Virtual Infrastructures

The true goal of a virtualized infrastructure is supporting the needs of the business. A key metric for this goal is end-user satisfaction. When virtualized systems are not running well, users will report these issues, particularly when outages or performance problems affect their ability to conduct business. Ideally, IT should be alerted regarding issues before users report them, but this is not always the case.

The original goal of virtualization revolved around lowering costs and server consolidation projects garnered the most attention as the means to achieve cost-reduction. With such a simplistic goal, performance and evaluation metrics were easy to determine, as there could be a direct and immediate analysis of return on investment calculations performed against that single goal.

Today's virtual infrastructures are much more complex than those early virtual infrastructures and have very different goals. Although these goals still include a cost component, an appropriately designed virtual infrastructure can and should have additional far-ranging benefits to an organization, including improved availability, more agile service provisioning and more efficient expansion of the supporting underlying infrastructure.

Furthermore, as virtualization initiatives have gathered momentum and virtual environments are becoming the norm rather than the exception, the infrastructure necessary to support these environments has exploded, creating new opportunities for inefficiency. Inefficiencies can be introduced in many forms, such as improper configuration, or can be as simple as under- or over-building the overall virtual architecture.

Only when a virtual infrastructure is optimized can an organization reap all of the potential

benefits—financial and operational—from the investment.

To accurately evaluate the operating environment, chief information officers (CIOs) need to understand the points at which inefficiency can be introduced and select the appropriate management and monitoring tools and reporting processes to battle these issues. Without a broad understanding of how to monitor a virtual environment and identify potential issues, CIOs cannot appropriately target monitoring dollars and could waste money while attacking problems that might not exist. Further, these wasted dollars can lead to wasted operational efficiency affecting the whole enterprise, not just the IT group.

It is prudent to have a high-level view of your virtual environment. This view helps you quantify the business benefit of the virtualization initiative and, once it is fully deployed, the ongoing operational status and cost savings of the virtual environment.

By now most organizations have embarked on some kind of virtualization initiative. As formerly physical-server-based applications are moved into the virtual environment, careful attention needs to be paid to application performance to ensure service levels remain at acceptable levels and the virtual architecture can handle the workload demand.

Although this is not strictly a virtualization monitoring and optimization issue, this assessment step helps organizations avoid lost time and service disruption that can result from poor pre-implementation planning.

Prior to virtualizing a new service, use monitoring tools for the existing workload to develop a baseline snapshot for determining the current virtual infrastructure has the resources necessary to support the additional workload. You can perform this step with little to no cost because operating systems have built-in tools to perform this kind of assessment, including:

• Windows: Performance Monitor—Provides an at-a-glance look at overall resource utilization.
• Windows: Resource Monitor—Provides in-depth statistics for every aspect of system performance including processor usage, memory utilization, storage I/O and much more.
• Linux: vmstat—Like Windows Performance Monitor, it reports information about running processes, memory usage, storage I/O and processor activity.
• Linux: iostat—Reports processor statistics and I/O statistics for storage.
• Microsoft Assessment and Planning Toolkit (MAP)—This free tool from Microsoft provides actionable intelligence that aids in planning migrations to Hyper-V. MAP enables network-wide data gathering and comprehensive data analysis, and its reporting tools can be used to better

understand the existing configuration and performance baselines of a physical environment so as to enable seamless migration to a virtual one.

In addition to these free tools, there are a number of third-party products and services available to help you in this phase of virtualization planning. Keep in mind, too, that even if you are well into your vrtualization initiative, don't forget pre-virtualization baseline practices as you continue to bring existing physical services into the fold.

CIOs should pay particular attention to virtualization planning. Failure to properly prepare for virtualization can result in your business suffering one or both of these two issues: 1) services run poorly and 2) IT's attention is drawn away from the business as additional effort and resources are spent correcting for planning deficiencies.

When there aren't enough resources to go around, services suffer. The same is true for the workloads placed on a virtual environment-if you fail to keep up with increasing workload demands, your users will notice and your business productivity will be negatively affected.

Once you're in production with your virtualization initiative, appropriate monitoring of available resources is one of the most important administrative functions you can make, but it's often overlooked. There are a number of tools on the market that can help you make sure available resources are sufficient to meet ongoing business demand.

At what point do resource constraints become an issue? The answer has dozens of dependencies such as organizational tolerance for risk related to failed hosts and the expected growth rate for virtual resource consumption. In the interest of simplicity, no baseline values are offered here for the following conservative recommendations.

Memory utilization: As individual hosts and overall clusters reach and begin to exceed 80% RAM utilization, consider adding more memory resources. This is particularly important in High Availability clusters. Failure to address a serious resource constraint could jeopardize the ability of the cluster to recover from a host failure. In other words, make sure your cluster has enough spare RAM to handle the loss of its member host with the most RAM. That way, you can be sure you can continue to adequately address business needs while maintaining high levels of availability.

Processor utilization: Occasional spikes are nothing to worry about as you monitor your virtual environment. Workloads can and will spike, and there needs to be enough overhead available

to accommodate these spikes and service workload needs. If you begin to see long periods of sustained high processor utilization across your hosts or clusters— say, 75% to 80% or more— you might want to consider additional processing resources.

Processor queue length: Regardless of processor utilization, as workloads share processing resources, there needs to be a scheduling commitment that ensures all requests are handled in a way that makes sense. Queue length is a metric that measures how many processor requests are outstanding at any time. Conventional wisdom dictates that this number should remain low because the higher the number, the longer it takes your environment to service requests. If this number remains higher than 5 or 6 on a regular basis, consider adding more physical processors.

Storage latency: Input/Output operations per second (IOPS) is commonly used as a storage performance metric. However, when it comes to monitoring overall ongoing storage performance, there are better metrics to use, such as read/write latency and storage queue length. Read and write latencies are separate counters that measure how long it takes for a storage system to respond to a particular request. When measured in milliseconds (ms), a single-digit latency value is considered good. If storage latency begins to exceed 20ms on a regular basis, additional inspections should be performed to determine if there is a need to add storage spindles or transmission pathways to the storage system.

Storage queue length: Like processors, storage can be monitored in a couple different ways. Storage also has a queue length metric that identifies the number of outstanding storage requests at any time. Storage queue length is a little less forgiving than processor queue length, however. If you are consistently seeing that the number of outstanding storage requests exceeds 3 or 4 for a single disk, consider adding disk resources.

Network utilization: Just about every server's front-end communication medium is Ethernet. A stalwart of the technology world, Ethernet just keeps getting better. But it, too, can get overwhelmed, especially as you converge traffic from many formerly separate services. Keep an eye on network performance in your environment, and bear in mind that Ethernet carries with it data encapsulation overhead, for both its own communication protocol and for application data transmitted via TCP/IP.

Dropped network packets: Even more important than utilization, an environment that is dropping a lot of packets indicates a problem somewhere along the line that needs to be corrected. This, too, can point to or create a performance issue.

Although performance problems can create difficulties for business users, nothing compares to the problems a business faces when there is a complete failure of all or part of the IT infrastructure. Many virtualization initiatives began life as cost-cutting measures and savvy organizations quickly discover that the new infrastructures, when architected properly, bring significant new availability opportunities. When a servicing host fails in a virtual environment, workloads on that host can be automatically and immediately restarted on an alternate host, resulting in little downtime.

For even higher levels of availability, hypervisors have built-in methods to run identical side-by-side workloads that can fail over to one another with no noticeable impact. Proper monitoring is the first step in ensuring that highly available workloads remain highly available. By itself, monitoring will not lead to higher levels of availability. Monitoring must be followed up with appropriate response from the IT department to resolve potential issues. My making sure that all systems are monitored, IT departments create conditions for success.

You might ask yourself, "Why shouldn't I just throw resources at the virtual environment to see how low I can get utilization patterns?"

It's all about balance and efficiency. As stated, many organizations moved to virtual infrastructures in order to make more efficient use of their hardware and get out of the "5% to 10%utilization" cycle. By simply throwing random resources at the environment, you're not targeting areas of need and are potentially wasting time and money by addressing areas that do not need attention. If your RAM utilization is 4%, you have plenty of room to grow, but if you have no intention of growing that is a lot of expensive RAM going to waste. Bear in mind that more is not always better and the optimal strategy is to instead right-size your virtual environment.

Change is often the biggest culprit when it comes to failures in the IT environment, and this presents a significant issue because change is a constant in technology operations. From adding resources to a particular workload to applying patches to operating systems to administering a formal technology equipment lifecycle, the IT environment undergoes change each and every day.

To control the change process, many organizations have implemented management frameworks, such as the Information Technology Infrastructure Library (ITIL). These frameworks mandate maintaining a configuration baseline, which is a recommended practice for any organization. A current configuration baseline is a record of the most recent operational state of a virtual environment. As changes are made to the environment, "drift" begins to appear and can be seen

by the delta between the configuration baseline and the current operating environment.

Beyond simple troubleshooting, compliance requirements typically dictate that a history be kept of changes to the environment. As such, collecting and storing change records moves beyond being a troubleshooting aid to becoming an integral part of an organization's compliance efforts.

It's possible to manually document change activity and reorient the environmental baseline from time to time, but it's much easier and more reliable to automate this data collection and analysis process. This is especially true as virtual environments grow in size and complexity. Tools can track change management information and provide CIOs with a business-centric view of the organization. For example, infrastructure views can be based on business unit, department, purpose, service level agreement or whatever grouping makes sense.