WHITE PAPER

October 2013

ConnectX®-3 Pro: Solving the NVGRE Performance Challenge

Objective

This white paper presents an overview of the rising need for overlay virtualized networks and the solution that the NVGRE technology provides to meet this need. It also examines the various performance challenges associated with NVGRE and presents Mellanox's new solution for resolving these challenges.

Background: The Need for Virtualized Overlay Networks

Over the past few years, cloud computing has grown at a tremendous rate, with worldwide spending on IT cloud services tripling since 2008 and expected to reach over $100 billion in 2014.[1] More specifically, Infrastructure as a Service (IaaS) is the fastest-growing segment of the public cloud services market, having grown over 45% in 2012 alone.[2]

IaaS allows multiple tenants to share system resources and infrastructure, which improves hardware utilization, thereby reducing the cost of the IT infrastructure, both at implementation and ongoing. Cloud computing also provides a measure of agility that simplifies the IT management process, provides additional control over proprietary data, and improves the end-user experience.

The IaaS segment of cloud computing is based on the concept of multiple tenants sharing the cloud infrastructure, enabled primarily by server virtualization. IaaS offers the following additional benefits to its consumers:

- Allows a company's IT department to focus on core competencies instead of assembling and maintaining network infrastructure
- Enables dynamic scaling of infrastructure services based on usage demands
- Reduces upfront investment costs, and limits ongoing costs to only OPEX
- Provides access to the infrastructure from anyhere in the world

©2013 Mellanox Technologies. All rights reserved.

IaaS providers need to support multiple tenancies on their datacenter infrastructure, creating the need to isolate each tenant within the network to provide the security and traffic isolation levels that independent infrastructure provides. This has typically been achieved through the use of VLANs, which segment the network into virtual network entities to provide security and network traffic control.

As the demand for cloud services continues to grow, many large consumers have outgrown the most basic solutions. For example, VLAN usage is limited to 4,096 entities (VLAN IDs), which, given the tremendous growth in cloud-based networks, is far from sufficient segmentation. A more scalable solution has become a necessity.

A number of solutions exist to address these issues, each with its own advantages and disadvantages. Technologies such as multiple VLANs (known as Q-in-Q) or MAC-in-MAC try to address the scalability concern by multiplying the number of possible VLAN IDs or MAC addresses that are used. Such technologies, however, remain in the Layer 2 (L2) domain, which means there are inherent challenges in scaling to a high number of entities. For example, Spanning Tree (STP), Rapid Spanning Tree (RTSP), or Multiple Spanning Tree (MSTP) are typically used to prevent loops, but half the connections in the "tree" are reserved (stand-by) for link failures. As a result, the network structure remains cumbersome.

An additional challenge in using new VLAN or MAC addresses stems from the need for configuration changes to the existing Layer-2 infrastructure, which complicates Virtual Machine (VM) and workload migration from server to server in the virtual domain.

As demonstrated in Figure 1, an approach that successfully addresses these challenges is to create a virtual network that transports data across existing Layer-3 (IP) infrastructure. A Layer-2 virtual overlay scheme builds a virtual L2 network on top of multiple L3 subnetworks using GRE tunnels between virtual machines (VMs), which operate in separate networks while operating as if they were attached to the same L2 subnet. By adding this L2 virtual overlay scheme on top of the L3 network, administrators are able to scale their cloud-based services without having to significantly reconfigure or add to existing infrastructure.

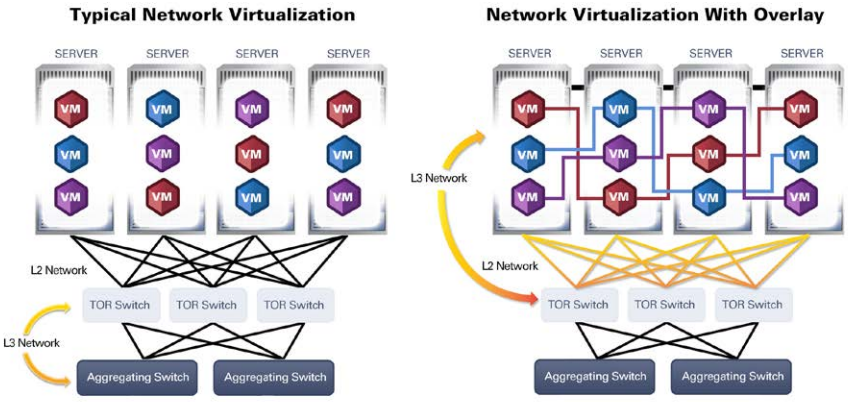

Figure 1. *Network Virtualization Before and After Overlay Network*

NVGRE Technology

In order for an overlay network to be of any use, a technology is required to encapsulate data such that it can tunnel into Layer 2 and be carried across Layer 3. One leading technology that provides this encapsulation and aims to resolve both the security and scalability issues is Network Virtualization using Generic Routing Encapsulation (NVGRE). NVGRE technology provides a solution for stretching the L2 network over the virtual L3 IP network.

©2013 Mellanox Technologies. All rights reserved.

Figure 2. *NVGRE Encapsulation*

The NVGRE concept is based on a new encapsulation for VM traffic in which a tunnel is created for the VM traffic using the GRE routing protocol. As Figure 2 shows, it encapsulates the VM's Layer 2 (Ethernet) traffic with new MAC and IP headers, and it adds an NVGRE header that includes a Tenant Network Identifier (TNI), a 24-bit identifier that radically extends the address space of the VLANs from 4,094 segments up to 16.7 million available IDs, solving the scalability issue.

The encapsulation consists of:

- An outer MAC address, which provides the physical destination and source addresses of the hypervisors or of intermediate L3 routers
- An optional 802.1q address to further delineate NVGRE traffic on the LAN
- Outer IP addresses, which are the IP addresses assigned to the hypervisors that are communicating over L3
- A GRE header that designates the TNI. Each TNI is associated with a specific GRE tunnel that identifies the tenant's unique virtual subnet within the cloud.

The encapsulation and decapsulation process is handled within the hypervisor, which connects the virtual machines with the IP network. The hypervisor, therefore, ensures that the virtual machines themselves are completely unaware of the GRE protocol that has been implemented to communicate between them.

Multicasting is yet another benefit of transporting messages via L3, as opposed to L2, which only offers broadcasting. The hypervisor determines whether the communicating virtual machines are within the same multicast group and thereby determines whether unicast or IP multicast is required. The hypervisor is able to differentiate between individual logical networks and to identify new virtual machines to be associated with multicast groups.

NVGRE is truly a hybrid solution, combining the benefits of L2, such as the ability to shift the location of VMs to maximize the efficiency of the datacenter, with the scalability realized by transporting via L3.

Considering its ability to address the challenges of VLAN grouping using a virtualization solution, an NVGRE solution is considered an ideal technology for network administrators working within a cloud computing environment.

NVGRE's Hidden Challenge

Although NVGRE solves scalability and security concerns, and although it does so at very little monetary expense, there are two major concerns that can impact IaaS performance:

1. In a traditional Layer-2 network, sizeable processing savings are realized by using CPU offloads. However, in an NVGRE setup, the classical offloading capabilities of the network interface controller (NIC), such as checksum offloading and large segmentation offloading (LSO), cannot be used because the inner packet is no longer accessible due to of the added layer of encapsulation. As such, additional CPU resources are required to perform tasks that would previously have been handled more efficiently by the existing NIC.
2. NVGRE restricts the ability to use Receive Side Scaling (RSS) to distribute traffic across multiple cores based on the inner packet, meaning that there is a severe reduction in bandwidth.

This unanticipated degradation in performance and increase in CPU overhead significantly offsets the many benefits of using NVGRE.

©2013 Mellanox Technologies. All rights reserved.

ConnectX-3 Pro Solves the Performance Challenge

While NVGRE offers significant benefits in terms of scalability and security of virtualized networks, it is not as valuable unless it maintains the availability of the CPU and a similar rate of throughput. However, because NVGRE uses an additional encapsulation layer, it requires a high level of CPU usage and prevents traditional offloads from reducing that workload. Throughput it also affected by this unfortunate and unintended consequence.

In order for NVGRE to be of real value, the extra CPU overhead it creates must be eliminated.

This can be achieved by supporting all existing hardware offloads in the network controllers. This includes:

- Allowing checksum to be performed on both the outer and inner headers
- Performing large segmentation offloading
- Handling Netqueue to ensure that virtual machine traffic is distributed between different CPU cores in the most efficient manner

Mellanox's ConnectX-3 Pro is the only NIC that currently handles the traditional offloads despite the extra encapsulation layer. Its impact on the CPU workload is dramatic. As Figure 3 shows, the CPU is freed up by as much as 73.9% when sending 64k messages, thanks to the addition of HW offloading performed by ConnectX-3 Pro.

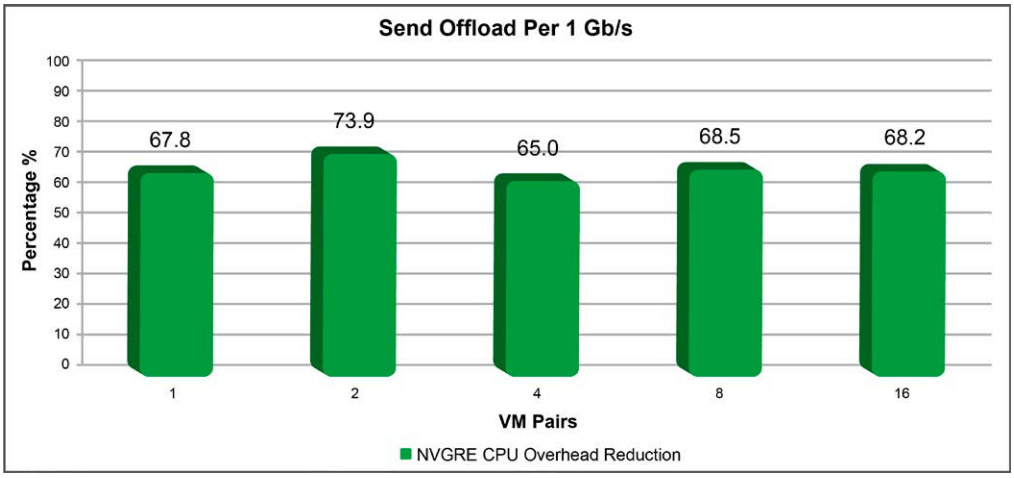

Figure 3. *NVGRE Send Offload*

While ConnectX-3 Pro's reduction of CPU overhead in receive operations is lower, it is not insignificant, lessening the strain by up to 26%, as seen in Figure 4.

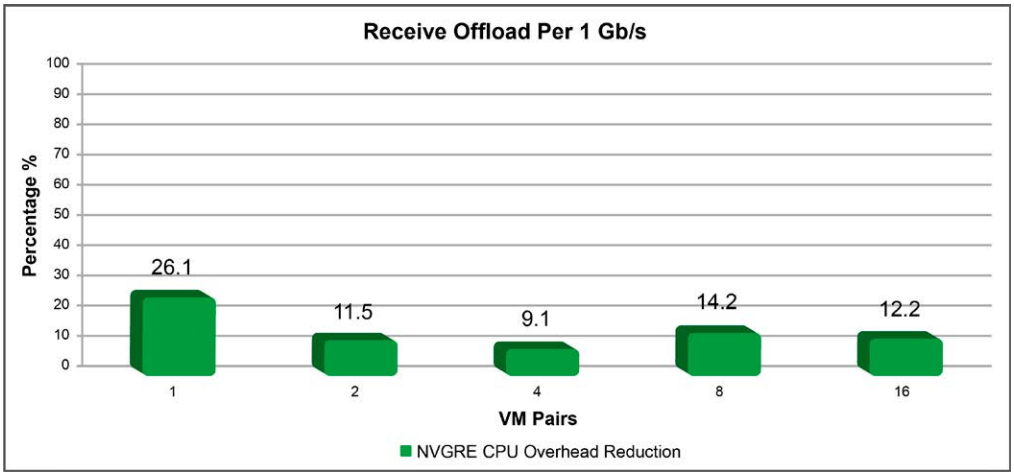

Figure 4. *NVGRE Receive Offload*

©2013 Mellanox Technologies. All rights reserved.

If one assumes that for every send operation there is a receive operation and vice versa, then it is fair to assume that a straight average (send + receive divided by 2) will provide the overall impact of ConnectX-3 Pro on CPU overhead. As Figure 5 illustrates, the average savings in CPU overhead is between 37-47% when using NVGRE with the ConnectX-3 Pro versus using NVGRE with other NICs.

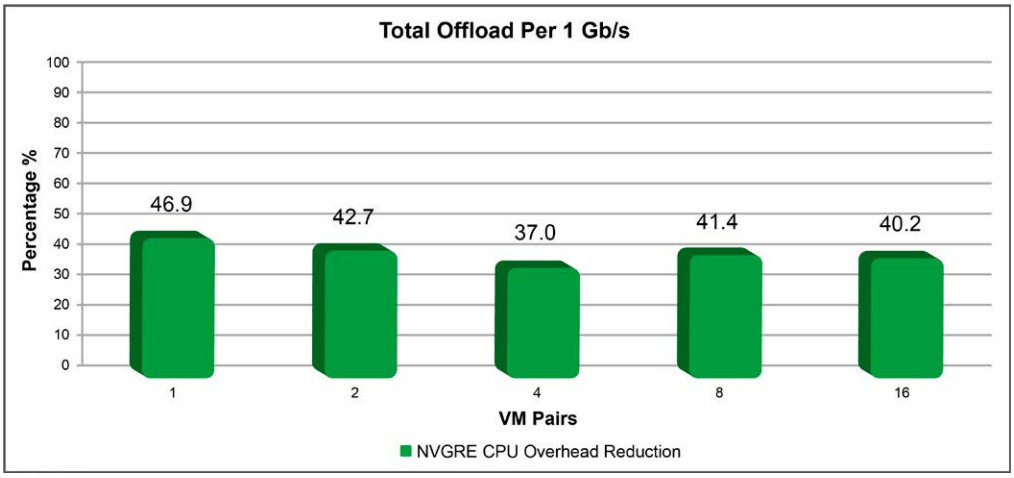

Figure 5. *NVGRE Average Total Offload*

Furthermore, ConnectX-3 Pro enables RSS to steer and distribute traffic based on the inner packet, and not , like other NICs, only the outer packet. The primary benefit of this is that it allows multiple cores to handle traffic, which then enables the interconnect to run at the intended line-rate bandwidth, not at reduced bandwidth as occurs when NVGRE is used without RSS.

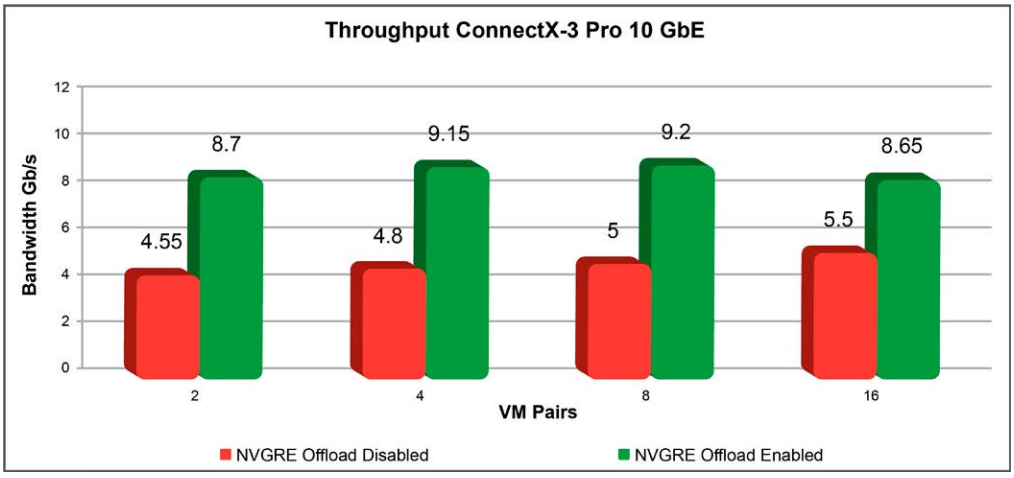

Figure 6. *Throughput With and Without NVGRE*

While other NICs see a dropoff of as much as 65% of the intended bandwidth when using NVGRE because of their inability to gain access to the inner packet for RSS to distribute traffic across multiple cores, ConnectX-3 Pro maintains near-line-rate throughput, as displayed in Figure 6.

ConnectX-3 Pro addresses the unintended performance degradation seen when using NVGRE in both CPU usage and throughput. By using ConnectX-3 Pro, the scalability and security benefits of NVGRE can be utilized without any decrease in performance and without a dramatic increase in CPU overhead.

©2013 Mellanox Technologies. All rights reserved.

Summary

Today's virtualized data centers suffer from a lack of scalability and security, restricting the ability to build a large-scale cloud-based network. NVGRE was designed to inexpensively resolve these security and scalability challenges by enabling a virtual overlay network without the need to reconfigure or add to the Layer 2 network.

While NVGRE is a step in the right direction, there are unforeseen performance challenges that are introduced from the loss of existing hardware offloads and from incompatibility with RSS.

In order to exploit the full potential of NVGRE, though, hardware that can properly address the performance challenges must be in place. This includes a network controller that not only supports NVGRE, but that can also access the inner packet to enable all existing hardware offloads and RSS traffic distribution. The only such NIC on the market at present is Mellanox's ConnectX-3 Pro.

About Mellanox

Mellanox Technologies (NASDAQ: MLNX) is a leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services.

More information is available at www.mellanox.com.

Works Cited

[1] Wall Street Journal, January 31, 2013: http://blogs.wsj.com/digits/2011/04/21/more-predictions-on-the-huge-growth-of-cloud-computing/

[2] Gartner, Inc., "Forecast Overview: Public Cloud Services, Worldwide, 2011-2016, 2Q12 Update"

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

© Copyright 2013. Mellanox Technologies. All rights reserved.
Mellanox, BridgeX, ConnectX, CORE-Direct, InfiniBridge, InfiniHost, InfiniScale, MLNX-OS, PhyX, SwitchX, Virtual Protocol Interconnect and Voltaire are registered trademarks of Mellanox Technologies, Ltd. Connect-IB, CoolBox, FabricIT, Mellanox Federal Systems, Mellanox Software Defined Storage, MetroX, MetroDX, Mellanox Open Ethernet, Open Ethernet, ScalableHPC, Unbreakable-Link, UFM and Unified Fabric Manager are trademarks of Mellanox Technologies, Ltd. All other trademarks are property of their respective owners.

15-1325WP
Rev 1.0