BigMemory Max Administrator Guide

Version 4.3.3

October 2016

TERRACOTTA

This document applies to Terracotta Server Array Version 4.3.3 and to all subsequent releases.

Specifications contained herein are subject to change and these changes will be reported in subsequent release notes or new editions.

Copyright © 2010-2016 Software AG, Darmstadt, Germany and/or Software AG USA Inc., Reston, VA, USA, and/or its subsidiaries and/or its affiliates and/or their licensors.

The name Software AG and all Software AG product names are either trademarks or registered trademarks of Software AG and/or Software AG USA Inc. and/or its subsidiaries and/or its affiliates and/or their licensors. Other company and product names mentioned herein may be trademarks of their respective owners.

Detailed information on trademarks and patents owned by Software AG and/or its subsidiaries is located at http://softwareag.com/licenses.

Use of this software is subject to adherence to Software AG's licensing conditions and terms. These terms are part of the product documentation, located at http://softwareag.com/licenses and/or in the root installation directory of the licensed product(s).

This software may include portions of third-party products. For third-party copyright notices, license terms, additional rights or restrictions, please refer to "License Texts, Copyright Notices and Disclaimers of Third Party Products". For certain specific third-party license restrictions, please refer to section E of the Legal Notices available under "License Terms and Conditions for Use of Software AG Products / Copyright and Trademark Notices of Software AG Products". These documents are part of the product documentation, located at http://softwareag.com/licenses and/or in the root installation directory of the licensed product(s).

Document ID: BMM-AG-433-20161018

Table of Contents

About the Terracotta Server Array....7
What is the Terracotta Server Array?....8
New for BigMemory Max 4.x....8
Definitions and Functional Characteristics....9

Terracotta Server Array Architecture....13
Terracotta Cluster in Development....14
Terracotta Cluster with Reliability using Fast Restart (FRS)....15
Fast Restart (FRS) Disk Compaction Strategies....17
Terracotta Cluster with High Availability....20
Failover Tuning for Guaranteed Consistency....25
Scaling the Terracotta Server Array....29

Configuring the Terracotta Server Array....33
About Terracotta Server Configuration....34
How Terracotta Servers Get Configured....34
How Terracotta Clients Get Configured....36
Configuration in a Development Environment....38
Configuration in a Production Environment....40
Binding Ports to Interfaces....42
Which Configuration?....43

Automatic Resource Management....45
What is Automatic Resource Management?....46
Eviction....46
Customizing the Eviction Strategy....50

Managing Near-Memory-Full Conditions....51
Behavior of the TSA under Near-Memory-Full Conditions....52
Restricted Mode Operations....53
Recovery....53

Monitoring Cluster Events....55
About Cluster Events....56
Event Types and Definitions....56

Backing Up Live In-Memory Data....63
About Live Backup....64
Creating a Backup....64
The Backup Directory....64
Restoring Data from a Backup....65

Clearing Data from a Terracotta Server....67

How to Clear Data from a Terracotta Server ... 68

Changing Topology of a Live Cluster ... 69
About Changing the Topology ... 70
Adding a New Server ... 70
Removing an Existing Server ... 71
Editing the Configuration of an Existing Server ... 71

Enabling Production Mode ... 73
Setting the Production Mode Property ... 74

Managing Distributed Garbage Collection ... 75
About Distributed Garbage Collection (DGC) ... 76
Running the Periodic Distributed Garbage Collection ... 76
Monitoring and Troubleshooting DGC ... 76

Starting the Terracotta Server as a Windows Service ... 77
Configuring the Terracotta Server to Run as a Service ... 78

Using BigMemory Hybrid ... 81
About BigMemory Hybrid ... 82
System Requirements ... 84
Hardware Capacity Guidelines ... 84
Configuring BigMemory Hybrid ... 84
Using the TMC with BigMemory Hybrid ... 85
Operator Events ... 86

Logging ... 87
SLFJ Logging ... 88
Recommended Logging Levels ... 88

Using Command Central to Manage Terracotta ... 89
Commands that Terracotta Supports ... 90
Configuration Types that Terracotta Supports ... 91
Lifecycle Actions for Terracotta ... 91
Run-time Monitoring Statuses for Terracotta ... 92
Run-time Monitoring States for Terracotta ... 92

Operational Scripts ... 93
Archive Utility (archive-tool) ... 94
Database Backup Utility (backup-data) ... 94
Backup Status (backup-status) ... 95
Cluster Thread and State Dumps (debug-tool, cluster-dump) ... 95
Distributed Garbage Collector (run-dgc) ... 96
Start and Stop Server Scripts (start-tc-server, stop-tc-server) ... 97
Server Status (server-stat) ... 98
Version Utility (version) ... 99

Terracotta Configuration Parameters..... 101
The Terracotta Configuration File..... 102
The Servers Parameters..... 105
/tc:tc-config/servers..... 105
/tc:tc-config/servers/server..... 105
/tc:tc-config/servers/server/data..... 106
/tc:tc-config/servers/server/logs..... 106
/tc:tc-config/servers/server/index..... 106
/tc:tc-config/servers/server/data-backup..... 106
/tc:tc-config/servers/server/tsa-port..... 107
/tc:tc-config/servers/server/jmx-port..... 107
/tc:tc-config/servers/server/tsa-group-port..... 107
/tc:tc-config/servers/server/management-port..... 107
/tc:tc-config/servers/server/security..... 108
/tc:tc-config/servers/server/security/ssl/certificate..... 108
/tc:tc-config/servers/server/security/keychain..... 108
/tc:tc-config/servers/server/security/auth..... 108
/tc:tc-config/servers/server/security/management..... 109
/tc:tc-config/servers/server/authentication..... 109
/tc:tc-config/servers/dataStorage..... 109
/tc:tc-config/servers/mirror-group..... 110
/tc:tc-config/servers/garbage-collection..... 111
/tc:tc-config/servers/restartable..... 112
/tc:tc-config/servers/client-reconnect-window..... 112
The Clients Parameters..... 112
/tc:tc-config/clients/logs..... 112

1 About the Terracotta Server Array

- What is the Terracotta Server Array? 8
- New for BigMemory Max 4.x 8
- Definitions and Functional Characteristics 9

What is the Terracotta Server Array?

The Terracotta Server Array (TSA) provides the platform for Terracotta products and the backbone for Terracotta clusters. A Terracotta Server Array can vary from a basic two-node tandem to a multi-node array providing configurable scale, high performance, and deep failover coverage.

The main features of the Terracotta Server Array include:

- **Distributed In-memory Data Management** - Manages 10-100x more data in memory than data grids
- **Scalability Without Complexity** - Simple configuration to add server instances to meet growing demand and facilitate capacity planning
- **High Availability** - Instant failover for continuous uptime and services
- **Configurable Health Monitoring** - Terracotta HealthChecker for inter-node monitoring. For information, see "Configuring the HealthChecker Properties" in the *BigMemory Max High-Availability Guide*.
- **Persistent Application State** - Automatic permanent storage of all current shared in-memory data
- **Automatic Node Reconnection** - Temporarily disconnected server instances and clients rejoin the cluster without operator intervention

New for BigMemory Max 4.x

The 4.x TSA is an in-memory data platform, providing faster, more consistent, and more predictable access to data. With resource management, if you have more data than memory available, the TSA protects itself from going over its limit through data eviction and throttling. In most cases, it will recover and come back to its normal working state automatically. In addition, four systems are available to protect data: the Fast Restart (FRS) feature, BigMemory Hybrid's use of SSD/Flash, active-mirror server groups, and backups.

Fast Restartability for Data Persistence

BigMemory's Fast Restart (FRS) feature is now integrated into the TSA, providing crash resilience with quick recovery, plus a consistent record of the entire in-memory data set, no matter how large. For more information about FRS, see "Fast Restartability" on page 15.

Hybrid Data Storage

"BigMemory Hybrid" extends BigMemory distributed in a Terracotta Server Array so that data can be stored across a hybrid mixture of RAM and SSD/Flash. This additional

storage is managed with the in-memory data as one TSA data set. For more information, see "Using BigMemory Hybrid" on page 81.

Resource Management

Resource management provides better control over the TSA's in-memory data through time, size, and count limitations. This enables automatic handling of, and recovery from, near-memory-full conditions. For more information, see "Automatic Resource Management" on page 45.

Predictable Eviction Strategy

Based upon user-configured time, size, and count limitations, the TSA's 3-pronged eviction strategy works automatically to ensure predictable behavior when memory becomes full. For more information, see "Eviction " on page 46.

Continuous Uptime

Improvements to provide continuous availability of data include flexibility in server startup sequencing, better utilization of extra mirrors in mirror groups, multi-stripe backup capability, optimizations to bulk load, and performance improvements for data access on rejoin. In addition, the TSA no longer uses Oracle Berkeley DB, enabling in-memory data to be ready for use much more quickly after any planned or unplanned restart.

Terracotta Management Console (TMC)

The expanded TMC replaces the Developer Console and Operations Center as the integrated platform for monitoring, managing, and administering all Terracotta deployments. There is also support for additional REST APIs for management and monitoring. For more information, start with the *Terracotta Management Console User's Guide*.

Additional Security Features

Active Directory (AD) and Lightweight Directory Access Protocol (LDAP) support on Terracotta servers, and custom SecretProvider on Terracotta clients. For more information, see the *BigMemory Max Security Guide*.

No More DSO, plus Simplified Configuration

DSO configuration has been deprecated, and the tc-config has a new format. Most of the elements are the same, but the structure is revised. For more information, see the *BigMemory Max Administrator Guide*.

Definitions and Functional Characteristics

The major components of a Terracotta installation are the following:

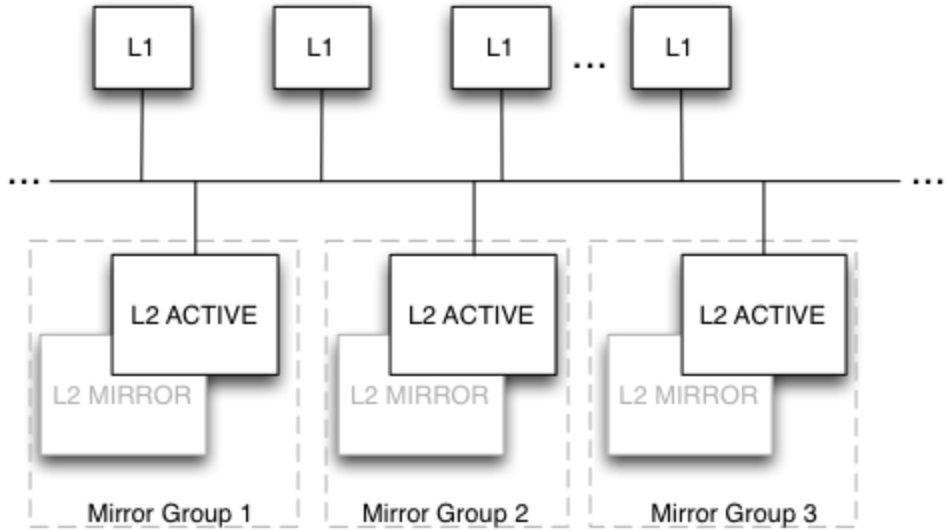
- **Cluster** - All of the Terracotta server instances and clients that work together to share application state or a data set.

- **Terracotta client** - Terracotta clients run on application servers along with the applications being clustered by Terracotta. Clients manage live shared-object graphs.
- **Terracotta server instance** - A single Terracotta server. An *active* server instance manages Terracotta clients, coordinates shared objects, and persists data. Server instances have no awareness of the clustered applications running on Terracotta clients. A mirror (sometimes called "hot standby") is a live backup server instance which continuously replicates the shared data of an active server instance, instantaneously replacing the active if the active fails. *Mirror servers add failover coverage within each mirror group*.
- **Terracotta mirror group** - A unit in the Terracotta Server Array. Sometimes also called a "stripe," a mirror group is composed of exactly one active Terracotta server instance and at least one mirror Terracotta server instance. The active server instance manages and persists the fraction of shared data allotted to its mirror group, while each mirror server in the mirror group replicates (or mirrors) the shared data managed by the active server. *Mirror groups add capacity to the cluster*. The mirror servers are optional but highly recommended for providing failover.
- **Terracotta Server Array** - The platform, consisting of all of the Terracotta server instances in a single cluster. Clustered data, also called in-memory data, or shared data, is partitioned equally among active Terracotta server instances for management and persistence purposes.

Tip: Nomenclature - This documentation may refer to a Terracotta server instance as L2, and a Terracotta client (the node running your application) as L1. These are the shorthand references used in Terracotta configuration files.

Figure 1 illustrates a Terracotta cluster with three mirror groups. Each mirror group has an active server and a mirror, and manages one third of the shared data in the cluster.

Fig. 1: A Server Array With 3 Mirror Groups



A Terracotta cluster has the following functional characteristics:

- Each mirror group automatically elects one active Terracotta server instance. There can never be more than one active server instance per mirror group, but there can be any number of mirrors. However, a performance overhead may become evident when adding more mirror servers due to the load placed on the active server by having to synchronize with each mirror.
- Every mirror group in the cluster must have a Terracotta server instance in active mode before the cluster is ready to do work.
- The shared data in the cluster is automatically partitioned and distributed to the mirror groups. The number of partitions equals the number of mirror groups. In Fig. 1, each mirror group has one third of the shared data in the cluster.
- Mirror groups cannot provide failover for each other. Failover is provided within each mirror group, not across mirror groups. This is because mirror groups provide scale by managing discrete portions of the shared data in the cluster -- they do not replicate each other. In Fig. 1, if Mirror Group 1 goes down, the cluster must pause (stop work) until Mirror Group 1 is back up with its portion of the shared data intact.
- Active servers are self-coordinating among themselves. No additional configuration is required to coordinate active server instances.
- Only mirror server instances can be hot-swapped in an array. In Fig. 1, the L2 MIRROR servers can be shut down and replaced with no affect on cluster functions. However, to add or remove an entire mirror group, the cluster must be brought down. Note also that in this case the original Terracotta configuration file is still in effect and no new servers can be added. Replaced mirror servers must have the same address (hostname or IP address). If you must swap in a mirror with a different configuration, see "Changing Topology of a Live Cluster" on page 69.

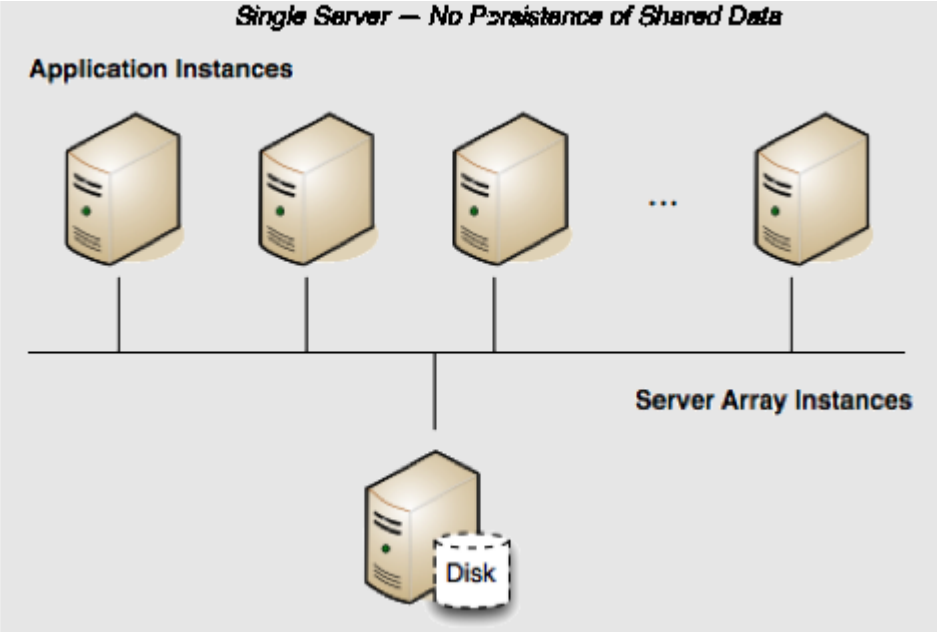
2 Terracotta Server Array Architecture

- Terracotta Cluster in Development 14
- Terracotta Cluster with Reliability using Fast Restart (FRS) 15
- Fast Restart (FRS) Disk Compaction Strategies 17
- Terracotta Cluster with High Availability 20
- Failover Tuning for Guaranteed Consistency 25
- Scaling the Terracotta Server Array 29

Terracotta Cluster in Development

Persistence: No | Failover: No | Scale: No

In a development environment, persisting shared data is often unnecessary and even inconvenient. Running a single-server Terracotta cluster without persistence is a good solution for creating an efficient development environment.



By default, a Terracotta server has Fast Restartability (*FRS*) disabled, which means it will not persist data after a restart. Its configuration could look like the following:

```
<?xml version="1.0" encoding="UTF-8" ?>
<tc:tc-config xmlns:tc="http://www.terracotta.org/config"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.terracotta.org/schema/terracotta-9.xsd">
 <servers>
   <server name="Server1">
     <data>/opt/terracotta/server1-data</data>
     <tsa-port>9510</tsa-port>
     <jmx-port>9520</jmx-port>
     <tsa-group-port>9530</tsa-group-port>
     <management-port>9540</management-port>
     <dataStorage size="4g">
        <offheap size="4g"/>
     </dataStorage>
   </server>
  </servers>
 ...
</tc:tc-config>
```

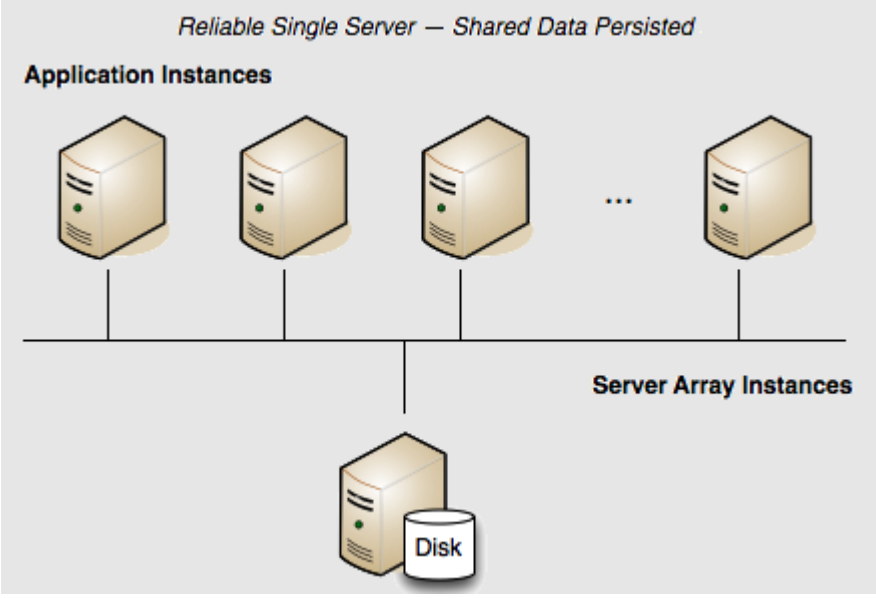
If this server goes down, the application state (all clustered data) in the shared memory is lost. In addition, when the server is up again, all clients must be restarted to rejoin the

cluster. Note that servers are required to run with off-heap, and that shared data is also lost.

Terracotta Cluster with Reliability using Fast Restart (FRS)

Persistence: Yes | Failover: No | Scale: No

The configuration above may be advantageous in development, but if shared in-memory data must be persisted, the server should be configured to use its local disk. Terracotta servers achieve data persistence with the Fast Restart (FRS) feature.



Fast Restartability

The Fast Restart (FRS) feature provides enterprise-ready crash resilience by keeping a fully consistent, real-time record of your in-memory data. After any kind of shutdown - planned or unplanned - the next time your application starts up, all of your BigMemory Max data is still available and very quickly accessible.

The Fast Restart feature persists the real-time record of the in-memory data in a Fast Restart store on the server's local disk. After any restart, the data that was last in memory (both heap and off-heap stores) automatically loads from the Fast Restart store back into memory. In addition, previously connected clients are allowed to rejoin the cluster within a window set by the `<client-reconnect-window>` element.

To configure the Terracotta server for Fast Restartability, add and enable the `<restartable>` element in the tc-config.xml.

```
<?xml version="1.0" encoding="UTF-8" ?>
<tc:tc-config xmlns:tc="http://www.terracotta.org/config"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.terracotta.org/schema/terracotta-9.xsd">
 <servers>
   <server name="Server1">
```

```
    <data>/opt/terracotta/server1-data</data>
    <tsa-port>9510</tsa-port>
    <jmx-port>9520</jmx-port>
    <tsa-group-port>9530</tsa-group-port>
    <management-port>9540</management-port>
    <dataStorage size="4g">
       <offheap size="4g"/>
    </dataStorage>
  </server>
  <!-- Fast Restartability must be added explicitly. -->
  <restartable enabled="true"/>
  <!-- By default the window is 120 seconds. -->
  <client-reconnect-window>120</client-reconnect-window>
 </servers>
 ...
</tc:tc-config>
```

Disk usage

Fast Restartability requires a unique and explicitly specified path. The default path is the Terracotta server's home directory. You can customize the path using the `<data>` element in the server's tc-config.xml configuration file.

The Terracotta Server Array can be configured to be restartable in addition to including searchable caches, but both of these features require disk storage. When both are enabled, be sure that enough disk space is available. The amount of disk storage required is dependent on the use case, taking into account aspects such as the number of searchable attributes, element size, read/write ratio and transaction rate. For this reason an exact figure cannot be given, but the amount of disk storage required can be a multiple of the amount of in-memory data. The recommended way to determine the true minimum disk space requirements is to perform multiple iterations of testing the real-world use case in full.

It is highly recommended to store the search index (`<index>`) and the Fast Restart data (`<data>`) on separate disks.

Client Reconnect Window

The `<client-reconnect-window>` does not have to be explicitly set if the default value is acceptable. However, in a single-server cluster, `<client-reconnect-window>` is in effect only if restartable mode is enabled.

Understanding how FRS works

FRS is a trade-off of space efficiency against performance. Because the Terracotta product is about performance, and in particular in-memory performance, the FRS feature purposely sacrifices on-disk space efficiency in order to avoid making the process (the Terracotta server) wait for more complex disk I/O. In periods of extensive writing (new entries, updated entries, deleted entries) to the system, the FRS transaction logs will grow extensively. This design preserves high, predictable performance (latencies).

Fast Restart (FRS) Disk Compaction Strategies

Since Terracotta is a product intended to maximize performance, a performance-focused "append-only-log" architecture is used for the FRS disk persistence technology.

Terracotta supports the following two compaction strategies:

1. Performance Based Compaction Policy - Favors predictable performance at the expense of unpredictable disk storage space requirements. This is the default strategy.
2. Size Based Compaction Policy - Favors predictable disk storage space requirements at the expense of unpredictable performance.

Overview of Behavior

The following process describes in general what you can expect to observe occurring on the file system in the configured <data> folder. This process is the same, regardless of which of the two compaction strategies you use:

1. Data is added to Terracotta cache(s).
2. An FRS file is created with a file name pattern similar to "seg<nnnn>.frs", where <nnnn> is a number. This file receives all current data writes.
3. As data is further added to Terracotta cache(s), the current FRS file will grow in size.
4. Once the current FRS file grows to some predetermined size, the file is closed.
5. Repeat from Step 2 above, where the file name's <nnnn> is incremented by 1.

It is important to note from the above process that any given FRS file only grows in size to some limit before it is closed and another file is created. This leads to increasing disk consumption.

Based on internal product heuristics, Terracotta will periodically try to "copy" living/valid data into newer FRS files. Once all the cached data in an older FRS file is either "dead" (i.e. the "time to live" has expired) or has been successfully copied to a newer FRS file, then that older FRS file will be deleted in its entirety and its associated disk space will be reclaimed.

In short, disk usage will accumulate to some amount, then some storage space will be "freed", and the process will repeat itself.

Note: It is very important to determine upper-bound used-disk space and ensure the environment always has enough free storage capacity to safely function. We recommend determining this upper-bound, then pad this figure by approximately 20% or more (to accommodate variations in usage), and set that as the "minimum required free disk space" for successful product operation.

Performance Based Compaction Policy (Default)

As noted above, the Performance Based Compaction Policy is the default compaction strategy because it provides the fastest predictable performance which is of paramount concern to "Big Data" users. This strategy frees up disk space associated with data which is either redundant or has been removed from cache. This strategy provides fast, predictable performance at the cost of large (and comparatively cheap) disk storage space. Put another way, this compaction strategy intentionally sacrifices disk space to provide maximal disk performance.

The amount of disk storage space that is required to successfully operate Terracotta when using this compaction strategy is wholly dependent upon a variety of use-case specific factors (such as data size, data volume, configured data lifetime, access patterns, etc.) which cannot reasonably be "guessed at" to determine the minimal FRS disk space requirements for a deployment. The only way to know how much disk space will be required for successful operation is to test the real use-case, under real load, and monitor the resulting disk usage.

The following are some observed real world examples. Please note that every use-case is different and that any particular usage could yield wildly different results; the following is intended to be used for illustrative purposes only to show what is possible as a normally expected disk usage.

Scenario: 50 MB Real Data

FRS was designed to efficiently manage Big Data use-cases; the side effect is that small data volumes typically exhibit a high "relative" disk usage cost.

With this use-case, there are approximately 50 MB of real data which is completely reloaded/refreshed every 24 hours from the backing data store. Based on the data lifetime requirements and access patterns the following behavior might result:

1. Each FRS file slowly grows in size to approximately 512 MB before another file is created.
2. The number of FRS files accumulates to an average of 14 FRS files (each at ~512 MB) before the configured data life-cycle coupled with the data access patterns allows FRS files to be freed.
3. All "old" FRS files are deleted except the most recent 3 FRS files.

Under normal operation, FRS disk usage could grow up to approximately 7.1 GB (14 files of 512 MB each) before 11 FRS files would be deleted (freeing about 5.6 GB) leaving 1.5 GB used on disk. Then the process would repeat, where the 1.5 GB grows to 7.1 GB, at which point 5.6 GB would be freed again. This pattern would repeat over the lifetime of this customer's deployment.

Scenario: 11 GB Real Data

With this use-case, there are approximately 11 GB of real data which is completely reloaded/refreshed every 24 hours from the backing data store. If the used disk-space is observed to peak at approximately 50 GB, we would recommend ensuring that there are 60 GB of free space (50 GB padded by 20%) dedicated to Terracotta data storage.

Scenario: 760 GB Real Data (4 Stripe Terracotta Server Array)

With this use-case, there are approximately 760 GB of real data distributed across 4 Terracotta Servers, with each Terracotta Server containing approximately 190 GB of that data. The majority of the time, this deployment is used for a read-heavy "pure caching" use-case which results in FRS disk usage peaking at about 500 GB on each Terracotta Server. For this usage, it should be safe to ensure that there is 600 GB (500 GB + 20%) of free disk storage capacity reserved for Terracotta usage.

However, once every 2 weeks, a write-heavy "data-reconciliation" process is executed which updates the entire 760 GB data-set (~190 GB per server). Due to the access patterns required to perform the data reconciliation during this bi-weekly process, FRS disk usage temporarily peaks during this window of time on each Terracotta Server at about 2000 GB (2 TB).

In summary, each Terracotta Server:

1. Stores approximately 190 GB of data.
2. Under normal operation requires at least 500 GB of free disk capacity.
3. Only during the bi-weekly data reconciliation operation, the minimum free disk space requirement grows from the normal 500 GB to 2 TB.

Based on the observed application requirements, it is recommended to allocate at least 2.4 TB (2 TB + 20%) of free disk storage capacity to safely operate during this bi-weekly process.

Optional Compaction Strategy (Size Based Compaction Policy)

If constraining FRS on-disk usage is more important than performance (i.e. predictable performance is not important), then the "Size Based Compaction Policy" can be used instead.

This optional compaction strategy is configured in the Terracotta Server's "tc-config.xml" by adding a new "tc-property" to the "tc-properties" section as follows:

```
<tc-properties>
    <property name="l2.frs.compactor.policy" value="SizeBasedCompactionPolicy"/>
</tc-properties>
```

By default, this compaction strategy will attempt to constrain the on-disk size to approximately 2 times the overall data size but is influenced by the FRS "segment size".

The default FRS segment file size is 512 MB (53687092 bytes), but this can be further tuned by changing the value used the following "tc-property":

```
<property name="l2.frs.io.nio.segmentSize" value="53687092"/>
<!-- 512MB = 32 * 1024 * 1024 -->
```

For small data sets, it might make sense to reduce the FRS segment file size relative to the amount of data being stored; such tuning will be an iterative process is dependent upon the data storage and access requirements of the application.

Configuration Example: 32 MB segment size

```
<tc-properties>
```

```
    <property name="l2.frs.compactor.policy" value="SizeBasedCompactionPolicy"/>
    <property name="l2.frs.io.nio.segmentSize" value="33554432"/>  <!-- 32MB = 32 * 1024 * 1024 -
</tc-properties>
```

Configuration Example: 256 MB segment size

```
<tc-properties>
    <property name="l2.frs.compactor.policy" value="SizeBasedCompactionPolicy"/>
    <property name="l2.frs.io.nio.segmentSize" value="268435456"/>  <!-- 256MB = 256 * 1024 * 102
</tc-properties>
```

Configuration (Advanced)

The vast majority of use-cases should be successfully handled using the above options; what follows is included for completeness. These additional tuning parameters are only applicable when the Size Based Compaction Policy is being used and should be used with great caution.

Note: The "value" of each parameter is expressed in percent as a fraction of 1.0 such that 100%=1.0, 85%=0.85, 50%=0.5, 5%=0.05, etc.

The following property controls when to trigger/start the on-disk data compaction process. The formula used to determine the trigger threshold is INMEMORY_DATA / DISK_USED.

```
<!-- start compaction when in-memory size is 50% of what on-disk size is,
or in other words when disk is twice as much -->
<property name="l2.frs.compactor.sizeBased.threshold" value="0.50"/>
```

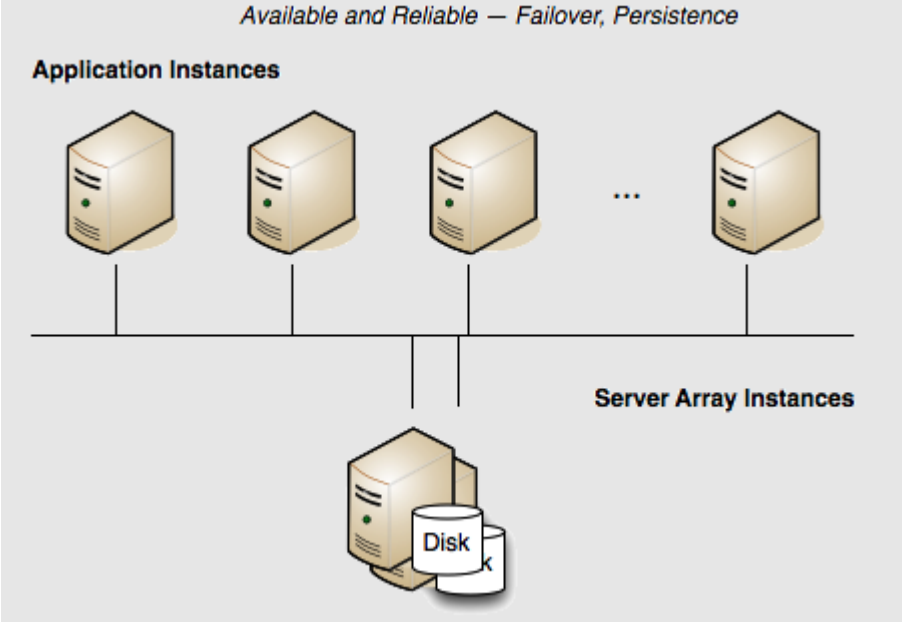
The following property controls how much data should be compacted at a time. By default, the Size Based Compaction Policy will attempt to free 5% of the on-disk stored data.

```
<property name="l2.frs.compactor.sizeBased.amount" value="0.05"/>
```

Terracotta Cluster with High Availability

Persistence: Yes | Failover: Yes | Scale: No

The example above presents a reliable but *not* highly available cluster. If the server fails, the cluster fails. There is no redundancy to provide failover. Adding a mirror server adds availability because the mirror serves as a "hot standby" ready to take over for the active server in case of a failure.



In this array, if the active Terracotta server instance fails, then the mirror instantly takes over and the cluster continues functioning. No data is lost.

The following Terracotta configuration file demonstrates how to configure this two-server array:

```
<?xml version="1.0" encoding="UTF-8" ?>
<tc:tc-config xmlns:tc="http://www.terracotta.org/config"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.terracotta.org/schema/terracotta-9.xsd">
 <servers>
   <server name="Server1">
     <data>/opt/terracotta/server1-data</data>
     <tsa-port>9510</tsa-port>
     <jmx-port>9520</jmx-port>
     <tsa-group-port>9530</tsa-group-port>
     <management-port>9540</management-port>
     <dataStorage size="4g">
        <offheap size="4g"/>
     </dataStorage>
   </server>
   <server name="Server2">
     <data>/opt/terracotta/server2-data</data>
     <tsa-port>9510</tsa-port>
     <jmx-port>9520</jmx-port>
     <tsa-group-port>9530</tsa-group-port>
     <management-port>9540</management-port>
     <dataStorage size="4g">
        <offheap size="4g"/>
     </dataStorage>
   </server>
   <restartable enabled="true"/>
   <client-reconnect-window>120</client-reconnect-window>
 </servers>
 ...
</tc:tc-config>
```

You can add more mirror servers to this configuration by adding more `<server>` sections. However, a performance overhead may become evident when adding more mirror servers due to the load placed on the active server by having to synchronize with each mirror.

Note: Terracotta server instances must not share data directories. Each server's `<data>` element should point to a different and preferably local data directory.

Starting the Servers

How server instances behave at startup depends on when in the life of the cluster they are started.

In a single-server configuration, when the server is started it performs a startup routine and then is ready to run the cluster (ACTIVE status). If multiple server instances are started at the same time, one is elected the active server (ACTIVE-COORDINATOR status) while the others serve as mirrors (PASSIVE-STANDBY status). The election is recorded in the servers' logs.

If a server instance is started while an active server instance is already running, it syncs up state from the active server instance before becoming a mirror. The active and mirror servers must always be synchronized, allowing the mirror server to mirror the state of the active. The mirror server goes through the following states:

1. **PASSIVE-UNINITIALIZED** - The mirror is beginning its startup sequence and is *not* ready to perform failover should the active fail or be shut down. The server's status light in the Terracotta Management Console (TMC) switches from red to orange.
2. **INITIALIZING** - The mirror is synchronizing state with the active and is *not* ready to perform failover should the active fail or be shut down. The server's status light in the TMC is orange.
3. **PASSIVE-STANDBY** - The mirror is synchronized and is ready to perform failover should the active server fail or be shut down. The server's status light in the TMC switches from orange to cyan.

The active server instance carries the load of sending state to the mirror during the synchronization process. The time taken to synchronize is dependent on the amount of clustered data and on the current load on the cluster. The active server instance and mirrors should be run on similarly configured machines for better throughput, and should be started together to avoid unnecessary sync ups.

The sequence in which servers startup does not affect data. Even if a former mirror server is initialized before the former active server, the mirror server's data is not erased. In the event that a mirror server went offline while the active server was still up, then when the mirror server returns, it remembers that it was in the mirror role. Even if the active server is offline at that point, the mirror server does not try to become the active. It waits until the active server returns, and clients are blocked from updating their data. When the active returns, it will restart the mirror. The mirror's data objects and indices

are then moved to the dirty-objectdb-backup directory, and the active syncs its data with the mirror.

Failover

If the active server instance fails and two or more mirror server instances are available, an election determines the new active. Successful failover to a new active takes place only if at least one mirror server is fully synchronized with the failed active server; successful client failover (migration to the new active) can happen only if the server failover is successful. Shutting down the active server before a fully-synchronized mirror is available can result in a cluster-wide failure.

If the `dataStorage` and/or `offheap` size on the mirror server is smaller than on the active server, then the mirror server will fail to start and the user will be alerted that the configuration is invalid. If there are multiple mirrors with differing amounts of storage configured, then the passive with the smallest `dataStorage` and `offheap` sizes (that are still greater than or equal to the active's `dataStorage` and `offheap` sizes) will be elected to be the new active.

Tip: Hot-Swapping Mirrors - A mirror can be hot-swapped if the replacement matches the original mirror's <server> block in the Terracotta configuration. For example, the new mirror should use the same host name or IP address configured for the original mirror. For information about swapping in a mirror with a different configuration, refer to "Changing Topology of a Live Cluster" on page 69.

Terracotta server instances acting as mirrors can run either in restartable mode or non-persistent mode. If a server instance running in restartable mode goes down, and a mirror takes over, the crashed server's data directory is cleared before it is restarted and allowed to rejoin the cluster. Removing the data is necessary because the cluster state could have changed since the crash. During startup, the restarted server's new state is synchronized from the new active server instance.

If both servers are down, and clustered data is persisted, the last server to be active will automatically be started first to avoid errors and data loss.

In setups where data is not persisted, meaning that restartable mode is not enabled, then no data is saved and either server can be started first.

Note: Under certain circumstances pertaining to server restarts, the data directory should be manually cleared. For more information, refer to "Clearing Data from a Terracotta Server" on page 67

A Safe Failover Procedure

To safely migrate clients to a mirror server without stopping the cluster, follow these steps:

1. If it is not already running, start the mirror server using the start-tc-server script. The mirror server must already be configured in the Terracotta configuration file.

2. Ensure that the mirror server is ready for failover (PASSIVE-STANDBY status). In the TMC, the status light will be cyan.
3. Shut down the active server using the stop-tc-server script.

   Note: If the script detects that the mirror server in STANDBY state isn't reachable, it issues a warning and fails to shut down the active server. If failover is not a concern, you can override this behavior with the `--force` flag.

   Clients will connect to the new active server.
4. Restart any clients that fail to reconnect to the new active server within the configured reconnection window.

The previously active server can now rejoin the cluster as a mirror server. If restartable mode had been enabled, its data is first removed and then the current data is read in from the now active server.

A Safe Cluster Shutdown Procedure

A safe cluster shutdown should follow these steps:

1. Shut down the mirror servers using the stop-tc-server script.
2. Shut down the clients. The Terracotta client will shut down when you shut down your application.
3. Shut down the active server using the stop-tc-server script.

To restart the cluster, first start the server that was last active. If clustered data is not persisted, any of the servers could be started first as no data conflicts can take place.

Split Brain Scenario

In a Terracotta cluster, "split brain" refers to a scenario where two servers assume the role of active server (ACTIVE-COORDINATOR status). This can occur during a network problem that disconnects the active and mirror servers, causing the mirror to both become an active server and open a reconnection window for clients (<client-reconnect-window>).

If the connection between the two servers is never restored, then two independent clusters are in operation. This is a split-brain situation. However, if the connection is restored, one of the following scenarios results:

- No clients connect to the new active server - The original active server "zaps" the new active server, causing it to restart, wipe its database, and synchronize again as a mirror.
- A minority of clients connect to the new active server - The original active server starts a reconnect timeout for the clients that it loses, while zapping the new active server. The new active restarts, wipes its database, and synchronizes again as a mirror. Clients that defected to the new active attempt to reconnect to the original

active, but if they do not succeed within the parameters set by that server, they must be restarted.

- A majority of clients connects to the new active server - The new active server "zaps" the original active server. The original active restarts, wipes its database, and synchronizes again as a mirror. Clients that do not connect to the new active within its configured reconnection window must be restarted.
- An equal number of clients connect to the new active server - In this unlikely event, exactly one half of the original active server's clients connect to the new active server. The servers must now attempt to determine which of them holds the latest transactions (or has the freshest data). The winner zaps the loser, and clients behave as noted above, depending on which server remains active. Manual shutdown of one of the servers may become necessary if a timely resolution does not occur.

In the case of split-brain occurrences it is imperative to confirm the integrity of shared data after such an event.

Note: Under certain circumstances pertaining to server restarts, the data directory should be manually cleared. For more information, refer to "Clearing Data from a Terracotta Server" on page 67.

Failover Tuning for Guaranteed Consistency

In a clustered environment any network, hardware, or other issues can result in an active node to partition from the cluster.

The detection of such a situation results in an Active-left event.

As described in the previous sections, the default behavior of a TSA would be that the remaining passive node will then run an election and, if not finding the active node (5 sec default), take over as Active.

While this configuration ensures a tight availability of the data, risks of experiencing a so-called split brain situation during such elections are increased.

In the case of a TSA, split brain would be a situation in which both nodes are acting as Active. Any further operations performed on the data are likely to result in inconsistencies.

So, depending on mission priorities, a cluster may also be configured to emphasize consistency.

Note: Using Failover Tuning requires the feature to be part of the license key. Contact your nearest Software AG representative in case of questions.

AVAILABILITY versus CONSISTENCY

AVAILABILITY	default setting is 5 seconds	Passives automatically become Actives
CONSISTENCY	explicit settings `<failover-priority>`	WAITING-FOR-PROMOTION state requesting operator to issue a failover-action command

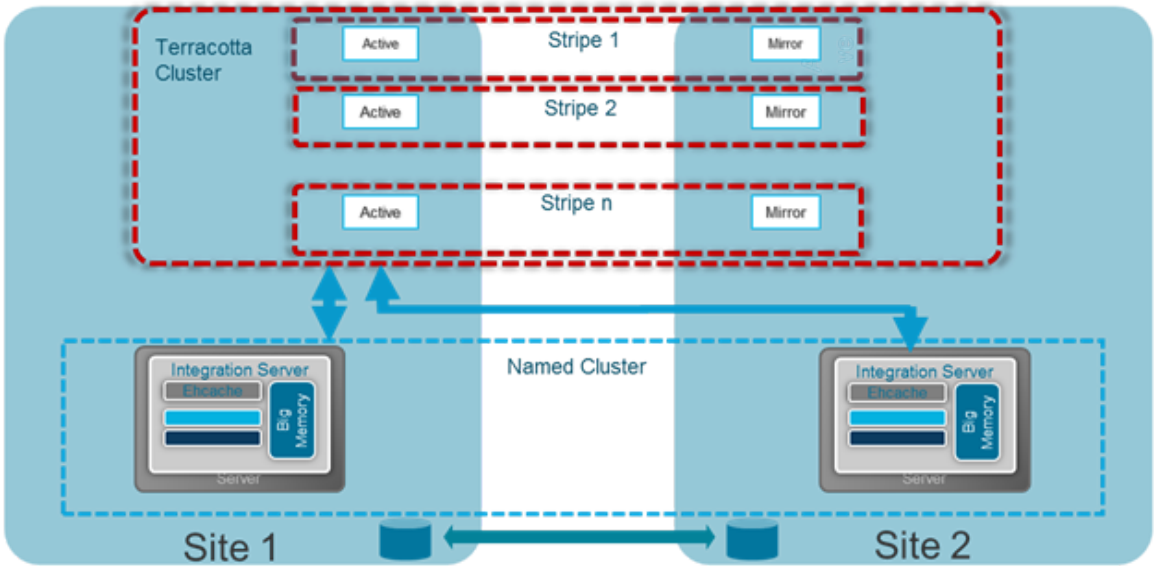
Note: Even in the absence of any configuration value, the default behavior is Availability.

Supported Configuration

Tuning the failover priority to CONSISTENCY can be applied to clusters consisting of up to two data centers.

Each mirror group in the cluster has one active and one mirror node. Mirror groups that have more than one mirror node are not supported.

A very common example scenario would be Integration Server instances (one per data center) that require a shared (Active-Active replicated) database.



How to Switch to failover-priority CONSISTENCY:

Since the default setting is AVAILABIILITY, changing this setting must be performed explicitly on the following property:

```
<tc-config>
```

```
 <servers>
  ...
  <failover-priority>
   CONSISTENCY
  </failover-priority>
  ...
 </servers>
</tc-config>
```

Note: In the CONSISTENCY scenario, no mirror group can have more than two servers configured, otherwise the startup would fail.

The CONSISTENCY setting has no effect in a single node stripe but will take effect for the modified stripe once a new server is added.

How Failover Works in Various Scenarios

If the Active Node Fails ?

If the active node fails, the mirror node automatically stops processing, but all data is preserved and there are no lost transactions. At this point, the mirror node can only determine that the connection to the active node has been lost, but cannot determine whether the active node has failed, or whether there is just a break in the network connection. The mirror node sets its own status to WAITING_FOR_PROMOTION and waits for human interaction to determine why the connection has been lost.

If the Mirror Node Fails ?

If the mirror node fails, the active node continues operation without interruption. Human intervention is required to restart the mirror node.

If the Network Connection Fails ?

If the network connection fails, the active node cannot determine if the mirror node is still operating. However, the active node will continue without interruption. The mirror node cannot determine if the active node is still operating, so the mirror node will proceed as if the active node is not operating.

See above for details of how the mirror node reacts in this case.

What Happens to Transactions in Transit During Failover?

Provided the client is running, transactions will not be lost as they will be replayed to the new Active.

Monitoring Using Server Stats Script

The `server-stat.[sh/bat]` script packaged with this software is the ideal tool to monitor a cluster configured for Fail-Over tuning. The script delivers information regarding the states of the clustered servers. The following block shows part of the output of `server-stat.[sh/bat]` in a live 2-node cluster:

```
g1s1.health: OK
g1s1.role: ACTIVE
g1s1.state: ACTIVE-COORDINATOR
g1s1.port: 9540
g1s1.group name: group1
g1s2.health: OK
g1s2.role: PASSIVE
g1s2.state: PASSIVE-STANDBY
```

```
g1s2.port: 9640
g1s2.group name: group1
```

In the case of an Active server crashing, the staus deliverd by server-stat.[sh/bat] will deliver an output as below:

```
localhost.health: unknown
localhost.role: unknown
localhost.state: unknown
localhost.port: 9540
localhost.group name: unknown
localhost.error: Connection refused to localhost:9540. Is the TSA running?
g1s2.health: OK
g1s2.role: WAITING-FOR-PROMOTION
g1s2.state: PASSIVE-STANDBY
g1s2.port: 9640
g1s2.group name: group1
```

The "role" field of the Passive server indicates that is waiting for promotion.

Monitoring Using REST Endpoints

The server-stat utility internally uses a rest endpoint on the servers to fetch the information shown in the output. The same REST endpoint can be addressed directly to get the same information using the following URL:

```
<server:mgmt-port>/tc-management-api/v2/local/stat
```

Pointing to this URL to a Passive node that is waiting for promotion would give a response such as below:

```
{"health":"OK","role":"WAITING-FOR-PROMOTION","state":"PASSIVE-STANDBY",
"managementPort":"<port number>","serverGroupName":"group1","name":"g1s2"}
```

The "role" attribute indicates that the Passive server is waiting for promotion.

Monitoring Using TMC

In the case of a partition of Active and Passive, TMC will receive operator events indicating that the Passive is waiting for promotion.

However, in the case of failure of the Active node acting as active-coordinator of the cluster, TMC will be unable to deliver any useful information on the cluster.

In all other cases, TMC provides accurate operator events.

CONSISTENCY: How to Start Up and What to Consider

The command for starting up a server can be extended by the flag `--active`

```
$KIT/server/bin/start-server.sh[bat] -f /path/to/tc-config.xml -n <server-name> --active
```

- If this flag is set, then that node will run an election and if won, will become the Active.
- If an Active is found during the election, then this flag is ignored and the node will join the cluster as a Passive.
- If this flag is not set at all, then this node will look for an Active until such is found and, in the case of an Active responding, join as a Passive.

CONSISTENCY: The fail-over-action Command

When a Passive standby running with failover-priority "CONSISTENCY" detects that a node has left, then it will move to WAITING-FOR-PROMOTION state.

This state ...

- raises operator alerts with an operator alert appearing in the TMC as well as in log files

  Note: TMC may not be accessible should stripe 0 Active be out of operation. Consequentially, integration of the alert into 3rd party software or accessing logs may be required.

- initiates continuous logging
- waits for an external trigger

This trigger is provided by the fail-over-action command that must be performed by an operator.

```
$KIT/server/bin/fail-over-action.sh[bat] -f /path/to/tc-config.xml -n <server-name> --promote|--r
```

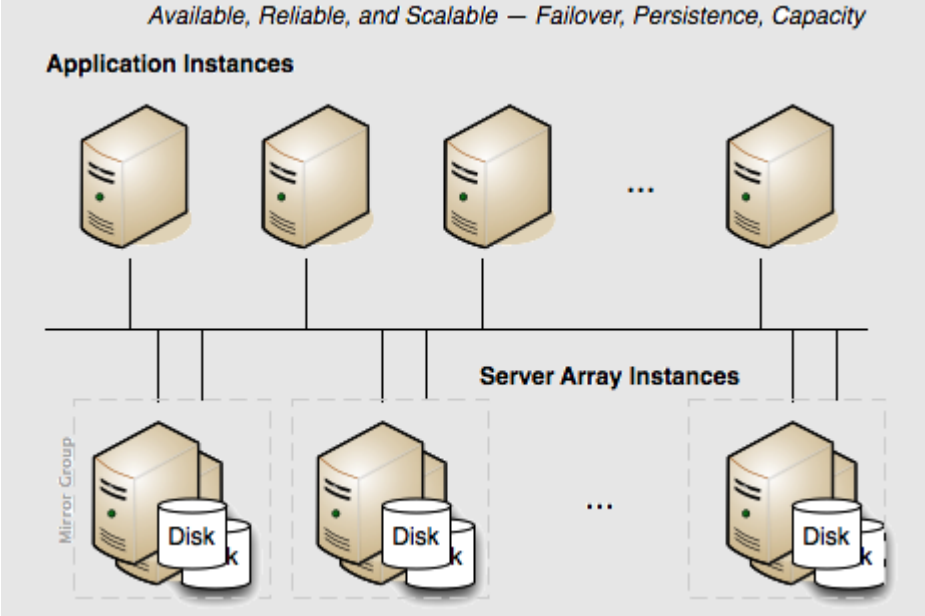
This command will call a REST endpoint to pass the fail-over action to the node specified by the `<server-name>`

Action	Description
`--promote`	The node will move to the ACTIVE_COORDINATOR state, provided the node was currently in the WAITING-FOR-PROMOTION state.
`--restart`	The node will log appropriately, shutdown, and mark the DB as dirty. The server will restart automatically.
`--failFast`	The node will log appropriately and shutdown without any changes to the database. The server will not restart automatically.

Scaling the Terracotta Server Array

Persistence: Yes | Failover: Yes | Scale: Yes

For capacity requirements that exceed the capabilities of a two-server active-mirror setup, expand the Terracotta cluster using a mirror-groups configuration. Using mirror groups with multiple coordinated active Terracotta server instances adds scalability to the Terracotta Server Array.



Mirror groups are specified in the `<servers>` section of the Terracotta configuration file. Mirror groups work by assigning group memberships to Terracotta server instances. The following snippet from a Terracotta configuration file shows a mirror-group configuration with four servers:

```
...
<servers>
  <mirror-group election-time="10" group-name="groupA">
    <server name="server1">
    ...
    </server>
    <server name="server2">
    ...
    </server>
  </mirror-group>
  <mirror-group election-time="15" group-name="groupB">
    <server name="server3">
    ...
    </server>
    <server name="server4">
    ...
    </server>
  </mirror-group>
  <restartable enabled="true"/>
</servers>
...
```

In this example, the cluster is configured to have two active servers, each with its own mirror. If server1 is elected active in groupA, server2 becomes its mirror. If server3 is elected active in groupB, server4 becomes its mirror. server1 and server3 automatically coordinate their work managing Terracotta clients and shared data across the cluster.

In a Terracotta cluster designed for multiple active Terracotta server instances, the server instances in each mirror group participate in an election to choose the active. Once every mirror group has elected an active server instance, all the active server instances in the cluster begin cooperatively managing the cluster. The rest of the server instances become

mirrors for the active server instance in their mirror group. If the active in a mirror group fails, a new election takes place to determine that mirror group's new active. Clients continue work without regard to the failure.

Note: Server vs. Mirror Group - Under `<servers>`, you may use either `<server>` or `<mirror-group>` configurations, but not both. All `<server>` configurations directly under `<servers>` work together as one mirror group, with one active server and the rest mirrors. To create more than one stripe, use `<mirror-group>` configurations directly under `<servers>`. The mirror group configurations then include one or more `<server>` configurations.

In a Terracotta cluster with mirror groups, each group, or "stripe", behaves in a similar way to an active-mirror setup (see "Terracotta Cluster with High Availability " on page 20. For example, when a server instance is started in a stripe while an active server instance is present, it synchronizes state from the active server instance before becoming a mirror. A mirror cannot become an active server instance during a failure until it is fully synchronized. If an active server instance running in restartable mode goes down, and a mirror takes over, the data directory is cleared before bringing back the crashed server.

Election Time

The `<mirror-group>` configuration allows you to declare the election time window. An active server is elected from the servers that cast a vote within this window. The value is specified in seconds and the default is 5 seconds. Network latency and the work load of the servers should be taken into consideration when choosing an appropriate window.

In the above example, the servers in groupA can take up to 10 seconds to elect an active server, and the servers in groupB can take up to 15 seconds.

Stripe and Cluster Failure

If the active server in a mirror group fails or is taken down, the cluster stops until a mirror takes over and becomes active (ACTIVE-COORDINATOR status).

However, the cluster cannot survive the loss of an entire stripe. If an entire stripe fails and no server in the failed mirror-group becomes active within the allowed window (based on the election-time setting), the entire cluster must be restarted.

3 Configuring the Terracotta Server Array

- About Terracotta Server Configuration 34
- How Terracotta Servers Get Configured 34
- How Terracotta Clients Get Configured 36
- Configuration in a Development Environment 38
- Configuration in a Production Environment 40
- Binding Ports to Interfaces 42
- Which Configuration? 43

About Terracotta Server Configuration

Terracotta XML configuration files set the characteristics and behavior of Terracotta server instances and Terracotta clients. The easiest way to create your own Terracotta configuration file is by editing a copy of one of the sample configuration files available with the Terracotta BigMemory Max kit.

Where you locate the Terracotta configuration file, or how your Terracotta server and client configurations are loaded, depends on the stage your project is at and on its architecture. This document covers the following cases:

- Development stage, 1 Terracotta server
- Development stage, 2 Terracotta servers
- Deployment stage

This document discusses cluster configuration in the Terracotta Server Array. To learn more about the Terracotta server instances, see "Terracotta Server Array Architecture" on page 13.

For a comprehensive and fully annotated configuration file, see config-samples/tc-config-reference.xml in the Terracotta kit.

How Terracotta Servers Get Configured

To configure the Terracotta server, create a tc-config.xml configuration file, or update the one that is provided in the config-samples/ directory of the BigMemory Max kit. For example:

```
<?xml version="1.0" encoding="UTF-8" ?>
<tc:tc-config xmlns:tc="http://www.terracotta.org/config"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.terracotta.org/schema/terracotta-9.xsd">
  <servers>
    <server host="localhost" name="My Server Name1">
      <!-- Specify the path where the server should store its data. -->
      <data>/local/disk/path/to/terracotta/server1-data</data>
       <!-- Specify the port where the server should listen for client
       traffic. -->
       <tsa-port>9510</tsa-port>
       <jmx-port>9520</jmx-port>
       <tsa-group-port>9530</tsa-group-port>
       <management-port>9540</management-port>
       <!-- Enable BigMemory on the server. -->
       <dataStorage size="800g">
          <offheap size="200g"/>
          <!-- Hybrid storage is optional. -->
          <hybrid/>
       </dataStorage>
     </server>
    <server host="localhost" name="My Server Name2">
      <data>/local/disk/path/to/terracotta/server2-data</data>
       <tsa-port>9510</tsa-port>
```

```
          <jmx-port>9520</jmx-port>
          <tsa-group-port>9530</tsa-group-port>
          <management-port>9540</management-port>
          <dataStorage size="200g">
             <offheap size="200g"/>
          </dataStorage>
        </server>
       <!-- Add the restartable element for Fast Restartability (optional). -->
       <restartable enabled="true"/>
     </servers>
     <clients>
       <logs>logs-%i</logs>
     </clients>
   </tc:tc-config>
```

To successfully configure a Terracotta Server Array using the Terracotta configuration file, note the following:

- Two or more servers should be defined in the <servers> section of the Terracotta configuration file.
- <tsa-port> is the port that the Terracotta server listens to for client traffic.
- <jmx-port> is the port that the Terracotta server's JMX Connector listens to.

  Note: Listening on the <jmx-port> is deprecated. Alternatively, use the monitoring features provided by the Terracotta Management Console (see the *Terracotta Management Console User Guide*) and the WAN Replication Service (see the *WAN Replication User Guide*).

  <jmx-port> is disabled by default. If you want to enable it, add `jmx-enabled="true"` in the <server> elements. For example:

  ```
  <server host="localhost" name="My Server Name1" jmx-enabled="true">
  ```

- <tsa-group-port> is the port used by the Terracotta server to communicate with other Terracotta servers.
- <management-port> is the port that the Terracotta Management Console (TMC) uses.
- Under <servers>, use either <server> or <mirror-group> configurations, but not a mixture. You may configure multiple servers or multiple mirror groups. <server> instances under <servers> work together as a mirror group. To create more than one stripe, use <mirror-group> instances.
- Terracotta server instances must not share data directories. Each server's <data> element should point to a different and preferably local data directory.
- For data persistence, configure fast restartability. Enabling <restartable> means that the shared in-memory data is backed up and, in case of failure, it is automatically restored. Setting <restartable> to "false" or omitting the <restartable> element are two ways to configure no persistence.
- Each server requires an off-heap store, which allows all data to be stored in-memory, limited only by the amount of memory in your server. The minimum <offheap> size is 4 GB. Additional hybrid storage is optional. Specify the <dataStorage> size and the

<offheap> size. The <offheap> size can be set to the amount of memory available in your server for data. If you enable <hybrid>, then the <dataStorage> size can exceed the <offheap> size.

- All servers and clients should be running the same version of Terracotta and Java.

Note: For more information about the Terracotta configuration file, see "Terracotta Configuration Parameters" on page 101.

Server Startup Behavior

At startup, Terracotta servers load their configuration from one of the following sources:

- A default configuration included with the Terracotta kit
- A local or remote XML file

These sources are explored below.

Default Configuration

If no configuration file is specified *and* no tc-config.xml exists in the directory in which the Terracotta instance is started, then default configuration values are used.

Local XML File (Default)

The file tc-config.xml is used by default if it is located in the directory in which a Terracotta instance is started *and* no configuration file is explicitly specified.

Local or Remote Configuration File

You can explicitly specify a configuration file by passing the -f option to the script used to start a Terracotta server. For example, to start a Terracotta server on UNIX/Linux using the provided script, enter:

```
start-tc-server.sh -f <path_to_configuration_file>
```

where <*path_to_configuration_file*> can be a URL or a relative directory path. In Microsoft Windows, use start-tc-server.bat.

Note: Cygwin (on Windows) is not supported for this feature.

How Terracotta Clients Get Configured

At startup, Terracotta clients load their configuration from one of the following sources:

- "Local or Remote XML File" on page 37
- "Terracotta Server" on page 37
- An Ehcache configuration file (using the "<terracottaConfig> element" on page 41) used with BigMemory Max and BigMemory Go.

- A Quartz properties file (using the `org.quartz.jobStore.tcConfigUrl` property) used with Quartz Scheduler.
- A Filter (in web.xml) element used with containers and Terracotta Web Sessions.
- The client constructor (TerracottaClient()) used when a client is instantiated programmatically.

Terracotta clients can load customized configuration files to specify <client> and <application> configuration. However, the <servers> block of every client in a cluster must match the <servers> block of the servers in the cluster. If there is a mismatch, the client will emit an error and fail to complete its startup. However, there are options you can set for server settings to override client settings. For details, see "How Server Settings Can Override Client Settings" in the *BigMemory Max Configuration Guide*.

Note: Error with Matching Configuration Files - On startup, a Terracotta client may emit a configuration-mismatch error if its <servers> block does not match that of the server it connects to. However, under certain circumstances, this error may occur even if the <servers> blocks appear to match.

The following suggestions may help prevent this error:

- Use `-Djava.net.preferIPv4Stack` consistently. If it is explicitly set on the client, be sure to explicitly set it on the server.
- Ensure the etc/hosts file does not contain multiple entries for hosts running Terracotta servers.
- Ensure that DNS always returns the same address for hosts running Terracotta servers.

Local or Remote XML File

See the discussion for local XML file (default) in "How Terracotta Servers Get Configured" on page 34.

To specify a configuration file for a Terracotta client, see "Clients in Development " on page 40.

Note: Fetching Configuration from the Server - On startup, Terracotta clients must fetch certain configuration properties from a Terracotta server. A client loading its own configuration will attempt to connect to the Terracotta servers named in that configuration. If none of the servers named in that configuration are available, the client cannot complete its startup.

Terracotta Server

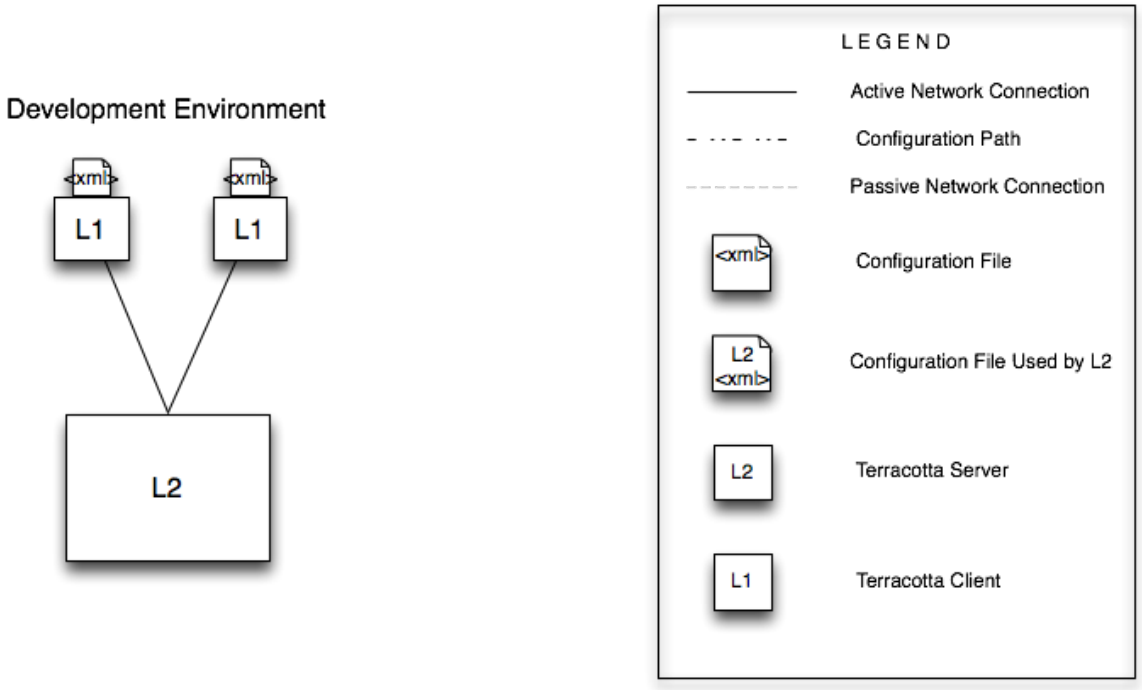
Terracotta clients can load configuration from an active Terracotta server by specifying its hostname and TSA port (see "Clients in Production" on page 41).

Configuration in a Development Environment

In a development environment, using a different configuration file for each Terracotta client facilitates the testing and tuning of configuration options. This is an efficient and effective way to gain valuable insight on best practices for clustering your application with Terracotta.

One-Server Setup in Development

For one Terracotta server, the default configuration is adequate.



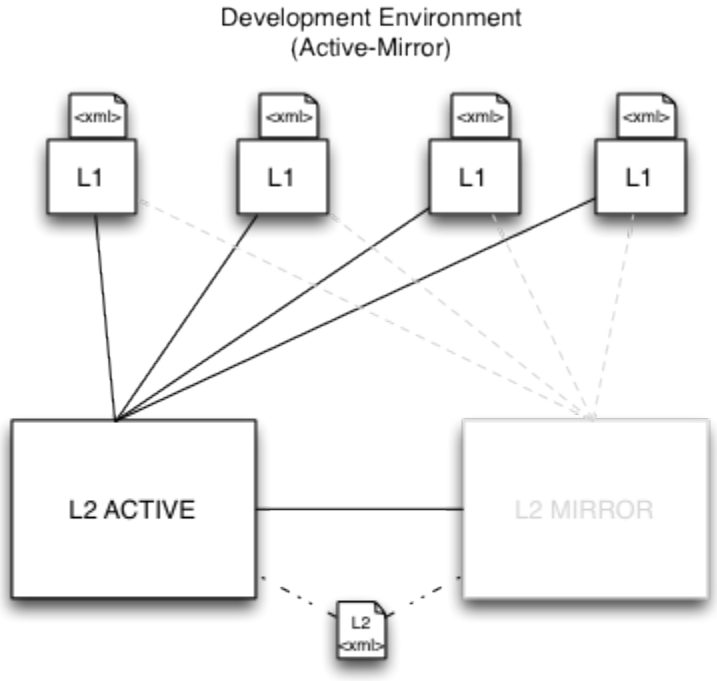
To use the default configuration settings, start your Terracotta server using the start-tc-server.sh (or start-tc-server.bat) script in a directory that does *not* contain the file tc-config.xml :

```
[PROMPT] ${TERRACOTTA_HOME}\bin\start-tc-server.sh
```

To specify a configuration file, use one of the approaches discussed in "How Terracotta Servers Get Configured" on page 34.

Two-Server Setup in Development

A two-server setup, sometimes referred to as an active-mirror setup, has one active server instance and one "hot standby" (the mirror) that should load the same configuration file.



The configuration file loaded by the Terracotta servers must define each server separately using <server> elements. For example:

```
<tc:tc-config xsi:schemaLocation="http://www.terracotta.org/schema/terracotta-9.xsd"
xmlns:tc="http://www.terracotta.org/config"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
...
<!-- Use an IP address or a resolvable host name for the host attribute. -->
  <server host="123.456.7.890" name="Server1">
...
  <server host="myResolvableHostName" name="Server2">
...
</tc:tc-config>
```

Assuming Server1 is the active server, using the same configuration allows Server2 to be the mirror and maintain the environment in case of failover. When running both servers on the same host, the <tsa-port> for each server should be different. The other ports are automatically generated from the <tsa-port>. Having a separate log location is a good idea, but failing to do so will not prevent the servers from starting. If the servers are set up to be restartable, setting the data directory to a different location would be a requirement.

Server Names for Startup

With multiple <server> elements, the name attribute may be required to avoid ambiguity when starting a server:

```
start-tc-server.sh -n Server1 -f <path_to_configuration_file>
```

In Microsoft Windows, use start-tc-server.bat.

For example, if you are running Terracotta server instances on the same host, you must specify the name attribute to set an unambiguous target for the script.

However, if you are starting Terracotta server instances in an unambiguous setup, specifying the server name is optional. For example, if the Terracotta configuration file specifies different IP addresses for each server, the script assumes that the server with the IP address corresponding to the local IP address is the target.

Clients in Development

You can explicitly specify a client's Terracotta configuration file by passing `-Dtc.config=path/to/my-tc-config.xml` when you start your application with the Terracotta client.

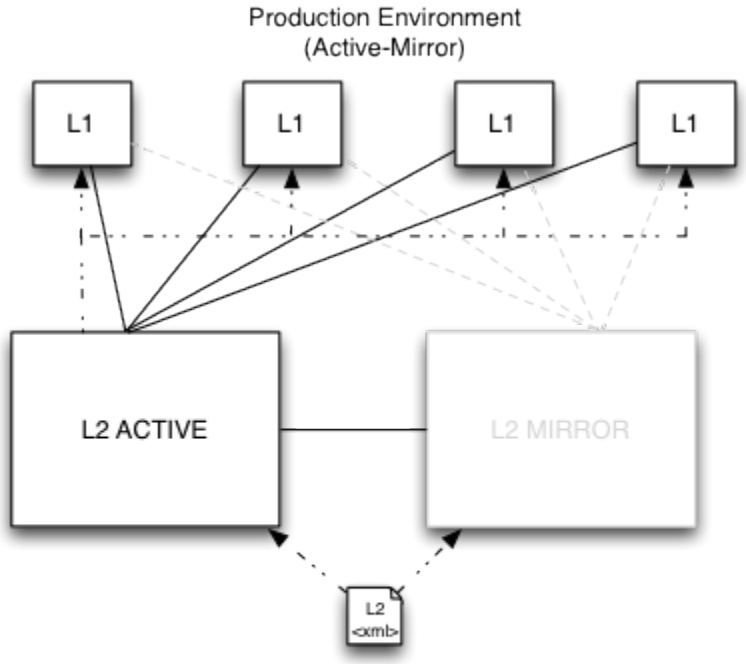
```
-Dtc.config=path/to/my-tc-config.xml -cp classes myApp.class.Main
```

where `myApp.class.Main` is the class used to launch the application you want to cluster with Terracotta.

If tc-config.xml exists in the directory in which you run Java, it can be loaded without `-Dtc.config`.

Configuration in a Production Environment

For an efficient production environment, it's recommended that you maintain one Terracotta configuration file. That file can be loaded by the Terracotta server (or servers) and pushed out to clients. While this is an optional approach, it's an effective way to centralize and decrease maintenance.



If your Terracotta configuration file uses "%i" for the hostname attribute in its server element, change it to the actual hostname in production. For example, if in development you used the following:

```
<server host="%i" name="Server1">
```

and the production host's hostname is myHostName, then change the host attribute to the myHostName:

```
 <server host="myHostName" name="Server1">
```

Clients in Production

For clients in production, you can set up the Terracotta environment before launching your application.

Setting Up the Terracotta Environment

To start your application with the Terracotta client using your own scripts, first set the following environment variables:

```
TC_INSTALL_DIR=<path_to_local_Terracotta_home>
TC_CONFIG_PATH=<path/to/tc-config.xml>
```

or

```
TC_CONFIG_PATH=<server_host>:<tsa-port>
```

where `<server_host>:<tsa-port>` points to the running Terracotta server. The specified Terracotta server will push its configuration to the Terracotta client.

Alternatively, a client can specify that its configuration come from a server by setting the `tc.config` system property:

```
-Dtc.config=serverHost:tsaPort
```

If more than one Terracotta server is available, enter them in a comma-separated list:

```
TC_CONFIG_PATH=<server_host1>:<tsa-port>,<server_host2>:<tsa-port>
```

If <server_host1> is unavailable, <server_host2> is used.

Terracotta Products

Terracotta products can set a configuration path using their own configuration files.

For BigMemory Max and BigMemory Go, use the `<terracottaConfig>` element in the Ehcache configuration file (ehcache.xml by default):

```
 <terracottaConfig url="localhost:9510" />
```

For Quartz, use the `org.quartz.jobStore.tcConfigUrl` property in the Quartz properties file (quartz.properties by default):

```
 org.quartz.jobStore.tcConfigUrl = /myPath/to/tc-config.xml
```

For Terracotta Web Sessions, use the appropriate elements in web.xml or context.xml (see the *Web Sessions User Guide*).

Binding Ports to Interfaces

Normally, the ports you specify for a server in the Terracotta configuration are bound to the interface associated with the host specified for that server. For example, if the server is configured with the IP address "12.345.678.8" (or a hostname with that address), the server's ports are bound to that same interface:

```
<server host="12.345.678.8" name="Server1">
 ...
 <tsa-port>9510</tsa-port>
 <jmx-port>9520</jmx-port>
 <tsa-group-port>9530</tsa-group-port>
 <management-port>9540</management-port>
</server>
```

However, in certain situations it may be necessary to specify a different interface for one or more of a server's ports. This is done using the `bind` attribute, which allows you bind a port to a different interface. For example, a JMX client may only be able connect to a certain interface on a host. The following configuration shows a JMX port bound to an interface different than the host's:

```
<server host="12.345.678.8" name="Server1">
 ...
 <tsa-port>9510</tsa-port>
 <jmx-port bind="12.345.678.9">9520</jmx-port>
 <tsa-group-port>9530</tsa-group-port>
 <management-port>9540</management-port>
</server>
```

Which Configuration?

Each server and client must maintain separate log directories. By default, server logs are written to %(user.home)/terracotta/server-logs and client logs to %(user.home)/terracotta/client-logs.

To find out which configuration a server or client is using, search its logs for an INFO message containing the text "Configuration loaded from".

4 Automatic Resource Management

- What is Automatic Resource Management? 46
- Eviction 46
- Customizing the Eviction Strategy 50

What is Automatic Resource Management?

Terracotta Server Array resource management involves self-monitoring and polling to determine the real-time size of the data set and assess the amount of memory remaining according to user-configured limitations. In-memory data can be managed from three directions:

1. **Time** - TTI/TTL settings can be configured to expire entries that will then be evicted by the new TSA eviction implementation. You can also configure caches so that their entries are eternal, or you can pin entries or caches so that they are never evicted.
2. **Size** - The total amount of BigMemory managed by the TSA can be configured using the `dataStorage` and `offheap` elements in the tc-config.xml file.
3. **Count** - The total number of entries per cache can be configured using the maxEntriesInCache attribute in the ehcache.xml file.

Eviction

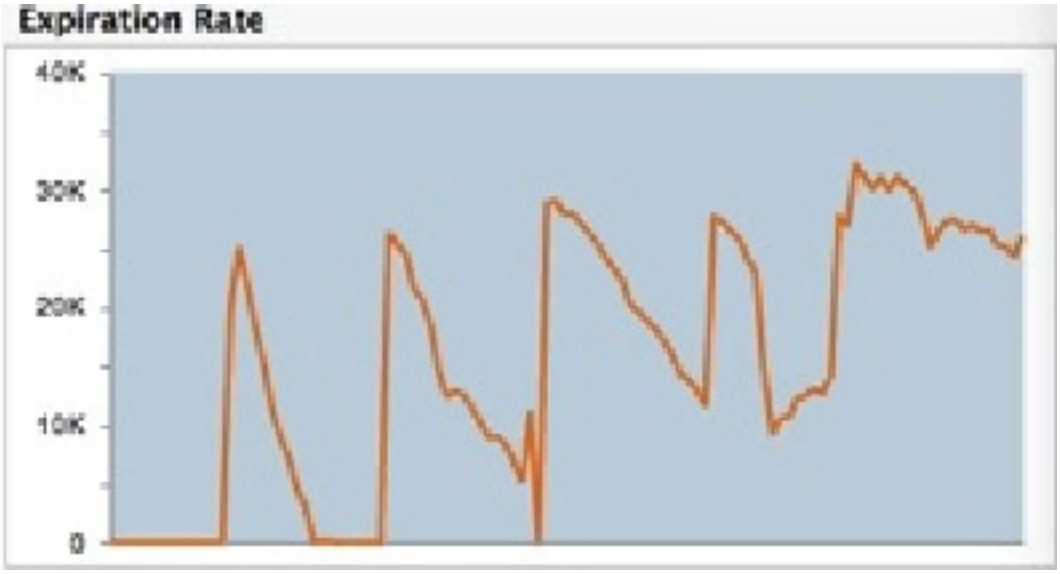
All data is kept in memory, and the TSA runs evictions in the background to keep the data set within its limitations. Eviction of entries from the data set reduces the amount of data before the memory becomes full. The criteria for an entry to be eligible for eviction are:

- It is not on a Terracotta client (L1).
- It is not pinned to a Terracotta server.
- It is held in a cache backed by a System of Record (SOR).

Note: Store vs. Cache - BigMemory's in-memory data is treated as a "store" when BigMemory owns the data, and as a "cache" when the data also resides in a System of Record (SOR). Generally, data that is created by BigMemory and run-time data created by your application are examples of data that is treated as a store. The TSA does not evict data stores because they are the only or primary records. The TSA can evict cached data because that data is backed up in an SOR. Distinctions in data structures are handled automatically.

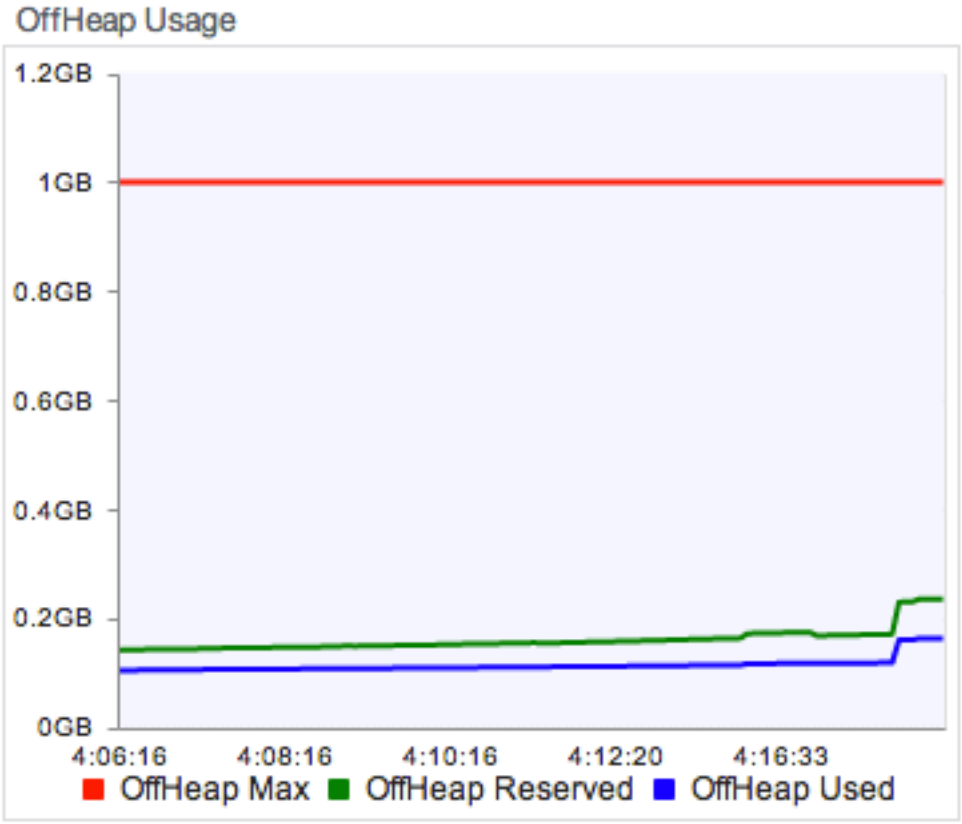
Eviction is done by the following evictors, which work together in the background:

1. The periodic evictor is activated on an as-needed basis. It removes expired entries based on TTI/TTL settings. The Server Expiration Rate graph in the TMC shows the activity of the periodic evictor.



2. The resource-based evictor is activated by the periodic TTI/TTL eviction scheduler, as well as by resource monitoring events. This evictor continuously polls BigMemory stores to check current resource usage. At approximately 10% usage of the disk as well as `dataStorage` and `offheap` sizes (configured in the tc-config.xml file), it starts looking for TTI/TTL-expired elements to evict. At approximately 80% usage, it evicts live as well as expired elements. If a monitored resource goes over its critical threshold, this evictor will work continually until the monitored resource falls below the critical threshold.

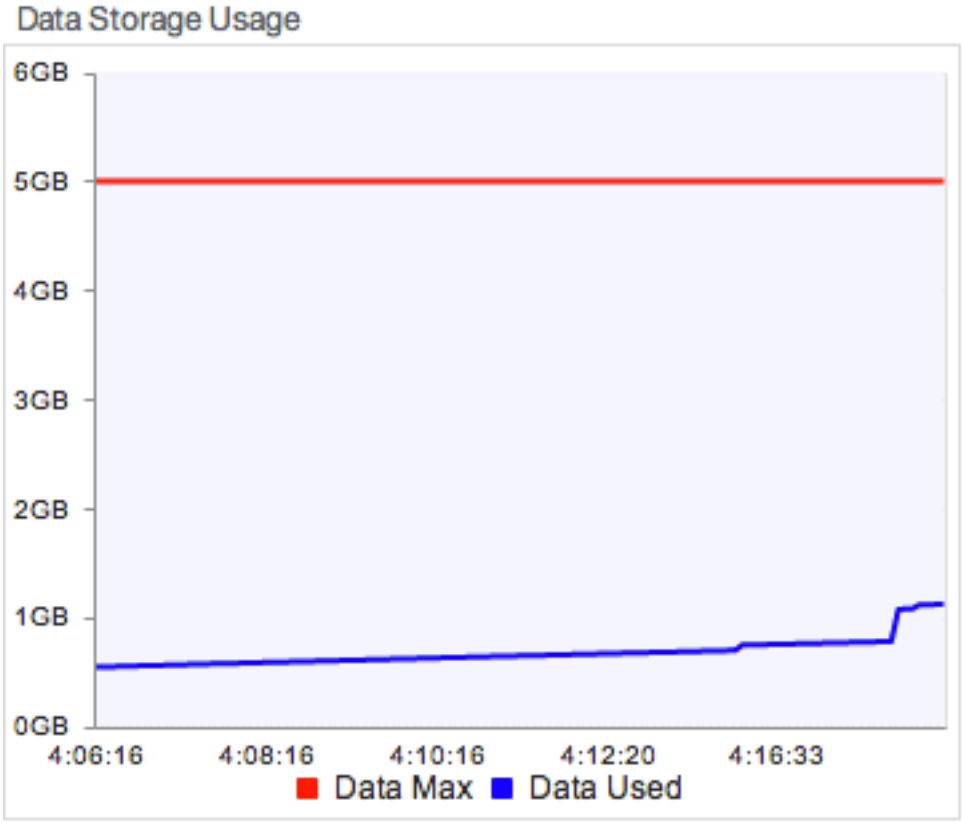
   This evictor monitors two off-heap thresholds -- used space and reserved space. Resource eviction is triggered if either the reserved or used space is above its threshold. Once resource eviction has started, both used and reserved spaces must fall below their respective thresholds before resource eviction ends.



The Offheap Usage graph in the TMC provides the following information:

- Off-heap Max is the configured `offheap` size
- Off-heap Reserved represents usage of the space that is reserved for the system
- Off-heap Used represents the amount of off-heap BigMemory that is in use

The Data Storage Usage graph in the TMC shows the configured `dataStorage` size and the amount in use. This usage includes off-heap and hybrid storage combined. (BigMemory Hybrid allows for a mix of a solid-state drive (SSD) with DRAM-based offheap storage.)



3. The capacity-based evictor is activated when a cache goes over its maximum count (as configured with `maxEntriesInCache`), plus an overshoot count, and it attempts to bring the size of the cache to the max capacity. The `maxEntriesInCache` attribute must be present in the Ehcache configuration (do not include `maxEntriesInCache` in your configuration if you do not want the capacity evictor to run). If `maxEntriesInCache` is not set, it gets the default value 0, which means that the cache is unbounded and will not undergo capacity eviction (but periodic and resource evictions are still allowed).

   The Server Eviction Rate graph in the TMC shows the activity of the resource and capacity evictors.

Customizing the Eviction Strategy

Based upon the three types of evictors, there are three strategies that you can employ for controlling the size of the TSA's data set:

1. Set the Time To Idle (TTI) or Time to Live (TTL) options for any entry in your data set. After the time has expired, the periodic evictor will clear the entry.
2. Set the `dataStorage` and `offheap` sizes to control how much BigMemory should be used before the resource-based evictor is activated.
3. Set the `maxEntriesInCache` attribute to control when the capacity-based evictor is activated.

5 Managing Near-Memory-Full Conditions

- Behavior of the TSA under Near-Memory-Full Conditions 52
- Restricted Mode Operations 53
- Recovery 53

Behavior of the TSA under Near-Memory-Full Conditions

In a near-memory-full condition, where evictions are not happening fast enough to keep the data set within its BigMemory size limitations, the TSA will put a throttle on operations for a temporary period while it attempts to automatically recover in the background. If unable to recover, the TSA will move into "restricted mode" to prevent out-of-memory errors. The Terracotta Management Console (TMC) uses events to report when the TSA enters restricted mode and allows you to execute additional recovery measures. See "Monitoring Cluster Events" on page 55.

Summary of TSA behavior in near-memory-full conditions:

- If usage reaches its critical threshold, T1, then it enters "throttle mode," where writes are slowed while the TSA attempts to evict eligible cache entries in order to bring memory usage within the configured range.
- If usage reaches its halt threshold, T2, then it enters "restricted mode," where writes are blocked, an exception is thrown, and operator intervention is needed to reduce memory usage.
- When usage falls below T1, then the TSA returns to normal operation.

	Throttle mode	Restricted mode
Entered when	Disk, `dataStorage` , and/or `offheap` usage crosses its critical threshold	Disk, `dataStorage` , and/or `offheap` usage crosses its halt threshold
Operator event	"TPS seems really low; marking us as being throttled'	"We're in restricted mode; waiting a while and retrying"
Data access	Modifications to in-memory data are slowed	Modifications to in-memory data are blocked
Allowed operations	All cache operations still allowed	Only gets, removes, and config changes are allowed
Actions	Evictions continue automatically in the background	Operator intervention required to make additional evictions
State of the data	Evictable data is still present	No more data present in memory that can be evicted

Throttle mode		Restricted mode
		by the evictor (all caches are pinned)
Recovery	Automatic	From the TMC or programmatically, clear caches and/or remove entries from a data set
Back to normal operation	As soon as the background evictions have time to catch up and reduce the data set to within its limitations	After user intervention clears space, the TSA will automatically continue with normal operation

Restricted Mode Operations

If the TSA is temporarily under restricted mode, any change to the data set which may result in increased resource utilization is not allowed, including all put and replace methods. Restricted mode does allow gets, removes, configuration changes, and other operations.

Recovery

Recovery from throttle mode is automatic, as soon as the background evictions have time to reduce the data set to within its limitations.

If the TSA enters restricted mode, operator events will be logged in the TMC, and user or programmatic intervention is necessary. In the TMC, you can initiate actions to manually reduce the data set. You can also anticipate operator events and use programmatic logic to respond appropriately.

The following actions are recommended for reducing the data set:

- Clear caches (from the TMC or programmatically)
- Remove entries from data sets programmatically

Note: Because eviction in restricted mode is resource-driven, changing TTI/TTL or maximum capacity will not move the TSA out of restricted mode.

To clear caches from the TMC, click the **Application Data** tab and the **Management** sub-tab. Each cache will have a clickable option to **Clear Cache**. Note that caution should be used when considering whether to clear a pinned cache.

6 Monitoring Cluster Events

- About Cluster Events 56
- Event Types and Definitions 56

About Cluster Events

Cluster events report topology changes, performance issues, and errors in operations. These events are logged by both Terracotta server (L2) and client (L1), and can also be viewed in the TMC (see the *Terracotta Management Console User's Guide*).

By default, the L2 stores a maximum of 100 events in memory, and this is the number pulled by the TMC. To edit that number, use the Terracotta property `l2.operator.events.store`. To set the property in the Terracotta configuration file, use:

```
<tc-properties>
  ...
  <property name="l2.operator.events.store" value="500" />
</tc-properties>
```

Event Types and Definitions

This section describes the types of events that can be found in logs or viewed in the TMC.

- **memory.longgc (Memory Manager)**
  - Level: WARN
  - Cause: A full garbage collection (GC) longer than the configured threshold has occurred.
  - Action: Reduce cache memory footprint in L1 (Terracotta client). Investigate issues with application logic and garbage creation.
  - Notes: The default critical threshold is 8 seconds, but it can be reconfigured in tc.properties using `longgc.threshhold`. For information about setting tc.properties, see the "Terracotta Configuration Parameters" on page 101.

    Occurrence of this event could help diagnose certain failures. For details, see "Configuring the HealthChecker Properties" in the *BigMemory Max High-Availability Guide*.
- **dgc.periodic.started (DGC)**
  - Level: INFO
  - Cause: Periodic distributed garbage collection (DGC), which was explicitly enabled in the configuration, has started a cleanup cycle.
  - Action: If periodic DGC is unneeded, disable it to improve overall cluster performance.
  - Notes: Periodic DGC, which is disabled by default, is mostly useful in the absence of automatic handling of distributed garbage.

- **dgc.periodic.finished (DGC)**
  - Level: INFO
  - Cause: Periodic DGC, which was explicitly enabled in the configuration, ended a cleanup cycle.
  - Action: If periodic DGC is unneeded, disable it to improve overall cluster performance.
  - Notes: Event message reads "DGC[{0}] finished. Begin Count : {1} Collected : {2} Time Taken : {3} ms Live Objects : {4}".
- **dgc.periodic.canceled (DGC)**
  - Level: INFO
  - Cause: Periodic DGC, wPeriodichich was explicitly enabled in the configuration, has been cancelled due to an interruption (for example, by a failover operation).
  - Action: If periodic DGC is unneeded, disable it to improve overall cluster performance.
  - Notes: Periodic DGC, which is disabled by default, is mostly useful in the absence of automatic handling of distributed garbage.
- **dgc.inline.cleanup.started (DGC)**
  - Level: INFO
  - Cause: L2 (Terracotta server) is starting up as ACTIVE with existing data, triggering inline distributed garbage collection (DGC).
  - Action: No action necessary.
  - Notes: Only seen when a server starts up as ACTIVE upon a recovery, using Fast Restartability.
- **dgc.inline.cleanup.finished (DGC)**
  - Level: INFO
  - Cause: Inline DGC operation completed.
  - Action: No action necessary.
  - Notes: Event message reads "Inline DGC [{0}] reference cleanup finished. Begin Count : {1} Collected : {2} Time Taken : {3} ms Live Objects : {4}".
- **dgc.inline.cleanup.canceled (DGC)**
  - Level: INFO
  - Cause: Inline DGC operation interrupted.
  - Action: Investigate any unusual cluster behavior or other events.
  - Notes: Possibly occurs during failover, but other events should indicate real cause.

- **topology.node.joined (Cluster Topology)**
  - Level: INFO
  - Cause: Specified node has joined the cluster.
  - Action: No action necessary.
  - Notes: None.
- **topology.node.left (Cluster Topology)**
  - Level: WARN
  - Cause: Specified node has left the cluster.
  - Action: Check why the node has left (for example: long GC, network issues, or issues with local node resources).
  - Notes: None.
- **topology.node.state (Cluster Topology)**
  - Level: INFO
  - Cause: L2 changing state (for example, from INITIALIZING to ACTIVE).
  - Action: Check to see that the state change is expected.
  - Notes: Event message reads "Moved to {0}", where {0} is the new state.
- **topology.handshake.reject (Cluster Topology)**
  - Level: ERROR
  - Cause: L1 is unsuccessfully trying to reconnect to cluster, but it has already been expelled.
  - Action: If the L1 does not go into a rejoin operation, it must be restarted manually.
  - Notes: Event message reads "An {0} client {1} tried to connect to {2} server. Connection refused!!"
- **topology.active.left (Cluster Topology)**
  - Level: WARN
  - Cause: Active server left the cluster.
  - Action: Check why the active L2 has left.
  - Notes: None.
- **topology.mirror.left (Cluster Topology)**
  - Level: WARN
  - Cause: Mirror server left the cluster.
  - Action: Check why the mirror L2 has left.

- Notes: None.

- **topology.zap.received (Cluster Topology)**
  - Level: CRITICAL
  - Cause: One L2 is trying to cause another L2 to restart ("zap").
  - Action: Investigate a possible "split brain" situation (a mirror L2 behaves as the ACTIVE) if the zapped L2 does not obey the restart order.
  - Notes: A "zap" operation happens only within a mirror group. Event message reads "SPLIT BRAIN, {0} and {1} are ACTIVE", where {0} and {1} are the two servers vying for the ACTIVE role.
- **topology.zap.accepted (Cluster Topology)**
  - Level: CRITICAL
  - Cause: The L2 is accepting the order to restart ("zap" order).
  - Action: Check the state of the zapped L2 to ensure that it restarts as a mirror, or manually restart it.
  - Notes: A "zap" order is issued only within a mirror group. Event message reads "{0} has more clients. Exiting!!", where {0} is the L2 that becomes the ACTIVE.
- **topology.db.dirty (Cluster Topology)**
  - Level: WARN
  - Cause: A mirror L2 is trying to join with data in place.
  - Action: If the mirror does not automatically restart and wipe its data, its data may need to be manually wiped and before it is restarted.
  - Notes: Restarted mirror L2s must wipe their data to resync with the active L2. This is normally an automatic operation that should not require action. Event message reads "Started with dirty database. Exiting!! Restart {0}", where {0} is the the mirror that is automatically restarting.
- **topology.config.reloaded (Cluster Topology)**
  - Level: INFO
  - Cause: Cluster configuration was reloaded.
  - Action: No action necessary.
  - Notes: None.
- **dcv2.servermap.eviction (DCV2)**
  - Level: INFO
  - Cause: Automatic evictions for optimizing Terracotta Server Array operations.
  - Action: No action necessary.

- Notes: Event message reads "DCV2 Eviction - Time taken (msecs)={0}, Number of entries evicted={1}, Number of segments over threshold={2}, Total Overshoot={3}".

- **system.time.different (System Setup)**
  - Level: WARN
  - Cause: System clocks are not aligned.
  - Action: Synchronize system clocks.
  - Notes: The default tolerance is 30 seconds, but it can be reconfigured in tc.properties using `time.sync.threshold`. For information about setting tc.properties, see "Terracotta Configuration Parameters" on page 101.

    Note that overly large tolerance can introduce unpredictable errors and behaviors.

- **resource.capacity.near (Resource)**
  - Level: WARN
  - Cause: L2 entered throttled mode, which could be a temporary condition (e.g., caused by bulk-loading) or could indicate insufficient allocation of memory.
  - Action: See "Managing Near-Memory-Full Conditions" on page 51.
  - Notes: After emitting this, L2 can emit `resource.capacityrestored` (return to normal mode) or `resource.fullcapacity` (move to restricted mode), based on resource availability. Event message reads "{0} is nearing capacity limit, performance may be degraded - {1}% usage", where {0} is the L2 identification and {1} is the % usage of the memory resources allocated to that L2.

- **resource.capacity.full (Resource)**
  - Level: ERROR
  - Cause: L2 entered restricted mode, which could be a temporary condition (e.g., caused by bulk-loading) or could indicate insufficient allocation of memory.
  - Action: See "Managing Near-Memory-Full Conditions" on page 51.
  - Notes: After emitting this, L2 can emit `resource.capacityrestored` (return to normal mode), based on resource availability. Event message reads "{0} is at over capacity limit, no further additive operations will be accepted - {1}% usage", where {0} is the L2 identification and {1} is the % usage of the memory resources allocated to that L2.

- **resource.capacity.restored (Resource)**
  - Level: INFO
  - Cause: L2 returned to normal from throttled or restricted mode.
  - Action: No action necessary.

- Notes: Event message reads "{0} capacity has been restored, performance has returned to normal - {1}% usage", where {0} is the L2 identification and {1} is the % usage of the memory resources allocated to that L2.

7 Backing Up Live In-Memory Data

- About Live Backup 64
- Creating a Backup 64
- The Backup Directory 64
- Restoring Data from a Backup 65

About Live Backup

Backups of the entire data set across all stripes (mirror groups) of the Terracotta Server Array can be made using the TMC Backup feature. This feature creates a time-stamped backup of each stripe's data, providing a snapshot of the TSA's in-memory data.

The Backup feature is available when fast restartability (FRS) is enabled for the TSA (`<restartable enabled="true"/>` in the tc-config.xml).

Creating a Backup

From the TMC, select the **Administration** tab and the **Backups** sub-tab. Click the **Make Backup** button to perform a backup. The TMC sends a backup request to all stripes in the cluster.

In order to capture a consistent snapshot of the in-memory data, the backup function creates a pause in transactions, allowing any unfinished transactions to complete, and then the backup is written. This allows the backup to be a consistent record of the entries in-memory, as well as search and other indices.

Note that when backing up a cluster, each stripe is backed up independently and at a slightly different time than the other stripes.

When complete, a window appears that confirms the backup was taken and provides the time-stamped file name(s) of the backup.

The Backup Directory

Backups are saved to the default directory data-backup, unless otherwise configured in the tc-config.xml. Terracotta automatically creates data-backup in the directory containing the Terracotta server's configuration file (tc-config.xml by default).

You can override the default directory by specifying a different backup directory in the server's configuration file using the `<data-backup>` property:

```
<servers>
   <server name="Server1">
      <data>/opt/terracotta/server1-data</data>
      <data-backup>path/to/my/backup/directory</data-backup>
      <offheap>
         <enabled>true</enabled>
         <maxDataSize>2g</maxDataSize>
      </offheap>
   </server>
   <restartable enabled="true"/>
</servers>
```

Restoring Data from a Backup

If the TSA fails, on restart it automatically restores data from its data directory, recreating the application state. If the current data files are corrupt or missing, or in other situations where an earlier snapshot of data is required, you can restore them from backups:

1. Shut down the Terracotta cluster.
2. (Optional) Make copies of any existing data files.
3. Delete the existing data files from your Terracotta servers.
4. Copy the backup data files to the directory from which you deleted the original (existing) data files.
5. Restart the Terracotta cluster.

8 Clearing Data from a Terracotta Server

- How to Clear Data from a Terracotta Server 68

How to Clear Data from a Terracotta Server

After a Terracotta server is restarted, under certain circumstances it will retain artifacts from previous runs and its data directory must be manually cleared. These circumstances include running with Fast Restartability disabled and BigMemory Hybrid enabled. This may also be the source of errors during "split brain" resolution or during a mirror server restart.

By default, the number of copies of a server's objectdb data that are retained is unlimited. Over time, and with frequent restarts, these copies may consume a substantial amount of disk space. You can manually delete these files, which are saved in the server's data directory under /dirty-objectdb-backup/dirty-objectdb-<timestamp>. You can also set a limit for the number of backups by adding the following element to the Terracotta configuration file's <tc-properties> block:

```
<property name="l2.nha.dirtydb.rolling" value="<myValue>" />
```

where <myValue> is an integer.

If you have Fast Restartability disabled and BigMemory Hybrid enabled, you can also automatically remove the older data before server shutdown by setting the property `l2.nha.dirtydb.autoDelete` to "true". It is important to note that a backup copy will still be made in the configured /dirty-objectdb-backup folder prior to the most recent "working" data being deleted. This can potentially consume all available disk space. If you want to disable the creation of backups altogether, you can set the property `l2.nha.dirtydb.backup.enabled` to "false".

9 Changing Topology of a Live Cluster

- About Changing the Topology 70
- Adding a New Server 70
- Removing an Existing Server 71
- Editing the Configuration of an Existing Server 71

About Changing the Topology

Using the TMC, you can change the topology of a live cluster by reloading an edited Terracotta configuration file.

Note the following restrictions:

- Only the removal or addition of <server> blocks in the <servers> or <mirror-group> section of the Terracotta configuration file are allowed.
- All servers and clients must load the same configuration file to avoid topology conflicts.

Servers that are part of the same server array but do not share the edited configuration file must have their configuration file edited and reloaded as shown below. Clients that do not load their configuration from the servers must have their configuration files edited to exactly match that of the servers.

Note: Changing the topology of a live cluster will not affect the distribution of data that is already loaded in the TSA. For example, if you added a stripe to a live cluster, the data in the server array would not be redistributed to utilize it. Instead, the new stripe could be used for adding new caches, while the original servers would continue to manage the original data.

Adding a New Server

To add a new server to a Terracotta cluster, follow these steps:

1. Add a new <server> block to the <servers> or <mirror-group> section in the Terracotta configuration file being used by the cluster. The new <server> block should contain the minimum information required to configure a new server. It should appear similar to the following, with your own values substituted:

```
<server host="myHost" name="server2" >
  <data>%(user.home)/terracotta/server2/server-data</data>
  <logs>%(user.home)/terracotta/server2/server-logs</logs>
  <tsa-port>9510</tsa-port>
  <management-port>9540</management-port>
</server>
```

2. Make sure you are connected to the TMC, and that the TMC is connected to the target cluster. See the *Terracotta Management Console User Guide* for more information on using the TMC.
3. With the target cluster selected in the TMC, click the **Administration** tab, then choose the panel.
4. Click **Reload**. A message appears with the result of the reload operation. A successful operation logs a message similar to the following:

```
2013-03-14 13:25:44,821 INFO - Successfully overridden server topology
```

```
from file at '/bigmemory-max-4/tc-config.xml'.
```

5. Start the new server.

Removing an Existing Server

To remove a server from a Terracotta cluster configuration, follow these steps:

1. Shut down the server you want to remove from the cluster. If you shutting down an active server, first ensure that a backup server is online to enable failover.
2. Delete the <server> block associated with the removed server from the Terracotta configuration file being used by the cluster. Make sure you are connected to the TMC, and that the TMC is connected to the target cluster. See the *Terracotta Management Console User's Guide* for more information on using the TMC.
3. With the target cluster selected in the TMC, click the **Administration** tab, then choose the **Change Topology** panel.
4. Click **Reload**. A message appears with the result of the reload operation. A successful operation logs a message similar to the following:

```
2013-03-14 13:25:44,821 INFO - Successfully overridden server topology
from file at '/bigmemory-max-4/tc-config.xml'.
```

The TMC will also display the event

```
Server topology reloaded from file at '/bigmemory-max-4/tc-config.xml'.
```

Editing the Configuration of an Existing Server

If you edit the configuration of an existing ("live") server and attempt to reload its configuration, the reload operation will fail. However, you can successfully edit an existing server's configuration by following these steps:

1. Remove the server by following the steps in "Removing an Existing Server " on page 71. Instead of deleting the server's <server> block, you can comment it out.
2. Edit the server's <server> block with the changed values.
3. Add (or uncomment) the edited <server> block.
4. In the TMC's **Change Server Topology** panel, click **Reload**. A message appears with the result of the reload operation.

Note: To be able to edit the configuration of an existing server, all clients must load their configuration from the Terracotta Server Array. Clients that load configuration from another source will fail to remain connected to the TSA due to a configuration mismatch.

10 Enabling Production Mode

- Setting the Production Mode Property 74

Setting the Production Mode Property

Production mode can be set by setting the Terracotta property in the Terracotta configuration:

```
<tc-properties>
  ...
  <property name="l2.enable.legacy.production.mode" value="true" />
</tc-properties>
```

Production mode requires the `--force` flag to be used with the stop-tc-server script if the target is an active server with no mirror.

11 Managing Distributed Garbage Collection

- About Distributed Garbage Collection (DGC) 76
- Running the Periodic Distributed Garbage Collection 76
- Monitoring and Troubleshooting DGC 76

About Distributed Garbage Collection (DGC)

There are two types of DGC: periodic and inline. The periodic DGC is configurable and can be run manually (see below). Inline DGC, which is an automatic garbage-collection process intended to maintain the server's memory, runs even if the periodic DGC is disabled.

Note that the inline DGC algorithm operates at intervals optimal to maximizing performance, and so does not necessarily collect distributed garbage immediately.

Running the Periodic Distributed Garbage Collection

The periodic DGC can be run in any of the following ways:

- run-dgc shell script - Call the run-dgc shell script to trigger DGC externally.
- JMX - Trigger DGC through the server's JMX management interface.

By default, DGC is disabled in the Terracotta configuration file in the `<garbage-collection>` section. However, even if disabled, it will run automatically under certain circumstances when clearing garbage is necessary but the inline DGC does not run (such as when a crashed server returns to the cluster).

Monitoring and Troubleshooting DGC

DGC events (both periodic and inline) are reported in a Terracotta server instance's logs. DGC events can also be monitored using the Terracotta Management Console.

If DGC does not seem to be collecting objects at the expected rate, one of the following issues may be the cause:

- Java GC is not able to collect objects fast enough. Client nodes may be under resource pressure, causing GC collection to run behind, which then causes DGC to run behind.
- Certain client nodes continue to hold references to objects that have become garbage on other nodes, thus preventing DGC from being able to collect those objects.

If possible, shut down all Terracotta clients to see if DGC then collects the objects that were expected to be collected.

12 Starting the Terracotta Server as a Windows Service

■ Configuring the Terracotta Server to Run as a Service 78

Configuring the Terracotta Server to Run as a Service

A Windows service supports scheduling and automatic start and restart. You might want to run the Terracotta Server Array or the Cross-Language Connector, which are Java applications, as a *Windows service*. If so, use the Service Wrapper located inside the kit at $installdir/server/wrapper (or, for 3.7, $installdir/wrapper).

Set JAVA_HOME

To start the service, set your `JAVA_HOME` in conf/wrapper-tsa.conf or conf/wrapper-clc.conf . For example:

```
set.JAVA_HOME=C:/Java/jdk1.7.0_21
```

The wrapper does not read your `JAVA_HOME` from the environment. For Windows, if you do not want to set it in the configuration file, comment it out and set `JAVA_HOME` in the registry instead.

Configuration Files

For the Cross-Language Connector, you need these configuration files:

- conf/cross-language-config.xml
- conf/ehcache.xml

For the TSA, you need this configuration file:

- conf/tc-config.xml

Overwrite those files with your own. If you want to change these file names, modify the names in the wrapper configurations.

Modify the TSA conf/wrapper-tsa.conf file to match the server name in your tc-config.xml:

```
set.SERVER_NAME=server0
```

where *server0* represents the name of the server you want to start.

Set Permissions

The services are controlled by an Administrator user, so you have to confirm for every action, such as install, start, stop, remove.

In addition, the Administrator user needs to have read/write permission for the "wrapper" directory.

Install and Start the Service

The wrapper service is located at $installdir/server/wrapper.

To install the service wrapper, run the script with the install parameter:

```
%> bin/tsa-service.bat install
```

Note: The examples in this section show the TSA script. For the Cross-Language Connector, use the clc-service or clc-service.bat script.

Then you can either start/stop the service:

```
%> bin/tsa-service.bat start
%> bin/tsa-service.bat stop
```

If you want to remove the service:

```
%> bin/tsa-service.bat remove
```

There are more commands available when you run the script without any parameter:

```
%> bin/tsa-service.bat
```

Changing Wrapper Configuration

There are comments in wrapper-tsa.conf and wrapper-clc.conf to explain each parameter. If you need to modify JVM system properties, classpath, or command line parameters, follow the current pattern. Pay close attention to their numerical order and parameter counts.

For more information, see http://wrapper.tanukisoftware.com/doc/english/properties.html.

13 Using BigMemory Hybrid

- About BigMemory Hybrid 82
- System Requirements 84
- Hardware Capacity Guidelines 84
- Configuring BigMemory Hybrid 84
- Using the TMC with BigMemory Hybrid 85
- Operator Events 86

About BigMemory Hybrid

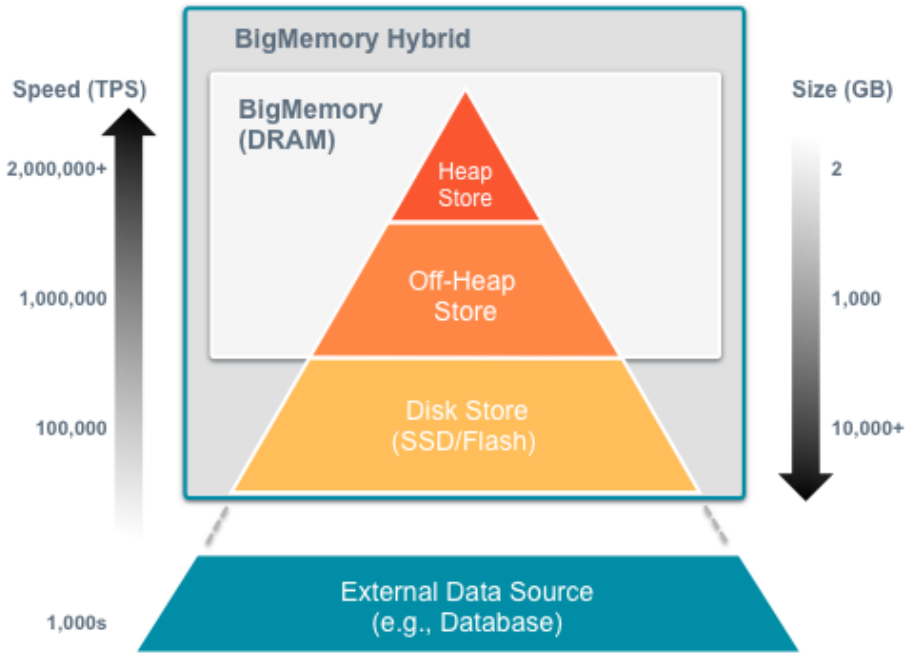
BigMemory Hybrid is an optional extension to BigMemory Max that:

- Enables scaling up to economical solid-state drive (SSD) flash memory motionless "disks" in conjunction with conventional dynamic random-access memory (DRAM) memory. This means that the cache size can exceed the available off-heap memory.
- Provides predictable low latency at very large scale.
- Manages the data flow seamlessly and automatically from DRAM to flash "disk" according to the size of the cache.
- Performs much faster than conventional hard disks, although not quite as fast as a pure DRAM in-memory solution.
- Supports searching, Fast Restart backup and recovery, Web Sessions, Quartz, and WAN replication.
- Works with industry-standard SSD devices from popular vendors, such as Fusion IO and Intel SSD.

How it works

For use in Terracotta servers, when using BigMemory Hybrid, all data is stored in SSD/Flash drives. The keys are stored in off-heap memory, providing optimal performance when using Hybrid (though there is an option to store cache keys on disk, but this comes with a performance penalty and is not recommended).

Figure 1. BigMemory Hybrid allows you to expand BigMemory in Terracotta servers, keeping more data closer to your application for increased transactions per second (TPS).



How is BigMemory Hybrid different than Overflow to Disk?

BigMemory Hybrid	Overflow to Disk
Available in version 4.1	Available in version 3.7
Leverages BigMemory's Fast Restart technology	Depends upon Berkeley DB to store data on disk
Manages all data in SSD/Flash for predictable performance. Only the cache keys are stored in off-heap memory.	If data does not fit in off-heap, it is pushed to disk, hence performance is less predictable
Optimized for SSD usage	No optimization done for SSDs

System Requirements

BigMemory Hybrid supports writing to one single mount, so all of the BigMemory Hybrid capacity must be presented to the Terracotta process as one continuous region, which can be a single device or a RAID.

The mount should be used exclusively for the Terracotta server process.

Note: System utilization is higher when using BigMemory Hybrid, and it is not recommended to run multiple servers on the same machine. Doing so could result health checkers timing out, and killing or restarting servers. Therefore, it is important to provision sufficient hardware, and it is highly recommended to deploy servers on different machines.

Hardware Capacity Guidelines

To account for the overhead necessary for consistent performance, the formulas below are suggested as initial starting points for sizing the amount of space allocated for BigMemory Hybrid operation.

Minimum SSD flash memory requirement = planned total data size * 3.2

Minimum DRAM requirement = planned maximum number of elements * (168 + key size)

Note: It is strongly recommended to configure enough offheap to accommodate all cache keys in DRAM.

Configuring BigMemory Hybrid

To configure BigMemory Hybrid, include the following elements in the tc-config.xml file:

- **dataStorage** - Specifies the maximum amount of data you plan to store on the server, using either DRAM alone or both DRAM and SSD flash memory.
- **offheap** - Specifies the maximum amount of data to hold in DRAM.
- **hybrid** - Enables the Hybrid option to use SSD flash memory in addition to off-heap DRAM.

For example:

```
<servers>
    ....
    <server host="hostname" name="server1">
        ...
        <dataStorage size="800g">
```

```
            <offheap size="200g"/>
            <hybrid/>
          </dataStorage>
    </server>
</servers>
```

For Terracotta servers, a minimum of 4 GB is recommended for the size attribute of the `offheap` element.

If the `hybrid` element is present, then the BigMemory Hybrid functionality is enabled. With Hybrid enabled, the value of the size attribute for the `dataStorage` element can exceed that of the size attribute for the `offheap` element. This enables SSD devices to supplement the DRAM and be many times larger than the DRAM.

If the `hybrid` element is absent, then BigMemory Hybrid functionality is off. With Hybrid off, the value of the size attribute for the `dataStorage` element must be less than or equal to the value of the size attribute for the `offheap` element. In this case, the `offheap` element is not required.

If the `dataStorage` element is absent, `dataStorage` size and `offheap` size default to 512 MB.

Although the `dataStorage` element is optional, if included, this element must have a value assigned to its size attribute.

Note: If you are migrating from BigMemory Max 4.0 to 4.1, the `dataStorage` element has replaced the `maxDataSize` element. The old element is still compatible for pure DRAM operation, but to enable Hybrid mode, you must use the new 4.1-compatible `dataStorage` element with the `hybrid` tag.

Disk Storage Path

BigMemory Hybrid requires a unique and explicitly specified path. The default path is the Terracotta server's home directory. You can customize the path using the `<data>` element in the server's tc-config.xml configuration file.

BigMemory Hybrid and Fast Restartability

If Fast Restartabillity is enabled, then if you have a restart, data will be loaded into BigMemory Hybrid in the same way as for BigMemory, with no difference in behavior or time required to get the system running again. See "Fast Restartability" on page 15.

If Fast Restartabillity is not enabled, then on restart, you will have some artifacts from the previous run left on disk, and you may want to remove them. For more information, see "Clearing Data from a Terracotta Server" on page 67.

Using the TMC with BigMemory Hybrid

When you use the Terracotta Management Console (TMC), you can see the effect of the BigMemory Hybrid feature in the Monitoring > Runtime Stats panel as "Data Storage

Usage". When the cache is operating at a steady state, the Data Used typically exceeds the OffHeap Max shown in the "OffHeap Usage" graph:

Operator Events

BigMemory Hybrid supports the existing operator events in the Terracotta Server Array (TSA), including

- a Throttle Mode when the amount of either RAM, Flash, or both is approaching its capacity limit. See "Managing Near-Memory-Full Conditions" on page 51.
- a Restricted Mode when the system has entered read-only mode.

For more information, see "Managing Near-Memory-Full Conditions" on page 51.

14 Logging

- SLFJ Logging 88
- Recommended Logging Levels 88

SLFJ Logging

BigMemory Max uses the SLF4J logging facade, so you can plug in your own logging framework. The following information pertains to Ehcache logging. For information about SLF4J in general, refer to the SLF4J website.

With SLF4J, users must choose a concrete logging implementation at deploy time. The options include Maven and the download kit.

Concrete Logging Implementation use in Maven

The maven dependency declarations are reproduced here for convenience. Add *one* of these to your Maven POM.

```
<dependency>
    <groupId>org.slf4j</groupId>
    <artifactId>slf4j-jdk14</artifactId>
    <version>1.5.8</version>
</dependency>
 <dependency>
    <groupId>org.slf4j</groupId>
    <artifactId>slf4j-log4j12</artifactId>
    <version>1.5.8</version>
</dependency>
 <dependency>
    <groupId>org.slf4j</groupId>
    <artifactId>slf4j-simple</artifactId>
    <version>1.5.8</version>
</dependency>
```

Concrete Logging Implementation use in the Download Kit

The slf4j-api jar is in the kit along with the BigMemory Max jars so that, if the app does not already use SLF4J, you have everything you need. Additional concrete logging implementations can be downloaded from SLF4J website.

Recommended Logging Levels

BigMemory Max seeks to trade off informing production-support developers of important messages and cluttering the log. ERROR messages should not occur in normal production and indicate that action should be taken.

WARN messages generally indicate a configuration change should be made or an unusual event has occurred. DEBUG and TRACE messages are for development use. All DEBUG level statements are surrounded with a guard so that no performance cost is incurred unless the logging level is set. Setting the logging level to DEBUG should provide more information on the source of any problems. Many logging systems enable a logging level change to be made without restarting the application.

15 Using Command Central to Manage Terracotta

- Commands that Terracotta Supports 90
- Configuration Types that Terracotta Supports 91
- Lifecycle Actions for Terracotta 91
- Run-time Monitoring Statuses for Terracotta 92
- Run-time Monitoring States for Terracotta 92

Commands that Terracotta Supports

Terracotta supports the Command Central CLI (command line interface) commands listed in the following table. In cases where there is Terracotta-specific information, the table lists where you can learn more about arguments and options that Terracotta supports or details about the actions Terracotta takes when you execute an `exec` command.

CLI Commands	Additional Information
`sagcc get configuration data`	For Terracotta-specific information about configuration types, see "Configuration Types that Terracotta Supports" on page 91.
`sagcc get configuration instances`	
`sagcc list configuration instances`	
`sagcc get configuration types`	
`sagcc list configuration types`	
`sagcc get diagnostic logs`	
`sagcc get monitoring`	
`sagcc get inventory components`	
`sagcc list inventory components`	
`sagcc exec lifecycle`	For Terracotta-specific information about using this command, see "Lifecycle Actions for Terracotta" on page 91.

For information about Command Central CLI commands, see the Command Central Help.

Configuration Types that Terracotta Supports

The Terracotta run-time component supports creating instances of the configuration types listed in the following table.

Configuration Type	Use to...
COMMON-CLUSTER	View the configuration for a cluster.
COMMON-MEMORY	Configure the JVM memory settings.
COMMON-PORTS	View the custom port configuration for a Terracotta Server Array instance.
TC-CONFIG	Configure the tc-config.xml file.
TC-SERVER-SERVER	Assign a server name to a Terracotta Server Array instance that matches the server name in the tc-config.xml file.

Lifecycle Actions for Terracotta

The following table lists the actions that Terracotta supports with the `sagcc exec lifecycle` command and the operation taken against Terracotta when an action is executed.

Important: You must execute all lifecycle operations in the correct order to ensure safe startup, shutdown, and restart for Terracotta Server Array cluster instances.

Action	Description
Start	Start a server instance that has stopped.
Stop	Stop a running server instance.
Restart	Restart a running server instance.

Run-time Monitoring Statuses for Terracotta

The following table lists the run-time statuses that the Terracotta run-time component can return in response to the `sagcc get monitoring runtimestatus` and `sagcc get monitoring state` commands, along with the meaning of each run-time status.

Run-time Status	Meaning
STOPPED	The server is not running.
ONLINE_MASTER	The server is running and is the master in its stripe.
ONLINE_SLAVE	The server is running and is a slave (mirror) in its stripe.
FAILED	The server was running, but crashed. A possible reason for this failure is that the server did not find the correct license.
UNRESPONSIVE	The server is running, but is not responding on the specified ports. A possible reason for this failure is that the server did not find the correct license.

Run-time Monitoring States for Terracotta

In response to the `sagcc get monitoring runtimestate` and `sagcc get monitoring state` commands, the Terracotta run-time component provides information about the following key performance indicators (KPIs):

KPI	Description
Live Object Count	The number of live objects in the cache.
Offheap	The off-heap storage capacity for the cache.
Write Operations Rate	The rate of write operations for the cache.

The KPIs apply to the whole Terracotta Server Array cluster.

A Operational Scripts

- Archive Utility (archive-tool) 94
- Database Backup Utility (backup-data) 94
- Backup Status (backup-status) 95
- Cluster Thread and State Dumps (debug-tool, cluster-dump) 95
- Distributed Garbage Collector (run-dgc) 96
- Start and Stop Server Scripts (start-tc-server, stop-tc-server) 97
- Server Status (server-stat) 98
- Version Utility (version) 99

Archive Utility (archive-tool)

The archive-tool is used to gather logs generated by a Terracotta server or client for the purpose of contacting Terracotta with a support query.

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\server\bin\archive-tool.bat <args>
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/server/bin/archive-tool.sh <args>
```

where `<args>` are:

- [-n] (No Data - excludes data files)
- [-c] (Client - include files from the client)
- [path to terracotta config xml file (tc-config.xml) or path to logs directory
- [output filename in .zip format]

Database Backup Utility (backup-data)

The backup utility creates a backup of the data being shared by your application by taking a snapshot of the data held by the Terracotta Server Array (TSA). Unless a different directory is specified in configuration, backups are saved to the default directory `${user.dir}/terracotta/backups`.

Configuring Backup

You can override this default behavior by specifying a different backup directory in the server's configuration file using the `<data-backup>` property:

```
<servers>
   <restartable enabled="true"/>
...
   <server host="%i" name="myServer">
   ...
     <data-backup>/Users/myBackups</data-backup>
   </server>
...
</servers>
```

Note that enabling `<restartable>` mode is required for using the backup utility.

Creating a Backup

The backup utility relies on the Terracotta Management Server (TMS) to locate and execute backups. The TMS must be running and connected to the TSA for the backup to take place. For more information about connecting the Terracotta Management Server to the TSA, see the *Terracotta Management Console User Guide*.

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\tools\management-console\bin\backup-data.bat <args>
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/tools/management-console/bin/backup-data.sh <args>
```

where `<args>` are:

- [l] <tms-host:port> – The host and port used to connect to the TMS. If omitted, `localhost:9889` is used by default.
- [u] <username> – If the TMS requires authentication, a username must be specified.
- [p] <password> – If the TMS requires authentication, a password must be specified.
- [a] <agentID> – Specify the agent ID of a TSA. The agent ID is set as a connection name when the connection to the TSA is configured on the TMS. If no agent ID is provided, the TMS returns a list of configured agent IDs.
- [k] This flag causes invalid TMS SSL certificates to be ignored.

For example, to initiate a backup on a cluster with the agent ID "someConnection":

```
${BIGMEMORY_HOME}/tools/management-console/bin/
backup-data.sh -l my-tms-host:9889 \ -u admin
-p admin -a someConnection -k
```

If initiation is successful, the script reports that the backup process has started. Once the backup is complete, the backup data files can be used to restore data in place of the current data files. For information about restoring data from a backup, see "Restoring Data from a Backup" on page 65.

Backup Status (backup-status)

The backup-status script is run from the tools/management-console/bin directory. This script complements the backup-data utility by checking on the status of executed backups for a specified cluster. For example, to return a list of backup operations on the agent myClusterAgent:

```
[PROMPT] ${BIGMEMORY_HOME}/tools/management-console/bin/backup-status
    -l http://myTMShost:9889 -a myClusterAgent
```

The backup-status script takes the same arguments as backup-data. For details, see "Database Backup Utility (backup-data)" on page 94.

Cluster Thread and State Dumps (debug-tool, cluster-dump)

The cluster and thread- and state-dump debug tools provide a way to easily generate debugging information that can be analyzed locally or forwarded to support personnel. These tools work against the Terracotta Management Server that is monitoring the target Terracotta cluster. All components must be running at the time a tool is used.

- `debug-tool` generates thread dumps for all nodes in the cluster, with each node's dump saved its log file. A flag is available for saving the thread dumps to a single zip file.
- `cluster-dump` provides a similar service, but adds each node's state, including information on locks. Note that these tools can generate a substantial amount of data.

Note: Server utility scripts do not work when a server is starting up or when a server is in the process of recovering using the Fast Restart feature.

For more information on operating these tools, run the associated script with the `-h` flag. For example:

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\tools\management-console\bin\debug-tool.bat -h
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/tools/management-console/bin/debug-tool.sh -h
```

Distributed Garbage Collector (run-dgc)

`run-dgc` is a utility that causes the specified cluster to perform distributed garbage collection (DGC). Use run-dgc to force a periodic DGC cycle in environments where inline DGC is not in effect. However, automated DGC collection is sufficient for most environments.

This utility relies on the Terracotta Management Server (TMS) to locate and execute backups. The TMS must be running and connected to the TSA for the DGC to be initiated. For more information about connecting the Terracotta Management Server to the TSA, see the *Terracotta Management Console User Guide*.

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\tools\management-console\bin\run-dgc.bat <args>
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/tools/management-console/bin/run-dgc.sh <args>
```

where `<args>` are:

- [l] <tms-host:port> – The host and port used to connect to the TMS. If omitted, `localhost:9889` is used by default.
- [u] <username> – If the TMS requires authentication, a username must be specified.
- [p] <password> – If the TMS requires authentication, a password must be specified.
- [a] <agentID> – Specify the agent ID of a TSA. The agent ID is set as a connection name when the connection to the TSA is configured on the TMS. If no agent ID is provided, the TMS returns a list of configured agent IDs.
- [k] This flag causes invalid TMS SSL certificates to be ignored.

Note: Two DGC cycles cannot run at the same time. Attempting to run a DGC cycle on a server while another DGC cycle is in progress generates an error

For more information on distributed garbage collection, see "Managing Distributed Garbage Collection" on page 75.

Start and Stop Server Scripts (start-tc-server, stop-tc-server)

Use the `start-tc-server` script to run the Terracotta Server, optionally specifying a configuration file:

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\server\bin\start-tc-server.bat ^
    [-n <name of server>] [-f <config specification>]
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/server/bin/start-tc-server.sh \
    [-n <name of server>] [-f <config specification>]
```

`<config specification>` can be one of:

- Path to configuration file
- URL to configuration file
- <server host>:<tsa-port> of another running Terracotta Server

Note the following:

- If no configuration is specified, a file named tc-config.xml in the current working directory will be used.
- If no configuration is specified and no file named tc-config.xml is found in the current working directory, a default configuration will be used.
- If no server is named, and more than one server exists in the configuration file used, an error is printed to standard out and no server is started.

Use the `stop-tc-server` script to cause the Terracotta Server to gracefully terminate:

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\server\bin\stop-tc-server.bat <host-name> <jmx-port> <args>
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/server/bin/stop-tc-server.sh <host-name> <jmx-port> <args>
```

where `<args>` are:

- [f] `<file-or-URL>` – Specifies the tc-config file to use, as a file path or URL. For an SSL-secured server, a valid path to the self-signed certificate must have been specified in the server's configuration file.
- [--force] – Force shutdown of the active server.

In production mode, if the stop-tc-server script detects that the mirror server in STANDBY state isn't reachable, it issues a warning and fails to shut down the active server. If failover is not a concern, you can override this behavior with the `--force` flag. For information about production mode, see "Enabling Production Mode" on page 73.

- [n] `<server-name>` – The name of the server to shut down. Defaults to the local host.
- [s] – If the server is secured with a JMX password, then a username and password must be passed into the script.
- [u] – Specify the JMX username. For an SSL-secured server, the user specified must have the "admin" role.
- [w] – Specify the JMX password.
- [k] – This flag causes invalid TMS SSL certificates to be ignored. Use this option to accept self-signed certificates (ones not signed by a trusted CA).

For more information, see "Setting up Server Security" in the *BigMemory Max Security Guide*.

Server Status (server-stat)

The status tool is a command-line utility for checking the current status of one or more Terracotta server instances.

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\server\bin\server-stat.bat <args>
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/server/bin/server-stat.sh <args>
```

where `<args>` are:

- [-s] host1,host2,... – Check one or more servers using the given hostnames or IP addresses using the default JMX port (9520).
- [-s] host1:9520,host2:9521,... – Check one or more servers using the given hostnames or IP addresses with JMX port specified.
- [-f] <path>/tc-config.xml – Check the servers defined in the specified configuration file.
- [-k] – This flag causes invalid TMS SSL certificates to be ignored. Use this option to accept self-signed certificates (ones not signed by a trusted CA).

The status tool returns the following data on each server it queries:

- *Health*– OK (server responding normally) or FAILED (connection failed or server not responding correctly).

- *Role* – The server's position in an active-mirror group. Single servers always show ACTIVE. Backups are shown as MIRROR or PASSIVE.
- *State* – The work state that the server is in. When ready, active servers should show ACTIVE-COORDINATOR, while mirror servers should show MIRROR-STANDBY or PASSIVE-STANDBY.
- *JMX port* – The TCP port the server is using to listen for JMX events.
- *Error* – If the status tool fails, the type of error.

Example

The following example shows usage of and output from the status tool.

```
[PROMPT] server-stat.sh -s myhost:9521
 localhost.health: OK
 localhost.role: ACTIVE
 localhost.state: ACTIVE-COORDINATOR
 localhost.jmxport: 9521
```

If no server is specified, by default the tool checks the status of localhost at JMX port 9520.

Version Utility (version)

The version tool is a utility script that outputs information about the BigMemory installation, including the version, date, and version-control change number from which the installation was created. When contacting Terracotta with a support query, please include the output from the version tool to expedite the resolution of your issue.

For Microsoft Windows:

```
[PROMPT] %BIGMEMORY_HOME%\server\bin\version.bat
```

For UNIX/Linux:

```
[PROMPT] ${BIGMEMORY_HOME}/server/bin/version.sh&
```

Use the following flags to produce more information:

- [r] – Produces detailed, raw information in a "property=value" format.
- [v] – Produces more detailed information.

B Terracotta Configuration Parameters

- The Terracotta Configuration File 102
- The Servers Parameters 105
- The Clients Parameters 112

The Terracotta Configuration File

This document is a reference to all of the Terracotta configuration elements in the Terracotta configuration file, which is named `tc-config.xml` by default.

You can use a sample configuration file provided in the kit as the basis for your Terracotta configuration. Some samples have inline comments describing the configuration elements. Be sure to start with a clean file for your configuration.

The Terracotta configuration XML document is divided into the sections <servers> and <clients>.

- The <servers> section contains parameters you use to configure the behavior and characteristics of the Terracotta Server Array and its component servers.
- The <clients> section contains parameters you use to configure client behavior.

Configuration Variables

Certain variables can be used that are interpolated by the configuration subsystem using local values:

Variable	Interpolated Value
%h	The fully-qualified hostname
%i	The IP address
%o	The operating system
%v	The version of the operating system
%a	The CPU architecture
%H	The home directory of the user running the application
%n	The username of the user running the application
%t	The path to the temporary directory (for example, /tmp on *NIX)
%D	Time stamp (yyyyMMddHHmmssSSS)

Variable	Interpolated Value
%(*system property*)	The value of the given Java system property

These variables can be used where appropriate, including for elements or attributes that expect strings or paths for values:

- the "name", "host" and "bind" attributes of the `<server>` element
- the password file location for JMX authentication
- client logs location
- server logs location
- server data location

Note: The variable %i is expanded into a value determined by the host's networking setup. In many cases that setup is in a `hosts` file containing mappings that may influence the value of %i. Test this variable in your production environment to check the value it interpolates.

Using Paths as Values

Some configuration elements take paths as values. Relative paths are interpreted relative to the current working directory (the directory from which the server was started). Specifying an absolute path is recommended.

Overriding tc.properties

Every Terracotta installation has a default `tc.properties` file containing system properties. Normally, the settings in `tc.properties` are pre-tuned and should not be edited.

If tuning is required, you can override certain properties in `tc.properties` using `tc-config.xml`. This can make a production environment more efficient by allowing system properties to be pushed out to clients with `tc-config.xml`. Those system properties would normally have to be configured separately on each client.

Setting System Properties in tc-config

To set a system property with the same value for all clients, you can add it to the Terracotta server's `tc-config.xml` file using a configuration element with the following format:

```
<property name="<tc_system_property>" value="<new_value>" />
```

All `<property />` tags must be wrapped in a `<tc-properties>` section placed at the beginning of `tc-config.xml`.

For example, to override the values of the system properties `l1.cachemanager.enabled` and `l1.cachemanager.leastCount`, add the following to the beginning of `tc-config.xml`:

```
<tc-properties>
  <property name="l1.cachemanager.enabled" value="false" />
  <property name="l1.cachemanager.leastCount" value="4" />
</tc-properties>
```

Override Priority

System properties configured in `tc-config.xml` override the system properties in the default `tc.properties` file provided with the Terracotta kit. The default `tc.properties` file should *not* be edited or moved.

If you create a *local*`tc.properties` file in the Terracotta `lib` directory, system properties set in that file are used by Terracotta and will override system properties in the *default*`tc.properties` file. System properties in the local `tc.properties` file are *not* overridden by system properties configured in `tc-config.xml`.

System property values passed to Java using -D override all other configured values for that system property. In the example above, if `-Dcom.tc.l1.cachemanager.leastcount=5` was passed at the command line or through a script, it would override the value in `tc-config.xml` and `tc.properties`. The order of precedence is shown in the following list, with highest precedence shown last:

1. default `tc.properties`
2. `tc-config.xml`
3. local, or user-created `tc.properties` in Terracotta `lib` directory
4. Java system properties set with -D

Failure to Override

If system properties set in `tc-config.xml` fail to override default system properties, a warning is logged to the Terracotta logs. The warning has the following format:

```
The property <system_property_name> was set by local settings to <value>.
This value will not be overridden to <value> from the tc-config.xml file.
```

System properties used early in the Terracotta initialization process may fail to be overridden. If this happens, a warning is logged to the Terracotta logs. The warning has the following format:

```
The property <system_property_name> was read before initialization completed.
```

The warning is followed by the value assigned to `<system_property_name>`.

Note: The property `tc.management.mbeans.enabled` is known to load before initialization completes and cannot be overridden.

The Servers Parameters

/tc:tc-config/servers

This section defines the Terracotta server instances present in your cluster. One or more entries can be defined, either directly under the <servers> element or in the "/tc:tc-config/servers/mirror-group" on page 110. If this section is omitted, Terracotta configuration behaves as if there's a single server instance with default values.

This section also defines certain global settings that affect all servers, including the attribute `secure`. This is a global control for enabling ("true") or disabling ("false" DEFAULT) SSL-based security for the entire cluster.

For information about SSL-based security, see the *BigMemory Max Security Guide*.

/tc:tc-config/servers/server

A server stanza encapsulates the configuration for a Terracotta server instance. The server element takes three optional attributes (see table below).

Attribute	Definition	Value	Default Value
host	The address of the machine hosting the Terracotta server	Host machine's IP address or resolvable hostname	Host machine's IP address
name	The symbolic name of the Terracotta server; can be passed to Terracotta scripts such as start-tc-server using `-n <name>`	user-defined string	<host>:<tsa-port>
bind	The network interface on which the Terracotta server listens cluster traffic; 0.0.0.0 specifies all interfaces	interface's IP address	0.0.0.0

Each Terracotta server instance needs to know which configuration it should use as it starts up. If the server's configured name is the same as the hostname of the host it runs on and no host contains more than one server instance, then configuration is found automatically.

Here is a sample configuration snippet

```
<server>
  <!-- my host is '%i', my name is '%i:tsa-port', my bind is 0.0.0.0 -->
  ...
</server>
<server host="myhostname">
  <!-- my host is 'myhostname', my name is 'myhostname:tsa-port',
       my bind is 0.0.0.0 -->
  ...
</server>
<server host="myotherhostname" name="server1" bind="192.168.1.27">
  <!-- my host is 'myotherhostname', my name is 'server1',
       my bind is 192.168.1.27 -->
  ...
</server>
```

/tc:tc-config/servers/server/data

This element specifies the path where the server should store its data for persistence.

Default: data (creates the directory `data` under the working directory)

/tc:tc-config/servers/server/logs

This section lets you declare where the server should write its logs.

Default: logs (creates the directory `logs` under the working directory)

You can also specify `stderr:` or `stdout:` as the output destination for log messages. For example:

```
<logs>stdout:</logs>
```

/tc:tc-config/servers/server/index

This element specifies the path where the server should store its search indexes.

Default: index (creates the directory `index` under the working directory)

/tc:tc-config/servers/server/data-backup

This element specifies the path where the server should store backups (if a backup call is initiated).

Default: data-backup (creates the directory `data-backup` under the working directory)

/tc:tc-config/servers/server/tsa-port

This section lets you set the port that the Terracotta server listens to for client traffic.

The default value of "tsa-port" is 9510.

Here is a sample configuration snippet:

```
<tsa-port>9510</tsa-port>
```

/tc:tc-config/servers/server/jmx-port

Note: Listening on the "jmx-port" is deprecated. Alternatively, use the monitoring features provided by the Terracotta Management Console (see the *Terracotta Management Console User Guide*) and the WAN Replication Service (see the *WAN Replication User Guide*).

"jmx-port" is disabled by default. To enable it, add `jmx-enabled="true"` to the `<server host>` sections of tc-config.xml. For example:

```
<server host="localhost" name="My Server Name1" jmx-
enabled="true">
```

This section lets you set the port that the Terracotta server's JMX Connector listens to.

The default value of "jmx-port" is 9520. If tsa-port was set to a value other than the default 9510, this port defaulted to the value of the tsa-port plus 10.

Here is a sample configuration snippet:

```
<jmx-port>9520</jmx-port>
```

/tc:tc-config/servers/server/tsa-group-port

This section lets you set the port that the Terracotta server uses to communicate with other Terracotta servers.

The default value of "tsa-group-port" is 9530. If tsa-port is set to a value other than the default 9510, this port defaults to the value of the tsa-port plus 20.

Here is a sample configuration snippet:

```
<tsa-group-port>9530</tsa-group-port>
```

/tc:tc-config/servers/server/management-port

This section lets you set the port that the Terracotta Management Console (TMC) uses.

The default value of "management-port" is 9540. If tsa-port is set to a value other than the default 9510, this port defaults to the value of the tsa-port plus 30.

Here is a sample configuration snippet:

```
<management-port>9540</management-port>
```

Prior to 4.2, the TMC used the ports specified in `/tc:tc-config/servers/server/tsa-port` (port 9510 by default) and `/tc:tc-config/servers/server/tsa-group-port` (port 9530 by default). Now the TMC uses only the port specified in `/tc:tc-config/servers/server/management-port`.

/tc:tc-config/servers/server/security

This section contains the data necessary for running a secure cluster based on SSL, digital certificates, and node authentication and authorization.

For more information, see "About Security in a Cluster" in the *BigMemory Max Security Guide*.

/tc:tc-config/servers/server/security/ssl/certificate

The element specifying certificate entry and location of the certificate store. The format is:

```
<store-type>:<certificate-alias>@</path/to/keystore.file>
```

The Java Keystore (JKS) type is supported by Terracotta 3.7 and higher.

/tc:tc-config/servers/server/security/keychain

This element contains the following subelements:

- `<class>` – Element specifying the class defining the keychain file. If a class is not specified, `com.terracotta.management.keychain.FileStoreKeyChain` is used.
- `<url>` – The URI for the keychain file. It is passed to the keychain class to specify the keychain file.
- `<secret-provider>` – The fully qualified class name of the user implementation of `com.terracotta.management.security.SecretProviderBackEnd`. This class can read and provide the keychain file.

/tc:tc-config/servers/server/security/auth

This element contains the following subelements:

- `<realm>` – Element specifying the class defining the security realm. If a class is not specified, `com.tc.net.core.security.ShiroIniRealm` is used.
- `<url>` – The URI for the Realm configuration (.ini) file. It is passed to the realm class to specify authentication file. Alternatively, URIs for LDAP or Microsoft Active directory can also be used if one of these schemes is implemented instead.

- `<user>` – The username that represents the server and is authenticated by other servers. This name is part of the credentials stored in the .ini file. The default value is "terracotta".

/tc:tc-config/servers/server/security/management

This element contains the subelements needed to allow the Terracotta Management Server (TMS) to make a secure connection to the TSA:

- `ia` – The HTTPS URL with the domain of the TMS, followed by the port 9443 and the path `/tmc/api/assertIdentity`.
- `timeout` – The timeout value (in milliseconds) for connections from the server to the TMS.
- `hostname` – Used only if the DNS hostname of the server does not match server hostname used in its certificate. If there is a mismatch, enter the DNS address of the server here.

/tc:tc-config/servers/server/authentication

Turn on JMX authentication for the Terracotta server. An empty tag (`<authentication />`) defaults to the standard Java JMX authentication mechanism referring to password and access files in `$JAVA_HOME/jre/lib/management`:

```
$JAVA_HOME/jre/lib/management/jmxremote.password
$JAVA_HOME/jre/lib/management/jmxremote.access
```

You must modify these files as follows (or, if none exist create them).

jmxremote.password: Add a line to the end of the file declaring a username and password followed by a carriage return:

```
secretusername secretpassword
```

jmxremote.access: Add the following line (with a carriage return) to the end of your file:

```
secretusername      readwrite
```

Be sure to assign the appropriate permissions to the file. For example, in *NIX:

```
$ chmod 500 jmxremote.password
$ chown <user who will run the server> jmxremote.password
```

For information on alternatives to JMX authentication, see the *BigMemory Max Security Guide*.

Note that version 4.x does not support HTTP authentication.

/tc:tc-config/servers/dataStorage

This configuration block includes the required `offheap` element, as well as the optional `hybrid` element.

- `<offheap>` must be configured for each server; the minimum amount is 4 GB.
- `<dataStorage>` specifies the maximum amount of data to store on each server, and represents the hybrid sum of off-heap DRAM plus flash SSD.
- `<hybrid/>`, if included, enables the Hybrid option.

Here is a sample configuration snippet:

```
<dataStorage size="800g">
   <offheap size="200g"/>
   <hybrid/>
</dataStorage>
```

If the `hybrid` element is present, then `dataStorage` size may exceed `offheap` size, however if the hybrid element is not present, then the `dataStorage` size must be less than or equal to the `offheap` size.

/tc:tc-config/servers/mirror-group

A mirror group is a *stripe* in a TSA, consisting of one active server and one or more mirror (or backup) servers. A configuration that does not use the <mirror-group> element would produce a one-stripe TSA:

```
<servers>
  <server name="A">
  ...
  </server>
  <server name="B">
  ...
  </server>
  <server name="C">
  ...
  </server>
  <server name="D">
  ...
  </server>
...
</servers>
```

One of the named servers would assume the role of active (the one started first or that wins the election), while the remaining servers become mirrors. Note that in a typical stripe, having only one or two mirrors is sufficient and less taxing on the active server's resources (as it needs to sync with each mirror).

The following example shows the same servers split into two stripes:

```
<servers>
  <mirror-group group-name="team1">
    <server name="A">
    ...
    </server>
    <server name="B">
    ...
    </server>
  </mirror-group>
  <mirror-group group-name="team2">
    <server name="C">
    ...
```

```
    </server>
    <server name="D">
    ...
    </server>
  </mirror-group>
...
</servers>
```

Each stripe will have one active and one mirror server.

Note: Under <servers>, you may use either <server> or <mirror-group> configurations, but not both. All <server> configurations directly under <servers> work together as one mirror group, with one active server and the rest mirrors. To create more than one stripe, use <mirror-group> configurations directly under <servers>. The mirror group configurations then include one or more <server> configurations

For more examples and information, see "Configuring the Terracotta Server Array" on page 33.

/tc:tc-config/servers/garbage-collection

This section lets you configure the periodic distributed garbage collector (DGC) that runs in the TSA. The DGC collects shared data made garbage by Java garbage collection.

For many use cases, there is no need to enable periodic DGC. For caches, the more efficient automatic inline DGC is normally sufficient for clearing garbage. In addition, certain read-heavy applications will never require the periodic DGC as little shared data becomes garbage.

However, concerning certain data structures, the periodic DGC may need to be enabled. Inline DGC may not be available for all data structures.

For more on how DGC functions, see "Managing Distributed Garbage Collection" on page 75.

Here is a configuration snippet:

```
<garbage-collection>
 <!--    Default: false -->
  <enabled>true</enabled>
  <!-- If "true", additional information is logged when a
     server performs distributed garbage collection.
     Default: false
  -->
  <verbose>false</verbose>
  <!-- How often should distributed garbage collection
     be performed, in seconds?
     Default: 3600 (60 minutes)
  -->
  <interval>3600</interval>
</garbage-collection>
```

/tc:tc-config/servers/restartable

The fast-restart persistence mechanism must be explicitly enabled for the TSA using this element:

```
<restartable enabled="true"/>
```

In case of TSA failure, fast-restart persistence allows the TSA to reload all shared cluster data.

To function, this feature requires <offheap> to be enabled on each server. To make backups of TSA data, the backup feature requires this feature to be enabled.

/tc:tc-config/servers/client-reconnect-window

This section lets you declare the window of time servers will allow disconnected clients to reconnect to the cluster as the same client. Outside of this window, a client can only rejoin as a new client. The value is specified in seconds and the default is 120 seconds.

If adjusting value, note that a too-short reconnection window can lead to unsuccessful reconnections during failure recovery, while a too-long window lowers the efficiency of the cluster since it is paused for the time the window is in effect.

For more information on how client and server reconnection is executed in a Terracotta cluster, and on tuning reconnection properties in a high-availability environment, see the *BigMemory Max High-Availability Guide*.

The Clients Parameters

/tc:tc-config/clients/logs

This section lets you configure where the Terracotta client writes its logs.

Here is a sample configuration snippet:

```
<!--
   This value undergoes parameter substitution before being used;
   thus, a value like 'client-logs-%h' would expand to
   'client-logs-banana' if running on host 'banana'. See the
   Product Guide for more details.
   If this is a relative path, then it is interpreted relative to
   the current working directory of the client (that is, the directory
   you were in when you started the program that uses Terracotta
   services). It is thus recommended that you specify an absolute
   path here.
   Default: 'logs-%i'; this places the logs in a directory relative
   to the directory you were in when you invoked the program that uses
   Terracotta services (your client), and calls that directory, for example,
   'logs-10.0.0.57' if the machine that the client is on has assigned IP
   address 10.0.0.57.
-->
<logs>logs-%i</logs>
```

You can also specify `stderr:` or `stdout:` as the output destination for log messages. For example:

```
<logs>stdout:</logs>
```

To set the logging level, see "Logging" on page 87.