

Phantom Scarcity

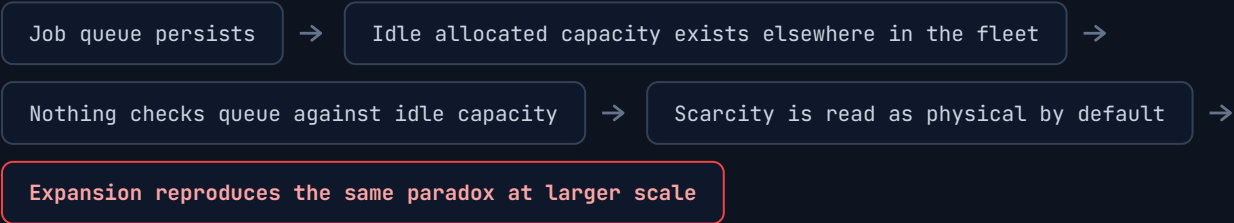
DEFINITION

The condition in which GPU jobs queue despite idle hardware because fragmentation and scheduling make capacity unallocatable.

FRAMEWORK FLOW

01 The Condition A job queue persists and is read as evidence of genuine GPU shortage — the default interpretation of "jobs are waiting for hardware."	02 The Boundary The queue is checked against allocatable-but-idle capacity elsewhere in the fleet — the point where "shortage" and "scheduling failure" become distinguishable, if anyone looks.	03 Failure State Recoverable yield loss — allocation drift, fragmentation, scheduling immaturity — is producing the appearance of scarcity that isn't physically real; the queue and the idle capacity coexist because nothing routes one to the other.	04 Consequence GPUs are purchased to relieve a queue that operational fixes could have cleared; because the root cause was never addressed, the next fleet expansion reproduces the same paradox at larger scale.
---	---	--	--

Failure State: Genuine Shortage Assumed, Never Verified



THE DIAGNOSTIC SIGNAL

SCARCITY OWNERSHIP Ask whether the queue is being read as proof of hardware shortage. It always is — confidently, and by default, since nothing else explains a queue as intuitively.	COMPOSITE OWNERSHIP Ask who checks the queue against allocatable-but-idle capacity before approving expansion. Nobody does, routinely — Phantom Scarcity exists because that check isn't a standing step anywhere.
TRIGGER CONDITION A persistent queue with zero idle allocated capacity elsewhere in the fleet never triggers this — Phantom Scarcity only activates when both conditions coexist.	REMEDIATION SHAPE Explicit Phantom Scarcity detection as a standing check before any expansion decision — classify the queue before buying hardware, not after.

DETECTION STATES

DETECTED Persistent queue coexists with allocatable idle capacity — the shortage is structural, produced by scheduling and allocation failure, not genuine demand.	PARTIAL Some idle capacity coexists with the queue, but not enough to fully explain it — both recoverable waste and real demand growth are likely present.
NOT DETECTED No meaningful idle capacity coexists with the queue — the shortage reflects genuine demand, and expansion is the correct response.	

ARCHITECTURAL RELATIONSHIPS

#89 Effective GPU Yield RELATED MODERATE Phantom Scarcity is the apparent hardware shortage produced when recoverable allocation losses — captured by Effective GPU Yield — are mistaken for insufficient physical capacity.	#90 Capacity Illusion Index RELATED MODERATE Phantom Scarcity is quantified by the Capacity Illusion Index — the percentage delta between provisioned GPU perception and effective usable yield, classified as Artificial, Mixed, or Credible.
#92 Queue–Idle Paradox RELATED MODERATE Co-resident in the GPU Utilization Analyzer — Phantom Scarcity and the Queue-Idle Paradox describe the same allocation failure from different angles.	#93 Fragmentation Tax RELATED MODERATE Co-resident in the GPU Utilization Analyzer — Fragmentation Tax is the direct cause of Phantom Scarcity.
#94 Economic Density Loss RELATED MODERATE Co-resident in the GPU Utilization Analyzer as paired scarcity/loss metrics.	
ANCHOR /gpu-utilization-analyzer/ FRAMEWORK HOME Vendor and hardware-class agnostic — the queue-vs-idle-capacity test applies to any provisioned GPU fleet regardless of scheduler or accelerator vendor.	

"A job queue proves nothing on its own. It's only a shortage once nobody can point to the idle capacity sitting next to it."