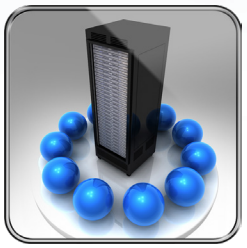
SOLUTION BRIEF

Accelerating Virtual Machine Migration over vSphere vMotion and Mellanox End-to-End 40GbE Interconnect Solutions



Introduction

Large virtualized environments, including those found in cloud infrastructures, require high I/O performance for VM-to-VM communication and for hypervisor services, such as live VM migration. With higher performance and more efficient networking, data center managers can migrate VMs to different physical servers faster, allowing them to meet stringent service level agreements (SLA) at a lower total cost of ownership (TCO). Faster live VM migration minimizes the server's "inactive" time and thus enables more jobs per second to run, which maximizes the virtualized infrastructure efficacy.

Mellanox end-to-end 40GbE solutions, including the SX1012 40GbE switch, the ConnectX-4 Lx NIC, and LinkX cables, enable the migration of a single large VM (with a large amount of memory) or a large set of VMs simultaneously in record time.

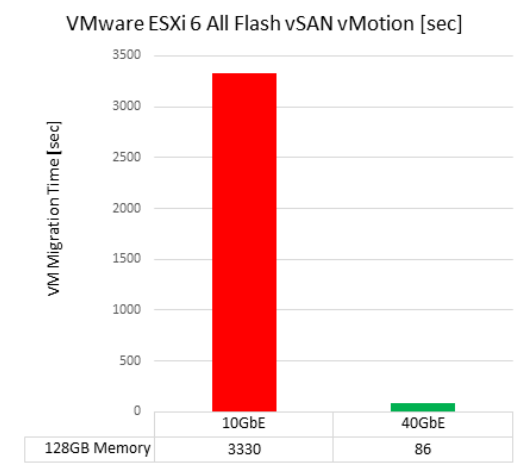


Figure 1. *38X vMotion acceleration over Linux guest OS running the VM with Microsoft FileIO workload*

The results show that running vSphere 6.0 vMotion over 40GbE vs. 10GbE accelerates the migration by 38 times when running over Linux guest OS and by greater than 5 times when running the VM over Windows guest OS. Both cases significantly boost the efficiency of the virtualized infrastructure.

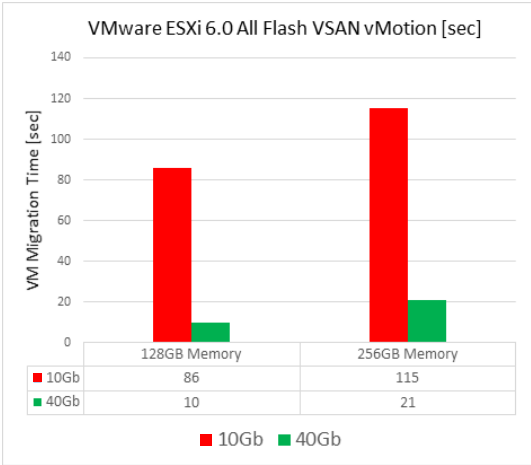


Figure 2. *5.5X vMotion acceleration over Windows guest OS running the VM with IOmeter workload*

Test Setup

The goal was to compare live migration of one VM between two ESX 6.0 servers, once over 40GbE versus over 10GbE.

The test setup consisted of four servers equipped with VMware's ESX 6.0 hypervisors, Mellanox inbox 10/40GbE driver (ESX 6.0 servers), and ConnectX-4 Lx 10/25/40/50GbE NIC. The ESX 6.0 servers were connected over two SwitchX-2 based SX1012 40GigE switches and LinkX copper cables. A fifth server ran VMware vCenter on top of Microsoft Windows Server 2012 R2 64-bit. VSAN 6.0 All Flash was used for storage, using 1 x PCIe 800GB SSD and 6 x 800GB SSD included in each of the four ESX 6.0 servers.

The ESX 6.0 servers had two memory configurations:

1. 384GB total memory with 1 VM configured as 128GB
2. 384GB total memory with 1 VM configured as 256GB

For each of these memory configurations, a vMotion test was performed, in which the entire memory was allocated to the VM. In real life, this case imitates, for example, a situation in which real-time received data records are written continuously into memory by the active VM.

Each test performed several iterations in order to verify the stability of the results.

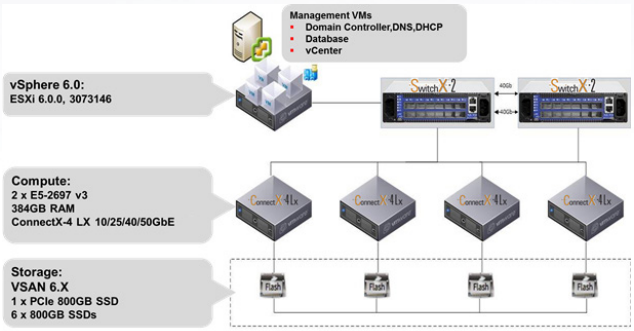


Figure 3. *The block diagram of the setup*

The Benchmark & Results

Since VSAN was used for storage, only the VM memory had to be migrated. In order to perform the job, vMotion had to execute the following tasks[1]:

1. Create a Shadow VM on the destination host.
2. Copy each memory page from the source server to the destination server over the vMotion network. This phase is known as preCopy.
3. Perform another pass over the VM's memory, copying any pages that changed (or became "dirty") during the previous preCopy iteration.
4. Continue this iterative memory copying until there are no changed pages (outstanding pages to be copied) remaining.
5. Stun the VM on the source and resume it on the destination.

This procedure works very well when each preCopy iteration takes less time to complete than the previous preCopy iteration. In such a case, the VM live migration will be converged. However, should the active VM modifications (or "dirtying") consist of a larger amount of memory than the amount that is being migrated, the migration process will not be converged. To avoid this, vMotion uses the Stun During Page-Send (SDPS) operation, which slows down the VM operation. Activating the SDPS will ensure that the memory modification rate is slower than the preCopy transfer rate and guarantees the convergence of the live migration process[2].

Our benchmark shows that when comparing live migration of an active VM that keeps dirtying the memory over different speed networks, the difference in migration time is higher than the proportional ratio between the network speeds. This is demonstrated very well in Figure 1 and Figure 2, in which the acceleration over 40GbE was 38 times faster than over 10GbE running over the Linux guest OS, and 5.5 times faster running over the Windows guest OS, respectively (both with active SDPS).

Summary

The performance results that have been achieved over Mellanox end-to-end 40GbE solutions show significant migration time acceleration, achieved due to higher I/O bandwidth delivery in comparison to a 10GbE. It is expected that even higher efficiency will be achieved over 50GbE and 100GbE using ConnectX-4 Lx 50GbE and ConnectX-4 100GbE respectively and connected over Mellanox Spectrum 10/25/40/50/100GbE switch and LinkX cables.

Increasing I/O demands in the cloud and Web 2.0 hyperscale computing can be overcome by using Mellanox end-to-end higher performance interconnect solutions, which improve the speed at which cloud service providers can provision new users with new VM and application requirements, and can meet their SLAs without impacting service to existing users. The paradigm changes from bulk live migration of a VM across physical servers for disaster recovery to scheduled migration for maintenance or load balancing, thereby enabling higher SLA achievements due to the faster migration, which improves the infrastructure efficiency and maximizes the ROI.

References:

[1] http://blogs.vmware.com/vsphere/2011/02/vmotion-whats-going-on-under-the-covers.html

[2] http://sparrowangelstechnology.blogspot.com/2012/11/vsphere-51-vmotion-best-practices.html

Mellanox TECHNOLOGIES
350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com