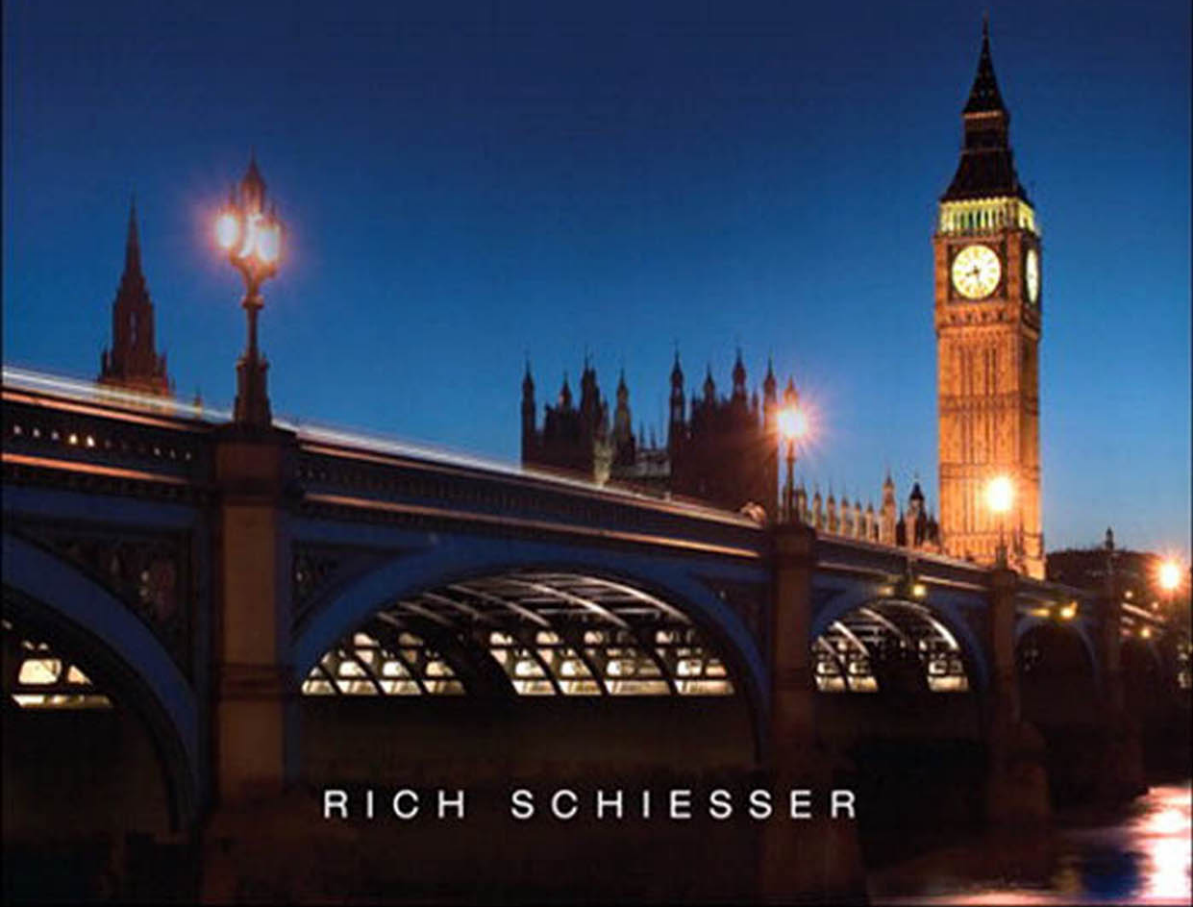
SECOND EDITION

IT Systems Management



RICH SCHIESSER

Praise for *IT Systems Management, Second Edition*

IT Systems Management, *Second Edition*, is one of those definitive books that will serve as the foundation for a whole new breed of IT professionals. Apart from the innovative content, instructors and students will appreciate the added features such as revision questions and answers per chapter, additional reading suggestions, and real-life experiences. The supporting material for instructors makes it quick and easy to develop courses for IT Systems Management.

Prof. Les Labuschagne
Director: School of Computing
University of South Africa (UNISA)

In Rich's second edition of *IT Systems Management*, he has built a true bridge between academia and today's IT professional working in a highly complex, ever changing environment. Now both student and practitioner have a common playbook available to guide us through the 21st century technology landscape. I strongly recommend this book as mandatory reading, whether on the front lines of support or for the senior executive working on navigating a long-term strategy for an organization.

Mike Marshall
Director, MMT Production Service
Association of Retired Persons (AARP)

Managing IT infrastructures has always been a complex undertaking and most managers had to learn its ins and outs through experience. Rich Schiesser's *IT Systems Management* text offers a much more humane approach in that one can gain the breath of exposure to the topic without suffering the negative consequences resulting from knowledge gaps. This text offers insights into this topic that only someone who had been intimately involved over a long period of time could provide. Students and technical managers alike can profit from this treatise. I highly recommend this as a source book for academic courses in this area. It has too long been an ignored topic.

Dr. Gary L. Richardson
Program Coordinator
Graduate Technology Project Management Program
University of Houston

The 2nd edition of *IT Systems Management* uses a managerial approach to handling IT systems. The techniques described in the book are appropriate for professionals in the Telephony and Information technology industries. Schiesser offers in-depth looks at processes and procedures, and provides "how to" approaches to customer

The first difference references the dual nature of performance and tuning—they are actually two related activities normally combined into one process. The performance activity has as its cornerstone the reporting of real-time performance monitoring and reporting as well as periodic management trend reports on a daily, weekly, or less-frequent basis. The tuning activity consists of a variety of analyses, adjustments, and changes to a whole host of parameters. All other systems management processes consist of primarily one major activity and they have one overall process owner. Because performance and tuning activities normally occur in five separate areas of the infrastructure, there is usually more than one subprocess owner depending on the degree to which the five areas are integrated. These multiple subprocess owners share ownership across several departments, whereas the other systems management processes have a centralized ownership within one infrastructure department.

The nature of the tasks differs in that performance and tuning is continuous and ongoing. Performance monitoring occurs for all of the time that networks and online systems are up and running. The other infrastructure processes tend to have definitive start and end dates for their various tasks. Changes are implemented, problems are resolved, and new applications become deployed.

Tuning a portion of the infrastructure environment to correct a performance problem can be a highly iterative activity, often requiring numerous trials and errors before the source of the problem is found. Other processes are seldom iterative because most of the tasks associated with them are of a one-time nature.

Processes such as change and problem management normally use a single database-oriented system to manage the activities. This becomes their main process tool. Process tools used with performance and tuning are large in number and diverse in application. This is due to the eclectic tuning characteristics of the various infrastructure resources. No single tool exists that can fine-tune the operating systems of mainframes, midranges, servers, networks, and desktop computers. Similarly, there are separate and highly specialized tools for defining databases, as well as for maintaining, reorganizing and backing up and restoring them.

Finally, there is the issue of metrics. The performance activity of this process relies heavily on a large variety of metrics and reports to identify, both proactively and reactively, trouble spots impacting online response, batch throughput, and web activity. Other processes use metrics to be sure, but not quite to the extent that performance and tuning does.

These differences are what sets this process apart from others and influences how the process is designed, implemented, and managed.

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.

The five infrastructure areas most impacted by performance and tuning are:

1. Servers
2. Disk storage
3. Databases
4. Networks
5. Desktop computers

The methodologies used to tune for optimum performance vary from one area to the other, lending to subprocesses and subprocess owners. But there are some generic characteristics of performance and tuning that apply to all five areas.

One of these is the false notion that the quickest and simplest way to solve a performance problem is to throw more hardware at it—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. These solutions may be quick and simple, but they are seldom the best approach. The relief they

may offer is seldom permanent and usually not optimal, and it is often hard to justify their cost in the future. The fallacy of a limited hardware solution lies in the fact that it fails to address the essence of performance management: that tuning is an ongoing activity in which the performance bottleneck is never truly eliminated, it's only minimized or relocated.

To further illustrate this point, consider the following scenario. 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. However, that leads to increased transaction arrival rates, which now clog the channels to the disk array. More channels are added, but that change now causes major traffic contention to the disk volumes. Volumes are added, but they saturate the cache, which, in turn, is increased. That increase saturates main memory, which is then increased and fully saturates the processors, which are upgraded to handle this, only to have the entire cycle start over in reverse order.

I admit that this is an exaggerated example, but it serves to reinforce two key points:

1. Performance and tuning are ongoing, highly iterative processes. Shops that treat them as occasional tasks usually end up with only occasionally good performance.
2. Overall planning of the type and volume of current and future workloads is essential to a robust process for performance and tuning.

Tuning is sometimes likened to trying to keep a dozen marbles on a piece of plate glass you are holding. You must constantly tilt the glass one way and then quickly to another to keep all the marbles on the surface. Distractions, unplanned changes, and unforeseen disruptions in the environment can easily cause a marble to fall off the glass, similar to what can happen with performance and tuning.

Preferred Characteristics of a Performance and Tuning Process Owner

Table 8-2 lists in priority order many of the preferred characteristics of a performance and tuning process owner. In reality, there may be two or

more subprocessors 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.

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.

Table 8–2 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

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 of 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 of 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

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. 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.

Real Life Experience—Swapping Club Requires an Explanation

A CIO at a major defense contractor took an active interest in online response times for his customers and wanted to know exactly when and why performance might be degrading. It was a tough sell for systems analysts to explain to him why response times increased beyond service levels whenever only a few users were signed on yet seemed to improve with more users.

The cause of this performance paradox was the way main memory swap space was set up. When many online users were signed on, almost all the available memory would be used for online transactions and very little swapping out to disk storage occurred. When only a few users signed on, much of memory was used for batch jobs and the online users would frequently get swapped out, causing slower response times.

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. Buffers are high-speed registers of main memory that store data being staged for input or output.

The concept of virtual storage is used to temporarily store small, frequently used portions of large programs in part of main memory to reduce time-consuming I/O operations to secondary storage. The portion of main memory set aside for this is called *swap space* because the program segments get swapped in and out of main memory. The **size of swap space** can be adjusted to match the profiles of application and database processing.

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

Performance metrics commonly collected in a server environment include:

1. Processor utilization percentages
2. The frequency of swapping in and out of main memory