Five best practices for VMware administrators: Microsoft Exchange on VMware

Scott Lowe

Founder and Managing Consultant
of the 1610 Group

Introduction

There was once a day when Microsoft specifically prohibited running Microsoft Exchange inside virtual environments due to the unknown ways by which the hypervisor-based environment could negatively impact the operation of Exchange. Those days are gone and Exchange is now fully supported as a first-class application in a virtual environment.

This acceptance of Exchange into the world of virtualization has taken place as a result of massive improvements in modern hypervisors coupled with massive improvements in Microsoft Exchange. Now, organizations using Exchange can easily leverage their hypervisor investments and implement Exchange just like any other application in the environment.

That said, it should never be forgotten that Exchange is a large, complex application upon which the business depends in order to conduct its operations. In most organizations, Exchange is considered a Tier 1 application deserving of top support and deserving of an operational environment that completely meets its needs.

When it comes to running Exchange in a vSphere-based environment, there are five critical items that VMware administrators should keep in mind in order to ensure that Exchange runs with maximum effectiveness. This guide is making an assumption that administrators are running Exchange Server 2010 with SP2 on vSphere 5 or higher.

Don't over allocate resources to Exchange virtual machines

Although it may be tempting, over allocating resources to an Exchange virtual machine can actually have a negative impact on the operation of both the Exchange server and the host environment.

Processing resources

Experienced vSphere administrators already know that over allocating vCPU resources in a virtual machine can actually have a negative impact on the performance of a virtual machine, particularly as a host gets busier. At first, it may seem a bit counterintuitive. After all, by over allocating processor resources, you're virtually guaranteeing that the virtual machine will always have the resources that it needs, right?

Unfortunately, this is not the case.

As administrators add more and more virtual CPUs to a virtual machine, at the same time, these administrators are instructing vSphere to hold on performing operations until matching physical resources can become available. For example, if an administrator creates a virtual machine with four virtual CPUs, then four physical processing resources – cores – need to be available when the virtual machine needs to run an operation. As the number of vCPUs increases, administrators may find CPU Ready Time, a metric, also increasing if the virtual machine has to wait for physical resources. CPU Ready Time is a measure of the amount of time between the point at which a virtual machine requests access to physical resources and the time at which the virtual machine is actually granted access to those resources.

From a best practices perspective, guidance on managing Exchange closely follows guidance for other workloads, but due to Exchange's need for high performance, it might be even more critical. Most important, don't make high-end assumptions regarding the number of virtual CPUs that should be allocated to an Exchange virtual machine. Instead, if you're unable to exactly target processing needs, start on the low end and, only as additional resources become necessary, add additional virtual CPUs to Exchange guests.

Here are some other considerations:

- When possible, try to limit the number of especially high resource consumers on hosts that will also support Exchange servers. This can be accomplished in a number of ways, but is mostly easily handled through the use of Distributed Resource Scheduler (DRS) rules. This step will help to reduce overall CPU contention and make it more likely that Exchange will be able to access the resources it needs when it needs them.

3

- While it may be tempting to use CPU reservations in order to make sure that Exchange has at least a minimum allocation of resources, this is not considered a best practice, unless an administrator must implement a reservation to meet an SLA. However, bear in mind that each time an administrator implements a CPU reservation, that chunk of processing power cannot be shared with other virtual machines that might be running on the host and, as such, may be wasted.

In addition to this guidance, make sure to use the tools that Microsoft makes available for assistance in sizing Exchange systems. Here are two such tools:

- [Understanding Processor Configurations and Exchange Performance](#)
- [Understanding Server Role Ratios and Exchange Performance](#)

RAM resources

Over the years, VMware has significantly improved vSphere's memory handling capabilities in order to maximize the number of virtual machines that can be run on a single host. Among vSphere's memory management techniques are:

- **Transparent Page Sharing.** Transparent Page Sharing (TPS) is a method by which RAM is deduplicated between virtual machines. For hosts that run many virtual machines with the same operating system, the savings can be substantial. TPS operates by waiting for host idle CPU time, at which point the system gathers pages from among virtual machines and shares them.

- **Guest Memory Balloon.** vSphere has a method by which memory can be borrowed from other virtual machines when it becomes necessary. When an administrator installs VMware Tools, one of the drivers that is installed is a memory ballooning driver. This driver is the method by which a virtual machine is able to share its memory with other virtual machines. Memory can be swapped into and out of the balloon as necessary.

- **Memory Compression.** Through Memory Compression, a vSphere server can fit multiple pages of RAM into fewer memory pages. This can help prevent a guest from having to swap to disk, which can lead to poor performance.

- **Memory Oversubscription.** Through a combination of all of these techniques and more, vSphere system administrators are able to allocate more RAM to virtual machines than is available in the host server.

- **Swapping to Disk.** Just like a Windows server will do when it runs low on RAM, a vSphere system can also swap to disk to reclaim some RAM, but since disk is much slower than RAM, swapping to disk carries with it a tremendous negative impact on overall performance.

Although Exchange is considered a full citizen when it comes to virtualization, there remain some recommendations that run counter to the overall philosophy that comes into play for many virtualization administrators. For many, the goal is to achieve the highest-level density as possible in the virtual environment. However, that guidance may result in an Exchange system that isn't running at optimal levels. The following best practices-based recommendations are made for vSphere administrators intending to support Exchange:

- **If using memory over-commitment, understand the limitations.** A memory reservation in vSphere can be used to ensure that a virtual machine is allocated a minimum amount of memory, but implementing such a reservation can create challenges for vMotion. When a vMotion takes place, vSphere must be able to locate a host that has enough available RAM to satisfy the defined memory commitment.

- **Closely monitor Exchange virtual machines and adjust RAM as necessary**. It may take some time, but Exchange is an application for which spending the time right-sizing RAM can pay off. Watch Exchange and determine its ongoing memory footprint and make RAM adjustments as necessary to meet performance needs. Make sure that the virtual machine has enough RAM such that internal swapping is not taking place.

As is the case for CPU resources, Microsoft makes available guidance for ensuring that RAM configurations will result in acceptable levels of performance:

- [Understanding Memory Configurations and Exchange Performance](#)

Consider multiple smaller virtual machines over fewer large ones

Many of vSphere's very best features, such as vMotion, Enhanced vMotion, Storage vMotion, Distributed Resource Scheduler, Fault Tolerance (FT) and High Availability (HA) tend to work more efficiently with smaller virtual machines. That's not to say that they don't work with large virtual machines; they simply work a bit better with smaller ones.

Some of these features have not been practical for Exchange environments until recently. For example, until only recently, Fault Tolerance supported only a single vCPU, which meant that the feature could not be used with substantial workloads.

As vSphere and Exchange administrators come together to design the Exchange environment, it's often desirable to break up the Exchange services among multiple smaller virtual machines than it is to attempt to operate fewer, larger virtual machines. By keeping virtual machines smaller, a number of operations, including workload migrations techniques (the various vMotions) work better and backups are easier to handle. In modern versions of Exchange, Microsoft has removed the need for the user to know the name of the server on which his mailbox is housed. As such, from a user perspective, multiple virtual machines won't be a problem.

In addition, multiple smaller virtual machines can improve availability opportunities. For example, when an administrator deploys multiple Hub Transport servers and multiple Client Access Servers, overall availability is improved.

Validate the Exchange environment design with Microsoft-provided tools

One way Microsoft helps administrators properly size their Exchange environments is by providing two tools that help Exchange administrators both size and test the environment where they will deploy Exchange. Although aimed at Exchange administrators, these two tools can be invaluable to the vSphere administrator responsible for developing and supporting the Exchange infrastructure. Both of these tools are intended to be run prior to placing Exchange into full production.

Exchange Server Jetstress 2010

The Exchange Server Jetstress provides an administrator with a way to test the stability and robustness of the storage system that will support the Exchange environment. Jetstress is intended to be used before any Exchange elements are installed. Better yet, Microsoft specifically supports Jetstress running inside a virtual machine under vSphere 4.1 and later.
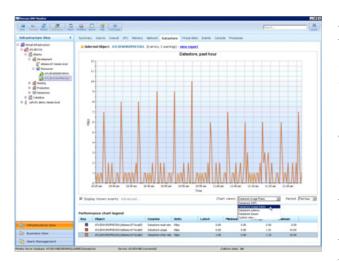
Jetstress is a complex tool, which, when used in conjunction with Performance Monitor, Event Viewer, and the Exchange Server Database Utilities (Eseutil.exe), provides vSphere administrators with comprehensive details about the operation of the storage infrastructure, a critical Exchange component.

Exchange Server Load Generator 2010

Exchange Load Generator is a tool that provides administrators with ways to measure the impact of MAPI (Outlook), OWA, ActiveSync, IMAP, POP and SMTP clients on Exchange servers. In short, these are the various ways by which users can connect to an Exchange server.

The Load Generator tool basically simulates various types of connectivity to an Exchange server. From the output, vSphere administrators can validate their deployments and begin to determine what kind of response time various clients may see when connecting to Exchange. Further, these values can help to identify an appropriate number of users per server, which can help vSphere administrators calculate an adequate number of virtual machines to build to support the Exchange environment.

Implement substantial monitoring and alerting

Any professional technology environment should include monitoring of both hosts and the individual workloads running on those hosts. With the addition of Exchange, which is considered by many to be a mission-critical communications service in the environment, ensuring that the infrastructure remains available is even more critical. Further, the monitoring solution should be able to go well beyond simple up and down monitoring by keeping tabs on a wide variety of metrics associated with a virtual environment. Virtualization has brought workloads together in ways that are unique, so ensuring that these workloads aren't interfering with one another is a key factor in an successful Microsoft Exchange experience.

In this space, vSphere administrators should look at Veeam ONE™, a comprehensive and powerful monitoring and reporting tool that supports both vSphere and Hyper-V environments. At the very least, Veeam ONE will point administrators in the right direction for troubleshooting efforts.

Stay current on guidance

Believe it or not, this paper may not have the definitive answers that you need in order to deploy your vSphere-based Exchange environment with maximum success. As new versions of Exchange are released and as new service packs are released, Microsoft makes further improvements in Exchange that may negate some existing best practices in favor of new ones.

For example, with the original release of Exchange 2010, Microsoft did not support vMotion for Exchange servers that were a part of a Database Availability Group (DAG). However, this capability is supported in Exchange Server 2010 SP2. The point here is simple: Just because something in this paper is true today doesn't mean it will be true tomorrow.

As such, it's critical that vSphere administrators stay as current as possible with guidance on an ongoing basis. Best practices change over time, so make it a best practice to stay abreast of best practices.

Data protection of a virtualized Exchange environment

One of the central benefits of virtualization with VMware vSphere is the increased protection options that come with the platform. Virtual machines are abstracted from the infrastructure, and critical applications within them such as Exchange need to be held to a high standard.

One approach to data protection, and specifically backups for vSphere virtual machines, is to take the agentless backup approach. That is good and works well, with many situations; in fact it is quite an agile process. It is important to remember the key requirements for Exchange backups like any other Tier 1 application:

- Application-consistent backups
- Exchange log pruning
- Granular recovery

vSphere administrators will appreciate having all of the critical requirements of the virtualized application (in this case Exchange); and having all of the benefits of a virtualization-specific backup solution.

Summary

There is no limit to the ways Exchange can be deployed into a virtual environment. However, doing so does require some additional planning on the part of the vSphere administrator to ensure overall success. The five tips presented in this paper are proven, real-world methods that can help ensure this success.

About the Author

Scott Lowe is the founder and managing consultant of the 1610 Group. He is a 17-year veteran of the IT world. Scott is also a frequent contributor on a wide range of topics to such outlets as TechRepublic, TechTarget and virtualizationadmin.com. He's a huge believer in the business benefits that can be had through the appropriate use of virtualization.

About Veeam Software

Veeam® Software develops innovative solutions for VMware backup, Hyper-V backup, and virtualization management. Veeam Backup & Replication™ is the #1 VM Backup solution. Veeam ONE™ is a single solution for real-time monitoring, resource optimization, documentation and management reporting for VMware and Hyper-V. Veeam extends deep VMware monitoring to Microsoft System Center with Veeam Management Pack™ (MP) and to HP Operations Manager with Veeam Smart Plug-In™ (SPI). Veeam also provides free virtualization tools. Learn more by visiting www.veeam.com.