Compute on the Side of the House

"Ah, watch out You might get what you're after" Talking Heads

THOMAS ROCHA III

MAY 07, 2026

Jensen Huang is right about the first half of the story. At the Morgan Stanley TMT Conference, Huang framed compute as foundational to the modern economy and tied the next phase of AI to agentic systems and efficiency measured in tokens per watt.

The Span, Nvidia, and PulteGroup story is what that thesis looks like when it leaves the keynote stage and hits the wall of a house.

According to Bisnow, Nvidia and PulteGroup are partnering with Span to test small data centers attached to new homes. Span, previously

known for smart electrical panels, is positioning this as a shift toward digital infrastructure: small Nvidia-powered nodes placed on residences and small businesses. Span's framing addresses what they call the "speed-to-power gap" for AI compute demand.

That is the important phrase. Not the model gap. Not the benchmark gap. Power. Location. Permitting. Grid attachment. Physical infrastructure.

Huang says compute is becoming foundational. Span's answer: if compute cannot get to power fast enough, move compute to where power exists. That sounds clever. It is also a near-perfect illustration of the failure pattern SSOAR is built around.

Once compute moves from hyperscale campuses into residential infrastructure, the hard problem is no longer how to run the workload. The hard problem becomes: who has authority over the live interaction among the homeowner, the builder, the smart panel, the utility, the compute operator, the AI workload, the grid, the insurer, the jurisdiction, the hardware vendor, and the customer whose data or inference task is being processed? The system does not have an answer.

The industry keeps mistaking capability for authority

A box with Nvidia GPUs can run workloads. A smart panel can meter power. A builder can install equipment. A cloud operator can route jobs. A utility can manage load. None of that establishes runtime authority.

This is the fourth C: Capability.

The mini-data-center narrative depends on a slide from capability into authority. The system can run compute, therefore it may be treated as authorized compute. The home has electrical capacity, therefore that capacity may be treated as available infrastructure. The homeowner opted in once; the runtime relationship may be treated as continuously

valid. Capability becomes agency, agency becomes authority, authority becomes assumed sovereignty. The same grammatical error infects agentic AI.

An aircraft may be capable of landing, but clearance is session-scoped, revocable, and specific to a maneuver. Capability survives. Admission can end.

A residential GPU node may be capable of running inference. That does not mean it has standing authority to consume power, process data, route workloads, or maintain compliance across changing conditions.

If the authority boundary is not explicit, capability will be treated as authority by default. That is how distributed systems get themselves into trouble.

The Three Cs

Compliance is the proof failure. Coordination is the scale failure. Concurrence is the timing failure. Together, they explain why putting compute closer to power does not solve the problem. It only relocates it.

Compliance: the proof failure

The residential node is not just hardware. It is an operational claim. Someone will have to prove, not merely assert: the homeowner consented; the device stayed within electrical limits; the workload stayed within permitted data boundaries; the insurer knew what risk it was underwriting; the jurisdiction allowed the installation; the node was disabled when policy, safety, weather, grid, ownership, or trust conditions changed.

Compliance is shifting from configuration to runtime behavior. The question is no longer whether the system was designed with reasonable

controls. The question is whether the system can prove what was allowed during execution, at the moment the thing happened.

A temporary equilibrium exists while two statements remain true: vendors cannot produce real-time proof of behavior, and authorities cannot yet require it. Once one credible system can produce runtime proof, the equilibrium changes.

Suppose the node runs inference during a grid curtailment event, or the homeowner sells the house, or the node processes data from a regulated customer. Did the operating authority update? Did the system know the jurisdiction of the data, the node, the operator, and the contractual boundary at the moment of routing? Post-hoc reconstruction will not be enough forever.

Coordination: the scale failure

One home node is a pilot. A million are a distributed control problem. Once the model scales, every node introduces coordination work: device provisioning, firmware updates, identity management, hardware attestation, physical security, grid scheduling, thermal management, homeowner support, utility coordination, insurance claims, workload placement, data residency, audit, decommissioning, local code compliance.

This is the Coordination Limit. The cost is not linear. Each node introduces new combinations of participants, modalities, features, authorities, and transport boundaries.

A residential AI node is not just a smaller data center. It is a boundary generator. Every boundary performs work: identity established, policy evaluated, state synchronized, authority reconciled. That work consumes compute, power, bandwidth, attention, legal and compliance overhead. It produces nothing the end user sees.

Huang's equation is about tokens per watt and available gigawatts. Necessary, but incomplete. Effective output is closer to compute capacity minus coordination overhead. If coordination overhead grows faster than useful workload capacity, the system becomes coordination-bound.

Putting GPUs on houses does not eliminate the coordination problem. It puts the coordination problem on the side of the house.

Concurrence: the timing failure

The hardest part is not that many parties are involved. The hardest part is that they change at the same time. A residential compute node is not static. Its authority state mutates continuously: grid, household load, weather, utility instructions, compute demand, network paths, device posture, firmware, insurance status, ownership, local law, AI workload routing, homeowner consent.

Two constraints are independent only when the interval between their state changes exceeds the time required to reconcile them. If state changes faster than reconciliation, independence collapses.

If workload placement moves inference to a home node at the same moment the utility issues a curtailment signal, the smart panel detects load change, the firmware monitor reports degraded state, and a data residency policy updates, which one wins? "Eventual consistency" is wrong for a live governed interaction. "Logs" is too late. "The cloud orchestrator decides" means the orchestrator has become de facto authority, whether anyone admitted it or not.

Edge compute promises proximity. It does not automatically provide authority. The closer compute moves to messy reality, the faster authority conditions mutate.

Failure domains this story activates

Efficiency Paradox. The project is sold as efficiency: faster deployment, cheaper compute, better use of existing capacity. Efficiency gains can be reversed by coordination overhead. The system may spend more on maintaining the distributed estate than it saves by avoiding a centralized buildout.

Zero Trust Security. A residential GPU node is a strange trust object. Owned by one party, installed by another, powered by another, insured by another, patched by another, scheduled by another, used to process workloads from still another. Zero Trust cannot be a login event here. It must be continuous authorization of each material action. Authentication is not authority.

Data Residency and Sovereignty. If AI workloads can be routed to distributed homes, "where was the data processed?" stops being a simple cloud-region question. Was it processed in a house? In which city? Under which utility? Under which state law? Data residency must be enforced at routing time, not reconstructed from storage records.

AI Coordination. Agentic AI makes the problem worse. Workloads will not be static batch jobs. Dynamic, agentic, multimodal, tool-using AI means more live decisions, more tool calls, more real-time placement decisions. The node is not merely executing inference. It is participating in an active interaction chain. A model cannot be the boundary for the actions it is trying to take.

Mobile and Edge Network Complexity. Edge compute is not just smaller cloud. It is cloud plus locality: codes, weather, grid constraints, physical access, household behavior, emergency response, jurisdiction, property ownership, human consent. The more local compute becomes, the more local authority matters.

Concurrency Control. The home wants power, the grid wants stability, the operator wants utilization, the homeowner wants incentives, the insurer wants bounded risk, the regulator wants compliance. Those

interests collide during operation. If the control model is last-write-wins, the system is not governed. It is improvised.

Accessibility and Public Safety. As soon as residential compute participates in communications, emergency services, medical systems, or public infrastructure, the failure is no longer merely commercial. A disabled user does not care that the GPU node, cloud orchestrator, network transport, and accessibility service were each locally compliant. The experience either remained accessible during the live interaction or it did not. All correct locally, system wrong globally.

Insurance is a proxy for authority

Who insures the device? The house? The workload? The data breach? Fire? Grid interaction? Damage caused by bad firmware? A claim when the homeowner used backup power at the same time the node was scheduled for compute?

If the answer is divided among different policies, each with different exclusions, the system is fragmented before it turns on. If no one can underwrite the full live interaction, no one really owns the full risk. The problem is not that every party lacks capability. The problem is that no layer owns the interaction boundary where capability becomes authorized action.

Why monitoring and orchestration do not solve this

The industry will answer: monitor it, orchestrate it, add policy, add telemetry, add AI supervisors. That is the additive trap. Every added layer becomes another participant in the coordination problem.

Monitoring sees, it does not govern. Telemetry reports, it does not authorize. Orchestration routes, it does not prove authority. A policy engine evaluates, it does not necessarily own the session. An AI supervisor reasons, it does not provide an external boundary.

The residential compute model needs one interaction-scoped authority boundary that can evaluate proposed changes before they become system truth.

What SSOAR would change

SSOAR does not require the residential node to become magical. It changes the location of authority. Instead of allowing each subsystem to act locally and reconcile afterward, SSOAR binds the live interaction as the governed object. A proposed workload assignment is a session mutation. A change in power state is a session constraint. A data residency requirement is an admissibility condition. A homeowner opt-out is an authority change. A firmware degradation is a trust mutation. An AI workload request is an action by a participant inside a governed interaction.

Current model: event occurs, subsystem acts, logs are gathered, compliance is reconstructed. SSOAR model: mutation is proposed, session authority evaluates, admissible state commits, inadmissible state is denied. That is not a product feature. It is an architectural correction.

Conclusion

This is not really about Span, PulteGroup, or even Nvidia. It is about what happens when AI infrastructure runs into physical limits and starts distributing itself into the built environment. The industry will keep doing this. Compute will move into homes, cars, factories, hospitals, cell sites, schools, and public infrastructure. Every move will be sold as solving a capacity problem. Every move will create an authority problem.

The question is: who governs the live interaction when compute, power, policy, data, trust, and human context mutate at the same time? If that question is not answered architecturally, it will be answered accidentally by whichever subsystem acts first. That is not governance. That is drift.

Huang is right that compute is foundational. Foundational compute does not become safe, compliant, secure, insurable, and scalable merely by being deployed. The more compute leaves the centralized campus and enters the lived world, the more authority has to travel with the interaction.

The Three Cs name the failure. The fourth C explains the category error. Capability is not authority.

A residential AI node may be capable of running compute. That does not mean the live interaction is governed. And if it is not governed, the future Huang is describing will not fail because the chips were too slow. It will fail because the system could connect everything, power everything, and process everything, while still being unable to prove who had authority when it mattered.