

TRILL working group
Internet Draft
Intended status: Standard Track
Expires: Sept 2011

L. Dunbar
Huawei

March 7, 2011

Directory Server Assisted TRILL edge
draft-dunbar-trill-server-assisted-edge-00.txt

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

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

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

This Internet-Draft will expire on September 7, 2011.

Copyright Notice

Copyright (c) 2011 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 BSD License.

Internet-Draft Directory Server Assisted TRILL edge

March 2011

Abstract

TRILL edge nodes currently learn the mapping between MAC address and its corresponding TRILL edge node address by observing the data packets traversed through.

This document describes why and how directory based server(s) can optimize TRILL network in data center environment.

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](#) 0.

Table of Contents

| | |
|---|-------------------|
| 1. Introduction | 2 |
| 2. Terminology | 3 |
| 3. Impact to TRILL by massive number of hosts | 3 |
| 4. Directory Server for TRILL in Data Center environment. | |
| 5. Conclusion and Recommendation | |
| 6. Manageability Considerations | |
| 7. Security Considerations | |
| 8. IANA Considerations | 5 |
| 9. Acknowledgments | 5 |
| 10. References | 5 |
| Authors' Addresses | 6 |
| Intellectual Property Statement | |
| Disclaimer of Validity | 6 |

[1. Introduction](#)

Data center networks are different from campus networks in several ways. Main differences include:

- VM (host) to server assignment is done by Server (or VM) Manager, which means that the host location is arranged by management system(s).
- Topology is based on racks and rows;

There could be massive number of virtual machines (hosts), but relatively small number of switches.

This draft describes why Data Center TRILL networks can be optimized by utilizing directory server based approach.

[2.](#) Terminology

Bridge: IEEE802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

DC: Data Center

EOR: End of Row switches in data center.

FDB: Filtering Database for Bridge or Layer 2 switch

ToR: Top of Rack Switch. It is also known as access switch.

VM: Virtual Machines

[3.](#) Impact to TRILL by massive number of hosts

In a virtualized data center, a VM may be placed on any physical server. A variety of algorithms can be applied to select the location of a VM. Resource aware algorithms (e.g. energy, bandwidth, etc,) will use a placement that satisfies the processing requirements of each VM but require the minimal number of physical servers and switching devices.

With this, and similar types of assignment algorithm, subnets tend to extend throughout the network. When this happens, the broadcast messages within each subnet will be flooded across the TRILL domain, which not only consumes a lot of bandwidth on links in TRILL domain, but also causes a TRILL edge port to learn all the hosts belonging to all the subnets which are enabled on the port. Even though a TRILL edge port is only supposed to learn the entries which communicate with hosts underneath, the frequent ARP/ND from all hosts within each subnet will always refresh the TRILL edge node's MAC<->TRILL-Edge mapping table.

Consider a data center with 80 rows, 8 racks per row and 40 servers per rack. There can be $80 \times 8 \times 40 = 25600$ servers. Suppose each server is virtualized to 20 VMs, there could be $25600 \times 20 = 512000$ hosts in this data center.

Let's consider a case that the TRILL edge starts at an Ingress port

of a TOR switch. Assuming there are 5 different VLANs enabled on the TRILL Ingress port (i.e. the 20 VMs in one server belong to 5 different VLANs) and each VLAN has 200 hosts, then the TRILL edge

port has to learn $5 \times 200 = 1000$ MAC&VLAN entries. Since there are 40 ports on the TOR, the total number of MAC&VLAN entries for the TOR switch is $1000 \times 40 = 40000$. Under this scenario, there will be 25600 entries in the TRILL routing domain if protection is not considered. When protection is considered, the number of ports in TRILL domain will double. That may be too many nodes for the IS/IS routing domain. Let's consider another case of TRILL edge starting at the End of Row switches. With the same assumption as before, there are $40 \times 20 = 800$ hosts to attached to each port of an EoR switch and $8 \times 800 = 6400$ hosts attached to an EoR switch. If all those 6400 hosts belong to 640 VLANs and each VLAN has 200 hosts, then the total number of MAC&VLAN entries to be learned by the TRILL edge (i.e. EoR) = $640 \times 200 = 128000$. Under this scenario, there will be $80 \times 8 = 640$ EoR ports in the TRILL routing domain when protection is not considered and 1280 EoR ports when protection is considered. However, the number of MAC&VLAN entries to be learnt by the TRILL edge node is very large.

4. Directory Server for TRILL in Data Center environment.

As described in the [Section 1](#), the VM placement to server/rack is orchestrated by Server (or VM) Management System(s). Therefore, there is a central location with the information on where each VM is placed. So it is relatively reliable to build a centralized (or distributed) directory server(s) who has the knowledge on where each VM is placed.

Here can be a procedure for TRILL edge node to utilize a Directory Server

TRILL edge node can simply intercept all ARP requests and forward them to the Directory Server,

The reply from the Directory Server can be the standard ARP reply with an extra field showing the TRILL egress node address

TRILL ingress node can cache the mapping

If TRILL edge node receives an unknown MAC-DA, it simply forwards the packet to the directory server. The directory

server can simply drop the frame if it doesn't have the information, or forward the frame to the correct egress node and send down a new mapping to the ingress Trill edge node.

Another approach is for Directory Server to pass down the MAC&VLAN mapping for all the hosts belonging to all the VLANs enabled on the TRILL edge port.

[5. Conclusion and Recommendation](#)

The traditional TRILL learning approach of observing data plane can no longer keep pace with the ever growing number of hosts in Data center.

Therefore, we suggest TRILL to consider directory assisted approach(es). This draft only introduces the basic concept of using directory assisted approach for TRILL edge nodes to learn the MAC to TRILL mapping. We want to get some working group consensus before drilling down to detailed steps required for the approach.

[6. Manageability Considerations](#)

This document does not add additional manageability considerations.

[7. Security Considerations](#)

This document has no additional requirement for security.

[8. IANA Considerations](#)

[9. Acknowledgments](#)

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

[10. References](#)

[ARMD-Problem] Dunbar, et,al, "Address Resolution for Large Data Center Problem Statement", Oct 2010.

[ARP reduction] Shah, et. al., "ARP Broadcast Reduction for Large Data Centers", Oct 2010

Dunbar

Expires Sept7, 2011

[Page 5]

Internet-Draft Directory Server Assisted TRILL edge

March 2011

Authors' Addresses

Linda Dunbar
Huawei Technologies
1700 Alma Drive, Suite 500
Plano, TX 75075, USA
Phone: (972) 543 5849
Email: ldunbar@huawei.com

Intellectual Property Statement

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary

rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Dunbar

Expires Sept7, 2011

[Page 6]

Internet-Draft Directory Server Assisted TRILL edge

March 2011

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

