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
Xpand Resiliency Background
Slices distributed

My Big Table

Replicas

Nodes

A key

Name: Joe
Last: Smith
State: GA

Hash: 0xFF452345A
Adding Nodes

Node #1

- slices
- replicas

Node #2

- slices
- replicas

Node #3

- slices
- replicas

NEW Node #4

- slices
- replicas
Adding Nodes
Self-healing

Node #1
- slices
- replicas

Node #2
- slices
- replicas

Node #3
- slices
- replicas

Node #4
- slices
- replicas
Self-healing - Temporary Failure

Node #1
- slices
- replicas

Node #2
- slices
- replicas

Node #3
- slices
- replicas

Node #4
- slices
- replicas

MariaDB
Self-healing

Node #1

Node #2

Node #3

slices
replicas
slices
replicas
slices
replicas
Self-healing

No transactions are blocked during these operations
Self-healing

Node #1

- Slices
- Replicas

Node #2

- Slices
- Replicas

Node #3

- Slices
- Replicas
Zones

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
Node #5
- slices
- replicas

Node #6
- slices
- replicas
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
Node #5
slices
replicas
Node #6
slices
replicas

MariaDB
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
Disaster Recovery
Disaster Recovery: Backups

Replicate to any cloud, any datacenter, anywhere

Asynchronous multi-point replication

XPAND Parallel Backup (up to 10x faster)
Disaster Recovery: Dual Region Replication
Disaster Recovery: Multi-region Replication
Xpand Multi-region
High Availability Cluster

MariaDB
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
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
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
Phase 1 - Near region cluster (Mid2023)

**Primary**
- Zone 1: MaxScale
- Zone 2: MariaDB Xpand node 1
  - tbl_books (slice 1, rr)
- Zone 3: MariaDB Xpand node 2
  - tbl_books (slice 2, rr)
- Zone 4: MariaDB Xpand node 3
  - tbl_books (slice 3, rr)
- Near Region: 20ms*

**Secondary**
- Zone 4: MaxScale
- Zone 5: MariaDB Xpand node 4
  - tbl_books (slice 1)
- Zone 6: MariaDB Xpand node 5
  - tbl_books (slice 2)
- Zone 7: MariaDB Xpand node 6
  - tbl_books (slice 3)
- Near Region: 100ms*

**Observer**
- Zone 7: MariaDB Xpand Obs node
- Remote: 100ms*
Phase 2 - Local reads in both regions

Primary

Zone 1
MaxScale
MariaDB Xpand node 1

tbl_books (slice 1)

Near Region

Zone 2
MaxScale
MariaDB Xpand node 2

tbl_books (slice 2)

20ms*

Zone 3
MaxScale
MariaDB Xpand node 3

tbl_books (slice 3)

Near Region

Secondary

Zone 4
MaxScale
MariaDB Xpand node 4

tbl_books (slice 1)

Near Region

Zone 5
MaxScale
MariaDB Xpand node 5

tbl_books (slice 2)

100ms*

Zone 6
MaxScale
MariaDB Xpand node 6

tbl_books (slice 3)

Observer

Zone 7
MaxScale
MariaDB Xpand Obs node

Remote
# Engineering Prototype Performance

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

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Concurrency</th>
<th>Throughput</th>
<th>Average Elapsed ms</th>
<th>stddev</th>
<th>p95_elapsed_ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>256</td>
<td>5412</td>
<td>47.29</td>
<td>6.14</td>
<td>54.25</td>
</tr>
<tr>
<td>12</td>
<td>512</td>
<td>10663</td>
<td>48</td>
<td>6.12</td>
<td>55.15</td>
</tr>
<tr>
<td>12</td>
<td>1024</td>
<td>20276</td>
<td>50.47</td>
<td>6.53</td>
<td>58.83</td>
</tr>
<tr>
<td>12</td>
<td>2048</td>
<td>28551</td>
<td>71.68</td>
<td>29.19</td>
<td>107.2</td>
</tr>
</tbody>
</table>
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