

Interaction Collapse Point

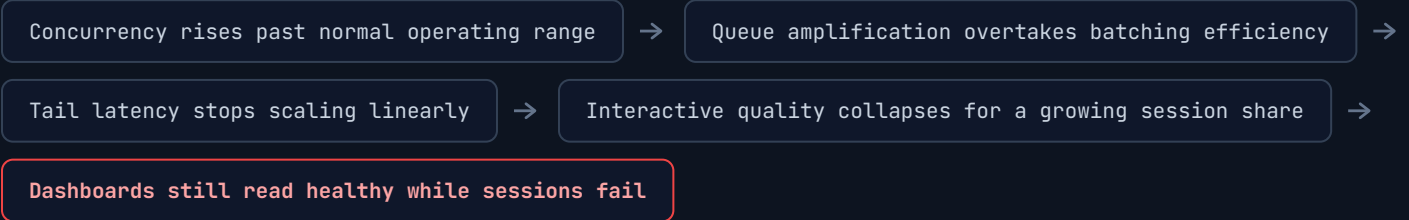
DEFINITION

The load level at which interactive inference latency degrades far enough to break the user interaction model.

FRAMEWORK FLOW

01 The Condition As concurrency rises on a serving deployment, first-token and inter-token latency begin increasing in a way that still looks like normal degradation.	02 The Boundary Past a specific concurrency knee, queue amplification overtakes batching efficiency and tail latency stops scaling linearly.	03 Failure State Interactive quality has already collapsed for a growing share of sessions while aggregate throughput dashboards still read healthy.	04 Consequence Teams discover the knee in production, during a launch, because GPU utilization and tokens-per-second gave no warning it had been crossed.
---	---	---	--

Failure State: Latency Holds Until It Doesn't



THE DIAGNOSTIC SIGNAL

EFFECTIVE INTERACTIVE CAPACITY The sessions actually served within your latency SLO — not the advertised peak concurrency target.	FIRST SATURATION DRIVER Which subsystem binds first: KV ceiling, decode saturation, prefill amplification, or queue amplification.
CONCURRENCY ELASTICITY Where the configured peak sits relative to effective capacity — Stable, Soft Saturation, Amplified Saturation, or Collapse Regime.	TRIGGER CONDITION The gap activates the moment offered load approaches served capacity — headroom, not raw scale, decides when.

ARCHITECTURAL RELATIONSHIPS

#96 Inference Saturation Curve RELATEDMODERATE Co-resident in the AI Inference Saturation Analyzer — the Interaction Collapse Point sits on the Inference Saturation Curve.	#97 Token Queue Amplification RELATEDMODERATE Co-resident in the Saturation Analyzer — Token Queue Amplification is the mechanism driving toward Interaction Collapse.
#98 Throughput Illusion RELATEDMODERATE Co-resident in the Saturation Analyzer — Throughput Illusion masks the approach to Interaction Collapse.	

ANCHOR /ai-inference-saturation-analyzer/

FRAMEWORK HOME

The concurrency knee where latency amplification overtakes batching efficiency, surfaced under stated serving assumptions.

"GPU utilization is not interactive capacity, and aggregate throughput is not responsiveness — the collapse point is where that difference stops being theoretical."