VMware vSphere 5.5 vMotion on EMC VPLEX Metro

Performance and Best Practices

TECHNICAL WHITE PAPER

Table of Contents

Executive Summary

As the scope of virtualized infrastructure continues to grow beyond a traditional single physical site to geographically dispersed multi-site environments, the ability to vMotion transparently across large distances will be a critical requirement. vMotion in vSphere 5.5 incorporates a number of performance enhancements that enable a seamless migration of even large virtual machines running enterprise workloads over a metro distance. In addition, vMotion in vSphere 5.5 offers support for EMC VPLEX Metro, which enables shared data access across metro distances.

A series of tests were conducted at VMware Performance labs, in cooperation with EMC, to characterize vMotion performance on a VMware vSphere® 5.5 virtual infrastructure that was stretched across two geographically dispersed datacenters using EMC VPLEX Metro. This white paper describes the testing methodology, workloads used, and the performance results. In addition, it outlines the best practices to follow when using vMotion in a VPLEX Metro environment.

Technology Overview

VMware vSphere® vMotion® and EMC VPLEX Metro technologies discussed in this white paper are briefly described in the sections below.

Note: While this paper specifically tests vSphere 5.5 vMotion with EMC VPLEX Metro, benchmarking similar metro deployment solutions that use an active-active replicated storage topology should show similar results. vSphere 5.5 supports a number of VMware vSphere Metro Storage Cluster (vMSC) solutions, which can be seen by searching the *VMware Compatibility Guide* on the keywords **metro esxi5.5** and the drop-down menu **Storage/SAN**.

vMotion in vSphere 5.5

vMotion—a key feature of the enterprise virtual infrastructure—enables live migration of a running virtual machine (VM) from one physical host to another. Migration of a VM with vMotion preserves the precise execution state of a virtual machine consisting of physical memory, storage, and virtual device state including CPU, network and disk adaptors, and SVGA. The VM continues to run throughout the migration process with minimal impact to the user workload and no disruption to network connectivity.

vMotion brings invaluable benefits to administrators because it helps prevent server downtime, enables troubleshooting, and provides flexibility. vMotion is a key enabler of a number of VMware technologies, including vSphere Distributed Resource Scheduler (DRS) and vSphere Distributed Power Management (DPM). These technologies together create a dynamic, automated, and self-optimizing datacenter.

Some of the features of vMotion include:

- Live migration of an entire virtual machine across vSphere hosts without any requirement for shared storage.
- A multi-NIC capability that transparently load-balances vMotion traffic over all of the vMotion-enabled NICs.
- Allows concurrent vMotions to reduce overall migration time in virtual machine evacuation scenarios.
- Enables live migration on metro networks with round-trip latencies of up to 10 milliseconds.

EMC VPLEX Metro

EMC VPLEX represents the next-generation architecture for data mobility and information access across geographically dispersed sites. It uses remote data replication techniques to allow servers at multiple sites to have simultaneous read/write access to block storage devices. EMC VPLEX includes a portfolio of replication solutions over any distance, using synchronous data replication for shorter distances, and asynchronous data replication for longer distances.

VPLEX Metro, part of the EMC VPLEX family, enables distributed and shared data access across two sites within a metro distance using synchronous data replication. In the VPLEX Metro environment, an I/O-update operation is not considered done until completion is confirmed at both the sites.

VPLEX Metro is a hardware appliance deployed between physical hosts and SAN arrays. VPLEX Metro appliances in the two data center sites work together to provide a coherent cache in both locations, sending updates across the link between the sites.

The unique characteristics of VPLEX Metro include:

- Provides a consistent view of LUNs between two VPLEX clusters separated by synchronous distances enabling new models of high availability and collaboration.
- Supports synchronous distributed volumes that mirror data between the two clusters using write-through caching.

Paired with vSphere vMotion, VPLEX Metro simplifies vMotion over metro distances because it removes the need to migrate virtual disks belonging to the virtual machine. This is because VPLEX appliances in both locations present the same identity to the LUNs, where VMs' virtual disks reside, to both source and destination vSphere hosts. As a result, vMotion views the underlying storage as shared storage, or exactly equivalent to a SAN that both source and destination hosts have access to. Hence, vMotion with VPLEX Metro is as easy as traditional vMotion that live migrates only the memory and device state of a virtual machine. This will significantly reduce vMotion duration and improves workload mobility while optimizing business and operational goals.

Performance Test Configuration and Methodology

The following sections describe the logical and physical layouts of the test-bed configuration and testing methodology.

Logical Layout

Figure 1 illustrates the deployment of the VPLEX Metro system used for vMotion testing. The figure shows two data centers, each with a vSphere host connected to a VPLEX Metro appliance.

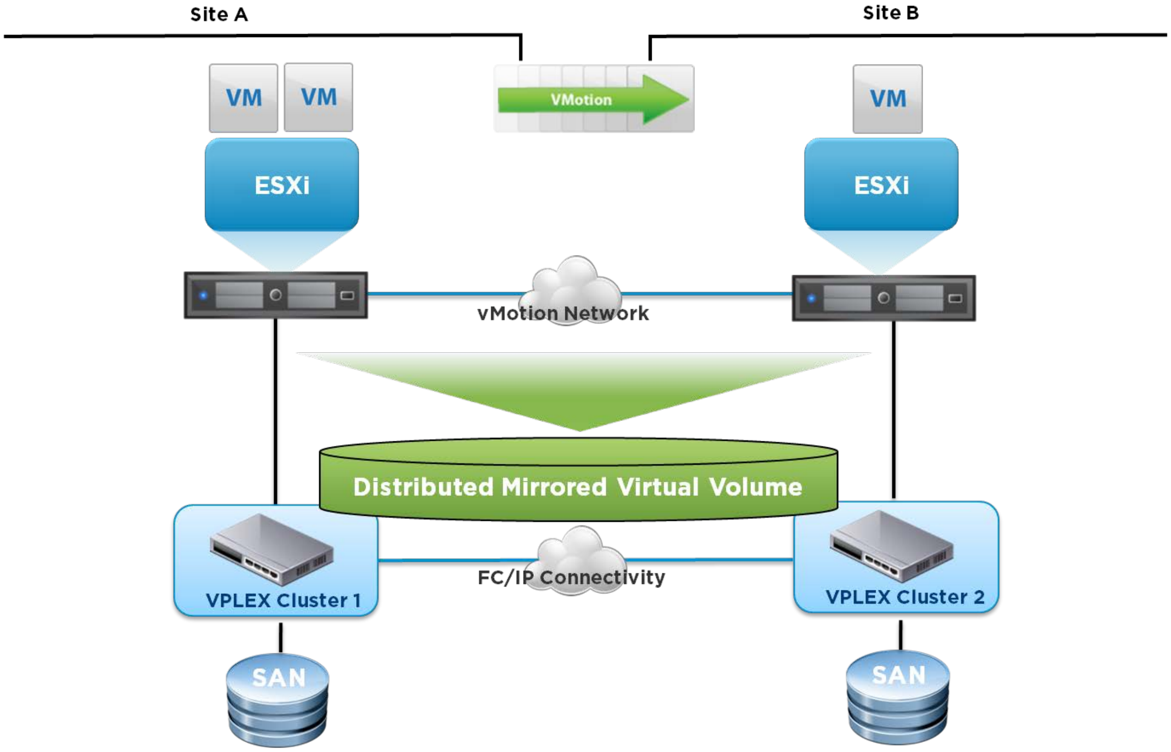

Figure 1. Logical layout of VPLEX Metro deployment between two data centers

The VPLEX virtual volumes presented to the vSphere hosts in each data center are synchronous, distributed volumes that mirror data between the two VPLEX clusters using write-through caching.

The connectivity between the VPLEX Metro appliances can be either Fibre Channel (FC) or IP-based. The two VPLEX Metro appliances in our test configuration used IP-based connectivity. The IP network connections between the two VPLEX Metro appliances passed through a WAN emulator (Dummynet) to induce latency. VPLEX uses the IP network between the VPLEX clusters primarily to mirror the writes at each cluster to the other cluster.

The vMotion network between the two ESXi hosts used a physical network link distinct from the VPLEX network.

Hardware Configuration

The VPLEX Metro test bed consisted of two identical VPLEX clusters, each with the following hardware configuration:

- Dell R610 host, 8 cores, 48GB memory, Broadcom BCM5709 1GbE NIC
- A single engine (two directors) VPLEX Metro IP appliance
- FC storage switch
- VNX array, FC connectivity, VMFS 5 volume on a 15-disk RAID-5 LUN

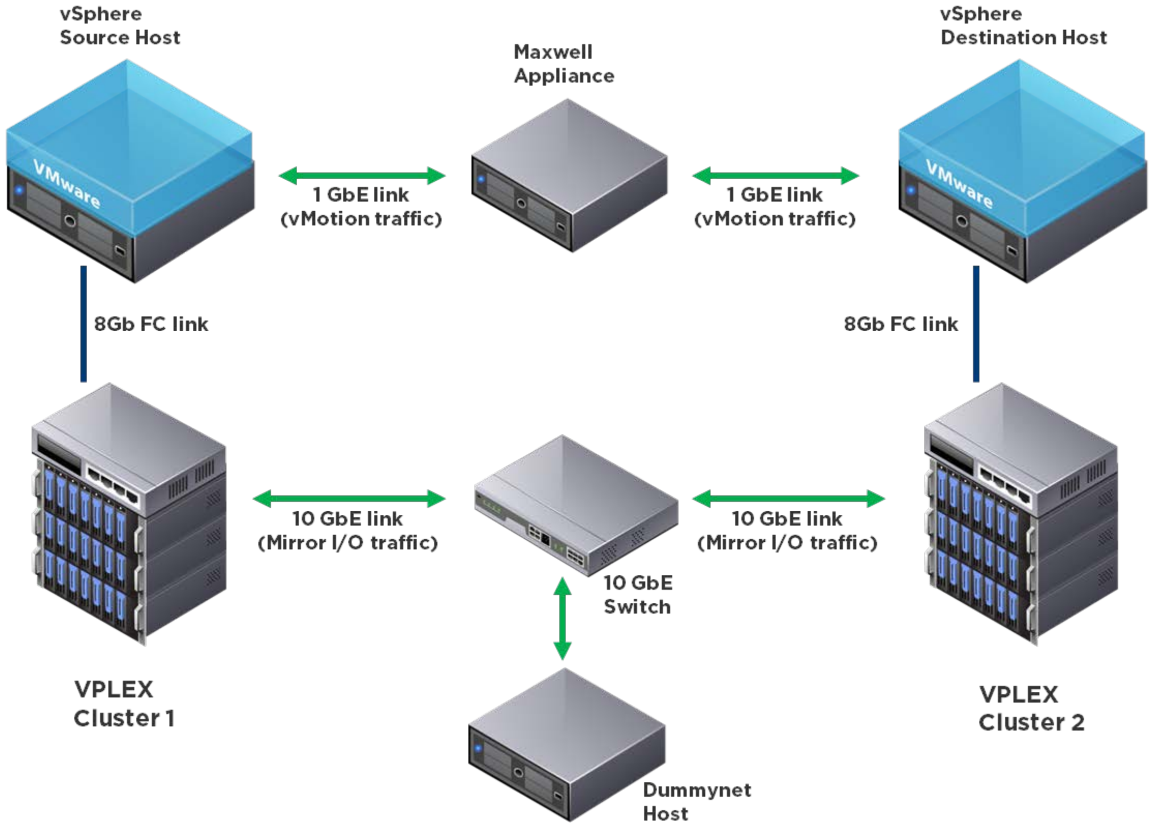

Figure 2. Physical layout of test-bed configuration

As shown in Figure 2, the vMotion traffic between the source and destination vSphere hosts passed through a Maxwell physical appliance to induce latency. We used a 1GbE link for the vMotion network to simulate a realistic bandwidth for Metro migrations. VPLEX WAN ports will only run at 10GbE, so a 10GbE Extreme Network switch was used to connect the WAN links of the two VPLEX clusters. A FreeBSD 9.0 host used the Dummynet network emulation tool to inject network latencies into the WAN link between the two VPLEX clusters.

Measuring vMotion Performance

The following metrics were used to understand the performance implications of vMotion:

- **Migration Time:** The total time taken for the migration to complete, beginning from the initiation of the migration.
- **Switch-Over Time:** The time during which the virtual machine is quiesced to enable virtual machine switchover from the source host to the destination host.
- **Guest Penalty:** The performance impact (latency and throughput) on the applications running inside the virtual machine during and after the migration.

To investigate the performance implications of vMotion on VPLEX Metro, we used different workloads including Iometer and an OLTP workload characterized by high consumption of CPU, memory, and storage resources.

vMotion Performance over VPLEX Metro Deployment

In this section, vMotion performance on a VPLEX Metro deployment under varying network latencies is examined.

Test Methodology

Load-Generation Software

The open-source Iometer was used as the benchmark tool. Iometer is widely used to measure and characterize I/O subsystem performance.

The test case used the following benchmark deployment scenario:

- Iometer ran in a virtual machine configured with two vCPUs and 4GB of memory.
- A Windows Server 2008 R2 virtual machine was configured with three virtual disks.
- The I/O pattern used the following profile: 50GB data disk, one worker thread, 100% write, 8KB I/Os, 100% random access, and 16 outstanding I/Os.

Tests varied the Dummynet delay parameter (used to inject latencies) to simulate a 5 millisecond and a 10 millisecond round-trip latency in a VPLEX Metro IP network. Similar latency was injected into the vMotion network as well. We also considered the baseline deployment scenario without any delay. The baseline deployment scenario had a round-trip latency of 0.5 milliseconds.

Test Results

In VPLEX Metro deployments, the write performance highly depends upon the VPLEX WAN round-trip-latency (RTT latency). The read performance is not impacted by the RTT latency. This is because the writes follow through the VPLEX Metro write-through cache model, which requires write I/Os to be mirrored on both VPLEX clusters. Synchronously replicating the data on a metro distance link will impact the write I/O latency, and thereby impact the write throughput. Accordingly, even though we used the same Iometer profile, we observed different write throughput in our test scenarios. Before the start of vMotion, the average write I/O operations per second in the baseline, 5ms, and 10ms test scenarios were about 2060, 1950, and 1350, respectively.

Figure 3 below shows the total duration and switch-over duration for various metro latency scenarios.

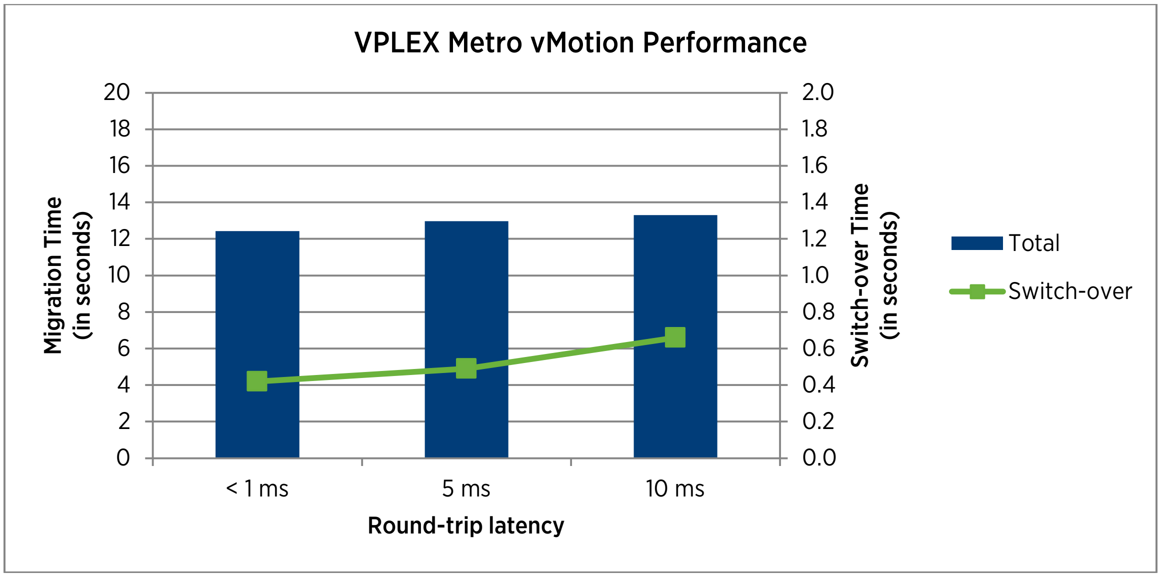

Figure 3. vMotion performance in VPLEX Metro scenarios

As shown in Figure 3, the vMotion duration in the 5ms and 10ms scenarios was nearly identical to the vMotion duration in the baseline scenario, thanks to the latency-aware optimizations in vMotion. The VM switch-over time was well below one second in all the scenarios. Although not shown in the figure, we found the impact on the write throughput during migration to be very minimal in all the scenarios.

These results indicate vMotion performs well on VPLEX Metro deployments with up to 10ms round-trip latency.

VPLEX Metro vMotion Performance in a Database Environment

Database workloads are widely acknowledged to be extremely resource intensive. They are characterized by a high consumption of CPU, memory, and storage resources, and so serve as useful tests of live migration performance. This study investigates the impact of VPLEX Metro live migration on Microsoft SQL Server online transaction processing (OLTP) performance.

Test Methodology

Load-Generation Software

The open-source DVD Store Version 2 (DS2) was used as the benchmark tool. DS2 simulates an online ecommerce DVD store, where customers log in, browse, and order products. The benchmark tool is designed to utilize a number of advanced database features, including transactions, stored procedures, triggers, and referential integrity. The main DS2 metric is orders per minute (OPM).

The DVD store benchmark driver was configured to enable fine-grained performance tracking, which helped to quantify the impact on SQL Server throughput (orders processed per second) during different phases of migration. Specifically, the source code of the `ds2xdriver.cs` file was edited with 1-second granularity, which resulted in the client reporting the performance data every 1 second (the default is 10 seconds).

The test case used the following SQL Server deployment scenarios:

- Microsoft SQL Server was deployed in a single virtual machine configured with four vCPUs and 8GB of memory.
- The DS2 workload used a database size of 50GB with 50 million customers.
- A benchmark load of 36 DS2 users was used.
- The virtual machine was configured with three virtual disks: a 50GB boot disk containing Windows Server 2008 R2 and Microsoft SQL Server, a 90GB database disk, and a 24GB log disk.

In our test scenario, a load of 36 DS2 users generated a substantial load on the virtual machine in terms of CPU, memory, and disk usage. The migration was initiated during the steady-state period of the benchmark, when the CPU utilization (esxtop %USED counter) of the virtual machine was close to 100%. The average read I/Os per second and the average write I/Os per second of the database disk were in the range of 1600 and 400, respectively.

Test Results

Figure 4 plots the performance of a SQL Server virtual machine in orders processed per second at a given time—before, during, and after VPLEX Metro vMotion with 10ms RTT latency.

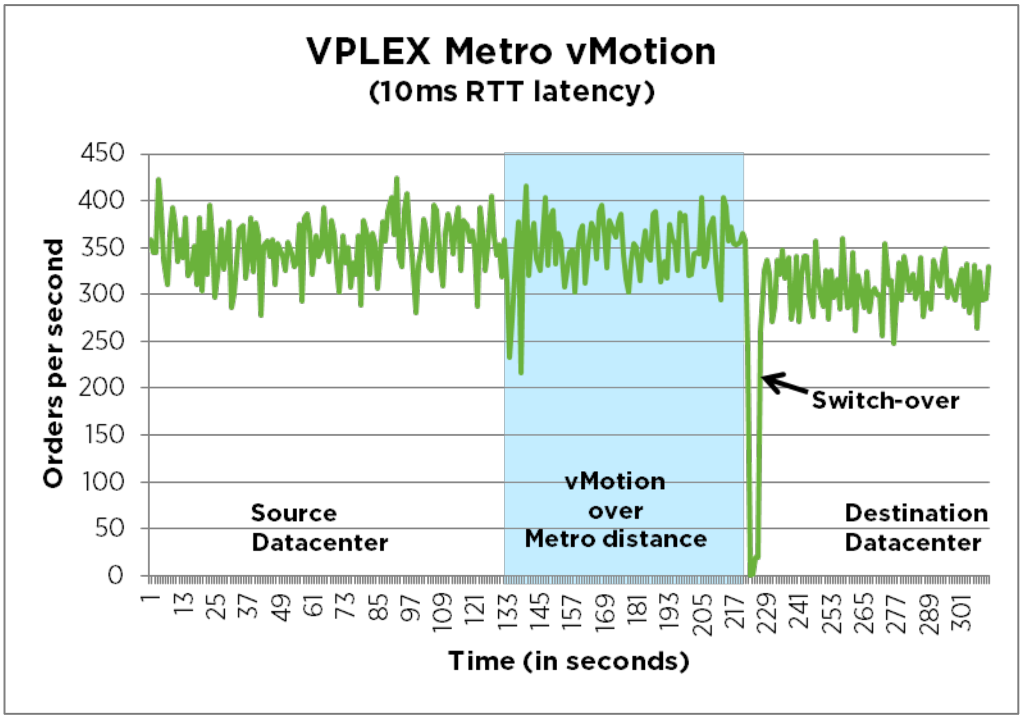

Figure 4. Performance of SQL Server VM before, during, and after VPLEX Metro vMotion

Figure 4 shows that vMotion started at about 131 seconds into the steady-state interval, and the total duration of vMotion was 88 seconds. As shown in the figure, the impact on SQL Server throughput was very minimal during vMotion. Despite the 10ms RTT latency, vMotion network bandwidth usage was close to line rate, thanks to the latency-aware optimizations in vMotion. A temporary drop in throughput occurred during the switch-over phase, when the virtual machine was quiesced on the source host and then resumed on the destination host.

The SQL Server throughput on the destination host was around 310 orders per second, compared to the throughput of 350 orders per second on the source host. This throughput drop after vMotion is due to the VPLEX inter-cluster cache coherency interactions and is expected. For some time after the vMotion, the destination VPLEX cluster continued to send cache page queries to the source VPLEX cluster and this has some impact on performance. The amount of time it takes to transfer all the metadata, including the ownership of cache pages to the destination VPLEX cluster, depends on the workload characteristics and how long the VM was running on the source VPLEX cluster prior to vMotion. After all the metadata is fully migrated to the destination cluster, we observed the SQL Server throughput increase to 350 orders per second, the same level of throughput seen prior to vMotion.

Results from these tests indicate the performance impact of vMotion on VPLEX Metro is very minimal even when running a highly I/O-intensive database application.

Performance Impact in Presence of Multiple Disks and Snapshots

In this section, we examine the impact of the number of disks and snapshots on VPLEX Metro vMotion performance.

Test Methodology

The test environment used the following VM configuration:

- An idle Ubuntu Linux VM with two vCPUs and 4GB of memory.
- The number of virtual disks of the VM varied from 3 VMDKs to 12 VMDKs.
- The number of snapshots of the VM varied from zero to two snapshots.

As in previous tests, the Dummynet delay parameter was used to simulate a 5ms and a 10ms round-trip latency in a VPLEX Metro IP network. Similar latency was injected into the vMotion network as well.

While the majority of the migration tests used idle VMs, we also included a few active VM migration tests to understand if the CPU and memory load on the VM had any additional impact on switch-over time. In the active VM tests, the VM ran a memory-intensive program that allocated 2GB memory in the virtual machine and dirtied the pages in an infinite loop.

Test Results

Figure 5 below shows the impact of the number of virtual disks and the RTT latency on the VM switch-over time.

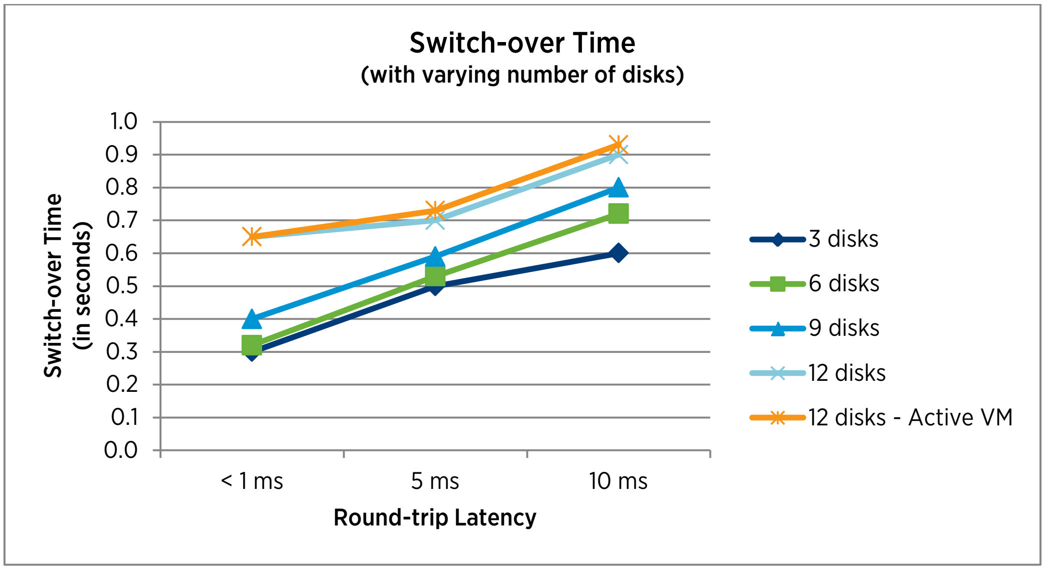

Figure 5. Virtual machine switch-over time in metro scenarios with varying numbers of disks

In all the test scenarios, there was a small, but steady increase in switch-over time as we increased the RTT latency in our test configuration. For instance, in the “3 virtual disks” test configuration, the switch-over duration increased from about 0.3 seconds in the baseline scenario (<1ms RTT latency) to about 0.6 seconds in the 10ms RTT latency scenario. This increase in switch-over time is a result of two contributing factors: RTT latency in the vMotion network and RTT latency in the VPLEX Metro IP network.

During the VM power-on, there was an exchange of a few RPC calls between the source and destination hosts over the vMotion network . Therefore, the RTT latency on the vMotion network resulted in higher RPC call overhead and hence contributed to an increase in the switch-over time. Similarly, the RTT latency on the VPLEX Metro IP network resulted in longer disk open times during the VM power-on and hence contributed to an increase in the switch-over time. Our findings show that the number of RPC exchanges is fixed, so the RPC overhead is only impacted by the RTT latency and not by the number of virtual disks. In comparison, the disk-open overhead is impacted by both the RTT latency and the number of virtual disks. Therefore, the disk open overhead can become a dominant factor in configurations with a large number of virtual disks.

It should be noted that these overheads observed in metro deployments are the same, regardless of the load on the VM. For instance, as shown in Figure 5, the switch-over durations seen in all the 12-disk idle VM tests were nearly identical to the switch-over durations seen in the 12-disk active VM tests.

Figure 6 below shows the impact of the number of snapshots and the RTT latency on the VM switch-over time.

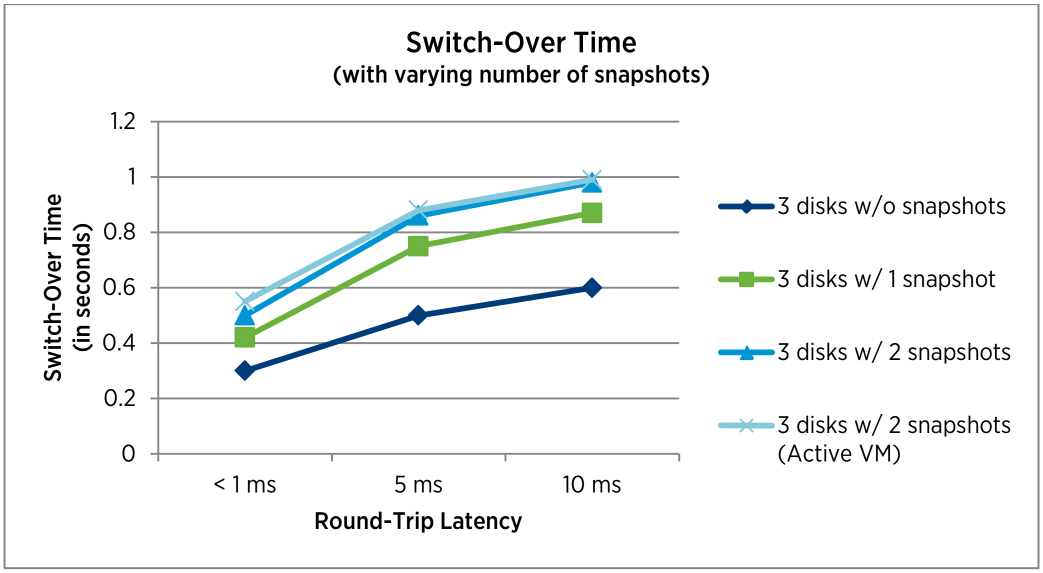

Figure 6. Virtual machine switch-over time in metro scenarios with varying numbers of snapshots

It is clear from the above figure that snapshots severely impact switch-over duration on metro deployments. In addition to the RPC and disk open overheads discussed earlier, snapshots lead to longer disk lock handoff during VM switch-over. For instance, in the “3 virtual disks w/ 2 snapshots” test configuration, the switch-over duration increased from about 0.5 seconds in baseline scenario (<1ms RTT latency) to about 0.98 seconds in the 10ms RTT latency scenario. Note that these overheads remain the same, regardless of the load on the VM. For instance, as shown in Figure 6, the switch-over durations seen in all the “3 disks w/ 2 snapshots (Idle VM)” tests were nearly identical to the switch-over durations seen in the “3 disks w/ 2 snapshots (Active VM)” tests.

Our findings clearly show that the presence of a large number of virtual disks and snapshots can have adverse impact on the switch-over duation in metro deployments. It is recommended to configure the VMs with only the required number of virtual disks and minimize the usage of snapshots to improve performance in metro deployments.

Best Practices

We recommend the following vMotion best practices for VPLEX Metro deployments These recommendations are equally applicable to any metro deployment solution that uses an active-active replicated storage topology.

- **Upgrade to vSphere 5.5.**

  The vSphere 5.5 release incorporates a number of vMotion performance enhancements for metro deployments.

- **Provision network interface cards (NICs) that support hardware TCP segmentation offload (TSO) for the vMotion network.**

  TSO is a technique of queuing up large buffers and letting the NIC split them into separate network packets, thus reducing the number of packets that need to be processed in the vSphere network stack and saving CPU. TSO serves well in metro latency environments where periods of high activity can overwhelm network queues, resulting in potential packet drops. Using NICs that support hardware TSO enables vMotion to maximize the network bandwidth usage on long latency networks even at default network queue depths.

- **Configure VMs with only the required number of virtual disks.**

  The presence of large numbers of virtual disks results in longer disk open times during the VM switch-over phase on metro deployments. Configure VMs with only the required number of virtual disks to improve switch-over performance in metro deployments.

- **Minimize the number of snapshots.**

  Snapshots are extremely helpful in protecting VMs against any unknown effects when patching operating systems or applications in the VM. However, it is not recommended to run production VMs off of snapshots on a permanent basis. Usage of snapshots results in higher I/O response times and lower application throughput. In addition, snapshots lead to longer disk lock handoff during VM switch-over. Be cognizant of the fact that the performance impact of these overheads is more pronounced on metro networks, and accordingly reduce or eliminate the use of snapshots on metro deployments.

- **Optimize VM switch-over time on metro networks.**

  Add the configuration parameter (`extension.convertonew = "FALSE")` to the virtual machine settings (VMX) through the vSphere Client interface. This option optimizes the opening of virtual disks during VM power-on time and thereby reduces switch-over time during vMotion. This option was found to be helpful in reducing the switch-over time on VPLEX Metro deployments when the VM is configured with a large number of disks or snapshots.

- **Ensure VPLEX best practices have been followed.**

  Ensure the optimal operation of VPLEX by adhering to the VPLEX configuration best practices. This topic is beyond the scope of this white paper.

Conclusion

As the virtualization landscape continues to evolve with new technological capabilities and an ever increasing number of virtualized workloads, VMware is continually innovating core vMotion functionality as well as leveraging external technologies to enable administrators to more effectively and efficiently manage their virtual environments.

In the vSphere 5.5 release, vMotion incorporates new enhancements, which include performance optimizations that enable seamless migration over EMC VPLEX Metro deployments without the need to transfer the virtual machine disk contents across geographically separated data centers during the migration. The performance results presented in this paper show VMware vMotion in vSphere 5.5 paired with EMC VPLEX Metro can provide workload federation over a metro distance by enabling administrators to dynamically distribute and balance the workloads non-disruptively across data centers.

About the Author

Sreekanth Setty is a staff member with the performance engineering team at VMware. His work focuses on investigating and improving the performance of VMware's virtualization stack, most recently in the live-migration area. He has published his findings in a number of white papers and has presented them at various technical conferences. He has a master's degree in computer science from the University of Texas, Austin.

Acknowledgements

The author would like to sincerely thank Ricardo Koller and Gabriel Tarasuk-Levin for their reviews and helpful suggestions to this paper. He also would like to extend thanks to his manager Bruce Herndon and other colleagues on his team.

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