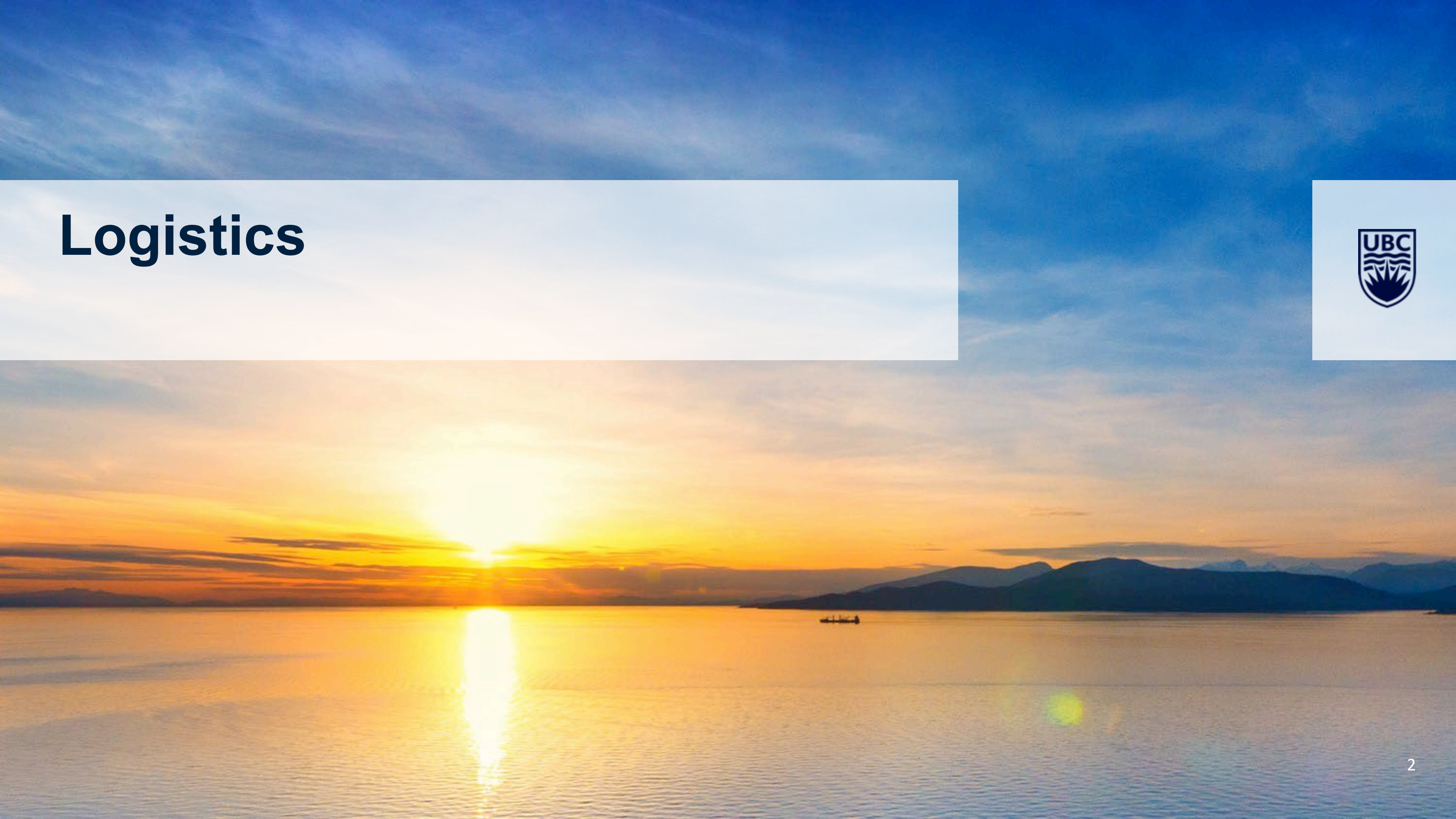
CPSC 416 Distributed Systems

Winter 2022 Term 2 (March 14, 2023)

Tony Mason (fsgeek@cs.ubc.ca), Lecturer



Logistics

Deadlines

Project 4 Released. Extended Due Date: March 20, 2023. Late Due: April 13, 2023.

Project 5 Released Due: April 13, 2023. **No extensions.**

All project work is due April 13, 2023. Late projects are scaled to 75% of the on-time max.

Final Exam: April 20, 2023, DMP 310, 08:30-11:00. Format TBA.

Deadlines

Alternate Path 1 & 2: Review in progress

- Piazza private threads need TLC
  - **Weekly updates due each Monday @ 23:59 PT**
- Final reports due no later than Thursday April 13, 2023 @ 23:59 PT
- Optional 10 min presentation April 13, 2023, up to 10 minutes.

Instructor Office Hours:

- Zoom Office Hours (Tuesday) @ 13:00-14:00
- Discord (Casual) Office Hours (Thursday) @ 14:00-15:00

TA Office Hours:

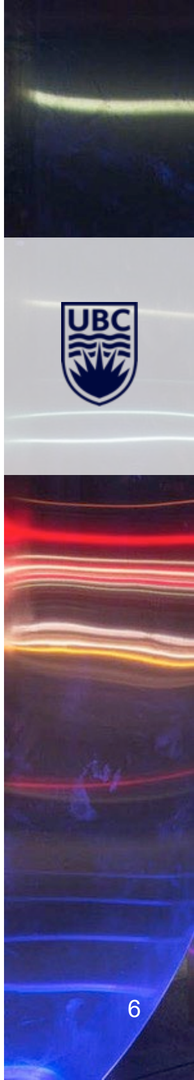
- Eric: Friday 9-11 am (in-person and Zoom)
- Japraj: Wednesday 3-5 pm (Zoom)
- Yennis: Thursday 2-4 (Zoom), Friday 2-4 (in-person)

Questions?

Questions about the class?

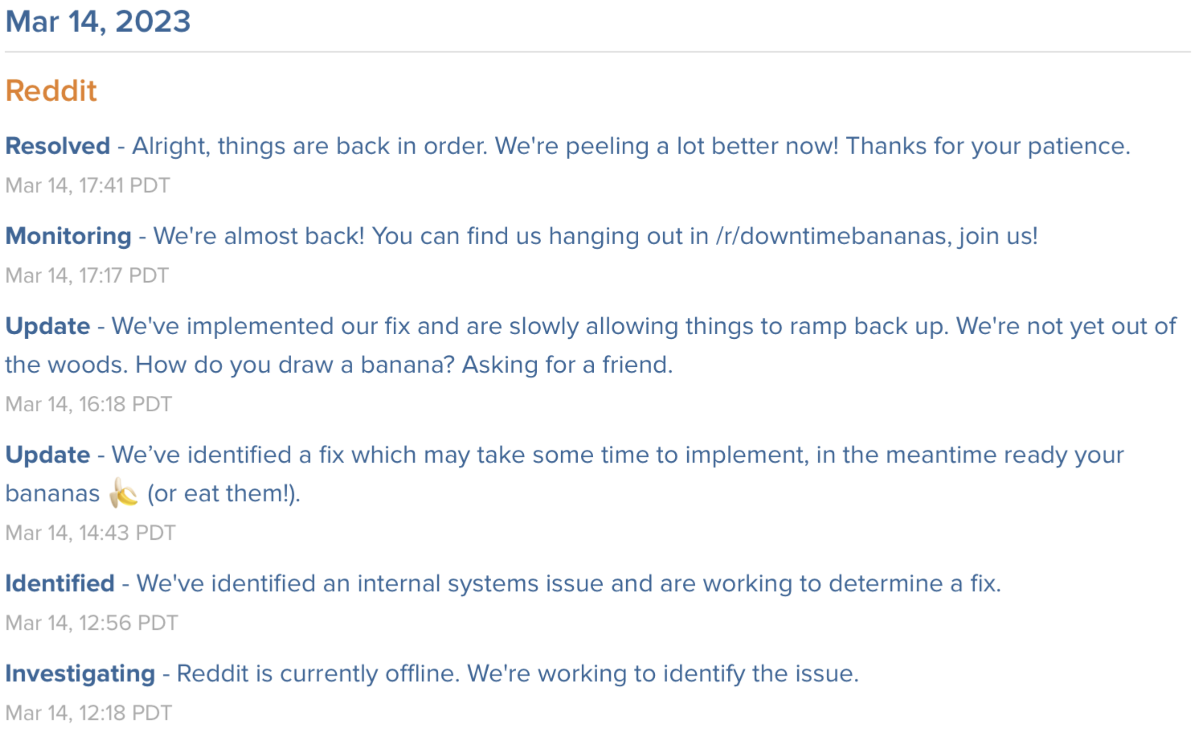
Questions about the previous lecture?

Funny stories to share?



Today’s Failure

Reddit was down!



Mar 14, 2023

Reddit

Resolved - Alright, things are back in order. We're peeling a lot better now! Thanks for your patience.

Mar 14, 17:41 PDT

Monitoring - We're almost back! You can find us hanging out in /r/downtimebananas, join us!

Mar 14, 17:17 PDT

Update - We've implemented our fix and are slowly allowing things to ramp back up. We're not yet out of the woods. How do you draw a banana? Asking for a friend.

Mar 14, 16:18 PDT

Update - We’ve identified a fix which may take some time to implement, in the meantime ready your bananas 🍌 (or eat them!).

Mar 14, 14:43 PDT

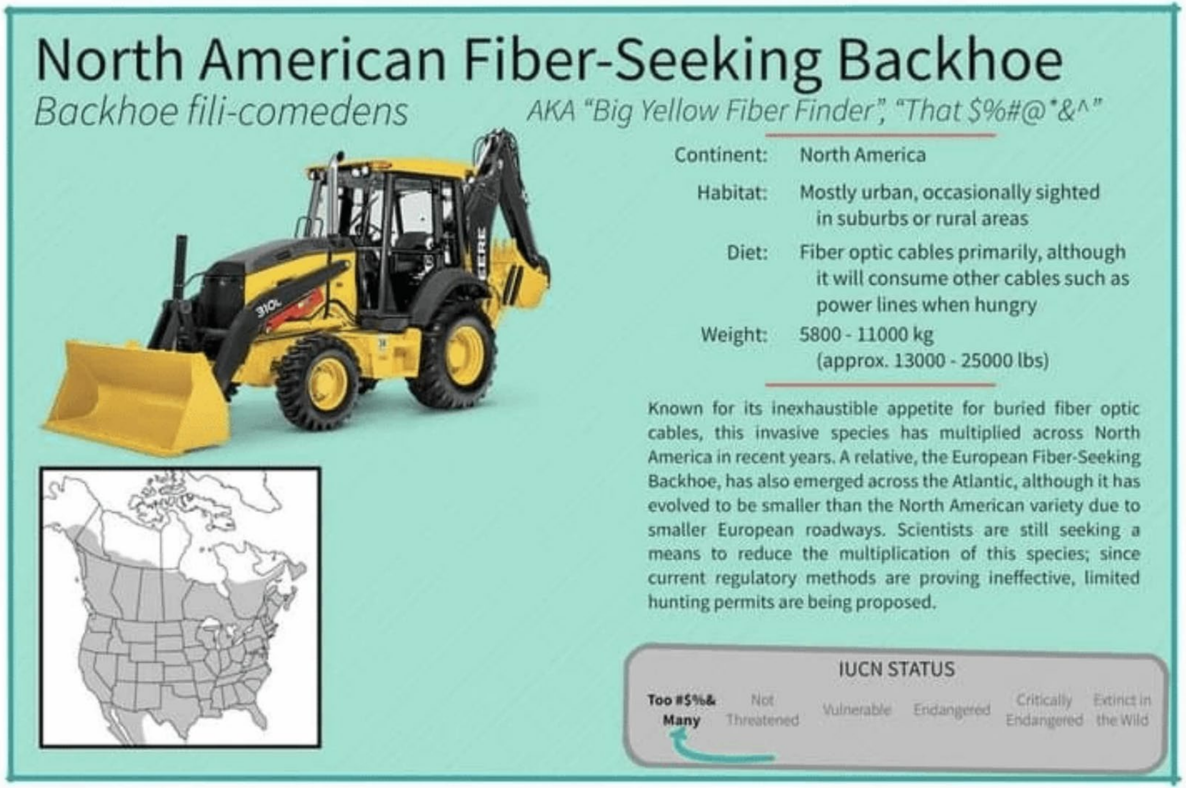
Identified - We've identified an internal systems issue and are working to determine a fix.

Mar 14, 12:56 PDT

Investigating - Reddit is currently offline. We're working to identify the issue.

Mar 14, 12:18 PDT

The culprit (according to Reddit)



North American Fiber-Seeking Backhoe

Backhoe fili-comedens

AKA "Big Yellow Fiber Finder", "That $%#@&^"*

Continent: North America

Habitat: Mostly urban, occasionally sighted in suburbs or rural areas

Diet: Fiber optic cables primarily, although it will consume other cables such as power lines when hungry

Weight: 5800 - 11000 kg (approx. 13000 - 25000 lbs)

Known for its inexhaustible appetite for buried fiber optic cables, this invasive species has multiplied across North America in recent years. A relative, the European Fiber-Seeking Backhoe, has also emerged across the Atlantic, although it has evolved to be smaller than the North American variety due to smaller European roadways. Scientists are still seeking a means to reduce the multiplication of this species; since current regulatory methods are proving ineffective, limited hunting permits are being proposed.

IUCN STATUS

Too #$%& Many · Not Threatened · Vulnerable · Endangered · Critically Endangered · Extinct in the Wild

Takeaways

Networks are *fragile*

Networks can *be fixed*

We design our distributed systems to handle breaks

- Partitions!

We design our distributed systems to handle healing

- De-partitions!
- Recovery

Project 5 Shard KV Store

Review: Consistent Hashing

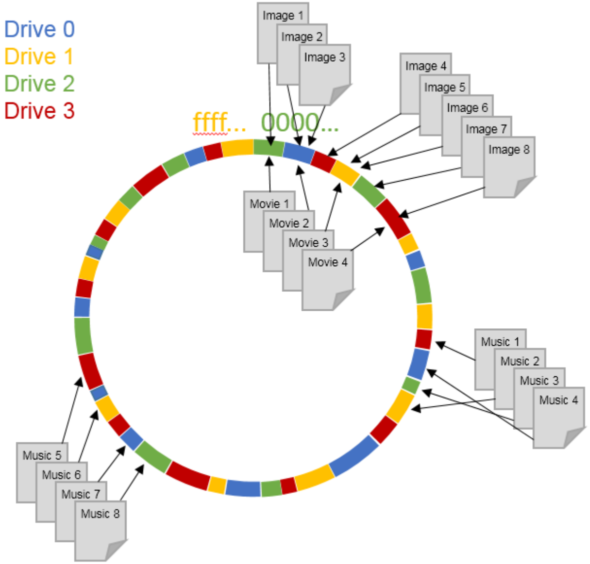
Problem: distribute requests in a changing population of servers.

Solution: Map keys to a location on the circle.

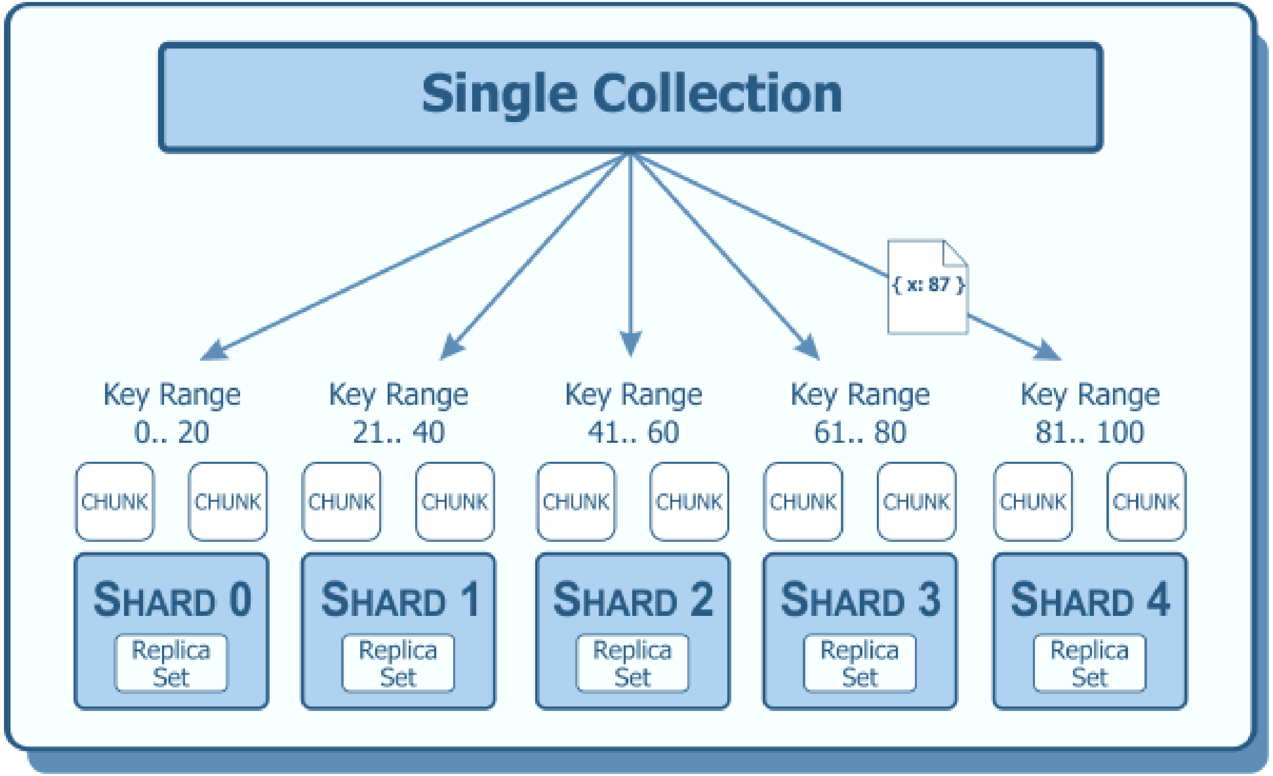
- Insert: can move keys from next node to current node
- Remove: can move keys from current node to next node.

Provides ability to dynamically add/remove servers (or clusters). Better load balancing, dynamic scaling.

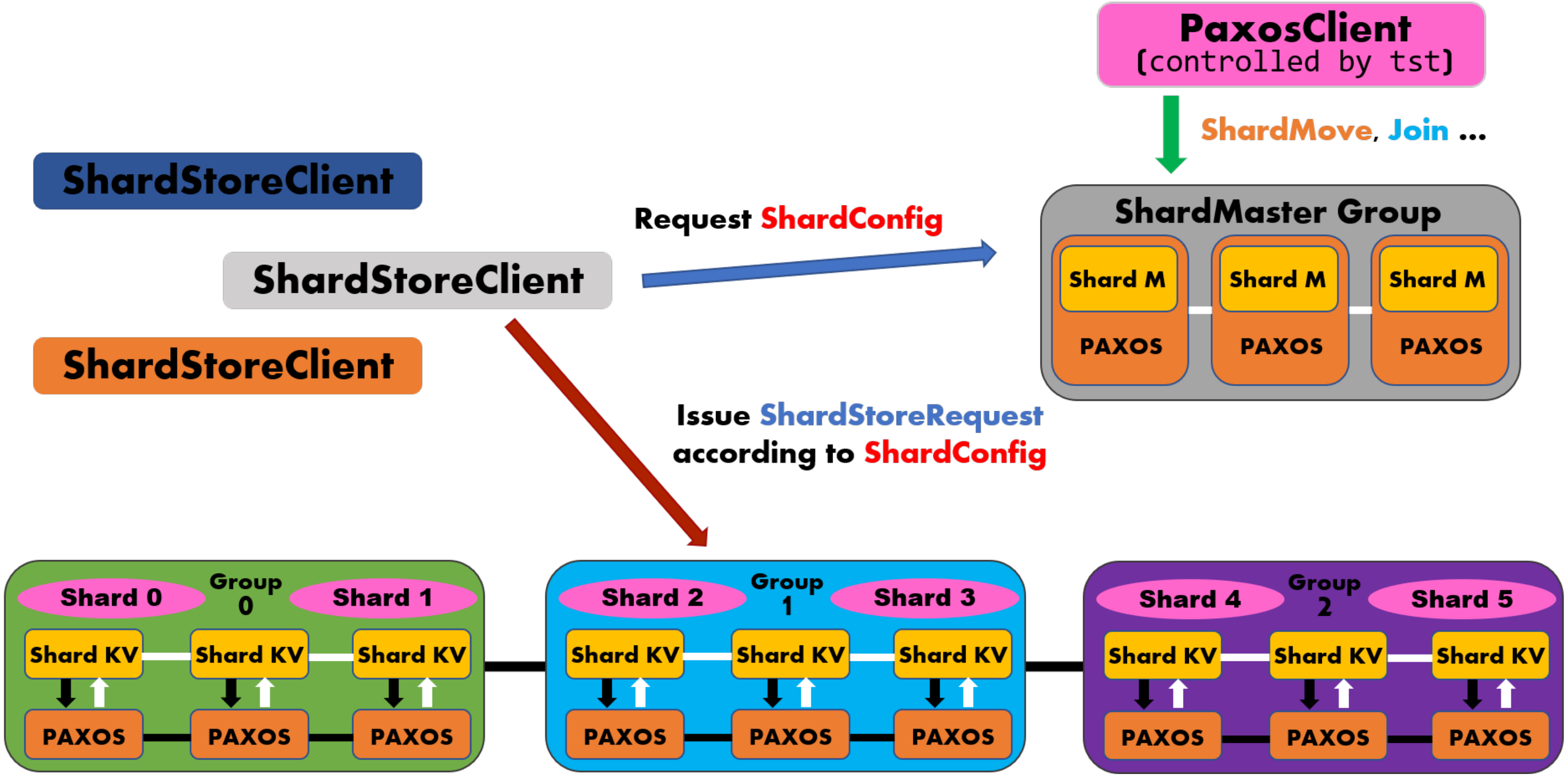
Video (CN 6250 Georgia Tech – Consistent Hashing)



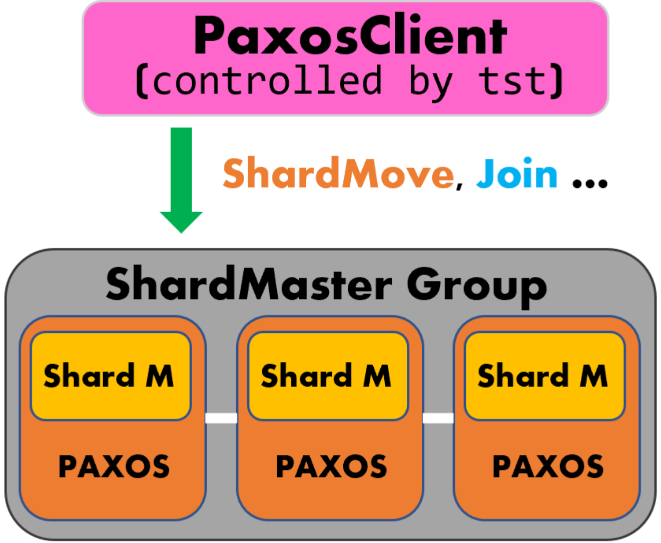
What is Sharding?



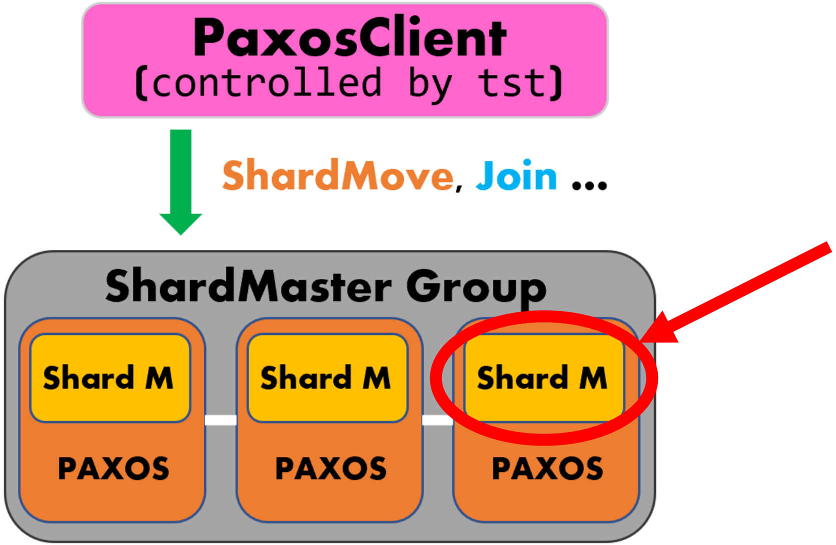
Project 5 Arch Overview



Part 1: Shard Master

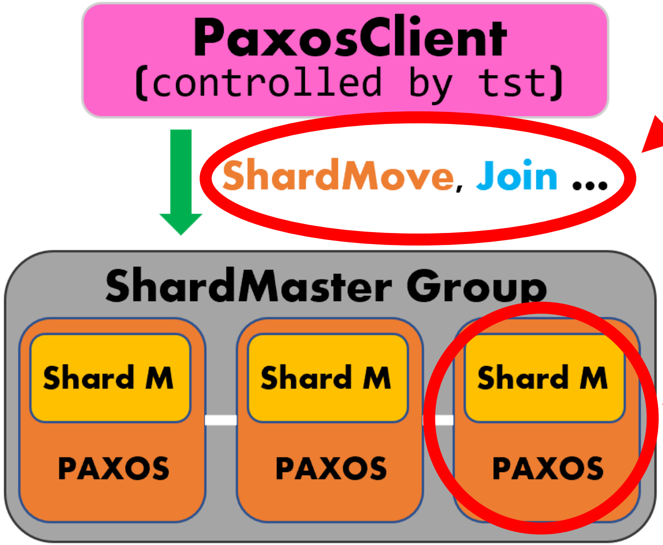


Part 1: Shard Master



- ShardMaster is an **Application** like KVStore
- P4 PAXOS provides fault-tolerance for applications
- PAXOS manages the order of `*PaxosRequest*`

Part 1: Shard Master



- \`*PaxosClient*\` generates commands - *AMOCommand*
- Issue \`*PaxosRequest*\`
- Handle \`*PaxosReply*\`

- Message handlers are in \`*PaxosServer*\`
- Take \`*PaxosRequest*\`
- Return \`*PaxosReply*\`

Server Client Messages

ShardStore Client

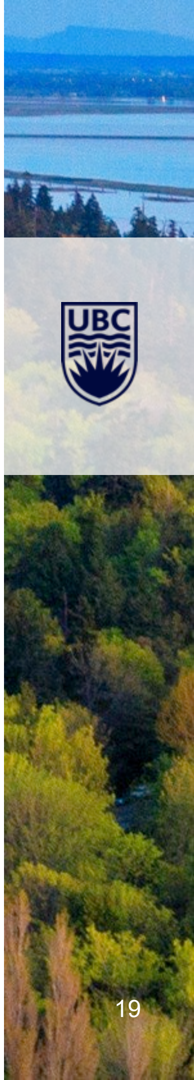
- `*ShardConfig*` from servers in ShardMaster PAXOS group - `*PaxosRequest*`
- `ShardStoreRequest` to ShardStoreServer group
- `ShardStoreReply` from ShardStoreServer group

ShardStore Server

- **ShardMaster server group**: `*ShardConfig*` - `*PaxosRequest*`
- **Clients**: `ShardStoreRequest` & `ShardStoreReply`
- **Other server groups**: shard move commands and acks, 2PC commands, …

Reconfiguration

- First joined group need to initial all shards
- Workflows:
  1. Receive new config
  2. Send local shards to other server groups
  3. Replicate received shards in PAXOS
  4. Update shards info (copy constructor)
  5. Send ack back to the sending groups (already received, already move to new config, …)
- Request *ShardConfig* one-by-one from *ShardMaster* group



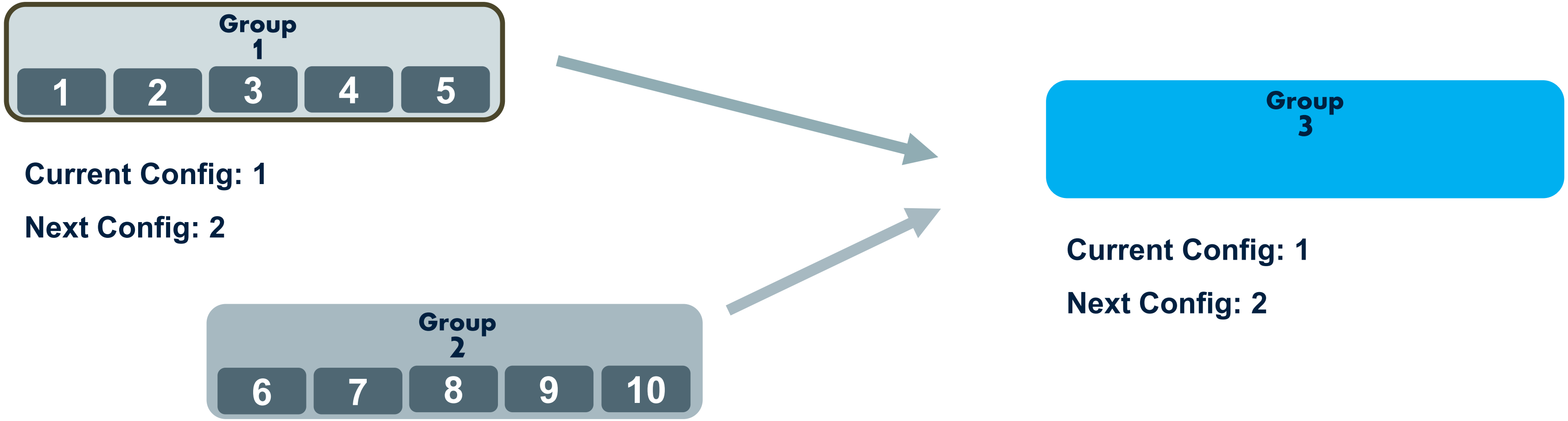
Reconfiguration

Shard M

Config:

1 - <Group 1: 1, 2, 3, 4, 5>, <Group 2: 6, 7, 8, 9, 10>

2 (Group 3 Join) – <Group 1: 1, 2, 3>, <Group 2: 6, 7, 8, 9>, <Group 3: 4, 5, 10>



Current Config: 1

Next Config: 2

Current Config: 1

Next Config: 2

Current Config: 1

Next Config: 2

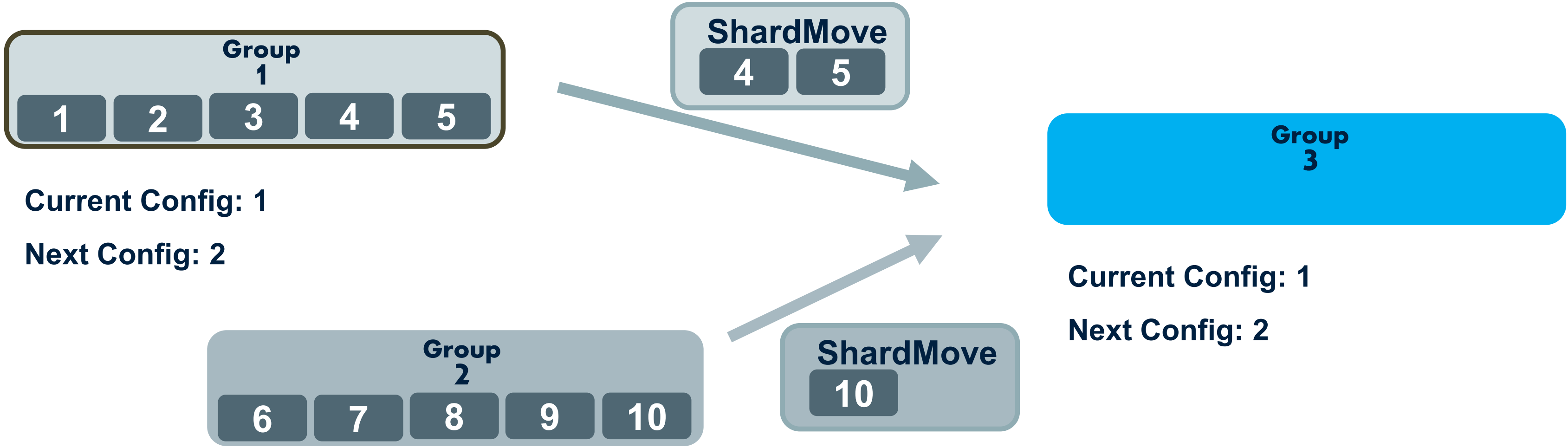
Reconfiguration

Shard M

Config:

1 - <Group 1: 1, 2, 3, 4, 5>, <Group 2: 6, 7, 8, 9, 10>

2 (Group 3 Join) – <Group 1: 1, 2, 3>, <Group 2: 6, 7, 8, 9>, <Group 3: 4, 5, 10>



Current Config: 1

Next Config: 2

Current Config: 1

Next Config: 2

Current Config: 1

Next Config: 2

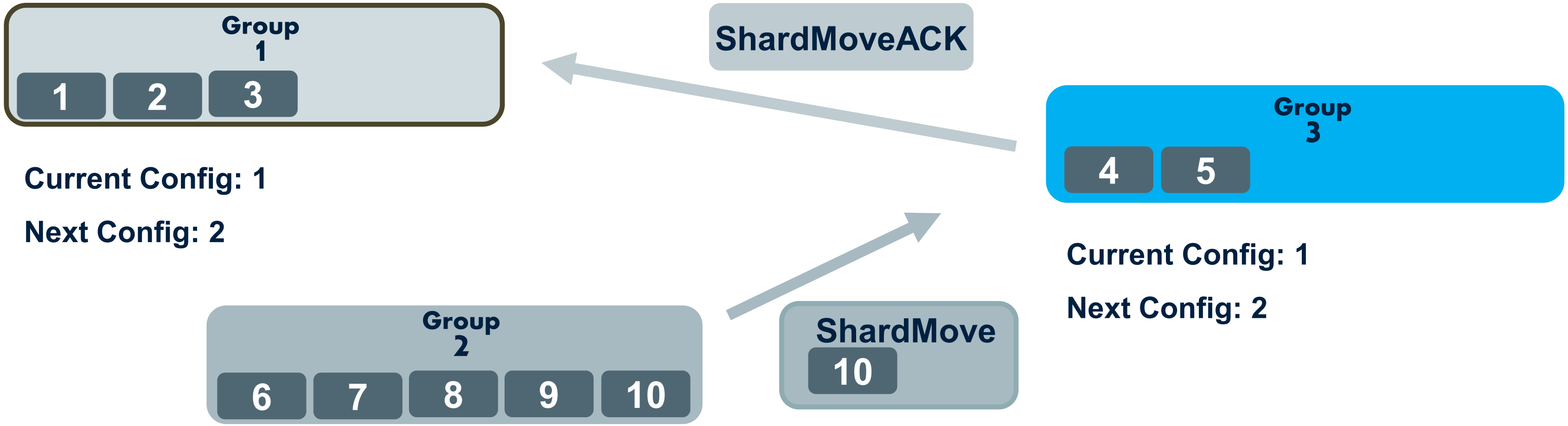
Reconfiguration

Shard M

Config:

1 - <Group 1: 1, 2, 3, 4, 5>, <Group 2: 6, 7, 8, 9, 10>

2 (Group 3 Join) – <Group 1: 1, 2, 3>, <Group 2: 6, 7, 8, 9>, <Group 3: 4, 5, 10>



Group 1

1 2 3

Current Config: 1

Next Config: 2

ShardMoveACK

Group 3

4 5

Current Config: 1

Next Config: 2

Group 2

6 7 8 9 10

ShardMove

10

Current Config: 1

Next Config: 2

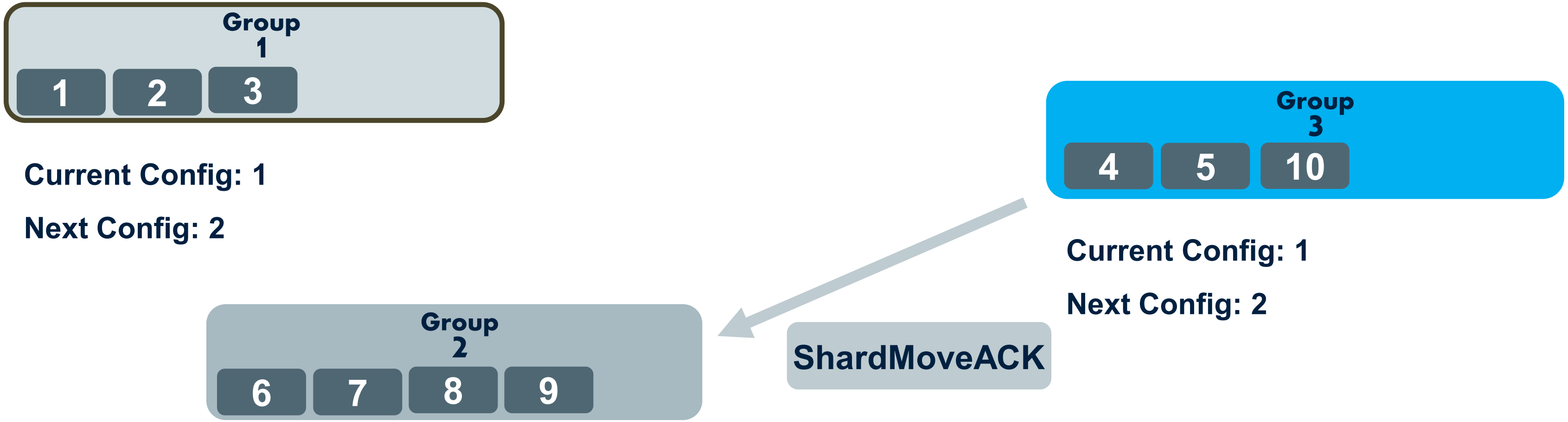
Reconfiguration

Shard M

Config:

1 - <Group 1: 1, 2, 3, 4, 5>, <Group 2: 6, 7, 8, 9, 10>

2 (Group 3 Join) – <Group 1: 1, 2, 3>, <Group 2: 6, 7, 8, 9>, <Group 3: 4, 5, 10>



Current Config: 1

Next Config: 2

Current Config: 1

Next Config: 2

Current Config: 1

Next Config: 2

Transaction

- Client select one group as the coordinator

- Shard-level locks

- Coordinator group sends `phase 1 – prepare` to followers (one-by-one with order)

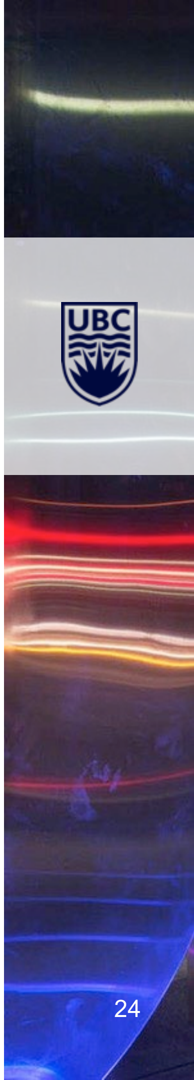
- Phase 1 response
  - All followers responds `phase 1 – ready`:

    Coordinator group sends `phase 2 – prepare`. All followers commit the change send `phase 2 – commit`.

  - One follower responds `phase 1 – abort`:

    Coordinator group sends `phase 2 – abort`. All previous ready followers abort the txn and proceeds to other pending txns.

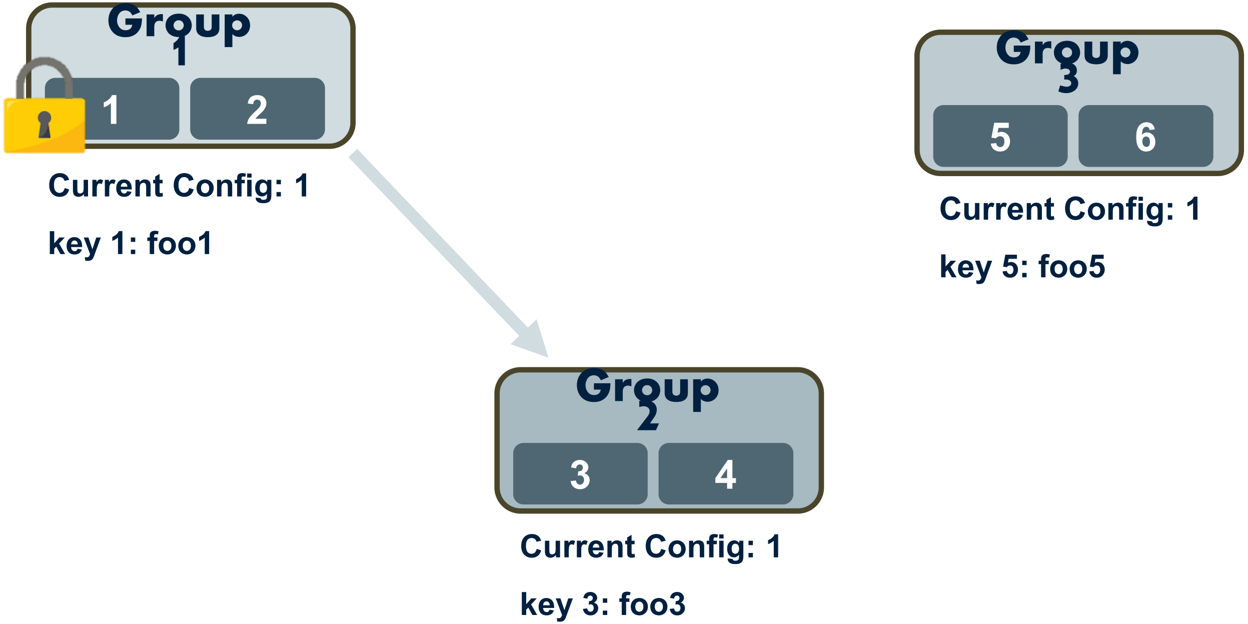
- Mechanism for detecting duplicated `phase 1 – prepare`



Transaction – Commit Case

TXN:

Get - <key 1, key 3, key 5>



Current Config: 1

key 1: foo1

Current Config: 1

key 3: foo3

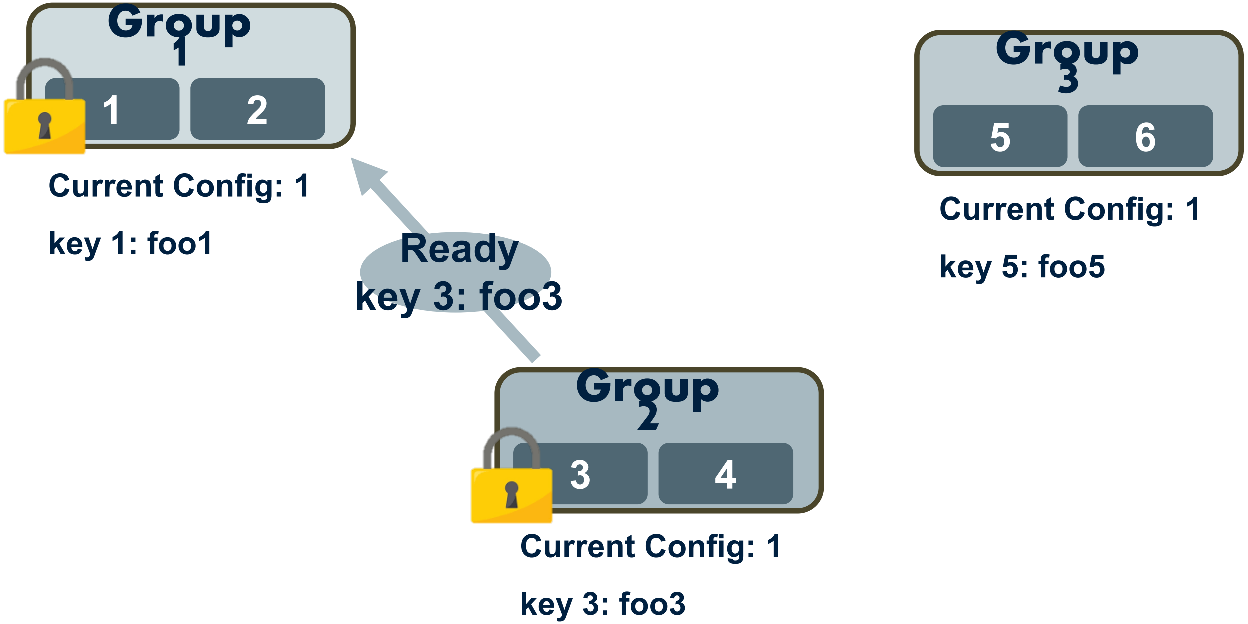
Current Config: 1

key 5: foo5

Transaction – Commit Case

TXN:

Get - <key 1, key 3, key 5>



Group 1

1

2

Current Config: 1

key 1: foo1

Ready

key 3: foo3

Group 2

3

4

Current Config: 1

key 3: foo3

Group 3

5

6

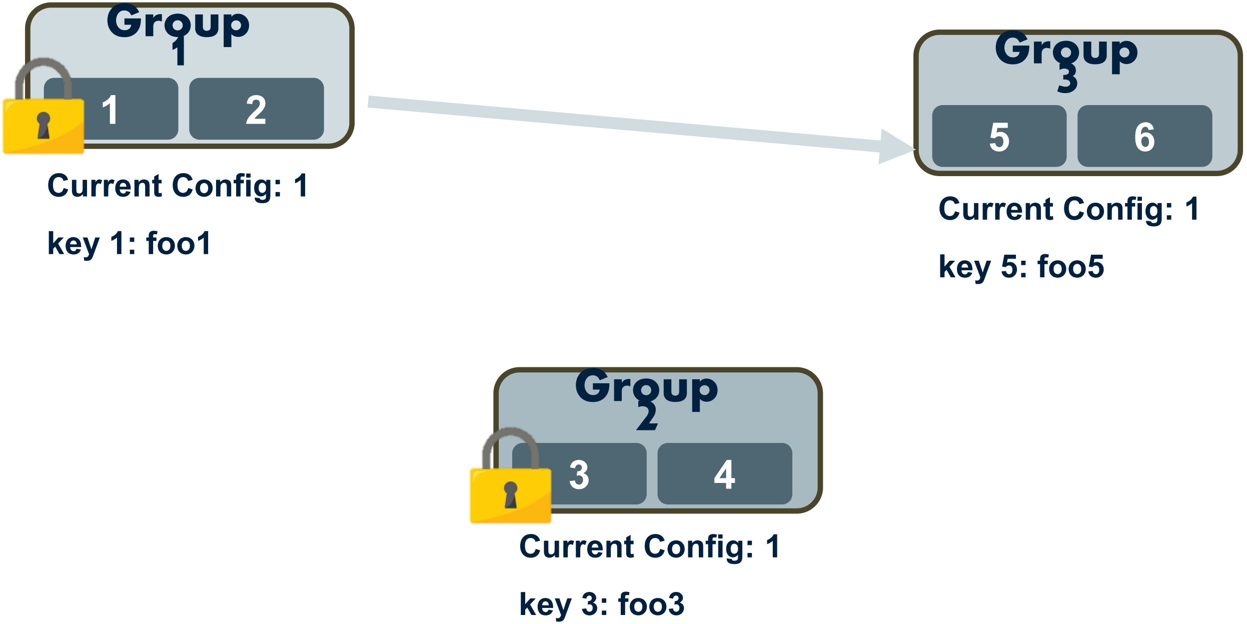
Current Config: 1

key 5: foo5

Transaction – Commit Case

TXN:

Get - <key 1, key 3, key 5>



Group 1

1 2

Current Config: 1

key 1: foo1

Group 3

5 6

Current Config: 1

key 5: foo5

Group 2

3 4

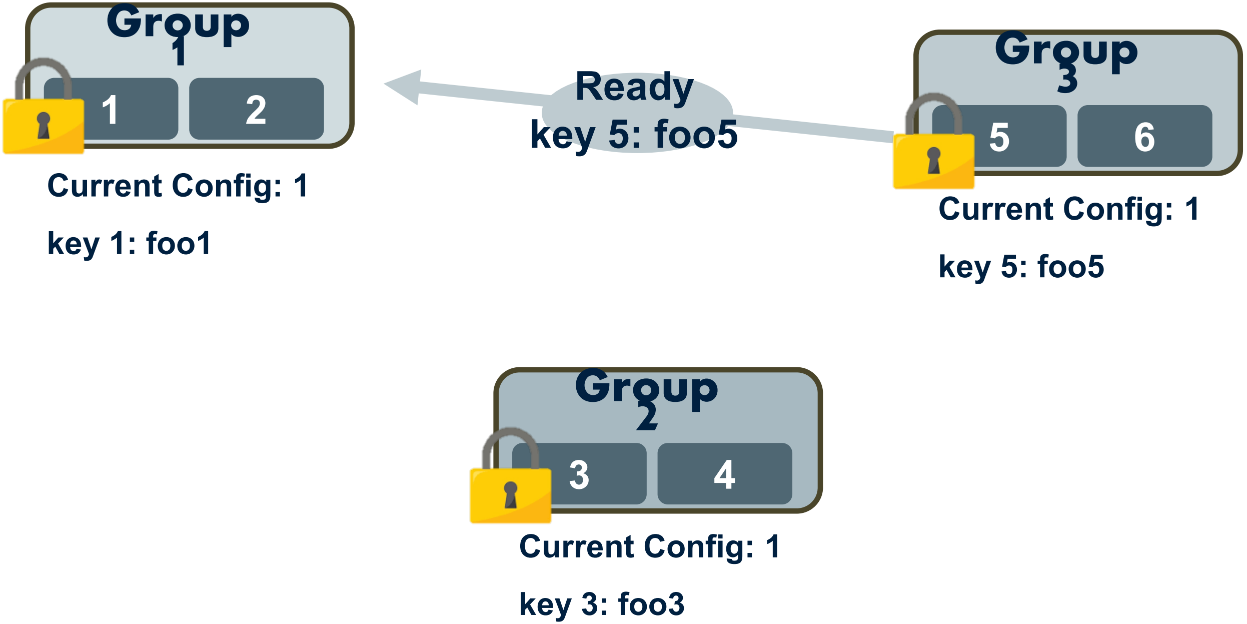
Current Config: 1

key 3: foo3

Transaction – Commit Case

TXN:

Get - <key 1, key 3, key 5>



Current Config: 1

key 1: foo1

Current Config: 1

key 5: foo5

Current Config: 1

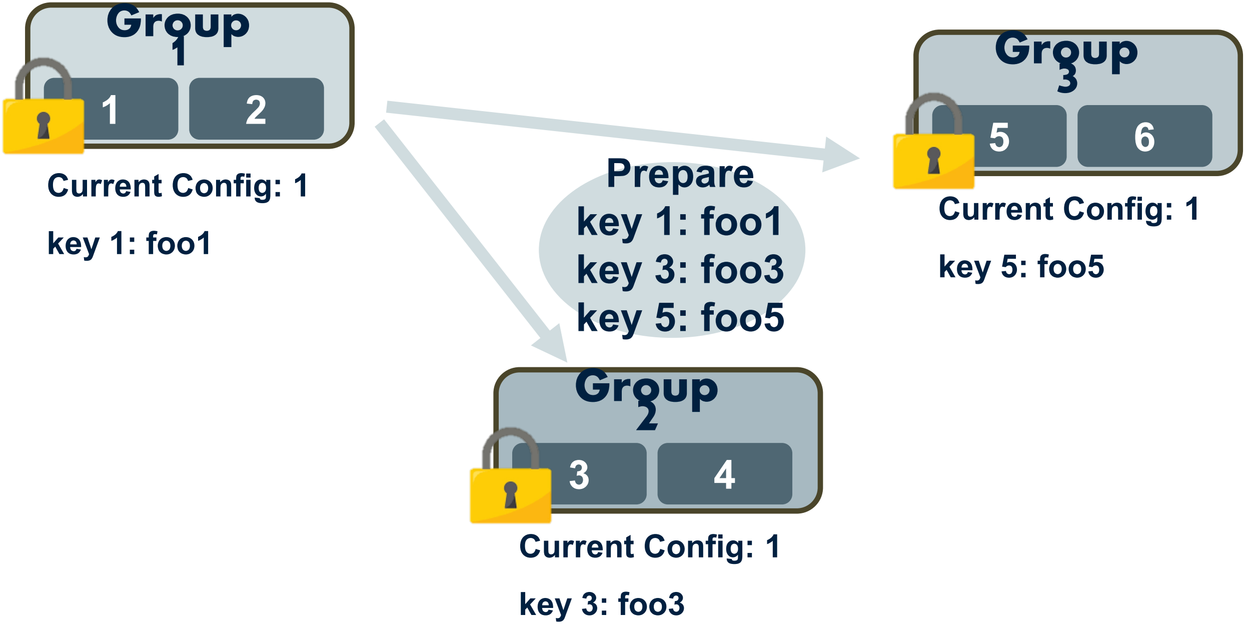
key 3: foo3

Transaction – Commit Case

TXN:

Get - <key 1, key 3, key 5>

Make result consistent and durable.



Group 1

1 2

Current Config: 1

key 1: foo1

Prepare
key 1: foo1
key 3: foo3
key 5: foo5

Group 2

3 4

Current Config: 1

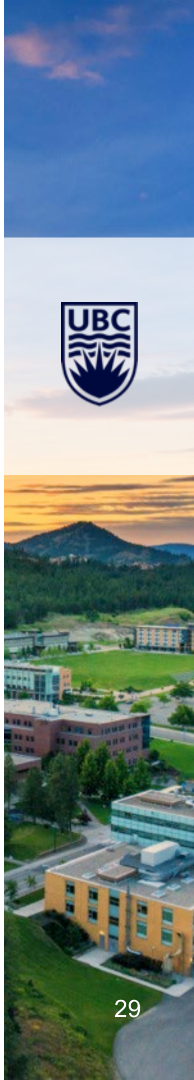
key 3: foo3

Group 3

5 6

Current Config: 1

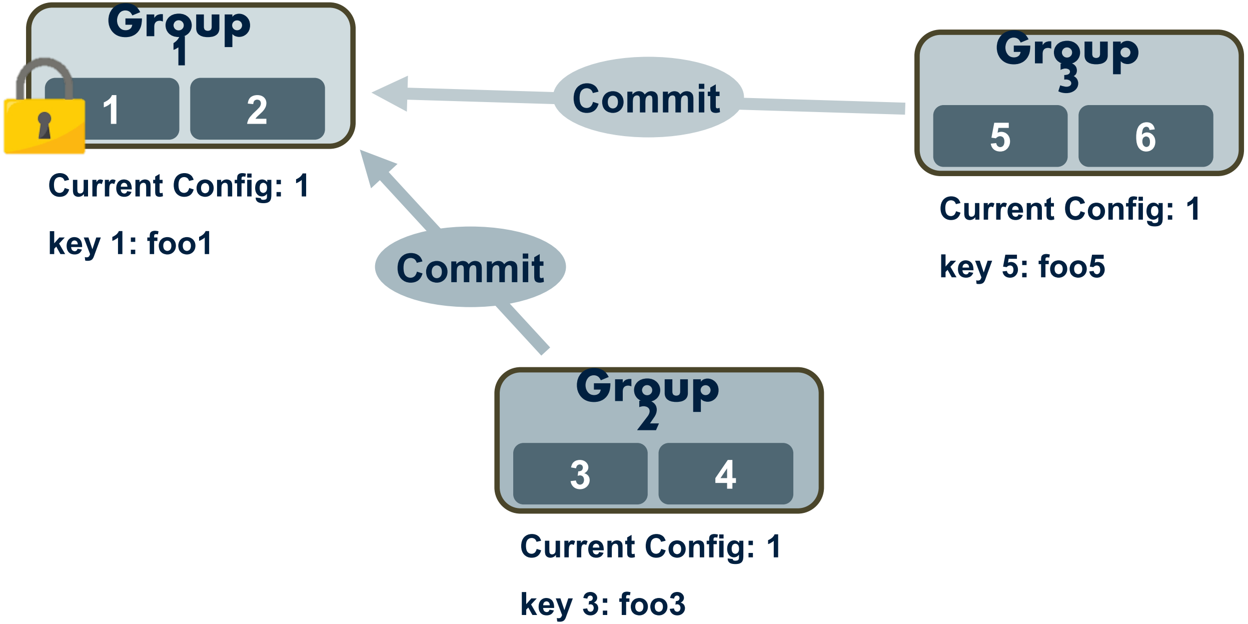
key 5: foo5



Transaction – Commit Case

TXN:

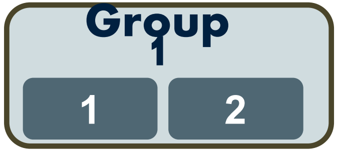
Get - <key 1, key 3, key 5>



Transaction – Commit Case

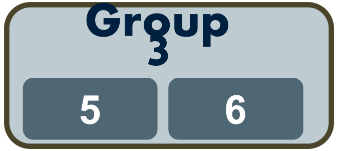
TXN:

Get - <key 1, key 3, key 5>



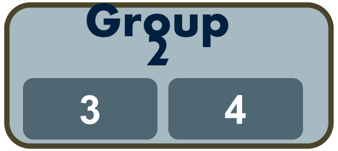
Current Config: 1

key 1: foo1



Current Config: 1

key 5: foo5



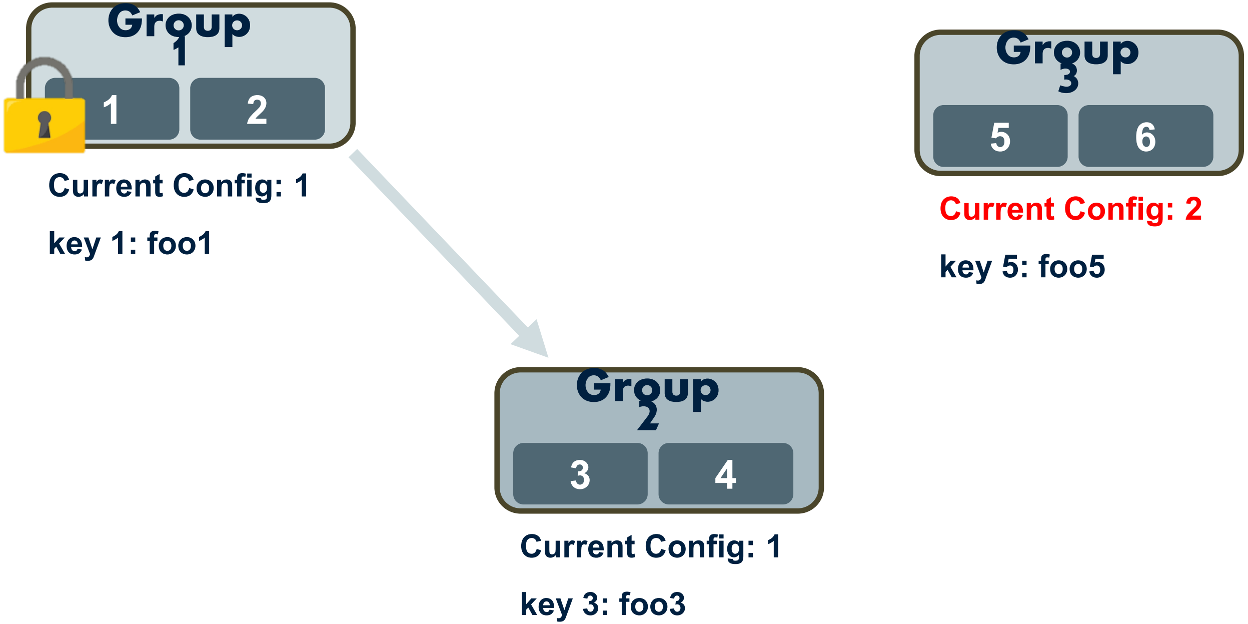
Current Config: 1

key 3: foo3

Transaction – Abort Case

TXN:

Get - <key 1, key 3, key 5>



Group 1

1

2

Current Config: 1

key 1: foo1

Group 3

5

6

Current Config: 2

key 5: foo5

Group 2

3

4

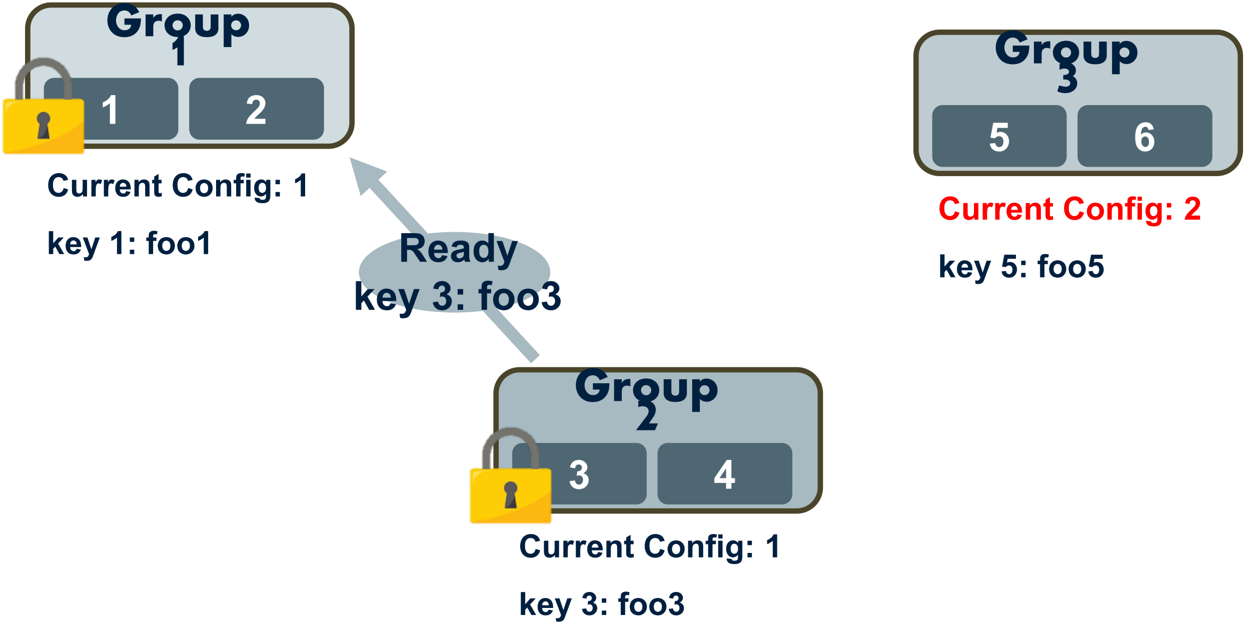
Current Config: 1

key 3: foo3

Transaction – Abort Case

TXN:

Get - <key 1, key 3, key 5>



Group 1

Current Config: 1

key 1: foo1

Ready
key 3: foo3

Group 2

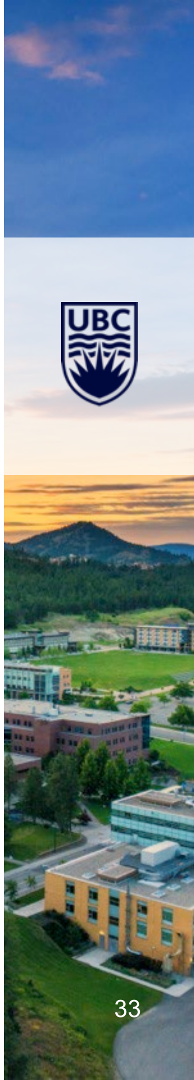
Current Config: 1

key 3: foo3

Group 3

Current Config: 2

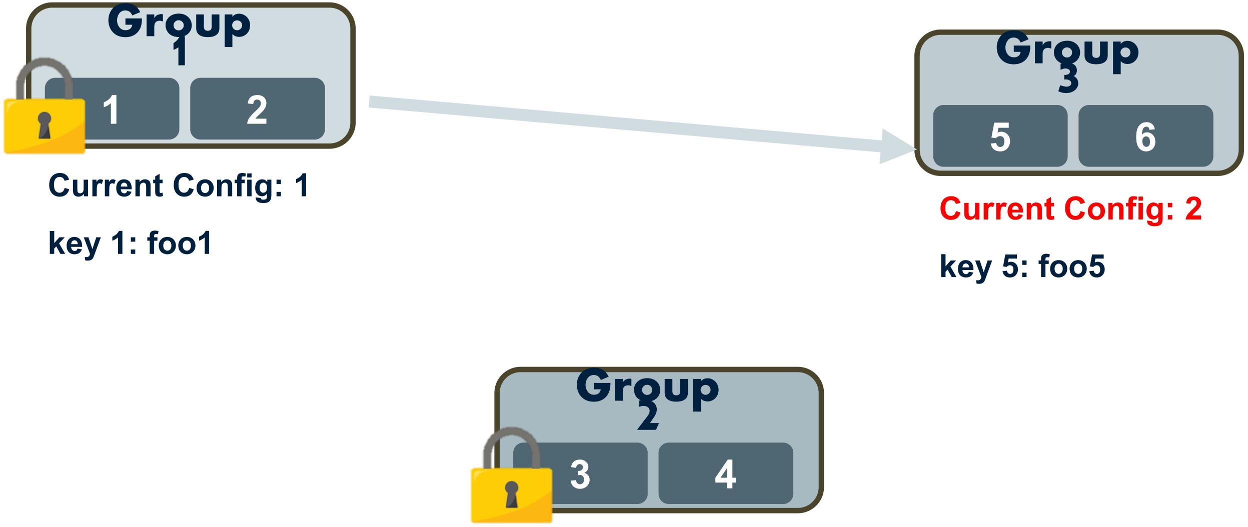
key 5: foo5



Transaction – Abort Case

TXN:

Get - <key 1, key 3, key 5>



Current Config: 1

key 1: foo1

Current Config: 2

key 5: foo5

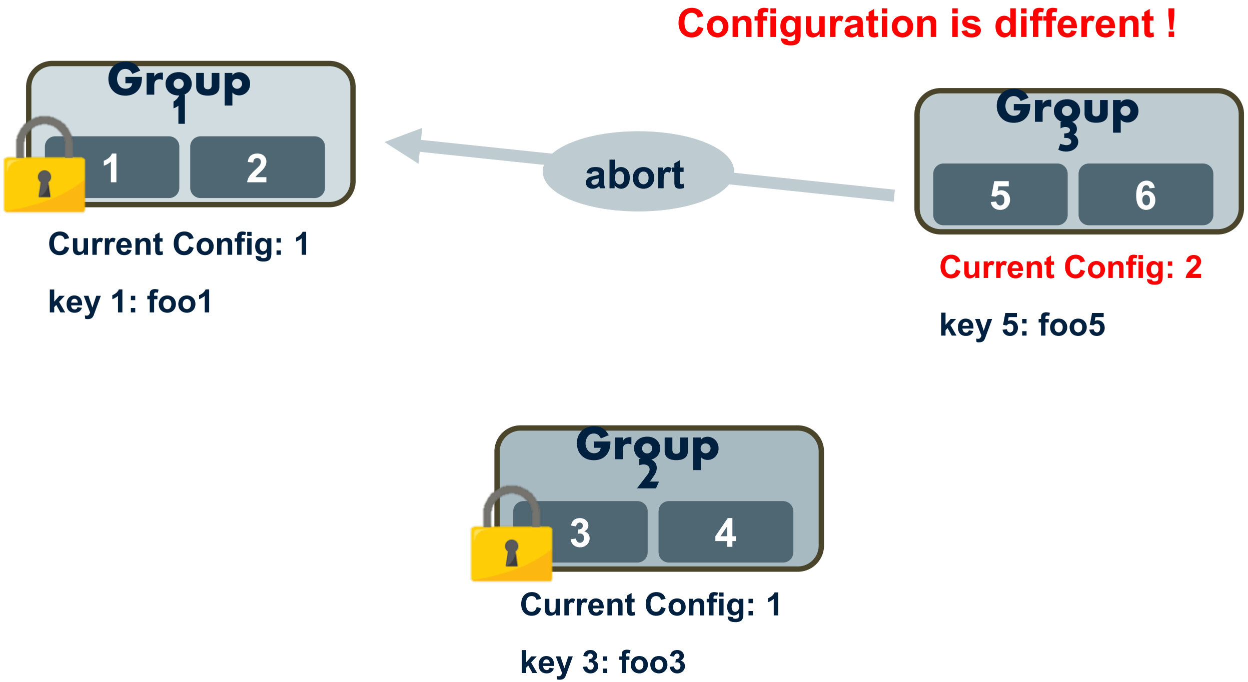
Current Config: 1

key 3: foo3

Transaction – Abort Case

TXN:

Get - <key 1, key 3, key 5>



Configuration is different !

Current Config: 1

key 1: foo1

Current Config: 2

key 5: foo5

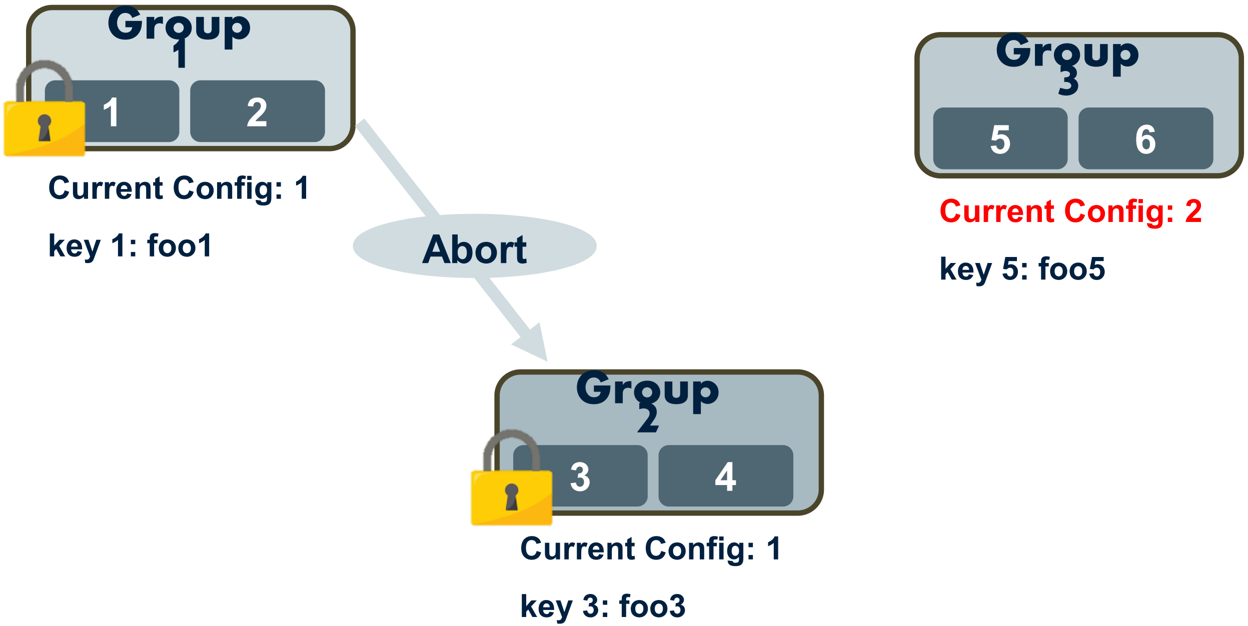
Current Config: 1

key 3: foo3

Transaction – Abort Case

TXN:

Get - <key 1, key 3, key 5>



Group 1

1 2

Current Config: 1

key 1: foo1

Abort

Group 2

3 4

Current Config: 1

key 3: foo3

Group 3

5 6

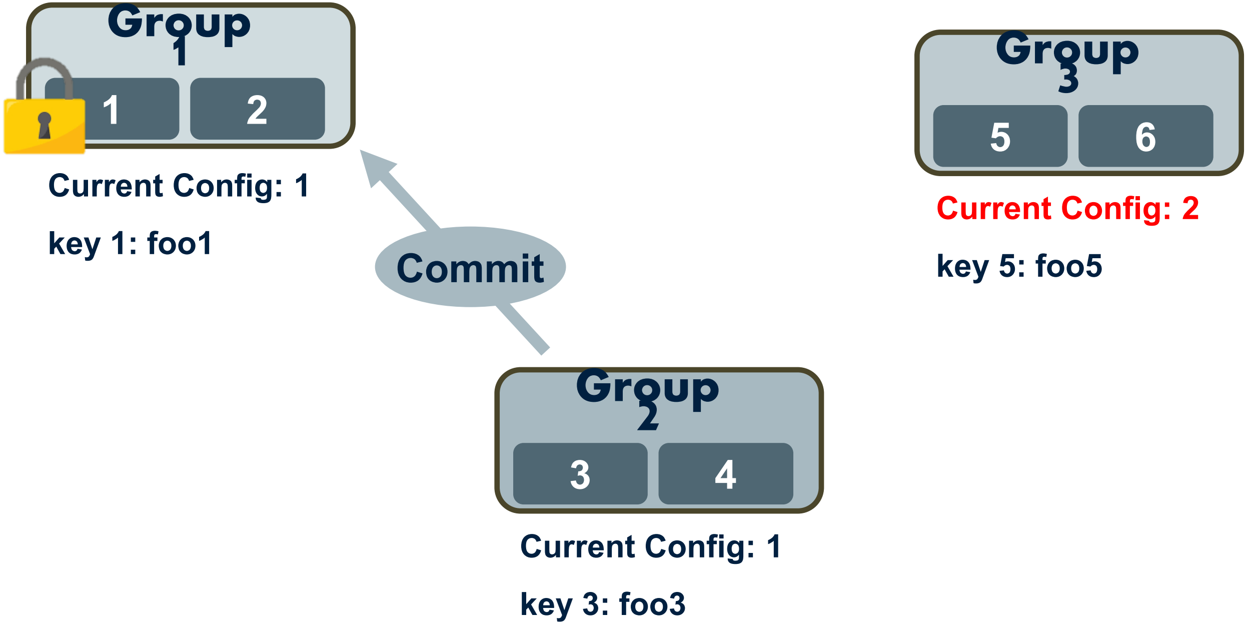
Current Config: 2

key 5: foo5

Transaction – Abort Case

TXN:

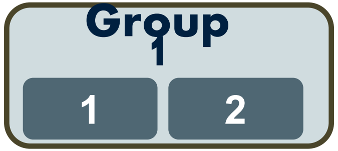
Get - <key 1, key 3, key 5>



Transaction – Abort Case

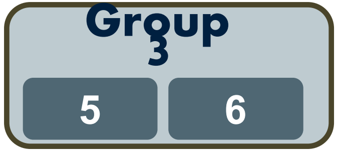
TXN:

Get - <key 1, key 3, key 5>



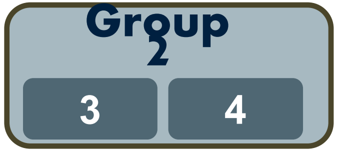
Current Config: 1

key 1: foo1



Current Config: 2

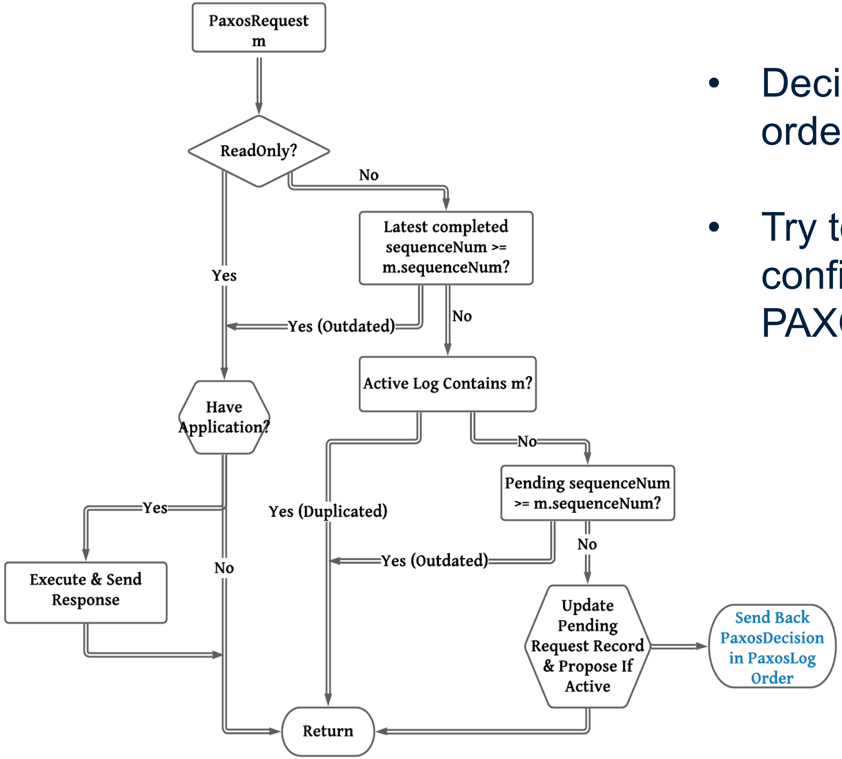
key 5: foo5



Current Config: 1

key 3: foo3

Reference PAXOS Note


PaxosRequest m
ReadOnly?
No
Yes
Latest completed sequenceNum >= m.sequenceNum?
Yes (Outdated)
No
Have Application?
Active Log Contains m?
No
Yes
Yes (Duplicated)
Pending sequenceNum >= m.sequenceNum?
Execute & Send Response
No
Yes (Outdated)
No
Update Pending Request Record & Propose If Active
Send Back PaxosDecision in PaxosLog Order
Return

- Decisions are only sent in the `PaxosLog` order. No duplicated decisions sent back.
- Try to tag the client requests with the current configuration number when proposing to PAXOS.

Questions?

THE UNIVERSITY OF BRITISH COLUMBIA