KubeCon | CloudNativeCon
North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

One VTOrc to Rule Them All

High Availability In a Distributed Database System

Deepthi Sigireddi, Manan Gupta

PlanetScale

Vitess

@vitessio

Deepthi Sigireddi

Technical Lead

@ATechGirl

Manan Gupta

Maintainer

@GuptaManan100

Vitess Overview

What is Vitess?

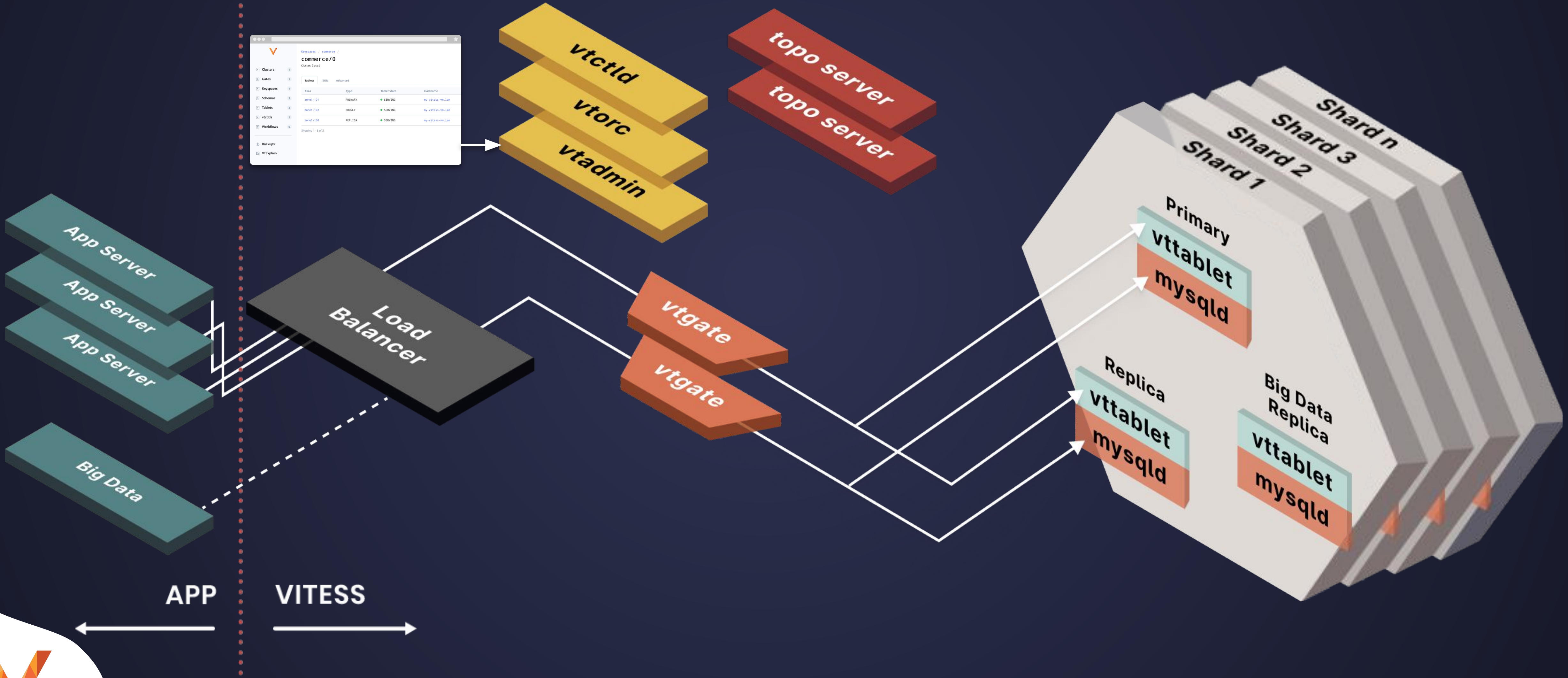
Cloud Native Database

Massively Scalable

Highly Available

MySQL 5.7/8.0 Compatible

Architecture



Keyspaces / commerce /

commerce/0

Cluster: local

Alias	Type	Tablet State	Hostname
zone1-101	PRIMARY	● SERVING	my-vitess-vm.lan
zone1-102	RDONLY	● SERVING	my-vitess-vm.lan
zone1-100	REPLICA	● SERVING	my-vitess-vm.lan

Showing 1 - 3 of 3

Vitess

@vitessio

KubeCon | CloudNativeCon North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

VTOrc Overview

Vitess

@vitessio

Problem Statement

- Resiliency to MySQL failures
- High availability
- Data Durability
- Minimize downtime / recovery time

Before there was VTOrc ...

- openark/orchestrator
- Integrated with Vitess
- enable_semi_sync flag on VTTablet
- Works well enough most of the time
- But not all the time

VTOrc

- VTOrc is now GA (v15)
- In production

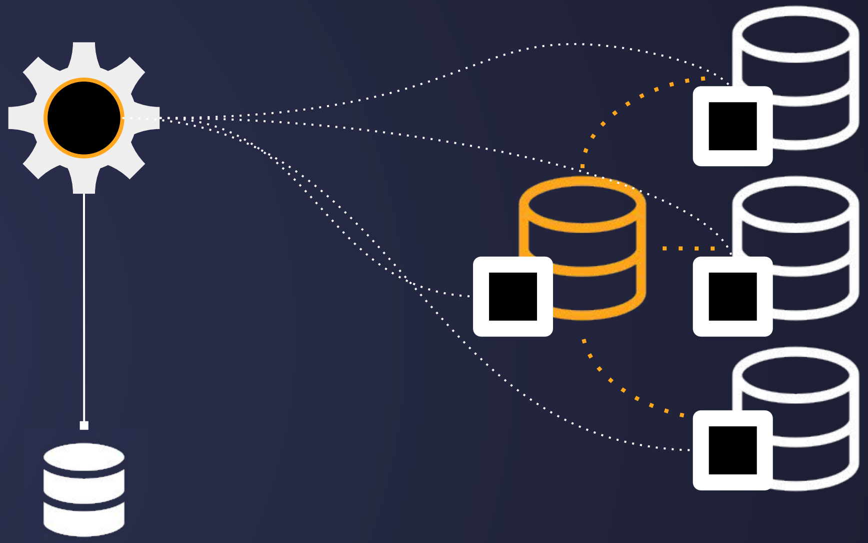
Using the Vitess Kubernetes operator and [..] orchestrator (VTOrc) has been very pleasant, as it removed a lot of operational overhead. Vitess seems rock solid so far, and we look forward to seeing what the future will bring for this awesome project.

- Principal Ops Architect, AAA gaming studio

Vitess

VTOrc

- VTOrc is the agent that detects failures
- Durability through Replication
  - Policies allow trade-offs
- High availability through failover
  - Planned / unplanned leader election



Design Principles

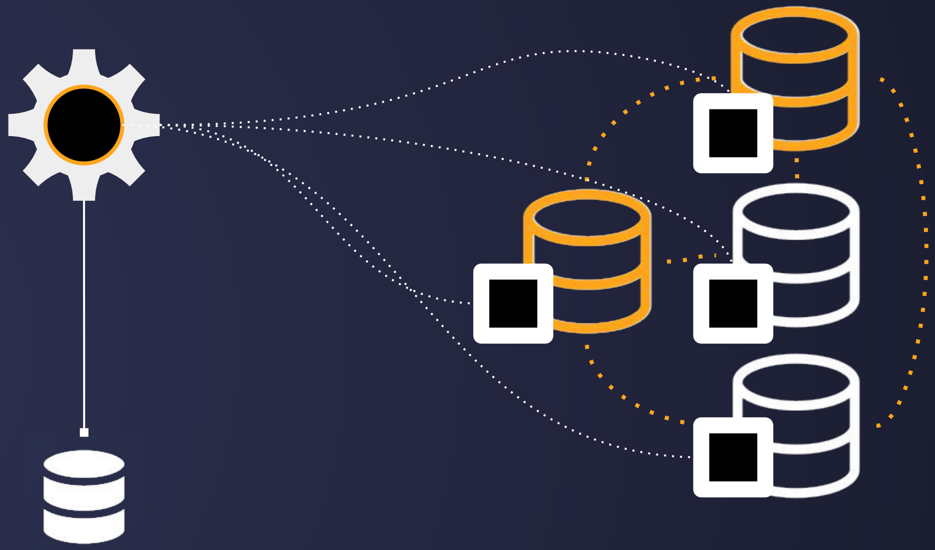
- Engineering approach
- Single leader system
- Fulfill requests while respecting durability policy
- Leader election process
  - Planned versus unplanned
- Forward Progress
- Race conditions

Leader Election

- Revocation
- Election
- Propagation

Planned Leader Election

- Revocation
  - Current leader is asked to step down
- Leader selection
  - A new leader is chosen
- Propagation
  - Completed requests

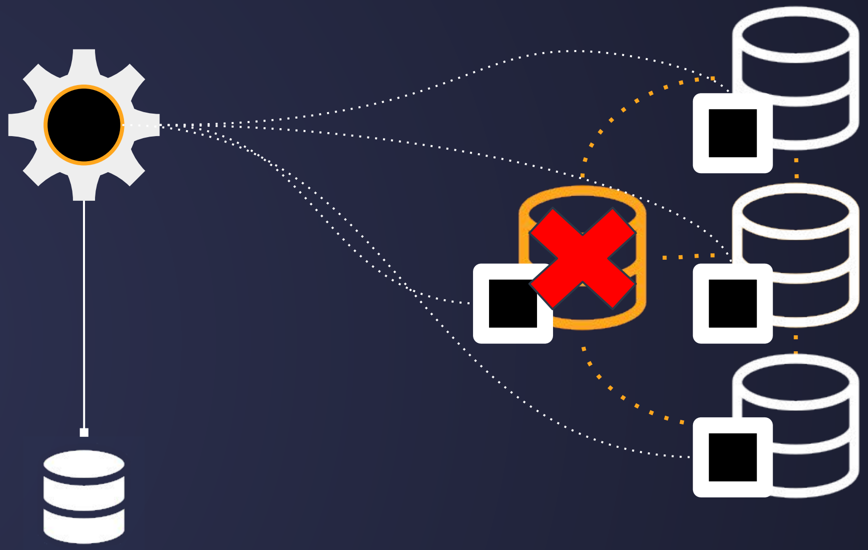


Vitess

@vitessio

Unplanned Leader Election

- Revocation
  - Reach “m” followers
- Leader selection
  - A new leader is chosen
  - Based on durability policy
- Propagation
  - Completed requests

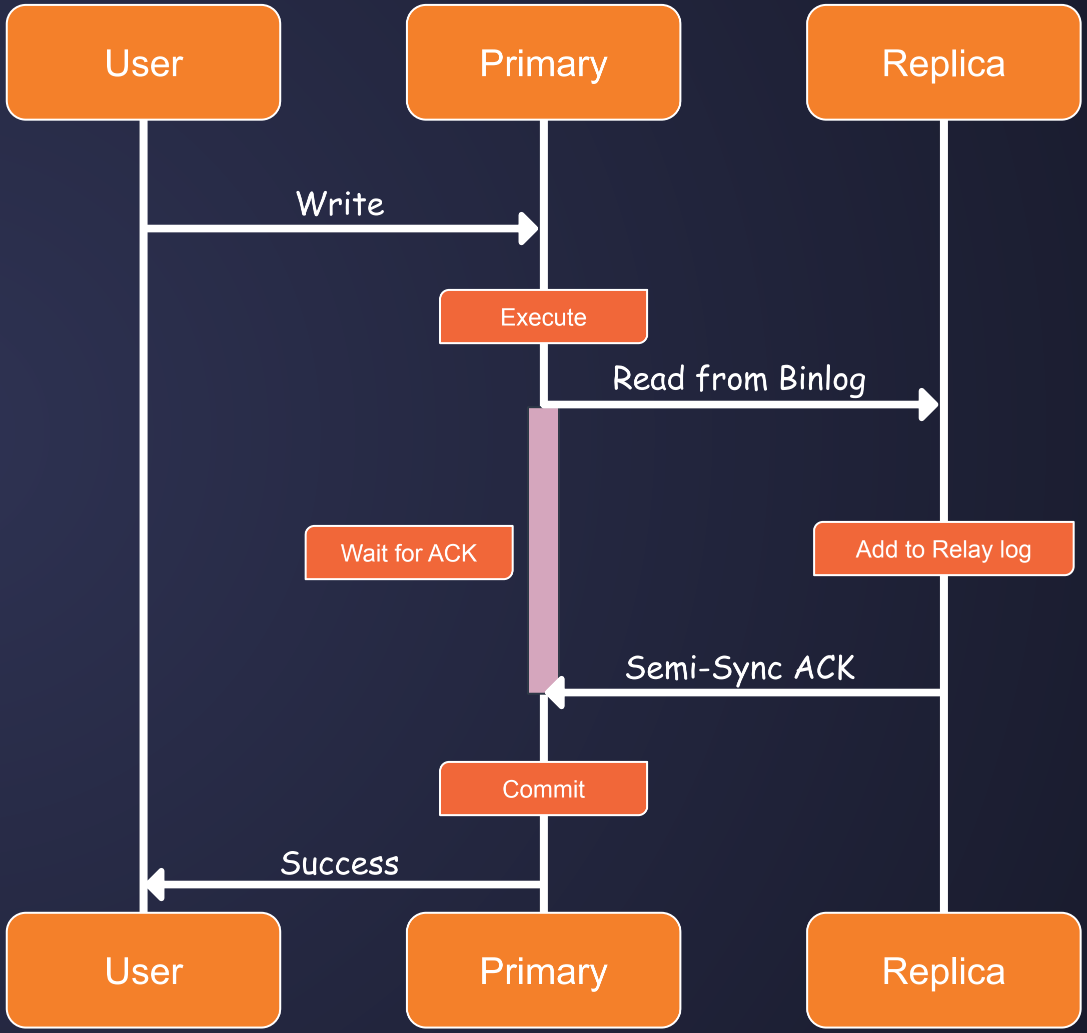


Vitess

@vitessio

Durability Policies & Semi-Sync

- Semi-Sync in MySQL
- Durability Policy
  - Who can be the primary?
  - How many semi-sync ACKs required for each primary?
  - Who can send these ACKs?
- Increased Flexibility



Vitess

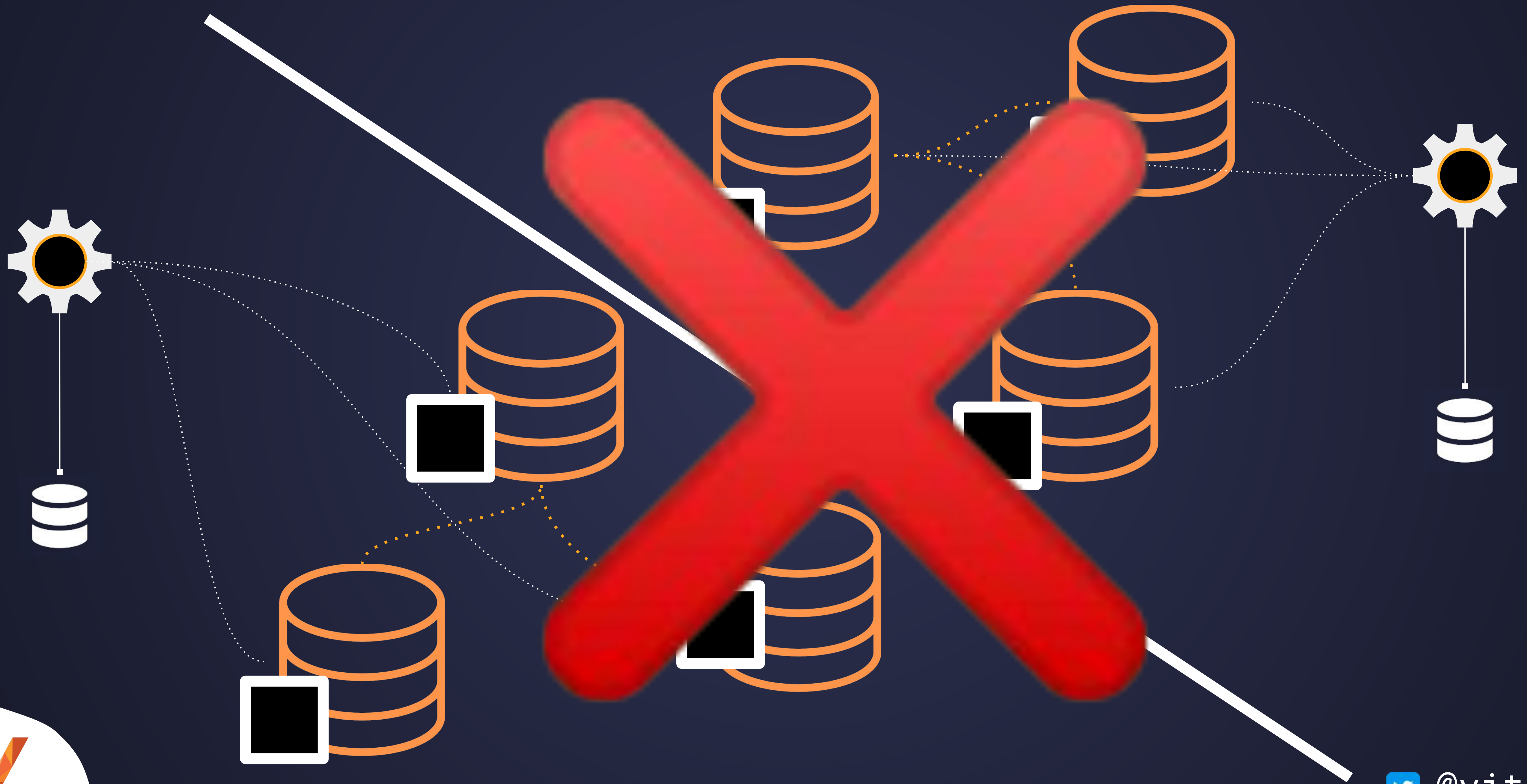
@vitessio

Revocation and Quorum

- What is “m”?
- How do we know we have reached sufficient tablets to guarantee safety?

- Intersecting Quorum
- Quorum for accepting transactions
- Quorum for revocation

Revocation and Quorum



Vitess

@vitessio

Revocation and Quorum

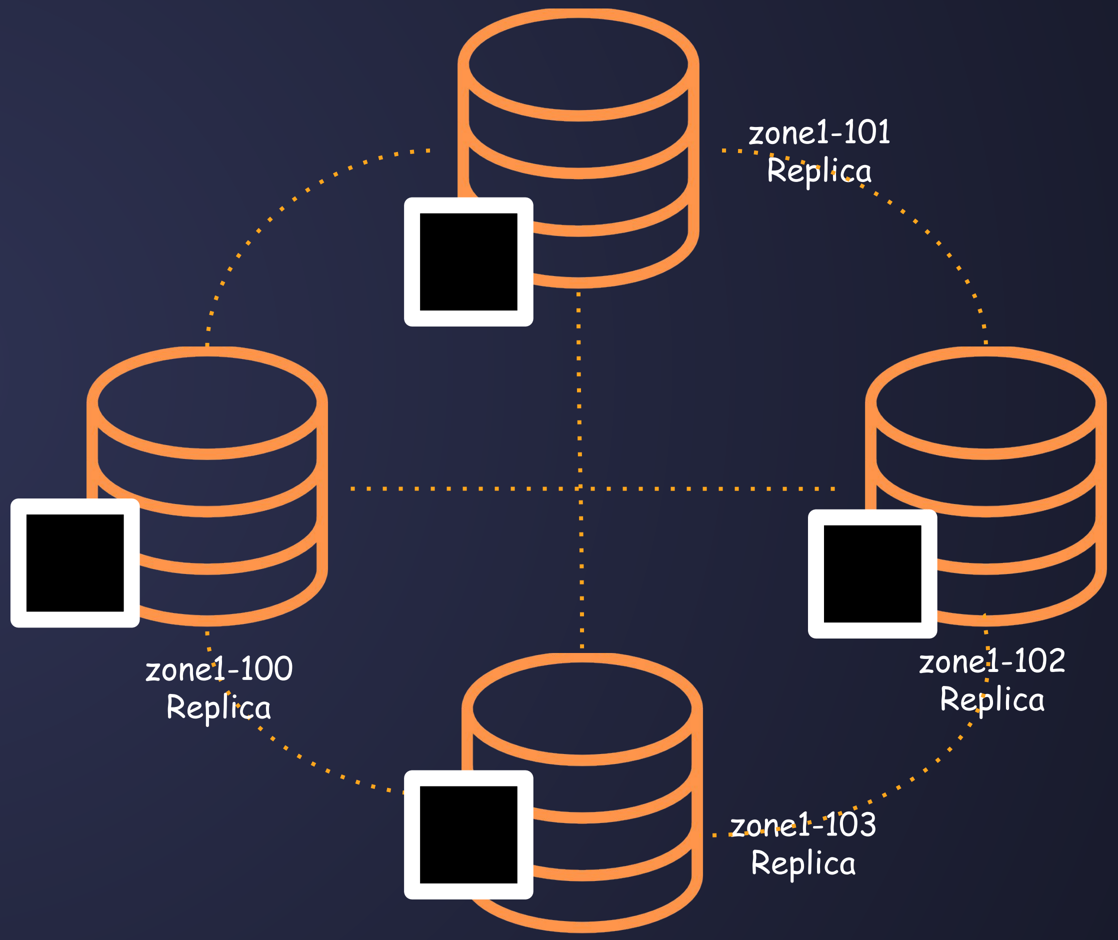
Vitess

@vitessio

Demo

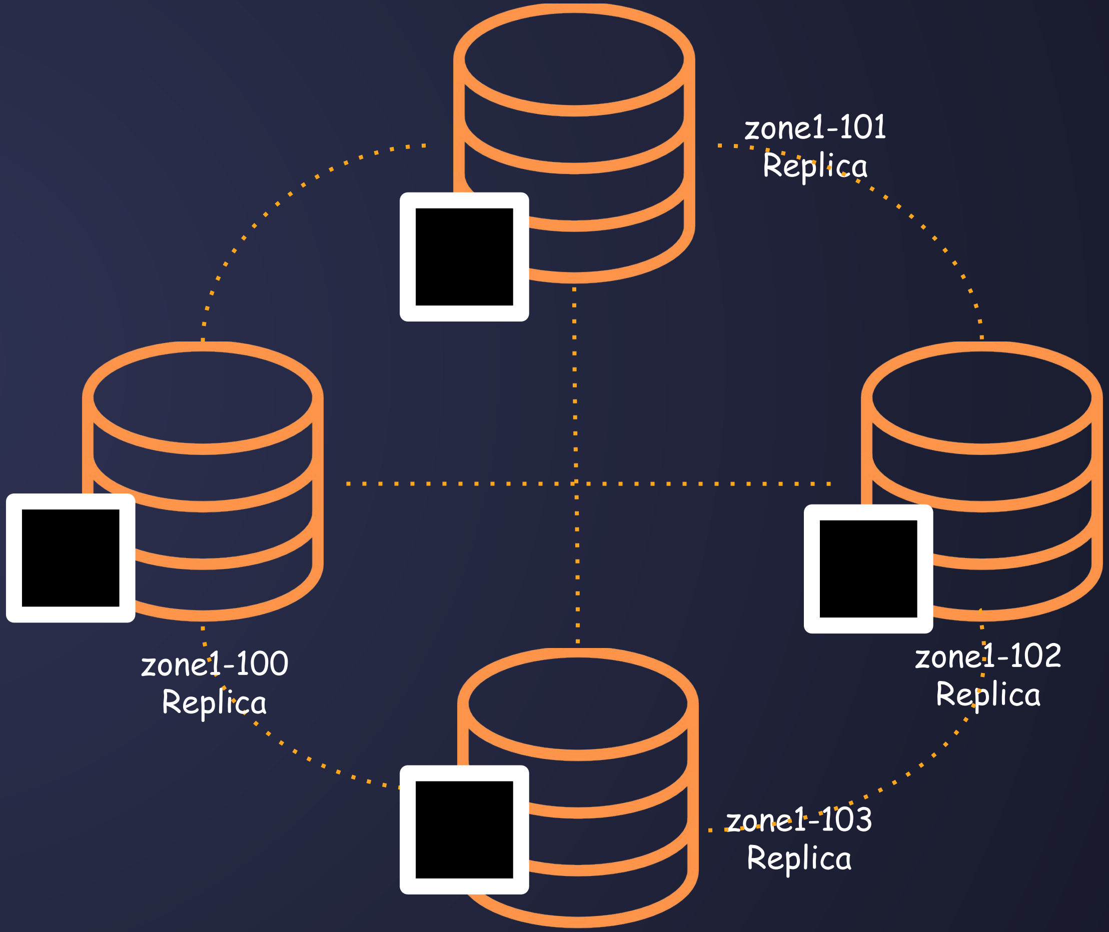
Semi-Sync Durability

- `semi_sync_ks`
- Durability Policy - semi-sync
  - Any replica can be the primary
  - 1 semi-sync ACK required
  - Any replica can send the ACK



Revocation

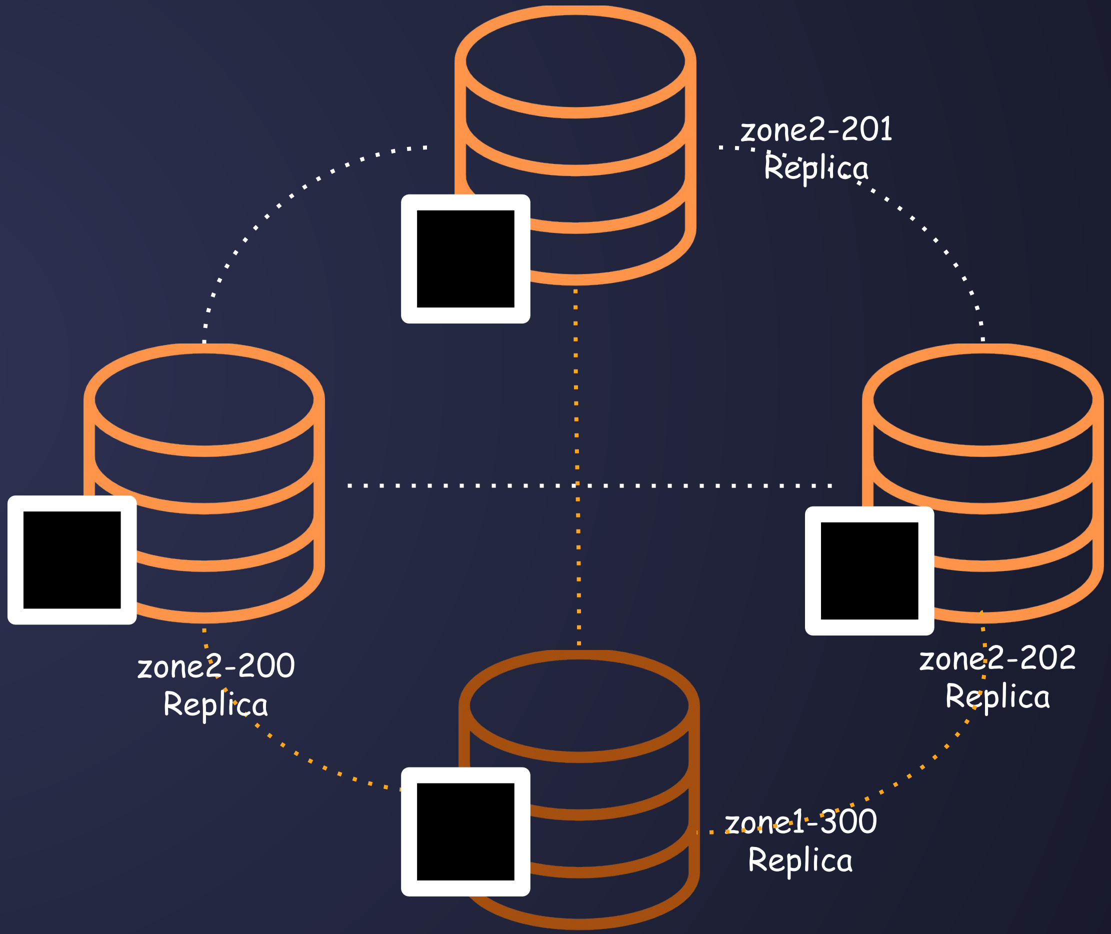
- **Quorums for Accepting Transactions**
  - [(**100**, 101), (**100**, 102), (**100**, 103)]
  - [(**101**, 100), (**101**, 102), (**101**, 103)]
  - [(**102**, 100), (**102**, 101), (**102**, 103)]
  - [(**103**, 100), (**103**, 101), (**103**, 102)]
- **Quorums for Revocations -**
  - [100, 103] ❌
  - [100, 102, 103] ✅
  - [100, 101, 102, 103] ✅
  - [101] ❌



Cross-Cell Durability

- `cross_cell_ks`
- Cell = Failure Domain

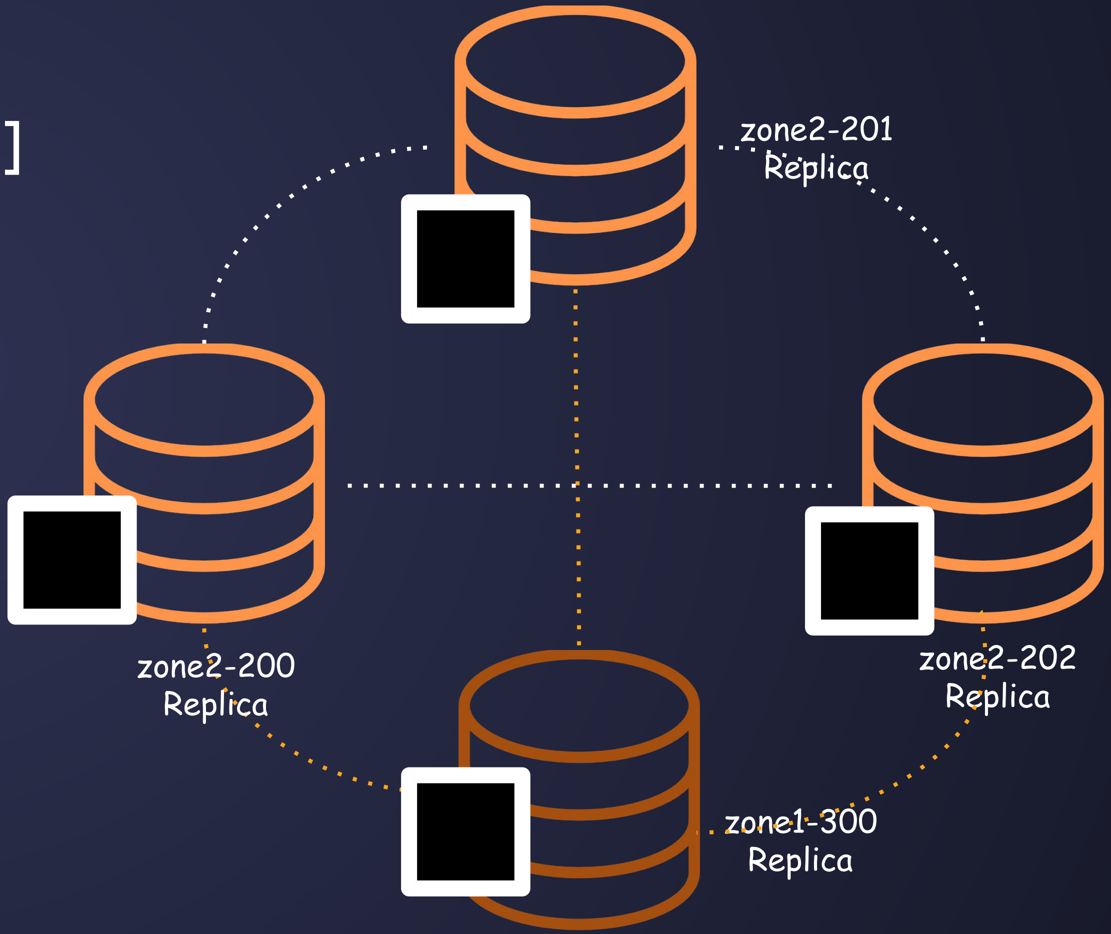
- Durability Policy - cross-cell
  - Any replica can be the primary
  - 1 semi-sync ACK required
  - A replica from a different cell can send an ACK



Revocation

- **Quorums for Accepting Transactions**
  - [(**300**, 200), (**300**, 201), (**300**, 202)]
  - [(**200**, 300)]
  - [(**201**, 300)]
  - [(**202**, 300)]

- **Quorums for Revocations -**
  - [200, 201] ❌
  - [300, 202, 201] ✅
  - [300, 201] ✅
  - [300, 200, 201, 202] ✅



Vitess

@vitessio

Custom Durability Policies

- Durability Policy
  - Who can be the primary?
  - How many semi-sync ACKs required for each primary?
  - Who can send these ACKs?
- Increased Flexibility

Vitess

@vitessio

Custom Durability Policies

- **Quorums for Accepting Transactions**
  - [(**A**, B, C), (**A**, B, D), (**A**, C, D)]
  - [(**B**, D), (**B**, A)]
  - [(**C**, B)]

- **Quorums for Revocations -**
  - [A, D] ❌
  - [A, B, D] ✅
  - [B, C] ✅
  - [C, D] ❌

Vitess

@vitessio

More Failure Scenarios

- Maintains the cluster’s desired state
- Primary is Read-Only
- Replica’s replication is stopped
- Replica is writable
- Semi-sync settings are incorrect
- Shard has no primary
- Primary is replicating from a different tablet

Vitess

Q & A

Vitess

Resources

Vitess: Introduction and New Features
Friday, Oct 28 11:55 am EDT

Blog Post Series

- https://planetscale.com/blog/blog-series-consensus-algorithms-at-scale-part-1

VTOrc Documentation

- https://vitess.io/docs/15.0/reference/vtorc/
- https://vitess.io/docs/15.0/user-guides/configuration-basic/vtorc/

https://sched.co/182J8

Vitess

@vitessio

Thank you!

Vitess

@vitessio