

Capacity Illusion Index

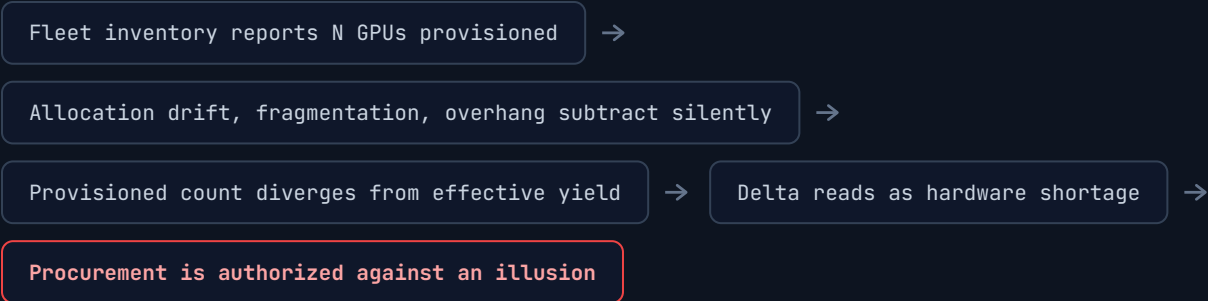
DEFINITION

A measure of how much apparent GPU capacity is illusory — provisioned and visible but not practically allocatable.

FRAMEWORK FLOW

01 The Condition Fleet inventory reports N GPUs provisioned — the number IT and finance both plan against as available capacity.	02 The Boundary Allocation drift, fragmentation, and reservation overhang each silently subtract from that count before any job runs — where "provisioned" and "effectively usable" stop being the same number.	03 Failure State The delta compounds unnoticed — e.g. 512 GPUs provisioned, 171 effective, a 66% Capacity Illusion Index — read as scarcity rather than as an artifact of allocation practice.	04 Consequence Procurement decisions are made against the provisioned number, not the effective one — capital is committed to scarcity that was never physically real.
---	--	---	---

Failure State: Scarcity That Isn't Physical



THE DIAGNOSTIC SIGNAL

PROVISIONING OWNERSHIP Ask whether the provisioned GPU count is treated as the true capacity ceiling. It always is — confidently, and it's the wrong ceiling.	COMPOSITE OWNERSHIP Ask who owns the gap between provisioned and effective capacity. Nobody does by design — the illusion isn't tracked as its own number in most fleet inventories.
TRIGGER CONDITION The index activates whenever allocation drift, fragmentation, or reservation overhang exist at all; a fleet with zero operational waste has an index of zero and never triggers it.	REMEDIATION SHAPE An explicit Capacity Illusion Index as a standing fleet metric, scarcity classification, recoverable-capacity quantification before any expansion decision — architectural, not procedural.

SCARCITY CLASSIFICATION

ARTIFICIAL SCARCITY The queue is fully explained by recoverable operational waste — allocation drift, fragmentation, or overhang. No genuine capacity shortage exists.	MIXED SCARCITY Both real demand growth and recoverable operational waste are present — expansion may be partly justified, but not at the full scale the queue implies.
CREDIBLE SCARCITY Provisioning, allocation, and demand are well-matched — the queue reflects genuine demand, and expansion is the correct response.	

ARCHITECTURAL RELATIONSHIPS

#89 Effective GPU Yield RELATED MODERATE Co-resident in the GPU Utilization Analyzer — Effective GPU Yield and Capacity Illusion Index are paired output metrics.	#91 Phantom Scarcity RELATED MODERATE Co-resident in the GPU Utilization Analyzer — Capacity Illusion Index quantifies Phantom Scarcity.
#92 Queue–Idle Paradox RELATED MODERATE Co-resident in the GPU Utilization Analyzer as paired capacity-illusion metrics.	#93 Fragmentation Tax RELATED MODERATE Co-resident in the GPU Utilization Analyzer — fragmentation is the mechanism behind the capacity illusion.
#94 Economic Density Loss RELATED MODERATE Co-resident in the GPU Utilization Analyzer as paired capacity/economic metrics.	
ANCHOR /gpu-utilization-analyzer/ FRAMEWORK HOME Vendor and hardware-class agnostic — the provisioned-vs-effective gap applies to any GPU fleet regardless of scheduler or accelerator vendor.	

"Every provisioned GPU is real. The capacity it represents on paper is not — and the gap between those two facts is the shortage everyone is planning around."