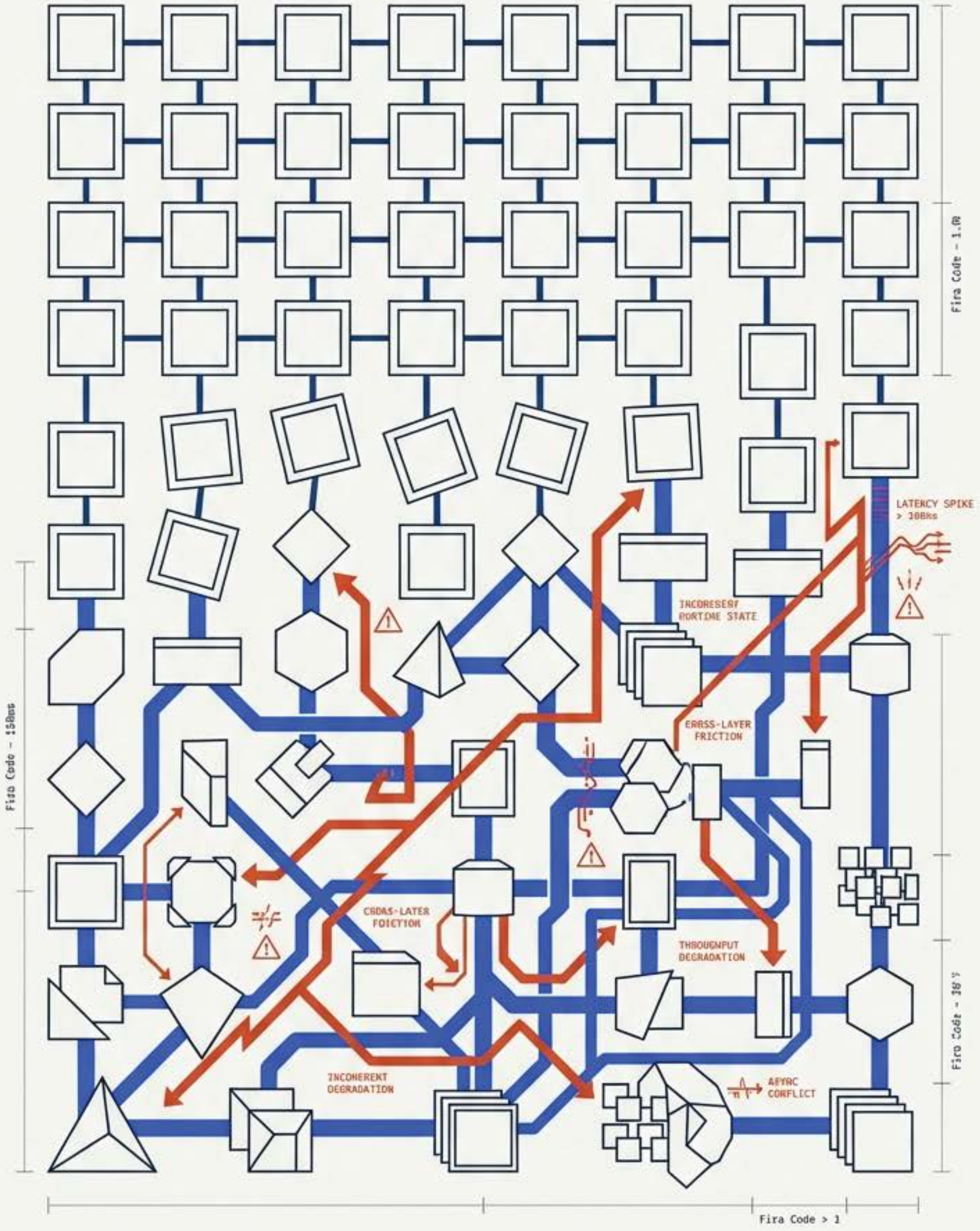
TECHNICAL WHITEPAPER SLIDE. PORTRAIT A4 LAYOUT

THE HIDDEN VARIABLE IN HETEROGENEOUS AI INFERENCE

The era of homogeneous compute is over. AI inference now spans mixed accelerators, disaggregated serving, and multi-cloud boundaries.

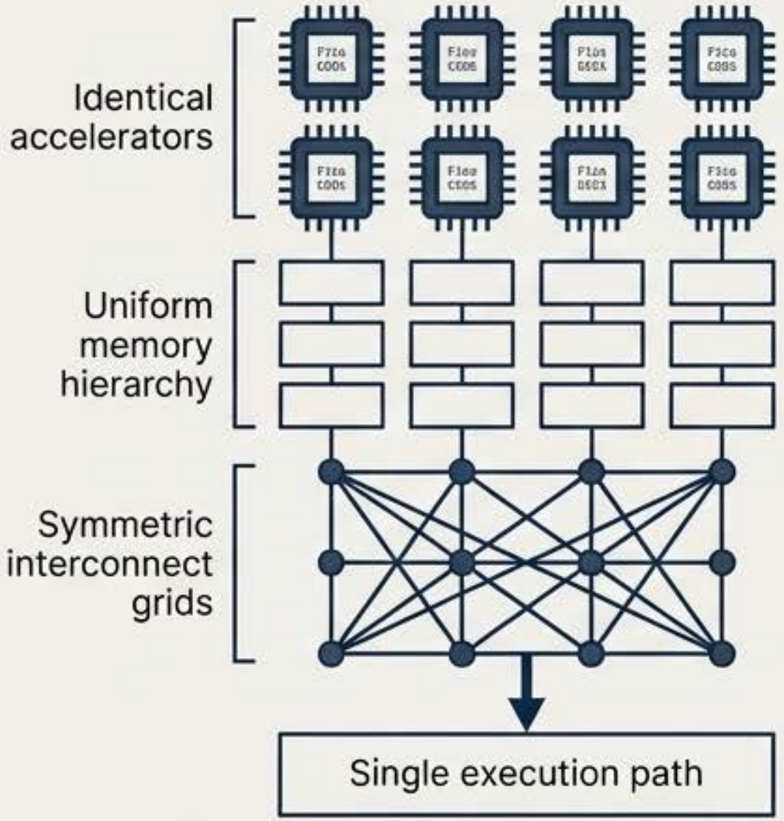
Classical layer-local metrics are failing to capture emergent cross-layer instability.

System degradation no longer starts with component failure—it starts with runtime incoherence.

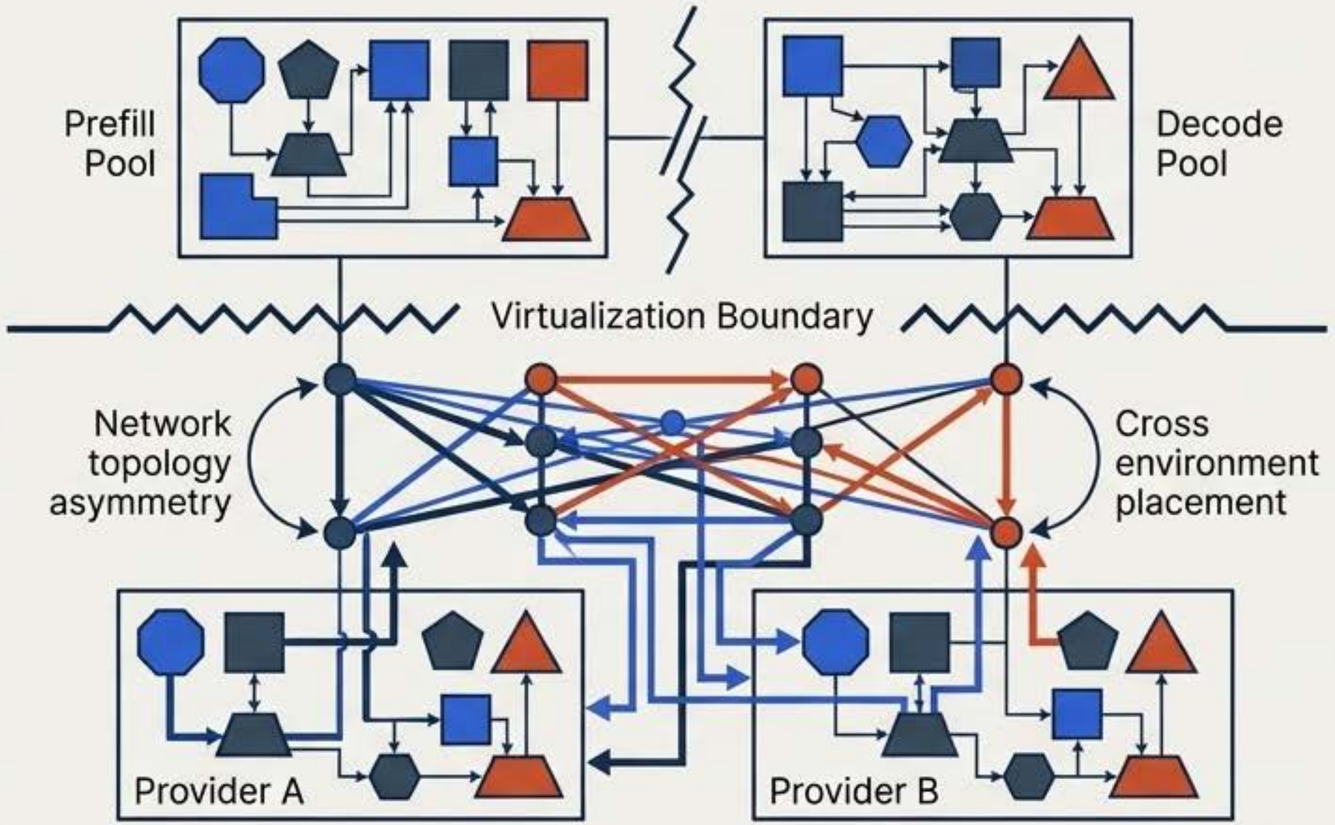


THE SHIFT TO HETEROGENEOUS INFERENCE FABRICS

LEGACY: HOMOGENEOUS COMPUTE



MODERN REALITY: HETEROGENEOUS FABRIC



Compute is no longer uniform. Modern infrastructure mixes **GPU generations, custom ASICs, and specialized NPUs**.

Execution paths are **disaggregated**. **Prefill and decode phases** are split across **distinct hardware pools** with **asymmetric memory hierarchies**.

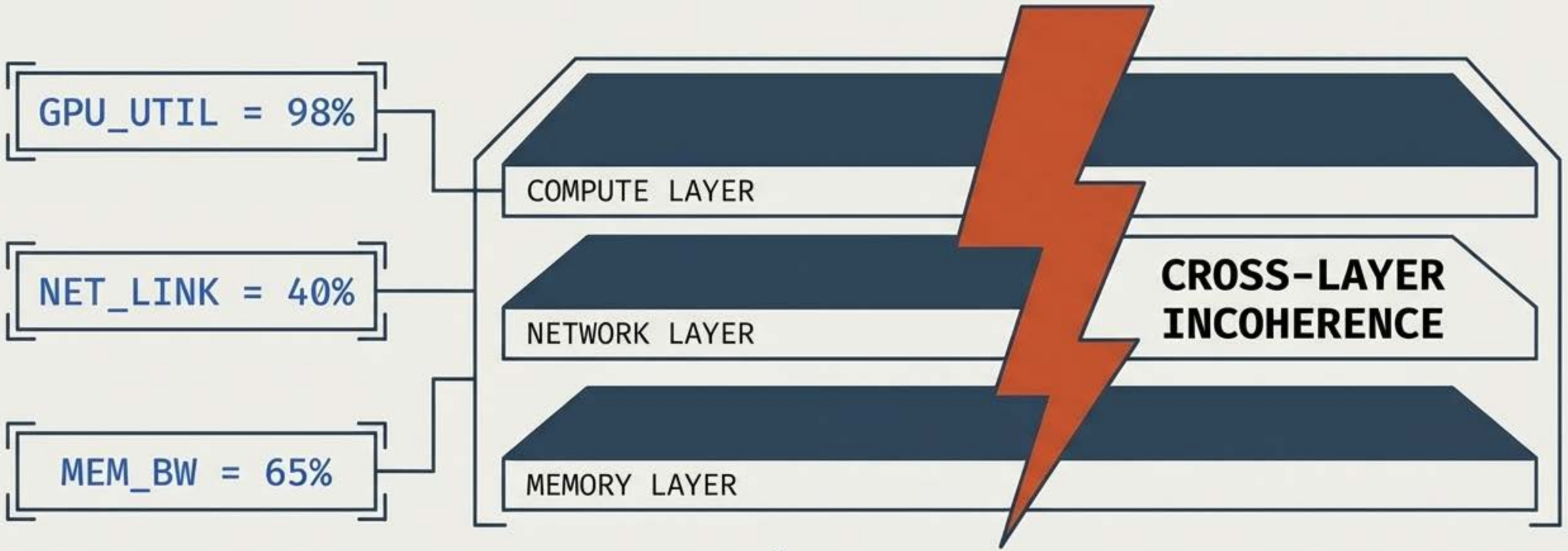
The operational default is diverse. **Multi-cloud, sovereign constraints, and mixed deployments** rewrite the execution surface.

THE MEASUREMENT GAP: YOUR METRICS ARE LYING

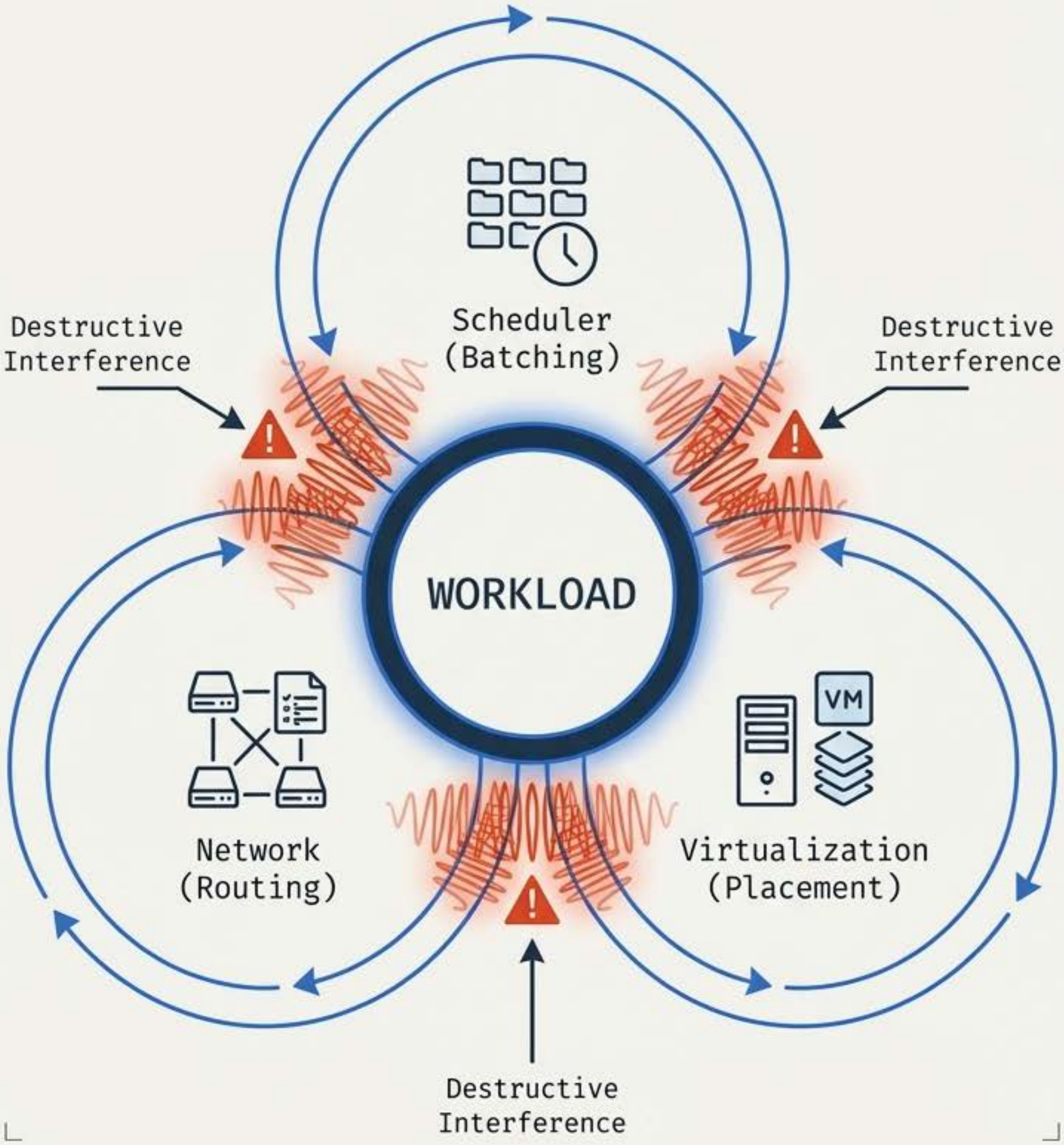
Dashboards mask structural behavior. Layer-local metrics fail to capture cross-layer dependencies.

High utilization often signals inefficiency. The system is compensating for cross-layer friction, not producing effective throughput.

Tail latency hides in blind spots. The slowest structurally exposed path dictates performance, entirely invisible to average metrics.



LOCAL OPTIMIZATION CAUSES GLOBAL INSTABILITY



Inter

Independent Tuning Creates Interference

- Independently optimizing schedulers, network routing, and batching creates cross-layer collisions.

Inter

Success at One Layer Breaks Another

- Aggressive batching alters memory pressure; tighter resource placement increases communication asymmetry.

Inter

Optimization Becomes the Threat

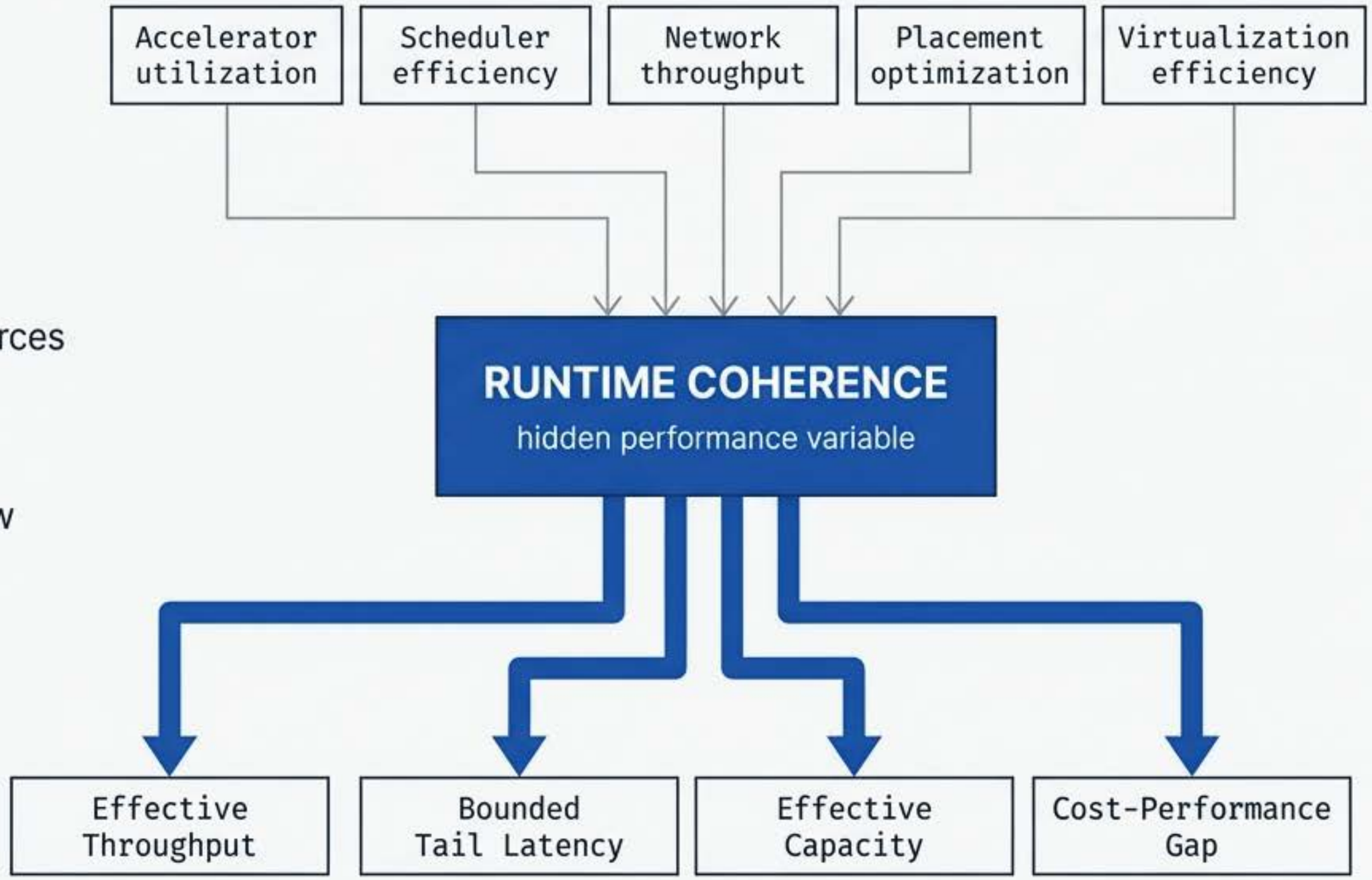
- Without a unified coherence model, local successes sharpen system-wide mismatch.

RUNTIME COHERENCE: THE FIRST-ORDER VARIABLE

The definition of stability.
Coherence is the degree to which hardware behavior, memory paths, topology, and placement remain mutually compatible.

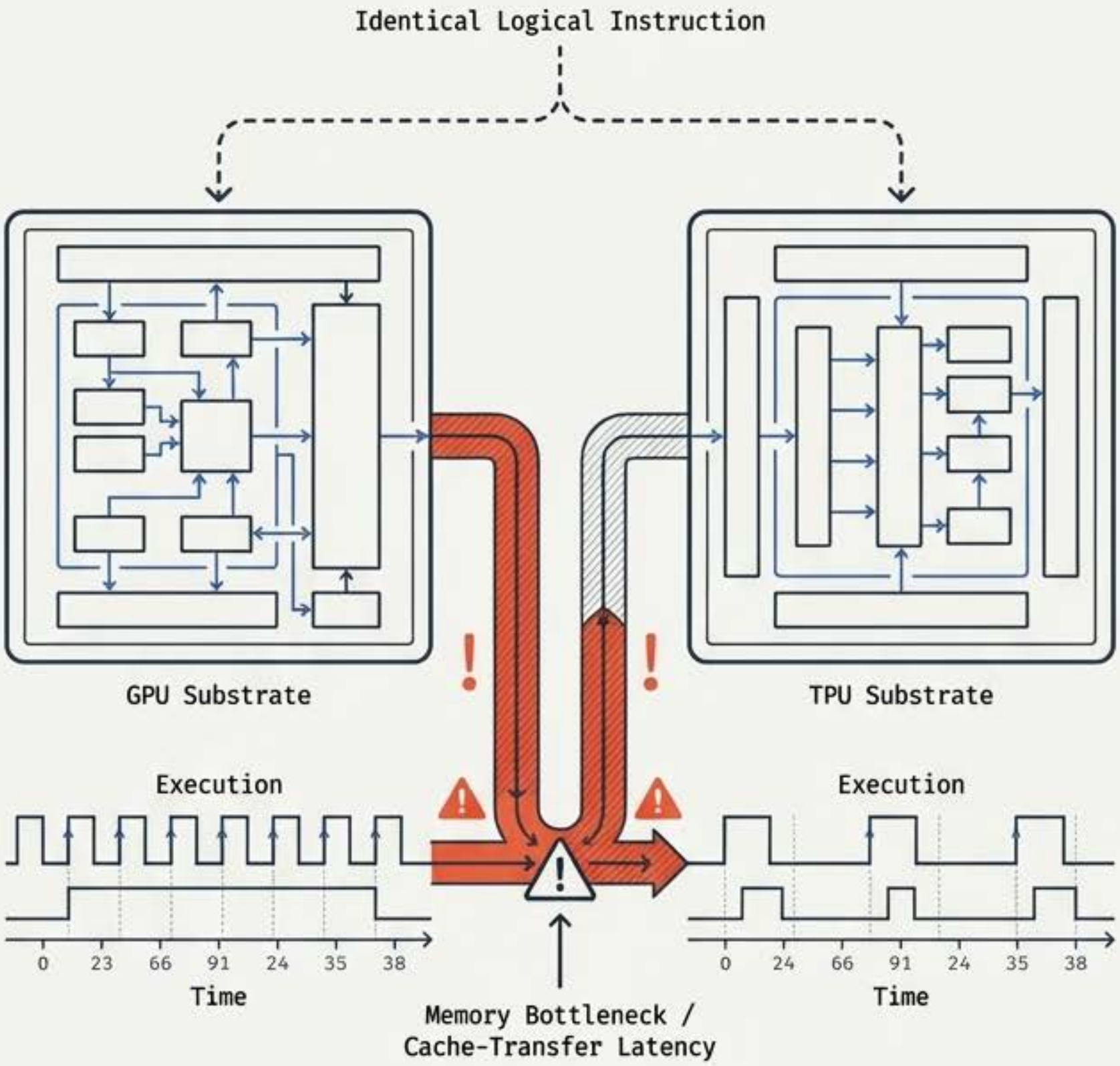
The prerequisite for capacity.
Without coherence, provisioned resources become topologically unreachable.

The core analytical shift.
We must stop measuring quantity (how much compute) and start measuring composition (how paths reshape coordination).



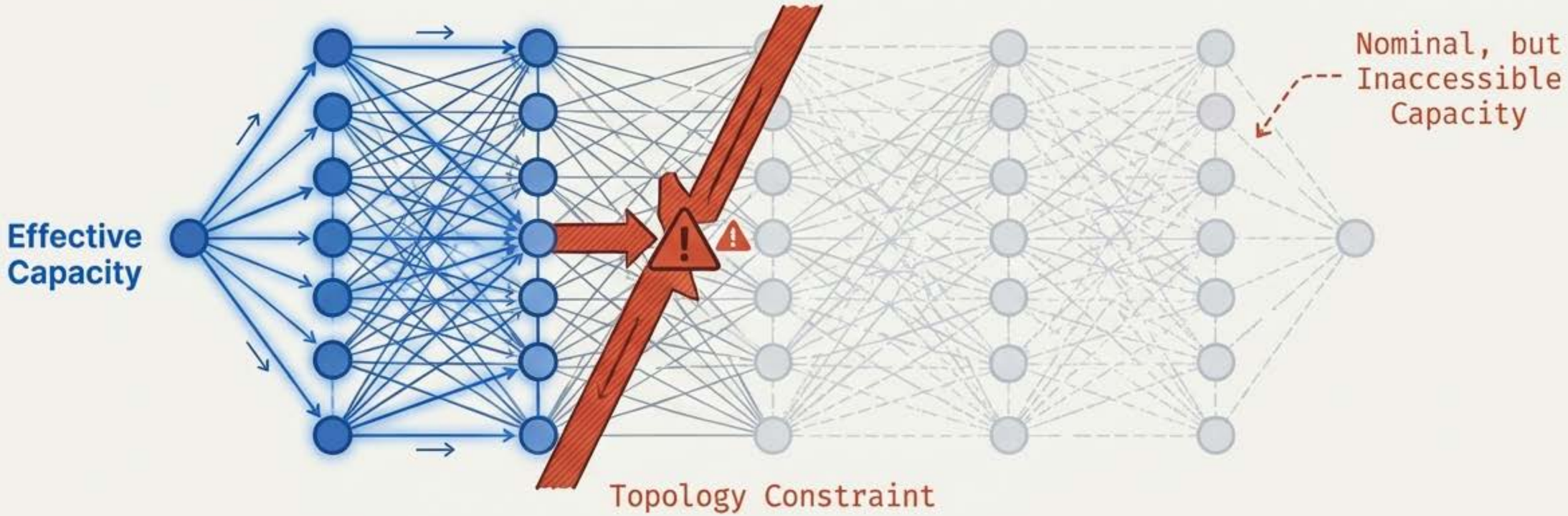
DOMAIN 1: ACCELERATOR CONTROL MISMATCH (AI.07)

- **Homogeneous control breaks heterogeneous hardware.** Applying uniform control logic across non-equivalent execution substrates creates structural fragility.
- **Hardware dictates runtime geometry.** Different accelerators impose strict, distinct constraints on batch formation and synchronization timing.
- **Memory asymmetry bottlenecks the pipeline.** Compute-heavy metrics hide cache-transfer latency when prefill and decode span mismatched hardware.



DOMAIN 2: NETWORK & PLACEMENT COUPLING (AI.11)

Physical capacity is conditional. The physical presence of compute means nothing if topological constraints block coherent serving. **Scale-out** ıaonesoncin. Topologies that handle the geometry. Dıcaggrogatily. Topologies that handle current traffic traffic singes fail under the non-linear stress patterns of targeted scale. **KV-cache continuity dictates placement.** Disaggregated serving forces heavy reliance on interconnects, turning network bounds into decode bounds.

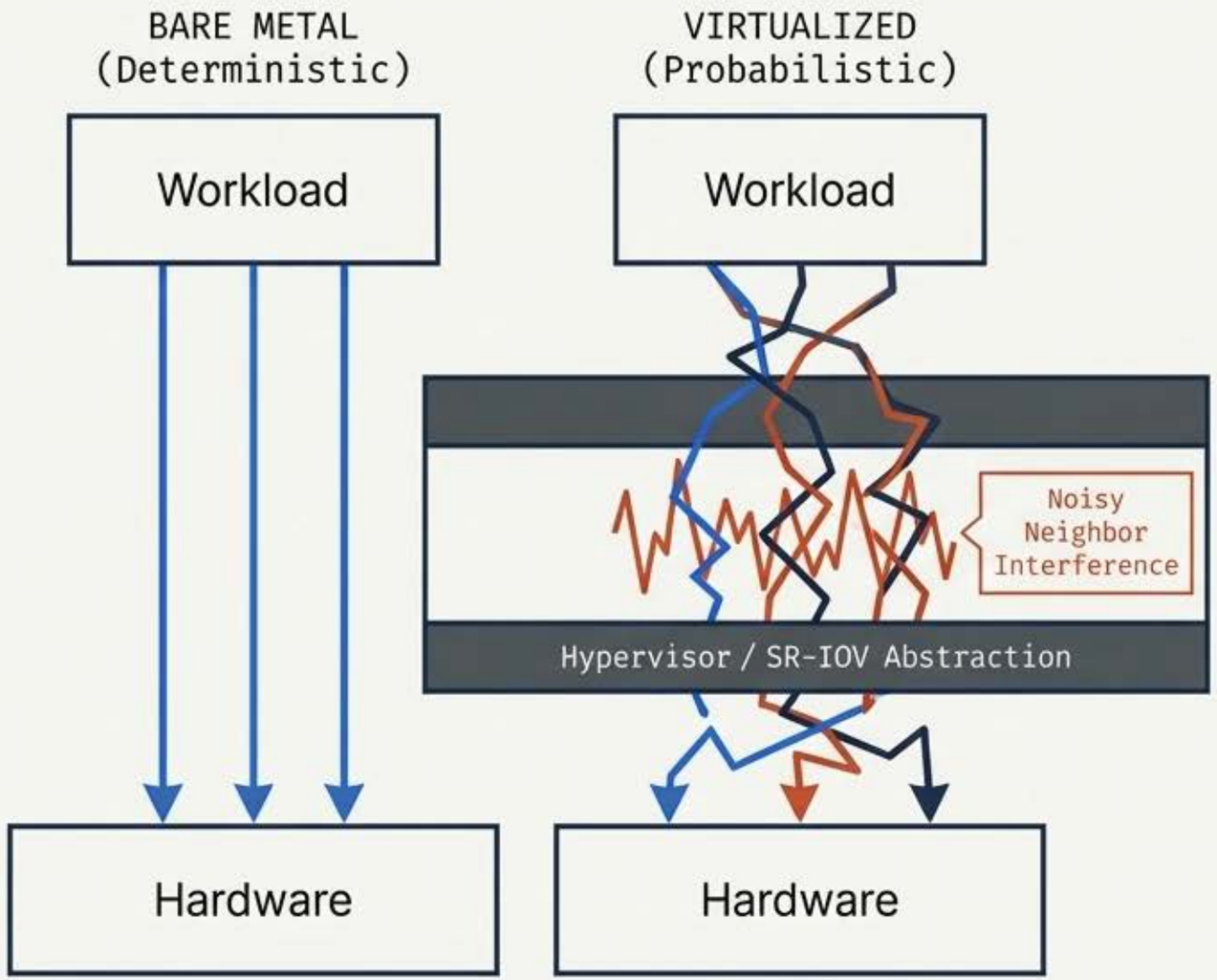


DOMAIN 3: VIRTUALIZATION-INDUCED DISTORTION (AI.14)

Abstraction rewrites physical coupling.
`SR-IOV`, `RDMA passthrough`, and `MIG partitioning` turn physical topology into a stochastic multi-tenant surface.

Virtualization breaks intent-to-execution mapping.
Abstraction layers mediate resource visibility, masking noisy neighbor cascades from workload-local telemetry.

Deterministic shifts to probabilistic.
Strict SLA guarantees become mathematically unachievable under unseen hypervisor contention.



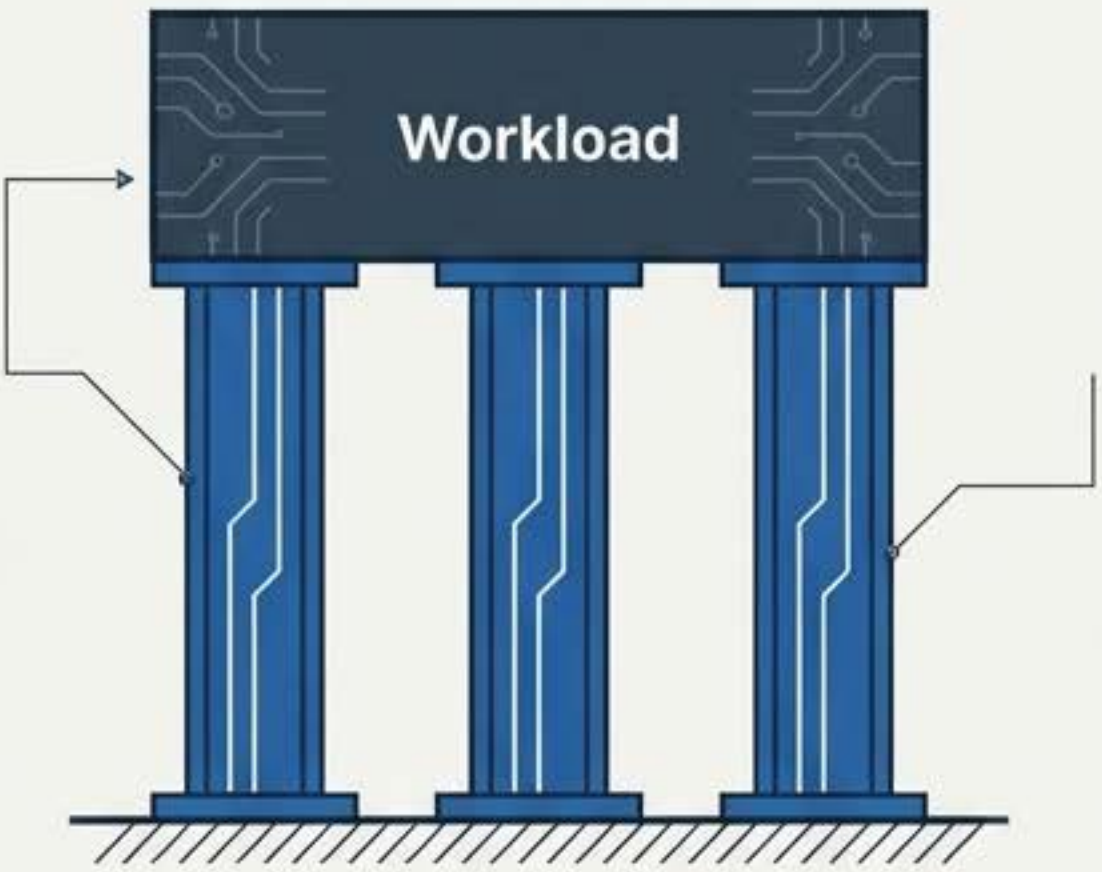
DOMAIN 4: MIGRATION RECONFIGURATION RISK (AI.20)

Code survives, physics change. **Moving workloads** across clouds preserves functional code but destroys the implicit **runtime geometry**.

Coupling patterns are **silently rewritten**. **Bare-metal** to **shared-fabric** transitions radically alter **latency distributions** and **storage coupling patterns**.

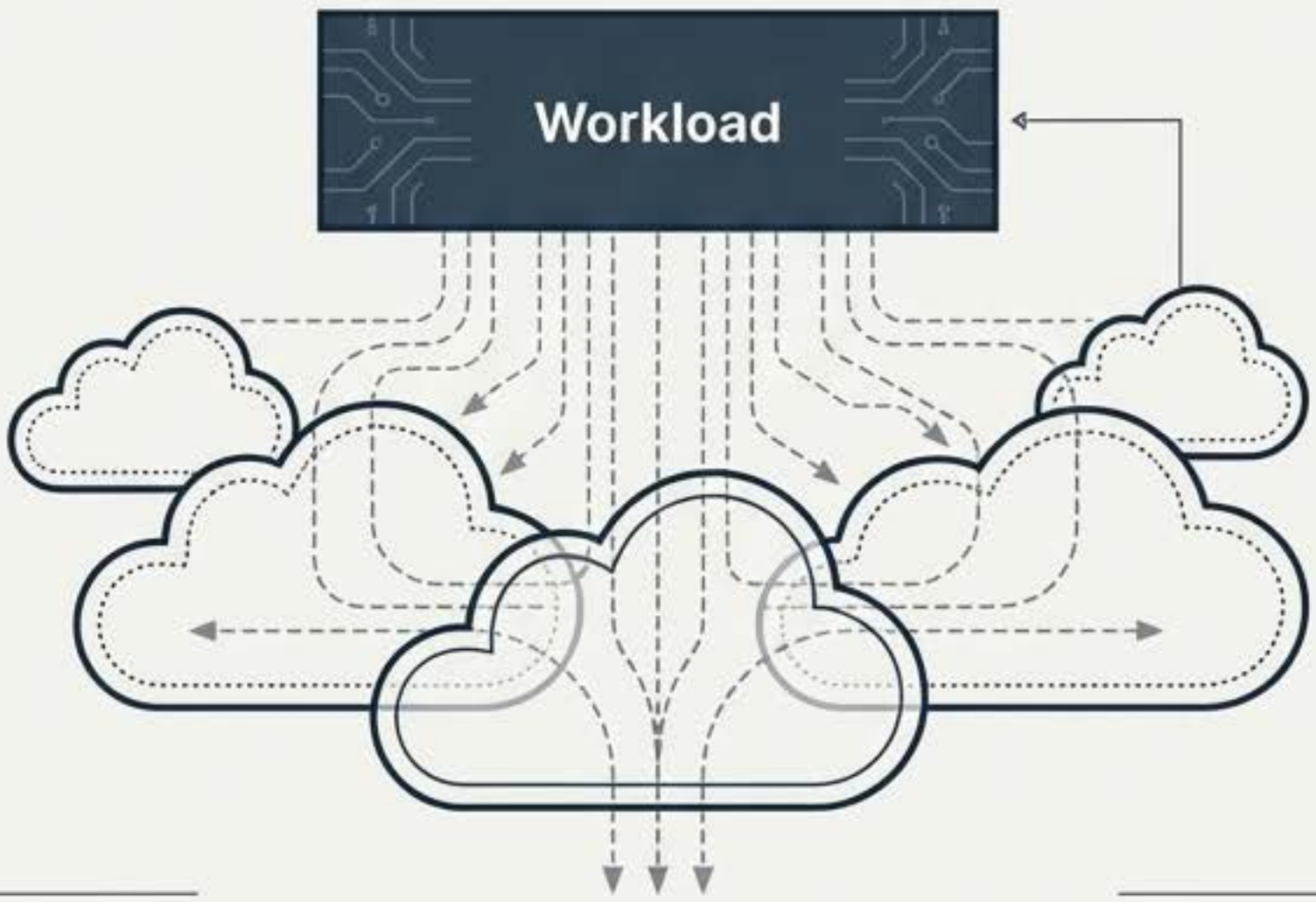
Schedulers become destabilizing. Logic optimized for **dedicated interconnects** becomes instantly hostile in **virtualized environments**.

SOURCE: ON-PREM TOPOLOGY



Structural Divergence:
Physics Reset

TARGET: CLOUD FABRIC



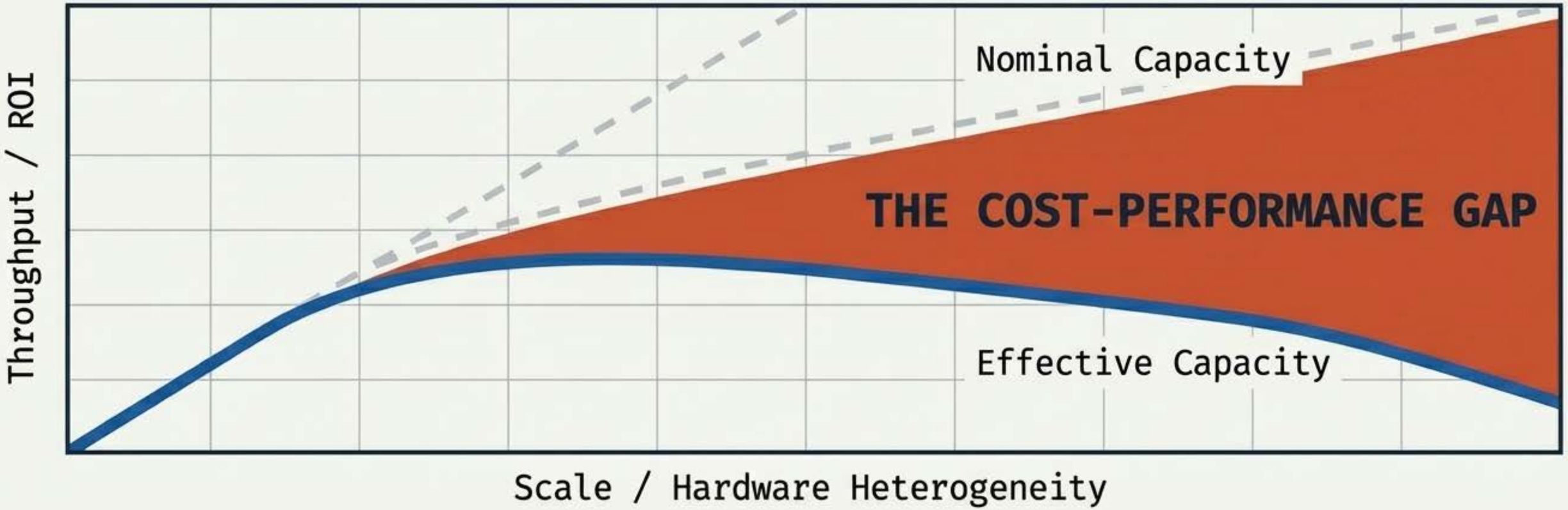
THE TAXONOMY OF STRUCTURAL INSTABILITY

Heterogeneous inference instability is not random. It manifests in five predictable structural modes. Component failure is rare; cross-layer incoherence is the operational default.

Instability Mode	Source Domain	Observable Consequence
Latency Asymmetry Drift	Accelerator Mismatch	Cumulative timing divergence across paths.
Memory-Path Incoherence	Hardware Divergence	Compute visible, but fundamentally memory-locked.
Capacity Inaccessibility	Network Coupling	Provisioned resources topologically unreachable.
Control Distortion	Virtualization	Hypervisor decouples steering intent from execution.
Reconfiguration Risk	Migration Transition	Environmental shift silently rewrites control geometry.

THE COST-PERFORMANCE COLLAPSE

- **Scale exposes the incoherence.** As heterogeneity scales, the gap between `Nominal Throughput` and `Effective Throughput` widens exponentially.
- **You are paying for chaos.** Hardware is fully billed, yet a decreasing fraction converts into service that meets strict `latency/SLA objectives`.
- **Adding hardware accelerates the collapse.** Adding compute without structural diagnostics only weaponizes your `cross-layer asymmetries`.



COHERENCE PRECEDES OPTIMIZATION

Stop tuning incoherent systems. Local optimization will only weaponize your cross-layer asymmetries.

Upgrade your observability. Map the runtime topology, trace `cross-layer coupling`, and measure `effective capacity` over nominal hardware presence.

Master the Heterogeneous Fabric. Organizations that enforce structural coherence extract value from `from diversity`; those that don't, pay for chaos.