TERRACOTTA
A SOFTWARE AG COMPANY

TERRACOTTA TECHNICAL WHITEPAPER:

BigMemory: In-Memory Data Management for the Enterprise

Abstract

As your application's data grows, maintaining scalability and performance is an increasing challenge. With BigMemory, you can perform real-time processing on terabytes of data in memory, cutting processing time from minutes to seconds, or less. The result is a 100% Java software solution for data-related performance and scalability problems that can easily be plugged into your data-intensive application today. Use BigMemory to give your Java applications instant access to terabytes of enterprise in-memory data storage for high performance at any scale, and to transform your business by creating new opportunities with this volume of data. BigMemory enables you to achieve this with a full set of enterprise data management capabilities including high availability, durability, consistency and transaction support, and monitoring.

Highlights

This paper covers the following features of BigMemory:

- **Performance at any scale**
  - In-memory data access is hundreds of times faster than disk or network-based storage
  - Keep your entire data set in memory to achieve maximum performance and scalability
- **Maximize server memory usage**
  - Utilize all of your server's memory for maximum performance and value
  - BigMemory keeps all of your data in memory without large a Java heap
- **Predictably with simplicity**
  - Dramatically decrease latency and increase throughput
  - BigMemory offers predictable data access time with low latency to meet your SLAs
- **Business in real time**
  - Accelerate business decisions based on your application data
  - Harness the hidden value in your enterprise data
- **Enterprise Data Management**
  - BigMemory "snaps in" to your enterprise applications
  - Capabilities include 100% reliability, high availability, transactions, consistency, and multi-data center support
- **Transformational Technology**
  - Build applications that turn high volumes of transactional data (or Big Data) into new opportunities and business advantage
- **A 100% pure Java software solution**
  -Seamlessly works with Java applications

Why In-Memory?

Applications need speed and scale in today's hyper-fast, need-it-now world. Competitive pressures and escalating customer expectations have put a premium on absolute application performance and throughput. Successful businesses must scale rapidly, which can stress traditional multi-tier application architectures.

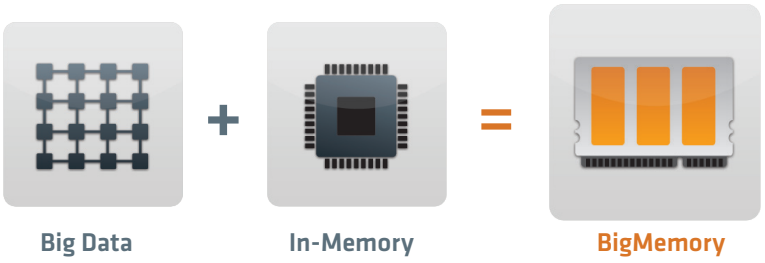
If your applications use data stored in a central resource, such as a database, web-service or other type of server, you know that indirect data access becomes a bottleneck that slows application performance dramatically. Alternatively, keeping data directly in the server's inexpensive RAM, where your application runs, offers the best performance and maximum value.

Why Big?

Data volumes continue to grow as applications become more connected and customer expectations increase. In fact, according to IDC[1] and Gartner[2] data volumes have been measured to increase ten-fold every five years, quickly outpacing the capability of existing technologies and even Moore's law. Since both business and customer value are often associated with the need for massive amounts of data, enterprises are always looking for a solution.

Additionally, there's a growing need to dynamically process this data in real time to derive even more value and use data in new ways. Traditional data storage technologies just aren't built to meet these requirements. See the *BigMemory Use Cases* section, later in this document, to see how companies are using BigMemory to transform what's possible to do.

Why BigMemory?



BigMemory allows you to store data where it's used: in memory. With BigMemory, you can easily store and manage your in-memory data in a reusable, standard way that simply plugs into your application. As a result, you can harness the value of massive amounts of data, with real-time processing, while keeping all of it in your server's RAM for maximum performance and scale. These performance improvements that result will continue to scale as your application's data set and user base grow. There's no simpler way to get predictable, and fast, access to large volumes of in-memory data.

In this paper, we'll explore the capabilities of BigMemory, such as its fast, low-latency data access, predictability, scalability, high availability, durability, consistency, and monitoring and management. We'll also cover how simple it is to get started and how it works with your existing hardware and developer skill set. Additionally, BigMemory supports the full consistency spectrum for data reliability and transaction support with maximum performance, including a high-availability configuration (no single point of failure), to meet enterprise requirements.

Using BigMemory requires no code changes, and only a few lines of configuration. Your application will see benefits immediately and in the future, as BigMemory allows you to scale up, scale out, and scale to the cloud.

1 Gantz, John F. The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011. Tech. An IDC White Paper - Sponsored by EMC. Web. <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>
2 Paquet, Raymond. "Technology Trends You Can't Afford to Ignore." Lecture. Gartner Webinar. Gartner.com. Gartner Inc., Jan. 2010. Web. <http://www.gartner.com/it/content/1258400/1258425/january_6_techtrends_rpaquet.pdf>.

Snap-In Performance and Scale

For the absolute best enterprise application performance, you need to store data where it's used and needed: in memory. Accessing data in your server's RAM is hundreds of time faster than disk or network-based access. Additionally, since inexpensive servers with hundreds of gigabytes of RAM are increasingly abundant, it makes sense to use as much of it as you can. With BigMemory, you can access and manage *all* of your enterprise data in memory to achieve the scale your application requires.

Capacity is virtually unlimited as you can affordably buy servers with large amounts of RAM and move even terabytes of data into memory. BigMemory allows you to scale up and out, utilizing your data in ways that weren't possible before BigMemory customers are seeing an immediate speed-up in application response times as well as increased throughput and transaction rates. Watch our customer videos at http://terracotta.org/resources/video to see firsthand how they are using BigMemory to transform their business.

Maximize Memory Use with BigMemory

Many in-memory data solutions—commercial or custom—store application data in the Java heap, and are therefore subject to garbage collection (GC) pauses. However, BigMemory is a pure-Java solution that enables in-memory data storage *off* the Java heap. As a result, this data is not visible to Java's garbage collector, and reduces the need for a large Java heap.

As your in-memory data demands grow, the Java heap can remain consistently small, further controlling and reducing garbage collection related pauses and latencies. Since the data still resides in-process, and in memory, accessing the data remains fast—around three orders of magnitude faster than a heap-based cache distributed across multiple Java virtual machines.

You can also achieve significant performance gains and cost savings by maximizing server density. BigMemory helps you achieve the most from a single server to avoid the costs that go along with scaling out across servers when it may not be needed. With BigMemory, you can scale across servers when you require it, while avoiding the complexity involved when you don't.

BigMemory gives Java applications instant access to a large memory footprint, but without the garbage collection cost. This solves three main problems commonly seen in Java applications with high data demands:

1. **Database-related delays:** keep application data in memory, eliminating costly database, web-server or other data access delays.
2. **Unpredictable GC latencies:** keep Java heap sizes small and eliminate GC latency
3. **Complicated deployment and management:** simply plug BigMemory into your application and keep hundreds of gigabytes or more of data cached in memory without the need to distribute that data across multiple Java instances.

For maximum value and flexibility, BigMemory is part of a tiered storage architecture designed to keep the most important data where it's needed most: in the local memory of your application. For high availability, consistency and scalability, all of BigMemory's in-memory data is available on demand from an external server array and an on-disk

backing store. Any portion of your server array can go offline at any time with no application downtime and no loss of data. BigMemory is the only data management solution that offers access to terabytes of in-memory data with this level of performance, availability and operational flexibility. Further, with BigMemory, you'll have enough headroom available to meet your growing data needs on the smallest hardware footprint possible.

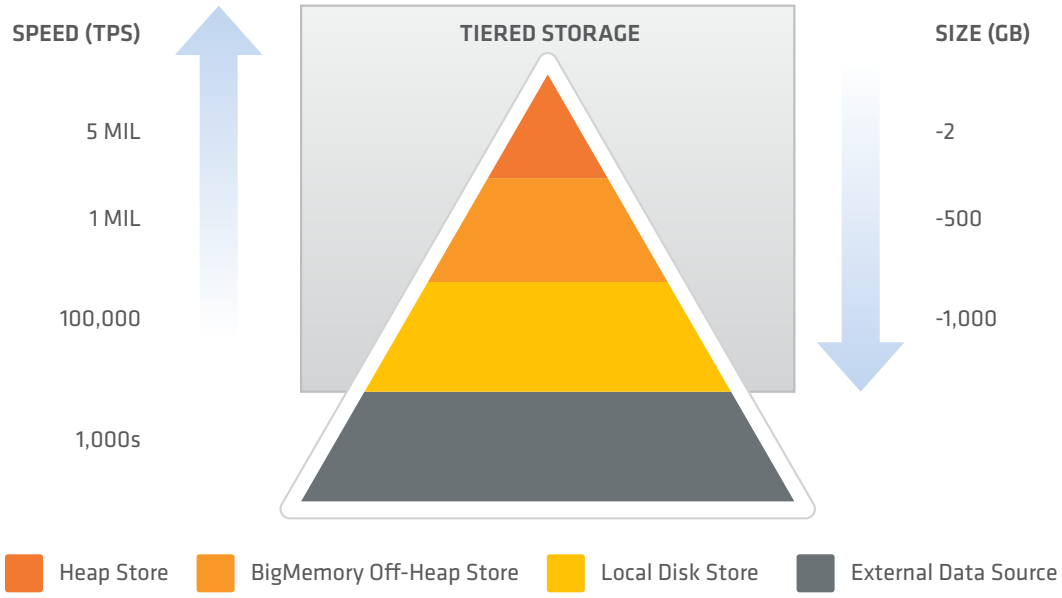


Figure 1

Predictability and Low Latency

The top-most layer represents the area within the Java heap that BigMemory keeps the most frequently-used data, allowing for read/write latencies less than 1 microsecond. The layer immediately below the heap represents BigMemory's in-memory cache that's a bit further away from the heap and hidden from the garbage collector, so that it never causes a pause in the JVM while it sits there resident. Caches hundreds of gigabytes in size can be accessed in less than 100 microseconds with no garbage collection penalties.

BigMemory maximizes the memory available to your application within each node in your application cluster. As a result, a large multi-terabyte in-memory data solution is instantly available to you at a fraction of the number of nodes. In customer deployments, we typically see the number of servers consolidate by a factor of four or more.

Why Predictability and Low Latency are Important

Many measure the performance of an enterprise application in terms of overall throughput achieved. This may be expressed as the number of *transactions-per-second* the application can process, and so on. While throughput is an important measurement, it can be misleading, because it's just an average of responses over time. For instance, systems that can process thousands of requests-per-seconds (or more) may have quite a few responses with up to a full second of latency, or greater—see Figure 2. Even though a majority of the requests—or transactions—are processed with low latency, the existence of a few with large latency represents outliers that may violate your user service level agreements (SLAs).

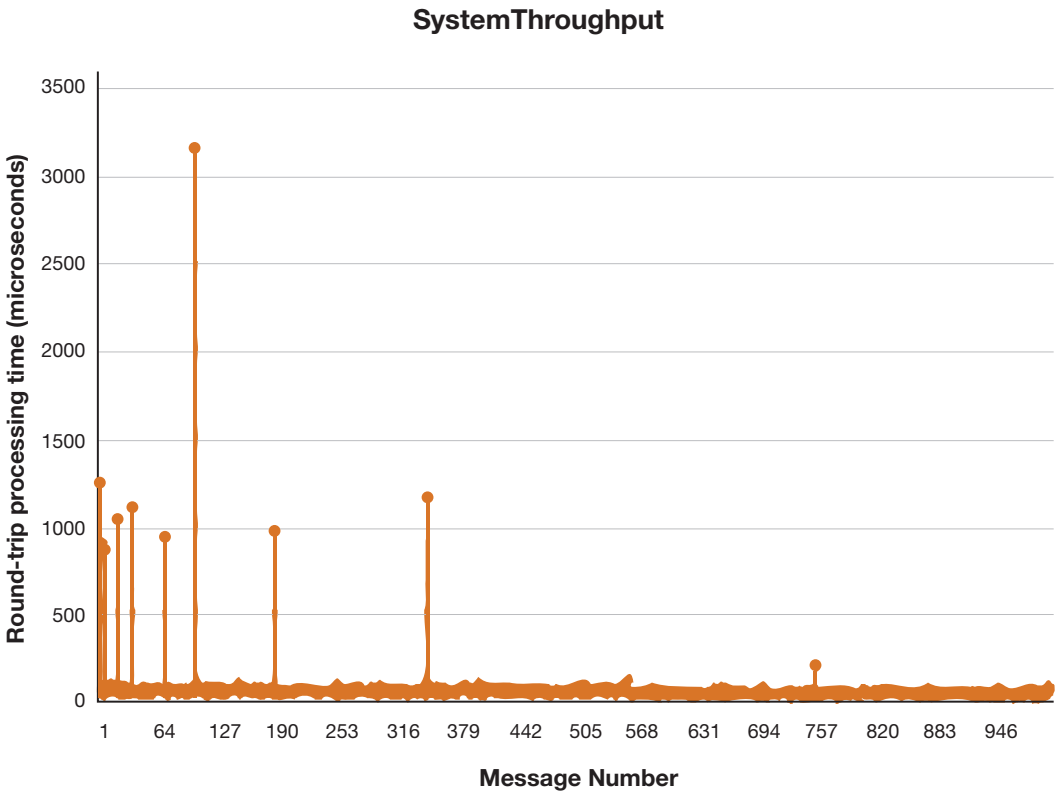


Figure 2

In a Java application, most latency outliers are due to long garbage collection pauses associated with a large Java heap. To achieve the best predictability in terms of absolute performance and latency (not just the average), you need to control and eliminate the unpredictable nature of the garbage collector. BigMemory does this by keeping all of your large data sets in memory, but outside of the Java heap. This eliminates both the performance bottlenecks of disk or network-based data storage, while eliminating the GC pauses associated with otherwise large Java heaps.

Because BigMemory lets you store increasingly large amounts of data in memory while keeping the Java heap small, application performance will continue to improve while latencies stay consistently small. In effect, BigMemory provides your application the best of all worlds when it comes to throughput, latency and scalability. This design is capable of delivering the appropriate balance of these three aspects of performance, including data consistency and correctness, according to the specific business needs of your application.

"We are able to reduce the heap size and get fast, local access for large amounts of data."

– Joe Caisse, CTO, News Digital Media

Business in Real Time

With all your data in memory available for real-time analysis, you can accelerate business decisions and satisfy customer requests much more quickly when compared to disk or network-based retrieval. Not only does this give you a potential competitive advantage and increased customer value, it allows you to harness value in your data that may have been hidden before.

For instance, data mining for valuable, marketable, data relationships and associations is often performed in off-line batch processes. With BigMemory's in-memory data management, this processing can be performed in real time while customers use your application. As a result, your application can personalize the user experience in real time, adding value to the user, and providing new opportunities to market to your customers' needs as they arise. For example, in a recent customer deployment, BigMemory reduced risk analysis calculation processing time from minutes to seconds, enabling them to act on their risk data during the customer transaction rather than after the fact.

BigMemory's support for in-memory data also makes it possible to instantly stream large amounts of data to mobile devices, without waiting for disk-based retrieval or impacting other users on the system sharing a smaller Java heap.

Enterprise Data Management

Keeping all of your enterprise data in memory is a big step towards achieving the highest performance and scalability. However, this alone isn't enough. An enterprise-grade in-memory solution requires a full suite of data management capabilities, such as scalability, high availability, data consistency, monitoring and multi-data center support.

Scalability

We've examined how BigMemory allows you to scale up and utilize all of the inexpensive RAM available in today's servers. With the Server Array, BigMemory also scales out across multiple servers for unlimited scalability and high availability.

The Server Array is an independently scalable set of storage servers that runs on commodity hardware. This array delivers enterprise-grade data management to BigMemory in the application tier. Each server in the array has an in-memory store and a disk-backed permanent store. Similar to RAID, this array is configured into groups of servers to form mirrored stripes. The data in the server array is partitioned across the existing stripes. Over time, more stripes may be added as needed to increase the total addressable cache size and I/O throughput.

"We needed a solution that would enable us to scale to millions of users, and banks of servers. With Terracotta, we've seen fantastic performance stability as we scale."

– Phil Sant, Omnifone

High Availability: Guaranteed Uptime and Data Access

To ensure maximum uptime and reliability, BigMemory runs in a high-availability configuration with no single point of failure. All in-memory data writes from the application layer to the server array are internally transactional and guaranteed. Any application server or server array node may be restarted or fail with no data loss.

The data management features of the Server Array provide a central authority that enables a number of runtime optimizations not available to other in-memory solutions. For example, transactions may be batched, folded and reordered at runtime to increase throughput. Latency is minimized, because no cross-node acknowledgements are required. For high availability, each node in the Server Array is transactionally mirrored. Should a server node in a stripe be restarted or fail, one of the mirrors will automatically take its place, ensuring maximum uptime and data reliability.

The Server Array architecture allows new stripes to be added without rehashing all of the existing stripes. As a result, new stripes can be brought online instantly. BigMemory with the Server Array offers a number of capabilities that allow instant-on server deployment:

- A bulk-loading mechanism that warms up new servers before adding them to the array, protecting the application from the runtime computational overhead and latency of in-memory data loading.
- A kick-start function makes new server array topology configurations instantly available to the application cluster. This means that when new servers are ready for service, the application can make use of their extra capacity immediately.

All of the high-availability features of BigMemory can scale geographically, across multiple data centers. This not only supports applications with extremely distributed architecture, but also disaster recovery in case you lose access to an entire data center of servers. BigMemory offers 100% reliability and high availability built into its architecture at all levels of scale with unmatched performance.

The Consistency Spectrum

Across the enterprise, there are typically requirements to support data access along a spectrum of consistency guarantees. This spectrum ranges from purely asynchronous operations suitable for read-only access to fully transactional access to business-critical data. Because the level of consistency affects throughput and latency and is dependent on the business rules of the application, BigMemory offers configurable consistency guarantees to different data sets in the same application.

At one end of the spectrum, BigMemory allows fully asynchronous access to cached data. This yields the highest throughput and lowest latency, but the lowest consistency guarantees. In the middle of the spectrum, BigMemory enforces synchronous access to cached data. This yields a balance between fast access to data while reading and a coherent, stable view of the data as it changes. At the far end of the consistency spectrum, BigMemory enforces fully transactional, XA compliant data access.

BigMemory ships with a default consistency setting that offers a balance of high consistency and high performance, but is easily configurable to suit the specific requirements of the application. No other in-memory data management solution supports the full range of the consistency spectrum all on the same architecture and deployment topology and within the same application using the same API. BigMemory is a single solution that delivers predictable and cost-effective performance at all levels of scale and consistency.

BigMemory Everywhere

Companies large and small have turned to BigMemory to accelerate business-critical, data-intensive applications and analyses. This is because BigMemory is broadly applicable, suitable for many types of applications across the enterprise.

In fact, you may be using BigMemory in common enterprise applications today without even knowing it. Have you booked a flight online? Some of our customers use BigMemory to speed up that highly transactional web-based process. Have you charged dinner on your credit card? BigMemory is currently being used for real-time fraud detection, scanning through hundreds of gigabytes of bank data in the second or two it takes to get an approval. Have you

streamed video to your mobile device? BigMemory provides the scale needed to support thousands of concurrent viewers without requiring a datacenter full of servers.

Here are just a few examples of how BigMemory is making a big difference for market leaders in a wide range of industries.

Financial Services

A global financial services firm processes trade reconciliations in a tight 4-hour window—accomplishing what canned database reports couldn't.

- Accelerate the processing of trade orders, credit card authorizations and other high-volume transactions
- Speed up large-scale data analysis for risk management, asset management or real-time fraud detection
- Provide fast, reliable access to aggregated account data on customer portals

Telecommunications

A major broadband carrier improved the performance of their billing system, boosting processing success rate from 80% to 99%.

- Speed up billing, subscriber provisioning and other high-volume transactions
- Improve call center efficiency with faster access to account data
- Expand subscriber self-service offerings with scalable online applications
- Implement a scalable, ultra-fast solution for network management

High-Tech, Internet and Online Gaming

A leading cloud service provider has achieved 100% uptime for its online meeting service by storing session data in memory.

- Accelerate searches, purchases, ad placements and other online transactions
- Handle spikes in demand and long-term traffic growth
- Ensure stable response times at any scale
- Boost service uptime

Entertainment and Media

A U.S. cable operator guarantees seamless TV viewing on the iPad with real-time user authentication.

- Scale web services to millions of concurrent users/viewers
- Display dynamic or aggregated data at lightning speed
- Ensure a superior user experience
- Reduce hardware costs

Travel, Transportation and Logistics

Europe's leading hotel portal ensures speedy online reservations—reducing database use by more than 50% at the same time.

- Increase throughput rates for booking, ticketing, and other high-volume transactions
- Instantly display such dynamic data as weather, delivery status, arrival time or traffic updates to enhance the value of customer portals

Government

A U.S. Government agency can now meet internal SLAs for three applications in two data centers.

- Improve the performance and scalability of mission-critical applications
- Support real-time data analysis and high-velocity data processing
- Deliver web services scalable to millions of citizens
- Comply with mandated SLA

Get Started

A thirty-day trial of BigMemory is available for download at http://terracotta.org/bigmemory. For more information on evaluating BigMemory or for pricing, please contact Terracotta sales.

Terracotta, Inc.
575 Florida St. Suite 100
San Francisco, CA 94110
USA

Product, Support, Training and Sales Information
sales@terracottatech.com

USA Toll Free
+1-888-30-TERRA

International
+1-415-738-4000

Terracotta China
china@terracottatech.com
+1-415-738-4088