

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: August 18, 2014

F. Mejia, Ed.  
AEPROVI  
R. Gagliano  
A. Retana  
Cisco Systems  
C. Martinez  
G. Rada  
LACNIC  
February 14, 2014

Implementing RPKI-based origin validation one country at a time. The  
Ecuadorian case study.

draft-fmejia-opsec-origin-a-country-00

## Abstract

One possible deployment strategy for BGP origin validation based on the Resource Public Key Infrastructure (RPKI) is the construction of islands of trust. This document describes the authors' experience deploying and maintaining a BGP origin validation island of trust in Ecuador.

## 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 August 18, 2014.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
1.1. Policer Network . . . . .	3
1.2. The resource holders . . . . .	4
1.3. RPKI certificate authorities and repository . . . . .	5
1.4. The technical support . . . . .	5
2. Objective . . . . .	5
3. Planning . . . . .	6
3.1. RPKI-based origin validation support . . . . .	6
3.2. Deploying a RPKI cache into the network . . . . .	7
3.3. Populating the RPKI database . . . . .	7
3.4. Action to take with NotFound and Invalid prefixes . . . . .	8
4. Deployment . . . . .	8
4.1. RPKI Validation servers . . . . .	9
4.2. Origin validation setting . . . . .	10
5. Training and RPKI signing event . . . . .	11
6. Outcome and post-event activities . . . . .	11
7. Lessons learned and best practices . . . . .	12
8. IANA Considerations . . . . .	13
9. Security Considerations . . . . .	13
10. Acknowledgements . . . . .	13
11. Informative References . . . . .	13
Appendix A. Router configuration templates . . . . .	14
Authors' Addresses . . . . .	17

## 1. Introduction

BGP origin validation based on RPKI [RFC6811] is in early stages of deployment. As with other new technologies, there are impediments to its global adoption as its full value is not yet perceived. Particularly, RPKI based origin validation involves on one side the creation of a large enough set of signed objects and on the other side the application of network policies based on these signed objects by network operators. An operator that does not see a large enough set of signed objects in the RPKI repository system is not encouraged to implement these set of policies. Conversely, IP address space resource holders that are not required by network operators (i.e. transit providers, peers or operators community in general) to create and maintain their RPKI objects have little incentive to do so.

To overcome this bootstrap problem, it is necessary to create a success story that brings enough value to both: network operators and resource holders. Moreover, one possible strategy for the adoption of a security technology is the creation of islands of trust where the technology is fully deployed in a reduced environment. In this direction, some organizations carried forward a full implementation of an island of trust in Ecuador. This was a multi stakeholder project where each party (resource holders, an Internet Exchange Point manager, a Regional Internet Registry and an equipment manufacturer) contributed to its success.

This document describes the experience of implementing RPKI-based origin validation in Ecuador and it is expected to be an useful guide to start other similar projects.

Below, it is described the different roles in the project and the involved parties.

### 1.1. Policer Network

In this document, the "Policer Network" is the networking infrastructure where the origin validation based on RPKI will be deployed to apply polices on BGP announcements. NAP.EC ([www.nap.ec](http://www.nap.ec)) was selected for this role.

NAP.EC is the Internet Exchange Point (IXP) in Ecuador with two Points of Presence (POPs): Quito (UIO) and Guayaquil (GYE). It has a BGP route-server in each location with a mandatory multilateral routing policy (i.e. all participants have a BGP session to the route-server). Each location uses a different IP address block and Autonomous System Number (ASN). NAP.EC is a meeting point where many organizations (Internet Service Providers, content providers, root servers, etc.) exchange routing information.

The participants connected to NAP.EC announce almost 100% of the total address space used in Ecuador (be believe it is 100% but we cannot be certain though). In some cases they announce their own address space and in some cases they are transit providers for their customers' resources.

AEPROVI ([www.aeprovi.org.ec](http://www.aeprovi.org.ec)) manages the NAP.EC infrastructure. It is a non-profit organization, based on membership and brings together around 30 Ecuadorian ICT-related companies. AEPROVI also has an excellent reputation as an innovator in the local networking community thanks to projects such as IPv6 adoption and CDN cache servers hosting. These projects have given the local community concrete value and have build the trust on the team that manages the local IXP.

Thanks to this trust, and to the fact that all local BGP announcements are performed through the route servers, NAP.EC is uniquely positioned to become the "policer network" for this project. At the same time, it can be said that implementing origin validation at the NAP.EC route servers is equivalent to implementing it for all inter-domain routing in the country.

## 1.2. The resource holders

In this document, an organization which operates their own IP prefixes is called resource holder or simply the holder. They may have resources allocated/assigned from a Regional Internet Registry (RIR) and/or legacy resources (if the allocation was done before RIR formation). Resource holders are responsible for creating the RPKI signed objects for this project.

NAP.EC routing tables involve a number of holders, including organizations like Internet Service Providers, content providers, universities, .ec domain administrator and root servers. Most of them are Ecuadorian companies and have received IP resources only from LACNIC, but some have both RIR and legacy resources. Moreover, a few holders are foreign companies and their resources are legacy or from other RIRs (e.g. root servers and content providers).

Not all resource holders are directly connected to the NAP.EC fabric; some have IP address resources but not an ASN and some others are small networks that receive traffic from other bigger networks. In this case their IP address prefixes are announced by their transit providers. One of the main challenges for this project was to identify all the resource holders that needed to be contacted and to encourage network administrators from these organizations to participate.

In addition, some resource holders are part of a larger (and sometimes international) organization, with strong change management processes. This means that any change on their configurations needed to be planned ahead of time and consulted outside of the country.

In NAP.EC - UIO, the routing table includes prefixes used in Ecuador and other countries.

In NAP.EC - GYE, the routing table includes prefixes from companies operating only in Ecuador.

For the project, the target was limited to prefixes used in Ecuador by Ecuadorian holders that had received resources from LACNIC until mid-2013.

### 1.3. RPKI certificate authorities and repository

The five Regional Internet Registries (RIRs) have a critical role in the RPKI trust model since they manage the trust anchors of the RPKI hierarchic design. Additionally, due to some reasons (e.g. economics, skills) the scenario where the Certification Authority (CA) certificate is hosted by a RIR will be the most popular for a long time, in which case, RIR's online software tools to manage RPKI objects are imperative.

The RIR-hosted RPKI CA model was used for this project. Local RPKI validation servers (validation and cache) were locally deployed. This means that all resource holders had to create and manage their RPKI signed objects using the online tools implemented by LACNIC and that the local validation servers retrieve these objects from the RIR's public global repositories. No local RPKI CA nor repository were configured.

LACNIC also runs a RPKI testbed (test CA with correspondent GUI and Trust Anchor material). This infrastructure was used during the training activity.

### 1.4. The technical support

RPKI and origin validation are in the early stage of deployment. Few people have full knowledge about its RFCs, the implementation support in different routers and the maintenance of RPKI signed objects. To involve trained people and train new ones is very important.

People from an equipment manufacturer (Cisco) contributed with support in the startup stage and to train the holders' staff. LACNIC's staff contributed developing new online RPKI tools and training about how to use them.

## 2. Objective

Considering all the definitions given during the introduction and after several discussions through face and online meetings among the involved parties, the following objective was agreed on:

"Deploy RPKI-based BGP origin validation in NAP.EC's route servers. For the success of the project, 80% of the Ecuadorian prefixes (both IPv4 and IPv6) received by those routers should have a valid origin."

In order to monitor the progress, NAP.EC - GYE was taken as reference because NAP.EC - UIO had non-Ecuadorian prefixes announced.

### 3. Planning

The project started with an initial idea from a very reduced number of enthusiasts that identified a suitable network (the island of trust), involved the appropriate organizations and set milestones in order to carry forward a full implementation of the technology. Into the process, all parties identified the gaps and proposed solutions to overcome them.

One point that it was wanted to guarantee is that we would be able to create the appropriate "buzz" around the project. So, a communication strategy should not be overlooked. In this case, LACNIC and AEPROVI signed a MoU in April 2013 and all parties (LACNIC, AEPROVI and Cisco) announced the project and issued a press release at the LACNIC event in May 2013.

Some points that required specific discussion by the core team included:

1. RPKI-based origin validation support in the route-servers equipments
2. How to deploy a RPKI cache into the Network
3. How to populate the RPKI database with the correct and necessary information
4. Action to take with NotFound and Invalid prefixes

#### 3.1. RPKI-based origin validation support

NAP.EC uses Cisco equipment. The project started with the initial idea of to implement origin validation into existing routers used as route servers, simply after a software update or upgrade. However, the vendor had no plans to support it in the existing platform. AEPROVI had future plans to carry forward a routers renewal, then this issue was overcome but it stopped the project for some time. Describing the equipment renewal process is beyond the scope of this document.

For Cisco equipment, the vendor has made available some online software tools to check the support. About origin validation, the routers must support: RTR protocol [RFC6810] and RPKI-based origin validation [RFC6811]. Moreover, among other things, four octects ASN support [RFC6793] and IPv6 routing support ([RFC2545] and some others) are mandatory in NAP.EC.

The selected routers were two Cisco ASR-1000 series routers (one for Quito and other for Guayaquil).

### 3.2. Deploying a RPKI cache into the network

Based on available resources and existing skills, it was decided to use Virtual Machines (VM) as RPKI caches, which would run GNU Linux.

The validating software is in the early stages of its development and there could be bugs or reliability problems, so it was decided using two different packages (processes) in each VM.

To ensure high availability, it was decided to deploy two VMs, each one in different host server.

There are no servers in Guayaquil, therefore both VMs would be in Quito within the NAP.EC management network and connect via the RTR protocol to the route-servers located in Quito and Guayaquil.

Additionally, the firewall rules allow RTR connections from the NAP.EC LANs to the RPKI validator servers in order to facilitate participants to perform origin validation in their edge equipments (if they wish to in the future).

### 3.3. Populating the RPKI database

The IP resource holders must create all needed RPKI data for the project, at least certificates and Route Origin Authorizations (ROAs). Moreover, the technical staff needed training about RPKI and origin validation because it is a new technology. Accordingly, a reasonable method to achieve it should be contrived.

It was decided to organize an event with two objectives: training and RPKI object signing. One key planning activity was to create the list of participants and to make sure that at least one participant per network had the authentication credentials to the LACNIC system to create its RPKI signed objects.

The target community was limited to Ecuadorian organizations that had received IP resources from LACNIC until mid-2013. That meant around fifty (50) organizations including Internet Services Providers, universities, banks, etc., or expressing it in prefixes: around 8600 IPv4 and 60 IPv6 blocks.

Some weeks before the deployment, there was informal dissemination meetings between the NAP.EC administrator and the participants. The project milestones were reported and the attendees received information about RPKI and origin validation for the first time. A

complete training was offered as a project milestone in the next few weeks.

All organizations were contacted and received an invitation to the event. More information about this can be found in Section 5.

#### 3.4. Action to take with NotFound and Invalid prefixes

Despite the efforts, the RPKI database information may be incomplete, therefore the routing tables often will have NotFound prefixes. Moreover, it is needed some time after the first contact with RPKI-based origin validation technology to fix possible errors (e.g. invalid prefixes) and to assess the impact. A strict policy of dropping prefixes did not seem convenient as a starting point for the project.

It was decided that NAP.EC proceeds as follows:

- At the beginning, the NAP.EC's routers only would monitor the RPKI origin state of prefixes without action.
- In the near future, NAP.EC administration might change the action based on results of the signing-party event and community consensus.

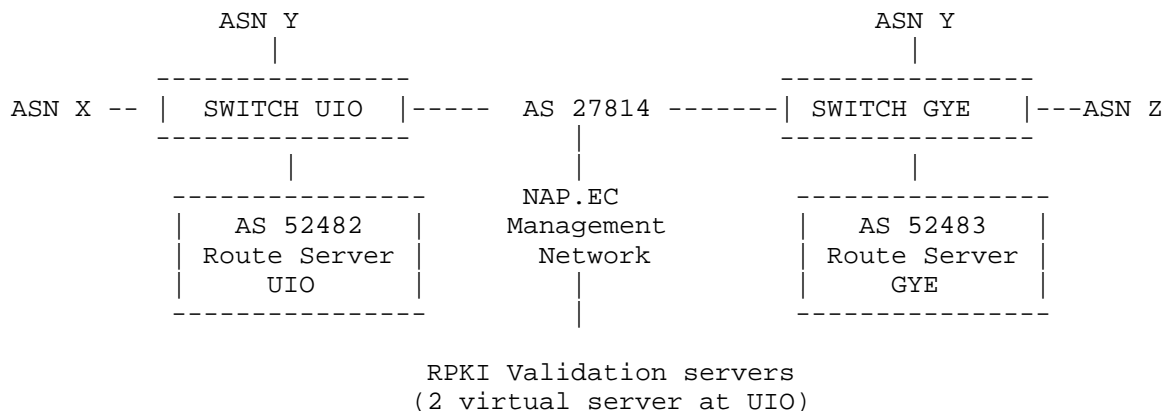
As part of the second stage, some days after the signing event, each prefix is being marked with a BGP community to identify its RPKI origin state, sending that information to the participants.

Finally, some months later, a date was set to begin applying a strict policy. The policy was defined as follows: dropping Invalid prefixes and setting a lower local preference for NotFound prefixes.

#### 4. Deployment

Following is the NAP.EC topology during the deployment:





#### 4.1. RPKI Validation servers

A virtual machine (named VM1) on VMware ESXi was deployed, running GNU Linux, Centos distribution. The other one (named VM2) was cloned from this one.

Each virtual machine has access to the Internet through 1 (one) ethernet interface with a public IPv4 address within the same subnet like this: 192.0.2.2/27 for VM1, 192.0.2.3/27 for VM2 and 192.0.2.1/27 for network gateway.

The 192.0.2.0/27 network is within AS 27814. AS 27814 contains NAP.EC monitoring equipment (among other things) and it is connected to NAP.EC - Quito and NAP.EC - Guayaquil.

Each VM has the following packages (installation and configuration guides of these packages are beyond the scope of this document):

- ntpd (as NTP client)
- iptables (as firewall)
- apache (as web server)
- validating software from RIPE (<http://www.ripe.net/lir-services/resource-management/certification/tools-and-resources>)
- validating software from the rpki.net project (<http://rpki.net/wiki/doc/RPKI/Installation>)

Each validating software was setup on a different port:

- validating software from RIPE, port 65001,

- validating software from the rpki.net project, port 65002.

Each validating software has a monitoring web page, each one was configured on different port:

- RIPE software, port 65081,
- rpki.net software, port 65082.

#### 4.2. Origin validation setting

Since the routers renewal, NAP.EC - Quito and NAP.EC - Guayaquil have each one a Cisco ASR-1000 series router as route server. These routers have IOS-XE version 3 which runs a process with IOS version 15.

First, it is required to configure communication via RTR protocol between the routers and the RPKI validation servers (caches). In this step, the necessary data are: IP addresses and service ports of the caches and the time which the router will re-query the cache (refresh time). Currently, RPKI information does not change quickly, therefore 600 seconds (10 minutes) may be considered enough for refresh time.

Cisco IOS 15 drops Invalid prefixes by default, but there is a command to avoid this behavior (ibgp bestpath prefix-validate allow-invalid). This must be applied while the policy is no action.

Later, a route-map is required to configure any action as applying different BGP local preference or marking each prefix with a BGP community based on its RPKI origin state (also send-community option must be enabled within the BGP session configuration):

The following community assignment policy was applied:

- <IXP-ASN>:21 --> Valid origin
- <IXP-ASN>:22 --> NotFound origin
- <IXP-ASN>:23 --> Invalid origin

Where <IXP-ASN> equals:

- 52482 for NAP.EC - Quito, or
- 52483 for NAP.EC - Guayaquil.

The template used in NAP.EC is in Appendix A.

## 5. Training and RPKI signing event

The event was called "Seminario sobre seguridad en el encaminamiento de Internet: BGP RPKI - Validacion de origen" and was scheduled for September 4-5, 2013. The agenda included theoretical and practical training, plus two time slots to sign RPKI objects: one at the end of the first day and other one during the second day.

Lack of training materials was a issue to overcome during preparatory work of the event. Some necessary activities were:

- The instructors (four people) prepared materials to cover topics such as BGP, RPKI, origin validation and the new NAP.EC platform.
- LACNIC's staff developed two new on-line tools: RPKI ROA wizard (<http://tools.labs.lacnic.net/roa-wizard/>) and RPKI announcement (<http://tools.labs.lacnic.net/announcement/>), further improved the demo environment of the RPKI system (<http://rpkidemo.labs.lacnic.net/>).
- Cisco's staff implemented a temporary virtualized network with many routers supporting RPKI and origin validation.

The event took place in a hotel and had Internet to access the training tools and the real LACNIC's hosted RPKI system.

Not all organizations sent a representative. The attendance represented around 80% of the target prefixes.

## 6. Outcome and post-event activities

Before the event, less 1% of the Ecuadorian prefixes were signed. At the start of the second day, less than 20% of the Ecuadorian prefixes were covered by a ROA. At the end of the event, around 80% of the Ecuadorian prefixes had a RPKI origin state as Valid.

MRTG graphs were implemented to monitor the amount of Valid, NotFound and Invalid prefixes after the event.

Feedback was received from attendees before closing the event. Some people recommended applying an acceptable policy in order do not waste the successful effort.

A few days after the event, some non-attending organizations were contacted by the NAP.EC administrator and meetings were coordinated for ROA creation. After these activities, almost 100% of Ecuadorian prefixes are covered for a ROA.

Communication activities performed after the event included:

- This document and presentation at relevant IETF Working Groups.
- Presentation at IEPG, LACNIC and other NOG events
- Publication at tech sites
- Note at local regulator newsletter
- Document and presentation at CITEC (Organization of American States)
- Blogging and social media in relevant platforms

As subsequent operational tasks, an update of validating software was performed. Overall, management has been simple and without major problems.

## 7. Lessons learned and best practices

- Implementation support needs to be verified in all target platforms.
- The IP resource holders community need RPKI-based origin validation training. Operators are less conservative than original though by organizers and once RPKI local space was full, support for removing invalid was unanimous.
- One day for a RPKI signing party is insufficient. the participants may not be confident about their skills or may need further authorization. Two days is a better practice (people need to sleep over what they learned the first day).
- From now on, when a new ISP wants to join NAP.EC, it receives information about RPKI-based origin validation and it is invited to create its ROAs.
- The event was a great opportunity to assemble the local community, particularly resource holders that had no previous participation at the local IXP.
- Initial work to have the "right people" in the room is a key to success. Particularly, operators need to have access to their RIR account.
- Post event communication needs to be discussed ahead of time.

## 8. IANA Considerations

No IANA requirements

## 9. Security Considerations

This document describes the experience of implementing a BGP origin validation island of trust in Ecuador. The actions taken are explicitly to be able to validate the origin in a BGP advertisement. There were no security-related issues identified during the deployment.

## 10. Acknowledgements

The authors wish to thank:

- all attendees at the training and RPKI signing event, without them this would not have happened.
- AEPROVI, LACNIC and Cisco for supporting the project.
- Arturo Servin for supporting the project from the start.
- Francisco Balarezo, Andres Piazza, Nicolas Fiumarelli and Chip Sharp as well as ISOC and Andean-Trade.

## 11. Informative References

- [RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, December 2012.
- [RFC6810] Bush, R. and R. Austein, "The Resource Public Key Infrastructure (RPKI) to Router Protocol", RFC 6810, January 2013.
- [RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, January 2013.

## Appendix A. Router configuration templates

## TEMPLATE 1

Policy: only marking prefixes based on RPKI origin state.

```
router bgp <IXP-ASN>

  bgp rpki server tcp 192.0.2.2 port 65001 refresh 600
  bgp rpki server tcp 192.0.2.2 port 65002 refresh 600
  bgp rpki server tcp 192.0.2.3 port 65001 refresh 600
  bgp rpki server tcp 192.0.2.3 port 65002 refresh 600
  !
  neighbor <neighbor-IPv4> remote-as <neighbor-IPv4-ASN>
  neighbor <neighbor-IPv4> version 4
  !
  neighbor <neighbor-IPv6> remote-as <neighbor-IPv6-ASN>
  neighbor <neighbor-IPv6> version 4
  !
  address-family ipv4
    bgp bestpath prefix-validate allow-invalid
    neighbor <neighbor-IPv4> send-community
    neighbor <neighbor-IPv4> route-map <route-map-name> out
  exit-address-family
  !
  address-family ipv6
    bgp bestpath prefix-validate allow-invalid
    neighbor <neighbor-IPv6> send-community
```

```
        neighbor <neighbor-IPv6> route-map <route-map-name> out
    exit-address-family
!
!
ip bgp-community new-format
!
!
route-map <route-map-name> permit 10
    match rpki valid
    set community <IXP-ASN>:21 no-export
!
route-map <route-map-name> permit 20
    match rpki not-found
    set community <IXP-ASN>:22 no-export
!
route-map <route-map-name> permit 30
    match rpki invalid
    set community <IXP-ASN>:23 no-export
!
```

## TEMPLATE 2

Policy: Dropping Invalid prefixes and setting lower local preference for NotFound prefixes.

```
router bgp <IXP-ASN>
    bgp rpki server tcp 192.0.2.2 port 65001 refresh 600
    bgp rpki server tcp 192.0.2.2 port 65002 refresh 600
```

```
bgp rpki server tcp 192.0.2.3 port 65001 refresh 600
bgp rpki server tcp 192.0.2.3 port 65002 refresh 600
!
neighbor <neighbor-IPv4> remote-as <neighbor-IPv4-ASN>
neighbor <neighbor-IPv4> version 4
!
neighbor <neighbor-IPv6> remote-as <neighbor-IPv6-ASN>
neighbor <neighbor-IPv6> version 4
!
address-family ipv4
    neighbor <neighbor-IPv4> send-community
    neighbor <neighbor-IPv4> route-map <route-map-name> out
exit-address-family
!
address-family ipv6
    neighbor <neighbor-IPv6> send-community
    neighbor <neighbor-IPv6> route-map <route-map-name> out
exit-address-family
!
!
ip bgp-community new-format
!
!
route-map <route-map-name> permit 10
```



```
match rpki valid

set community <IXP-ASN>:21 no-export

!

route-map <route-map-name> permit 20

match rpki not-found

set local-preference 50

set community <IXP-ASN>:22 no-export

!

!
```

#### Authors' Addresses

Fabian Mejia (editor)  
AEPROVI  
Av. Republica de El Salvador N34-211  
Quito  
EC

Email: [fabian@aeprovi.org.ec](mailto:fabian@aeprovi.org.ec)

Roque Gagliano  
Cisco Systems  
Avenue des Uttins 5  
Rolle 1180  
Switzerland

Email: [rogaglia@cisco.com](mailto:rogaglia@cisco.com)

Alvaro Retana  
Cisco Systems  
7025 Kit Creek Rd.  
Research Triangle Park, NC 27617  
US

Email: [aretana@cisco.com](mailto:aretana@cisco.com)

Carlos Martinez  
LACNIC

Email: carlos@lacnic.net

Gerardo Rada  
LACNIC

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: August 17, 2014

Camilo Cardona  
IMDEA Networks/UC3M  
Pierre Francois  
IMDEA Networks  
Paolo Lucente  
Cisco Systems  
February 13, 2014

Making BGP filtering a habit: Impact on policies  
draft-ietf-grow-filtering-threats-02

Abstract

Network operators define their BGP policies based on the business relationships that they maintain with their peers. By limiting the propagation of BGP prefixes, an autonomous system avoids the existence of flows between BGP peers that do not provide any economical gain. This draft describes how unexpected traffic flows can emerge in autonomous systems due to the filtering of overlapping BGP prefixes by neighboring domains.

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 August 17, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Filtering overlapping prefixes . . . . .	3
2.1. Local filtering . . . . .	3
2.2. Remotely triggered filtering . . . . .	6
3. Uses of overlapping prefix filtering that create unexpected traffic flows . . . . .	6
3.1. Unexpected traffic Flows . . . . .	7
3.1.1. Unexpected traffic flows caused by local filtering of overlapping prefixes . . . . .	8
3.1.2. Unexpected traffic flows caused by remotely triggered filtering of overlapping prefixes . . . . .	12
4. Techniques to detect unexpected traffic flows caused by filtering of overlapping prefixes . . . . .	15
4.1. Being the 'victim' of unexpected traffic flows . . . . .	15
4.2. Being a contributor to the existence of unexpected traffic flows in other networks . . . . .	15
5. Techniques to counter unexpected traffic flows due to the filtering of overlapping prefixes . . . . .	16
5.1. Reactive counter-measures . . . . .	17
5.2. Anticipant counter-measures . . . . .	18
5.2.1. Access lists . . . . .	18
5.2.2. Automatic overlapping prefix filtering . . . . .	19
5.2.3. Neighbor-specific forwarding . . . . .	19
6. Conclusions . . . . .	19
7. References . . . . .	20
7.1. References . . . . .	0
7.2. URIs . . . . .	20
Authors' Addresses . . . . .	20

## 1. Introduction

It is common practice for network operators to propagate overlapping prefixes along with the prefixes that they originate. It is also possible for some Autonomous Systems (ASes) to apply different policies to the overlapping (more specific) and the covering (less specific) prefix. Some ASes can even benefit from filtering the overlapping prefixes.

BGP makes independent, policy driven decisions for the selection of the best path to be used for a given IP prefix. However, routers

must forward packets using the longest-prefix-match rule, which "precedes" any BGP policy (RFC1812 [1]). Indeed, the existence of a prefix  $p$  that is more specific than a prefix  $p'$  in the Forwarding Information Base (FIB) will let packets whose destination matches  $p$  be forwarded according to the next hop selected as best for  $p$  (the overlapping prefix). This process takes place by disregarding the policies applied in the control plane for the selection of the best next-hop for  $p'$  (the covering prefix). When an Autonomous System filters overlapping prefixes and forwards packets according to the covering prefix, the discrepancy in the routing policies applied to covering and overlapping prefixes can create unexpected traffic flows that infringe the policies of other ASes still holding a path towards the overlapping prefix.

This document presents examples of such cases and discusses solutions to the problem. The objective of this draft is to shed light on the use of prefix filtering by making the routing community aware of the cases where the effects of filtering might turn to be negative for the business of Internet Service Providers (ISPs).

The rest of the document is organized as follows: Section 2 illustrates the motivation to filter overlapping prefixes. In Section 3, we provide some scenarios in which the filtering of overlapping prefixes lead to the creation of unexpected traffic flows on other ASes. Section 4 and Section 5 discuss some techniques that ASes can use for, respectively, detect and react to unexpected traffic flows.

## 2. Filtering overlapping prefixes

There are several scenarios where filtering an overlapping prefix is relevant to the operations of an AS. In this section, we provide examples of these scenarios. We differentiate cases in which the filtering is performed locally from those where the filtering is triggered remotely. These scenarios will be used as a base in Section 3 for describing side effects bound with such practices.

### 2.1. Local filtering

Let us first analyze the scenario depicted in Figure 1. AS1 and AS2 are two autonomous systems spanning a large geographical area and peering in 3 different physical locations. Let AS1 announce prefix 10.0.0.0/22 over all peering links with AS2. Additionally, let us define that there is part of AS1's network which exclusively uses prefix 10.0.0.0/24 and which is closer to a peering point than to others.

To receive the traffic destined to prefix 10.0.0.0/24 on the link closer to this subnet, AS1 could announce the overlapping prefix only over this specific session. At the time of the establishment of the peering, it can be defined by both ASes that hot potato routing would happen in both directions of traffic. In other words, it was agreed that each AS will deliver the traffic to the other AS on the nearest peering link. In this scenario, it becomes relevant to AS2 to enforce such practice by detecting the described situations and automatically issuing the appropriate filtering. In this case, by implementing these automatic procedures, AS2 would legitimately detect and filter prefix 10.0.0.0/24.



Figure 1: Basic scenario of local filtering

Local filtering could be required in other cases. For example, a dual homed AS receiving an overlapping prefix from only one of its providers. Figure 2 depicts a simple example of this case.

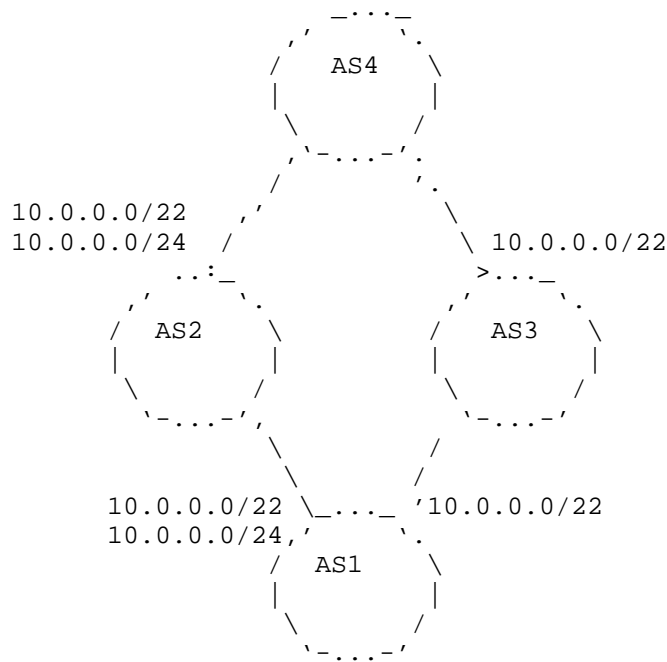


Figure 2: Basic scenario of local filtering

In this scenario, prefix 10.0.0.0/22 is advertised by AS1 to AS2 and AS3. Both ASes propagate the prefix to AS4. Additionally, AS1 advertises prefix 10.0.0.0/24 to AS2, which subsequently propagates the prefix to AS4.

It is possible that AS4 resolves to filter the more specific prefix 10.0.0.0/24. One potential motivation could be the economical preference of the path via AS2 over AS3. Another feasible reason is the existence of a technical policy by AS4 of aggregating incoming prefixes longer than /23.

The above examples illustrate two of the many motivations to configure routing within an AS with the aim of ignoring more specific prefixes. Operators have reported applying these filters in a manual fashion [3]. The relevance of such practice led to investigate automated filtering procedures in I-D.WHITE [2].

## 2.2. Remotely triggered filtering

ISPs can tag the BGP paths that they propagate to neighboring ASes with communities, in order to tweak the propagation behavior of the ASes that handle these paths [1].

Some ISPs allow their direct and indirect customers to use such communities to let the receiving AS not export the path to some selected neighboring AS. By combining communities, the prefix could be advertised only to a given peer of the AS providing this feature. Figure 3 illustrates an example of this case.

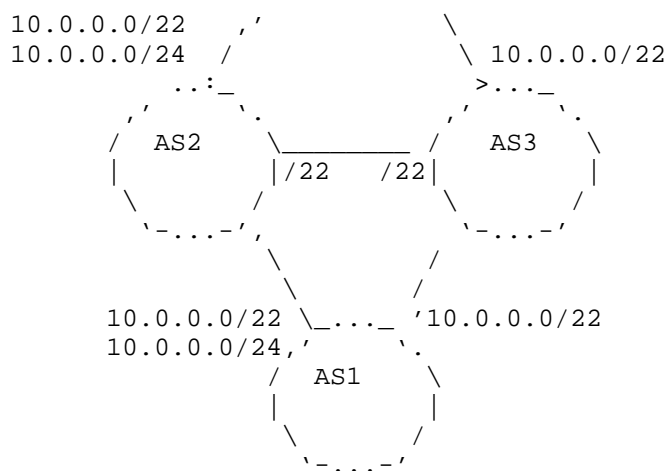


Figure 3: Remote triggered filtering

AS2 and AS3 are peers. Both ASes are providers of AS1. For traffic engineering purposes, AS1 could use communities to prevent AS2 from announcing prefix 10.0.0.0/24 to AS3.

Such technique is useful for operators to tweak routing decisions in order to align with complex transit policies. We will see in later sections that by producing the same effect as filtering, they can also lead to unexpected traffic flows at other, distant, ASes.

## 3. Uses of overlapping prefix filtering that create unexpected traffic flows

In this section, we define the concept of unexpected traffic flows and describe three configuration scenarios that lead to their creation. Note that these examples do not capture all the cases where such issues can take place.



### 3.1. Unexpected traffic Flows

The BGP policy of an Internet Service provider includes all actions performed over its originated routes and the routes received externally. One important part of the BGP policy is the selection of the routes that are propagated to each neighboring AS. One of the goals of these policies is to allow ISPs to avoid transporting traffic between two ASes without economical gain. For instance, ISPs typically propagate to their peers only routes coming from its customers (RFC4384 [3]). We briefly illustrate this operation in Figure 4. In the figure, AS2 is establishing a settlement free peering with AS1 and AS3. AS2 receives prefix P3/p3, from AS3. AS2, however, is not interested in transporting traffic from AS1 to AS3, therefore it does not propagate the prefix to AS1. In the figure, we also show a customer of AS2, AS4, which is announcing prefix P4/p4. AS2 propagates this prefix to AS1.

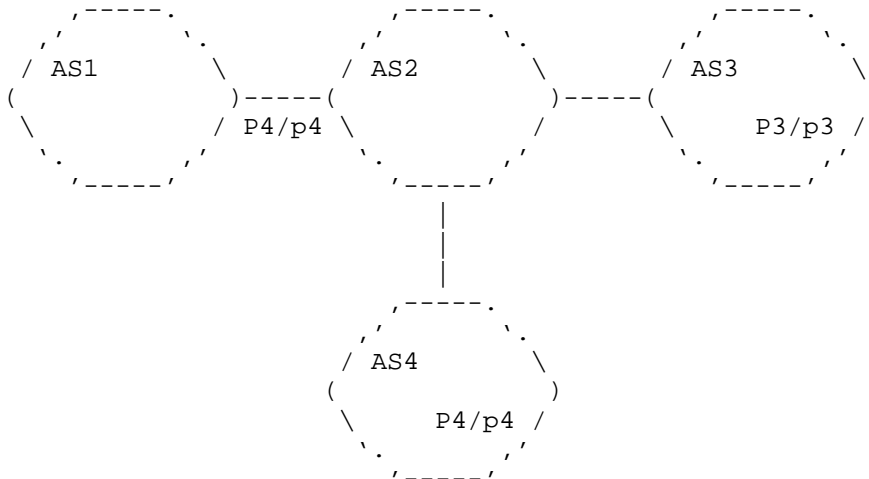


Figure 4: Prefix exchange among four autonomous systems

Although ISPs usually implement the aforementioned policies, unexpected traffic flows may still appear. In Figure 4, unexpected traffic flows are created, when, despite AS2's policy, traffic arriving from peer AS1 is received and transported to AS3 by AS2. These types of traffic flows can arise due to a number of reasons. Specifically, in this document we explain how the filtering of overlapping prefixes might cause unexpected traffic flows on ASes. We provide examples of these cases in the next sections.

### 3.1.1.1. Unexpected traffic flows caused by local filtering of overlapping prefixes

In this section, we describe cases in which an AS locally filters an overlapping prefix. We show that, depending on the BGP policies applied by surrounding ASes, this decision can lead to unexpected traffic flows.

#### 3.1.1.1.1. Initial setup

We start by describing the basic scenario of this case in Figure 5.

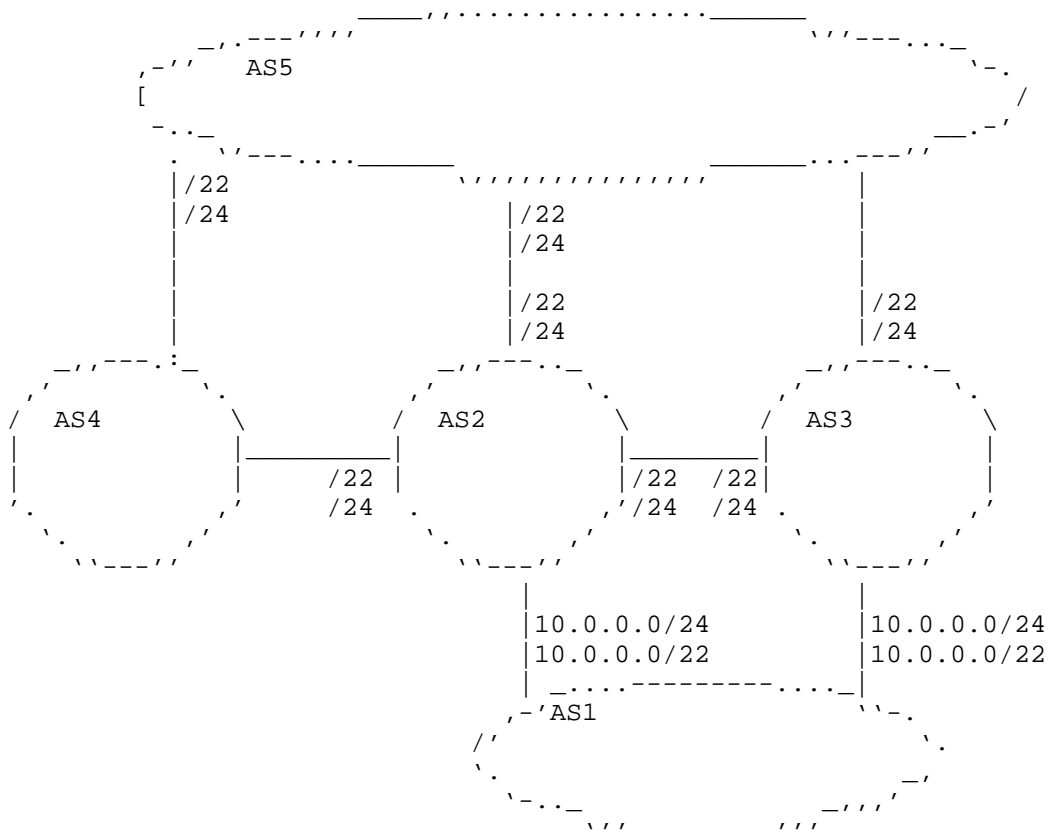


Figure 5: Initial Setup Local

AS1 is a customer of AS2 and AS3. AS2, AS3, and AS4 are customers of AS5. AS2 is establishing a peering with AS3 and AS4. AS1 is announcing a covering prefix, 10.0.0.0/22, and an overlapping prefix

10.0.0.0/24 to its providers. In the initial setup, AS2 and AS3 announce the two prefixes to their peers and transit providers. AS4 receives both prefixes from its peer (AS2) and transit provider (AS5). We will consider that AS5 chooses as best path to AS1 the one received from AS3.

#### 3.1.1.2. Unexpected traffic flows by local filtering - Case 1

In the next scenarios, we show that if AS4 filters the incoming overlapping prefix from AS5, there is a situation in which unexpected traffic flows are created on other ASes.

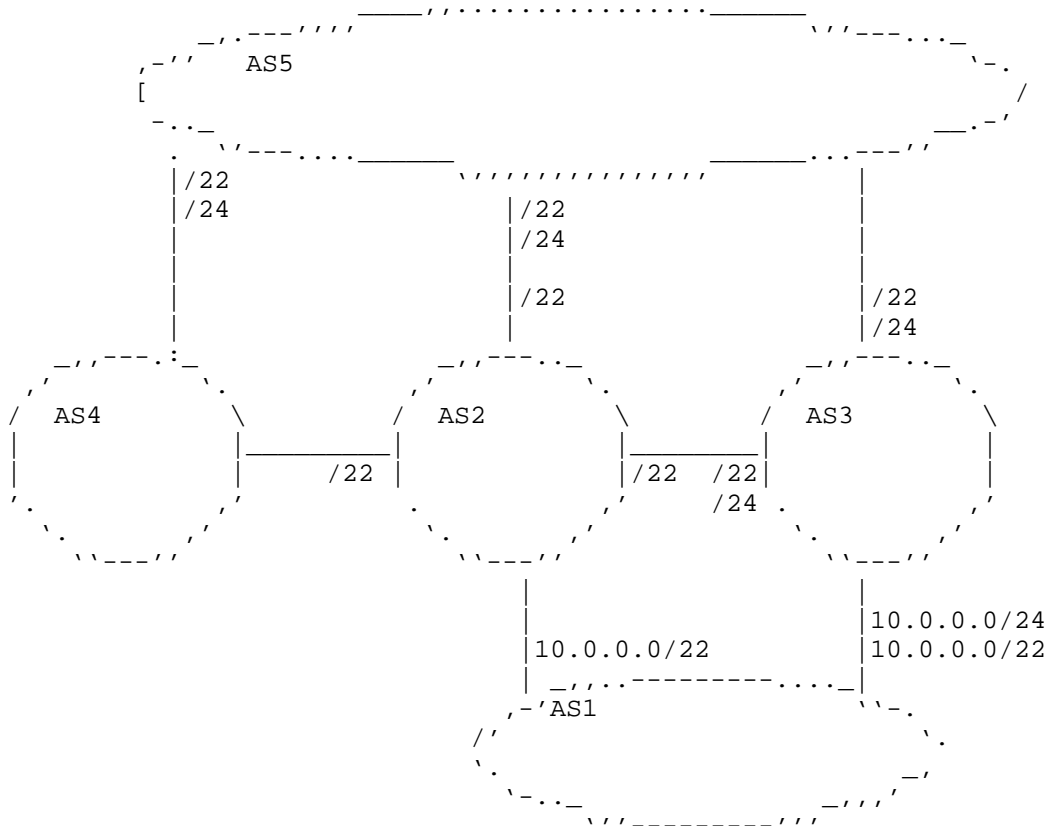


Figure 6: Unexpected traffic flows by local filtering - Case 1

Let us assume the scenario illustrated in Figure 6. For this case, AS1 only propagates the overlapping prefix to AS3. AS4 receives the overlapping prefix only from its transit provider, AS5.

AS4 now is in a situation in which it would be favorable for it to filter the announcement of prefix 10.0.0.0/24 from AS5. Subsequently, traffic from AS4 to prefix 10.0.0.0/24 is forwarded towards AS2. Because AS2 receives the more specific prefix from AS3, traffic from AS4 to prefix 10.0.0.0/24 follows the path AS4-AS2-AS3-AS1. AS2's BGP policies are implemented to avoid using itself to exchange traffic between AS4 and AS3. However, due to the discrepancies of routes from the overlapping and covering prefixes, unexpected traffic flows between AS4 and AS3 still exist on AS2's network. This situation is economically detrimental for AS2, since it forwards traffic from a peer to a non-customer neighbor.

#### 3.1.1.3. Unexpected traffic flows by local filtering - Case 2

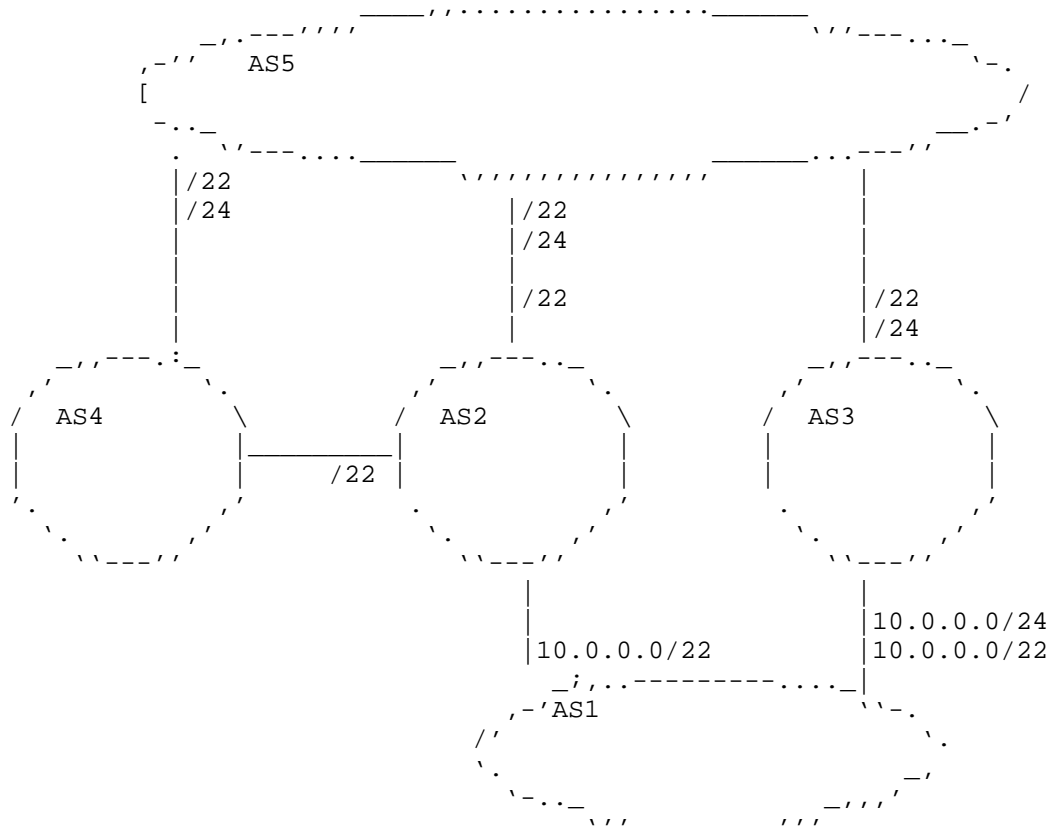


Figure 7: Unexpected traffic flows after local filtering - Case 2

Let us assume a second case where AS2 and AS3 are not peering and AS1 only propagates the overlapping prefix to AS3. AS4 receives the overlapping prefix only from its transit provider, AS5. This case is illustrated in Figure 7.

Similar to the scenario described in Section 3.1.1.2, AS4 is in a situation in which it would be favorable to filter the announcement of prefix 10.0.0.0/24 from AS5. Subsequently, traffic from AS4 to prefix 10.0.0.0/24 would be forwarded towards AS2. Due to the existence of a route to prefix 10.0.0.0/24, AS2 receives the traffic heading to this prefix from AS4 and sends it to AS5. This situation creates unexpected traffic flows that contradict AS2's BGP policy,

since the AS ends up forwarding traffic from a peer to a transit network.

### 3.1.2. Unexpected traffic flows caused by remotely triggered filtering of overlapping prefixes

We present a configuration scenario in which an AS, using the mechanism described in Section 2.2, informs its provider to selectively propagate an overlapping prefix, leading to the creation of unexpected traffic flows in another AS.

#### 3.1.2.1. Initial setup

Let AS1 be a customer of AS2 and AS3. AS1 owns 10.0.0.0/22, which it advertises through AS2 and AS3. Additionally, AS2 and AS3 are peers.

Both AS2 and AS3 select AS1's path as best, and propagate it to their customers, providers, and peers. Some remote ASes will route traffic destined to 10.0.0.1 through AS2 while others will route traffic through AS3.

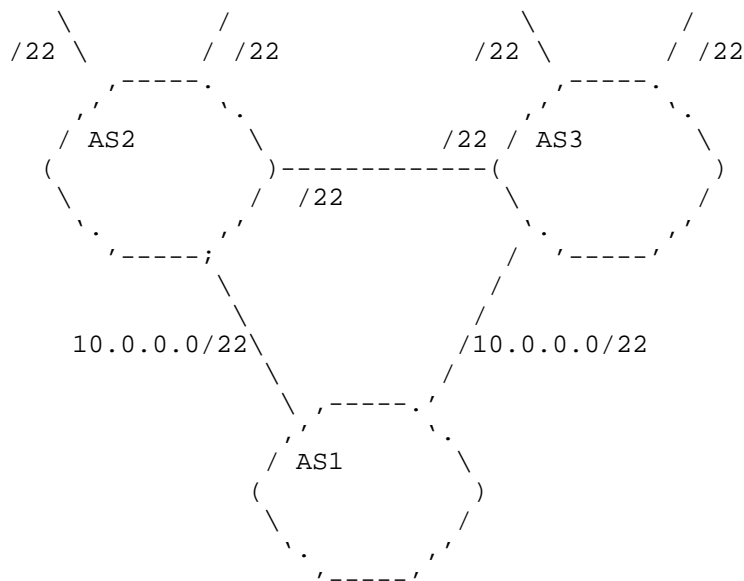


Figure 8: Example scenario

### 3.1.2.2. Injection of an overlapping prefix

Let AS1 advertise 10.0.0.0/24 over AS3 only. AS3 would propagate this prefix to its customers, providers, and peers, including AS2.

From AS2's point of view, the path towards 10.0.0.0/24 is a "peer path" and AS2 will only advertise it to its customers. ASes in the customer branch of AS2 will receive a path to the /24 that contains AS3 and AS2. Some multi-homed customers of AS2 may also receive a path through AS3, but not through AS2, from other peering or provider links. Any remote AS that is not lying in the customer branch of AS2, will receive a path for 10.0.0.0/24 through AS3 and not through AS2.

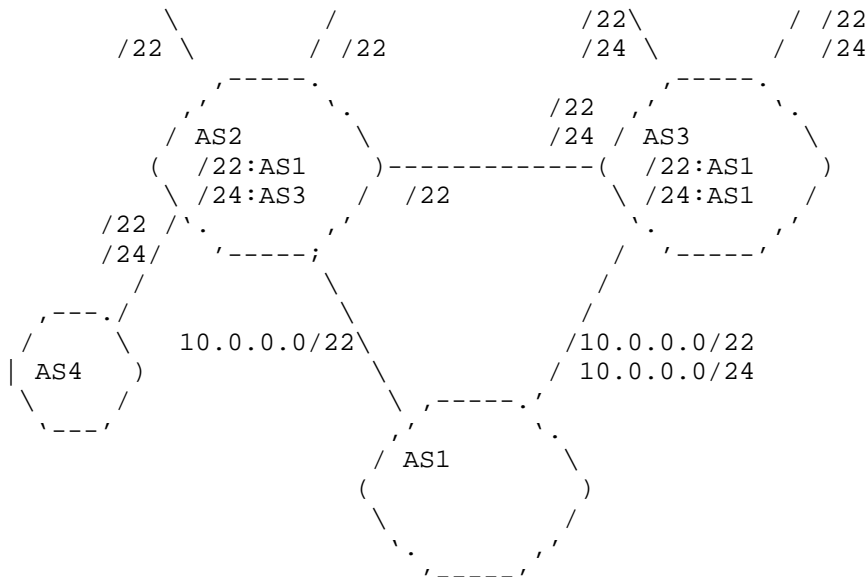


Figure 9: Injection of overlapping prefix

AS2 only receives traffic destined to 10.0.0.0/24 from its customers, which it forwards to its peer AS3. Routing is consistent with usual Internet Routing Policies in this case. AS3 could receive traffic destined to 10.0.0.0/24 from its customers, providers, and peers, which it directly forwards to its customer AS1.

### 3.1.2.3. Creation of unexpected traffic flows by limiting the scope of the overlapping prefix

Now, let us assume that 10.0.0.0/24, which is propagated by AS1 to AS3, is tagged to have AS3 only propagate that path to AS2, using the techniques described in Section 2.2.

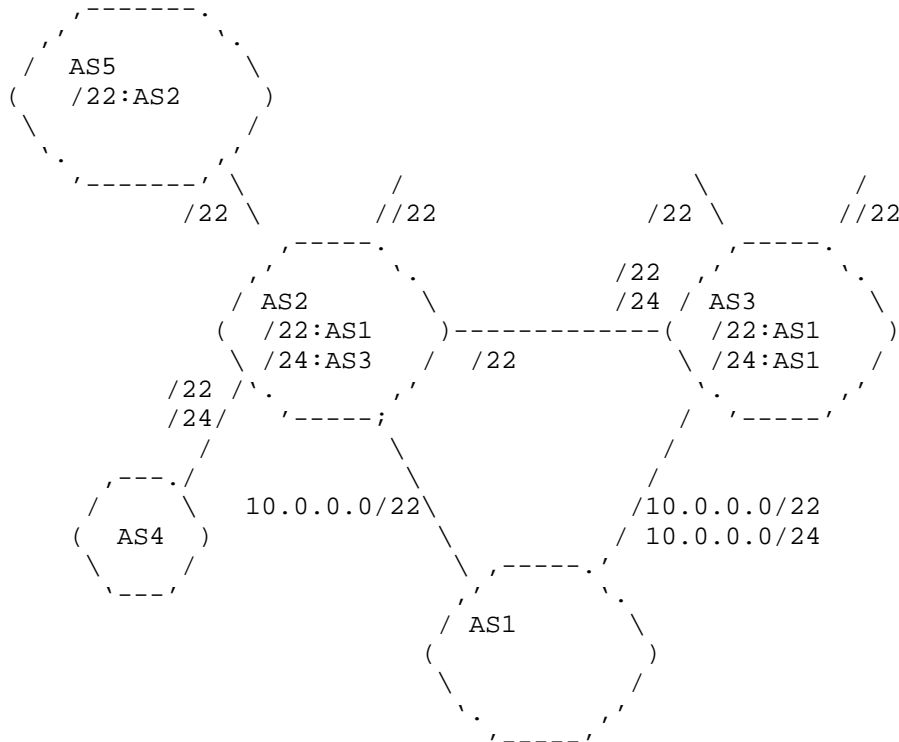


Figure 10: More Specific Injection

From AS2's point of view, such a path is a "peer path" and will only be advertised by AS2 to its customers.

ASes that are not customers of AS2 will not receive a path for 10.0.0.0/24. These ASes will forward packets destined to 10.0.0.0/24 according to their routing state for 10.0.0.0/22. Let us assume that AS5 is such an AS, and that its best path towards 10.0.0.0/22 is through AS2. Then, packets sent towards 10.0.0.1 by AS5 will eventually reach AS2. However, in the data-plane of the nodes of AS2, the longest prefix match for 10.0.0.1 is 10.0.0.0/24, which is reached through AS3, a peer of AS2. Since AS5 is not in the customer



branch of AS2, we are in a situation in which traffic flows between non-customer ASes take place in AS2.

#### 4. Techniques to detect unexpected traffic flows caused by filtering of overlapping prefixes

We differentiate the techniques available for detecting unexpected traffic flows caused by the described scenarios from the cases in which the interested AS is the victim or contributor of such operations.

##### 4.1. Being the 'victim' of unexpected traffic flows

To detect if unexpected traffic flows are taking place in its network, an ISP can monitor its traffic data and validate if any flow entering the ISP network through a non-customer link is forwarded to a non-customer next-hop.

As mentioned in Section 3.1, unexpected traffic flows might appear due to different situations. To discover if the problem arose after the filtering of prefixes by neighboring ASes, an operator can analyze available BGP data. For instance, an ISP can seek for overlapping prefixes for which the next-hop is through a provider (or peer), while the next-hop for their covering prefix(es) is through a client. Direct communication or looking glasses can be used to check whether non-customer neighboring ASes are propagating a path towards the covering prefix and not the path towards the overlapping prefix. This situation should trigger a warning, as this would mean that ASes in the surrounding area of the current AS are forwarding packets based on the routing entry for the less specific prefix only.

##### 4.2. Being a contributor to the existence of unexpected traffic flows in other networks

It can be considered problematic to be causing unexpected traffic flows on other ASes. This situation may appear as an abuse to the network resources of other ISPs.

There may be justifiable reasons for one ISP to perform filtering, either to enforce established policies or to provide prefix advertisement scoping features to its customers. These can vary from trouble-shooting purposes to business relationships implementations. Restricting such features for the sake of avoiding the creation of unexpected traffic flows is not a practical option.

Traffic data does not help an ISP detect that it is acting as a contributor of the creation of the unexpected traffic flow. It is thus advisable to obtain as much information as possible about the

Internet environment of the AS and assess the risks of filtering overlapping prefixes before implementing them.

Monitoring the manipulation of the communities that implement the scoping of prefixes is recommended to the ISPs that provide these features. The monitored behavior should then be compared with their terms of use.

5. Techniques to counter unexpected traffic flows due to the filtering of overlapping prefixes

Network Operators can adopt different approaches with respect to unexpected traffic flows. We classify these actions according to whether they are anticipant or reactive.

Reactive approaches are those in which the operator tries to detect the situations via monitoring and solve unexpected traffic flows, manually, on a case-by-case basis.

Anticipant or preventive approaches are those in which the routing system will not let the unexpected traffic flows actually take place when the configuration scenario is set up.

We use the scenario depicted in Figure 11 to describe these two kinds of approaches. Based on our analysis, we observe that anticipant approaches can be complex to implement and can lead to undesired repercussions. Therefore, we conclude that the reactive approach is the more reasonable recommendation to deal with unexpected flows.

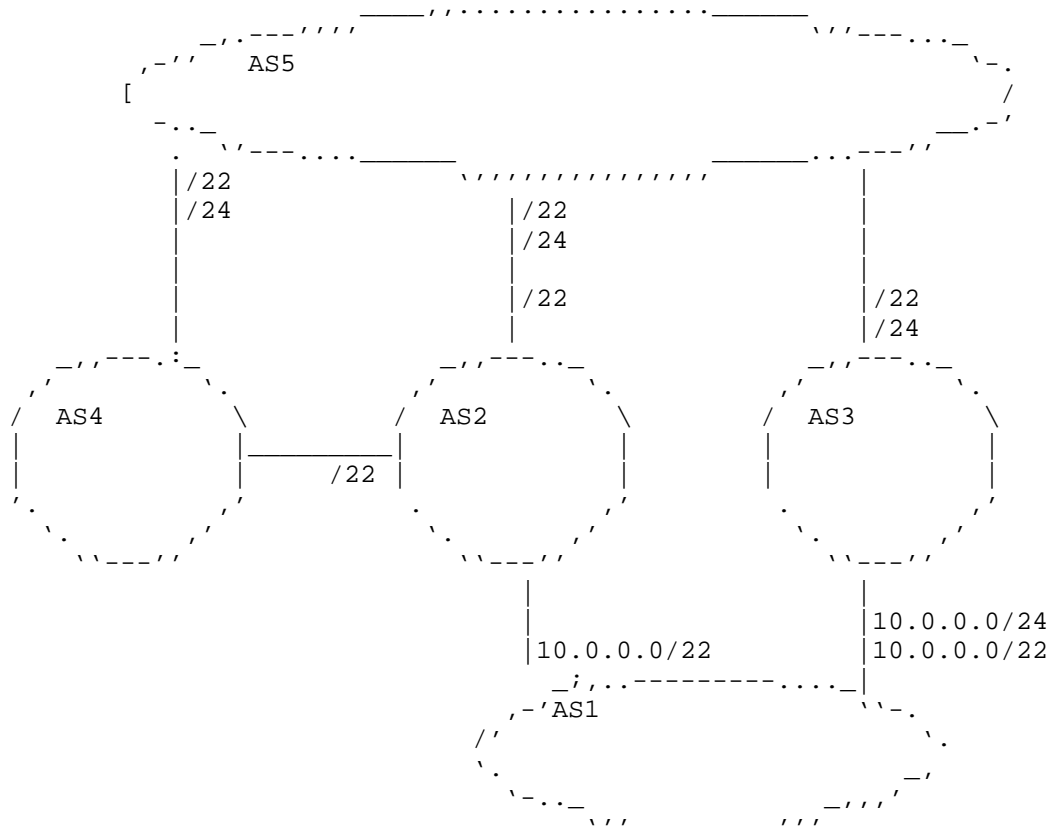


Figure 11: Anticipant counter-measures - Base example

### 5.1. Reactive counter-measures

An operator who detects unexpected traffic flows originated by any of the cases described in Section 3 can contact the ASes that are likely to have performed the propagation tweaks, inform them of the situation, and persuade them to change their behavior.

If the situation remains, the operator can implement prefix filtering in order to stop the unexpected flows. The operator can decide to perform this action over the session with the operator announcing the overlapping prefix or over the session with the neighboring AS from which it is receiving the traffic. Each of these options carry a different repercussion for the affected AS. We describe briefly the two alternatives.

- o An operator can decide to stop announcing the covering prefix at the peering session with the neighboring AS from which it is receiving traffic to the overlapping prefix. In the example of Figure 11, AS2 would filter out the prefix 10.0.0.0/22 from the eBGP session with AS4. In this case, all the traffic heading to the prefix 10.0.0.0/22 from AS1 would not longer traverse AS2. AS2 should evaluate if solving the inconvenient originated by the unexpected traffic flows are worth the loss of this traffic share.
- o An operator can decide to filter-out the concerned overlapping prefix at the peering session over which it was received. In the example of Figure 11, AS2 would filter out the incoming prefix 10.0.0.0/24 from the eBGP session with AS5. As a result, the traffic destined to that /24 would be forwarded by AS2 along its link with AS1, despite the actions performed by AS1 to have this traffic coming in through its link with AS3. However, as AS2 will no longer possess a route to the overlapping prefix, it risks losing the traffic share from customers different from AS1 to that prefix. Furthermore, this action can generate conflicts between AS2 and AS1, since AS2 does not follow the policy expressed by AS1 in its BGP announcements.

It is possible that the behavior from the neighboring AS that is causing the unexpected traffic flows opposes the peering agreement. In this case, an operator can account the amount of traffic that has been subject to the unexpected flows and charge the peer for that traffic. That is, the operator can claim that it has been a provider of that peer for the traffic that transited between the two ASes.

## 5.2. Anticipant counter-measures

### 5.2.1. Access lists

An operator can configure its routers to install dynamically an access-list made of the prefixes towards which the forwarding of traffic from that interface would lead to unexpected traffic flows. In the example of Figure 11, AS2 would install an access-list denying packets matching 10.0.0.0/24 associated with the interface connecting to AS4. As a result, traffic destined to that prefix would be dropped, despite the existence of a valid route towards 10.0.0.0/22.

Note that this technique actually lets packets destined to a valid prefix be dropped while they are sent from a neighboring AS that cannot know about policy conflicts and hence had no means to avoid the creation of unexpected traffic flows.

### 5.2.2. Automatic overlapping prefix filtering

As described in Section 3, filtering of overlapping prefixes can in some scenarios lead to unexpected traffic flows. Nevertheless, depending on the autonomous system implementing such practice, this operation can prevent these cases. This can be illustrated using the example described in Figure 11: if AS2 or AS3 filter prefix 10.0.0.0/24, there would be no unexpected traffic flow in AS2. Nevertheless, as described in Section 5.1, the filtering of overlapping prefixes can generate conflicts between AS1 and AS2, since AS2 would not forward traffic according to AS1's policy. Additionally, AS2 can lose traffic share for the overlapping prefix from customers different from AS1.

### 5.2.3. Neighbor-specific forwarding

An operator can technically ensure that traffic destined to a given prefix will be forwarded from an entry point of the network based only on the set of paths that have been advertised over that entry point.

As an example, let us analyze the scenario of Figure 11 from the point of view of AS2. The edge router connecting to the AS4 forward packets destined to prefix 10.0.0.0/24 towards AS5. Likewise, it will forward packets destined to prefix 10.0.0.0/22 towards AS1. The router, however, only propagates the path of the covering prefix (10.0.0.0/22) to AS4. An operator could implement the necessary techniques to force the edge router to forward packets coming from AS4 based only on the paths propagated to AS4. Thus, the edge router would forward packets destined to 10.0.0.0/24 towards AS1 in which case no unexpected traffic flow would occur.

Different techniques could provide the functionality just described; however, their technical implementation can be complex to design and operate. [2] describes an approach to implement this behavior. Similar to the solution described in Section 5.2.2, this approach could create conflicts between AS2 and AS1, since the traffic forwarding performed by A2 goes against the policy of AS1.

## 6. Conclusions

In this document, we described threats to policies of autonomous systems caused by the filtering of overlapping prefixes performed by external networks. We provide examples of scenarios in which unexpected traffic flows are caused by these practices and introduce some techniques for their detection and prevention. Analyzing the different options for dealing with this kind of problems, we recommend potential victims to implement monitoring systems that can

detect them and react to them according to the specific situation. Although we observe that there are reasonable situations in which ASes could filter overlapping prefixes, we encourage that network operators implement this type of filters only after considering the cases described in this document.

## 7. References

- [1] Donnet, B. and O. Bonaventure, "On BGP Communities", ACM SIGCOMM Computer Communication Review vol. 38, no. 2, pp. 55-59, April 2008.
- [2] Vanbever, L., Francois, P., Bonaventure, O., and J. Rexford, "Customized BGP Route Selection Using BGP/MPLS VPNs", Cisco Systems, Routing Symposium <http://www.cs.princeton.edu/~jrex/talks/cisconag09.pdf>, October 2009.
- [3] "INIT7-RIPE63", <<http://ripe63.ripe.net/presentations/48-How-more-specifics-increase-your-transit-bill-v0.2.pdf>>.

### 7.2. URIs

- [1] <http://www.ietf.org/rfc/rfc1812.txt>
- [2] <http://tools.ietf.org/html/draft-white-grow-overlapping-routes-02>
- [3] <http://www.ietf.org/rfc/rfc4384.txt>

### Authors' Addresses

Camilo Cardona  
IMDEA Networks/UC3M  
Avenida del Mar Mediterraneo, 22  
Leganes 28919  
Spain  
  
Email: [juancamilo.cardona@imdea.org](mailto:juancamilo.cardona@imdea.org)

Pierre Francois  
IMDEA Networks  
Avenida del Mar Mediterraneo, 22  
Leganes 28919  
Spain  
  
Email: [pierre.francois@imdea.org](mailto:pierre.francois@imdea.org)

Paolo Lucente  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [plucente@cisco.com](mailto:plucente@cisco.com)

Internet Engineering Task Force  
Internet-Draft  
Updates: 4271 (if approved)  
Intended status: Standards Track  
Expires: January 7, 2016

W. George  
Time Warner Cable  
S. Amante  
Apple, Inc.  
July 6, 2015

Autonomous System Migration Mechanisms and Their Effects on the BGP  
AS\_PATH Attribute  
draft-ietf-idr-as-migration-06

Abstract

This draft discusses some existing commonly-used BGP mechanisms for ASN migration that are not formally part of the BGP4 protocol specification. It is necessary to document these de facto standards to ensure that they are properly supported in future BGP protocol work.

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 7, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of



the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
1.2. Documentation note . . . . .	3
2. ASN Migration Scenario Overview . . . . .	3
3. External BGP Autonomous System Migration Mechanisms . . . . .	5
3.1. Modify Inbound BGP AS_PATH Attribute . . . . .	5
3.2. Modify Outbound BGP AS_PATH Attribute . . . . .	7
3.3. Implementation . . . . .	8
4. Internal BGP Autonomous System Migration Mechanisms . . . . .	9
4.1. Internal BGP AS Migration . . . . .	10
4.2. Implementation . . . . .	12
5. Additional Operational Considerations . . . . .	13
6. IANA Considerations . . . . .	14
7. Security Considerations . . . . .	14
8. Acknowledgements . . . . .	14
9. References . . . . .	14
9.1. Normative References . . . . .	14
9.2. Informative References . . . . .	15
Appendix A. Implementation report . . . . .	15
Authors' Addresses . . . . .	16

## 1. Introduction

This draft discusses some existing commonly-used BGP mechanisms for Autonomous System Number (ASN) migration that are not formally part of the BGP4 [RFC4271] protocol specification. These mechanisms are local to a given BGP Speaker and do not require negotiation with or cooperation of BGP neighbors. The deployment of these mechanisms do not need to interwork with one another to accomplish the desired results, so slight variations between existing vendor implementations exist, and will not necessarily be harmonized due to this document. However, it is necessary to document these de facto standards to ensure that new implementations can be successful, and any future protocol enhancements to BGP that propose to read, copy, manipulate or compare the AS\_PATH attribute can do so without inhibiting the use of these very widely used ASN migration mechanisms.

The migration mechanisms discussed here are useful to ISPs and organizations of all sizes, but it is important to understand the business need for these mechanisms and illustrate why they are so critical for ISPs' operations. During a merger, acquisition or divestiture involving two organizations it is necessary to seamlessly migrate both internal and external BGP speakers from one ASN to a

second ASN. The overall goal in doing so is to simplify operations through consistent configurations across all BGP speakers in the combined network. In addition, given that the BGP Path Selection algorithm selects routes with the shortest AS\_PATH attribute, it is critical that the ISP does not increase AS\_PATH length during or after ASN migration, because an increased AS\_PATH length would likely result in sudden, undesirable changes in traffic patterns in the network.

By default, the BGP protocol requires an operator to configure a router to use a single remote ASN for the BGP neighbor, and the ASN must match on both ends of the peering in order to successfully negotiate and establish a BGP session. Prior to the existence of these migration mechanisms, it would have required an ISP to coordinate an ASN change with, in some cases, tens of thousands of customers. In particular, as each router is migrated to the new ASN, to avoid an outage due to ASN mismatch, the ISP would have to force all customers on that router to change their router configurations to use the new ASN immediately after the ASN change. Thus, it becomes critical to allow the ISP to make this process a bit more asymmetric, so that it could seamlessly migrate the ASN within its network(s), but allow the customers to gradually migrate to the ISP's new ASN at their leisure, either by coordinating individual reconfigurations, or accepting sessions using either the old or new ASN to allow for truly asymmetric migration.

#### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

#### 1.2. Documentation note

This draft uses Autonomous System Numbers (ASNs) from the range reserved for documentation as described in RFC 5398 [RFC5398]. In the examples used here, they are intended to represent Globally Unique ASNs, not private use ASNs as documented in RFC 6996 [RFC6996] section 5.

#### 2. ASN Migration Scenario Overview

The use case being discussed here is an ISP merging two or more ASNs, where eventually one ASN subsumes the other(s). In this use case, we will assume the most common case where there are two ISPs, A and B, that prior to the ASN migration use AS 64500 and 64510, respectively. AS 64500 will be the permanently retained ASN used across the consolidated set of both ISPs network equipment, and AS 64510 will be

retired. Thus, at the conclusion of the ASN migration, there will be a single ISP A' with all internal BGP speakers configured to use AS 64500. To all external BGP speakers, the AS\_PATH length will not be increased.

In this same scenario, AS 64496 and AS 64499 represent two separate customer networks: C and D, respectively. Originally, customer C (AS 64496) is attached to ISP B, which will undergo ASN migration from AS 64510 to AS 64500. Furthermore, customer D (AS 64499) is attached to ISP A, which does not undergo ASN migration since the ASN for ISP A will remain constant, (AS 64500). Although this example refers to AS 64496 and 64499 as customer networks, either or both may be settlement-free or other types of peers. In this use case they are referred to as "customers" merely for convenience.

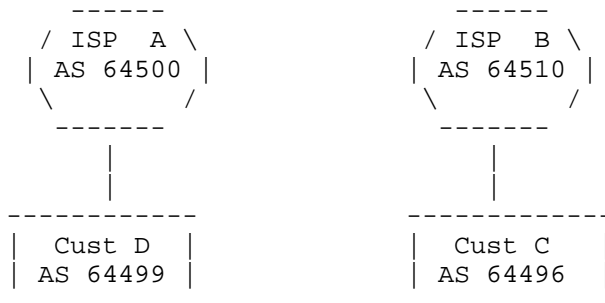


Figure 1: Before Migration

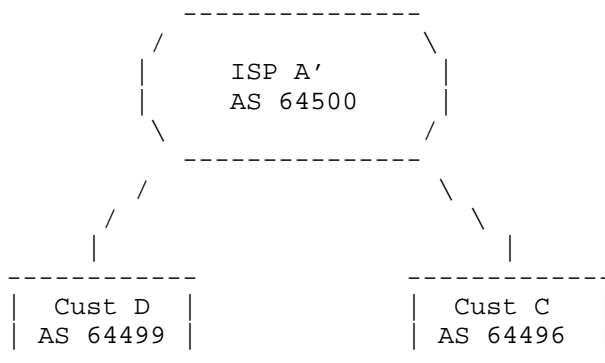


Figure 2: After Migration

The general order of operations, typically carried out in a single maintenance window by the network undergoing ASN migration (ISP B), are as follows. First, ISP B will change the global BGP ASN used by

a Provider Edge (PE) router, from ASN 64510 to 64500. At this point, the router will no longer be able to establish eBGP sessions toward the existing Customer Edge (CE) devices that are attached to it and still using AS 64510. Second, since ISP B needs to do this without coordinating the simultaneous change of its ASN with all of its eBGP peers, ISP B will configure two separate, but related ASN migration mechanisms discussed in this document on all eBGP sessions toward all CE devices. These mechanisms enable the router to establish BGP neighbors using the legacy ASN, modify the AS\_PATH attribute received from a CE device when advertising it further, and modify AS\_PATH when transmitted toward CE devices to achieve the desired effect of not increasing the length of the AS\_PATH.

At the conclusion of the ASN migration, the CE devices at the edge of the network are not aware of the fact that their upstream router is now in a new ASN and do not observe any change in the length of the AS\_PATH attribute. However, after the changes discussed in this document are put in place by ISP A', there is a change to the contents of the AS\_PATH attribute to ensure the AS\_PATH is not artificially lengthened while these AS migration parameters are used.

In this use case, neither ISP is using BGP Confederations RFC 5065 [RFC5065] internally.

### 3. External BGP Autonomous System Migration Mechanisms

The following section addresses optional capabilities that are specific to modifying the AS\_PATH attribute at the Autonomous System Border Routers (ASBRs) of an organization, (typically a single Service Provider). This ensures that external BGP customers/peers are not forced to make any configuration changes on their CE routers before or during the exact time the Service Provider wishes to migrate to a new, permanently retained ASN. Furthermore, these mechanisms eliminate the artificial lengthening of the AS\_PATH both transmitted from and received by the Service Provider that is undergoing AS Migration, which would have negative implications on path selection by external networks.

#### 3.1. Modify Inbound BGP AS\_PATH Attribute

The first instrument used in the process described above is called "Local AS". This allows the router to supersede the globally configured ASN in the "My Autonomous System" field of the BGP OPEN [RFC4271] with a locally defined AS value for a specific BGP neighbor or group of neighbors. This mechanism allows the PE router that was formerly in ISP B to establish an eBGP session toward the existing CE devices using the legacy AS, AS 64510. Ultimately, the CE devices (i.e.: customer C) are completely unaware that ISP B has reconfigured

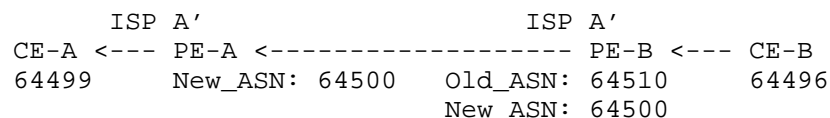
its router to participate as a member of a new AS. Within the context of the former ISP B PE router, the second effect this specific mechanism has on AS\_PATH is that, by default, it prepends all received BGP UPDATES with the legacy AS of ISP B: AS 64510, while advertising it (Adj-RIB-Out) to other BGP speakers (A'). Within the Loc-RIB on ISP B prior to the migration, the AS\_PATH of route announcements received from customer C would appear as: 64496, whereas the same RIB on ISP A' (ISP B routers post-migration) would contain AS\_PATH: 64510 64496.

A second instrument, referred to as "No Prepend Inbound", is enabled on PE routers migrating from ISP B. The "No Prepend Inbound" capability causes ISP B's routers to not prepend the legacy AS, AS 64510, when advertising UPDATES received from customer C. This restores the AS\_PATH within ISP A' for route announcements received from customer C so that it is just one ASN in length: 64496.

In the direction of CE -> PE (inbound):

1. "Local AS": Allows the local BGP router to generate a BGP OPEN to an eBGP neighbor with the old, legacy ASN value in the "My Autonomous System" field. When this capability is activated, it also causes the local router to prepend the <old\_ASN> value to the AS\_PATH when installing or advertising routes received from a CE to iBGP neighbors inside the Autonomous System.
2. "No Prepend Inbound (of Local AS)": the local BGP router does not prepend <old\_ASN> value to the AS\_PATH when installing or advertising routes received from the CE to iBGP neighbors inside the Autonomous System

PE-B is a PE that was originally in ISP B, and has a customer eBGP session to CE-B. PE-B has had its global configuration ASN changed from AS 64510 to AS 64500 to make it part of the permanently retained ASN. This now makes PE-B a member of ISP A'. PE-A is a PE that was originally in ISP A, and has a customer peer CE-A. Although its global configuration ASN remains AS 64500, throughout this exercise we also consider PE-A a member of ISP A'.



Note: Direction of BGP UPDATE as per the arrows.

Figure 3: Local AS and No Prepend BGP UPDATE Diagram

As a result using both the "Local AS" and "No Prepend Inbound" capabilities on PE-B, CE-A will see an AS\_PATH of: 64500 64496. CE-A will not receive a BGP UPDATE containing AS 64510 in the AS\_PATH. (If only the "Local AS" mechanism was configured without "No Prepend Inbound" on PE-B, then CE-A would have seen an AS\_PATH of: 64500 64510 64496, which results in an unacceptable lengthening of the AS\_PATH). NOTE: If there are still routers in the old ASN (64510), it is possible for them to accept these manipulated routes (i.e. those with 64510 removed from the AS\_PATH by this command) as if they have not already passed through their ASN, potentially causing a loop, since BGP's normal loop-prevention behavior of rejecting routes that include its ASN in the path will not catch these. Careful filtering between routers remaining in the old ASN and routers migrated to the new ASN is necessary to minimize the risk of routing loops.

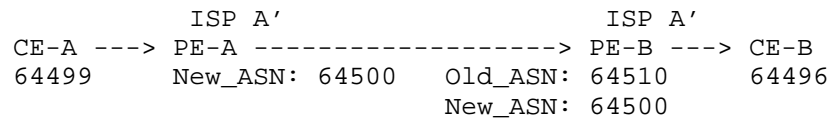
### 3.2. Modify Outbound BGP AS\_PATH Attribute

The two aforementioned mechanisms, "Local AS" and "No Prepend Inbound", only modify the AS\_PATH Attribute received by the ISP's PE's in the course of processing BGP UPDATES from CE devices when CE devices still have an eBGP session established with the ISPs legacy AS (AS64510).

In some existing implementations, "Local AS" and "No Prepend Inbound" does not concurrently modify the AS\_PATH Attribute for BGP UPDATES that are transmitted by the ISP's PE's to CE devices. In these implementations, with "Local AS" and "No Prepend Inbound" used on PE-B, it automatically causes a lengthening of the AS\_PATH in outbound BGP UPDATES from ISP A' toward directly attached eBGP speakers, (Customer C in AS 64496). The externally observed result is that customer C, in AS 64496, will receive the following AS\_PATH: 64510 64500 64499. Therefore, if ISP A' takes no further action, it will cause an unacceptable increase in AS\_PATH length within customer's networks directly attached to ISP A'.

A tertiary mechanism, referred to as "Replace Old AS", is used to resolve this problem. This capability allows ISP A' to prevent routers from appending the globally configured ASN in outbound BGP UPDATES toward directly attached eBGP neighbors that are using the "Local AS" mechanism. Instead, only the old (or previously used) AS will be prepended in the outbound BGP UPDATE toward the customer's network, restoring the AS\_PATH length to what it was before AS Migration occurred.

To re-use the above diagram, but in the opposite direction, we have:



Note: Direction of BGP UPDATE as per the arrows.

Figure 4: Replace AS BGP UPDATE Diagram

By default, without the use of "Replace Old AS", CE-B would see an AS\_PATH of: 64510 64500 64499. After ISP A' changes PE-B to use "Replace Old AS", CE-B would receive an AS\_PATH of: 64510 64499, which is the same AS\_PATH length pre-AS migration.

### 3.3. Implementation

The mechanisms introduced in this section MUST be configurable on a per-neighbor or per neighbor group (i.e. a group of similar BGP neighbor statements that reuse some common configuration to simplify provisioning) basis to allow for maximum flexibility. When the "Local AS" capability is used, a local ASN will be provided in the configuration that is different from the globally-configured ASN of the BGP router. To implement this mechanism, a BGP speaker SHOULD send BGP OPEN [RFC4271] (see section 4.2) messages to the configured eBGP peer(s) using the local ASN configured for this session as the value sent in "My Autonomous System". The BGP router SHOULD NOT use the ASN configured globally within the BGP process as the value sent in "My Autonomous System" in the OPEN message. This will avoid causing the eBGP neighbor to unnecessarily generate a BGP OPEN Error message "Bad Peer AS". This method is typically used to re-establish eBGP sessions with peers expecting the legacy ASN after a router has been moved to a new ASN.

Implementations MAY support a more flexible model where the eBGP speaker attempts to open the BGP session using either the ASN configured as "Local AS" or the globally configured AS as discussed in BGP Alias (Section 4.2). If the session is successfully established to the globally configured ASN, then the modifications to AS\_PATH described in this document SHOULD NOT be performed, as they are unnecessary. The benefit to this more flexible model is that it allows the remote neighbor to reconfigure to the new ASN without direct coordination between the ISP and the customer.

Note that this procedure will vary slightly if the locally or globally configured ASN is a 4-octet ASN. See section 3 of [RFC4893].

When the BGP router receives UPDATES from its eBGP neighbor configured with the "Local AS" mechanism, it processes the UPDATE as described in RFC4271 section 5.1.2 [RFC4271]. However the presence of a second ASN due to "Local AS" adds the following behavior to processing UPDATES received from an eBGP neighbor configured with this mechanism:

1. Internal: the router SHOULD append the configured "Local AS" ASN in the AS\_PATH attribute before installing the route or advertising the UPDATE to an iBGP neighbor. The decision of when to append the ASN is an implementation detail outside the scope of this document. Some considerations factoring into this decision include consistency in the AS\_PATH throughout the AS, and implementation of the loop detection mechanism.
2. External: the BGP router SHOULD first append the globally configured ASN to the AS\_PATH immediately followed by the "Local AS" value before advertising the UPDATE to an eBGP neighbor.

Two options exist to manipulate the behavior of the basic "Local AS" mechanism. They modify the behavior as described below:

1. "No Prepend Inbound" - When the BGP router receives inbound BGP UPDATES from its eBGP neighbor configured with this option, it MUST NOT append the "Local AS" ASN value in the AS\_PATH attribute when installing the route or advertising that UPDATE to iBGP neighbors, but it MUST still append the globally configured ASN as normal when advertising the UPDATE to other local eBGP neighbors (i.e. those natively peering with the globally configured ASN).
  2. "Replace Old AS", (outbound) - When the BGP router generates outbound BGP UPDATES toward an eBGP neighbor configured with this option, the BGP speaker MUST NOT append the globally configured ASN from the AS\_PATH attribute. The BGP router MUST append only the configured "Local AS" ASN value to the AS\_PATH attribute before sending the BGP UPDATES outbound to the eBGP neighbor.
4. Internal BGP Autonomous System Migration Mechanisms

The following section describes mechanisms that assist with a gradual and least service impacting migration of Internal BGP sessions from a legacy ASN to the permanently retained ASN. The following mechanism is very valuable to networks undergoing AS migration, but its use does not cause changes to the AS\_PATH attribute.



#### 4.1. Internal BGP AS Migration

In this case, all of the routers to be consolidated into a single, permanently retained ASN are under the administrative control of a single entity. Unfortunately, the traditional method of migrating all Internal BGP speakers, particularly within larger networks, is both time consuming and widely service impacting.

The traditional method to migrate Internal BGP sessions was strictly limited to reconfiguration of the global configuration ASN and, concurrently, changing all iBGP neighbors' remote ASN from the legacy ASN to the new, permanently retained ASN on each router within the legacy AS. These changes can be challenging to swiftly execute in networks with more than a few dozen internal BGP routers. There is also the concomitant service interruptions as these changes are made to routers within the network, resulting in a reset of iBGP sessions and subsequent route reconvergence to reestablish optimal routing paths. Operators often cannot make such sweeping changes given the associated risks of a highly visible service interruption; rather, they require a more gradual method to migrate Internal BGP sessions, from one ASN to a second, permanently retained ASN, that is not visibly service-impacting to its customers.

With the "Internal BGP AS Migration" mechanism described herein, it allows an Internal BGP speaker to form a single iBGP session using either the old, legacy ASN or the new, permanently retained ASN. The benefits of using this mechanism are several fold. First, it allows for a more gradual and less service-impacting migration away from the legacy ASN to the permanently retained ASN. Second, it (temporarily) permits the coexistence of the legacy and permanently retained ASN within a single network, allowing for uniform BGP path selection among all routers within the consolidated network.

The iBGP router with the "Internal BGP AS Migration" capability enabled allows the receipt of a BGP OPEN message with either the legacy ASN value or the new, globally configured ASN value in the "My Autonomous System" field of the BGP OPEN message from iBGP neighbors. It is important to recognize that enablement of the "Internal BGP AS Migration" mechanism preserves the semantics of a regular iBGP session (i.e. using identical ASNs). Thus, the BGP attributes transmitted by and the acceptable methods of operation on BGP attributes received from iBGP sessions configured with "Internal BGP AS Migration" capability are no different than those exchanged across an iBGP session without "Internal BGP AS Migration" configured, as defined by [RFC4271] and [RFC4456].

Typically, in medium to large networks, BGP Route Reflectors [RFC4456] (RRs) are used to aid in reduction of configuration of iBGP

sessions and scalability with respect to overall TCP (and, BGP) session maintenance between adjacent iBGP routers. Furthermore, BGP Route Reflectors are typically deployed in pairs within a single Route Reflection cluster to ensure high reliability of the BGP Control Plane. As such, the following example will use Route Reflectors to aid in understanding the use of the "Internal BGP AS Migration" mechanism. Note that Route Reflectors are not a prerequisite to enable "Internal BGP AS Migration" and this mechanism can be enabled independent of the use of Route Reflectors.

The general order of operations is as follows:

1. Within the legacy network, (the routers comprising the set of devices that still have a globally configured legacy ASN), one member of a redundant pair of RRs has its global configuration ASN changed to the permanently retained ASN. Concurrently, the "Internal BGP AS Migration" capability is enabled on all iBGP sessions on that device. This will comprise Non-Client iBGP sessions to other RRs as well as Client iBGP sessions, typically to PE devices, both still utilizing the legacy ASN. Note that during this step there will be a reset and reconvergence event on all iBGP sessions on the RRs whose configuration was modified; however, this should not be service impacting due to the use of redundant RRs in each RR Cluster.
2. The above step is repeated for the other side of the redundant pair of RRs. The one alteration to the above procedure is that the "Internal BGP AS Migration" mechanism is now removed from the Non-Client iBGP sessions toward the other (previously reconfigured) RRs, since it is no longer needed. The "Internal BGP AS Migration" mechanism is still required on all RRs for all RR Client iBGP sessions. Also during this step, there will be a reset and reconvergence event on all iBGP sessions whose configuration was modified, but this should not be service impacting. At the conclusion of this step, all RRs should now have their globally configured ASN set to the permanently retained ASN and "Internal BGP AS Migration" enabled and in use toward RR Clients.
3. At this point, the network administrators would then be able to establish iBGP sessions between all Route Reflectors in both the legacy and permanently retained networks. This would allow the network to appear to function, both internally and externally, as a single, consolidated network using the permanently retained network.
4. To complete the AS migration, each RR Client (PE) in the legacy network still utilizing the legacy ASN is now modified.

Specifically, each legacy PE would have its globally configured ASN changed to use the permanently retained ASN. The ASN configured within the PE for the iBGP sessions toward each RR would be changed to use the permanently retained ASN. It is unnecessary to enable "Internal BGP AS Migration" mechanism on these migrated iBGP sessions. During the same maintenance window, External BGP sessions would be modified to include the above "Local AS", "No Prepend" and "Replace Old AS" mechanisms described in Section 3 above, since all of the changes are service interrupting to the eBGP sessions of the PE. At this point, all PEs will have been migrated to the permanently retained ASN.

5. The final step is to excise the "Internal BGP AS Migration" configuration from the Router Reflectors in an orderly fashion. After this is complete, all routers in the network will be using the new, permanently retained ASN for all iBGP sessions with no vestiges of the legacy ASN on any iBGP sessions.

The benefit of using the aforementioned "Internal BGP AS Migration" capability is that it is a more gradual and less externally service-impacting change to accomplish an AS migration. Previously, without "Internal BGP AS Migration", such an AS migration change would carry a high risk and need to be successfully accomplished in a very short timeframe (e.g.: at most several hours). In addition, it would likely cause substantial routing churn and rapid fluctuations in traffic carried -- potentially causing periods of congestion and resultant packet loss -- during the period the configuration changes are underway to complete the AS Migration. On the other hand, with "Internal BGP AS Migration", the migration from the legacy ASN to the permanently retained ASN can occur over a period of days or weeks with reduced customer disruption. (The only observable service disruption should be when each PE undergoes the changes discussed in step 4 above.)

#### 4.2. Implementation

The mechanism introduced in this section MUST be configurable on a per-neighbor or per neighbor group basis to allow for maximum flexibility. When configured with this mechanism, a BGP speaker MUST accept BGP OPEN and establish an iBGP session from configured iBGP peers if the ASN value in "My Autonomous System" is either the globally configured ASN or a locally configured ASN provided when this capability is utilized. Additionally, a BGP router configured with this mechanism MUST send its own BGP OPEN [RFC4271] (see section 4.2) using either the globally configured or the locally configured ASN in "My Autonomous System" as follows. To avoid potential deadlocks when two BGP speakers are attempting to establish a BGP

peering session and are both configured with this mechanism, the speaker SHOULD send BGP OPEN using the globally configured ASN first, and only send a BGP OPEN using the locally configured ASN as a fallback if the remote neighbor responds with the BGP error "Bad Peer AS". In each case, the BGP speaker MUST treat UPDATES sent and received to this peer as if this was a natively configured iBGP session, as defined by [RFC4271] and [RFC4456].

Note that this procedure will vary slightly if the locally or globally configured ASN is a 4-octet ASN. See section 3 of [RFC4893].

## 5. Additional Operational Considerations

This document describes several mechanisms to support ISPs and other organizations that need to perform ASN migrations. Other variations of these mechanisms may exist, for example, in legacy router software that has not been upgraded or reached End of Life, but continues to operate in the network. Such variations are beyond the scope of this document.

Companies routinely go through periods of mergers, acquisitions and divestitures, which in the case of the former cause them to accumulate several legacy ASNs over time. ISPs often do not have control over the configuration of customers' devices (i.e.: the ISPs are often not providing a managed CE router service, particularly to medium and large customers that require eBGP). Furthermore, ISPs are using methods to perform ASN migration that do not require coordination with customers. Ultimately, this means there is not a finite period of time after which legacy ASNs will be completely expunged from the ISP's network. In fact, it is common that legacy ASNs and the associated External BGP AS Migration mechanisms discussed in this document can and do persist for several years, if not longer. Thus, it is prudent to plan that legacy ASNs and associated External BGP AS Migration mechanisms will persist in a operational network indefinitely.

With respect to the Internal BGP AS Migration mechanism, all of the routers to be consolidated into a single, permanently retained ASN are under the administrative control of a single entity. Thus, completing the migration from iBGP sessions using the legacy ASN to the permanently retained ASN is more straightforward and could be accomplished in a matter of days to months. Finally, good operational hygiene would dictate that it is good practice to avoid using "Internal BGP AS Migration" capability over a long period of time for reasons of not only operational simplicity of the network, but also reduced reliance on that mechanism during the ongoing

lifecycle management of software, features and configurations that are maintained on the network.

## 6. IANA Considerations

This memo includes no request to IANA.

## 7. Security Considerations

This draft discusses a process by which one ASN is migrated into and subsumed by another. This involves manipulating the AS\_PATH Attribute with the intent of not increasing the AS\_PATH length, which would typically cause the BGP route to no longer be selected by BGP's Path Selection Algorithm in others' networks. This could result in sudden and unexpected shifts in traffic patterns in the network, potentially resulting in congestion.

Given that these mechanisms can only be enabled through configuration of routers within a single network, standard security measures should be taken to restrict access to the management interface(s) of routers that implement these mechanisms. Additionally, BGP sessions SHOULD be protected using TCP Authentication Option [RFC5925] and the Generalized TTL Security Mechanism [RFC5082]

## 8. Acknowledgements

Thanks to Kotikalapudi Sriram, Stephane Litkowski, Terry Manderson, David Farmer, Jaroslaw Adam Gralak, Gunter Van de Velde, Juan Alcaide, Jon Mitchell, Thomas Morin, Alia Atlas, Alvaro Retana, and John Scudder for their comments.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

## 9.2. Informative References

- [ALU] Alcatel-Lucent, "BGP Local AS attribute", 2006-2012, <[https://infoproducts.alcatel-lucent.com/html/0\\_add-h-f/93-0074-10-01/7750\\_SR\\_OS\\_Routing\\_Protocols\\_Guide/BGP-CLI.html#709567](https://infoproducts.alcatel-lucent.com/html/0_add-h-f/93-0074-10-01/7750_SR_OS_Routing_Protocols_Guide/BGP-CLI.html#709567)>.
- [CISCO] Cisco Systems, Inc., "BGP Support for Dual AS Configuration for Network AS Migrations", 2003, <[http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/iproute\\_bgp/configuration/xe-3s/asr1000/irg-xe-3s-asr1000-book/irg-dual-as.html](http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/iproute_bgp/configuration/xe-3s/asr1000/irg-xe-3s-asr1000-book/irg-dual-as.html)>.
- [JUNIPER] Juniper Networks, Inc., "Configuring the BGP Local Autonomous System Attribute", 2012, <[http://www.juniper.net/techpubs/en\\_US/junos13.3/topics/concept/bgp-local-as-introduction.html](http://www.juniper.net/techpubs/en_US/junos13.3/topics/concept/bgp-local-as-introduction.html)>.
- [RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.
- [RFC5398] Huston, G., "Autonomous System (AS) Number Reservation for Documentation Use", RFC 5398, December 2008.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.
- [RFC6996] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", BCP 6, RFC 6996, July 2013.

## Appendix A. Implementation report

As noted elsewhere in this document, this set of migration mechanisms has multiple existing implementations in wide use.

- o Cisco [CISCO]
- o Juniper [JUNIPER]
- o Alcatel-Lucent [ALU]

This is not intended to be an exhaustive list, as equivalent features do exist in other implementations, however the authors were unable to find publicly available documentation of the vendor-specific implementation to reference.

#### Authors' Addresses

Wesley George  
Time Warner Cable  
13820 Sunrise Valley Drive  
Herndon, VA 20171  
US

Phone: +1 703-561-2540  
Email: wesley.george@twcable.com

Shane Amante  
Apple, Inc.  
1 Infinite Loop  
Cupertino, CA 95014  
US

Email: samante@apple.com

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: August 10, 2014

P. Lapukhov  
Facebook  
A. Premji  
Arista Networks  
J. Mitchell, Ed.  
Microsoft Corporation  
February 6, 2014

Use of BGP for routing in large-scale data centers  
draft-lapukhov-bgp-routing-large-dc-07

Abstract

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

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 10, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.



This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Network Design Requirements . . . . .	4
2.1. Bandwidth and Traffic Patterns . . . . .	4
2.2. CAPEX Minimization . . . . .	4
2.3. OPEX Minimization . . . . .	5
2.4. Traffic Engineering . . . . .	5
2.5. Summarized Requirements . . . . .	5
3. Data Center Topologies Overview . . . . .	6
3.1. Traditional DC Topology . . . . .	6
3.2. Clos Network topology . . . . .	7
3.2.1. Overview . . . . .	7
3.2.2. Clos Topology Properties . . . . .	8
3.2.3. Scaling the Clos topology . . . . .	9
3.2.4. Managing the Size of Clos Topology Tiers . . . . .	10
4. Data Center Routing Overview . . . . .	10
4.1. Layer 2 Only Designs . . . . .	11
4.2. Hybrid L2/L3 Designs . . . . .	11
4.3. Layer 3 Only Designs . . . . .	12
5. Routing Protocol Selection and Design . . . . .	12
5.1. Choosing EBGp as the Routing Protocol . . . . .	13
5.2. EBGp Configuration for Clos topology . . . . .	14
5.2.1. Example ASN Scheme . . . . .	14
5.2.2. Private Use BGP ASNs . . . . .	15
5.2.3. Prefix Advertisement . . . . .	16
5.2.4. External Connectivity . . . . .	17
5.2.5. Route Summarization at the Edge . . . . .	18
6. ECMP Considerations . . . . .	19
6.1. Basic ECMP . . . . .	19
6.2. BGP ECMP over Multiple ASNs . . . . .	20
6.3. Weighted ECMP . . . . .	20
6.4. Consistent Hashing . . . . .	21
7. Routing Convergence Properties . . . . .	21
7.1. Fault Detection Timing . . . . .	21
7.2. Event Propagation Timing . . . . .	22
7.3. Impact of Clos Topology Fan-outs . . . . .	22
7.4. Failure Impact Scope . . . . .	23

7.5. Routing Micro-Loops . . . . .	24
8. Additional Options for Design . . . . .	25
8.1. Third-party Route Injection . . . . .	25
8.2. Route Summarization within Clos Topology . . . . .	25
8.2.1. Collapsing Tier-1 Devices Layer . . . . .	25
8.2.2. Simple Virtual Aggregation . . . . .	27
8.3. ICMP Unreachable Message Masquerading . . . . .	27
9. Security Considerations . . . . .	28
10. IANA Considerations . . . . .	28
11. Acknowledgements . . . . .	28
12. References . . . . .	28
12.1. Normative References . . . . .	28
12.2. Informative References . . . . .	29
Authors' Addresses . . . . .	31

## 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.

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

This document first presents an overview of network design requirements and considerations for large-scale data centers. Then traditional hierarchical data center network topologies are contrasted with Clos networks that are horizontally scaled out. 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, the document reviews some additional considerations and design options.

## 2. Network Design Requirements

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

### 2.1. Bandwidth and Traffic Patterns

The primary requirement when building an interconnection network for large number of servers is to accommodate application bandwidth and latency requirements. Until recently it was quite common to see the majority of traffic entering and leaving the data center, commonly referred to as "north-south" traffic. As a result, 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 line-cards or fabrics or replacing the device with one with higher port density.

Today many large-scale data centers host applications generating significant amounts of server-to-server traffic, which does not egress the DC, commonly referred to as "east-west" traffic. Examples of such applications could be compute clusters such as 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 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 hence there is a need to constantly drive down the cost of individual network elements. This can be accomplished in two ways:

- o Unifying all network elements, preferably using the same hardware type or even the same device. This allows for volume pricing on bulk purchases.
- o 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 could be expensive, provided that larger amount of elements will statistically fail more often. Having a simpler design and operating using a limited software feature-set minimizes software issue related failures.

An important aspect of OPEX minimization is reducing size of failure domains in the network. Ethernet networks are known to be susceptible to broadcast or unicast traffic storms that have dramatic impact on network performance and availability. The use of a fully routed design significantly reduces the size of the data-plane failure domains - 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 control-plane protocols that are expected to have less chances of 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 load-balancing layer horizontally, by adding more of the uniform nodes and distributing incoming traffic across these nodes. In situation like this, an ideal choice would be to use network infrastructure itself to distribute traffic across a group of load-balancers. The combination of Anycast prefix advertisement [RFC4786] and Equal Cost Multipath (ECMP) functionality can be used to accomplish this goal. To allow for more granular load-distribution, it is beneficial for the network to support the ability to perform controlled per-hop traffic engineering. For example, it is beneficial to directly control the ECMP next-hop set for Anycast prefixes at every level of network hierarchy.

### 2.5. Summarized Requirements

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

- o 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.
- o REQ2: Define a narrow set of software features/protocols supported by a multitude of networking equipment vendors.
- o REQ3: Choose a routing protocol that has a simple implementation in terms of programming code complexity and ease of operational support.
- o REQ4: Minimize the failure domain of equipment or protocol issues as much as possible.
- o REQ5: Allow for traffic engineering, preferably via explicit control of the routing prefix next-hop using built-in protocol mechanics.

### 3. Data Center Topologies Overview

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

#### 3.1. Traditional DC Topology

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

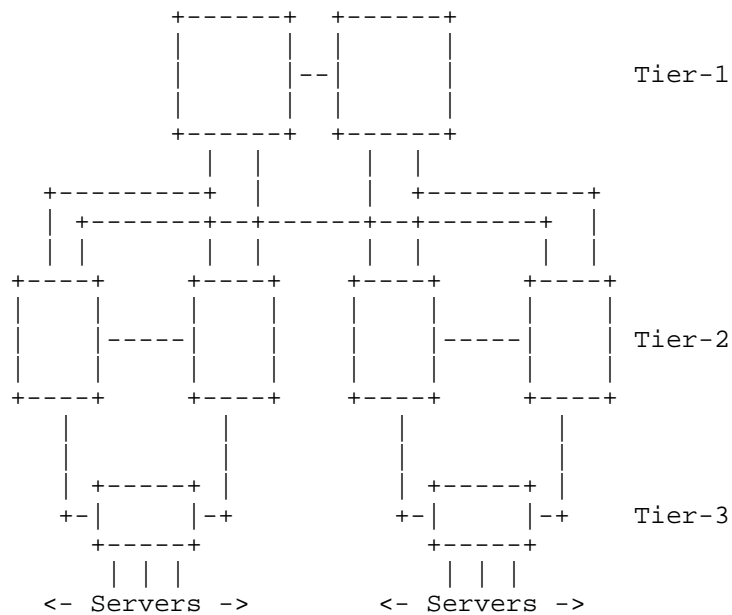


Figure 1: Typical DC network topology

### 3.2. Clos Network topology

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

#### 3.2.1. Overview

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

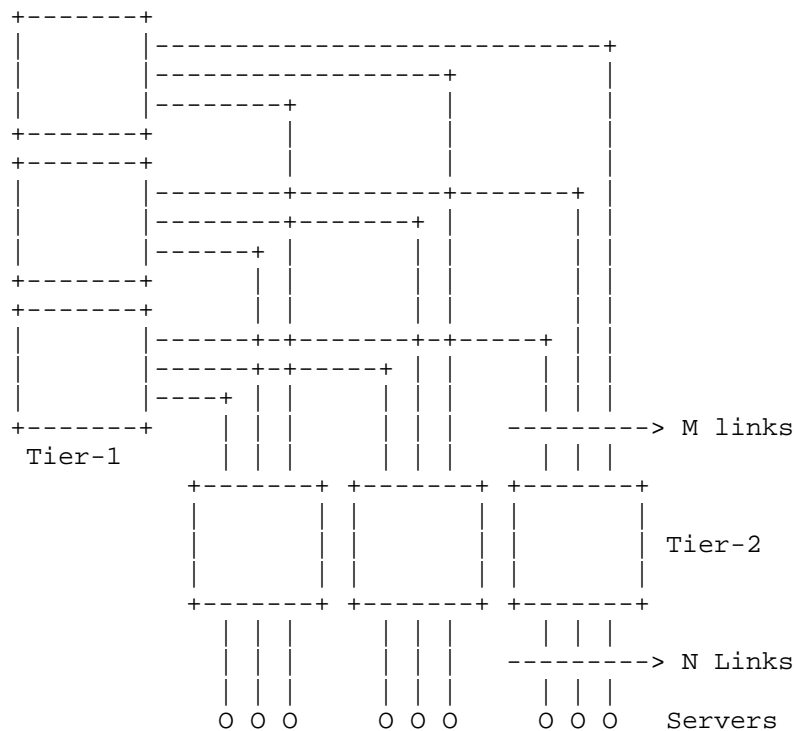


Figure 2: 3-Stage Folded Clos topology

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:

- o The topology is fully non-blocking (or more accurately: non-interfering) if  $M \geq N$  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.
- o Utilizing this topology requires control and data plane supporting ECMP with the fan-out of  $M$  or more.

- o Tier-1 switches have exactly one path to every server in this topology. This is an important property that makes route summarization impossible in this topology (see Section 8.2 below).
- o Traffic flowing from server to server is load-balanced over all available paths using ECMP.

### 3.2.3. Scaling the Clos topology

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

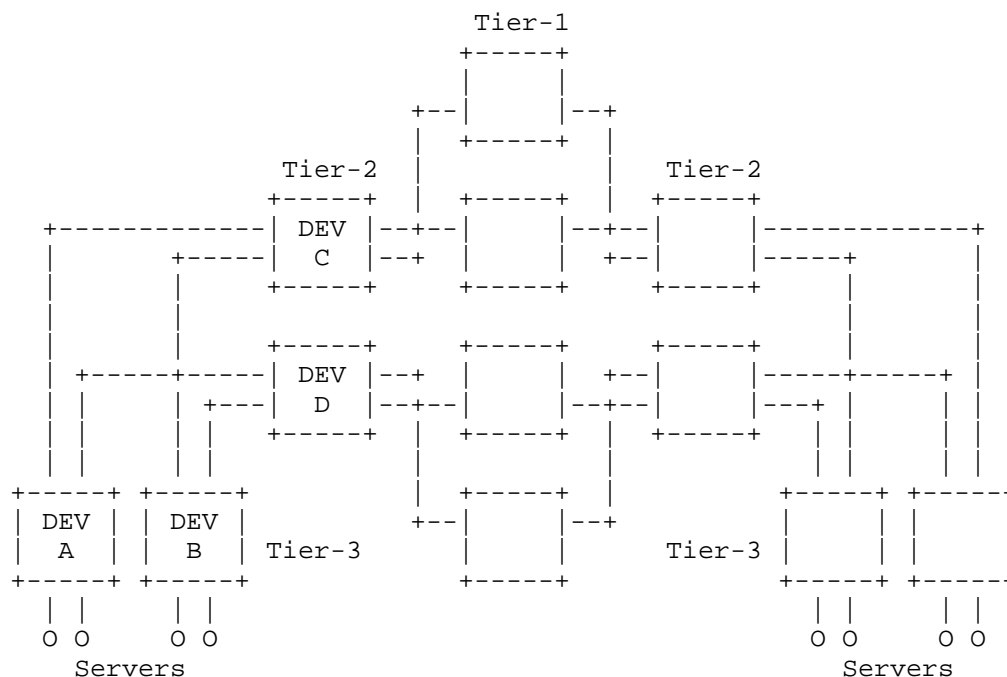


Figure 3: 5-Stage Clos topology

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

In practice, the Tier-3 layer of the network, which are typically top of rack switches (ToRs), is where oversubscription is introduced to



allow for packaging of more servers in the data center while meeting the bandwidth requirements for different types of applications. The main reason to limit oversubscription at a single layer of the network is to simplify application development that would otherwise need to account for multiple bandwidth pools: within rack (Tier-3), between racks (Tier-2), and between 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 Clos topology by a power of two. To understand how this could be done, take Tier-1 as an example. Every Tier-2 device connects to a single group of Tier-1 devices. If half of the ports on each of the Tier-1 devices are not being used then it is possible to reduce the number of Tier-1 devices by half and simply map two uplinks from a Tier-2 device to the same Tier-1 device that were previously mapped to different Tier-1 devices. This technique maintains the same bisectional bandwidth while reducing the number of elements in the Tier-1 layer, thus saving on CAPEX. The tradeoff, in this example, is the reduction of maximum DC size in terms of overall server count by half.

In this example, Tier-2 devices will be using two parallel links to connect to each Tier-1 device. If one of these links fails, the other will pick up all traffic of the failed link, possibly resulting in heavy congestion and quality of service degradation if the path determination procedure, does not take bandwidth amount into account. To avoid this situation, parallel links can be grouped in link aggregation groups (LAGs, such as [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 re-balanced over the multitude of remaining paths that equals the number of Tier-1 devices. This example is using two links for simplicity it should be noted, that 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 L2/L3 and Layer 3 only.

#### 4.1. Layer 2 Only Designs

Originally most data center designs used Spanning-Tree Protocol (STP) 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 routed protocols or supported it with additional licensing fees, which played a part in the design choice. Although many enhancements have been made through the introduction of Rapid Spanning Tree Protocol and Multiple Spanning Tree Protocol that increase convergence, stability and load balancing in larger topologies many of the fundamentals of the protocol limit its applicability in large scale DC's. STP and its newer variants use an active/standby approach to path selection and are therefore hard to deploy in horizontally scaled topologies described in Section 3.2. Further, operators have had many experiences with large failures due to issues caused by improper cabling, misconfiguration, or flawed software on a single device. These failures regularly 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 downside of this approach is proprietary nature of such extensions.

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

Finally, neither TRILL nor M-LAG approach eliminate the fundamental problem of the shared broadcast domain, that is so detrimental to the operations of any Layer 2, Ethernet based solutions.

#### 4.2. Hybrid L2/L3 Designs

Operators have sought to limit the impact of data-plane faults and build larger scale topologies through implementing routing protocols in either the Tier-1 or Tier-2 parts of the network and dividing the Layer-2 domain into numerous, smaller domains. This design has allowed data centers to scale up, but at the cost of complexity in

the network managing multiple protocols. For the following reasons, operators have retained Layer 2 in either the access (Tier-3) or both access and aggregation (Tier-3 and Tier-2) parts of the network:

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

#### 4.3. Layer 3 Only Designs

Network designs that leverage IP routing down to Tier-3 of the network have gained popularity as well. The main benefit of these designs is improved network stability and scalability, as a result of confining L2 broadcast domains. Commonly an IGP such as OSPF [RFC2328] is used as the primary routing protocol in such a design. As data centers grow in scale, and server count exceeds tens of thousands, such fully routed designs have become more attractive.

Choosing a Layer 3 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 also develop new solutions to meet some of the requirements that previously have driven large Layer 2 domains.

#### 5. Routing Protocol Selection and Design

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

### 5.1. Choosing EBGW 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 EBGW at the device bordering the WAN or Internal BGP (IBGP) throughout, this document proposes the use of an EBGW only design.

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

- o BGP is perceived as a "WAN only protocol only" and not often considered for enterprise or data center applications.
- o BGP is believed to have a "much slower" routing convergence compared to IGPs.
- o BGP deployment within an Autonomous System typically assumes the presence of an IGP for next-hop resolution.
- o 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:

- o BGP has less complexity within its protocol design - internal data structures and state-machines are simpler when compared to a link-state IGP such as OSPF. For example, instead of implementing adjacency formation, adjacency maintenance and/or flow-control, BGP simply relies on TCP as the underlying transport. This fulfills REQ2 and REQ3.
- o 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 EBGW only design. In contrast, the event propagation scope of a link-state IGP is an entire area, regardless of the failure type. This meets REQ3 and REQ4. It is worth mentioning that all widely deployed link-state IGPs also feature periodic refreshes of routing

information, while BGP does not expire routing state, even if this rarely causes significant impact to modern router control planes.

- o BGP supports third-party (recursively resolved) next-hops. This allows for manipulating multi-path to be non-ECMP based or forwarding based on application-defined forwarding paths, through establishment of a peering session with an application "controller" which can inject routing information into the system, satisfying REQ5. OSPF provides similar functionality using concepts such as "Forwarding Address", but with more difficulty in implementation and lack of protocol simplicity.
- o Using a well-defined BGP 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 BGP ASN allocation scheme. In a link-state IGP accomplishing the same goal would require multi-(instance/topology/processes) support, typically not available in all DC devices and quite complex to configure and troubleshoot. Using a traditional singleflooding domain, which most DC designs utilize, under certain failure conditions may pick up unwanted lengthy paths, e.g. traversing multiple Tier-2 devices.
- o EBGW configuration that is implemented with minimal routing policy is easier to troubleshoot for network reachability issues. In most implementations, it is straightforward to view contents of BGP Loc-RIB and compare it to the router's RIB. Also every BGP neighbor has corresponding Adj-RIB-In and Adj-RIB-Out structures with incoming and outgoing NRI information that can be easily correlated on both sides of a BGP session. Thus, BGP satisfies REQ3.

## 5.2. EBGW 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 (5 stages in unfolded state).

### 5.2.1. Example ASN Scheme

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

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

- o 16-bit (two octet) BGP ASNs are used, since these are widely supported and have better vendor interoperability.
- o Private BGP ASNs from the range 64512-65534 are used so as to avoid ASN conflicts.
- o A single BGP ASN is allocated to all of the Clos topology's Tier-1 devices.
- o Unique BGP ASN is allocated per each group of Tier-2 devices.
- o Unique BGP ASN is allocated to every Tier-3 device (e.g. ToR) in this topology.

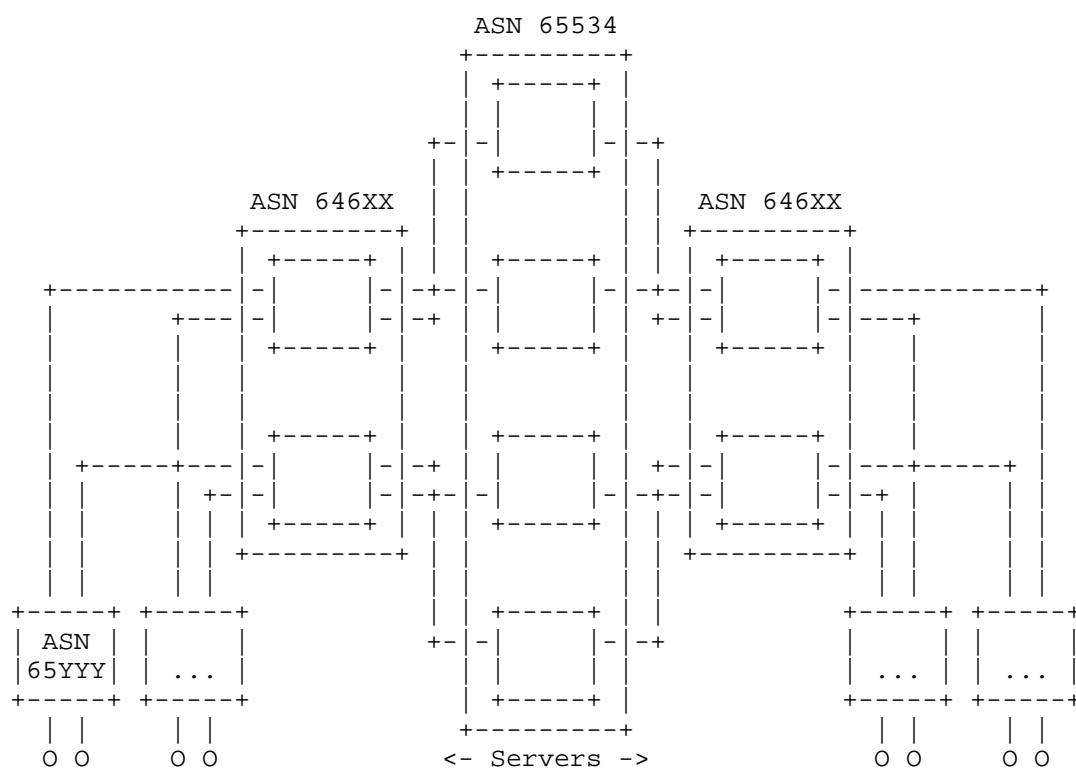


Figure 4: BGP ASN layout for 5-stage Clos

#### 5.2.2. Private Use BGP ASNs

The original range of Private Use BGP ASNs [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 BGP ASNs 65001, 65002 ... 65032 could be used within every individual cluster and assigned to Tier-3 devices.

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

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

#### 5.2.3. Prefix Advertisement

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

- o Do not advertise any of the point-to-point links into BGP. Since the EBGP based design changes the next-hop address at every device, distant networks will automatically be reachable via the advertising EBGP peer and do not require reachability to these prefixes. However, this may complicate operational troubleshooting or monitoring systems if the addresses are not reachable: e.g. using the popular "traceroute" tool will display IP addresses that are not reachable.
- o Advertise point-to-point links, but summarize them on every device. This requires an address allocation scheme such as allocating a consecutive block of IP addresses per Tier-1 and Tier-2 device to be used for point-to-point interface addressing to the lower layers (Tier-2 uplinks will be numbered out of Tier-1 addressing and so forth).

Server subnets on Tier-3 devices must be announced into BGP without using route 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 using a simpler bypass topology, e.g. a "ring" as described in [FB4POST], but such a topology adds extra hops and has very limited bisection bandwidth, in addition requiring special tweaks to make BGP routing work - such as possibly splitting every device into an ASN on its own. In Section 8.2 another, less intrusive, method for performing a limited form route summarization in Clos networks and the associated trade-offs are described.

#### 5.2.4. External Connectivity

A dedicated cluster (or clusters) in the Clos topology could be used for the purpose of connecting to the Wide Area Network (WAN) edge devices, or WAN Routers. Tier-3 devices in such 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:

- o Hide network topology information when advertising paths to WAN routers, i.e. remove Private BGP ASNs from the AS\_PATH attribute. This is typically done to avoid ASN number collisions between different data centers. An implementation specific BGP feature typically called "Remove Private AS" is commonly used to accomplish this. Depending on implementation, the feature should strip a contiguous sequence of private ASNs found in AS\_PATH attribute prior to advertising the path to a neighbor. This assumes that all BGP ASN's used for intra data center numbering are from the private ASN range. The process for stripping the private ASNs is not currently standardized, but most implementations commonly follow the logic described in [REMOVE-PRIVATE-AS] vendor's document.
- o Originate a default route to the data center devices. This is the only place where default route can be originated, as route summarization is risky for the "scale-out" topology. Alternatively, Border Routers may simply relay the default route learned from WAN routers. Advertising the default route from Border Routers requires that all Border Routers to be fully connected to the WAN Routers upstream, to provide resistance to a



single-link failure causing the black holing of traffic. To prevent chance of operator or implementation error that may impact EBGp sessions to the WAN routers simultaneously (although these scenarios are not planned for by many operators since they represents a multiple failure) it is more desirable to take this approach rather than introducing complicated conditional default origination schemes provided by some implementations.

#### 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 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 or other prefixes. However, as discussed before, 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:

- o 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 underneath the Border Routers.
- o 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 devices would have to connect to at least two Border Routers. This puts additional requirements on the port count for Tier-1 devices and Border Routers, potentially making it a non-uniform, larger port count, device with the other devices in the Clos. 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 layer.

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 non-uniform

topology as additional links have to be provisioned on some network devices.

## 6. ECMP Considerations

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

### 6.1. Basic ECMP

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

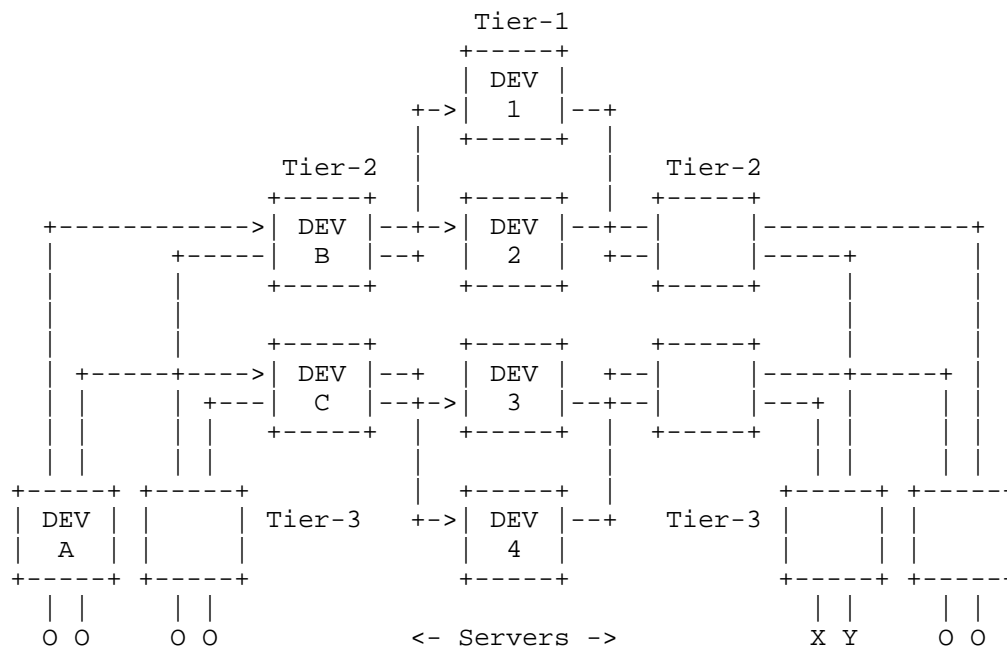


Figure 5: ECMP fan-out tree from A to X and Y

The ECMP requirement implies that the BGP implementation must support multi-path fan-out for up to the maximum number of devices directly attached at any point in the topology in upstream or downstream

direction. Normally, this number does not exceed half of the ports found on a device in the topology. For example, an ECMP 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 multitude of Tier-1 devices if route summarization at Border Router level is implemented as described in Section 5.2.5. If a device's hardware does not support wider ECMP, logical link-grouping (link-aggregation at layer 2) could be used to provide "hierarchical" ECMP (Layer 3 ECMP followed by Layer 2 ECMP) to compensate for fan-out limitations. Such approach, however, increases the risk of flow polarization, as less entropy will be available to the second stage of ECMP.

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

## 6.2. BGP ECMP over Multiple ASNs

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

## 6.3. Weighted ECMP

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

signaled using the technique described in [I-D.ietf-idr-link-bandwidth].

#### 6.4. Consistent Hashing

It is often desirable to have the hashing function used to ECMP to be consistent (see [CONS-HASH]), to minimizing the impact on flow to next-hop affinity changes when a next-hop is added or removed to 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 detrimental effect of 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. Notice that the usual downside of consistent hashing is increased load on hardware resource utilization, as typically more space is required to implement a consistent-hashing region.

### 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 update upon failure of the associated link.

#### 7.1. Fault Detection Timing

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

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

Ethernet technologies may support failure signaling or detection standards such as [IEEE8021AG] and [IEEE8023AH], which may make failure detection more robust. Alternatively, some platforms may support Bidirectional Forwarding Detection (BFD) [RFC5880] to allow for sub-second failure detection and fault signaling to the BGP process. However, use of either of these presents additional requirements to vendor software and possibly hardware, and may contradict REQ1. Until recently with [I-D.ietf-bfd-on-lags], BFD also did not allow detection of a single member link failure on a LAG, which would limit its usefulness in some designs.

## 7.2. Event Propagation Timing

In this design the impact of BGP Minimum Route Advertisement Interval (MRAI) timer (See section 9.2.1.1 of [RFC4271]) should be considered. 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 has either one path or N paths for the same prefix, where N is a significantly large number, e.g. N=32 (the ECMP fan-out). Therefore, if a path fails there is either no backup path at all, or the backup is readily available in BGP Local RIB. In the former case, the BGP withdrawal announcement will propagate un-delayed and trigger re-convergence 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 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 WITHDRAW message to all upstream Tier-1 devices, and Tier-1 devices will relay this message to all downstream Tier-2 devices (except for the originator). Tier-2 devices other than the one originating the WITHDRAW should then wait for ALL adjacent Tier-1 devices to send a WITHDRAW message before it removes the affected prefixes and sends corresponding WITHDRAW downstream to connected 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 long as multiple seconds. In order to avoid such 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.

#### 7.4. Failure Impact Scope

A network is declared to converge in response to a failure once all devices within the failure impact scope are notified of the event and have re-calculated their RIB's and consequently updated their FIB's. Larger failure impact scope typically means slower convergence since more devices have to be notified, and additionally results in a less stable network. In this section 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 the FIB. However, many failures will not result in such a wide impact. There are two main failure types where impact scope is reduced:

- o 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, which 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 prefix withdrawal process will only affect Tier-2 devices directly connected to the affected Tier-1 device. The Tier-2 devices receiving the BGP UPDATE messages withdrawing prefixes will simply have to update their ECMP groups. The Tier-3 devices are not involved in the re-convergence process.
- o 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, but may receive "implicit withdraws" as described above.

Even though in case of such failures multiple IP prefixes will have to be reprogrammed in the FIB, it is worth noting that ALL of these prefixes share a single ECMP group on Tier-2 device. Therefore, in the case of implementations with a hierarchical FIB, only a single change has to be made to the FIB. Hierarchical FIB here means FIB structure where the next-hop forwarding information is stored separately from the prefix lookup table, and the latter only store pointers to the respective forwarding information.

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

#### 7.5. Routing Micro-Loops

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

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

## 8. Additional Options for Design

### 8.1. Third-party Route Injection

BGP allows for a "third-party", i.e. directly attached, BGP speaker to inject routes anywhere in the network topology, meeting REQ5. This can be achieved by peering via a multihop BGP session with some or even all devices in the topology. Furthermore, BGP diverse path distribution [RFC6774] could be used to inject multiple BGP next hops for the same prefix to facilitate load-balancing, or using the BGP ADD-PATH capability [I-D.ietf-idr-add-paths] if supported by the implementation. Unfortunately, in many implementations ADD-PATH has been found to only support IBGP properly due to the use cases it was originally optimized for, which limits the "third-party" peering to iBGP only, if the feature is used.

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 parallel 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.

Route summarization would be possible with a small modification to the network topology, though the trade-off would be reduction of the total size of the network as well as network congestion under specific failures. This approach is very similar to the technique described above, which allows Border Routers to summarize the entire data-center address space.

#### 8.2.1. Collapsing Tier-1 Devices Layer

In order to add more paths between Tier-1 and Tier-3 devices, group Tier-2 devices into pairs, and then connect the pairs to the same group of Tier-1 devices. This is logically equivalent to "collapsing" Tier-1 devices into a group of half the size, merging the links on the "collapsed" devices. The result is illustrated in



Figure 6. For example, in this topology DEV C and DEV D connect to the same set of Tier-1 devices (DEV 1 and DEV 2), whereas before they were connecting to different groups of Tier-1 devices.

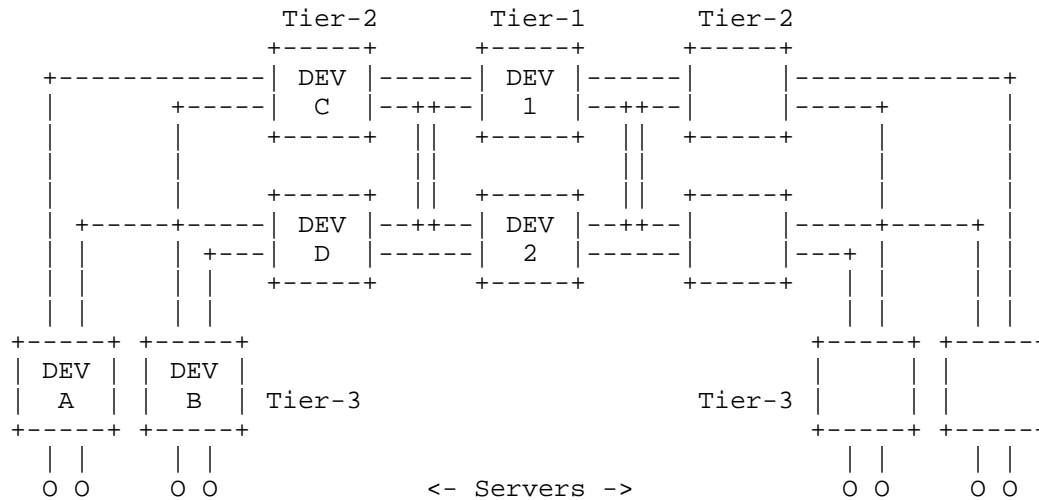


Figure 6: 5-Stage Clos topology

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 not possible to advertise a summary route covering prefixes for a single cluster from Tier-2 devices since each of them has only a single path down to this prefix. It would require dual-homed servers to accomplish that. Also note that this design is only resilient to single link failure. It is possible for a double link failure to isolate a Tier-2 device from all paths toward a specific Tier-3 device, thus causing a routing black-hole.

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

Another problem is traffic re-balancing under link failures. Since 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 of 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 pressure, 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, looking from the perspective of a Tier-3 devices, all routes learned from upstream Tier-2's, including the default route, will share the same set of BGP next-hops, provided that there is no failures in the network. This makes it possible to use the technique similar to 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 need to be programmed into 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 for detection of mismatches in the next-hop sets if a particular link fails, changing the next-hop set for a specific prefix.

Re-stating once again, this technique does not reduce the amount of control plane state (i.e. BGP UPDATES/BGP LocRIB sizing), but only allows for more efficient FIB utilization, by spotting more specific prefixes that share their next-hops with less specifics.

### 8.3. ICMP Unreachable Message Masquerading

This section discusses some operational aspects of not advertising point-to-point link subnets into BGP, as previously outlined 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 the IP addresses of the same device pointing to the same name. The connectivity then can be made by resolving this name to the "primary" IP address of the devices, e.g. its Loopback interface, which is always advertised into BGP.

However, this create dependency on DNS subsystem, which may happen to 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 off of the device with the "primary" IP address of the device. Specifically, the ICMP Destination Unreachable Message (type 3) codes 3 (port unreachable) and ICMP Time Exceeded (type 11) code 0, which are involved in proper working 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.

## 9. Security Considerations

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

## 10. IANA Considerations

This document includes no request to IANA.

## 11. Acknowledgements

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

## 12. References

### 12.1. Normative References

- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC6996] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", BCP 6, RFC 6996, July 2013.

## 12.2. Informative References

- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006.
- [RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, December 2006.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC6325] Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (Rbridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC6774] Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, November 2012.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, December 2012.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.
- [RFC6769] Raszuk, R., Heitz, J., Lo, A., Zhang, L., and X. Xu, "Simple Virtual Aggregation (S-VA)", RFC 6769, October 2012.
- [I-D.ietf-idr-add-paths]  
Walton, D., Retana, A., Chen, E., and J. Scudder,  
"Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-09 (work in progress), October 2013.
- [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.

[I-D.ietf-bfd-on-lags]

Bhatia, M., Chen, M., Boutros, S., Binderberger, M., and J. Haas, "Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces", draft-ietf-bfd-on-lags-04 (work in progress), December 2013.

[GREENBERG2009]

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

[IEEE8021AG]

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

[IEEE8023AH]

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

[INTERCON]

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

[ALFARES2008]

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

[IANA.AS]

IANA, , "Autonomous System (AS) Numbers", January 2014, <<http://www.iana.org/assignments/as-numbers/>>.

[IEEE8023AD]

IEEE 802.3ad, , "IEEE Standard for Link aggregation for parallel links", October 2000.

[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](http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080093f27.shtml)>.

[FB4POST]

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

[JAKMA2008]

Jakma, P., "BGP Path Hunting", 2008, <[https://blogs.oracle.com/paulj/entry/bgp\\_path\\_hunting](https://blogs.oracle.com/paulj/entry/bgp_path_hunting)>.

[CONS-HASH]

Wikipedia, , "Consistent Hashing",  
<[http://en.wikipedia.org/wiki/Consistent\\_hashing](http://en.wikipedia.org/wiki/Consistent_hashing)>.

#### Authors' Addresses

Petr Lapukhov  
Facebook  
1 Hacker Way  
Menlo Park, CA 94025  
US

Email: [petr@fb.com](mailto:petr@fb.com)

Ariff Premji  
Arista Networks  
5453 Great America Parkway  
Santa Clara, CA 95054  
US

Email: [ariff@aristanetworks.com](mailto:ariff@aristanetworks.com)  
URI: <http://aristanetworks.com/>

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

Email: [Jon.Mitchell@microsoft.com](mailto:Jon.Mitchell@microsoft.com)