Orchestrating Stateful Workloads on Kubernetes

An Apache Cassandra Use Case

Yannis Zarkadas <yanniszark@arrikto.com>
Software Engineering Intern, Arrikto

National Technical University of Athens CSLab

Arrikto

Athens Kubernetes Meetup - March 12, 2019

Problem Statement

- Great database
- Difficult to manage

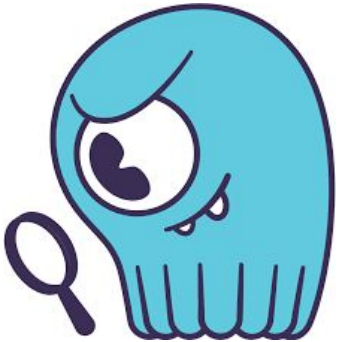
- Great workload management platform

Can we leverage Kubernetes to write a great management layer for Apache Cassandra ?

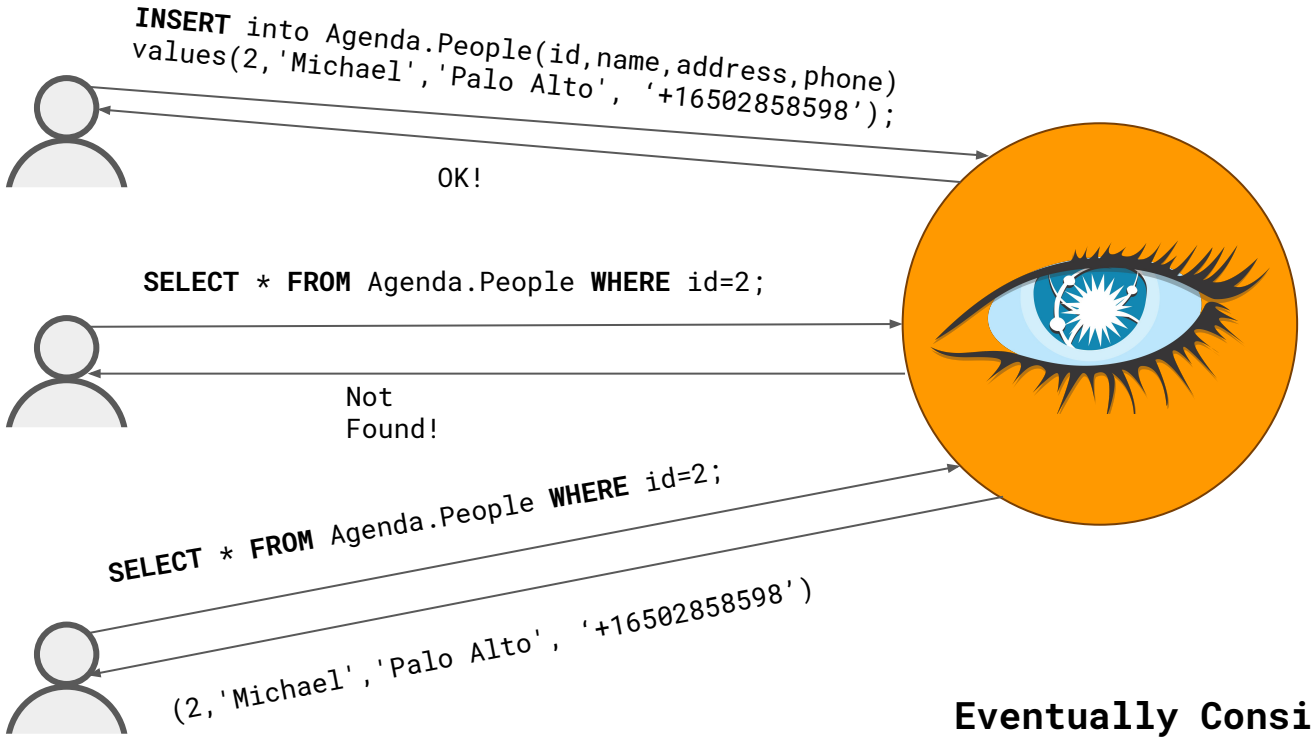
Collaborations

- Cassandra Operator is part of Rook.io
  - CNCF-Incubating project
  - Supports Ceph, CockroachDB, Minio, EdgeFS and others
  - Healthy community
  - Testing framework with Jenkins integration
  - Reusable functionality across storage providers
  - Code reviews by industry experts

- Scylla Operator is based on the Cassandra Operator
  - ScyllaDB is a super-fast database implementing the Cassandra API
  - Based their own official operator on the Cassandra Operator

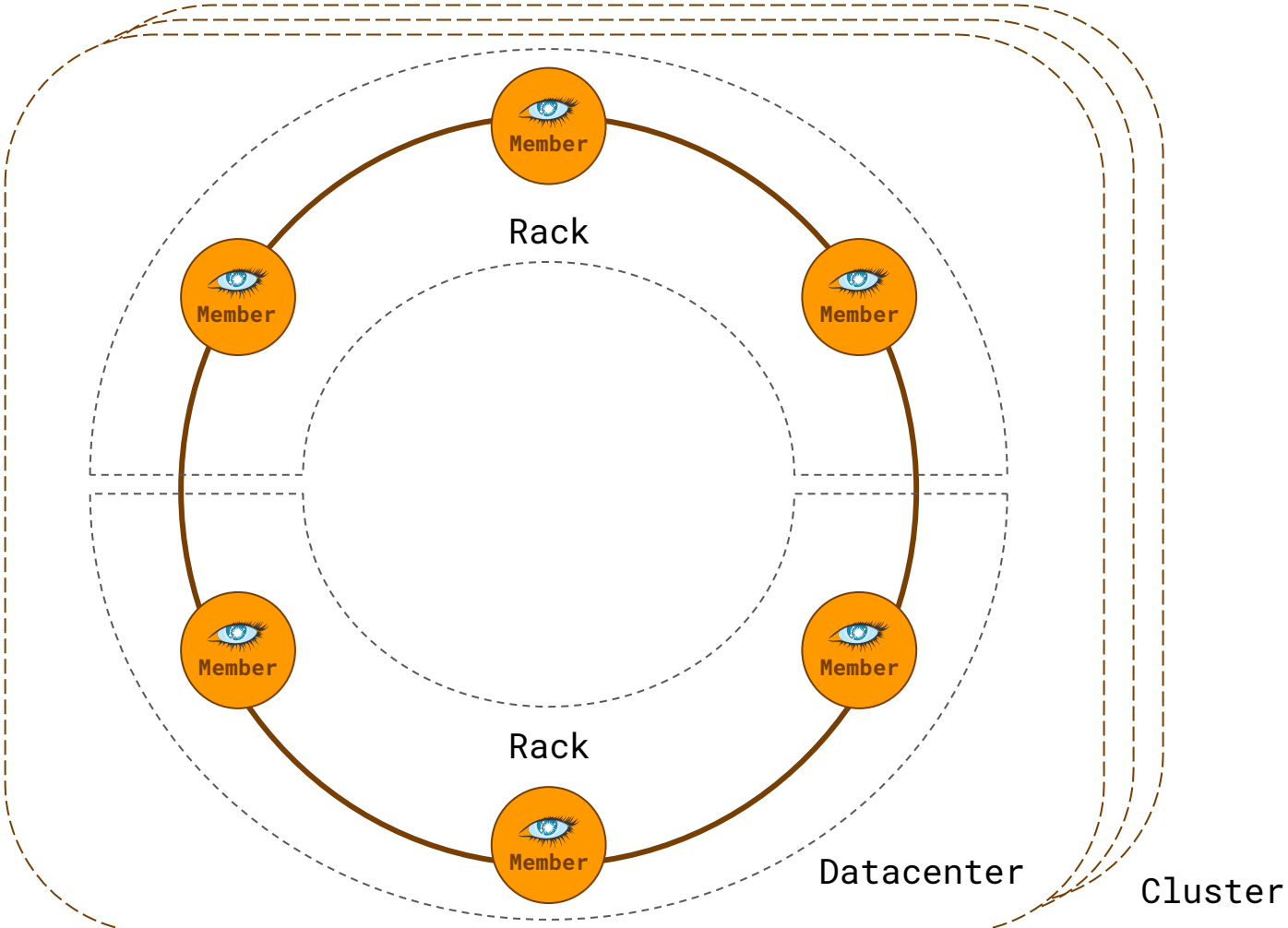


Apache Cassandra Overview



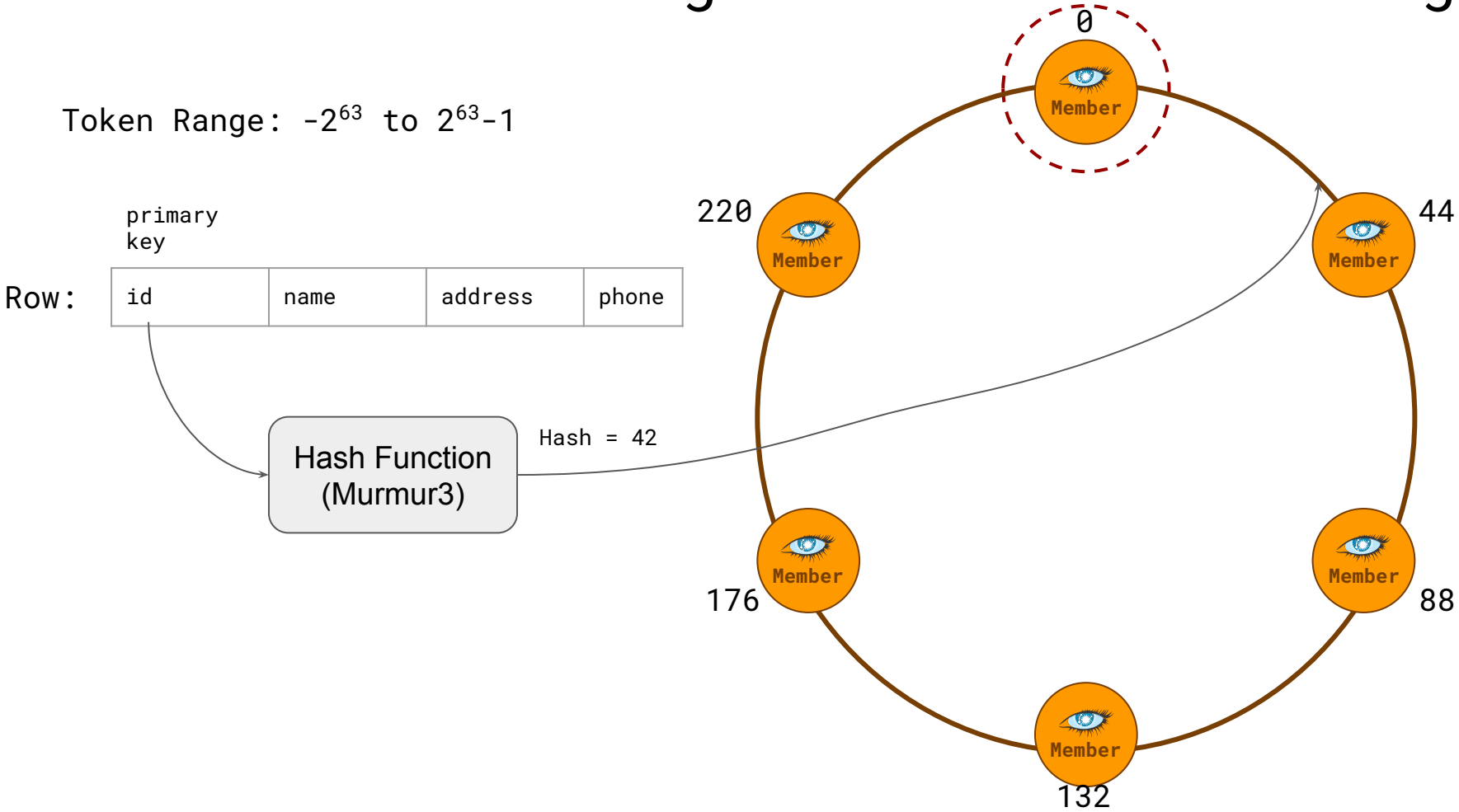
Eventually Consistent

Distributed Architecture



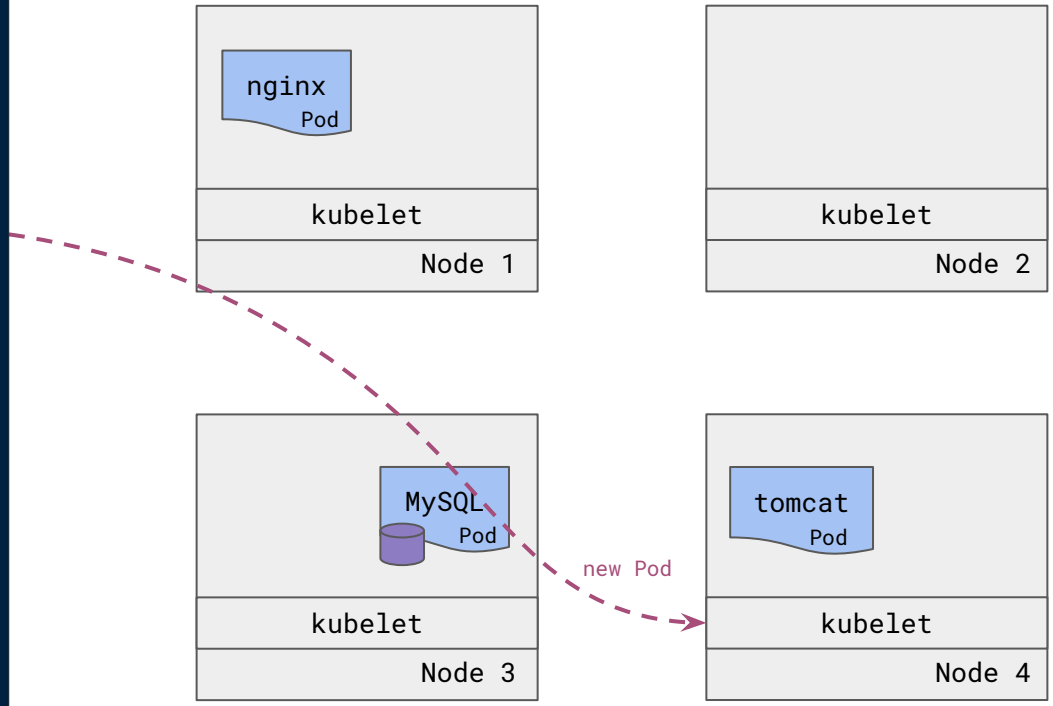
The Cassandra Ring - Data Partitioning

Token Range: -2^{63} to $2^{63}-1$



Pod

```yaml
kind: Pod
metadata:
  name: web-server
  labels:
    app: my-web-app
spec:
  image: tomcat:7
  ports:
    - containerPort: 80
      name: "http server"
  nodeName:
```



StatefulSet

Deploys and scales stateful software.
Provides guarantees for:

- **Pod uniqueness**
  - At most 1 of each Pod exists at any given time
- **Pod ordering**
  - Rolling Update and Deployment
- **Persistent network and storage identity**
  - DNS record and own Persistent Volume

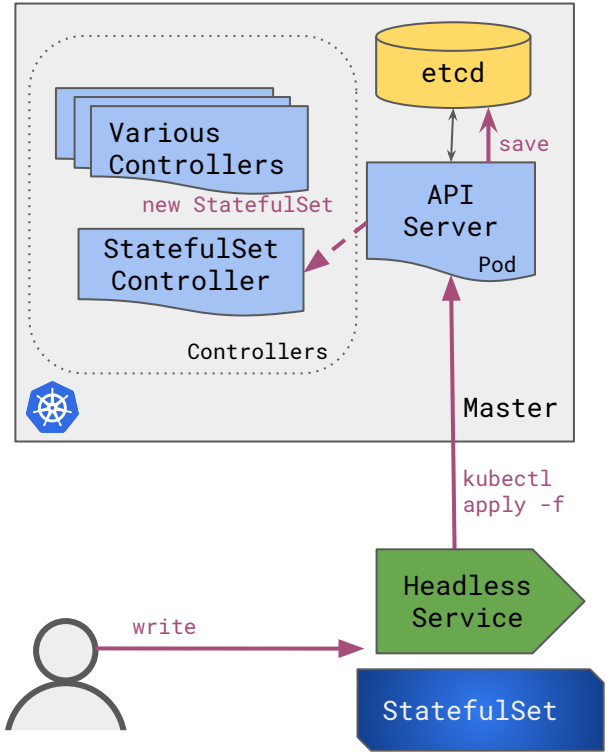
```yaml
kind: Service
metadata:
  name: cassandra
spec:
  clusterIP: None
  ports:
  - port: 9042
  selector:
    app: cassandra
```

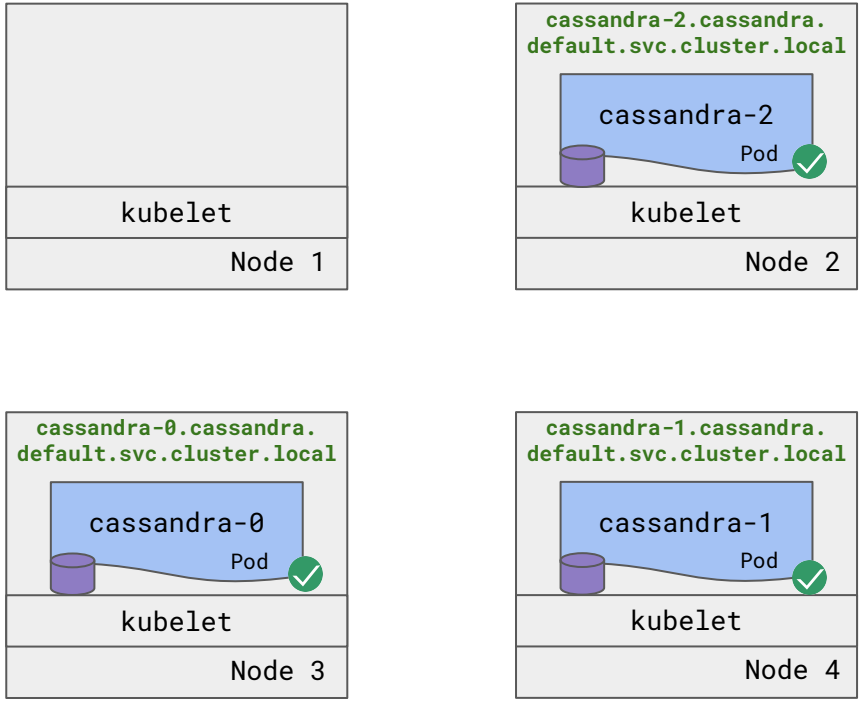
```yaml
kind: StatefulSet
metadata:
  name: cassandra
spec:
  serviceName: cassandra
  replicas: 3
  selector:
    matchLabels:
      app: cassandra
  template:
    metadata:
      labels:
        app: cassandra
    spec:
      containers:
      - name: cassandra
        image: gcr.io/google-samples/cassandra:v13
        ports:
          - name: cql
            containerPort: 9042
        volumeMounts:
        - name: cassandra-data
          mountPath: /cassandra_data
  volumeClaimTemplates:
  - metadata:
      name: cassandra-data
    spec:
      storageClassName: local-disks
      resources:
        requests:
          storage: 500Gi
```

network identity

storage identity

StatefulSet Controller





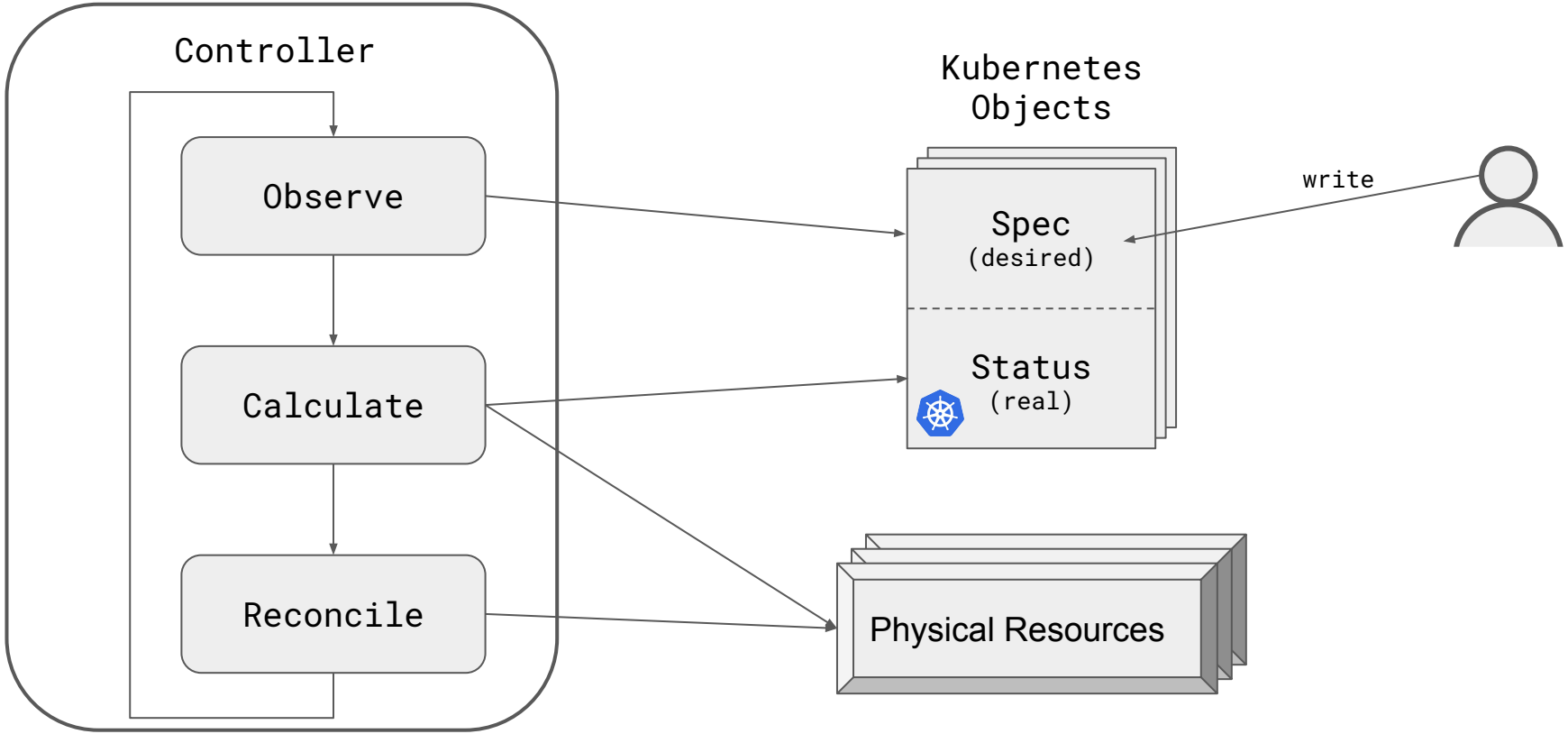
spec.replicas: 3

status.replicas: 3

status.readyReplicas: 3

Controller Pattern

Used ***everywhere*** in Kubernetes



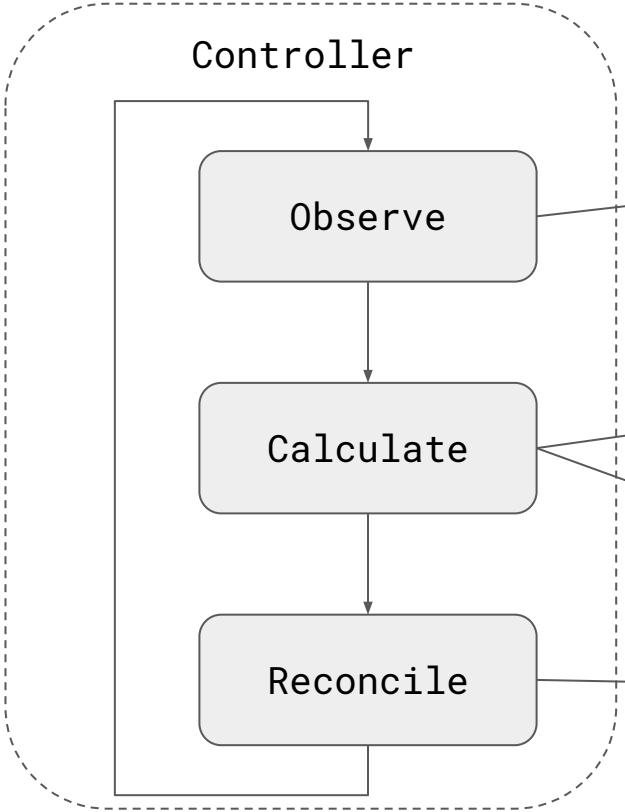
Custom Resource Definition

- Registers a REST endpoint in the API-Server
- We can use that endpoint to store custom Kubernetes Objects
- Compatible with kubectl
  - kubectl get clusters.cassandra.rook.io

```yaml
# Cassandra Cluster CRD
apiVersion: apiextensions.k8s.io/v1beta1
kind: CustomResourceDefinition
metadata:
  name: clusters.cassandra.rook.io
spec:
  group: cassandra.rook.io
  names:
    kind: Cluster
    listKind: ClusterList
    plural: clusters
    singular: cluster
  scope: Namespaced
  version: v1alpha1
```

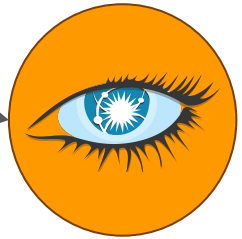
The Operator Pattern

Operator = Controller(s) + CRD(s)




```yaml
apiVersion: cassandra.rook.io/v1alpha1
kind: Cluster
metadata:
  name: my-cassandra-cluster
spec:
  version: 3.11.1
  datacenter:
    name: europe-west-1
    racks:
      - name: europe-west1-b
        members: 3
status:
  racks:
    europe-west1-b:
      members: 2
      readyMembers: 1
```

write

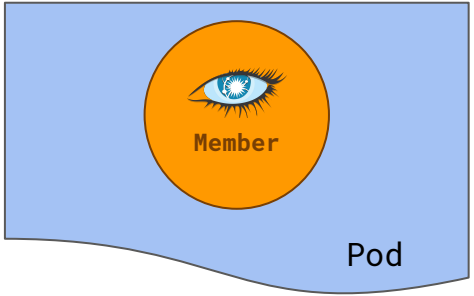


StatefulSet Caveats

StatefulSet: Confined to 1 Rack

kubernetes

Member	Pod
Rack	StatefulSet
Datacenter	StatefulSet
Cluster	StatefulSet

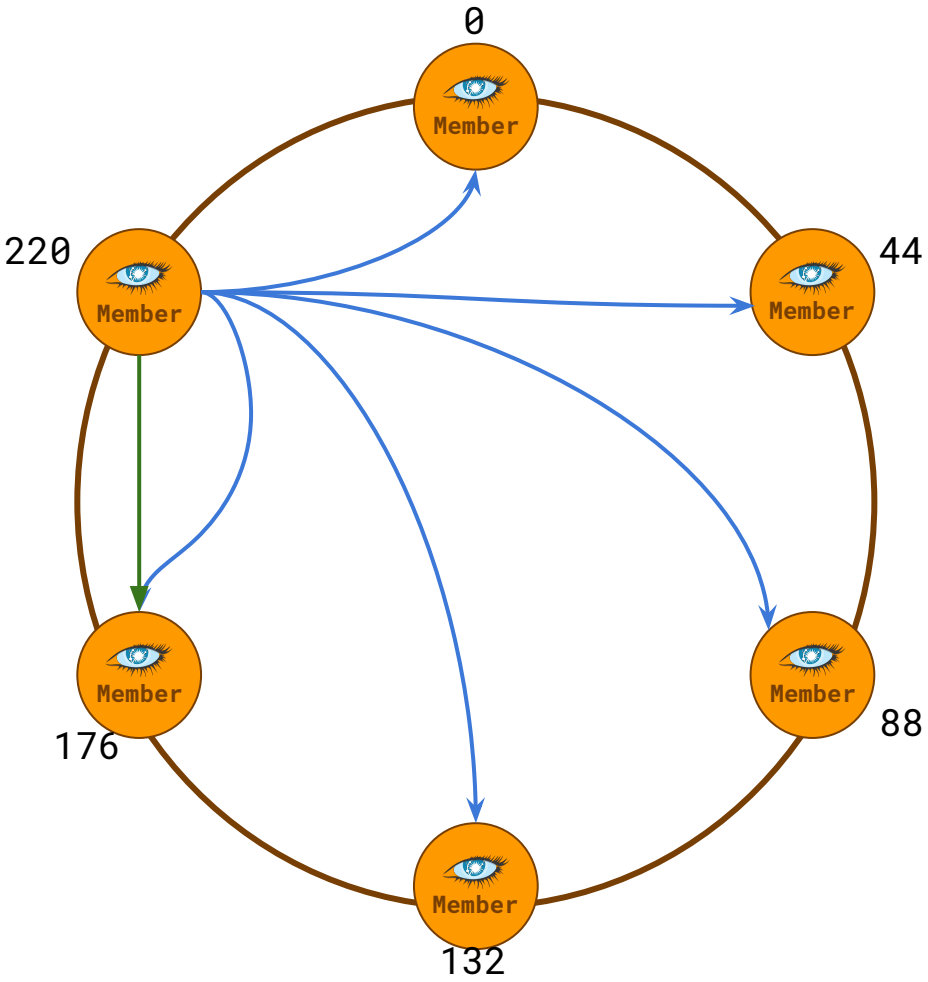


Multip[illegible]cks ?
Multi[illegible]acenters?

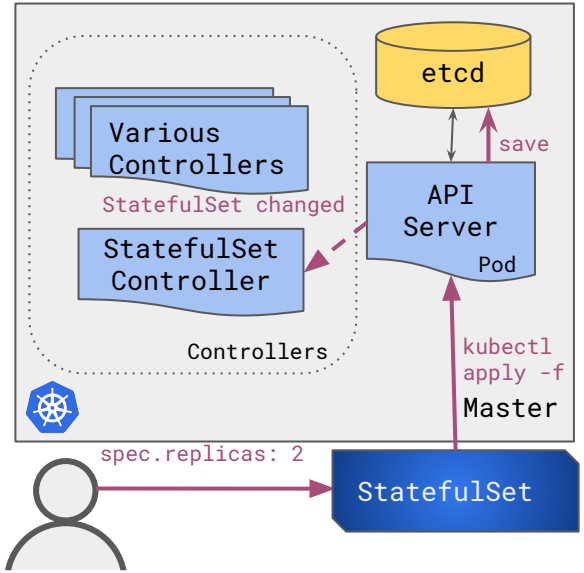
Safe Scale Down

- Want to leave
  - nodetool decommission
- Stream data
- Leave

Cassandra Ring	
member-0	Up
member-1	Up
member-2	Up
member-3	Up
member-4	Up

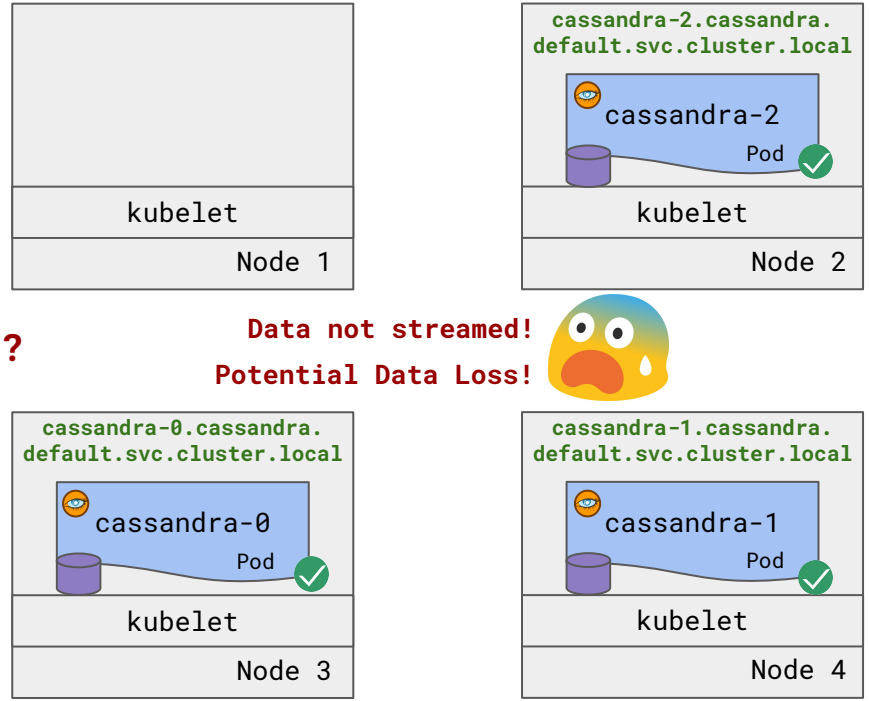


StatefulSet: Unsafe Scale Down



Cassandra Ring	
cassandra-0	Up
cassandra-1	Up
cassandra-2	Down

Scale Down?

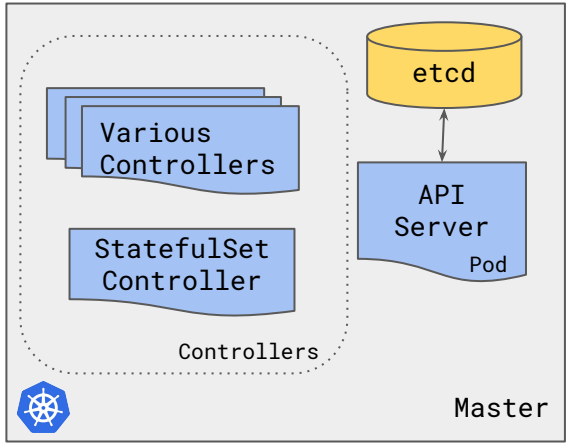


spec.replicas: 2

status.replicas: 3

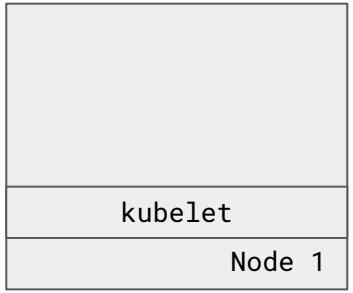
status.readyReplicas: 3

StatefulSet: Cannot track Member identity

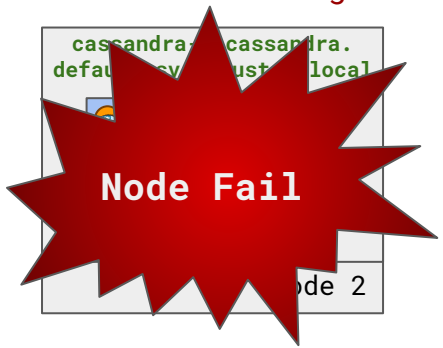


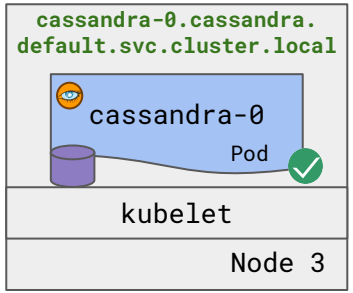
Replace Member? Add new Member?

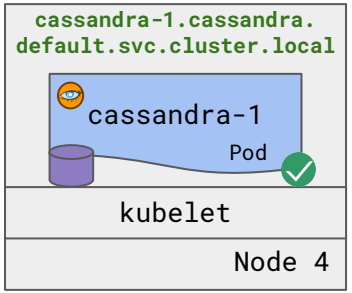
Must know Member identity beforehand!



Member Joining







Vanilla Solution: StatefulSet

Problems with:

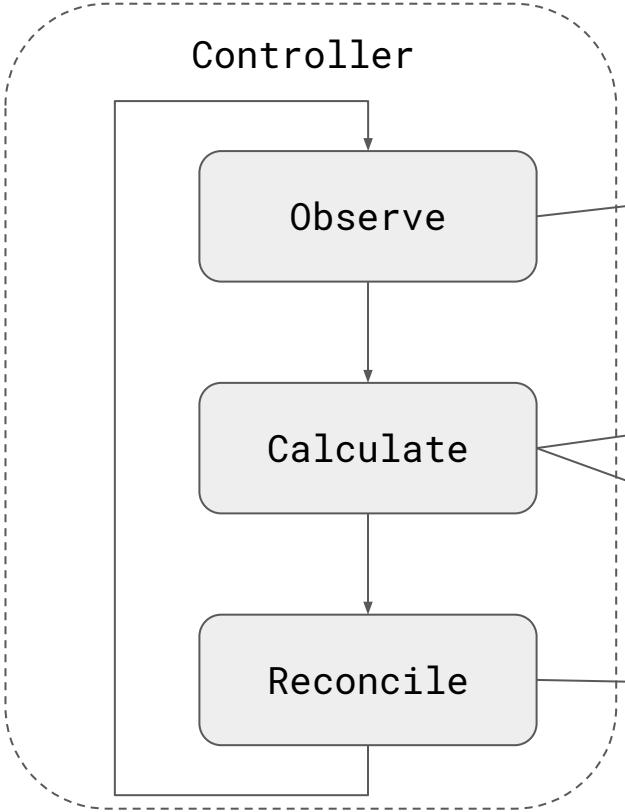
- Seeds
- Multi-zone deployment
- Scale Down
- Loss of Persistence
- Backups/Restores
- Extensibility

What if we could create management software in the image of Kubernetes Controllers?

Design

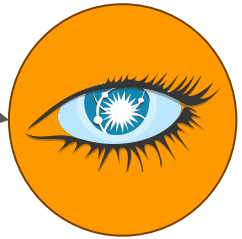
Our goal

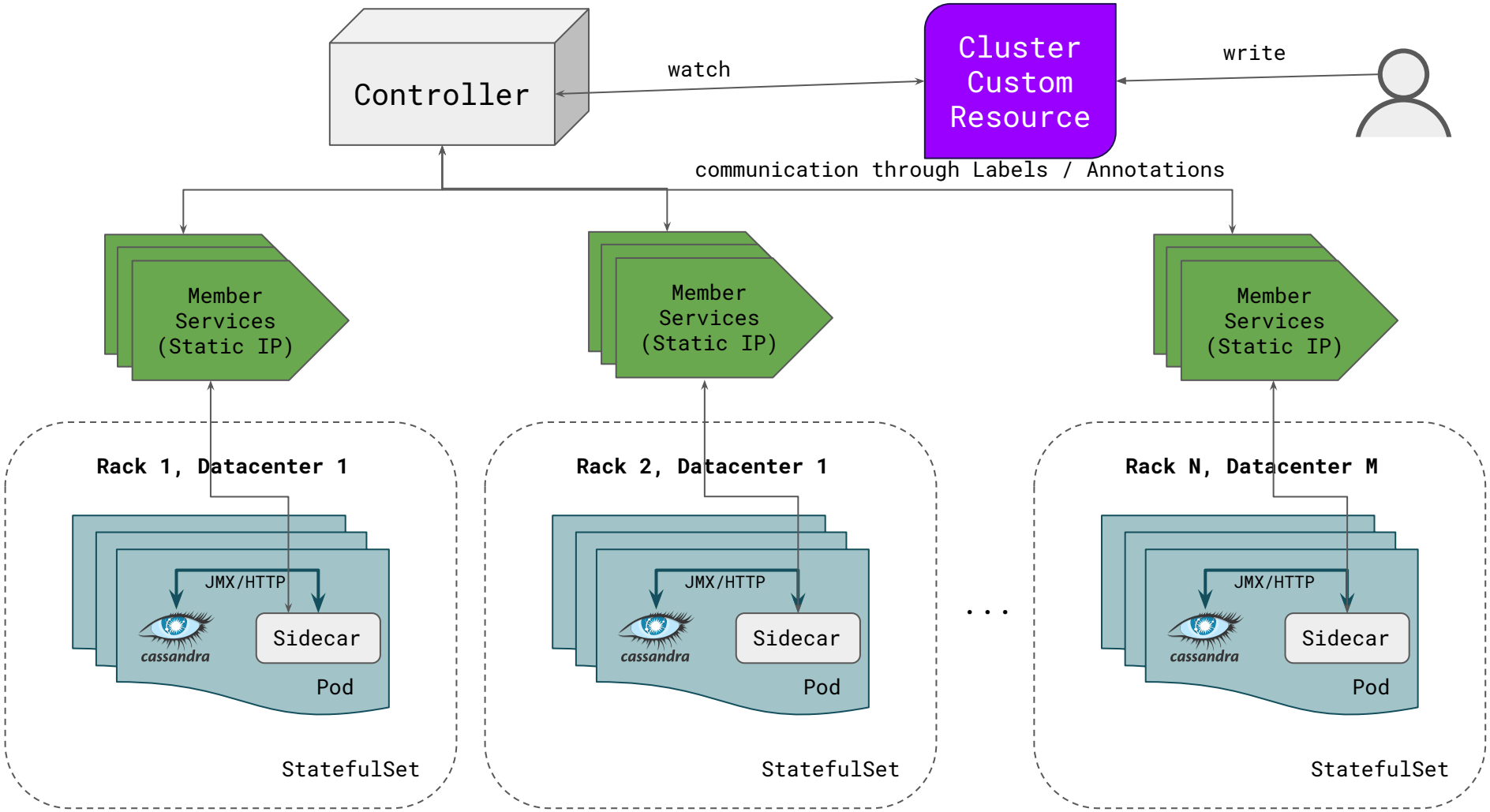
Operator = Controller(s) + CRD(s)




```yaml
apiVersion: cassandra.rook.io/v1alpha1
kind: Cluster
metadata:
  name: my-cassandra-cluster
spec:
  version: 3.11.1
  datacenter:
    name: europe-west-1
    racks:
      - name: europe-west1-b
        members: 3
status:
  racks:
    europe-west1-b:
      members: 2
      readyMembers: 1
```

write




Controller
watch
Cluster Custom Resource
write
communication through Labels / Annotations
Member Services (Static IP)
Member Services (Static IP)
Member Services (Static IP)
Rack 1, Datacenter 1
Rack 2, Datacenter 1
Rack N, Datacenter M
JMX/HTTP
cassandra
Sidecar
Pod
JMX/HTTP
cassandra
Sidecar
Pod
JMX/HTTP
cassandra
Sidecar
Pod
. . .
StatefulSet
StatefulSet
StatefulSet

Mapping of Abstractions

cassandra	kubernetes
Member	Pod
Rack	StatefulSet
Datacenter	StatefulSets
Cluster	Cluster Custom Resource

```yaml
apiVersion: cassandra.rook.io/v1alpha1
kind: Cluster
metadata:
  name: my-cassandra-cluster
spec:
  version: 3.11.1
  datacenter:
    name: europe-west-1
    racks:
      - name: europe-west1-b
        members: 3
status:
  racks:
    europe-west1-b:
      members: 2
      readyMembers: 1
```

Sidecar

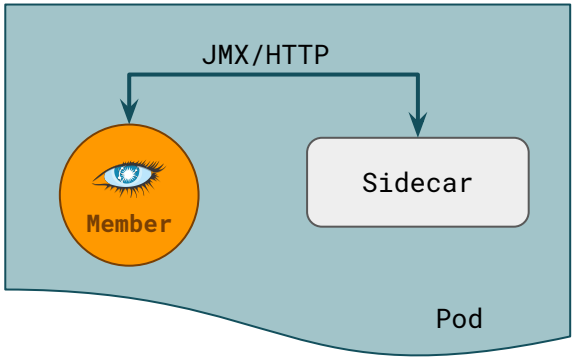
CRD + Controller + Sidecar

Sidecar needed to:

- Setup config files
- Install plugins at startup
- Backup and Restore functionality
- Future extensibility

Communicating with Cassandra:

- JMX interface requires Java
- Use Jolokia HTTP/JMX Bridge



An Alternative to DNS Records

- What if we could have static IPs?

Much Requested Feature ->

Sticky IPs for StatefulSet #28969

Open bprashanth opened this issue on Jul 14, 2016 · 59 comments

Services already have a static IP, called ClusterIP.

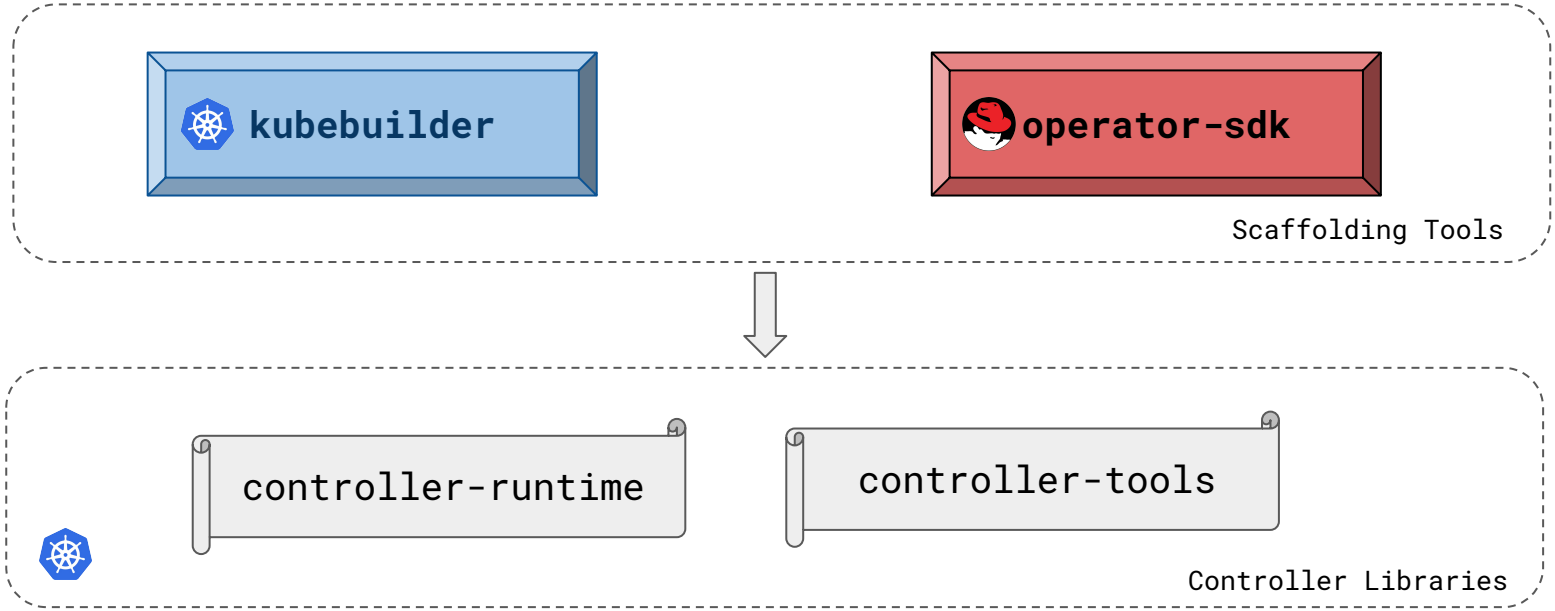
Solution: **ClusterIP Service per Pod**

Drawbacks? :

- Performance: iptables can handle a few hundred Members, IPVS can handle thousands with no problem.
- ClusterIP CIDR Depletion: Usually a /12 IP Block, so plenty of addresses.

Implementation

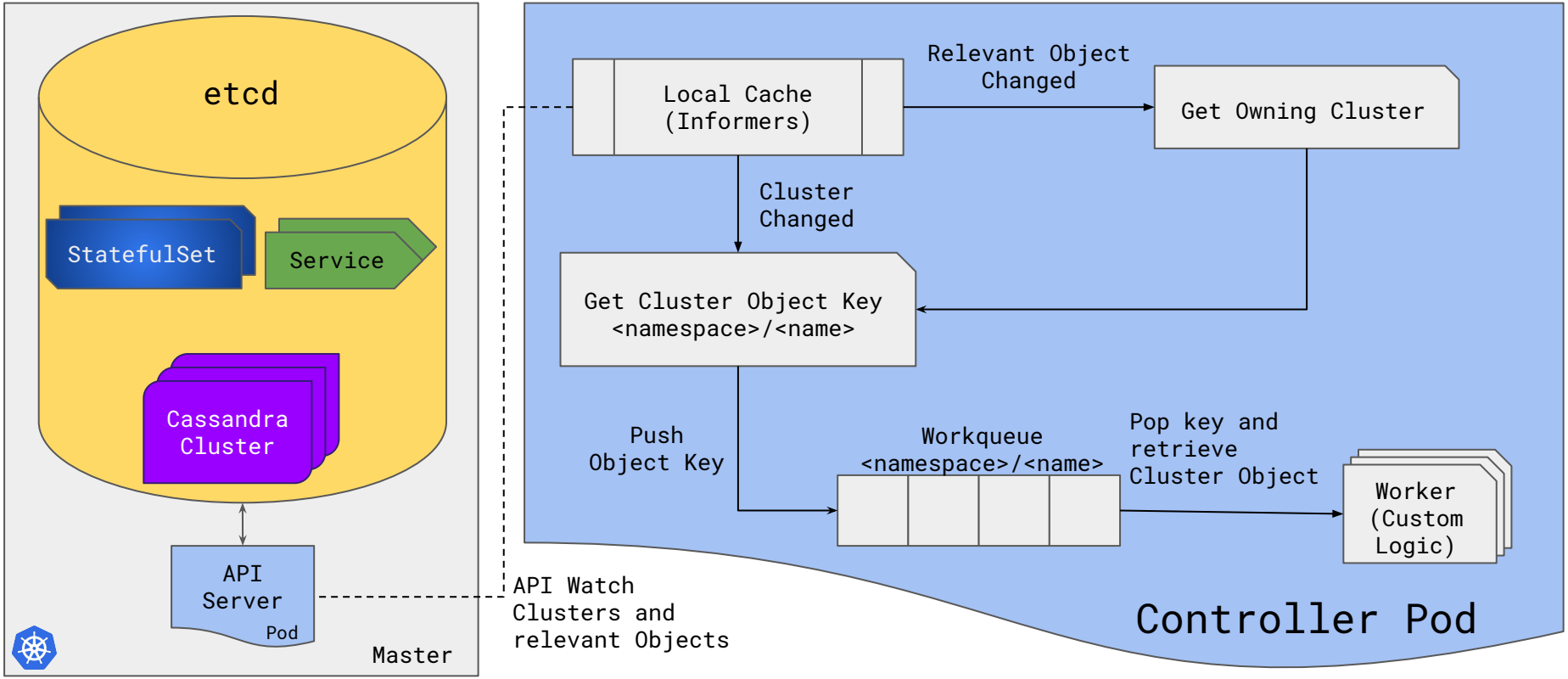
Controller Libraries and Frameworks



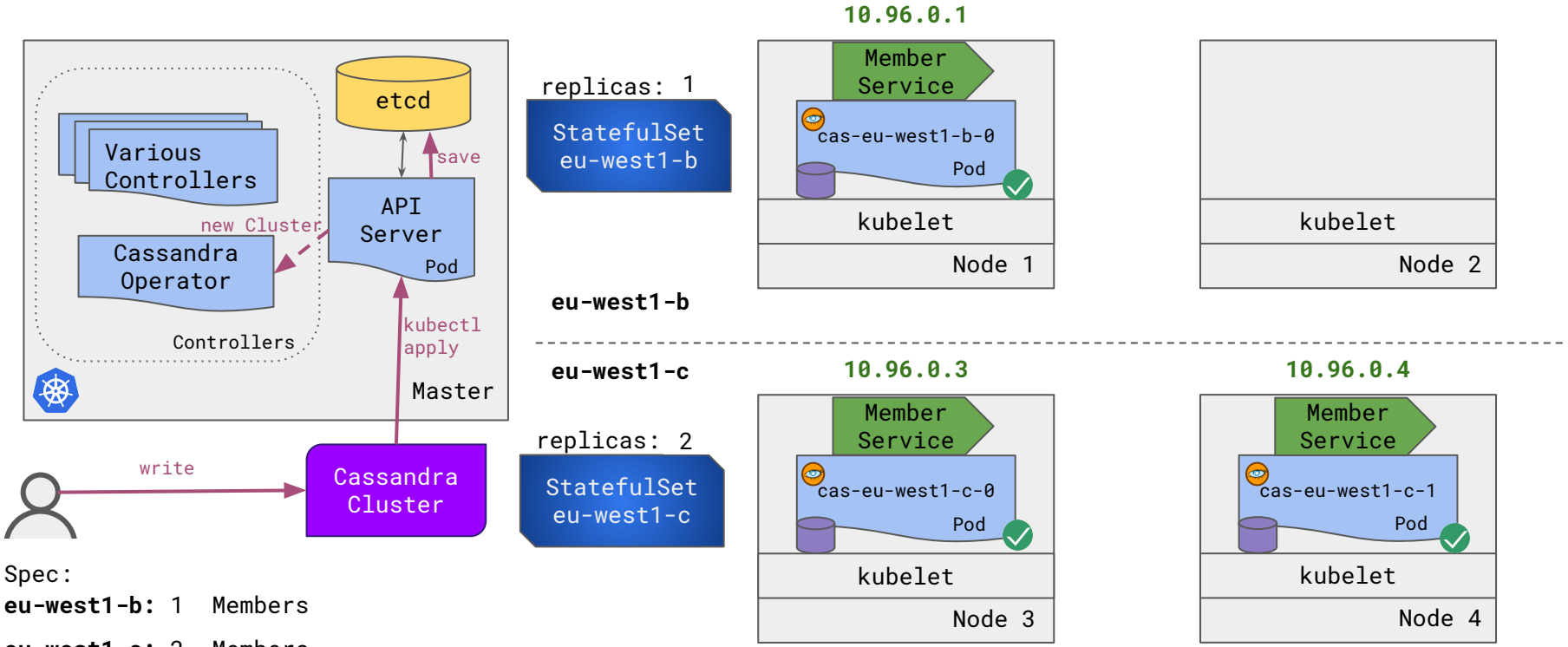
Recommended for writing Controllers now.

Not an option when Cassandra Operator was developed.

Structure of a Kubernetes Controller



Cluster Creation & Scale Up



Spec:

eu-west1-b: 1 Members

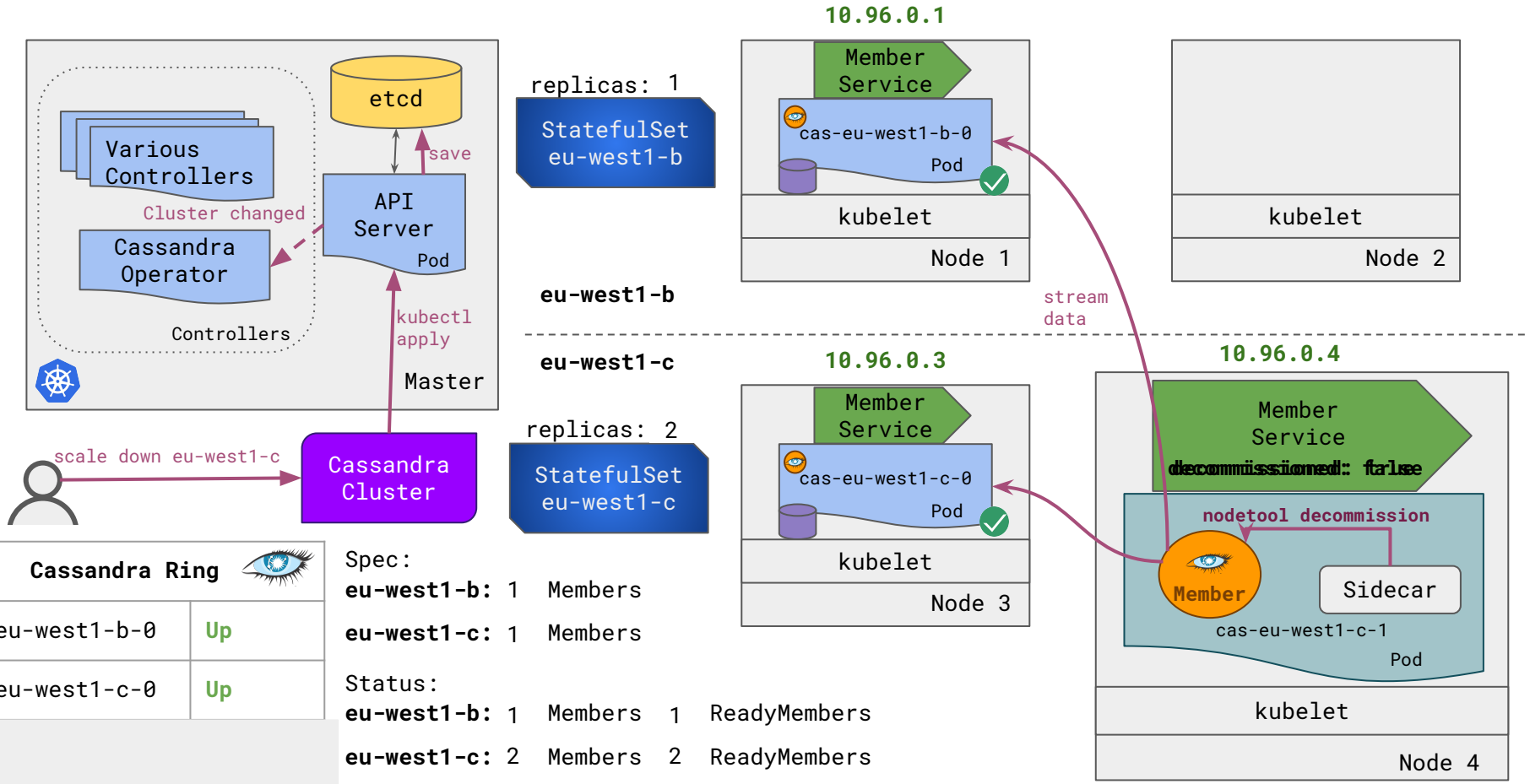
eu-west1-c: 2 Members

Status:

eu-west1-b: 1 Members 1 ReadyMembers

eu-west1-c: 2 Members 2 ReadyMembers

Scale Down



Cassandra Ring	
cas-eu-west1-b-0	Up
cas-eu-west1-c-0	Up

Spec:

eu-west1-b: 1 Members

eu-west1-c: 1 Members

Status:

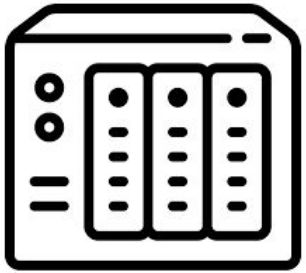
eu-west1-b: 1 Members 1 ReadyMembers

eu-west1-c: 2 Members 2 ReadyMembers

Local Storage vs Network Attached

Local NVME SSD

- Fast
- Ephemeral



Network Attached Storage
(AWS EBS, Google Persistent Disk)

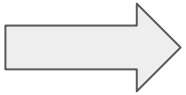
- Slow
- Fault-tolerant

Cassandra handles replication => Use Local Storage!

v1.10: Local Persistent Volumes in Beta

Local Storage Failure Scenarios

- Disk Misbehaves
  - Block errors
  - Deteriorating performance



- Pod still runs
- Unhandled by K8s

- Disk Fails
  - Mount Point Disappears

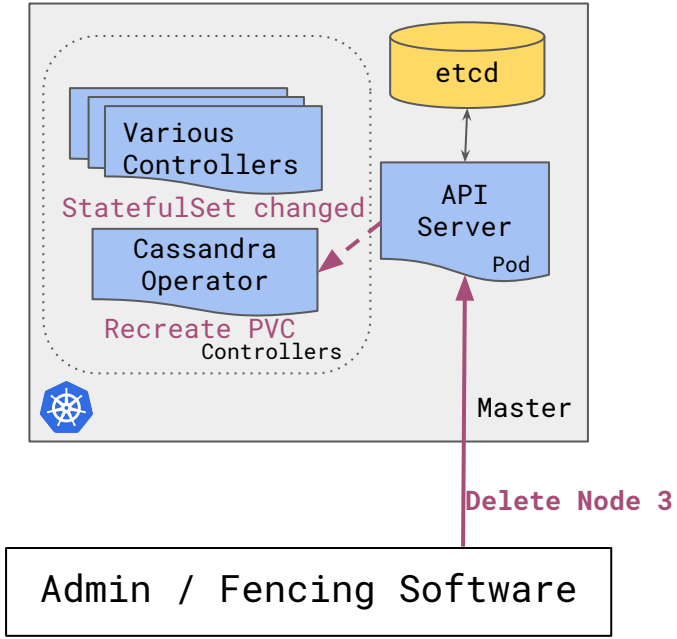
- Pod fails to start
- Unhandled by K8s

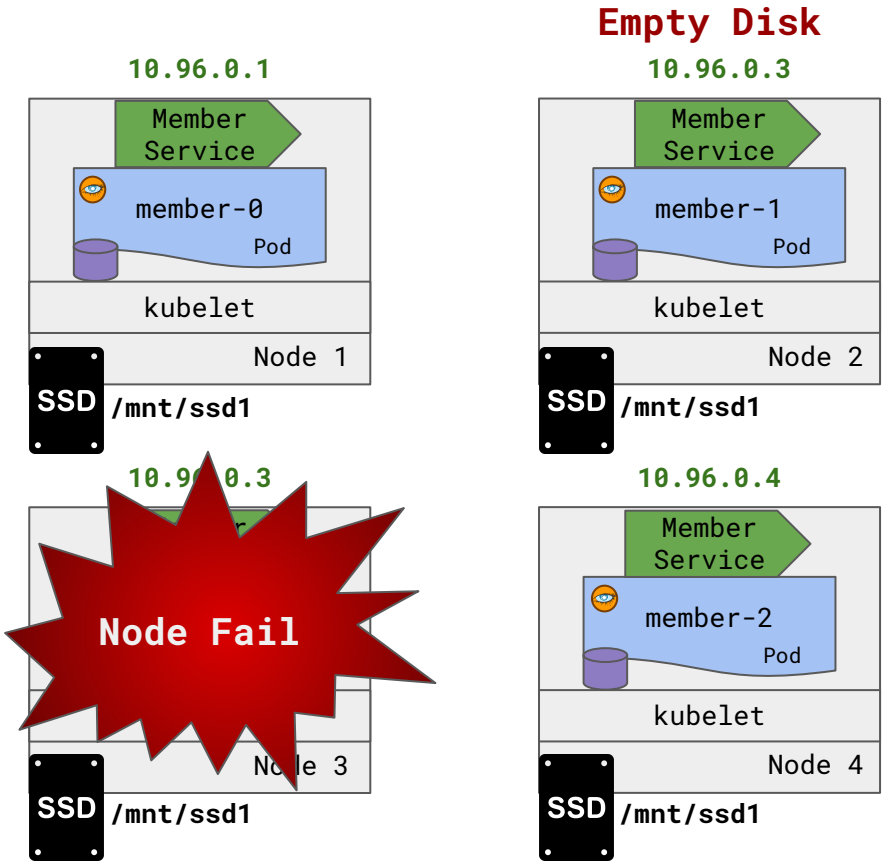
Common in the Cloud!

- Node Fails
  - With Disk on it

- Pod fails to be scheduled
- Unhandled by K8s

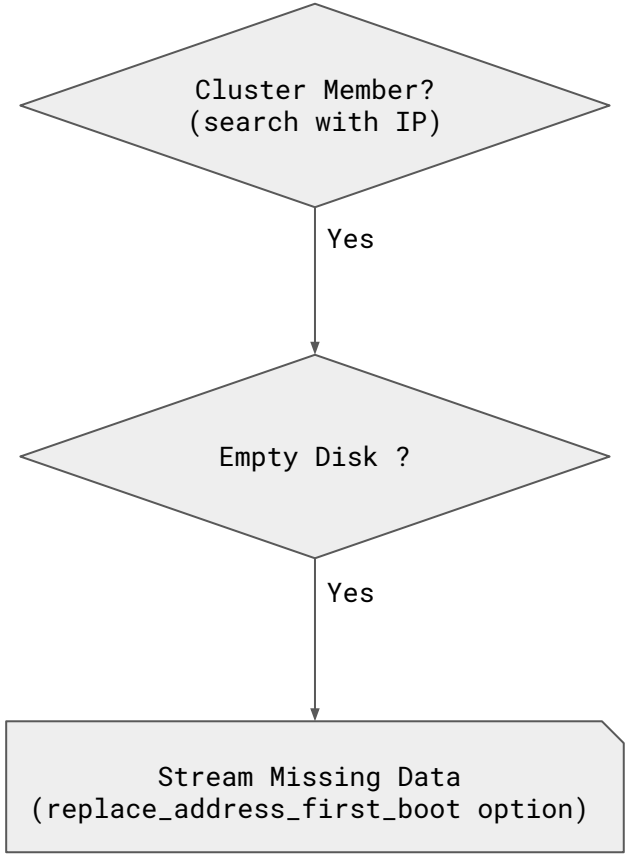
Node Fail

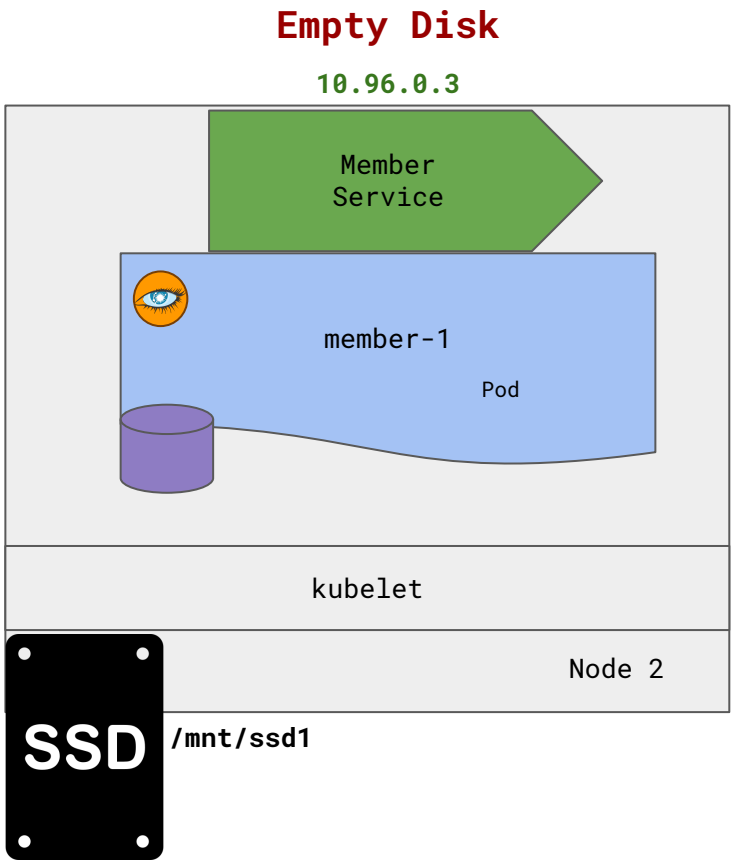

etcd
Various Controllers
StatefulSet changed
Cassandra Operator
API Server
Pod
Recreate PVC
Controllers
Master
Delete Node 3
Admin / Fencing Software


Empty Disk
10.96.0.1
Member Service
member-0
Pod
kubelet
Node 1
SSD
/mnt/ssd1
10.96.0.3
Member Service
member-1
Pod
kubelet
Node 2
SSD
/mnt/ssd1
Node Fail
Node 3
SSD
/mnt/ssd1
10.96.0.4
Member Service
member-2
Pod
kubelet
Node 4
SSD
/mnt/ssd1

Node Fail

Algorithm:





Demo

Future Work

Cassandra Operator

- Backups and Restores
  - Backup CRD and Backup Controller
- Multi-Region Clusters
  - Very early support in Kubernetes
  - Can be worked around using LoadBalancer per Pod
- Monitoring, Repairs, Better UX

Local Storage

- Persistent Volume Monitoring
  - Proposal in Pull #1484 kubernetes/community
  - Issue #10 in kubernetes-sigs/sig-storage-local-static-provisioner

Arrikto

ROOK

SCYLLA.