Network Working Group   M. Sridharan
Internet Draft   A. Greenberg
Intended Category: Informational   N. Venkataramiah
Expires: August 25, 2013   Y. Wang
                                                           Microsoft
                                                            K. Duda
                                                      Arista Networks
                                                           I. Ganga
                                                              Intel
                                                             G. Lin
                                                               Dell
                                                         M. Pearson
                                                     Hewlett-Packard
                                                          P. Thaler
                                                           Broadcom
                                                        C. Tumuluri
                                                             Emulex
                                                   February 25, 2013

         NVGRE: Network Virtualization using Generic Routing Encapsulation
                draft-sridharan-virtualization-nvgre-02.txt

Status of this Memo

   This memo provides information for the Internet Community. It does
   not specify an Internet standard of any kind; instead it relies on a
   proposed standard. Distribution of this memo is unlimited.

Copyright Notice

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html

This document is subject to BCP 78 and the IETF Trust's Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with
respect to this document.

This Internet-Draft will expire on August 24, 2013.

Abstract

This document describes the usage of Generic Routing Encapsulation
(GRE) header for Network Virtualization (NVGRE) in multi-tenant
datacenters. 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.

Table of Contents

1. Introduction

   Conventional data center network designs cater to largely static
   workloads and cause fragmentation of network and server capacity
   [VL2, COST-CCR]. There are several issues that limit dynamic
   allocation and consolidation of capacity. Layer-2 networks use 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 TRILL that address this problem
   [TRILL].

   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 run a given tenant's services. The current VLAN limits
   theoretically allow for 4K such subnets to be created. The size of
   the broadcast domain is typically restricted due to the overhead of
   broadcast traffic (e.g., ARP). The 4K VLAN limit 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 workload's addresses. This can be
   achieved today by stretching VLANs however when workloads migrate
   the network needs to be reconfigured which is typically error prone.
   By decoupling the workload's location on the LAN from its network
   address, the network administrator configures the network once and
   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 concurrent VLANs,
c) preserving Layer-2 semantics for services and allowing them to
retain their addresses as they move within and across data centers,
and d) providing broadcast isolation as workloads move around
without burdening the network control plane.

1.1. Terminology

Refer to [NVO3 Framework] or [NVO3 PA&S]

NVE: Network Virtualization Endpoint

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 [RFC2119].

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

3. Network Virtualization using GRE

This section describes Network Virtualization using GRE, NVGRE.
Network virtualization involves creating virtual Layer 2 and/or
Layer 3 topologies on top of an arbitrary physical Layer 2/Layer 3
network. Connectivity in the virtual topology is provided by
tunneling Ethernet frames in IP over the physical network. Virtual
broadcast domains are realized as multicast distribution trees. The
multicast distribution trees are analogous to the VLAN broadcast
domains. A virtual Layer 2 network can span multiple physical
subnets. Support for bi-directional IP unicast and multicast
connectivity is the only requirement from the underlying physical
network to support unicast communications within a virtual network.
If the operator chooses to support broadcast and multicast traffic
in the virtual topology the physical topology must support IP
multicast. The physical network, for example, can be a conventional
hierarchical 3-tier network, a full bisection bandwidth Clos
network, or a large Layer 2 network with or without TRILL support.

Every virtual Layer-2 network is associated with a 24 bit
identifier, called Virtual Subnet Identifier (VSID). A 24 bit VSID
allows up to 16 million virtual subnets in the same management
domain in contrast to only 4K achievable with VLANs. Each VSID
represents a virtual Layer-2 broadcast domain and routes can be

configured for communication between virtual subnets. The VSID can
be crafted in such a way that it uniquely identifies a specific
tenant's subnet. The VSID is carried in an outer header allowing
unique identification of the tenant's virtual subnet to various
devices in the network.

GRE is a proposed IETF standard [RFC 2784, RFC 2890] 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-tenanct-aware
tools for traffic analysis, traffic inspection, and monitoring.

The following sections detail the packet format for NVGRE, describe
the functions of a 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. Any physical server or network device can
be a 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

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

```
    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 Ethernet Header:                  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 (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|  IHL  |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)         |   Reserved    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   Inner Ethernet Header
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 (Inner) Destination MAC Address              |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |(Inner)Destination MAC Address | (Inner)Source MAC Address    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 (Inner) Source MAC Address                   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |Optional Ethertype=C-Tag 802.1Q| PCP |0| VID set to 0         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |      Ethertype 0x0800         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

(Continued on the next page)

Inner IPv4 Header:
```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version|  IHL  |Type of Service|          Total Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Identification        |Flags|      Fragment Offset    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time to Live |    Protocol   |         Header Checksum        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Source Address                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Destination Address                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Options               |     Padding         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Original IP Payload                        |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                    GRE Encapsulation Frame Format

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 Ethernet address is set to the MAC address
of the nexthop IP address for the destination NVE. The destination
endpoint may or may not be on the same physical subnet. The outer
VLAN tag information is optional and can be used for traffic
management and broadcast scalability.

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 address
associated with an NVGRE endpoint, with policy controlling the
choice of PA to use for a given CA.

The GRE header:

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

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

- Virtual Subnet ID (VSID): The first 24 bits are used for VSID
  as shown in Figure 1.
- FlowID: The last 8 bits of the Key field are (optional)
  FlowID, which can be used to add per-flow entropy within the
  same VSID, where the entire Key field (32-bit) is used for
  ECMP purposes by switches or routers in the physical network
  infrastructure. If a FlowID is not generated, the FlowID field
  MUST be set to all zero.

o The protocol type field in the GRE header is set to 0x6558
(transparent Ethernet bridging) [ETHTYPES].

The inner headers (headers of the GRE payload):

o The inner Ethernet frame comprises of an inner Ethernet header
followed by the 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 FCS is not encapsulated.

o Inner VLAN tag: The inner Ethernet header of NVGRE SHOULD NOT
contain inner VLAN Tag. When an NVE performs NVGRE encapsulation, it
SHOULD remove any existing VLAN Tag before encapsulating NVGRE
headers. If a VLAN-tagged frame arrives encapsulated in NVGRE, then
the decapsulating NVE SHOULD drop the frame.

**4. o 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).NVGRE Deployment Consideration

4.1. Broadcast and Multicast Traffic

The following discussion applies if the network operator chooses to
support broadcast and multicast traffic. Each virtual subnet is
assigned an administratively scoped multicast address to carry
broadcast and multicast traffic. All traffic originating from within
a VSID is encapsulated and sent to the assigned multicast address.
As an example, the addresses can be derived from a administratively
scoped multicast address as specified in RFC 2365 for IPv4
(organization Local Scope 239.192.0.0/14), or an Organization-Local
scope multicast address for IPv6 as specified in RFC 4291. This
provides a wide range of address choices. Purely from an efficiency
standpoint for every multicast address that a tenant uses the
network operator may configure a corresponding multicast address in
the PA space. To support broadcast and multicast traffic in the
virtual topology the physical topology must support IP multicast.
Depending on the hardware capabilities of the physical network

   devices multiple virtual broadcast domains may be assigned the same
   physical IP multicast address. For interoperability reasons, a
   future version of this draft will specify a standard way to map VSID
   to IP multicast address.

4.2. 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. Bridging uses
   the outer Ethernet encapsulation for scope on the LAN. The only
   assumption is bi-directional 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.

4.3. IP Fragmentation

   RFC 2003 section 5.1 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 future version of this draft will clarify the details around
   setting the DF bit on the outer IP header as well as maintaining per
   destination NVGRE endpoint MTU soft state so that ICMP Datagram Too
   Big messages can be exploited. Fragmentation behavior when tunneling
   non-IP Ethernet frames in GRE will also be specified in a future
   version.

4.4. Address/Policy Management & Routing

   Address acquisition is beyond the scope of this document and can be
   obtained statically, dynamically or using stateless address auto-
   configuration. CA and PA space can be either IPv4 or IPv6. In fact
   the address families don't have to match, for example, CA can be
   IPv4 while PA is IPv6 and vice versa. The isolation policies MUST be
   explicitly configured in the NVGRE endpoint. A typical policy table
   entry consists of CA, MAC address, VSID and optionally, the specific
   PA if more than one PA is associated with the NVGRE endpoint. If
   there are multiple virtual subnets, explicit routing information
   MUST be configured along with a default gateway for cross-subnet
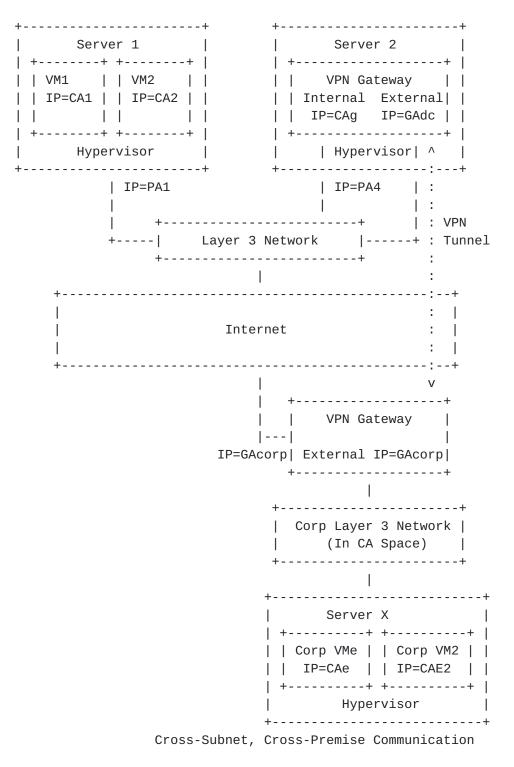   communication. Routing between virtual subnets can be optionally

   handled by the NVGRE endpoint acting as a gateway. If
   broadcast/multicast support is required the NVGRE endpoints MUST
   participate in IGMP/MLD for all subscribed multicast groups.

4.5. 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 a 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 GW 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 VMe | | Corp VM2 | |
                      | | IP=CAe   | | IP=CAE2  | |
                      | +----------+ +----------+ |
                      |         Hypervisor      |
                      +-------------------------+
            Cross-Subnet, Cross-Premise Communication
```

   The flow here is similar to the unicast traffic flow between VMs,
   the key difference in this case 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 VPN gateway is provisioned
   per-tenant for communication back to the enterprise. The example

illustrates an outbound connection between VM1 inside the datacenter
and VMe inside the enterprise network. The outbound packet from CA1
to CAe when it hits the hypervisor on Server 1 matches the default
gateway rule as CAe is not part of the tenant virtual network in the
datacenter. The packet is encapsulated and sent to the PA of 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
tunnel back to the enterprise network. As the communication here is
external to the datacenter 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 which point the VPN decapsulates the packet and sends
it inside the enterprise where the CAe is routable on the network.
The reverse path is similar once the packet hits the enterprise VPN
gateway.

4.6. 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 performs translation between the
virtualized world and the Internet, for example, the NVGRE endpoint
can be part of a load balancer or a NAT. Section 4 has more
discussions around building GRE gateways.

4.7. Management and Control Planes

There are several protocols that can manage and distribute policy;
however this document does not recommend any one mechanism.
Implementations SHOULD choose a mechanism that meets their scale
requirements.

4.8. 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 GRE, 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-tenancy aware services inside the
data center. As outlined earlier it is imperative to exploit
multiple paths inside the network through techniques such as Equal
Cost Multipath (ECMP). The Key field may provide additional entropy

to the switches to exploit path diversity inside the network. One
such example could be to use the upper 8 bits of the Key field to
add flow based entropy and tag all the packets from a flow with an
entropy label. A diverse ecosystem play is expected to emerge as
more and more devices become multitenancy 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 PA to be used.

It is expected that communication can span multiple data centers and
also cross the virtual to physical boundary. Typical scenarios that
require virtual-to-physical communication includes 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
GRE aware. IPsec Security Associations (SA) 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.

4.9. Network Scalability with GRE

One of the key benefits of using GRE is the IP address scalability
and in turn MAC address table scalability that can be achieved.
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 a NVGRE endpoint in a hypervisor it is
possible to scale significantly. This framework allows for location
information to be preconfigured inside a NVGRE endpoint 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 efficiently
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.

6. IANA Considerations

None.

7. References

7.1. Normative References

[RFC 2119] Bradner, S., "Key words for use in RFCs to Indicate
     Requirement Levels", BCP 14, RFC 2119, March 1997.

[ETHTYPES] ftp://ftp.isi.edu/in-notes/iana/assignments/ethernet-
numbers

7.2. Informative References

[VL2] A. Greenberg et al, "VL2: A Scalable and Flexible Data Center
     Network", Proc. SIGCOMM 2009.

[COST-CCR] A. Greenberg et al, "The Cost of a Cloud: Research
          Problems in the Data Center", ACM SIGCOMM Computer
          Communication Review.

8. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

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

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

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

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

   Geng Lin
   Dell
   One Dell Way
   Round Rock, TX 78682
   Email: geng_lin@dell.com

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

   Patricia Thaler
   Broadcom Corporation
   3151 Zanker Road

San Jose, CA 95134
Email: pthaler@broadcom.com

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

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

Yu-Shun Wang
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
Email: yushwang@microsoft.com