Copyright Statement

Chapter 7: Optimizing Virtualization Performance

The primary goal for most IT departments is to wring every last bit of available performance out of their existing investments. Clearly, this is one of the reasons that virtualization is such an easy decision. By combining workloads onto physical servers that are typically under-utilized, organizations can get instant benefits without large hardware-related expenditures. Some might be happy with the gains that they've realized. But idealistic IT managers will want to know: How can we further improve performance? That's the subject of this chapter.

Much of the complexity of managing IT environments is related to the many types of supported hardware, software, and operating systems (OSs). For example, the effort required to manage direct-attached hard disks on individual servers can limit scalability. An ideal solution for IT departments is to create a single pool of hardware resources. Workloads can then be deployed into the pool without having to worry about all the individual configuration details.

In many ways, attempting to optimize virtualization performance is analogous to going from "good" to "better" (rather than from "bad" to "good"). For example, an individual physical server might be running at 60 percent average utilization when supporting virtual machines, rather than at 10 percent average utilization when supporting only a single workload. The primary goal of improving virtual machine performance is to push the capacity-related limits of a computer without adversely affecting real-world performance for users. Applications and services should still be able to meet service level expectations.

This chapter will discuss details about how to manage this balancing act. I began the discussion in Chapter 6, when talking about ways in which you can monitor the performance of virtual and physical systems. In this chapter, you'll learn how to apply that information to make better decisions about where and how to deploy virtual machines. The goal is to optimize resource utilization and virtual machine performance.

Developing a Performance Optimization Approach

When working in complex environments, it's often all too tempting to take a haphazard and reactive approach to resolving performance issues. Is a particular virtual machine adequately sized with respect to its physical memory allocation? If the allocated amount is too low, users will notify the IT department of slow performance. If it's too high, you might never detect it at all. Before diving into technical considerations, it's important to develop an organized and systematic approach to identifying and resolving performance-related inefficiencies. Figure 7.1 provides an example of the typical steps that should be included in a performance optimization plan.

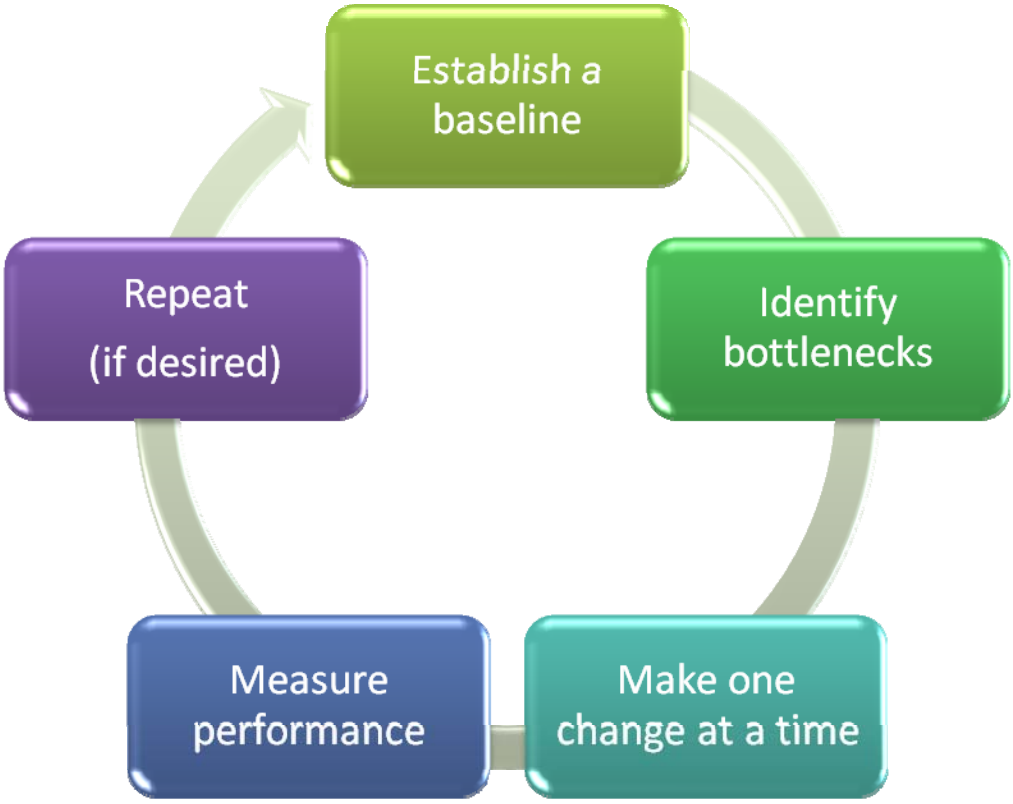

Figure 7.1: Developing a performance-optimization process.

The process step begins with establishing a baseline of current performance. It's very difficult to accurately identify the source of a performance issue without gathering statistics. Having baseline performance details is also a vital aspect of determining the effects of changes.

Defining the Starting Point

The second step involves identifying bottlenecks. If the term bottleneck is defined as the slowest step in a given process, then there will always be some resource that is constrained. In the case of virtualization hosts, the most common limitations are related to CPU, memory, disk, and network details. Often the symptoms are inter-related. For example, if a particular host server is exhibiting a large amount of disk-related activity, the root cause might be long or frequent file transfer operations. It could also be a symptom that is secondary to having a large amount of swapping to disk. In that case, the best solution would be to increase the total amount of physical memory on the server and rebalancing memory allocations on the servers.

Making and Monitoring Changes

Once a potential performance limitation is identified, IT staff can determine whether a change is required. Here is another part of the performance optimization process that is often ignored. Often, when a problem occurs, it's tempting to make multiple changes at a time. For example, if a virtualized workload is running too slowly, systems administrators might move virtual hard disk files to other volumes, reconfigure the total amount of memory, and change CPU allocation priorities all at the same time. In some cases, the net effect of these changes might be an increase in overall performance. But the effects of the individual changes might be different. Figure 7.2 shows an example where some changes actually *decreased* performance, although the sum of the changes was positive. In this situation, performance can said to have been improved but not optimized.

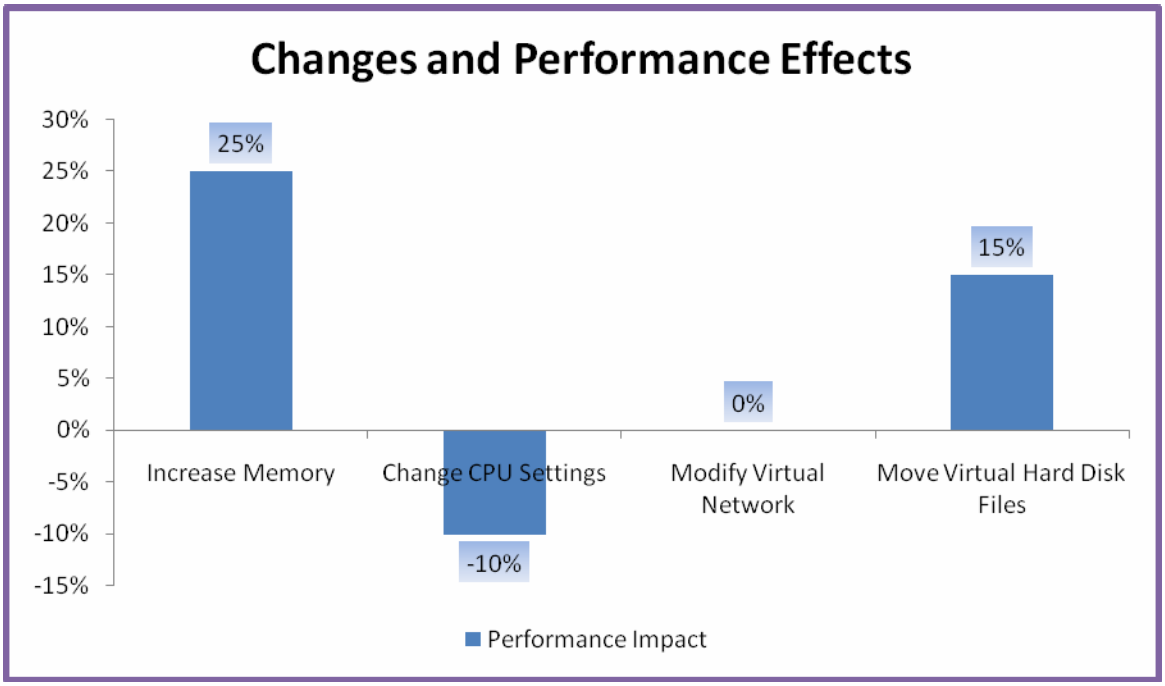

Figure 7.2: Complicated effects of making multiple changes at a time.

After a change has been made, it's important to retest the system. In the case of performance troubleshooting, the original symptom was likely to be related to the processing of a transaction, system throughput, or response times. At the very least, systems administrators should verify that the initial issue was addressed. Sometimes, re-measuring performance will help identify other potential problems.

Settling for "Good Enough"

As the process cycle indicates, performance optimization can potentially be a never-ending process. Again, if a bottleneck is defined as the slowest portion of a particular system, then the best one can hope for is to move the bottleneck (rather than to remove it entirely). From a practical standpoint, organizations will often see diminishing returns when repeating the optimization process for the same workload. Figure 7.3 shows an example. The primary factors to consider include the overall costs of improving performance versus the potential benefits.

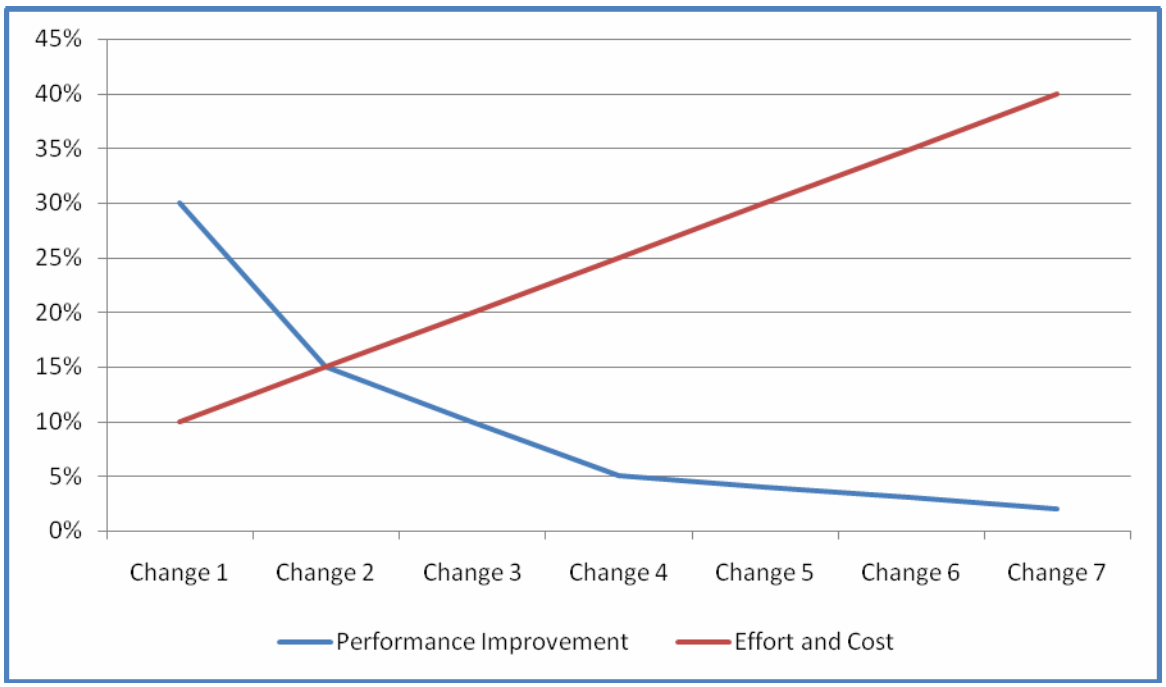

Figure 7.3: Diminishing returns from performance optimization.

In many cases, the biggest gains can be realized by detecting and addressing the major bottlenecks. Through the initial iterations of the optimization process, changes can be relatively easy and result in significant performance gain. Thus, given the fact that IT departments have limited time and human resources, when is the process complete? The general answer to the question is that optimization can stop when performance meets users' requirements. Rather than searching for theoretical maximum performance numbers, IT organizations should settle for "good enough" workload response times and throughput. One way of defining this goal is through the use of Service Level Agreements (SLAs).

📖 SLAs were covered in Chapter 6.

Optimizing Virtual Machine Placement

In traditional IT architectures, the unwritten rule was often "one application per server." Computers would usually have specifications that were driven by the software that they were intended to run. Scaling down usually wasn't an option. Scaling up could often be performed by upgrading the server's hardware. And scaling out was done by distributing the application across servers (such as in a multi-tier application scenario) or by using technologies such as clustering. Figure 7.4 shows examples of these two approaches.

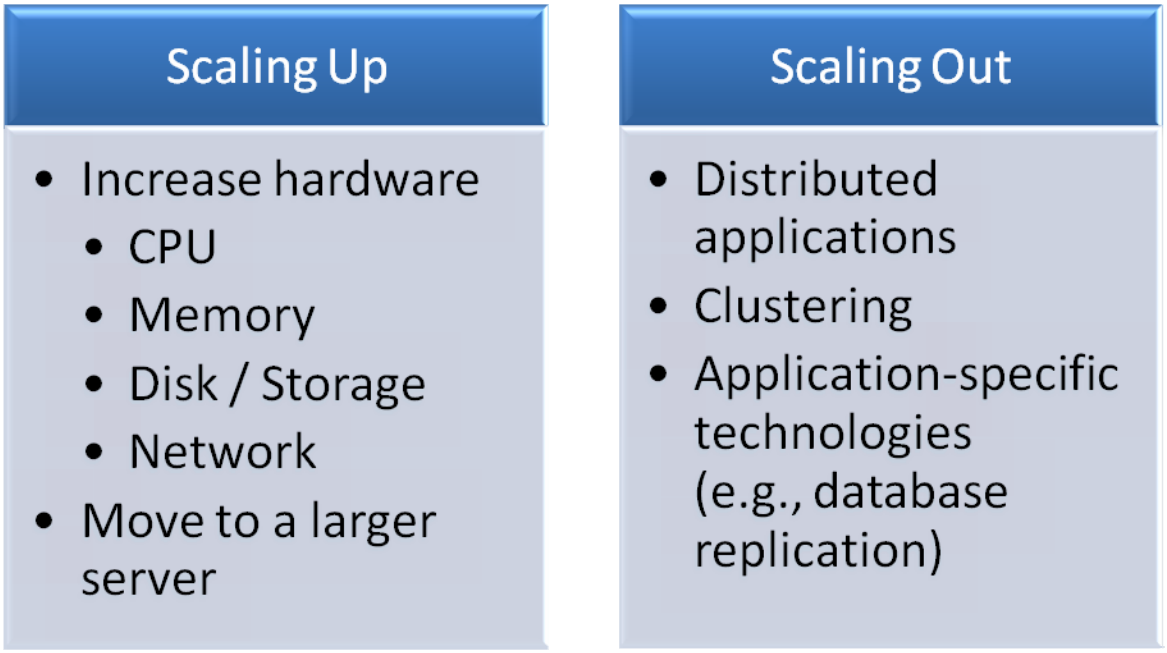

Figure 7.4: Comparing scalability approaches for physical servers.

Organizations that have incorporated virtualization still have many of these options. But they also have the ability to move workloads between physical host computers. In "classic" IT architectures, workloads are tightly tied to individual computers. Due to complex configurations and a strong coupling between OSs, applications, and physical hardware, the process of managing the infrastructure can be difficult. The ideal solution for IT departments is to treat all their server assets as a single, large pool of resources. Workloads can then be deployed onto those systems. And, best of all, IT departments no longer have to worry about managing individual disks, network adapters, and other settings.

This approach can help increase server utilization and enables IT agility by simplifying the process of "scaling" virtual machines rather than physical ones. But decision-making can be more complicated. The challenge is in determining the optimal placement of virtual machines, given a set of host computers. In this section, I'll cover several factors that should be considered when determining where virtual machines should be placed for optimal performance.

Determining Workload Importance

When it comes to running multiple virtual machines on a single host server, not all virtual machines are created equally. By default, virtualization management platforms will provide an equal priority to each of the virtual machines that is located on a host computer. A much better configuration, in most cases, is to place priorities on each of the virtual OSs, applications, and services that are required.

The first step in determining the optimal configuration settings for a particular virtual machine is to determine its importance to the organization is a whole. Figure 7.5 provides an example of typical roles that virtual machines can play in a production environment.

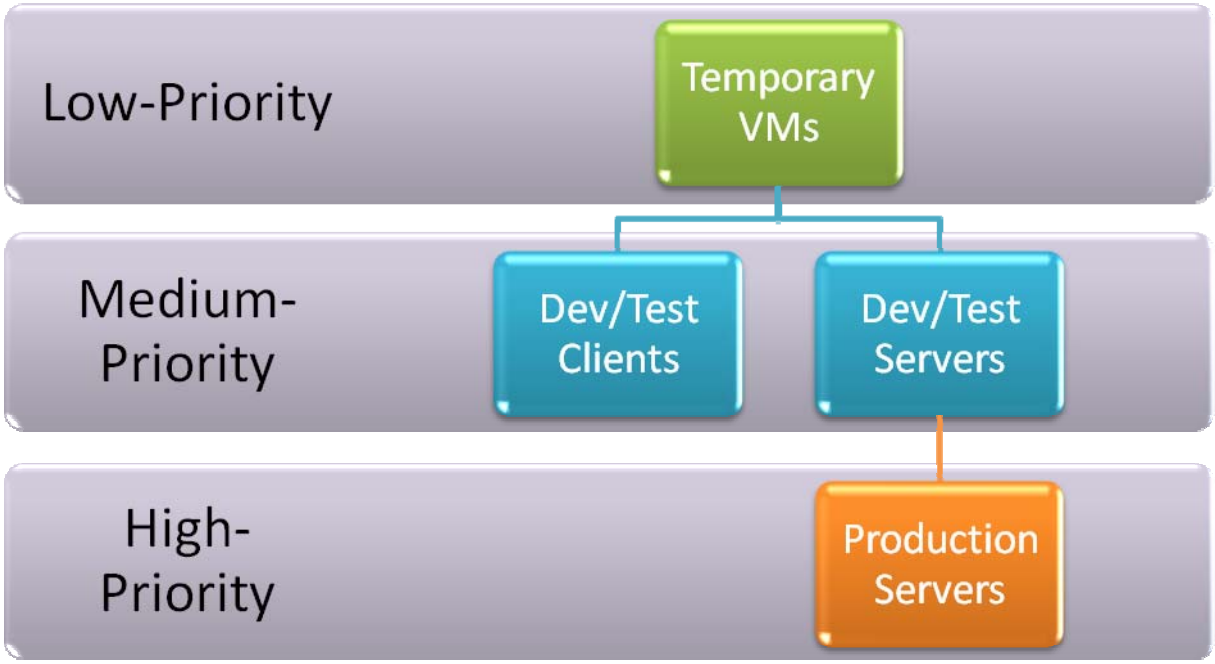

Figure 7.5: Determining the importance of various workloads.

At the lowest end of the spectrum are temporary virtual machines that might have been created for compatibility testing or related purposes. These virtual machines generally should not be present on production servers. If they are, they should be given fewer system resources (especially when memory and CPU time are constraints). Moving higher in the order of importance, there are virtual machines that are used for development and testing purposes. At the highest level are those virtual machines that are running mission-critical workloads.

There are several factors that should be taken into account when determining the importance of virtual machines:

- Workload purpose—The purpose of the actual workload can vary significantly. In some cases, an application might be used for development and testing purposes. In another case, it might be relied upon as a critical portion of an important business process.
- Performance requirements—IT organizations often have many types of customers, each with their own perspective on acceptable performance. In cases in which performance (as measured by response times, throughput, or other criteria) is critical, the workload should be given a higher importance.
- Number of users—Typically, applications and services that are relied upon by larger numbers of users are considered more important than those that will affect only a few users.
- Downtime costs—Most production workloads will have an associated downtime cost—the amount of revenue or business lost based on the lack of availability of the functionality. The higher the downtime cost, the higher should be the associated importance of the virtual machine.

When combined, these factors can help IT organizations determine the true importance of each virtual machine. In some cases, it might make sense to place a numerical priority (for example, from 1 to 10) on each type of virtual machine.

Defining Host Computer Requirements

Although it is clear that the purpose and importance of virtual machines will vary, there are often associated requirements that are important to consider for host systems. Some host servers might be protected using a wide array of high-availability and performance options. Others might be running without redundancy. Figure 7.6 provides an example of some of the optional features for host servers.

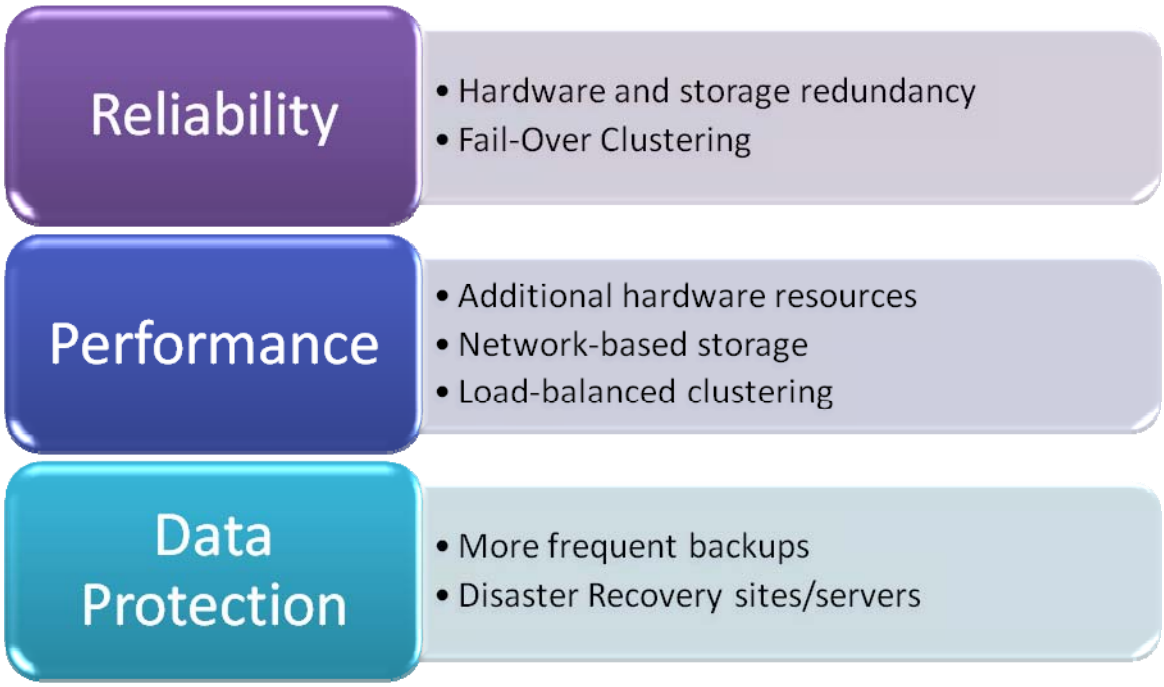

Figure 7.6: Examples of host technologies that can improve service quality.

In an ideal world, all servers in the data center will have all these features. In reality, it is a simple fact that higher data levels of performance and data protection will result in higher costs. The result is that IT departments often need to choose which servers are most important. Accordingly, the most important virtual machines should be placed on servers that have high performance, data protection, and reliability features. Lower-priority virtual machines can be moved to other servers to reduce overall costs.

Calculating Resource Requirements

One of the most valuable benefits related to working with heterogeneous environments is that of flexibility. When virtual platforms are used, systems administrators have the ability to relocate workloads to different systems quickly and easily. That raises the question of how to best arrange the workloads. One method involves profiling applications to determine their types of resource usage. Table 7.1 provides an example.

Workload	CPU Utilization	Memory Utilization	Disk Utilization	Network Utilization
Public Web Server	Low	Low	Low	High
Web Application Server (Internal)	Medium	Medium	Low	Medium
Middle-Tier Server (ERP Application)	Medium	High	Low	Low
Database Server (ERP Application)	High	High	High	Medium

Table 7.1: Workload characterization based on resource utilization.

The table shows various types of hypothetical workloads, along with a high-level categorization of resource usage. For simplicity, only CPU, memory, disk, and network utilization are being measured. The source of this information is usually performance data that is collected over time. For example, if an application is running on an existing physical server, various tools and methods can be used to determine its requirements. For new applications, load-testing can be used to obtain estimated resource usage.

📖 For more information about performance testing, see Chapter 6.

Wherever possible, resource utilization should be measured using numerical statistics. Table 7.2 provides examples for a sample type of workload, along with related comments. In this case, performance characteristics for a sample database server are determined. Useful details include the average resource utilization over a period of time (such as an hour, a day, or a week) as well as peak utilization.

Resource	Average Values	Peak Values	Performance Notes
CPU	45%	90%	Peak utilization during end-of-day reporting
Memory	4.8GB	12.0GB	More memory decreases average transaction times
Disk	2.0MBps	7.0MBps	Additional disk load occurs during backups
Network	3.0Mbps	45.0Mbps	Actual throughput based on network congestion

Table 7.2: Performance characteristics for a hypothetical database server.

These types of performance measurements can be generated both for workloads running on physical servers and for those running within virtual machine environments.

Combining Compatible Workloads

When organizations have performance-related details, they can use them to make better decisions about virtual machine "compatibility." In general, it is best to match up the types of workloads that do not compete for the same resources. For example, an application that places a heavy load on the disk subsystem would best be matched up with other applications that have heavy CPU requirements. Figure 7.7 provides an example.

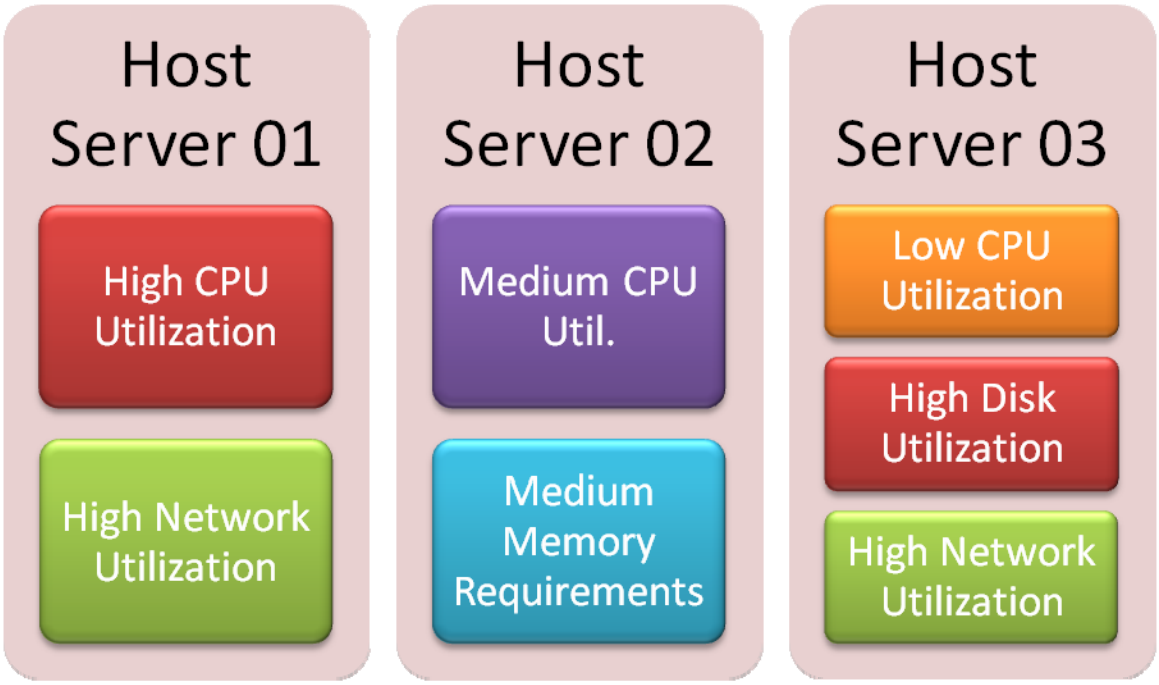

Figure 7.7: Allocating virtual machines based on resource requirements.

From a performance standpoint, this distribution of virtual machines helps ensure that most of a server's performance potential is being realized. It also helps ensure that performance bottlenecks caused by resource contention are minimized. In real-world situations, it might be difficult to accurately determine the ideal configuration. Later in this chapter, I'll discuss how automation can assist in making better virtual machine distribution decisions.

Understanding Resource Management Approaches

There are two main goals of performance optimization with relation to virtualization. The first goal is that of load-balancing. In this approach, IT departments will attempt to keep an even level of utilization across servers. This is often known as load-balancing because virtual machines are placed and moved so that average load is kept balanced across the data center. The primary advantage of this approach is that it allows additional capacity for new virtual machines. For example, if servers are 65 percent utilized on average, it is often simple to calculate the number and types of virtual machines that can be added to the environment. Additionally, the load-balancing optimization approach helps avoid potential performance problems on virtual machines because each server will generally have additional capacity that is available to accommodate spikes and contention problems.

Another performance optimization approach is to focus on maximized resource utilization on each physical server. The goal is to ensure that server investments are being used optimally. A typical example is the case of using virtualization for server consolidation. An organization that has decided to combine many workloads on a smaller number of physical computers will need to determine how many new machines are needed. In this case, placing the maximum number of virtual machines on a host can result in significant cost savings. The primary drawback of optimizing resource utilization on each server is that this approach can make managing performance more difficult. For example, if several virtual machines require a large amount of CPU time, they're more likely to hit a constraint if the server was initially 85 percent utilized versus if a server was initially 60 percent utilized. Figure 7.8 provides a summary of the different approaches with respect to one type of resource—CPU utilization.

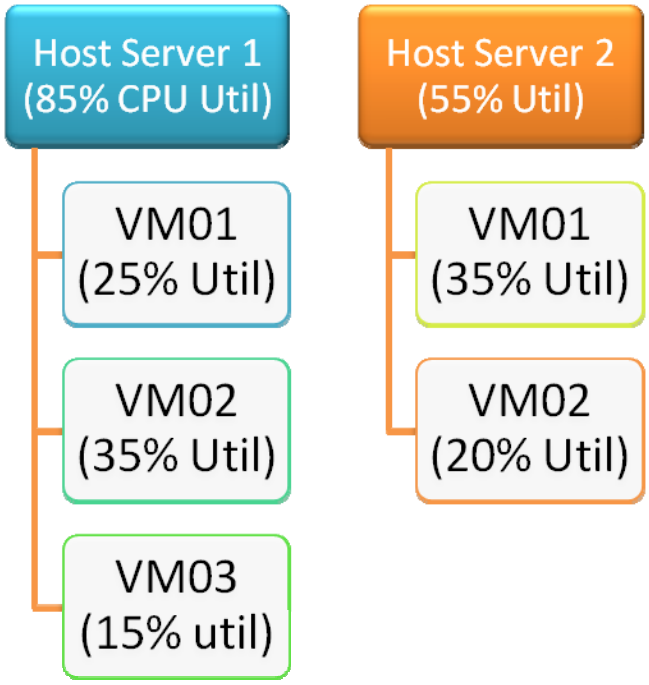

Figure 7.8: Load-balancing vs. utilization optimization approaches.

Maintaining Optimal Performance

Modern IT departments are expected to be dynamic and agile. The ability to quickly adapt to changing business and technical requirements can provide a significant strategic advantage to the entire enterprise. In fact, one of the primary motivators for adopting virtualization technology is the ability to quickly and easily move workloads throughout the data center.

Organizations can benefit from implementing a process of regularly reviewing resource utilization and making changes as required. Figure 7.9 shows possible steps in the process.

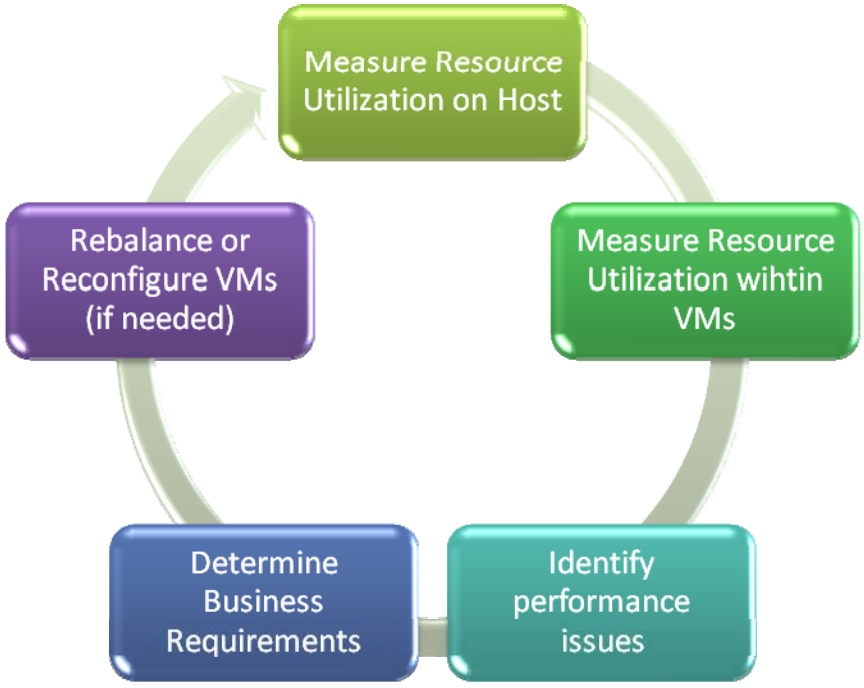

Figure 7.9: A process for maintaining optimal performance over time.

The interactions between host and virtual machine resource utilization can sometimes be difficult to manage because organizations have options about how they can adapt to changes. In this section, I'll cover details about deciding how to maintain optimal performance in a virtualized infrastructure.

Reasons for Performance Optimization

The process of maintaining optimal virtual machine placement is rarely a one-time effort. There are many reasons that performance requirements change over time:

- Changes in user activity—Over time, some applications will see increases in the number of users, while others might be used less frequently.
- Changes in virtual machine importance—As business directions change, new levels of importance might be placed on certain workloads. For example, a public Web server might experience significantly more activity during a marketing campaign.
- Unexpected usage patterns—Most applications experience spikes in utilization at particular times. When these patterns change, they can cause contention for host server resources. The end result is often system slowdowns that are noticeable to end users.
- Changes to SLAs—IT organizations that have implemented SLAs for their internal or external customers often have to react to changing terms. Frequently, customers will need to ensure additional performance, availability, and uptime.

All these changes can have significant effects on virtualization performance. Consequently, they often require changes to the configuration of placement of virtual machines.

Resource Optimization within a Virtual Machine

The first, and often easiest, approach to meeting changing technical requirements is to optimize the configuration of a virtual machine. Often, the changes can be accommodated through the reconfiguration of settings at the level of the host computer's virtualization layer. For example, most virtualization platforms provide the ability to easily reallocate resources such as CPU time and committed physical memory. The changes can be made quickly and easily by knowledgeable systems administrators. As the methods of managing resources differ for various virtualization platforms, expertise is required when performing the process manually. The assumption, of course, is that the host server has adequate unused capacity to meet these requirements.

Moving Virtual Machines

In other cases, it can be necessary to move a virtual machine to another host server. This is often done when server hardware is being upgraded or when virtual machines simply "outgrow" their current physical host. Physical machines have limitations on hardware specifications, so moving a virtual machine is sometimes unavoidable. Usually, the process of moving a virtual machine is significantly more complicated than that of reconfiguring it. The factors that must be considered include:

- Downtime—Although some platforms provide for performing a "live migration" of a virtual machine, there is usually at least a temporary slowdown in performance. Other platforms will require that the virtual machine be paused or shut down while its data files are moved between host servers.
- Migrating virtualization settings—Most virtual machine settings are stored in configuration files that can be reallocated to another host server. There are cases, however, in which the settings might not be compatible due to the configuration of the destination host computer. For example, physical disk paths might be different due to different logical volume structures. Systems administrators must take this into account when planning to move a virtual machine.
- Managing network settings—Depending on the details of the data center infrastructure, it's likely that a virtual machine's new host computer will be connected to a different switch or router. Additionally, virtual switch or virtual network settings might be different on the destination server.

These are just a few of the primary issues that can affect moving virtual machines between servers.

Moving Between Physical and Virtual Environments

Although virtualization technology can be used to meet the needs of a broad range of different types of workloads, it is not always the ideal solution. For example, if a particular application or service requires the full CPU, memory, disk, and network capabilities of a physical computer, it might be best to run it directly on the server hardware. By removing the virtualization layer, performance often improves. There are, however, drawbacks. For example, the ability to easily move the workload and to create snapshots or to roll back to an earlier point in time will be lost.

Over time, some virtual workloads might outgrow their virtual machines and need to run directly on server hardware. In other cases, applications that are deployed on physical hardware will be found to be under-utilizing that computer. To manage these situations, IT departments must consider moving workloads between physical and virtual environments. There are three common operations:

- Physical-to-Virtual (P2V)—Moving a workload from a physical computer to a virtual machine environment
- Virtual-to-Physical (V2P)—Moving a workload from a virtual machine environment to running directly on a physical computer
- Virtual-to-Virtual (V2V)—This operation is usually performed to move a virtual machine that has been built for one platform to another virtualization solution. The ability to perform V2V conversions also enables the creation and maintenance of only a single set of base virtual machine images that can then be deployed to many different target host servers.

Although these operations can be performed manually, they often take a significant amount of time and effort. Details such as OS configurations, application settings, and service details must be retained. Any errors can result in downtime. Fortunately, automated tools are available for performing these types of conversions.

Choosing an Optimization Approach

So far, we have discussed two approaches to maintaining virtual machine performance. The first involves reconfiguring virtual machine configuration settings on its current host server. The second approach involves moving the virtual machine to another physical host server. A good general rule is that organizations should reconfigure virtual machines whenever they *can*, and that they should move virtual machines whenever they *must*. Figure 7.10 compares the characteristics of these approaches.

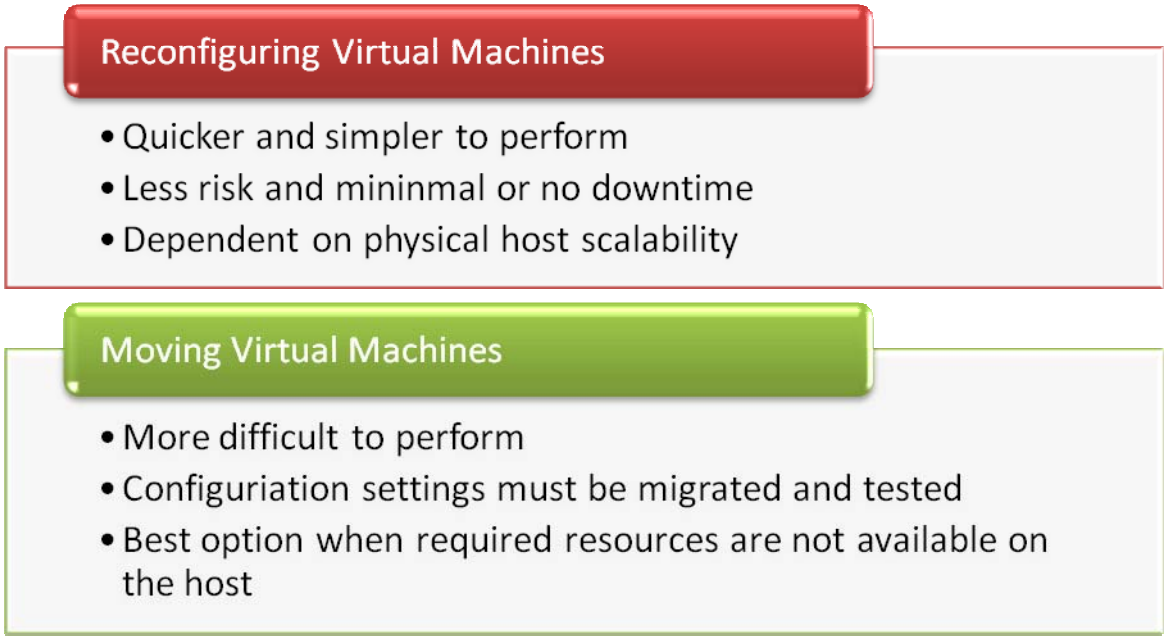

Figure 7.10: Comparing virtual machine optimization approaches.

Benefits of Automation

Many IT organizations tend to address performance issues on a reactive basis. That is, they wait until a problem is detected or reported by users and then manually perform the required changes. There are several problems with this approach. The first is that users often experience increased transaction times and loss of service when server resources are exhausted. That leads to frustration and potentially more serious problems such as downtime or even data loss. The other issue with reactive performance optimization is that systems administrators must spend a significant amount of time and effort in trying to make the best decisions about virtual machine configuration and placements. Later in this chapter, I'll cover ways in which automated performance management can simplify these tasks.

Managing System Resources

From a technical standpoint, modern computers have been designed with many features for increasing overall application performance. Some of these features are related to scalability (such as increasing the amount of physical memory), while others are based on implementing faster components (such as high-rotational-speed hard disks). In this section, I'll provide some suggestions about how systems administrators can optimize resource utilizations. Many of these methods apply equally to physical and virtual machines.

CPU Utilization

One of the most frequently measured statistics on virtual and physical computers is CPU utilization. Most OSs provide a simple method of viewing current CPU utilization. Often, there are ways to track utilization over time to establish an average. From the standpoint of the physical server, there are several ways to improve virtualization performance. Using dual- and multi-core CPUs can provide significant benefits. Most virtualization platforms spread virtual machine requests across different threads. These threads can then be allocated to specific CPUs to improve performance and reduce context-switching. Additionally, the CPU type can affect performance. AMD-V and Intel VT are two technologies that place virtualization-related extensions directly on the CPU. These processor extensions (which are sometimes referred to as *hardware-assisted virtualization*) have become commonplace on new server and desktop computers.

From the standpoint of virtual machines, administrators can balance resource allocations based on several factors. Some platforms allow for dedicating all or part of a specific CPU to a particular virtual machine. This can help reduce the overhead caused by switching processor context on heavily used systems. Additionally, it is possible to define priorities for virtual machines. At a simplified level, virtual machines can be given relative priorities (such as a single number between 1 and 100). When CPU utilization is maximized on the host server, virtual machines will be given processing time based on their importance. Another approach is to specifically place restrictions on the minimum and maximum amount of CPU resources that can be used by a specific workload. This can help avoid problems in which a failed guest OS is monopolizing host resources.

Memory Allocation

One of the primary scalability constraints related to the number of virtual machines that can be placed on a host computer is physical memory. Although some virtualization platforms allow for "over-committing" memory, others require that each virtual machine be given a dedicated amount that must be consumed if the virtual machine is running. Memory is a precious resource on physical host servers, so it's important to monitor the actual memory usage within virtual machines. Figure 7.11 shows an example of various configurations that might need adjustments. Based on this information, virtual machines can be reconfigured to optimally match allocated memory versus required memory.

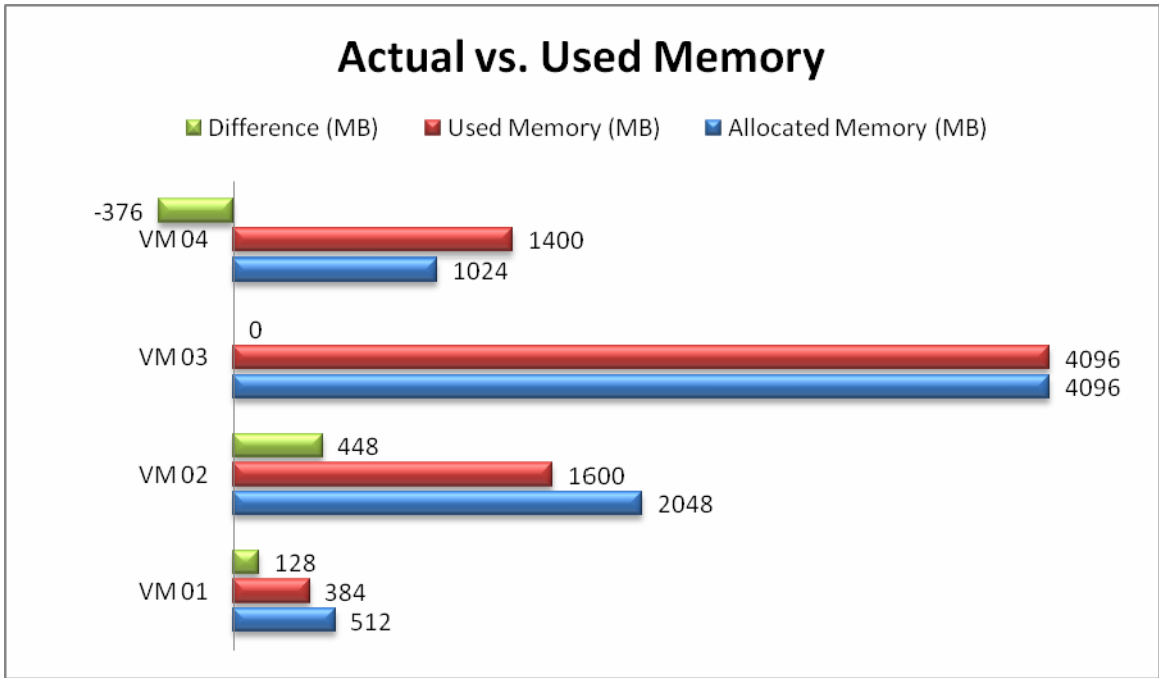

Figure 7.11: Allocated vs. used memory in a variety of virtual machines.

In addition to reconfiguring memory settings at the virtualization layer, it is often possible to reconfigure guest OSs, applications, and services. For example, application servers and database servers have several settings that can affect how memory is used by processes.

Disk Performance

When it comes to managing virtual machines, storage is often an important consideration. Apart from having adequate disk space to manage and maintaining all the required virtual hard disk files, disk read and write activity often experiences significant contention. The virtualization layer must coordinate disk operations, so bottlenecks frequently occur at this point. One of the best ways to optimize disk performance within a single host server is to distribute virtual hard disk files across physical disks. Most production data center servers will be configured with some type of direct-attached disk storage. RAID technology helps improve reliability (through redundancy) and performance (through striping of data across multiple physical devices). When servers have multiple physical arrays or disks, it is best to minimize disk contention by separating virtual hard disks and their associated undo or snapshot files.

IT organizations can also increase performance, scalability, and manageability by moving to network-based storage options. The most common approaches are to use Network-Attached Storage (NAS) devices or to invest in Storage Area Networks (SANs). Another useful option is to use iSCSI, which allows for performing block-level disk I/O over standard Ethernet connections. This provides the combined benefits of SANs (such as block-level disk access) and NAS devices (the ability to leverage copper-based network infrastructure).

Network Utilization

Almost every production server, OS, application, and service will rely upon network resources. There are numerous methods of achieving scalability and reliability from network infrastructures. With respect to virtualization, the first optimization step involves identifying the specific requirements for a virtual workload. Some virtual machines will require direct access to the Internet, while others will not. Some internal servers might be able to communicate only on virtual network connections. Most virtualization platforms allow virtual networks to appear as physical ones within the virtual machine. Actual data transfers, however, are performed in-memory.

On the host server, network performance can be improved by using NIC adapter teaming. Various vendors offer solutions that allow multiple network interfaces (or multiple ports on the same interface card) to be combined. The goal is to achieve reliability (through automatic failover) and scalability (through dynamic load-balancing). Having multiple independent network adapters can help segregate traffic. For example, internal- and external-facing workloads can use separate physical NIC devices to reduce contention. Another common technique is to dedicate one or more network connections solely for the purpose of performing backups. Overall, these approaches can often overcome many potential network I/O bottlenecks.

Automating Performance Management

Up to this point, the focus of this chapter has been on ways in which organizations can make the most of their IT investments. Some of the required tasks are listed in Figure 7.12.

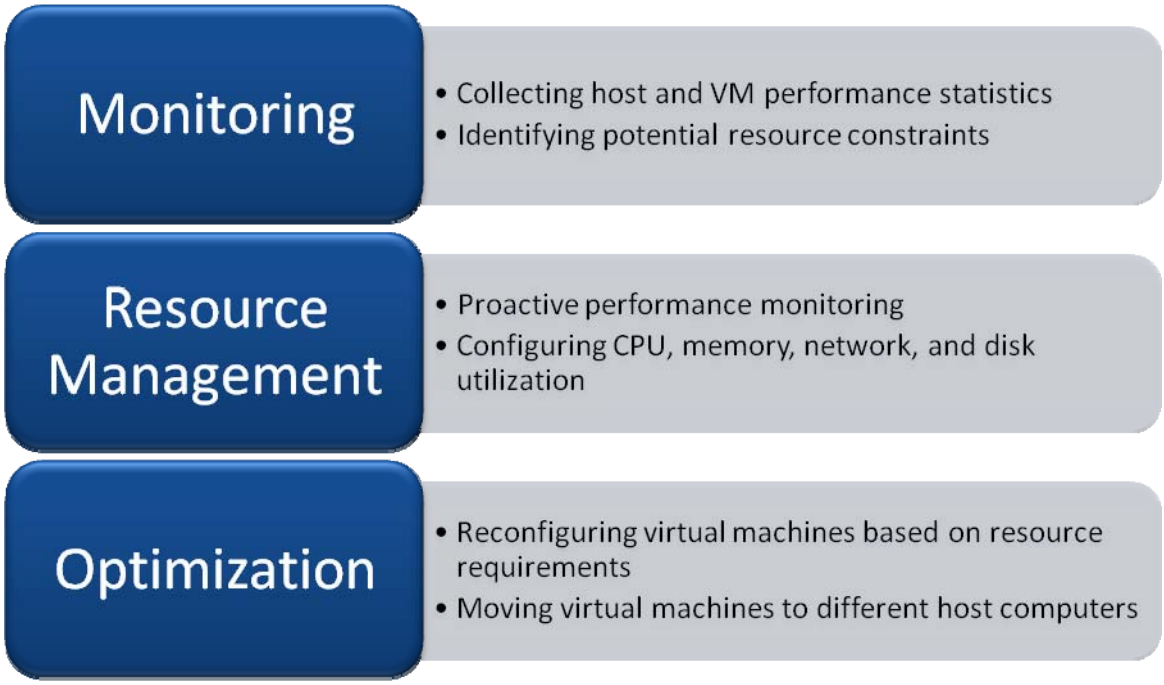

Figure 7.12: Typical virtualization performance optimization tasks.

Although the benefits of performing these tasks are readily apparent, it can often take a significant amount of time and effort to implement them. Most IT organizations face constraints based on budgets, human resources, and expertise. For these reasons, performance optimization is often given a lower priority to other tasks. An excellent way of monitoring and optimizing performance is through the use of automated performance management tools. In this section, I'll outline some of the ways in which they can help maximize IT resource utilization.

Monitoring Heterogeneous Environments

In typical data center environments, IT organizations have defined standard configurations for which OSs and hardware can be deployed into production. This helps ensure consistency and can make the systems easier to support. In virtual machine environments, however, the list of potential OSs is much longer. From a performance standpoint, systems administrators must have expertise in monitoring and optimizing settings for each of these different platforms.

An automated performance management solution can reduce this burden by providing automatic methods of collecting resource utilization information. Ideally, administrators will only need to use a single tool and user interface to keep track of CPU, memory, disk, and network utilization throughout the entire environment. Additionally, the details can be tracked over time and stored in a central database for easy access.

The solution should provide a consistent way of managing workloads that have been deployed in a variety of configurations. The environments can range from virtual deployments to physical deployments. A unified view that also includes clusters and various virtualization platforms can help reduce administrative effort.

Dynamic Resource Reallocation

Many common tasks related to managing virtualization resources can be automated. For example, when a particular virtual machine experiences significant levels of swapping, it is likely that it can benefit from more physical memory. An automated performance optimization solution can react to this situation by reconfiguring the virtual machine. Usage patterns change frequently and with little advanced warning, so the ability to constantly monitor, reevaluate, and reconfigure settings is essential to maintaining optimal utilization.

Reporting

Organizations that depend on virtualization technology often find that they have a need to answer important questions related to capacity planning and actual resource usage. Common questions include:

- What is my current average resource utilization in the data center?
- How much additional capacity is available on my current servers?
- How many more servers will I need to add in the next year?
- Which workloads are using the most resources?
- Which host servers are most/least utilized?
- What types of servers and virtualization configurations work best for my applications?

This information is used for budgeting and purchasing decisions, which can affect the entire organization. An automated performance management solution should provide the ability to collect, organize, and analyze the usage of physical and virtual resources throughout the environment. The details can then be presented in a report that will give significant insight into business and technical managers. Additionally, these details can be tracked over time to evaluate the overall efficiency of the IT department.

Choosing a Virtualization Approach

Organizations have a broad range of different types of IT solutions from which they can choose. OS virtualization is a standard approach that is used to provide compatibility with the widest array of workloads. Other options include hardware-level virtualization and application virtualization. In cases in which workloads can fully utilize a physical computer's resources, they should be run directly on a physical server. By eliminating the virtualization overhead, users will often see the best performance with this method. Finally, in some cases, applications must be scaled out. Clustering is a good option, assuming that it is supported by the underlying application or service platform.

Each of these approaches has its own performance-related benefits and drawbacks (for details, see Chapter 3). Insight that is gained by automated virtualization management solutions can help determine the best approach for a particular application or service. The end result is increased utilization and scalability in the data center.

Features to Look For

All but the smallest of virtualization implementations will need some form of automation. IT departments often find themselves replacing their "server sprawl" issue with the proliferation of virtual machines throughout the organization. Those virtual machines quickly become mission-critical, and the organization cannot afford to experience downtime or performance problems. All of this leads to the conclusion that automation is necessary. Overall, there are numerous features that are important when considering an investment in automated performance management. Figure 7.13 highlights the features that organizations should look for.

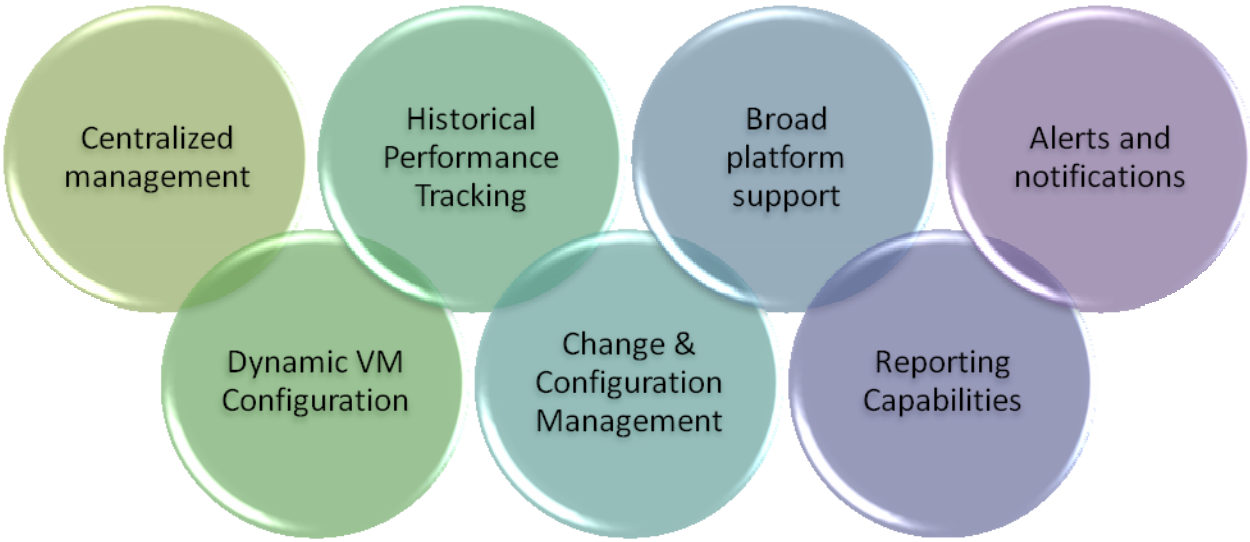

Figure 7.13: Features to look for in an automated performance optimization tool.

Summary

The focus of this chapter is on optimizing performance in a virtualized environment. The primary goal for many IT environments is to create a pool of computing resources into which various workloads can be seamlessly deployed. First, I covered details related to developing a solid performance optimization approach. These rules will be helpful for technical staff that is tasked with detecting and troubleshooting performance issues. Next, I focused on the topic of optimizing virtual machine placement. The goal is to improve resource utilization throughout the data center while still maintaining adequate performance for all workloads. Good decisions will take into consideration workload priorities, host computer requirements, virtual workload requirements, and the performance "compatibility" of different types of applications and services.

Performance optimization is an ongoing task, and it often requires re-evaluating deployment decisions. Changes in usage patterns and other details often require the IT department to quickly adapt. The two main approaches involve reconfiguring virtual machines and moving virtual machines between host servers. It might also become necessary to perform conversions between physical and virtual environments. From a technical standpoint, production deployments need to make efficient use of several computing resources: CPU, memory, disk, and network subsystems. I presented several ways of optimizing performance in these areas.

Finally, it's difficult to overlook the fact that monitoring and optimizing performance in typical IT environments can require significant time, effort, and expertise. A good solution is automation. I provided details related to helpful features in automated performance optimization tools. These include the support for a broad variety of platforms and environment types, the ability to dynamically reallocate resources, and powerful reporting capabilities. Overall, through the use of automation, organizations can ensure that they are maximizing the value of their physical and virtual investments.