Structural Diagnostic Report

ai.01 — Interconnect Stability Control

Scenario S2: Latency-Critical Inference

Document Type:	Structural Diagnostic Scenario Report
Application ID:	ai.01
Scenario ID:	ai.01.S2
Schema Version:	0.5.1

Source: SORT AI Structural Diagnostics Demo
Application: ai.01 Interconnect Stability Control
Scenario: S2
Version: 1.0.1
Generated: 2025-01-18
Web: https://independent-research-systems-modeling.com

Scope and Limits: *This report presents a structural diagnostic scenario analysis based on pre-computed, normalized projection runs. It is not a complete Architecture Risk Assessment and does not contain implementation guidance.*

1. Scenario Overview

System Class

Globally distributed inference serving infrastructure with hard SLA constraints and continuous availability requirements.

Scale Abstraction

SLA-adjacent operation with hundreds of inference replicas experiencing bursty load patterns and increasing tail latency trends.

Operational Context

Continuous serving with tensor-parallel inference and request batching. P99 latency targets under sustained high throughput requirements. Metric-driven load balancing with retry logic for fault tolerance.

2. Observed Structural Pattern

The following structural effects emerge from the interaction of correctly functioning components under variable load conditions:

- Tail latency growth originates from coupling between replica states under bursty load, not from individual replica overload or capacity exhaustion.
- Load balancing decisions based on average-case metrics systematically miss structural coupling patterns that manifest at distribution tails.
- Retry logic designed for fault tolerance amplifies rather than resolves coupling-induced delays, creating secondary load that compounds the original effect.
- SLA compliance masks escalating structural costs until safety margin exhaustion, at which point the problem becomes acute rather than gradual.
- Transient coupling effects between replicas during load bursts create tail latency signatures that persist after load normalizes.

3. Stability Assessment

Baseline Structural Condition

SLA nominally met but safety margins eroding progressively. Cost trajectory unsustainable with expenses growing faster than demand. Stability reserve diminishing over time.

Observed Instability Class

Marginal — characterized by compliant SLA metrics with hidden cost accumulation and shrinking operational buffers.

Post-Projection Stability Class

Stable — tail distribution stabilized through coupling-aware load distribution. Safety margins restored and cost trajectory normalized.

Transition Type

Gradual stabilization through structural intervention at the routing layer.

4. Aggregated Indicators

All values are normalized ratios. No absolute values or reconstructable parameters are provided.

Indicator	Baseline	Comparison	Direction
Cost per Request Ratio	1.34	1.02	Improvement
Effective Capacity Utilization	0.71	0.88	Improvement
Tail Latency Growth Rate	0.28	0.07	Improvement
SLA Margin Erosion Rate	0.19	0.04	Improvement
Retry Amplification Factor	1.42	1.08	Improvement
Coupling-Induced Delay Fraction	0.31	0.09	Improvement

5. Interpretation

Systemic Relevance

The observed instability pattern is systemically relevant because it represents a structural cost accumulation mechanism that operates independently of demand growth. Costs rise not because more requests arrive, but because the system compensates for structural instability through overprovisioning and retry processing. This creates an economic trajectory that cannot be corrected through demand management or capacity planning alone.

Detection Challenge

This instability class remains undetected in practice because the primary monitoring signal — SLA compliance — continues to show success while structural costs accumulate invisibly. Average-case metrics and compliance checks are designed to detect service degradation, not economic inefficiency. The problem only becomes visible when safety margins are exhausted and SLA breaches begin, at which point significant resources have already been consumed by compensatory mechanisms.

Growing tail latency, rising per-request costs, and shrinking safety margins appear as separate trends rather than manifestations of a single structural cause.

6. Decision Relevance

If inference serving shows growing tail latency and rising costs despite stable average metrics and maintained SLA compliance, the underlying cause is likely structural coupling that overprovisioning will not resolve.

Additional capacity may temporarily restore margins but increases the coupling surface area between replicas, potentially accelerating the eventual instability.

Structural visibility into replica coupling dynamics enables intervention at the load distribution layer, addressing the root cause rather than compensating for its symptoms through capacity increases.

Related Document: SORT AI Interconnect Application Context Brief

---

Such structural findings are typically contextualized through a scoped architecture risk assessment.