IDR
Internet-Draft
Intended status: Informational
Expires: January 8, 2013

P. Lapukhov
Microsoft Corp.
A. Premji
Arista Networks
July 7, 2012

Using BGP for routing in large-scale data centers
draft-lapukhov-bgp-routing-large-dc-00

Abstract

Some service providers build and operate data centers at the size exceeding 100,000 servers. In this document, those data-centers are referred to as "large-scale" to differentiate them from more common smaller infrastructures. The data centers of that scale have unique set of network design requirement, with primary focus on operational simplicity and stability.

This document attempts to summarize the authors' experiences in designing and supporting large data centers, using BGP as the only control-plane protocol. The intent is to describe 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 8, 2013.

Table of Contents

1. Introduction

This document presents a practical routing design to be used in large-scale data centers, sometimes called hyperscale or warehousr-scale. The most distinctive characterstic of these data center is having 100,000 or more end hosts connected to the network. While historically only a few companies have been operating networks of that scale, recent trend in building large cloud data centers re-ignated interest in network designs to support deployment of this scale. In contrary to more traditional data center designs, the approach proposed in this document does not depend on large Layer 2 domains and instead uses routing at every level of the network. The reason to make that choice is based on the unique set of design requirements, with primary focus on cost reduction. Furthermore, analyzing the requirements the conclusion is that BGP best suits to accomplish this goal due primarily to its simplicity and broad vendor support.

2. Traditional data center designs

This section provides an overview of two types of traditional data center designs - Layer-2 and fully routed Layer-3 topologies.

2.1. Layer 2 Designs

In the networking industry, common design choice for data centers is using a mix of Ethernet-based Layer 2 technologies. Network topology typically looks like a tree with redundant uplinks and three levels of hierarchy (see Figure 1) commonly named Core, Aggregation and Access. To accommodate bandwidth demands, every next level has higher port density and bandwidth capacity. In this document, the topology layers will be referenced as "tiers", e.g. Tier 1, Tier 2 and Tier 3 instead of Core, Aggregation or Access layers.

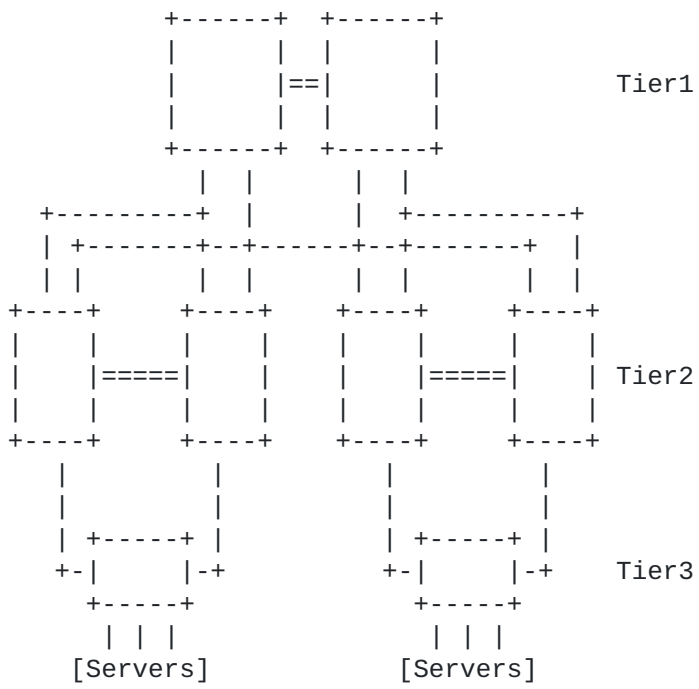

Figure 1: Typical Data Center network layout

IP routing is normally used only at the upper layers of the topology, e.g. Tier 1 or Tier 2. The main reasons for introducing such large (sometimes called stretched) Level 2 domains, are the following:

- Supporting legacy applications that may require direct Layer 2 adjacency or use non-IP protocols
- Seamless mobility for virtual machines, to allow preserving IP address when a virtual machine changes physical host
- Simplified IP addressing - less IP subnets is required for the data-center
- Application load-balancing may require direct Layer 2 adjacency to perform some functions such as Level 2 Direct Server Return (DSR)

2.2. Fully routed network designs

Network designs that leverage IP routing down to the access layer (Tier 3) of the network gained some popularity, mostly due to improved network stability, scalability (by means of information hiding) and convergence times. A common choice of routing protocol for data center designs would be an IGP, such as OSPF or ISIS. As data centers grow in scale, and server count exceeds tens of thousands, those fully routed designs become more attractive.

BGP is the de-facto standard protocol for routing on the Internet, having wide support from network equipment vendors and being well-understood by network engineers world-wide. However, it is not common to see BGP being used in data centers that employ fully routed network design. There multiple reasons for that:

- BGP is perceived as "WAN protocol only" and often not being considered for enterprise or data center application
- BGP is believed to converge "slower" than traditional IGPs
- BGP is assumed to have a dependency on the presence of an IGP, which assists with recursive next-hop resolution
- BGP require a lot of configuration efforts as it does not support any form of neighbor auto-discovery

In this document we argue benefits of choosing BGP as the single routing protocol, including acceptable convergence time.

3. Document structure

The remaining of this document is organized as following. First the design requirements for large scale data centers are presented. Next, the document gives an overview of Clos network topology and its properties. After that, the arguments for selecting BGP as the single routing protocols are presented. Finally, the document goes over design detail and specific BGP policy features.

4. Network design requirements

This section describes and summarizes network design requirement for a large-scale data center.

4.1. Traffic patterns

The primary requirement when building an interconnection network for large number of servers is accommodating application bandwidth and latency requirements. For long period of time, it was common to see traffic flowing mainly to and from the data center. There were no intense (highly meshed flows) traffic patterns between the machines within the same tier. As a result, traditional "tree" topology was sufficient to accommodate data flow, even with high oversubscription ratios in network equipment. If more bandwidth was required, it was added by "scaling up" the network elements, e.g. by adding more line-cards or replacing existing devices with higher capacity switches.

In contrast, large-scale data centers often host applications that generate large amount of server to server traffic, also known as

"east-west" traffic. Examples of such applications could be compute clusters such as Hadoop or live virtual machine migration in "cloud" data-centers. Scaling up traditional tree topology to match those bandwidth demands becomes either too expensive or impossible due to physical limitation.

4.2. CAPEX minimization

Cost of networking component alone (CAPEX) constitutes about 10-15% of total data center cost [GREENBERG2009]. Still, absolute numbers are significant, and hence the need to constantly drive cost of networking elements down. This is normally accomplished in two ways:

- Unifying network elements, preferably using the same hardware type or even the same device. This allows for bulk purchases with discounted pricing.
- Driving costs down by introducing diversity of networking vendors that may supply equipment for data center network

In order to allow for vendor diversity, it is important to minimize the feature requirements for network equipment software. In addition, the above strategy means that network equipment vendor choice may change often, or that the network may have to be multi-vendor and interoperability becomes critical.

4.3. OPEX minimization

Operating large scale infrastructure could be expensive, provide that larger amount of elements will statistically fail more often. Therefore, it is important to operate on the simplest software and feature set possible.

An important aspect of OPEX minimization is reducing size of failure domains in the network. Ethernet data-plane is known to be susceptible to massive impact due to broadcast or unicast storms. The use of fully routed designs reduces the size of data-plane failure domains, but at the time introduces the problem of distributed control-plane failures. This requirement calls for simpler control-plane protocols that are expected to have less chances of network meltdown.

4.4. Traffic Engineering

In any data center, application load-balancing is critical function performed by network devices. Traditionally, load-balancers are deployed as dedicated devices in traffic forwarding path. A common problem is scaling load-balancers under growing traffic demand. Preferable solution would be able to scale load-balancing layer

horizontally, by adding more of the uniform nodes and distributing incoming traffic across them.

In situation like this, an ideal choice would be using network infrastructure to distribute traffic across a group of load-balancers. A combination of features such Anycast prefix advertisement [RFC4786] along with Equal Cost Multipath (ECMP) functionality could be used to accomplish this. To allow for more granular load-distribution, it is beneficial for the network to support the ability to perform controlled per-hop traffic engineering.

5. Requirement List

This section summarizes the requirements in a list, based on the analysis made before

- REQ1: Select a network topology where capacity could be scaled "horizontally" by adding more links and network switches of the same type, without requiring upgrading the network elements themselves.
- REQ2: Define a narrow set of software features/protocols supported by multitude of networking equipment vendors.
- REQ3: Among the network protocols, select those having simpler implementation in terms of minimal programming code complexity.
- REQ4: The selected network routing protocol should support per-hop change of forwarding behavior.

6. Network topology

This section outlines the most common choice for horizontally scalable topology in large scale data centers.

6.1. Clos topology overview

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

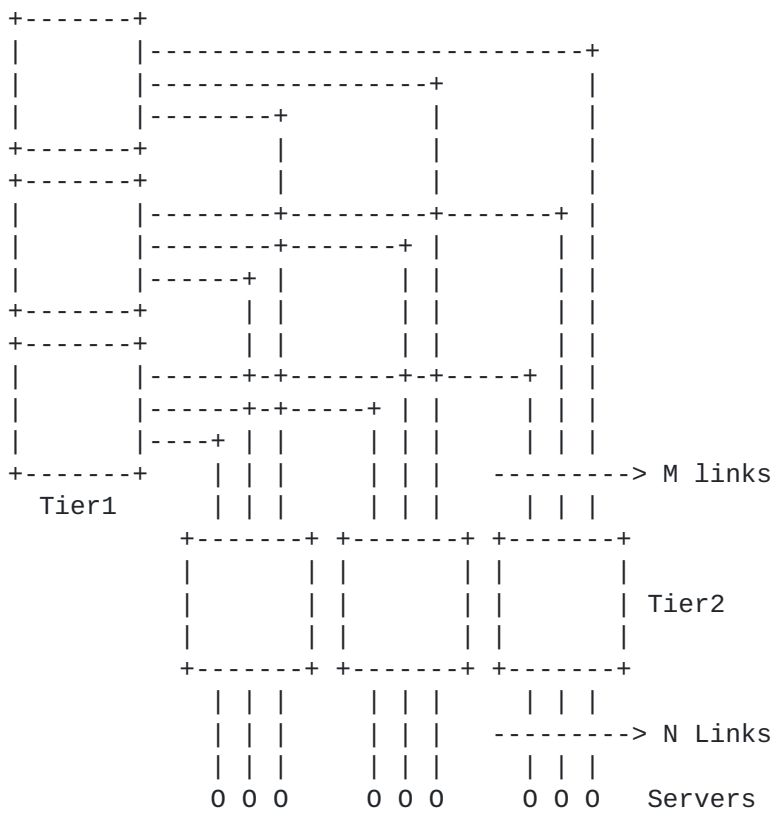

Figure 2: 3-Stage Folded Clos topology

In the networking industry, a topology like this is sometimes referred to as "Leaf and Spine", where Spine is the name for the middle stage of the Clos topology (Tier 1) and Leaf is the name of input/output stage (Tier 2). However, for consistency, the document will be using "Tier n" notation.

6.2. Clos topology properties

The following are some key properties of the Clos topology:

- 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 Tier 2 switch, as shown on Figure 2
- Implementing Clos topology requires a routing protocol supporting ECMP with the fan-out of M or more
- Every Tier 1 device has exactly one path to every end host (server) in this topology

o Traffic flowing from server to server is naturally load-balanced over all available paths using simple ECMP behavior

6.3. Scaling Clos topology

Clos topology could be scaled either by increasing network switch radix or adding more stages, e.g. moving to a 5-stage Clos, as illustrated on Figure 3 below:

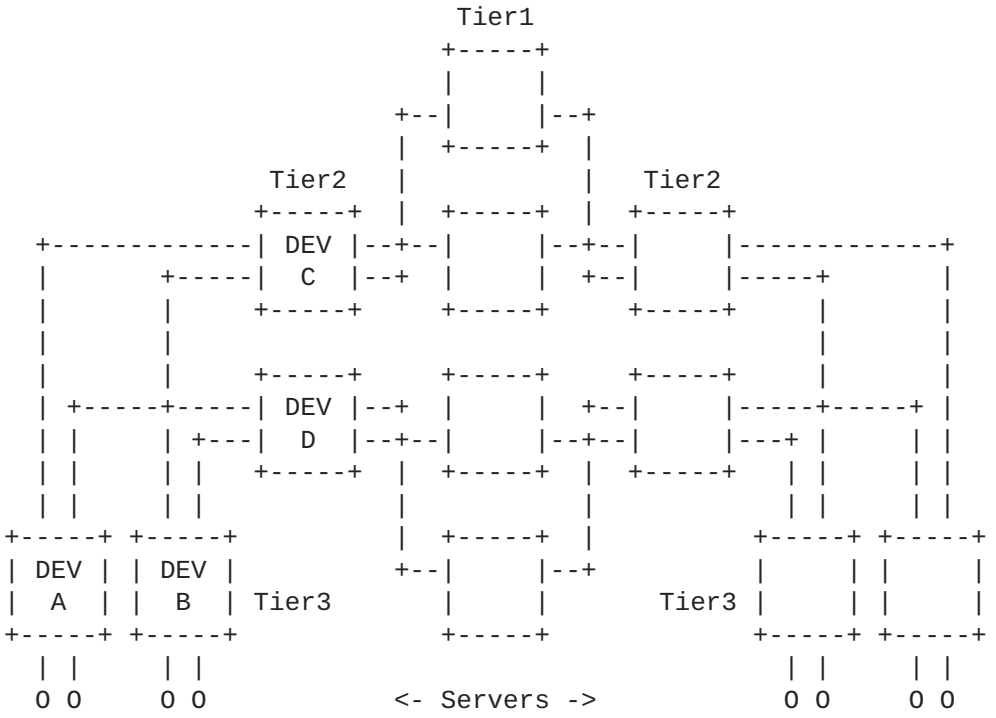

Figure 3: 5-Stage Clos topology

The topology on Figure 3 is built from switches with radix 4 and provides full bisection bandwidth to all connected servers. We'll be referring to the collection of directly connected Tier 2 and Tier 3 switches as "cluster" in this document. For example, devices A, B, C, and D on Figure 3 form a cluster.

In practice, Tier 3 level of the network (typically top of rack switches, or ToRs) often introduces oversubscription to allow for packaging more servers in data center. The main reason to oversubscribe only at a single layer of the network is to simplify application development that would need to account for two bandwidth pools: within the same access switch (e.g. rack) and outside of the local switch. Oversubscription, however, does not affect routing

design and hence not considered in more details in this document.

7. Routing design

This section discusses the motivation for choosing BGP as the routing protocol and BGP configuration for routing in Clos topology.

7.1. Choosing the routing protocol

The set of requirement provide above calls for a single routing protocol (REQ2) in the data center to reduce complexity and interdependencies. While it's common to rely on an IGP in this situation, the document proposes to use BGP only. The advantages of using BGP are argued below.

- BGP has less complexity in protocol design - internal data structures and state-machines are simpler when compared to a link-state IGP. For example, as opposed to implementing adjacency formation and maintenance, flow-control, etc. BGP simply relies on TCP as the underlying transport. This also simplified protocol testing and fulfills REQ1 and REQ2.
- BGP information flooding overhead is less when compared to link-state IGPs. Indeed, since every BGP router normally re-calculates and propagates best-paths only, a network failure is masked as soon as BGP speaker finds an alternate path. On contrary, event propagation scope of a link-state IGP is single area/domain, regardless of the failure type. Furthermore, all well-known link-state IGPs feature periodic refresh updates, while BGP does not expire routing state.
- BGP supports third-party (recursively resolved) next-hops, which allows for injecting custom routing paths into any device in the network, using eBGP multi-hop peering session. This satisfied REQ4 stated above. Some IGPs, such as OSPF, support similar functionality using special concepts such as "Forwarding Address", but do not satisfy other requirement, such as protocol simplicity.
- BGP is easier to troubleshoot, mostly because of simplified protocol mechanics and database structures that directly map to forwarding tables structure. For example, it is straightforward to dump contents of LocRIB and compare it to the router's RIB and FIB. As another example, BGP routing updates translate directly into NLRI information, as compared to LSA/LSP information that describes network topology. Thus BGP fully satisfies REQ3.

7.2. BGP configuration for Clos topology

This section provides configuration guidelines for a 5-stage Clos topology. It is easy to reduce it to a 3-Stage Clos configuration,

and having topology that has more than 5 stages is very uncommon due to high link density of associated designs.

7.2.1. BGP Autonomous System numbering layout

The diagram below illustrates suggests BGP Autonomous System Number (BGP ASN) allocation scheme. The following is a list of guiding principles:

- All BGP peering sessions are external BGP (eBGP) established over direct point-to-point links interconnecting the network switches.
- 16-bit (two octet) BGP ASNs are used, for the reason of wider vendor support and better vendor interoperability (e.g. no need to support BGP capability negotiation).
- Private BGP ASNs from the range 64512-64534 are used for the reasons of avoiding ASN conflicts and being able to use BGP private ASN stripping feature (see below).
- A single BGP ASN is allocated to the Clos middle stage ("Tier 1"), e.g. ASN 64534 on Figure 4
- Unique BGP ASN is allocated per every group of "Tier 2" switches. All Tier 2 switches in the same group share the BGP ASN.
- Unique BGP ASN is allocated to every Tier 3 switch (e.g. ToR) in this topology.

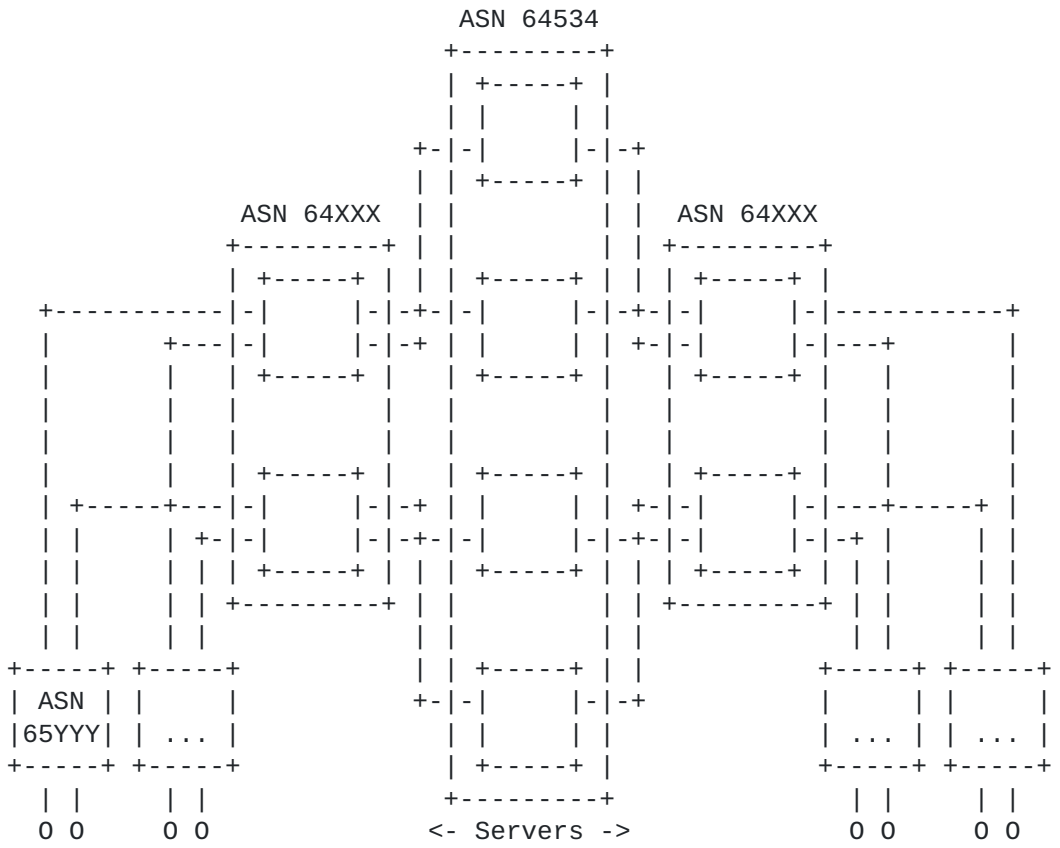

Figure 4: BGP ASN layout for 5-stage Clos

7.2.2. Non-unique private BGP ASN's

The use of private BGP ASNs limits to a range of 1022 unique numbers. It is possible that the number of network switches could exceed this value, and such situation requires a workaround. One approach could be re-using the private ASN's assigned to Tier 3 switches across different clusters. For example, private BGP ASN's 65001, 65003 ... 65032 could be used within every individual cluster to be assigned to Tier 3 switches.

To avoid Tier 3 route discards on the Tier 3 switches sharing the same ASN due to AS PATH loop prevention, upstream eBGP sessions on Tier 3 switches must be configured with so-called "AllowAS In" feature. This BGP policy feature allows accepting device's own ASN in incoming BGP path advertisements. Introduction of this feature does not create the opportunity for permanent routing loops under misconfiguration since AS PATH is always increments when routes are propagated from tier to tier.

Another solution to this problem would be switching over to using four-octet (32-bit) BGP ASNs. However, there is no explicitly reserved private ASN range in four-octet numbering, but a work is in progress to request such an allocation in [I-D.mitchell-idr-as-private-reservation]. This will also require vendors to implement specific policy features, such as private AS removal from AS-PATH attribute.

7.2.3. Prefix advertisement

Clos topology has large number of point-to-point links and associated prefixes. Advertising all of them into BGP may create FIB sizing issues, and there are two possible solutions to overcome this:

- Do not advertise any of the point-to-point links into BGP. Since eBGP peering changes next-hop address at every node, this will not create any reachability issues for subnets advertised from Tier 3 switches.
- Advertising point-to-point links, but summarizing them on every advertising device. This requires proper address allocation, for example allocating a consecutive block of IP addresses per Tier 1 and Tier 2 device to be used for point-to-point interface numbering.

Server facing subnets on Tier 3 switches are announced into BGP without using summarization on Tier 2 and Tier 1 switches. Summarizing subnets in the Clos topology will result in route black-holing under a single link failure (e.g. between Tier 2 and Tier 3 switch) and hence must be avoided. The use of peer links within the same tier to resolve the black-holing problem is undesirable due to O(N^2) complexity of the peering mesh and waste of ports on the switches.

7.2.4. External connectivity

A dedicate cluster (or clusters) in Clos topology could be selected solely for the purpose of connecting to the Wide Area Network (WAN) edge devices, which we will call WAN Routers. Tier 3 switches in such cluster would be replaced with WAN Routers, but eBGP peering will be used as usual, though WAN routers are likely to belong to a public ASN.

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

- o Hide network topology information when advertising paths to WAN routers, i.e. remove some BGP AS-PATH information. This is typically done to avoid BGP ASN number collisions across the data centers. A BGP policy feature called "Remove Private AS" is commonly used to accomplish this. This feature strips 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 private range.
- o Originate a default route to the data center devices. This is the only place where default route could 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.

7.3. ECMP Considerations

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

7.3.1. Basic ECMP

ECMP is the key load-sharing mechanism leveraged by Clos topology. Effectively, every lower-tier switch will use all of its directly attached upper-tier devices to load-share traffic to the same prefix. Number of ECMP paths between two input/output switches in Clos topology equals to the number of the switches 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.

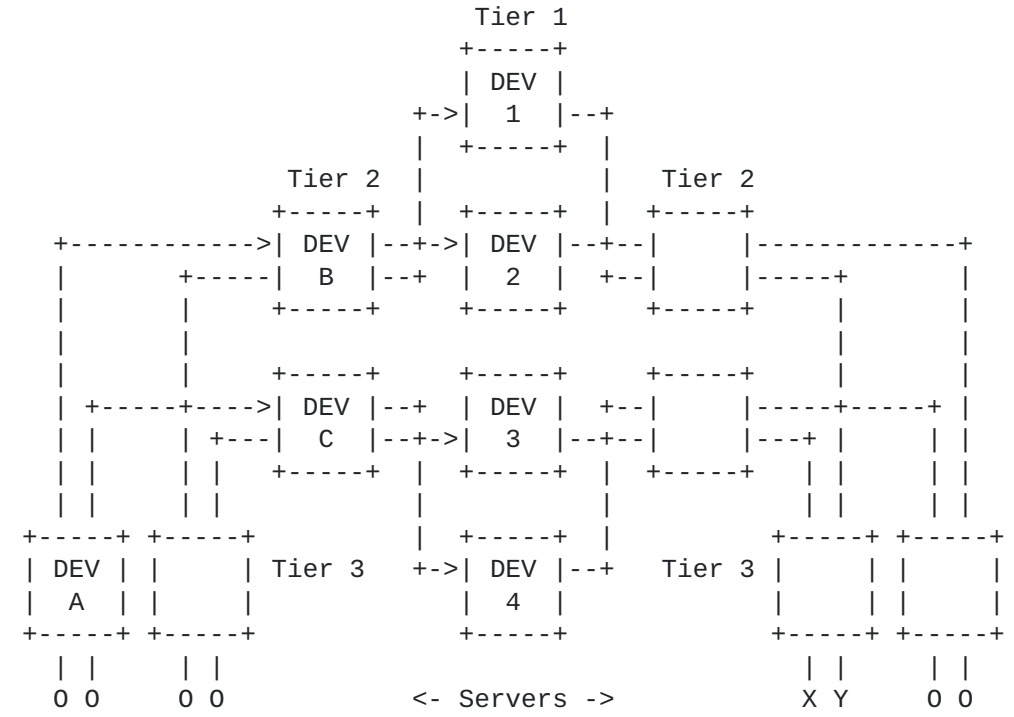

Figure 5: ECMP fan-out tree from A to X and Y

The ECMP requirement implies that BGP implementation must support multi-path fan-out for up to the maximum number of devices directly attached at any point in the topology. Normally, this number does not exceed half of the ports found on a switch in the topology, e.g. 32 for a 64-port switch.

Most implementations declare paths to be equal from 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 automatically assumed to be zero (or otherwise the same value across all paths). Loop prevention is assumed to be handled by BGP best-path selection process.

7.3.2. BGP ECMP over multiple ASN

For the purpose of application load-balancing purposes same prefix could be advertised from multiple Tier-3 switches. From the perspective of other devices, such prefix would have BGP paths with different AS PATH attribute values, though having the same AS PATH length. BGP implementation must support load-sharing for the paths having different AS PATH attribute values with equal attribute length. This feature is sometimes known as "AS PATH multipath relax"

and effectively allows for ECMP to be done across different neighboring ASNs.

7.4. BGP convergence properties

This section reviews routing convergence properties of BGP in the proposed design. A case is made that sub-second convergence is achievable provided that implementation supports fast BGP peering session shutdown upon failure of an associated link.

7.4.1. Convergence timing

BGP typically relies on IGP to route around link/node failures inside an AS, and implements either polling based or event-driven mechanism to obtain updates on IGP state changes. The proposed routing design lacks any IGP, so the only mechanism that could be used for fault detection is BGP keep-alive packet exchange.

Relying purely on BGP keep-alive packets may result in high convergence delays, on the order of multiple seconds (normally, the minimum recommended BGP hold time value is 3 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 majority of the links in modern data centers are point to point fiber connections, a physical failure translates into interface going down within order of milliseconds, and trigger BGP re-convergence. Furthermore, popular link technologies, such as 10Gbps Ethernet, may support simple form of OAM for failure signaling such as [FAULTSIG10GE], which makes failure detection more robust. Alternatively, as opposed to relying on physical layer for fault signaling, some platforms may support Bidirectional Forwarding Detection ([RFC5880]) to allow for sub-second failure detection and fault signaling to BGP process. This, however, presents additional requirements to vendor software and possibly hardware, and may contradict REQ1.

7.4.2. Failure impact scope

BGP is inherently a distance-vector protocol, and as such some of failures could be masked if the local node can immediately find a backup path. Worst case is that all devices would have to either withdraw a prefix completely, or update the ECMP paths in the FIB. That fault domain cannot be reduced by using summarization, since using this technique may create route black-holing issues as mentioned previously.

7.4.3. Third-party route injection

BGP allows for a third-party BGP speaker (not necessarily directly attached to the network devices) to inject routes at any point of network topology. This could be achieved by peering an external speaker using eBGP multi-hop session with some or even all devices in the topology. Furthermore, BGP diverse path distribution [I-D.ietf-grow-diverse-bgp-path-dist] could be used to inject multiple next-hop for the same prefix and facilitate load-balancing. Using that technique, it is possible to implement unequal-cost load-balancing across multiple clusters in the data-center, by associating the same prefix with next-hops mapping to different clusters.

For example, 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 switches. The resulting forwarding table programming could provide desired traffic proportion distribution among different clusters.

8. Security Considerations

The design does not introduce any special security concerns others than normally associated with BGP deployments. For control plane security, BGP peering sessions could be authenticated using TCP MD5 signature extension header [RFC2385]. Furthermore, BGP TTL security [I-D.gill-btsh] could be used to reduce the risk of session spoofing and TCP SYN flooding attacks against the control plane.

9. IANA Considerations

There are no considerations associated with IANA for this document.

10. Acknowledgements

This publication summarizes work of many people who participated in developing, testing and deploying the proposed design. Their names, in alphabetical order, are George Chen, Parantap Lahiri, Dave Maltz, Edet Nkposong, Robert Toomey, and Lihua Yuan. Authors would also like to thank Jon Mitchell for reviewing and providing valuable feedback on the document.

11. Informative References

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

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

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

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

[I-D.ietf-grow-diverse-bgp-path-dist] Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.", draft-ietf-grow-diverse-bgp-path-dist-07 (work in progress), May 2012.

[I-D.mitchell-idr-as-private-reservation] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", draft-mitchell-idr-as-private-reservation-00 (work in progress), June 2012.

[I-D.gill-btsh] Gill, V., Heasley, J., and D. Meyer, "The BGP TTL Security Hack (BTSH)", draft-gill-btsh-02 (work in progress), May 2003.

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

[FAULTSIG10GE] Frazier, H. and S. Muller, "Remote Fault & Break Link Proposal for 10-Gigabit Ethernet", September 2000.

Authors' Addresses

Petr Lapukhov
Microsoft Corp.
One Microsfot Way
Redmond, WA 98052
US

Phone: +1 425 7032723 X 32723
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/