

Settlement Layers, Multi-Agent Workflows and Stabilization

Governing fast decisions and slower consequences

V2 Reconstruction • Conceptual chapter • September 2026

Summary

An enterprise can think faster than it can fulfill its commitments. Agentic systems sharpen that difference: a decision may be generated rapidly, a service may acknowledge it seconds later, and its operational consequences may remain unfinished for hours or days. GUDIYA proposes that Grid approval and runtime stabilization explicitly govern this interval between initiation and verified settlement.

This chapter introduces five settlement-time bands, explains how a multi-agent Intent-Group can cross them, and specifies the information needed to manage unfinished commitments. The bands are a proposed architectural convention, not an established standard or a proof of stability. Their purpose is to make temporal dependencies visible enough to govern and test.

1 An enterprise that runs on several clocks

Imagine a distributor preparing for a busy working day. Its agents continually examine orders and stock availability. Its order services confirm transactions in seconds. Warehouse allocation takes longer. Planning and reconciliation run overnight. Suppliers deliver over several days. These activities belong to one enterprise, but they do not share one settlement clock.

A purchasing agent notices a shortage and submits an order. Seconds later, the order service accepts it. The stock display still shows the shortage because no goods have arrived. If the agent treats that unchanged display as evidence that nothing happened, it may order again. Another agent may reach the same conclusion independently. Every order can be valid and authorized while the aggregate obligation becomes excessive.

The relevant state is therefore richer than current inventory. It includes accepted orders, reserved capacity, expected arrivals, unresolved exceptions and cancellations not yet confirmed. A dashboard showing only completed effects leaves the enterprise unable to reason reliably about consequences already in motion.

This is a candidate mechanism for what GUDIYA calls AI-Vertigo: agents lose operational orientation when their decisions outrun visibility into the effects of previous decisions. Different cycle times alone do not establish instability. The hypothesis concerns delayed feedback combined with inadequate commitment tracking, repeated intervention or conflicting corrective actions.

2 What settlement means

Settlement time is the elapsed time from an accepted commitment until its specified completion conditions are met and sufficiently verified for dependent actions to proceed. The definition must identify the relevant boundary. A reservation can settle before delivery; delivery can settle before financial reconciliation. None implies that every downstream consequence has ended.

Three times must be distinguished. Computation time measures how long an agent takes to produce a decision. Response time measures how long an interface takes to return a result. Settlement time measures how long the agreed operational commitment takes to reach its verified completion condition. A successful response may acknowledge acceptance without certifying settlement.

An operation also has an initiation frequency: how often new commitments are accepted. A five-day fulfillment process may safely accept many independent orders concurrently. Conversely, a rapid local action may have irreversible consequences. Neither long duration nor high speed, by itself, determines risk.

3 Five settlement layers

The proposed enterprise view uses non-overlapping bands. These layers describe operations and milestones rather than the permanent identity of an application or agent. Thresholds should be calibrated to the domain and measured workload.

Layer	Settlement interval	Illustrative boundary
1 Immediate	≤ 100 microseconds	A bounded local state update is complete.
2 Rapid	> 100 microseconds to ≤ 10 seconds	An order or reservation is confirmed.
3 Operational	> 10 seconds to ≤ 1 hour	An allocation or provisioning workflow completes.

Layer	Settlement interval	Illustrative boundary
4 Daily	> 1 hour to ≤ 24 hours	A planning or reconciliation cycle completes.
5 Extended	> 24 hours	Delivery or a multiday approval completes.

Layer 1 is not a claim that general-purpose LLM inference, distributed consensus or physical business fulfillment normally completes within 100 microseconds. It provides a band for qualifying local operations. Whether a particular operation belongs there requires measurement and a precise completion criterion.

One inventory application may contain Layer-2 acknowledgements, Layer-3 allocations, Layer-4 reconciliations and Layer-5 replenishment. The resulting architecture is a dependency graph annotated with settlement bands, not a mandatory five-tier deployment stack. Operations need not pass through every layer in sequence.

Expected bands and actual durations must be tracked separately. A normally rapid transaction may become overdue under congestion. The Grid must preserve its original deadline, show its current age and reassess its risk. Relabeling it into a slower band must not erase a breached commitment. Layer 5 also requires finite deadlines or review checkpoints; it must not become an indefinite waiting room.

4 Inter-timescale rendezvous

A rendezvous between different settlement layers must establish what can proceed before completion. The downstream participant should convey acceptance capacity, pending status, expected completion, freshness of evidence and the rules for retries or revisions. Backpressure is the protective signal that limits further initiation when downstream capacity is constrained; it is not itself the instability.

Fast cognition can continue inside a bounded action envelope. An agent may repeatedly evaluate demand while being allowed only one unresolved replenishment commitment for the same business need. It can revise a forecast without creating another order. Where additional independent demand is genuine, the system can admit it against remaining capacity rather than imposing a blanket wait.

Dependency conditions should be milestone-specific. Dispatch may require verified inventory allocation but need not wait for end-of-day accounting. A financial close may require both shipment and invoice evidence. Excessively broad barriers would serialize the enterprise unnecessarily; permissive barriers would release work before its prerequisites exist.

Pending, failed and unknown are different states. A lost acknowledgement does not prove that an operation failed. If an agent cannot establish whether a request took effect, it must reconcile by commitment identity or use an approved recovery procedure before creating a replacement. A timeout signals that expected evidence has not arrived; it does not automatically authorize duplication.

5 Settlement-aware Intent-Group approval

When a multi-agent Intent-Group is submitted to the Grid, it should include a Settlement Profile for every consequential operation and its dependencies. Agent-level defaults may help discovery, but they cannot replace operation-level declarations. Approval should cover the proposed composition as well as the individual participants.

Profile field	Approval question
---------------	-------------------

Profile field	Approval question
Commitment identity and owner	Can the Grid follow one business commitment across agents, retries and handoffs?
Completion condition and evidence	What proves this milestone has settled, and who can attest to it?
Expected band and timing	What are the duration range, deadline and review checkpoints?
Dependencies and versions	Which milestones must settle, and is the state used to decide still applicable?
Outstanding limits	How much quantity, value, capacity or irreversible exposure may remain pending?
Retry and revision policy	Which retries repeat the same intent, and which changes need new approval?
Recovery and enforcement	Who may pause initiation, reconcile uncertainty, compensate or authorize restart?

Declared timing is a claim requiring evidence from tests, operating history or a conservative initial envelope. Uncertain or novel operations should receive restricted approval with observation and escalation conditions. A changed supplier, new tool, altered dependency or increased commitment size may require reassessment even when the nominal layer remains unchanged.

Approval is not permanent permission to initiate at maximum speed. The initial envelope defines limits; runtime telemetry can narrow it. Local enforcement should remain effective when regional communication is delayed. A central Grid round trip is not feasible for every Layer-1 event: preauthorized bounds and local checks must protect fast paths, with escalation and aggregation at slower timescales.

6 A worked enterprise example

Consider a hypothetical distributor whose approved replenishment envelope permits at most 100 units outstanding for a particular item and destination. This number is illustrative. A supplier order for 60 units is accepted within seconds, while delivery is expected in three days. The acceptance milestone is Layer 2; fulfillment is Layer 5.

Another agent observes the same shortage. The shared commitment record shows 60 units already pending. If the second proposal requests 50 additional units, it exceeds the remaining 40-unit allowance. The Grid can reject it, admit a smaller justified quantity, or send the proposed expansion for review. The unchanged on-hand inventory does not restore the consumed allowance.

On the second day, 20 units arrive and their receipt is verified. With no other changes, 40 units remain outstanding and 60 units of the original outstanding-order allowance are available. That arithmetic does not itself justify another order: the new inventory, revised demand, supplier capacity and other constraints must still be evaluated.

Now suppose the remaining delivery becomes uncertain. The agent proposes an emergency order elsewhere. The original 40 units remain an obligation until cancellation or another disposition is confirmed. A cancellation request alone does not release that exposure. If both shipments might arrive, DDMC must count that possibility before approving the substitute.

This example makes the distinction tangible. The system does not stop an agent from thinking. It governs the rate and size of commitments whose consequences the enterprise must absorb. The policy also prevents several individually compliant Intent-Groups from independently consuming the same shared allowance.

7 The five stabilization disciplines

Discipline	Settlement-layer responsibility
Admittance Control	Admit operations and new dependencies only within supported timing and exposure limits.
DDMC	Estimate current and projected margins, including pending commitments, age and uncertain outcomes.
DLPC	Choose the producer, dependency or shared resource where intervention can best reduce harmful accumulation.
Cognitive Rx	Apply pacing, batching, consolidation, reconciliation or bounded compensation, then verify effects.
Circuit Breaking	Stop a hazardous initiation path when margins cannot be preserved; protect evidence and safe completion paths.

These functions must distinguish stopping new work from abandoning work already accepted. Disconnecting all communication can prevent receipt confirmations, cancellation acknowledgements or recovery actions. A selective circuit break may block new purchases while preserving status queries and verified cancellation channels. Restart should require reconciliation and explicit readmission conditions.

The stabilizer also creates consequences. Throttling a supplier interface may miss a fulfillment deadline; cancelling one workflow may release resources another workflow still assumes are reserved. Cognitive Rx therefore needs its own settlement profile, scope limits and evidence of effect. The intervention cannot be treated as instantly successful merely because its command was issued.

8 What DDMC should measure

For a bounded scope such as an item, production line or business service, a simple accounting relation is useful: outstanding exposure equals previous exposure plus newly accepted obligations, minus verified releases. Releases may follow completion or confirmed cancellation. Compensation counts as a release only to the extent that it actually removes the relevant obligation or impact.

Available margin can then be expressed as the allowed exposure minus outstanding exposure minus a reserve for uncertainty. These quantities must share meaningful units. Orders, dollars, warehouse slots and recovery effort should not be casually added together. DDMC should retain a vector of constraints and show which one currently limits admission.

The Grid should monitor pending quantity and value, age by deadline, projected completion, overdue proportion, evidence freshness, duplicate attempts and recovery backlog. A small queue of irreversible commitments may be more consequential than a large queue of harmless queries. Means alone can hide long-tail settlement delays and synchronized deadline failures.

Forecasts should examine the relevant settlement horizon. A ten-second service may require seconds of look-ahead; a three-day supplier commitment requires a longer forecast and intermediate checkpoints. No universal 30-second horizon can cover all five bands. A longer horizon also increases uncertainty, which should reduce confidence in available margin rather than produce false precision.

9 Established foundations and the proposed extension

This problem did not begin with agentic AI. Reactive Streams explicitly addresses asynchronous processing with non-blocking backpressure [1]. AWS documents idempotent API design as a way to make retries safer [2]. These are important foundations: GUDIYA should reuse them rather than claim that older IT never managed timing or unfinished work.

Transport-level deduplication does not by itself establish that two differently worded agent requests serve the same business intent. Conversely, two similar requests may represent legitimate independent demand. GUDIYA therefore proposes to associate protocol-level controls with domain-defined commitment identity, settlement evidence and shared operating envelopes. The semantic matching must itself be tested and governed.

McKinsey Global Institute describes an AI economy whose capabilities, infrastructure, organizations and institutions evolve at different speeds [3]. This provides broader motivation for examining timescale mismatch. It does not establish the five bands or validate a Grid implementation. The chapter's contribution is a proposed enterprise design lens and a set of requirements that can be evaluated experimentally.

10 How the proposal can be tested

Begin with one workflow that spans several settlement bands. Compare the existing implementation with a strong baseline using ordinary backpressure, retry limits and idempotency. Then add explicit business commitment tracking, dependency-aware approval and dynamic exposure limits. Keep workloads and failure scenarios comparable so improvements are not merely the result of allowing less useful work.

Test delayed and lost acknowledgements, partial completion, duplicate initiation, stale state, competing Intent-Groups, supplier substitution and interruption of Grid connectivity. Include recovery after circuit breaking and test whether local limits remain coherent when distributed observers disagree. A margin that several nodes spend independently is not an enforceable shared limit.

Measure duplicate obligations, maximum outstanding exposure, missed deadlines, recovery time, useful throughput and unnecessary rejection. The proposal is strengthened if it reduces damaging accumulation without disproportionate loss of service. It is weakened if existing controls perform equally well or if its coordination overhead creates new failures. Such results should refine the bands and policies rather than be hidden by the terminology.

11 A settlement-aware enterprise

The enterprise does not need every participant to operate at one speed. It needs each participant to understand which consequences remain unfinished, what capacity they consume and what evidence allows the next dependent action. Settlement layers make those questions visible in architecture; Intent-Group profiles carry them into approval; the Grid Loop maintains them during execution and recovery.

Proposed GUDIYA principle: authorization to initiate an action is not evidence that its consequences have settled. Fast cognition may continue, but new commitments must remain bounded by the coupled enterprise's capacity to execute, verify and recover from what is already in motion.

References and scope

[1] Reactive Streams. Official project and specification overview. <https://www.reactive-streams.org/> — establishes the backpressure foundation cited in Section 9.

[2] Malcolm Featonby. Making retries safe with idempotent APIs. Amazon Builders' Library. <https://d1.awsstatic.com/builderslibrary/pdfs/making-retries-safe-with-idempotent-apis-malcolm-featonby.pdf> — establishes the retry and idempotency foundation cited in Section 9.

[3] McKinsey Global Institute. The AI economy: Interconnected forces, feedback loops, and speeds of change. September 10, 2026. <https://www.mckinsey.com/mgi/our-research/the-ai-economy-interconnected-forces-feedback-loops-and-speeds-of-change> — contextual motivation from the discussion reviewed in this research exchange.

The five settlement bands, Settlement Profile and Grid approval requirements in this chapter are GUDIYA proposals developed in this research conversation. Enterprise scenarios and numeric limits are illustrative, not reported deployments or validated performance claims.

The Gudiya Effect

An IT transaction may complete while its consequences continue unfolding. A sales order can reserve inventory, trigger production, commit transport capacity, and create delivery obligations across enterprises, industries, and countries. The proposed Gudiya-Effect encompasses the initiating action and its attributable downstream consequences within a declared scope. It establishes continuity of observation and responsibility beyond the moment an application reports success.

As the AI economy crosses organizational and national boundaries, no single application or enterprise necessarily possesses visibility of the complete journey. The Federated Gudiya Grid is therefore a suitable proposed architecture for tracking a Gudiya-Effect from initiation through settlement and accountable closure. Participating domains would retain their operational authority and records while exchanging the identifiers, obligation states, evidence, and stability signals needed to manage cross-boundary dependencies. Federation must establish shared meanings and responsibilities while respecting access controls and jurisdictional boundaries.

Each effect instance receives a Gudiya-Effect number, while the Gudiya-cookie provides continuity across participating agents and linked records. Its effect graph connects actions, dependencies, outstanding obligations, and evidence of fulfillment. A single effect may span several settlement layers: an order accepted in milliseconds may initiate a reservation settled in seconds, dispatch completed in hours, and delivery verified days later. The federated view helps prevent rapid upstream activity from overwhelming slower downstream fulfillment.

Settlement management asks: *What remains to be fulfilled or resolved?* Stabilization asks: *Can this effect, together with other effects already unfolding, remain within the ecosystem's operating margins?* Local Grid components would manage local conditions; coordination across domains would address shared constraints and propagating consequences.

“Cradle to grave” means continuity through a defined lifecycle, including accountable handoffs—not unlimited surveillance of every possible consequence. Closure requires agreed evidence and explicit handling of unresolved obligations or residual effects. Where visibility is incomplete, the Grid must preserve that uncertainty. The architectural aim is to ensure that responsibility survives changes of agent, system, enterprise, and jurisdiction until the effect's declared obligations and consequences have been appropriately resolved.

<https://www.manhattanproject20.com/blogs/the-gudiya-effect>

