etcd

Overview and Future

Jonathan Boulle
CoreOS
@baronboulle

What is etcd?

/etc

distributed

hence, the name...

a clustered

key-value store

GET and SET operations

a **building block** for higher order systems

primitives for building reliable distributed systems

What is etcd?

- distributed `/etc`
    - cluster-level configuration
- clustered key-value store
- primitives for building reliable distributed systems
    - distributed locking system
    - distributed scheduling system
- supports a lot of large distributed applications
    - SkyDNS, Kubernetes, CloudFoundry, ...

History of etcd

History of etcd

- 2013.8 Alpha release (v0.x)
- 2015.2 Stable release (v2.0+)
  - stable replication engine (new Raft implementation)
  - stable v2 API
- 2016.? (v3.0+)
  - efficient, powerful API
  - highly scalable backend

etcd today

- Production-ready!
  - long running failure injection tests
  - no known data loss issues
  - no known inconsistency issues
  - used in critical CoreOS systems like locksmith and fleet
  - trusted by Google, Pivotal, compose and many more

Why build etcd?

Why build etcd?

- CoreOS mission: "Secure the internet"
- Updating servers = rebooting servers
- Move towards new application container paradigm
- Need a:
  - shared configuration store (for service discovery etc)
  - distributed lock manager (to co-ordinate reboots)
- Existing solutions were inflexible (undocumented binary API), difficult to configure

Why use etcd?

Why use etcd?

- Highly available
- Highly reliable
- Strong consistency guarantees
- Simple, fast HTTP API
- Open source

"For the most critical data of a distributed system"

How does etcd work?

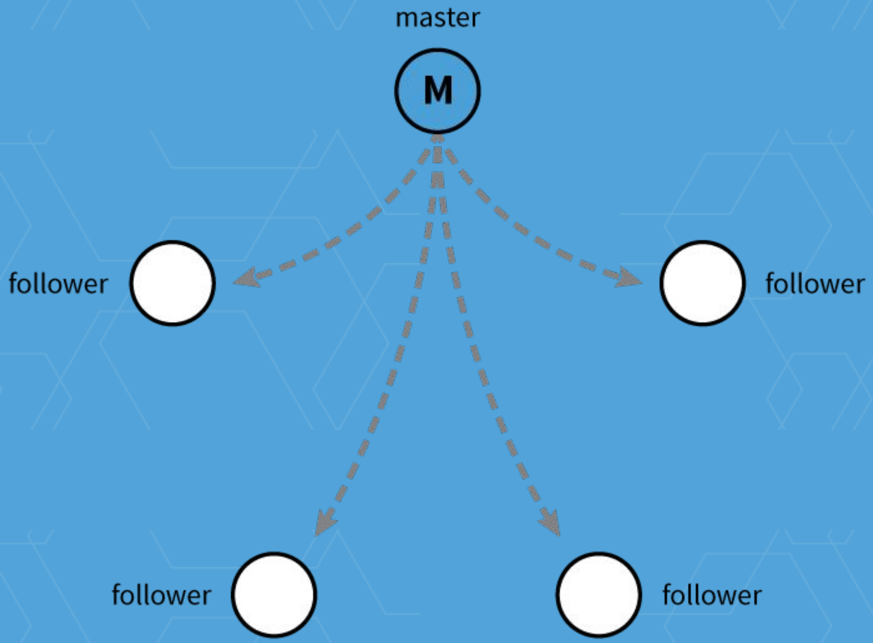
How does etcd work?

- **Raft** consensus algorithm
  - Using a replicated log to model a state machine
  - *"In Search of an Understandable Consensus Algorithm"* (Ongaro, 2014)

- Three key concepts
  - *Leaders*
  - *Elections*
  - *Terms*

How does etcd work?

- The cluster *elects* a leader for every given *term*
- All log appends (--> state machine changes) are decided by that leader and propagated to followers
- Much much more at http://raft.github.io/



How does etcd work?

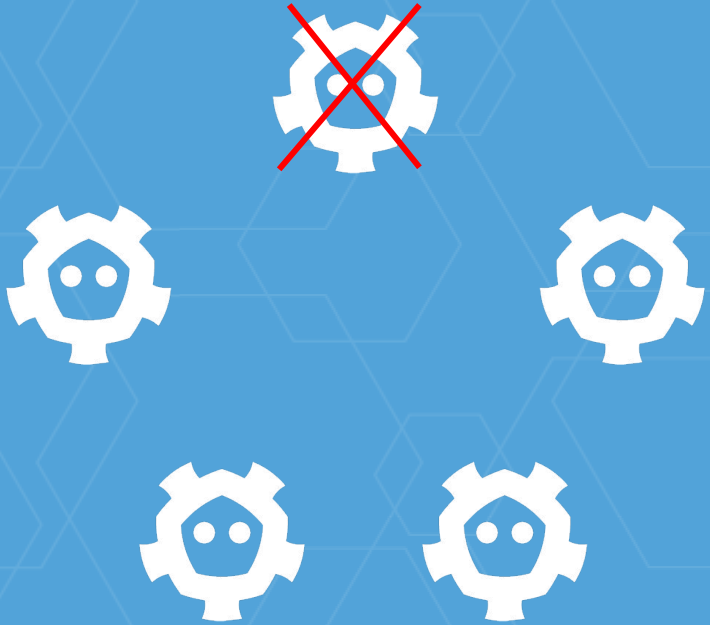
- Written in Go, statically linked
- `/bin/etcd`
  - daemon
  - 2379 (client requests/HTTP + JSON API)
  - 2380 (peer-to-peer/HTTP + protobuf)
- `/bin/etcdctl`
  - command line client
- `net/http, encoding/json, golang/protobuf, ...`

etcd basics

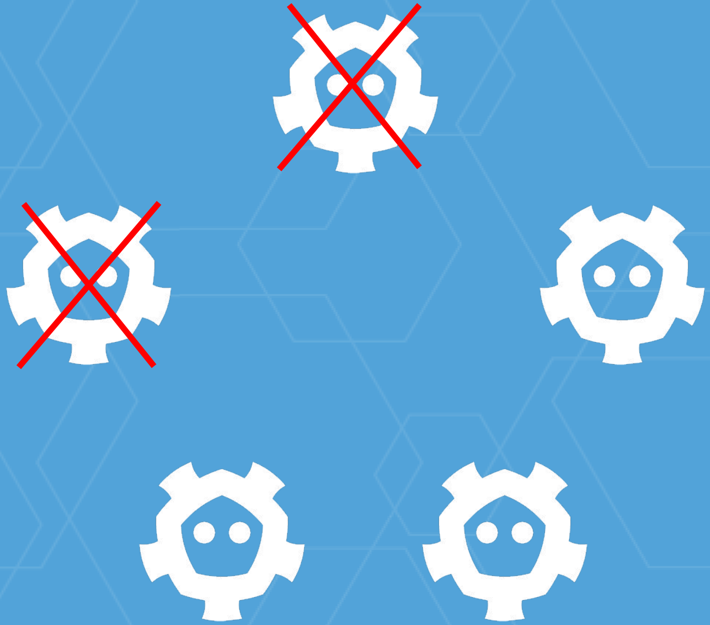
clusters

Available

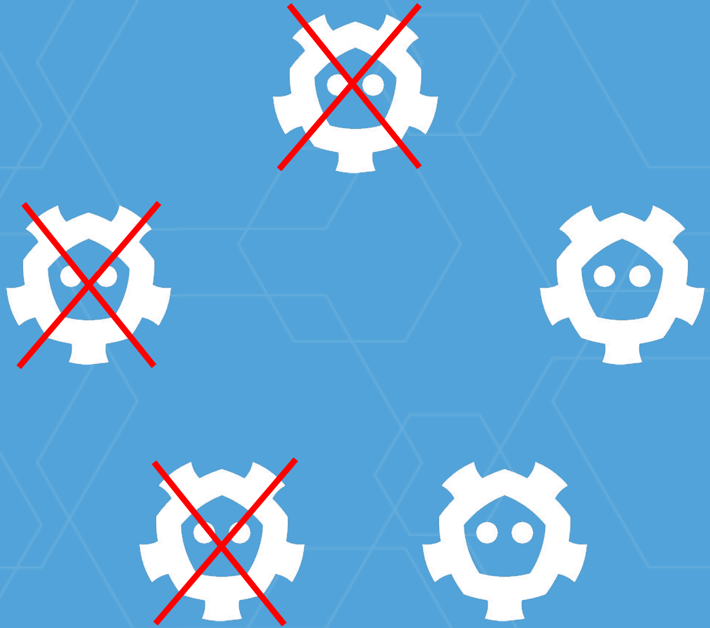
Available



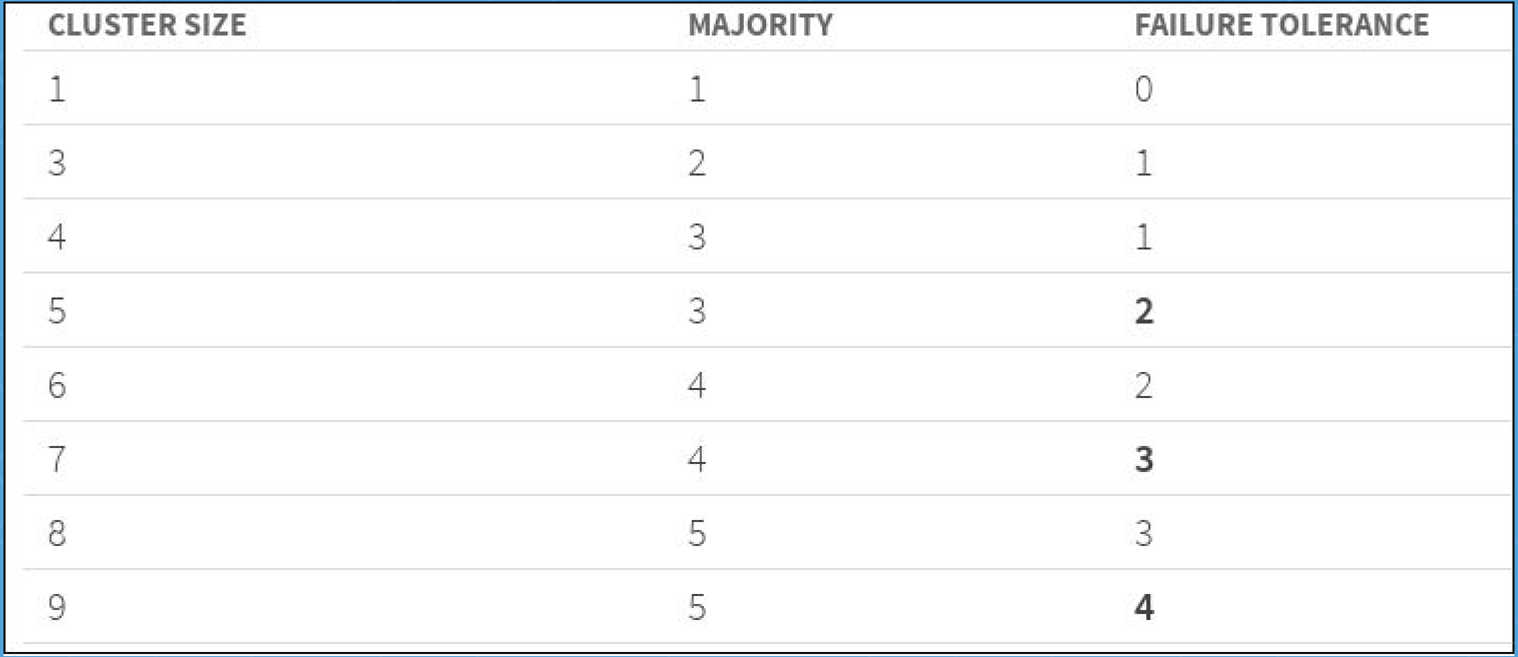
Available



Unavailable



etcd cluster failure tolerance



CLUSTER SIZE	MAJORITY	FAILURE TOLERANCE
1	1	0
3	2	1
4	3	1
5	3	**2**
6	4	2
7	4	**3**
8	5	3
9	5	**4**

etcd basics

API

Simple HTTP API (v2)

- `GET /v2/keys/foo`
  - Get the value of a key
- `GET /v2/keys/foo?wait=true`
  - Wait for changes on key foo
  - Receive notifications on every change
- `PUT /v2/keys/foo -d value=bar`
  - Set the value of a key
- `DELETE /v2/keys/foo`
  - Delete a key

Compare-and-Swap

```
PUT /v2/keys/foo?prevValue=bar -d
value=ok
```

```
CAS(/foo, bar, ok)

if /foo == bar
   set(/foo, ok)
else
   do nothing
```

Compare-and-Delete

```
DELETE /v2/keys/foo?prevValue=bar
```

```
CAD(/foo, bar)

if /foo == bar
    delete(/foo)
else
    do nothing
```

Simple HTTP API (v2)

Native Go bindings

```
import "github.com/coreos/etcd/client"


cl := client.New(client.Config{})
kapi := client.NewKeysAPI(cl)
kapi.Set("foo", "bar", ...)
```

etcd apps

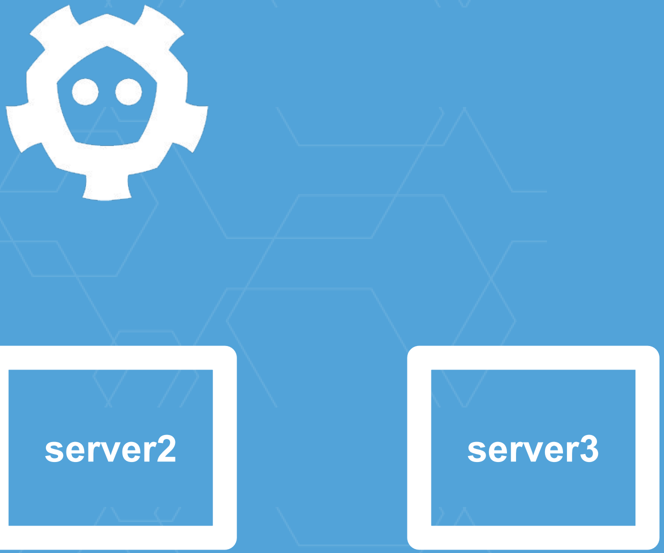
etcd apps

locksmith

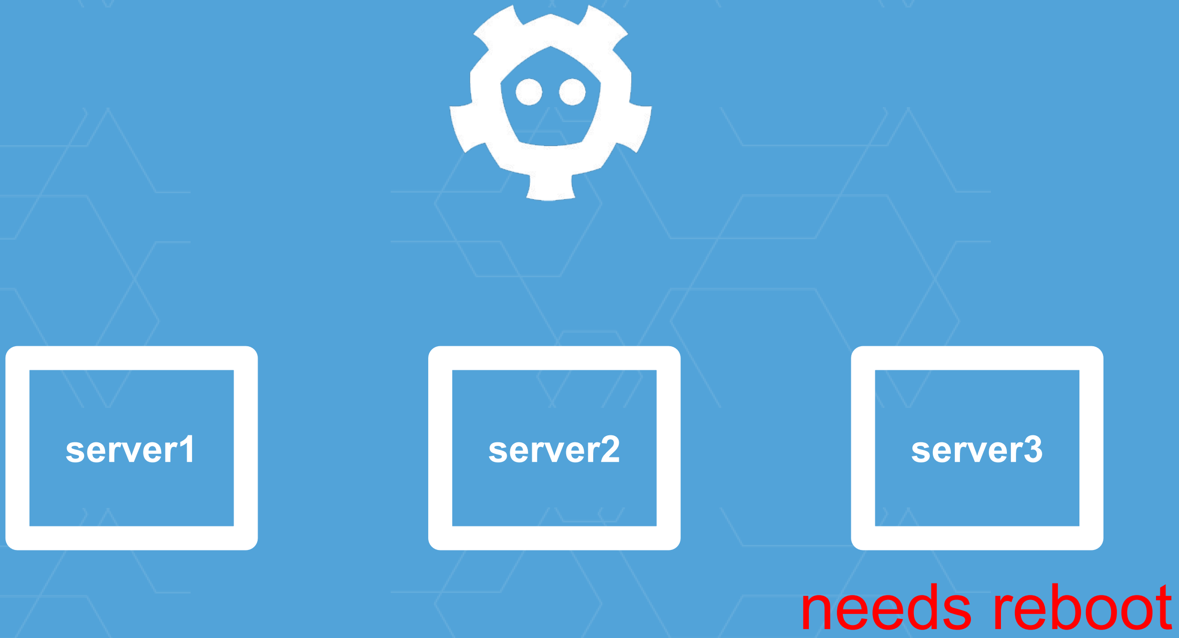
locksmith

- cluster wide reboot lock
  - "semaphore for reboots"
- CoreOS updates happen *automatically*
  - prevent all the machines restarting at once...

Cluster Wide Reboot Lock



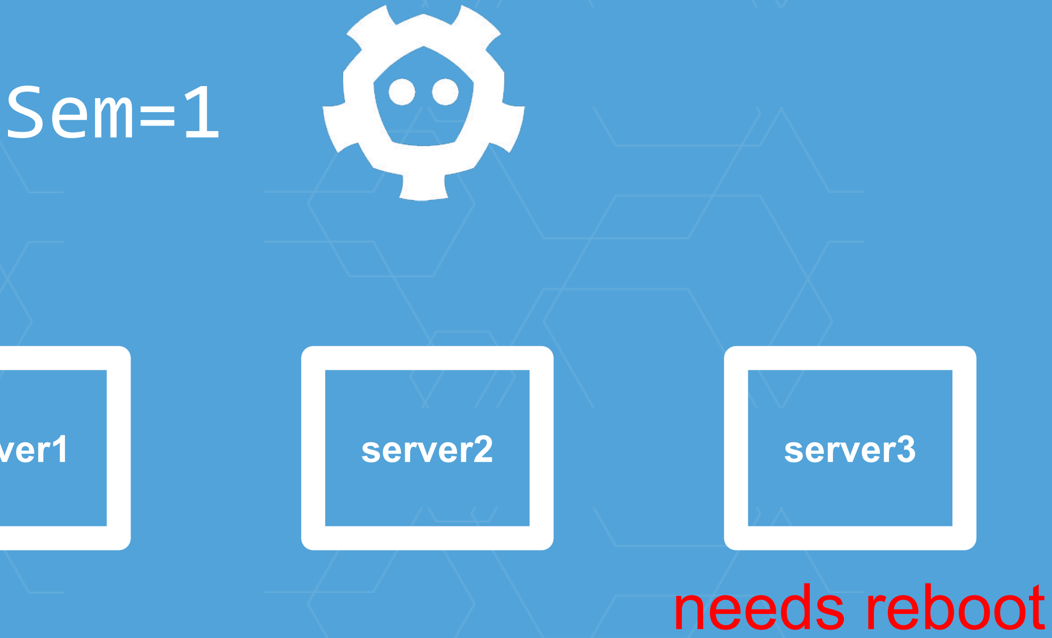
Cluster Wide Reboot Lock



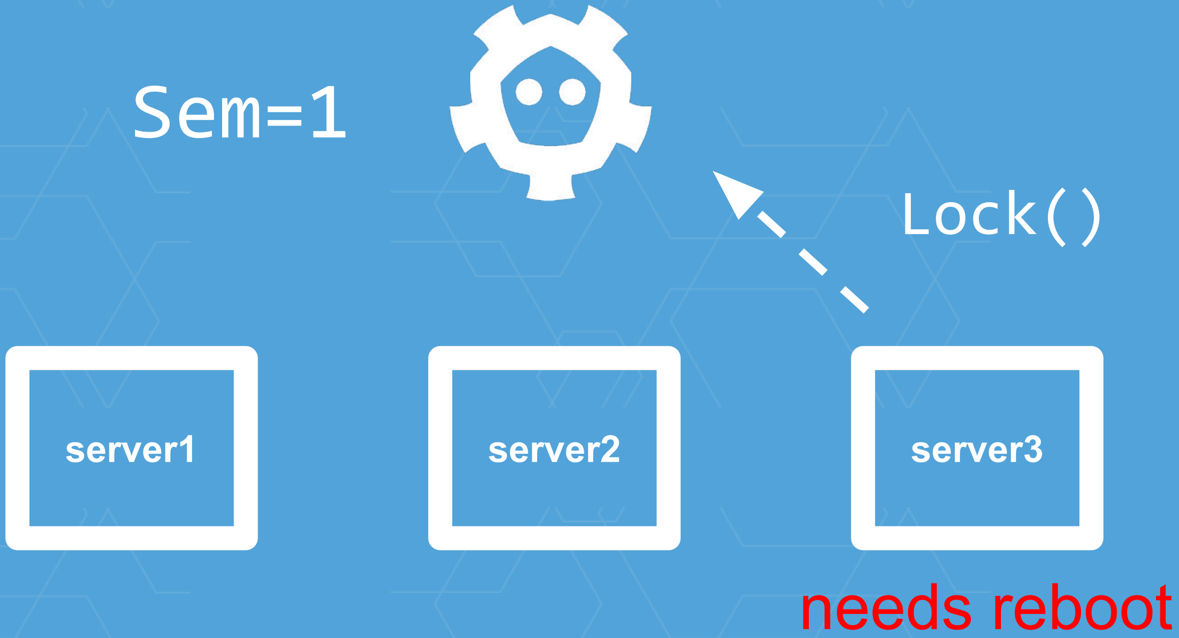
Cluster Wide Reboot Lock

- Need to reboot? Decrement the semaphore key (*atomically*) with etcd
- `manager.Reboot()` and wait...
- After reboot, increment the semaphore key in etcd (*atomically*)

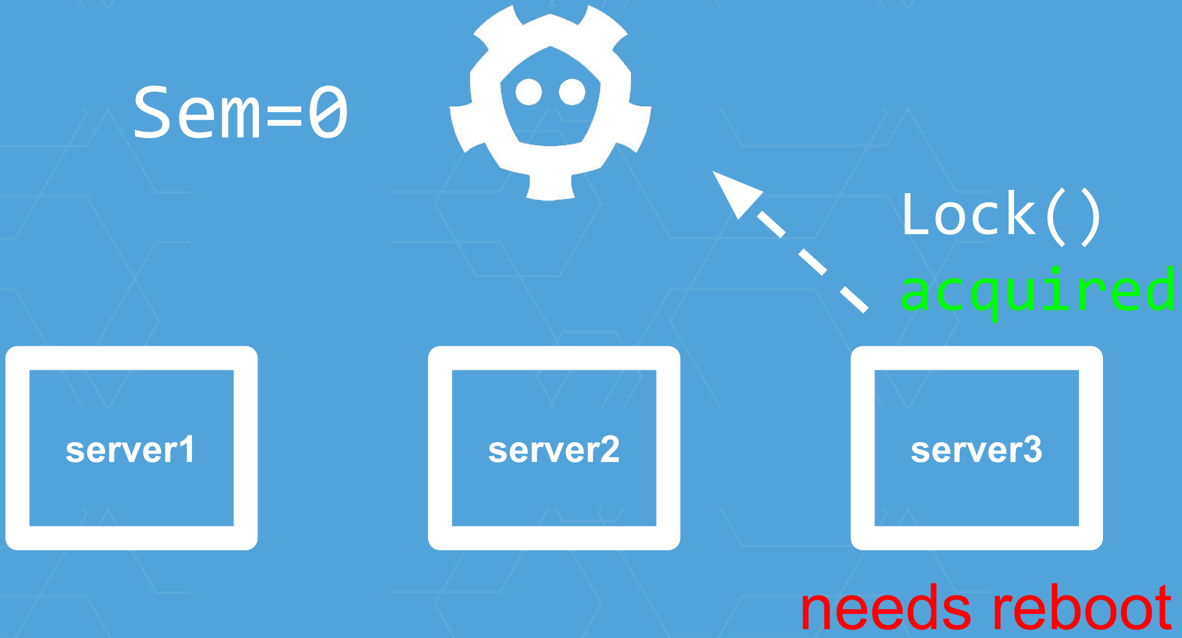
Cluster Wide Reboot Lock



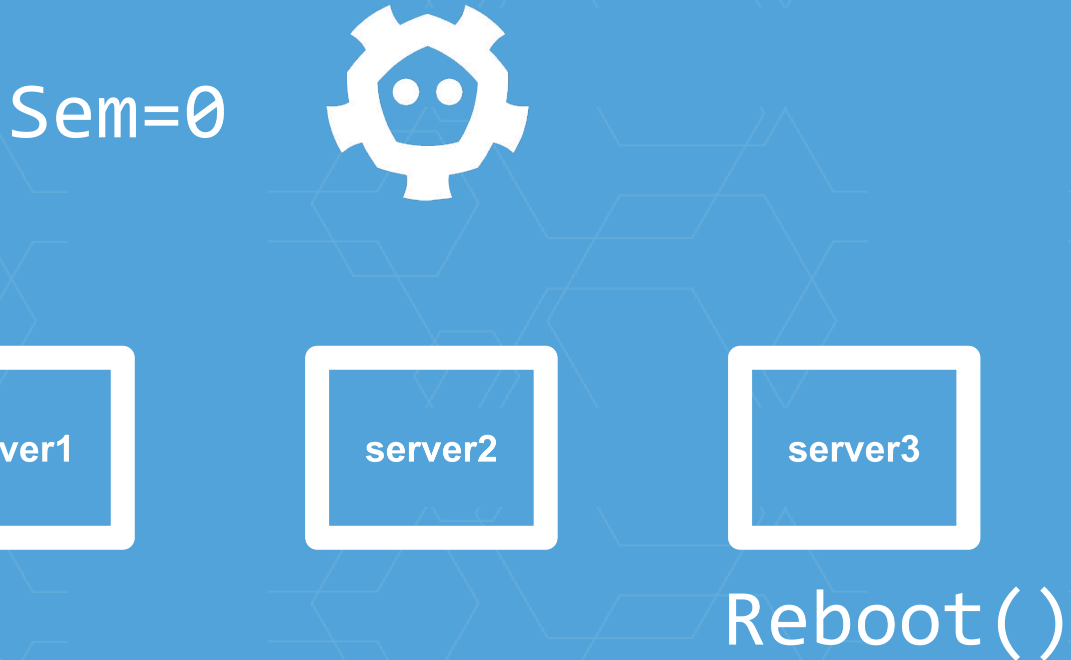
Cluster Wide Reboot Lock



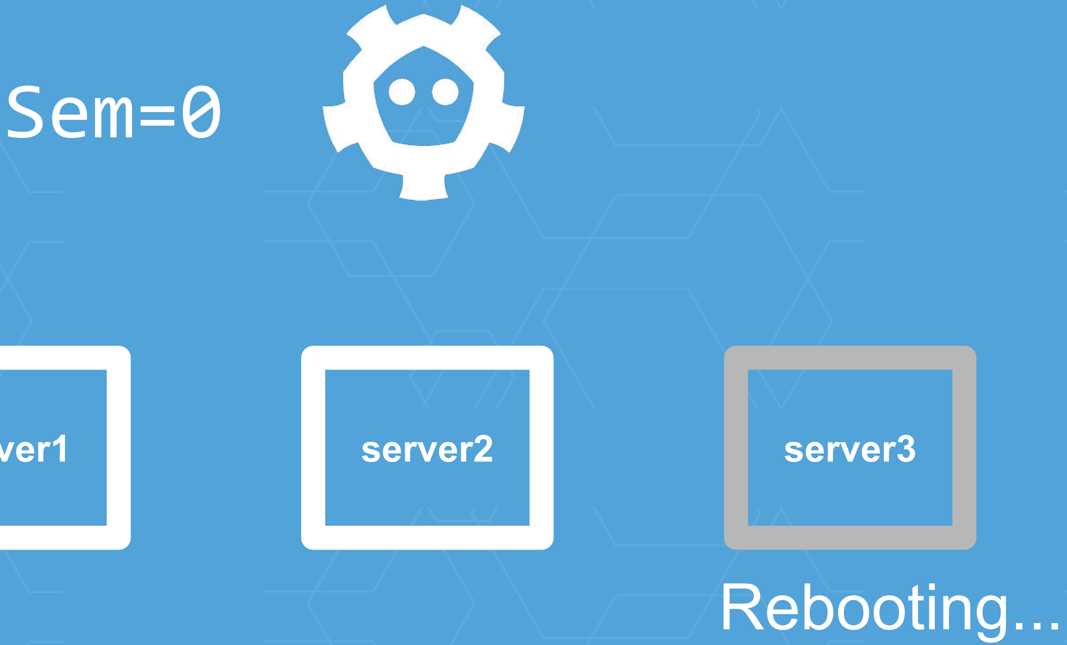
Cluster Wide Reboot Lock



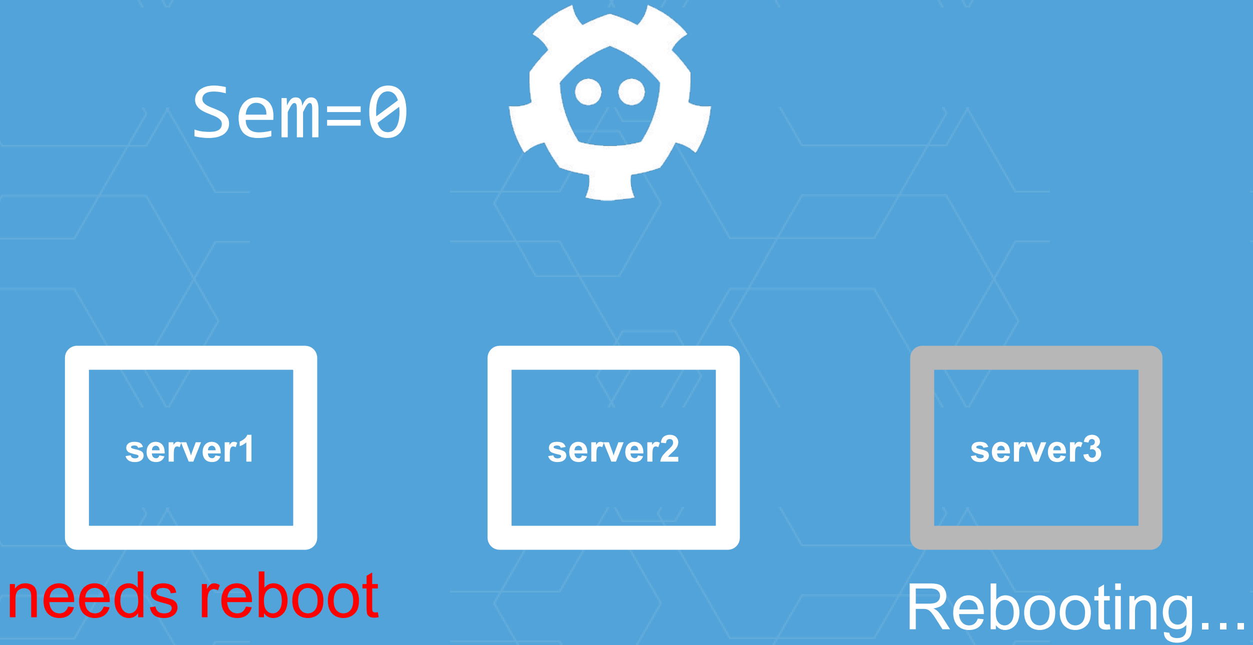
Cluster Wide Reboot Lock



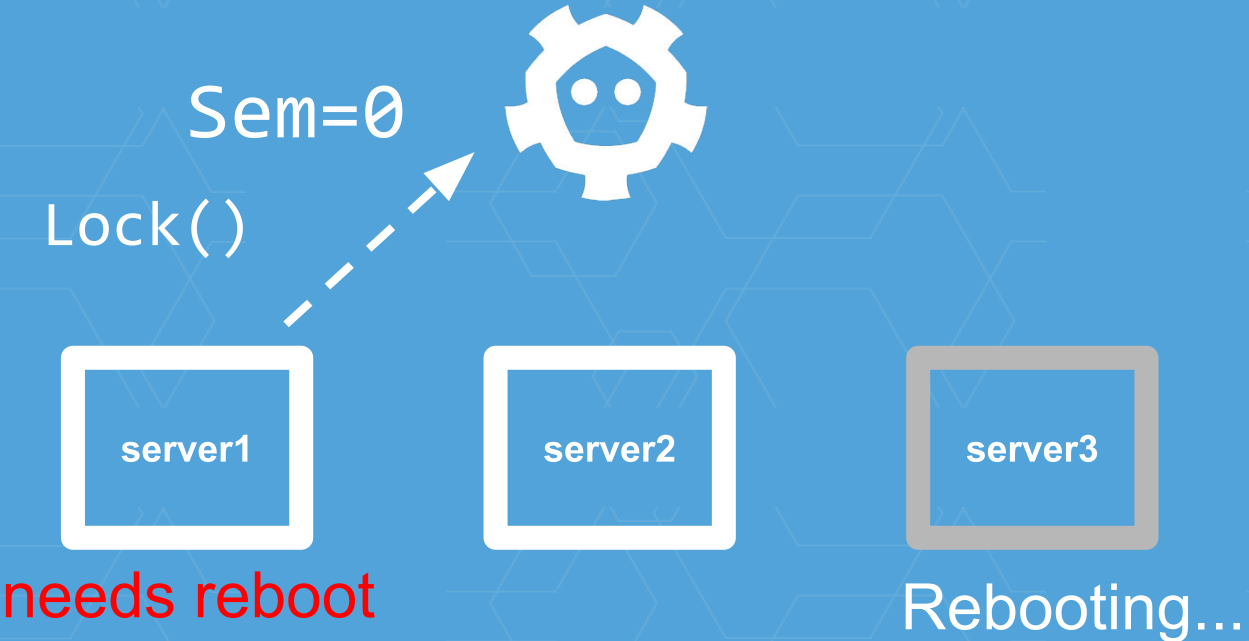
Cluster Wide Reboot Lock



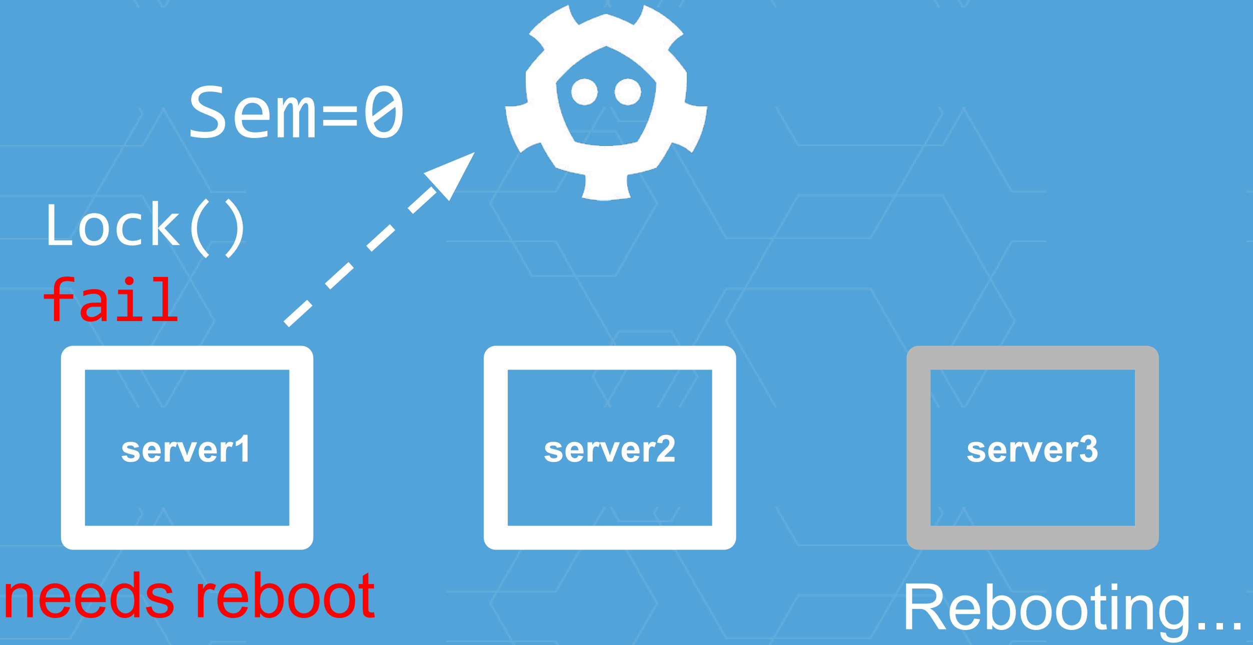
Cluster Wide Reboot Lock



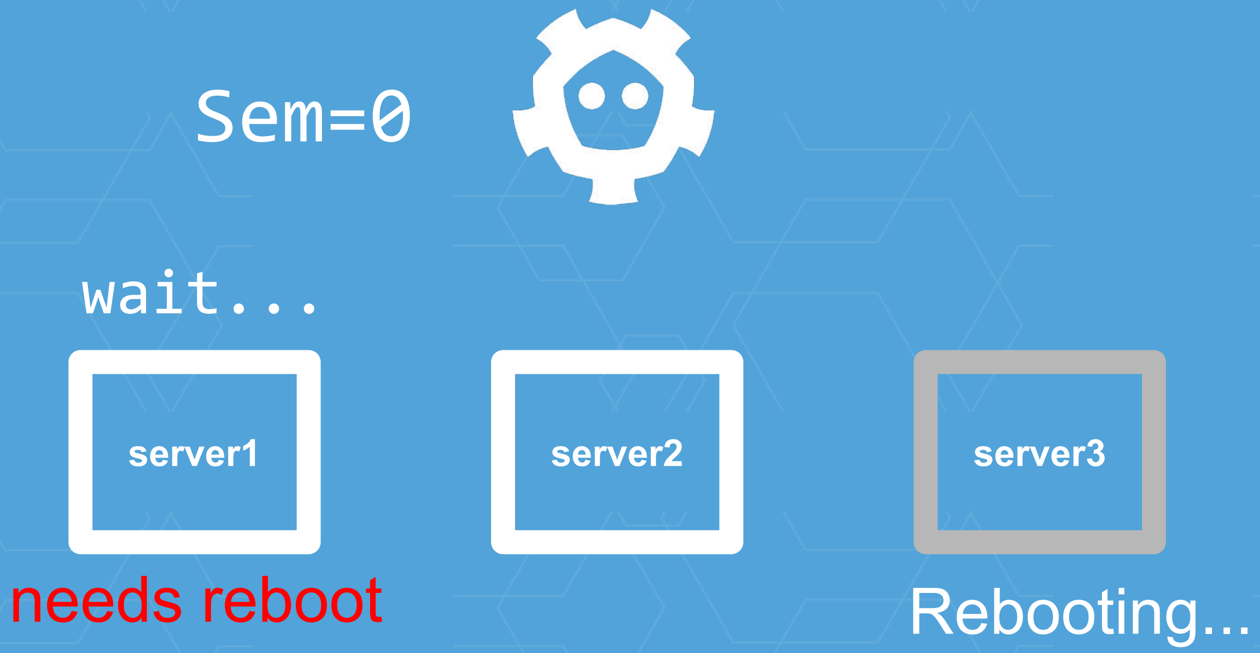
Cluster Wide Reboot Lock



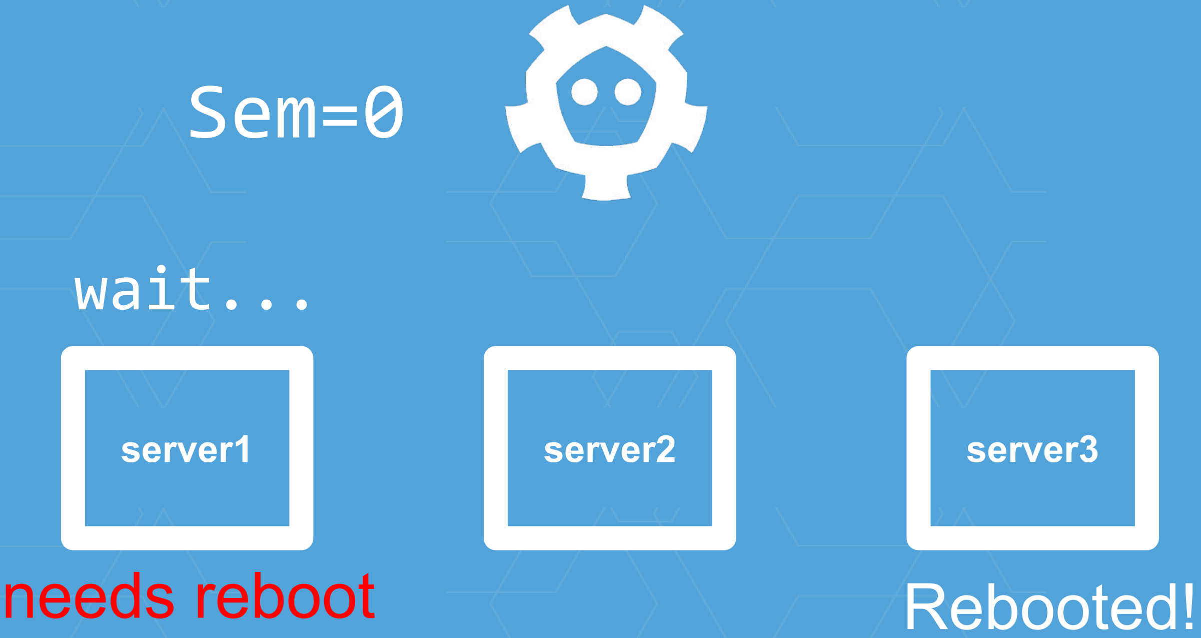
Cluster Wide Reboot Lock



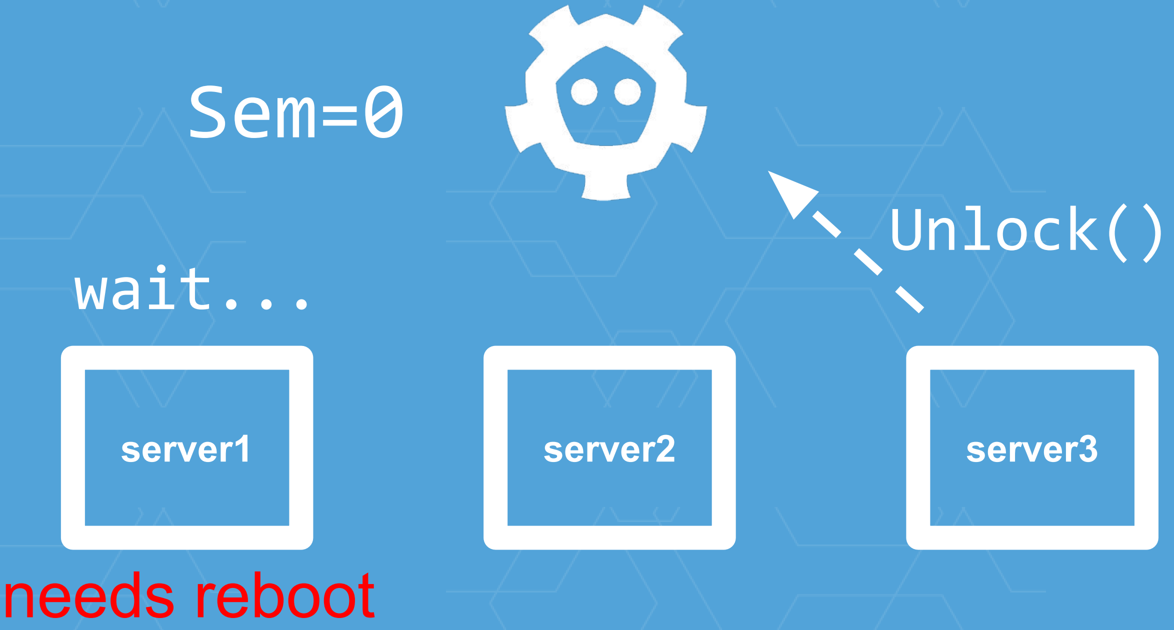
Cluster Wide Reboot Lock



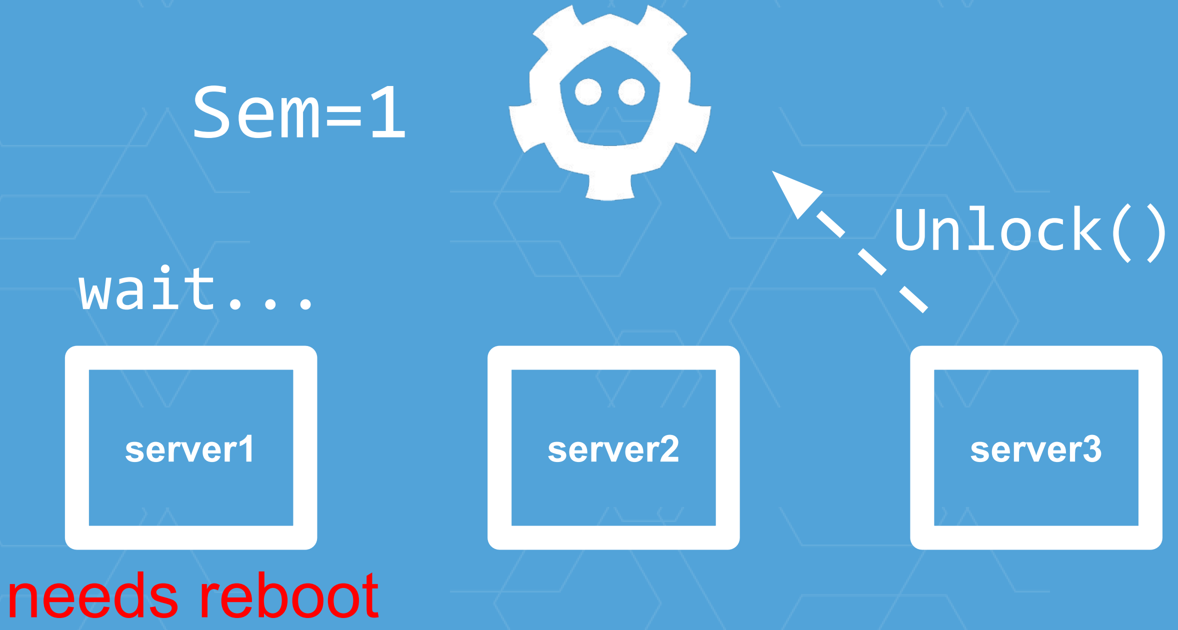
Cluster Wide Reboot Lock



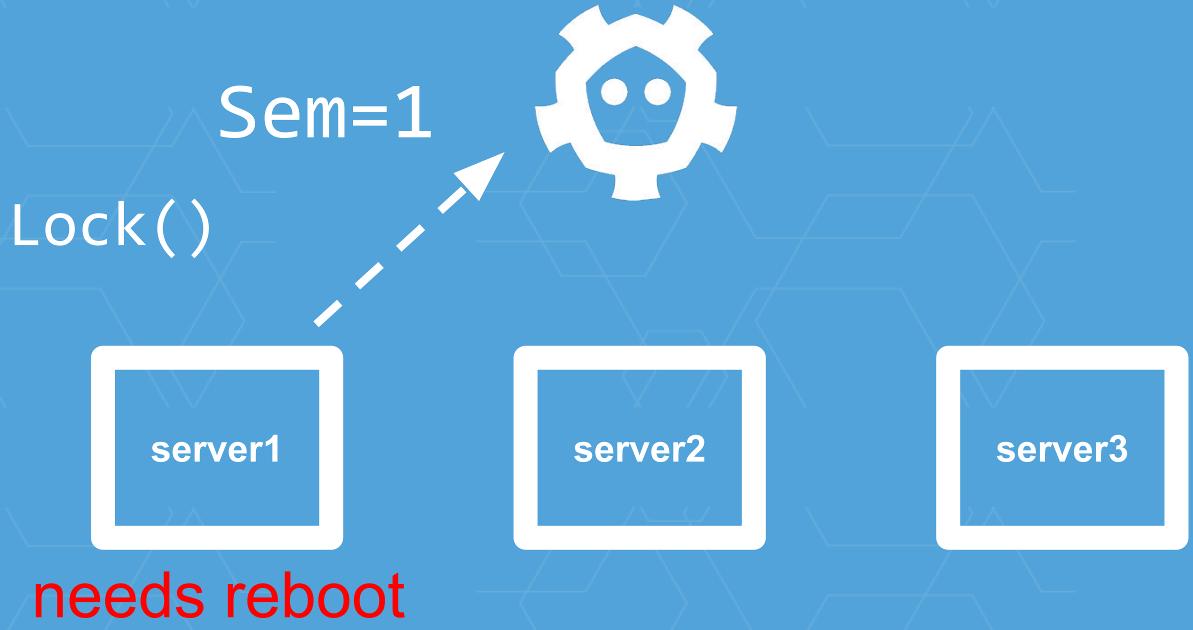
Cluster Wide Reboot Lock



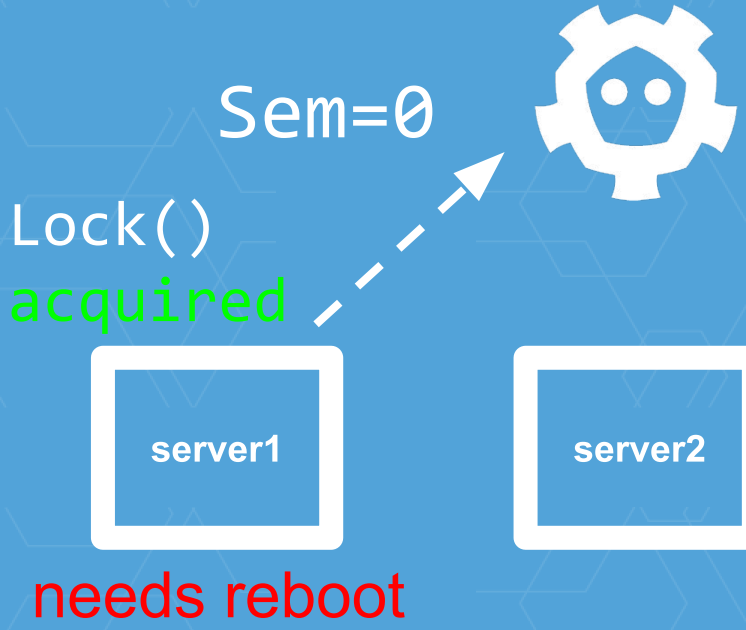
Cluster Wide Reboot Lock



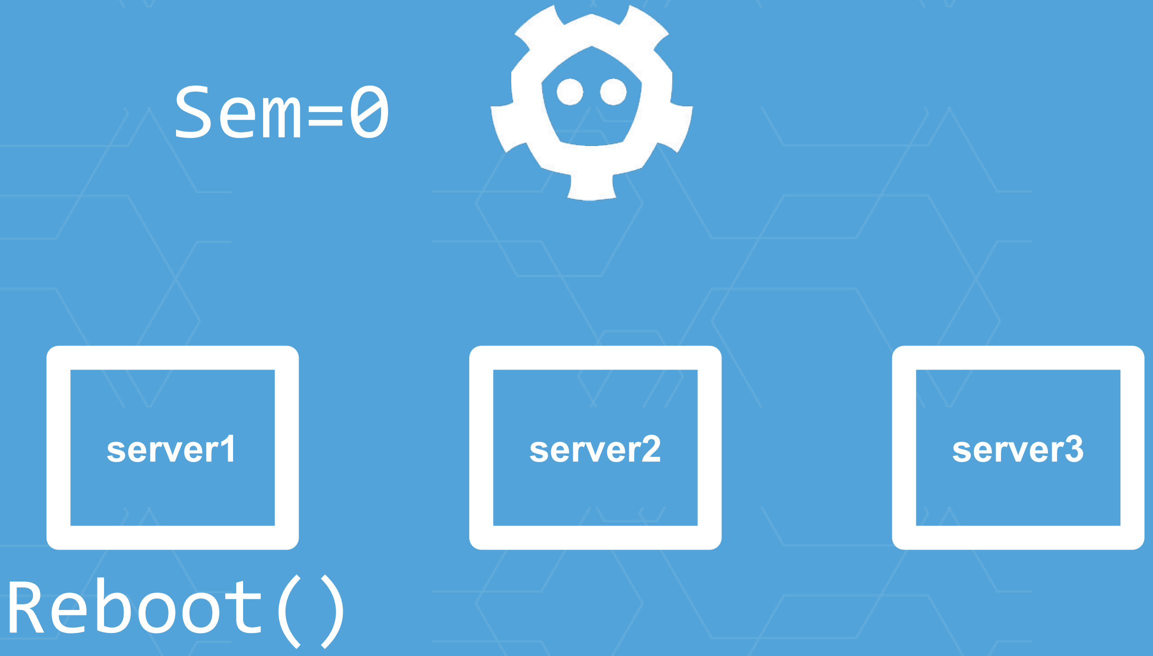
Cluster Wide Reboot Lock



Cluster Wide Reboot Lock



Cluster Wide Reboot Lock



etcd apps

flannel

flannel

- virtual (overlay) network, providing a subnet to each host, and handling all routing
- for use with Kubernetes etc
- uses etcd to store network configuration, allocated subnets, and auxiliary data (host's IP).

etcd apps

skydns

skydns

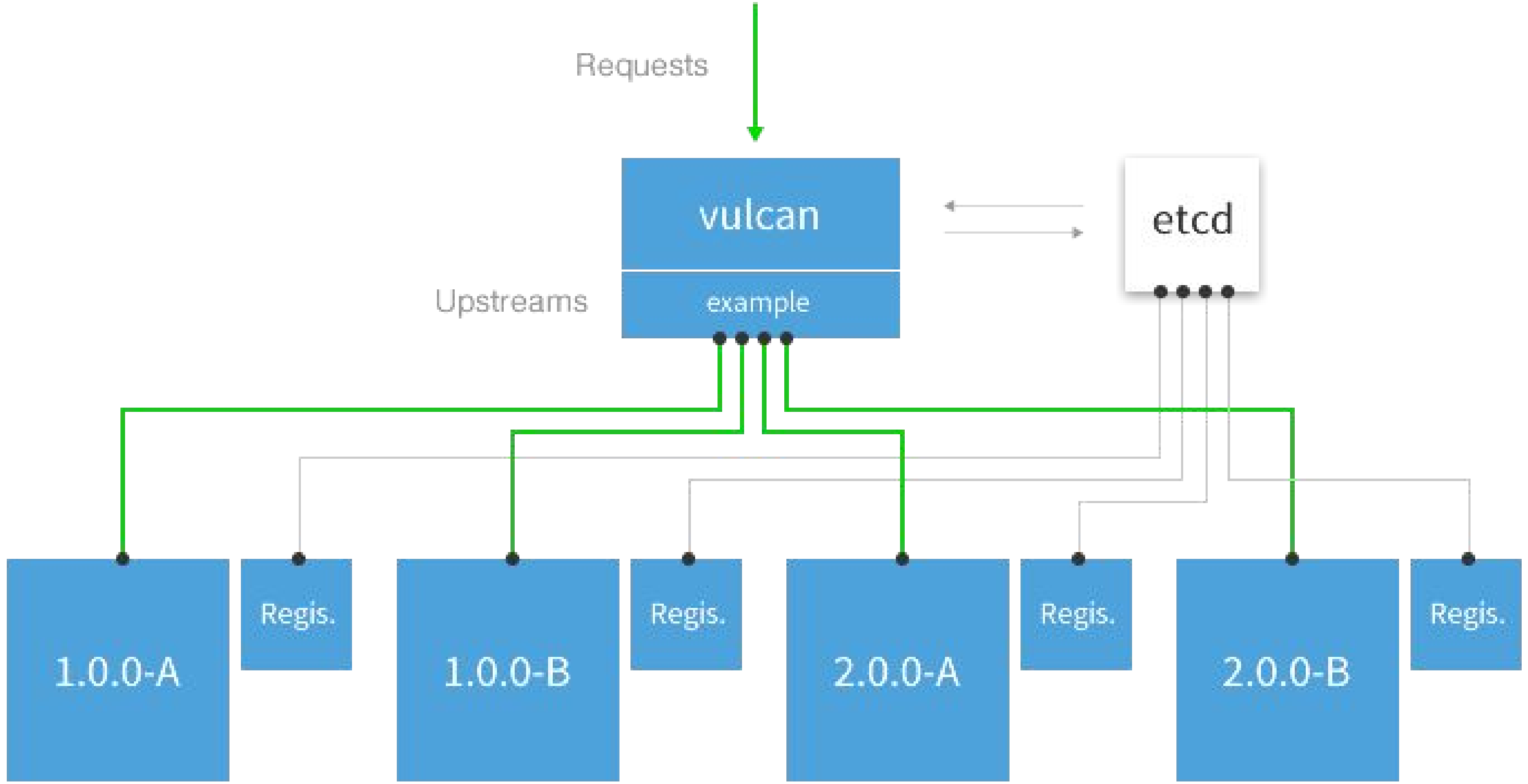
- Service discovery and DNS server
- backed by etcd for all configuration/records

etcd apps

vulcand

vulcand

- "programmatic, extendable proxy for microservices"
- HTTP load balancer
- etcd for storing all configuration


Requests
vulcan
Upstreams
example
etcd
1.0.0-A
Regis.
1.0.0-B
Regis.
2.0.0-A
Regis.
2.0.0-B
Regis.

etcd apps

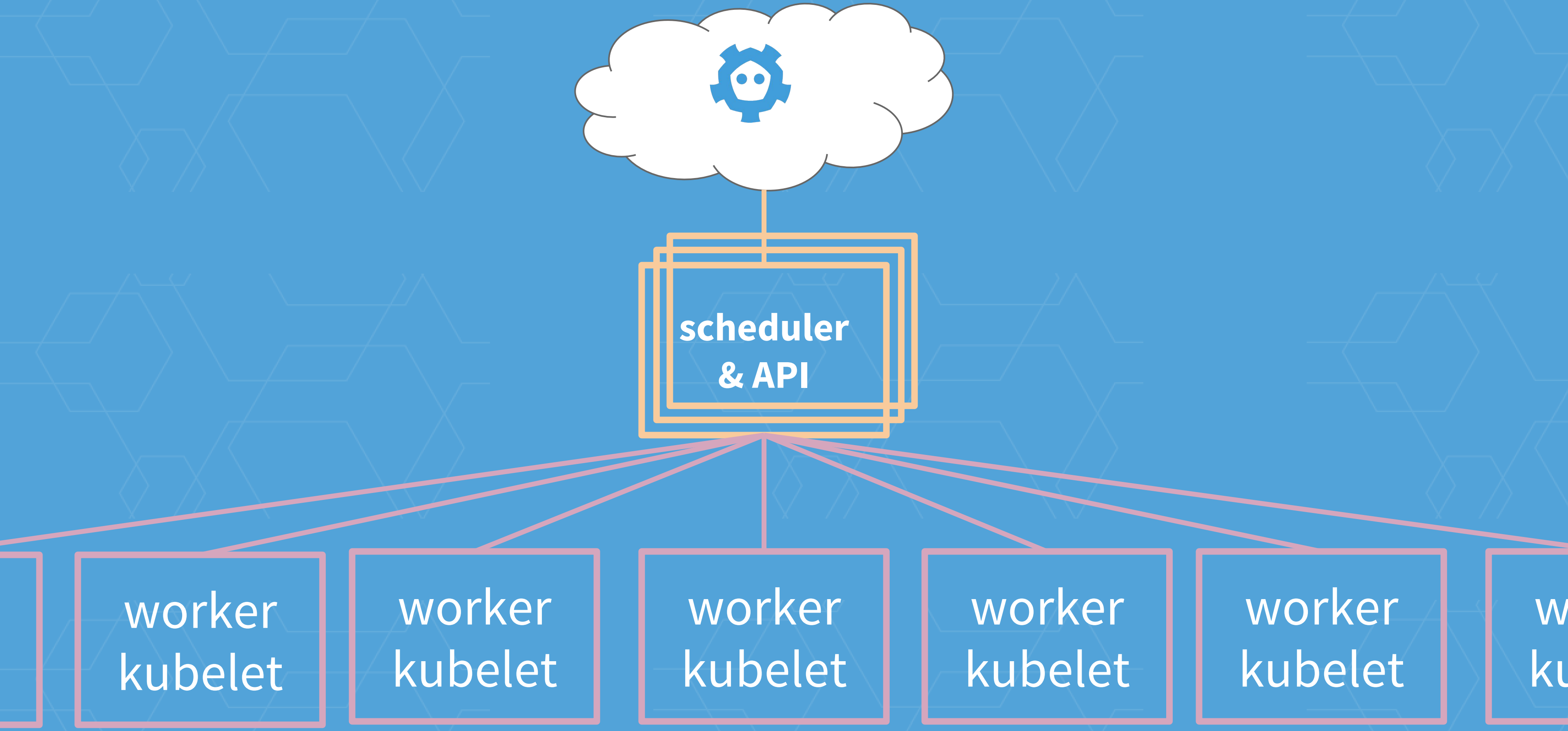
confd

confd

- simple configuration templating
- for "dumb" applications
- watch etcd for changes, render templates with new values, reload applications

etcd apps

Kubernetes


scheduler
& API
worker
kubelet
worker
kubelet
worker
kubelet
worker
kubelet
worker
kubelet

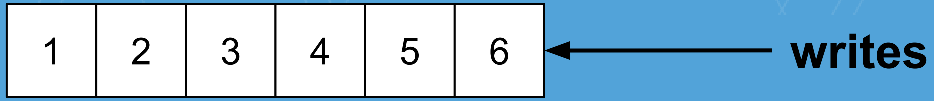
Scaling etcd

Scaling etcd to the next level

- Recent improvements
    - Asychronous snapshots
    - Request pipelining
- Future improvements
    - v3 and beyond

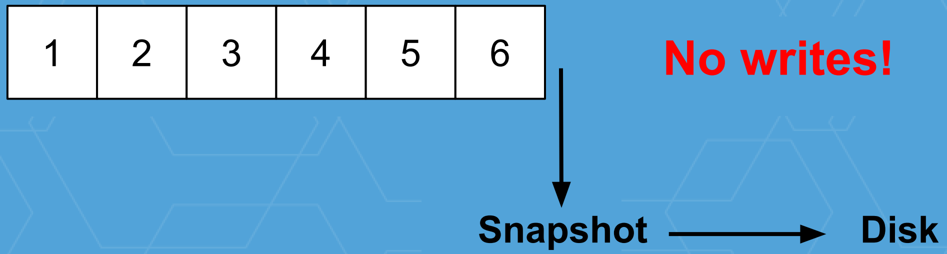
Recent improvements (v2)

- Asynchronous snapshotting
  - append-only log-based system
    - with more and more writes, will grow indefinitely...
  - so, snapshot, purge log



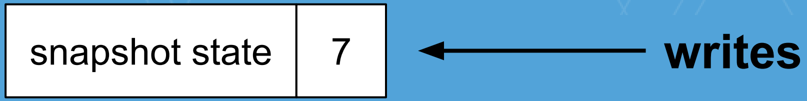
Recent improvements (v2)

- Asynchronous snapshotting
  - before: stop the world, save snapshot to disk
  - problem: this blocks all writes
  - can only resume writing when snapshot is done



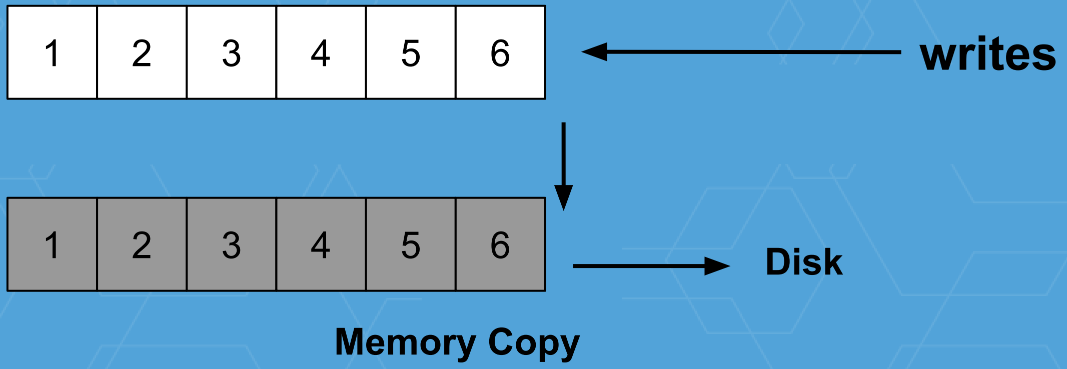
Recent improvements (v2)

- Asynchronous snapshotting
  - before: stop the world, save snapshot to disk
  - problem: this blocks all writes
  - can only resume writing when snapshot is done



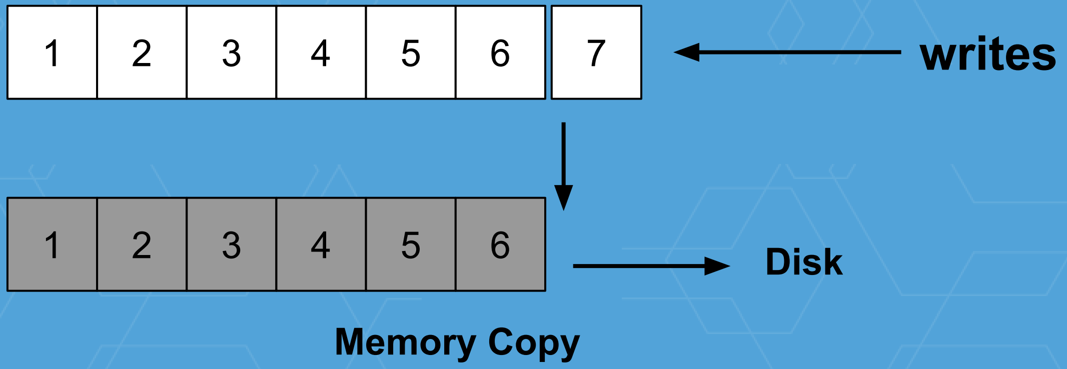
Recent improvements (v2)

- Asynchronous snapshotting
  - now: in-memory copy, write copy to disk
  - can continue serving writes while snapshotting



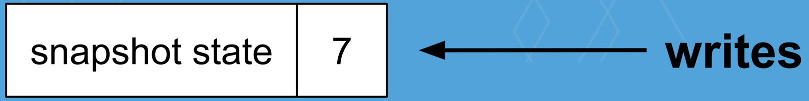
Recent improvements (v2)

- Asynchronous snapshotting
  - now: in-memory copy, write copy to disk
  - can continue serving writes while snapshotting



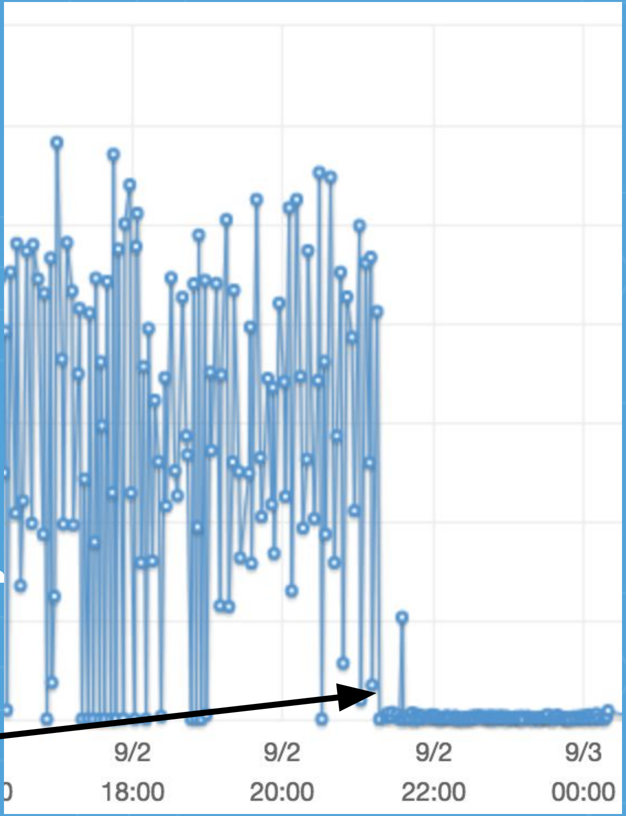
Recent improvements (v2)

- Asynchronous snapshotting
    - now: in-memory copy, write copy to disk
    - can continue serving writes while snapshotting



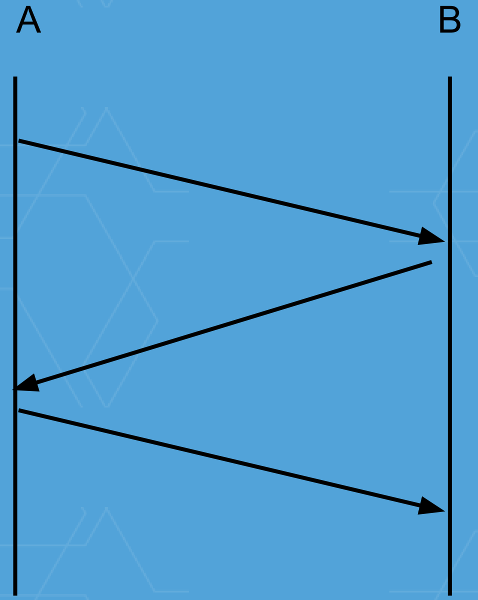
Recent improvements (v2)

- Asynchronous snapshotting in the real world
    - Discovery service (production etcd cluster serving https://discovery.etcd.io)
    - *`500 Internal Server Error`* spikes dropped to zero



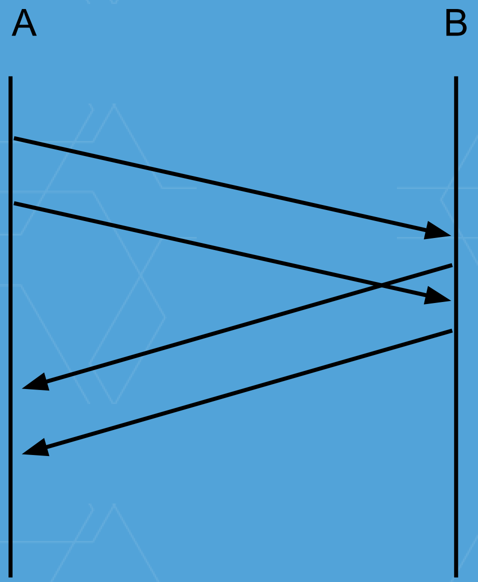
Recent improvements (v2)

- raft pipelining
  - raft is based around a series of RPCs (e.g. `AppendEntry`)
  - etcd previously used *synchronous RPCs*
  - send next message only after getting previous reply
  - safe, but low throughput



Recent improvements (v2)

- raft pipelining
  - etcd now uses *RPC pipelining*
  - send series of messages without waiting for replies
  - optimistic, recover from failures



Future improvements (v3)

"Scaling etcd to thousands of nodes"

etcd v3.0

- Efficient and powerful API
- Disk-backed storage
- Incremental snapshots

The future (v3)

- Efficient and powerful API
  - flat binary key space
  - multi-object transaction
  - native leasing API
  - native locking API
  - gRPC (HTTP2 + protobuf)

Key space

- Flat binary key-value space
  - `coreos=awesome`
  - `coreos/etcd=kv`
  - `coreos/rkt=container`
- Keep it as simple as possible
  - want hierarchy?
    - build your own layer on top of kv

v3 API

- Put
  - foo=bar
- Get
- Range (consistent multi-get)
  - single key: foo
  - prefix: foo->fop (exclude)
  - range: foo->foo1
- Delete Range
  - same as range

KV API

```
KV.Put("foo", "bar")
KV.Get("foo")
KV.Range("foo", "foo10")
KV.Delete("foo")
KV.DeleteRange("foo", "foo10")
```

v3 API

- Mini transaction
  - two phases
    - *compare*
    - *execution* (either *success* or *failure*)
  - compare on value, index, etc.
  - execute a list of basic operations

v3 API

- Mini transaction
    - compare and swap
        - *compare*: foo=bar
        - *success*: foo=bar2
    - multiple object transaction
        - *compare*: cond1=true && cond2=true
        - *success*: pass=true
        - *failure*: pass=false

Mini Transaction

```
Tx.If(
    Compare(Value("foo"), ">", "bar"),
    Compare(Version("foo"), "=", 2),
    ...
).Then(
    OpPut("ok","true")...
).Else(
    OpPut("ok","false")...
).Commit()
```

v3 API

- Watch
  - support multiple keys and prefixes per stream
    - `watchKey(foo)`
    - `watchPrefix(coreos)`
  - support watch from historical point
    - `watchKey(foo, index_of_an_hour_ago)`
    - user-driven history compaction

gRPC

- Efficient
  - multiple streams share one TCP connection
  - compacted encoding format (protobuf)
- Rich generated libraries in tens of languages
  - Go, Java, Python, C++...

The future (v3)

- Incremental snapshot
  - only save the delta instead of full data set
  - less I/O and CPU cost per snapshot
  - no bursty resource usage, more stable performance

The future (v3)

- Disk backend
  - keep the cold historical data on disk
  - keep the hot data in memory
  - support "entire history" watches
  - user-facing compaction API

The future (v3)

- Upstream recipes for common usage patterns
  - Leases: attaching *ownership* to keys
  - Leader election
  - Locking resources

Lease

```
l := lease.Create(10*second)

kv.Put("foo", "bar", l.ID)

// key will be removed without keeping
// alive the lease
go KeepAlive(l.id)
```

Watch

```
events, err := watcher.Watch("foo")
if err != nil {
  // handle error
}
for r := <- responses {
  // consume received events
}
```

Recipes

Leader election

- `election.Elect("eFoo"), election.Resign("eFoo")`

Locking

- `locking.Acquire("lFoo"), locking.Release("lFoo")`

FEST
CoreOS

coreos.com/fest

May 9 & 10, 2016 | Berlin, Germany

- Early bird tickets
- Sponsorships are still available
- Submit a talk before February 29th!

Thanks!

etcd

CoreOS

Join us!
github.com/coreos/etcd

We are hiring!
coreos.com/careers

extra slides

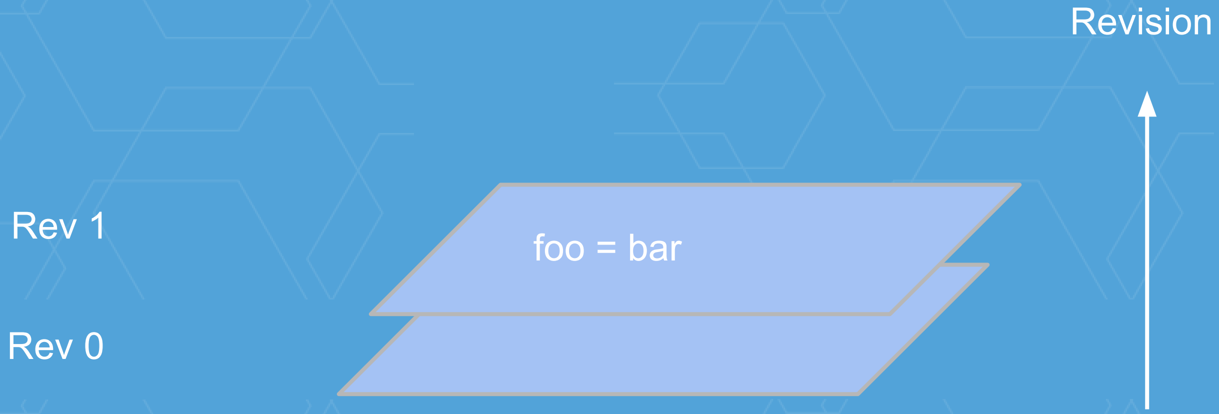
(did I talk too fast?)

MVCC

Revision
Rev 0

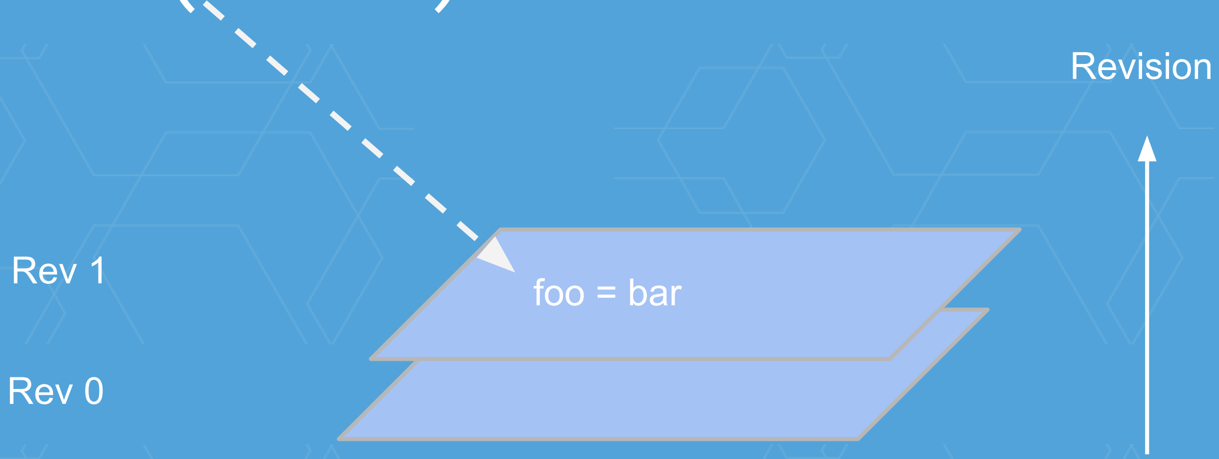
MVCC

```
KV.Put("foo", "bar") -> increase Rev
```



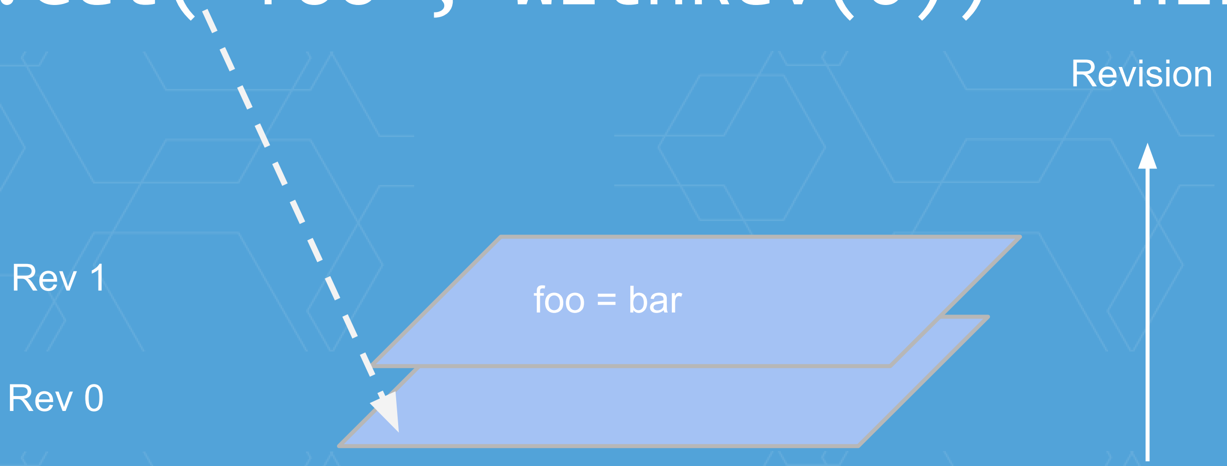
MVCC

```
KV.Get("foo") = "bar"
```



MVCC

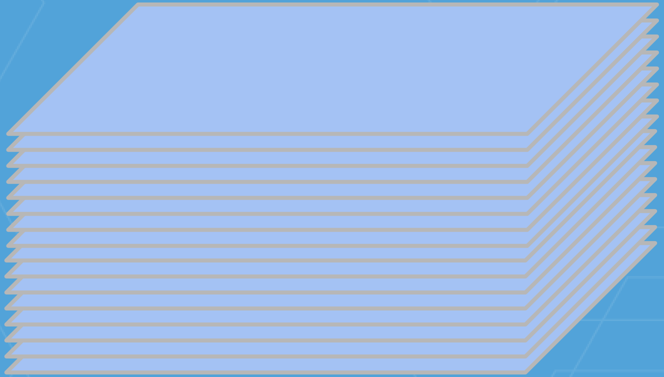
```
KV.Get("foo", WithRev(0)) = nil
```



Compaction

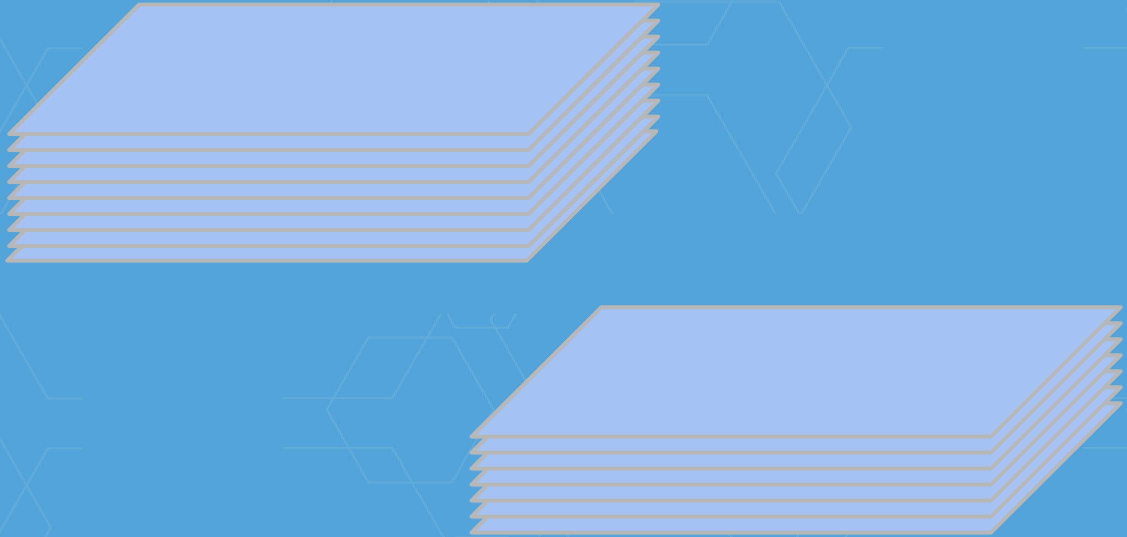
Why?

- Too many revisions



Compaction

Analyse the old revisions to be compacted



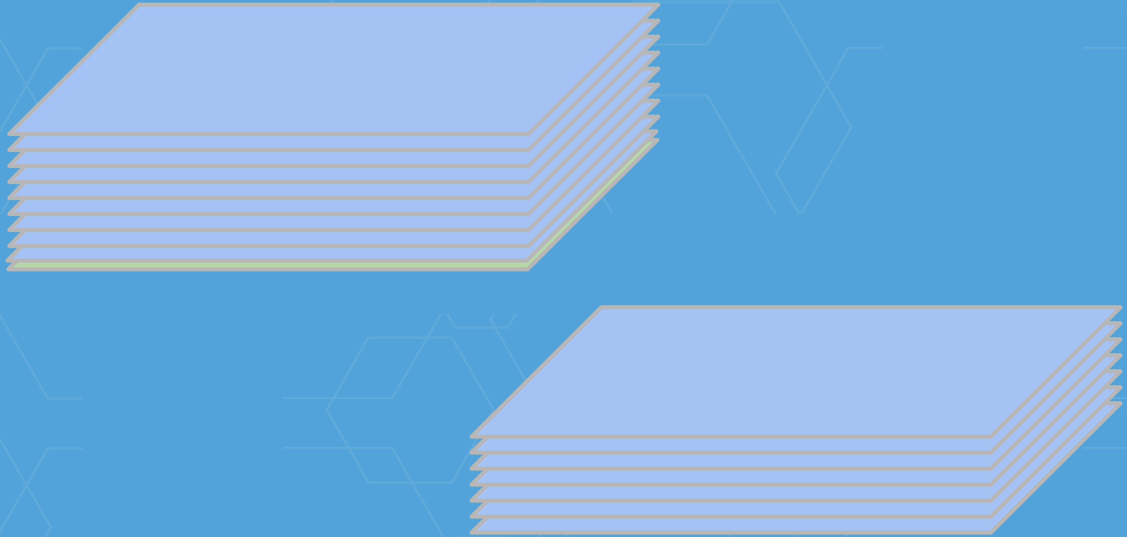
Compaction

Rule 1: the key with tombstone can be removed

Rule 2: keep the latest version of a non-tombstone key

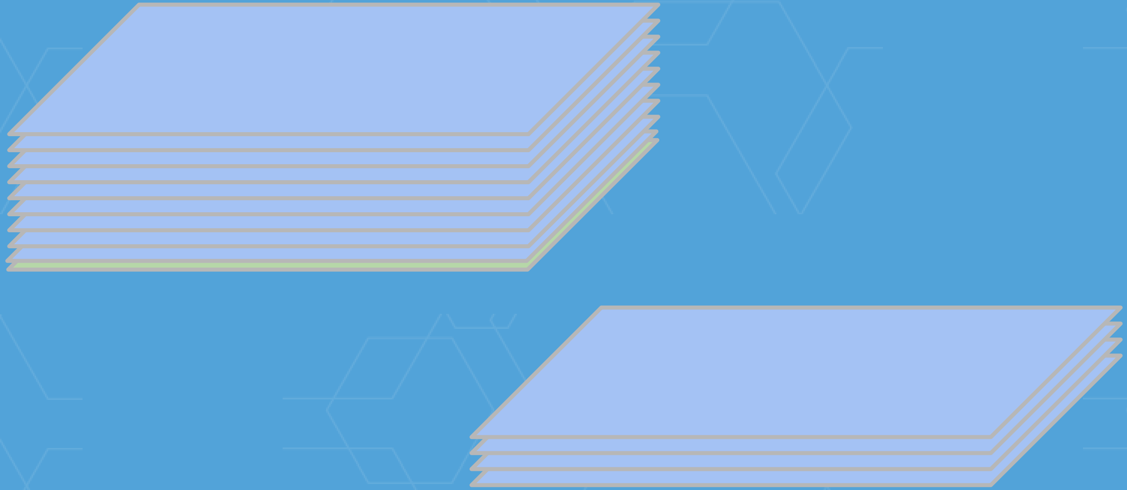
Compaction

Clean up old revisions in background



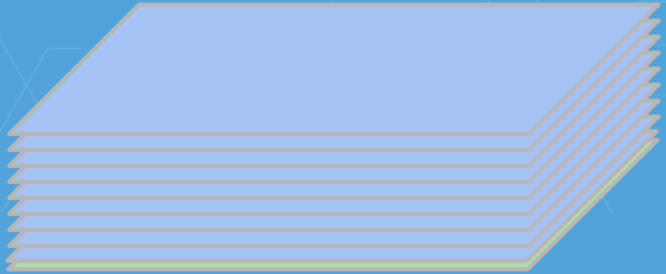
Compaction

Clean up old revisions in background



Compaction

Done!

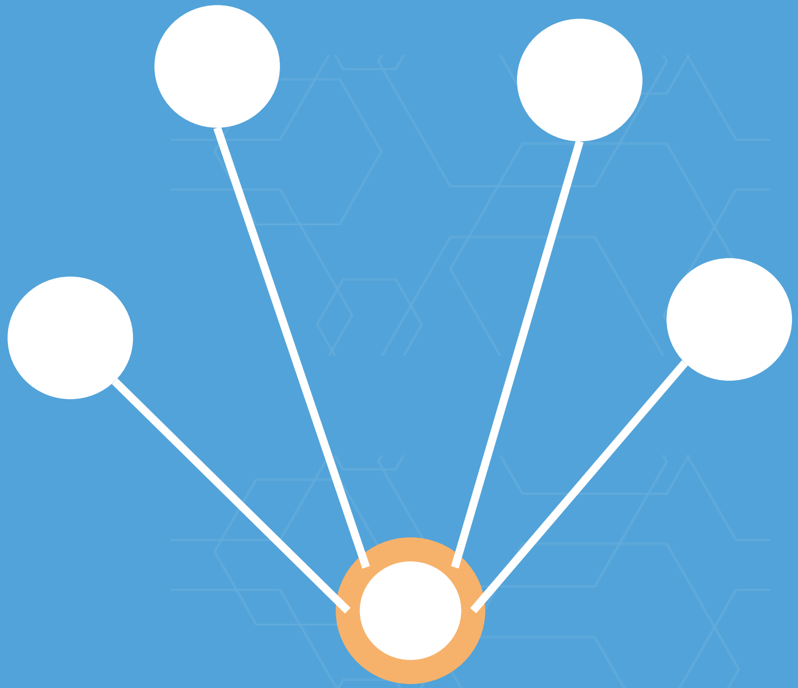


etcd basics

fault tolerance

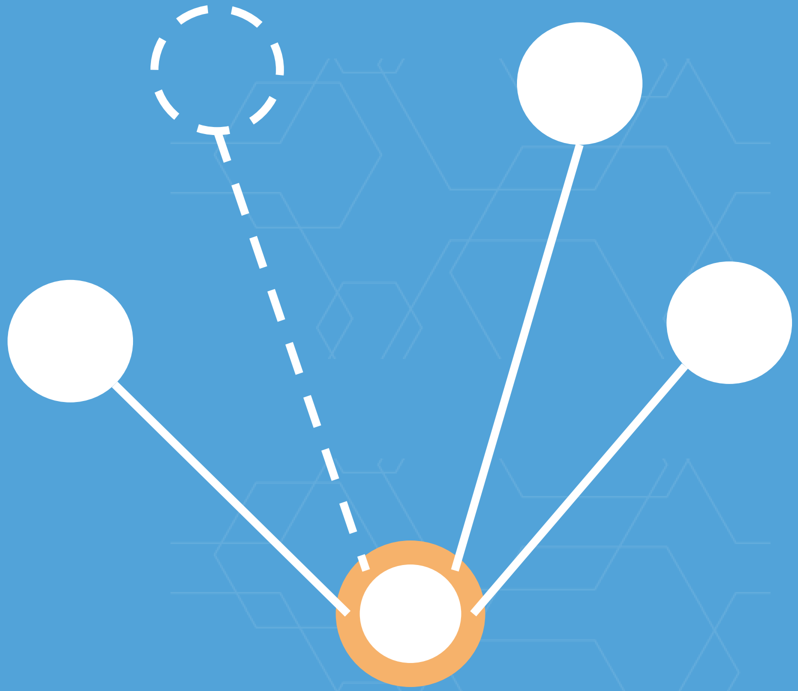
Available

Leader
Follower



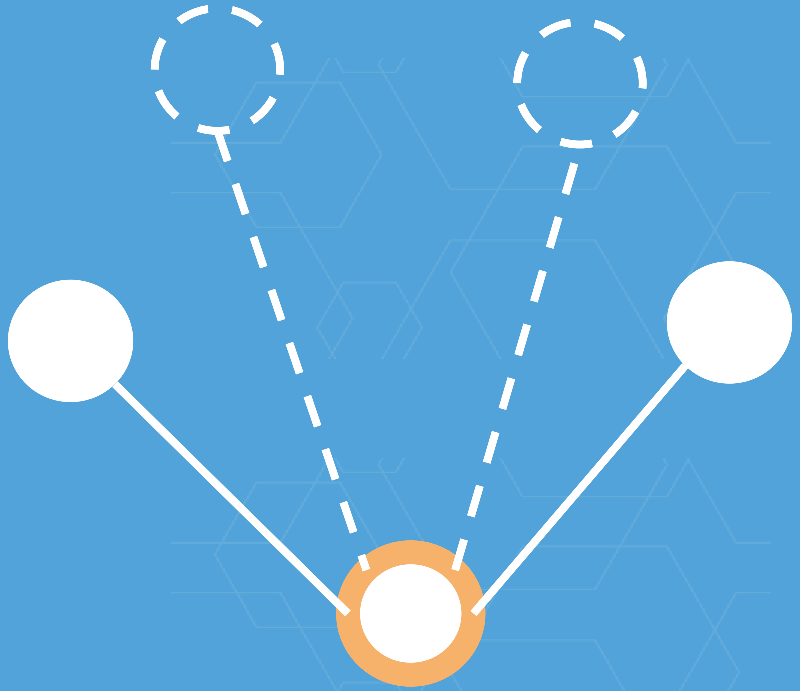
Available

Leader
Follower

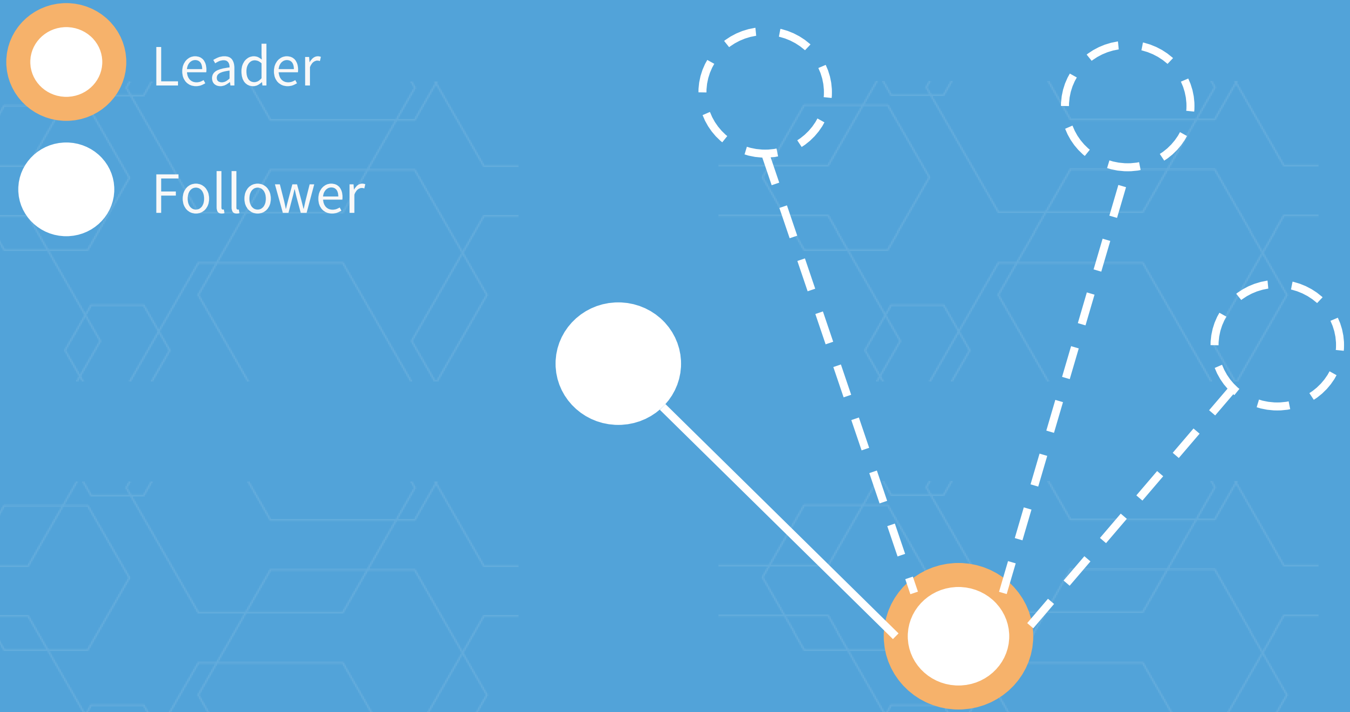


Available

Leader
Follower



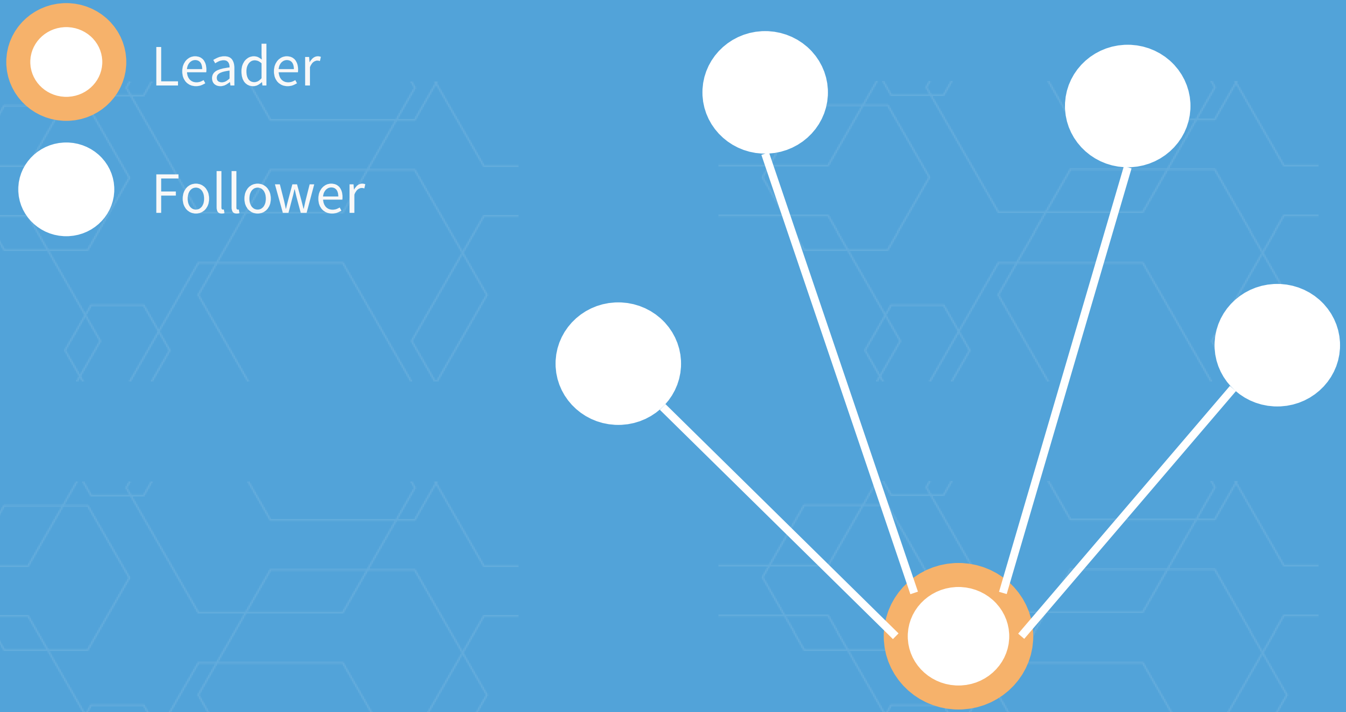
Unavailable



etcd basics

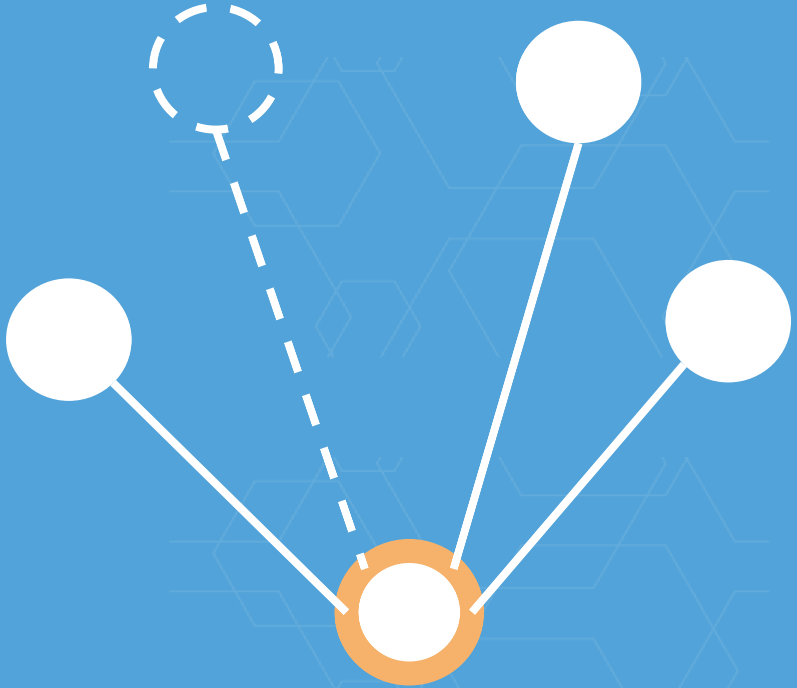
leader fault tolerance

Available



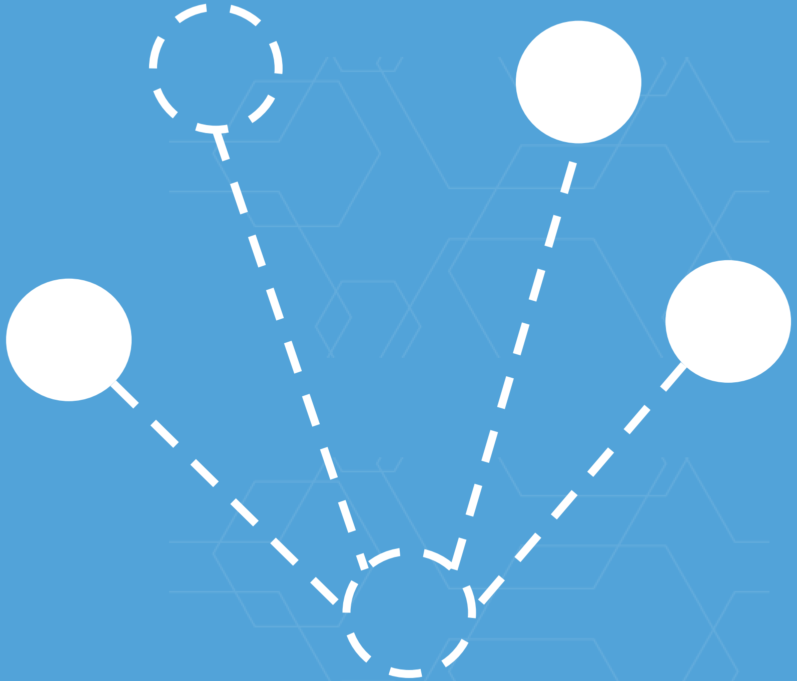
Available

Leader
Follower



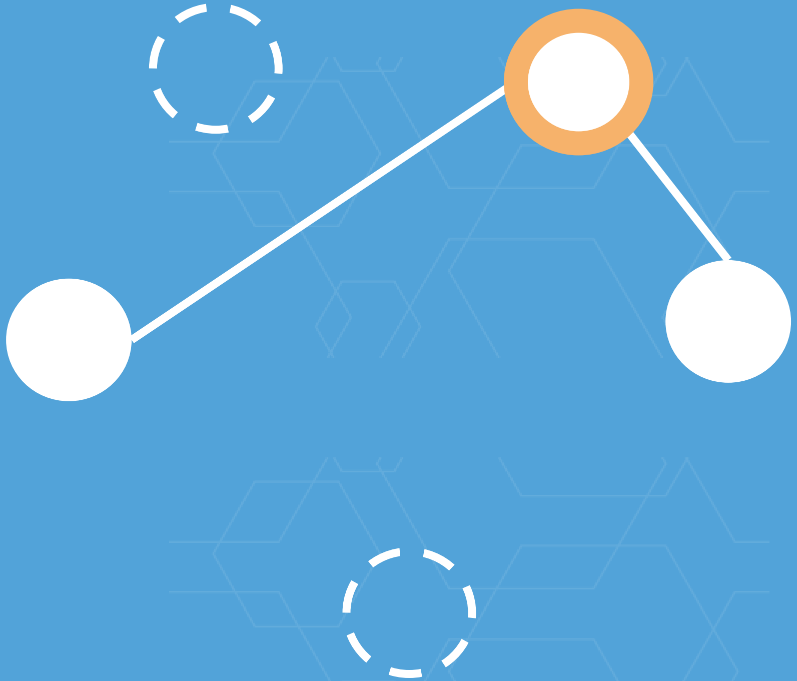
Temporarily Unavailable

Leader
Follower

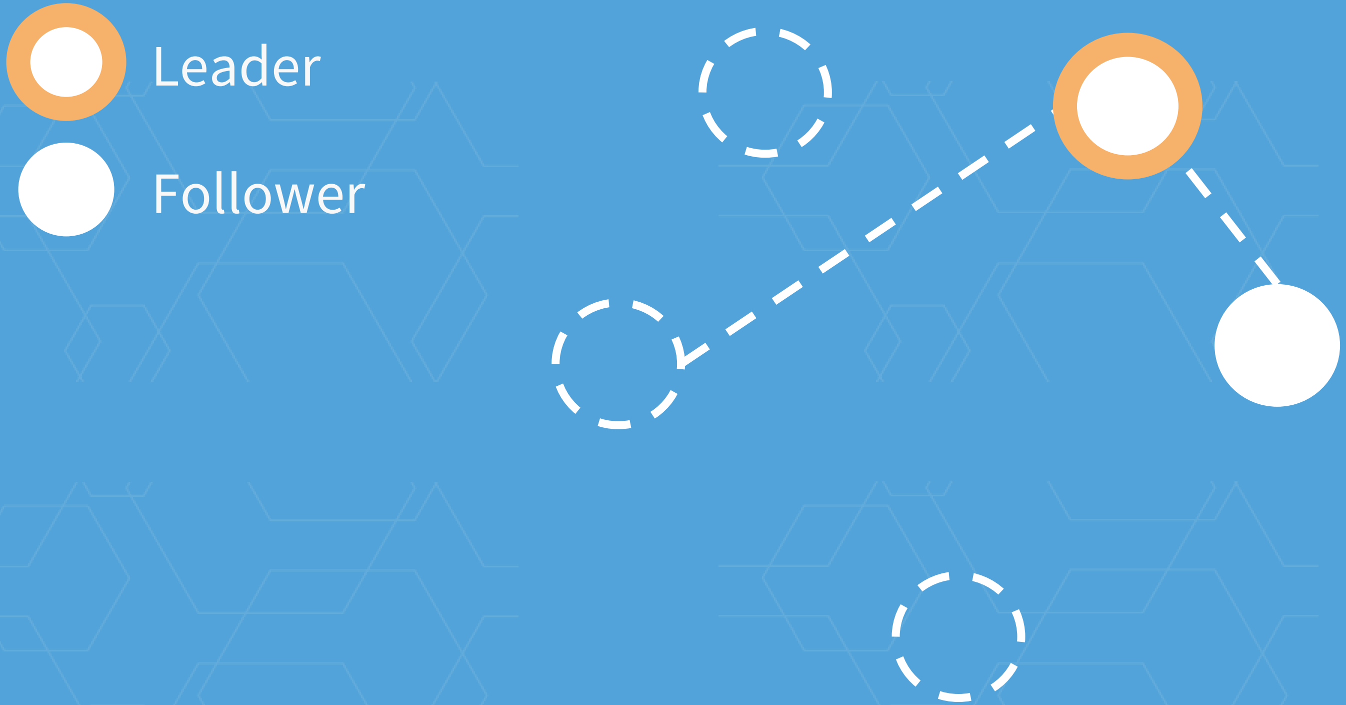


Available

Leader
Follower



Unavailable



etcd durability

wal, snapshots, backups

etcd bootstrap

discovery, static

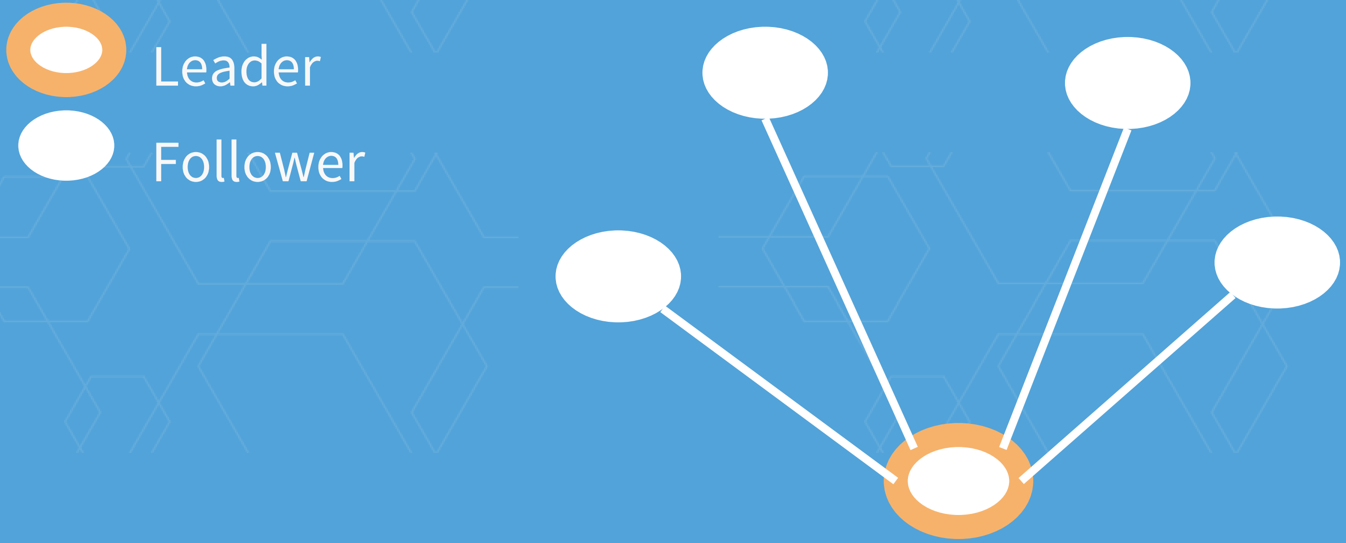
```
$ curl discovery.etcd.io/new?size=5
https://discovery.etcd.io/6eadeac2...
```

discovery

discovery

discovery

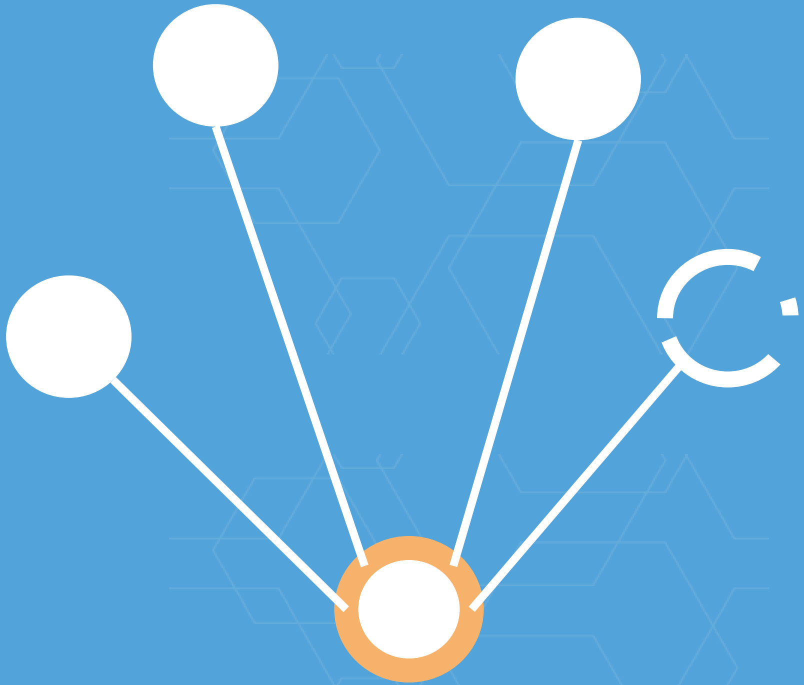
discovery

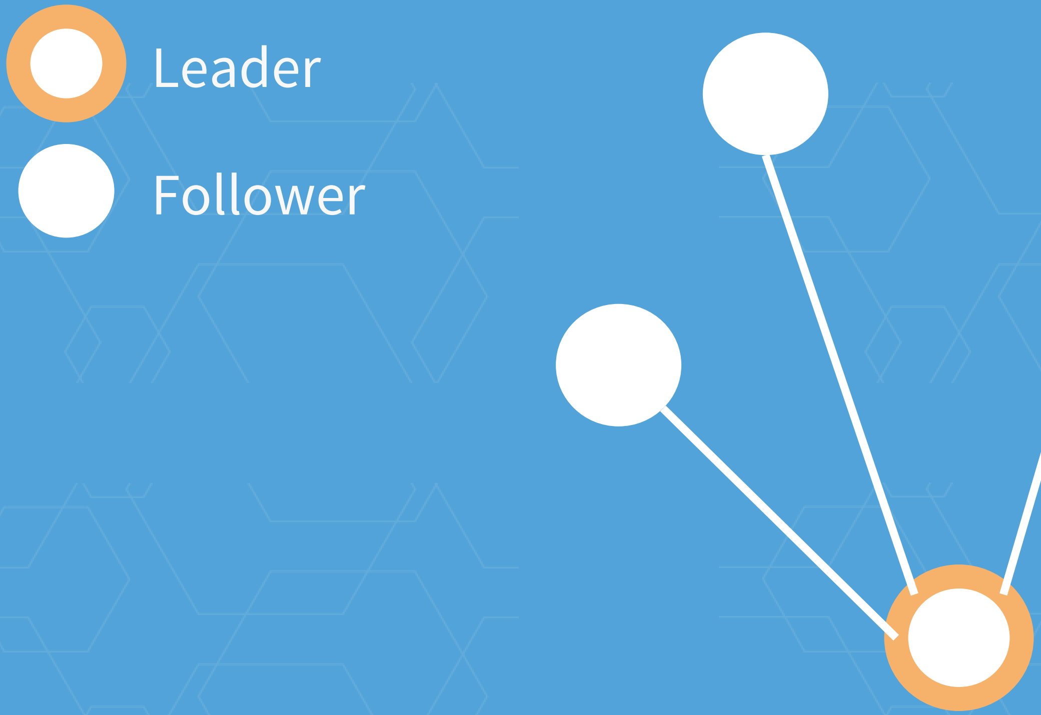


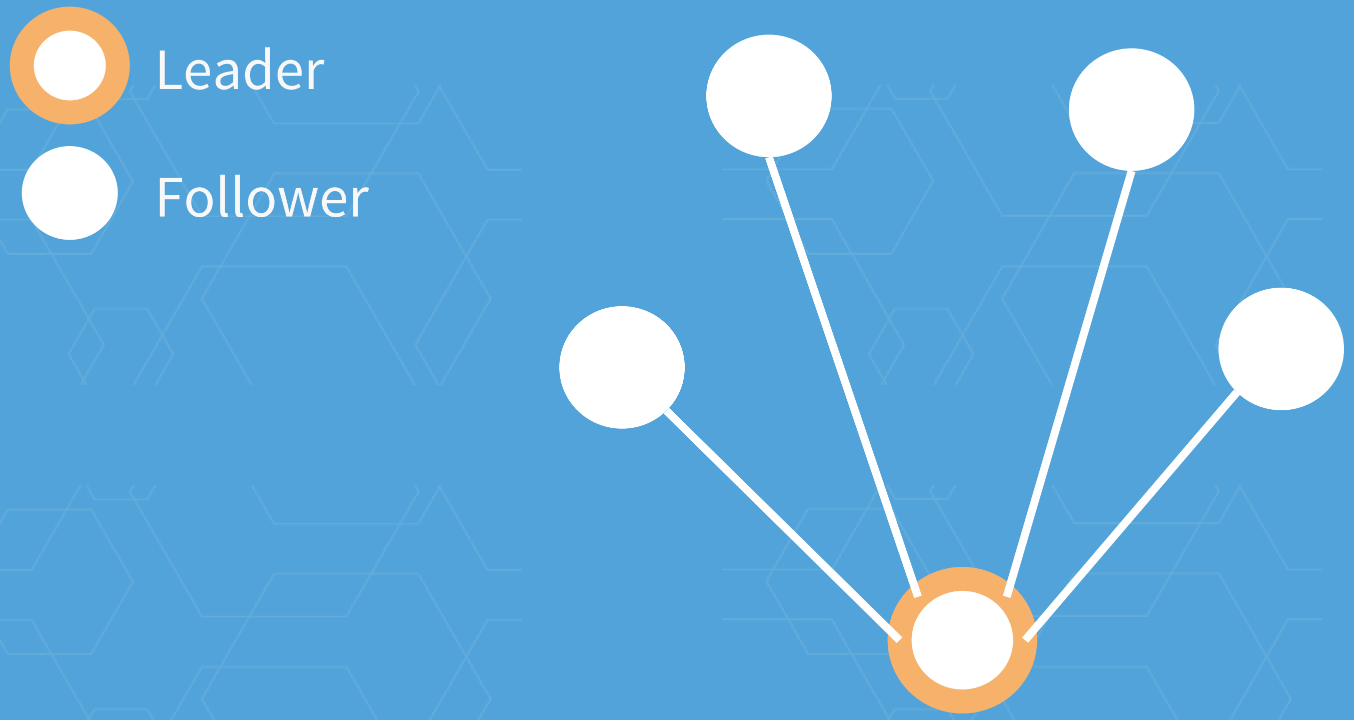
etcd reconfig

live addition and removal

Leader
Follower




Leader
Follower


Leader
Follower