Application Performance Management for Dummies

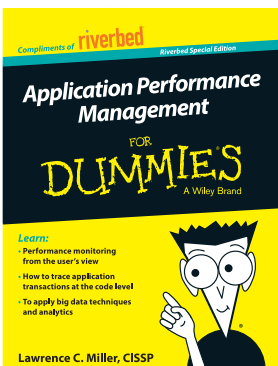**Compliments of** riverbed
Riverbed Special Edition

Application Performance Management
FOR DUMMIES
A Wiley Brand

Learn:
- Performance monitoring from the user's view
- How to trace application transactions at the code level
- To apply big data techniques and analytics

Lawrence C. Miller, CISSP

Learn best practices for APM:

- Performance monitoring from the user's view
- How to trace application transactions at the code level
- How to apply big data techniques and analytics

DOWNLOAD THE eBOOK

WHAT CUSTOMERS ARE ACHIEVING:

"42% of surveyed IT organizations' mean-time-to-resolution is 10 times faster or better since implementing Riverbed® SteelCentral™ AppInternals"

Source: TechValidate TVID:A39-F20-FC7

DOWNLOAD THE eBOOK

Java Performance Optimization

By Pierre-Hugues Charbonneau

Java is among the most widely used programming languages in the software development world today. Java applications are used within many verticals (banking, telecommunications, healthcare, etc.), and in some cases each vertical suggests a particular set of design optimizations. Many performance-related best practices are common to applications of all kinds. The purpose of this Refcard is to help developers improve application performance in as many business contexts as possible by focusing on the JVM internals, performance tuning principles and best practices, and how to make use of available monitoring and troubleshooting tools.

It is possible to define "optimal performance" in different ways, but the basic elements are: the ability of a Java program to perform its computing tasks within the business response time requirements, and the ability of an application to fulfill its business functions under high volume, in a timely manner, with high reliability and low latency. Sometimes the numbers themselves become patternized: for some major websites, a page response time of 500ms maximum per user function is considered optimal. This Refcard will include target numbers when appropriate, but in most cases you will need to decide these on your own, based on business requirements and existing performance benchmarks.

JVM INTERNALS

FOUNDATIONS

CODE COMPILATION AND JIT

Java byte code interpretation is clearly not as fast as native code executed directly from the host. In order to improve performance, the Hotspot JVM looks for the busiest areas of byte code and compiles these into native, more efficient, machine code (adaptive optimization). Such native code is then stored in the code cache in non-heap memory.

Note: most JVM implementations offer ways to disable the JIT compiler (Djava.compiler=NONE). You should only consider disabling such crucial optimization in the event of unexpected JIT problems such as JVM crashes.

The following diagram illustrates the Java source code, just-in-time compilation processes and life cycle.
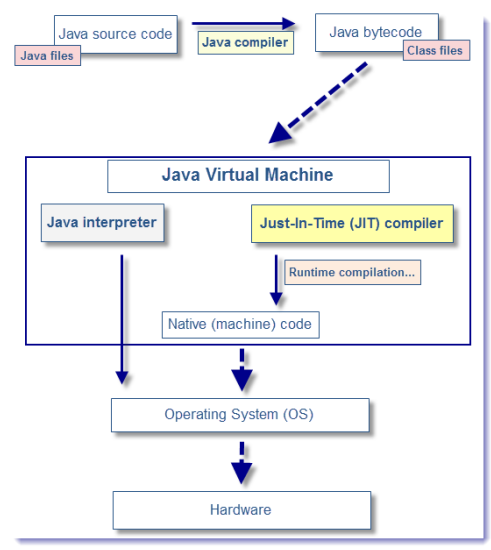
MEMORY SPACES

The HotSpot Java Virtual Machine is composed of the following memory spaces.

| MEMORY SPACE	DESCRIPTION
Java Heap	Primary storage of the Java program class instances and arrays.
Permanent Generation Metaspace (JDK 1.8)	Primary storage for the Java class metadata. NOTE: Starting with Java 8, the PermGen space is replaced by the Metaspace and using native memory, similar to the IBM JVM.
Native Heap (C-Heap)	Native memory storage for the Threads, Stack, code cache including objects such as MMAP files and third party native libraries.

CLASS LOADING

Another important feature of Java is its ability to load your compiled Java classes (bytecode) following the start-up of the JVM. Depending on the size of your application, the class loading process can be intrusive and significantly degrade the performance of your application under high load following a fresh restart. This short-term penalty can also be explained by the fact that the internal JIT compiler has to start over its optimization work following a restart.

It is important to note that several improvements were introduced since JDK 1.7, such as the ability for the default JDK class loader to better load classes concurrently.

HOT SPOTS

AREA OF CONCERN	RECOMMENDATION
Performance degradation following a JVM restart.	Avoid deploying an excessive amount of Java classes to a single application classloader (ex: very large WAR file).
Excessive class loading contention (thread lock, JAR file searches...) observed at runtime, degrading the overall performance.	Profile your application and identify code modules performing dynamic class loading operations too frequently. Look aggressively for non-stop class loading errors such as ClassNotFoundException and NoClassDefFoundError. Revisit any excessive usage of the Java Reflection API and optimize where applicable.
java.lang.OutOfMemoryError: PermGen space error or native memory leak observed.	Revisit the sizing of your JVM Permanent Generation and / or native memory capacity, where applicable. Analyze your application class loaders and identify any source of metadata memory leak.
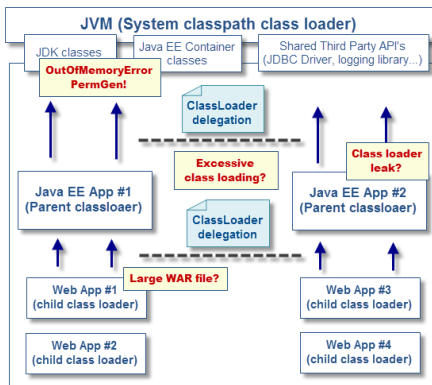
TROUBLESHOOTING & MONITORING

GOAL	RECOMMENDATION
Keep track of the Java classes loaded to the different class loaders.	Profile your application using a Java profiler of your choice such as JProfiler or Java VisualVM . Focus on class loader operations and memory footprint. Enable class loading details via –verbose:class. For the IBM JVM, generate multiple Java core snapshots and keep track of the active class loaders and loaded classes.

| Investigate suspected source(s) of class metadata memory leak(s).	Profile your application and identify the possible culprit(s). Generate and analyze JVM heap dump snapshots with a primary focus on ClassLoader and java.lang.Class instances.
Ensure a proper Permanent Generation / Metaspace and native memory sizing.	Closely monitor your PermGen, Metaspace and native memory utilization, and adjust the maximum capacity where applicable. Analyze your application class loaders size and identify opportunities to reduce the metadata footprint of your applications, where possible.

GARBAGE COLLECTION

The Java garbage collection process is one of the most important contributing factors for optimal application performance. In order to provide efficient garbage collection, the Heap is essentially divided into sub areas.

HEAP AREAS

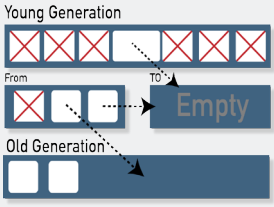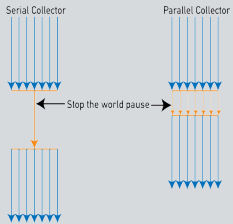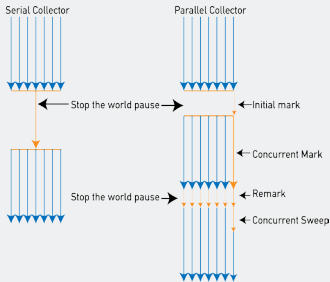| AREA	DESCRIPTION
Young Generation (nursery space)	Part of the heap reserved for allocation of new or short-lived objects. Garbage is collected by a fast but stop-the-world YG collector. Objects that have lived long enough in the young space are promoted to the old space NOTE: It is important to realize that an excessive size and / or GC frequency of the YG space can significantly affect the application response time due to increased JVM pause time.
Old Generation (tenured space)	Part of the heap reserved for long-lived objects. Garbage is usually collected by a parallel or mostly concurrent collector such as CMS or gencon (IBM JVM). Performance Tip: It is very important to choose and test the optimal GC policy for your application needs. For example, switching to a "mostly" concurrent GC collector such as CMS or G1 may significantly improve your application average response time (reduced latency).

Oracle HotSpot Heap Structure

riverbed
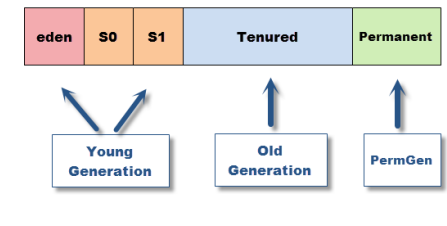
GC COLLECTORS

Choosing the right collector or GC policy for your application is a determinant factor for optimal application performance, scalability and reliability. Many applications are very sensible to response time latencies, requiring the use of mostly concurrent collectors such as the HotSpot CMS or the IBM GC policy `balanced`.

As a general best practice, it is highly recommended that you determine most suitable GC policy through proper performance and load testing. A comprehensive monitoring strategy should also be implemented in your production environment in order to keep track of the overall JVM performance and identify future areas for improvement.

GC	ARGUMENTS	DESCRIPTION
Serial Collector	-XX:+UseSerialGC (Oracle HotSpot)	Both Young and Old collections are done serially, using a single CPU and in a stop-the-world fashion. **NOTE:** This policy should only be used by client-side applications not sensitive to JVM pauses.
Parallel Collector	-XX:+UseParallelGC -XX:+UseParallelOldGC (Oracle HotSpot) -Xgcpolicy:optthruput (IBM J9, single space, stop-the-world)	Designed to take advantage of available CPU cores. Both Young and Old collections are done using multiple GC threads (via –XX:ParallelGCThreads=n), thus better leveraging the available CPU cores from the host. **NOTE:** While the collection time can be reduced significantly, applications with large heap size are still exposed to large and stop-the-world old collections and affecting the response time.
Mostly concurrent collectors (low-latency collectors)	Concurrent Mark-Sweep	Designed to minimize impact on application response time associated with Old generation stop-the-world collections. Most of the collection of the old generation using the CMS collector is done concurrently with the execution of the application. **NOTE:** The YoungGen collections are still stop-the-world events, thus requiring proper fine-tuning in order to reduce the overall JVM pause time.

GARBAGE FIRST (G1) COLLECTOR

The HotSpot G1 collector is designed to meet user-defined garbage collection (GC) pause time goals with high probability, while achieving high throughput.

This latest HotSpot collector essentially partitions the heap into a set of equal-sized heap regions, each a contiguous range of virtual memory. It concentrates its collection and compaction activity on the areas of the heap that are likely to be full of reclaimable objects (garbage first), or in other words on areas with the least amount of "live" objects.

Oracle recommends the following use cases or candidates for using the G1 collector, especially for existing applications currently using either the CMS or parallel collectors:

- Designed for applications that require large heaps (>= 6 GB) with limited GC latency (pause time <= 0.5 second).

- More than 50% of the Java heap is occupied with live data (objects that cannot be reclaimed by the GC).

- The rate of object allocation rate or promotion varies significantly.

- Undesired long garbage collection or compaction pauses (longer than 0.5 to 1 second).
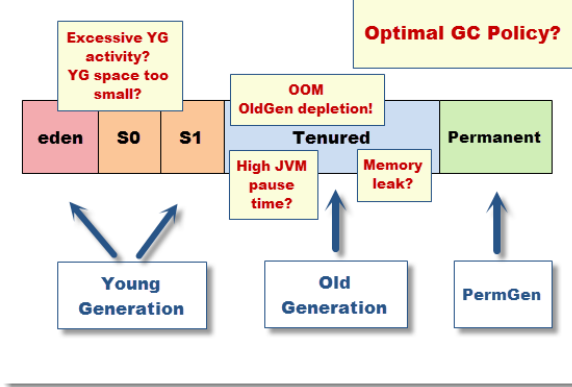
G1 Heap Allocation

JAVA HEAP SIZING

It is important to realize that no GC policy can save your application from an inadequate Java heap sizing. Such exercise involves configuring the minimum and maximum capacity for the various memory spaces such as the Young and Old generations, including the metadata and native memory capacity. As a starting point, here are some recommended guidelines:

- Choose wisely between a 32-bit or 64-bit JVM. If your application needs more than 2 GB to run with acceptable JVM pause time due to a large live data footprint, consider using a 64-bit JVM.

- Remember that the application is king: make sure that you profile it and adjust the heap sizing based on our application memory footprint. It is always recommended to measure the live data footprint through performance and load testing.

- A larger heap is not always better or faster: do not over-tune the Java heap. In parallel of JVM tuning, identify opportunities to reduce or "spread" your application memory footprint in order to keep the average JVM pause time < 1 %.

- For a 32-bit JVM, consider a maximum heap size of 2 GB in order to leave some memory from the address space to the metadata and native heap.

- For 64-bit JVM's, explore vertical and horizontal scaling strategies instead of simply attempting to expand the Java heap size beyond 15 GB. Such an approach very often provides better throughput, better leverages the hardware, and increases your application fail-over capabilities.

- Do not re-invent the wheel: take advantage of the multiple open source and commercials troubleshooting and monitoring tools available. The APM (Application Performance Management products have evolved significantly over the past decade.

riverbed

HOT SPOTS

Oracle HotSpot Heap Structure

TROUBLESHOOTING & MONITORING

GOAL	RECOMMENDATION
Measure and monitor your application YoungGen and OldGen memory footprint, including the GC activity. Determine the right GC policy and Java heap size for your application. Fine-tune your application memory footprint such as live objects.	Profile and monitor your application using a Java profiler of your choice such as JProfiler, Java VisualVM, or other commercial APM products. Enable the JVM GC activity logging via –verbose:gc. You can also use tools such as GCMV (GC Memory Visualizer) in order to assess your JVM pause time and memory allocation rate. **Performance Tip:** an excessive memory allocation rate may indicate a need to perform vertical and/or horizontal scaling, or to decouple your live data across multiple JVM processes. For your long-lived objects or long-term live data, consider generating and analyzing JVM heap dump snapshots. Heap dump analysis is also very useful at optimizing your application memory footprint (retention). **Performance Tip:** Since going from a 32-bit to a 64-bit machine increases heap requirement for an existing Java application by up to 1.5 times (bigger ordinary object pointers), it is very important to use -XX:+UseCompressedOops in Java version prior to 1.7 (which is now default). This tuning argument greatly alleviates the performance penalty associated with a 64-bit JVM.
Investigate OutOfMemoryError problems and suspected source(s) of OldGen memory leak.	Profile your application for possible memory leaks using tools such as Java VisualVM or Plumbr (Java memaory leak detector). **Performance Tip:** Focus your analysis on the biggest Java object accumulation points. It is important to realize that reducing your application memory footprint will translate in improved performance due to reduced GC activity. Generate and analyze JVM heap dump snapshots using tools such as Memory Analyzer.

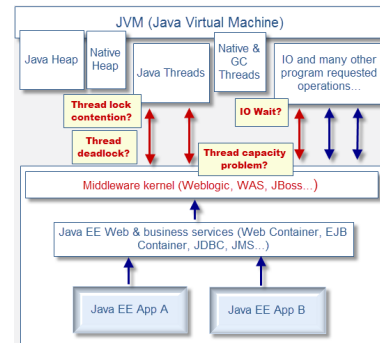## JAVA CONCURRENCY

Java concurrency can be defined as the ability to execute several tasks of a program in parallel. For large Java EE systems, this means the capability to execute multiple user business functions concurrently while achieving optimal throughput and performance.

Regardless of your hardware capacity or the health of your JVM, Java concurrency problems can bring any application to its knees and severely affect the overall application performance and availability.

THREAD LOCK CONTENTION

Thread lock contention is by far the most common Java concurrency problem that you will observe when assessing the concurrent threads health of your Java application. This problem will manifest itself by the presence of 1…n BLOCKED threads (thread waiting chain) waiting to acquire a lock on a particular object monitor. Depending onthe severity of the issue, lock contention can severely affect your application response time and service availability.

Example: Thread lock contention triggered by non-stop attempts to load a missing Java class (ClassNotFoundException) to the default JDK 1.7 ClassLoader.

```
ult task-15" prio=6 tid=0x0000000014849800 nid=0x2050 waiting for monitor entry [0x00000
a.lang.Thread.State: BLOCKED (on object monitor)
java.lang.ClassLoader.loadClass(ClassLoader.java:403)
aiting to lock <0x00000000ab84c0c8> (a java.lang.Object)      Waiting to acquire Object
sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:308)    monitor 0x00000000ab84c0c8
java.lang.ClassLoader.loadClass(ClassLoader.java:356)           held by another Thread
org.jboss.modules.ConcurrentClassLoader.performLoadClass(ConcurrentClassLoader.java:371)
org.jboss.modules.ConcurrentClassLoader.loadClass(ConcurrentClassLoader.java:119)
java.lang.Class.forName0(Native Method)
java.lang.Class.forName(Class.java:186)      Attempting to load a class via
org.jboss.tools.examples.rest.MemberResourceREST$          Class.forName()...         ResourceRE$
org.jboss.tools.examples.rest.MemberResourceRESTService$Proxy$_$$_WeldClientProxy.System(
sun.reflect.GeneratedMethodAccessor15.invoke(Unknown Source)
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
java.lang.reflect.Method.invoke(Method.java:601)
..............
```

It is highly recommended that you aggressively assess the presence of such a problem in your environment via proven techniques such as Thread Dump analysis. Typical root causes of this issue can vary from abuse of plain old Java synchronization to legitimate IO blocking or other non-thread safe calls. Lock contention problems are often the "symptoms" of another problem.

JAVA-LEVEL DEADLOCKS

True Java-level deadlocks, while less common, can also greatly affect the performance and stability of your application. This problem is triggered when two or more threads are blocked forever, waiting for each other. This situation is very different from other more common "day-to-day" thread problems such as lock contention, threads waiting on blocking IO calls etc. A true lock-ordering deadlock can be visualized as per below:
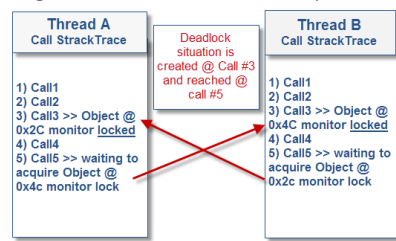
The Oracle HotSpot and IBM JVM implementations provide deadlock detectors for most scenarios, allowing you to quickly identify the culprit threads involved in such condition. Similar to lock contention troubleshooting, it is recommended to use techniques such as thread dump analysis as a starting point.

```
Found one Java-level deadlock:          Java-level deadlock detected!
============================
"pool-1-thread-5":
  waiting to lock monitor 0x04d29604 (object 0x2705cc50, a eu.javaspec:
  which is held by "pool-1-thread-1"
"pool-1-thread-1":
  waiting to lock monitor 0x04d2896c (object 0x2705cc58, a eu.javaspec:
  which is held by "pool-1-thread-2"
"pool-1-thread-2":
  waiting to lock monitor 0x0256e45c (object 0x2705cc60, a eu.javaspec:
  which is held by "pool-1-thread-3"
"pool-1-thread-3":
  waiting to lock monitor 0x0256e3f4 (object 0x2705cc68, a eu.javaspec:
  which is held by "pool-1-thread-4"
"pool-1-thread-4":
  waiting to lock monitor 0x04d2966c (object 0x2705cc70, a eu.javaspec:
  which is held by "pool-1-thread-5"
```

Once the culprit code is identified, solutions involve addressing the lock-ordering conditions and/or using other available concurrency programming techniques from the JDK such as java.util.concurrent. locks.ReentrantLock, which provides methods such as tryLock(). This approach gives Java developers much more flexibility and ways to prevent deadlock or thread lock "starvation."

CLOCK TIME AND CPU BURN

In parallel withthe JVM tuning, it is also essential that you review your application behavior, more precisely the highest clock time and CPU burn contributors.

When the Java garbage collection and thread concurrency are no longer a pressure point, it is important to drill down into your application code execution patterns and focus on the top response time contributors, referred as clock time. It is also crucial to review the CPU consumption of your application code and Java threads (CPU burn). High CPU utilization (> 75%) should not be assumed to be "normal" (good physical resource utilization). It is often the symptom of inefficientimplementation and/ or capacity problems. For large Java EE enterprise applications, it is essential to keep a safe CPU buffer zone in order to deal with unexpected load surges.

Stay away from traditional tracing approaches such as adding response time "logging" in your code. Java profiler tools and APM solutions exist precisely to help you with this type of analysis and in a much more efficient and reliable way. For Java production environments lacking a robust APM solution, you can still rely on tools such Java VisualVM, thread dump analysis (via multiple snapshots) and OS CPU per Thread analysis.

Finally, do not try to address all problems at the same time. Start by building a list of your top five clock time and CPU burn contributors and explore solutions.
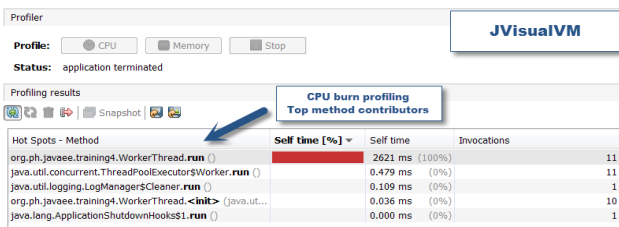
APPLICATION BUDGETING

Other important aspects of your Java applications performance are stability and reliability. This is particularly important for applications operating under a SLA umbrella with typical availability targets of 99.9%. These systems require a high fault-tolerant level, with strict application and resource budgeting in order to prevent domino effect scenarios. This approach prevents for example one business process from using all available physical, middleware, or JVM resources.

HOT SPOTS
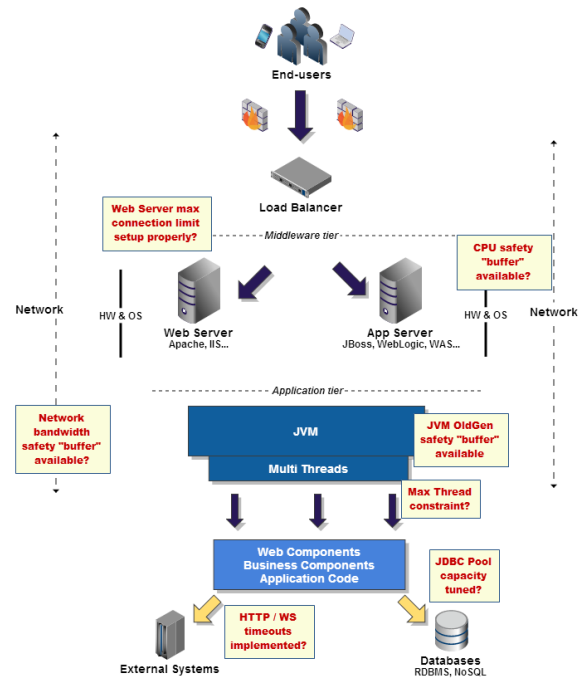
TIMEOUT MANAGEMENT

Lack of proper HTTP/HTTPS/TCP IP timeouts between your Java application and external systems can lead to severe performance degradation and outage due to middleware and JVM threads depletion (blocking IO calls). Proper timeout implementation will prevent Java threads from waiting for too long in the event of major slowdown of your external service providers.

```
"http-0.0.0.0-8443-102" daemon prio=3 tid=0x022a6400 nid=0x1bd runnable [0x78efb00
  java.lang.Thread.State: RUNNABLE
 at java.net.SocketInputStream.socketRead0(Native Method)
 at java.net.SocketInputStream.read(SocketInputStream.java:129)
 at com.sun.net.ssl.internal.ssl.InputRecord.readFully(InputRecord.java:293)
 at com.sun.net.ssl.internal.ssl.InputRecord.read(InputRecord.java:331)
 at com.sun.net.ssl.internal.ssl.SSLSocketImpl.readRecord(SSLSocketImpl.java:789)
 - locked <0xdd0ed968> (a java.lang.Object)           STUCK thread due to
 at com.sun.net.ssl.internal.ssl.SSLSocketImpl.readD   lack of HTTPS timeout
 at com.sun.net.ssl.internal.ssl.AppInputStream.read   with a remote Web
 - locked <0xdd0eda88> (a com.sun.net.ssl.internal.s   Service provider
 at java.io.BufferedInputStream.fill(BufferedInputSt      (blocking IO)
 at java.io.BufferedInputStream.read1(BufferedInputS
 at java.io.BufferedInputStream.read(BufferedInputSt
 - locked <0xddb1f6d0> (a java.io.BufferedInputStrea
 at sun.net.www.http.HttpClient.parseHTTPHeader(HttpClient.java:687)
 at sun.net.www.http.HttpClient.parseHTTP(HttpClient.java:632)
 at sun.net.www.http.HttpClient.parseHTTP(HttpClient.java:652)
 ..................
```

TOOLS

GOALS	RECOMMENDED TOOLS
Pro-active and real-time performance monitoring, tuning, alerting, trending, capacity management and more	Enterprise APM solutions **NOTE:** APM solutions provide tools allowing you to achieve most of the following Java performance goals out-of-the-box
Performance and load testing	Commercial performance testing solutions Apache JMeter http://jmeter.apache.org/

riverbed

JVM garbage collection assessment, memory allocation rate and troubleshooting	Oracle Java VisualVM http://docs.oracle.com/javase/8/docs/technotes/guides/visualvm/intro.html http://java.dzone.com/articles/profile-your-applications-java Oracle Java Mission Control http://www.oracle.com/technetwork/java/javaseproducts/mission-control/java-mission-control-wp-2008279.pdf http://www.oracle.com/technetwork/java/javase/jmc53-release-notes-2157171.html IBM Monitoring and Diagnostic Tools for Java (via IBM Support Assistant tool) http://www-01.ibm.com/software/support/isa/ JVM verbose:gc logs JVM argument : -verbose:gc http://docs.oracle.com/javase/8/docs/technotes/tools/windows/java.html IBM GCMV https://www.ibm.com/developerworks/java/jdk/tools/gcmv/
JVM heap and class metadata memory leak analysis	Oracle Java VisualVM and Oracle Java Mission Control IBM Monitoring and Diagnostic Tools for Java Memory Analyzer (heap dump analysis, hprof and phd formats) https://www.eclipse.org/mat/ https://www.ibm.com/developerworks/java/jdk/tools/memoryanalyzer/ Plumbr (Java memory leak detector) https://plumbr.eu/ jmap (heap histogram and heap dump generation) http://www.oracle.com/technetwork/java/javase/tooldescr-136044.html#gbdid JVM verbose:class logs JVM argument : -verbose:class IBM Java core file analysis (via kill -3 <PID>)

JVM memory profiling and heap capacity sizing	Oracle Java VisualVM and Java Mission Control IBM Monitoring and Diagnostic Tools for Java Java profilers (JProfiler, YourKit) http://en.wikipedia.org/wiki/JProfiler http://www.yourkit.com/ Memory Analyzer (heap dump and application memory footprint analysis)
JVM and middleware concurrency troubleshooting such as thread lock contention and deadlocks	Oracle Java VisualVM and Oracle Java Mission Control (threads monitoring, thread dump snapshots) jstack, native OS signal such as kill -3 (thread dump snapshots) http://www.oracle.com/technetwork/java/javase/tooldescr-136044.html#gblfh IBM Monitoring and Diagnostic Tools for Java NOTE: Proper knowledge on how to perform a JVM thread dump analysis is highly recommended
Java application clock time analysis and profiling	Oracle Java VisualVM and Oracle Java Mission Control (build-in profiler, sampler and recorder) Java profilers (JProfiler, YourKit)
Java application and threads CPU burn analysis	Oracle Java VisualVM and Oracle Java Mission Control (CPU profiler) Java profilers (JProfiler, YourKit) NOTE: You can also fall back on JVM thread dump and OS CPU per Thread analysis, if necessary
Java IO and remoting contention analysis, including timeout management assessment and tuning	Oracle Java VisualVM and Oracle Java Mission Control (threads monitoring, thread dump snapshots) jstack, native OS signal such as kill -3 (thread dump snapshots) IBM Monitoring and Diagnostic Tools for Java NOTE: Proper knowledge on how to perform a JVM thread dump analysis is highly recommended
Middleware, Java EE container tuning such as threads, JDBC data sources and more.	Oracle Java VisualVM and Oracle Java Mission Control (extra focus on exposed Java EE container runtime MBeans) Java EE container administration and management console

ABOUT THE AUTHOR

Pierre-Hugues ('P-H') Charbonneau has worked as an IT architect for CGI Canada for the last 11 years. He specializes in production system troubleshooting, middleware, JVM tuning and capacity analysis. P-H is the creator and primary author of Java EE Support Patterns, a technology blog dedicated to Java EE and Java technologies, focusing on memory leaks, coding best practices, anti-patterns, and troubleshooting techniques. He also runs a YouTube channel offering interactive videos and tutorials on Java and middleware. In his free time he enjoys cinema, sports, nutrition and spending quality time with his family.

RECOMMENDED BOOK

Java Performance is the most comprehensive book on the subject, covering all aspects of performance tuning at every stage. Includes sections on JVM performance monitoring, systematic profiling, HotSpot tuning, bencharking, web application and services performance, Enterprise Java Beans performance, tips and tricks, and more.

BUY NOW

BROWSE OUR COLLECTION OF 250+ FREE RESOURCES, INCLUDING:

RESEARCH GUIDES: Unbiased insight from leading tech experts
REFCARDZ: Library of 200+ reference cards covering the latest tech topics
COMMUNITIES: Share links, author articles, and engage with other tech experts

JOIN NOW

DZone