The Floor Nobody Measures

Better inference requires more orchestration. More orchestration requires session authority. The energy saved by getting that right is not a side effect. It is the point.

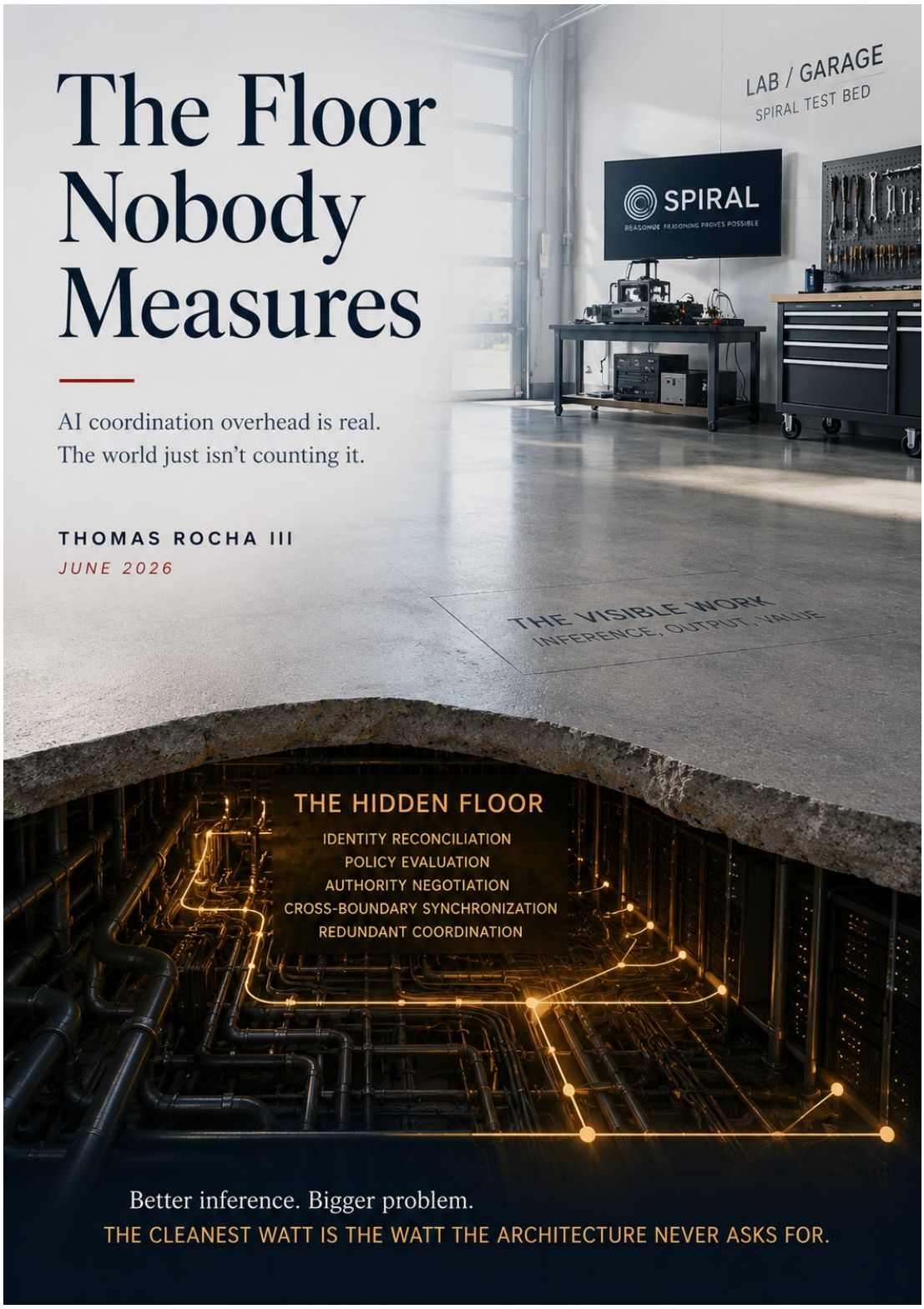
THOMAS ROCHA III

JUN 25, 2026

Abstract

The AI energy debate has two positions, and both are correct as far as they go. The public says data centers are consuming resources that communities never agreed to absorb. The industry says efficiency is improving, and tokens per watt is the right metric. Neither side is wrong. Neither side is measuring the floor.

The floor is coordination overhead: identity reconciliation, policy evaluation, authority negotiation, and cross-boundary synchronization. It fires before the model runs. It fires between traces in a parallel inference pipeline. It fires at every surface the session crosses, independently, multiplicatively, before a single useful token is produced. No chip roadmap moves it. No renewable energy purchase eliminates it. No efficiency gain at the model layer reaches it



The Floor Nobody Measures

AI coordination overhead is real.
The world just isn't counting it.

THOMAS ROCHA III

JUNE 2026

Better inference. Bigger problem.

THE CLEANEST WATT IS THE WATT THE ARCHITECTURE NEVER ASKS FOR.

On June 22, 2026, Stanford published SPIRAL: LEARNING TO SEARCH AND AGGREGATE, one of the most architecturally honest papers the field has produced about where inference is going. SPIRAL proves that better reasoning requires more orchestration. More orchestration, on fragmented infrastructure, multiplies the coordination surface with every additional trace. Session governance collapses that surface from a product to a sum. The watts not burned are not offset. They are not balanced. They are simply not produced.

This essay explains why the floor exists, what SPIRAL proves about where it is going, and why session governance is an energy argument, not just a control-plane argument.

The AI energy debate has two sides, and neither one is wrong.

The public argument: data centers are consuming power, water, tax subsidies, and local grid capacity at a rate that communities did not agree to and cannot easily reverse. The IEA projects data center electricity consumption will roughly double to 945 TWh by 2030. Seven in ten Americans now oppose an AI data center in their local area. Oregon, Texas, and Virginia are already showing the shape of organized resistance at scale.

The industry argument: efficiency is improving. Tokens per watt is the governing metric at every serious infrastructure conversation. Hardware is getting cheaper. Models are getting more efficient. Jensen Huang formalized the frame at GTC 2026: Revenue = (Tokens per Watt) × (Available Gigawatts). The left side is improving. The engineers are not wrong.

Both sides are measuring the visible load.

Neither one is measuring the floor.

What the Dyno Held Fixed

On June 22, 2026, researchers at Stanford published SPIRAL: Sequential-Parallel-Aggregative Reinforcement Learning. It is the most architecturally honest paper the field has produced about where inference is going.

SPIRAL names what serious practitioners already know. A single chain of thought is not enough for hard problems. The answer is search: sample parallel traces independently, reason sequentially inside each, then aggregate across all of them into a final response. The paper demonstrates up to 11x scaling efficiency over sequential-only methods and 15% higher performance when all three compute primitives are scaled together. The results are real. The engineering is serious.

SPIRAL is also a dyno result.

The AI Hot Rod essay from April described the distinction. The dyno measures the engine under controlled load: temperature-controlled room, smooth drum, single measured resistance, clean tires. The numbers are honest. The dyno does not lie. It measures exactly what it measures, under exactly the conditions it creates.

Read what SPIRAL held fixed while the traces ran. One authority domain. One participant set. One modality. One transport. Fixed policy context throughout. The parallel traces fired inside a closed environment where reconciliation between them was never required. The aggregation step operated over traces that already shared a session context by experimental design. Token budgets were equalized across methods to isolate the reasoning improvement. That is correct experimental design.

A skeptical reader will note that SPIRAL's traces are all running inside one system, one operator, one deployment. There are no cross-boundary surfaces in a math reasoning benchmark. That observation is accurate and is precisely the point. The lab is a single-operator context. The street is not. The moment SPIRAL-style inference runs inside an enterprise agent pipeline, each trace that calls a tool, reads from

memory, delegates to a sub-agent, or invokes a third-party API has crossed a surface the lab never encountered. Production does not hold those surfaces fixed. The dyno does not measure them. The street charges for them on every trace, independently, before the aggregation step even begins.

It is also precisely why the result does not transfer directly to the street.

What the Street Does to the Formula

The Coordination Limit essays established the cost structure of fragmented AI infrastructure. When a system coordinates across independently governed participants, modalities, features, authorities, and transports, the coordination cost scales as a Cartesian product of those dimensions:

$C_{frag} = k \times P \times M \times F \times A \times T$

In an interaction-scoped architecture, where all five dimensions bind to a single persistent session identity, the same surfaces collapse to a sum:

$C_{ssoar} = k \times (P + M + F + A + T)$

In plain English: fragmented systems re-check the same interaction at every boundary. Session-governed systems establish the boundary once and carry it through the full interaction.

The difference between a product and a sum. At the IoE scale, that difference is the difference between buildable and not. Part II of that sequence added the system-class point: optimization reduces the constant k, but it does not change the function. You can make the product smaller. You cannot make it a sum by improving the engine.

Now apply that formula to a SPIRAL session running on the street.

In the lab, the parallel traces share a session context by design. On the street, each parallel trace is a participation act. It crosses whatever authority and transport boundaries the session spans. If the underlying infrastructure is fragmented, each trace inherits the full coordination surface of the session. The reconciliation does not divide across traces. It does not average out. It fires independently for each one, and then fires again at the aggregation step, which conditions on all of them simultaneously.

In a fragmented production architecture, SPIRAL-style inference can approach n repetitions of the full coordination surface, plus aggregation overhead across those traces. The formula grows faster than the trace count.

The efficiency gain from SPIRAL operates above the floor. The floor is coordination overhead. SPIRAL does not touch the floor. It improves what runs above it.

That is the gap neither side of the energy debate is measuring.

The Green Case

Here is the claim stated plainly, without qualification:

Whatever AI does next, session governance makes it greener.

Not because anyone designed it to be. Because the math works the same way regardless of motivation. A session boundary established once and held across the full interaction, across all parallel traces, across the aggregation step, across every tool call and delegation and retry, collapses the coordination product to a sum. The product surfaces that would have fired independently do not fire. The watts those surfaces would have consumed are not consumed. They are not offset. They are not balanced by renewable purchases. They are simply not burned.

This is a structural reduction, not a behavioral one. It does not depend on the model being more efficient. It does not depend on the hardware being newer. It does not depend on the grid being cleaner. It is a consequence of the architecture, and it applies at every layer: SPIRAL-style inference, agentic pipelines, multi-party real-time sessions, IoE deployments. Every interaction that crosses independently governed surfaces carries this overhead. Every interaction that does not has already paid the lower cost.

A reasonable challenge: what fraction of current session energy is coordination overhead versus useful compute? The honest answer is that the fraction is not fixed. It expands with every dimension added to the session. More traces, more tools, more agents, more jurisdictions, more policy domains: each one adds a surface to the product. The coordination fraction of a simple single-trace session is small. The coordination fraction of a SPIRAL-style multi-trace agentic pipeline crossing tool boundaries, memory systems, and third-party APIs is not. The argument is not that current deployments are wasting half their energy on reconciliation. The argument is that the fraction grows faster than the compute as inference complexity increases, and the industry's own roadmap is increasing inference complexity as fast as it can.

The more sophisticated the inference becomes, the larger the delta.

SPIRAL-style reasoning, by design, multiplies the number of participation acts per session. More traces, more aggregation steps, more intermediate state, more authority decisions per final answer. On fragmented infrastructure, each of those acts inherits the full coordination surface. The product grows with every additional trace. On session-scoped infrastructure, the boundary is established once. Additional traces do not add coordination surfaces. They share the one that already exists.

This means SPIRAL makes the green case stronger, not weaker. The usual efficiency-versus-performance tradeoff does not apply here.

Better reasoning, more orchestration, more traces: all of it increases the delta between fragmented and session-scoped coordination cost. Session governance gets greener as inference gets better. The two curves move in the same direction.

The industry's roadmap is producing better engines. Better engines on the current infrastructure floor increase the gap that session governance closes.

The Question Nobody Is Asking

The energy debate is asking: how many data centers do we need, and how efficient are the models inside them?

Those are the right questions for the engine and the fuel. They are not the right questions for the floor.

The floor is coordination overhead: identity reconciliation, policy evaluation, authority negotiation, cross-boundary synchronization. It fires before the model runs. It fires between traces in a SPIRAL pipeline. It fires at every surface the session crosses, independently, multiplicatively, before a single useful token is produced. No chip roadmap moves that floor. No renewable energy purchase eliminates it. No efficiency gain at the model layer reaches it.

The Coordination Limit sequence estimated that at IoE scale, coordination overhead under the current architecture approaches 90 to 100 gigawatts before payload runs. Those are projections, not measured values, and they depend on assumptions about deployment scale and session complexity that will sharpen as the architecture matures. The structural conclusion holds regardless of the constant: a product dominates a sum at scale, and the current architecture is running the product.

The answer the public wants is not more data centers. The answer the industry needs is not just better chips. The standard hyperscaler

response, renewable energy purchases and carbon removal credits, addresses the carbon intensity of the watts that are consumed. It does not eliminate the watts. Session governance eliminates watts that would otherwise be consumed. One intervention cleans the fuel. The other reduces the burn. Both matter. Only one addresses the floor.

The answer that addresses the floor is session governance: a boundary that binds all five dimensions to a single persistent identity, collapses the product to a sum, and makes the coordination overhead measurable, attributable, and eliminable.

SPIRAL proves the inference shape. The inference shape proves the urgency. The urgency is not just about performance.

Current production deployments are simpler than SPIRAL's full pipeline. That is not a reason to wait. Infrastructure decisions made in 2026 shape the cost and carbon profile for the decade that follows. The time to establish session governance as the coordination substrate is before SPIRAL-style inference becomes the production default, not after the architecture is already built around the wrong floor. The window is open. The direction is visible. The watts that will be burned when that architecture arrives at scale are not yet committed.

It is about the watts that are burning before the answer starts, and the far larger number that will burn if the floor is never addressed.