Internet Engineering Task Force (IETF)
Request for Comments: 7938
Category: Informational
ISSN: 2070-1721

P. Lapukhov
Facebook
A. Premji
Arista Networks
J. Mitchell, Ed.

August 2016

Use of BGP for Routing in Large-Scale Data Centers

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 document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7938.

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 presented in this document is based on operational experience with data centers built to support large-scale distributed software infrastructure, such as a web search engine. The primary requirements in such an environment are operational simplicity and network stability so that a small group of people can effectively support a significantly sized network.

Experimentation and extensive testing have shown that External BGP (EBGP) [RFC4271] is well suited as a stand-alone routing protocol for these types of data center applications. This is in contrast with more traditional DC designs, which may use simple tree topologies and rely on extending Layer 2 (L2) 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 exploring 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 [CLOS1953] that are horizontally scaled

out. This is followed by arguments for selecting EBGP with a Clos topology as the most appropriate routing protocol to meet the requirements and the proposed design is described in detail. Finally, this document reviews some additional considerations and design options. A thorough understanding of BGP is assumed by a reader planning on deploying the design described within the document.

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 a 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, commonly referred to as "north-south" traffic. Traditional "tree" topologies were sufficient to accommodate such flows, even with high oversubscription ratios between the layers of the network. If more bandwidth was required, it was added by "scaling up" the network elements, e.g., by upgrading the device's linecards 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, which does not egress the DC, commonly referred to as "east-west" traffic. Examples of such applications could be computer clusters such as Hadoop [HADOOP], massive data replication between clusters needed by certain applications, or virtual machine migrations. Scaling traditional tree topologies to match these bandwidth demands becomes either too expensive or impossible due to physical limitations, e.g., port density in a switch.

2.2. CAPEX Minimization

The Capital Expenditures (CAPEX) associated with the network infrastructure alone constitutes about 10-15% of total data center expenditure (see [GREENBERG2009]). However, the absolute cost is significant, and hence 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 volume pricing on bulk purchases and reduced maintenance and inventory costs.

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

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

2.3. OPEX Minimization

Operating large-scale infrastructure can be expensive as a 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 Operational Expenditure (OPEX) minimization is reducing the size of failure domains in the network. Ethernet networks are known to be susceptible to broadcast or unicast traffic storms that can have a dramatic impact on network performance and availability. The use of a fully routed design significantly reduces the size of the data-plane failure domains, i.e., limits them to the lowest level in the network hierarchy. However, such designs introduce the problem of distributed control-plane failures. This observation calls for simpler and less control-plane protocols to reduce protocol interaction issues, reducing the chance of a network meltdown. Minimizing software feature requirements as described in the CAPEX section above also reduces testing and training requirements.

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 the load-balancing layer horizontally, by adding more of the uniform nodes and distributing incoming traffic across these nodes. In situations 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 the 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 some 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 looks like an (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 the 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.

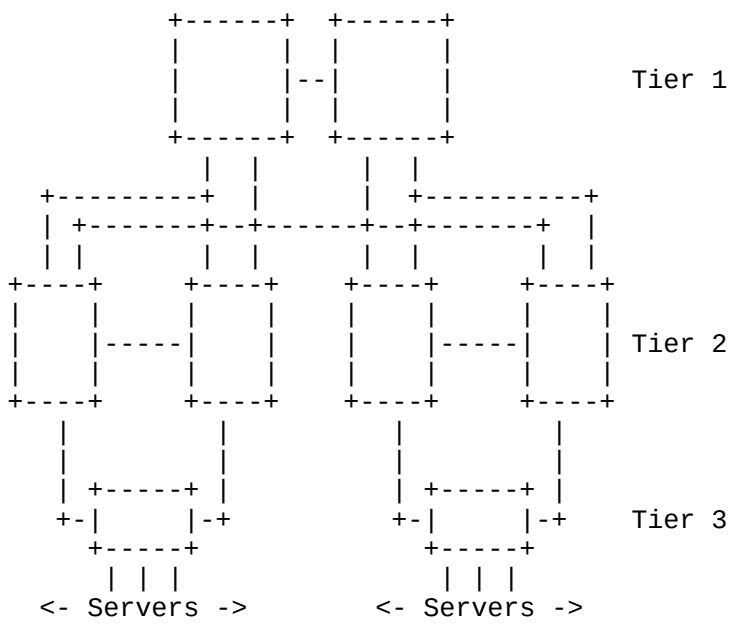

Figure 1: Typical DC Network Topology

Unfortunately, as noted previously, it is not possible to scale a tree-based design to a large enough degree for handling large-scale designs due to the inability to be able to acquire Tier 1 devices with a large enough port density to sufficiently scale Tier 2. Also, continuous upgrades or replacement of the upper-tier devices are required as deployment size or bandwidth requirements increase, which is operationally complex. For this reason, REQ1 is in place, eliminating this type of design from consideration.

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. Overview

A common choice for a horizontally scalable topology is a folded Clos topology, sometimes called "fat-tree" (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 folded Clos topology satisfies REQ1 and

facilitates REQ2. See Figure 2 below for an example of a folded 3-stage Clos topology (3 stages counting Tier 2 stage twice, when tracing a packet flow):

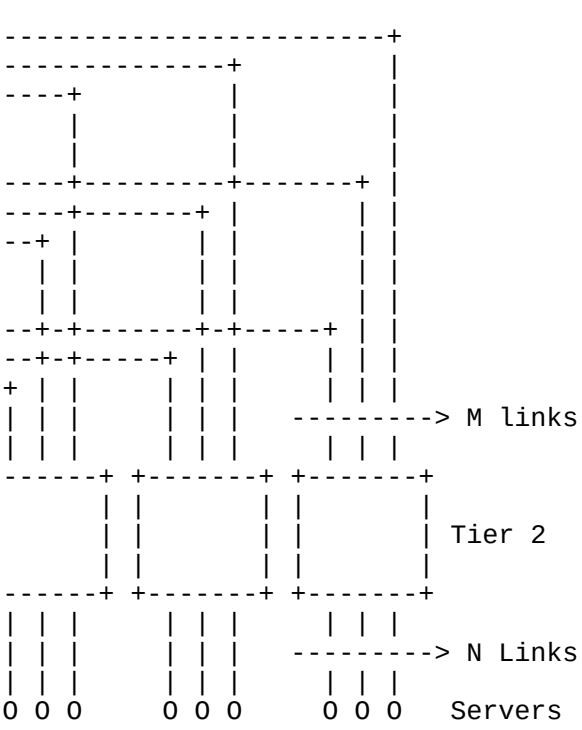

Figure 2: 3-Stage Folded Clos Topology

This topology is often 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 control and data-plane support for ECMP with a fan-out of M or more.

- Tier 1 switches have exactly one path to every server in this topology. This is an important property that makes route summarization dangerous in this topology (see Section 8.2 below).

- 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 by adding more stages, e.g., moving to a 5-stage Clos, as illustrated in Figure 3 below:

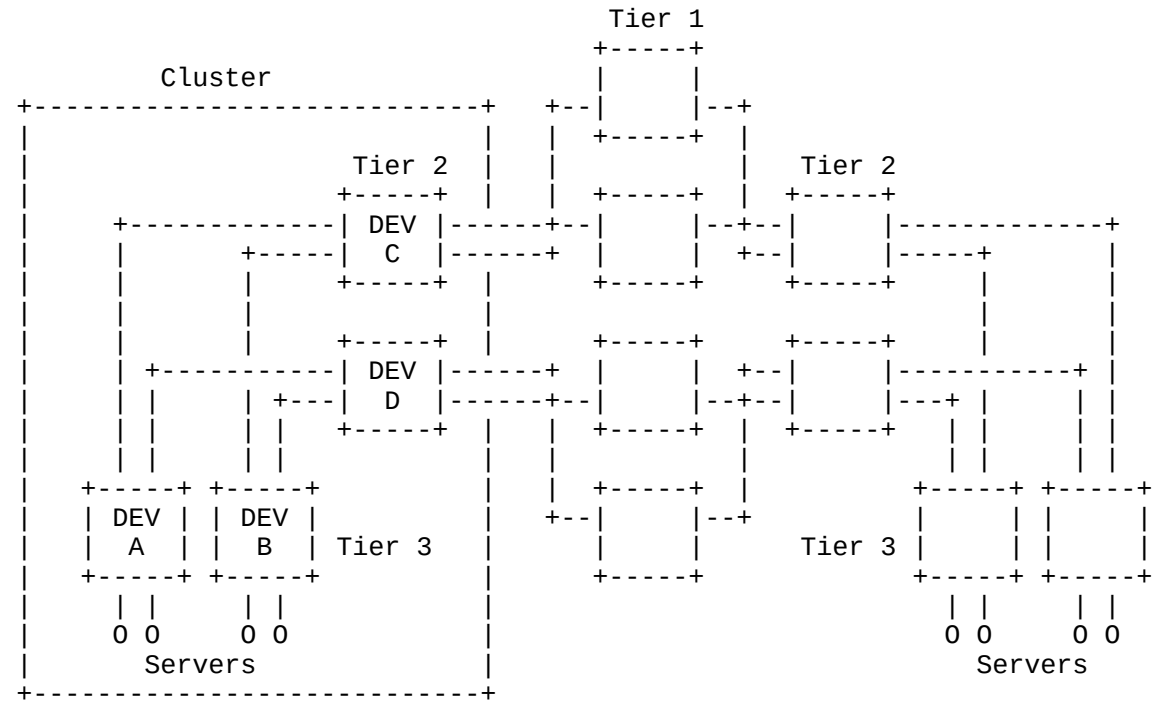

Figure 3: 5-Stage Clos Topology

The small example of topology in Figure 3 is built from devices with a port count of 4. 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. The

concept of a cluster may also be a useful concept as a single deployment or maintenance unit that can be operated on at a different frequency than the entire topology.

In practice, Tier 3 of the network, which is 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 clusters (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 a Clos topology by a factor 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 bandwidth while reducing the number of elements in Tier 1, 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, possibly resulting in heavy congestion and quality of service degradation if the path determination procedure does not take bandwidth amount into account, since the number of upstream Tier 1 devices is likely wider than two. To avoid this situation, parallel links can be grouped in link aggregation groups (LAGs), e.g., [IEEE8023AD], 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 rebalanced over the multitude of remaining paths that equals the number of Tier 1 devices. This example is using two links for simplicity, having more links in a bundle will have less impact on capacity upon a member-link failure.

4. Data Center Routing Overview

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

4.1. L2-Only Designs

Originally, most data center designs used Spanning Tree Protocol (STP) originally defined in [IEEE8021D-1990] for loop-free topology creation, typically utilizing variants of the traditional DC topology described in Section 3.1. At the time, many DC switches either did not support Layer 3 routing protocols or supported them 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 (RSTP) in the latest revision of [IEEE8021D-2004] and Multiple Spanning Tree Protocol (MST) specified in [IEEE8021Q] that increase convergence, stability, and load-balancing in larger topologies, many of the fundamentals of the protocol limit its applicability in large-scale DCs. STP and its newer variants use an active/standby approach to path selection, and are therefore hard to deploy in horizontally scaled topologies as 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 affected 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. Various enhancements to link-aggregation protocols such as [IEEE8023AD], generally known as Multi-Chassis Link-Aggregation (M-LAG) made it possible to use Layer 2 designs with active-active network paths while relying on STP as the backup for loop prevention. The major downsides of this approach are the lack of ability to scale linearly past two in most implementations, lack of standards-based implementations, and the added failure domain risk of syncing state between the devices.

It should be noted that building large, horizontally scalable, L2-only networks without STP is possible recently through the introduction of the Transparent Interconnection of Lots of Links (TRILL) protocol in [RFC6325]. TRILL resolves many of the issues STP has for large-scale DC design however, due to the limited number of implementations, and often the requirement for specific equipment that supports it, this has limited its applicability and increased the cost of such designs.

Finally, neither the base TRILL specification nor the M-LAG approach totally eliminate the problem of the shared broadcast domain that is so detrimental to the operations of any Layer 2, Ethernet-based solution. Later TRILL extensions have been proposed to solve the this problem statement, primarily based on the approaches outlined in [RFC7067], but this even further limits the number of available interoperable implementations that can be used to build a fabric. Therefore, TRILL-based designs have issues meeting REQ2, REQ3, and REQ4.

4.2. Hybrid L2/L3 Designs

Operators have sought to limit the impact of data-plane faults and build large-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 managing multiple network protocols. For the following reasons, operators have 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 a different Tier 3 switch.
- Simplified IP addressing = less IP subnets are 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). See [L3DSR].
- Continued CAPEX differences between L2- and L3-capable switches.

4.3. L3-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 Interior Gateway Protocol (IGP) such as Open Shortest Path First (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.

Choosing a L3-only design greatly simplifies the network, facilitating the meeting of REQ1 and REQ2, and has widespread 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 develop new solutions to meet some of the requirements that previously had driven large Layer 2 domains by using various overlay or tunneling techniques.

5. Routing Protocol 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-Domain Routing in 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 (some of which are interrelated):

- 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 compared to IGPs.
- Large-scale BGP deployments typically utilize an IGP for BGP next-hop resolution as all nodes in the IBGP topology are not directly connected.
- BGP is perceived to require significant configuration overhead and does not support 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 in parts of its protocol design -- internal data structures and state machine are simpler as compared to most link-state IGPs 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 an EBGP-only design. In contrast, the event propagation scope of a link-state IGP is an entire area, regardless of the failure type. In this way, BGP better meets REQ3 and REQ4. It is also worth mentioning that all widely deployed link-state IGPs feature periodic refreshes of routing information while BGP does not expire routing state, although this rarely impacts modern router control planes.

- BGP supports third-party (recursively resolved) next hops. This allows for manipulating multipath to be non-ECMP-based or forwarding-based on application-defined paths, through establishment of a peering session with an application "controller" that 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 far less control of information propagation scope.

- Using a well-defined Autonomous System Number (ASN) allocation scheme and standard AS_PATH loop detection, "BGP path hunting" (see [JAKMA2008]) can be controlled and complex unwanted paths will be ignored. See Section 5.2 for an example of a working ASN allocation scheme. In a link-state IGP, accomplishing the same goal would require multi-(instance/topology/process) 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 the BGP Loc-RIB and compare it to the router's Routing Information Base (RIB). Also, in most implementations, an operator can view every BGP neighbors Adj-RIB-In and Adj-RIB-Out structures, and therefore incoming and outgoing Network Layer Reachability Information (NLRI) information 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 reference to the 5-stage Clos topology (in unfolded state).

5.2.1. EBGP Configuration Guidelines and Example ASN Scheme

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

- EBGP single-hop sessions are established over direct point-to-point links interconnecting the network nodes, no multi-hop or loopback sessions are used, even in the case of multiple links between the same pair of nodes.

- Private Use ASNs from the range 64512-65534 are used to avoid ASN conflicts.

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

- A unique ASN is allocated to each set of Tier 2 devices in the same cluster.

- A unique ASN is allocated to every Tier 3 device (e.g., ToR) in this topology.

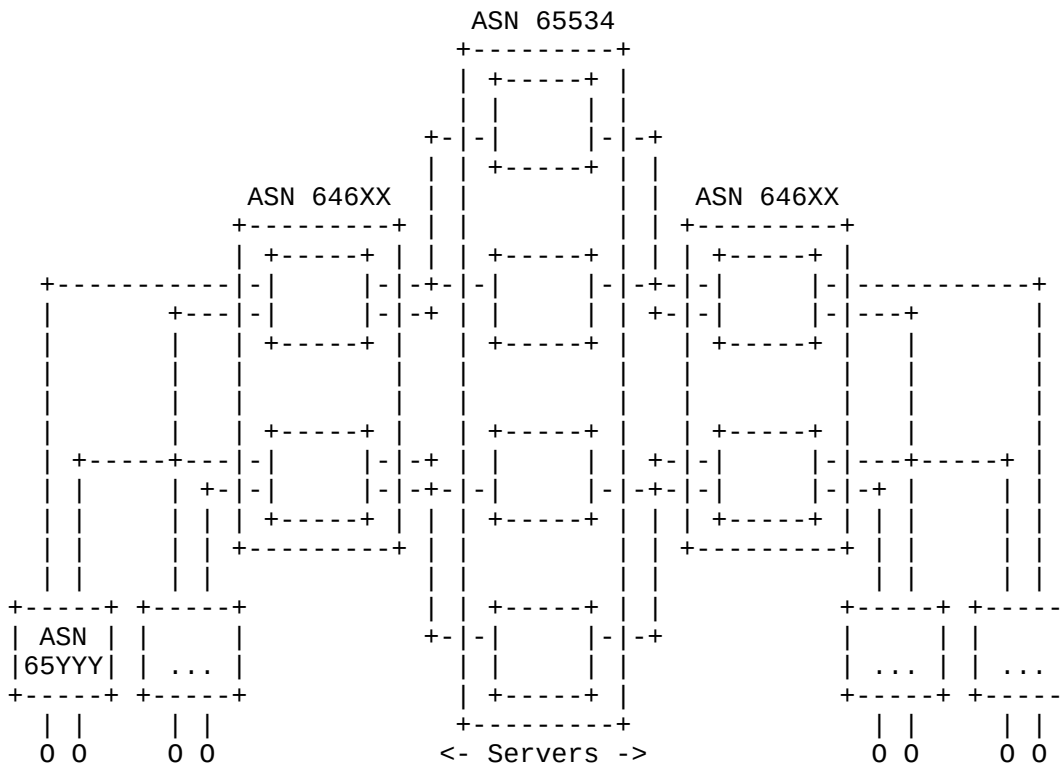

Figure 4: BGP ASN Layout for 5-Stage Clos

5.2.2. Private Use ASNs

The original range of Private Use ASNs [RFC6996] 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 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 [ALLOWASIN] that allows accepting a device's own ASN in received route advertisements. Although this feature is not standardized, it is widely available across multiple vendors implementations. Introducing this feature does not make routing loops more likely in the design since the AS_PATH is being added to by routers at each of the topology tiers and AS_PATH length is an early tie breaker in the BGP path selection

process. Further loop protection is still in place at the Tier 1 device, which will not accept routes with a path including its own ASN. Tier 2 devices do not have direct connectivity with each other.

Another solution to this problem would be to use Four-Octet ASNs ([RFC6793]), where there are additional Private Use ASNs available, see [IANA.AS]. Use of Four-Octet ASNs puts additional protocol complexity in the BGP implementation and should be balanced 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. When supported, ensure that deployed implementations are able to remove the Private Use ASNs when external connectivity (Section 5.2.4) to these ASNs is required.

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 Forwarding Information Base (FIB) overload 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 operations or monitoring: e.g., using the popular "traceroute" tool will display IP addresses that 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 allocated from Tier 1 address blocks and so forth).

Server subnets on Tier 3 devices must be announced into BGP without using route summarization on Tier 2 and Tier 1 devices. Summarizing subnets in a Clos topology results in route black-holing under a single link failure (e.g., between Tier 2 and Tier 3 devices), and hence 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. An alternative to the full mesh of peer links would be to use a simpler bypass topology, e.g., a "ring" as

described in [FB4POST], but such a topology adds extra hops and has limited bandwidth. It may require special tweaks to make BGP routing work, e.g., splitting every device into an ASN of its own. Later in this document, Section 8.2 introduces a less intrusive method for performing a limited form of route summarization in Clos networks and discusses its associated tradeoffs.

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 Use ASNs [RFC6996] from the AS_PATH attribute. This is typically done to avoid ASN number collisions between different data centers and also to provide a uniform AS_PATH length to the WAN for purposes of WAN ECMP to anycast prefixes originated in the topology. 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 Use ASNs found in an AS_PATH attribute prior to advertising the path to a neighbor. This assumes that all ASNs used for intra data center numbering are from the Private Use ranges. The process for stripping the Private Use ASNs is not currently standardized, see [REMOVAL]. However, most implementations at least follow the logic described in this vendor’s document [VENDOR-REMOVE-PRIVATE-AS], which is enough for the design specified.

- Originate a default route to the data center devices. This is the only place where a default route can be originated, as route summarization is risky for the unmodified Clos 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 be fully connected to the WAN Routers upstream, to provide resistance to a single-link failure causing the black-holing of traffic. To prevent black-holing in the situation when all of the EBGP sessions to the WAN routers fail simultaneously on a given device, it is more desirable to readvertise the default route rather than originating the default route via complicated conditional route origination schemes provided by some implementations [CONDITIONALROUTE].

5.2.5. Route Summarization at the Edge

It is often desirable to summarize network reachability information prior to advertising it to the WAN network due to the high amount of IP prefixes originated from within the data center in 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 prefixes. However, as discussed in Section 5.2.3, the proposed network design does not allow for route summarization 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 using any other "peer-mesh" topology, such as ring or hub-and-spoke. Configure BGP accordingly on all Border Leafs to exchange network reachability information, e.g., by adding a mesh of IBGP sessions. The interconnecting peer links need to be appropriately sized for traffic that will be present in the case of a device or link failure in the mesh connecting the Border Routers.

- Tier 1 devices may have additional physical links provisioned toward the Border Routers (which are Tier 2 devices from the perspective of Tier 1). Specifically, if protection from a single link or node failure is desired, each Tier 1 device 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 nonuniform, larger port count, device compared with the other devices in the Clos. This also reduces the number of ports available to "regular" Tier 2 switches, and hence the number of clusters that could be interconnected via Tier 1.

If any of the above options are implemented, it is possible to perform route summarization 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 nonuniform 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. The number of ECMP paths between any two Tier 3 devices in Clos topology is equal to the number of the devices in the middle stage (Tier 1). For example, Figure 5 illustrates a 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.

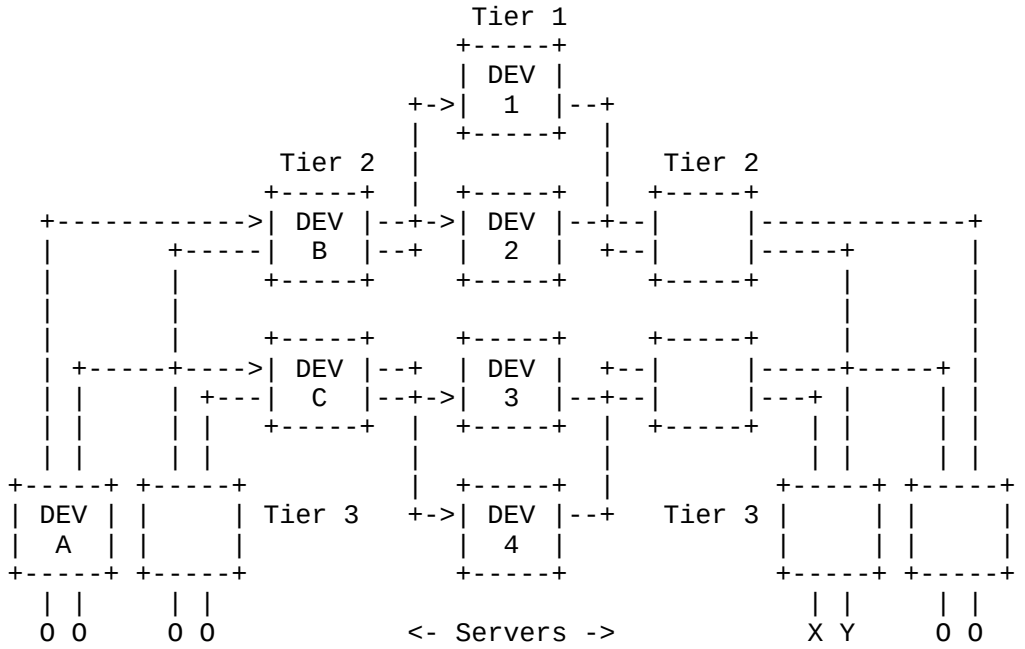

Figure 5: ECMP Fan-Out Tree from A to X and Y

The ECMP requirement implies that the BGP implementation must support multipath fan-out for up to the maximum number of devices directly attached at any point in the topology in the 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 fan-out 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 a 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 coupled with Layer 2 ECMP) to compensate for fan-out limitations. However, this approach increases the risk of flow polarization, as less entropy will be available at the second stage of ECMP.

Most BGP implementations declare paths to be equal from an ECMP perspective if they match up to and including step (e) in 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, such as the MULTI_EXIT_DISC (MED) attribute and origin code. For historical reasons, it is also useful to not use 0 as the equalized MED value; this and some other useful BGP information is available in [RFC4277]. 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 ASN in the path) are 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 the above-mentioned paths. This feature is sometimes known as "multipath relax" or "multipath multiple-AS" and effectively allows for ECMP to be done across different neighboring ASNs if all other attributes are equal as already 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 multi-hop 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 [LINK].

6.4. Consistent Hashing

It is often desirable to have the hashing function used for ECMP to be consistent (see [CONS-HASH]), to minimize the impact on flow to next-hop affinity changes when a next hop is added or removed to an ECMP group. This could be used if the network device is used as a load balancer, mapping flows toward multiple destinations -- in this case, losing or adding a destination will not have a detrimental effect on currently established flows. One particular recommendation on implementing consistent hashing is provided in [RFC2992], though other implementations are possible. This functionality could be naturally combined with weighted ECMP, with the impact of the next hop changes being proportional to the weight of the given next hop. The downside of consistent hashing is increased load on hardware resource utilization, as typically more resources (e.g., Ternary Content-Addressable Memory (TCAM) space) are required to implement a consistent-hashing function.

7. Routing Convergence Properties

This section reviews routing convergence properties in the proposed design. A case is made that sub-second convergence is achievable if the implementation supports fast EBGP peering session deactivation and timely RIB and FIB updates 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 remaining mechanisms that could be used for fault detection are BGP keep-alive time-out (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, on 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 "fast fallover". Since links in modern data centers are predominantly point-to-point fiber connections, a physical interface failure is often detected in milliseconds and subsequently triggers a BGP reconvergence.

Ethernet links may support failure signaling or detection standards such as Connectivity Fault Management (CFM) as described in [IEEE8021Q]; this 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, the use of either of these presents additional requirements to vendor software and possibly hardware, and may contradict REQ1. Until recently with [RFC7130], BFD also did not allow detection of a single member link failure on a LAG, which would have limited its usefulness in some designs.

7.2. Event Propagation Timing

In the proposed design, the impact of the BGP MinRouteAdvertisementIntervalTimer (MRAI timer), as specified in Section 9.2.1.1 of [RFC4271], should be considered. Per the standard, it is required for BGP implementations to space out consecutive BGP UPDATE messages by at least MRAI seconds, which is often a configurable value. The 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 when a BGP speaker "waits" for the new path to be learned from its peers and has no local backup path information.

In a Clos topology, each EBGP speaker typically has either one path (Tier 2 devices don’t accept paths from other Tier 2 in the same cluster due to same ASN) or N paths for the same prefix, where N is a significantly large number, e.g., N=32 (the ECMP fan-out to the next tier). Therefore, if a link fails to another device from which a path is received there is either no backup path at all (e.g., from the perspective of a Tier 2 switch losing the link to a Tier 3 device), or the backup is readily available in BGP Loc-RIB (e.g., from the perspective of a Tier 2 device losing the link to a Tier 1 switch). In the former case, the BGP withdrawal announcement will propagate without delay and trigger reconvergence on affected devices. In the latter case, the best path will be re-evaluated, and the local ECMP group corresponding to the new next-hop set will be changed. If the BGP path was the best path selected previously, an "implicit withdraw" will be sent via a BGP UPDATE message as described as Option b in Section 3.1 of [RFC4271] due to the BGP AS_PATH attribute changing.

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 UPDATE messages to all upstream Tier 1 devices, withdrawing the affected prefixes. The Tier 1 devices, in turn, will relay these messages to all downstream Tier 2 devices (except for the originator). Tier 2 devices other than the one originating the UPDATE should then wait for ALL upstream Tier 1 devices to send an UPDATE message before removing the affected prefixes and sending corresponding UPDATE downstream to connected Tier 3 devices. If the original Tier 2 device or the relaying Tier 1 devices introduce some delay into their UPDATE message announcements, the result could be UPDATE message "dispersion", that could be as long as multiple seconds. In order to avoid such a behavior, BGP implementations must support "update groups". The "update group" is defined as a collection of neighbors sharing the same outbound policy -- the local speaker will send BGP updates to the members of the group synchronously.

The impact of such "dispersion" grows with the size of topology fan-out and could also grow under network convergence churn. Some operators may be tempted to introduce "route flap dampening" type features that vendors include to reduce the control-plane impact of rapidly flapping prefixes. However, due to issues described with false positives in these implementations especially under such "dispersion" events, it is not recommended to enable this feature in this design. More background and issues with "route flap dampening" and possible implementation changes that could affect this are well described in [RFC7196].

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 recalculated their RIBs and consequently updated their FIBs. Larger failure impact scope typically means slower convergence since more devices have to be notified, and results in a less stable network. In this section, we describe BGP’s advantages over link-state routing protocols in reducing failure impact scope for a Clos topology.

BGP behaves like a distance-vector protocol in the sense that only the best path from the point of view of the local router is sent to neighbors. As such, some failures are masked if the local node can immediately find a backup path and does not have to send any updates further. Notice that in the worst case, all devices in a data center

topology have to either withdraw a prefix completely or update the ECMP groups in their FIBs. 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 the affected ECMP groups, removing the failed link. There is no need to send new information to downstream Tier 3 devices, unless the path was selected as best by the BGP process, in which case only an "implicit withdraw" needs to be sent and this should not affect forwarding. 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 a 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 reconvergence 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 a non-local cluster. The Tier 3 devices are once again not involved in the reconvergence process, but may receive "implicit withdraws" as described above.

Even in the case of such failures where 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 a Tier 2 device. Therefore, in the case of implementations with a hierarchical FIB, only a single change has to be made to the FIB. "Hierarchical FIB" here means FIB structure where the next-hop forwarding information is stored separately from the prefix lookup table, and the latter only stores pointers to the respective forwarding information. See [BGP-PIC] for discussion of FIB hierarchies and fast convergence.

Even though BGP offers reduced failure scope for some cases, further 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 failure impact scope on the control plane is the network as a whole -- for instance, in the 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 Clos topology.

7.5. Routing Micro-Loops

When a downstream device, e.g., Tier 2 device, loses all paths 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 where a Tier 2 switch loses a prefix, but a Tier 1 switch still has the path pointing to the Tier 2 device; this results in a transient micro-loop, since the Tier 1 switch will keep passing packets to the affected prefix back to the Tier 2 device, and the Tier 2 will bounce them back again using the default route. This micro-loop will last for the time it takes the upstream device to fully update its forwarding tables.

To minimize impact of such micro-loops, Tier 2 and Tier 1 switches can be configured with static "discard" or "null" routes that will be more specific than the default route for prefixes missing during network convergence. For Tier 2 switches, the discard route should be a summary route, covering all server subnets of the underlying Tier 3 devices. For Tier 1 devices, the discard route should be a summary covering the server IP address subnets 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", i.e., a directly attached BGP speaker, to inject routes anywhere in the network topology, meeting REQ5. This can be achieved by peering via a multi-hop 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 [RFC7911] if supported by the implementation. Unfortunately, in many implementations, ADD-PATH has been found to only support IBGP properly in the use cases for which it was originally optimized; this limits the "third-party" peering to IBGP only.

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 Summarization within Clos Topology

As mentioned previously, route summarization 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 redundant paths between network elements, e.g., 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.

If any technique to summarize within the topology is planned, modeling of the routing behavior and potential for black-holing should be done not only for single or multiple link failures, but also for fiber pathway failures or optical domain failures when the topology extends beyond a physical location. Simple modeling can be done by checking the reachability on devices doing summarization under the condition of a link or pathway failure between a set of devices in every tier as well as to the WAN routers when external connectivity is present.

Route summarization would be possible with a small modification to the network topology, though the tradeoff 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.

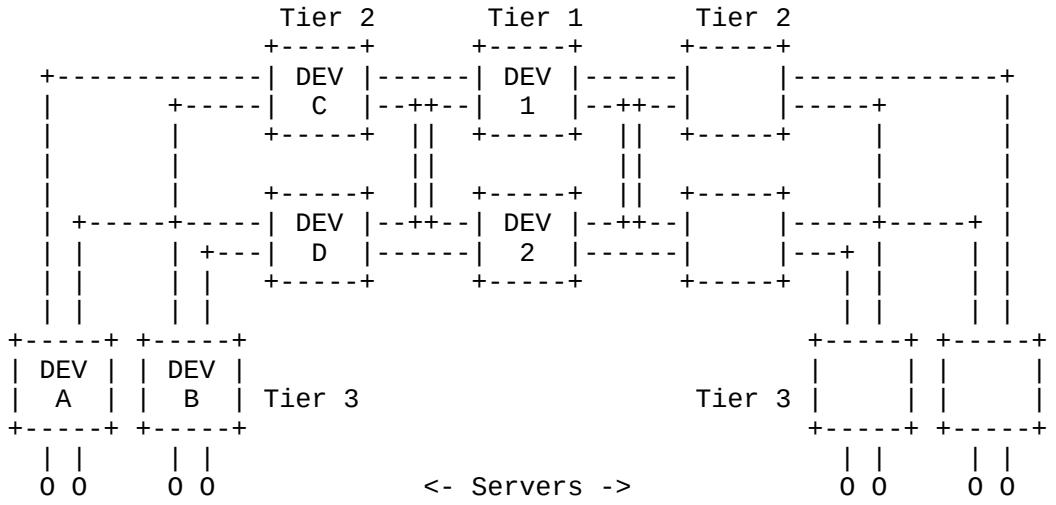

Figure 6: 5-Stage Clos Topology

Having this design in place, 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 still 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 failures. 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.

A result of the proposed topology modification would be a reduction of the port capacity of Tier 1 devices. 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 rebalancing under link failures. Since there 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 the link utilization on the remaining link.

8.2.2. Simple Virtual Aggregation

A completely different approach to route summarization is possible, provided that the main goal is to reduce the FIB size, while allowing the control plane to disseminate full routing information. Firstly, it could be easily noted that in many cases multiple prefixes, some of which are less specific, share the same set of the next hops (same ECMP group). For example, from the perspective of Tier 3 devices, all routes learned from upstream Tier 2 devices, including the default route, will share the same set of BGP next hops, provided that there are no failures in the network. This makes it possible to use the technique similar to that described in [RFC6769] and only install the least specific route in the FIB, ignoring more specific routes if they share the same next-hop set. For example, under normal network conditions, only the default route needs to be programmed into the FIB.

Furthermore, if the Tier 2 devices are configured with summary prefixes covering all of their attached Tier 3 device's prefixes, the same logic could be applied in Tier 1 devices as well and, by induction to Tier 2/Tier 3 switches in different clusters. These summary routes should still allow for more specific prefixes to leak to Tier 1 devices, to enable detection of mismatches in the next-hop sets if a particular link fails, thus changing the next-hop set for a specific prefix.

Restating once again, this technique does not reduce the amount of control-plane state (i.e., BGP UPDATEs, BGP Loc-RIB size), but only allows for more efficient FIB utilization, by detecting more specific prefixes that share their next-hop set with a subsuming less specific prefix.

8.3. ICMP Unreachable Message Masquerading

This section discusses some operational aspects of not advertising point-to-point link subnets into BGP, as previously identified as an option in Section 5.2.3. The operational impact of this decision could be seen when using the well-known "traceroute" tool. Specifically, IP addresses displayed by the tool will be the link's point-to-point addresses, and hence will be unreachable for management connectivity. This makes some troubleshooting more complicated.

One way to overcome this limitation is by using the DNS subsystem to create the "reverse" entries for these point-to-point IP addresses pointing to the same name as the loopback address. The connectivity then can be made by resolving this name to the "primary" IP address

of the device, e.g., its Loopback interface, which is always advertised into BGP. However, this creates a dependency on the DNS subsystem, which may be unavailable during an outage.

Another option is to make the network device perform IP address masquerading, that is, rewriting the source IP addresses of the appropriate ICMP messages sent by the device with the "primary" IP address of the device. Specifically, the ICMP Destination Unreachable Message (type 3) code 3 (port unreachable) and ICMP Time Exceeded (type 11) code 0 are required for correct operation of the "traceroute" tool. With this modification, the "traceroute" probes sent to the devices will always be sent back with the "primary" IP address as the source, allowing the operator to discover the "reachable" IP address of the box. This has the downside of hiding the address of the "entry point" into the device. If the devices support [RFC5837], this may allow the best of both worlds by providing the information about the incoming interface even if the return address is the "primary" IP address.

9. Security Considerations

The design does not introduce any additional security concerns. General BGP security considerations are discussed in [RFC4271] and [RFC4272]. Since a DC is a single-operator domain, this document assumes that edge filtering is in place to prevent attacks against the BGP sessions themselves from outside the perimeter of the DC. This may be a more feasible option for most deployments than having to deal with key management for TCP MD5 as described in [RFC2385] or dealing with the lack of implementations of the TCP Authentication Option [RFC5925] available at the time of publication of this document. The Generalized TTL Security Mechanism [RFC5082] could also be used to further reduce the risk of BGP session spoofing.

10. References

10.1. Normative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC6996] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", BCP 6, RFC 6996, DOI 10.17487/RFC6996, July 2013, <http://www.rfc-editor.org/info/rfc6996>.

10.2. Informative References

[ALFARES2008]
Al-Fares, M., Loukissas, A., and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture", DOI 10.1145/1402958.1402967, August 2008, <http://dl.acm.org/citation.cfm?id=1402967>.

[ALLOWASIN]
Cisco Systems, "Allowas-in Feature in BGP Configuration Example", February 2015, <http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/112236-allowas-in-bgp-config-example.html>.

[BGP-PIC] Bashandy, A., Ed., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", Work in Progress, draft-ietf-rtgwg-bgp-pic-02, August 2016.

[CLOS1953] Clos, C., "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal, Vol. 32(2), DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953.

[CONDITIONALROUTE]
Cisco Systems, "Configuring and Verifying the BGP Conditional Advertisement Feature", August 2005, <http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/16137-cond-adv.html>.

[CONS-HASH]
Wikipedia, "Consistent Hashing", July 2016, <https://en.wikipedia.org/w/index.php?title=Consistent_hashing&oldid=728825684>.

[FB4POST] Farrington, N. and A. Andreyev, "Facebook’s Data Center Network Architecture", May 2013, <http://nathanfarrington.com/papers/facebook-oic13.pdf>.

[GREENBERG2009]
Greenberg, A., Hamilton, J., and D. Maltz, "The Cost of a Cloud: Research Problems in Data Center Networks", DOI 10.1145/1496091.1496103, January 2009, <http://dl.acm.org/citation.cfm?id=1496103>.

[HADOOP] Apache, "Apache Hadoop", April 2016, <https://hadoop.apache.org/>.

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

[IEEE8021D-1990] IEEE, "IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges", IEEE Std 802.1D, DOI 10.1109/IEEESTD.1991.101050, 1991, <http://ieeexplore.ieee.org/servlet/opac?punumber=2255>.

[IEEE8021D-2004] IEEE, "IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges", IEEE Std 802.1D, DOI 10.1109/IEEESTD.2004.94569, June 2004, <http://ieeexplore.ieee.org/servlet/opac?punumber=9155>.

[IEEE8021Q] IEEE, "IEEE Standard for Local and Metropolitan Area Networks: Bridges and Bridged Networks", IEEE Std 802.1Q, DOI 10.1109/IEEESTD.2014.6991462, <http://ieeexplore.ieee.org/servlet/opac?punumber=6991460>.

[IEEE8023AD] IEEE, "Amendment to Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications - Aggregation of Multiple Link Segments", IEEE Std 802.3ad, DOI 10.1109/IEEESTD.2000.91610, October 2000, <http://ieeexplore.ieee.org/servlet/opac?punumber=6867>.

[INTERCON] Dally, W. and B. Towles, "Principles and Practices of Interconnection Networks", ISBN 978-0122007514, January 2004, <http://dl.acm.org/citation.cfm?id=995703>.

[JAKMA2008] Jakma, P., "BGP Path Hunting", 2008, <https://blogs.oracle.com/paulj/entry/bgp_path_hunting>.

[L3DSR] Schaumann, J., "L3DSR - Overcoming Layer 2 Limitations of Direct Server Return Load Balancing", 2011, <https://www.nanog.org/meetings/nanog51/presentations/Monday/NANOG51.Talk45.nanog51-Schaumann.pdf>.

[LINK] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, draft-ietf-idr-link-bandwidth-06, January 2013.

[REMOVAL] Mitchell, J., Rao, D., and R. Raszuk, "Private Autonomous System (AS) Removal Requirements", Work in Progress, draft-mitchell-grow-remove-private-as-04, April 2015.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <http://www.rfc-editor.org/info/rfc2328>.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, DOI 10.17487/RFC2385, August 1998, <http://www.rfc-editor.org/info/rfc2385>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <http://www.rfc-editor.org/info/rfc4272>.

[RFC4277] McPherson, D. and K. Patel, "Experience with the BGP-4 Protocol", RFC 4277, DOI 10.17487/RFC4277, January 2006, <http://www.rfc-editor.org/info/rfc4277>.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <http://www.rfc-editor.org/info/rfc4786>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <http://www.rfc-editor.org/info/rfc5082>.

[RFC5837] Atlas, A., Ed., Bonica, R., Ed., Pignataro, C., Ed., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, DOI 10.17487/RFC5837, April 2010, <http://www.rfc-editor.org/info/rfc5837>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <http://www.rfc-editor.org/info/rfc5925>.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <http://www.rfc-editor.org/info/rfc6325>.

[RFC6769] Raszuk, R., Heitz, J., Lo, A., Zhang, L., and X. Xu, "Simple Virtual Aggregation (S-VA)", RFC 6769, DOI 10.17487/RFC6769, October 2012, <http://www.rfc-editor.org/info/rfc6769>.

[RFC6774] Raszuk, R., Ed., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, DOI 10.17487/RFC6774, November 2012, <http://www.rfc-editor.org/info/rfc6774>.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <http://www.rfc-editor.org/info/rfc6793>.

[RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, DOI 10.17487/RFC7067, November 2013, <http://www.rfc-editor.org/info/rfc7067>.

[RFC7130] Bhatia, M., Ed., Chen, M., Ed., Boutros, S., Ed., Binderberger, M., Ed., and J. Haas, Ed., "Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces", RFC 7130, DOI 10.17487/RFC7130, February 2014, <http://www.rfc-editor.org/info/rfc7130>.

[RFC7196] Pelsser, C., Bush, R., Patel, K., Mohapatra, P., and O. Maennel, "Making Route Flap Damping Usable", RFC 7196, DOI 10.17487/RFC7196, May 2014, <http://www.rfc-editor.org/info/rfc7196>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <http://www.rfc-editor.org/info/rfc7911>.

[VENDOR-REMOVE-PRIVATE-AS] Cisco Systems, "Removing Private Autonomous System Numbers in BGP", August 2005, <http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080093f27.shtml>.

Acknowledgements

This publication summarizes the 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. The authors would also like to thank Linda Dunbar, Anoop Ghanwani, Susan Hares, Danny McPherson, Robert Raszuk, and Russ White for reviewing this document and providing valuable feedback, and Mary Mitchell for initial grammar and style suggestions.

Authors' Addresses

Petr Lapukhov
Facebook
1 Hacker Way
Menlo Park, CA 94025
United States of America

Email: petr@fb.com

Ariff Premji
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
United States of America

Email: ariff@arista.com
URI: http://arista.com/

Jon Mitchell (editor)

Email: jrmitche@puck.nether.net