On Exploiting Locality for Generalized Consensus

Sebastiano Peluso, Alexandru Turcu, Roberto Palmieri and Binoy Ravindran
ECE Department
Virginia Tech
{peluso,talex,robertop,binoy}@vt.edu

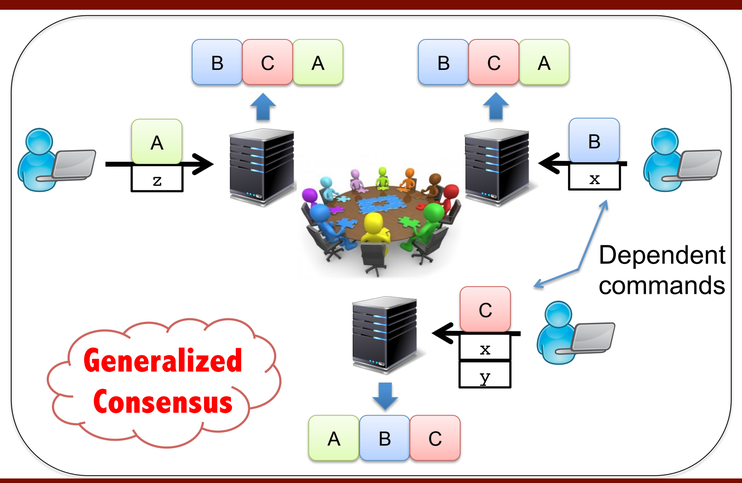
VirginiaTech — Invent the Future

Hyflow — Systems Software Research Group
http://www.hyflow.org

The Problem of Generalized Consensus

Generalized Consensus

- Proposers submit commands
- Acceptors agree on accepting equivalent sequences of commands



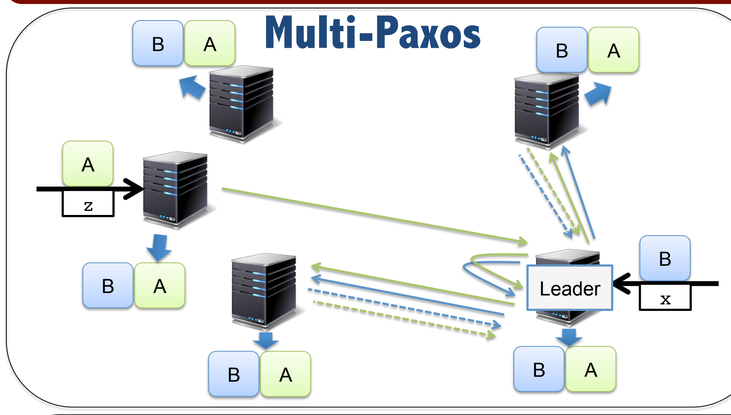
Our Challenge

Jointly ensuring the following features:

- Avoiding a designated leader
- Accepting commands in 2 communication delays (with high probability)
- Relying on the minimal quorum size of $\left\lfloor \frac{N}{2} \right\rfloor + 1$, where the maximum number of faulty nodes is $\left\lceil \frac{N}{2} \right\rceil$
- Avoiding the exchange of command dependencies

From 3 to 2 Communications Delays - Single vs. Multiple Leaders

Multi-Paxos



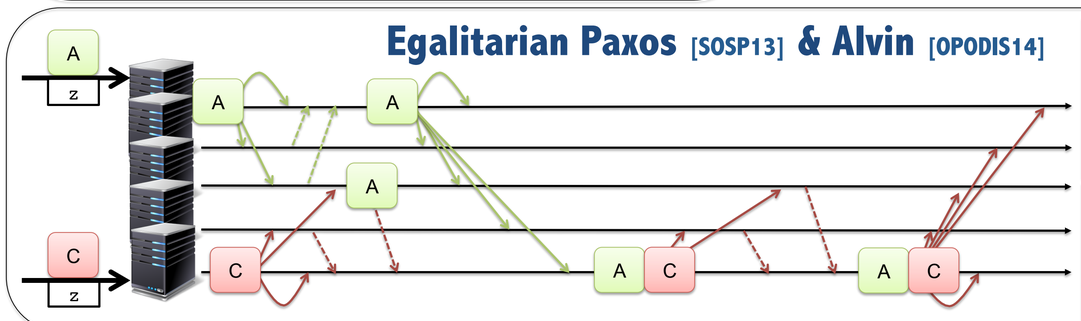
Pros
- Quorum size: $\left\lfloor \frac{N}{2} \right\rfloor + 1$
- No exchange of dependencies

Cons
- 3 communication delays
- Single leader as a bottleneck
- Do not exploit commutativity

Generalized Paxos

- In fast rounds commands are accepted in 2 communication delays in case of no concurrent and conflicting commands:
  - Proposers can bypass the leader
- A Classic Paxos round is needed if the fast round fails:
  - The single leader has to recover from failure
- Bigger quorums are required to allow decisions in 2 communication delays

Egalitarian Paxos [SOSP13] & Alvin [OPODIS14]



Pros
- 2 communication delays if no conflicts
- Exploit commutativity
- Multiple Leaders

Cons
- Exchange of dependencies
- Quorum size: $\left\lceil \frac{3}{4}N - 1 \right\rceil$

Pros
- 2 communication delays if no conflict
- Exploit commutativity

Cons
- Single leader is still a bottleneck in case of conflicts
- Quorum size: $\left\lceil \frac{3}{4}N \right\rceil$

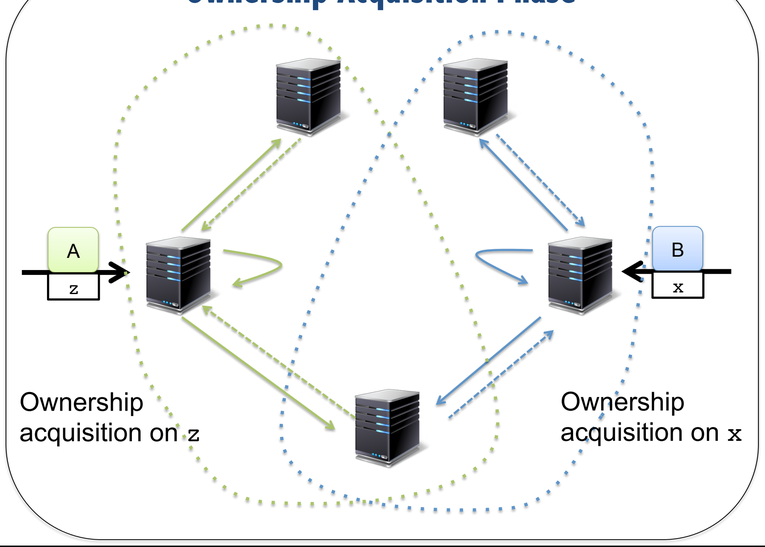
idea

Generalized Consensus is worth in case of locality, i.e., low inter-node contention.
If so a node could autonomously decide for its own commands most of the time!

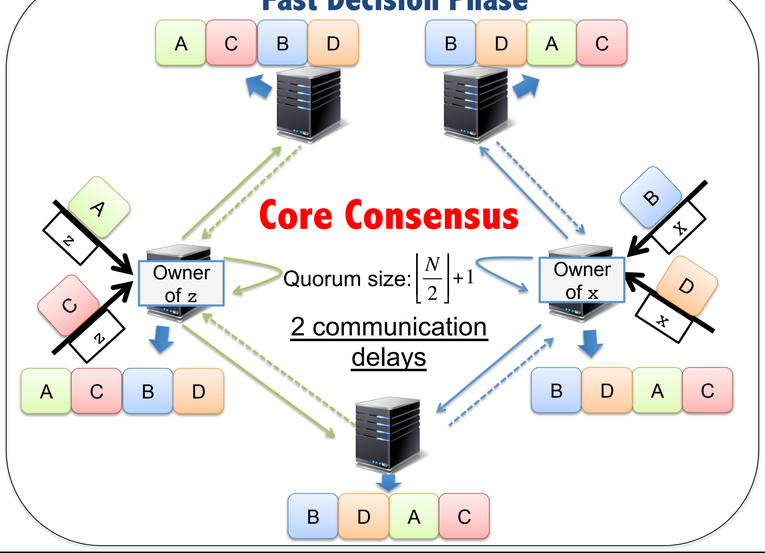
idea

Our Contribution: M²Paxos

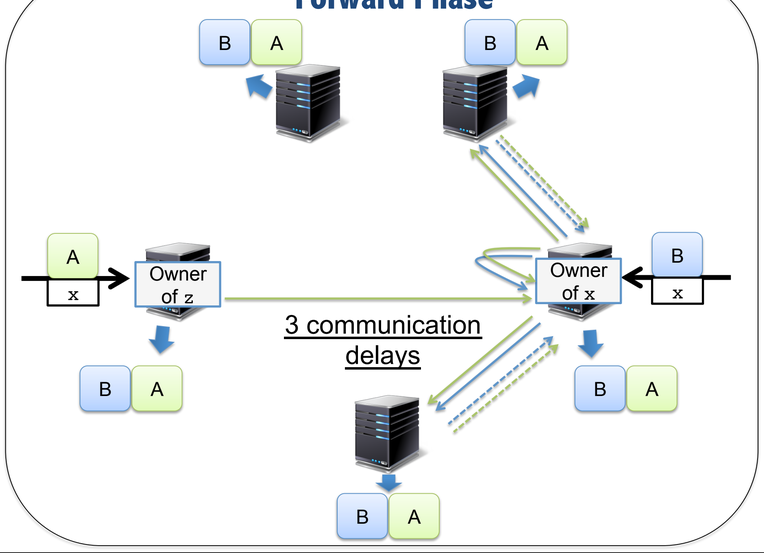
Ownership Acquisition Phase



Fast Decision Phase



Forward Phase

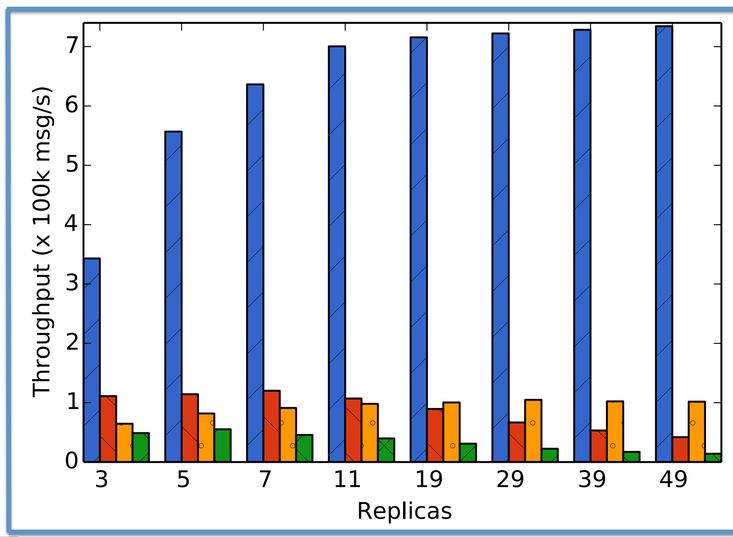


Executed with low probability in case of low inter-node conflict

Commands on z (resp. x) are ordered by the owner of z (resp. x)

In case of 1 requested owner, no ownership acquisition is needed.

Preliminary Results



Evaluation under the most favorable conditions

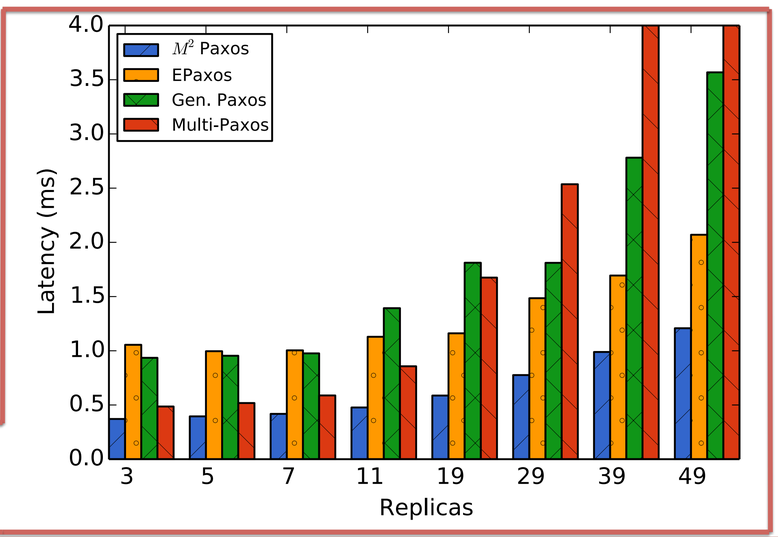
100% locality: a command proposed by a replica can only conflict with commands proposed by the same replica.

Platform

- Up to 49 replicas on Amazon EC2
- c3.4xlarge replica: Intel Xeon 2.8 GHz, 16 cores, 30GB RAM
- Network bandwidth: 7900 Mbps

Maximum attainable throughput varying the number of replicas.

Median latency without batching network messages.



ACKNOWLEDGEMENTS: This work is supported in part by US National Science Foundation under grant CNS-1217385