



MAY 16-18, 2017

MIAMI, FL

Tomcat Cluster

Keiichi Fujino

Agenda

- About me
- Tomcat Clustering Overview
- Session Replication
- Cluster Channel Component
- Monitoring Cluster components

About me

- Keiichi Fujino
- I live in Japan
- Software Engineer since 2002
- Apache Tomcat committer since 2010
- kfujino@apache.org

Clustering Overview

What is Cluster?

- Performance improvement
- High availability

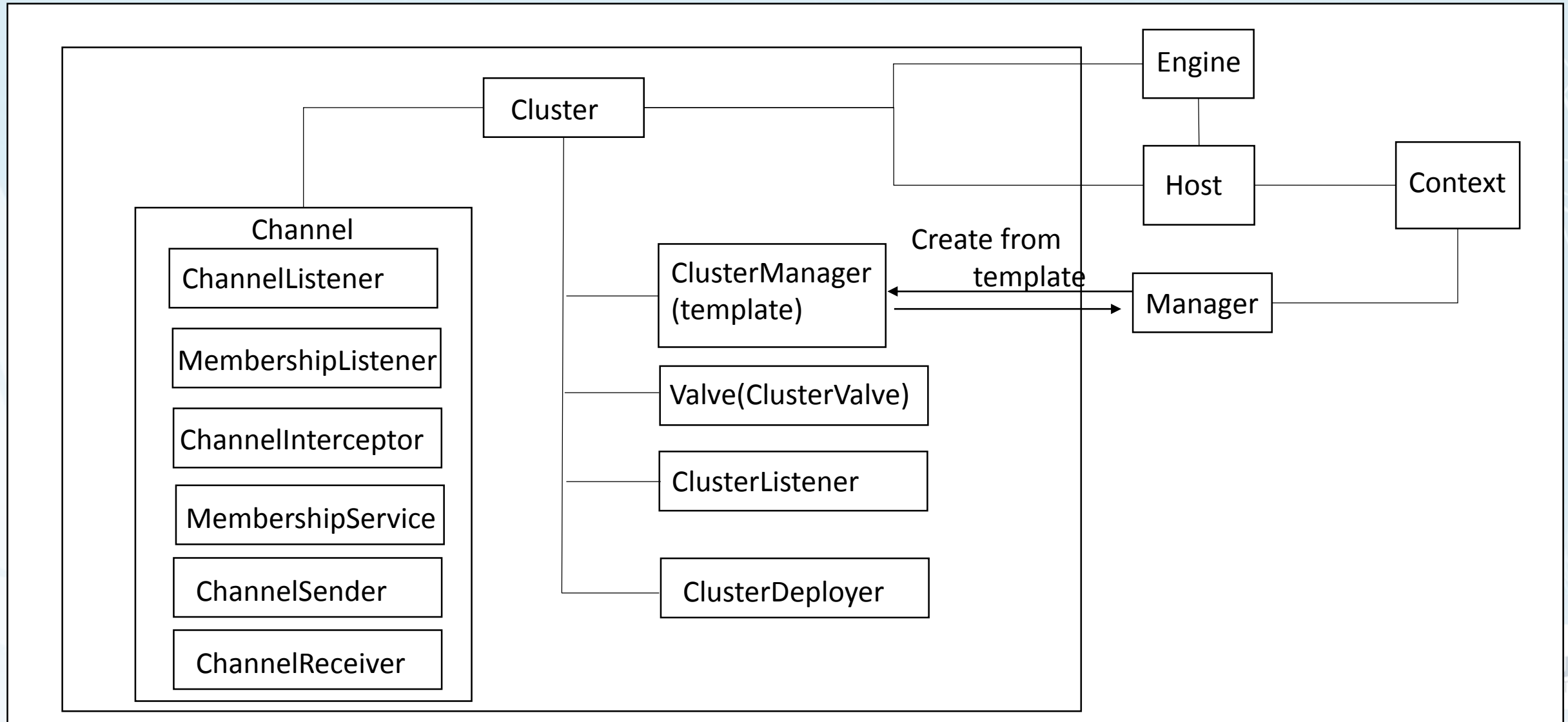
Tomcat Clustering

- Cluster membership/Grouping
- Session Replication

Load balancing is not a Tomcat features

- Use `mod_jk` / `mod_proxy_balancer`

Cluster Architecture



- Cluster
 - The main component of Tomcat Cluster
- Cluster Manager
 - The session manager for the session replication
- Valve(Cluster Valve)
 - The same as usual Valve
 - Added to the request processing pipeline automatically

- Cluster Deployer
 - Sharing of WAR files among cluster nodes
- ClusterListener
 - Listen cluster messages and events
- Channel
 - Performs messaging and grouping among the cluster nodes

Session Replication



MAY 16-18, 2017
MIAMI, FL

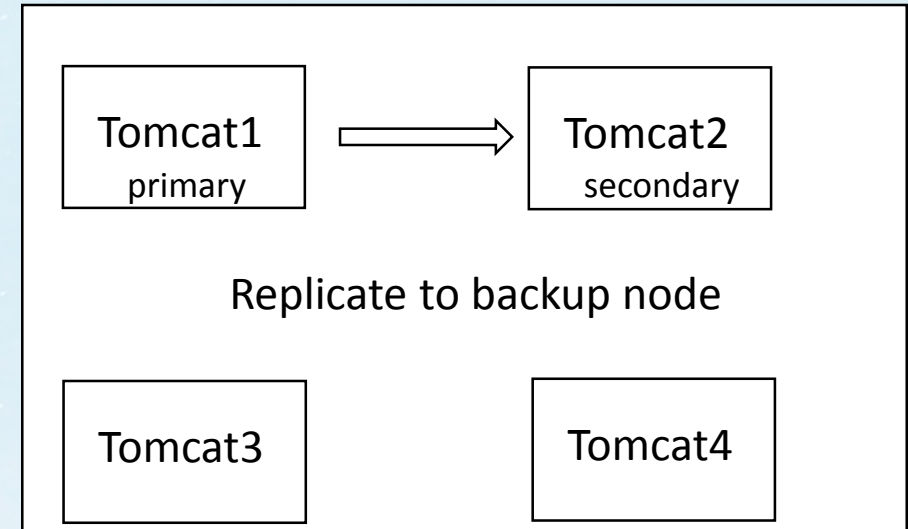
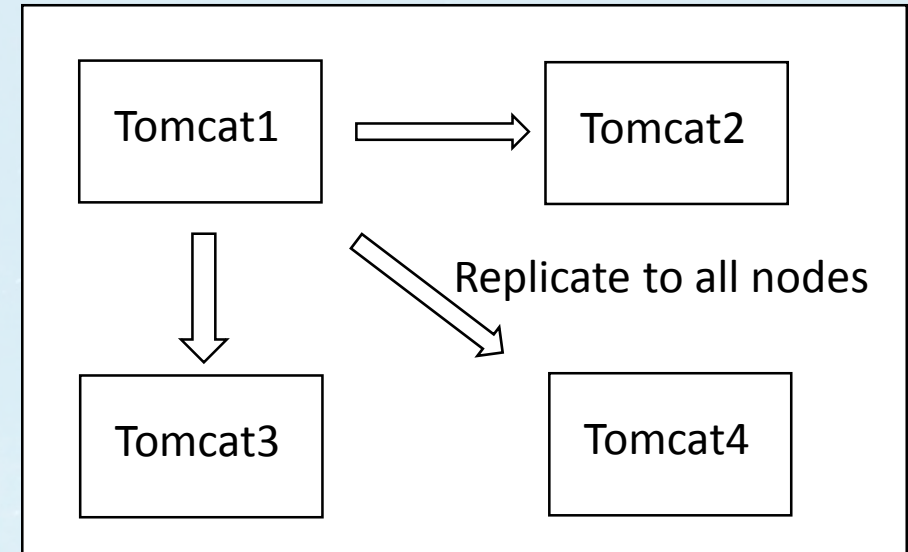
Session Replication

Session Replication

Implementations

- All-to-All session replication
 - DeltaManager (Default)

- Primary-Secondary session replication
 - BackupManager



Use constraints

- Make sure that your web.xml describe the `<distributable/>` element
- Session attributes must implement `java.io.Serializable`
- sticky session
 - If you use the BackupManager, This is required

How to configure

- Define `<Cluster>` element inside The Engine or Host element.
- Configure Cluster Manager
 - DeltaManager or BackupManager
- Configure Channel components
- Enable `org.apache.catalina.ha.tcp.ReplicationValve`
- Enable `org.apache.catalina.ha.session.ClusterSessionListener`
 - If DeltaManager use

Delta Replication

Delta Replication

- Replicate only the changes of session
 - Not all session data
- Replicate all changes of session at the time of end of request
 - Not replication per change of session

Delta Replication

Register Delta Info

- *Add ATTRIBUTE(Attr_A, Value_A)*
- *Add ATTRIBUTE(Attr_B, Value_B)*
- *Remove ATTRIBUTE(Attr_A)*
- *Add ATTRIBUTE(Attr_B, Value_BB)*

Default : recordAllActions=false

recordAllActions=true

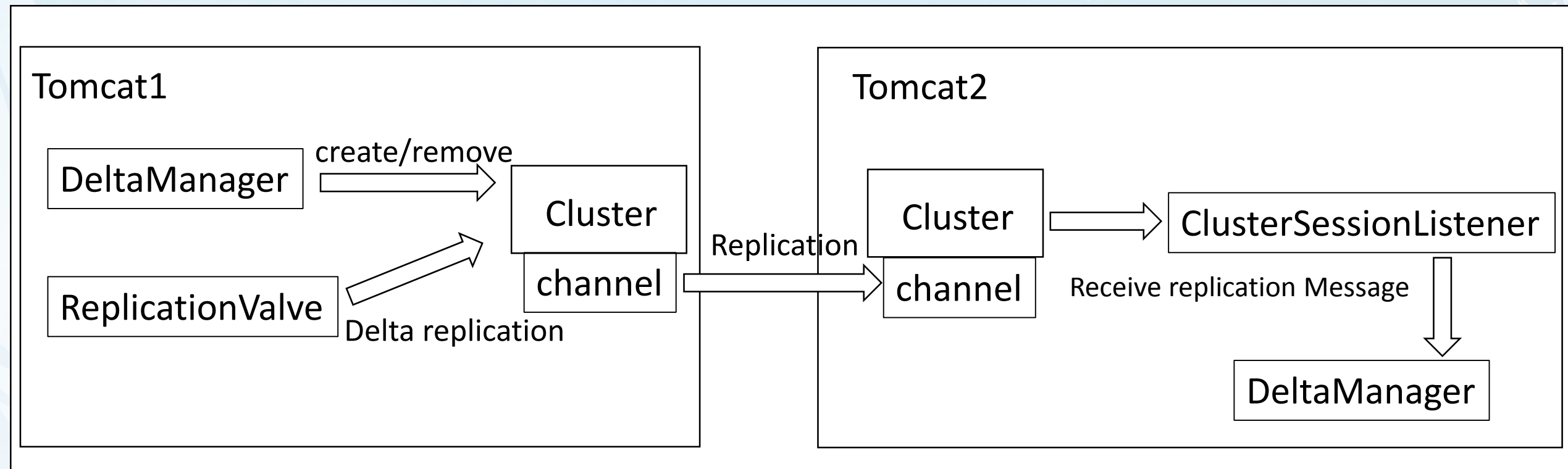
TYPE	ACTION	NAME	VALUE
ATTRIBUTE	SET	Attr_A	Value_A
ATTRIBUTE	SET	Attr_B	Value_B
ATTRIBUTE	REMOVE	Attr_A	null
ATTRIBUTE	SET	Attr_B	Value_BB

TYPE	ACTION	NAME	VALUE
ATTRIBUTE	SET	Attr_A	Value_A
ATTRIBUTE	SET	Attr_B	Value_B
ATTRIBUTE	REMOVE	Attr_A	null
ATTRIBUTE	SET	Attr_B	Value_BB

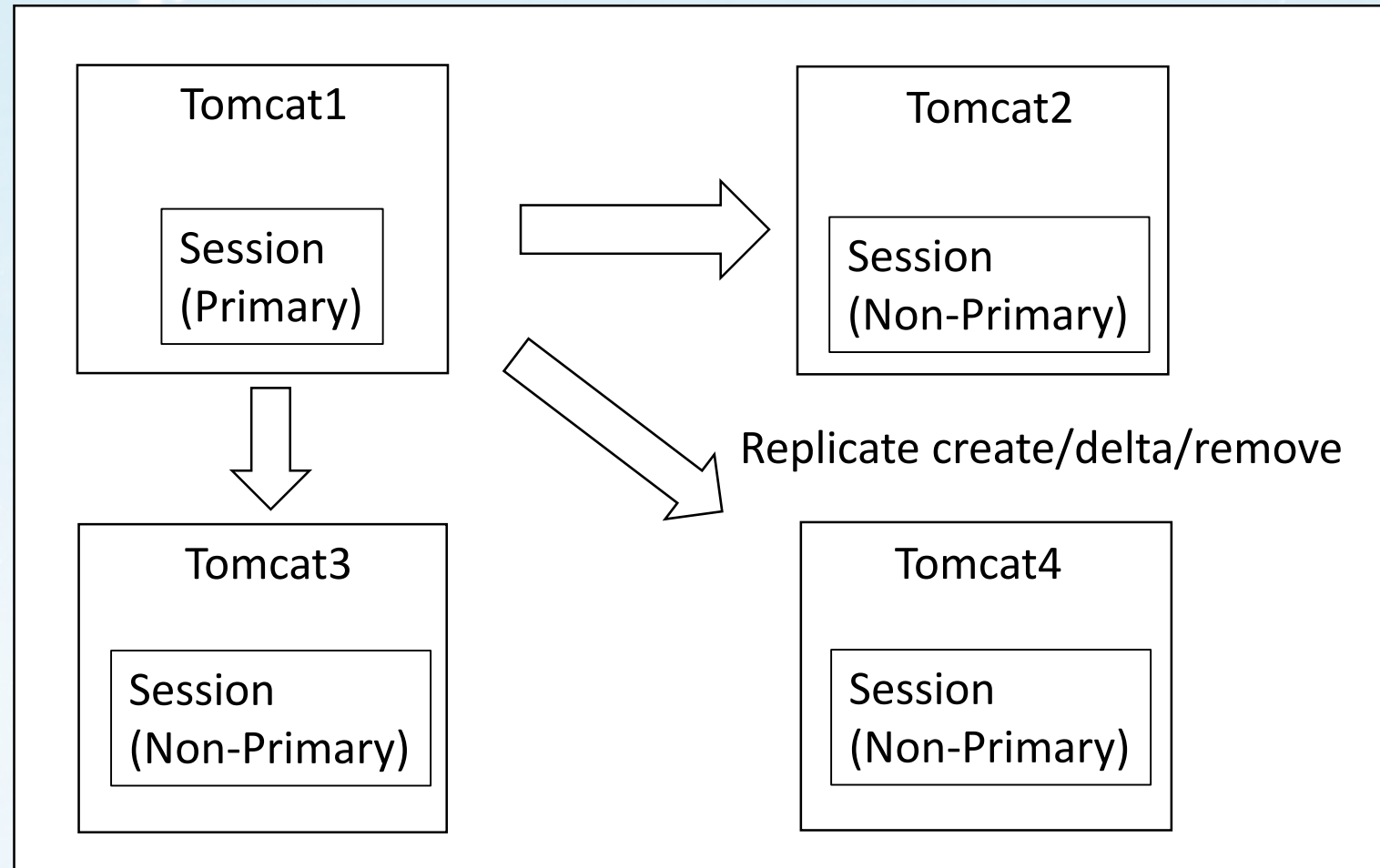
DeltaManager

- All-to-All session replication
- Default Session Manager in Cluster environment
- For small cluster

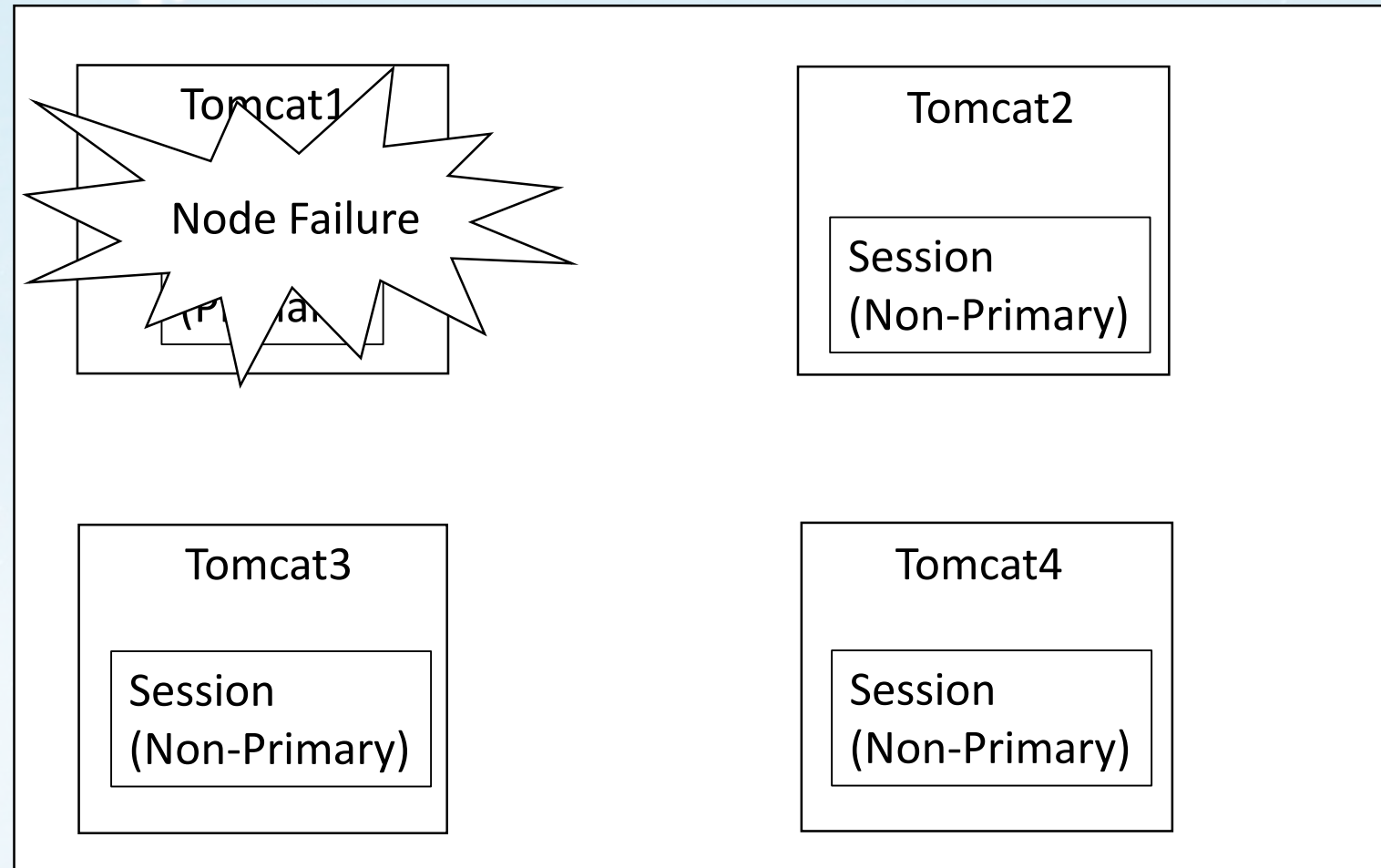
Architecture



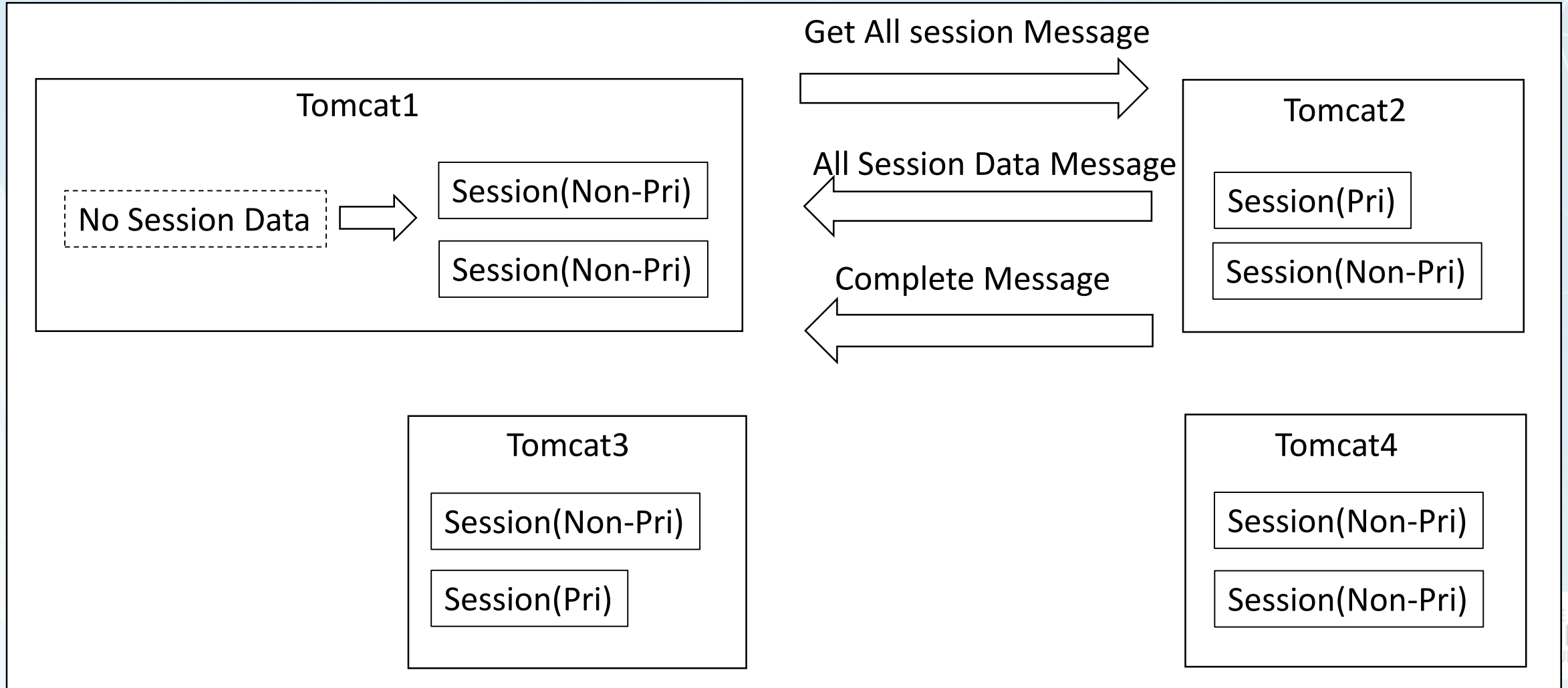
Behavior



Node Failure



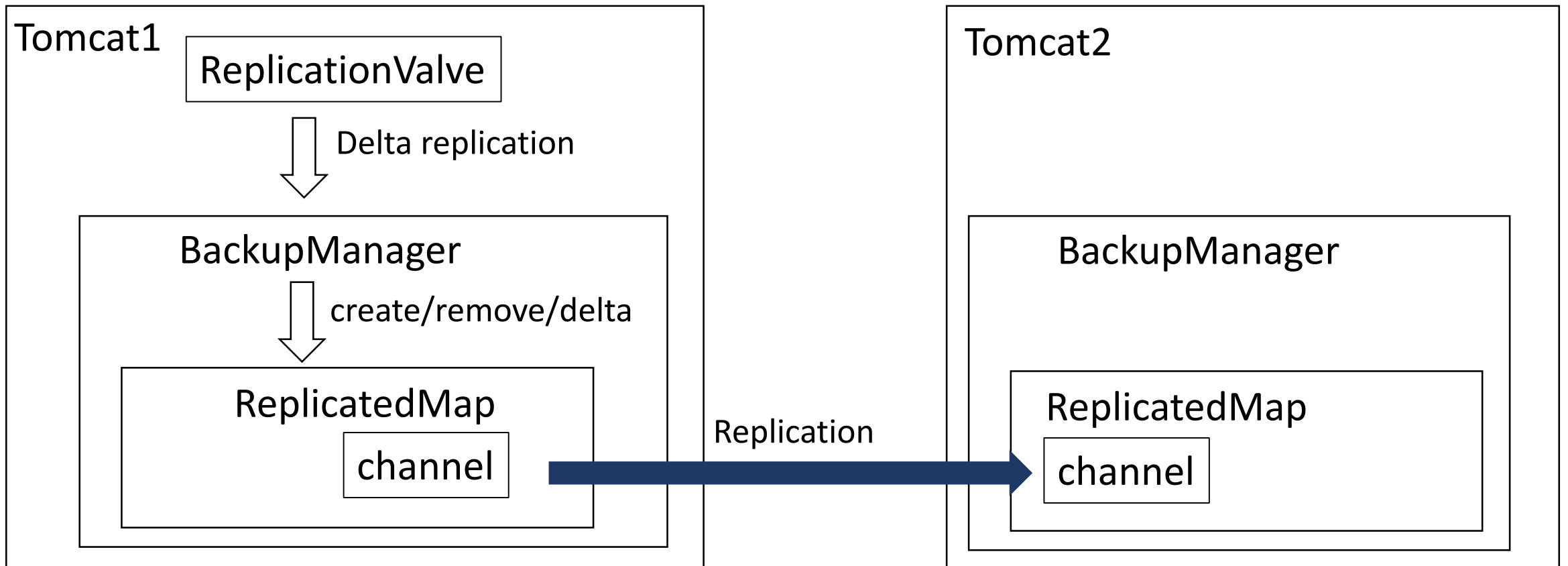
Node Recovery



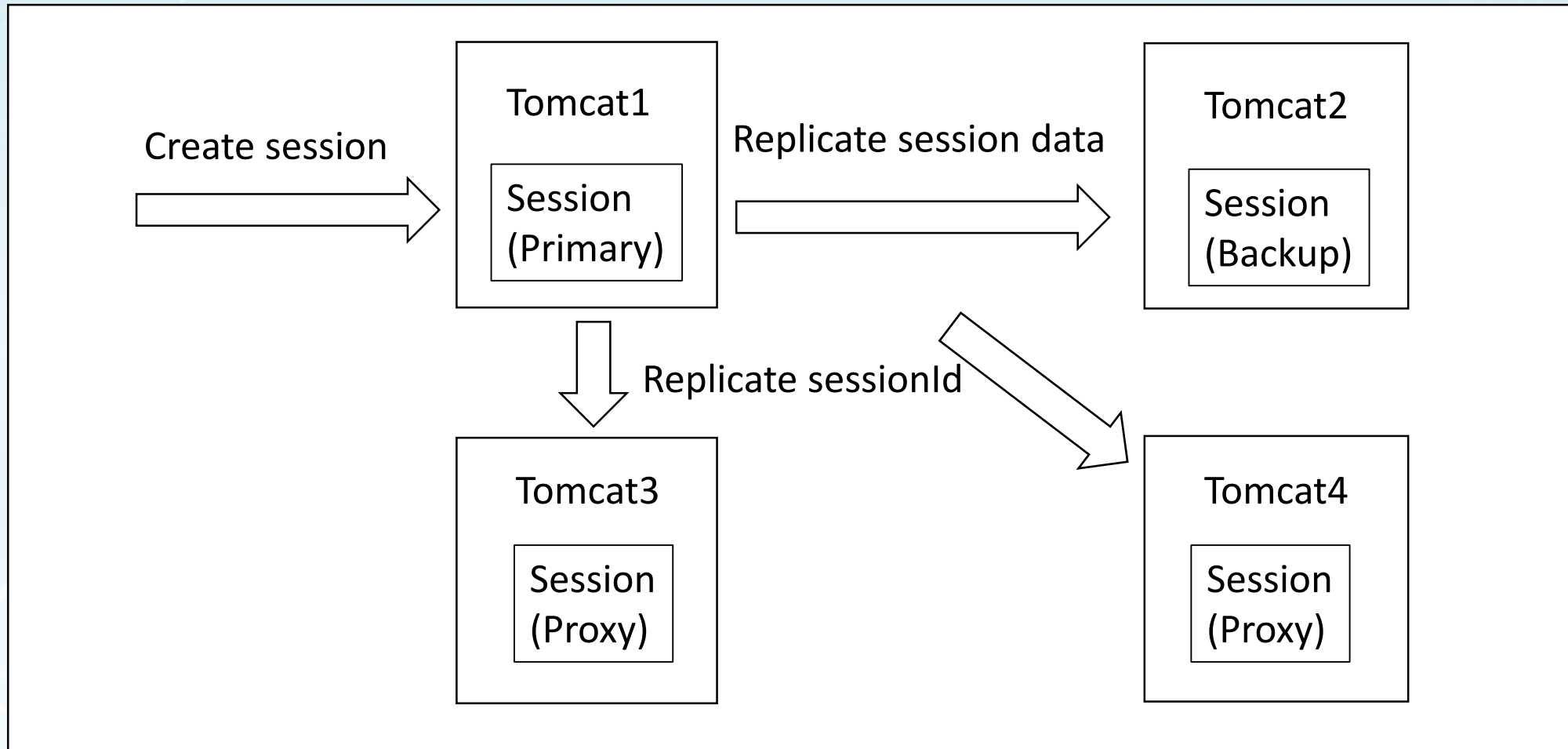
BackupManager

- Primary-Secondary session replication
- For large cluster

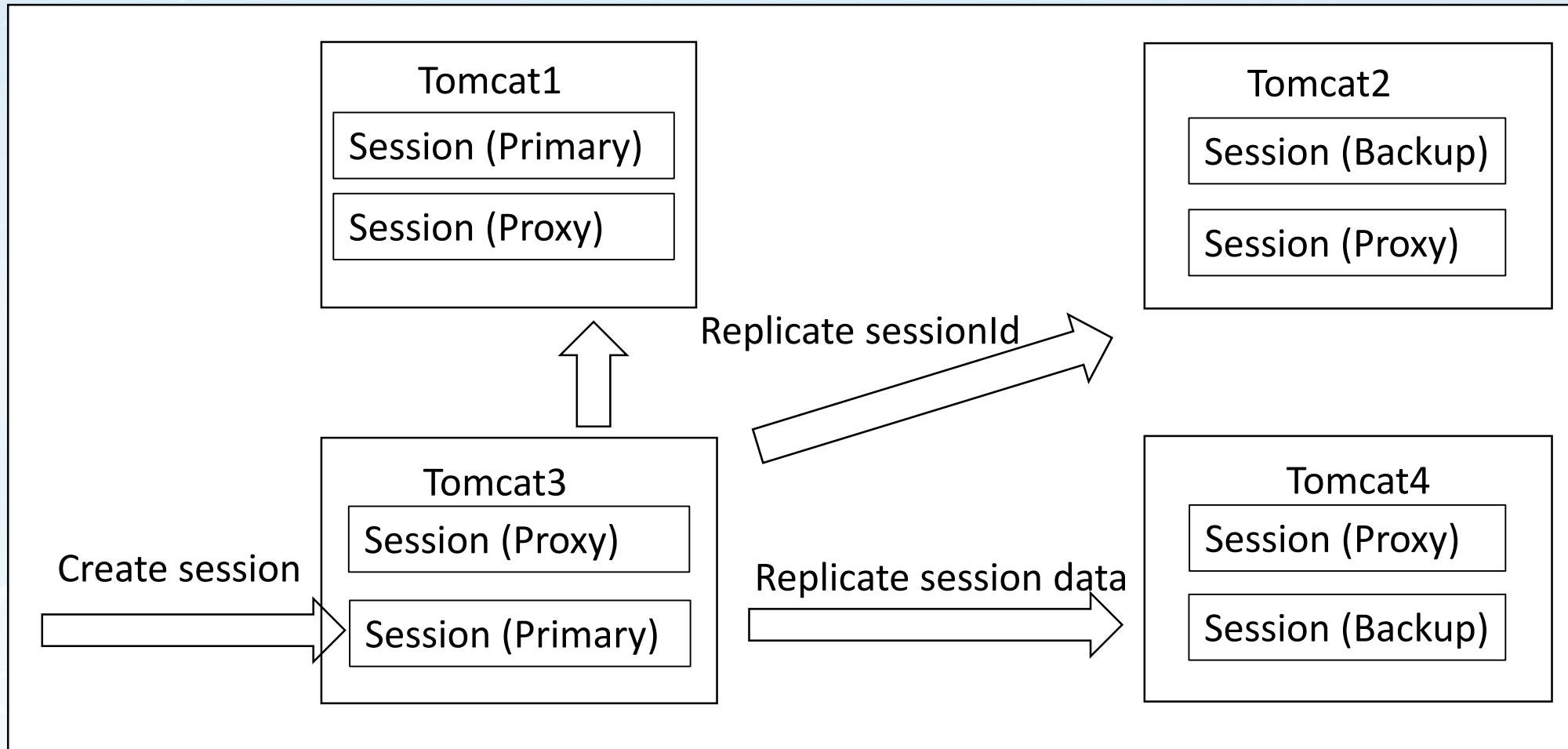
Architecture



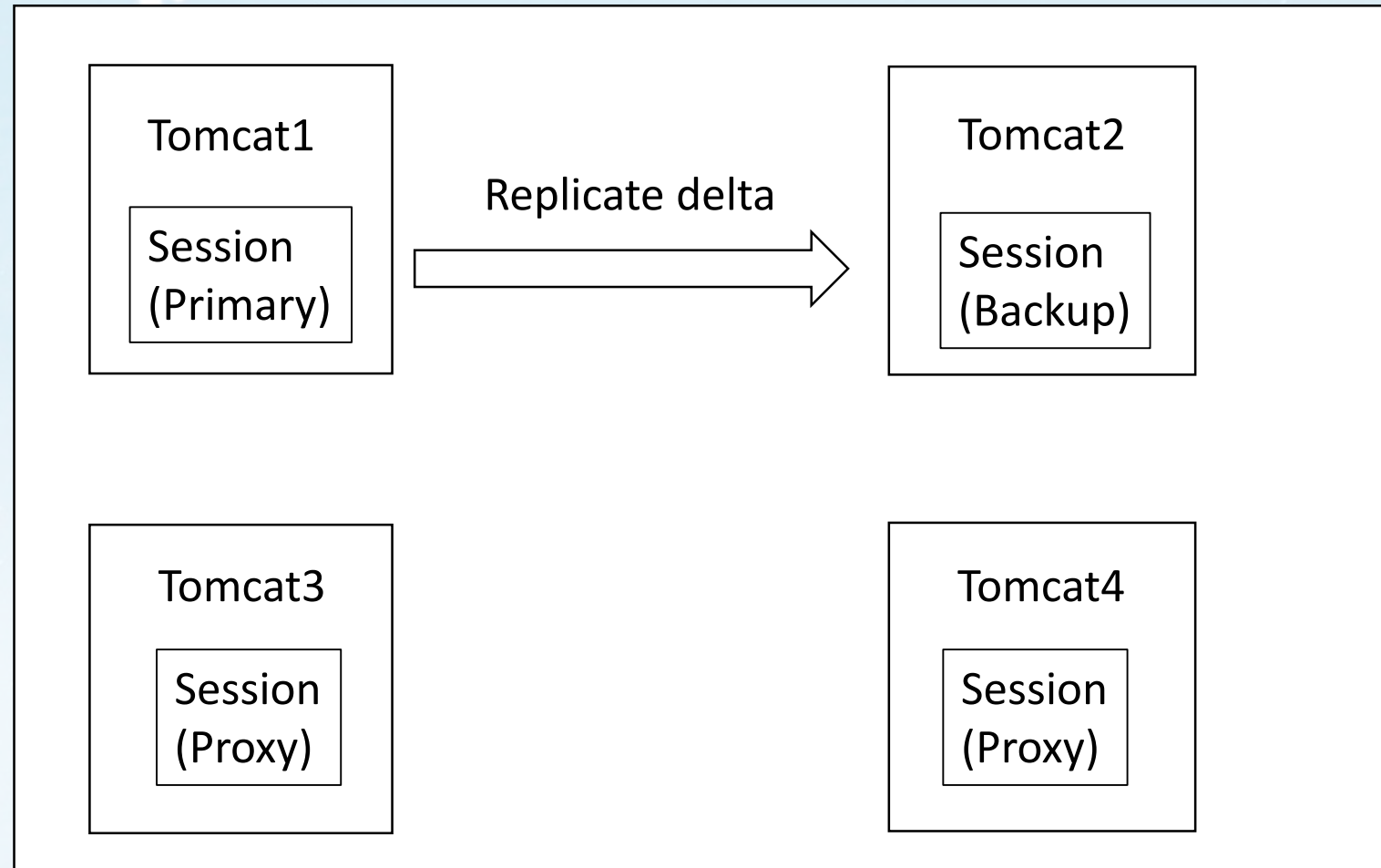
Behavior of Create



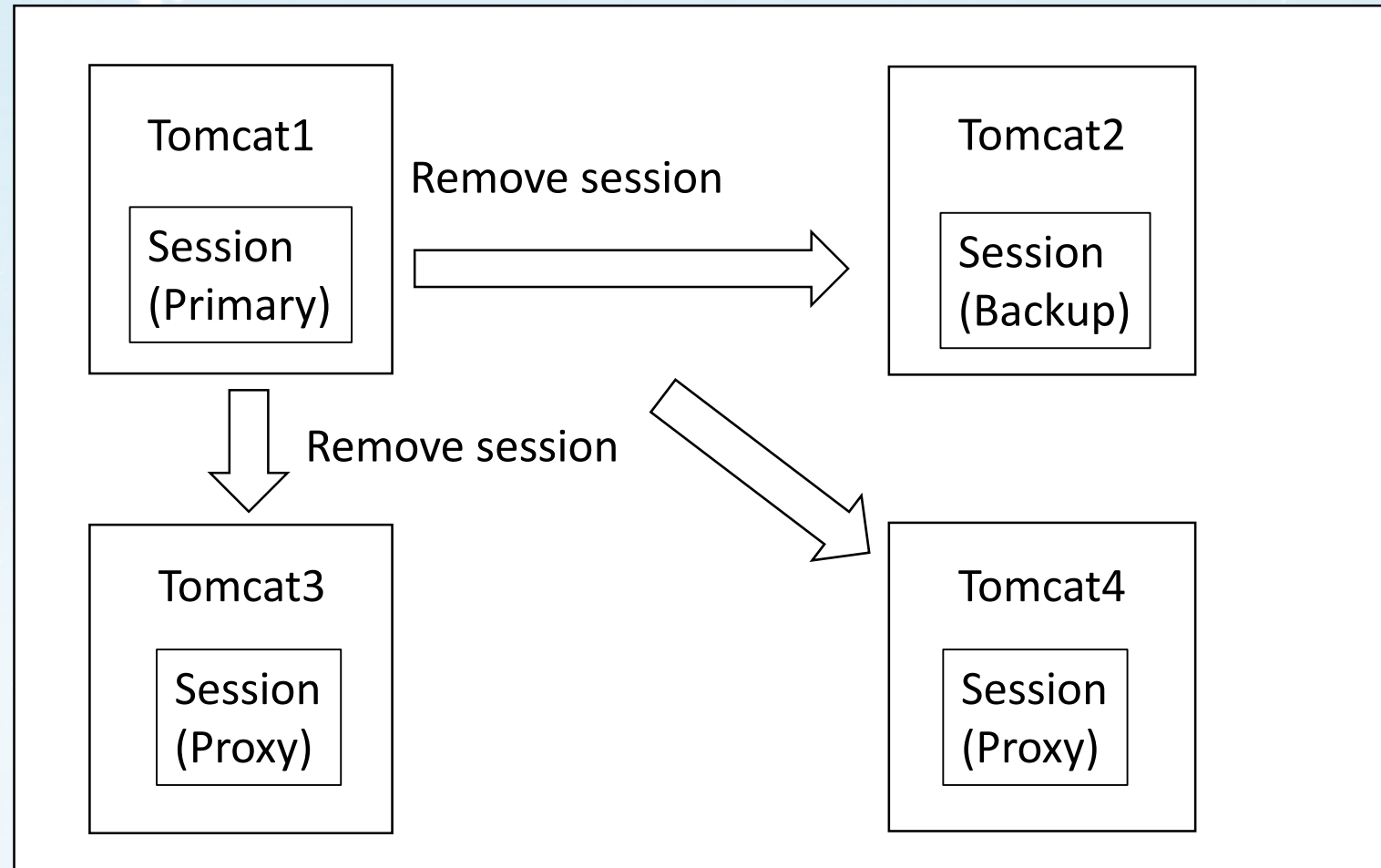
Behavior of Create



Behavior of Delta

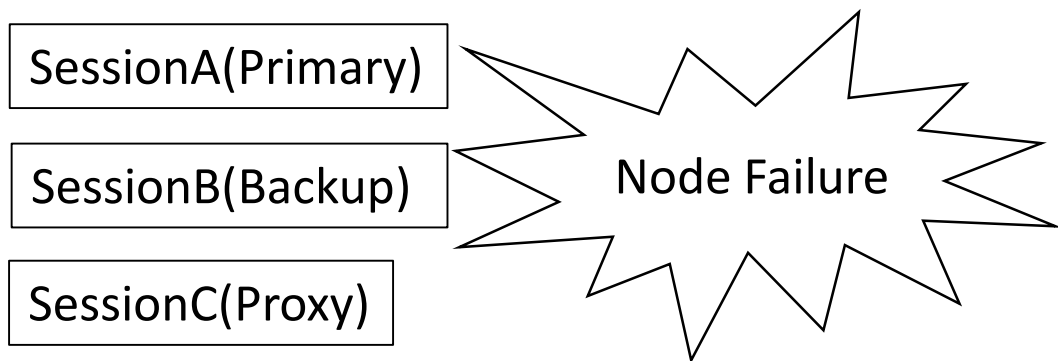


Behavior of Remove

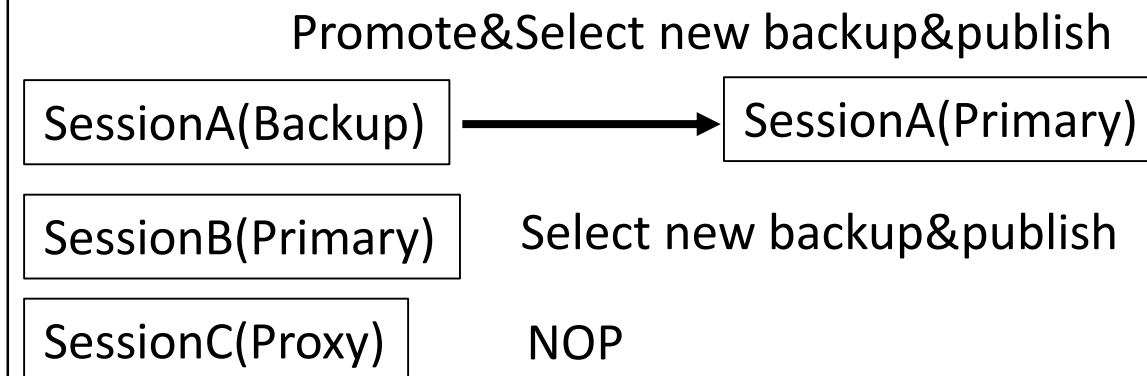


Node Failure

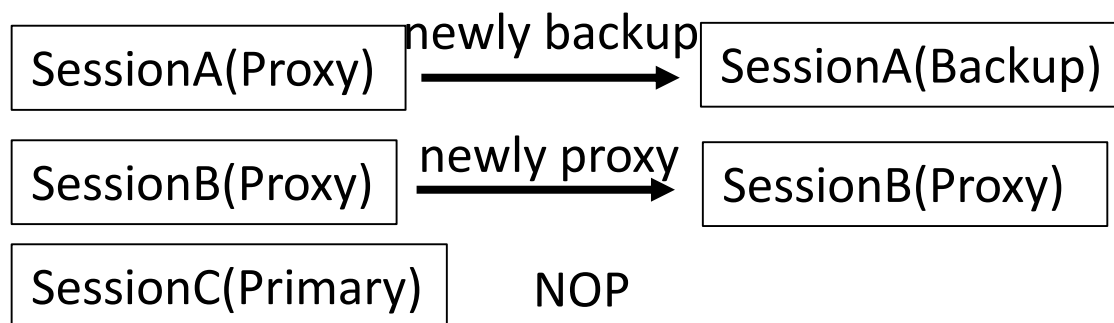
Tomcat1



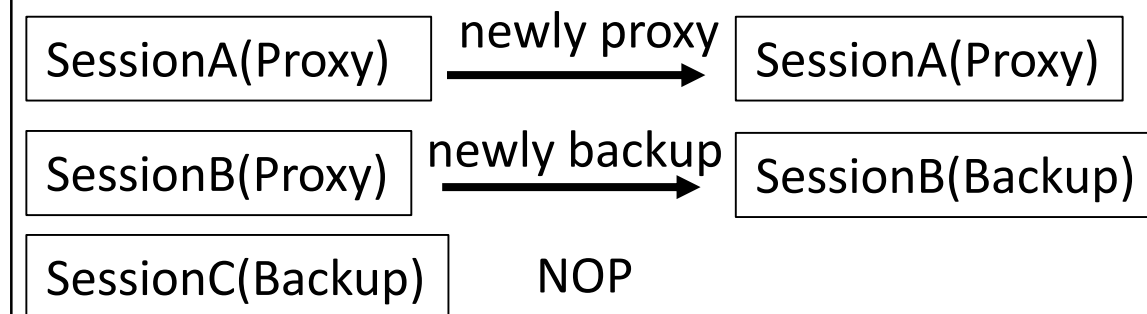
Tomcat2



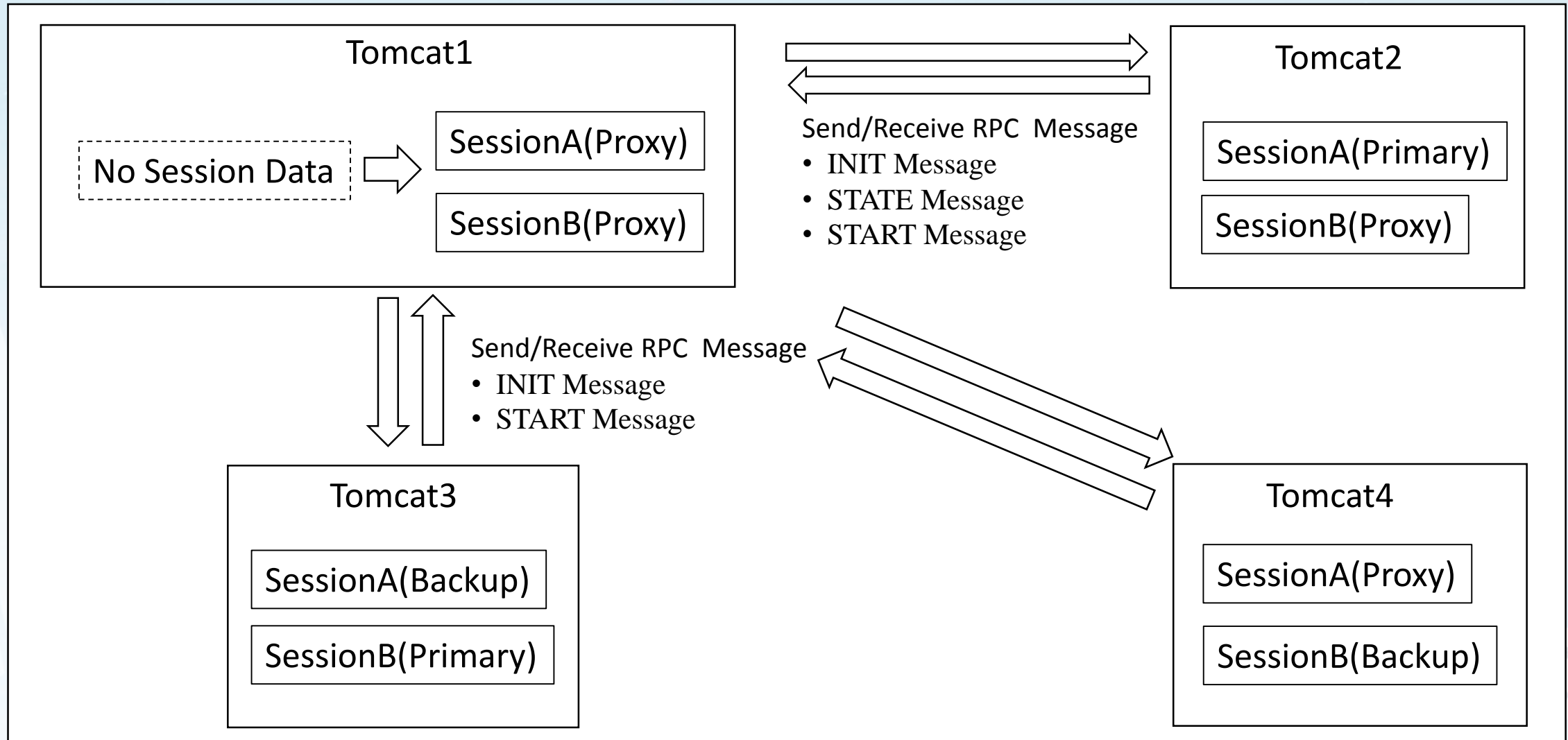
Tomcat3



Tomcat4



Node Recovery



Channel



MAY 16-18, 2017
MIAMI, FL

Cluster Channel

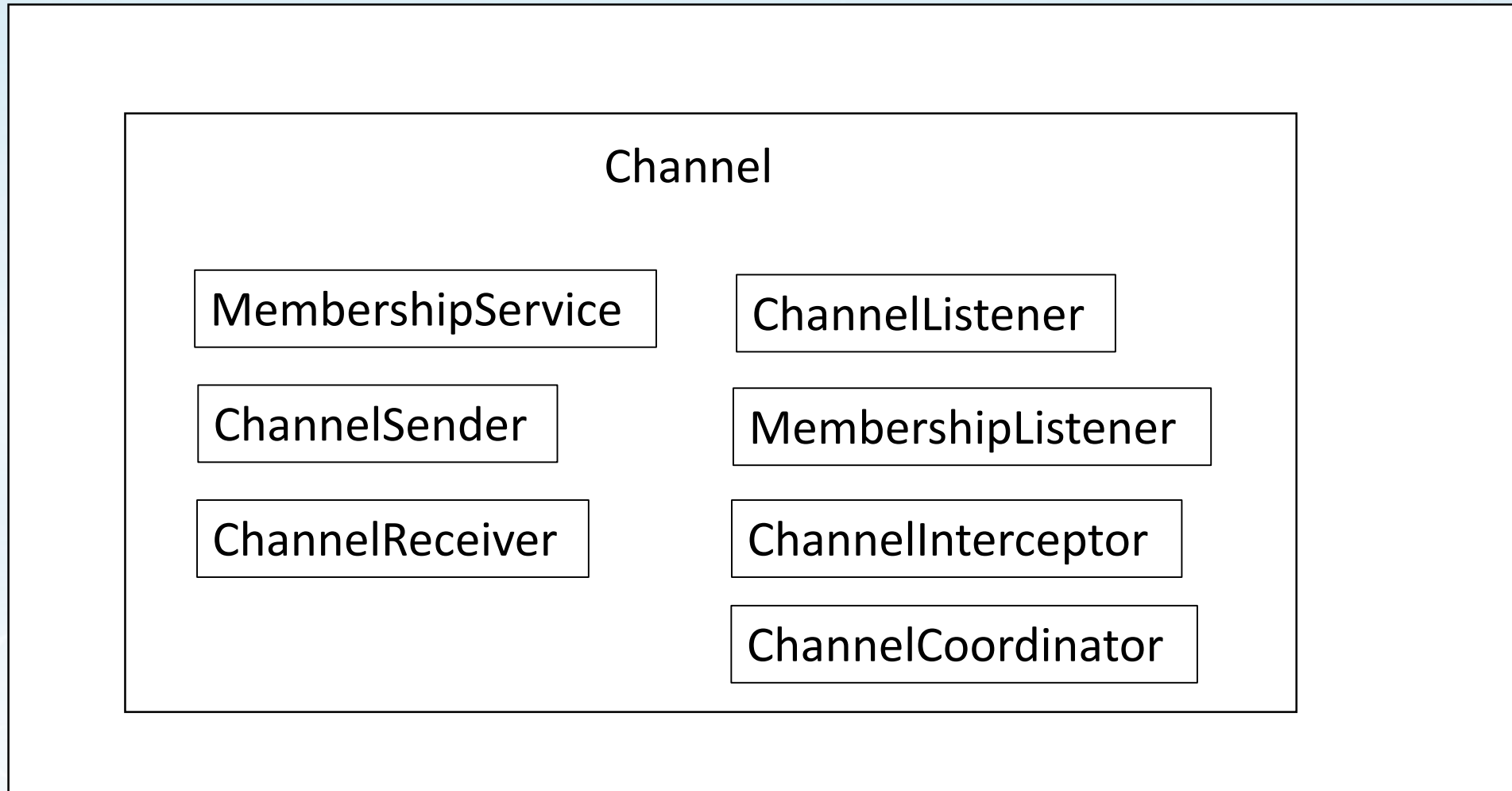
What is Channel ?

- Messaging & Grouping component

Responsibility

- Build Membership
- Send channel messages
- Receive channel messages

Channel Components



- Channel
 - The main component of channel
 - `org.apache.catalina.tribes.group.GroupChannel` only
- MembershipService
 - The component which build a cluster group
 - Start a multicast receiver thread and a multicast sender thread

- ChannelSender
 - Send channel messages to other nodes
 - Sender Queue size is specified in poolSize attribute
- ChannelReceiver
 - Receive channel messages from other nodes
 - Tuning of the maxThreads attribute depends on send option of channel message
 - Synchronization mode
 - Need to align maxThreads with poolSize
 - Asynchronous mode
 - do not need to align the maxThreads with poolSize.

Channel Components

- ChannelListener
 - Listen received channel messages
- MembershipListener
 - Listen add/remove of cluster members

- ChannelInterceptor
 - Intercept a channel message and a member detection
 - There are many configurable implementations
- ChannelCoordinator
 - Special ChannelInterceptor
 - Coordinates the ChannelInterceptors

Sample config

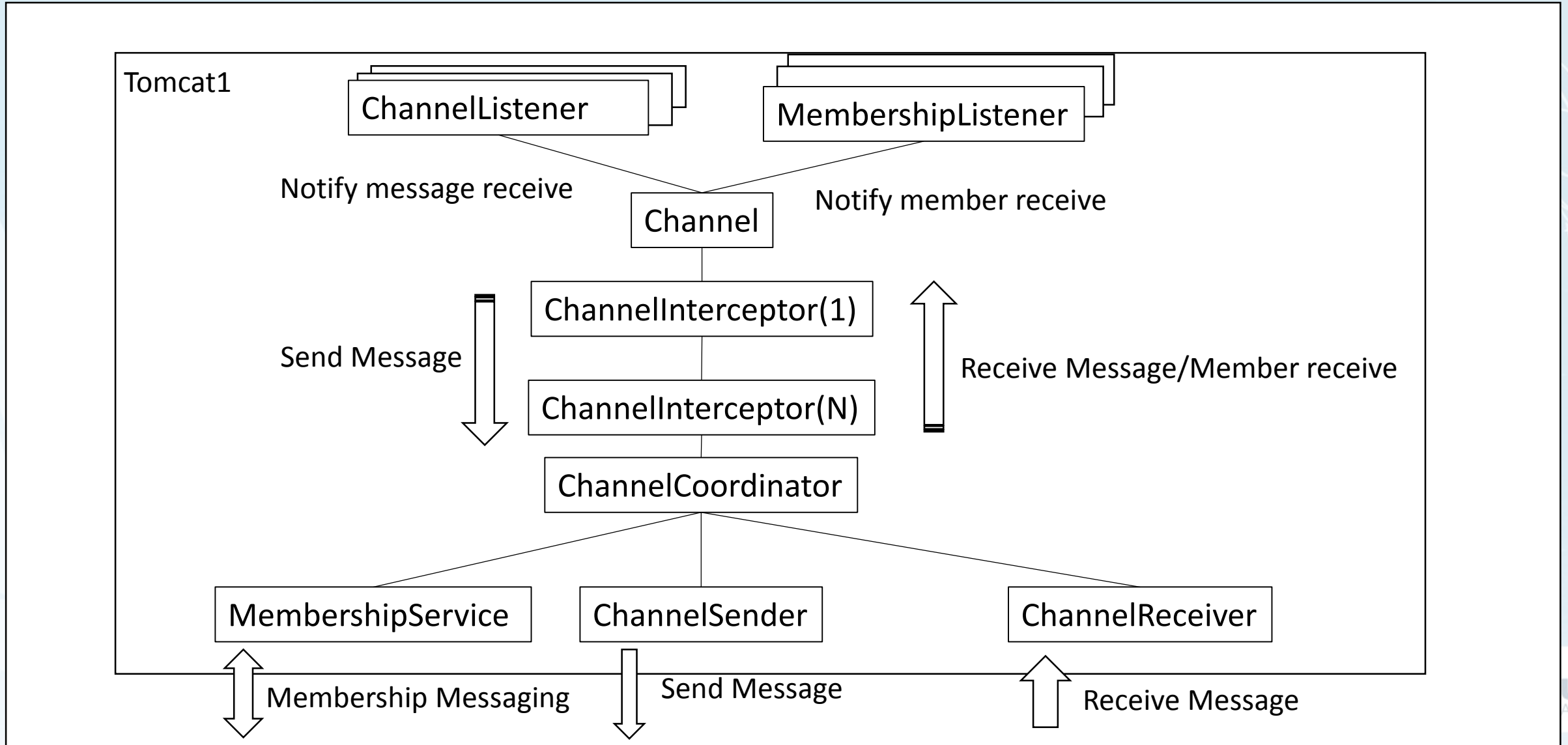
```
<Cluster className="org.apache.catalina.ha.tcp.SimpleTcpCluster">
... ..
<Channel className="org.apache.catalina.tribes.group.GroupChannel">
  <Membership className="org.apache.catalina.tribes.membership.McastService"
    address="229.0.0.61" port="45564" frequency="500" dropTime="4000" />

  <Receiver className="org.apache.catalina.tribes.transport.nio.NioReceiver"
    address="auto" port="4004" autoBind="100" maxThreads="25"/>

  <Sender className="org.apache.catalina.tribes.transport.ReplicationTransmitter">
    <Transport className="org.apache.catalina.tribes.transport.nio.PooledParallelSender"
      timeout="5000" poolSize="25"/>
  </Sender>

  <Interceptor className="org.apache.catalina.tribes.group.interceptors.MessageDispatch15Interceptor" />
  <Interceptor className="org.apache.catalina.tribes.group.interceptors.TcpFailureDetector"/>
</Channel>
... ..
</Cluster>
```

Channel Architecture



TCPFailureDetector

- Main Features
 - Intercept memberDisappeared events.
 - Member check when send errors.
- Other Features
 - Manage membership in static membership

MessageDispatchInterceptor

- Asynchronous send message
- Use Thread Pooling
 - maxThreads
 - maxSpareThreads
- You must set the send options asynchronous.
 - Cluster's channelSendOptions to asyn mode(8)
 - BackupManager's mapSendOptions to asyn mode(8)

StaticMembershipInterceptor

- The static membership instead of multicast
- Make sure
 - Disable multicast membership
 - Cluster's channelStartOptions attribute = 3
 - Enable TcpPingInterceptor for nodes failure detection
 - Enable TcpFailureDetector for membership management
 - The order is
 - TcpPingInterceptor
 - TcpFailureDetector
 - StaticMembershipInterceptor

```
<Interceptor className="org.apache.catalina.tribes.group.interceptors.TcpPingInterceptor"/>
<Interceptor className="org.apache.catalina.tribes.group.interceptors.TcpFailureDetector"/>
<Interceptor className=
    "org.apache.catalina.tribes.group.interceptors.StaticMembershipInterceptor">
  <LocalMember className="org.apache.catalina.tribes.membership.StaticMember"
    uniqueId="{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,1}"/>
  <Member className="org.apache.catalina.tribes.membership.StaticMember"
    port="4010" host="hostA"
    uniqueId="{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,2}" />
  <Member className="org.apache.catalina.tribes.membership.StaticMember"
    port="4010" host="hostB"
    uniqueId="{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,3}" />
</Interceptor>
```

ThroughputInterceptor

- Measuring the send and receive of channel messages

DomainFilterInterceptor

- Filter the members that join cluster group by the domain

Monitoring Cluster



MAY 16-18, 2017
MIAMI, FL

Monitoring Cluster

Monitoring Cluster

How to monitor Tomcat cluster.

- Logging
- JMX

Monitoring your Cluster with Logging

- To track channel messages for Debugging
 - Use the Tomcat JULI
 - Key: `org.apache.catalina.tribes.MESSAGES`
 - Level: `FINEST`

```
# FOR DEBUG
```

```
10catalina.org.apache.juli.AsyncFileHandler.formatter = org.apache.juli.VerbatimFormatter
```

```
10catalina.org.apache.juli.AsyncFileHandler.level = FINEST
```

```
10catalina.org.apache.juli.AsyncFileHandler.directory = ${catalina.base}/logs
```

```
10catalina.org.apache.juli.AsyncFileHandler.prefix = MESSAGES.
```

```
10catalina.org.apache.juli.AsyncFileHandler.bufferSize = -1
```

```
org.apache.catalina.tribes.MESSAGES.level = FINEST
```

```
org.apache.catalina.tribes.MESSAGES.handlers = 10catalina.org.apache.juli.AsyncFileHandler
```

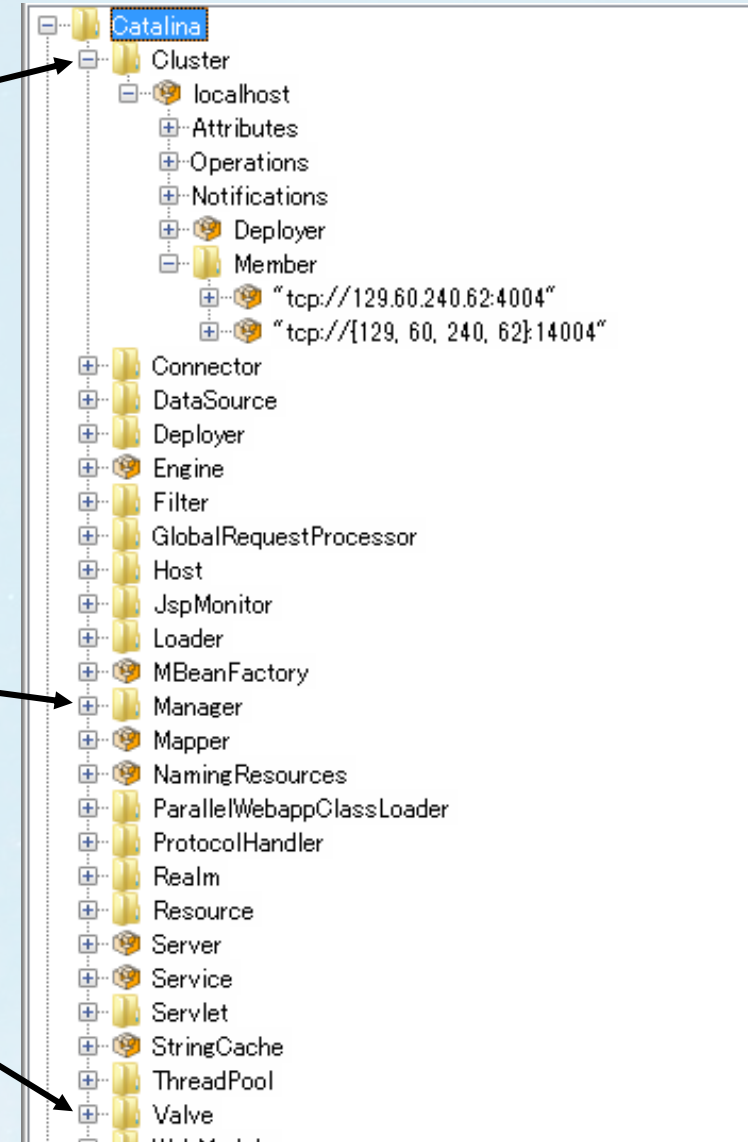
Monitoring your Cluster with Logging

- ThroughputInterceptor
 - Report the throughput statistics.
 - Default interval is to report every 10000 messages.
 - Key: org.apache.catalina.tribes.group.interceptors.ThroughputInterceptor
 - Level: INFO

```
org.apache.catalina.tribes.group.interceptors.ThroughputInterceptor.report ThroughputInterceptor Report[
  Tx Msg:60,024 messages
  Sent:53.30 MB (total)
  Sent:53.31 MB (application)
  Time:16.84 seconds
  Tx Speed:3.16 MB/sec (total)
  TxSpeed:3.16 MB/sec (application)
  Error Msg:0
  Rx Msg:60,048 messages
  Rx Speed:0.23 MB/sec (since 1st msg)
  Received:53.28 MB]
```

Monitoring your Cluster with JMX

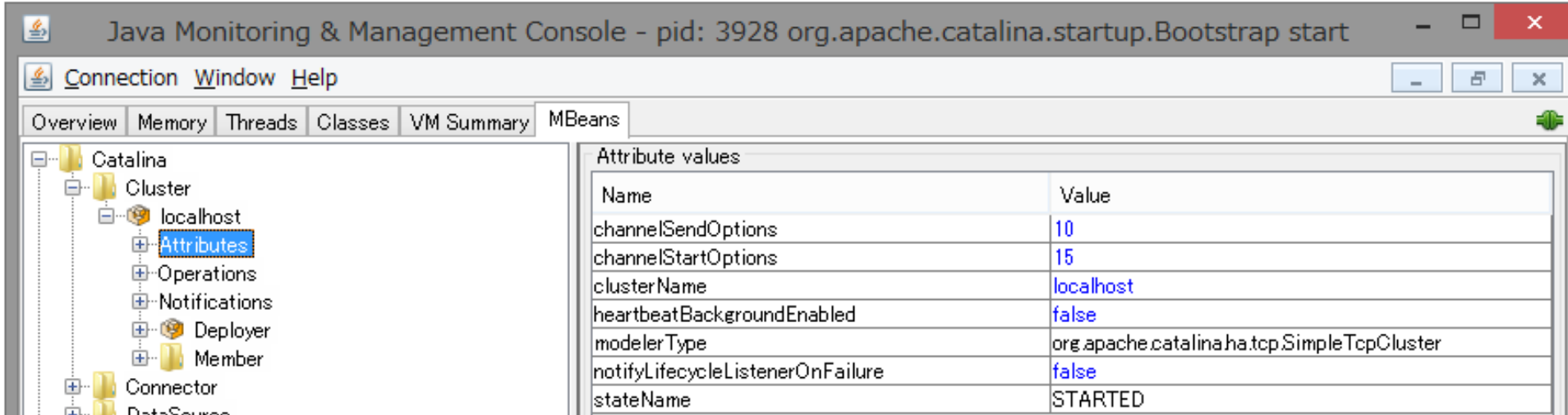
- Catalina Domain
 - Cluster Mbeans
 - Cluster Mbean
 - Deployer Mbean
 - Member MBeans
 - (Cluster)Manager Mbean
 - (Cluster)Valve MBean



Monitoring Cluster

Cluster Mbean

- Cluster settings



The screenshot shows the Java Monitoring & Management Console window. The title bar reads "Java Monitoring & Management Console - pid: 3928 org.apache.catalina.startup.Bootstrap start". The "MBeans" tab is selected, showing a tree view on the left and a table of attribute values on the right.

Tree View:

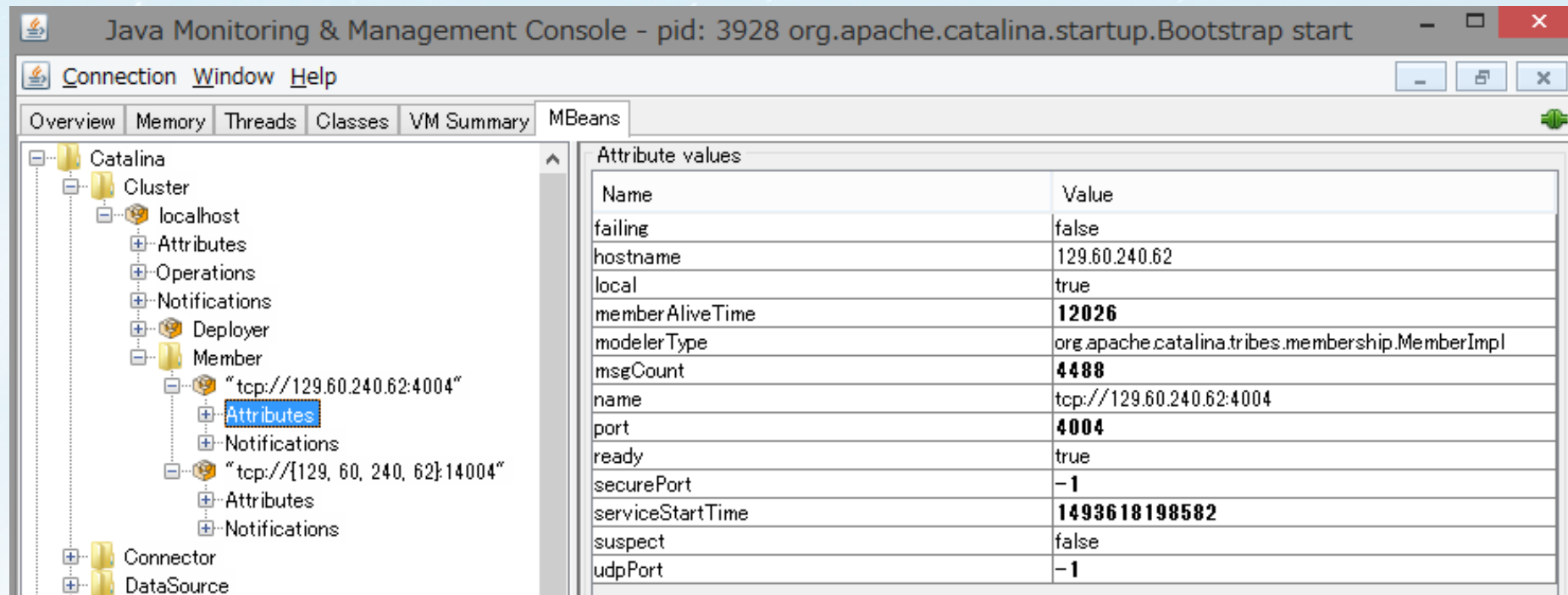
- Catalina
 - Cluster
 - localhost
 - Attributes
 - Operations
 - Notifications
 - Deployer
 - Member
 - Connector
 - DataSource

Attribute values table:

Name	Value
channelSendOptions	10
channelStartOptions	15
clusterName	localhost
heartbeatBackgroundEnabled	false
modelerType	org.apache.catalina.ha.tcp.SimpleTcpCluster
notifyLifecycleListenerOnFailure	false
stateName	STARTED

Member Mbeans

- All Cluster members that have been joining same cluster group



The screenshot displays the Java Monitoring & Management Console window. The title bar reads "Java Monitoring & Management Console - pid: 3928 org.apache.catalina.startup.Bootstrap start". The interface includes a menu bar with "Connection", "Window", and "Help". Below the menu bar are tabs for "Overview", "Memory", "Threads", "Classes", "VM Summary", and "MBeans". The "MBeans" tab is active, showing a tree view on the left and a table of attribute values on the right.

The tree view on the left shows the following structure:

- Catalina
 - Cluster
 - localhost
 - Attributes
 - Operations
 - Notifications
 - Deployer
 - Member
 - "tcp://129.60.240.62:4004"
 - Attributes
 - Notifications
 - "tcp://{129, 60, 240, 62}:14004"
 - Attributes
 - Notifications
 - Connector
 - DataSource

The table of attribute values on the right is as follows:

Name	Value
failing	false
hostname	129.60.240.62
local	true
memberAliveTime	12026
modelerType	org.apache.catalina.tribes.membership.MemberImpl
msgCount	4488
name	tcp://129.60.240.62:4004
port	4004
ready	true
securePort	-1
serviceStartTime	1493618198582
suspect	false
udpPort	-1

Monitoring Cluster

(Cluster)Manager Mbean

- Session Information

The screenshot displays the JMX console interface. On the left, a tree view shows the hierarchy of MBeans under the 'Catalina' domain. The path is: Catalina > Manager > localhost > /test > Attributes. The 'Attributes' folder is selected, and its contents are shown in a table on the right.

Name	Value
accessTimeout	5000
activeSessions	0
activeSessionsFull	0
className	org.apache.catalina.ha.session.BackupManager
duplicates	0
expiredSessions	0
invalidatedSessions	java.lang.String [0]
mapName	/test-map
mapSendOptions	10
maxActive	0
maxActiveSessions	-1
modelerType	org.apache.catalina.ha.session.BackupManager
name	/test
notifyListenersOnReplication	true
processExpiresFrequency	6
processingTime	0
recordAllActions	false
rejectedSessions	0
rpcTimeout	55000
secureRandomAlgorithm	SHA1PRNG
secureRandomClass	
secureRandomProvider	
sessionAttributeNameFilter	
sessionAttributeValueClassNameFilter	
sessionAverageAliveTime	0
sessionCounter	0
sessionMaxAliveTime	0
stateName	STARTED
terminateOnStartFailure	false
warnOnSessionAttributeFilterFailure	false

Monitoring Cluster

(Cluster)Valve MBean

- ReplicationValve settings

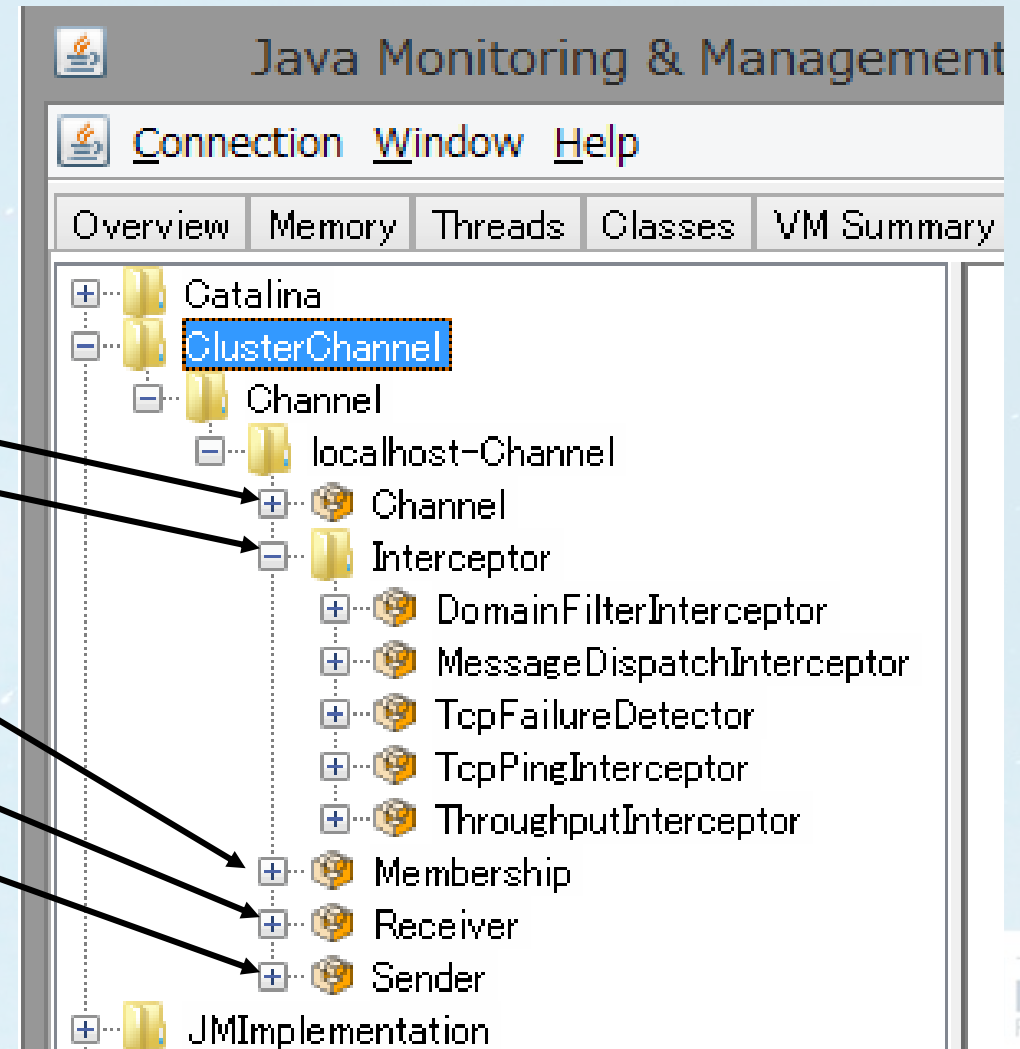
The screenshot shows the Java Monitoring & Management Console window. The title bar reads "Java Monitoring & Management Console - pid: 3928 org.apache.catalina.startup.Bootstrap start". The "MBeans" tab is selected, and the left-hand tree view shows the hierarchy: Service > Valve > localhost > /test > ReplicationValve. The "Attributes" sub-tab is selected for the ReplicationValve MBean. The right-hand pane displays the following attribute values:

Name	Value
asyncSupported	true
doProcessingStats	false
filter	*%gif *%js *%.jpeg *%.jpg *%.png *%.htm *%.html *%.css *%.txt
lastSendTime	0
modelerType	org.apache.catalina.ha.tcp.ReplicationValve
nrOfCrossContextSendRequests	0
nrOfFilterRequests	0
nrOfRequests	0
nrOfSendRequests	0
primaryIndicator	false
primaryIndicatorName	org.apache.catalina.ha.tcp.isPrimarySession
stateName	STARTED
totalRequestTime	0
totalSendTime	0

A "Refresh" button is located at the bottom right of the console window.

Monitoring your Cluster with JMX

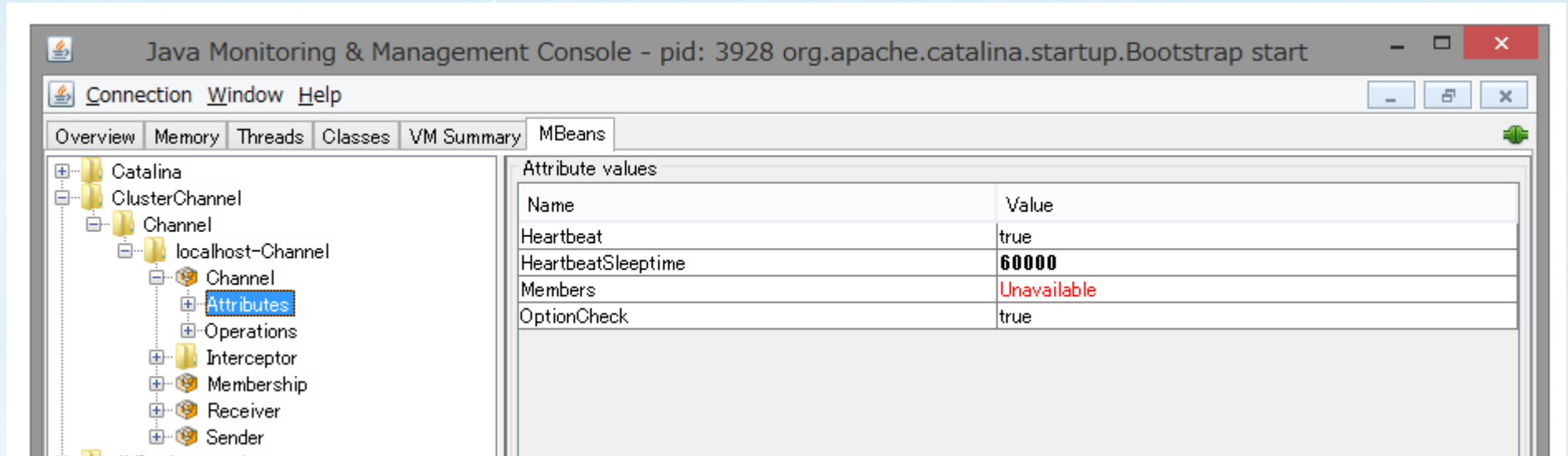
- ClusterChannel Domain
 - Channel Mbeans
 - Channel Mbean
 - Interceptor Mbeans
 - Membership Mbean
 - Receiver Mbean
 - Sender MBean



Monitoring Cluster

Channel Mbean

- Channel settings

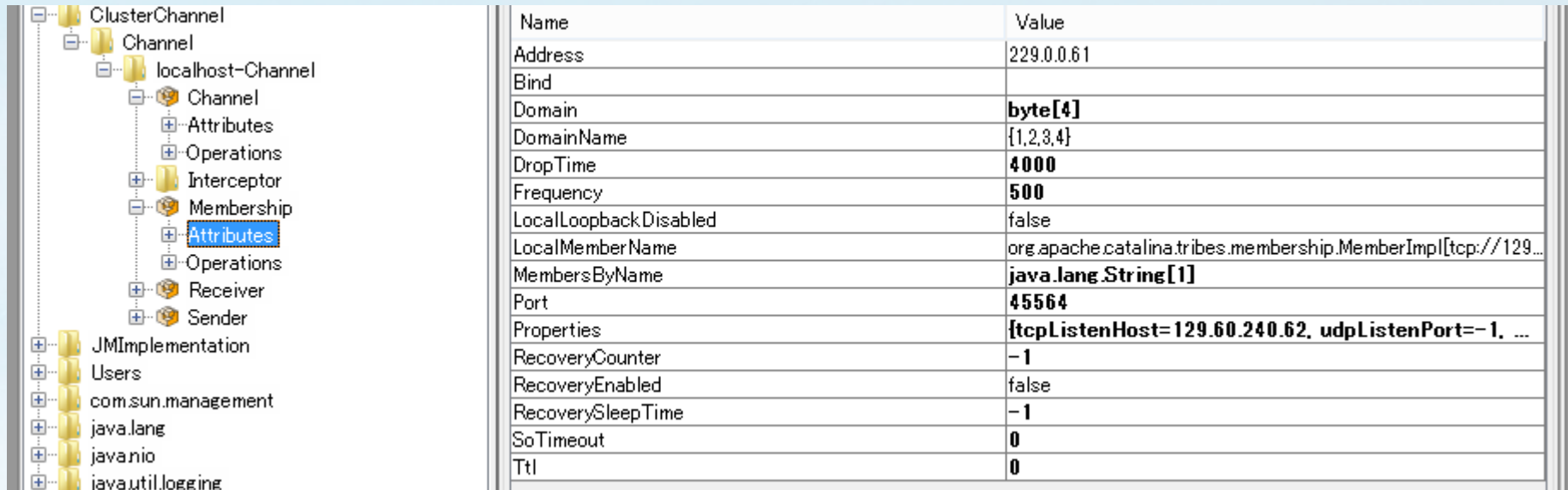


The screenshot shows the Java Monitoring & Management Console window. The title bar reads "Java Monitoring & Management Console - pid: 3928 org.apache.catalina.startup.Bootstrap start". The "MBeans" tab is selected, and the tree view on the left shows the hierarchy: Catalina > ClusterChannel > Channel > localhost-Channel > Channel > Attributes. The "Attributes" node is selected, and the right pane displays the following table:

Attribute values	
Name	Value
Heartbeat	true
HeartbeatSleeptime	60000
Members	Unavailable
OptionCheck	true

Membership Mbean

- Membership Settings



The screenshot displays the JMX console interface. On the left, a tree view shows the hierarchy of MBeans: ClusterChannel > Channel > localhost-Channel > Channel > Membership. The 'Attributes' sub-tree under 'Membership' is expanded and highlighted in blue. On the right, a table lists the properties of the selected Membership Mbean.

Name	Value
Address	229.0.0.61
Bind	
Domain	byte[4]
DomainName	{1,2,3,4}
DropTime	4000
Frequency	500
LocalLoopbackDisabled	false
LocalMemberName	org.apache.catalina.tribes.membership.MemberImpl[tcp://129...
MembersByName	java.lang.String[1]
Port	45564
Properties	{tcpListenHost=129.60.240.62, udpListenPort=-1, ...
RecoveryCounter	-1
RecoveryEnabled	false
RecoverySleepTime	-1
SoTimeout	0
Ttl	0

Monitoring Cluster

Membership Mbean

- Membership Operations

The screenshot displays the JMX console interface. On the left, a tree view shows the hierarchy of MBeans, with the path `ClusterChannel > Channel > localhost-Channel > Channel > Operations` selected. The right pane shows the `Membership` MBean's operations. The `hasMembers` operation is shown as a `boolean` type with no parameters. The `findMemberByName` operation is shown as a `Member` type with one parameter of type `String`.

```
ClusterChannel
├── Channel
│   ├── localhost-Channel
│   │   ├── Channel
│   │   │   ├── Attributes
│   │   │   ├── Operations
│   │   │   ├── Interceptor
│   │   │   ├── Membership
│   │   │   │   ├── Attributes
│   │   │   │   └── Operations
│   │   │   ├── Receiver
│   │   │   └── Sender
│   │   └── JMImplementation
│   └── Users
├── com.sun.management
├── java.lang
├── java.nio
└── java.util.logging
```

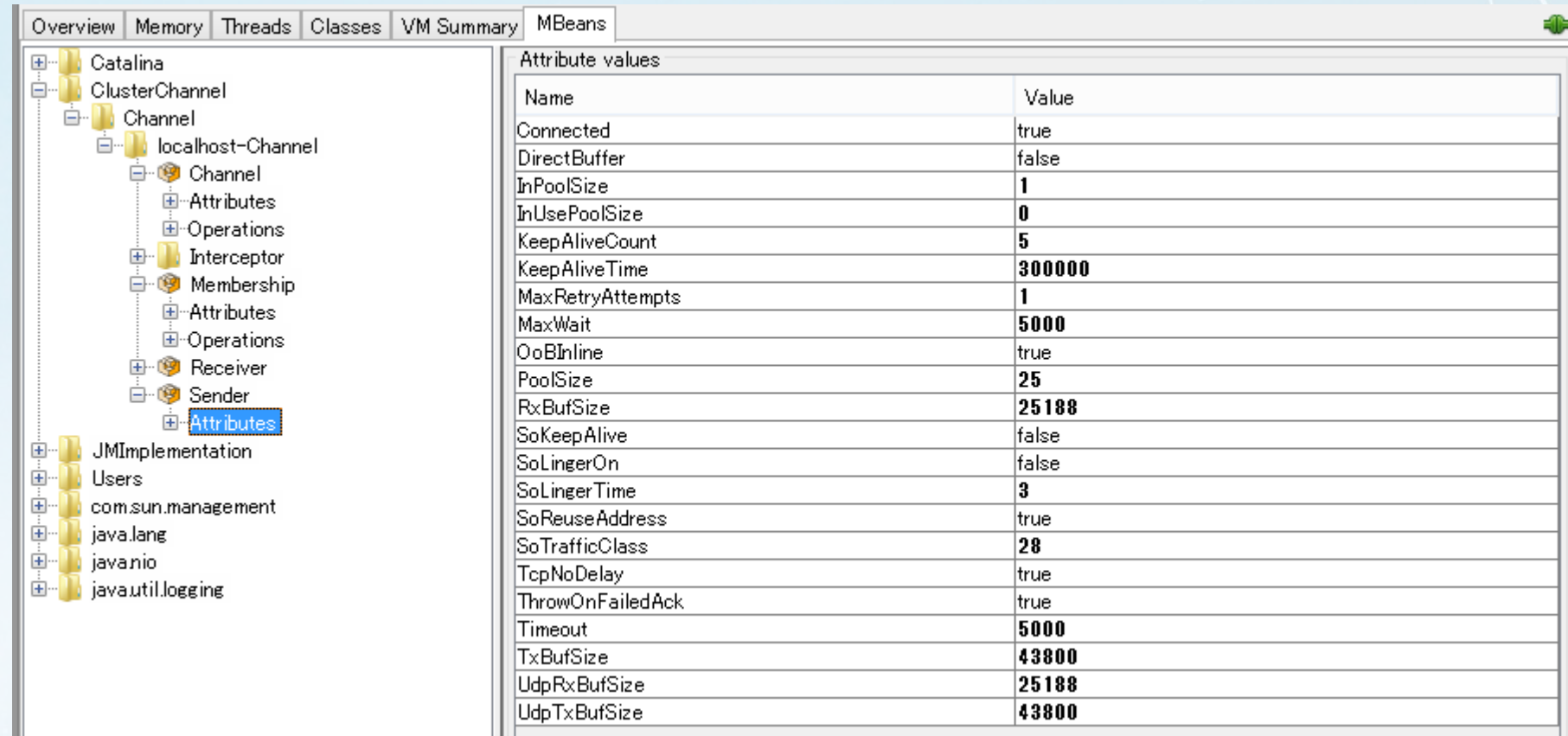
boolean `hasMembers` ()

Member `findMemberByName` (p1 String)

Monitoring Cluster

Sender MBean

- Settings and Stats info



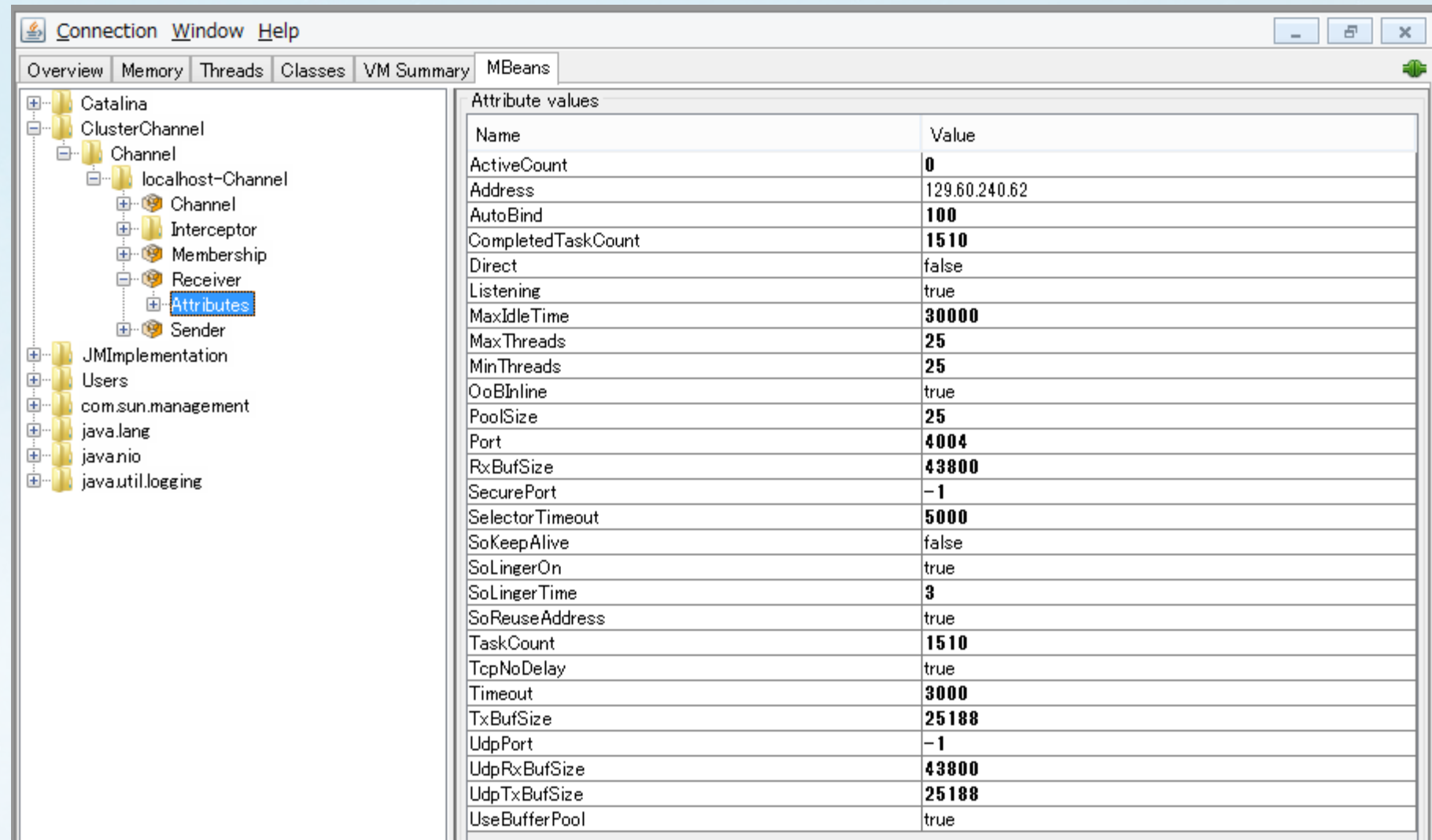
The screenshot displays the JMX console interface. The left pane shows a tree view of the MBean hierarchy, with the path `Catalina > ClusterChannel > Channel > localhost-Channel > Channel > Sender` selected. The right pane shows the 'Attribute values' table for the selected MBean.

Name	Value
Connected	true
DirectBuffer	false
InPoolSize	1
InUsePoolSize	0
KeepAliveCount	5
KeepAliveTime	30000
MaxRetryAttempts	1
MaxWait	5000
OoBInline	true
PoolSize	25
RxBufSize	25188
SoKeepAlive	false
SoLingerOn	false
SoLingerTime	3
SoReuseAddress	true
SoTrafficClass	28
TcpNoDelay	true
ThrowOnFailedAck	true
Timeout	5000
TxBufSize	43800
UdpRxBufSize	25188
UdpTxBufSize	43800

Monitoring Cluster

Receiver Mbean

- Settings and Stats



The screenshot shows the JMX console interface with the following structure:

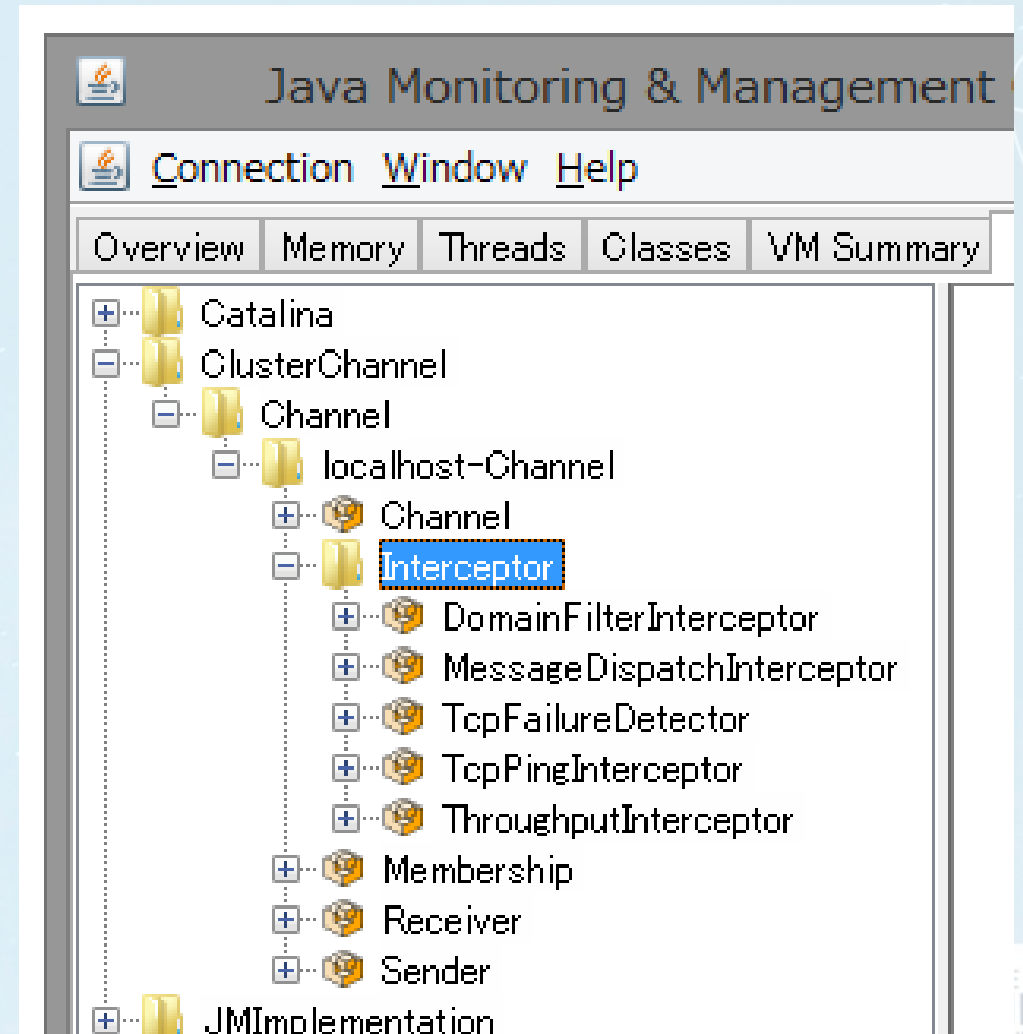
- Catalina
 - ClusterChannel
 - Channel
 - localhost-Channel
 - Channel
 - Interceptor
 - Membership
 - Receiver
 - Attributes
 - Sender
- JMImplementation
- Users
- com.sun.management
- java.lang
- java.nio
- java.util.logging

The 'Attributes' tab is selected, displaying the following table:

Name	Value
ActiveCount	0
Address	129.60.240.62
AutoBind	100
CompletedTaskCount	1510
Direct	false
Listening	true
MaxIdleTime	30000
MaxThreads	25
MinThreads	25
OoBInline	true
PoolSize	25
Port	4004
RxBufSize	43800
SecurePort	-1
SelectorTimeout	5000
SoKeepAlive	false
SoLingerOn	true
SoLingerTime	3
SoReuseAddress	true
TaskCount	1510
TcpNoDelay	true
Timeout	3000
TxBufSize	25188
UdpPort	-1
UdpRxBufSize	43800
UdpTxBufSize	25188
UseBufferPool	true

Interceptor Mbeans

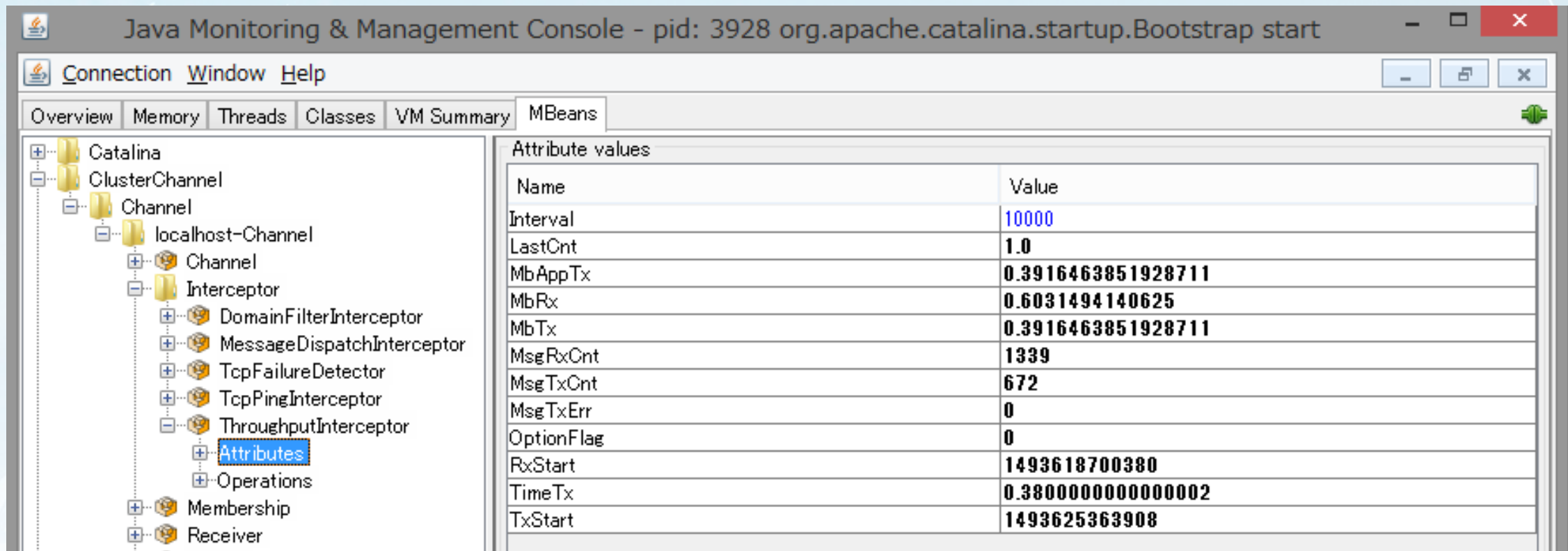
- All interceptor that are used in channel
- Implements MBeans of all commonly used Interceptors but It does not implement all Interceptor MBeans.



Monitoring Cluster

ThroughputInterceptor

- The throughput statistics



The screenshot shows the Java Monitoring & Management Console (JMX Console) for a Catalina server. The 'MBeans' tab is selected, and the 'ThroughputInterceptor' MBean is expanded. The 'Attributes' sub-tab is active, displaying a table of performance metrics.

Name	Value
Interval	10000
LastOnt	1.0
MbAppTx	0.3916463851928711
MbRx	0.6031494140625
MbTx	0.3916463851928711
MsgRxOnt	1339
MsgTxOnt	672
MsgTxErr	0
OptionFlag	0
RxStart	1493618700380
TimeTx	0.38000000000000002
TxStart	1493625363908

Monitoring Cluster

TcpFailureDetector

- Settings and Member check by TCP

Java Monitoring & Management Console - pid: 3928 org.apache.catalina.startup.Bootstrap start

Connection Window Help

Overview Memory Threads Classes VM Summary MBeans

Catalina

- ClusterChannel
 - Channel
 - localhost-Channel
 - Channel
 - Interceptor
 - DomainFilterInterceptor
 - MessageDispatchInterceptor
 - TcpFailureDetector**
 - Attributes
 - Operations
 - TcpPingInterceptor

Attribute values

Name	Value
ConnectTimeout	1000
OptionFlag	0
PerformReadTest	false
PerformSendTest	true
ReadTestTimeout	5000
RemoveSuspectsTimeout	300

Catalina

- ClusterChannel
 - Channel
 - localhost-Channel
 - Channel
 - Interceptor
 - DomainFilterInterceptor
 - MessageDispatchInterceptor
 - TcpFailureDetector
 - Attributes
 - Operations**

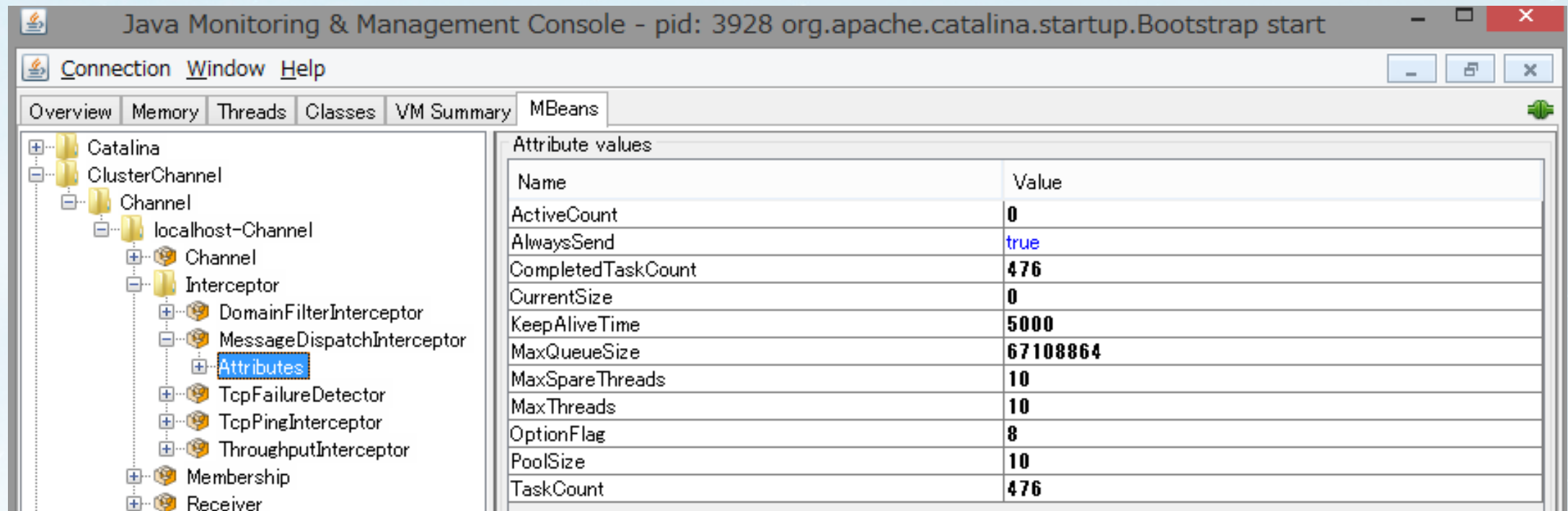
Operation invocation

void checkMembers (p1 true)

Monitoring Cluster

MessageDispatchInterceptor

- Settings and Stats info



Java Monitoring & Management Console - pid: 3928 org.apache.catalina.startup.Bootstrap start

Connection Window Help

Overview Memory Threads Classes VM Summary MBeans

Catalina

- ClusterChannel
 - Channel
 - localhost-Channel
 - Channel
 - Interceptor
 - DomainFilterInterceptor
 - MessageDispatchInterceptor
 - Attributes
 - TcpFailureDetector
 - TcpPingInterceptor
 - ThroughputInterceptor
 - Membership
 - Receiver

Attribute values

Name	Value
ActiveCount	0
AlwaysSend	true
CompletedTaskCount	476
CurrentSize	0
KeepAliveTime	5000
MaxQueueSize	67108864
MaxSpareThreads	10
MaxThreads	10
OptionFlag	8
PoolSize	10
TaskCount	476

Monitoring Cluster

It's important to note that

- Channel Mbeans are supported in 9.0.0.M20 and later
- This feature has not back-ported into Tomcat8.5 yet.
- The ReplicationMap MBean is T.B.D

Questions?

Thank You

Tomcat Cluster

Keiichi Fujino