<http://www.oracle.com/webfolder/technetwork/tutorials/obe/java/gc01/index.html>

### Java Overview

Java is a programming language and computing platform first released by Sun Microsystems in 1995. Java is composed of a number of key components that, as a whole, create the Java platform.

### Java Runtime Environment

Java Runtime Environment (JRE) consists of the Java Virtual Machine (JVM), Java platform core classes, and supporting Java platform libraries. All three are required to run Java applications on your computer.

### Java Programming Language

Java is an object-oriented programming language that includes the following features:

a) Platform Independence - Java applications are compiled into bytecode which is stored in class files and loaded in a JVM. Since applications run in a JVM, they can be run on many different operating systems and devices.

b) Object-Oriented - Java is an object-oriented language that take many of the features of C and C++ and improves upon them.

c) Automatic Garbage Collection - Java automatically allocates and deallocates memory so programs are not burdened with that task.

d) Rich Standard Library - Java includes a vast number of premade objects that can be used to perform such tasks as input/output, networking, and date manipulation.

### Java Development Kit

Java Development Kit (JDK) is a collection of tools for developing Java applications. With the JDK, you can compile programs written in the Java Programming language and run them in a JVM. In addition, the JDK provides tools for packaging and distributing your applications.

The JDK and the JRE share the Java Application Programming Interfaces (Java API). The Java API is a collection of prepackaged libraries developers use to create Java applications. The Java API makes development easier by providing the tools to complete many common programming tasks including string manipulation, date/time processing, networking, and implementing data structures (e.g., lists, maps, stacks, and queues).

### Java Virtual Machine

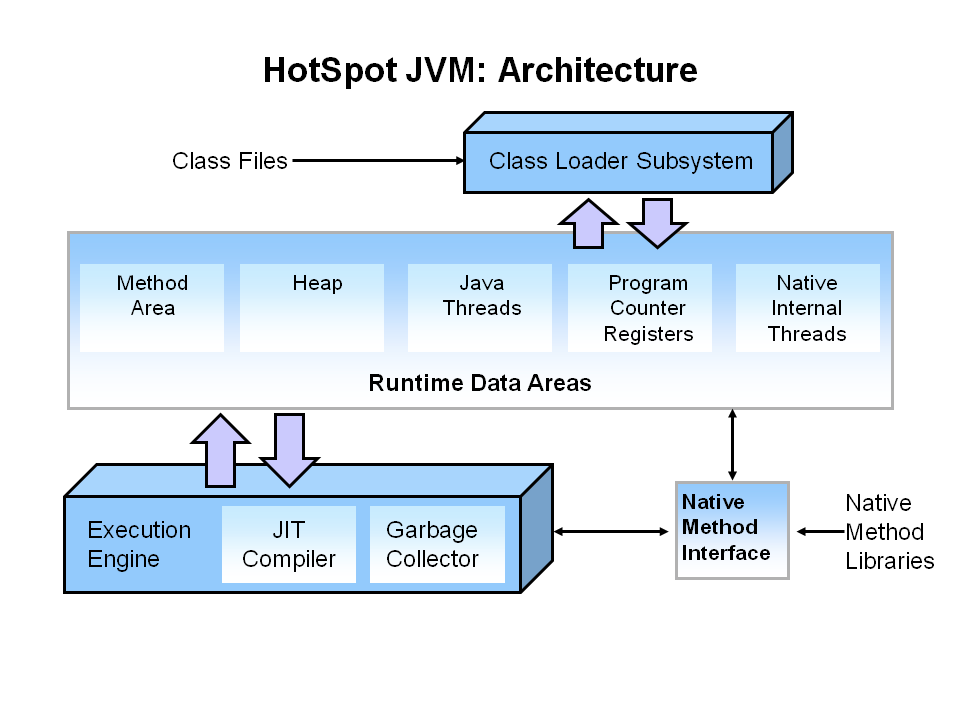
The Java Virtual Machine (JVM) is an abstract computing machine. The JVM is a program that looks like a machine to the programs written to execute in it. This way, Java programs are written to the same set of interfaces and libraries. Each JVM implementation for a specific operating system, translates the Java programming instructions into instructions and commands that run on the local operating system. This way, Java programs achieve platform independence.

The Java virtual machine knows nothing of the Java programming language, only of a particular binary format, the class file format. A class file contains Java virtual machine instructions (or bytecodes) and a symbol table, as well as other ancillary information.

For the sake of security, the Java virtual machine imposes strong syntactic and structural constraints on the code in a class file. However, any language with functionality that can be expressed in terms of a valid class file can be hosted by the Java virtual machine.

### Hotspot Architecture

The HotSpot JVM possesses an architecture that supports a strong foundation of features and capabilities and supports the ability to realize high performance and massive scalability. For example, the HotSpot JVM JIT compilers generate dynamic optimizations. In other words, they make optimization decisions while the Java application is running and generate high-performing native machine instructions targeted for the underlying system architecture. In addition, through the maturing evolution and continuous engineering of its runtime environment and multithreaded garbage collector, the HotSpot JVM yields high scalability on even the largest available computer systems.

There are three components of the JVM that are focused on when tuning performance. The Heap is where your object data is stored. This area is then managed by the Garbage Collector selected at startup. Most tuning options relate to sizing the heap and choosing the most appropriate garbage collector for your situation. The JIT compiler also has a big impact on performance but rarely requires tuning with the newer versions of the JVM.

### Performance Basics

Typically, when tuning a Java application, the focus is on one of two main goals: responsiveness or throughput.

### Responsiveness

Responsiveness refers to how quickly an application or system responds with a requested piece of data. Examples include:

* How quickly a desktop UI responds to an event
* How fast a website returns a page
* How fast a database query is returned

For applications that focus on responsiveness, large pause times are not acceptable. The focus is on responding in short periods of time.

### Throughput

Throughput focuses on maximizing the amount of work by an application in a specific period of time. Examples of how throughput might be measured include:

* The number of transactions completed in a given time.
* The number of jobs that a batch program can complete in an hour.
* The number of database queries that can be completed in an hour.

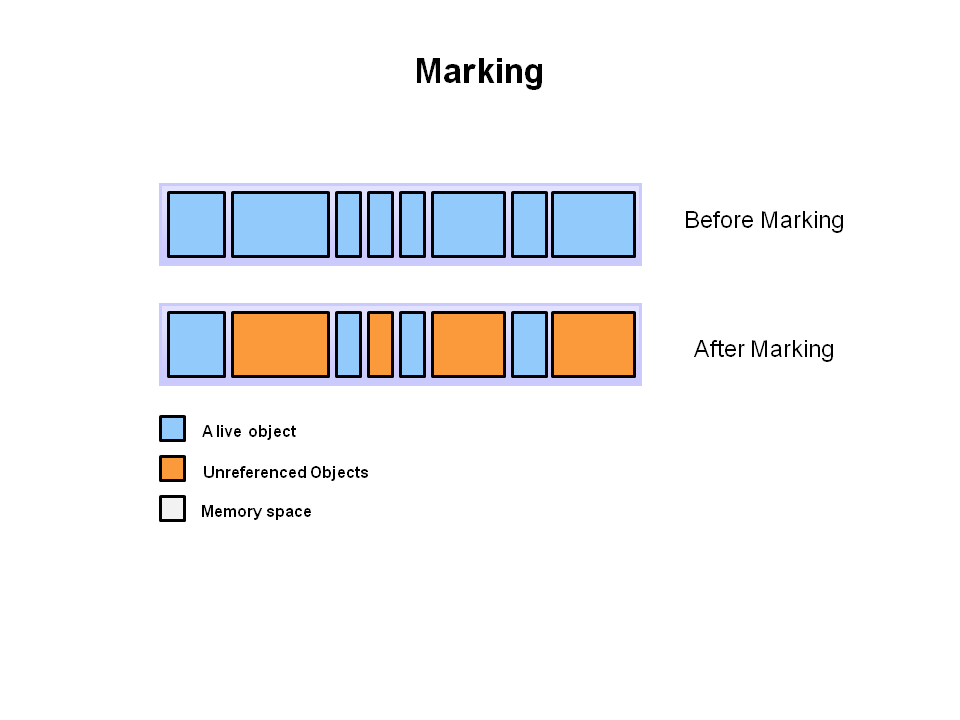
High pause times are acceptable for applications that focus on throughput. Since high throughput applications focus on benchmarks over longer periods of time, quick response time is not a consideration.

### Automatic Garbage Collection

Automatic garbage collection is the process of looking at heap memory, identifying which objects are in use and which are not, and deleting the unused objects. An in use object, or a referenced object, means that some part of your program still maintains a pointer to that object. An unused object, or unreferenced object, is no longer referenced by any part of your program. So the memory used by an unreferenced object can be reclaimed.

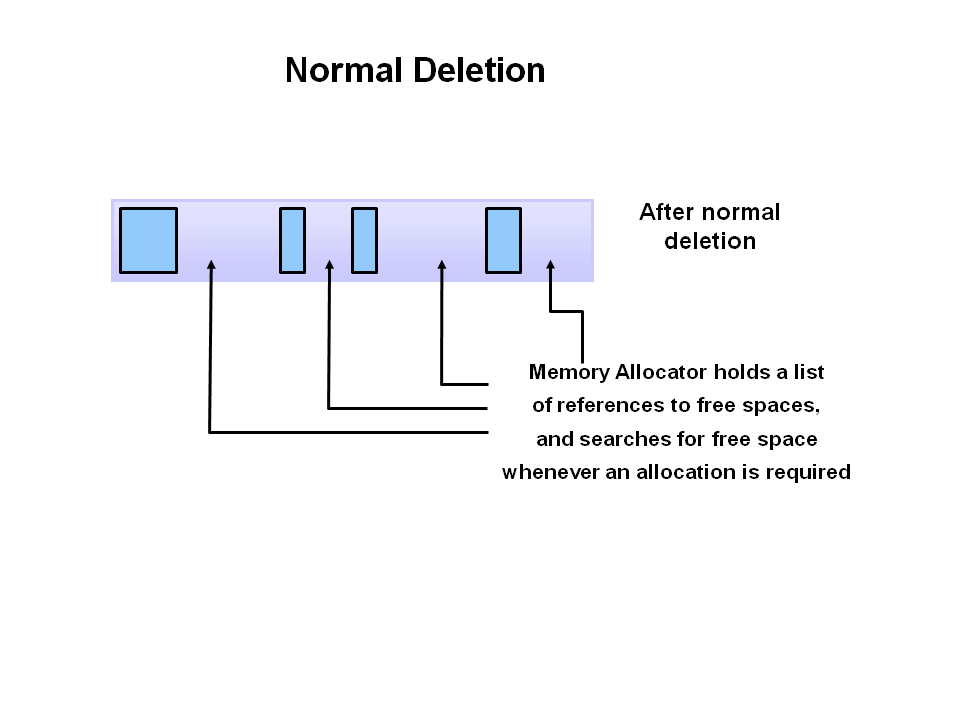
#### Step 1: Marking

The first step in the process is called marking. This is where the garbage collector identifies which pieces of memory are in use and which are not.

Referenced objects are shown in blue. Unreferenced objects are shown in gold. All objects are scanned in the marking phase to make this determination. This can be a very time consuming process if all objects in a system must be scanned.

#### Step 2: Normal Deletion

Normal deletion removes unreferenced objects leaving referenced objects and pointers to free space.

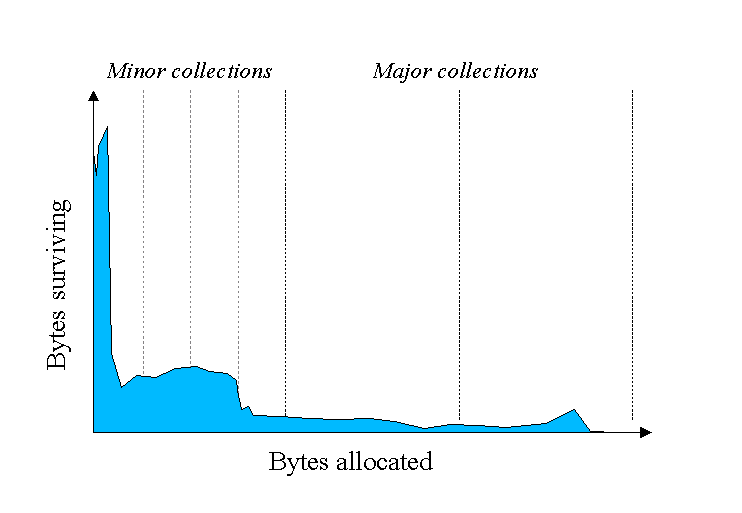
The memory allocator holds references to blocks of free space where new object can be allocated.

#### Step 2a: Deletion with Compacting

To further improve performance, in addition to deleting unreferenced objects, you can also compact the remaining referenced objects. By moving referenced object together, this makes new memory allocation much easier and faster.

### Generational Garbage Collection

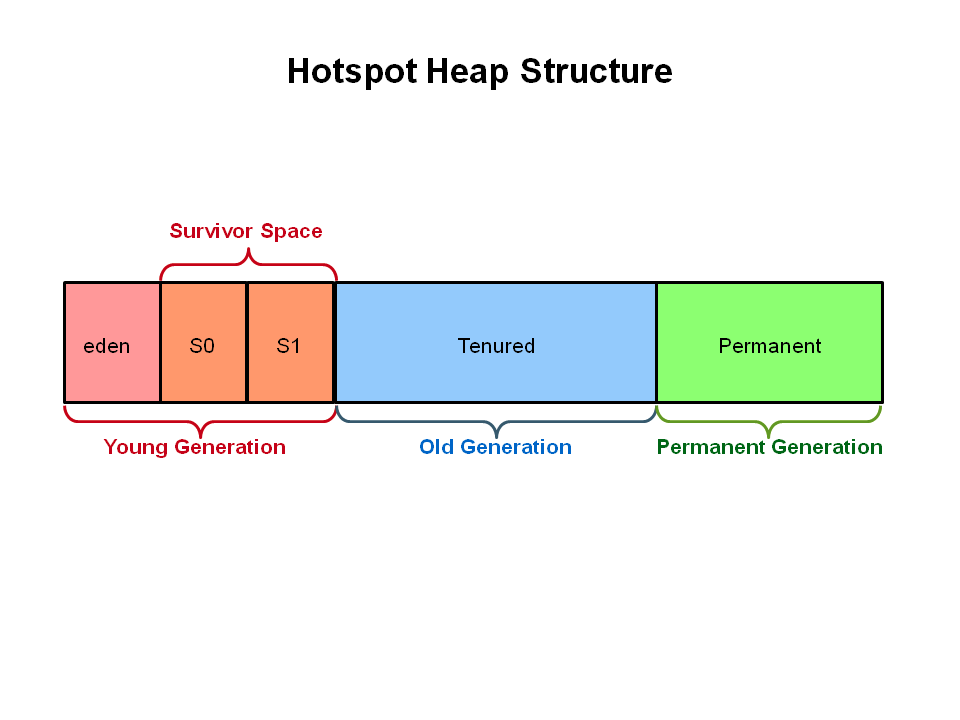
As stated earlier, having to mark and compact all the objects in a JVM is inefficient. As more and more objects are allocated, the list of objects grows and grows leading to longer and longer garbage collection time. However, empirical analysis of applications has shown that most objects are short lived.

Here is an example of such data. The Y axis shows the number of bytes allocated and the X access shows the number of bytes allocated over time.

As you can see, fewer and fewer objects remain allocated over time. In fact most objects have a very short life as shown by the higher values on the left side of the graph.

### JVM Generations

The information learned from the object allocation behavior can be used to enhance the performance of the JVM. Therefore, the heap is broken up into smaller parts or generations. The heap parts are: Young Generation, Old or Tenured Generation, and Permanent Generation.

The **Young Generation** is where all new objects are allocated and aged. When the young generation fills up, this causes a **minor garbage collection**. Minor collections can be optimized assuming a high object mortality rate. A young generation full of dead objects is collected very quickly. Some surviving objects are aged and eventually move to the old generation.

**Stop the World Event** - All minor garbage collections are "Stop the World" events. This means that all application threads are stopped until the operation completes. Minor garbage collections are always Stop the World events.

The **Old Generation** is used to store long surviving objects. Typically, a threshold is set for young generation object and when that age is met, the object gets moved to the old generation. Eventually the old generation needs to be collected. This event is called a **major garbage collection**.

Major garbage collection are also Stop the World events. Often a major collection is much slower because it involves all live objects. So for Responsive applications, major garbage collections should be minimized. Also note, that the length of the Stop the World event for a major garbage collection is affected by the kind of garbage collector that is used for the old generation space.

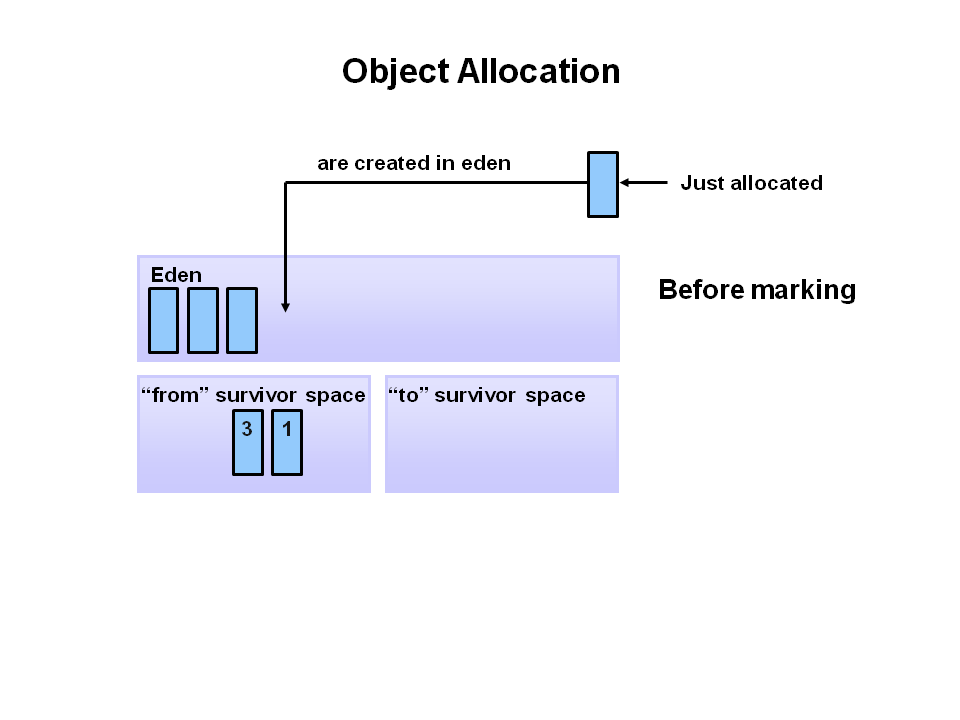
The **Permanent generation** contains metadata required by the JVM to describe the classes and methods used in the application. The permanent generation is populated by the JVM at runtime based on classes in use by the application. In addition, Java SE library classes and methods may be stored here.

Classes may get collected (unloaded) if the JVM finds they are no longer needed and space may be needed for other classes. The permanent generation is included in a full garbage collection.

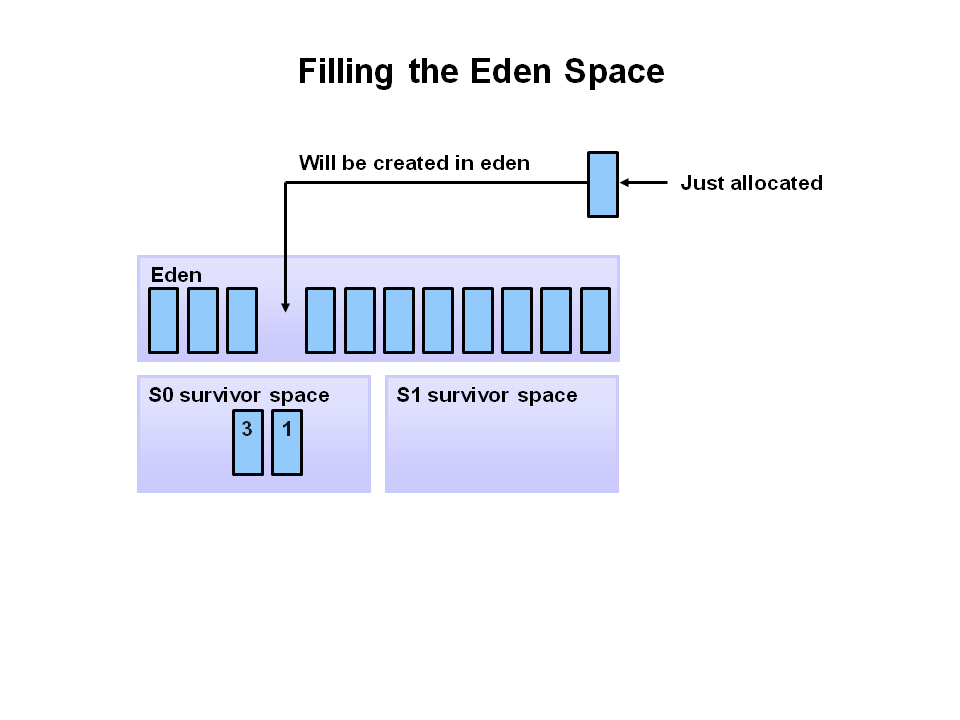
### Generational Garbage Collection Process

Now that you understand why the heap is separted into different generations, it is time to look at how exactly these spaces interact. The pictures that follow walks through the object allocation and aging process in the JVM.

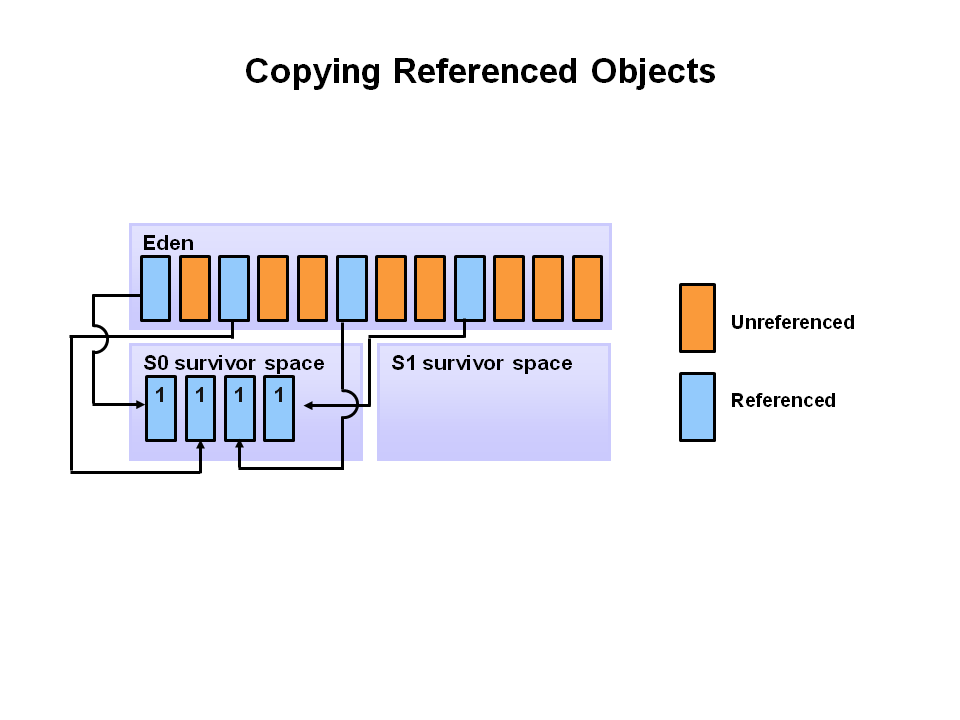
1. First, any new objects are allocated to the eden space. Both survivor spaces start out empty.

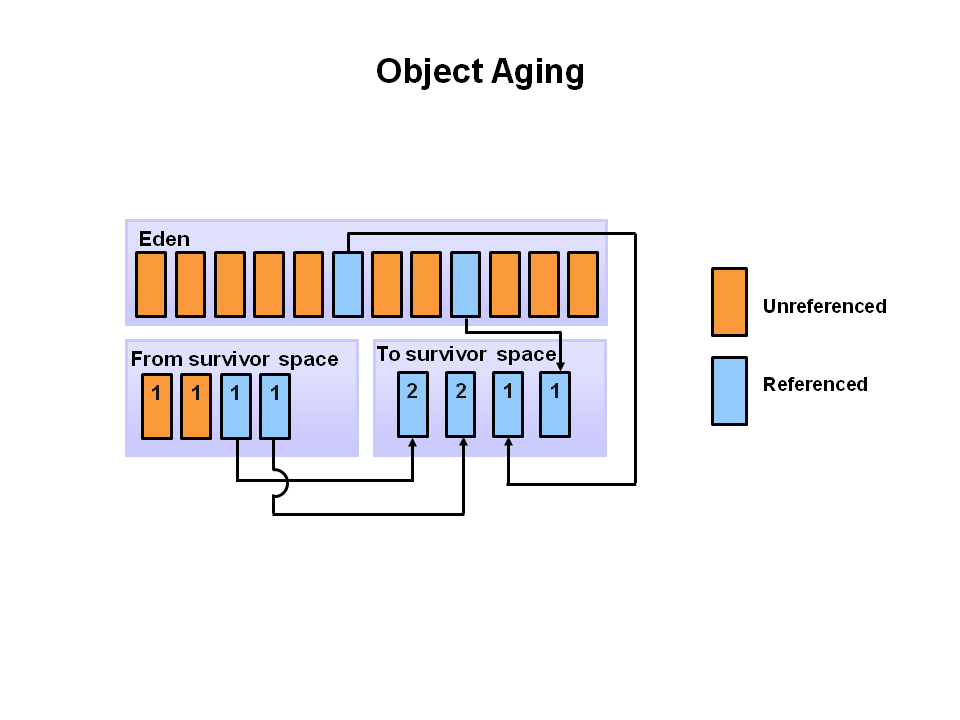


2. When the eden space fills up, a minor garbage collection is triggered.

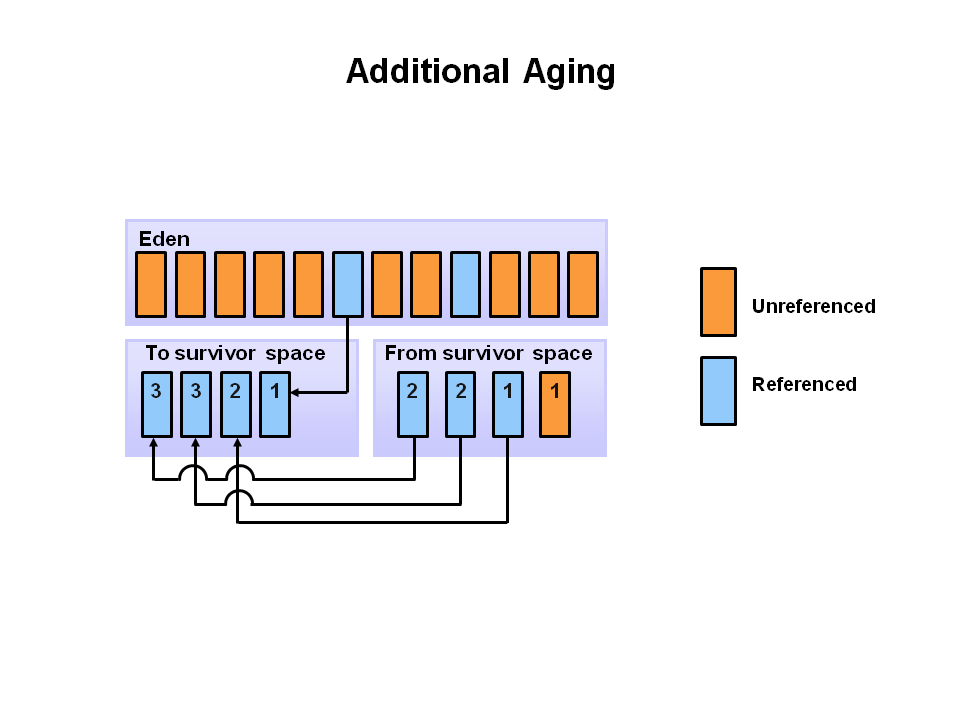


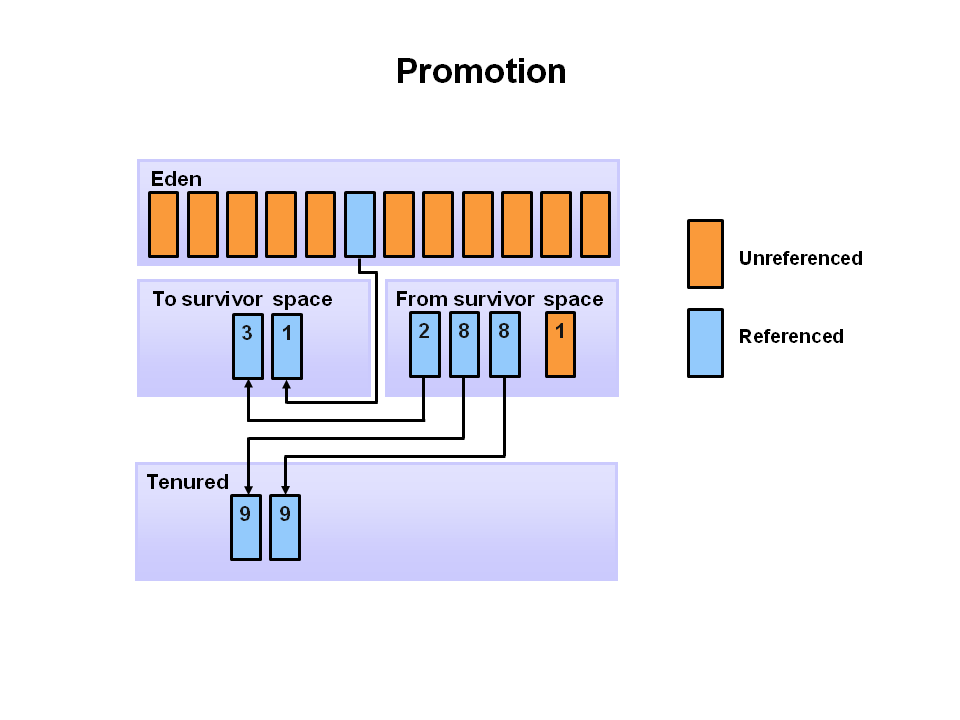
3. Referenced objects are moved to the first survivor space. Unreferenced objects are deleted when the eden space is cleared.

4. At the next minor GC, the same thing happens for the eden space. Unreferenced objects are deleted and referenced objects are moved to a survivor space. However, in this case, they are moved to the second survivor space (S1). In addition, objects from the last minor GC on the first survivor space (S0) have their age incremented and get moved to S1. Once all surviving objects have been moved to S1, both S0 and eden are cleared. Notice we now have differently aged object in the survivor space.

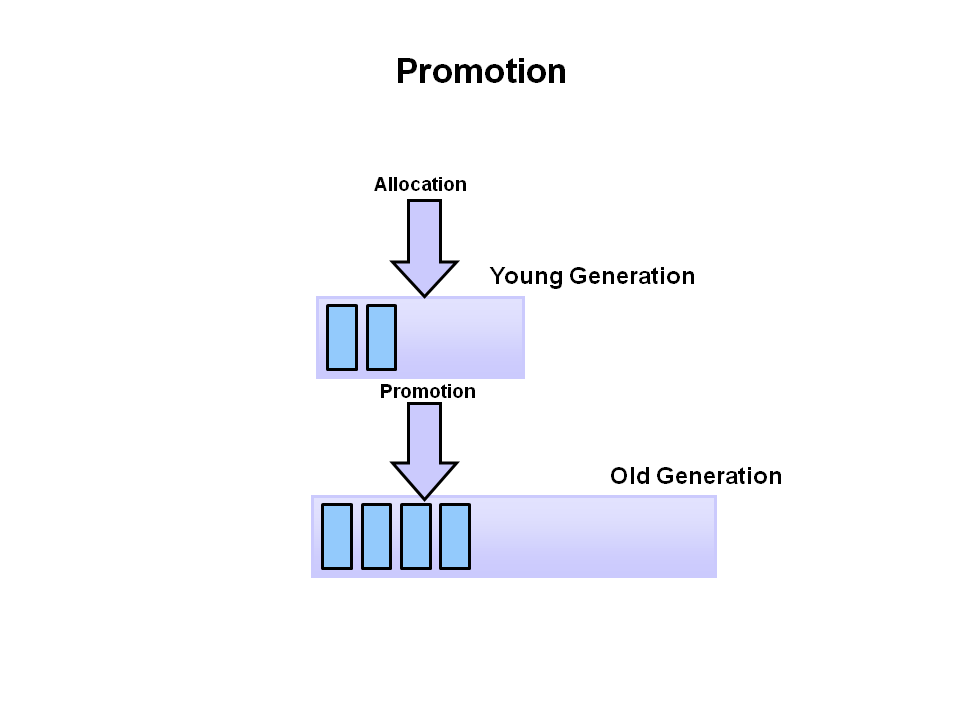


5. At the next minor GC, the same process repeats. However this time the survivor spaces switch. Referenced objects are moved to S0. Surviving objects are aged. Eden and S1 are cleared.

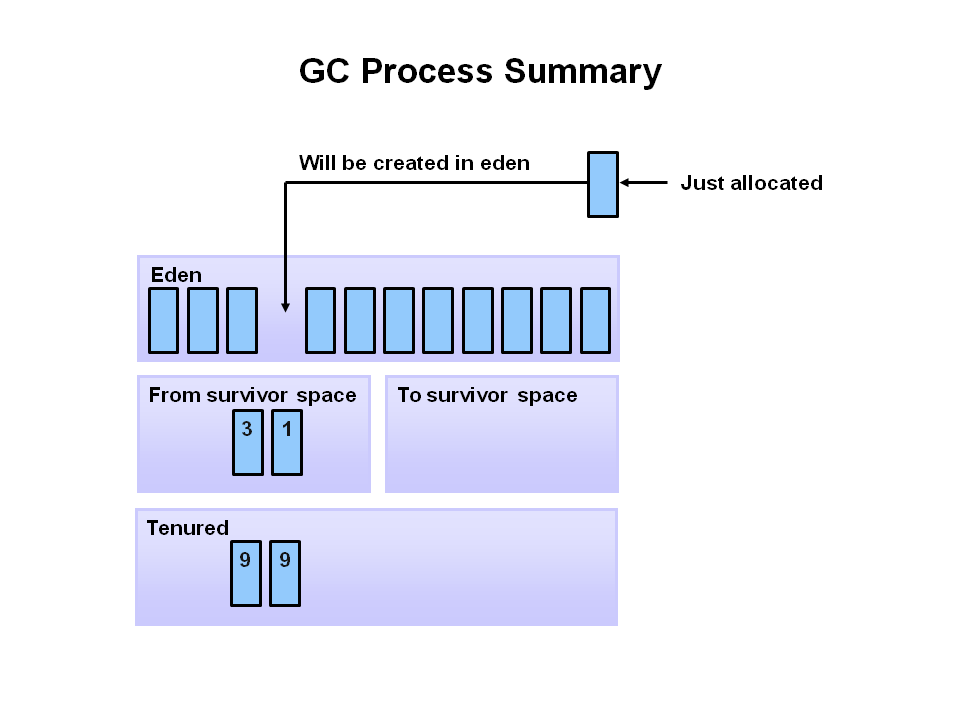
6. This slide demonstrates promotion. After a minor GC, when aged objects reach a certain age threshold (8 in this example) they are promoted from young generation to old generation.



7. As minor GCs continue to occure objects will continue to be promoted to the old generation space.



8. So that pretty much covers the entire process with the young generation. Eventually, a major GC will be performed on the old generation which cleans up and compacts that space.



### Java Garbage Collectors

In this section, we will learn about the garbage collectors available for Java and the command line switches we need to select them.

### Common Heap Related Switches

| **Switch** | **Description** |
| --- | --- |
| -Xms | Sets the initial heap size for when the JVM starts. |
| -Xmx | Sets the maximum heap size. |
| -Xmn | Sets the size of the Young Generation. |
| -XX:PermSize | Sets the starting size of the Permanent Generation. |
| -XX:MaxPermSize | Sets the maximum size of the Permanent Generation |

### The Serial GC

The serial collector is the default for client style machines in Java SE 5 and 6. With the serial collector, both minor and major garbage collections are done serially (using a single virtual CPU). In addition, it uses a mark-compact collection method. This method moves older memory to the beginning of the heap so that new memory allocations are made into a single continuous chunk of memory at the end of the heap. This compacting of memory makes it faster to allocate new chunks of memory to the heap.

#### Usage Cases

The Serial GC is the garbage collector of choice for most applications that do not have low pause time requirements and run on client-style machines. It takes advantage of only a single virtual processor for garbage collection work (therefore, its name). Still, on today's hardware, the Serial GC can efficiently manage a lot of non-trivial applications with a few hundred MBs of Java heap, with relatively short worst-case pauses (around a couple of seconds for full garbage collections).

Another popular use for the Serial GC is in environments where a high number of JVMs are run on the same machine (in some cases, more JVMs than available processors!). In such environments when a JVM does a garbage collection it is better to use only one processor to minimize the interference on the remaining JVMs, even if the garbage collection might last longer. And the Serial GC fits this trade-off nicely.

To enable the Serial Collector use: **-XX:+UseSerialGC**

### The Parallel GC

The parallel garbage collector uses multiple threads to perform the young genertion garbage collection. By default on a host with N CPUs, the parallel garbage collector uses N garbage collector threads in the collection. The number of garbage collector threads can be controlled with command-line options: **-XX:ParallelGCThreads=<desired number>**

On a host with a single CPU the default garbage collector is used even if the parallel garbage collector has been requested. On a host with two CPUs the parallel garbage collector generally performs as well as the default garbage collector and a reduction in the young generation garbage collector pause times can be expected on hosts with more than two CPUs.

#### Usage Cases

The Parallel collector is also called a throughput collector. Since it can use multilple CPUs to speed up application throughput. This collector should be used when a lot of work need to be done and long pauses are acceptable. For example, batch processing like printing reports or bills or performing a large number of database queries.

The Parallel GC comes in two flavors -

#### -XX:+UseParallelGC

With thie -XX:+UseParallelGC option, we get a multi-thread young generation collector with a single-threaded old generation collector. The option also does single-threaded compaction of old generation.

#### -XX:+UseParallelOldGC

With the -XX:+UseParallelOldGC option, the GC is both a multithreaded young generation collector and multithreaded old generation collector. It is also a multithreaded compacting collector. HotSpot does compaction only in the old generation. Young generation in HotSpot is considered a copy collector; therefore, there is no need for compaction.

Compacting describes the act of moving objects in a way that there are no holes between objects. After a garbage collection sweep, there may be holes left between live objects. Compacting moves objects so that there are no remaining holes. It is possible that a garbage collector be a non-compacting collector. Therefore, the difference between a parallel collector and a parallel compacting collector could be the latter compacts the space after a garbage collection sweep. The former would not.

### Concurrent Mark Sweep (CMS) Collector

The Concurrent Mark Sweep (CMS) collector (also referred to as the concurrent low pause collector) collects the tenured generation. It attempts to minimize the pauses due to garbage collection by doing most of the garbage collection work concurrently with the application threads. Normally the concurrent low pause collector does not copy or compact the live objects. A garbage collection is done without moving the live objects. If fragmentation becomes a problem, allocate a larger heap.

**Note:** CMS collector on young generation uses the same algorithm as that of the parallel collector.

#### Usage Cases

The CMS collector should be used for applications that require low pause times and can share resources with the garbage collector. Examples include desktop UI application that respond to events, a webserver responding to a request or a database responding to queries.

To enable the CMS Collector use: **-XX:+UseConcMarkSweepGC**

To set the number of threads use: **-XX:ParallelCMSThreads=<desired number>**

[**http://www.oracle.com/technetwork/tutorials/tutorials-1876574.html**](http://www.oracle.com/technetwork/tutorials/tutorials-1876574.html)

### The G1 Garbage Collector

The Garbage-First (G1) collector is a server-style garbage collector, targeted for multi-processor machines with large memories. It meets garbage collection (GC) pause time goals with a high probability, while achieving high throughput. The G1 garbage collector is fully supported in Oracle JDK 7 update 4 and later releases.

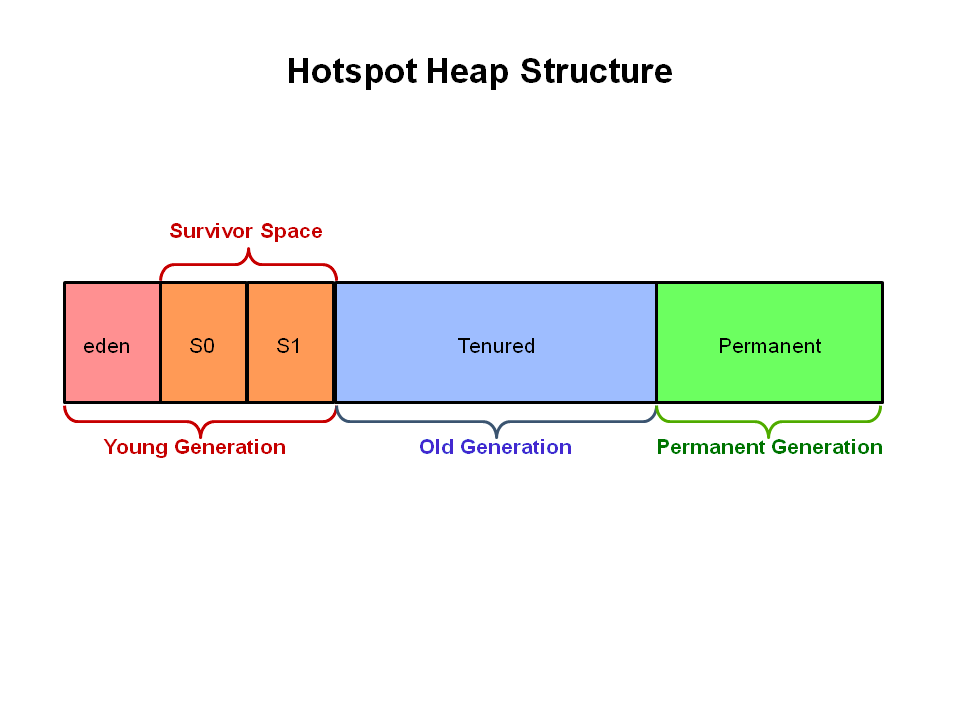
The G1 collector is designed for applications that:

* Can operate concurrently with applications threads like the CMS collector.
* Compact free space without lengthy GC induced pause times.
* Need more predictable GC pause durations.
* Do not want to sacrifice a lot of throughput performance.
* Do not require a much larger Java heap.

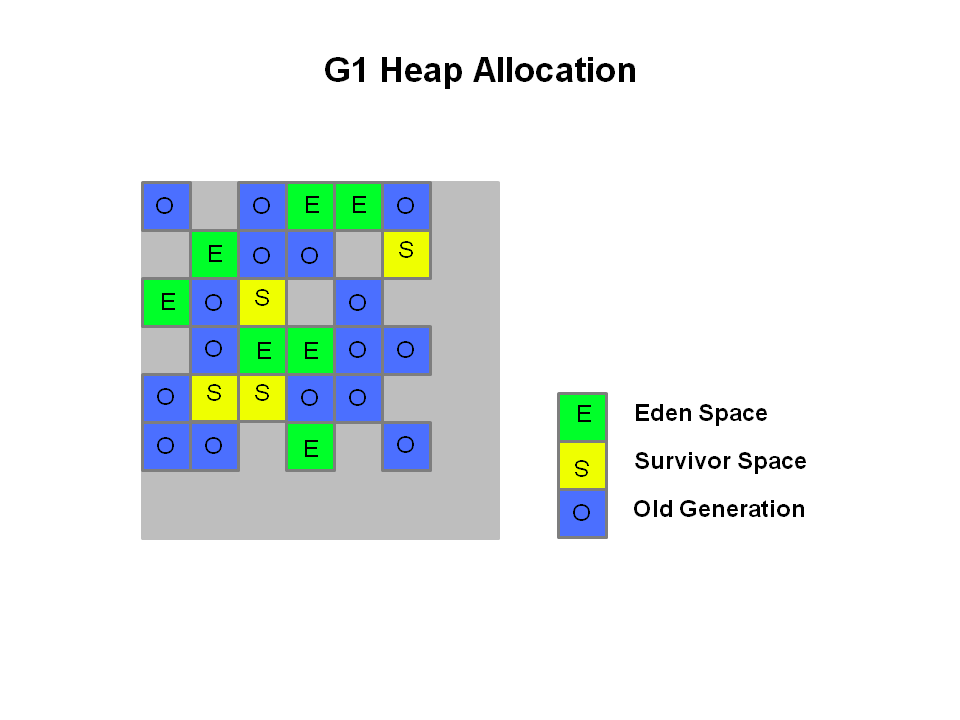
G1 is planned as the long term replacement for the Concurrent Mark-Sweep Collector (CMS). Comparing G1 with CMS, there are differences that make G1 a better solution. One difference is that G1 is a compacting collector. G1 compacts sufficiently to completely avoid the use of fine-grained free lists for allocation, and instead relies on regions. This considerably simplifies parts of the collector, and mostly eliminates potential fragmentation issues. Also, G1 offers more predictable garbage collection pauses than the CMS collector, and allows users to specify desired pause targets.

### G1 Operational Overview

The older garbage collectors (serial, parallel, CMS) all structure the heap into three sections: young generation, old generation, and permanent generation of a fixed memory size.

All memory objects end up in one of these three sections.

The G1 collector takes a different approach.

The heap is partitioned into a set of equal-sized heap regions, each a contiguous range of virtual memory. Certain region sets are assigned the same roles (eden, survivor, old) as in the older collectors, but there is not a fixed size for them. This provides greater flexibility in memory usage.

When performing garbage collections, G1 operates in a manner similar to the CMS collector. G1 performs a concurrent global marking phase to determine the liveness of objects throughout the heap. After the mark phase completes, G1 knows which regions are mostly empty. It collects in these regions first, which usually yields a large amount of free space. This is why this method of garbage collection is called Garbage-First. As the name suggests, G1 concentrates its collection and compaction activity on the areas of the heap that are likely to be full of reclaimable objects, that is, garbage. G1 uses a pause prediction model to meet a user-defined pause time target and selects the number of regions to collect based on the specified pause time target.

The regions identified by G1 as ripe for reclamation are garbage collected using evacuation. G1 copies objects from one or more regions of the heap to a single region on the heap, and in the process both compacts and frees up memory. This evacuation is performed in parallel on multi-processors, to decrease pause times and increase throughput. Thus, with each garbage collection, G1 continuously works to reduce fragmentation, working within the user defined pause times. This is beyond the capability of both the previous methods. CMS (Concurrent Mark Sweep ) garbage collector does not do compaction. ParallelOld garbage collection performs only whole-heap compaction, which results in considerable pause times.

It is important to note that G1 is not a real-time collector. It meets the set pause time target with high probability but not absolute certainty. Based on data from previous collections, G1 does an estimate of how many regions can be collected within the user specified target time. Thus, the collector has a reasonably accurate model of the cost of collecting the regions, and it uses this model to determine which and how many regions to collect while staying within the pause time target.

**Note:** G1 has both concurrent (runs along with application threads, e.g., refinement, marking, cleanup) and parallel (multi-threaded, e.g., stop the world) phases. Full garbage collections are still single threaded, but if tuned properly your applications should avoid full Gcs.

#### G1 Footprint

If we migrate from the ParallelOldGC or CMS collector to G1, you will likely see a larger JVM process size. This is largely related to "accounting" data structures such as Remembered Sets and Collection Sets.

**Remembered Sets** or RSets track object references into a given region. There is one RSet per region in the heap. The RSet enables the parallel and independent collection of a region. The overall footprint impact of RSets is less than 5%.

**Collection Sets** or CSets the set of regions that will be collected in a GC. All live data in a CSet is evacuated (copied/moved) during a GC. Sets of regions can be Eden, survivor, and/or old generation. CSets have a less than 1% impact on the size of the JVM.

#### Usage Cases

The first focus of G1 is to provide a solution for users running applications that require large heaps with limited GC latency. This means heap sizes of around 6GB or larger, and stable and predictable pause time below 0.5 seconds.

Applications running today with either the CMS or the ParallelOldGC garbage collector would benefit switching to G1 if the application has one or more of the following traits.

* + Full GC durations are too long or too frequent.
  + The rate of object allocation rate or promotion varies significantly.
  + Undesired long garbage collection or compaction pauses (longer than 0.5 to 1 second)

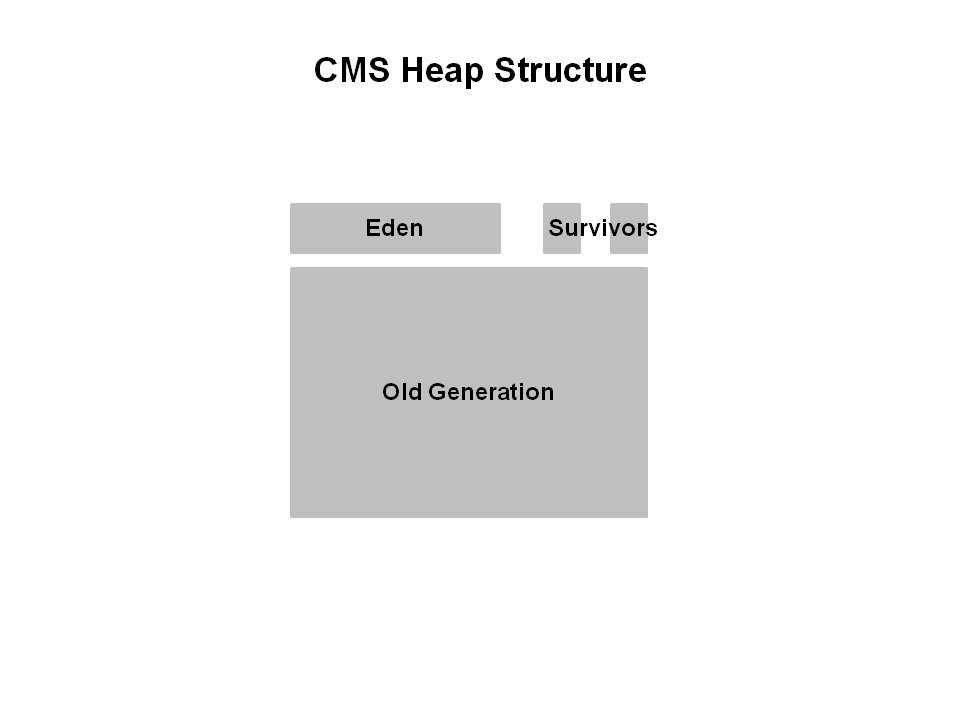
**Note:** If you are using CMS or ParallelOldGC and your application is not experiencing long garbage collection pauses, it is fine to stay with your current collector. Changing to the G1 collector is not a requirement for using the latest JDK.

### CMS Garbage Collection Steps

Next, let's review CMS Collector operations step by step.

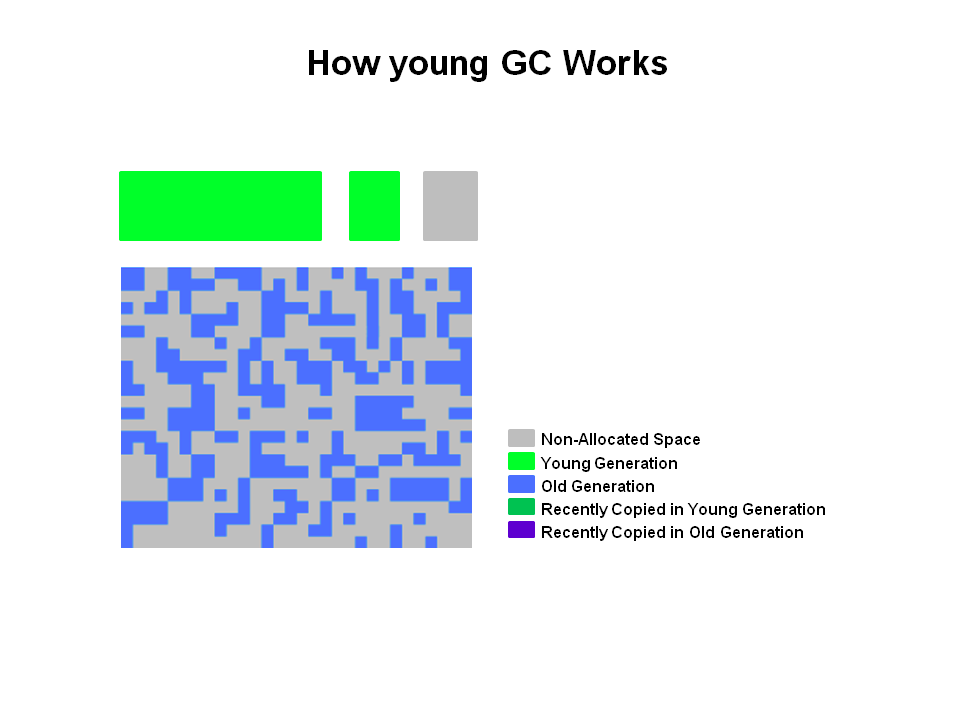
1. **Heap Structure for CMS Collector**

The heap is split into three spaces.

Young generation is split into Eden and two survivor spaces. Old generation is one contiguous space. Object collection is done in place. No compaction is done unless there is a full GC.

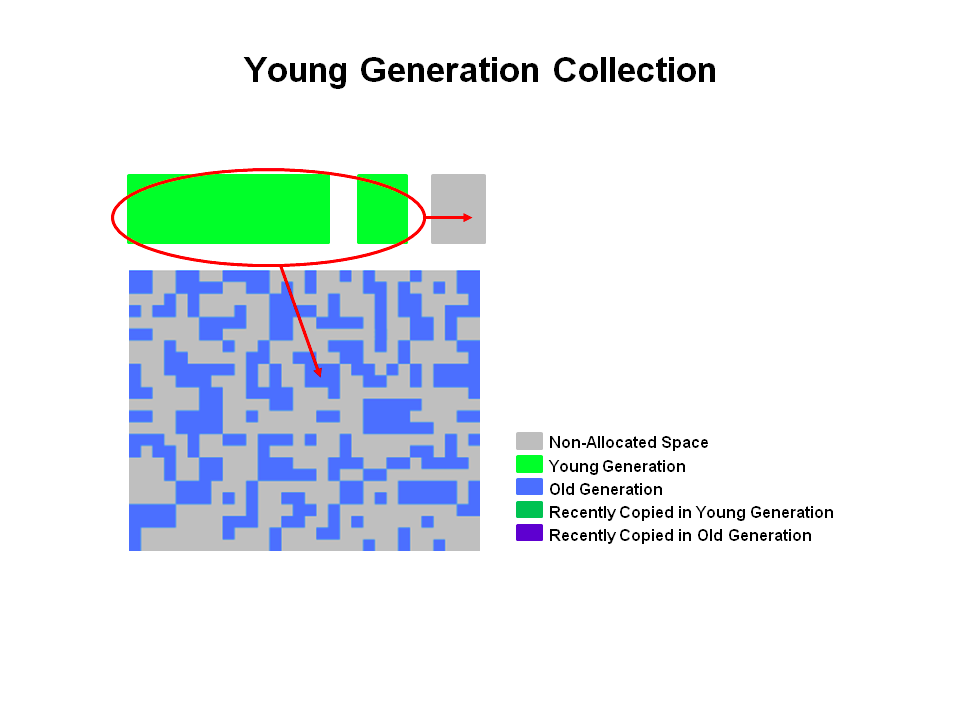
2. **How Young GC works in CMS**

The young generation is colored light green and the old generation in blue. This is what the CMS might look like if your application has been running for a while. Objects are scattered around the old generation area.

With CMS, old generation objects are deallocated in place. They are not moved around. The space is not compacted unless there is a full GC.

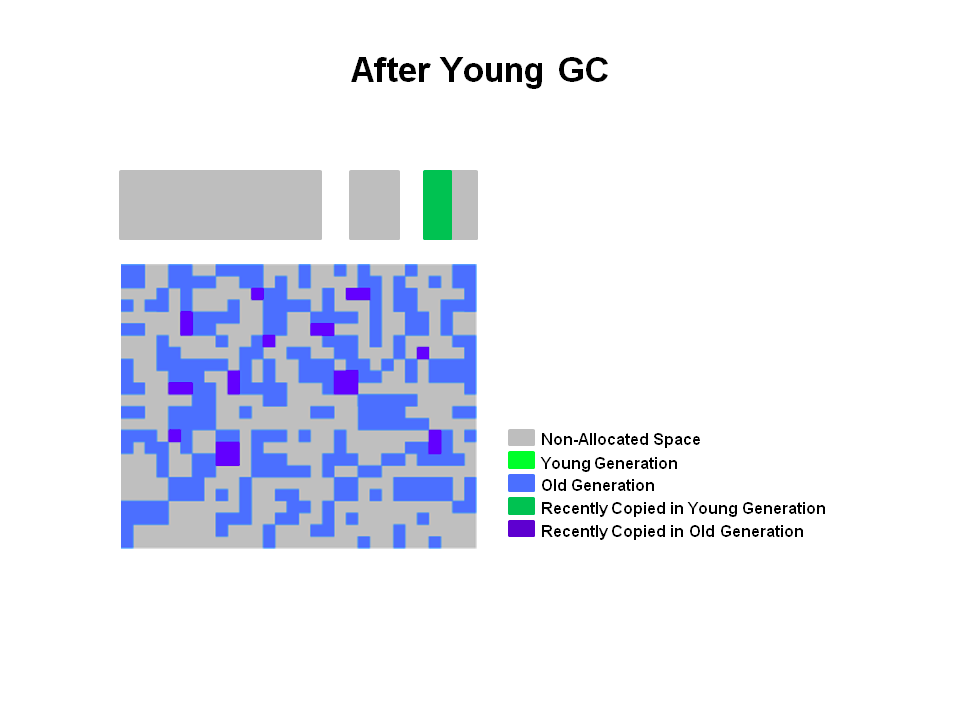
3. **Young Generation Collection**

Live objects are copied from the Eden space and survivor space to the other survivor space. Any older objects that have reached their aging threshold are promoted to old generation.



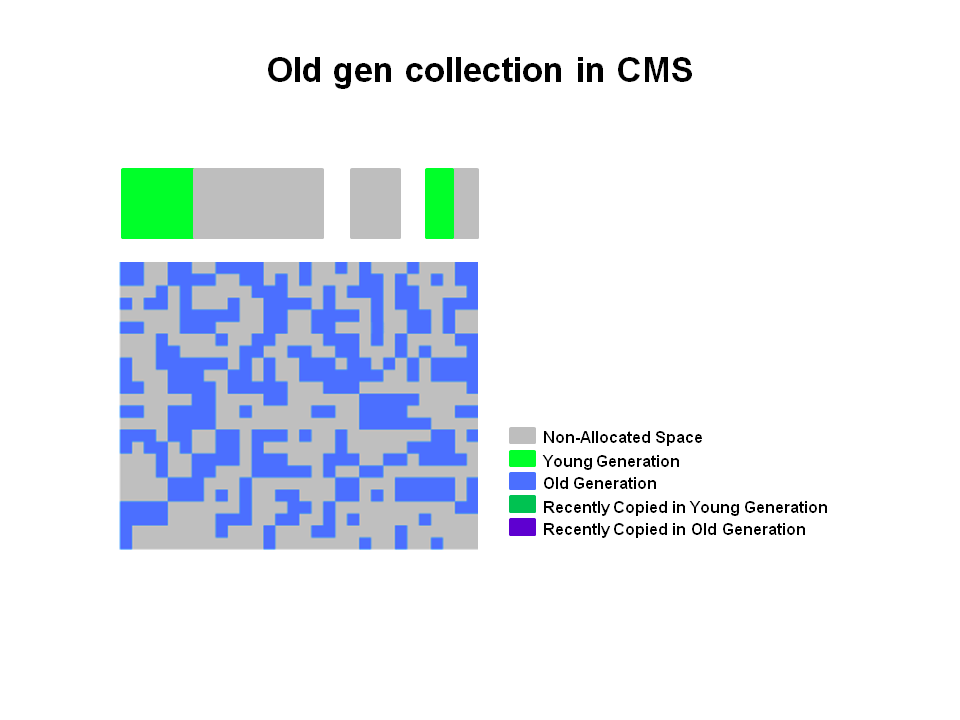
4. **After Young GC**

After a young GC, the Eden space is cleared and one of the survivor spaces is cleared.

Newly promoted objects are shown in dark blue on the diagram. The green objects are surviving young generation objects that have not yet been promoted to old generation.

5. **Old Generation Collection with CMS**

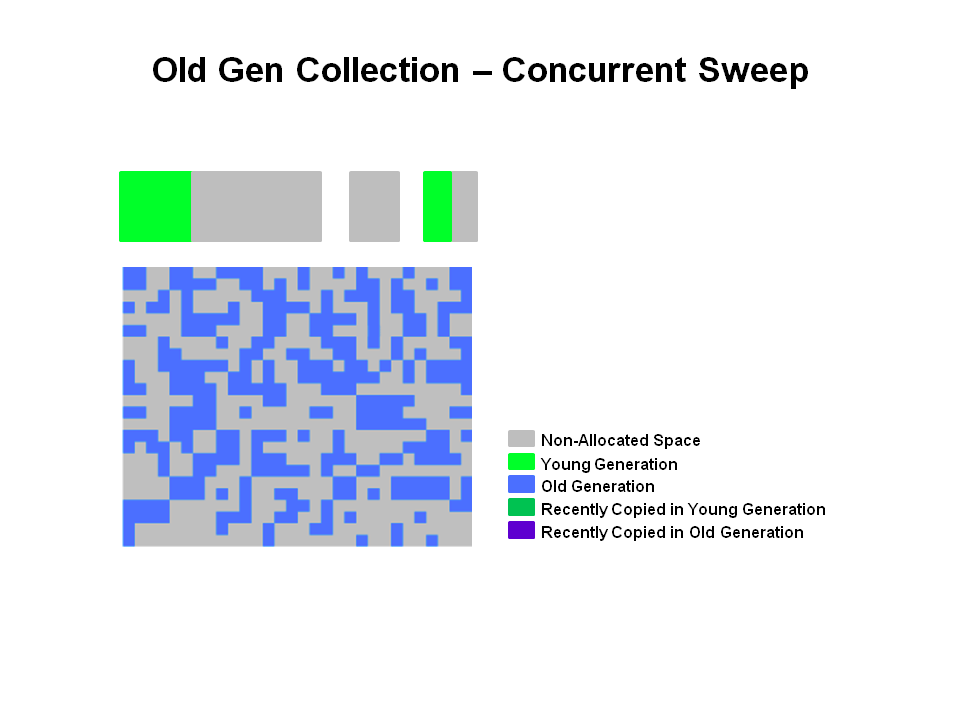
Two stop the world events take place: initial mark and remark. When the old generation reaches a certain occupancy rate, the CMS is kicked off.



(1) Initial mark is a short pause phase where live (reachable) objects are marked. (2) Concurrent marking finds live objects while the application continues to execute. Finally, in the (3) remark phase, objects are found that were missed during (2) concurrent marking in the previous phase.

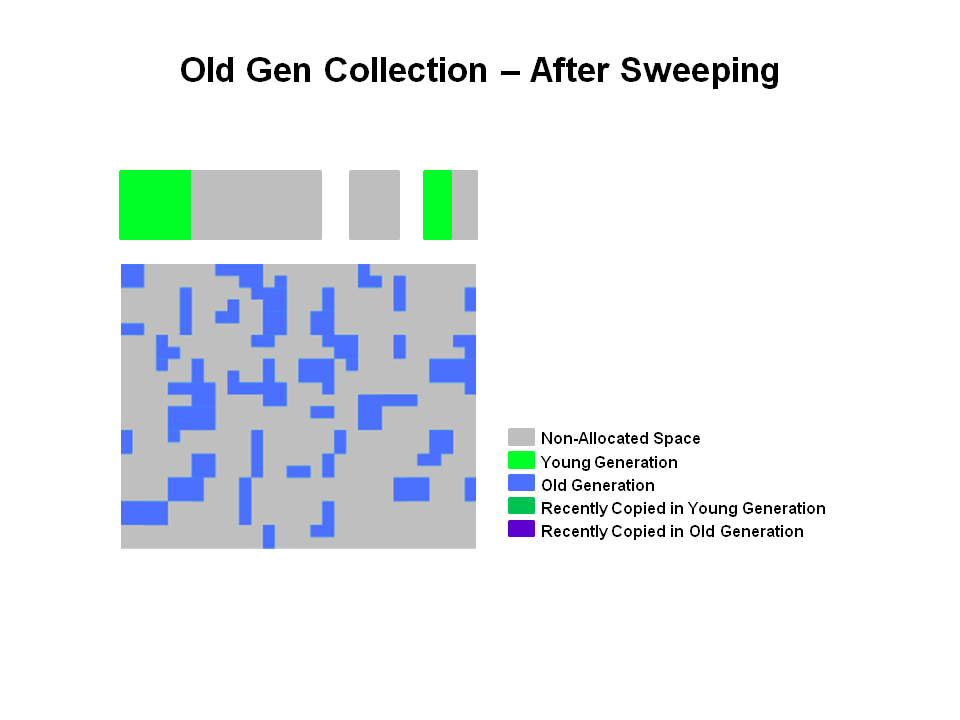
6. **Old Generation Collection - Concurrent Sweep**

Objects that were not marked in the previous phase are deallocated in place. There is no compaction.

**Note:** Unmarked objects == Dead Objects

7. **Old Generation Collection - After Sweeping**

After the (4) Sweeping phase, you can see that a lot of memory has been freed up. You will also notice that no compaction has been done.



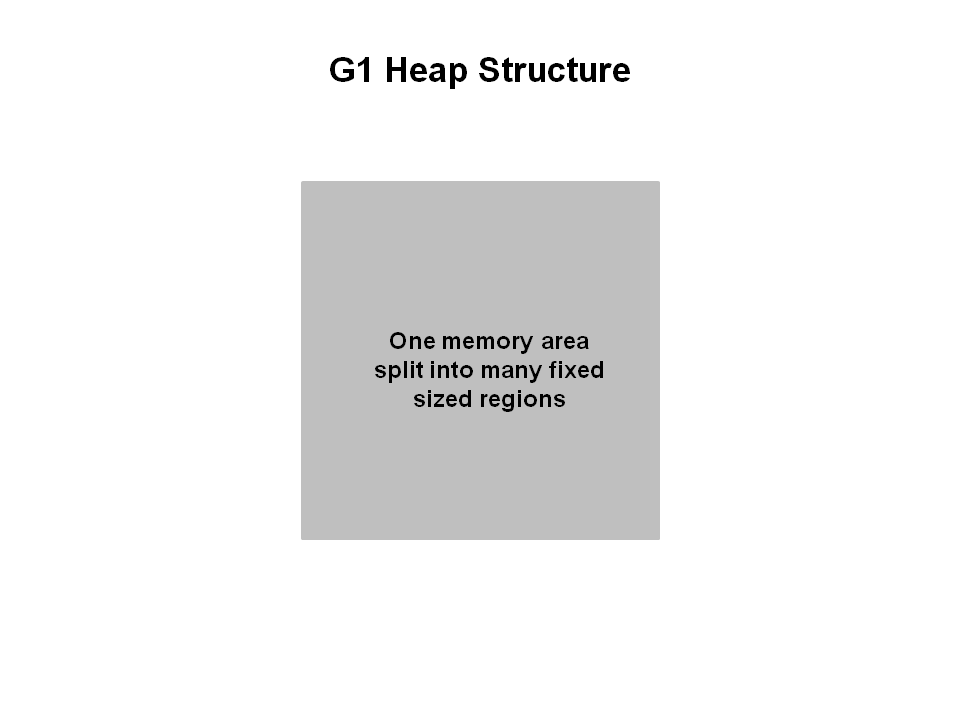
Finally, the CMS collector will move through the (5) resetting phase and wait for the next time the GC threshold is reached.

### G1 Garbage Collector Step by Step

The G1 collector takes a different approach to allocating the heap. The pictures that follow review the G1 system step by step.

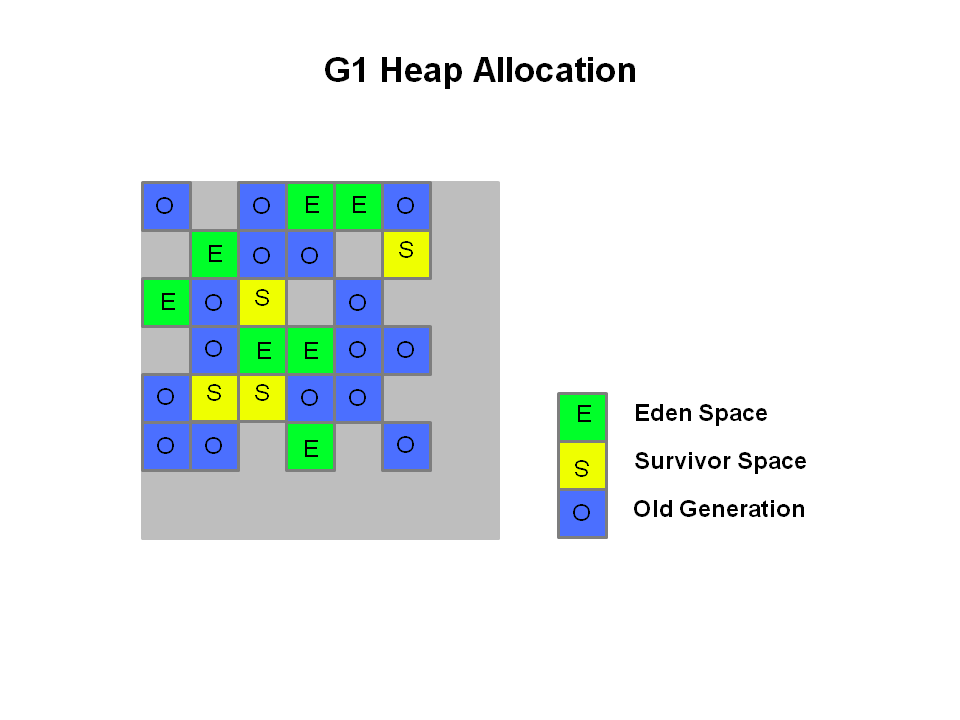
1. **G1 Heap Structure**

The heap is one memory area split into many fixed sized regions.

Region size is chosen by the JVM at startup. The JVM generally targets around 2000 regions varying in size from 1 to 32Mb.

2. **G1 Heap Allocation**

In reality, these regions are mapped into logical representations of Eden, Survivor, and old generation spaces.



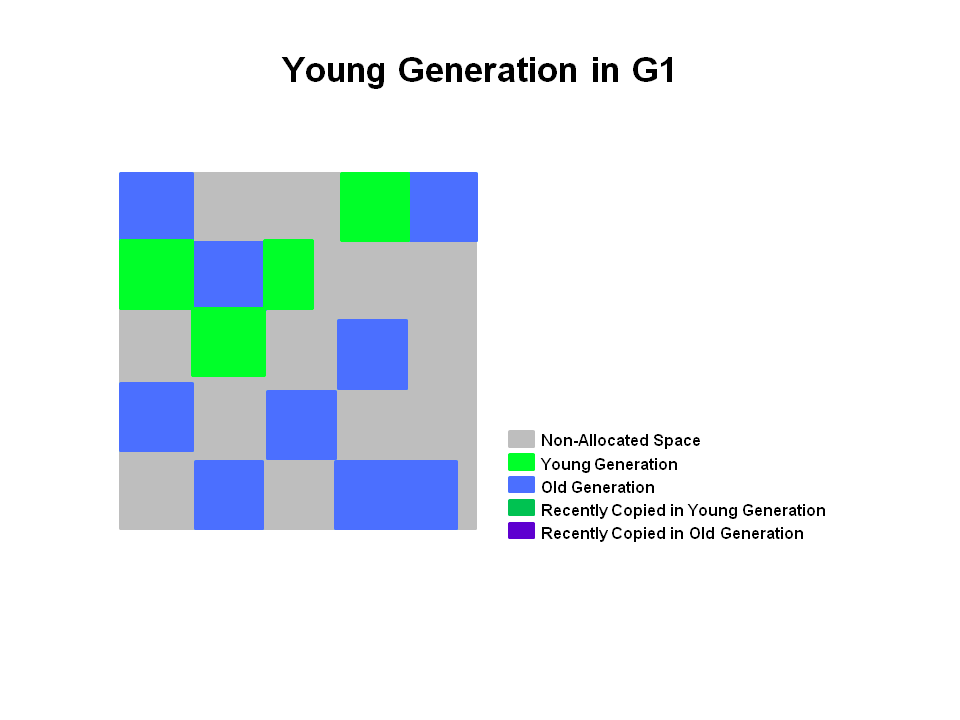
The colors in the picture shows which region is associated with which role. Live objects are evacuated (i.e., copied or moved) from one region to another. Regions are designed to be collected in parallel with or without stopping all other application threads.

As shown regions can be allocated into Eden, survivor, and old generation regions. In addition, there is a fourth type of object known as Humongous regions. These regions are designed to hold objects that are 50% the size of a standard region or larger. They are stored as a set of contiguous regions. Finally the last type of regions would be the unused areas of the heap.

**Note:** At the time of this writing, collecting humongous objects has not been optimized. Therefore, you should avoid creating objects of this size.

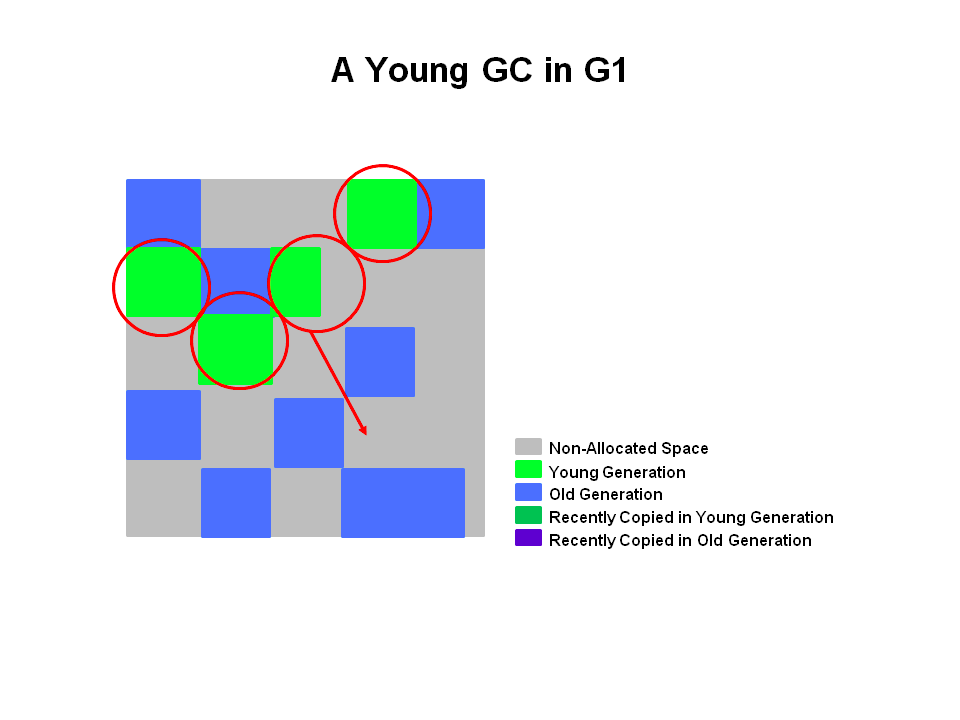
3. **Young Generation in G1**

The heap is split into approximately 2000 regions. Minimum size is 1Mb and maximum size is 32Mb. Blue regions hold old generation objects and green regions hold young generation objects.

Note that the regions are not required to be contiguous like the older garbage collectors.

4. **A Young GC in G1**

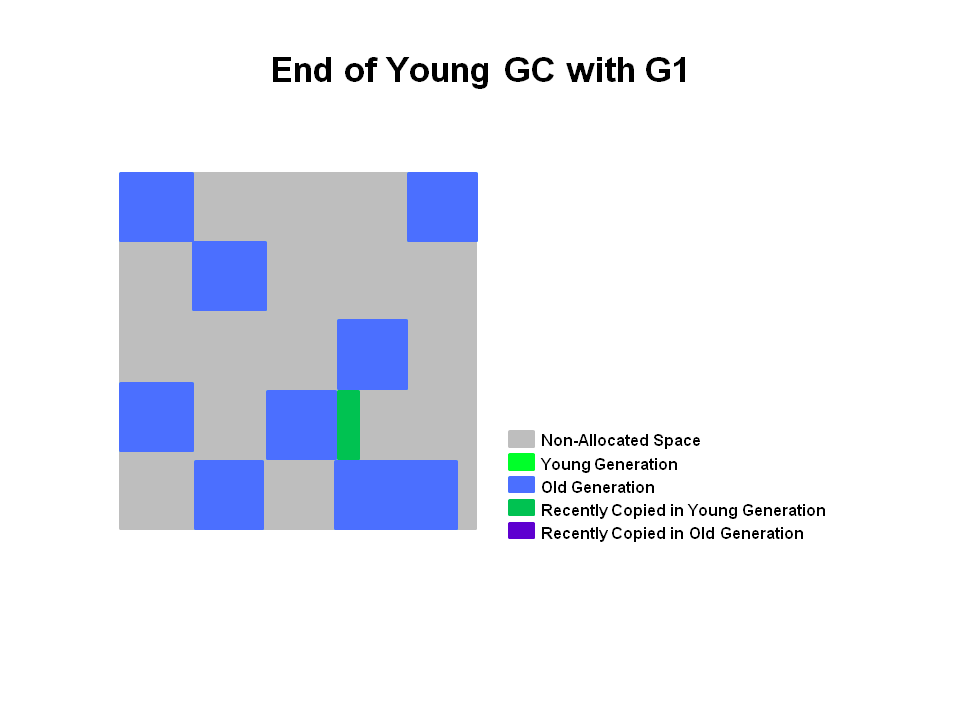
Live objects are evacuated (i.e., copied or moved) to one or more survivor regions. If the aging threshold is met, some of the objects are promoted to old generation regions.

This is a stop the world (STW) pause. Eden size and survivor size is calculated for the next young GC. Accounting information is kept to help calculate the size. Things like the pause time goal are taken into consideration.

This approach makes it very easy to resize regions, making them bigger or smaller as needed.

5. **End of a Young GC with G1**

Live objects have been evacuated to survivor regions or to old generation regions.

Recently promoted objects are shown in dark blue. Survivor regions in green.

In summary, the following can be said about the young generation in G1:

* The heap is a single memory space split into regions.
* Young generation memory is composed of a set of non-contiguous regions. This makes it easy to resize when needed.
* Young generation garbage collections, or young GCs, are stop the world events. All application threads are stopped for the operation.
* The young GC is done in parallel using multiple threads.
* Live objects are copied to new survivor or old generation regions.

### Old Generation Collection with G1

Like the CMS collector, the G1 collector is designed to be a low pause collector for old generation objects. The following table describes the G1 collection phases on old generation.

#### G1 Collection Phases - Concurrent Marking Cycle Phases

The G1 collector performs the following phases on the old generation of the heap. Note that some phases are part of a young generation collection.

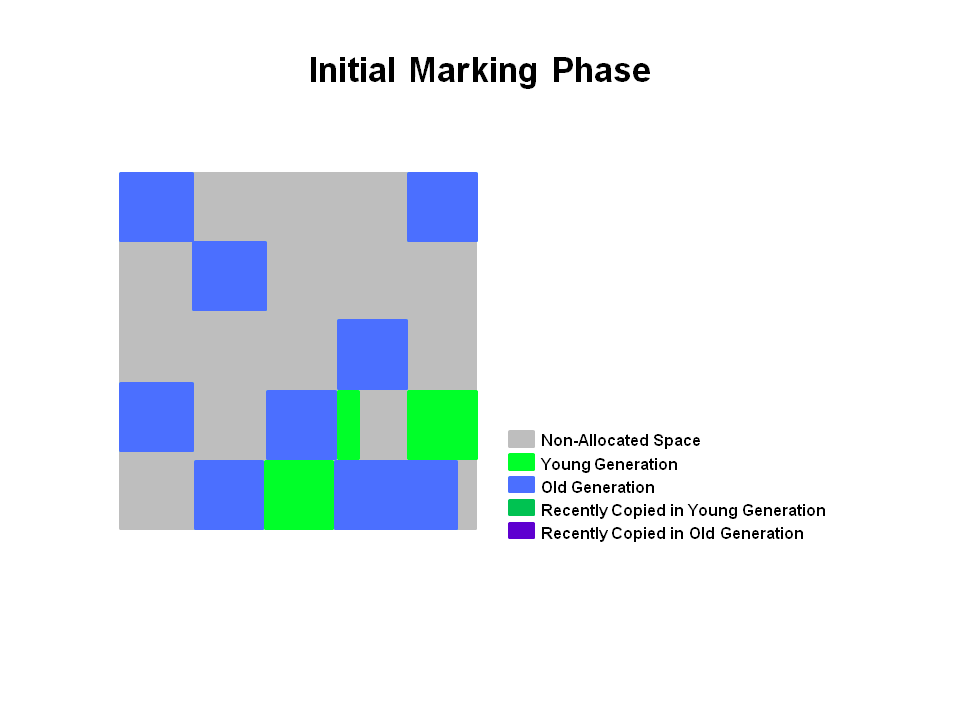
| Phase | Description |
| --- | --- |
| (1) Initial Mark *(Stop the World Event)* | This is a stop the world event. With G1, it is piggybacked on a normal young GC. Mark survivor regions (root regions) which may have references to objects in old generation. |
| (2) Root Region Scanning | Scan survivor regions for references into the old generation. This happens while the application continues to run. The phase must be completed before a young GC can occur. |
| (3) Concurrent Marking | Find live objects over the entire heap. This happens while the application is running. This phase can be interrupted by young generation garbage collections. |
| (4) Remark *(Stop the World Event)* | Completes the marking of live object in the heap. Uses an algorithm called snapshot-at-the-beginning (SATB) which is much faster than what was used in the CMS collector. |
| (5) Cleanup *(Stop the World Event and Concurrent)* | * Performs accounting on live objects and completely free regions. (Stop the world) * Scrubs the Remembered Sets. (Stop the world) * Reset the empty regions and return them to the free list. (Concurrent) |
| (\*) Copying *(Stop the World Event)* | These are the stop the world pauses to evacuate or copy live objects to new unused regions. This can be done with young generation regions which are logged as [GC pause (young)]. Or both young and old generation regions which are logged as [GC Pause (mixed)]. |

#### G1 Old Generation Collection Step by Step

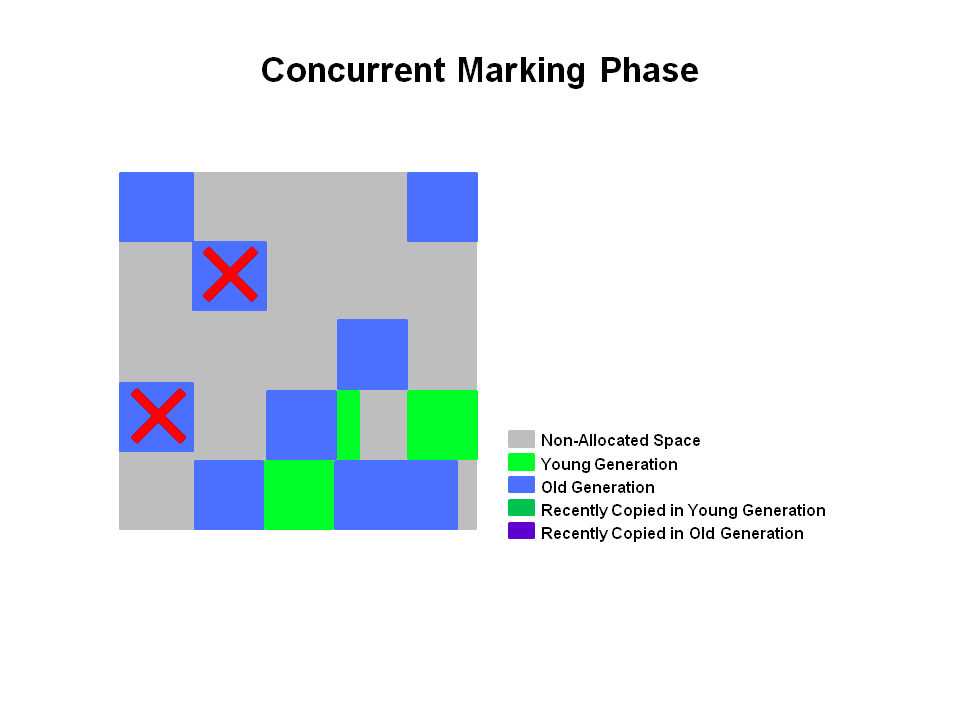
With the phases defined, let's look at how they interact with the old generation in the G1 collector.

6. **Initial Marking Phase**

Initial marking of live object is piggybacked on a young generation garbage collection. In the logs this is noted as GC pause (young)(inital-mark).

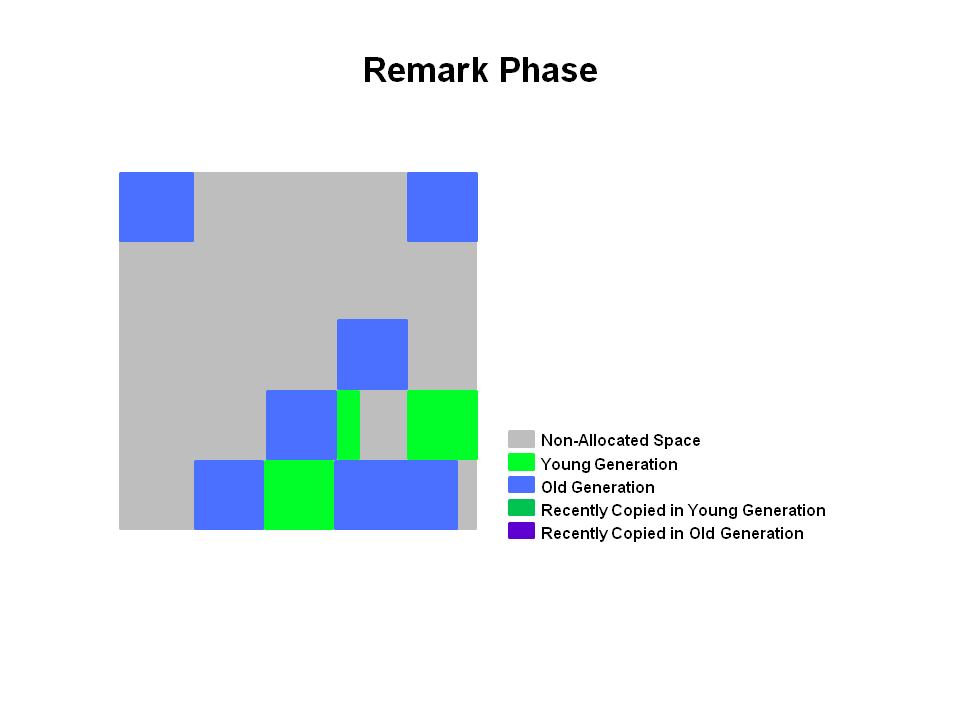
7. **Concurrent Marking Phase**

If empty regions are found (as denoted by the "X"), they are removed immediately in the Remark phase. Also, "accounting" information that determines liveness is calculated.



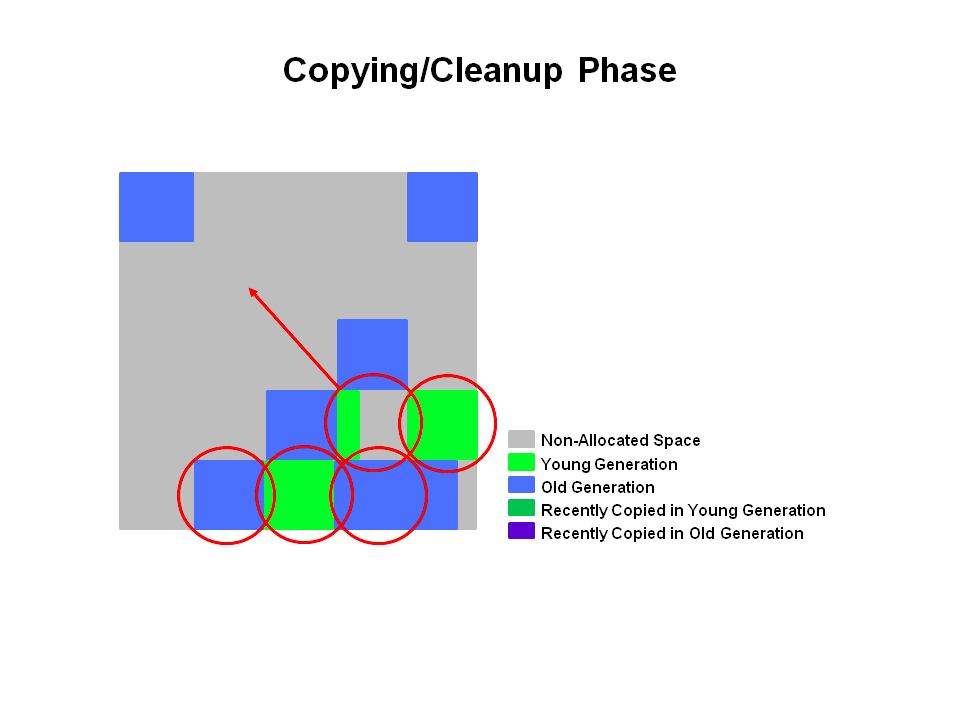
8. **Remark Phase**

Empty regions are removed and reclaimed. Region liveness is now calculated for all regions.



9. **Copying/Cleanup Phase**

G1 selects the regions with the lowest "liveness", those regions which can be collected the fastest. Then those regions are collected at the same time as a young GC. This is denoted in the logs as [GC pause (mixed)]. So both young and old generations are collected at the same time.



10. **After Copying/Cleanup Phase**

The regions selected have been collected and compacted into the dark blue region and the dark green region shown in the diagram.

#### Summary of Old Generation GC

In summary, there are a few key points we can make about the G1 garbage collection on the old generation.

* Concurrent Marking Phase
  + Liveness information is calculated concurrently while the application is running.
  + This liveness information identifies which regions will be best to reclaim during an evacuation pause.
  + There is no sweeping phase like in CMS.
* Remark Phase
  + Uses the Snapshot-at-the-Beginning (SATB) algorithm which is much faster then what was used with CMS.
  + Completely empty regions are reclaimed.
* Copying/Cleanup Phase
  + Young generation and old generation are reclaimed at the same time.
  + Old generation regions are selected based on their liveness.

To enable the G1 Collector use: **-XX:+UseG1GC**

**-XX:+UseG1GC** - Tells the JVM to use the G1 Garbage collector.

**-XX:MaxGCPauseMillis=200** - Sets a target for the maximum GC pause time. This is a soft goal, and the JVM will make its best effort to achieve it. Therefore, the pause time goal will sometimes not be met. The default value is 200 milliseconds.

**-XX:InitiatingHeapOccupancyPercent=45** - Percentage of the (entire) heap occupancy to start a concurrent GC cycle. It is used by G1 to trigger a concurrent GC cycle based on the occupancy of the entire heap, not just one of the generations. A value of 0 denotes 'do constant GC cycles'. The default value is 45 (i.e., 45% full or occupied).

#### Best Practices

**Do not Set Young Generation Size**

Explicitly setting young generation size via -Xmn meddles with the default behavior of the G1 collector.

* G1 will no longer respect the pause time target for collections. So in essence, setting the young generation size disables the pause time goal.
* G1 is no longer able to expand and contract the young generation space as needed. Since the size is fixed, no changes can be made to the size.

**Response Time Metrics**

Instead of using average response time (ART) as a metric to set the XX:MaxGCPauseMillis=<N>, consider setting value that will meet the goal 90% of the time or more. This means 90% of users making a request will not experience a response time higher than the goal. Remember, the pause time is a goal and is not guaranteed to always be met.

**What is an Evacuation Failure?**

A promotion failure that happens when a JVM runs out of heap regions during the GC for either survivors and promoted objects. The heap can't expand because it is already at max. This is indicated in the GC logs when using -XX:+PrintGCDetails by **to-space overflow**. This is expensive!

* GC still has to continue so space has to be freed up.
* Unsuccessfully copied objects have to be tenured in place.
* Any updates to RSets of regions in the CSet have to be regenerated.
* All of these steps are expensive.

**How to avoid Evacuation Failure**

To avoid evacuation failure, consider the following options.

* Increase heap size
  + Increase the **-XX:G1ReservePercent=n**, the default is 10.
  + G1 creates a false ceiling by trying to leave the reserve memory free in case more 'to-space' is desired.
* Start the marking cycle earlier
* Increase the number of marking threads using the **-XX:ConcGCThreads=n** option.

#### Complete List of G1 GC Switches

This is the complete list of G1 GC switches. Remember to use the best practices outlined above.

|  |  |
| --- | --- |
| Option and Default Value | Description |
| -XX:+UseG1GC | Use the Garbage First (G1) Collector |
| -XX:MaxGCPauseMillis=n | Sets a target for the maximum GC pause time. This is a soft goal, and the JVM will make its best effort to achieve it. |
| -XX:InitiatingHeapOccupancyPercent=n | Percentage of the (entire) heap occupancy to start a concurrent GC cycle. It is used by GCs that trigger a concurrent GC cycle based on the occupancy of the entire heap, not just one of the generations (e.g., G1). A value of 0 denotes 'do constant GC cycles'. The default value is 45. |
| -XX:NewRatio=n | Ratio of new/old generation sizes. The default value is 2. |
| -XX:SurvivorRatio=n | Ratio of eden/survivor space size. The default value is 8. |
| -XX:MaxTenuringThreshold=n | Maximum value for tenuring threshold. The default value is 15. |
| -XX:ParallelGCThreads=n | Sets the number of threads used during parallel phases of the garbage collectors. The default value varies with the platform on which the JVM is running. |
| -XX:ConcGCThreads=n | Number of threads concurrent garbage collectors will use. The default value varies with the platform on which the JVM is running. |
| -XX:G1ReservePercent=n | Sets the amount of heap that is reserved as a false ceiling to reduce the possibility of promotion failure. The default value is 10. |
| -XX:G1HeapRegionSize=n | With G1 the Java heap is subdivided into uniformly sized regions. This sets the size of the individual sub-divisions. The default value of this parameter is determined ergonomically based upon heap size. The minimum value is 1Mb and the maximum value is 32Mb. |

### Logging GC with G1

This section provides a quick overview of the switches you can use to collect data and the information that is printed in the logs.

#### Setting the Log Detail

You can set the detail to three different levels of detail.

**(1) -verbosegc** (which is equivalent to **-XX:+PrintGC**) sets the detail level of the log to *fine*.

**Sample Output**

[GC pause (G1 Humongous Allocation) (young) (initial-mark) 24M- >21M(64M), 0.2349730 secs]

[GC pause (G1 Evacuation Pause) (mixed) 66M->21M(236M), 0.1625268 secs]

**(2) -XX:+PrintGCDetails** sets the detail level to *finer*.

The options shows the following information:

* Average, Min, and Max time are displayed for each phase.
* Root Scan, RSet Updating (with processed buffers information), RSet Scan, Object Copy, Termination (with number of attempts).
* Also shows “other” time such as time spent choosing CSet, reference processing, reference enqueuing and freeing CSet.
* Shows the Eden, Survivors and Total Heap occupancies.

**Sample Output**

[Ext Root Scanning (ms): Avg: 1.7 Min: 0.0 Max: 3.7 Diff: 3.7]

[Eden: 818M(818M)->0B(714M) Survivors: 0B->104M Heap: 836M(4096M)->409M(4096M)]

**(3) -XX:+UnlockExperimentalVMOptions -XX:G1LogLevel=finest** sets the detail level to its *finest*. Like finer but includes individual worker thread information.

[Ext Root Scanning (ms): 2.1 2.4 2.0 0.0

Avg: 1.6 Min: 0.0 Max: 2.4 Diff: 2.3]

[Update RS (ms): 0.4 0.2 0.4 0.0

Avg: 0.2 Min: 0.0 Max: 0.4 Diff: 0.4]

[Processed Buffers : 5 1 10 0

Sum: 16, Avg: 4, Min: 0, Max: 10, Diff: 10]

#### Determining Time

A couple of switches determine how time is displayed in the GC log.

**(1) -XX:+PrintGCTimeStamps** - Shows the elapsed time since the JVM started.

**Sample Output**

1.729: [GC pause (young) 46M->35M(1332M), 0.0310029 secs]

**(2) -XX:+PrintGCDateStamps** - Adds a time of day prefix to each entry.

2012-05-02T11:16:32.057+0200: [GC pause (young) 46M->35M(1332M), 0.0317225 secs]