

Throughput Illusion

DEFINITION

The appearance of healthy aggregate throughput while per-request latency and interactivity have already collapsed.

FRAMEWORK FLOW

01 The Condition Aggregate tokens-per-second is measured across all concurrent sessions, so it keeps climbing even as individual sessions degrade.	02 The Boundary Past the concurrency knee, per-session latency crosses the point of usability while the aggregate metric is still rising or holding steady.	03 Failure State The dashboard reports a busy, productive cluster while a growing share of users experience a service that is functionally broken.	04 Consequence Capacity decisions get made off a metric that measures volume, not experience — the gap surfaces as a user complaint, not an alert.
---	--	---	---

Failure State: Busy Is Not the Same as Responsive



THE DIAGNOSTIC SIGNAL

THROUGHPUT ILLUSION INDEX The percentage delta between advertised aggregate throughput and throughput actually delivered within the latency SLO.	PHANTOM THROUGHPUT DETECTION Whether aggregate tokens/sec stays high while interactive latency has already collapsed.
GPU YIELD DEPENDENCY Effective interactive capacity is bounded above by effective GPU yield — the illusion compounds when both are read in isolation.	SLO-BOUND MEASUREMENT Throughput only means something when measured against sessions actually served within SLO, not against raw volume.

ARCHITECTURAL RELATIONSHIPS

#95 Interaction Collapse Point RELATEDMODERATE Co-resident in the Saturation Analyzer — Throughput Illusion masks the approach to Interaction Collapse.	#96 Inference Saturation Curve RELATEDMODERATE Co-resident in the Saturation Analyzer as paired curve/illusion metrics.
#97 Token Queue Amplification RELATEDMODERATE Co-resident in the Saturation Analyzer — queue amplification is masked by the throughput illusion until it isn't.	

ANCHOR /ai-inference-saturation-analyzer/

FRAMEWORK HOME

High aggregate tokens/sec masking unusable interactive latency — the gap between a cluster that is busy and a cluster that is responsive.

"A cluster that is busy and a cluster that is responsive are not the same claim — the Throughput Illusion is the gap between them."