Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: March 4, 2012

T. Narten, Ed.
IBM
M. Sridharan
Microsoft

September 2011

Problem Statement: Using L3 Overlays for Network Virtualization draft-narten-nvo3-overlay-problem-statement-00

Abstract

This document lays out the case for developing L3 overlays to provide network virtualization. In addition, the document describes what issues need to be resolved in order to produce an interoperable standard. The goal is lead to scalable, interoperable implementations of virtualization overlays based on standardized encapsulation methods and control plane approaches.

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-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 March 4, 2012.

Table of Contents

1. Introduction

Server virtualization is increasingly becoming the norm 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 data security, reduced user downtime, reduced power usage, etc.

Server virtualization is driving and accentuating scaling limitations in existing datacenter networks. Placement and movement of VMs in a network effectively requires that VM IP addresses be fixed and static. While a VM's IP address could be updated to always be consistent with the VM's current point of attachment, changing the VM's address makes it difficult for the VM to be reached by clients, and breaks any open TCP connections. Thus, from an IP perspective, VM's can generally only move within a single IP subnet. Having VMs move to different subnets (while retaining their previous IP addresses) requires taking additional steps at the IP routing level to ensure that traffic continues to reach the VM at its new location. In practice, this leads to a desire for larger and flatter L2 networks, so that a given VM can be placed (or moved to) anywhere within the datacenter, without being constrained by subnet boundary concerns.

The general scaling problems of large, flat L2 networks are well known. In traditional data center architectures using Spanning Trees, the size of a layer 2 domain is generally limited to two tiers (access and aggregation), limiting the number of hosts within a single L2 domain. Current network deployments, however, are experiencing additional and new pain points. The increase in both the number of physical machines, and the number of VMs per physical machine has lead to MAC address explosion whereby switches need to have increasingly large forwarding tables to handle the traffic they switch. Furthermore, the dynamic nature of VM creation, deletion and movement between between servers is leading to sub-optimal broadcast domain expansion. Finally, the 4094 VLAN limit is no longer sufficient in a shared infrastructure servicing multiple tenants.

This document outlines the problems encountered in scaling L2 networks in a datacenter and makes the case that an overlay based approach, where individual L2 networks are implemented within individual L3 "domains" provide a number of advantages over current approaches. The goal is lead to scalable, interoperable

implementations of virtualization overlays based on standardized encapsulation methods and control plane approaches.

Section 2 describes the problem space in detail. Section 3 provides a general discussion of overlays. Section 4 covers standardization issues and possible work areas. Section 5 summarizes the benefits of an overlay approach. Section 6 and 7 discuss related work and further work.

2. Problem Details

2.1. Limitations Imposed by Spanning Tree and VLAN Spaces

Years of operational experience with Spanning Tree Protocol has led to best practices that limit topologies to two tiers, access and aggregation, with the STP root at the aggregation switch. Such topologies limit both the number of servers that can be connected to the same layer 2 domain, and the amount of East-West bandwidth available between servers connected to different access switches. Newer initiatives like Trill overcome these topology restrictions and allow high cross-sectional bandwidth between servers as well as large layer 2 domains, but require new networking equipment. Another way to expand the layer 2 domains, but stay within the topology restrictions imposed by using STP, is to use Overlay Transport Virtualization [I-D.hasmit-otv] to interconnect the individual STP topologies.

Yet another approach for increasing East-West bandwidth is to constrain the layer 2 domain to stay within the access switch, and terminate layer 3 in the access switch as well. All communications between servers connected to different access switches happens over IP. However, this is incompatible with the desire to be able to move VMs anywhere within the datacenter while maintaining Layer 2 adjacency between the VMs.

Another characteristic of Layer 2 data center networks is their use of Virtual LANs (VLANs) to provide broadcast isolation. A 12-bit VLAN ID is used in the Ethernet data frames to divide the larger Layer 2 network into multiple broadcast domains. VLANs have worked well in smaller data centers which are limited to less than 4094 VLANs. The growing demand for multitenancy, e.g., for clouds, further accelerate the need for larger VLAN limits, as discussed below. Finally, some switching equipment supports less than the 4094 maximum limit of VLANs.

2.2. Multitenant Environments

Cloud computing involves on-demand elastic provisioning of resources for multitenant environments. The most common example of cloud computing is the public cloud, where a cloud service provider offers these elastic services to multiple customers over the same infrastructure.

Isolation of network traffic by tenant could be done via Layer 2 or Layer 3 networks. For Layer 2 networks, VLANs are often used to segregate traffic - so a tenant could be identified by its own VLAN, for example. 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. Note that there is a proposal in the Trill working group to increase the VLAN space from 12 to 24 bits using two concatenated 12-bit tags [I-D.eastlake-trill-rbridge-fine-labeling]. Additionally, IEEE 802.1aq has defined double tagging.

Layer 3 networks are not a complete solution for multi tenancy either. Two tenants might use the same set of Layer 3 addresses within their networks which requires the cloud provider to provide isolation in some other form. Further, requiring all tenants to use IP excludes customers relying on direct Layer 2 or non-IP Layer 3 protocols for inter VM communication, which puts limitations on moving some applications onto a virtualized server environment.

2.3. Stretching of L2 Domain

Another use case is cross pod expansion. A pod typically consists of one or more racks of servers with its associated network and storage connectivity. Tenants may start off on a pod and, due to expansion, require servers/VMs on other pods, especially the case when tenants on the other pods are not fully utilizing all their resources. This use case requires a "stretched" Layer 2 environment connecting the individual servers/VMs.

2.4. Inadequate MAC Table Sizes in Switches

Today's virtualized environments place additional demands on the MAC address tables of layer 2 switches. Instead of just one MAC address per server link, the switching infrastructure now has to learn the MAC 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 switched infrastructure. This places a much larger demand on the switches' MAC table capacity compared to non-virtualized environments.

If the table overflows, the switches do not learn new entries until idle entries age out. This leads to flooding the frames over the entire VLAN in accordance with IEEE standards.

2.5. Decoupling Logical and Physical Configuration

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 (e.g., by using Trill or OTV). However, in order to limit the broadcast domain of each VLAN, all VLANs should not be configured to flow to every server in the datacenter. When workloads migrate, the physical network (e.g., server access lists) need to be reconfigured which is typically time consuming and error prone. By decoupling the network addresses used by the VMs when communicating with each other from the network addresses of the servers they are currently hosted on, the network administrator can configure the network once and not every time a service migrates. This decoupling enables any server to become part of any server resource pool.

3. Overlay Network Framework

The idea behind overlays is straightforward. Take the set of machines that are allowed to communicate with each other and group them into a high-level construct called a domain. A domain could be one L2 VLAN, a single IP subnet, or just an arbitrary collection of machines. The domain identifies the set of machines that are allowed to communicate with each other directly, and provides isolation from machines not within the same domain. The overlay connects the machines of a particular domain together. A switch connects each machine to its domain, accepting ethernet frames from attached VMs and encapsulating them for transport across the IP overlay. An egress switch decapsulates the frame and delivers it to the target VM.

3.1. Overlay Design Characteristics

There are existing layer 2 overlay protocols in existence, but they were not necessarily designed to solve the problem in the environment of a highly virtualized datacenter. Below are some of the characteristics of environments that must be taken into account by the overlay technology:

1. Highly distributed systems. The overlay should work in an environment where there could be many thousands of access switches (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 segments with sparse connectivity. Each overlay segment could be highly dispersed inside the datacenter. Also, along with expectation of many overlay segments, the number of end systems connected to any one segment is expected to be relatively low; Therefore, the percentage of access switches participating in any one overlay segment would also be expected to be low.

3. Highly dynamic end systems. End systems connected to segments can be very dynamic, both in terms of creation/deletion/power-on/off and in terms of mobility across the access switches.

4. Work with existing, widely deployed network Ethernet switches and IP routers without requiring wholesale replacement.

5. Network infrastructure administered by a single administrative domain. This is consistent with operation within a datacenter, and not across the internet.

6. Low access switch overhead / simple implementation. With the requirement to support very large numbers of access switches, the resource requirements on each switch should not be intensive both in terms of memory footprint or processing cycles. This also means consideration for hardware offload.

4. Standardization Issues for Overlay Networks

To provide a robust and interoperable overlay solution, a number of issues need to be considered. First, an overlay header is needed for transporting encapsulated Ethernet frames across the IP network to their ultimate destination within a specific domain. To provide multi-tenancy, the overlay header needs a field to identify which domain an encapsulated packet belongs to. Consequently, some sort of Domain Identifier is needed. VXLAN [I-D.mahalingam-dutt-dcops-vxlan] uses a 24-bit VXLAN Network Identifier (VNI), while NVGRE [I-D.sridharan-virtualization-nvgre] uses a 24-bit Tenant Network Identifier (TNI).

The details of a specific overlay header format need to be worked out. Questions to be resolved include whether to use existing

standard formats such as GRE [RFC2784] [RFC2890], or to define one specifically tailored to meet the requirements of an overlay network. One issue concerns whether to use UDP (in order to facilitate transport through middlebox devices) or to build directly on top of IP as GRE does. If the primary deployment environment is datacenters, and paths will not generally cross the public internet, middlebox traversal may not be a significant concern. Additionally, the encapsulated payload could include a full Ethernet header, including source and destination MAC addresses, VLAN information, etc., or some subset thereof. Finally, the encapsulation will need to consider whether inclusion of a checksum is necessary or whether it imposes unacceptable overhead (when encapsulated packets are themselves IP, they will carry their own end-to-end checksums). Note that GRE supports an optional checksum. In IPv4, UDP checksums can be disabled by setting the UDP checksum field to zero, but checksums must always be included in IPv6. The 6man working group is currently considering relaxing the IPv6 UDP checksum requirement [I-D.ietf-6man-udpzero].

Separate from encapsulation, an address mapping system is needed to map the destination address as specified by the originating VM into the egress IP address of the router to which the Ethernet frame will be tunneled. VXLAN uses a "learning" approach for this, similar to what L2 bridges use. NVGRE has not indicated how it proposes to perform address mapping, leaving details for a later document. Other approaches are possible, such as managing mappings in a centralized mapping system. Use of a centralized mapping system would require development of a protocol for distributing address mappings from the controller to the switches where encapsulation takes place.

Another aspect of address mapping concerns the handling of broadcast and multicast frames, or the delivery of unicast packets when no mapping exists. One approach is to flood such frames to all machines belonging to the domain. Both VXLAN and NVGRE suggest associating an IP multicast address taken from the network's infrastructure as a way of connecting together all the machines belonging to the same domain. All VMs within a domain can be reached by sending encapsulated packets to the domain's IP multicast address. One issue with this approach, however, is that some existing implementations may be challenged in supporting large numbers of multicast groups.

Another issue is whether fragmentation is needed. Whenever tunneling is used, one faces the potential problem that the packet plus encapsulation overhead will exceed the MTU of the path to the egress router. Fragmentation could be left to IP, could be done at the overlay level in a more optimized fashion or could be left out altogether, if it is believed that datacenter networks can be engineered to prevent MTU issues from arising.

Related to fragmentation is the question of how best to handle Path MTU issues, should they occur. Ideally, the original source of any packet (i.e, the sending VM) would be notified of the optimal MTU to use. Path MTU problems occurring within an overlay network would result in ICMP MTU exceeded messages being sent back to the egress tunnel switch at the entry point of the overlay. If the switch is embedded within a hypervisor, the hypervisor could notify the VM of a more appropriate MTU to use. It may be appropriate to specify a set of best practices for implementers related to the handling of Path MTU issues.

Both VXLAN and NVGRE assume that all machines belonging to the same domain are on the same IP subnet and reflect one L2 broadcast domain. This means that machines on different subnets cannot communicate directly, and must (conceptually) go through a router that is also part of the domain. The result can be suboptimal forwarding, compared to the case where traffic is tunneled directly between the two machines. Thus, another issue to consider is the handling of machines belonging to the same domain, but residing on different IP subnets. While one solution may simply be to assign a /0 subnet mask to the entire domain (so that all machines are on the same subnet), this could require a configuration change on individual VM images, which may be undesirable in some deployments.

Finally, successful deployment of an overlay approach will likely require appropriate Operations, Administration and Maintenance (OAM) facilities.

5. Benefits of an Overlay Approach

A key aspect of overlays is the decoupling of the "virtual" MAC and IP addresses used by VMs from the physical network infrastructure and the infrastructure IP addresses used by the datacenter. If a VM changes location, the switches at the edge of the overlay simply update their mapping tables to reflect the new location of the VM within the data center's infrastructure space. Because IP is used, a VM can now be located anywhere in the data center 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.

Multitenancy is supported by isolating the traffic of one domain from traffic of another. Traffic from one domain cannot be delivered to another domain without (conceptually) exiting the domain and entering the other domain. Likewise, external communications (from a VM within a domain to a machine outside a domain) is handled by having an ingress switch forward traffic to an external router, where an

egress switch decapsulates a tunneled packet and delivers it to the router for normal processing. This router is external to the overlay, and behaves much like existing external facing routers in datacenters today.

The use of a large (e.g., 24-bit) domain identifiers would allow 16 million distinct domains within a single datacenter, eliminating current VLAN size limitations.

Using an overlay that sits above IP allows for the leveraging of the full range of IP technologies, including quality-of-service (QoS) and Equal Cost Multipath (ECMP) routing for load balancing across multiple links.

Overlays are designed to handle the common case of a set of VMs placed within a single L2 broadcast domain. Such configurations include VMs placed within a single VLAN or IP subnet. All the VMs would be placed into a common overlay domain.

6. Related Work

6.1. 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 multitenancy), 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 trade offs 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.

6.2. Trill

Trill is an L2 based approach aimed at improving deficiencies and limitations with current Ethernet networks. While Trill provides a good approach to improving current Ethernets, it is entirely L2 based. Trill was not originally designed to scale to the sizes that current data centers need to scale to. [RFC6325] explicitly says:

> The TRILL protocol, as specified herein, is designed to be a Local Area Network protocol and not designed with the goal of scaling beyond the size of existing bridged LANs.

That said, approaches to extend Trill are currently under investigation [I-D.eastlake-trill-rbridge-fine-labeling]

6.3. L2VPNs

The IETF has specified a number of approaches for connecting L2 domains together as part of the L2VPN Working Group. That group, however is 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 be used within data centers where the overlay network is managed by the datacenter operator, rather than by an outside party. While overlays can run across the Internet as well, they will extend well into the datacenter itself (e.g., up to and including hypervisors) and include large numbers of machines within the datacenter 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 physical switches to which VMs connect) will be part of the overlay network and be responsible for encapsulating and decapsulating packets.

6.4. Proxy Mobile IP

Proxy Mobile IP [RFC5213] [RFC5844] makes use of the GRE Key Field [RFC5845] [RFC6245], but not in a way that supports multitenancy.

6.4.1. LISP

[Placeholder for LISP]

7. 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 datacenters with a large flat L2 network. Further analysis is needed to characterize expected improvements.

8. Summary

This document has argued that network virtualization using L3 overlays addresses a number of issues being faced as data centers scale in size. In addition, careful consideration of a number of issues would lead to the development of interoperable implementation of virtualization overlays.

9. Acknowledgments

This document incorporates significant amounts of text from [I-D.mahalingam-dutt-dcops-vxlan]. Specifically, much of Section 2 is incorporated verbatim from Section 3 of [I-D.mahalingam-dutt-dcops-vxlan]. The authors of that document include Mallik Mahalingam (VMware), Dinesh G. Dutt (Cisco), Kenneth Duda (Arista Networks), Puneet Agarwal (Broadcom), Lawrence Kreeger (Cisco), T. Sridhar (VMware), Mike Bursell (Citrix), and Chris Wright (Red Hat).

Additional text in Section 2 was taken from [I-D.sridharan-virtualization-nvgre]. The authors of that document include Murari Sridharan (Microsoft), Kenneth Duda (Arista Networks), Ilango Ganga (Intel), Albert Greenberg (Microsoft), Geng Lin (Dell), Mark Pearson (Hewlett-Packard), Patricia Thaler (Broadcom), Chait Tumuluri (Emulex), Narasimhan Venkataramiah (Microsoft) and Yu-Shun Wang (Microsoft).

Helpful comments and improvements to this document have come from Dinesh Dutt, Ariel Hendel, Vinit Jain and Larry Kreeger.

10. IANA Considerations

This memo includes no request to IANA.

11. Security Considerations

TBD

12. Informative References

[I-D.eastlake-trill-rbridge-fine-labeling]
Eastlake, D., Zhang, M., Agarwal, P., Dutt, D., and R. Perlman, "RBridges: Fine-Grained Labeling", draft-eastlake-trill-rbridge-fine-labeling-01 (work in

progress), July 2011.

[I-D.hasmit-otv]
Grover, H., Rao, D., Farinacci, D., and V. Moreno, "Overlay Transport Virtualization", draft-hasmit-otv-03 (work in progress), July 2011.

[I-D.ietf-6man-udpzero]
Fairhurst, G. and M. Westerlund, "IPv6 UDP Checksum Considerations", draft-ietf-6man-udpzero-03 (work in progress), April 2011.

[I-D.mahalingam-dutt-dcops-vxlan]
Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-00 (work in progress), August 2011.

[I-D.sridharan-virtualization-nvgre]
Sridharan, M., Duda, K., Ganga, I., Greenberg, A., Lin, G., Pearson, M., Thaler, P., Tumuluri, C., Venkataramaiah, N., and Y. Wang, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-00 (work in progress), September 2011.

[RFC2661] Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., and B. Palter, "Layer Two Tunneling Protocol "L2TP"", RFC 2661, August 1999.

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, September 2000.

[RFC5213] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.

[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.

Authors' Addresses

Thomas Narten (editor)
IBM

Email: narten@us.ibm.com

Murari Sridharan
Microsoft

Email: muraris@microsoft.com