

The GPU Market is Broken

A Workload-Specific Approach to Performance Transparency

OCP as a Path to Standardization

Ash Chary, Andrew Nelson, PhD | FLOPS | team@flopsindex.com

PROPOSED FRAMEWORK

1

Canonical Workloads

Standardized benchmarks with production conditions

2

Infrastructure Metadata

Structured context explaining performance

3

Results Aggregation

Anonymized data with k-anonymity protection

OCP PROVIDES: TRUST, NEUTRALITY, GOVERNANCE

Proposed Pilot:

Framework Components: workload suite, metadata schema, operator-submitted results

OCP: governance, neutrality, operator coordination

90 days | 3 operators | 2 workloads

Deliverable: Initial workload-specific performance benchmarks with OCP validation

We are seeking collaborators across operators and the OCP community

THE PROBLEM (\$15B WASTED ANNUALLY)

Benchmarks Don't Reflect Reality

MLPerf: 8%

Variation

Reality: 35%

cost variation

Observed in Production

60K+ GPUs

5 Operators

23% Mismatch

40-60% Utilization

Derived Framework

Workload-dependent performance

Multi-system bottlenecks

Benchmark mismatch

FUTURE-PROOF BY DESIGN

Hardware Evolves!

GPUS → TPUs → ASICs → Custom Silicon

Consumables-Based Model

Power | Utilization | Throughput

Comparable Across Architectures

Vendor-neutral performance evaluation



29-30 April, 2026
Barcelona, Spain

Leading the Future of AI.