Microsoft Exchange Server 2010 Performance on VMware vSphere® 5

Performance Study

TECHNICAL WHITE PAPER

Table of Contents

Introduction

Deploying Microsoft Exchange Server in a virtualized environment is becoming a common practice in today's IT industry. There is a high demand for information on how virtualizing Microsoft Exchange Server 2010, which can be a demanding and critical application, benefits from the latest hypervisor technologies provided by VMware vSphere® 5. This paper details experiments that demonstrate the efficient performance and scalability of Exchange Server 2010 when running inside a virtual machine in a VMware vSphere 5 environment.

Key performance requirements for Exchange Server are the efficiency of both single and multiple Exchange Server 2010 virtual machines, and migration using VMware vSphere® vMotion® and VMware vSphere® Storage vMotion® operations with no measurable impact to end-user experience. Acceptable single– and multiple–virtual machine performance is required to maintain a positive user experience. The high performance of vMotion and Storage vMotion features is key to assisting virtualization and allowing Exchange Server administrators to maintain a highly available environment for their users.

This paper addresses all three of the following performance characteristics:

- The performance implications of scaling up an Exchange Server 2010 mailbox server virtual machine in vSphere 5 in comparison to vSphere 4.1 (increasing the number of virtual CPUs in the virtual machine)
- The performance of scaling out with multiple Exchange Server 2010 virtual machines (increasing the number of virtual machines to the host)
- Data showing significant performance improvements of vMotion and Storage vMotion features in vSphere 5

Experimental Configuration and Methodology

The performance and sizing studies were done in VMware® internal labs. The purpose of the tests was to measure, analyze and understand the performance of Exchange Server in a vSphere 5 virtual environment. In the following sections, the test-bed configuration used is described in detail and the test tools are discussed. Finally, a description of the experiments is presented.

Test-Bed Configuration

A Dell PowerEdge R910 server was configured with four Intel Xeon Processors X7550 and 256GB of physical memory. Exchange Server 2010 was installed on Windows Server 2008 R2 in a virtual machine. The virtual environment used VMware vSphere 5 to virtualize the Windows/Exchange environment.

Here is the detailed test-bed configuration:

vSphere host – the system under test (SUT)

- Server: Dell PowerEdge R910 rack server
- Processors: Four Intel Xeon Processors X7550 (Nehalem-EX) @ 2.00GHz
- Memory: 256GB, 1,066MHz, DDR3 quad-ranked RDIMMs
- HBA: QLogic QLE2562-SP 8Gb dual-port Fibre Channel host bus adapter
- vMotion network interface controller: Intel Ethernet Server Adapter X520-SR2, 10Gb

Storage

- Storage enclosure: EMC CX3-40 with two storage processors
- Hard drives: 150 146GB 15K RPM drives
- Disk configuration for Exchange Server: 134 disks for the database and 16 for logs

Client

- Client machine: HP ProLiant DL580 server
- Processors: Two quad-core Intel Xeon Processors X7350, 2.93 GHz
- Memory: 64GB DDR2 DIMM
- Operating system: Microsoft Windows Server 2008 R2 Datacenter Edition (64-bit)
- Application: Microsoft Exchange Load Generator 2010 (64-bit)

Server hosting peripheral Windows and Exchange Server roles

(Microsoft Active Directory, DNS, Exchange Server Client Access and Hub Transport roles)

- Server: Dell PowerEdge R710 server
- Processors: Two Intel Xeon Processors X5570 @ 2.93GHz
- Memory: 96GB DDR3 DIMM
- Operating system: Microsoft Windows Server 2008 R2 Datacenter Edition (64-bit)
- Application: Exchange Server 2010 Client Access and Hub Transport roles were installed for only the scaling up of the Exchange Server 2010 mailbox server–role experiments

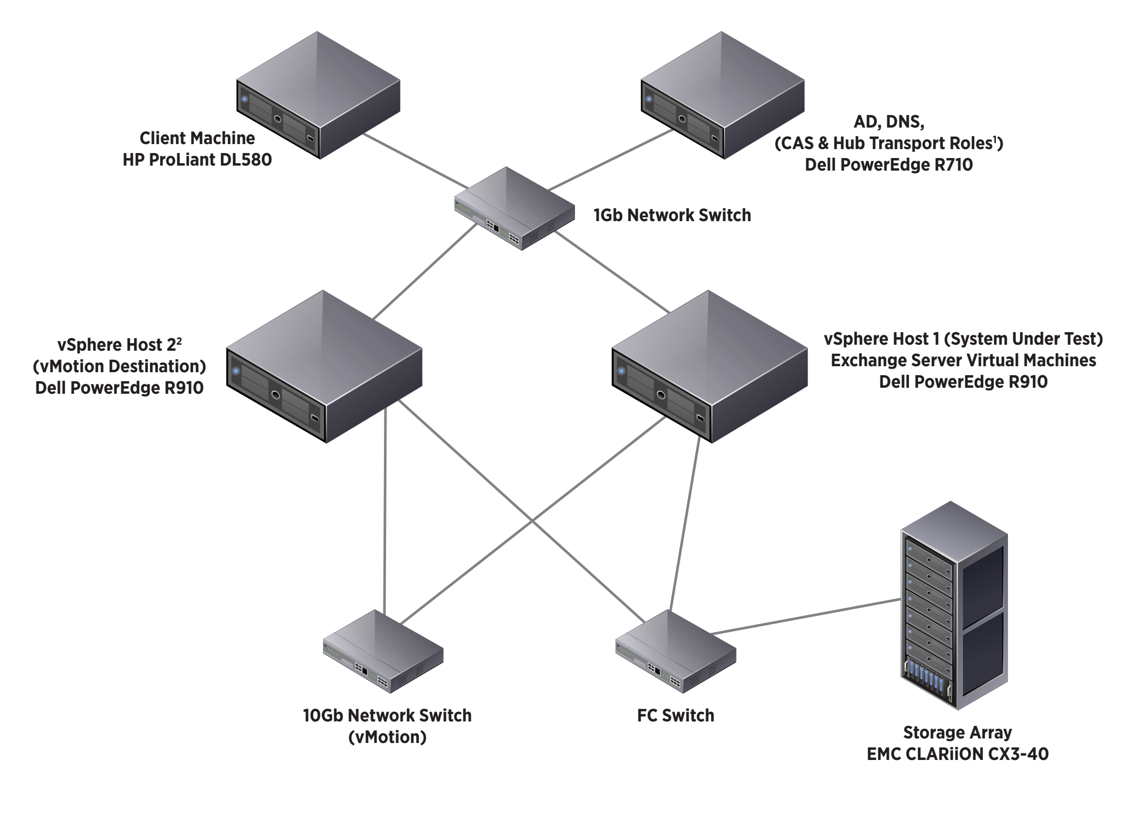

Figure 1. Test-Bed Layout

[1] The Client Access and Hub Transport combined server roles were installed only for the mailbox server virtual machine scale-up experiments.

[2] The vSphere Host 2 was online for the vMotion experiments only.

Test and Measurement Tools

The Microsoft Exchange Load Generator 2010 (LoadGen) runs on a client system and simulates the messaging load against Exchange Server deployments. LoadGen measures user response times for each of the different types of operations that it runs against the Exchange Server systems. In this case, LoadGen was configured to simulate Outlook 2007 online clients using the LoadGen very heavy–user profile, with 150 messages sent and received per day per user. Each mailbox user was initialized with 100MB data in the mailbox before starting the tests. LoadGen tests included a variety of Exchange Server transactions and were run for an 8-hour period to represent a typical workday. Quality of service (QoS) is determined by industry consensus, which indicates that testing should focus on the 95th-percentile transaction latency of the sendmail transaction to measure mailbox performance. The value of this result, which represents the maximum time needed to send an email for 95% of the transactions, should be reported as being below 500ms to represent an acceptable user experience.

The LoadGen benchmark models the normal daily email usage of real users and provides an estimate of the number of users a system can support. Although this benchmark is widely used to measure the performance of Exchange Server platforms, as with all benchmarks, the results might not match the specifics of your environment.

Microsoft Windows Perfmon (http://technet.microsoft.com/en-us/library/bb490957.aspx) is used to collect the latency and status counters of Exchange Server client tasks in the client machine.

For vSphere 5, esxtop was used to record both vSphere and virtual machine–related performance counters. Esxtop was configured to log CPU, memory, disk, network and system counters during LoadGen runs.

Test Cases and Test Method

The configuration of user and resource (processor and memory) sizing was based on Microsoft's recommendations (http://technet.microsoft.com/en-us/library/dd346699.aspx and http://technet.microsoft.com/en-us/library/dd346700.aspx). One thousand very heavy users were configured per processor, and memory was sized to 4GB plus 6MB per mailbox for the Exchange Server mailbox virtual machine.

The following are the three primary objectives in performing these experiments:

1. Compare the scale-up performance of vSphere 5 with that of vSphere 4.1 in a single Exchange Server virtual machine. This test included the following experiments:
    - vSphere 4.1 scale-up experiments with 2, 4 and 8 vCPUs.
    - vSphere 5 scale-up experiments with 2, 4, 8 and 12 vCPUs (an additional 12–virtual processor scenario was conducted). vSphere 5 has increased the number of supported virtual processors per virtual machine to 32 vCPUs, which enables the test bed to utilize the recommended maximum (12) number of processor cores for a virtualized Exchange Server mailbox server.

The following table shows the number of CPUs, memory size and number of very heavy users for each scenario in the scale-up experiments:

NUMBER OF VIRTUAL CPUs	MEMORY SIZE (GB)	TOTAL NUMBER OF HEAVY USERS
2	16	2,000
4	28	4,000
8	52	8,000
12	76	12,000

Table 1. Number of vCPUs and Memory Size per Number of Users

2. Understand the scale-out performance of Exchange Server on vSphere 5 by increasing the number of virtual machines. The Exchange Server environment was scaled out by deploying one combined Client Access/Hub Transport virtual machine for every mailbox server virtual machine in a 1:1 ratio as recommended in the "Performance and Scalability" section for Exchange Server 2010 on Microsoft TechNet (http://technet.microsoft.com/en-us/library/ee832795.aspx).

   The two types of virtual machine configurations for the scale-out tests are:

   - Mailbox server virtual machine with four vCPUs and 28GB memory to support 4,000 users
   - Client Access and Hub Transport combined-role server virtual machines with four vCPUs and 8GB memory

To demonstrate scale-out performance, the number of Exchange Server virtual machines was increased to eight (four mailbox server virtual machines plus four Client Access and Hub Transport combined-role virtual machines) with 16,000 Exchange Server very heavy users.

The following table shows the number of CPUs, memory size and number of very heavy users for each scenario in the scale-out experiments:

NUMBER OF MAILBOX AND CAS/HT VIRTUAL MACHINES	TOTAL NUMBER OF vCPUs IN VIRTUAL MACHINES	TOTAL MEMORY (GB) IN VIRTUAL MACHINES	TOTAL NUMBER OF VERY HEAVY USERS
1+1	8	36	4,000
2+2	16	72	8,000
3+3	24	108	12,000
4+4	32	144	16,000

Table 2. Number of Virtual Machines and vCPUs, Memory Size, and Number of Users in Scale-Out Scenarios

In each of the scalability test scenarios, Exchange Server transaction latencies and system CPU utilization were used to measure performance. Transaction latency directly predicts user experience, and CPU utilization reflects efficiency. The tests were run for 8 hours per LoadGen guidelines.

3. Measure and compare the impact to end users during vMotion and Storage vMotion migration on vSphere 5 and vSphere 4.1. We chose the 16,000 very heavy–users load with eight virtual machines hosted in a vSphere server (Host 1) for the premigration state. The vMotion operations occurred after the LoadGen tests had reached a stable state 2 hours into the test run. A 10Gb bandwidth link was configured for a dedicated vMotion connection. The following two vMotion migration experiments were conducted:

   vMotion: Migrate a four-vCPU/28GB-memory mailbox server virtual machine from vSphere Host 1 to vSphere Host 2.

   Storage vMotion: Migrate a 350GB Exchange Server database disk containing 2,000 user mailboxes from a current (source) datastore to another (target) datastore. Both datastores are six-disk LUNs with separate spindles, and both datastores exist on the same EMC CX3-40 CLARiiON array.

Experimental Results and Performance Analysis

Results are presented to show the scale-up performance of a single virtual machine, scale-out performance of many virtual machines, and the impact to end-user response time during vMotion and Storage vMotion migrations of an Exchange Server virtual machine and an Exchange Server database disk.

Single–Virtual Machine Performance

The same server and virtual machine configurations were used for the vSphere 5 and vSphere 4.1 environments, with scale-up tests conducted moving from two to eight virtual processors. An additional 12–virtual processor scenario was conducted in the vSphere 5 environment. This test is not possible in vSphere 4.1, due to its limit of eight vCPUs.

Exchange Latencies

Among all of LoadGen's reported transaction latencies, sendmail is the most prominent. Its latency is considered a yardstick of user responsiveness. The 95th percentile for sendmail latencies being at or below 500ms is considered acceptable user responsiveness. Sendmail is therefore used to compare performance across the various configurations.

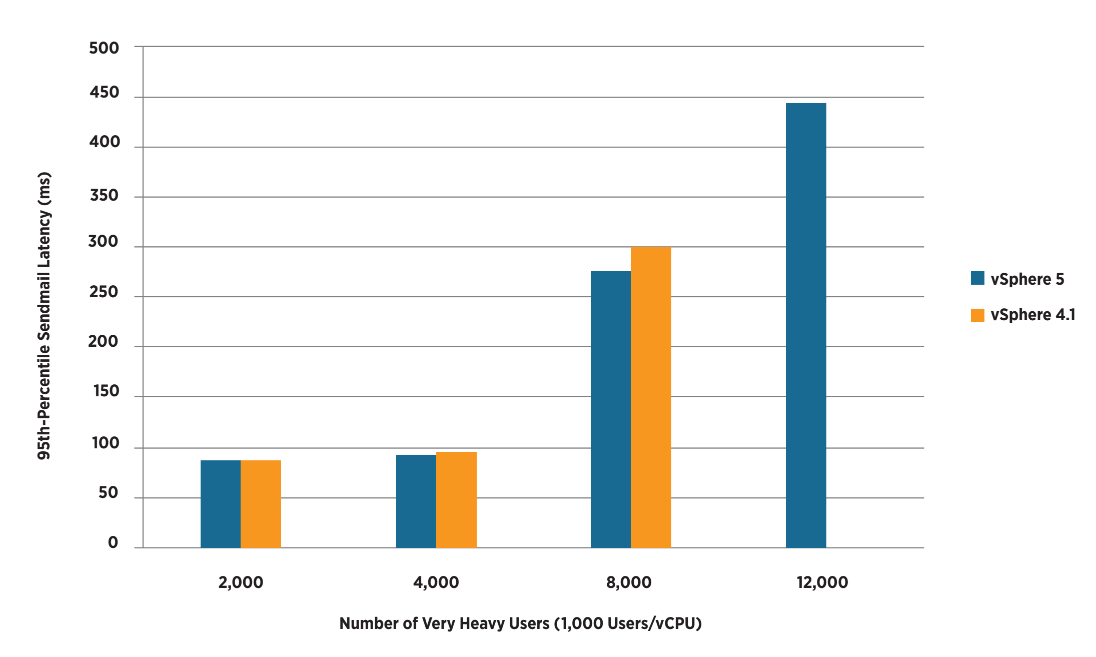

Figure 2. 95th-Percentile Sendmail Latency (vSphere 5 in Comparison to vSphere 4.1)

The vSphere 5 and vSphere 4.1 Exchange Server environments offered the same 84ms latency in the 2-vCPU virtual machine scenario. As the numbers of virtual processors increased, the latencies of vSphere 5 outperformed those of vSphere 4.1 by 6% in the 4–vCPU virtual machine and by 13% in the 8–vCPU virtual machine scenarios. The 95th-percentile sendmail latency of the 12–vCPU virtual machine scenario in vSphere 5 is 444ms, which is much lower than the standard maximum of 500ms.

Processor Utilization

The overall vSphere host CPU utilization was measured for each of the test cases. The CPU utilization is an important metric to evaluate efficiency. For these results, 100% is used to represent the power of a CPU core. The total CPU power of the 32 processor–based vSphere server used in these tests is 3,200%.

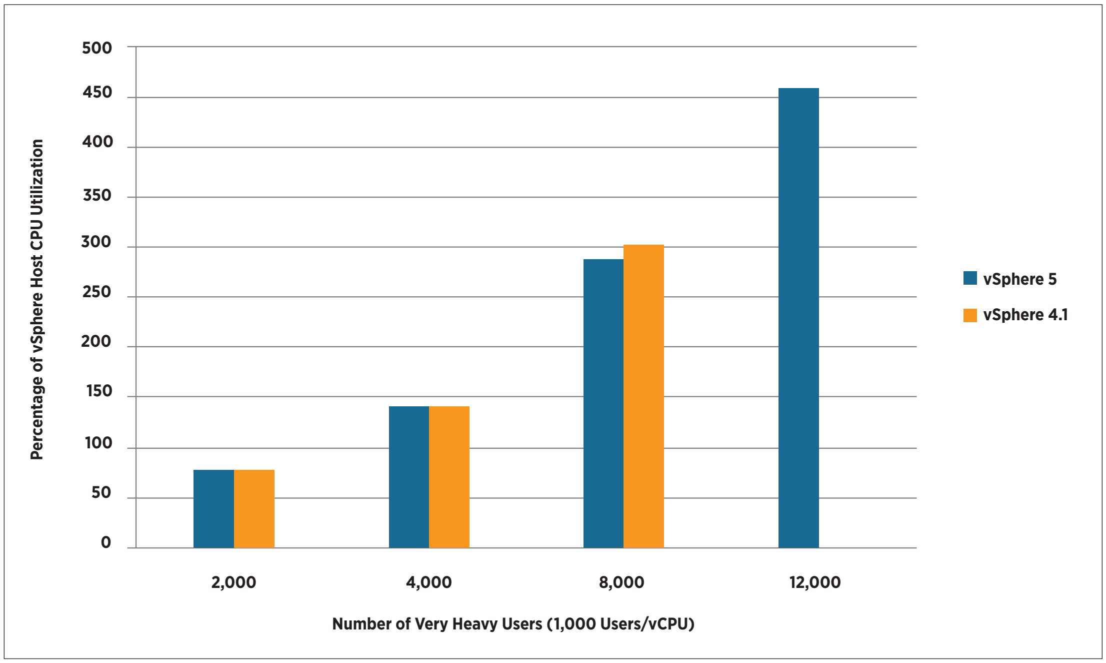

Figure 3. Percentage of CPU Utilization (vSphere 5 in Comparison to vSphere 4.1), 100% = 1 CPU

Figure 3 shows that the vSphere 5 and vSphere 4.1 Exchange Server environments have similar CPU consumption for supporting either 2,000 or 4,000 users. The CPU utilization of the vSphere 5 environment for supporting 8,000 users was 5% less than that of vSphere 4.1 (reduced to 288% from 301%). With 12,000 users on the vSphere 5 environment, the CPU utilization increase was linear, specifically 5.9 times the CPU utilization of the 2,000-users case. The 459% CPU utilization for the 12–virtual processor mailbox virtual machine supporting 12,000 users consumed less than 15% of the total host processing capacity.

Multiple–Virtual Machine Performance

Multiple–virtual machine environments increase load. They also stress vSphere storage and networking stacks as well as its resource management (CPU, memory scheduler) modules. The Exchange Server virtualization environment was scaled out to demonstrate up to 16,000 very heavy users in eight Exchange Server virtual machines. Each set of Client Access/Hub Transport combined virtual machine and mailbox server virtual machine was configured to support 4,000 very heavy users.

Exchange Latencies

Figure 4 depicts the 95th-percentile sendmail transaction latencies for the multiple virtual machines running concurrently.

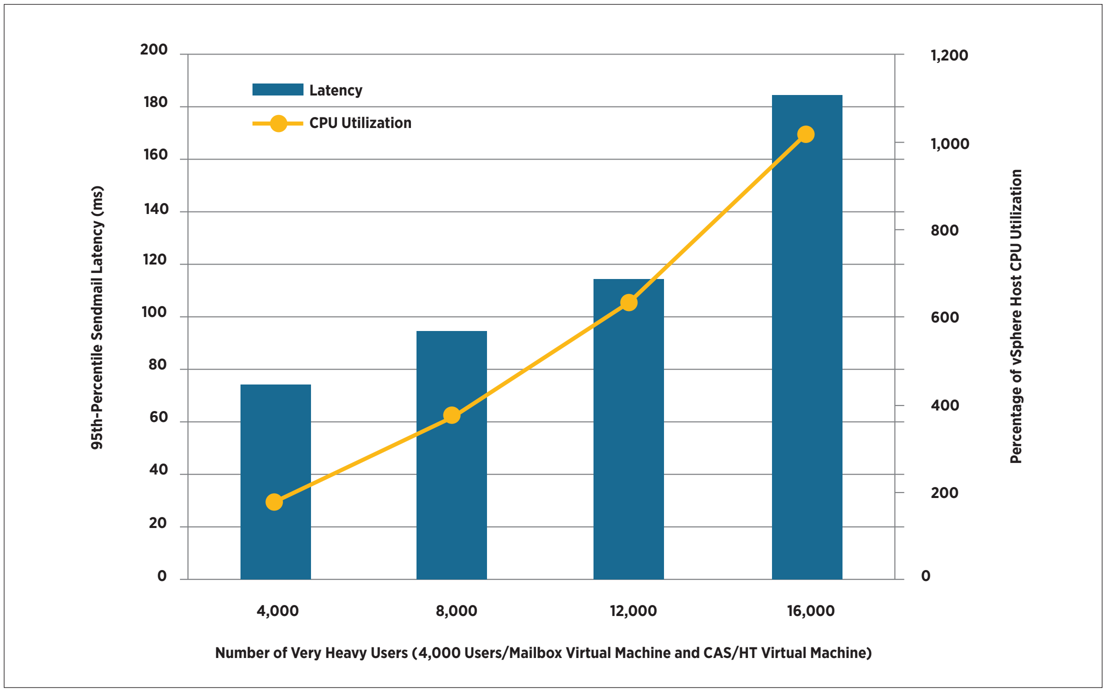

Figure 4. 95th-Percentile Sendmail Latency and CPU Utilization for Scale-Out

The 95th-percentile sendmail latency at 16,000 users in eight Exchange Server virtual machines was 184ms, far below the 500ms threshold recommendation. The scenarios with 4,000, 8,000 and 12,000 users got superior latencies, equal to or less than 114ms.

The 16,000–very heavy users load in the vSphere 5 environment consumed 1,022% CPU, which is only 32% of the host processing capacity. CPU utilization increased at a greater than linear rate when testing reached 16,000 users. Investigation into the CPU usage of each virtual machine revealed that additional CPU usage was due to more application instructions between the Client Access and Hub Transport combined-role virtual machines.

vMotion and Storage vMotion

vMotion provides live migrations of virtual machines across physical servers. Storage vMotion offers live migrations of virtual machine disk files across physical storage locations. These are virtualization capabilities that give administrators a unique flexibility in managing their environments. This section discusses the details of vMotion and Storage vMotion experiments and the performance improvements in vSphere 5.

The following three metrics are used to measure Exchange Server user impact during migration:

Migration time: The total time to complete a migration. During the migration, any moving virtual machine or virtual machine disk could have slightly slower responses to requests.

Maximum task queue length: A LoadGen performance counter presents the number of Exchange Server tasks currently in the queue (waiting) for dispatching to Exchange Servers. The maximum task queue length during a migration period was recorded as an important factor in measuring the impact on the end-user experience. Normally, the task queue length is between 0 and 1 during the LoadGen run in standard experiments without any migration.

Number of task exceptions: A LoadGen performance counter presents the number of task executions that resulted in a fatal exception, typically due to lack of response from Exchange Servers.

The following two experiments were conducted:

1. Use vMotion to move a mailbox server virtual machine from vSphere Host 1 to vSphere Host 2.
2. Use Storage vMotion to move a 350GB Exchange Server database disk containing 2,000 user mailboxes from a current datastore to another datastore on the same EMC CX3-40 CLARiiON array.

vMOTION	MIGRATION TIME (SECONDS)	MAXIMUM TASK QUEUE LENGTH	NUMBER OF TASK EXCEPTIONS
vSphere 5	47	219	0
vSphere 4.1	71	294	0

Table 3. vMotion Performance Comparison Between vSphere 5 and vSphere 4.1

The migration time of an Exchange Server mailbox server virtual machine using vMotion was significantly reduced by 34%, from 71 seconds in vSphere 4.1 to 47 seconds in vSphere 5. The maximum task queue length was also reduced, from 294 to 219, which further shows that Exchange Server users can get a better response time during the migration period in the vSphere 5 environment. There were no reported task exceptions during migrations, which means no Exchange Server task was dropped in either the vSphere 5 or vSphere 4.1 environment.

STORAGE vMOTION	MIGRATION TIME (SECONDS)	MAXIMUM TASK QUEUE LENGTH	NUMBER OF TASK EXCEPTIONS
vSphere 5	728	1	0
vSphere 4.1	822	2	0

Table 4. Storage vMotion Performance Comparison Between vSphere 5 and vSphere 4.1

Table 4 shows that the migration time of a 350GB Exchange Server database disk was reduced by 11%, from 822 seconds in vSphere 4.1 to 728 in vSphere 5. The value of the maximum task queue length was maintained at 1 in vSphere 5 and 2 in vSphere 4.1. No task exceptions were reported in either environment.

Conclusion

The following results presented in this paper show that Microsoft Exchange Server 2010 running on VMware vSphere 5 achieves excellent performance and scalability:

- Exchange Server sendmail 95th-percentile latencies remain lower than the standard requirement of 500ms for both single–virtual machine and multiple–virtual machine environments.
- For 16,000 very heavy users, transaction latencies were only 184ms and consumed less than 32% CPU utilization on the host.

These low reported latencies for 95% of the mailboxes provide an outstanding response time for Exchange Server users. The low CPU utilization on the host leaves huge room for further user growth.

This paper also demonstrates that in addition to the exhibited low latencies and low CPU utilization in comparison to vSphere 4.1, vSphere 5 performance resulted in the following:

- A 34% reduction in vMotion migration time for an Exchange Server mailbox server virtual machine
- An 11% reduction in Storage vMotion migration time for a 350GB Exchange Server database disk

With these migration time reductions, the response time for Exchange Server tasks during the migration period also improved. These high-performance features—vMotion and Storage vMotion—are the keys to maintaining uninterrupted service for Exchange Server users if any migration occurs during load balancing or hardware maintenance, or even from a failing server.

The performance results and analysis demonstrate that vSphere 5 is an ideal virtualization platform for small, medium or large Exchange Server production deployments.

About the Author

Vincent Lin is a performance engineer at VMware. In this role, his primary focus is to evaluate and help improve the performance of VMware products in better supporting key enterprise applications. Prior to coming to VMware, Vincent was a principal performance engineer at Oracle. He received a Master of Science degree from the University of West Florida.

Acknowledgements

The author would like to thank the VMware development teams for their efforts in developing and optimizing vSphere. He would also like to thank the following members of the VMware performance team who have helped with the performance analysis and with the creation of this document: Julie Brodeur, Seongbeom Kim, Todd Muirhead, Aravind Pavuluri, Leah Schoeb and Sreekanth Setty.

vmware®

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com