Realtime publishers

The Definitive Guide™ To

Virtual Platform Management

sponsored by

ca technologies

Anil Desai

Copyright Statement

Chapter 6: Monitoring Virtualization Performance

Organizations that choose to invest in virtualization technology have often based their decisions on several important assumptions. The first is that by combining multiple independent workloads on a single physical computer, they will be able to better utilize the resources of this computer. The second assumption is that the performance of their virtualized production applications and services will be adequate to meet the needs of users. Although these goals are certainly achievable for the vast majority of IT projects, there are many factors that must be taken into account.

Usage patterns for workloads can change dramatically over time. A small application that was originally only intended for use by several people in the Accounting department might now be relied upon by dozens of end users. Alternatively, some applications and virtual machines might be granted too many resources. Regardless of the reason, these changes raise the importance of managing the performance of virtual workloads and the physical hosts on which they are running.

The focus of this chapter is on monitoring performance for virtualized infrastructures. It will begin by presenting potential business benefits of implementing proactive performance monitoring. It will then present details related to how you can monitor physical and virtual workloads, including which statistics are useful. Then, I'll switch more to the management side of IT by looking at the many benefits of implementing and monitoring Service Level Agreements (SLAs). Next, I'll look at ways in which IT staff can predict virtualization performance through benchmarks and other tests. Finally, the chapter will conclude with a discussion of ways in which performance management can be automated.

Business Benefits of Performance Monitoring

Overall, the goal of tracking system performance is to maximize the business benefits of utilizing virtualization technology. Although IT organizations' initial deployment of virtual machines can provide significant business and technical advantages, there is often room for improvement. Before we look at the technical details of implementing performance management, it's worthwhile to look at some of the applications and benefits of proactively monitoring system performance. In most cases, it will take additional effort from systems administrators and IT staff to perform these actions, so IT organizations should consider the value of their efforts.

Capacity Planning

Capacity planning involves determining future business and technology needs and comparing them with the current environment. IT departments often face a difficult challenge in making these predications. On one hand, they're faced with limitations on budgets, physical resources, expertise, and personnel to build out their infrastructures. On the other hand, business units often require new servers and systems to be deployed quickly in order to support new initiatives. In many situations, systems administrators tend to wait until the last minute to decide whether the purchase of new capital equipment is necessary. This can make it difficult for accounting staff and technical personnel to meet their goals.

In the "old days" of computing, it was common for organizations to use a standard rule of "one application per server." Although this approach often led to underutilized computers and the problem of server sprawl, it made planning somewhat simpler. For example, if the Marketing department decided that it needed to deploy a new Customer Relationship Management (CRM) software application, part of the evaluation and purchasing process would be to acquire the necessary hardware. The specifications for the hardware were often provided by the vendor of the product or may have been determined through basic testing. Regardless of the benefits or drawbacks, the decisions were easier—new applications meant new servers. To find available capacity, IT staff would look for old applications or machines that were no longer needed and re-purpose them.

When running in a virtualized infrastructure, it is often more difficult to determine the total amount of available capacity. New workloads do not necessarily require new servers. Instead, it's important to look at the actual utilization of resources on each host system. A simple example is the need to support a new workload that requires 2.0GB of physical memory. IT staff could look at the configuration of existing hardware to find any host server that has this additional capacity. Other statistics can often be more complicated to calculate. For example, CPU time, disk utilization, and network utilization can vary significantly over time. Performance considerations should include average and peak utilization of these resources.

Through the use of performance monitoring, IT departments can get a clearer picture of the total computing capabilities that they are supporting. To make better placement decisions, they must consider how resources on physical servers are being used. Based on that data, they can make better decisions about which resources might be needed in the future. Predictions related to capacity can be helpful to businesses in numerous ways. Figure 6.1 provides an overview.

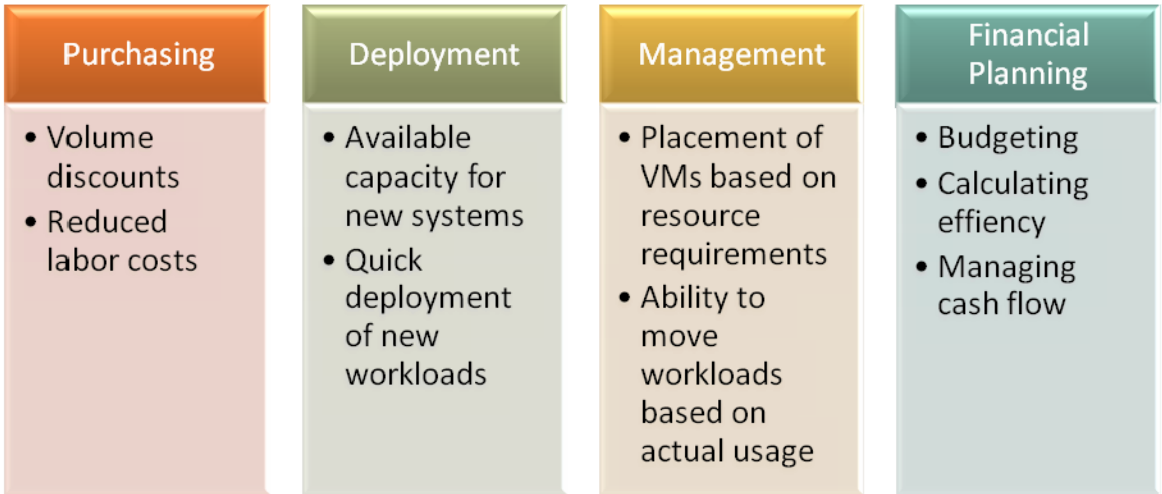

Figure 6.1: Benefits of performing capacity planning.

For example, when purchasing physical server hardware, it's often more cost-effective and convenient to acquire numerous servers at a time, as many vendors offer volume discounts. Purchases might be made regularly to meet the needs for the month, the quarter, or the year (depending on the needs of the organization). The benefit is that new workloads can be deployed almost instantly, when the need arises. Although there will always be some variability between actual versus estimated utilization, virtualization technology allows IT departments to move workloads based on requirements. Finally, from the standpoint of financial management and budgeting, it's often helpful to be able to predict the cost of additional resources that might be required in the future.

Maximizing Hardware Utilization

For many businesses, the most immediate benefit of moving workloads to a virtual environment is that of reducing costs. By consolidating different types of applications and services onto physical servers that have additional capacity, businesses' power, cooling, and physical space requirements can be lowered dramatically. However, the primary goal for achieving optimal efficiency involves predicting which workloads and settings will provide the best utilization levels. This can often be a balancing act: If too few virtual machines are placed on a system, server resources may be left underutilized. But if the host computer is over-worked, it's possible that the performance of one or more virtual machines will be unacceptable. Figure 6.2 illustrates the situation.

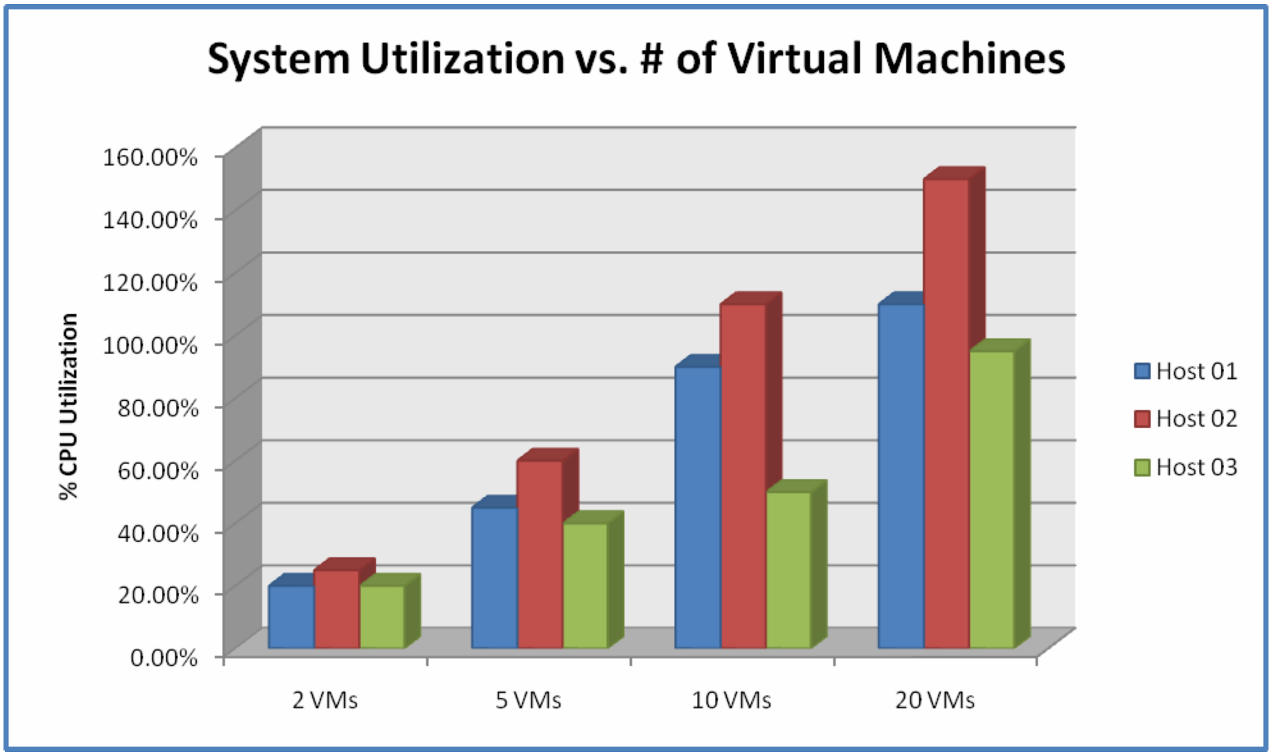

Figure 6.2: A graph of system utilization based on the number of virtual machines.

Of course, in real production environments, virtual machines will be running many different kinds of functions. A Web server virtual machine may have very low disk I/O but might require significant memory resources and CPU time during periods of peak load. Another workload—perhaps a virtualized departmental file server—might require significant disk time but have low CPU and memory requirements. These two virtual machines would be good candidates for being consolidated onto the same physical host server. The net result is that more of the physical server's resources could be utilized. To make these types of decisions, it's important to monitor actual server performance statistics. Subsystems such as CPU, memory, disk, and network components have theoretical and practical limitations, and without accurate statistics, it is difficult to predict which workloads are the most compatible.

Tracking the End-User Experience

The ultimate goal for many IT service and application deployments is to ensure that end users are able to complete their jobs. From a performance standpoint, this can be measured based on application response times and throughput. When virtualized programs start to run slowly, the result is a less-than-ideal experience for end users. In some cases, this will result in a call to the service desk, which in turn will require a systems administrator to make changes to system settings. For example, the amount of physical memory that is allocated to a virtual machine can be increased (assuming that unallocated RAM is available) or the virtual machine could be moved to another server (if it is not). Alternatively, workloads might be moved to other physical host servers to reduce contention for system resources. Usually, the changes will require some amount of downtime and disruption to users.

Overall, the goal for IT departments should be to proactively detect performance-related problems. Often, by monitoring CPU, memory, disk, network, and other related statistics, they can determine when resources are being overused. Data based on performance trends can be used to predict when a virtual machine or physical host server will need to be reconfigured. This additional information can enable IT staff to make performance-related modifications during scheduled downtime windows or during periods of low usage, thereby reducing disruptions to end users.

Ensuring Reliability and Availability

Whether a production service or application is running directly on a physical server or within a virtual machine environment, it is often relied upon by many users in the environment. That places the IT management focus on ensuring the availability of these important programs. When working in a physical environment, availability is often measured in terms of "uptime," expressed as a percentage (for example, 99.9%). There are numerous ways to monitor the health of a server, including the monitoring of event logs and network scanning tools.

For the most part, the same methods and techniques apply to ensuring availability in a virtualized environment. Virtual machines run guest operating systems (OSs), so they usually have methods for ensuring that the system is working adequately. In many cases, however, there are unique performance-related issues that might cause problems in reliability. For example, a lack of memory could lead to slow response times. Often, this situation will be perceived by users to be "downtime" with relation to the systems they're trying to use.

By implementing proactive performance management, IT organizations can reduce or minimize downtime and availability issues. For example, if memory consumption or CPU utilization numbers are consistently higher than expected, the virtualized workload can be moved to another host. Or it can be reconfigured with more appropriate settings based on actual usage. Compared with the process of moving applications between physical servers (or upgrading the hardware configuration of a host machine), the corrective steps can be simple.

Monitoring Virtualization Performance

There is a management aphorism that states that, “If you can’t measure it, you can’t manage it.” The idea is that, to improve operations, you must have a good idea of the current state of the environment. This is certainly true when it comes to managing virtualized environments. Although the process of managing physical servers can be time consuming and tedious, keeping track of virtual machines can be even more challenging. Virtual machines can be deployed in a matter of minutes, and they can easily be moved between servers. Compared with the reliability of visually inspecting server racks, new considerations must be taken into account.

Goals of Monitoring Performance

Before moving into the details of ways in which virtualization performance can be monitored, it will be helpful to look at the potential goals and benefits. Overall, performance monitoring can affect many areas of operations in a typical IT environment. Figure 6.3 provides an overview of these areas, along with examples.

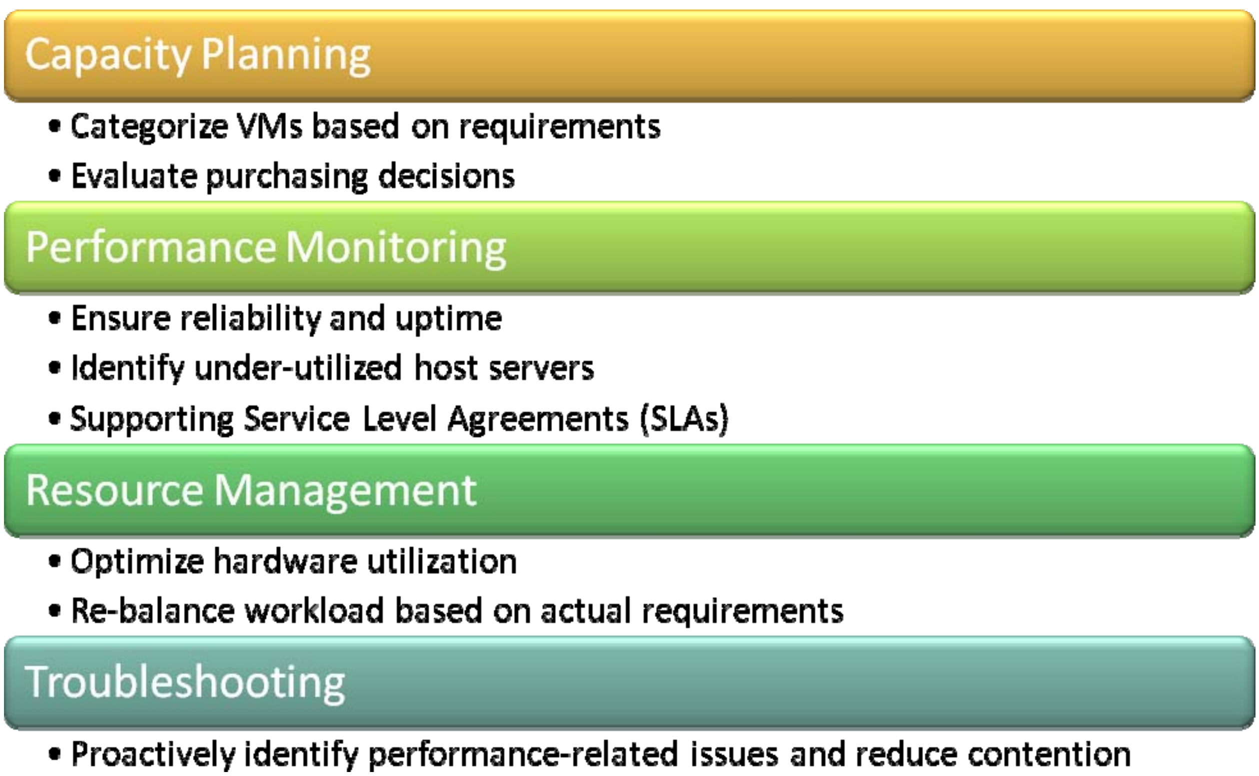

Figure 6.3: Goals and benefits of monitoring virtualization performance.

In addition to these goals and benefits, performance monitoring can help answer important questions related to these areas. Examples include:

- Capacity Planning:
  - What is the maximum number of virtual machines I can support based on current hardware?
  - How many virtual machines can I run on a specific host?
  - What types of investments will I need to meet future needs?
- Performance
  - How can I identify potential performance issues before they affect production operations?
  - Which types of workloads are best-suited for running within a virtual environment?
- Scalability
  - What are the practical limitations on virtual machine performance?
  - Which is the best virtualization approach for a particular type of workload?

Understanding Unique Virtualization Issues

Many of the goals and technical details related to monitoring performance for virtual machines are similar to those of monitoring physical computers. Supporting virtualization does, however, add new challenges that must be addressed. First, there is the issue of running multiple independent workloads on the same hardware. In general, virtual machines are unaware of each other, and it is up to the virtualization layer to coordinate and complete requests. Apart from standard virtualization overhead, the guest OS and applications often try to monopolize basic system resources. Additionally, it's common for IT environments to support many types of OSs and applications—often on the same host server. This requires technical expertise and training in order to make the most of each particular situation. Finally, the large number of virtual machines in many environments makes it difficult to keep track of all the systems that are deployed. Often, IT departments may be unaware of new virtual machines until they receive performance complaints from users.

Statistics to Measure

Most OSs provide built-in methods of collecting performance-related data. When monitoring physical servers and virtual machines, systems administrators need to know which types of statistics they should be most interested in. The primary types of resources that effect modern servers include CPU, memory, hard disk, and network components. In addition, specific applications and hardware can create bottlenecks. Figure 6.4 provides an overview of types of statistics that should be measured.

> It's important to note that specific terminology might differ between the host OS, the guest OS, and the virtualization layer. For example, paging and caching might be measured and reported in a different manner on Windows and Linux-based platforms.

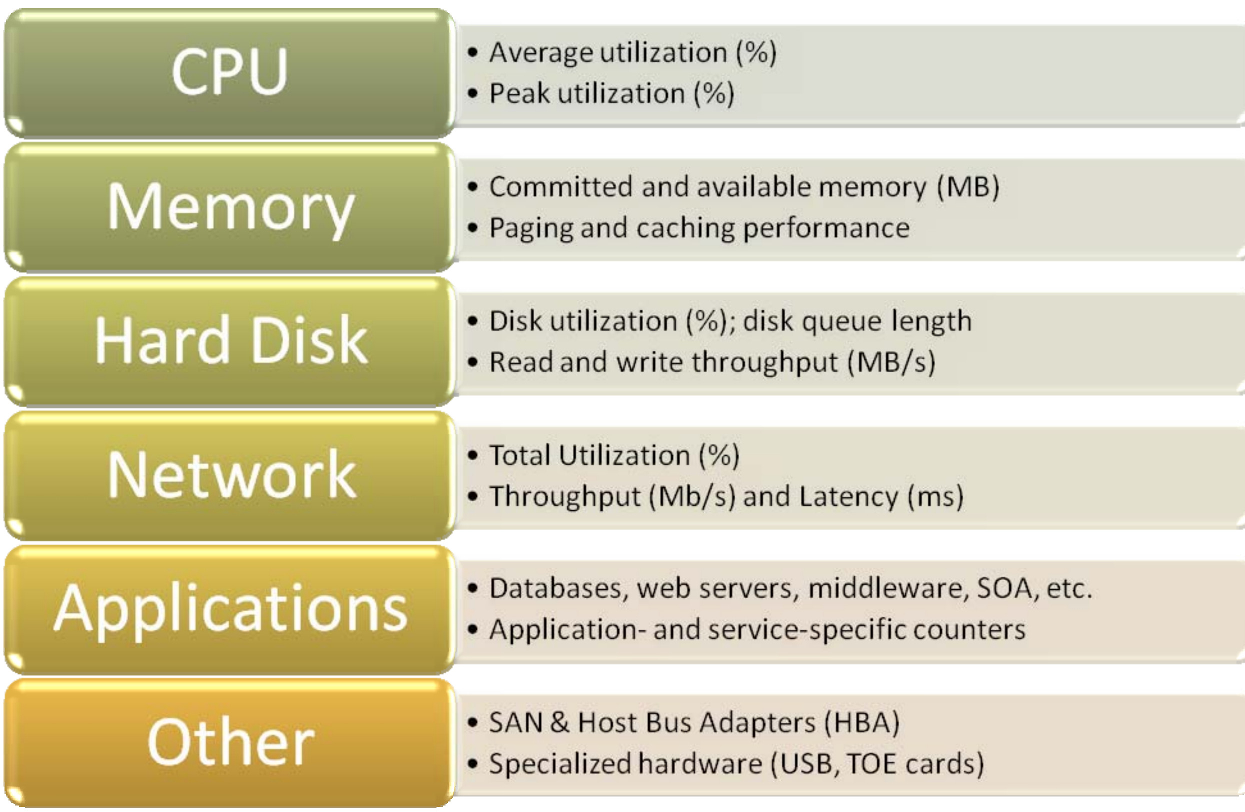

Figure 6.4: Commonly monitored performance statistics.

When this list of performance information is multiplied by the number of physical and virtual servers in an IT environment, the amount of data that must be analyzed can be overwhelming. One common approach that is used by IT departments is to address performance issues as they arise. For example, if several virtual machines on a particular host appear to slow at a particular time of day, administrators might start collecting details related to CPU utilization on the server. The information that is collected can help determine whether a physical upgrade or server reconfiguration operations are required.

Monitoring Host Performance

There are two main levels at which virtualization performance can be measured. The first is at the level of the host computer and host OS. The primary goal is to get an overview of total hardware resource utilization on the physical computer itself. If, for example, CPU utilization remains relatively low on the server, it is possible that the computer has additional capacity for supporting more virtual machines. Figure 6.5 shows an example of the result of tracking these statistics over time.

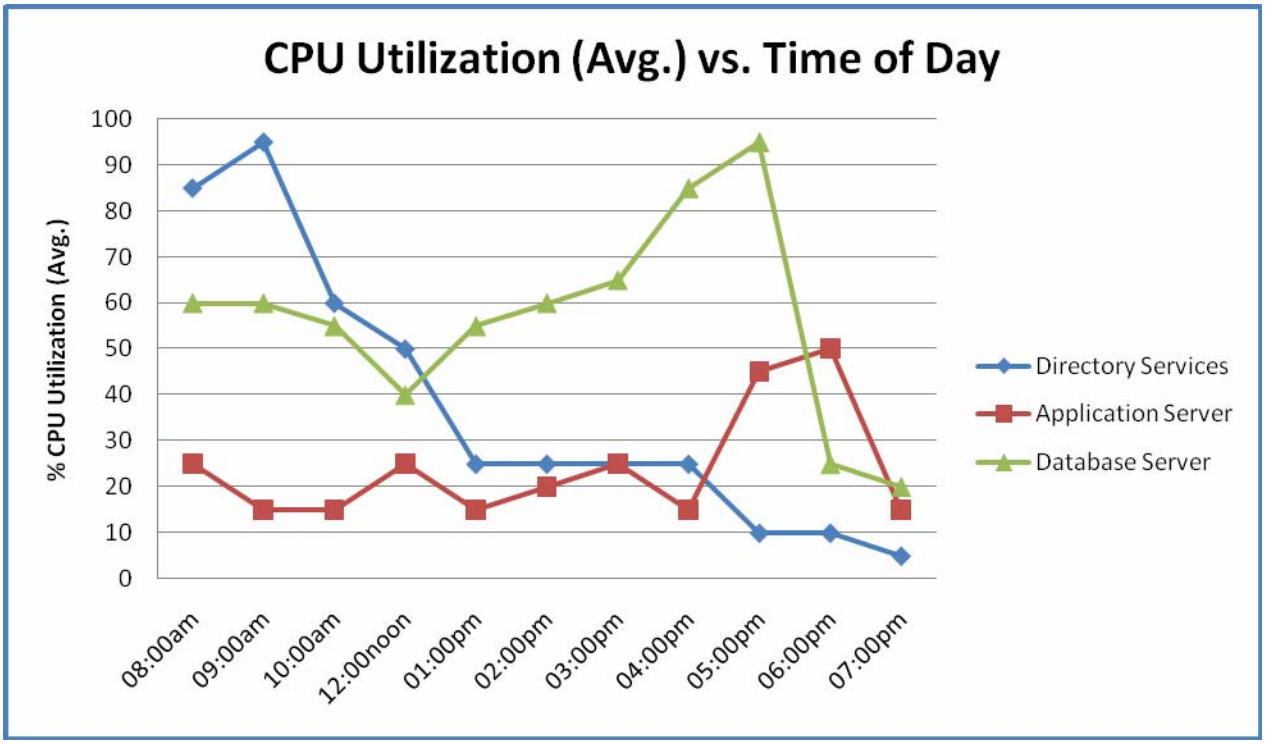

Figure 6.5: Monitoring average CPU utilization vs. time of day.

Another example is measuring disk-related performance. If several virtual workloads require significant disk access, it is likely that disk controllers or physical disks themselves are causing a bottleneck. This situation suggests that virtual machines should be moved to other servers or the storage system should be upgraded.

Monitoring performance at the host level is also useful for establishing a baseline of overall system utilization over time. Performance tools can often be configured to track important details and store them to a file or central location for analysis. For example, the Windows System Monitor tool can write data to a binary file that can later be analyzed by IT staff.

Monitoring Guest Performance

The second approach to monitoring virtualization performance is at the level of the guest OS. Although monitoring host performance provides an aggregate view of hardware resource utilization, it does not specify which virtual machine(s) might be causing the majority of the load on the system. Often, the goal of monitoring guest OSs is to identify the root cause of resource usage or to determine the specific requirements for a virtualized workload. Figure 6.6 provides an example.

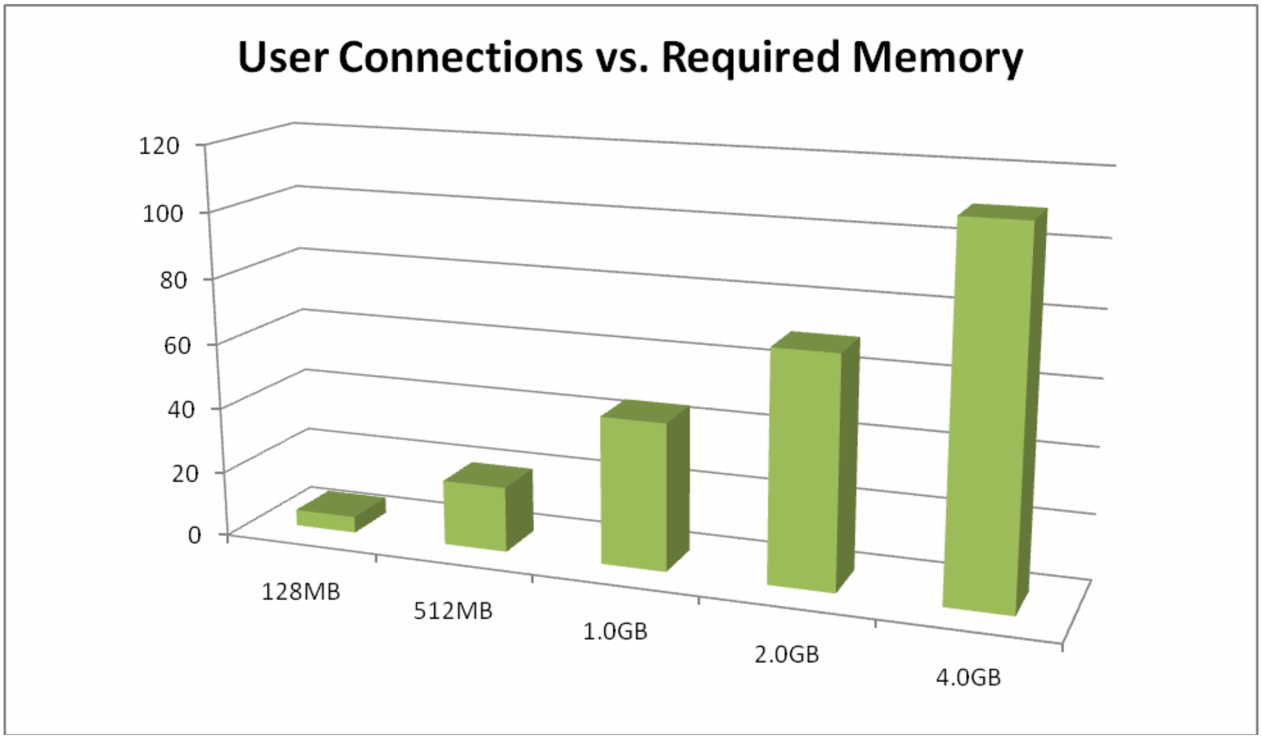

Figure 6.6: Measuring memory requirements against number of concurrent user connections for a virtualized workload.

In this chart, a particular virtual machine requires different amounts of physical memory based on the number of active user connections. For situations in which there will only be a small number of active users, relatively little memory can be allocated to the virtual machine. As the needs increase, it is likely that memory-related paging will occur, resulting in a slowdown of the virtual workload. This could also end up adversely affecting the performance of other virtual machines on the system. One potential resolution to the problem is to increase the amount of memory that is dedicated for the virtual machine.

Guest-level monitoring is also useful for troubleshooting specific performance issues and for isolating the root cause. For example, a particular virtual machine might be monopolizing overall CPU time due to a failed application or service. The host would only show that overall CPU utilization was high, while examining specific guests would help pinpoint the actual source of the problem.

Implementing SLAs

IT departments often face numerous challenges when attempting to meet the needs of businesses that they support. From a high-level management standpoint, IT departments are often seen as "cost centers" rather than as strategic partners that can help the company meet its goals. Much of the reason for this perception is a lack of alignment between IT departments' goals and the goals of the departments they support. Without adequate communications about expectations, it is difficult for IT staff to prioritize their technology investments in a way that is consistent with the approach of the rest of the organization. Additionally, IT-related costs often represent a significant proportion of total expenditures, so it is important to align business and technology.

Various business models also rely upon service-level understandings. For example, an Internet Service Provider (ISP) might guarantee a certain level of peak bandwidth to be available. They might also include uptime guarantees, along with associated penalties. Hosting providers often base their pricing on expected availability of their systems as well as metrics such as the amount of time it would take to failover to a disaster recovery site.

Virtualization Service Level Challenges

Although typical disconnects between IT departments and the users they support are often independent of the technology itself, virtualization technology can present new issues. First, in many organizations, there is a lack of trust in virtual infrastructures. IT management can readily see the many benefits of combining workloads through the use of virtual machines, but the rest of the organization might not be so quick to reach the same conclusion. There is often a perception that virtual machines will perform noticeably slower than their physical counterparts. Many business units have become accustomed to the "one server per application or service" model. The downsides include server sprawl and added data center costs—two problems that are often primary IT challenges. Nevertheless, the result is resistance toward using virtualization.

SLA Goals

SLAs are policies that are put in place to help document and measure mutually agreed-upon levels of service. The primary goal is to align IT departments' directions with the business users they support. The focus of the underlying technology used to meet these goals is determined by the IT department. However, the actual goals themselves—such as performance, risk of data loss, reliability, and availability—are communicated based on metrics.

Defining SLA Approaches

When defining an SLA, it's most important to begin with focusing on business requirements. This can often be difficult, especially when business leaders have a semi-technical background. For example, a Marketing VP might request that a certain application or service be clustered for high availability. The business goal is to implement high availability, and the focus of the SLA should be on that. The best technical implementation should be determined by IT departments that may, in some cases, find that clustering is not the best solution.

Perhaps the most important aspect of the SLA development process is that it should involve all areas of the organization. This includes executive management, representative end users, and business management staff. In addition, IT representatives with technology and management expertise should be included. The act of creating a new SLA should be based on a give-and-take negotiation. For example, when asked how much downtime or data loss is acceptable for a particular application or service being considered, the initial reaction from business stakeholders is "zero." However, when factoring in IT constraints and considerations (such as limited budgets, physical hardware, personnel, and expertise), most organizations will find that it is worthwhile to make tradeoffs. For example, an amount of data loss equal to 15 minutes of time might be acceptable if the cost savings are many thousands of dollars.

Creating a New SLA

It is a good idea to develop a process for creating new SLAs within an organization. Figure 6.7 provides an overview of typical steps in the process of creating a new SLA.

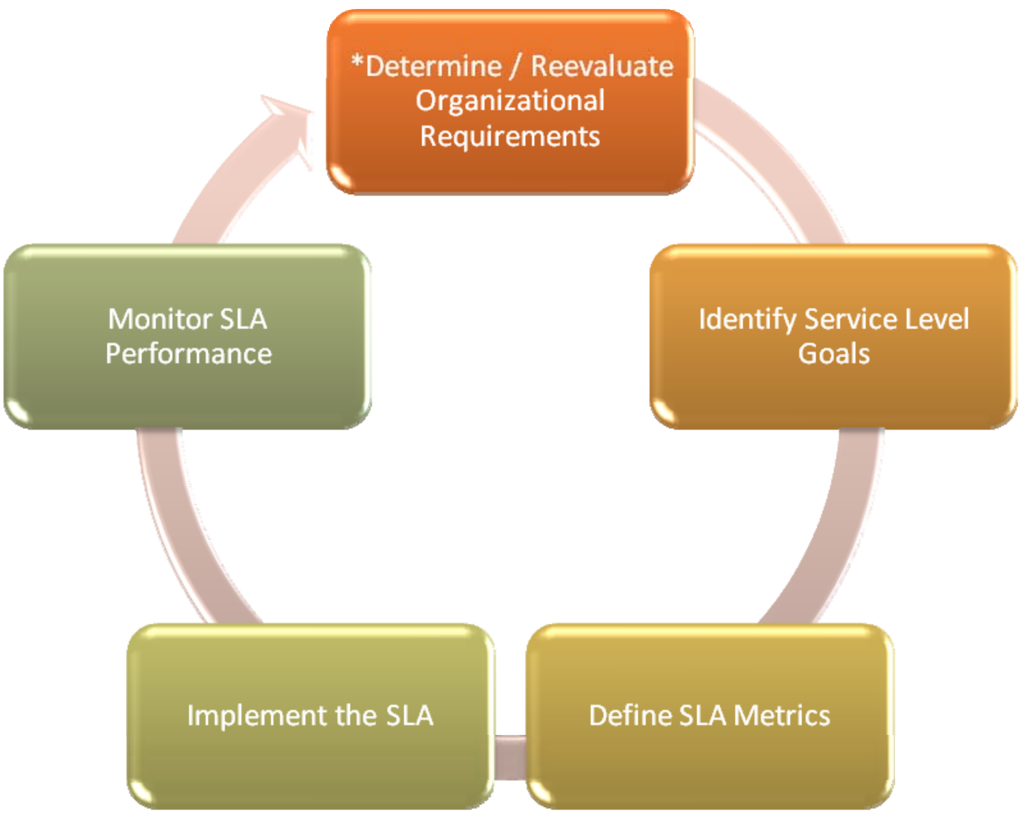

Figure 6.7: Steps in the process of creating new SLAs.

The first important step involves determining the specific requirements of the organization. All too often, IT departments skip this step and focus instead on the technology itself. The process may begin with identifying strategic business initiatives as outlined by executive management, or it may attempt to address a common user complaint that has been recognized by the IT department. Regardless of the source of the issue, it should be one that management and users agree is worth spending time and effort to solve.

Developing SLA Metrics

The next step of the SLA creation process involves identifying target levels of service that might be required. Some examples from throughout an organization might include:

- Increased uptime for a specific application or service
- Quick resolution of high-priority Help desk issues
- Rapid deployment of new virtual machines
- Improved performance for several mission-critical workloads

These relatively high-level statements can greatly help organizations determine their priorities. Once the problems or areas for improvement are identified, it's time to look at quantifying the desired levels of service. Often, this begins with a statement of an ideal environment, and then factors in various IT-related constraints. Examples might include:

- Limit scheduled downtime for Application A to 15 minutes per month
- Resolve 95% of Severity 1 Help desk issues within 2 hours
- Deploy new basic virtual machines within 4 business hours of the initial request
- Reduce response times for Application B so that the majority of common transactions are completed within 2 seconds

The primary benefit of these statements is that they can be measured. For example, downtime can be measure in minutes (or, hopefully, seconds). And, virtual machine deployment times can be tracked across the organization. Table 6.1 provides a summary of typical SLA-related metrics and service-level goals.

SLA Area	Metrics	Goal	Notes / Terms
CRM Application Uptime	Percent availability	99.9% availability	• Excludes planned downtime for maintenance operations and downtime due to unrelated network issues • Major application updates might require additional planned downtime
Service Desk: Level 1 Issue Resolution	Issue Resolution Time	4 business hours	• Include definition of "Level 1 Issues"
Service Desk: Level 2 Issue Resolution	Issue Resolution Time	8 business hours	• Time is measured from original submission of issue to the Service desk • Include definition of "Level 2 Issues"
Engineering: New Server Deployments (Physical machine)	Time to Deployment	3 days	• Time is measured from when formal change request has been approved • SLA applies only to servers that will be hosted within the data center
Engineering: New Server Deployments (Virtual machine)	Time to Deployment	2 hours	• Virtual machines must use one of the three standard configuration profiles • Time is measured from when formal change request has been approved

Table 6.1: Examples of virtualization SLA metrics and goals.

Virtualization-Related Services

IT departments can be seen as service providers that are attempting to meet the needs of their "customers." The customers themselves might range from internal business units and employees to external partners. A focus on virtualization technology often comes with several types of services that are important considerations. Figure 6.8 provides an overview.

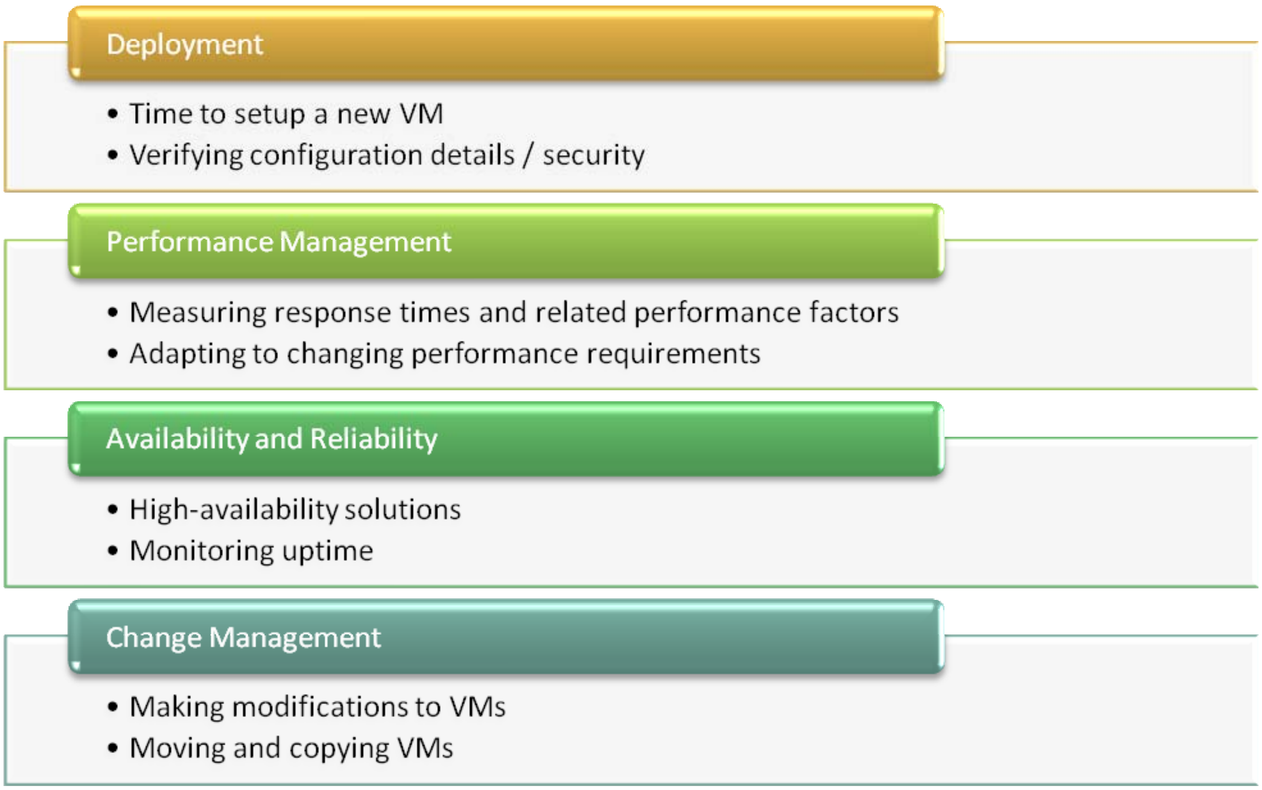

Figure 6.8: Examples of typical virtualization-related services.

The performance of providing each of these types of services can usually be measured using SLA metrics. For example, the response time for a particular transaction can be simulated and the duration of the process could be recorded. This measurement may be made many times during a typical day. For example, a public e-commerce organization might want to ensure that the time it takes to place an online order never exceeds 15 seconds. Sample transactions can be performed throughout the day to help ensure that the system always meets this requirement.

Implementing and Monitoring SLAs

Once service-level metrics and goals have been properly defined, IT organizations can begin to work on the technical implementation of SLA management. A common first step is to upgrade the current infrastructure to support specific requirements. For example, if higher availability is required, perhaps a standby server connected by a high-speed, low-latency connection is in order. In other cases, technical staff might need to be retrained or existing processes modified to meet expectations.

In addition to technology and process considerations, it is important to have a method of tracking SLA-related metrics. In some cases, monitoring of SLA-related performance will need to be somewhat manual. Examples might include Help desk resolution times (which require users to manually record the amount of time it took to close an issue). More commonly, however, automated performance management tools can be used to simplify the process of tracking SLA-related performance. Figure 6.9 provides an example of automated steps that may be performed.

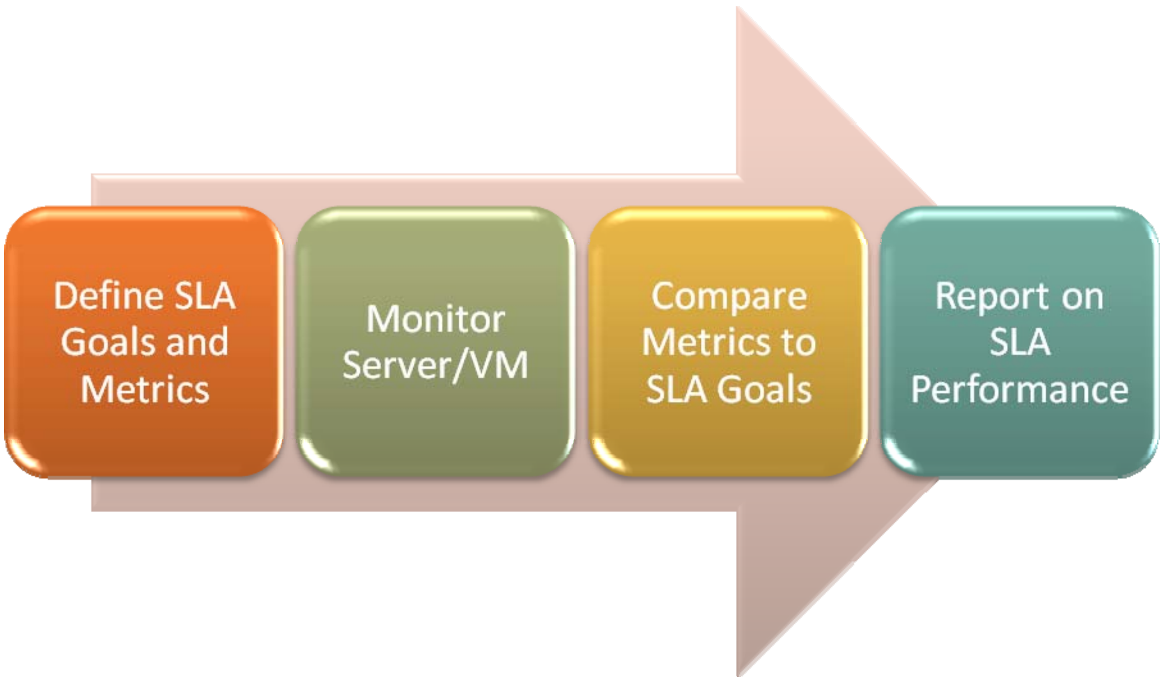

Figure 6.9: Steps in an automated SLA performance management process.

The first step involves defining the actual SLA goals and metrics within an automated performance management solution. Usually, the types of thresholds that are defined can be quantified. For example, the average amount of CPU utilization might be correlated with transaction processing times. By periodically measuring transaction times, the system can use this information to compare against SLA goals. Organizations can then generate reports based on this data. For maximum visibility, IT organizations can make SLA-based reports available on a company intranet or other shared location. This helps all areas of the organization monitor critical IT functions.

Responding to SLA Issues

In most IT environments, it's simply a matter of time before one or more SLA-defined goals are not met. In some cases, hardware limitations or unpredicted usage patterns might be the root cause. It is important to be able to respond to these issues and to identify their root causes. Often, performance monitoring can help. For example, if a particular application started responding too slowly, the likely cause could have a number of factors. By combining details related to CPU, memory, disk, and network performance, likely causes can be determined and resolved quickly. In more serious cases (such as repeated SLA violations), the IT department may need to consider the purchase of additional host server hardware or upgrades to current systems.

Reviewing and Updating SLAs

It is important for organizations to consider SLAs as changeable over time. Companies need to respond quickly to changes in business and technical requirements. This can result in an unexpected increase in the use of a particular application or service and corresponding decreases in other areas. The levels of required service at a particular point in time may not correspond to what is most important several months later. For these reasons, organizations should allot time to regularly review current SLA terms and details. If high availability is no longer a primary concern for an application, the IT department can save significant costs by configuring it with lower importance. The hardware, software, and labor investments can then be reallocated to other more valuable systems. Overall, the goal should be to regularly review SLAs to ensure that they are still aligned with end-users' requirements.

Reducing Costs with SLAs

The process of developing, implementing, and monitoring SLAs can be significant in organizations that are putting them in place for the first time. Although the organizational benefits of better aligning IT departments with the rest of the business are usually evident, there are often more tangible benefits. Table 6.2 provides an example of ways in which organizations can use SLA-based details as the basis for reducing costs.

Product or Service	Proposed Investment	Current Cost or Service Level	New Cost or Service Level	Benefit
Virtual Machine Deployment	Investment in automated deployment tools	$120 / VM	$40 / VM	• $80 savings per virtual machine deployed • Reduced virtual machine deployment time
Performance Monitoring	Automated performance management tools	65% avg. server utilization	85% avg. server utilization	• Verify performance of existing virtual machines • Automated corrective actions for performance issues
Host Server Deployment	Investment in automated server deployment and configuration tools	$450 / server	$125 / server	• $325 savings per server deployed • Quicker server deployment
Average Time to Complete Basic Virtual Machine Management Requests	Purchase of automated Help desk system	~4 hours	~1.25 hours	• Reduction in average resolution time by ~3.75 hours per issue • Ability to report on problem resolution • Option to create a self-help knowledge base
Server Patch Management	Automated security management solution.	~ 12 hours per server per year	~ 2 hours per server per year	• Reduction in time and effort required to maintain servers by approximately 10 hours per year • Decreased latency between patch release and patch deployment • Auditing of server configurations

Table 6.2: Ways in which organizations can reduce costs through the use of automation.

Although the numbers provided in the table are hypothetical, the main idea is that IT management can use SLA requirements to identify areas of improvement. Usually, there is an upfront investment, but the cost savings are easy to identify. For example, in the case of deploying new virtual machines, the use of automated deployment tools can reduce the rollout costs significantly. Additionally, new virtual machines can be rolled out more quickly, benefiting end users that often make requests. Similarly, there are numerous benefits to automating configuration management and patch deployment.

Overall, SLAs are a great way for IT departments to manage many of the virtualization-related challenges that they face. Through the use of well-defined business requirements, the entire organization can synchronize its time and effort on the requirements that have the most value. Performance monitoring is a critical component of the process, as it helps ensure that SLA goals are being met.

Testing Virtualization Performance

When considering performance management, it is often important to be able to determine the capacity of a particular system. So far, the focus of this chapter has been on monitoring current performance statistics in a production environment. But what about situations in which IT departments need to predict future performance? Clearly, it’s not a good idea to place a system in production and let end users determine whether it’s fast enough. Common questions are related to determining the actual physical configuration of a host server in order to meet particular requirements. The hardware specifications, costs, and data center requirements can vary significantly based on the type of solution that is selected. Common options include rack-mounted servers, blade servers, and clustering solutions.

To make accurate predictions about expected performance on both physical servers and virtual ones, IT departments should be prepared to implement some type of performance testing. Figure 6.10 provides an overview of the different types of performance testing approaches, along with their pros and cons. Next, let’s look at the details of how each type of test is performed.

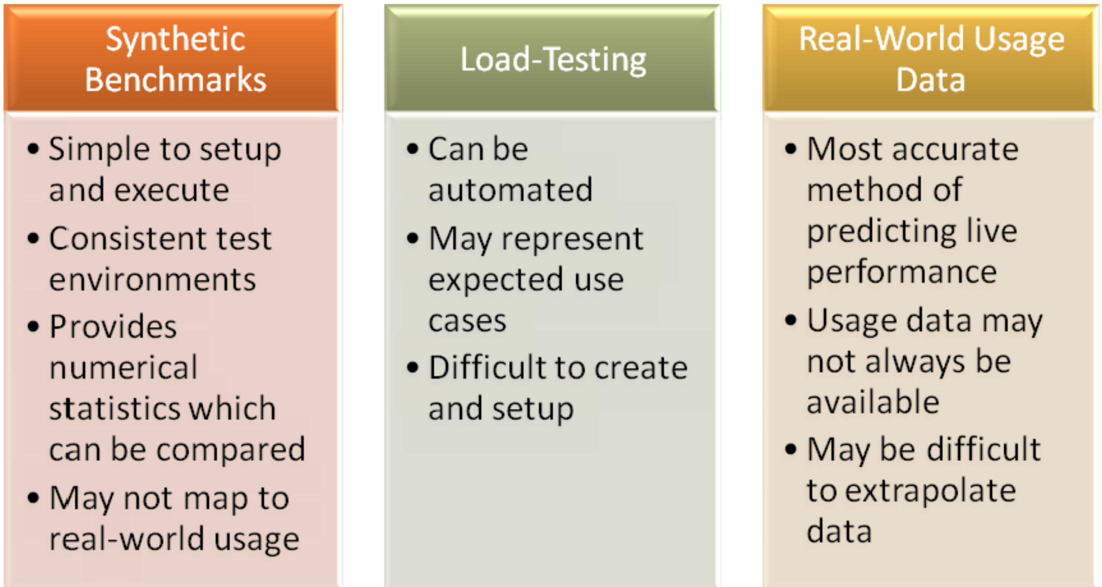

Figure 6.10: Various approaches to performance testing.

Synthetic Benchmarks

One of the simplest methods of performance testing is through the use of synthetic benchmarking applications. Numerous benchmarking products are available on the market. The tests themselves usually involve stressing one or more components of a system. With respect to virtualization, the tests can be run at the host level to provide details related to total system capabilities. For example, the maximum amount of disk throughput can be measured. If details related to a workload's expected disk utilization are known, they can be compared with the maximum capabilities of the hardware.

Synthetic benchmarks can also be executed within a guest OS. A common reason to do so is to get an idea of overall virtualization overhead. Table 6.3 provides a hypothetical comparison of performance statistics for specific tests that are run directly on the physical hardware and those that are run within a virtual machine. Although the numbers will vary significantly based on the type of test that is run, it's sometimes useful to determine expected overhead under certain extreme conditions.

Category	Test	Physical Machine	Virtual Machine
CPU	CPU (Average utilization for a sample workload)	34%	45%
	Floating-Point Unit Benchmark	5,400 MFLOPs	4,900 MFLOPs
Disk	Throughput (Maximum)	12MBps	7MBps
	Throughput (Average)	2.0MBps	1.5MBps
Memory	Bandwidth (Peak)	8	12
	Maximum Committed Memory	88MB	88MB
Network	Transfer Rate (Average)	4.0Mbps	2.8Mbps

Table 6.3: Comparing synthetic benchmark results for physical and virtual workloads on the same computer

The ability to quickly run tests and compare results across different systems is the primary benefit of using synthetic benchmarks. This can help IT departments in the procurement process by comparing overall performance of several brands and types of server hardware. In addition, it can help compare the relative performance of different virtualization products. The primary drawback of synthetic benchmarks is that it is often difficult (and inaccurate) to extrapolate these statistics to real-world usage patterns.

Load Testing

Once a workload is in production, systems administrators can get a good idea of resource utilization by monitoring the system "in action." To minimize risks, a similar approach can be taken prior to a production deployment. The purpose of load testing is to simulate user activity for an application or service. Then, while the system is performing typical activities and transactions, administrators can measure the usage of resources. This approach can be used for workloads that are running on physical servers as well as those that are running within a virtual machine.

The process of performing load testing can be significantly more difficult than performing a synthetic benchmark. In some cases, in-house application developers or third parties may have created usage benchmarks that approximate expected patterns of behavior. In such cases, load is generated against the application and details such as throughout and transaction latency can be recorded. In the case of Web-based applications, there are several testing suites that can be used to simulate Web-based activity. They can perform actions such as placing an item in an online shopping cart and completing the checkout process. Ideally, the tests can be rerun numerous times and by simulating different numbers of users on each test pass.

In some cases, application-specific load testing can provide a good indication of actual production performance. In other cases, however, it may be difficult to determine how users will interact with a new program. Overall, however, this approach can provide valuable insight into how a system will behave under stress.

Real-World Usage Information

The most valid type of performance test is one that is based on actual production usage data. This approach is most commonly used when IT departments are considering moving a current workload that is running directly on physical hardware into a virtual machine environment. Typical questions involve the performance overhead of virtualization and ensuring reliability and response times for the application or service. Figure 6.11 provides an overview of the typical steps that are required to perform real-world testing.

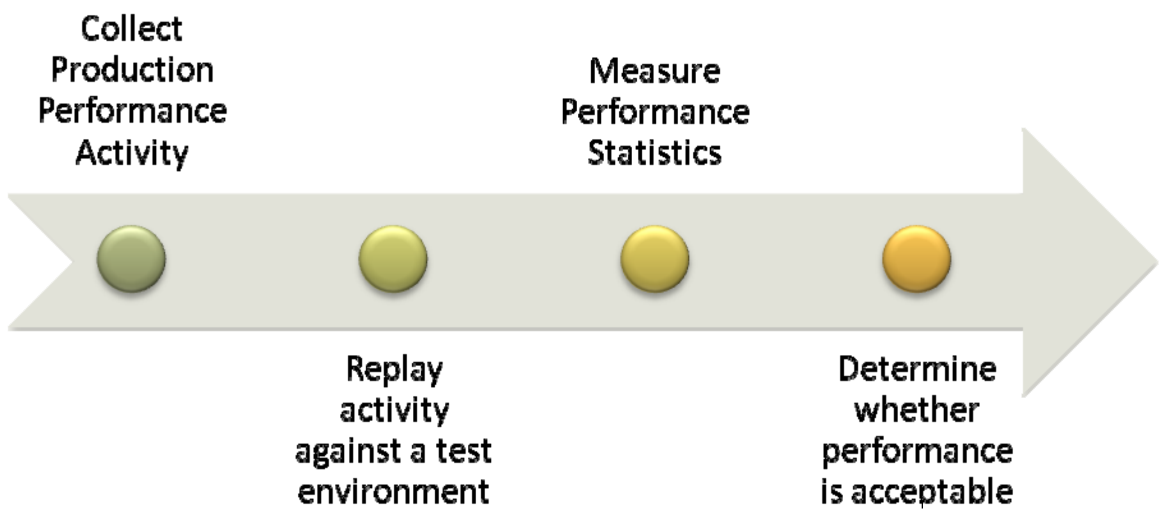

Figure 6.11: Performing real-world performance tests.

The first step generally involves collecting production system activity. This can sometimes be a challenge, but various tools allow for monitoring or profiling applications while they are running. For example, a database server platform could use a monitoring tool to collect all the queries that are being executed against a production server. The second step involves replaying the workload against a test environment. For example, if the IT department is hoping to perform a physical-to-virtual (P2V) migration of the existing application, a new virtual environment can be set up.

Although the activity is being replayed in the test environment, administrators can collect standard performance statistics. For example, memory, network, CPU, and disk systems can be examined for potential bottlenecks. Ideally, this step will also include monitoring of the "end-user experience." Finally, based on this data, technical staff can determine whether the application can be moved to a virtual machine.

Overall, there are several methods of implementing performance testing for current or prospective workloads. Each has its own benefits and limitations but all of them lead to making better-educated decisions related to virtualization. In addition, the testing process often provides additional insight into the behavior of systems. For example, there may be sub-optimal configuration settings that are limiting scalability. Or, some types of issues might only occur during periods of high load.

Automating Performance Monitoring

In relatively small IT environments, it might be possible for systems administrators to manually track performance details for several host servers. As long as the configurations are fairly consistent and systems administrators have the required expertise, it is manageable. However, in most environments, many considerations significantly limit the scalability of this approach. Figure 6.12 provides a list of important considerations to keep in mind.

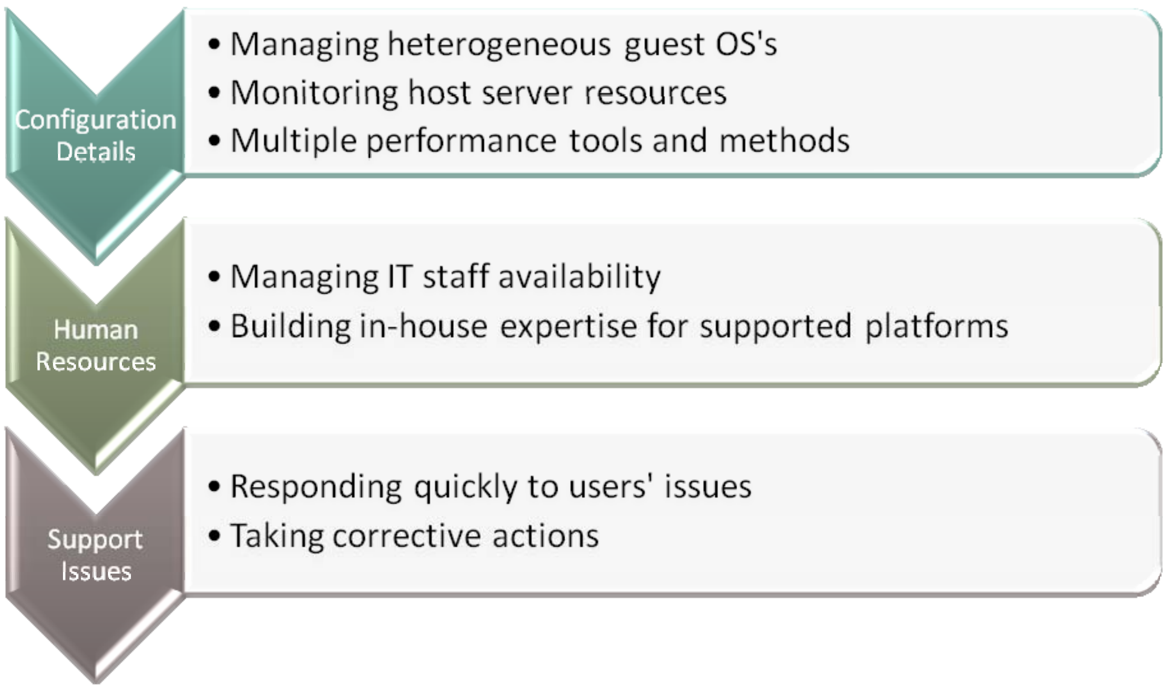

Figure 6.12: Common constraints related to performance management.

Clearly, these limitations can greatly affect the quality of overall systems management, especially in environments that supports hundreds of thousands of virtual machines. Monitoring performance is one area in which organizations can benefit significantly. They can reduce costs, improve scalability, make better use of hardware investments, and respond to problems much more quickly. Automated performance management tools generally include several valuable features (see Figure 6.13).

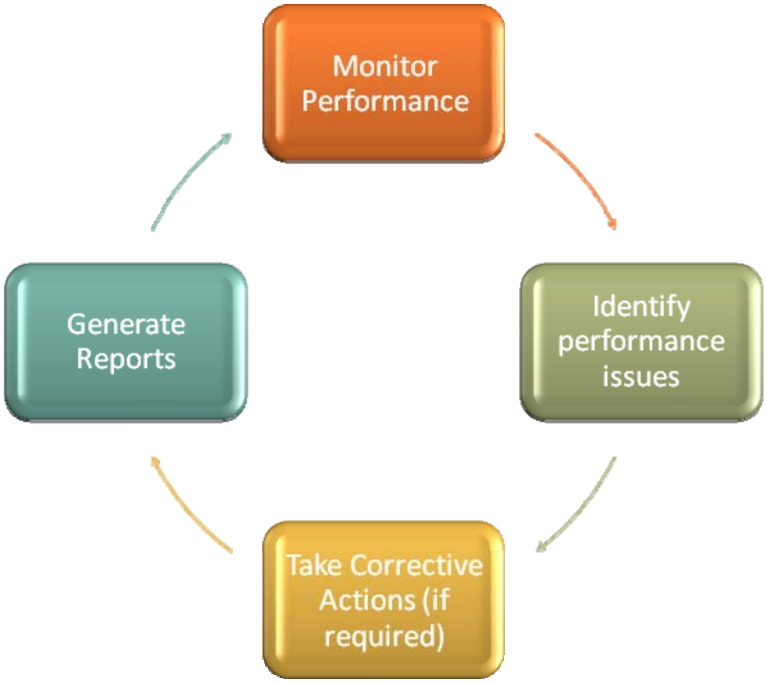

Figure 6.13: Benefits of using automatic performance management tools.

From a monitoring standpoint, automated products often have the ability to collect performance data from a wide variety of different types of systems. Organizations will typically support several virtualization platforms that are running on different host server types, so this is an important aspect. Additionally, variations in supported guest OS types can be large, and relying on individuals' expertise can severely limit the quality of monitoring.

> Chapter 7 will present more details related to what to look for in automated performance management solutions that support virtualization.

Summary

The focus of this chapter is the methods and details related to monitoring performance in virtualized environments. The discussion started by enumerating the business benefits of performance monitoring. Examples include capacity planning, increasing hardware utilization, and ensuring a positive end-user experience. The chapter then presented details related to how performance monitoring can be implemented in a mixed physical and virtual environment. Details include determining what to monitor and how to use guest- and host-level statistics.

The next section covered SLAs—an IT management approach that helps ensure that IT departments are meeting the needs of their internal and external customers. Through the use of SLAs, organizations can help ensure that their business and technical directions are aligned. I included several virtualization-related examples and ways in which SLAs can improve communications and lower overall costs.

Performance monitoring can actually be carried out proactively through the use of performance testing. I presented several ways in which IT staff can help predict how a workload will perform prior to putting it into production. Finally, I concluded the chapter with ways in which the process of performance monitoring can be automated. The net result is a decrease in costs and increases in uptime, reliability, and system utilization. Chapter 7 will use this information to identify ways in which system resources can be better utilized to provide the best possible end-user experience.