Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 10, 2013

Yantao Sun
Beijing Jiaotong University
Xiaoli Song
Bin Liu
ZTE Inc.
Qiang Liu
Jing Cheng
Beijing Jiaotong University

July 9, 2012

MatrixDCN: A New Network Fabric for Data Centers
draft-sun-matrix-dcn-00

Abstract

This document introduces describes the requirement of today's data centers and a new type of network topology called MatrixDCN (matrix data center network). MatrixDCN is used to deploy large scale data center network, which can support more than 100 thousands of servers in one data center without network bandwidth bottleneck.

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 January 10, 2013.

Copyright Notice

Table of Contents

1. Introduction

Traditional network topology is a tree-like fabric composed by routers and switches, in which network is divided into 3 layers including core layer, aggregation layer and access layer. All servers are connected to the access switches in access layer.

This kind of topology has some problems using in data center network. Firstly, it constrains the scale of data center network. When the scale of network expands, those routers in core layer are apt to be the bandwidth bottleneck in the whole network, since more packages need to be routed between different layer-2 domains through core routers. Secondly, data center network is divided into many small layer-2 domains by core routers, which conduces to that VM's migration is limited to only one layer-2 domains. And last, it's difficult to exploit mass of redundant links between switches as STP protocol must be used to ensure no loops in layer-2 network to avoid broadcast storm.

To solve above problems, new network architectures should be introduced to data centers. The new architecture requires the following features: 1) support multiple paths to eliminate bandwidth bottleneck; 2) has regular network topology with good extendibility and maintainability; 3) support VM migration in the entire network; 4) has enough VLANs and permit any endpoints to compose one VLAN.

1.1. Acronyms & Definitions

DCN - Data Center Network

AS - Access Switch

RS - Row Switch

CS - Column Switch

PDU - Protocol Data Unit

OSPF - Open Shortest Path First Routing Protocol

VLAN - Virtual Local Area Network

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 fabric

In the researches on data center networks, fat-tree fabric attracts a great many of attentions. Fat-tree is on kind of multi-root tree and a lot of important research works and some practices have been done based on fat-tree fabric. In this proposal, we introduce another variety of multi-root tree called MatrixDCN for building large-scale data center network. Similar with fat-tree, MatrixDCN has regular topology, multiple paths, special addressing and routing matched with its topology.

MatrixDCN can support more than 100 thousands of physical servers in a single data center network. And furthermore, it can support virtual machine migration in whole network and huge number of tenants' isolation by modest modification.

3.1. Components

In MatrixDCN, there are 3 types of network devices, called Row Switch (RS), Column Switch (CS) and Access Switch (AS). AS switches are deployed as a matrix with multiple rows and columns. For example, one 8 X 8 matrix has 8 rows and 8 columns and 64 AS switches. For a RS, it is deployed on the head of one row and links all the AS switches together in this row, and for a CS, it is deployed on the head of one column and links all the AS switches together in this column. Every AS connects with all the RS and CS switches located at the row and column head of it. Figure 1 is an example of 2 X 2 MatrixDCN.

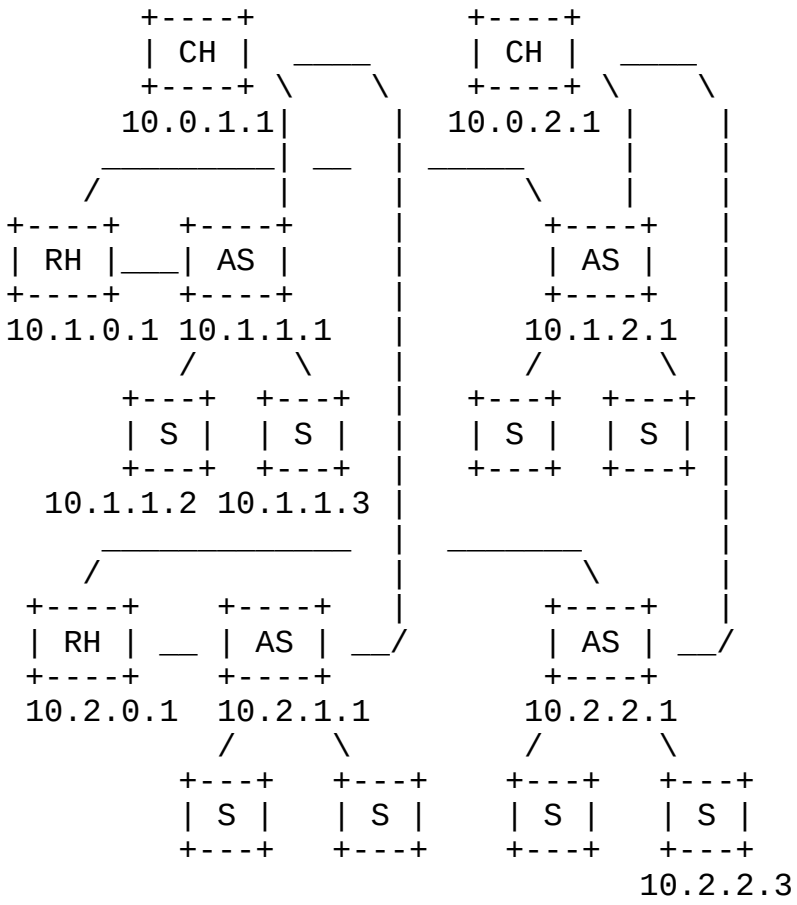


Figure 1: demonstration topology

3.2. Multiple Paths

To eliminate bandwidth bottleneck in data center networks, MatrixDCN can deploy multiple RS switches in one row and CS switches in one column. Thus, we can have more links between AS and RS/CS, which means more bandwidth, between an AS and its RS and CS switches. If bandwidth between AS and RS(called row bandwidth), bandwidth between AS and CS(called column bandwidth) and access bandwidth of AS are equal, this network is approximately non-blocking.

3.3. Addressing

In MatrixDCN, all the devices including servers and switches are assigned an IP address according to their position in the network. Suppose an AS is located at the mth row and nth column, its IP Address is 10.m.n.1/24 and the servers connected with it are set to 10.m.n.x/24. For RS switches located at the head of mth row, their IP will be set as 10.m.0.x/16. For CS switches located at the head of mth column, their IP will be set as 10.0.m.x/255.0.255.0.

4. Routing

The routing is very simple for MatrixDCN, as its topology is very regular and every switch can obtain the knowledge of the entire topology without exchanging link states since the device position is encoded in its address. All the switches in MatrixDCN have routing ability. These switches store routing entry using the standard routing table. The structure of routing table is illustrated in table 1.

Table standard IP routing table

Subnet	Address	Subnet mask	Next hop	Cost	Create/update time
10.1.0.0	255.255.0.0	10.1.1.1			
10.2.0.0	255.255.0.0	10.1.2.1			

4.1. Routing of RS

When packets arrive at RS switch, RS will determine the next hop of a packet by the 3rd number of its destination IP address based on IP routing table. If this number is k, the packet will be send to the kth AS switch (next hop) on the row of this RS. The routing table on RS switch of the ith row looks like below:

Destination/Subnet mask	Next hop
10.X.1.X/255.0.255.0	10.i.1.1
10.X.2.X/255.0.255.0	10.i.2.1
10.X.3.X/255.0.255.0	10.i.3.1
.....	

4.2. Routing of CS

When packets arrive at CS switch, CS will determine the next hop of a packet by the 2rd part of its destination IP address based on IP routing table. If this number is k, the packet will be send to the kth AS switch (next hop) on the column of this CS. The routing table on CS switch of the ith row looks like below:

Destination/Subnet mask	Next hop
10.1.X.X/255.255.0.0	10.1.Col.1
10.2.X.X/255.255.0.0	10.2.Col.1
10.3.X.X/255.255.0.0	10.3.Col.1
.....	

4.3. Routing of AS

When packets arrive at AS switch, AS should determine the next hop is whether RS or CS switch for every packet. For a packet, if its destination IP address is on the same row, it will be sent to the RS switch, and if its destination IP address is on the same column, it will be sent to the CS switch. For the packet whose destination is on different row and column, it can be sent to either RS or CS, and for the packet with destination on the same row and column, it is forwarded through level-2 switching without routing.
The routing table of the CS on the crossing position of the ith row and the jth column looks like below:

Destination/Subnet mask	Next hop
10.Row.0.0/255.255.0.0	10.Row.0.X1
10.Row.0.0/255.255.0.0	10.Row.0.X2
.....	
10.0.Col.0/255.0. 255.0	10.0.Col.X1
10.0.Col.0/255.0. 255.0	10.0.Col.X2
.....	
10.0.0.0/255.0.0.0	10.Row.0.X1
10.0.0.0/255.0.0.0	10.Row.0.X2
.....	
10.0.0.0/255.0.0.0	10.0.Col.X1
10.0.0.0/255.0.0.0	10.0.Col.X2
.....	

4.4. Construction of Routing Table

To build the routing table for switches, the connection relationship between adjacent switches should be learned automatically, and to learn connection relationship, every switch will send Hello PDU to all of its active ports periodically. Hello PDU is encapsulated in a UDP packet. A well-known UDP port will be obtained from IANA.

4.4.1. Construction for CS

For a CS switch, such as 10.0.n.x, which will receive Hello PDUs from all the AS switches on the same column, its routing table is built according to the following rules:

When the CS received a Hello PDU from 10.m.n.1, a routing entry that destination is "10.m.0.0/255.255.0.0" and next hop is "10.m.n.1" will be added/refreshed to its routing table. If Hello PDU can’t be received in set time, the corresponding routing entry will be deleted.

4.4.2. Construction for RS

For a RS switch, such as 10.m.0.x, which will receive Hello PDUs from all the AS switches on the same row, its routing table is built according to the following rules:

When the RS received a Hello PDU from 10.m.n.1, a routing entry that destination is "10.0.n.0/255.0.255.0" and next hop is "10.m.n.1" will be added/refreshed to its routing table. If Hello PDU can’t be received in set time, the corresponding routing entry will be deleted.

4.4.3. Construction for AS

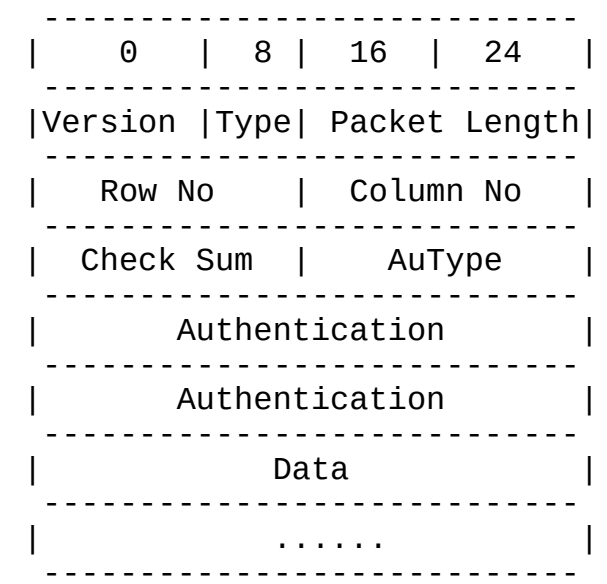
For an AS switch, such as 10.m.n.1, which will receive Hello PDUs from all the RS switches of the mth row and all the CS switches of the nth column, its routing table is built according to the following rules:

If a Hello PDU received from a RS 10.m.0.x, two routing entry "10.m.0.0/255.255.0.0 10.m.0.x" and "10.0.0.0/255.0.0.0 10.m.0.x" will be added/refreshed to its routing table.

If a Hello PDU received from a CS 10.0.n.x, two routing entry "10.0.n.0/255.255.0.0 10.0.0.x" and "10.0.0.0/255.0.0.0 10.m.0.x" will be added/refreshed to its routing table.

If Hello PDU can’t be received in set time, the corresponding routing entry will be deleted.

4.5. PDU Format



Version: version number of MatrixDCN Routing Protocol.

Type: PDU packet type. If value is 1, this is a Hello PDU. If value is 2, this is a Link State Advertisements PDU to notice link fault knowledge.

Packet Length: the total length of the PDU including PDU head and data.

Row No and Colum No: The position of this switch.

Check Sum: Check sum for the total PDU.

AuType: Authentication type. 0: no authentication, 1: Plaintext Authentication, 2: MD5 Authentication.

Authentication: Authentication infomation. 0: undefined, 1: Key, 2: key ID, MD5 data length and packet number. MD5 data is appended to the back of the packet.

AuType and Authentication can refer to the definition of OSPF packet.

4.6. Fault Tolerance

Any network node or link fault will conduct communication break, so fault tolerance must be considered in a usable routing protocol. To do this, switches in MatrixDCN should learn the whole network state through Link State Advertisements PDU. The more details will be elaborated in the next version of this document.

5. VM Migration

Subnetting and location-related addressing make possible to build large-scale data center networks, but limit the migration of VMs. To solve this problem, Overlay or IP tunneling technology is introduced to data center networks. By small extension to AS, MatrixDCN can support seamless VM migration in the entire data center without any modification to above routing procedure. The feature of regular topology in MatrixDCN has been considered into this solution and more detail would be specialized in another document.

6. Multiple tenants

The present VLAN protocol 802.1q can’t satisfy the requirement of data center networks as it can only support only about 4000 VLANs. VXLAN and NVGRE are two similar but competitive draft protocols for solving this problem and can be used in MatrixDCN. Moreover, another solution in consideration of the regular topology will be discussed in other document.

7. Deployment Scenarios

MatrixDCN can be used to deploy large-scale data center network. Suppose AS switch has 40 down-link ports with 1G bits speed and 8 up-link ports with 10G bits speed, RS and CS switch has 40 ports with 10G bits speed, those switches are currently main stream switches used in data center, we can build a MatrixDCN with 40 rows X 40 columns. In this MatrixDCN, every row has 4 RS switches and every column has 4 CS switches. For every AS, 4 up-link ports link RS switches and 4 up-link ports link CS switches, and its 40 down-link ports link servers. Thus, this MatrixDCN can contain up to 64,000 servers using 1600 AS switches, 160 RS switches and 160 CS switches. The average available bandwidth for every server is about 1000M bits.

8. Security Considerations

The protection for routing information and isolation for networks of different tenants (VLAN) should be considered in this protocol.

9. Conclusion

Today’s data center produces some new requirement to networks, such as no-blocking, seamless VM migration and multiple tenants.This document introduces MatrixDCN, a new network fabric for data centers.

MatrixDCN can support up to 100 thousands of servers without network bandwidth bottleneck. This fabric is very simple and extendable, and its routing is very effective. Furthermore, this fabric has some advantages on supporting VM Migration and one-to-many or many-to-many traffic in cloud computing.

10. Reference

[RFC2328] J. Moy, "OSPF Version 2", RFC2338, Apr. 1998.

[FAT-TREE] M. Al-Fares, A. Loukissas, and A. Vahdat. "A Scalable, Commodity, Data Center Network Architecture",In ACM SIGCOMM 2008.

11. Acknowledgments

Thanks for Yuhua Wei, Lizhong Jin, Jinghai Yu, and Zhui GUo. They give important advices for this document.

Authors' Addresses

Yantao Sun
Beijing Jiaotong University
No.3 Shang Yuan Cun, Hai Dian District
Beijing 100044
China

Phone:
Email: ytsun@bjtu.edu.cn

Xiaoli Song
ZTE Inc.
ZTE Plaza, No.19 East Huayuan Road,Haidian District
Beijing 100191
China

Email: song.xiaoli@zte.com.cn

Bin Liu
ZTE Inc.
ZTE Plaza, No.19 East Huayuan Road,Haidian District
Beijing 100191
China

Email: liu.bin21@zte.com.cn

Qiang Liu
Beijing Jiaotong University
No.3 Shang Yuan Cun, Hai Dian District
Beijing 100044
China

Email: liuq@bjtu.edu.cn

Jing Cheng
Beijing Jiaotong University
No.3 Shang Yuan Cun, Hai Dian District
Beijing 100044
China

Email: yourney.j@gmail.com