

Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: February 11, 2013

T. Narten, Ed.  
IBM  
D. Black  
EMC  
D. Dutt  
  
L. Fang  
Cisco Systems  
E. Gray  
Ericsson  
L. Kreeger  
Cisco  
M. Napierala  
AT&T  
M. Sridharan  
Microsoft  
August 10, 2012

Problem Statement: Overlays for Network Virtualization  
draft-narten-nvo3-overlay-problem-statement-04

## Abstract

This document describes issues associated with providing multi-tenancy in large data center networks that require an overlay-based network virtualization approach to addressing them. A key multi-tenancy requirement is traffic isolation, so that a tenant's traffic is not visible to any other tenant. This isolation can be achieved by assigning one or more virtual networks to each tenant such that traffic within a virtual network is isolated from traffic in other virtual networks. The primary functionality required is provisioning virtual networks, associating a virtual machine's virtual network interface(s) with the appropriate virtual network, and maintaining that association as the virtual machine is activated, migrated and/or deactivated. Use of an overlay-based approach enables scalable deployment on large network infrastructures.

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Draft

Overlays for Network Virtualization

August 2012

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 11, 2013.

#### Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

<a href="#">1.</a>	Introduction . . . . .	<a href="#">4</a>
<a href="#">2.</a>	Problem Areas . . . . .	<a href="#">5</a>
<a href="#">2.1.</a>	Need For Dynamic Provisioning . . . . .	<a href="#">5</a>
<a href="#">2.2.</a>	Virtual Machine Mobility Limitations . . . . .	<a href="#">6</a>
<a href="#">2.3.</a>	Inadequate Forwarding Table Sizes in Switches . . . . .	<a href="#">6</a>
<a href="#">2.4.</a>	Need to Decouple Logical and Physical Configuration . . . . .	<a href="#">7</a>
<a href="#">2.5.</a>	Need For Address Separation Between Tenants . . . . .	<a href="#">7</a>
2.6.	Need For Address Separation Between Tenant and Infrastructure . . . . .	<a href="#">7</a>
<a href="#">2.7.</a>	IEEE 802.1 VLAN Limitations . . . . .	<a href="#">8</a>
<a href="#">3.</a>	Network Overlays . . . . .	<a href="#">8</a>
<a href="#">3.1.</a>	Benefits of Network Overlays . . . . .	<a href="#">9</a>
<a href="#">3.2.</a>	Communication Between Virtual and Traditional Networks . . . . .	<a href="#">10</a>
<a href="#">3.3.</a>	Communication Between Virtual Networks . . . . .	<a href="#">11</a>
<a href="#">3.4.</a>	Overlay Design Characteristics . . . . .	<a href="#">11</a>
<a href="#">3.5.</a>	Overlay Networking Work Areas . . . . .	<a href="#">12</a>
<a href="#">4.</a>	Related IETF and IEEE Work . . . . .	<a href="#">14</a>
<a href="#">4.1.</a>	L3 BGP/MPLS IP VPNs . . . . .	<a href="#">14</a>
<a href="#">4.2.</a>	L2 BGP/MPLS IP VPNs . . . . .	<a href="#">15</a>
<a href="#">4.3.</a>	IEEE 802.1aq - Shortest Path Bridging . . . . .	<a href="#">15</a>
<a href="#">4.4.</a>	ARMD . . . . .	<a href="#">15</a>
<a href="#">4.5.</a>	TRILL . . . . .	<a href="#">15</a>
<a href="#">4.6.</a>	L2VPNs . . . . .	<a href="#">16</a>
<a href="#">4.7.</a>	Proxy Mobile IP . . . . .	<a href="#">16</a>
<a href="#">4.8.</a>	LISP . . . . .	<a href="#">16</a>
<a href="#">5.</a>	Further Work . . . . .	<a href="#">16</a>
<a href="#">6.</a>	Summary . . . . .	<a href="#">17</a>
<a href="#">7.</a>	Acknowledgments . . . . .	<a href="#">17</a>
<a href="#">8.</a>	IANA Considerations . . . . .	<a href="#">17</a>
<a href="#">9.</a>	Security Considerations . . . . .	<a href="#">17</a>
<a href="#">10.</a>	Informative References . . . . .	<a href="#">17</a>
<a href="#">Appendix A.</a>	Change Log . . . . .	<a href="#">19</a>
<a href="#">A.1.</a>	Changes from -01 . . . . .	<a href="#">19</a>
<a href="#">A.2.</a>	Changes from -02 . . . . .	<a href="#">19</a>

<a href="#">A.3.</a> Changes from -03 . . . . .	<a href="#">20</a>
Authors' Addresses . . . . .	<a href="#">20</a>

## [1.](#) Introduction

Data Centers are increasingly being consolidated and outsourced in an effort, both to improve the deployment time of applications as well as reduce operational costs. This coincides with an increasing demand for compute, storage, and network resources from applications. In order to scale compute, storage, and network resources, physical resources are being abstracted from their logical representation, in what is referred to as server, storage, and network virtualization. Virtualization can be implemented in various layers of computer systems or networks

The demand for server virtualization is increasing in data centers. With server virtualization, each physical server supports multiple virtual machines (VMs), each running its own operating system, middleware and applications. Virtualization is a key enabler of workload agility, i.e., allowing any server to host any application and providing the flexibility of adding, shrinking, or moving services within the physical infrastructure. Server virtualization provides numerous benefits, including higher utilization, increased security, reduced user downtime, reduced power usage, etc.

Multi-tenant data centers are taking advantage of the benefits of server virtualization to provide a new kind of hosting, a virtual hosted data center. Multi-tenant data centers are ones where individual tenants could belong to a different company (in the case of a public provider) or a different department (in the case of an internal company data center). Each tenant has the expectation of a level of security and privacy separating their resources from those

of other tenants. For example, one tenant's traffic must never be exposed to another tenant, except through carefully controlled interfaces, such as a security gateway.

To a tenant, virtual data centers are similar to their physical counterparts, consisting of end stations attached to a network, complete with services such as load balancers and firewalls. But unlike a physical data center, end stations connect to a virtual network. To end stations, a virtual network looks like a normal network (e.g., providing an ethernet or L3 service), except that the only end stations connected to the virtual network are those belonging to a tenant's specific virtual network.

A tenant is the administrative entity that is responsible for and manages a specific virtual network instance and its associated services (whether virtual or physical). In a cloud environment, a tenant would correspond to the customer that has defined and is using a particular virtual network. However, a tenant may also find it useful to create multiple different virtual network instances.

Hence, there is a one-to-many mapping between tenants and virtual network instances. A single tenant may operate multiple individual virtual network instances, each associated with a different service.

How a virtual network is implemented does not generally matter to the tenant; what matters is that the service provided (L2 or L3) has the right semantics, performance, etc. It could be implemented via a pure routed network, a pure bridged network or a combination of bridged and routed networks. A key requirement is that each individual virtual network instance be isolated from other virtual network instances.

For data center virtualization, two key issues must be addressed. First, address space separation between tenants must be supported. Second, it must be possible to place (and migrate) VMs anywhere in the data center, without restricting VM addressing to match the subnet boundaries of the underlying data center network.

This document outlines the problems encountered in scaling the number of isolated networks in a data center, as well as the problems of managing the creation/deletion, membership and span of these networks and makes the case that an overlay based approach, where individual

networks are implemented within individual virtual networks that are dynamically controlled by a standardized control plane provides a number of advantages over current approaches. The purpose of this document is to identify the set of problems that any solution has to address in building multi-tenant data centers. With this approach, the goal is to allow the construction of standardized, interoperable implementations to allow the construction of multi-tenant data centers.

[Section 2](#) describes the problem space details. [Section 3](#) describes overlay networks in more detail. Sections [4](#) and [5](#) review related and further work, while [Section 6](#) closes with a summary.

## [2.](#) Problem Areas

The following subsections describe aspects of multi-tenant data center networking that pose problems for network infrastructure. Different problem aspects may arise based on the network architecture and scale.

### [2.1.](#) Need For Dynamic Provisioning

Cloud computing involves on-demand provisioning of resources for multi-tenant environments. A common example of cloud computing is the public cloud, where a cloud service provider offers elastic

services to multiple customers over the same infrastructure. In current systems, it can be difficult to provision resources for individual tenants in such a way that provisioned properties migrate automatically when services are dynamically moved around within the data center to optimize workloads.

### [2.2.](#) Virtual Machine Mobility Limitations

A key benefit of server virtualization is virtual machine (VM) mobility. A VM can be migrated from one server to another, live, i.e., while continuing to run and without needing to shut it down and restart it at the new location. A key requirement for live migration is that a VM retain critical network state at its new location, including its IP and MAC address(es). Preservation of MAC addresses may be necessary, for example, when software licenses are bound to

MAC addresses. More generally, any change in the VM's MAC addresses resulting from a move would be visible to the VM and thus potentially result in unexpected disruptions. Retaining IP addresses after a move is necessary to prevent existing transport connections (e.g., TCP) from breaking and needing to be restarted.

In traditional data centers, servers are assigned IP addresses based on their physical location, for example based on the Top of Rack (ToR) switch for the server rack or the VLAN configured to the server. Servers can only move to other locations within the same IP subnet. This constraint is not problematic for physical servers, which move infrequently, but it restricts the placement and movement of VMs within the data center. Any solution for a scalable multi-tenant data center must allow a VM to be placed (or moved) anywhere within the data center, without being constrained by the subnet boundary concerns of the host servers.

### [2.3.](#) Inadequate Forwarding Table Sizes in Switches

Today's virtualized environments place additional demands on the forwarding tables of switches in the physical infrastructure. Instead of just one link-layer address per server, the switching infrastructure has to learn addresses of the individual VMs (which could range in the 100s per server). This is a requirement since traffic from/to the VMs to the rest of the physical network will traverse the physical network infrastructure. This places a much larger demand on the switches' forwarding table capacity compared to non-virtualized environments, causing more traffic to be flooded or dropped when the number of addresses in use exceeds a switch's forwarding table capacity.

### [2.4.](#) Need to Decouple Logical and Physical Configuration

Data center operators must be able to achieve high utilization of server and network capacity. For efficient and flexible allocation, operators should be able to spread a virtual network instance across servers in any rack in the data center. It should also be possible to migrate compute workloads to any server anywhere in the network while retaining the workload's addresses. In networks using VLANs,

moving servers elsewhere in the network may require expanding the scope of the VLAN beyond its original boundaries. While this can be done, it requires potentially complex network configuration changes and can conflict with the desire to bound the size of broadcast domains, especially in larger data centers.

However, in order to limit the broadcast domain of each VLAN, multi-destination frames within a VLAN should optimally flow only to those devices that have that VLAN configured. When workloads migrate, the physical network (e.g., access lists) may need to be reconfigured which is typically time consuming and error prone.

An important use case is cross-pod expansion. A pod typically consists of one or more racks of servers with its associated network and storage connectivity. A tenant's virtual network may start off on a pod and, due to expansion, require servers/VMs on other pods, especially the case when other pods are not fully utilizing all their resources. This use case requires that virtual networks span multiple pods in order to provide connectivity to all of its tenant's servers/VMs. Such expansion can be difficult to achieve when tenant addressing is tied to the addressing used by the underlay network or when it requires that the scope of the underlying L2 VLAN expand beyond its original pod boundary.

#### [2.5.](#) Need For Address Separation Between Tenants

Individual tenants need control over the addresses they use within a virtual network. But it can be problematic when different tenants want to use the same addresses, or even if the same tenant wants to reuse the same addresses in different virtual networks. Consequently, virtual networks must allow tenants to use whatever addresses they want without concern for what addresses are being used by other tenants or other virtual networks.

#### [2.6.](#) Need For Address Separation Between Tenant and Infrastructure

As in the previous case, a tenant needs to be able to use whatever addresses it wants in a virtual network independent of what addresses the underlying data center network is using. Tenants (and the underlay infrastructure provider) should be able use whatever

addresses make sense for them, without having to worry about address



collisions between addresses used by tenants and those used by the underlay data center network.

### [2.7.](#) IEEE 802.1 VLAN Limitations

VLANs are a well known construct in the networking industry, providing an L2 service via an L2 underlay. A VLAN is an L2 bridging construct that provides some of the semantics of virtual networks mentioned above: a MAC address is unique within a VLAN, but not necessarily across VLANs. Traffic sourced within a VLAN (including broadcast and multicast traffic) remains within the VLAN it originates from. Traffic forwarded from one VLAN to another typically involves router (L3) processing. The forwarding table look up operation is keyed on {VLAN, MAC address} tuples.

But there are problems and limitations with L2 VLANs. VLANs are a pure L2 bridging construct and VLAN identifiers are carried along with data frames to allow each forwarding point to know what VLAN the frame belongs to. A VLAN today is defined as a 12 bit number, limiting the total number of VLANs to 4096 (though typically, this number is 4094 since 0 and 4095 are reserved). Due to the large number of tenants that a cloud provider might service, the 4094 VLAN limit is often inadequate. In addition, there is often a need for multiple VLANs per tenant, which exacerbates the issue. The use of a sufficiently large VNID, present in the overlay control plane and possibly also in the dataplane would eliminate current VLAN size limitations associated with single 12-bit VLAN tags.

### [3.](#) Network Overlays

Virtual Networks are used to isolate a tenant's traffic from that of other tenants (or even traffic within the same tenant that requires isolation). There are two main characteristics of virtual networks:

1. Virtual networks isolate the address space used in one virtual network from the address space used by another virtual network. The same network addresses may be used in different virtual networks at the same time. In addition, the address space used by a virtual network is independent from that used by the underlying physical network.
2. Virtual Networks limit the scope of packets sent on the virtual network. Packets sent by end systems attached to a virtual network are delivered as expected to other end systems on that virtual network and may exit a virtual network only through controlled exit points such as a security gateway. Likewise,

---

packets sourced from outside of the virtual network may enter the virtual network only through controlled entry points, such as a security gateway.

### [3.1.](#) Benefits of Network Overlays

To address the problems described in [Section 2](#), a network overlay model can be used.

The idea behind an overlay is quite straightforward. Each virtual network instance is implemented as an overlay. The original packet is encapsulated by the first-hop network device. The encapsulation identifies the destination of the device that will perform the decapsulation before delivering the original packet to the endpoint. The rest of the network forwards the packet based on the encapsulation header and can be oblivious to the payload that is carried inside.

Overlays are based on what is commonly known as a "map-and-encap" architecture. There are three distinct and logically separable steps:

1. The first-hop overlay device implements a mapping operation that determines where the encapsulated packet should be sent to reach its intended destination VM. Specifically, the mapping function maps the destination address (either L2 or L3) of a packet received from a VM into the corresponding destination address of the egress device. The destination address will be the underlay address of the device doing the decapsulation and is an IP address.
2. Once the mapping has been determined, the ingress overlay device encapsulates the received packet within an overlay header.
3. The final step is to actually forward the (now encapsulated) packet to its destination. The packet is forwarded by the underlay (i.e., the IP network) based entirely on its outer address. Upon receipt at the destination, the egress overlay device decapsulates the original packet and delivers it to the intended recipient VM.

Each of the above steps is logically distinct, though an implementation might combine them for efficiency or other reasons. It should be noted that in L3 BGP/VPN terminology, the above steps are commonly known as "forwarding" or "virtual forwarding".

The first hop network device can be a traditional switch or router or the virtual switch residing inside a hypervisor. Furthermore, the

endpoint can be a VM or it can be a physical server. Examples of architectures based on network overlays include BGP/MPLS VPNs [[RFC4364](#)], TRILL [[RFC6325](#)], LISP [[I-D.ietf-lisp](#)], and Shortest Path Bridging (SPB-M) [[SPBM](#)].

In the data plane, a virtual network identifier (or VNID), or a locally significant identifier, can be carried as part of the overlay header so that every data packet explicitly identifies the specific virtual network the packet belongs to. Since both routed and bridged semantics can be supported by a virtual data center, the original packet carried within the overlay header can be an Ethernet frame complete with MAC addresses or just the IP packet.

The use of a sufficiently large VNID would address current VLAN limitations associated with single 12-bit VLAN tags. This VNID can be carried in the control plane. In the data plane, an overlay header provides a place to carry either the VNID, or an identifier that is locally-significant to the edge device. In both cases, the identifier in the overlay header specifies which virtual network the data packet belongs to.

A key aspect of overlays is the decoupling of the "virtual" MAC and/or IP addresses used by VMs from the physical network infrastructure and the infrastructure IP addresses used by the data center. If a VM changes location, the overlay edge devices simply update their mapping tables to reflect the new location of the VM within the data center's infrastructure space. Because an overlay network is used, a VM can now be located anywhere in the data center that the overlay reaches without regards to traditional constraints implied by L2 properties such as VLAN numbering, or the span of an L2 broadcast domain scoped to a single pod or access switch.

Multi-tenancy is supported by isolating the traffic of one virtual network instance from traffic of another. Traffic from one virtual network instance cannot be delivered to another instance without (conceptually) exiting the instance and entering the other instance via an entity that has connectivity to both virtual network instances. Without the existence of this entity, tenant traffic remains isolated within each individual virtual network instance.

Overlays are designed to allow a set of VMs to be placed within a single virtual network instance, whether that virtual network provides a bridged network or a routed network.

### [3.2.](#) Communication Between Virtual and Traditional Networks

Not all communication will be between devices connected to virtualized networks. Devices using overlays will continue to access

Narten, et al.

Expires February 11, 2013

[Page 10]

---

Internet-Draft

Overlays for Network Virtualization

August 2012

devices and make use of services on traditional, non-virtualized networks, whether in the data center, the public Internet, or at remote/branch campuses. Any virtual network solution must be capable of interoperating with existing routers, VPN services, load balancers, intrusion detection services, firewalls, etc. on external networks.

Communication between devices attached to a virtual network and devices connected to non-virtualized networks is handled architecturally by having specialized gateway devices that receive packets from a virtualized network, decapsulate them, process them as regular (i.e., non-virtualized) traffic, and finally forward them on to their appropriate destination (and vice versa). Additional identification, such as VLAN tags, could be used on the non-virtualized side of such a gateway to enable forwarding of traffic for multiple virtual networks over a common non-virtualized link.

A wide range of implementation approaches are possible. Overlay gateway functionality could be combined with other network functionality into a network device that implements the overlay functionality, and then forwards traffic between other internal components that implement functionality such as full router service, load balancing, firewall support, VPN gateway, etc.

### [3.3.](#) Communication Between Virtual Networks

Communication between devices on different virtual networks is handled architecturally by adding specialized interconnect functionality among the otherwise isolated virtual networks. For a virtual network providing an L2 service, such interconnect functionality could be IP forwarding configured as part of the "default gateway" for each virtual network. For a virtual network

providing L3 service, the interconnect functionality could be IP forwarding configured as part of routing between IP subnets or it can be based on configured inter-virtual network traffic policies. In both cases, the implementation of the interconnect functionality could be distributed across the NVEs, and could be combined with other network functionality (e.g., load balancing, firewall support) that is applied to traffic that is forwarded between virtual networks.

### [3.4.](#) Overlay Design Characteristics

There are existing layer 2 and layer 3 overlay protocols in existence, but they do not necessarily solve all of today's problem in the environment of a highly virtualized data center. Below are some of the characteristics of environments that must be taken into account by the overlay technology:

Narten, et al.

Expires February 11, 2013

[Page 11]

---

Internet-Draft

Overlays for Network Virtualization

August 2012

1. Highly distributed systems. The overlay should work in an environment where there could be many thousands of access devices (e.g. residing within the hypervisors) and many more end systems (e.g. VMs) connected to them. This leads to a distributed mapping system that puts a low overhead on the overlay tunnel endpoints.
2. Many highly distributed virtual networks with sparse membership. Each virtual network could be highly dispersed inside the data center. Also, along with expectation of many virtual networks, the number of end systems connected to any one virtual network is expected to be relatively low; Therefore, the percentage of access devices participating in any given virtual network would also be expected to be low. For this reason, efficient delivery of multi-destination traffic within a virtual network instance should be taken into consideration.
3. Highly dynamic end systems. End systems connected to virtual networks can be very dynamic, both in terms of creation/deletion/power-on/off and in terms of mobility across the access devices.
4. Work with existing, widely deployed network Ethernet switches and IP routers without requiring wholesale replacement. The first hop device (or end system) that adds and removes the overlay header will require new equipment and/or new software.

5. Work with existing data center network deployments without requiring major changes in operational or other practices. For example, some data centers have not enabled multicast beyond link-local scope. Overlays should be capable of leveraging underlay multicast support where appropriate, but not require its enablement in order to use an overlay solution.
6. Network infrastructure administered by a single administrative domain. This is consistent with operation within a data center, and not across the Internet.

### [3.5.](#) Overlay Networking Work Areas

There are three specific and separate potential work areas needed to realize an overlay solution. The areas correspond to different possible "on-the-wire" protocols, where distinct entities interact with each other.

One area of work concerns the address dissemination protocol an NVE uses to build and maintain the mapping tables it uses to deliver encapsulated packets to their proper destination. One approach is to build mapping tables entirely via learning (as is done in 802.1

networks). But to provide better scaling properties, a more sophisticated approach is needed, i.e., the use of a specialized control plane protocol. While there are some advantages to using or leveraging an existing protocol for maintaining mapping tables, the fact that large numbers of NVE's will likely reside in hypervisors places constraints on the resources (cpu and memory) that can be dedicated to such functions. For example, routing protocols (e.g., IS-IS, BGP) may have scaling difficulties if implemented directly in all NVEs, based on both flooding and convergence time concerns. An alternative approach would be to use a standard query protocol between NVEs and the set of network nodes that maintain address mappings used across the data center for the entire overlay system.

From an architectural perspective, one can view the address mapping dissemination problem as having two distinct and separable components. The first component consists of a back-end "oracle" that is responsible for distributing and maintaining the mapping information for the entire overlay system. The second component

consists of the on-the-wire protocols an NVE uses when interacting with the oracle.

The back-end oracle could provide high performance, high resiliency, failover, etc. and could be implemented in significantly different ways. For example, one model uses a traditional, centralized "directory-based" database, using replicated instances for reliability and failover. A second model involves using and possibly extending an existing routing protocol (e.g., BGP, IS-IS, etc.). To support different architectural models, it is useful to have one standard protocol for the NVE-oracle interaction while allowing different protocols and architectural approaches for the oracle itself. Separating the two allows NVEs to transparently interact with different types of oracles, i.e., either of the two architectural models described above. Having separate protocols could also allow for a simplified NVE that only interacts with the oracle for the mapping table entries it needs and allows the oracle (and its associated protocols) to evolve independently over time with minimal impact to the NVEs.

A third work area considers the attachment and detachment of VMs (or Tenant End Systems [[I-D.lasserre-nvo3-framework](#)] more generally) from a specific virtual network instance. When a VM attaches, the Network Virtualization Edge (NVE) [[I-D.lasserre-nvo3-framework](#)] associates the VM with a specific overlay for the purposes of tunneling traffic sourced from or destined to the VM. When a VM disconnects, it is removed from the overlay and the NVE effectively terminates any tunnels associated with the VM. To achieve this functionality, a standardized interaction between the NVE and hypervisor may be needed, for example in the case where the NVE resides on a separate

device from the VM.

In summary, there are three areas of potential work. The first area concerns the oracle itself and any on-the-wire protocols it needs. A second area concerns the interaction between the oracle and NVEs. The third work area concerns protocols associated with attaching and detaching a VM from a particular virtual network instance. All three work areas are important to the development of scalable, interoperable solutions.

## [4.](#) Related IETF and IEEE Work

The following subsections discuss related IETF and IEEE work in progress, the items are not meant to be complete coverage of all IETF and IEEE data center related work, nor are the descriptions comprehensive. Each area is currently trying to address certain limitations of today's data center networks, e.g., scaling is a common issue for every area listed and multi-tenancy and VM mobility are important focus areas as well. Comparing and evaluating the work result and progress of each work area listed is out of scope of this document. The intent of this section is to provide a reference to the interested readers.

### [4.1.](#) L3 BGP/MPLS IP VPNs

BGP/MPLS IP VPNs [[RFC4364](#)] support multi-tenancy address overlapping, VPN traffic isolation, and address separation between tenants and network infrastructure. The BGP/MPLS control plane is used to distribute the VPN labels and the tenant IP addresses which identify the tenants (or to be more specific, the particular VPN/VN) and tenant IP addresses. Deployment of enterprise L3 VPNs has been shown to scale to thousands of VPNs and millions of VPN prefixes. BGP/MPLS IP VPNs are currently deployed in some large enterprise data centers. The potential limitation for deploying BGP/MPLS IP VPNs in data center environments is the practicality of using BGP in the data center, especially reaching into the servers or hypervisors. There may be computing work force skill set issues, equipment support issues, and potential new scaling challenges. A combination of BGP and lighter weight IP signaling protocols, e.g., XMPP, have been proposed to extend the solutions into DC environment [[I-D.margues-end-system](#)], while taking advantage of building in VPN features with its rich policy support; it is especially useful for inter-tenant connectivity.

### [4.2.](#) L2 BGP/MPLS IP VPNs

Ethernet Virtual Private Networks (E-VPNs) [[I-D.ietf-l2vpn-evpn](#)] provide an emulated L2 service in which each tenant has its own



Ethernet network over a common IP or MPLS infrastructure and a BGP/MPLS control plane is used to distribute the tenant MAC addresses and the MPLS labels that identify the tenants and tenant MAC addresses. Within the BGP/MPLS control plane a thirty two bit Ethernet Tag is used to identify the broadcast domains (VLANs) associated with a given L2 VLAN service instance and these Ethernet tags are mapped to VLAN IDs understood by the tenant at the service edges. This means that the limit of 4096 VLANs is associated with an individual tenant service edge, enabling a much higher level of scalability. Interconnection between tenants is also allowed in a controlled fashion.

VM Mobility [[I-D.raggarwa-data-center-mobility](#)] introduces the concept of a combined L2/L3 VPN service in order to support the mobility of individual Virtual Machines (VMs) between Data Centers connected over a common IP or MPLS infrastructure.

#### [4.3.](#) IEEE 802.1aq - Shortest Path Bridging

Shortest Path Bridging (SPB-M) is an IS-IS based overlay for L2 Ethernets. SPB-M supports multi-pathing and addresses a number of shortcomings in the original Ethernet Spanning Tree Protocol. SPB-M uses IEEE 802.1ah MAC-in-MAC encapsulation and supports a 24-bit I-SID, which can be used to identify virtual network instances. SPB-M is entirely L2 based, extending the L2 Ethernet bridging model.

#### [4.4.](#) ARMD

ARMD is chartered to look at data center scaling issues with a focus on address resolution. ARMD is currently chartered to develop a problem statement and is not currently developing solutions. While an overlay-based approach may address some of the "pain points" that have been raised in ARMD (e.g., better support for multi-tenancy), an overlay approach may also push some of the L2 scaling concerns (e.g., excessive flooding) to the IP level (flooding via IP multicast). Analysis will be needed to understand the scaling tradeoffs of an overlay based approach compared with existing approaches. On the other hand, existing IP-based approaches such as proxy ARP may help mitigate some concerns.

#### [4.5.](#) TRILL

TRILL is an L2-based approach aimed at improving deficiencies and limitations with current Ethernet networks and STP in particular.

Although it differs from Shortest Path Bridging in many architectural and implementation details, it is similar in that it provides an L2-based service to end systems. TRILL as defined today, supports only the standard (and limited) 12-bit VLAN model. Approaches to extend TRILL to support more than 4094 VLANs are currently under investigation [[I-D.ietf-trill-fine-labeling](#)]

#### [4.6.](#) L2VPNs

The IETF has specified a number of approaches for connecting L2 domains together as part of the L2VPN Working Group. That group, however has historically been focused on Provider-provisioned L2 VPNs, where the service provider participates in management and provisioning of the VPN. In addition, much of the target environment for such deployments involves carrying L2 traffic over WANs. Overlay approaches are intended to be used within data centers where the overlay network is managed by the data center operator, rather than by an outside party. While overlays can run across the Internet as well, they will extend well into the data center itself (e.g., up to and including hypervisors) and include large numbers of machines within the data center itself.

Other L2VPN approaches, such as L2TP [[RFC2661](#)] require significant tunnel state at the encapsulating and decapsulating end points. Overlays require less tunnel state than other approaches, which is important to allow overlays to scale to hundreds of thousands of end points. It is assumed that smaller switches (i.e., virtual switches in hypervisors or the adjacent devices to which VMs connect) will be part of the overlay network and be responsible for encapsulating and decapsulating packets.

#### [4.7.](#) Proxy Mobile IP

Proxy Mobile IP [[RFC5213](#)] [[RFC5844](#)] makes use of the GRE Key Field [[RFC5845](#)] [[RFC6245](#)], but not in a way that supports multi-tenancy.

#### [4.8.](#) LISP

LISP[[I-D.ietf-lisp](#)] essentially provides an IP over IP overlay where the internal addresses are end station Identifiers and the outer IP addresses represent the location of the end station within the core IP network topology. The LISP overlay header uses a 24-bit Instance ID used to support overlapping inner IP addresses.

### [5.](#) Further Work

It is believed that overlay-based approaches may be able to reduce

the overall amount of flooding and other multicast and broadcast related traffic (e.g, ARP and ND) currently experienced within current data centers with a large flat L2 network. Further analysis is needed to characterize expected improvements.

There are a number of VPN approaches that provide some if not all of the desired semantics of virtual networks. A gap analysis will be needed to assess how well existing approaches satisfy the requirements.

## 6. Summary

This document has argued that network virtualization using overlays addresses a number of issues being faced as data centers scale in size. In addition, careful study of current data center problems is needed for development of proper requirements and standard solutions.

Three potential work were identified. The first involves the interaction that take place when a VM attaches or detaches from an overlay. A second involves the protocol an NVE would use to communicate with a backend "oracle" to learn and disseminate mapping information about the VMs the NVE communicates with. The third potential work area involves the backend oracle itself, i.e., how it provides failover and how it interacts with oracles in other domains.

## 7. Acknowledgments

Helpful comments and improvements to this document have come from John Drake, Ariel Hendel, Vinit Jain, Thomas Morin, Benson Schliesser and many others on the mailing list.

## 8. IANA Considerations

This memo includes no request to IANA.

## 9. Security Considerations

TBD

## 10. Informative References

[I-D.fang-vpn4dc-problem-statement]

Napierala, M., Fang, L., and D. Cai, "IP-VPN Data Center

Narten, et al.

Expires February 11, 2013

[Page 17]

---

Internet-Draft

Overlays for Network Virtualization

August 2012

Problem Statement and Requirements",

[draft-fang-vpn4dc-problem-statement-01](#) (work in progress),  
June 2012.

[I-D.ietf-l2vpn-evpn]

Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F.,  
Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN",  
[draft-ietf-l2vpn-evpn-01](#) (work in progress), July 2012.

[I-D.ietf-lisp]

Farinacci, D., Fuller, V., Meyer, D., and D. Lewis,  
"Locator/ID Separation Protocol (LISP)",  
[draft-ietf-lisp-23](#) (work in progress), May 2012.

[I-D.ietf-trill-fine-labeling]

Eastlake, D., Zhang, M., Agarwal, P., Perlman, R., and D.  
Dutt, "TRILL: Fine-Grained Labeling",  
[draft-ietf-trill-fine-labeling-01](#) (work in progress),  
June 2012.

[I-D.kreeger-nvo3-overlay-cp]

Kreeger, L., Dutt, D., Narten, T., Black, D., and M.  
Sridhavan, "Network Virtualization Overlay Control  
Protocol Requirements", [draft-kreeger-nvo3-overlay-cp-01](#)  
(work in progress), July 2012.

[I-D.lasserre-nvo3-framework]

Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y.  
Rekhter, "Framework for DC Network Virtualization",  
[draft-lasserre-nvo3-framework-03](#) (work in progress),  
July 2012.

[I-D.raggarwa-data-center-mobility]

Aggarwal, R., Rekhter, Y., Henderickx, W., Shekhar, R., and L. Fang, "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", [draft-raggarwa-data-center-mobility-03](#) (work in progress), June 2012.

- [RFC2661] Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., and B. Palter, "Layer Two Tunneling Protocol "L2TP"", [RFC 2661](#), August 1999.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", [RFC 4364](#), February 2006.
- [RFC5213] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", [RFC 5213](#), August 2008.

Narten, et al.

Expires February 11, 2013

[Page 18]

---

Internet-Draft

Overlays for Network Virtualization

August 2012

- [RFC5844] Wakikawa, R. and S. Gundavelli, "IPv4 Support for Proxy Mobile IPv6", [RFC 5844](#), May 2010.
- [RFC5845] Muhanna, A., Khalil, M., Gundavelli, S., and K. Leung, "Generic Routing Encapsulation (GRE) Key Option for Proxy Mobile IPv6", [RFC 5845](#), June 2010.
- [RFC6245] Yegani, P., Leung, K., Lior, A., Chowdhury, K., and J. Navali, "Generic Routing Encapsulation (GRE) Key Extension for Mobile IPv4", [RFC 6245](#), May 2011.
- [RFC6325] Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", [RFC 6325](#), July 2011.
- [SPBM] "IEEE P802.1aq/D4.5 Draft Standard for Local and Metropolitan Area Networks -- Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks, Amendment 8: Shortest Path Bridging", February 2012.

## [Appendix A](#). Change Log

### [A.1](#). Changes from -01

1. Removed [Section 4.2](#) (Standardization Issues) and [Section 5](#)

(Control Plane) as those are more appropriately covered in and overlap with material in [[I-D.lasserre-nvo3-framework](#)] and [[I-D.kreeger-nvo3-overlay-cp](#)].

2. Expanded introduction and better explained terms such as tenant and virtual network instance. These had been covered in a section that has since been removed.
3. Added [Section 3.3](#) "Overlay Networking Work Areas" to better articulate the three separable work components (or "on-the-wire protocols") where work is needed.
4. Added section on Shortest Path Bridging in Related Work section.
5. Revised some of the terminology to be consistent with [[I-D.lasserre-nvo3-framework](#)] and [[I-D.kreeger-nvo3-overlay-cp](#)].

#### [A.2.](#) Changes from -02

1. Numerous changes in response to discussions on the nvo3 mailing list, with majority of changes in [Section 2](#) (Problem Details) and [Section 3](#) (Network Overlays). Best to see diffs for specific

Narten, et al.

Expires February 11, 2013

[Page 19]

---

Internet-Draft

Overlays for Network Virtualization

August 2012

text changes.

#### [A.3.](#) Changes from -03

1. Too numerous to enumerate; moved solution-specific descriptions to Related Work section. Pulled in additional text (and authors) from from [[I-D.fang-vpn4dc-problem-statement](#)], numerous editorial improvements.

#### Authors' Addresses

Thomas Narten (editor)  
IBM

Email: narten@us.ibm.com

David Black

EMC

Email: david.black@emc.com

Dinesh Dutt

Email: ddutt.ietf@hobbesdutt.com

Luyuan Fang  
Cisco Systems  
111 Wood Avenue South  
Iselin, NJ 08830  
USA

Email: lufang@cisco.com

Eric Gray  
Ericsson

Email: eric.gray@ericsson.com

Narten, et al.

Expires February 11, 2013

[Page 20]

---

Internet-Draft

Overlays for Network Virtualization

August 2012

Lawrence Kreeger  
Cisco

Email: kreeger@cisco.com

Maria Napierala  
AT&T  
200 Laurel Avenue  
Middletown, NJ 07748  
USA

Email: mnapierala@att.com

Murari Sridharan  
Microsoft

Email: muraris@microsoft.com