Demartek

Evaluation of HPE 25GbE RDMA Benefits in Virtualized Environments

Exploring the benefits of switching from 10GbE to 25GbE with RDMA: SMB and RDMA over 25GbE work together to reduce live virtual machine migration times in Hyper-V.

Executive Summary

Virtualization and 10GbE network connections have become the norm in today's datacenter; however, 10GbE no longer provides the necessary bandwidth to support today's growth in certain virtualization environments. Therefore, businesses need to upgrade from 10GbE to 25GbE with Remote Direct Memory Access (RDMA). The days of over-provisioning are gone. New virtual machines (VMs) are added regularly to servers resulting in fewer processor resources available to process network packets through the TCP/IP stack. Furthermore, the VMs often drive aggregate demand for bandwidth to levels higher than 10GbE can provide. This can be observed when a database application such as Microsoft SQL Server or Oracle is running in a VM alongside other virtualized production workloads. Hyperconverged Infrastructure (HCI) and Storage Spaces Direct (S2D) clusters can also exhibit this condition.

HPE commissioned Demartek to explore the benefits of

> **Upgrading from 10GbE to 25GbE technology**

> **Utilizing RDMA in virtualized environments**

First, we will explore where 10GbE is today and what 25GbE with RDMA has to offer. Then we will show an example of how 25GbE with RDMA improves performance, resulting in dramatic time-savings during live VM migrations.

Key Findings

25GbE Benefits

> ***Faster*** – 25GbE increases the clock rate, enabling 2.5 times the amount of data to be delivered in a single lane.

> ***Density*** – 25GbE uses pre-existing IEEE standards, increasing switch density while using the same fiber-optic cables as 10GbE.

> ***Efficient*** – 25GbE has a lower cost per unit bandwidth than 10GbE.

Live Migrations with 25GbE RDMA

> RDMA usage in a real-world 25GbE high CPU utilization environment produced a ***60% reduction in time to complete*** live migrations as compared to non-RDMA connections.

> 25GbE usage in a real-world high CPU utilization RDMA environment produced a ***30% reduction in time to complete over 10GbE***.

© 2018 Demartek

Evaluation of HPE 25GbE RDMA Benefits in Virtualized Environments

25GbE with RDMA and the Benefits over 10GbE

Today's 10GbE

10GbE switch and adapter ports operate at a clock rate of 10.31 GHz. For years, this has been the fastest single-lane clock rate available for Ethernet. To achieve higher rates, such as 40GbE and 100GbE, multiple lanes of 10GbE have to be run in parallel. For example, 40GbE is composed of four 10GbE lanes using a Quad Small Form-factor Pluggable (QSFP) cable and connector, and 100GbE is composed of ten lanes of 10GbE using a specialized cable and connector.

The throughput increases from these multi-lane, parallel connections results in increasing complexity of cables, connectors, and setup. As shown in **Figure 1**, each lane of a 40GbE connection can be configured individually with a breakout cable used to connect to four separate 10GbE ports. Alternatively, the 40GbE port could be configured as one port and a single cable used to connect to another 40GbE port. Multi-lane cables are more expensive, and extra knowledge is needed to connect and set up higher speeds correctly.

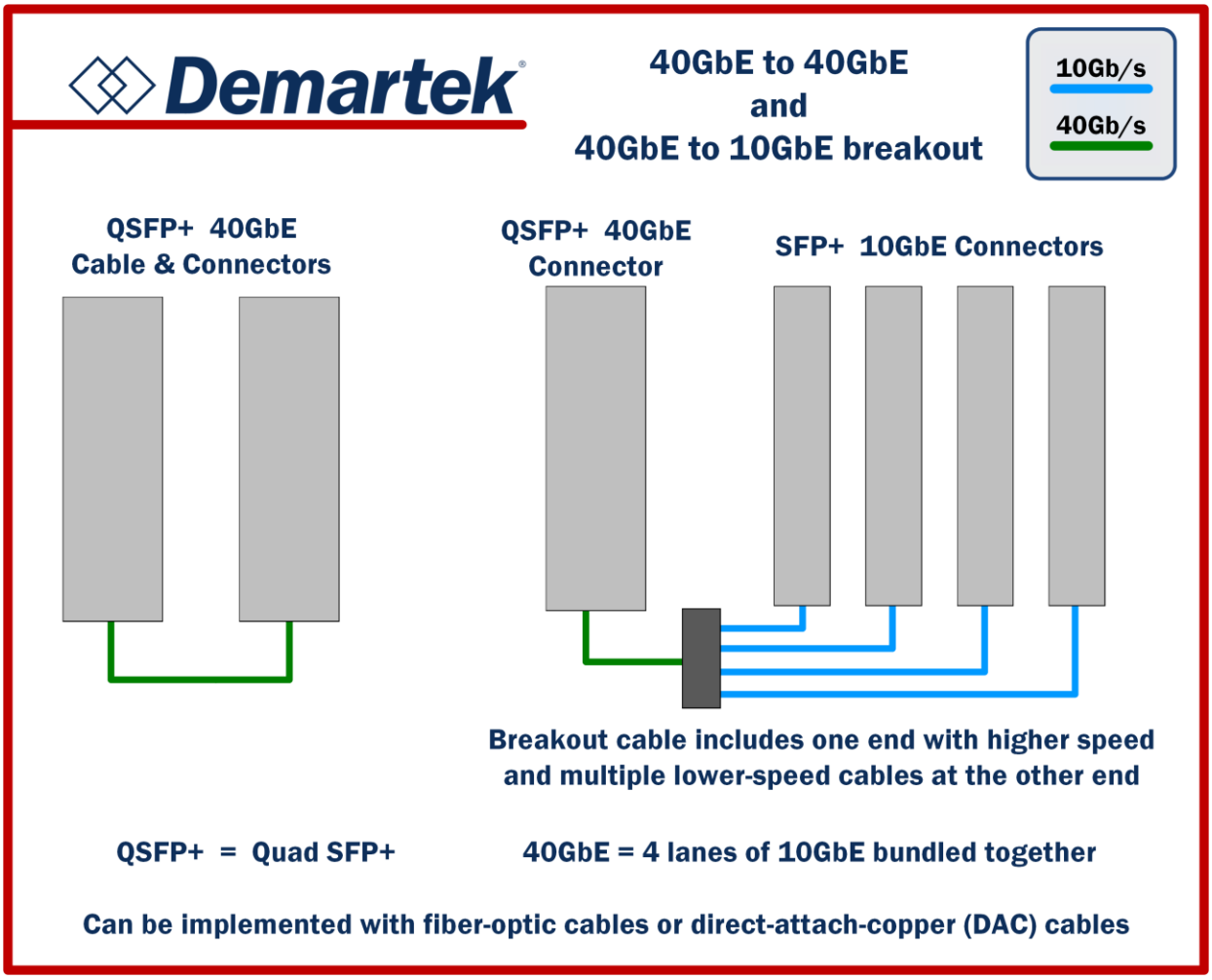

***Figure 1** – 40GbE to 40GbE and 40GbE to 10GbE Breakout Cables and Connectors*

© 2018 Demartek

Benefits of 25GbE

25GbE operates at a single-lane clock rate of 25.78 GHz, which delivers 2.5 times the data throughput rate of 10GbE. The extra throughput is obtained through the clock speed, not by adding extra lanes. Using 25GbE technology, higher speeds such as 50GbE and 100GbE can be achieved by running either two or four of these lanes together in parallel. 50GbE has more throughput than 40GbE with half the number of lanes. With four lanes of 25GbE, 100GbE can be achieved using QSFP connectors compared to 40GbE with four lanes of 10GbE technology.

Using fewer lanes for high-throughput connections increases switch density in the larger ecosystem. More high-throughput ports can be serviced with the same number of lanes. In addition, a single 25GbE port can replace two 10GbE ports that a server was previously using, reducing the number of ports required for each server.

It has been estimated that using 25GbE instead of 40GbE and 10GbE connections will result in a lower cost per unit bandwidth. While the per connection cost might be higher for 25GbE, the effective cost is less in terms of price and power consumption to achieve 25Gbps of bandwidth than 10GbE technology provides. It is also estimated that in the future, 25GbE prices will drop to be the same as 10GbE.

While 25GbE is relatively new, the adoption rate is expected to be much faster than the adoption rate for 10GbE. 25GbE technology uses 25GbE SFPs, known as SFP28, and these are backwards compatible with 10GbE SFPs.

Cabling for 25GbE

There are differences in cabling between 10GbE technology and 25GbE technology. For a given cable type, the distance supported by 25GbE is generally less than the distance supported by 10GbE technology.

For multi-mode fiber-optic cables, the Telecommunications Industry Association (TIA) Engineering Committee TR-42 states that OM3 is the minimum requirement and OM4 is recommended.

For passive direct attached copper (DAC), the lengths supported for 25GbE are less than for 10GbE. Furthermore, DAC cables require different transceivers for 25GbE technology.

The cable distances supported are described in Figure 2 below.

Fiber-optic	OM1	OM2	OM3	OM4
10GbE	33m	82m	300m	400m
25GbE	–	20m	70m	100m

Passive DAC	Distance
10GbE	Up to 7m
25GbE	Up to 5m

Figure 2 *– Cable Distances for 10GbE & 25GbE*

© 2018 Demartek

Evaluation of HPE 25GbE RDMA Benefits in Virtualized Environments

Benefits of RDMA

RDMA is the remote memory management capability that allows server-to-server data movement directly between the application memory of each without any CPU involvement. RDMA bypasses the normal system software network stack components and the multiple buffer copy operations that are normally performed when TCP/IP is used for typical network traffic.

This elimination of buffer copy operations reduces overall CPU consumption and improves the latency of the host software stack since it uses fewer instructions to complete a data transfer. **Figure 3** below illustrates the difference between the traditional TCP/IP stack on the left and the RDMA stack on the right.

Today's RDMA Network Interface Cards (RNICs) will usually support at least one, and in some cases two, implementations of RDMA. One implementation is RDMA over Converged Ethernet (RoCE). The other is Internet Wide-area RDMA Protocol (iWARP). Either protocol can be used by Microsoft SMB Direct, a file transfer protocol that supports the use of RDMA that has been available since Windows Server 2012 debuted. Microsoft Hyper-V, by providing support for SMB Direct, is the first hypervisor that enables live migrations using an RDMA-enabled protocol. We expect other hypervisors to follow suit in the near future, leveraging the performance benefits of RDMA in a virtualized environment.

The Need for 25GbE with RDMA in Today's Datacenter

In the past, putting one or two 10GbE ports on a server with a couple of applications would ensure that there was more network bandwidth available than the server would need. That is no longer the case. With today's highly virtualized environments, we have many virtual machines sharing the same network ports. Each virtual machine might be hosting multiple services. Bandwidth requirements can easily exceed what 10GbE can provide. A standard 10GbE link will not be sufficient, and without RDMA, it may be difficult to achieve full 25GbE throughput. We recommend that RDMA be enabled as the default technology for highly virtualized environments.

With the current environments, every resource, not just networking, is optimized and fully utilized, including the processor. With higher-throughput networking connections, a higher demand is placed on an already taxed processor. Utilization of RDMA is necessary in order to negate this effect and leave the processor resources with the VMs and hypervisor, instead of investing more CPU into the networking.

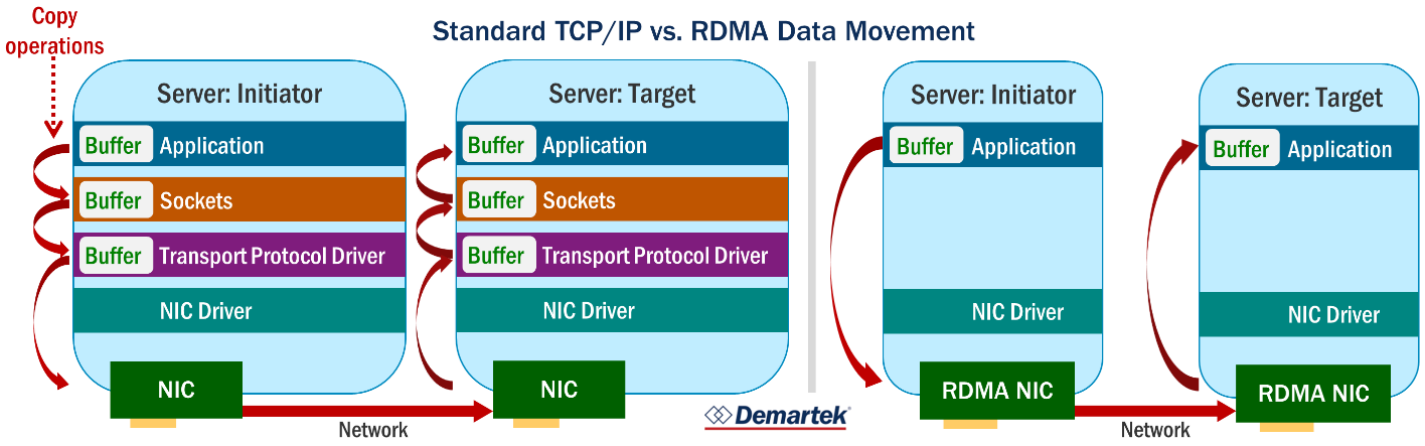

***Figure 3** – Data Movement for the Standard TCP/IP stack vs. the RDMA stack*

© 2018 Demartek

Configuration for 25GbE with RDMA

Server – 2x HPE ProLiant DL380 Gen10

> 2x Intel Xeon Gold 6142 2.6GHz, 32 total cores, 64 total threads
> HPE Ethernet 10/25Gb 2-port 622FLR-SFP28 Converged Network Adapter
> 128 GB RAM, DDR4, 2400 MHz
> Windows Server 2016
> Iometer 1.1.0

Ethernet Switch

> Mellanox SN2410 25/100GbE

Test Environment

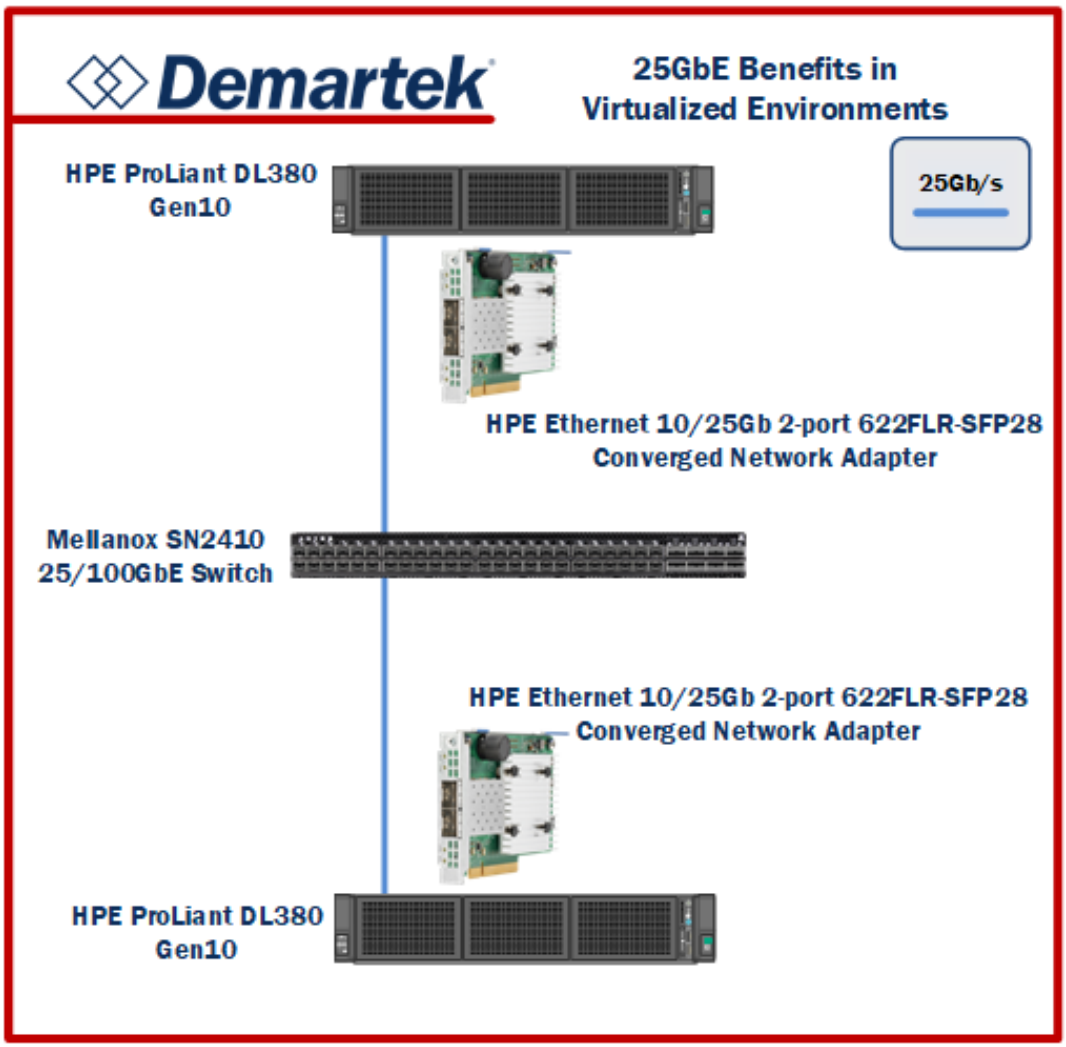

***Figure 4** – Configuration Diagram for 25GbE*

© 2018 Demartek

The Base Environment for SSD (Low CPU Utilization Tests)

One Windows Server 2016 VM was created on an SSD on each server. One was a fileserver while the other was a file client that connected to the fileserver and ran a file client emulation workload using Iometer. This was to designed provide some baseline VM traffic to run during our migrations. The workload was large block, 80% read, and generated approximately 160MBPS of throughput before migrations started.

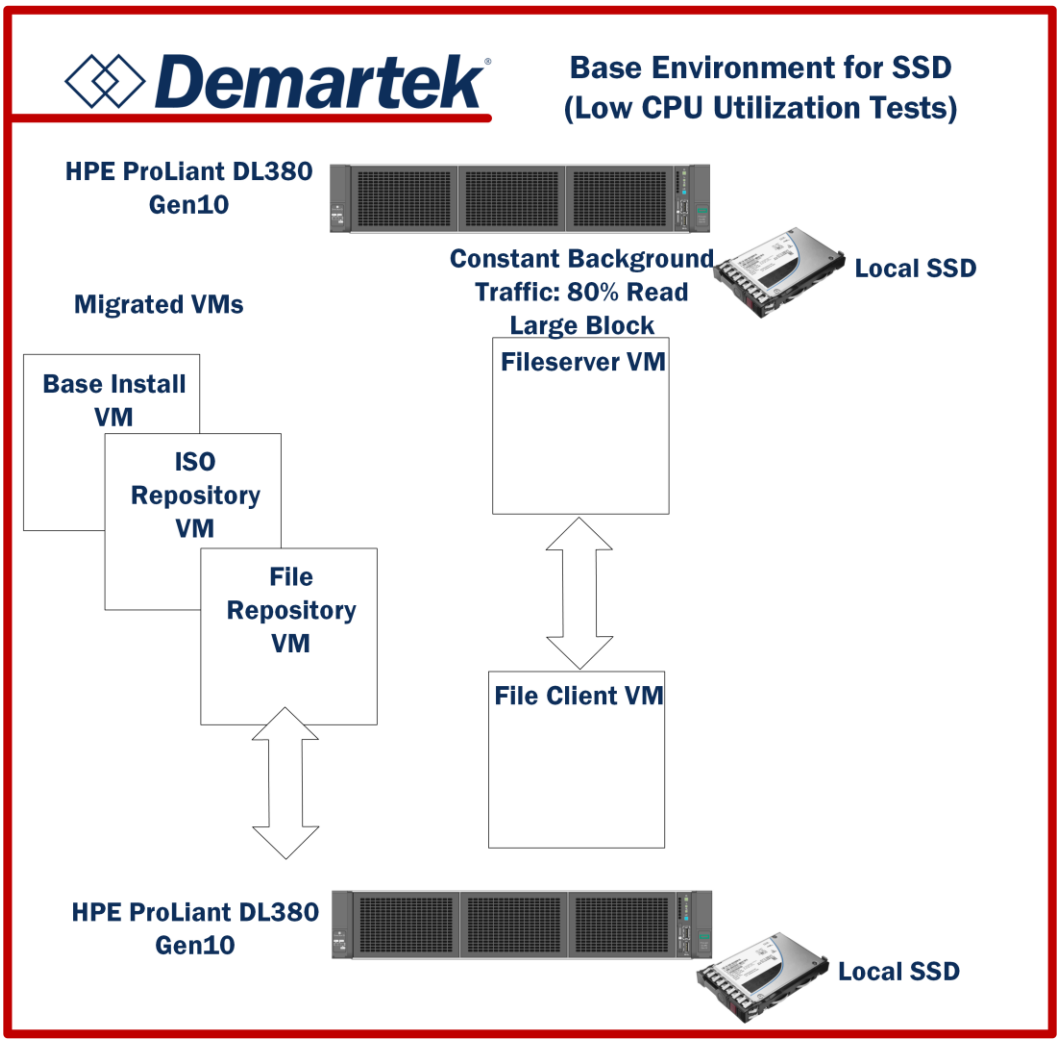

Figure 5 *– Base Environment for SSD (Low CPU Utilization Tests)*

© 2018 Demartek

Evaluation of HPE 25GbE RDMA Benefits in Virtualized Environments

The Base Environment for NVMe (High CPU Utilization Tests)

Four Windows Server 2016 VMs were created on a striped NVMe storage pool on each server. Two were made into a fileserver, and two were made into a file client. Iometer was used to run a 100% read small block workload between all file clients and their corresponding fileservers on the opposite server. The Sysinternals group at Microsoft provides an application known as CPUSTRES.EXE that simulates high CPU usage by a user mode process. CPUSTRES.EXE was used on all fileservers and file clients to create more CPU utilization. Total CPU utilization on both hosts with this setup was approximately 87%.

Test Setup and VM Configuration

A domain was created and both servers were added to the domain to enable VM migration between the servers. Three Windows Server 2016 VMs were created, and each was allocated a VHDX (virtual hard disk) with 127GB of available space. The VMs resided on the local NVMe Storage Pool. With just the OS installed, the VHDX was only 11.4GB. Files were added to one VM to make it an ISO repository, bringing the size to 109GB. The other VM had fileserver files added, bringing the size to 89.1 GB. These three VMs were migrated repeatedly between the machines during testing. For each test, the base install was migrated to validate the test setup. Then the two larger VMs were migrated simultaneously. The start and end time of the migrations were logged using a script that provided a date stamp, executed the PowerShell commands to perform the migration, and then provided a date stamp once the PowerShell commands were complete.

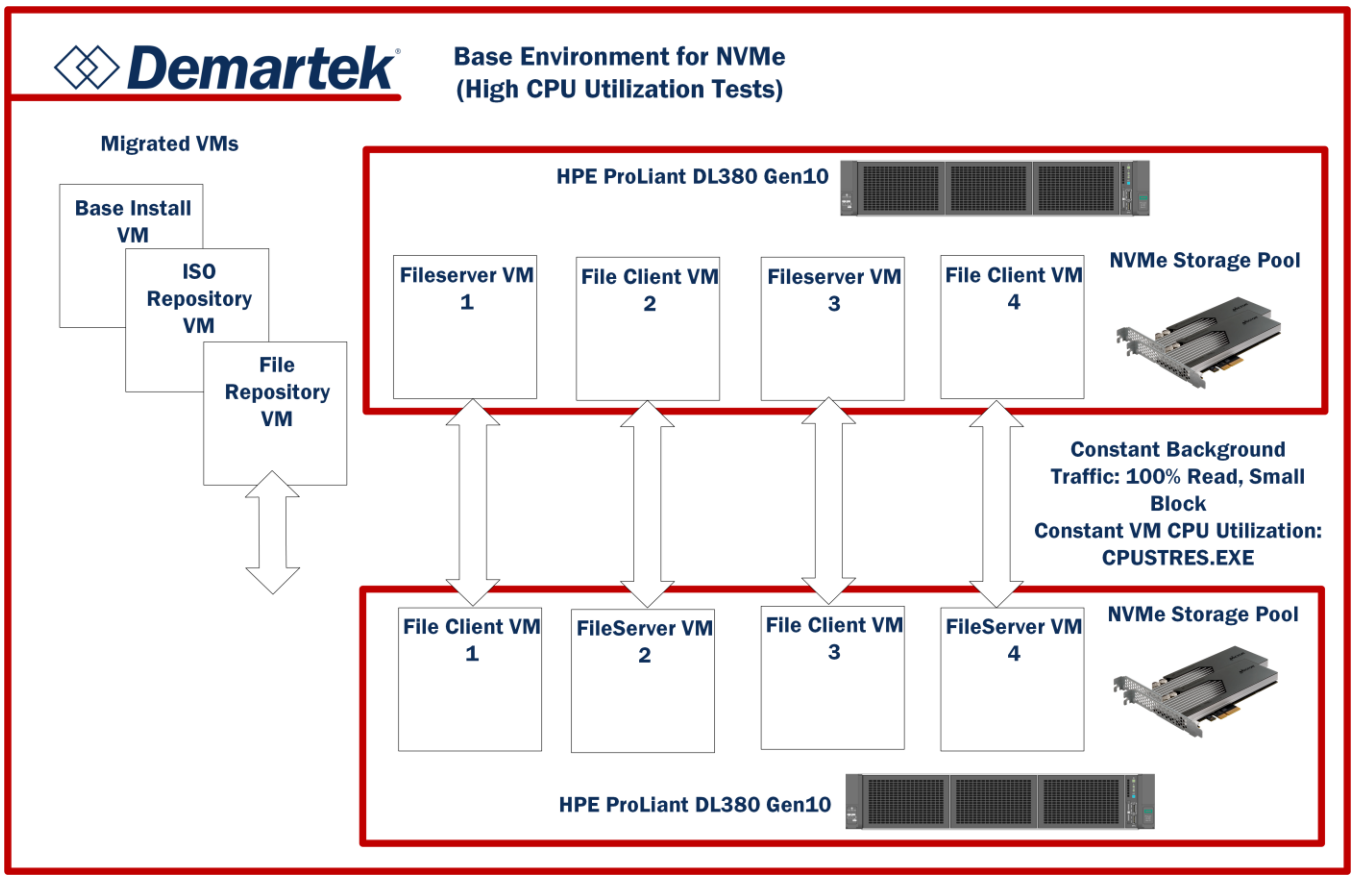

Figure 6 *– Base Environment for NVMe (High CPU Utilization Tests)*

© 2018 Demartek

Test Results with SSD Storage (Low CPU Utilization / Storage Bottleneck)

Tests were completed using 25GbE technology, then the speed was changed to 10GbE and the tests were repeated. Six tests were completed for each configuration and the results were averaged. The lack of difference in time to complete between 10GbE and 25GbE as shown in Figures 7, 8, and 9 is most likely due to a bottleneck in the local SSD used to store the virtual hard drives rather than the limitations of the 10GbE adapter. The lack of difference between RDMA verses no RDMA connections as shown in Figures 7, 8, and 9 is due to the server CPU not being taxed enough for the reduced CPU overhead of RDMA to be needed. This underscores how RDMA and higher network speeds typically only show benefits in highly optimized servers with a higher percentage of CPU utilization.

It should be mentioned that when the VMs were being migrated from the server hosting the file client, the time to complete was consistently longer than the time to complete when the VMs were being migrated back from the fileserver. See Figures 8 and 9 for comparison.

The background workload caused the increased time to complete. Six more tests were completed without background workload and we found that the times to complete did not vary with direction of migration when this factor was eliminated.

This data illustrates an important point. With the two other VMs on the system utilizing very little CPU and with a storage bottleneck in the SSD, insufficient benefits were derived from the 25GbE RDMA configuration.

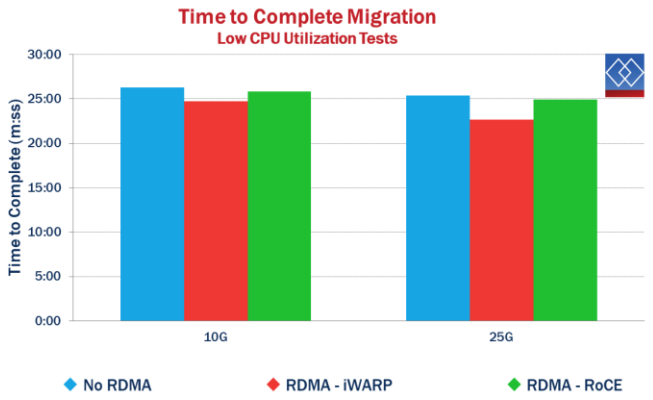

***Figure 7** – Time to Complete Migration, Low CPU Utilization Tests*

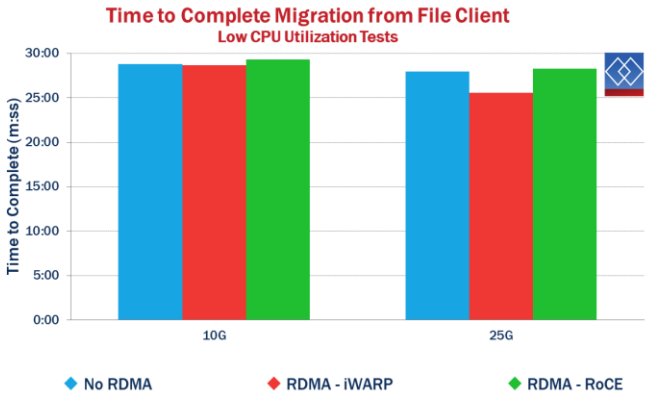

***Figure 8** – Time to Complete Migration from File Client, Low CPU*

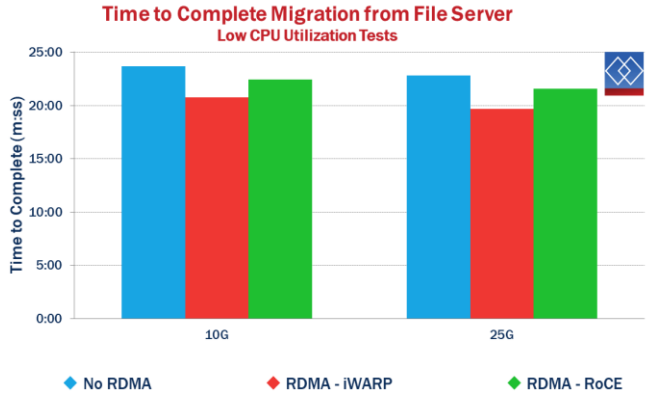

***Figure 9** – Time to Complete Migration from File Server, Low CPU*

© 2018 Demartek

Test Results with NVMe (High CPU Utilization Tests)

Six tests were completed for each configuration and the results were averaged. As each server had an equal number of background VMs – two fileservers and two file clients – the time to complete did not vary with migration direction as it did in the previous tests. Using an NVMe pool instead of an SSD, should remove the virtual hard drive storages bottleneck.

We can see in Figure 10 below that with significant CPU utilization by the VMs and no storage bottleneck, the advantage of RDMA becomes apparent. This CPU-taxed setup with virtualized environment is typical of what is found in today's datacenter. With RDMA, the performance improvement with 25GbE also becomes apparent.

Using 10GbE technology, the performance gains were strong with RDMA, with an approximately 47% reduction in the time to complete over the non-RDMA configuration. However, with 25GbE technology, the RDMA technology yielded an approximately ***60% reduction in the time to complete over non-RDMA*** configurations.

Comparing 10GbE technology with RDMA to 25GbE technology with RDMA, the 25GbE RDMA completed the work approximately ***30% faster than 10GbE with RDMA***.

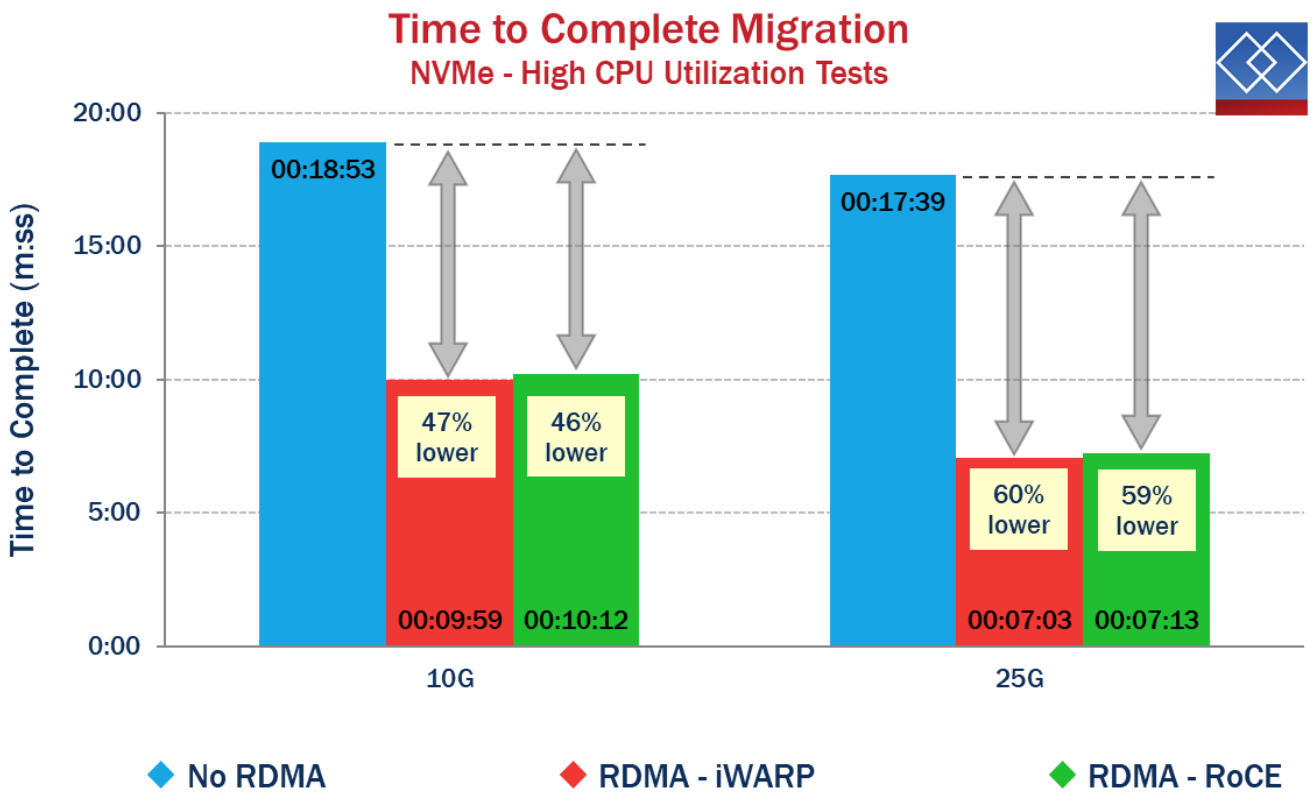

Figure 10 *– Time to Complete for NVMe, High CPU Utilization Tests*

© 2018 Demartek

Evaluation of HPE 25GbE RDMA Benefits in Virtualized Environments

Demartek
March 2018

Summary and Conclusion

In a server where resources are under-utilized, 25GbE and RDMA will show little performance benefit. In a virtualized environment where the majority of server resources are utilized, however, RDMA allows heavy network use without incurring a processor penalty and enables the server to more fully utilize the performance available from higher-speed connections.

In our heavily-utilized systems, 25GbE and RDMA yielded a ***60% reduction in time to complete*** the VM migrations over the non-RDMA configuration. This was also approximately ***30% faster than the 10GbE technology*** with RDMA enabled.

As system architects design servers with more VM density and higher performing NVMe storage, 25GbE with RDMA will soon replace 10GbE as the standard. Designing solutions today with RDMA-enabled 25GbE adapters will allow support for increased VM densities and for software designed storage implementations that can utilize high performance NVMe storage.

The most current version of this report is available at http://www.demartek.com/Demartek_HPE_25GbE_RDMA_Evaluation_2018-03.html on the Demartek website.

HPE is a registered trademark of HPE Corporation and/or its affiliates in the United States, certain other countries and/or the EU.

Demartek is a registered trademark of Demartek, LLC.

All other trademarks are the property of their respective owners.

© 2018 Demartek