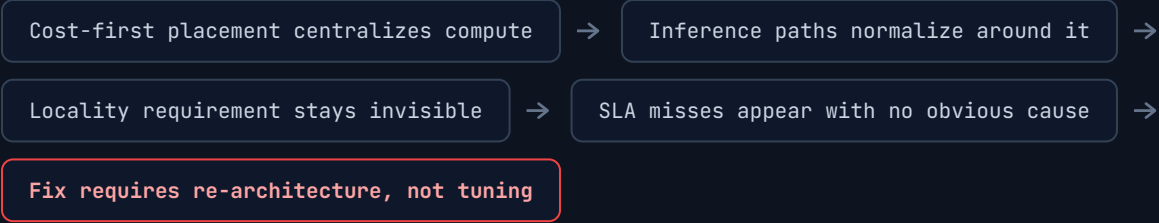
DEFINITION

The accumulated performance and cost penalty incurred when AI placement decisions defer latency until it must be bought back through architecture change.

FRAMEWORK FLOW

01 The Condition Cost is the default optimization axis for AI placement. Centralized GPU clusters win on utilization economics; locality is deferred to "a later phase when requirements are better understood."	02 The Boundary That later phase rarely arrives. Inference paths, data gravity, and application behavior normalize around the centralized topology before locality becomes operationally visible.	03 Failure State Debt accrues invisibly, then surfaces as degraded user experience, SLA misses in specific workload paths, and timeout increases with no obvious architectural cause.	04 Consequence The placement topology is already deeply embedded by the time the signal appears. Remediation is not an optimization exercise. It is a re-architecture.
---	--	--	---

Failure State: Debt That Compounds Before It's Visible



THE DIAGNOSTIC SIGNAL

WORKLOAD CLASSIFICATION Latency-elastic workloads tolerate distance; latency-critical workloads collapse under multi-hop topology. The sensitivity is invisible until the workload is under production load.	COMPOUNDING EFFECT Latency is not additive across pipeline hops. A 40ms retrieval delay shifts the guardrail window and amplifies through every chained model call downstream.
TRIGGER CONDITION The debt stays invisible for the entire deferral period. It surfaces only when an SLA breach or a timeout spike forces the topology question open.	REMEDICATION SHAPE Retrofitting means changing topology and reconciling every downstream dependency — timeout budgets, retry logic, SLAs, observability — that normalized around the original one.

ARCHITECTURAL RELATIONSHIPS

#82

False Completion

RELATED

MODERATE

Both named in the same AI Placement Latency/Cost post — Latency Debt is the accrued cost; False Completion is the metric failure that hides it.

ANCHOR /ai-placement-latency-cost-tradeoff/

FRAMEWORK HOME

Placement decisions look reversible only until inference paths, data gravity, and runtime behavior normalize around whichever topology was chosen first.

"AI placement decisions look reversible during design. They become operationally permanent once runtime behavior hardens around them."