BGP

Border Gateway Protocol

BGP Course Outline

- BGP Basics – Why BGP, Autonomous Systems
- BGP Best Path Selection
- BGP Monitoring Protocol - BMP
- EBGP – Basics of EBGP
- EBGP Traffic Engineering – Inbound and Outbound
- GAO – Rexford Model – Inter-domain Policies for Business Economics
- EBGP Multipath, IBGP Multipath and EIBGP Multipath
- Inter domain routing – Settlement Free Peering, Partial Peering, Remote Peering and IP Transit
- BGP Route Servers and the Looking Glasses
- ISP Tiers – Tier 1 , 2 , 3 Type Providers

BGP Course Outline

- IBGP – Basics of IBGP
- BGP Route Reflectors
- Route Reflector Design Options
- Centralized vs. Distributed, Inline and Offline Route Reflector Design
- BGP Add-Path , BGP Shadow Route Reflectors/Distributing Diverse Paths and Shadow Sessions
- Unique RD per VRF per PE Approach for path distribution
- Comparison between BGP Add-path, Shadow RR, Shadow Sessions and Unique RD per VRF per PE
- BGP RT – Route Target Constraint
- BGP Optimal Route Reflection – ISP Case Study
- BGP Confederations and Confederation Design
- Full mesh IBGP vs. Route Reflector vs. Confederation
- Full mesh IBGP to BGP RR Migration

BGP Course Outline

- BGP – IGP Interactions – Blackhole avoidance
- BGP – MPLS Interactions
- BGP LU – Labeled Unicast – RFC 3107 , RFC 8277
- BGP LS – BGP Link State
- BGP EPE – Egress Peer Engineering – Traditional EPE and Modern Egress Peer Engineering with EPE/SDN Controller
- BGP RTBH , Source and Destination based Remotely Triggered Blackholing and Scrubbing for DDOS mitigation with BGP
- BGP Flowspec – RFC 5575

BGP Course Outline

- BGP Session Culling and Alternative Approaches , OXC
- BGP Graceful Restart, BGP Graceful Shutdown and BGP Administrative Shutdown Communication
- AIGP – Accumulated IGP Metric Attribute for BGP
- BGP MED vs. AIGP
- EBGP Default Route Propagation Behavior without Policies– RFC 8212
- BGP Information Security – 6 Different Types BGP Route Leaks
- 512K Incident Types 6 BGP Route Leak Contribution to it!
- BGP Information Security - Sub-Prefix Hijacking, Exact Prefix Hijacking , IP Squatting , Path-Shortening, Prefix Filters, IRR , RPSL , RPKI – Resource Public Key Infrastructure, PeeringDB, BGPSEC , Origin and Path Validation, AS-Cones , ASPA

BGP Course Outline

- BGP in the Datacenter
- BGP in the Wide Area Networks
- Flat, Hierarchical and Generalized FIB Architecture and impact on BGP Prefix Independent Convergence
- BGP PIC – Prefix Independent Convergence , BGP PIC Core and BGP PIC Edge
- Case Studies
- BGP vs. IGP Comparison
- BGP in the CCIE , CCDE exam
- Summary
- BGP Quiz – BGP Questions and the Answers
- - Bonus - BGP vs. other Hyper Scale Datacenter Routing Protocol Comparisons : RIFT , Open Fabric and BGP SPF

BGP – Border Gateway Protocol Basics

Why BGP ?

- If the requirement is to use a routing protocol on the Public Internet then only choice is Border Gateway Protocol aka BGP

- BGP is the most scalable routing protocol and considered as very robust as it runs over TCP and TCP is inherently reliable

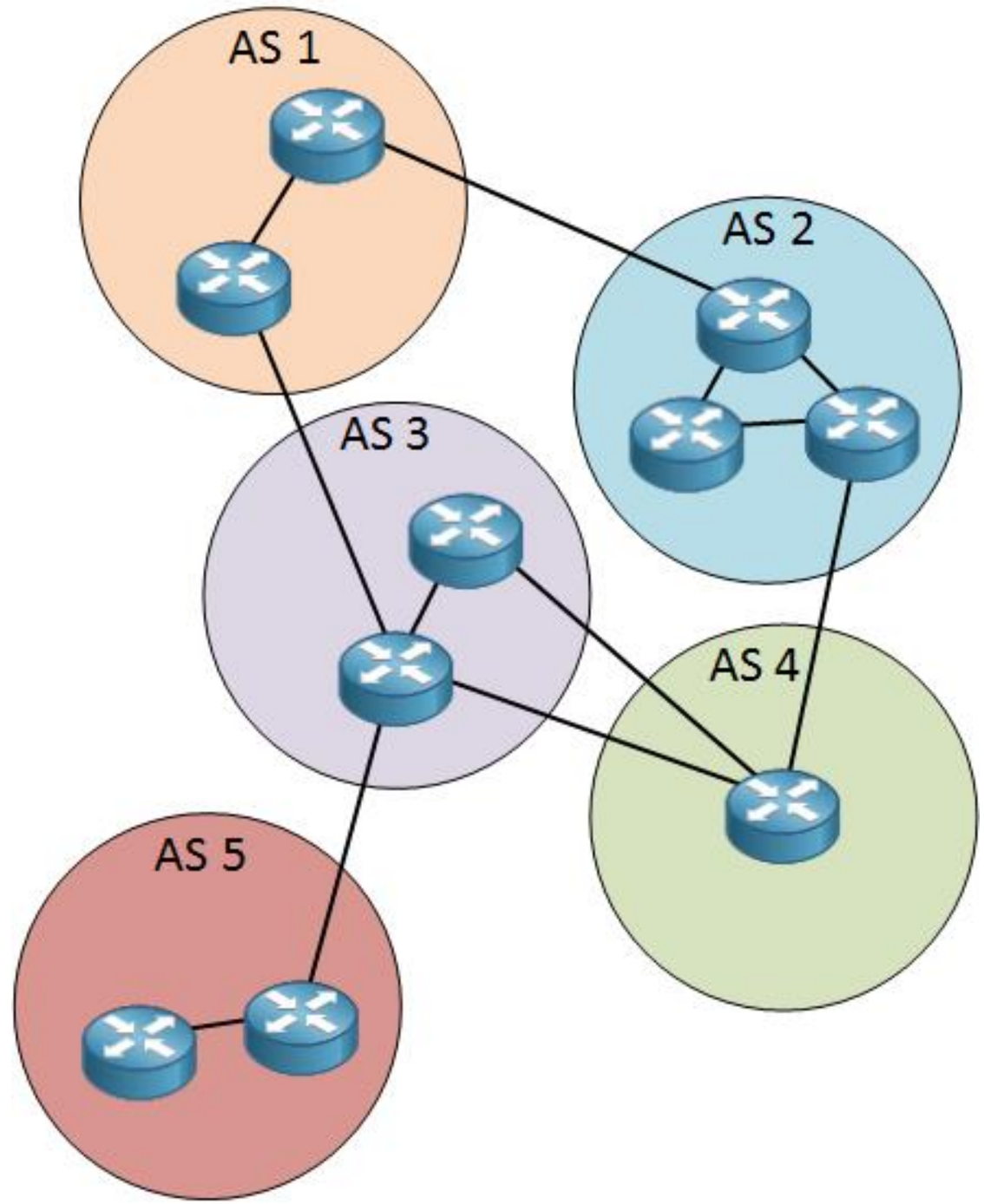
- BGP is a multi protocol , with the new NLRI it can carry many address families. Today almost a 20 different NLRI is carried over BGP. New AFI, SAFI is defined for the new address families (https://orhanergun.net/tag/multi-protocol-bgp/)

BGP – Border Gateway Protocol Basics

Autonomous System

An Autonomous System (AS) is a collection of routers whose prefixes and routing policies are under common administrative control. This could be a network service provider, a large company, a university, a division of a company, or a group of companies

An Exterior Gateway Protocol (EGP) is a routing protocol that handles routing between Autonomous Systems (inter-AS routing). BGP version 4, the Border Gateway Protocol, is the standard EGP for inter-AS routing


AS 1
AS 2
AS 3
AS 4
AS 5

BGP – Border Gateway Protocol Basics

- EBGP and IBGP are our main focus. If the BGP connection between two different Autonomous Systems, it is called EBGP (External BGP).

- If BGP is used inside an Autonomous System, so same AS number is used between the BGP nodes, then the connection is called IBGP (Internal BGP)

BGP Best Path Selection

- Unlike IGP protocols, BGP doesn’t use link metrics for the best path selection. Instead it uses many attributes for the best path selection. This allows creating complex BGP policies

- BGP is a policy based protocol which provides IP based Traffic Engineering inside an Autonomous System

- In fact IGPs don’t support traffic engineering like the BGP does

BGP Best Path Selection

- BGP path vector protocol which has many similarities with the Distance Vector protocols such as EIGRP

- For example in EBGP and IBGP, always one best path is chosen and placed in the Routing table, this path is advertised to the other BGP neighbor. This might create sub optimal routing design or slow BGP convergence as we will see later in the BGP course

- There might be vendor specific attributes such as Weight attribute. Also there are some intermediary steps which is not used commonly. Below is the BGP best path selection criteria list

BGP best path selection steps

- BGP next hop has to be reachable
- Longest match wins
- Weight
- Local Preference
- As-Path
- As-Path
- Origin
- MED
- Prefer EBGP over IBGP
- Lowest IGP metric to the BGP next hop(Hot Potato)
- Multipath
- Lastly prefer lowest neighbor address

BGP Best Path Selection

- Local Preference is used to send traffic on outbound direction. When prefixes are received from BGP neighbor, default local preference value is 100

- Local preference value can be changed, higher local preference value is preferred to lower value

- If same prefix is received from two BGP neighbors, neighbor which has higher local preference value is chosen by BGP as a best path and used to send traffic from Autonomous System to the other Autonomous Systems

- For incoming traffic from other Autonomous Systems to Local Autonomous System, BGP MED Attribute, AS-Path Prepending and Community Attribute techniques can be used

- All these techniques will be explained later in the EBGP topic

BMP - BGP Monitoring Protocol

- BMP is defined in RFC 7854
- It is used to monitor BGP sessions
- Until BMP, information about BGP session was received with CLI which can be a CPU intensive
- BMP is an automated way of collecting the BGP data from the routers

BMP - BGP Monitoring Protocol

- BMP client (monitored router) peers with several BGP speaking routers (BGP peers). The BMP client establishes a monitoring session to one or more BMP collectors (monitoring devices)

- The client encapsulates BGP messages from one or more BGP peers into a single TCP stream to one or more BMP collectors

- BMP collectors store data in a database thus automated programs or scripts can access the database and process this data

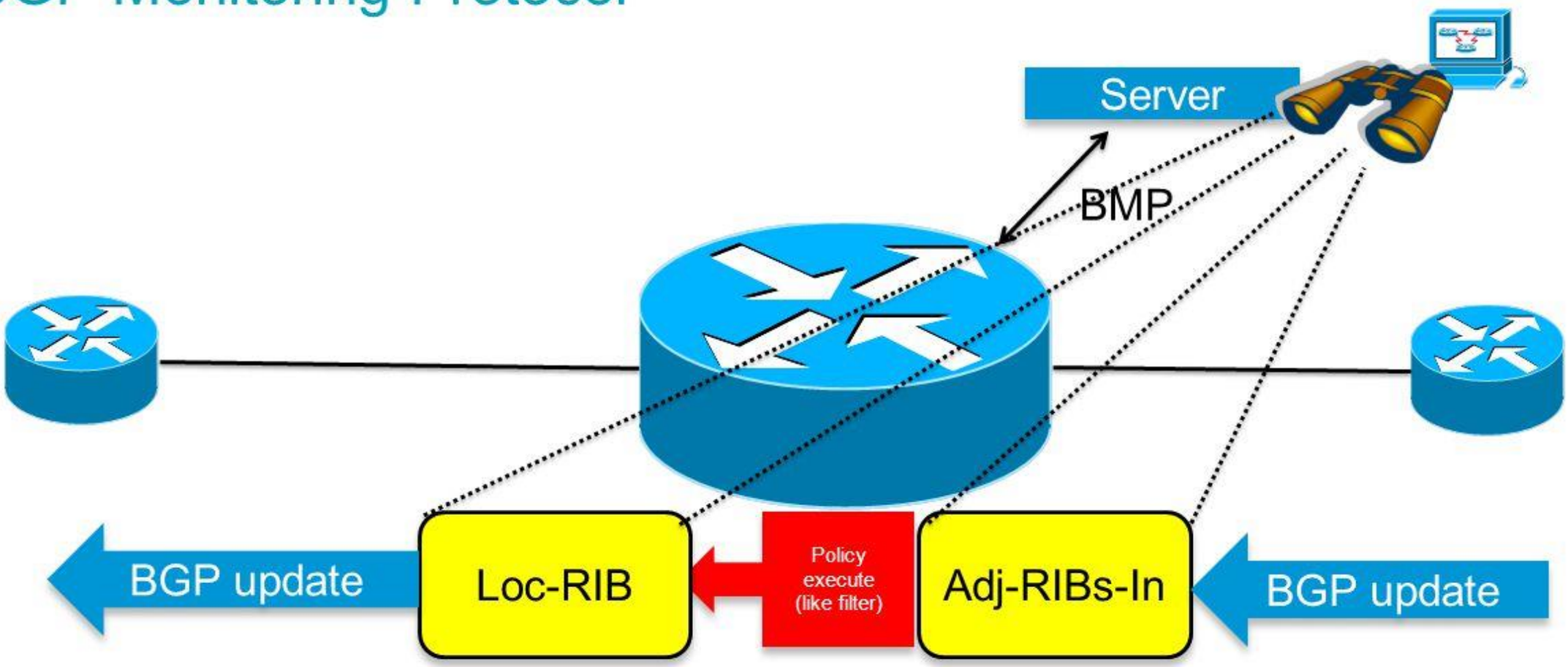
BMP - BGP Monitoring Protocol

- BMP provides an access to the Adjacency-RIB-In database of router

- The Adj-RIBs-In stores unprocessed routing information received from BGP peers. Network operator then has the unedited access to the routing information sent from BGP peers to the BMP client

- BMP also provides a periodic dump of statistics that can be used for further analysis

BGP Monitoring Protocol



BMP - BGP Monitoring Protocol

- BMP operates over TCP
- When a TCP connection is established, BMP messages are being sent from the BMP client to a BMP collector
- No BMP message is ever sent from the collector to the client

Which information are sent with BMP?

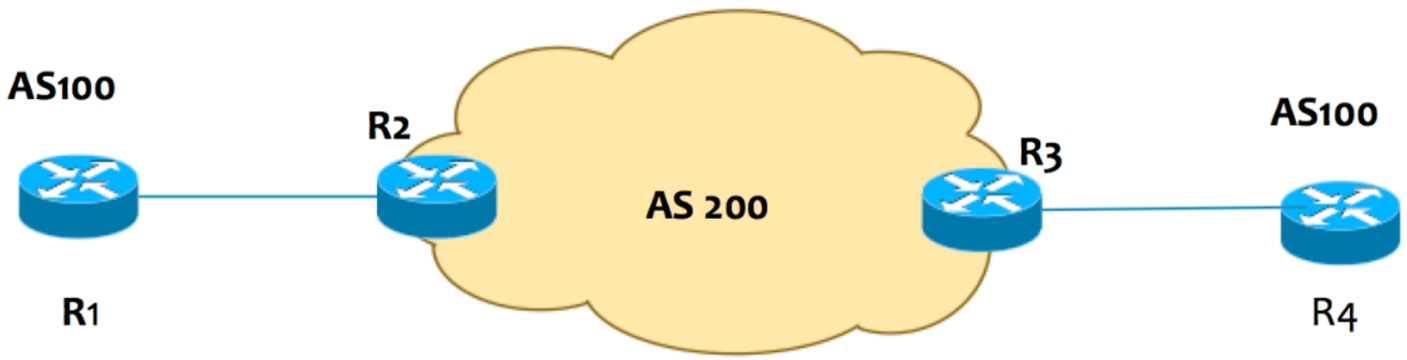
- An initial dump of the current BGP table, called *route monitoring*
- Peer down notification, including a code indicating why the peer went down
- Stat reports, including number of prefixes rejected by inbound policy, number of duplicate prefixes, number of duplicate withdraws, etc.
- Peer up notification

What you cannot do with BMP?

- You can’t monitor outgoing routes (Adj-RIB-Out)
- You can’t monitor the best BGP routes (Loc-RIB)
- You can’t monitor why prefixes were rejected (Post-Policy routes)
- You can’t see why routes didn’t win the best path selection
- There are two drafts in IETF as BMP extension, one is for Outgoing routes (Adj-RIB-OUT) and another for best BGP routes (Loc-RIB)

EBGP

- EBGP is used between two different Autonomous Systems, loop prevention in EBGP is done by the AS Path attribute, that’s why it is a mandatory BGP attribute
- If BGP node sees its own AS Path in the incoming BGP update message, BGP message is rejected



EBGP

EBGP Traffic Engineering

- BGP traffic engineering is to send and receive the network traffic based on customer business and technical requirements

- For example link capacities might be different, one link might be more stable than the other, or monetary costs of the links might be different

- In all these cases , customer may want to optimize their incoming and outgoing traffic

EBGP Traffic Engineering

- Network traffic flows in two directions ; Incoming and outgoing

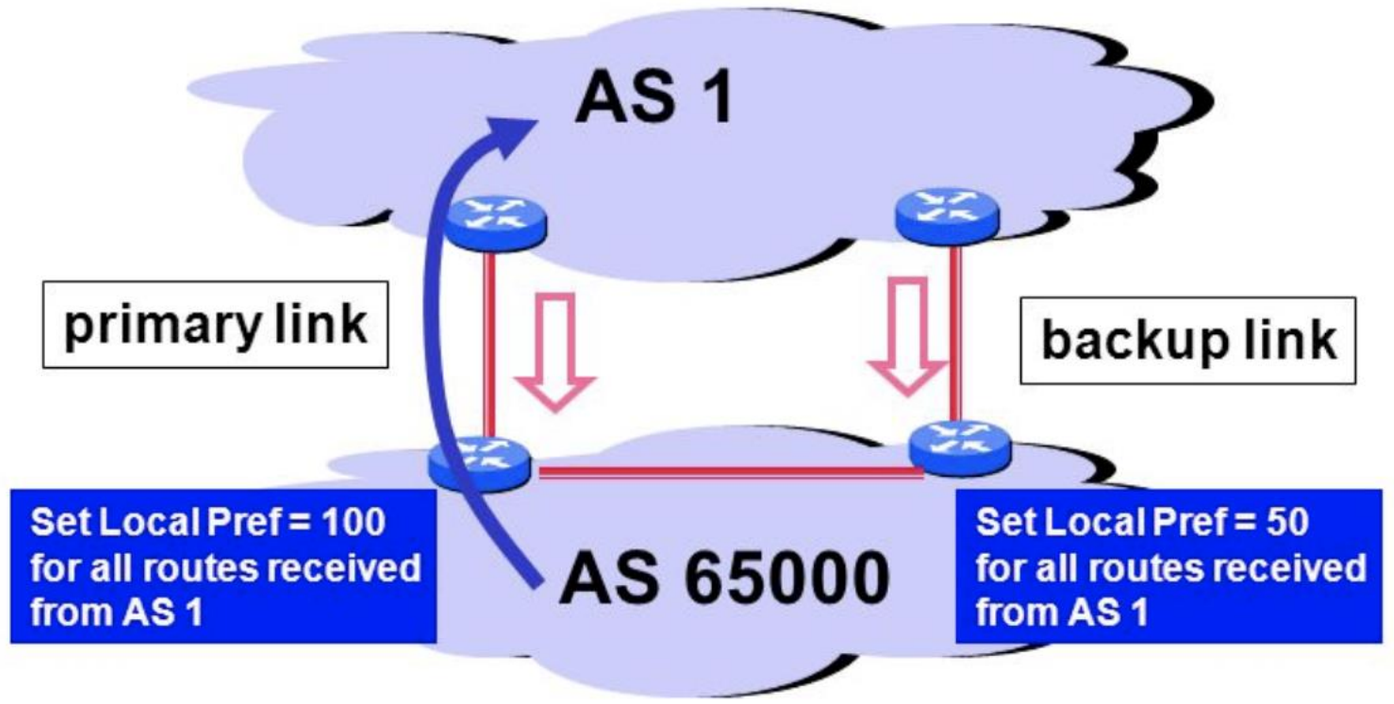
- Incoming traffic engineering refers receiving traffic into the Local Autonomous System from one of the many available paths or receiving specific application/services traffic from any path

EBGP Traffic Engineering

- Outbound Traffic Engineering: Refers sending the traffic from Local AS to the other Autonomous Systems from one of the many paths or sending specific application/services traffic to other AS from any path

- For the BGP outgoing traffic, commonly, local preference attribute is used

EBGP Outbound Traffic Engineering



AS 65000 has two paths to AS1, by increasing Local Preference on one of the links, AS 65000 sends all outbound traffic from the AS over that path

EBGP Outbound Traffic Engineering

- BGP weight attribute can be used for the outgoing traffic engineering as well but don’t forget that it is local to the router which mean is not propagated between the IBGP neighbors and it is Cisco preparatory, no vendor interoperability

GAO – Rexford Model

- Service Providers chooses to send the traffic for their customer prefixes over the customer link first, then peering links, lastly through upstream ISP. Because they want to utilize the customer link as much as possible to charge more money

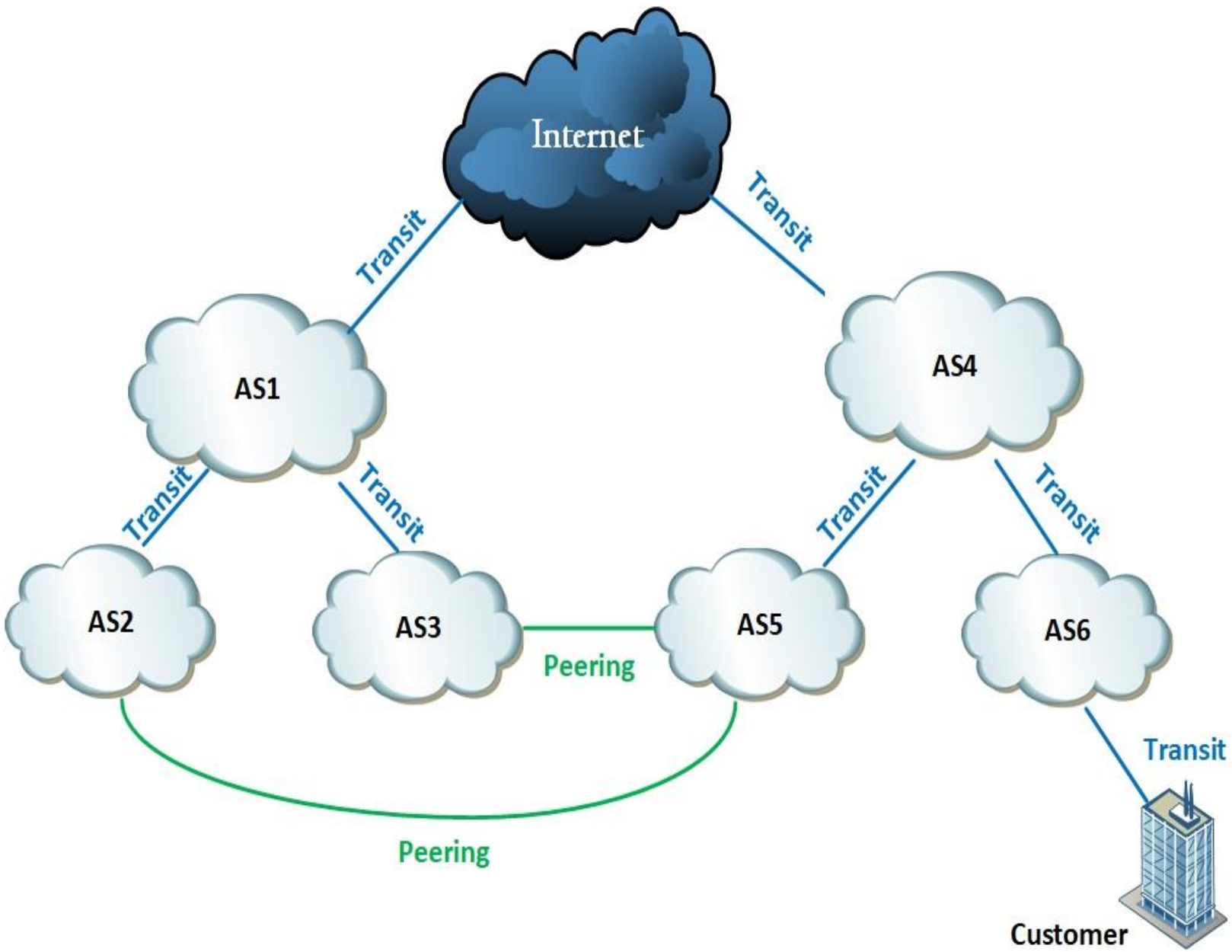
- Utilizing Customer link can generate revenue , Settlement Free Peering Link or Upstream ISP link usage cost money, thus technical solution provided by GAO and Rexford is used to realize this business requirement

GAO – Rexford Model

- Service Providers may implement Local Preference attribute to achieve this

- Basic local preference policy could be; Local Preference 100 towards Customer, local Preference 90 towards peering link and Local Preference 80 towards upstream ISP

GAO – Rexford Model



EBGP Inbound Traffic Engineering

- **BGP Inbound traffic engineering can be achieved in multiple ways:**

1. MED (BGP External metric attribute)
2. AS-Path prepending
3. BGP Community attribute

EBGP Inbound Traffic Engineering with MED

- Creating an inter domain policy with the MED attribute is not a good practice

- MED attribute is used between two Autonomous system. If the same prefix is coming from two different AS to the 3rd AS, although you can use always-compare MED feature, it is not good practice to enable this feature since it can cause BGP MED Oscillation problem

Don’t compare BGP MED when the prefixes are received from two different AS !

- As per RFC 4451 – BGP MED Considerations : BGP speakers often derive MED values by obtaining the IGP metric associated with reaching a given BGP NEXT_HOP within the local AS. This allows MEDs to reasonably reflect IGP topologies when advertising routes to peers. While this is fine when comparing MEDs between multiple paths learned from a single AS, it can result in potentially "weighted" decisions when comparing MEDs between different autonomous systems.

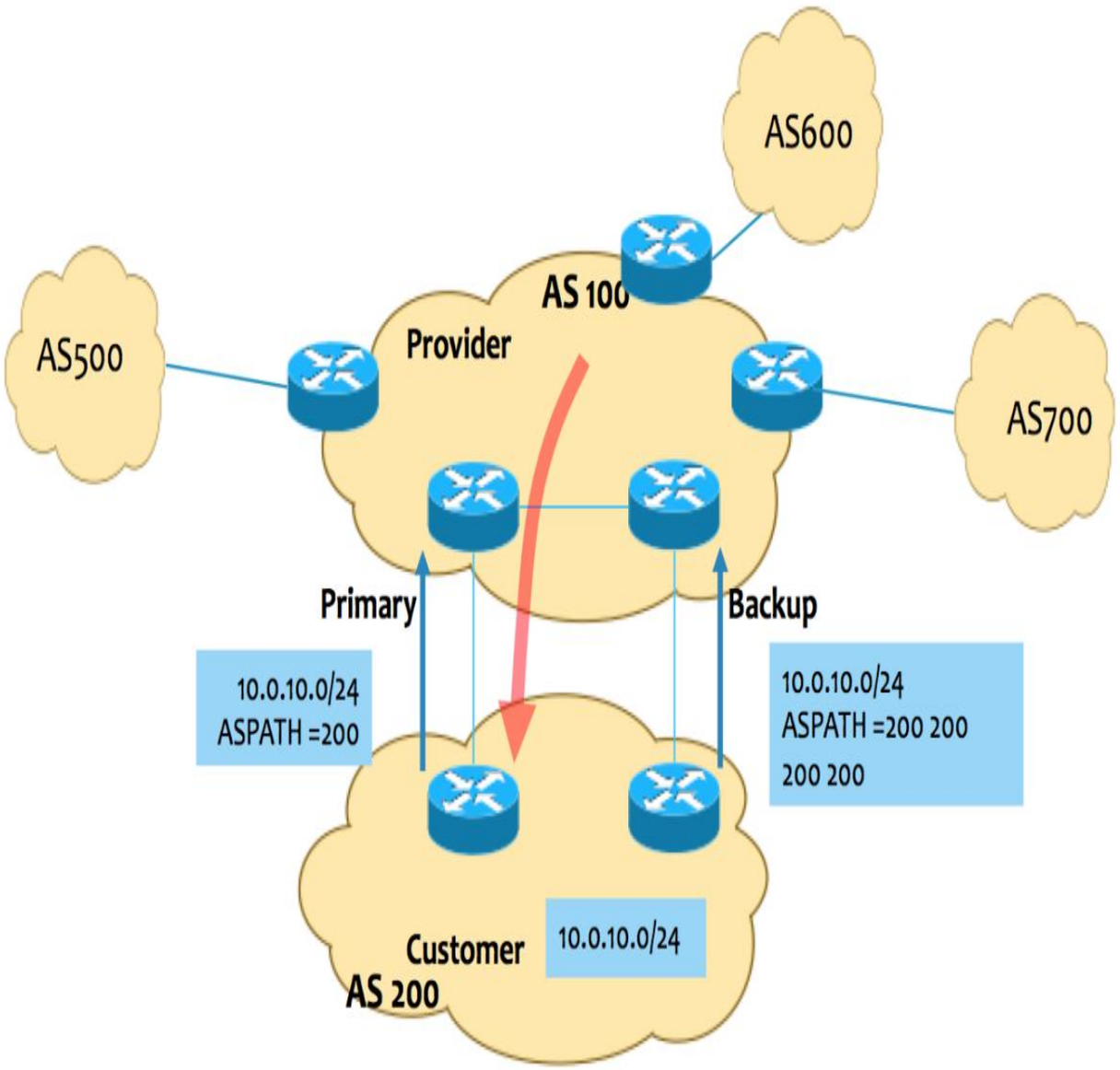
- This is most typically the case when the autonomous systems use different mechanisms to derive IGP metrics for BGP MEDs, or when they perhaps even use different IGP protocols with vastly contrasting metric spaces (e.g., OSPF vs. traditional metric space in IS-IS)

EBGP Inbound Traffic Engineering with AS-Path Prepending

- BGP As-path is a mandatory BGP attribute which has to be sent in every BGP message. BGP as-path prepending is one of the BGP traffic engineering methods

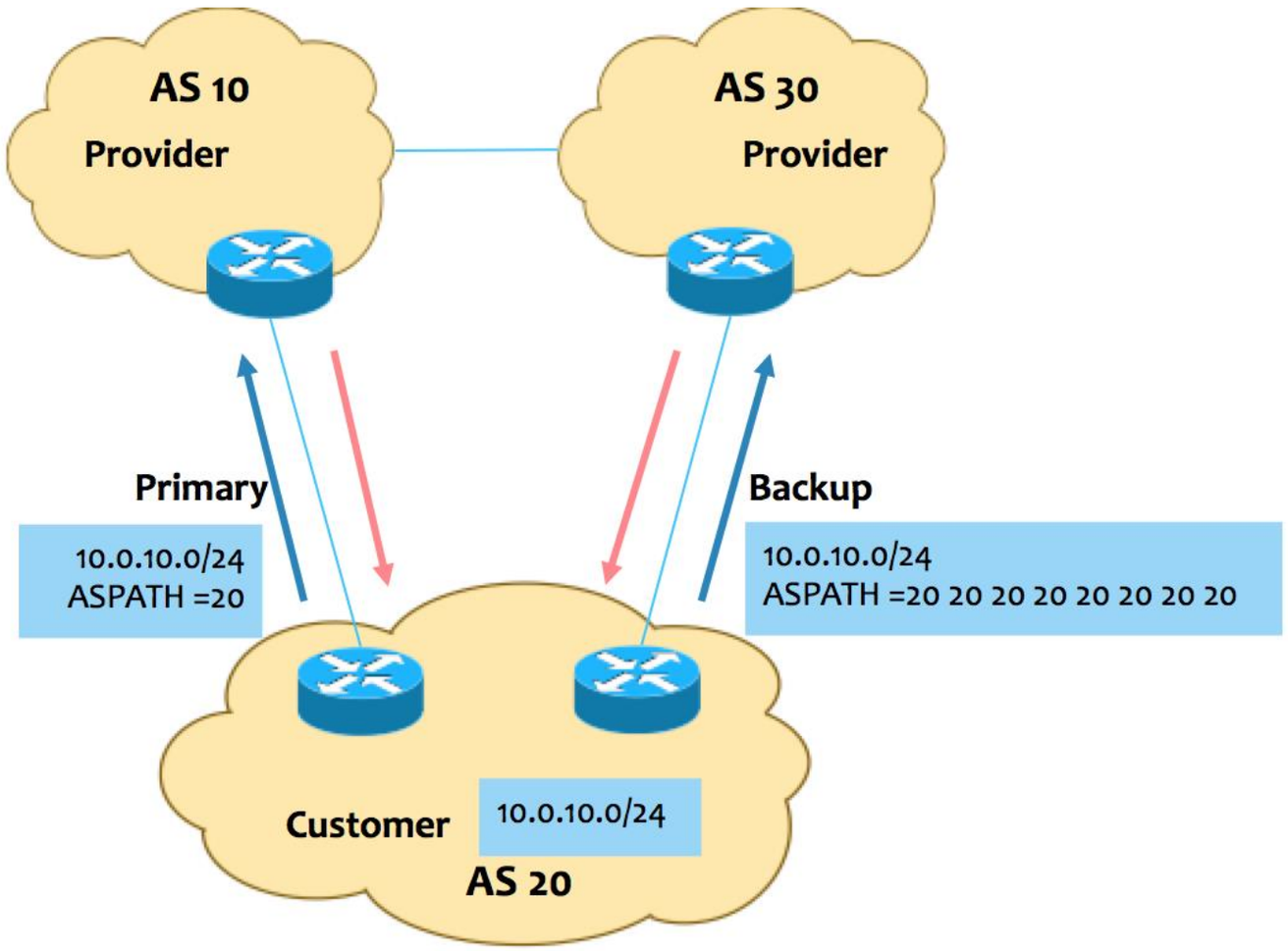
- BGP As-path prepending is used to influence inbound traffic to the company. BGP As-path prepending is used in active-standby link scenarios. When there are two BGP neighborship which prefix will be advertised, one link for set of prefixes or maybe all the prefixes can be used as backup. In this case, one way to achieve this setup is using BGP AS-path prepending.

EBGP Inbound Traffic Engineering with AS-Path Prepending



Customer AS 200 wants to use one of the links as backup. 10.0.10.0/24 prefix is sent via backup link with the 3 prepend. Thus AS path is seen through the backup link by the upstream service provider which is AS 100 as ‘ 200 200 200 200 ‘

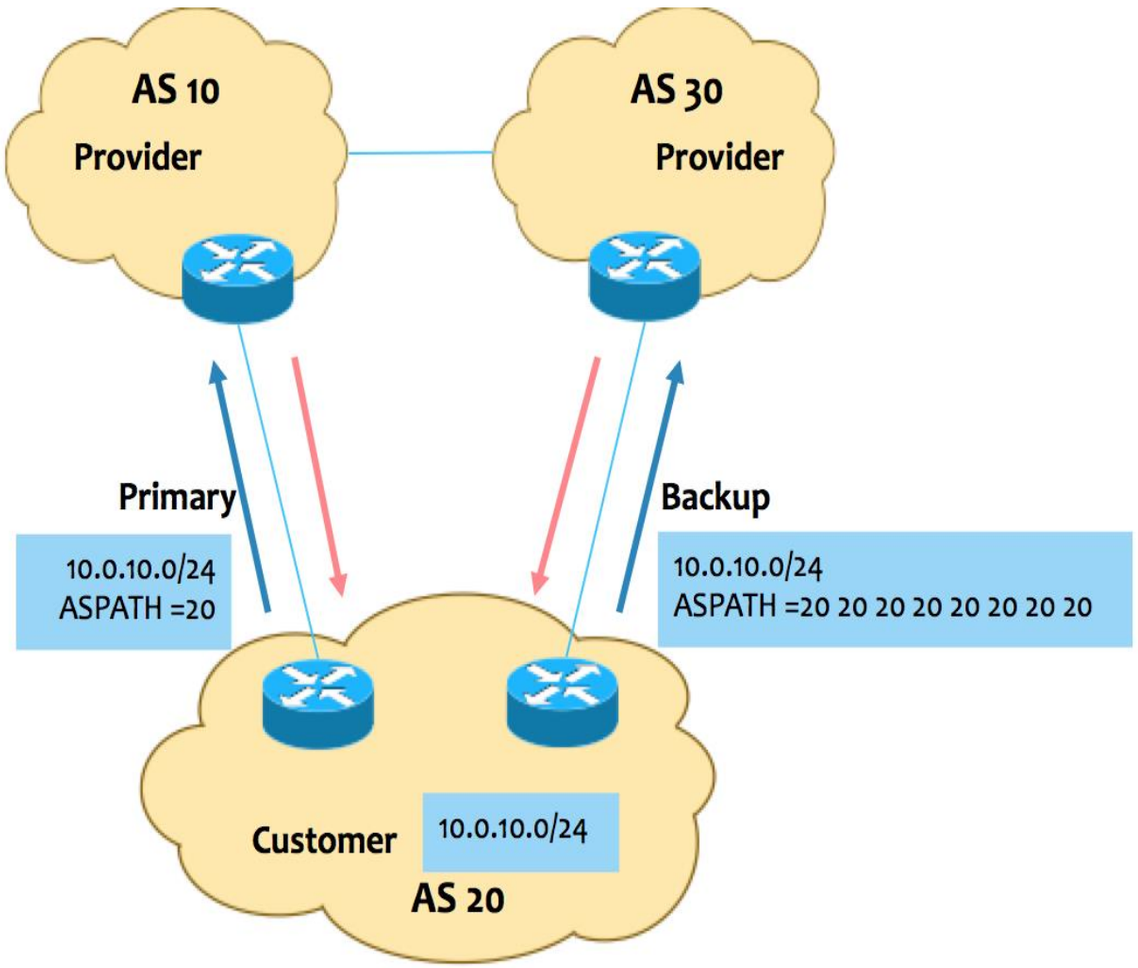
AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !



There are some challenges with BGP as-path prepending when it is used in multi-homed BGP setup

- Customer AS 20 is connected to two Service Providers. Customer is sending 10.0.10.0/24 prefix to both ISP
- They are advertising this prefix to their upstream ISPs and also each other through BGP peering
- AS 30 wants to be used as backup. Thus Customer is sending the 10.0.10.0/24 prefix towards AS30 with As-path prepends. Customer prepends its own AS path with 7 more AS
- You might think that link from AS 30 won't be used anymore so it will be used as backup. But that's not totally true !

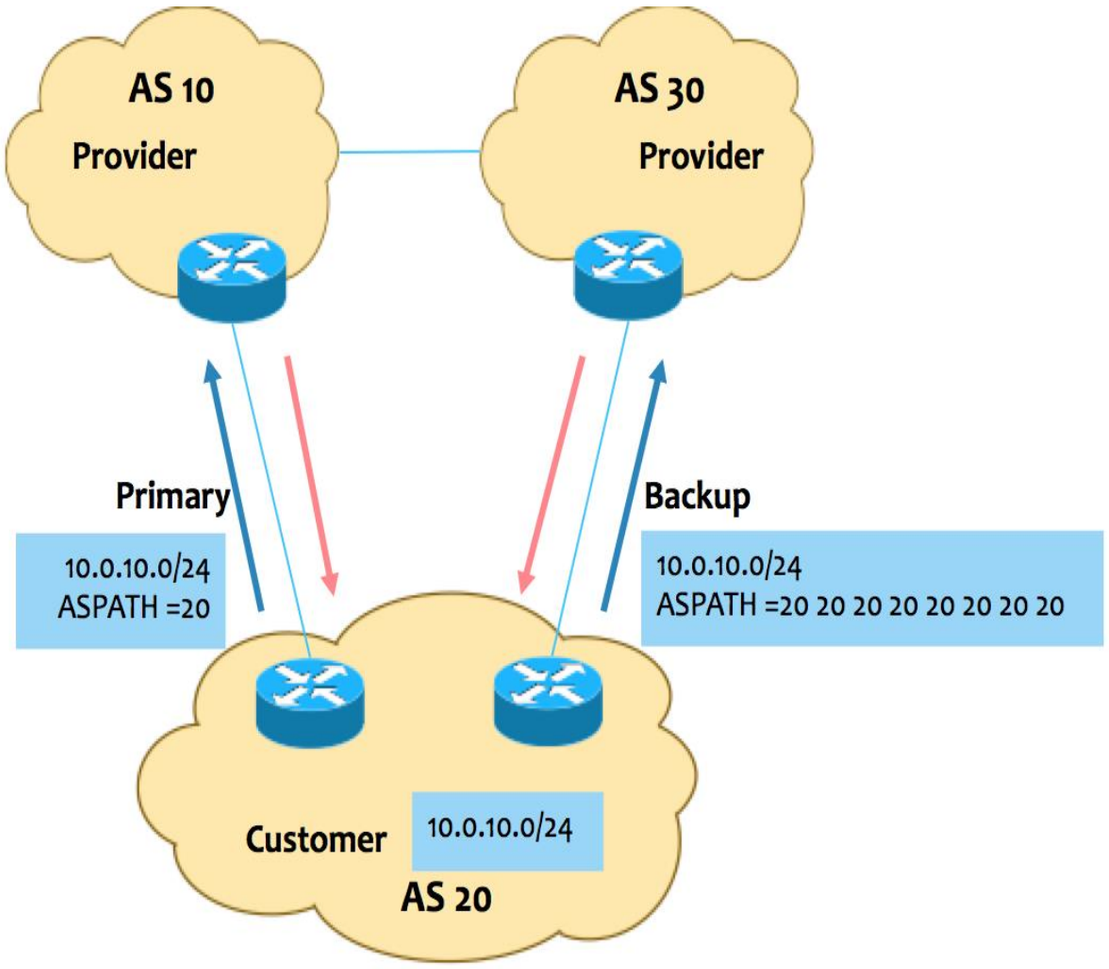
AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !



- Traffic from their upstream ISPs will go to the AS 10 because all the other ASes over Internet will see the advertisement from AS 30 with lots of prepends. So far so good

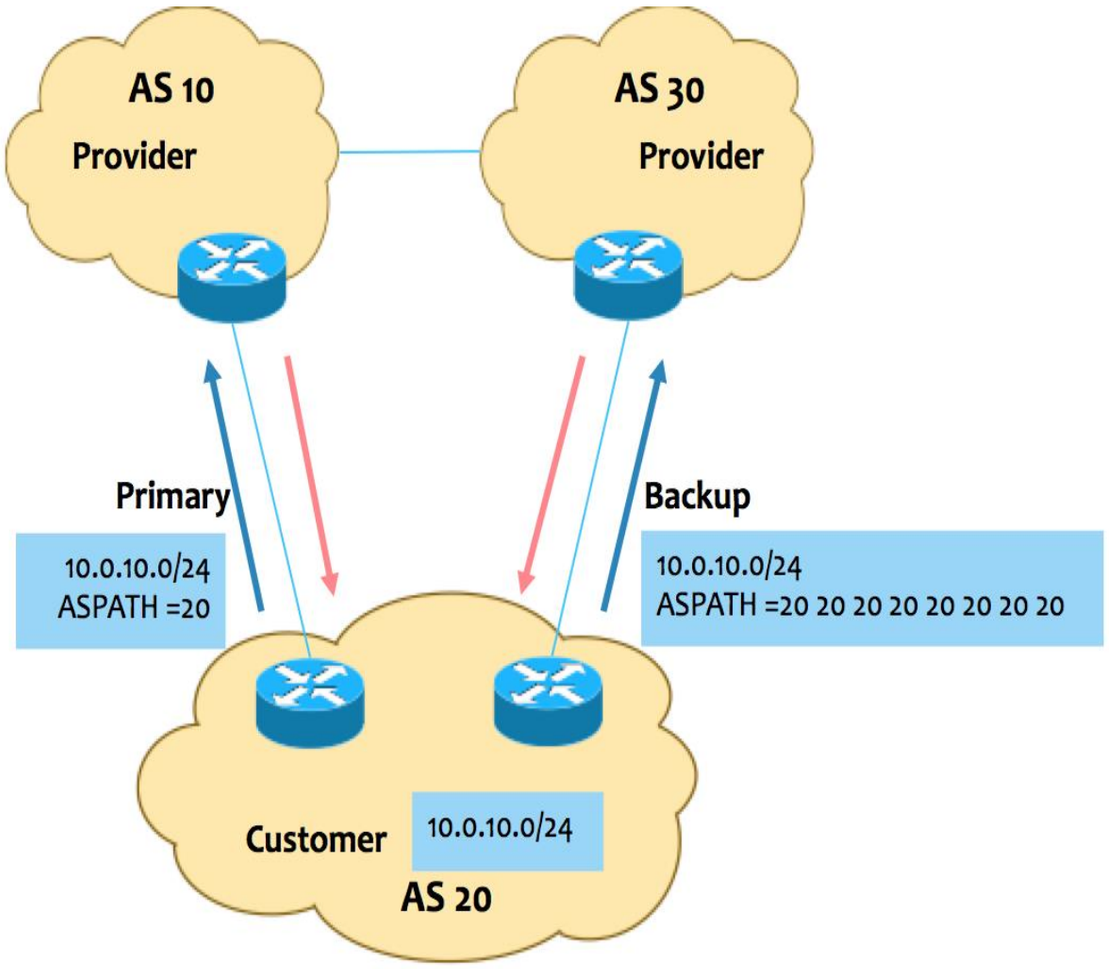
- But all the customers of AS 30 will still send the traffic for 10.0.10.0/24 prefix over the link which wants to be used as backup, although AS 30 learns 10.0.10.0/24 prefix over BGP peering link with AS 10 as well, its upstream providers as well.

AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !



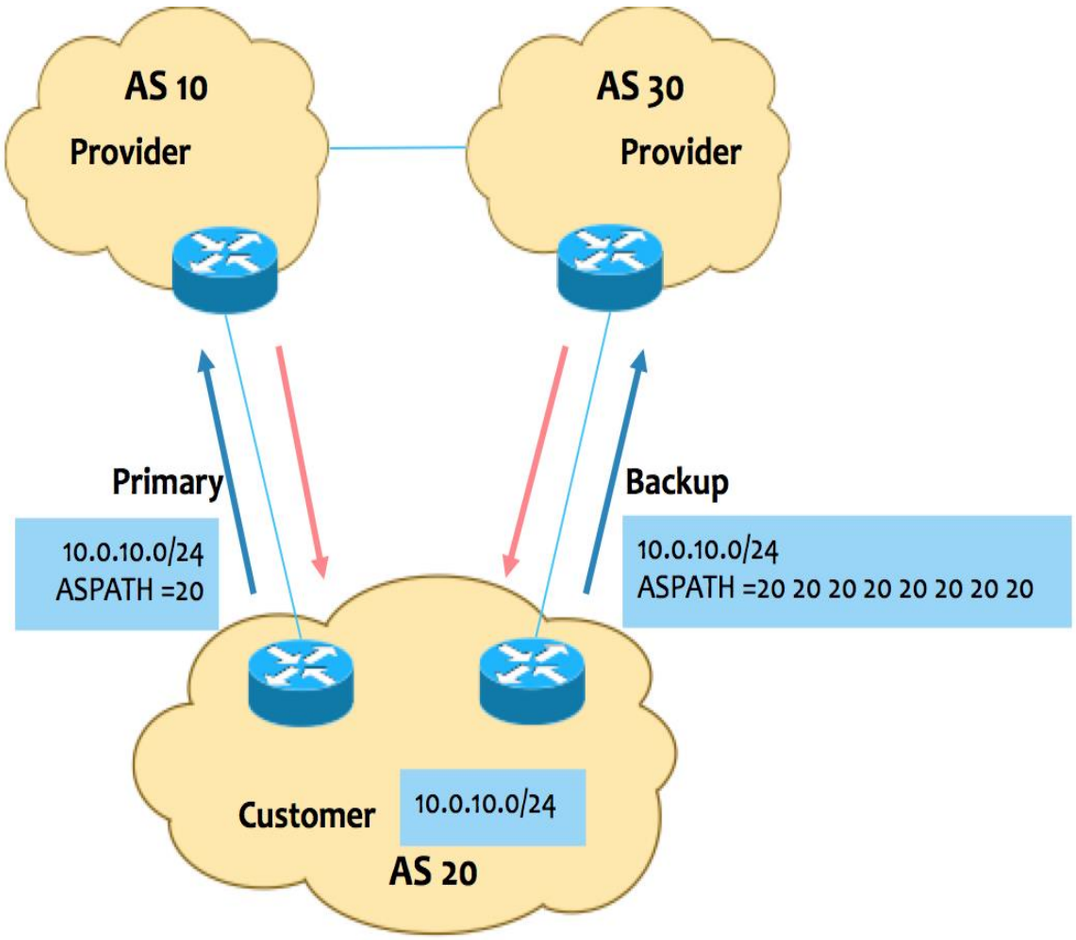
- Service Providers chooses to send the traffic for their customer prefixes over the customer link first, then peering links, lastly through upstream ISP. Because they want to utilize the customer link as much as possible to charge more money

AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !



- Service Providers implement Local Preference attribute to achieve this. Basic local .preference policy could be; Local Preference 100 towards Customer, local Preference 90 towards peering link and Local Preference 80 towards upstream ISP

AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !

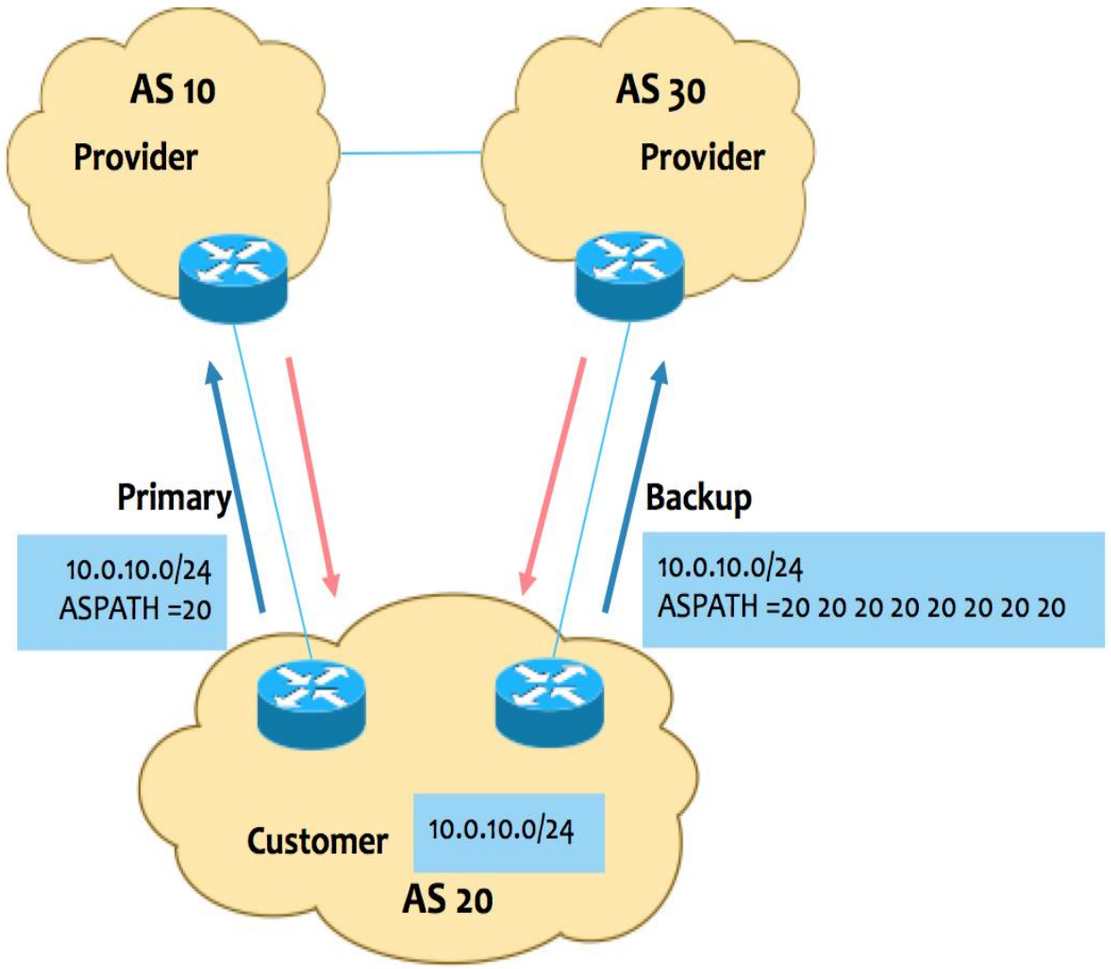


- Customer of AS 30 would still use customer link for 10.0.10.0/24 prefix although customer wants that link to be used as backup
- AS 20 is sending that prefix with AS-path-prepends but service provider implements Local Preference for that prefix

- Since Local Preference attribute is more important in the BGP best path selection process, if the traffic comes to any of the BGP routers of AS 30, it is sent through customer link. Not through BGP peering link with AS 10 or any upstream provider of AS 30

This problem can be solved with BGP community

EBGP Inbound Traffic Engineering with Community Attribute



Instead of prepending AS,BGP community attribute technique should be used instead of prepending AS, if the topology is multi homed BGP!

- AS 20 sends 10.0.10.0/24 prefix with the BGP community which changes local preference value of AS 30, link between customer and AS 30 is not used anymore.

- As an example AS 20 could send the community as 30:70 which reduces the Local Preference to 70 for the AS 20 prefixes over the customer BGP session, AS 30 would start to use BGP peer link to reach to 10.0.10.0/24 prefix

EBGP Inbound Traffic Engineering with Community Attribute

- Community attribute is sent over the BGP session between BGP Peers. Upon receiving the prefixes BGP peer can take an action for their predefined communities

- ISPs publish their supported community attribute values. For example they can say that if my customer send the prefixes with the attached 5000:110 community I will apply Local preference 110 towards that circuit.

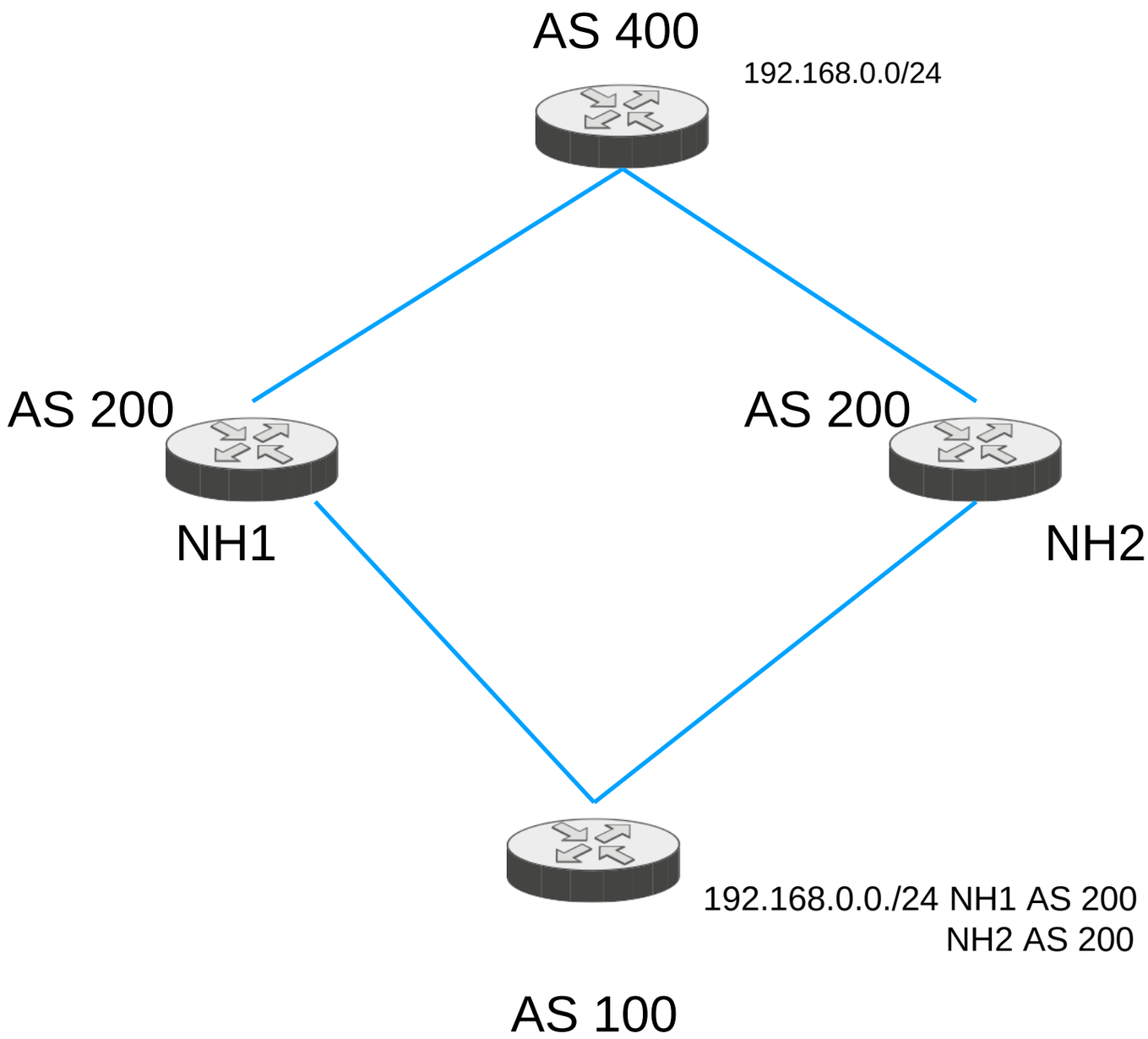
(Level 3 (Tier 1 ISP) Community Values and Corresponding Local Preference)

BGP Multipath

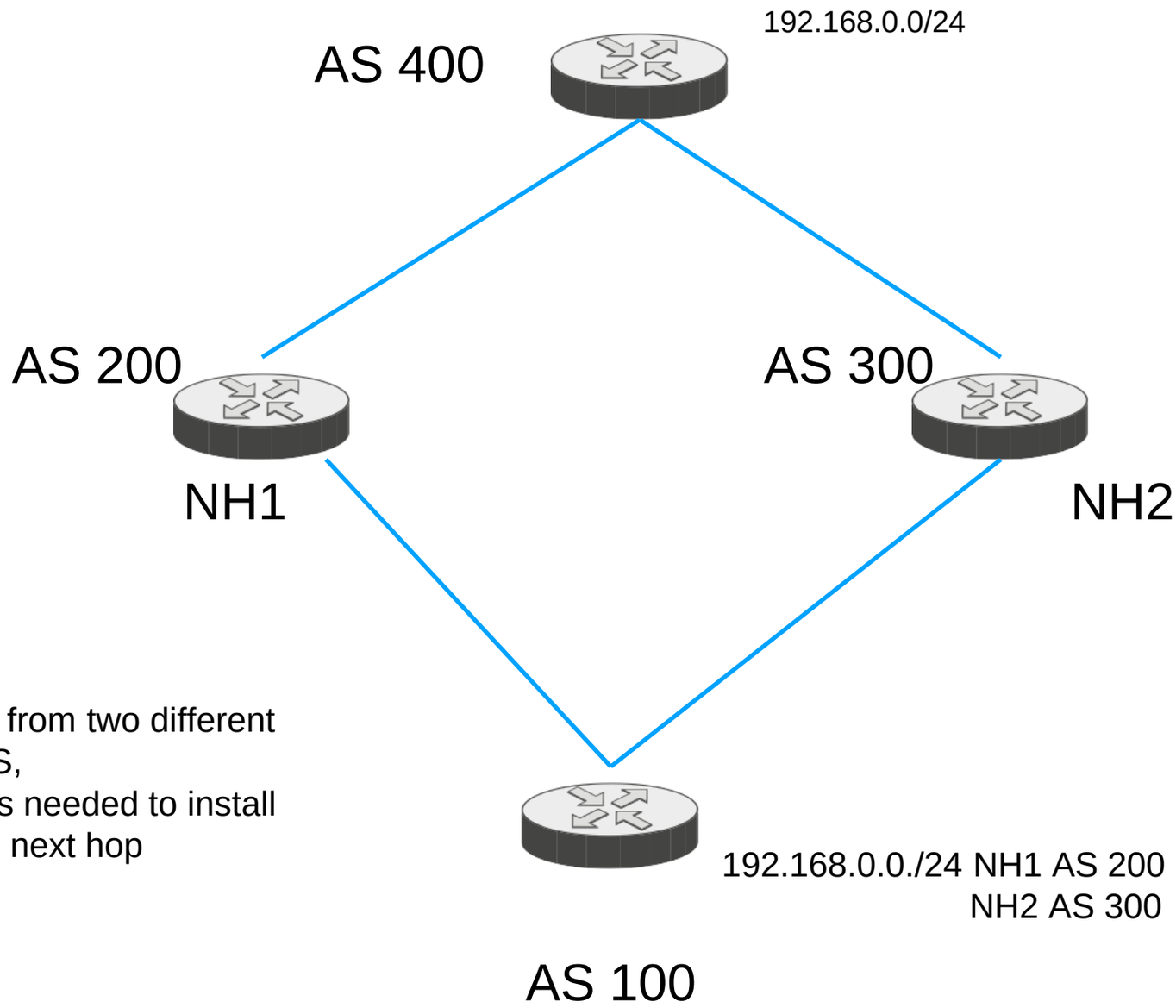
- BGP by default installs only single path in IBGP and EBGP deployment
- If prefixes is learned via multiple path, BGP supports multipath for IBGP , EBGP or across both IBGP and EBGP via EIBGP Multipath feature
- Multipath feature should be enabled manually

EBGP Multipath

EBGP Multipath works by default only if next hops from the single EBGP AS



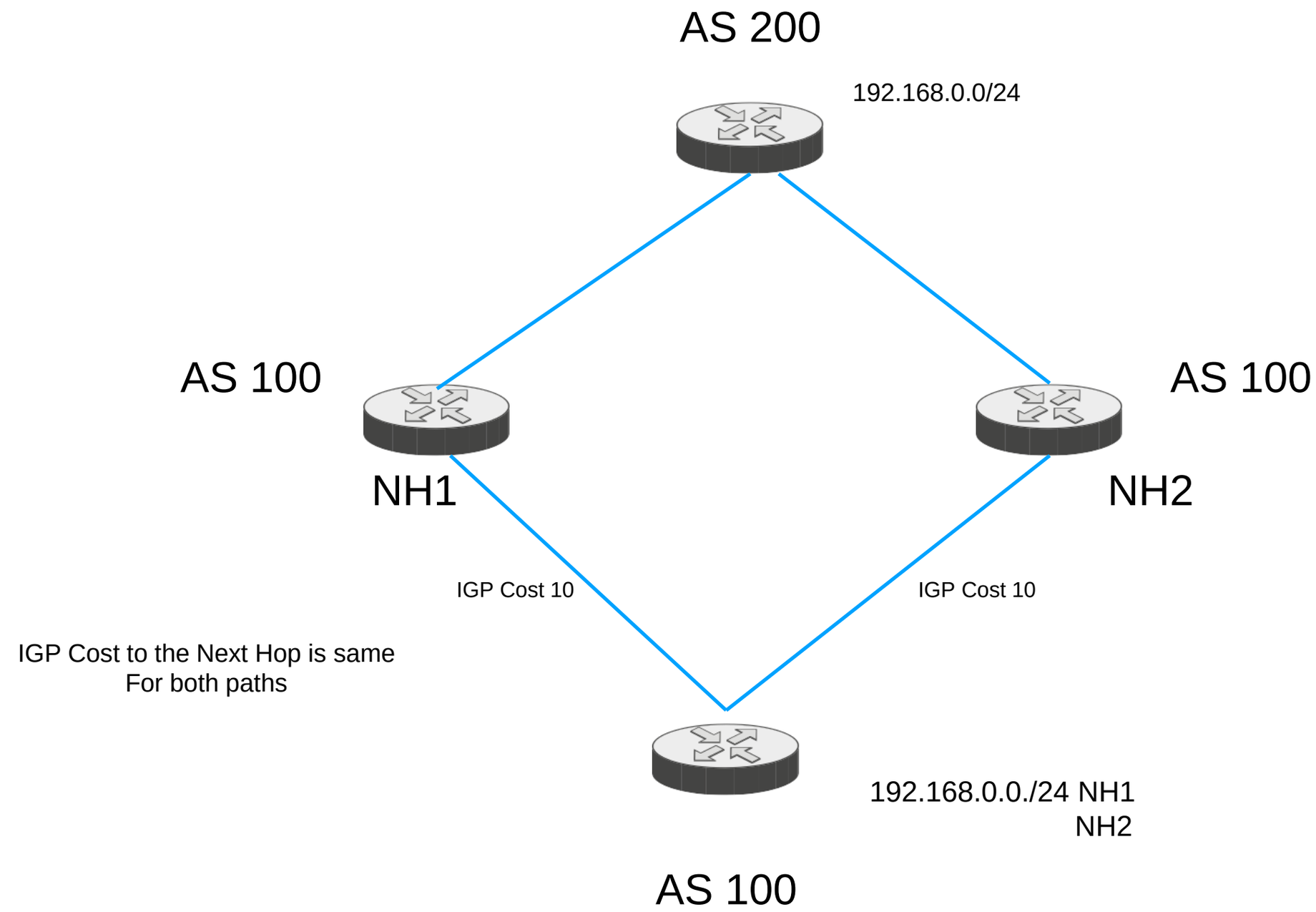
Two different AS requires ‘ multipath-relax ‘



When prefix is advertised from two different EBGP AS,
as-path relax command is needed to install the prefix via two next hop

“bgp bestpath **as-path** multipath-**relax**”

IBGP Multipath



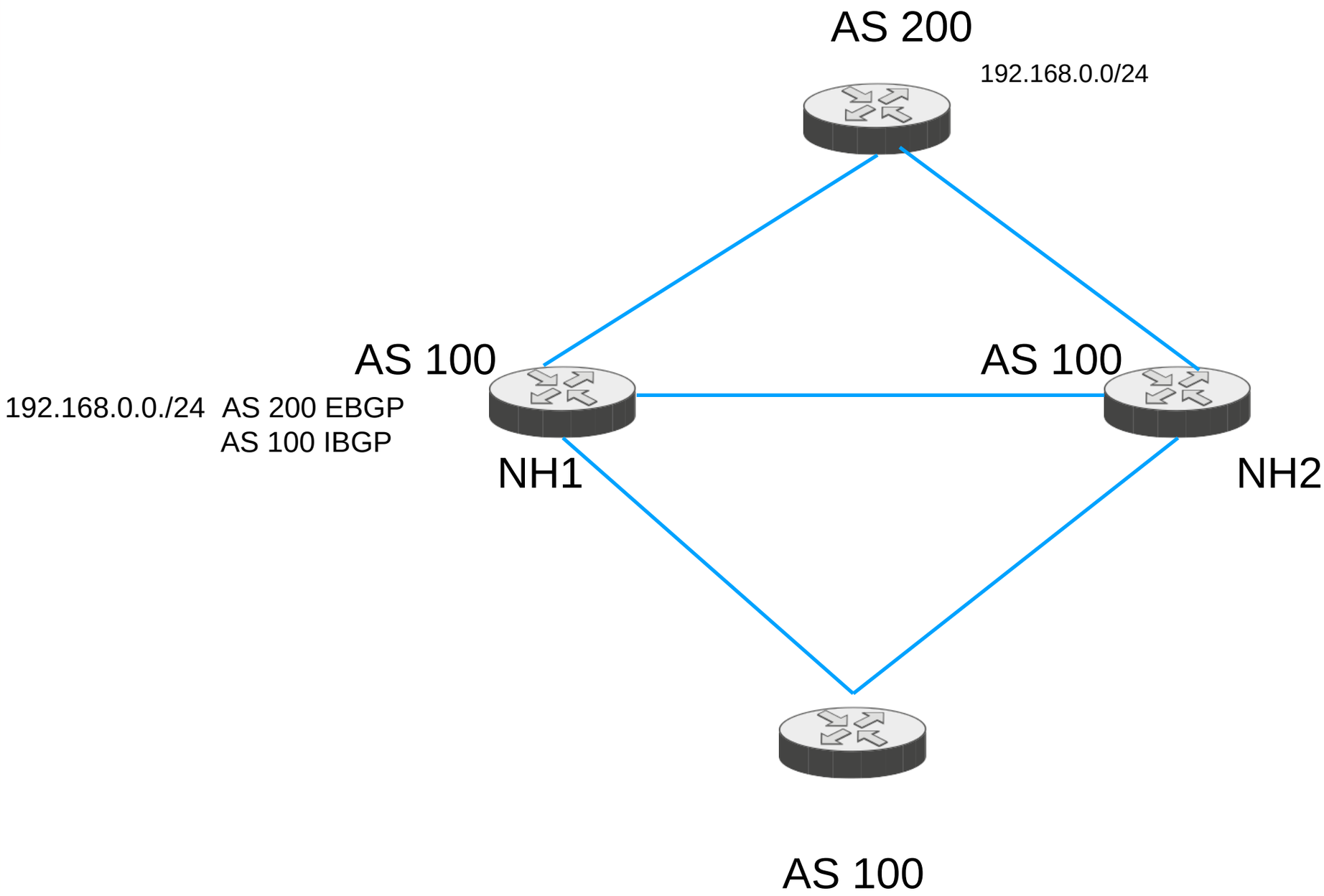
EIBGP Multipath

- BGP Best path selection algorithm prefers EBGP paths over IBGP paths

- This prevents having both IBGP and EBGP prefixes to be installed in the routing table at the same time

- EIBGP multipath feature allows same prefix to be installed both with IBGP and EBGP next hops

EIBGP Multipath



EIBGP Multipath

- EIBGP Multipath feature can create routing loop , that is why important to understand it from design point of view

- It is typically used in MPLS L3 VPN deployments

Inter-domain Routing

- To understand BGP peering, first we must understand how network is connected to each other on the Internet

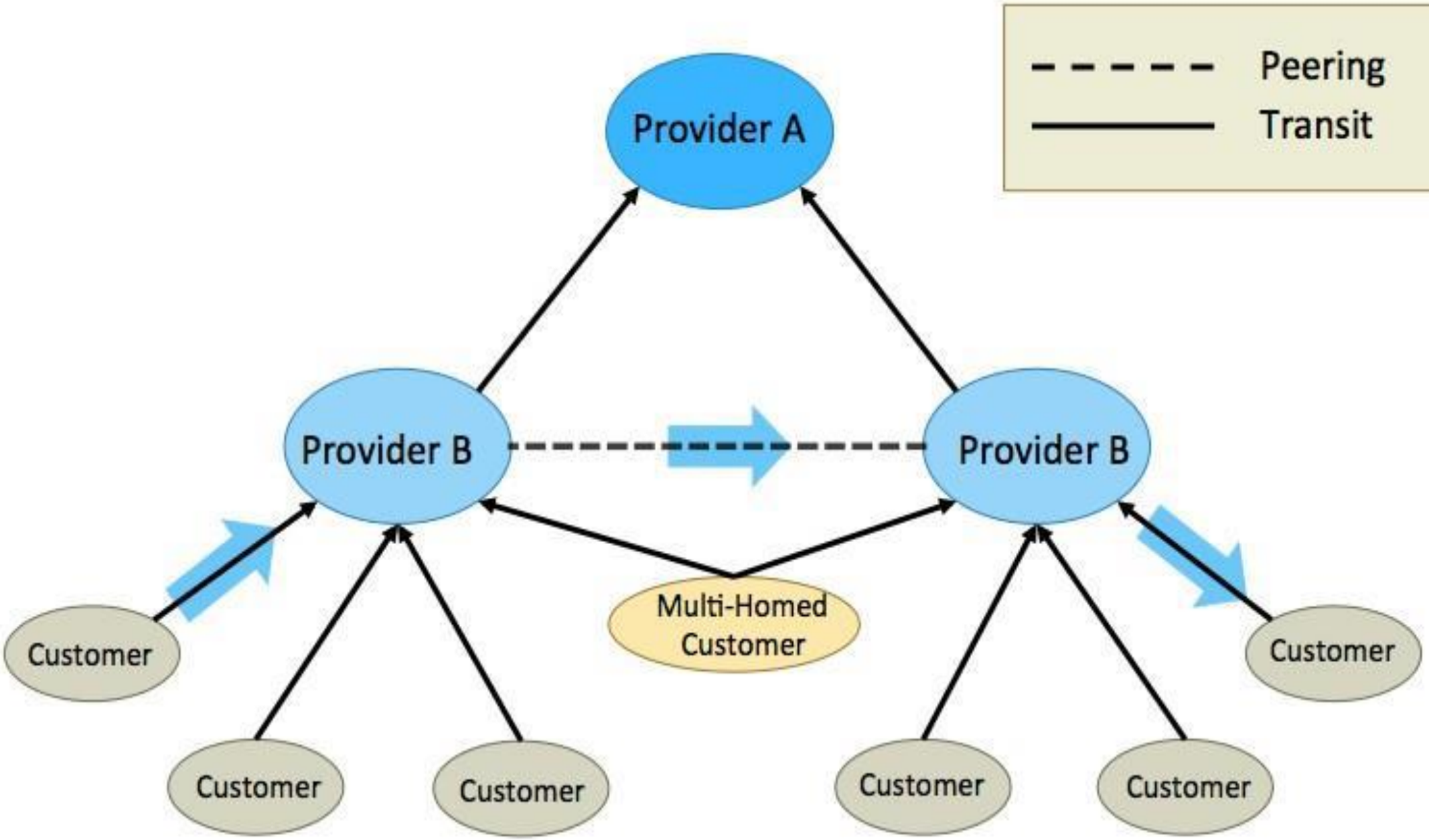
- The Internet is a collection of many individual networks, which interconnect with each other under the common framework of ensuring global reachability between any two points

There are 3 primary relationships for this interconnection

- **Provider** – Typically someone you pay money to, who has the responsibility of routing your packets to/from the entire Internet

- **Customer** – Typically someone who pays you money, with the expectation that you will route their packets to/from the entire Internet

- **Peers** – Two networks who get together and agree to exchange traffic between each others' networks, typically for free. There are two types of peering in general , Private and Public which will be explained later


Peering
Transit
Provider A
Provider B
Provider B
Multi-Homed
Customer
Customer
Customer
Customer
Customer
Customer
Customer

Settlement Free Peering

- Peering is a BGP session between the two Routers. When different companies have Peering with each other, they exchange network traffic over the peering session

There are three reason to have BGP peering on Internet:

- Company wants to receive an Internet service

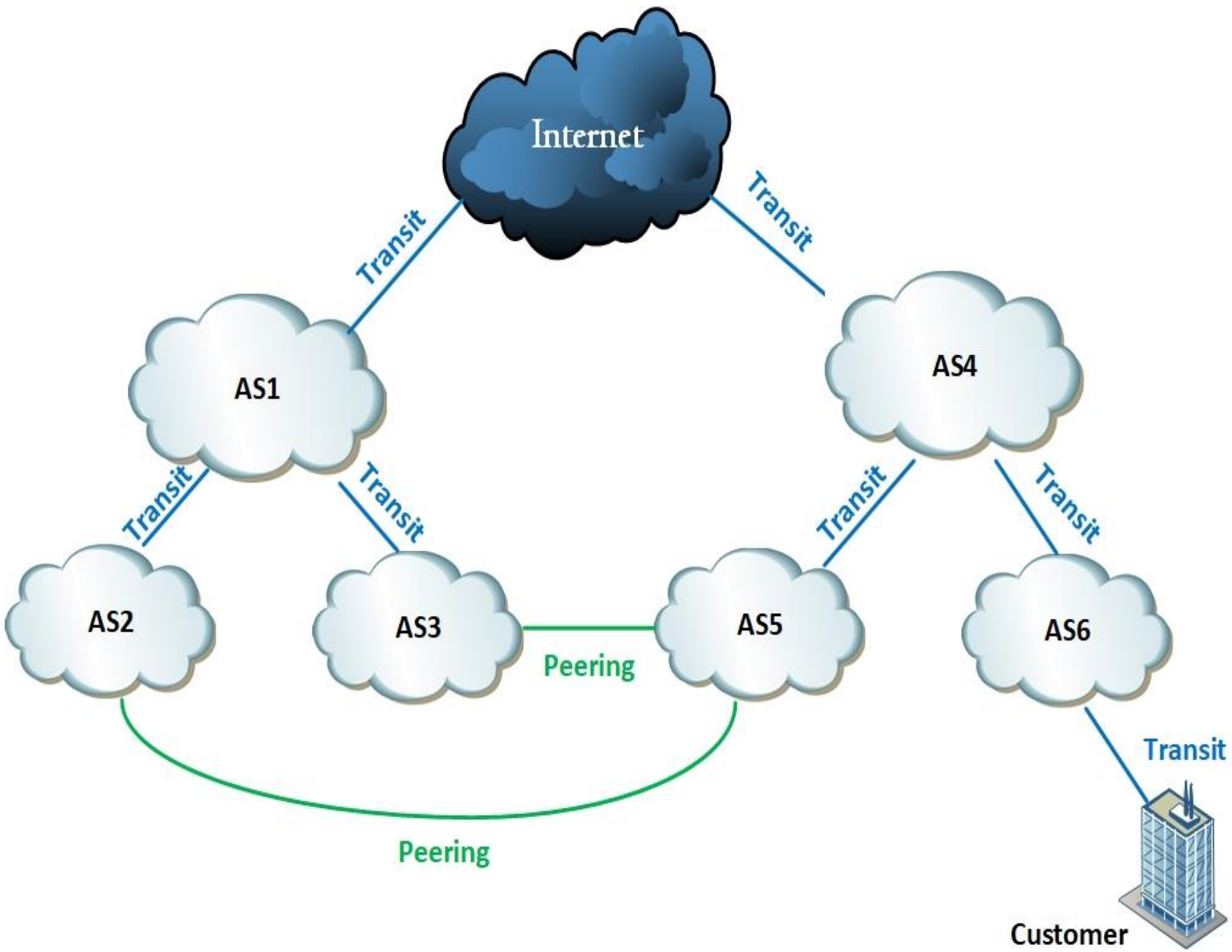
- Company wants to sell an Internet service

- Two companies exchange their customer prefixes and exchange network traffic but don’t pay each other, which is called Settlement Free Peering

- Settlement Free Peering is also referred as Settlement Free Interconnection and here onwards, to make it short, SFI term will be used

- SFI is an agreement between different Service Providers. It is an EBGP neighborship between different Service Providers to send BGP traffic between them without paying upstream Service Provider

Business relationship between the networks.



Private BGP peering

- Private Peering is a direct interconnection between two networks, using a dedicated transport service or fiber. It is also known as bilateral peering in the industry. May also be called a Private Network Interconnect, or PNI.

- Inside a datacenter this is usually a dark-fiber cross-connect. May also be a Telco-delivered circuit as well.

- If there is big amount of traffic between two networks, private peering makes more sense than public peering (Avoid some extra hops/network devices)

Private Peering

- Private peering can be setup inside Internet Exchange Point as well (Internet Exchange Point will be explained)

- Larger companies generally use Private peering rather than Public peering since they want to select who they are going to be peer with and the amount of traffic between them is large, they don’t want to exchange traffic with everyone by joining to the Public Peering

Public BGP Peering

- Typically, public peering is done at the Internet Exchange Point. BGP Route servers are used in public peering to improve scalability

- Multilateral Peering is commonly adopted in Public Peering (Not only two network peer with each other but it can be hundreds of networks)

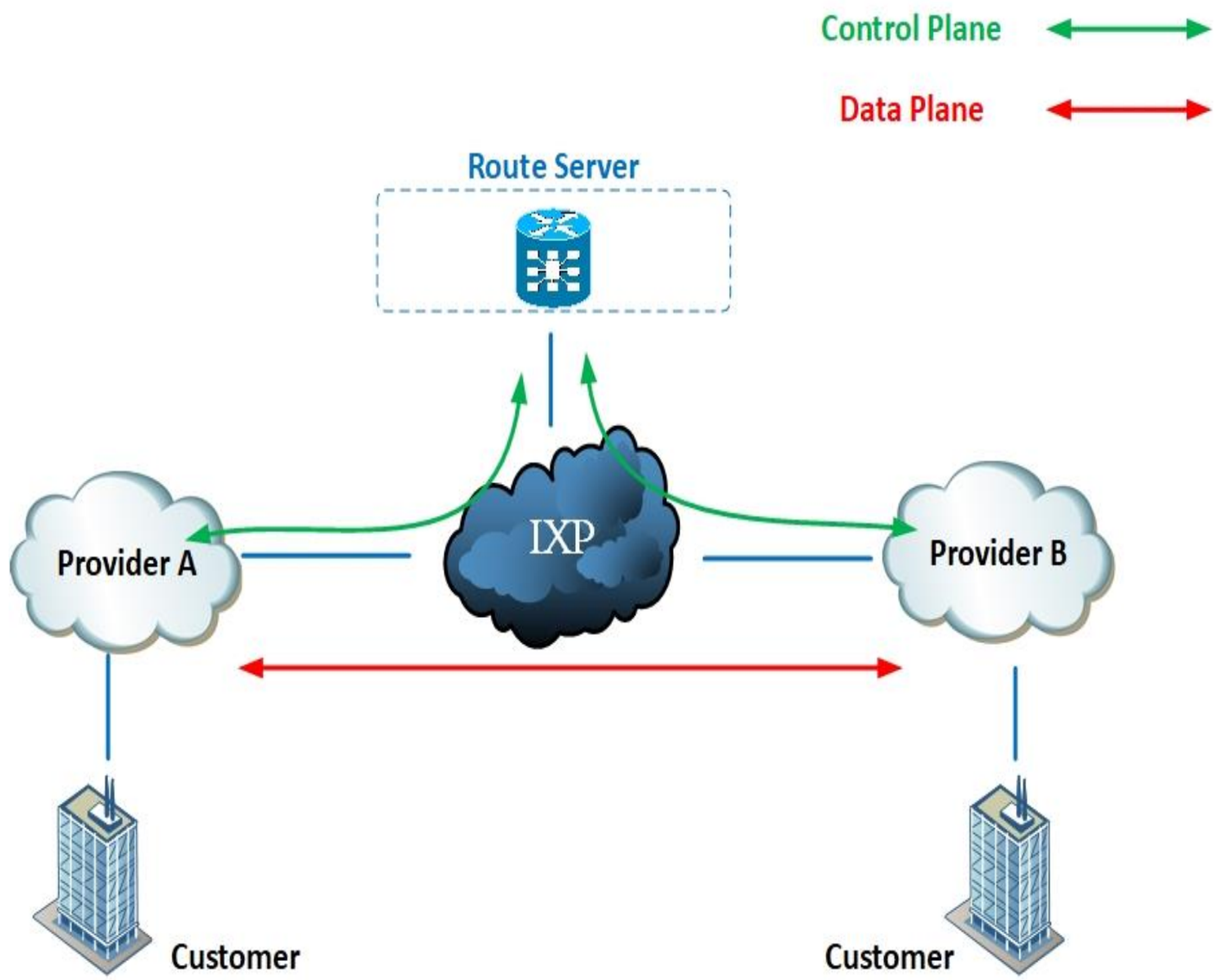
BGP Route Server

- BGP Route Server is used at the Internet Exchange Point to simplify BGP Peering process. Instead of managing, maintaining hundreds of Peering sessions in large Internet Exchange Point, BGP Route Server is used

- Every BGP speaking router has a BGP session with BGP Route Server

- Route Server doesn’t change the BGP Attributes, although the type of BGP Peering session is EBGP (Similar to Route Reflector in IBGP)

BGP Route Server



BGP Route Server

- BGP Route Server doesn’t change the next-hop to itself, thus it is used only as a Control Plane device, not a Data Plane. Which means, actual traffic is passed between the companies that participant to the Public Peering Internet Exchange Point , traffic doesn’t go through the Route Server

- It is very similar to BGP Router Reflector which is used in IBGP topologies. The difference is, BGP Route Server is used in EBGP

Bilateral Peering

- When two networks negotiate with each other and establish a peering session directly, this is called Bilateral Peering. Generally done when there is a big amount of traffic between two networks

- Also Tier 1 Operators just do Bilateral peering as they don’t want to peer with anyone other than other Tier 1 Operators. Rest of the companies are their potential customers, not their peers

Multilateral Peering

- Bilateral peering offers the most control, but some networks with very open peering policies may wish to simplify the process, and simply “connect with everyone”. To help facilitate this, many Exchange Points offer “multilateral peering exchanges”, or an “MLPE”

- An MLPE is typically an exchange point that offers “route-server”, allowing a member to establish a single BGP session and receive routes from every other member connected to the MLPE

Multilateral Peering

- Effectively, connecting to the MLPE is the same as agreeing to automatically peer with everyone else connected to the MLPE, without requiring the configuration of a BGP session for every peer

- Basically, Public Peering and MLPE is almost the same thing and used mostly interchangeably

Looking Glass

- It is a server commonly deploy by an IXP to provide a view to the prefixes in the specific IXP

- It gives publicly available information so any network owner can check the available prefixes in the IXP before they decide to join to that particular IXP

Looking Glass

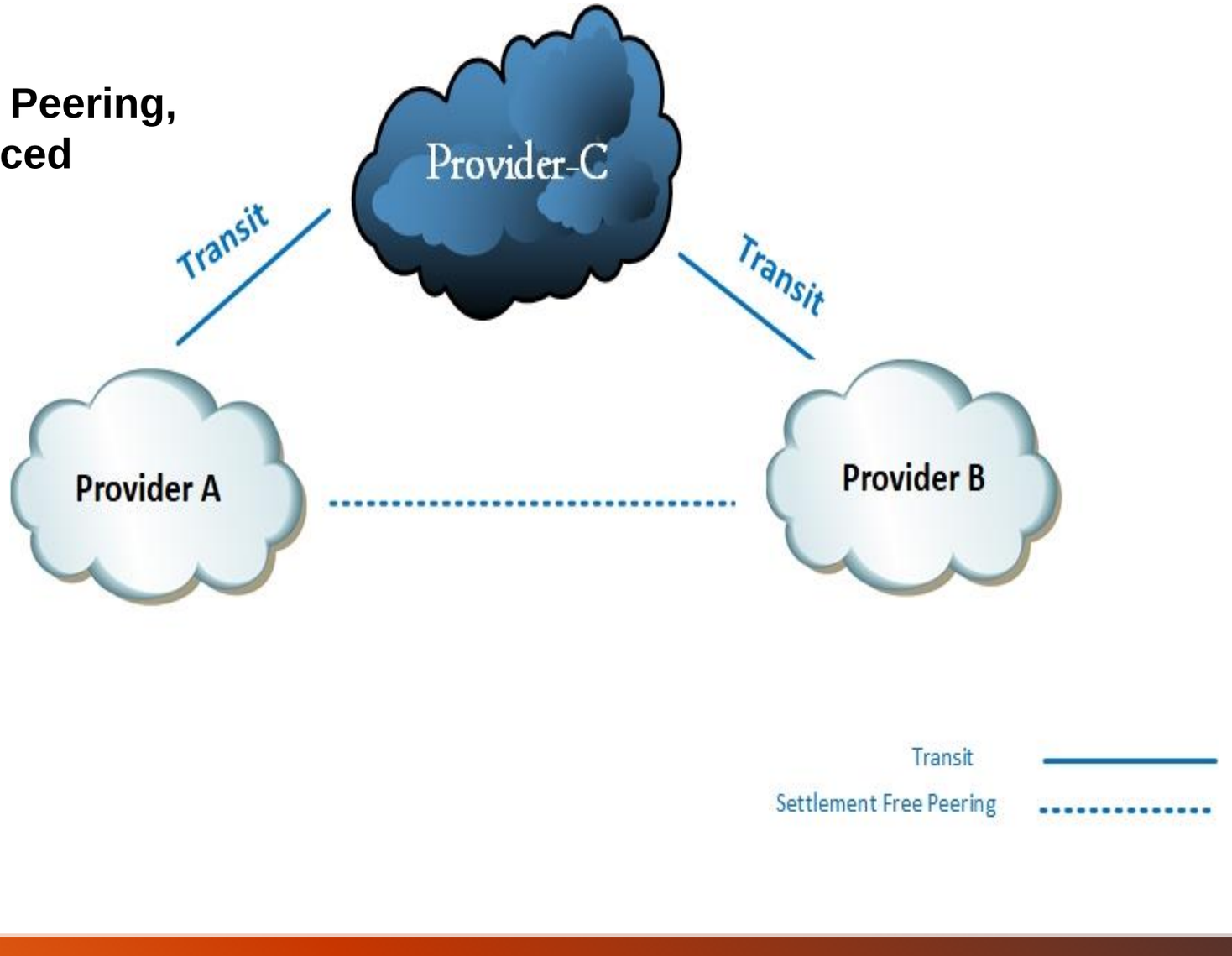
- There are many publicly available Looking Glasses in the World. They are configured as read-only so person who wants to check the particular looking glass cannot change the BGP routing information

- From http://www.bgplookingglass.com you can see the Looking Glass database

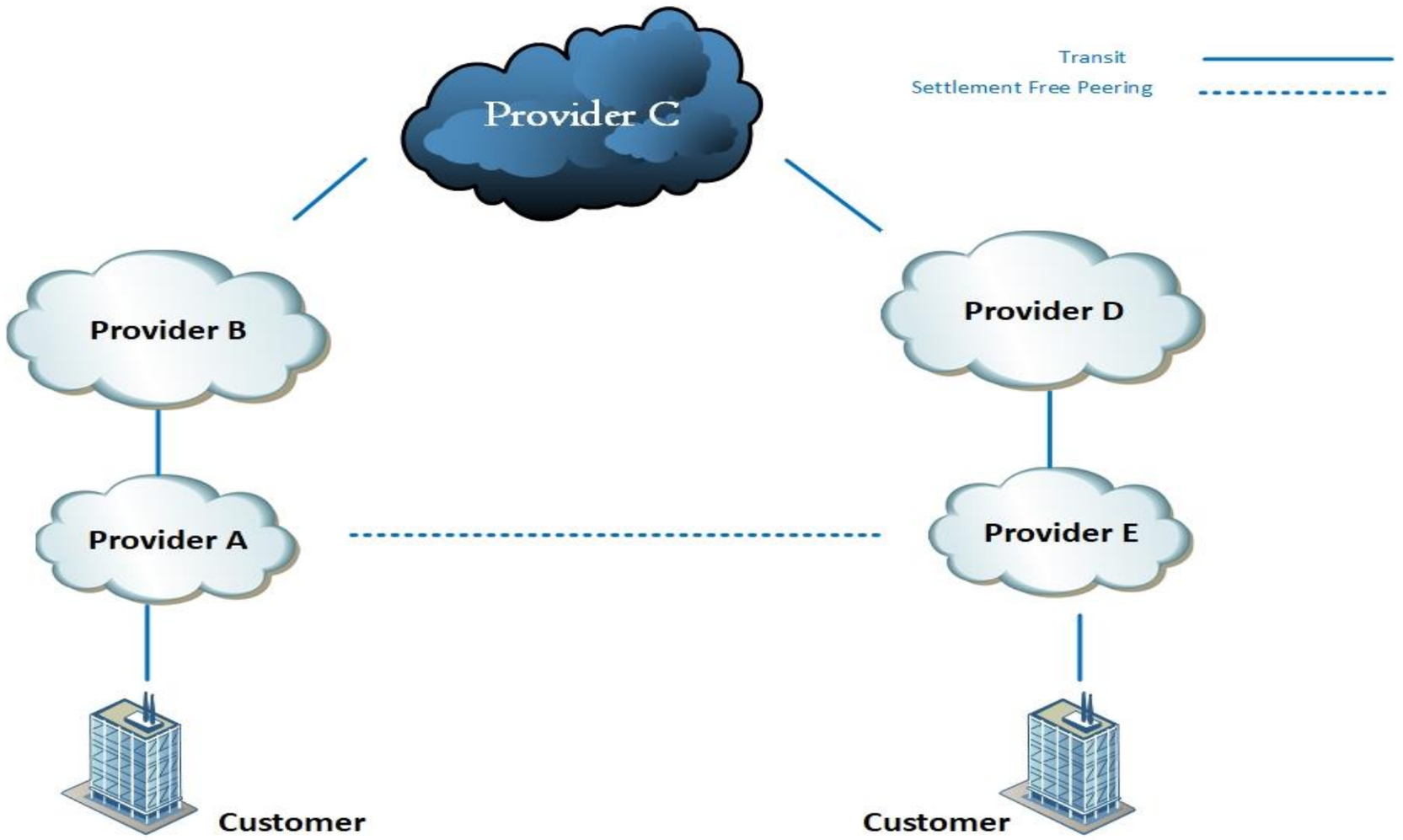
Benefits of Settlement Free Peering

- Reduced operating cost: A transit provider is not being paid to deliver some portion of your traffic. Peering traffic is free!

Through Settlement Free Peering,
Transit Cost is reduced



Improved routing: By directly connecting with another network with whom you exchange traffic, you are eliminating a middle-man and potential failure point



Settlement Free Peering Benefits

- Distribution of traffic: By distributing traffic over interconnections with many different networks, the ability to scale is potentially improved

- Almost every country has Internet Exchange Point where Service Providers, Content Networks, CDNs, Enterprises, Mobile Operators, Carriers, TLD (Top Level Domains) and Root DNS Servers can meet

What is IXP (Internet Exchange Point)?

- A layer 2 network where multiple network entities meet, for the purposes of interconnection and exchanging traffic with one another

- Internet Exchange Points start with a single Layer 2 switch at one location. Networks peer with each other in this facility

- When the number of participant grows, more switches are added at that location and more locations are added to the IXP itself. For example, AMS-IX in Netherlands have many places, inside many Datacenters and each Datacenter they have more than one switch for the Settlement Free Interconnection

What is IXP (Internet Exchange Point)?

- Often referred to as an Internet Exchange (IX), or “public peering”
- Today most Exchange Points are Ethernet based LANs, where all members sharing a common broadcast domain, and each member is given a single IP per router out of a common IP block (such as a /24)
- Typically done at the carrier neutral datacenters and many cases Datacenter owners provide racks for the peering fabric for free

Why Networks Peer at the IXP?

- An Exchange Point acts as a common gathering point, where networks who want to peer can find each other

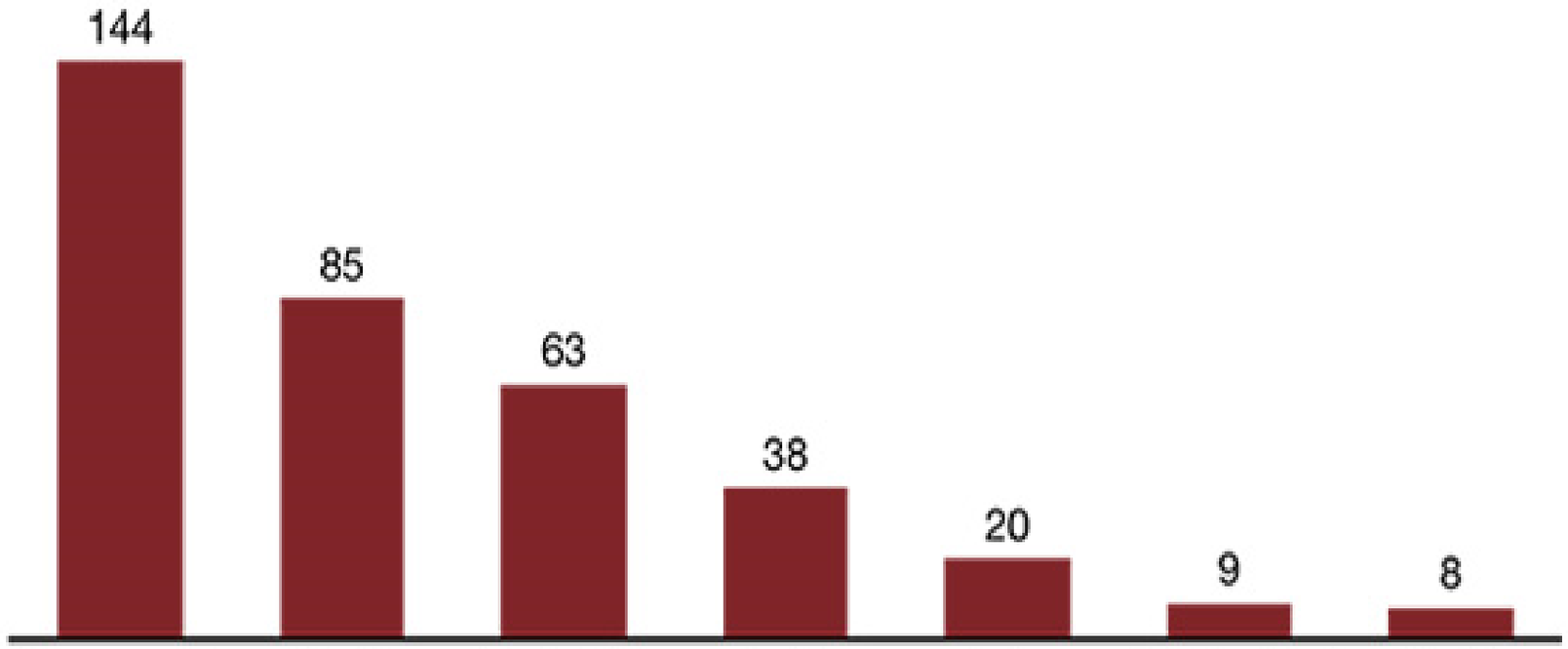
- A network new to peering will typically go to an exchange point as their first step, and be able to find so many other like-minded networks interested in peering with them

- The more members an exchange point has, the more attractive it becomes to new members looking to interconnect with the most other networks. This is called as “critical mass”

Where are the Internet Exchange Points?

- Most of the IXPs in the World in Europe. There are many IXP in North America as well
- IXPs in the Europe work mainly based on Membership model, IXPs in the U.S work based on Commercial model. There are exceptions in each case though
- Most European IXPs grew from non-commercial ventures, such as research organizations. Most African IXPs were established by ISP Associations and Universities

Number of IXPs in the World



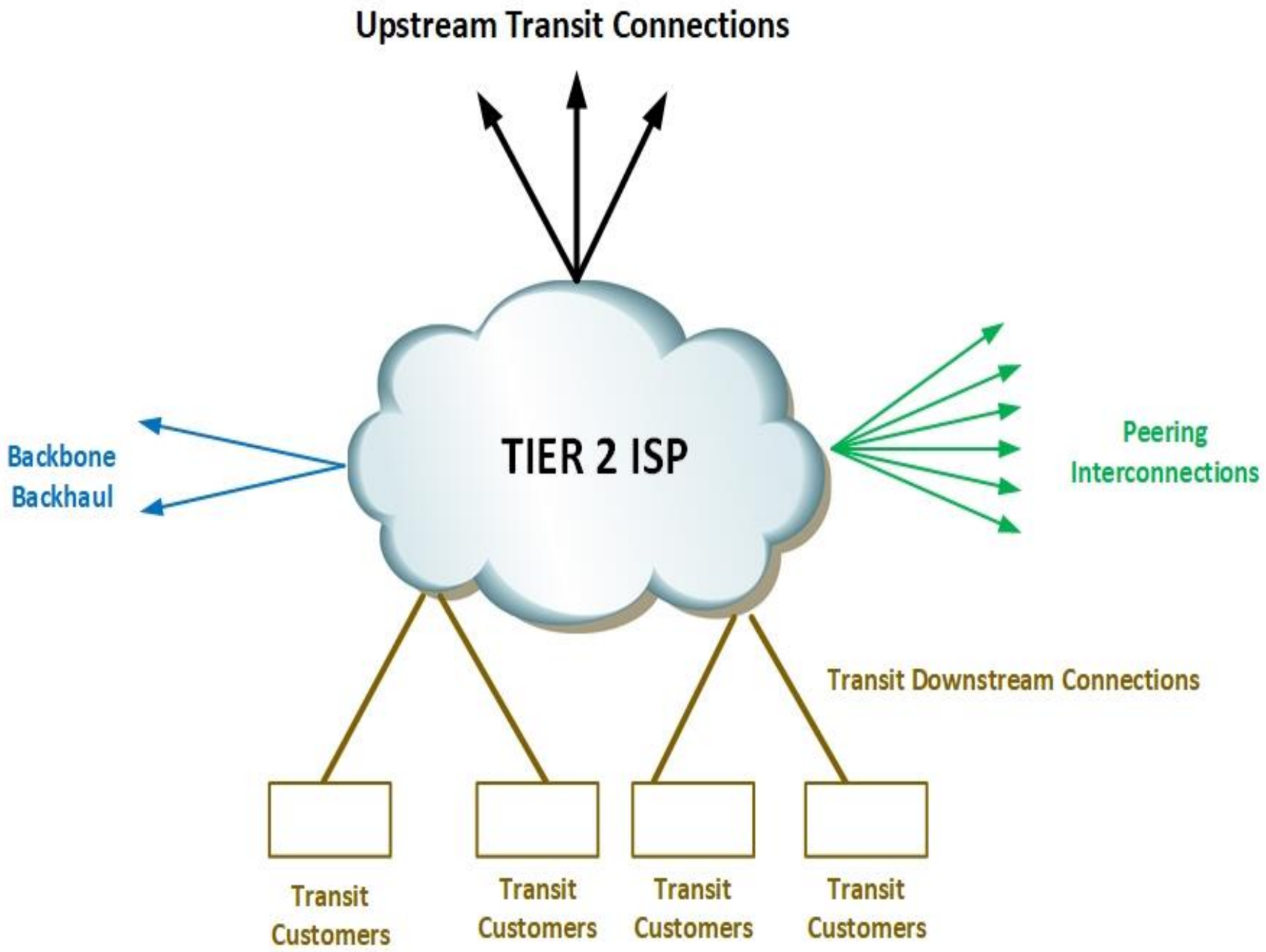
Source: ITU; Packet Clearing House; The World Bank, World Development Indicators

ISP Tiers – Tier 1, Tier 2, Tier 3 ISP

- Tier 1 Service Provider is a network which does not purchase transit from any other network and peers with every other Tier 1 network to maintain global reachability

- Tier 2 Service Provider is a network with transit customers and some peering, but still buys full transit to reach some portion of the Internet

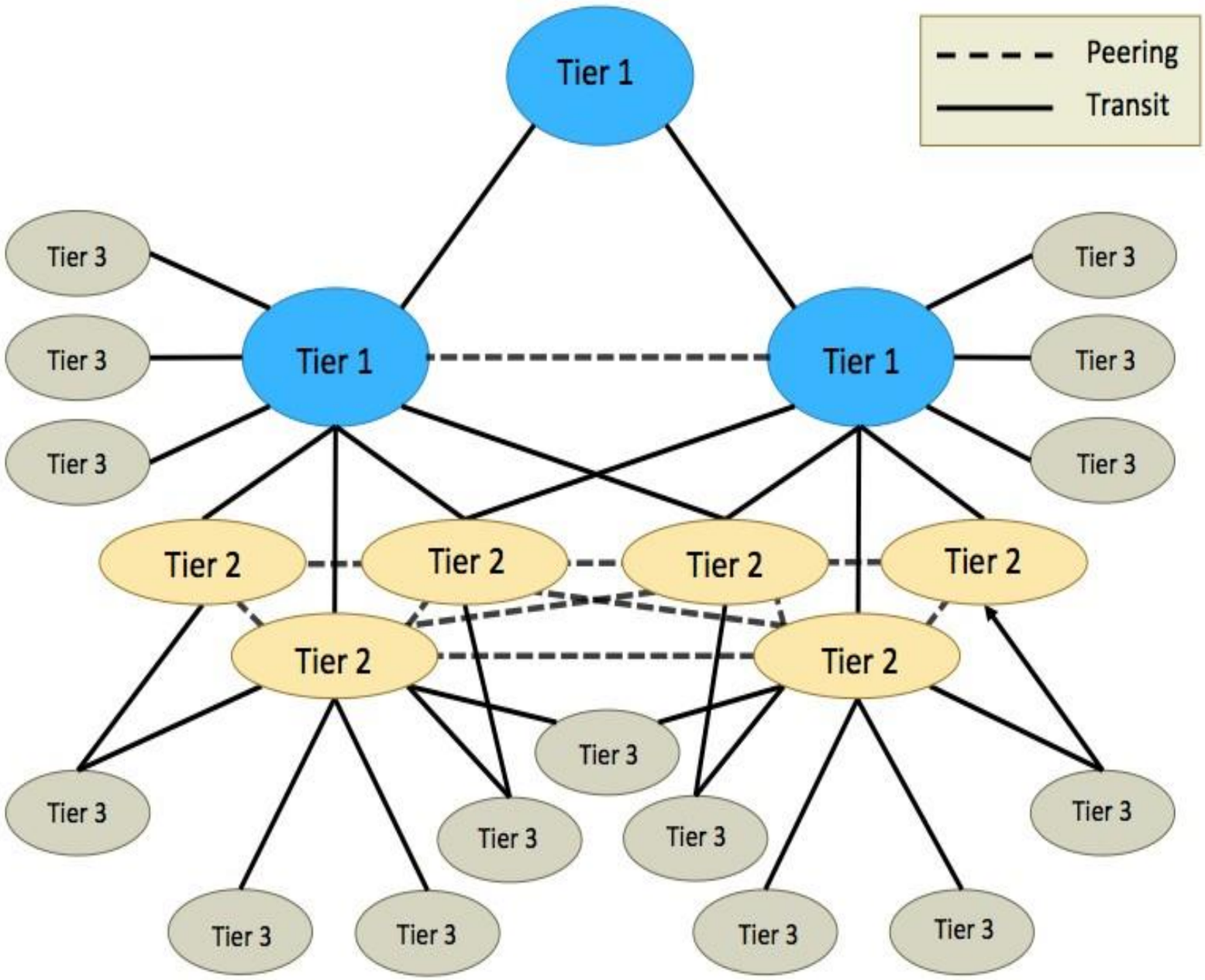
Tier 2 ISP and its Connections



Tier 3 ISP

- Tier 3 Service Provider is considered as stub network. They are generally considered as local/access ISPs. They don’t sell any IP Transit service to anyone. Sometimes, Tier 3 ISP definition is used to describe Enterprise, SMB or End Users

Tier 1 , Tier 2 and Tier 3 ISP Relationship

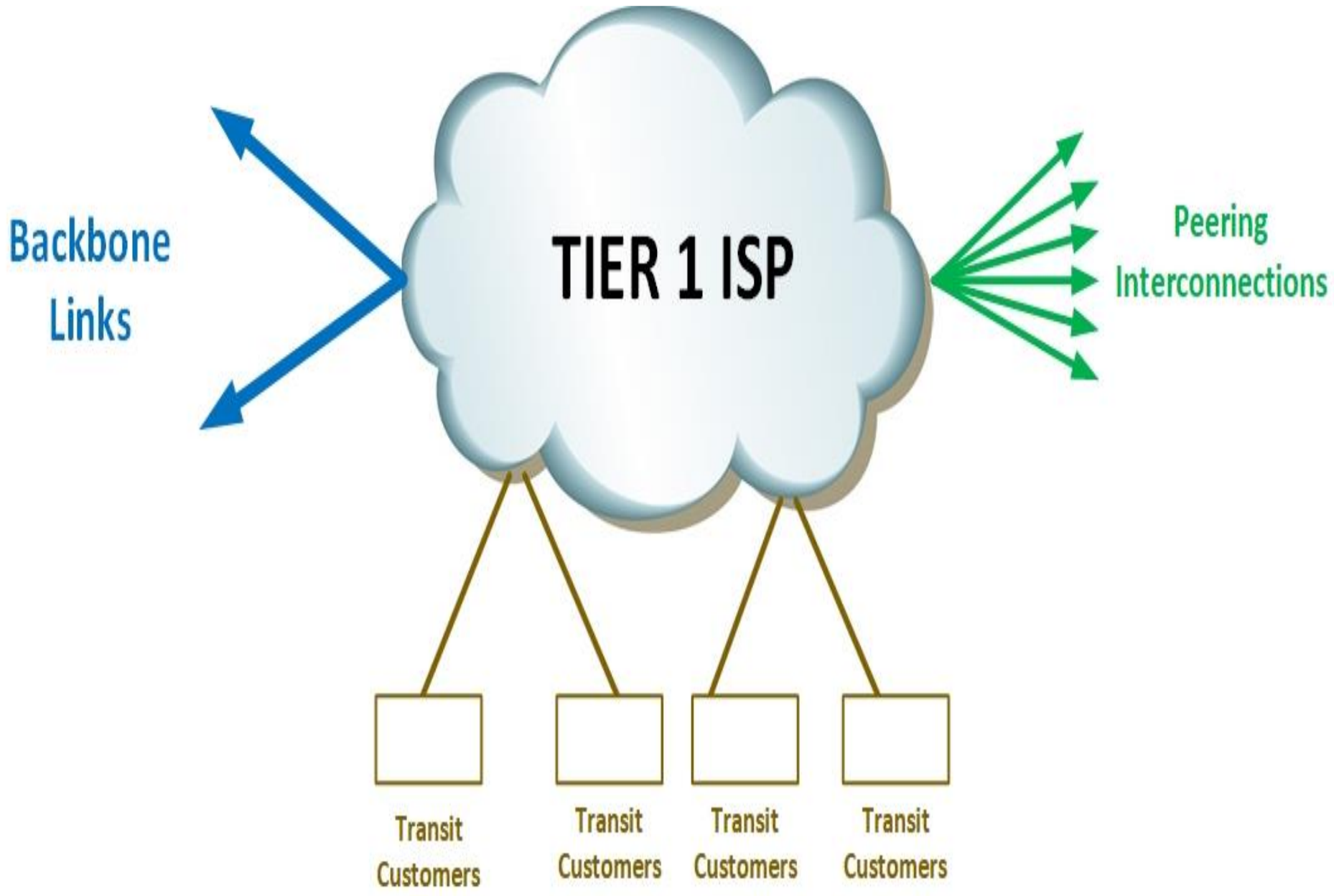


Tier 1 Service Providers

- A Tier 1 ISP is an ISP that has access to the entire Internet Region solely via its settlement free peering relationship

- Tier 1 ISPs only peer with other Tier 1 ISPs and sometimes with CDN and the Search Engines. They don’t have any Transit ISP but they are the top tier ISP

Tier 1 ISP and its Connections

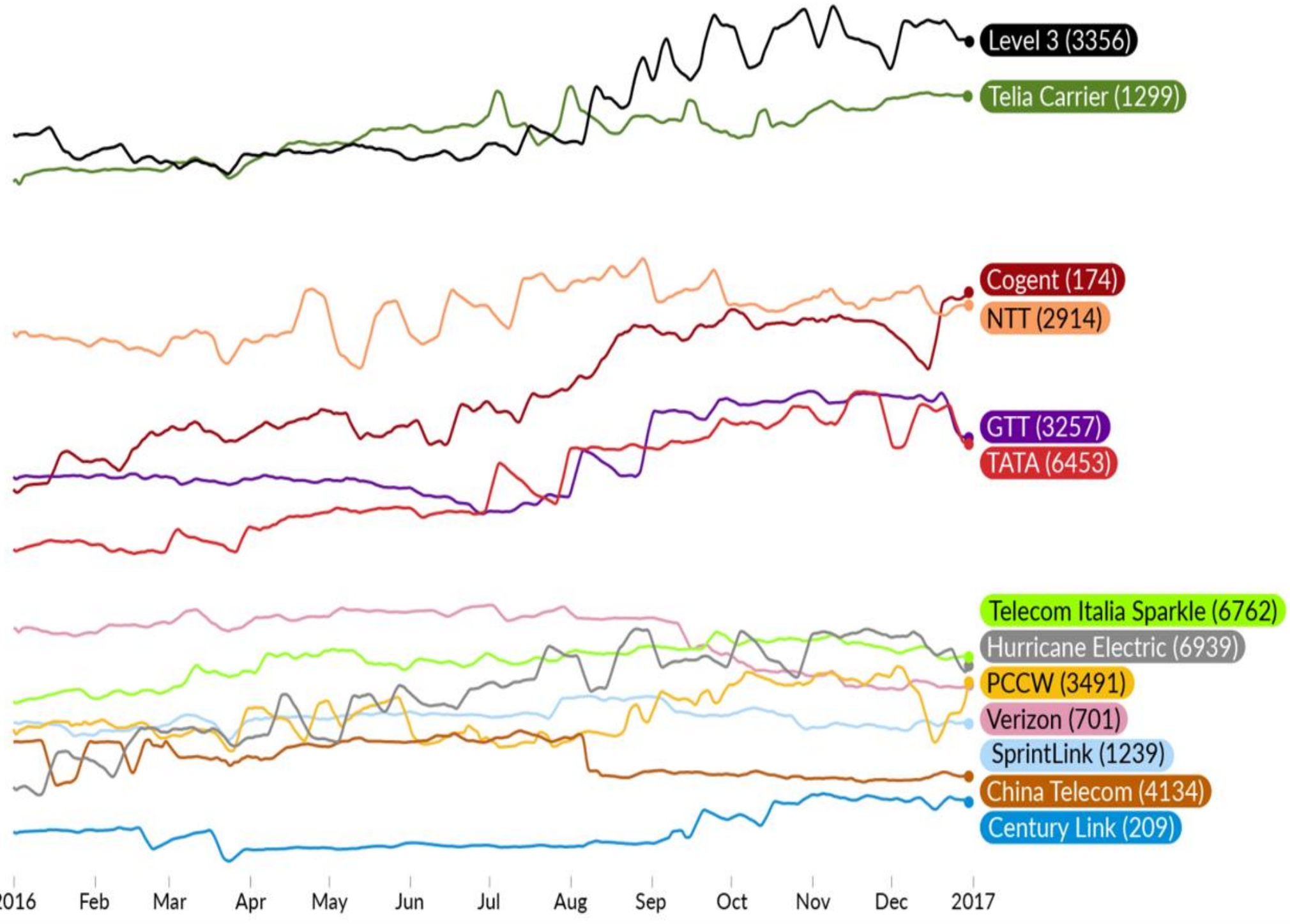


Who are Global Tier 1 ISPs in the World?

- As of 2016, there are 13 Tier 1 ISP which don’t have any transit provider

- Baker’s Dozen is considered as Tier 1 ISP List and every year list is updated with the ISP ranking. List is provided by measuring the Transit IP Space of each ISP

2016 Baker's Dozen Tier 1 ISP Rankings



IP Transit

- IP Transit is the service of allowing traffic from another network to cross or "transit" the provider’s network, usually used to connect a smaller Internet Service Provider to the rest of the Internet

- It’s also known as Internet Transit. ISPs simply connect their network to their Transit Provider and pay the Transit Provider, which will do the rest

Selling an IP Transit Service in the IXP

- Selling an IP Transit Service in the IXP is common

- Many big Service Providers such as Tier 1 or Regional Tier 2 Providers join an IXP as they see IXP is not only peering point but a location where they can sell their IP Transit Services

Selling an IP Transit Service in the IXP

- Although some IXPs don’t allow selling or buying an IP Transit, there is no real control mechanism which can prevent this situation

- When companies have a peering, they still receive an IP Transit Service (Except Tier 1s) and they use IP Transit as backup connection

BGP Soft Reconfiguration and Route Refresh

- BGP is a policy based protocol and we use inbound and outbound filters with the attributes. BGP updates are kept in many different places in the router.

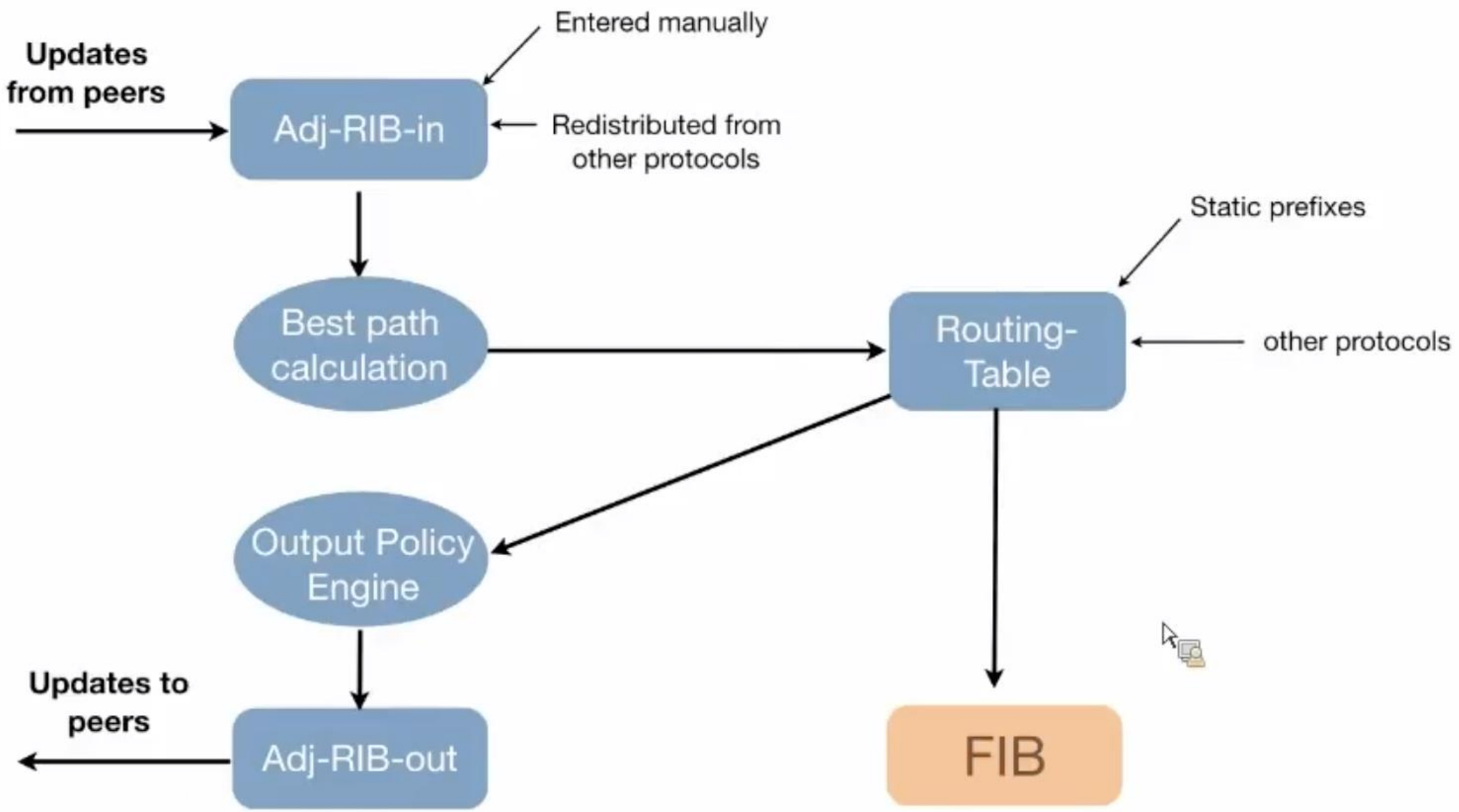
- BGP RIB which is routing table of BGP , RIB which is a router’s general routing table created by all the routing protocols , FIB which is a forwarding table which is data plane

BGP Soft Reconfiguration and Route Refresh

- In addition to BGP RIB, BGP uses adjacency RIB-IN and RIB-OUT databases in the Routers

- All the prefixes from the remote BGP neighbor is placed in the BGP RIB-IN database first.

BGP Soft Reconfiguration and Route Refresh



BGP Soft Reconfiguration and Route Refresh

- Then inbound filter is applied , if we want to allow them, then prefix is taken into BGP RIB database

- If we enable BGP Soft Reconfiguration Inbound , we keep received prefixes in the BGP RIB-IN database, if it is not enabled, we ignore them

BGP Soft Reconfiguration and Route Refresh

- That's why if BGP soft reconfiguration inbound is enabled, even if you filter the prefixes after receiving from the neighboring BGP device , you can still reach them for maybe troubleshooting purposes

- It helps you to verify whether your filter is working correctly

BGP Soft Reconfiguration and Route Refresh

- But obviously this is memory intensive since you keep those prefixes in BGP RIB-IN database in addition to BGP RIB database

- BGP Route refresh works in a different way to accomplish the same task. Still filter is applied for the incoming or outgoing prefixes

- With Route Refresh, you don’t keep the prefixes in the separate databases

- You either take them into BGP RIB database or ignore entirely after filtering

- Thus memory consumption is more efficient

BGP Soft Reconfiguration and Route Refresh

- Don’t forget that Router Memories are expensive

IBGP

- IBGP is used inside an Autonomous system. In order to prevent routing loop, IBGP requires BGP nodes to have full mesh interconnections among them.

- This rule is not required in EBGP because routing loop prevention is done by checking the AS number in the AS path in EBGP. In IBGP, AS number is not sent between the BGP neighbors

- Full mesh IBGP sessions may create configuration complexity and resource problem due to high number of BGP sessions in large scale BGP deployment

IBGP

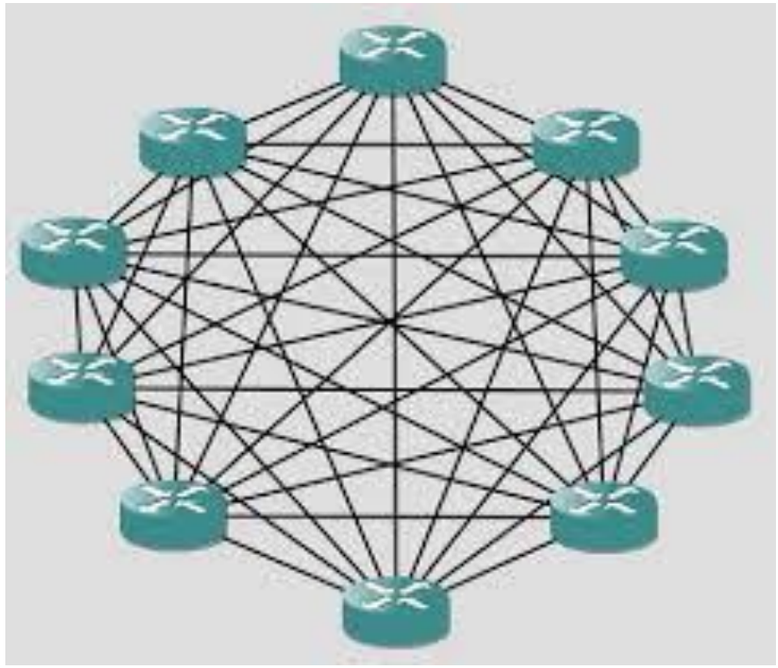
- Route reflectors and confederations can be used to reduce the sessions on each router. Number of sessions and configuration can be reduced by the route reflectors and confederations but they both have important design considerations

- Confederations divide the autonomous system to smaller sub-Autonomous systems

- Confederations give the ability to have ebgp rules between Sub-ASes. Also inside each Sub-AS, different IGP can be used. Also merging company’s scenarios is easier with Confederation than Route Reflectors

BGP Route Reflectors

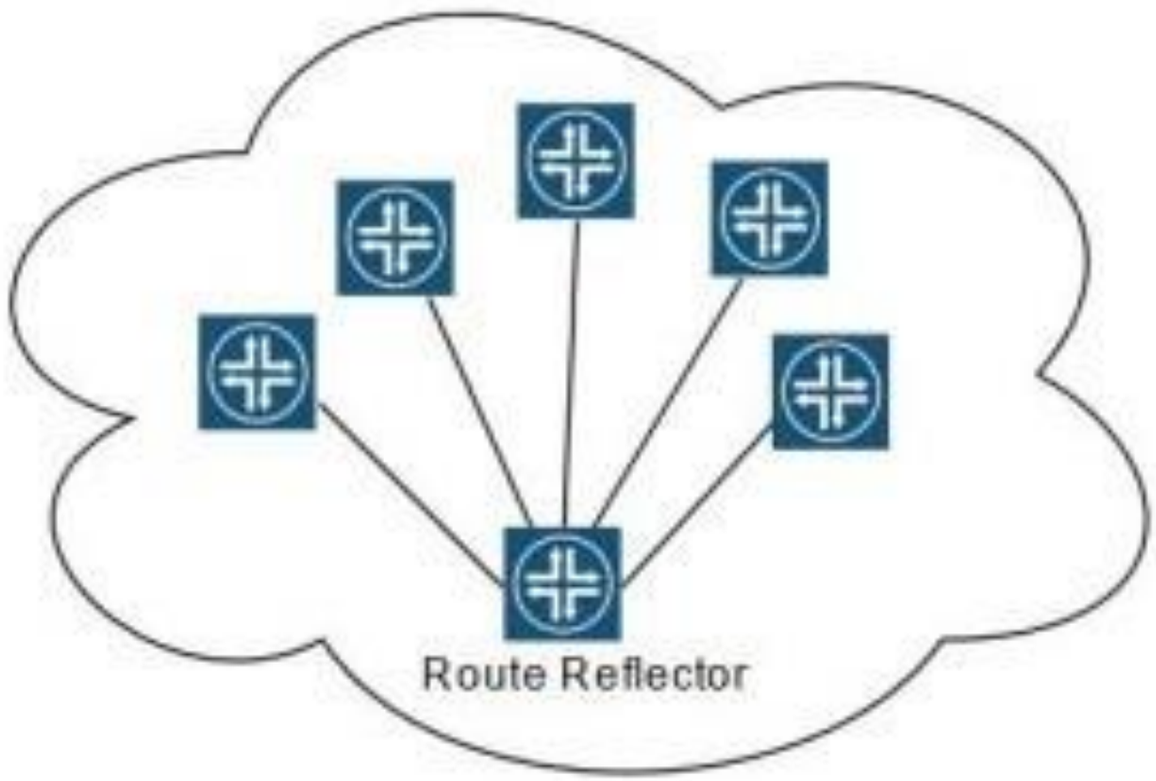
- It is used to avoid Full Mesh connectivity requirement in IBGP



- There are many design caveats when BGP RR is used, most important ones are BGP RR can create sub optimal routing, increases convergence time, reduces redundancy

BGP RR Creates Logical Hub and Spoke Topology

- It creates a logical Hub and Spoke Topology. Each BGP RR Client has a BGP session with only BGP RR, not with each other.
- Thus Full Mesh IBGP topology become Hub and Spoke IBGP Topology with BGP RR



BGP RR Path Selection and Distribution

- Route reflectors choose the best path to the exit point based on their perspective, and not the client's perspective.

- A path to the exit point of the network for a certain prefix can be optimal for the Route Reflector based on its lowest IGP metric to the exit point, but this might not be true from the client's perspective.

- Route Reflectors only advertise one path as their best path for a prefix and don't advertise any other paths to their clients.

BGP RR Sub-optimal Routing

- With this Route Reflectors behavior which removes additional BGP advertisements to the control plane of its clients, an issue of suboptimal routing will occur for Route Reflector clients.

- This is because the Route Reflector client will not have all the available routes and it cannot compare the IGP metric of every path in order to determine the shortest path.

BGP RR Sub-optimal Routing

- Sub optimality in reflecting the path from the RR to the clients usually happens when the Route Reflector is not topologically near its clients. This sub optimality is more seen when RRs are not in the forwarding path, especially in virtual RR’s that are completely out of path.

BGP RR Works Based On

- Route selection is based on their point of view
- RRs propagate only the best path over their sessions by hiding other paths.
- This might not be the best path according to the client's point of view.
- RRs run best path algorithm and advertise only one update to their clients, which may result in suboptimal routing.
- RR's are usually deployed based on exit points in the network

BGP RR vs. Regular BGP Speakers Best Path Selection

- Route Reflectors use the same best-path selection process as normal BGP speakers do. When receiving the same prefix coming from multiple peers, the tiebreaker decision process is done:
- Highest LOCAL_PREF
- Locally originated via network/aggregate or redistributed via IGP.
- Shortest AS-Path
- Lowest origin type
- Lowest MED
- eBGP paths over iBGP
- Path with the lowest IGP metric to the BGP next hop

- If all the steps before the 7th step are equal, then step 7 will be the deciding factor for the best path for the Route Reflector. So, the preferred path will be the lowest IGP metric to the BGP next hop.

- By default, Route Reflector’s only advertise the best path to their clients, so in case of the tiebreaker explained above, the traffic will be send to the exit point with the lowest cost/shortest path possible.

In-band vs. Out-band (Inline vs. Offline) BGP RR for Optimal Routing

- In-band Route Reflectors usually have better view of the IGP topology of the network than out-of-band Route Reflectors, so they can advertise better optimal best paths to their clients

- This due to proximity to the RR Clients, with inbound BGP RR, RR is topologically deployed closer to the RR Clients

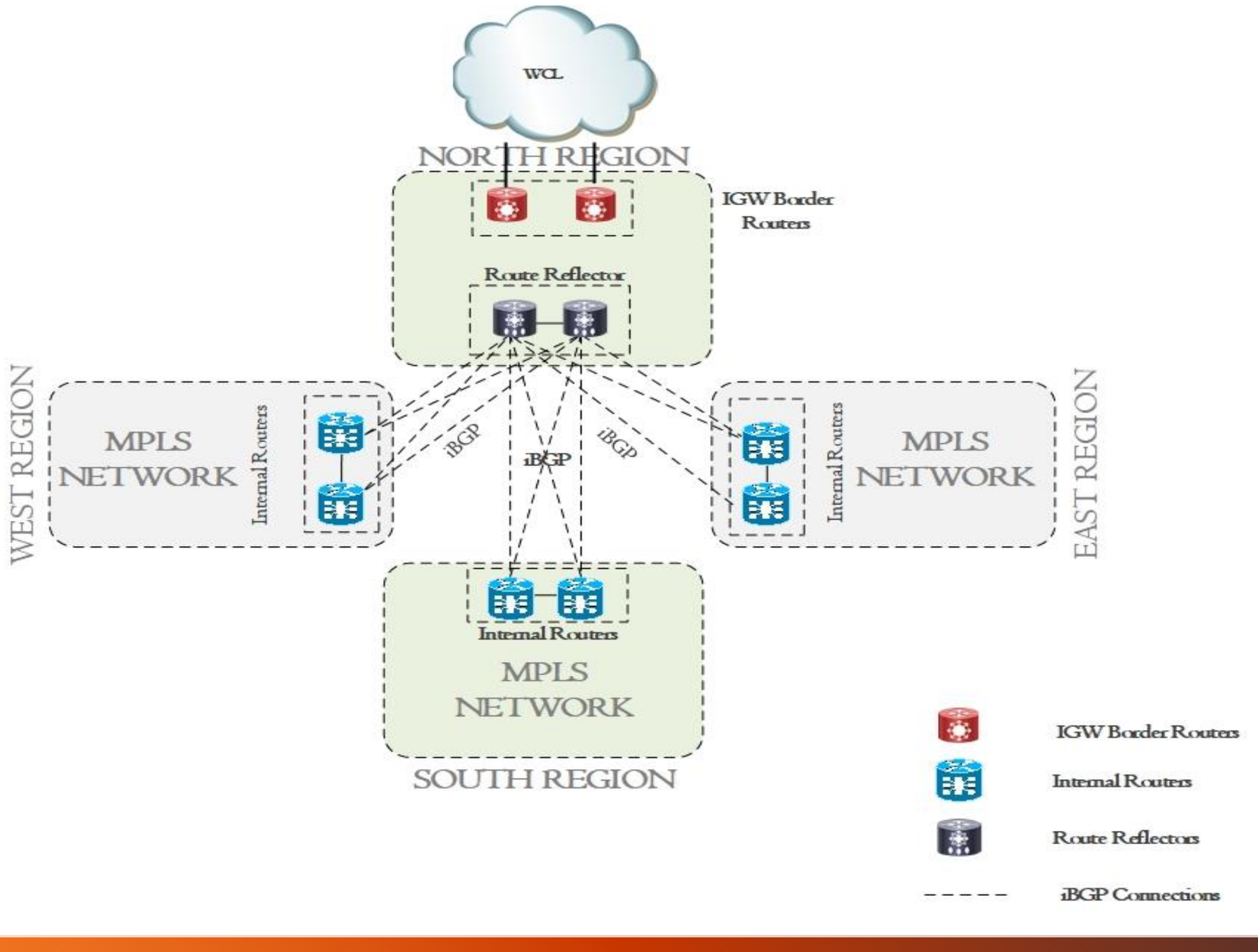
BGP Route Reflector Design Options

- BGP Route Reflectors can be deployed in a distributed or centralized way

- Distributed BGP Route Reflectors provide optimal routing compare to Centralize design

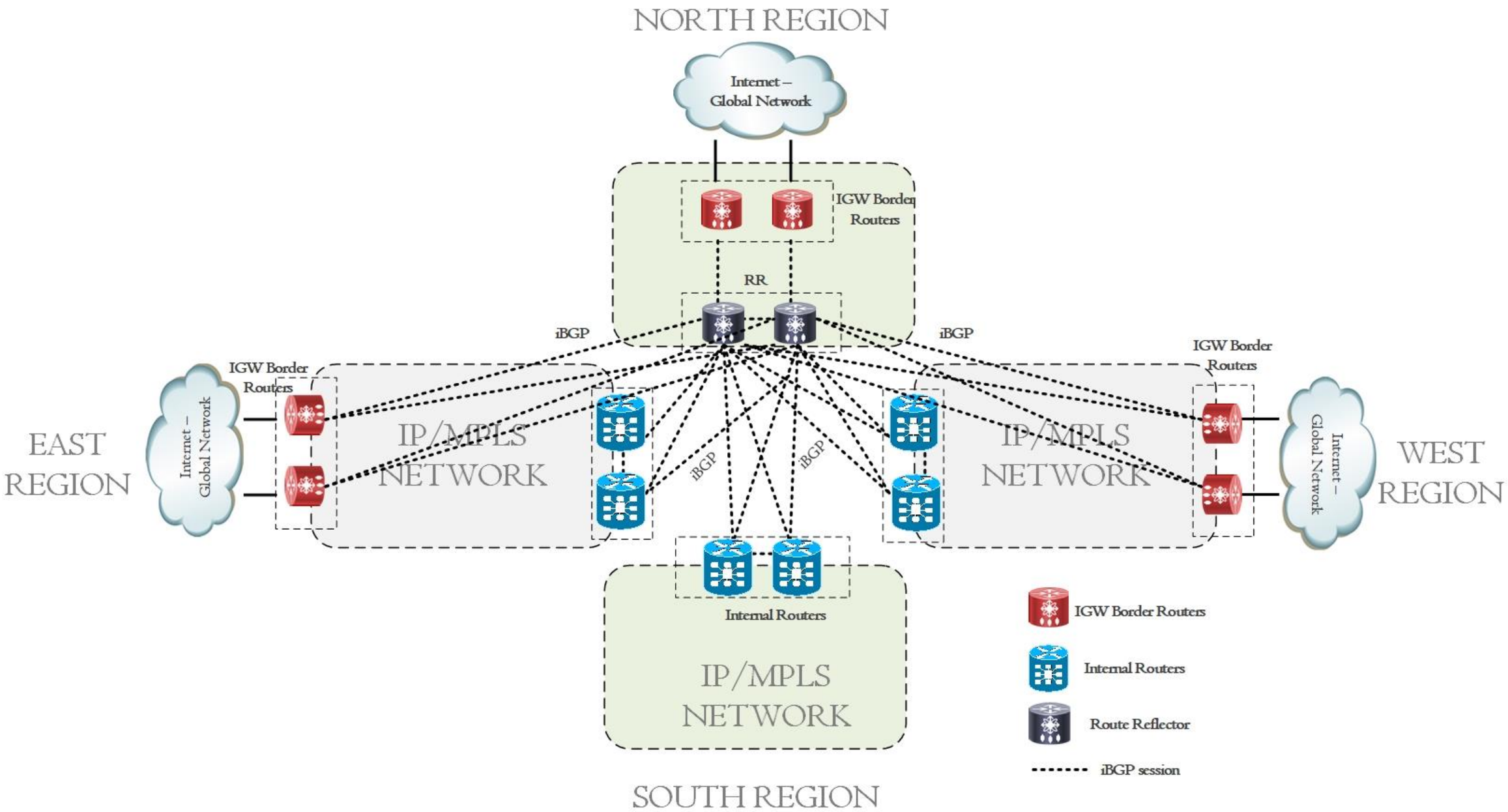
- Centralized BGP Route Reflectors can be used together with BGP ORR (Optimal Route Reflection) to provide optimal routing for RR Clients

Centralized BGP RR Design

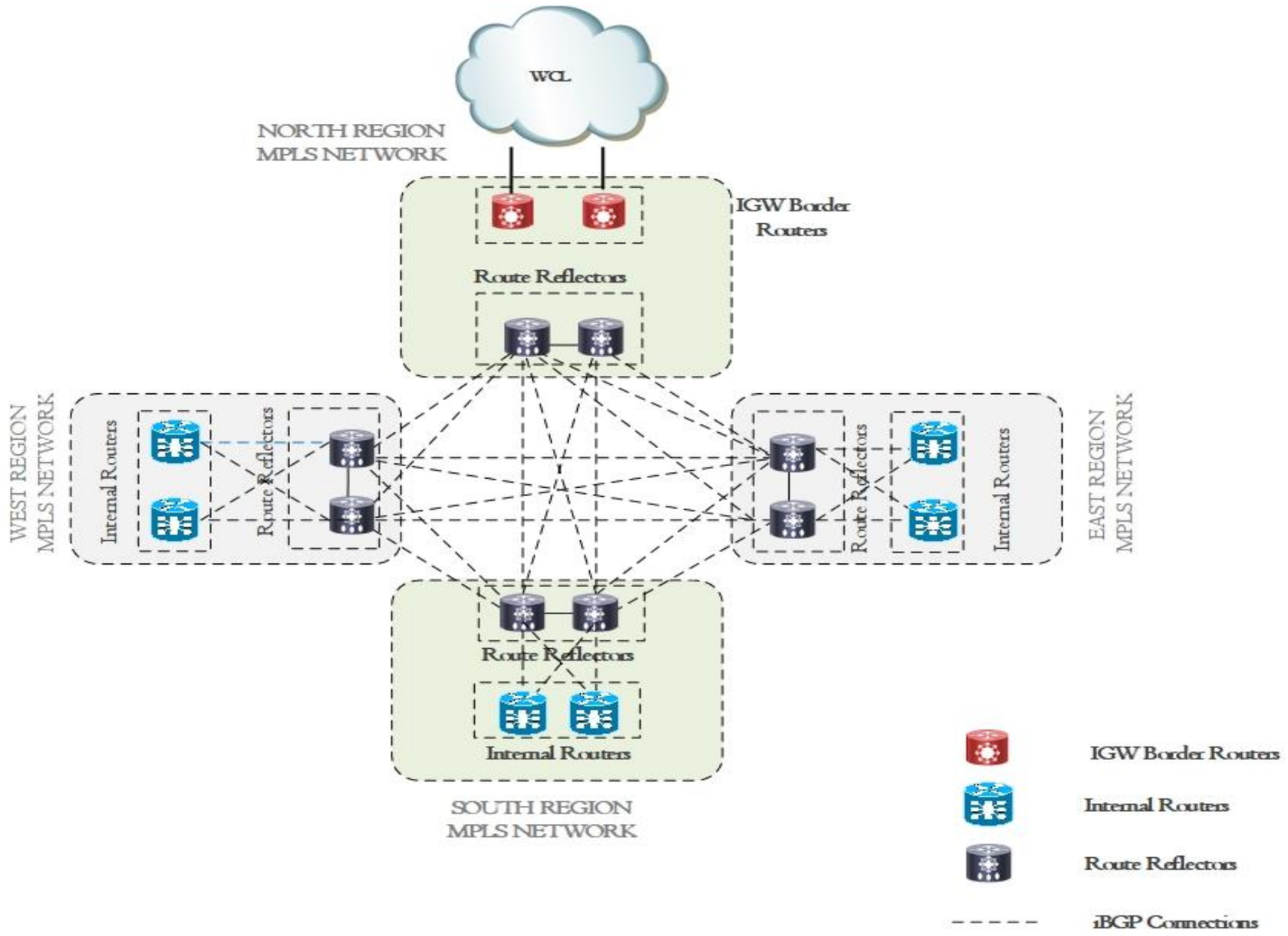


Centralized BGP RR Design

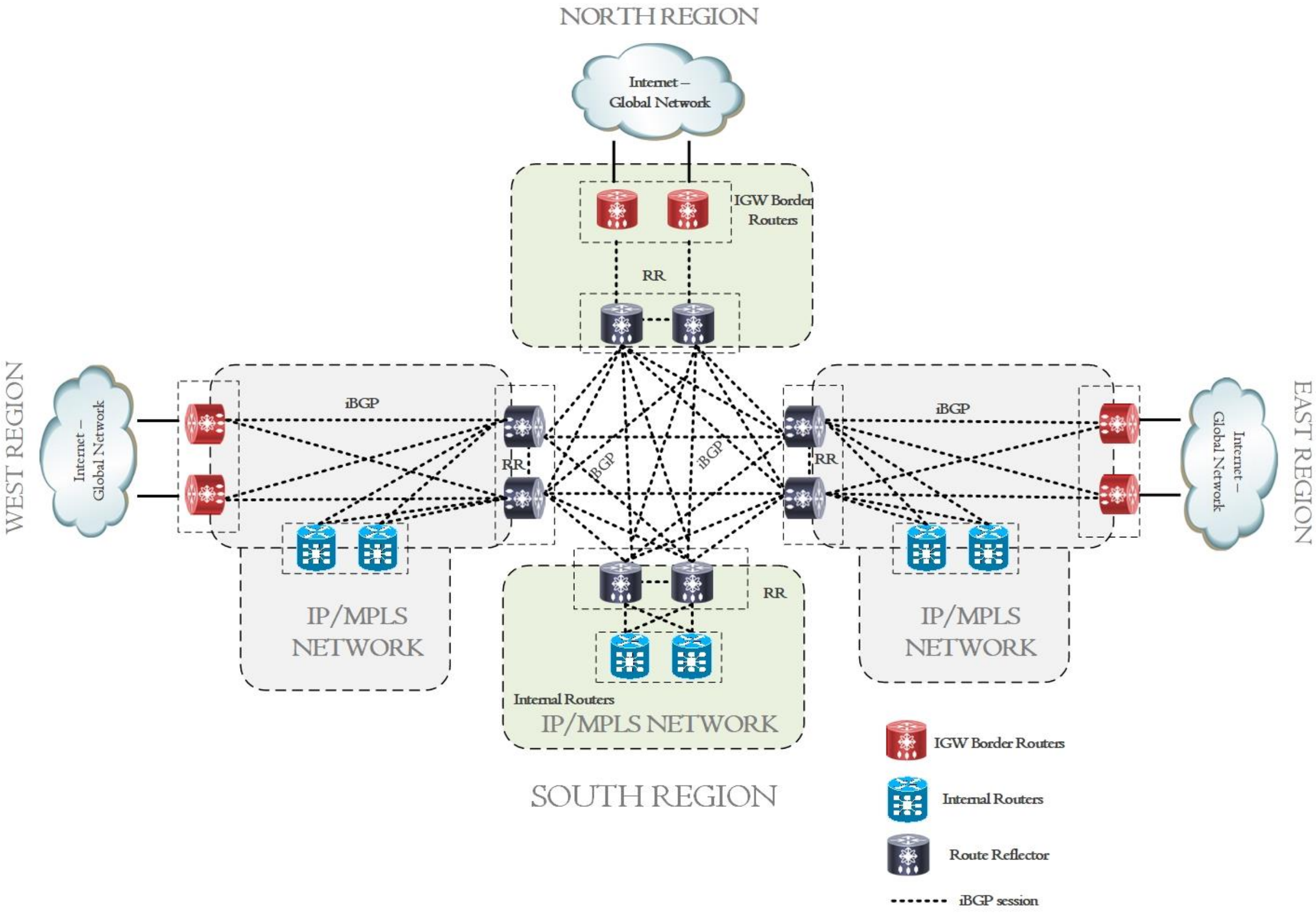
In Centralized RR Design, even though there are Internet exit from East and West Region as well, all Internal routers use North Region Internet connectivity as RR's point of view North Region Internet Gateways have lowest IGP cost



Distributed BGP RR Design



Distributed BGP RR Design



BGP Route Reflector Cluster

- Route reflectors create a hub and spoke topology from the control plane standpoint. RR is the hub and the clients are the spokes.

- RRs and RR Clients form a cluster.

- We should have more than one RR in a cluster for redundancy and load sharing

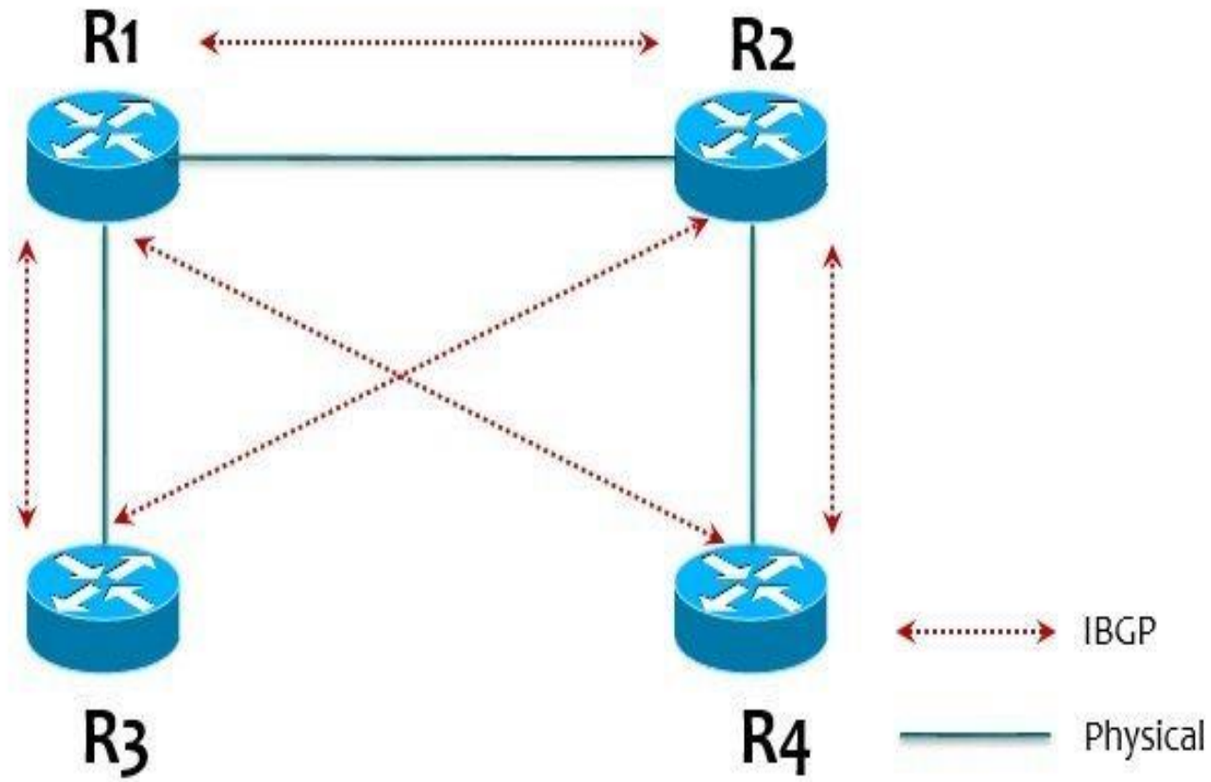
BGP RR Cluster

Assume that we use both route reflectors as cluster ID 1.1.1.1 which is R1's router ID.

R1 and R2 receive routes from R4.
R1 and R2 receive routes from R3.

Both R1 and R2 as route reflectors appends 1.1.1.1 as cluster ID attributes that they send to each other.
However, since they use same cluster, they discard the routes of each other.

That's why, if RRs use the same cluster ID, RR clients have to connect to both RRs.



- BGP Route reflector cluster is the collection of BGP Route reflector and Route reflector clients

- The RR uses Cluster ID for the loop prevention RR Clients don’t know which cluster they belong to

- Using same BGP Cluster ID is good for resource consumption but bad for fast convergence

Changing the BGP Route Reflector Behavior via BGP Shadow RR and BGP Add-Path

- If you want to send more than one best path by the BGP Route Reflectors for multi pathing or fast reroute purpose then below are the approaches.

- Unique RD per VRF per PE. Unique Route Distinguisher is used per VRF per PE. No need, Add-Path, Shadow RRs, Diverse Paths. But only applicable in MPLS VPNs.

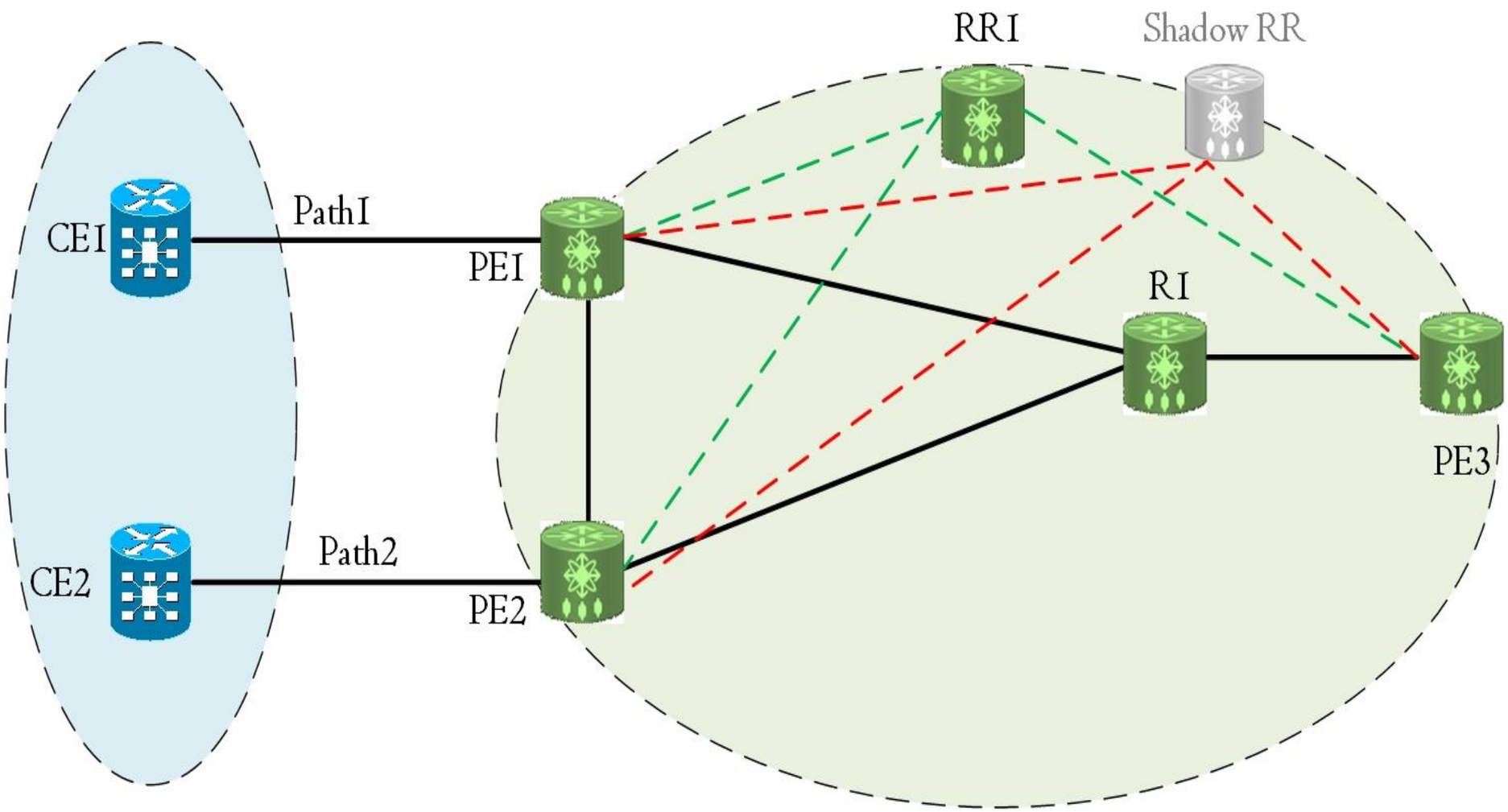
- BGP Add-Path

- BGP Shadow Route Reflectors

- BGP Shadow Sessions

SHADOW ROUTE REFLECTORS

Shadow Route reflectors; you have two Route reflectors, one route reflector sends best path, second one calculate the second best and sends the second best path.



BGP SHADOW ROUTE REFLECTORS

- Shadow Route reflector deployments don’t require MPLS in the network

- Shadow RR is using a different RR node to advertise the second best path, Shadow Session approach is the same as Shadow RR, except no need to have a separate box , using the same box but maybe having a virtual context on the RR node for the Shadow sessions

BGP SHADOW ROUTE REFLECTORS

With Shadow Route Reflector second IBGP session is created between Shadow RR and PE3. Over the second IBGP session, second best is sent. This session is called shadow Route reflector sessions.

BGP Add-Path

- With Shadow RR or Shadows sessions, there are secondary IBGP sessions between RR and PEs. But same behavior can be achieved with BGP ADD-Path without extra IBGP session

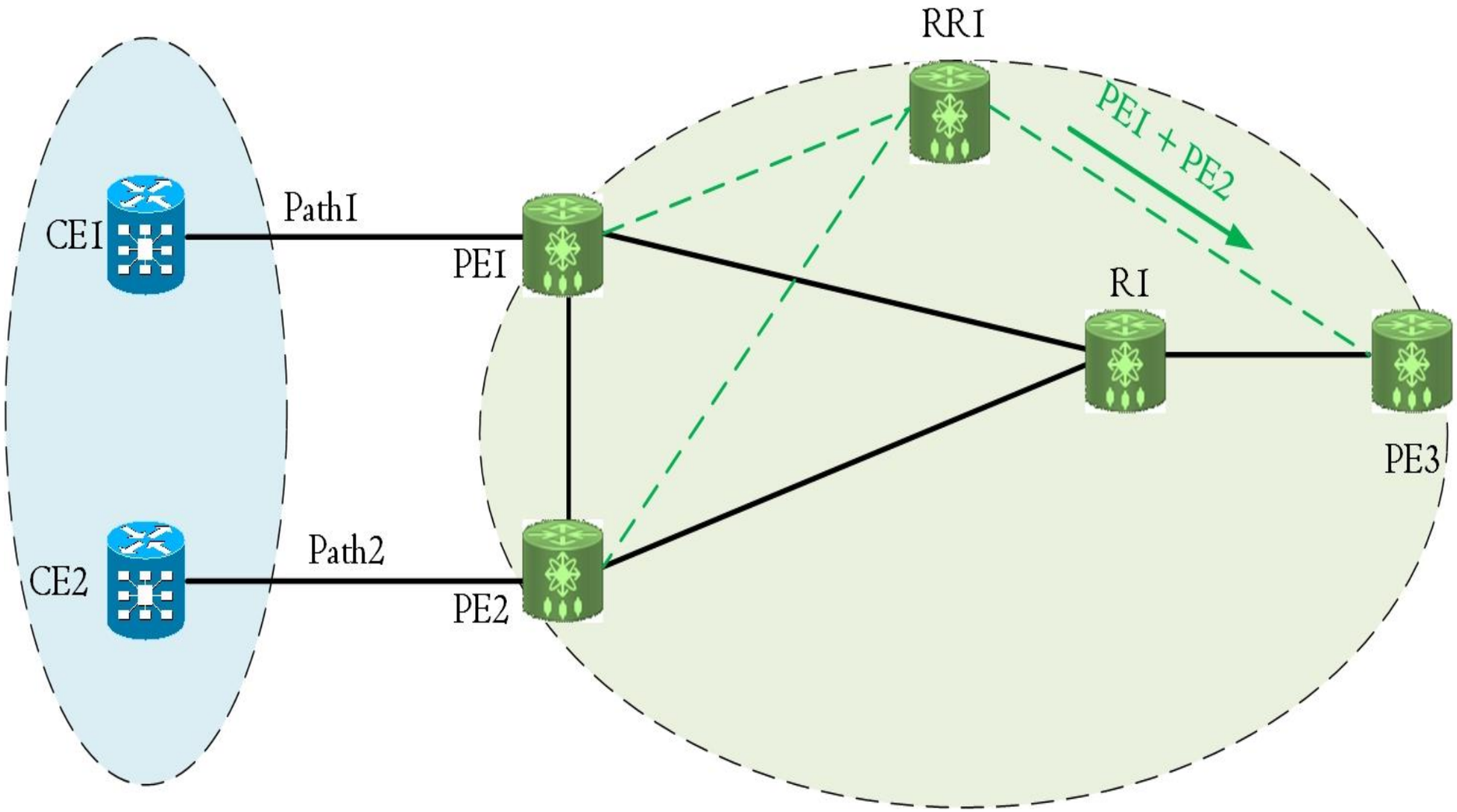
- Add-path uses path-identifier to distinguish the different next hops over one IBGP session

BGP Add-Path

- In IBGP, if multiple paths are sent over the same BGP session, last one is kept by the receiving BGP speaker, because for the first one implicit withdrawn is sent, if route is completely gone then BGP Explicit withdrawn is sent

- With Add-Path, withdrawn is not sent thus receiving BGP router keeps all the paths and can make a best path selection based on its own view

BGP Add-Path

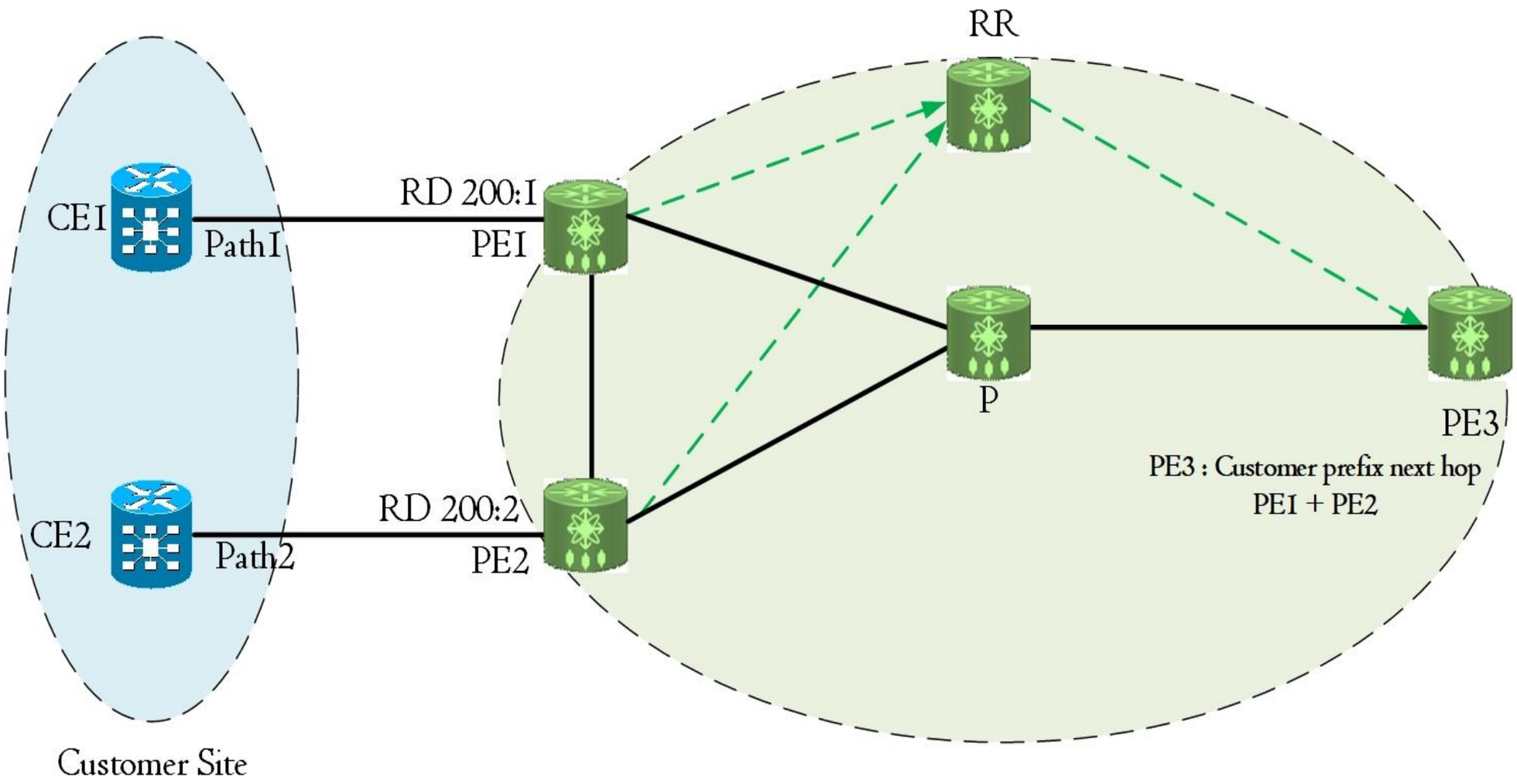


BGP Route Reflector behavior in MPLS VPN Using Unique RD per VRF per PE Approach

- When RR is used in MPLS VPN, unique RD (Different RD) is configured on the PE routers to advertise unique VPN prefixes

- When VPN RR receives the prefixes, since there will be unique VPN prefix for the same customer prefix, RR doesn’t perform best path selection, it reflects both prefixes with their own next hops

BGP Route Reflector behavior in MPLS VPN Using Unique RD per VRF per PE Approach



Comparison between BGP Add-path, Shadow RR, Shadow Sessions and Unique RD per VRF per PE

Design Concern	BGP Add Path	BGP Shadow RR	BGP Shadow Sessions	MPLS Unique RD per PE per VRF
Best in MPLS	No	No	No	Yes
How many IBGP Session between RR and RR-Client	One IBGP session, Path IDs are different for different next-hop	One session per Route reflectos. If there is only one more Shadow RR which sends second best path, two IBGP sessions on the RR Client, one for each RR	One session per next-hop. Only one RR but multiple separate IBGP session is required between RR and RR Client	One IBGP session between VPN RR and RR Client, different RDs make the same IP prefixes unique
Resource Requirement	Best	Worst, requires separate RR and IBGP session per next-hop	Better than Shadow RR because doesn't requre separate Route reflector, worse than ADD path because require extra IBGP session per next-hop	Same as Add-path, doesn't require extra IBGP session or Route Reflector
Migration of existing Route Reflectors	Very hard,all Route Reflectors and clients need to be upgraded to support Add-path	Easy, only Route Reflector code needs to be upgraded	Easy,only Route Reflector code needs to be upgraded	Easiest because there is no upgrade on any device. Only unique/separate Route Distinguisher needs to be configured on the Pes per VRF
Standard Protocol	Yes IETF Standard	Yes IETF Standard	Yes IETF Standard	Yes IETF Standard
Stuff Experince	Not well known	Not well known	Not well known	Known
Troubleshooting	Hard,default behaviour of BGP which is advertising only one best path is changing. Operation stuff needs to learn new troubleshooting skill	Easy	Easy	Easy
IPv6 Support	Yes	Yes	Yes	Yes
Provisioning	Easy, only one IBGP session between Route reflector and the client	Hard, one IBGP session per next-hop	Hard, one IBGP session per next-hop	Easiest,only the consideration is to have unique RD per VRF per PE

BGP RR Benefits

- **Main benefit of BGP Route Reflector is Scalability**

- BGP Route Reflector reduces the total number of BGP sessions in the network and also reduces the number of BGP session per router

- BGP RR simplifies the configuration of the BGP routers

BGP RR Benefits

- Route Reflector hides the available paths. This is benefit for some networks, problem for the others

- BGP RR provides RBAC Opportunity, Low level engineers can maintain the RR Client and only Advanced level engineers can touch the Route Reflectors

BGP RR Problems

- It hides the path. In the previous slide, we mentioned it is as benefit. If resource utilization is concern, it is benefit, other than that it is a problem for suboptimal routing and other requirements

- BGP RR prevents Fast Reroute which might be requirements for some networks

- BGP RR increases Control Plane Convergence time

BGP RR Problems

- BGP RR can create sub optimal routing

- BGP RR can be a single point of failure if it is not designed correctly

BGP Add-Path and BGP ORR Requirement

- Reducing routing churns via oscillation, faster convergence, better load sharing and availability are some advantages of BGP Add-Path

- Improved Path Diversity is another benefit from this solution, which can bring effective BGP level load and fast connectivity restoration (ex. BGP PIC - Prefix Independent Convergence for faster convergence-FRR)

Memory consumption on the edge devices with BGP Add-Path

- By expanding the network to more exit point peering connections, which can result in getting same routes from more peers (especially when receiving full routing tables)

- More paths and lots of updates are advertised to clients, so the number of BGP announcements will increase for Route Reflector clients, which might lead to significant memory problems on the edge devices.

- Introducing a large number of BGP states to all routers will create a lot of entry on the Route Reflector clients BGP Table. Some clients might not support Add-Path, others that support, might not have enough capacity.

How can optimal routing with BGP can be guaranteed?

- Add-path is a BGP capability, which mean it needs to be agreed between RR and RR Client. Upgrading both RR and RR Client might take so much time to migrate the BGP Software to one which supports BGP Add-Path feature as there might be so many Edge device in the network that acts as RR Client

- If all available next hops won’t be advertise how can optimality can be guaranteed?

- ANSWER is BGP ORR (Optimal Route Reflection)

BGP Optimal Route Reflection – BGP ORR

- Optimal Route Reflection is a IETF Draft but there are many vendor implementation as of 2019

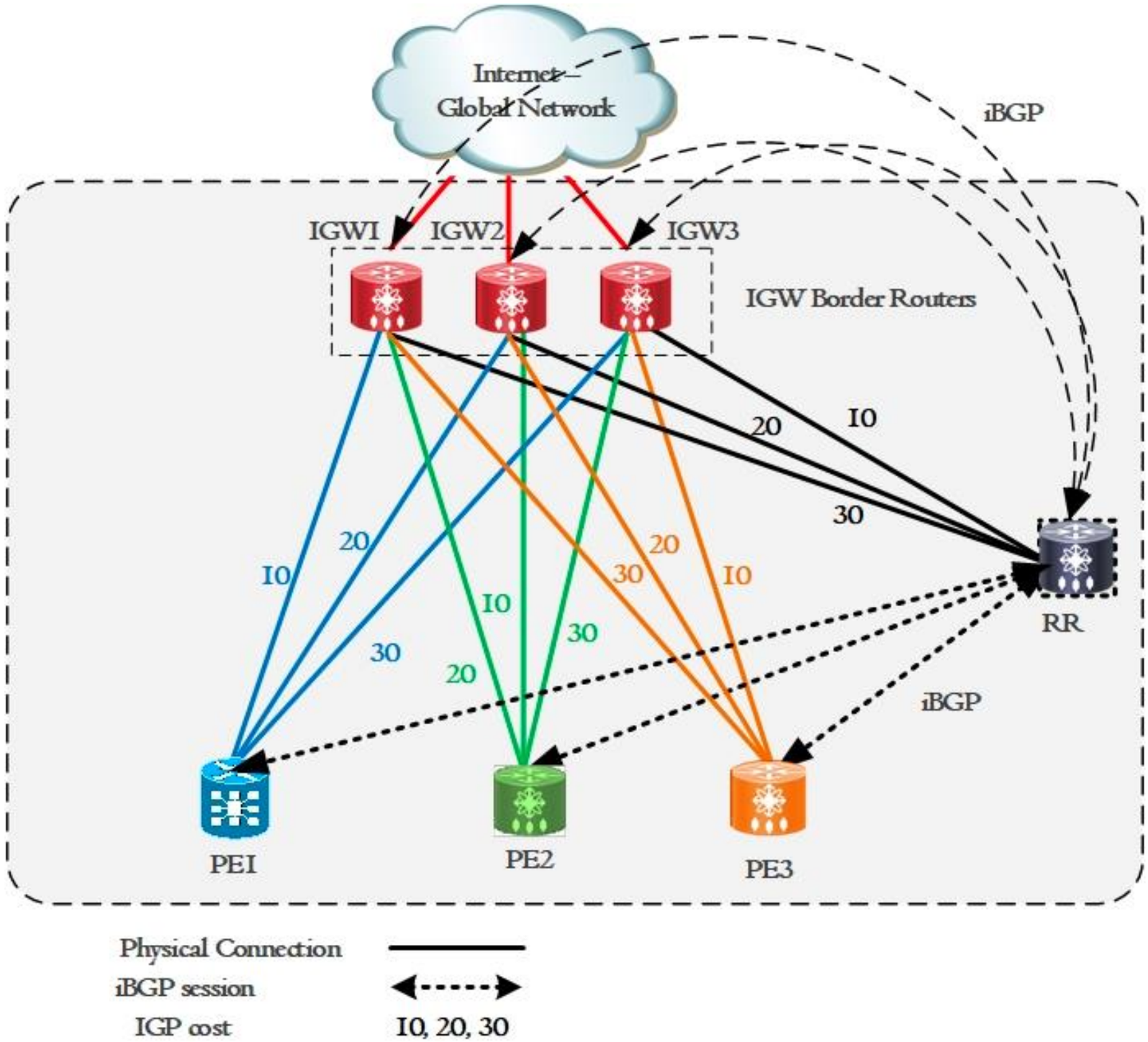
- Based on this solution, the RR will do the optimal path selection based on each client's point of view. It runs SPF calculation with their clients as the root of the tree and calculates the cost to the BGP next-hop based on this view

BGP Optimal Route Reflection – BGP ORR

- From the logical point of view, the Route Reflector position is virtualized, making it independent of its RR-Clients

- With ORR Route Reflectors location would be independent from the selection process of the best-path. Each ingress BGP border router can have a different exit point to the transit providers, for the same prefix for example

BGP Optimal Route Reflection – BGP ORR



Requirements for BGP ORR

- Link-state routing protocol is required in the network for the Route Reflectors to have a complete view of the network topology based on the IGP perspective. No changes are required to be done by the clients

- ORR is applicable only when BGP path selection algorithm is based on IGP metric to BGP next hop, so the path will be the lowest metric for getting the Internet traffic out of the network as soon as possible

BGP ORR is not an Alternative but Complementary to the BGP Add-Path

- ORR is not an alternative to BGP Add-Path or other methods for Path Diversity, though it is an alternative to provide optimal routing

- ORR can be used together with ADD-PATH to improve the quality of multiple advertisements, to propagate the route that can be the best path. Also, it can add resiliency and faster re-convergence for the network. For example, by receiving 4 paths from exit point peers across the network, it will choose the best path plus the 3 other paths based on the IGP cost. So, it’s a true way to add resiliency through add-path

How BGP ORR Works?

- With ORR, at the 1st step, the topology data is acquired via ISIS, OSPF, or BGP-LS. The Route Reflector will then have the entire IGP Topology, so it can run its own computations (SPF) with the client as the root. There could be as many rSPF (Reverse SPF) run based on the number of RR clients, which can increase the CPU load on the RR

- So, a separate RIB for each of the clients/groups of clients is kept by the RR. BGP NLRI and next-hop changes trigger ORR SPF calculations. Based on each next-hop change, the SPF calculation is triggered on the Route Reflector

How BGP ORR Works?

- The Route Reflectors should have complete IGP view of the network topology for ORR, so a link-state routing protocol is required to be used in the network. OSPF/IS-IS can be used to build the IGP topology information

- IGP is great for link state distribution within a routing domain or an autonomous system but for link state distribution across routing domains EGP is required. BGP-LS provides such capability at high scale by carrying the link state information from IGP protocols as part of BGP protocol messages

How BGP ORR Works?

- Route Reflectors keeps track of which route it has sent to each client, so it can resend a new route based on changes in the network topology (BGP/IGP changes reachability). The Route Reflector function is 1 process per route but the ORR function is 1 process per route per client router

- ORR brings the flexibility to place the Route Reflector anywhere in the topology, which provides Hot Potato Routing, supports resiliency via ORR Groups, requires no support from clients and finally brings much better output when is used with ADD-PATH

Different types of ORR (Optimal Route Reflection) Deployments

1. Optimal BGP path selection based on client IGP perspective

2. Optimal BGP Path Selection Based on Policy

1. Optimal BGP path selection based on client IGP perspective

- Optimal BGP path selection is done Based on the Client's IGP Perspective, and not the RR's IGP perspective. To reduce the SPF calculation overhead on the RR, Optimization such as partial and incremental SPF can be used

2. Optimal BGP Path Selection Based on Policy

- This solution is based on User Defined Policy. The clients will always send traffic to a specific exit point of the network regardless of how the topology looks like

- For example, one of the Policy methods can be using for the customers who pay more and gets SLA (Can be classified and marked with BGP Communities), so the traffic can be sent to particular Internet region and particular Transit Operator, instead of doing Hot Potato routing

Same or Different BGP RR for Different Services (Different BGP Address Families)

- For the different address families, different set of Route reflectors can be used, this avoids fate sharing

For example if IPv4 RR is attacked, VPN customers may not be impacted if different sets of RR is used

Route Target (RT) Constraints

- If you are using VPN Route reflectors , you can use multiple Route reflectors for different prefixes if scalability is a concern

- Based on Route Targets, we can use Route Reflector Group-1 to serve odd Route Target values, Route Reflector Group-2 to serve even Route target values

Route Target (RT) Constraints

- In this solution PEs send all the RT values to both Route Reflector Groups. They receive and process all the prefixes but based on odd/even ownership they filter the unwanted ones. But processing the prefixes which will be filtered anyway is not efficient way

- Instead Route Target Constraints should be deployed so PEs can signal to the RR their desired Route Target values

- RR sends to the clients the prefixes which are asked with the RT values send to them by the clients

BGP Confederations

- RFC 5065 describes the use of Autonomous System Confederations for BGP

- BGP confederations help with this scalability issue by allowing the engineer to subdivide the autonomous system into smaller sub-autonomous systems

- There are generally two design methods when considering BGP confederations

Different BGP Confederation Designs

- Same IGP (OSPF , IS-IS , EIGRP etc.) in each Sub-AS
- Different IGP in the sub-AS
- There are pros and cons of each method as usual
- Implementing BGP confederation significantly reduces the total number of BGP sessions

Different BGP Confederation Designs

- Implementing BGP confederations involves quite a change to BGP configurations and the architecture itself, adding more complexity to achieve stable and scalable BGP design

- Migrating a network to a BGP confederation will be disruptive. Routers that are part of a sub-AS will need to change their BGP configuration to use the sub-AS instead of the real AS numbers

How BGP Confederation Works

- BGP routers within a sub-AS peering are IBGP peers
- BGP routers in different sub-AS are EBGP peers which means that the AS number is prepended when an update travels between the sub-AS
- If a router has to send an update towards its IBGP neighbor within a sub-AS, it will not change the AS_PATH attribute
- BGP between the sub-ASs is called as intra-confederation EBGP

BGP Confederation Route Preference – Best Path Selection

- EBGP routes that are exchanged between the sub-ASs are also known as confederation external routes, which are preferred over IBGP routes when it comes to best path selection

- If BGP has to choose between two paths to the same destination, one path leading inside the sub-AS, and another outside the sub-AS but within confederation, it will choose the external path – towards the neighboring sub-AS.

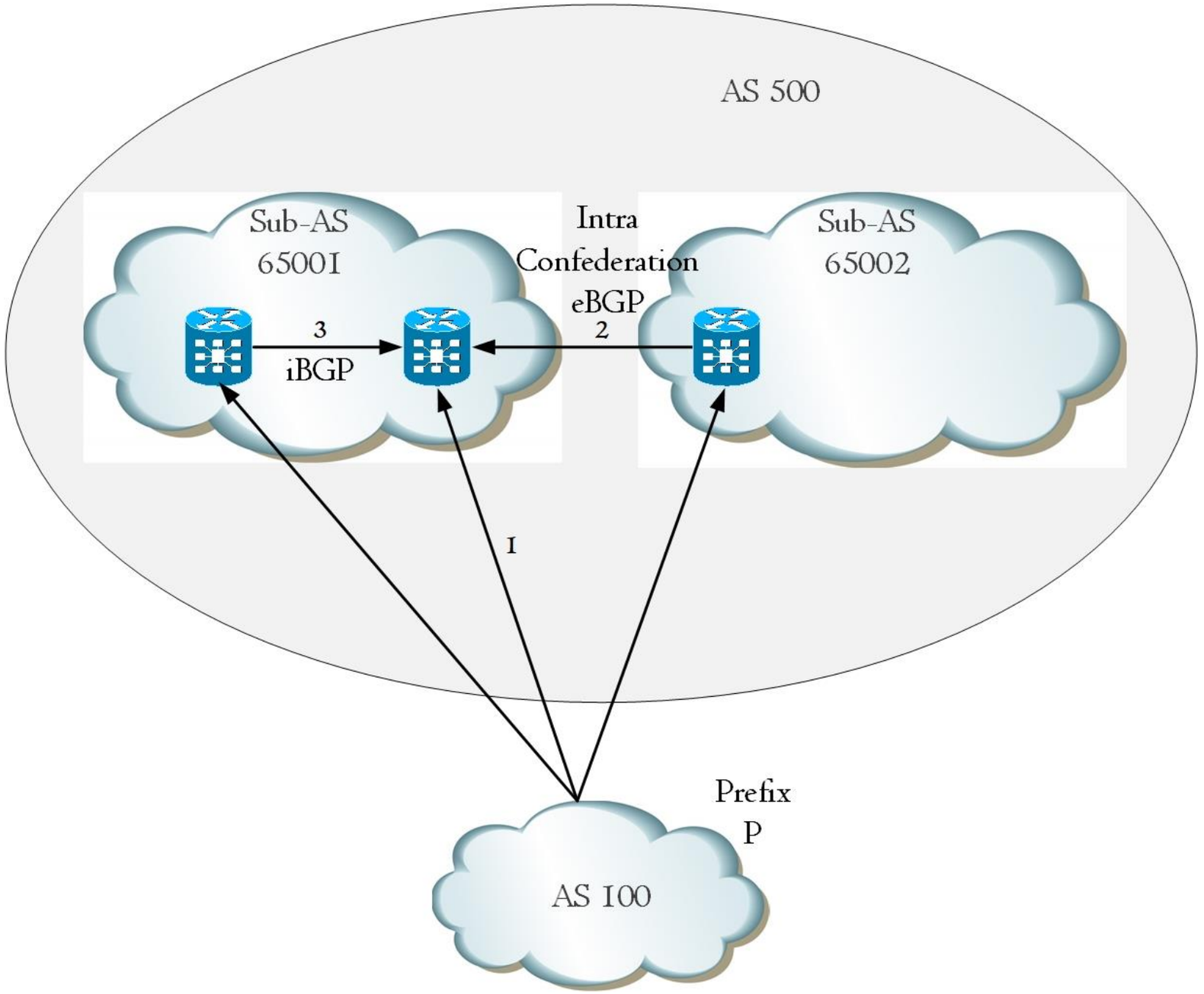
- If it has to choose between confederation EBGP route and EBGP route that leads outside the confederation, BGP will choose the second one

BGP Confederation Route Preference – Best Path Selection

If same prefix learned over real EBGP , intra-confederation EBGP and IBGP within sub-AS, preference will be:

1. Real EBGP Connection (Confederation AS- ID)
2. Intra-Confederation Connection
3. IBGP Connection (Route is learned from an IBGP neighbor within sub-AS)

BGP Confederation Route Preference – Best Path Selection



How BGP Confederation Works

- Between Confederation EBGP and Real EBGP, there are some differences

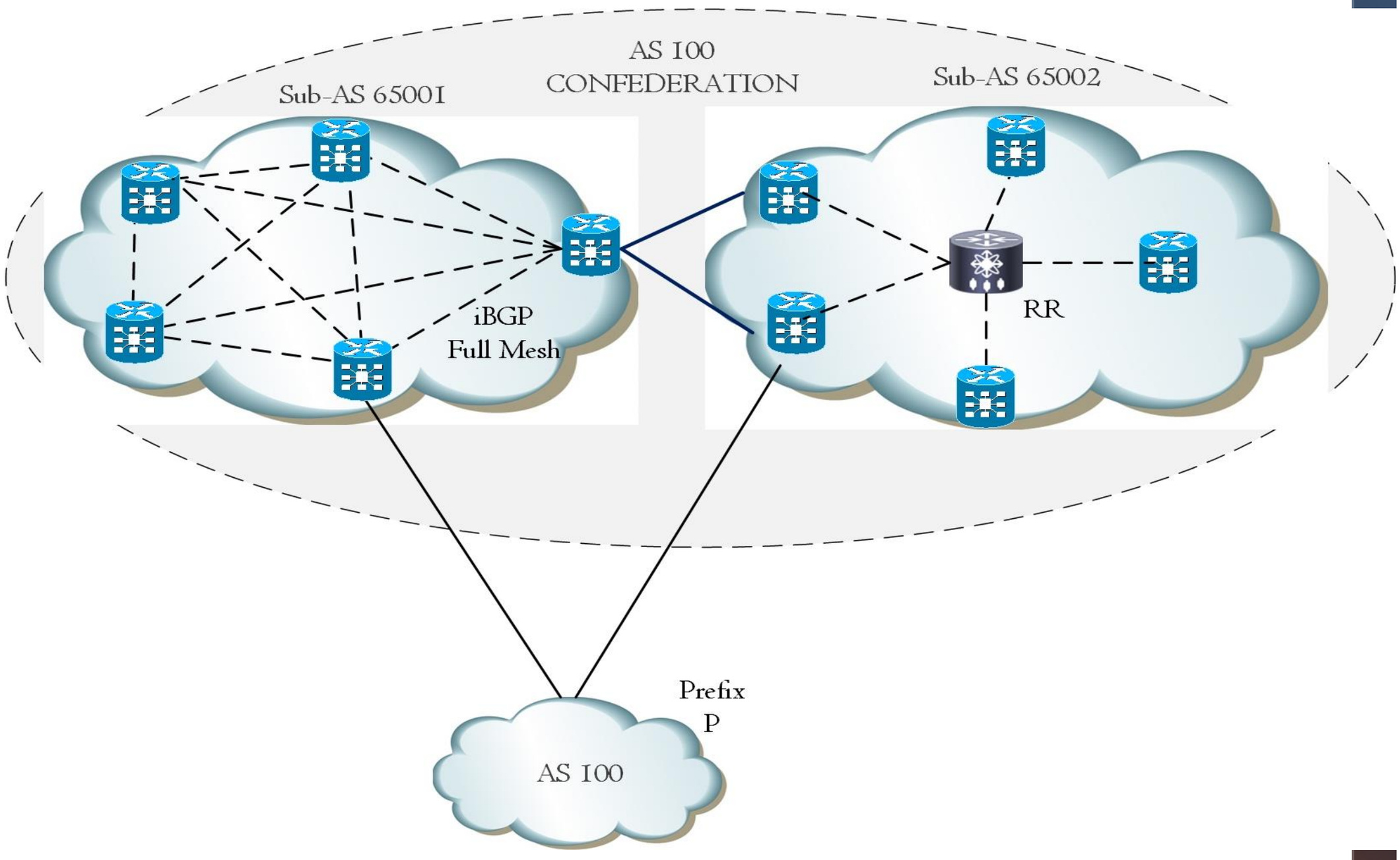
- With BGP confederation EBGP session, MED, local preference and the next-hop are sent unmodified, this is similar to how IBGP works, but AS-Path attribute is changed

How BGP Confederation Works

- Confederation sub-ASs exchange routing information as if they are using IBGP, and the only attribute that changes is AS_PATH. In other words, EBGP behaves like IBGP when implemented inside a confederation

- Because the next-hop is sent unchanged, either an IGP needs to run across the entire confederation or the border routers need to set the next-hop to themselves

BGP Confederation



How BGP Confederation Works

- To make it appear as one AS to all real EBGP peers, the sub-AS in the AS path need to be stripped when sending updates to its peers

- One of the advantages of running a confederation is that a policy can be applied for the sub-AS which does not apply for the entire real AS

- For example prefixes can be sent between the BGP peers within sub-AS with the NO_EXPORT_SUBCONFED community and policy can be distributed within the sub-AS but not outside the sub-AS

BGP Confederation Routing Loop Avoidance

- An EBGP connection between sub-ASs also serves as kind of a loop-avoidance mechanism

- If in the previous topology route is learned from AS 64512 is advertised somehow back to AS 64512, routing update is not accepted by the Originator AS

- This is done with an AS_CONFED_SEQ parameter inside AS-Path Attribute

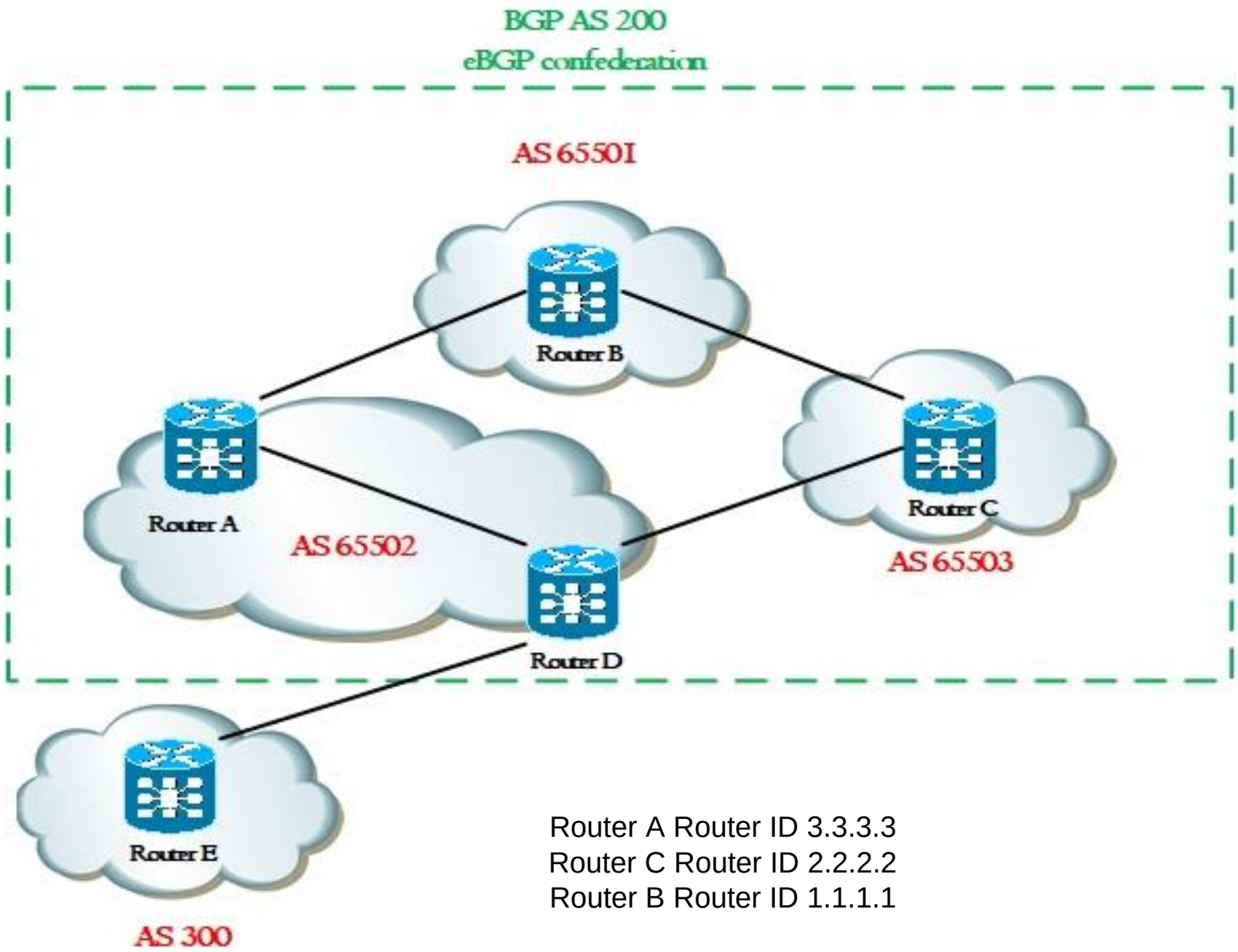
Routing loop avoidance in BGP Confederation

- Based on RFC 5065 section
  When comparing routes using AS_PATH length, CONFED_SEQUENCE and CONFED_SETs SHOULD NOT be counted

- BGP is using the AS_CONFED_SEQ portion of the AS path attribute for routing loop control inside the confederation. However, it's not being used as criteria for BGP path selection inside the confederation

- Let’s have a look at next page for the example

In the picture below all other BGP attributes are identical, BGP chose the path it received from the router with the lowest BGP router ID. Since Router C has a lower router ID (2.2.2.2) than Router **A** (3.3.3.3), Router B choses the path through Router C as best, even though as-path length is longer through the Router C for the destination at AS 300 (2 AS vs. 1)



Router A Router ID 3.3.3.3
Router C Router ID 2.2.2.2
Router B Router ID 1.1.1.1

When to choose BGP Confederation instead of BGP RR?

- The main difference between BGP RR and Confederation is that a confederation may contain different IGPs, adding more flexibility to scaling your network

- Therefore, choosing a Confederation over BGP RR would be more appropriate in case your IGP is exceeding its scalability limit and becomes unmanageable, and you would like to manage many independent ASs, each of which may run a different IGP

Full mesh IBGP to RR Migration

Drivers

- Number of IBGP sessions are too much
- Some devices cannot manage receiving the route from multiple next hops
- Less number of devices to touch when New BGP routers are added to the network

Full mesh IBGP to RR Migration

Preparation Steps:

- Verify that BGP next hops are reachable via IGP (Example Loopbacks)
- You may want to have OOB (Out of Band) Management access
- Schedule the migration during maintenance windows

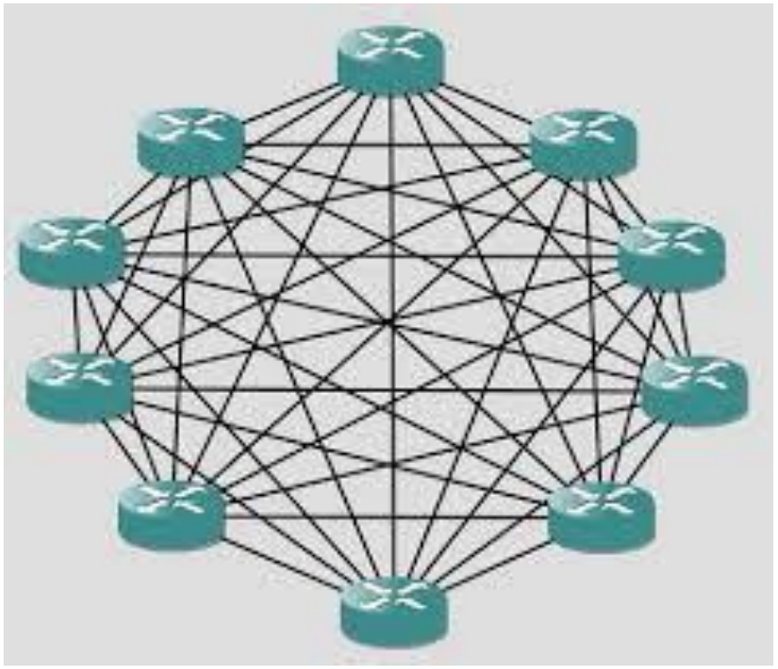
Full mesh IBGP to RR Migration

Preparation Steps:

- Migrate one POP location during one maintenance window to minimize downtime risk
- Have a proper rollback plan and execute if things go wrong in time

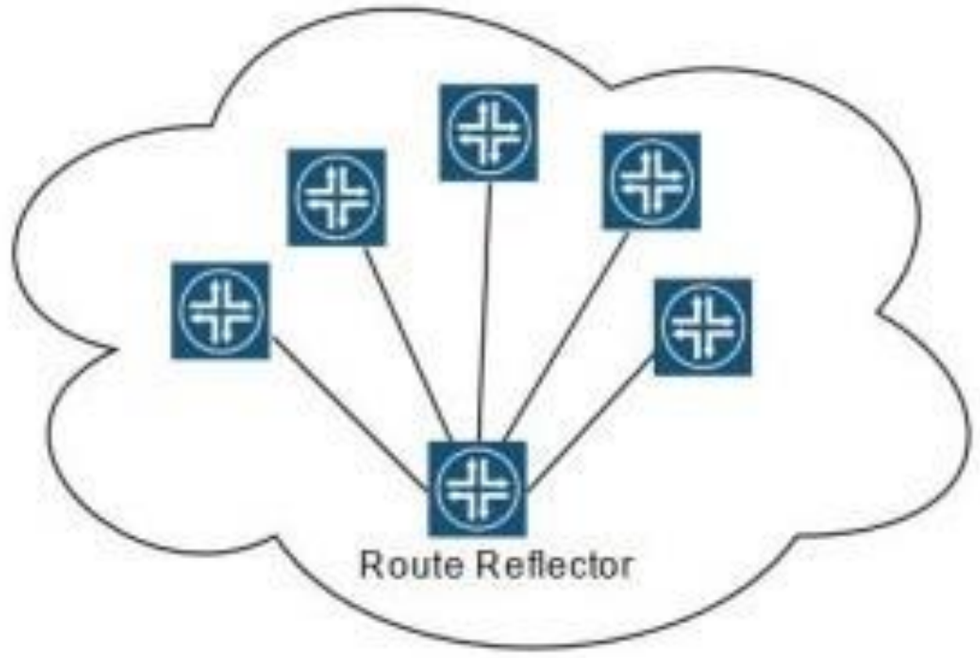
Full mesh IBGP to RR Migration

BEFORE



(Full Mesh IBGP)

AFTER



(BGP Route Reflector)

Full mesh IBGP to RR Migration

Migration Procedure

- Migration depends on if new routers will be deployed as RR or some of the existing routers will be migrated to become RR

- If new routers are added as an RR, downtime can be further minimized

Full mesh IBGP to RR Migration

- Let’s examine what happens when existing routers are used as an RR

- In this case, Core Routers are selected as an RR, they are migrated first

- Make sure there are two Core routers to minimize downtime

Full mesh IBGP to RR Migration

- First migrate the Core routers that will become Route Reflectors and then migrate RR client routers one at a time

- In order to do this, one of the two Core routers advertise into the IGP that it shouldn’t be used as a Transit (OSPF Max-metric , IS-IS Overload)

Full mesh IBGP to RR Migration

- When all edge BGP routers (RR Clients) are migrated to both of the RRs, their full mesh IBGP sessions should be removed

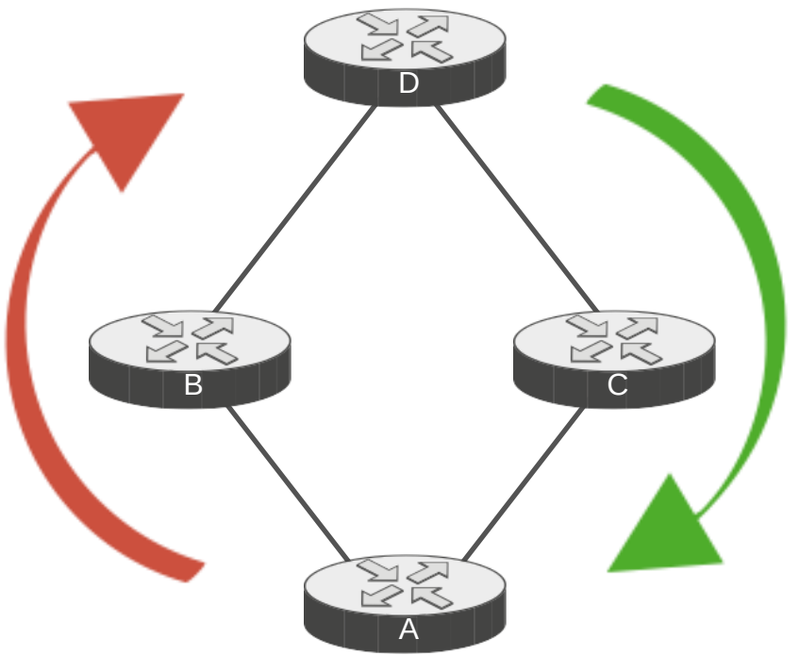
- After removing the Full Mesh IBGP sessions, BGP reachability to the prefixes should be verified again

BGP – IGP Interactions

- When a router needs to be restarted or maybe having a resource issue, operator may want to remove it from the network path

- In order to do this without losing packet, a router signal it’s IGP neighbor that they shouldn’t send the traffic towards it anymore

BGP – IGP Interactions



IS-IS Overload Bit
OSPF Max-Metric Router LSA
If BGP as an IGP , BGP Graceful Shutdown Community
EIGRP Stub Feature

Node B or C removal can create packet loss
if nodes don’t signal their neighbors with increased IGP metric

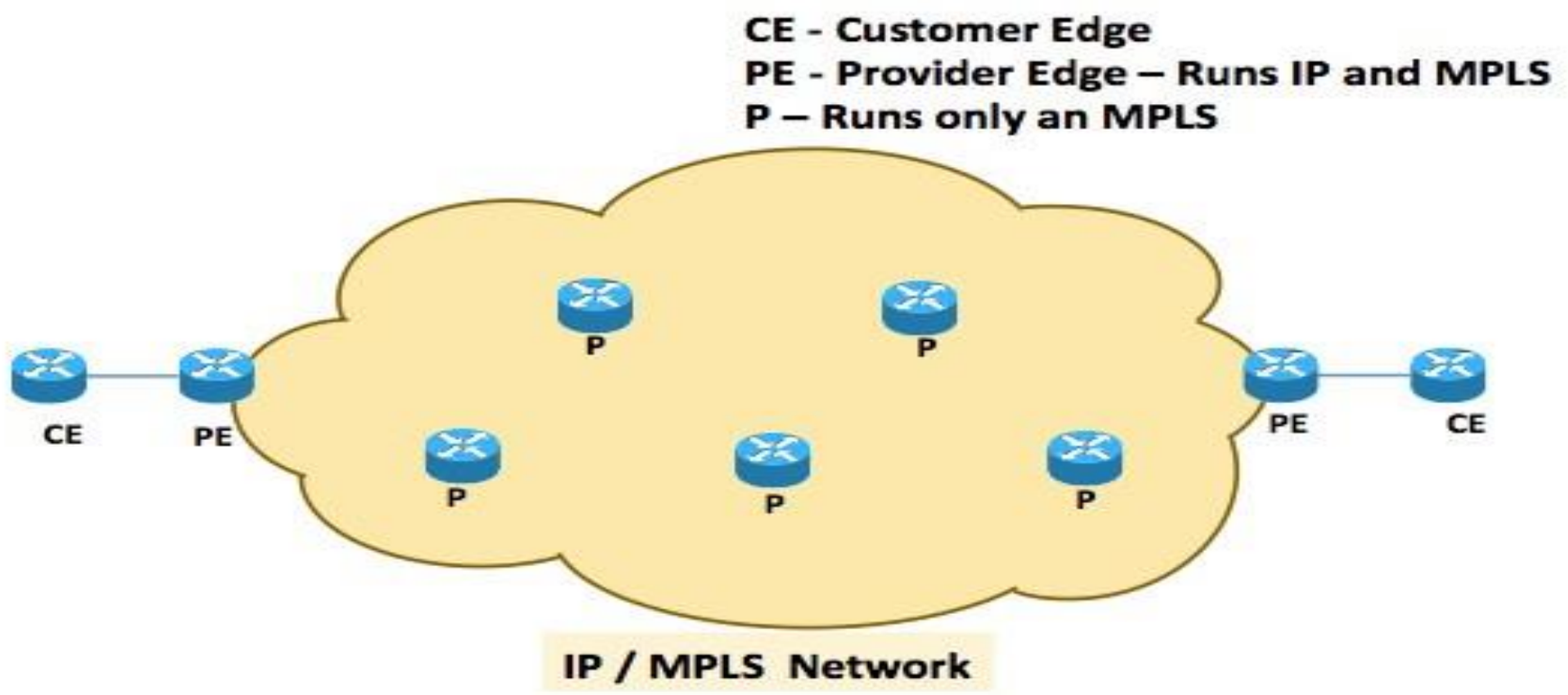
BGP – MPLS Interactions

- When BGP and MPLS is used together, generally it is used for VPN services

- MPLS removes the requirement of having BGP in the Mid/Core routers

- This phenomena is known in design as ‘ BGP Free Core ‘ design

BGP – MPLS Interactions



BGP Only needs to run at the PE routers in the above topology

BGP Labeled Unicast (BGP LU)

- BGP LU (Labeled Unicast) is used in the multi domain networks to connect the domains or in a single domain to advertise MPLS labels for the BGP next hops

- It was specified first in RFC 3107 ‘ Carrying Label Information in BGP-4 ‘ and then updated with RFC 8277

BGP Labeled Unicast (BGP LU)

- Practical deployment cases with BGP LU is Inter-AS Option C , Carrier Supporting Carrier and Seamless/Unified MPLS Scenarios

- BGP LU is used both as an Intra AS (With Seamless MPLS) and Inter-AS Routing Architecture (Inter-AS Option C , CSC)

BGP Labeled Unicast (BGP LU)

- With BGP LU , BGP sends the IPv4 prefix + Label (Address Family Identifier (AFI) 1 and Subsequent Address Family Identifier (SAFI) 4) , which is different than sending a Label for the VPN prefix (Ex: MPLS L3 VPN)

- We will discuss Seamless MPLS , Inter AS MPLS VPNs and Carrier Supporting Carrier (CSC) in the MPLS lesson

BGP LS (Link State and TE Information Distribution using BGP

- RSVP-TE have been providing resource allocation and provide an LSP with the distributed path computation algorithm (CSPF) since decades

- In order to overcome Bin Packing , Dead Lock or Network wide optimal traffic engineering, centralized controllers have been used for a long time

- RFC 7752 specifies the details of **North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP**

BGP LS (Link State and TE Information Distribution using BGP

- PCE (Path Computation Element) is an SDN controller which provides optimal path computation in Multi Area and Multi AS (Autonomous System) deployments

- It requires Link State and Traffic Engineering attributes such as Link coloring , SRLG , reserved bandwidth etc., from the network

- Link state IGP protocols (OSPF, IS-IS) can be used for this purpose but they are considered chatty and non-scalable

BGP LS (Link State and TE Information Distribution using BGP

- BGP LS is used to distribute Link state information and traffic engineering attributes from the network nodes to the Centralized TE controller

BGP EPE (Egress Peer Engineering)

- Monetary cost , latency and packet loss are important parameters for the Quality of User Experience and Traffic Engineering can be done by optimizing any of the above parameters

- BGP NLRIs don’t provide the information about the cost, latency or loss of the path or exit point for the destinations

BGP EPE (Egress Peer Engineering)

- The data-plane interconnection link (NNI) and control-plane (eBGP) direct connection between two ASs allows Internet traffic to travel between the two, usually as part of a formal agreement called *peering*

- This peering can be settlement free based or settlement based (Ex : IP Transit)

BGP EPE (Egress Peer Engineering)

- The selection of the best exit link for a given destination prefix selection and the enforcement of this selection on a network is not simple task. This is because the decision for one prefix might impact other traffic by changing the utilization of the NNI link and potentially leading to overload

- This is an end to end traffic engineering requirement !

Traditional EPE and the Limitations

- Traditionally, SPs use a policy to manipulate the BGP attributes contained in NLRIs received from a peer. This policy-based manipulation is usually performed on the Egress ASBR, but sometimes also on a route reflector (RR) and the Ingress ASBR

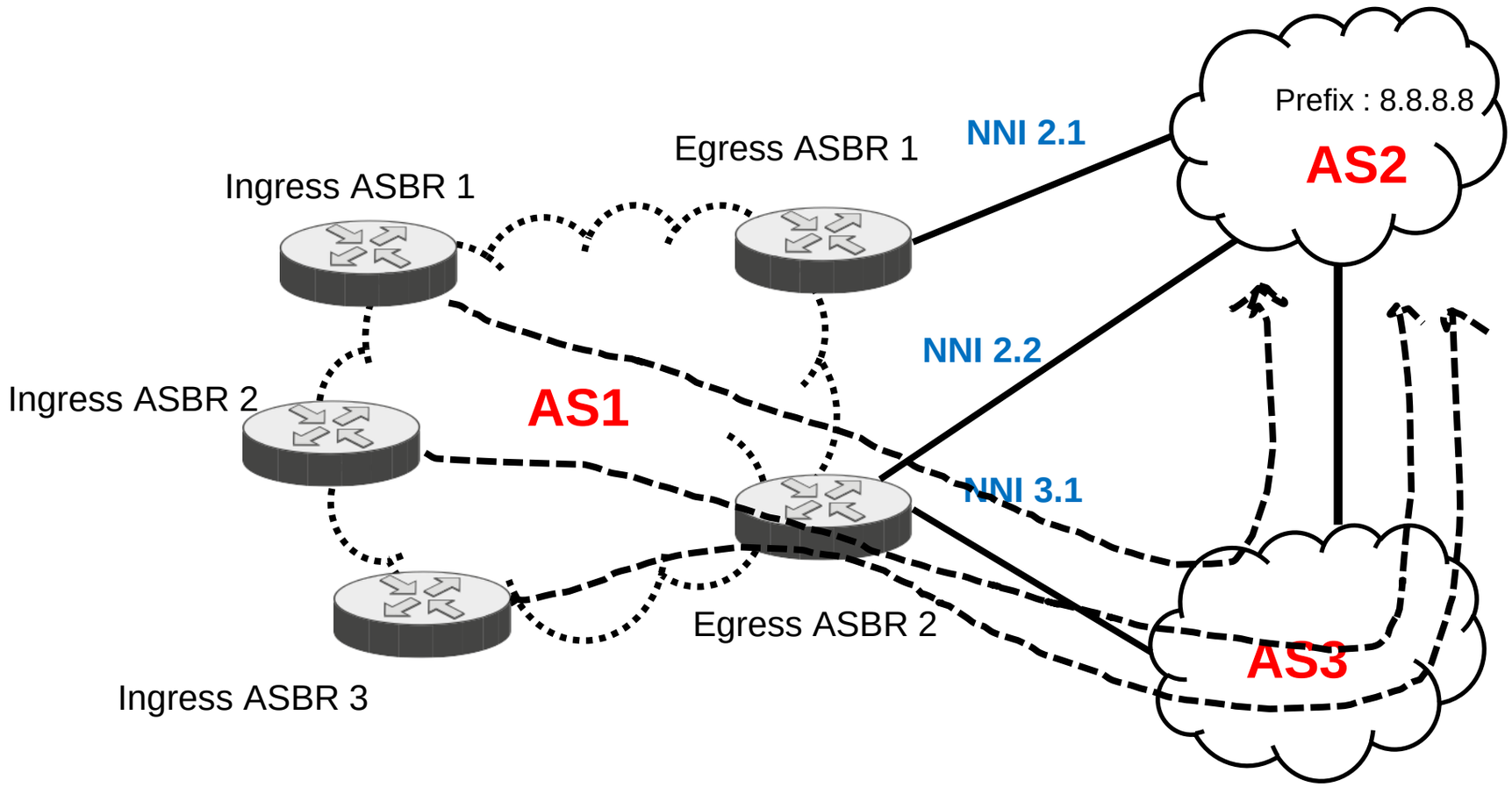
Traditional EPE and the Limitations

- This traditional technique provides some level of flexibility and control on how traffic leaves the SP and AS

- However, it is also limited by the BGP path selection algorithm and the fact that the results apply to all traffic for given prefix, regardless of the traffic's origin (Doesn't matter which Ingress ASBR sends the traffic)

Traditional EPE Example

Assume 8.8.8.8 is in AS 2 , AS2 and AS3 are peer
AS1 learns the prefix from both AS2 and AS3
If on Egress ASBR2 Local preference is higher for 8.8.8.8 towards AS3, NNI3.1 link is used by every Ingress ASBR



Modern – Better way of EPE

- If EPE had the ability to distribute traffic among several egress links based not only on destination address, but also by considering the ingress ASBR (or ingress port etc.), this would provide much finer granularity and also bandwidth management

- This would be especially true if EPE were combined with traffic statistics and centralized optimization

Modern – Better way of EPE

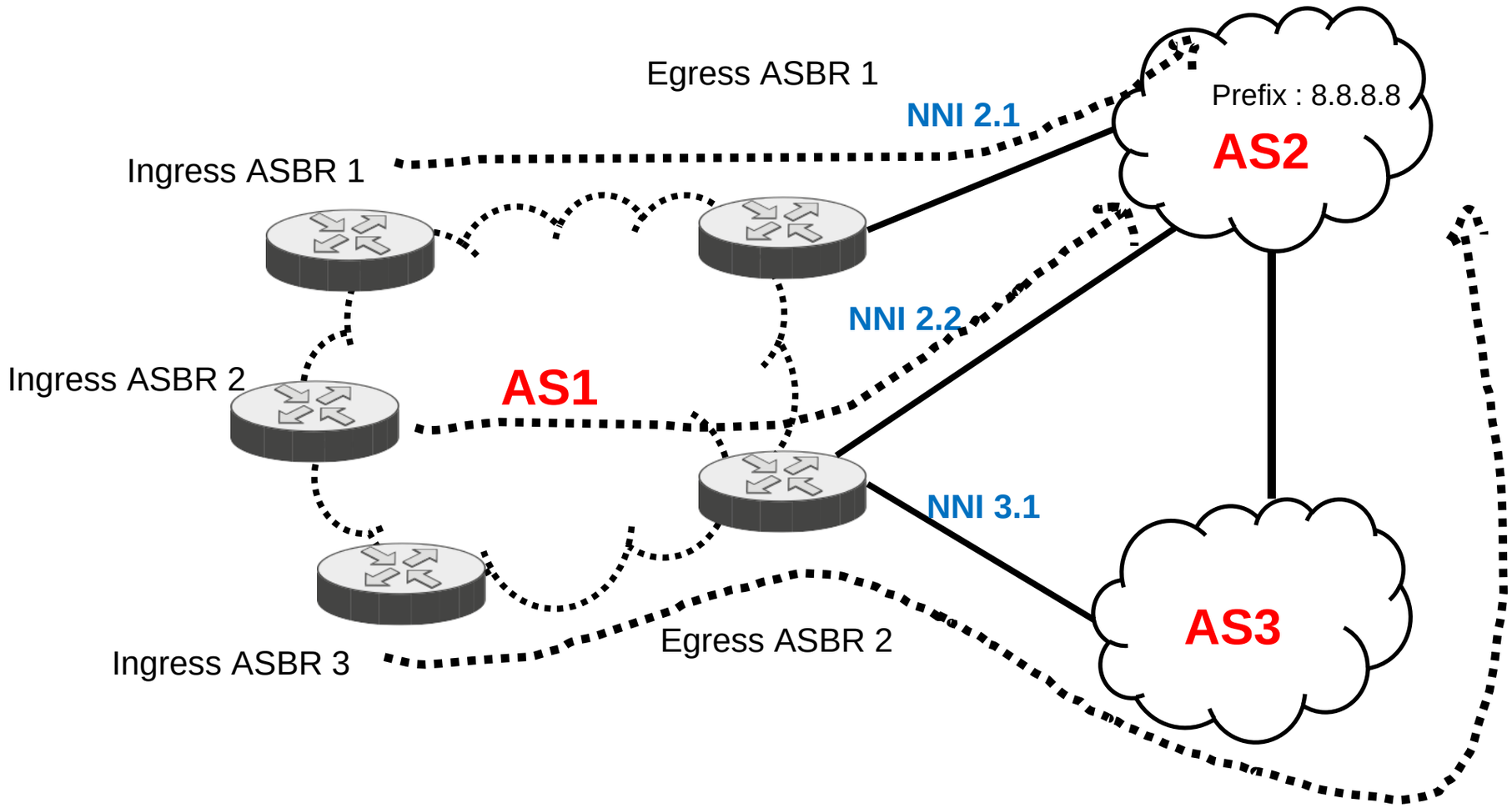
- The EPE solution should direct traffic for a given prefix that enters the network on a particular Ingress ASBR to a particular egress NNI (Egress Link) on a particular Egress ASBR

Modern – Better way of EPE

Better Policy for utilization of both internal and inter-domain resources (NNI Bandwidth etc.) can be as follows for AS 1

- [Ingress ASBR1, 8.8.8.8] to NNI 2.1
- [Ingress ASBR2, 8.8.8.8] to NNI 2.2
- [Ingress ASBR3, 8.8.8.8] to NNI 3.1

Modern – Better way of EPE



Modern EPE Requirements

- The utilization data for each of NNI. This data is provided by traditional or modern telemetry infrastructure (for example, SNMP interface statistics)

- The reachability information for destination IP prefixes This information is provided by eBGP advertisement from peer ASs

Modern EPE Requirements

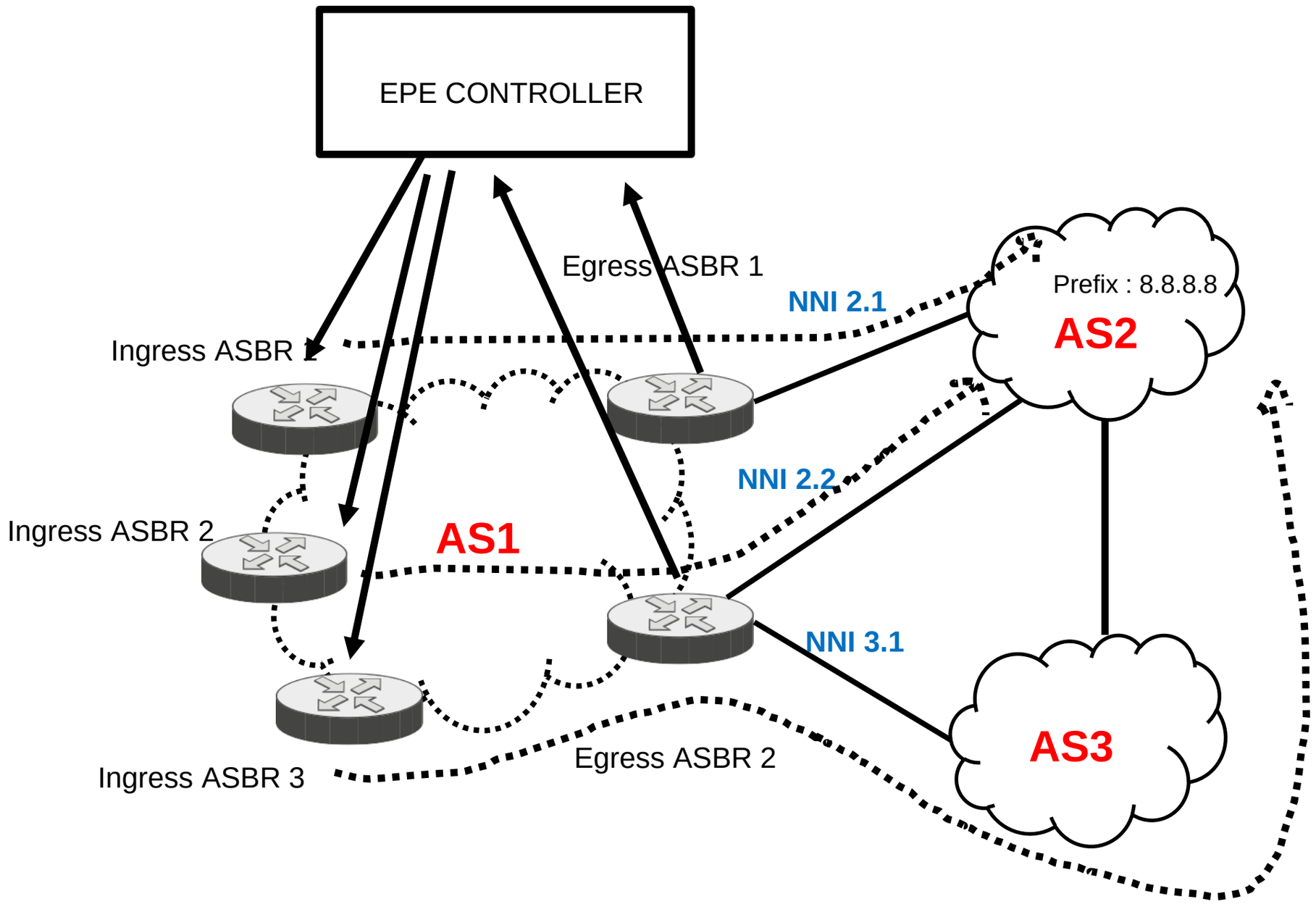
- The fine-grain partitioning of egress traffic into “flows” with information about the traffic volume carried by each

- An Egress Peering Engineering (EPE) controller that executes some logic to map these “flows” to NNI in a globally optimal way (Distributed TE cannot provide global optimality, Bin Packing , Dead-Lock are known problems)

Modern EPE Requirements

- A network infrastructure that allows forwarding traffic from an ingress AS Border Router (ASBR) in the service provider network to the designated egress NNI, as determined by the EPE controller

Modern – Better way of EPE



BGP RTBH (Remotely Triggered Blackholing)

- Remotely triggered blackholing is used for DDOS prevention for a long time

- DDOS attacks have an economical impact

- According to NBC News article, More than 40% of DDOS Attacks cost $1 million per day

BGP RTBH (Remotely Triggered Blackholing)

- Remote Triggered Blackhole is a technique which is used to mitigate DDOS attack dynamically

- Before RTBH, customer used to call Operator when there is an attack, Operator NOC engineer used to connect to the attacked network, trace the source of the attack, place the filters accordingly and attack goes away

BGP RTBH (Remotely Triggered Blackholing)

- Manual operation is open to configuration mistakes, cannot scale in large networks and between the attack and the required action, services stay down

BGP RTBH (Remotely Triggered Blackholing)

There are two types of RTBH

- Destination based RTBH
- Source based RTBH

Destination BGP RTBH (Remotely Triggered Blackholing)

- First RTBH idea was Destination based RTBH

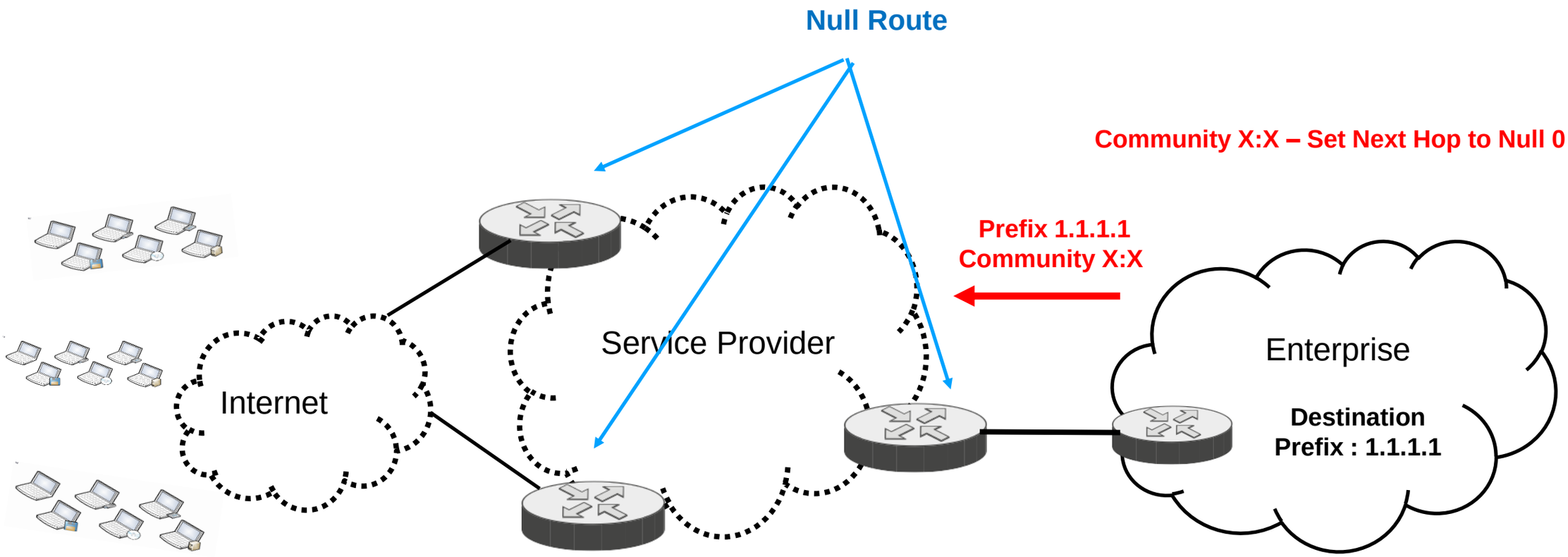
- With this technique, SP and the Customer agree on the discard community

Destination BGP RTBH (Remotely Triggered Blackholing)

- When there is an attack to the server, victim (customer) send the server prefix with the previously agreed community value

- When SP receives the update with that community, action is set to next hop to null, so packet is dropped before reaching to the customer link

Destination BGP RTBH (Remotely Triggered Blackholing)



Destination BGP RTBH (Remotely Triggered Blackholing)

- Problem with this attack, server will not be reachable from the legitimate sources too

- Attack is completed but at least the other services might stay up

- Also customer might change the IP address of the attacked server in DNS, which might take time to propagate this though

Destination BGP RTBH (Remotely Triggered Blackholing)

- RFC 3882 covers Destination based RTBH

- Better than manual processing

- Requires pre-configuration of null route on all edge routers in the SP network

Source BGP RTBH (Remotely Triggered Blackholing)

- RFC 5635 brings the idea of Source RTBH

- Instead of customer specifying the attacked system IP address to the SP, customer calls SP that they are under attack

- By combining uRPF and discard route (null route) configuration, based on the attack source, DDOS is mitigated (In theory)

uRPF and S/RTBH

- uRPF is Unicast Reverse Path Forwarding (Not Filtering☺) is defined in RFC 3704

- It is designed to limit the impact of DDOS attacks, by denying traffic with spoofed addresses access to the network

uRPF and S/RTBH

- Routers make their forwarding decisions based on Destination IP Address

- With uRPF, router looks at the Source IP address as well

- There are four types of uRPF : Strict, Loose , Feasible Path and VRF mode

Strict and Loose Mode uRPF

- Routers look at Source IP and then the Routing Table

- If source is reachable via the input interface , then it is forwarded, otherwise packet is dropped , this mode is called Strict mode uRPF

- If source is reachable via any route in the routing table then it is forwarded, otherwise packet is dropped, this mode is called Loose mode uRPF

Strict and Loose Mode uRPF

- uRPF Strict mode is generally used for Single Homed Customers

- For Multihomed customers, traffic can come from different interfaces, thus uRPF Loose mode should be used

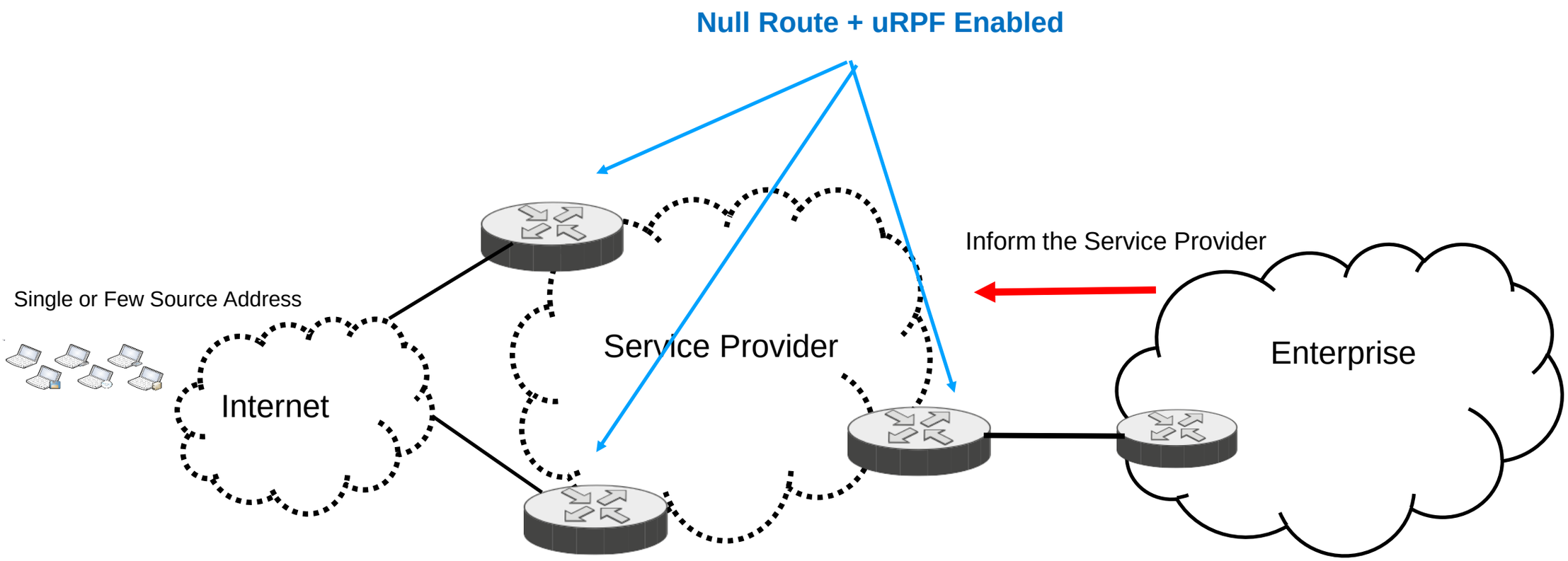
- For Source Based RTBH, Source addresses are injected into the BGP and with uRPF check traffic is sent to the blackhole (Null)

Strict and Loose Mode uRPF

- BGP Injector Machine is used to quickly detect source addresses and placed into BGP

- Exa BGP is commonly used as an Injector tool

Source BGP RTBH (Remotely Triggered Blackholing)



Source BGP RTBH (Remotely Triggered Blackholing)

- Advantage of Source RTBH over Destination RTBH is customer's (victim) destination address is still usable

- This method only useful if source is single or few addresses

BGP Flowspec

- Defined in RFC 5575
- New AFI and SAFI is defined for BGP Flowspec
- AFI/SAFI : 1/133 : Unicast Traffic Filtering Applications
- Instead of blackholing entire IP address, there are many fields in the flow which can be used for filtering

BGP Flowspec Matching Fields

- Type 1: Destination Prefix
- Type 2 : Source Prefix
- Type 3 : IP Protocol
- Type 4 : Source or Destination Port
- Type 5 : Destination Port
- Type 6 : Source Port

BGP Flowspec

- Flow routes are automatically validated against unicast routing information or via routing policy

- When validated, firewall filter is created based on match and action criteria

BGP Flowspec Matching Fields

- Type 7 : ICMP Type
- Type 8 : ICMP Code
- Type 9 : TCP Flags
- Type 10 : Packet Length
- Type 11 : DSCP
- Type 12 : Fragment Encoding

BGP Flowspec Actions

- After identifying the flow, any of the below actions can be taken:
  - Rate limit to 0 (Drop the traffic)
  - Remarking
  - Redirect to VRF (Route Target)
  - Sampling the Traffic (Monitoring)

BGP Flowspec Vendor Support

- BGP Flowspec is supported by Commercial vendors and the Open Source implementation

- Commercial vendors , Arbor Peakflow , Juniper DDOS Secure , Alcatel Lucent , JunIper JUNOS , Cisco on ASR and CSR

BGP Flowspec and other DDOS Mitigation Approaches

- BGP Flowspec, compare to other methods such as manual filtering , Destination or Source based RTBH , is more granular

- BGP Flowspec can filter the traffic based on many other fields in the IP Packet , still provides same level of automation as RTBH

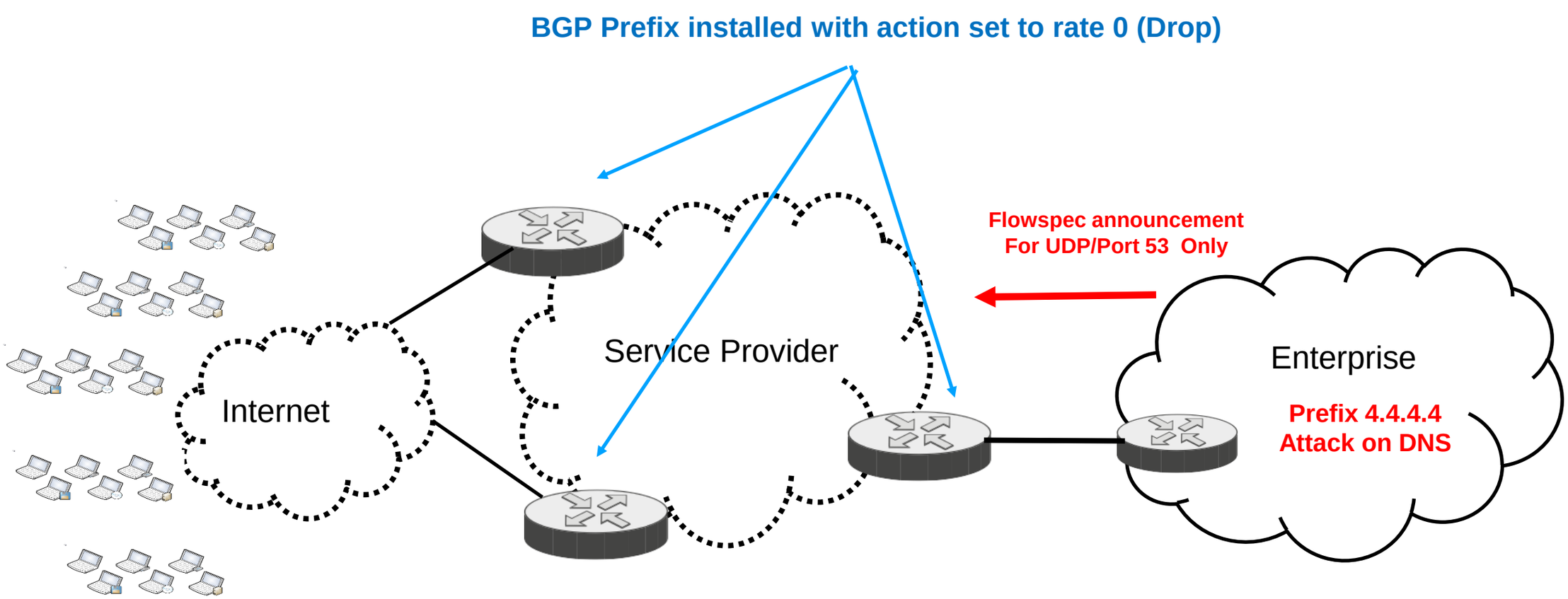
- Automation means, distributing the filtering policy rules to the edge nodes in the network, based on the matching criteria

How BGP Flowspec Works?

- Service Provider allows Customer to advertise Flowspec routes

- When attack starts towards customer (victim) IP address, customer initiates the filter , not just entire IP address but based on the attack type (DNS attack , NTP attack etc.)

BGP Flowspec Example



Who initiates Flowspec filters, SP or Customers?

- With BGP Flowspec, customer can initiate the filtering rules by advertising flowspec routes , saying to the SP that filter UDP port 53 for their specific IP address

- This is not mandatory though. If SP doesn’t provide this option, maybe because they don’t trust their customer expertise , SP may want customer to call them or through customer portal etc. when they are under attack and SP can initiate Flowspec filters , this option requires manual operation thus can take more time to stop attack

BGP Session Culling

- BGP session culling mitigates the negative impact of maintenance activities on IP networks, specifically on the IXP (Internet Exchange Points)

- RFC 8327 published for the BGP Session Culling

BGP Session Culling

- The approach is to ensure BGP-4 sessions that will be affected by maintenance are forcefully torn down before the actual maintenance activities start on the lower layers, such as physical layer

BGP Session Culling

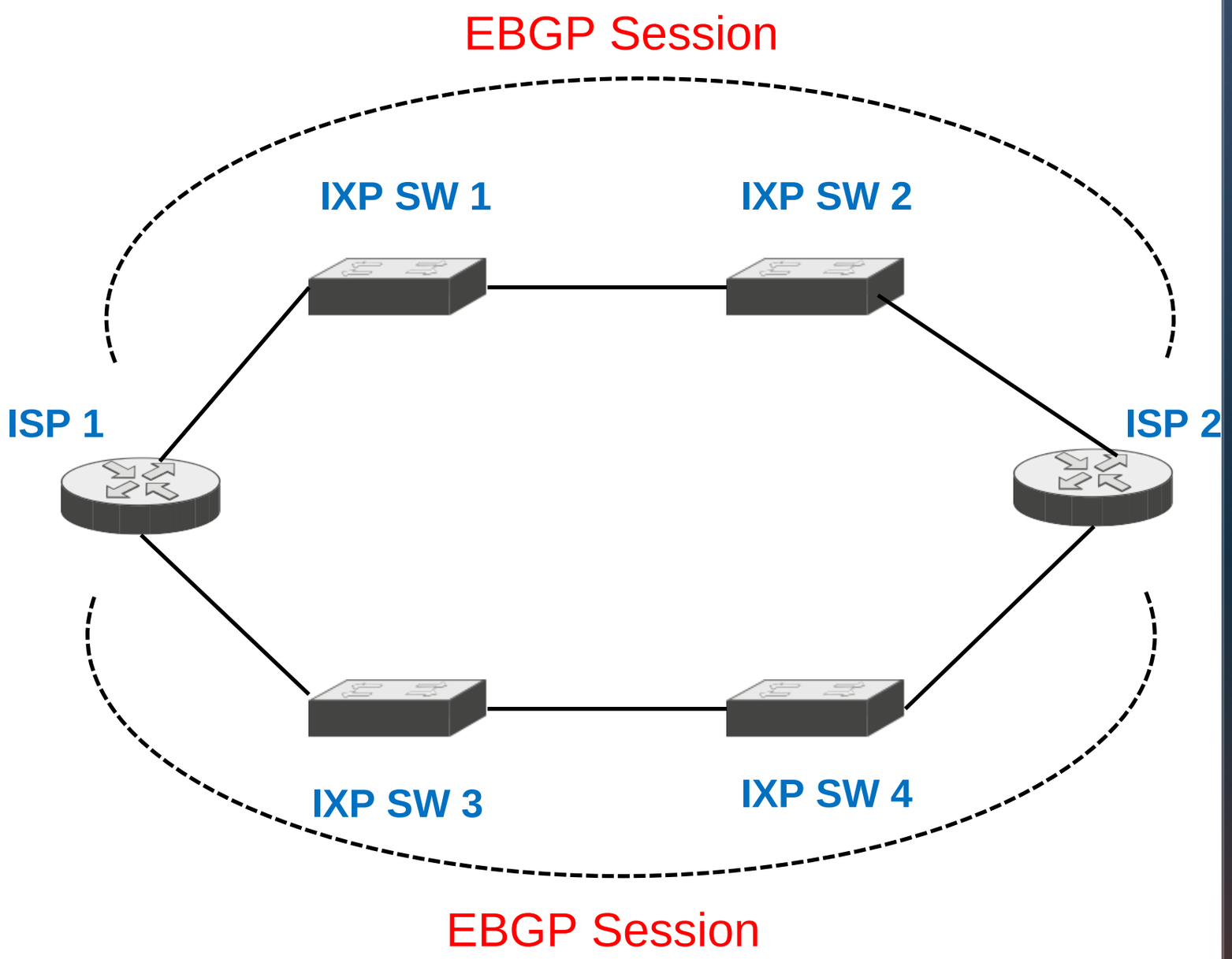
- BGP Session Culling minimizes the amount of disruption that lower- layer network maintenance activities cause, by making BGP speakers preemptively converge onto alternative paths while the lower-layer network's forwarding plane remains fully operational

BGP Session Culling

There are two paths between ISP1 and ISP2
Two EBGP sessions are created between two operators

BGP Session Culling is basically a Layer4 Access list (ACL) applied on the IXP Layer2 ports before maintenance. The ACL blocks traffic to and from IXP subnet, BGP port TCP 179, allowing all other traffic

The ACL causes that BGP Hold timer expires, BGP sessions are culled down and end-user traffic can be rerouted over alternative paths. Afterwards maintenance is commenced.



BGP Session Culling Alternatives

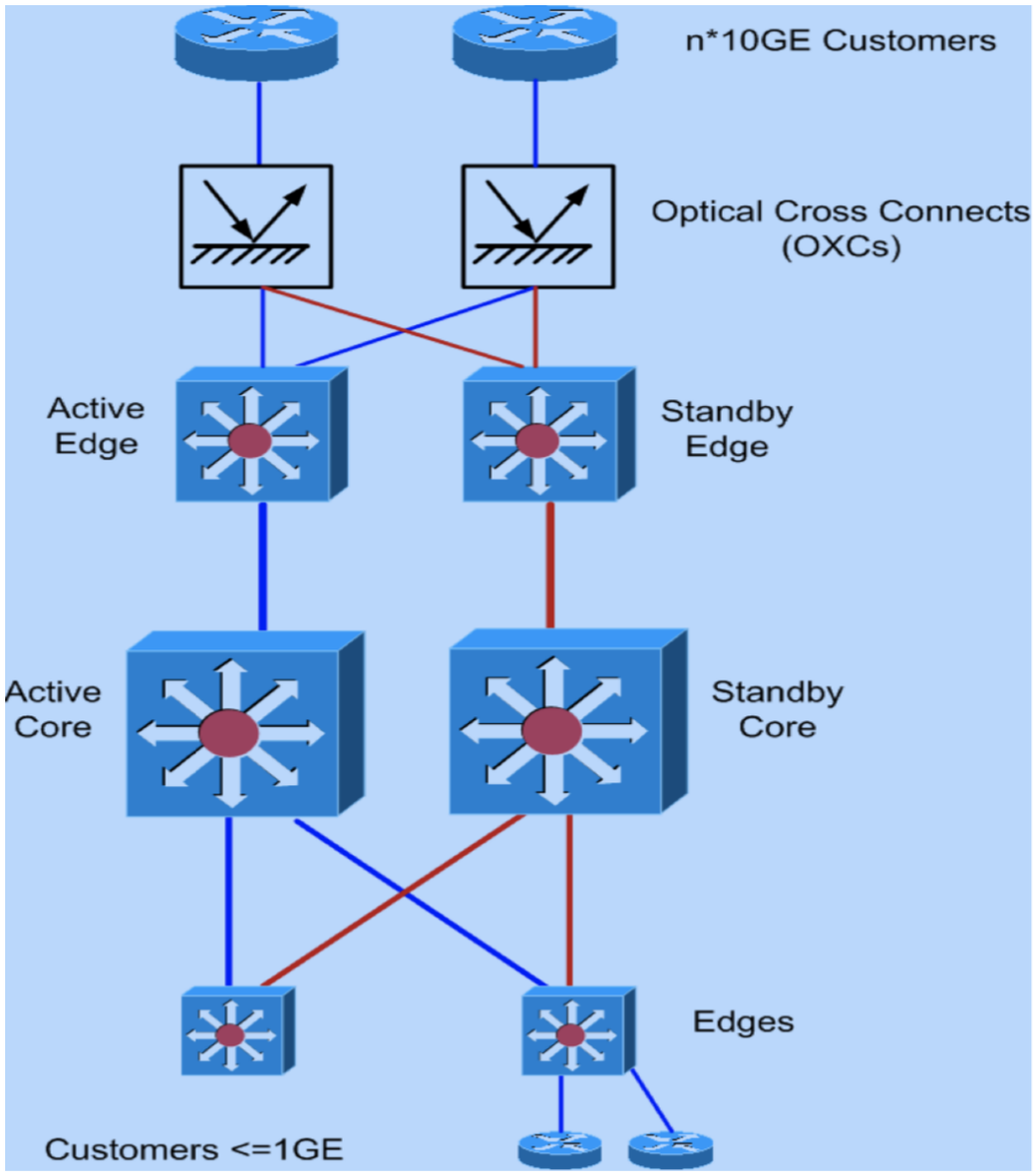
BGP Session Culling is described as Involuntary BGP Session Teardown technique in the RFC 8327. Lower layer network operator , such as IXP , cannot access to the IXP member routers , thus cannot bring the BGP session down which would help to avoid dataplane traffic loss

Alternative to Involuntary BGP Session Teardown, BGP Session can be teardown voluntarily. In this case, ISP shutdown the BGP session, IXP operator monitors the dataplane and when the traffic reaches to a minimum, IXP operator starts maintenance on the switches

Some IXPs use Optical Switch and terminate their IXP members to the Optical Cross Connect (OXC) Switch which is connected to the Fabric Switches

BGP Session Culling Alternatives

- AMS-IX (Amsterdam Internet Exchange) uses Optical Cross Connect Switch to have redundancy. It helps to redirect traffic to the second plane when there is a maintenance activity on the first plane



AMS-IX Customer to Fabric Connections

BGP Session Culling Summary

- BGP session culling gains popularity and is applied at more and more IXPs
- It mitigates a negative impact of maintenance activities while requiring no input from the ISPs.
- It is a IETF standard
- Alternative method : Voluntary BGP Session Teardown, Optical Cross Connect (OXC) Switches
- Session Culling is a Involuntary BGP Session teardown technique

BGP Graceful Restart, BGP Graceful Shutdown and BGP Administrative Shutdown Communication

- RFC 4724 specifies BGP Graceful Restart procedure
- RFC 8326 specifies BGP Graceful Shutdown
- RFC 8203 specifies BGP Administrative Shutdown Communication
- All three are separate mechanisms, provide different functionalities

BGP Graceful Restart

- GR ensures normal forwarding of data during the restart of routing protocols to prevent interruption of key services

- Graceful Restart is available today for OSPF, ISIS, EIGRP, LDP and BGP. Standards are defined for OSPF, ISIS, BGP and LDP to ensure vendor interoperability

- GR is usually used when the active route processor (RP) fails because of a software or hardware error, or used by an administrator to perform the master/slave switchover

- GR is known as NSF (Non Stop Forwarding)

BGP Graceful Restart

- During BGP peer relationship establishment, devices negotiate GR capabilities by sending supported GR capabilities to each other

- Dual processor systems which support Stateful Switch Over (SSO) or In-Service Software Upgrades (ISSU) can continue to forward traffic while restarting the control plane on the second processor

BGP Graceful Restart

- Usually, when BGP on a router restarts, all the BGP peers detect that the session went down and then came up. This "down/up" transition results in a "routing flap" and causes BGP route re-computation, generation of BGP routing updates, and unnecessary churn to the forwarding tables

BGP Graceful Restart

- The Graceful Restart is a BGP capability which is used by a BGP speaker to indicate its ability to preserve its forwarding state during BGP restart

- While control plane restarting, routers can forward the traffic, routes are marked as stale and removed after BGP session is re-established

BGP Graceful Shutdown

- BGP Graceful Shutdown is a well known BGP community

- BGP Graceful Shutdown is used for the planned maintenance activities

- The well-known community allows implementers to provide an automated graceful shutdown mechanism that does not require any router reconfiguration at maintenance time

BGP Graceful Shutdown

- Loss comes from transient lack of reachability during BGP convergence that follows the shutdown of an EBGP peering session between two Autonomous System Border Routers (ASBRs)

- Graceful Shutdown is used when GR is not applicable, for example during the maintenance forwarding/data plane might be impacted

BGP Graceful Shutdown

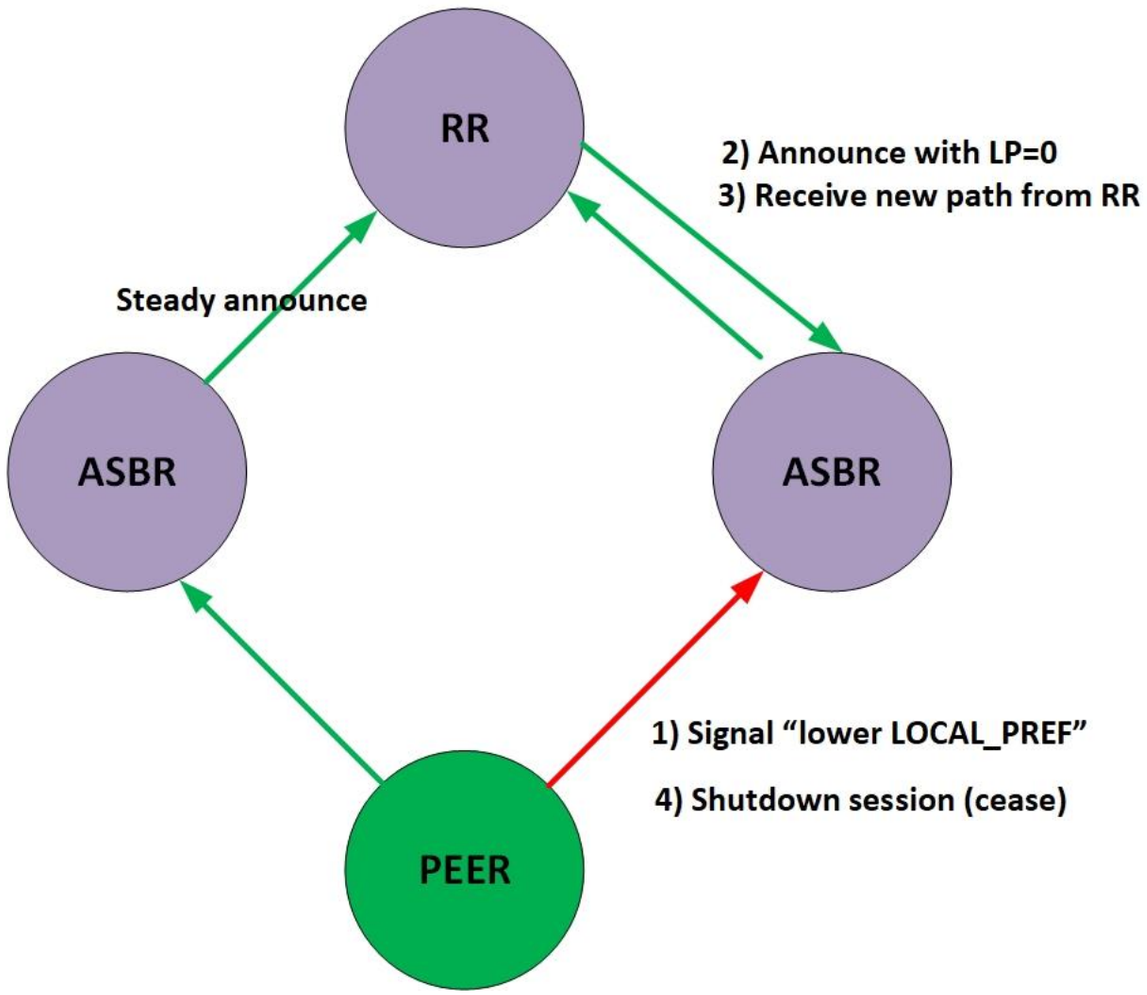
- Graceful Shutdown can be applied to reduce or avoid packet loss for outbound and inbound traffic flows initially forwarded along the peering link to be shut down

BGP Graceful Shutdown

- In both Autonomous Systems (ASes), Graceful Shutdown trigger rerouting to alternate paths if they exist within the AS while allowing the use of the old path until alternate ones are learned. This ensures that routers always have a valid route available during the convergence process

BGP Graceful Shutdown

Graceful Shutdown triggers “Path Hunting”



- Initiated by the operator on the router before maintenance by sending the GRACEFUL_SHUTDOWN well-known community (65535:0 as per IANA)
- Receiving EBGP peer sets LOCAL_PREFERENCE to 0 and selects paths to route traffic away from the initiator, (similar to setting overload in ISIS)
- When BGP session goes down, minimizes impact to traffic because alternate paths have already been installed.

BGP Administrative Shutdown Communication

- BGP Shutdown Communication is specified in RFC 8203

According to RFC 8203:

- BGP Administrative Shutdown Communication enhances the BGP Cease NOTIFICATION message "Administrative Shutdown" and "Administrative Reset" subcodes for operators to transmit a short freeform message to describe why a BGP session was shutdown or reset

BGP Administrative Shutdown Communication

- Operators, before the BGP session shutdown, inform the BGP peers with BGP Administrative Shutdown Communication procedure

BGP Administrative Shutdown Communication

- If a BGP speaker decides to terminate its session with a BGP neighbor, and it sends a NOTIFICATION message with the Error Code "Cease" and Error Subcode "Administrative Shutdown" or "Administrative Reset", it MAY include an UTF-8 encoded string

- They can send each other ‘ happy face ‘ emoji ☺

BGP Administrative Shutdown Communication

- Operators are encouraged to use the Shutdown Communication to inform their peers of the reason for the shutdown of the BGP session and include out-of-band reference materials

BGP Administrative Shutdown Communication

- An example of a useful Shutdown Communication would be: "[TICKET-1-1438367390] software upgrade; back in 2 hours"

- "[TICKET-1-1438367390]" is a ticket reference with significance to both the sender and receiver, followed by a brief human-readable message regarding the reason for the BGP session shutdown followed an indication about the length of the maintenance. The receiver can now use the string 'TICKET-1-1438367390' to search in their email archive to find more details

AIGP – Accumulated IGP Metric Attribute for BGP

- AIGP is specified in RFC 7311
- IGPs are designed to run within a single administrative domain and they make path-selection decision based on metric value
- BGP is an inter AS domain routing protocol and there is no inter AS metric which can be used for end to end shortest path selection
- AIGP is an Optional and Non-Transitive BGP Attribute

AIGP – Accumulated IGP Metric Attribute for BGP

- BGP is designed to provide routing over a large number of independent ASs with limited or no coordination among respective administrations. BGP does not use metrics in the path selection decisions

AIGP – Accumulated IGP Metric Attribute for BGP

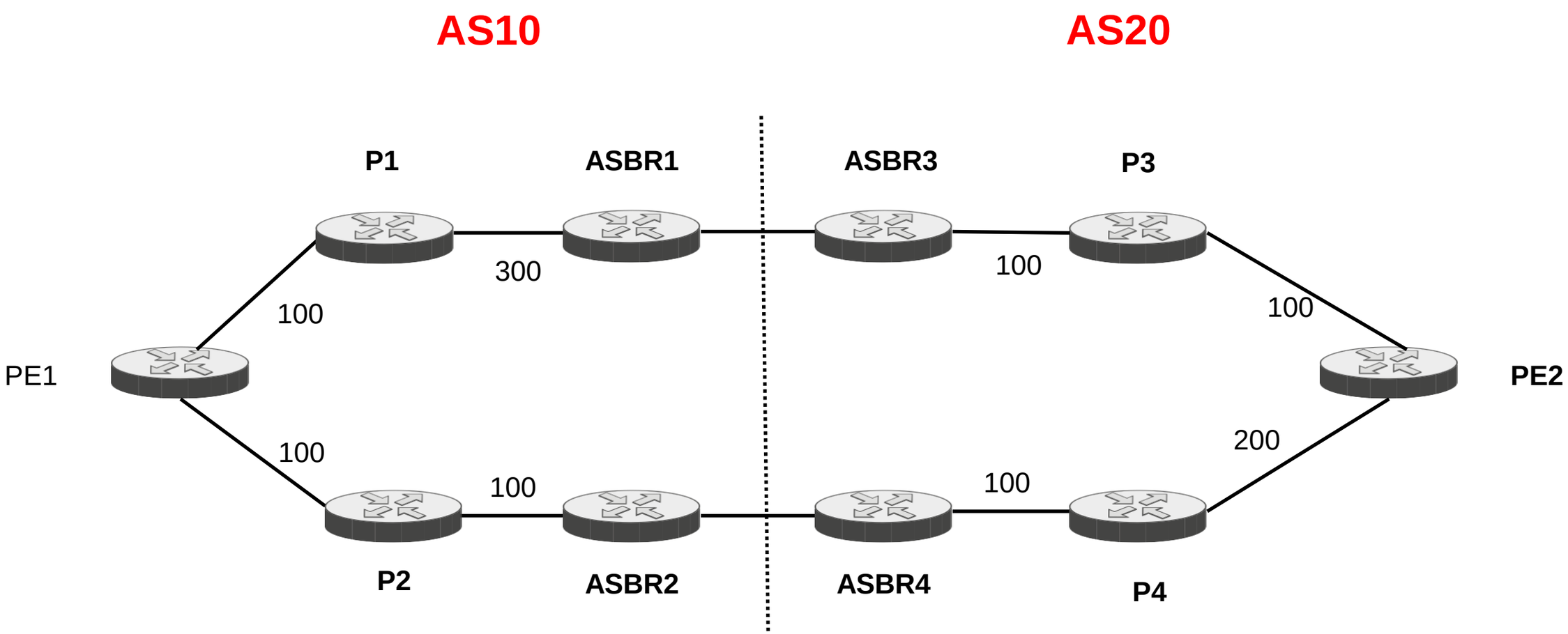
- The accumulated IGP (AIGP) metric attribute for BGP enables deployment in which a single administration can run several contiguous BGP ASs.
- Such deployments allow BGP to make routing decisions based on the IGP metric
- In such networks, it is possible for BGP to select paths based on metrics as is done by IGPs.
- In this case, BGP chooses the shortest path between two nodes, even though the nodes might be in two different ASs

AIGP – Accumulated IGP Metric Attribute for BGP

- AIGP impacts the BGP best-route decision process
- The AIGP attribute preference rule is applied after the local-preference rule.
- The AIGP distance is compared to break a tie in the BGP best path selection

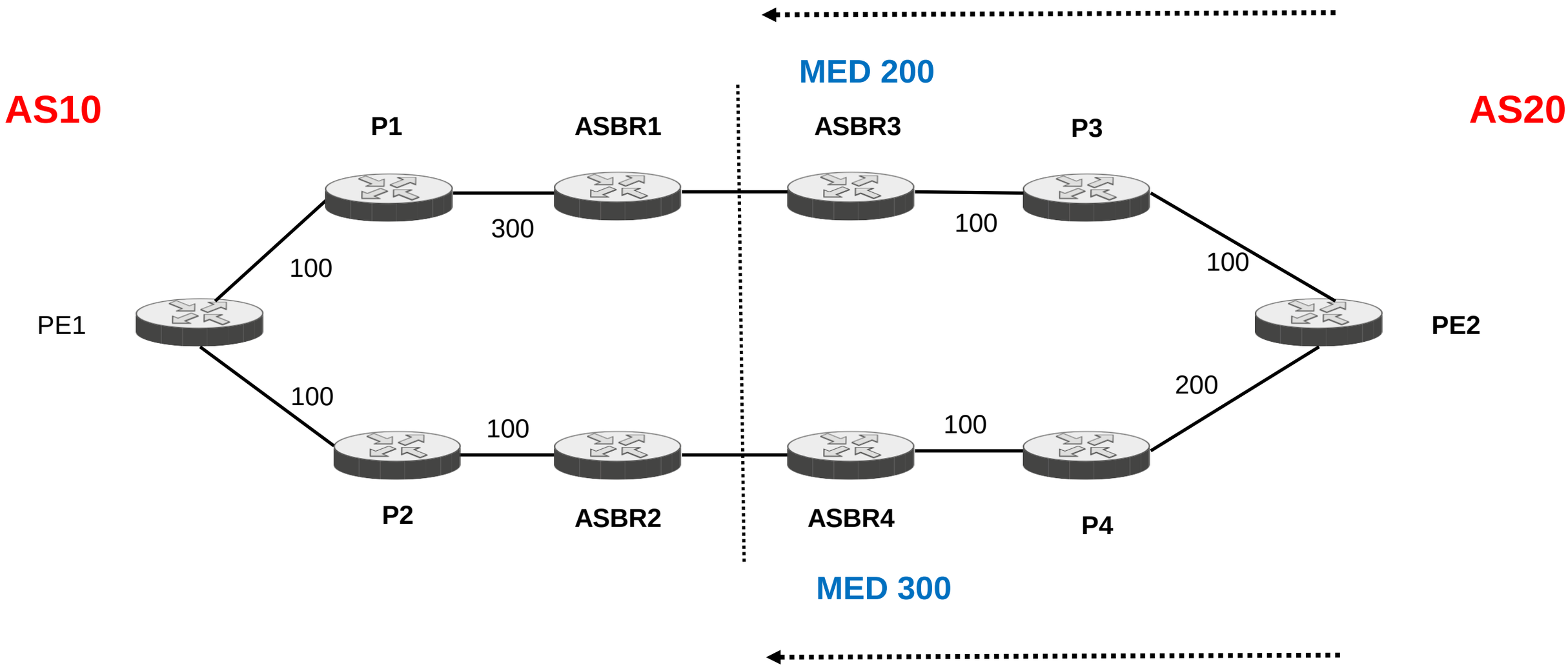
How AIGP Works

Bottom Path End to End IGP Cost is PE1 to ASBR2 + ASBR 4 to PE2 = 500
Top Path End to End IGP Cost is PE1 to ASBR 1 + ASBR3 to PE2 = 600
Thus better path is selected by PE1 to reach PE 2 if AIGP is enabled



BGP MED vs. AIGP Metric Attributes

Top Path AIGP Metric is 600 , BGP MED is 200
Bottom Path AIGP Metric is 500 , BGP MED is 300
If AIGP wouldn't be used, Top Path would be selected as Best ,
that is not end to end optimal path



EBGP Default Route Propagation Behavior without Policies– RFC 8212

- This is specified in RFC 8212 and updates RFC 4271

- Behavior of EBGP route propagation without Import and Export Route Policy created many problems on Internet

- With RFC 8212 routes are neither imported nor exported unless specifically enabled by configuration

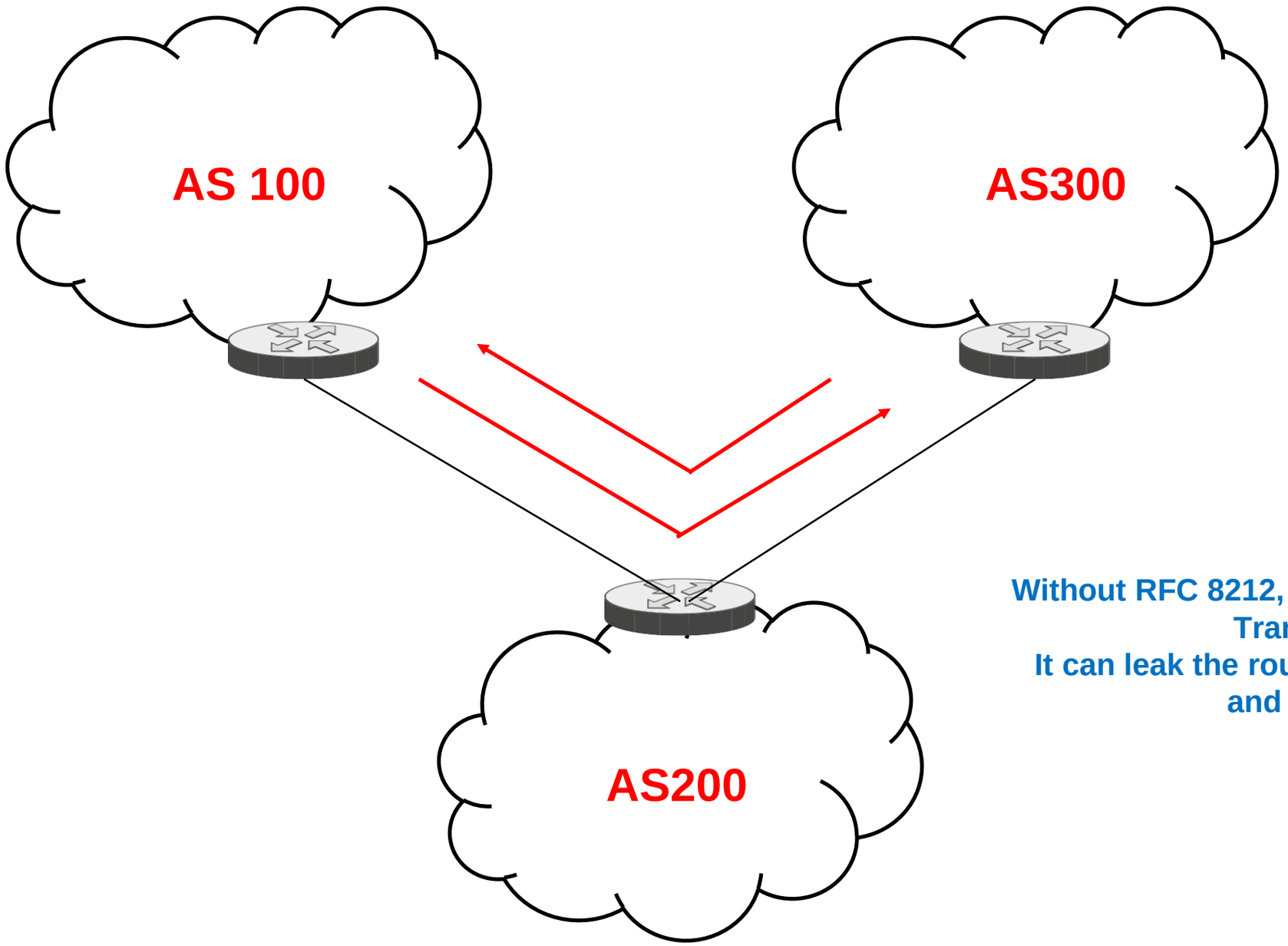
EBGP Default Route Propagation Behavior without Policies– RFC 8212

- Many deployed BGP speakers send and accept any and all route announcements between their BGP neighbors by default

- This behavior results with Route Leaks which will be covered in next topic

- Route Leak in general resulting in routing of traffic through an unexpected path

EBGP Default Route Propagation Behavior without Policies– RFC 8212



Without RFC 8212, AS 200 can become a Transit AS
It can leak the routes between AS 100 and AS 300

EBGP Default Route Propagation Behavior without Policies– RFC 8212

- As the Internet has become more densely interconnected, the risk of a misbehaving BGP speaker poses significant risks to Internet routing (Route Leak , Configuration mistakes etc.)

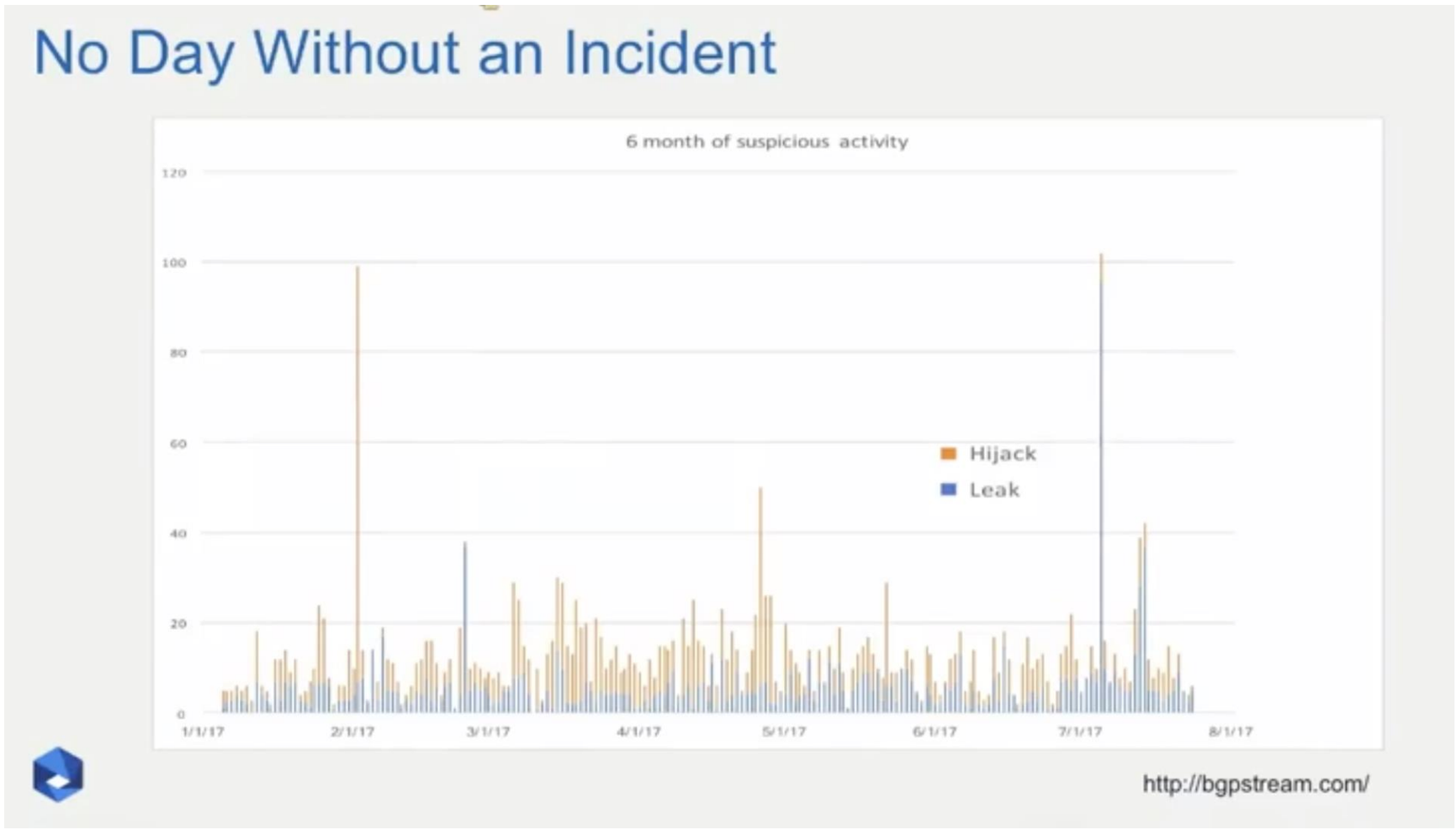
- RFC 8212 intends to improve this situation by requiring the explicit configuration of both BGP Import and Export Policies for any External BGP (EBGP) session such as customers, peers, or confederation boundaries for all enabled address families

BGP Information Security

- Border Gateway Protocol (BGP) is based entirely on trust between networks
- No built-in validation that updates are legitimate
- The chain of trust spans whole Internet , continents
- Lack of reliable data creates big issues !

BGP Information Security

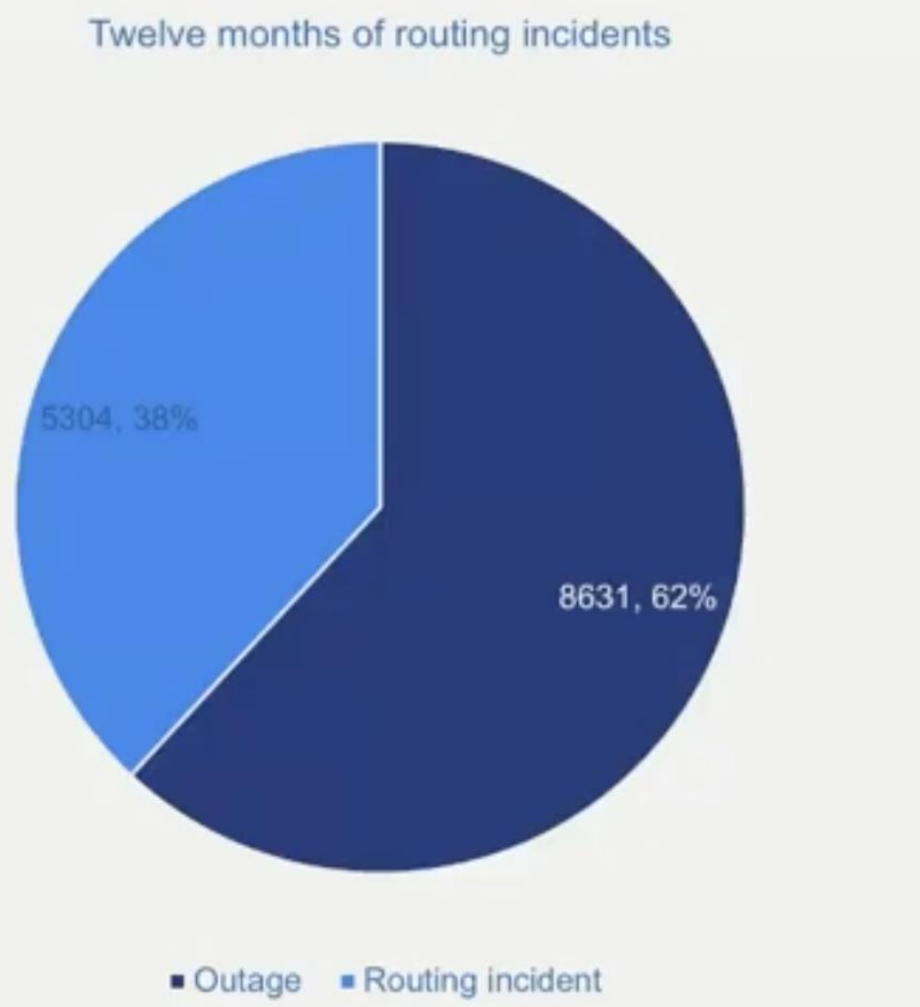
- We will deal with BGP Route Leaks, Different Types of BGP Hijackings and IP Address Spoofing



BGP Information Security

Global stats

- 13,935 total incidents (either outages or attacks like route leaks and hijacks)
- Over 10% of all Autonomous Systems on the Internet were affected
- 3,106 Autonomous Systems were a victim of at least one routing incident
- 1,546 networks caused at least one incident



BGP Information Security

We will be talking about many tools:

- Prefix Filtering , Distribute Lists, AS-PATH Filtering
- RPKI ROA, Validation, IRR Toolset, BGPQ3
- BGPSEC

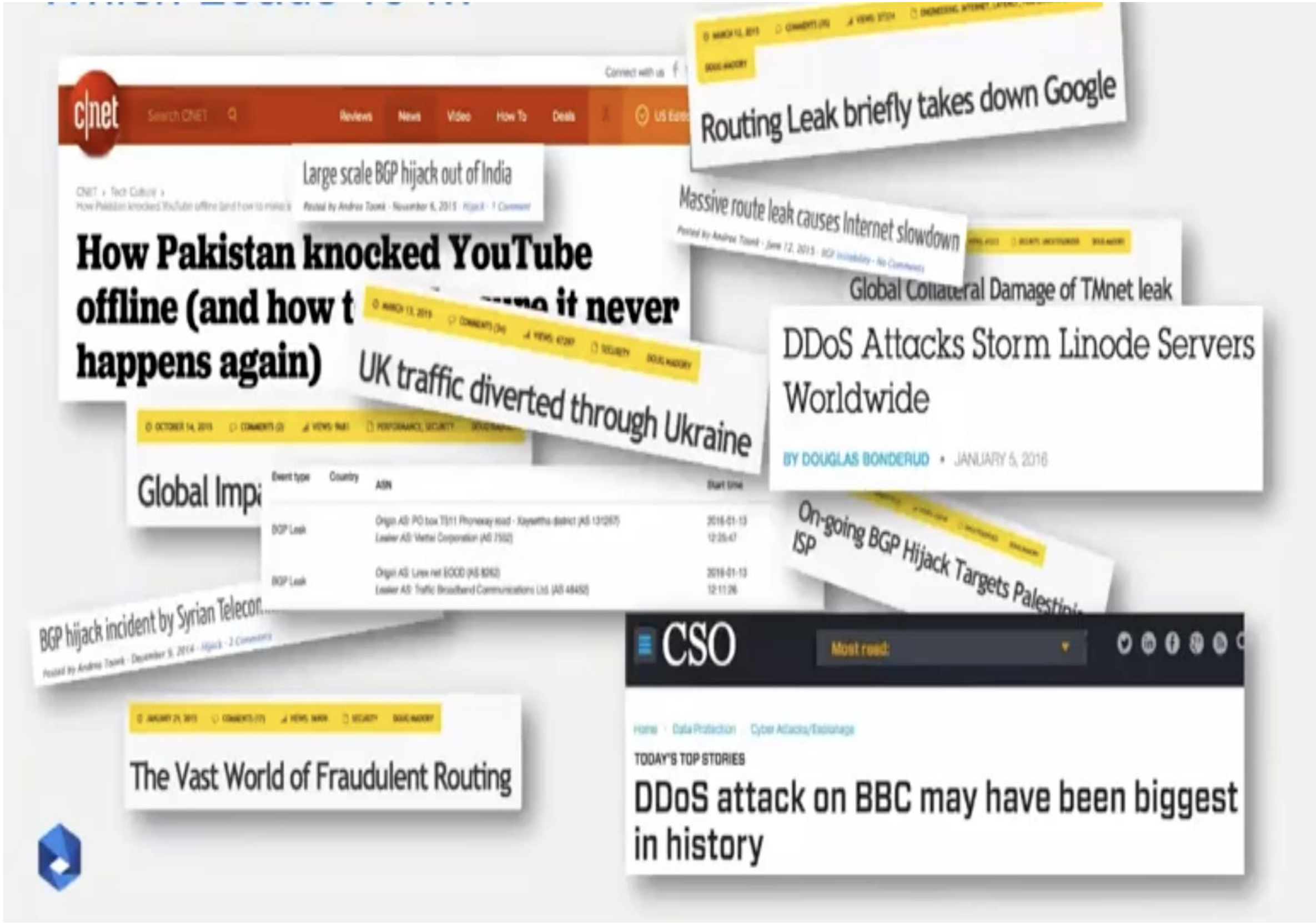
BGP Information Security

<u>Biggest Problems with current approaches for BGP Security:</u>

- Not enough deployment
- Lack of reliable data

Routing Incidents Cause Real World Problems

- Insecure routing is one of the most common paths for malicious threats
- Attacks can take anywhere from hours to months to even recognize
- Inadvertent error can take entire countries offline, while attackers steal money or data (Current Bitcoin Attack)


Routing Leak briefly takes down Google
Large scale BGP hijack out of India
How Pakistan knocked YouTube offline (and how t
happens again)
Massive route leak causes Internet slowdown
Global Collateral Damage of TMnet leak
UK traffic diverted through Ukraine
DDoS Attacks Storm Linode Servers Worldwide
BY DOUGLAS BONDERUD • JANUARY 5, 2016
Global Impa
Event type
Country
ASN
BGP Leak
BGP Leak
On-going BGP Hijack Targets Palestini
ISP
BGP hijack incident by Syrian Telecom
The Vast World of Fraudulent Routing
CSO
Most read:
Home · Data Protection · Cyber Attacks/Espionage
TODAY'S TOP STORIES
DDoS attack on BBC may have been biggest in history

We will deal with BGP Route Leaks, Different Types of BGP Hijacking and IP Address Spoofing

The Threats: What's Happening?

Event	Explanation	Repercussions	Solution
Prefix/Route Hijacking	A network operator or attacker impersonates another network operator, pretending that a server or network is their client.	Packets are forwarded to the wrong place, and can cause Denial of Service (DoS) attacks or traffic interception.	Stronger filtering policies
Route Leak	A network operator with multiple upstream providers (often due to accidental misconfiguration) announces to one upstream provider that is has a route to a destination through the other upstream provider.	Can be used for traffic inspection and reconnaissance.	Stronger filtering policies
IP Address Spoofing	Someone creates IP packets with a false source IP address to hide the identity of the sender or to impersonate another computing system.	The root cause of reflection DDoS attacks	Source address validation

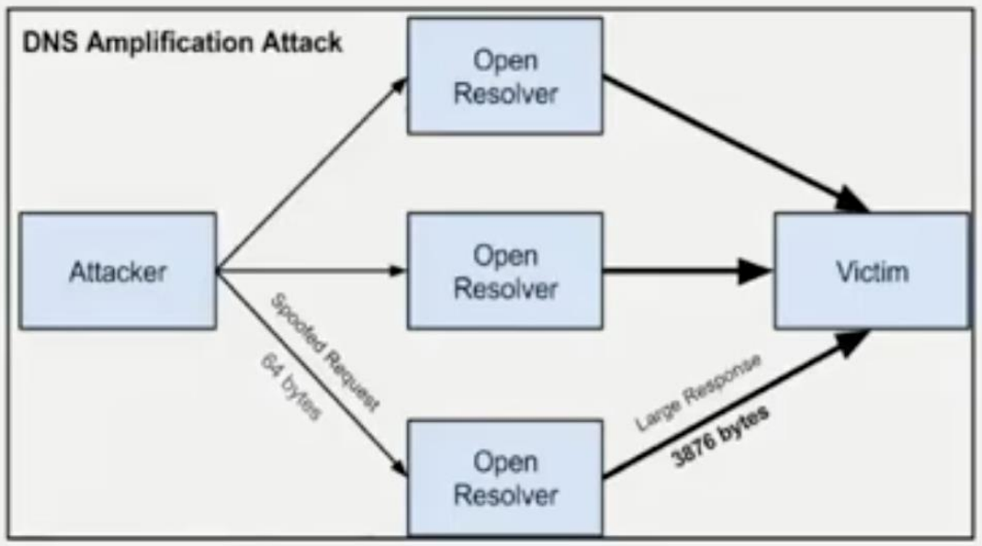
IP Address Spoofing

IP Address Spoofing

IP address spoofing is used to hide the true identity of the server or to impersonate another server. This technique can be used to amplify an attack.

Example: DNS amplification attack. By sending multiple spoofed requests to different DNS resolvers, an attacker can prompt many responses from the DNS resolver to be sent to a target, while only using one system to attack.

Fix: Source address validation: systems for source address validation can help tell if the end users and customer networks have correct source IP addresses (combined with filtering).



BGP Information Security

- BGP Information Security deals with the anything related with the network traffic, it doesn’t deal with the BGP Transport Security, such as TCP AO , MD5 Authentication etc.

BGP Information Security – Route Leaks

- BGP Route Leak is dangerous for the network traffic in many ways

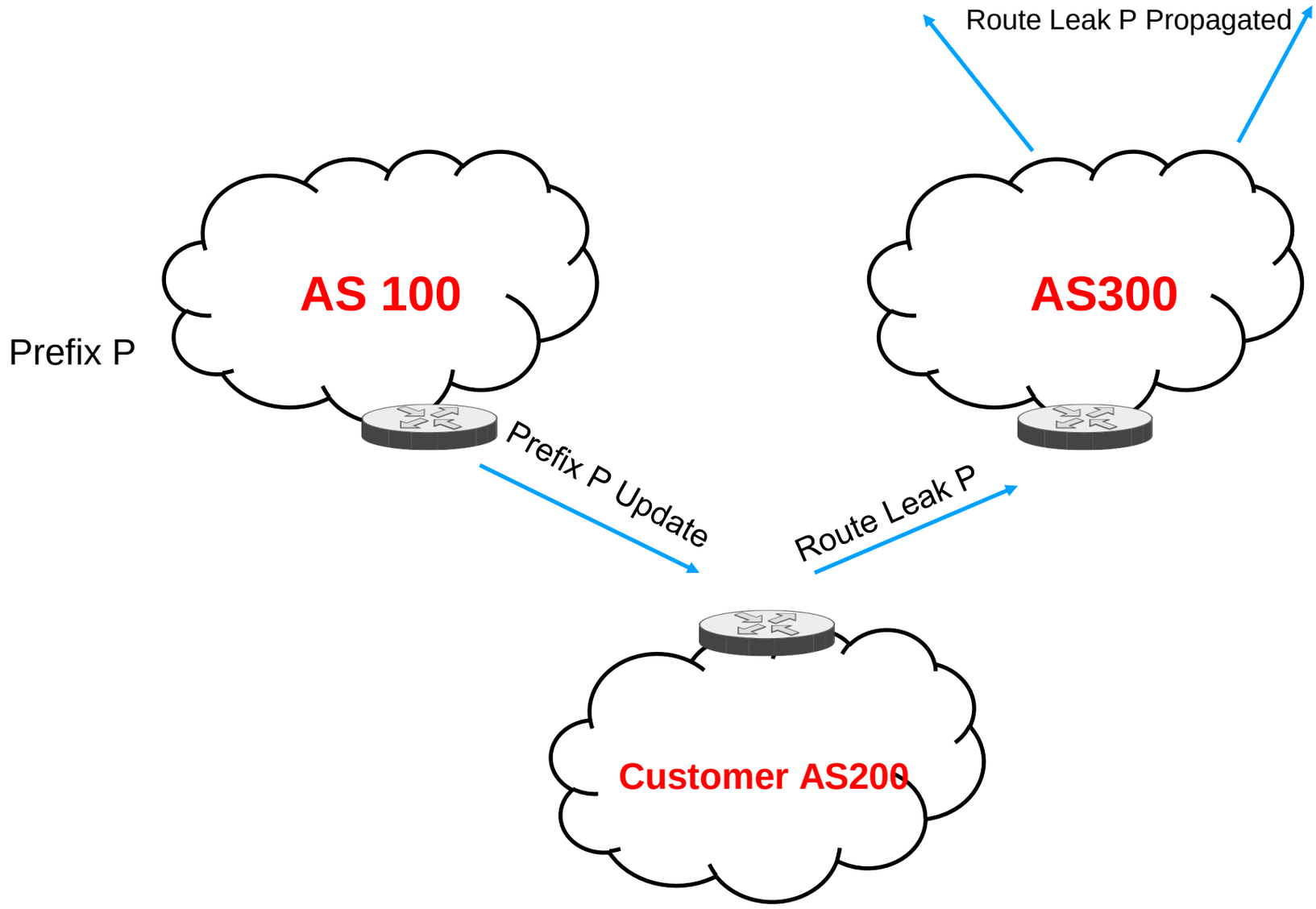
- BGP Route Leak happens mostly due to miss-configuration but some BGP Route Leak Types is considered as an Attack

BGP Information Security – Route Leaks

- It can create blackhole , extra latency , packet loss , thus will have bad effect for the customer experience

- The result of a route leak can be redirection of traffic through an unintended path that may enable eavesdropping or traffic analysis, or simply blackhole network traffic as there is no enough capacity on the network which leaked the prefixes

BGP Information Security – Route Leaks



BGP Information Security – Route Leaks

- RFC 7908 highlights the Problem Definition and Classification of BGP Route Leaks

- Formal definition of BGP Route Leak is the propagation of routing announcement(s) beyond their intended scope. That is, an announcement from an Autonomous System (AS) of a learned BGP route to another AS is in violation of the intended policies of the receiver, the sender, and/or one of the ASes along the AS path

BGP Route Leak Types

Based on RFC 7908, several Route Leak Type is defined

1. Hairpin Turn with Full Prefix
2. Lateral ISP-ISP-ISP Leak
3. Leak of Transit Provider Prefixes to Peer
4. Leak of Peer Prefixes to Transit Provider
5. Prefix Re-origination with Data Path to Legitimate Origin
6. Accidental Leak of Internal Prefixes and More-Specific Prefixes

BGP Route Leak - Hairpin Turn with Full Prefix

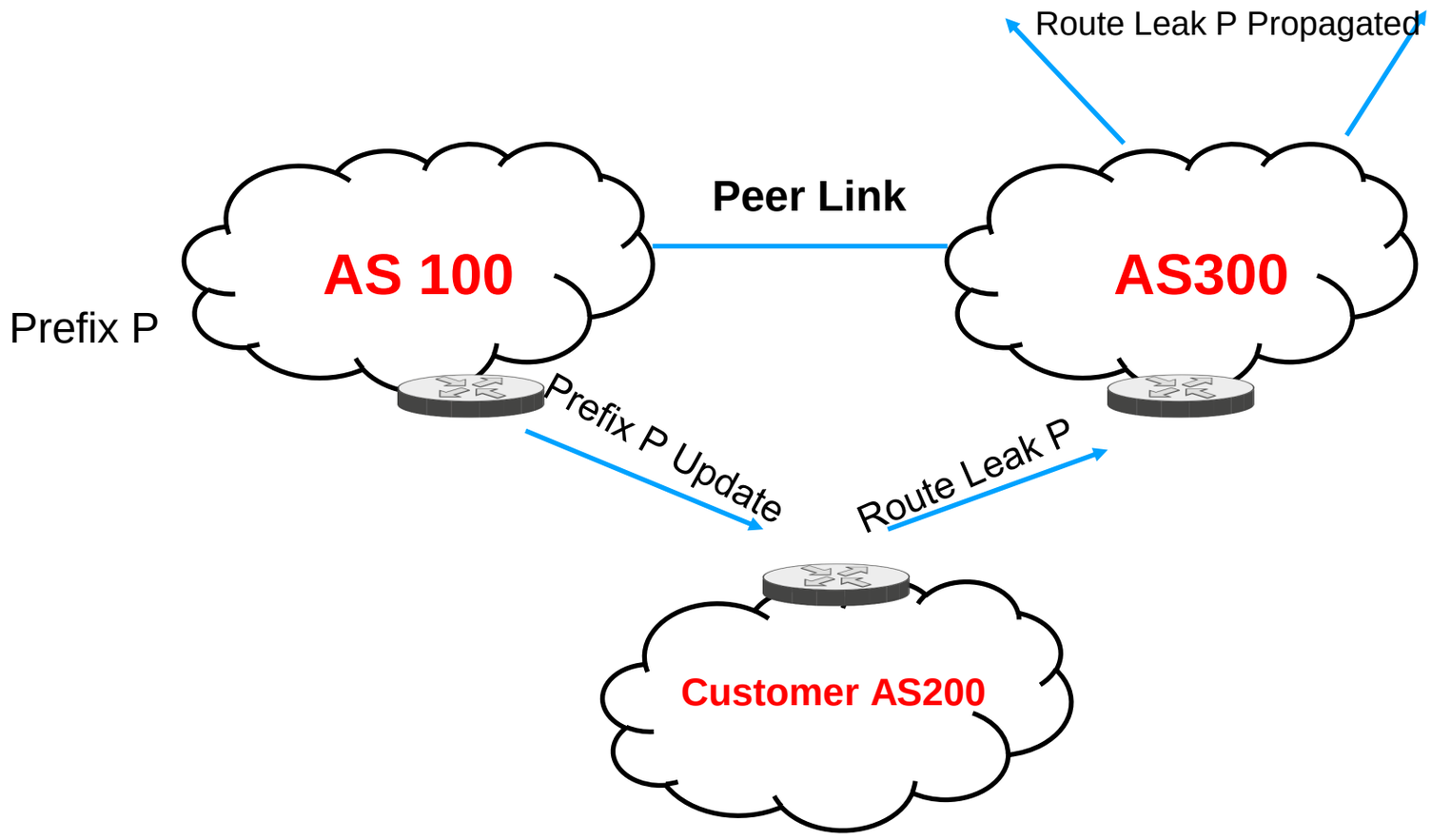
- A multihomed AS learns a route from one upstream ISP and simply propagates it to another upstream ISP

- It should be noted that leaks of this type are often accidental (not malicious)

- The leak often succeeds (the leaked update is accepted and propagated) because the second ISP prefers customer announcement over peer announcement of the same prefix

BGP Route Leak - Type 1 - Hairpin Turn with Full Prefix

AS 300 Receives Prefix P from, both Customer AS 200 and Peer AS 100 As 300 Prefers AS 200 because based on GAO-Rexford Model, Customer is preferred to Peer announcements



BGP Route Leak - Type 2 – Lateral ISP-ISP-ISP Leak

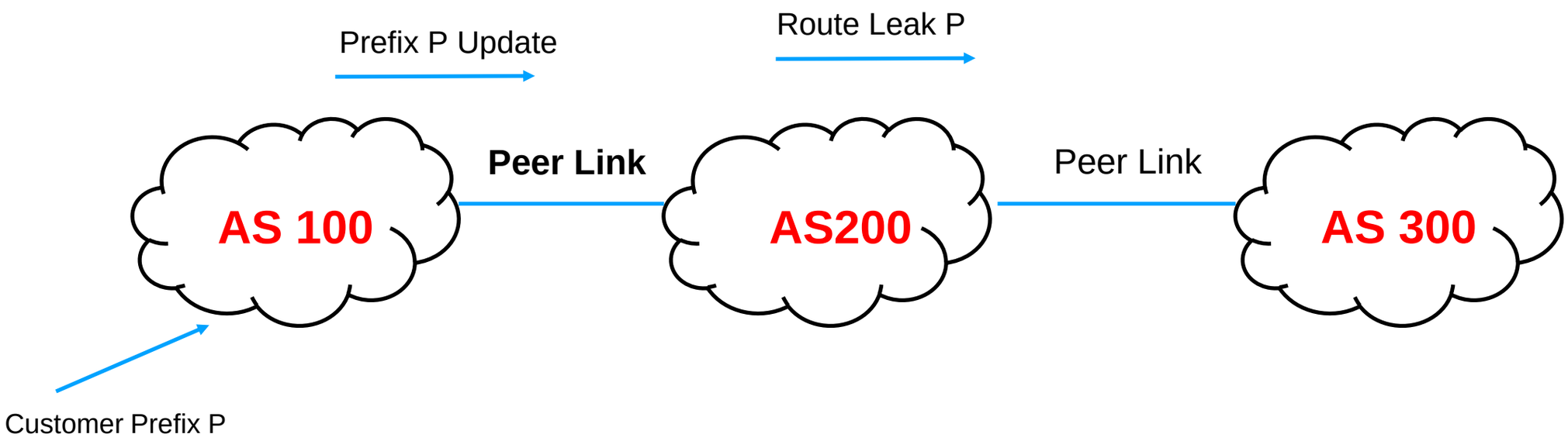
- The term "lateral" here is synonymous with "non-transit" or "peer-to-peer”

- This type of route leak typically occurs when, for example, three sequential ISP peers (ISP-A, ISP-B, and ISP-C) are involved, and ISP-B receives a route from ISP-A and in turn leaks it to ISP-C

BGP Route Leak - Type 2 – Lateral ISP-ISP-ISP Leak

- The typical routing policy between laterally (i.e., non-transit) peering ISPs is that they should only propagate to each other their respective customer prefixes

BGP Route Leak - Type 2 – Lateral ISP-ISP-ISP Leak

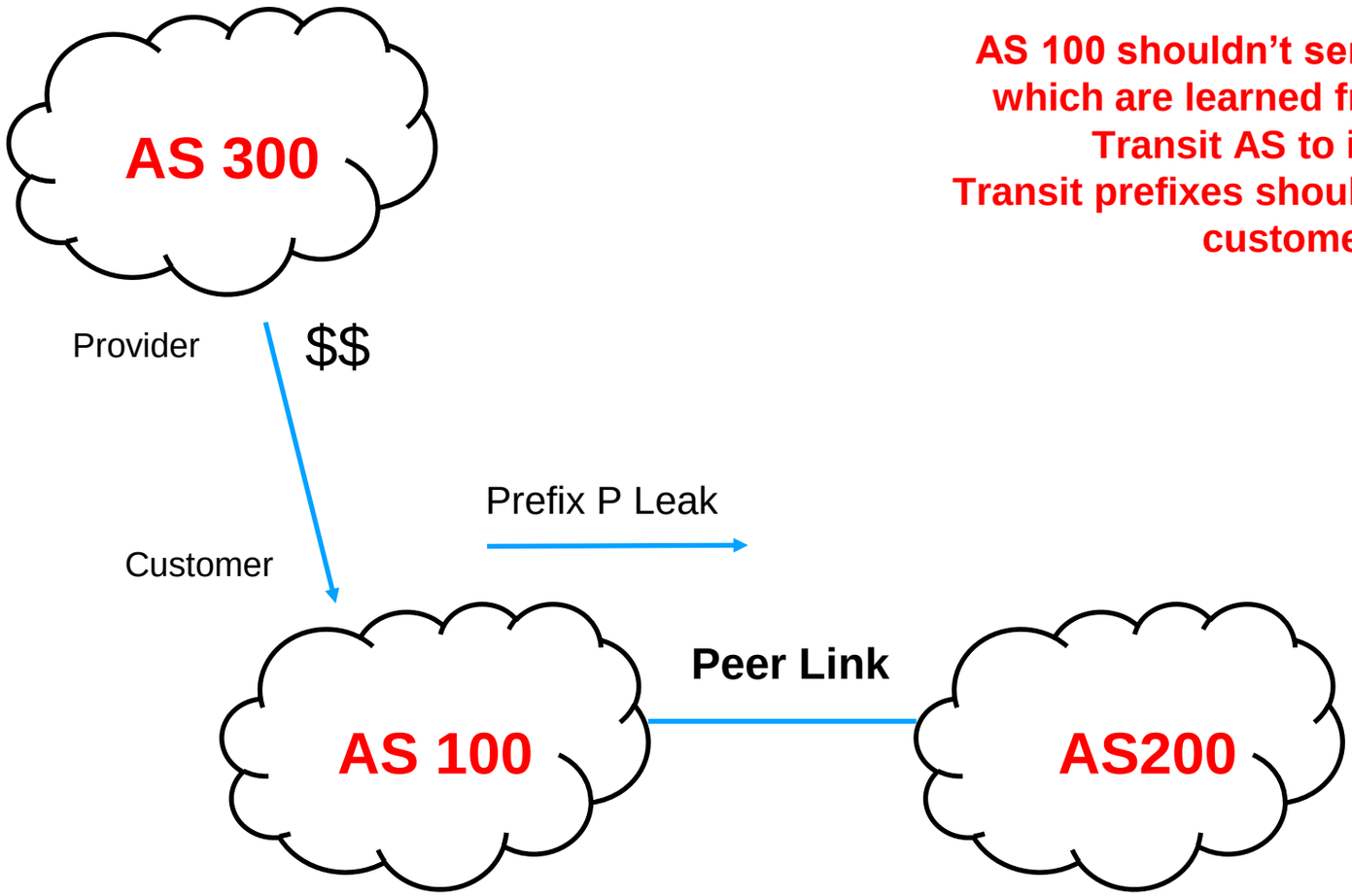


AS 200 shouldn’t send Customer Prefix of AS 100 to AS 300, It is a Route Leak

BGP Route Leak – Type 3 – Leak of Transit Provider Prefixes to Peer

- This type of route leak occurs when an offending AS leaks routes learned from its transit provider to a lateral (i.e., non-transit) peer

BGP Route Leak – Type 3 – Leak of Transit Provider Prefixes to Peer



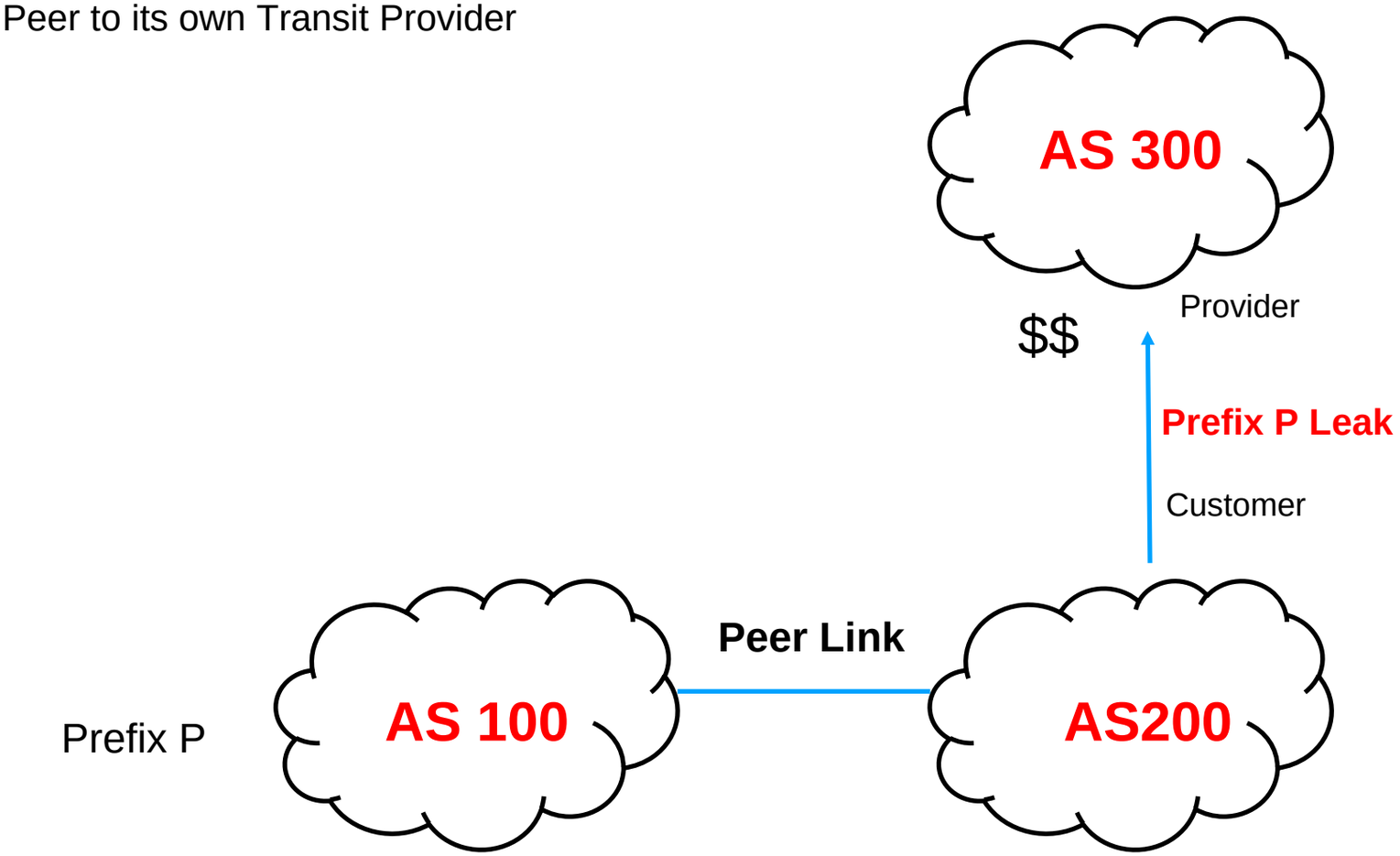
AS 100 shouldn't send the prefixes which are learned from upstream Transit AS to its Peers
Transit prefixes should be sent to the customers

BGP Route Leak - **Type 4 - Leak of Peer Prefixes to Transit Provider**

- This type of route leak occurs when an offending AS leaks routes learned from a lateral (i.e., non-transit) peer to its (the AS's) own transit provider

BGP Route Leak - **Type 4 - Leak of Peer Prefixes to Transit Provider**

AS 200 shouldn’t send the Prefixes which are learned from its Settlement Free Peer to its own Transit Provider



AS 300 is Transit Provider of AS200
AS 100 and AS 200 are Settlement Free Peers

BGP Route Leak - **Type 4 - Leak of Peer Prefixes to Transit Provider**

- Type 4 BGP Route Leak incidents commonly seen on the Internet

- Some Examples are :

- Axcelx-Hibernia route leak of Amazon Web Services (AWS) prefixes causing disruption of AWS and a variety of services that run on AWS

- Hathway-Airtel route leak of 336 Google prefixes causing widespread interruption of Google services in Europe and Asia

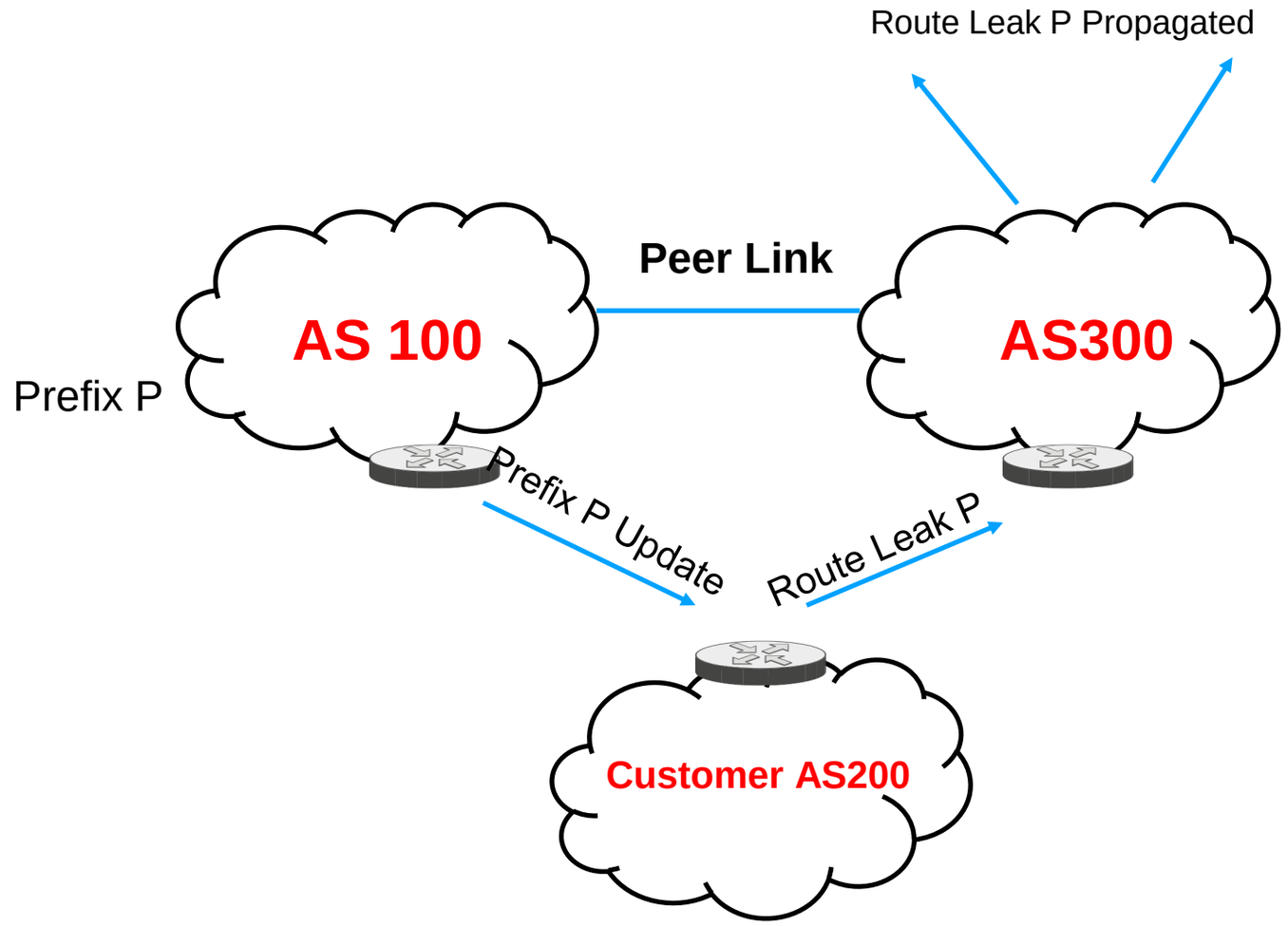
- Moratel-PCCW route leak of Google prefixes causing Google's services to go offline

BGP Route Leak - **Type 5: Prefix Re-origination with Data Path to Legitimate Origin**

- A multihomed AS learns a route from one upstream ISP and announces the prefix to another upstream ISP as if it is being originated by it (i.e., strips the received AS path and re-originates the prefix). This can be called re-origination or mis-origination

BGP Route Leak - **Type 5: Prefix Re-origination with Data Path to Legitimate Origin**

Similar to Type 1 Route Leak but in Type 5, Customer AS 200 strips the AS 100 which is the Originator AS from the AS Path



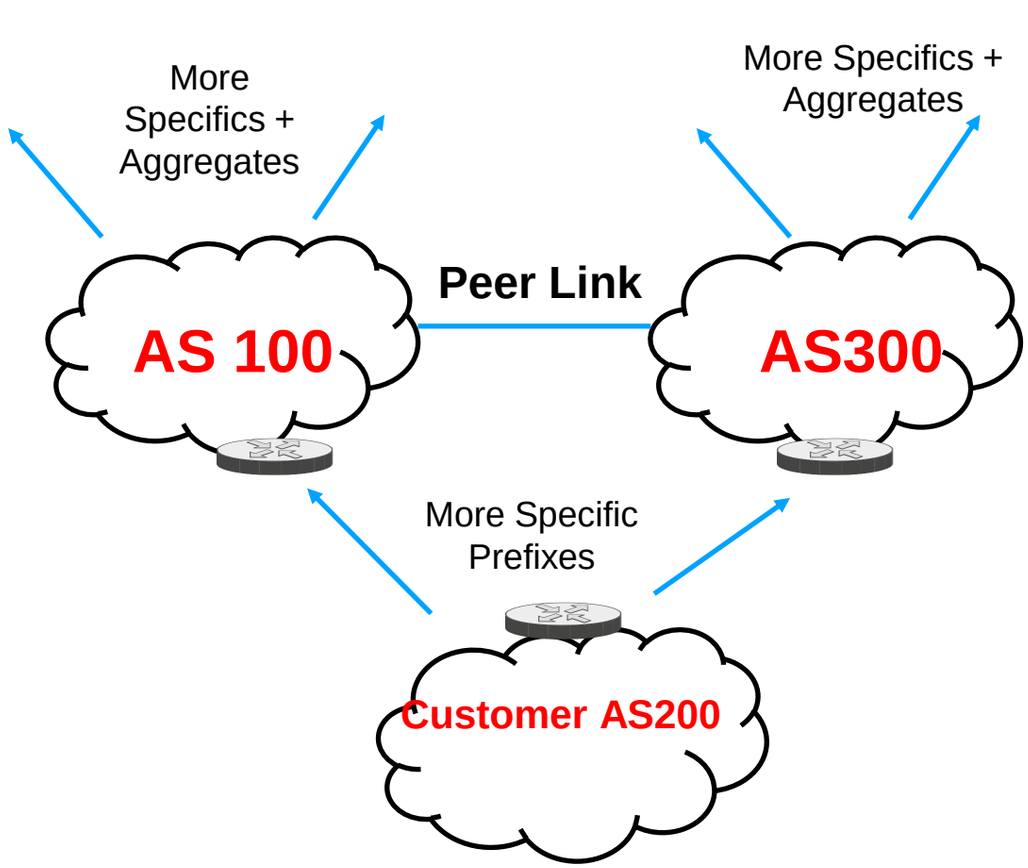
BGP Route Leak - **Type 6: Accidental Leak of Internal Prefixes and More-Specific Prefixes**

- An offending AS simply leaks its internal prefixes to one or more of its transit-provider ASes and/or ISP peers. The leaked internal prefixes are often more-specific prefixes subsumed by an already announced, less-specific prefix

- Partly because of this route leak , 512k incident happened in August 2014

BGP Route Leak - **Type 6: Accidental Leak of Internal Prefixes and More-Specific Prefixes**

Transit ISPs accept More Specific announcements from downstream AS and advertises the more specifics to DFZ



BGP Route Leak - **Type 6: Accidental Leak of Internal Prefixes and More-Specific Prefixes**

- Internal Prefix Route leak generally short lived route leaks

- It disappears after short amount of time

- In August 2014, 512k incident happened this will be explained next!

512K Incident, Type 6 BGP Route Leak Contribution to it!

- The 12th August 2014 was widely reported as a day when the Internet collapsed

- What was happening was that the Internet’s growth had just exceeded the default configuration limits of certain models of network switching equipment

- 0800 UTC on that day, when the Internet was flooded with 22,000 new prefixes, which were withdrawn very rapidly thereafter

512K Incident, Type 6 BGP Route Leak Contribution to it!

- All these routes shared a common origin, AS 701, and were all more specifics of already announced aggregate routes

- The announcements were short-lived, and were withdrawn soon after their announcement

512K Incident, Type 6 BGP Route Leak Contribution to it!

- This was a Type 6 Route Leak by AS 701, announced 22k more specific prefixes to the Internet. This amount of extra prefixes suddenly exceeds the 512000 prefixes in the Default Free Zone

- This triggered some BGP session restarts due to tripping some maximum prefix threshold values that was installed by network operators

Preventing BGP Route Leaks

- Because BGP is founded on trust and thus insecure, it can be extremely difficult to quickly resolve a route leak affecting your network, as you'll need to convince other networks to choose the legitimate route over the incorrect one

- While you won't have complete control in a route leak situation, you do have some options to fight with ongoing route leaks

Preventing BGP Route Leaks

- IRR Route Objects, RPKI ROAs and BGPSEC all contributed to prevent Route Leak

- Depends on the Type of Route Leaks and percentage of the implementation and whether upstream ISP deployed these methods, IRR, RPKI and even BGPSEC may not be effective

- Most people deploy Static Prefix Filters but when the network gets larger, maintaining prefix list becomes easily cumbersome

BGP Hijacking

- Hijacking occurs when an attacker claims to own a prefix or sub-prefix that belongs to another AS causing redirection of routes from the AS to the attacker

- Attackers hijack prefixes to produce different malicious activities. For example, the hijacker can blackhole all traffic to the victim causing a DoS for that network

- In another scenario, the attacker becomes a man-in-the-middle, intercepting the traffic without affecting victim reachability

BGP Hijacking

- Phishing attacks can also be done by hijacking a prefix through redirecting traffic to an incorrect destination

- Additionally, the attacker can use stolen IP addresses to send spam

- There are examples of each of these with BGP Hijacks

Some BGP Hijacking Incidents

Hijacking for Monetary Gain

- Between October 2013 – May 2014 , Canadian Hijacker ISP originated more specifics (Sub-Prefix Hijack) to stole Bitcoin
- Hijacked AWS, LeaseWeb and couple other OTT Subnets
- It was towards IX Peers at TORIX (Toronto IX)
- $83,000 was stolen

Some BGP Hijacking Incidents

Hijacking for Censorship

- March 28 – 30 Election in Turkey
- Turk Telekom brought up DNS servers and redirected DNS traffic
- Man in the Middle affected Google, Level3, OpenDNS etc.

Some BGP Hijacking Incidents

Hijacking for Spamming

- Using un-announced address space to send spam
- Find unused space & announce for few hours and send spam
- Rinse and Repeat
- This technique is known as IP Squatting

Types of BGP Hijacks

BGP Hijacking are classified into 4 Subtypes:

- Hijacking a prefix
- Hijacking a prefix and its AS
- Hijacking a sub-prefix
- Hijacking a sub-prefix and its AS

BGP Prefix Hijacking

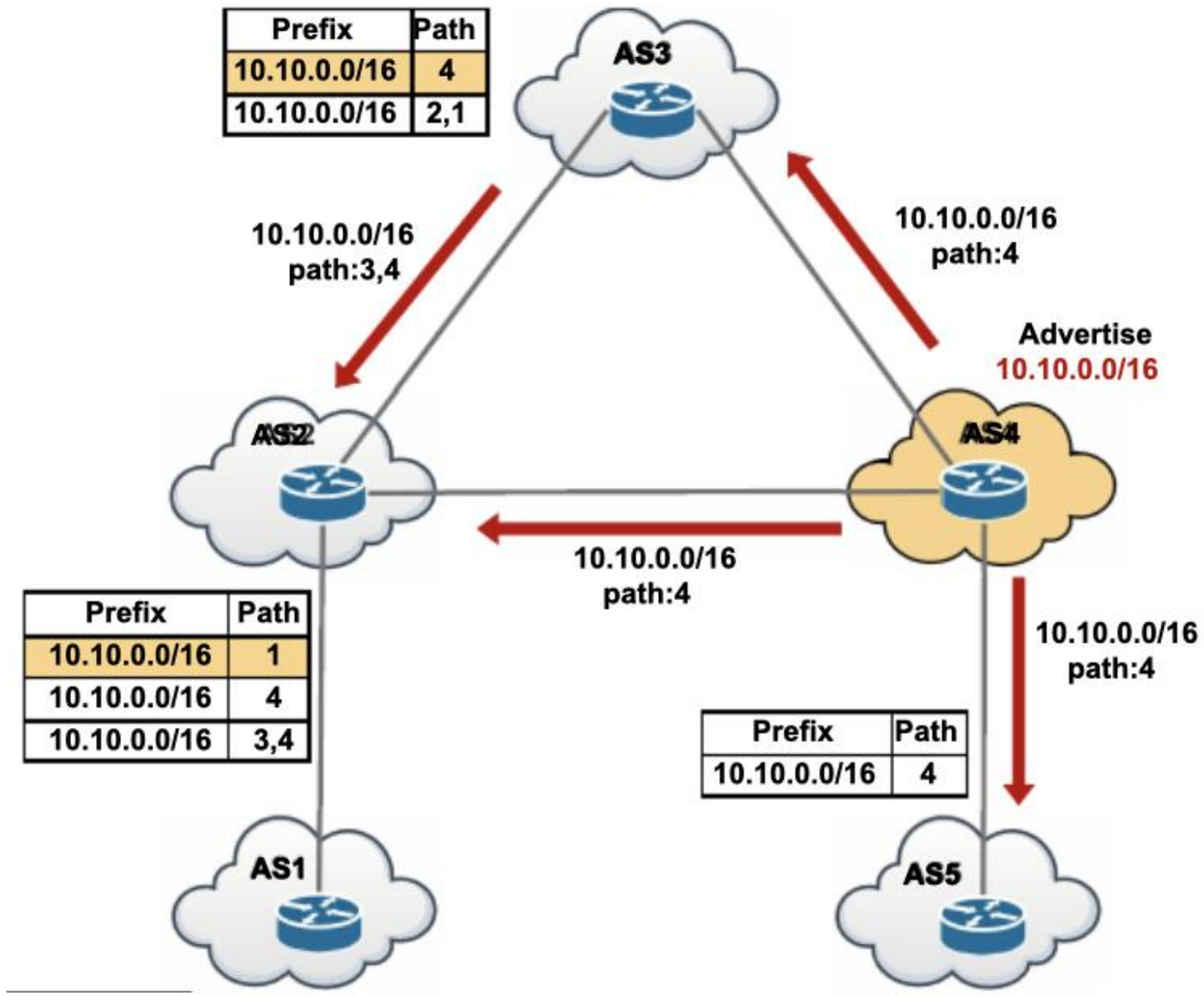
- In this type of hijack, an attacker configures its BGP router to announce a prefix belonging to another AS

- BGP allows any BGP speaker to announce any route regardless of whether the route actually exists or not

- Attacker’s neighbors will adopt it as a new route

BGP Prefix Hijacking

AS4 sends hijacked prefix 10.10.0.0/16
Legitimate Origin is AS1
AS3 and AS5 believes that AS4 owns the prefix
AS2 doesn't change the best path since both AS1 and AS4 has same AS Path length



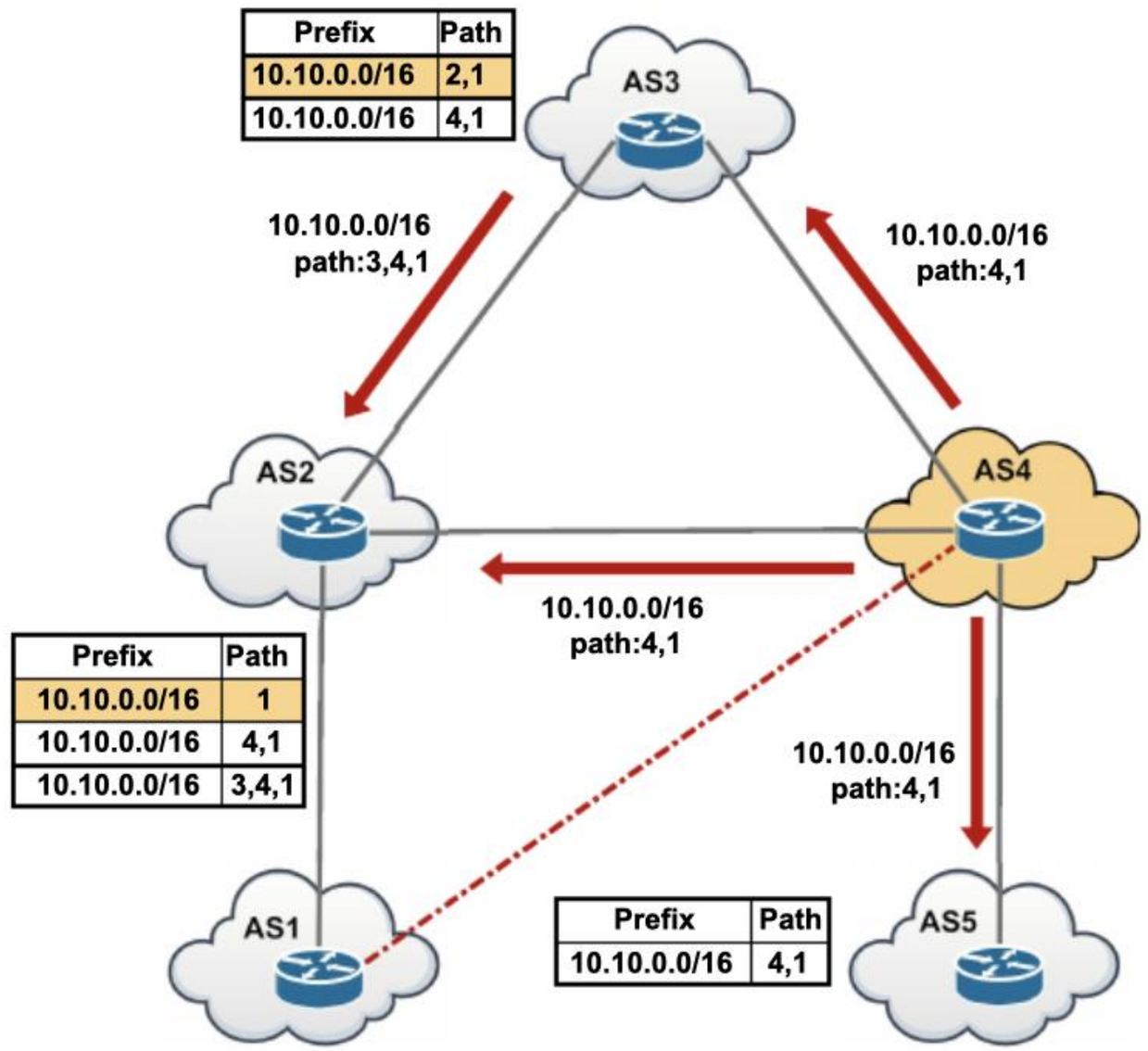
BGP Prefix Hijacking

- BGP Prefix Hijacking is commonly known as BGP Exact Prefix Hijacking (Exactly same prefix length with the victim AS is announced)

- Since shortest route paths are typically preferred, only a part of the Internet that is closer to the hijacker (in number of AS-hops) switches to route paths towards the hijacker

Prefix and its AS Hijack

Origin validation can stop Prefix Hijack but not this one, because
With Prefix and its AS Hijack, attacker manipulates the path as it is connected to the Origin
AS4 sends an announcements to the AS 2 , 3 and 5 as [4, 1]
AS2 and AS3 doesn't use AS4, but AS5 still does



BGP Sub-prefix Hijacking

- With BGP Sub-prefix Hijacking, an attacker announces a sub-prefix that belongs to a victim AS

- BGP selects the most specific address or longest address match. For example, a BGP router will select a specific address such as 10.10.0.0/24 over a more general address such as 10.10.0.0/16

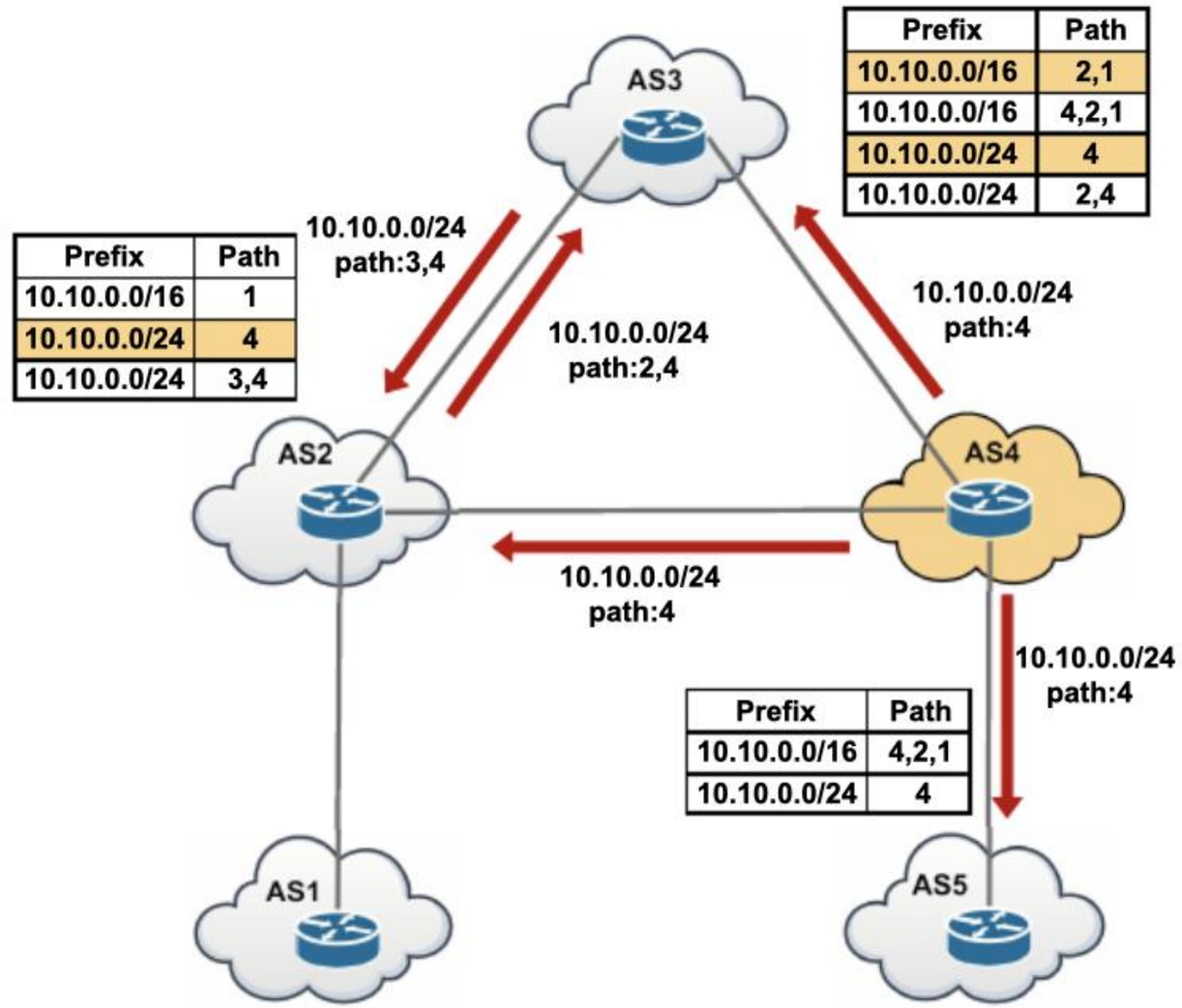
BGP Sub-prefix Hijacking

- This is the most widely propagated type of hijacking since all ASes between the attacker and the victim are affected

- This type of hijacking can be globally propagated when there is no other advertisement or filtering for this route

BGP Sub-prefix Hijacking

AS4 announces a prefix 10.10.0.0/24 which is a part of the prefix 10.10.0.0/16 owned by AS1
All ASes in this example chooses AS 4 as an Origin AS



Famous Sub-Prefix Hijack

- 2008 Pakistan Telecom and Youtube
- Pakistan Telecom wanted to block access to Youtube by announcing /24 of Youtube's /22 blocks unintentionally
- Good percentage of Youtube traffic was redirected to Pakistan Telecom
- In 2010 China Telecom leaked 50k prefixes, 2017 Russian SP leaked 80 important prefixes

BGP Sub-Prefix and its AS Hijack

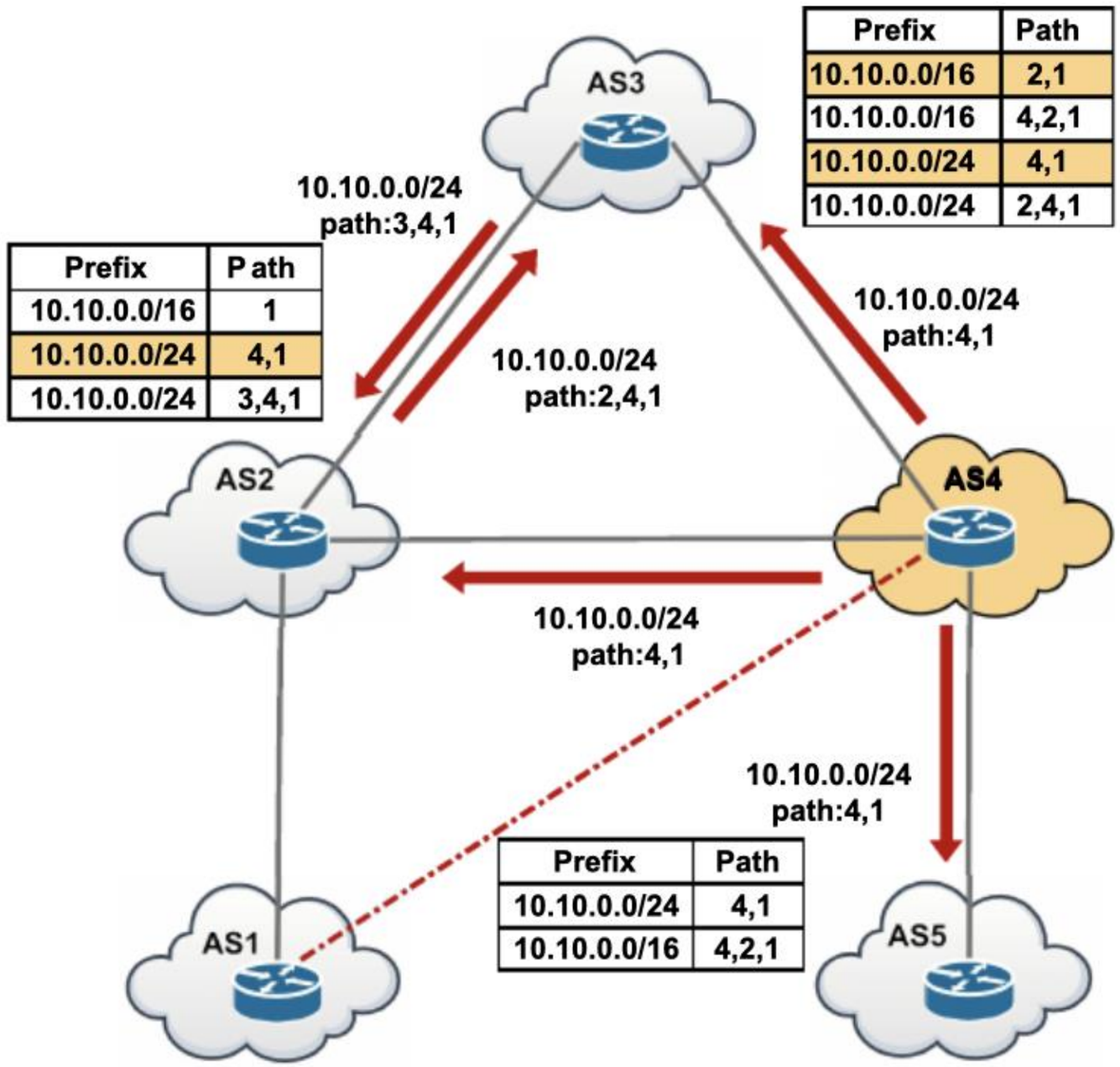
- Attacker Announces a fake path to a subnet of a target prefix

- Using a fake path with sub-prefix hijack represents a critical challenge for detection as the attacker does not claim to own a full prefix length

- This attack type is considered as most difficult to detect

BGP Sub-Prefix and its AS Hijack

AS 2 , 3, and 5 believe that AS4 is the Origin AS for the prefix 10.10.0.0/24



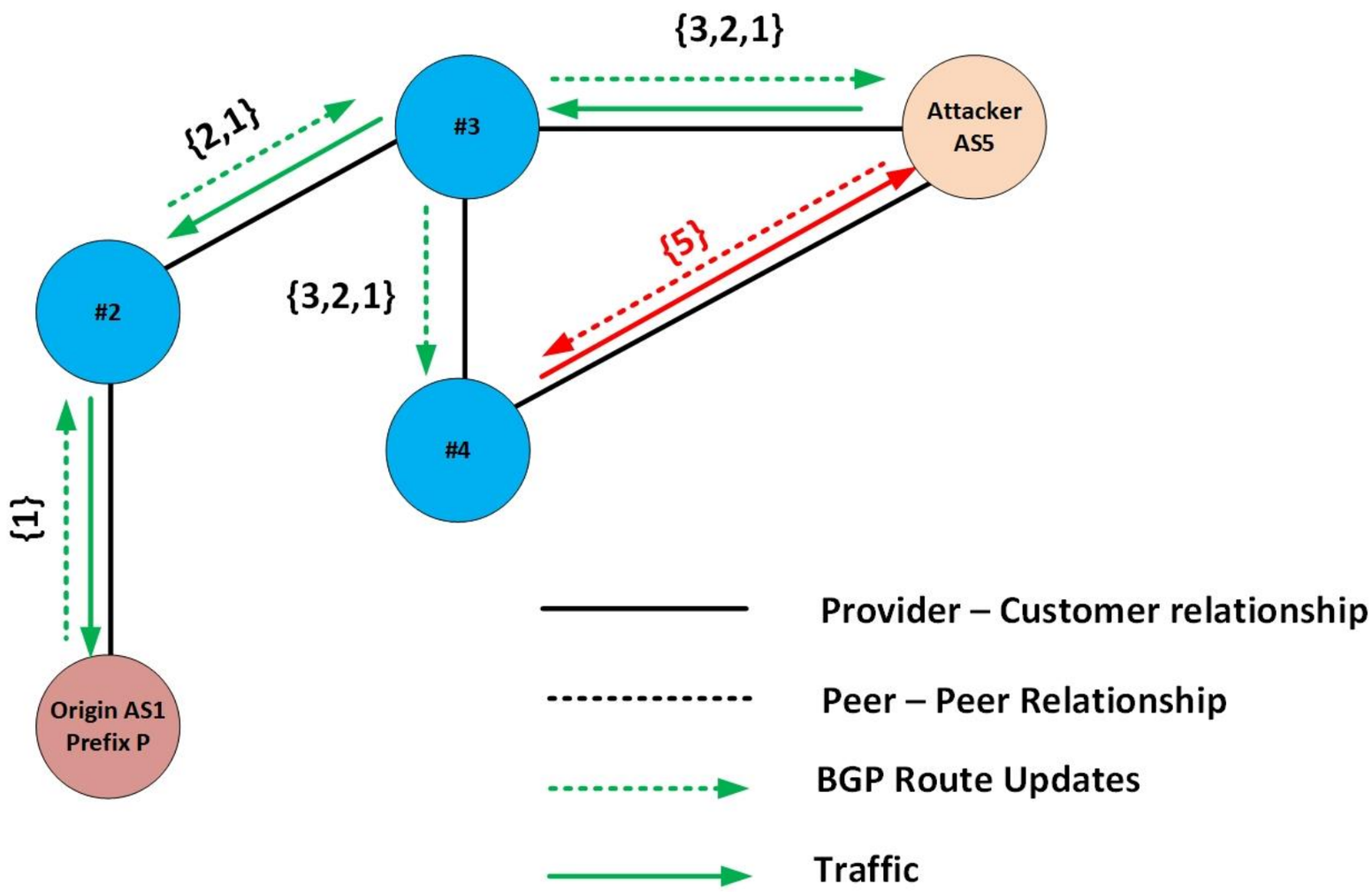
What is Traffic Interception?

- After hijacking the prefix, malicious AS can also forward the hijacked traffic back to real destination and this type of attack is called Traffic Interception

- As the traffic reach to the destination, connectivity is not disrupted and interception is transparent to the victim Ases

- This type of attack can lead to a Man-in-the-middle attack which allows a malicious AS to eavesdrop or modify the traffic

What is Traffic Interception?



How to detect and mitigate Route Leaks, Hijacks and Interceptions?

- IP hijacking can be prevented to some extent by means of filters and hijack detection systems

- Announcements by customer ASes and peer ASes in which prefixes are out of the allocated range can be filtered

- If route filters at the links between providers and their customers are properly configured in order to prevent customer ASes from advertising the routes for the prefixes which do not belong to them

Challenges with Prefix Filters to mitigate Hijacking

<u>According to following reasons, Prefix Filters is insufficient and difficult:</u>

- To install ingress filters, it is not always possible for providers to know which prefixes are assigned to which customers. If customers have multiple providers, they may have different address prefixes from different providers

Challenges with BGP Prefix Filters to mitigate Hijacking

- And enforcing ingress filters in peering edges is also difficult as it is not knowable that peer ASes allocate which addresses to their customers

- Even if route filters are installed in ingress points, when there is one provider that does not practice route filtering, IP hijacking becomes possible

- Because of these reasons, building Filters by asking to the customer is not practical, instead IRR and RPKI repositories are used to generate filters, in the next section, Alternatives to the Filter will be explained

Alternate methods to build BGP Prefix Filters

- In addition to Filters, IRR , RPKI and BGPSEC will help to mitigate/stop Hijacks , Route Leaks and Interception

- IRR and RPKI have same goals which is Origin AS Validation, BGPSEC provides both Origin and Path Validation

- IRR and RPKI databases can be used to automatically generate prefix filters with the help of some tools

Alternate methods to build BGP Prefix Filters

- There are some tools designed to work with IRR policies to automatically generate ingress and egress filters by parsing aut-num object, commonly used tools are IRRToolset, BGPQ3 and IRRPT

- Network operator register their announcements in the form of ROA objects and they are used by other operators either to generate filters or to validate announcements using more advanced techniques such as RPKI-to-Router protocol

Important to know about IRR

- IRR, RPKI , BGPSEC and Filters are not mutually exclusive. In a given network all four approaches can be used. In fact having Filter and RPKI at the same time are seen as common deployment as of 2020

Three Source Database for Routing Security

- There are three databases and repositories which should be used by the network operators to document routing policy and maintain contact information.

- These are IRR, RPKI and PeeringDB

- You can publish your contact information, create route objects and Routing Registry through these systems

- You will learn all three in the next sections

Three Source Database for Routing Security

- Most of the Hijack cases can be solved with Origin Validation

- IRR Route Objects and RPKI ROAs are used to provide Origin Validation

- Cases which cannot be solved with Origin Validation might require Path Validation and there are couple solutions that provides Path Validation for BGP routes

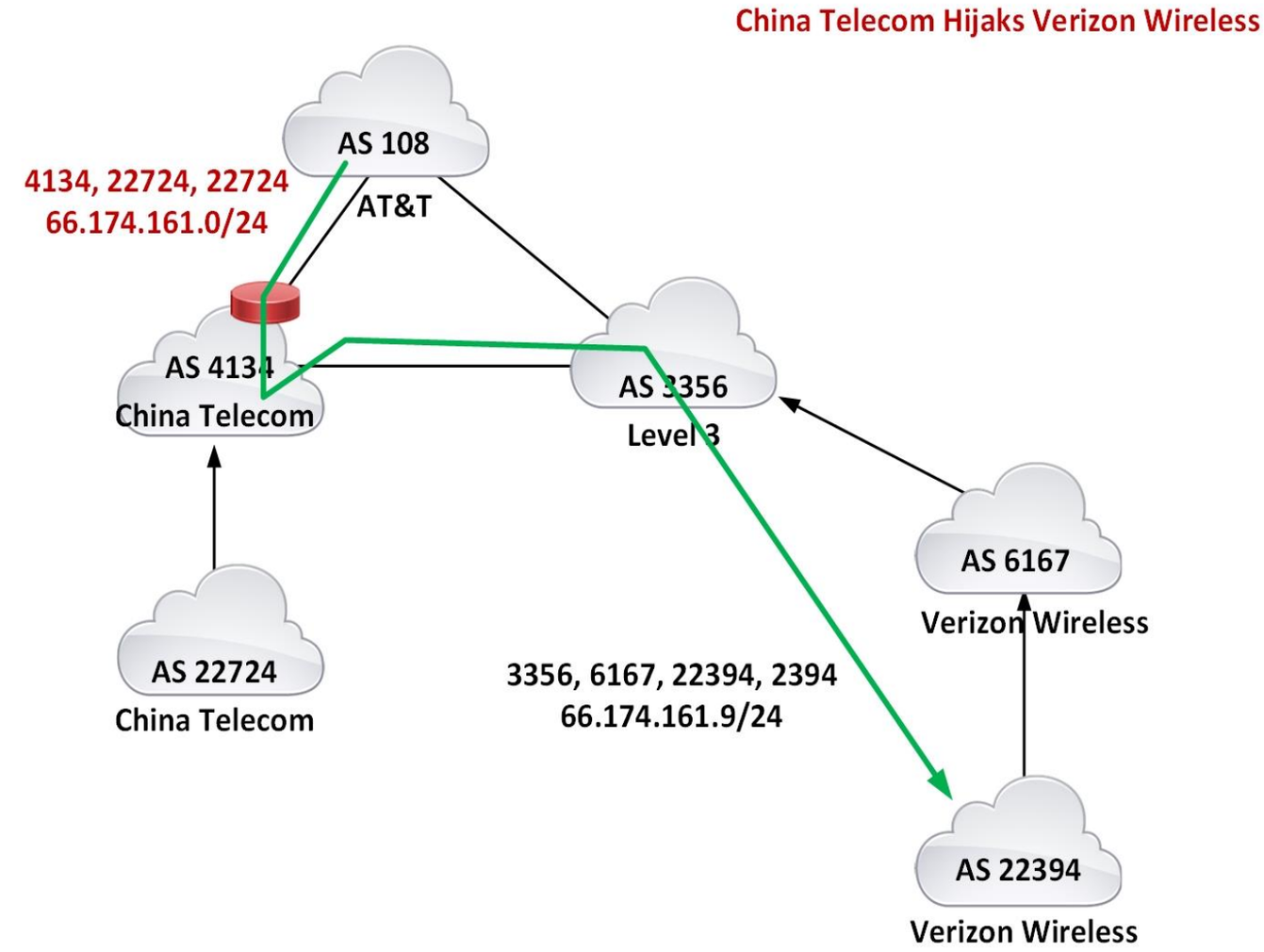
What is Origin Validation?

- Origin Validation is a reliable way of telling whether a BGP Route Announcement is authorized by the legitimate holder of the address space

- RPKI ROAs and IRR Route Objects are used for Origin Validation

Origin Validation Example

In this example China Telecom hijacked Verizon Wireless 66.174.161.0/24
If Origin Validation technique would be used, ATT wouldn't prefer China Telecom (Because of shorter AS path) because China Telecom is not authorized to originate that prefix



Origin Validation Example

- Current Origin Validation mechanism is RPKI (Resource Public Key Infrastructure)

- Hijack in previous example could be prevented if Victim (Verizon Wireless) would generate RPKI ROA and if AT&T would verify the ROA and prefer cryptographically signed path

- In the next section, RPKI will be explained and how it provides Origin Validation will be shown

RPKI – Resource Public Key Infrastructure

- Resource Public Key Infrastructure (RPKI) is a specialized PKI that aims to improve the security of the Internet routing system, specifically the BGP

- It does this through the issuing of X.509-based resource certificates to holders of IP addresses and AS numbers in order to prove assignment of these resources

- These certificates are issued to Local Internet Registries (LIRs) by one of the five Regional Internet Registries (RIRs)

RPKI – Resource Public Key Infrastructure

- APNIC, ARIN, LACNIC, AFRINIC and RIPE are the five RIR who are responsible for allocation and assignment of these resources in their service regions

- As of 2019, 15% - 20% of the prefixes in DFZ have ROAs

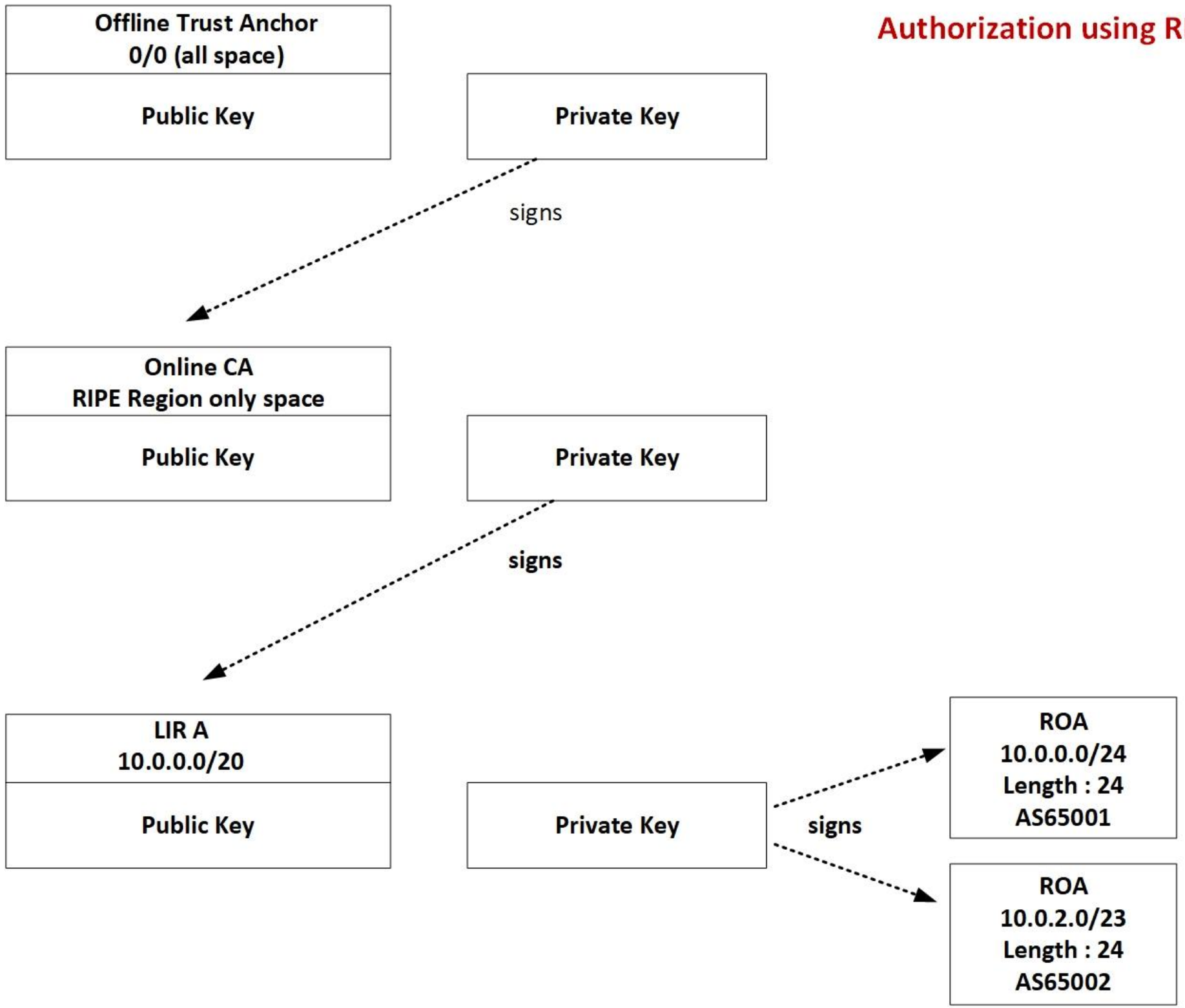
- Compare to Route Objects in IRR, this is very small percentage

RPKI – Resource Public Key Infrastructure

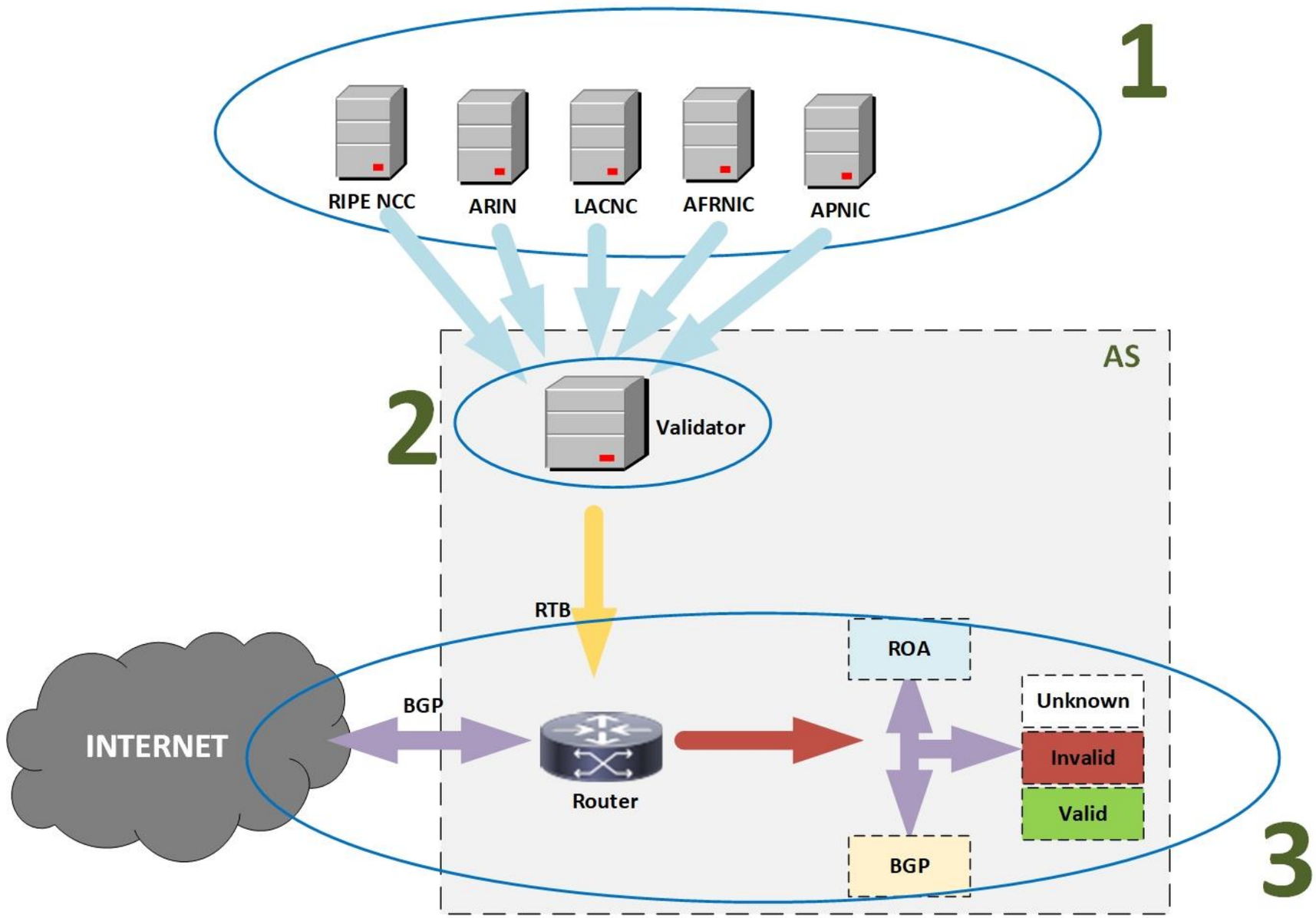
- Resource certificates allow LIRs (ISPs) to generate Route Origin Authorizations (ROAs) which attest to which networks (specifically AS numbers) are authorized to originate which ranges of IP addresses

- This then allows other networks to determine whether route announcements are valid and should therefore be accepted, thus reducing the likelihood of fake routes being propagated across the Internet

Authorization using RPKI



RPKI Overview



RPKI – Resource Public Key Infrastructure

- Each RIR acts as a CA and trust anchor for the resources assigned within their service regions

Configuring RPKI on Router Example

```
route-map rpki-loc-pref permit 10
 match rpki invalid
 set local-preference 90
!
route-map rpki-loc-pref permit 20
 match rpki not-found
 set local-preference 100
!
route-map rpki-loc-pref permit 30
 match rpki valid
 set local-preference 110
```

```
router bgp 64500
 bgp log-neighbor-changes
 bgp rpki server tcp 10.1.1.6 port 8282 refresh 5
 network 192.0.2.0
 neighbor 10.1.1.2 remote-as 64510
 neighbor 10.1.1.2 route-map rpki-loc-pref in
```

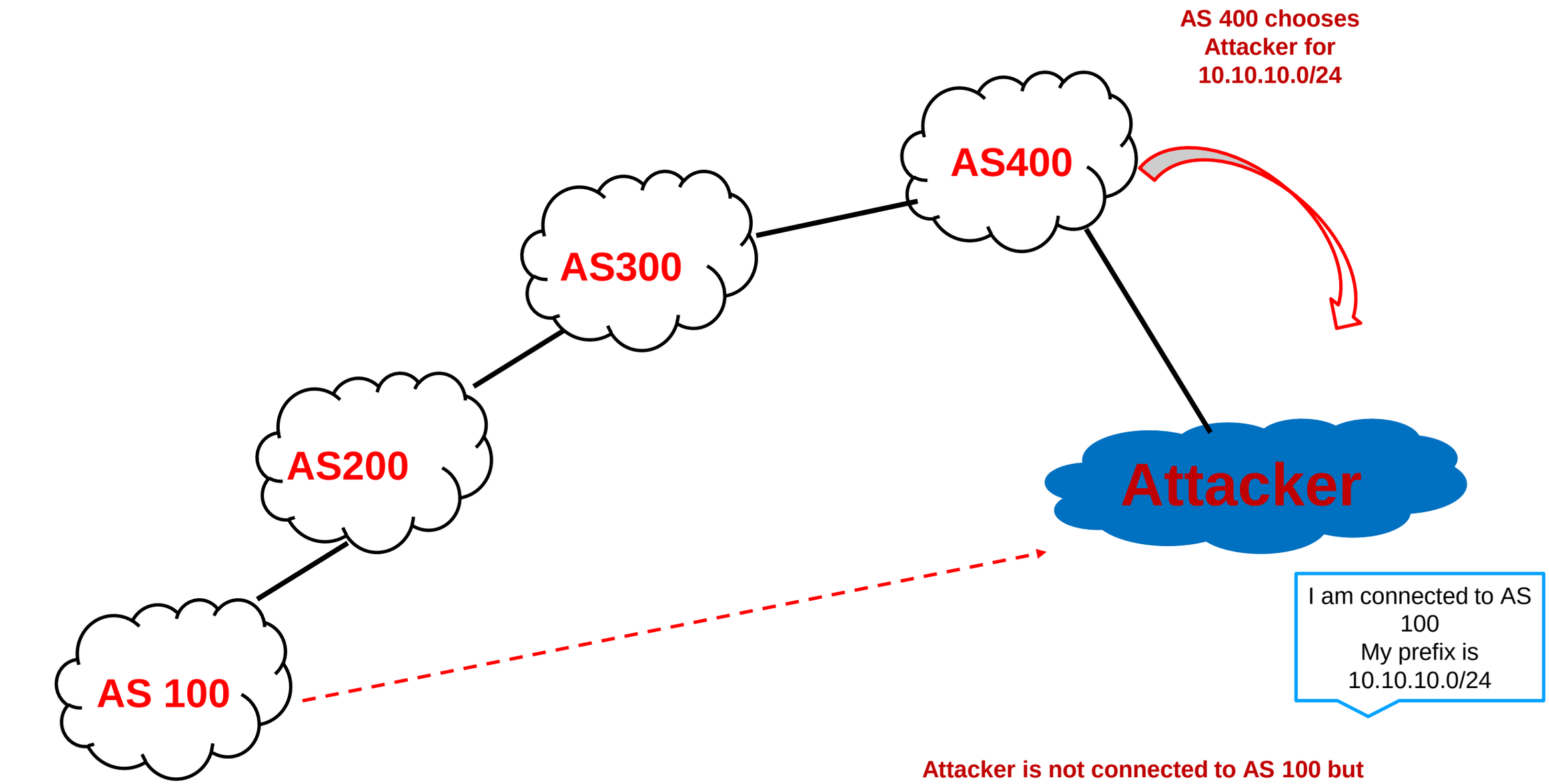
BGP Hijacks is solved with RPKI

- Exact Prefix Hijacks and Sub-Prefix Hijacks can be solved with RPKI but still attacker can launch a path-shortening attack as we will see in the next example

- AS-Path Shortening is an advertising the prefix with the legitimate Origin AS but shorten it, so other ASes can see insecure path is the best due to shorter AS-Path

- Generally AS-Path Shortening is considered as Malicious attack, not as ‘Misconfiguration’

Path-Shortening Attack cannot be solved with RPKI



RPKI without BGPSEC?

- RPKI provides Origin Validation, not Path Validation

- If BGPSEC is not available, lack of path validation can be resolved with densely peering (Google and Akamai has more than 130 peering facilities in common)

- This prevents Path shortening attacks which cannot be solved with RPKI, thus densely peering is considered an advantage on top of RPKI

Do everyone really need an RPKI?

- Because of centralization of the Web, if a few largest companies deploy RPKI, millions of people benefit from RPKI

- These are large content providers (OTTs) such as Google, Akamai, Cloudflare, Amazon, Facebook, Microsoft etc.

RPKI Summary

- It is available since 2011
- RPKI is a security framework for verifying the association between resource holders and their Internet resources
- Attached digital certificates to network resources upon request that lists all resources held by the member
- -- AS Numbers
- --IP Addresses
- Operators associate these two resources through ROA
- Provides Origin Validation, not Path Validation
- Offline cryptography , no Router resource issue

Last but not least RPKI doesn’t prevent Route Leaks

IRR – Internet Routing Registry

- IRR is a public database of Internet route object. IRRs are used for determining and sharing route and other related information used for configuring routers

- Using IRR, networks exchange their routing policies with each other

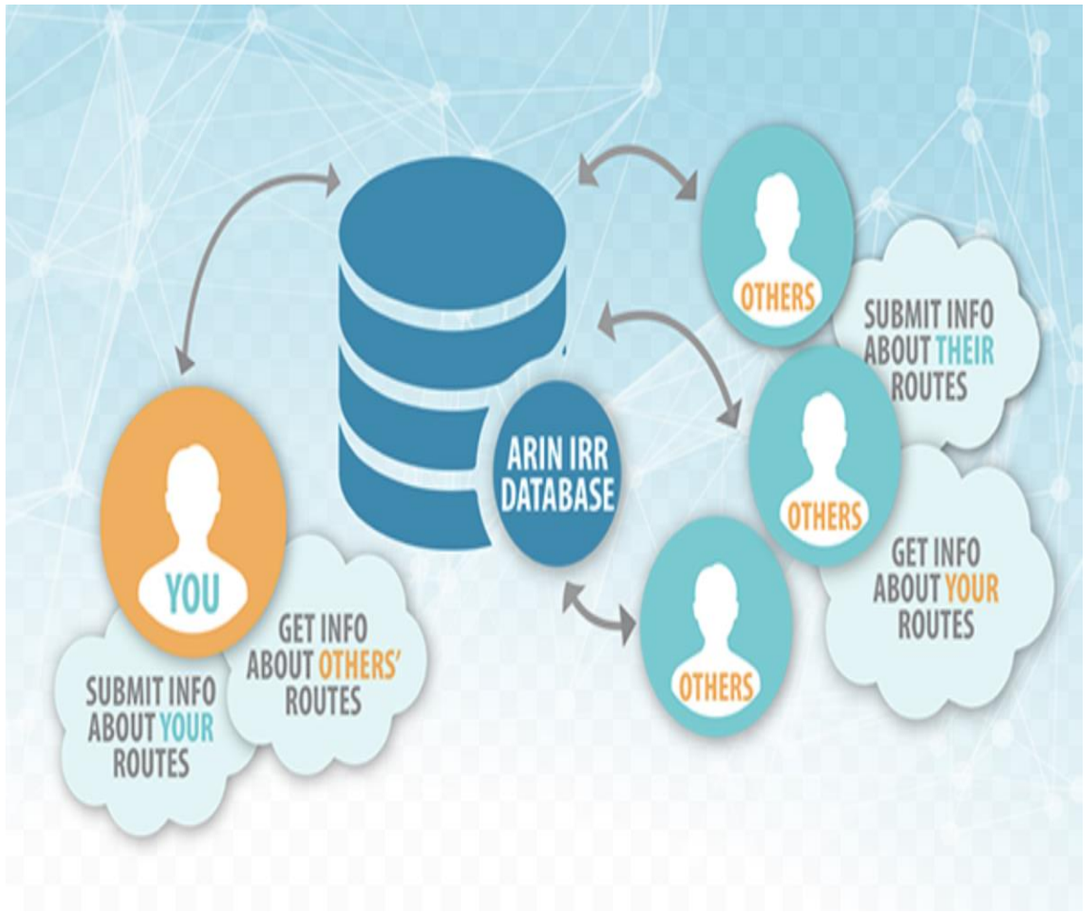
- If the Regional Internet Registry (RIR) in your region operates an IRR it should be used to document the network routing policy and related route announcements

Example of BGP Routing Policy

- Who are my BGP peers ? Customers, peers, upstream
- What routes are originated by each neighbor?
- What routes are Imported from each neighbor?
- What routes are exported to each neighbor?
- What routes are preferred when multiple routes exist?
- What to do if no route exists?

IRR – Internet Routing Registry

- Route Objects are created in the IRR database



- Route Objects consist of IP Prefix , AS Number which announces the prefix and Origin of the Registry (Example RIPE NCC)

IRR – Internet Routing Registry

- When you want to create a Route Object, you should be the Prefix Holder at least, some RIRs require ASN owner to approve Route Object creation (Ex: RIPE NCC)

- Difference between Route Objects in IRR and the RPKI ROA is, ROA digitally signs the prefixes and the Maximum length of the prefix can be specified with ROAs, not with the Route Objects

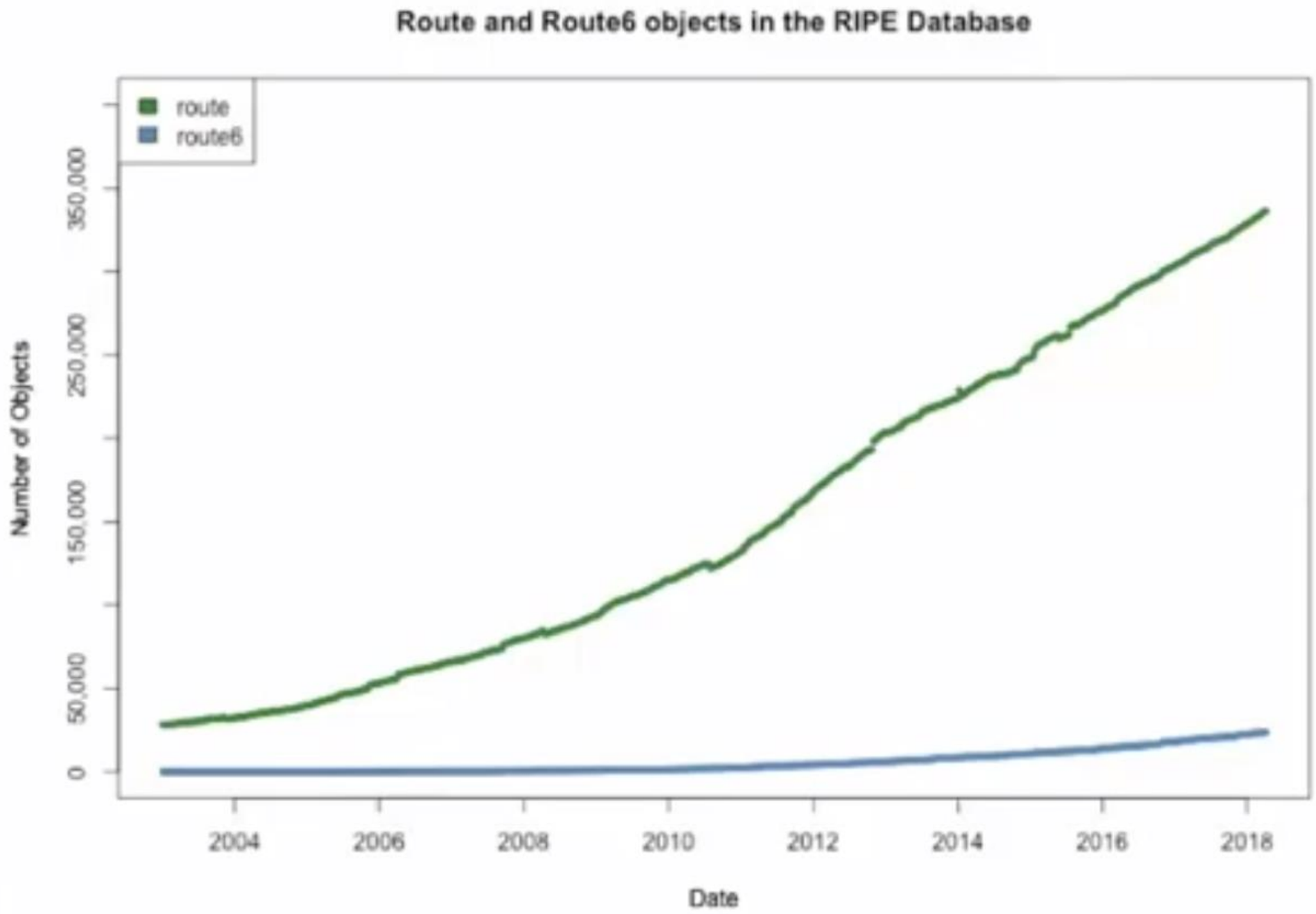
IRR – Internet Routing Registry

- There are many IRR (34 as of 2019), most widely used ones are RIPE Database and RADB

- **Based on Internet Society below IRR should be used for given regions:**

Region	Preferred IRR	Alternative IRR
America	ARIN	RADb/NTTCOM
Africa	AFRINIC	RADb/NTTCOM
Asia Pacific	APNIC	RADb/NTTCOM
Europe	RIPE NCC	RADb/NTTCOM
Latin America and Caribbean	RADb	NTTCOM

Number of Route and Route 6 Objects in RIPE Database



IRR – Internet Routing Registry

- IRR supposed to verify the holdership, which mean they need to verify whether the ASNs which announce the prefixes are valid owner of the prefixes

IRR – Internet Routing Registry

- Unfortunately lots of IRR don’t verify the holdership

- This create legitimacy issue as people can create fake announcements through IRR databases

How IRR Works?

- Routing policy information is documented in RPSL (Routing Policy Specification Language)

- Information related to an Internet resource (AS number, customer cone ,routes etc.) , or supporting functions are contained within RPSL objects and stored in IRR Database

- There are many Route Objects in IRR but we will cover three types of Objects defined in IRR Database

How IRR Works?

IRR Objects:

- AUT-NUM
- Route/Route6
- AS-SET

- In addition to these objects, other IRR objects are; inet-rtr, peering-set, rtr-set, filter-set. Each object has its own purpose. Together they express routing policies

IRR – Aut-Num Object

- Aut-num objects contain the registration details of the Autonomous System Number by the RIR

```
aut-num:        AS64500
descr:          Provider 64500
remarks:        ++ Customers ++
mp-import:      from AS64501 accept
                AS64501
mp-export:      to AS64501 announce ANY
mp-import:      from AS64502 accept
                AS64502
mp-export:      to AS64502 announce ANY
remarks:        ++ Peers ++
mp-import:      from AS64511 accept
                AS64511:AS-A::
mp-export:      to AS64511 announce
                64500:AS-ALL
remarks:        ++ Transit ++
mp-import:      from AS64510 accept ANY
                except FLTR-BOGONs
mp-export:      to AS64510 announce
AS64500:AS-ALL
```

Here is an example of a basic aut-num object registered by AS64500.

IRR – Route/Route6 Object

- Route/route6 objects contain routing information for IPv4/IPv6 address space resources. They show what routes that an AS originates

```
route:              2001:db8:1000::/36
descr:              Provider 64500
origin:             AS64500
mnt-by:             MAINT-AS64500
created:            2012-10-27T12:14:23Z
last-modified:      2016-02-27T12:33:15Z
source:             RIPE
```

The route6 object above shows that AS64500 is allowed to announce the address prefix /36.

IRR- AS-SET Object

- AS-SET Object in IRR is used to describe your network's customer-cone which is a set of ASNs that are owned by your customers

```
as-set:          AS64500:AS-CUSTOMERS
descr:           AS64500 regional customers
members:         AS64501, AS64502
tech-c:          EXAMPLE1-AP
admin-c:         EXAMPLE2-AP
mnt-by:          MAINT-AS64500
last-modified:   2008-09-04T06:40:26Z
source:          APNIC
```

This example contains all of AS64500's customers which include **AS64501** and **AS64502**.

PeeringDB

- PeeringDB is an Open Source database for networks to share their peering information and other relevant information amongst each other

- Networks are responsible for maintaining their own records

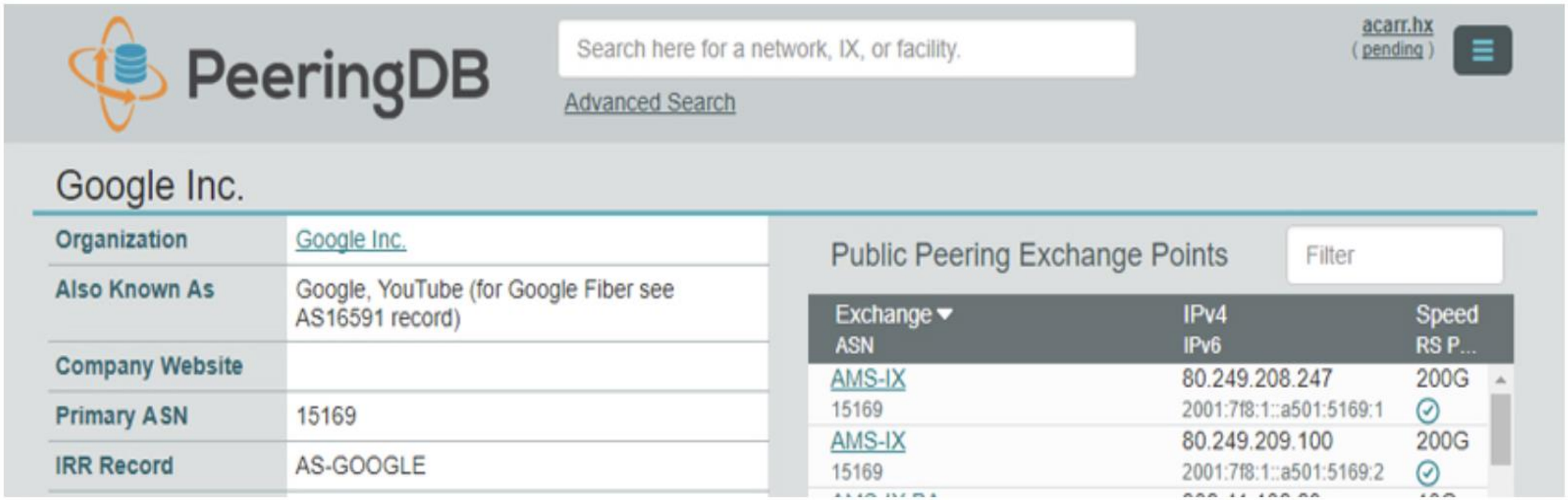
- A PeeringDB record allows you to consolidate your network information in a single location, and look up information about other networks

PeeringDB

- PeeringDB records are used to supplement routing information stored in RPKI and IRR repositories

- PeeringDB allows you to publish information to let other networks know about your network, lets other network know how to contact you, and it is the first place when deciding where and whom to peer with

PeeringDB



In order to use PeeringDB, you must register an account. After registering an account, you should request affiliation to your organization, or ASN to unlock full access to data (for example, non-public network contact information).

The example above shows the PeeringDB record for Google Inc. You can see information such as the name of its IRR record, primary ASN, and public peering exchange points.

BGPSEC - Cryptographic Path Validation

- The community has considered a number of solutions that can eliminate the attacks that can be launched against the RPKI

- Building on the RPKI's guarantees that a BGP route has an authorized origin AS, BGPSEC also provides **path validation**

- BGPSEC is an IETF Standard, RFC 8205

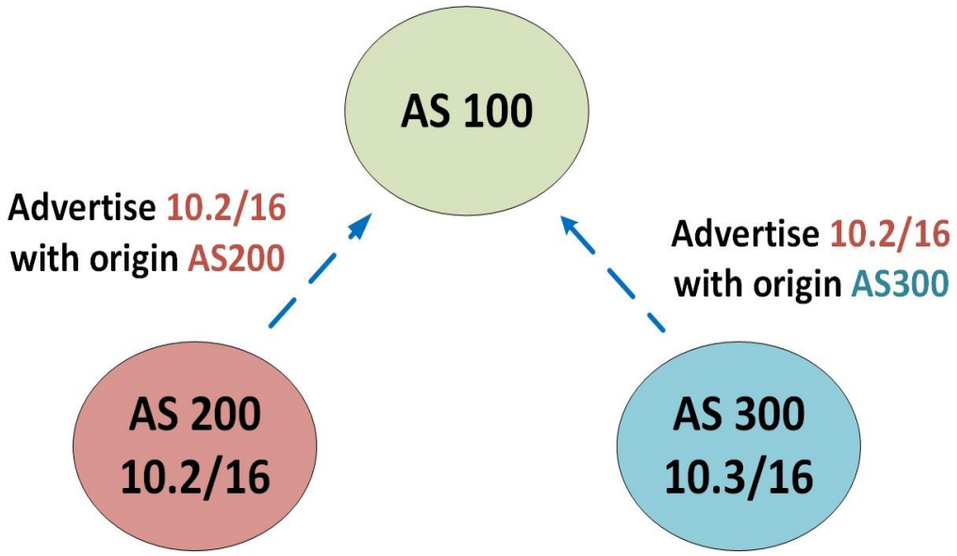
BGPSEC - Cryptographic Path Validation

- BGPSEC builds on RPKI by adding cryptographic signatures to BGP messages

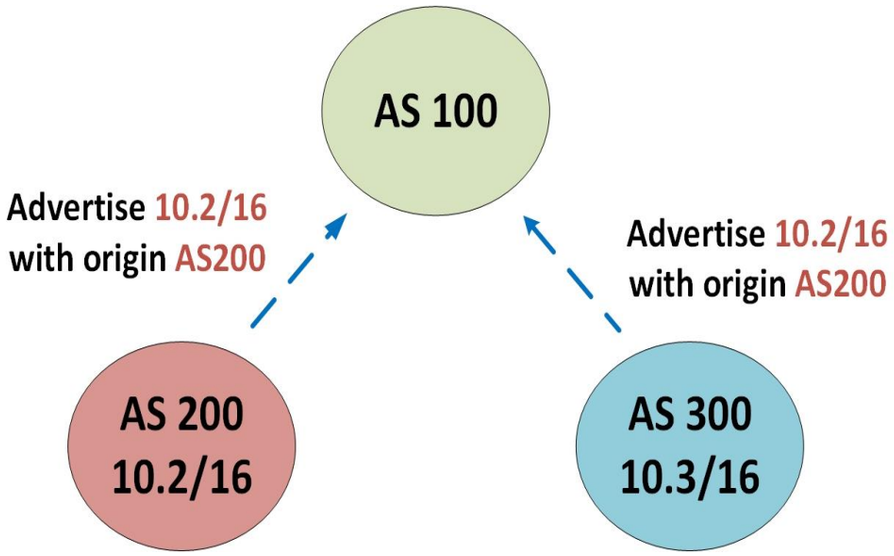
- It requires each AS to digitally sign each of its BGP messages

- The signature on a BGPSEC message covers (1) the prefix and AS-level path; (2) the AS number of the AS receiving the BGPSEC message; and includes (3) all the signed messages received from the previous ASes on the path

BGPSEC solves some issues which RPKI cannot solve



Origin Validation - RPKI



Path Validation - BGPSEC

BGPSEC Challenges

- RPKI requires an Offline Cryptography but BGPSEC is
  an online cryptographic protocol; routers must cryptographically sign and verify every BGP message they send

BGPSEC Challenges

- This high computational overhead, which could require routers to be upgraded with crypto hardware accelerators, could slow down BGPSEC deployment

BGPSEC Challenges

- AS cannot validate the correctness of an AS-level path (and therefore filter bogus routes) unless all the ASes on the path have applied their signatures to the message

- This means the security benefits of BGPSEC apply only after every AS on the path has deployed BGPSEC

Protocol Downgrade Attack

If BGPSEC is partially deployed, it can cause Protocol Downgrade attack, which is selecting insecure short path over secure long path due to BGP Policies of the companies (GAO-Rexford)

Protocol Downgrade Attack

- BGPSEC is changing the way of how BGP works

- It replaces as-path attribute with bgpsec-path attribute

- If there is a partial BGPSEC deployment (current status of Internet), BGPSEC and Legacy routers need to communicate with each other

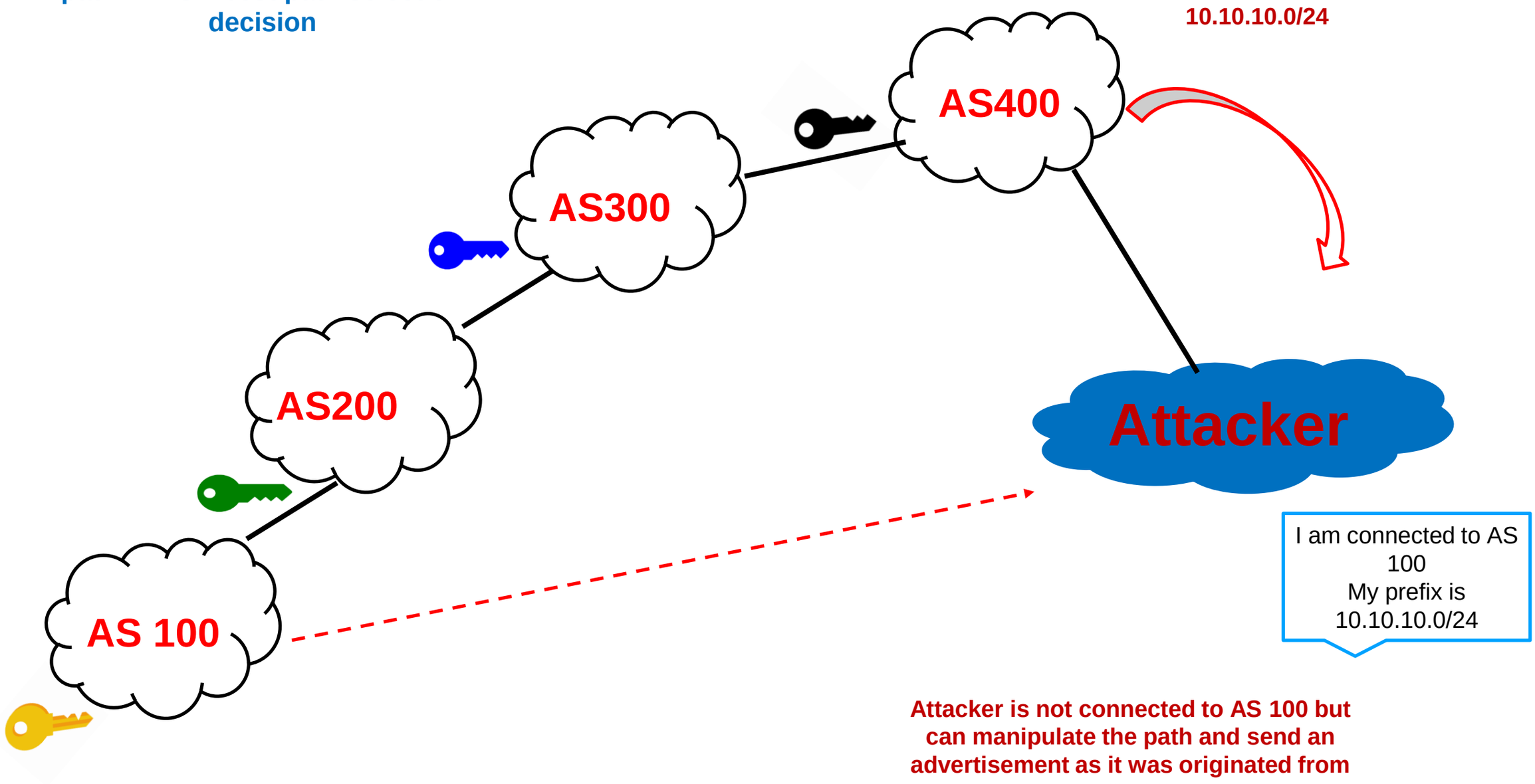
Protocol Downgrade Attack

- To communicate with Legacy routers, BGPSEC speaking routers must receive and send ‘insecure’ routes

- Last but not least, if an AS considers AS-Path length to be selected before Secure route via BGPSEC, short insecure path is preferred to longer secure paths , let’s have a look at the example

Only Attacker is not secure , even if all ASes deploy BGPSEC, still 10.10.10.0/24 is chosen through Attacker AS, as other ASes consider AS-path length is preferred to secure path in BGP best path selection decision

AS 400 chooses Attacker AS to reach 10.10.10.0/24



Attacker is not connected to AS 100 but can manipulate the path and send an advertisement as it was originated from AS 100

Prefix 10.10.10.0/24

BGPSEC Summary

- It is an IETF Standard
- Almost no deployment yet
- Provides Origin and Path Validation
- Requires all ASes to deploy, not much partial deployment benefit
- Protocol downgrade attacks are possible
- Requires BGP messages to change, router hardware to change, so much changes..

BGP in the Datacenter

- BGP is used on the Internet facing routers which are mostly placed in the Datacenter but this is not the topic of BGP for the Internet (Inter-domain routing)

- BGP can be used as an IGP in the Datacenter and we will cover it

- RFC 7938 – Use of BGP for Routing in Large Scale Datacenters cover the usage of BGP in DC networks

BGP in the Datacenter – Why not other IGPs?

- OSPF and IS-IS were considered and used by some of the Web-scale companies initially

- Lack of Multiprotocol support by OSPF , lack of good Open source implementation for link state protocols and most importantly flooding and blast radius impact of link state protocols were the biggest factors of using BGP as IGP in the Datacenters

Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Massively Scale Datacenters or commonly known as Hyper Scale or Warehouse-Scale Datacenters can have 100s of thousands of servers

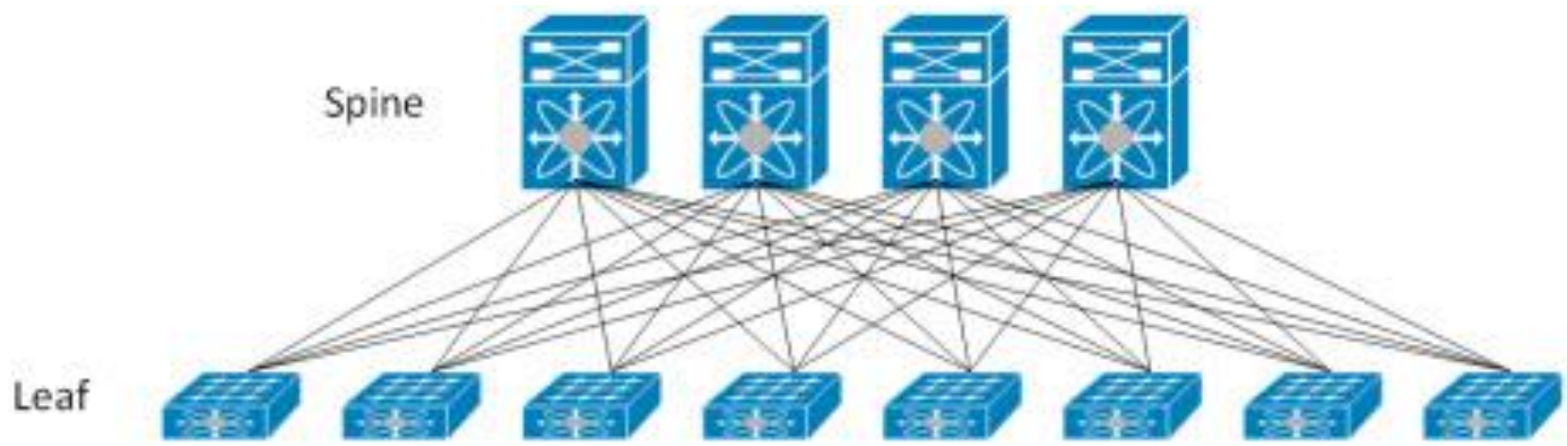
These datacenters require :

- 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

- This requires CLOS topology

CLOS Topology

- In the CLOS topology, there are Leaf and Spine Switches
- There is no shortcut links between Leaf or Spine switches



Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Narrow set of software features/protocols supported by a multitude of networking equipment vendors. (Open Source Implementation is preferred , Exa BGP , FRRouting etc.)

- Routing protocol that has a simple implementation in terms of programming code complexity and ease of operational support (State Machine of BGP vs. IGP protocols)

Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Minimize the failure domain of equipment or protocol issues as much as possible (IGP Flooding scope, Periodic database refresh vs. BGP Incremental Update)

- Allow for some traffic engineering, preferably via explicit control of the routing prefix next hop using built-in protocol mechanics (BGP 3rd part next hop allows some level of Traffic Engineering)

Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Also in the hyper scale DC requirements, having protocol synchronization (Inter-dependency between protocol) is not wanted, thus having a BGP as single protocol is an important parameter to have simple and less OPEX design

Counter Arguments for using BGP in the Datacenter as Routing Protocol

- BGP is perceived as a "WAN-only protocol“ and not often considered for enterprise or data center applications
- BGP is believed to have a "much slower" routing convergence compared to IGPs
- BGP is perceived to require significant configuration overhead and does not support neighbor auto-discovery

Counter Arguments for BGP in DC – It is for WAN

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

- This is partly through, thus there are some tweaks for BGP to use in the Datacenter as Routing Protocol

- In the WAN networks, expectation from BGP is stability, in the Datacenter Stability is important but rapid notification and convergence is more important

Counter Arguments for BGP in DC – It is Slow

- BGP is believed to have a "much slower" routing convergence compared to IGPs, timers can be tuned, BGP can converge much faster than conventional thoughts

For Fast BGP convergence there are in general three timers

- Minimum Route Advertisement Interval
- Keepalive and Hold Timers

Counter Arguments for BGP in DC – It is Slow MRAI Timer

- BGP has MRAI per neighbor

- Events within this minimum interval window are collected together and sent at one shot when the minimum interval expires

- This is essential for the most stable code, but it also helps prevent unnecessary processing in the event of multiple updates within a short duration such as link flaps

Counter Arguments for BGP in DC – It is Slow
MRAI Timer

- The default value for this interval is 30 seconds for EBGP peers, and 0 seconds for IBGP peers. However, waiting 30 seconds between updates is not necessary for a densely connected network such as CLOS topology

- 0 is the more appropriate choice for MRAI timers in EBGP in the DC, because we're not dealing with routers across administrative domains when we deploy EBGP in the Datacenter

Counter Arguments for BGP in DC – It is Slow Keepalive and Hold Timer

- By default, the keepalive timer is 60 seconds and the hold timer is 180 seconds

- This means that a node sends a keepalive message for a session every minute. If the peer does not see a single keepalive message for three minutes, it declares the session dead

Counter Arguments for BGP in DC – It is Slow Keepalive and Hold Timer

- By default, for EBGP sessions for which the peer is a single routing hop away, if the link fails, this is detected and the session is reset immediately

- What the keepalive and hold timers do is to catch any software errors while the link is up but has become one-way due to an error, such as in cabling issue

Counter Arguments for BGP in DC – It is Slow Keepalive and Hold Timer

- Some operators enable Bidirectional Forwarding Detection (BFD) for sub-second detection of errors due to cable issues. However, to catch errors in the BGP process itself, you need to adjust these timers

- Inside the data center, three minutes is too much. Recommended values configured inside the data center are 3 seconds for keepalive and 9 seconds for the hold timer

Counter Arguments for BGP in DC – Lack of BGP Neighbor Auto-Discovery

- There are two IETF draft now for BGP neighbor auto discovery

1. BGP LLDP Peer Discovery
2. BGP Neighbor Discovery Draft

Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- In BGP, neighbor adjacency is configured manually, putting a neighbor IP address and ASN into the BGP

- Neighbor Auto Discovery is a desired future to reduce OPEX when BGP is used inside the Datacenter

Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- This IETF draft can change the behavior of manual adjacency setup, allowing the BGP adjacency on the point-to-point links to be established automatically, using the LLDP protocol

- LLDP is an Industry standard

Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- It is intended to replace proprietary Cisco Discovery Protocol (CDP)

- With LLDP, network devices such as switches operating on Layer 2 (Link layer) of OSI model can collect upper layer information of neighboring device, such as IP address, OS version etc.

Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- BGP peering discovery using LLDP could be used in large Layer 3 data centers where eBGP is being used as a single routing protocol

- Deployment of BGP with enabled BGP peering discovery using LLDP in large data centers in the future would significantly lower the BGP configuration overhead

Counter Arguments for BGP in DC – BGP Neighbor Discovery Draft

- This approach suggests the change way of BGP hello message

- Instead of using other protocols such as LLDP, the draft introduces a new BGP Hello message

Counter Arguments for BGP in DC – BGP Neighbor Discovery Draft

- The message is sent periodically on the interfaces where BGP neighbor auto-discovery is enabled to the multicast IP address using UDP port 179

- The hello message contains ASN of the sender along with IP address, router id etc

BGP Path Hunting

- When BGP is used in the DC, based on the AS allocation, it might suffer from the BGP Path Hunting behavior

- BGP Path Hunting will slow down the convergence based on the topology and ASN allocation schema when there is a failure in the network, link or node failure

- Let’s have a look at the details of BGP Path Hunting before start talking about BGP ASN allocation

BGP Path Hunting

- Without topology information a Router does not know the physical link state of every other node in the network, it doesn’t know whether the route is truly gone (because the node at the end went down itself) or is reachable via some other path

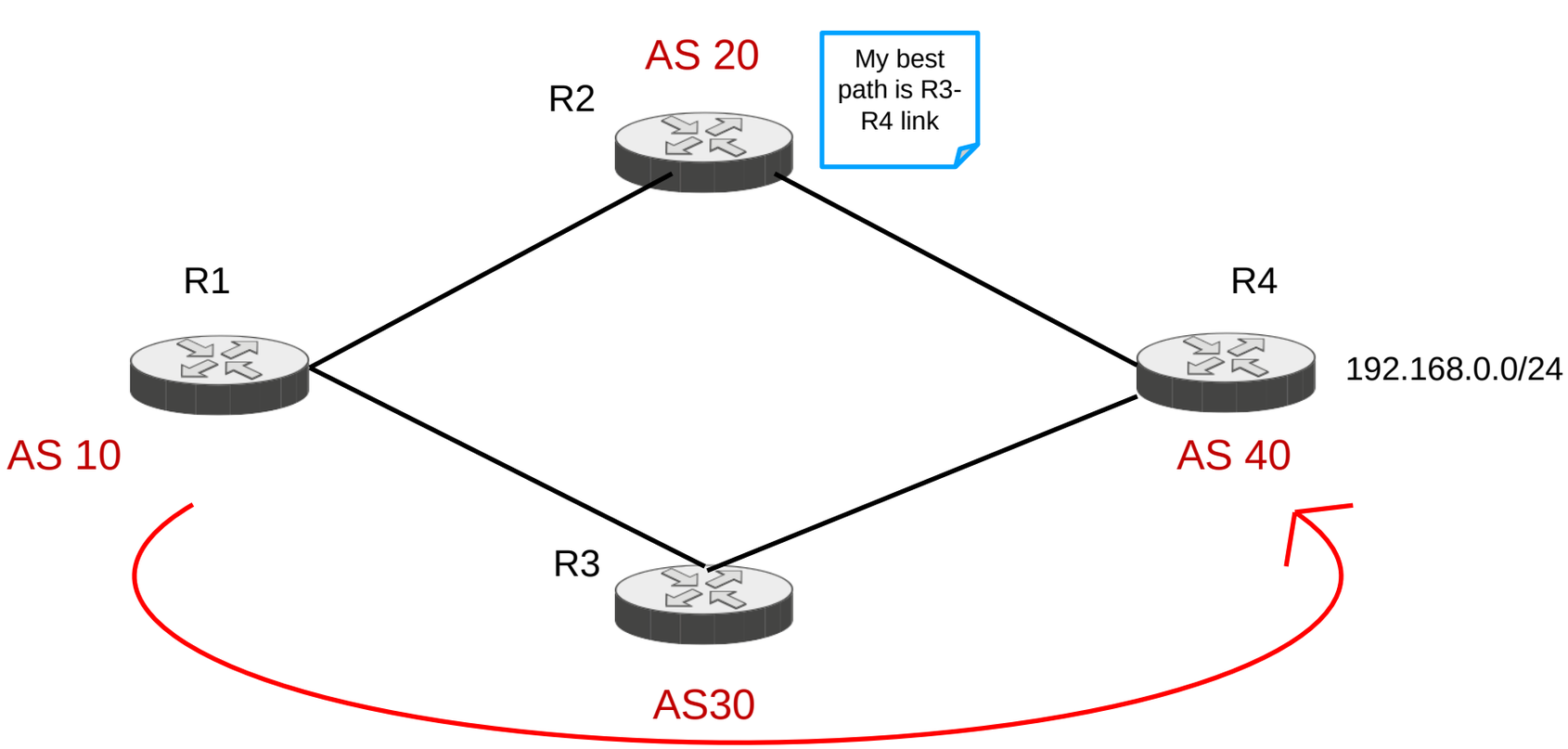
- That’s why a Router proceeds to hunt down reachability to the destination via all its other available paths. This is called path hunting

BGP Path Hunting

Let’s assume R1 selected best path to 192.168.0.0/24 as R3

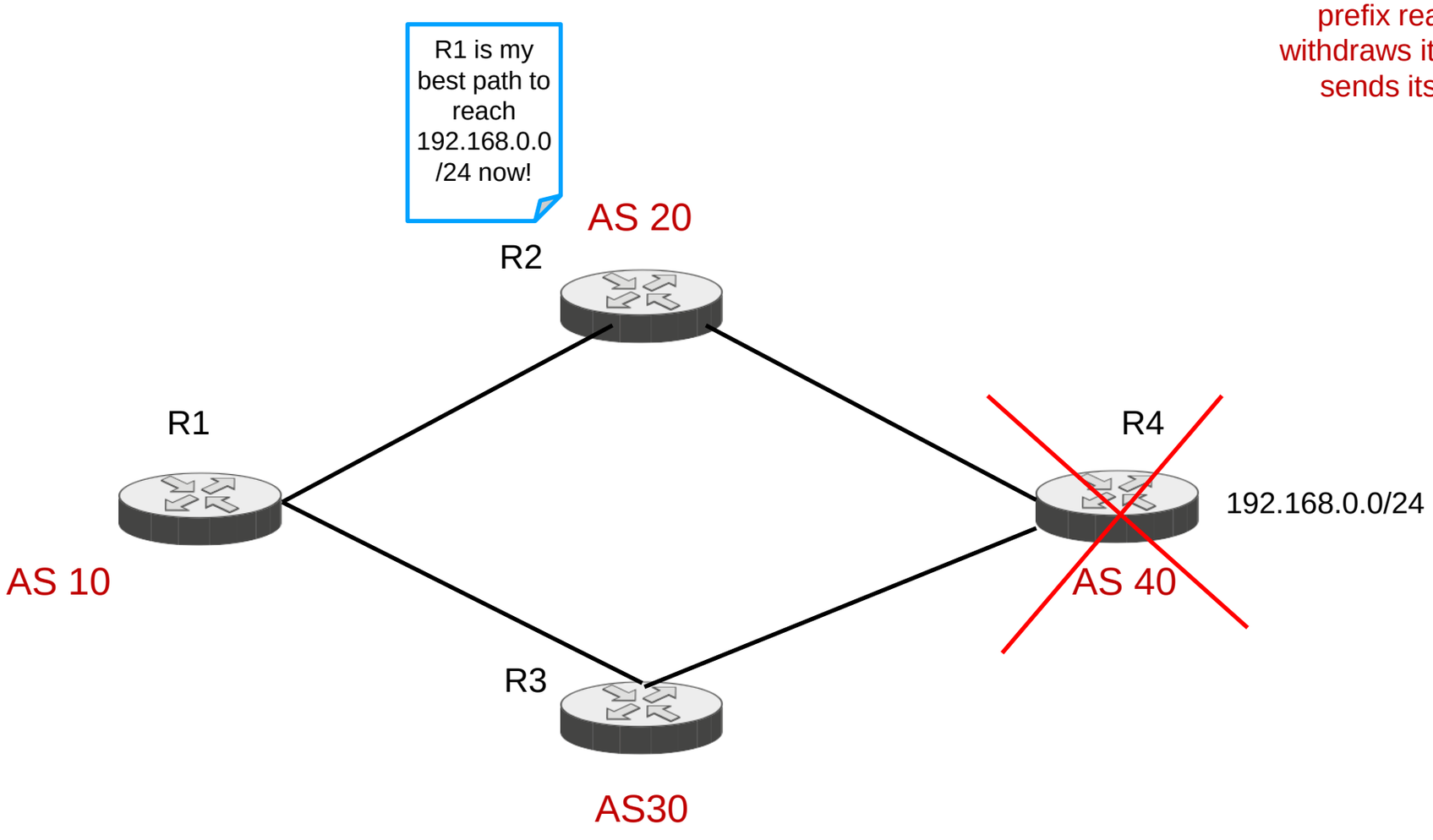
R1 advertises [R1 R3 R4] As-Path to R2

R2 accepts the advertisement but doesn’t use it , as R2 has shorter path to 192.168.0.0/24



BGP Path Hunting

Now, when the Router R4 fails, R2 loses its best path to 192.168.0.0/24, and so it re-computes its best path via R1, AS_PATH [R1, R3, R4] and sends this message to R1.
R2 also sends a route withdrawal message for the prefix to R1
When R3's withdrawal for the prefix reaches R1, R1 also withdraws its route to prefix, and sends its withdrawal to R2



BGP Path Hunting

- EBGP AS number allocation will trigger path hunting when there is a failure to the destination

- Path Hunting will slow down the convergence which is not good for the Datacenter BGP

BGP Path Hunting

- Path Hunting in BGP is a normal process for convergence, you cannot say I don’t want Path Hunting, it is how protocol works (Similar to EIGRP)

- We will look at next how ASN allocation should happen to reduce convergence impact of BGP Path Hunting behavior when EBGP is used inside the Datacenter

ASN Numbering Schema when EBGP is used inside Datacenter

- All ToR (Top of Rack) switches (Sometimes referred as Tier 3 switches) are assigned their own ASN , unique ASN

- Leaf Switches (Tier 2 Switches) inside a pod have a same ASN, but leaves in different pod have a unique ASN.POD sometimes referred as Cluster

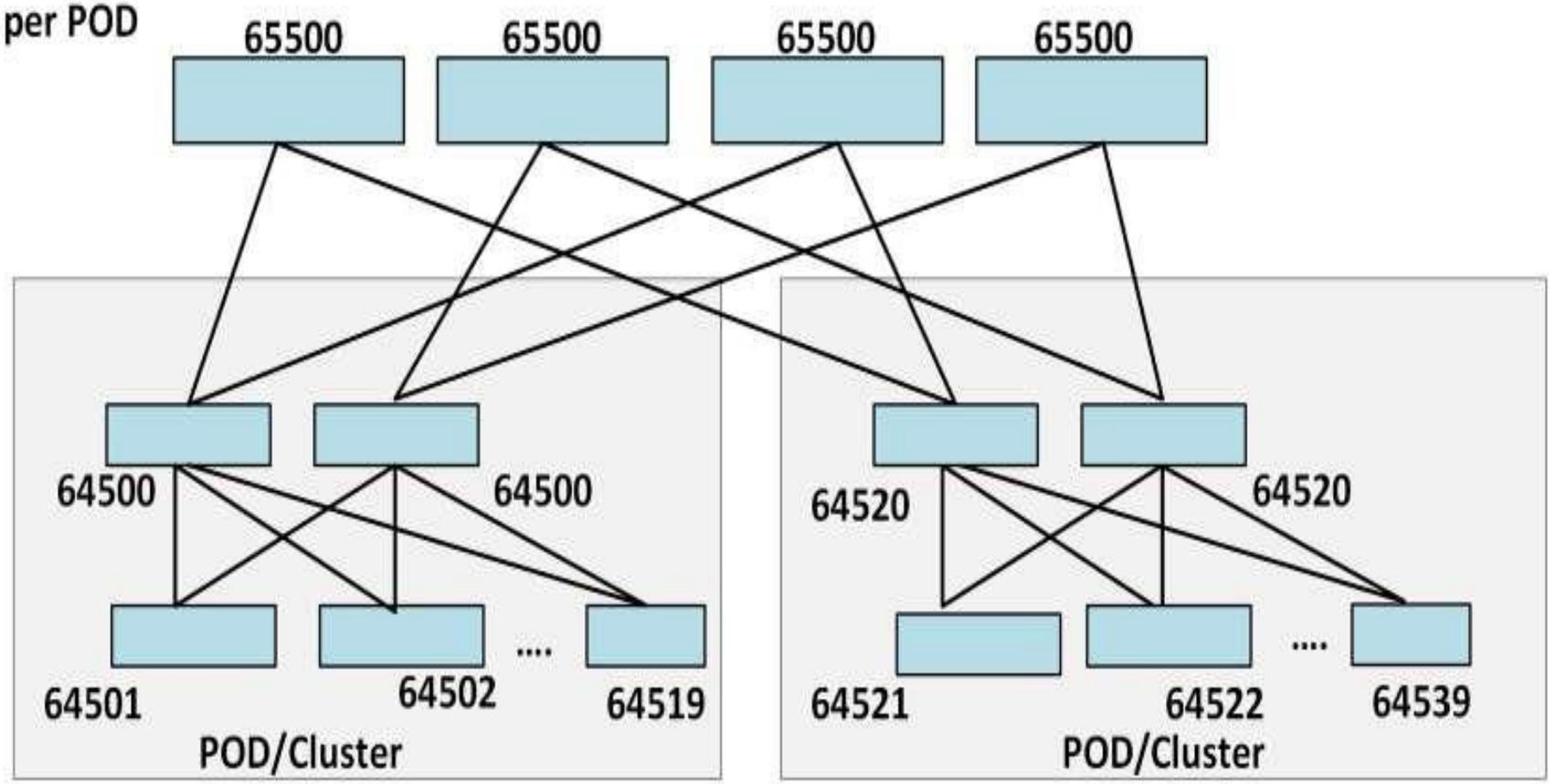
- Spines share a common ASN, Spine switches sometimes referred as Tier 1 devices

Recommended BGP ASN Allocation Schema for 3 Tier CLOS Networks

Unique AS per ToR

Same AS on Leaf switches per POD

Same AS on all SPINES



Possible Problems with BGP ASN Allocation Schema

- If 2 byte ASN space is used, Private ASN range is recommended as there might be mistake to leak Public ASN to the Peers or Transit

- Since Operator knows most of the public AS numbers of the large companies, allocating private AS number for DC usage is better for troubleshooting

Possible Problems with BGP ASN Allocation Schema

- 2 byte ASN space Private range is limited to 1023 AS

- There might be much more TOR switches in the DC than 1023 AS

- Two Options : Either use 4 byte ASN or assign same AS numbers on different POD/Cluster’s TOR switches

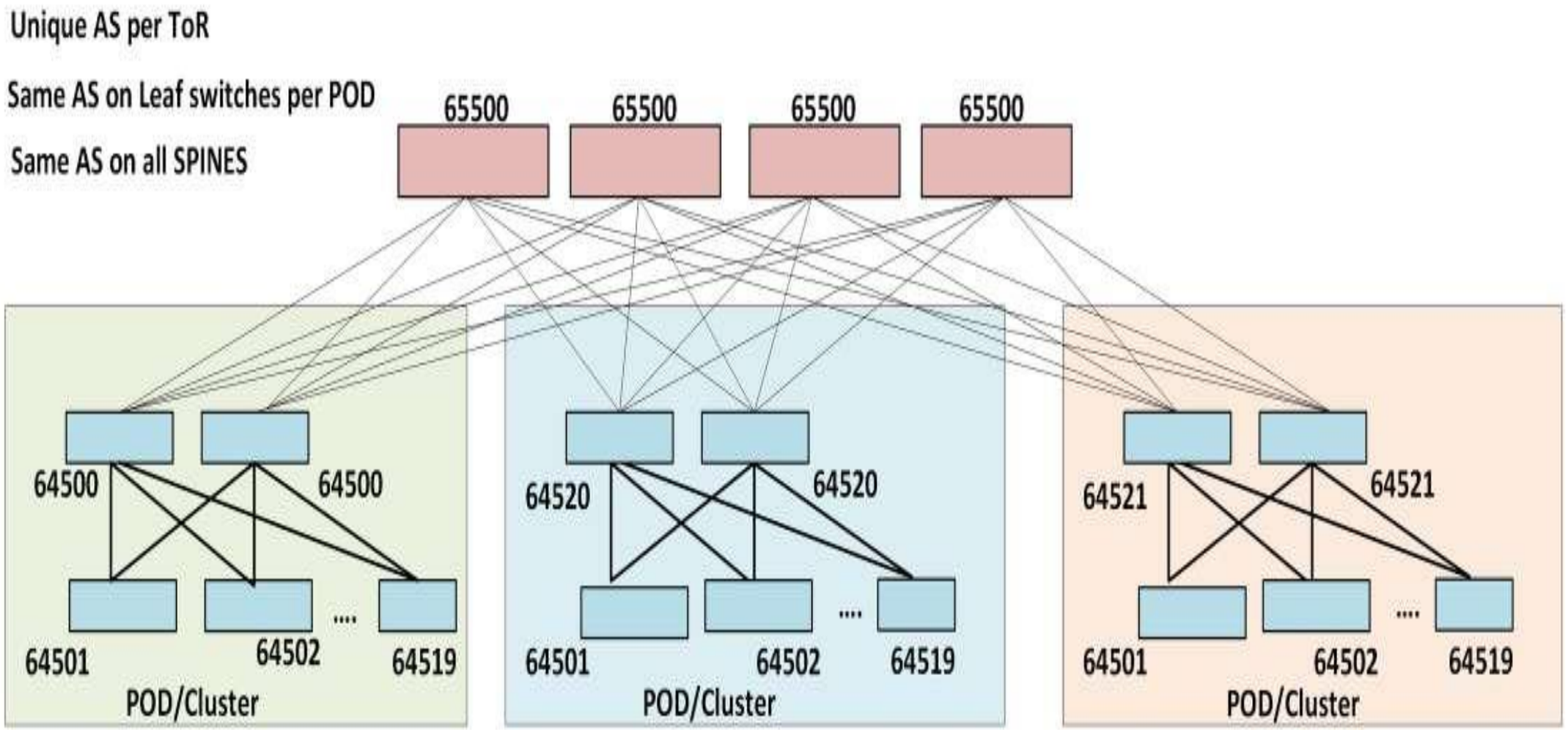
Possible Problems with BGP ASN Allocation Schema

- 4 byte ASN is not still supported by some BGP implementations thus it might limit the vendor selection of DC equipment

- Also using 4 bye ASNs puts additional protocol complexity

- If same AS numbers are used in the different Cluster TOR switches, when there is a traffic between separate Cluster TOR, Allow-as in feature is required

Recommended BGP ASN Allocation when Private 2 Byte Private ASN is used



Allow AS-IN feature should be set on each ToR

Allow AS-IN feature should be set on each ToR

Allow AS-IN feature should be set on each ToR

BGP in the Wide Area Network

- BGP is designed first for the Wide Area Network and today it is mostly still used for the Wide Area Network

- When it is first invented, it was for the IPv4 Unicast address family but today BGP supports 20 different address families which are mostly used for the WAN use cases

- There are some Datacenters (Hyper-scale/MSDC) which use BGP in their Datacenter as IGP

BGP in the Wide Area Network

- BGP is used for the Inter-domain routing mainly
- It is the only protocol which is used in the Global Internet
- It is the most scalable routing protocol and for the Inter-domain routing scalability is must as there are 800k+ prefixes in the Global routing table which is commonly known as Default Free Zone (DFZ)

BGP in the Wide Area Network

- Stability in the Default Free Zone is important since the failure impacts so many networks and so many routers in the Internet

- Policy is the key function of BGP which companies in the Internet express their business intent with BGP

BGP in the Wide Area Network

- BGP is used in the WAN network not just for the Internet but for VPNs (MPLS Layer 2 and Layer 3 VPNs) commonly as well

- When it is used in the WAN, stability and multi-protocol capability is very important along with the Traffic engineering and reliability

- In the Datacenter, instead of Stability and multi-protocol capability , Fast convergence , scalability , policy control are the key parameters expected from BGP

BGP PIC – Prefix Independent Convergence

- BGP PIC is used to make the re-convergence after topology change independent of the number of BGP prefixes

- BGP PIC is an evolution of Forwarding Plane to have BGP Fast Convergence

Three Different FIB Architecture

- BGP PIC has two modes , BGP PIC Core and BGP PIC Edge

- BGP PIC Core provides fast convergence in case of Core Link or Node Failure, IGP converges is important for BGP PIC Core

- BGP PIC Edge provides fast convergence in case of Edge Link or Node Failure

Three Different FIB Architecture

- For us to understand both BGP PIC Core and Edge and how they provide Fast Convergence for the failure scenarios, let’s have a look at different FIB architectures in the routers

- These are Flat FIB , Hierarchical FIB and Generalized FIB

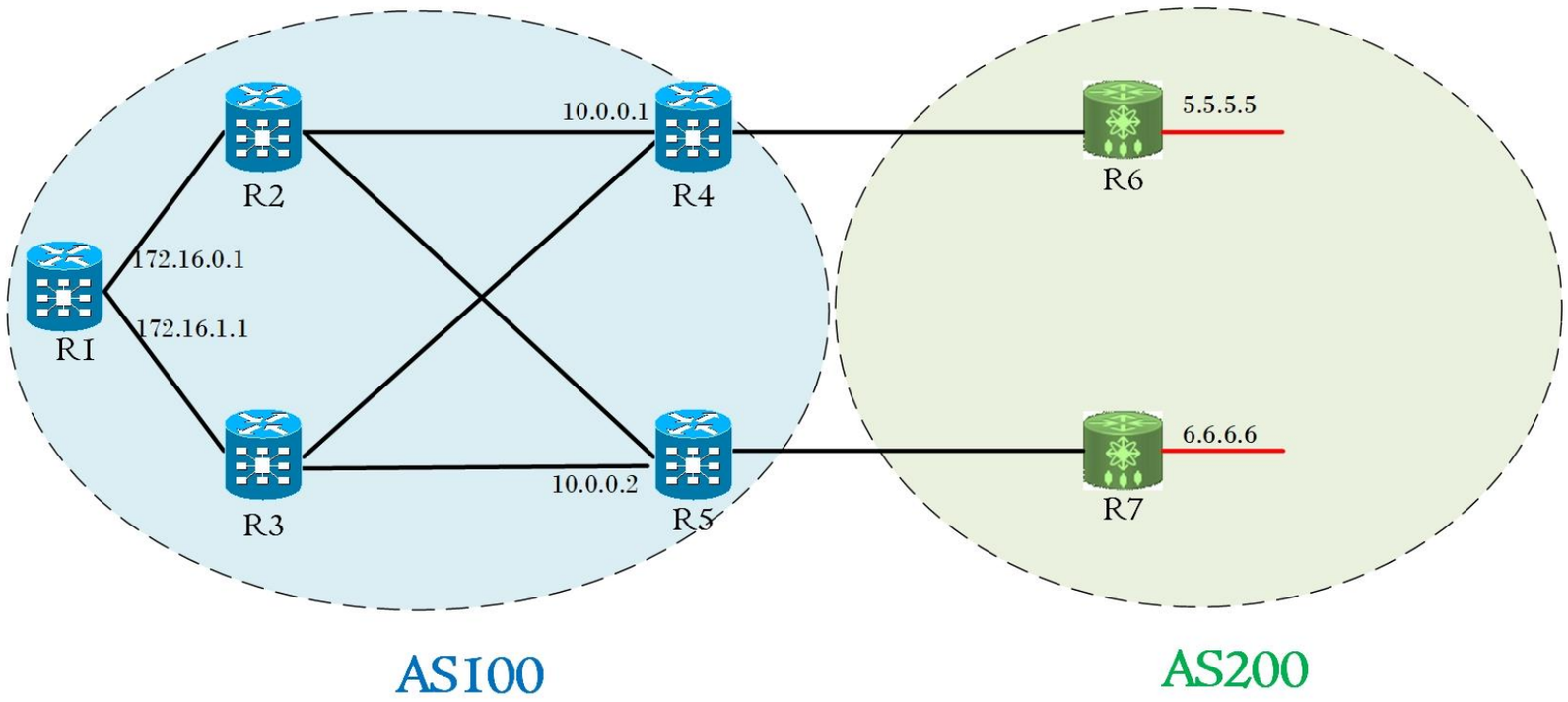
Three Different FIB Architecture – Flat FIB Architecture

R1's RIB

Prefix	Next Hop or Outgoing Interface	Protocol
5.5.5.5	10.0.0.1	BGP
6.6.6.6	10.0.0.2	BGP
10.0.0.1	172.16.0.1	IGP
10.0.0.1	172.16.1.1	IGP
10.0.0.2	172.16.0.1	IGP
10.0.0.2	172.16.1.1	IGP

R1's FIB

Prefix	Next Hop or Outgoing Interface
5.5.5.5	172.16.0.1
6.6.6.6	172.16.1.1
10.0.0.1	172.16.0.1
10.0.0.1	172.16.1.1
10.0.0.2	172.16.0.1
10.0.0.2	172.16.1.1



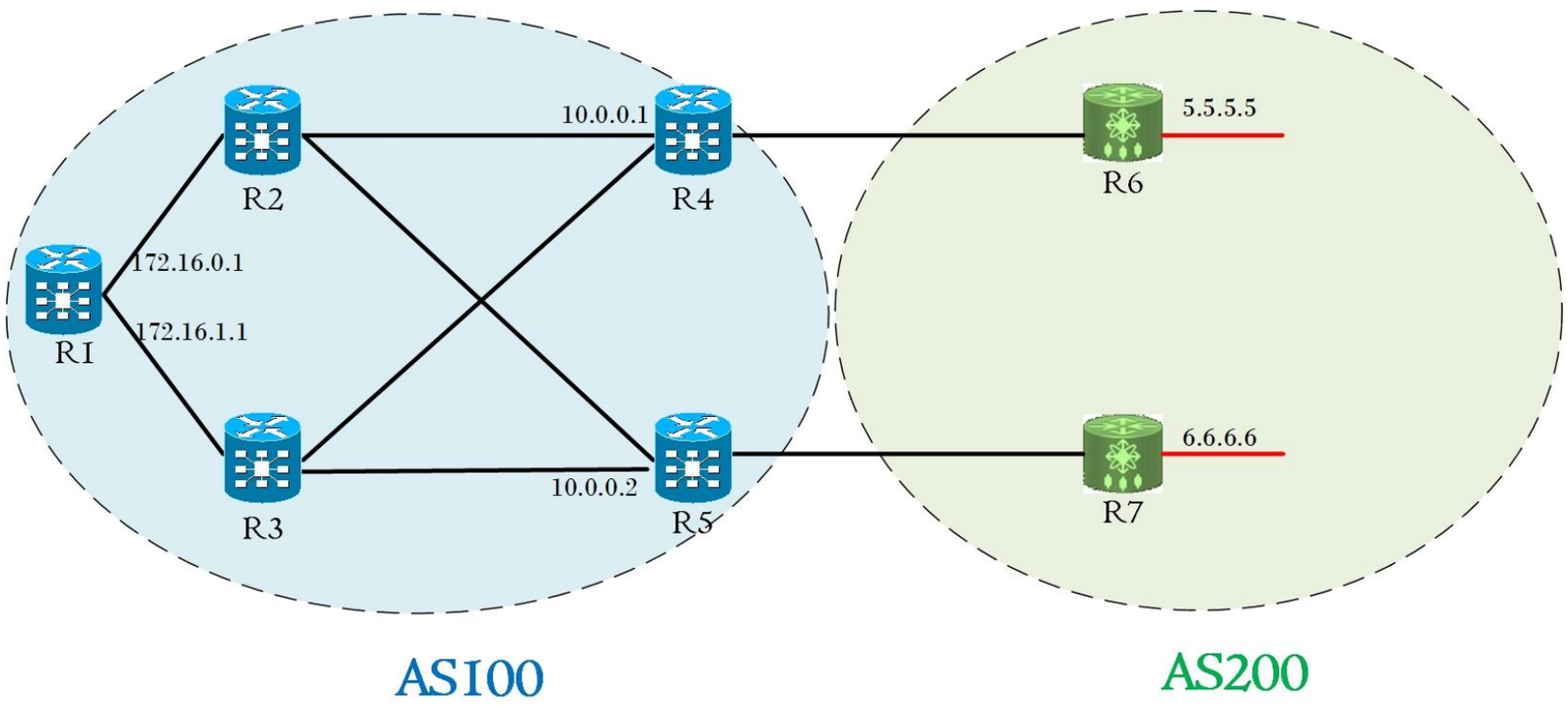
Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE

R1's RIB

Prefix	Next Hop or Outgoing Interface	Protocol
5.5.5.5	10.0.0.1	BGP
6.6.6.6	10.0.0.2	BGP
10.0.0.1	172.16.0.1	IGP
10.0.0.1	172.16.1.1	IGP
10.0.0.2	172.16.0.1	IGP
10.0.0.2	172.16.1.1	IGP

R1's FIB

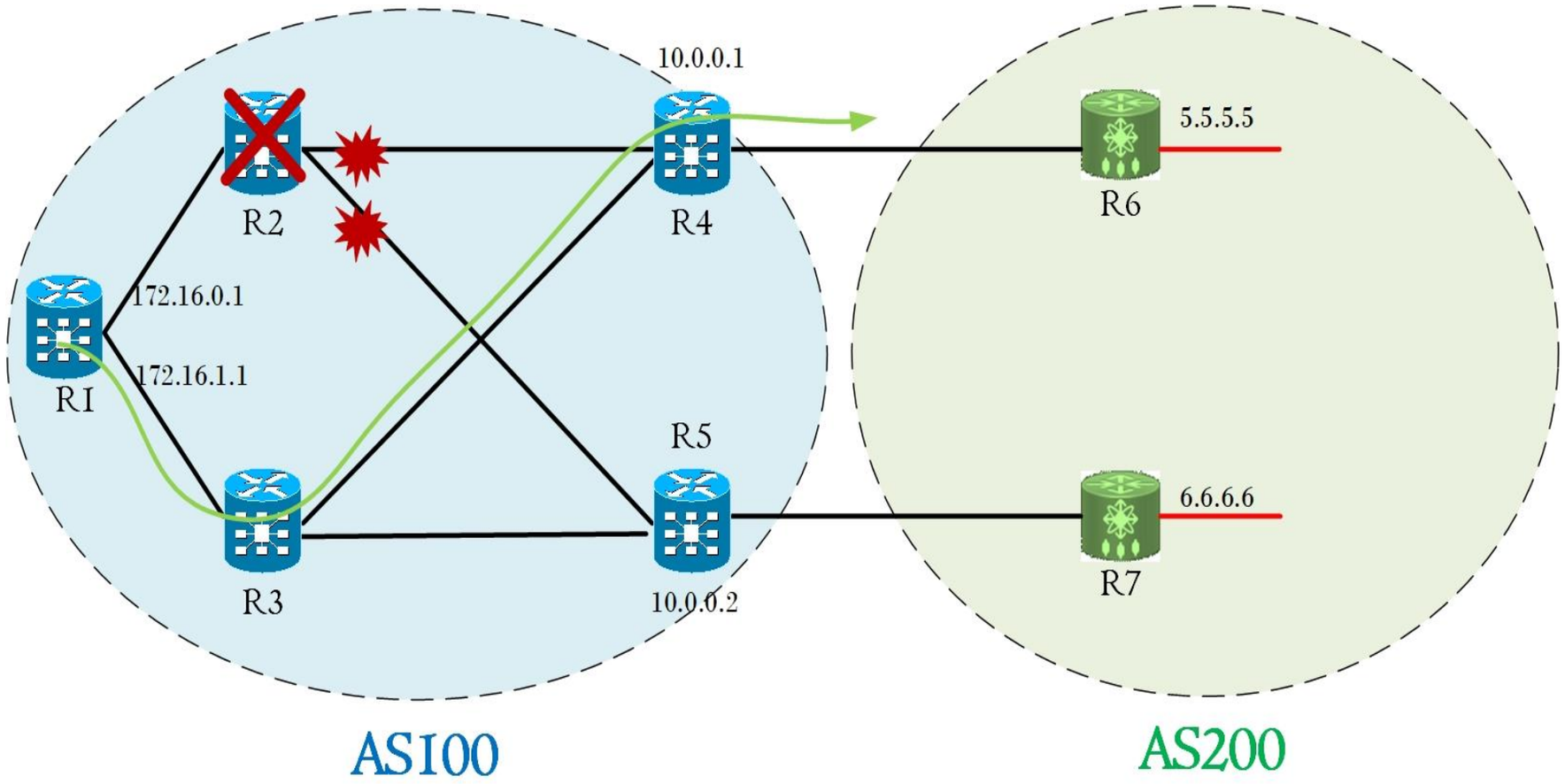
Prefix	Next Hop or Outgoing Interface
5.5.5.5	10.0.0.1
6.6.6.6	10.0.0.2
10.0.0.1	172.16.0.1
10.0.0.1	172.16.1.1
10.0.0.2	172.16.0.1
10.0.0.2	172.16.1.1



With Hierarchical FIB
R1's FIB has recursive route for BGP Next hop
Thus both RIB and FIB has Recursion

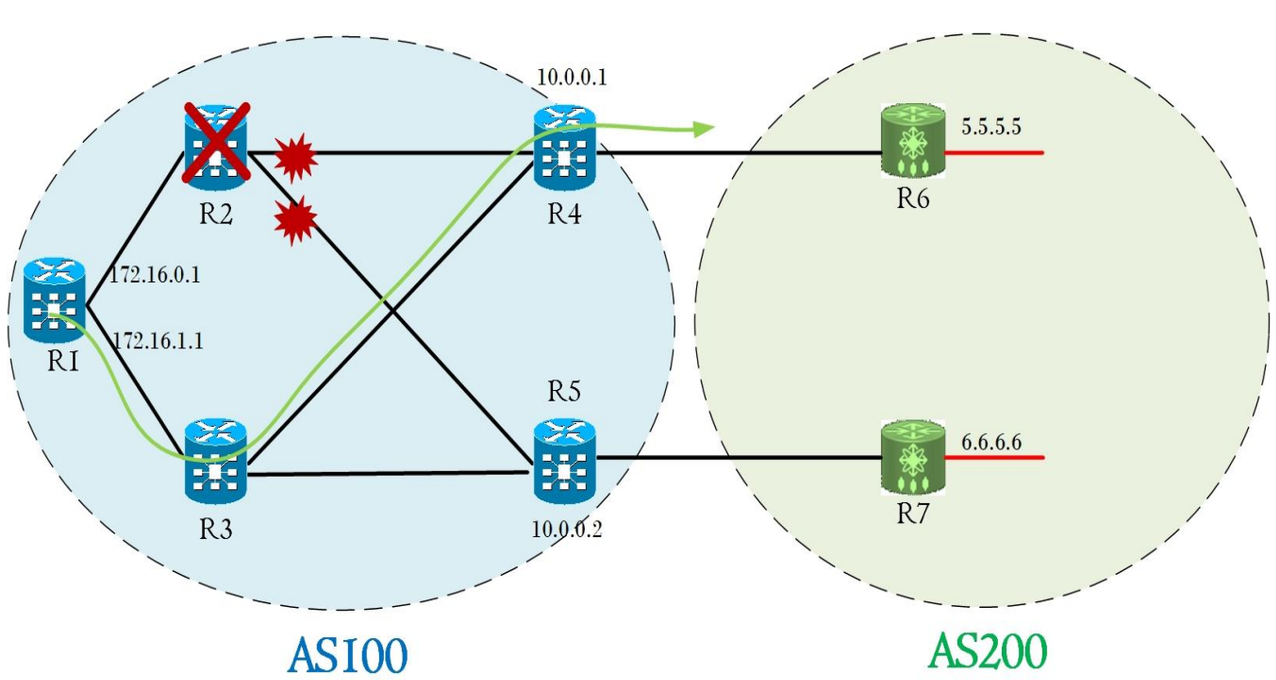
Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE

- Hierarchical FIB helps for the BGP PIC Core, when there is a Core Link or Device Failure, BGP next hop doesn't change, FIB just changes the IGP next hop if there is one available



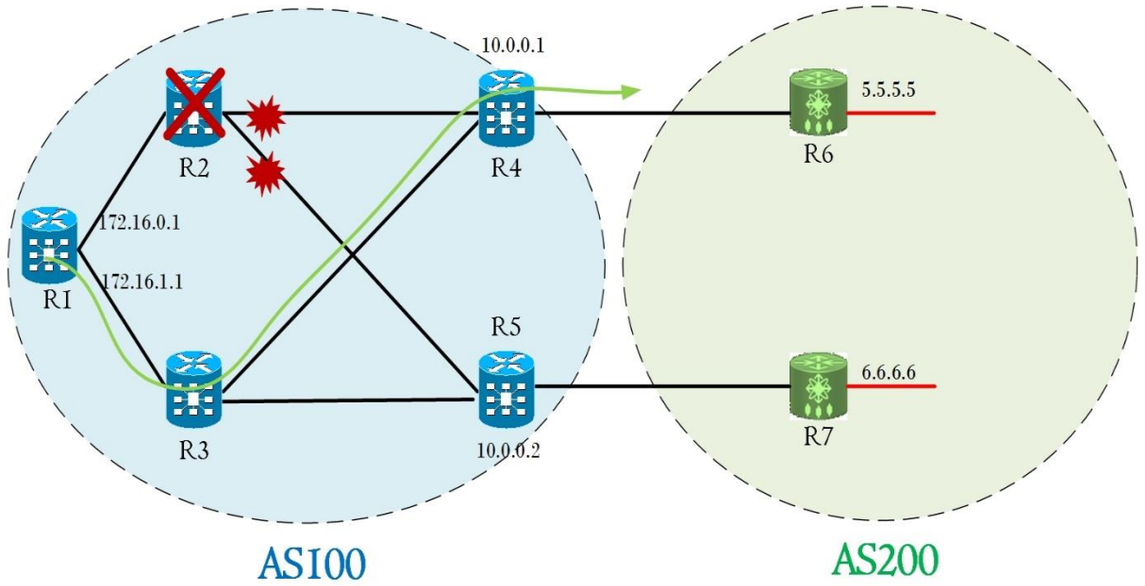
Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE

5.5.5.5 and 6.6.6.6 as a BGP Next Hop stays up, FIB just changes the IGP next hop which is used to reach BGP Next Hop



Prefix	Next Hop or Outgoing Interface
5.5.5.5	10.0.0.1
6.6.6.6	10.0.0.2
~~10.0.0.1~~	~~172.16.0.1~~
10.0.0.1	172.16.1.1
~~10.0.0.2~~	~~172.16.0.1~~
10.0.0.2	172.16.1.1

Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE



- Hierarchical FIB is only helpful for BGP PIC Core not for BGP PIC Edge

- For BGP PIC Core, R1 doesn't need to have two BGP next hop for external prefix

- For BGP PIC Edge , two BGP Next hop for external prefix is needed

Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

- Generalized FIB Architecture relies on a concept, sharing BGP Path List for the external prefixes and IGP Path List for the BGP Next Hops

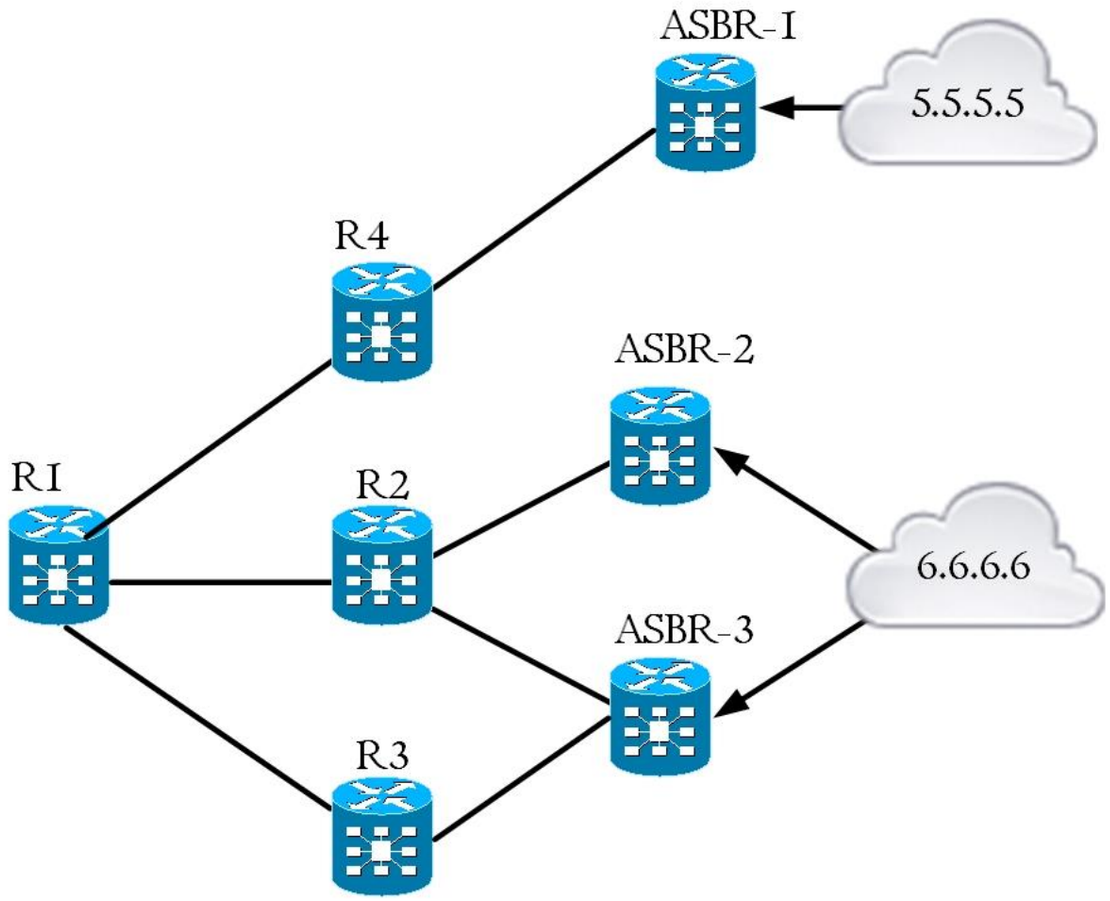
- External BGP Prefixes will share the multiple BGP Next hops and BGP next hop will share multiple IGP Next hops

Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

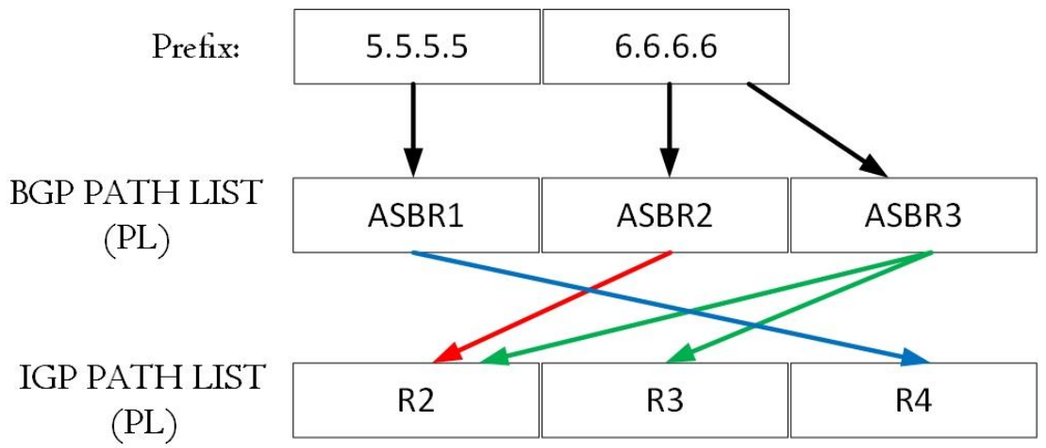
- Generalized FIB is an evolution of Hierarchical FIB and helps for the Core and Edge link or node failure situations

- Many vendors provide Generalized FIB hardware

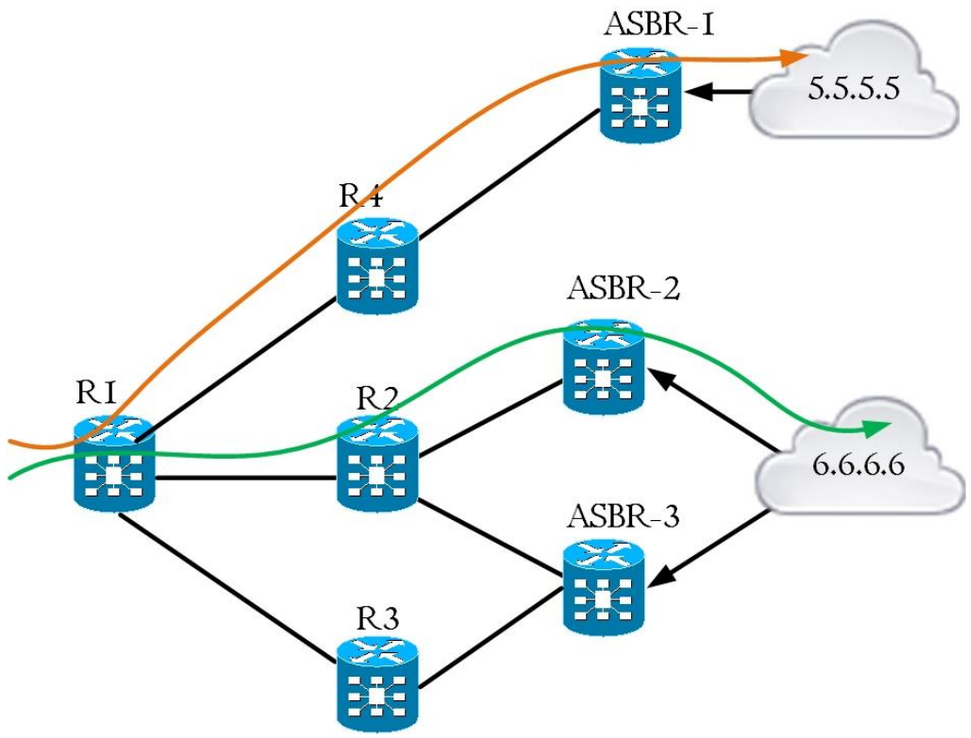
Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE



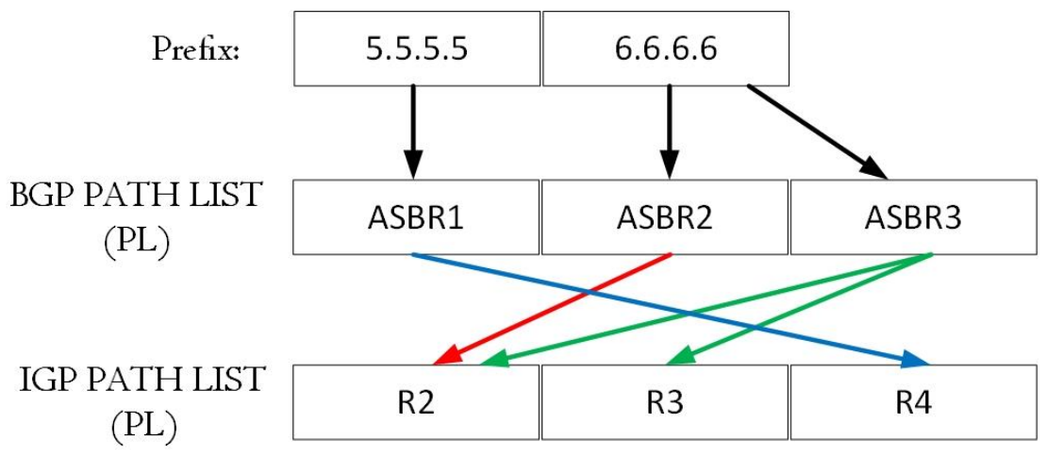
Let's assume R1 selects ASBR 2 as the next hop to reach 6.6.6.6
Thus in the BGP PL, ASBR 2 will be Primary path and ASBR 3 will be secondary path



Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

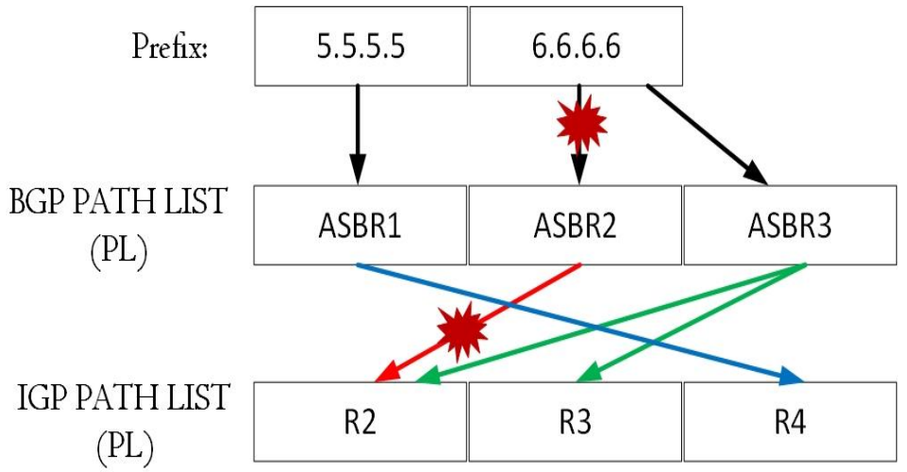


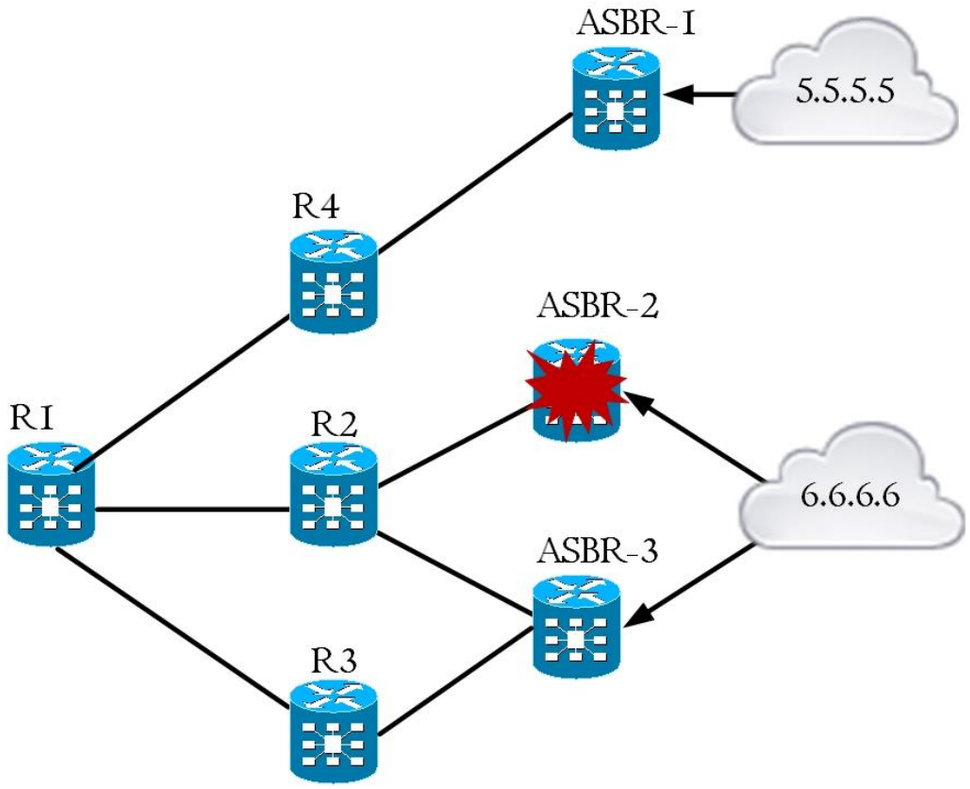
Let's assume R1 selects ASBR 2 as the next hop to reach 6.6.6.6
Thus in the BGP PL, ASBR 2 will be Primary path and ASBR 3 will be secondary path

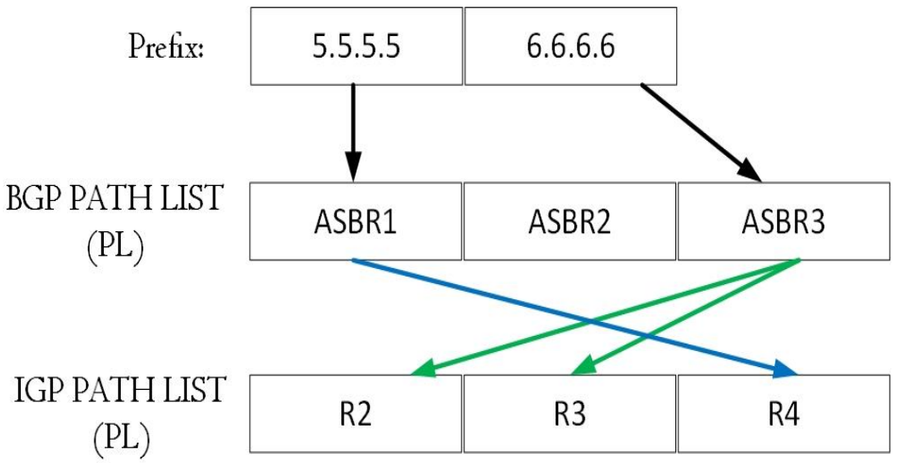


Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

If ASBR 2 fails, IGP detects the failure and R2 is removed from IGP Path List and FIB Backwalking process will start removing any BGP Next hop from the BGP Path List if there is no available next hop in the IGP Path List pointing for it

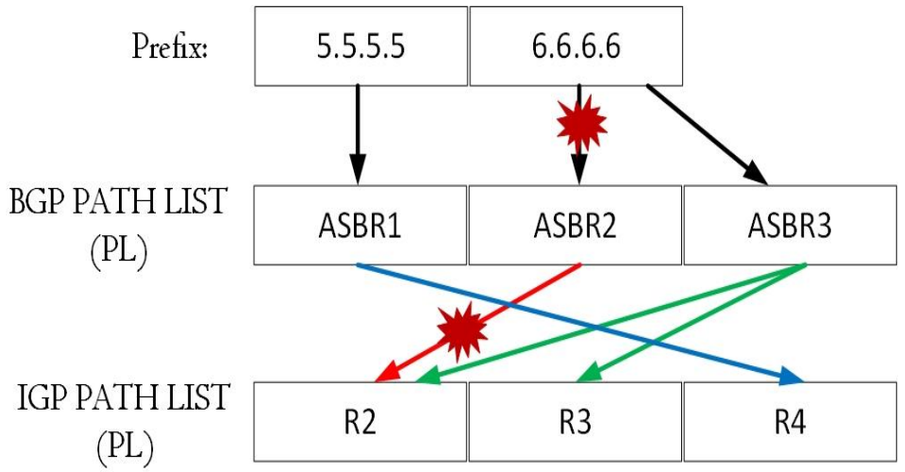


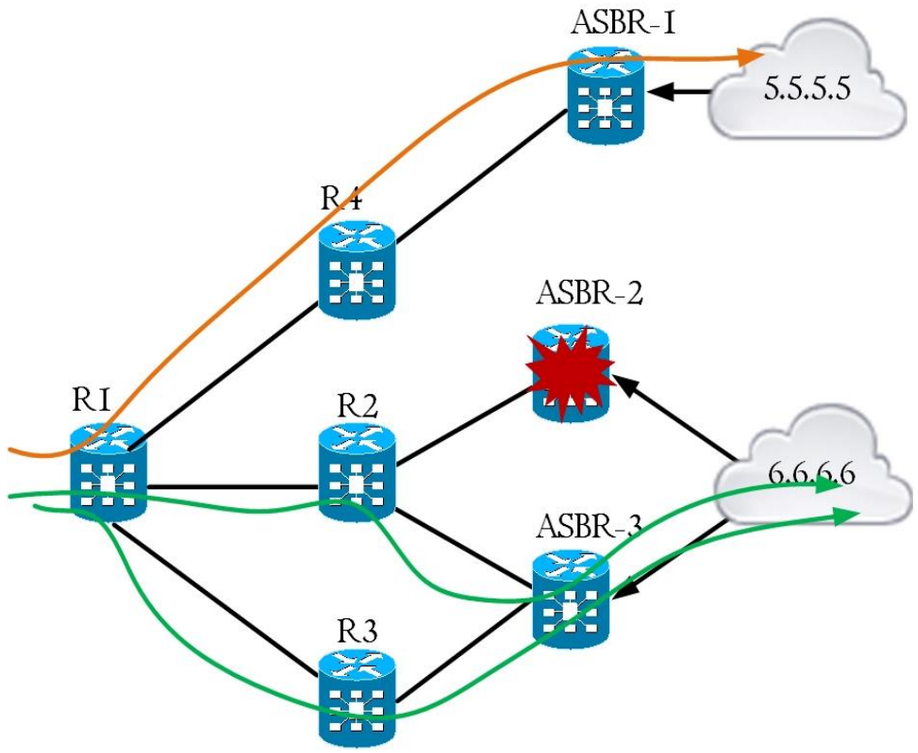


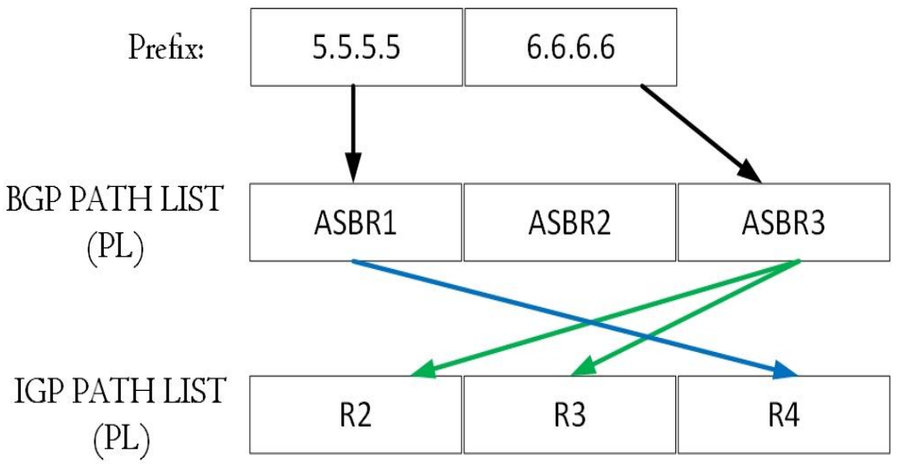


Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

ASRB3 becomes best path immediately after ASBR 2 is removed from the BGP Path List.
Backwalking process to find whether there is any BGP next hop for the external prefixes in the BGP Path List takes only couple milliseconds







BGP PIC Core and Edge Summary

- BGP PIC Core and Edge provides fast convergence incase of Core and Edge link or node failure situations

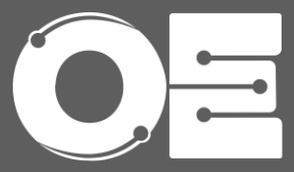
- Flat FIB points direct Layer 2 adjacency not to the BGP next hop or BGP Path List for the external prefixes, thus incase of failure, all the prefixes need to be refreshed with the new next hop which can take so much time

- Hierarchical and Generalized FIB helps for the BGP PIC Core and Edge scenarios

BGP PIC Core and Edge Summary

- If there is no alternate IGP next hop, BGP PIC Core is useless, if there is no alternate BGP next hop, BGP PIC Edge is useless as well

- BGP PIC provides fast dataplane convergence , control plane convergence will follow after that

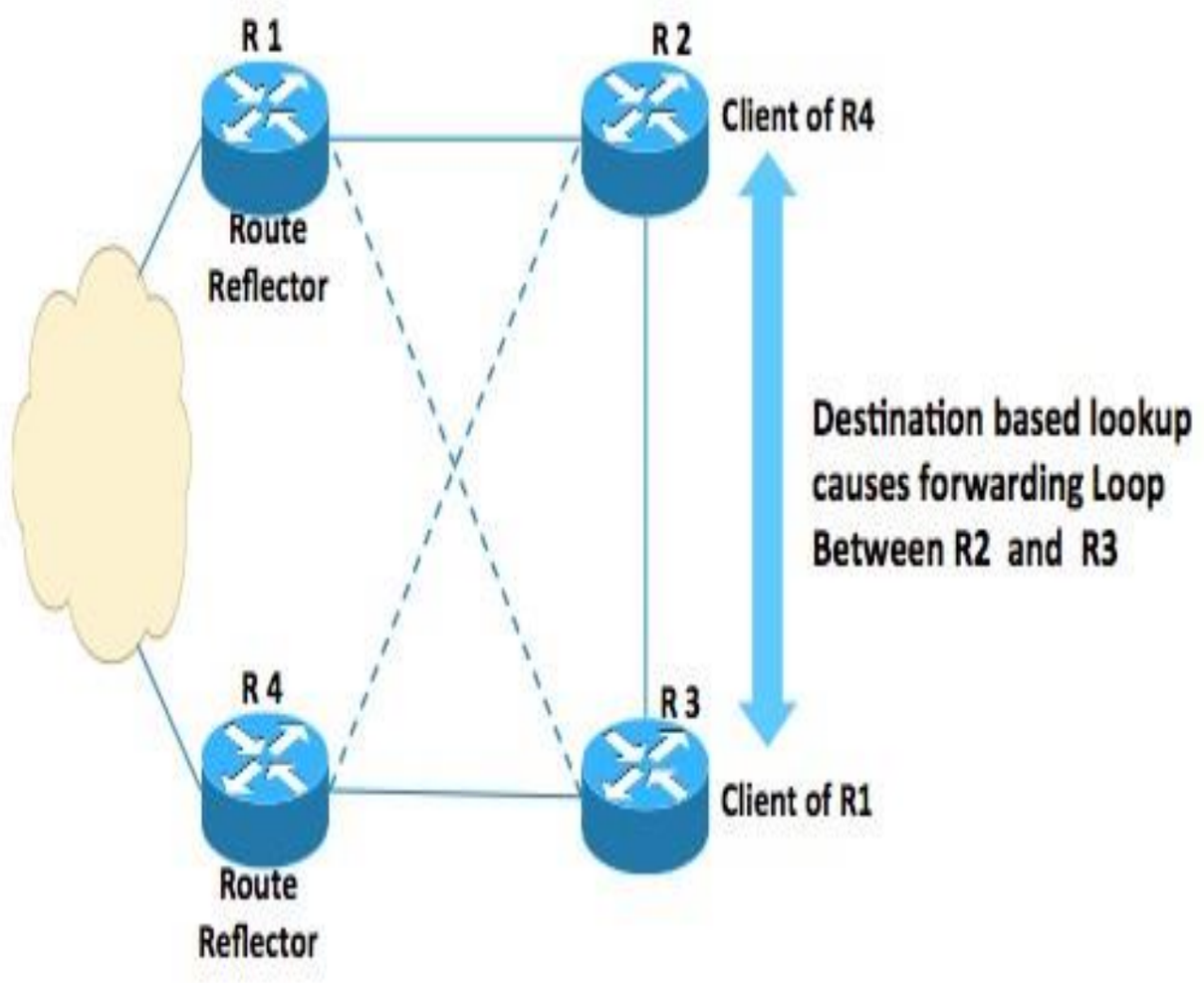


BGP Case Studies

Border Gateway Protocol

BGP Routing Loop with IP Route Reflector

In the diagram below; R2 is route reflector client of R4, R3 is route reflector client of R1.



1. MPLS or any tunneling mechanism is not enabled. What is the problem with this design ?

2. Would you have the problem if MPLS is enabled ?

R3 should be a client of R4 instead of R1 . R2 should be a client of R1 instead of R4. Then, we wouldn't have this problem

- Permanent forwarding loop will occur. (Not micro-loop which is resolved automatically when the topology converged).

- Suppose prefix A is coming from the cloud to the route reflectors

- Route reflectors will reflect to their clients by putting as next-hop themselves

- When the packet comes to R2 for example, R2 will do the IP based destination lookup for the prefix A and find the next hop as R4 so it will send the packet to R3

- Because R3 is the only physical path towards R4

- When R3 receives the packet, It will do the destination based lookup for prefix A then it will find next hop R1

- To reach R1, R3 will send the packet to R2

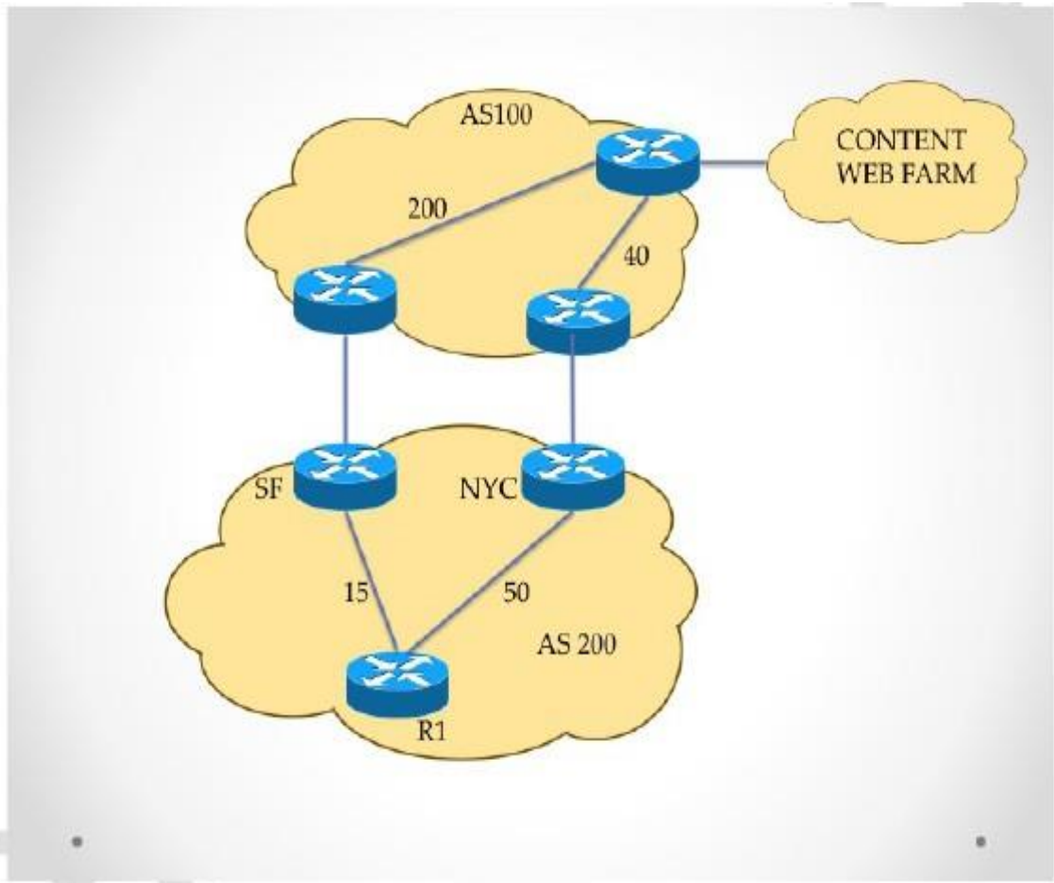
R2 will do the lookup for prefix A and send it to R2 , R3 will send it back. Packet will loop between R2 and R3

If MPLS would be enabled, we wouldn’t have the same behavior since when R2 do the destination lookup for the prefix A, it will find the next hop R4 but in order to reach to R4, it would push the transport label

When R3 receives the packet from R2, R3 wouldn’t do the IP based lookup but MPLS label lookup so it would swap the incoming label from R2 to outgoing label towards R4

BGP Hot Potato Routing

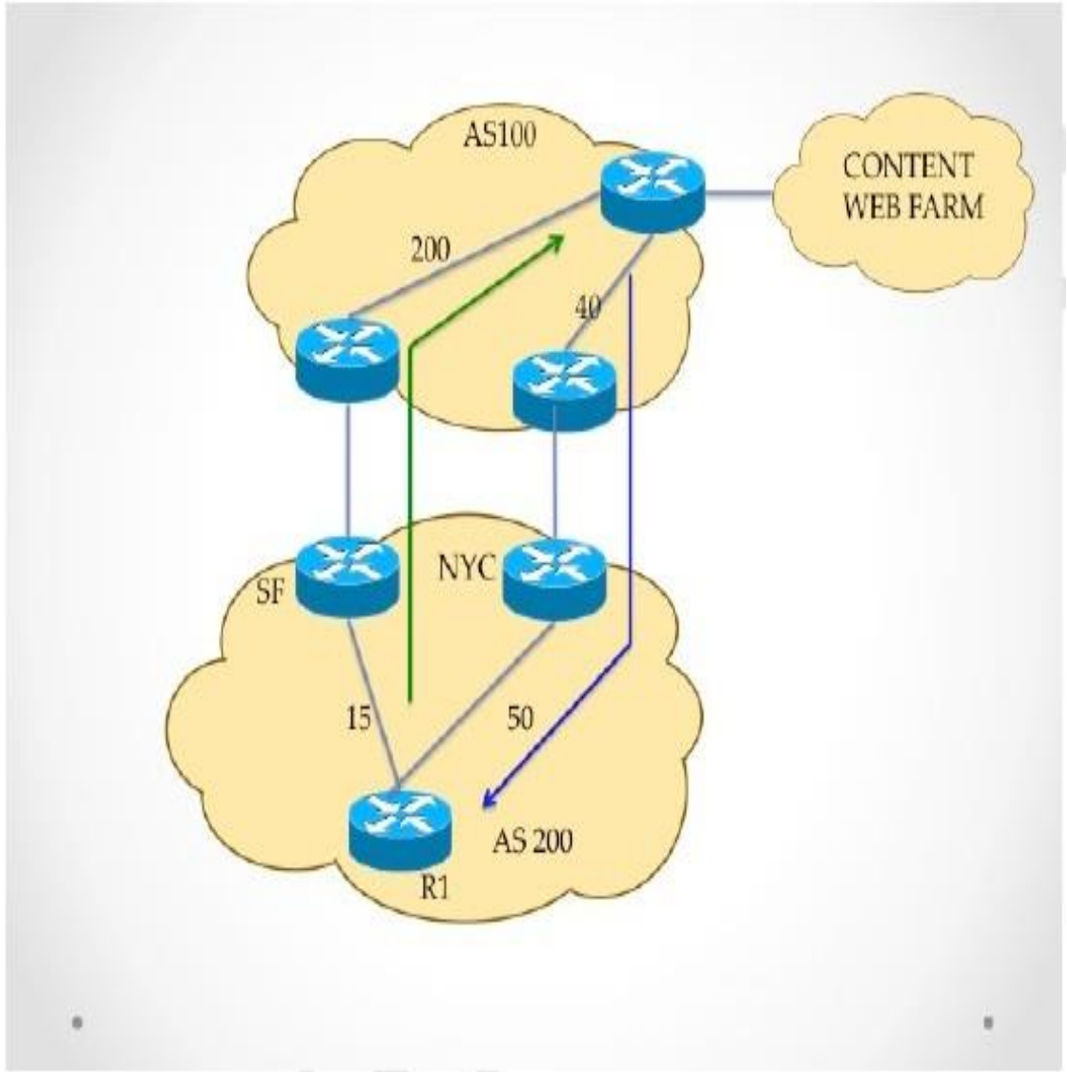
- AS200 is a customer service provider of AS100 transit Service provider. Customers of AS200 is trying to reach a web page located behind AS100. AS200 is not implementing any special BGP policy. What would be the ingress and egress traffic for AS 200 ?



- Topology picture depicts the AS 100 and AS 200 connections. They have a BGP peer (Customer- Transit) relationship on two locations. San Francisco and New York

- IGP distances are shown in the diagram. Since there is no any special BGP policy (Local pref, MED, AS-Path is the same , Origin and so on) , Hot Potato rule will apply so egress path will be chosen from AS 200 and AS100 based on IGP distances

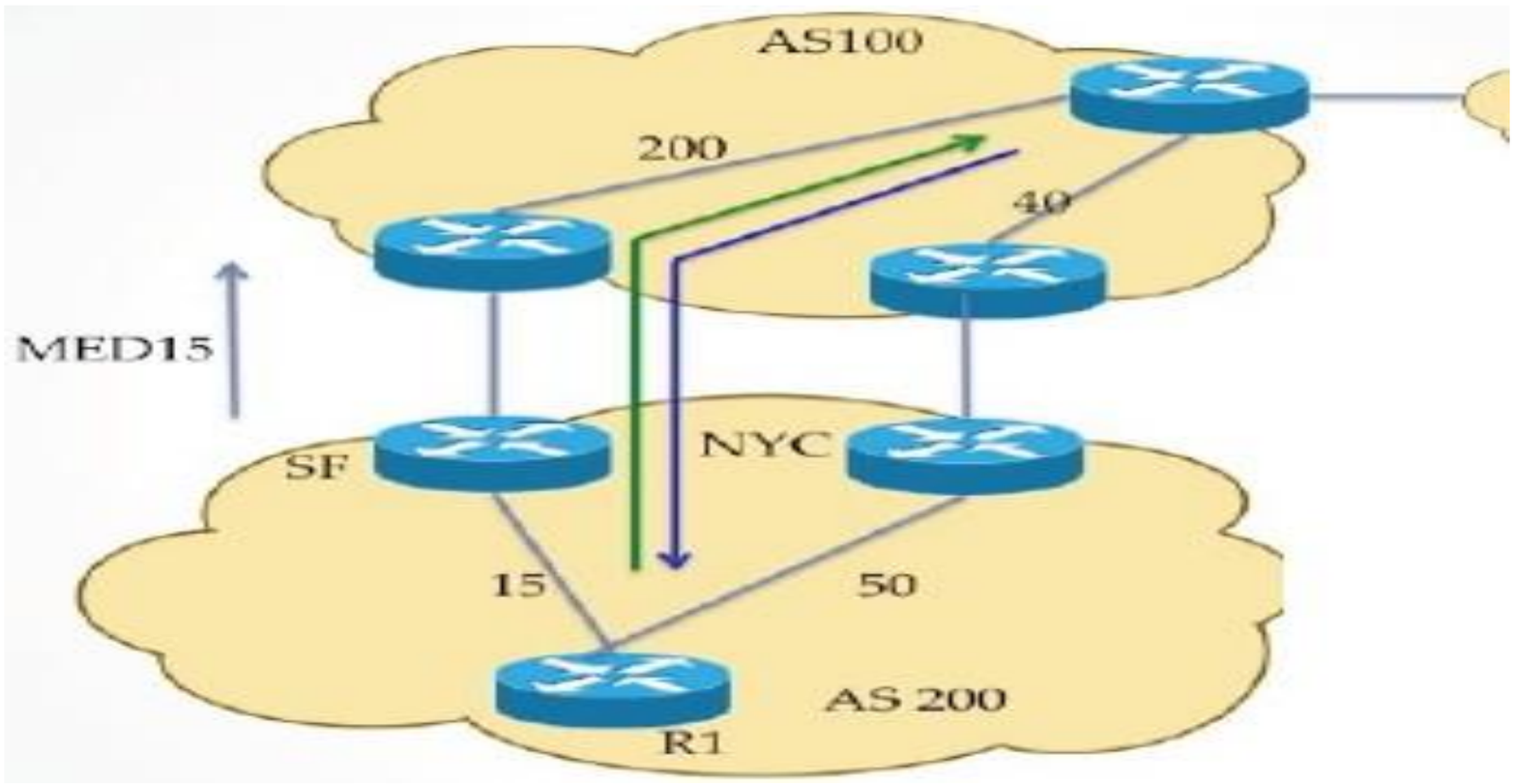
- Egress traffic from AS 200 is the green arrow in the below diagram, since SF path is shorter IGP distance. Ingress traffic to AS200 from AS 100 is the blue arrow, since NYC connection from AS100 shorter IGP distance (40 vs. 200)



- AS 200 is complaining from the performance and they are looking for a solution to fix the above behavior. What would you suggest to AS200 ?

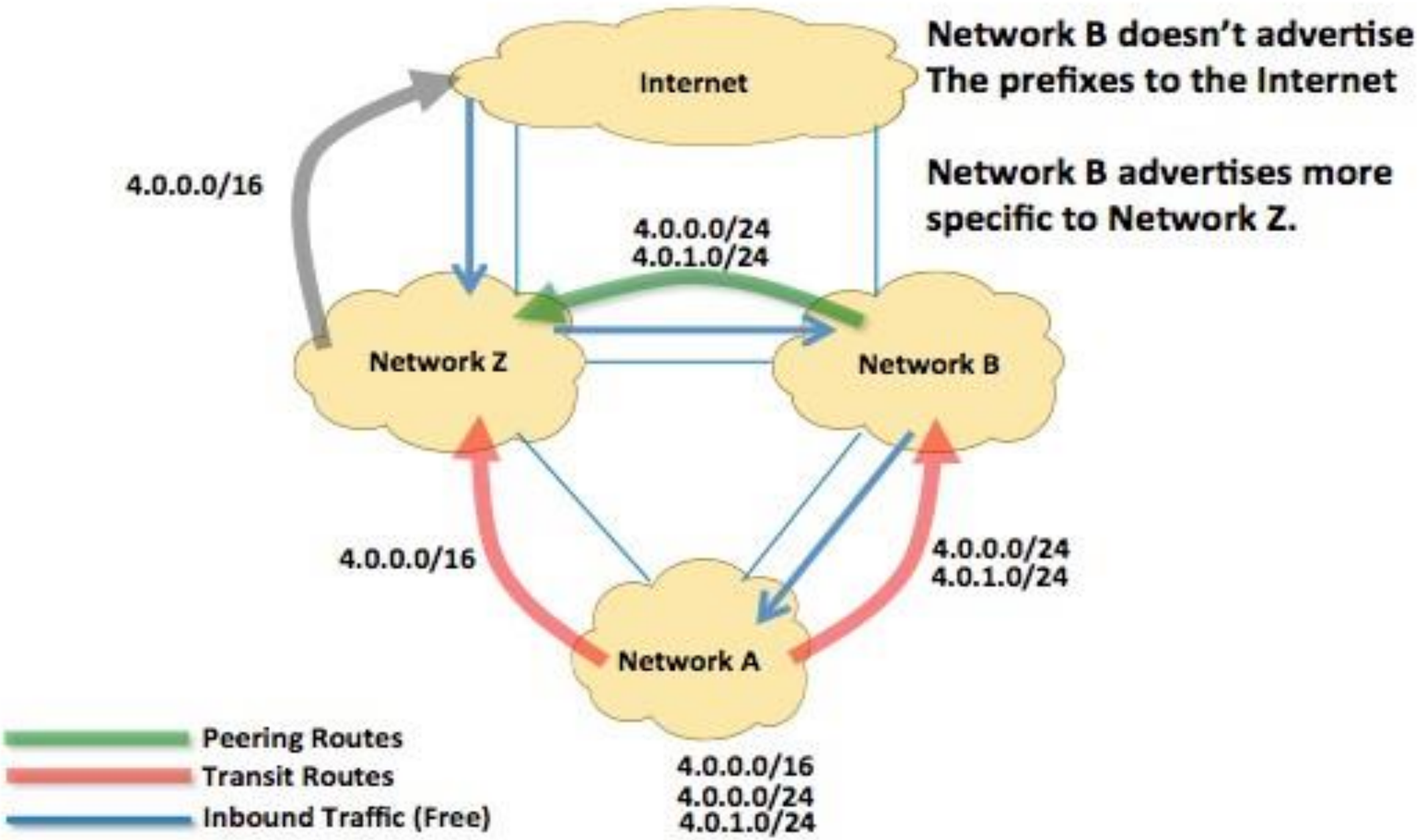
- Customer AS200 should force AS100 for cold potato routing. By forcing for cold potato routing ,AS 100 has to carry the Web content traffic to the closest exit point to AS200, which is San Francisco.

That's why AS200 is sending its prefixes from SF with lower MED than NYC as depicted in the below diagram



How to use Transit service for Free with BGP Jack Move

- Network A is a customer of Network Z, Network B is a peer of Network Z.
- Network A becomes transit customer of Network B.
- Network A announces 4.0.0.0/16 aggregate to Network Z and more specific prefixes,4.0.0.0/24 and 4.0.1.0/24 to Network B. Network B sends more specific to its peer Z.
- Network Z only announces the aggregate to the world. What is the impact of this design ?
- How can it be fixed ?


Network B doesn't advertise
The prefixes to the Internet
Network B advertises more
specific to Network Z.
Internet
4.0.0.0/16
4.0.0.0/24
4.0.1.0/24
Network Z
Network B
4.0.0.0/16
4.0.0.0/24
4.0.1.0/24
Network A
Peering Routes
Transit Routes
Inbound Traffic (Free)
4.0.0.0/16
4.0.0.0/24
4.0.1.0/24

- As it is depicted on the above diagram, Network B doesn’t announce the specific to the world. As a result traffic from internet to Network A goes through Network Z and then through Network B over peer link

- Network A doesn’t have to pay its provider Network Z. This is known as Jack Move. Here Network A and Network pull the Jack Move on network Z

- As we already saw before in the peering section, most if not all networks prefer customer over peer and it is implemented with local preference

- But here customer (Network A) is sending aggregates only to Network Z but more specific routes are coming from Peer network, Network B

- Prefix length overrides the local preference during forwarding

- If Network Z watch for peers advertising more specific of routes for the routes learned from the customers, it is the only way to prevent this

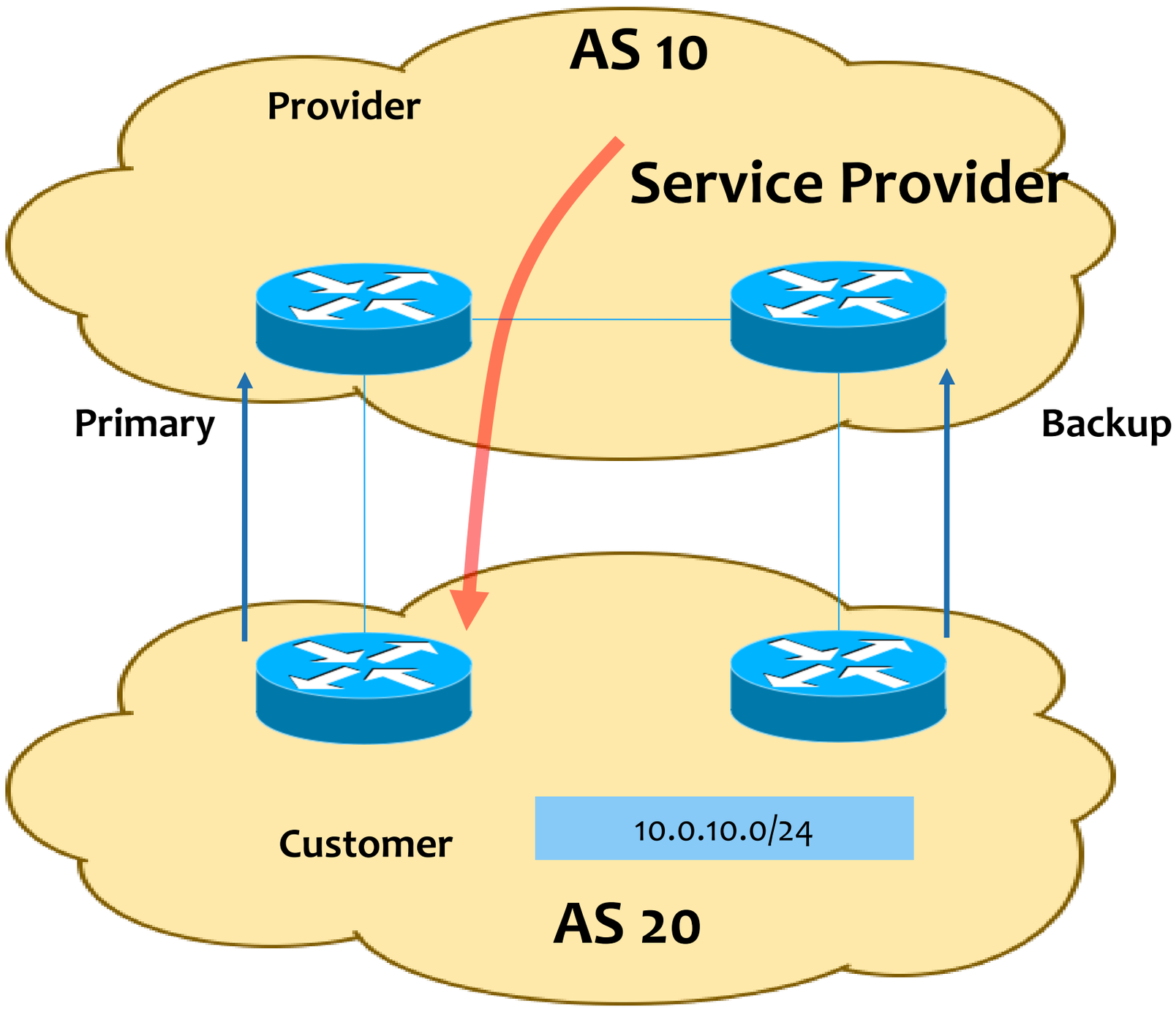
BGP Unintended Behaviors (aka BGP Wedgies)

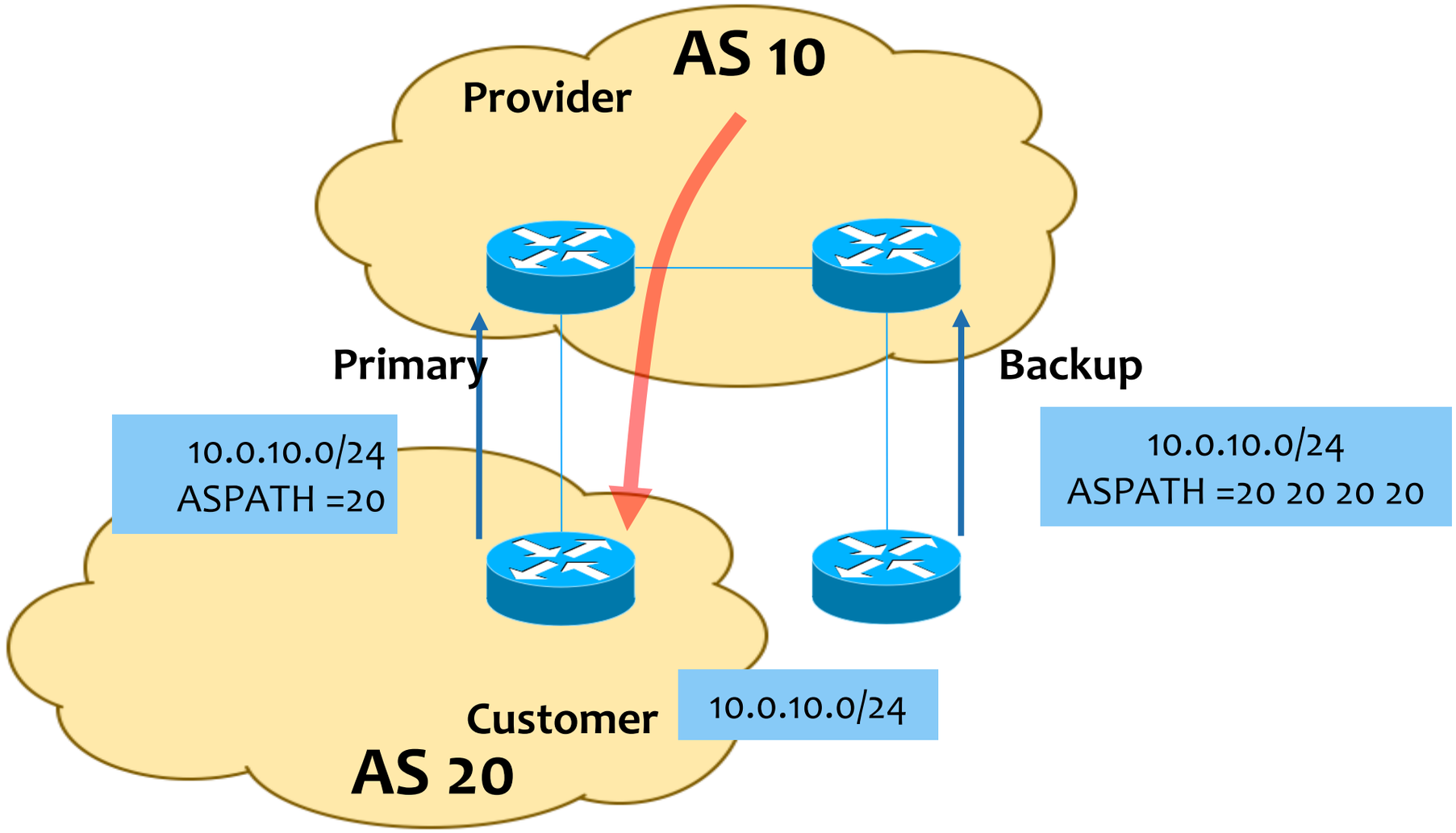
- Customer is running a BGP session with 1 service provider, they are considering to receive a transit service from the second Service Provider as well though

- Customer is using their own AS number which is AS20

- They have 2 connections to their service provider and as it seems in the topology left path will be used as primary for their incoming traffic.

Question 1: How can you achieve this ?

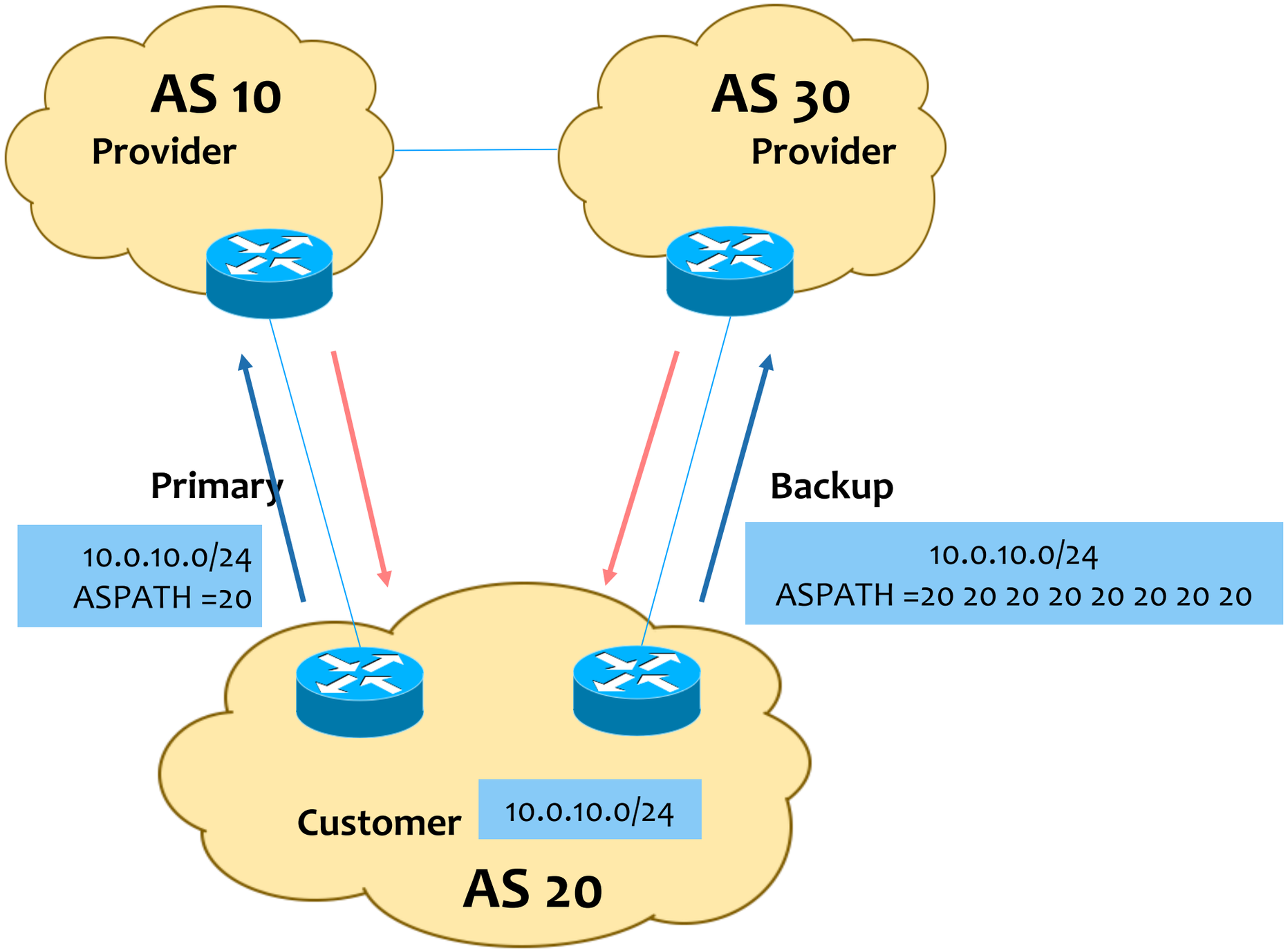

AS 10
Provider
Service Provider
Primary
Backup
Customer
10.0.10.0/24
AS 20



Prepending will (usually) force inbound traffic from AS 10 to take primary link

- Customer purchased a new link from the second service provider which uses AS number 30 and decommissioned one of it’s link from the old service provider

- They want to use second service provider link as backup link. They learned from the early experience the as-path prepending trick

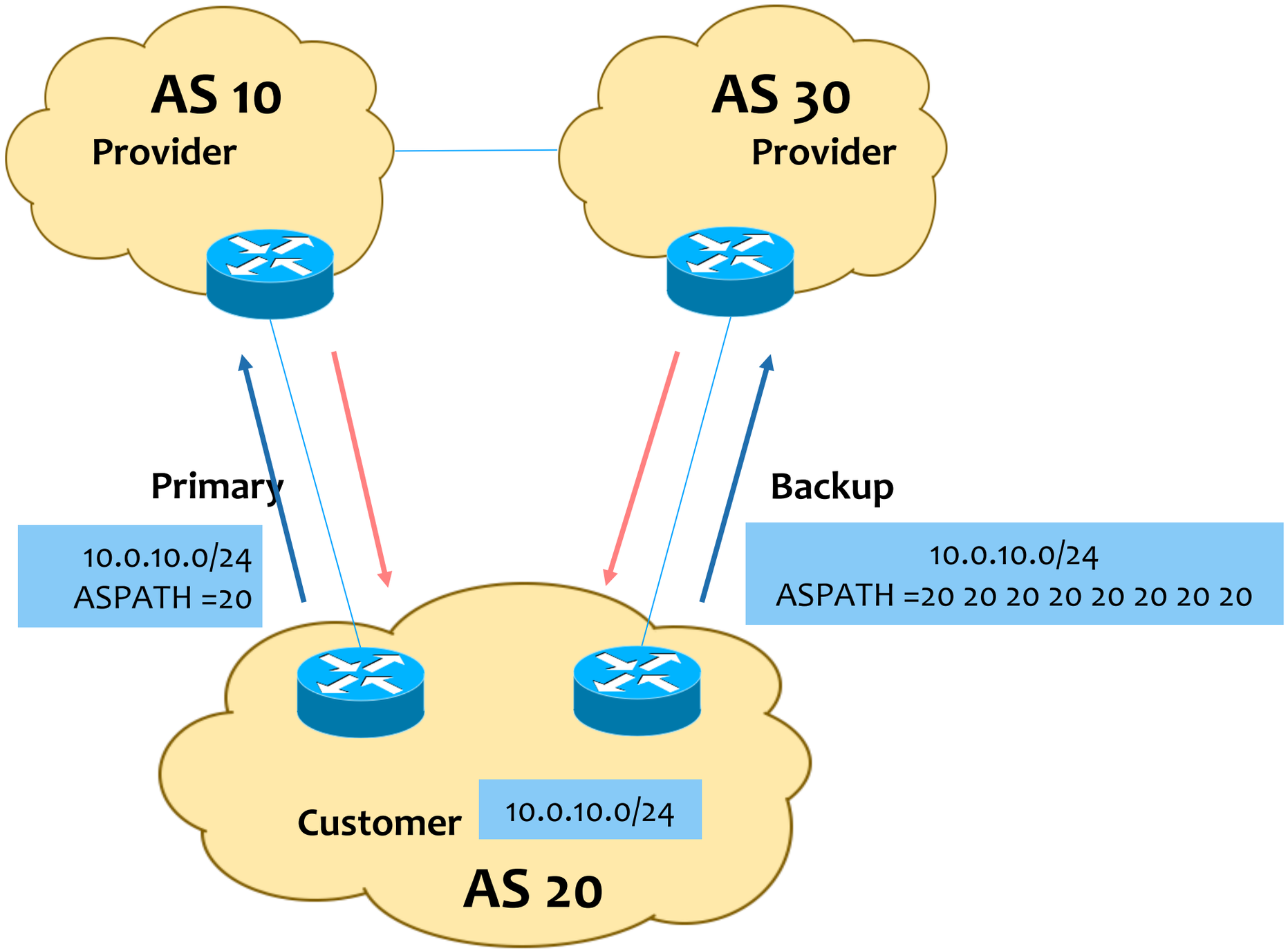

AS 10
Provider
AS 30
Provider
Primary
10.0.10.0/24
ASPATH =20
Backup
10.0.10.0/24
ASPATH =20 20 20 20 20 20 20 20
Customer
10.0.10.0/24
AS 20

- Question 2: Is there a problem with this design ?

- Yes

- No

Question 3 : What is the problem, how you can solve it ?

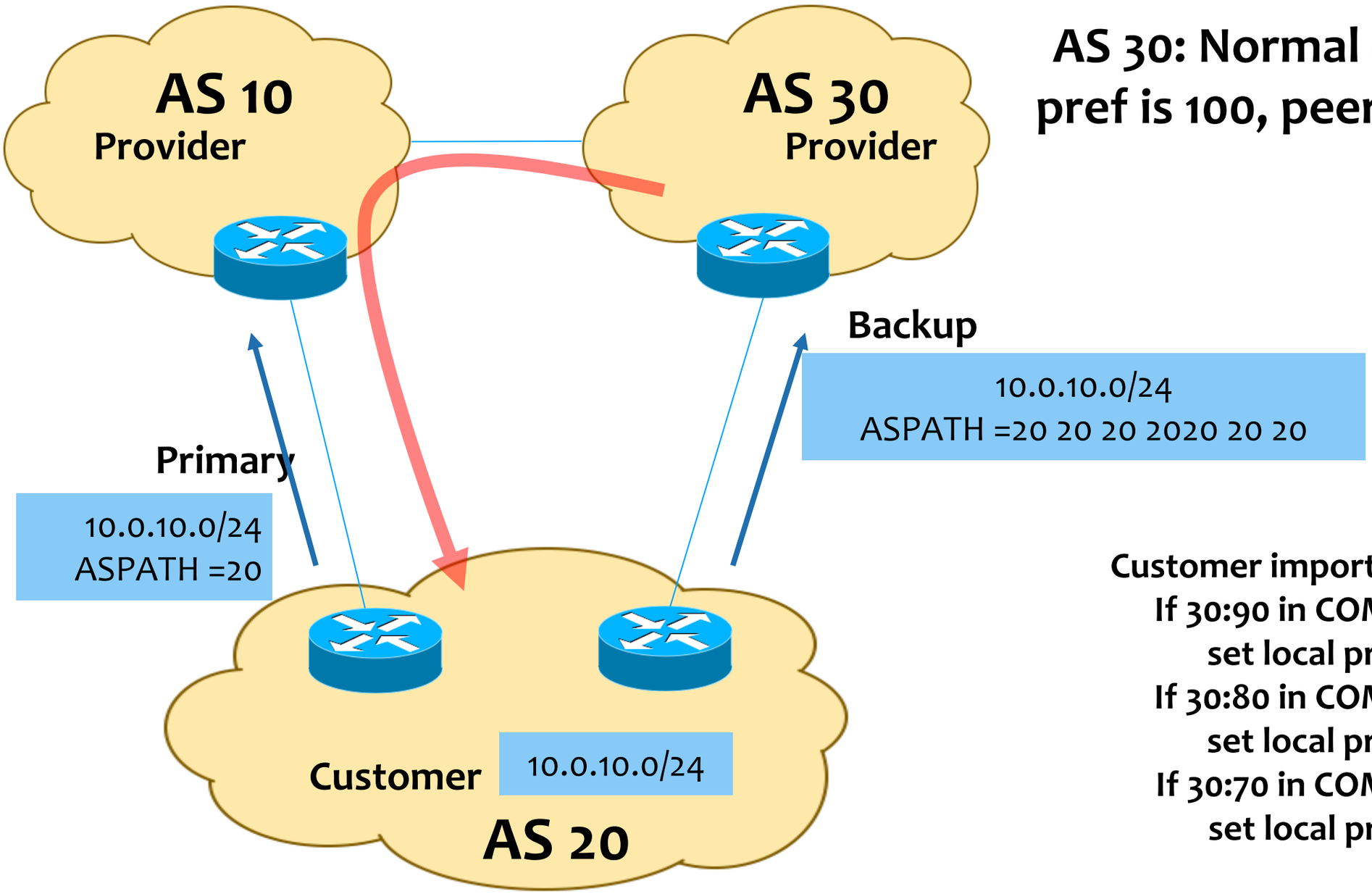

AS 10
Provider
AS 30
Provider
Primary
10.0.10.0/24
ASPATH =20
Backup
10.0.10.0/24
ASPATH =20 20 20 20 20 20 20 20
Customer
10.0.10.0/24
AS 20

- There is a problem with the design since customer wants to use second service provider as backup. AS-Path prepending in this way is often used as a form of load Balancing

- BUT AS 30 will send traffic on “Backup” link, because it prefers customer routes due to higher local preference Service providers use on the customer link than the peer link and local preference is considered before ASPATH Length, so as-path prepending is not effected in this design

- Solution is to use communities

COMMUNITY 30:80 is okay to send the traffic through peer



AS 30: Normal customer local pref is 100, peer local pref is 90

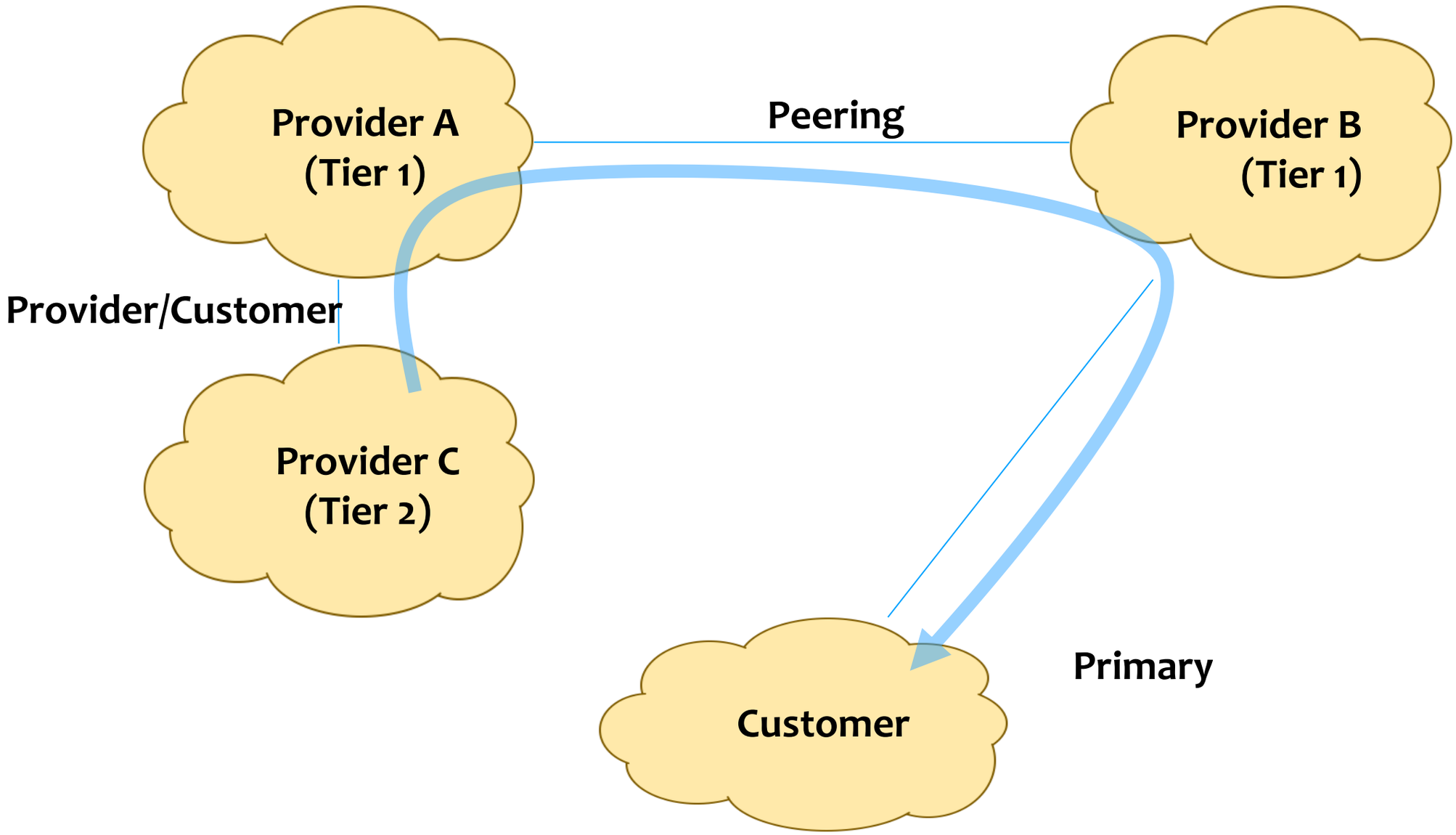
Customer import policy at AS 30:
If 30:90 in COMMUNITY then
set local preference to 90
If 30:80 in COMMUNITY then
set local preference to 80
If 30:70 in COMMUNITY then
set local preference to 70

GAO – REXFORD MODEL

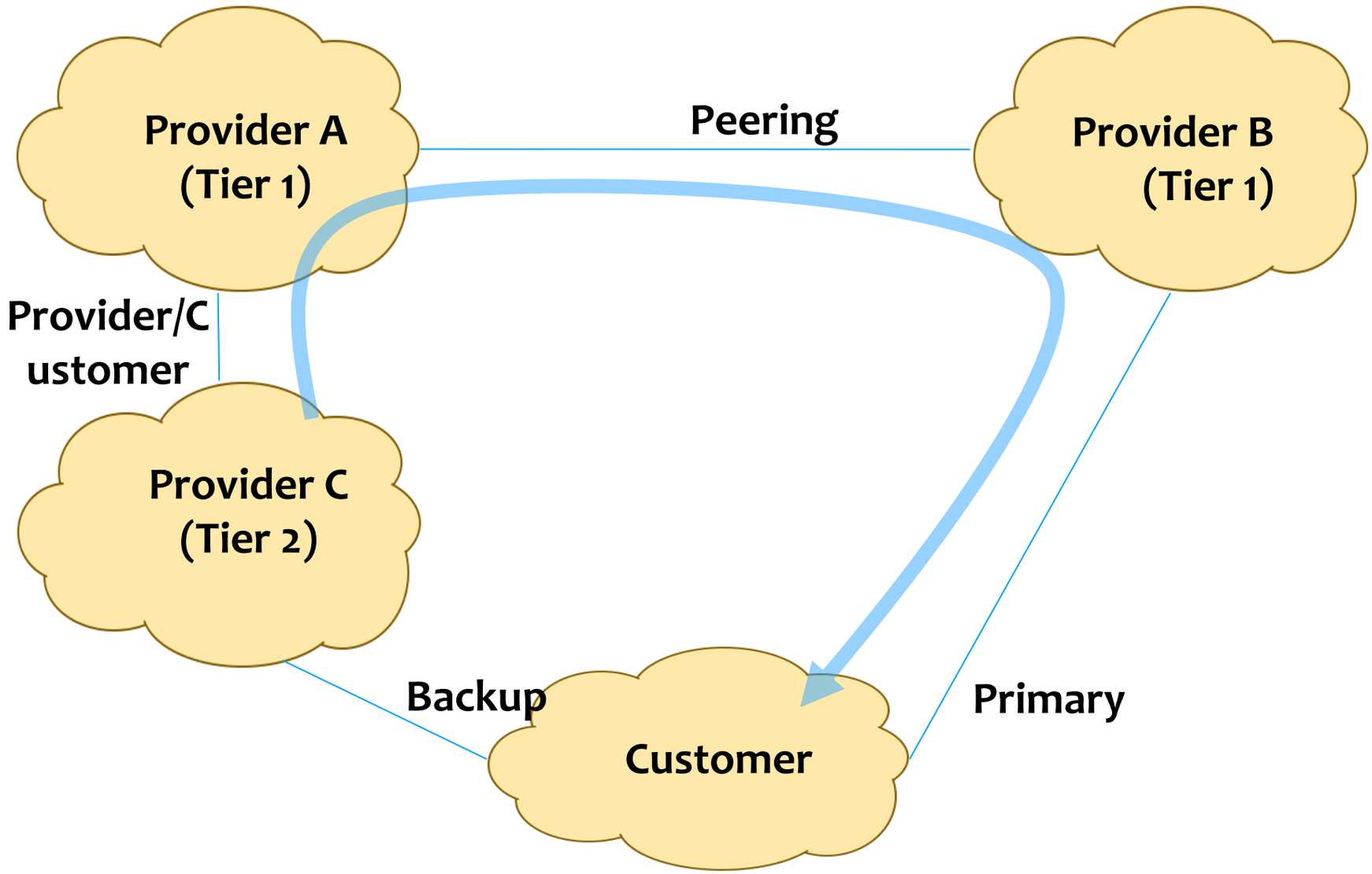
- Question 5: What if customer uses second service provider link as primary and the old provider secondary and the second provider peering connection as depicted in the below topology ?

- Does community help ?

Now, customer wants a backup link to C….

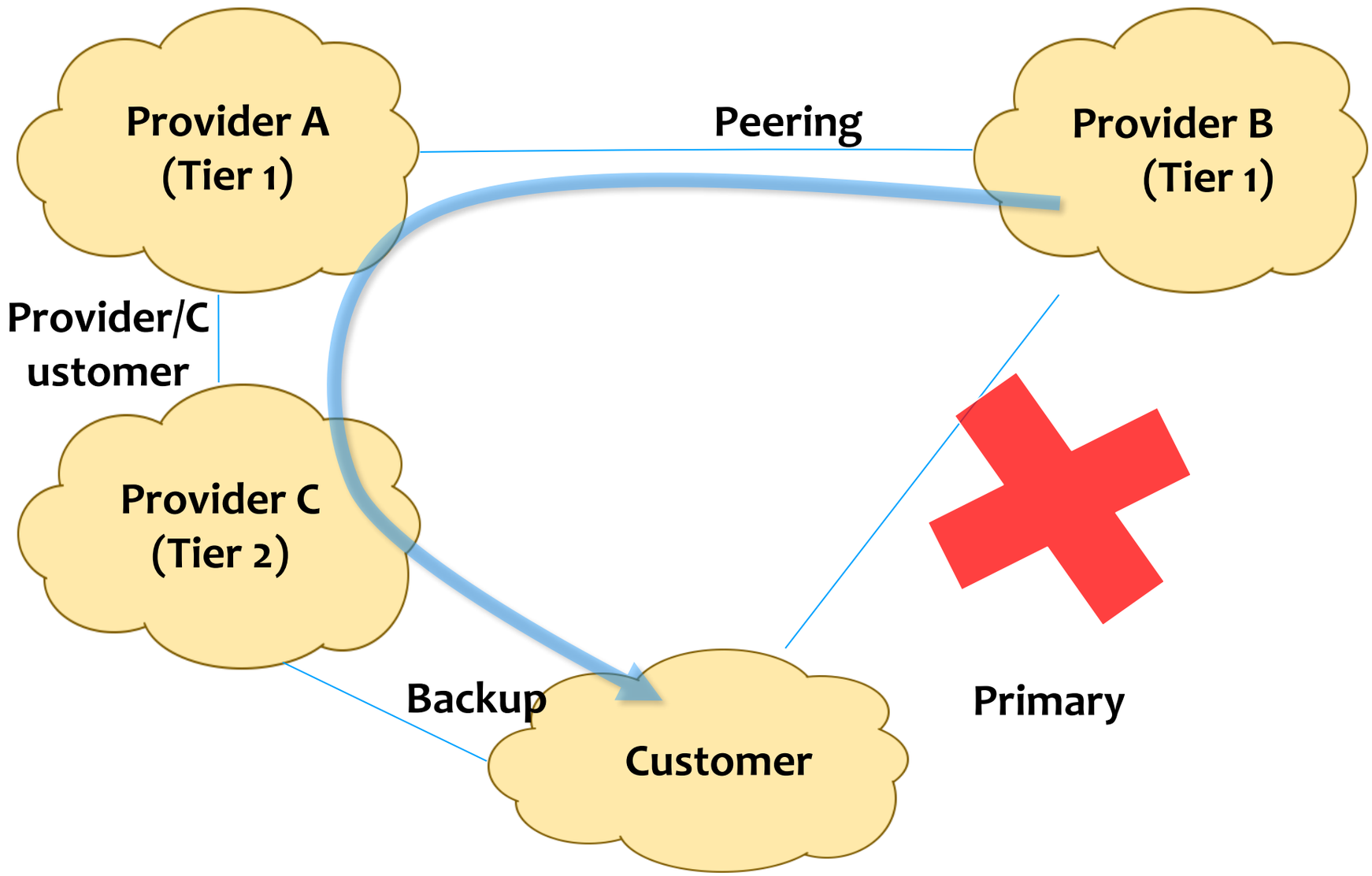


Customer sends community to Provider C, in order to use Provider B as backup



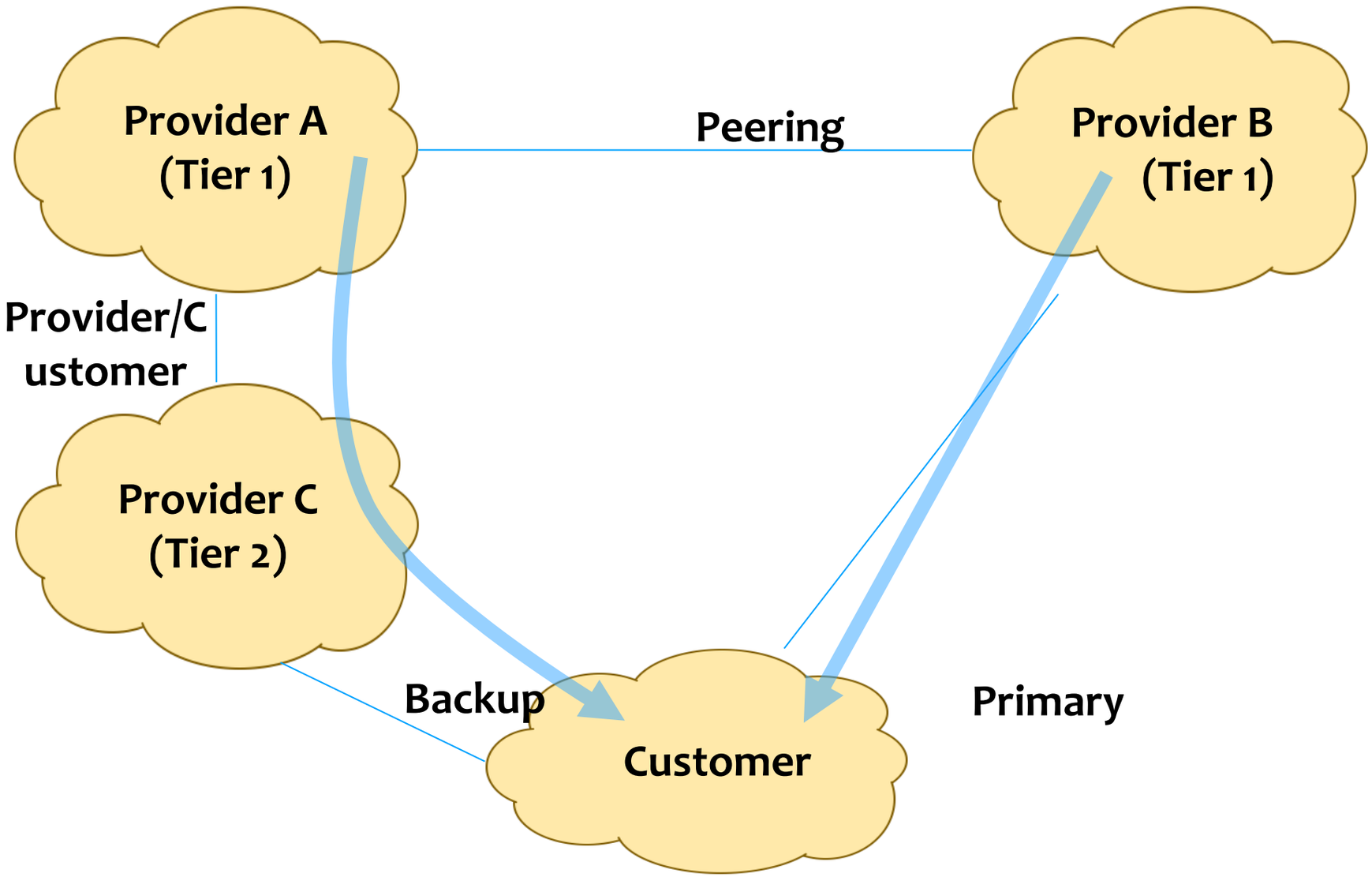
YES IT HELPS , NOW PROVIDER B CAN BE USED AS PRIMARY

- Question 6: What happens if Primary link fails ?



BACKUP LINK IS INSTALLED AND CAN BE USED BY THE CUSTOMER

- Question 7: What happens when the primary link comes back ?

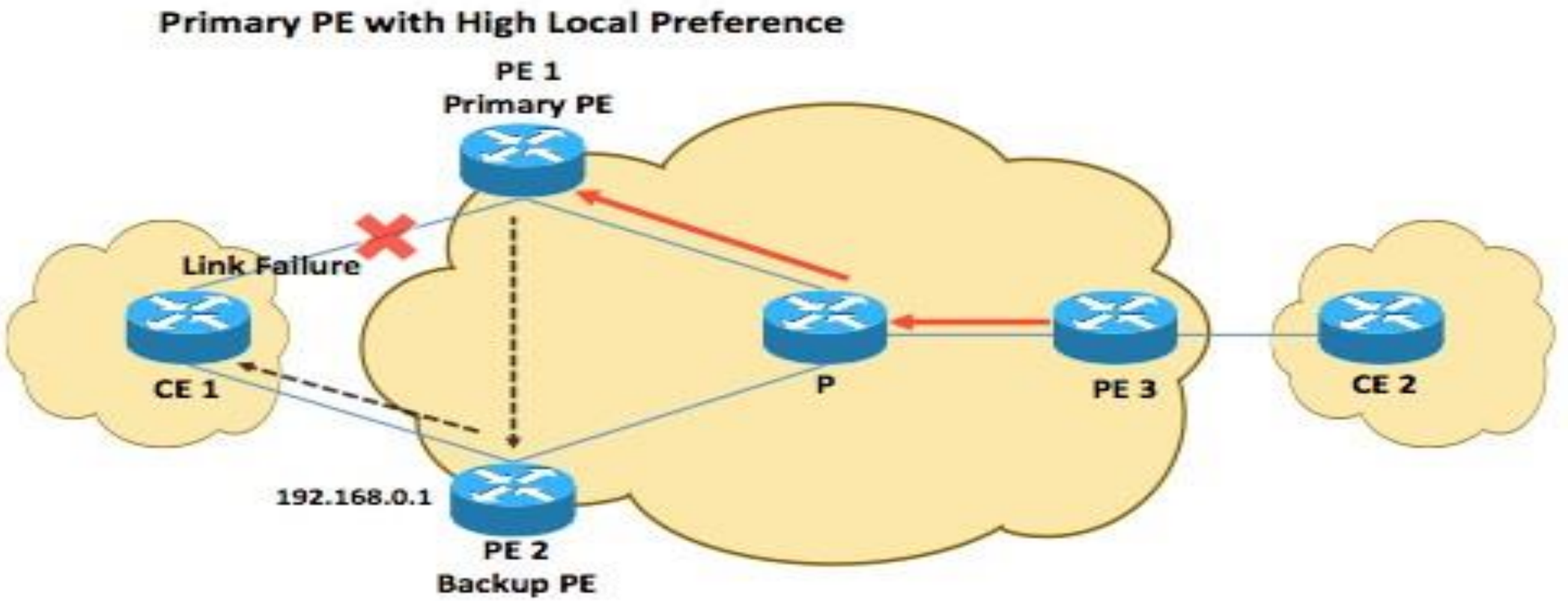

Provider A
(Tier 1)
Peering
Provider B
(Tier 1)
Provider/C
ustomer
Provider C
(Tier 2)
Backup
Primary
Customer

- When primary link comes back, both path is used for the incoming traffic anymore

- Because Provider A continue to choose to send Provider C since the community attribute is sent by Customer to Provider C, not to Provider A

- Solution to fix it, either Provider C will send a Provider A for its customer a community attribute, or Backup BGP link will be resetted when primary link comes back

BGP Best External



Without Best External PE3 cannot learn the standby/backup path.
Best External is used for Active/Standby customer
Link topologies on the Service Provider.
Customer doesn't need to run BGP Best External

In the above picture:

- eBGP sessions exist between the provider edge (PE) and customer edge (CE) routers
- PE1 is the primary router and has a higher local preference setting
- Traffic from CE2 uses PE1 to reach router CE1
- PE1 has two paths to reach CE1
- CE1 is dual-homed with PE1 and PE2
- PE1 is the primary path and PE2 is the backup path

PE1 and PE2 are configured with the BGP Best External feature. BGP computes both the best path (the PE1–CE1 link) and a backup path (PE2) and installs both paths into the RIB and FIB

The best external path (PE2) is advertised to the peer routers, in addition to the best path

BGP Route Reflector Clusters

- Question : Customer wants to use two BGP Route reflector for the redundancy but they don’t know the design best practices whether they should use same or different BGP Route Reflector Cluster ID ? Can you help them ?

- Yes
- No

Should you use different or same Cluster IDs if you have more than one RR in BGP design?

- Almost always use same RR. With different cluster IDs on RR, you will accept and keep the prefixes on the RR. Those prefixes will never be used. But with same cluster ID, prefixes will not be accepted since the ID is the same, this will reduce the resource consumption.

Do you need to install all the prefixes into the RIB and FIB ?

NO ! If RR is in the data path !.

Is there any Exception?

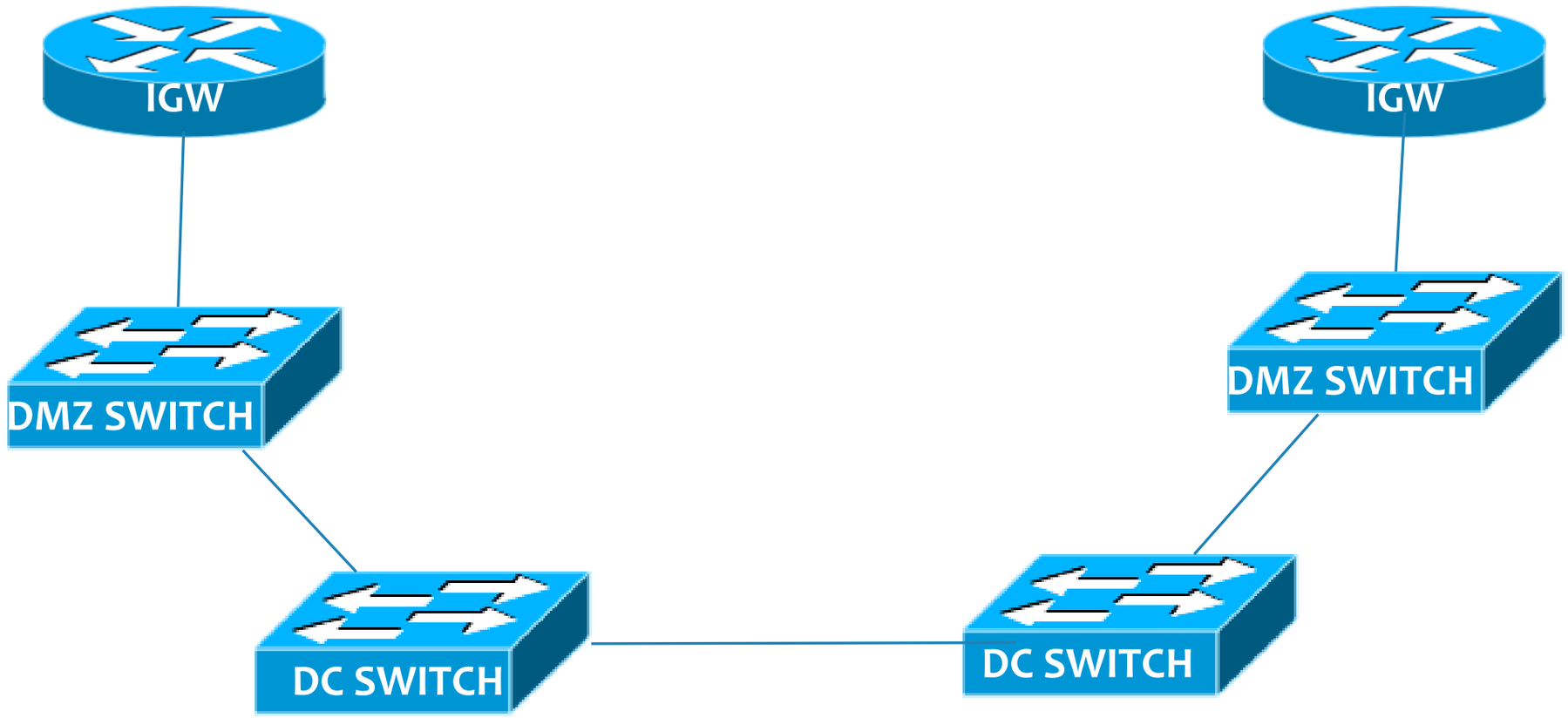
Yes, for example Seamless/Unified MPLS scenario

IBGP over Physical and Tunnel Connections

- Enterprise company has two datacenters. They have 200 remote offices and all the locations access to the internet from the Datacenters

- They recently had an outage on the internet circuit and all the Internet sessions from the remote offices which uses that link wad dropped

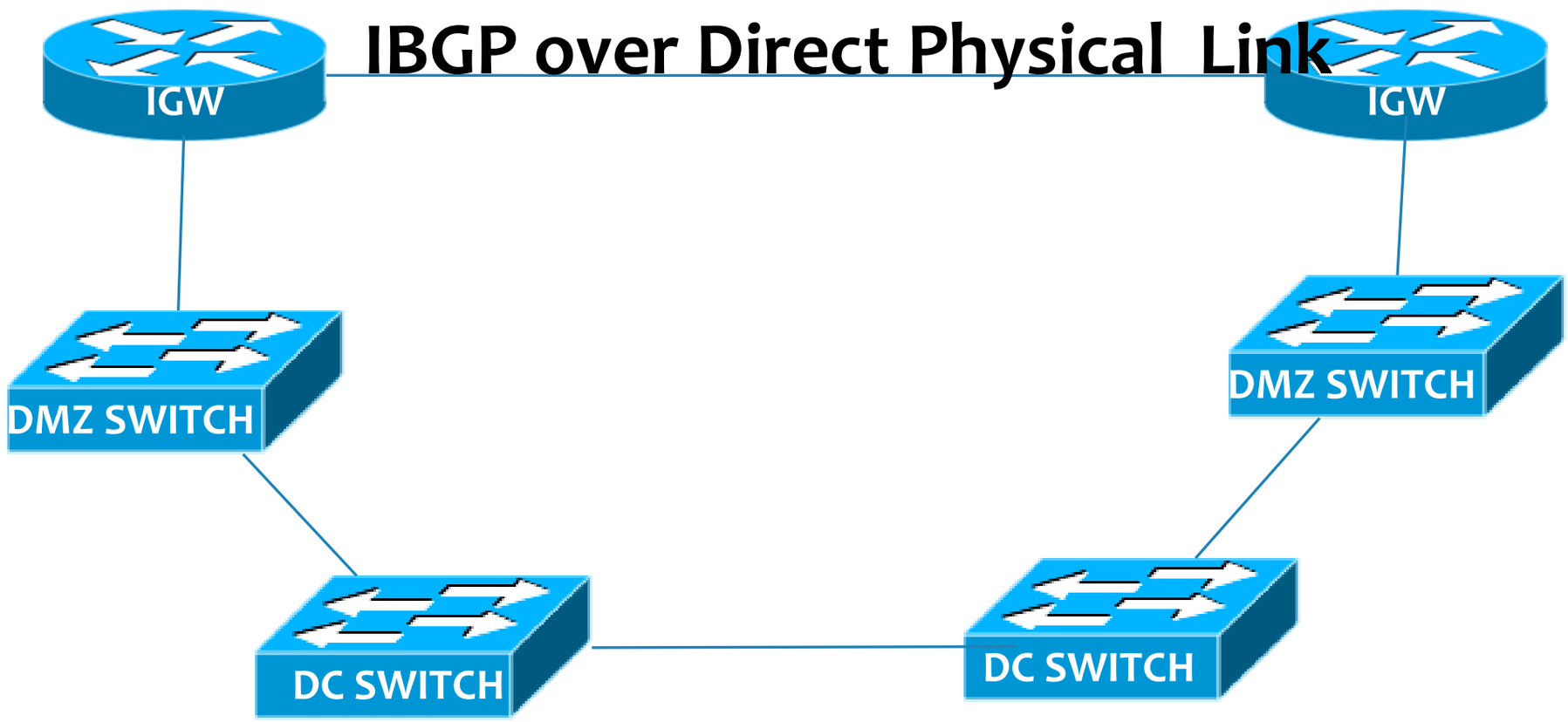
- What are the solutions to prevent the session failure in case of a link failure on the Internet Gateways of this company ?



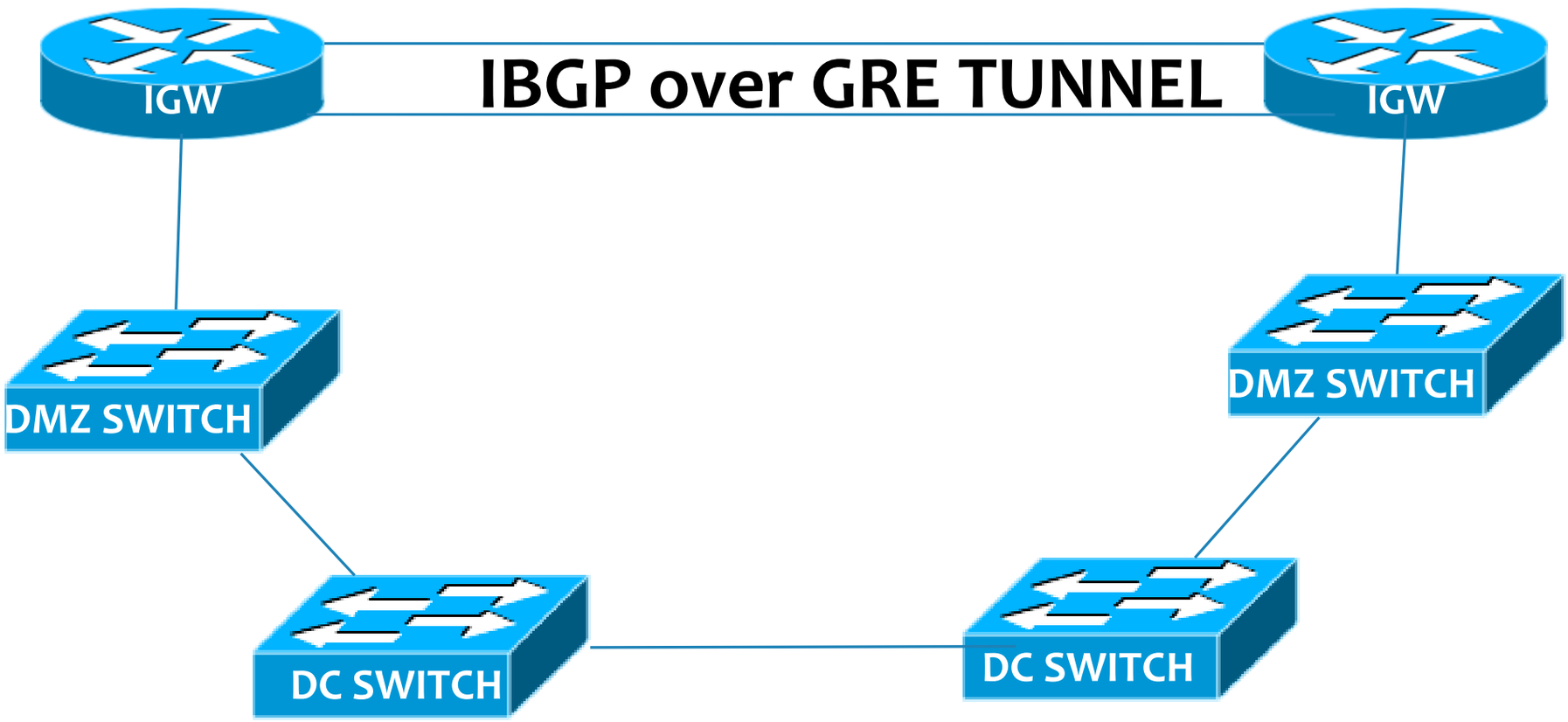
DATACENTER 1

DATACENTER 2

Solution 1 – Best option but can be costly. Budget might be concern, also deployment might take longer compare to other solutions



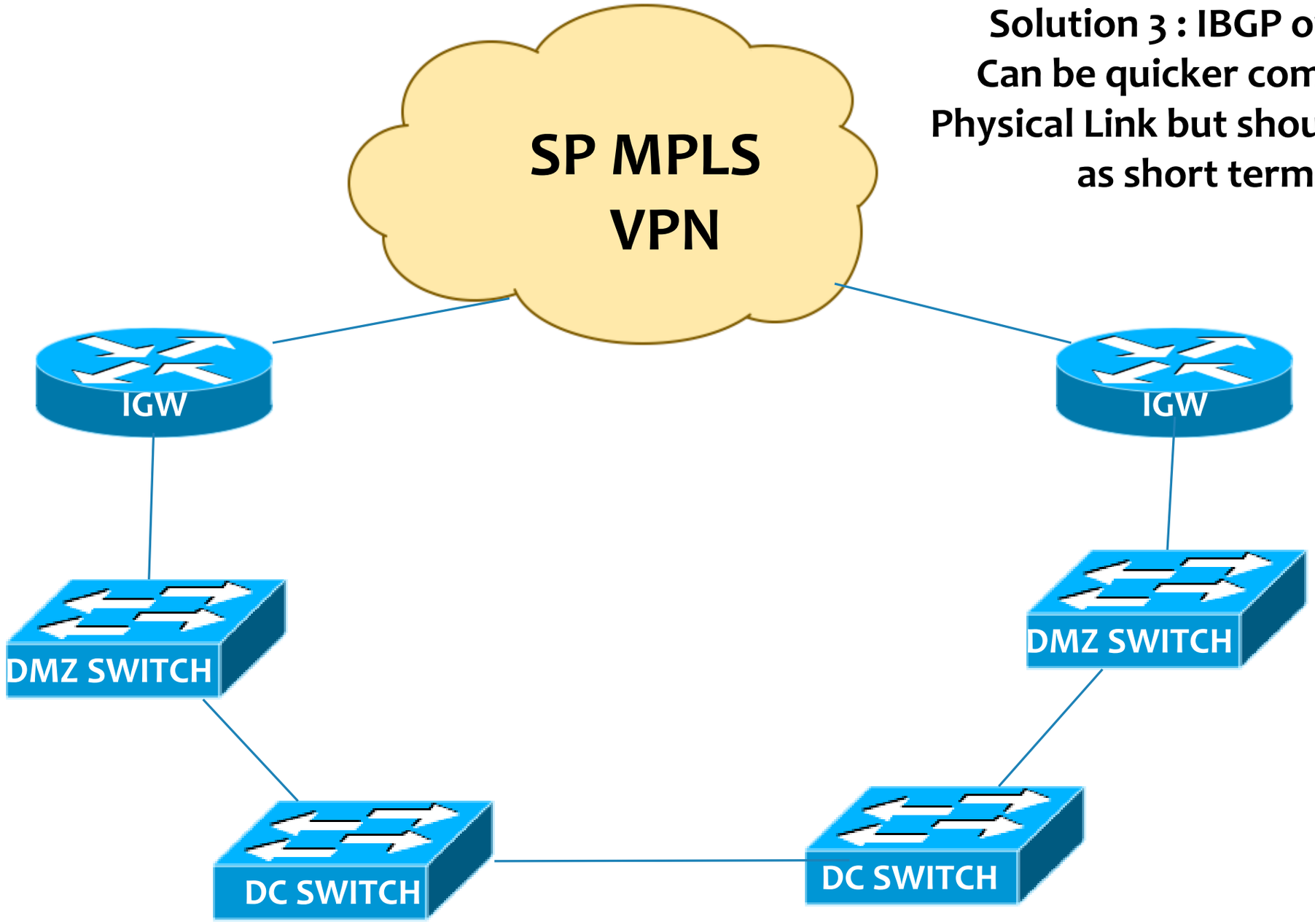
Solution 2 – Fastest option , don't require Service Provider Interaction.
It should be used as a short term solution



DATACENTER 1

DATACENTER 2

Solution 3 : IBGP over MPLS VPN
Can be quicker compare to Direct Physical Link but should be considered as short term solution

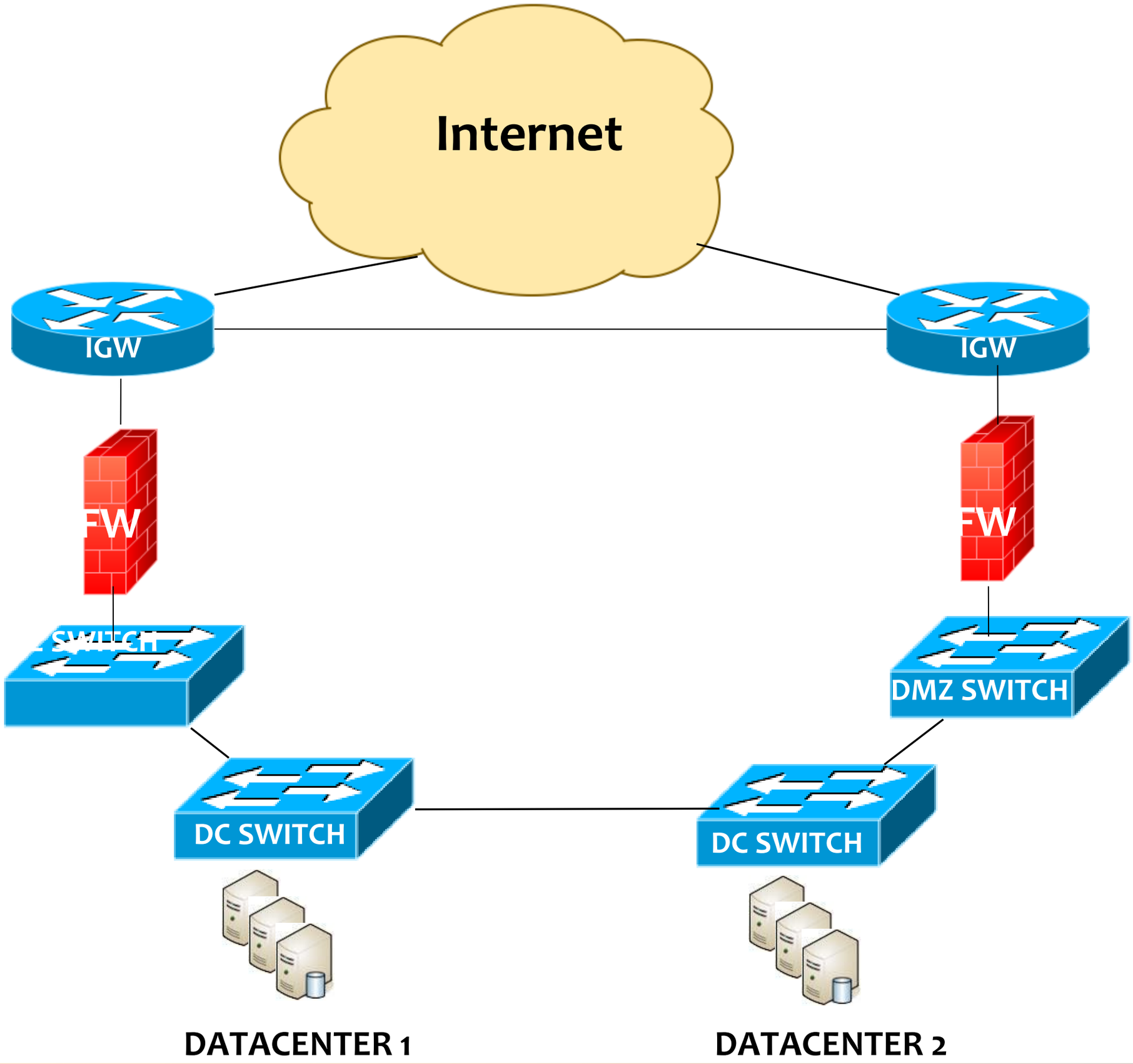


DATACENTER 1

DATACENTER 2

Enterprise Edge Design with BGP

- U.S based Enterprise e-commerce company is designing a new network. They have two datacenters and both datacenters will host many servers.
- There are 1000km between the two datacenters.
- Their networking team knows that for the best user performance traffic should be symmetric between servers and the users/clients.
- In addition to datacenter interconnect link they have direct physical connection
- Based on the below topology what might be the issue ?


Internet
IGW
IGW
FW
FW
SWITCH
DMZ SWITCH
DC SWITCH
DC SWITCH
DATACENTER 1
DATACENTER 2

- It is already given in the requirements that traffic from DC1 should come back to DC1 directly. Asymmetric traffic cause firewall to drop all the traffic

- So if the users are accessing to DC1 servers it should go back from the DC1. Classical design for this, servers uses DC switch as default gateway. DC switches receive default route redistributed to their IGP from BGP by the IGW. And IGP cost is used to reach to the closest IGW by the DC switches

- Incoming traffic can be a problem whenever there is a stateful device in the path

- In the above topology if traffic comes to DC1 it has to go back from DC1 and vice versa, it is not only for asymmetric flow on the Firewalls, Load Balancers and so on but also important to avoid hairpin

- If traffic destined to DC1 comes to DC2, it has to go through direct physical internet link to the DC1, this adds additional latency and consume unnecessary bandwidth

- Question 2: How does company achieve symmetric traffic flow so they don’t have any traffic drop or performance issue ?

- They can split their public IP space to half and advertise specifics from each datacenters and summary from both datacenters as a backup in case first DC IGW link or node fails

- Imagine they have /23 address space, they can dive 2x/24 and advertise each /24 from local datacenters only and /23 from both datacenters. Since their upstream SP will profer longest match routing over any other BGP attribute, traffic returns to the location where it is originated

BGP vs. IGP Comparison

orhanergun.net	OSPF	IS-IS	EIGRP	BGP
Scalablability	2 tier hierarchy , less scalable	2 tiers hierarchy , less scalable	Support many tiers and scalable	Most scalable routing protocol
Working on Full Mesh	Works well with mesh group	Works well with mesh group	Works very poorly, and there is no mesh group	Works very poorly, but RR removes the requirement
Working on a Ring Topology	Its okay	Its okay	Not good if ring is big due to query domain	Good with Route Reflector
Working on Hub and Spoke	Works poorly, require a lot of tuning	Works bad requires tuning	Works very well. It requires minimum tuning	IBGP works very well with Route Reflector
Fast Reroute Support	Yes - IP FRR	Yes - IP FRR	Yes - IP FRR and Feasible Successor	Requires BGP PIC + NHT + Best external + Add-Path
Suitable on WAN	Yes	Yes	Yes	Yes, but in very large scale or when policy is needed
Suitable on Datacenter	DCs are full mesh. So, No	DCs are full mesh so No	DCs are full mesh so no	Yes, in large scale DC and it is not uncommon
Suitable on Internet Edge	No it is designed as an IGP	No it is designed as an IGP	No, it is designed as an IGP	Yes, it is designed to be an Inter domain protocol
Standard Protocol	Yes IETF Standard	Yes IETF Standard	No, there is a draft but lack of Stub feature	Yes, IETF Standar
Stuff Experince	Very well known	Not well known	Well known	Not well known
Overlay Tunnel Support	Yes	Doesn't support IP tunnels	Yes	Yes
MPLS Traffic Engineering Support	Yes with CSPF	Yes, with CSPF	No	No
Security	Less secure	More secure since it is on layer2	Less secure	Secure since it runs on TCP
Suitable as Enterprise IGP	Yes	No, it lacks Ipsec	Yes	Not exactly, very large scale networks only
Suitable as Service Provider IGP	Yes	Definitely	No, it doesn't support Traffic Engineering	Maybe in the datacenter but not as an IGP
Complexity	Easy	Easy	Easy	Complex
Policy Support	Good	Good	Not so Good	Very good
Resource Requirement	SPF requires more processing power	SPF requires more processing power	DUAL doesn't need much power	Requires a lot of RAM and decent CPU
Extendibility	Not good	Good, thanks to TLV support	Good, thanks to TLV support	Very good, it supports 20 + address families
IPv6 Support	Yes	Yes	Yes	Yes
Default Convergece	Slow	Slow	Fast with Feasible Successor	Very slow
Training Cost	Cheap	Cheap	Cheap	Moderate
Troubleshooting	Easy	Very easy	Easy	Moderate
Routing Loop	Good protection	Good protection	Open to race condition	Good protection

BGP in the CCDE Exam

- IBGP will be main focus for the CCDE Exam
- BGP Route Reflectors and the different BGP Route Reflector Design Options
- BGP Confederation and using multiple Sub AS for the multinational providers and so on.
- Multi Protocol BGP should be understood very well

BGP in the CCDE Exam

- Different NLRI, especially IPv6, IPv6 and VPN address families with BGP should be understood

- BGP AS Migration needs to be understood, changing one AS with another one , maybe in a merger or acquisition scenario with Local AS feature

- BGP Traffic Engineering with BGP Path Attributes. Local Pref, MED, As-Path Prepend and Community etc.

BGP Summary

- BGP Use Cases , Theory
- EBGP Loop Prevention , EBGP Traffic engineering , BGP Path Attributes
- EBGP – Inter-domain routing, IXP , Peering , IP Transit Business
- Service Provider Business models – GAO- Rexford model and default SP BGP policy

BGP Summary

- IBGP , Full Mesh IBGP, Route Reflectors, Confederation and their design options

- BGP Add-Path , Shadow RR , Unique RD per VRF per PE

- BGP Egress Peer Engineering

- BGP RTBH and BGP Flowspec

BGP Summary

- BGP Information Security , Route Leaks , BGP Hijacks, Route Origin Validation and Path Validation
- IRR, Peeringdb, RPKI and BGPSEC
- BGP in the Datacenter – IBGP vs. EBGP , EBGP ASN Allocation, BGP Path Hunting
- BGP PIC – Flat , Hierarchical and Generalized FIB Architecture , PIC Edge and PIC Core
- BGP vs. IGP Comparison

BGP Quiz !!

Questions and the Answers

BGP Quiz - Question 1

Which of the below option is the reason to run IBGP? (Choose Two)

A. It is used for the reachability between PE devices in MPLS network

B. It is used to carry EBGP prefixes inside an Autonomous System

C. It is used with Route Reflectors for the scalability reason in large scale networks

D. It is used to prevent failures outside your network from impacting your internal network operation

BGP Quiz – Answer 1

One of the correct answers of this question is to carry EBGP prefixes inside an Autonomous system.

IGP is used for the reachability between PE devices in an MPLS network.
Option C is valid but not the correct answer, because; question is asking the reasons, not the best practices.

Option D is one of the correct answers as well because with IBGP, internal network is protected from the outside failures by separating the local failure domains.
That’s why; answers of this question are B and D.

BGP Quiz - Question 2

Which of the below options are true for the BGP Route Reflectors? (Choose Three)

A. Route Reflectors provide scalability in large scale network design

B. Route Reflectors hide the available paths

C. Route Reflectors selects and advertise only the best path to Route Reflector clients

D. Route Reflectors can be placed anywhere in the IP backbone as an IPv4 RR

BGP Quiz - Answer 2

Route reflectors are used to improve scalability of the BGP design in large-scale deployments.

Route reflectors hide the available path information by selecting and advertising only the best path to the clients.

Thus the correct answer of this question is A, B and C.

Option D is wrong because, Route Reflectors should follow the physical topology in an IP backbone, it cannot be placed everywhere, careful planning is required. Otherwise forwarding loop occurs as it was explained in one of the case studies in the BGP chapter.

BGP Quiz - Question 3

Which below attributes are commonly used for BGP path manipulation? (Choose Three)

A. Local Preference

B. Origin

C. As-Path

D. Community

E. Weight

BGP Quiz – Answer 3

Origin is not used commonly for the BGP path manipulation. Weight is Cisco preparatory and it is only local to the routers. It shouldn’t be used for path manipulation.

BGP path manipulation was explained in detail in BGP chapter.

Answer of this question is A, C and D.

BGP Quiz - Question 4

Which of the below options is used in the Public Internet Exchange Points to reduce configuration overhead on the BGP devices?

A. BGP Route Reflectors

B. BGP Prefix Lists

C. BGP Route Servers

D. BGP Map Servers

BGP Quiz – Answer 4

There is nothing called BGP Map Servers. In the Public Internet Exchange points BGP Route Servers are used to reduce configuration overhead.

They improve scalability. Very similar to Route Reflectors but Route Reflectors are used in IBGP, not in the Public Exchange Points. That’s why answer of this question is C.

BGP Quiz - Question 5

Which below options are true for the BGP Confederation? (Choose Three)

A. It is done by creating Sub-Autonomous system

B. It is easier to migrate from full-mesh IBGP, compare to BGP Route Reflectors

C. Between Sub Autonomous Systems mostly EBGP rules apply

D. Compare to BGP Route Reflector design, it is less commonly deployed in the networks

BGP Quiz - Answer 5

From the migration point of view, Full mesh IBGP to BGP Confederation is harder, compare to BGP Route Reflectors. Thus Option B is invalid.

All the other options are correct thus the answer of this question is A, C and D

BGP Quiz - Question 6

Which below option is used for inbound BGP path manipulation? (Choose Three)

A. Local Preference

B. MED

C. As-Path prepending

D. Community

E. Hot Potato Routing

BGP Quiz - Answer 6

Hot Potato Routing and Local Preference are used for Outbound BGP Path manipulation as explained in the BGP chapter in detail.

MED should be used if there is only one upstream ISP but still it is used for inbound path manipulation. AS-Path prepending and the communities are used for the multihoming connections as well.

That’s why; answer of this question is B, C and D.

BGP Quiz - Question 7

What does MP-BGP (Multi Protocol BGP) mean?

A. BGP implementation which can converge less than a second

B. BGP implementation which is used in Service Provider networks

C. BGP implementation which can carry multiple BGP Address Families

D. BGP implementation which is used in Enterprise Networks

BGP Quiz - Answer 7

MP-BGP (Multi Protocol BGP) is the BGP implementation, which can carry multiple Address Families. BGP in 2016, can carry more than 20 different Address Families such as IPv4 Unicast, IPv6 Unicast, IPv4 Multicast, L2 VPN, L3VPN, Flowspec and so on.

That’s; why; answer of this question is C.

BGP Quiz - Question 8

- **What does Hot Potato Routing mean?**

A. Sending the traffic to the most optimum exit for the neighboring AS

B. Sending the traffic to the closest exit to the neighboring AS

C. By coordinating with the neighboring AS, sending traffic to the closest exit point

D. It is the other name of BGP Multipath

BGP Quiz - Answer 8

Hot Potato Routing means, sending the traffic to the closest exist point from the Local Autonomous system to the neighboring Autonomous System by taking the IGP metric into consideration

There is no coordination between the Autonomous System in Hot Potato Routing definition. But Coordination with the Hot Potato Routing case study was provided in the BGP chapter.

That’s why; answer of this question is B

BGP Quiz - Question 9

Fictitious Service Provider is considering providing an availability SLA for their MPLS VPN customers. They want to provide sub second convergence in case link or node failure scenarios.
What would you suggest to this company to achieve their goal? (Choose Two)

A. Implementing BFD

B. Implementing BGP PIC Core and Edge

C. Implementing BGP Route Reflectors

D. Implementing IGP FRR

BGP Quiz - Answer 9

They should implement BGP PIC features to protect BGP from the link or node failure. Especially Edge node failures, even if MPLS Traffic Engineering or IP FRR deployed, couldn’t be recovered in sub second.

Since BGP PIC convergence is mostly depends on IGP convergence as well, deploying IGP FRR (Fat Reroute) provides a necessary infrastructure for the BGP PIC.

They should be deployed together. BFD is just a failure detection mechanism. IGP Convergence is depends on many other parameters tuning

That’s why; answer of this question is B and D

BGP Quiz - Question 10

- **With which below options, internal BGP speaker can receive more than one best path even if BGP Route Reflectors are deployed? (Choose Three)**

A. BGP Shadow RR

B. BGP Shadow Sessions

C. BGP Add-path

D. BGP Confederation

E. BGP Multipath

BGP Quiz - Answer 10

Shadow Sessions, Shadow RR and BGP Add-path design provides more than best path to the internal BGP speaker even if BGP Route Reflectors are deployed.

BGP Multipath requires more than one best path and all the path attributes to be the same. Thus it requires one of the above mechanisms. BGP Confederation doesn’t provide this functionality.

That’s why; answer of this question is A, B and C.

BGP Quiz - Question 11

Which below option is recommended to send more than one best path to the VPN PEs in the MPLS VPN deployment if VPN Route Reflectors are deployed?

A. BGP Add-path

B. BGP Shadow RR

C. BGP Full Mesh

D. Unique RD per VRF per PE

BGP Quiz - Answer 11

BGP Add-path, BGP Shadow RR and Sessions deployments are suitable for the IP backbones.

If there is an MPLS backbone, configuring unique RD per VRF per PE is best and recommended design option since there is no software or hardware upgrade, no additional BGP sessions and so on.

That’s why the answer of this question is D

BGP Quiz - Question 12

What are the reasons to send more than one BGP best path in IP and MPLS deployment? (Choose Four)

A. BGP Multipath

B. BGP Fast Reroute

C. BGP Multihop

D. Preventing Routing Oscillation

E. Optimal BGP routing

BGP Quiz - Answer 12

There are many reasons to send more than one BGP best path in both IP and MPLS deployments.

These are; avoiding routing oscillations, BGP Multipathing, Fast convergence/Fast Reroute and Optimal Routing.

Sometimes for the optimal routing, just sending more than one BGP best path is not enough but may require all available paths though.

That’s why, answer of this question is A, B, D and E

BGP Quiz - Question 13

What is the drawback of sending more than one BGP best path in BGP?

A. More resource usage

B. Sub Optimal Routing

C. Slower Convergence

D. Security Risk

BGP Quiz - Answer 13

Sending more than one BGP best path requires more memory, CPU, network bandwidth, thus more resource usage in the network.

As a rule of thumb, whenever more information is sent, it consumes more resource, may provide optimal routing, better high availability, better convergence.

All other options are wrong, except Option A

BGP Quiz - Question 14

What below options are the advantages of Full Mesh IBGP design compare to BGP Route Reflector design? (Choose Four)

A. It can provide more optimal routing compare to Route Reflector design

B. It can provide faster routing convergence compare to Route Reflector design

C. It provides better resource usage compare to Route Reflector design

D. It can provide better protection against route churn

E. Multipath information is difficult to propagate in a route reflector topologies

BGP Quiz - Answer 14

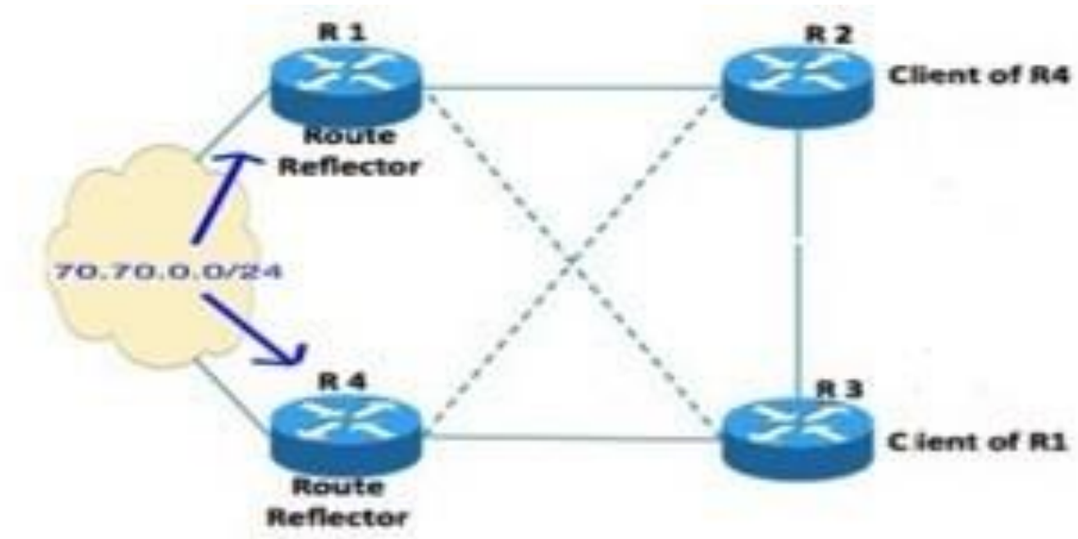
Although there are advantages of using BGP Route Reflectors, there are many drawbacks as well. Probably it is more harmful than deploying Full Mesh IBGP if the requirement is optimal routing, faster convergence and avoiding route churns.

Sending multiple paths is difficult since it requires Shadow Sessions, RR or Add-path deployments in Route Reflector topologies.

Full Mesh IBGP design consumes more device and network resources and requires more configurations on the devices compare to Route Reflector design. That’s why the answer of this question is A, B, D and E

BGP Quiz - Question 15

In the below topology IP backbone is shown. R2 is the RR client of R4 and R3 is the RR client of R1.



What is the next hop of R2 and R3 for the 70.70.0.0/24 prefix?

- A. R1 is the next hop of R2, R4 is the next hop of R3
- B. R1 is the next hop of R3, R4 is the next hop of R2
- C. R2 is the next hop of R3, R3 is the next hop of R2
- D. R4 is the next hop of both R2 and R3

BGP Quiz - Answer 15

Since it is given as IP backbone, IP destination based lookup is done for the BGP prefixes.
Sine BGP prefixes require recursion and IGP next hop needs to be found for the BGP prefixes, R2's and R3's IGP next hops for the BGP prefixes should be found.

On R2, For the BGP next hop of 70.70.0.0/24 BGP prefix is R4. R2 can only reach R4 through R3.

Thus, R2's IGP next hop is R3. It applies for the R3.
R2's IGP next hop is R3 and R3's IGP next hop is R2. That's why the answer of this question is C.

BGP Quiz - Answer 15

Please note that in this topology BGP Route Reflectors don’t follow the physical topology, which is against to BGP Route Reflector design requirement in IP networks.

That’s why, in this design between R2 and R3, routing loop occurs.

Correct design is R2 should be the Route Reflector client of R1 and R3 should be the Route Reflector client of R4

BGP Quiz - Question 16

What can be the problem with BGP design in the Enterprise if there are more than one datacenter?

A. Convergence is very slow

B. Asymmetric routing issues if there are stateful devices

C. Route Reflector deployment is harder compare to SP deployment

D. Traffic flow cannot be optimized

BGP Quiz - Answer 16

All the options are wrong except Option B.

Asymmetric can be a problem in Enterprise design, which has stateful devices as it was explained in the BGP chapter. Because stateful devices require symmetric routing for the flow information and firewalls, load balancers, IDS/IPS are common elements at the Internet edge or within the datacenters in Enterprise design.

In the Service Providers, CGN (LSN) is deployed to overcome IPv4 exhaustion problem. These nodes also require symmetric routing.
Answer of this question is B

BGP Quiz - Question 17

Which below option is true for the VPN Route Reflectors in MPLS deployments? (Choose Two)

A. It can be deployed in centralize place

B. It doesn’t have to follow physical topology, can have more flexible placement compare to IP Route Reflectors

C. It is best practice to use VPN Route Reflectors for the IP Route Reflectors as well

D. It always provides most optimal path to the Route Reflector clients

BGP Quiz - Answer 17

VPN Route reflector can be deployed in the centralized placed and they have more flexible placement advantage compare to the IP Route Reflector.
The reason is there is no IP destination based lookup in the MPLS networks.

Thus there is no layer 3 routing loop problem as in the case of IP Route Reflector which was explained in the Answer 15.

BGP Quiz - Answer 17

It is not best practice to deploy IP and VPN services on the same node. Reason will be explained in Answer 18.
VPN RR, similar to IP RR, cannot always provide most optimal path to their clients. Because they selects the BGP best path from their point view, not from their clients point of view.

That’s why the answer of this question is A and B

BGP Quiz - Question 18

What can be the problem with using IP and VPN Route Reflector on the same device? (Choose Two)

A. Attack for the Internet service can affect VPN Customers

B. Attack for the VPN service can affect Internet Customers

C. Scalability of the Route Reflectors are reduced

D. They have to participate in the IGP process

BGP Quiz - Answer 18

When a Route Reflector is used for more than one service, it is called Multi Service Route Reflector. The problem of using Internet and VPN services on the same BGP Route Reflector is Fate Sharing

Internet based attacks can affect VPN customers and any problem on the VPN service users affect Internet customers. Also in case of failure, all the customers fail.

BGP Quiz - Answer 18

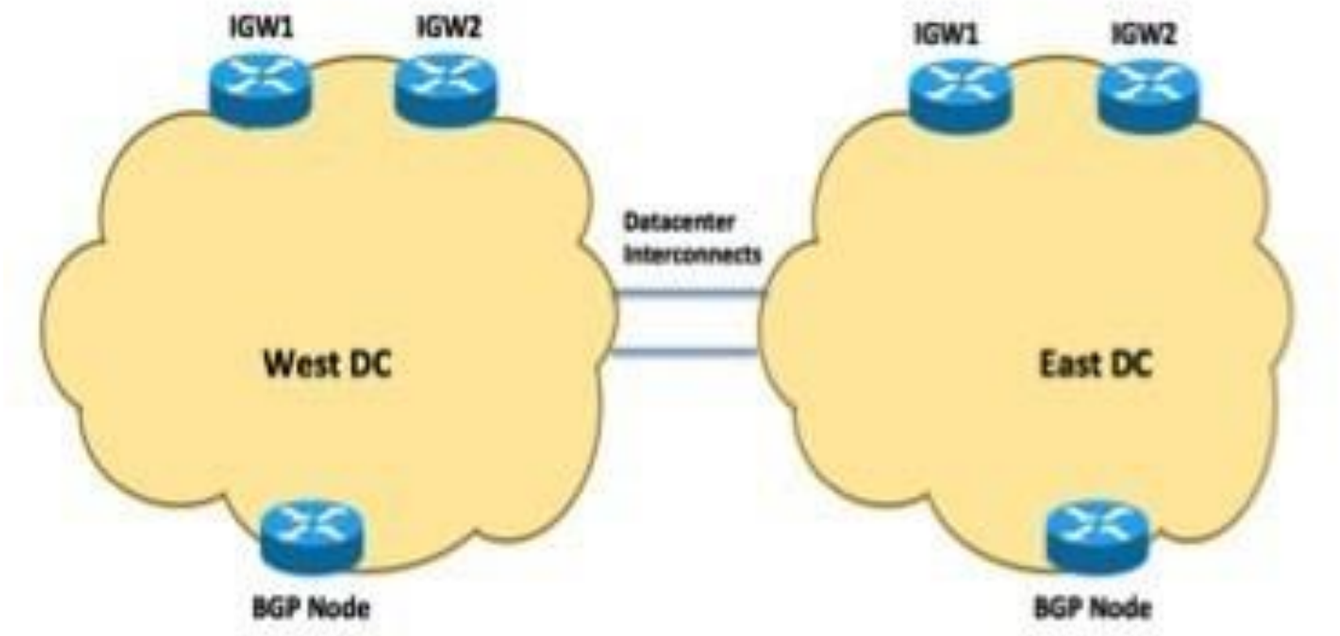
Thus using a separate BGP Route Reflector per service is a best practice. Using Multi Service RRs don’t reduce the scalability. And when using multi service RRs, they still don’t have to participate in IGP process

They can be designed as inline RR that participates IGP process in specific design such as Seamless MPLS.
Answer of this question is A and B

BGP Quiz - Question 19

In the below topology there are two datacenters of the Service Provider. If the requirement were to provide closest exit for the Route Reflector clients, in which datacenter would you deploy the Route Reflectors?

A. In West DC
B. In East DC
C. Doesn’t matter the placement
D. Both in East and West DC



BGP Quiz - Answer 19

Route Reflectors should be placed in both East and West DC. Otherwise Route Reflector would choose the best path from their point of view and would send the best path to the Route Reflector Clients from their best path.

If RR would be placed in West DC, all BGP RR Clients in East DC would choose the West DC IGW (Internet Gateways) as exit point and vice versa.

Thus the correct answer of this question is D

BGP Quiz - Question 20

Which below options are true for the BGP PIC deployment? (Choose Two)

A. BGP PIC can provide sub second convergence even if there are millions of prefixes in the routing table

B. BGP edge devices don’t have to receive more than one best path for BGP PIC Edge to work

C. BGP PIC Edge can protect both from Edge link and Node failure

D. BGP PIC has to work with BGP Add-Path

BGP Quiz - Answer 20

BGP edge nodes have to receive more than one best path for BGP PIC Edge operation. This was explained in the BGP chapter in detail. BGP Add-Path is one of the mechanisms, which is used to send multiple paths even RR is deployed in the network.

But BGP Add-Path is not mandatory for BGP PIC.

BGP PIC Edge can protect from both Edge link and node failures and can provide sub second convergence even if there are millions of prefixes.

That’s why the correct answer of this question is A and C

BGP Quiz - Question 21

Which below option provide Route Origin Validation ?

A. IRR

B. Peeringdb

C. RPKI

D. BGPSEC

E. Flowspec

BGP Quiz - Answer 21

IRR is used to provide information for filtering on the Internet facing routers but doesn’t provide Origin validation

Flowspec is not a validation mechanism, it provides protection for DDOS attack

BGPSEC provides path validation

RPKI is the only correct answer, it provides Origin Validation which mean, by creating ROAs, networks can validate the prefixes if they are coming from correct origin AS

BGP Quiz - Question 22

What are the important consideration when EBGP is used in the Massively Scale Datacenters?

A. BGP Path Hunting

B. ASN allocation schema

C. Whether there is a BGP Route Server

D. Traffic Engineering

E. Add-path or Shadow RR deployment

BGP Quiz - Answer 22

Add-path or Shadow RR are not a concern for EBGP deployments, question is asking EBGP deployment

Also in the Datacenter BGP Route Server is not used, it is used in the IXP networks

In the MSDC environments, ASN numbering schema is important to advertise routes , for example allow-as in might be required, or due to different ASN design BGP Path hunting can be a problem

BGP Quiz - Answer 22

Also traffic engineering with BGP is one of the main reasons why BGP is used inside the Massively Scale Data Centers. Between different TOR and Leaf or Leaf and Spine switches, for different type of applications or different type of traffic class, such as Elephant and Mice flows, DC operators provide different paths for different applications and the services.

Answer of this question is A. B and D

BGP – Extra Study Resources

Books :

- http://www.amazon.com/BGP-Design-Implementation-Randy-Zhang/dp/1587051095/ref=sr_1_1?ie=UTF8&qid=1436564612&sr=8-1&keywords=bgp+design+and+implementation

- **Videos** :

- https://www.nanog.org/meetings/nanog38/presentations/dragnet.mp4
  https://www.youtube.com/watch?v=txtiNFyvWjQ

- **Articles** :

- https://www.nanog.org/meetings/nanog51/presentations/Sunday/NANOG51.Talk3.peering-nanog51.pdf

- http://ripe61.ripe.net/presentations/150-ripe-bgp-diverse-paths.pdf

- http://orhanergun.net/2015/05/bgp-pic-prefix-independent-convergence/

- http://orhanergun.net/2015/01/bgp-route-flap-dampening/

- https://www.nanog.org/meetings/nanog48/presentations/Tuesday/Raszuk_To_AddPaths_N48.pdf

- http://orhanergun.net/2015/03/bgp-design-quiz/
-
- http://packetpushers.net/bgp-rr-design-part-1/
-
- http://packetpushers.net/bgp-rr-design-part-2/
-
- https://tools.ietf.org/html/draft-ietf-idr-bgp-optimal-route-reflection-10
-
- http://arxiv.org/pdf/0907.4815.pdf
-
- http://www.scn.rain.com/~neighorn/PDF/Traffic_Engineering_with_BGP_and_Level3.pdf
-
- http://packetpushers.net/bgp-path-huntingexploration/