Software

Best Practices for Oversubscription of CPU, Memory and Storage in vSphere Virtual Environments

How far can oversubscription be taken safely?

Written by Scott D. Lowe

Abstract

One of the benefits of virtualization is that it enables administrators to efficiently share host resources among different applications. In fact, administrators often oversubscribe the physical resources on a host in order to maximize the number of workloads that can run on a host. But how much oversubscription is too much? This paper discusses what oversubscription is and why it is used, explores the pros and cons of the practice, and proposes some ideas about the point at which oversubscription becomes dangerous

Introduction

Virtualization enables data centers to focus on the needs of business applications, but it can make resource allocation more challenging.

One of the great features of virtualization is the ability to run many disparate workloads on a single host server, thereby maximizing the utilization of that host server. In doing so, organizations have been able to reinvent the modern data center. Whereas data centers of ten years ago tended to be server-centric places, modern data centers revolve around the needs of line-of-business applications, including ensuring that those applications remain highly available and able to survive the loss of host servers.

Virtualization has changed the data center dynamic in many other ways as well. While workloads used to be confined to the hardware on which they were originally installed, in a modern data center, workloads are fluid; they flow from host to host based on sets of administrator-defined rules as well as in reaction to changes in the host environment. The fluidic nature of the modern data center has added new challenges to resource allocation, but over the years, both free and paid tools have been introduced to assist administrators in their resource planning efforts.

Administrators can oversubscribe the physical resources on a host in order to maximize the number of workloads that can run on it. In other words, they can assign to virtual machines, in aggregate, more resources than are actually available on the host.

Sizing resources in a virtualized environment

Virtualization makes hardware sizing more complex—but offers opportunities for improving data center efficiency. The rise of virtualization has also enabled the use of hardware in ways that were never envisioned even just ten years ago. In those days, administrators purchased servers sized to support the peak needs of a single application, and that sizing included a projection of the resources the application would likely need over the life of the server hardware. Because many servers were deployed with just a single application, resource planning was relatively simple. In modern data center environments, which are heavily virtualized, resource planning takes on new complexity: because a wide array of I/O patterns will be present on single pieces of hardware, administrators need insight into how individual applications interact with the rest of the environment.

This blending of I/O in a heavily virtualized environment has also created a significant opportunity for efficiency in the data center. Whereas administrators used to size individual servers based on the needs of a single application, the mixed nature of I/O in a virtual environment enables sharing of resources with different peak needs. As a result, there is opportunity for administrators to very efficiently share host resources among different applications.

Overprovisioning is useful—but how much is too much?
Even with virtualization, administrators still sometimes overprovision resources and size individual virtual machines (VMs) to meet peak demands, so there are often resources that go unused in a virtual machine. vSphere provides a number of powerful methods for sharing idle resources with other running workloads. In addition, administrators can oversubscribe the physical resources on a host in order to maximize the number of workloads that can run on a host. In other words, they can assign to virtual machines, in aggregate, more resources than are actually available on the host. For example, suppose a host has 96 GB of physical RAM. Under the right circumstances, an administrator might assign 128 GB of RAM to all of the virtual machines running on that host.

But just how far can this oversubscription be taken? The limits depend on a number of factors. This paper discusses oversubscription in general, explores its pros and cons, and proposes some ideas about the point at which oversubscription becomes dangerous.

Resource management and oversubscription

What is oversubscription?
Oversubscription in vSphere refers to various methods by which more resources than are available on the physical host can be assigned to the virtual servers that are supported by that host. In general, administrators have the ability to oversubscribe processing, memory and storage resources in virtual machines.

Refusing to oversubscribe resources is the safest choice, but often wastes resources. Different administrators have different opinions on the wisdom of oversubscribing physical resources. Many administrators prefer to assign only those resources that are physically available to support all of the running workloads. This is the safest option as it ensures that, in general, all running virtual machines will always have the resources they need.

However, in days of physical servers, it was not uncommon to find that physical servers rarely made use of all of their resources. From a processor standpoint, utilization averaged only 5–15 percent, meaning that there was a whole lot of room for growth.

Oversubscription maximizes the value of resources—but introduces risk of the host not having enough resources to service all its VMs.
While virtual machines are generally more right-sized than their physical counterparts were in the past, there is still room to grow built in, especially when particular workloads are idle. Many administrators see this as an opportunity to make use of those idle resources in order to maximize virtual machine density on a host. However, with oversubscription, administrators are basically assigning to virtual machines more resources than are actually available on the host. In other words, if all of the virtual machines suddenly requested access to all of their allocated resources, the host would not have enough resources to service the needs.

Resource oversubscription, while it does increase virtual machine density, carries with it some risks. Once a particular resource is finally exhausted, if that resource happens to be oversubscribed, stability issues can occur and major performance problems can be introduced affecting all of the workloads running on the host server.

Before we discuss more about oversubscription, it's important to understand vSphere manages the three basic types of resources:

- Processing resources
- Memory resources
- Storage resources

How vSphere manages processing resources

How physical resources are represented on a vSphere host
In vSphere, administrators assign CPUs to virtual machines in order to support the workload needs of each individual virtual machine. These virtual processing resources are pulled from the host's available physical CPUs. The number of physical CPUs that are present in hosts is dependent on a couple factors.

In vSphere, a physical CPU (pCPU) refers to:

- **When hyperthreading is not present or enabled**: A single physical CPU core
- **When hyperthreading is present and enabled**: A single logical CPU core

Here are two examples:

- If a host has two eight core processors and hyperthreading is either not supported or not enabled, that host has sixteen physical CPUs (8 cores x 2 processors).
- If a host has two eight core processors and hyperthreading is enabled, that host has thirty-two physical CPUs (8 cores x 2 processors x 2 threads per core).

How those resources are presented to virtual machines
In a virtual machine, processors are referred to as virtual CPUs (vCPUs). When an administrator adds vCPUs to a virtual machine, each of those vCPUs is assigned to a pCPU, although the actual pCPU may not always be the same. There must be enough pCPUs available to support the number of vCPUs assigned to an individual virtual machine or that virtual machine will not boot.

However, that doesn't mean that administrators are limited to just the number of pCPUs in the host. On the contrary, there is no 1:1 ratio between the number of vCPUs that can be assigned to virtual machines and the number of physical CPUs in the host. In fact, as of vSphere 5.0, there is a maximum of 25 vCPUs per physical core, and administrators can allocate up to 2,048 vCPUs to virtual machines on a single host.

How vSphere manages memory resources

vSphere uses a number of techniques to maximize the use of RAM in a virtual environment:

- **Transparent page sharing (TPS)**—In most virtual environments, administrators run many copies of the same operating system. In these cases, there is a lot of duplication of memory pages in host memory. Transparent page sharing is basically a

Oversubscription is an opportunity to maximize VM density on a host. But it introduces risks: if all of the VMs suddenly request access to all of their allocated resources, the host will not have enough resources to service the needs.

The number of vCPUs that can be assigned to VMs is not limited to the number of physical CPUs in the host.

form of memory deduplication—vSphere combines multiple identical memory pages into just one and frees the remaining pages up for other uses. TPS has an almost imperceptible impact on the performance of the host.

- **Memory ballooning**—When the VMware Tools are installed inside a guest virtual machine, a memory balloon driver is installed as well. This driver acts as a Windows process, and the OS can use its normal memory management techniques to assign idle or unused memory pages to the driver. The balloon driver then "pins" those pages and reports this back to the hypervisor. If the host becomes low on physical memory, guest memory pages are assigned to the balloon driver, and the host can then reclaim these memory pages in order to address the needs of other virtual machines that may need the RAM.

  In this way, when a particular virtual machine has RAM to spare, it can transparently share that RAM with other virtual machines on the same host, enabling the host to achieve yet higher levels of VM density. Whereas TPS is a memory deduplication technique, the ballooning process brings to RAM a sort of thin provisioning capability. The ballooning process does require some processing overhead, which is usually imperceptible in the performance of the guest and host. However, in extreme cases, ballooning can cause swapping inside the OS.
- **Memory compression**—Introduced in vSphere 4.1, memory compression can, in some cases, replace the costly swapping process. With this technique, rather than memory pages being swapped to disk on a per-VM basis, the memory pages are compressed and placed into a compression cache on disk. When the need arises to swap to return to RAM a page that would have been swapped, the page is instead retrieved from the cache and uncompressed. While this process is less costly than swapping to disk, it does still carry something of a performance hit.
- **Swapping to disk**—Swapping to disk is the hypervisor's last-ditch effort to retrieve enough physical RAM to satisfy the needs of workloads running on a host. Swapping is a process by which the hypervisor moves the least used memory pages to disk. Those memory pages are still accessible, but when they are required, they must be retrieved from disk. Swapping will noticeably degrade the overall performance of the host.

  Because vSphere's other memory management techniques are so good, swapping usually takes place only on seriously overcommitted hosts, although swapping can also be caused by resource pool constraints or memory limits configured on a virtual machine. In addition, if a VM does not have VMware Tools installed or VMware Tools is not running, the ballooning process would get skipped completely and the system will go straight to swapping.

It should be noted that neither swapping nor compression take place unless there is a memory contention issue on the host, or in the situations discussed with regard to swapping. In most environments, memory contention issues that result in swapping or compression should be avoided since this situation means that the host has basically run out of RAM.

How vSphere manages storage resources

Storage is the third piece of the resource puzzle in a vSphere environment. Storage resources can also be oversubscribed through what have become very common resource allocation techniques.

Thin provisioning is the most common technique for overprovisioning storage resources.

The most common technique for overprovisioning storage is a process known as thin provisioning. In many cases, when an administrator allocates storage to a virtual machine, more storage than is absolutely necessary

is allocated. After all, it's reasonable to expect that the virtual machine will continue to need additional disk space as time goes on.

Thin provisioning operates as follows: When an administrator provisions the total disk space for the virtual machine, the virtual machine is told that it has access to the entirety of the allocated space. In reality, however, vSphere gives the virtual machine only the space that it is actually consuming. For instance, if an administrator allocates 200 GB to a new virtual machine, but that virtual machine is using only 40 GB, the remaining 160 GB remain available for allocation to other virtual machines. As a virtual machine requires more space, vSphere provides additional chunks to that virtual machine, up to the size of the disk that was originally allocated.

By using thin provisioning, administrators can create virtual machines with virtual disks of a size that is necessary in the long term, without having to immediately commit the total disk space that is necessary to support that allocation. In many tests, it has been shown that thin provisioning carries with it only a very slight—almost negligible—performance impact. Accordingly, thin provisioning has become a common, acceptable and often recommended method for managing storage capacity.

Some storage devices have additional features, such as data compression and deduplication, that enable additional levels of oversubscription. For the purposes of this paper, however, the focus is on the hypervisor, so only thin provisioning will be discussed.

Getting insight into resource usage in your environment

Now that we have see how resources are managed in a vSphere environment, let's move on to oversubscribing those resources. In order to determine whether resources are overcommitted, the administrator needs a monitoring tool such as the free vOPS™ Server Explorer tool from Dell®. One of its several utilities is Environment Explorer, which provides administrators with a high-level view of resource usage in the environment. As shown in Figure 1, Environment Explorer shows resource utilization as a percentage of actual physical resources, making the utility a perfect fit for exploring the topic of resource over-commitment.

Now let's explore some guidelines for oversubscribing each of the three types of resources: processing, memory and storage.

> Swapping to disk is the hypervisor's last-ditch effort to retrieve enough physical RAM to satisfy the needs of workloads running on a host. Swapping will noticeably degrade the overall performance of the host.

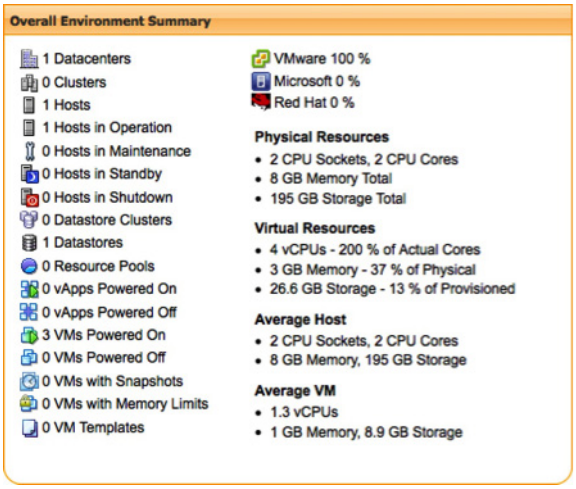

Figure 1. Environment Explorer (part of the free vOPS Server Explorer tool) provides a high-level view of resource usage.

By using thin provisioning, administrators can create VMs with virtual disks of a size that is necessary in the long term, without having to immediately commit the total disk space that is necessary to support that allocation.

Oversubscribing processing resources

The common wisdom

As mentioned earlier, in vSphere 5, every physical processor core can support up to 25 vCPUs. However, for every additional workload beyond a 1:1 vCPU to pCPU ratio, the vSphere hypervisor needs to invoke processor scheduling in order to distribute processor time to virtual machines that need it. For example, if an administrator has created a vCPU to pCPU ratio of 5:1, then each processor is supporting five vCPUs.

Experts in the field disagree about the proper rules of thumb when it comes to vCPU to pCPU ratio. However, there are two items on which just about everyone agrees:

- **Start with one vCPU per virtual machine**. Most experts agree that administrators should create new virtual machines with just one vCPU and add virtual vCPUs as needs dictate. As vCPUs are added, the virtual machine is tied to requiring processor time from the host. Whenever the virtual machine needs to perform an operation, it has to wait for a number of physical CPUs equal to the number of assigned vCPUs to be available. So, as administrators add more vCPUs to a virtual machine, there is an increased risk of poorer overall performance.
- **The vCPU to pCPU ration is workload dependent**. While 1:1 vCPU to pCPU assignment is sometimes advocated, other ratios are common. Although vSphere 5 supports a ratio of up to 25:1, your ability to achieve a high ratio will depend on the kinds of workloads you need to support. If the host is supporting lots of virtual machines, each with only meager processing needs, the vCPU to pCPU ratio could be quite high. If, however, the host is running a number of processor intensive workloads, the ratio may be much smaller.

Metrics to watch

The following metrics will help you maintain a vCPU to pCPU ratio that makes efficient use of resources while still allowing workloads to run well:

- Inside virtual machines
  - CPU Utilization—When average CPU usage remains high, it's time to add an additional vCPU to the VM.
- On the host:
  - CPU Ready—CPU Ready is, by far, the most important gauge of overall host health standpoint with regard to CPU. CPU Ready indicates the length of time that a VM is waiting for enough physical processors to become available in order to meet its demands. For example, if a VM is allocated four vCPUs, this metric will show the length of time that the VM waited for four corresponding pCPUs to become available at the same time.
  - CPU Utilization—The overall CPU usage on the host server is also critical because it enables an administrator to understand just how much work the host server is doing.

Real-world observations and advice

Virtual resources
4 vCPUs - 200 % of Actual cores
3 GB Memory - 37 % of Physical
26.6 GB Storage - 13 % of Provisioned

Figure 2. A lab with a vCPU to pCPU ratio of 2;1

Online forums are filled with questions from users requesting insight into acceptable vCPU to pCPU ratios in a real-world environment. While some responses continue to advocate for a 1:1 ratio, from a pure density standpoint, 1:1 should be considered a worst-case scenario. Figure 2 shows a lab with a ratio of 2:1.

Some respondents indicate that they have received guidance that suggests no more than a 1.5:1 vCPU to pCPU ratio, but guidance from industry experts suggests that vSphere real-world numbers are in the 10:1 to 15:1 range. Still others indicate that VMware itself has a recommended ratio range of 6:1 to 8:1.

The Dell white paper, "Demystifying CPU Ready (% RDY) as a Performance Metric," establishes the following vCPU:pCPU guidelines:

- 1:1 to 3:1 is no problem.
- 3:1 to 5:1 may begin to cause performance degradation.
- 6:1 or greater is often going to cause a problem.

In addition, keeping the CPU Ready metric at 5 percent or below is considered a best practice.

The actual achievable ratio in a specific environment will depend on a number of factors:

- **vSphere version**—The vSphere CPU scheduler is always being improved. The newer the version of vSphere, the more consolidation that should be possible.
- **Processor age**—Newer processors are much more robust than older ones, so organizations with newer processors should be able to achieve higher processor ratios.
- **Workload type**—Different kinds of workloads on the host will result in different optimal ratios.

vScope Explorer, another utility included in vOPS Server Explorer, can help you investigate performance metrics, including CPU Ready, at both the host and virtual machine levels, so you can determine whether your vCPU to pCPU ratio is too high.

In addition, Environment Explorer identifies where host processor resources are overcommitted, so you'll know where perform additional analysis to determine if that over-commitment is causing performance issues. As the "% of actual cores" metric begins to surpass 500 percent, carefully monitor CPU Ready and general workload performance to ensure that business needs are being met.

Oversubscribing memory resources

The common wisdom

Oversubscribing RAM is one of the more controversial resource oversubscription options. Whereas CPU and storage resources are often overcommitted, there seems to be some conservatism when it comes to overcommitting RAM.

Metrics to watch

On a host server, administrators need to watch the amount of RAM actually in use by virtual machines. As the actual RAM in use approaches 100 percent, either add additional RAM to the server or migrate workloads to hosts that have more available RAM.

Real-world observations and advice

In order to maximize VM density and ensure that the environment remains operational, it's important to monitor the actual memory utilization. That's why Environment Explorer displays RAM usage (the "% of physical" metric) using the amount of RAM actually provisioned to each virtual machine, rather than the amount of RAM actually being used by virtual machines once all of vSphere's various memory management techniques are taken into consideration.

Environment Explorer (part of Dell's free vOPS Server Explorer tool) provides a high-level view of resource usage.

Most experts agree that administrators should create new virtual machines with just one vCPU and add virtual vCPUs as needed.

The level of over-commitment possible depends on one primary factor: how much memory deduplication can take place by virtue of the fact that there are many similar workloads running on the host. The greater the level of disparity between running workloads, the less memory consolidation that can take place and the less density that can be enjoyed. Here are some observations about what others are doing and recommending with regard to memory over-commitment:

- Many administrators refuse to oversubscribe RAM at all.
- Some administrators prefer to not exceed 125 percent of physical memory, feeling that going beyond that limit carries unacceptable risk.
- If every workload on the server is identical, much higher over-commitment levels are possible.
- Many other administrators simply spot-check host memory usage, but don't regularly scan for over-commitment levels.

Oversubscribing storage resources

The common wisdom

It has become commonplace to oversubscribe storage resources using thin provisioning. This technique offers many benefits; the primary one is maximizing the use of the organization's storage capacity. Plus, thin provisioning also helps IT in two ways. First, it enables administrators to give a virtual machine all of the storage it will ever need without having to constantly watch to see if it needs more space. Second, thin provisioning can reduce conflicts between IT and other teams: application owners can request all of the storage they like and storage administrators—knowing full well that the request is too high—can simply grant the request without worrying about wasting that over-requested storage.

However, thin provisioning also carries with it some challenges. While it can make life easier on a daily basis, it does add some complexity, and if administrators aren't careful, they can introduce major availability issues. If the storage oversubscription results in the storage volume running out of space, the VMs will still think they have available disk space to use but there won't be any space available. This can cause a serious outages that can result in data loss and costly recovery. So if you use thin provisioning, be sure to monitor carefully.

Metrics to watch

To mitigate the risks associated with thin provisioning, you need to keep a close eye on the amount of free space in a datastore. As a datastore gets low on space, proactively add space to the datastore or use Storage vMotion to move one of the virtual machines to a different datastore that has enough available capacity to serve the needs of the workloads.

Real-world observations and advice

Thin provisioning is well represented in Environment Explorer, although it's displayed only in aggregate. In the example in Figure 2, 26.8 GB of storage is currently in use—13 percent of what's actually provisioned to the three virtual machines that are powered on. As the "% of provisioned" metric approaches 100 percent, take care to ensure that additional physical resources are made available.

Conclusion

Overprovisioning of processing, memory and storage can help you maximize resource utilization in your virtual environment. But you want to overprovision in such as way that you can also maintain high performance. Using the techniques and real-world advice presented in this white paper, along with the right tools, you can balance these needs and overprovision wisely.

For More Information

About Dell

Dell Inc. (NASDAQ: DELL) listens to customers and delivers worldwide innovative technology, business solutions and services they trust and value. For more information, visit www.dell.com.

If you have any questions regarding your potential use of this material, contact:

Dell Software
5 Polaris Way
Aliso Viejo, CA 92656
www.dell.com
Refer to our Web site for regional and international office information.