

Effective GPU Yield

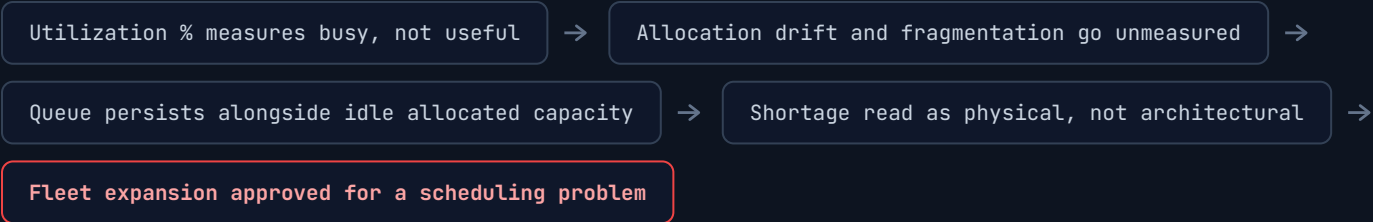
DEFINITION

The fraction of purchased GPU capacity that produces useful work after scheduling, fragmentation, and idle losses.

FRAMEWORK FLOW

#1 The Condition GPU utilization % is read as the health signal — averaged across allocated cards, blind to whether allocated capacity is doing useful work.	#2 The Boundary Allocation drift, fragmentation, and scheduling immaturity diverge from the utilization number — the point where "GPUs are busy" and "GPUs are producing useful work" separate.	#3 Failure State 62% average utilization plus a persistent job queue reads as a hardware shortage; Effective GPU Yield is actually far lower — the gap is operational, not physical.	#4 Consequence Fleet expansion is approved to relieve a queue that recoverable operational waste could have cleared — capital committed to a scheduling problem, not a hardware one.
---	--	---	---

Failure State: Yield Collapse Misread as Hardware Shortage



THE DIAGNOSTIC SIGNAL

YIELD OWNERSHIP Ask any allocated GPU whether it's busy. The answer is yes — confidently, and correctly. Utilization percent never lies about activity.	COMPOSITE OWNERSHIP Ask who owns the gap between busy and useful. The honest answer is nobody — dashboards report utilization because it's easy to measure, not because it answers the yield question.
TRIGGER CONDITION A persistent job queue coexisting with allocated-but-idle GPUs is the activation condition. A cluster with genuine 1:1 demand-to-capacity never triggers the gap at all.	REMEDIATION SHAPE Scheduler maturity upgrade, workload class isolation, reservation right-sizing, an explicit yield metric standing alongside utilization — architectural, not a bigger fleet.

THE THREE YIELD DISCOUNTS

ALLOCATION-UTILIZATION GAP The gap between capacity allocated to workloads and the compute those workloads actually consume — the largest and least visible of the three discounts.	FRAGMENTATION TAX Stranded fraction of each GPU from whole-card allocation to sub-card workloads that require only a fraction of the card's compute or memory capacity.
RESERVATION OVERHANG Committed GPU capacity that outlives the workload that justified it — a compounding cost obligation rather than a one-time loss.	

ARCHITECTURAL RELATIONSHIPS

#90 Capacity Illusion Index 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 — low yield and Phantom Scarcity describe the same underlying allocation failure.
#92 Queue–Idle Paradox RELATED MODERATE Co-resident in the GPU Utilization Analyzer — yield loss and the Queue–Idle Paradox are the same symptom measured two ways.	#93 Fragmentation Tax RELATED MODERATE Co-resident in the GPU Utilization Analyzer — Fragmentation Tax is the primary cause of reduced Effective GPU Yield.
#94 Economic Density Loss RELATED MODERATE Co-resident in the GPU Utilization Analyzer — Economic Density Loss is the dollar expression of reduced yield.	
ANCHOR /gpu-utilization-analyzer/ FRAMEWORK HOME Vendor and hardware-class agnostic — the yield/discount model applies to any provisioned GPU fleet regardless of scheduler or accelerator vendor.	

"Every allocated GPU can correctly report that it is busy. Nobody can explain why the queue never clears — not because anyone neglected the question, but because utilization was never built to answer it."