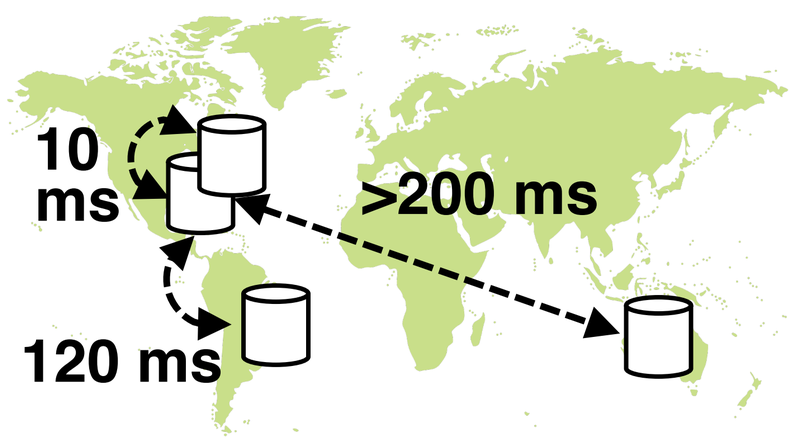
Towards Enabling All Feasible Latency–Cost Tradeoffs in Geo-Distributed Storage

Muhammed Uluyol, Anthony Huang, Ayush Goel,
Mosharaf Chowdhury, and Harsha V. Madhyastha

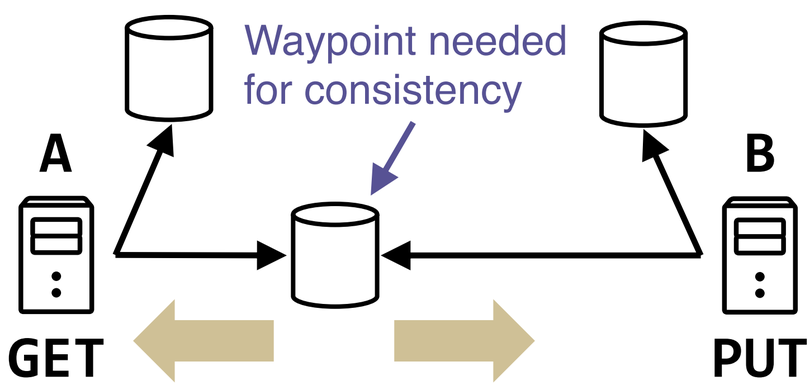
Goal: Best latency–cost trade-off

Setting: Non-uniform RTTs

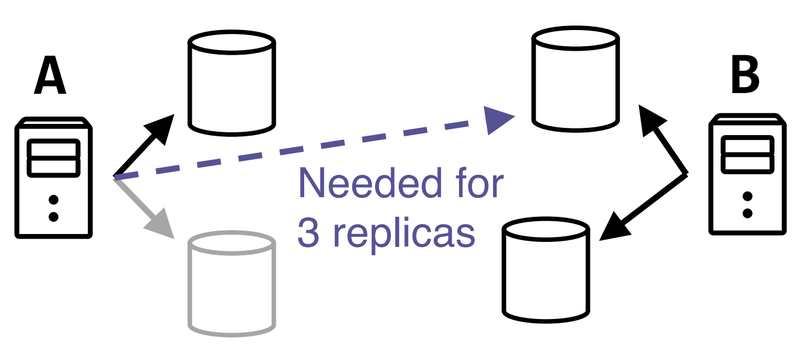


Minimize Latency ≠ Minimize #RTTs

Read–write latency trade-off



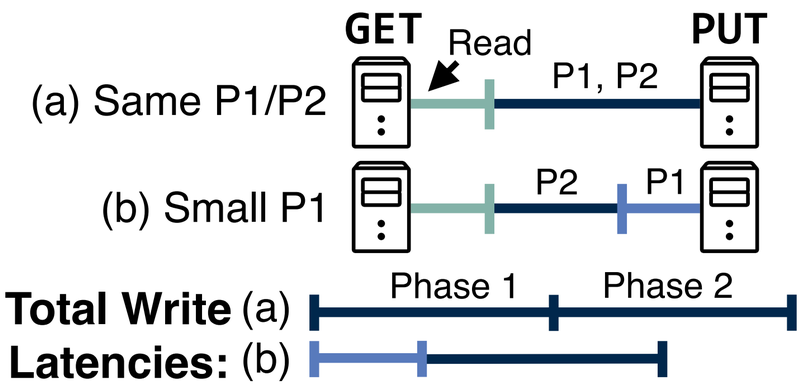
Read latency–cost trade-off



- Handle F failures → quorum size ≥ F+1
- More client data centers → more data sites needed for low read latency

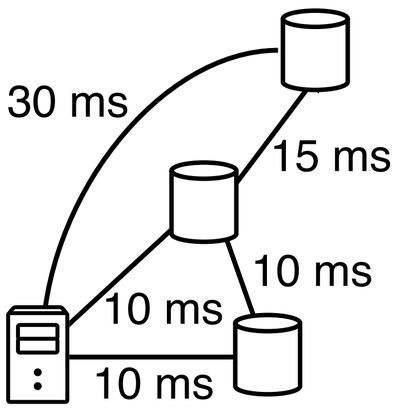
Low Latency Reads & Writes

Small Phase 1 Quorums



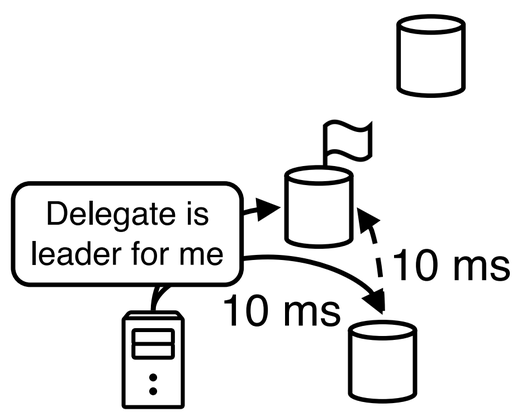
- Paxos writes are 2 phases (P1 & P2)
- Waypoint needed only across P1 & P2 → read quorums in P1 for faster writes

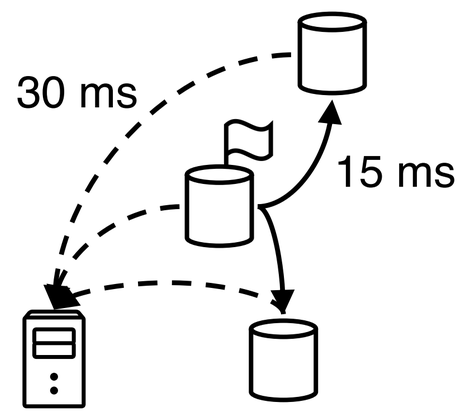
Inter-DC Latencies



Delegating Phase 2

Quorum sizes: **P1=2, P2=3**
Overall 4 replicas

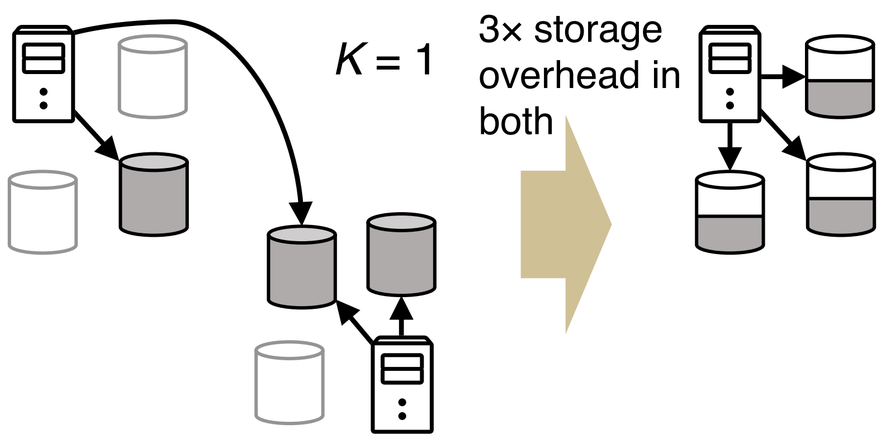




Avoid trip back to client → faster writes. Example: 65 ms (Pando) versus 60 ms (ideal)

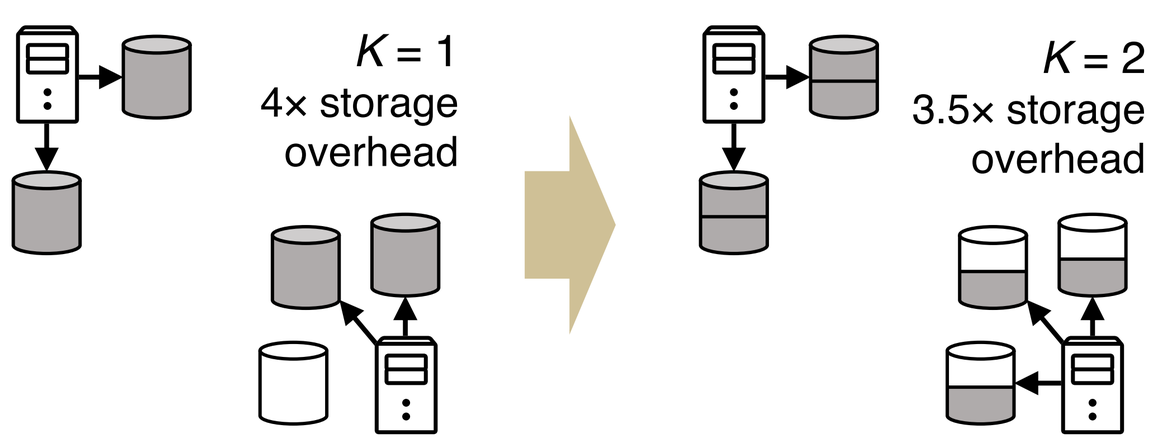
Improving Cost Efficiency

Lowering Read Latency w/ Erasure Coding



K base splits → more data sites close by. Quorum size ≥ F+K

Selectively Co-locating Data Splits

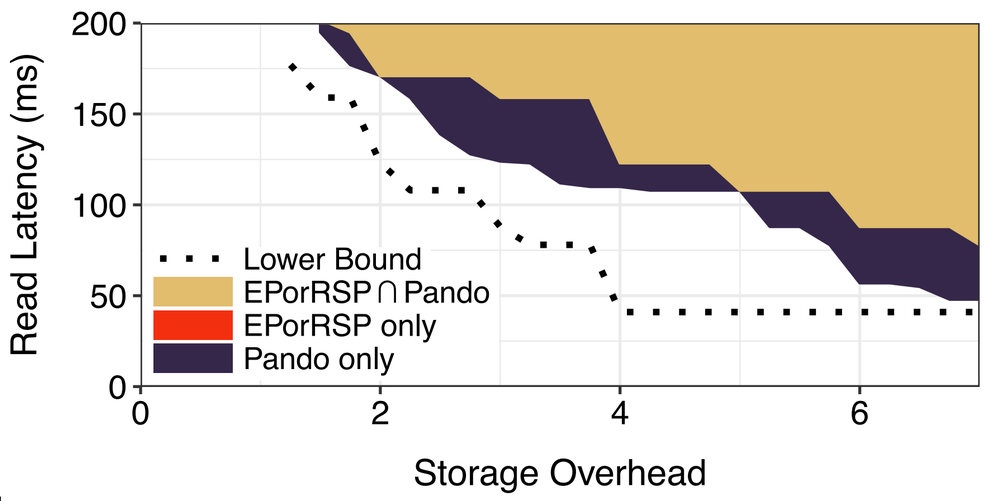


- Few nearby DCs → increase cost only in that region
- Account for multiple splits in same failure domain

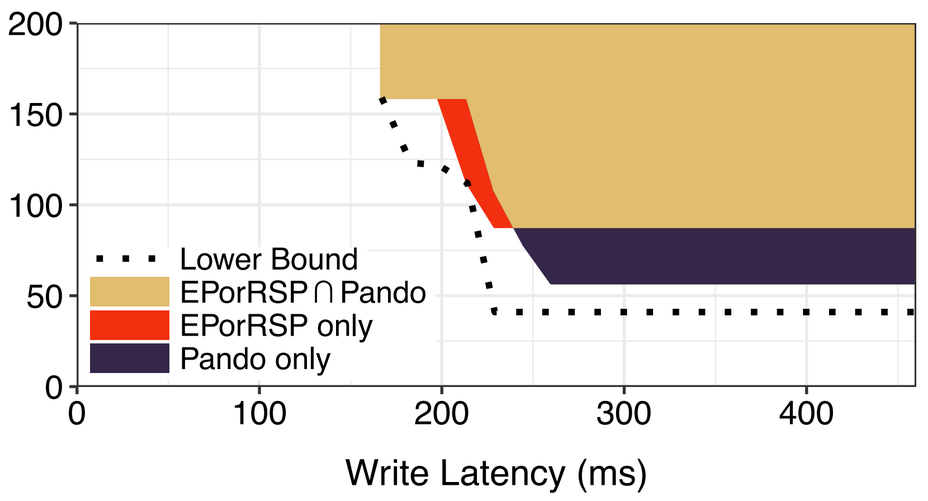
Preliminary Results

Setup

- Clients in 5 data centers
- Write latency ≤ 300 ms
- Storage overhead ≤ 6×
- Chosen data sites minimize read latency across clients
- Compare to one-round EPaxos (EP) and erasure-coded RS-Paxos (RSP)



Better read–storage trade-off, up to 30% lower read latency



10% higher min write latency, 30% lower read latency