Gangigunta, Amarnath (CORP)

From: Gangigunta, Amarnath (CORP)
Sent: Sunday, May 29, 2022 6:02 PM
To: Gangigunta, Amarnath (CORP)
Subject: Performance Engineering Basics

Performance testing Vs Engineering :

- Performance Testing is a quality check of the application in terms of the application's respor handling capacity. Performance Testing verifies how a system will perform under productior anticipate issues that might arise during heavy load conditions. On the other hand, Perform aims to design the application by keeping the performance metrics in mind and also to disc issues early in the development cycle.

Software architecture:

Before going to explain the types of architecture firstly you need to understand the different l foundation of software architecture. There are four types of layer:

- **Presentation Layer:** This layer is responsible to display the user interface and manage use
- **Application Layer:** Application layer (also known as Business Layer) has all the business lo policies. The application layer is a bridge between the presentation layer and the data laye
- **Data Layer:** This layer is responsible for storing the data.
- **Service Layer:** This layer is responsible to define and implement the service interface and Service layer communicates with the application layer

One Tier: MP3 player, MS Office
Two tier: Client(Presentation and app) – server(Database) application

Performance metrics:

Client Tier:

- TCP connection time
- HTML resources load time
- CSS files load time
- Images load time
- JavaScript file load time
- HTTP response time
- HTTP response status

Web tier:

- Cache Hit length
- Request Queued
- Number of HTTP connections

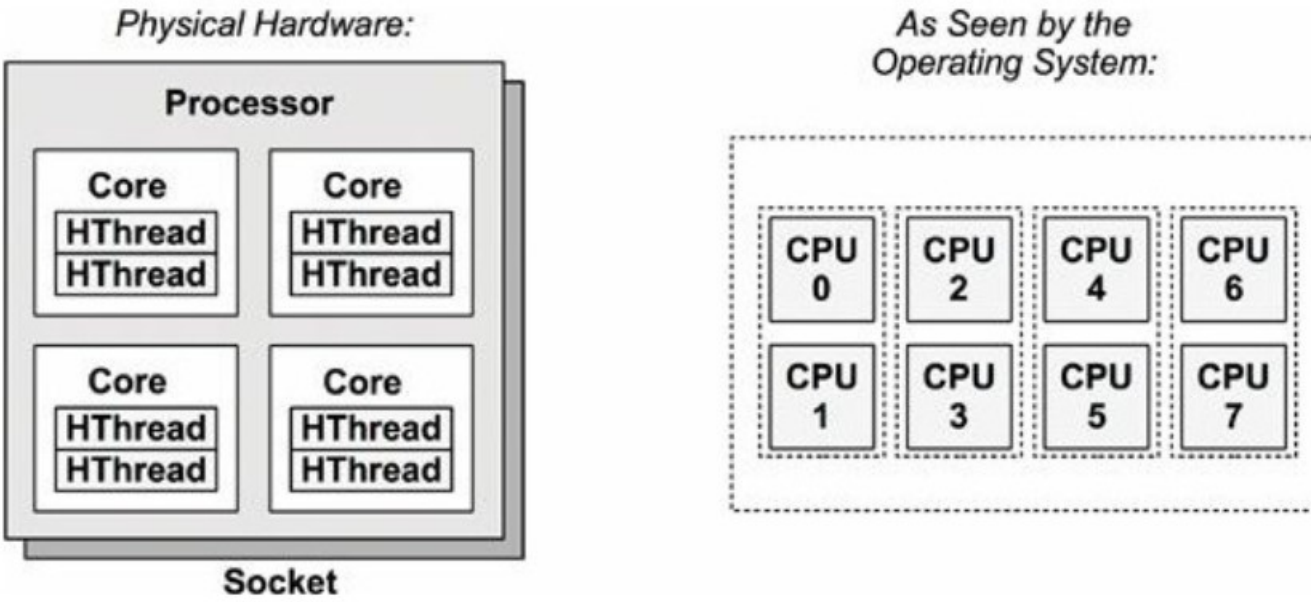
Application tier:

- Connection Time
- Connection Wait Time
- Total Threads
- Memory Use
- Active Transactions
- Transactions suspended
- Rolled Back transactions
- Timeouts
- Servlet Errors

Database Tier:

- Query throughput
- Query execution performance
- Queue Length
- Connections
- Buffer pool usage

Physical Vs Logical CPU:



Now that we know the required terms, a **Physical CPU** means the actual Physical Co is present on a processor. The multithreading being enabled on the core or not doesn't the count of Physical CPUs present on a Processor. In short, Physical CPU correlates actual Physical Cores present on a processor. In the above diagram there are 4 Physic Cores present on the Single Processor.

Logical CPU intern refers to the ability of each core doing 2 or more tasks simultane This is achieved by enabling hyperthreading on the cores. Each single physical core c divided in to multiple logical core by enabling hyperthreading on them. In the above diagram, although there are only 4 Physical cores, the system is tricked to look at it as Logical CPUs, by enabling hyperthreading on each core.

The above diagram correlates to below details :

1 Processor, 4 Physical cores, 8 Logical Cores. In other words - 1P4C - with 2 threads core.

Agent Vs Agent Less monitoring tools:

Difference Between Agent-based and Agent-less Performance Monitoring Tool:

- **Proprietary Software:**
  - *Agent-based:* Need to install on each server which is under monitoring
  - *Agent-less:* No need to install on each server
- **Central System:**
  - *Agent-based:* Along with agents, there is one central system (machine) named Master o This machine communicates with all the agents and collects the data.
  - *Agent-less:* There is only one central system which directly communicates with the serve monitoring) and collects the data.
- **Installation**:
  - *Agent-based:* It requires the installation of a central system with the additional installati each server.
  - *Agent-less:* It requires the installation of a central system only.

Factors that affect performance:

Code:

- memory leaks
- array bound errors
- inefficient buffering
- too many processing cycles
- a larger number of HTTP transactions
- too many file transfers between memory and disk
- inefficient session state management
- thread contention due to maximum concurrent user
- poor architecture sizing for peak load
- inefficient SQL statements
- lack of proper indexing on the database tables
- an inappropriate configuration of the servers

Network:

- Older or unoptimized network infrastructure
- Slow web site connections lead to network traffic and hence poor response time
- Imbalanced load on servers affecting the performance

Garbage Collection:

A Garbage Collector is a Java program which tracked the referenced (live) objects and allowed them t
the heap memory whereas the memory of the unreferenced (dead) objects is reclaimed and reused fo
object allocation. This method of reclaiming the unused memory is known as Garbage Collection. In (

Garbage collection is related to memory management and helps to improve the performance
Garbage collection in a system should not be too quick or too late. Too many GC cycles degra
performance of the system and causing the spike in CPU whereas a delay in GC cycles leads to

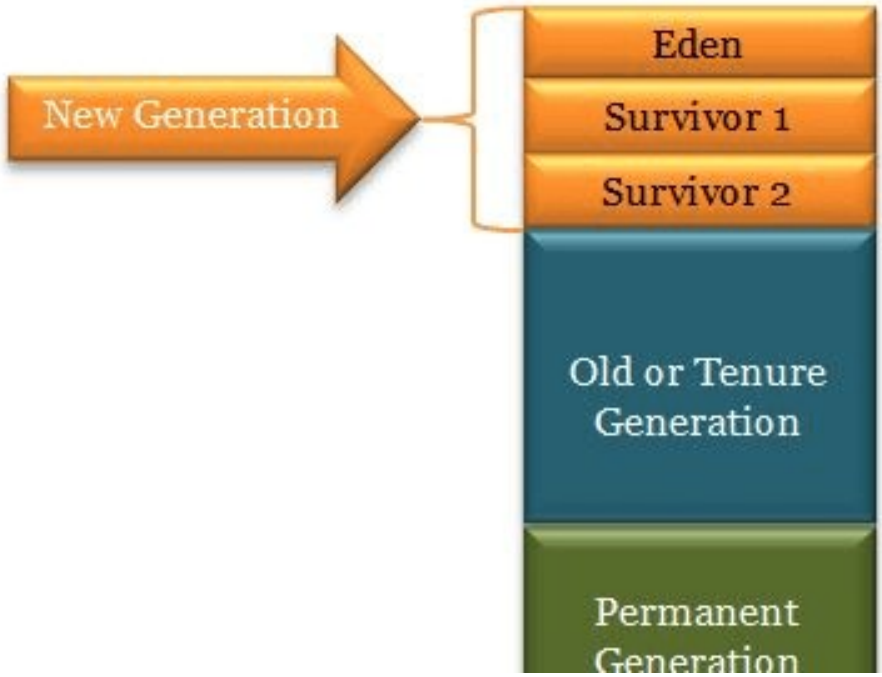
Some important terms:

1. *Live Object:* The object which is referenced by another object
2. *Dead Object:* The object which is not referenced by any other object and unreachable
3. *Daemon Thread:* Garbage collection process is carried out by Daemon Thread
4. *System.gc():* It is used to invoke the garbage collector and on invocation, the garbage collec reclaim the unused memory space. The System.gc() is a static method.

Memory heap:

What is Memory Heap?

Heap is a portion of memory which is used for dynamic allocation. The blocks of the memory in the
frequently allocated and freed based on the status of the object. The allocation and release of memo
place in an arbitrary order and hence the unreferenced or dead objects exist in the heap. The memo
unreferenced objects is reclaimed by the Java program called Garbage Collector.



Garbage Collection Process:

Promotion:

There is one more process which takes place along with all above-mentioned activities i.e. Object Pr
After each minor GC, the objects become old and their age keeps on increasing. When aged objects
certain age threshold they are promoted from the young generation to the old generation. This pro
as Promotion.

In our example you can see object A, G, J and O are promoted to the old generation after 'n' numbe
GC executions. The promotion process leads to fill old generation memory space. One important po
down here some of the objects like String and Array objects are directly created in tenured generati
space.

Types of GC:

Serial: Uses only one thread for GC
Parallel: multiple threads for GC
Concurrent mark sweep (CMS) GC:

The Concurrent Mark Sweep (CMS) garbage collector collects the tenured generation. It is also calle concurrent low pause collector. It attempts to minimize the pauses due to garbage collection by doi the garbage collection work concurrently with the application threads. Due to this reason, the CMS more CPU than other GCs. If you can allocate more CPU for better performance then CMS garbage preferred choice over the parallel collector. CMS garbage collector uses multiple threads to scan the

G1 (Garbage First) Garbage Collector : G1 GC provides benefit where multi-processor machines along with large memory space are available.

Stop the world event:

Stop the World Event - All minor garbage collections are "Stop the World" events. This means that all application thr 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 a when that age is met, the object gets moved to the old generation. Eventually the old generation needs to be collected. event is called a ***major garbage collection***.

Heap Dump:

JVM heap dump is an informative data (sometimes it is called a snap) of Java heap memory which co low-level details about java objects and classes allocated in the Java heap at the moment when the d fetched or snapshot is triggered. Generally, a heap dump is triggered after the full GC ran so that the contains the information about the remaining objects in the heap. A heap dump contains the inform

Thread Dump:

Similar to Heap Dump, Thread Dump is a snapshot of the status of all the threads at a particular time

1. Thread:

A thread is a series of statements which can be processed in parallel in the same program to achieve concurrency. All the Java threads are managed by JVM (Java Virtual Machine) which are mapped wit operating system thread, also called as Native Thread. Threads help to allow multiple activities withi process and act like lightweight processes. Each thread has a unique identifier, name, and category.

2. Multithreading:

Java is a multi-threaded application that allows multiple thread execution at any particular time. A we uses tens to hundreds of threads to process a large number of concurrent users and achieve multithr the problem occurs when two or more threads utilize the same resources and leads deadlock situatic

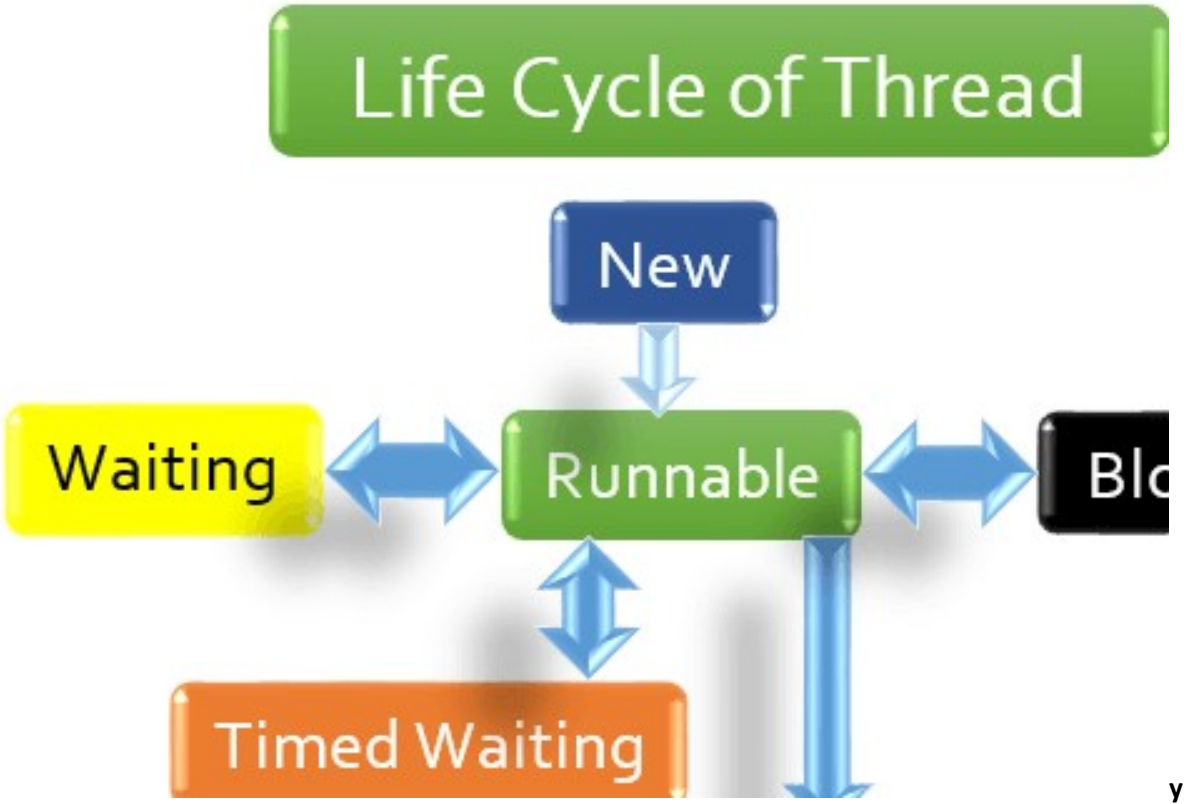
3. Deadlock:

Deadlock is a situation when two or more threads are waiting for the other threads to complet order to complete their own tasks. To identify the root cause of the deadlock, a thread dump i

4. Thread Dump:

Re-iterating the definition again. Thread Dump is a snapshot of the status of all the threads at helps to find out what every thread in the JVM is doing at a particular point in time. Are some deadlock? Are all the threads running? Which threads are waiting for the blocked resources et questions can be answered by analysing in the thread dump.

Life Cycle of a Thread:



y

Thread Dump Analyzer: fastThread

Blocked Thread Scenario

To continue the thread dump analysis, now we will move to the next part which will cover the analysi threads. The meaning of blocked thread is that a thread locks a resource for a long time and restrict threads to use that particular resource, hence the restricted threads are moved into the Blocked state as Blocked Threads. Let's try to understand with an example. Refer to the below fastThread dashboar

Deadlock Scenario

Deadlock is a situation when the locked resources are required by some other threads who lo resources which are needed by other threads. Such a cyclic condition restrict to complete the the threads, resulting in none of the thread completes its execution and system is halted. Refe

AWR :

Automatic Workload Repository (also caller AWR) report provides information about the database bottleneck.

Dump vs AWR:

The difference between the dump (heap or thread) and Automatic Workload Repository repor reports are the actual snapshots taken at a particular time whereas AWR report is a comparisc snapshots taken at the different timestamps.

AWR Features:

The Automatic Workload Repository report provides:

- Wait-events causing a delay
- Highlights query taking long Elapsed time or execution time
- Report on CPU utilization and memory usage
- Blocking sessions and many other important stats which we will be discussed in later posts

Server Tuning Tips:

Web server: Apache, ASP, IIS