EGALITARIAN PAXOS: THERE IS MORE CONSENSUS IN EGALITARIAN PARLIAMENTS

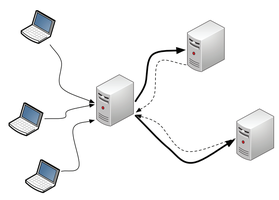
Iulian Moraru, David Andersen (Carnegie Mellon University), Michael Kaminsky (Intel)

PAXOS OVERVIEW

- State Machine Replication:
  - Fault tolerance through redundancy
  - All replicas execute the same commands in the same order
- Tolerates F failures with 2F+1 replicas
- Replicas can fail by crashing (non-Byzantine)

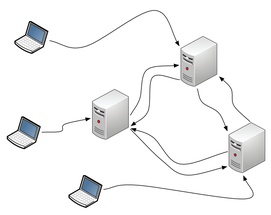
BOTTLENECK IN (MULTI-)PAXOS

- Leader brokers all communication with clients
  - Handles O(N) messages per command
- State machine unavailable until new leader is elected after a failure



EGALITARIAN PAXOS

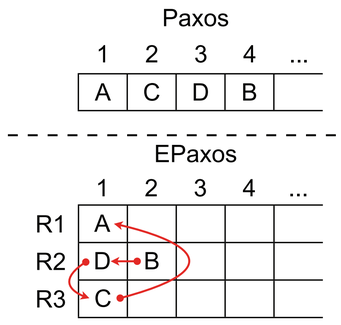
- Clients submit commands to <u>any</u> replica
  - No contention for instances
- Available without interruption if F+1 replicas are non-faulty (2F+1 replicas total)



EXECUTION ALGORITHM

- Performed independently on each replica:
  1. Wait until command C is committed
  2. Build C's dependency graph recursively
  3. Find strongly connected components (SCCs)
  4. Execute:
     - Execute SCCs in inverse topological order
     - Execute commands within each SCC in increasing sequence number order

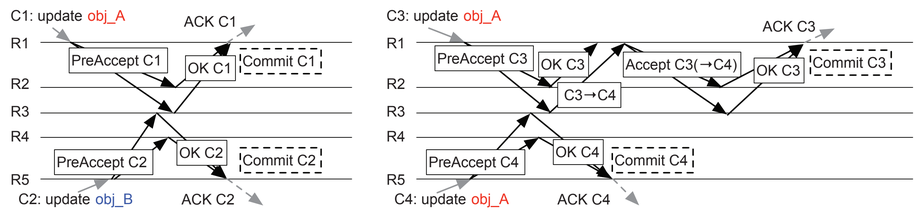
INTUITION



- Pre-ordered instance space
- Instances ordered at commit time
- Ordering attributes chosen along with commands

COMMIT ALGORITHM

- Order only commands that *interfere*
- Ordering attributes:
  1. Dependency list
  2. Approximate sequence number



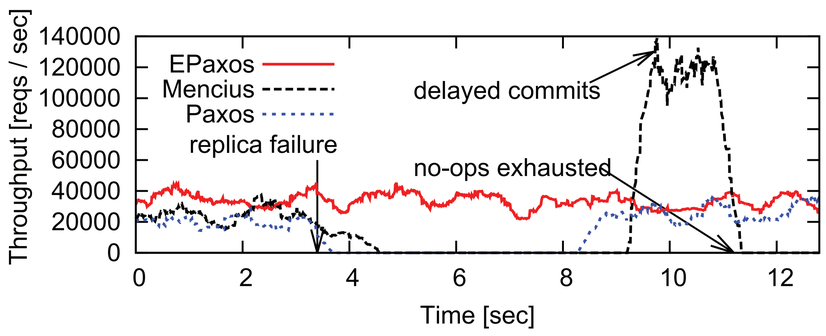
OPTIMIZED EPAXOS

- Can we commit after only one round?
  - Yes (fast path), if enough acceptors agree on the same attributes

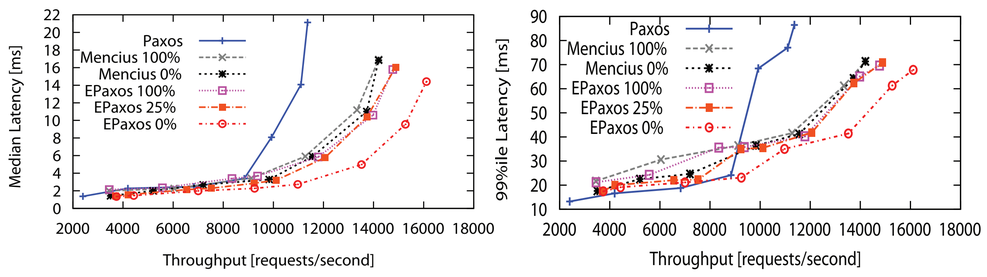
    Fast quorum size = $F + \lfloor (F+1)/2 \rfloor$
- Optimal for 3 and 5 replica setups
- Better than Fast/Generalized Paxos by 1

EVALUATION

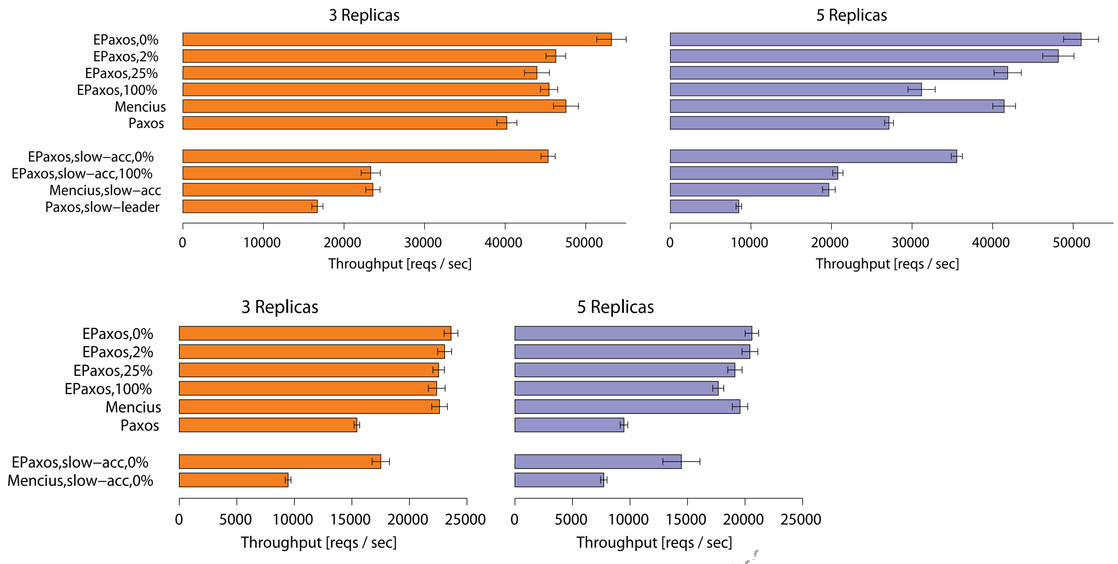
AVAILABILITY



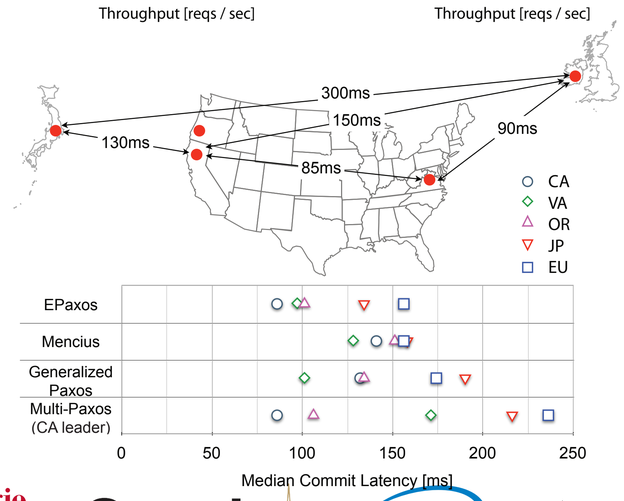
LATENCY VS. THROUGHPUT



THROUGHPUT



WIDE-AREA COMMIT LATENCY



CONCLUSIONS

- High throughput due to load balancing
- Optimal commit latency in wide area when tolerating 1 and 2 faults
- Constantly available if majority of replicas alive
- Better handling of slow replicas than previous Paxos versions

Intel Science & Technology Center for Cloud Computing

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