Egalitarian Paxos

Iulian Moraru, David Andersen, Michael Kaminsky* (CMU, *Intel)

EPAXOS OVERVIEW

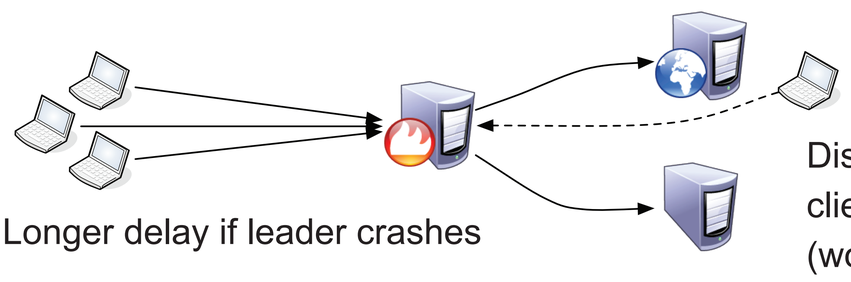
- Optimal latency in the wide-area for 1 and 2 failures
- Higher throughput in a cluster (due to load balancing)
- Constant availability if majority of replicas alive
- Better handling of slow replicas

PAXOS OVERVIEW

- State Machine Replication
  - › Fault tolerance through redundancy
  - › All replicas execute commands in same order
- Tolerates F non-Byzantine faults with 2F + 1 replicas

BOTTLENECK IN (MULTI-)PAXOS

Clients send commands to a stable leader

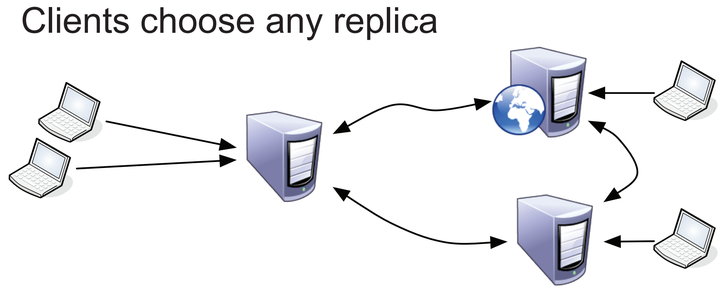


Longer delay if leader crashes

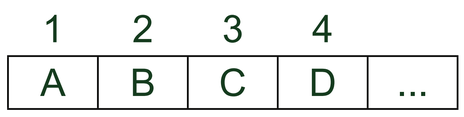
Distant clients too (worse latency)

EGALITARIAN PAXOS

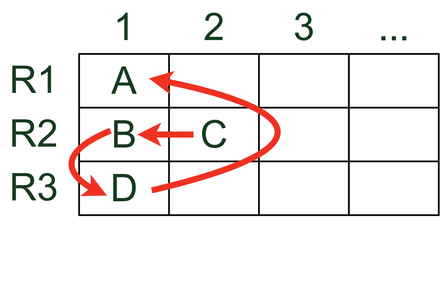
Clients choose any replica



INTUITION

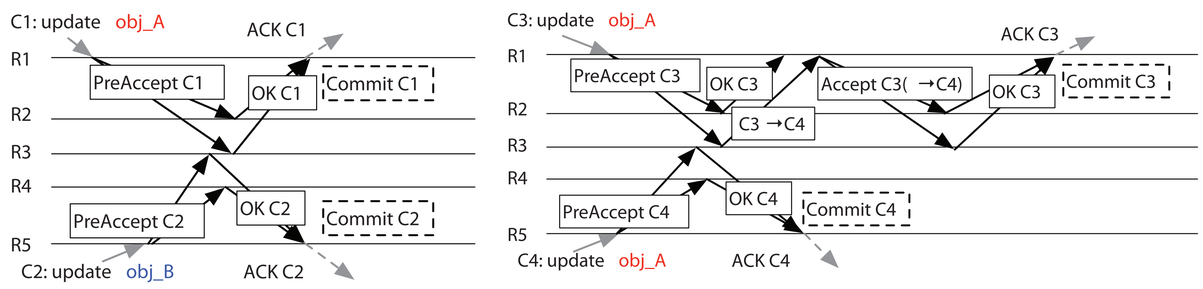


Paxos: Pre-ordered instance space



EPaxos: Instances ordered at commit time (ordering constraints)

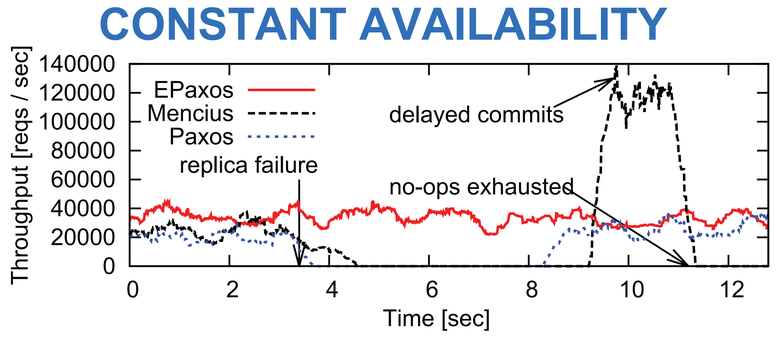
COMMIT ALGORITHM



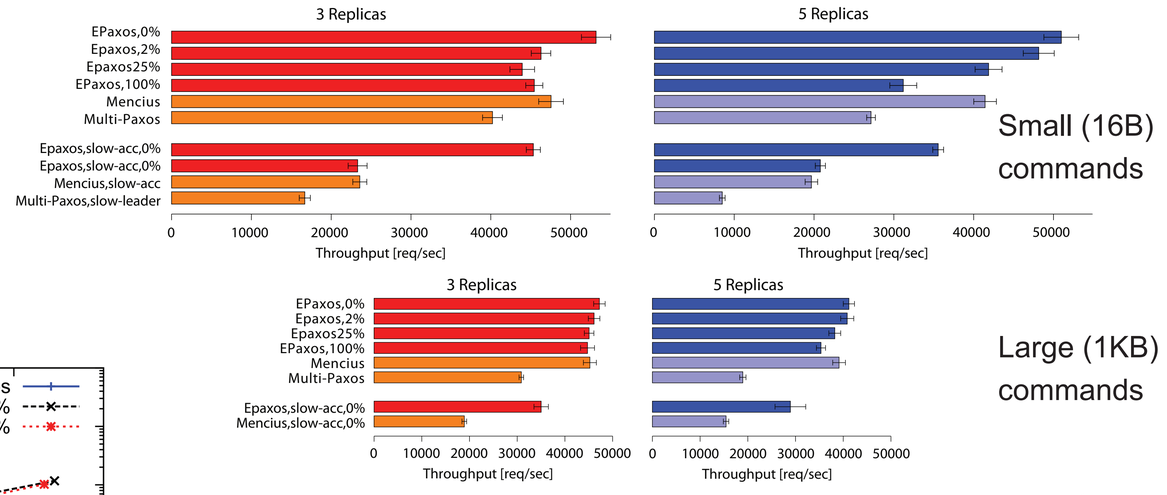
- Order only commands that interfere
- Ordering constraints: dependency list, approx. seq. #
- Fast path (1RTT) quorum: F + ⌈F/2⌉
  - › Better than Generalized Paxos by 1
- Slow path (2RTT) quorum: F + 1

EVALUATION

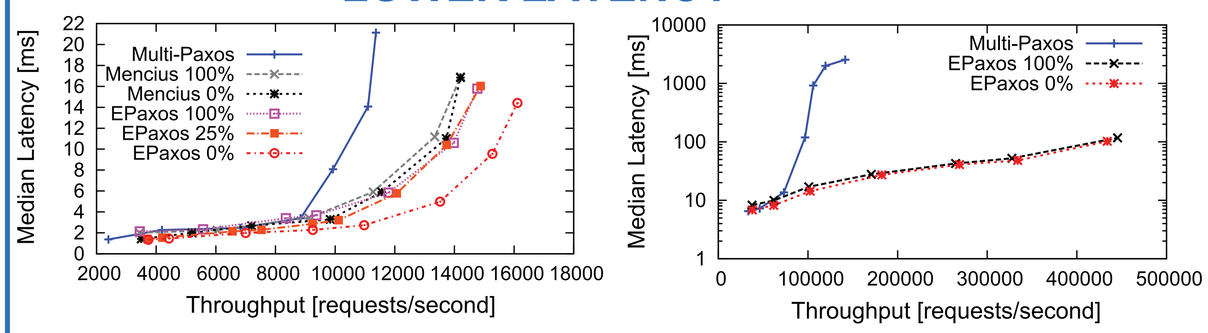
CONSTANT AVAILABILITY



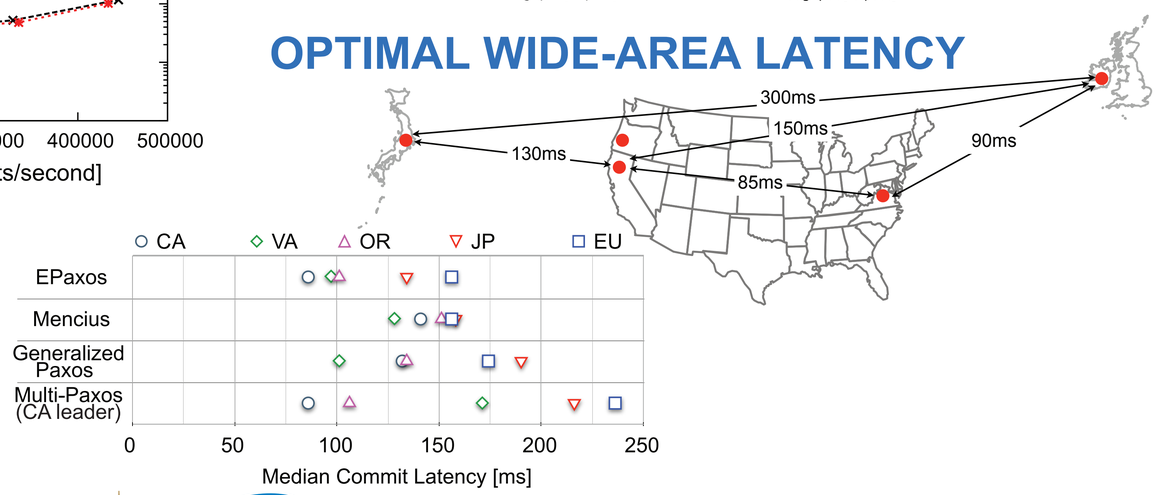
HIGHER THROUGHPUT



LOWER LATENCY



OPTIMAL WIDE-AREA LATENCY



Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley · University of Washington