VKernel
is now a part of Dell

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

Pros and cons of oversubscription and how far it should be taken before it becomes dangerous

By Scott D. Lowe

WHITE PAPER

Contents

Introduction

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 and ensuring that these applications remain highly available and able to survive the loss of host servers.

Virtualization has, in fact, completely 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 the market intended to assist administrators in their resource planning efforts.

However, the rise of virtualization has also enabled hardware usage 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 for how much in the way of resources the application would likely need over the life of the server hardware. Because many servers were deployed with just a single application, it made resource planning relatively simple. In modern data center environments, which are heavily virtualized, resource planning takes on new complexity due to the fact that a wide array of I/O patterns will be present on single pieces of hardware. Administrators must gain 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 single application needs, 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.

Moreover, it’s important to keep in mind that, even with virtualization, administrators still overprovision resources sometimes and size individual virtual machines to meet peak demands. As such, there are often resources that go unused in a virtual machine. vSphere provides a number of powerful methods through which to share idle resources with other running workloads. In fact, in addition to being able to share idle resources, in order to maximize the number of workloads that can run on a host, administrators can oversubscribe the physical resources that exist on a host. In other words, administrators have the ability to assign in aggregate to virtual machines 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? In reality, the limits depend on a number of factors. This paper discusses overprovisioning in general, the pros and cons of oversubscription and, based on a number of real world discussions, proposes some ideas about the point at which oversubscription becomes dangerous.

Resource Management and 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.

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 recalling the physical server days, 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% to 15%, meaning that there was a whole lot of room for growth.

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 discussing more about oversubscription, it's important to understand the resource management features that are built into vSphere.

Processing Resource Management

In vSphere, administrators are accustomed to assigning CPUs to virtual machines in order to support the workload needs of that 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 order to establish a baseline understanding of physical and virtual processor management, you must understand that in vSphere, a physical CPU (often abbreviated 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 some 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).

With an understanding of how physical resources are represented on a vSphere host, the discussion turns to how those processing 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.

Memory Resource Management

vSphere uses a number of techniques to maximize the use of RAM in a virtual environment. Here is a list of those techniques and a brief description of what each one does:

- **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 form of memory deduplication in which vSphere combines multiple identical memory pages into just one and frees the remaining pages up for other uses. From a performance impact perspective, TPS has an almost imperceptible impact on the host.
- **Memory ballooning** - When the VMware Tools are installed inside a guest virtual machine, a memory balloon driver is installed along with the other Tools components. This driver acts as a Windows process, allowing the OS to use its normal memory management techniques to assign idle/unused memory pages to it. The balloon driver then "pins" those pages and reports this back to the hypervisor. In the event that the host becomes low on physical memory, guest memory pages are assigned to this balloon driver. Once that assignment takes place, 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 virtual machine 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** - In vSphere 4.1, VMware introduced the concept of memory compression, which can, in some cases, replace the costly swapping process. With this technique, rather than swapping memory pages 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, it is instead retrieved from this cache and un-

compressed. 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. 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 due to memory limits configured on a virtual machine. Likewise, 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. Swapping is a process by which the hypervisor moves the least used memory pages to disk. Those memory pages are still accessible, but when required, must be retrieved from disk. When it comes to the impact on performance, swapping has an extremely high cost and will noticeably degrade the overall performance of the host.

It should be noted that neither swapping nor compression take place unless there is a memory contention issue on the host, or under 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.

Storage Resource Management

Storage is the third piece of the resource puzzle in a vSphere environment and also carries with it the ability to be oversubscribed through what have become very common resource allocation techniques.

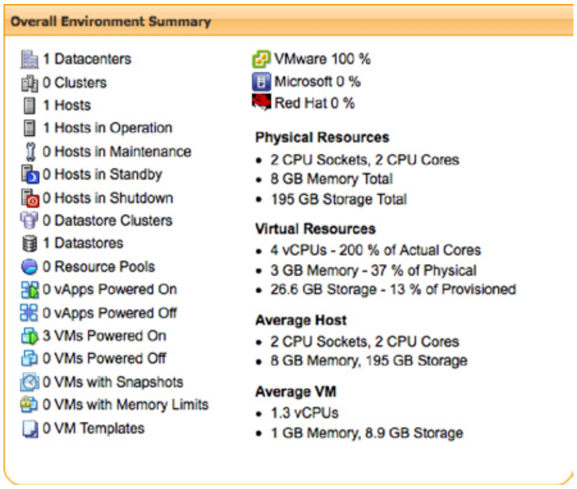
The most common technique by which storage can be overprovisioned is through a process known as thin provisioning. In many cases, when an administrator allocates storage to a virtual machine, more storage than is absolutely necessary is allocated. After all, it's reasonable to expect that the virtual machine will continue to need additional disk space as time goes on. Thin provisioning operates thusly: 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 only gives the virtual machine the space that it is actually consuming. So, if an administrator allocates 200 GB to a new virtual machine, but that virtual machine is only using 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 to the point of being negligible—performance impact. As such, thin provisioning has become a common, acceptable and often recommended method to best manage storage capacity.

Note also that some storage devices have additional features that may allow for additional levels of oversubscription. Such features include data compression and deduplication. For the purposes of this paper, however, the focus is on the hypervisor, so only thin provisioning will be discussed.

Resource Overcommitment

With an understanding for how resources are managed in a vSphere environment, the discussion moves to oversubscribing those resources. For the purposes of this paper, the assumption is that oversubscription is an acceptable practice. In order to determine whether or not resources are overcommitted, a monitoring tool needs to be used. In this paper, the free vOPS Server Explorer tool from Dell is in use. This tool has multiple utilities built-in, Environment Explorer shown to the left is one of the utilities in vOPS Server Explorer that provides administrators with a high level look at resource usage in the environment. As indicated in the figure, resource utilization as a percentage of actual physical resources is displayed, making Environment Explorer a perfect fit when it comes to exploring the issue of resource overcommitment.



Overall Environment Summary

- 1 Datacenters
- 0 Clusters
- 1 Hosts
- 1 Hosts in Operation
- 0 Hosts in Maintenance
- 0 Hosts in Standby
- 0 Hosts in Shutdown
- 0 Datastore Clusters
- 1 Datastores
- 0 Resource Pools
- 0 vApps Powered On
- 0 vApps Powered Off
- 3 VMs Powered On
- 0 VMs Powered Off
- 0 VMs with Snapshots
- 0 VMs with Memory Limits
- 0 VM Templates

VMware 100 %
Microsoft 0 %
Red Hat 0 %

Physical Resources

- 2 CPU Sockets, 2 CPU Cores
- 8 GB Memory Total
- 195 GB Storage Total

Virtual Resources

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

Average Host

- 2 CPU Sockets, 2 CPU Cores
- 8 GB Memory, 195 GB Storage

Average VM

- 1.3 vCPUs
- 1 GB Memory, 8.9 GB Storage

Oversubscribing Processing Resources

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. So, if an administrator has created a 5:1 vCPU to pCPU ratio, then each processor is supporting five vCPUs.

In reviewing general guidance for this paper, there is obvious disagreement as to what constitutes a rule of thumb when it comes to vCPU to pCPU ratio. There were two items on which just about everyone agrees:

- Start with 1 vCPU per virtual machine. Most experts agree that, when creating a new virtual machine, administrators should create that virtual machine with just one vCPU and, as needs dictate, add virtual vCPUs. 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, it's common to find a different ratio in place. That said, although vSphere 5 supports a ratio of up to 25:1, the ability to achieve a high ratio is very dependent on the kinds of workloads that are being supported. 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, that ratio may be much smaller.

Metrics to Watch

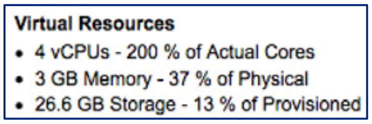
There are a number of metrics to watch with regard to CPU that will help administrators maintain a balance of vCPU to pCPU ratios that make more efficient use of resources while still allowing workloads to run well.

- Inside virtual machines
  - CPU Utilization- This metric will allow an administrator to make a determination about when it's time to add an additional vCPU to a virtual machine. It's time to add an additional vCPU to a virtual machine when average CPU usage remains running at high levels.
- On the host
  - CPU Ready - From an overall host health standpoint with regard to CPU, this metric is, by far, the most important gauge. CPU Ready is a metric used to determine the length of time that a virtual machine is waiting for enough physical processors to become available in order to meet the demands of the virtual machine. If a virtual machine is

allocated four vCPUs, this metric will indicate the length of time that the virtual machine 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 as it allows an administrator to understand just how much work the host server is doing.

Real World Observations



Virtual Resources

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

Various 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. In the diagram to the left, note that this particular lab has a current 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 real world recommended ratio range of 6:1 to 8:1.

In this Dell white paper, the following vCPU:pCPU guidelines are established:

- 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

Additional guidance suggests that keeping the CPU Ready metric at 5% 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 and, with newer processors, organizations should be able to achieve higher processor ratios.
- **Workload Type** - Different kinds of workloads on the host will result in different possible ratios.

vScope Explorer is another utility included in vOPS Server Explorer which can help look at performance problems, including CPU Ready, at both the host and virtual machine levels to help identify if vCPU to pCPU ratio is too high.

Furthermore, Environment Explorer identifies where host processor resources are overcommitted, pointing administrators to a place to perform additional analysis to determine if that over commitment is cause for any current performance issues. As the "% of actual cores" metric begins to surpass 500%, administrators should more carefully monitor CPU Ready and general workload performance to ensure that business needs are being met.

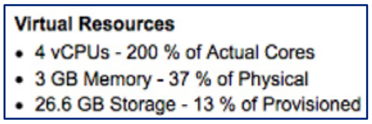
Oversubscribing Memory Resources

Oversubscribing RAM is, perhaps, one of the more controversial resource oversubscription options out there. 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%, either additional RAM needs to be added to the server or workloads need to be migrated to hosts that have more available RAM.

Real World Observations



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

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

The level of over commitment possible is dependent 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.

In reviewing 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% of physical, feeling that going beyond this metric 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 overcommitment levels.

Oversubscribing Storage Resources

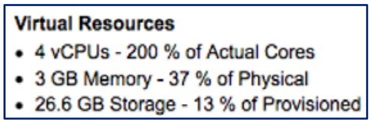
It has become commonplace to oversubscribe storage resources through the use of thin provisioning. Of course, oversubscription carries with it a number of pros and cons. The primary item in the pros column is the administrator’s ability to maximize the use of the organization’s storage capacity. Further, thin provisioning provides an administrator with a way 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. Additionally, thin provisioning can reduce conflict on the IT team. 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 and, while it can make life easier on a daily basis, it does add some complexity. First and foremost, if administrators aren’t careful, they can introduce major availability issues. If the storage oversubscription results in the storage volume is running out of space, the VMs still think they have available disk space to use but there isn’t any space available. This can cause a serious outages with data loss and costly recovery as an outcome if you don´t monitor carefully.

Metrics to Watch

To mitigate this problem, administrators using thin provisioning need to keep a close eye on the amount of free space in a datastore. As a datastore gets low on space, the administrator needs to 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



Virtual Resources

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

In Environment Explorer, thin provisioning is well represented, although it’s displayed only in aggregate. In the example in this paper, of the 195 GB of storage available to be used by virtual machines, only 26.8 GB is currently in use. In reality, much more than that is provisioned to the three virtual machines that are powered on. As administrators watch Environment Explorer and the “% of provisioned” metric approaches 100%, care should be taken to ensure that additional physical resources are made available.

Conclusion

Virtualization allows great flexibility and the ability to maximize resource utilization on a host server through overprovisioning and oversubscription. Through the various methods and real world possibilities covered in this white paper, administrators have the best practices needed to leverage oversubscription of CPU, memory and storage to maximize utilization while maintaining performance in their virtual environments.

This paper also explored Dell's free vOPS Server Explorer tool that provides at-a-glance information about oversubscription as well as performance monitoring of your virtual environment.