WHITE PAPER

NVGRE Overlay Networks: Enabling Network Scalability for a Cloud Infrastructure
Advanced Management Solutions

Table of contents

Executive Summary ................................................................. 3
Cloud Computing Growth ......................................................... 3
Cloud Computing Infrastructure Optimization .......................... 3
The Economics Driving Overlay Networks ............................... 4
Cloud Computing’s Network Scalability Challenge ................. 4
  Limitations to security (network isolation) using 802.1Q VLAN grouping .................................................. 5
  Capacity expansion across server racks or server clusters (use cases) ................................................................. 5
  Limitations of Spanning Tree Protocol (STP) .............................. 5
  Infrastructure over provisioning capital expenditure (CAPEX) ............................................................ 5
Addressing Network Scalability with NVGRE .......................... 6
  NVGRE—a solution to the network scalability challenge ......... 6
  What is NVGRE? ................................................................. 6
  NVGRE Ethernet frame encapsulation .................................... 7
  VM-to-VM communications in an NVGRE environment .......... 7
NVGRE—Enabling Network Scalability in the Cloud ............... 9
Looking Ahead ........................................................................... 10
NVGRE—Microsoft’s Call to Action .......................................... 10
Conclusion .................................................................................. 10
Executive Summary

The paradigm shift from a virtualization driven data center toward an infrastructure for cloud computing includes shifting the focus from applications consolidation to tenant consolidation. While virtualization has reduced the cost and time required to deploy a new application from weeks to minutes, reducing the costs from thousands of dollars to a few hundred, reconfiguring the network for a new or migrated virtual workload can take a week and cost thousands of dollars. Scaling existing networking technology for multi-tenant infrastructures requires solutions enabling virtual machine (VM) communication and migration across Layer 3 boundaries without impacting connectivity while ensuring isolation for hundreds of thousands of logical network segments and maintaining existing VM addresses and Media Access Control (MAC) addresses. NVGRE, an IETF proposed standard co-authored by Microsoft, Emulex and other leading organizations address these challenges.

Cloud Computing Growth

Cloud computing adoption, both public and private, continues unabated with server revenues in 2014 forecasted to grow to $12.5 Billion (IDC) and Cloud Services becoming a $35 Billion/year market by 2013 (Cisco).

Almost 80 percent of respondents to a recent study (ESG) indicated having three or more data centers worldwide, and 54 percent have started or completed data center consolidation projects. This trend toward data center consolidation is a key driver of new requirements for multi-tenant cloud based private infrastructures that are manageable and scalable.

Andrew McAfee of MIT’s Sloan School describes cloud computing a “deep and permanent change in how computing power is generated and consumed.”

Cloud computing, with any service model approach, public, private or hybrid, delivers multiple benefits including:

- Extracting increased hardware efficiencies from server consolidation via virtualization.
- Enabling IT agility through optimized placement or relocation of workloads.
- Optimizing IT capacity through distribution of workloads between private and public infrastructure.
- Ensuring business continuity and disaster recovery.

Server virtualization has laid the foundation for the growth of cloud computing, with hypervisor offerings such as Microsoft Hyper-V.

Cloud Computing Infrastructure Optimization

Ensuring optimized growth of cloud computing infrastructure, and achieving investment return, rests on three fundamental tenets – based on optimization of three data center infrastructures:

**Server**—The cloud is a natural evolution of server virtualization and advances including powerful multi-core processors, server clusters/pods and standardized low cost servers; together, they offer a plethora of options to design an optimized core cloud infrastructure.

**Storage**—Storage virtualization, de-duplication, on-demand and simplified provisioning, disaster recovery and scalable file storage using the Hadoop File Storage System (HDFS) provide the infrastructure fabric capabilities to optimally design for cloud based storage.

**Network**—Forrester’s research has indicated that 50 percent of companies that didn’t invest in the network suffered from degradation of services or had noticeable increases in operational costs. Advances in NIC adapter and Ethernet switching technologies are simultaneously driving the optimization of data center networking for cloud infrastructure.

The growth of 10Gb Ethernet (10GbE) networking, accelerating with the launch of Intel’s Xeon® E5 (Romley) processors, coupled with other adapter card advances including SR-IOV (Single Root I/O Virtualization) and Emulex’s vEngine™ protocol offload technology, available with the XE4310R Ethernet controller, enables greater VM consolidation on a host server without impacting networking I/O throughput.
However, when customers deploy private or hybrid clouds, a good architecture will always have spare capacity or spare hosts (over provision) for failover. A primary reason is that a VM can only communicate or be seamlessly migrated within a Layer 2 (L2) domain (L2 subnet).

Network optimization and network scalability will be critical to achieve the full potential of cloud based computing.

The Economics Driving Overlay Networks

Overlay Networks are being driven by, simply put, the need to reduce the costs of deploying a new server workload and managing it. Here is an interesting analysis from a presentation at the Open Networking Summit in the spring of 2012:¹ Server virtualization has drastically reduced the cost and time to deploy a new server workload. Before virtualization, it would take ten weeks and cost $10,000, plus software costs, to deploy a new server workload. With virtualization, a new server workload can be deployed in a few minutes at a deployment cost of $300, plus software costs. But, the network hasn’t been virtualized, and the network really hasn’t changed during the last ten years. From a network management perspective, it will take five days and cost $1,800 to reconfigure the network devices, including routers, IP and MAC addresses, firewalls, intrusion detection systems, etc. Overlay Networks is the first step to Network Virtualization and promises to significantly reduce the costs of network management for virtualized workloads and hybrid cloud deployments.

Cloud Computing’s Network Scalability Challenge

This paradigm shift from a virtualization-driven data center to an infrastructure architecture for supporting cloud computing will also shift the focus from applications consolidation to tenant consolidation. A tenant can be a department or a Line of Business (LoB) in a private cloud or a distinct corporate entity in a public or hybrid cloud.

For a cloud infrastructure, the importance of network scalability extends beyond the immediately obvious bandwidth driven economics, to that of increasing VM density on the physical host server and extending the reach of tenants.

Let us examine challenges to achieving network scale in a cloud based data center. Reviewing the OSI 7-layer architectural model in Figure 1 below, we will want to focus on the critical communication layers 2 and 3 as they pertain to VM-to-VM communications and VM migration.

Specifically, we will examine how to ensure intra-tenant VM-to-VM communications isolation (security) while enabling a tenant’s resource to seamlessly scale at layer 2 (L2) and, as equally important, address the challenges of elastically adding additional resources at layer 3 (L3).

¹ http://opennetsummit.org
**Limitations to security (network isolation) using 802.1Q vLAN grouping**

VMs belonging to a group (Departmental, LoB, Corporation, etc.) each require Virtual Local Area Network tagging (vLAN tagging) based L2 segment isolation for their network traffic (since VMs communicate at L2). However, the IEEE 802.1Q standard, with its 12-Bit namespace for a vLAN Identifier (VID), allows only 4094 vLANs (Note: The maximum number of vLANs that can be supported is 4094 rather than 4096, as the VID values of 0 and FFF are reserved).

A Top-of-Rack switch may connect to dozens of physical servers, each hosting a dozen or more VMs that can each belong to at least one vLAN. A data center can easily contain enough switches, resulting in the vLAN ID limit of 4094 being exceeded quickly.

Furthermore, VMs running a single application may reside in geographically disparate data centers. These VMs may have to communicate via an L2 data link, so vLAN identifiers must remain unique across geographies.

**Capacity expansion across server racks or server clusters (use cases)**

A single rack with multiple host servers, or multiple server racks residing in a common cluster or pod, effectively providing a small virtualized environment for a single tenant. Each server rack, may constitute an L2 subnet, while multiple host server racks in a cluster will likely be assigned their own L3 network.

Expanding (elastic) demand for additional computing resource by the tenant may require more servers or VMs on a different rack, within a different L2 subnet, or on a different host cluster, in a different L3 network. These additional VMs need to communicate with the applications VM on the “primary” rack or cluster.

A second use case is the need to migrate a VM(s) to another server cluster, that is, inter-host workload mobility to optimize usage of the underlying server hardware resources. This may be driven by scenarios such as:

- Migration from, and shut down of, underutilized hosts and consolidating on another more fully utilized host to reduce energy costs.
- Decommissioning one or more older hosts and bringing up workloads on newly added infrastructure.

These use cases require “stretching” the tenant’s L2 subnet to connect the servers or VM since VM-to-VM communications and/or VM migration between hosts currently requires the hosts to be in the same L2 subnet.

**Limitations of Spanning Tree Protocol (STP)**

STP effectively limits the number of VMs in a vLAN or IP network segment to 200-300 resulting in:

- “Wasted” links where links in the networking infrastructure are effectively turned off to avoid connectivity loops.
- Failures or changes in link status that result in loss of several seconds to re-establish connectivity – an unacceptable delay in many current applications scenarios.
- Limited ability to provide multipathing and resulting networking resiliency due to its (STP’s) requirement for a single active path between network nodes.

**Infrastructure over provisioning capital expenditures (CAPEX)**

*Emulex has learned from working with end users and talking with industry leaders that it isn’t uncommon to over provision a data center by as much as 10 percent to provide spare capacity for server failover. This over provisioning practice means data center managers allocate one spare physical host server for every 11, based on an approximate 23:1 consolidation ratio with the L2 domain limited by STP of 254 devices (network nodes).* Over provisioning spans the entire infrastructure and covers servers, racks, power and cooling, networking, administration and management, and some storage. NVGRE, on the other hand, breaks the STP boundary, allowing data center administrators to reduce their over provisioning to a much lower number. Since servers are fairly reliable, if the annual failure rate for servers is two percent, then the cost of the over provisioned infrastructure can be reduced by a factor of five. Optimizing the failover infrastructure is another opportunity for NVGRE to significantly reduce IT costs.
In summary, the current practices of host-centric segmentation based on vLANs/L2 subnets, with server clusters separated by L3 physical IP networks limit today’s multi-tenant cloud infrastructure from meaningful network scaling for longer term growth. Figure 2 describes these network scaling challenges associated with the deployment of a cloud based server infrastructure.

Addressing Network Scalability with NVGRE

NVGRE—a solution to the network scalability challenge

Microsoft and Emulex, along with other industry leaders, have outlined a vision for a new virtualized Overlay Network architecture that enables the efficient and fluid movement of virtual resources across cloud infrastructures, leading the way for large scale and cloud scale VM deployments. NVGRE (Network Virtualization using Generic Routing Encapsulation), is a proposed new IETF standard co-authored by Microsoft, Emulex and HP amongst a group of other leading companies. The NVGRE standard enables Overlay (virtual) Networks, and focuses on eliminating the constraints described earlier, thus enabling virtualized workloads to seamlessly communicate or move across server racks, clusters and data centers.

What is NVGRE?

At its core, NVGRE is simply an encapsulation of an Ethernet L2 Frame in IP that enables the creation of virtualized L2 subnets that can span physical L3 IP networks.

Looking back at the use case described earlier, NVGRE enables the connection between two or more L3 networks and makes it appear like they share the same L2 subnet. This now allows inter-VM communications or VM migrations across L3 networks while operating as if they were attached to the same L2 subnet.

NVGRE is an L2 overlay scheme over an L3 network.
NVGRE Ethernet frame encapsulation — A unique 24-bit ID called a Tenant Network Identifier (TNI) is added to the L2 Ethernet frame, using the lower 24 bits of the GRE Key field. This new 24-bit TNI now enables more than 16 million L2 (logical) networks to operate within the same administrative domain, a scalability improvement of many orders of magnitude over the 4,094 vLAN segment limit discussed above.

The L2 frame with GRE encapsulation is then encapsulated with an outer IP header and finally an outer MAC address. A simplified representation of the NVGRE frame format and encapsulation is shown in Figure 3 below.

NVGRE is a tunneling scheme, using the GRE routing protocol as defined by RFC 2784 and extended by RFC 2890. Each TNI is associated with an individual GRE tunnel and uniquely identifies, as its name suggests, a cloud tenant’s unique virtual subnet. NVGRE thus isn’t a new “standard” as such, since it uses the already established GRE protocol between hypervisors.

VM-to-VM communications in an NVGRE environment — The value of NVGRE technology is in enabling inter-VM communications and VM migration for hosts in different L2 subnets as discussed earlier. A simplified explanation of the process is described below and demonstrated in Figures 4 and 5.

The NVGRE endpoint, residing in the server (or a switch or physical 10GbE NIC) encapsulates the VM traffic, adding the 24-Bit TNI, and sends it through a GRE tunnel. At the destination, the endpoint de-encapsulates the incoming packets and presents the destination VM with the original Ethernet L2 packet. (NOTE: The VM originating the communication is unaware of this TNI tag.)

The current IETF draft specification explicitly avoids specifying the method of obtaining a destination VM address in the Ethernet frame, leaving that detail to a future version of the specification or new network management techniques such as Software Defined Networking (SDN).

The inner IP address is called the Customer Address (CA). The outer IP address is called the Provider Address (PA). When an NVGRE endpoint needs to send a packet to the destination VM, it needs to know the PA of the destination NVGRE endpoint. The mechanism to maintain a mapping between PAs and CAs will be more fully defined in additional white papers.
Figure 4

Physical L3 Network

L2 subnet cannot span Layer 3

Figure 5

Legend

TNI1

TNI2
NVGRE—Enabling Network Scalability in the Cloud

We have seen how NVGRE technology addresses some of the key tactical challenges associated with networking within a cloud infrastructure. Figure 6, a simplified representation, demonstrates the benefits of adopting NVGRE.

In summary, with NVGRE:

- The 4094 vLAN limit for ensuring computing privacy is addressed with the NVGRE 24-bit TNI construct that enables 16 million isolated tenant networks. VMs need to be on the same TNI to communicate with each other. This delivers the isolation demanded in a multi-tenant architecture keeping associated VMs within the same TNI.

- A multi-tenant cloud infrastructure is now capable of delivering “elastic” capacity service by enabling additional application VMs to be rapidly provisioned in a different L3 network which can communicate as if they are on a common L2 subnet.

- Overlay Networking overcomes the limits of STP and creates very large network domains where VMs can be moved anywhere. This also enables IT teams to reduce over provisioning to a much lower percentage which can save a lot of money. For example, deploying an extra server for 50 servers, over provisioning is reduced to two percent (from 10 percent estimated before). As a result, data centers can save as much as eight percent of their entire IT infrastructure budget with NVGRE Overlay Networking.

- Overlay Networking can make hybrid cloud deployments simpler to deploy because it leverages the ubiquity of IP for data flows over the WAN.

- VMs are uniquely identified by a combination of their MAC addresses and TNI. Thus it is acceptable for VMs to have duplicate MAC addresses, as long as they are in different tenant networks. This simplifies administration of multi-tenant customer networks for the Cloud service provider.

- Finally, NVGRE is an evolutionary solution, already supported by switches and next-generation hypervisors, not requiring “forklift” hardware upgrades thus easing and hastening the adoption of the technology.
Looking Ahead

Overlay Networking implemented in software, as currently defined, is incompatible with current vEngine™ (and similar) hardware based protocol offload technologies, so network performance goes down, more CPU cycles are consumed, and the VM-host consolidation ratio is reduced. This fundamentally impacts the core economics of virtualization and can easily consume 10 to 20 percent of CPU cycles of a host.

While the current NVGRE draft specification leaves out many of the details for managing the Control Plane, it outlines the critical piece for interoperability—the NVGRE frame format. As for the control plane, there will be several tools available to manage the networking piece. The fundamental philosophy is that when a VM moves, a VM management tool orchestrates its movement. Since this tool knows the complete lifecycle of a VM, it can provide all of the control plane information required by the network. This builds on the trend that the control plane migrates away from the hypervisor’s embedded vSwitch to the physical 10GbE NIC, with the NIC becoming more intelligent, and a critical part of the networking infrastructure.

Therefore, in order for NVGRE to be leveraged without a notable performance impact an intelligent NIC needs to be deployed that supports:

- NVGRE-offload without disrupting current protocol offload capabilities.
- API integration to apply the correct control plane attributes to each flow for Overlay Networking.
- APIs to expose the presence and connectivity of Overlay Networks to network management tools.

NVGRE - Microsoft’s Call to Action

- NVGRE is a pivotal technology that enables Microsoft’s HyperV for Hybrid Cloud environments
  - Required for scalable virtual hybrid cloud environments
  - Solves VM migration problem between separate L2 domains
- NVGRE will be released with Windows Server 2012 with support for hardware offloads
  - Microsoft will have a Hardware Qualified List for NICs and LOMs (including the servers that support them)
  - Microsoft call to action: NVGRE offloads be required for NIC hardware
  - Any encapsulation breaks stateless offloads, NICs/LOMs must become GRE aware
  - OEMs need to add NVGRE as a requirement to their NIC/LOM RFQs

Conclusion

By extending the routing infrastructure inside the server, Microsoft is enabling the creation of programmatic dynamic Overlay Networks that can easily scale from small to large to cloud enabling multi-tenant architectures. Network manageability can be moved to the edge, and the network management layer becomes virtualization aware and can adapt and manage the network as VMs start up, shut down, and move around.

NVGRE, a new standard from Microsoft and Emulex to create Overlay Networks, focuses on eliminating the last constraint to enable workloads to seamlessly move across racks, host clusters and data centers.

This is a very important step towards lowering the cost of the data center infrastructure. Given the cost and complexity of traditional networks to deploy and manage virtual servers, Overlay Network protocols such as NVGRE can reduce the time and cost to deploy and manage a virtual workload throughout its lifecycle.