IDR
Internet-Draft
Intended status: Informational
Expires: October 10, 2013

P.L. Lapukhov
Microsoft Corp.
A.P. Premji
Arista Networks

April 08, 2013

Using BGP for routing in large-scale data centers
draft-lapukhov-bgp-routing-large-dc-04

Abstract

Some service providers build and operate data centers that support over 100,000 servers. In this document, such data centers are referred to as "large-scale" to differentiate them from smaller infrastructures. The environments of this scale have a unique set of network requirements, with emphasis on operational simplicity and network stability.

This document summarizes ideas and experience of many people involved in designing and operating large scale data centers using BGP as the only control-plane protocol. The intent here is to report 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 October 10, 2013.

Table of Contents

1. Introduction

This document describes a practical routing design that can be used in large-scale data centers. Such data centers, also known as hyper-scale or warehouse-scale data centers, have a unique attribute of supporting over a 100,000 end hosts. In order to accommodate networks of such scale, operators are revisiting networking designs and platforms to address this need.

This design described in this document is based upon the operational experience with data centers built to support online applications, such as Web search engines. The primary requirement in such environments is operational simplicity and network stability, in order to allow for a small group of people to support large network infrastructure.

After experimentation and extensive testing, the final design decision was made to use fully routed option with BGP as the control-plane protocol. This is in contrast with more traditional data center designs, which rely heavily on extending Layer 2 domains across multiple network devices. This document elaborates the network design requirements that led to this choice and presents detailed aspects of the BGP routing design.

2. Document structure

The remaining of this document is organized as following. First, the document gives a quick overview of the more traditional data center network designs, and analyzes reasons that often made designers ignore using BGP for data center routing in the past. Next, the design requirements for large scale data centers are presented and briefly discussed. Following this, the document gives an overview of Clos network topology and its properties. After that, the arguments for selecting BGP as the routing protocol for data center are presented. Finally, the document discusses the design in more details and covers specific BGP features used for the network configuration, as well as analyzes some properties of the proposed design.

3. 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.

3.1. Layer 2 designs

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

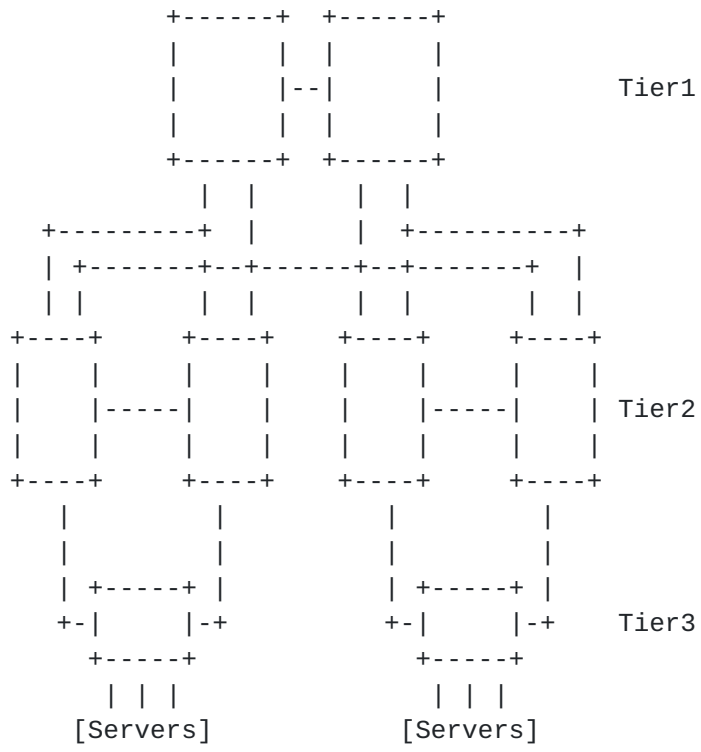


Figure 1: Typical Data Center network layout

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

- Supporting legacy applications that may require direct Layer 2 adjacency or use non-IP protocols

- Seamless mobility for virtual machines, to allow the preservation of IP addresses when a virtual machine moves across physical hosts

- Simplified IP addressing - less IP subnets is required for the data center

- Application load-balancing may require direct Layer 2 reachability to perform certain functions such as Layer 2 Direct Server Return (DSR)

3.2. Fully routed network designs

Network designs that leverage IP routing down to the access layer (Tier-3) of the network have gained popularity as well. The main benefit of such designs is improved network stability and scalability, as a result of confining L2 broadcast domains. 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, such fully routed designs become more attractive.

Although BGP is the de-facto standard protocol for routing on the Internet, having wide support from both the vendor and service provider communities, it is not generally deployed in data centers for a number of reasons:

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

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

- BGP deployment within an Autonomous System (iBGP mesh) is assumed to have a dependency on the presence of an IGP, which assists with recursive next-hop resolution.

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

In this document we demonstrate a practical approach for using BGP as the single routing protocol for data center networks.

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 to accommodate application bandwidth and latency requirements. Until recently it was quite common to see traffic flows mostly entering and leaving the data center (also known as north-south traffic) There were no intense highly meshed flows or traffic patterns between the machines within the data center. As a result, traditional "tree" topologies were sufficient to accommodate such flows, even with high oversubscription ratios in network equipment. If more bandwidth was required, it was added by "scaling up" the network elements, e.g. by upgrading the switch line-cards or fabrics.

In contrast, large-scale data centers often host applications that generate significant 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 migrations. Scaling up traditional tree topologies to match these bandwidth demands becomes either too expensive or impossible due to physical limitations.

4.2. CAPEX minimization

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

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

- Driving costs down using economic principles, by introducing multiple network equipment vendors.

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

4.3. OPEX minimization

Operating large scale infrastructure could be expensive, provided that larger amount of elements will statistically fail more often. Having a simpler design and operating using a limited software feature-set ensures that failures will mostly result from hardware malfunction and not software issues.

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

4.4. Traffic Engineering

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

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

5. Requirement List

This section summarizes the list of requirements, based on the discussion so far:

- REQ1: Select a network topology where capacity could be scaled "horizontally" by adding more links and network switches of the same type, without requiring an upgrade to the network elements themselves.

- REQ2: Define a narrow set of software features/protocols supported by a multitude of networking equipment vendors.

- REQ3: Among the network protocols, choose the one that has a simpler implementation in terms of minimal programming code complexity.

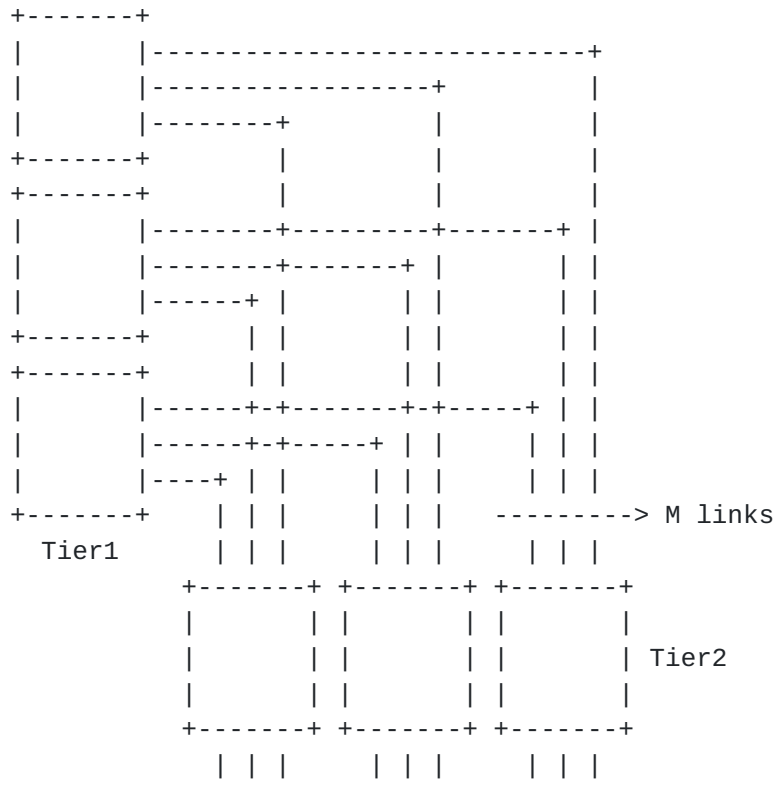
- REQ4: The network routing protocol should allow for explicit control of the routing prefix next-hop set using built-in protocol mechanics.

6. Network topology

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

6.1. Clos topology overview

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



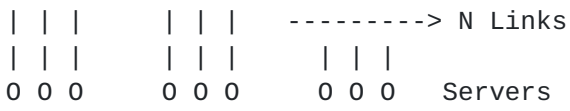


Figure 2: 3-Stage Folded Clos topology

In the networking industry, a topology like this is sometimes referred to as "Leaf and Spine" 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). However, for uniformity, we will continue to refer to these layers using the "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

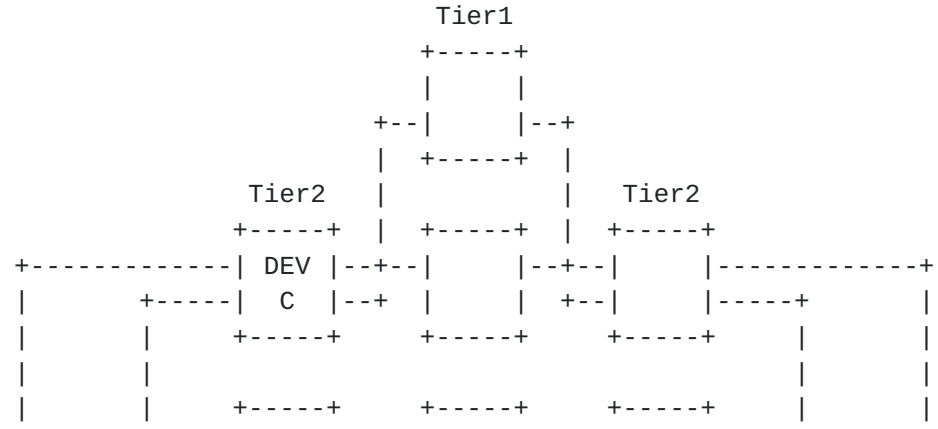
- Implementing Clos topology requires a routing protocol supporting ECMP with the fan-out of M or more

- Every Tier-1 device has exactly one path to every end host (server) in this topology

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

6.3. Scaling Clos topology

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



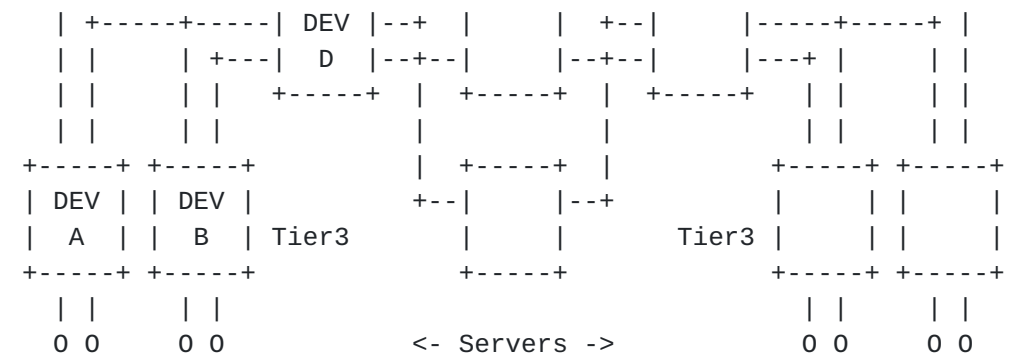


Figure 3: 5-Stage Clos topology

The topology on Figure 3 is built from switches with port count of 4 and provides full bisection bandwidth to all connected servers. We will refer to the collection of directly connected Tier-2 and Tier-3 switches along with their attached servers as a "cluster" in this document. For example, devices A, B, C, and D on Figure 3 form a cluster along with the servers they connect to.

In practice, the Tier-3 level of the network (typically top of rack switches, or ToRs) is where oversubscription is introduced to allow for packaging of more servers in data center. 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 two bandwidth pools: within the same access switch (e.g. rack) and outside of the local switch. Since oversubscription itself does not have any effect on the routing design, we will not be discussing it further in this document

6.4. Managing the size of Clos topology tiers

If the data-center network size is small, it is possible to reduce the number of devices in Tier-1 or Tier-2 of Clos topology by a factor of two, four or any other power of two. To understand how this could be done, let's take Tier-1 as example. Notice that every Tier-2 switch connects to a single group of Tier-1 switches. Imagine that half of the ports on each of Tier-1 switch is not being used. Then, it is possible to reduce the number of Tier-1 switches by half, and simply map two uplinks from a Tier-2 device to the same Tier-1 device - those uplinks were previously mapped to different Tier-1 devices. This technique maintains the same bisection bandwidth, but reduces the size of Tier-1 layer, thus saving on CAPEX. The tradeoff, naturally, is reduction of maximum data-center size in terms of server count by half, if Tier-1 size is reduced in this manner.

As a result of this transformation, Tier-2 switches will be using two or more parallel links to connect to each Tier-1 switch. If one of these links fails, the other one will pick up all traffic of the failed link, possible resulting in heavy congestion and quality of service degradation. To avoid this situation, the parallel links could be grouped in link aggregation groups (LAGs) with the configuration setting that takes the whole bundle down, upon a single link failure. Any equivalent technique that enforces "fate sharing" on the parallel links could be used in place of LAGs to achieve the same effect. As a result of such fate-sharing, traffic from two or more failed links will be re-balanced over the multitude of remaining paths, which is normally much higher than two, as it equals to the number of Tier-1 switches.

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 requirements discussed earlier call for a single routing protocol (REQ2) to reduce complexity and interdependencies. While it is common to rely on an IGP in this situation, the document proposes the use of BGP only. The advantages of using BGP are discussed below.

- BGP inherently has less complexity within its protocol design - internal data structures and state-machines are simpler when compared to a link-state IGP, such as OSPF. For example, instead of implementing adjacency formation, adjacency maintenance and/or flow-control, BGP simply relies on TCP as the underlying transport. This fulfills REQ1 and REQ2.

- BGP information flooding overhead is less when compared to link-state IGPs. Indeed, since every BGP router typically re-calculates and propagates best-paths only, a network failure is masked as soon as the BGP speaker finds an alternate path, which often exists in highly symmetric topologies, such as Clos. In contrary, the event propagation scope of a link-state IGP is single flooding domain, regardless of the failure type. Furthermore, even though this does not cause any significant impact on the modern routers, it is worth mentioning that 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. This allows for ECMP or forwarding based on application-defined

forwarding paths, by establishing an eBGP multi-hop peering session with the application "controller". This satisfied REQ4 stated above. Some IGPs, such as OSPF, support similar functionality using concepts such as "Forwarding Address", but do not satisfy other requirement, e.g. protocol simplicity.

- It is easy to lay down BGP ASN allocation scheme such that "BGP path hunting" is well-controlled, and complex unwanted paths are ignored. See below Section 7.2 for an example of such ASN allocation scheme. Such policy could not be enforced on a link-state IGP, and in result, under certain failure conditions, it may pick up unwanted lengthy paths, e.g. traverse multiple Tier-2 devices.

- Plain BGP configuration, without routing policies, is easier to troubleshoot for network reachability issues. For example, it is straightforward to dump contents of BGP Loc-RIB and compare it to the router's RIB and, possibly, FIB. Furthermore, every BGP neighbor has corresponding Adj-RIB-In and Adj-RIB-Out structures with incoming/outgoing NRLI information that could be easily correlated on both sides of the BGP peering session. Thus, BGP fully satisfies REQ3.

7.2. BGP configuration for Clos topology

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

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 guidelines that can be used:

- All BGP peering sessions are external BGP (eBGP) established over direct point-to-point links interconnecting the network nodes.

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

- Private BGP ASNs from the range 64512-64534 are used so as to avoid ASN conflicts. The private ASN stripping feature can be leveraged as a result (see below).

- A single BGP ASN is allocated to the Clos topology's middle stage ("Tier-1"), e.g. ASN 64534 as shown in Figure 4

- Unique BGP ASN is allocated per each group of "Tier-2" switches (e.g. aggregation switches).

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

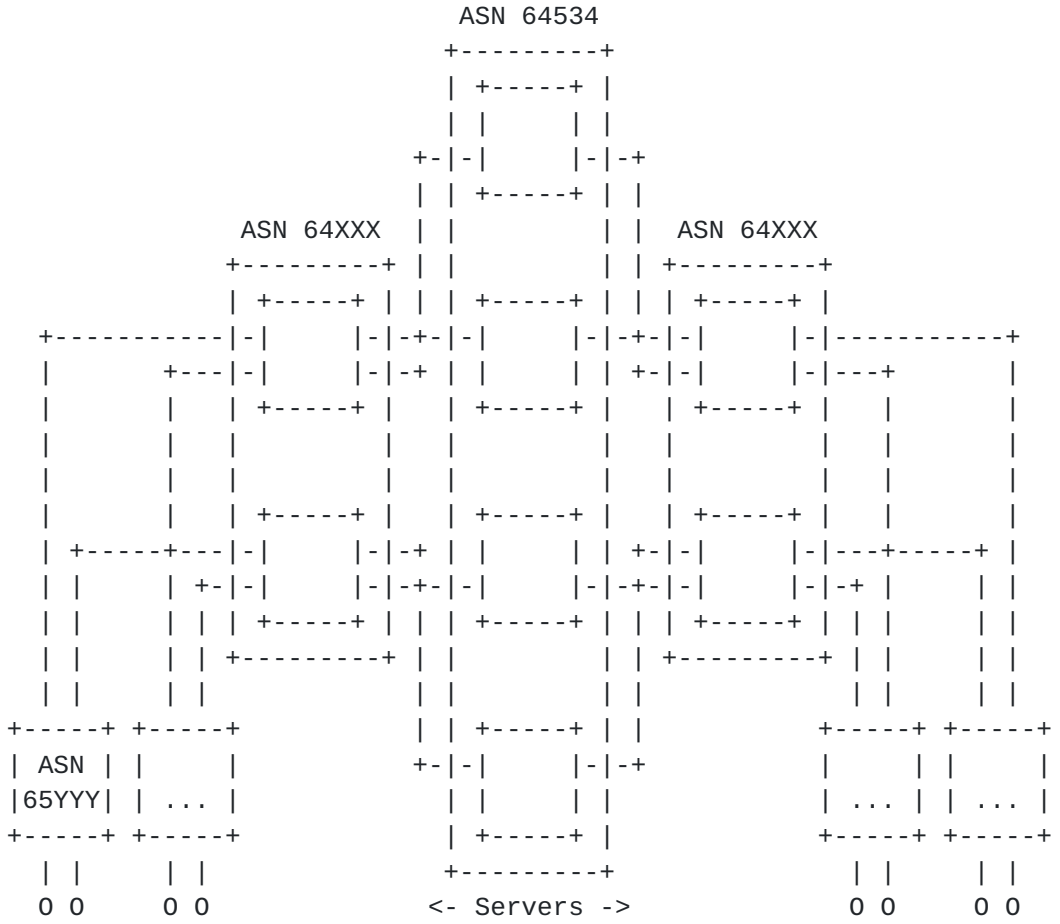


Figure 4: BGP ASN layout for 5-stage Clos

7.2.2. Non-unique private BGP ASN's

The use of private BGP ASNs limits to the usable range of 1022 unique numbers. Since it is very likely that the number of network switches could exceed this number, a workaround is required. One approach would be to re-use the private ASN's assigned to the Tier-3 switches across different clusters. For example, private BGP ASN's 65001, 65002 ... 65032 could be used within every individual cluster and assigned to Tier-3 switches.

To avoid route suppression due to AS PATH loop detection mechanism in BGP, upstream eBGP sessions on Tier-3 switches must be configured with the "AllowAS In" feature that allows accepting a device's own ASN in received route advertisements. Introducing this feature does not create an opportunity for routing loops under misconfiguration since the AS PATH is always incremented when routes are propagated between topology tiers.

Another solution to this problem would be to using four-octet (32-bit) BGP ASNs. However, there are no reserved private ASN range in the four-octet numbering scheme although efforts are underway to support this, see [I-D.mitchell-idr-as-private-reservation]. This will also require vendors to implement specific policy features, such as four-octet private AS removal from AS-PATH attribute.

7.2.3. Prefix advertisement

A Clos topology features a large number of point-to-point links and associated prefixes. Advertising all of these routes into BGP may create FIB overload conditions in the network devices. There are two possible solutions that can help prevent such FIB overload:

- Do not advertise any of the point-to-point links into BGP. Since eBGP peering changes the next-hop address anyways at every node, distant networks will automatically be reachable via the advertising eBGP peer

- 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 addressing.

Server facing subnets on Tier-3 switches must be announced into BGP without using route aggregation on Tier-2 and Tier-1 switches. Summarizing subnets in a Clos topology will result in route black-holing under a single link failure (e.g. between Tier-2 and Tier-3 switch) 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 switches. However, see the section Section 8 below for a method of performing route summarization in Clos networks and associated trade-offs.

7.2.4. 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 switches 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.

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

- Hide network topology information when advertising paths to WAN routers, i.e. remove private BGP ASNs from the AS-PATH attribute. This is typically done to avoid BGP ASN number collisions between different data centers. A BGP policy feature called "Remove Private AS" is commonly used to accomplish this. This feature strips a contiguous sequence of private ASNs found in AS-PATH attribute prior to advertising the path to a neighbor. This assumes that all BGP ASN's used for intra data center numbering are from the private ASN range.

- Originate a default route to the data center devices. This is the only place where default route 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. Notice that advertising the default route from Border Routers requires that all Border Routers to be fully connected to the WAN Routers upstream, to provide resistance to a single-link failure.

7.2.5. Route aggregation at the network edge

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

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

- Interconnect the Border Routers using a full-mesh of physical links or by using additional aggregation switches, forming hub-and-spoke topology. Next, build a full-mesh of iBGP sessions between all Border Routers to allow for sharing of specific network prefixes. Notice that in this case the interconnecting peer links need to be appropriately sized depending on the amount

of traffic that is planned to be taken in case of a device or link failure underneath the Border Routers.

- Tier-1 devices may have additional physical links running toward the Border Routers (which are Tier-2 devices in essence). Specifically, if a protection from a single node/link failure is desired, each Tier-1 devices would have to connect to at least two Border Routers. This puts additional requirements on the port count for Tier-1 devices and Border Routers, likely requiring the use of a different router model for Border Routers.

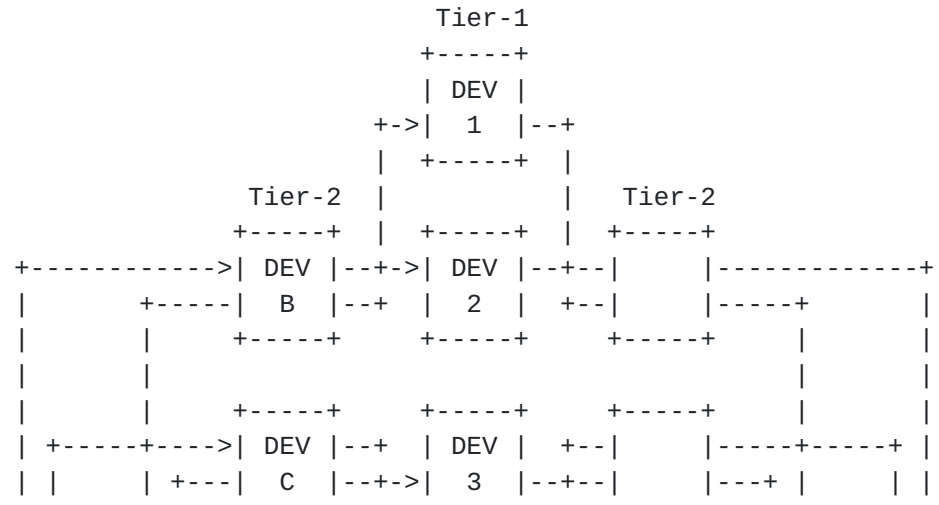
If any of the above option is implemented, it is possible to perform route aggregation at the Border Routers toward the WAN network core, without risking routing black-hole condition under a single link failure. Both of those options would result in non-uniform topology, as additional links have to be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

7.3. ECMP Considerations

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

7.3.1. Basic ECMP

ECMP is the fundamental load-sharing mechanism used by a Clos topology. Effectively, every lower-tier switch will use all of its directly attached upper-tier devices to load-share traffic destined to the same IP prefix. Number of ECMP paths between any two Tier-3 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.



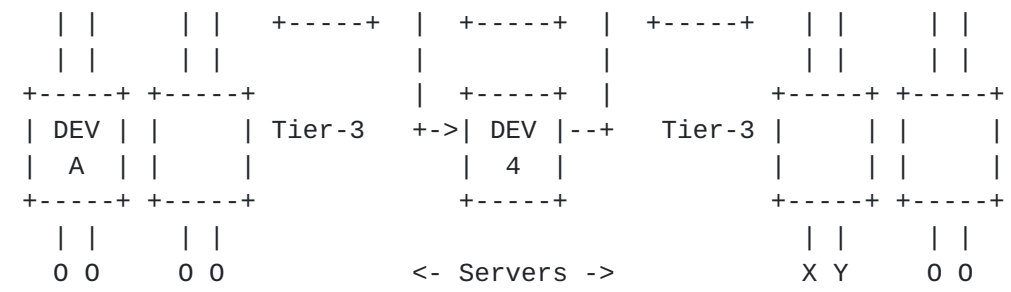


Figure 5: ECMP fan-out tree from A to X and Y

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

Most BGP implementations declare paths to be equal from ECMP perspective if they match up to and including step (e) Section 9.1.2.2 of [RFC4271]. In the proposed network design there is no underlying IGP, so all IGP costs are automatically assumed to be zero (or otherwise the same value across all paths). Loop prevention is assumed to be handled by the BGP best-path selection process, specifically by comparing the AS_PATH lengths.

7.3.2. BGP ECMP over multiple ASN

For application load-balancing purposes it is desirable to have the same prefix advertised from multiple Tier-3 switches. 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, the BGP implementations must support load-sharing over above-mentioned paths. This feature is sometimes known as "AS PATH multipath relax" and effectively allows for ECMP to be done across different neighboring ASNs.

7.3.3. 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 are to be injected using remote BGP speaker (central agent) over an eBGP multihop or iBGP session (see Section 7.5 for more information on third-party route injection. Sigaling wise, the weight-distribution for multiple BGP paths could be done using the technique described in [I-D.ietf-idr-link-bandwidth].

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 deactivation upon failure of an associated link.

7.4.1. Fault detection timing

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

Relying solely on BGP keep-alive packets may result in high convergence delays, in the order of multiple seconds (commonly, the minimum recommended BGP hold timer 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 fail-over". Since the majority of the links in modern data centers are point-to-point fiber connections, a physical interface failure if often detected in milliseconds and subsequently triggers a BGP re-convergence.

Furthermore, popular link technologies, such as 10Gbps Ethernet, may support a 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 (BFD) ([RFC5880]) to allow for sub-second failure detection and fault signaling to the BGP process. This presents additional requirements to vendor software and possibly hardware, and may contradict REQ1.

7.4.2. Event propagation timing

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

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

7.4.3. Impact of Clos topology fan-outs

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

The impact of such "dispersion" grows with the size of topology fan-out, and could become more noticeable under network convergence churn.

7.4.4. Failure impact scope

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

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. The worst case is that ALL devices in data center topology would have to either withdraw a prefix completely, or update the ECMP groups in the FIB. However, many failures will not result in such 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, Tier-2 device will simply have to update its ECMP group, removing the failed link. There is no need to send new information to the downstream Tier-3 devices. The affected Tier-1 device will lose the only path available to reach a particular cluster and will have to withdraw the affected prefixes. Such prefix withdrawal process will only affect Tier-2 switches directly connected to the affected Tier-1 device. In turn, the Tier-2 devices receiving BGP UPDATE message withdrawing prefixes will simply have to update their ECMP groups for affected prefixes. The Tier-3 devices will not be involved in re-convergence process.

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

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

Even though BGP offers some failure scope reduction, reduction of the fault domain using summarization is not always possible with the proposed design, since using this technique may create route black-holing issues as mentioned previously. Thus, the worst control-plane failure impact scope is the network as a whole, e.g. in a case of a link failure between Tier-2 and Tier-3 switches. However, the amount of affected prefixes in this case would be much less, as compared to a failure in the upper layers of a Clos network topology. Finally, it is worth pointing that the property of having such large failure scope is not a result of choosing BGP, but rather a result of using the "scale-out" Clos topology.

7.4.5. Routing micro-loops

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

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

7.5. Third-party route injection

BGP allows for a third-party BGP speaker (not necessarily directly attached to the network devices) to inject routes anywhere in the network topology. This could be achieved by peering an external speaker using an 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 to facilitate load-balancing, or using the BGP Add-Path extension (see [I-D.walton-bgp-add-paths]) if supported by the implementation.

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.

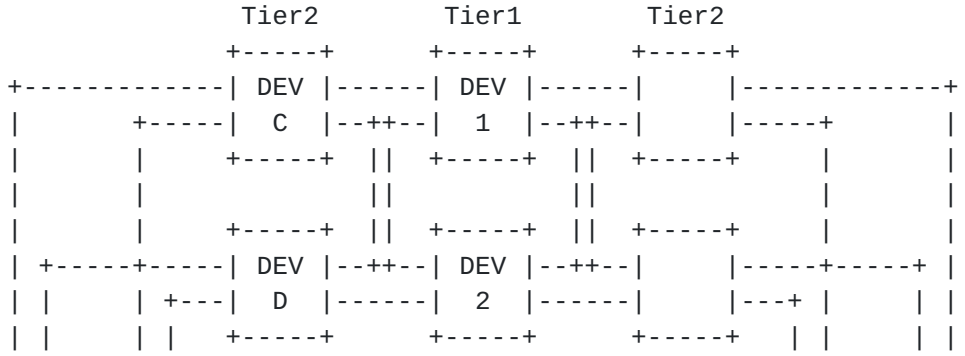
8. Adding route aggregation to Clos topology

As mentioned previously, route aggregation is not possible in "native" Clos topology, since it makes network susceptible to route black-holing under a single link failure. The main problem is limited number of parallel paths between network elements, for example only a single path between any pair of Tier-1 and Tier-3 switches. However, some operators may find route aggregation desirable to improve network control plane stability.

With this said, it is possible to change the network topology design and allow for route aggregation, though the trade-off would be reduced size of the total network, and network congestion under specific failures. This approach is very similar to the techique described above, to allow Border Routers to summarize the data-center address space.

8.1. Collapsing Tier-1 switches layer

In order to add more paths between Tier-1 and Tier-3 switches, imagine that we group Tier-2 switches in pairs, and connect the pair to the same group of Tier-1 switches. This is logically equivalent to "collapsing" Tier-1 switches into a group of half size, merging the links on the "collapsed" devices. The result is illustrated on the figure Figure 6 below. 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 there connecting to different groups of Tier-1 devices.



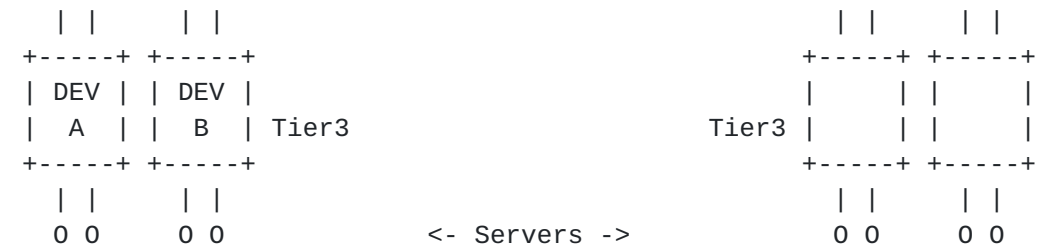


Figure 6: 5-Stage Clos topology

With this design, Tier-2 devices may not advertise a default route only 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 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 failure - it is possible for a double link failure to isolate Tier-2 device from all paths toward a select Tier-3 device, thus causing routing black-hole.

8.2. Implications of the network design change

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

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

9. Security Considerations

The design does not introduce any additional security concerns. 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.

10. IANA Considerations

This document includes no request to IANA.

11. Acknowledgements

This publication summarizes work of many people who participated in developing, testing and deploying the proposed network design. 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 Linda Dunbar, Susan Hares and Jon Mitchell for reviewing the document and providing valuable feedback.

12. 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-08 (work in progress), July 2012.

[I-D.mitchell-idr-as-private-reservation] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", draft-mitchell-idr-as-private-reservation-01 (work in progress), August 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.

[I-D.walton-bgp-add-paths] Walton, D., "Advertisement of Multiple Paths in BGP", draft-walton-bgp-add-paths-06 (work in progress), July 2008.

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

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

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

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

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

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/