Application Context Brief

Runtime Control Coherence
in Large-Scale AI Systems

Gregor Herbert Wegener
Independent Researcher, SORT Framework

Companion document to the article:
"SORT-AI: Runtime Control Coherence in Large-Scale AI Systems"

Not a product description. No implementation. No vendor assumptions.

1. Executive Context

Large-scale AI systems are increasingly constrained not by raw compute capacity or interconnect bandwidth, but by the economic consequences of incoherent control decisions across distributed runtime layers. Unlike interconnect stability, which concerns physical communication and synchronization constraints, runtime control coherence addresses conflicts between autonomous decision-making mechanisms operating on shared infrastructure. As system complexity grows and multiple autonomous control mechanisms operate concurrently, performance degradation and cost escalation often emerge long before any technical failure is detected. These effects typically arise from conflicts between schedulers, orchestrators, runtime engines, model-level control mechanisms, and policy enforcement layers, where misaligned control intents and competing optimization objectives remain largely invisible to classical monitoring and performance analysis. As a result, organizations experience rising cost per performance, non-deterministic behavior, and diminished reproducibility without a clear diagnostic explanation. This brief exists to contextualize this problem space from a decision-oriented perspective and to clarify why runtime control coherence has become a first-order economic and governance concern. It complements the accompanying article by translating its core insights into a concise framework for architectural and strategic decision-making, and serves as the logical counterpart to the Interconnect Stability analysis that addresses physical and topological coupling effects.

2. Problem Statement

The Problem This Brief Addresses

Runtime instability in large-scale AI systems does not arise solely from component failures, interconnect limitations, or software defects, but from structural conflicts between multiple autonomous control loops operating on shared infrastructure without mutual coordination. As workloads scale across distributed environments, control decisions made by schedulers, orchestrators, runtime execution engines, model-level mechanisms, and safety or policy layers interact in ways that were neither anticipated nor designed for. Each control loop pursues its own optimization objective, reacts on its own timescale, and actuates through its own mechanisms, yet all operate on the same underlying operator graph and compete for the same resources.

These effects typically manifest as soft degradation rather than hard failure. Systems continue to run, jobs complete, and monitoring dashboards remain nominal, yet effective throughput declines, latency variance increases, and execution outcomes become non-reproducible. Classical metrics such as utilization, queue depth, or error rates capture symptoms in isolation, but fail to expose the structural origins of these degradations. Control loops may each report healthy status according to their own objectives while their combined behavior produces global inefficiency. As

a consequence, instability is often misclassified as transient load variation, software inefficiency, or resource contention rather than recognized as a systemic property of the control architecture.

The economic outcome is a growing divergence between provisioned capacity and realized performance, accompanied by the accumulation of ghost costs. Ghost costs represent economic expenditures that arise from control incoherence but are not attributable to any identifiable fault: accelerator time consumed without proportional progress, redundant execution driven by uncoordinated retries, and engineering effort spent diagnosing phantom issues that arise from emergent interactions rather than identifiable defects. Over-provisioning may temporarily mask symptoms, but it expands the control surface on which incoherent interactions can occur and amplifies cost per performance over time. This brief addresses this gap by framing runtime control coherence as an architectural and economic problem that requires structural analysis of control-plane interactions rather than incremental capacity expansion.

3. Explicit Scope Boundaries

This brief is intentionally limited in scope. It is not a product description, nor does it introduce a monitoring tool, diagnostic software, or deployable system component. No implementation details, reference architectures, or operational blueprints are proposed or implied. The purpose of this document is analytical and contextual rather than technical or prescriptive.

The analysis does not require access to internal systems, proprietary telemetry, or sensitive operational data. All considerations are derived from publicly observable system characteristics and generally applicable assumptions about large-scale distributed execution environments with layered control architectures. Consequently, no non-disclosure agreement or privileged access is necessary to engage with the concepts presented here.

Furthermore, the brief is vendor-agnostic and does not assume a specific scheduler, orchestration platform, runtime framework, or policy enforcement mechanism. The structural phenomena discussed are intended to apply across a broad class of AI systems where multiple autonomous control loops operate concurrently, independent of implementation choices. This explicit boundary ensures that the brief serves as a neutral framework for understanding control coherence risk and its decision relevance, rather than as an implicit endorsement of any particular technology or solution.

4. Intended Audience

This brief is intended for stakeholders responsible for the design, operation, and economic efficiency of large-scale AI systems operating under layered control architectures. It is particularly relevant to platform engineering leadership and MLOps teams who manage distributed training and inference environments where multiple control mechanisms interact, as well as runtime owners tasked with maintaining stability and predictability under increasing system complexity.

The analysis also addresses architecture review boards and risk committees that evaluate proposed changes, assess scaling decisions, and approve architectural directions. For cost, efficiency, and capacity planning stakeholders, the brief provides a structured lens through which rising cost per performance and ghost cost accumulation can be understood beyond conventional utilization metrics. Finally, the perspective is applicable to government, defense, and regulated compute environments, where reliability, auditability, and economic predictability are critical, and where systems may appear operationally healthy while exhibiting significant hidden inefficiencies and governance blind spots arising from the inability to reconstruct control-plane decision chains.

5. Decision Relevance

Decisions This Context Supports

The perspective outlined in this brief is designed to support strategic and architectural decision-making in environments where traditional performance indicators no longer provide sufficient

guidance. It helps clarify when system behavior becomes unpredictable as a result of structural conflicts between autonomous control loops, and why additional hardware, isolated optimizations, or capacity expansion often fail to restore economic efficiency when the underlying control architecture remains incoherent.

By reframing stability as a control coherence concern rather than a purely operational metric, the analysis enables decision-makers to reason about system behavior at the level where costs are actually incurred. This includes understanding when over-provisioning serves only as a compensatory measure for unresolved control conflicts that expand rather than contract with additional resources, and which diagnostic questions should be asked before committing to further capacity expansion. The intent is not to prescribe specific technical outcomes, but to make structurally relevant trade-offs visible so that decisions can be taken with greater clarity, economic awareness, and governance accountability.

6. Relation to Further Analysis

This document is intended to provide orientation within the relevant problem space by framing control coherence risks and their economic consequences at an abstract, system-independent level. It does not constitute a system-specific assessment, nor does it attempt to evaluate individual architectures, configurations, or operational practices. Instead, it establishes a shared analytical context that can be used to determine whether deeper, targeted analysis is warranted.

This brief serves as the logical counterpart to the Application Context Brief on Interconnect Stability. While the Interconnect Stability analysis addresses physical and topological coupling effects arising from communication patterns and synchronization dependencies, the present brief addresses logical and control-plane coupling effects arising from conflicts between autonomous control mechanisms. The two analyses are intentionally non-overlapping but composable, and their combined use enables a comprehensive evaluation of structural risk across both hardware coupling and software control domains. Where Interconnect Stability addresses how systems fail due to physical coupling, Runtime Control Coherence explains why systems fail even when physical coupling is no longer the dominant constraint.

Relation to Architecture Risk Briefings

Architecture Risk Briefings build on the structural perspective outlined here by examining specific runtime classes and control architectures under explicitly stated assumptions. These briefings remain implementation-agnostic, do not require access to internal systems or proprietary data, and avoid prescriptive recommendations. Their purpose is to make control coherence risks, ghost cost exposure, and auditability gaps visible at the architectural level, not to modify or redesign existing systems. Such briefings are conducted as paid analytical engagements and are intended to support internal evaluation and decision-making processes.

Contact

Gregor Herbert Wegener
Independent Researcher, SORT Framework

LinkedIn: linkedin.com/in/gregorwegener
ORCID: 0009-0008-1791-7487
Email: gregor.wegener@gmail.com

Reference: SORT Whitepaper v6 (DOI).