Paxos made Live

How Google employs paxos to build a replicated log

Qing Zheng

15799 – Adv Topics in DB Systems

Distributed Consensus

Distributed Consensus

Lock Service

- Exclusive Access
- Synchronization
- ...

Distributed Consensus

Lock Service

- Exclusive Access
- Synchronization
- ...

Name Service

- Primary Copy
- Partition Table
- Leader / Master
- Membership
- Global Metadata
- ...

Chubby

- Help clients ...
  - synchronize activities
  - agree on basic information about their environment

What should Chubby Offer?

What should Chubby Offer?

- Agreement

What should Chubby Offer?

- Agreement
- High Throughput

What should Chubby Offer?

- Agreement
- ~~High Throughput~~

What should Chubby Offer?

- Agreement
- ~~High Throughput~~
- Massive Storage

What should Chubby Offer?

- Agreement
- ~~High Throughput~~
- ~~Massive Storage~~

What should Chubby Offer?

- Agreement
- ~~High Throughput~~
- ~~Massive Storage~~
- Availability

What should Chubby Offer?

- Agreement
- ~~High Throughput~~
- ~~Massive Storage~~
- Availability
- Reliable and Fault Tolerant

Why use Paxos?

Why use Paxos?

- Safety
  - bad things never happen

Why use Paxos?

- Safety
  - bad things never happen
- Liveness
  - good things eventually happen
    - as long as only 1 proposer exists eventually

Why use Paxos?

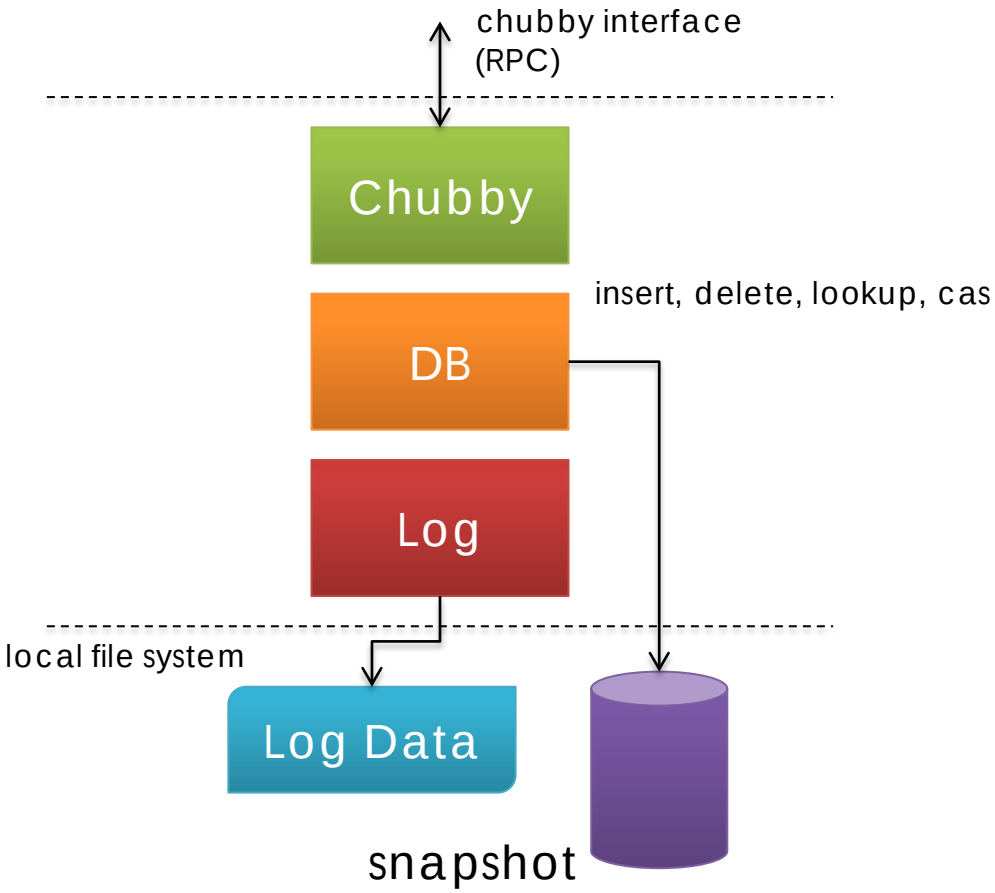
- Safety
  - bad things never happen
- Liveness
  - good things eventually happen
    - as long as only 1 proposer exists eventually
- Fault-Tolerant
  - won't block
    - as long as a majority of nodes are still live

Why use Paxos?

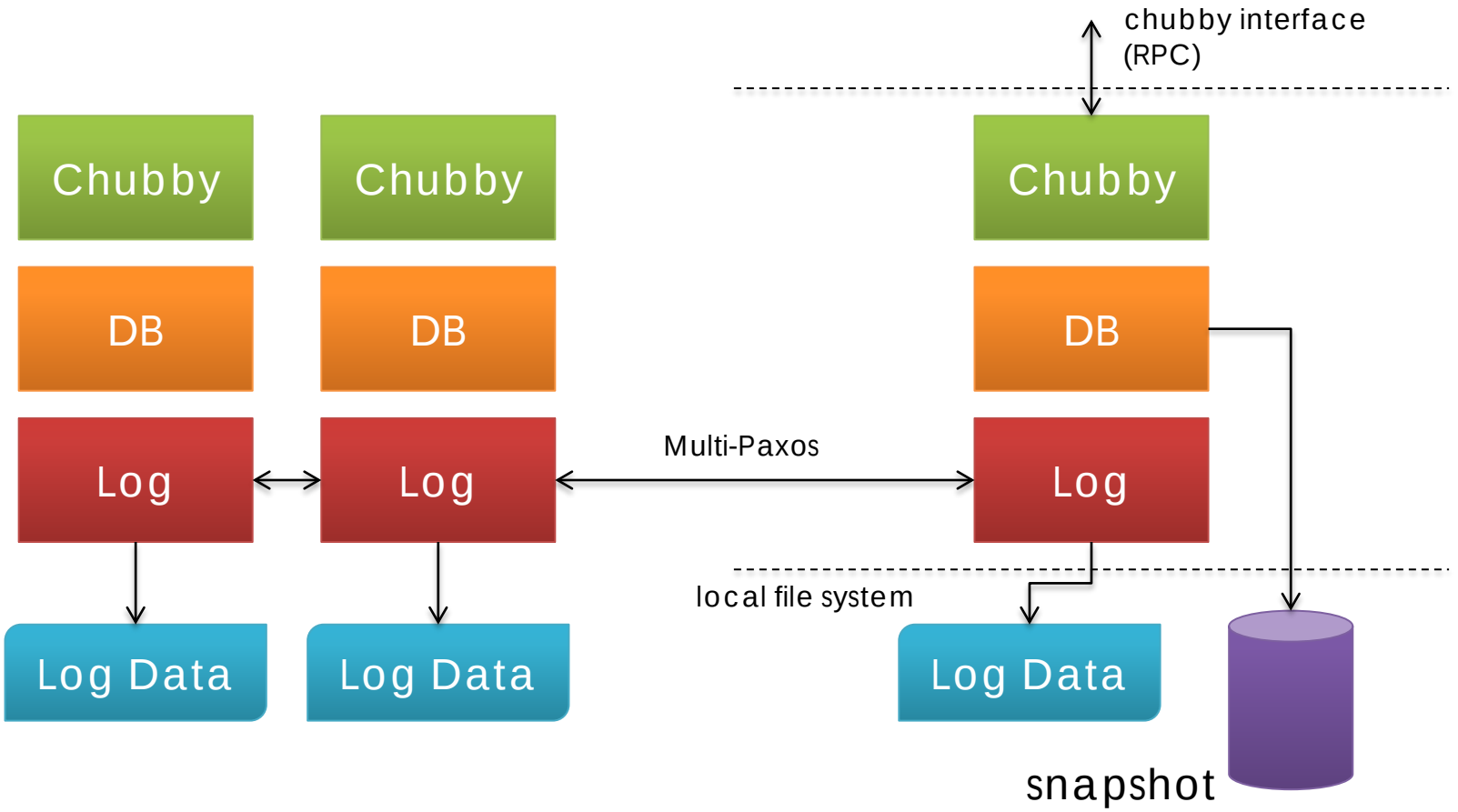
- Safety
  - bad things never happen
- Liveness
  - good things eventually happen
    - as long as only 1 proposer exists eventually
- Fault-Tolerant
  - won't block
    - as long as a majority of nodes are still live

No other choices ...

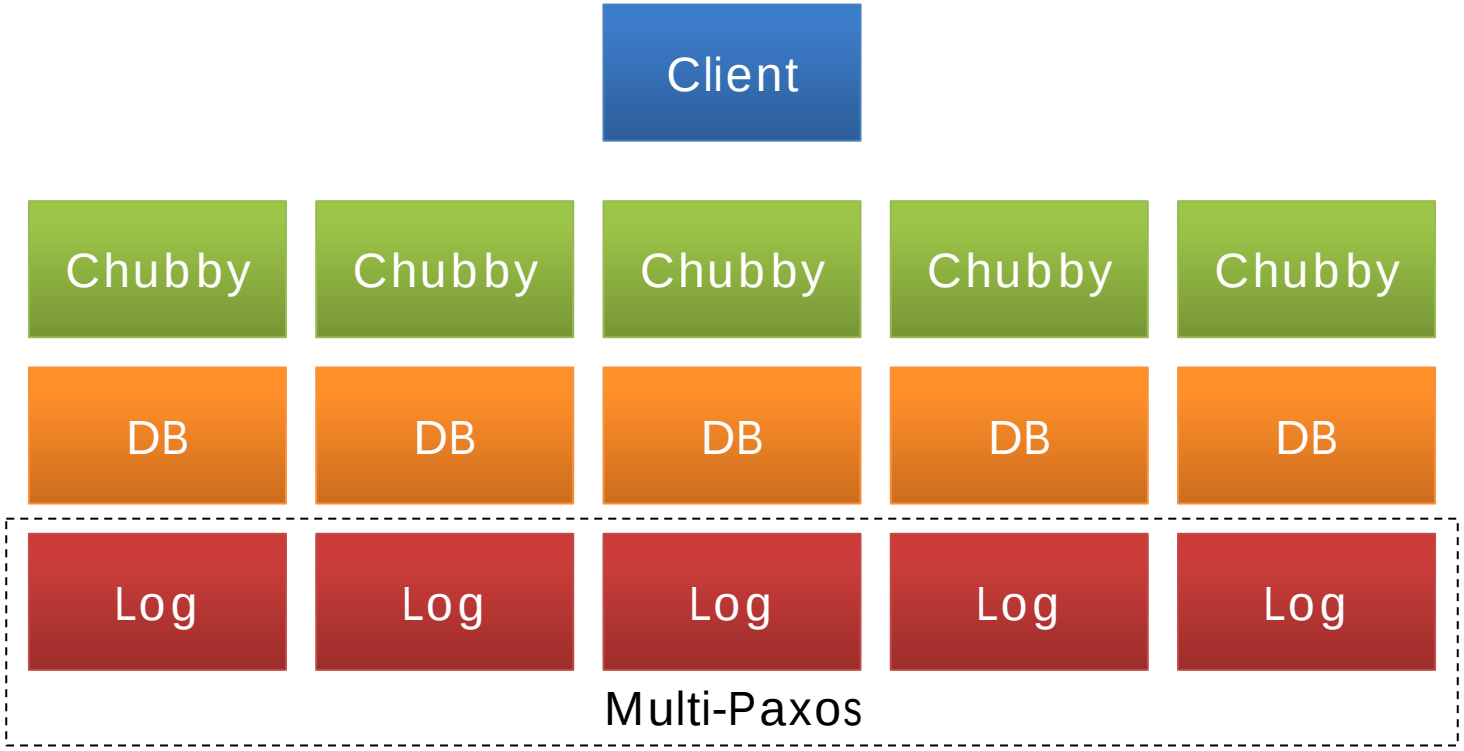
Chubby Overview



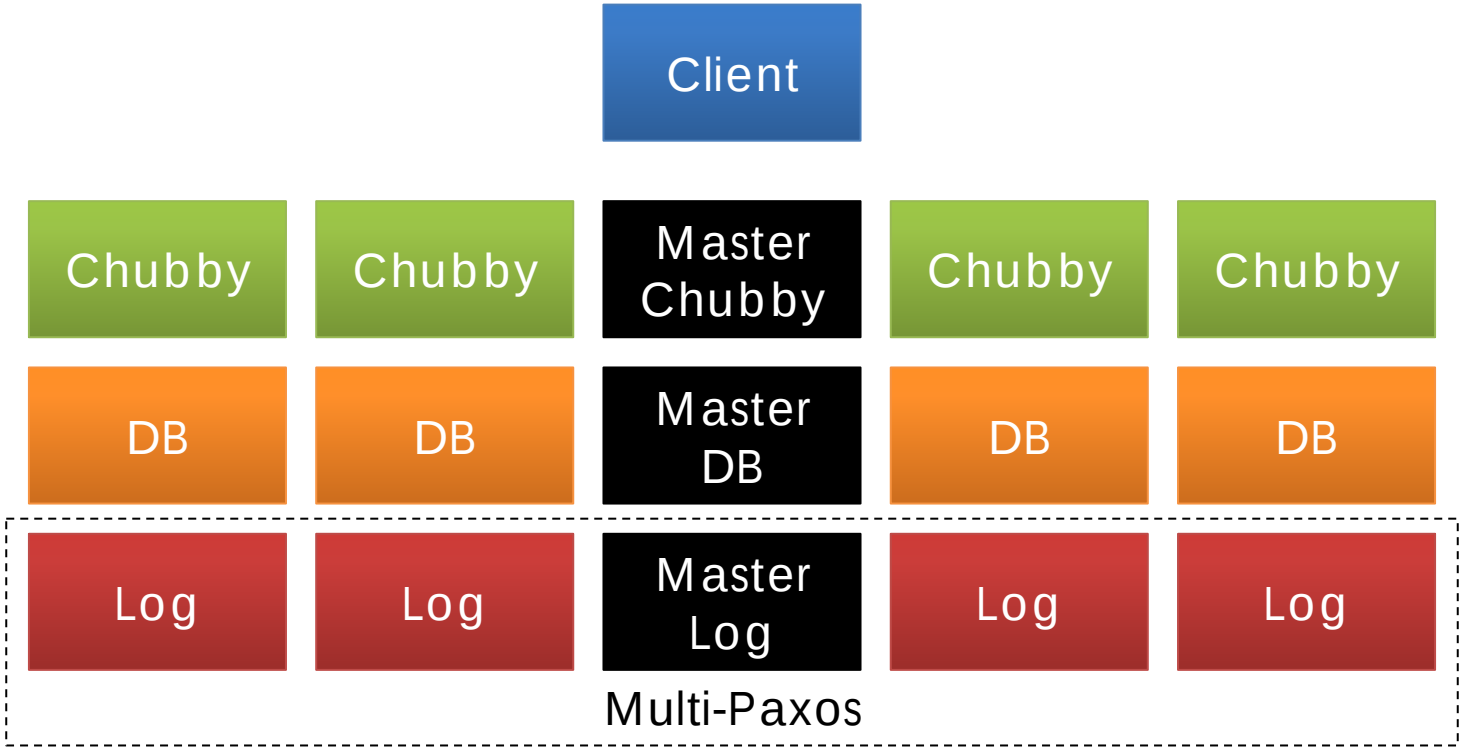
Chubby Overview



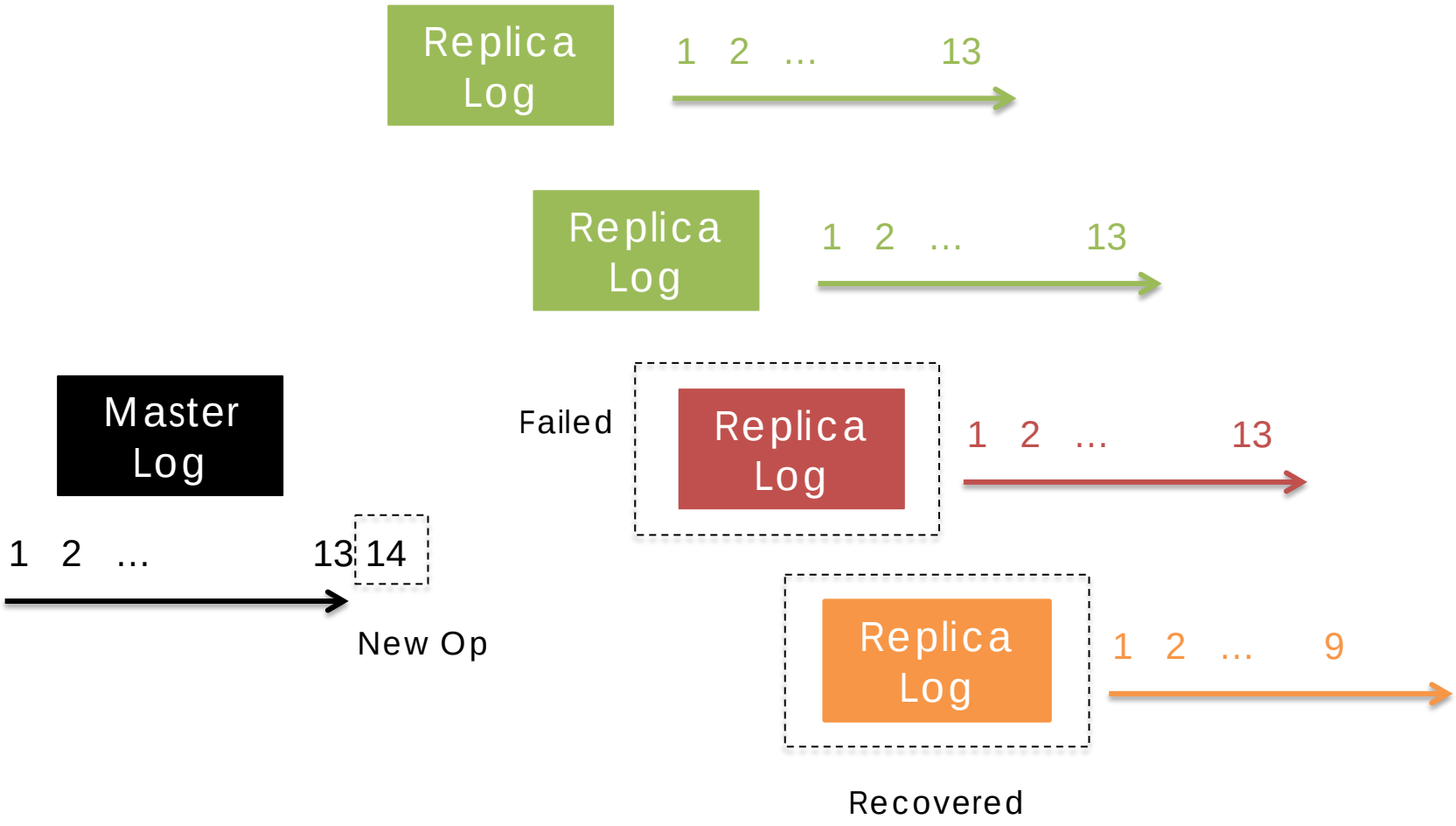
Chubby Overview



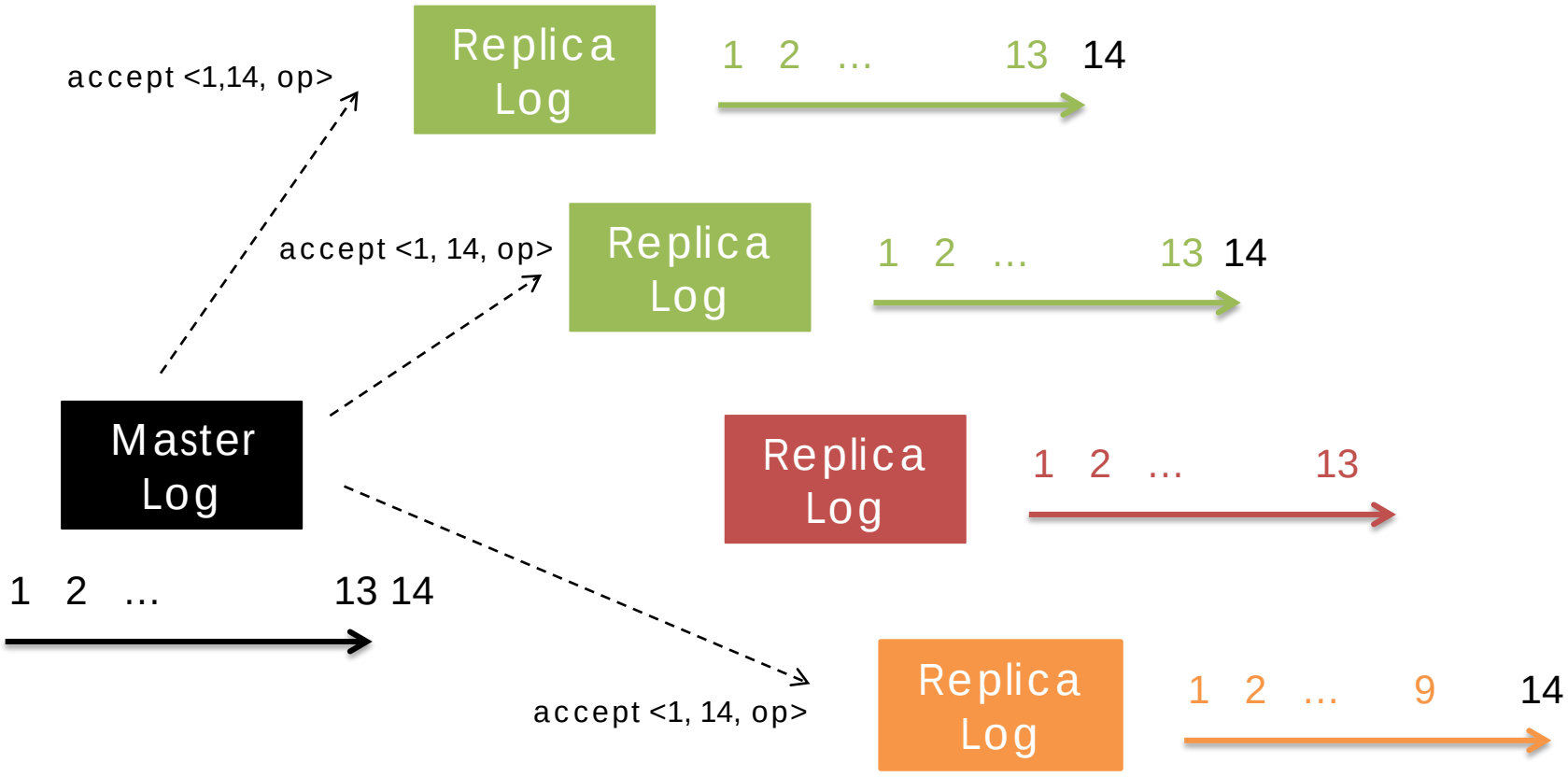
Chubby Overview



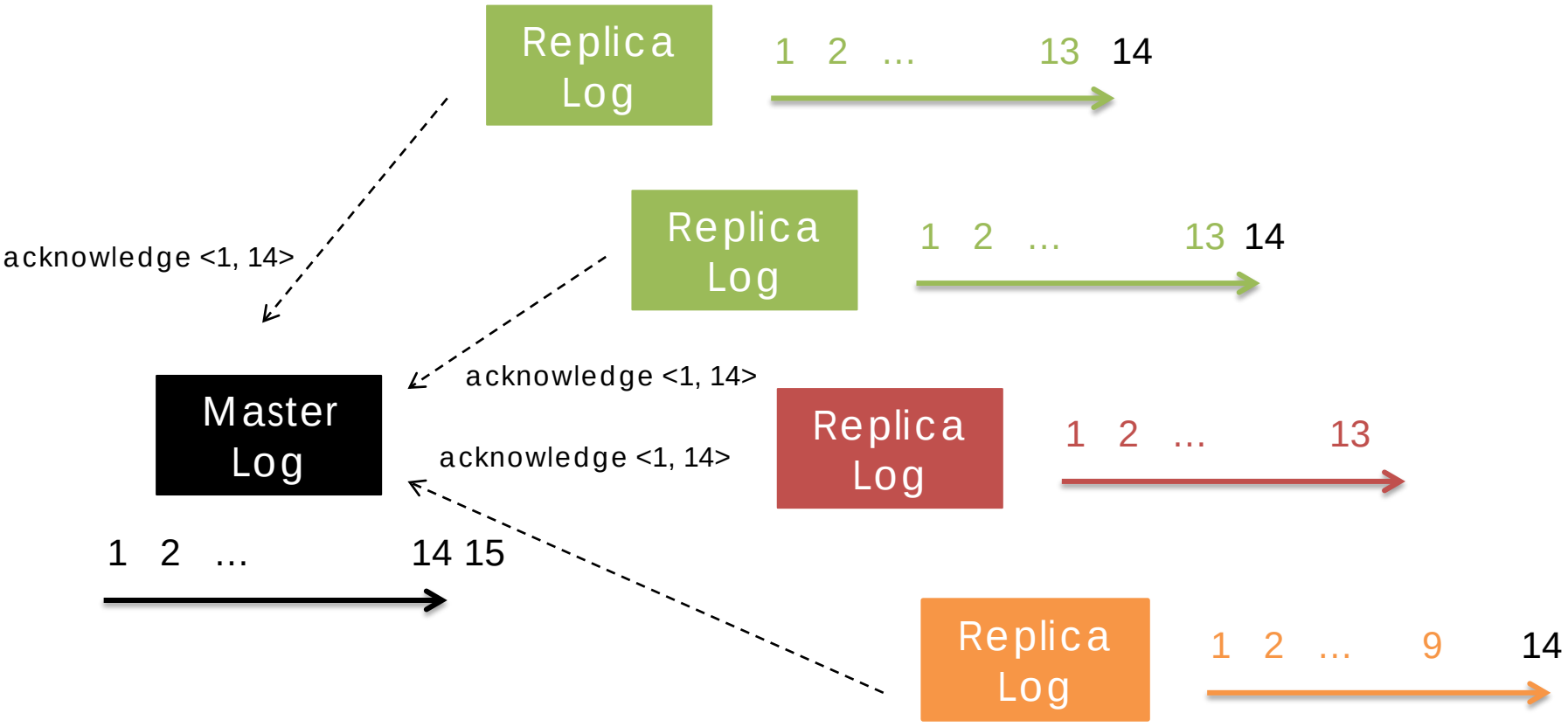
Multi-Paxos


Replica Log
1 2 ... 13
Replica Log
1 2 ... 13
Master Log
1 2 ... 13 14
New Op
Failed
Replica Log
1 2 ... 13
Replica Log
1 2 ... 9
Recovered

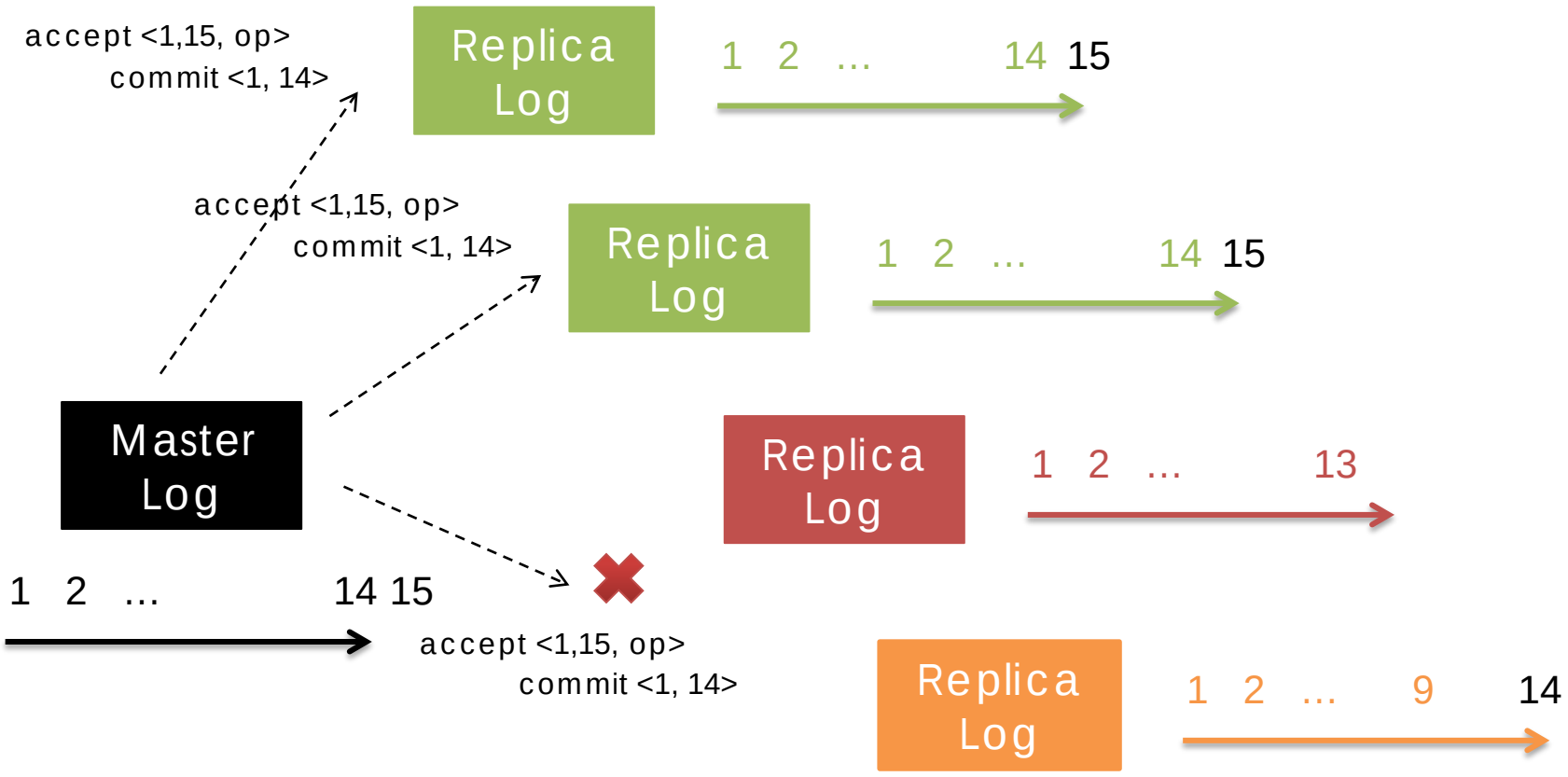
Multi-Paxos


accept <1,14, op>
Replica Log
1 2 ... 13 14
accept <1, 14, op>
Replica Log
1 2 ... 13 14
Master Log
Replica Log
1 2 ... 13
1 2 ... 13 14
accept <1, 14, op>
Replica Log
1 2 ... 9 14

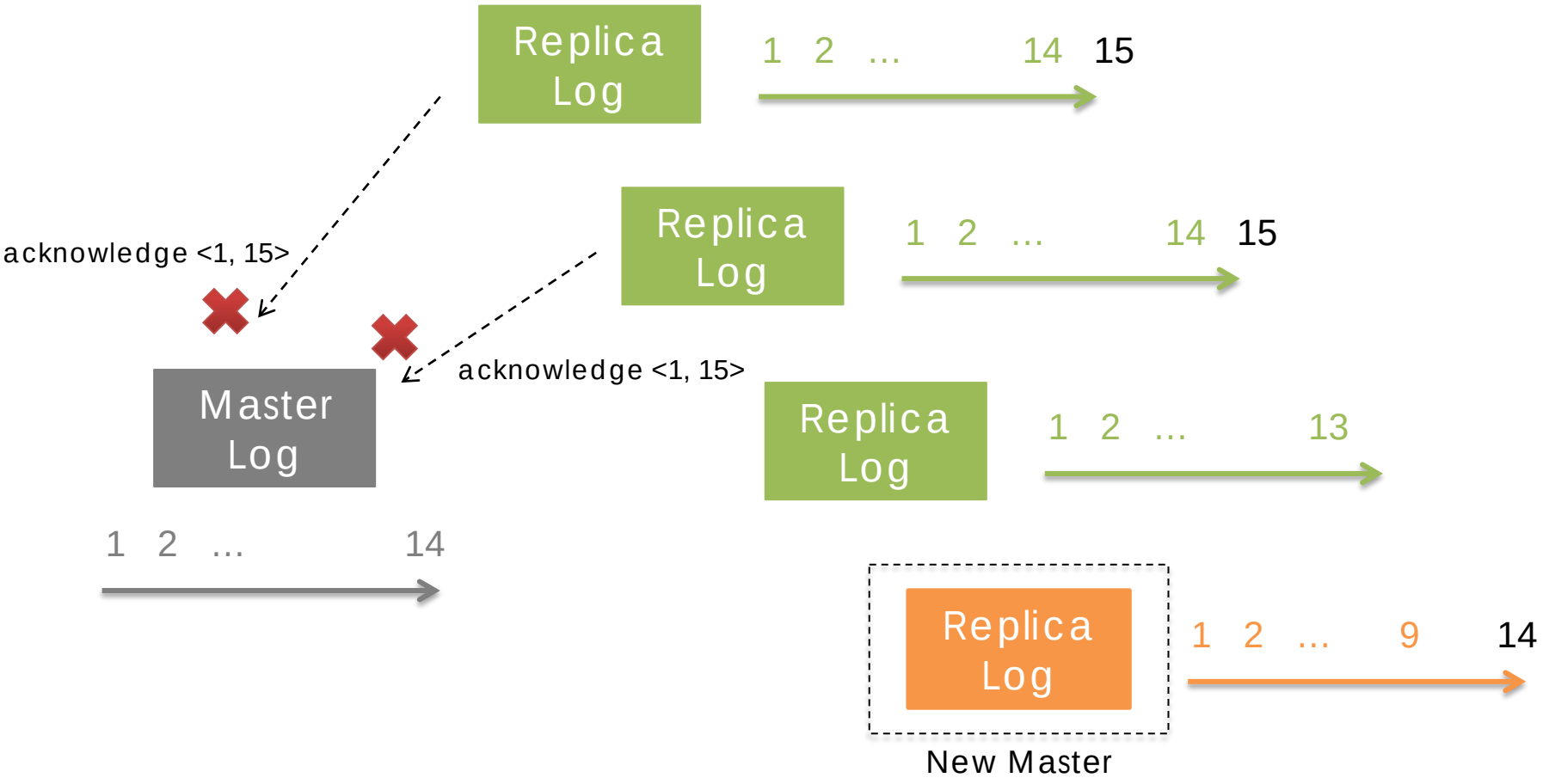
Multi-Paxos


Replica Log
1 2 ... 13 14
Replica Log
1 2 ... 13 14
acknowledge <1, 14>
Master Log
acknowledge <1, 14>
Replica Log
1 2 ... 13
acknowledge <1, 14>
1 2 ... 14 15
Replica Log
1 2 ... 9 14

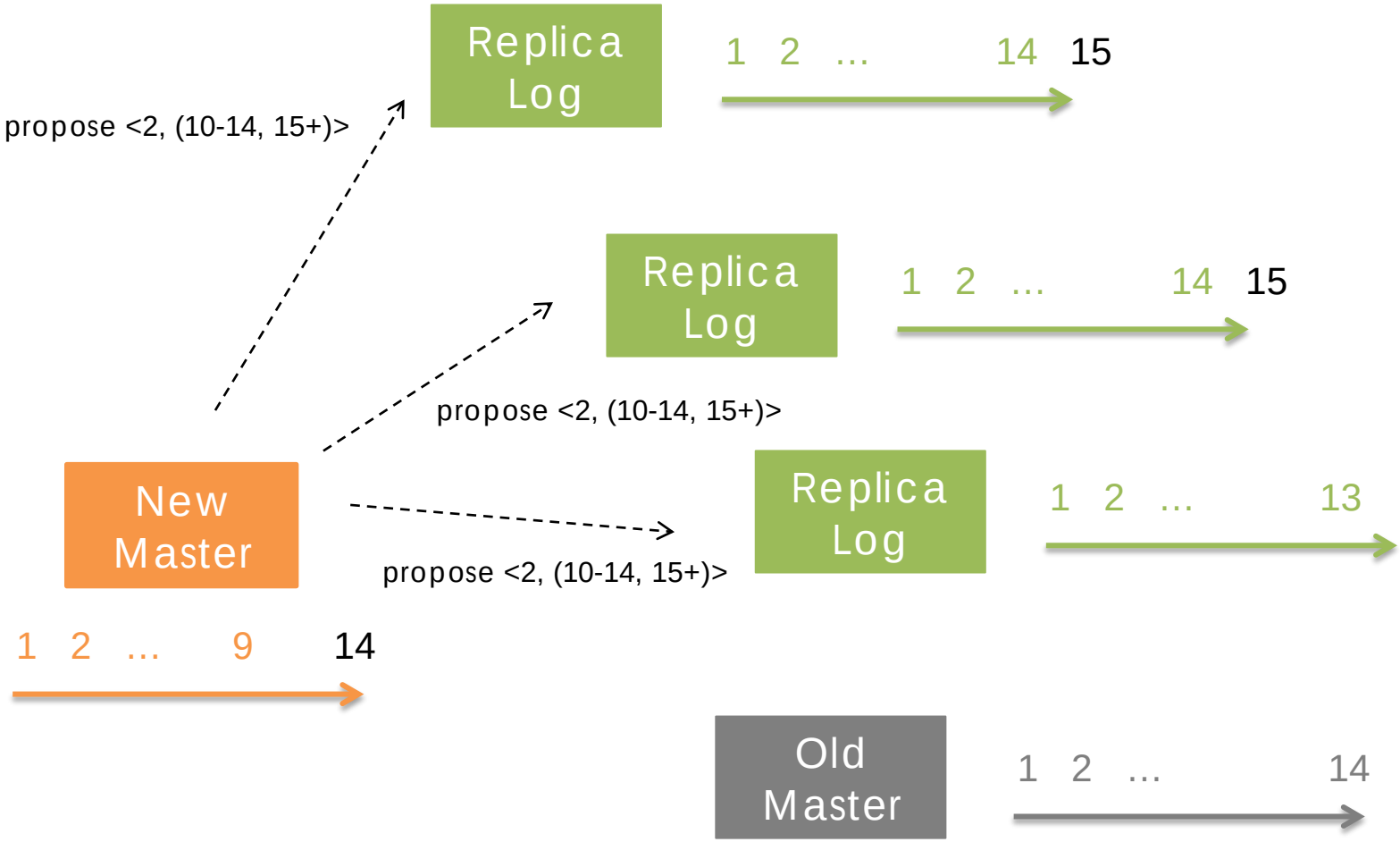
Multi-Paxos


accept <1,15, op>
commit <1, 14>
Replica Log
1 2 ... 14 15
accept <1,15, op>
commit <1, 14>
Replica Log
1 2 ... 14 15
Master Log
Replica Log
1 2 ... 13
1 2 ... 14 15
accept <1,15, op>
commit <1, 14>
Replica Log
1 2 ... 9 14

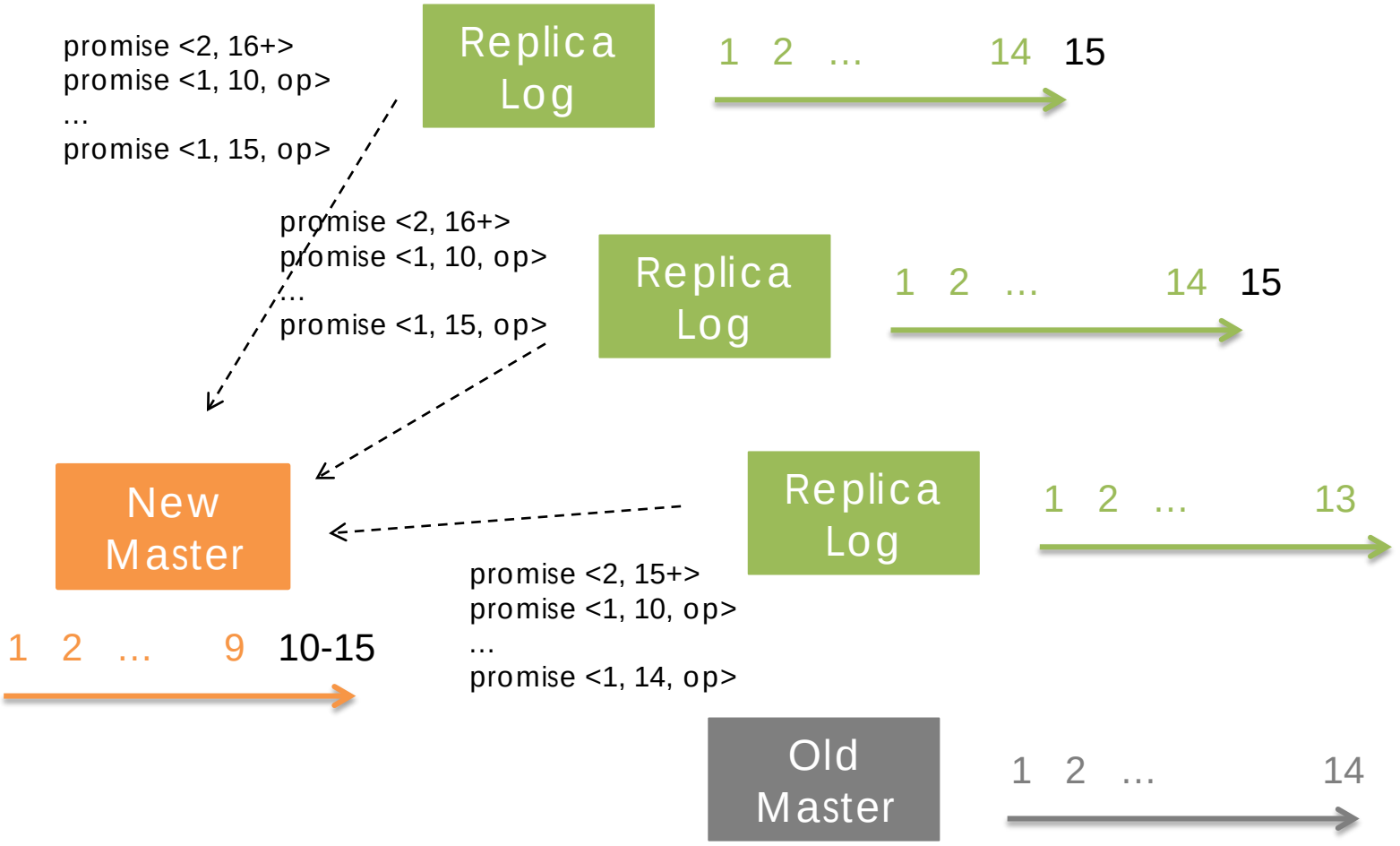
Multi-Paxos


Replica Log
1 2 ... 14 15
Replica Log
1 2 ... 14 15
acknowledge <1, 15>
acknowledge <1, 15>
Master Log
Replica Log
1 2 ... 13
1 2 ... 14
Replica Log
1 2 ... 9 14
New Master

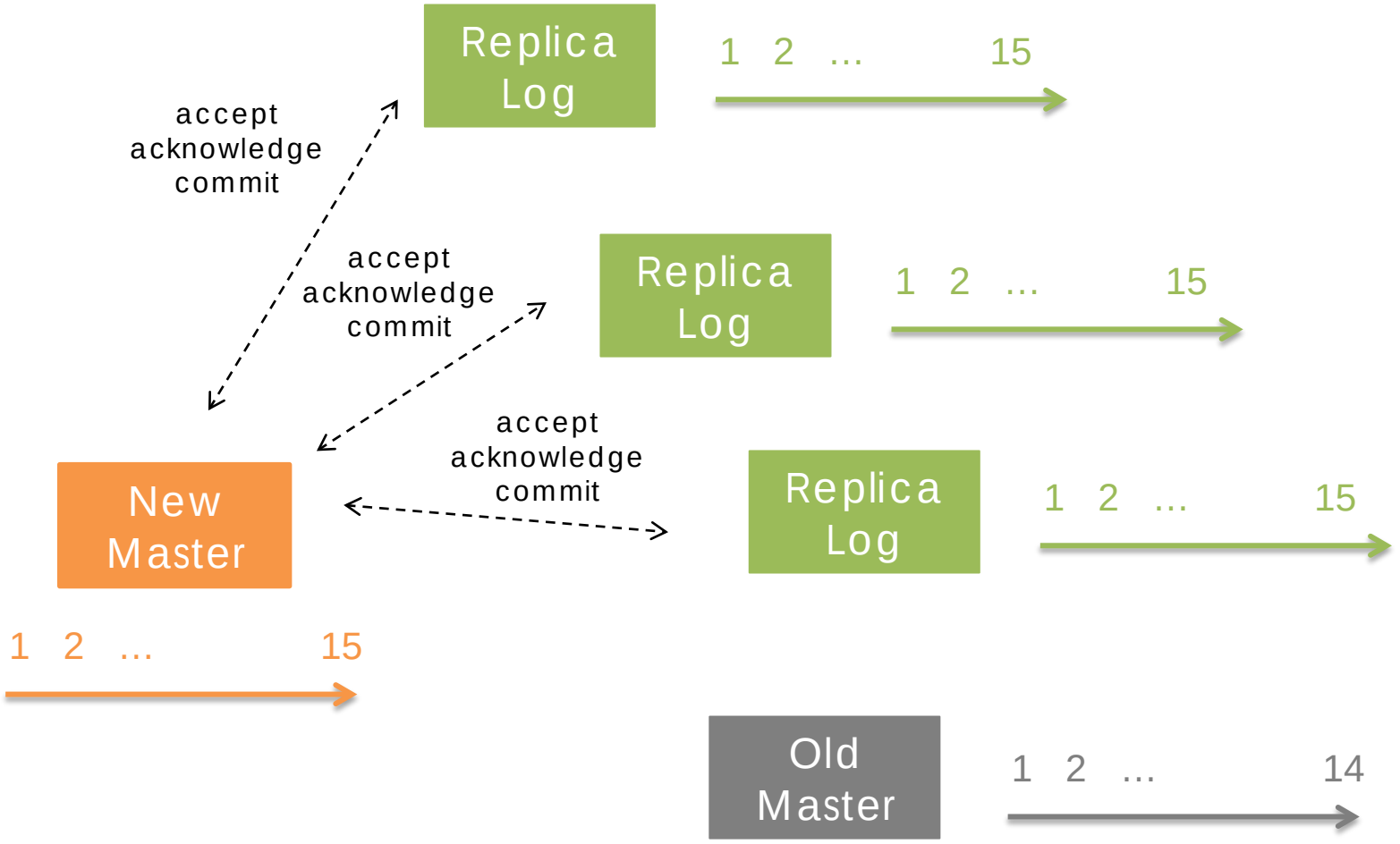
Multi-Paxos


Replica Log
1 2 ... 14 15
propose <2, (10-14, 15+)>
Replica Log
1 2 ... 14 15
propose <2, (10-14, 15+)>
New Master
propose <2, (10-14, 15+)>
Replica Log
1 2 ... 13
1 2 ... 9 14
Old Master
1 2 ... 14

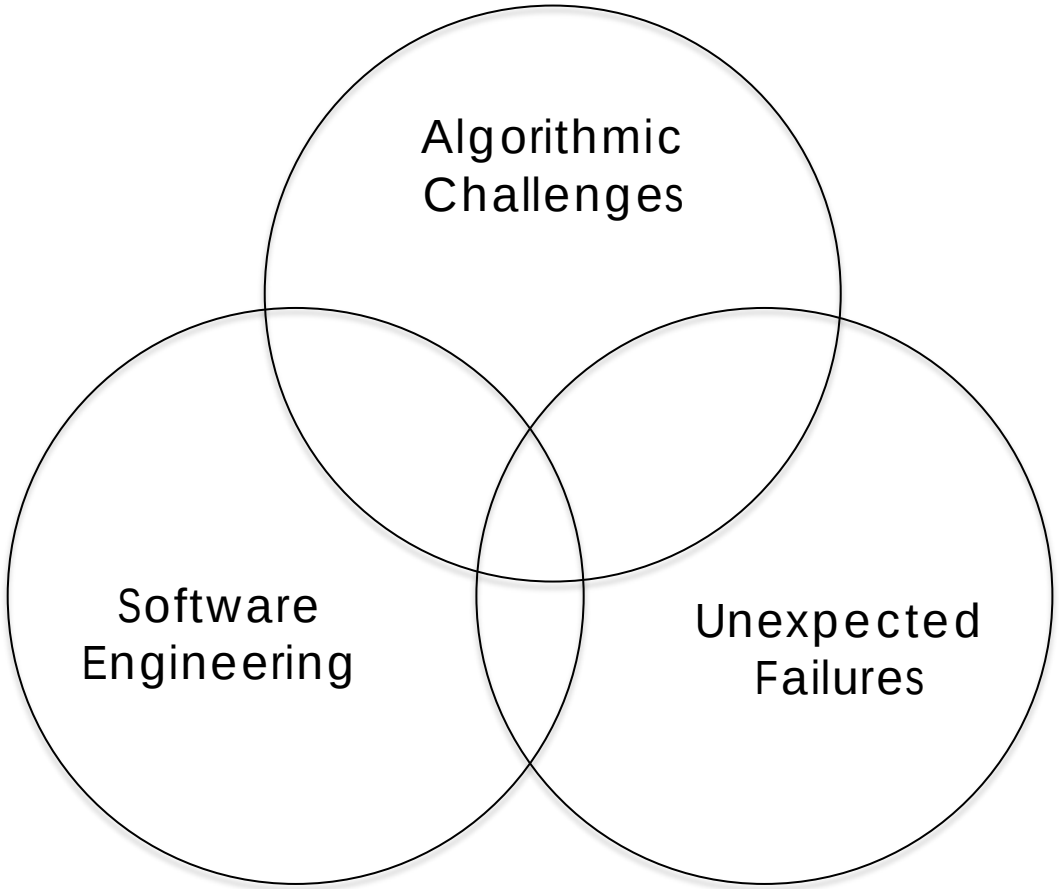
Multi-Paxos


promise <2, 16+>
promise <1, 10, op>
...
promise <1, 15, op>
Replica Log
1 2 ... 14 15
promise <2, 16+>
promise <1, 10, op>
...
promise <1, 15, op>
Replica Log
1 2 ... 14 15
New Master
Replica Log
1 2 ... 13
promise <2, 15+>
promise <1, 10, op>
...
promise <1, 14, op>
1 2 ... 9 10-15
Old Master
1 2 ... 14

Multi-Paxos


Replica Log
1 2 ... 15
accept
acknowledge
commit
accept
acknowledge
commit
Replica Log
1 2 ... 15
accept
acknowledge
commit
New Master
Replica Log
1 2 ... 15
1 2 ... 15
Old Master
1 2 ... 14

Implementing Paxos



Algorithmic Challenges

- Disk Corruption
- Master Leases
- Group Membership

Software Engineering

- Compiler Support
- Runtime Checking
- Testing

Unexpected Failures

- OS Bugs
- Script Bugs
- Rollback Errors
- System Upgrade

Measurements

Test	# workers	file size (bytes)	Paxos-Chubby (100MB DB)	3DB-Chubby (small database)	Comparison
Ops/s Throughput	1	5	91 ops/sec	75 ops/sec	1.2x
Ops/s Throughput	10	5	490 ops/sec	134 ops/sec	3.7x
Ops/s Throughput	20	5	640 ops/sec	178 ops/sec	3.6x
MB/s Throughput	1	8 KB	345 KB/s	172 KB/s	2x
MB/s Throughput	4	8 KB	777 - 949 KB/s	217 KB/s	3.6 - 4.4x
MB/s Throughput	1	32 KB	672 - 822 KB/s	338 KB/s	2.0 - 2.4x

Reference

- Tushar D. Chandra, Robert Griesemer, and Joshua Redstone. 2007. Paxos made live: an engineering perspective. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing* (PODC '07).

Thank you.