

Token Queue Amplification

DEFINITION

The compounding growth of queued token work under load — small arrival increases amplify into large queue and latency growth.

FRAMEWORK FLOW

01 The Condition As offered load approaches served capacity, requests begin waiting rather than executing immediately.	02 The Boundary Past that point, the waiting line itself — not the underlying work — becomes the dominant source of latency.	03 Failure State Small increases in offered load produce disproportionate increases in tail latency, because queueing delay compounds nonlinearly near capacity.	04 Consequence Tail latency is manufactured by the queue, not the compute, so adding raw throughput without changing admission or batching policy doesn't fix it.
---	---	---	--

Failure State: The Line, Not the Work



THE DIAGNOSTIC SIGNAL

QUEUE AMPLIFICATION SIGNAL The quantified latency multiplier at offered peak, expressed against utilization ρ and burst-adjusted load.	UTILIZATION PROXIMITY How close offered load sits to served capacity — proximity, not just absolute load, drives amplification.
DOMINANT MECHANISM TEST Whether tail latency is being manufactured by the queue or by the underlying compute — the two require different fixes.	TRIGGER CONDITION Activation begins as offered load approaches served capacity, independent of how that capacity was originally sized.

ARCHITECTURAL RELATIONSHIPS

#95 Interaction Collapse Point RELATED MODERATE Co-resident in the Saturation Analyzer — Token Queue Amplification is the mechanism driving toward Interaction Collapse.	#96 Inference Saturation Curve RELATED MODERATE Co-resident in the Saturation Analyzer — queue amplification is the mechanism shaping the saturation curve.
#98 Throughput Illusion RELATED MODERATE 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

The nonlinear latency escalation that begins once offered load approaches served capacity — the dominant mechanism behind the Interaction Collapse Point.

"Tail latency is manufactured by the waiting line, not the work — queue amplification is where that distinction stops being academic."