

Building Resilient In-Memory Data Grids

...

June 19, 2019

52.60 Minutes

About the presenters



Ranvirsinh Raol

Technical Lead @ Capital One



Terry Walters

Senior Solution Architect @ Hazelcast

Let's start with basics

Definition of *Resilience*:

Resilience in software describes its ability to withstand stress and other challenging factors to continue performing its core functions and avoid loss of data.

“Resilience is the ability to provide required capability in the face of adversity.”

“Resilience is the ability of system to absorb external stress.”

Achieving resilience include...

- ❑ Recognize, Avoid & Build foresight
- ❑ Defend & Withstand
- ❑ Recover
- ❑ Adapt & Evolve

What we are going to cover today...

Context : Distributed Map holding Large quantity of Data in public cloud

- ❑ Infrastructure based Resiliency
- ❑ Resilient Connectivity
- ❑ Data Resilience
- ❑ Operational Resilience
- ❑ Monitoring

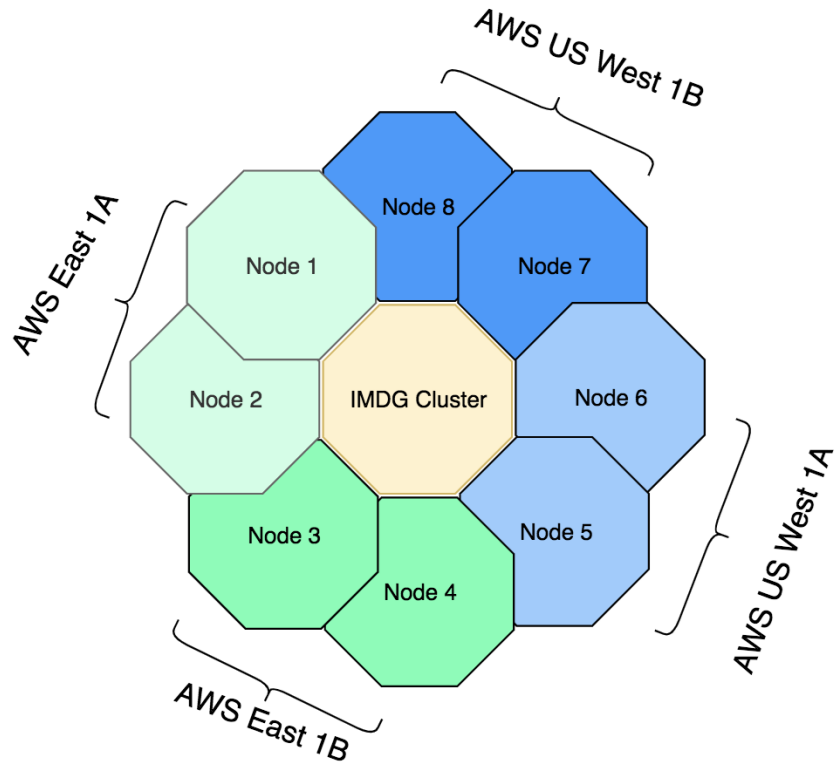
Resiliency Scope

Infrastructure based Resiliency

- ❑ How the Data is spread across the Systems/ Machines & Data Centers...
- ✓ Isolate your data from machine(s) or Datacenter outage.

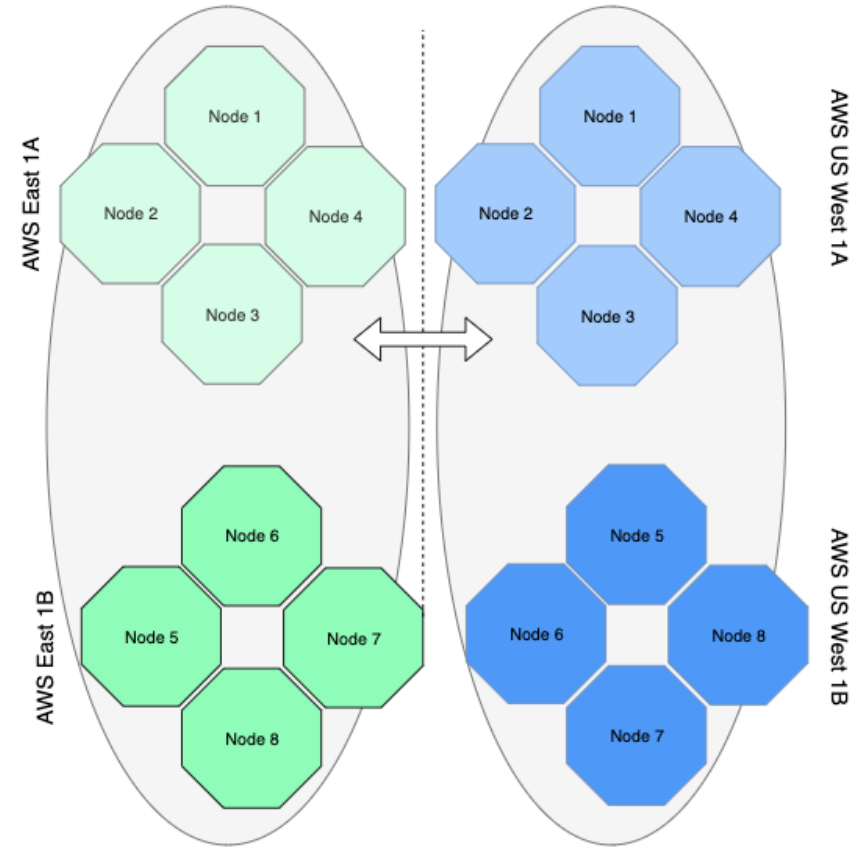
Infrastructure based
Resiliency

Cross Region IMDG Cluster



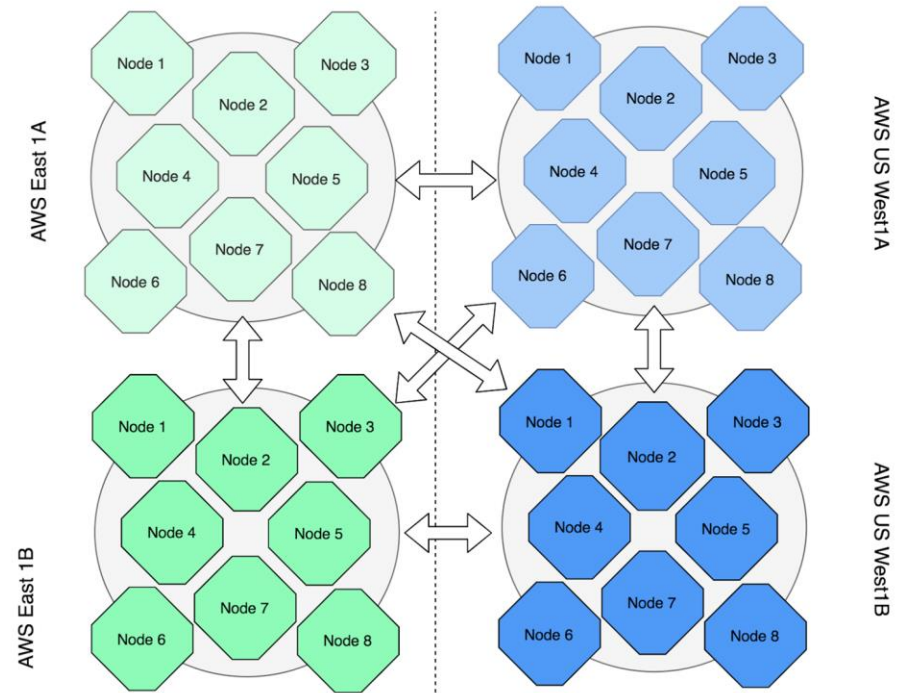
Infrastructure based
Resiliency

IMDG Cluster per Region










Infrastructure based Resiliency

IMDG Cluster per AZ



Infrastructure based Resiliency

	Single Cluster - xRegion	Multi-Cluster - Per AZ	Multi-Cluster - Per Region
Latency			
Replication Overhead	N/A		
Data Consistency			

Resiliency Scope

Infrastructure based Resiliency

- ❑ How the Data is spread across the Systems/ Machines & Data Centers...
- ✓ Isolate your data from machine(s) or Datacenter outage.

Resilient Connectivity

- ❑ How clients connect to the Hazelcast Clusters...
- ✓ Expect connectivity failures & Re-Connect automatically.

Resilient Client Connectivity

- TCP
- Multicast
- AWS Cloud Discovery
- GCP Cloud Discovery
- Apache jclouds® Cloud Discovery
- Azure Cloud Discovery
- Zookeeper Cloud Discovery
- Consul Cloud Discovery
- etcd Cloud Discovery
- Hazelcast for PCF
- Hazelcast OpenShift Integration
- Eureka Cloud Discovery
- Heroku Cloud Discovery
- Kubernetes Cloud Discovery

EC2 based Discovery

```
<hazelcast>
  <network>
    <join>
      <multicast enabled="false"/>
      <aws enabled="true">
        <access-key>my-access-key</access-key>
        <secret-key>my-secret-key</secret-key>
        <region>us-west-1</region>
        <security-group-name>hazelcast</security-group-name>
        <tag-key>aws-test-cluster</tag-key>
        <tag-value>cluster1</tag-value>
        <hz-port>5701-5708</hz-port>
      </aws>
    </join>
  </network>
</hazelcast>
```

<https://github.com/hazelcast/hazelcast-aws/blob/master/README.md>

Resiliency Scope

Infrastructure based Resiliency

- ❑ How the Data is spread across the Systems/ Machines & Data Centers...
- ✓ Isolate your data from machine(s) or Datacenter outage.

Resilient Connectivity

- ❑ How clients connect to the Hazelcast Clusters...
- ✓ Expect connectivity failures & Re-Connect automatically.

Data Resilience

- ❑ How do we keep the Data within cluster resilient?
- ✓ Build controls for failure & configure cluster to manage failure.

Single/Multi Node Failure Scenario

Data Resilience - backup-count

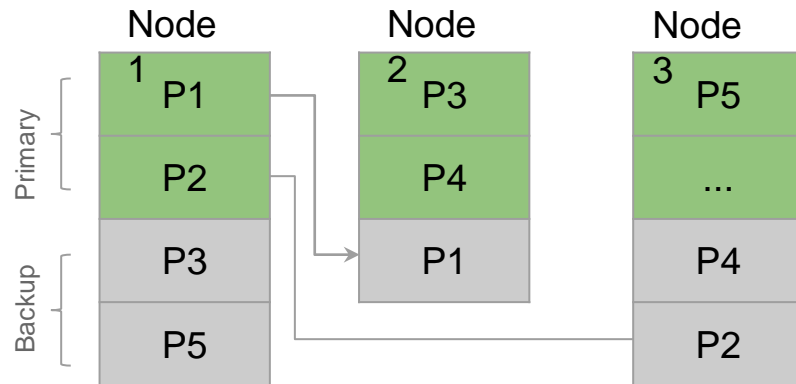
Data is recovered using the backups from cluster.

Two types of backups as described below:
sync (backup operations block operations until backups are successfully copied to backup members) and *async*.

Default:

- Backup operations are synchronous
- Distributed maps have one backup

*Backups increase memory usage since they are also kept in memory.



Backup Configuration

```
<hazelcast>
...
<map name="default">
  <backup-count>0</backup-count>
  <async-backup-count>1</async-backup-count>
</map>
...
</hazelcast>
```

<https://docs.hazelcast.org/docs/3.12.1/manual/html-single/index.html#creating-a-member-for-map-backup>

Network Split Scenario

Data Resilience - Split-Brain Protection

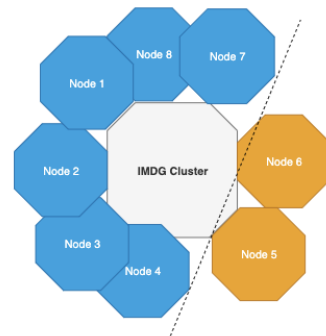
“Network partitioning is a network failure that causes the members to split into multiple groups such that a member in a group cannot communicate with members in other groups.”

Refer: [link](#)

Split-Brain Protection - Set the minimum number of members for the cluster to continue operating

Example : Cluster with size of 7 & a split-brain happens,

- Only the sub-cluster of four members is allowed to be used.
- the sub-clusters of sizes 1, 2 and 3 are prevented from being used.



Network partition

Split brain protection for map

```
<hazelcast>
...
<quorum name="quorumRuleWithFourMembers" enabled="true">
  <quorum-size>4</quorum-size>
</quorum>
<map name="default">
  <quorum-ref>quorumRuleWithFourMembers</quorum-ref>
</map>
...
</hazelcast>
```

<https://docs.hazelcast.org/docs/3.12.1/manual/html-single/index.html#configuring-split-brain-protection>

Resiliency Scope

Infrastructure based Resiliency

- ❑ How the Data is spread across the Systems/ Machines & Data Centers...
- ✓ Isolate your data from machine(s) or Datacenter outage.

Resilient Connectivity

- ❑ How clients connect to the Hazelcast Clusters...
- ✓ Expect connectivity failures & Re-Connect automatically.

Data Resilience

- ❑ How do we keep the Data within cluster resilient?
- ✓ Build controls for failure & configure cluster to manage failure.

Operational Resilience

- ❑ How do we recover from failures?
- ✓ Build operational processes for backup & recovery.

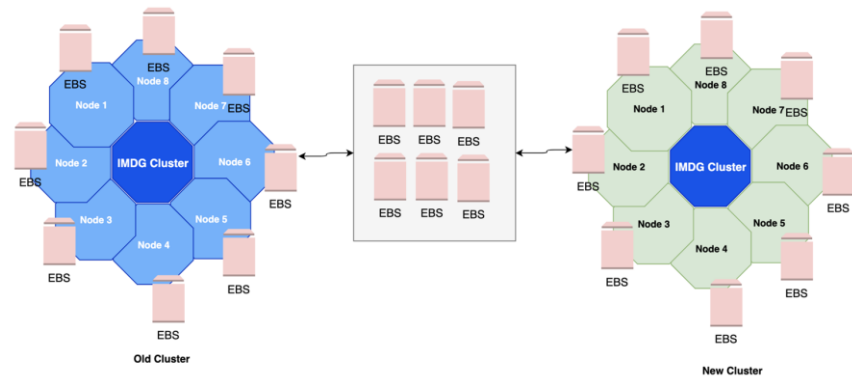
$$P_M = -K_M \left(e^{-\frac{I \cdot C \cdot U + F}{F_M}} - 1 \right)$$

Operational Resilience - Recovery via Hot Restart

Provides fast cluster restarts by storing the states of the cluster members on the disk.

1. Shutdown Hazelcast Cluster
2. Attach EBS Volume to New Hazelcast Cluster
3. Start New Hazelcast Cluster

fsync: Data is persisted to the disk during update before operation returns successful response



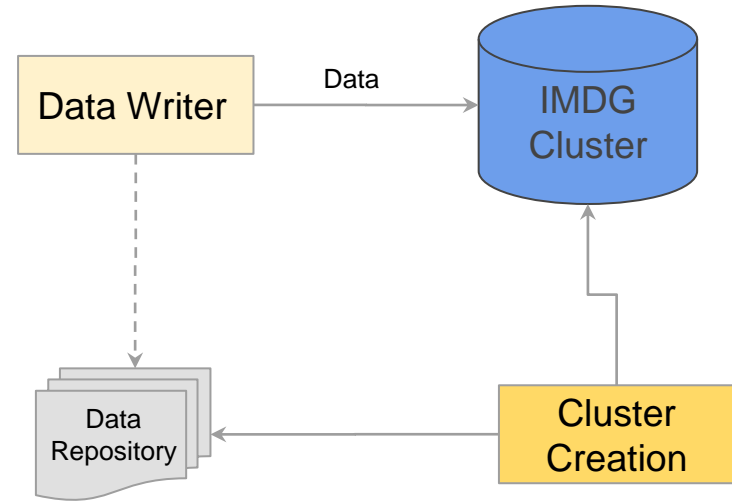
Hot-Restart Configuration

```
<hot-restart-persistence enabled="true">
  <base-dir>/mnt/hot-restart</base-dir>
  <parallelism>1</parallelism>
  <validation-timeout-seconds>120</validation-timeout-seconds>
  <data-load-timeout-seconds>900</data-load-timeout-seconds>
  <cluster-data-recovery-policy>FULL_RECOVERY_ONLY</cluster-data-recovery-policy>
  <auto-remove-stale-data>true</auto-remove-stale-data>
</hot-restart-persistence>
...
<map name="test-map">
  <hot-restart enabled="true">
    <fsync>false</fsync>
  </hot-restart>
</map>
```

<https://docs.hazelcast.org/docs/3.12.1/manual/html-single/index.html#creating-a-member-for-map-backup>

Operational Resilience - Recovery via Seed Data from System of Record

- Keep Master Data available
- Load data after new cluster is formed



Operational Resilience - Recovery via WAN Replication/Sync

Hazelcast WAN Replication allows to replicate data over WAN environments e.g. between different regions.

Define connectivity either via “Static endpoints” or “Discovery SPI”

Use case-

- Replicate data during BAU Operations
- Sync entire map to seed data

Add WAN Replication Configuration

Config Name:	<input type="text" value="the-wan-cluster"/>	Class Name:	<input type="text" value="com.hazelcast.enterprise.wan.replication.WanBatch"/>
Target Group Name:	<input type="text" value="Cluster-2"/>	Group Password:	<input type="password" value="*****"/>
Queue Capacity:	<input type="text" value="10000"/>	Endpoints:	<input type="text" value="127.0.0.1:5715"/>
Batch Max Delay(ms):	<input type="text" value="2000"/>	Batch Size:	<input type="text" value="500"/>
Response Timeout(ms):	<input type="text" value="60000"/>	Acknowledge Type:	<input type="text" value="ACK_ON_RECEIPT"/>
Full Queue Behavior:	<input type="text" value="DISCARD_AFTER_MUTATION"/>		

[+ Add Configuration](#)

Start WAN Sync

Select WAN Configuration

Select Target

Select Map

Sync

<https://docs.hazelcast.org/docs/3.12.1/manual/html-single/index.html#wan>

Resiliency Scope

Infrastructure based Resiliency

- ❑ How the Data is spread across the Systems/ Machines & Data Centers...
- ✓ Isolate your data from machine(s) or Datacenter outage.

Resilient Connectivity

- ❑ How clients connect to the Hazelcast Clusters...
- ✓ Expect connectivity failures & Re-Connect automatically.

Data Resilience

- ❑ How do we keep the Data within cluster resilient?
- ✓ Build controls for failure & configure cluster to manage failure.

Operational Resilience

- ❑ How do we recover from failures?
- ✓ Build operational processes for backup & recovery.

Monitoring

- ❑ How do we measure resiliency ?
- ✓ Leverage tools & build dashboard & Alerts.

ManCenter Console

Features:

- Monitor and Manage cluster
- Analyze Data Distribution
- Browse your data structures
- View Configuration
- Take thread dumps from members
- etc

GC Major
Count

1

GC Major
Time(ms)

23

GC Minor
Count

6

GC Minor
Time(ms)

27

CPU Utilization

Node

1min

5min

15min

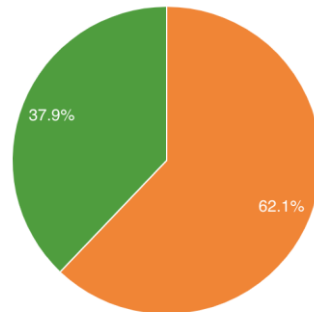
Utilization (%)

10.113.219.7:5701

0.42

0.45

0.46



map other free

JMX Beans

Parameters to be added to Mencoder -

-D Hazelcast.mc.jmx.enabled=true -

D Hazelcast.mc.jmx.port=9000 -

D com.sun.management.jmxremote.ssl=false

Top Latency items:

- MaxGetLatency
- MaxPutLatency
- MaxRemoveLatency

Average Latency items:

- AvgGetLatency
- AvgPutLatency
- AvgRemoveLatency

WanConfigs

- getOutboundQueueSize(<Publisher ID>)

pid: 16317 Launcher

Overview Memory Threads Classes VM Summary **MBeans**

Attribute value

Name	Value
HeapUsedMemory	133529424

Refresh

MBeanAttributeInfo

Name	Value
Attribute:	
Name	HeapUsedMemory
Description	HeapUsedMemory
Readable	true
Writable	false
Is	false
Type	long

Descriptor

Name	Value
Attribute:	
openType	javax.management.openmbe...
originalType	long

Tree View:

- JMImplementation
 - ManagementCenter
 - ManagementCenter[dev1]
 - Members
 - "10.113.219.7:5701"
 - Attributes
 - OwnedPartitionCount
 - ConnectedClientCount
 - Master
 - HeapUsedMemory**
 - HeapFreeMemory
 - HeapMaxMemory
 - HeapTotalMemory
 - NativeMaxMemory
 - NativeCommittedMemory
 - NativeUsedMemory
 - NativeFreeMemory
 - Services
 - WanConfigs
 - dev1
 - com.sun.management
 - java.lang
 - java.nio
 - java.util.logging

Appendix