V20PCA107 - IT INFRASTRUCTURE MANAGEMENT 2.0

UNIT - IV_WEEK - 10

Performance and Tuning

Introduction

Performance tuning is primarily a matter of resource management and correct system-parameter setting. Tuning the workload and the system for efficient resource use consists of the Identifying the workloads on the system. Setting objectives: Determining how the results will be measured. Performance and tuning are a methodology to maximize throughput (the amount of a product or service that a company can produce and deliver to a client within a specified period) and minimize response times of batch jobs, online transactions, and Internet activities.

Definition of Performance and Tuning

The previous discussion on what sets the performance and tuning process apart from other infrastructure processes leads us to the following formal definition of performance and tuning.

Performance and Tuning

Performance and tuning is a methodology to maximize throughput and minimize response times of batch jobs, online transactions, and Internet activities. Our definition highlights the two performance items that most IT customers want their systems to have: maximum throughput and minimal response times. These two characteristics apply to all platforms and to all forms of service, whether batch jobs, online transactions, or Internet activities.

In keeping with our focusing first on people and processes ahead of technology, we identify and examine issues associated with performance and tuning rather than provide a shopping list of the plethora of technology products and tools used to actually

do the tuning. Performance specialists often find many ways to successfully tune each of the five major resource environments by deploying appropriate staff and by implementing proper procedures before applying technology.

Within each of the five areas, we discuss some of the meaningful metrics that world-class infrastructures use to measure and improve performance and tuning. We conclude by offering assessment worksheets for evaluating a company's performance and tuning process, both with and without weighting factors.

Differences between the Performance and Tuning Process and Other Infrastructure Processes

But the performance and tuning process has more differences from the others than is typical, and the differences themselves are more significant than usual. Understanding these differences can help to define the overall process and select suitable process owners. Summarizes these differences.

Table: Summary of Performance and Tuning Process Differences

Performance and tuning	Other processes
• Consists primarily of two major activities • Normally has multiple sub process owners • Shares ownership across multiple departments • Task have continuous and ongoing nature • Process is highly iterative • Process tools vary widely among the resource environments • Process utilizes a large number of differentiated metrics(identify risk areas, priorities initiatives, set targets)	• Consists primarily of one major activity • Normally has one overall process owner • Centralizes ownership in a single department • Tasks have definitive start and end dates • Processes are rarely iterative • Process tools are usually shared across departments • Processes utilizes a small number of similar metrices

Solution for Performance Problem

- The quickest and simplest way to solve a performance problem is to throw more hardware at it. For Example,
  - The system is running slow, so upgrade the processors
  - If swap rates are too high, just add more memory.
  - For cache hits too low or disk extents too high, simply buy more cache and disk volumes
- The tuning is an ongoing activity in which the performance bottleneck is never truly eliminated; it's only minimized or relocated.
  - Example: Hundreds of new desktop computers have been added to a highly used application.
  - The resulting slow response is attributed to lack of adequate bandwidth on the network, so bandwidth is added.

Preferred Characteristics of a Performance and Tuning Process Owner

- There may be two or more sub processors whose high-priority characteristics vary depending on the area in which they are working.

- For example, an individual selected as the process owner for only the network area should have as a high priority knowledge of network software and components,
- But knowledge of systems software and components need only be a medium priority.
- The reverse would apply to an individual selected as the process owner for only the server area.

Table: Prioritized Characteristics for a Performance and Tuning Process Owner

Characteristic	Priority
1. Knowledge of systems software and components	High
2. Knowledge of network software and components	High
3. Knowledge of software configurations	High
4. Knowledge of hardware configurations	High
5. Ability to think and act tactically	High
6. Knowledge of applications	Medium
7. Ability to work effectively with developers	Medium
8. Knowledge of desktop hardware and software	Medium
9. Knowledge of power and air conditioning	Medium
10. Ability to meet effectively with customers	Low
11. Ability to promote teamwork and cooperation	Low
12. Ability to manage diversity	Low

Knowledge of software and hardware configurations is of a high priority regardless of the area in which a process owner works, as is the ability to think and act tactically. The relationships of the process owners to the application developers can be key ones in that performance problems sometimes appear to be infrastructure related only to be traced to application coding problems, or vice versa. A good working knowledge of critical applications and the ability to work effectively with developers is at minimum a medium priority for process owners working with application systems.

Performance and Tuning Applied to the Five Major Resource Environments

Now let's look at how the performance and tuning process applies to each of the five major resource environments found within a typical infrastructure: servers, disk storage, databases, networks, and desktop computers. Since we are focusing first on people and processes ahead of technology, we will identify and examine issues associated with performance and tuning rather than talk about all the technology products and tools used to actually do the tuning.

Server Environment

The first of the five infrastructure areas affected by performance and tuning covers all types and sizes of processor platforms, including mainframe computers, midrange computers, workstations, and servers. For simplicity, we refer to all these platforms as servers. The following list details the major performance issues in a server environment:

1. Processors
2. Main memory
3. Cache memory
4. Number and size of buffers
5. Size of swap space
6. Number and type of channels

Processors - The number and power of **processors** influence the rate of work accomplished for processor-oriented transactions. Processors are the central components (also called the *central processing units*) of a digital computer that interpret instructions and process data. At its core, a processor adds and compares binary digits. All other mathematical and logical functions stem from these two activities.

For optimal performance, processor utilization rates should not exceed 80 percent. Tools are available to measure real-time utilizations of processors, **main memory**, and channels.

Cache memory is available on most mainframe computers and on some models of servers, offering an additional means of tuning transaction processing. Cache memory differs from main memory in the following manner.

It is slower than main memory but faster than secondary storage. One of the most effective ways to improve disk-storage performance is to utilize cache memory in RAID-type disk arrays. (RAID stands for redundant array of independent — originally, inexpensive— disks.

The reason cache is so effective at improving disk performance is that, for an overwhelming number of input or output transactions to and from disks, it eliminates the time-consuming steps of seeking and searching for data out on disk volumes.

It accomplishes this using ingenious pre-fetch algorithms that anticipate which data will be requested next and then preload it into the high-speed cache memory.

A pre-fetch algorithm is a set of software instructions that analyses the patterns of disk storage read-request activity to anticipate and pre-load, or pre-fetch, expected records of data.

Main memory is extremely fast circuitry (directly attached to the processor) that stores instructions, data, and most of the operating system software, all of which are likely to be used immediately by the processor.

Cache memory is slightly slower memory (directly attached to main memory) that stores instructions and data about to be used. Cache is much faster (and more expensive) than secondary storage such as disks and tape.

The number and size of buffers - Buffers are high-speed registers of main memory that store data being staged for input or output. The number and size of buffers assigned for processing of I/O operations can trade off the amount of memory available for processor-oriented transactions.

The size of swap space -The portion of main memory set aside(allocate) for storing small frequently used part of programs from main memory to secondary storage is called swap space. The size of swap space can be adjusted to match the profiles of application and database processing.

Number and Type of Channels

Channels are physical cables that connect the main memory to external I/O devices such as disk drives, tape drives, and printers. The rate of processing I/O operations is also determined by the number and speed of channels connecting servers to external disk equipment.

Performance metrics in a server environment include:

1. Processor utilization percentages (percentage of time the processor spentdoing work)

2. The frequency of swapping in and out of main memory

3. Percentage of hits to main memory cache

4. The length and duration of processing queues

5. The percentage of utilization of channels

6. The amount of processor overhead (work) used to measure performance

Disk Storage Environment

The second type of infrastructure resources impacted by performance and tuning is disk-storage equipment.

The following list indicates the major performance issues in a disk-storage environment:

1. Cache memory

2. Volume groups

3. Striping

4. Storage area networks

5. Network -attached storage

6. Extents

7. Fragmentation

Mapping out volume groups is another effective way to improve performance. Volume groups are a logical grouping of physical disk drives. It reduces seek-and-search times by combining frequently used groups of physical volumes into one logical group.

Example: By arranging the data of customers using credit and debit card in a single disk is more efficient than spreading across different disks.

Striping is a performance-improvement technique in which long blocks of data that are to be read or written sequentially are stored across multiple drives, usually within a logical volume group. This is done to increase the number of data paths and transfer rates and to allow for the simultaneous reading and writing to multiple disks.