Independent Submission
Request for Comments: 7637
Category: Informational
ISSN: 2070-1721
September 2015

P. Garg, Ed.
Y. Wang, Ed.
Microsoft

NVGRE: Network Virtualization Using Generic Routing Encapsulation

Abstract

This document describes the usage of the Generic Routing Encapsulation (GRE) header for Network Virtualization (NVGRE) in multi-tenant data centers. Network Virtualization decouples virtual networks and addresses from physical network infrastructure, providing isolation and concurrency between multiple virtual networks on the same physical network infrastructure. This document also introduces a Network Virtualization framework to illustrate the use cases, but the focus is on specifying the data-plane aspect of NVGRE.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7637.

Copyright Notice

Table of Contents

1. Introduction

Conventional data center network designs cater to largely static workloads and cause fragmentation of network and server capacity [6] [7]. There are several issues that limit dynamic allocation and consolidation of capacity. Layer 2 networks use the Rapid Spanning Tree Protocol (RSTP), which is designed to eliminate loops by blocking redundant paths. These eliminated paths translate to wasted capacity and a highly oversubscribed network. There are alternative approaches such as the Transparent Interconnection of Lots of Links (TRILL) that address this problem [13].

The network utilization inefficiencies are exacerbated by network fragmentation due to the use of VLANs for broadcast isolation. VLANs are used for traffic management and also as the mechanism for providing security and performance isolation among services belonging to different tenants. The Layer 2 network is carved into smaller-sized subnets (typically, one subnet per VLAN), with VLAN tags configured on all the Layer 2 switches connected to server racks that host a given tenant's services. The current VLAN limits theoretically allow for 4,000 such subnets to be created. The size of the broadcast domain is typically restricted due to the overhead of broadcast traffic. The 4,000-subnet limit on VLANs is no longer sufficient in a shared infrastructure servicing multiple tenants.

Data center operators must be able to achieve high utilization of server and network capacity. In order to achieve efficiency, it should be possible to assign workloads that operate in a single Layer 2 network to any server in any rack in the network. It should also be possible to migrate workloads to any server anywhere in the network while retaining the workloads' addresses. This can be achieved today by stretching VLANs; however, when workloads migrate, the network needs to be reconfigured and that is typically error prone. By decoupling the workload's location on the LAN from its network address, the network administrator configures the network once, not every time a service migrates. This decoupling enables any server to become part of any server resource pool.

The following are key design objectives for next-generation data centers:

a) location-independent addressing

b) the ability to a scale the number of logical Layer 2 / Layer 3 networks, irrespective of the underlying physical topology or the number of VLANs

c) preserving Layer 2 semantics for services and allowing them to retain their addresses as they move within and across data centers

d) providing broadcast isolation as workloads move around without burdening the network control plane

This document describes use of the Generic Routing Encapsulation (GRE) header [3] [4] for network virtualization. Network virtualization decouples a virtual network from the underlying physical network infrastructure by virtualizing network addresses. Combined with a management and control plane for the virtual-to-physical mapping, network virtualization can enable flexible virtual machine placement and movement and provide network isolation for a multi-tenant data center.

Network virtualization enables customers to bring their own address spaces into a multi-tenant data center, while the data center administrators can place the customer virtual machines anywhere in the data center without reconfiguring their network switches or routers, irrespective of the customer address spaces.

1.1. Terminology

Please refer to RFCs 7364 [10] and 7365 [11] for more formal definitions of terminology. The following terms are used in this document.

Customer Address (CA): This is the virtual IP address assigned and configured on the virtual Network Interface Controller (NIC) within each VM. This is the only address visible to VMs and applications running within VMs.

Network Virtualization Edge (NVE): This is an entity that performs the network virtualization encapsulation and decapsulation.

Provider Address (PA): This is the IP address used in the physical network. PAs are associated with VM CAs through the network virtualization mapping policy.

Virtual Machine (VM): This is an instance of an OS running on top of the hypervisor over a physical machine or server. Multiple VMs can share the same physical server via the hypervisor, yet are completely isolated from each other in terms of CPU usage, storage, and other OS resources.

Virtual Subnet Identifier (VSID): This is a 24-bit ID that uniquely identifies a virtual subnet or virtual Layer 2 broadcast domain.

2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [1].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lowercase uses of these words are not to be interpreted as carrying the significance defined in RFC 2119.

3. Network Virtualization Using GRE (NVGRE)

This section describes Network Virtualization using GRE (NVGRE). Network virtualization involves creating virtual Layer 2 topologies on top of a physical Layer 3 network. Connectivity in the virtual topology is provided by tunneling Ethernet frames in GRE over IP over the physical network.

In NVGRE, every virtual Layer 2 network is associated with a 24-bit identifier, called a Virtual Subnet Identifier (VSID). A VSID is carried in an outer header as defined in Section 3.2. This allows unique identification of a tenant's virtual subnet to various devices in the network. A 24-bit VSID supports up to 16 million virtual subnets in the same management domain, in contrast to only 4,000 that is achievable with VLANs. Each VSID represents a virtual Layer 2 broadcast domain, which can be used to identify a virtual subnet of a given tenant. To support multi-subnet virtual topology, data center administrators can configure routes to facilitate communication between virtual subnets of the same tenant.

GRE is a Proposed Standard from the IETF [3] [4] and provides a way for encapsulating an arbitrary protocol over IP. NVGRE leverages the GRE header to carry VSID information in each packet. The VSID information in each packet can be used to build multi-tenant-aware tools for traffic analysis, traffic inspection, and monitoring.

The following sections detail the packet format for NVGRE; describe the functions of an NVGRE endpoint; illustrate typical traffic flow both within and across data centers; and discuss address/policy management, and deployment considerations.

3.1. NVGRE Endpoint

NVGRE endpoints are the ingress/egress points between the virtual and the physical networks. The NVGRE endpoints are the NVEs as defined in the Network Virtualization over Layer 3 (NVO3) Framework document [11]. Any physical server or network device can be an NVGRE endpoint. One common deployment is for the endpoint to be part of a hypervisor. The primary function of this endpoint is to encapsulate/decapsulate Ethernet data frames to and from the GRE tunnel, ensure Layer 2 semantics, and apply isolation policy scoped on VSID. The endpoint can optionally participate in routing and function as a gateway in the virtual topology. To encapsulate an Ethernet frame, the endpoint needs to know the location information for the destination address in the frame. This information can be provisioned via a management plane or obtained via a combination of control-plane distribution or data-plane learning approaches. This document assumes that the location information, including VSID, is available to the NVGRE endpoint.

3.2. NVGRE Frame Format

The GRE header format as specified in RFCs 2784 [3] and 2890 [4] is used for communication between NVGRE endpoints. NVGRE leverages the Key extension specified in RFC 2890 [4] to carry the VSID. The packet format for Layer 2 encapsulation in GRE is shown in Figure 1.

```
   Outer Ethernet Header:
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                (Outer) Destination MAC Address                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |(Outer)Destination MAC Address |  (Outer)Source MAC Address    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                  (Outer) Source MAC Address                   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |Optional Ethertype=C-Tag 802.1Q| Outer VLAN Tag Information    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       Ethertype 0x0800        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

   Outer IPv4 Header:
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |Version|  HL   |Type of Service|          Total Length         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Identification        |Flags|      Fragment Offset    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  Time to Live | Protocol 0x2F |         Header Checksum       |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      (Outer) Source Address                   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                  (Outer) Destination Address                  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

   GRE Header:
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |0| |1|0|   Reserved0     | Ver |   Protocol Type 0x6558        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               Virtual Subnet ID (VSID)        |    FlowID     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   Inner Ethernet Header
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                (Inner) Destination MAC Address                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |(Inner)Destination MAC Address |  (Inner)Source MAC Address    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                  (Inner) Source MAC Address                   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       Ethertype 0x0800        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Inner IPv4 Header:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version|  HL   |Type of Service|          Total Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Identification        |Flags|      Fragment Offset    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time to Live |    Protocol   |         Header Checksum       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Source Address                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Destination Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Options                    |    Padding    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Original IP Payload                      |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: GRE Encapsulation Frame Format

Note: HL stands for Header Length.

The outer/delivery headers include the outer Ethernet header and the outer IP header:

- o The outer Ethernet header: The source Ethernet address in the outer frame is set to the MAC address associated with the NVGRE endpoint. The destination endpoint may or may not be on the same physical subnet. The destination Ethernet address is set to the MAC address of the next-hop IP address for the destination NVE. The outer VLAN tag information is optional and can be used for traffic management and broadcast scalability on the physical network.

- o The outer IP header: Both IPv4 and IPv6 can be used as the delivery protocol for GRE. The IPv4 header is shown for illustrative purposes. Henceforth, the IP address in the outer frame is referred to as the Provider Address (PA). There can be one or more PA associated with an NVGRE endpoint, with policy controlling the choice of which PA to use for a given Customer Address (CA) for a customer VM.

In the GRE header:

- o The C (Checksum Present) and S (Sequence Number Present) bits in the GRE header MUST be zero.

- The K (Key Present) bit in the GRE header MUST be set to one. The 32-bit Key field in the GRE header is used to carry the Virtual Subnet ID (VSID) and the FlowID:

  - Virtual Subnet ID (VSID): This is a 24-bit value that is used to identify the NVGRE-based Virtual Layer 2 Network.

  - FlowID: This is an 8-bit value that is used to provide per-flow entropy for flows in the same VSID. The FlowID MUST NOT be modified by transit devices. The encapsulating NVE SHOULD provide as much entropy as possible in the FlowID. If a FlowID is not generated, it MUST be set to all zeros.

- The Protocol Type field in the GRE header is set to 0x6558 (Transparent Ethernet Bridging) [2].

In the inner headers (headers of the GRE payload):

- The inner Ethernet frame comprises an inner Ethernet header followed by optional inner IP header, followed by the IP payload. The inner frame could be any Ethernet data frame not just IP. Note that the inner Ethernet frame's Frame Check Sequence (FCS) is not encapsulated.

- For illustrative purposes, IPv4 headers are shown as the inner IP headers, but IPv6 headers may be used. Henceforth, the IP address contained in the inner frame is referred to as the Customer Address (CA).

3.3. Inner Tag as Defined by IEEE 802.1Q

The inner Ethernet header of NVGRE MUST NOT contain the tag as defined by IEEE 802.1Q [5]. The encapsulating NVE MUST remove any existing IEEE 802.1Q tag before encapsulation of the frame in NVGRE. A decapsulating NVE MUST drop the frame if the inner Ethernet frame contains an IEEE 802.1Q tag.

3.4. Reserved VSID

The VSID range from 0-0xFFF is reserved for future use.

The VSID 0xFFFFFF is reserved for vendor-specific NVE-to-NVE communication. The sender NVE SHOULD verify the receiver NVE's vendor before sending a packet using this VSID; however, such a verification mechanism is out of scope of this document. Implementations SHOULD choose a mechanism that meets their requirements.

4. NVGRE Deployment Considerations

4.1. ECMP Support

Equal-Cost Multipath (ECMP) may be used to provide load balancing. If ECMP is used, it is RECOMMENDED that the ECMP hash is calculated either using the outer IP frame fields and entire Key field (32 bits) or the inner IP and transport frame fields.

4.2. Broadcast and Multicast Traffic

To support broadcast and multicast traffic inside a virtual subnet, one or more administratively scoped multicast addresses [8] [9] can be assigned for the VSID. All multicast or broadcast traffic originating from within a VSID is encapsulated and sent to the assigned multicast address. From an administrative standpoint, it is possible for network operators to configure a PA multicast address for each multicast address that is used inside a VSID; this facilitates optimal multicast handling. Depending on the hardware capabilities of the physical network devices and the physical network architecture, multiple virtual subnets may use the same physical IP multicast address.

Alternatively, based upon the configuration at the NVE, broadcast and multicast in the virtual subnet can be supported using N-way unicast. In N-way unicast, the sender NVE would send one encapsulated packet to every NVE in the virtual subnet. The sender NVE can encapsulate and send the packet as described in Section 4.3 ("Unicast Traffic"). This alleviates the need for multicast support in the physical network.

4.3. Unicast Traffic

The NVGRE endpoint encapsulates a Layer 2 packet in GRE using the source PA associated with the endpoint with the destination PA corresponding to the location of the destination endpoint. As outlined earlier, there can be one or more PAs associated with an endpoint and policy will control which ones get used for communication. The encapsulated GRE packet is bridged and routed normally by the physical network to the destination PA. Bridging uses the outer Ethernet encapsulation for scope on the LAN. The only requirement is bidirectional IP connectivity from the underlying physical network. On the destination, the NVGRE endpoint decapsulates the GRE packet to recover the original Layer 2 frame. Traffic flows similarly on the reverse path.

4.4. IP Fragmentation

Section 5.1 of RFC 2003 [12] specifies mechanisms for handling fragmentation when encapsulating IP within IP. The subset of mechanisms NVGRE selects are intended to ensure that NVGRE-encapsulated frames are not fragmented after encapsulation en route to the destination NVGRE endpoint and that traffic sources can leverage Path MTU discovery.

A sender NVE MUST NOT fragment NVGRE packets. A receiver NVE MAY discard fragmented NVGRE packets. It is RECOMMENDED that the MTU of the physical network accommodates the larger frame size due to encapsulation. Path MTU or configuration via control plane can be used to meet this requirement.

4.5. Address/Policy Management and Routing

Address acquisition is beyond the scope of this document and can be obtained statically, dynamically, or using stateless address autoconfiguration. CA and PA space can be either IPv4 or IPv6. In fact, the address families don’t have to match; for example, a CA can be IPv4 while the PA is IPv6, and vice versa.

4.6. Cross-Subnet, Cross-Premise Communication

One application of this framework is that it provides a seamless path for enterprises looking to expand their virtual machine hosting capabilities into public clouds. Enterprises can bring their entire IP subnet(s) and isolation policies, thus making the transition to or from the cloud simpler. It is possible to move portions of an IP subnet to the cloud; however, that requires additional configuration on the enterprise network and is not discussed in this document. Enterprises can continue to use existing communications models like site-to-site VPN to secure their traffic.

A VPN gateway is used to establish a secure site-to-site tunnel over the Internet, and all the enterprise services running in virtual machines in the cloud use the VPN gateway to communicate back to the enterprise. For simplicity, we use a VPN gateway configured as a VM (shown in Figure 2) to illustrate cross-subnet, cross-premise communication.

```
   +-----------------------+        +-----------------------+
   |       Server 1        |        |       Server 2        |
   | +--------+ +--------+ |        | +-------------------+ |
   | | VM1    | | VM2    | |        | |    VPN Gateway    | |
   | | IP=CA1 | | IP=CA2 | |        | | Internal  External| |
   | |        | |        | |        | |  IP=CAg   IP=GAdc | |
   | +--------+ +--------+ |        | +-------------------+ |
   |       Hypervisor      |        |     | Hypervisor| ^   |
   +-----------------------+        +-------------------:---+
               | IP=PA1                   | IP=PA4    | :
               |                          |           | :
               |     +-------------------------+      | : VPN
               +-----|     Layer 3 Network     |------+ : Tunnel
                     +-------------------------+        :
                                  |                     :
        +-----------------------------------------------:--+
        |                                               :  |
        |                     Internet                  :  |
        |                                               :  |
        +-----------------------------------------------:--+
                                  |                     v
                                  |   +-------------------+
                                  |   |    VPN Gateway    |
                                  |---|                   |
                             IP=GAcorp| External IP=GAcorp|
                                      +-------------------+
                                                |
                                    +-----------------------+
                                    |  Corp Layer 3 Network |
                                    |      (In CA Space)    |
                                    +-----------------------+
                                                |
                                   +---------------------------+
                                   |       Server X            |
                                   | +----------+ +----------+ |
                                   | | Corp VMe1| | Corp VMe2| |
                                   | |  IP=CAe1 | |  IP=CAe2 | |
                                   | +----------+ +----------+ |
                                   |         Hypervisor        |
                                   +---------------------------+
```

Figure 2: Cross-Subnet, Cross-Premise Communication

The packet flow is similar to the unicast traffic flow between VMs; the key difference in this case is that the packet needs to be sent to a VPN gateway before it gets forwarded to the destination. As part of routing configuration in the CA space, a per-tenant VPN gateway is provisioned for communication back to the enterprise. The example illustrates an outbound connection between VM1 inside the data center and VMe1 inside the enterprise network. When the outbound packet from CA1 to CAe1 reaches the hypervisor on Server 1, the NVE in Server 1 can perform the equivalent of a route lookup on the packet. The cross-premise packet will match the default gateway rule, as CAe1 is not part of the tenant virtual network in the data center. The virtualization policy will indicate the packet to be encapsulated and sent to the PA of the tenant VPN gateway (PA4) running as a VM on Server 2. The packet is decapsulated on Server 2 and delivered to the VM gateway. The gateway in turn validates and sends the packet on the site-to-site VPN tunnel back to the enterprise network. As the communication here is external to the data center, the PA address for the VPN tunnel is globally routable. The outer header of this packet is sourced from GAdc destined to GAcorp. This packet is routed through the Internet to the enterprise VPN gateway, which is the other end of the site-to-site tunnel; at that point, the VPN gateway decapsulates the packet and sends it inside the enterprise where the CAe1 is routable on the network. The reverse path is similar once the packet reaches the enterprise VPN gateway.

4.7. Internet Connectivity

To enable connectivity to the Internet, an Internet gateway is needed that bridges the virtualized CA space to the public Internet address space. The gateway needs to perform translation between the virtualized world and the Internet. For example, the NVGRE endpoint can be part of a load balancer or a NAT that replaces the VPN Gateway on Server 2 shown in Figure 2.

4.8. Management and Control Planes

There are several protocols that can manage and distribute policy; however, it is outside the scope of this document. Implementations SHOULD choose a mechanism that meets their scale requirements.

4.9. NVGRE-Aware Devices

One example of a typical deployment consists of virtualized servers deployed across multiple racks connected by one or more layers of Layer 2 switches, which in turn may be connected to a Layer 3 routing domain. Even though routing in the physical infrastructure will work without any modification with NVGRE, devices that perform specialized processing in the network need to be able to parse GRE to get access to tenant-specific information. Devices that understand and parse the VSID can provide rich multi-tenant-aware services inside the data center. As outlined earlier, it is imperative to exploit multiple paths inside the network through techniques such as ECMP. The Key field (a 32-bit field, including both the VSID and the optional FlowID) can provide additional entropy to the switches to exploit path diversity inside the network. A diverse ecosystem is expected to emerge as more and more devices become multi-tenant aware. In the interim, without requiring any hardware upgrades, there are alternatives to exploit path diversity with GRE by associating multiple PAs with NVGRE endpoints with policy controlling the choice of which PA to use.

It is expected that communication can span multiple data centers and also cross the virtual/physical boundary. Typical scenarios that require virtual-to-physical communication include access to storage and databases. Scenarios demanding lossless Ethernet functionality may not be amenable to NVGRE, as traffic is carried over an IP network. NVGRE endpoints mediate between the network-virtualized and non-network-virtualized environments. This functionality can be incorporated into Top-of-Rack switches, storage appliances, load balancers, routers, etc., or built as a stand-alone appliance.

It is imperative to consider the impact of any solution on host performance. Today's server operating systems employ sophisticated acceleration techniques such as checksum offload, Large Send Offload (LSO), Receive Segment Coalescing (RSC), Receive Side Scaling (RSS), Virtual Machine Queue (VMQ), etc. These technologies should become NVGRE aware. IPsec Security Associations (SAs) can be offloaded to the NIC so that computationally expensive cryptographic operations are performed at line rate in the NIC hardware. These SAs are based on the IP addresses of the endpoints. As each packet on the wire gets translated, the NVGRE endpoint SHOULD intercept the offload requests and do the appropriate address translation. This will ensure that IPsec continues to be usable with network virtualization while taking advantage of hardware offload capabilities for improved performance.

4.10. Network Scalability with NVGRE

One of the key benefits of using NVGRE is the IP address scalability and in turn MAC address table scalability that can be achieved. An NVGRE endpoint can use one PA to represent multiple CAs. This lowers the burden on the MAC address table sizes at the Top-of-Rack switches. One obvious benefit is in the context of server virtualization, which has increased the demands on the network infrastructure. By embedding an NVGRE endpoint in a hypervisor, it is possible to scale significantly. This framework enables location information to be preconfigured inside an NVGRE endpoint, thus allowing broadcast ARP traffic to be proxied locally. This approach can scale to large-sized virtual subnets. These virtual subnets can be spread across multiple Layer 3 physical subnets. It allows workloads to be moved around without imposing a huge burden on the network control plane. By eliminating most broadcast traffic and converting others to multicast, the routers and switches can function more optimally by building efficient multicast trees. By using server and network capacity efficiently, it is possible to drive down the cost of building and managing data centers.

5. Security Considerations

This proposal extends the Layer 2 subnet across the data center and increases the scope for spoofing attacks. Mitigations of such attacks are possible with authentication/encryption using IPsec or any other IP-based mechanism. The control plane for policy distribution is expected to be secured by using any of the existing security protocols. Further management traffic can be isolated in a separate subnet/VLAN.

The checksum in the GRE header is not supported. The mitigation of this is to deploy an NVGRE-based solution in a network that provides error detection along the NVGRE packet path, for example, using Ethernet Cyclic Redundancy Check (CRC) or IPsec or any other error detection mechanism.

6. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[2] IANA, "IEEE 802 Numbers", <http://www.iana.org/assignments/ieee-802-numbers>.

[3] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, DOI 10.17487/RFC2784, March 2000, <http://www.rfc-editor.org/info/rfc2784>.

[4] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, DOI 10.17487/RFC2890, September 2000, <http://www.rfc-editor.org/info/rfc2890>.

[5] IEEE, "IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q.

[6] Greenberg, A., et al., "VL2: A Scalable and Flexible Data Center Network", Communications of the ACM, DOI 10.1145/1897852.1897877, 2011.

[7] Greenberg, A., et al., "The Cost of a Cloud: Research Problems in Data Center Networks", ACM SIGCOMM Computer Communication Review, DOI 10.1145/1496091.1496103, 2009.

[8] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <http://www.rfc-editor.org/info/rfc4291>.

[9] Meyer, D., "Administratively Scoped IP Multicast", BCP 23, RFC 2365, DOI 10.17487/RFC2365, July 1998, <http://www.rfc-editor.org/info/rfc2365>.

[10] Narten, T., Ed., Gray, E., Ed., Black, D., Fang, L., Kreeger, L., and M. Napierala, "Problem Statement: Overlays for Network Virtualization", RFC 7364, DOI 10.17487/RFC7364, October 2014, <http://www.rfc-editor.org/info/rfc7364>.

[11] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <http://www.rfc-editor.org/info/rfc7365>.

[12] Perkins, C., "IP Encapsulation within IP", RFC 2003, DOI 10.17487/RFC2003, October 1996, <http://www.rfc-editor.org/info/rfc2003>.

[13] Touch, J. and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement", RFC 5556, DOI 10.17487/RFC5556, May 2009, <http://www.rfc-editor.org/info/rfc5556>.

Contributors

Murari Sridharan
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
United States
Email: muraris@microsoft.com

Albert Greenberg
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
United States
Email: albert@microsoft.com

Narasimhan Venkataramiah
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
United States
Email: navenkat@microsoft.com

Kenneth Duda
Arista Networks, Inc.
5470 Great America Pkwy
Santa Clara, CA 95054
United States
Email: kduda@aristanetworks.com

Ilango Ganga
Intel Corporation
2200 Mission College Blvd.
M/S: SC12-325
Santa Clara, CA 95054
United States
Email: ilango.s.ganga@intel.com

Geng Lin
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States
Email: genglin@google.com

Mark Pearson
Hewlett-Packard Co.
8000 Foothills Blvd.
Roseville, CA 95747
United States
Email: mark.pearson@hp.com

Patricia Thaler
Broadcom Corporation
3151 Zanker Road
San Jose, CA 95134
United States
Email: pthaler@broadcom.com

Chait Tumuluri
Emulex Corporation
3333 Susan Street
Costa Mesa, CA 92626
United States
Email: chait@emulex.com

Authors' Addresses

Pankaj Garg (editor)
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
United States
Email: pankajg@microsoft.com

Yu-Shun Wang (editor)
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
United States
Email: yushwang@microsoft.com