



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

The Hyperscale Uncertainty Principle

Debugging Tail Latency in a Trillion-Object System

Yashraj Kakkad
Software Engineer, Google Photos





KubeCon



CloudNativeCon

Europe 2026

Exabytes of Data Trillions of Objects



KubeCon



CloudNativeCon

Europe 2026



Relational Metadata ↔ Object Storage



Designing for Balance



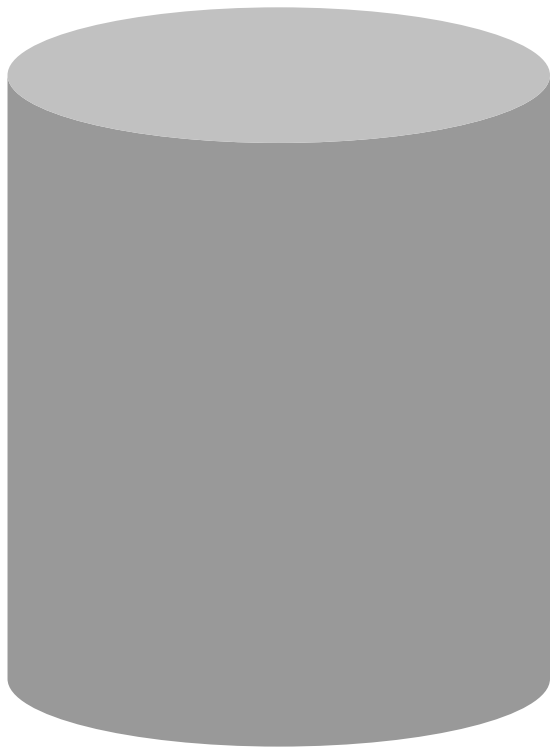
KubeCon



CloudNativeCon

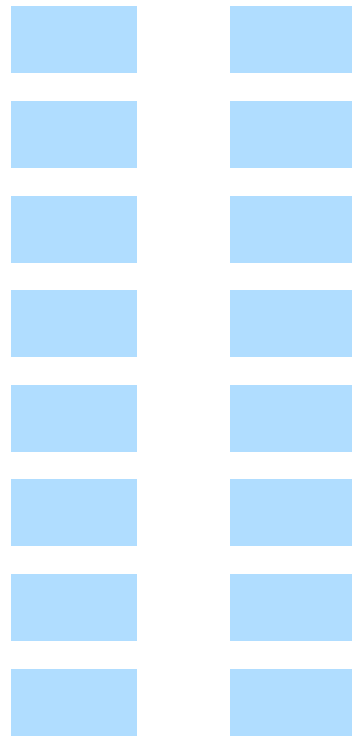
Europe 2026

Exabyte Namespace



Identity-Based Logical Split

16 Balanced Partitions



Reality Check



KubeCon



CloudNativeCon

Europe 2026

16 "Identical" Partitions

Strict 3-Day SLA



Hypothesis: It's just Data Skew



KubeCon



CloudNativeCon

Europe 2026

“The slow partitions must have more data”

Identifying the Ghost



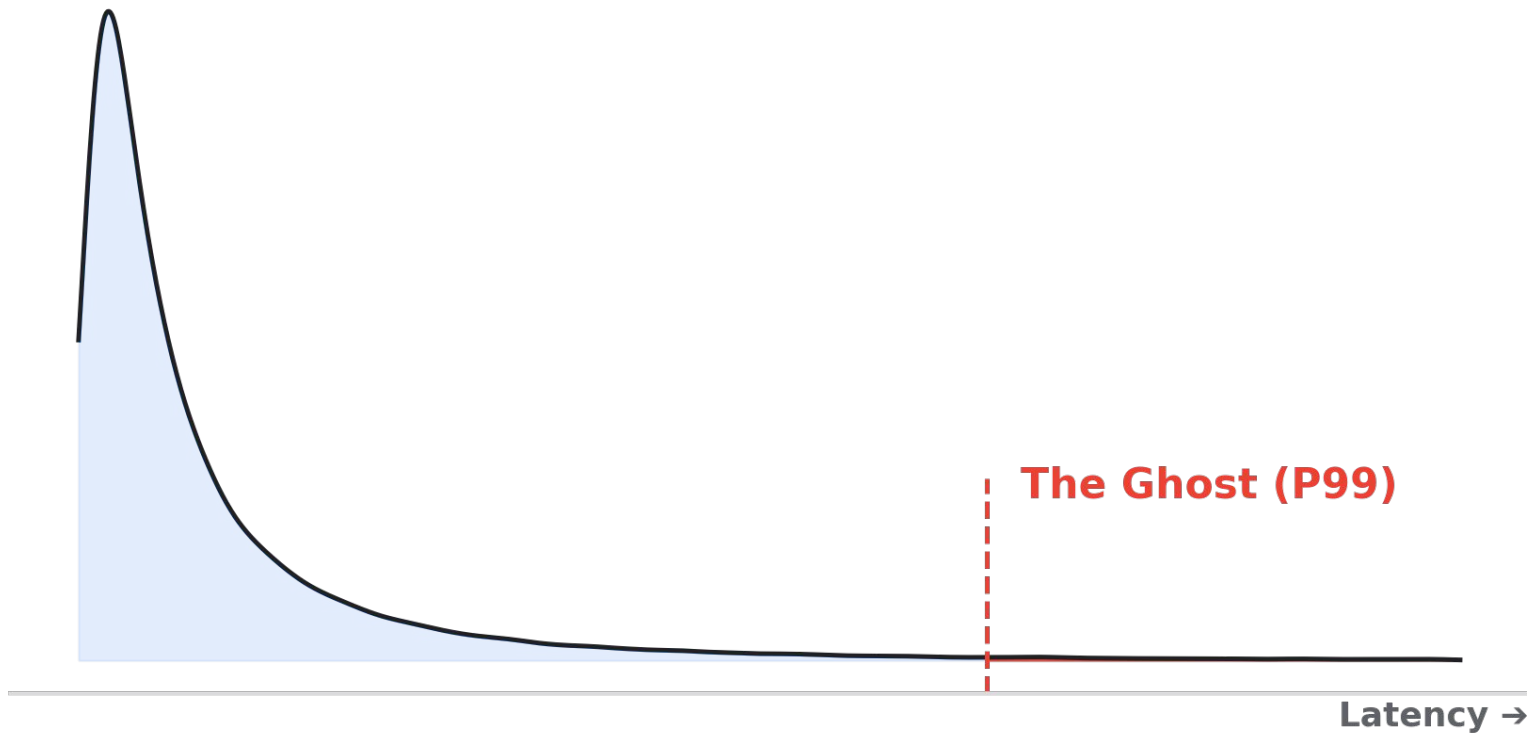
KubeCon



CloudNativeCon

Europe 2026

Volume



The Math vs. The Metal

The Ideal (The Math)

$$\text{Ideal Shards} = 3N \ln N$$

To hide the P99 tail across N workers

The Reality (The Gap)

REQUESTED:

450,000 Shards

RECEIVED:

~30,000 Shards

Result: Massive, unsplittable blocks

Request Hedging - Self Inflicted DDoS?

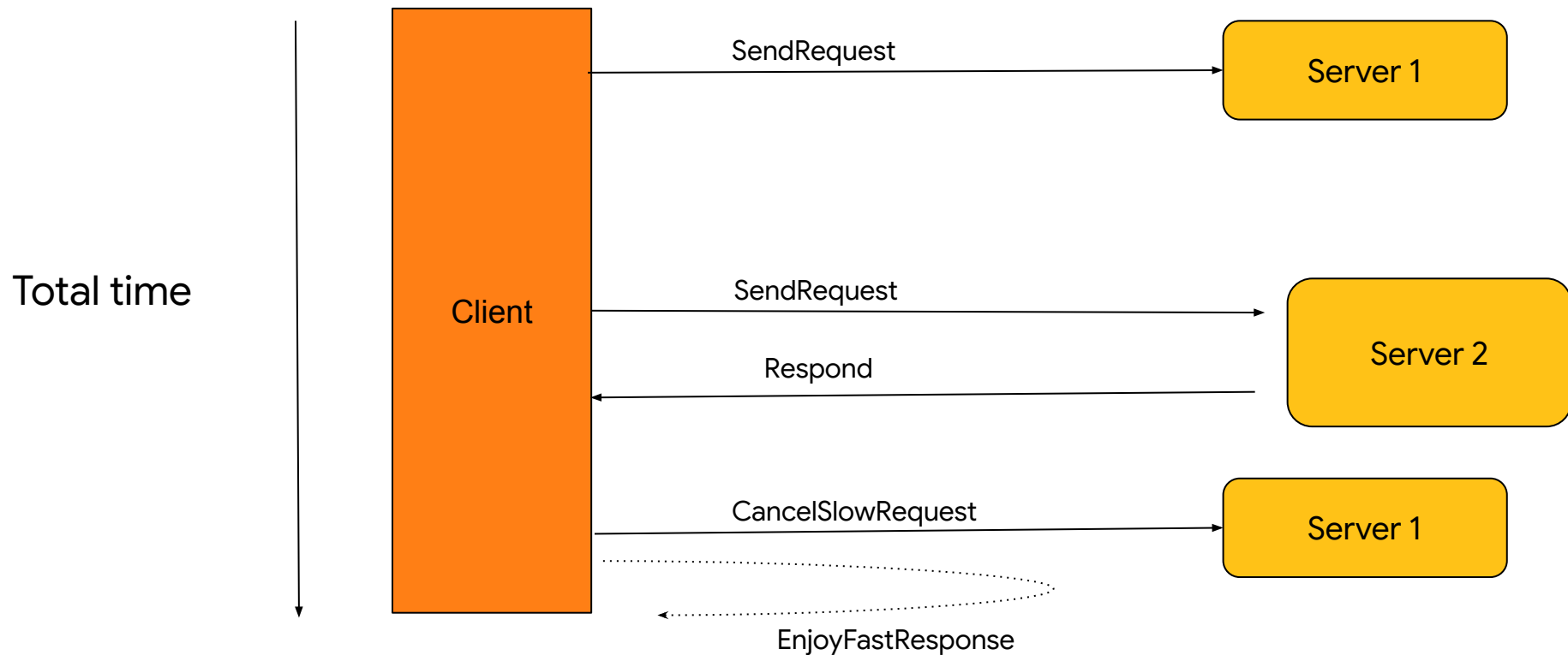


KubeCon



CloudNativeCon

Europe 2026



Logical Boundaries are Illusions

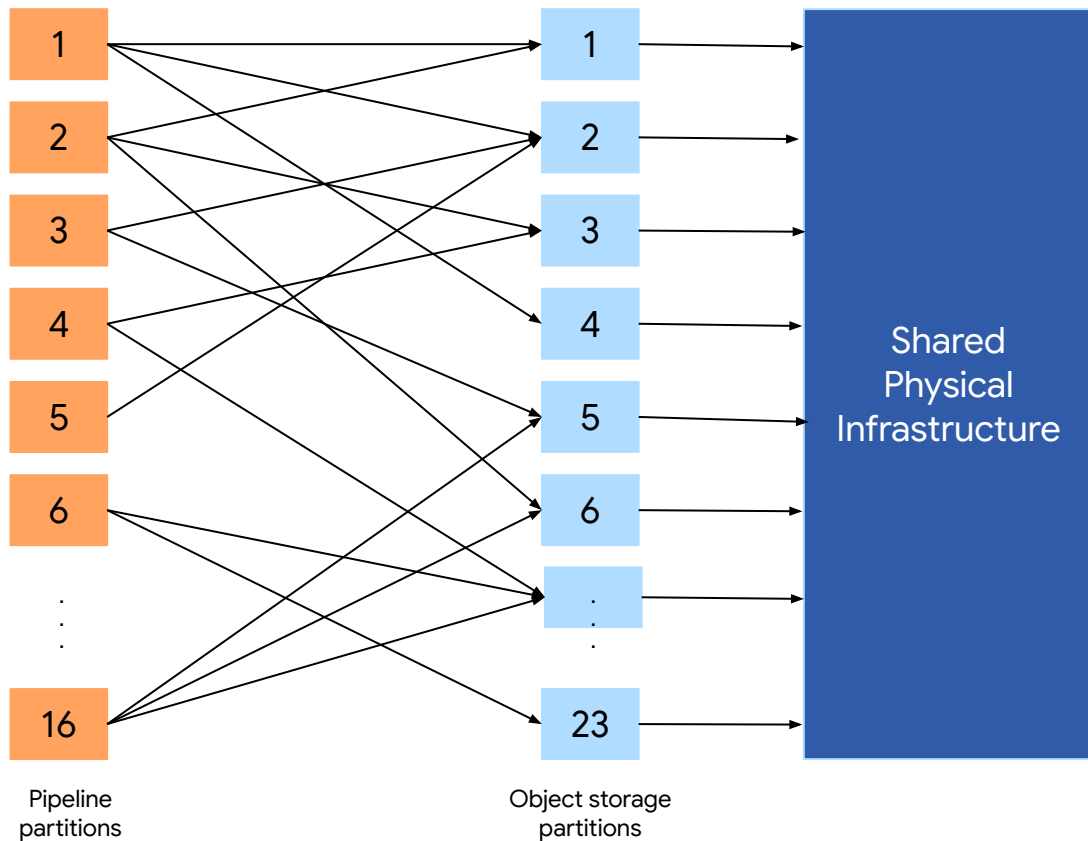


KubeCon



CloudNativeCon

Europe 2026



Solution: Liquid Sharding



KubeCon

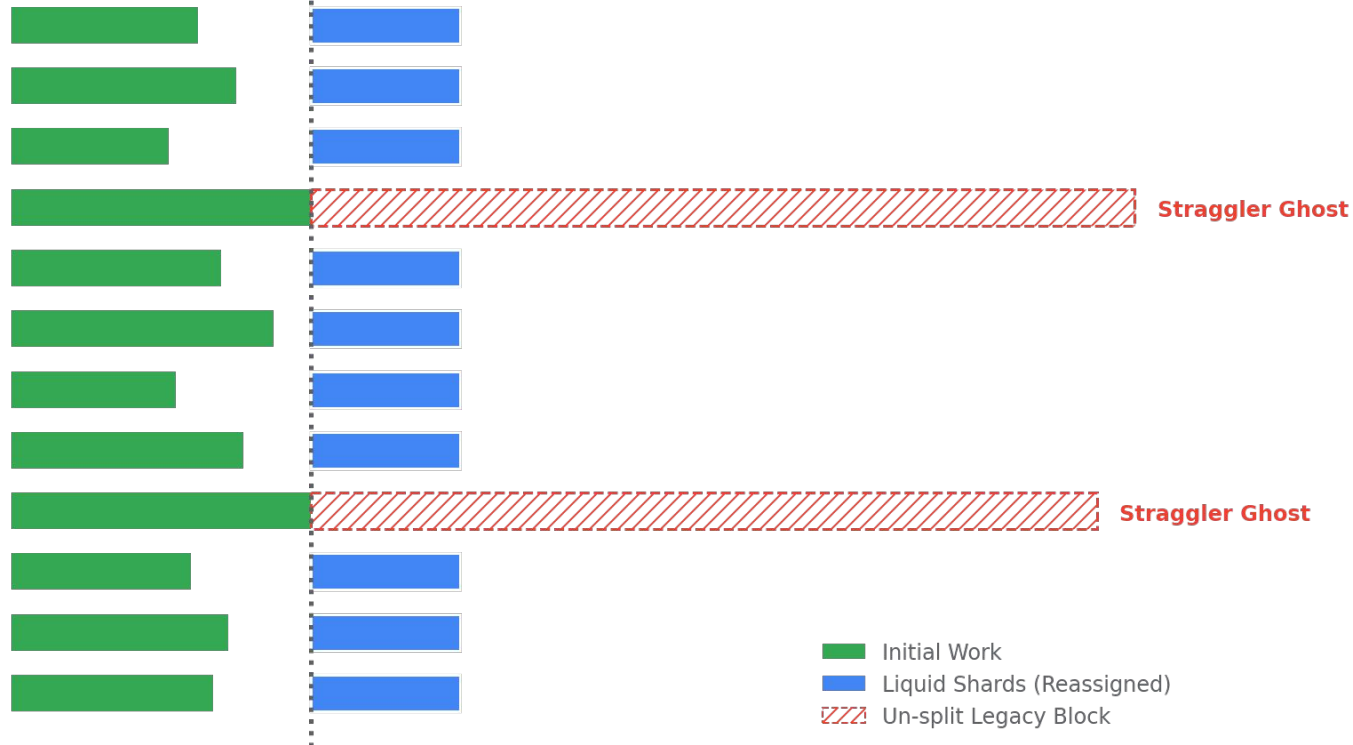


CloudNativeCon

Europe 2026

Worker Pool

Dynamic Work Rebalancing Triggered



Execution Time →

The Expectation

"Best-Effort" Compute = Massive Savings

The Exabyte Reality

>70% Waste Ratio
O(XXX) TiB RAM
O(XXXk) GCUs
50+ Preemptions/Shard

Hard Engineering Truths (Takeaways)



KubeCon



CloudNativeCon

Europe 2026

Architect for the P99 Tail
Respect the Bare Metal
Beware the ROI of Preemption

Yashraj Kakkad
Software Engineer, Google



@KakkadYashraj



in/yashraj

Slides, Math & References



Acknowledgements

Thank you to Ben Appleton for his insights on the math of stragglers, and the rest of the Photos Storage Infrastructure team.