Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 16, 2014

P. Lapukhov
Microsoft Corporation
A. Premji
Arista Networks
J. Mitchell, Ed.
Microsoft Corporation

July 15, 2013

Use of BGP for routing in large-scale data centers
draft-lapukhov-bgp-routing-large-dc-05

Abstract

Some network operators build and operate data centers that support over one hundred thousand servers. In this document, such data centers are referred to as "large-scale" to differentiate them from smaller infrastructures. Environments of this scale have a unique set of network requirements with an emphasis on operational simplicity and network stability. This document summarizes operational experience in designing and operating large-scale data centers using BGP as the only routing protocol. The intent is to report on a proven and stable routing design that could be leveraged by others in the industry.

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 16, 2014.

Copyright Notice

Copyright (c) 2013 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

1. Introduction

This document describes a practical routing design that can be used in a large-scale data center ("DC") design. Such data centers, also known as hyper-scale or warehouse-scale data centers, have a unique attribute of supporting over a hundred thousand servers. In order to accommodate networks of this scale, operators are revisiting networking designs and platforms to address this need.

The design described in this document is based on operational experience with data centers built to support large scale Web infrastructure. The primary requirements in such environments are operational simplicity and network stability so that a small group of people can effectively support a large network infrastructure.

After experimentation and extensive testing, Microsoft choose to use an end to end routed network infrastructure with External BGP (EBGP) [RFC4271] as the only routing protocol for some of its DC deployments. This is in contrast with more traditional DC designs, which may use more hierarchical topologies and rely on extending Layer 2 domains across multiple network devices. This document elaborates on the requirements that led to this design choice and presents details of the EBGP routing design as well as explores ideas for further enhancements.

This document first presents an overview of network design requirements and considerations for large-scale data centers. Then traditional hierarchical data center network topologies are contrasted with Clos networks that are horizontally scaled out. Arguments for selecting EBGP with a Clos topology as the most appropriate routing protocol to meet the requirements are presented. Then the design is described in detail. Finally some additional considerations and options are presented.

2. Network Design Requirements

This section describes and summarizes network design requirements for large-scale data centers.

2.1. Bandwidth and Traffic Patterns

The primary requirement when building an interconnection network for large number of servers is to accommodate application bandwidth and latency requirements. Until recently it was quite common to see the majority of traffic entering and leaving the data center (also known as north-south traffic). As a result, traditional "tree" topologies were sufficient to accommodate such flows, even with high oversubscription ratios in network equipment. If more bandwidth was required, it was added by "scaling up" the network elements, especially at the Tier-1 layer, e.g. by upgrading the device's line-cards or fabrics or replacing the device with one with higher port density.

Today many large-scale data centers host applications generating significant amounts of server to server traffic, traveling between various Tier-2 or Tier-3 devices but without egressing the DC, also known as "east-west" traffic. Examples of such applications could be compute clusters such as Hadoop, large amounts of replication traffic between clusters needed by certain applications, or virtual machine migrations. Scaling up traditional tree topologies to match these bandwidth demands becomes either too expensive or impossible due to physical limitations.

2.2. CAPEX Minimization

The cost of the network infrastructure alone (CAPEX) constitutes about 10-15% of total data center expenditure (see [GREENBERG2009]). However, the absolute cost is significant, and there is a need to constantly drive down the cost of individual network elements. This can be accomplished in two ways:

- Unifying all network elements, preferably using the same hardware type or even the same device. This allows for bulk purchases with discounted pricing.

- Driving costs down using competitive pressures, by introducing multiple network equipment vendors.

In order to allow for vendor diversity, it is important to minimize the software feature requirements for the network elements. Furthermore, this strategy provides maximum flexibility of vendor equipment choices while enforcing interoperability using open standards.

2.3. OPEX Minimization

Operating large scale infrastructure could be expensive, provided that larger amount of elements will statistically fail more often. Having a simpler design and operating using a limited software feature-set minimizes software issue related failures.

An important aspect of OPEX minimization is reducing size of failure domains in the network. Ethernet networks are known to be susceptible to broadcast or unicast traffic storms that have dramatic impact on network performance and availability. The use of a fully routed design significantly reduces the size of the data-plane failure domains (e.g. limits them to Tier-3 devices only). However, such designs also introduce the problem of distributed control-plane failures. This observation calls for simpler control-plane protocols that are expected to have less chances of network meltdown.

2.4. Traffic Engineering

In any data center, application load-balancing is a critical function performed by network devices. Traditionally, load-balancers are deployed as dedicated devices in the traffic forwarding path. The problem arises in scaling load-balancers under growing traffic demand. A preferable solution would be able to scale load-balancing layer horizontally, by adding more of the uniform nodes and distributing incoming traffic across these nodes.

In situation like this, an ideal choice would be to use network infrastructure itself to distribute traffic across a group of load-balancers. The combination of Anycast prefix advertisement [RFC4786] and Equal Cost Multipath (ECMP) functionality can be used to accomplish this goal. To allow for more granular load-distribution, it is beneficial for the network to support the ability to perform controlled per-hop traffic engineering. For example, it is beneficial to directly control the ECMP next-hop set for Anycast prefixes at every level of network hierarchy.

2.5. Summarized Requirements

This section summarizes the list of requirements outlined in the previous sections:

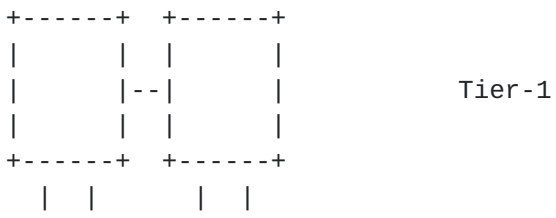
- REQ1: Select a topology that can be scaled "horizontally" by adding more links and network devices of the same type without requiring upgrades to the network elements themselves.
- REQ2: Define a narrow set of software features/protocols supported by a multitude of networking equipment vendors.
- REQ3: Choose a routing protocol that has a simple implementation in terms of programming code complexity and ease of operational support.
- REQ4: Minimize the failure domain of equipment or protocol issues as much as possible.
- REQ5: Allow for traffic engineering, preferably via explicit control of the routing prefix next-hop using built-in protocol mechanics.

3. Data Center Topologies Overview

This section provides an overview of two general types of data center designs - hierarchical (also known as tree based) and Clos based network designs.

3.1. Traditional DC Topology

In the networking industry, a common design choice for data centers typically look like a (upside-down) tree with redundant uplinks and three layers of hierarchy namely core, aggregation/distribution and access layers (see Figure 1). To accommodate bandwidth demands, each higher layer, from server towards DC egress or WAN, has higher port density and bandwidth capacity where the core functions as the "trunk" of the tree based design. To keep terminology uniform and for comparison with other designs, in this document these layers will be referred to as Tier-1, Tier-2 and Tier-3 "tiers" instead of Core, Aggregation or Access layers.



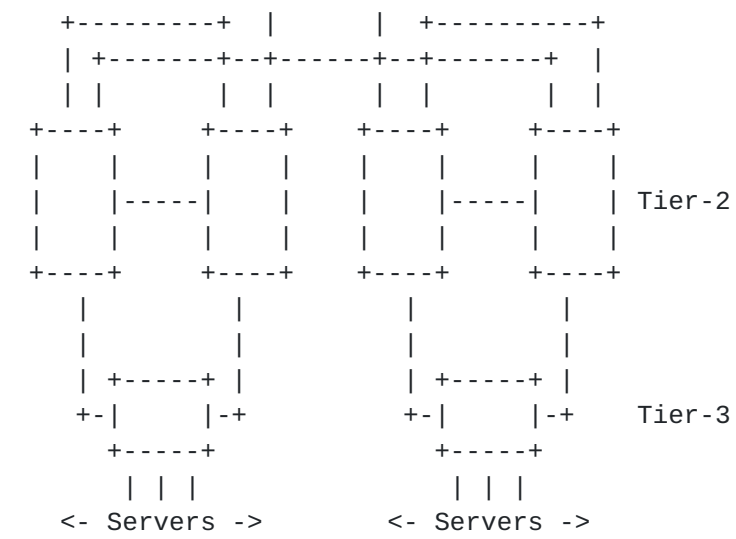


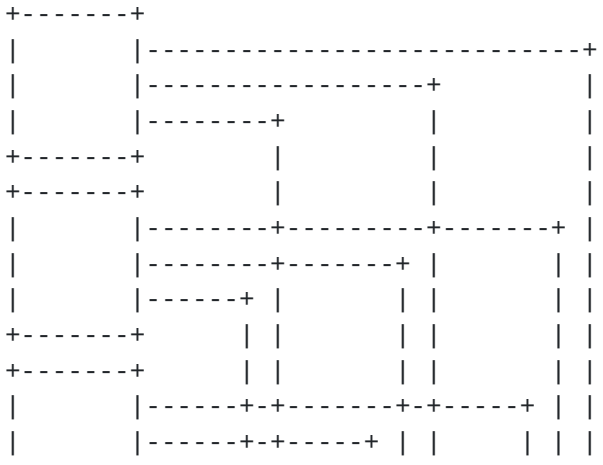
Figure 1: Typical DC network topology

3.2. Clos Network topology

This section describes a common design for horizontally scalable topology in large scale data centers in order to meet REQ1.

3.2.1. Clos Topology Overview

A common choice for a horizontally scalable topology is a folded Clos topology, sometimes called "fat-tree" (see, for example, [INTERCON] and [ALFARES2008]). This topology features an odd number of stages (sometimes known as dimensions) and is commonly made of uniform elements, e.g. network switches with the same port count. Therefore, the choice of Clos topology satisfies both REQ1 and also facilitates REQ2. See Figure 2 below for an example of a folded 3-stage Clos topology:



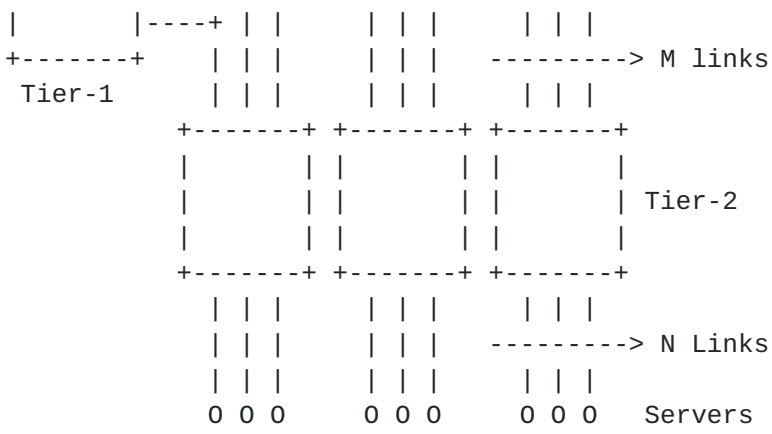


Figure 2: 3-Stage Folded Clos topology

This topology is sometimes also referred to as a "Leaf and Spine" network, where "Spine" is the name given to the middle stage of the Clos topology (Tier-1) and "Leaf" is the name of input/output stage (Tier-2). For uniformity, this document will refer to these layers using the "Tier-n" notation.

3.2.2. Clos Topology Properties

The following are some key properties of the Clos topology:

- The topology is fully non-blocking (or more accurately: non-interfering) if M >= N and oversubscribed by a factor of N/M otherwise. Here M and N is the uplink and downlink port count respectively, for a Tier-2 switch as shown in Figure 2

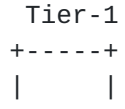
- Utilizing this topology requires an control and data plane supporting ECMP with the fan-out of M or more

- Tier-1 switches have exactly one path to every server in this topology

- Traffic flowing from server to server is load-balanced over all available paths using ECMP

3.2.3. Scaling the Clos topology

A Clos topology can be scaled either by increasing network element port density or adding more stages, e.g. moving to a 5-stage Clos, as illustrated in Figure 3 below:



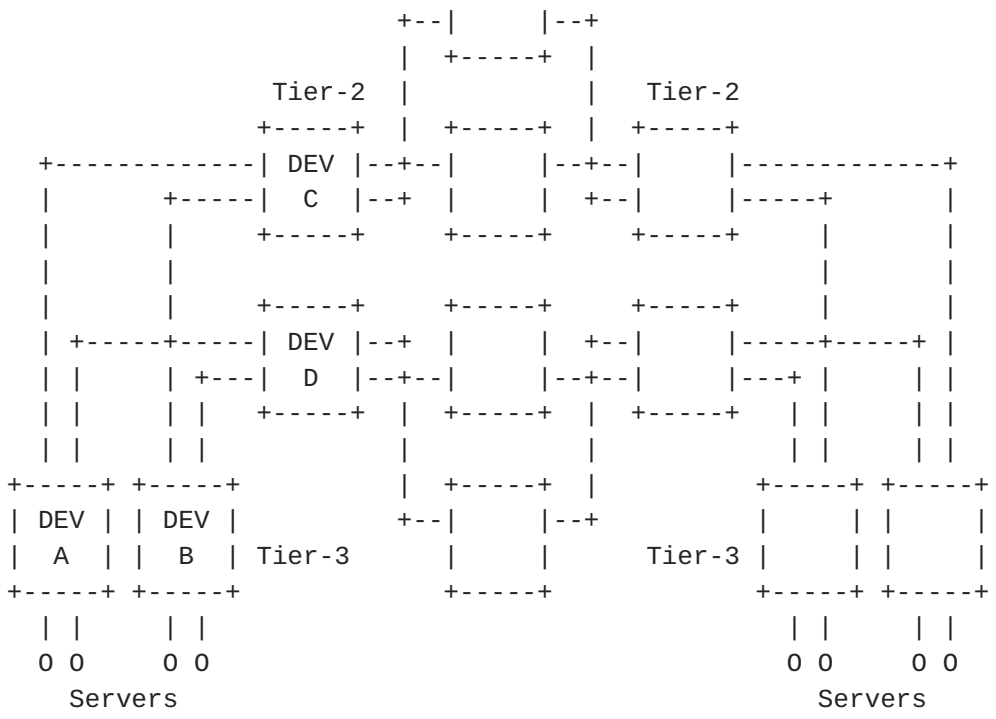


Figure 3: 5-Stage Clos topology

The small example topology on Figure 3 is built from devices with a port count of 4 and provides full bisectional bandwidth to all connected servers. In this document, one set of directly connected Tier-2 and Tier-3 devices along with their attached servers will be referred to as a "cluster". For example, DEV A, B, C, D, and the servers that connect to DEV A and B, on Figure 3 form a cluster.

In practice, the Tier-3 layer of the network, which are typically top of rack switches (ToRs), is where oversubscription is introduced to allow for packaging of more servers in the data center while meeting the bandwidth requirements for different types of applications. The main reason to limit oversubscription at a single layer of the network is to simplify application development that would otherwise need to account for multiple bandwidth pools: within rack (Tier-3), between racks (Tier-2), and between cluster (Tier-1). Since oversubscription does not have a direct relationship to the routing design it is not discussed further in this document.

3.2.4. Managing the Size of Clos Topology Tiers

If a data-center network size is small, it is possible to reduce the number of switches in Tier-1 or Tier-2 of Clos topology by a power of two. To understand how this could be done, take Tier-1 as an example. Every Tier-2 device connects to a single group of Tier-1 devices. If half of the ports on each of the Tier-1 devices are not

being used then it is possible to reduce the number of Tier-1 devices by half and simply map two uplinks from a Tier-2 device to the same Tier-1 device that were previously mapped to different Tier-1 devices. This technique maintains the same bisectional bandwidth while reducing the number of elements in the Tier-1 layer, thus saving on CAPEX. The tradeoff, in this example, is the reduction of maximum DC size in terms of overall server count by half.

In this example, Tier-2 devices will be using two parallel links to connect to each Tier-1 device. If one of these links fails, the other will pick up all traffic of the failed link, possible resulting in heavy congestion and quality of service degradation if the path determination procedure, does not take bandwidth amount into account. To avoid this situation, parallel links can be grouped in link aggregation groups (LAGs) with widely available implementation settings that take the whole bundle down upon a single link failure. Equivalent techniques that enforce "fate sharing" on the parallel links can be used in place of LAGs to achieve the same effect. As a result of such fate-sharing, traffic from two or more failed links will be re-balanced over the multitude of remaining paths that equals the number of Tier-1 devices. Although this example is using 2 for simplicity, reduced impact of bandwidth capacity can be achieved for a link or device failure with a larger fan-out.

4. Data Center Routing Overview

This section provides an overview of three general types of data center protocol designs - Layer 2 only, Hybrid L2/L3 and Layer 3 only.

4.1. Layer 2 Only Designs

Originally most data center protocol designs used Spanning-Tree Protocol (STP) for loop free topology creation, typically utilizing variants of the typical DC topology described in Section 3.1. At the time, many DC switches either did not support Layer 3 routed protocols or supported it with additional licensing fees, which played a part in the design choice. Although many enhancements have been made through the introduction of Rapid Spanning Tree Protocol and Multiple Spanning Tree Protocol that increase convergence, stability and load balancing in larger topologies many of the fundamentals of the protocol limit its applicability in large scale DC's. STP and its newer variants use an active/standby approach to path selection and are therefore hard to deploy in horizontally scaled topologies described in Section 3.2. Further, operators have had many experiences with large failures due to issues caused by improper cabling, misconfiguration, or flawed software on a single device. These failures regularly impacted the entire spanning-tree

domain and were very hard to troubleshoot due to the nature of the protocol. For these reasons, and since almost all DC traffic is now IP therefore requiring a Layer 3 routing protocol at the network edge for external connectivity, designs utilizing STP usually fail all of the requirements of large scale DC operators.

It should be noted that building large, horizontally scalable, Layer 2 only networks without STP is possible recently through the introduction of TRILL [RFC6325]. TRILL resolves many of the issues STP has for large scale DC design however currently the maturity of the protocol, limited number of implementations, and requirement for new equipment that supports it has limited it's applicability and increased the cost of such designs.

4.2. Hybrid L2/L3 Designs

Operators have sought to limit the impact of STP failures and build larger scale topologies through implementing routing protocols in either the Tier-1 or Tier-2 parts of the network and dividing the Layer-2 domain into numerous, smaller domains. This design has allowed data centers to scale up, but at the cost of complexity in the network managing multiple protocols. For the following reasons, operators have still retained Layer 2 in either the access (Tier-3) or both access and aggregation (Tier-3 and Tier-2) parts of the network:

- Supporting legacy applications that may require direct Layer 2 adjacency or use non-IP protocols
- Seamless mobility for virtual machines that require the preservation of IP addresses when a virtual machine moves to different access switch
- Simplified IP addressing = less IP subnets is required for the data center
- Application load-balancing may require direct Layer 2 reachability to perform certain functions such as Layer 2 Direct Server Return (DSR)
- Continued CAPEX differences between Layer-2 and Layer-3 capable switches

4.3. Layer 3 Only Designs

Network designs that leverage IP routing down to Tier-3 of the network have gained popularity as well. The main benefit of these designs is improved network stability and scalability, as a result of confining L2 broadcast domains. Commonly an IGP such as OSPF [RFC2328] is used as the primary routing protocol in such a design. As data centers grow in scale, and server count exceeds tens of thousands, such fully routed designs have become more attractive. Many vendors pricing has also changed to support this model so that data center class switches often do not cost more whether running traditional Layer 2 or Layer 3 control plane protocols.

Choosing a Layer 3 only design greatly simplifies the network, facilitating the meeting of REQ1 and REQ2, and has wide spread adoption in networks where large Layer 2 adjacency and larger size Layer 3 subnets are not as critical compared to network scalability and stability. Application providers and network operators continue to also develop new solutions to meet some of the requirements that previously have driven large Layer 2 domains.

5. Routing Protocol Selection and Design

In this section the motivations for using External BGP (EBGP) as the single routing protocol for data center networks having a Layer 3 protocol design and Clos topology are reviewed. Then, a practical approach for designing an EBGP based network is provided.

5.1. Choosing EBGP as the Routing Protocol

REQ2 would give preference to the selection of a single routing protocol to reduce complexity and interdependencies. While it is common to rely on an IGP in this situation, sometimes with either the addition of EBGP at the device bordering the WAN or Internal BGP (IBGP) throughout, this document proposes the use of an EBGP only design.

Although EBGP is the protocol used for almost all inter-provider routing on the Internet and has wide support from both vendor and service provider communities, it is not generally deployed as the primary routing protocol within the data center for a number of reasons:

- BGP is perceived as a "WAN only protocol only" and not often considered for enterprise or data center applications.

- BGP is believed to have a "much slower" routing convergence than traditional IGPs.

- BGP deployment within an Autonomous System typically assumes the presence of an IGP for next-hop resolution.

- BGP is perceived to require significant configuration overhead and does not support any form of neighbor auto-discovery.

This document discusses some of these perceptions, especially as applicable to the proposed design, and highlights some of the advantages of using the protocol such as:

- BGP has less complexity within its protocol design - internal data structures and state-machines are simpler when compared to a link-state IGP such as OSPF. For example, instead of implementing adjacency formation, adjacency maintenance and/or flow-control, BGP simply relies on TCP as the underlying transport. This fulfills REQ2 and REQ3.

- BGP information flooding overhead is less when compared to link-state IGPs. Since every BGP router calculates and propagates only the best-path selected, a network failure is masked as soon as the BGP speaker finds an alternate path, which exists when highly symmetric topologies, such as Clos, are coupled with EBGP only design. In contrast, the event propagation scope of a link-state IGP is an entire area, regardless of the failure type. This meets REQ3 and REQ4. It is worth mentioning that all widely deployed link-state IGPs also feature periodic refreshes of routing information, while BGP does not expire routing state, even if this rarely causes significant impact to modern router control planes.

- BGP supports third-party (recursively resolved) next-hops. This allows for manipulating multi-path to be non-ECMP based or forwarding based on application-defined forwarding paths, through establishment of a peering session with an application "controller" which can inject routing information into the system, satisfying REQ5. OSPF provides similar functionality using concepts such as "Forwarding Address", but with more difficulty in implementation and lack of protocol simplicity.

- Using a well-defined BGP ASN allocation scheme and standard AS_PATH loop detection, "BGP path hunting" can be controlled and complex unwanted paths will be ignored. See Section 5.2 for an example of a working scheme. In a link-state IGP accomplishing the same goal would require multi-(instance/topology/processes) support, typically not available in all DC devices and quite complex to configure and troubleshoot. Using a traditional single flooding domain, which most DC designs utilize, under certain failure conditions may pick up unwanted lengthy paths, e.g. traversing multiple Tier-2 devices.

- EBGP configuration that is implemented with minimal routing policy is easier to troubleshoot for network reachability issues. In most implementations, it is straightforward to view contents of BGP Loc-RIB and compare it to the router's RIB. Also every BGP neighbor has corresponding Adj-RIB-In and Adj-RIB-Out structures with incoming and outgoing NRLI information that can be easily correlated on both sides of a BGP session. Thus, BGP satisfies REQ3.

5.2. EBGP Configuration for Clos topology

Clos topologies that have more than 5 stages are very uncommon due to the large numbers of interconnects required by such a design. Therefore, the examples below are made with regards to the 5 stage Clos topology (unfolded).

5.2.1. Example ASN Scheme

The diagram below illustrates an example ASN allocation scheme. The following is a list of guidelines that can be used:

- Only EBGP sessions established over direct point-to-point links interconnecting the network nodes.

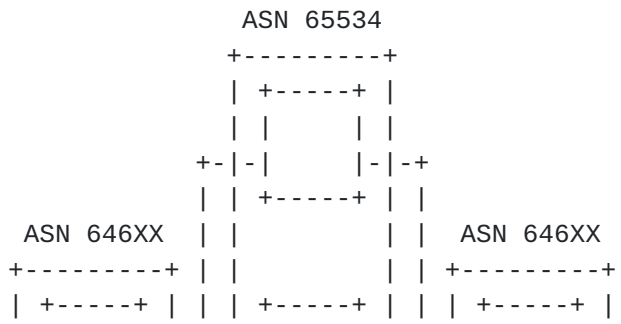
- 16-bit (two octet) BGP ASNs are used, since these are widely supported and have better vendor interoperability (e.g. no need to support BGP capability negotiation).

- Private BGP ASNs from the range 64512-65534 are used so as to avoid ASN conflicts.

- A single BGP ASN is allocated to all of the Clos topology's Tier-1 devices

- Unique BGP ASN is allocated per each group of Tier-2 devices

- Unique BGP ASN is allocated to every Tier-3 device (e.g. ToR) in this topology.



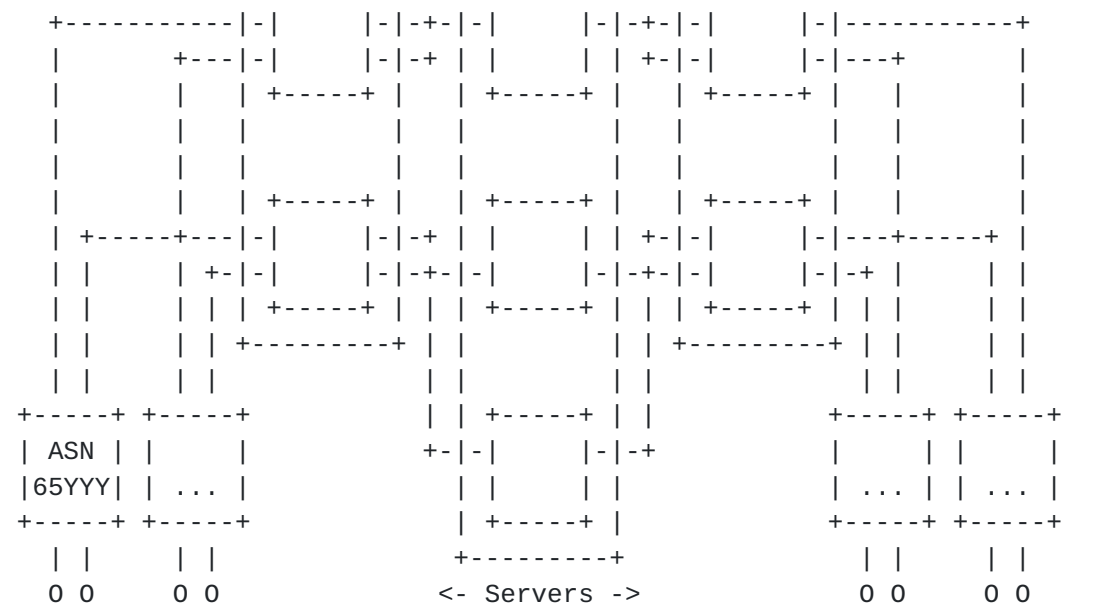


Figure 4: BGP ASN layout for 5-stage Clos

5.2.2. Private Use BGP ASNs

The original range of Private Use BGP ASNs[I-D.ietf-idr-as-private-reservation] limited operators to 1023 unique ASNs. Since it is quite likely that the number of network devices may exceed this number, a workaround is required. One approach is to re-use the ASNs assigned to the Tier-3 devices across different clusters. For example, Private Use BGP ASNs 65001, 65002 ... 65032 could be used within every individual cluster and assigned to Tier-3 devices.

To avoid route suppression due to the AS_PATH loop detection mechanism in BGP, upstream EBGP sessions on Tier-3 devices must be configured with the "AllowAS In" feature that allows accepting a device's own ASN in received route advertisements. Introducing this feature does not create an opportunity for routing loops under misconfiguration since the AS_PATH is always incremented when routes are propagated between topology tiers. Loop protection is also in place at the Tier-1 device which does not accept routes with a path including its own ASN.

Another solution to this problem would be using Four-Octet BGP ASNs [RFC6793], where there are additional Private Use ASN's available, see [IANA.AS]. Use of Four-Octet BGP ASNs put additional protocol complexity in the BGP implementation so should be considered against the complexity of re-use when considering REQ3 and REQ4. Perhaps more importantly, they are not yet supported by all BGP implementations, which may limit vendor selection of DC equipment.

5.2.3. Prefix Advertisement

A Clos topology features a large number of point-to-point links and associated prefixes. Advertising all of these routes into BGP may create FIB overload conditions in the network devices. Advertising these links also puts additional path computation stress on the BGP control plane for little benefit. There are two possible solutions:

- Do not advertise any of the point-to-point links into BGP. Since the EBGP based design changes the next-hop address at every device, distant networks will automatically be reachable via the advertising EBGP peer and do not require reachability to these prefixes. However this may complicate operational troubleshooting or monitoring systems if the addresses are not reachable.

- Advertise point-to-point links, but summarize them on every device. This requires an address allocation scheme such as allocating a consecutive block of IP addresses per Tier-1 and Tier-2 device to be used for point-to-point interface addressing to the lower layers (Tier-2 uplinks will be numbered out of Tier-1 addressing and so forth).

Server subnets on Tier-3 devices must be announced into BGP without using route aggregation on Tier-2 and Tier-1 devices. Summarizing subnets in a Clos topology will result in route black-holing under a single link failure (e.g. between Tier-2 and Tier-3 devices) and must be avoided. The use of peer links within the same tier to resolve the black-holing problem by providing "bypass paths" is undesirable due to O(N^2) complexity of the peering mesh and waste of ports on the devices. In Section 8.2 a method for performing route summarization in Clos networks and the associated trade-offs is described.

5.2.4. External Connectivity

A dedicated cluster (or clusters) in the Clos topology could be used for the purpose of connecting to the Wide Area Network (WAN) edge devices, or WAN Routers. Tier-3 devices in such a cluster would be replaced with WAN Routers, and EBGP peering would be used again, though WAN routers are likely to belong to a public ASN if Internet connectivity is required in the design.

The Tier-2 devices in such a dedicated cluster will be referred to as "Border Routers" in this document. These devices have to perform a few special functions:

- Hide network topology information when advertising paths to WAN routers, i.e. remove Private BGP ASNs from the AS_PATH attribute.

This is typically done to avoid ASN number collisions between different data centers. An implementation specific BGP feature typically called "Remove Private AS" is commonly used to accomplish this. Depending on implementation, the feature should strip a contiguous sequence of private ASNs found in AS_PATH attribute prior to advertising the path to a neighbor. This assumes that all BGP ASN's used for intra data center numbering are from the private ASN range.

- Originate a default route to the data center devices. This is the only place where default route can be originated, as route summarization is highly undesirable for the "scale-out" topology. Alternatively, Border Routers may simply relay the default route learned from WAN routers. Advertising the default route from Border Routers requires that all Border Routers to be fully connected to the WAN Routers upstream, to provide resistance to a single-link failure causing the black holing of traffic. To prevent chance of operator or implementation error that may impact EBGP sessions to the WAN routers simultaneously (although these scenarios are not planned for by many operators since they represents a multiple failure) it may be more desirable to take this approach than introducing complicated conditional default origination schemes provided by some implementations.

5.2.5. Route Aggregation at the Edge

It is often desirable to aggregate network reachability information prior to advertising it to the WAN network due to high amount of IP prefixes originated from within the data center with a fully routed network design. For example, a network with 2000 Tier-3 devices will have at least 2000 servers subnets advertised into BGP, along with the infrastructure or other prefixes. However, as discussed before, the proposed network design does not allow for route aggregation due to the lack of peer links inside every tier.

However, it is possible to lift this restriction for the Border Routers, by devising a different connectivity model for these devices. There are two options possible:

- Interconnect the Border Routers using a full-mesh of physical links or by using additional aggregation devices, forming hub-and-spoke topology. Next, build a full-mesh of IBGP sessions between all Border Routers to allow for sharing of specific network prefixes. Notice that in this case the interconnecting peer links need to be appropriately sized for the amount of traffic that will be present in the case of a device or link failure underneath the Border Routers.

o Tier-1 devices may have additional physical links running toward the Border Routers (which are Tier-2 devices in essence). Specifically, if protection from a single link or node failure is desired, each Tier-1 devices would have to connect to at least two Border Routers. This puts additional requirements on the port count for Tier-1 devices and Border Routers, potentially making it a non-uniform, larger port count, device with the other devices in the Clos.

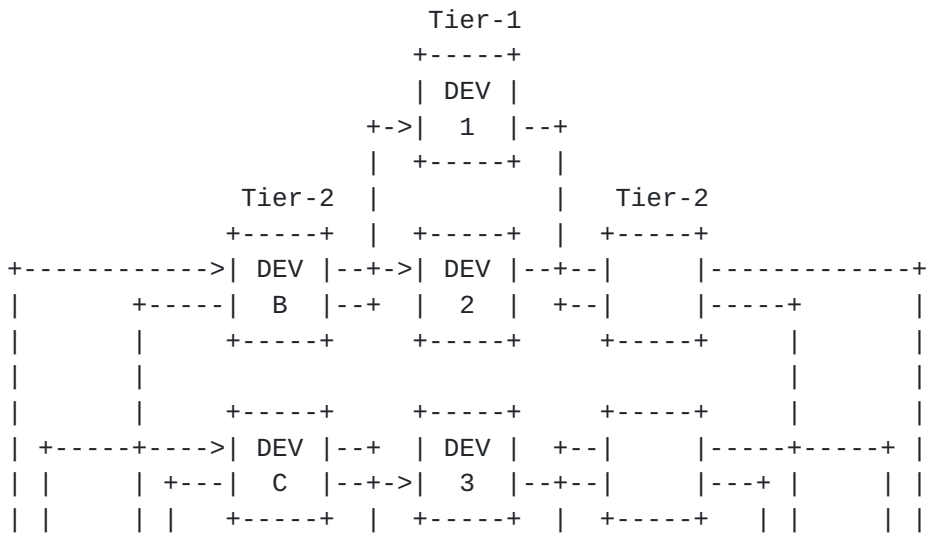
If any of the above options are implemented, it is possible to perform route aggregation at the Border Routers toward the WAN network core without risking a routing black-hole condition under a single link failure. Both of the options would result in non-uniform topology as additional links have to be provisioned on some network devices.

6. ECMP Considerations

This section covers the Equal Cost Multipath (ECMP) functionality for Clos topology and discusses a few special requirements.

6.1. Basic ECMP

ECMP is the fundamental load-sharing mechanism used by a Clos topology. Effectively, every lower-tier device will use all of its directly attached upper-tier devices to load-share traffic destined to the same IP prefix. Number of ECMP paths between any two Tier-3 devices in Clos topology equals to the number of the devices in the middle stage (Tier-1). For example, Figure 5 illustrates the topology where Tier-3 device A has four paths to reach servers X and Y, via Tier-2 devices B and C and then Tier-1 devices 1, 2, 3, and 4 respectively.



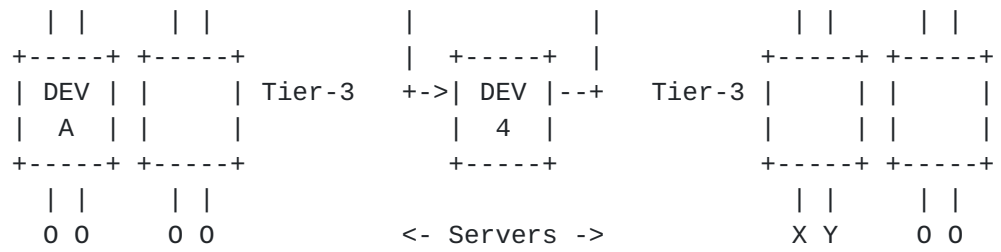


Figure 5: ECMP fan-out tree from A to X and Y

The ECMP requirement implies that the BGP implementation must support multi-path fan-out for up to the maximum number of devices directly attached at any point in the topology in upstream or downstream direction. Normally, this number does not exceed half of the ports found on a device in the topology. For example, an ECMP max-path of 32 would be required when building a Clos network using 64-port devices. The Border Routers may need to have wider fan-out to be able to connect to multitude of Tier-1 devices if route summarization at Border Router level is implemented as described in Section 5.2.5. If a device's hardware does not support wider ECMP, logical link-grouping (link-aggregation at layer 2) could be used to provide "hierarchical" ECMP (Layer 3 ECMP followed by Layer 2 ECMP) to compensate for fan-out limitations. Such approach, however, increases the risk of flow polarization, as less entropy will be available to the second stage of ECMP.

Most BGP implementations declare paths to be equal from ECMP perspective if they match up to and including step (e) Section 9.1.2.2 of [RFC4271]. In the proposed network design there is no underlying IGP, so all IGP costs are assumed to be zero or otherwise the same value across all paths and policies may be applied as necessary to equalize BGP attributes that vary in vendor defaults, as has been seen occasionally with MED and origin code. Routing loops are unlikely due to the BGP best-path selection process which prefers shorter AS_PATH length, and longer paths through the Tier-1 devices which don't allow their own AS in the path and have the same ASN are also not possible.

6.2. BGP ECMP over Multiple ASNs

For application load-balancing purposes it is desirable to have the same prefix advertised from multiple Tier-3 devices. From the perspective of other devices, such a prefix would have BGP paths with different AS_PATH attribute values, while having the same AS_PATH attribute lengths. Therefore, BGP implementations must support load-sharing over above-mentioned paths. This feature is sometimes known as "multipath relax" and effectively allows for ECMP to be done across different neighboring ASNs if all other attributes are equal as described in the previous section.

6.3. Weighted ECMP

It may be desirable for the network devices to implement weighted ECMP, to be able to send more traffic over some paths in ECMP fan-out. This could be helpful to compensate for failures in the network and send more traffic over paths that have more capacity. The prefixes that require weighted ECMP would have to be injected using remote BGP speaker (central agent) over a multihop session as described further in Section 8.1. If support in implementations is available, weight-distribution for multiple BGP paths could be signaled using the technique described in [I-D.ietf-idr-link-bandwidth].

7. BGP Convergence of Design

This section reviews routing convergence properties in the proposed design. A case is made that sub-second convergence is achievable provided that the implementation supports fast EBGP peering session deactivation and timely RIB and FIB update upon failure of the associated link.

7.1. Fault Detection Timing

BGP typically relies on an IGP to route around link/node failures inside an AS, and implements either a polling based or an event-driven mechanism to obtain updates on IGP state changes. The proposed routing design does not use an IGP, so the only mechanisms that could be used for fault detection are BGP keep-alive process (or any other type of keep-alive mechanism) and link-failure triggers.

Relying solely on BGP keep-alive packets may result in high convergence delays, in the order of multiple seconds (on many BGP implementations the minimum configurable BGP hold timer value is three seconds). However, many BGP implementations can shut down local EBGP peering sessions in response to the "link down" event for the outgoing interface used for BGP peering. This feature is sometimes called as "fast fall-over". Since links in modern data centers are often point-to-point fiber connections, a physical

interface failure is often detected in milliseconds and subsequently triggers a BGP re-convergence.

Ethernet technologies may support failure signaling or detection standards such as [IEEE8021AG] and [IEEE8023AH], which may make failure detection more robust. Alternatively, some platforms may support Bidirectional Forwarding Detection (BFD) [RFC5880] to allow for sub-second failure detection and fault signaling to the BGP process. However use of either of these presents additional requirements to vendor software and possibly hardware, and may contradict REQ1.

7.2. Event Propagation Timing

In this design the impact of BGP Minimum Route Advertisement Interval (MRAI) timer (See section 9.2.1.1 of [RFC4271]) should be considered. It is required for BGP implementations to space out consecutive BGP UPDATE messages by at least MRAI seconds, which is often a configurable value. Notice that initial BGP UPDATE messages after an event carrying withdrawn routes are commonly not affected by this timer. The MRAI timer may present significant convergence delays if a BGP speaker "waits" for the new path to be learned from peers and has no local backup path information.

In a Clos topology each EBGP speaker has either one path only or N paths for the same prefix, where N is a significantly large number, e.g. N=32. Therefore, if a path fails there is either no backup at all, or the backup is readily available in BGP Loc-RIB. In the first case, the BGP withdrawal announcement will propagate un-delayed and trigger re-convergence on affected devices. In the second case, only the local ECMP group needs to be changed.

7.3. Impact of Clos Topology Fan-outs

Clos topology has large fan-outs, which may impact the "Up->Down" convergence in some cases, as described in this section. In a situation when a link between Tier-3 and Tier-2 device fails, the Tier-2 device will send BGP WITHDRAW message to all upstream Tier-1 devices, and Tier-1 devices will relay this message to all downstream Tier-2 devices. A Tier-2 device other than the one originating the WITHDRAW should wait for ALL adjacent Tier-1 devices to send a WITHDRAW message before it removes the affected prefixes and sends WITHDRAW downstream to Tier-3 devices. If the original Tier-2 device or the relaying Tier-1 devices introduce some delay into their announcements, the result could be WITHDRAW message "dispersion", that could be as much as multiple seconds. In order to avoid such behavior, BGP implementations must support "update groups", where a BGP message is built once for a group of neighbors, which typically

must have the same outgoing policy, which will receive this update and then advertised synchronously to all neighbors.

The impact of such "dispersion" grows with the size of topology fan-out and could also grow under network convergence churn.

7.4. Failure Impact Scope

A network is declared to converge in response to a failure once all devices within the failure impact scope are notified of the event and have re-calculated their RIB's and consequently FIB's. Larger failure impact scope typically means slower convergence since more devices have to be notified, and additionally results in a less stable network. In this section BGP's advantages over link-state routing protocols in reducing failure impact scope when implemented in a Clos topology are described.

BGP is similar to a distance-vector protocol as only the best path from the point of view of the local router is sent to neighbors and routers do not maintain a full view of the topology. As such, some failures are masked if the local node can immediately find a backup path. In the worst case ALL devices in a data center topology have to either withdraw a prefix completely or update the ECMP groups in the FIB. However, many failures will not result in such a wide impact. There are two main failure types where impact scope is reduced:

- Failure of a link between Tier-2 and Tier-1 devices: In this case, a Tier-2 device will update its ECMP group, removing the failed link. There is no need to send new information to downstream Tier-3 devices. The affected Tier-1 device will lose the only path available to reach a particular cluster and will have to withdraw the associated prefixes. Such prefix withdrawal process will only affect Tier-2 devices directly connected to the affected Tier-1 device. The Tier-2 devices receiving the BGP UPDATE messages withdrawing prefixes will simply have to update their ECMP groups. The Tier-3 devices are not involved in the re-convergence process.

- Failure of a Tier-1 device: In this case, all Tier-2 devices directly attached to the failed node will have to update their ECMP groups for all IP prefixes from non-local cluster. The Tier-3 devices are once again not involved in the re-convergence process.

Even though in case of such failures multiple IP prefixes will have to be reprogrammed in the FIB, it is worth noting that ALL of these prefixes share a single ECMP group on Tier-2 device. Therefore, in

the case of implementations with a hierarchical FIB, only a single change has to be made to the FIB.

Even though BGP offers some failure scope reduction, reduction of the fault domain using summarization is not always possible with the proposed design, since using this technique may create routing black-holes as mentioned previously. Therefore, the worst control-plane failure impact scope is the network as a whole, for instance in a case of a link failure between Tier-2 and Tier-3 devices. The amount of impacted prefixes in this case would be much less than in the case of a failure in the upper layers of a Clos network topology. The property of having such large failure scope is not a result of choosing EBGP in the design but rather a result of using the "scale-out" Clos topology.

7.5. Routing Micro-Loops

When a downstream device, e.g. Tier-2 device, loses a path for a prefix, it normally has the default route pointing toward the upstream device, in this case the Tier-1 device. As a result, it is possible to get in the situation when Tier-2 switch loses a prefix, but Tier-1 switch still has the path which results in transient micro-loop, since Tier-1 switch will keep passing packets to the affected prefix back to Tier-2 device, and Tier-2 will bounce it back again using the default route. This micro-loop will last for the duration of time it takes the upstream device to fully update its forwarding tables.

To minimize impact of the micro-loops, Tier-2 and Tier-1 switches can be configured with static "discard" or "null0" routes that will be more specific than the default route for specific prefixes missing during network convergence. For Tier-2 switches, the discard route should be an aggregate route, covering all server subnets of the underlying Tier-3 devices. For Tier-1 devices, the discard route should be an aggregate covering the server IP address subnet allocated for the whole data-center. Those discard routes will only take precedence for the duration of network convergence, until the device learns a more specific prefix via a new path.

8. Additional Options for Design

8.1. Third-party Route Injection

BGP allows for a "third-party", or not directly attached, BGP speaker to inject routes anywhere in the network topology, meeting REQ5. This can be achieved by peering using a multihop BGP session with some or even all devices in the topology. Furthermore, BGP diverse path distribution [RFC6774] could be used to inject multiple BGP next

hops for the same prefix to facilitate load-balancing, or using the BGP ADD-PATH capability [I-D.ietf-idr-add-paths] if supported by the implementation. Unfortunately in many implementations ADD-PATH has been found to only support IBGP properly due to the use cases it was originally optimized for.

To implement route injection in the proposed design a third-party BGP speaker may peer with Tier-3 and Tier-1 switches, injecting the same prefix, but using a special set of BGP next-hops for Tier-1 devices. Those next-hops are assumed to resolve recursively via BGP, and could be, for example, IP addresses on Tier-3 devices. The resulting forwarding table programming could provide desired traffic proportion distribution among different clusters.

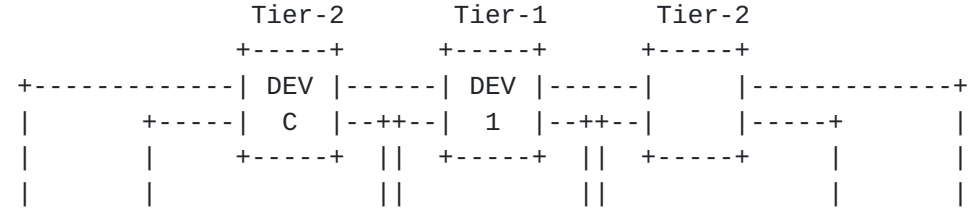
8.2. Route Aggregation within Clos Topology

As mentioned previously, route aggregation is not possible within the proposed Clos topology since it makes the network susceptible to route black-holing under single link failures. The main problem is the limited number of parallel paths between network elements, such as when there is only a single path between any pair of Tier-1 and Tier-3 devices. However, some operators may find route aggregation desirable to improve control plane stability.

By changing the network topology route aggregation can be allowed, if necessary, though the trade-off would be reduction of the total size of the network as well as network congestion under specific failures. This approach is very similar to the technique described above, which allows Border Routers to summarize the entire data-center address space.

8.2.1. Collapsing Tier-1 Devices Layer

In order to add more paths between Tier-1 and Tier-3 devices, group Tier-2 devices into pairs, and then connect the pairs to the same group of Tier-1 devices. This is logically equivalent to "collapsing" Tier-1 devices into a group of half the size, merging the links on the "collapsed" devices. The result is illustrated in Figure 6. For example, in this topology DEV C and DEV D connect to the same set of Tier-1 devices (DEV 1 and DEV 2), whereas before they were connecting to different groups of Tier-1 devices.



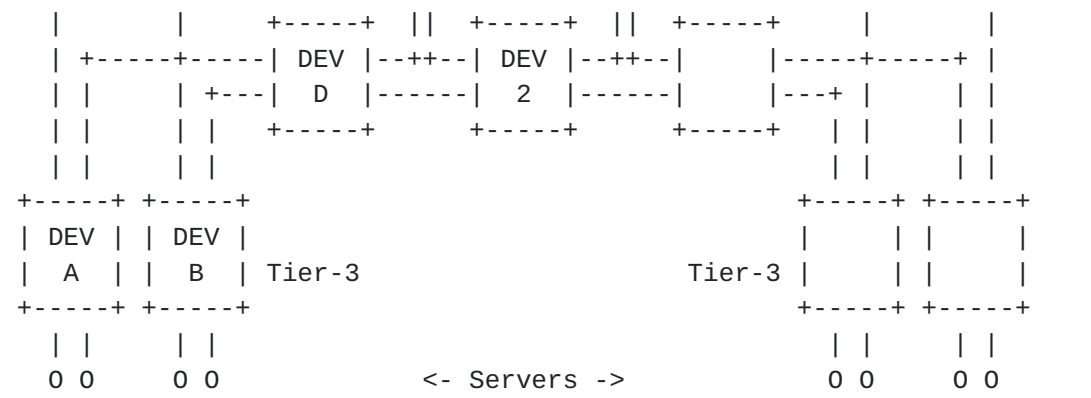


Figure 6: 5-Stage Clos topology

In this design choice, Tier-2 devices may be configured to advertise only a default route down to Tier-3 devices. If a link between Tier-2 and Tier-3 fails, the traffic will be re-routed via the second available path known to a Tier-2 switch. It is not possible to advertise a summary route covering prefixes for a single cluster from Tier-2 devices since each of them has only a single path down to this prefix. It would require dual-homed servers to accomplish that. Also note that this design is only resilient to single link failure. It is possible for a double link failure to isolate a Tier-2 device from all paths toward a specific Tier-3 device, thus causing a routing black-hole.

8.2.2. Implications of Collapsing Tier-1 Devices Layer

As mentioned already, a result of the proposed topology modification would be reduction of Tier-1 devices port capacity. This limits the maximum number of attached Tier-2 devices and therefore will limit the maximum DC network size. A larger network would require different Tier-1 devices that have higher port density to implement this change.

Another problem is traffic re-balancing under link failures. Since three are two paths from Tier-1 to Tier-3, a failure of the link between Tier-1 and Tier-2 switch would result in all traffic that was taking the failed link to switch to the remaining path. This will result in doubling of link utilization on the remaining link.

9. Security Considerations

The design does not introduce any additional security concerns. General BGP security considerations are discussed in [RFC4271] and [RFC4272]. Furthermore, the Generalized TTL Security Mechanism [RFC5082] could be used to reduce the risk of BGP session spoofing.

10. IANA Considerations

This document includes no request to IANA.

11. Acknowledgements

This publication summarizes work of many people who participated in developing, testing and deploying the proposed network design, some of whom were George Chen, Parantap Lahiri, Dave Maltz, Edet Nkposong, Robert Toomey, and Lihua Yuan. Authors would also like to thank Linda Dunbar and Susan Hares for reviewing the document and providing valuable feedback and Mary Mitchell for grammar and style suggestions.

12. References

12.1. Normative References

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[I-D.ietf-idr-as-private-reservation] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", draft-ietf-idr-as-private-reservation-05 (work in progress), May 2013.

12.2. Informative References

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, December 2006.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC6325] Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC6774] Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, November 2012.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, December 2012.

[I-D.ietf-idr-add-paths] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-08 (work in progress), December 2012.

[I-D.ietf-idr-link-bandwidth] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-06 (work in progress), January 2013.

[GREENBERG2009] Greenberg, A., Hamilton, J., and D. Maltz, "The Cost of a Cloud: Research Problems in Data Center Networks", January 2009.

[IEEE8021AG] IEEE 802.1Q, ., "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", October 2012.

[IEEE8023AH] IEEE 802.3, ., "IEEE Standard for Information technology - Local and metropolitan area networks - Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications", December 2008.

[INTERCON] Dally, W. and B. Towles, "Principles and Practices of Interconnection Networks", ISBN 978-0122007514, January 2004.

[ALFARES2008] Al-Fares, M., Loukissas, A., and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture", August 2008.

[IANA.AS] IANA, ., "Autonomous System (AS) Numbers", July 2013, <http://www.iana.org/assignments/as-numbers/>.

Authors' Addresses

Petr Lapukhov
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
US

Phone: +1 425 703 2723
Email: petrlapu@microsoft.com
URI: http://microsoft.com/

Ariff Premji
Arista Networks
5470 Great America Parkway
Santa Clara, CA 95054
US

Phone: +1 408 547 5699
Email: ariff@aristanetworks.com
URI: http://aristanetworks.com/

Jon Mitchell (editor)
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
US

Email: Jon.Mitchell@microsoft.com