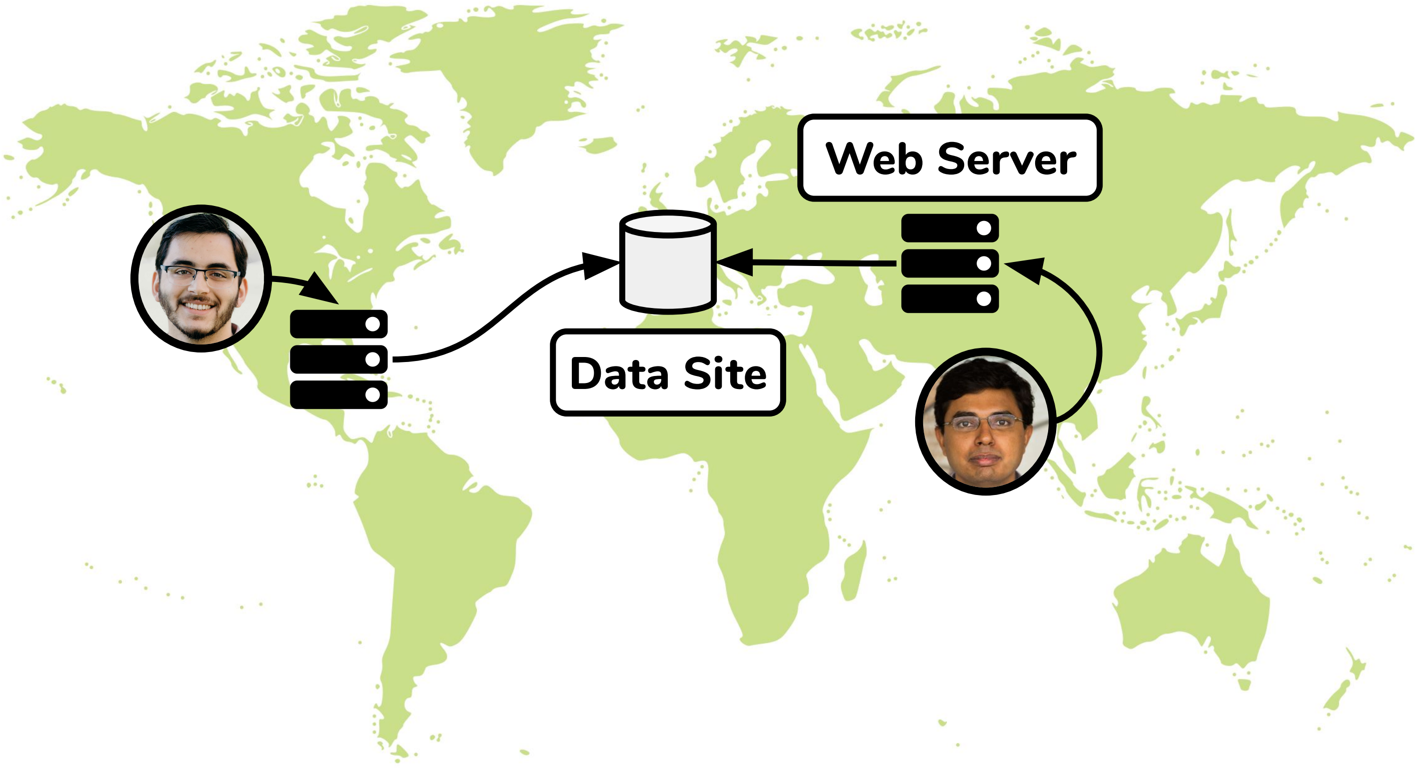
Near-Optimal Latency Versus Cost Tradeoffs in Geo-Distributed Storage

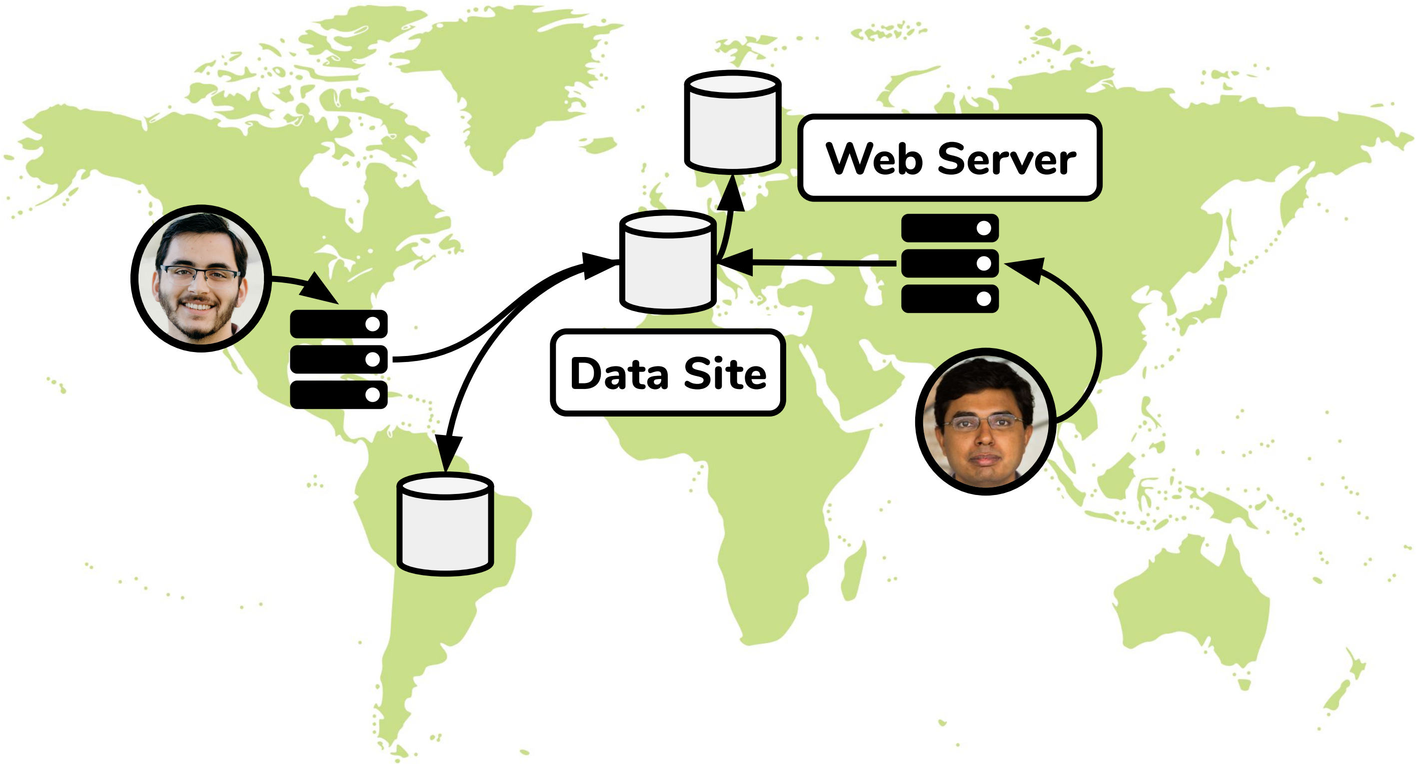
Muhammed Uluyol, Anthony Huang, Ayush Goel,
Mosharaf Chowdhury, Harsha V. Madhyastha

University of Michigan

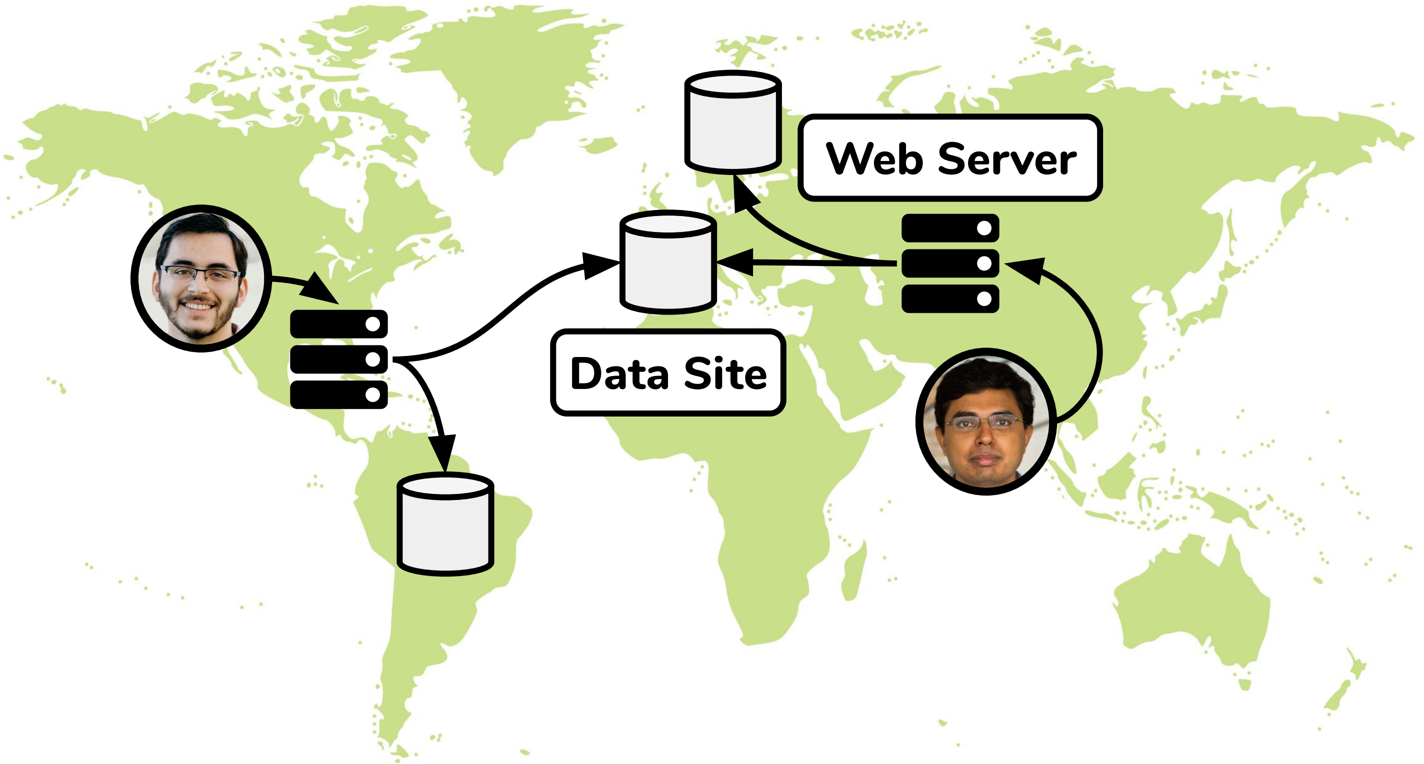
Distribute Web Servers for Interactive Latency



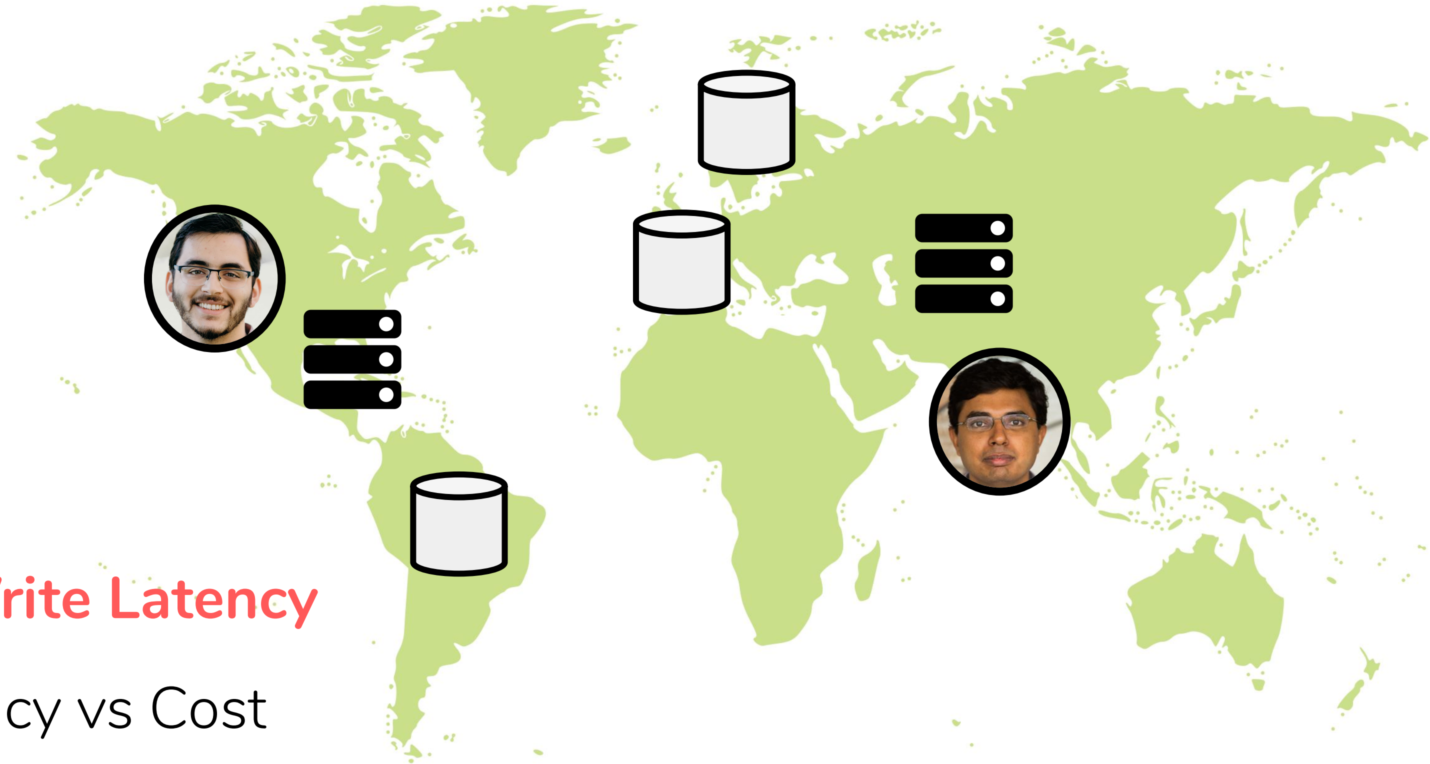
Distribute Data for Availability



Distribute Data for Availability and Latency

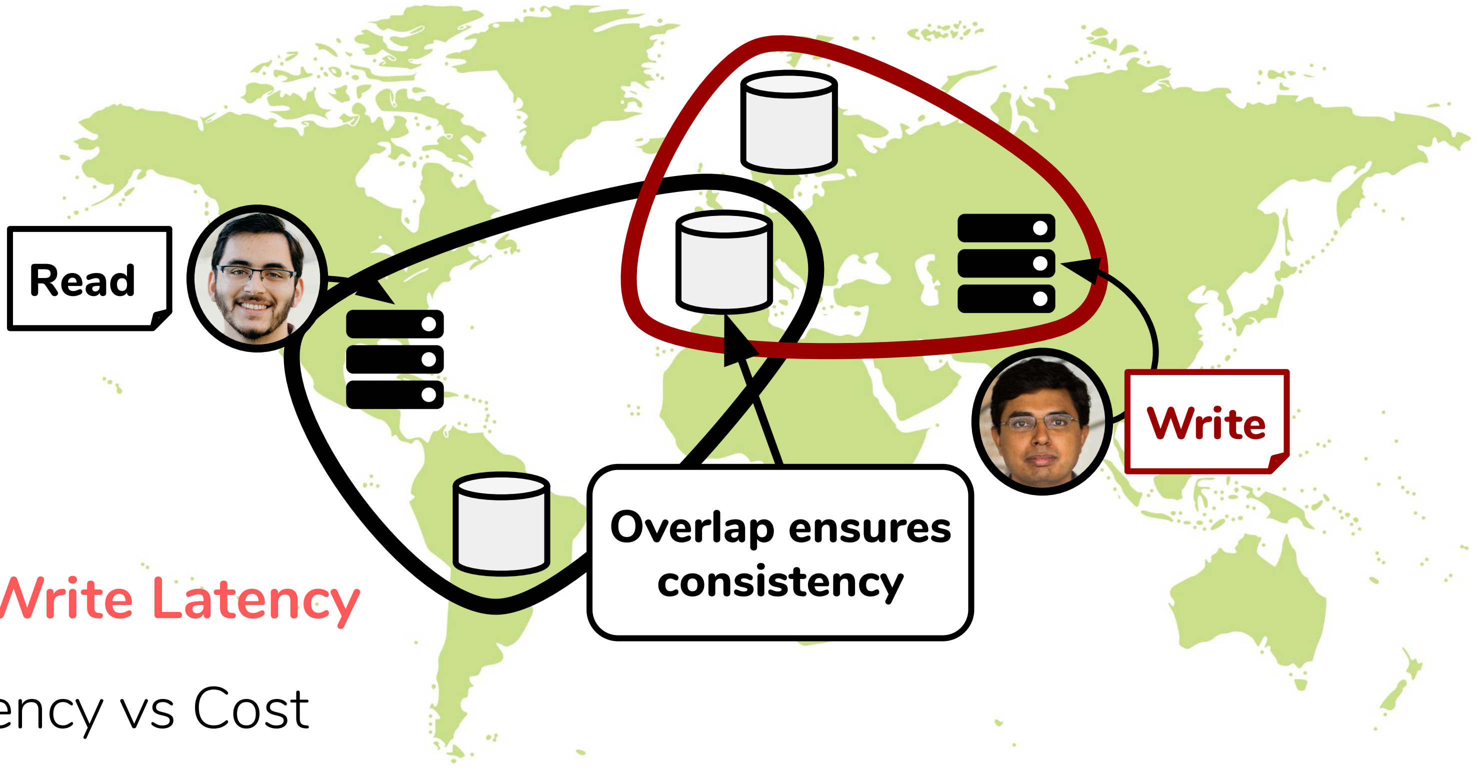


Linearizability Imposes Unavoidable Trade-offs



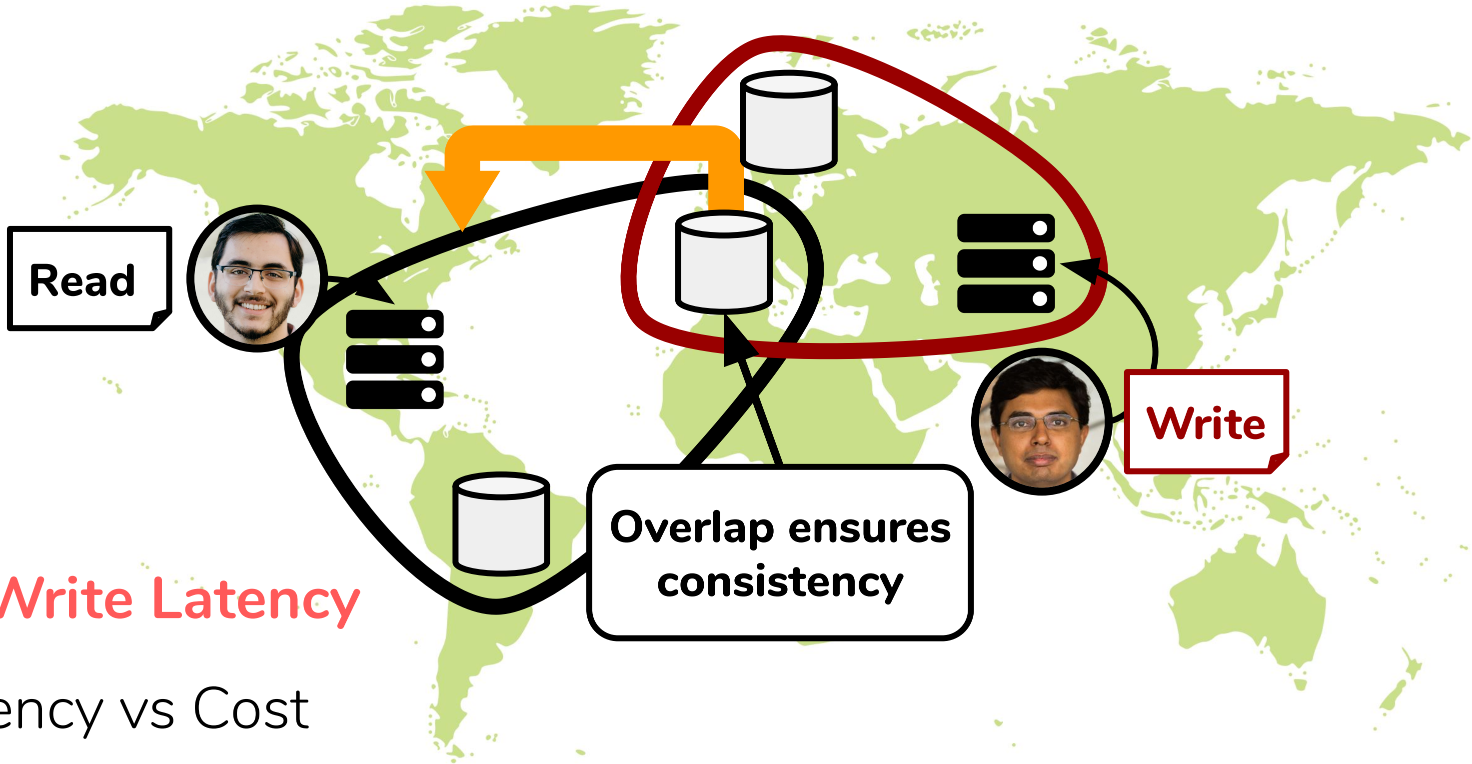
- Read vs Write Latency
- Read Latency vs Cost

Linearizability Imposes Unavoidable Trade-offs



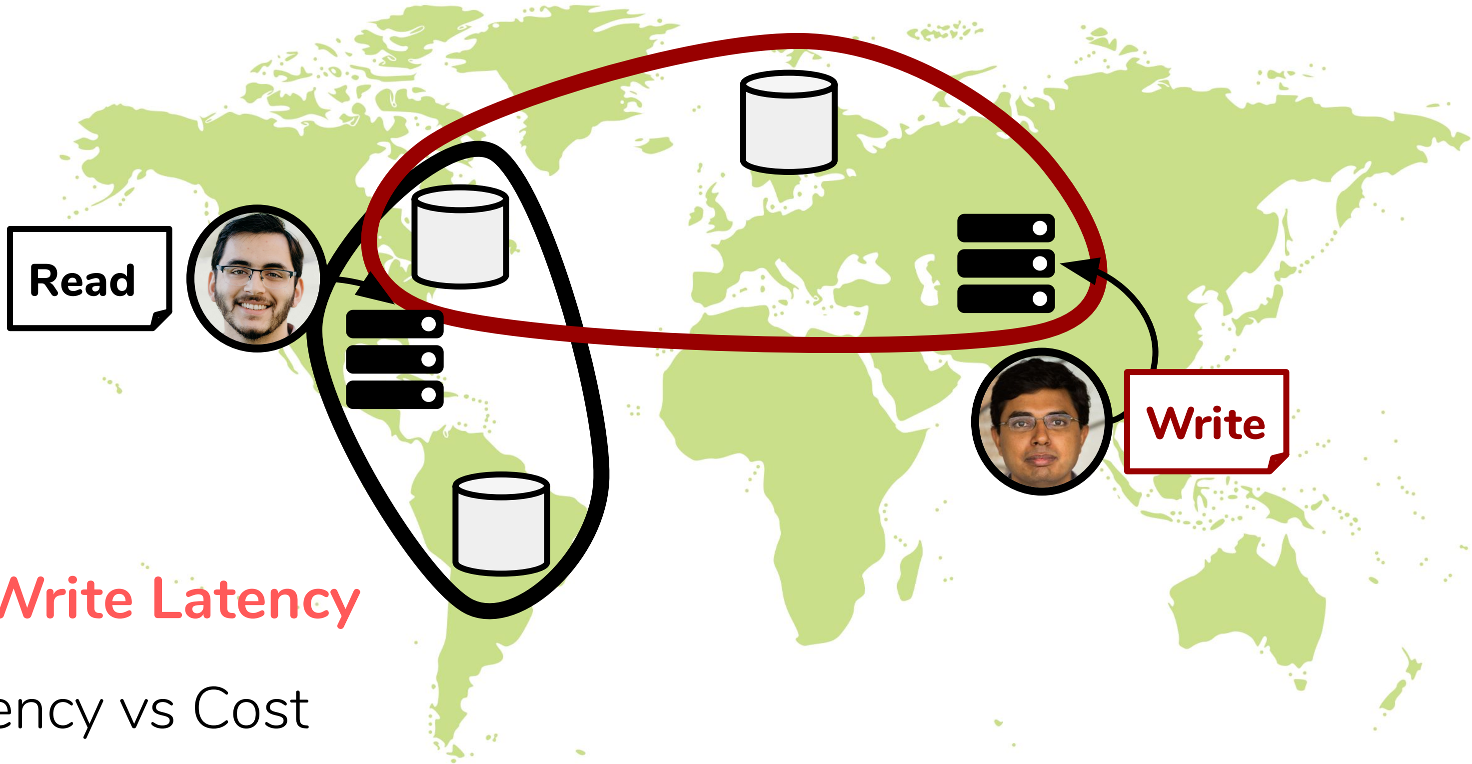
- Read vs Write Latency
- Read Latency vs Cost

Linearizability Imposes Unavoidable Trade-offs



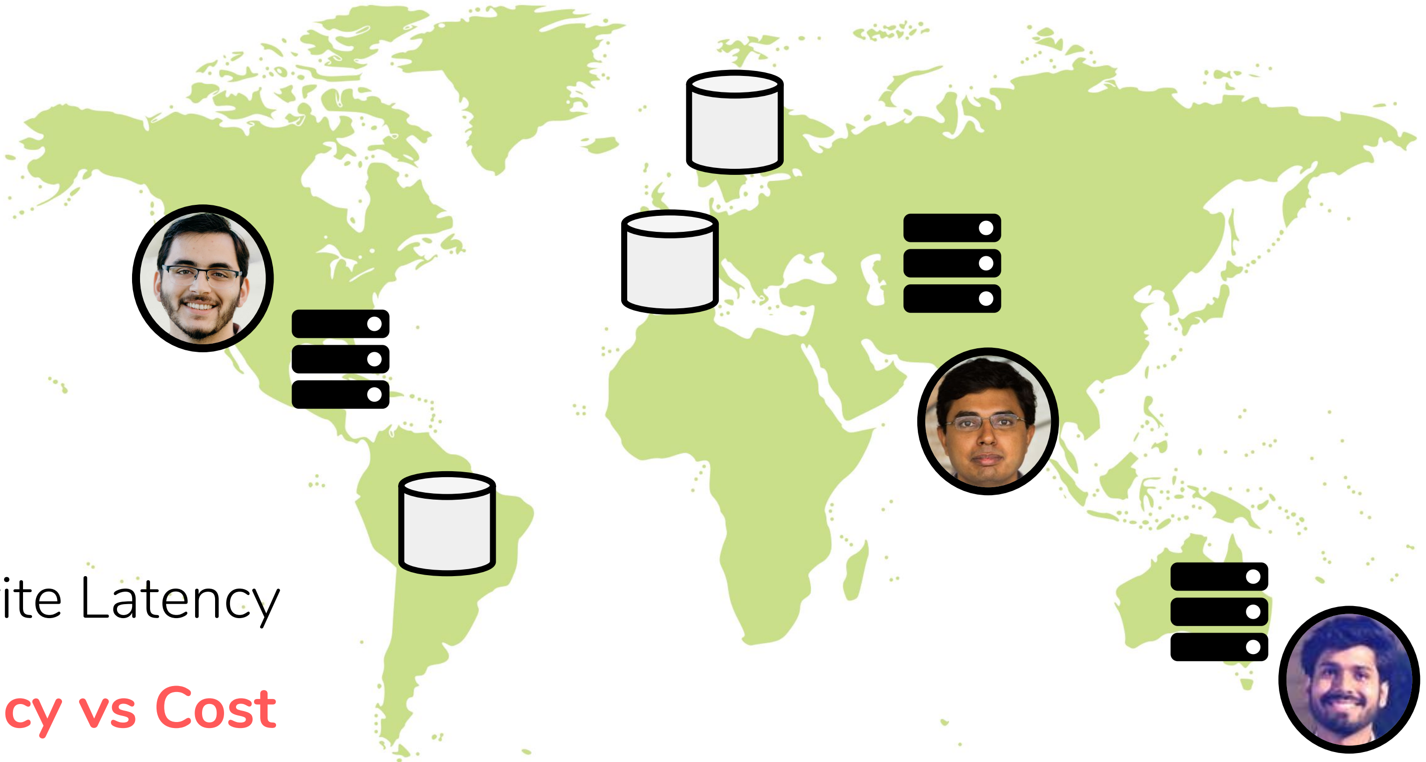
- Read vs Write Latency
- Read Latency vs Cost

Linearizability Imposes Unavoidable Trade-offs



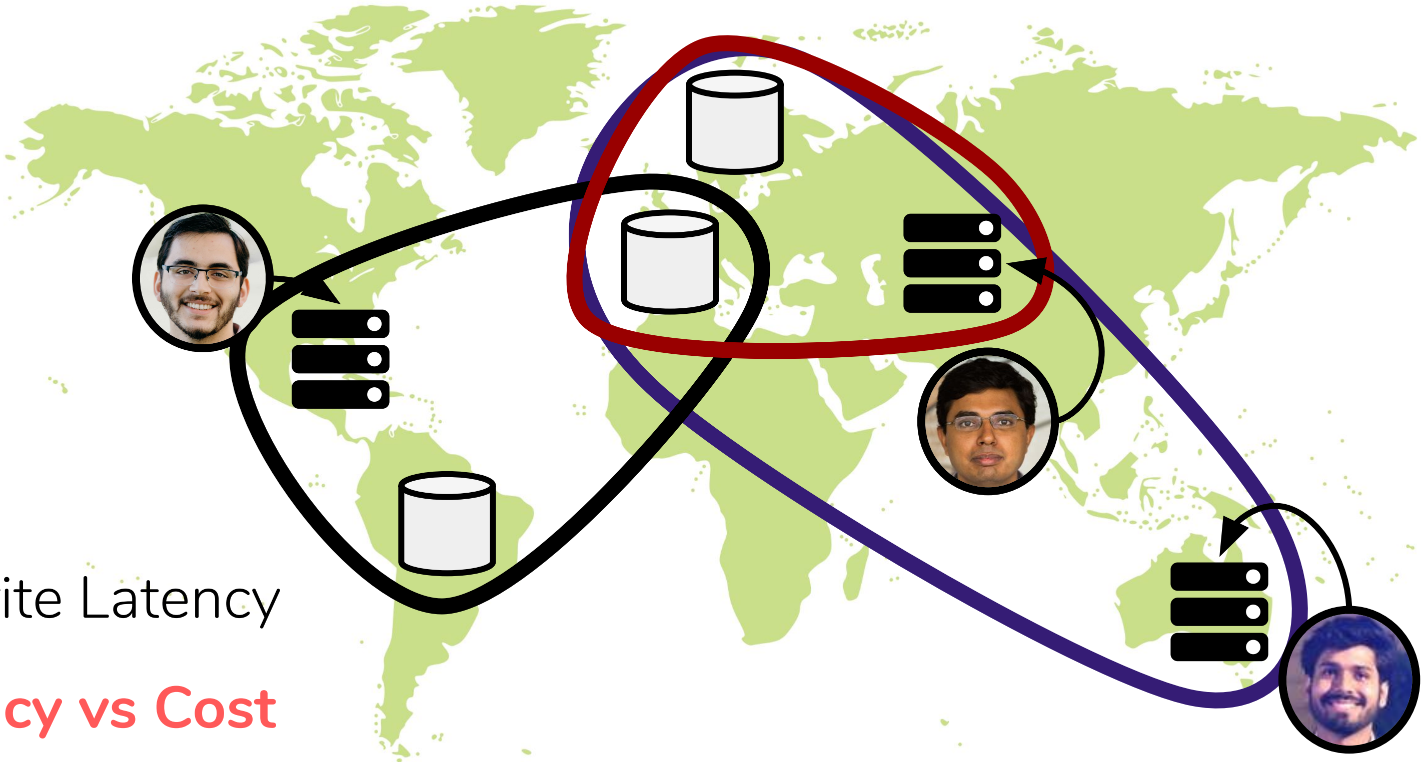
- Read vs Write Latency
- Read Latency vs Cost

Linearizability Imposes Unavoidable Trade-offs



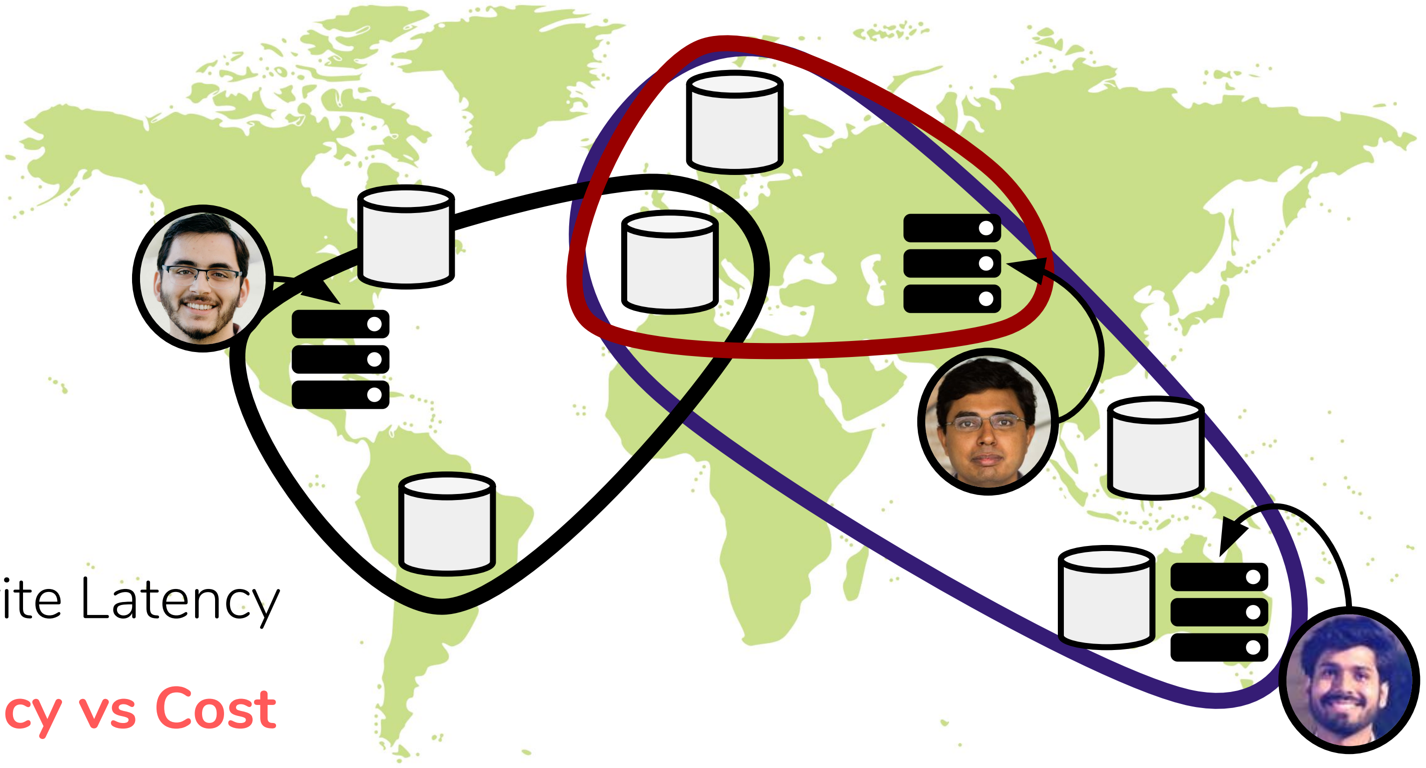
- Read vs Write Latency
- **Read Latency vs Cost**

Linearizability Imposes Unavoidable Trade-offs



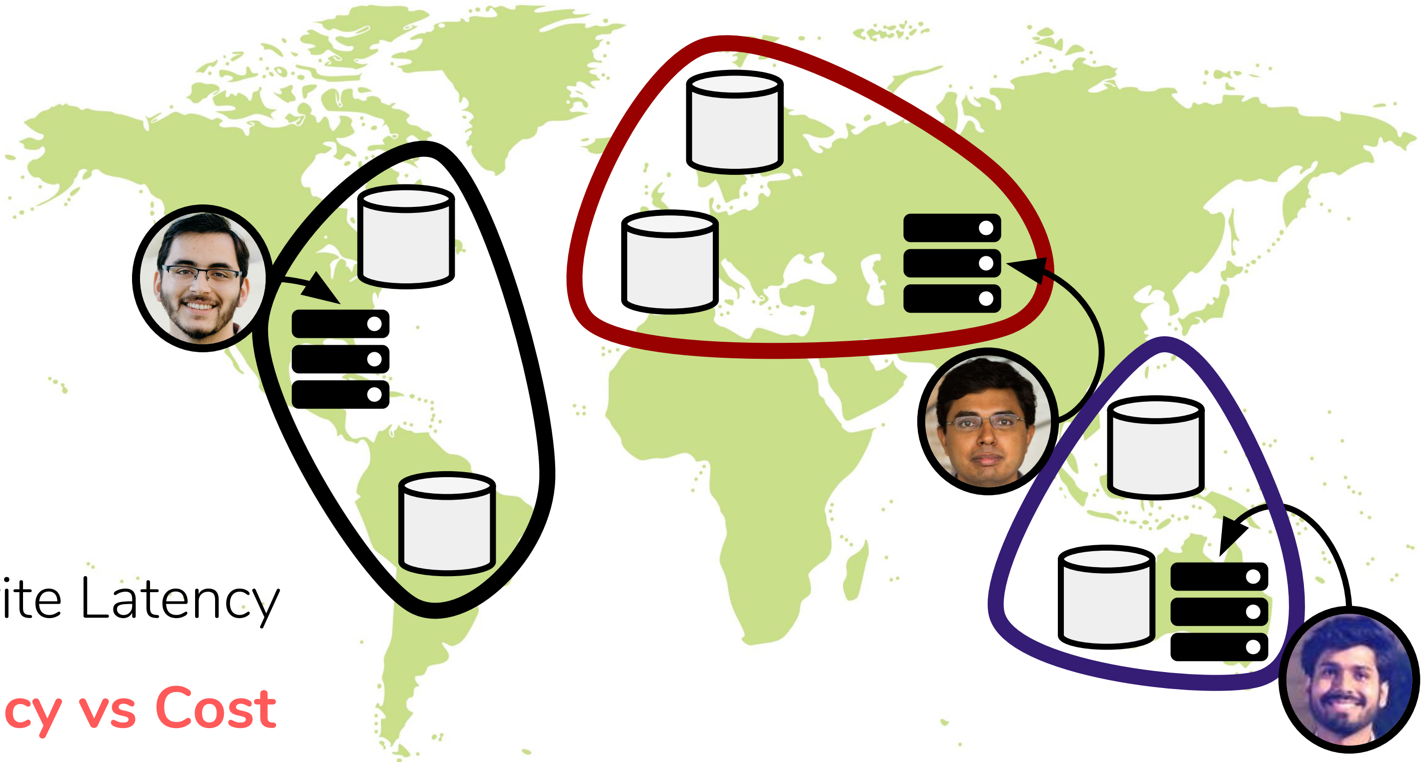
- Read vs Write Latency
- **Read Latency vs Cost**

Linearizability Imposes <u>Unavoidable</u> Trade-offs



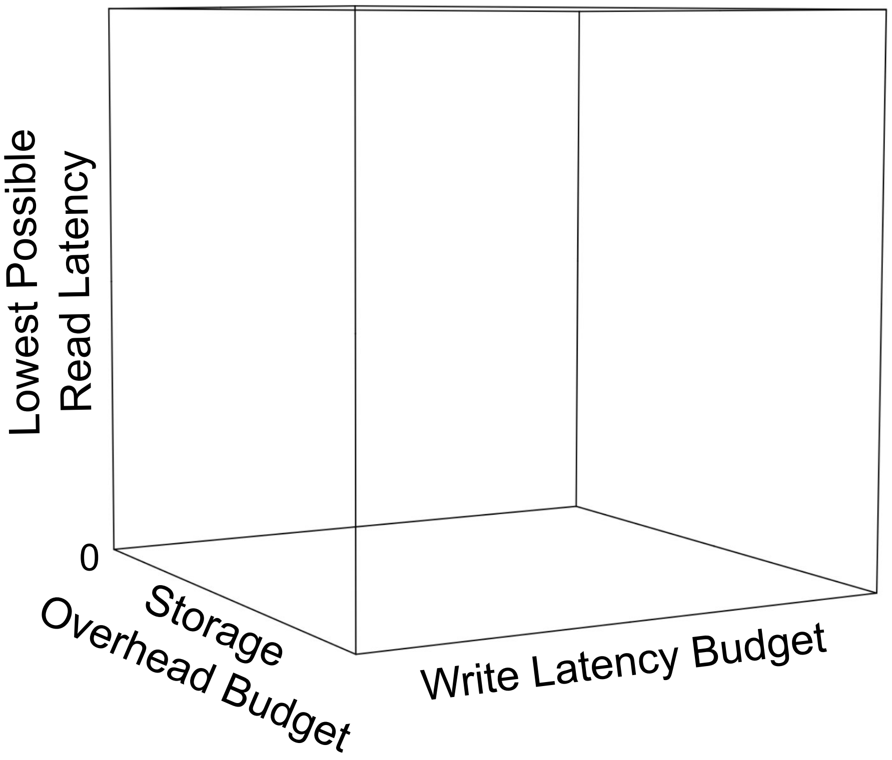
- Read vs Write Latency
- **Read Latency vs Cost**

Linearizability Imposes Unavoidable Trade-offs

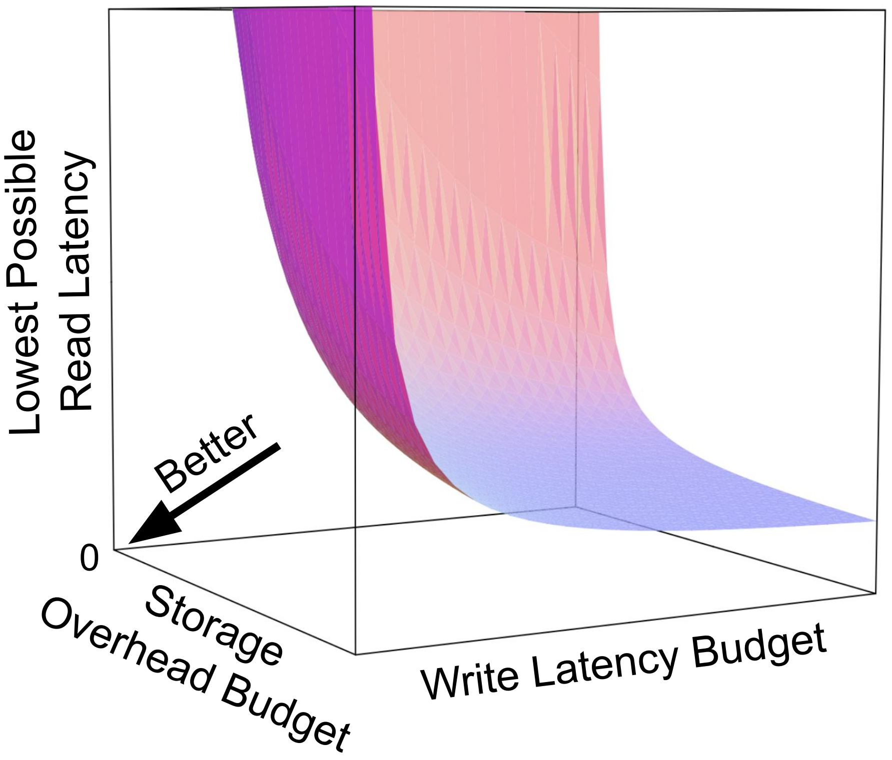


- Read vs Write Latency
- **Read Latency vs Cost**

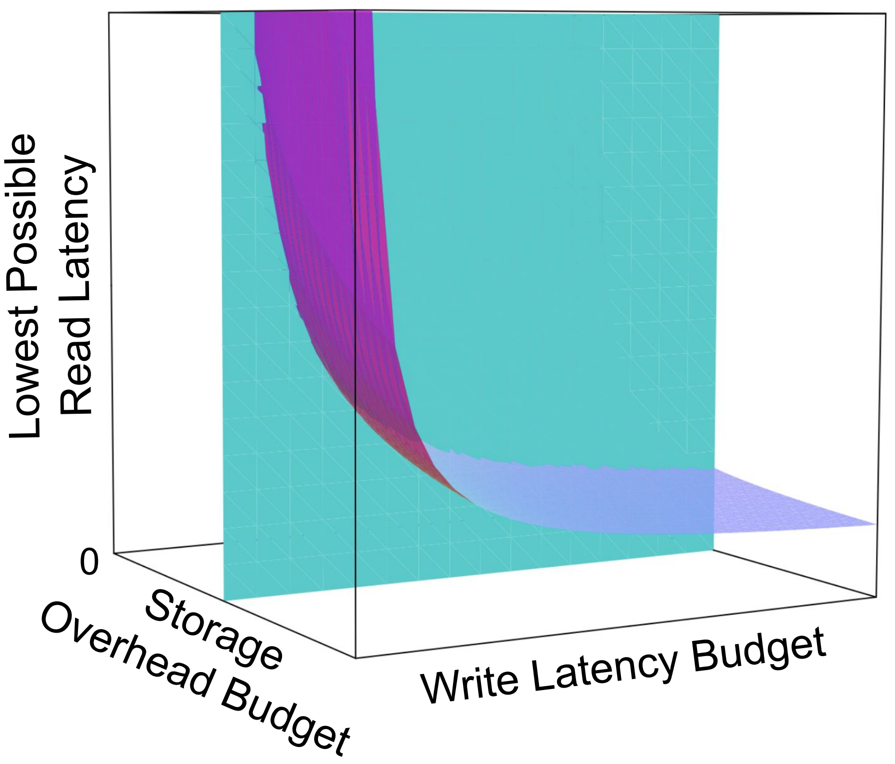
How Do Existing Approaches Perform?

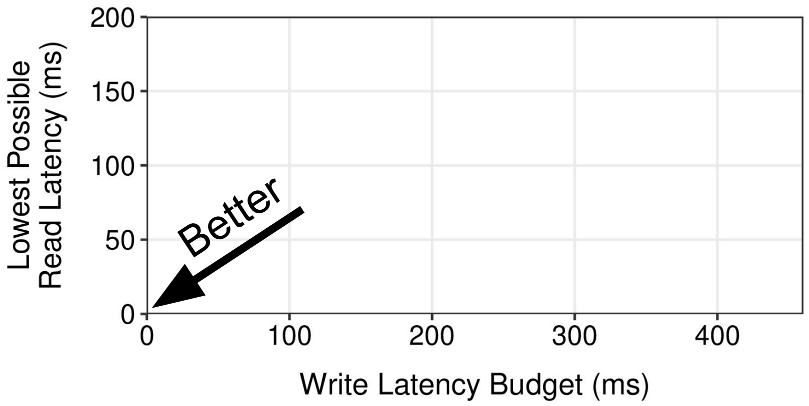


How Do Existing Approaches Perform?

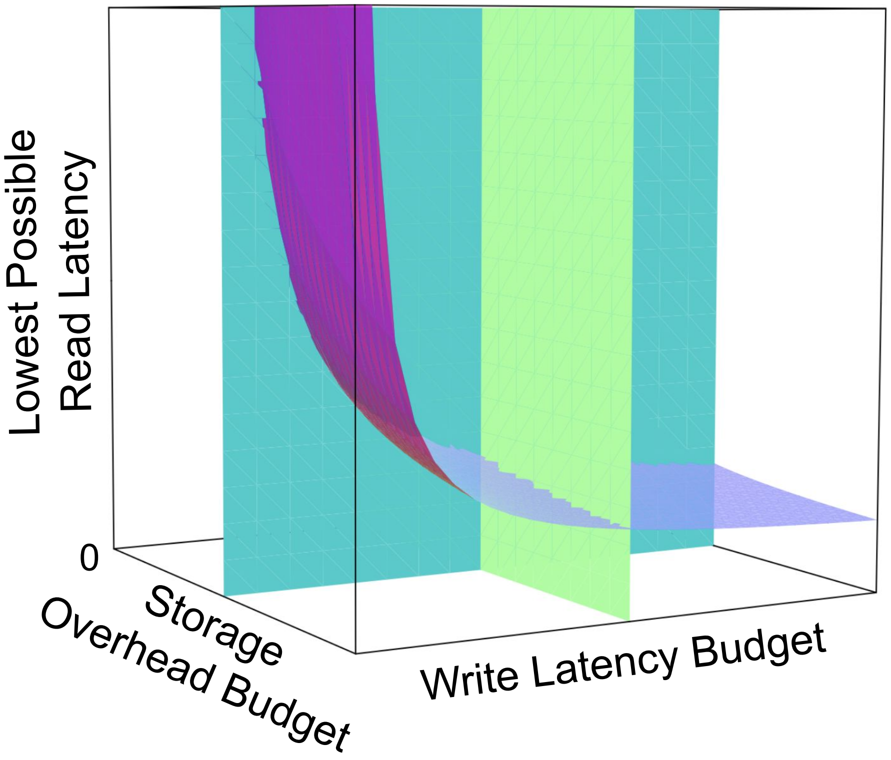


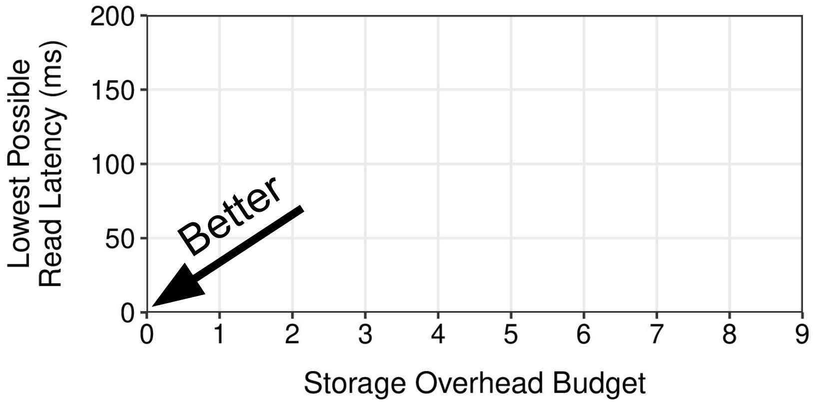
How Do Existing Approaches Perform?

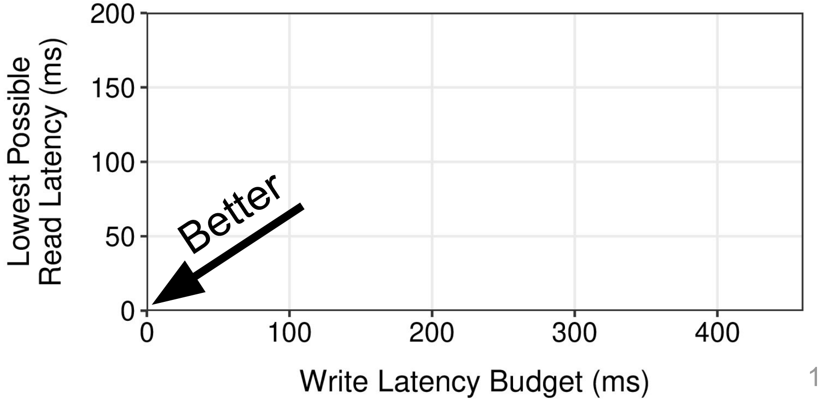




How Do Existing Approaches Perform?

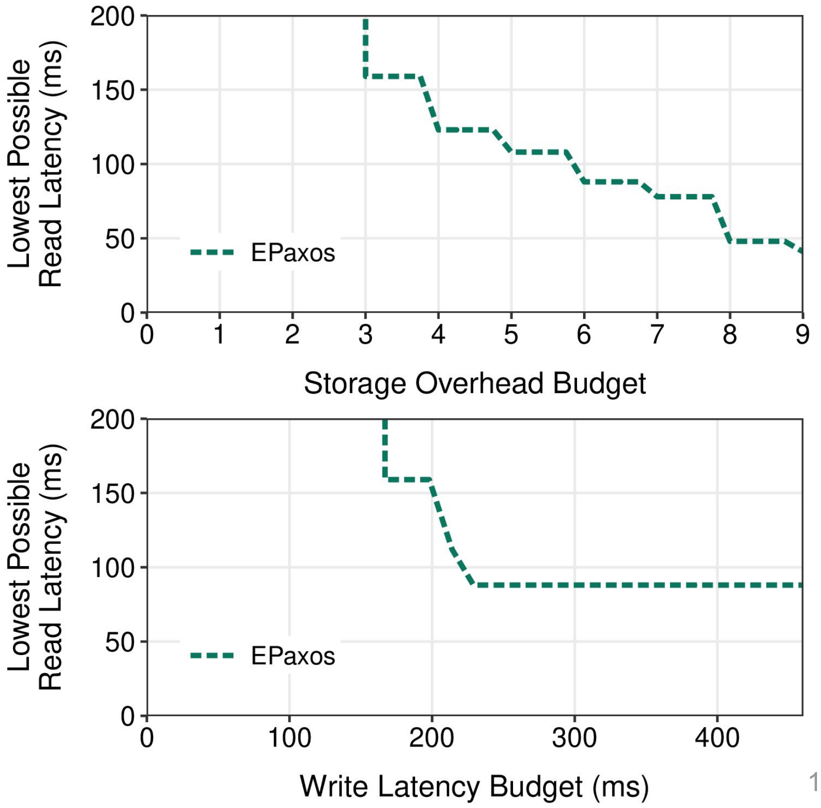






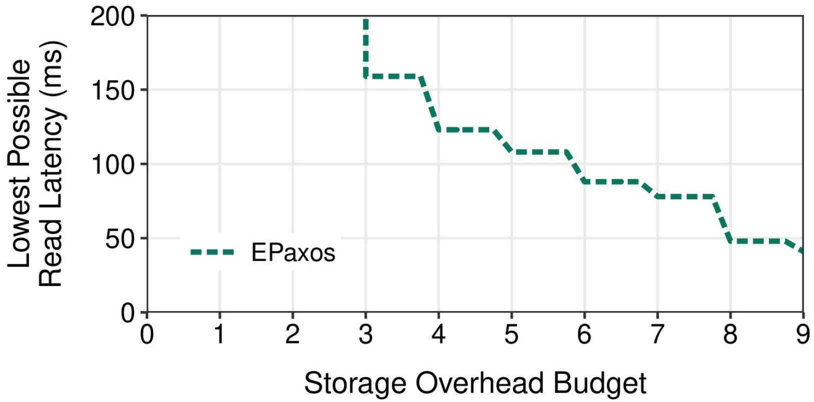
How Do Existing Approaches Perform?

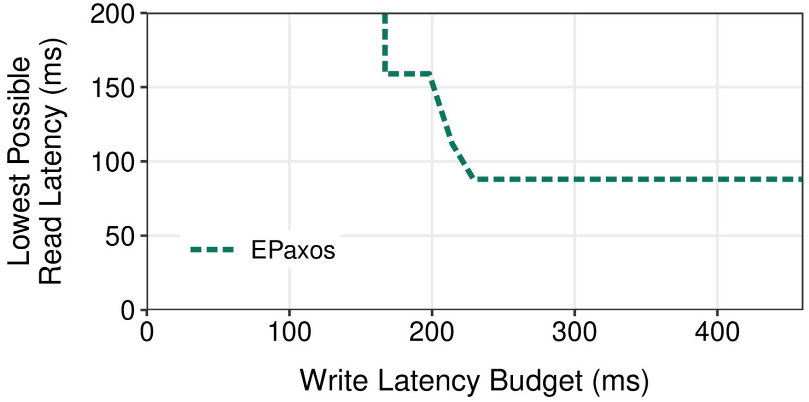
- EPaxos: state-of-the-art geo-replication protocol



How Do Existing Approaches Perform?

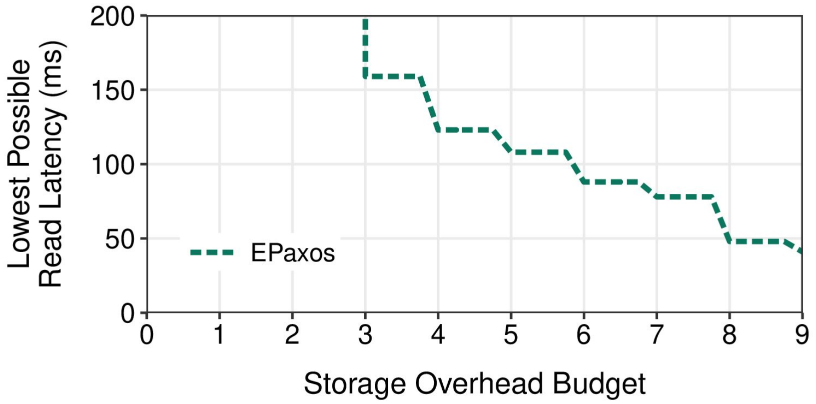
- EPaxos: state-of-the-art geo-replication protocol
- Compare with estimate of theoretical lower bound

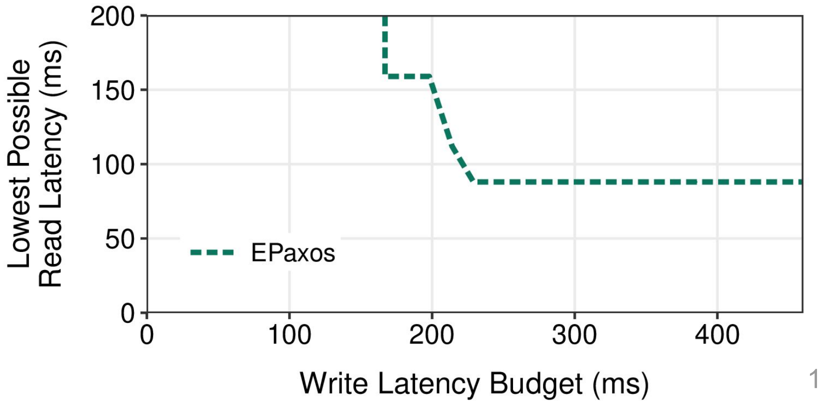




How Do Existing Approaches Perform?

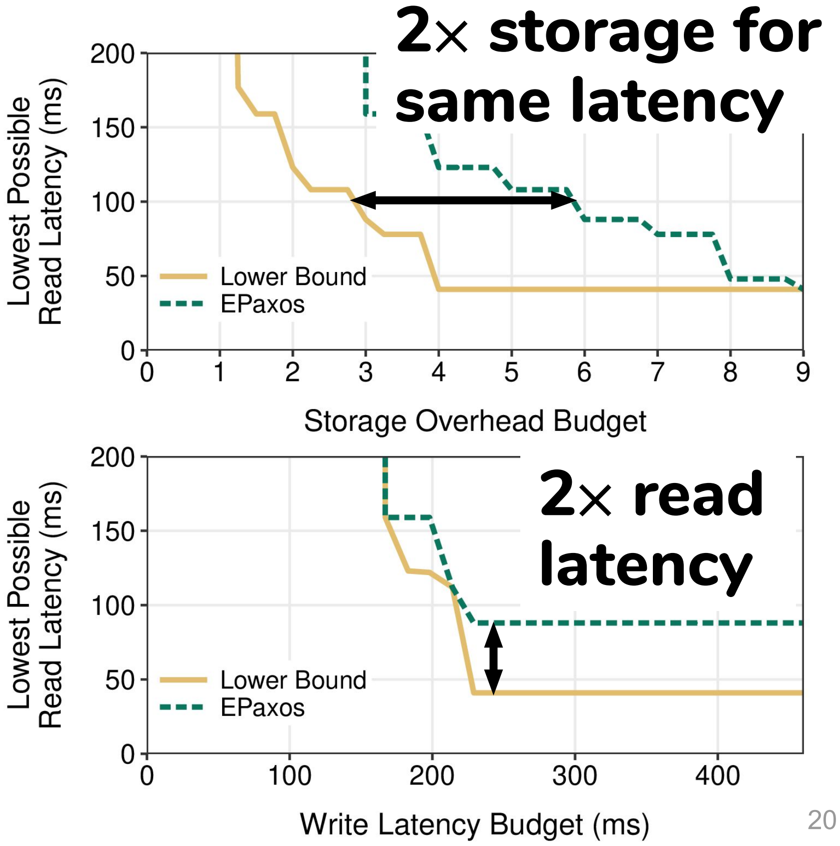
- EPaxos: state-of-the-art geo-replication protocol
- Compare with estimate of theoretical lower bound
  - No particular protocol
  - Respects consistency and fault-tolerance constraints





How Do Existing Approaches Perform?

- EPaxos: state-of-the-art geo-replication protocol
- Compare with estimate of theoretical lower bound
  - No particular protocol
  - Respects consistency and fault-tolerance constraints



How Do Existing Approaches Perform?

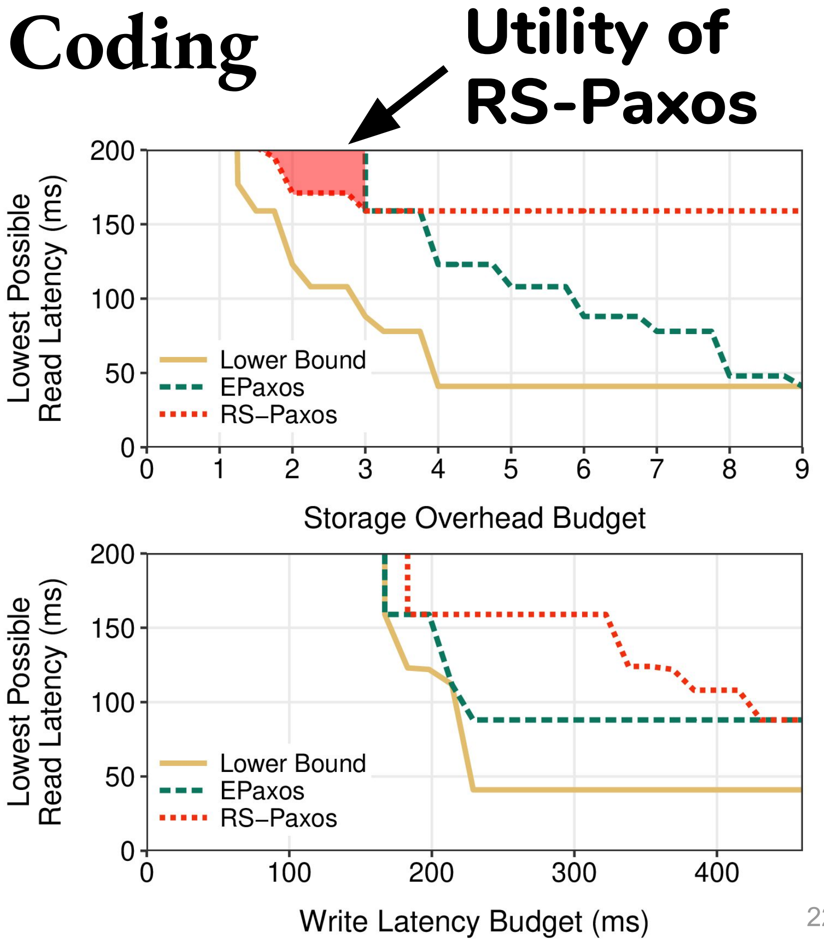
- EPaxos: state-of-the-art geo-re…
- Compa… theoret…
  - No p…
  - Resp… fault-tolerance constraints

Core problem: replication

Each site stores a full copy

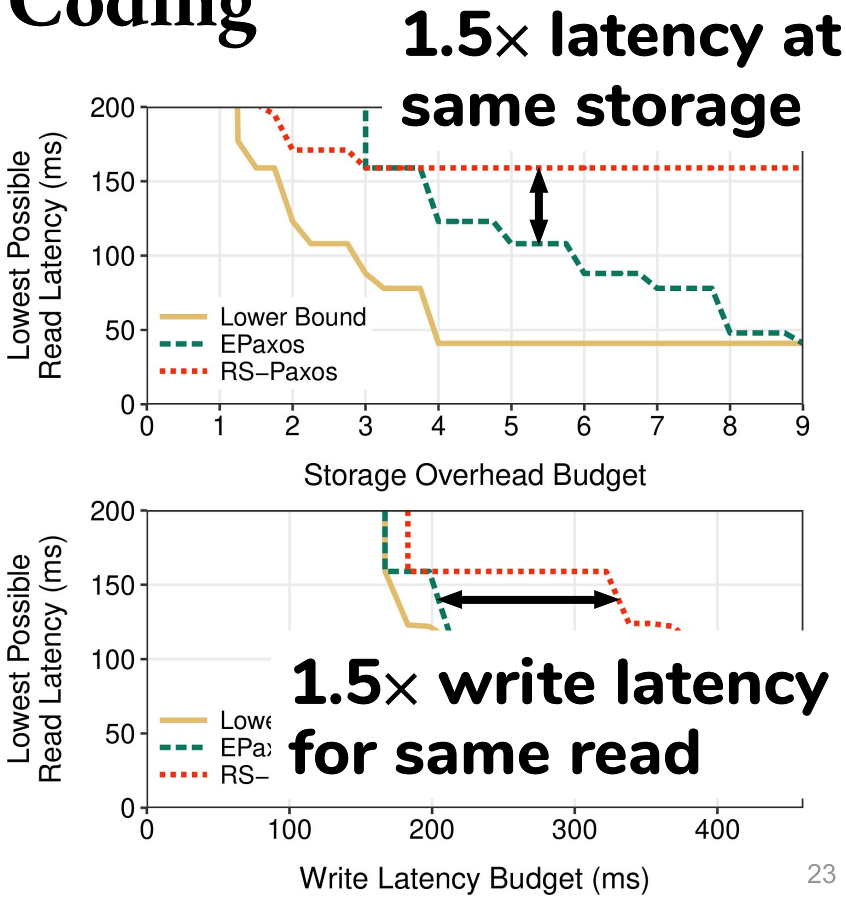
Lowering Cost with Erasure Coding

- Each site stores 1/kth of the data
- RS-Paxos: Paxos on erasure-coded data



Lowering Cost with Erasure Coding

- Each site stores 1/kth of the data
- RS-Paxos: Paxos on erasure-coded data

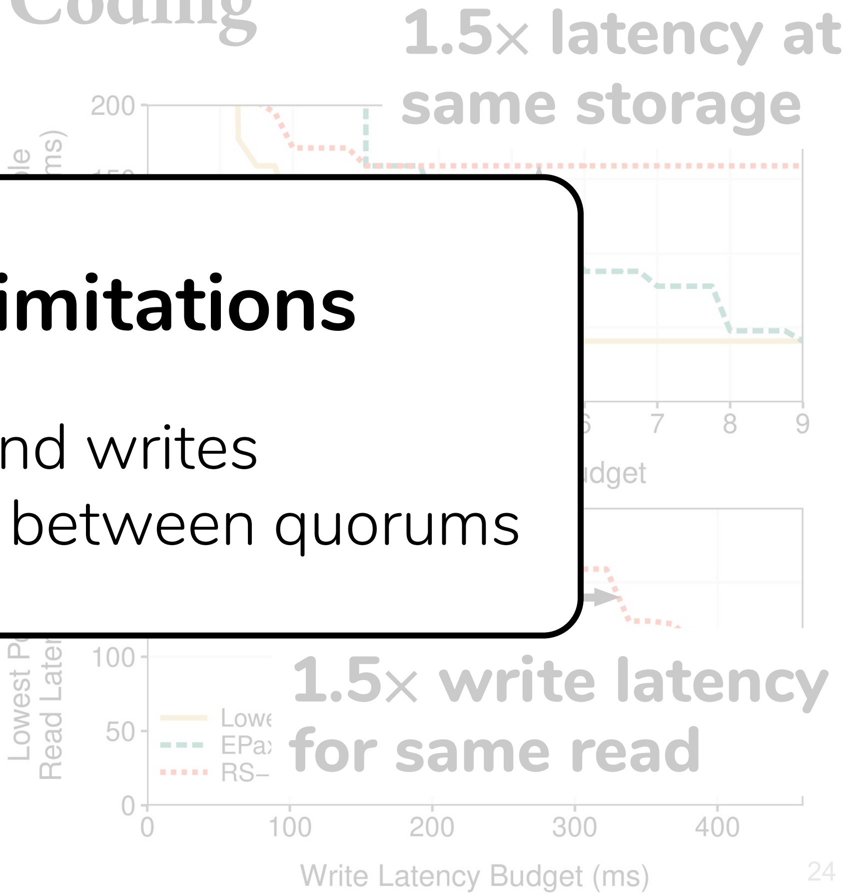


Lowering Cost with Erasure Coding

- Each site stores 1/kth of the data
- RS-Pax… erasure…

RS-Paxos Limitations

- Two-round writes
- k-site intersection between quorums

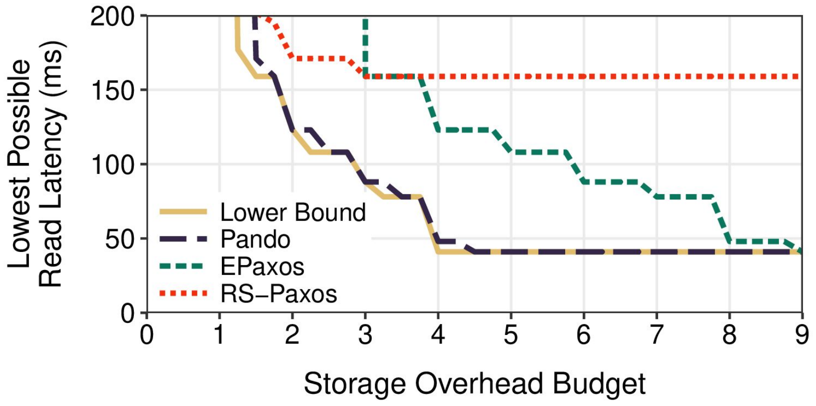


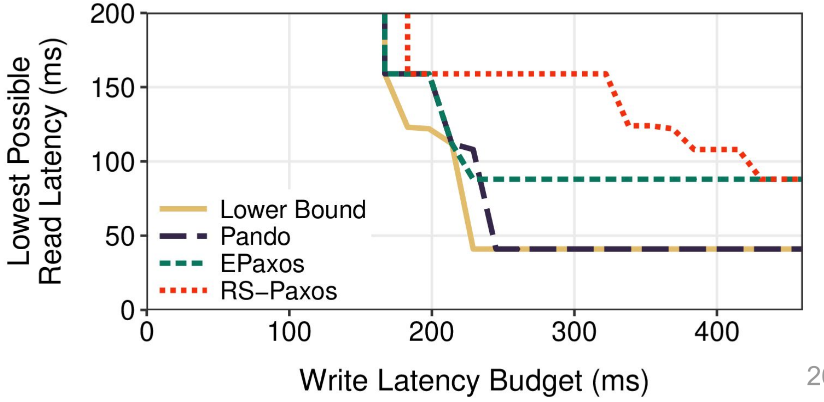
Recap of the Problem

- Want to spread data across DCs, but constraints that impose trade-offs
- State-of-the-art falls short of the optimal
- Use erasure coding $\rightarrow$ hurts latency

Pando: Near-Optimal Trade-off

- ~~Two-round writes~~
  Approximates latency of one-round writes
- ~~k-site intersection between quorums~~
  1-site intersection (common-case)





Paxos Review

Paxos Made Moderately Complex

ROBBERT VAN RENESSE and DENIZ ALTINBUKEN, Cornell University

This article explains the full reconfigurable multidecree Paxos (or multi-Paxos) protocol. Paxos is by no means a simple protocol, even though it is based on relatively simple invariants. We provide pseudocode and explain it guided by invariants. We initially avoid optimizations that complicate comprehension. Next we discuss liveness, list various optimizations that make the protocol practical, and present variants of the protocol.

Categories and Subject Descriptors: C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*Network operating systems*; D.4.5 [**Operating Systems**]: Reliability—*Fault-tolerance*

General Terms: Design, Reliability

Additional Key Words and Phrases: Replicated state machines, consensus, voting

ACM Reference Format:
Robbert van Renesse and Deniz Altinbuken. 2015. Paxos made moderately complex. ACM Comput. Surv. 47, 3, Article 42 (February 2015), 36 pages.
DOI: http://dx.doi.org/10.1145/2673577

1. INTRODUCTION

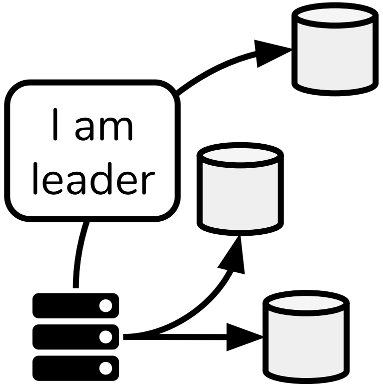
Paxos [Lamport 1998] is a protocol for state machine replication in an asynchronous environment that admits crash failures. It is useful to consider the terms in this sentence carefully:

42

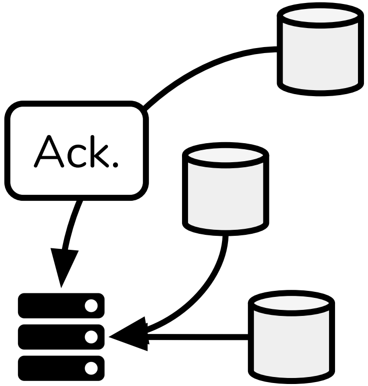
Paxos Review

- 2-Phase writes: first become leader

30 ms

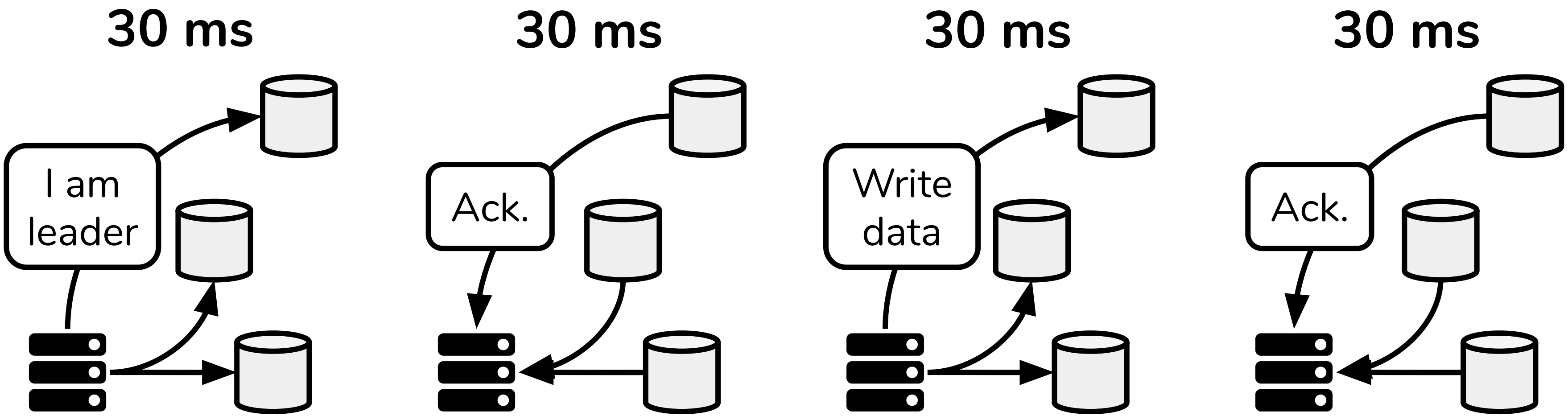


30 ms



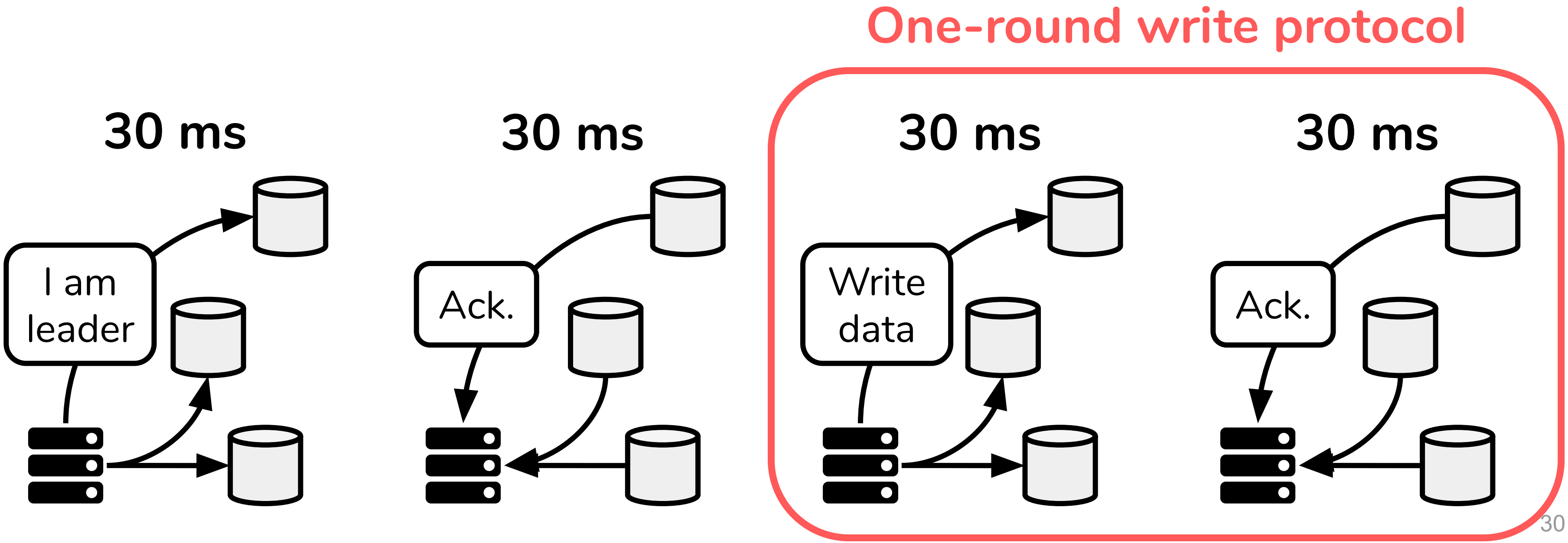
Paxos Review

- 2-Phase writes: first become leader, then write



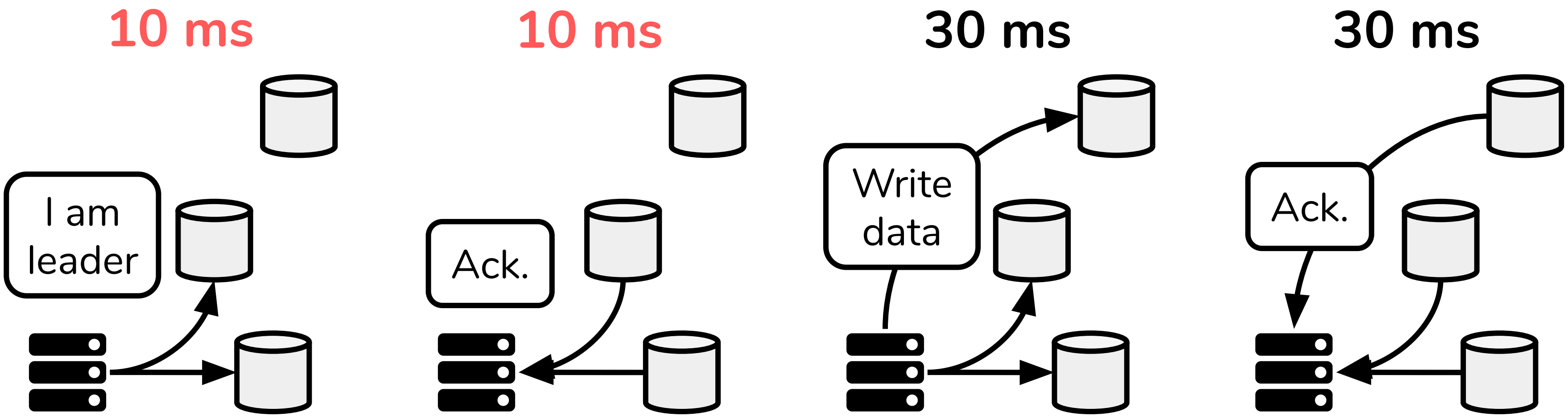
Paxos Review

- 2-Phase writes: first become leader, then write



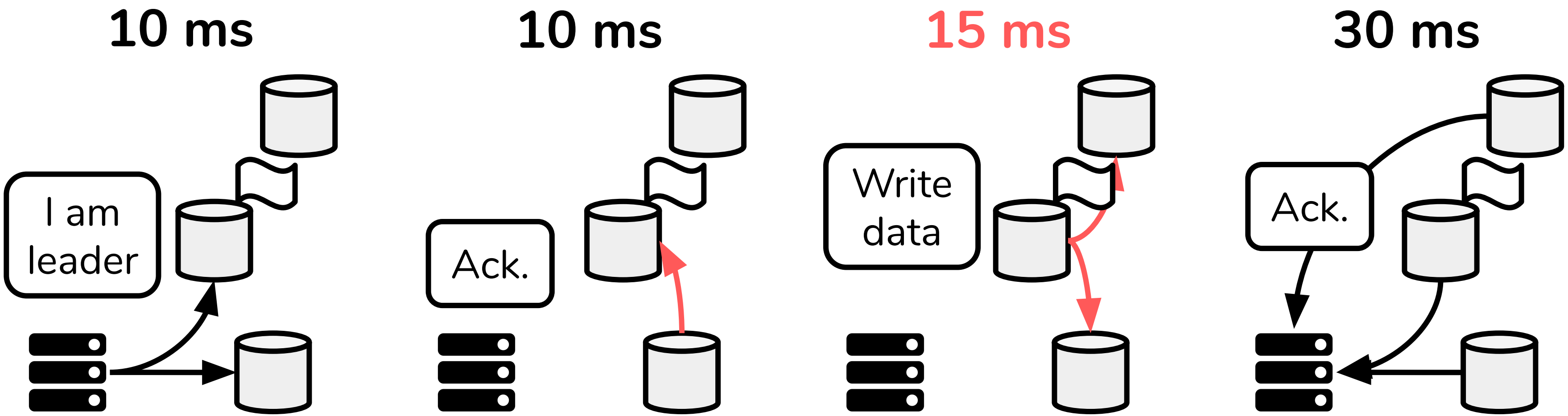
Quickly Executing 2-Phase Writes

- Step 1: faster Phase 1
  - Flexible Paxos [OPODIS'16]: need Phase 1, 2 quorums to intersect
  - **Phase 1 quorums need not overlap**



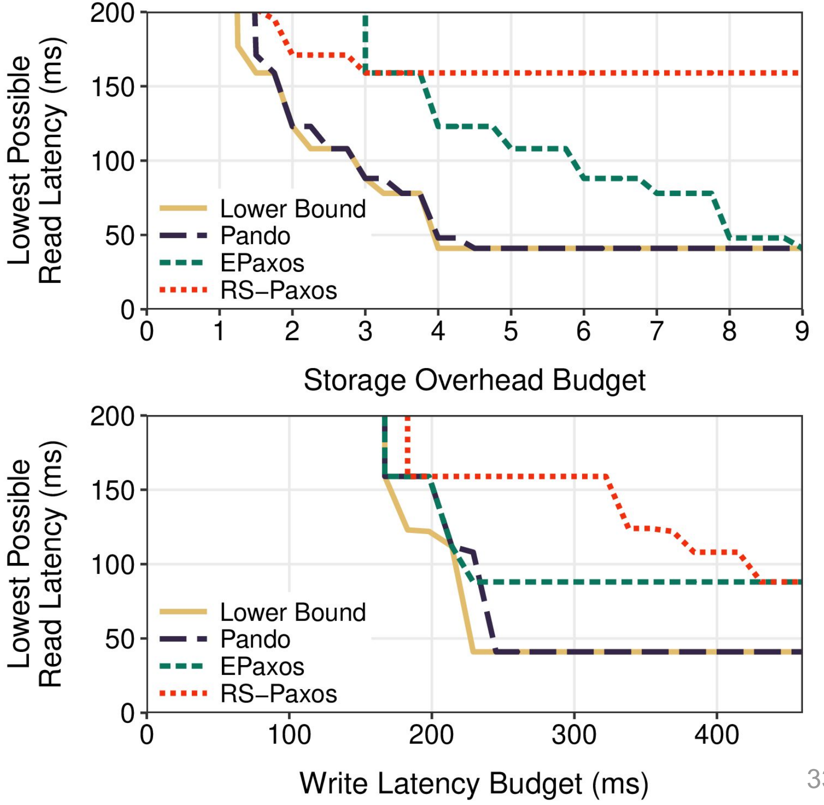
Quickly Executing 2-Phase Writes

- Step 1: faster Phase 1
- Step 2: overlap latency cost of Phase 1 with Phase 2
  - RPC Chains [NSDI'09]: **start Phase 2 at a delegate**



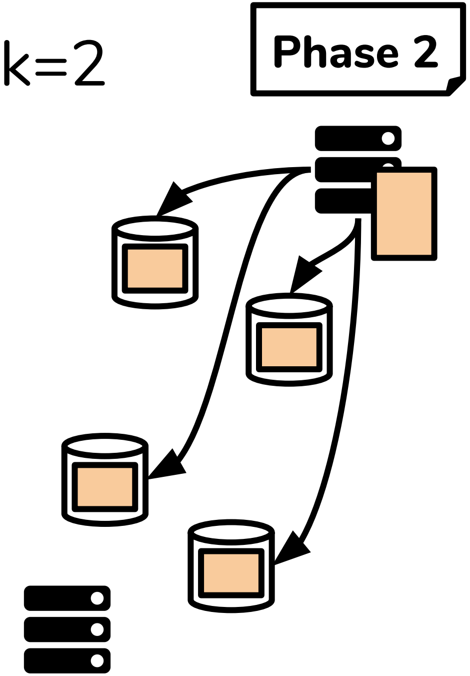
Pando: Near-Optimal Trade-off

- ~~Two-round writes~~
  Approximates latency of one-round writes
- ~~k-site intersection between quorums~~
  1-site intersection (common-case)



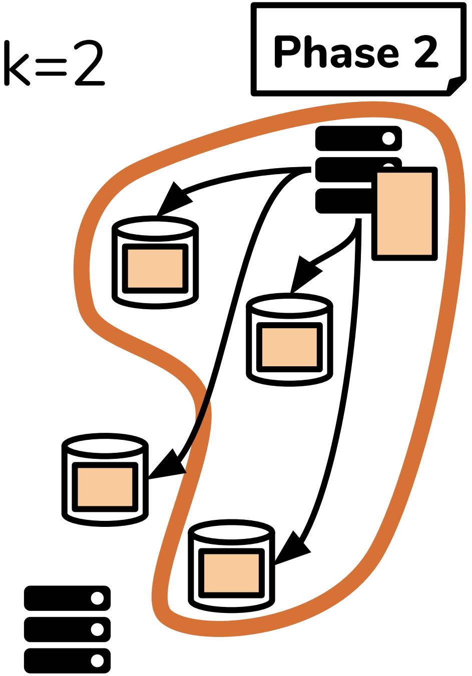
Write to All

k=2

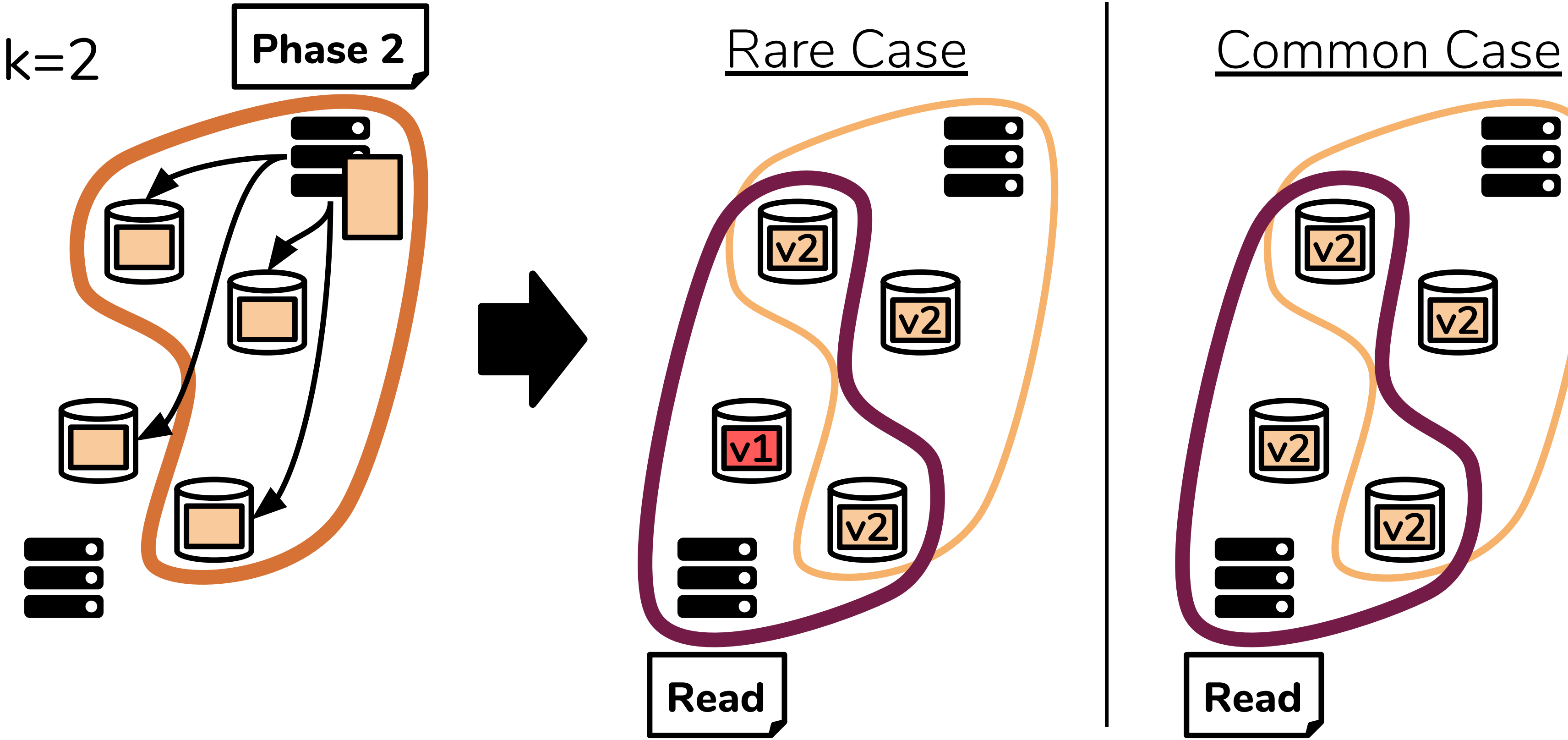


Write to All, Wait for Quorum

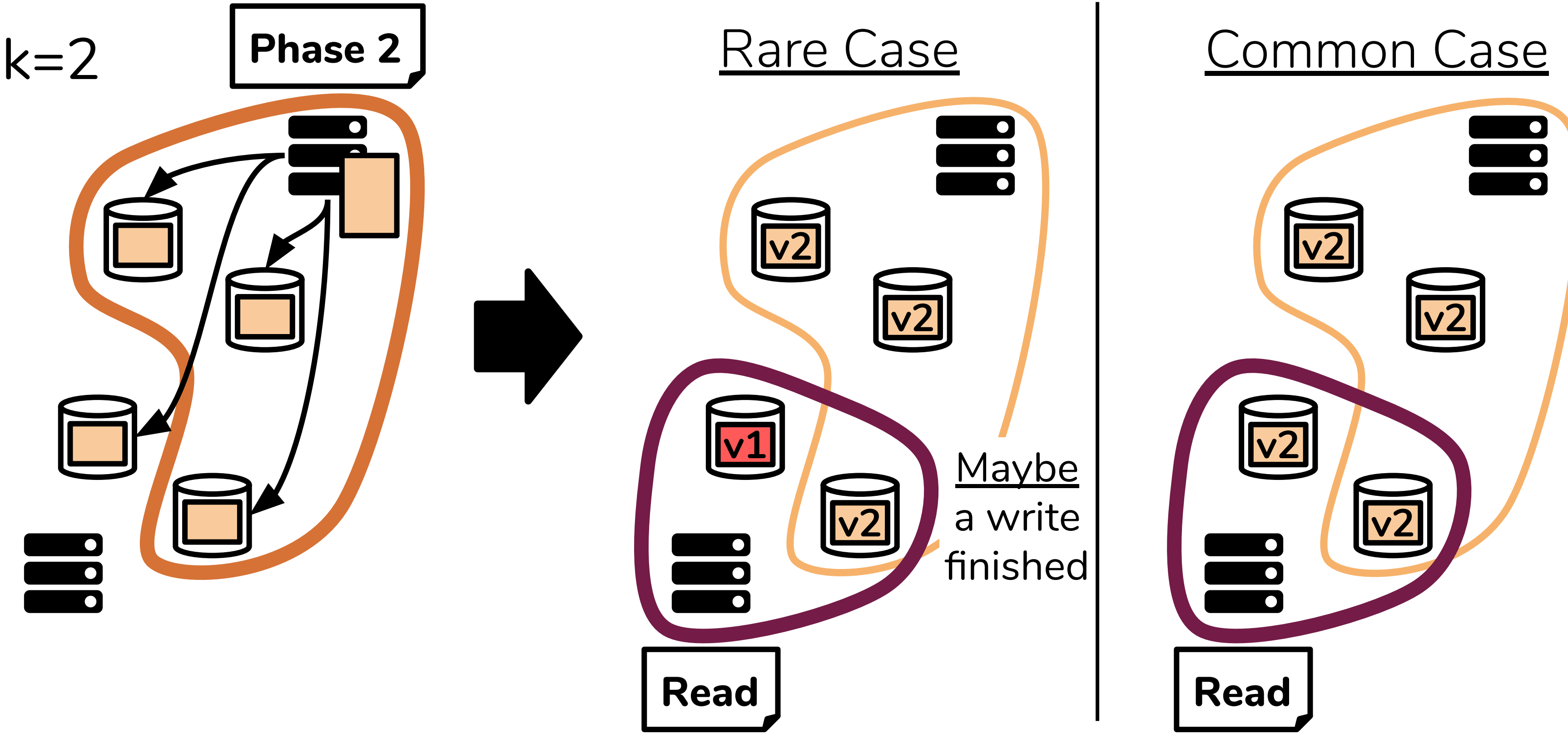
k=2



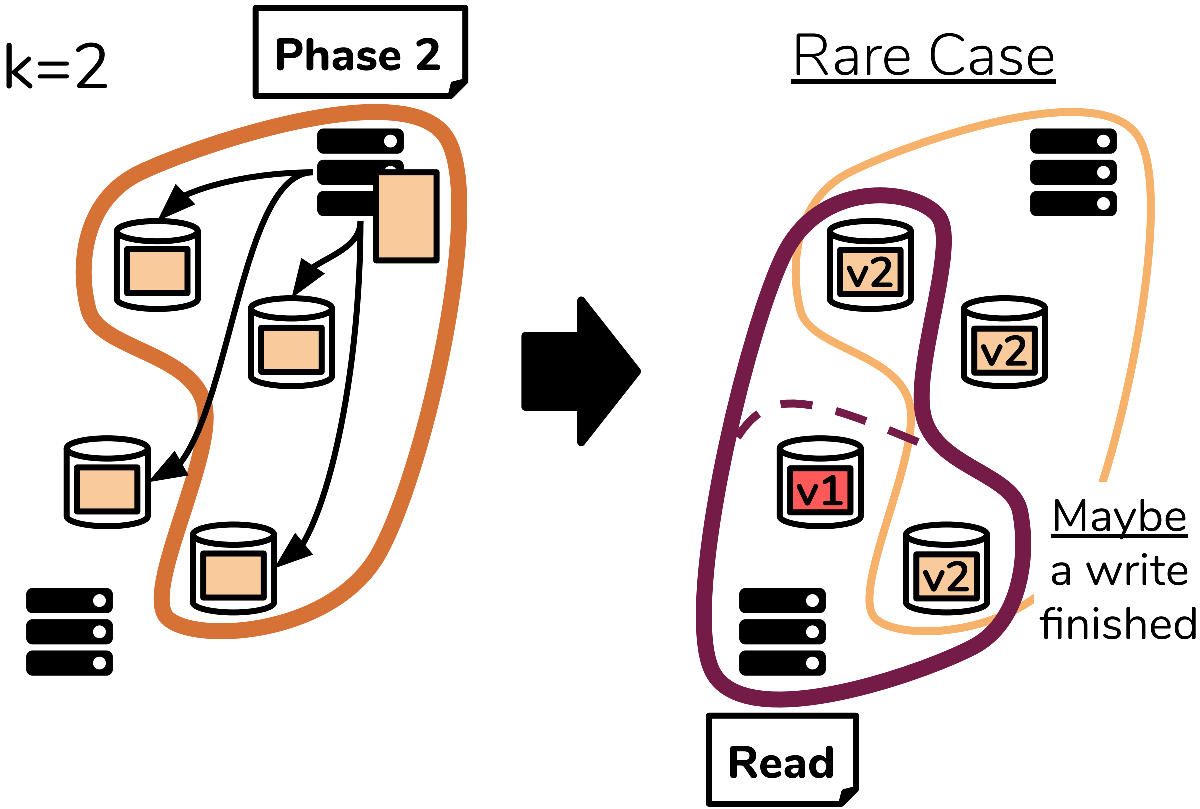
Write to All, Wait for Quorum



Achieving 1-Site Intersection

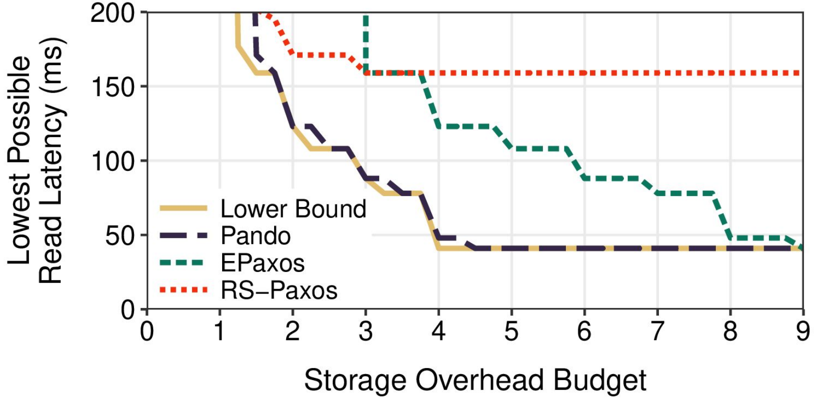


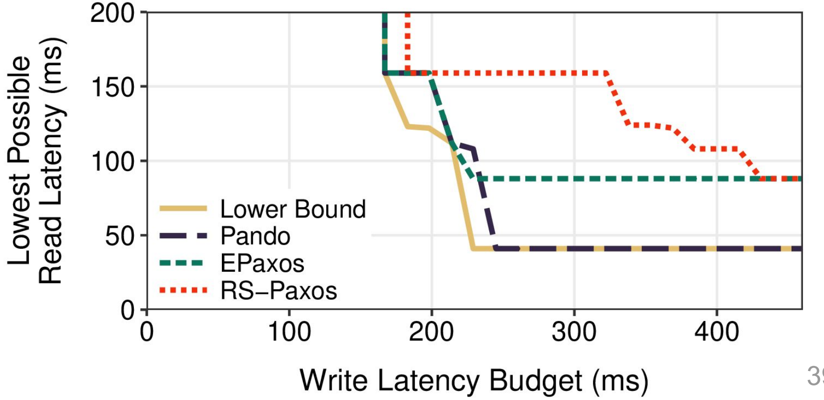
Achieving 1-Site Intersection



Pando: Near-Optimal Trade-off

- ~~Two-round writes~~
  Approximates latency of one-round writes
- ~~k-site intersection between quorums~~
  1-site intersection (common-case)





Pando: Near-Optimal Trade-off

✓ ~~Two-round writes~~

Ap…

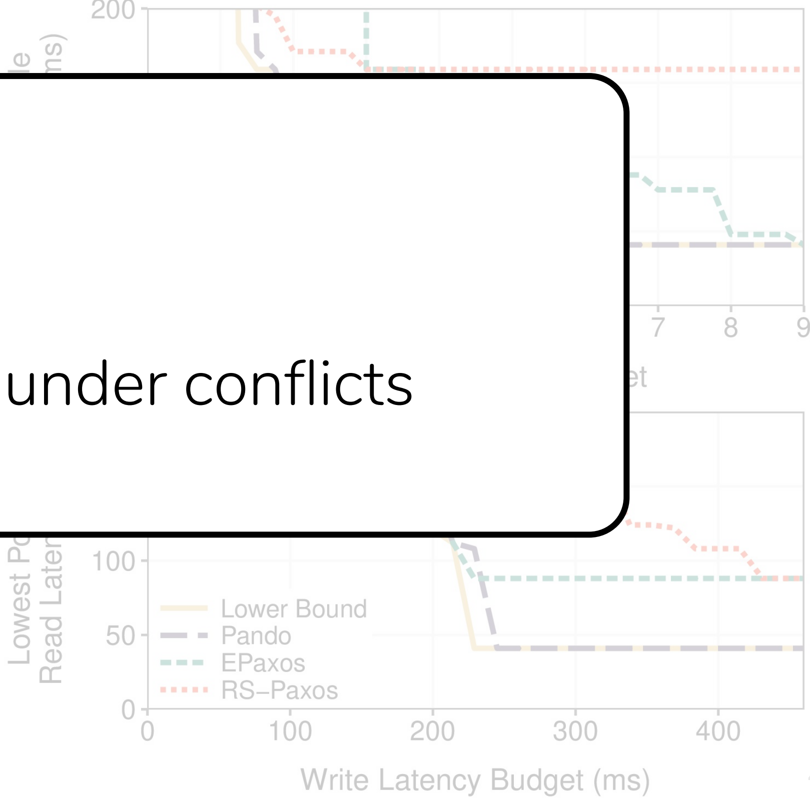
on…

✓ ~~k-s…~~

~~qu…~~

1-site intersection

(common-case)



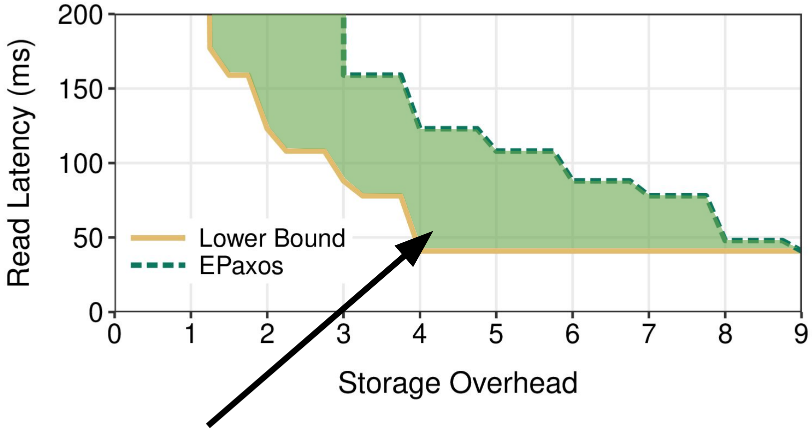
See paper:

- Correctness
- Bounding latency under conflicts

Evaluation: Proximity to Lower Bound

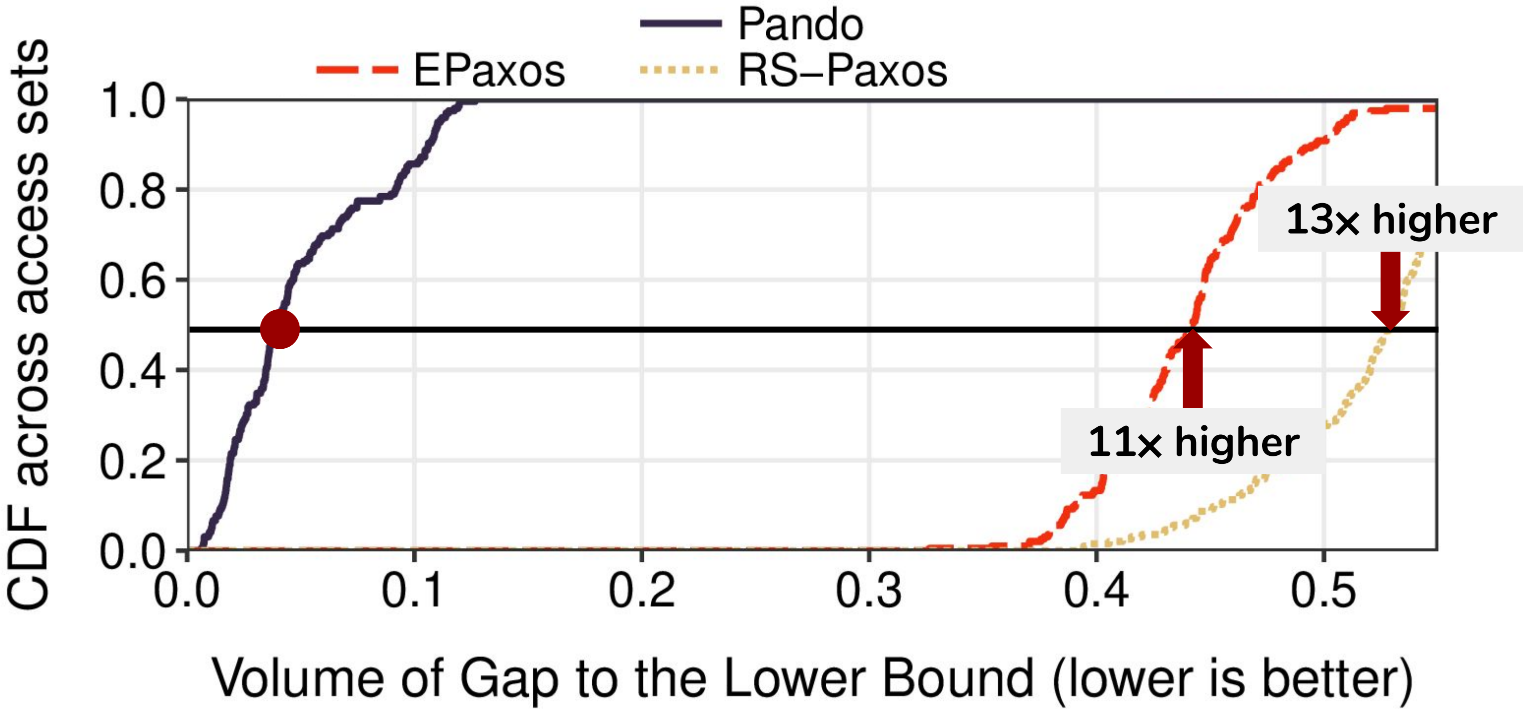
- Access set: DCs hosting web servers reading/writing data
- MIP solver selects data sites to minimize latency
- 500 access sets

Sample access set

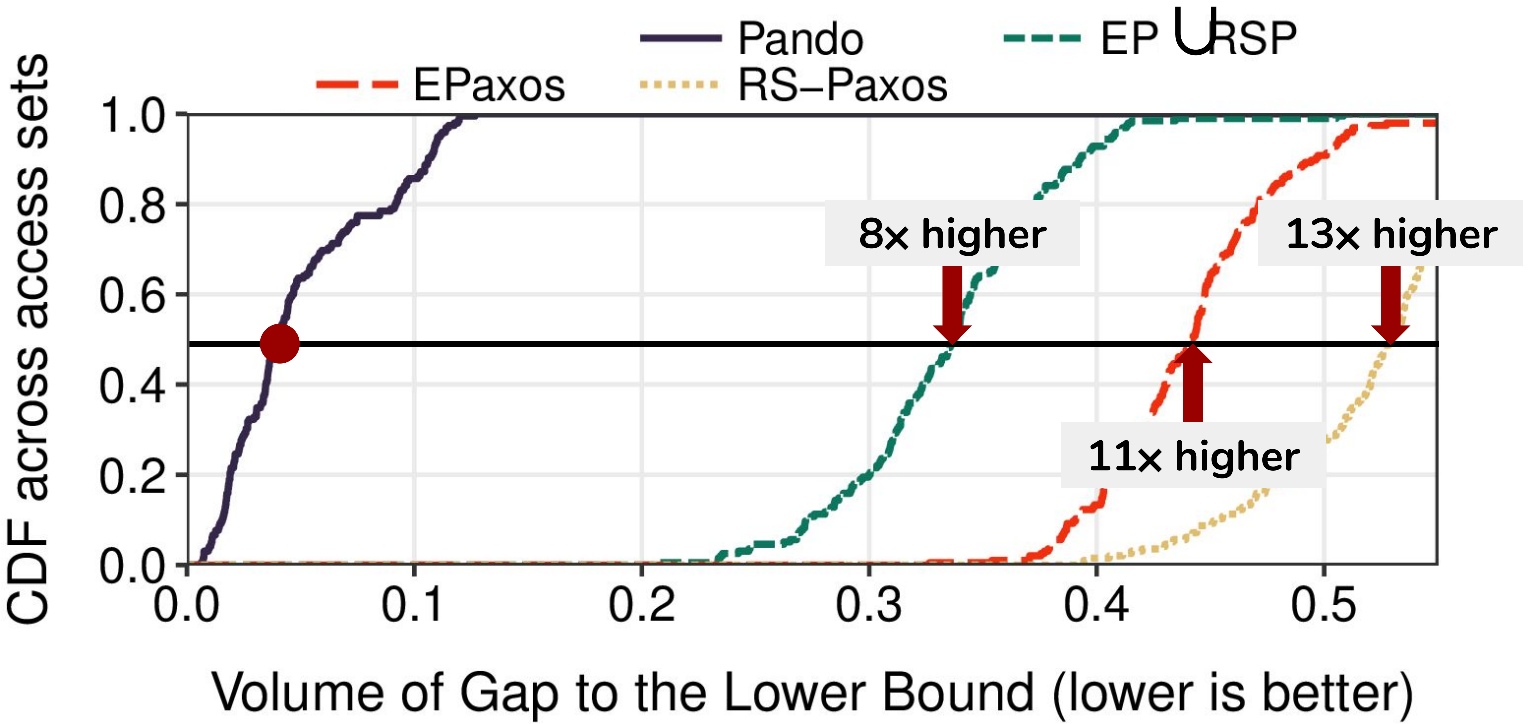


Measure gap

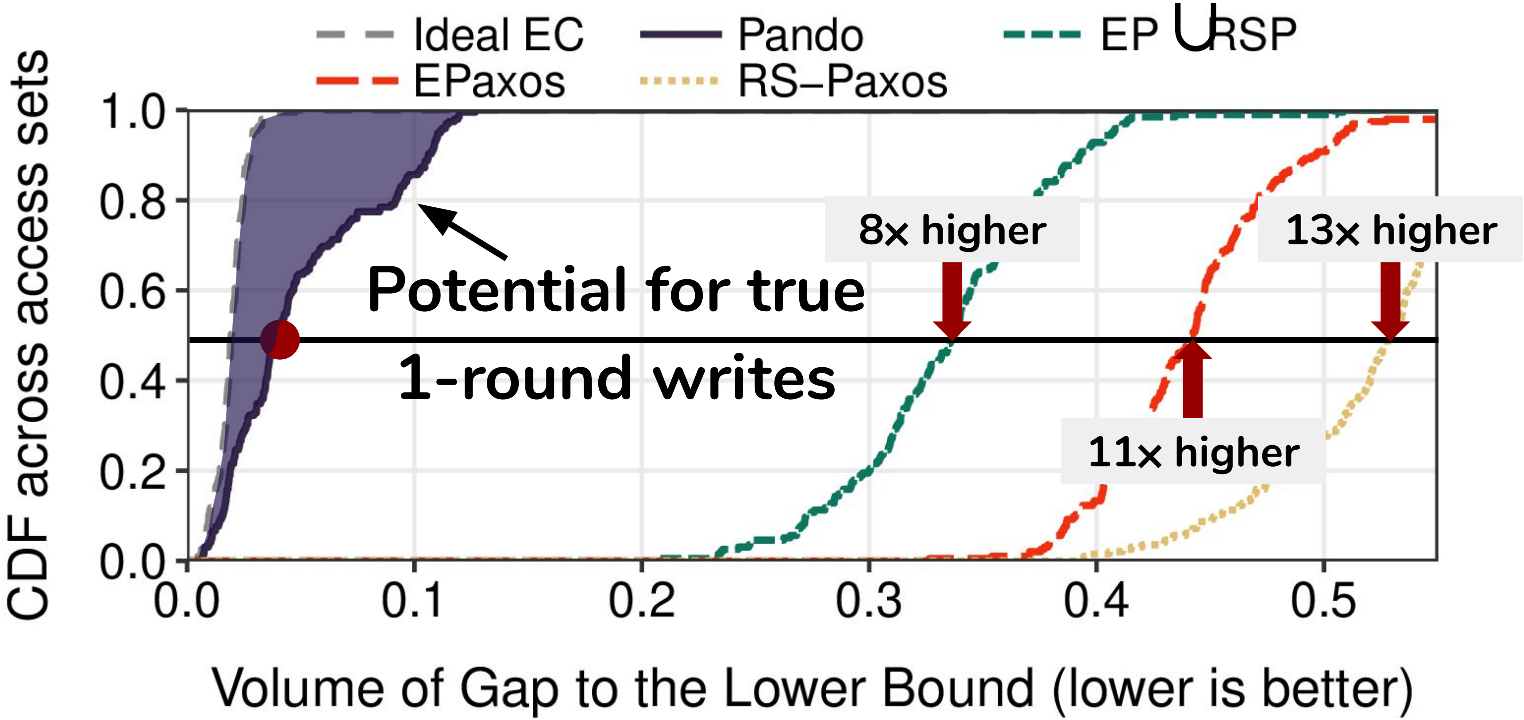
Pando is Close to the Lower Bound



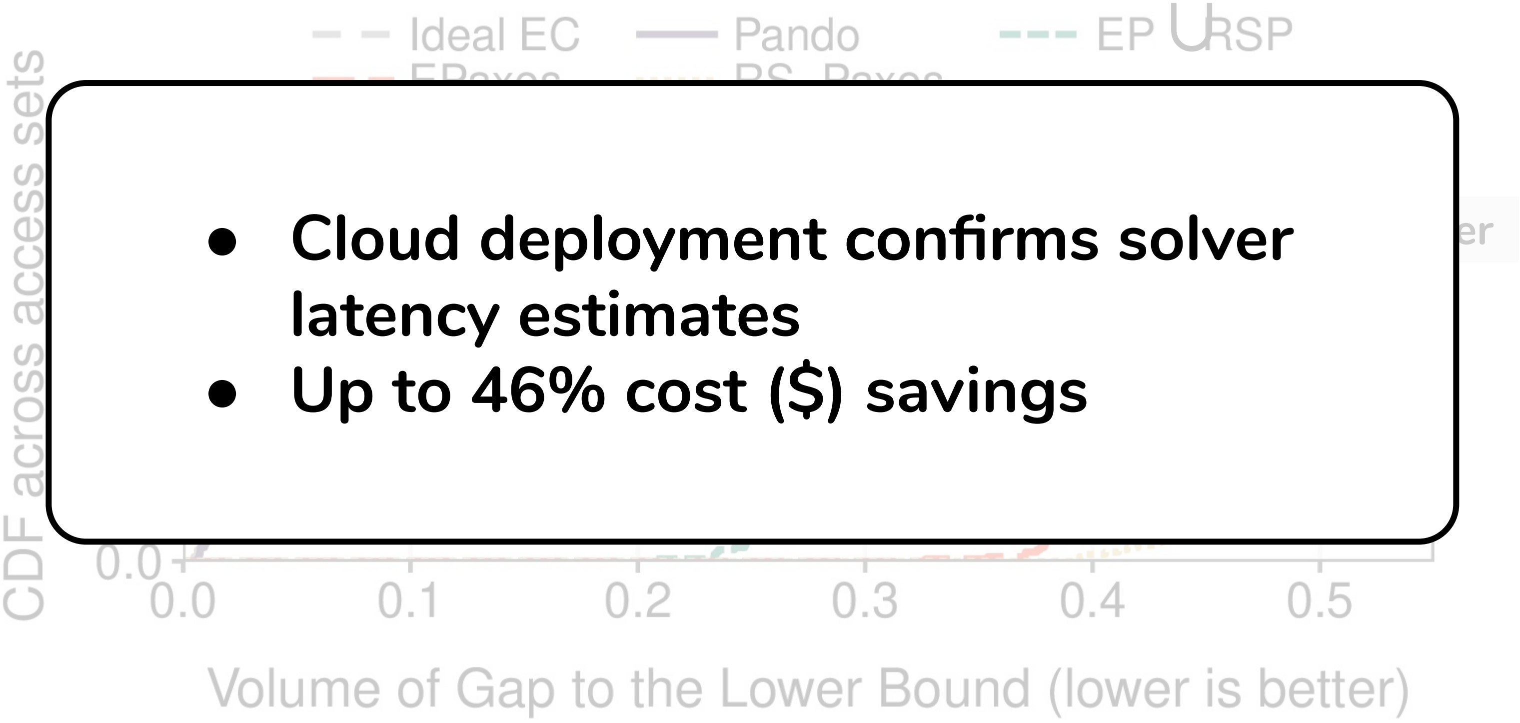
Pando is Close to the Lower Bound



Pando is Close to the Lower Bound



Pando is Close to the Lower Bound



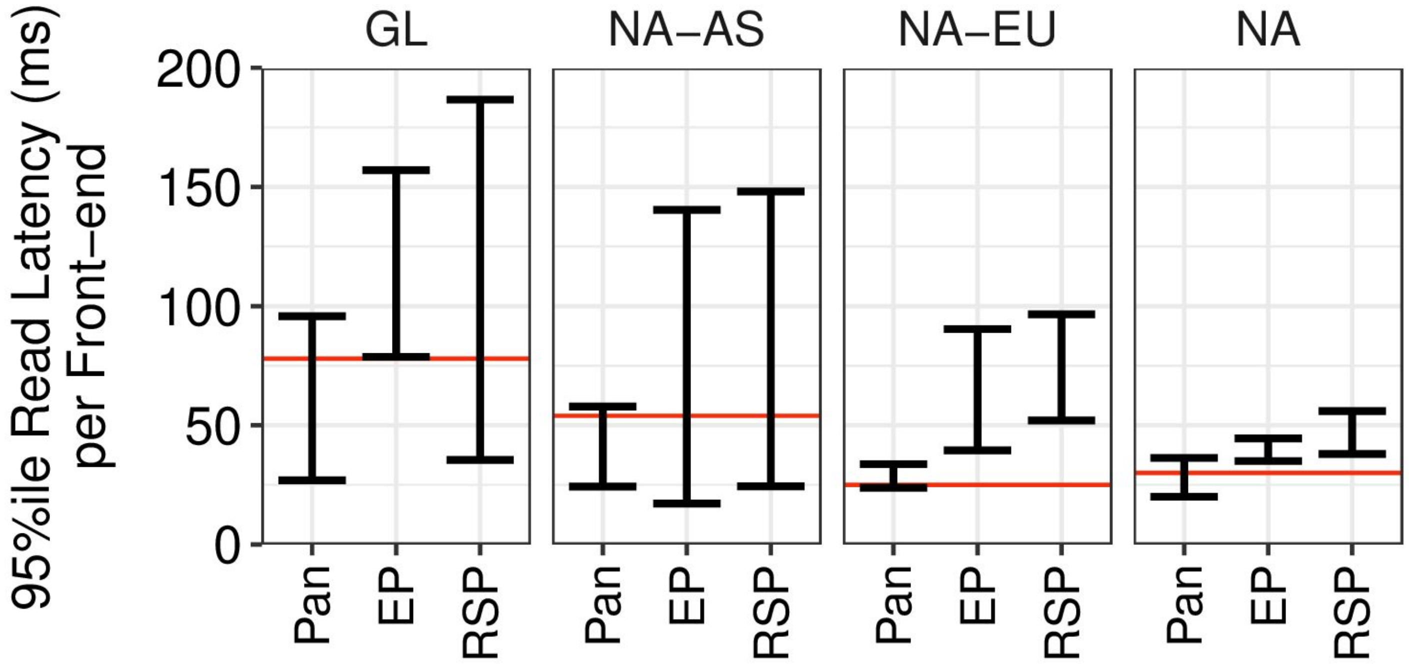
- **Cloud deployment confirms solver latency estimates**
- **Up to 46% cost ($) savings**

Conclusion

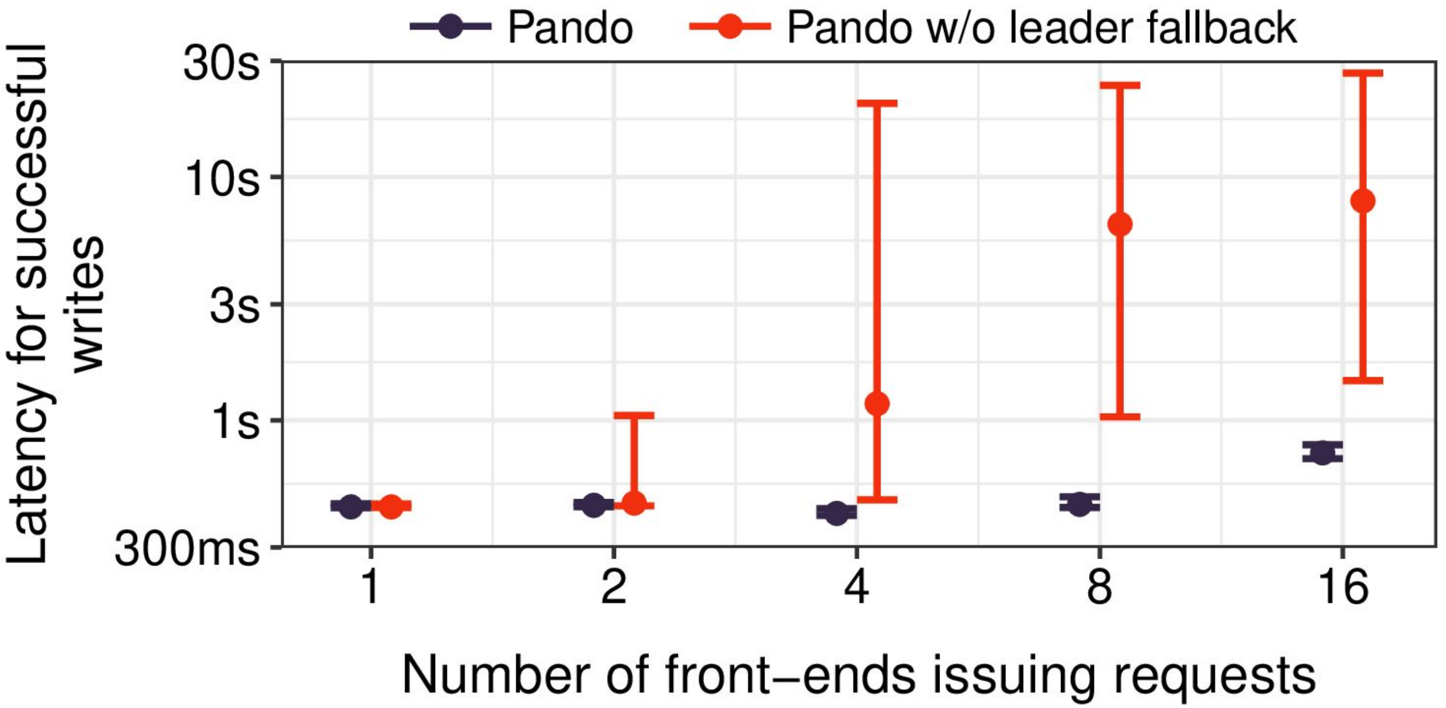
- Pando: linearizability across geo-distributed DCs
- Achieves a near-optimal read–write–storage trade-off
  - Allow for erasure-code data to minimize cost
  - Rethink how to use Paxos in the wide-area setting

Backup Slides

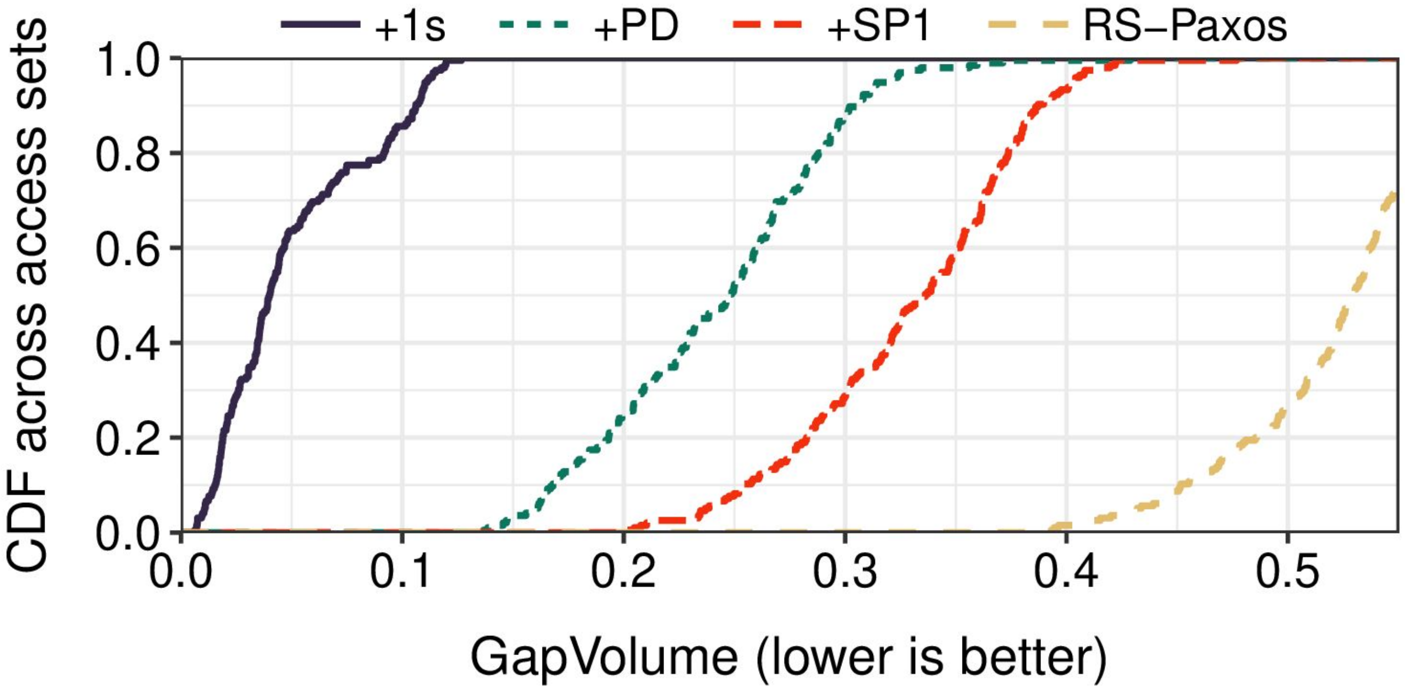
Deployment Latency



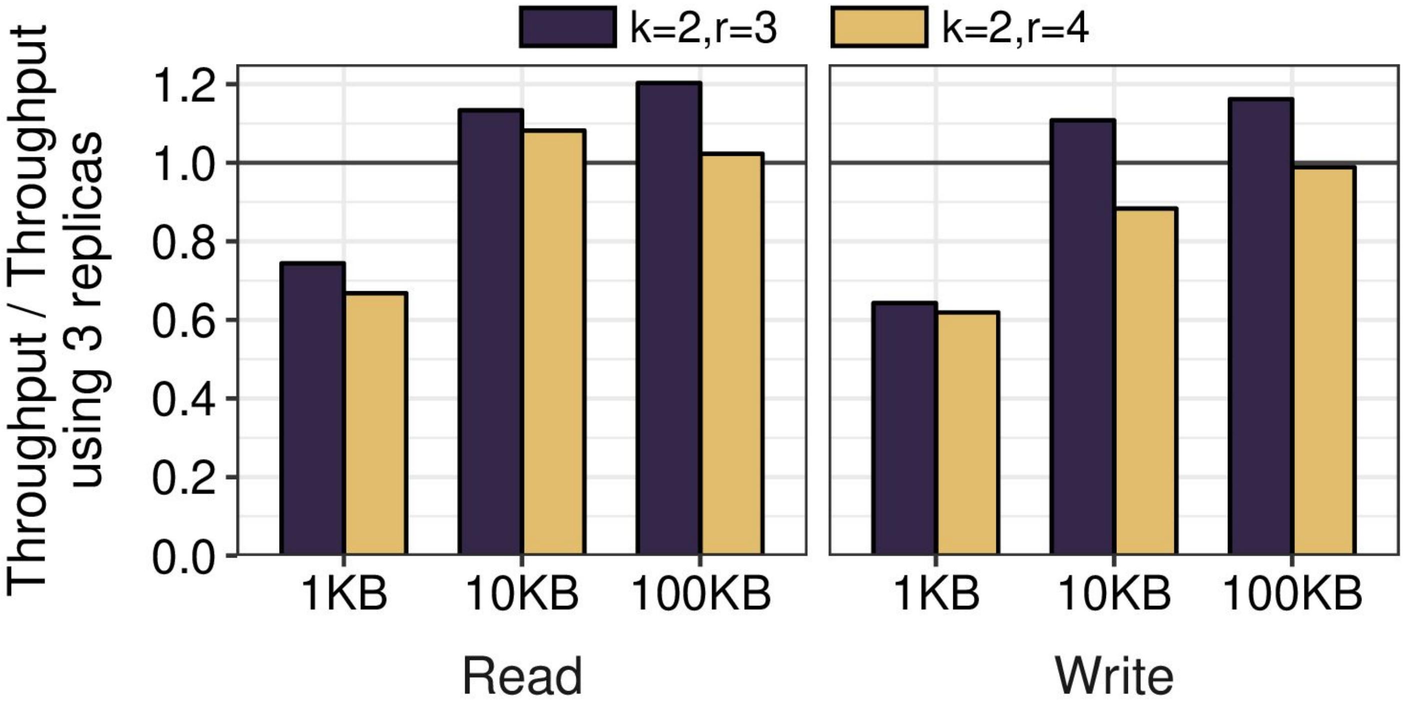
Latency Under Conflicts



Contributions of Each Technique



Throughput



k=2,r=3
k=2,r=4
Throughput / Throughput using 3 replicas
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1KB
10KB
100KB
1KB
10KB
100KB
Read
Write

Read Latency After Failure