KubeCon | CloudNativeCon

North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

Multi-Cluster Stateful Set Migration: A Solution to Upgrade Pain

Matt Schallert (@schallert)
Peter Schuurman (@pwschuurman)

Multi-Cluster Migration

KubeCon | CloudNativeCon

North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

October 24-28, 2021

Matt Schallert
Chronosphere

Peter Schurman
Software Engineer
Google

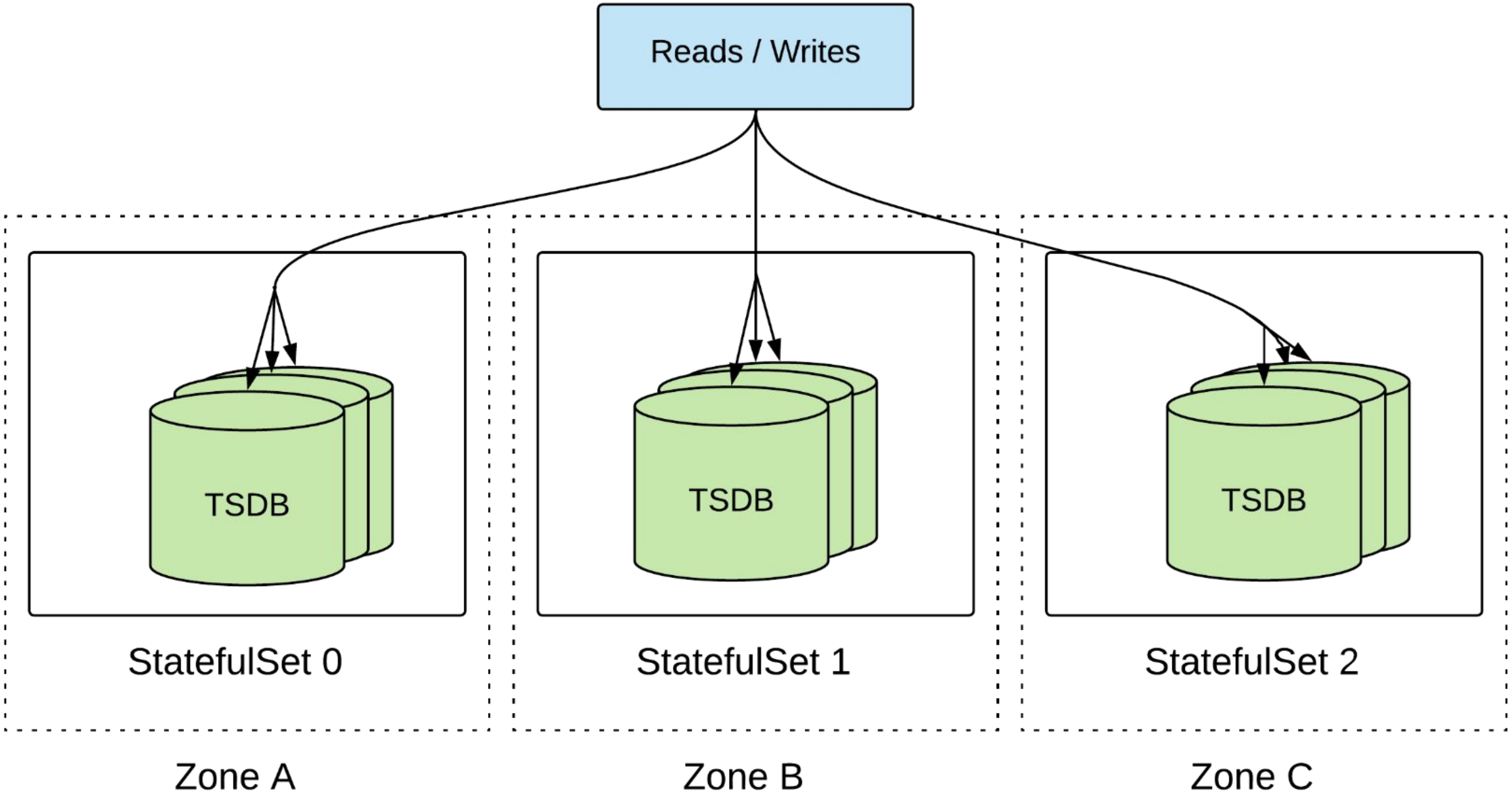
Outline

- Kubernetes at Chronosphere
- Cross-Cluster Migration Challenges
- Cross-Cluster Building Blocks
- Demo

Kubernetes at Chronosphere

- Hosted observability platform. High SLA.
- Thousands of nodes across many clusters, multiple regions.
- Mix of stateless + stateful workloads.
- Primary stateful workload: metrics datastore.

TSDB Architecture



Stateful Operations

- Robust workflows for maintenance operations.
  - Migrating node pools, storage resize, upgrades.
- No solution for migrating stateful workloads between Kubernetes clusters.
  - Lots of hacks and orphaned StatefulSets.

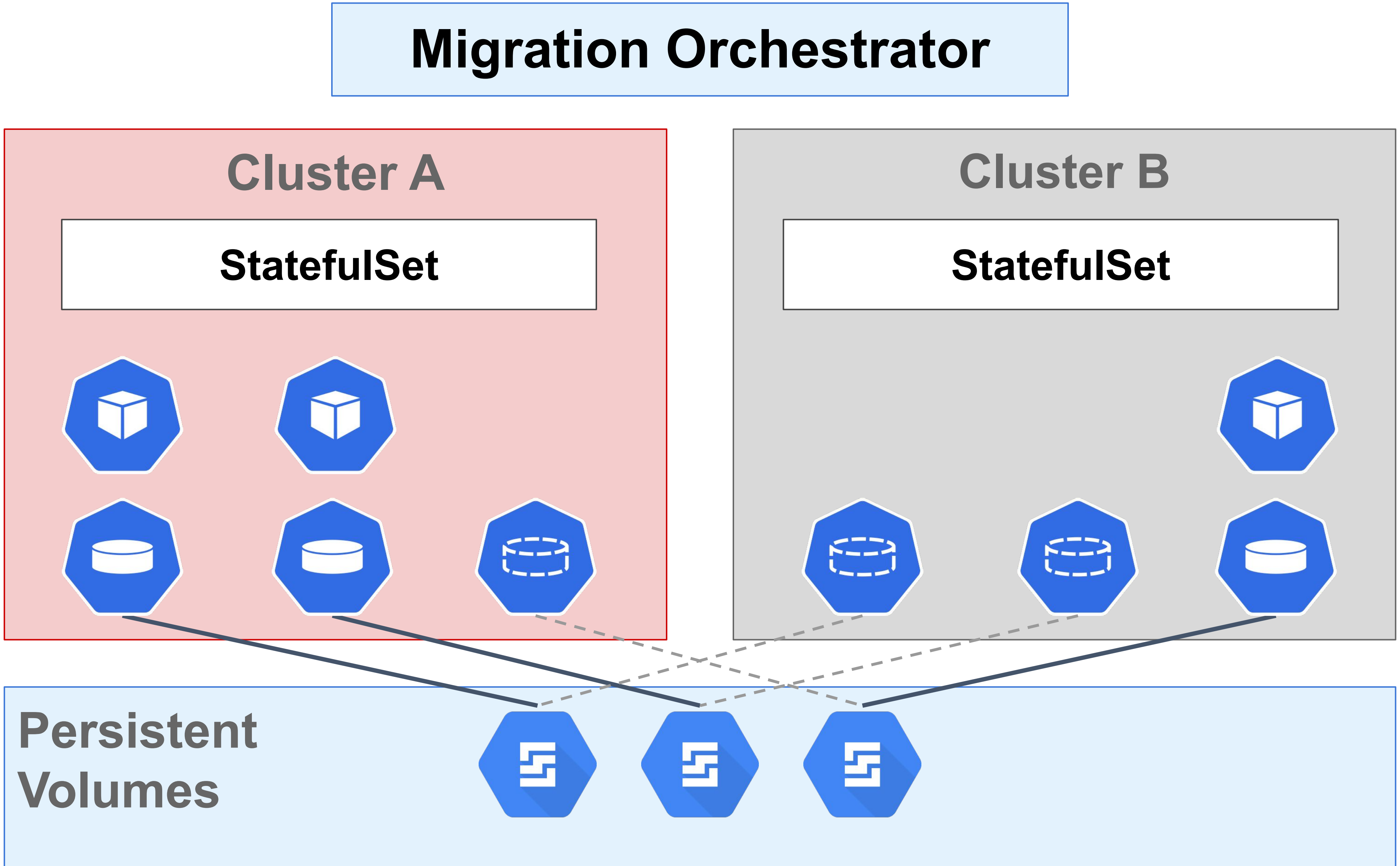
Cross-Cluster Migration Use Cases

- Cluster capacity balancing + reducing blast radius of failures.
- Migrating to new regions (data sovereignty, user latency).
- Features only available on new clusters.
- Low-level cluster changes.
    - Swapping out dataplane on a running cluster == changing plane engine mid-flight.

Cross-Cluster Migration Challenges

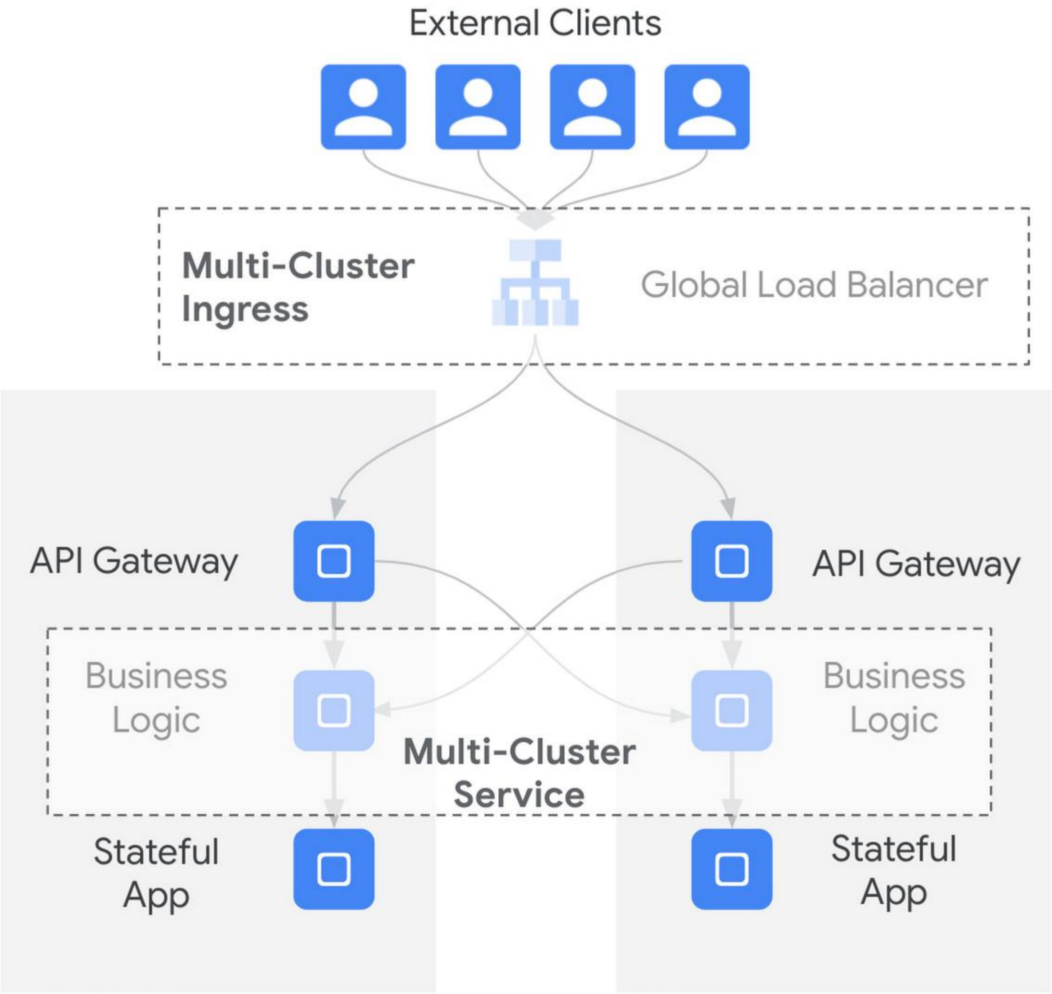
- Lots of solutions for cross-cluster stateless workloads.
    - Multi-cluster services, multi-cluster ingress.
- Stateful story is less certain.
- TSDB architecture means a multi-cluster LB isn't enough.
    - Client-side quorum requires connection to each individual node.

Migration Challenges: Overview



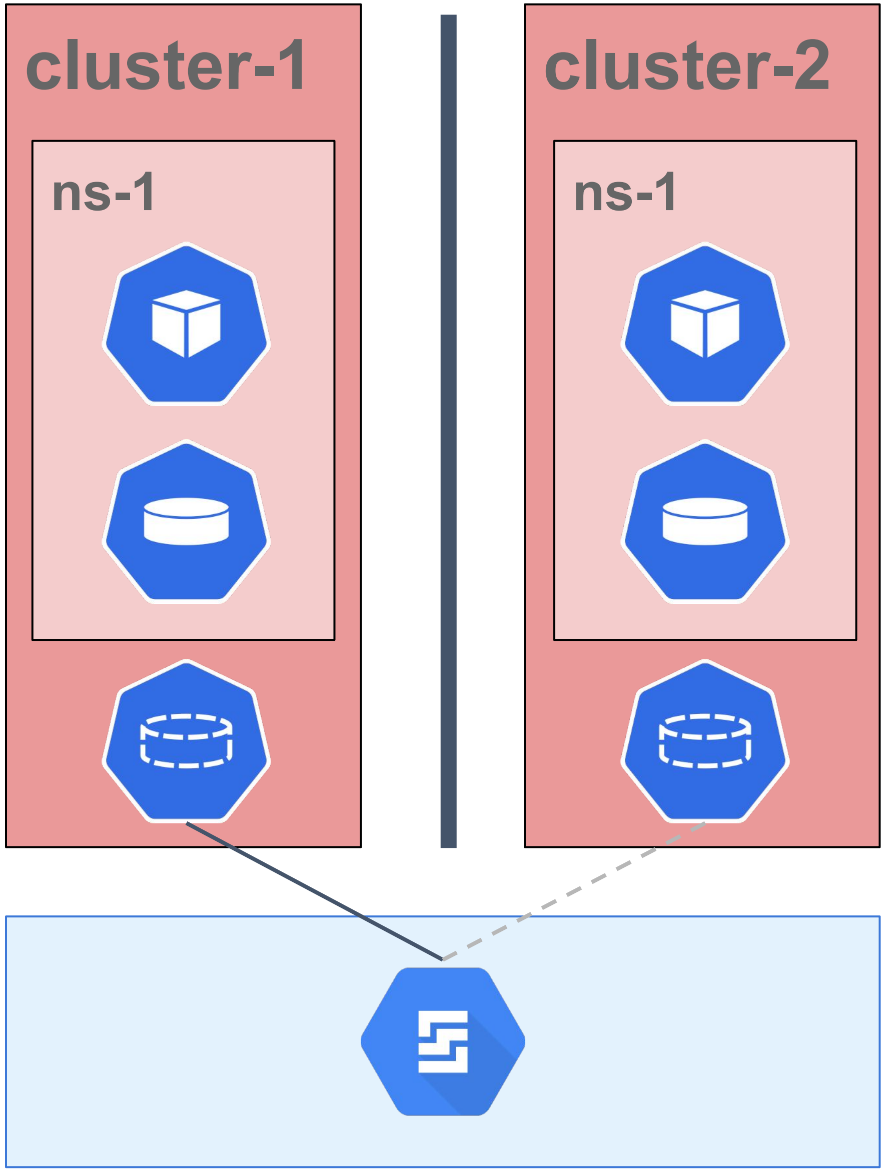
Migration Challenges: Network

- Clients need transparent access to updated endpoints
- Application needs to be accessible from peers
- Applications should have minimal changes to business logic, operation within a cluster should match operation across clusters



Migration Challenges: Storage

- Data layer should be accessible across clusters. Relying on application layer to replicate data over the network can be costly
- PVs are global resources in a single cluster
- Across clusters, Apiserver can't enforce PV <-> PVC uniqueness



Migration Challenges: Orchestration

- Replicas need to follow storage
- Disruption Budget needs to be respected
- Orchestration must move in lockstep with network endpoint propagation
- Operators need to be in sync with orchestration

Building Blocks: Multi-Cluster Services

- KEP-1645: Multi-Cluster Services
- Specification for cross-cluster domain naming
- Solves peer discovery between application replicas, and individual addressing of database replicas

Building Blocks: KEP-3335

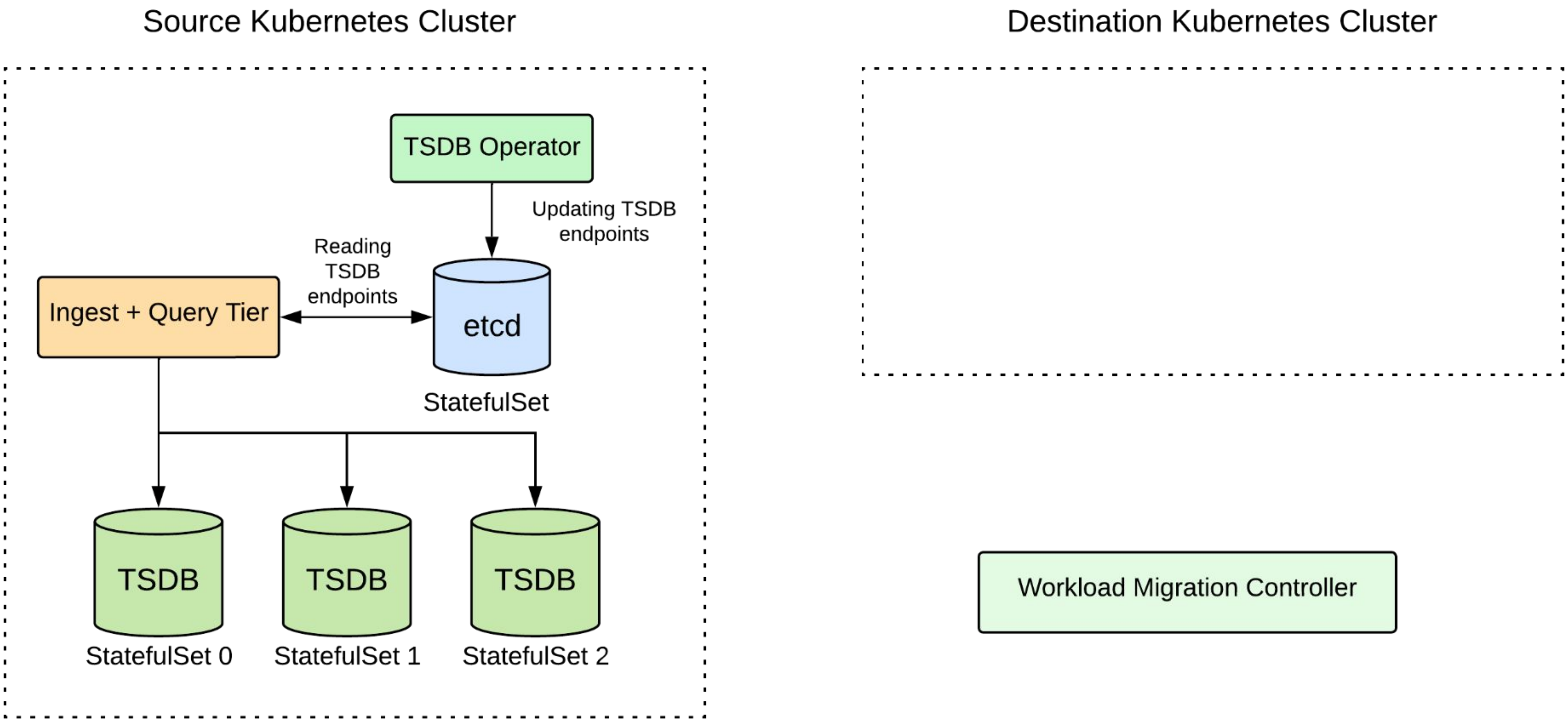
- KEP-3335: StatefulSet Slice
- Granular orchestration of StatefulSet Replicas
- Allows for scaling down from one StatefulSet and scaling up to another StatefulSet

Building Blocks: Tying Pieces Together

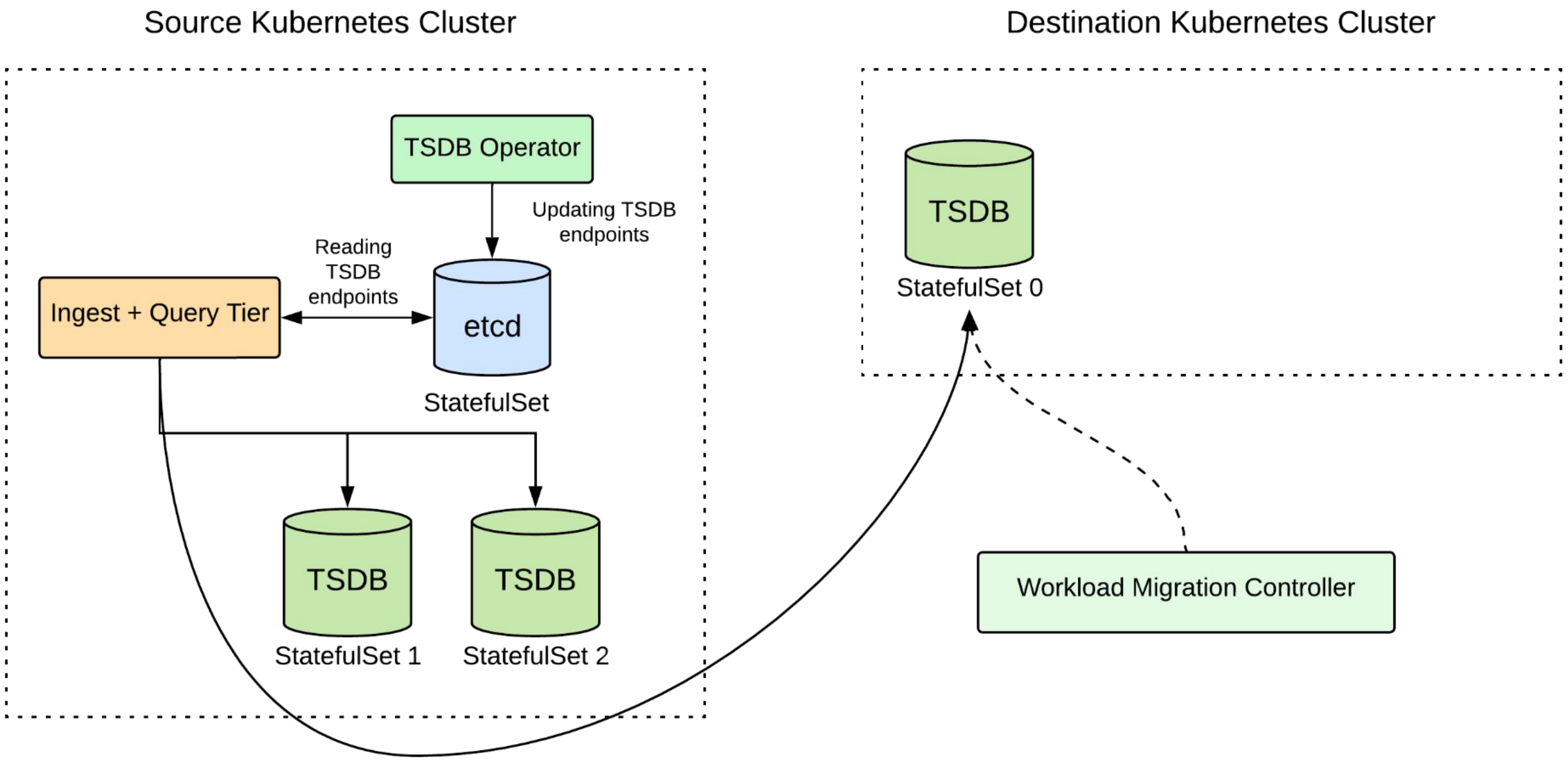
- Migration requires coordination of building blocks
- Setting up applications to be multi-cluster transparent
- Ensuring applications properly express health metrics
- Setting up CI/CD and Operators to be aware of a migration
- Moving StatefulSet dependencies (ConfigMaps, PV/PVC) prior to moving StatefulSet replicas

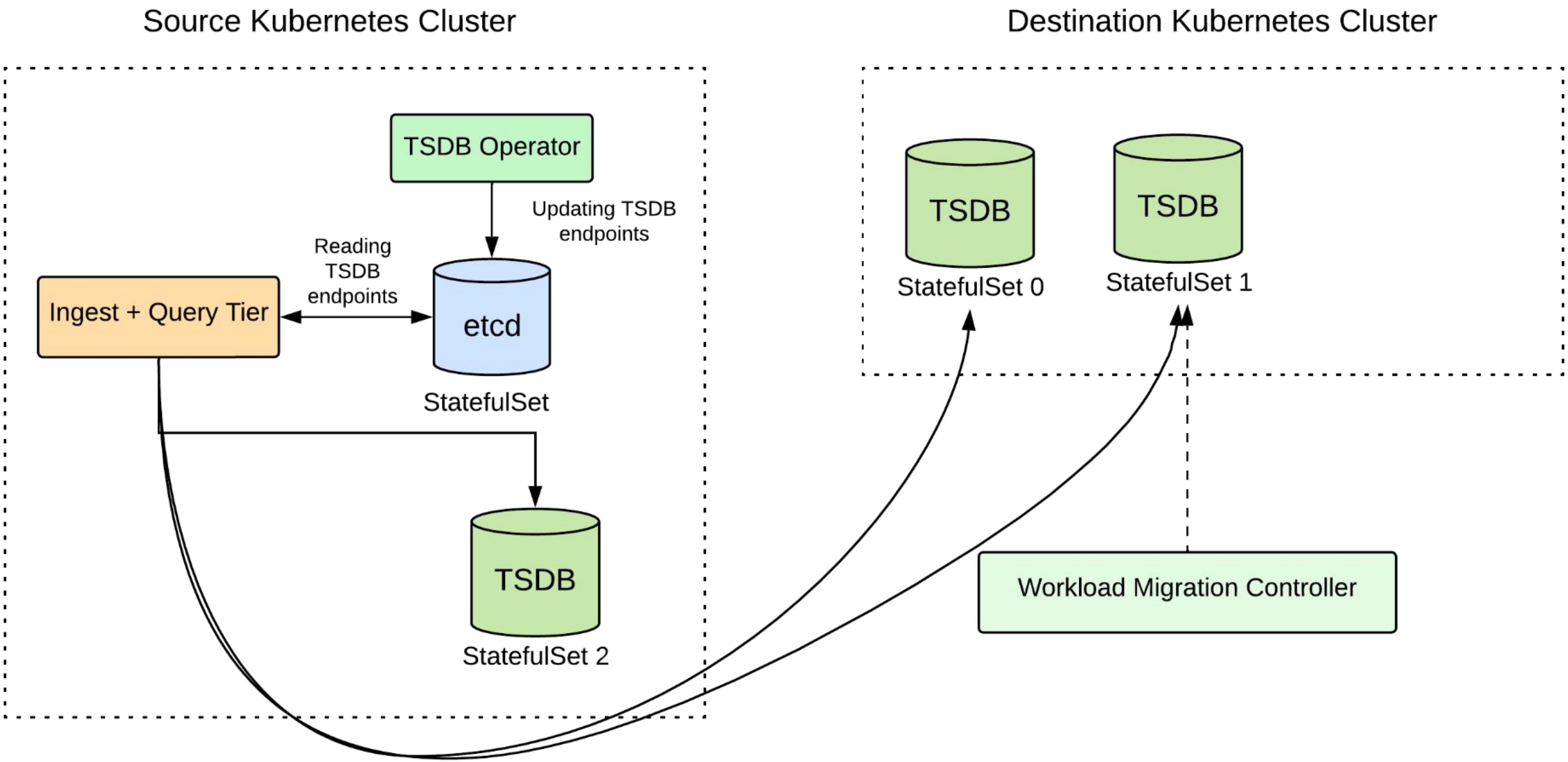
Demo

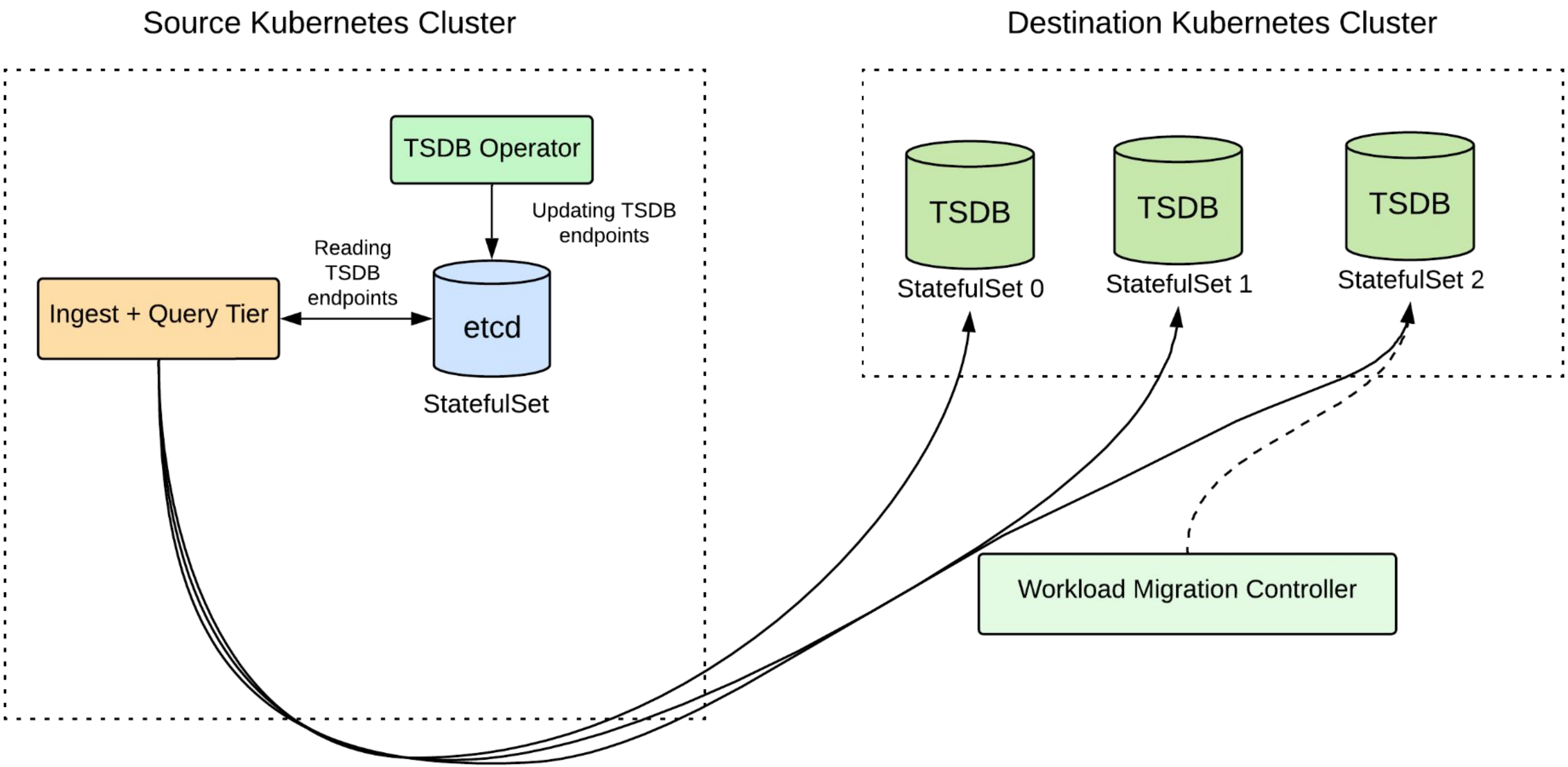
- M3DB Migration across clusters (Video)
- Four StatefulSets
  - Three application StatefulSets (one-per-zone)
  - One placement database (etcd)
- Networking: Multi-Cluster Services on GKE
- Orchestration: StatefulSet Slices on GKE
- Storage Migration: Orchestration of StatefulSets and PV/PVC references



Source Kubernetes Cluster
Destination Kubernetes Cluster
TSDB Operator
Updating TSDB endpoints
Reading TSDB endpoints
Ingest + Query Tier
etcd
StatefulSet
TSDB
TSDB
TSDB
StatefulSet 0
StatefulSet 1
StatefulSet 2
Workload Migration Controller


Source Kubernetes Cluster
Destination Kubernetes Cluster
TSDB Operator
Updating TSDB endpoints
Reading TSDB endpoints
Ingest + Query Tier
etcd
StatefulSet
TSDB
StatefulSet 1
TSDB
StatefulSet 2
TSDB
StatefulSet 0
Workload Migration Controller


Source Kubernetes Cluster
Destination Kubernetes Cluster
TSDB Operator
Updating TSDB endpoints
Reading TSDB endpoints
Ingest + Query Tier
etcd
StatefulSet
TSDB
StatefulSet 2
TSDB
StatefulSet 0
TSDB
StatefulSet 1
Workload Migration Controller



chronosphere

What’s Next?

- Safety: Protecting applications across clusters
- Speed: Aligning update unavailability budget with failure domains
- Data Flexibility: Moving data across regions
- Operator Compatibility: Supporting general operators to be multi-cluster aware

Please scan the QR Code above to leave feedback on this session