WPaxos: A Multileader WAN Paxos

Murat Demirbas
University at Buffalo, SUNY

[2017-10-09 Mon]

Flexible quorums

Revisiting the Paxos Foundations: A Look at Summer Internship Work at VMware Research. Heidi Howard, Dahlia Malkhi, Sasha Spiegelman. 2017.

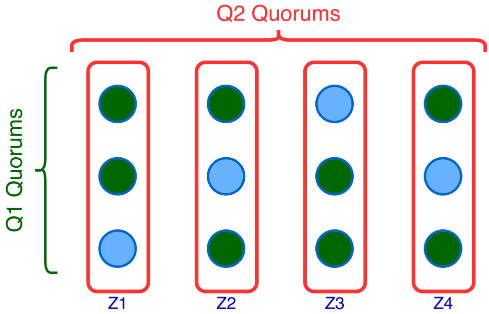
In Paxos, we can weaken "all quorums should intersect" to "only quorums from different phases should intersect".

- Majority quorums are not needed, if phase-1 quorums (Q1) intersect with phase-2 quorums (Q2)
- This allows trading off Q1 and Q2 sizes to improve performance, e.g., $Q1=10$, $Q2=3$, $N=12$
- This can be applied in a grid layout. Rows and columns act as Q1 and Q2. E.g., $Q1=4 + Q2=3 < N=12$

WPaxos: Layout the grid across WAN

In WPaxos quorums, each column coincides in a zone.

- Q1 quorums span across all the zones
- Q2 quorums map to a column=zone, making phase-2 of the protocol operate in LAN without a need for WAN communication.

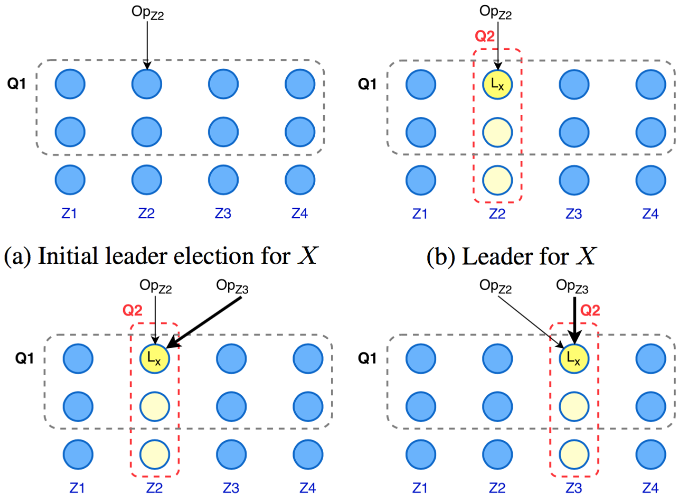


WPaxos is multileader

- Each node may act as a leader for a subset of objects
- Each object gets its own commit log with separate ballot and slot numbers, allowing for per-object linearizability
- Per object ballot&slot instead of per leader ballot&slot

- The nodes may *steal* ownership/leadership of objects from each other using phase-1 of Paxos executed over Q1
- Then the node commits the updates to the objects over the corresponding Q2, and can execute phase-2 multiple times until another node steals the object

WPaxos operation



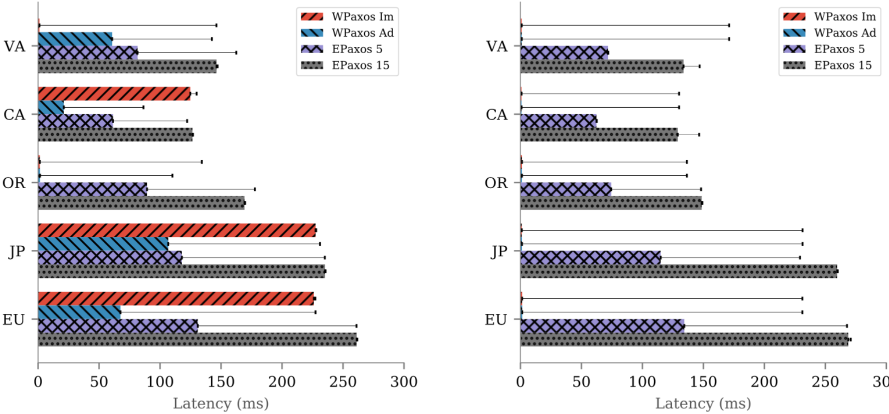
(a) Initial leader election for X

(b) Leader for X

(c) Heavy cross-region traffic

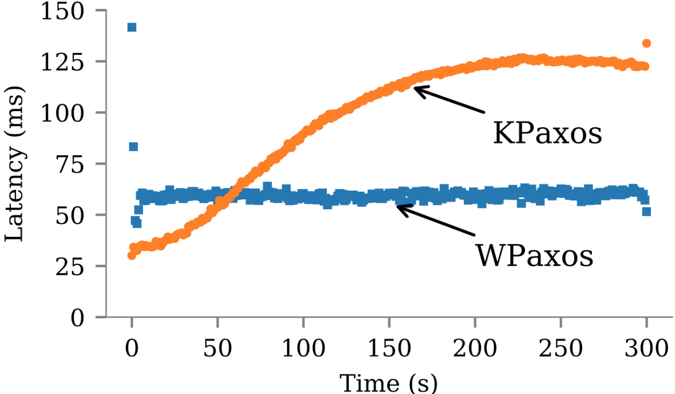
(d) Object is migrated

Median and 95th percentile for 70% & 90% locality workloads



Shifting locality workload

Statically key-partitioned Paxos (KPaxos) starts in the optimal state with most of the requests done on the local objects. The access locality is gradually shifted by changing the mean of the locality distributions at a rate of 2 objects/sec.



Questions?

WPaxos is available at `http://github.com/ailidani/paxi`

ePaxos

In order to eliminate the single leader bottleneck, EPaxos proposes a leaderless Paxos protocol where any replica at any zone can propose and commit commands opportunistically provided the commands are non-interfering. This opportunistic commit protocol requires an agreement from a fast-quorum of roughly 3/4ths of the acceptors.
For a deployment of size $2F+1$, fast-quorum is $F+\lfloor\frac{F+1}{2}\rfloor$, which means that WAN latencies are still incurred.