Flexible Paxos

Christian Smith and Huanyu Zhao
Authors: Heidi Howard, Dahlia Malkhi, Alexander Spiegelman
Shortcomings of Paxos

- Paxos requires majority agreement in both phases
- This limits a system’s
  - Latency/throughput
  - Scalability
  - Availability
Paxos has two phases:

- **Phase 1**
  - Establish one proposer as a leader

- **Phase 2**
  - Propose and decide on a value

- Each requires agreement from a **quorum**
  - Not necessarily a majority in Flexible Paxos
Multi-Paxos

- Run phase 1 to elect a leader
- Run phase 2 one or more times
  - Decide on a sequence of values
- Run phase 1 when a new leader is needed
Flexible Paxos

- Flexible Paxos allows us to:
  - Decide with less than majority agreement
  - Tolerate $\geq \lfloor n/2 \rfloor - 1$ failures

**Flexible Paxos: Quorum Intersection Revisited**

Heidi Howard$^{1,2}$, Dahlia Malkhi$^1$, and Alexander Spiegelman$^{1,3}$

1. VMware Research, Palo Alto, US
dahliamalkhi@gmail.com
2. University of Cambridge Computing Laboratory, Cambridge, UK
heidi.howard@cl.cam.ac.uk
3. Viterbi Dept. of Electrical Engineering, Technion Haifa, 32000, Israel
sashas@tx.technion.ac.il
How Flexible Paxos Works

- All quorums intersect in Paxos
  - Hence, majority agreement
- This is not actually necessary!
- P1 and P2 quorums must intersect

\[ \forall Q_1, Q_2 : Q_1 \cap Q_2 \neq \emptyset \]
A Simple Application

- Consider a system with $n = 6$ nodes
- Phase 1 quorums: 4 nodes
- Phase 2 quorums: 3 nodes
- All phase 1 and phase 2 quorums intersect
- But phase 2 quorums require 1 less node
- Works for any even $n$
Flexible Paxos

We can replace the majority quorum system of Paxos with any quorum system that guarantees that phase 1 and phase 2 quorums intersect.
A Few Quorum Systems

- Majority quorums
- Simple quorums
- Grid quorums
- Group quorums

We can replace the majority quorum system of Paxos with any quorum system that guarantees that phase 1 and phase 2 quorums intersect.
Simple quorums

- Can choose any quorum sizes $|Q_1|$ and $|Q_2|$ such that $|Q_1| + |Q_2| > n$
- General strategy:
  - Make $|Q_2|$ smaller and increase $|Q_1|$ accordingly
  - Phase 2 is more frequent than phase 1 in Multi-Paxos

$n = 5$
$|Q_1| = 4$
$|Q_2| = 2$
Simple quorums example

- Improved fault tolerance, speed in phase 2
- Worse in phase 1
- Good tradeoff in Multi-Paxos

\[ |Q_1| + |Q_2| > n \]
\[ n = 9 \]
\[ |Q_1| = 7 \]
\[ |Q_2| = 3 \]
Grid quorums

- Organize nodes into a grid
- Phase 1 quorum: any full row
- Phase 2 quorum: any full column
Grid quorums

- Smaller quorums in both phases
- Variable fault tolerance

![Diagram showing grid quorums with phase 1 and phase 2 quorums marked]
Grid quorums

- Do we actually need a column for phase 2 quorums?
- Only actually need a node from each row
  - Doesn’t matter if they’re in the same column
Group quorums

- Generalization of observation on last slide
- Divide replicas into arbitrary or natural groups
- Phase 1 quorum: All in one group
- Phase 2 quorum: One from each group
Group quorums

- Like grid quorums, fault tolerance varies
Safety

- Once a value has been decided, no different value can be decided.

- Recall that every proposal has a proposal number.

- Formally:
  - $v$ is decided with proposal number $p$.
  - $v'$ is decided with proposal number $p' > p$.
  - Then $v = v'$.
Safety

- Proof is by contradiction
- Assume $v \neq v'$
- $Q_{p,2}$: phase 2 quorum of proposal $p$
- $Q_{p',1}$: phase 1 quorum of proposal $p'$
- There is some acceptor $a$ in both
  - All phase 1 and phase 2 quorums intersect
Safety

- Acceptor $a$ must receive $\text{prepare}(p')$ after $\text{propose}(p, v)$
  - (since $p' > p$)

- After enumerating through several cases:
  - Conclude the proposer of $p'$ must propose value $v$

- This contradicts our assumption that $v \neq v'$

$$a \in Q_{p,2} \cap Q_{p',1}$$
Safety: The Main Takeaway

We can prove that the safety requirement is satisfied with only the fact that phase 1 and phase 2 quorums must intersect.
Prototype

Implement a naïve FPaxos by modifying LibPaxos3

Modification over the size of Q1 and Q2
Performance of FPaxos

(a) Performance of FPaxos and LibPaxos3 with 5 replicas.

(b) Performance of FPaxos and LibPaxos3 with 8 replicas.
Further More

- A quorum of replicas vs. all replicas
- FPaxos handles more failures (Q2 size 4)
- Only naïve implementation
Enhancements

All Q2s vs Q2s with smaller proposal numbers

Case: 100f replicas with Q2 f+1
Related Works

- Cheap Paxos
- Fast Paxos
- Ring-Paxos
- ...

SMR reconfiguration
Thank you!