Xpand Multi-region High Availability

Matt White
Xpand Director of Engineering
MariaDB Corporation

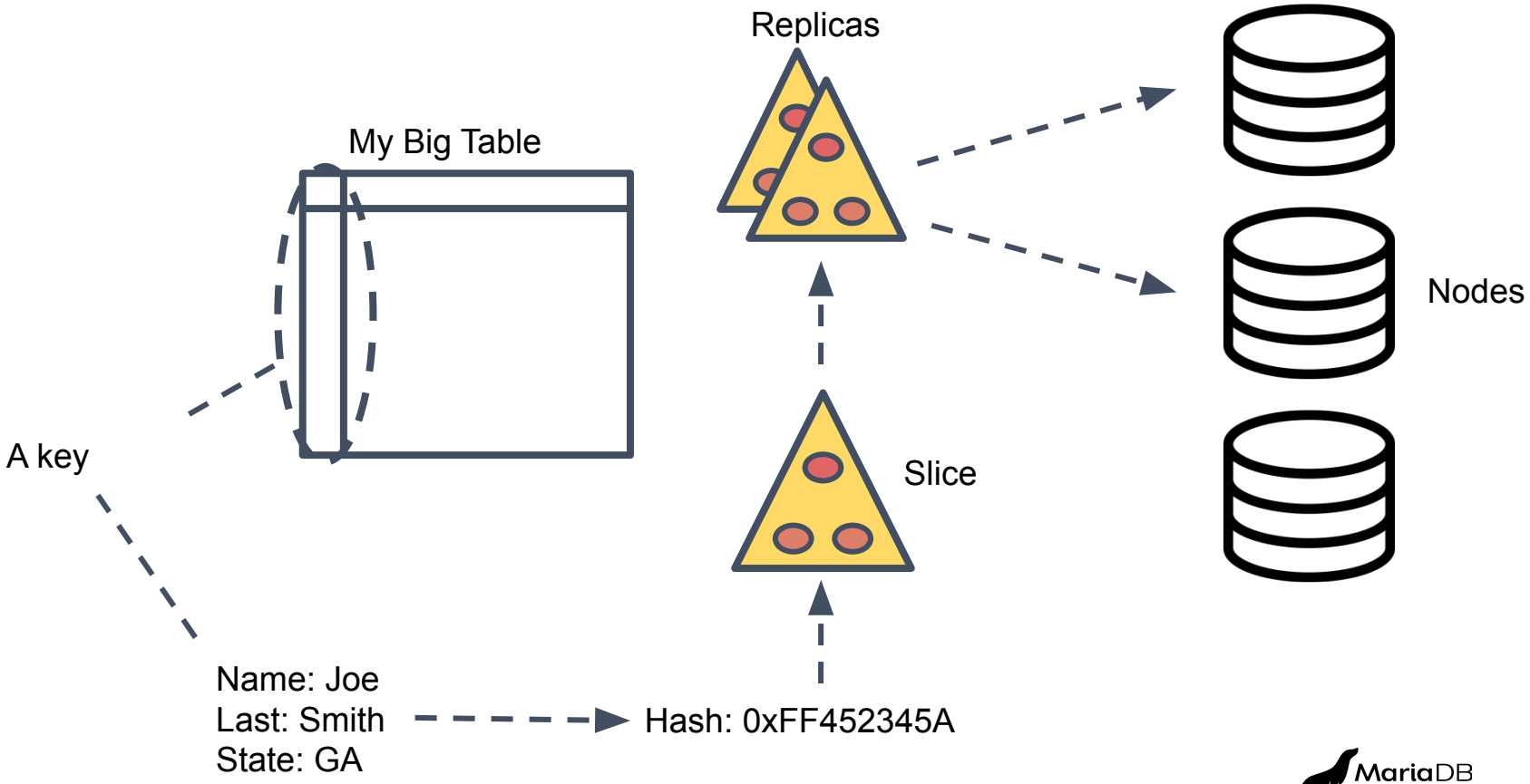
Overview

- Resiliency Background
  - Node Failure
  - Zone Failure
- Disaster Recovery
  - High Speed Backup
  - Asynchronous Replication
- Multi-Region Resiliency

MariaDB

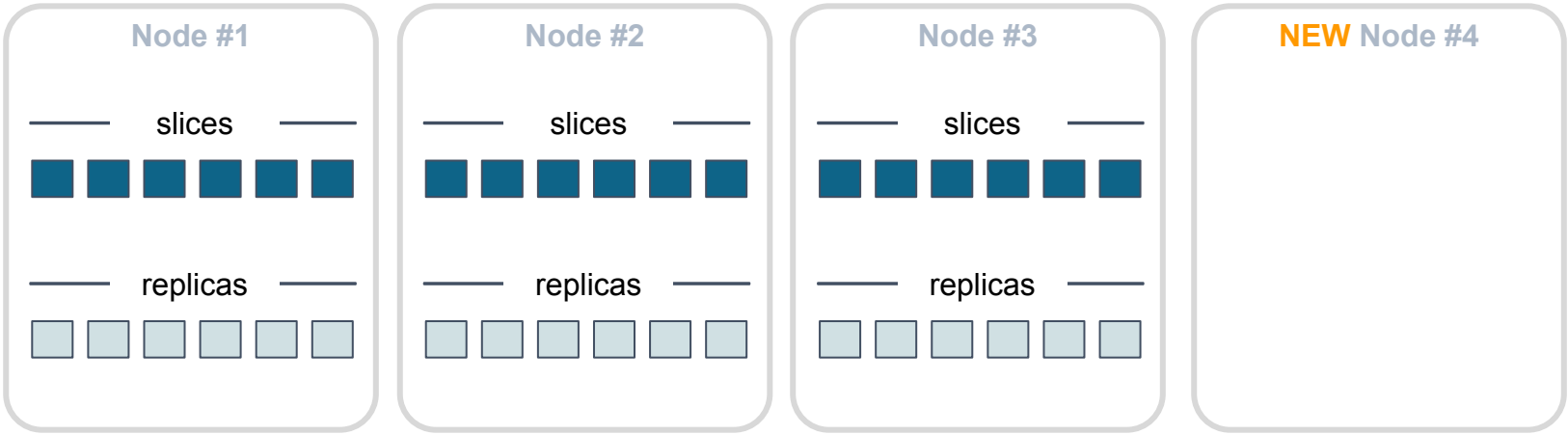
Xpand Resiliency Background

Slices distributed



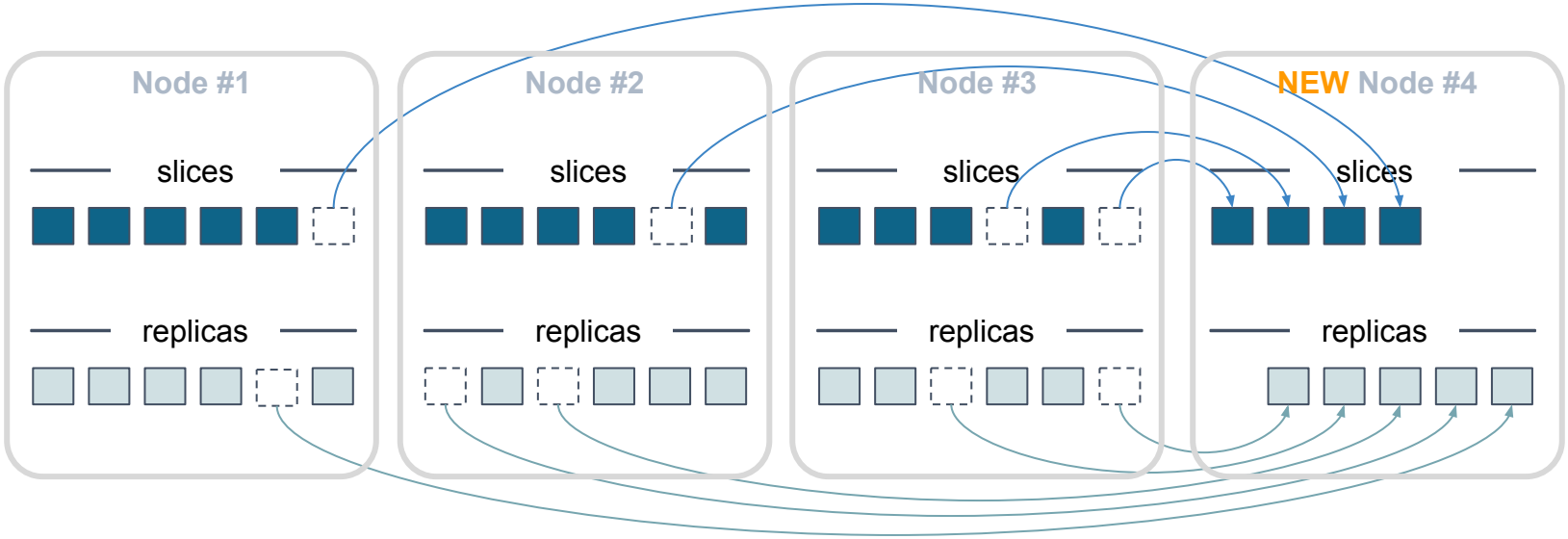
MariaDB

Adding Nodes



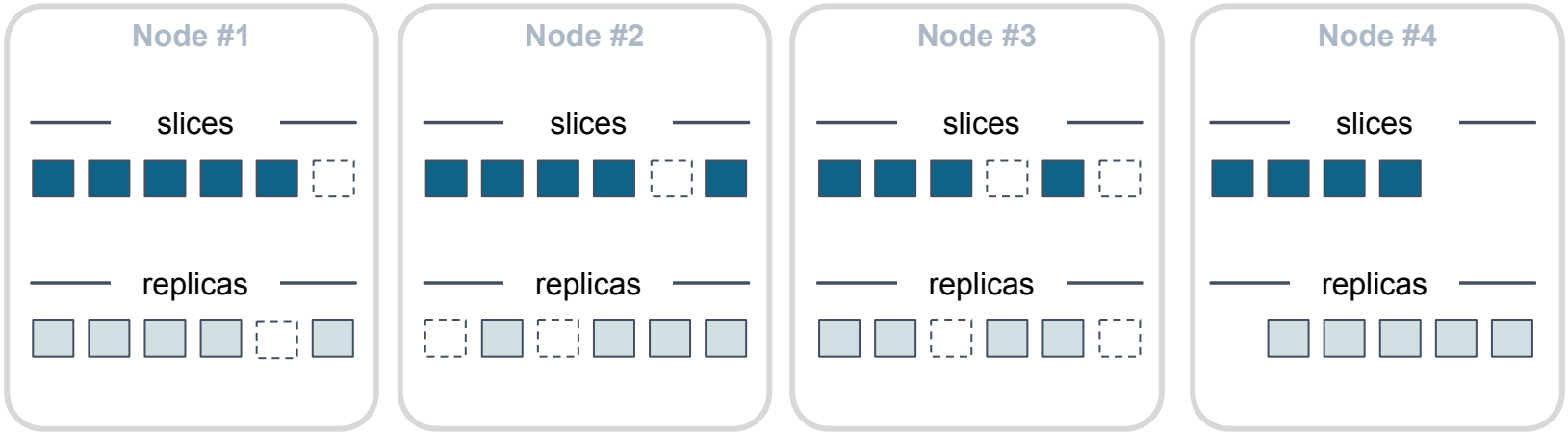
MariaDB

Adding Nodes



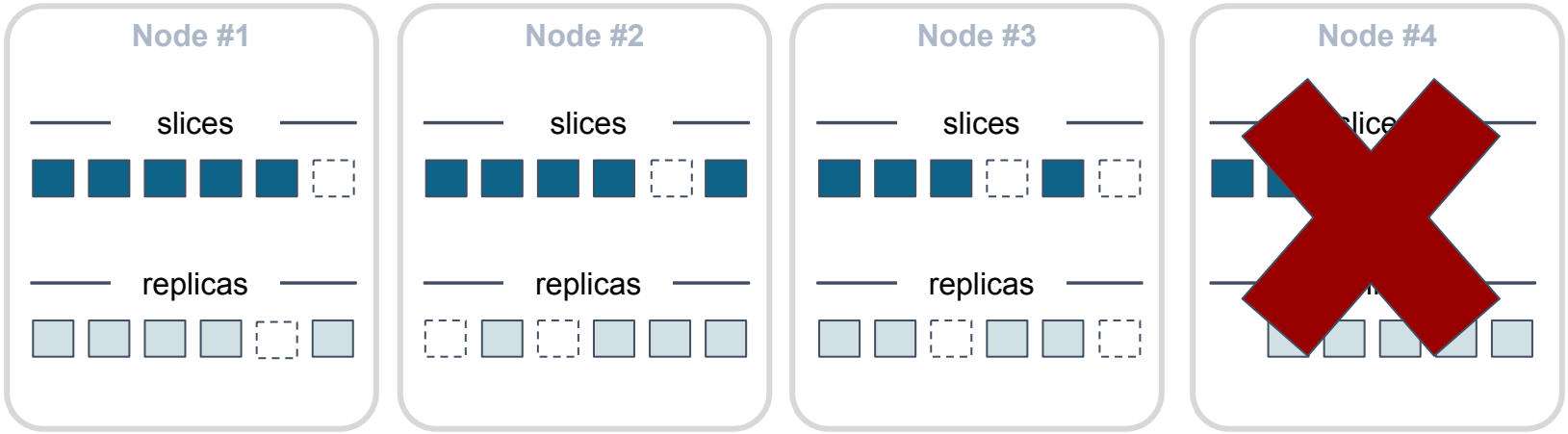
MariaDB

Self-healing

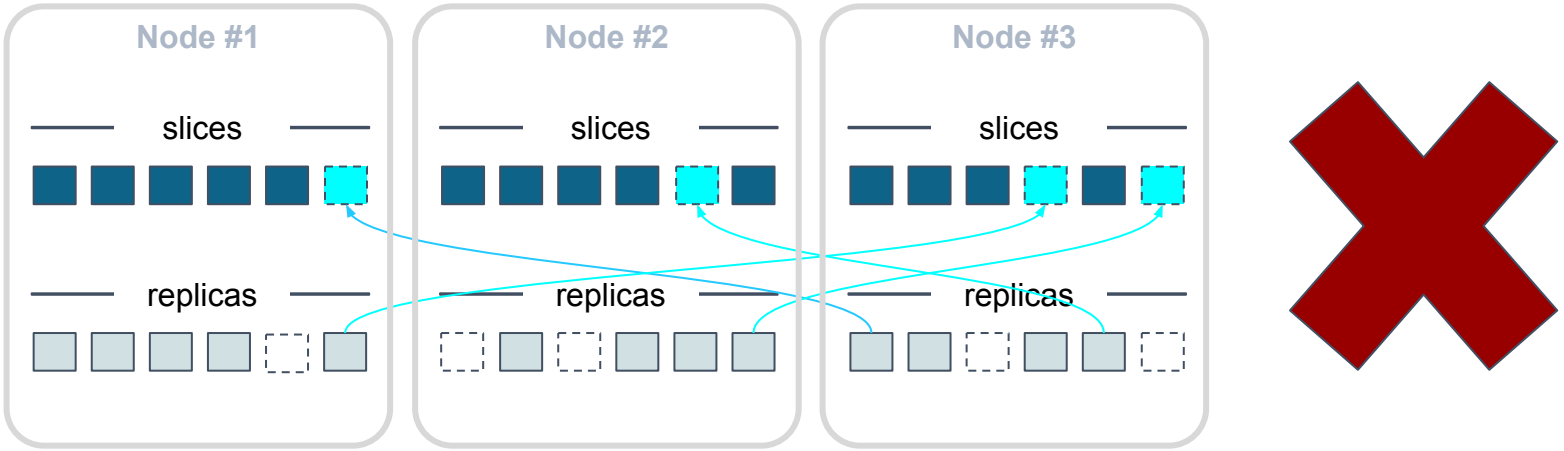


MariaDB

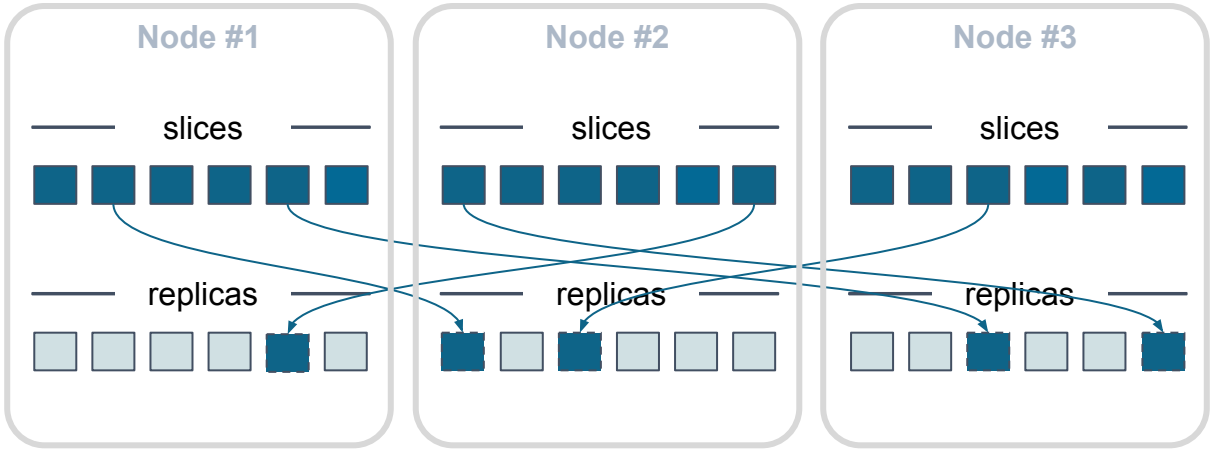
Self-healing - Temporary Failure



Self-healing



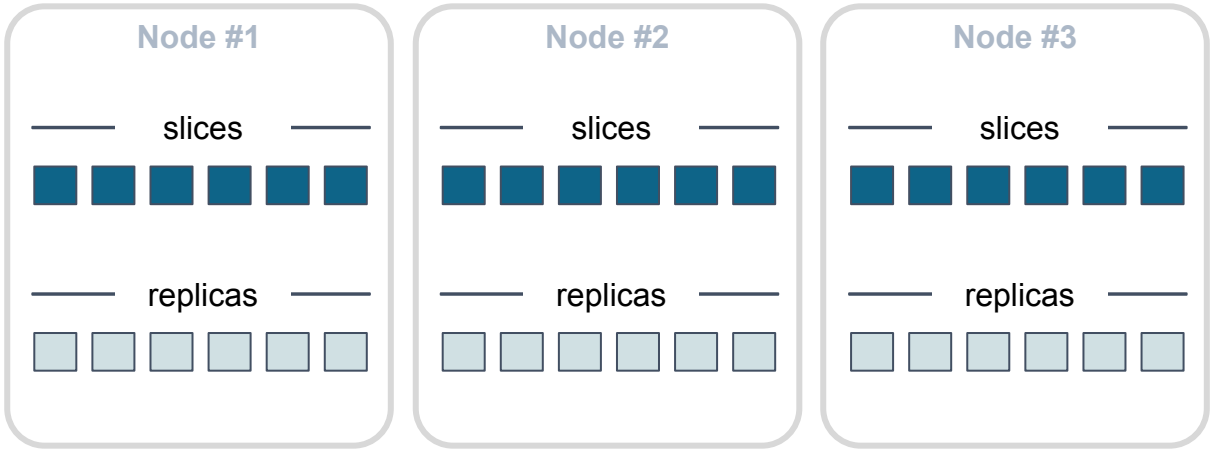
Self-healing



No transactions are blocked during these operations

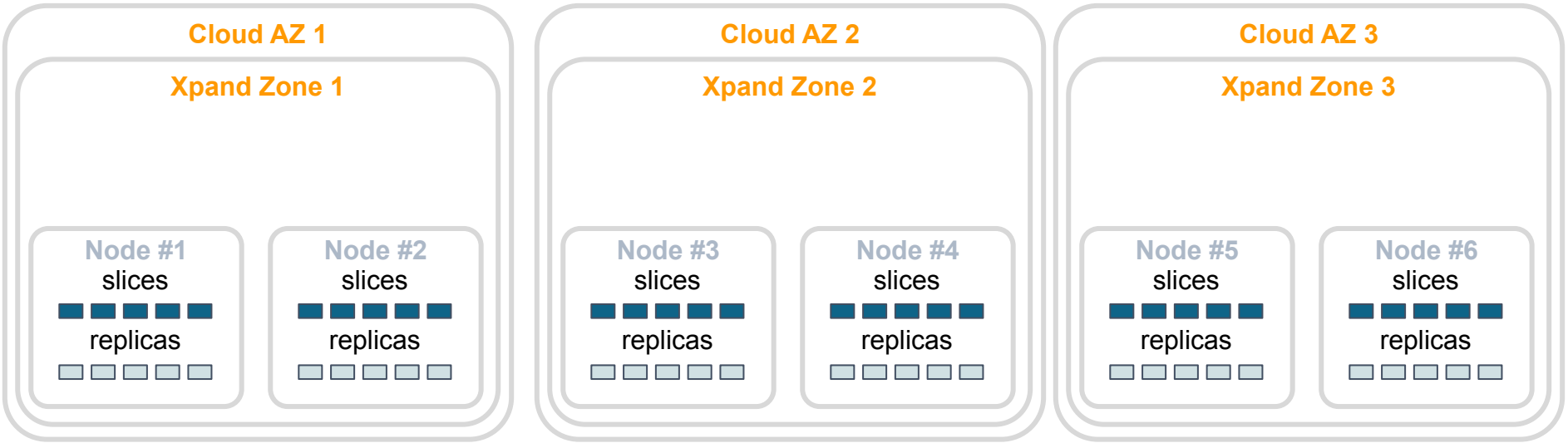
MariaDB

Self-healing



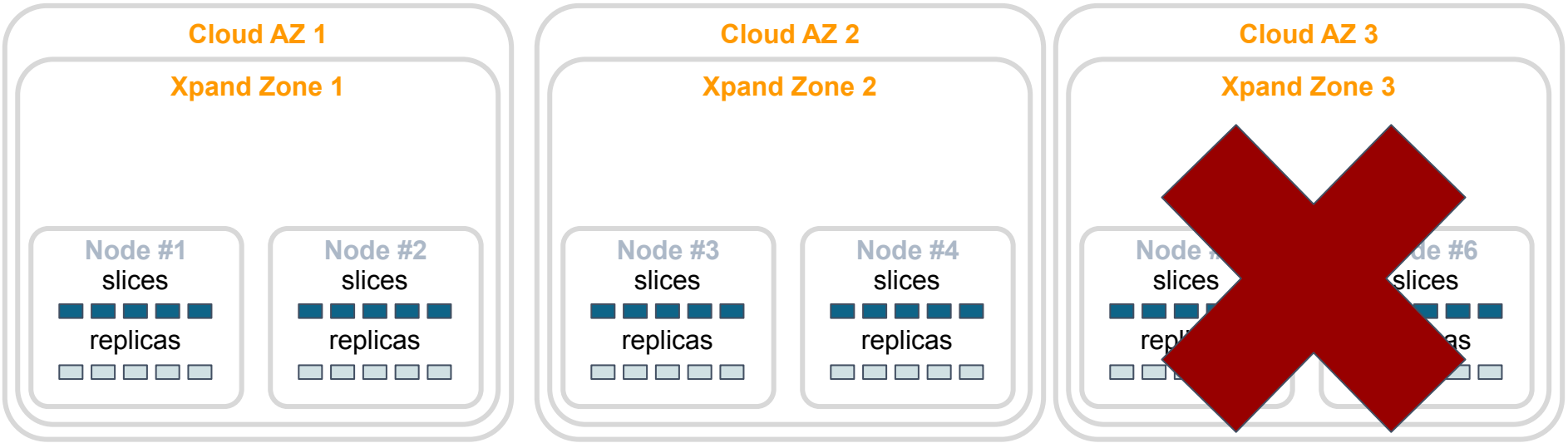
MariaDB

Zones



MariaDB

Zone outage


Cloud AZ 1
Xpand Zone 1
Node #1
slices
replicas
Node #2
slices
replicas
Cloud AZ 2
Xpand Zone 2
Node #3
slices
replicas
Node #4
slices
replicas
Cloud AZ 3
Xpand Zone 3
slices
slices

Key points

- DBA can specify
  - Initial number of slices
  - Max size of slice before re-slicing
  - Number of replicas (max number of failures to survive)
  - Use of zones or not
- Xpand self-manages
  - Hash ranges
  - Node assignment (***and re-assignment***)
    - Enforce resiliency policies
    - Workload balancing
    - Simplicity of administration

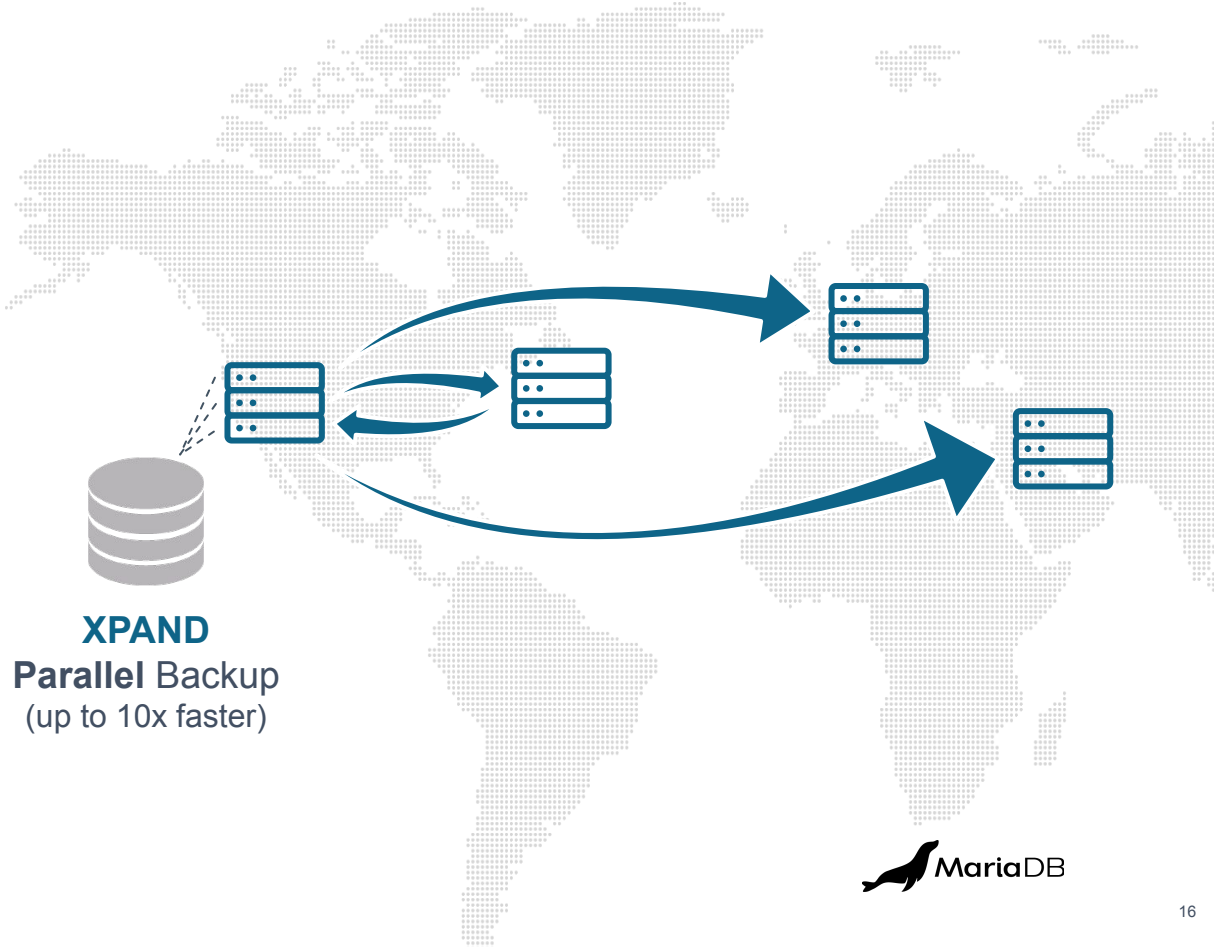
MariaDB

Disaster Recovery

Disaster Recovery: Backups

Replicate to any cloud, any datacenter, anywhere

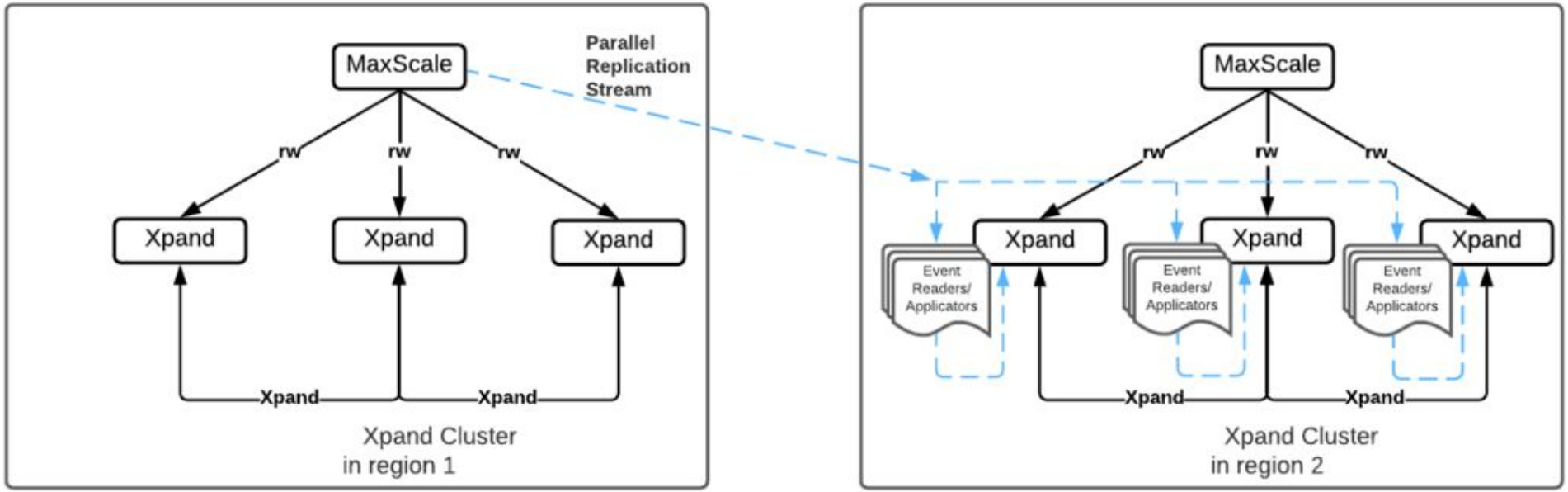
Asynchronous multi-point replication



XPAND
Parallel Backup
(up to 10x faster)

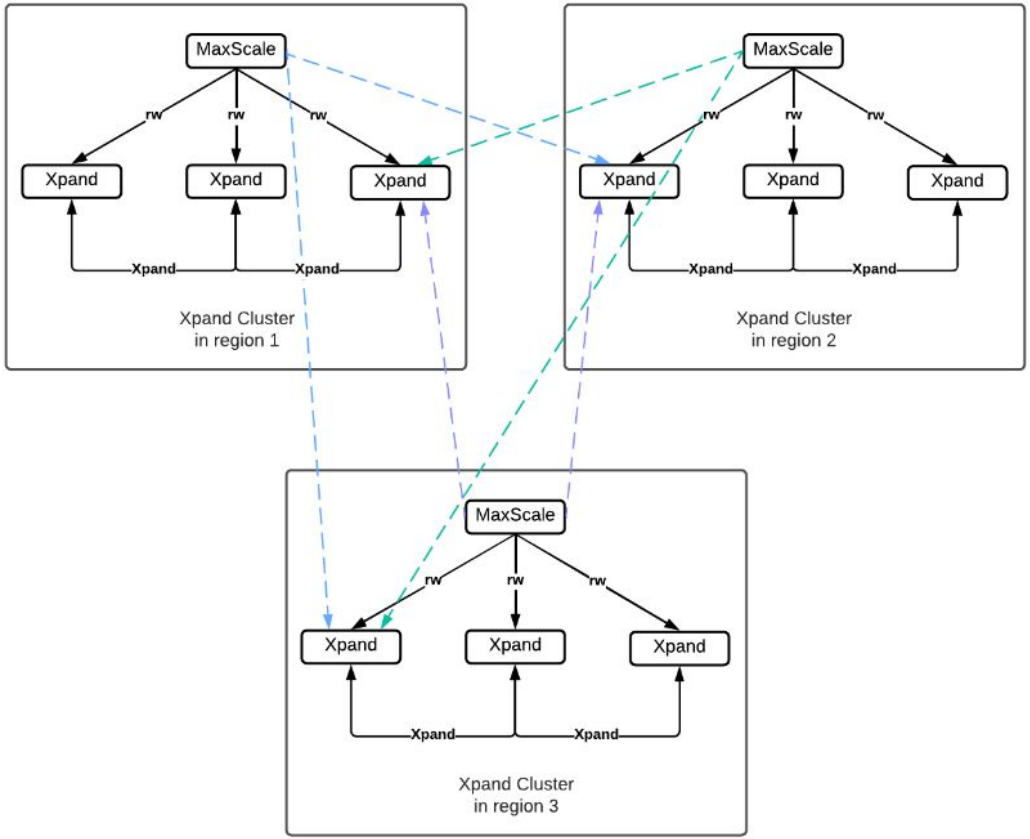
MariaDB

Disaster Recovery: Dual Region Replication



MariaDB

Disaster Recovery: Multi-region Replication



MariaDB

Xpand Multi-region High Availability Cluster

Paris Is Drowning: GCP's Region Failure in Age of Operational Resilience

The time is coming, and maybe sooner than we think, when regulators will require a standardized approach to resilience in the name of public good.

Apr 27th, 2023 1:15pm by Michelle Gienow

A Major Outage At AWS Has Caused Chaos At Amazon's Own Operations, Highlighting Cloud Computing Risks

MariaDB

High Availability Requirements

- RPO-0 - Recovery Point Objective - zero data loss
- Current state
  - Xpand provides RPO-0 for failure ***within a region*** (Node, zone domain failure)
  - Region domain failure RPO > 0
  - Complicated failover processes
    - Promote secondary cluster
    - Investigate potential data loss
- Objective
  - Survive **region** failure with RPO-0 (Recovery Point Objective - zero data loss)
  - Simplified, automatic failover

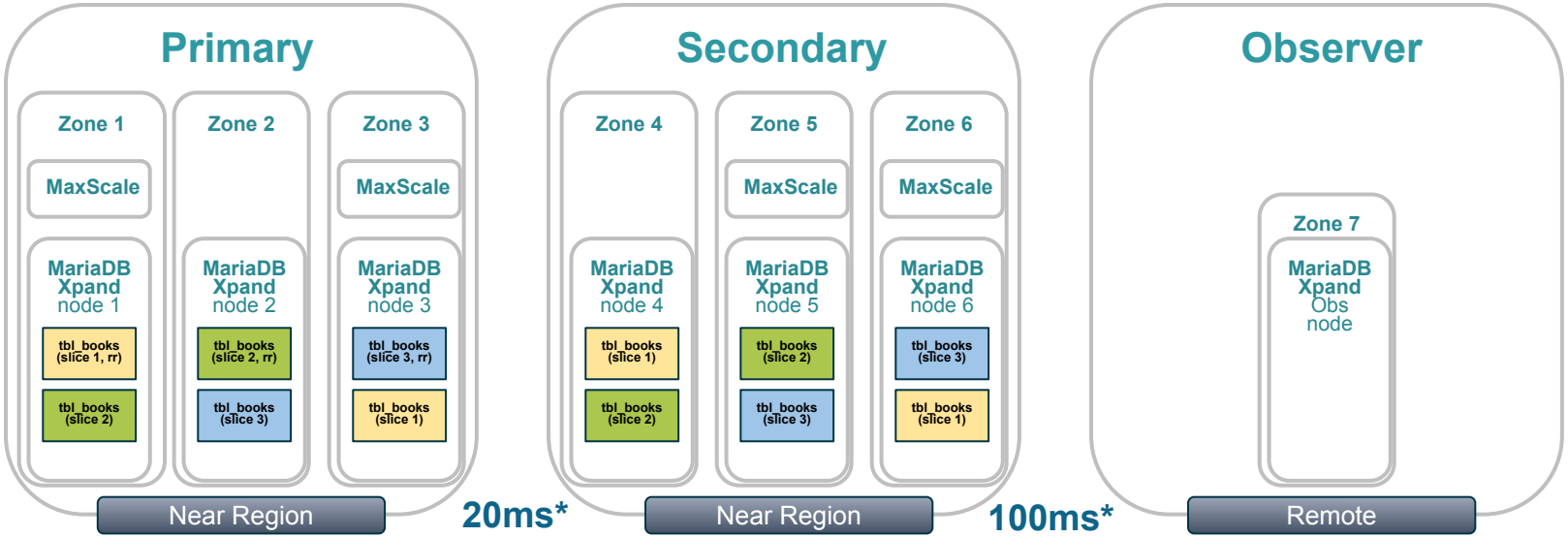
MariaDB

Our Solution

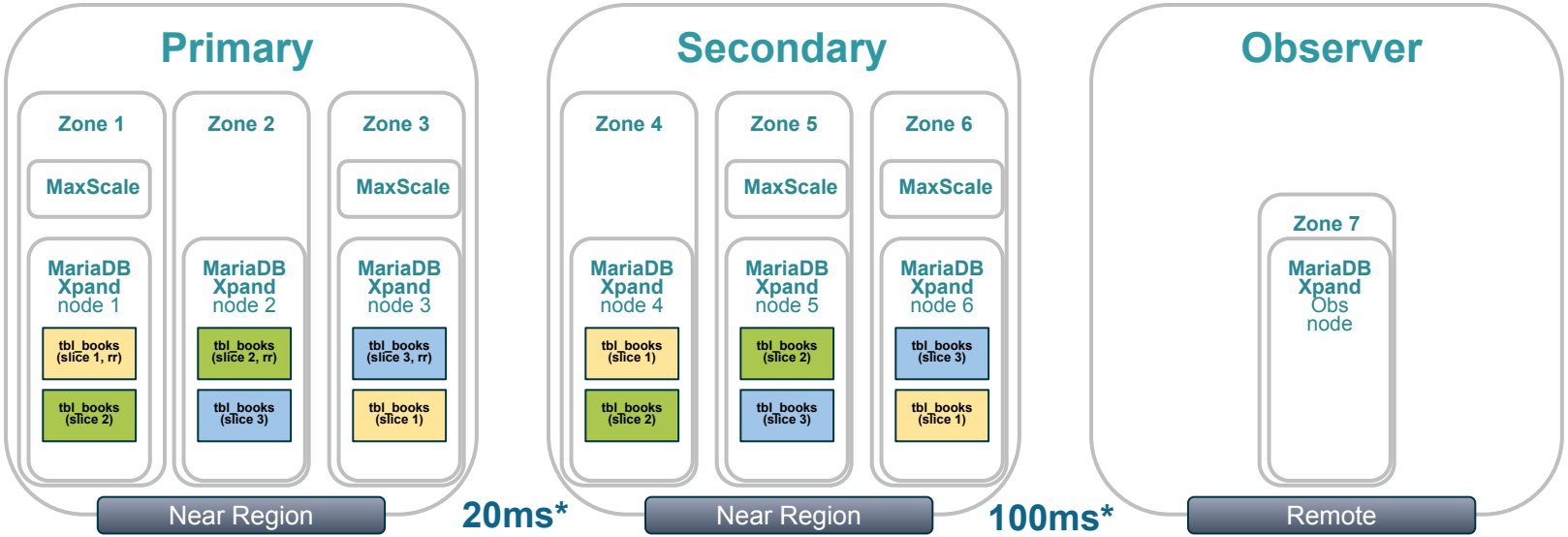
- Deliver RPO-0 for region failures
  - Consensus based redundancy management with a near stand-by region
  - Remote observer region to arbitrate surviving region and prevent split-brain
  - Applications connect to a primary MaxScale directed to the primary Xpand region
  - Writes performed concurrently in both regions (stand-by writes replicas)
  - Reads performed in primary region
  - Primary may be divided into availability zones for additional resilience
- Automated failover for region failure
  - Xpand nodes in secondary region form cluster with arbiter node providing quorum
  - Applications fail over to a secondary MaxScale which connects to the secondary region

MariaDB

Phase 1 - Near region cluster (Mid2023)



Phase 2 - Local reads in both regions



Engineering Prototype Performance

6 AWS-East1, 6 AWS-East2 Sysbench 90:10

Nodes	Concurrency	Throughput	Average Elapsed ms	stddev	p95_elapsed_ms
12	256	5412	47.29	6.14	54.25
12	512	10663	48	6.12	55.15
12	1024	20276	50.47	6.53	58.83
12	2048	**28551**	71.68	29.19	**107.2**

MariaDB

Development

- New cluster management DDL
  - CREATE/CHANGE/DROP REGION [PRIMARY, SECONDARY, OBSERVER]
  - CHANGE REGION [PRIMARY, SECONDARY, OBSERVER]
  - CREATE/CHANGE/RENAME/DROP ZONE [in REGION]
  - CHANGE NODE TO [zone]
- Resiliency changes
  - Assignment of data, acceptors becomes region aware
  - When uneven replicas requested (maxfailures 2, 3 replica) overload primary

Summary

- Phase 1 planned for Mid-2023
- Extend inherent resiliency architecture to region
  - Nodes assigned to zones
  - Zones assigned to regions
  - Primary/secondary region assigned
  - Data and acceptors automatically distributed (and redistributed) across regions
- On region failure
  - Observer node automates identification of surviving region
  - Transactions satisfied by surviving nodes in surviving region
  - Data reprotected in surviving region

MariaDB