Practical Cisco Certified Design Expert

CCDE

Orhan
Ergun

CCDE
#2014:17
CCIE #26567

- Cisco Certified Design Expert Practical/Lab exam
  - It is a computer network design exam, very different than CCIE which is an operational exam
  - Prerequisite: Student must pass CCDE Written exam to be able to attend CCDE Practical exam
  - Throughout the course CCDE Practical exam and CCDE Lab exam wordings will be used interchangeably

- CCDE Practical Exam details

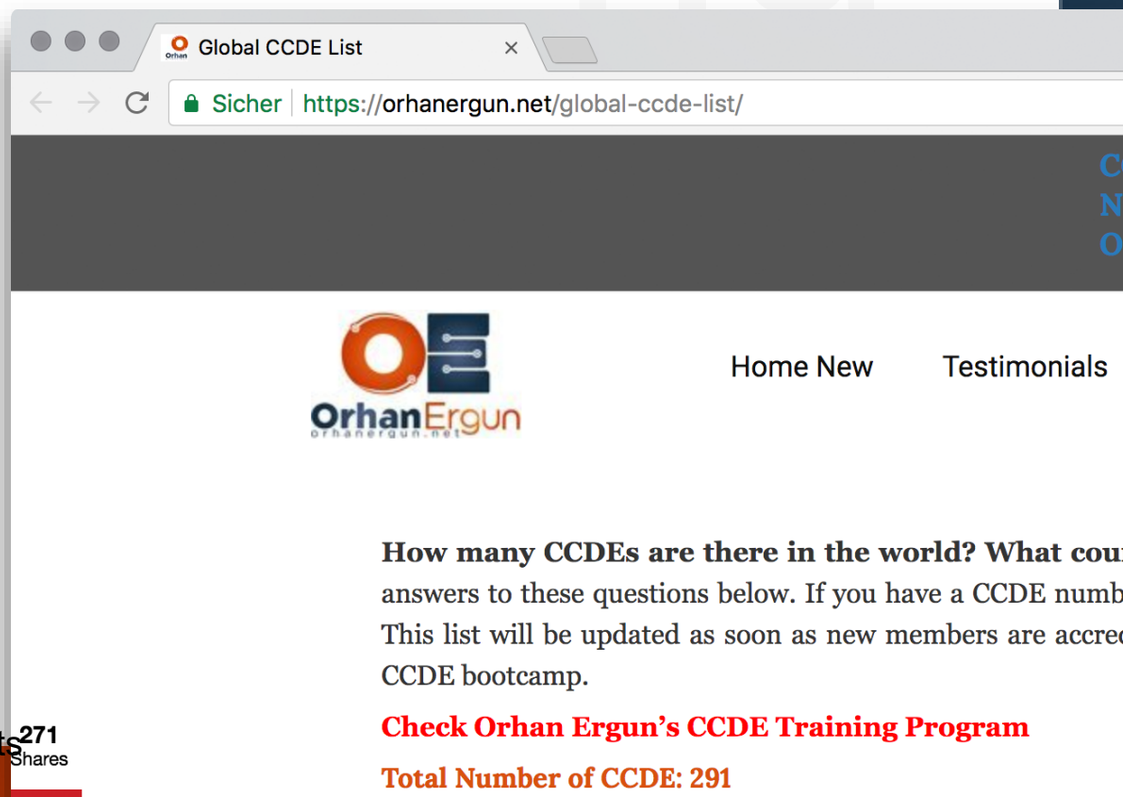
- Four scenario’s over 8 Hours
- 1 hour lunch break after first 4 hours
- Two scenarios every 4 hours
- Business Challenges – Background info, emails, text and diagram
- There is no configuration for any vendor equipment

- Vendor agnostic exam, but still some technologies relevant to Cisco (erg. HSRP, GLBP, EIGRP, DMVPN,GETVPN)
    - Exam score will be made available in 8 to 12 weeks as per Cisco
- Results are announced on Pearson Vue website and CCO account
    - Reading intensive exam. Must skim through some material in the scenario.
- Analyze, Design, Implement and Optimize are the 4 job tasks

- Analyzing the design is the most critical and hardest part
  - Most students get the lowest score from the analyzing the design part
- Exam score is provided based on these job tasks
  - Exam is given every 3 months in the Pearson Professional Center's (275 locations around the World)

- During the exam, pen and paper will be provided, also you will be able to use the computer to take notes.
  - Unlike CCIE exams, Internet is not open, so you can't take any help, you are alone there!
- You can go to the washroom, don't worry 💻
  - Passing score is around 75 – 80 %

- Less than 400 CCDE in the world as of 2018 , Most of them in U.S and as a company, work for Cisco
- https://orhanergun.net/global-ccde-list
- CCDE exam cancelled on May 2017 due to leak
- Almost 30 people lost their certificates as they involved in cheating as per Cisco
- New scenarios have been added to the exam since the cancellation, new topics have been introduced with the new scenarios
- All new topics will be covered in the course as well



CCDE Task Domains

- Four task domains in general candidates will encounter in the CCDE exam. One or more tasks can be seen in 1 scenario.

1. Merge & Divest
2. Add Technologies
3. Replace Technology
4. Scaling

task domains

Merge and Divest

- During company Mergers or Divest infrastructure of the company these are the questions you should be looking for

  - What are the best method for the migration? Phased approached or everything in one long operation?
  - Where will be the first place in the network for migration? Core to Edge? Edge to Core?
  - What happens to the Routing, IGP, BGP? Is ship in the night approach suitable?
  - Which type of security infrastructure overall network will support?

Merge and Divest

- During company Mergers or Divest infrastructure of the company these are the questions you should be looking for

  - Will merged network have IPv6? Multicast?
  - What is the new capacity requirement?
  - How will be the merged network monitored?
  - When you divest the network, where will be the datacenters? Head Quarters and connections from the branches to these locations?

Adding Technologies

- If you are adding new technologies onto an existing network these questions you should be asking.

- What can be broken? Does this technology affect others in the network?

- What does this technology provide? Do you really need it?

Adding Technologies

- If you are adding new technologies onto an existing network these questions you should be asking.
  - What are the alternatives ?
  - Which additional information do you need to deploy this technology?

Every new technology adds some amount of complexity so consider complexity vs benefits of the technology tradeoff !

Replacing Technologies

- If you are replacing one technology with the other, these questions you should be asking.

- Is this change really needed? Is there a valid business case?

- What is the potential impact to overall network?

Replacing Technologies

- If you are replacing one technology with the other, these questions you should be asking.
  - What will be the migration steps?
  - Is there a budget constraint?
  - Does this new technology require a learning curve ?

Scaling

- Hierarchy is common way to provide scaling
- Scaling is growing the network, merging, divesting, upgrading without taking it completely down or waiting for flag day

Scaling

- Which method is better to provide scaling?
  - OSPF Multi area design vs. single area design, can we scale OSPF network with some feature (prefix-suppression, OSPFv3 etc.) instead of multi area OSPF design

- Which place in the network to provide scalability? Where you should place an OSPF ABR?

CCDE Written Recommended Reading List

- CCDE Streamlined Study Resources prepared by Orhan Ergun, Elaine Lopes, Andre Laurent and Virgilio Spaziani can be found from the below link.

https://learningnetwork.cisco.com/community/certifications/ccde/written_exam/study-material

These resources are useful for the CCDE Practical exam as well

Course Outline

Layer 2 Technologies 5%

- Spanning Tree CST, PVST+, RSTP, RPVST+, MST
- Vlan and Flow Based Load Balancing
- LAG, MC-LAG, VSS, VPC
- TRILL, Fabricpath, PB, PBB, SPB
- Layer 2 Fast recovery mechanisms: G.8032, REP
- First Hop Redundancy Protocols HSRP, VRRP, GLBP
- Layer 2 and Layer 3 interaction
- VLAN, VTP, DTP Technologies
- Layer 2 Traffic Engineering

Course Outline

- TSN – Time Sensitive Networking IEEE 802.1AS
- First Hop Security Mechanisms
- Layer 2 and 3 access design
- Case Studies
- Layer 2 Technologies in the CCDE Exam
- Summary
- Bonus Materials

Course Outline

Layer 3 Routing 15-25%

- OSPF Theory
- OSPF Fast Convergence, Scalability. Multi Area OSPF Design
- Fast Reroute with OSPF
- Overlay Technologies and OSPF (GRE, mGRE, DMVPN, GETVPN LISP)
- OSPF in the Datacenter, Enterprise and Service Provider Networks
- OSPF Design Best Practices
- OSPF Advantages and Disadvantages
- Case Studies
- OSPF in the CCDE Exam
- Summary
- Bonus Materials

Course Outline

IS-IS

- IS-IS Theory
- IS-IS Fast Convergence, Scalability.
- Multi Level IS-IS Design
- Fast Reroute with IS-IS
- Overlay Technologies and IS-IS (GRE, mGRE, DMVPN, LISP)
- IS-IS in the Datacenter, Enterprise and Service Provider Networks
- IS-IS Design Best Practices
- IS-IS Advantages and Disadvantages
- Case Studies
- IS-IS in the CCDE Exam
- Summary
- Bonus Materials

Course Outline

EIGRP

- EIGRP Theory
- EIGRP Fast Convergence, EIGRP Scalability
- Overlay Technologies and EIGRP (GRE, mGRE, DMVPN and LISP)
- EIGRP Design Best Practices
- EIGP Advantages and Disadvantages
- Case Studies
- EIGRP in the CCDE exam
- Summary
- Bonus Materials

BGP Course Outline

- BGP Basics – Why BGP, Autonomous Systems
- BGP Best Path Selection
- BGP Monitoring Protocol - BMP
- EBGP – Basics of EBGP
- IBGP Multipath , EBGP Multipath , EIBGP Multipath
- Inter domain routing – Settlement Free Peering, IP Transit
- ISP Tiers – Tier 1 , 2 , 3 Type Providers
- IBGP – Basics of IBGP
- BGP Route Reflectors
- Route Reflector Design Options
- BGP Optimal Route Reflection – ISP Case Study
- BGP Confederations
- Full mesh IBGP vs. Route Reflector vs. Confederation
- Full mesh to RR Migration
- RR to Confederation Migration

BGP Course Outline

- BGP – IGP Interactions – Blackhole avoidance
- BGP – MPLS Interactions
- BGP LU – Labeled Unicast
- BGP LS – BGP Link State
- BGP Segment Routing
- BGP EPE – Egress Peer Engineering
- BGP Flowspec
- BGP Session Culling
- AIBGP – Accumulated IBGP
- BGP Selective Blackholing
- BGP Route Propagation Behavior – RFC 8212
- BGP Security - Route Leak, Hijacking, RPKI, Origin and Path Validation

BGP Course Outline

- BGP in the Datacenter
- BGP in the WAN
- BGP PIC – Prefix Independent Convergence
- Case Studies
- BGP vs. IGP Comparison
- BGP in the CCDE exam
- Summary
- Bonus Materials

Course Outline

MPLS and Applications 25%

- MPLS Basics
- LDP vs. RSVP
- Layer 2 MPLS VPN - VPWS, VPLS, EVPN, PBB-EVPN, VXLAN-EVPN
- Layer 3 MPLS VPN
- Inter-AS MPLS VPNs
  - Option A, B , C, D
- MPLS Traffic Engineering – SLA to customer, Network Utilization, FRR Use cases
- Inter Area/AS Traffic Engineering
- PCE – PCEP

Course Outline

MPLS and Applications 25%

- Carrier Supporting Carrier
- Seamless/Unified MPLS
- MPLS Transport Profile
- GMPLS – Generalized MPLS
- RMR – Resilient MPLS Ring
- Segment Routing
- MPLS in the Mobile networks
- MPLS in the WAN networks

Course Outline

MPLS and Applications 25%

- MPLS in the Core networks
- MPLS in the Datacenter
- MPLS in the CCDE Exam
- Case Studies
- Summary
- Bonus Materials

Course Outline

VPN Design 10%

- VPN Theory, Overlay and Underlay Concepts
- GRE
- mGRE
- IPSEC
- DMVPN
- GETVPN
- L2TPv3
- LISP
- OTV

outline

Course Outline

VPN Design 10%

- VXLAN
- NVGRE
- STT
- GENEVE
- VPN Comparisons
- VPN Quiz
- VPN technologies in the CCDE Exam
- Summary
- Bonus Materials

Course Outline

QOS Design 10%

- IP QoS
- MPLS QoS
- Diffserv MPLS Tunneling Models – Uniform, Short Pipe and Pipe Models and Comparison
- QoS in the CCDE Exam
- Summary
- Bonus Materials

Course Outline

Security Design 10%

- Infrastructure ACL, URPF (Strict, Loose, Flexible Path), RTBH, Control and Management Plane Protection
- DNS Security, DNS Sinkhole
- IGP Routing Security
- Security in the CCDE Exam
- Summary
- Bonus Materials

Course Outline

Multicast Design 10%

- Multicast Theory , History and Benefits
- PIM Deployment Models – ASM, SSM and Bidir
- MPLS Multicast Rosen GRE, MLDP, P2MP RSVP
- AMT – Automatic Multicast Tunneling
- BIER – Bit Indexed Multicast Replication
- Multicast in the CCDE Exam
- Case Studies
- Summary
- Bonus Materials

Course Outline

IPV6 Design 5%

- IPv6 Theory, Address Planning
- IPv6 Transition Mechanisms
- Dual Stack, Tunneling , Translation: 6rd, 6to4, Toredo, LISP, DS-Lite, MAP-E, MAP-T, 464-XLAT
- IPv6 Tunnel Brokers
- Carrier Grade NAR/Large Scale NAT

Course Outline

Evolving Technologies

- SDN
- Openflow
- NFV
- IOT
- OpenStack
- NETCONF/YANG
- Edge and Fog Computing
- AI, Machine Learning, Deep Learning
- Summary
- Bonus Materials

LAYER 2 TECHNOLOGIES

Agenda

- Ethernet and it’s control planes
- What is control and data plane?
- Vlan and Flow Based Load Balancing
- Spanning Tree CST, PVST+, RSTP, RPVST+, MST
- LAG, MC-LAG, VSS, VPC
- VLAN, VTP and Trunking
- First Hop Redundancy Protocols HSRP, VRRP, GLBP
- HSRP, VRRP and GLBP Comparison
- Layer 2 and Layer 3 interaction
- Layer 2 Traffic Engineering
- Layer 2 and 3 access design
- Case Studies
- Layer 2 Technologies in the CCDE Exam
- Summary

Agenda

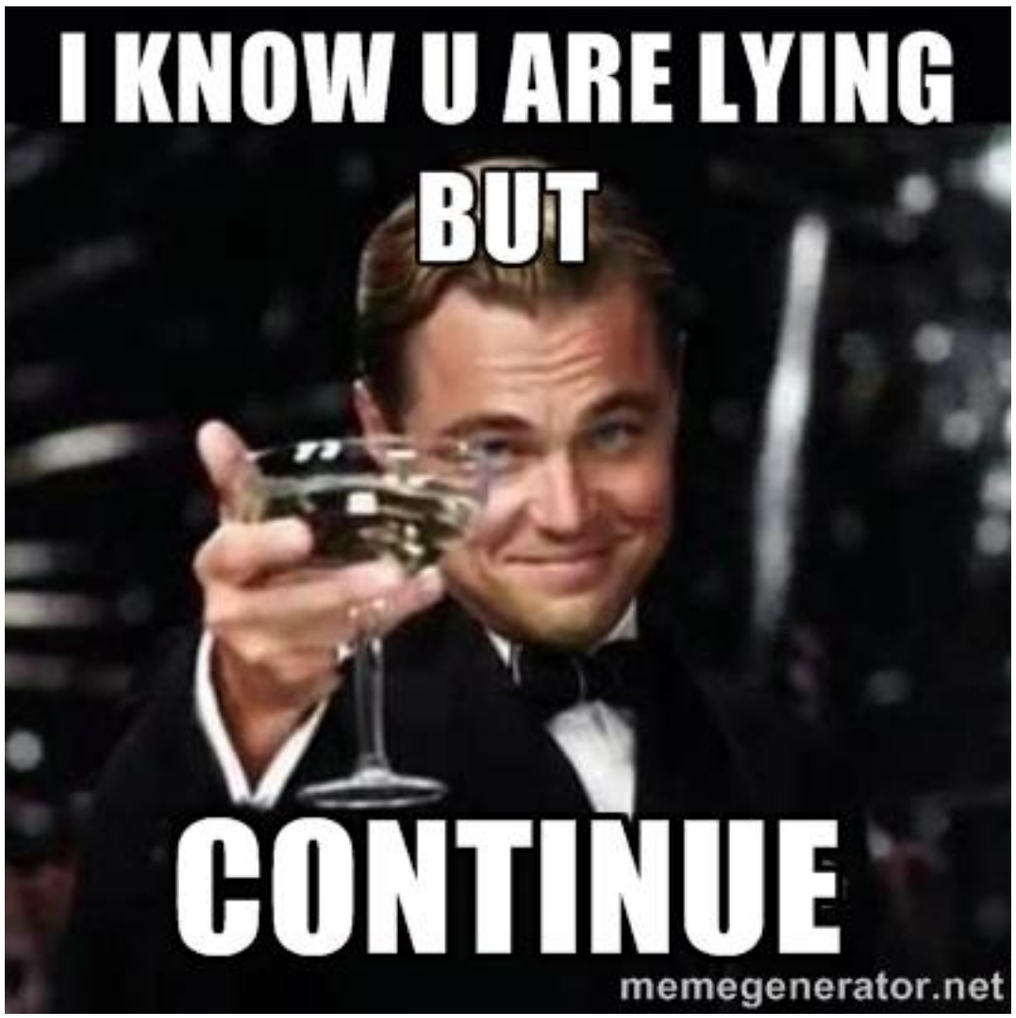
- Bonus Materials
- Bonus - TRILL, Fabricpath, PB, PBB, SPB
- Bonus - Layer 2 Fast recovery mechanisms: G.8032, REP
- Bonus - TSN – Time Sensitive Networking IEEE 802.1AS
- Bonus - First Hop Security Mechanisms

Ethernet and it’s control planes

- Ethernet is a most common layer 2 protocol in today Campus, datacenter and WAN networks.
  - Spanning tree its legacy control plane. Although today there are many control plane for Ethernet such as SPB, Trill, SPB-TE, Fabric path, spanning tree is by far mostly used control plane mechanism.
- That’s why it is important to understand spanning tree from the design point of view.

Ethernet and it's control planes

- Is Spanning Tree really the most common control plane?



Ethernet and it's control planes

- Hyper giant datacenters, as an example, FAMGA, uses IGP or BGP in the datacenter due to scale.
    - If Layer 3 fabric is not needed, then LAG and MC-LAG is the most common control plane for link bundling mechanism in any decent size datacenter.
- Spanning tree is used as a backup mechanism even if LAG and MC-LAG is used.

Control and Data Plane

- Control plane refers to all the functions and processes that determine which path to use in the network. Routing protocols, spanning tree, LDP etc. are examples.
  - Control plane packets are destined to or locally originated by the networking devices, such as routers, switches itself.
- Data plane refers to all the functions and processes that forward packets/frames from one interface to another.

Control and Data Plane

- As for the data plane, sometimes called the Forwarding Plane, this is basically anything that goes ‘through’ the networking device, and not ‘to’ the networking device.
  - Management plane is all the functions you use to control and monitor devices.
- Data plane is usually called user plane in the mobile network business. Also, the term forwarding plane is occasionally used

Control and Data Plane

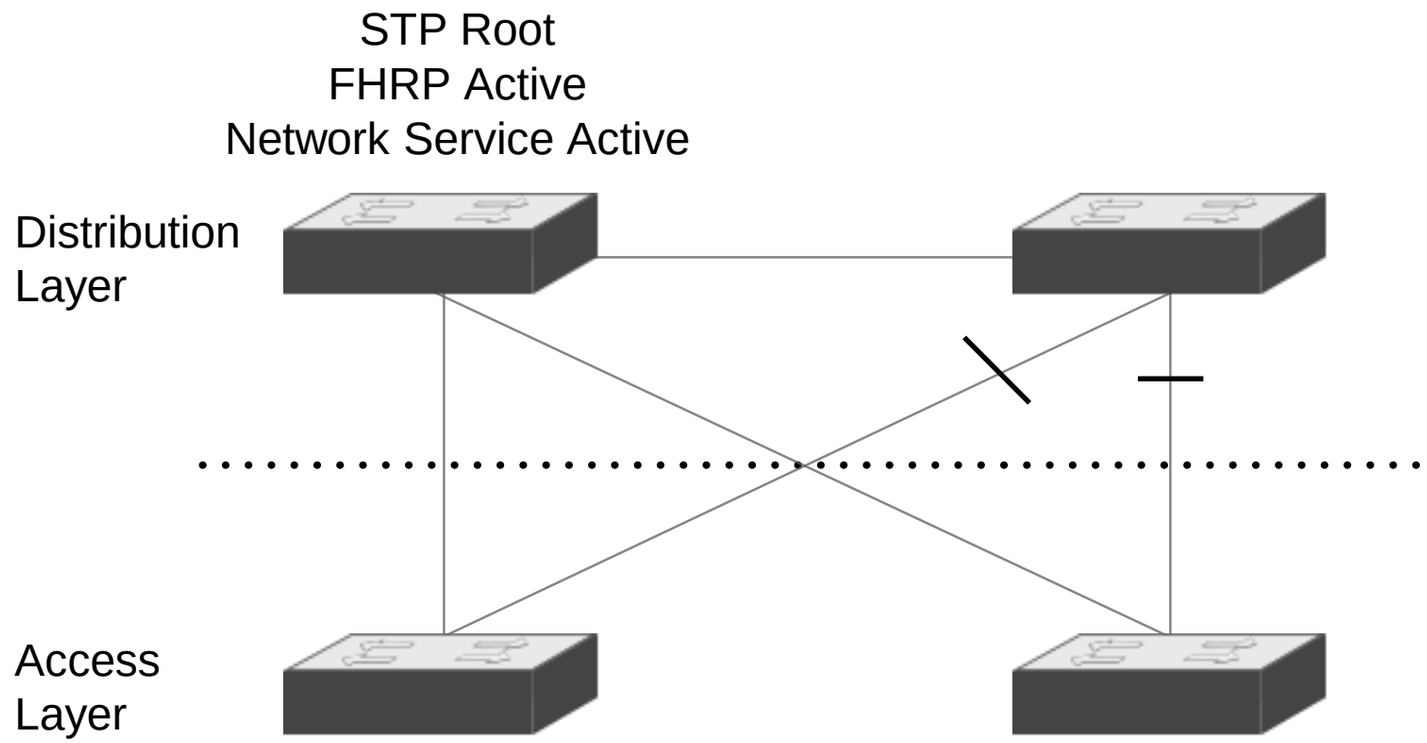
- Control Plane learning is using control plane protocols to advertise reachability information.

- Data plane learning is used in Layer 2 in general, networking devices examine the packets/frames and learn MAC to interface binding.

pla
ne

Vlan and Flow Based Load Balancing

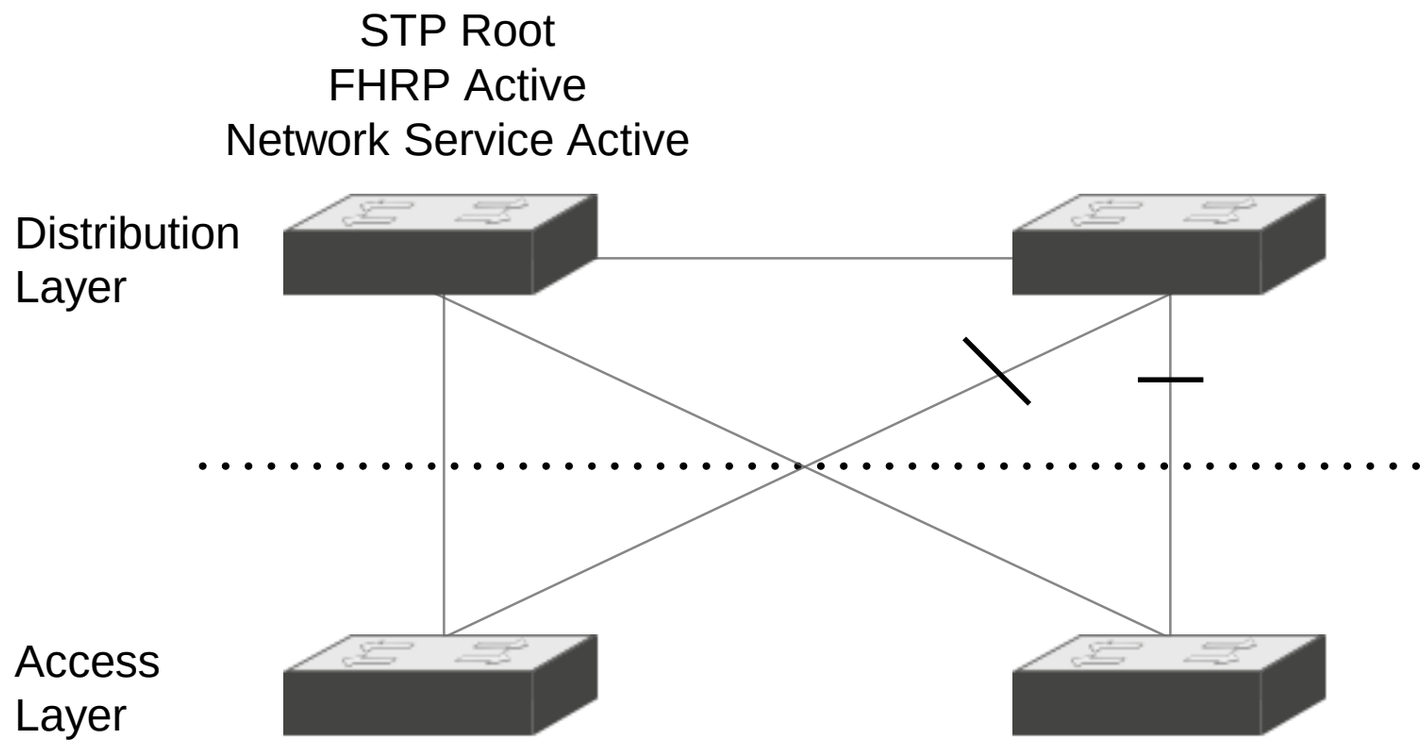
- Vlan based load balancing allow the switch to be active layer 3 gateway for only some set of Vlans and other switch stays as standby, and for the different set of Vlans standby switch acts as an active switch and active switch acts as standby.



Triangle Spanning Tree Topology

Vlan and Flow Based Load Balancing

- Flow based load balancing mean is to allow both gateway switches and the links to the gateway switches to be used as an active-active for the same Vlan.



Triangle Spanning Tree Topology

Vlan and Flow Based Load Balancing

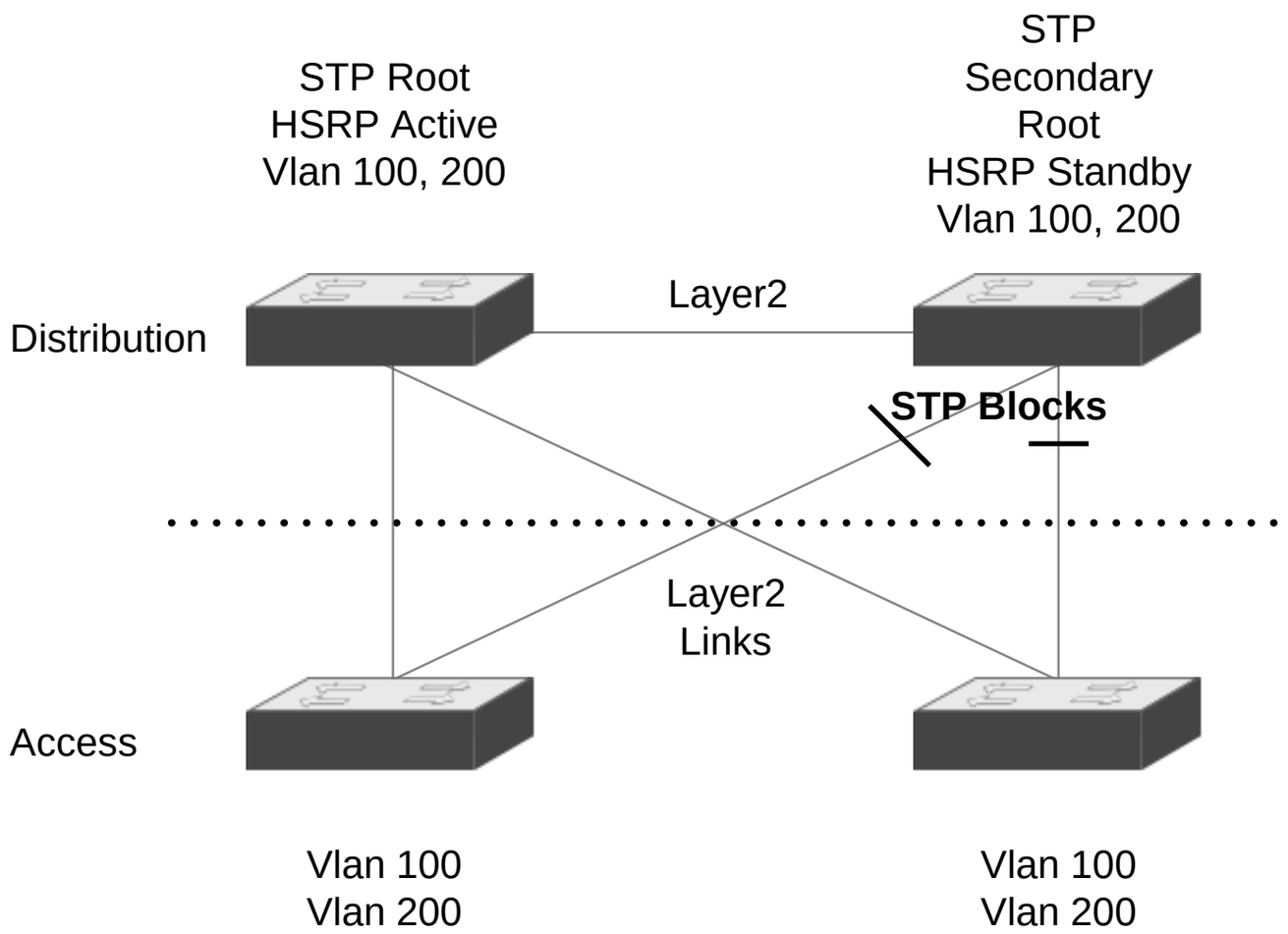
- In VPLS and EVPN, Vlan and flow based load balancing concepts will be revisited
  - For flow based load balancing, Multi Chassis Link Aggregation (MC-LAG) should be enabled. In that case, all of the links between the switches is placed in a bundle and can be utilized regardless of spanning tree's existence. Spanning tree behaves as one link to the logical link aggregation bundle.

Vlan and Flow Based Load Balancing

- In flow-based load balancing, hosts in the same Vlan can use both links at the same time. In spanning tree this is not possible.
    - From the access layer switches to the distribution layer switches Multi-chassis Link aggregation group bundle can be activated, in that case flow based load balancing can be possible.

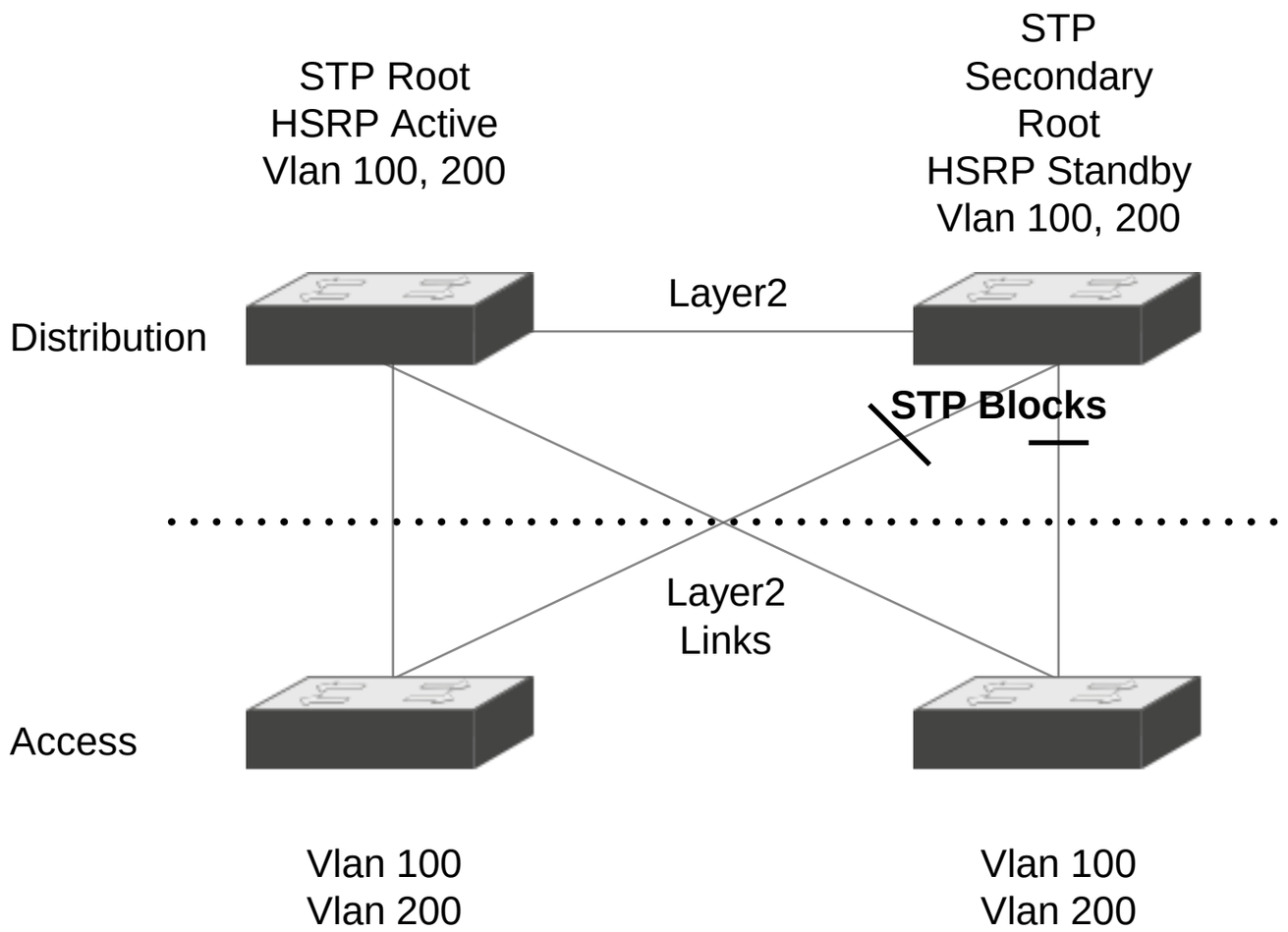
Spanning Tree

- Spanning tree is a control plane mechanism for Ethernet. It is used to create a layer 2 topology (A tree) by placing the root switch on top of the tree.



Spanning Tree

- Since classical Ethernet works based on data plane learning and Ethernet frames don't have TTL for loop prevention, loop is prevented by blocking the links by the spanning tree.



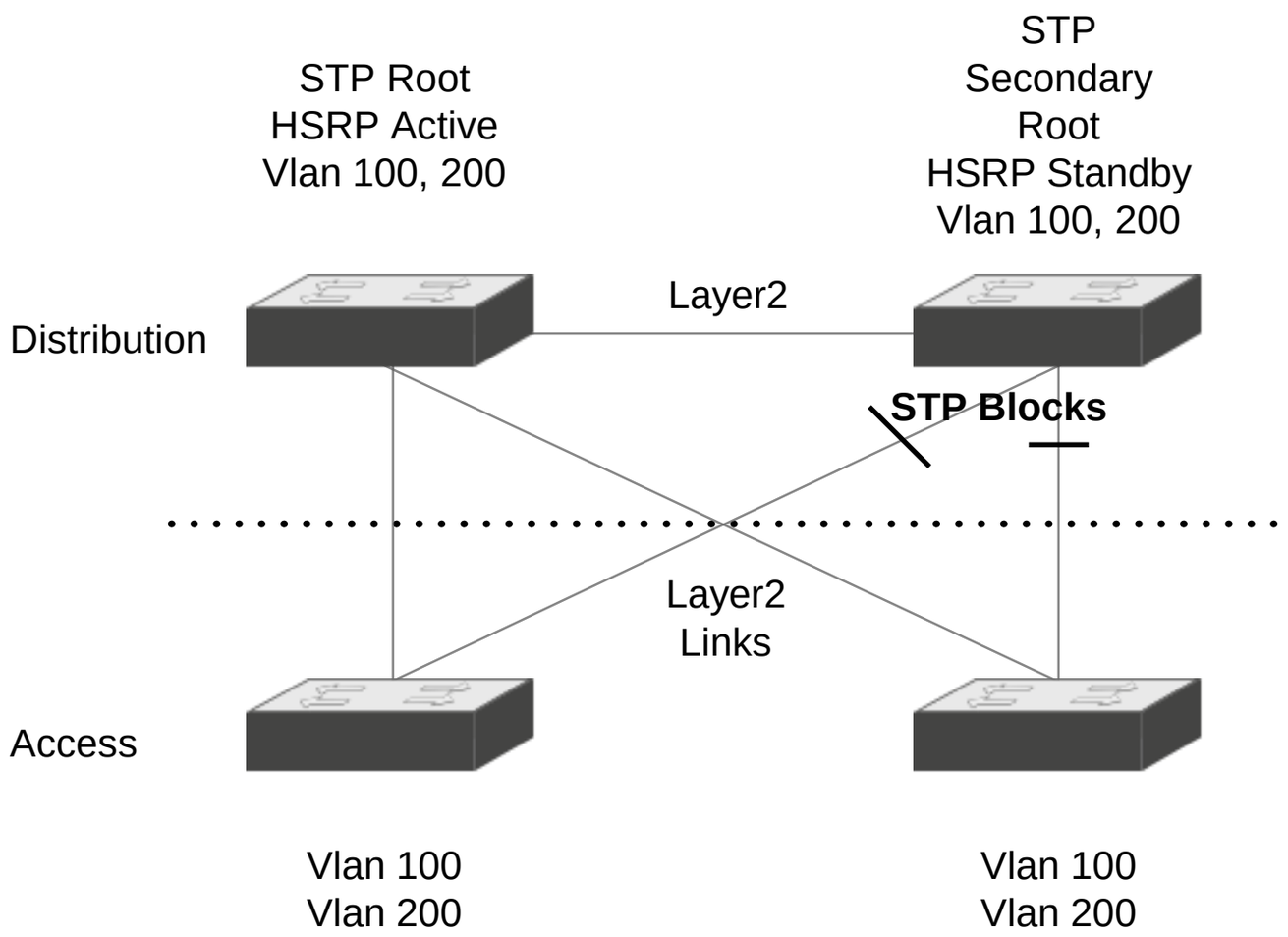
Spanning Tree

- Loop has to be mitigated but blocking links don’t allow using all the links actively. Spanning tree doesn’t provide multipathing.

- As soon as spanning tree detects a loop, it blocks some links in the topology to prevent the loop.

Spanning Tree Root Switch Selection

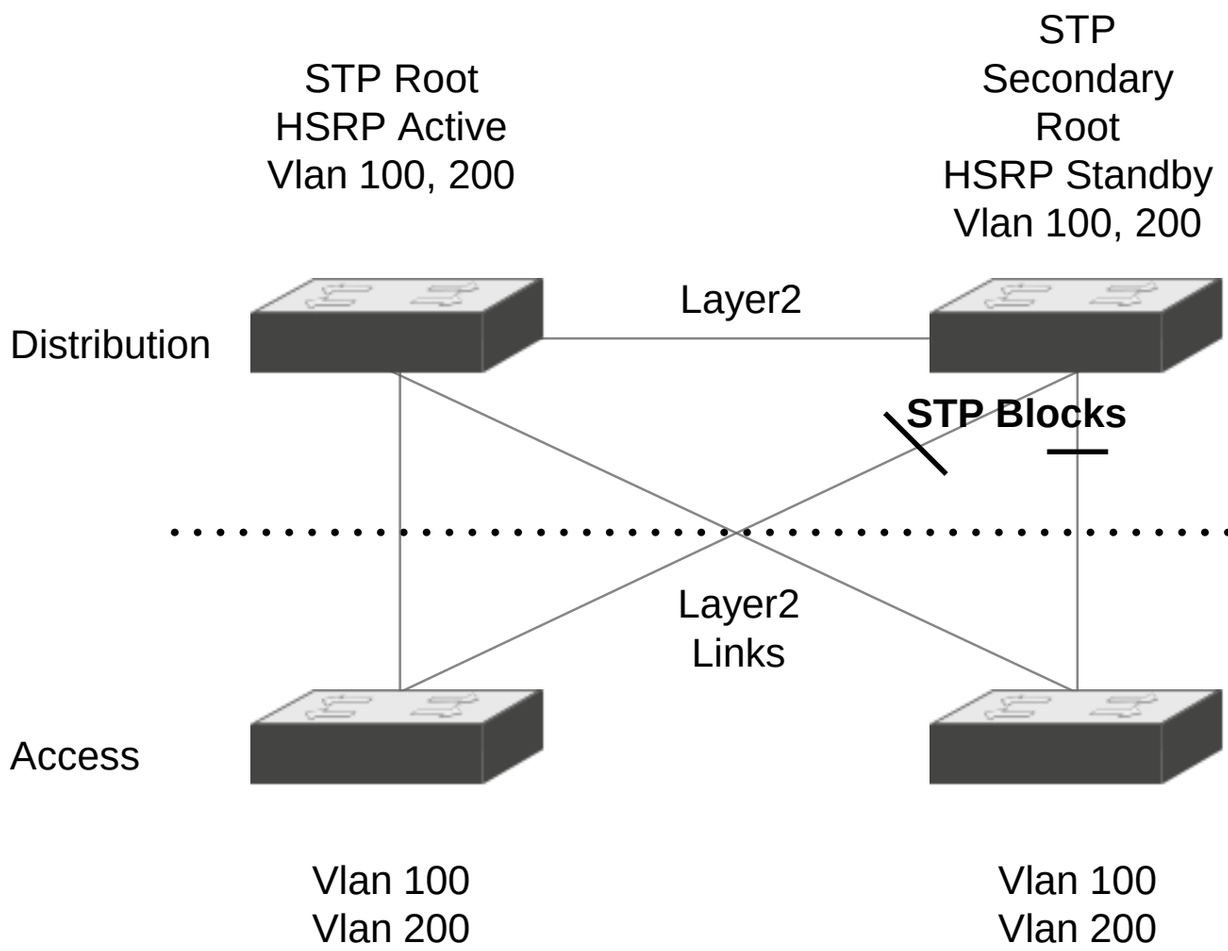
- One switch is selected as root switch.
- Root switch is selected based on priority, if priority is not set manually, switch has the lowest MAC address becomes root switch.



Spanning Tree Root Switch Selection

- Setting root switch priority manually provides determinism and it is good thing.

- If it is configured manually Newly added switch don’t change the forwarding topology.



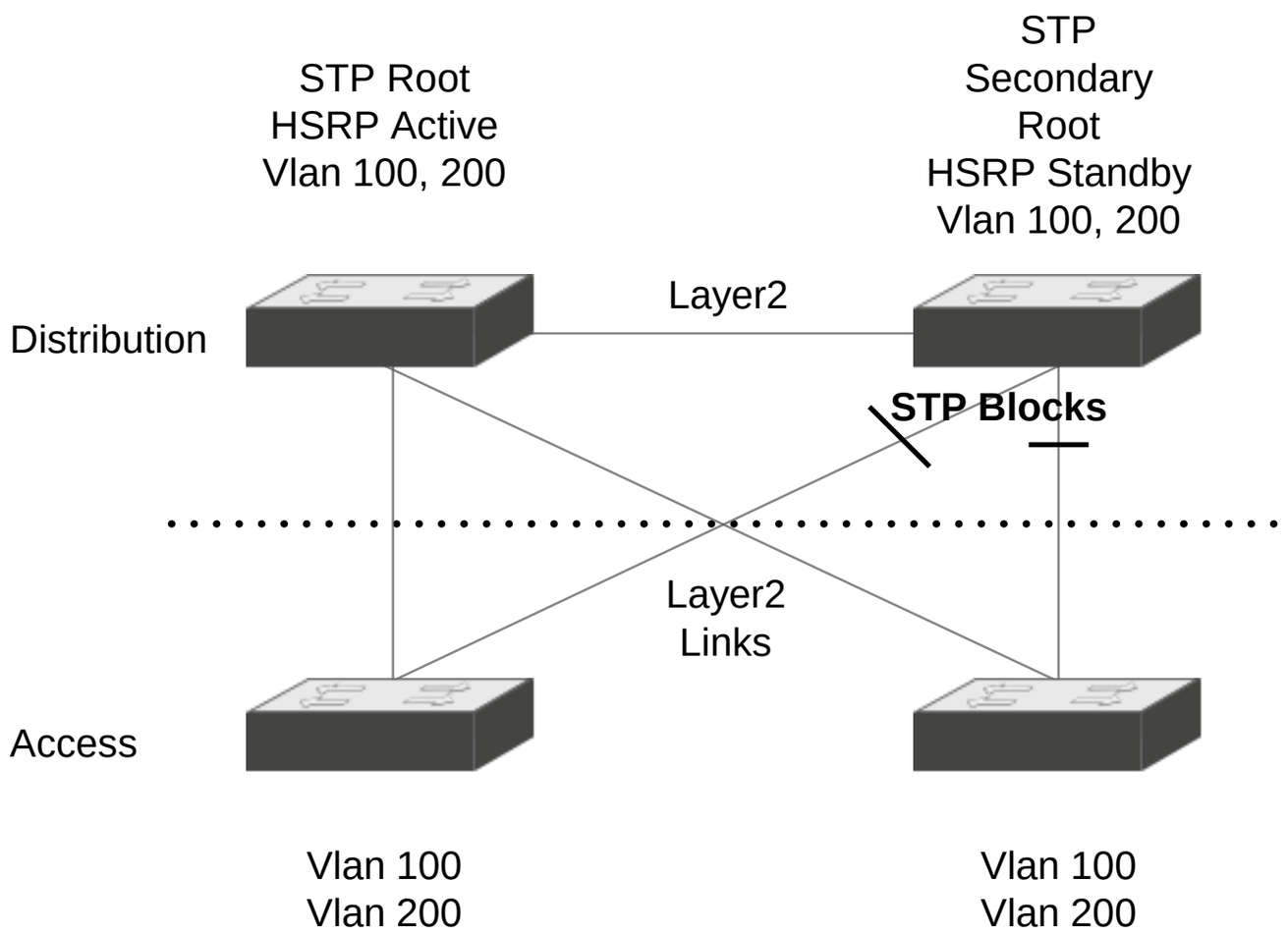
Spanning Tree and Load Balancing Capability

- Two common load balancing techniques are in the Layer 2 networks; Vlan based and Flow based load balancing
  - Spanning tree doesn’t allow the flow based load balancing
- Some implementations of spanning tree allows only Vlan based load balancing. Some of them allow only active-standby redundancy

Spanning Tree and Load Balancing Capability

- CST (Common Spanning Tree) 802.1d which is classical/legacy spanning tree; supports only one instance for all vlans. It doesn't support Vlan load balancing.

- CST supports only Active – Standby redundancy, all Vlans have one root switch and one backup root switch



Spanning Tree and Load Balancing Capability

- Take advantage of Vlan load balancing instead of Active-Standby, with Vlan load balancing you can use your available uplink capacity. It is called bisectional bandwidth as well

- Vlan load balancing can be cumbersome, operationally hard but gives advantage of using all uplinks.

Spanning Tree Toolkit

- The following enhancements to 802.1(d,s,w) comprise the spanning-tree toolkit:
  - PortFast—Allows the access port bypass the listening and learning phases
- UplinkFast-Provides 3-to-5 second convergence time after a link failure.

Spanning Tree Toolkit

- Backbone Fast—Cuts convergence time by MaxAge for indirect failure.

- Loop Guard—Prevents the alternate or root port from being elected unless Bridge Protocol Data Units (BDPUs) are present.

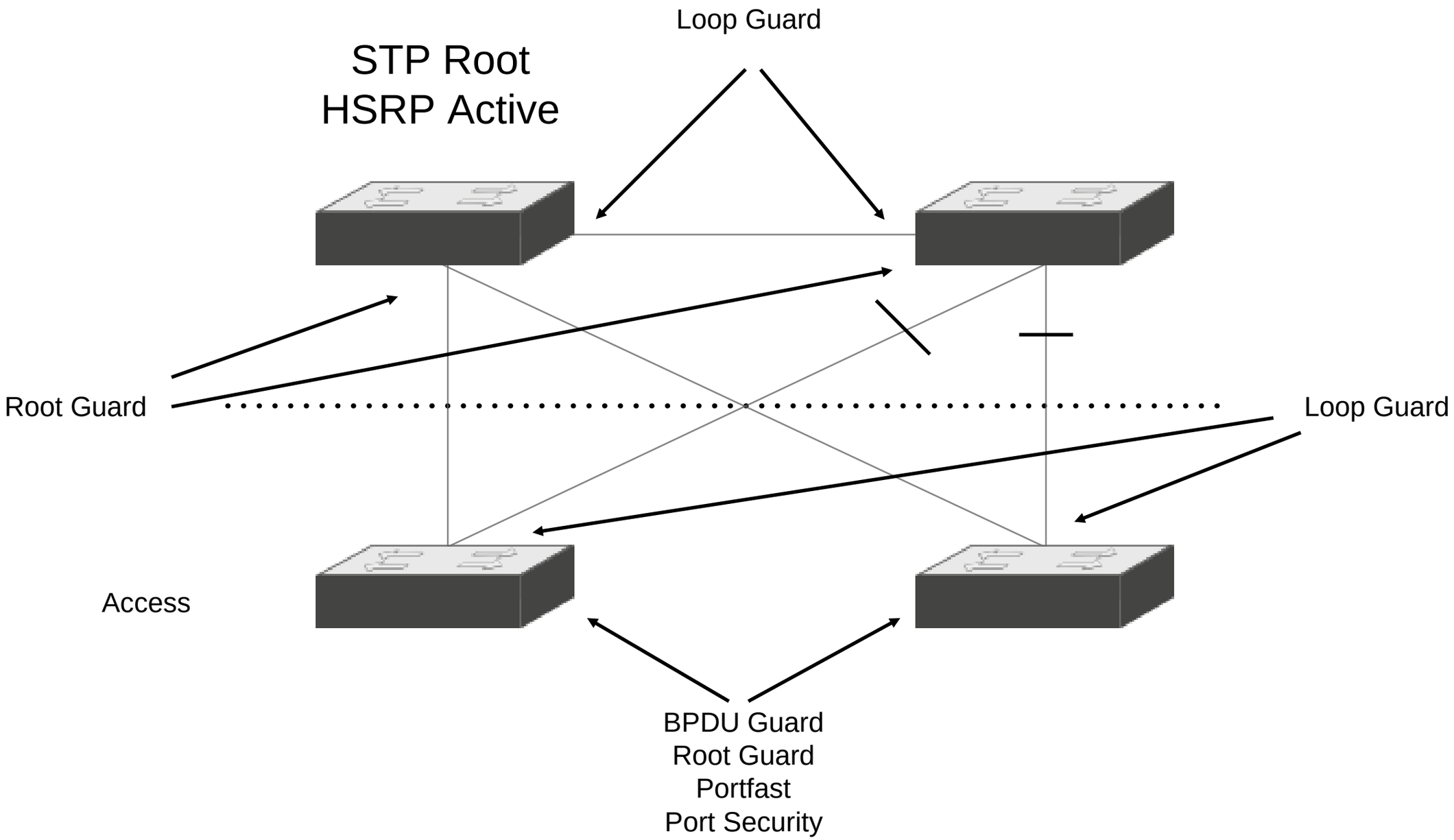
- Root Guard—Prevents external switches from becoming the root. Provides determinism.

Spanning Tree Toolkit

- BPDU Guard—Disables a PortFast-enabled port if a BPDU is received.

- BPDU Filter—Prevents sending or receiving BPDUs on PortFast-enabled ports.

Spanning Tree Toolkit Placement



Spanning Tree Toolkit Placement

- PVSTP+ Cisco's 802.1d implementation, supports one to one instance to Vlan mapping.

- Enhancements to PVSTP provide good optimizations, but it has slow convergence compared to MST and RSTP and cannot scale as MST.

Spanning Tree Toolkit Placement

- MST 802.1s is the industry standard. Convergence is like RSTP, proposal and agreement mechanism. Group of vlans are mapped to spanning tree instance.

- So if you have 100 Vlans you don’t need to have 100 Instance as in the case of RPVST+ thus reduces CPU and memory requirements on the switches, so provides scalability.

Spanning Tree Toolkit Placement

- With the region support, MST can be used between data centers. But still spanning tree domain is limited to local data center. Think of it as an OSPF multi area.

- MST supports large number of VLANs so that’s why it might be suitable to large data centers or service provider access networks if uses QinQ, 802.1ah Provider bridging PB or Mac in Mac 802.1aq Provider Backbone Bridging PBB.

Spanning Tree Toolkit Placement

- MST is used still in many datacenters because of its large scale layer 2 support. Also capability of having different MST regions on different datacenter allows spanning tree BPDU’s to be limited to individual data center.

Spanning Tree Best Practices

- Use RSTP or RPVST+ for fast convergence for direct and indirect failures.
  - Use MST for scaling. If you have large scale Vlan deployment and CPU is a concern, you can take advantage of grouping vlans to MST instance.
- Don’t use 802.1d, CST. If you will deploy standard base spanning tree mechanism, use RSTP or MST.

Spanning Tree Best Practices

- Always enable spanning tree on the access facing ports to protect the network from intentional or unintentional attacks.
  - Port-security is used as a spanning tree loop avoidance mechanism at the edge of the layer 2 campus Ethernet networks.
- For multipath support, enable LAG with Spanning Tree

Odd – Even Vlan Load Balancing

- For ease of troubleshooting, you can use one distribution switch as primary root switch for odd Vlans; other distribution as primary root switch for even Vlans, it gives predictability.

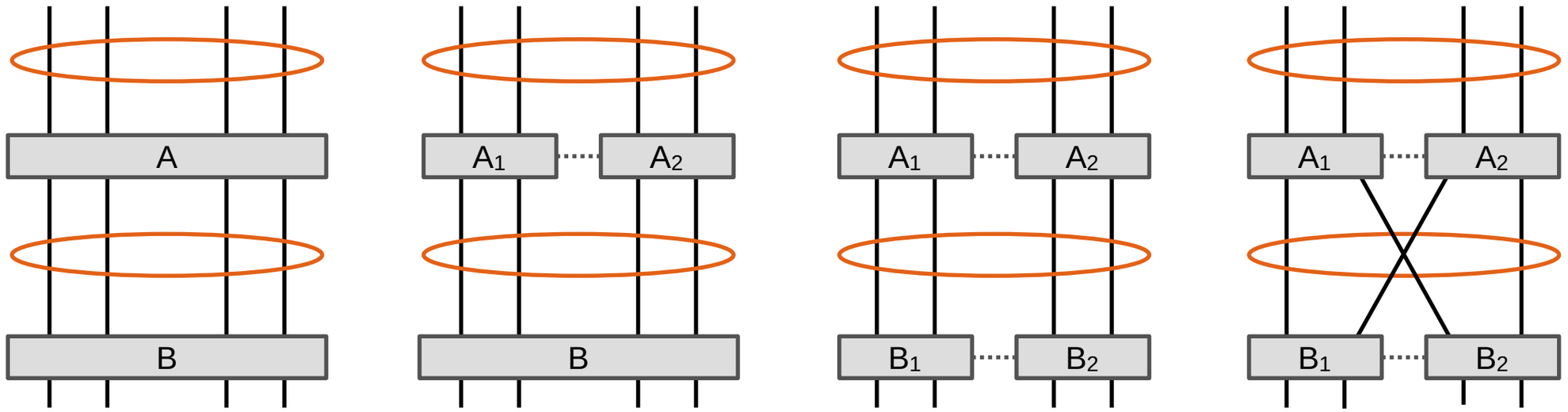
odd

even

LAG, MC-LAG, VSS and VPC

- LAG (Link Aggregation Group) is an IEEE 802.1AX-2008 standard.
  - It is used to place multiple Ethernet links in a bundle.
- Allows one or more links to be aggregated together to form a Link Aggregation Group, such that a network device can treat the Link Aggregation Group as if it were a single link.

LAG, MC-LAG, VSS and VPC



(1) LAG (2) MLAG+LAG (3) MLAG+MLAG (4) High Availability

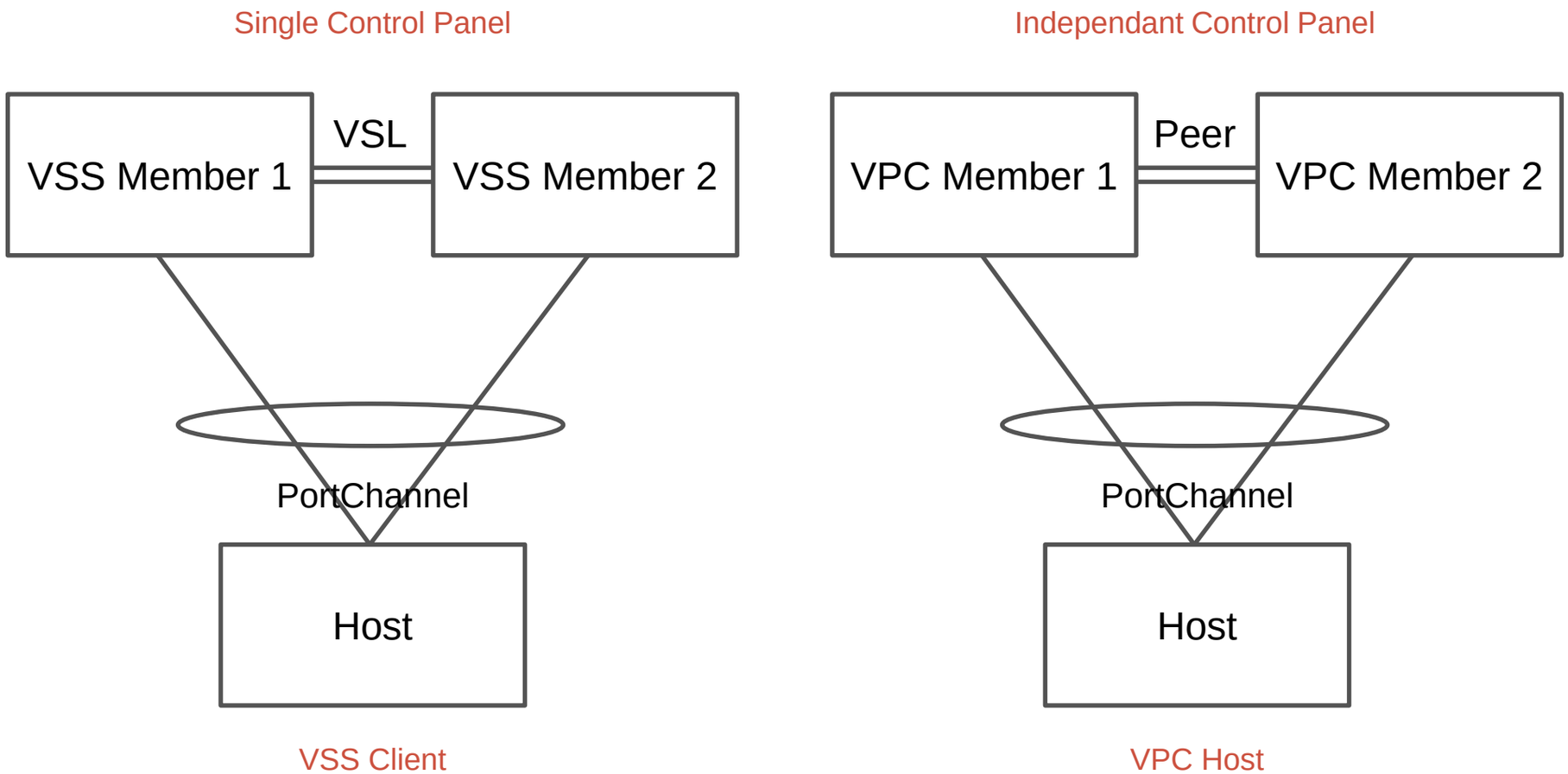
- Caveats:
- Flows are mapped to the physical link by the hashing algorithm of the network device in per flow load balancing.
  - Per packet load balancing can cause reordering problem at the destination due to jitter.
- Each Individual flow can reach up to physical link speed, for example if flow generates 1,5 Gbps traffic but it is mapped to 1 Gbps physical link, capacity need to be increased.

VENDOR	IMPLEMENTATION NAME
Cisco NEXUS	VPC
Cisco Catalyst 6500	VSS
Juniper	MC-LAG
Ericsson	MC-LAG

VSS and VPC

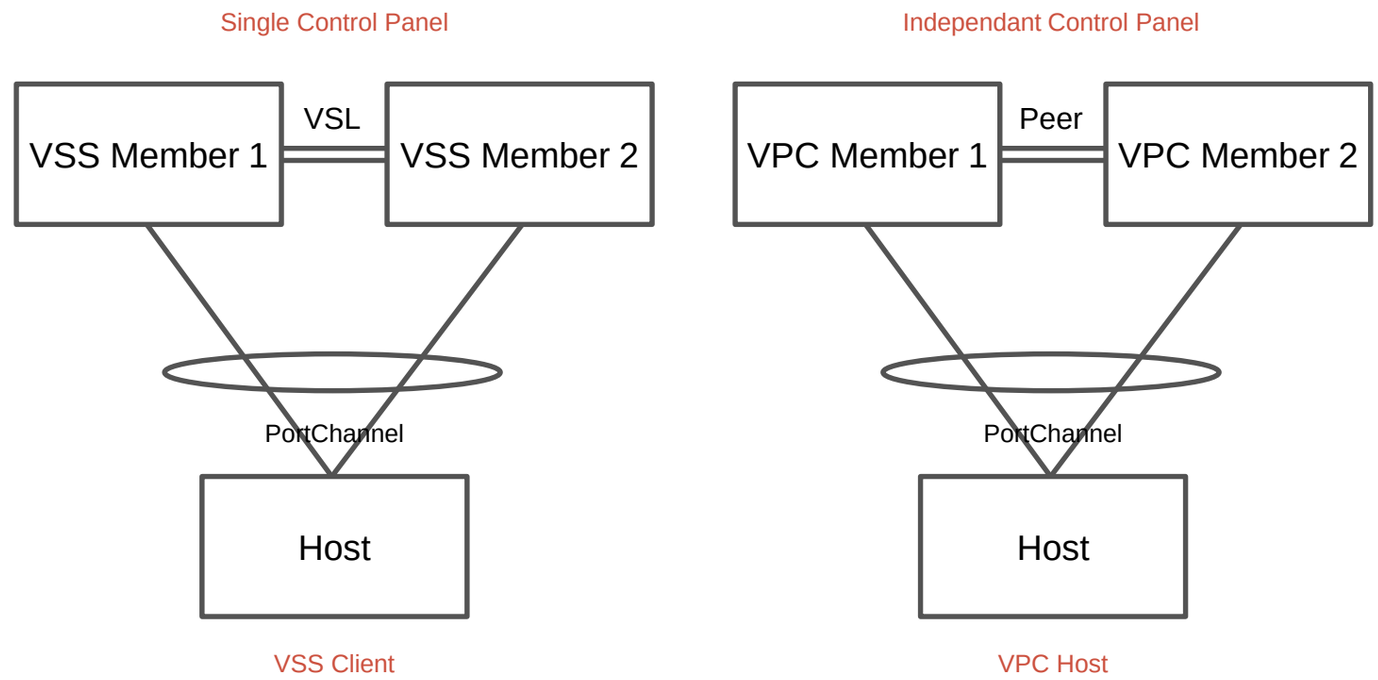
- VSS – Virtual Switching System switches work as single control plane.
  - VPC – Virtual Port Channel switches work as individual control plane.
- Thus you don’t have to run HSRP in VSS but run in VPC.
  - Both VSS and VPC provides multipathing, flow based load balancing and eliminates blocked links.
- Both provides device level redundancy to the downstream switches or hosts.

VSS and VPC



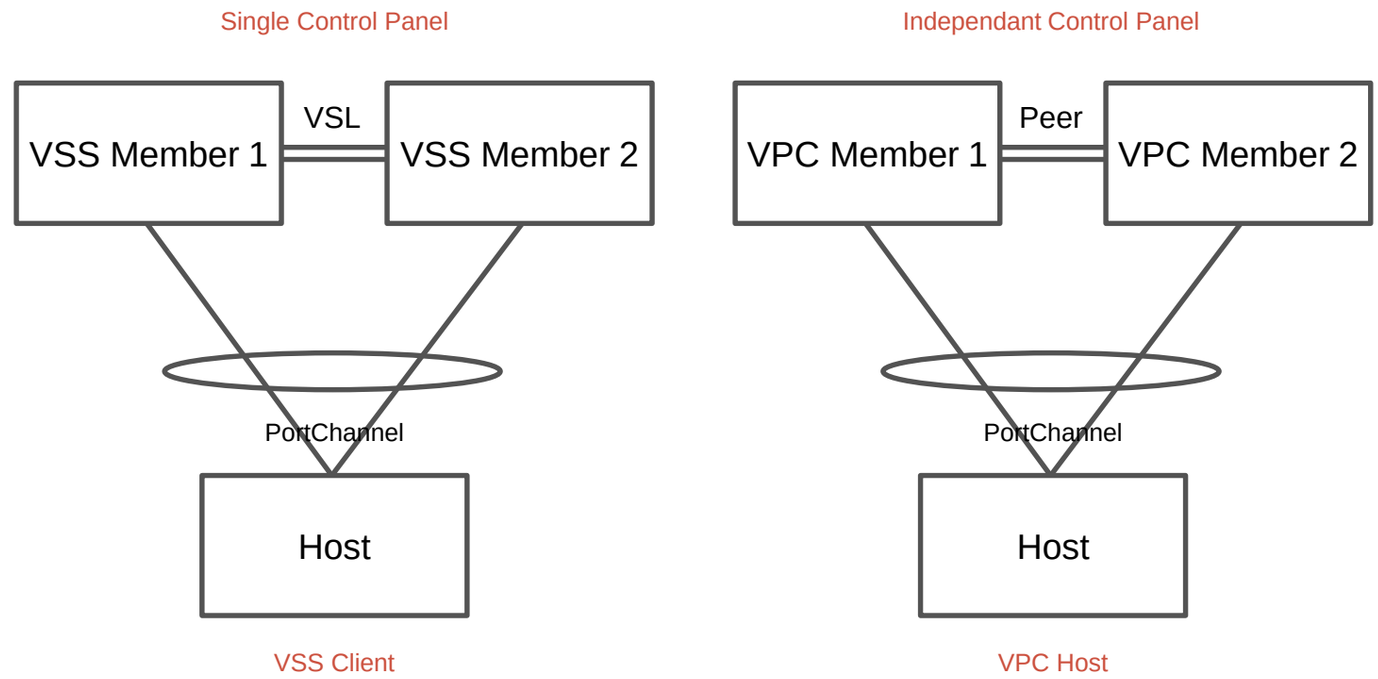
VSS and VPC

- VSS and VPC enabled switches use Cisco Preparatory protocol between to send control messages such as system IDs
- Downstream switch sees two VSS or VPC switches as one switch



VSS and VPC

- Downstream switch is not aware that it is connected to two different switches, because VSS or VPC members send a logical system ID instead of each individual switch physical System ID



VSS and VPC

- Both of these technologies are Cisco preparatory but as it is shown before other networking vendors provide the same functionality with the different names

VLAN, VTP and TRUNKING Best Practices

- VTP is not recommended anymore because of configuration complexity and the catastrophic failure. In other word, small mistake on the VTP configuration can take whole network down. So benefits of VTP might be too costly.
  - Risk of VTP based incident is greater than benefit of VTP.
- If VTP will be used, VTP Transparent mode is recommended practice because it decreases the potential for operational error.

VLAN, VTP and TRUNKING Best Practices

- Always configure VTP Domain name & Password when VTP is used.
- Manually prune unused VLANs from trunked interfaces to avoid broadcast propagation.
- Don't keep default VLAN as native VLAN, it protects from VLAN hopping attacks.
- Disable trunks on host ports.

VLAN, VTP and TRUNKING Best Practices

- Don’t put so many host in one Vlan, keep it small to provide manageable fault domain. In the same Vlan all broadcast, unknown unicast packets have to be processed by all the nodes.

- If fast convergence is the requirement don’t use Dynamic Trunking Protocol, it slows down the convergence since switches negotiate the trunking mode.

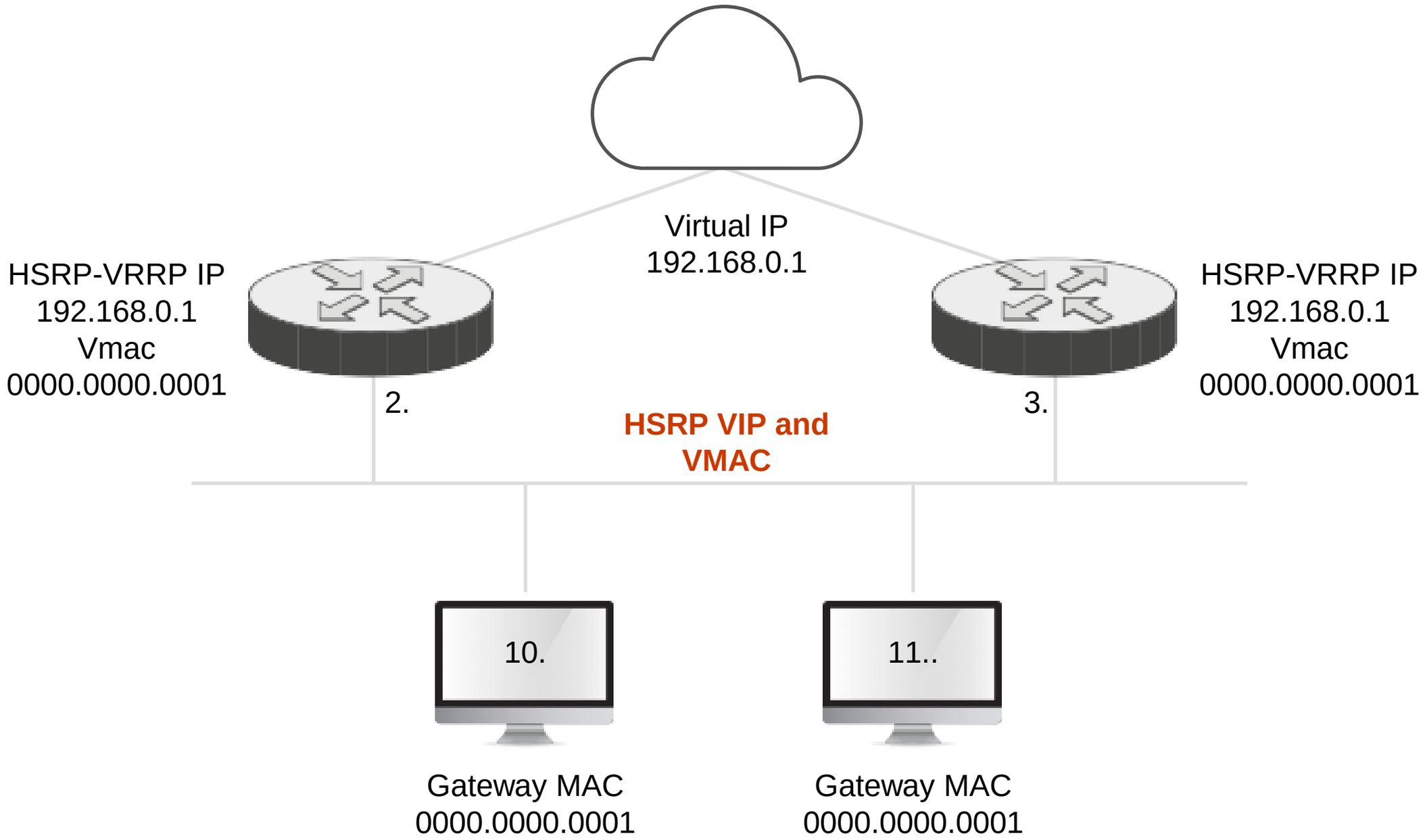
First Hop Redundancy Protocols

- Three commonly used first hop redundancy protocols are HSRP, VRRP and GLBP.
  - All of them provide device level redundancy in Layer 2 access networks, if topology is layer 3, we don't have any of these protocols.
- HSRP and GLBP are the Cisco preparatory protocols but VRRP is IETF standard, so use VRRP if you need multivendor or interoperability.

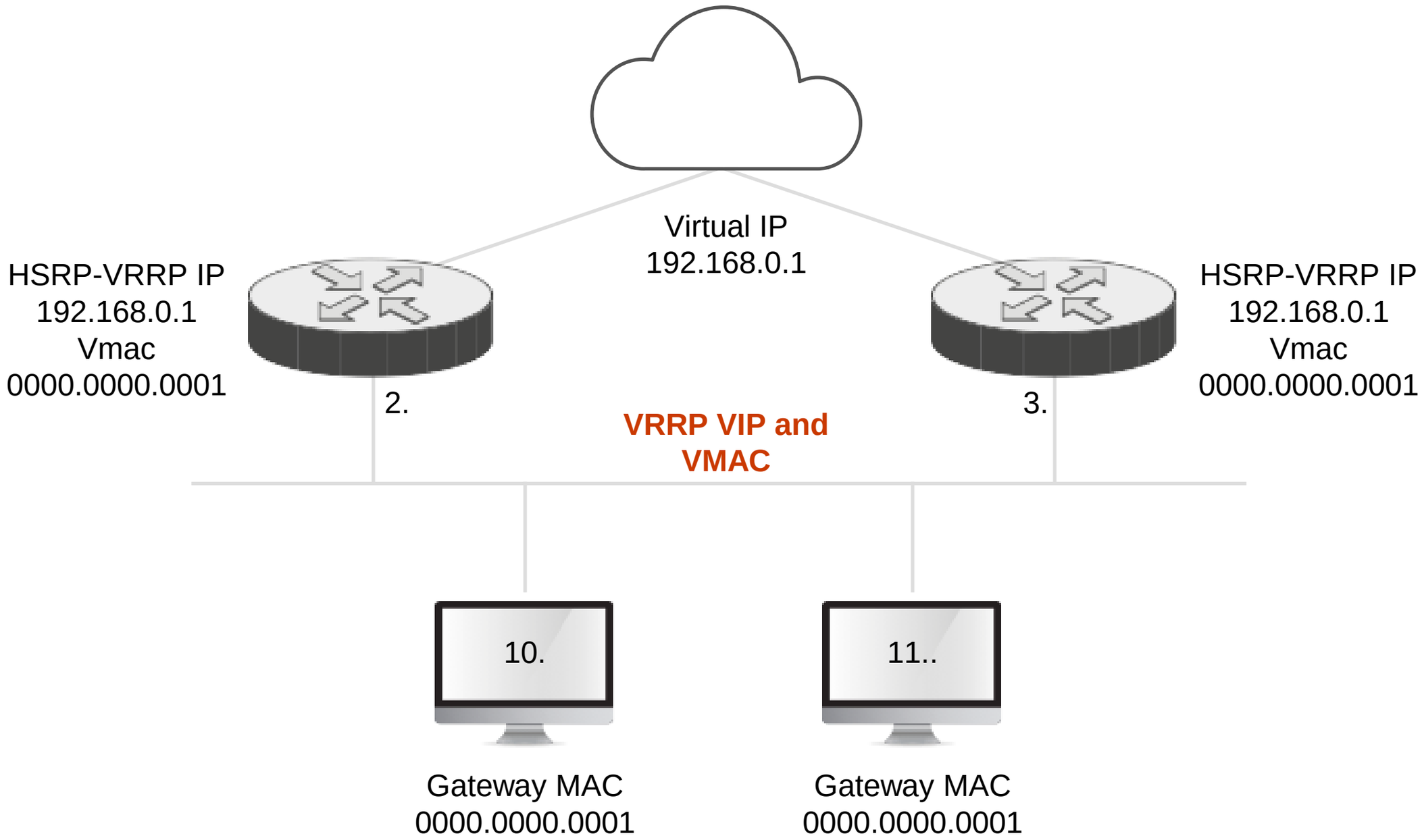
First Hop Redundancy Protocols

- HSRP and VRRP use 1 Virtual IP and 1 Virtual MAC address for gateway functionality.
  - Hosts always have the same Virtual IP address in HSRP, VRRP and GLBP.
- Virtual MAC address doesn’t change in HSRP, VRRP and GLBP in case of a failure.

HSRP


Virtual IP
192.168.0.1
HSRP-VRRP IP
192.168.0.1
Vmac
0000.0000.0001
2.
3.
HSRP-VRRP IP
192.168.0.1
Vmac
0000.0000.0001
HSRP VIP and
VMAC
10.
11..
Gateway MAC
0000.0000.0001
Gateway MAC
0000.0000.0001

VRRP


Virtual IP
192.168.0.1
HSRP-VRRP IP
192.168.0.1
Vmac
0000.0000.0001
2.
3.
HSRP-VRRP IP
192.168.0.1
Vmac
0000.0000.0001
VRRP VIP and
VMAC
10.
11..
Gateway MAC
0000.0000.0001
Gateway MAC
0000.0000.0001

GLBP

- GLBP uses 1 Virtual IP and several Virtual MAC address. For the clients ARP requests, different virtual MAC addresses are given thus network based load balancing can be achieved.
    - But still each individual client uses same device as its default gateway.
- Different clients use different device as their default gateway.

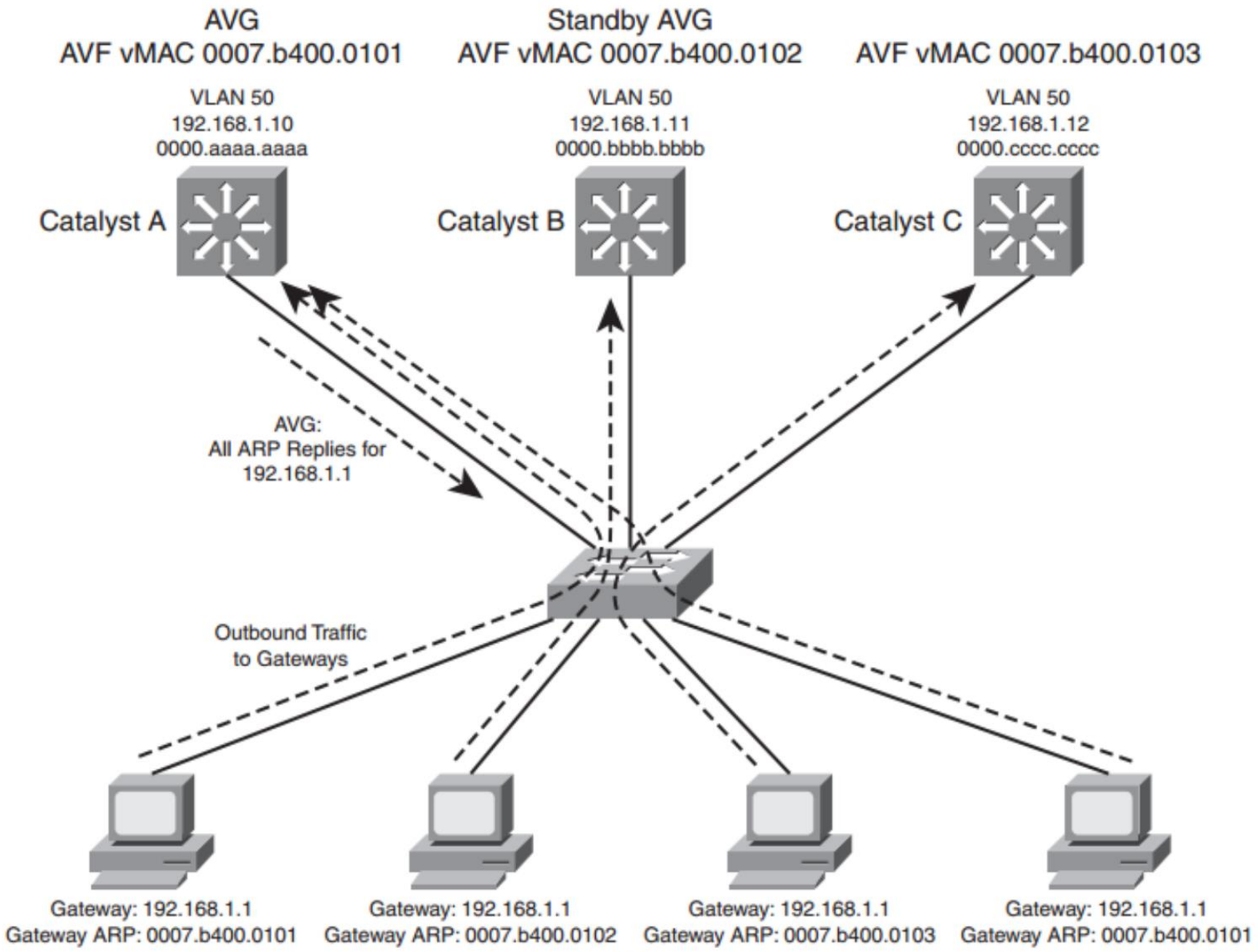
GLBP

- There are two terminologies.
  - AVF – Active Virtual Forwarder and AVG – Active Virtual Gateway.
- Each member of GLBP is AVF.

GLBP

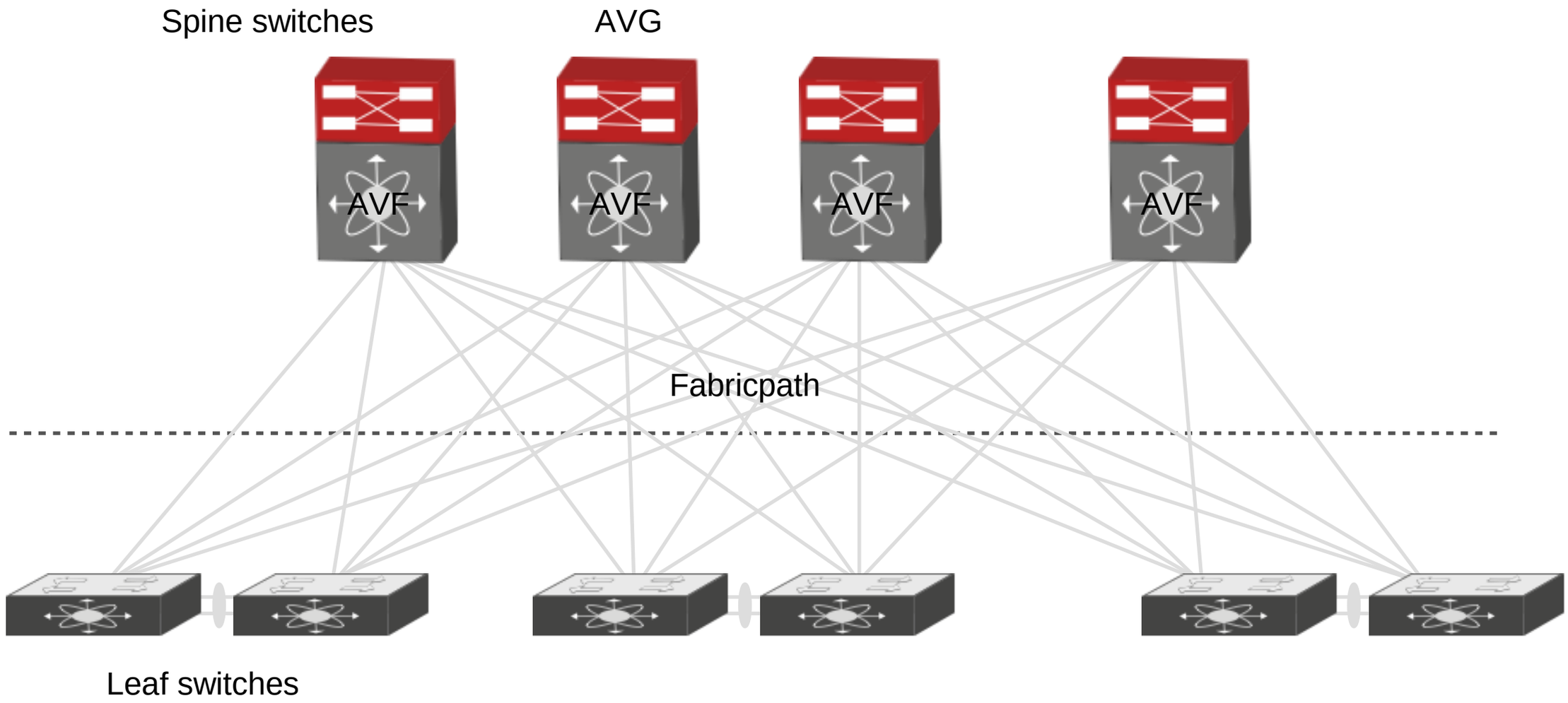
- AVG assigns a virtual MAC address to each member of GLBP group. The AVG also answers ARP requests for the Virtual IP address.
  - Each router is an AVF which forwards traffic received on VIP and vMAC.
- Different deployment options, by default Round Robin but it supports Weight as well. So AVG can send different amount of traffic to the different AVFs based on the configuration.

GLBP Operation – How it works



GLBP on Leaf and Spine Topology

- One popular design with GLBP and fabricpath which can provide up to 4 active virtual forwarder on spine switches



GLBP Operation - How it works

- GLBP might be suitable for campus but not for Internet Edge since the firewall uses same IGW as its default gateway by using same IP address.

glb
p

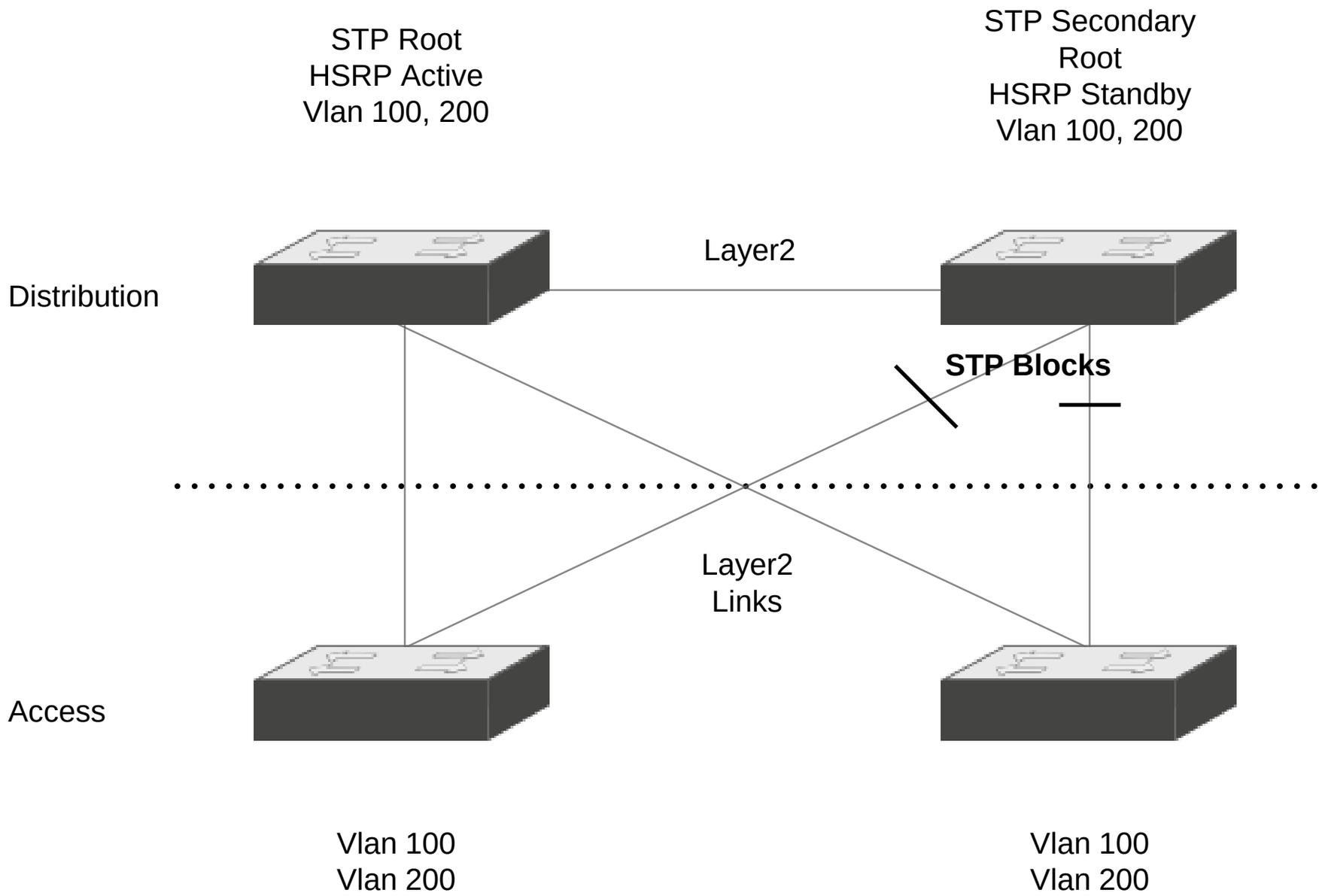
HSRP - VRRP - GLBP Comparison

orhanergun.net	HSRP	VRRP	GLBP
Suitable on LAN	Yes	Yes	Yes
Suitable on Datacenter	Yes if layer 3 access is not used	Yes if layer 3 access is not used	Yes if layer 3 access is not used
Suitable on Internet Edge	Yes but there might be better options such as routing with the Firewall or router behind the firewall	Yes but there might be better options such as routing with the Firewall or router behind the firewall	No, it creates polarization issue. This explained in detail on Orhan Ergun's CCDE Course
Standard Protocol	No,Cisco proprietary	Yes IETF Standard	No,Cisco proprietary
Preemption Support by default	No,You need to configure it manually and preemption is important to avoid suboptimal traffic flow	Yes it is enabled by default and you can disable it on many vendor implementation	No,You need to configure it manually and preemption is important to avoid suboptimal traffic flow
Virtual IP and MAC	1 Virtual IP and 1 Virtual MAC	1 Virtual IP and 1 Virtual MAC	1 Virtual IP and Multiple Virtual MACs
Stuff Experince	Very well known	Well known	Not well known
Flow based load balancing	No	No	Yes,Active Virtual Gateway responds ARP requests with different Active Virtual Forwarder in an indivudial vlan
Vlan based load balancing	Yes with HSRP Groups	Yes with HSRP Groups	Yes with GLBP Groups
Transport Protocol	Multicast	Multicast	Multicast
Default Convergece	Slow - 10 seconds	Fastest but still slow for some applications - 3 seconds	Slow - 10 seconds
Security	MD5 Authentication	MD5 Authentication	MD5 Authentication
More than 2 device support	Yes	Yes	Yes
IPv6 Support	Yes	Yes with VRRP v3	Yes
Active Active Node Support	Yes with Anycast HSRP	Yes	Yes

Layer 2 and Layer 3 Interaction

- One important factor to take into account when tuning HSRP is its preemptive behavior.
  - Preemption causes the primary device to retake the primary role when it comes back online after a failure or maintenance event.
- But HSRP by default is not preemptive, manually needs to be configured.
  - VRRF by default comes with preemption.

Layer 2 and Layer 3 Interaction



Layer 2 and Layer 3 Interaction

- Preemption is the desired behavior because the STP/RSTP root should be the same device as the HSRP primary for a given subnet or VLAN.

- If HSRP and STP/RSTP are not synchronized, the interconnection between the distribution switches can become a transit link, and traffic takes a multi-hop L2 path to its default gateway.

Layer 2 and Layer 3 Interaction

- HSRP preemption needs to be aware of switch boot time and connectivity to the rest of the network. It is possible for HSRP neighbor relationships to form and preemption to occur before the primary switch has L3 connectivity to the core. If this happens, traffic can be dropped until full connectivity is established. This is HSRP to Layer 3 interaction.

Layer 2 and Layer 3 Interaction Summary

- We have two interactions here, one for Spanning Tree and HSRP, another HSRP and Layer 3.

Layer 2 Traffic Engineering

- Traffic engineering is a mechanism to utilize network bandwidth efficiently.
  - Mostly used in MPLS Traffic Engineering.
- Engineers may not aware that they are doing TE in Layer 2 networks.
  - There is no layer 2 traffic engineering topic in the books but it is done in the networks intentionally or unintentionally.

Layer 2 Traffic Engineering

- BGP, IGP, MPLS, Layer 2, we always try to do traffic engineering to use bandwidth efficiently, send different traffic to different links etc.
  - Voice traffic towards low latency links, data traffic through high capacity high latency links
- Vlan load balancing, VTP pruning, HSRP Groups are the examples of Layer 2 Traffic Engineering

Layer 2 Access Design

- If access and distribution layer connection is based on layer 2, then this topology is called as layer 2 access designs.

- It can be implemented as looped or loop free topology.

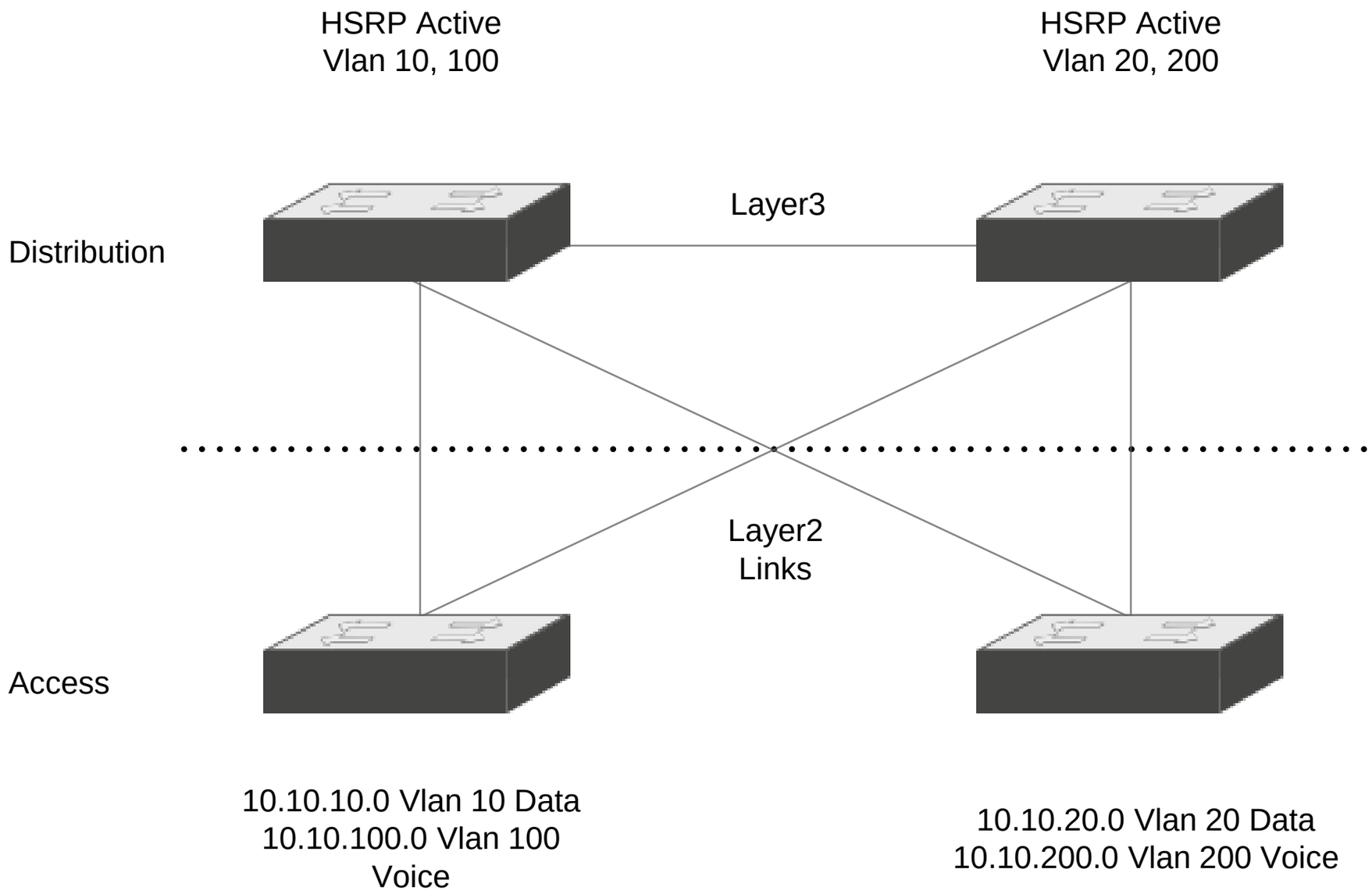
Layer 2 Access Design

- In loop free design, the link between distribution layer switches is layer 3 and same Vlan is not used in different access switches thus there is no loop in the topology so spanning tree doesn’t block any link.

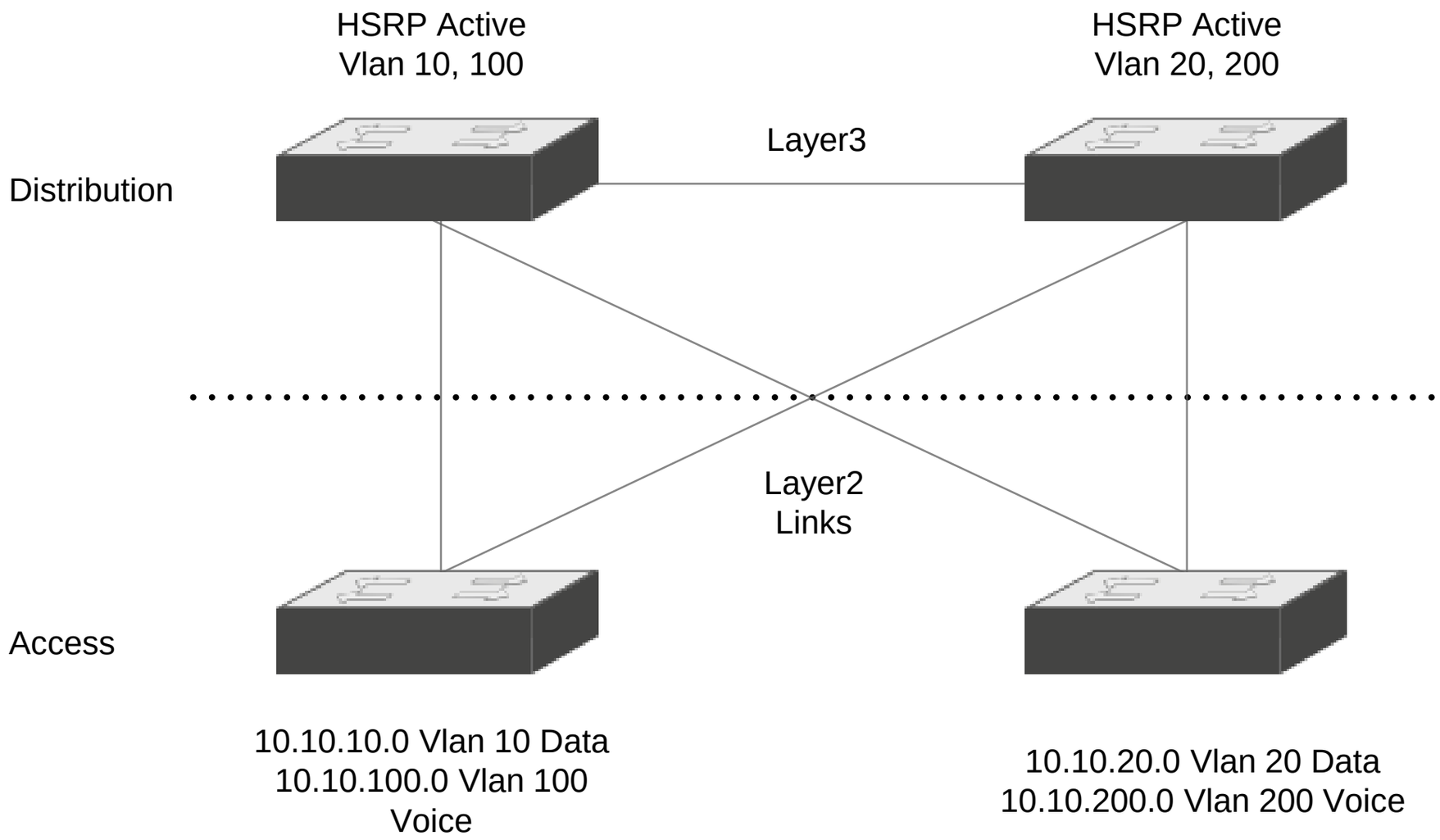
Layer 2 Access Design

- In both Layer 2 Looped and Layer 2 Loop free design, We need to have FHRP since we want to have more than one distribution switch for redundancy.

Layer 2 Loop Free Topology



Layer 2 Loop Free Topology

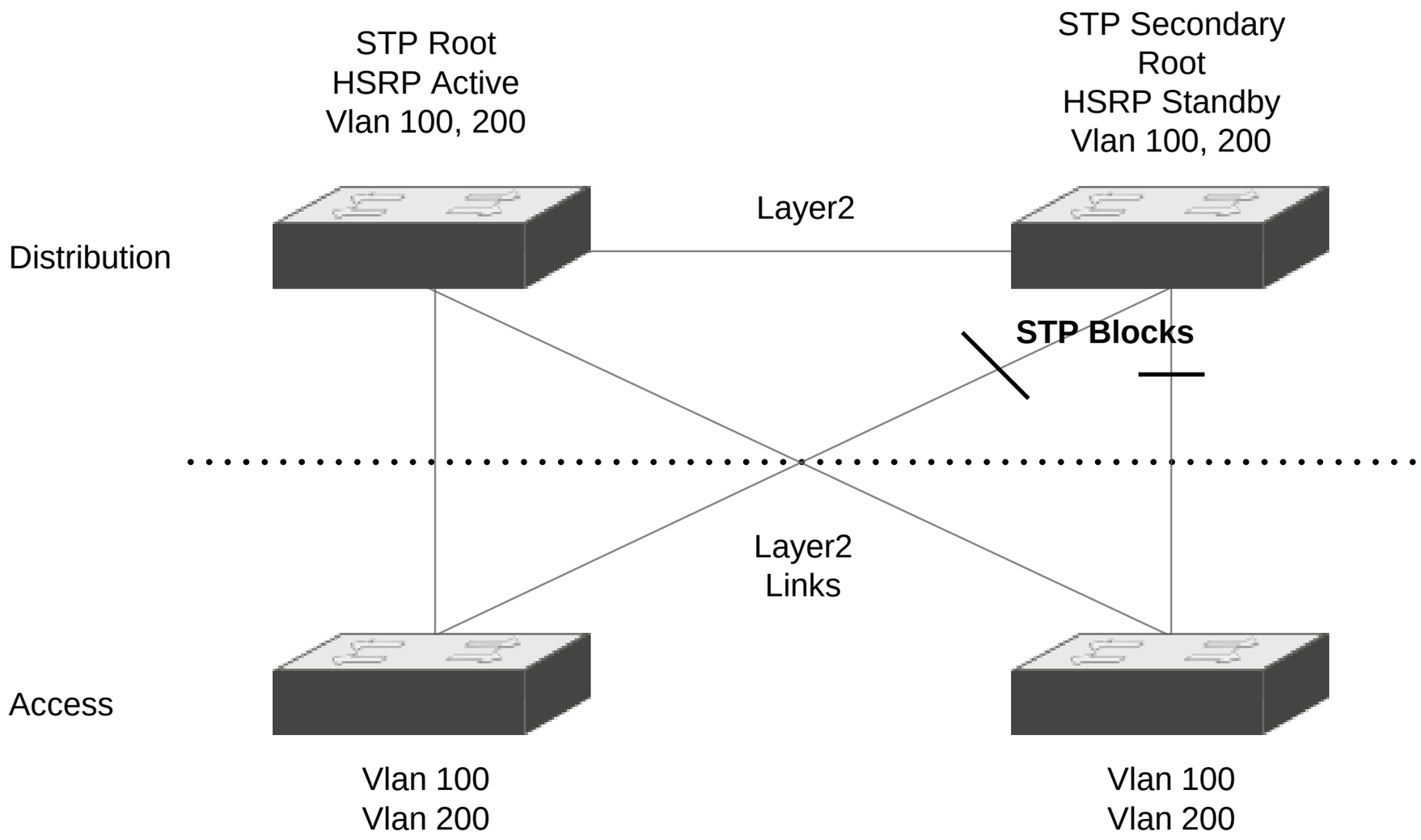


- For loop free topology FHRP BPDUs travel through access switch links.

Layer 2 Loop Free Topology

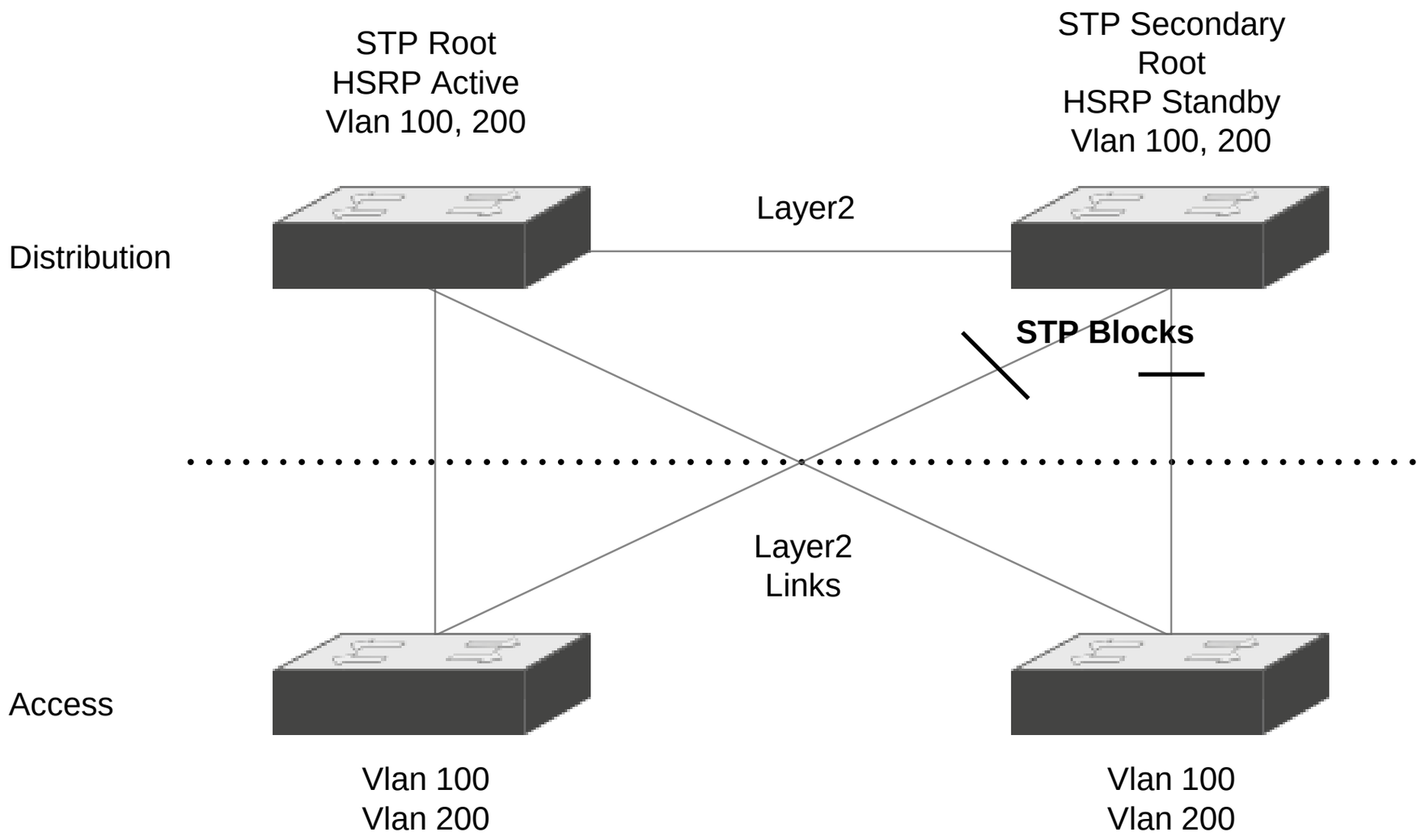
- In looped design, the link between distribution layer switches is layer 2 and same Vlan is used on different access switches so spanning tree will block one of the links to prevent loop.

Layer 2 Looped Topology



- We want to align STP Root with FHRP active and if we have network services device such as Firewalls, we want to align active firewalls with STP and FHRP

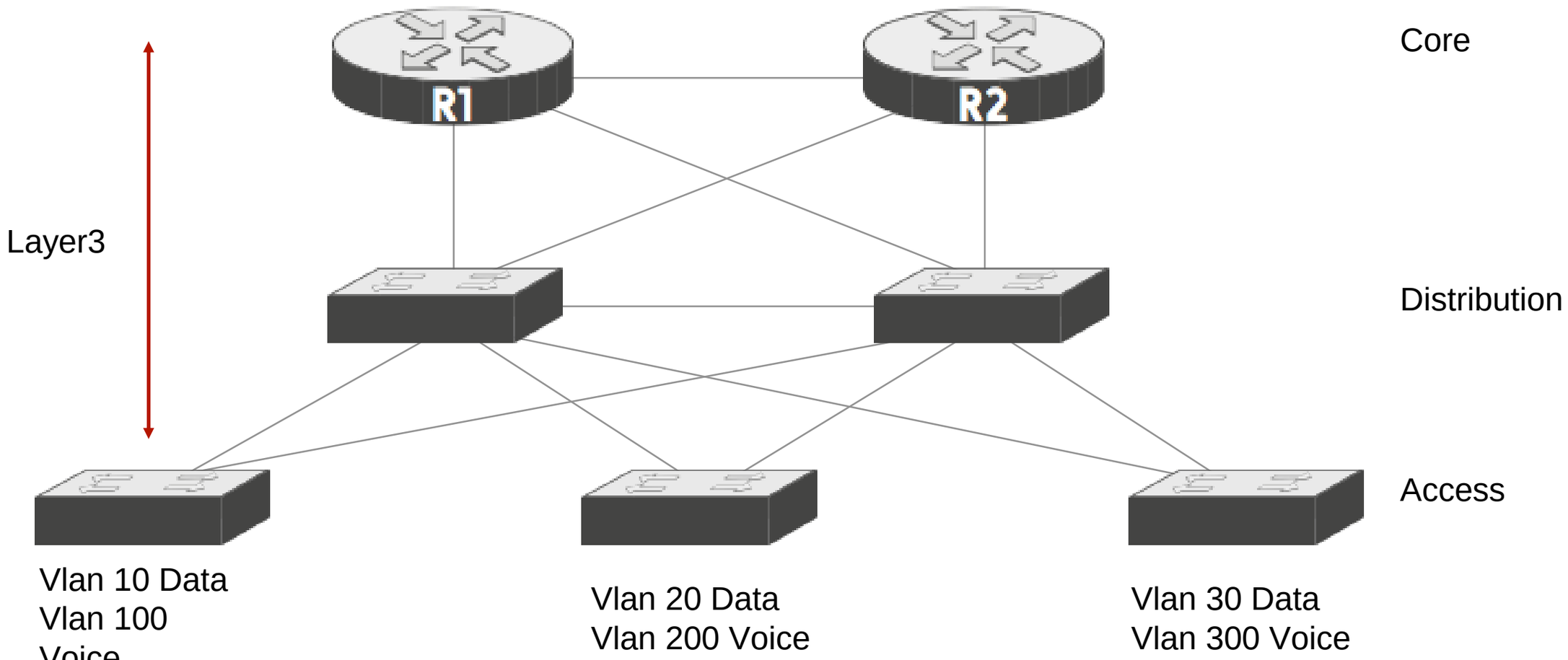
Layer 2 Looped Topology



- Same Vlan can be used on every access switch.
- Bi-sectional bandwidth usage is low

Layer 3 Access Design/Routed Access Design

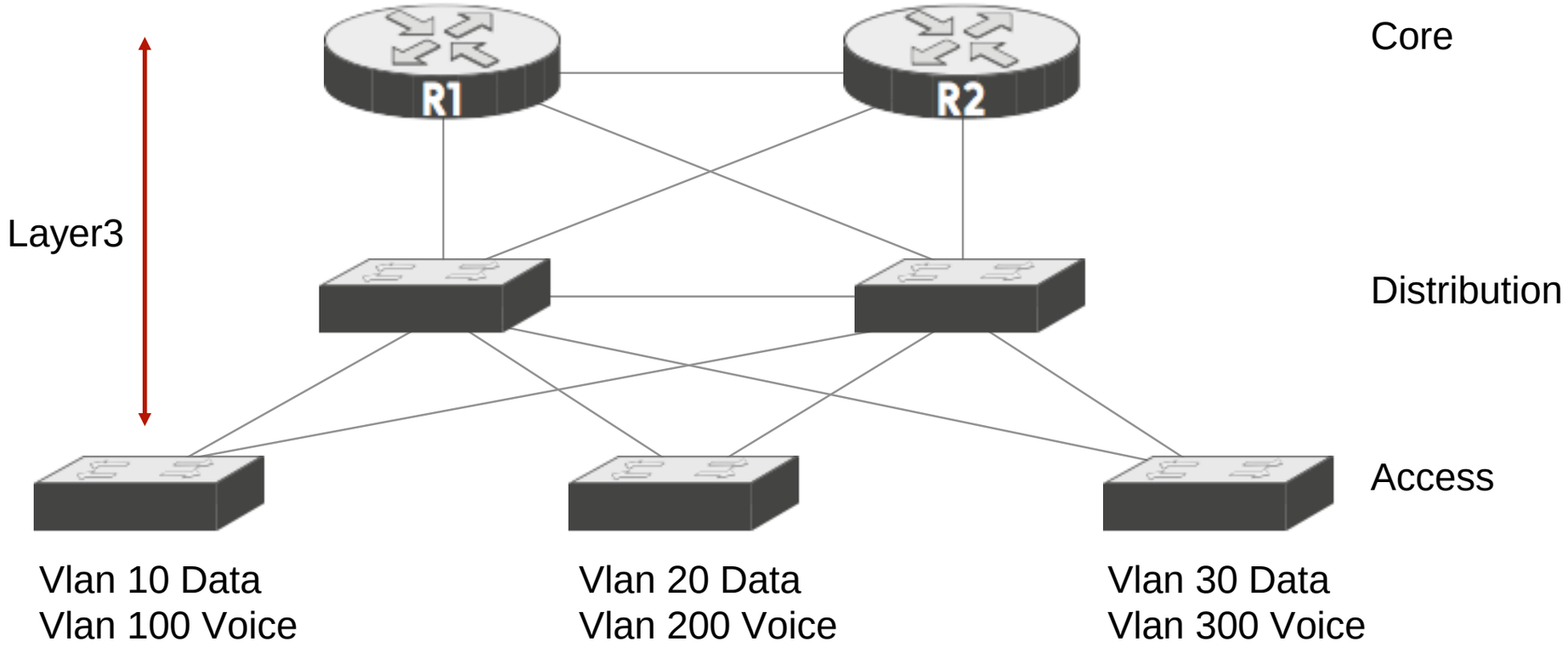
- It is also known as routed access design.
- The connections between access and distribution layer switches are layer 3, so first hop gateway of clients is access layer switch.



Unique Voice and Data Vlans

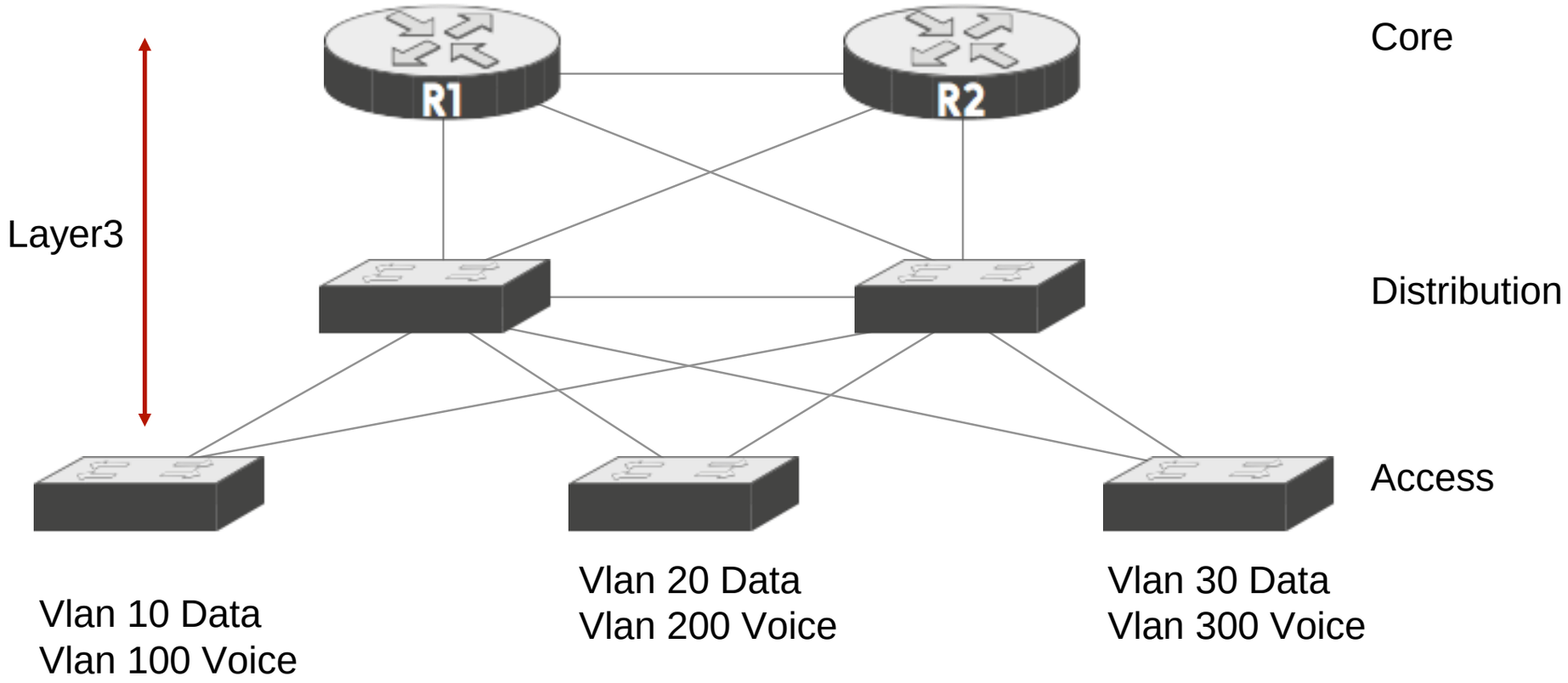
Layer 3 Access Design/Routed Access Design

- No need to have any first hop redundancy protocol since the access layer switch is the first hop gateway.
- There is no spanning tree in this design as there is no layer 2 link between the network devices



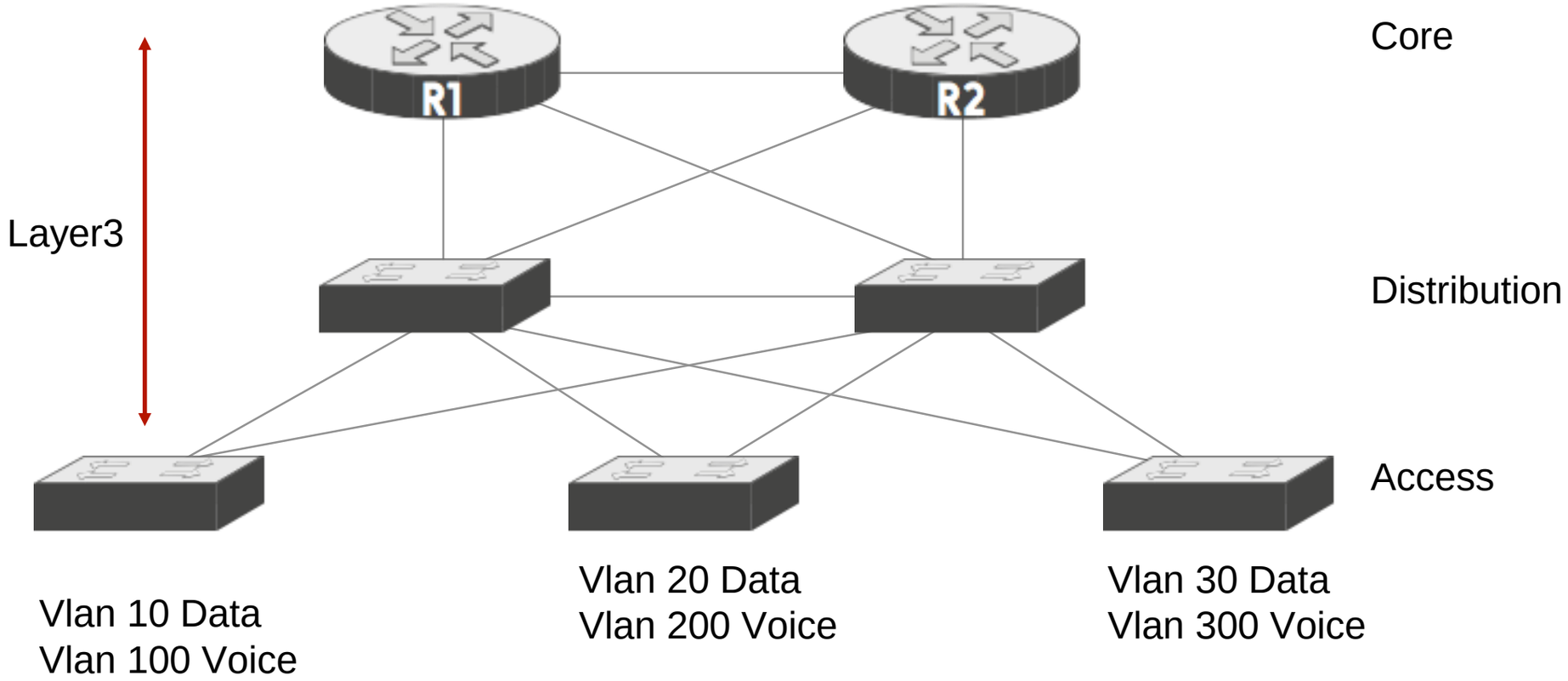
Layer 3 Access Design/Routed Access Design

- We can take advantage of fast convergence since we can use any IGP protocols between access and distribution layer and we can tune it, of course tuning protocol convergence time comes with its cost.



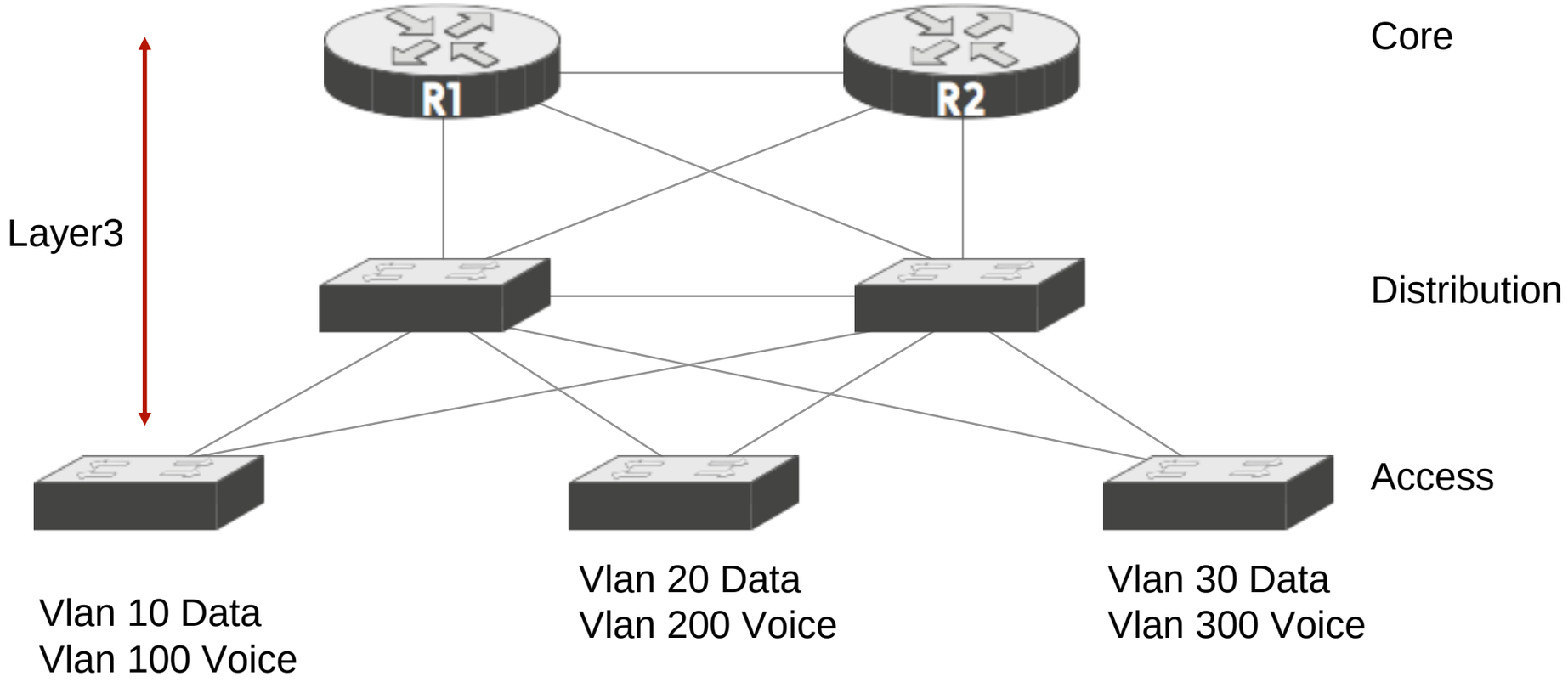
Layer 3 Access Design/Routed Access Design

- Tuning routing protocol for faster convergence, may impact overall stability of the network. You might have false positive.



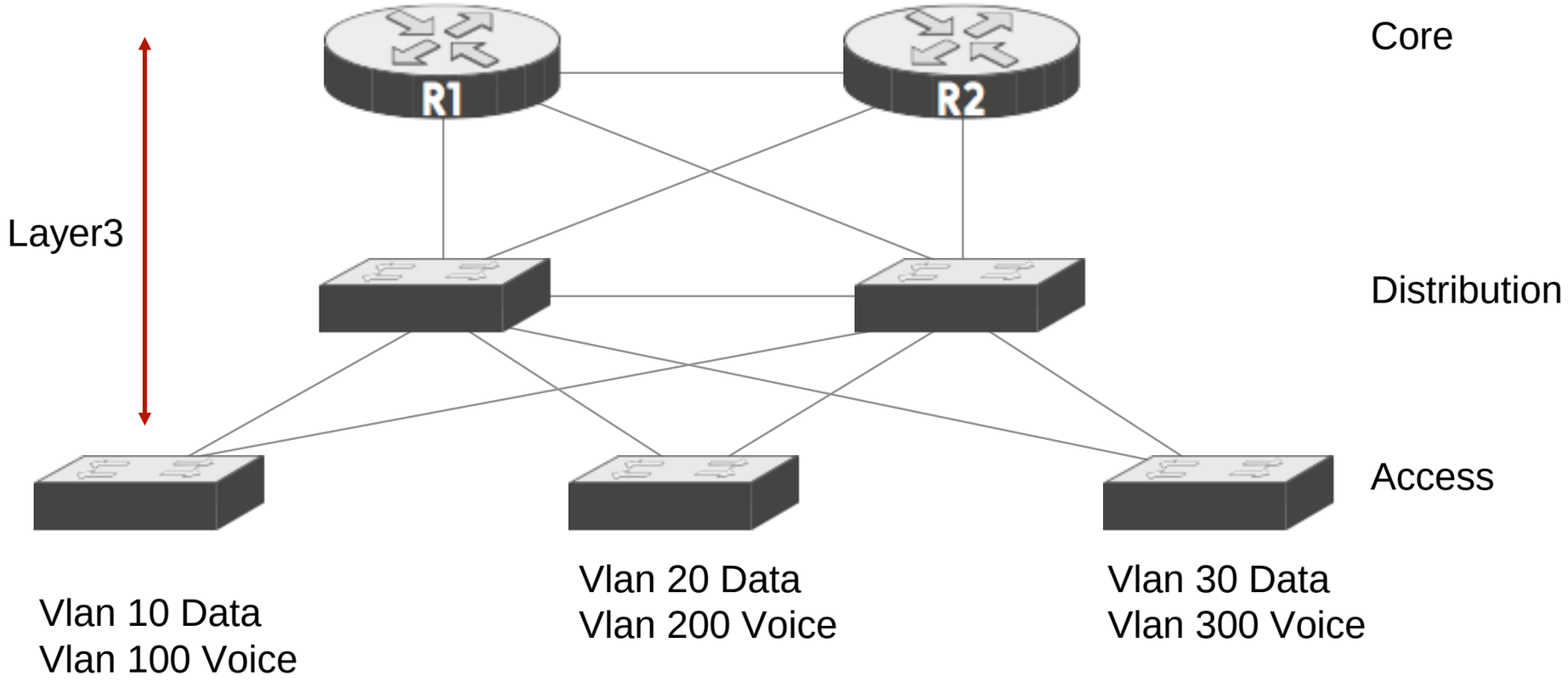
Layer 3 Access Design/Routed Access Design

- Also configuration will be much more complex.



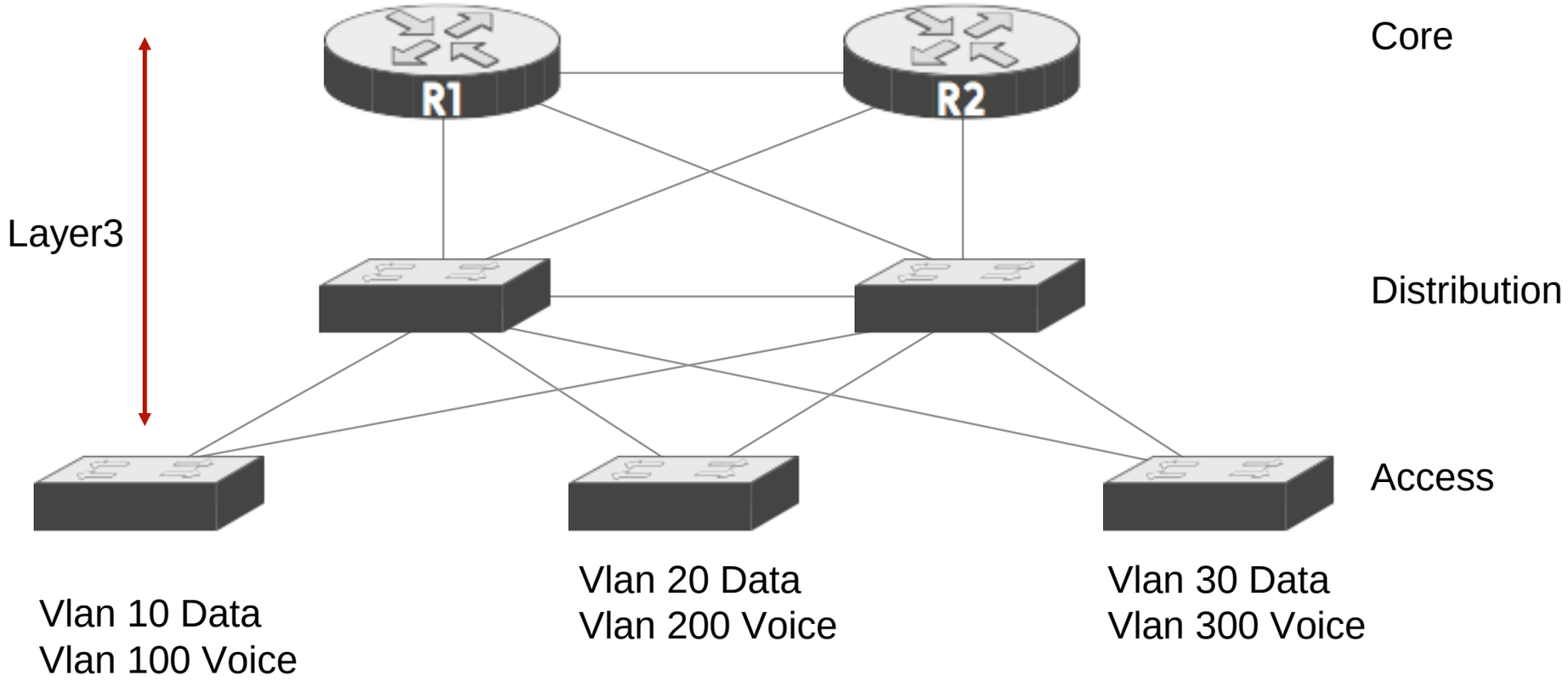
Layer 3 Access Design/Routed Access Design

- Although, there is no spanning tree anymore, still you may want to protect user site loop by enabling spanning tree at the edge towards user.



Layer 3 Access Design/Routed Access Design

- The drawback of this design, same Vlan cannot be used on the different access layer switches, at least for the campus network.



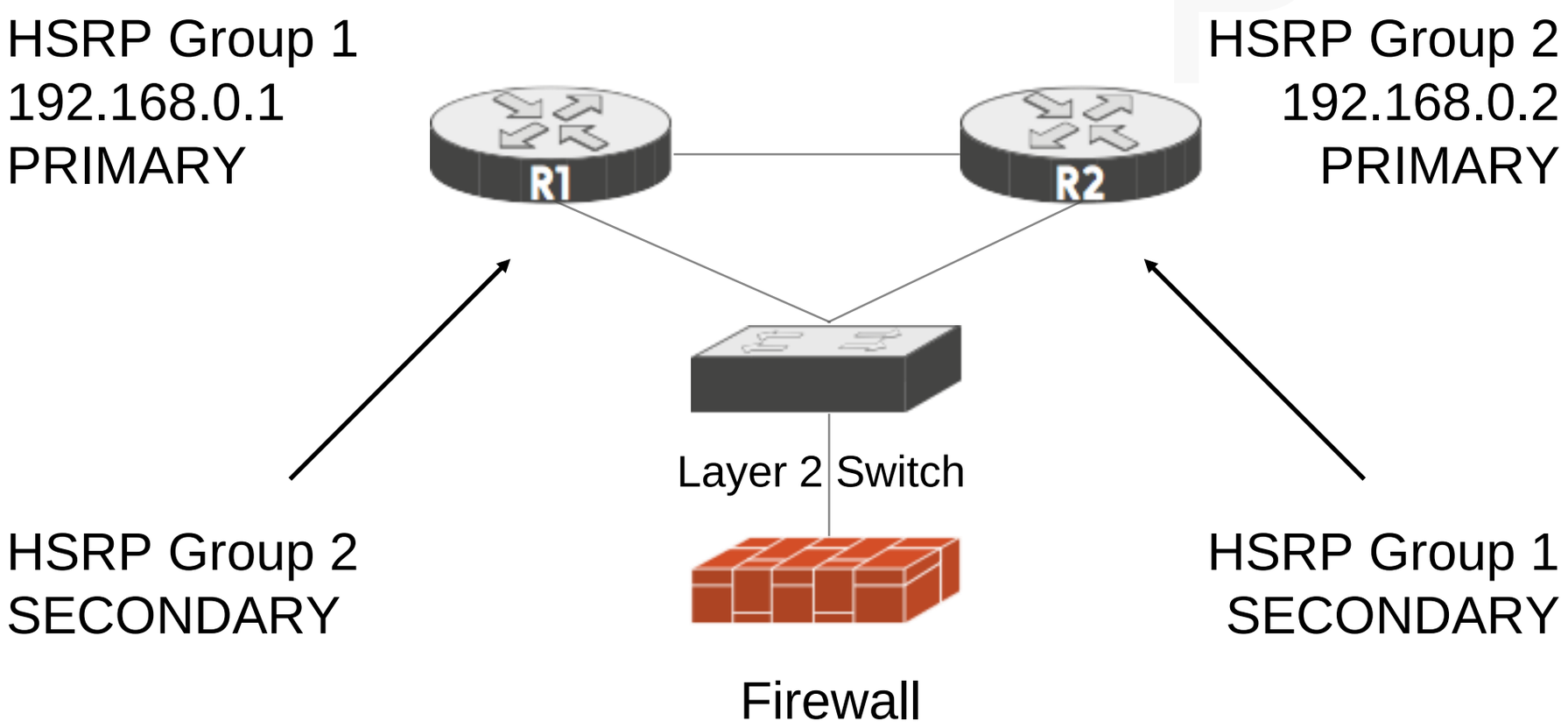
Layer 3 Access Design/Routed Access Design

- Host based overlays can be considered as similar to routed access design, in that case, since it is targeted for the datacenter and the vlan extension might be the requirement, host based overlays such as Vxlan, NvGRE, STT and Geneve support this.

- Host based overlays such as VXLAN, NvGRE, STT and Geneve will be explained in the VPN Design Course.

LAYER 2 CASE STUDIES

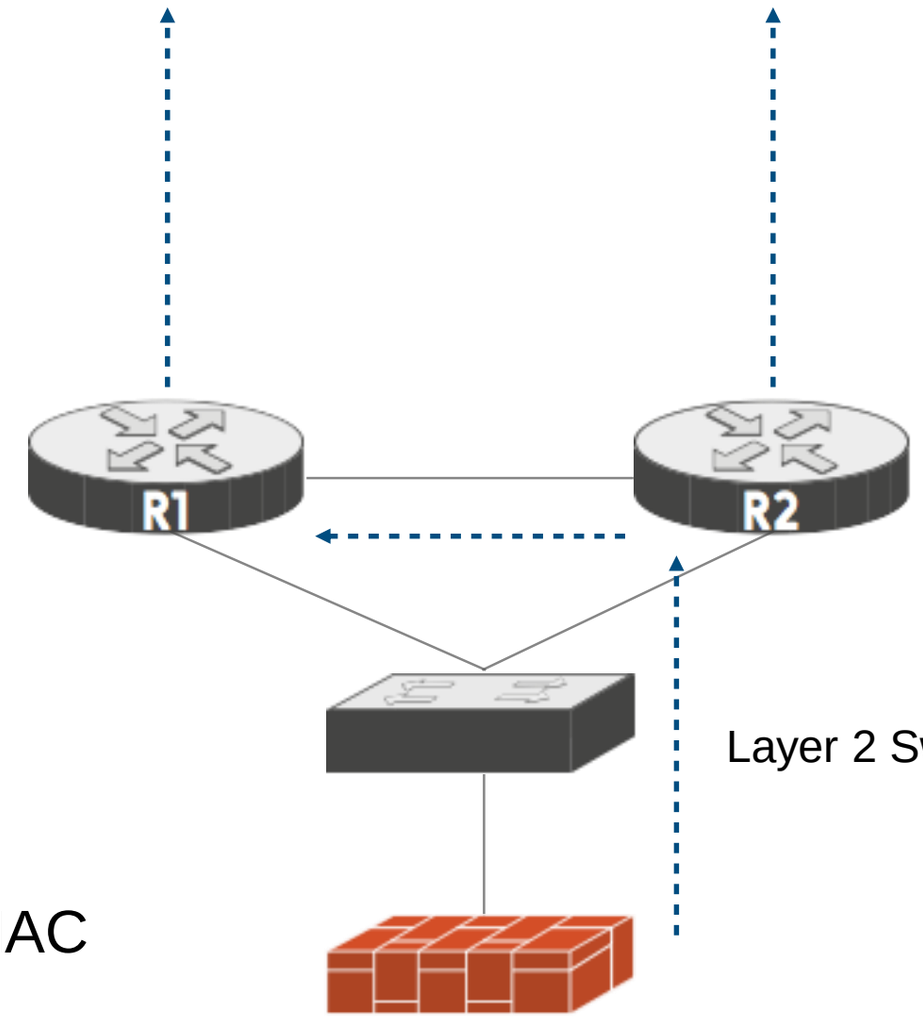
- Which one is more suitable for the internet edge, HSRP or GLBP?



- Create two HSRP Groups both routers; each router is active for one of the HSRP Groups egress from firewall: Static routes on FW to HRSP Group; BGP handles outbound forwarding.

- On the firewall default route is pointed to the both internet getaways. (We divide the default route to the half actually.)
  - First half of the default route is sent to the HSRP Group 1 address route outside 0.0.0.0 128.0.0.0 192.168.0.1
- Second half of the default route is sent to the HRSP Group 2 address route outside 128.0.0.0 128.0.0.0 192.168.0.2

Virtual MAC
0000.0000.0001



Virtual MAC
0000.0000.0002

Default Gateway MAC
0000.0000.0002

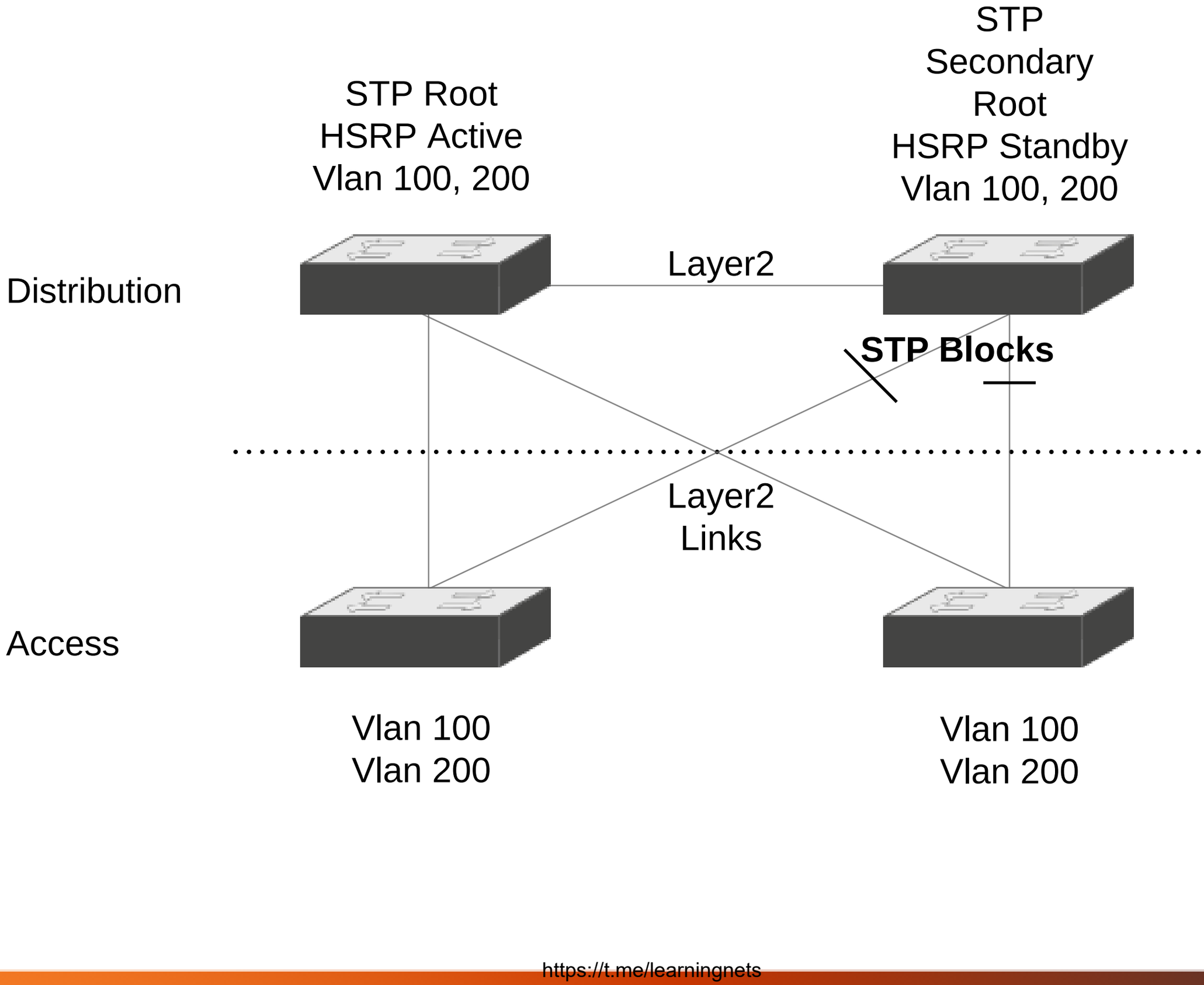
- What about Gateway Load Balancing Protocol (GLBP)?

GL
BP

- The firewall will perform ARP and the AVG (Active Virtual Gateway) will respond with Virtual MAC of either R1 or R2. Traffic is now polarized to a single link. More specific routes and use of local Preference is required for forwarding on both links.

- From the case study above, we can see that although HSRP might seem configuration wise more complex, traffic will not be polarized as in the case of GLBP.
  - In the GLBP case, one of the links from firewall to the Internet Gateway is not used. Only one link will be used.
- This may not be seen as a problem, because it also provides predictability.

Access Topologies



Q1: What is the name of this topology?
Q2: Is HRRP or GLBP more suitable. Why?

- The topology is called Layer 2 looped topology since the connection between two distribution layer switches is layer2. Once it is layer2, spanning tree has to block one link which is far from the root switch to prevent forwarding loop.

- Otherwise if you create a loop in Ethernet networks, since Ethernet doesn’t have TTL field in the header, unless you take manual action to stop loop, it continues to increase CPU of the device and the utilize the links.

- Different layer 2 protocols on Ethernet layer 1 media can work differently. Trill, Fabric path, SPB are some of them which you can find an article about on the website can use all the links without blocking.

- GLBP provides flow based load-balancing. If you are working in design field, you might heard this term often.

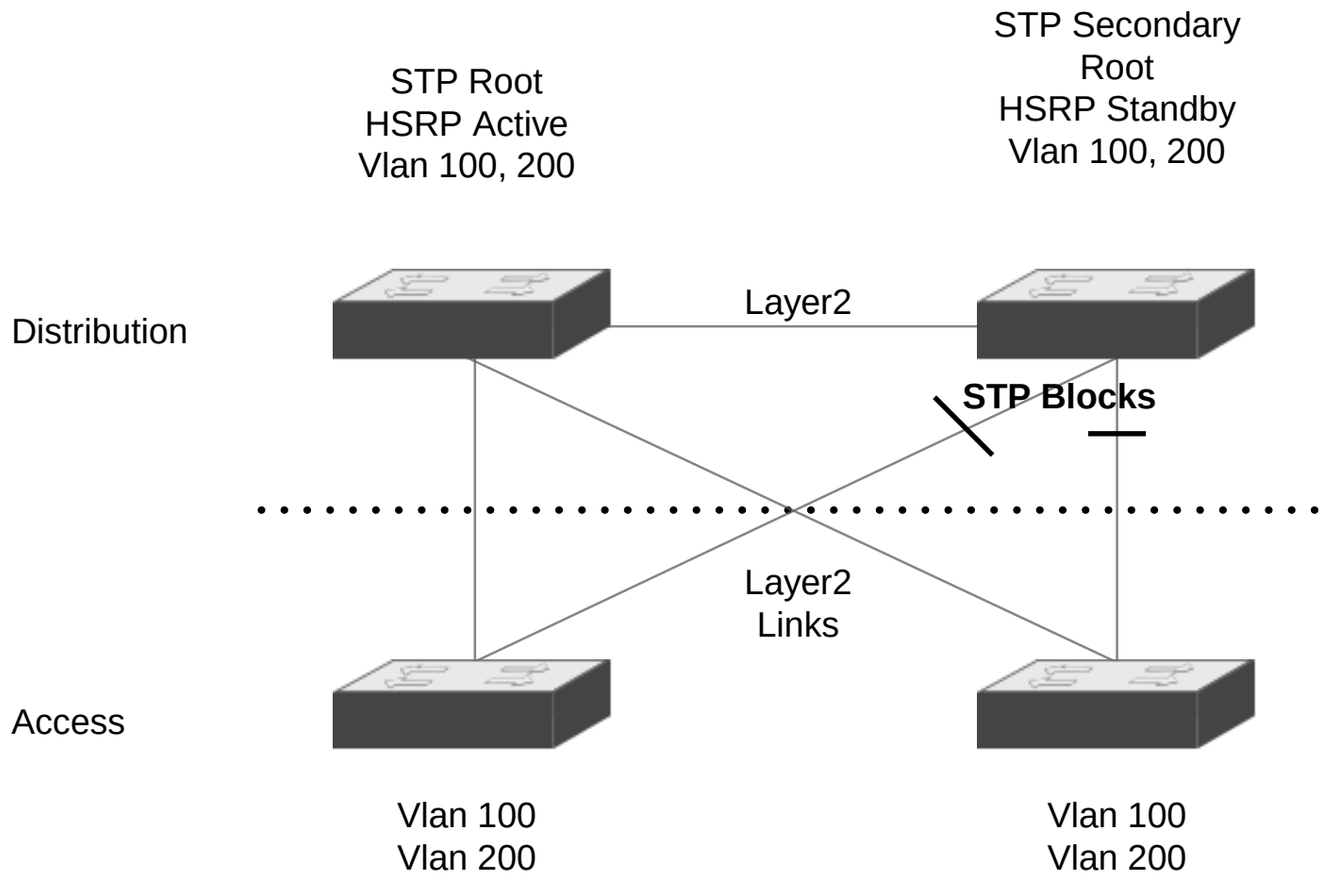
- Two common load balancing techniques are in the Layer 2 networks; Vlan based and Flow based load balancing.

- Vlan based load balancing allow the switch to be active layer 3 gateway for only some set of Vlans and other distribution stays as standby, and for the different set of Vlans standby switch acts as an active switch and active switch acts as standby. HSRP and VRRP works in this way.
  - Van 100 HSRP active gateway can be the left distribution switch, different Vlan let's say Vlan 101 can be the right distribution switch.
- Flow based load balancing mean is to allow both distribution switches to be used as an active-active for the same Vlan.

- Some users from the particular vlan use one distribution switch as an active default gateway and other users within the same vlan use previously standby switch as an active switch.
    - In this way you can use both distribution switches as active-active and you can utilize all the links in the layer 2 networks but supporting this configuration instead of using GLBP is more complex from the design point of view.
- If you want both right and left distribution switches to be used active-active for the same Vlan, let’s say Vlan 100, then you need to use GLBP.

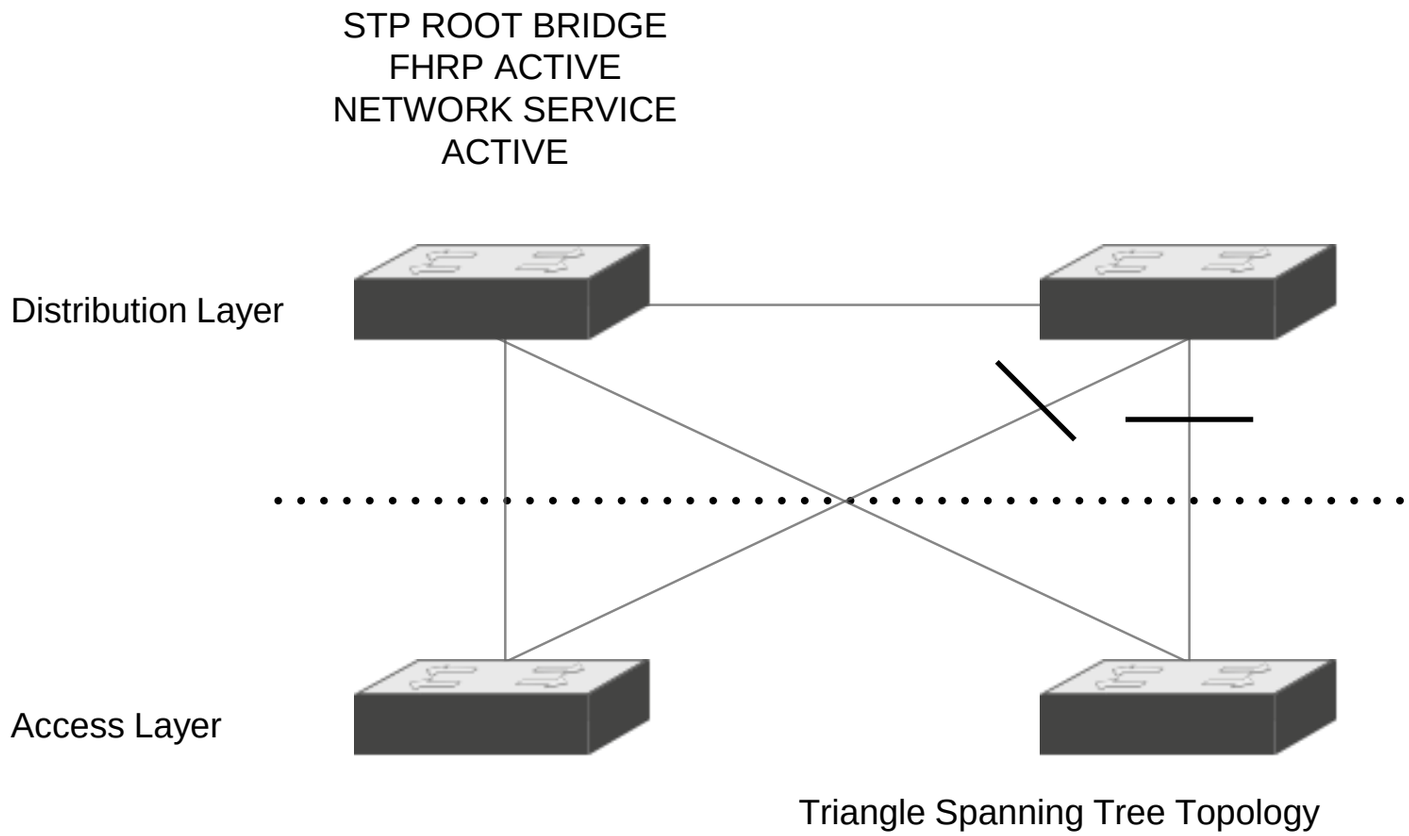
But spanning tree shouldn’t block the Layer 2 links. How you can achieve this?

- One way to change the inter distribution link to Layer 3. In that way none of the access layer links between access and distribution layer switches will be block, thus you can use all the uplinks.



- On the above topology if you use GLBP, since the right access to distribution link will be blocked, all the user traffic from the right access switch will go first to left distribution switch then through the interconnect link traffic will go to the right distribution switch since right distribution switch as an Active GLBP virtual forwarder replies to the ARP packets. That's why in this way always sub optimal path is used.

Why root switch is placed in the distribution layer instead of access?



Triangle Spanning Tree Topology

- In the figure above Access and Distribution layer switches are shown.

QUESTION

- Why do we always place Spanning tree root bridge and First hop redundancy protocol gateway at the distribution layer?

- Is it better if spanning tree root switch would be placed in the access layer?

SOLUTION

- Traffic pattern in the campus networks always in North-South direction. In two or three layer designs, layer 2 and layer 3 is placed on the distribution layer.

- Distribution layer is used for scalability, modularity and hierarchy.

- With distribution layer, any access layer switches can be upgraded smoothly. Also functions are shared between the access and distribution layer devices.
  - Access layer provides edge functions such as filtering, client access, QoS, security features.
- Distribution layer responsible from the route and traffic/speed aggregation.

- Layer 3 starts at the distribution layer. Thus first hop redundancy protocols are enabled at the distribution layer.

  - Since the user traffic from the campus environment reach to Internet, Servers are in the datacenter network which is a centralize place, traffic pattern is in north south direction, not east-west.

Thus it is logical to place spanning tree root, first hop redundancy protocol gateway at the top position at the network.

- One can question then in the three layer hierarchy, can we put the root functionality into the Core layer?
  - Answer is yes but it may not be a good design. In that case, layer 2 domain would be much larger and we always want to keep layer 2 domain small unless the application requires it to be much larger such as Vmotion, layer 2 extension and so on.

First Hop Redundancy Protocol Damage

- An Enterprise company will implement first hop redundancy protocol on their distribution switches.

- The requirement is if the failure happen at the distribution switches, they don’t want all the users in a given vlan is effected from the failure. Some users in that Vlan should still be able to operate.

Question: Which first hop redundancy company should use and why?

- As it is indicated in the book earlier, only one device is used as an active gateway with HSRP and VRRP.

- If failure happens standby device takes responsibility and even with fast hellos and BFD still will be down time. During convergence clients traffic will be effected.

- But with GLBP, in a given Vlan, there can be more than 2 active gateways. That's why clients traffic can be divided among the active gateways.
  - If failure happens in a GLBP enabled network, only some of the clients traffic in a given vlan is effected. If there is two active gateways, only half of them will be affected.
- Thus for the requirements given in the question, GLBP is the best choice.

Port-Security comes to the rescue

- In the conference room of the company, contractors connected a device, which doesn't generate spanning tree BPDU with two ports to the existing switch environment.

QUESTION 1: What would be the implication of this?
QUESTION 2: How can future problems be mitigated?

- This problem has happened in the early days of networking. Hubs don’t generate a spanning tree BPDUs. If you connect a hub with two ports to a switch, forwarding loop occurs.

- In order to stop it, you can remove one of the cable for sure. But if the contractor would know the implication probably they wouldn’t connect it at the first place.

- That’s why a future which can prevent a loop should be in place in advance.
  - BPDU Guard and BPDU Filter are the two features, which react based on bpdu.
- BPDU Guards shutdown the switch port if spanning tree bpdu is received from the port.
  - BPDU Guard doesn’t shutdown the port but can give some information about the bpdu.

- But in the requirement of this case study it is clearly told that bpdu is not generated. In this case Port-security helps.
  - Port-security doesn’t allow two mac addresses to be shown on the two ports of a given switch. If it happens, port-security feature shutdown the ports.
- That’s why it is one of the best practices to enable port-security not only as security feature but spanning tree feature as well.

Layer 2 Looped Design Use Case

- Where would Layer2 looped design be better from the Layer 2 campus network design point of view?

Answer : In an environment where layer 2 Vla needs to be spanned. Classical example is the datacenter.

- In the data centers hosts (Specifically Virtual Machines) can move between access switches. Vlans should be spread on those switches.
  - Also it is very common in the campus environment where WLAN is used commonly on every access switches.
- In these environments where layer 2 needs to be extended on many access switches, layer 2 looped design is the only design option.

NETWORK DESIGN TOOLS & THE BEST PRACTICES

Agenda

- Reliability
- Resilience
- Fast Convergence and Fastreroute
- Scalability
- Cost :Opex and Capex
- Flexibility
- Modularity

Agenda

- Design Considerations for Network Mergers and the Acquisition
- High Availability
- Convergence
- Some Routing Best Practices
- Load Balancing
- Redistribution
- Optimal Routing
- Network Topologies
- Security
- Simplicity and Complexity

- There are design tools, which we should consider for every design. LAN, WAN and the data center where this common design tools and attributes should be considered

Reliability

- Within the reasonable amount of time, which depends on the application type and architecture, delivering the legitimate packets from source to destination.

reli
abil
ity

  - This time is known as delay or latency and it is one of the packet delivery parameters. Consistency of delay known as jitter and it is very important for some type of applications such as voice and video, jitter is our second delivery parameters

Reliability

- Third packet delivery parameter is packet loss or drop; especially voice and video traffic is more sensitive to packet loss compare to data traffic.
  - Packet loss is application dependent and some applications are very drop/packet loss sensitive.

reli
abil
ity

Effect of packet loss on IPTV Service

Reliability

- General accepted best practices for the delay, jitter and packet loss ratio has been defined and knowing and considering them is important from the network design point of view.

reli
abil
ity

- For example for the voice packets one way delay which is also known as ‘mouth to ear’ delay should be less than 150ms.

Reliability

- Reliability should not be considered only at the link level. Network links, devices such as switches, routers, firewalls, application delivery controllers, servers, storage systems and others should be reliable; also component of these devices needs to be reliable.

reli
abil
ity

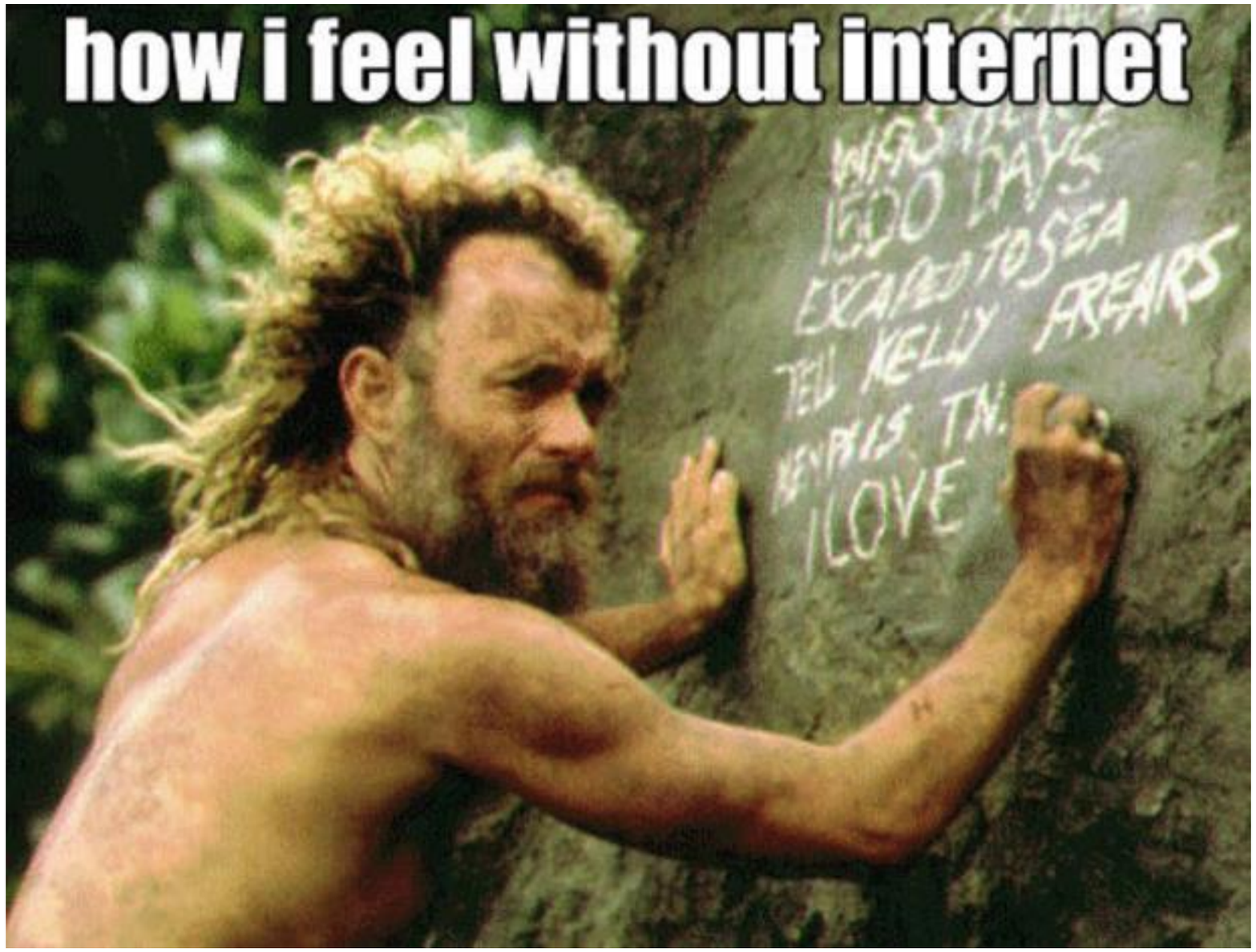
Reliability

- For example, if you will carry the voice traffic over unreliable serial links, you may likely encounter packet drops because of link flaps. Best practice is to carry voice traffic over the low latency links, which don’t have packet loss and the latency.

reli
abil
ity

- If you have to utilize those cheaper unreliable links such as Internet, you should carry the Data traffic over them.

Internet is not reliable, it’s no effort (not a best effort) but we can’t live without internet!



Reliability

- But actually whichever device, link or component you choose, essentially they will fail.

reli
abil
ity

- Vendors share their MTBF (Meantime between failure) numbers. You can choose the best reliable devices, links, component, protocols and architecture; you need to consider unavoidable failure. This brings us the resiliency.

Resiliency

- Resiliency is how the network behaves once the failure happen. Is that highly available, will it convergence and when?
    - Resilience can be considered as combination of high availability and convergence.
- In order any network design to be resilient, it should be redundant and converge fast enough to avoid application timeout.

Resiliency

- Thus, resiliency is interrelated with redundancy and the fast convergence/fast reroute mechanisms.
    - Every component and every device can and eventually will fail, thus system should be resilient enough to re converge/recover to a previous state. As it is stated above; Resiliency can be achieved with redundancy.
- But how much redundancy is best for the resiliency?

Sometimes one is enough - But in networking, generally two is company, three is crowded!

Fast Convergence & Reroute

- Network reliability is an important measure for deployability of sensitive applications.

- When a link, node or SRLG failure occurs in a routed network, there is inevitably a period of disruption to the delivery of traffic until the network reconverges on the new topology.

Fast Convergence & Reroute

- Fast reaction is essential for the failed element. There are two approaches for the fast reaction: Fast convergence and fast reroute.

- When a failure occur four steps are necessary for the convergence.

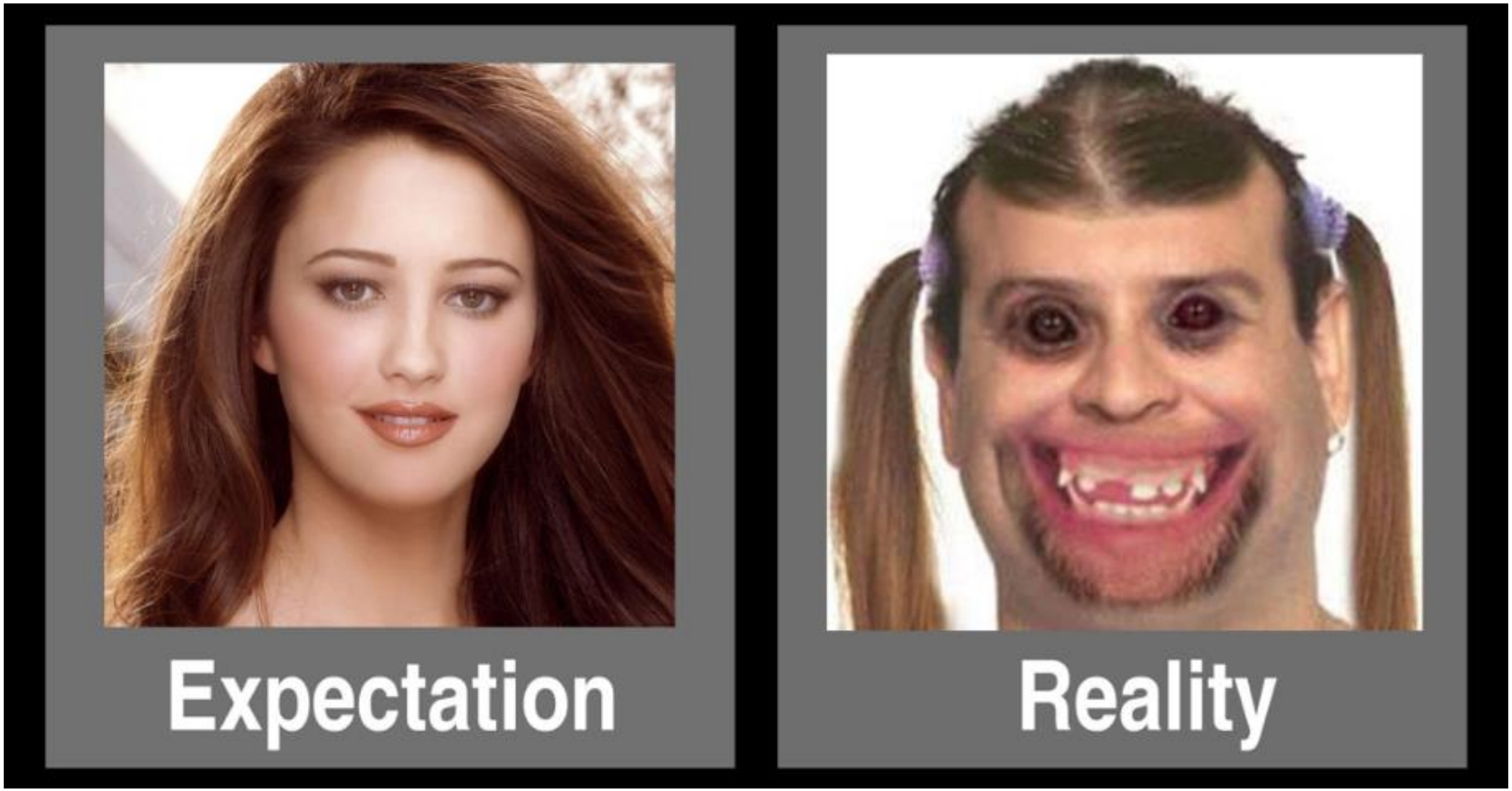
  1. Failure detection
  2. Failure propagation
  3. New information process
  4. Update new route into RIB/FIB

Fast Convergence & Reroute

- For fast convergence, these steps may need to be tuned. Although the RIB/FIB update is hardware dependent, the network operator can configure all other steps.

rer
out
e

- One thing always needs to be kept in mind; you might expect to have Fast convergence and fast reroute but reality it can affect network stability.



Fast Convergence & Reroute

- Unlike fast convergence, for the fast reroute, routes are pre-computed and pre-programmed into the router RIB/FIB.

- There are many Fast Reroute mechanisms available today. Most known ones are; Loop Free Alternate (LFA), Remote Loop Free Alternate (rLFA), MPLS Traffic Engineering Fast Reroute and Segment Routing Fast Reroute.

Scalability

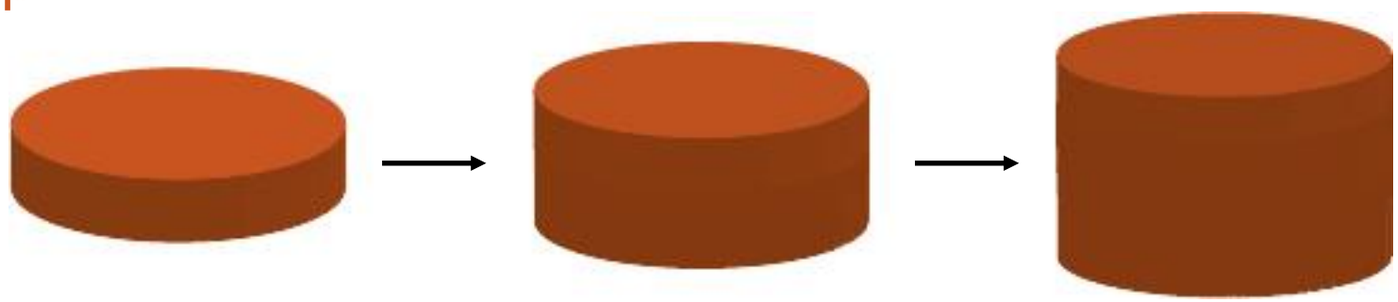
- Scalability is the ability to change, modify or remove the part of entire system without having a huge impact on the overall design.

sca
labi
lity

- There are two scalability approaches for the IT systems. These approaches are scale up or scale out and implies for the Network, Compute, Storage, Application, Database and many other systems.

Scale up & Scale out

Scale up



Scale out

Scalability

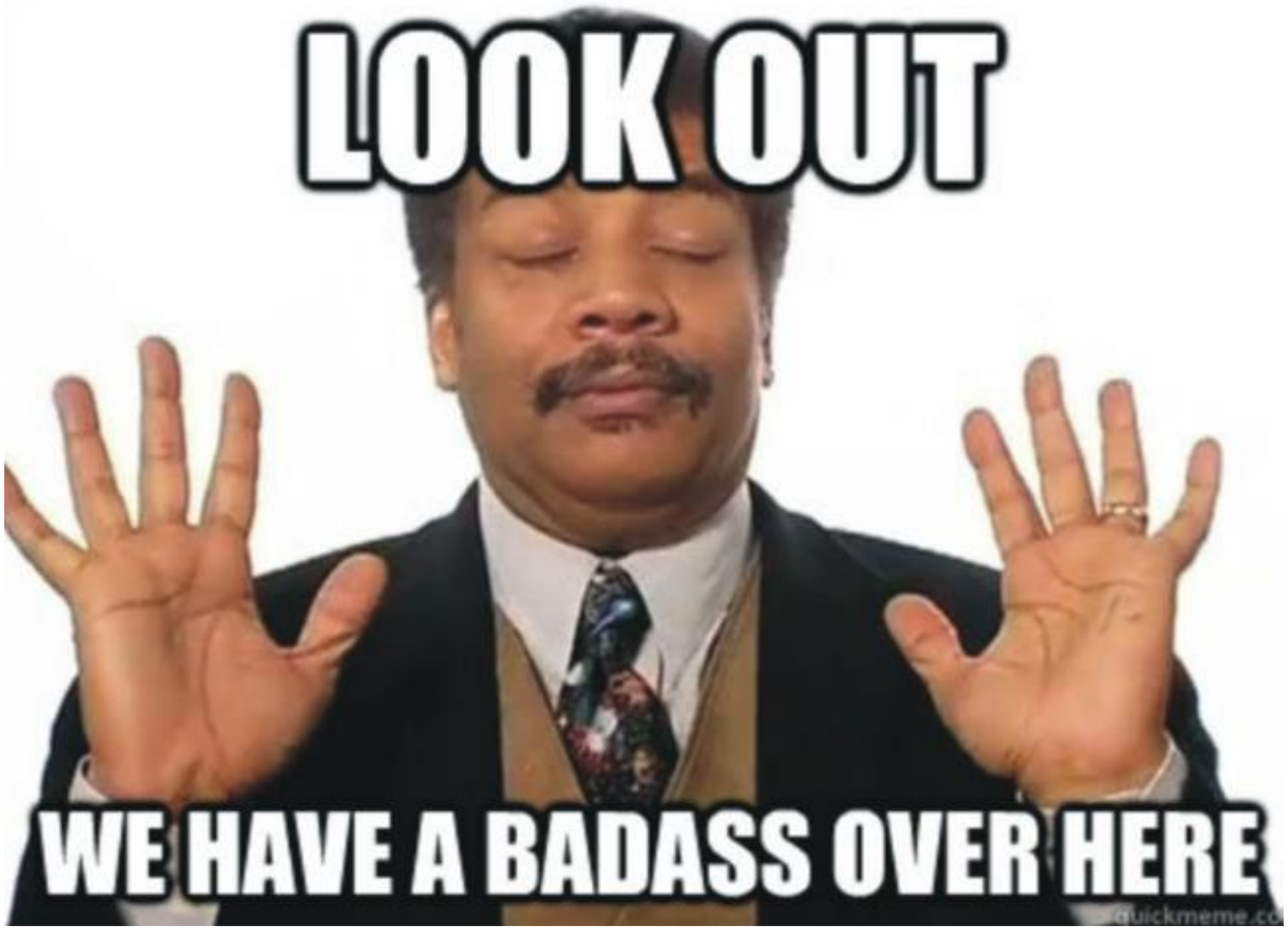
- Scalability through scaling up the system can be defined as to increase the existing system resources without adding a new system.
    - Consider scale out application architecture, if application can be run over the two different servers, we can do some maintenance on the one of the servers without affecting the user experience.
- Instead of having a highly redundant one physical chassis based router, having two routers.

Scalability

- Scalable systems also should be manageable. While growing in size, if system starts to become non manageable, it will affect inversely the scalability of the system.
    - In order to perform any change, network shouldn’t need flag days, long and frequent maintenance windows because of operationally complex environment. (OPEX)
- This brings us to the next design tools, which is COST.

Cost-Opex & Capex

- Cost is generally afterthought in network design. But most of the time it is very important constraint in the projects. If we breakdown the components of cost we have OPEX and CAPEX.
  - OPEX refers to operational expenses such as support, maintenance, labor, bandwidth and utilities.
- Creating a complex network design may show off your technical knowledge but it can also cause unnecessary complexity making it harder to build, maintain, operate and manage the network.



Cost-Opex & Capex

- A well- designed network reduces OpEx through improved network uptime (which in turn can avoid or reduce penalties related to outages), higher user productivity, ease of operations, and energy savings.
    - Consider creating the simplest solution that meets the business requirements.
- In the CCDE exam, which solution is best from the OPEX point of view type of questions are common!

Cost-Opex & Capex

- CapEx refers to the upfront costs such as purchasing equipment, inventory, acquiring intellectual property or real estate.

cost
ope
xan
dca
pex

- A well-thought design provides longer deployment lifespan, investment protection, network consolidation and virtualization, producing non-measurable benefits such as business agility and business transformation and innovation, thus reducing risk and lowering costs in the long run.

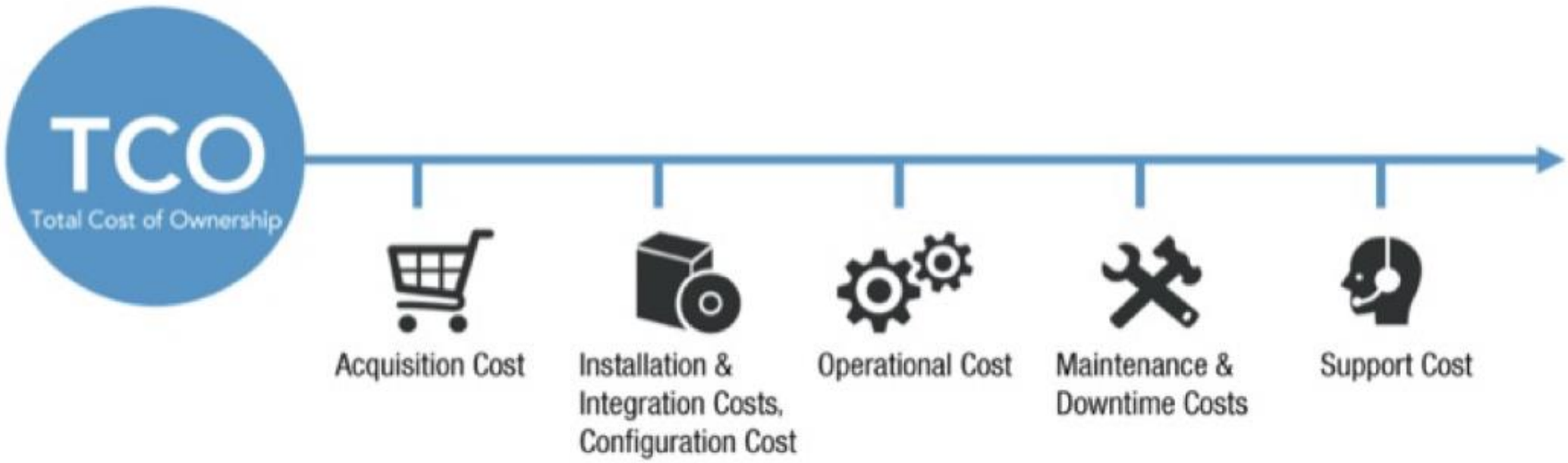
TCO - Total cost of ownership

- Last metric in the COST constraint is TCO (Total cost of ownership)

cost
ope
xan
dca
pex

- TCO is a better metric than pure CapEx to evaluate network cost, as it considers CapEx plus OpEx. Make your network designs cost-effective in the long run and do more with less by optimizing both CapEx and OpEx.

TCO - Total cost of ownership



Flexibility

- Flexibility refers to the ability of a network design to adapt to business changes, which can come in a planned or unplanned way.

flexibility

- There are a few constants in life: death, taxes and change – not much we can do about the first two, but certainly we can influence how to adapt to change.

Flexibility

- Merger, acquisition, divesture can happen anytime in any business. How your network will react to these rapid changes?

flexibility

- You can make network design more flexible by making it more modular.

Modularity

- Modularity means to divide the network by functions or policy boundaries, making it replicable (for example on branches) and thus easier to scale and operate, and enabling business continuity.

modularity

- In three ways you can make design modular so you can provide flexibility.

Modularity

- Choose the physical topology: Some topologies such as hierarchical or leaf & spine are more conducive to allow for modules than others (fully meshed, for example).

Modularity

- Split functions or geographies: Separate campus, branches, data center and applications, Internet, network management systems, and security policy boundaries to make each function easier to expand, upgrade, enhance or change. Make them small enough to ease replication.

Modularity

- Break it into smaller pieces: Create smaller fault domains so that a failure on a part of the network doesn't propagate to other parts, by subdividing the functions as appropriate.

Modularity

- Modular design allows different modules to be managed by different teams. In the Service Provider networks this is common. Access, Aggregation and Core networks are modular and they generally managed by individual teams.

Modularity

- Modular design reduces deployment time since the same configuration is used for the new module, same physical topologies are used and so on.

Design Considerations For Network Mergers & Acquisitions

- Network mergers and acquisitions are the processes, which can be seen in any type of businesses.
    - As a network designer, our job to identify the business requirements of both existing networks and the merged network and finding best possible technical solutions for the business.

Design Considerations For Network Mergers & Acquisitions

- Network mergers and acquisitions are also called as Network integration.

- Business and network analysis and technical information gathering are the key steps and there are many questions which need to be asked and answered should be well understood.

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Business analysis and information gathering.
    - Applications of the company, at least the business critical applications should be understood and analyze very well.
- What are the capabilities of these applications and what are the requirements from the existing network.

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- What type of WAN, LAN and DC infrastructure each network is using?

- What is the convergence time of the network and what is the required convergence time of any single component failure?

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Analyzing the design for network mergers and acquisitions is not different than analyzing the design for the greenfield network.

- Application and business requirements are always the most important, technology is second. Alternative technologies always can be found.

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Where will be the first place in the network for the merger?
- What happens to the routing, IGP, BGP? Is a “ship in the night” approach suitable or redistribution is better for routing protocol merger?

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Which type of security infrastructure will merged network support?

- What is the Quality of Service Policies of the companies? Will final merged network support Quality of Service or through bandwidth everywhere?

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Will merged network have IPv6?
    - Does one of the networks require Multicast? Will merged network support Multicast?
- What is the new capacity requirement of the merged network?

Key Points For Any Type Of Network Mergers & Acquisitions Projects

- How will be the merged network monitored? Do exist Network Management tools capable to support all the technologies/protocols of the both network?

- When you divest the network, where will the datacenters be? Can you decommission any datacenter, POP location for cost optimization?

High Availability

- Availability of a system is mainly measured with two parameters. Mean time between failure (MTBF) and Mean time to repair (MTTR). MTBF is calculated as average time between failures of a system. MTTR is the average time required to repair a failed component (link, node, device in networking terms).

High Availability

- Too much redundancy increases the MTTR of the system (Router, Switch or overall network) thus inversely effect the availability.

- Most failures are caused by human error. In fact, the estimated range is between 70 and 80%.

High Availability

- Not every network needs 5x9 or 6x9 availability. Before deciding upon the availability level of a network design, understand the application requirements and the place where the design will be applied on the network.

Convergence

- Don’t use Routing Protocol hellos for the Layer 3 routing failure detection, at least don’t tune them aggressively, instead leave with the default. Use BFD whenever possible for failure detection in Layer 3.

con
ver
gen
ce

- BFD supports all routing protocols except RIP. It supports LDP and MPLS Traffic Engineering as well.

Convergence

- If you can detect the failure in Layer 1, then don’t enable BFD (Bidirectional Forwarding Detection) as well.

- Pooling-based mechanisms are always slower than event-driven mechanisms. For example, Layer 1 loss of signal will be much faster than BFD hellos; Automatic Protection Switching (APS) on SDH links are always faster than BFD hellos for failure detection.

Convergence

- Distance Vector Protocol converge time is the same as Link-State Routing Protocols. If there is a feasible successor in the EIGRP topology, EIGRP by default converges faster than other routing protocols.

convergence

- BGP doesn’t have to converge slowly. Understanding the data plane and control plane convergence difference is important for network designers.

Convergence

- BGP route reflector is not always the best solution. It hides the available alternate next-hops, slows down the network convergence, and requires route reflector engineering thus requires stuff experience.

convergence

Some Routing Best Practices

- In modern platforms there are software and hardware forwarding information tables. Software forwarding is a very resource-intensive task; utilize hardware forwarding if you need better performance.

Some Routing Best Practices

- Multi-area OSPF or multi-level IS-IS design is not always necessary but you should know what business problem you are trying to solve. Resiliency? Opex? Security? Reliability? Scalability? In general Scalability is considered as a reason to deploy Multi Area OSPF or Multi Level IS-IS. These two topics will be covered in great detail in the related lessons.

Some Routing Best Practices

- Try to find a way to deploy any technology, feature, or protocol with the least amount of configuration possible. If you can achieve the same result with lesser configuration steps, prefer that one.

  - For example instead of having a BGP configuration for each BGP peer, use peer-group template.

Load Balancing

- Load balancing and load sharing is not the same thing. Load sharing terminology should be used for the routers or switches but load balancing requires more intelligence such as Load balancer.

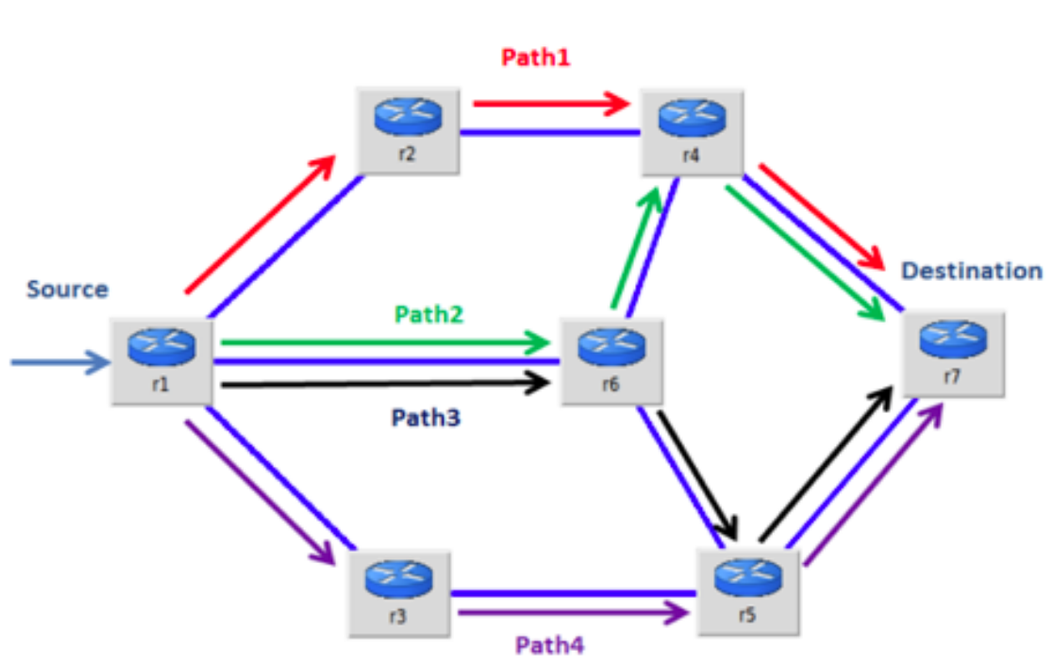
load balancing

- Load balancing is any intelligence feature that devices need to support, such as destination health check, considering destination device resource utilization, the number of connections, etc.

ECMP (Equal Cost Multipath)

- All routing protocols, except BGP by default supports ECMP (Equal Cost Multipath).
    - OSPF and IS-IS can do the unequal cost load sharing with the help of MPLS-Traffic Engineering only.

load balancing



Redistribution

- You may need to redistribute routing protocols. You may have a partner networks or BGP into IGP for default route advertisement.

redistribution

- Redistribution should be used in conjunction with the filtering mechanisms such as route tags.

Redistribution

- Keep in mind that these mechanisms increase overall complexity of the network.

redistribution

- Also be aware of routing loops during redistribution operation. Two-way redistribution is the place where routing loops are most likely to occur. And most common prevention for routing loop in this case is to use route tags.

Redistribution

- Redistribution between routing protocols does not happen directly; routes are installed in RIB and pull from the RIB to other protocol.

redistribution

- So route should be in the RIB to be redistributed. A classic example of this is BGP.

Redistribution

- If avoidable, don’t use redistribution. Managing redistribution can be very complex!

redistribution

Optimal Routing

- Overlay protocols should follow the underlay protocol to avoid sub optimal routing and traffic blackholing. In other word, they should synchronize.

optimal

- Control plane state is the aggregate amount of information carried by the control plane through the network in order to produce the forwarding table at each device.

Optimal Routing

- This added complexity, in turn, adds to the burden of monitoring, understanding, troubleshooting, and managing the network.

optimal

- We don’t configure the networks; we configure the networking devices (Routers, Switches etc.)

Optimal Routing

- It is a good idea to create a small failure domain in Layer 2 and Layer 3, but you must be aware of suboptimal routing and black holes.

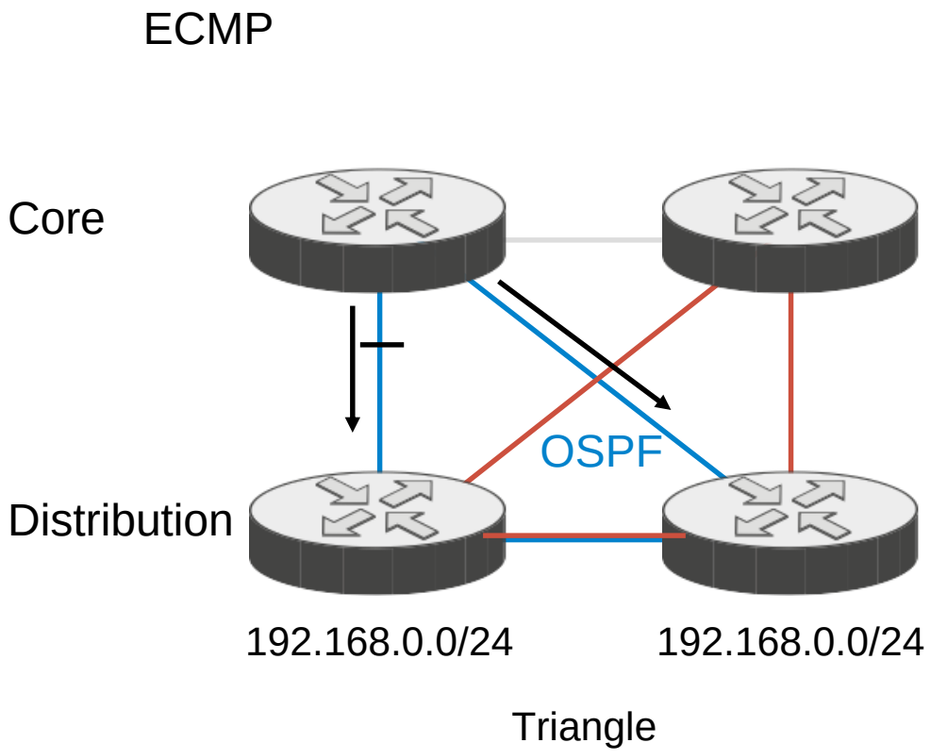
optimal

Network Topologies

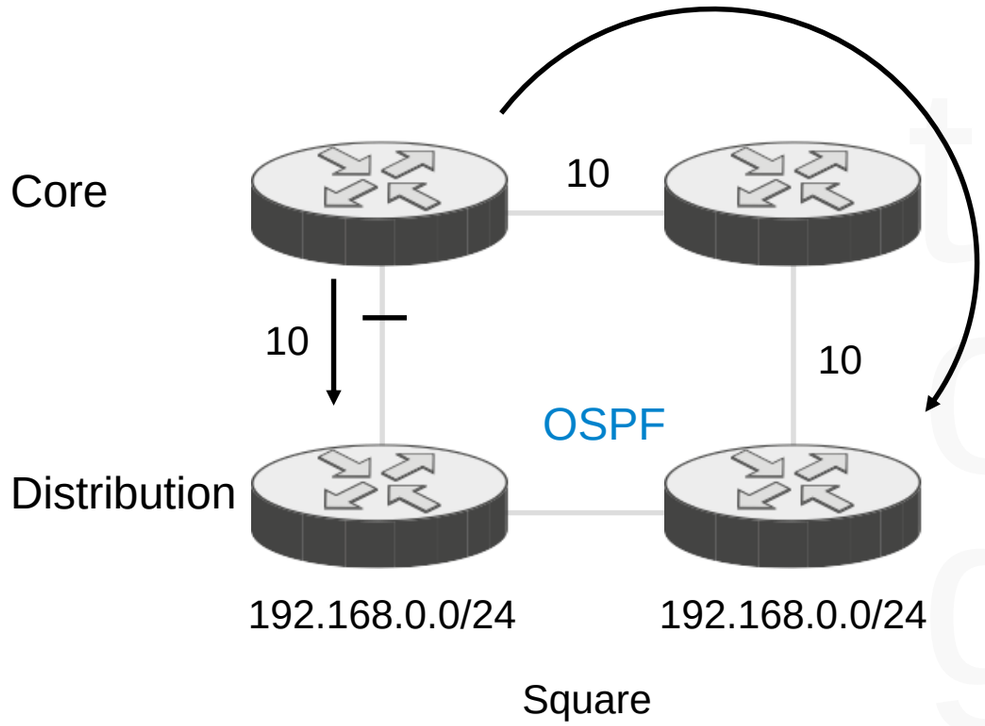
- Intelligence should be at the edge of the networks and the network core should be as simple as possible.
  - In different places of the network, different physical topologies can be seen.
- Physical and logical topologies can be totally different, example: on top of physical ring network, you can have full mesh IBGP

Network Topologies

- Try to create triangle physical topology instead of a square. In the case of link failure, triangle topologies converge faster than squares.



Triangle



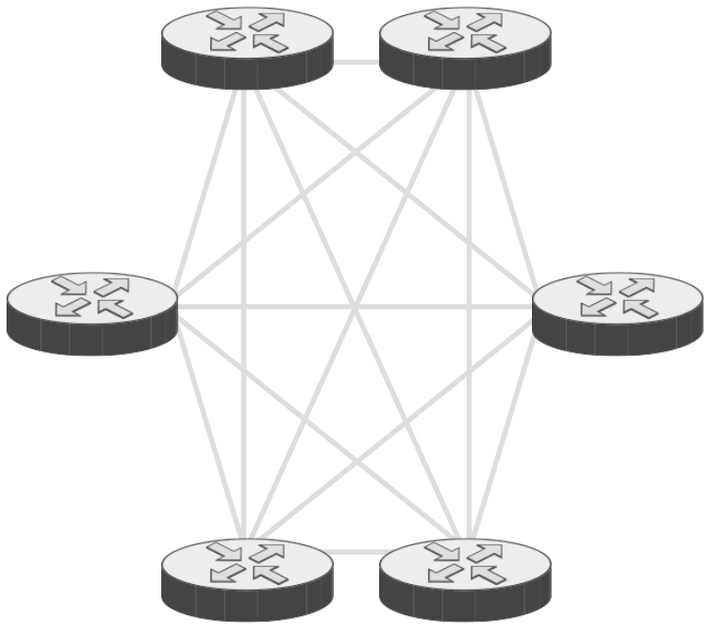
Square

Network Topologies

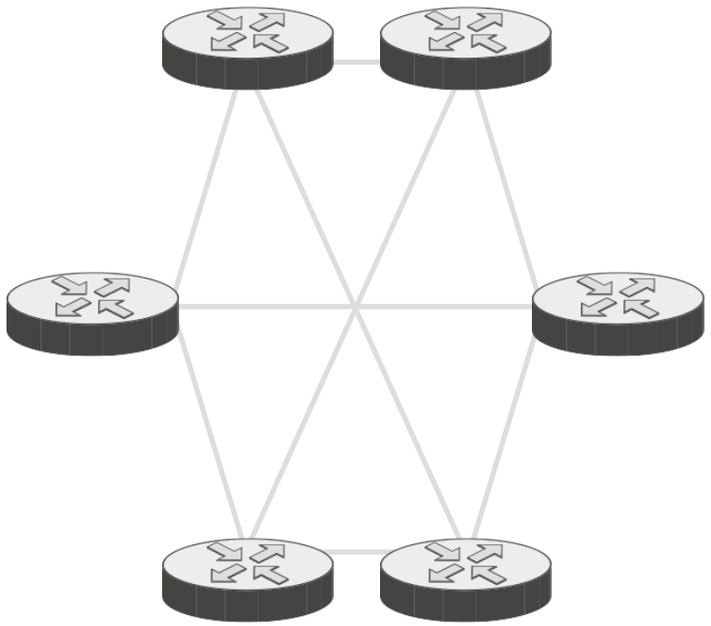
- Ring topology is the most difficult of all the routing protocols from the viewpoint of convergence and optimality.

  - Simply adding some links and creating partial-mesh topology instead of ring provides a more optimal path, better resource usage, and faster convergence time in case of link or node failure.

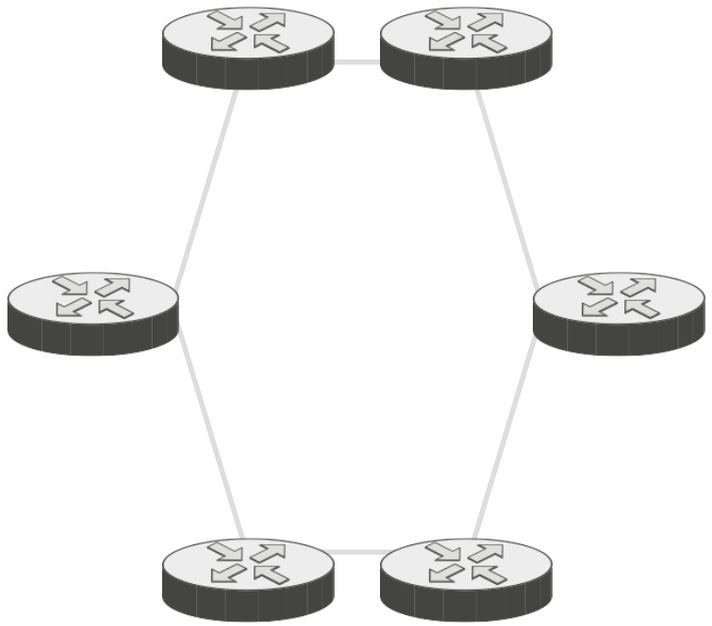
Common Access & WAN Topologies



Full Mesh Topology

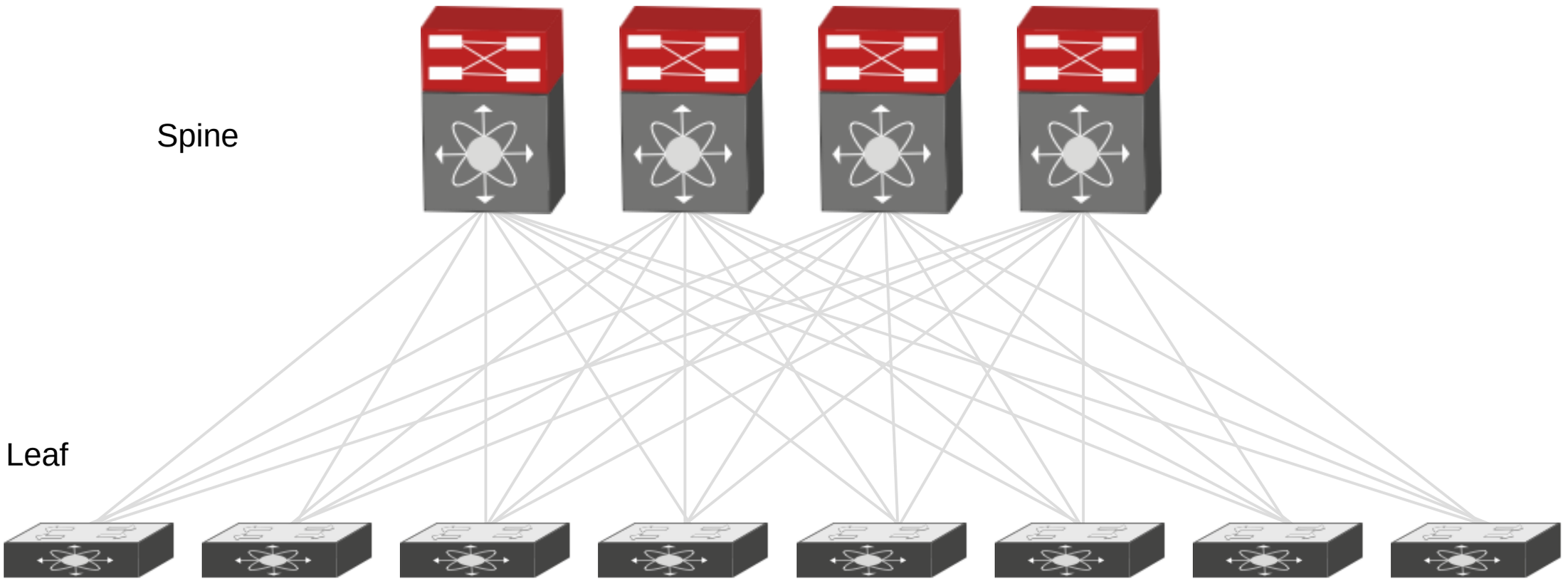


Partial Mesh Topology

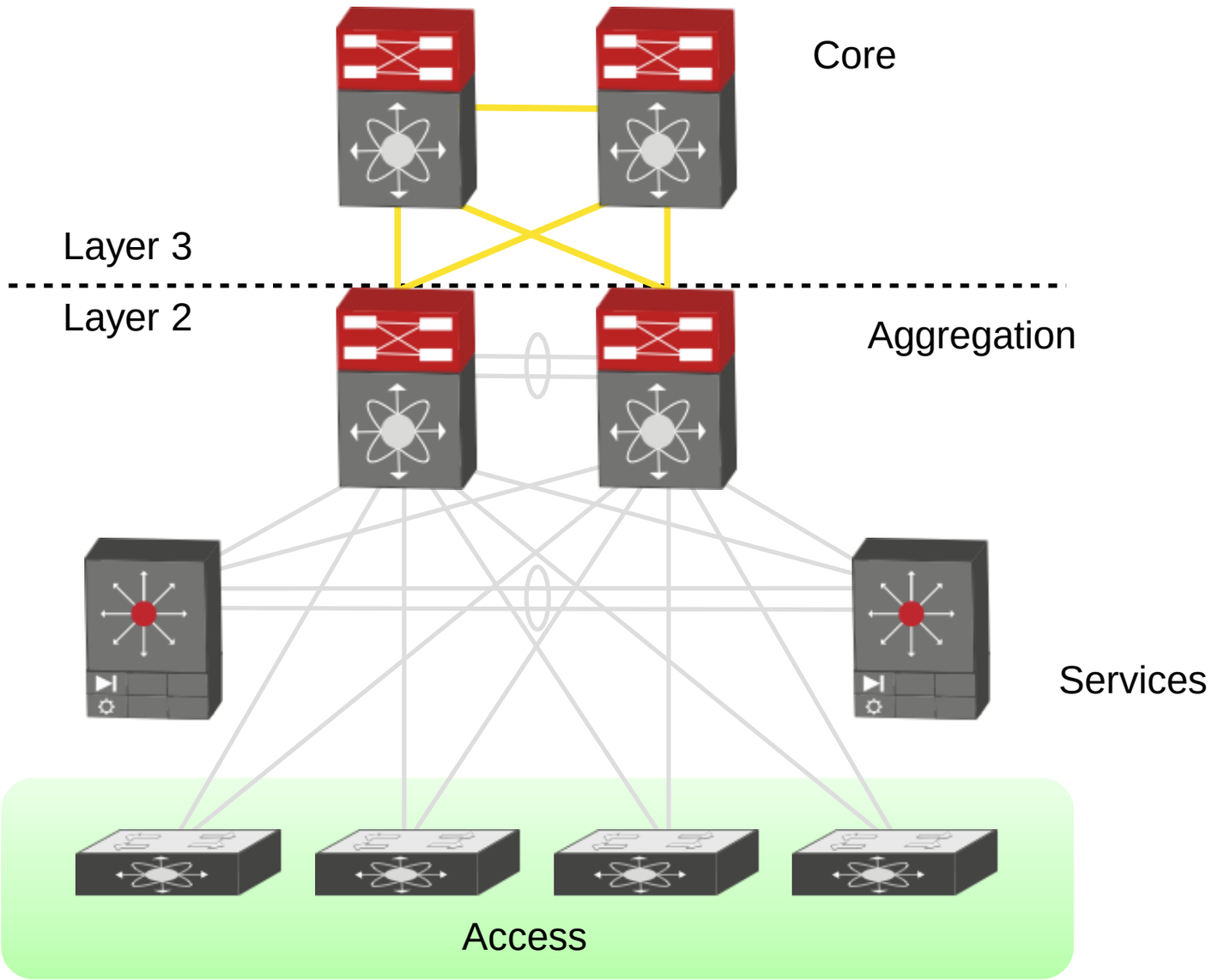


Ring Topology

DC Topologies Clos



DC Topologies Three Tiers



- Full mesh topologies are the most expensive topologies since every device needs to be connected to each other.

- When there is a big traffic demand between any two locations in a ring topology based network, direct link is added between the two locations. Then topology becomes partial mesh.

Security

- Enabling a new feature such as IPv6 or Multicast on part of the network can open the rest of the network to security attacks.
  - Network Address Translation is not a security mechanism.
- MPLS VPNs is equally secure with ATM and TDM based networks.

Security

- Predictable network design is good for security. Removing unnecessary complexity from the design makes network more predictable thus more secure.

security

  - Don’t enable the routing protocols on the user/customers ports; otherwise bad routing entries can be injected into the routing system.

Security

- Always use BGP hardening features on the Internet Gateways.

security

  - BGP Ingress filtering provides protection against route leaking and hijacking.

Security

- More security devices is not necessarily mean more secure network.
  - Stateful security devices such as Firewall, IPS/IDS, Load Balancer requires symmetric routing. This makes network design harder.

Security

- Stateful devices in the datapath can be a bottleneck. Putting a big datacenter firewall vs. smaller virtual firewalls per host or VM is a good example for this.

Simplicity & Complexity

- There are two types of reachability information (MAC or IP) learning mechanisms: control plane learning and data plane learning. Classical Ethernet and VPLS are examples of data plane learning. Routing protocols, LISP, DMVPN are the technologies that use control plane to build reachability information.

Push & Pull Based Control Plane Protocols

- There are generally two types of control plane learning: push based and pull based. Routing protocols are the example of a push based control plane, since routing neighbors send reachability information t each other by default when the adjacency is set up. LISP and DMVPN are pull based control plane mechanisms since devices don t by default send reachability information to each other. Instead, they send to this information to a centralized node.

Simplicity & Complexity

- If you need a robust network some complexity is necessary.
  - You should separate necessary complexity from unnecessary complexity. If you need redundancy, dual redundancy is generally good enough. You add unnecessary complexity by adding a third level of redundancy.

Simplicity & Complexity

- In your design, the motto should be “two’s company, three’s a crowd”. Complexity is inescapable, but unnecessary complexity should be avoided.

Simplicity & Complexity

- Network design is about managing the tradeoffs between different design goals.
  - Not all network design has to have fast convergence, maximum resiliency characteristics, or be scalable.
- Complexity can be shifted between the physical network, operators, and network management systems; overall complexity of the network is reduced by taking the human factor away

Simplicity & Complexity

- SDN helps to reduce overall network complexity by shifting some responsibility from the human operators.

- Don’t use cutting-edge technologies just to show off your skills! Remember, things that may seem simple in actuality might be very complex.

Simplicity & Complexity

- Which one is salt and which one is pepper? It needs to be so simple to understand for a basic object, doesn't it?

- Network design is exactly the same. There are many technologies interact with each other.
- Although each one might be simple, end result may not be so simple to predict.

Simplicity & Complexity

- Features can be intended for robustness, but instead create fragility. The impact may not be seen immediately, but it can be huge. In design this is known as the Butterfly Effect.

“ A butterfly flapping its wings in South America can affect the weather in Central Park.”

Simplicity & Complexity

- Last but not least, know the purpose of your design. Do you know why you are doing whatever you are planning to do ? Is there a valid business requirement? Does it make life easier? What is the purpose?

Imagine you designed below teapot, can you use it?
Let me know if you can.

NETWORK DESIGN TOOLS & THE BEST PRACTICES

QUIZ

Question 1

In the below figure, two routers are connected through two links. OSPF routing protocol is running over the links.

Which below statement is true for the below figure?

A. Adding more links between the routers increase routing table size

B. IS-IS would be a better option

C. More links provide better resiliency

D. More links provide faster convergence

E. More links provide better security

Answer 1

A. Adding more links between the routers increase routing table size

Adding more links don’t provide better security. Resiliency depends on redundancy, convergence and reliable packet delivery. More links don’t necessarily provide better resiliency.

General rule of thumb, 2 links is best for resiliency. We cannot know whether IS-IS would be better since there is no other requirement.

Option A is definitely correct. More links increase routing table size since OSPF is running on individual links and more links means more routing table entry.

If there would be an etherchannel between the routers and OSPF would run on top of that link, adding more link wouldn’t increase the routing table size.

Question 2

Which below technologies provide fast failure detection? (Choose two)

A. BFD

B. Routing fast hellos

C. Loopguard

D. SPF Timers

E. BGP Scanner time

Answer 2

A. BFD

B. Routing fast hellos

Loopguard, SPF timers and BGP Scanner Timers are not used for fast failure detection. BGP Scanner time for example is 60seconds and reducing can create 100% CPU utilization. Thus better and newer approach Next Hop Tracking is used in BGP, as it will be explained in the BGP Chapter.

Routing Protocols hellos can be tuned to provide fast failure detection and the purpose of BFD is to provide fast failure detection.

Question 3

Which of the below protocols support BFD for fast failure detection? (Choose all that apply)

A. Static Routing

B. OSPF

C. IS-IS

D. EIGRP

E. BGP

F. RIP

Answer 3

- A. Static Routing
- B. OSPF
- C. IS-IS
- D. EIGRP
- E. BGP

All the routing protocols above except RIP support BFD as it was mentioned in this chapter. They can register to BFD process for fast failure detection. In case of failure BFD inform these protocols to tear down the routing session.

RIPv2 on the other hand supports BFD.

Question 4

What are the bene ts of having modular network design? (Choose two)

A. Each module can be designed independently from each other

B. Each module can be managed by different team in the organization

C. Each module can have a separate routing protocol

D. Each module can have different security policy

Answer 4

A. Each module can be designed independently from each other

B. Each module can be managed by different team in the organization

If the design supports modularity, then each module can be designed independently, In access, aggregation, core module for example, access network can be hub and spoke, distribution can be full mesh and core network can be partial mesh.

Also commonly in the service provider networks, access and core team are the separate business units and modularity provides this opportunity. Or in large Enterprises, different team can managed the different geographical areas of the network, which has been designed by considering modularity.

Modularity is not done to have different routing protocols and companies should deploy common security policies across all domains.

Question 5

Which below statements are true for the network design? (Choose two)

A. Predictability increases security

B. Every networks need 5x9 or 6x9 High Availability

C. Using more than routing protocol in the network increases availability

D. Modular network design reduces deployment time

Answer 5

A. Predictability increases security

D. Modular network design reduces deployment time

As it was explained in this chapter, modular network design reduces deployment time. And predictability increases security. Predictable networks also reduces troubleshooting time thus increases high availability.

Not every networks need 5x9 or 6x9 high availability. Using more than one routing protocol in the network, if there is mandatory reason such as partner network requirement, is not a good design.

Question 6

If there is two-way redistribution between routing protocols, how can routing loop is avoided?

A. Deploying Spanning Tree

B. Deploying Fast Reroute

C. Implementing Route tags

D. Only one way redistribution is enough

Answer 6

C. Implementing Route tags

As it was explained in the redistribution part of the Best Practices chapter of the book, route tags are the common method to prevent routing loops if redistribution is done at multiple locations between the protocols.

Question 7

Which below statements are true for the network design? (Choose Three)

A. Using triangle topology instead of square reduces convergence time that's why it is recommended.

B. Full mesh topology is the most expensive topology to create.

C. Using longer and complex configurations always better so people can understand how good network designer you are.

D. Sub optimal routing is always bad so avoid route summarization whenever you can since it can create sub optimal routing.

E. Network complexity can be reduced by utilizing SDN technologies.

Answer 7

A. Using triangle topology instead of square reduces convergence time that's why it is recommended.

B. Full mesh topology is the most expensive topology to create.

E. Network complexity can be reduced by utilizing SDN technologies.

Network complexity can be reduced by utilizing SDN technologies as it was explained in this chapter. It helps to shift the configuration task from the human to the software. That's why Option E is one of the correct answers.

Route summarization can create sub optimal routing but sub optimal routing is not always bad. For some type of traffic in the network, optimal routing may not be required at all. And just because we might have sub optimal routing, we shouldn't avoid doing summarization. That's why Option D is incorrect.

It should be obvious that Option C doesn't make sense.

Option A and B are also correct. Triangle topology reduces convergence time and full mesh topologies are the most expensive topologies.

Question 8

What is the key benefit of hierarchical network design?

A. less Broadcast traffic

B. Increased flexibility and modularity

C. Increased security

D. Increased availability

Answer 8

B. Increased flexibility and modularity

Hierarchical design may not be redundant and highly available. Also it doesn’t bring additional security but key benefit of it is flexibility and modularity as it was explained in the Best Practices chapter.

Question 9

If routing summarization is done which below statements are valid for the link state protocols? (Choose Two)

A. Convergence will be slower

B. Sub optimal routing may occur

C. Traffic blackholing may occur

D. Routing table size grows

Answer 9

B. Sub optimal routing may occur

C. Traffic blackholing may occur

As it was explained in the chapter, when route summarization is done routing table size gets smaller which makes converges faster. It can create sub optimal routing and traffic might be blackholed in some failure scenarios.

Question 10

What would be the impact of doing summarization at the aggregation layer in three-tier hierarchy? (Choose Two)

A. Core network can be simplified, it doesn’t have to keep all Access network routes.

B. If you have summary in the aggregation layer, core can be collapsed with aggregation layer.

C. Access network changes don’t affect the core network.

D. Aggregation is the user termination point and summarization shouldn’t be made at aggregation layer.

Answer 10

A.Core network can be simplified, it doesn’t have to keep all Access network routes.

C.Access network changes don’t affect the core network.

In three-layer hierarchy aggregation layer is the natural summarization point. When the summarization is done at the aggregation layer, core layer is simplified and the access network changes don’t affect the core layer.

Collapsing the core is not the result of summarization since the main reason of using core layer is physical scaling requirement. With summarization physical requirements don’t go away.

Aggregation layer is not the user termination point. User termination is the access layer responsibility, thus Option D is incorrect.

Question 11

Which routing protocol supports unequal cost multi path routing?

A. OSPF

B. IS-IS

C. EIGRP

D. RIPv2

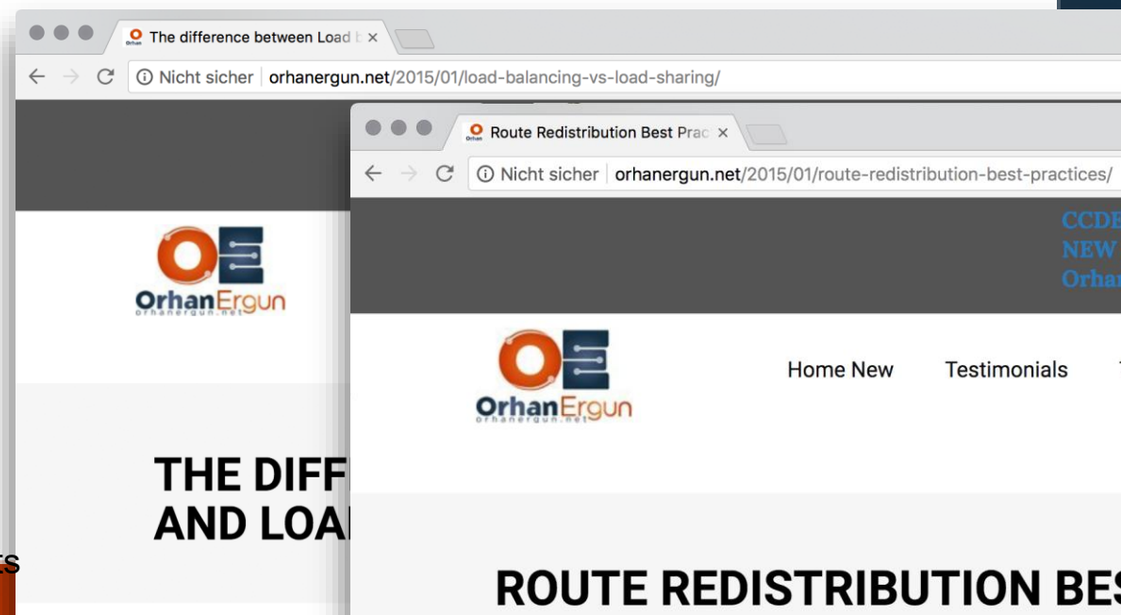
Answer 11

B.EIGRP

In the above question, all the routing protocols are dynamic routing protocols and among them only EIGRP supports unequal cost multi path routing. And as it was explained in the chapter, with MPLS Traffic engineering tunnels only, OSPF and IS-IS can support unequal cost multipath.

Extra Study Resources

- Videos:
- http://ripe61.ripe.net/archives/video/19/
- Articles :
- http://orhanergun.net/2015/01/route-redistribution-best-practices/
- https://tools.ietf.org/html/draft-ietf-ospf-omp-02
- https://www.ietf.org/rfc/rfc3439.txt
- http://orhanergun.net/2015/01/load-balancing-vs-load-sharing/



OSPF

Open Shortest Path First

Agenda

- OSPF Theory
- OSPF Fast Convergence
- Convergence and Micro-loop
- OSPF Scalability, Multi Area OSPF Design
- Fast Reroute with OSPF
- Overlay Technologies and OSPF (GRE, mGRE, DMVPN, LISP)
- OSPF in the Datacenter Networks
- OSPF in the Service Provider Networks

Agenda

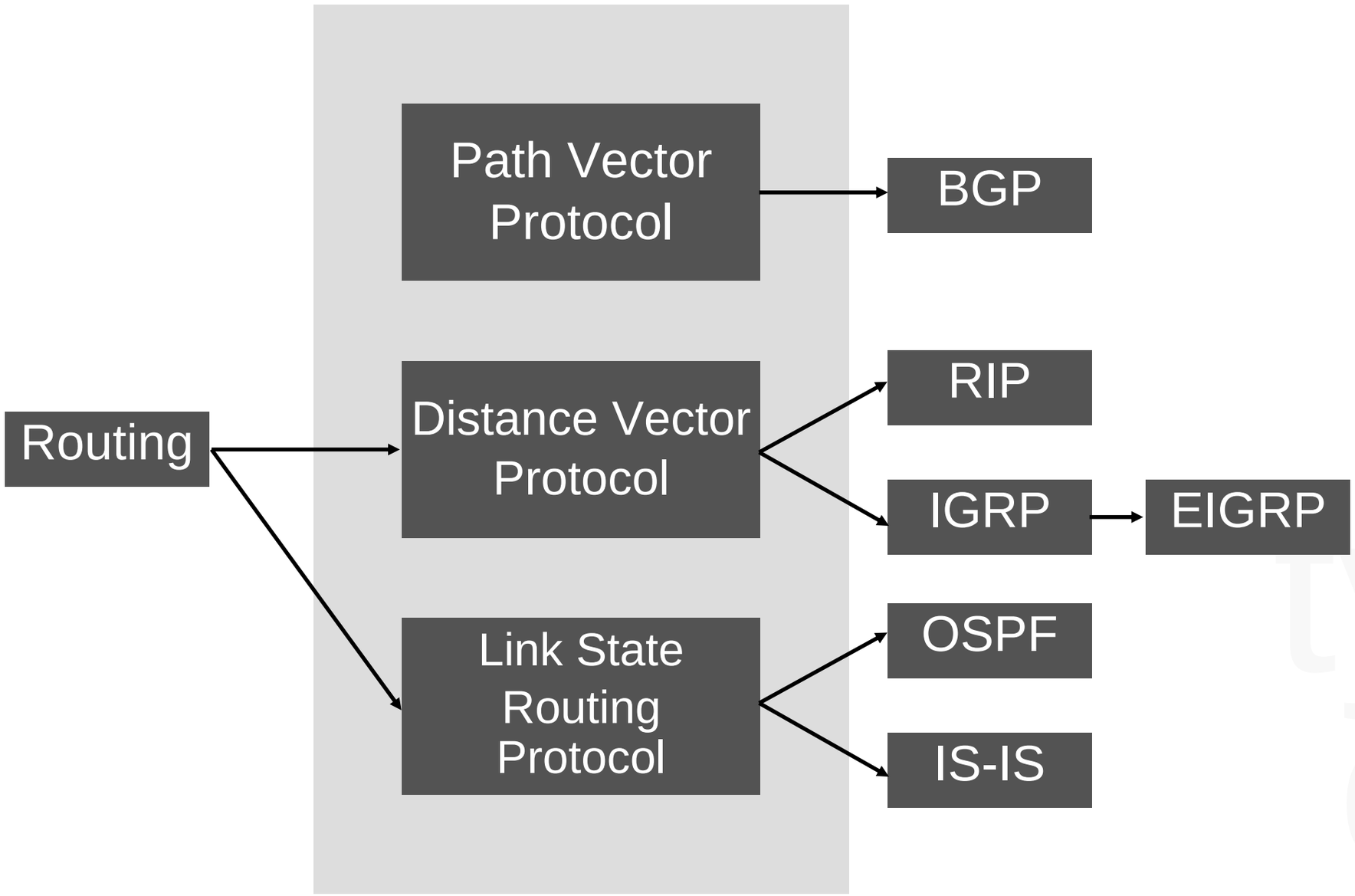
- OSPF Design Best Practices
- OSPF Advantages and Disadvantages
- OSPF Frequently Asked Questions – How many Routers in an OSPF Area, How many ABR per Area ?
- Case Studies
- OSPF in the CCDE Exam
- Summary
- Bonus Materials

Theory

- If the requirements is to have MPLS Traffic Engineering, Standard based and Enterprise level protocol then only choice is OSPF.

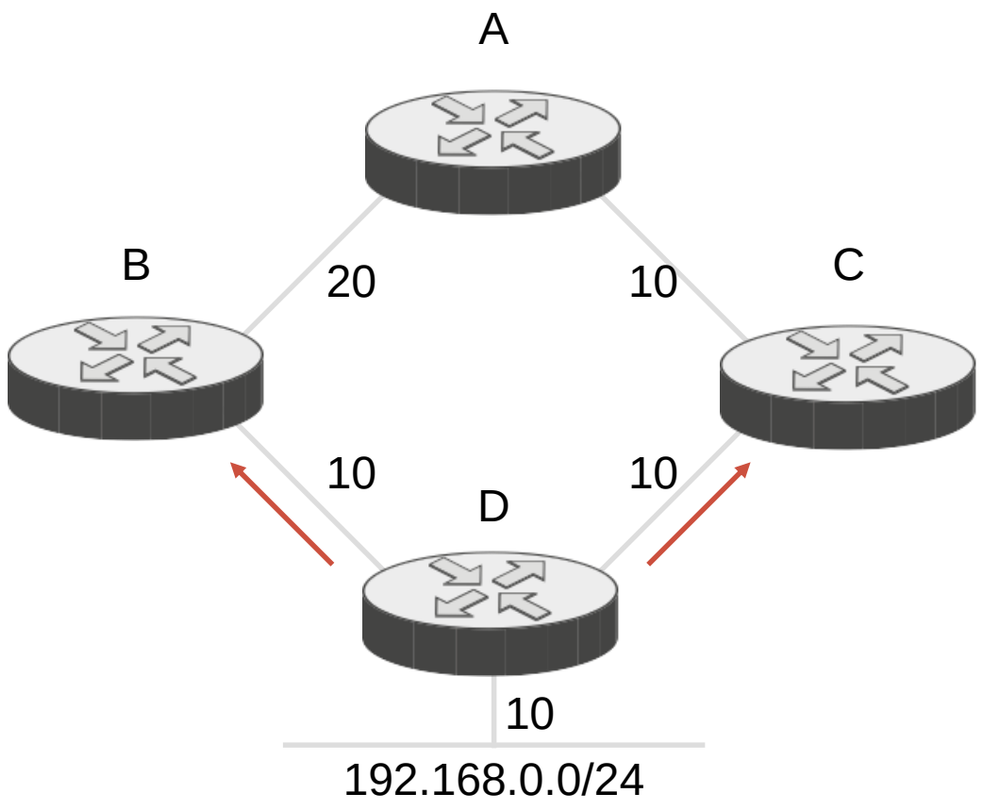
- OSPF as a link state protocol has many similarities with IS-IS but if the requirements is to run IPsec, since IS-IS doesn’t work on top of IP, it is not well suited for Enterprise environment.

3 Types of Routing Protocols



LSA Flooding

- In a link state protocols, each router advertises the state of its links to every other router in the Network

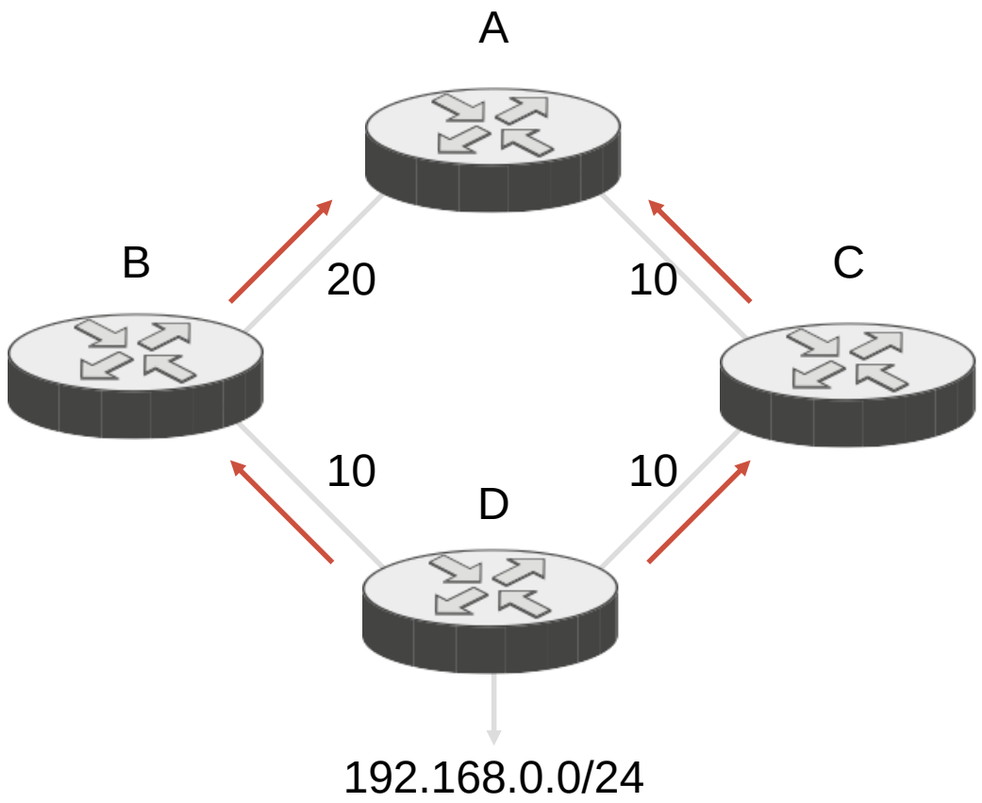


- D determines that it is connected to 192.168.0.0/24 with metric 10
- Connected to B with metric 10
- Connected to B with metric 10

- D advertises this information describing all of its links to its neighbors B and C

lsa

LSA Flooding



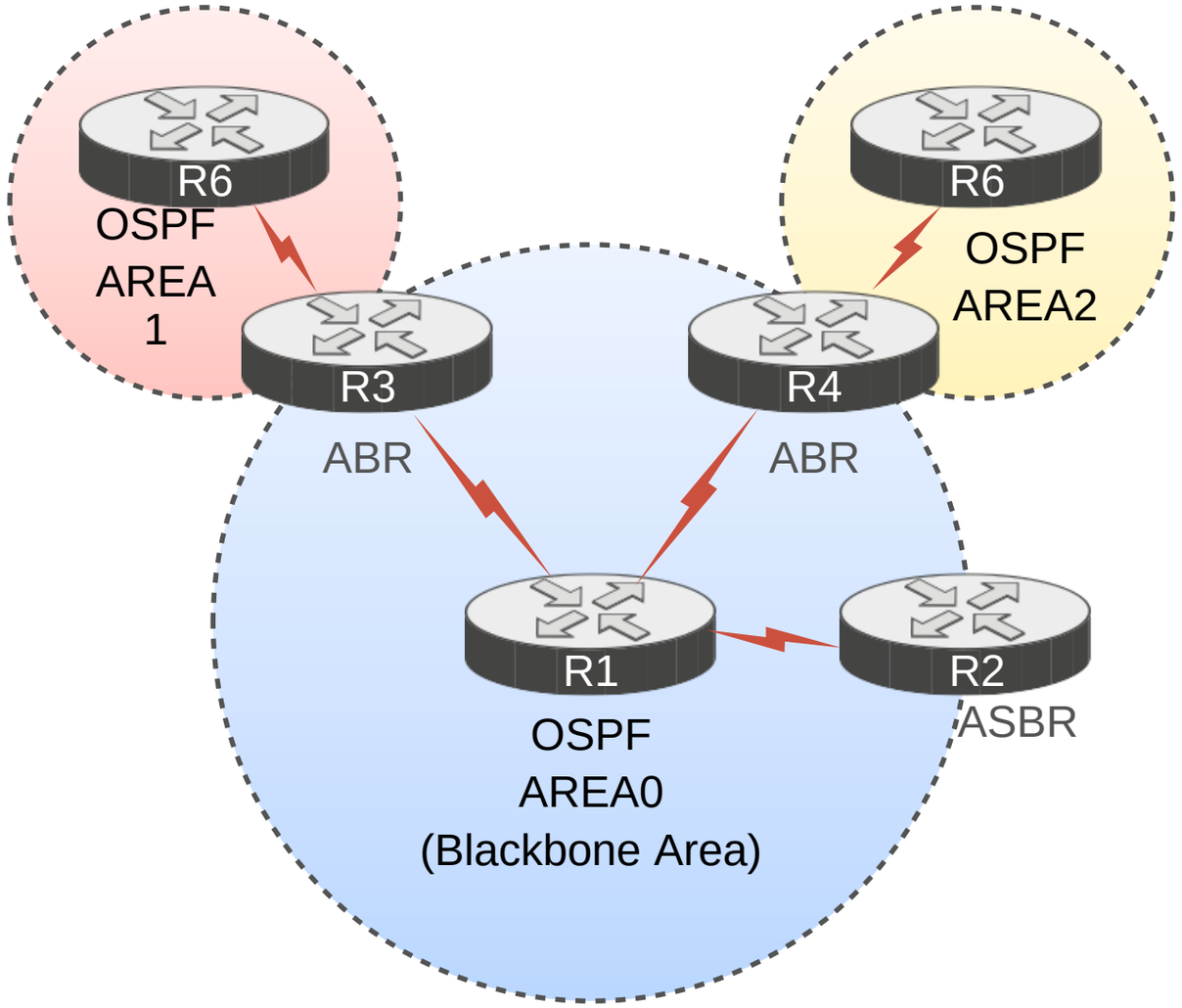
- This process of recording and re-transmitting is called flooding.
- Since information is flooded within a link State network, every router should have the same information about the network (How it looks like).

ABR (Area Border Router) and ASBR (Autonomous System Border Router)

- When scaling become an issue network is broken into separate flooding domains, which we call it areas.

- The router connecting the two area is called an Area Border Router (ABR).

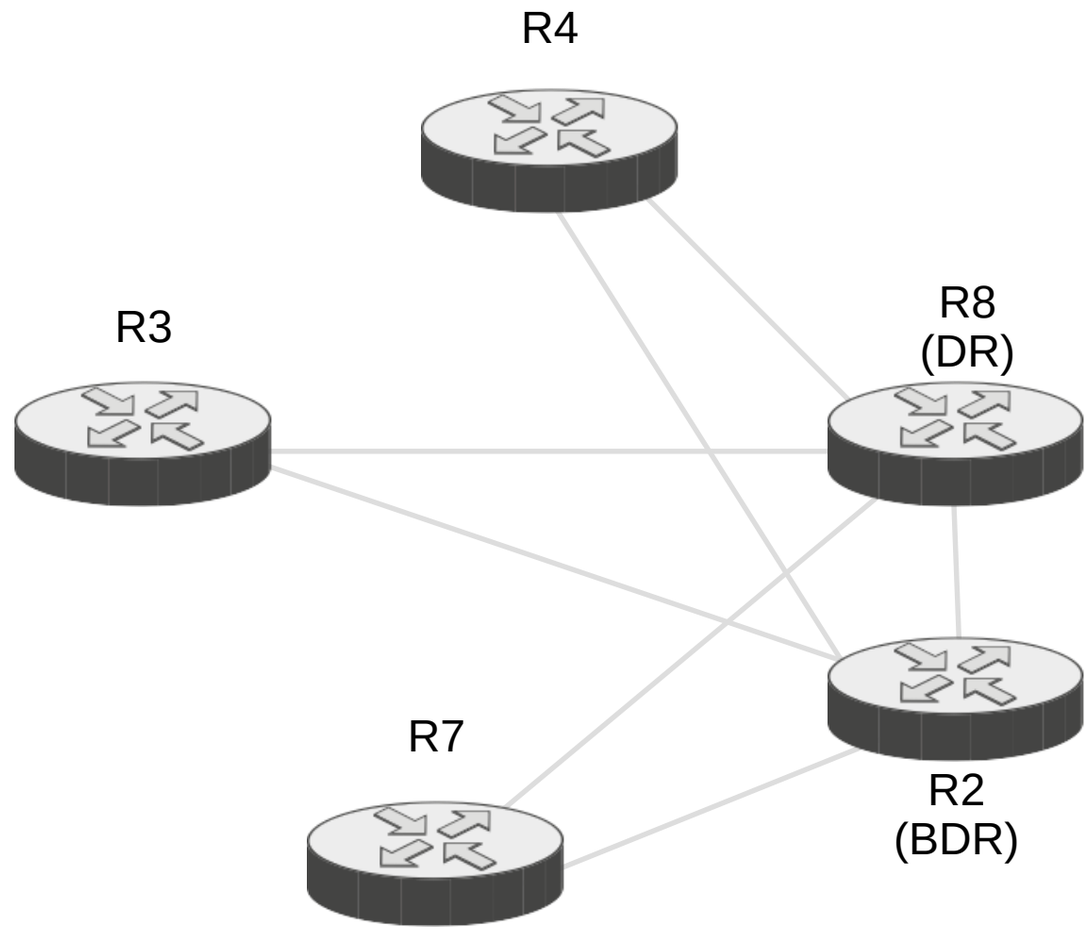
- The router connecting the network to the other networks is called ASBR.



- In a particular area every routers have identical topology map. Every router knows which network behind which router and their metrics.
    - OSPF, unlike EIGRP and IS-IS works differently on different media. On broadcast network DR (Designated Router) creates a pseudo node to avoid unnecessary flooding.

OSPF DR & BDR

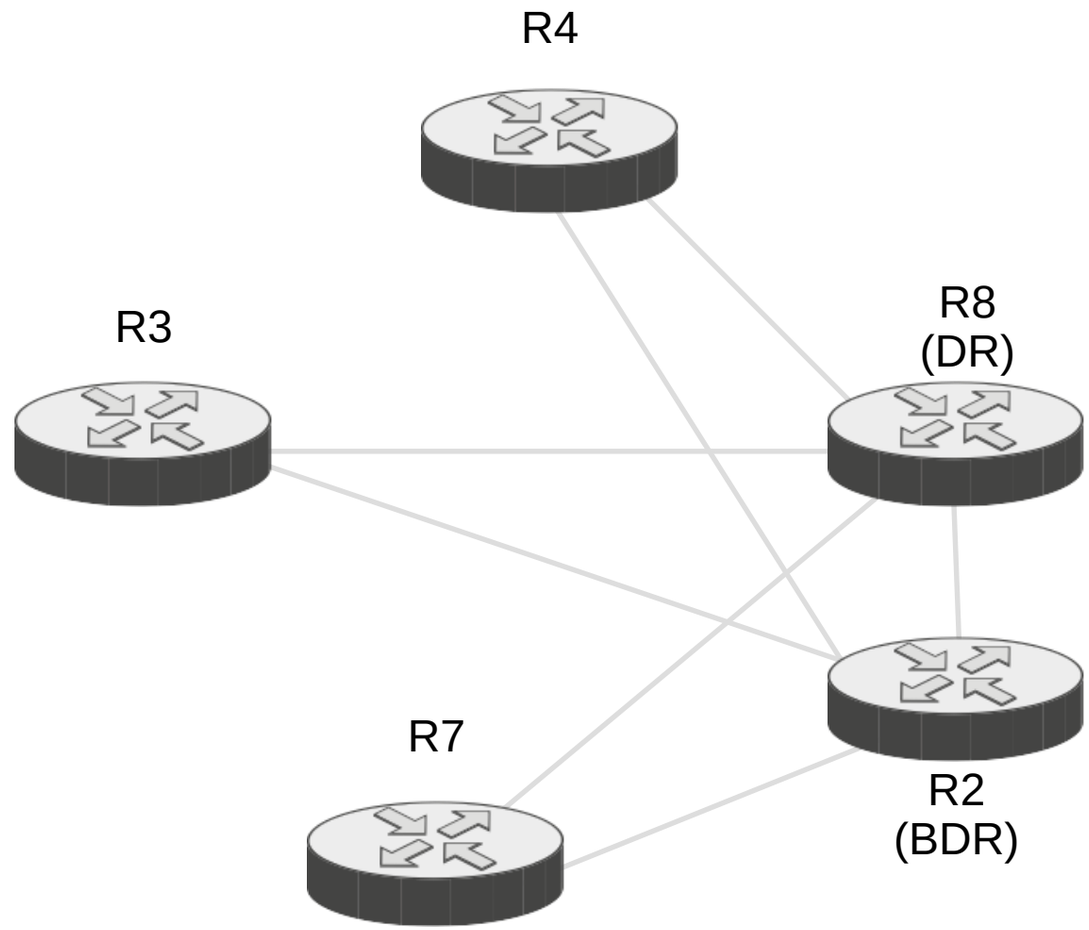
- DR creates type 2 LSA (Network LSA) to inform the connected routers on the broadcast network.
- Highest priority OSPF router on the broadcast segment wins the Designated Router (DR) election. If priorities are the same then highest router ID wins the DR election. On every broadcast segment there can be only 1 DR.



OSPF DR & BDR

- Each router in multi access segment creates a OSPF neighbourship with DR and BDR only

- Unlike IS-IS, there is Backup Designated Router (BDR) in OSPF.



DR & BDR Question

- Is there a DR and BDR in the below topology?

Direct back to back connection

- We can only have scalable, resilient, fast-converged OSPF design when we understand OSPF LSAs and Area types and their restrictions

- There are 11 types of LSAs and 5 of them are important for the OSPF routing protocol design.

5 Critical LSAs for OSPF Design

	Description
1.Router	• Router information • Connections to other routers • Connections to links (link states)
2.Pseudonode	• Pseudonode information • Connections to routers • Connections to broadcast link
3.Summary	• Destinations reachable within an area (flooding domain)
4.Border Router	• Cost to reach a router advertising external routing information (an ASBR) • Generated by the ABR
5.External Destination	• Cost to reach a destination which is external to the OSPF flooding domain (outside the local autonomous system)

lsa

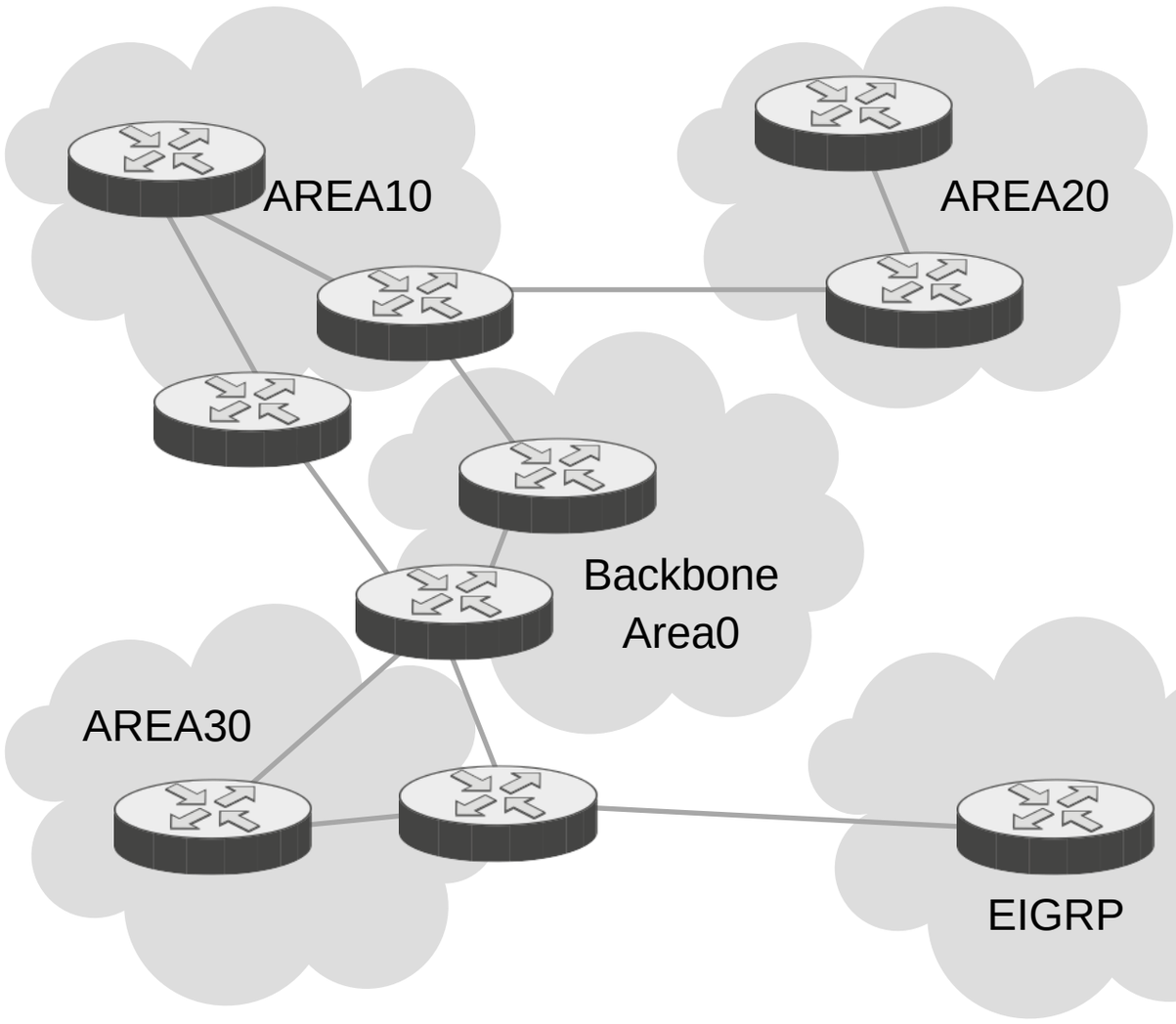
OSPFv2 all LSA Types

LSA Type	Description
1	Router LSA
2	Network LSA
3 and 4	Summary LSAs
5	AS external LSA
6	Multicast OSPF LSA
7	Defined for NSSAs
8	External attribute LSA for Border Gateway Protocol (BGP)
9, 10, 11	Opaque LSAs

OSPF Area Types

Area	Restriction
Normal	None
Stub	No type 5 ASYExternal LSAs allowed
Totally Stubby	No type 3, 4 or 5 LSA allowed except the default summary route
NSSA	No type 5 ASYExternal LSAs allowed, but type 7 LSAs that convert to type 5 at the NSSA ABR can traverse
NSSA Totally Stubby	No type 3, 4 or 5 LSAs allowed except the default summary route, but type 7 LSAs that convert to type 5 at the NSSA ABR are allowed

All routers in an area must have same LSDB (Link State Database)



- OSPF uses two level hierarchical model.
- Areas are use for scalability.
- Regular, Stub, Totally Stub, NNSA and Totally NSSA Areas.
- Router keeps separate link state database for each area which it belongs.
- LSA flooding is bounded by area, outside of an area Type 1 and Type 2 LSA is not sent.
- SPF calculation is performed independently for each area.
- All routers belonging the same area should have identical link state database.

OSPF Fast Convergence

- Network convergence is the between the failure event and the recovery.
    - Through the path all the routers process the event and update their routing and forwarding table.

- Thus; there are 4 steps for convergence in general:

    1. Failure Detection
    2. Failure Propagation
    3. Processing the new information
    4. Routing and Forwarding table update

OSPF Fast Convergence

- Convergence is a control plane events and for IGPs it can take seconds; BGP routers which have full internet routing table, control plane convergence can take minutes.

- Protection is a data plane recovery mechanism. As soon as failure is detected and propagated to the nodes, data plane can react and a backup path can be used. A backup path should be calculated and installed in routing and forwarding table before the failure event.

Convergence Tools

- DETECTION

  Carrier Delays
  Debounce Timers
  Bidirectional Forwarding Detection-BFD
  Protocol Hello/Dead Timers

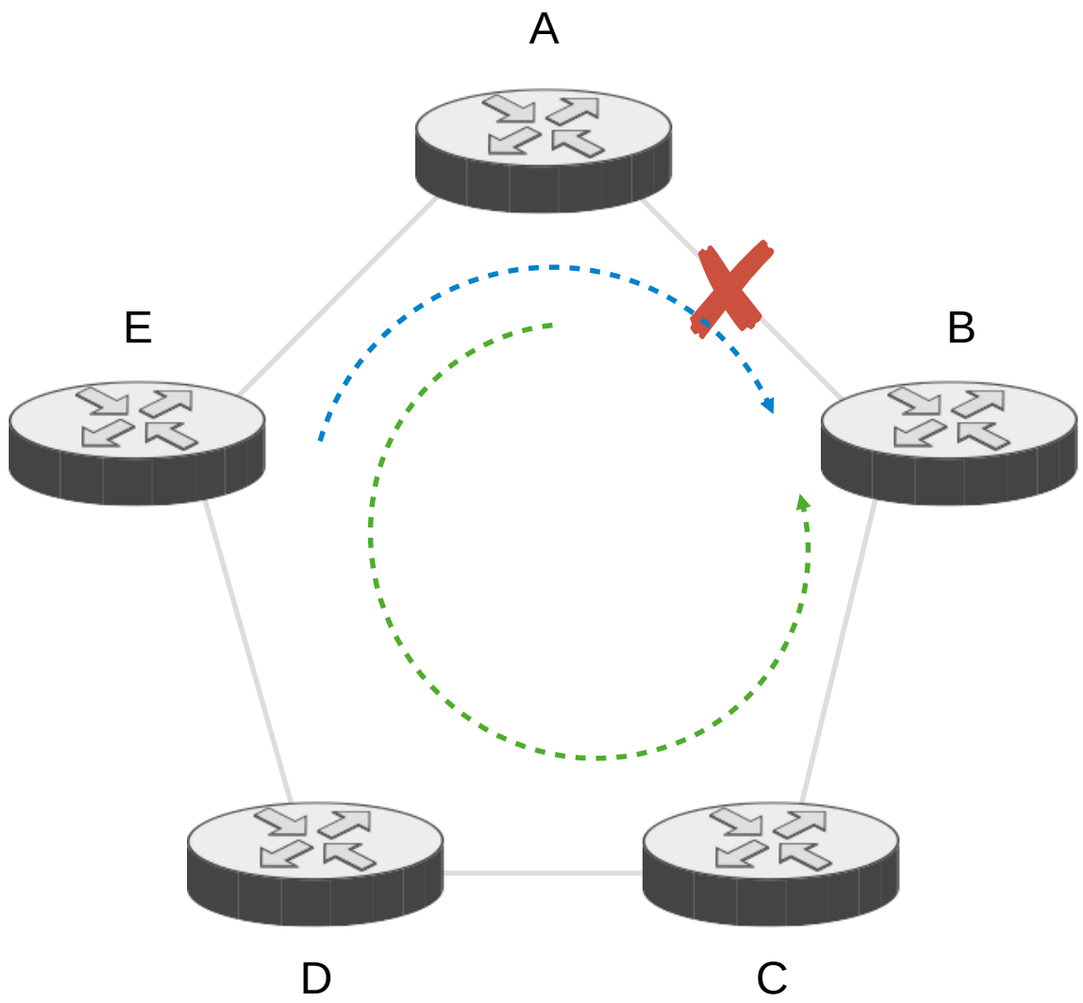
- PROPAGATION

  Interface event dampening LSA Pacing

- PROCESSING

  Full, Partial and Incremental SPF
  MinLSA Arrival Interval
  LSA and SPF Throttling timers

Convergence & Micro-loop



OSPF Scalability and Multi Area Design

- To reduce the impact of flooding and provide scalability Multi Area OSPF Design can be used.
  - With the today hardware 100s of OSPF routers can be placed in an OSPF area.
- OSPF Multi-Area design is not the only tool to provide scalability.

OSPF Scalability and Multi Area Design

- OSPF prefix-suppression feature provides scalability through removing point-to-point links from the Type 1 LSA, thus LSDB and routing table size is reduced.

- Also OSPF Database-filter (similar to IS-IS mesh group) reduces the flooding between the routers in a full-mesh topology, thus provides scalability

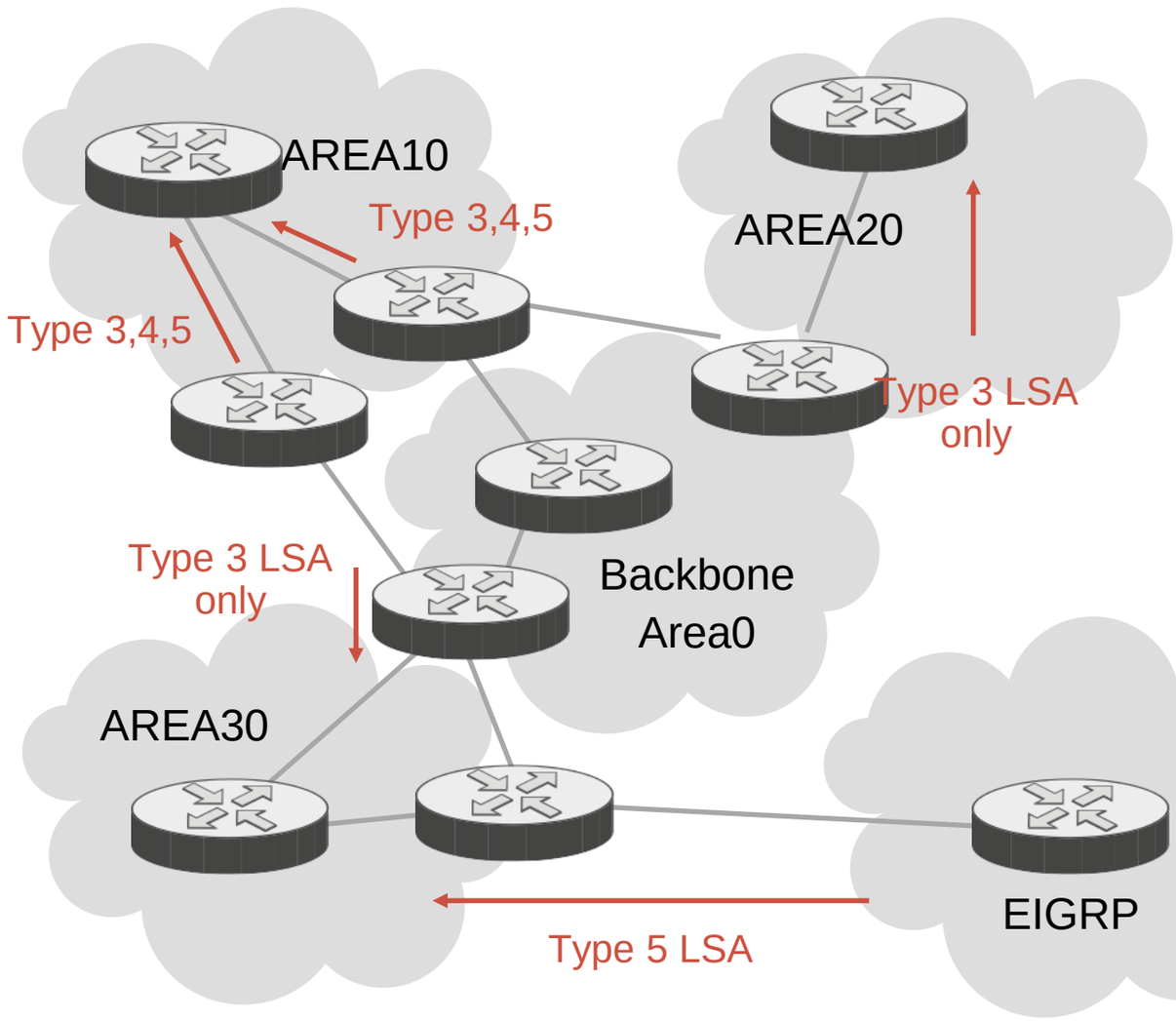
OSPF Scalability and Multi Area Design

- Number of routers in an OSPF domain may impact scalability.
    - Problem with the number of routers in OSPF domain is the Router LSA size.
- Each additional link and subnet makes Router LSA bigger and when it exceed the interface MTU, packet is fragmented. You should always avoid fragmentation.

Multi Area OSPF – Fault Isolation

- Special areas such as Stub and NSSA in OSPF provides fault isolation.
  - When fault is isolated, adding a link or node or changing the metric in one area doesn't cause Full SPF calculation in other OSPF areas.
- This is important for scaling.

What is the problem with the below area design?



- Area 10 is regular area thus all the LSA types including type 3 and 5 are allowed.
- ABRs create Type 4 LSA into an Area 10.
- Area 20 is Stub Area. That's why only Type 3 LSA is allowed.
- Type 5 LSA is not allowed in Stub Area
- Thus type 4 is not generated as well.

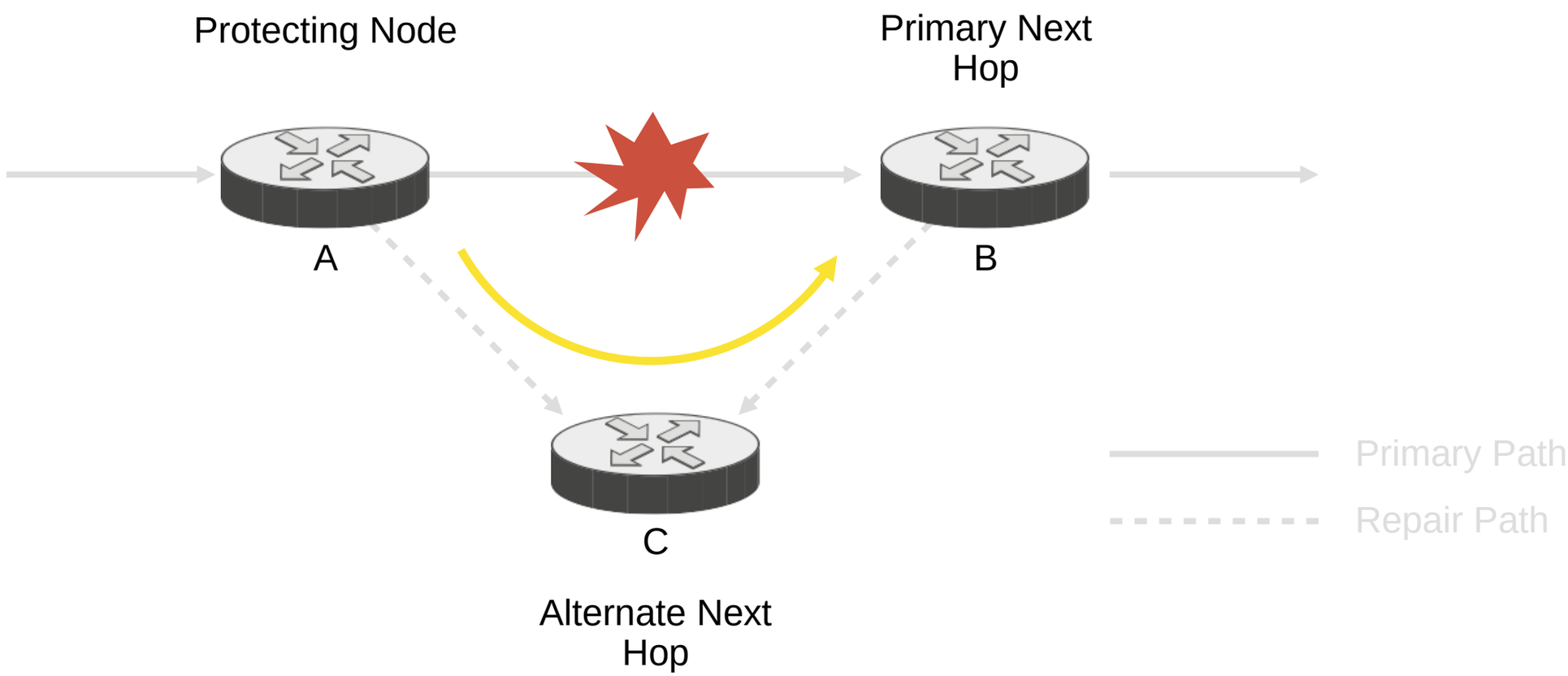
- ABR has to have a connection to more than one area, and at least one area should be in Area 0 (Backbone Area) but even creating a loopback interface and placing it into a Area 0 makes that router an ABR.

Area	LSAs Allowed
Backbone	1, 2, 3, 4, 5
Regular	1, 2, 3, 4, 5
Stub	1, 2, 3
Totally Stubby	1, 2, Default 3
Not So Stubby	1, 2, 3, 4, 7

Fast reroute with OSPF

- Fast reroute is done by placing the alternate route in RIB and FIB
    - Alternate/backup route is not used while primary link is up and running.
- OSPF FRR can be done with LFA, Remote LFA, Segment Routing FRR, RSVP-TE FRR.

Convergence & Micro-loop



- OSPF Fast reroute can provide 50ms convergence time which cannot be done by tuning SPF parameters, link failure detection tuning with BFD etc.

- Fast reroute is proactive recovery, fast convergence is reactive recovery technique.

- Proactive recovery mean, calculating and installing the backup path into the RIB and FIB before the failure event.

Overlay Technologies and OSPF (GRE, MGRE, DMVPN, GETVPN, LISP)

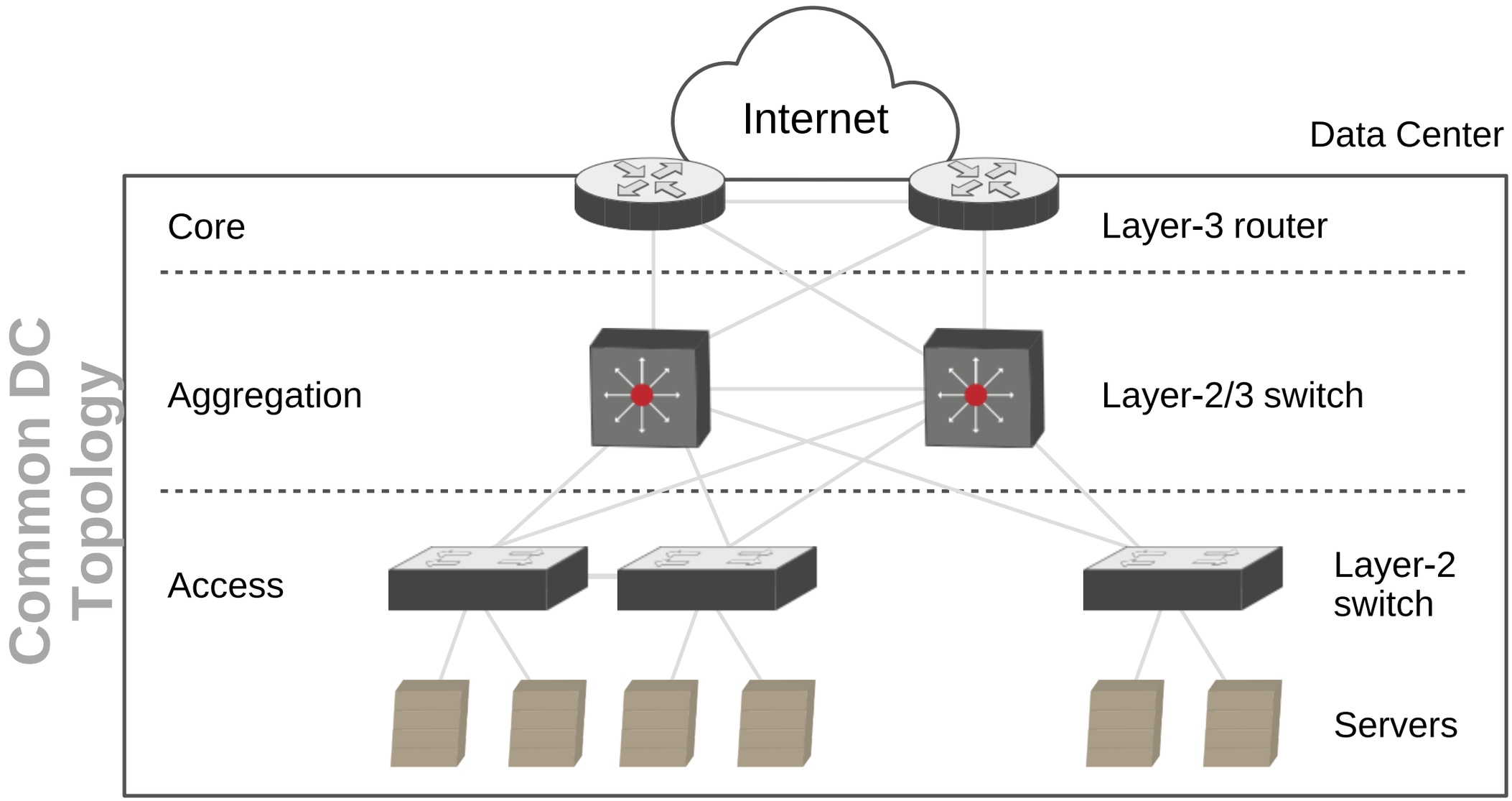
- OSPF can work on top of many overlay technologies.
  - GRE, MGRE, DMVPN, GETVPN and LISP can be used to create overlay/VPN in the networks.
- OSPF can be used for these overlay mechanisms as an underlay infrastructure routing protocol.

Overlay Technologies and OSPF (GRE, MGRE, DMVPN, GETVPN, LISP)

- OSPF works over GRE, MGRE and DMVPN
  - OSPF doesn’t work over GETVPN and LISP, because both are tunnelless VPN mechanisms, routing protocols can be an underlay for them but not an overlay
- OSPF with GRE is not scalable for large scale deployment but scaling limitation comes from GRE, it is not the OSPF problem, MGRE provides scalability with OSPF even in large scale deployment.

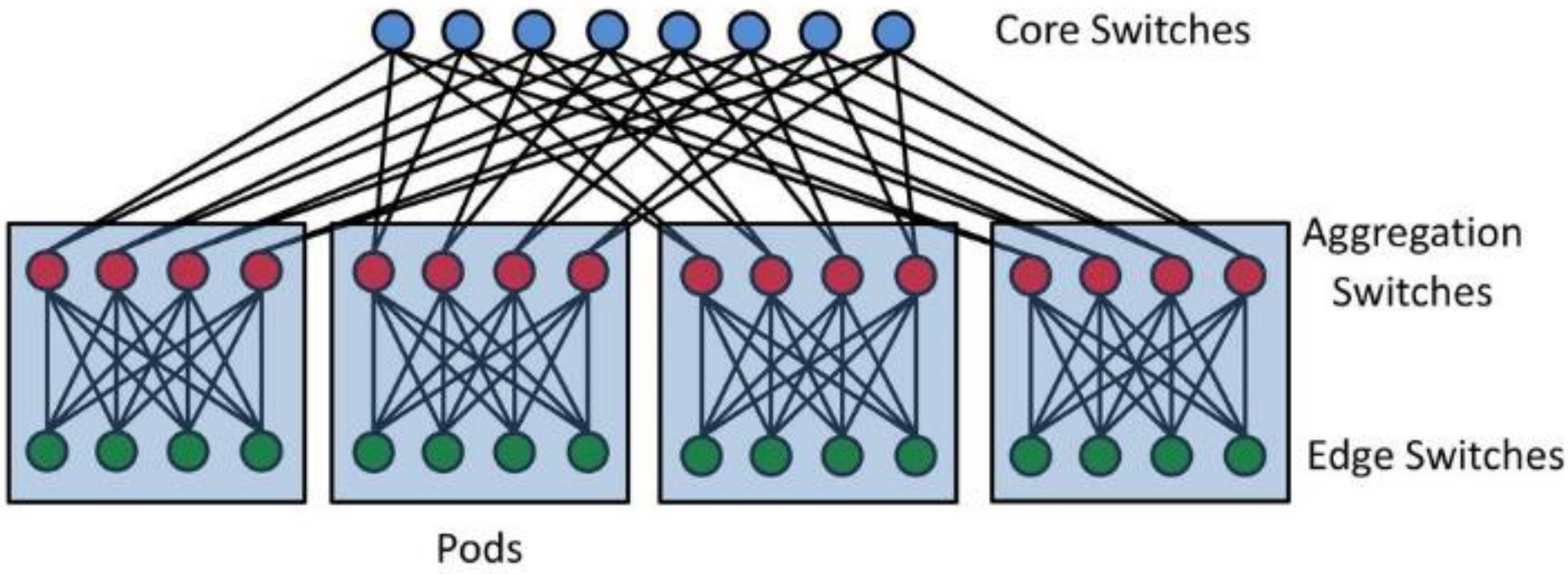
OSPF in the Datacenter

- OSPF can be used at the DC edge to advertise DC prefixes to the WAN and Campus network



- Also OSPF can be used as a Datacenter Fabric Protocol.
    - Datacenters are very densely connected networks, thus OSPF flooding creates scalability problem.
- Large scale Datacenters mainly use CLOS (Leaf and Spine) topology, depends on scale, multi stage CLOS topologies are used.

3 stage CLOS topology



- The Fabric provides basic connectivity, with possibility to carry one or more overlays
- The Fabric MAY provide interconnect facility for other fabrics.
- The Fabric MUST support non equidistant end-points.
- The Fabric MUST support Spine and Leaf [CLOS] + isomorphic topologies within its network.
- The Fabric MAY support non Spine and Leaf topologies

- The Fabric SHOULD support 250k routes @ 5k fabric nodes with convergence time below 250ms.
- The Fabric SHOULD support 500k routes @ 7.5k fabric nodes with convergence time below 500ms.
- The Fabric SHOULD support 1M routes @ 10k fabric nodes with convergence time below 1s.

- The Fabric routing protocol MUST support load balancing using ECMP, wECMP and UCMP.
- The Fabric routing protocol MUST support and provide facility for topology-specific algorithms that enable correct operations in that specific topology.
- The Fabric routing protocol SHOULD support route scale and convergence times of a Fabric mentioned above.
- The Fabric routing protocol SHOULD support ECMP as wide as 256 paths.
- The Fabric routing protocol MUST support various address families that covers IP as well as MPLS forwarding.
- The Fabric routing protocol MUST support Traffic Engineering paths that are host and/or router based paths.

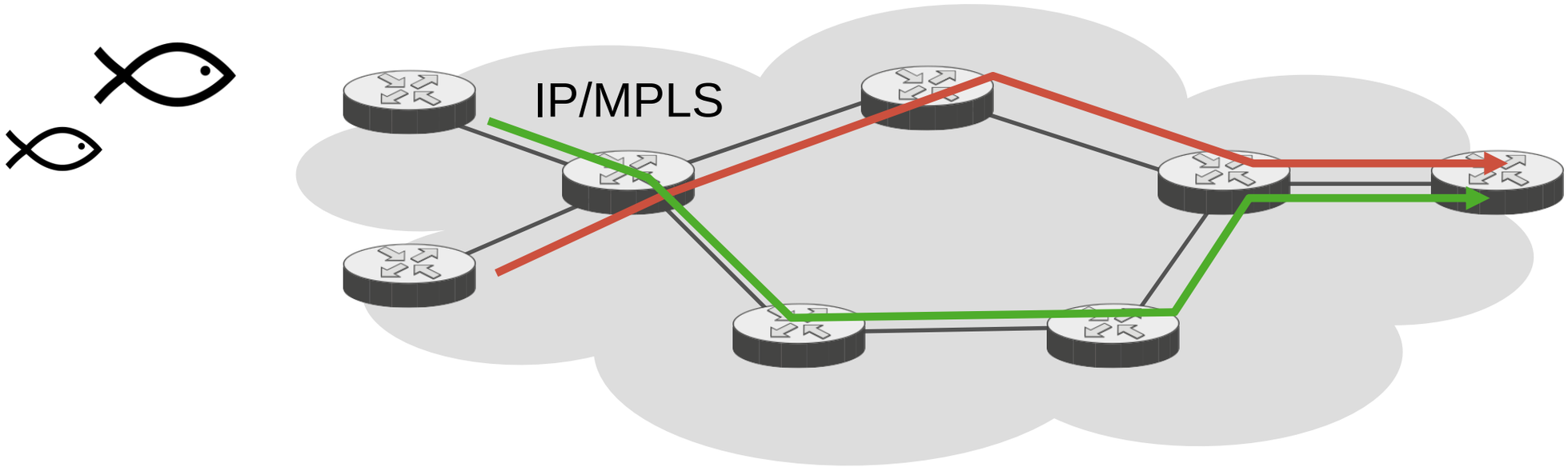
- The Fabric routing protocol MUST support Zero Touch Provisioning (ZTP).
- The Fabric routing protocol MUST support Neighbor Discovery to facilitate ZTP.
- The Fabric routing protocol MUST be able to leverage BFD [RFC5880] for neighbor state.
- The Fabric routing protocol MUST be able to support real time state notifications of routes and its neighbors state to facilitate control plane telemetry.
- The Fabric routing protocol MUST be able handle commission/decommission of a node as well as any node restart with a minimal data plane impact.

OSPF in the Service Provider Networks

- OSPF is very commonly used in the Service Provider networks, especially in the Middle East and Europe, many Service Providers use OSPF in their network, IS-IS is found in U.S Service Provider networks commonly.
    - OSPF is used in Core Networks mostly but some providers extend OSPF to the Aggregation and even to the access domains.
- In Seamless MPLS/Unified MPLS architecture, OSPF in the access network usage will be explained in detail.

OSPF and MPLS Traffic Engineering is used together in many SP networks

- OSPF is used to create shortest path routing but many Service Providers use OSPF with MPLS Traffic Engineering so they don't just use shortest path between their nodes.



Classical Fish Diagram of MPLS Traffic Engineering.
Without MPLS TE, IGP protocols always chooses shortest path.
Source routing is not possible with IGP protocols.

- OSPF is used to carry the Service Provider network device prefixes in the SP networks, not the customer routes.
    - Customer routes are carried within BGP.
- OSPF is used in Service Provider network as a PE-CE routing protocol if SP is providing MPLS L3 VPN , or mobile operators are using MPLS L3 VPN at their 3G UMTS and 4G LTE sites in Unified/MPLS architecture.

OSPF Design Best Practices

- Unless there is a valid reason, don’t deploy Multi Area OSPF, keep the design simple, it provides better convergence, less configuration on the ABR nodes and optimal traffic flow.

- Don’t enable OSPF on the customer facing ports, for MPLS L3 VPN PE-CE protocol, enable prefix limit, authentication and control plane policing

OSPF Design Best Practices

- Use OSPF Prefix-suppression feature to remove infrastructure links from the Type 1 (Router) LSA, it provides scalability if necessary.

- Always start deploying OSPF Area 0 (Backbone Area), it will provide easier migration when multi area OSPF design is necessary.

OSPF Design Best Practices

- Use OSPF network type ‘ point to point ‘, it removes the Types 2 LSA from LSDB, thus better for troubleshooting and high availability also it is good for fast convergence.

- If there is DR in the OSPF domain, make sure you don’t have performance problem with it.

OSPF Design Best Practices

- Summarization removes reachability information and it can be done on either ABR for summary LSA or at ASBR for External Type 5 LSA.
    - Summarization may break the MPLS LSP, since LDP cannot have aggregated FEC unless the RFC 5283 – LDP Extension to Inter Area LSP is in use.

OSPF Design Best Practices

- If PE loopback mask is /24, OSPF advertises it as /32 but LDP assigns a label for /24, since there is a mismatch between two control plane protocols (LDP and OSPF), packet is dropped. Because OSPF advertises loopback interfaces as /32. They should follow each other.
    - Either OSPF network type should be point to point to advertise loopback as /32 so routing table and LDP is same, or use /32 loopback subnet mask.

OSPF Design Best Practices

- Don’t redistribute full Internet routing table to OSPF.
    - OSPF in the large scale datacenter has flooding issue, database filter-out can be used to remove the topology information, towards downstream TOR switches.
- If you need to deploy Multi Area Design , know that it can create suboptimal routing in many topologies.

OSPF Design Best Practices

- Don’t deploy more than two ABRs for redundancy, two is enough.
    - ABRs slow down the convergence.
- Don’t carry customer prefixes with the infrastructure OSPF in Service Provider networks, customer prefixes should be carried in BGP.

OSPF Design Best Practices

- OSPF Fast convergence might bring unstability to network, make sure timers are tuned accordingly for the fast convergence.

- OSPF Fast reroute with LFA may not cover every topology, especially ring will not be protected, you may need to deploy Remote LFA or MPLS TE FRR for that , if topology is partial/full mesh, OSPF and LFA is enough to provide FRR for links or prefixes.

OSPF Design Best Practices

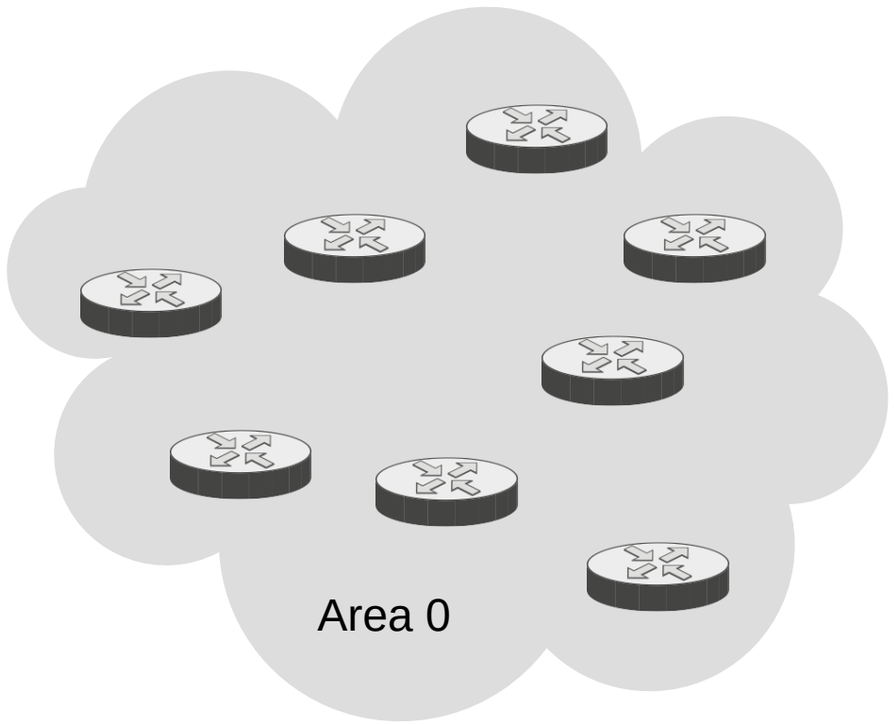
- OSPF doesn’t use TLV encoding, it is not extendable, required OSPFv3 for IPv6 for example.
    - OSPF has 11 Type of LSA, compare to 2 Levels of IS-IS it is considered as more complex.
- Each OSPF LSA has a separate header, IS-IS TLVs share common LSP header, thus OSPF is seen as less scalable.

OSPF Design Best Practices

- OSPF needs an IP address for adjacency, IS-IS doesn’t require an IP Address for neighborship, remote attack to the IS-IS is hard if not impossible, thus IS-IS is seen more secure compare to OSPF.
    - OSPF provides MPLS TE supports, similar to IS-IS, but distance vector protocols don’t.
- OSPF is a good protocols for those who look Enterprise level and standard base protocol.

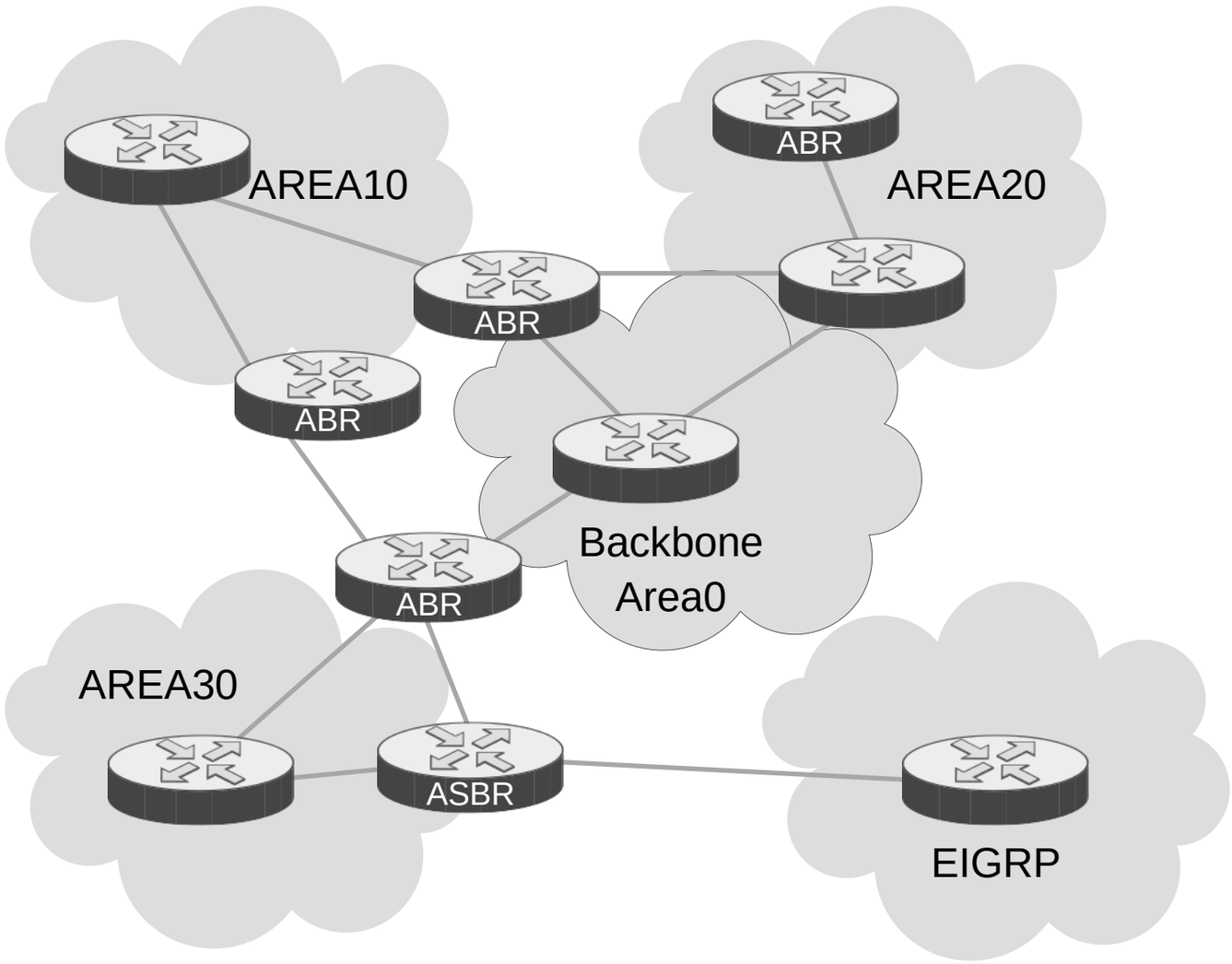
OSPF Frequently Asked Questions

- How many routers should be in one OSPF Area ?



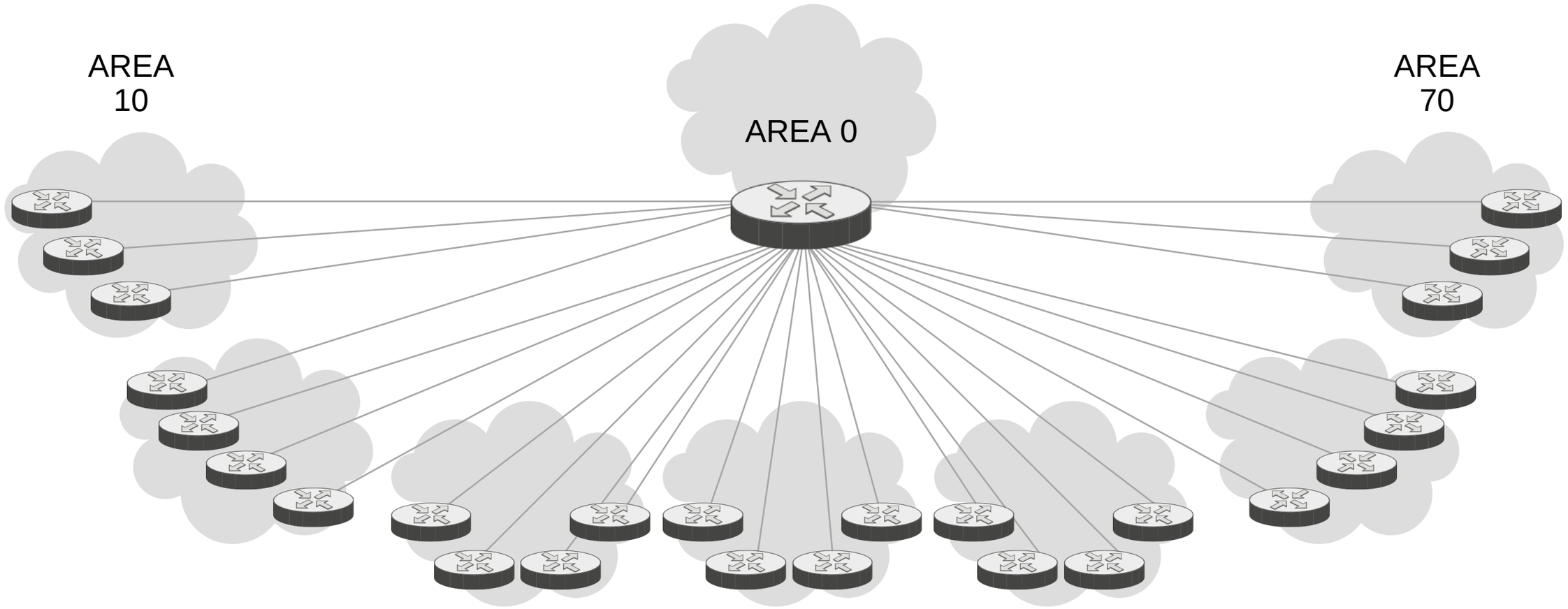
- Number of neighbor is more important question which we should ask.
- Always try to keep router LSA under the MTU size to avoid fragmentation.
- Routers cannot deal with fragmentation and reassembly well.

How many ABR (Area Border Router) per OSPF Area ?



- In the previous diagram, there are 2 ABRs in Area 10. For the redundancy and optimal traffic flow, two is always enough.
    - More ABRs will create more Type 3 LSA replication within the back bone and non-back bone areas.
- In large scale OSPF design, number of ABRs will have an huge impact on number of prefixes.

How many OSPF area is suitable per OSPF ABR?

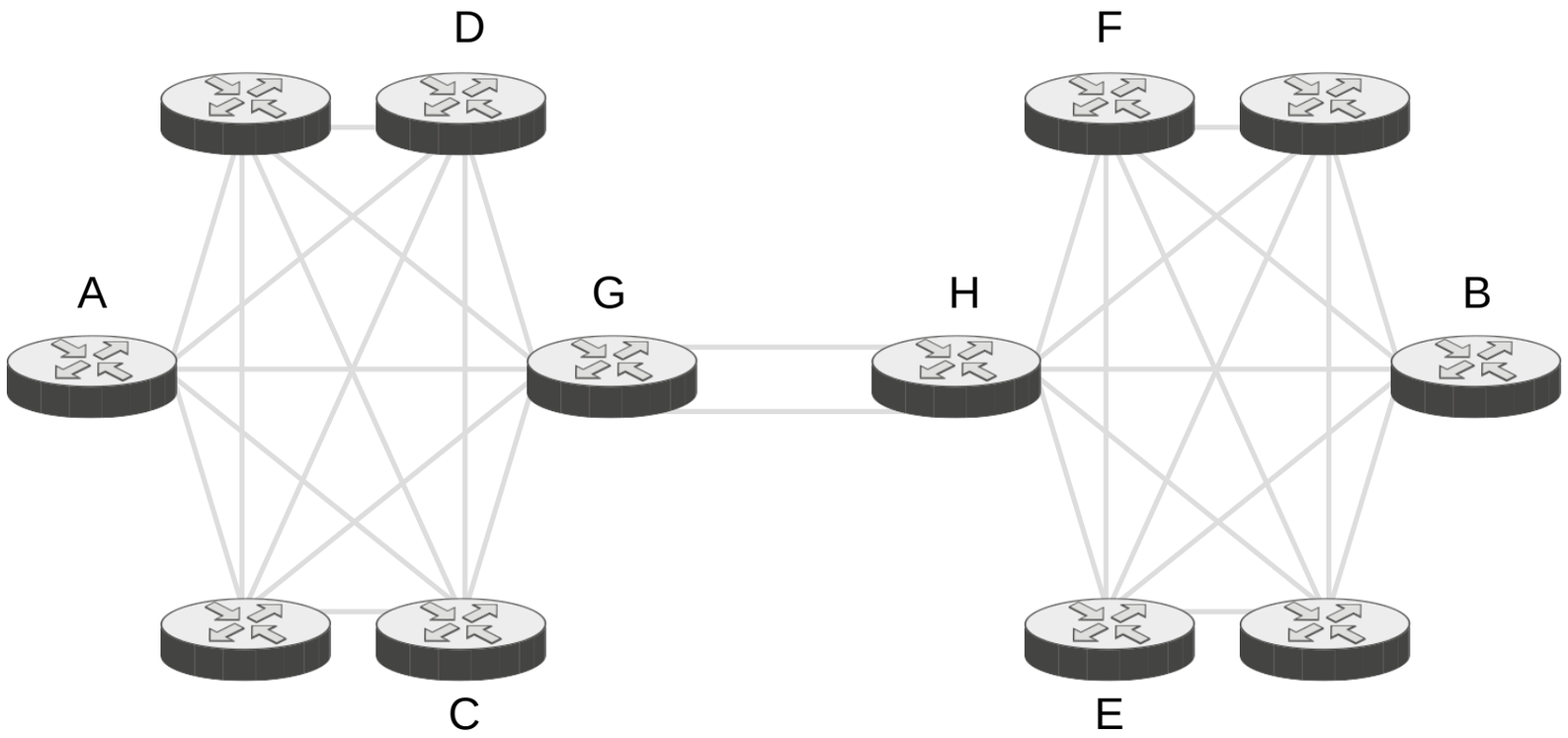


- More Areas per ABR might create a resource problem on the ABR.
- Much more Type 3 LSA will be generated by the ABR
- Between the Areas there will not be Type 1 or Type 2 LSA, Type 1 and Type 2 LSA stays in the area and the reachability information is sent as Type 3 LSA between the Areas.

OSPF CASE STUDIES

ABR Placement

- Where should we place an ABR in the below topology. Why?

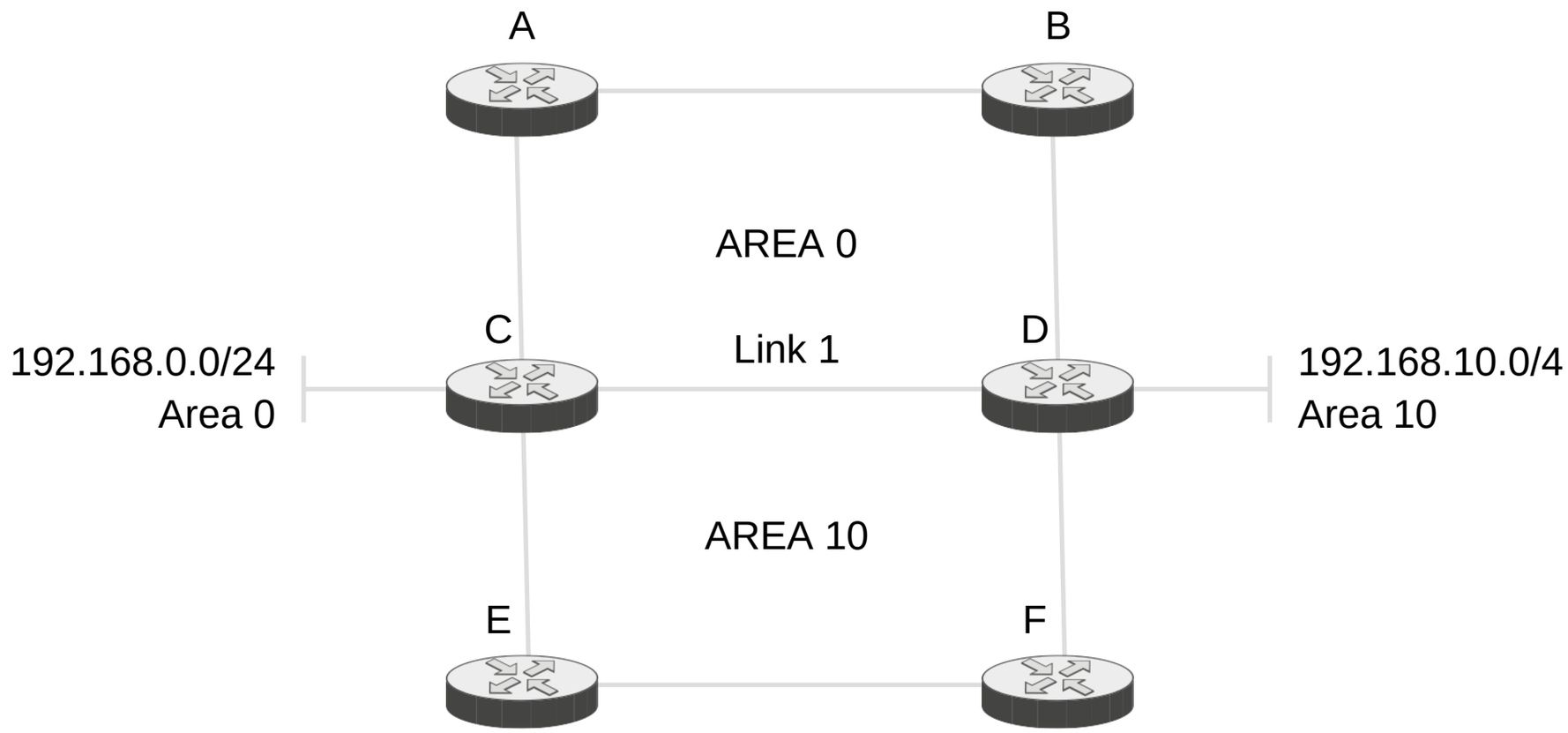


ABR Placement

- Between Router A and Router B there are 1800 different paths. (5x6) x 2 (5x6) If we would put all of them in a same area we would have flooding, convergence, resource utilization, troubleshooting problems.
    - If we use Router G or Router H as an ABR, we will have only 32 paths max (5x6) +2 between Router A and B, this will greatly reduce the load on the resources, reduces the overall complexity thus makes troubleshooting easier.
- Put ABR always a place where you can separate the complex topologies.

Multi-Area OSPF Adjacency

- What is the path from Router C to 192.168.10.0/24 and path from Router D to 192.168.0.0/24 networks? Is there a problem with the path? Why? What is the possible solution?



Multi-Area OSPF Adjacency

- If Link 1 is in area 0, router C will choose an path through E, F, and D to 192.168.10.0/24 rather than Link1.
  - This is because OSPF always prefers intra-area routes over inter-area routes.
- If Link 1 is put in area 10, router D will choose an path through B, A, and C to 192.168.0.0/24 with the same reason.

Multi-Area OSPF Adjacency

- This is suboptimal. Placing link into Area 1 and creating virtual link was the temporary solution. Also in this solution for each additional non-backbone area new OSPF adjacency is required.

- Real solution to this: RFC 5185 -OSPF Multi Area Adjacency.

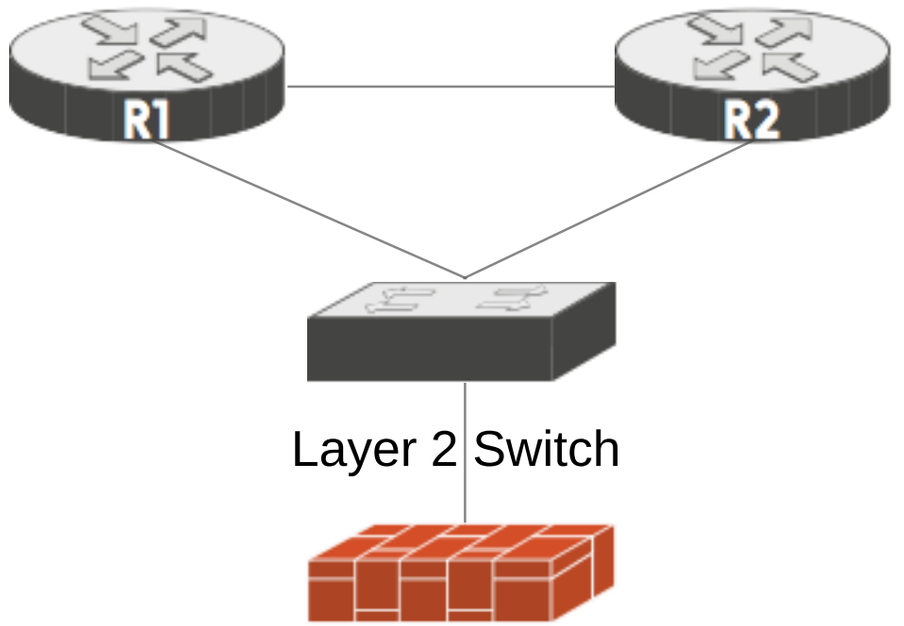
Multi-Area OSPF Adjacency

- Over 1 OSPF adjacency multiple area can be allowed with the RFC 5185.
    - Below is a sample configuration from the Cisco device which supports RFC 5185.

```
rtr-C(config)# interface Ethernet 0/0
rtr-C(config-if)# ip address 192.168.12.1 255.255.255.0
rtr-C(config-if)# ip ospf 1 area 0
rtr-C(config-if)# ip ospf network point-to-point
rtr-C(config-if)# ip ospf multi-area 2
```

NSSA at the Internet Edge

- Enterprise company wants to run OSPF at the Internet edge between their Internet Gateway routers and the firewalls, which type of OSPF area is most suitable in this design and why?



NSSA at the Internet Edge

- Solution: If OSPF is used at the Internet Edge, IGW(Internet Gateways) don’t need to have full OSPF routing table.
    - Using Stub or NSSA areas is most suitable. Firewalls only need a default routes from the Internet Gateways.
- Default route, partial route or even full route can be received from the BGP neighbor but only default route is needed by the firewalls.

NSSA at the Internet Edge

- It is good practice to redistribute default route from BGP to OSPF.
  - If the link fails between the customer and the service provider, BGP goes down and default route is removed from the OSPF as well.
- Only NSSA allows redistribution into an OSPF Stub areas.
  - That’s why, if OSPF will be implemented NSSA would be the most suitable area types on the Internet Edge.

OSPF and BGP Interaction

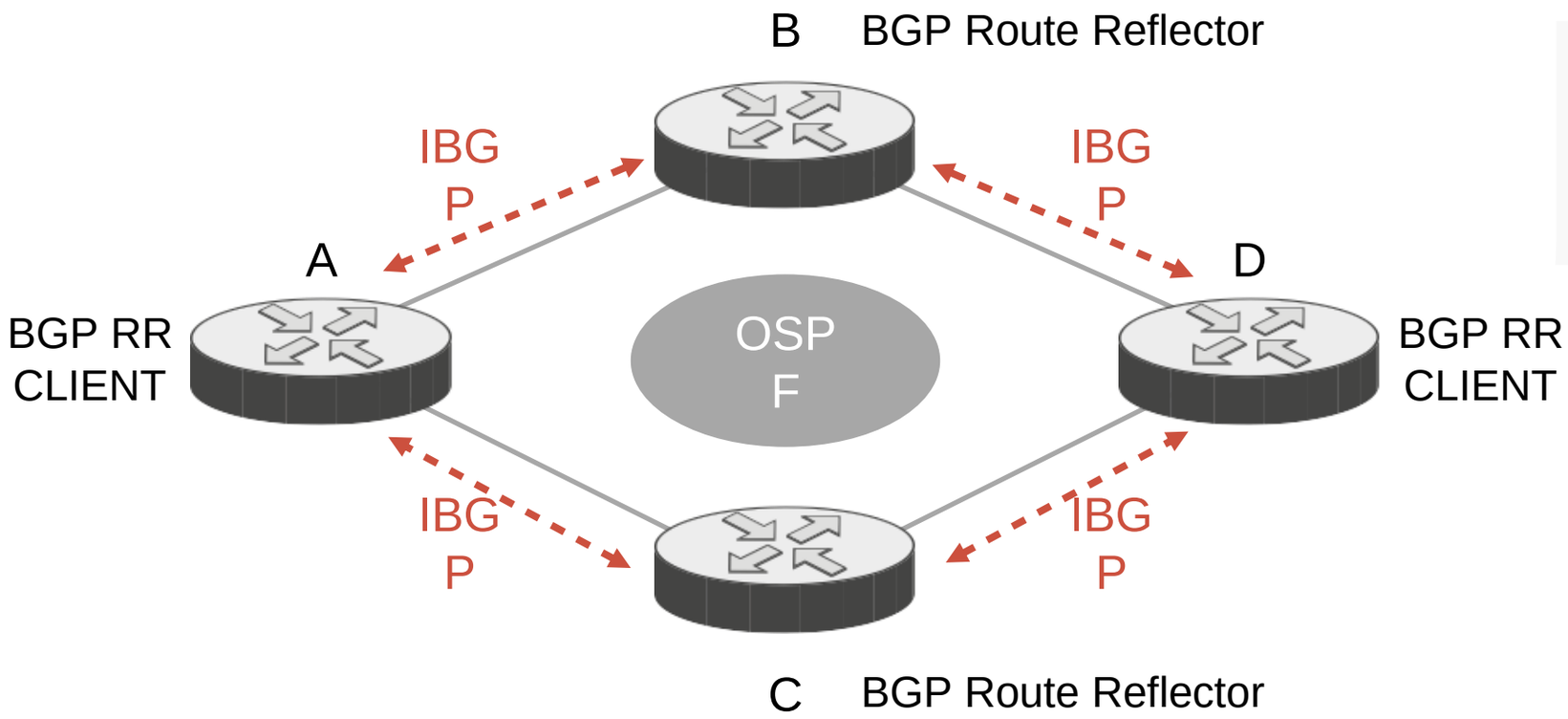
- OSPF is running as an IGP protocol in the below network. Also there is no MPLS in the core and all routers run BGP.

For scaling purpose company decided to use BGP Route Reflector design.

Router B and C are the Route Reflectors and Router A and D are the Route Reflector clients

Company wants to perform maintenance on the Router B but they don’t want to have any downtime

What would be your design recommendation ?



OSPF and BGP Interaction

- BGP as an overlay protocol needs next hop reachability. Static routing or the dynamic routing protocol is used to create an underlay network infrastructure for the overlay protocols such as BGP, LDP, PIM and so on.

- In this case study one of the routers which is in the path towards BGP next hop will be reloaded. We might have two problems here.

OSPF and BGP Interaction

- When Router B is reloaded traffic is going to Router B shouldn’t be dropped. Router B should signal the other OSPF routers. This signaling is done with OSPF Stub Router advertisement feature.
    - ‘ max-metric router-lsa ‘ is used by OSPF for graceful restart
- IGP always converges faster than BGP.
    - Second problem is when the Router B comes back, BGP traffic towards Router B will be black holed, because IGP process of Router B will converge faster than its BGP.

OSPF and BGP Interaction

- IGP should wait to BGP. Router B should take the BGP traffic once BGP prefixes installed in the routing table.
  - This is done with the OSPF Stub router advertisement feature as well.
- ‘ max-metric router-lsa on-startup wait-for-bgp ‘ is used by OSPF, so until BGP process is converged, OSPF doesn’t take traffic.
  - In this case study, with the OSPF Stub router advertisement feature, other OSPF routers are signaled for Graceful restart and also OSPF.

Case Study Key Point

- OSPF interacts with many protocols in the network such as spanning tree, BGP, MPLS and so on. Understanding the impact of such an interaction is the first step for the robust network design.

OSPF in the CCDE Exam

- OSPF Areas and LSA Types should be known very well.
    - ABR placement is an important topic, When there is a DC, Branches, WAN modules, where ABR will be placed?
- OSPF in an MPLS VPN environment , superbackbone, sham-link, route advertisement should be expected.
    - In general OSPF Scaling, Multi Area design needs to be understood very well.

Summary

- Link state protocols behaviors are explained
- OSPF Fast Convergence and Fast Reroute
- OSPF Scalability,Multi Area OSPF Design
- Overlay Technologies and OSPF (GRE, mGRE, DMVPN, LISP)
- OSPF in the Datacenter Networks
- OSPF in the Service Provider Networks
- OSPF Design Best Practices
- OSPF Advantages and Disadvantages
- OSPF Frequently Asked Questions – How many Routers in an OSPF Area, How many ABR per Area?
- OSPF in the CCDE Exam

OSPF

Open Shortest Path First

QUIZ

Question 1

How many routers can be placed in any given OSPF area?

A. 50

B. 100

C. 250

D. Less than 50

E. It depends

Answer 1

E. It depends

As it is explained in the OSPF chapter, you cannot have a numeric answer for this question.

There is no numeric answer of this question. It depends on how many links each router have, stability of the links, hardware resources such as CPU and Memory of the routers and physical topology of the network.

For example in full mesh topology, every router is connected to each other and number of links is too much compare to ring or partial topologies.

Thus, in one OSPF network you may place 50 routers in one OSPF area, but other OSPF network can have 100s of routers in one area.

Question 2

Why many different types of LSAs are used in OSPF? (Chose all that apply)

A. Provides Scalability

B. Allow Multi-Area OSPF design

C. Provides fast convergence

D. Provides High Availability

E. Better Traffic Engineering

Answer 2

A. Provides Scalability

B. Allow Multi-Area OSPF design

Question here is asking the reason of having multiple different types of OSPF LSAs. As you have seen in the OSPF chapter there are 11 different types of OSPF LSAs.

Although there are other reasons to use OSPF LSAs, two important ones are scalability and Multi-Area design. They don’t help for fast convergence or high availability LSAs are not related with High Availability or Fast convergence. Although MPLS Traf c engineering can use OSPF Opaque LSAs for the distributed CSPF calculation, CSPF is not mandatory and many networks which have MPLS Traf c engineering uses Of ine Path calculation tool such as Cariden Mate.

Question 3

What does topology information mean in OSPF?

A. IP addresses of the directly connected interface.

B. IP addresses of the loopback interfaces of all the routers.

C. Provides an IP reachability information and the metric of all the physical and logical interfaces.

D. Provides a graph of the OSPF network by advertising connection information such as which router is connected to which one and the metric of the connections.

Answer 3

D. Provides a graph of the OSPF network by advertising connection information such as which router is connected to which one and the metric of the connections

There are two type of information is provided in link state protocols: Topology and reachability information.

Reachability information means IP addresses of the physical or logical interfaces of the routers. Topology information explains, which router is connected to which one, what is the OSPF metric value between them, thus provide a graph of the OSPF network.

Based on this information every router runs SPF algorithm to nd a shortest path to each and every destination in the network.

Question 4

Why more than one Area is used in an OSPF network?

A. They are used for high availability.

B. They are used for easier troubleshooting.

C. They are used to provide scalability by having smaller flooding domains.

D. Since topology information is not shared between OSPF areas, they provide better security.

Answer 4

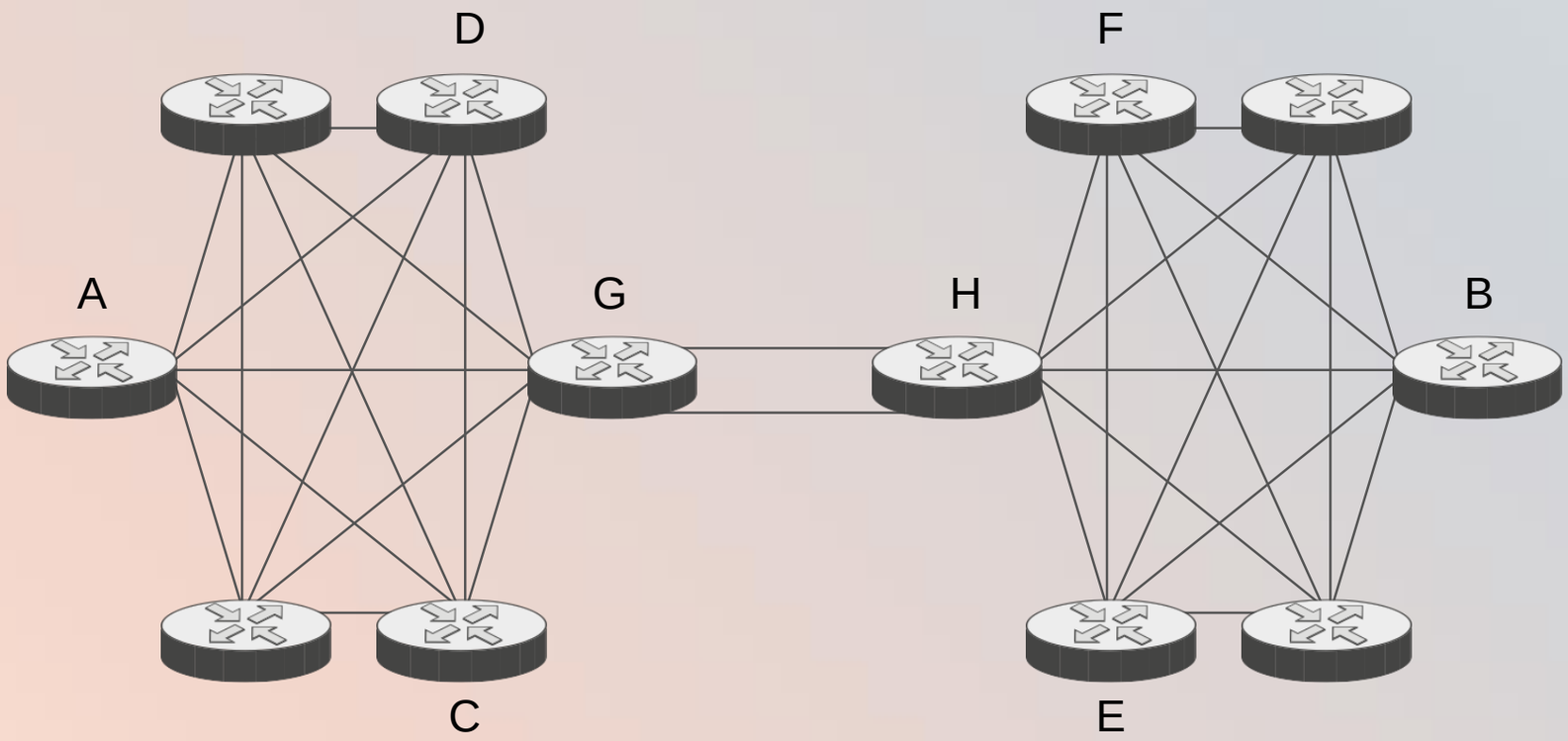
C. They are used to provide scalability by having smaller flooding domains.

OSPF areas are used mainly for scalability. Having smaller domain means, keeping topology information in an area and not sending between the areas. More than one area doesn’t provide high availability and doesn’t make troubleshooting easier.

Also in OSPF having more than one area doesn’t prevent a route to be propagated to other areas by default, it requires manual configuration and even in that case it doesn’t bring extra security.

Question 5

Which router in the below topology should be an ABR?



A. G or H B. A or B C. C or D D. E or F E. G

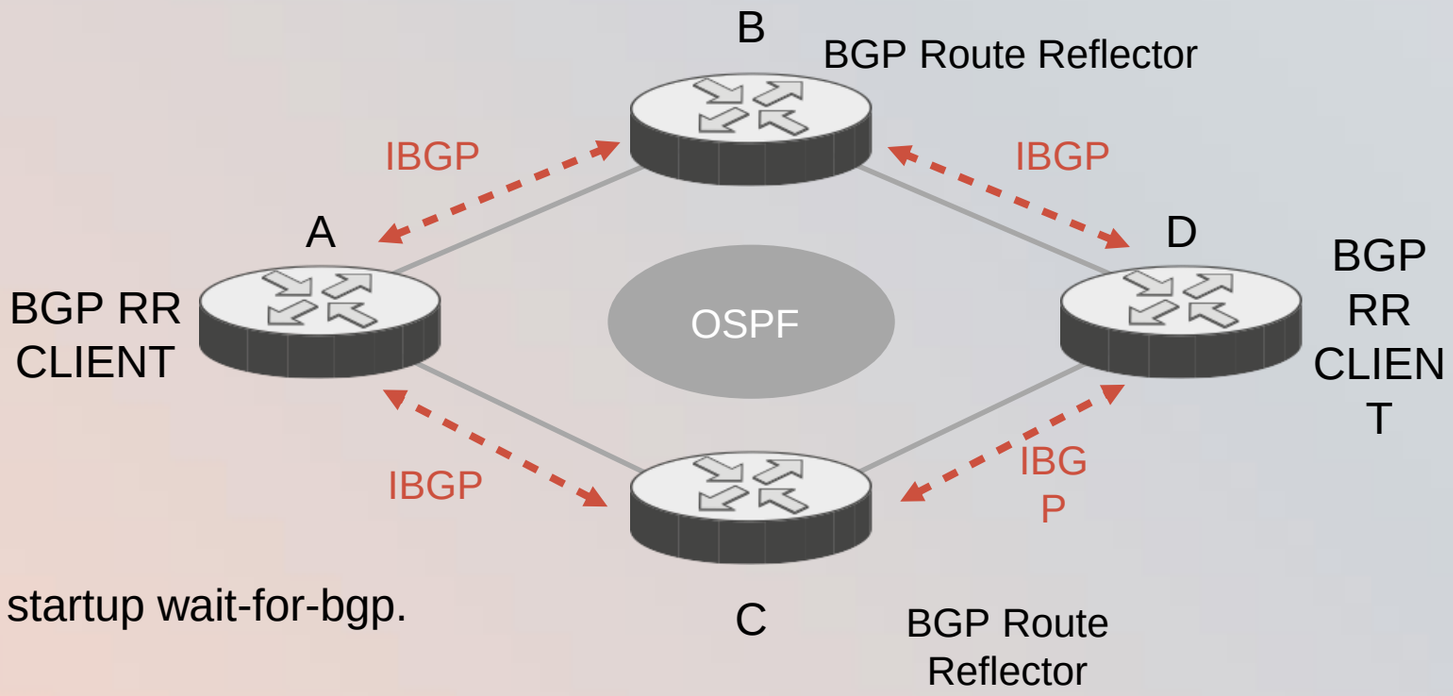
Answer 5

A. G or H

Router G or H should be an ABR to separate two full mesh topology from each other. Otherwise each router in the top full mesh network would run full SPF algorithm for each other router in the below full mesh network in case link failure, metric change or when new link or pre x is added.

Question 6

In the below topology, Router B needs to be reloaded. Network operator doesn't want any traffic loss during and after Router B's maintenance operation. Which feature should be enabled on the Router B?



A. Max-metric router-lsa on startup wait-for-bgp.

B. OSPF prefix-list.

C. Type2-lsa on-startup wait-for-bgp.

D. IGP LDP synchronization.

Answer 6

A. Max-metric router-lsa on startup wait-for-bgp.

BGP as an overlay protocol needs next hop reachability. Static routing or the dynamic routing protocol is used to create an underlay network infrastructure for the overlay protocols such as BGP, LDP, PIM and so on.

One of the routers in the forwarding path towards BGP next hop will be reloaded. We might have two problems here.

When Router B is reloaded, traf c is going to Router B shouldn’t be dropped. Router B should signal the other OSPF routers.

This signaling is done with OSPF Stub Router advertisement feature. ‘ Max-metric router-lsa ‘ is used by OSPF for graceful restart. Second problem is when the Router B comes back; BGP traffic towards Router B will be black holed, because IGP process of Router B will converge faster than its BGP.

IGP should wait to BGP. Router B should take the BGP traf c once BGP pre xes installed in the routing table.

This is done with the OSPF Stub router advertisement feature as well.

Question 7

How many levels in OSPF hierarchy used ?

A. One

B. Two

C. Three

D. As many as possible

Answer 7

B. Two

OSPF supports two level of hierarchy. Hierarchy is common network design term, which is used to identify the logical boundaries. Backbone area and Non-Backbone areas are the only two areas, which are supported by OSPF, thus it supports only two level of hierarchy.

Question 8

Which below options are correct for OSPF ABR?
(Choose all that apply)

A. It slows down the convergence.

B. It generates Type 4 LSA in Multi Area OSPF design.

C. It does translation between Type7 to Type 5 in NSSA area.

D. It does translation between Type 5 to Type 7 in NSSA area.

E. It prevents topology information between OSPF areas.

Answer 8

A. It slows down the convergence.

B. It generates Type 4 LSA in Multi Area OSPF design.

C. It does translation between Type7 to Type 5 in NSSA area.

E. It prevents topology information between OSPF areas.

OSPF ABR slows down the network convergence. Because it needs to calculate for each Type 1 and Type 2 LSAs, corresponding Type 3 LSAs and send its connected OSPF areas.

OSPF ABR generates Type 4 LSAs in Multi Area OSPF Design. When ABR receives the external pre xes in an Area, it translates Type 1 LSAs of the ASBR to Type 4 LSA and sends it to the other areas.

In NSSA Area, ABR translates Type 7 LSA to Type 5 LSA, but there is no Type 5 to Type 7 LSA translation. It is not allowed.

Topology information is not sent between the OSPF Areas, ABR stops topology information.

Thus the answer of this question is A- B – C- E.

Question 9

Why Designated Router is used in OSPF network?

A. It is used to have an ABR in the network

B. It is used to create topology information

C. It is used to centralize the database, instead of keeping distributed OSPF link state database in every node

D. It is used to avoid flooding information between each device in multi access OSPF network

Answer 9

D. It is used to avoid flooding information between each device in multi access OSPF network

Designated Router (DR) is used to avoid flooding information between each OSPF device in Multi-Access networks such as Ethernet or Frame Relay.

Routers only send their update to DR and DR floods this information to the every router in the segment. Multicast Group addresses 2224.0.0.5 and 224.0.0.6 is used for communication in IPv4.

Question 10

Which below feature is used to avoid blackholing when OSPF and LDP are used together?

A. OSPF Fast Reroute.

B. OSPF Multi Area Design.

C. IGP LDP Synchronization.

D. Converging OSPF faster than LDP in case of failure.

Answer 10

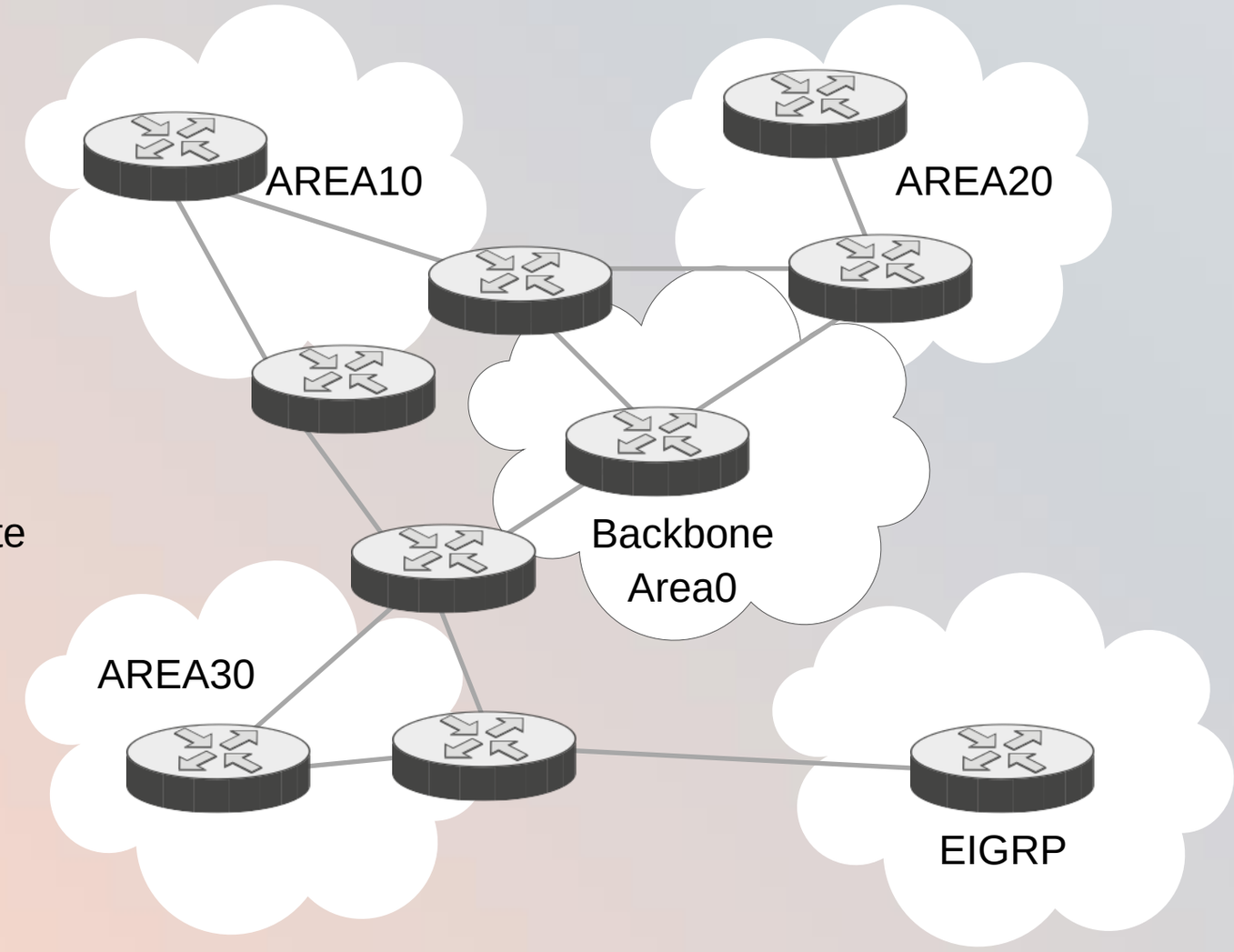
C. IGP LDP Synchronization.

The problem occurs when link or node fails when OSPF and LDP is used together. It also occurs when IS-IS and LDP is together and the IG-LDP synchronization provides a label for the IGP pre xes in the Label database, otherwise since IGP converge rst and then LDP, packets would be blackholed.

Chicken and egg problem is solved and blackholing is avoided.

Question 11

Which below option is correct for the given topology?



A. Area 20 has to be Stub area.

B. Sending default route might create suboptimal routing for internal Area 20 routers.

C. ABR of Area 20 has to be Designated Router.

D. Area 20 doesn’t receive Type 1 and Type 2 LSAs from the other areas.

Answer 11

D. Area 20 doesn’t receive Type 1 and Type 2 LSAs from the other areas.

Area 20 can be any type of OSPF area since there is no given requirement.

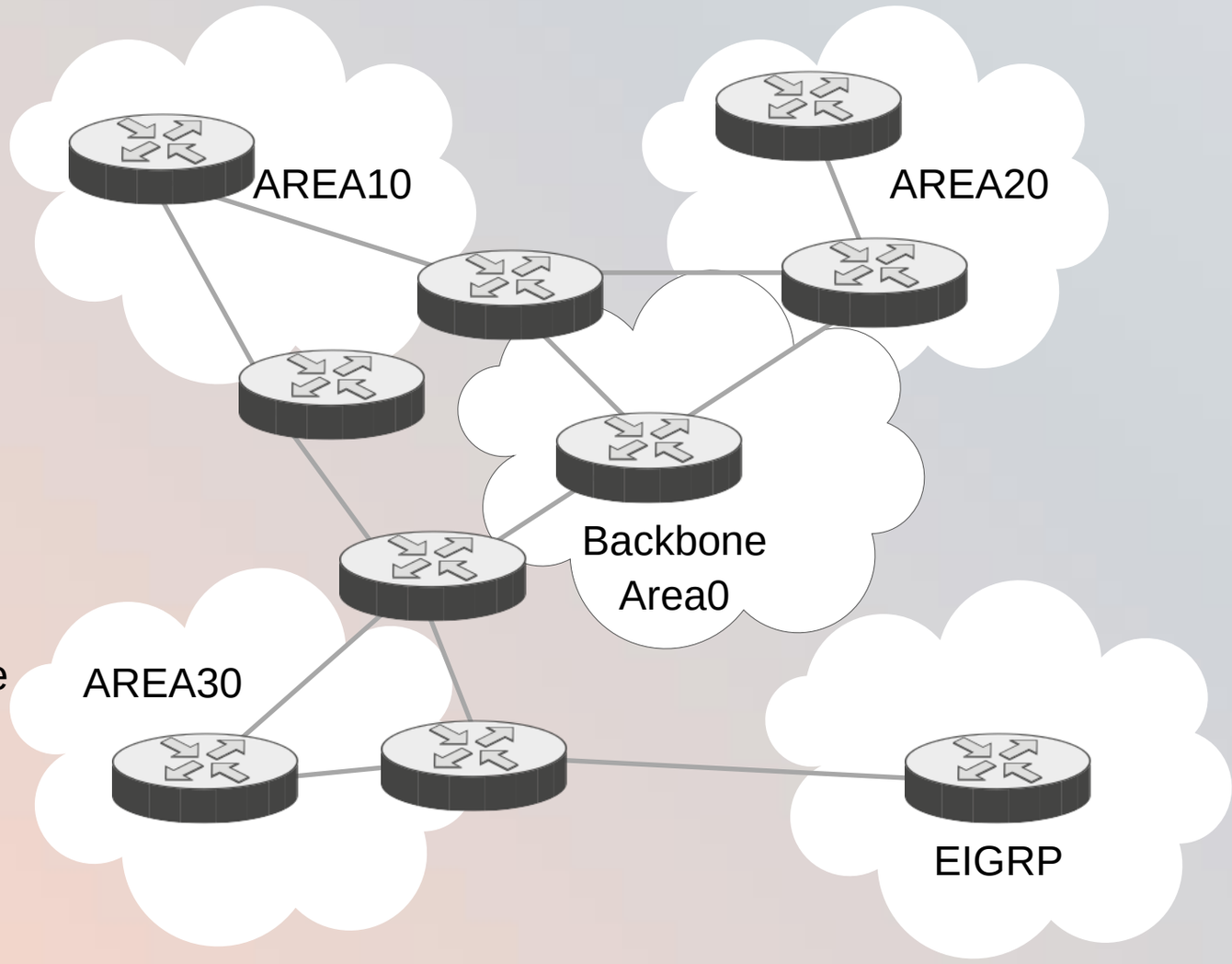
Sending default route cannot create suboptimal routing because there is only one exit point from the Area 20. Sub optimal routing can only be created if there is more than one exit from the Area.

ABR of Area 20 doesn’t have to DR. In fact, DR and ABR shouldn’t be the same router. Since both operations are resource intensive and separating these two ask is a best practice.

Type 1 and Type 2 LSAs cannot be received from the other Areas because topology information is not allowed between the OSPF areas and in OSPFv2 Type 1 and Type 2 LSAs carry topology information in addition to reachability information.

Question 12

In the below topology Area 30 is an NSSA area. Which below option is true?



A. There will not be any Type 3 LSA in Area 30.

B. ABR of Area 30 will translate Type 7 LSA to Type 5 LSA.

C. There will not be any Type 1 or Type 2 LSA.

D. EIGRP prfixes will not be allowed in Area 30.

Answer 12

B. ABR of Area 30 will translate Type 7 LSA to Type 5 LSA.

Since Area 30 is an NSSA area; there will be Type 3 LSA, that's why Option A is incorrect. There will be Type 1 and Type 2 LSA, but not from the other Areas.

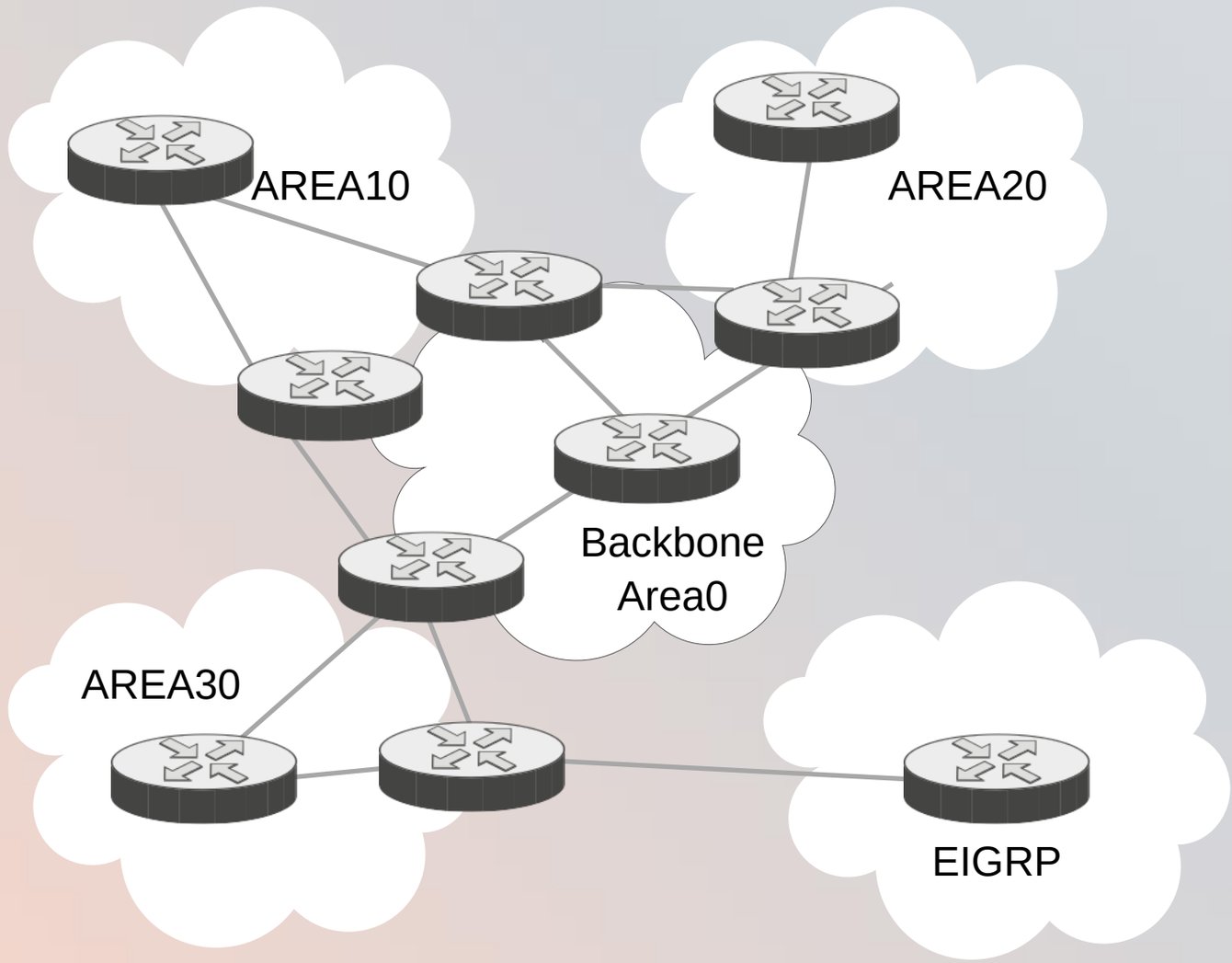
In Are 30, every router generates Type 1 LSAs, and of there is multi- access network, the DR will generate Type 2 LSA as well.

EIGRP preFIxes will be allowed and they will be seen as Type 7 LSA in the Area 30.

Only Option B is correct, because ABR of Area 30 translate Type 7 LSA which is the EIGRP pre xes to Type 5 LSA send them to the network.

Question 13

In the below topology Area 10 is Totally NSSA Area. Which below option is true?



A. Area 10 will not have any Type 1 or Type 2 LSA.

B. Area 10 will not have EIGRP prefixes.

C. Area 10 cannot reach to EIGRP prefixes.

D. Both ABRs of Area 10 will do the Type 7 to Type 5 translation.

Answer 13

B. Area 10 will not have EIGRP prefixes.

Area 10 will be able to reach EIGRP network through default route even if it is Totally NSSA. But Area 10 devices cannot have specific EIGRP prefixes because Type 3, 4, 5 LSAs are not allowed in Totally NSSA Area. Answer of this question is B.

Question 14

Which below topology, OSPF is worse than EIGRP in large-scale implementation?

A. Full Mesh

B. Partial Mesh

C. Hub and Spoke

D. Ring

Answer 14

C. Hub and Spoke

In Full Mesh physical topology, Mesh Group feature allows only two routers to flood LSAs into the area. Mesh Group is supported by both OSPF and IS-IS.

This brings scalability into OSPF.

Ring and Partial mesh topologies are hard for all the routing protocols. Ring and Partial mesh are cheaper to build but convergence, optimal routing and fast reroute is very hard in Ring and Partial mesh.

EIGRP is best in Hub and Spoke topology from the scalability point of view, because it doesn’t require so many configurations for its operation. OSPF on the other hand, requires a lot of tuning for its operation in Large scale Hub and spoke topology.

Question 15

Why OSPF is used as an Infrastructure IGP in an MPLS VPN environment?

A. To carry the customer prefixes.

B. Reachability between the MPLS VPN endpoints.

C. OSPF is not used in MPLS VPN environment as an Infrastructure IGP protocol but BGP is used.

D. LDP requires OSPF as an IGP.

Answer 15

B. Reachability between the MPLS VPN endpoints.

LDP requires IGP yes but it is not relevant. It could be EIGRP or IS-IS as well.

And the purpose of OSPF or any other IGP.as an Infrastructure protocol is to carry the loopback interface addresses of the MPLS VPN endpoints.

So the OSPF is used for reachability between the VPN endpoints (PE devices) in SP networks. OSPF is not used to carry the customer prefixes as an Infrastructure IGP.

Knowing the difference between the Infrastructure IGP and the PE-CE IGP protocol in MPLS VPN is important. This will be explained in detail in the MPLS chapter.

Question 16

Which OSPF feature in MPLS VPN PE-CE is used to ensure MPLS service is always chosen as primary link?

A. OSPF max-metric

B. OSPF prefer-primary path

C. OSPF sham-link

D. Passive-interface

E. Virtual link

Answer 16

C OSPF sham-link

Even domain IDs are the same in both site of the MPLS VPN, without sham-link feature only Type 3 LSA can be received from the PE by CE.

Sham-link is used to receive Type 1 LSA and even if there is a backup connection between the CEs, only changing cost on either PE-CE or CE-CE link make MPLS link as primary.

OSPF as a PE-CE protocol will be explained in detail in the MPLS chapter.

Question 17

Which below options are correct for OSPF?
(Choose all that apply)

A. OSPFv2 doesn’t support IPv6 so when IPv6 is needed, OSPFv3 is necessary.

B. OSPF virtual link shouldn’t be used as permanent solution is OSPF design.

C. OSPF and BGP are the two separate protocols so when OSPF cost changes, it doesn’t affect BGP path selection.

D. OSPF can carry the label information in Segment Routing so LDP wouldn’t be necessary.

E. OSPF unlike EIGRP, supports MPLS Traffic Engineering with dynamic path calculation.

Answer 17

A.OSPFv2 doesn’t support IPv6 so when IPv6 is needed, OSPFv3 is necessary.

B.OSPF virtual link shouldn’t be used as permanent solution is OSPF design.

D.OSPF can carry the label information in Segment Routing so LDP wouldn’t be necessary.

E.OSPF unlike EIGRP, supports MPLS Traf c Engineering with dynamic path calculation.

Only incorrect option of this question is C. although they are two separate protocols; changing the OSPF metric can affect the best BGP exit point.

Taking IGP cost into consideration to calculate best path for the BGP prefixes is called Hot Potato Routing.

Changing IGP metric can affect BGP best path.

Question 18

What is the reason to place all routers in Area 0/Backbone Area, even at the begiining in OSPF design?

A. You cannot place routers in non-backbone area without backbone area.

B. Type 3 LSAs should be received from the ABR.

C. Future Multi Area design migration can be easier.

D. It is not a best practice to place all the routers in Area 0 in Flat/Single OSPF area design.

Answer 18

C Future Multi Area design migration can be easier.

In OSPF design, all the routers can be placed in any Non-Backbone area. If you have 50 routers in your network, you can place all of them in Area 100 for example.

But having the routers in OSPF Backbone area (Area 0) from the early stage of network design provides easier migration to Multi Area OSPF design.

This is true for the IS-IS as well. In IS-IS you can have all the routers in the network in Level 1 domain. But having them in Level 2 allows easier Multi-Level IS-IS design if it is required in the future. This will be explained in the IS-IS chapter with the case study.

Question 19

In OSPFv2 which LSA types cause Partial SPF run? (Choose Three)

A. Type 1

B. Type 2

C. Type 3

D. Type 4

E. Type 5

Answer 19

C. Type 3

D. Type 4

E. Type 5

In OSPFv2, Type 3, 4 and 5 causes Partial SPF run. Not full SPF. Partial SPF is less CPU intensive process compare to Full SPF run.

Question 20

Based on which design attributes, number of maximum routers change in OSPF area?

A. It depends on how many area is in the OSPF domain.

B. Maximum number of routers in OSPF area should be around 50.

C. Depends on link stability, physical topology, number of links, hardware resources, rate of change in the network.

D. If there are two or more ABRs, number can be much more

Answer 20

C Depends on link stability, physical topology, number of links, hardware resources, rate of change in the network.

Depends on link stability, physical topology, number of links on the routers, hardware resources and rate of change in the network. If some links flap all the time, this affects the routers resources and the scalability of the network.

Question 21

How many OSPF ABR routers should be in place in OSPF by keeping also redundancy in mind?

A. One

B. Two

C. Three

D. If the number of routers in an area is too much, it can be up to 8 ABRs

Answer 21

B. Two

In large-scale OSPF design, the number of ABRs will have a huge impact on the number of prefixes. Thus having two ABRs is good for redundancy for the critical sites.

For example some of the remote offices or POP locations may not be critical as other locations and having only one ABR in those locations, can be tolerated by the company.

In this case that specific location may have only one ABR as well.

Keep in mind that; two is company, three is crowded in design.

Question 22

What are the most important reasons of route summarization in OSPF? (Choose Two)

A. In order to reduce the routing table size so routers have to store and process less information.

B. In order to increase the availability of the network.

C. Increase the security of the routing domain.

D. In order to reduce the impact of topology changes.

E. In order to provide an optimal routing in the network.

Answer 22

A. In order to reduce the routing table size so routers have to store and process less information.

D. In order to reduce the impact of topology changes.

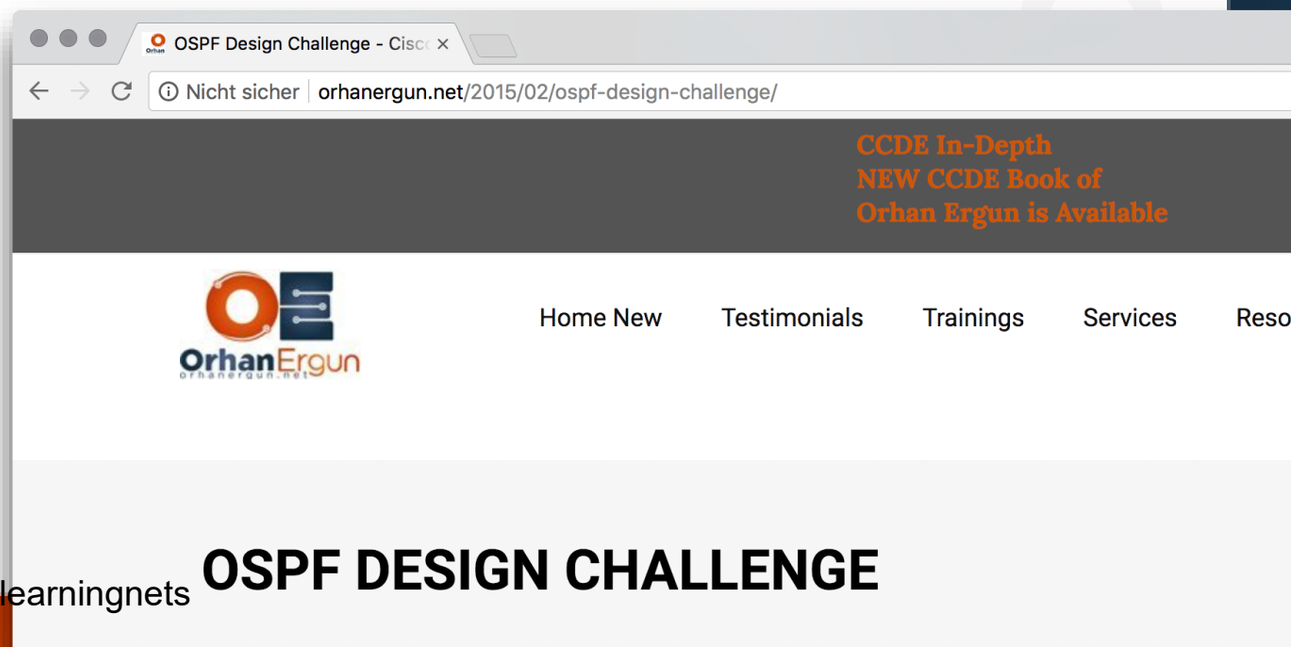
If there is route summarization, sub optimal routing might occur as it was explained in the OSPF chapter. Thus Option E is incorrect.

Availability and security doesn’t increase with route summarization. But topology change affects is de nitely reduced.

Also the routing table size is reduced and this provides better memory and CPU utilization, fast convergence and better troubleshooting.

Extra Study Resources

- Books :
- http://www.amazon.com/OSPF---Choosing-Large-Scale-Networks/dp/0321168798/ref=sr_1_1?ie=UTF8&qid=1436566360&sr=8-1&keywords=ospf+and+is-is
- Videos :
- Ciscolive Session – BRKRST -2337
- Articles :
- http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_16-2/162_lsp.html
- http://orhanergun.net/2015/02/ospf-design-challenge/
- https://tools.ietf.org/html/rfc4577



IS-IS

Intermediate System to Intermediate System

Agenda

- IS-IS Theory
- IS-IS Fast Convergence
- Convergence and Micro-loop
- Fast Reroute with IS-IS
- IS-IS Scalability, Multi Level IS-IS Design
- Overlay Technologies and IS-IS (GRE, mGRE, DMVPN, LISP)
- IS-IS in the Datacenter Networks
- IS-IS in the Service Provider Networks
- IS-IS Design Best Practices

Agenda

- Design Examples with IS-IS Areas and Levels
- IS-IS BGP Interaction and Overload Bit
- Multi Level IS-IS and MPLS Interaction
- Fast Service Restoration with IS-IS and LDP
- IS-IS in a Full-Mesh Topology
- IS-IS Advantages and Disadvantages
- Case Studies
- IS-IS in the CCDE Exam
- Summary
- Bonus Materials

Theory

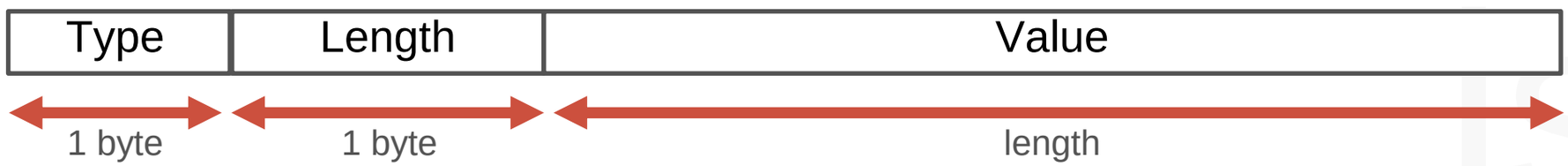
- IS-IS is a link-state routing protocol, similar to OSPF. If you are looking for Service Provider grade, MPLS Traffic Engineering support, extendible routing protocol for easier future migration then only choice is IS-IS.

- Commonly used in Service Providers, Datacenter (as an underlay) and some large Enterprise networks.

Flexibility in terms of tuning : TLV based protocols allow many parameters to be tuned and extendable

- IS-IS TLV Codes - Specified in RFC 1195

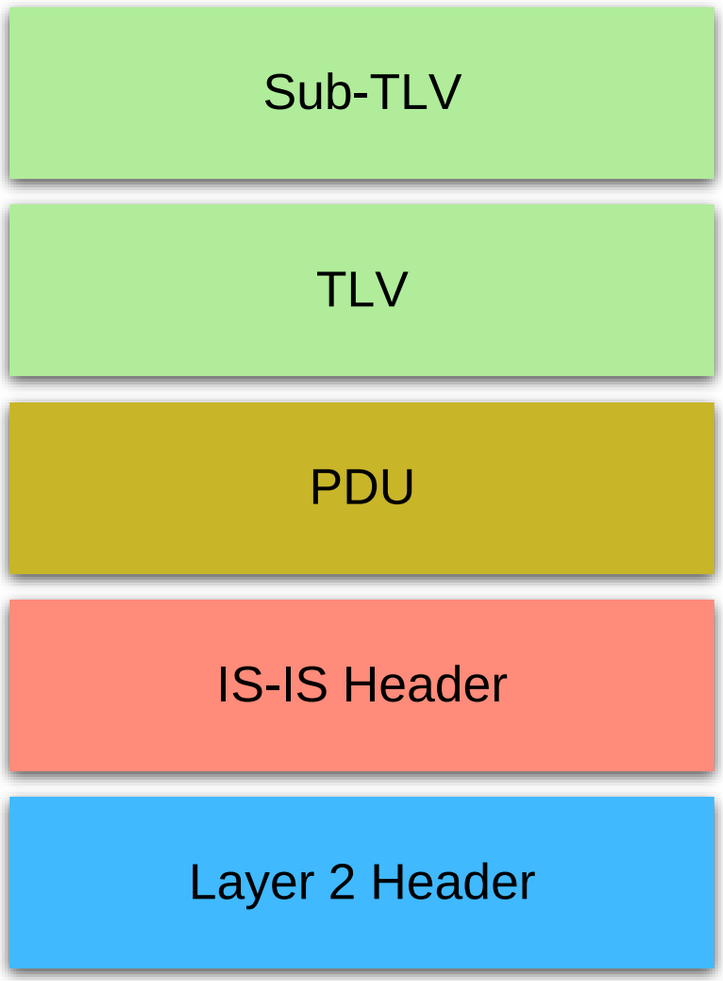
Type	TLV	Description
128	IP Internal Reachability Information	List IP addr/mask within the routing domain
129	Protocols Supported	Protocols supported by the originator (v4, v6,…)
130	IP External Reachability Information	List IP addr/mask external to the routing domain
131	Inter-Domain Routing Protocol Information	Carry information from external routing protocols transparently through the IS-IS domain
132	IP Interface Address	IP address of the interface out which the PDU was sent
133	Authentication Information	Authentication type and information



- You don’t need totally different protocol to support new extensions. In IS- IS IPv6, MTR and many other protocol just can be used with additional TLVs.

1. IPv6 Address Family support (RFC 2308)
2. Multi-Topology support (RFC 5120)
3. MPLS Traffic Engineering (RFC 3316)

TLVs are carried in LSP



- IS-IS is a Layer 2 protocol and is not encapsulated in IP, thus it is hard if not impossible to attack Layer2 networks remotely, IS-IS is considered as more secure than OSPF.

- IS-IS uses NET (Network Entity Title) address similar to OSPF Router ID.

- IP support to IS-IS is added by the IETF after ISO invented it for the CLNS. If IS-IS is used together with IP, it is called Integrated IS-IS.

- IS-IS doesn’t require IP address for the neighborship.

- ISPs commonly choose addresses as follows:
  1. First 8 bits – pick a number (49 used in these examples)
  2. Next 16 bits – area ID
  3. Next 48 bits – router loopback address (6 bytes, every 4 numbers is 2 bytes)
  4. Final 8 bits (2 Numbers) is 00 on the routers

Example

1. NET: 49.0001.1921.6800.1001.00 49.0001 is the IS-IS Area ID
2. 192.168.1.1(Router loopback) in Area1
3. 00 is the NSEL

IS-IS vs. OSPF Terminology

IS-IS & OSPF Terminology

OSPF	IS-IS
Host	End System (ES)
Router	Intermediate System (IS)
Link	Circuit
Packet	Protocol Data Unit (PDU)
Designated router (DR)	Designated IS (DIS)
Backup DR (BDR)	N/A (no BDIS is used)
Link-State Advertisement (LSA)	Link-State PDU (LSP)
Hello packet	IIH PDU
Database Description (DBD)	Complete sequence number PDU (CSNP)

IS-IS vs. OSPF Terminology

IS-IS & OSPF Terminology

OSPF	IS-IS
Area	Sub domain - Level
Non-backbone area	Level-1 domain
Backbone area	Level-2 domain (backbone)
	L1L2 router
Area border Router (ABR)	
Autonomous System Boundary Router (ASBR)	Any IS

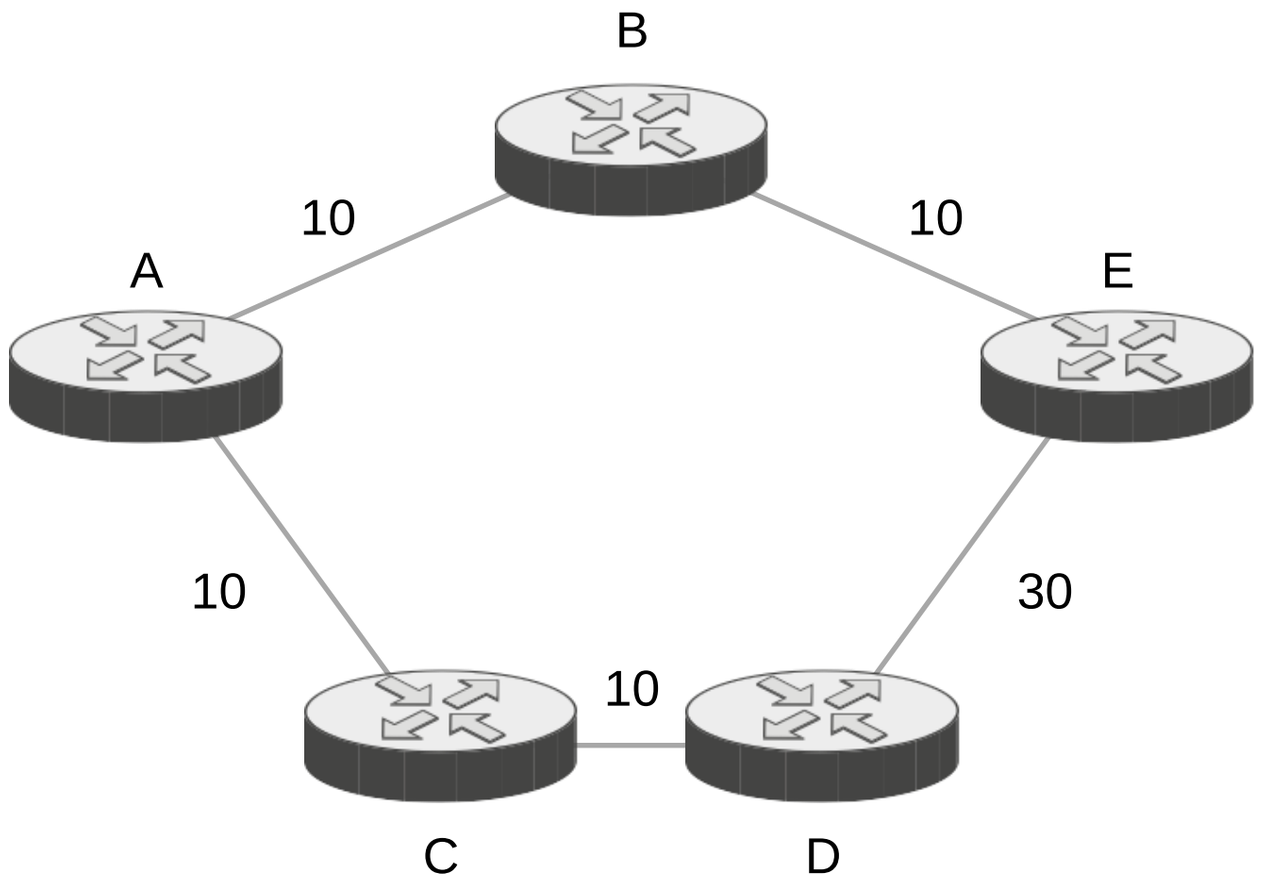
IS-IS Fast Convergence

- Similar convergence characteristic with OSPF, SPF,LSA Throttling timers, LSA Pacing, Processing and Propagation timers are tunable in IS-IS as well

- 4 Steps of convergence is applicable to IS-IS too

1. Detection
2. Propagation
3. Finding a new path
4. Installing new path to RIB and FIB

Convergence and Micro-loop



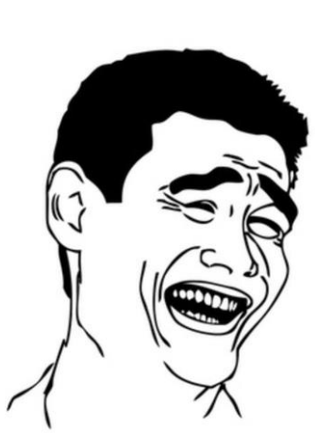
Fast Reroute with IS-IS

- IS-IS similar to OSPF, supports MPLS Traffic Engineering Fast Reroute, IP Traffic Engineering (LFA, Remote LFA) and also supports Segment Routing Fast Reroute.

- Since IS-IS is used in many large Service Providers and FRR is mainly SP requirement, FRR features with IP and MPLS are first introduced in IS-IS.

IS-IS Scalability

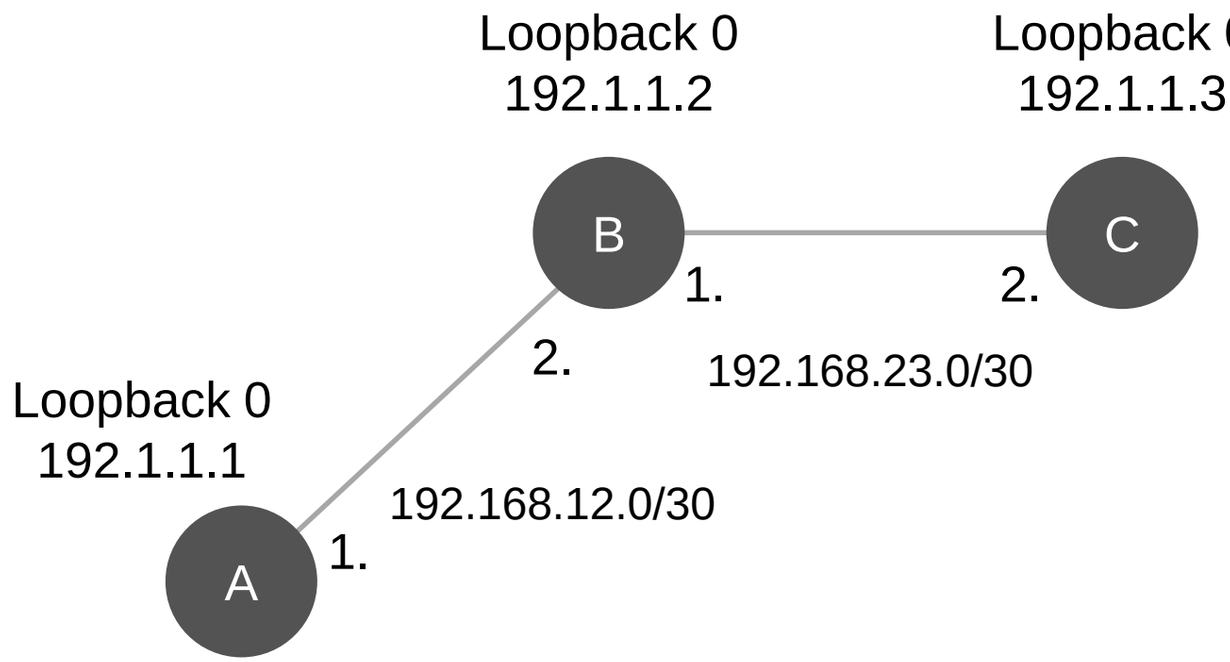
- IS-IS TLVs don’t require separate header, OSPF LSAs does, thus IS-IS is considered more scalable compare to OSPF.



ONE LSP

IS-IS	OSPF
TLV	LSA1
TLV	LSA2
TLV	LSA3
TLV	LSA4
TLV	LSA5
TLV	LSA7
TLV	LSAx

- IS-IS prefix suppression is supported as well, ‘advertise passive only‘ command, loopbacks are set as passive interface, transit links are removed from the TLV 128 (IP Reachability) or 135 (Extended IP Reachability/Wide Metric).



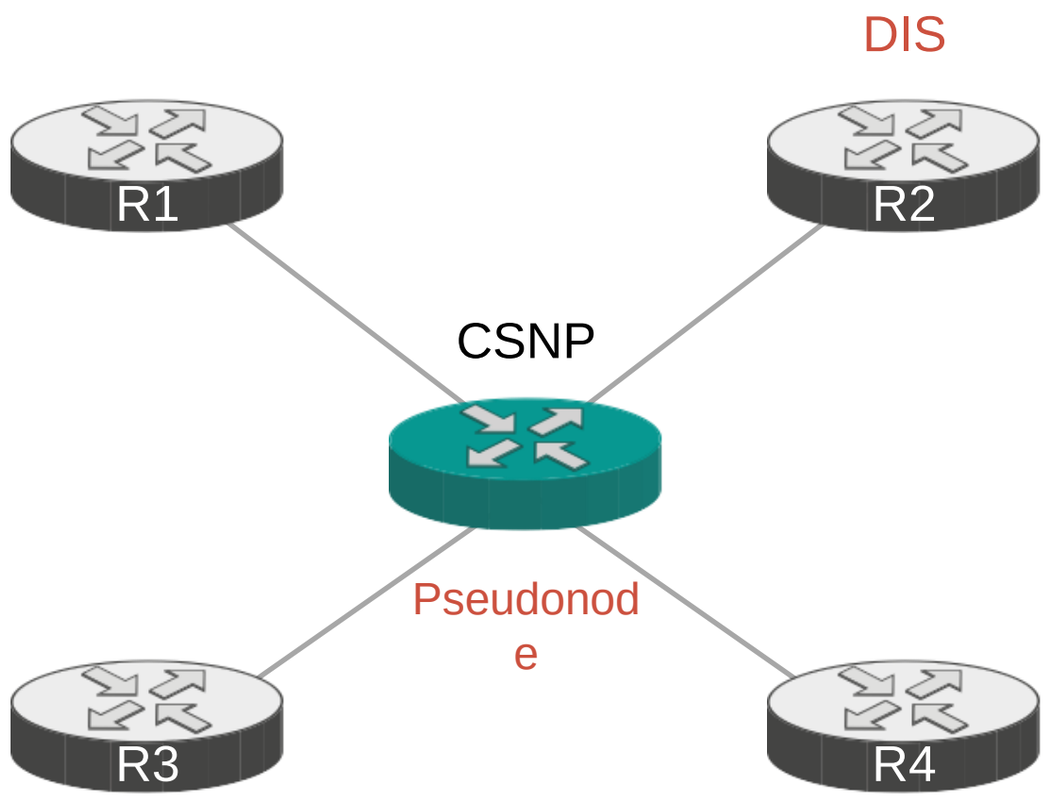
- IS-IS Prefix Suppression similar to OSPF Prefix Suppression, removed transit link IP reachability information from the LSDB and Routing Table.

- Prefix Suppression not only helps for scaling through smaller LSDB and Routing Table but also decreases convergence time, which means provides faster convergence.

IS-IS Flooding in LAN

- IS-IS Multi-Access Segment, similar to OSPF ,elects Designated Router which is called DIS (Designated Intermediate System)
    - Based on the priority (Highest Priority wins, otherwise highest MAC Address) is elected as DIS.
- DIS creates Pseudonode which changes the full mesh LAN topology to the set of p2p links.

IS-IS Flooding in LAN



flooding

IS-IS Flooding in LAN

- DIS advertises CSNP (Complete Sequence Number PDU) messages to the other routers in the multi access segments very often, so if any router doesn’t receive LSP, they can learn it from the DIS.

- In IS-IS each Router (IS) forms an adjacency with each other in broadcast link (Unlike OSPF). In OSPF broadcast link, routers form an adjacency with DR and BDR only.

flo
odi
ng

IS-IS Flooding in LAN

- DIS doesn’t advertise IP reachability information of the attached router, OSPF DR does with Type 2 LSA.

- when new routers come up, it can preempt the existing DIS, that’s why IS-IS DIS election is deterministic, in OSPF, even higher.

flooding

IS-IS Flooding in LAN

- There is no Backup DIS concept within IS-IS, because DIS sends CSNP very often, so routers get the latest info all the time and all routers are neighbor of each other, so they send the LSPs to each other anyway, DIS is like a backup mechanism to ensure receiving LSP.

- DIS concept in Broadcast link provides scalability.

IS-IS Flooding in LAN

- SPF (Dijkstra) runs when topology has to be calculated (SPF Tree).
    - PRC (Partial Route Calculation) runs when IP Routing information has to be calculated.
- If a router (IS) receives an LSP where only IP information has changed, it will run PRC only (Less CPU compare to SPF), thus better compare to OSPF.

flooding

- Scalability can be achieved through IS-IS Multi level design as well which we will discuss next.

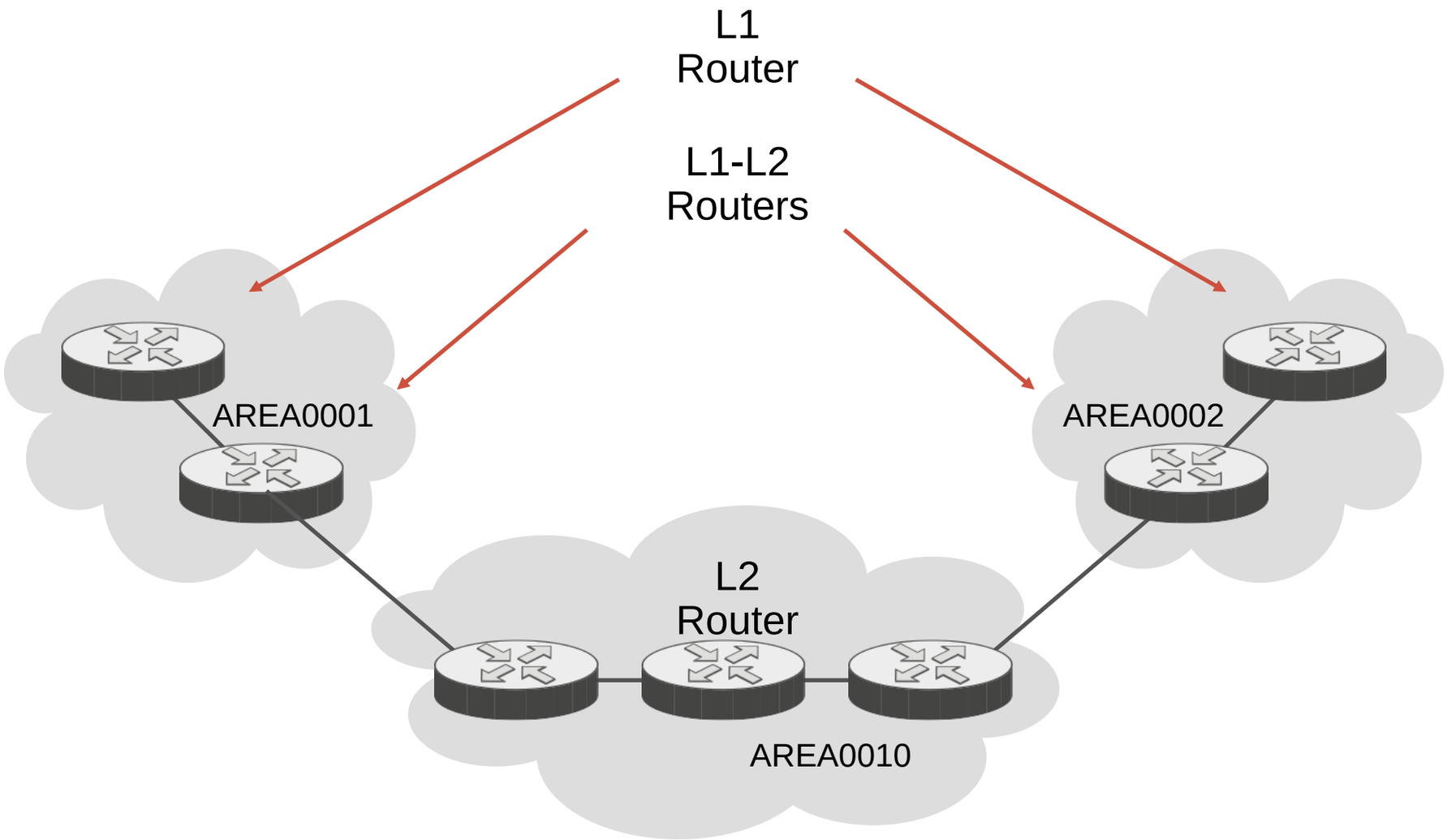
Multi Level IS-IS Design

- IS-IS has two Levels : Level 2 and Level 1
  - Levels are similar to Backbone Area and Non-Backbone areas of OSPF.
- Level 2 IS-IS is similar to OSPF Backbone, Level 1 IS-IS is similar to OSPF Non-Backbone Area.

Multi Level IS-IS Design

- If the Area ID is the same on the 2 routers, they can setup both L1 and L2 adjacency.
    - If Area ID is different they can only setup a L2 IS-IS adjacency.
- There is no backbone area in IS-IS as in the case of OSPF. There is only contiguous Level2 routers. Level 2 domain have to be contiguous.
    - But still for the new learners, IS-IS level 2 domain can be considered similar to OSPF backbone area.

IS-IS Router Types



IS-IS Router Types

- There are three type of routers in IS-IS
  - Level 1 Router:
    1. Can only form adjacencies with Level 1 routers within the same area
    2. LSDB only carries intra area information
    3. Use the closest Level 2 router to exit the area
    4. This may result in suboptimal routing

IS-IS Router Types

- There are three type of routers in IS-IS
    - Level 2 Router:
        1. Can form adjacencies in multiple areas
        2. Exchange information about the whole network

IS-IS Router Types

- There are three type of routers in IS-IS
    - Level 1-2 Router:
        1. These routers keep separate LSDB for each level, 1 for Level 1database, 1 for level 2 databases.
        2. These routers allow L1 routers to reach to other L1 in different area via the L2 topology.

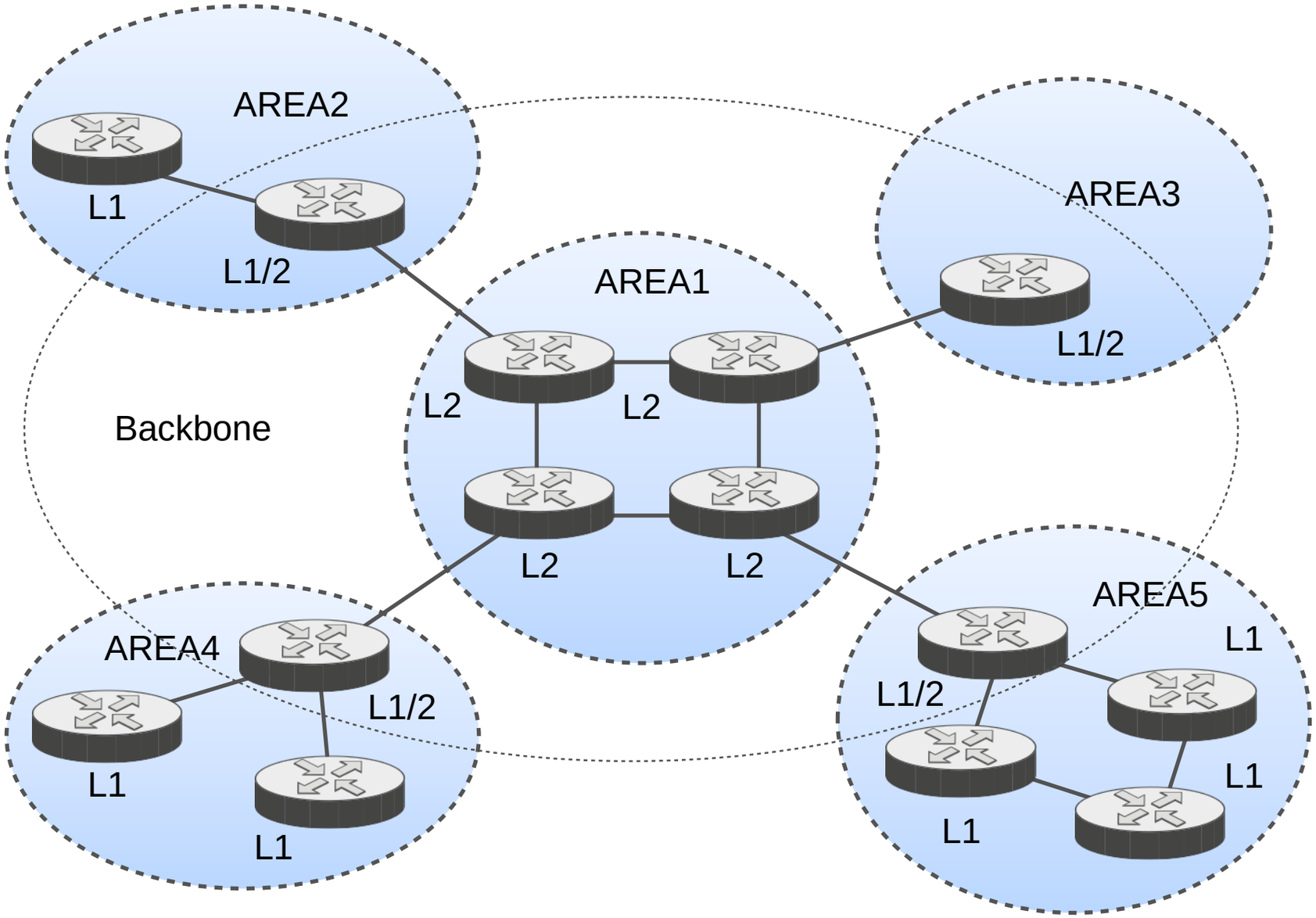
ATT (Attached) Bit and Default Route

- Level 1 routers look at the ATT bit in L1 LSP of L1-L2 routers.
    - And use it as a default route to reach the closest Level 1-2 router in the area. This can create suboptimal routing which we will see later.

ATT (Attached) Bit and Default Route

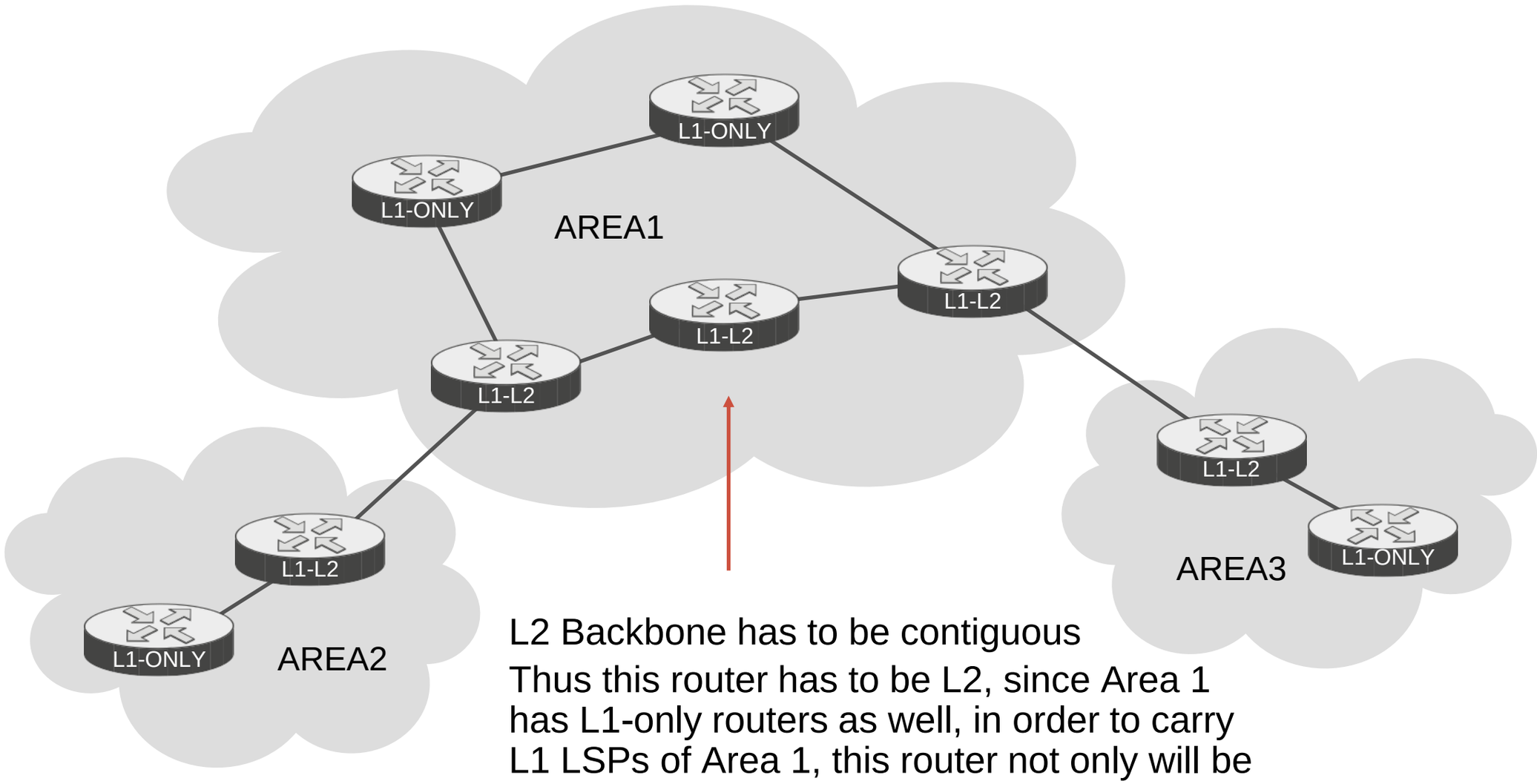
- L1 domain is similar to OSPF Totally NSSA Area since L1 domain doesn’t accept anything other than default route from the Level 2 domain and redistribution is allowed into the L1 domain.

ATT (Attached) Bit and Default Route



Hierarchy Levels

- Level 1, Level 2 and Level 1-2 Routers



L2 Backbone has to be contiguous

Thus this router has to be L2, since Area 1 has L1-only routers as well, in order to carry L1 LSPs of Area 1, this router not only will be L2 but it will be L1-L2.

Overlay Technologies and IS-IS (GRE, MGRE, DMVPN, GETVPN, LISP)

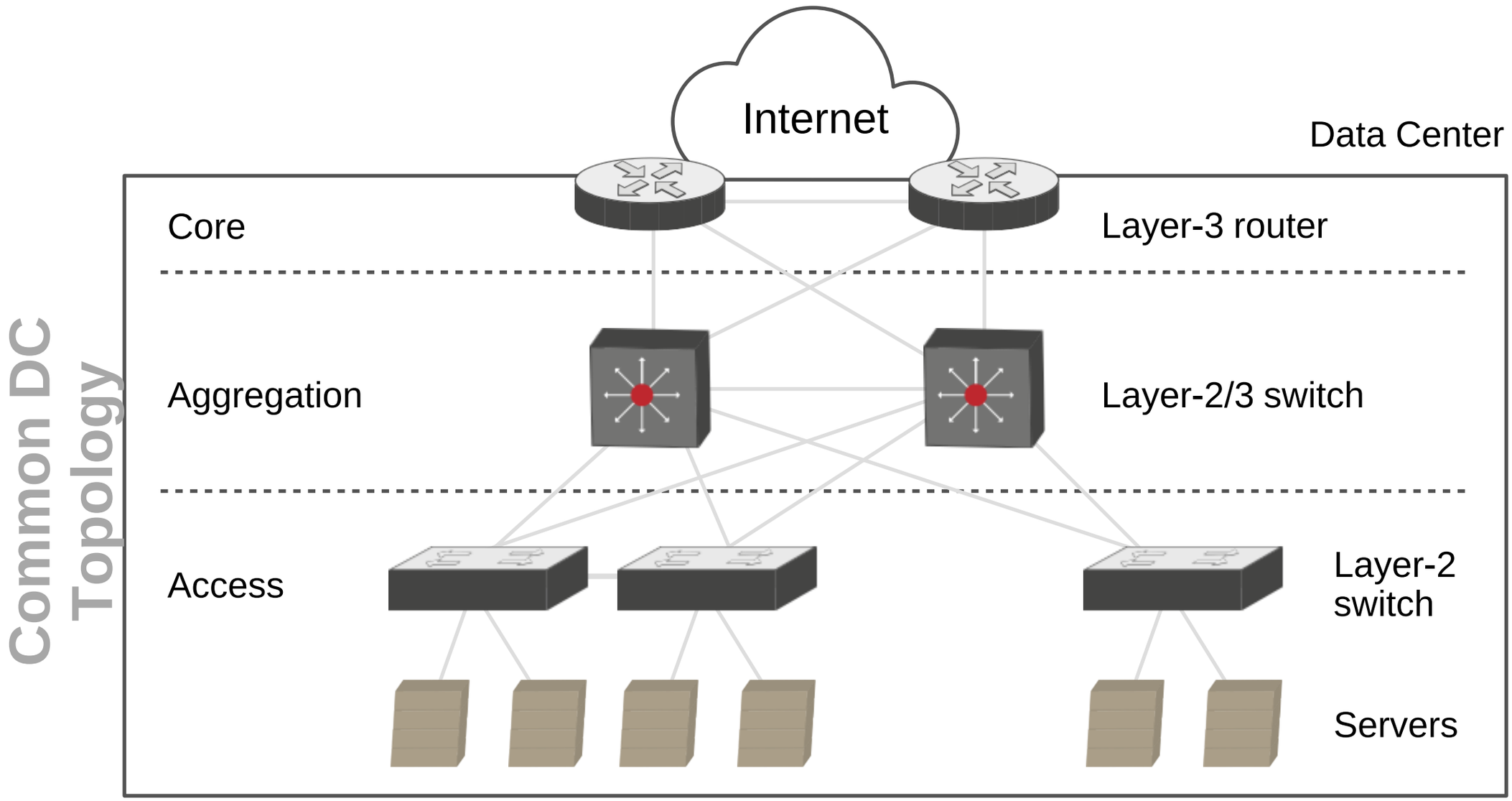
- IS-IS can work on top of only some overlay technologies.
    - GRE, MGRE, DMVPN, GETVPN and LISP can be used to create overlay/VPN in the networks.
- IS-IS can be used for these overlay mechanisms as an underlay infrastructure routing protocol.

Overlay Technologies and IS-IS (GRE, MGRE, DMVPN, GETVPN, LISP)

- IS-IS works over GRE, MGRE.
    - IS-IS doesn’t work over GETVPN and LISP , because both are tunnelless VPN mechanisms, routing protocols can be an underlay for them but not an overlay.
- IS-IS doesn’t work over DMVPN, because IS-IS is layer 2 protocol but DMVPN is an IP Overlay mechanism not layer 2 overlay mechanism.
    - IS-IS with GRE is not scalable for large scale deployment but scaling limitation comes from GRE, it is not the IS-IS problem, MGRE provides scalability with IS-IS even in large scale deployment.

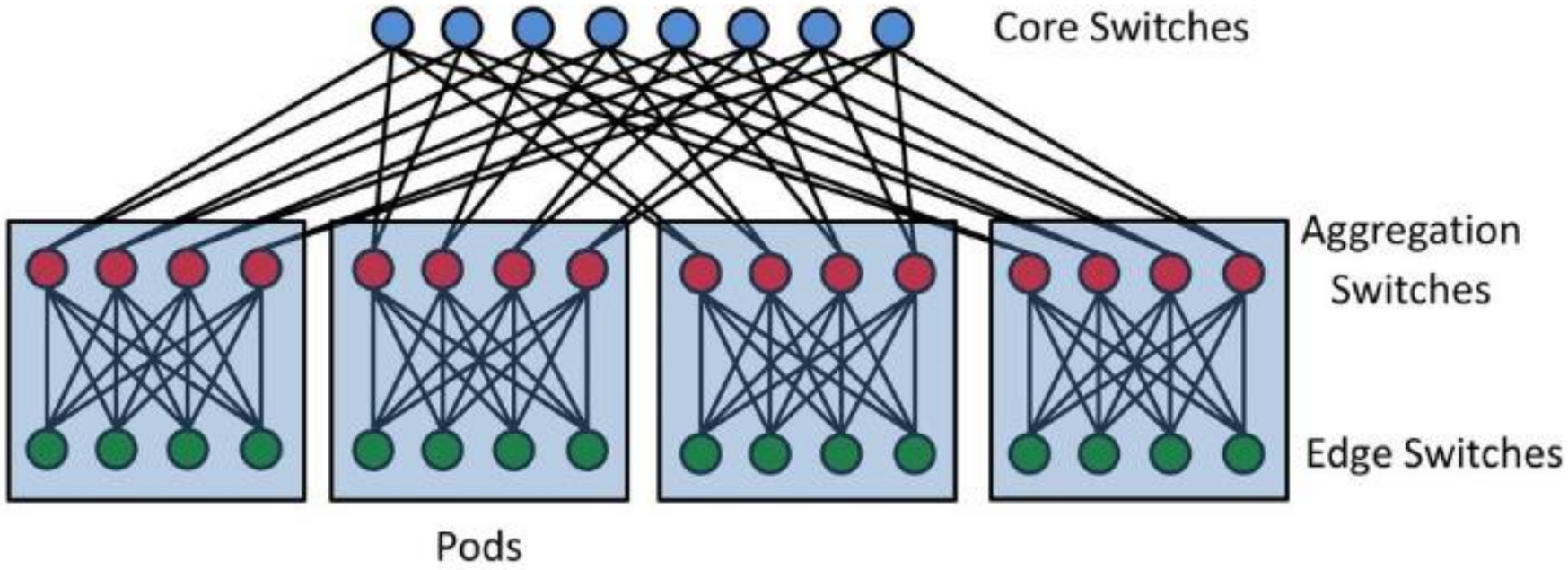
IS-IS in the Datacenter

- IS-IS can be used at the DC edge to advertise DC prefixes to the WAN and Campus network.



- Also IS-IS can be used as a Datacenter Fabric Protocol.
  - Datacenters are very densely connected networks, thus IS-IS flooding creates scalability problem.
- Large scale Datacenters mainly use CLOS (Leaf and Spine) topology, depends on scale, multi stage CLOS topologies are used.

3 stage CLOS topology



- To create DC Fabric, IS-IS is used in some Datacenters as a layer 3 fabric protocol.
    - IS-IS is used for many Layer 2 Fabric protocols as an underlay routing protocol. Some of these L2 Fabric protocols are TRILL, SPB, FabricPath and OTV.
- Large scale Datacenter requirements are discussed currently in the IETF and protocol related ones are:

- The Fabric provides basic connectivity, with possibility to carry one or more overlays.

1. The Fabric MAY provide interconnect facility for other fabrics.
2. The Fabric MUST support non equidistant end-points.
3. The Fabric MUST support Spine and Leaf [CLOS] + isomorphic topologies within its network.
4. The Fabric MAY support non Spine and Leaf topologies
5. The Fabric SHOULD support 250k routes @ 5k fabric nodes with convergence time below 250ms.
6. The Fabric SHOULD support 500k routes @ 7.5k fabric nodes with convergence time below 500ms.
7. The Fabric SHOULD support 1M routes @ 10k fabric nodes with convergence time below 1s.

8. The Fabric routing protocol MUST support load balancing using ECMP, wECMP and UCMP.
   The Fabric routing protocol MUST support and provide facility for topology-specific algorithms that enable correct operations in that specific topology.
9. The Fabric routing protocol SHOULD support route scale and convergence times of a Fabric mentioned above.
   The Fabric routing protocol SHOULD support ECMP as wide as 256 paths.
10. The Fabric routing protocol MUST support various address families that covers IP as well as MPLS forwarding.
11. The Fabric routing protocol MUST support Traffic Engineering paths that are host and/or router based paths.

- The Fabric routing protocol MUST be able to leverage BFD [RFC5880] for neighbor state.

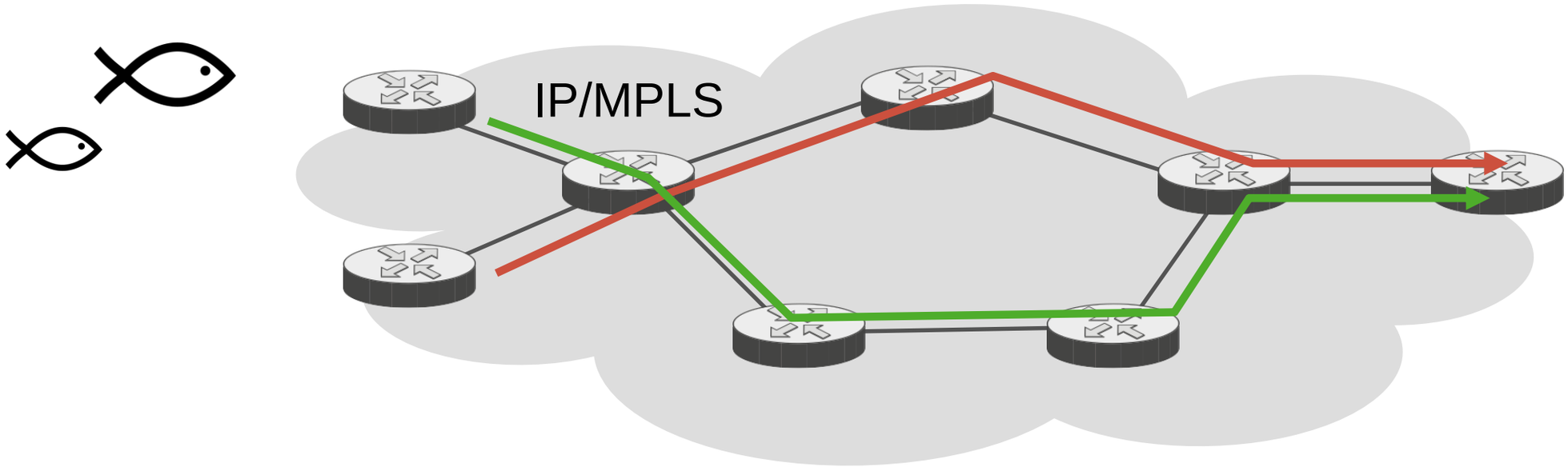
The Fabric routing protocol MUST be able handle commission/decommission of a node as well as any node restart with a minimal data plane impact.

IS-IS in the Service Provider Networks

- IS-IS is very commonly used in the Service Provider networks, in the Middle East and Europe, many Service Providers use OSPF in their network, IS-IS is found in U.S Service Provider networks commonly.
    - IS-IS is used in Core Networks mostly but some providers extend IS-IS to the Aggregation and even to the access domains.
- In Seamless MPLS/Unified MPLS architecture, IS-IS in the access, aggregation and core network usage will be explained in detail

IS-IS and MPLS Traffic Engineering is used together in many SP networks

- IS-IS is used to create shortest path routing but many Service Providers use IS-IS with MPLS Traffic Engineering so they don't just use shortest path between their nodes.



Classical Fish Diagram of MPLS Traffic Engineering.
Without MPLS TE, IGP protocols always chooses shortest path.
Source routing is not possible with IGP protocols.

- IS-IS is used to carry the Service Provider network device prefixes in the SP networks, not the customer routes.
    - Customer routes are carried within BGP

- IS-IS, in theory can be used in Service Provider network as a PE-CE routing protocol if SP is providing MPLS L3 VPN, or mobile operators are using MPLS L3 VPN at their 3G UMTS and 4G LTE sites in Unified/MPLS architecture.

- Most of the Service Providers who run IS-IS in their network as an infrastructure IGP protocol, have flat (single tier) Level 2 IS-IS design.

IS-IS Design Best Practices

- Unless there is a valid reason, don’t deploy Multi Level IS-IS, keep the design simple, it provides better convergence, less configuration on the L1L nodes and optimal traffic flow.

  - If there are low end devices which have low CPU and low memory, they can be placed in a L1 domain in Multi Level Large Scale IS-IS design.

IS-IS Design Best Practices

- Don’t enable IS-IS on the customer facing ports, for MPLS L3 VPN PE-CE protocol, enable prefix limit, authentication and control plane policing.

best practices

IS-IS Design Best Practices

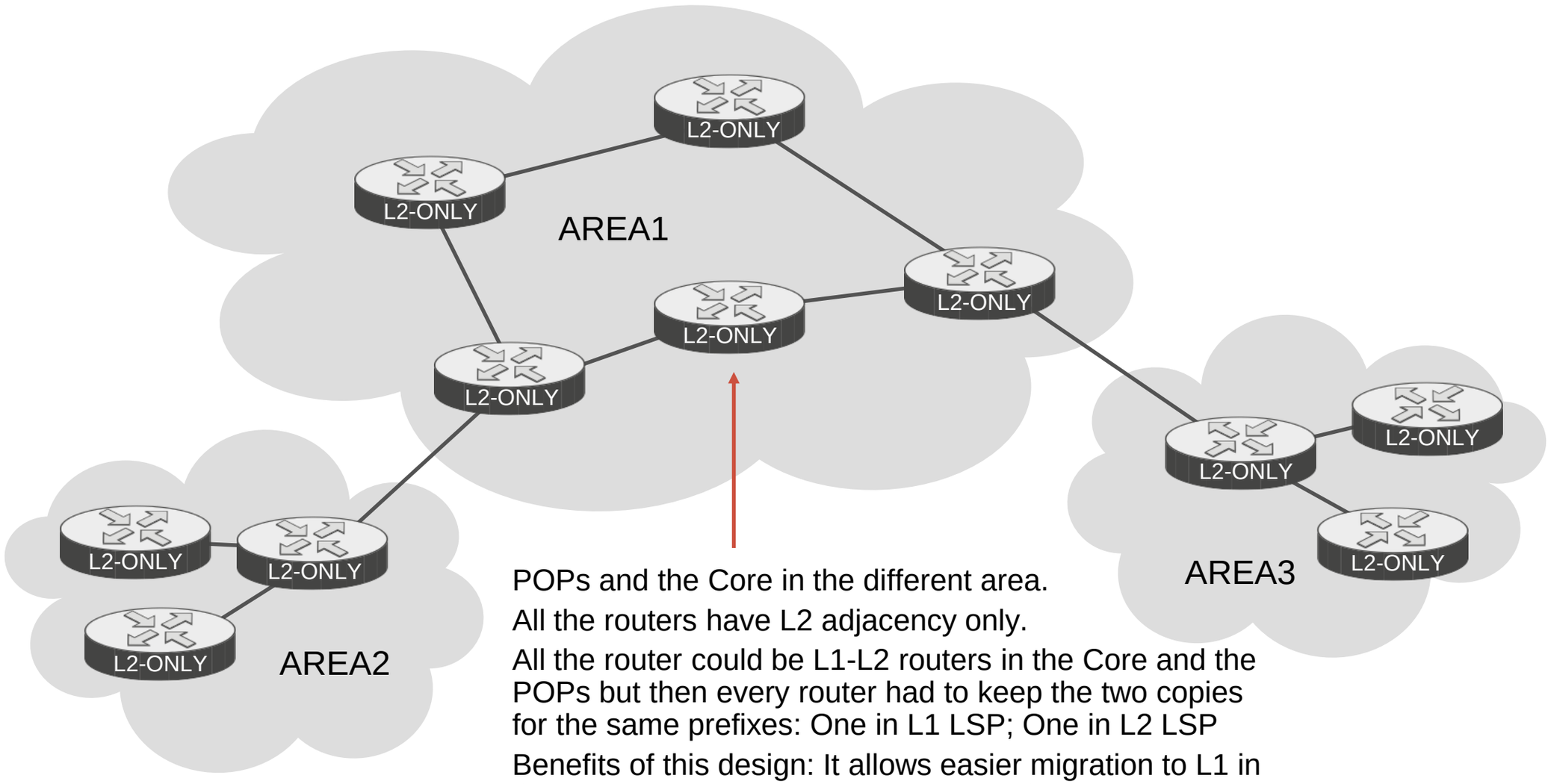
- Use IS-IS Prefix-suppression feature to remove infrastructure/transit links from the TLV, it provides scalability in large scale IS-IS design.

  - Always start green field network design with Level 2 IS-IS, it will provide easier migration when multi level IS-IS design is necessary.

- If you start green field network design with L1-L2 then all the routers have to keep two databases for every prefix. So it is resource intensive without additional benefit

Level 2 to IS-IS to Multi-Level IS-IS Design requires only edge nodes to be touched

- Area Design L2 in the POP and Core. POPs are in different Areas



POPs and the Core in the different area.

All the routers have L2 adjacency only.

All the router could be L1-L2 routers in the Core and the POPs but then every router had to keep the two copies for the same prefixes: One in L1 LSP; One in L2 LSP

Benefits of this design: It allows easier migration to L1 in the POP and L2 in the Core Multi Level IS-IS design if scalability is required in the future. There is no ATT bit since there is no L1 router in the design.

- When there is DIS in the IS-IS domain, make sure you don’t have performance problem with it.

- Summarization removes reachability information and it can be done on both L1L2 router or L1 internal router in Multi Level IS-IS design.

  - By default, only default router is sent into L1 domain in Multi Level IS-IS design, this breaks the MPLS LSP, since LDP cannot have aggregated FEC unless the RFC 5283 – LDP Extension can be used.

- Don’t redistribute full Internet routing table to IS-IS.
    - IS-IS in the large scale datacenter has flooding issue, mesh-group can be used to remove the topology information.
- If you need to deploy Multi Level IS-IS Design, know that it can create suboptimal routing in many topologies, we will see a case study for this issue, in this lesson.
    - Sub optimal routing is not always bad, just know the application requirements. Some application can tolerate and you can have low end devices in L1 areas.You can place edge and core in Level 1.

- Don’t deploy more than two L1L2 Routers (IS) for redundancy, two is enough.
  - L1-L2 Routers slow down the convergence.
- Don’t carry customer prefixes with the infrastructure IS-IS in Service Provider networks, customer prefixes should be carried in BGP.

- IS-IS Fast convergence might bring unstability to network, make sure timers are tuned accordingly for the fast convergence.

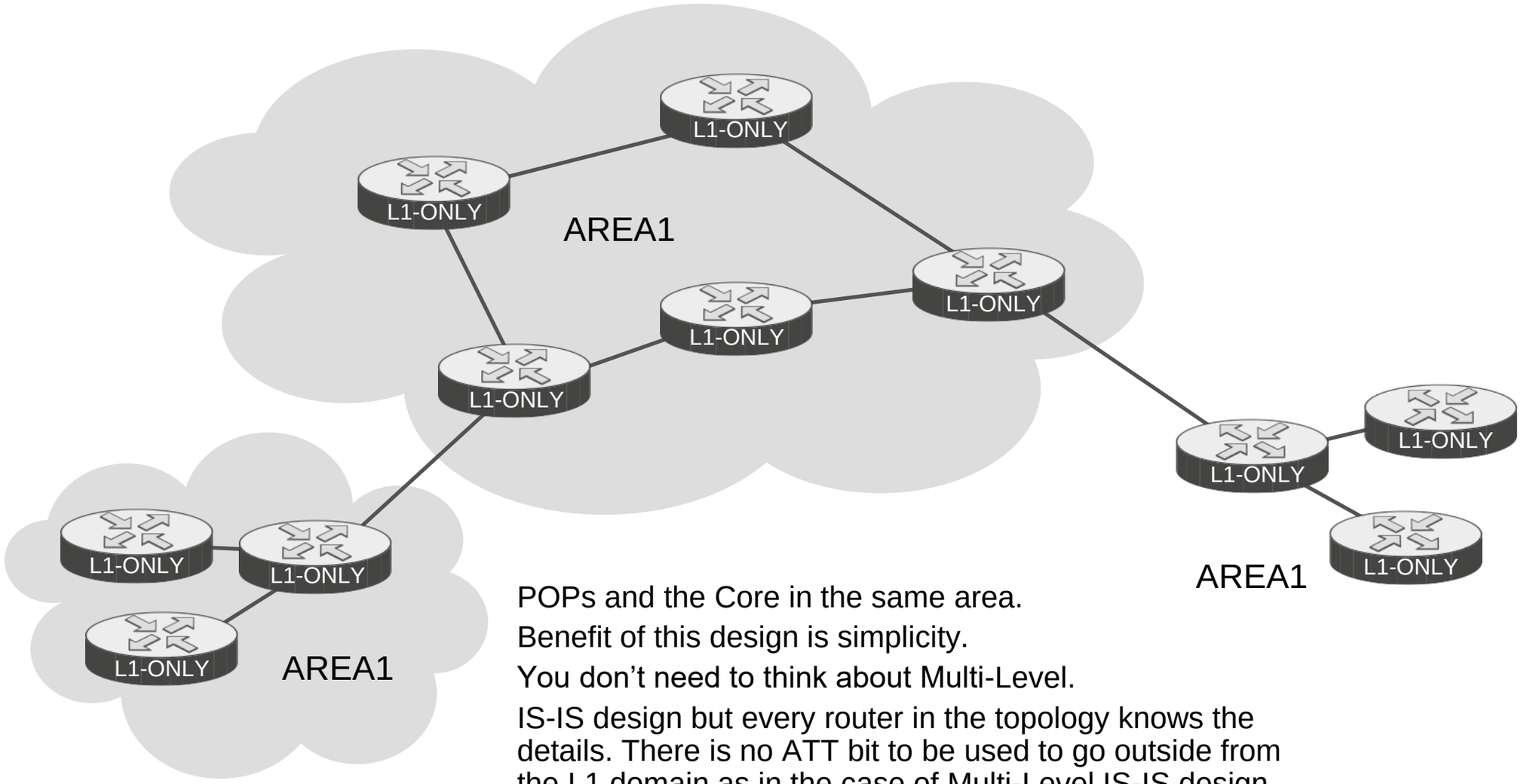
  - IS-IS Fast reroute with LFA may not cover every topology, especially ring will not be protected, you may need to deploy Remote LFA or MPLS TE FRR for that, if topology is partial/full mesh, OSPF and LFA is enough to provide FRR for links or prefixes.

Design Examples with IS-IS Areas and Levels

- You can place edge and core in Level 1 or Level 2 with same area.
    - You can place edge and core in level 2 but in different areas which can make future multi-level migration easier.
- Or You can place the POPs in Level 1 areas and core in Level2. This can create a suboptimal routing but provides excellent scalability.

L1 in the POP and Core

- Area Design L1 in the POP and Core.



POPs and the Core in the same area.

Benefit of this design is simplicity.

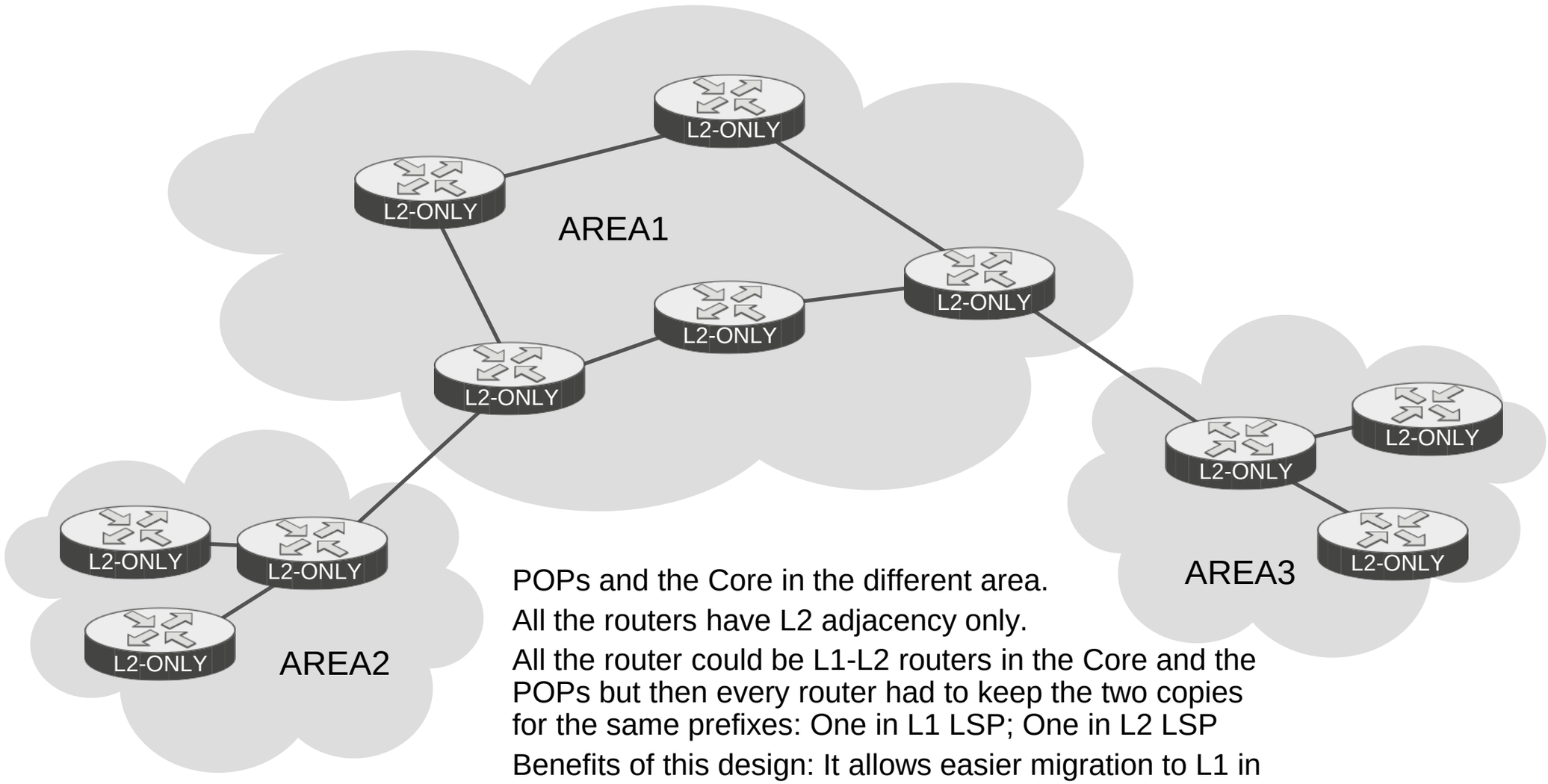
You don’t need to think about Multi-Level.

IS-IS design but every router in the topology knows the details. There is no ATT bit to be used to go outside from the L1 domain as in the case of Multi-Level IS-IS design.

If you design L1 everywhere, migration to a Multi-Level IS-IS design is harder compare to L2 everywhere.

L2 in the POP and Core

- Area Design L2 in the POP and Core. POPs are in different Areas.



POPs and the Core in the different area.

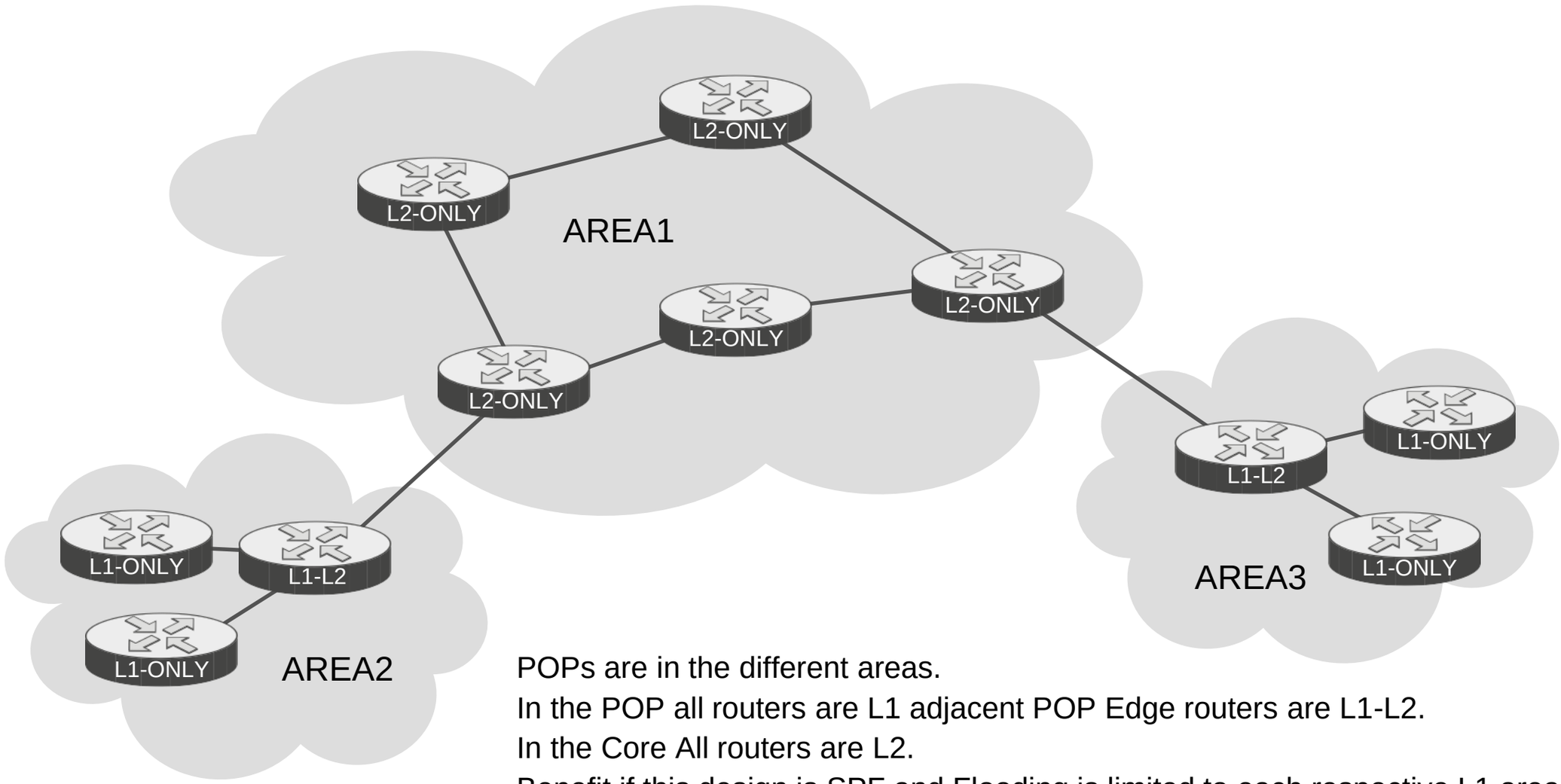
All the routers have L2 adjacency only.

All the router could be L1-L2 routers in the Core and the POPs but then every router had to keep the two copies for the same prefixes: One in L1 LSP; One in L2 LSP

Benefits of this design: It allows easier migration to L1 in the POP and L2 in the Core Multi Level IS-IS design if scalability is required in the future. There is no ATT bit since there is no L1 router in the design.

L1 at the Edge, L2 in the Core with Different Areas

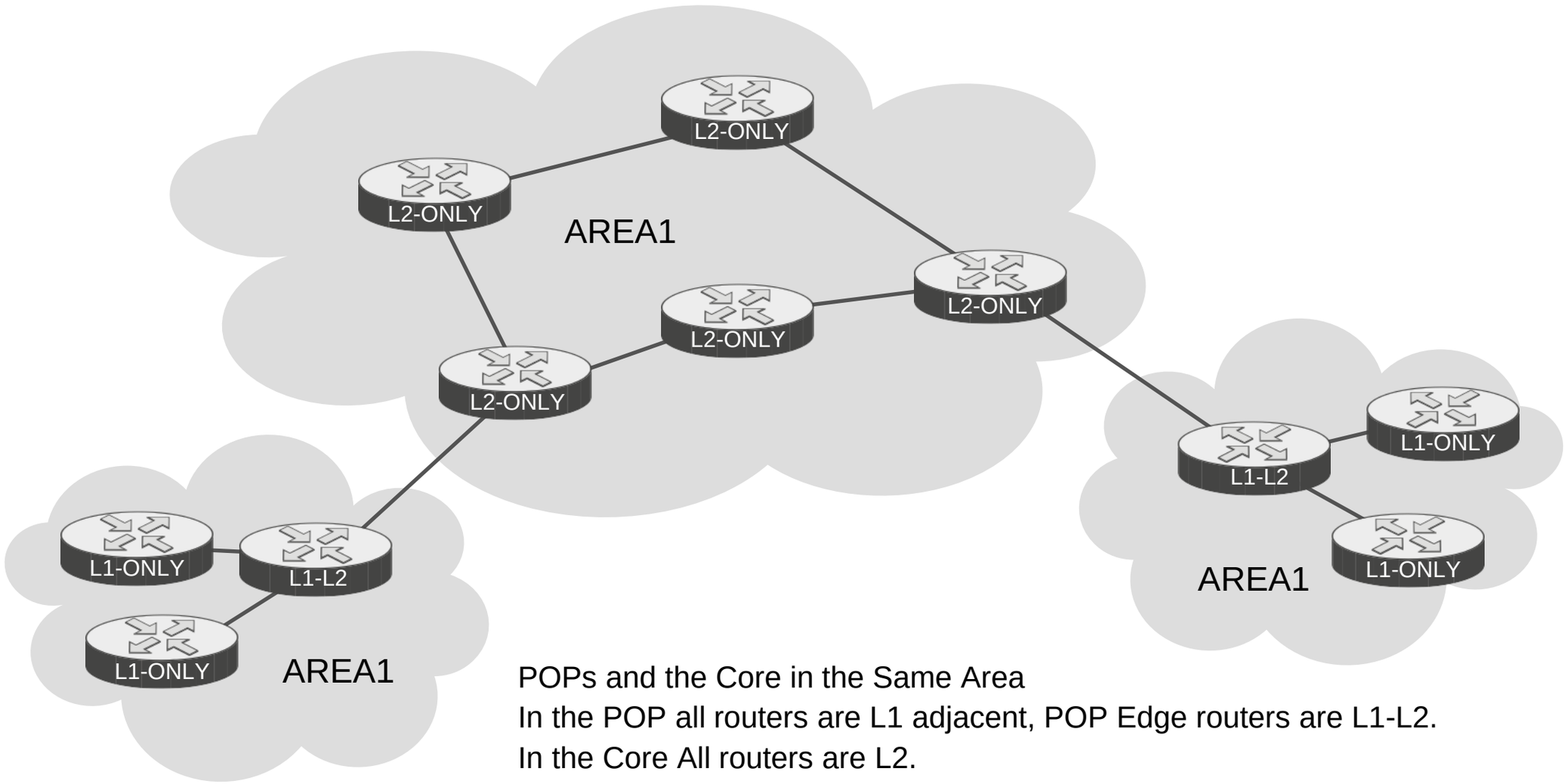
- L1 Only POPs L2 Only Core.



POPs are in the different areas.
In the POP all routers are L1 adjacent POP Edge routers are L1-L2.
In the Core All routers are L2.
Benefit if this design is SPF and Flooding is limited to each respective L1 areas.
L1 routers only know the ATT bit for outside of their areas.
If scalability is the business requirement, this is the ultimate design.

L1 at the Edge, L2 in the Core with Same Areas

- L1 Only POPs L2 Only Core.



POPs and the Core in the Same Area

In the POP all routers are L1 adjacent, POP Edge routers are L1-L2.

In the Core All routers are L2.

This design doesn't work as L1 Only Routers cannot receive a default route through ATT bit from L1 –L2 routers, as L1-L2 routers are not in the different areas, thus they cannot set ATT bit in their L1 LSP.

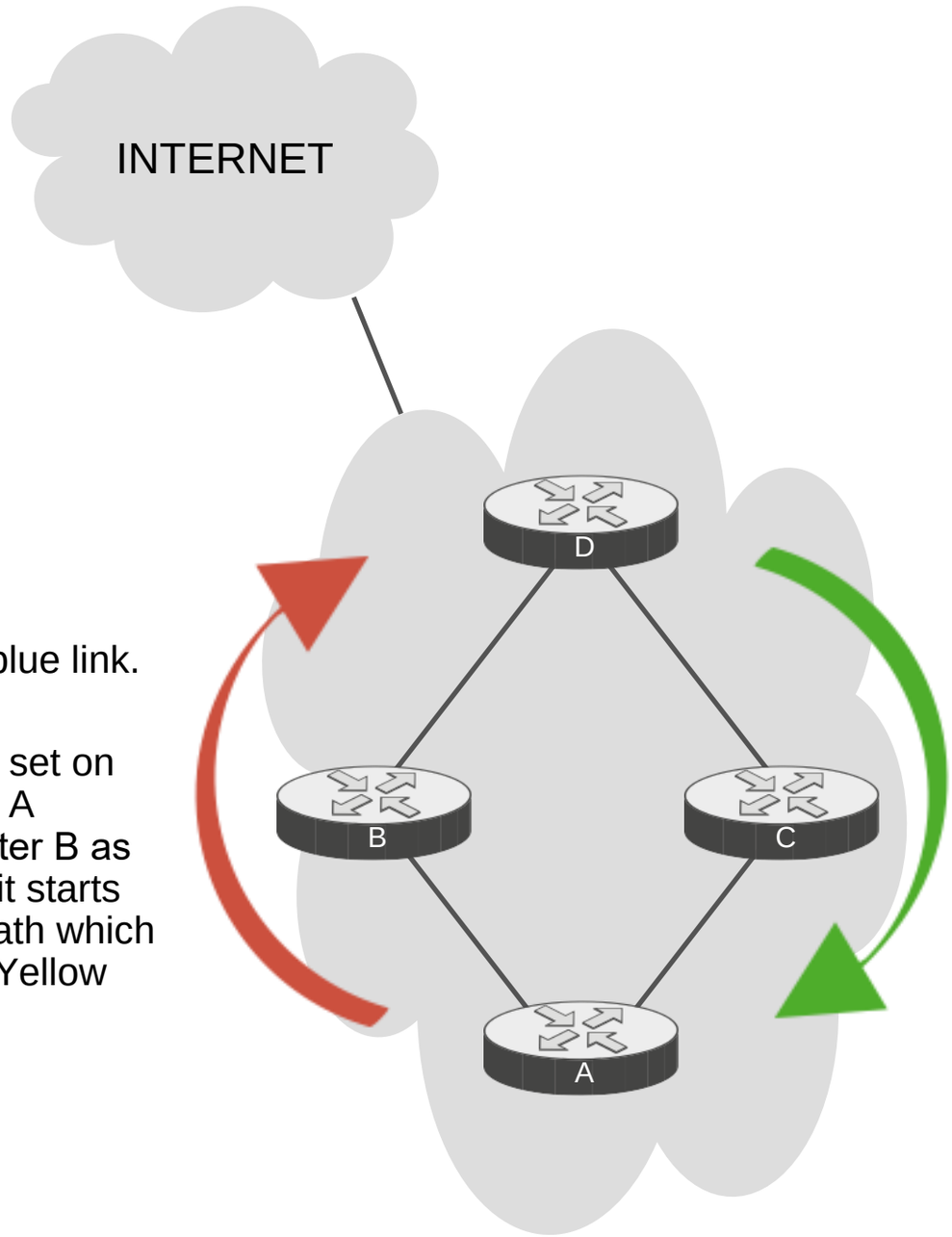
IS-IS - BGP Interaction and Overload Bit

- Network design, especially large scale network design is all about managing interaction and understand the tradeoffs. IS-IS as an IGP provides an underlay infrastructure for BGP, MPLS overlays.

- Below is the interaction of IS-IS with BGP. Overload bit is set to signal the other routers, so the router is not used as transit router.

IS-IS - BGP Interaction

- Interaction with BGP



Primary path is blue link.

If Overload bit is set on Router B Router A doesn't use Router B as transit node but it starts to use backup path which is depicted with Yellow link.

Overload bit is used to avoid black holing during transient network events.

Once overload bit is set, router is not used as a transit node by the other.

Same behavior in OSPF is achieved with OSPF max-metric router LSA feature.

With that feature OSPF node flood its link with maximum metric so it is not used as a transit node.

IS-IS and MPLS Interaction

- When IS-IS multi level design is used in MPLS enabled networks, MPLS LSP is broken since LDP does not have capability of Aggregated FECs.

- In other term, If LDP is used for label distribution, all the routers in the IS-IS network have to have /32 loopback address (If Loopback interfaces are configured as /32, if loopback interfaces are configured as /24 , LDP can assign a label for the /24 loopback address) of the PE devices in their routing table. Otherwise label is not assigned.

Fortunately there are at least three ways to fix it.

1. You can leak the loopback of PEs into L1 domain.

2. RFC 5283 (LDP extensions) allows aggregated FECs. So you don’t need /32 in the routing table to assign a label for the MPLS FEC (Forwarding Equivalence Class).

Fortunately there are at least three ways to fix it.

3. RFC 3107, BGP+Label. IPv4 Label is assigned by the BGP. It is not a label for the VPN prefixes, it is a label for the IP prefixes.

Seamless MPLS uses this concept. Inter domain Hierarchical MPLS LSP is created and all the PE loopbacks are carried through Multi Protocol BGP. Seamless MPLS will be explained in the MPLS chapter.

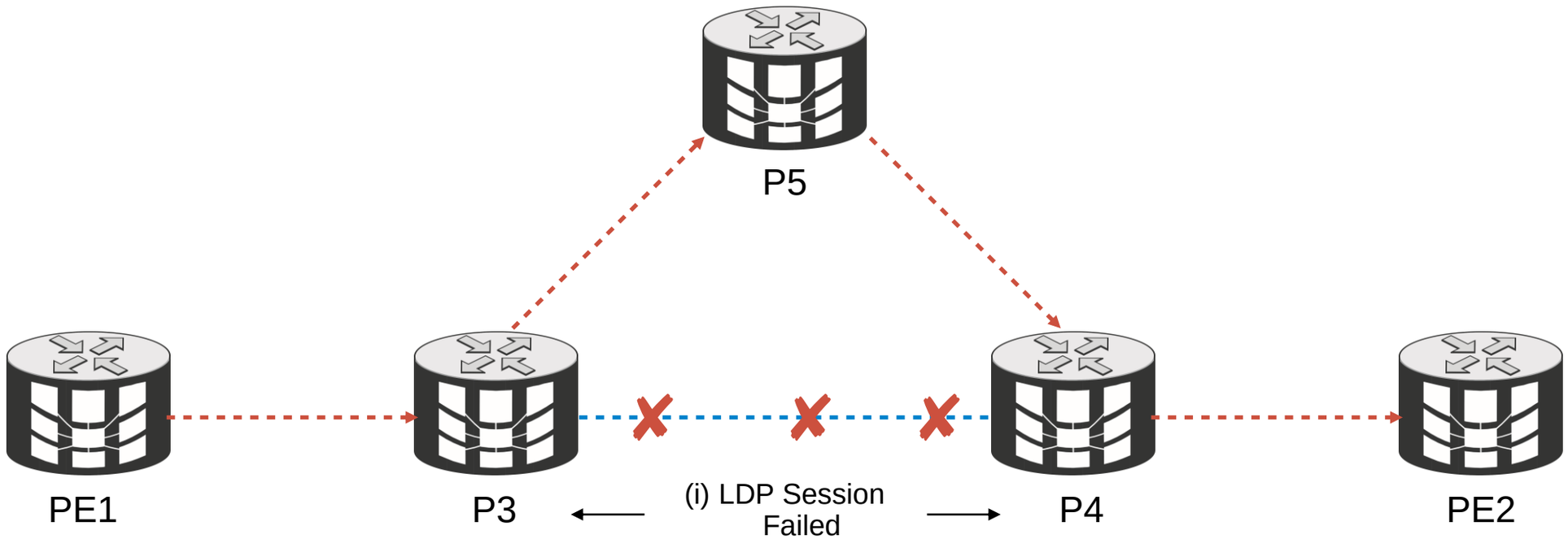
- Previous example require IGP and LDP control plane to have same prefixes in RIB and LFIB respectively.

- IGP and LDP dataplane should be following each other as well, similar to IGP, BGP synchronization, we have IGP-LDP Synchronization which I will explain next.

Fast Service Restoration with IS-IS and LDP

IGP - LDP Sync

- When there is link up event or LDP session down event, IGP advertises max/metric and indicates that traffic should follow the alternate links.
  - This solution is called, cost-out procedure as well and very similar to IGP-BGP synchronization.
- Problem happens because LDP converge after IGP when link comes up, this creates black holing.



(ii)IGP Assign Max metric for this link

(iii)R3 IGP is using link to R5 to reach R4

(iv)LDP follows IGP, and LSP diverted to R3-R5-R4 path

Fast Service Restoration with IS-IS and LDP

- With IGP-LDP Sync feature, LDP session goes down, when it comes up, before IGP advertise regular metric for the newly established link, LDP should converge. When LDP converge, it signals IGP to advertise the link metric as regular.

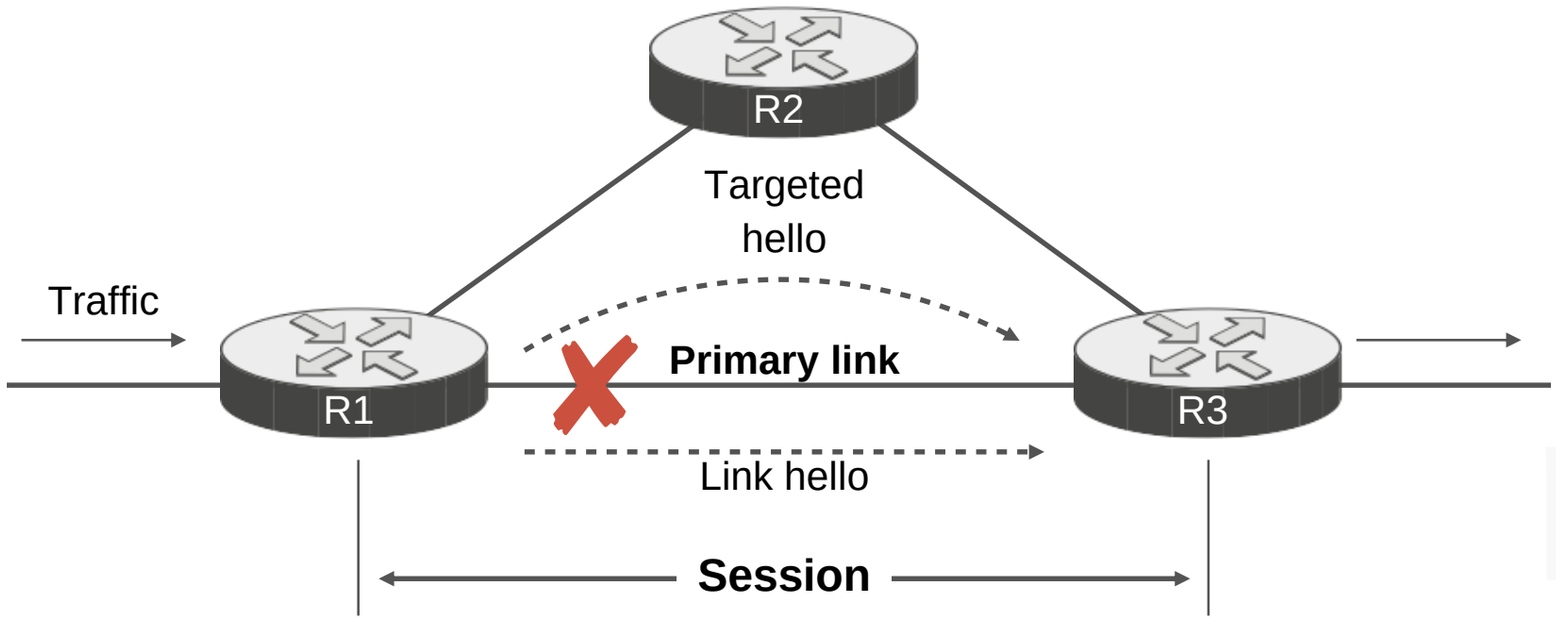
Fast Service Restoration with IS-IS and LDP

LDP Session Protection

- When a link comes up, IP converges earlier and much faster than MPLS LDP and may result in MPLS traffic loss until MPLS convergence. If a link flaps, the LDP session will also flap due to loss of link discovery.
    - LDP initiates backup targeted hellos automatically for neighbors for which primary link adjacencies already exist. These backup targeted hellos maintain LDP sessions when primary link adjacencies go down.
- We have always up LDP session with LDP Session Protection feature.

Fast Service Restoration with IS-IS and LDP

- The primary link adjacency between R1 and R3 is directly connected link and the backup; targeted adjacency is maintained between R1 and R3.



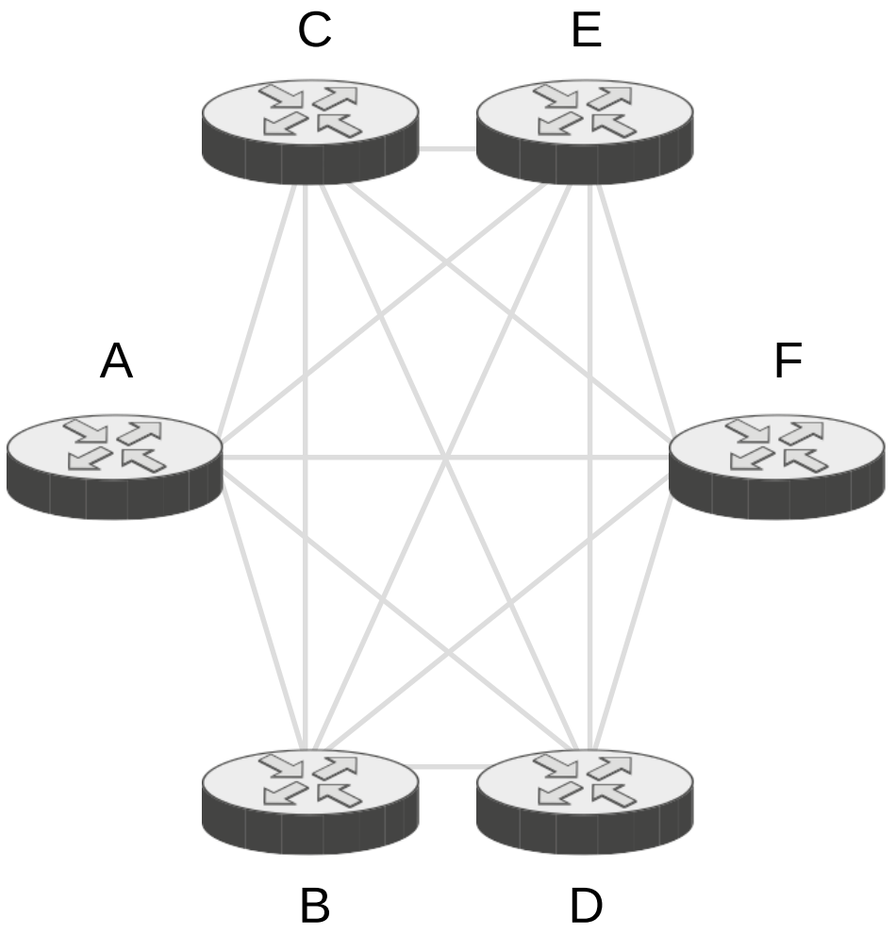
Fast Service Restoration with IS-IS and LDP

- With LDP Session Protection, if the direct link fails, LDP link adjacency is destroyed, but the session is kept up and running using targeted hello adjacency (through R2). When the direct link comes back up, there is no change in the LDP session state and LDP can converge quickly and begin forwarding MPLS traffic.

Fast Service Restoration with IS-IS and LDP

- Both IGP - LDP Synchronization and LDP Session Protection features provides fast service restoration for MPLS Enabled services in case LDP problem or link up events.
  - With IGP-LDP Sync case, LDP gets the help of IGP to avoid traffic loss, with Session Protection, there is no need to sync IGP and LDP but Label bindings are kept in LIB even if link goes down.

IS-IS in a Full Mesh Topology



Full Mesh Topology

- Mesh group is one mechanism that reduces the amount of flooding in a full-mesh topology. With mesh group, we can designate couple routers in the topology to flood.

- Those routers will be responsible for flooding event. In this topology, Router A and B can be assigned as the flooders. Important consideration is to select powerful devices as flooders in the mesh group since their duty is to flood the LSPs in all the networks.

IS-IS in a Full Mesh Topology

- This is the formula: for N routers, there are (N) (N-1)/2 links; If there is only 2 routers in the topology, the total number of link between them is 1; if 3 routers, there are 3 links; if 4 routers, there are 6 links; and if 5 routers, there are 10 links.

mesh topology

- Because in the above topology there are 6 routers, there are 15 links. Even if one loopback is added to any one of these routers, that loopback information is flooded in all the routers over all the links.

IS-IS Advantages and the Disadvantages

disad vanta ges

- IS-IS useS TLV encoding, it is an extendable protocol, don’t require new version of protocol for IPv6 for example.

- OSPF has 11 Type of LSA, compare to 2 Levels of IS-IS thus IS-IS is considered as less complex.

- Each OSPF LSA has a separate header, IS-IS TLVs share common LSP header, thus IS-IS is considered as more scalable.

adva ntag es

IS-IS Advantages and the Disadvantages

disad vanta ges

- IS-IS doesn’t require an IP Address for neighborship, remote attack to the IS-IS is hard if not impossible, thus IS-IS is considered as more secure compare to OSPF and EIGRP.

- IS-IS provides MPLS TE supports, similar to OSPF, but distance vector protocols don’t.

adva ntag es

- IS-IS is a good protocols for those who look for SP level and standard base protocol.

IS-IS Advantages and the Disadvantages

disad vanta ges

- IS-IS doesn’t run on top of IP, so it may not be suitable overlay protocol for many type of VPNs, for example DMVPN.

adva ntag es

- IS-IS is used as an underlay routing protocol for many L2 Fabric protocols in datacenter networks, because IS-IS doesn’t require IP address for it’s operation

IS-IS CASE STUDIES

OSPF to IS-IS Migration

- Fastnet is a fictitious service provider which has some security problems with their internal IGP routing recently. They want to deploy IPv6 on their network very soon but they don’t want to run two different routing protocol one for existing IPv4 and new IPv6 architecture.

- They currently have OSPF routing protocol deployed in their network.

OSPF to IS-IS Migration

- Fastnet knows that adding a new feature in IS-IS by the vendors are faster than OSPF in general. Also thanks to the TLV structure of IS-IS, when they need additional feature, IS-IS can easily be extendable.

mig
rati
on

- Also since the majority of the service providers historically use IS-IS for their core IGP routing protocol, Fastnet decided to migrate their IGP from OSPF to IS- IS.

OSPF to IS-IS Migration

- Please provide a migration plan for Fastnet for smooth transition. Fastnet will plan all their activity during a maintenance window.

mig
rati
on

- Fastnet has been using flat OSPF design but they want flexible IS-IS design which allows Fastnet to migrate multi-level IS-IS in the future.

High Level Migration Plan from OSPF to IS-IS for FASTNET

- Below are the migration steps for the migration. Ship in the night approach will be used. Both routing protocols will be running on the network at the same time during migration.

mig
rati
on

1. Verify OSPF configuration and operation
2. Deploy IS-IS over entire network
3. Set OSPF admin distance to be higher than IS-IS
4. Check for leftovers in OSPF
5. Remove OSPF from entire network
6. Verify IGP and related (BGP) operation

Detail OSPF to IS-IS Migration Steps

1. Verify OSPF configuration and operation Check if there is any routing table instabilities. Next hop values for the BGP are valid and reachable Check OSPF routing table, record the number of prefixes

Detail OSPF to IS-IS Migration Steps

2. Deploy IS-IS over entire network

- Use wide metrics for IS-IS, this will allow IPv6, MPLS Traffic Engineering and new extensions.

- Deploy L2 only IS-IS since they want flexibility, L2 only reduces the resource requirement on the routers and allows easier migration to multi-level IS-IS.

Detail OSPF to IS-IS Migration Steps

2. Deploy IS-IS over entire network

- Deploy both IPv4 and IPv6.
- Deploy IS-IS passive interface at the edge links, these links should be carried in IBGP. Also prefix suppression can be used to carry infrastructure links in IBGP but these are not a requirement of Fastnet.
- Make sure the IS-IS LSDB is consistent with OSPF routing table

Detail OSPF to IS-IS Migration Steps

3. Set OSPF admin distance to be higher than IS-IS Increase the AD of OSPF across entire network

Detail OSPF to IS-IS Migration Steps

4. Set OSPF admin distance to be higher than IS-IS Increase the AD of OSPF across entire network.

   - In this step all the prefixes in the routing table should be learned by IS-IS. If there is any OSPF prefixes, we should find out why they are there. You can compare the ‘show ip OSPF neighbor’ with ‘show is-is neighbor’ so should be the same number of neighbors for both.

- If not the same number of neighbors, fix the problem.

Detail OSPF to IS-IS Migration Steps

5. Remove OSPF from entire network.

- All the OSPF processes can be removed.

- If there is interface specific configuration such as Traffic Engineering configuration (metric, cost), authentication should removed as well.

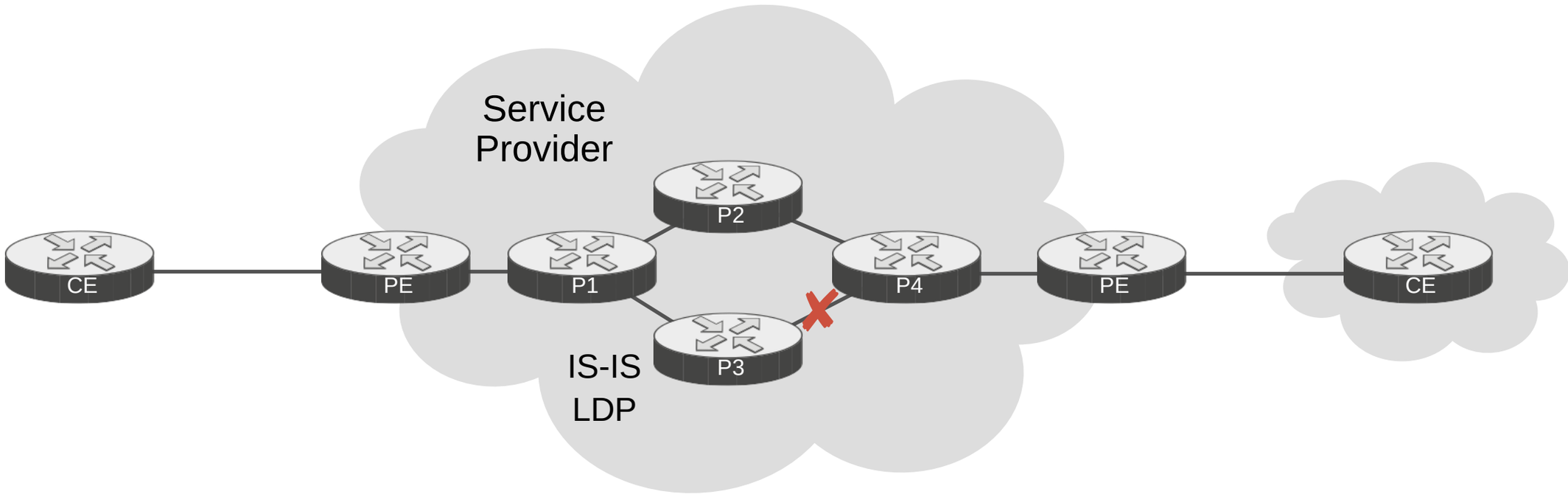
Detail OSPF to IS-IS Migration Steps

6. Verify IGP and Related operation.

- Entire network should be functioning over IS-IS Verify IBGP sessions.
- Verify MPLS sessions.
- Verify customer and edge link prefixes .

Enjoy the party!

IS-IS and MPLS Interaction - Avoiding Black-holing



- In the topology above, IS-IS is running in the network of the service provider. For the transport label distribution or topmost label/tunnel label, LDP is used even though RSVP or segment routing could be used as well.

QUESTIONS

1. What happens if P3-P4 link fails?
2. Do you need to know the level of IS-IS network to provide a solution?
3. What would be your design recommendation to ensure high availability on this network?

ANSWER 1

- If any link fails in the MPLS networks, IGP should not converge on the failed link before getting green light from the LDP.

- Also, if P3-P4 link fails in the topology shown above, P1-P2-P4 link is used. If the link comes up and if IGP converges before LDP converge, P3 cannot create a label for the prefixes; it sends the regular IP packet to P4. In fact, P4 drops the packets because it cannot recognize the CE (Customer).

ANSWER 2

- It doesn’t matter which IS-IS Level (L1 or L2) is used to provide a solution for this problem.
  - Here the question is seeking if you know the solution already.
- This type of questions will be asked in the CCDE Practical exam and the task domain will be Analyzing the design.

ANSWER 3

- If IGP-LDP synchronization feature is enabled, P3 and P4 signals their neighbor not to use P3-P4 link unless LDP converges. IGP signals the other nodes in the routing domain to use alternate links by costing-out the failed link.

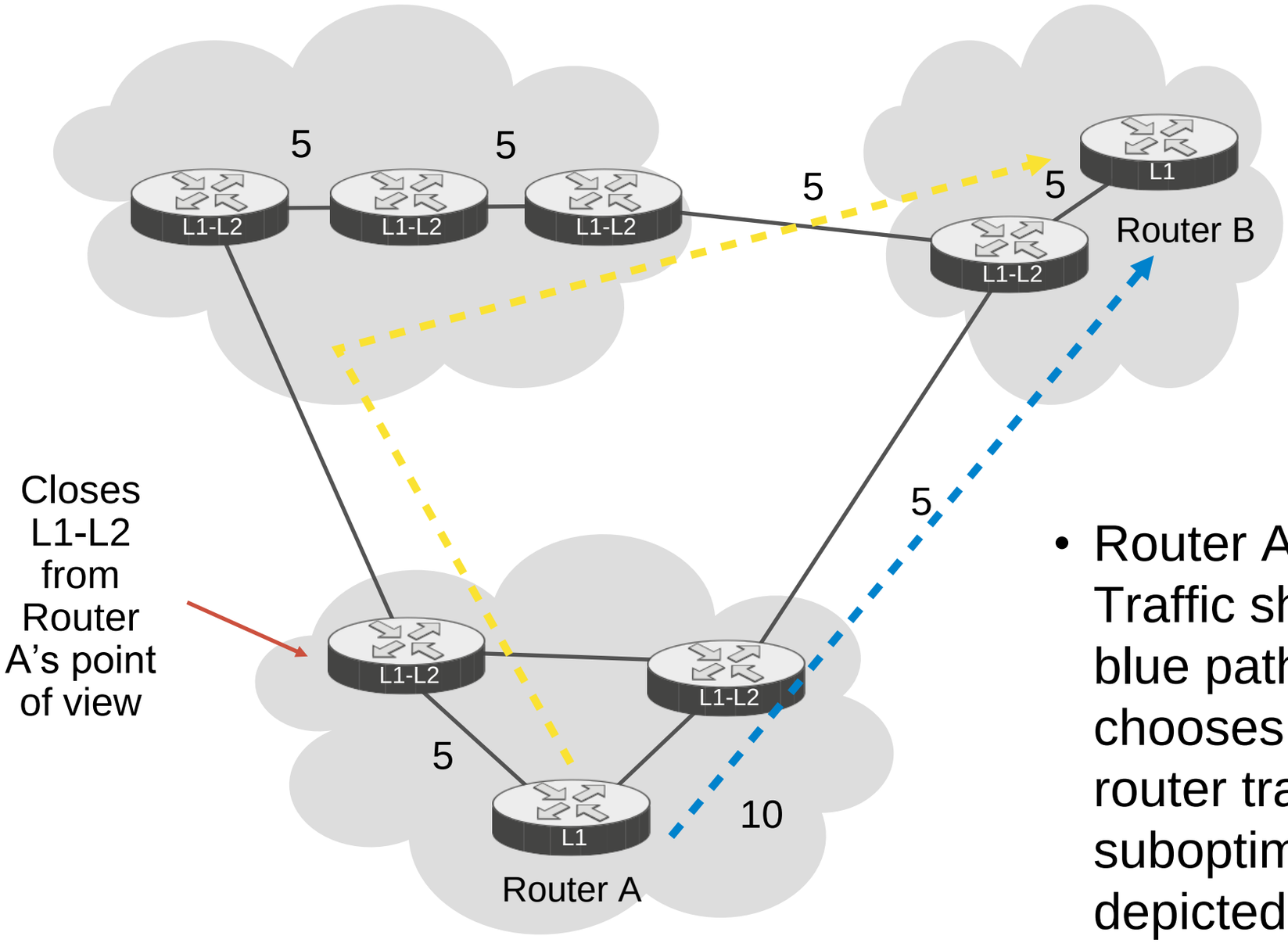
- Also, LDP session protection would help to avoid black holing in this question and would provide faster service restoration.

- Protocol interaction is for the optimal routing design. If overlay protocols doesn't follows the underlay protocols or physical topology, most of the time, sub optimal routing, blackhole, routing or forwarding loop occurs.
    - In order to avoid it, synchronization should be enabled. So far in this class you have seen, Spanning Tree- FHRP, IGP – BGP, IGP-MPLS interactions with the case studies.
- More interactions for different technologies with the case studies will be provided in the later sessions.

Suboptimal Routing with Multi Level IS-IS Design

- Service Provider decided to migrate from Flat IS-IS design to Multi Level IS-IS design.
    - They have for the redundancy two L1L2 routers in each POP location.
- They discovered that Multicast application performance between their users in different locations have performance issue. They checked their new IS-IS design and found some issues. What can be possible problems?

Suboptimal Routing with Multi Level IS-IS Design



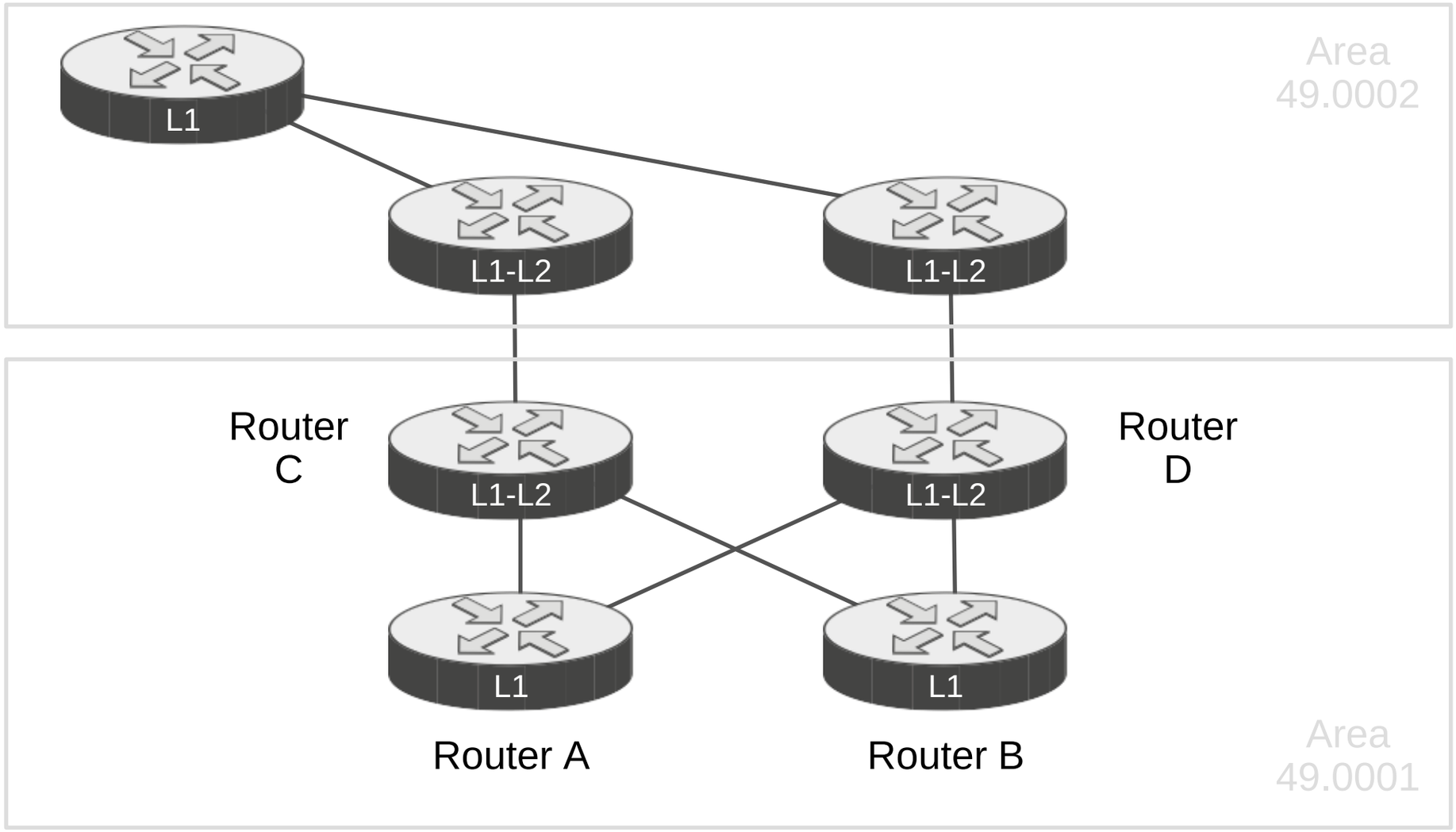
- Router A to Router B Traffic should follow the blue path. If Router A chooses the left L1-L2 router traffic follows the suboptimal path which is depicted with yellow arrow.

IS-IS Level Engineering

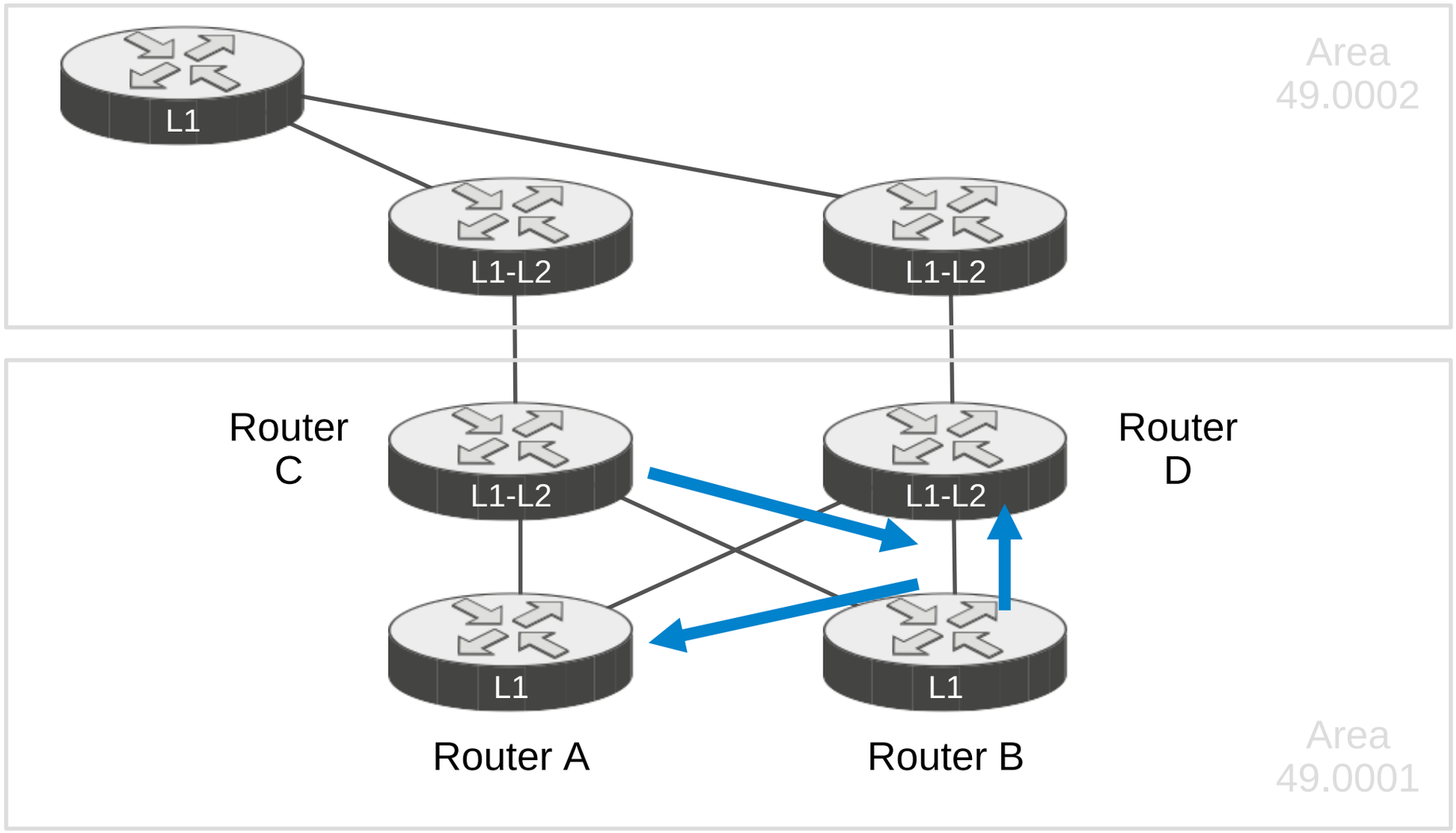
- In the below topology for the simplicity only the one of the region of the company is shown.
  - Level 1 routers use L1L2 routers to reach to the rest of the network.
- More interactions for different technologies with the case studies will be provided in the later sessions.

IS-IS Level Engineering

- Question 1: What would happen if the link between Router A and Router C fails?



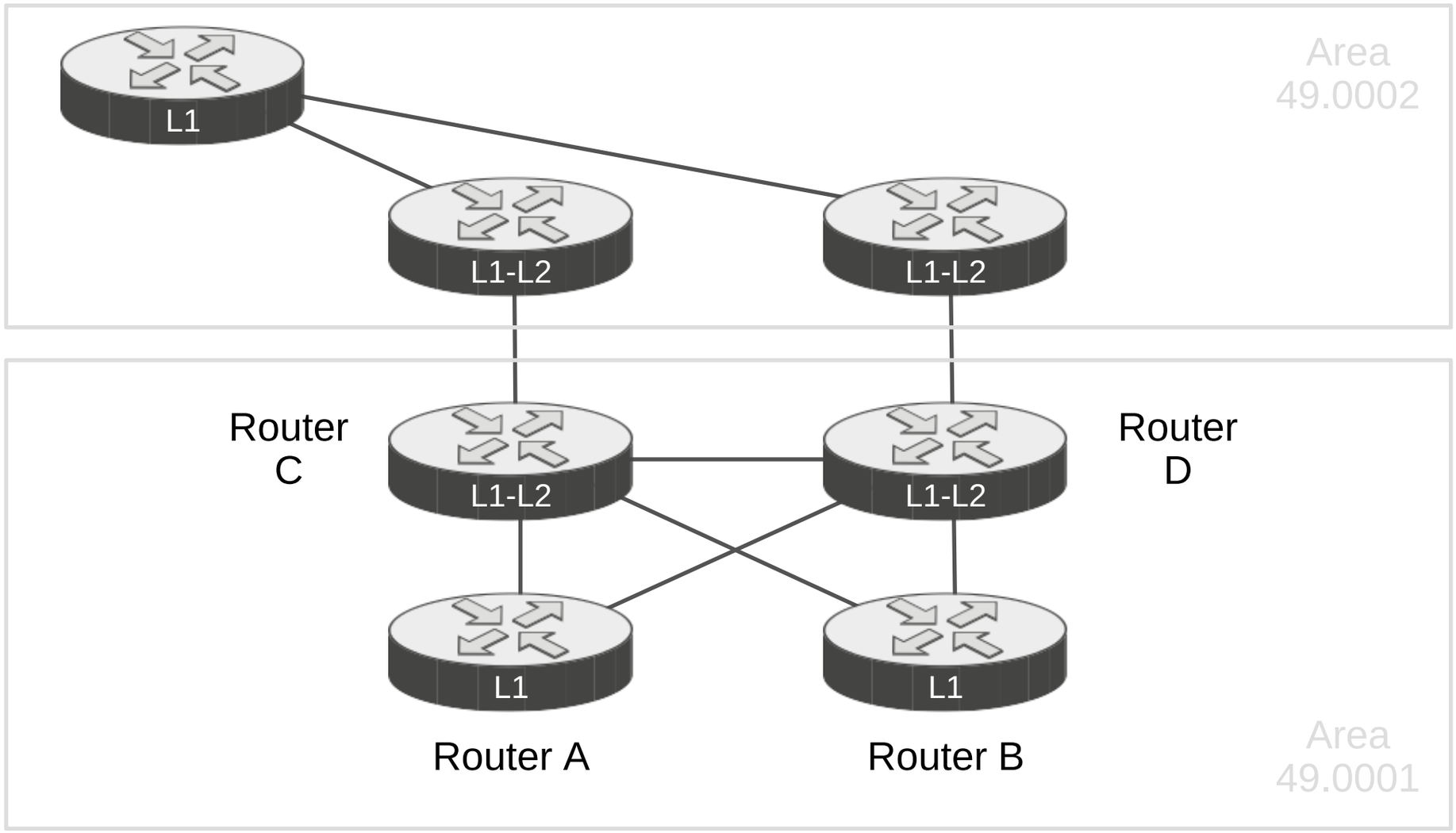
IS-IS Level Engineering



IS-IS Level Engineering

- When Router A to Router C link fails, the traffic from L1L2 router flows through other L1 router. This obviously creates sub optimal traffic flow but it can be tolerable since the faulty link is repaired after some time.
    - But also the L1 router (Router B in this topology) might have performance and bandwidth issue since it was handling half of the traffic before the failure.
- The solution is to connect a direct link between the L1L2 routers as shown in the below topology.

IS-IS Level Engineering



IS-IS Level Engineering

- Question 1: Which ISIS level new link between L1L2 routers should be placed into?

IS-IS Level Engineering

- Answer 1: It should be L1L2 link. If it would be only L2, Router C learns all the prefix of Router A from Router B as L1 LSP and from direct link as L2 LSP. Since L1 is preferred over L2, suboptimal path is still used.

That’s why the best solution is to place the direct inter L1L2 link into L1L2.

IS-IS in the CCDE Exam

- IS-IS Levels and the Areas
- IS-IS interaction with MPLS, using IS-IS in an Inter-AS MPLS VPNs
- IS-IS and other routing protocols redistribution
- IS-IS and MPLS Traffic Engineering
- IS-IS and other protocol comparison to answer when to select IS-IS

Summary

- IS-IS Theory - Link State - Areas and Levels - TLV and Hierarchy
- Fast Convergence and Fast Reroute with IS-IS
- IS-IS Levels and different POP and Core Designs
- IS-IS Scalability - DIS in LAN, Prefix Suppression and Multi-Level IS-IS
- IS-IS and Overlay Protocols
- IS-IS in the Datacenter and SP Networks
- IS-IS , BGP and MPLS Interaction - Fast Service Restoration through IGP-LGP Sync and LDP Session Protection

IS-IS

Intermediate System to Intermediate System

QUIZ

Question 1

Which OSPF Area is similar to IS-IS Level 1 sub domain?

A. Backbone Area

B. Stub Area

C. Totally Stub Area

D. Totally NNSA Area

Answer 1

D. Totally NNSA Area

Answer of this question is D. Because IS-IS level 1 domain allows route redistribution and only the default route is sent from the L2 domain. This was explained in the IS-IS chapter.

Question 2

If two IS-IS devices are connected to an Ethernet switch. Which below option provides fastest down detection to the IGP process?

A. Tuned IS-IS LSP Timers

B. BFD

C. Tuned IS-IS SPF Timers

D. IS-IS Hello Timers

Answer 2

B-BFD

Tuning LSP and SPF timers can improve the convergence of IS-IS in case of a failure but they don’t provide fast failure detection.

Reducing the hello timers can provide shorter failure detection time but cannot be tuned as much as BFD. Also since there is an Ethernet switch in between, port-failure event cannot trigger remote port interface down event. BFD is a best solution, especially if there is a node, which prevents end-to-end failure signaling between two devices.

Question 3

Why IS-IS overload bit is set in IGP – BGP synchronization?

A. In order to prevent routing loop.

B. In order to prevent traffic loss which can be caused by black-holing.

C. In order to prevent routing oscillation.

D. For fast convergence.

Answer 3

B- In order to prevent traffic loss which can be caused by black-holing.

As it was explained in the IS-IS chapter, it uses to signal the other routers so the node is not used as transit. If node would be eligible to be used as primary path, blackhole would occur since BGP and IGP converges times are not the same.

IGP should wait BGP before staring to accept network traffic.

Question 4

Which of the below mechanisms are used to slow down the distribution of topology information caused by a rapid link flaps in IS-IS? (Choose Two)

A. ISPF

B. Partial SPF

C. Exponential Back Off

D. LSA Throttiling

E. SPF Throttling

Answer 4

C. Exponential Back Off

D. LSA Throttiling

Exponential back off mechanism is used in OSPF and IS-IS to protect the routing system from the rapid link flaps. Also LSA throttling timers can be tuned to protect the routing system from these types of failures.

But LSA throttling timers tuning also will affect on convergence so careful monitoring is necessary if there is IS-IS fast convergence requirement in design.

Question 5

When would it be required to leak the prefixes from Level 2 to Level 1 subdomain? (Choose Two)

A. When optimal routing is necessary from the Level 1 routers towards the rest of the network.

B. When MPLS PE devices are configured in Level 1 domain.

C. When ECMP is required from Level 1 domain to the rest of the network.

D. When unequal cost load balancing is required between L1 internal routers and the L1-L2 routers.

Answer 5

A. When optimal routing is necessary from the Level 1 routers towards the rest of the network.

B. When MPLS PE devices are configured in Level 1 domain.

Unequal cost load balancing is not supported in IS-IS. Even if you leak the prefixes it won't work. ECMP is done by hop by hop. Even L2 prefixes are not leaked into the L1 domain; still internal L1 domain routers can do the ECMP towards L1-L2 routers if there is more than one L1-L2 router. But L1-L2 routers may not do ECMP. Thus Option C is incorrect.

When MPLS PE is inside L1 domain, LDP cannot assign a label to the PE loopbacks since the remote loopbacks are not known. Internal L1 routers only learn default route as it was explained in the IS-IS chapter.

And whenever optimal routing is required, of there is available, more specific information can help for that.

Question 6

How many level of hierarchy is supported by IS-IS?

A. One

B. Two

C. Three

D. As many as possible

Answer 6

B. Two

IS-IS supports two level of hierarchy. Hierarchy is common network design term, which is used to identify the logical boundaries.

IS-IS Level 1 and IS-IS Level 2 domains provide maximum two levels of hierarchy. Level 2 IS-IS domain is similar to Backbone area in OSPF, Level 1 IS-IS domain is similar to Totally NSSA area in OSPF.

Question 7

If some prefixes are leaked from the IS-IS level 2 domain into level 1 domain, how IS-IS prevents those prefixes to be advertised back in Level 2 domain?

A. Route tag should be used.

B. ATT bit prevents prefixes to be advertised back in Level 2 domain.

C. U/D bit is used to prevent prefixes to be advertised back in Level 2 domain.

D. They wouldn’t be advertised back in Level 2 domain anyway.

Answer 7

C. U/D bit is used to prevent prefixes to be advertised back in Level 2 domain.

If some reason some prefixes are leaked from Level 2 into level 1, U/D bit in IS-IS prevents those prefixes to be advertised back into IS-IS level 2 domain. This is an automatic process, doesn’t require configuration. It is a loop prevention mechanism in IS-IS route leaking.

Question 8

Which below mechanism is used in IS-IS full mesh topologies to reduce the LSP flooding?

A. Elect a DIS and Backup DIS.

B. Use IS-IS Mesh Group.

C. Use DR and BDR.

D. Deploy Multi Level IS-IS design.

Answer 8

B. Use IS-IS Mesh Group.

Full mesh topology could be in any level, either Level 1 or Level 2 in multi level design. Thus having Multi level design won't help for LSP flooding if the topology already in any particular level.

It is similar to have BGP Confederation for scalability but still in sub AS you have to configure full mesh IBGP or for scalability you implement Route Reflector inside confederation sub AS. There is no Backup DIS in IS-IS, there is only a DIS (Designated Intermediate System), thus the Option a is incorrect. DR and BDR is an OSPF feature not the IS-IS.

Question 9

If an IS-IS router is connected to three links and redistributing 100 EIGRP prefixes into the domain, and the design is at/single level IS-IS design, how many IS-IS LSP is seen in the domain?

A. 100 IS-IS LSP

B. 3 IS-IS LSP

C. 300 IS-IS LSP

D. 1 IS-IS LSP

Answer 9

D.1 IS-IS LSP

There will be different TLVs for internal and external routes but there will be only 1 IS-IS LSP for the domain. If there would be multi level IS- IS design two LSP would be seen but since the question says that it is a at/single level deployment, there will be only 1 IS-IS LSP, either L1 or L2.

Question 10

Which below statements are correct for IS-IS design?

A. Topology information is not advertised between IS-IS levels.

B. Starting with Flat/Single Level 2 IS-IS design makes the possible future IS-IS deployment easier.

C. IS-IS level 2 route is preferred over level 1 route in IS-IS.

D. IS-IS uses DIS and Backup DIS on the multi access links.

Answer 10

A. Topology information is not advertised between IS-IS levels.

B. Starting with Flat/Single Level 2 IS-IS design makes the possible future IS-IS deployment easier.

There is no backup DIS in IS-IS, thus Option D is incorrect.

IS-IS level 1 routes are preferred over IS-IS level 2 routes. Similar to OSPF intra area routes preferred over Inter Area routes. Thus option C is incorrect as well.

IS-IS – Study Resources:

- Books :
- http://www.amazon.com/--Deployment-IP-Networks/dp/0201657724/ref=sr_1_1?ie=UTF8&qid=1436565940&sr=8-1&keywords=is-is+russ+white
- Videos
- Ciscolive Session – BRKRST – 2338
- Podcast :
- http://packetpushers.net/show-89-ospf-vs-is-is-smackdown-where-you-can-watch-their-eyes-reload/

EIGRP

Enhanced Interior Gateway Routing Protocol

Agenda

- EIGRP Theory
- EIGRP Fast Convergence
- EIGRP Fast Reroute and Difference between EIGRP FS and EIGRP FRR
- EIGRP Scalability
- Overlay Technologies and EIGRP (GRE, mGRE, DMVPN and LISP)
- EIGRP in the Datacenter
- EIGRP in the Service Provider
- EIGRP Design Best Practices
- EIGRP Advantages and Disadvantages
- Case Studies
- EIGRP in the CCDE exam
- Summary
- Bonus Materials

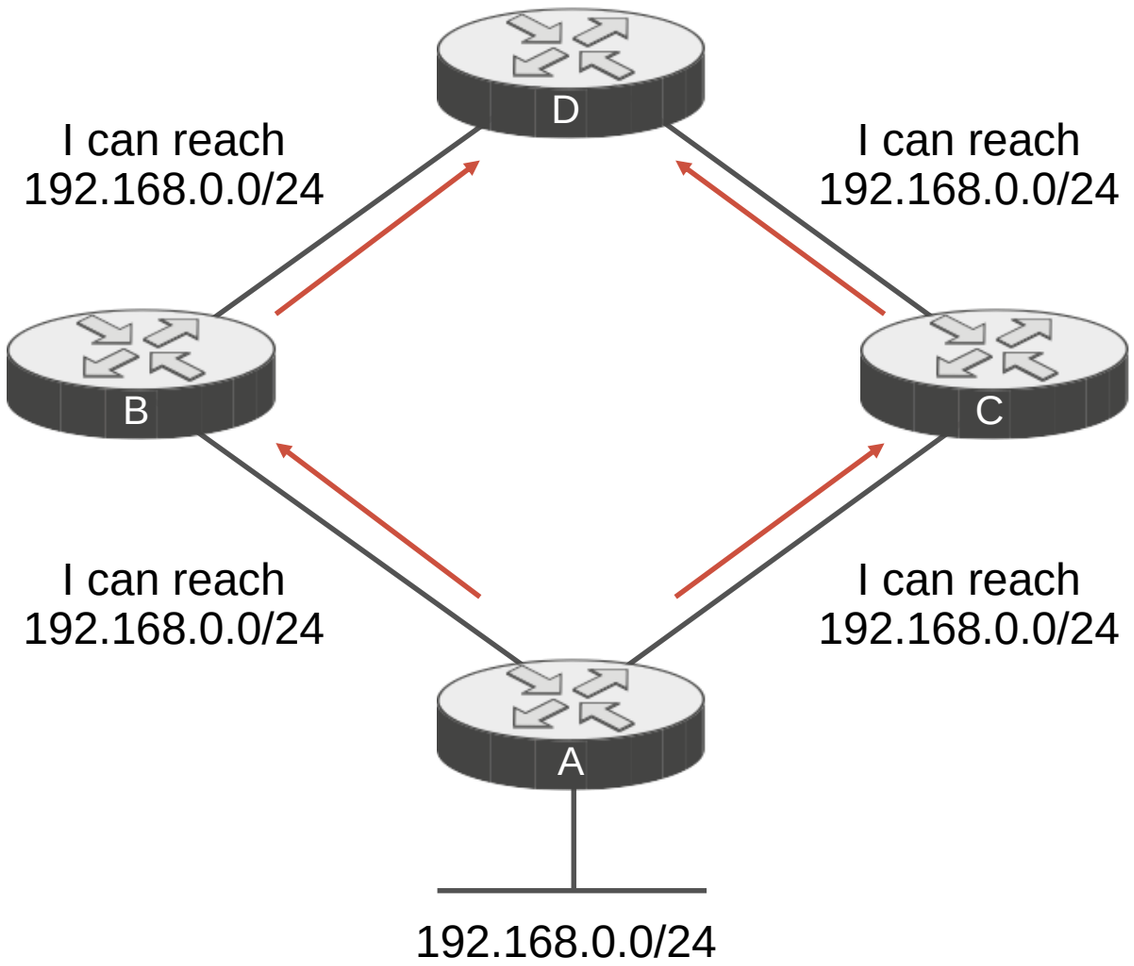
Theory

- If the requirement is to use Enterprise level, scalable, minimal configuration for the Hub and Spoke topology such as DMVPN, then only choice is the EIGRP.
  - Two routers become neighbor when they see each other's hello packets, 224.0.0.10
- EIGRP is distance vector protocol and unlike OSPF and IS-IS, topology information is not carried between the routers.
  - EIGRP routers only know the networks/subnets which their neighbors advertise to them

EIGRP RFC

- EIGRP is Cisco preparatory protocol, May 2016, RFC (Informational Track) 7868 published.
  - In EIGRP RFC, EIGRP Stub feature has been not shared.
- Vendor interoperability is an issue with EIGRP, when you want to deploy different vendor equipment, you may have an issue.

EIGRP Routers don’t have topology information of the network



In Distance vector protocols topology information is hidden beyond the next hop.

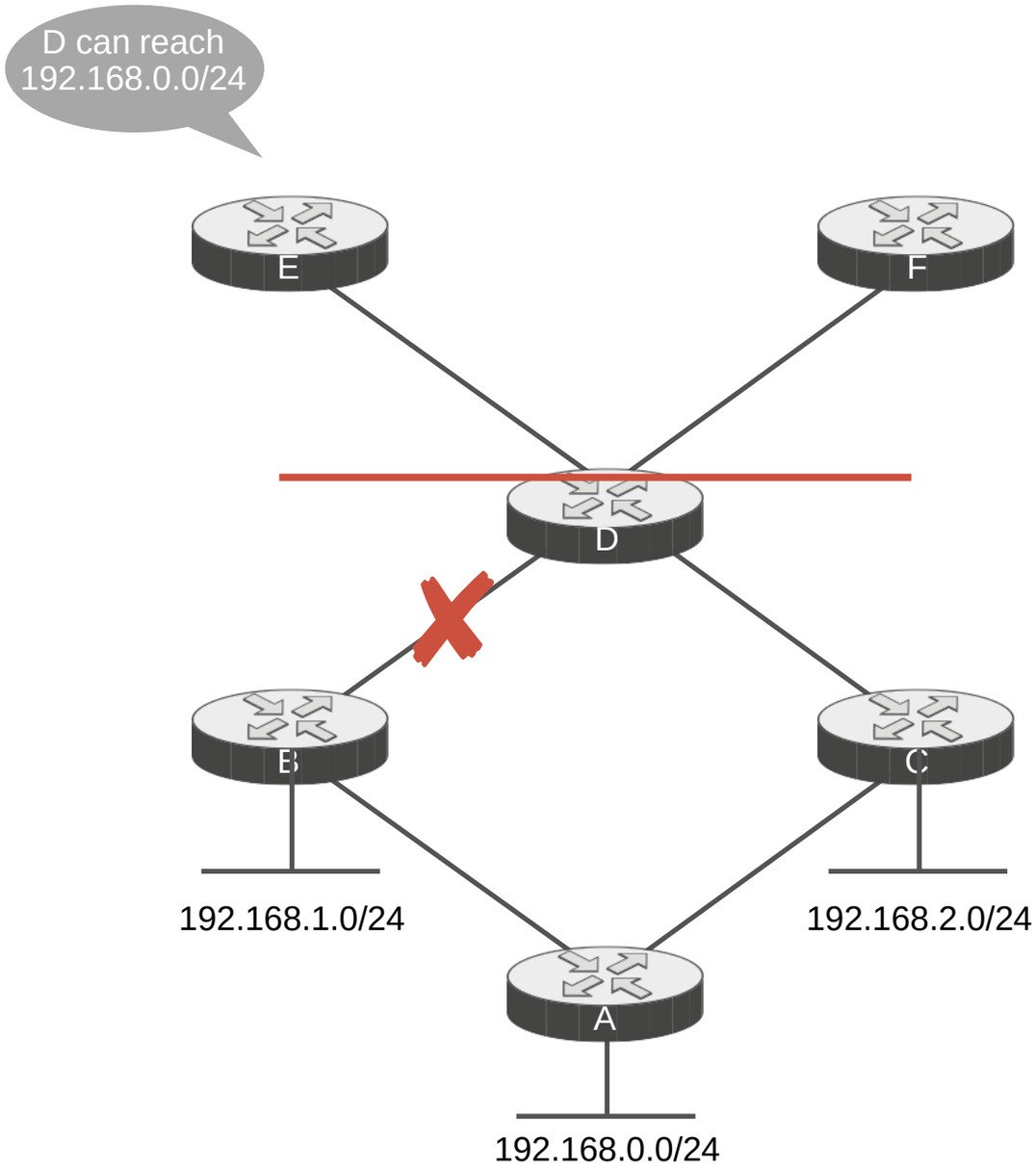
EIGRP only knows prefix and next hop information.

In this topology A advertises that it has 192.168.0.0/24 network.

Router B and C only advertise to Router D tat they can reach 192.168.0.0/24, not that they are connected to Router a which is then connected to Router A.

Router D learn to reach to 192.168.0.0/24, it can use Router B or Router C, but Router D doesn’t know which routers or Connections exist beyond Router B and Router B.

EIGRP Routers don’t have topology information of the network



For E and F Topology information is hidden here.

They only know that 192.168.0.0/24 can be reached via Router D.

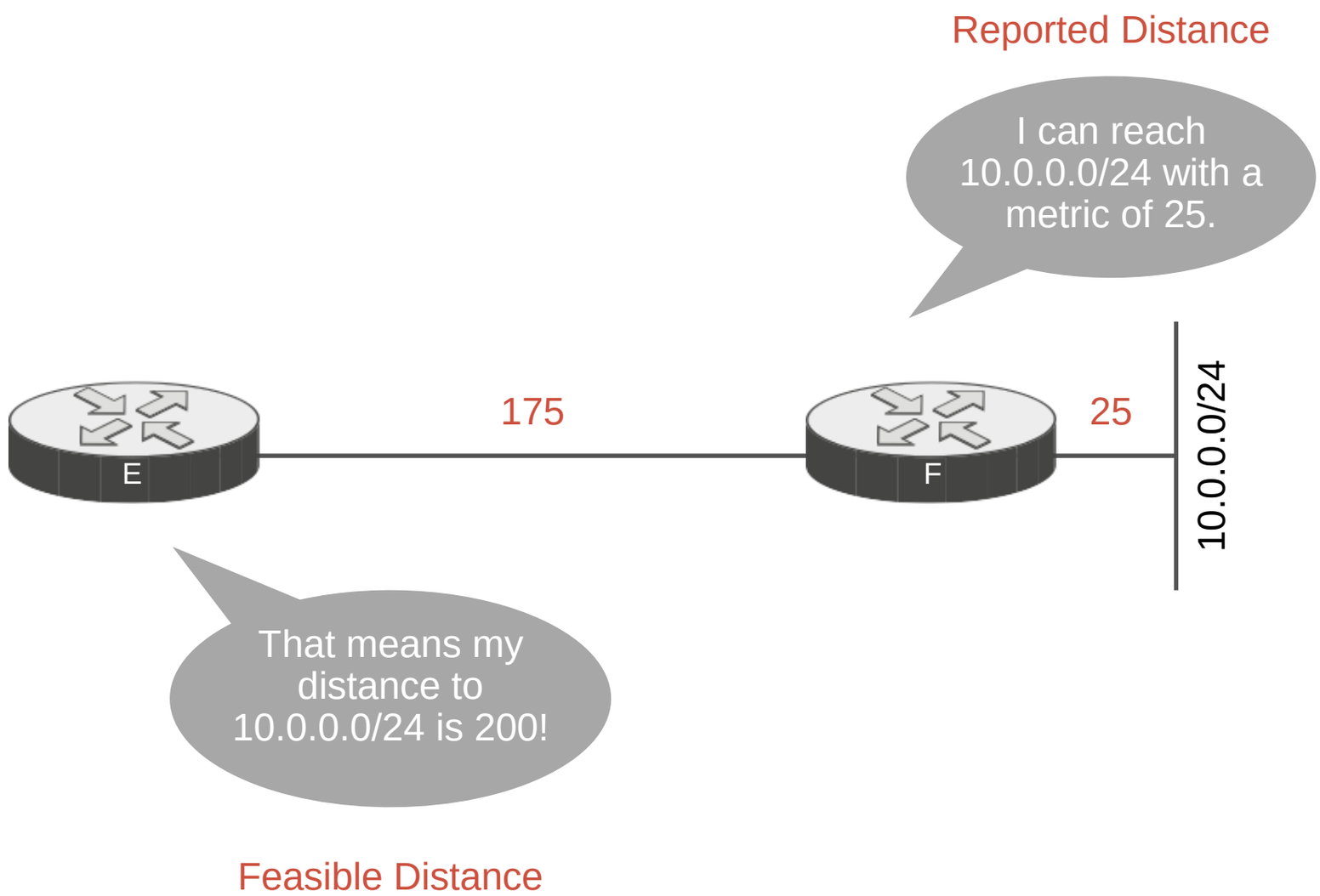
In this network, if all three subnets are summarized at the Router D and send to Router E and F, when link B-D fails (or any link between the routers in this topology) since Router D still can reach to all subnets via alternate links Router E and F doesn’t need to know all the specific subnets. Because only the exit point Router E and Router F is Router D. Summarization may create suboptimal routing if there is more than one exit point though. (This is general rule for all the IGPs)

- As a distance vector protocol, nodes in EIGRP domain don't keep the topology information of all the other nodes. Instead they trust what their neighbors tell them.

EIGRP Terminology

- Feasible distance is the best path, primary path Successor is the next-hop router for the route.
  - Each EIGRP router advertise its routes with a metric called Reported distance, also known as advertised distance.
- Feasible successors are the routers which satisfy the feasibility condition. These are the Backup routers.

Feasible Distance and Reported Distance



•Feasible successors are placed in EIGRP topology table.

- Reported distance is the feasible distance of the neighboring router.

feasible

- Having feasible successor provides fast convergence to the EIGRP, which will be explained in EIGRP Scalability.

feasible

EIGRP Metric Calculation

•Classical EIGRP uses 5 K Values for metric calculation.

- K1 through K5

1. K1 - Lowest bandwidth of route
2. K2 - Worst Load on route based on packet rate
3. K3 - Cumulative interface delay of route
4. K4 - Worst reliability of route based on keep alive
5. K5 - Smallest MTU in path

me
tric

EIGRP Metric Calculation

- EIGRP named mode introduced 6th K value which is defined for jitter and energy control.

  - In both classical EIGRP and named mode, only K1 and K3, Bandwidth and Delay attributes are used for metric calculation.

EIGRP Metric Calculation

- Minimum bandwidth and cumulative delay along the path are considered for the EIGRP metric.

  - If there will be traffic engineering, which one you should change? Bandwidth or Delay? Why?

me
tric

EIGRP Metric Calculation

- EIGRP will not form neighborship if K values are mismatch.
  - EIGRP will not form neighborship if AS numbers are mismatch.

me
tric

EIGRP Fast Convergence and Fast Reroute

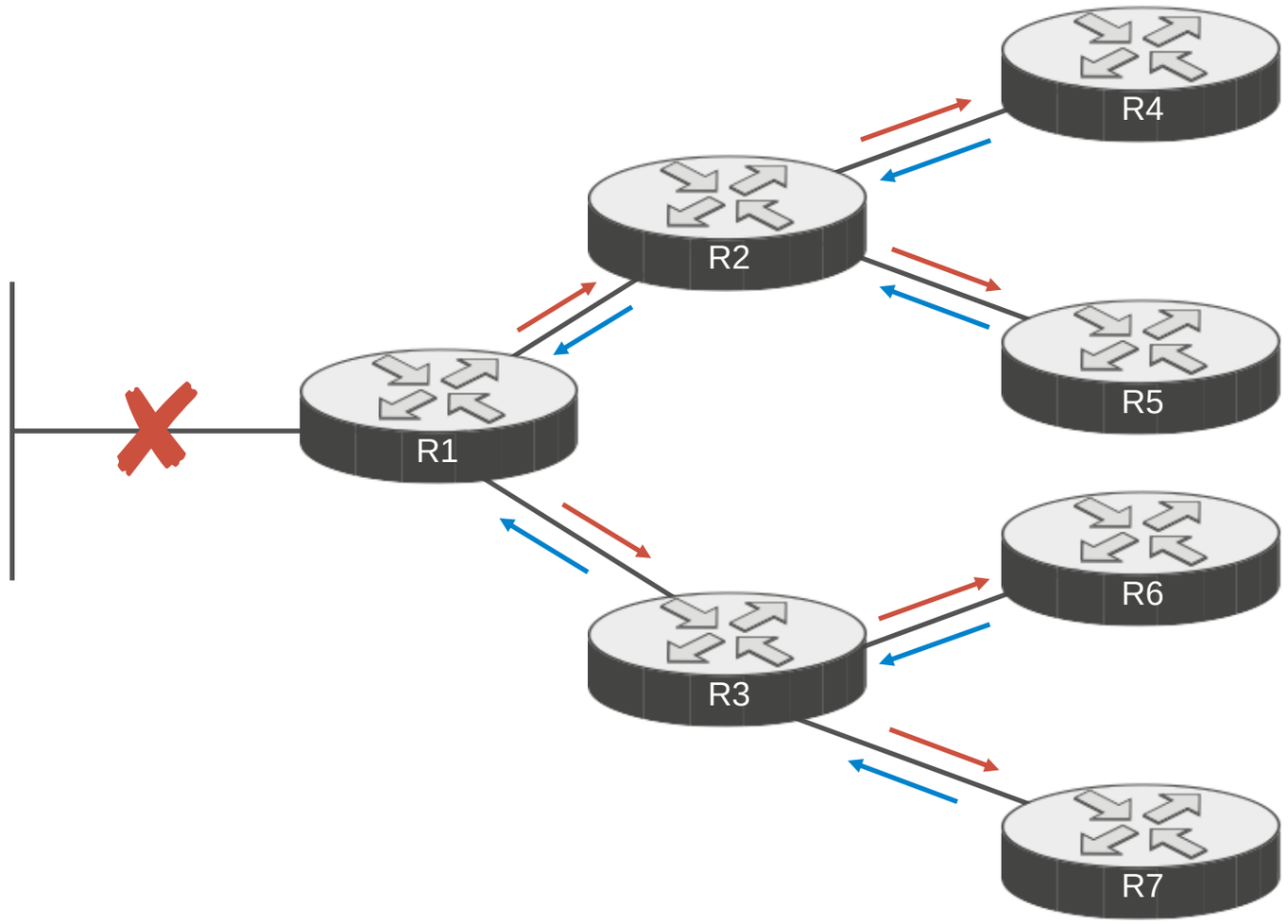
- We can use some techniques to converge faster than default.
  - Fast failure detection is very important.
- After detection, failure information is propagated to the rest of the network.

EIGRP Fast Convergence and Fast Reroute

- If there is a feasible successor, without running DUAL algorithm, backup route is taken from the EIGRP topology table and placed into the RIB and FIB.

- If there is no feasible successor (backup route), prefixes are marked as active and EIGRP query is sent to all EIGRP neighbors of the node.

EIGRP Query is sent to all EIGRP neighbors

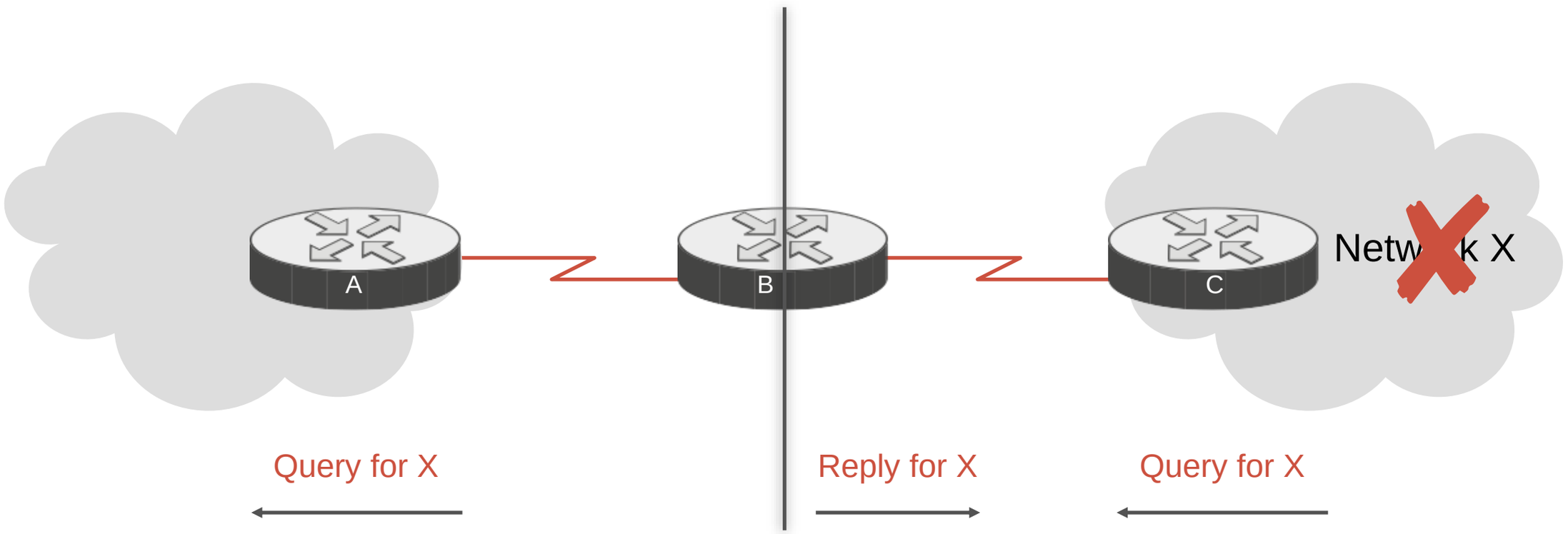


EIGRP Query is sent to all EIGRP neighbors

- The router will have to get all of the replies from its neighbors before the router calculates the new path.
  - If any neighbor fails to reply the query within three minutes, this route is stuck in active (SIA), and the router resets the neighbor which fails to reply.
- Solution is to limit query range which is also known as EIGRP query domain!

EIGRP Query Scope in multi AS scenario

EIGRP queries are not bounded by AS Boundary.
Queries from AS 1 will be propagated to AS 2 in below topology



EIGRP Query Scope in multi AS scenario

- EIGRP Feasible successor helps for fast convergence.
  - For fast failure detection, BFD can be used.
- EIGRP supports IP FRR as well, in this case, feasible successor routes are not only placed in EIGRP topology table, but also in RIB and FIB as a standby route.

EIGRP Query Scope in multi AS scenario

- Having EIGRP Feasible Successor (Backup Route) is important for the fast convergence, but how feasible successor is calculated?
  - Answer is, backup route should satisfy the feasibility condition!
- What is EIGRP feasibility condition?

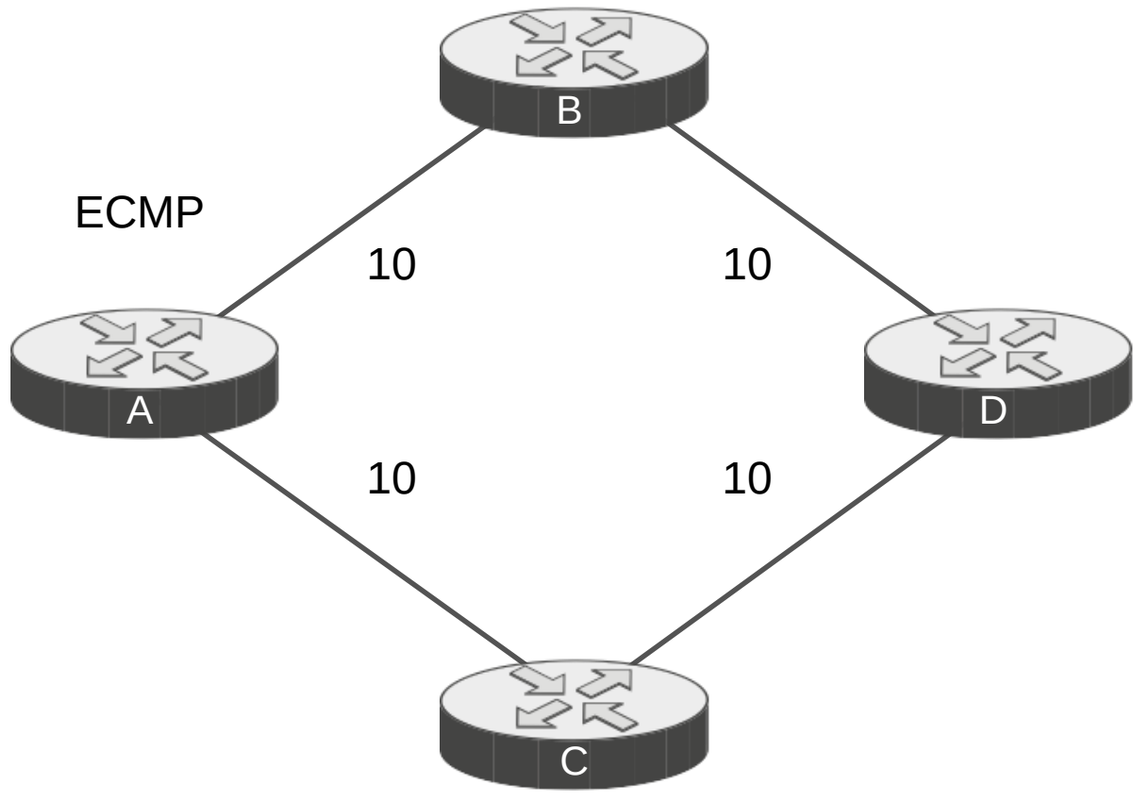
EIGRP Feasibility Condition

- If node A'S next-hop router's reported distance is less than the feasible distance of node A than A's backup router is loop free. And it can be used as a backup and placed in the topology table.

feasibility

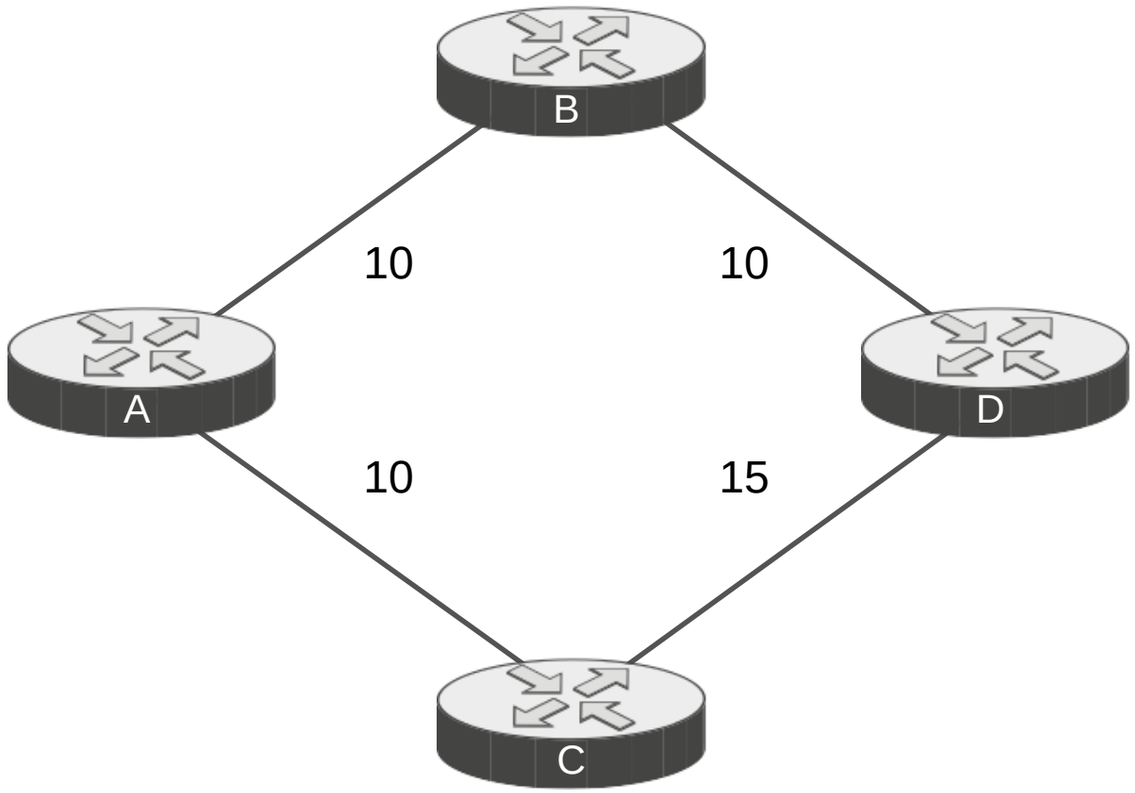
- In order the path to be placed in the EIGRP topology table, it has to satisfy the EIGRP feasibility condition.

EIGRP Feasibility Condition



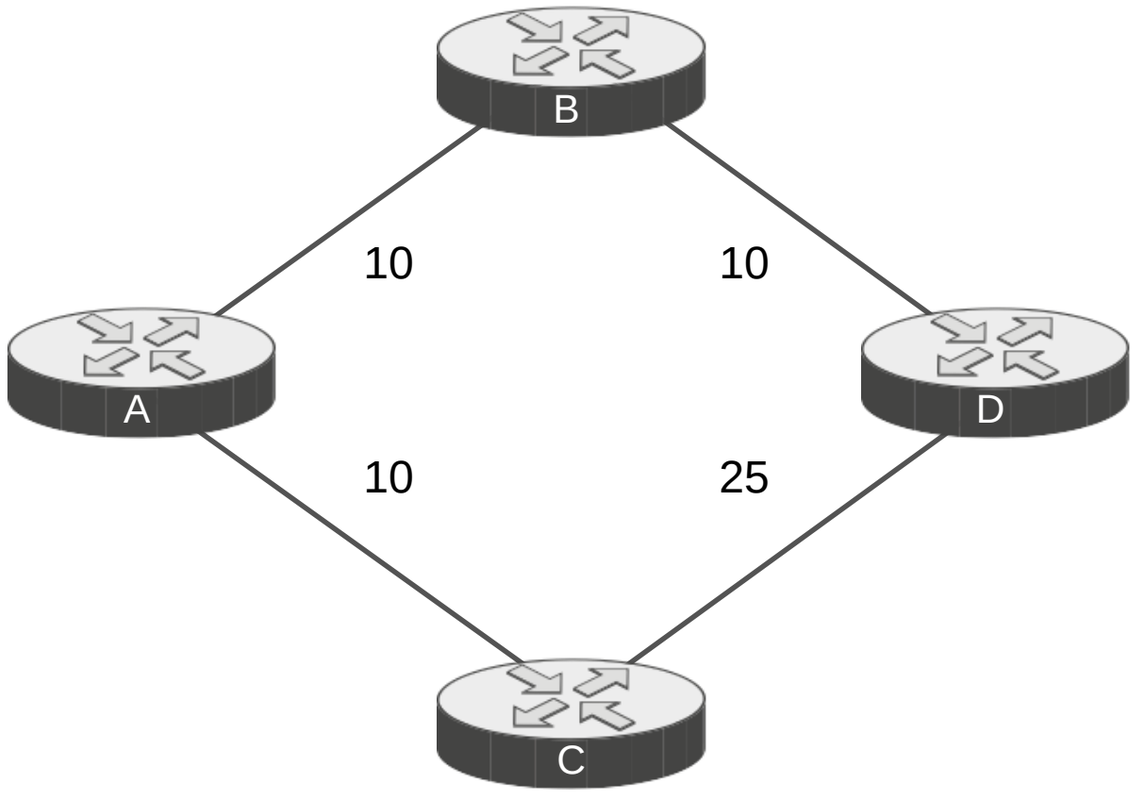
From the Router A’s point of view, Router B and Router C is the Equal Cost routers, so both ABD and ACD path can be used. Router A installs both Router B and Router C in not only EIGRP topology table but also in the routing table.

EIGRP Feasibility Condition



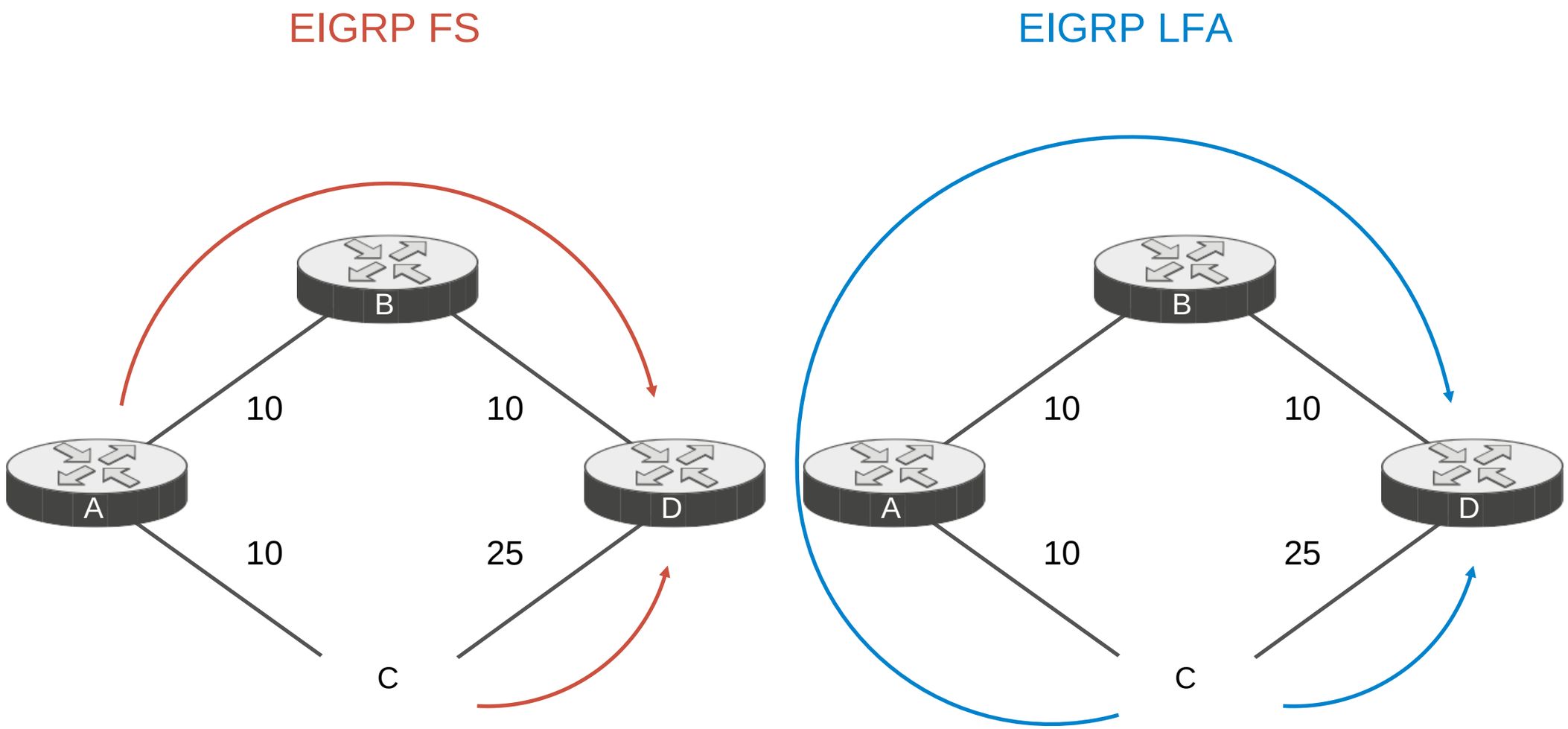
Router C to Router D link cost is 15. In order to satisfy feasibility condition for the Router A, Router C - Router D link cost should be smaller than, A- B – D total cost. Since 15 < 10 + 10, Router C can be used as a backup router by the Router A to Reach Router D

EIGRP Feasibility Condition



Router C to Router D link cost is 25. In order to satisfy feasibility condition for the Router A, Router C - Router D link cost should be smaller than, A- B – D total cost Since 25 > 10(A-B) + 10(B-D), Router C can not be used as a backup router by the Router A to Reach Router D. What if C-D would be 20?

Difference between EIGRP FS and EIGRP LFA Calculation



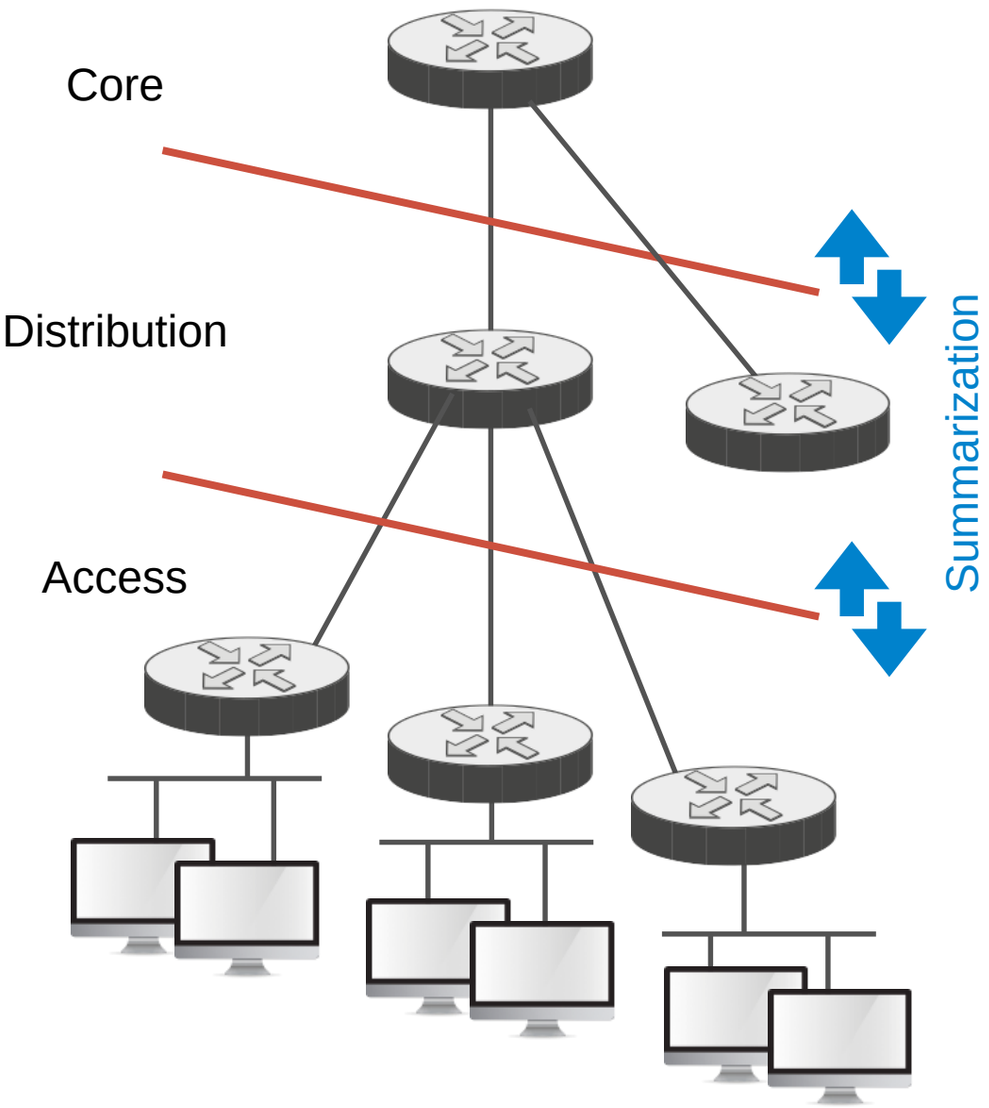
EIGRP Scalability

- EIGRP supports manual summarization in any interface at any router in the network, because there is no topology information.

scalability

- Unlike OSPF and IS-IS, EIGRP is not limited with two tier hierarchy, summarization can be done at multiple points.

EIGRP Summarization can be done at multiple points, such as both Core and Distribution



EIGRP Summarization can be done at multiple points, such as both Core and Distribution

•For EIGRP, limiting query domain is very important for scalability.

- Query domain can be kept small by introducing summarization and filtering.

summarization

EIGRP Summary and Filters

•Summary and Filters reduces the query domain to one hop beyond.

- Summary and Route filters help for scalability but the problem, nodes still send and receive the EIGRP query in case link or node fails.

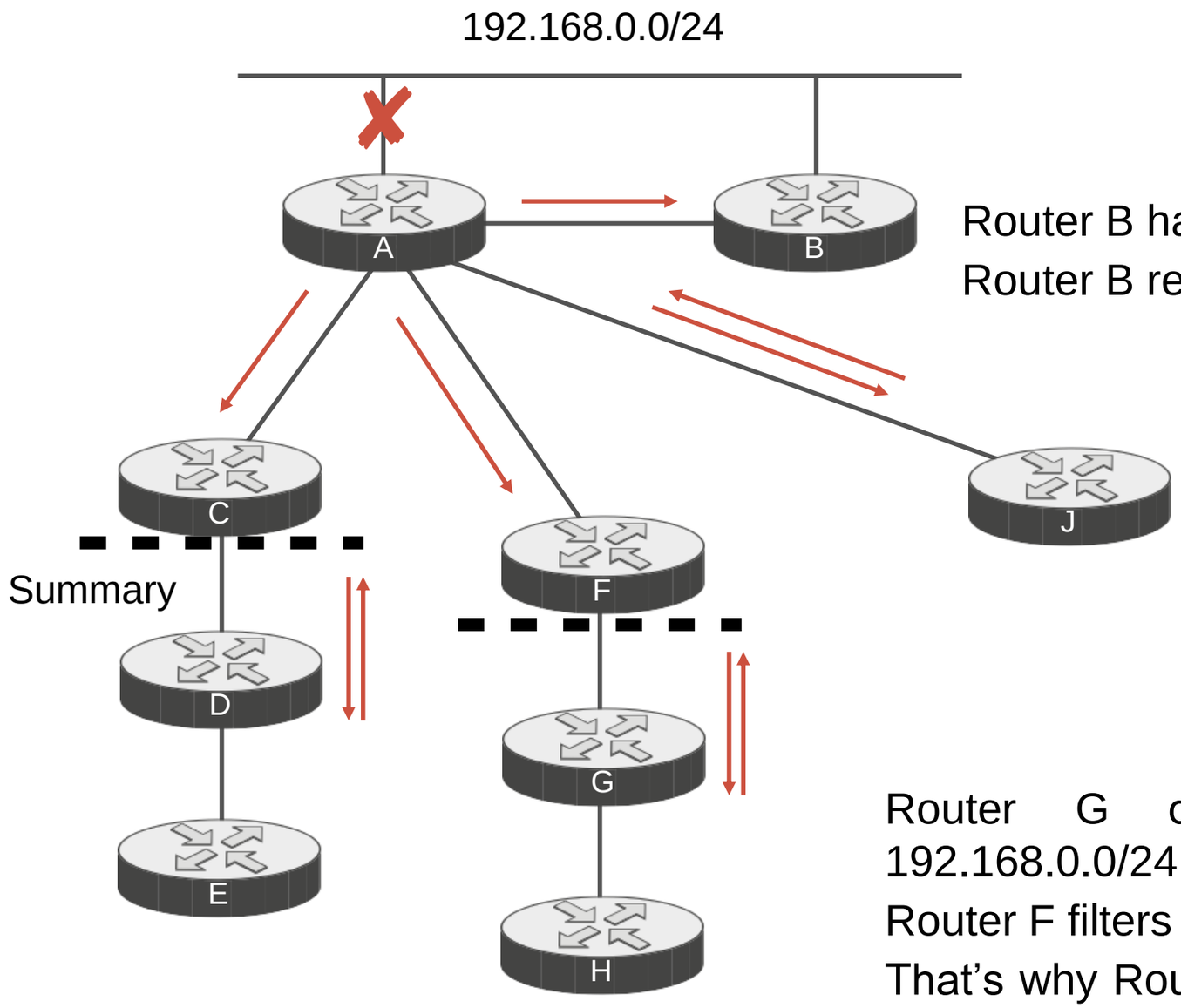
•If you want to stop receiving EIGRP query, then EIGRP stub feature should be enabled.

filt
ers

When EIGRP node loses the Connection to the prefixes. If there is no feasible successor installed in EIGRP topology database.

Router is marked as active and EIGRP query is sent to every neighbor.

Router D doesn't know the 192.168.0.0/24 network. Router C sends a summary 192.168.0.0/16. That's why it replies without asking to Router E.



Router B has an alternate path.
Router B replies immediately.

Router J doesn't have any EIGRP neighbor. It replies the Query immediately.

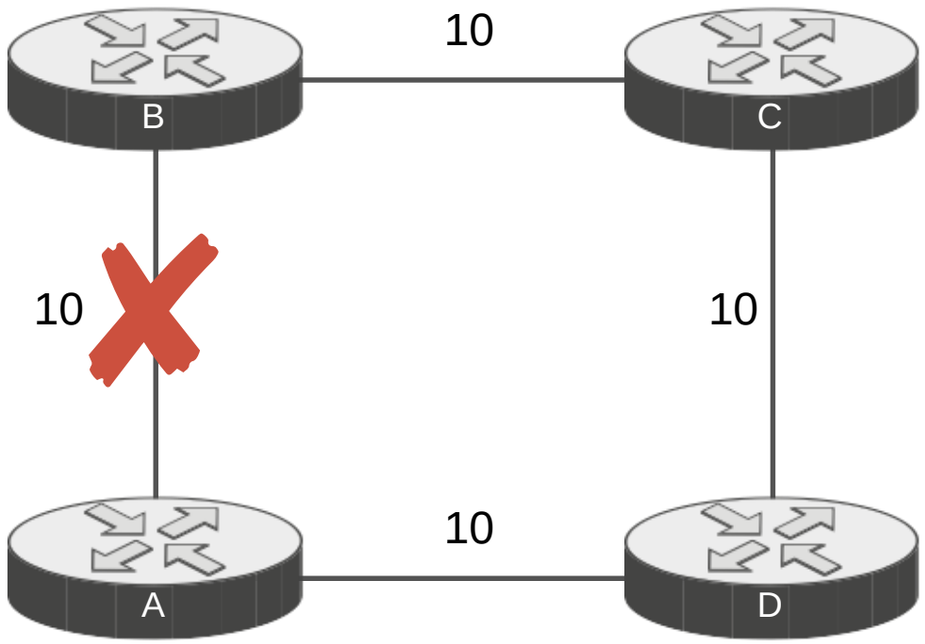
Router G doesn't know the 192.168.0.0/24 network.

Router F filters the 192.168.0.0/24

That's why Router G replies without asking to Router H.

Query Domain Range in Ring Topology

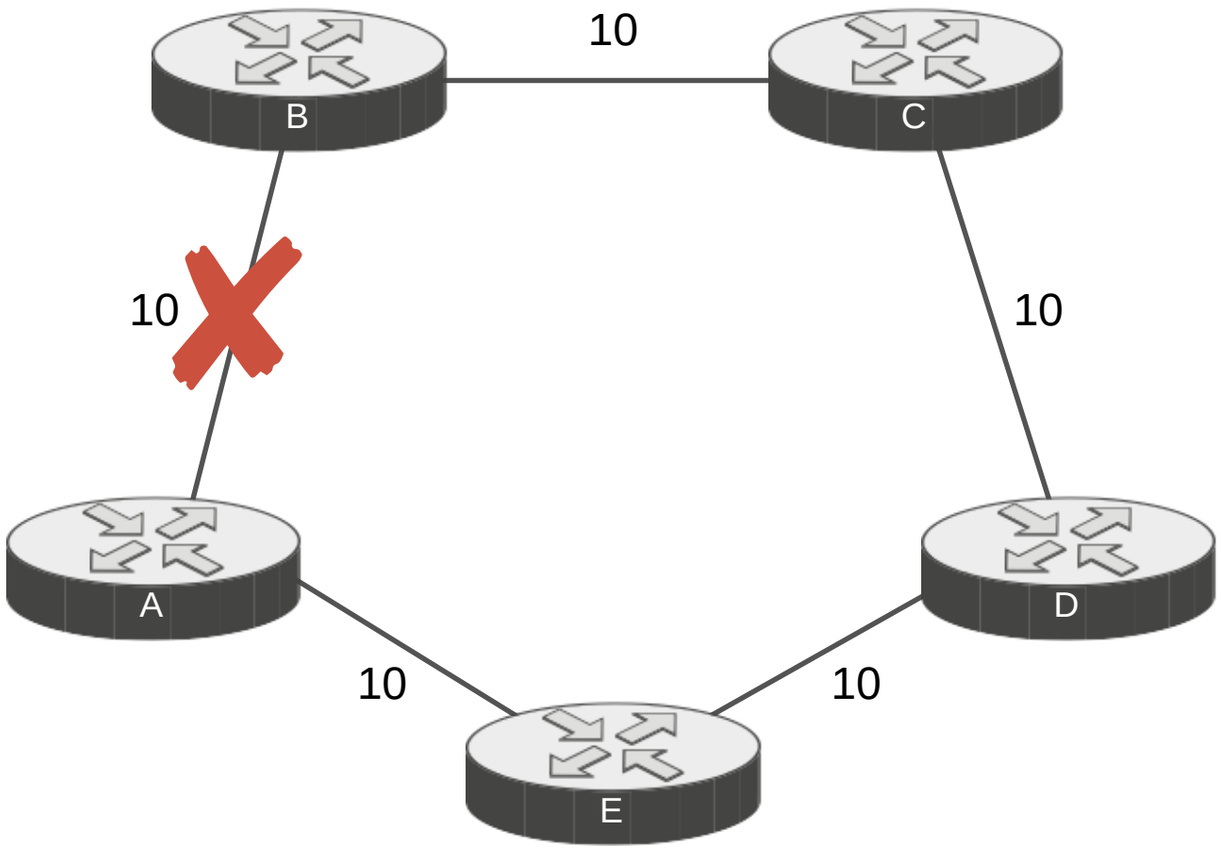
- EIGRP is best in Hub and Spoke topology but not good at Rings. See the below examples.



in this network, if A-B link fails, Router A sends a query to Router D, Router D sends a query to Router C.

Siince Router C is not using Router D to reach Router B, query stops at Router C and Router C replies that it has an Alternate path.

Query Domain Range in Ring Topology



in this network, if A-B link fails, Router A sends a query to Router E, Router E sends a query to Router D and Router D sends a query to Router C.

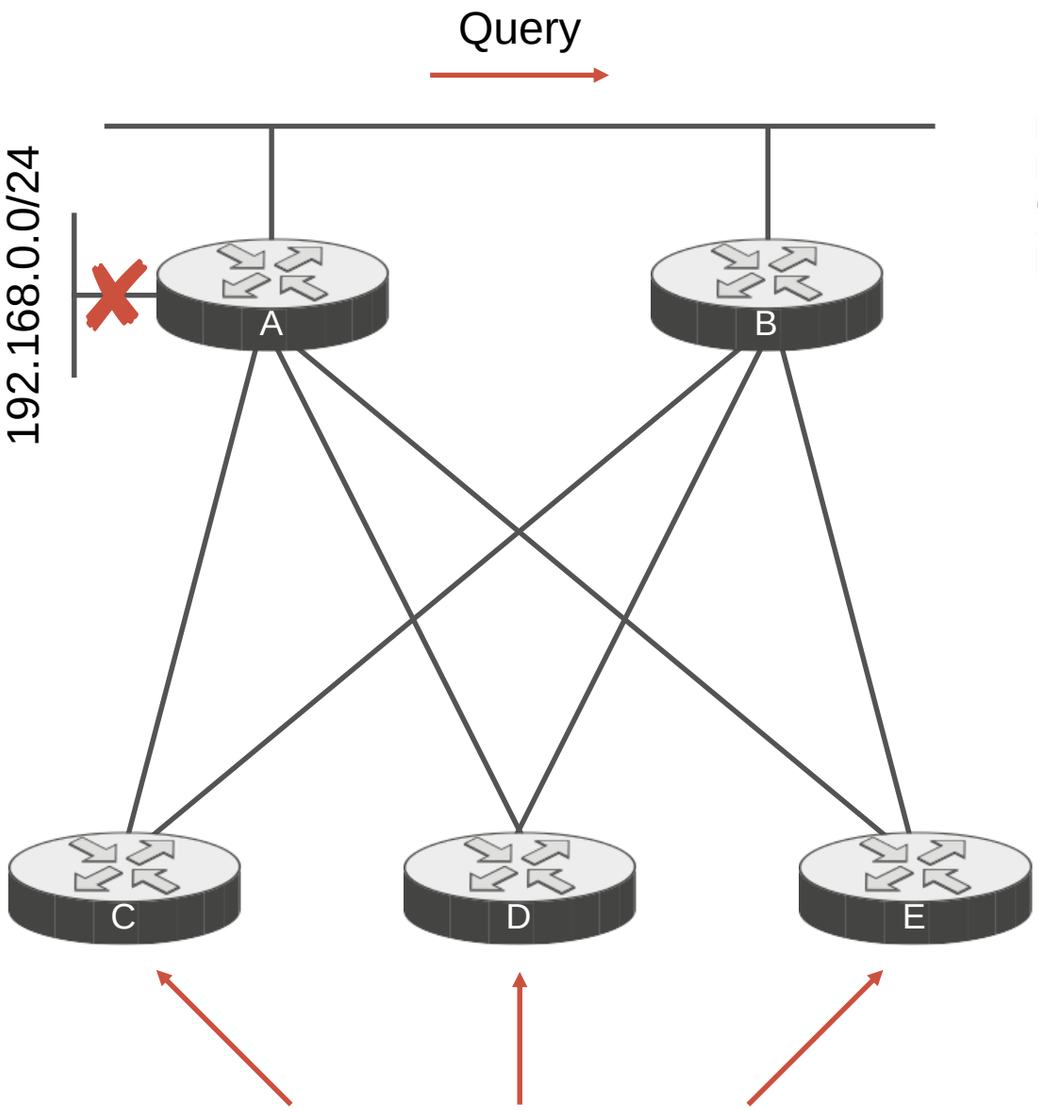
Query range in this network is three hops.

Ring topology is very challenging for all IGP protocols including EIGRP.

If this topology would be triangle instead of ring EIGRP could find a feasible successor and convergence would be very fast.

- In Hub and Spoke, Spokes should be a STUB. This is similar behavior of BGP transit AS. In order to prevent being Transit AS as a customer, Customer only send his AS path to the upstream and filters all the other AS.

EIGRP Stub allows the router to be not queried, thus router does not advertise the routes to peer if route is learned from another peer.



Always use EIGRP Stub in Hub and Spoke topology Spoke sites shouldn't be used transit site

EIGRP STUB

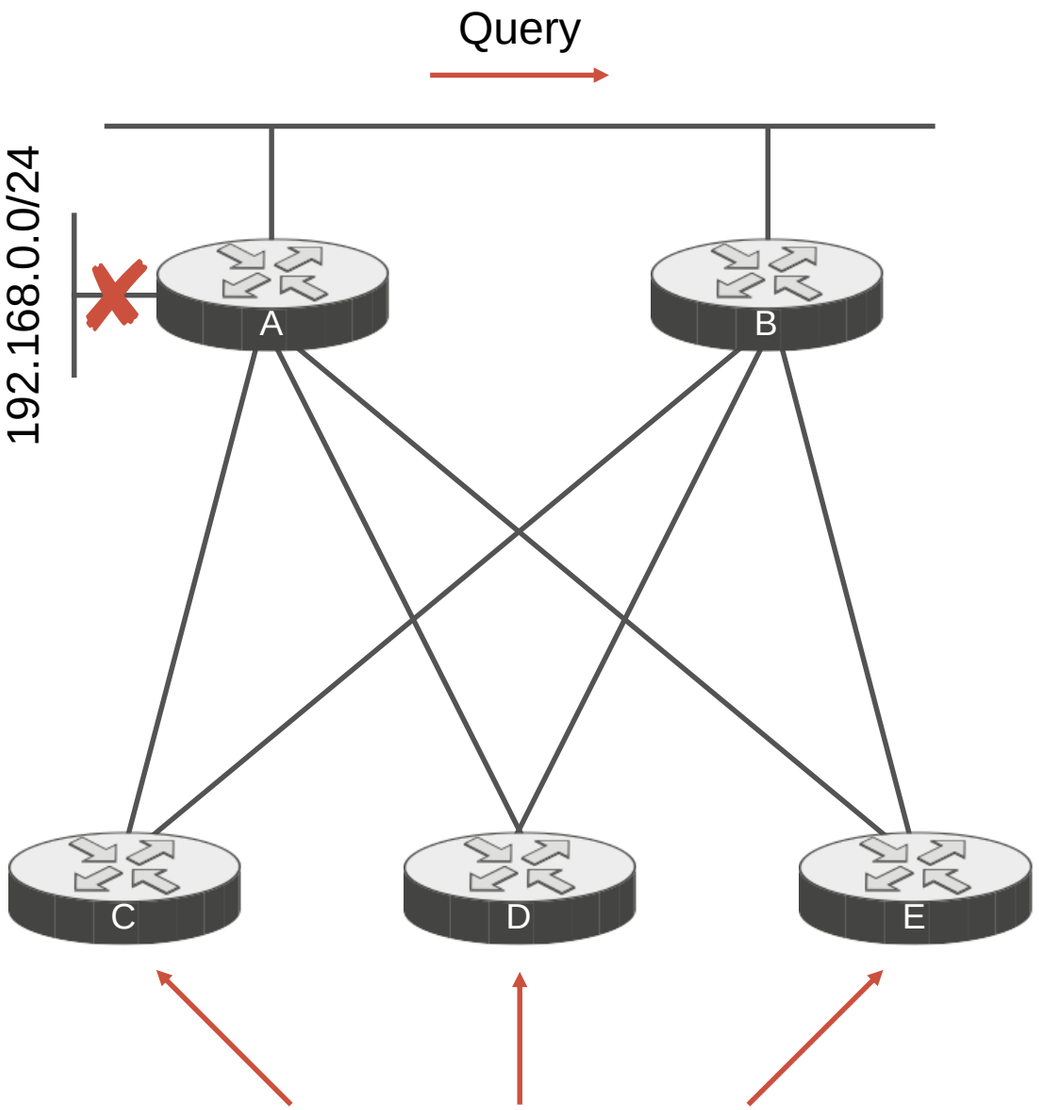
When EIGRP Stub feature is enabled spoke sites are not used as transit site.

Also Hub site doesn't even send an EIGRP query if 192.168.0.0/24 network fails as in the picture Router A sends a query only to Router B. This helps for the convergence, provide faster convergence.

If EIGRP Stub wouldn't be enabled but filtering or summarization is enabled at the Hub sites, spokes sites still would receive a query and they would process.

This might create a resource problem on the Hub for large scale Hub and Spoke deployment.

Access network always should be configured as stub



Always use EIGRP Stub in Hub and Spoke topology Spoke sites shouldn’t be used transit site

EIGRP is the most stable IGP protocol in Hub and Spoke Topologies.

Spoke sites have two connections for redundancy not to be used as transit between Hub sites.

Having a direct link between the Hub sites is important.

Otherwise Spoke sites might be used as transit nodes.

EIGRP Scalability Summary

•Limiting query domain is important for scalability.

- Summary, Filtering and most importantly EIGRP Stub feature limits the EIGRP query range/scope.

•EIGRP allows manual summarization at each hop, any interface at any router, no need an ABR, L1L2 router as in the case with OSPF and IS-IS.

EIGRP Scalability Summary

•Ring topology is hardest topology for EIGRP scalability.

- Installing more Feasible Successors increase resource usage but provides fast convergence.

Overlay Technologies and EIGRP (GRE, MGRE, DMVPN, GETVPN, LISP)

- EIGRP can work on top of many overlay technologies.
  - GRE, MGRE, DMVPN, GETVPN and LISP can be used to create overlay/VPN in the networks.
- EIGRP can be used for these overlay mechanisms as an underlay infrastructure routing protocol.

Overlay Technologies and EIGRP (GRE, MGRE, DMVPN, GETVPN, LISP)

- EIGRP works over GRE, MGRE and DMVPN.

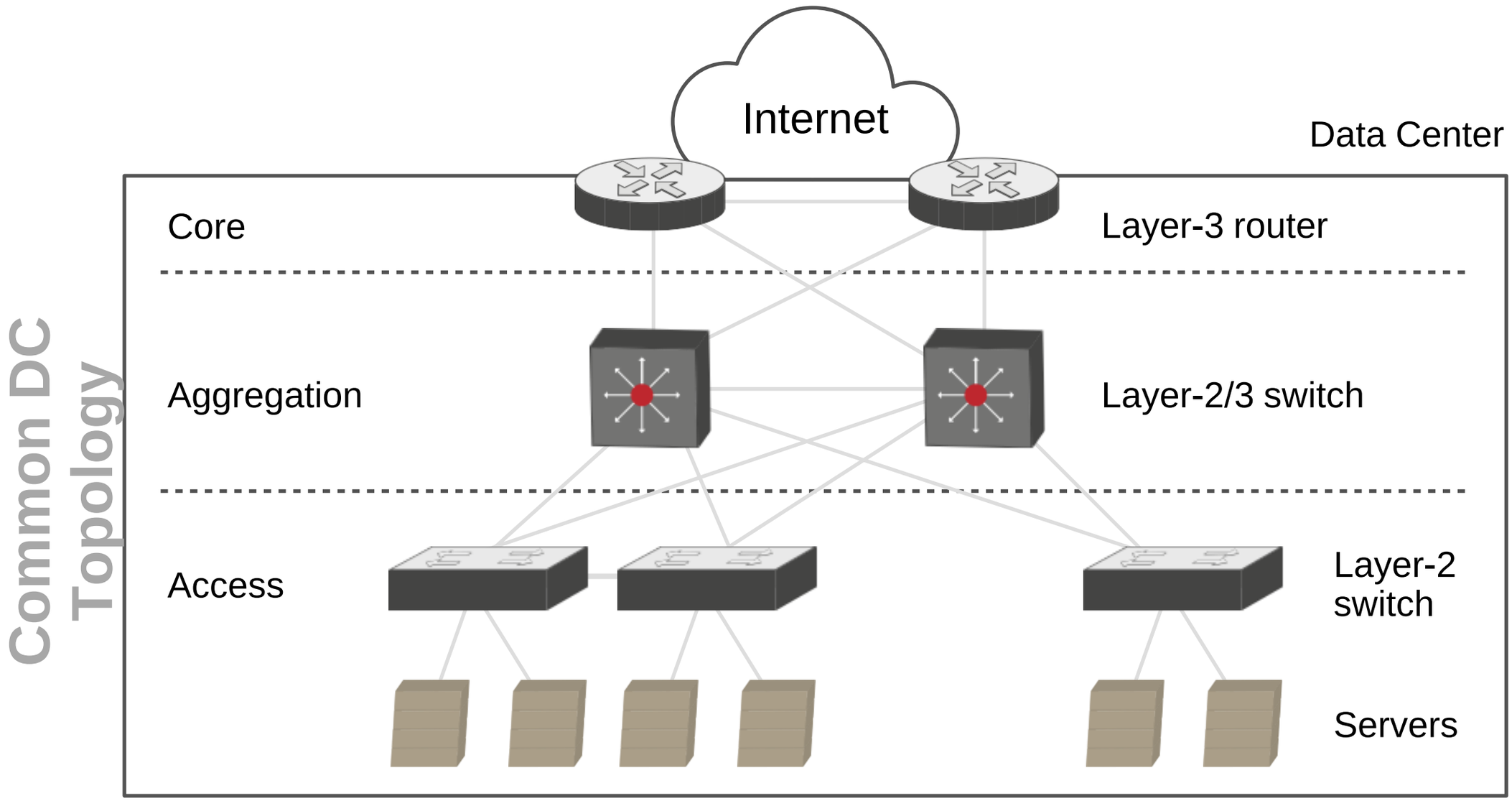
- EIGRP doesn't work over GETVPN and LISP, because both are tunnelless VPN mechanisms, routing protocols can be an underlay for them but not an overlay.

Overlay Technologies and EIGRP (GRE, MGRE, DMVPN, GETVPN, LISP)

- EIGRP is one of the best mechanisms which work over DMVPN in large scale designs.
  - EIGRP with GRE is not scalable for large scale deployment but scaling limitation comes from GRE, it is not the EIGRP problem, MGRE provides scalability with EIGRP even in large scale deployment.
- EIGRP can work as a PE-CE protocol in MPLS L3 VPN deployment which we will talk in the MPLS lesson.

EIGRP in the Datacenter

- EIGRP can be used at the DC edge to advertise DC prefixes to the WAN and Campus network.



EIGRP in the Datacenter

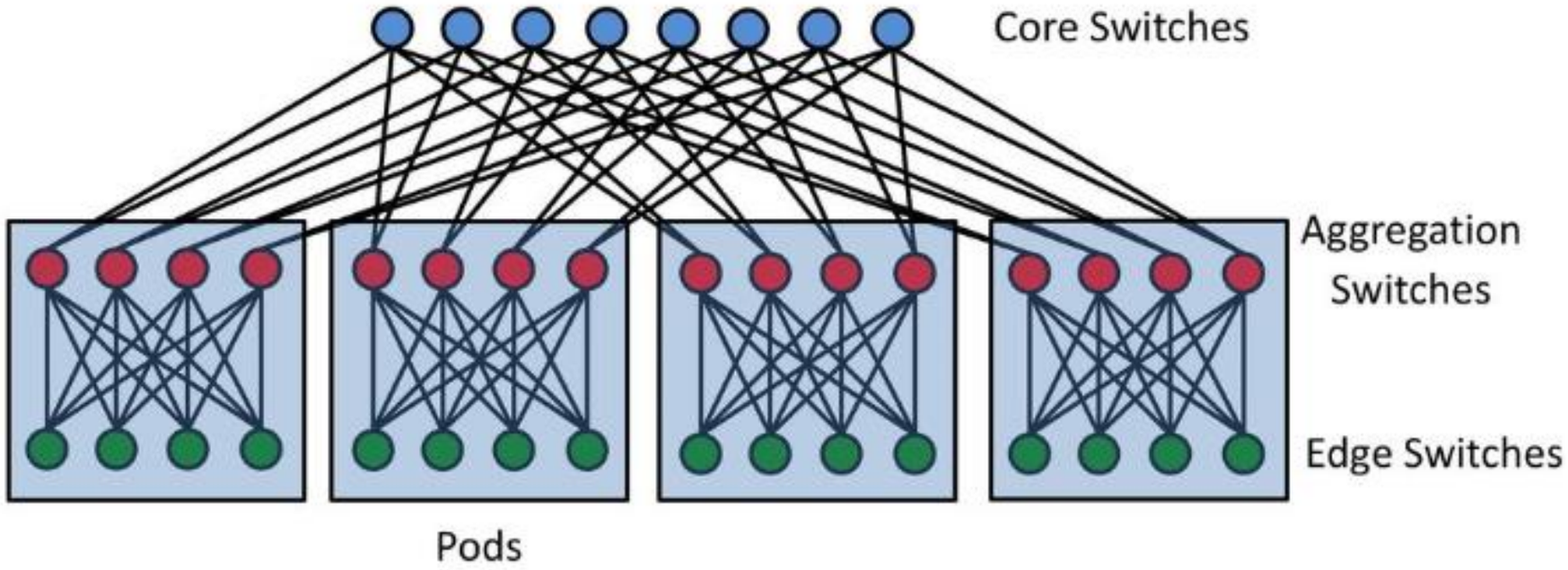
- Also EIGRP can be used as a Datacenter Fabric Protocol, but this is not common.

- Datacenters are very densely connected networks, from Core towards access, summarization would be highly necessary as Access switches (TOR in CLOS topology) don’t need to have full routing table and they cannot due to resource limit.

EIGRP in the Datacenter

- Large scale Datacenters mainly use CLOS (Leaf and Spine) topology, depends on scale, multi stage CLOS topologies are used

3 stage CLOS topology



- To create DC Fabric, EIGRP can be used in some Datacenters as a layer 3 fabric protocol, but as it is stated before, EIGRP is not common in the Datacenters as a L3 fabric protocol.

- EIGRP is used mainly on the WAN, classical example EIGRP over DMVPN.

EIGRP in the Service Provider Networks

- EIGRP is very uncommon in the Service Provider networks.
  - Due to lack of MPLS Traffic Engineering Support and multi-vendor interoperability issues, Service Providers use Standard based protocols.
- Service Providers are heterogeneous environment (Many vendors are used in the network) interoperability between the network equipments is critical.

EIGRP Design Best Practices

- Enable summarization in large scale deployment, it will provide scalability.

- Limit the EIGRP query scope, it will help for fast convergence, availability, troubleshooting and scalability.

EIGRP Design Best Practices

- Enable EIGRP Stub feature at the Spokes site in Hub and Spoke deployment, EIGRP stub routers will not be used as a transit node, it will help for fast convergence and scalability as well.

EIGRP Design Best Practices

- When you need to do traffic engineering, change delay attribute, not bandwidth, delay is cumulative, changing delay mat not have an impact for the other paths, also bandwidth is used for QoS, RSVP and other applications as well.

EIGRP Design Best Practices

- Don’t lower the EIGRP fast hellos for faster failure detection, use BFD instead.

EIGRP Advantages and the Disadvantages

•EIGRP is considered more flexible than OSPF and IS-IS because it supports manual summarization in any interface at any router in the network.

disa
dva
ntag
es

- EIGRP, in addition to ECMP, support UCMP as well.

adva
ntag
es

EIGRP Advantages and the Disadvantages

- EIGRP supports Stub feature which protects Spoke routers to be a transit router, this is not easily possible with OSPF.

disa
dva
ntag
es

- EIGRP is better suitable in large scale Hub and Spoke design than OSPF and IS-IS.

adva
ntag
es

EIGRP Advantages and the Disadvantages

Disadvantages

- There is no topology information, end to end path may not be optimal.
- EIGRP is Cisco preparatory, vendor interoperability may be an issue.

Advantages

- With Feasible Successor support, it converges faster than OSPF and IS-IS when there is no ECMP with them.

EIGRP Advantages and the Disadvantages

•There is no topology information, EIGRP cannot be used in distributed traffic engineering.

disadvantages

- Well known protocol by many network engineers and commonly deployed in many Enterprise networks.

advantages

EIGRP CASE STUDIES

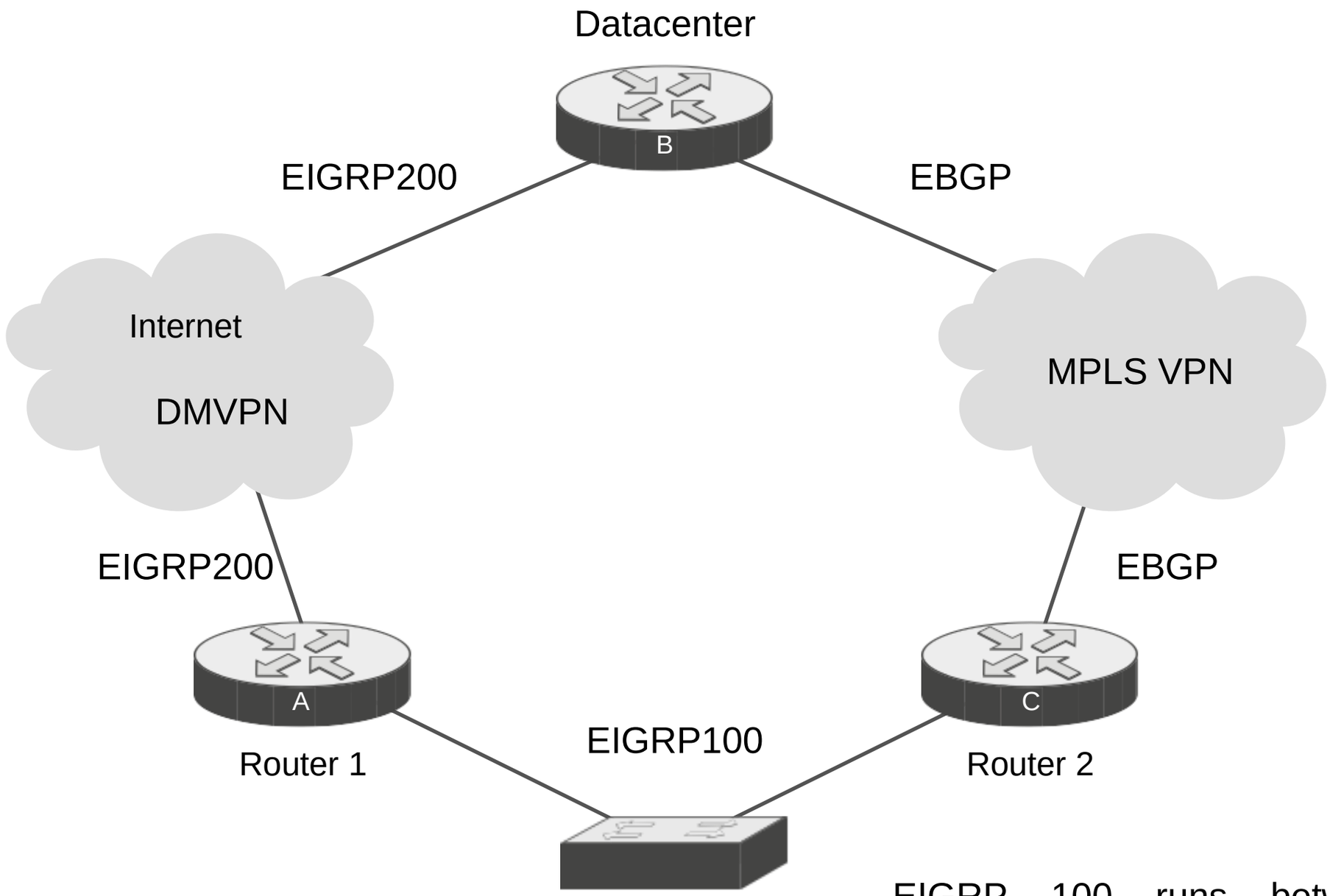
MPLS VPN and DMVPN with EIGRP

- Enterprise company wants to use MPLS L3 VPN (Right one) as its primary path between their Remote office and the Datacenter.
  - Customer uses EIGRP and EIGRP AS 100 for the Local Area Network inside the office.
- They want to use their DMVPN network as a backup path.
  - Customer runs EIGRP AS 200 over DMVPN.

MPLS VPN and DMVPN with EIGRP

- Service Provider doesn't support EIGRP as a PE-CE protocol, only static routing and BGP.
  - Customer selected to use BGP instead of static routing since cost community attribute might be used to carry EIGRP metric over the MP- BGP session of service provider.
- Redistribution is needed on the R2 between EIGRP and BGP (Two ways).
  - Since customer uses different EIGRP AS numbers for the LAN and DMVPN networks, redistribution is need on R1 too.

Network Topology



EIGRP 100 runs between switch and remote office routers

Question 1

- Should customer use same EIGRP AS (Autonomous System) on the DMVPN network and the their office LAN? What is the problem with that design?

Answer 1: No. They shouldn't.

- Since Customer requirement is to use MPLS VPN as primary path, if the customer runs same EIGRP AS on Local Area Network and over DMVPN, EIGRP routes is seen as internal from DMVPN but external from MPLS VPN.
  - Internal EIGRP is preferred over external because of Admin Distance, customer should use different AS numbers.
- DMVPN could be used as a primary path for some delay, jitter, loss insensitive traffic but customer didn't specify that.

Question 2

- When company changed the EIGRP AS on DMVPN and started to use different EIGRP AS on the DMVPN and the LAN, which path is used between the Remote Offices and the datacenter?

Answer 2

- Since redistribution is done on R1 and R2, remote switch and datacenter devices see the routes both from DMVPN and BGP as EIGRP external. Then the metric is compared. If the metric (Bandwidth and Delay in EIGRP) is the same, both path can be used (Equal Cost Multipath-ECMP).

Question 3

- Does result fits for the customer traffic requirement?

Answer 3

- You should first remember what was the customer expectation for the links, they want to use MPLS VPN for all their application as primary path.

  - That's why answer is Yes, it satisfy the customer requirement. Because if customer uses different EIGRP AS on LAN and DMVPN, with just metric adjustment, MPLS VPN path can be used as primary with just the metric arrangement.

Question 4

- What happens when the primary MPLS VPN link goes down?

Answer 4

•Traffic from remote office to the datacenter goes through Switch - R1 - DMVPN path. From the datacenter, since those will not be known through MPLS VPN when it fails, only DMVPN link is used.

- So DMVPN link is used as primary when the failure happens.

Question 5

- What happens when failed MPLS VPN link comes back?

Answer 5

- R2 receives the datacenter prefixes over MPLS VPN path via EBGP, also from R1 via EIGRP. Once the link comes back, datacenter prefixes will still be received via DMVPN and MPLS VPN and appears on the office switch as an EIGRP external.

- Since metric was arranged previously to make MPLS VPN path as primary, no further action is required.

Answer 5

- This is tricky part. If it is Cisco switches or from other vendor which uses BGP weight attribute into consideration for the best path selection, then redistributed prefixes weight would be higher than the prefixes which are received through MPLS VPN so R2 uses Switch-R1 DMVPN path which violates the customer expectation.

EIGRP Stub and Caveats

- Haleo is a fictitious Enterprise company which uses DMVPN Wide Area Network technology over the Internet.
  - They have only 1 datacenter and the WAN routers which connect remote branches to the datacenter and the campus network are terminating many remote branches.
- For the redundancy and capacity purpose Haleo is using two DMVPN hub routers.
  - They know that EIGRP is best IGP for the large scale DMVPN design but recently they had some issues with their WAN network.

Question 1

•Which additional information do you need from Haleo to address their issue?

1. Datacenter network topology
2. Which routing protocol Haleo is using
3. Encryption method of Haleo
4. Routing configuration
5. Wan Network Topology

Answer 1

•Lets look at the options

- Datacenter Network topology: We don’t need this information since in the background information, we are told that problem on the Wan network.

•Which routing protocol Haleo is using: We don’t need it as well since in the background information it is already stated as EIGRP.

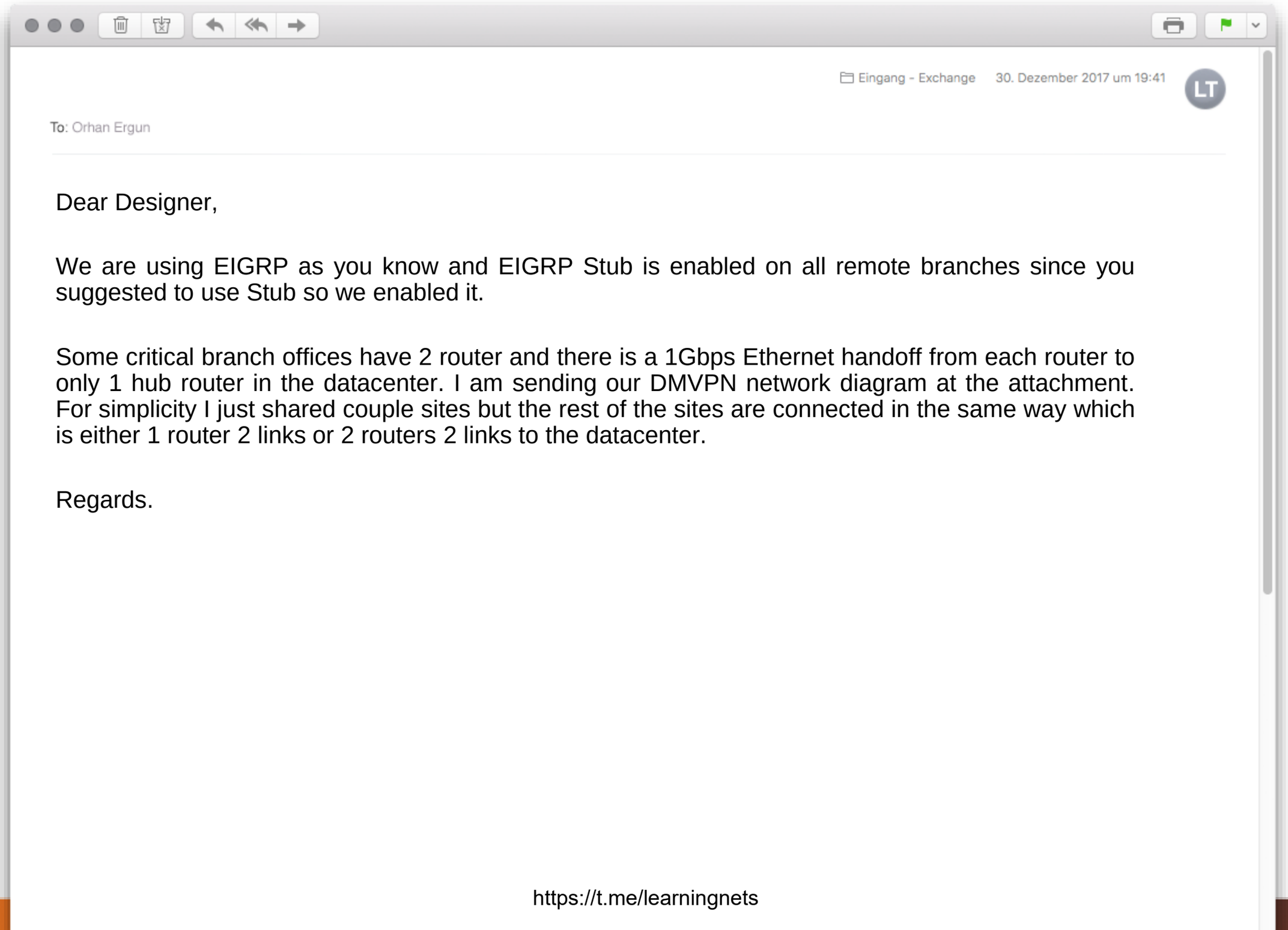
- Encryption method: Nothing said about encryption, we can’t assume that they are using encryption over DMVPN network since DMVPN doesn’t require encryption.

Answer 1

- Routing configuration: We know that Haleo is using EIGRP but we don’t know if they have EIGRP Stub on the edges , whether split horizon is disabled on the Hub, so we need to know their EIGRP features.

- Wan network topology: We need Wan network topology since we don’t know how many routers in the branches, how they are connected, background information only mention that Haleo is using two hub routers.

Answer 1



Eingang - Exchange 30. Dezember 2017 um 19:41

LT

To: Orhan Ergun

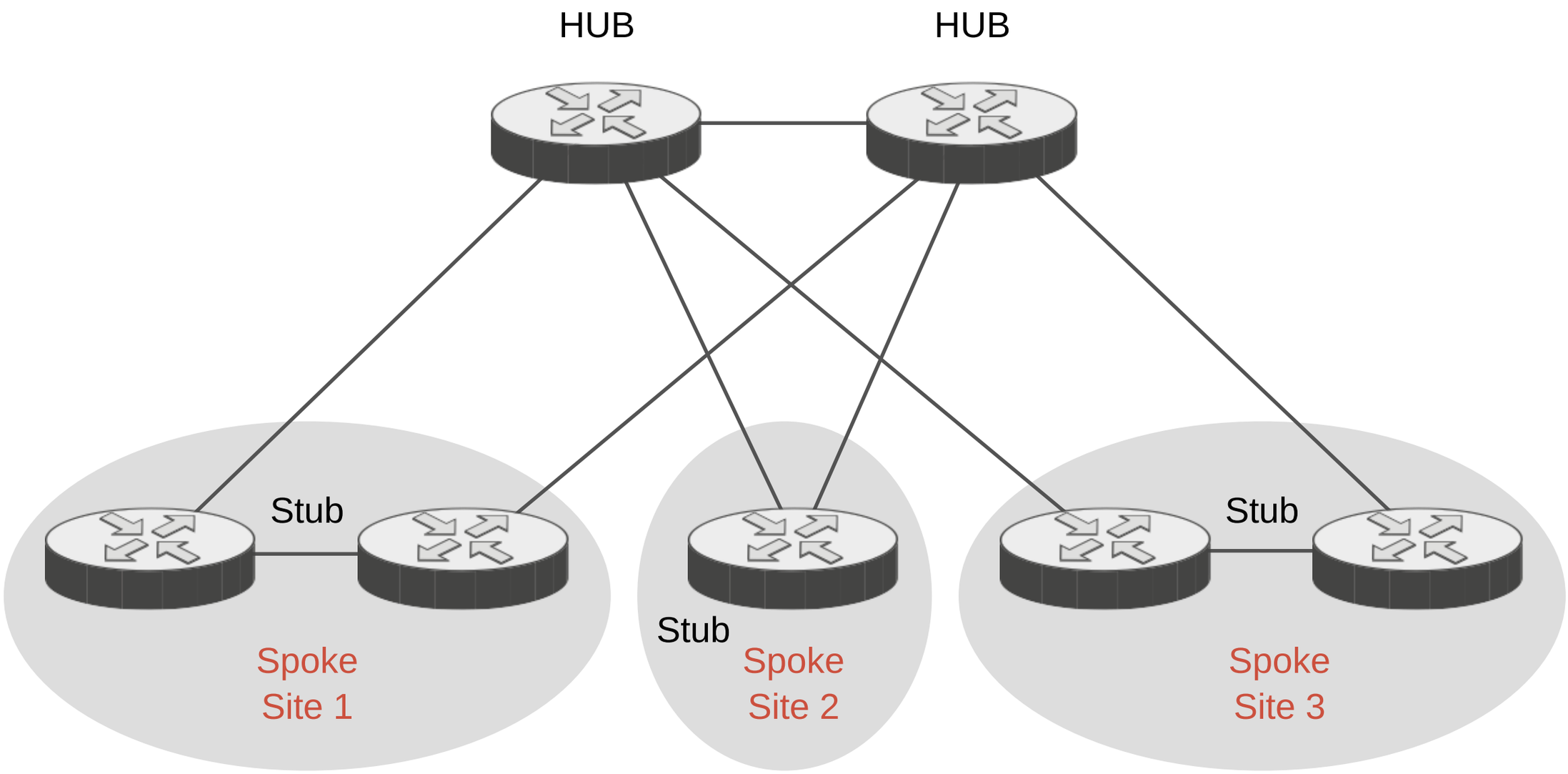
Dear Designer,

We are using EIGRP as you know and EIGRP Stub is enabled on all remote branches since you suggested to use Stub so we enabled it.

Some critical branch offices have 2 router and there is a 1Gbps Ethernet handoff from each router to only 1 hub router in the datacenter. I am sending our DMVPN network diagram at the attachment. For simplicity I just shared couple sites but the rest of the sites are connected in the same way which is either 1 router 2 links or 2 routers 2 links to the datacenter.

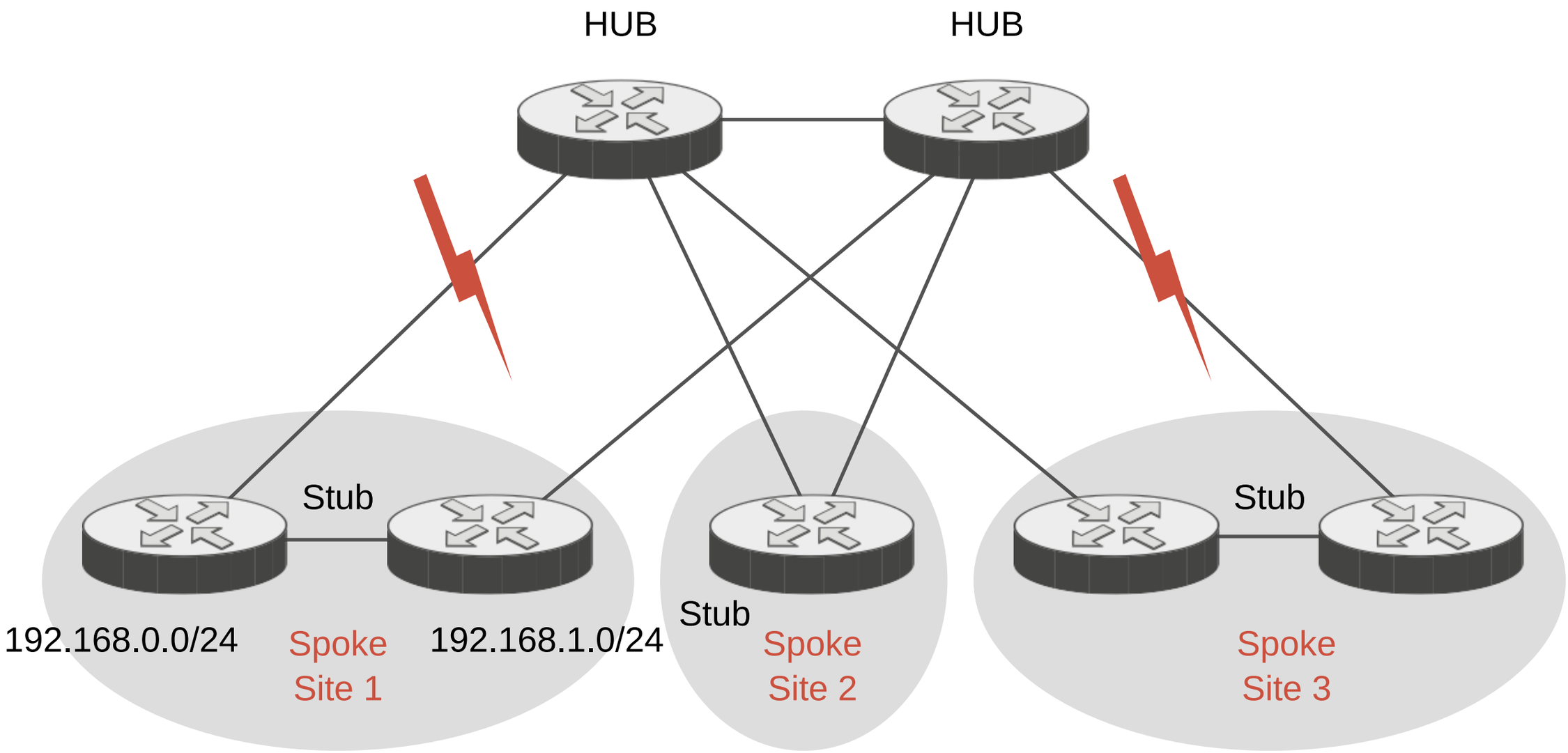
Regards.

Customer Topology



Question 2

- Based on the provided information what might be the problem? How it can be solved?



Answer 2

- Since the spoke routers are running as EIGRP stub, they don't send the prefixes which are learned from each other to the Hubs.
  - That's why If the link between Hub and the Spoke sites which have two routers fails, router is isolated from the rest of the network.
- Spokes in the Spoke site 1 send their network to each other. So 192.168.0.0/24 and 192.168.1.0/24 learned by both spokes but since they are EIGRP Stub, they don't send the learned routes to the hub.

Answer 2

- If the Hub and Spoke link is failed in the Spoke Site 1, 192.168.0.0/24 network will not be reachable anymore.
  - Same think for the Spoke Site 3 since that site also has 2 routers and EIGRP stub is enabled.
- The solution is to enable EIGRP Stub leaking. In DMVPN it is good to send summary or default route to the Spokes by the Hubs.
  - Spoke should send the routes which they learn from each other to the Hub and also should send the routes which they learn from Hub to the each other. In this way sites which have more than 1 router which has EIGRP Stub configuration doesn't have an issue in case of any failure.

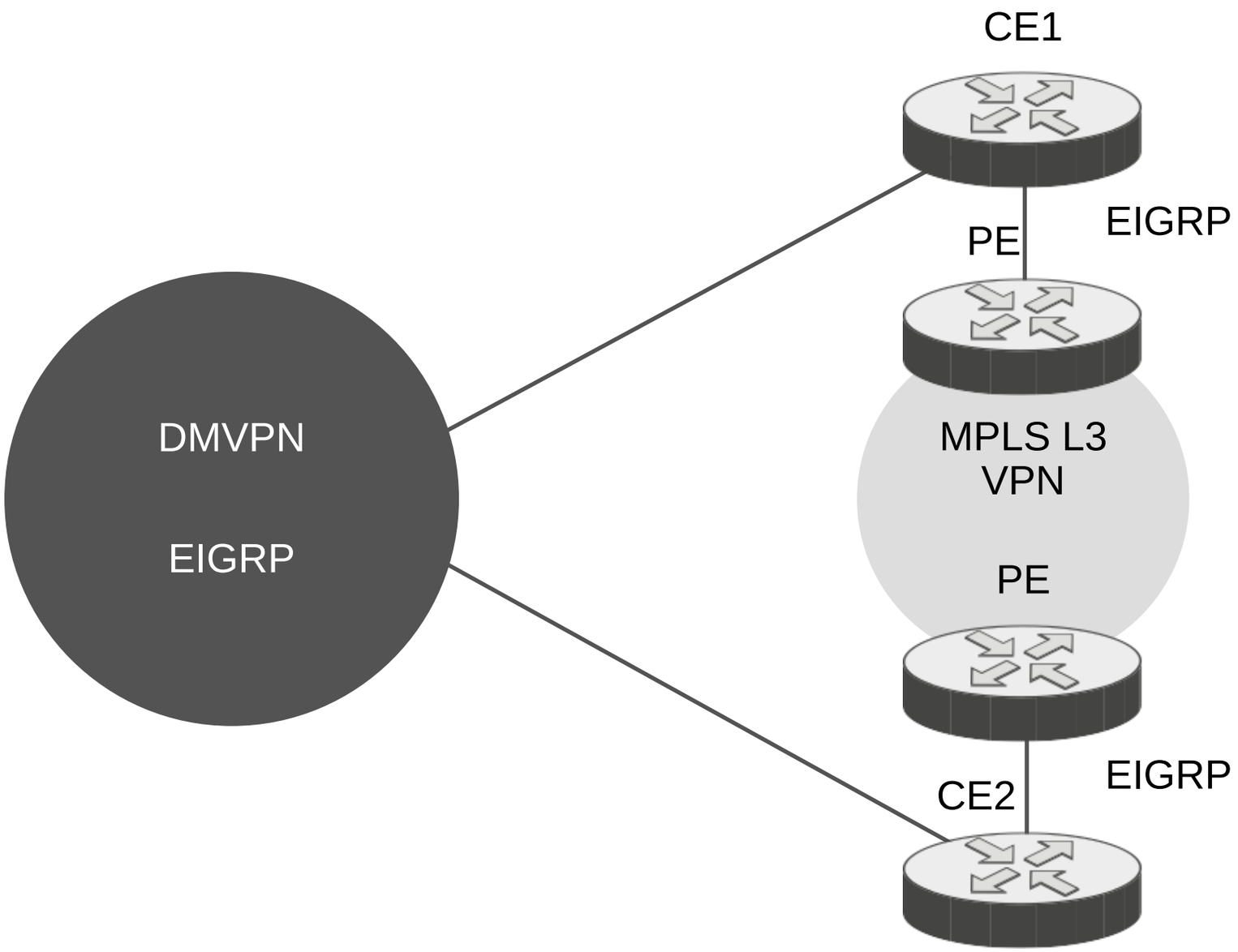
Race Condition and Routing Loop

- Enterprise company is using EIGRP on their network. They have MPLS VPN and also DMVPN network for their WAN network.
  - MPLS service is layer 3 VPN, managed by the Service Provider.
- Company has 30ms end to end latency SLA for their MPLS VPN network and all their application's fits to his service level agreement.
  - That's why they want to use MPLS VPN for their primary path and in case it fails, DMVPN should be used as backup.

Race Condition and Routing Loop

- Recently they had an outage and when they researched, they realized that the convergence of their EIGRP was too higher than their expectation.
  - They are asking your help to address the issue and provide a solution.
- They send their topology at the attachment.

Network Topology

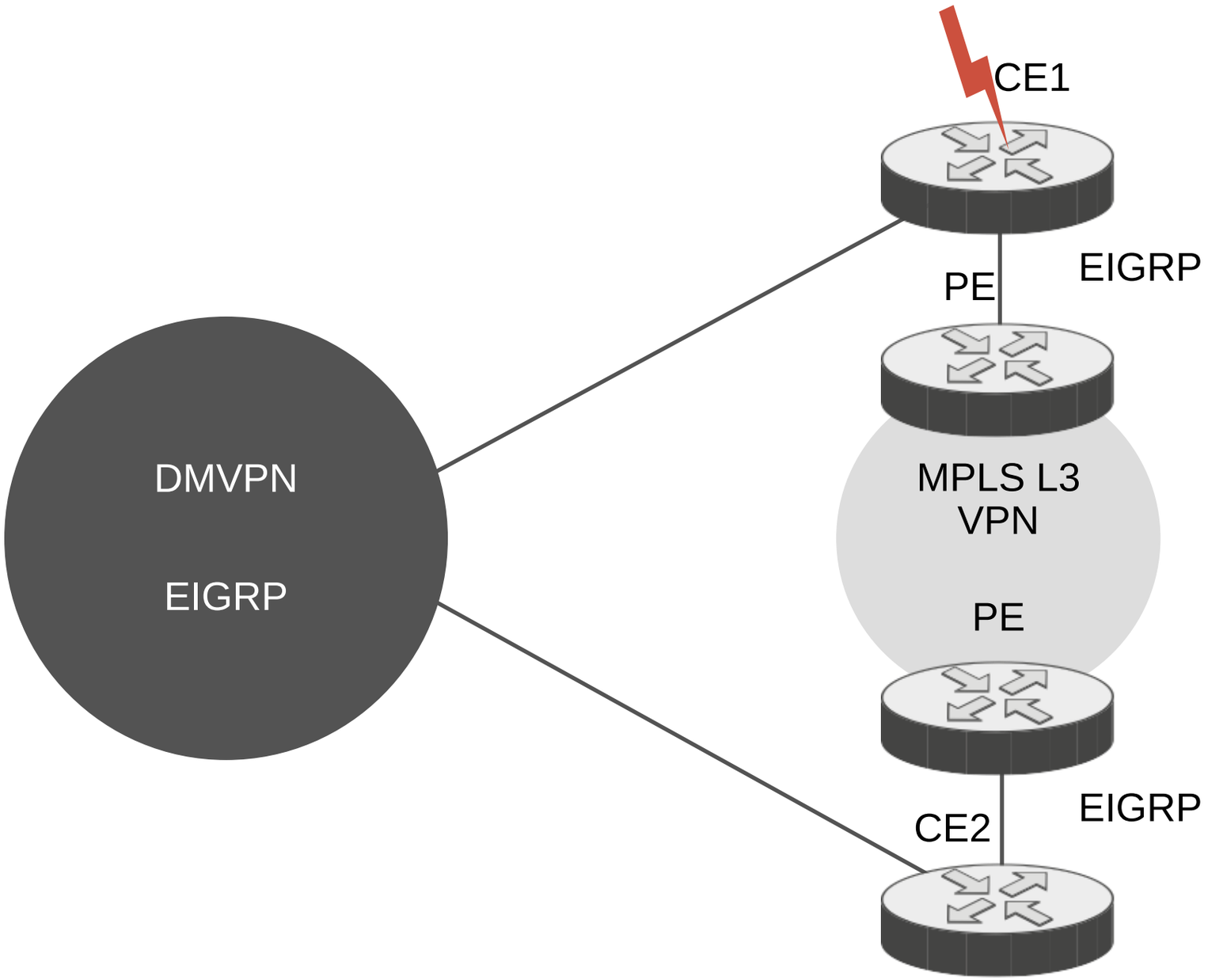


Race Condition and Routing Loop

•EIGRP because it is a distance vector protocol, suffers from two problems.

- In case of failure, you will see either suboptimal routing or black hole situation, both due to the EIGRP race condition.

•In the topology below if the loopback interface of CE1 fails, depending on the timing of EIGRP and MP-BGP update, you might have routing loop.



- CE1 sends an EIGRP queries to PE1 and CE2 asking it's loopback prefix. Network as we know from the scenario is using MPLS VPN as primary path.
  - That's why PE1 prefers the prefix via CE1, since it learns the prefix from the CE1 it assumes there are no alternate path.
- PE 1 doesn't send queries further and reply with prefix unreachable to CE1.
  - When PE1 stops learning CE'1 loopbacks from EIGRP, it removes it from its BGP table as well (EIGRP to MP-BGP redistribution).

•When PE1 removes the prefix from its BGP routing table it sends a BGP withdrawal to the PE2.

- PE1 sent previously an EIGRP query to the CE2, and CE2 would propagate it.

•Here the EIGRP Race condition can be a big issue.

- Depending on the arrival of EIGRP query and the BGP Withdrawal message to the BGP, persistent routing loop occurs.

- If BGP withdrawal over the MPLS VPN backbone via MP-BGP comes faster than EIGRP query through CE1 - CE2 and eventually to the PE2, everything is fine.

  - Because PE2 removes the prefix from BGP table and stops redistributing into the EIGRP (MP-BGP to EIGRP Redistribution).

- But if EIGRP query from the CE2 comes faster than BGP withdrawal to the PE2.

- Since PE2 has a path to the CE1 over the MPLS VPN, it replies to the CE2 that It has an alternate path, this reply goes up to the originator which is CE1 back and CE1 sends it to the PE1, in return PE1 sends it to the PE2.

  - CE1 now thinks that CE2 has an alternate path for its own loopback.

- Meanwhile PE2 receives the BGP withdraw which was previously sent by PE1.

  - PE2 sends an EIGRP query to CE2 so CE2 again things that it is not reachable so it sends a EIGRP query to CE1.

EIGRP in the CCDE Exam

- EIGRP and VPN interaction, such as EIGRP, MPLS L3 VPN, DMVPN case study.
  - EIGRP and other routing protocols interaction, redistribution, extending EIGRP to the OSPF or IS-IS domains or vice-versa
- EIGRP Stub and other scalability features.

Summary

- EIGRP Theory – Distance Vector, No-topology, terminology
- EIGRP Fast convergence and Fast Reroute
- EIGRP Scalability – Query Scope, Summary, Filter and Stub
- EIGRP in the Datacenter and Service Provider Networks
- EIGRP Advantages – Disadvantages – Summarization at every hop, scalability, Cisco preparatory

EIGRP

Enhanced Interior Gateway Routing Protocol

QUIZ

Question 1

Which below technology provides similar functionality with EIGRP Feasible Successor?

A. ISPF

B. Partial SPF

C. Loop Free Alternate Fast Reroute

D. OSPF Stub Areas

E. IS-IS Level 1 Domain

Answer 1

C. Loop Free Alternate Fast Reroute

Although EIGRP convergence was not explained in the EIGRP chapter, it is important to mention here. EIGRP Feasible successor is the backup path, which satisfies the feasibility condition.

Which means it should satisfy the EIGRP's loop free backup path condition.

There is no ISPF, Partial SPF, PRC or SPF in EIGRP. These algorithms are used in link state protocols.

Answer of this question is LFA FRR, which is one of the IP Fast Reroute mechanisms. IP FRR mechanisms will be explained in the MPLS traffic engineering section later in the book.

Question 2

How many levels of hierarchy is supported in EIGRP?

A. One

B. Two

C. Three

D. Unlimited

Answer 2

D.Unlimited

Unlike OSPF and IS-IS, there is no limit in EIGRP. OSPF and IS-IS support two levels of hierarchy as it was explained earlier.

There is no topology information in EIGRP and summarization can be done anywhere in EIGRP. Unlimited level of hierarchy is possible with EIGRP, that’s why answer of this question is D.

Question 3

In the below topology R3 is configured as EIGRP Stub. If the link between R1 and R2 fails, which below statements are true for the below topology? (Choose Two)

A. Since R3 is configured as EIGRP Stub, R3 is not used as transit for the traffic between R1 and R2.

B. R1 can reach the networks behind R2 through R3.

C. R3 has all the R1 and R2 prefixes in its routing table.

D. R3 doesn’t have R1 and R2 prefixes.

E. R2 can reach the networks behind R1 through R3.

Answer 3

A. Since R3 is configured as EIGRP Stub, R3 is not used as transit for the traffic between R1 and R2.

C. R3 has all the R1 and R2 prefixes in its routing table.

As it was explained in the EIGRP Stub section in this chapter, when node is configured as EIGRP stub, it is not used as transit node anymore.

Question is asking when the R1 and R2 link fails, whether R3 will be transit node. No, it will not be transit node.

Which mean, R1 cannot reach R2 through R3 and R2 cannot reach R1 through R3. That’s why B and E are incorrect.

R3 has all the prefixes of R1 and R2 even if it is configured as EIGRP Stub.

That’s why the correct answer of this question is A and C.

Question 4

Which below option is considered as loop free path in EIGRP?

A. If reported distance is less than feasible distance.

B. If reported distance is same as the feasible distance.

C. If reported distance is higher than feasible distance.

D. If administrative distance is higher than feasible distance.

Answer 4

A. If reported distance is less than feasible distance.

In order a path to be chosen as loop free alternate which means satisfy the EIGRP feasibility condition as it was explain in the EIGRP chapter of the book, reported distance has to be less than feasible distance. That’s why the answer of the question is A.

Question 5

What happens if the backup path satisfies the feasibility condition? (Choose Two)

A. It is placed in link state database.

B. It is advertised to the neighbors.

C. It is placed in the topology table.

D. It can be used as unequal cost path.

E. It is placed in the routing table.

Answer 5

C. It is placed in the topology table.

D. It can be used as unequal cost path.

EIGRP database is called Topology database. Link state database is used in link state protocols.

If backup path satisfies feasibility condition, it is placed in topology table, not in routing table. If it would be best path (successor) or equal cost path, it would be placed in routing table. But since question says, backup path, it is only placed in EIGRP topology database.

Since it is not the best path, it is not advertised to the neighbors.

With ‘variance’ command, it can be used as unequal cost path and can be placed in the routing table.

Question 6

Which below statements are true for EIGRP Summarization? (Choose Two)

A. EIGRP Auto-summarization is on by default for all the Internal and External routes.

B. EIGRP Route summarization can reduce the query domain which helps for convergence.

C. EIGRP Route Summarization can reduce the query domain which can prevent Stuck in Active problem.

D. Summarization cannot be done at each hop in EIGRP.

Answer 6

B. EIGRP Route summarization can reduce the query domain which helps for convergence.

C. EIGRP Route Summarization can reduce the query domain which can prevent Stuck in Active problem.

Summarization can be done at each hop in EIGRP. This is different than OSPF and IS-IS. Auto-Summarization is not enabled for all the routes by default in EIGRP. Summarization helps to reduce query domain boundary, which in turn help for convergence, SIA problem, troubleshooting and so on.

Question 7

Which below statement is true for EIGRP queries? (Choose Two)

A. EIGRP queries always send.

B. Limiting EIGRP query domain helps for scalability.

C. If summarization is configured, EIGRP query is not sent.

D. If filtering is configured, EIGRP query is not sent.

E. If EIGRP Stub is configured, EIGRP query is not sent.

Answer 7

B. Limiting EIGRP query domain helps for scalability.

E. If EIGRP Stub is configured, EIGRP query is not sent.

If EIGRP Stub is configured, as it was explained before, EIGRP query is not sent. With summarization and filtering still EIGRP query is sent. EIGRP query domain size affects scalability. If the query domain size is reduced, scalability increases.

Question 8

Why passive interface should be enabled on the access/customer ports?

A. To prevent injecting the customer prefixes to the network.

B. To reduce the size of the routing table.

C. For the fast convergence.

D. For higher availability.

Answer 8

A. To prevent injecting the customer prefixes to the network.

Passive interface should be used on all hosts, access and customer ports. Otherwise security attack can happen and prefixes can be injected into the routing domain. It doesn't provide faster convergence. And the reason to disable routing protocols on the customer/access ports is not to reduce routing table size.

Question 9

If the path in the network will be changed by changing the EIGRP attribute, which below statement would you recommend as a network designer?

A. Bandwidth should be changed.

B. Delay should be changed.

C. Reliability should be changed.

D. PBR should be configured.

Answer 9

B. Delay should be changed.

PBR is not an EIGRP attribute. Reliability is not used for EIGRP path selection. Bandwidth and Delay attributes are used for EIGRP path selection and metric is calculated based on these two parameters.

But, since bandwidth can be used by many applications such as QoS, RSVP-TE and so on it should be changed, otherwise other things in the network can change too.

Also since the minimum bandwidth is used for path calculation, changing bandwidth can affect entire network design. Not only the path, which we want.

On the other hand, delay is additive and changing it can only affect the path, which we want.

Question 10

When EIGRP is used as MPLS VPN PE-CE routing protocol, which below mechanism helps for loop prevention even if there is a backdoor link?

A. Up/Down bit

B. Sham link

C. Site of Origin

D. Split Horizon

Answer 10

C. Site of Origin

EIGRP Site of Origin is used to prevent loop even if there is a backdoor link. Backdoor link causes race condition in MPLS VPN topologies and it can create sub optimal routing and routing loop.

It will be explained in the MPLS VPN section in the MPLS chapter in detail.

EIGRP – Study Resources:

- Books :
- http://www.amazon.com/EIGRP-Network-Design-Solutions-Definitive/dp/1578701651/ref=sr_1_1?ie=UTF8&qid=1436565482&sr=8-1&keywords=eigrp
- Videos :
- Ciscolive Session – BRKRST -2336
- Podcast :
- http://packetpushers.net/show-144-open-eigrp-with-russ-white-ciscos-donnie-savage/
- Articles :
- http://www.cisco.com/c/en/us/td/docs/ios/12_0s/feature/guide/eigrpstb.html
- http://www.cisco.com/c/en/us/td/docs/ios xml/ios/iproute_eigrp/configuration/xe-3s/ire-xe-3s-book/ire- ipfrr.html

IPv6

- Business Drivers
- Transition Mechanisms
- Key Points in IPv6/Summary

IPv6 Business Drivers

- **IPv4 address exhaustion**
- **Business Continuity (E-Commerce)**
- **Easier Network Mergers and Acquisitions (No overlap, NAT etc.)**
- **Government IT Strategy and Regulations (Government mandates, if Private companies want to work with them)**
- **Infrastructure Evolution**

IPv6 Transition Mechanisms

- If the underlay transport is MPLS; best methods are 6PE and 6VPE
- If the underlay transport is IP; dual stack, tunneling and translation are the options
- Depends on the company, their customer requirements and many other factors, one method might be better than other.
- Is Really Dual Stack best deployment method ?

- The only available public IP addresses are IPv6 addresses. But vast majority of the content is still working on IPv4. How IPv6 users can connect to the IPv4 world and How IPv4 users can reach to the IPv6 content. This is accomplished with the IPv6 transition technologies.

- Probably the IPv6 transition technologies is a misleading term. Because; IPv4 infrastructure is not removed with these technologies. Thus probably the IPv6 integration technologies is a better term.

- But still throughout this section I will be using IPv6 transition technologies.

- There are three types of IPv6 Transition Technology. 1. Dual Stack
- • IPv6 + IPv4
  The entire infrastructure is running both IPv4 and IPv6.
- 2. Tunnels
- IPV6 - IPv4 – IPv6
- IPv4 – IPv6 – IPv4
- Two IPv6 islands communicate over IPv4 part of the network or two IPv4 islands communicate over IPv6 part of the network.
- 3. Translation
- IPv6 – IPv4 (NAT64)
- Private IPv4 – Public IPv4 (NAT44)
- With translation, IPv6 only device can communicate with IPv4 only device. But they think that they communicate with the same version of device.

Key points in IPv6

- You don’t have to deploy IPv6 everywhere from day 1
- Assessment and planning is key for IPv6 design
- Stateful IPv6 translation mechanisms have many challenges such as asymmetric routing, logging issues, single point of failure and so on
- Dual stack still requires IPv4 address on the CPE ! It is against to IPv4 exhaustion issue

- You can start core to edge (You have a time), Edge to core (rely on tunneling) or Internet edge (e-commerce)

- 6PE and 6VPE is best transition mechanisms for the MPLS networks

- Running IPv6 together with IPv4 doesn’t create a problem for IPv4 infrastructure but still memory and CPU of the devices need to be tracked.

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 Multipath, IBGP Multipath and EIBGP Multipath
- Inter domain routing – Settlement Free Peering, IP Transit
- ISP Tiers – Tier 1 , 2 , 3 Type Providers
- IBGP – Basics of IBGP
- BGP Route Reflectors
- Route Reflector Design Options
- BGP Optimal Route Reflection – ISP Case Study
- BGP Confederations
- Full mesh IBGP vs. Route Reflector vs. Confederation
- Full mesh to RR Migration
- RR to Confederation Migration

BGP Course Outline

- BGP – IGP Interactions – Blackhole avoidance
- BGP – MPLS Interactions
- BGP LU – Labeled Unicast
- BGP LS – BGP Link State
- BGP Segment Routing
- BGP EPE – Egress Peer Engineering
- BGP Flowspec
- BGP Session Culling
- AIBGP – Accumulated IBGP
- BGP Selective Blackholing
- BGP Route Propagation Behavior – RFC 8212
- BGP Security - Route Leak, Hijacking, RPKI, Origin and Path Validation

BGP Course Outline

- BGP in the Datacenter
- BGP in the WAN
- BGP PIC – Prefix Independent Convergence
- Case Studies
- BGP vs. IGP Comparison
- BGP in the CCDE exam
- Summary
- Bonus Materials

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

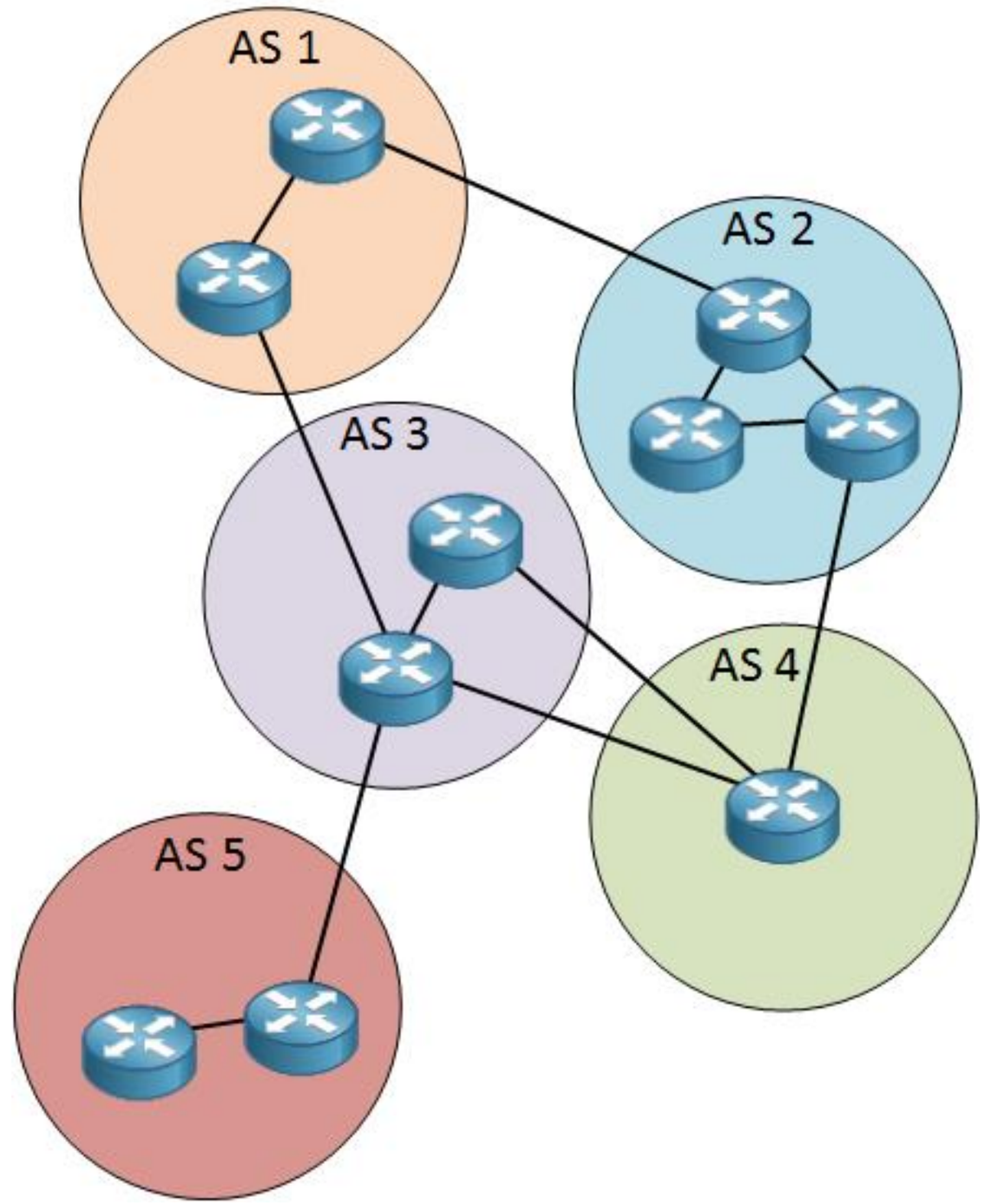
- BGP is a multi protocol , with the new NLRI it can carry many address families. Today almost a 20 different NLRI is carried over BGP. New AFI, SAFI is defined for the new address families (https://orhanergun.net/tag/multi-protocol-bgp/)

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



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 protocols 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

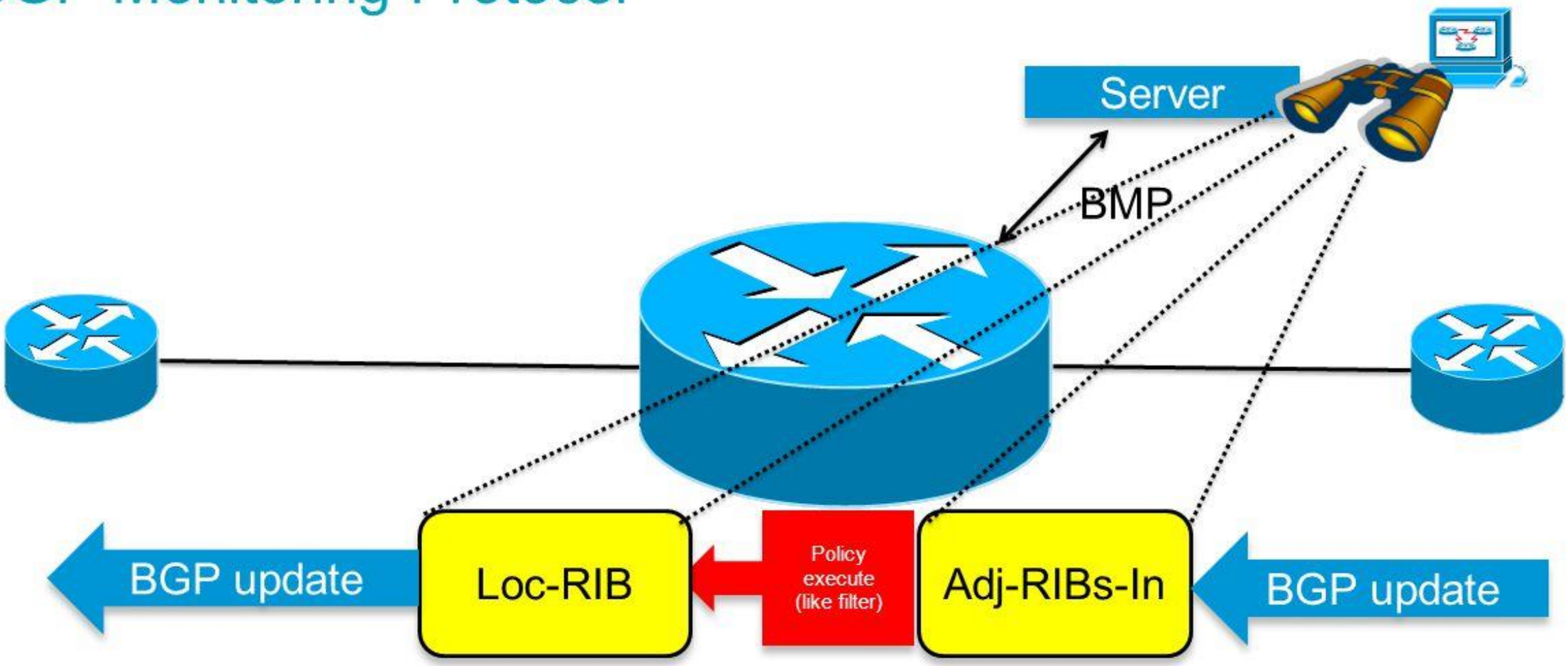
BMP - BGP Monitoring Protocol

- 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



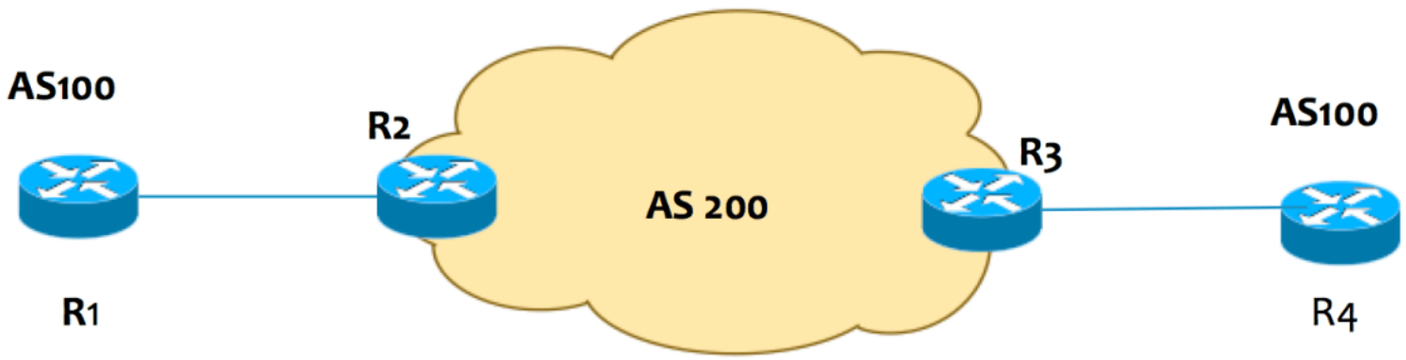
BMP - BGP Monitoring Protocol

- 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

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

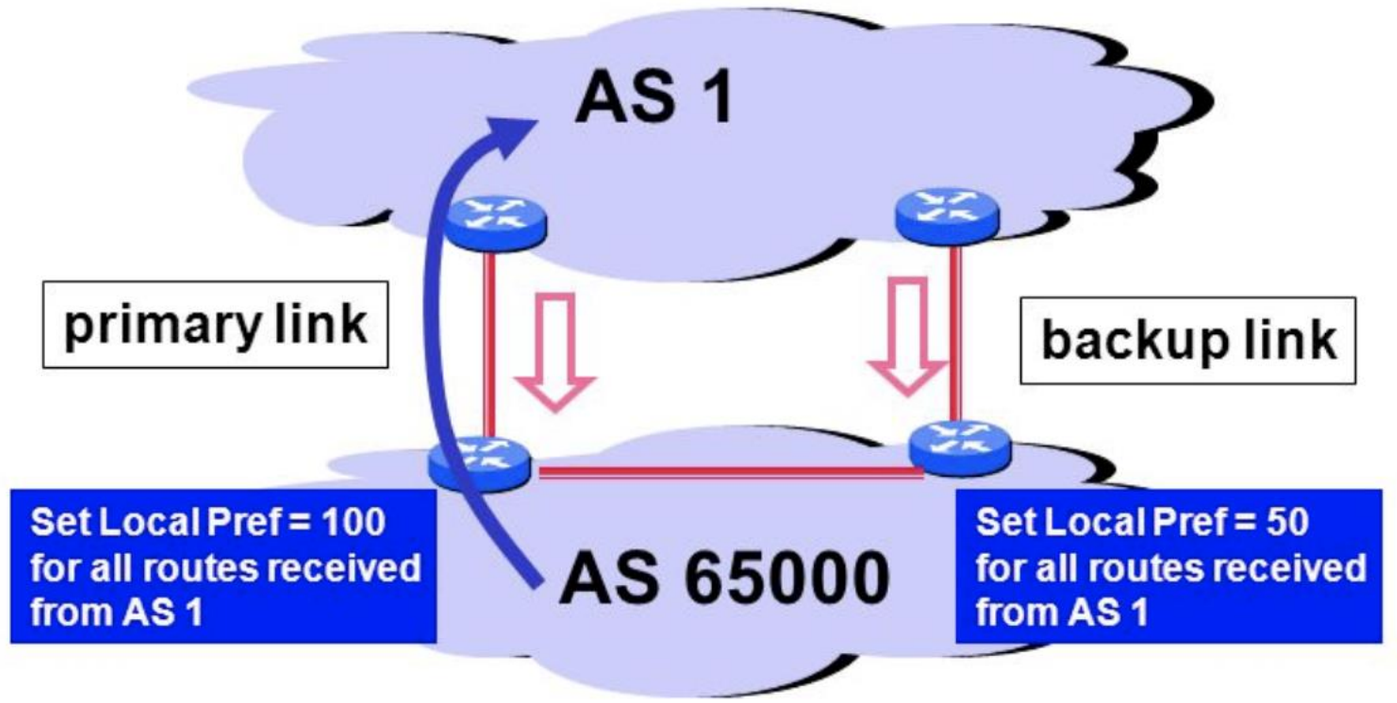
- For example link capacities might be different, one link might be more stable than the other, or monetary costs of the links might be different

- 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
- 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

Don’t use BGP Weight Attribute !

- 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

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 !

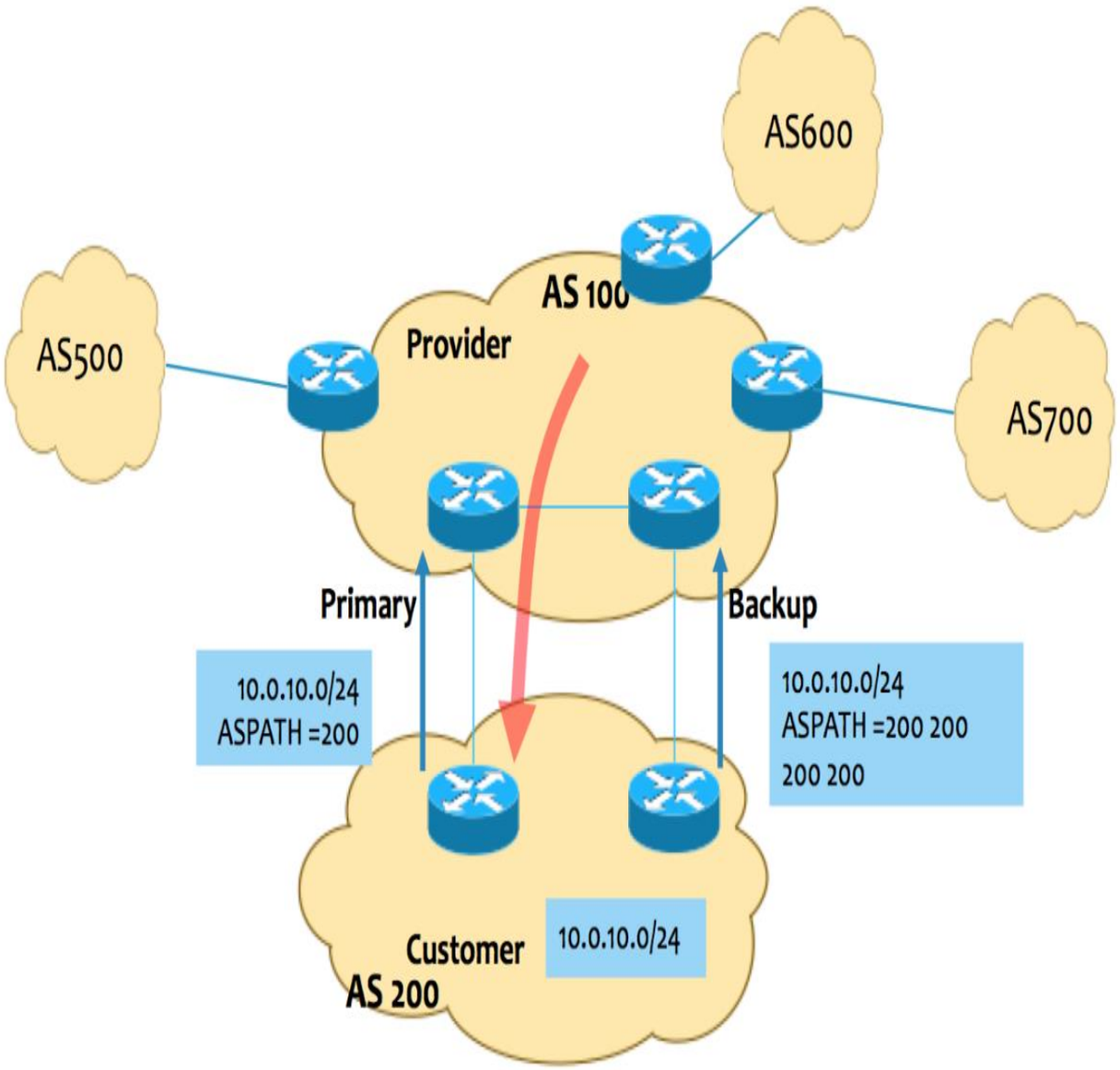
- As per RFC 4451 – BGP MED Considerations : BGP speakers often derive MED values by obtaining the IGP metric associated with reaching a given BGP NEXT_HOP within the local AS. This allows MEDs to reasonably reflect IGP topologies when advertising routes to peers. While this is fine when comparing MEDs 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 or 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

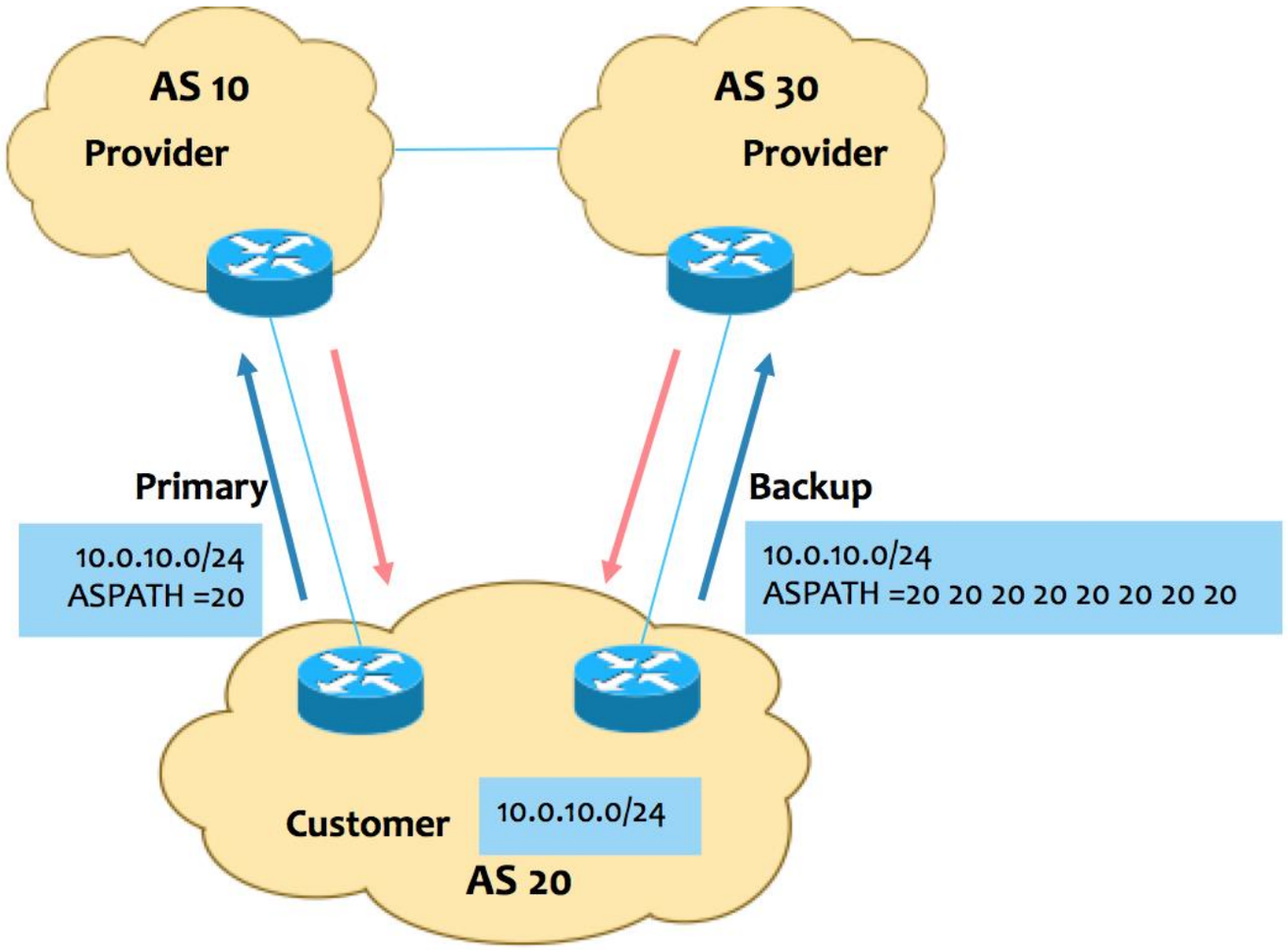
- BGP As-path prepending is used to influence inbound traffic to the company. BGP As-path prepending is used in active-standby link scenarios. When there are two BGP neighborship which prefix will be advertised, one link for set of prefixes or maybe all the prefixes can be used as backup. In this case,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

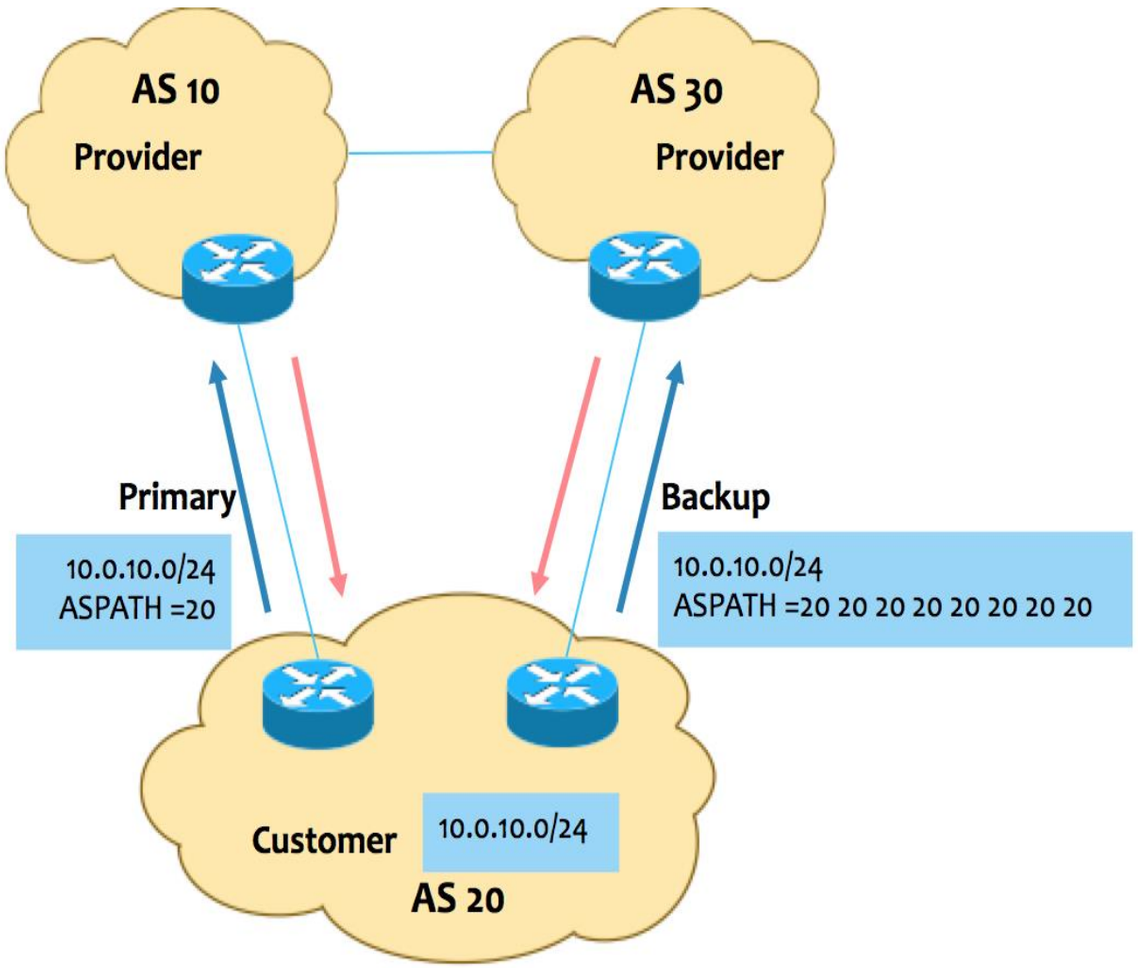
AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !



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 !

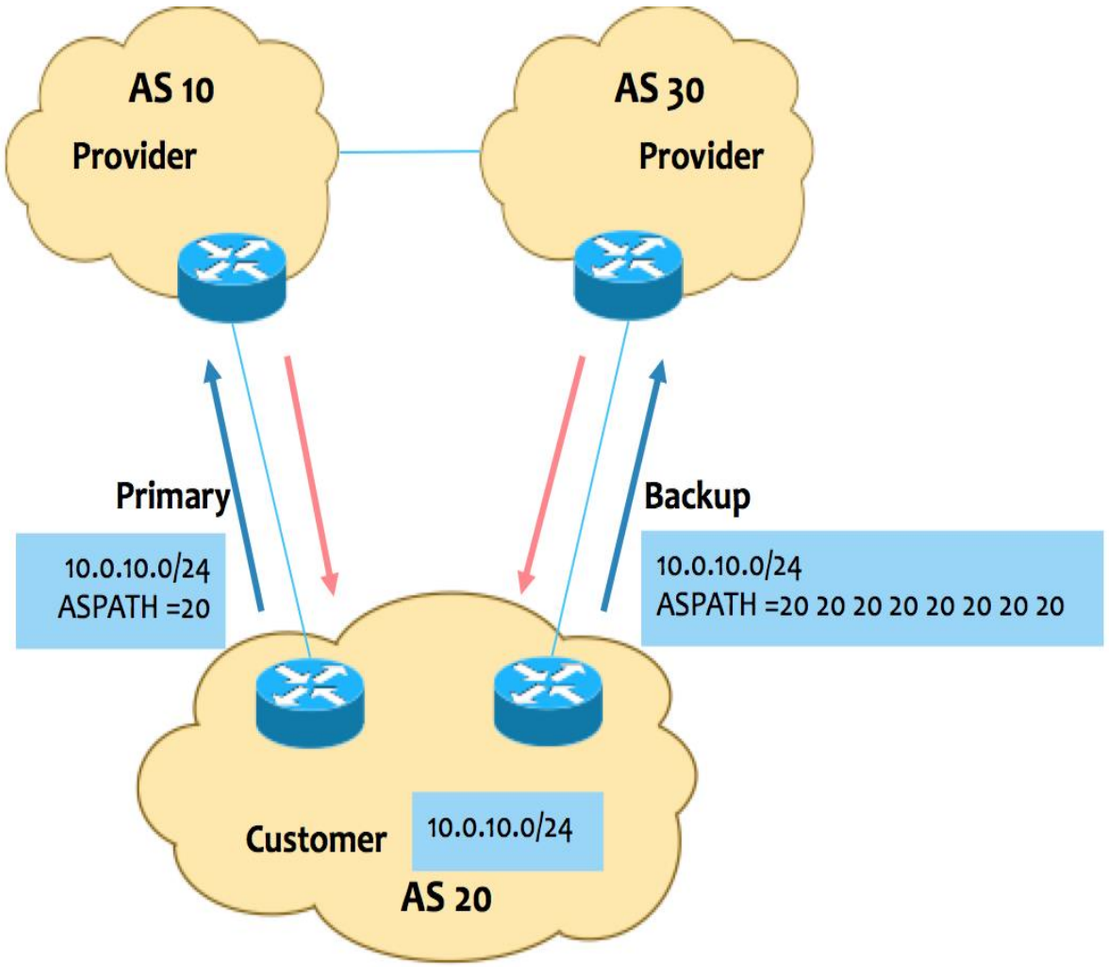
AS-Path Prepending will not work in some cases for EBGP Inbound Traffic Engineering !



- 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 !

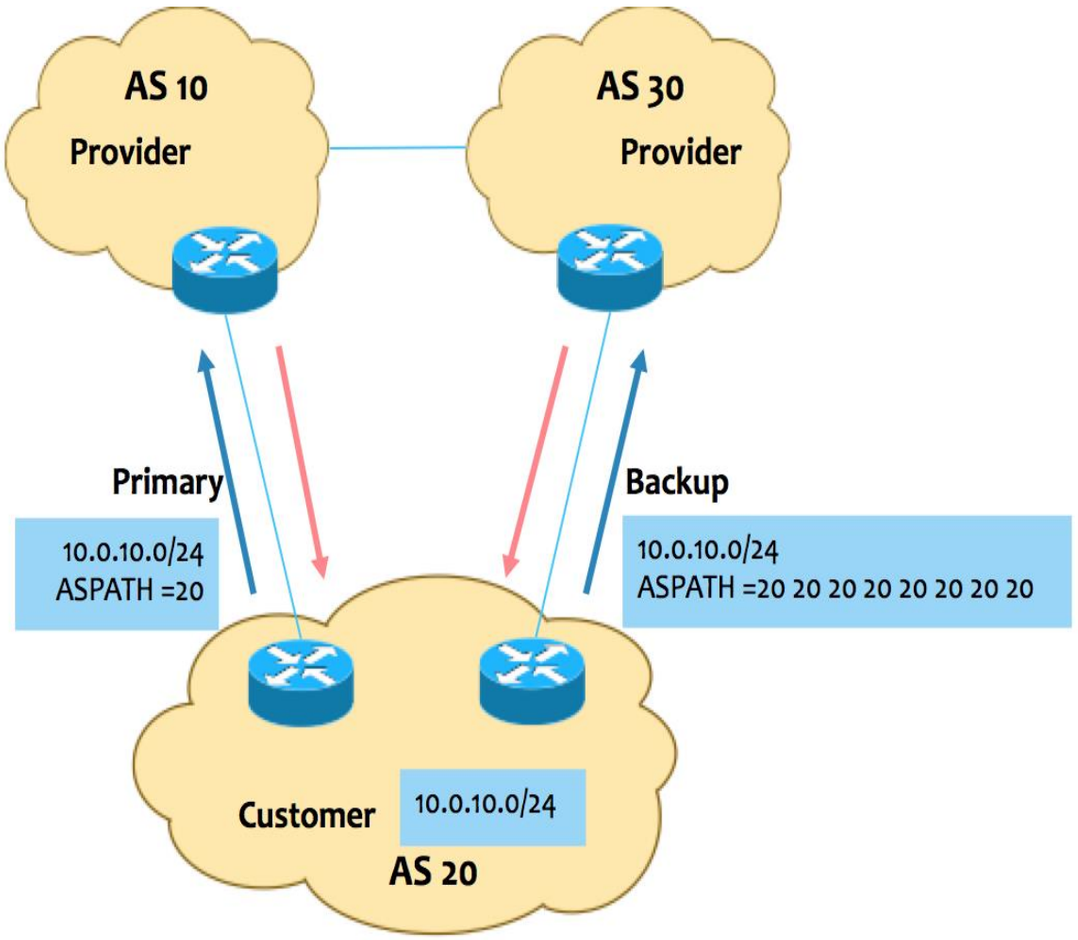


- Service Providers always chooses to send the traffic for their customer prefixes over the customer link first, then peering links, lastly through upstream ISP. Because they want to utilize the customer link as much as possible to charge more money

- 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 !

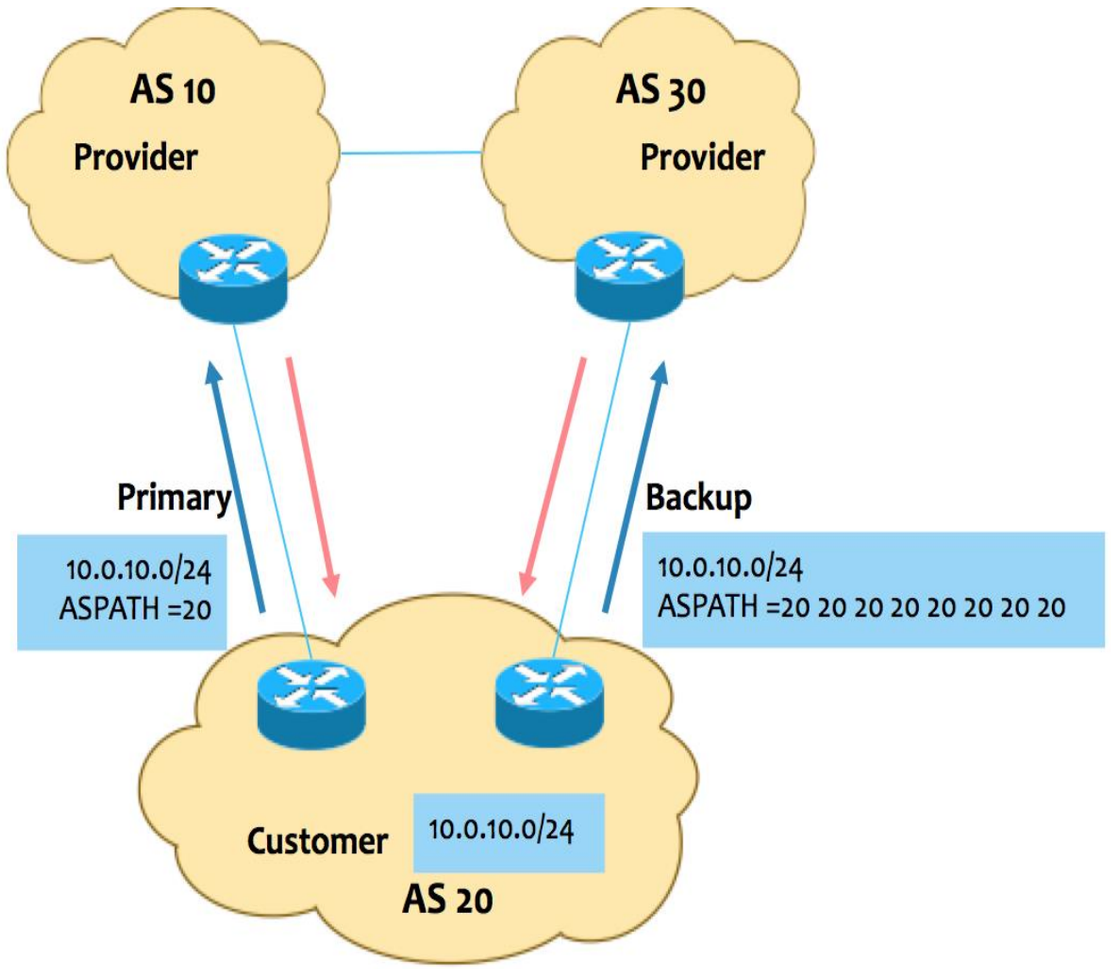


- Customer of AS 30 would still use customer link for 10.0.10.0/24 prefix although customer wants that link to be used as backup
- AS 20 is sending that prefix with AS-path-prepends but service provider implements Local Preference for that prefix

- 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



- 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

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

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.

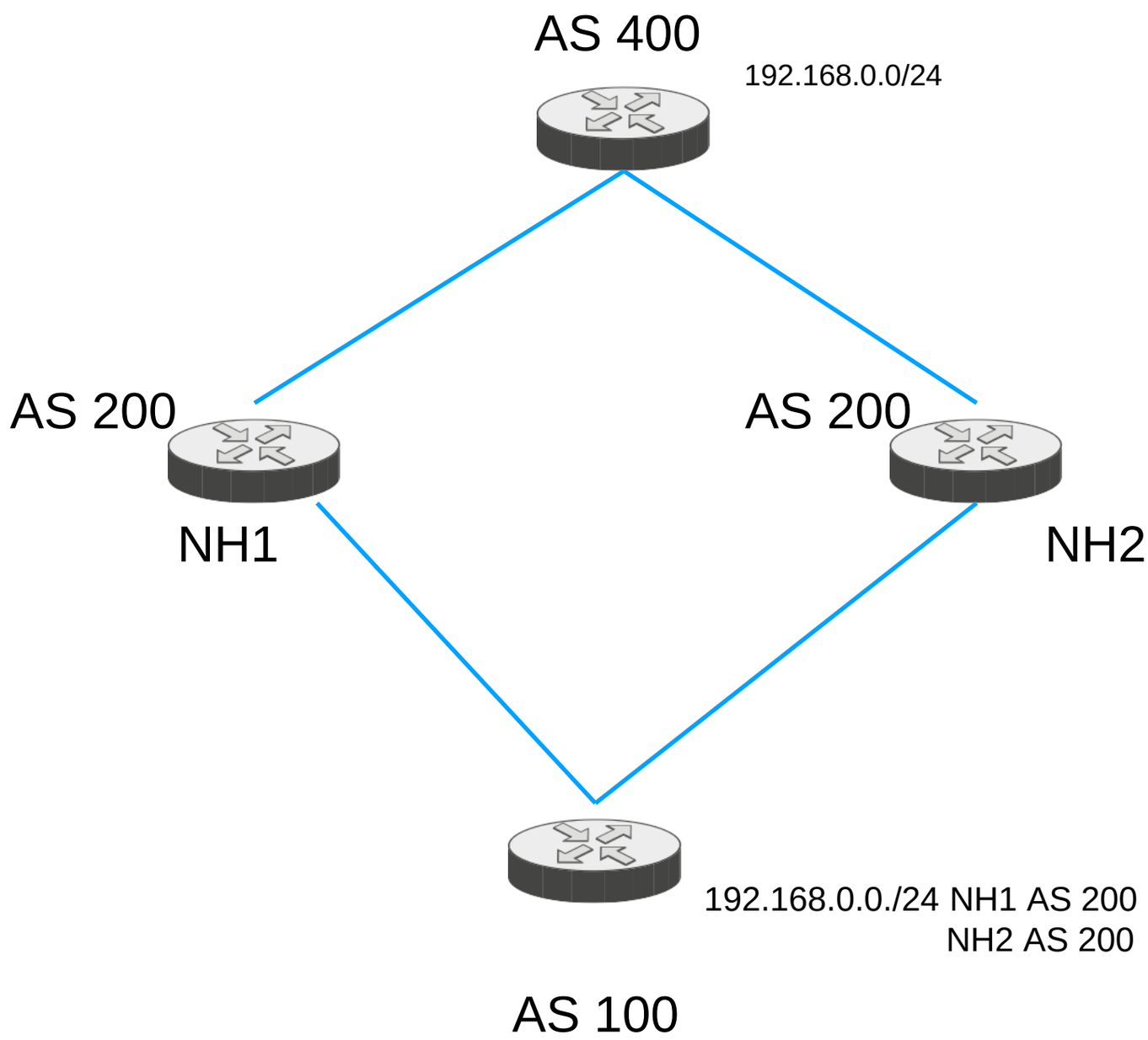
(Level 3 (Tier 1 ISP) Community Values and Corresponding Local Preference)

BGP Multipath

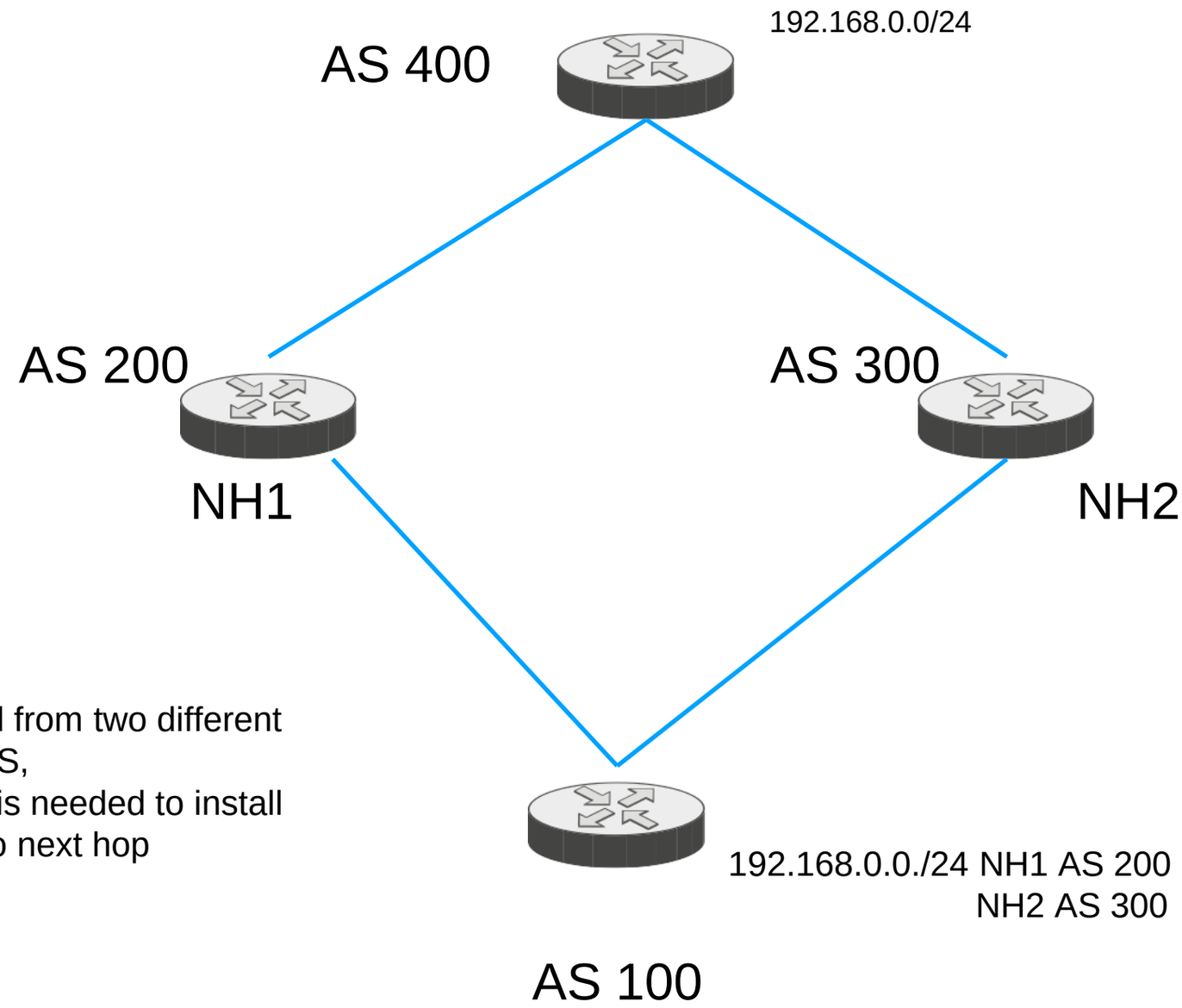
- BGP by default installs only single path in IBGP and EBGP deployment
- If prefixes is learned via multiple path, BGP supports multipath for IBGP , EBGP or across both IBGP and EBGP via EIBGP Multipath feature
- Multipath feature should be enabled manually

EBGP Multipath

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



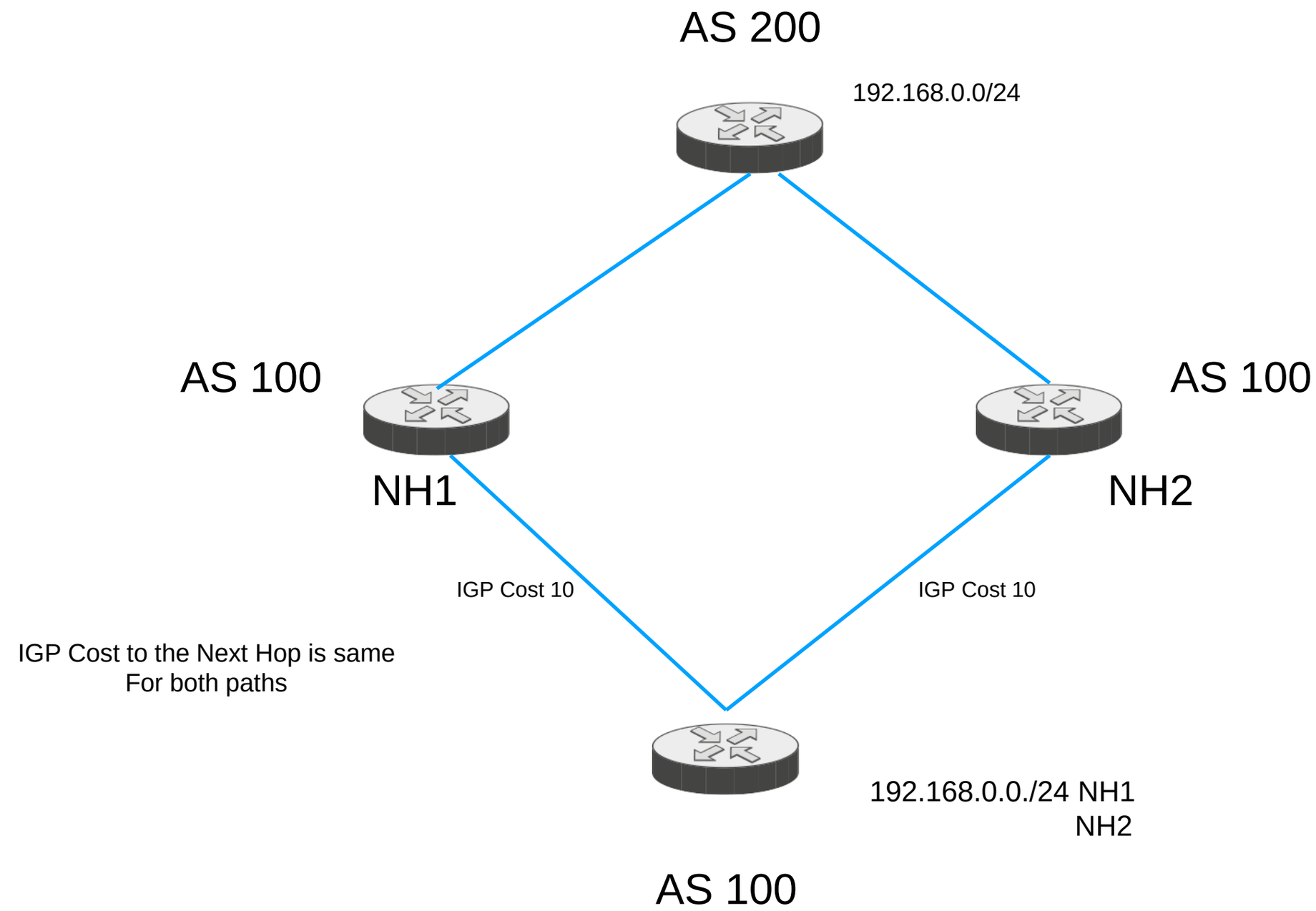
Two different AS requires ‘ multipath-relax ‘



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

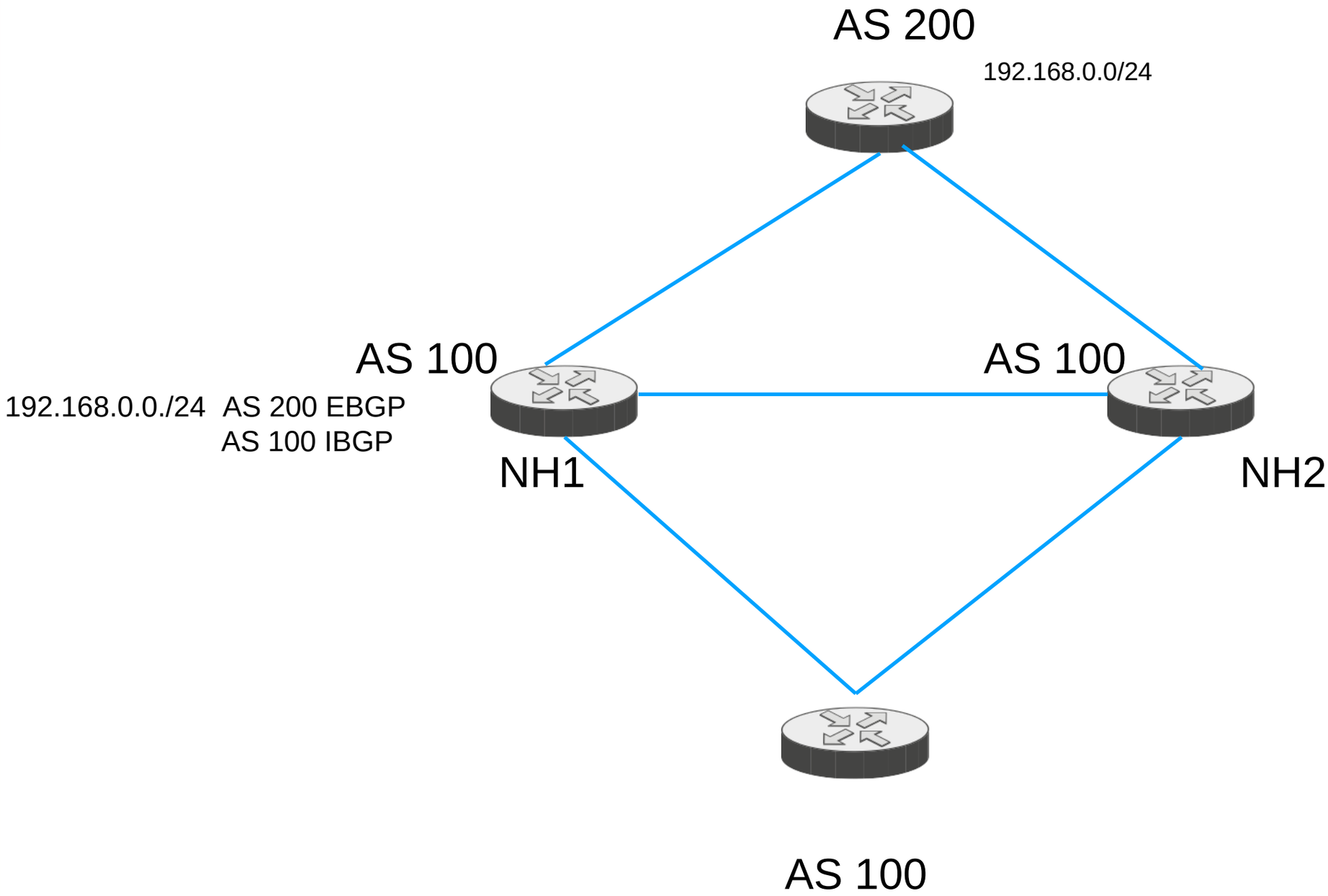
“bgp bestpath **as-path** multipath-**relax**”

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 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

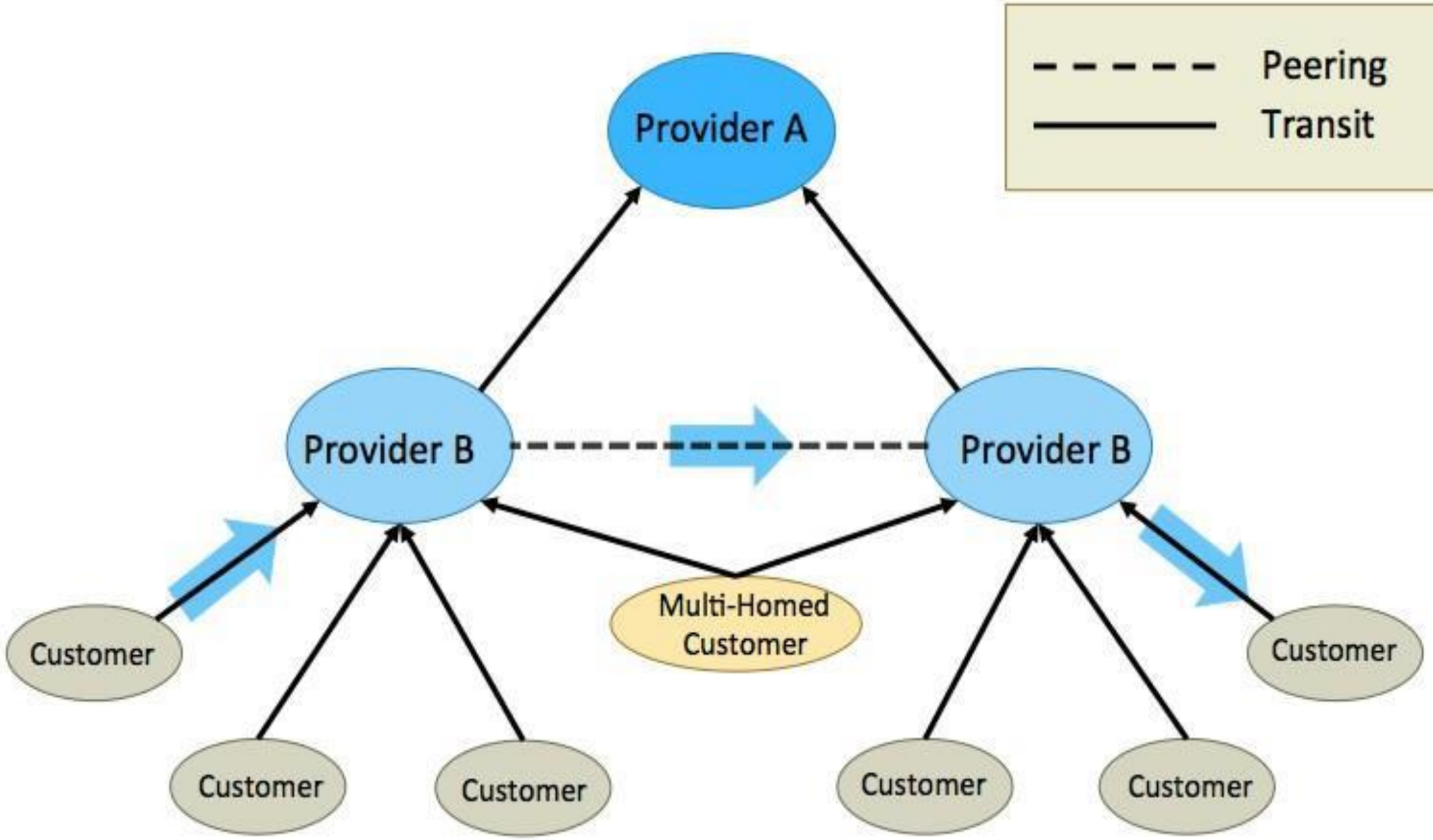
- The Internet is a collection of many individual networks, which interconnect with each other under the common framework of ensuring global reachability between any two points

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



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

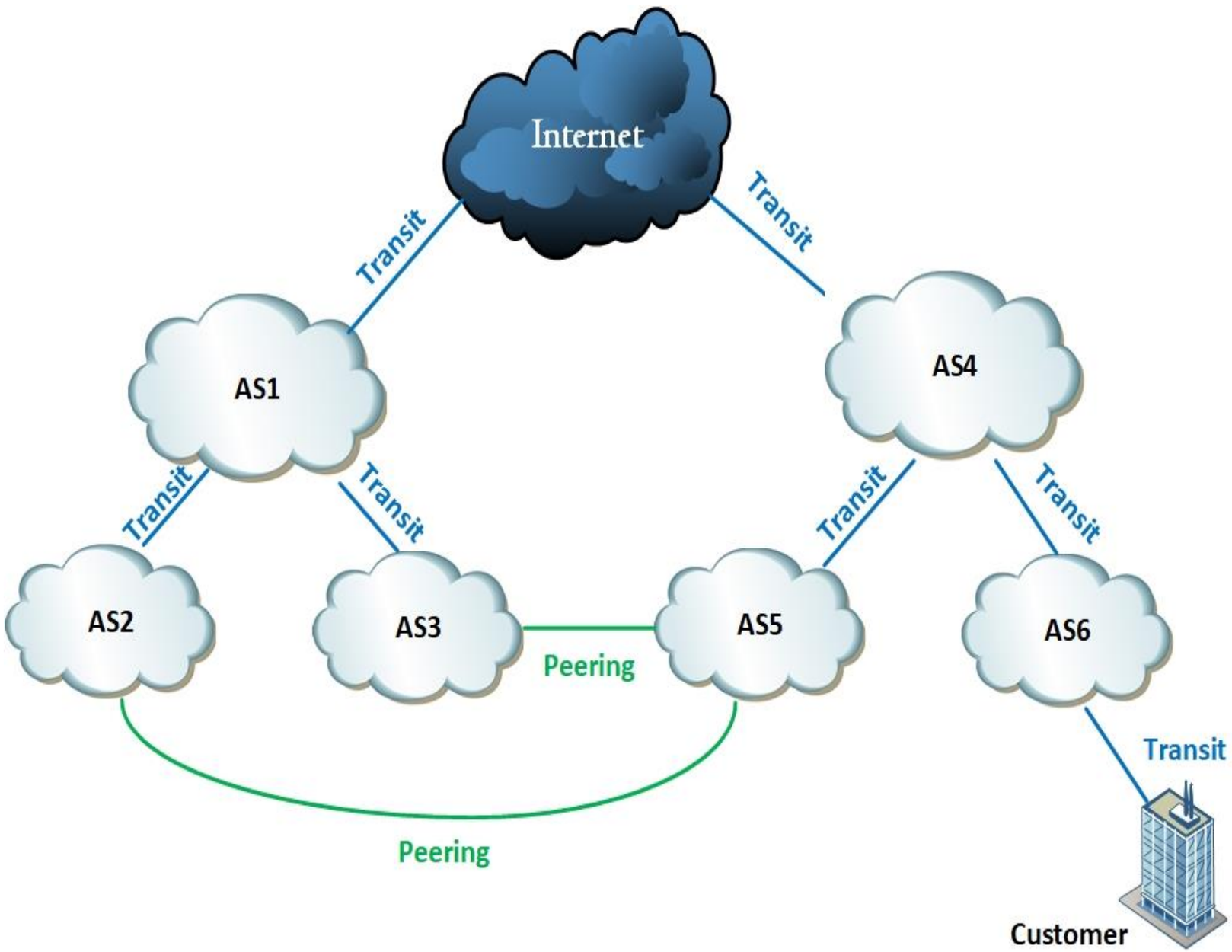
- Company wants to sell an Internet service

- 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

Private Peering

- Private peering can be setup inside Internet Exchange Point as well

- 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 are 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

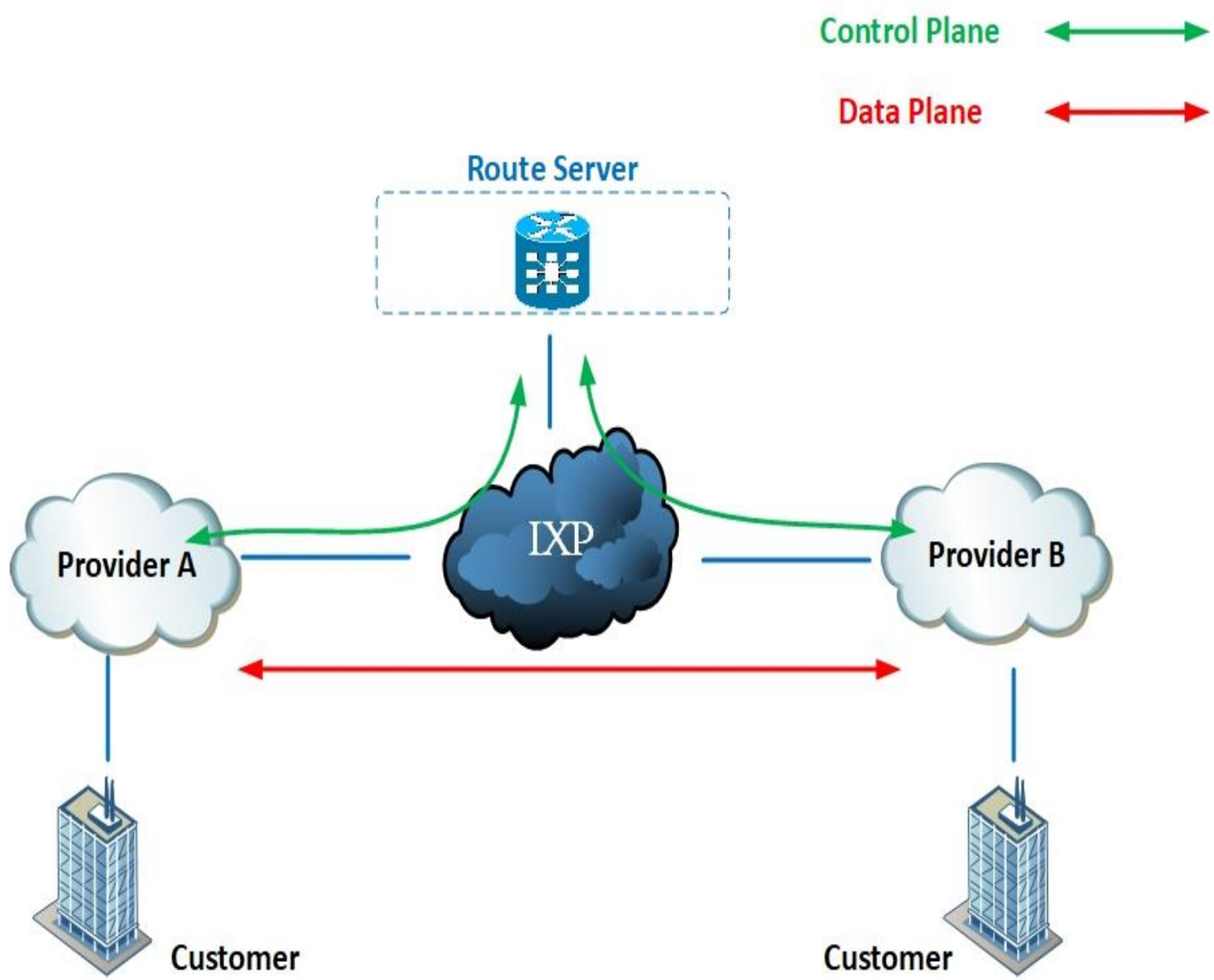
BGP Route Server

- 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

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

- 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

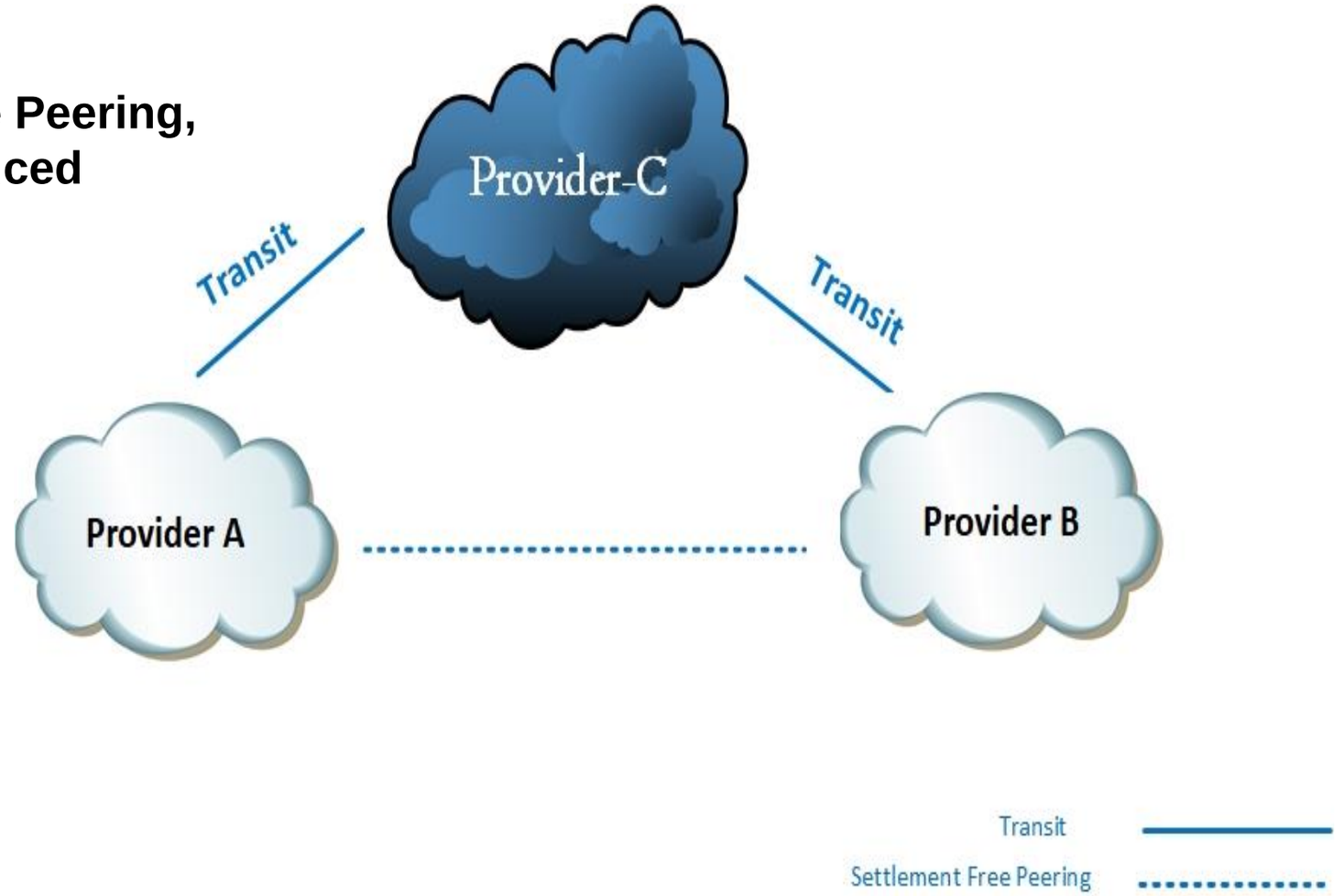
- There are many publicly available Looking Glasses in the World. They are configured as read-only so person who wants to check the particular looking glass cannot change the BGP routing information

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

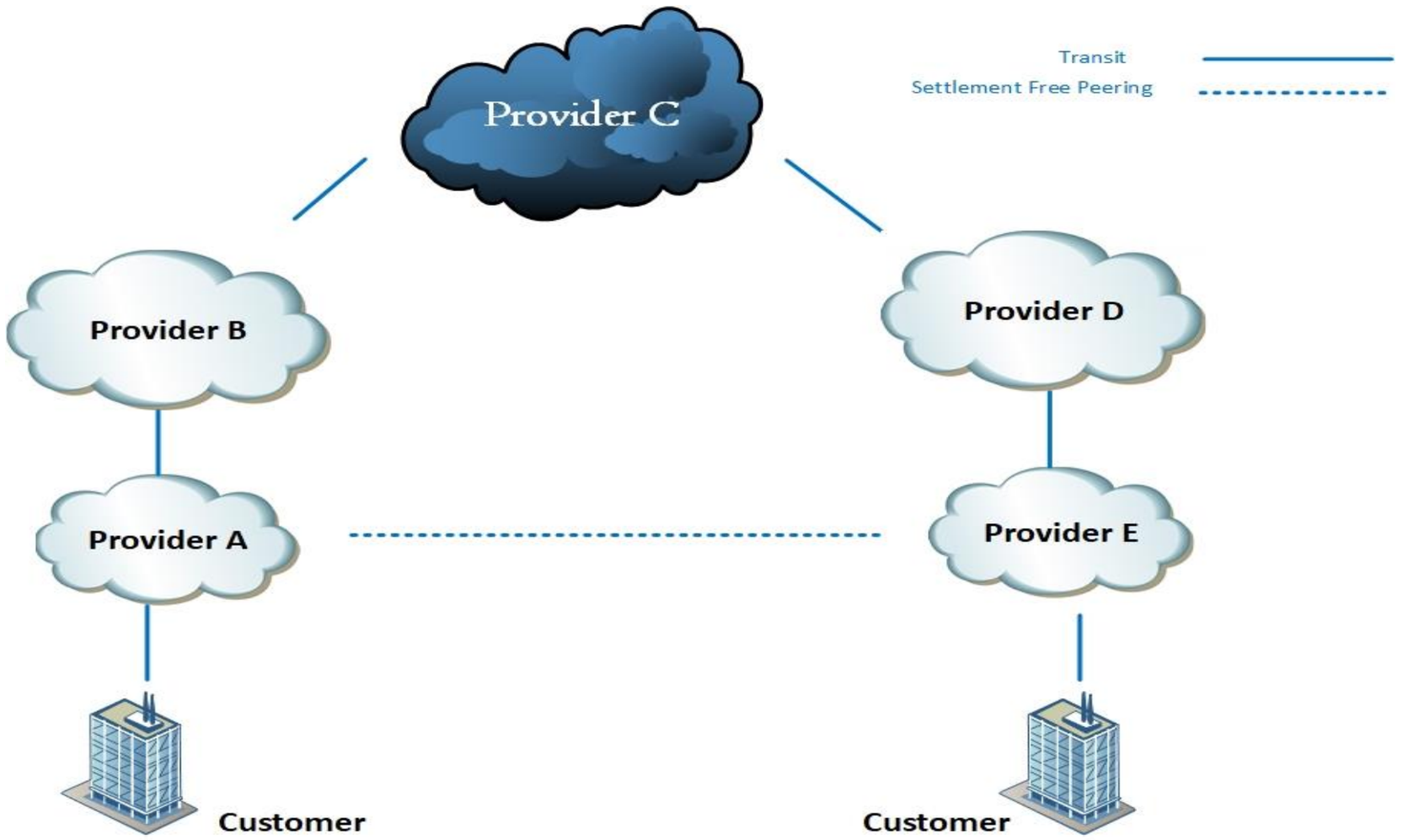
Benefits of Settlement Free Peering

- 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

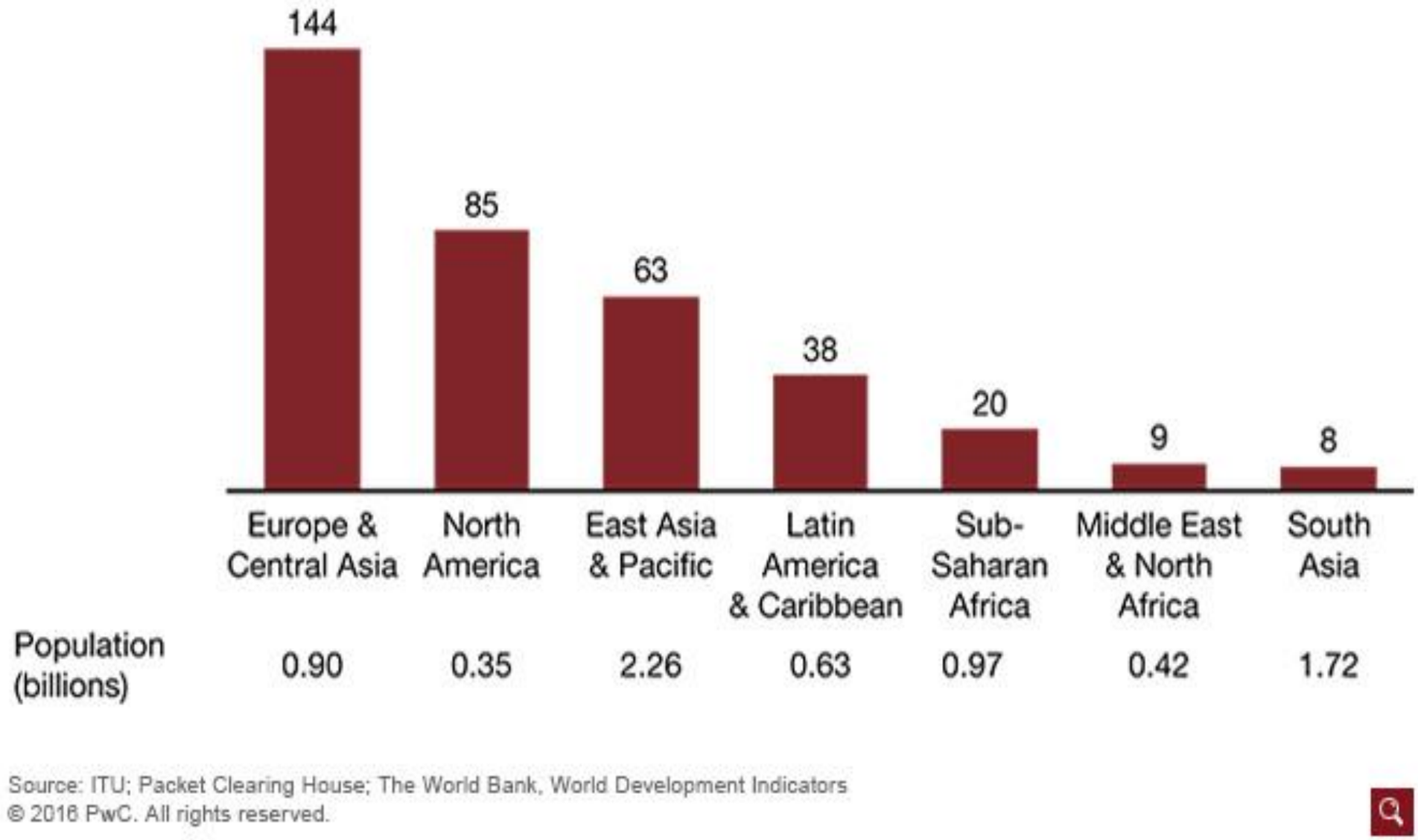
- A network new to peering will typically go to an exchange point as their first step, and be able to so many other like-minded networks interested in peering with them

- 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



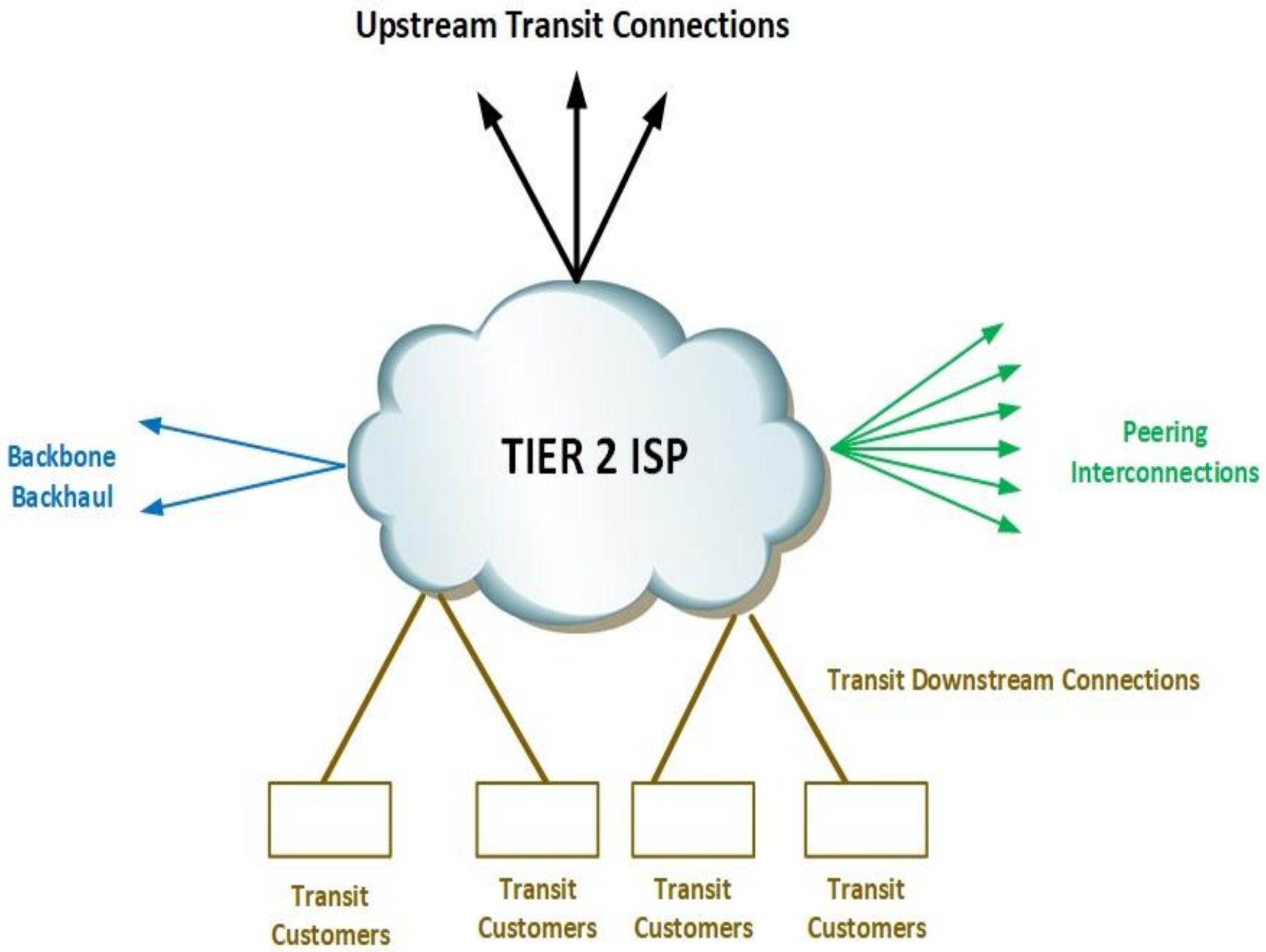
Source: ITU; Packet Clearing House; The World Bank, World Development Indicators

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

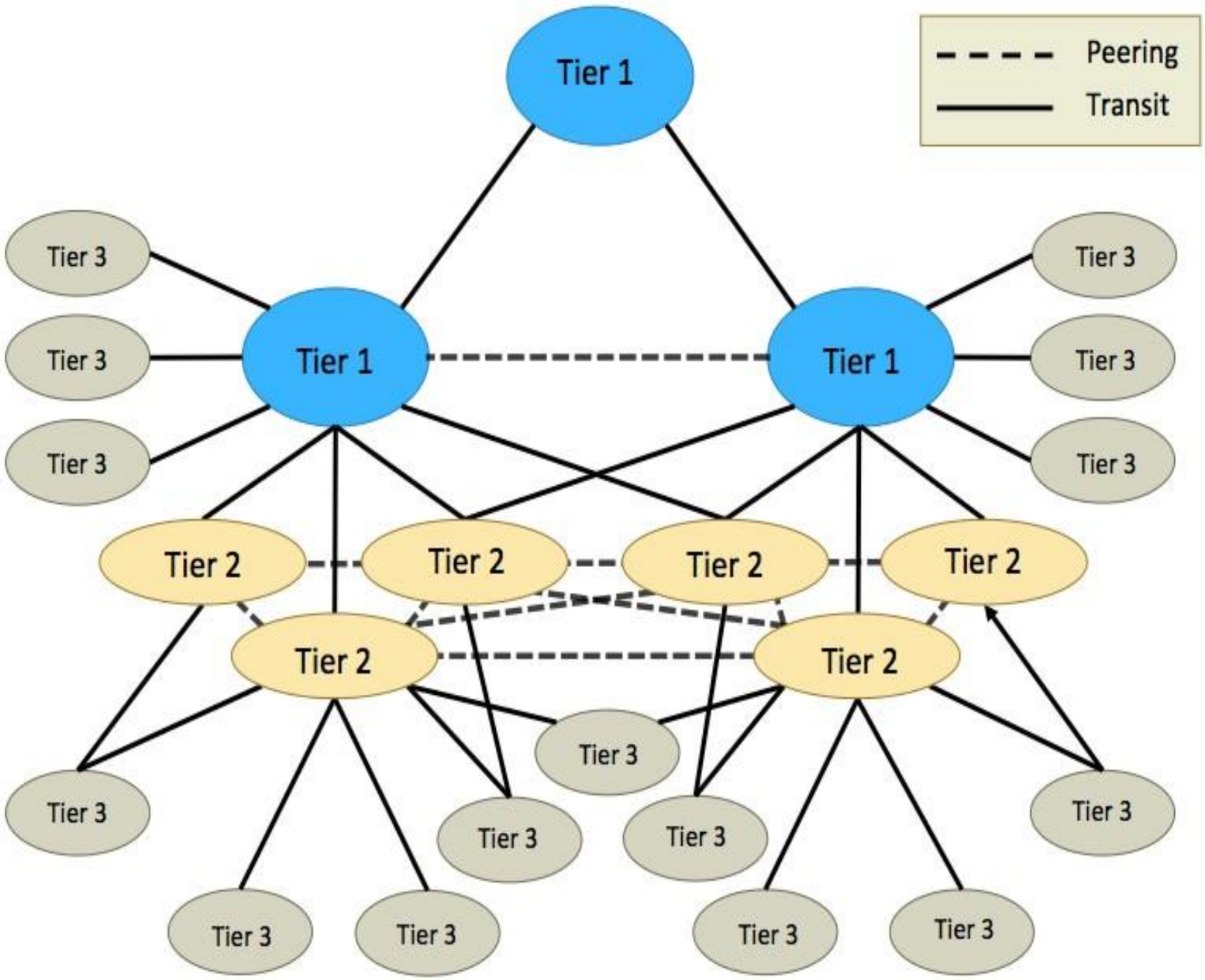
Tier 2 ISP and its Connections



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

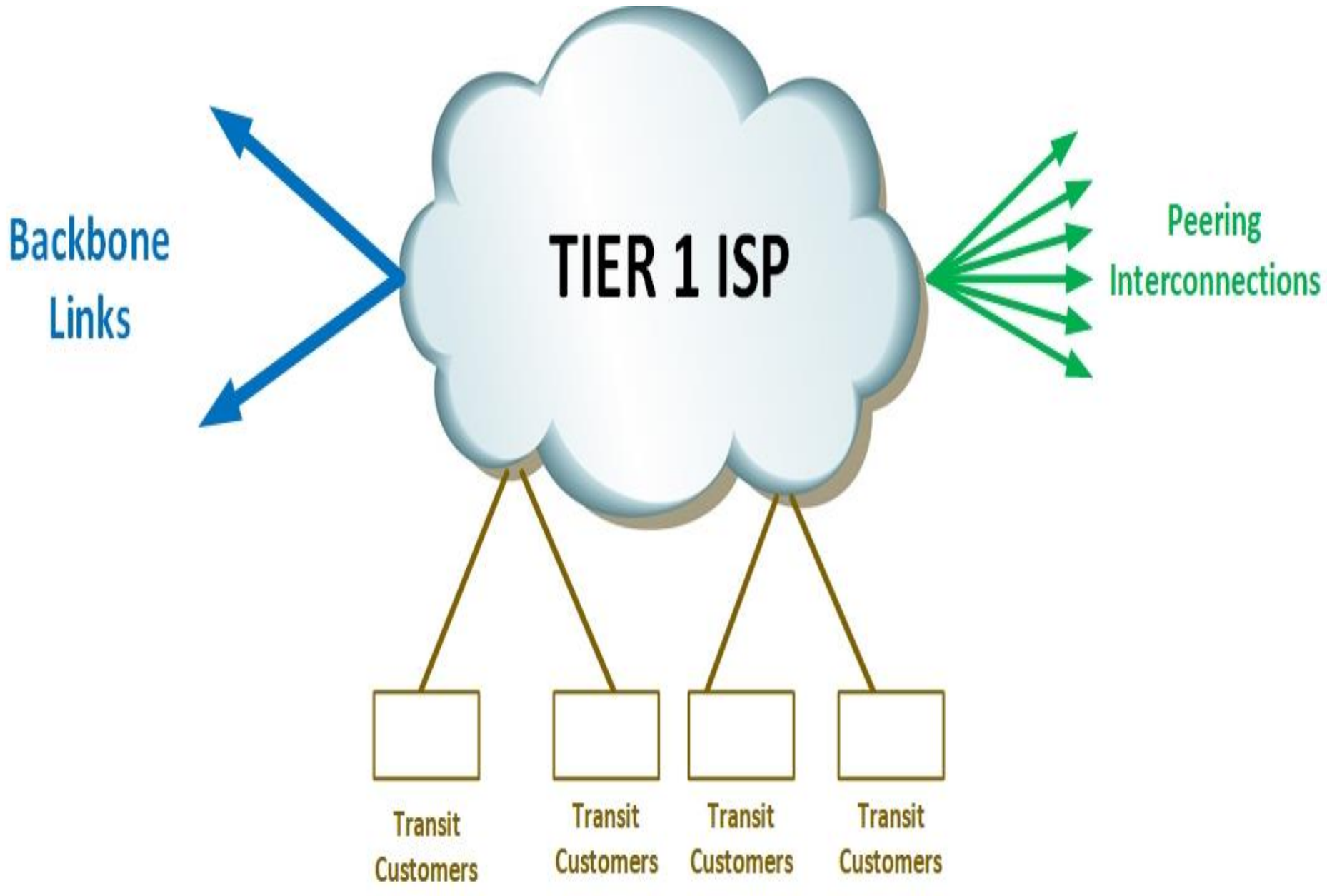


Tier 1 Service Providers

- 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

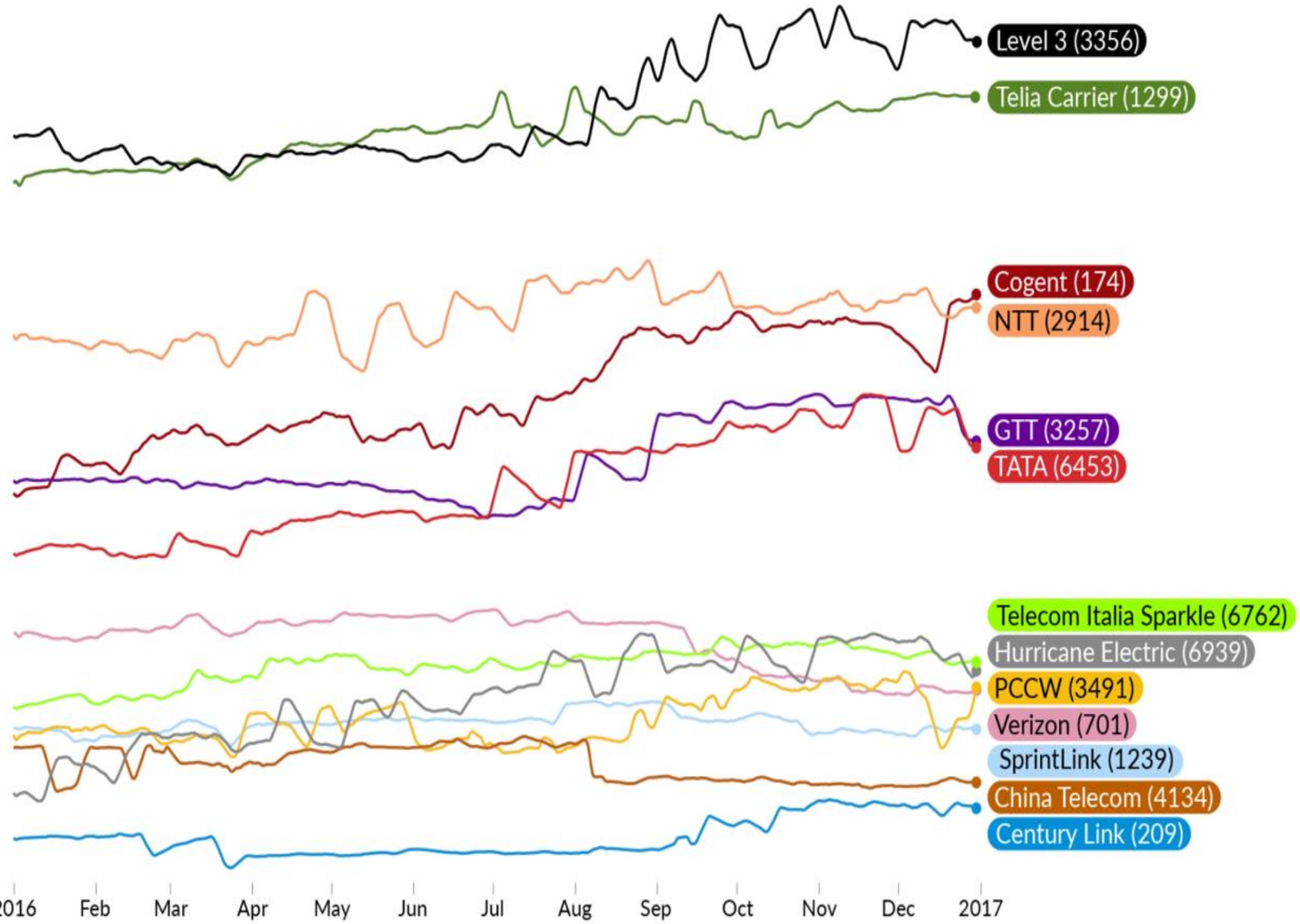


Who are Global Tier 1 ISPs in the World?

- As of 2017, 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, there are BGP use 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

- 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

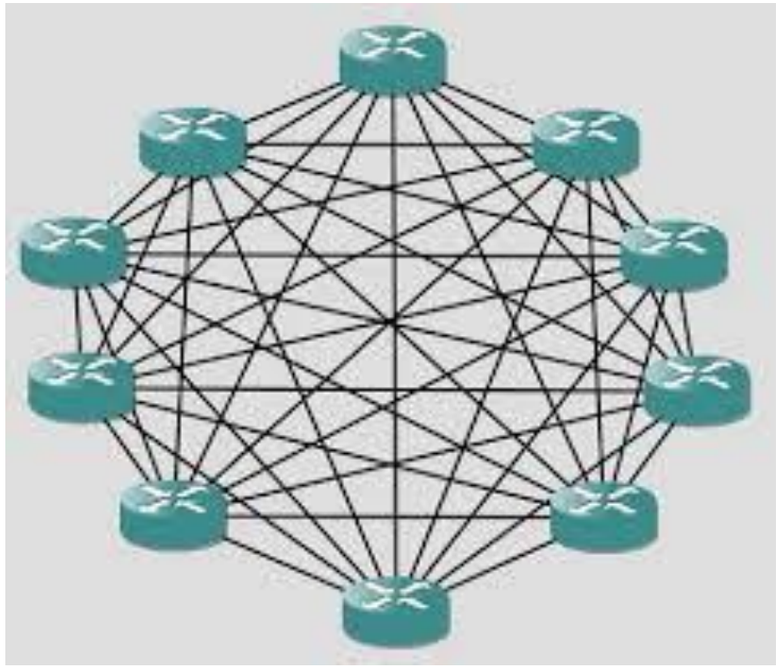
- Route reflectors and confederations can be used to reduce the sessions on each router. Number of sessions and configuration can be reduced by the route reflectors and confederations but they both have important design considerations

- 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

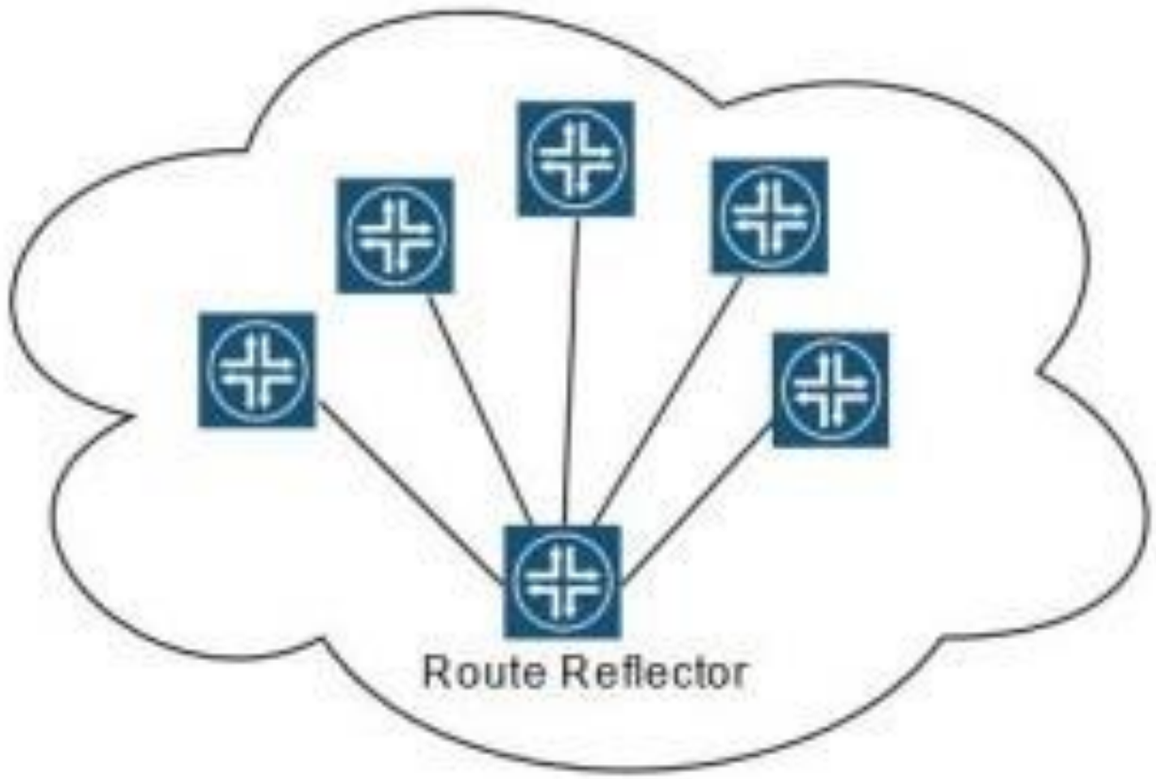
- It is used to avoid Full Mesh connectivity requirement in IBGP



- 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 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

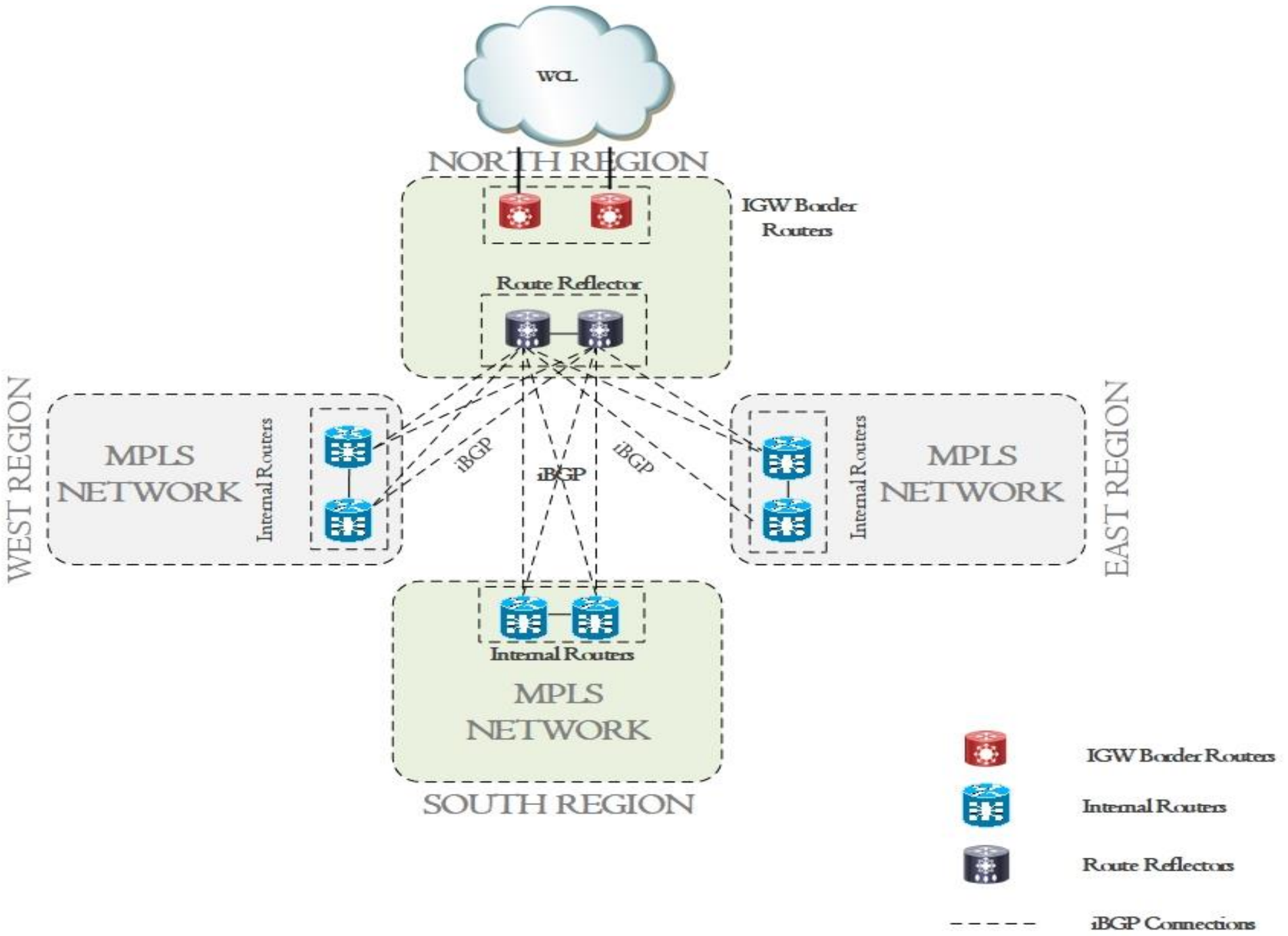
BGP Route Reflector Design Options

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

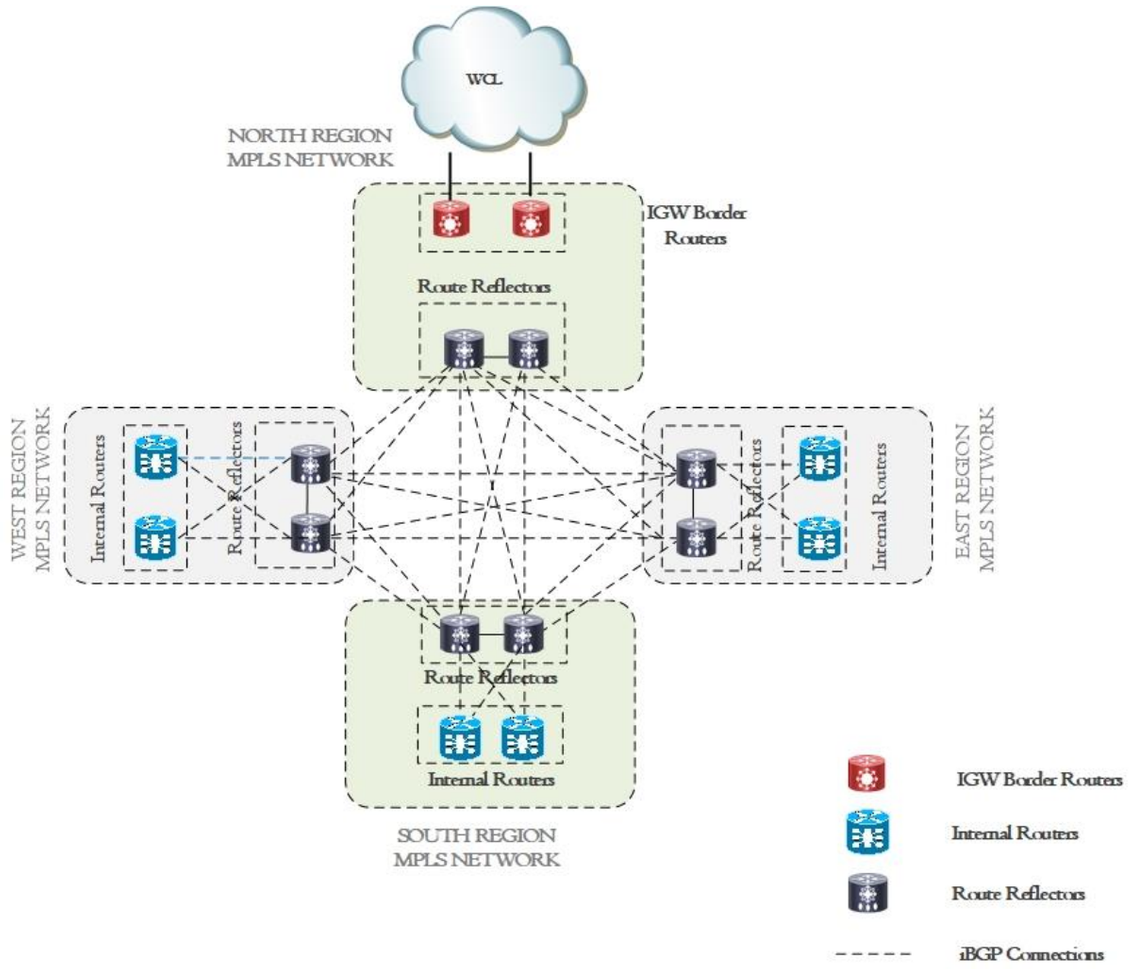
- Distributed BGP Route Reflectors provide optimal routing compare to Centralize design

- Distributed BGP Route Reflectors can be used together with BGP ORR (Optimal Route Reflection) to provide optimal routing

Centralized 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

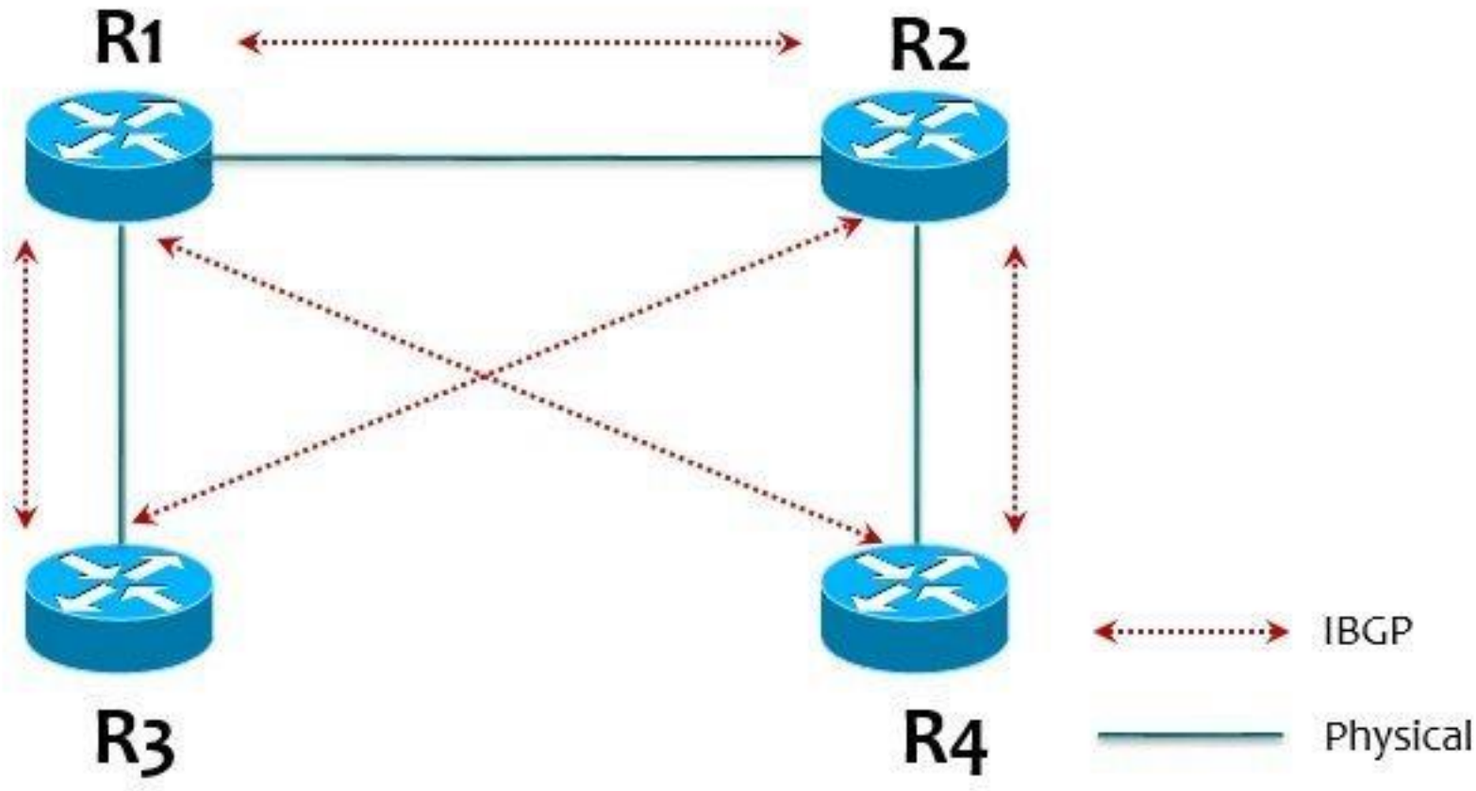
BGP RR Cluster

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.

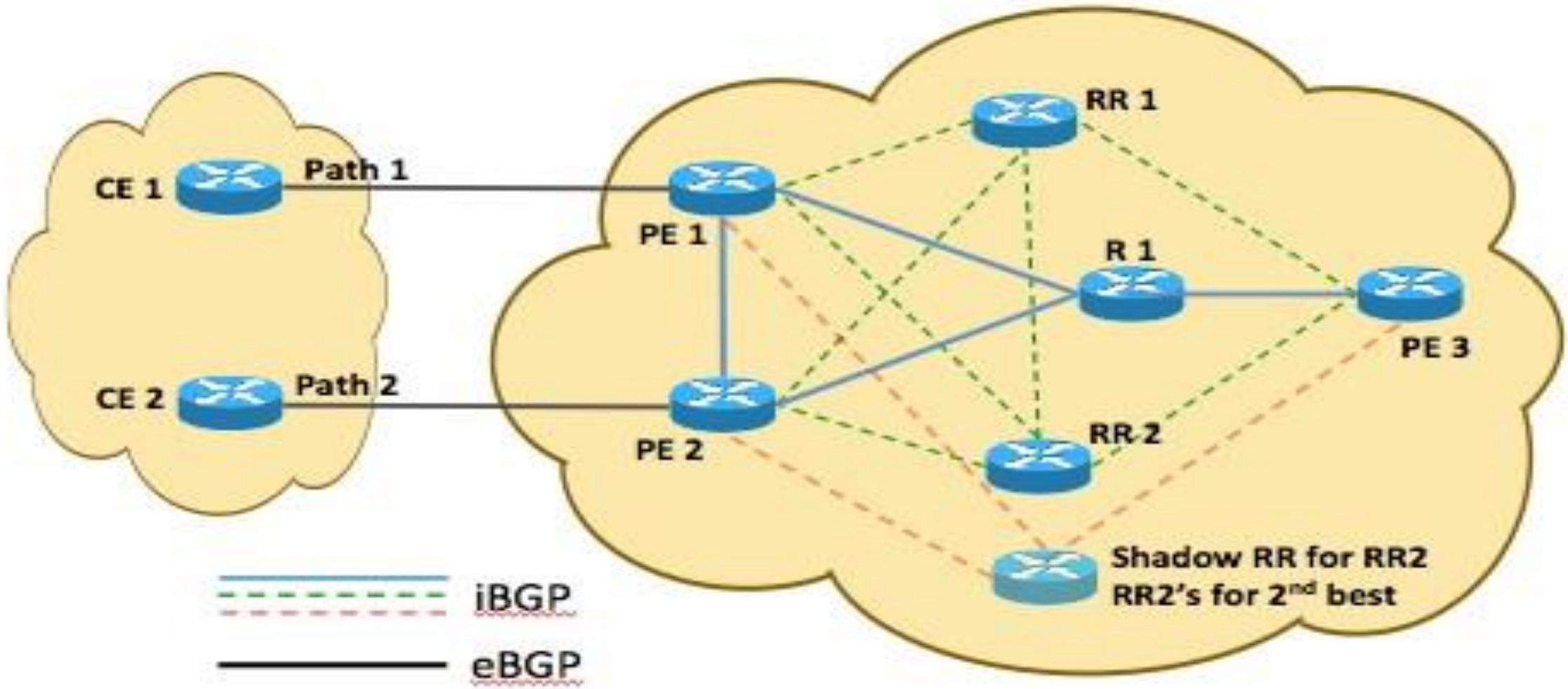
- Unique RD per VRF per PE. Unique Route Distinguisher is used per VRF per PE. No need, Add-Path, Shadow RRs, Diverse Paths. But only applicable in MPLS VPNs.

- BGP Add-Path

- BGP Shadow Route Reflectors

- BGP Shadow Sessions

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



R1 and R2 is used for redundancy and they advertise the best path to the PE3 RR2' calculates and advertises only the second best path.

- In the topology above, path P1 and P2 is learned by both RR1 and RR2. But customer sends lower MED on path P2 to use their links active/standby

- In order to send both paths to the RRs, BGP best external is enabled on PE1 and PE2, thus RR1 and RR2 receives both P1 and P2 paths. Since BGP MED is lower from the P2 path, RR1 and RR2 choose PE2 as best exit. That’s why they advertise only PE2 as best path towards R3

By deploying RR2’, we can send the second best which is path towards PE1 towards PE3

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

- Shadow Sessions – Second IBGP session can be created between RRs and PE. PE is used here as a general term for edge BGP node. Shadow RR and shadow sessions design don’t require MPLS in the network

On the above topology, second sessions can be created between RR1,RR2 and PE3. Over the second IBGP session, second best can be sent. This session is called shadow Route reflector sessions.

➢ BGP Add-Path

BGP Shadow Route Reflector and 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.

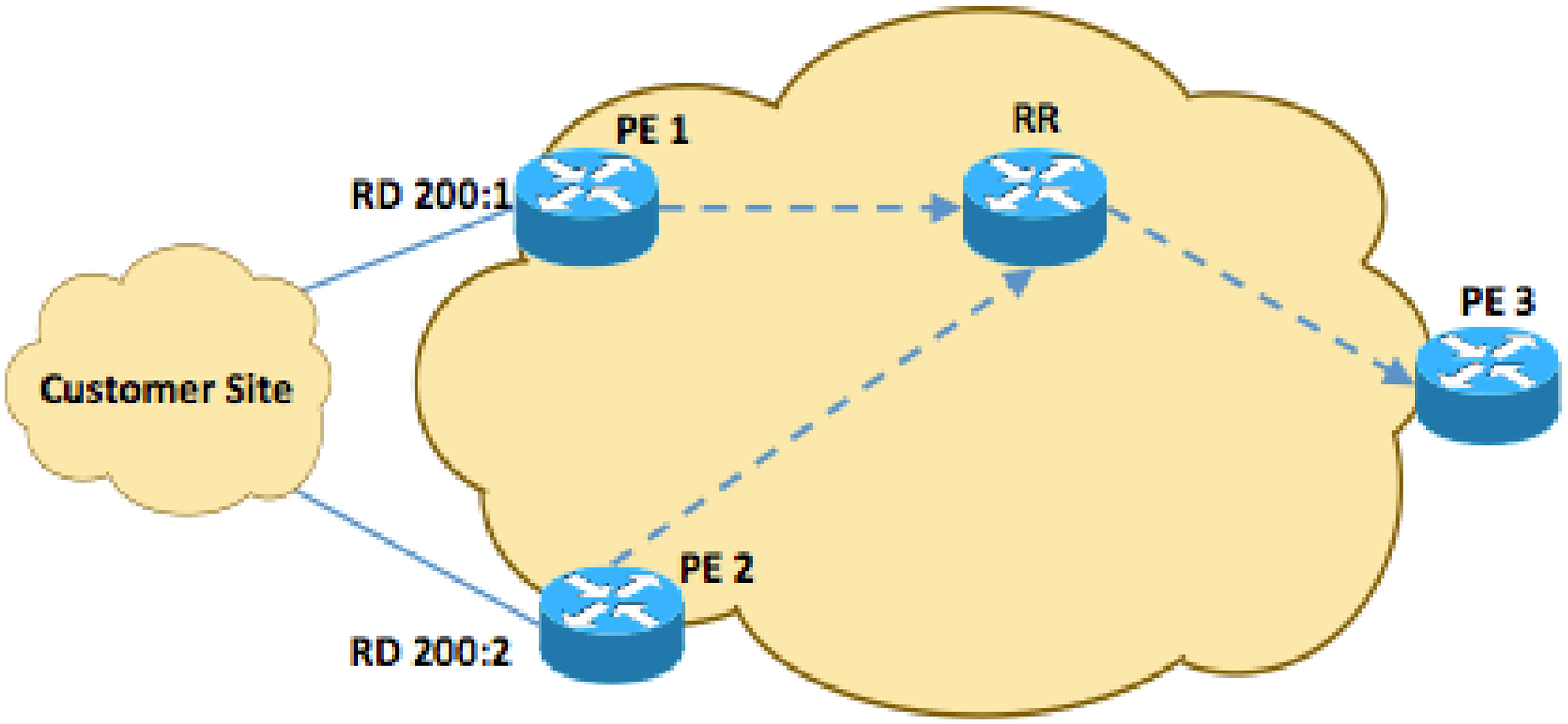
- In IBGP, if multiple paths are sent over the same BGP session, last one is kept by the receiving BGP speaker, because for the first one implicit withdrawn is sent

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

If BGP topology is not full mesh but there is Route Reflector then in MPLS/VPN environment unique RD is configured on the PEs to advertise different VPN prefixes to the Route Reflectors

Because RD is different the VPNv4 prefixes are different. Both PE1 and PE2's advertisements are reflected to PE3. Now on the PE3 Load sharing or Fast Reroute (BGP PIC) is possible



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.
- But, there is one more thing which I will explain with the below topology.
- Route Reflector hides the available paths. This is benefit for some networks, problem for the others.
- BGP RR provides RBAC Opportunity

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 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

- 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

- So, the 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

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

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

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

- It can be used together 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 rSPFs (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

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

- 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

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 neigbor within sub-AS)

How BGP Confederation Works

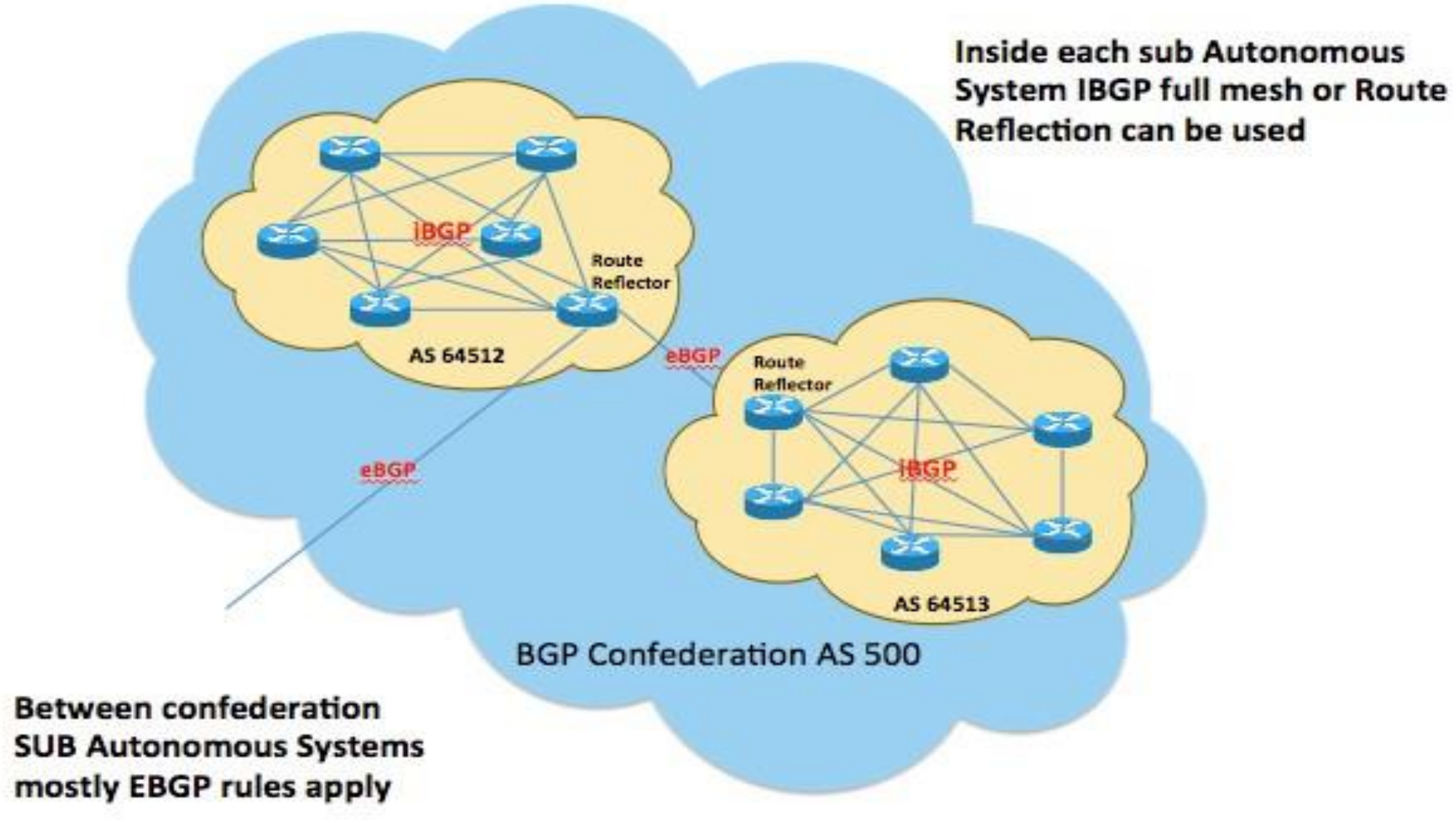
- Between Confederation EBGP and Real EBGP, there are some differences

- 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 changed, 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

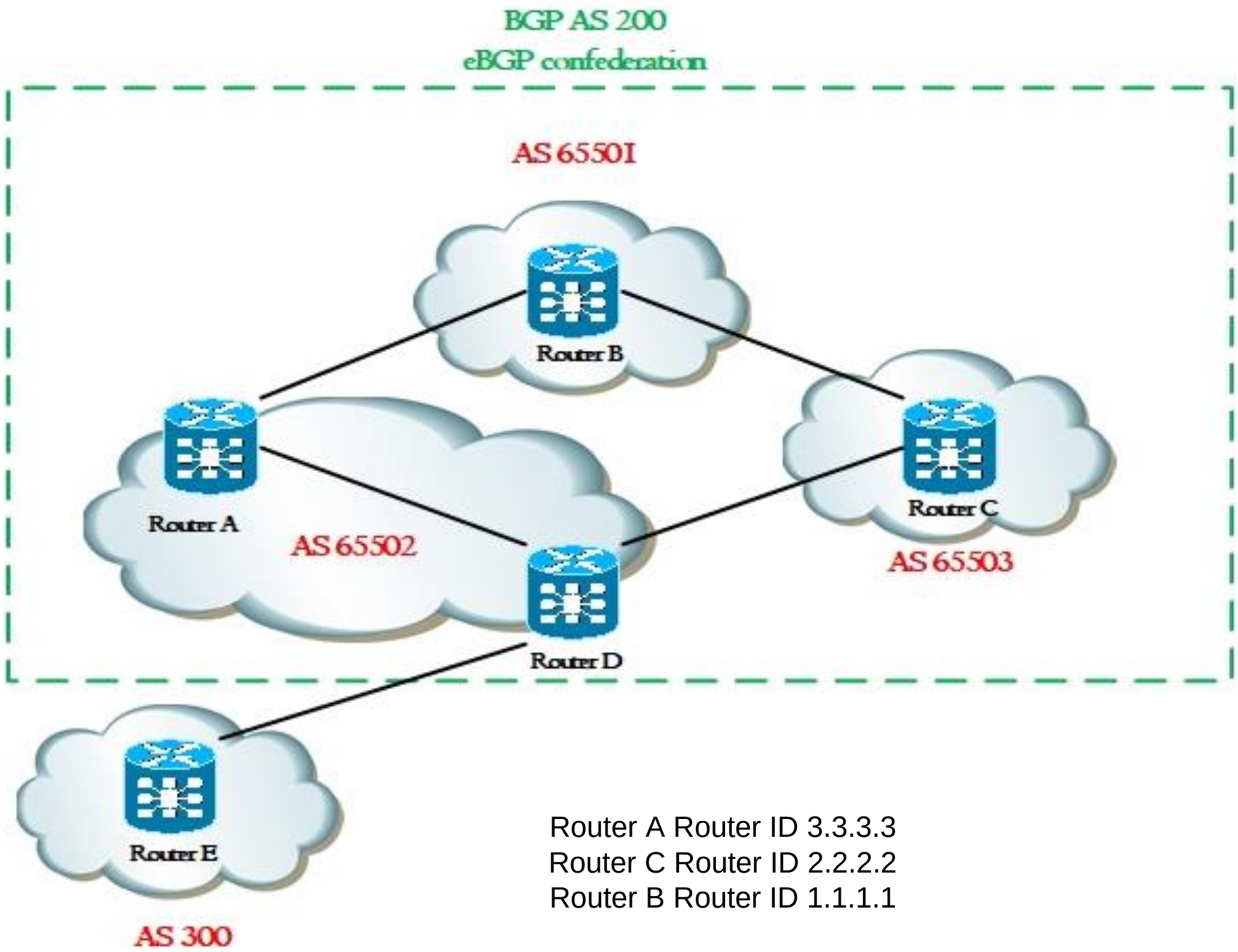
Routing loop avoidance in BGP Confederation

- 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), it chose the path through Router C as best, even though as-path length is longer through the Router C for the destination at AS 300



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

RT Constraints

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

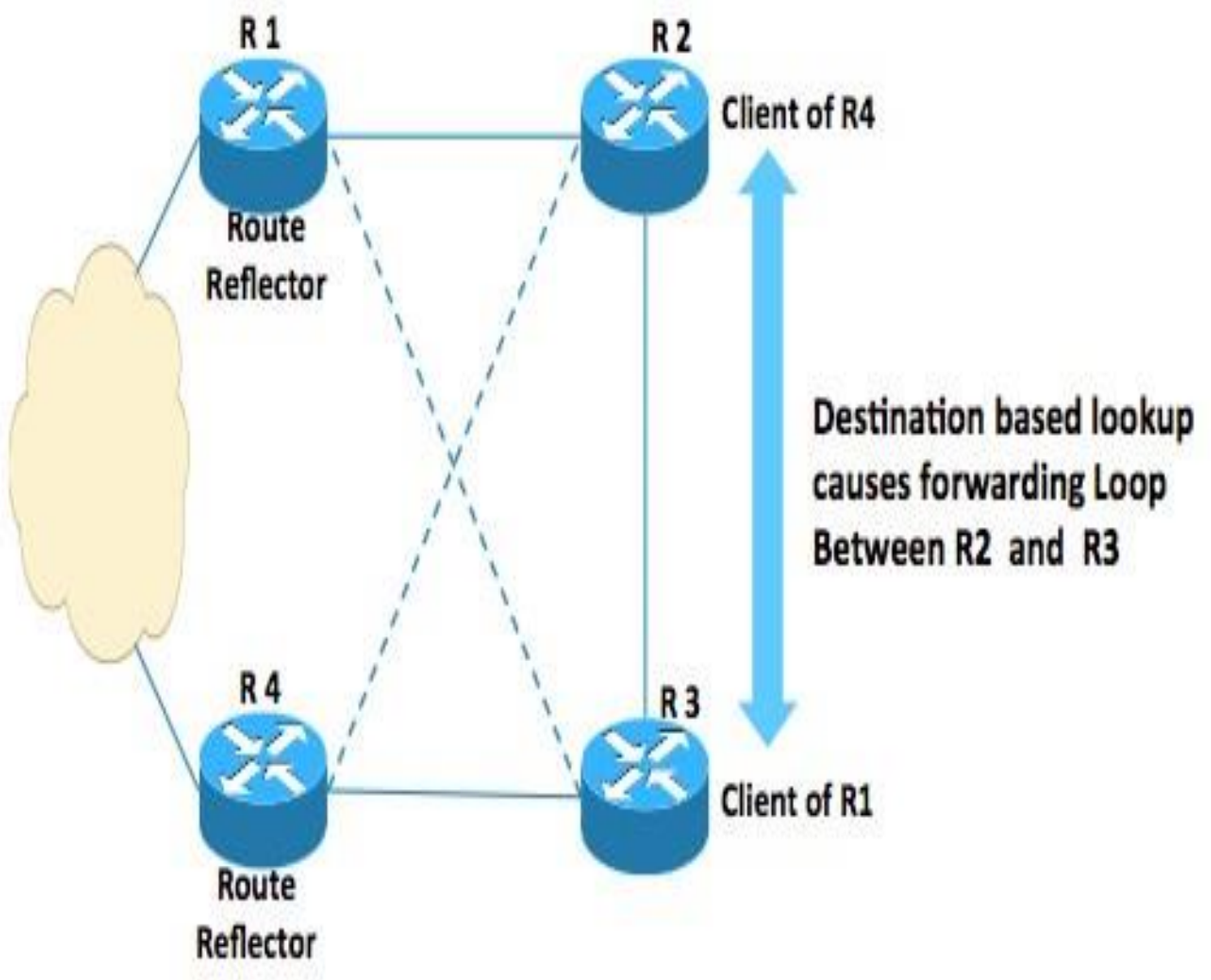
- Based on Route Targets, we can use Route Reflector Group-1 to serve odd Route Target values, Route Reflector Group-2 to serve even Route target values.

- 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 Route reflectors can signal to the PEs which are route reflector clients their desired Route Target values.

BGP Case Study – 1

- 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.

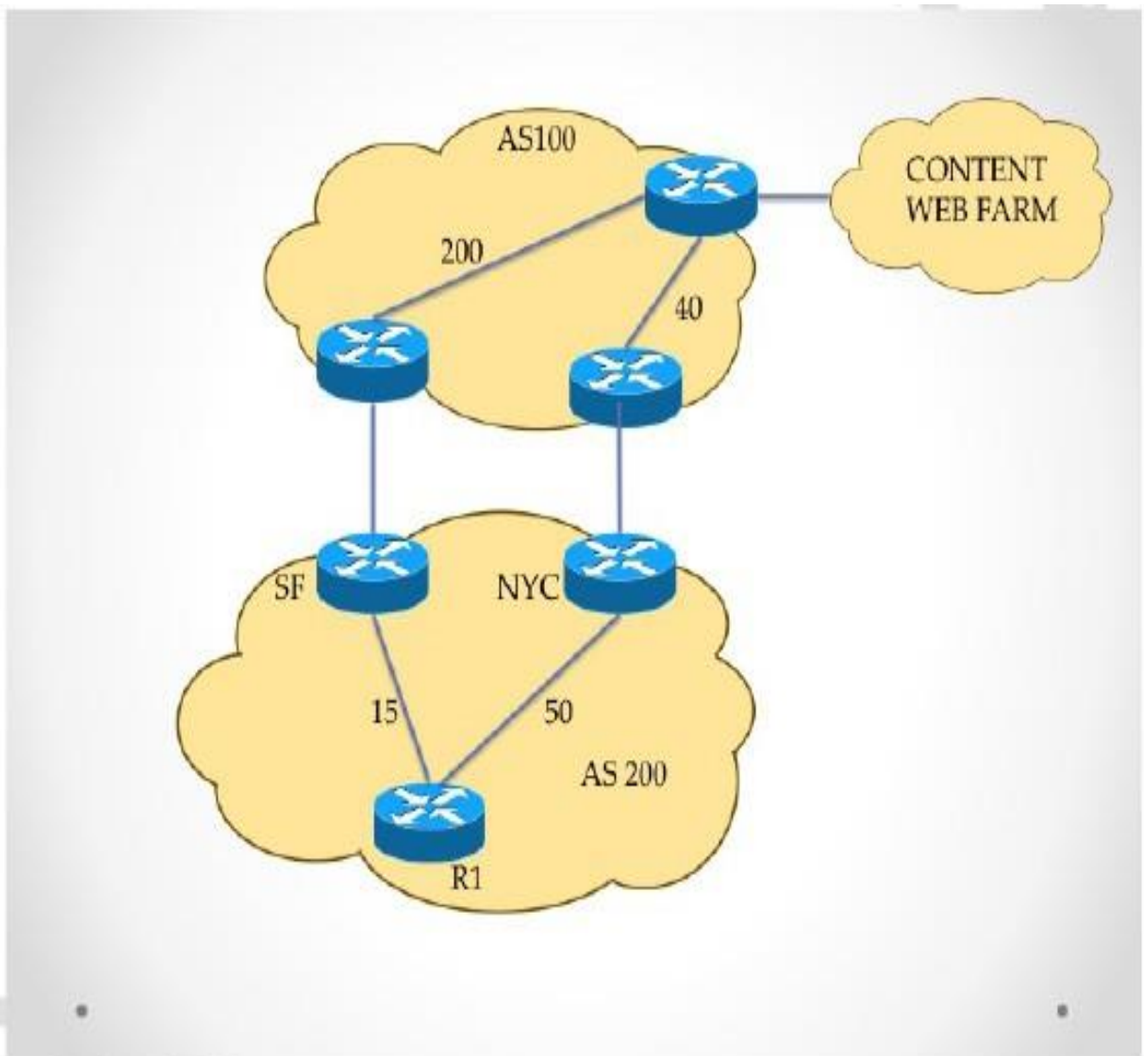
- R2 will do the lookup for prefix A and send it to R2 , R3 will send it back. Packet will loop between R2 and R3.

- 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 Case Study – 2

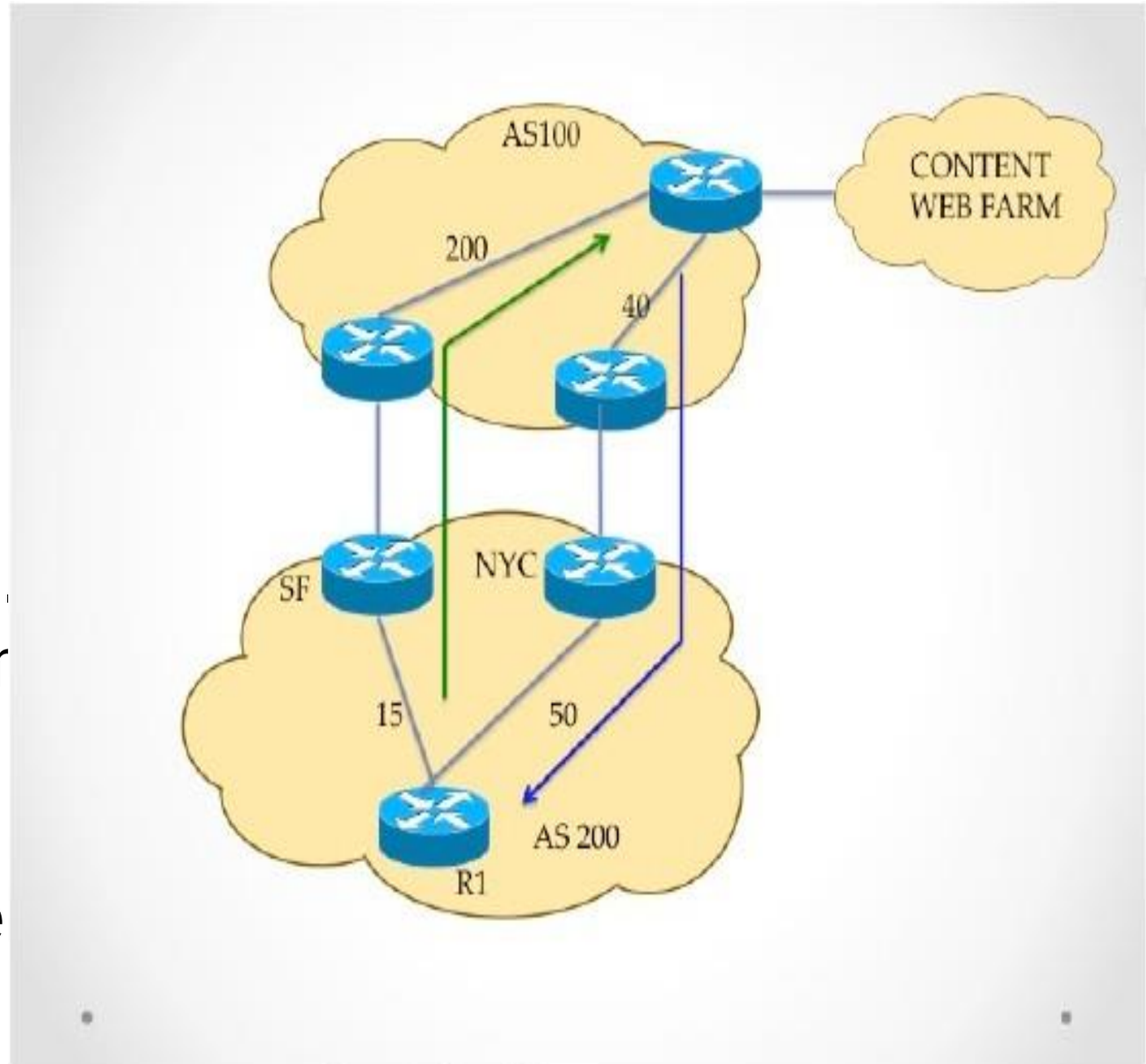
- AS200 is a customer service provider of AS100 transit Service provider. Customers of AS200 is trying to reach a web page located behind AS100. AS200 is not implementing any special BGP policy. What would be the ingress and egress traffic for AS 200 ?



- Above 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.

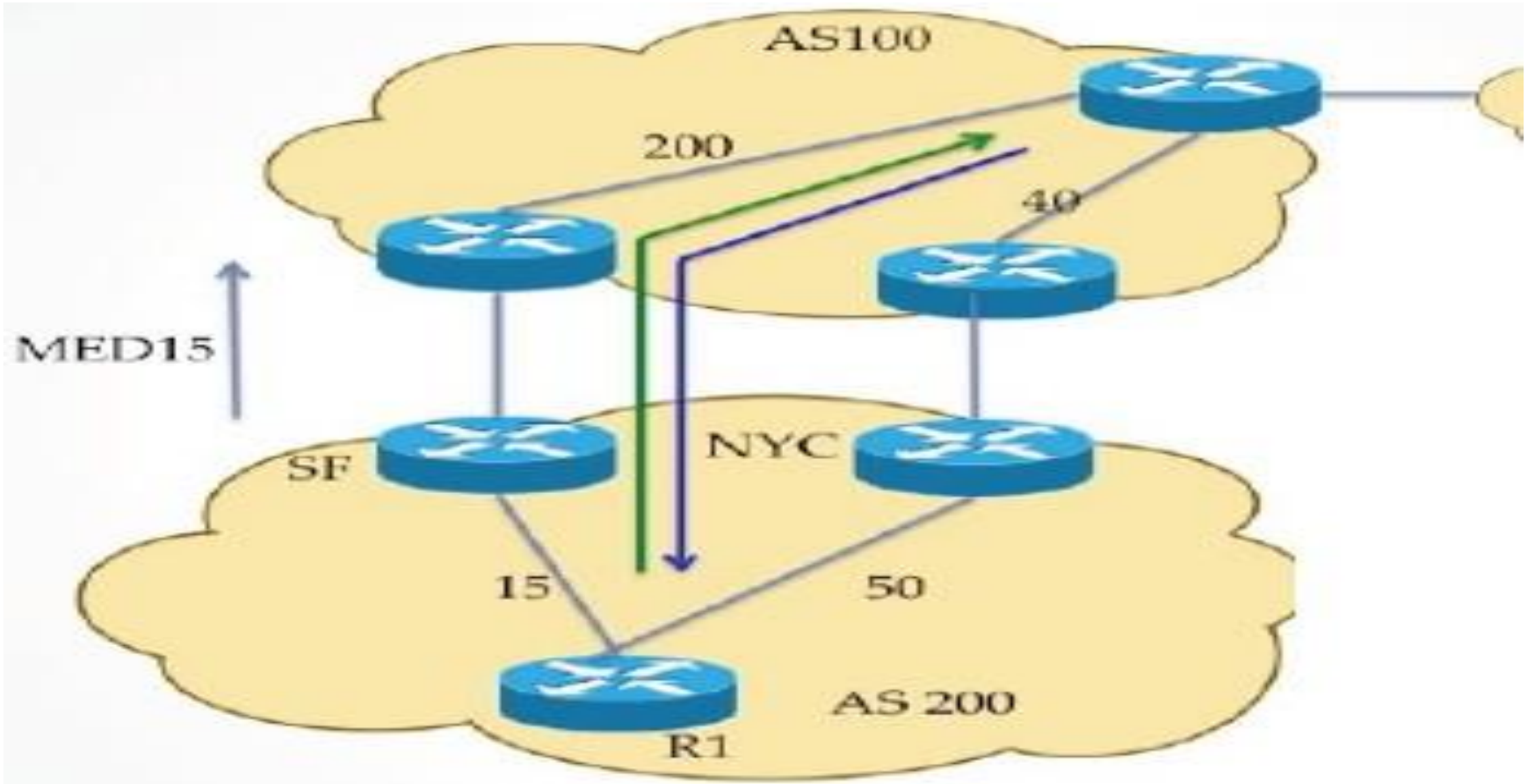
- Egress traffic from AS 200 is the green arrow in the below diagram, since SF path is shorter IGP distance. Ingress traffic to AS200 from AS 100 is the blue arrow, since NYC connection from AS100 shorter IGP distance (40 vs. 200)



➢ 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 ?

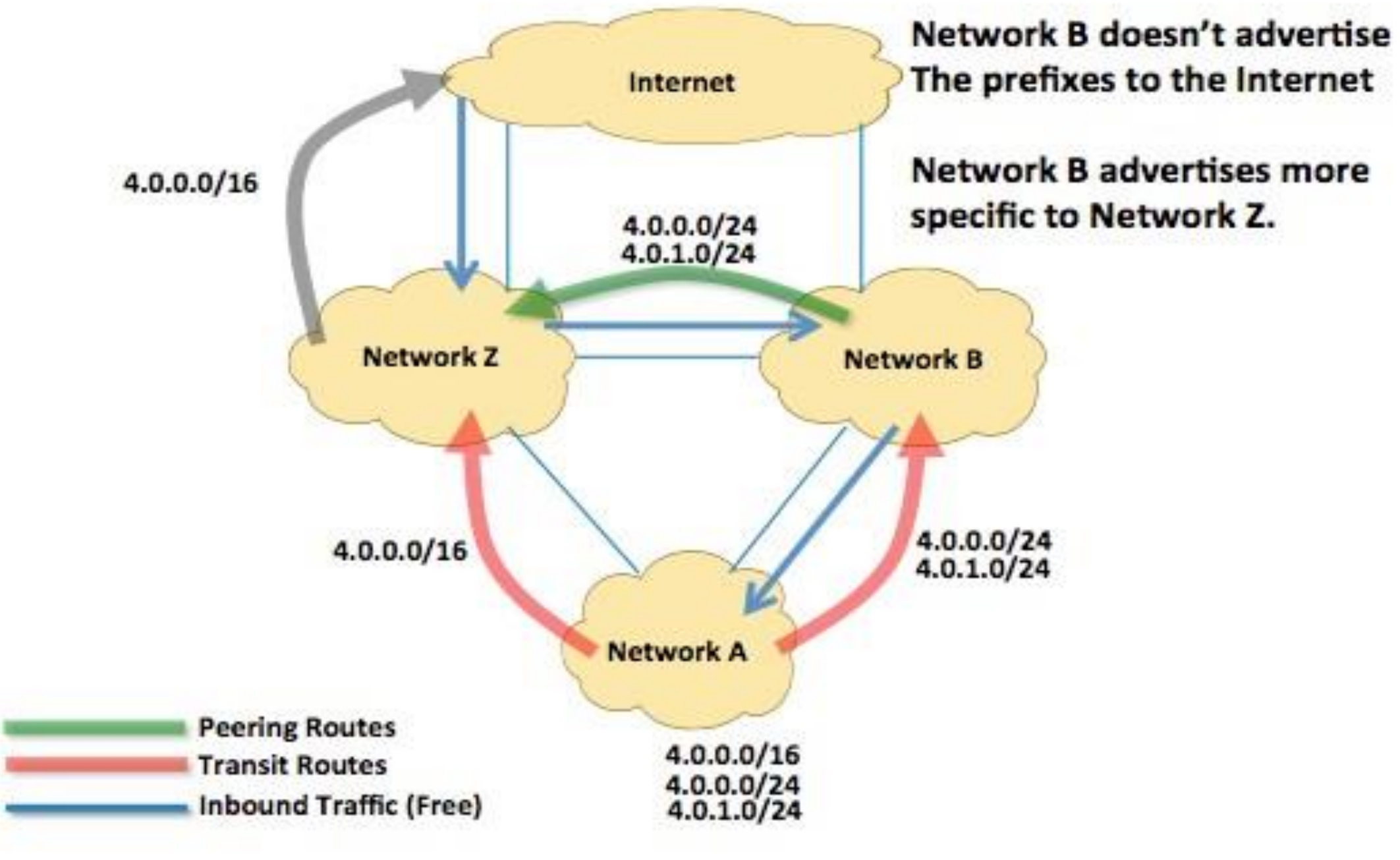
❖ Customer AS200 should force AS100 for cold potato routing. By forcing for cold potato routing ,AS 100 has to carry the Web content traffic to the closest exit point to AS200, which is San Francisco.

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



BGP Case Study – 3

- 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 ?



- 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.

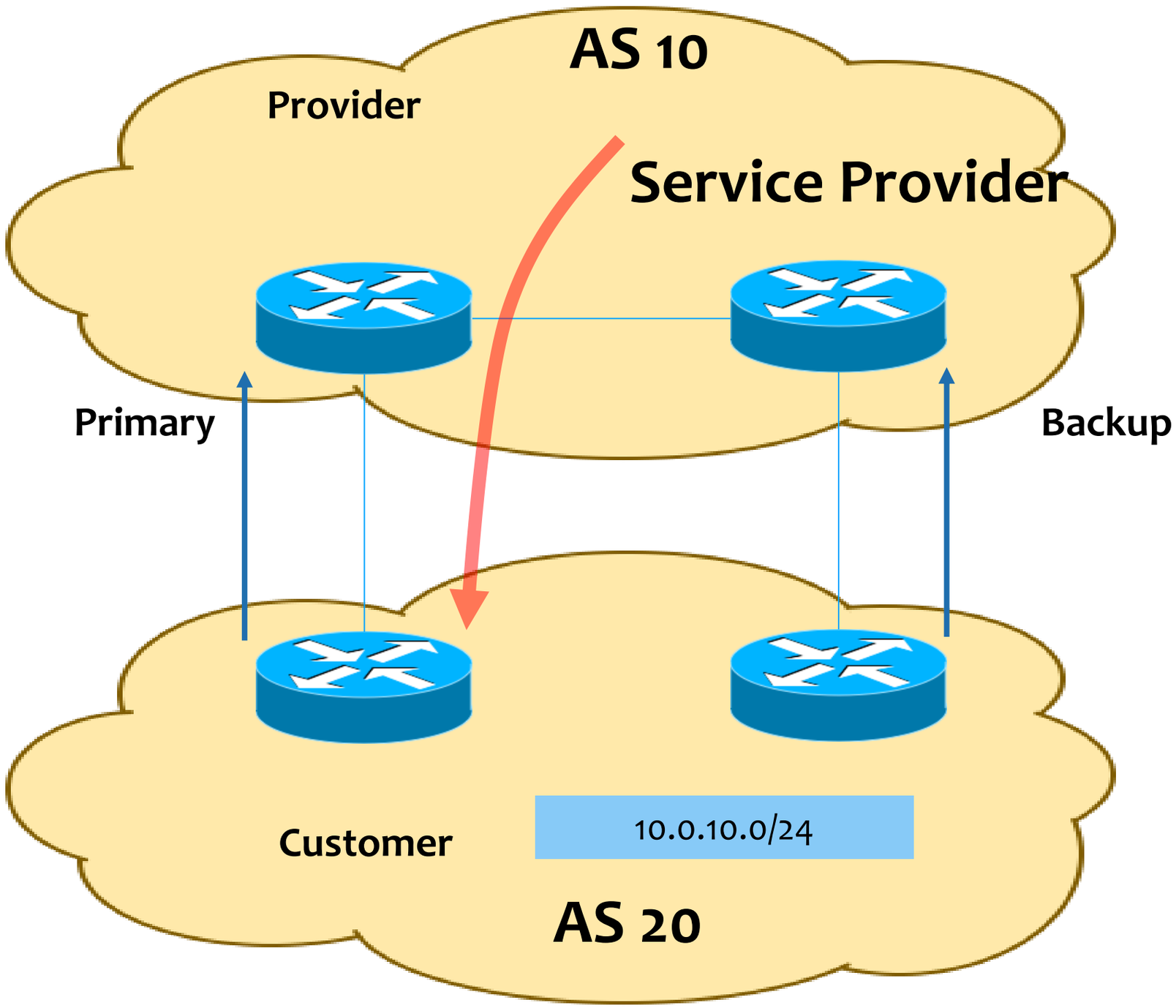
BGP Case Study – 4

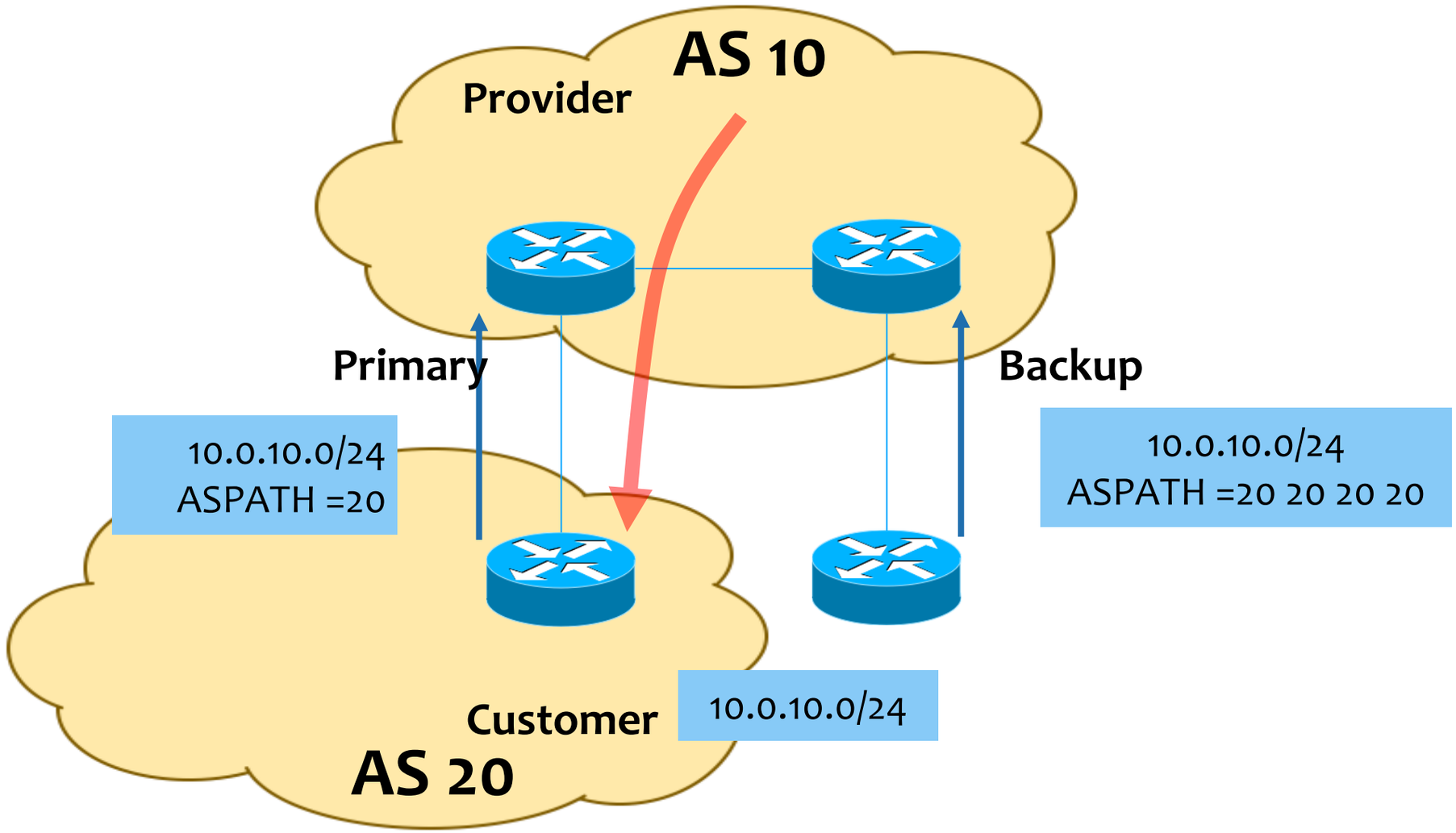
- Customer is running a BGP session with 1 service provider, they are considering to receive a transit service from the second Service Provider as well though

- 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 ?

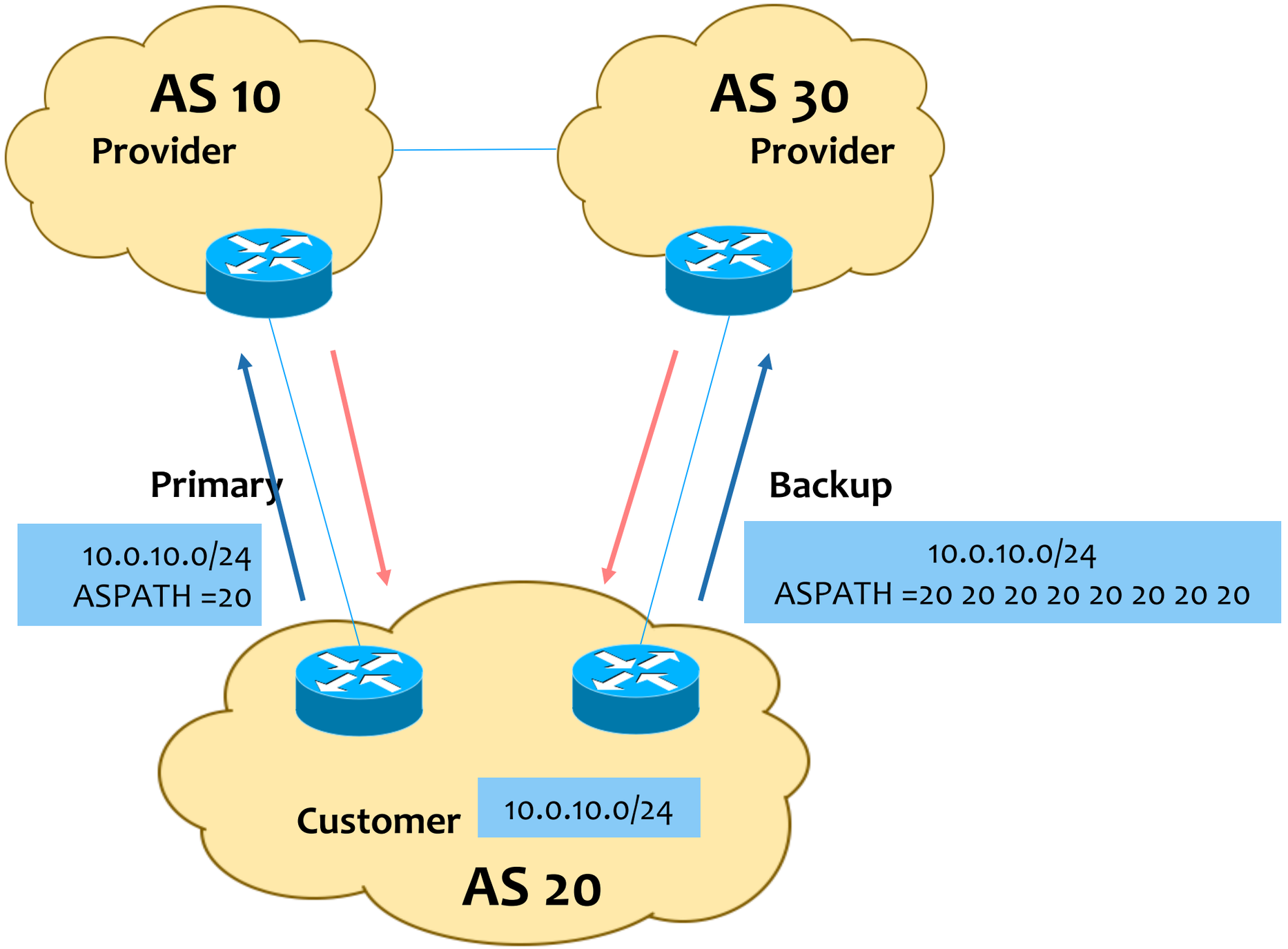




Prepending will (usually) force inbound traffic from AS 10 to take primary link.

- 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.

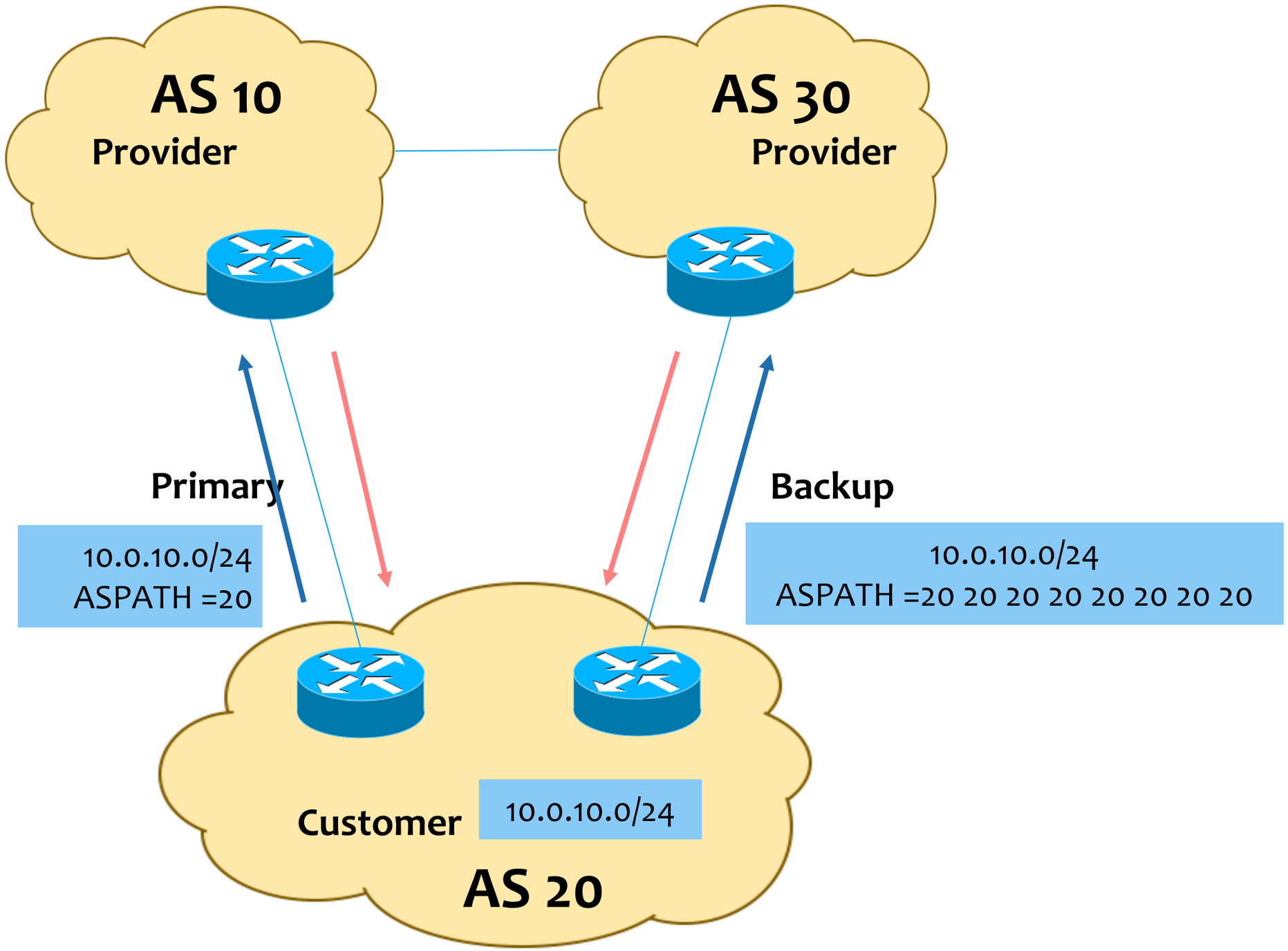
- They want to use second service provider link as backup link. They learned from the early experience the as-path prepending trick



- Question 3: Is there a problem with this design ?

- Yes
- No

Question 4 : What is the problem, how you can solve it ?

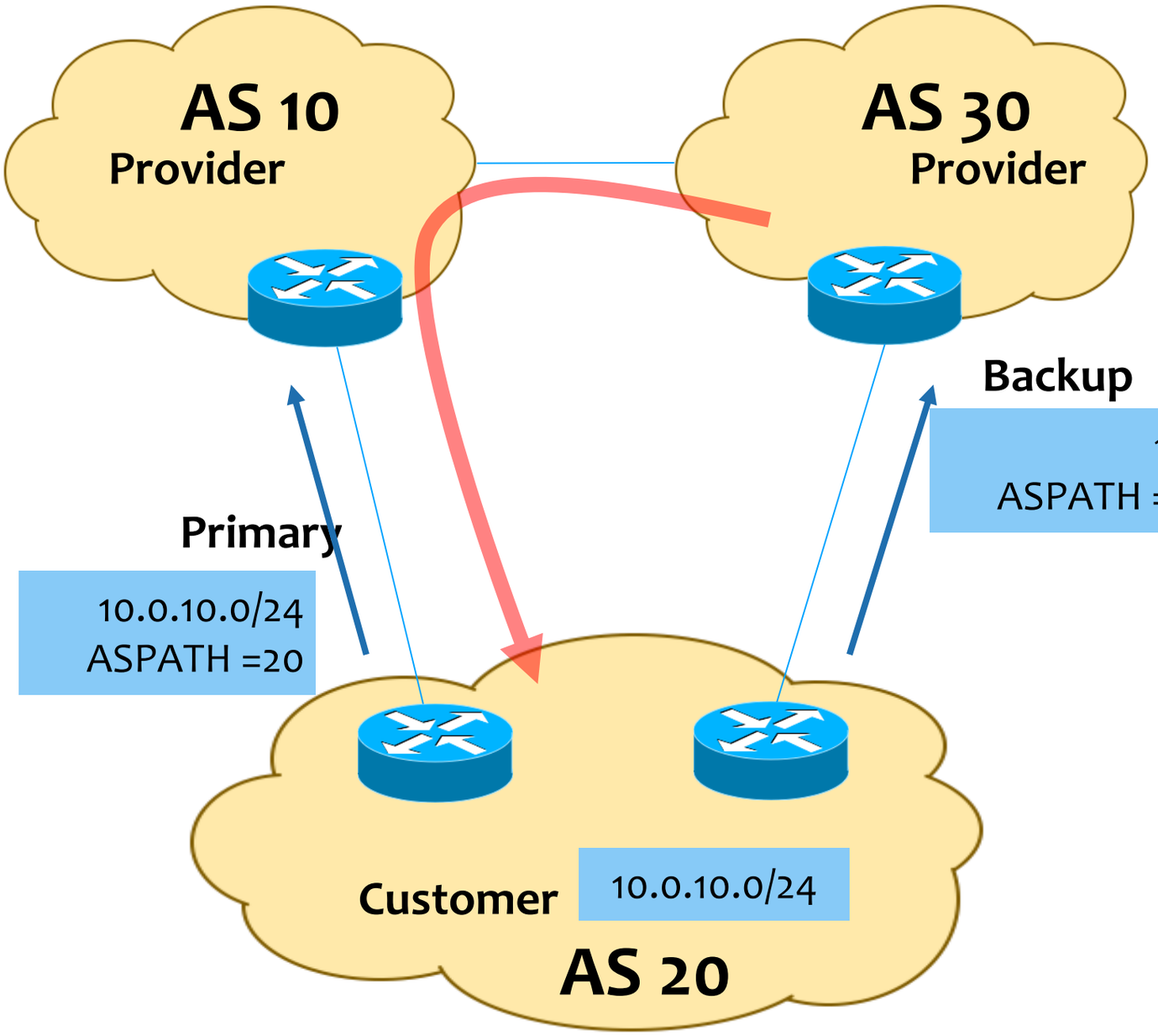


- There is a problem with the design since customer wants to use second service provider as backup. AS-Path prepending in this way is often used as a form of load Balancing

- 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



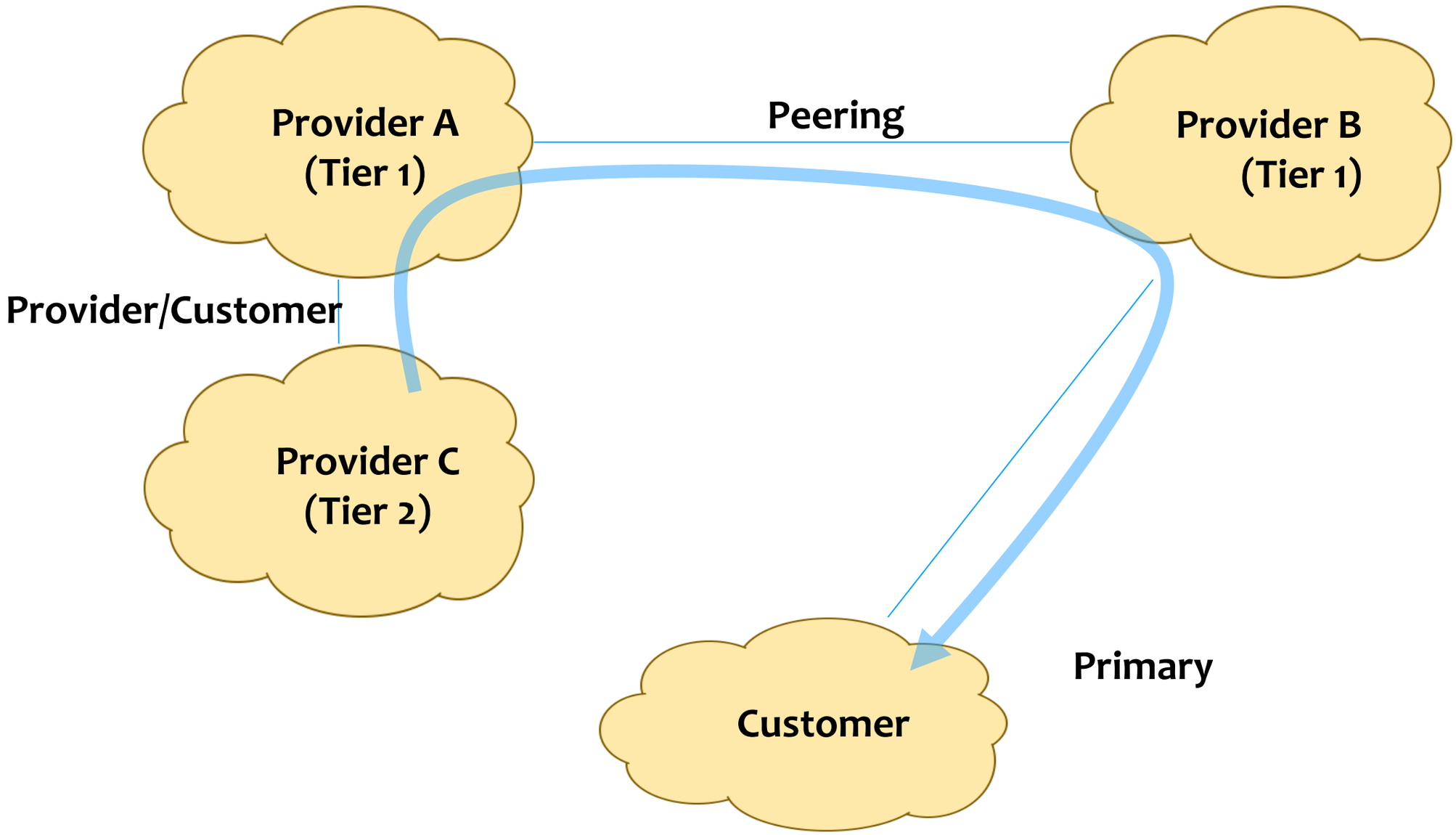
AS 30: Normal customer local pref is 100, peer local pref is 90

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

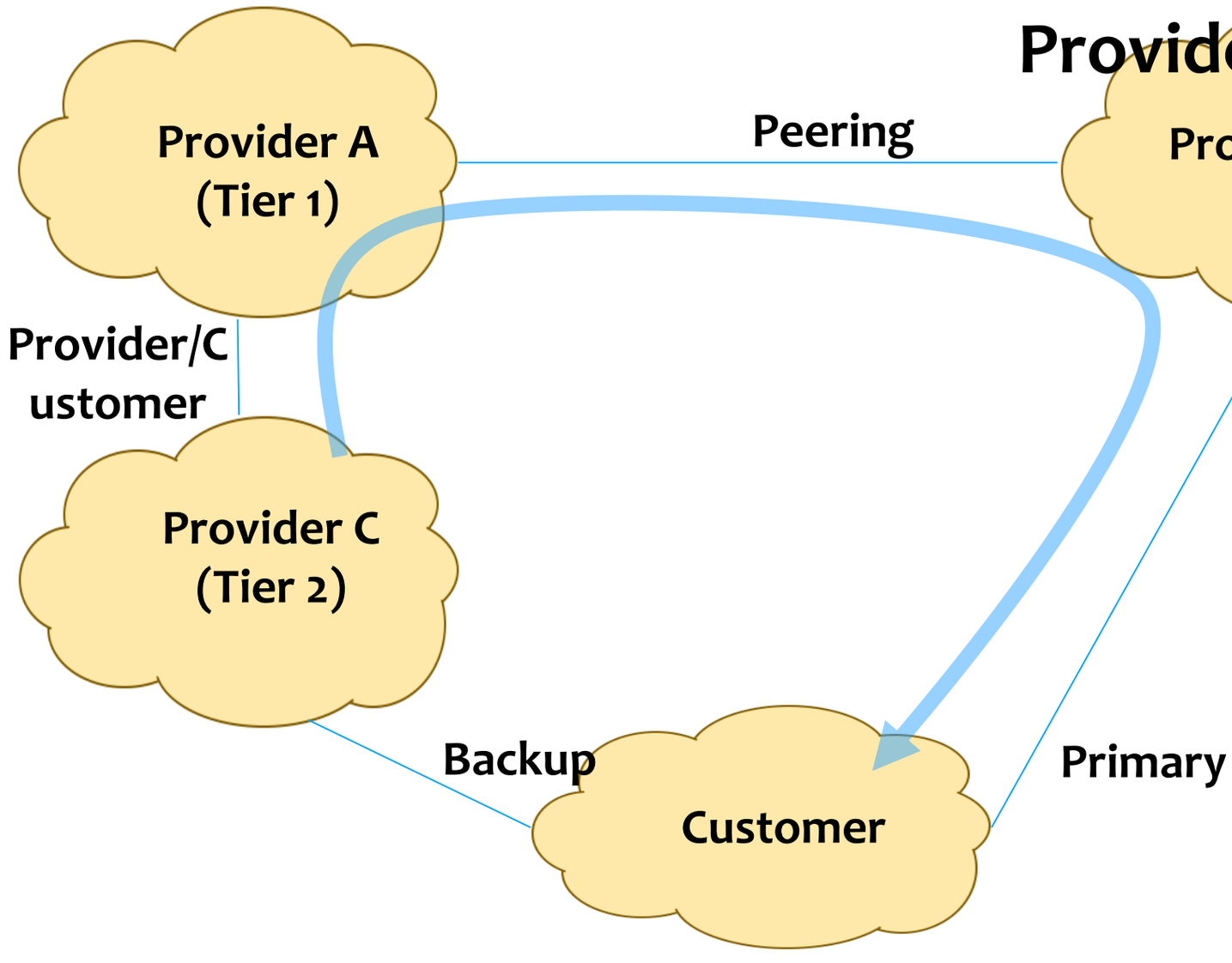
- Question 5: What if customer uses second service provider link as primary and the old provider secondary and the second provider peering connection as depicted in the below topology ?

- Does community help ?

Now, customer wants a backup link to C….

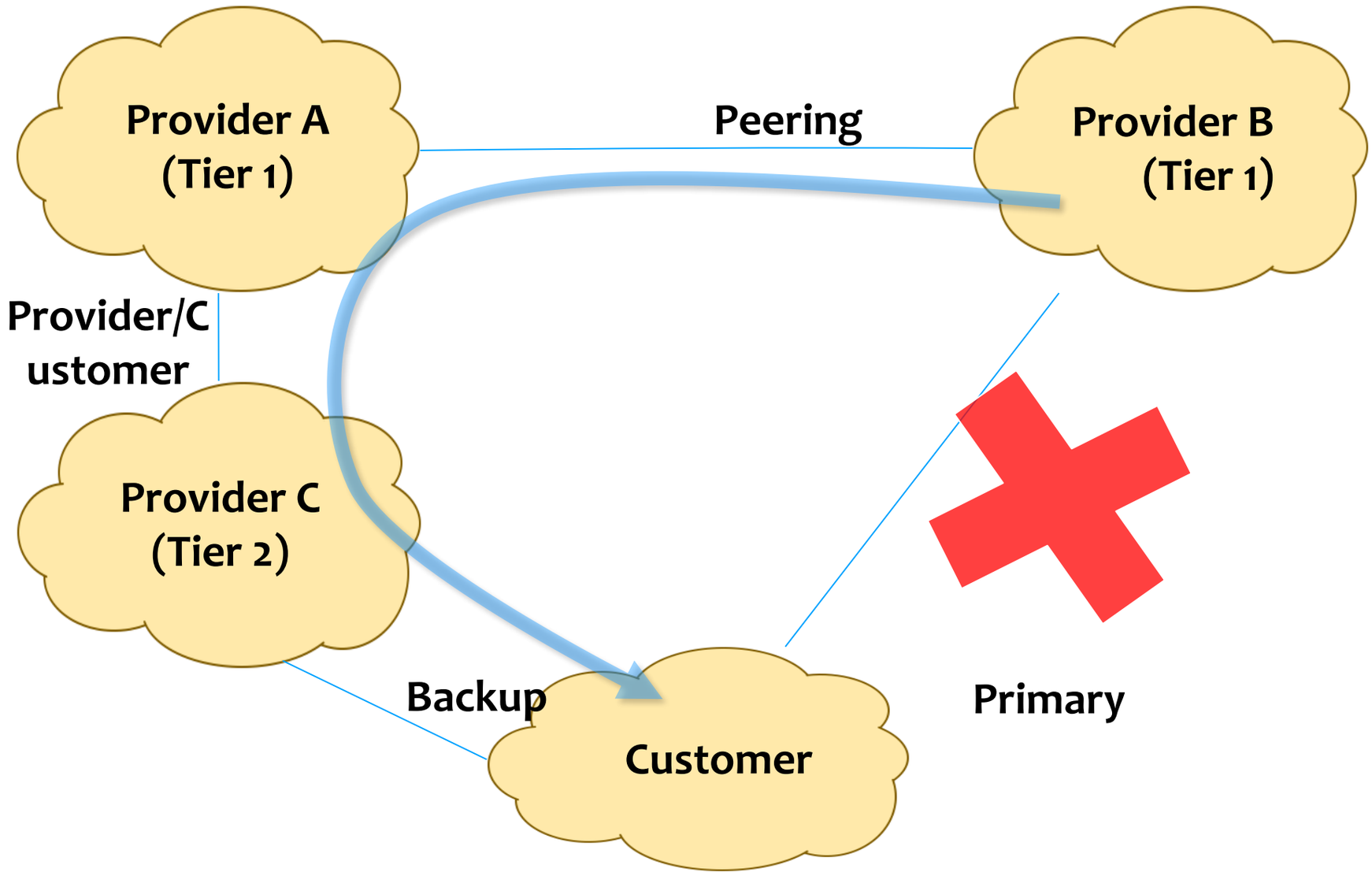


Customer sends community to Provider C, in order to use Provider B as backup



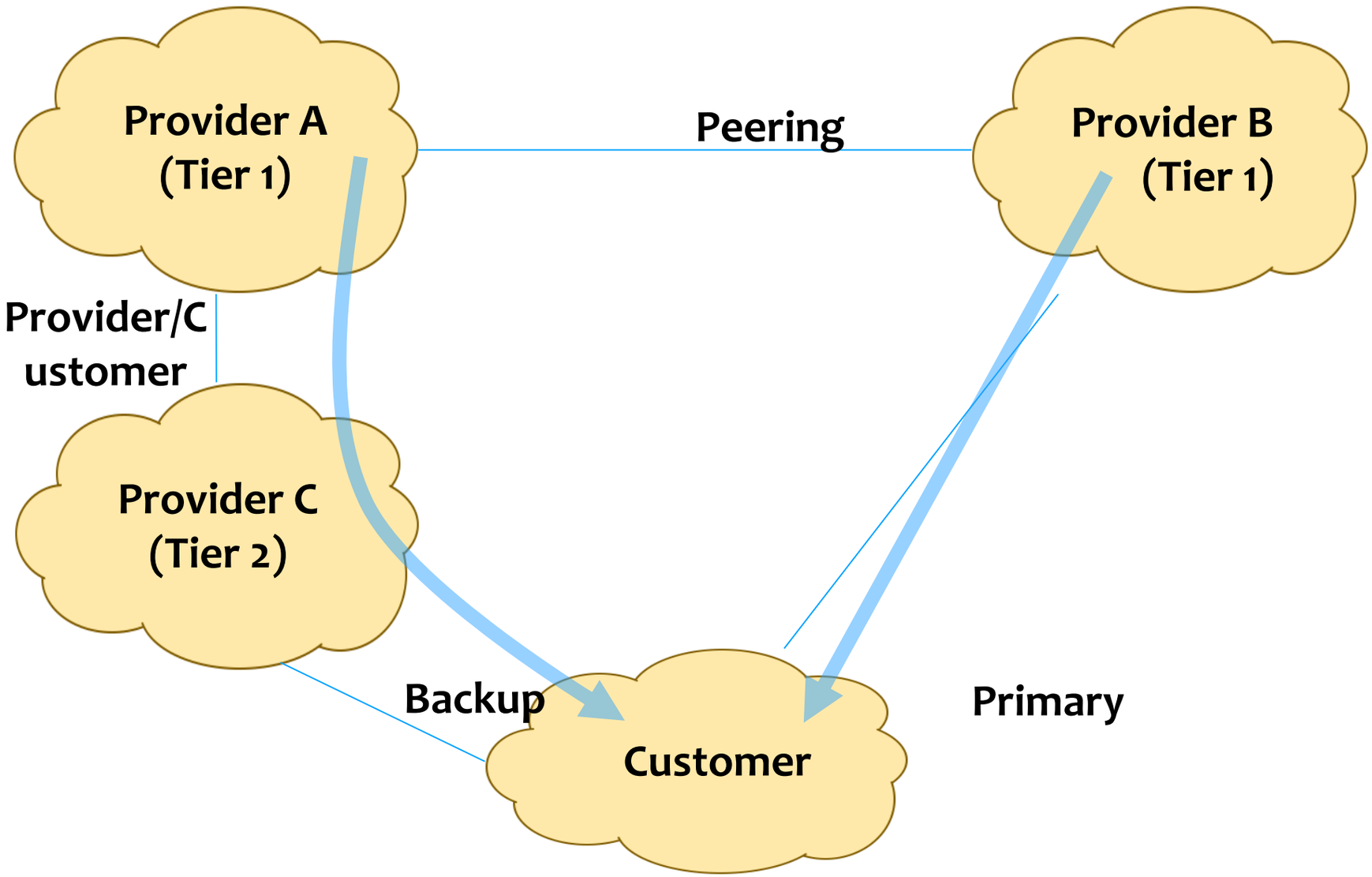
YES IT HELPS , NOW PROVIDER B CAN BE USED AS PRIMARY

- 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 ?



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

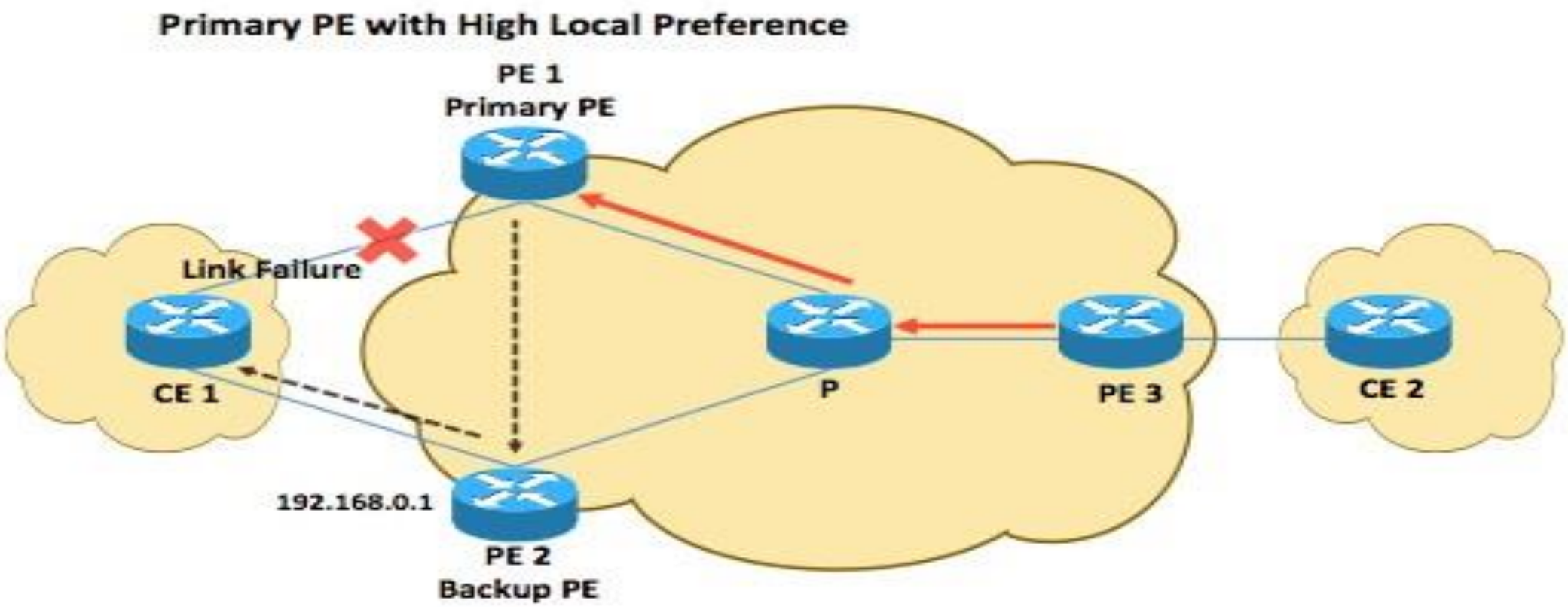
- Because Provider A continue to choose to send Provider C since the community attribute is sent by Customer to Provider C, not to Provider A

- 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 Case Study – 5

- Question : Where is BGP Best External feature used? What is the benefit of using it ?

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 Case Study - 6

- 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 ?

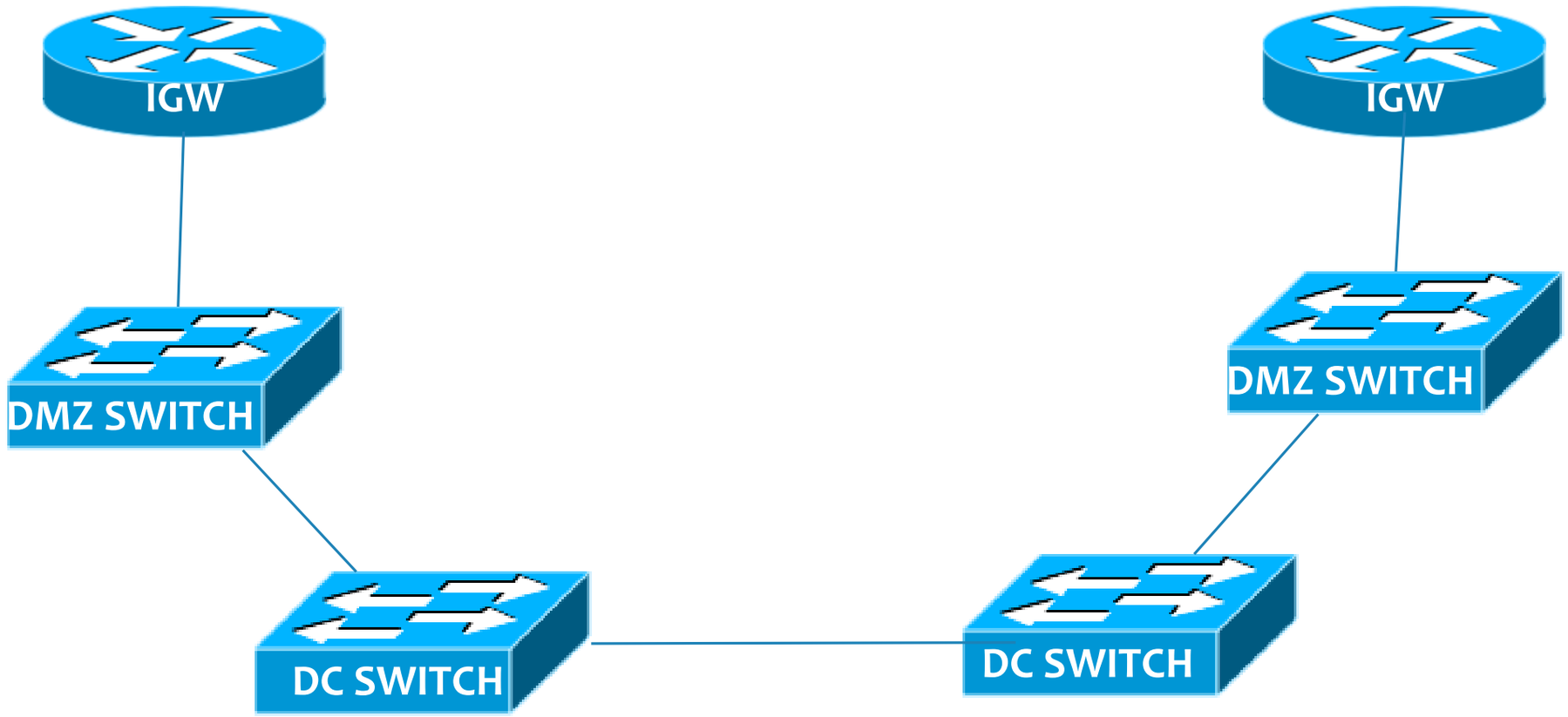
NO ! If RR is in the data path !.

Is there any Exception?

Yes, for example Seamless/Unified MPLS scenario

BGP Case Study – 7

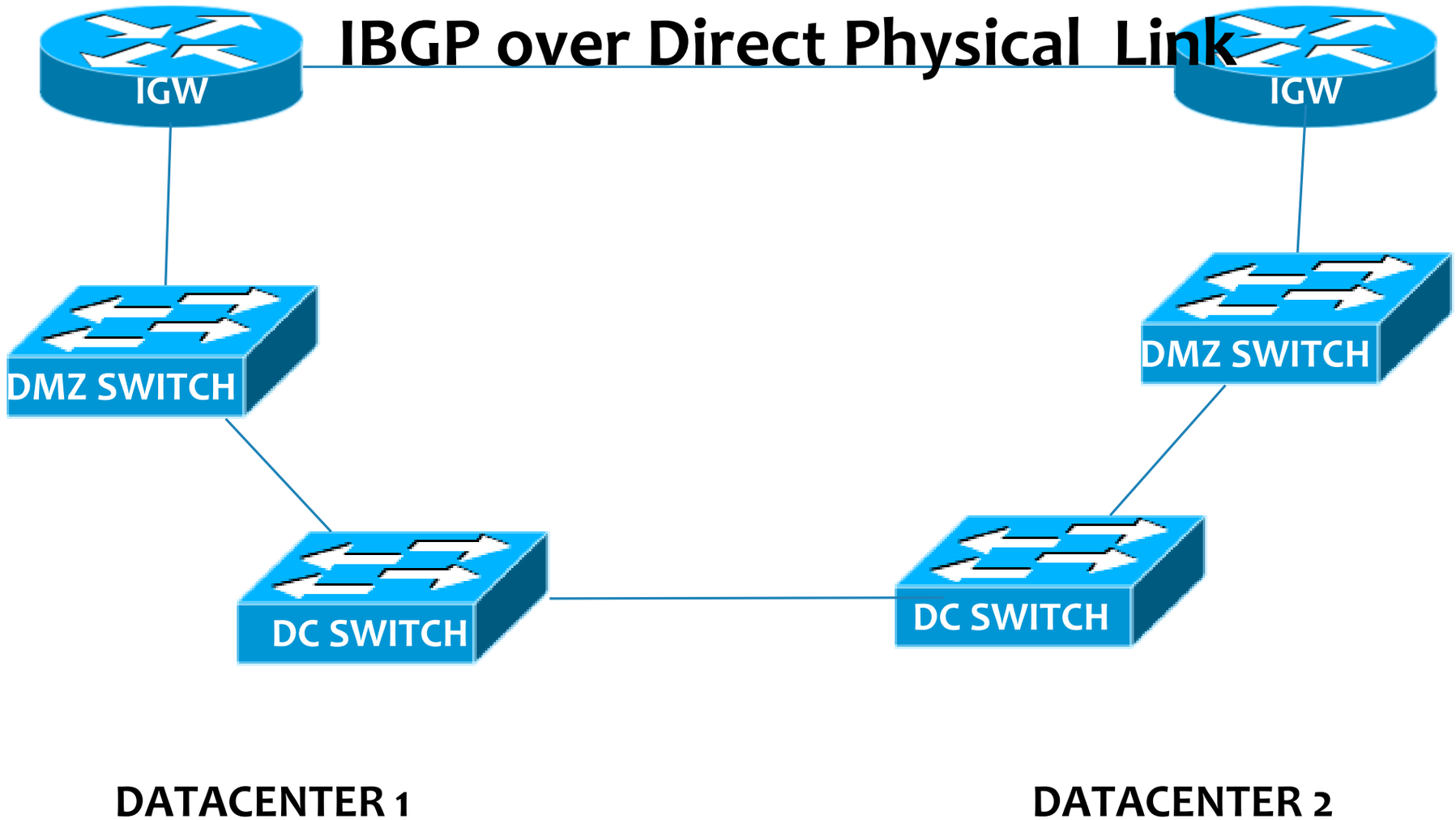
- Enterprise company has two datacenters. They have 200 remote offices and all the locations access to the internet from the Datacenters
- They recently had an outage on the internet circuit and all the Internet sessions from the remote offices which uses that link wad dropped
- What are the solutions to prevent the session failure in case of a link failure on the Internet Gateways of this company ?



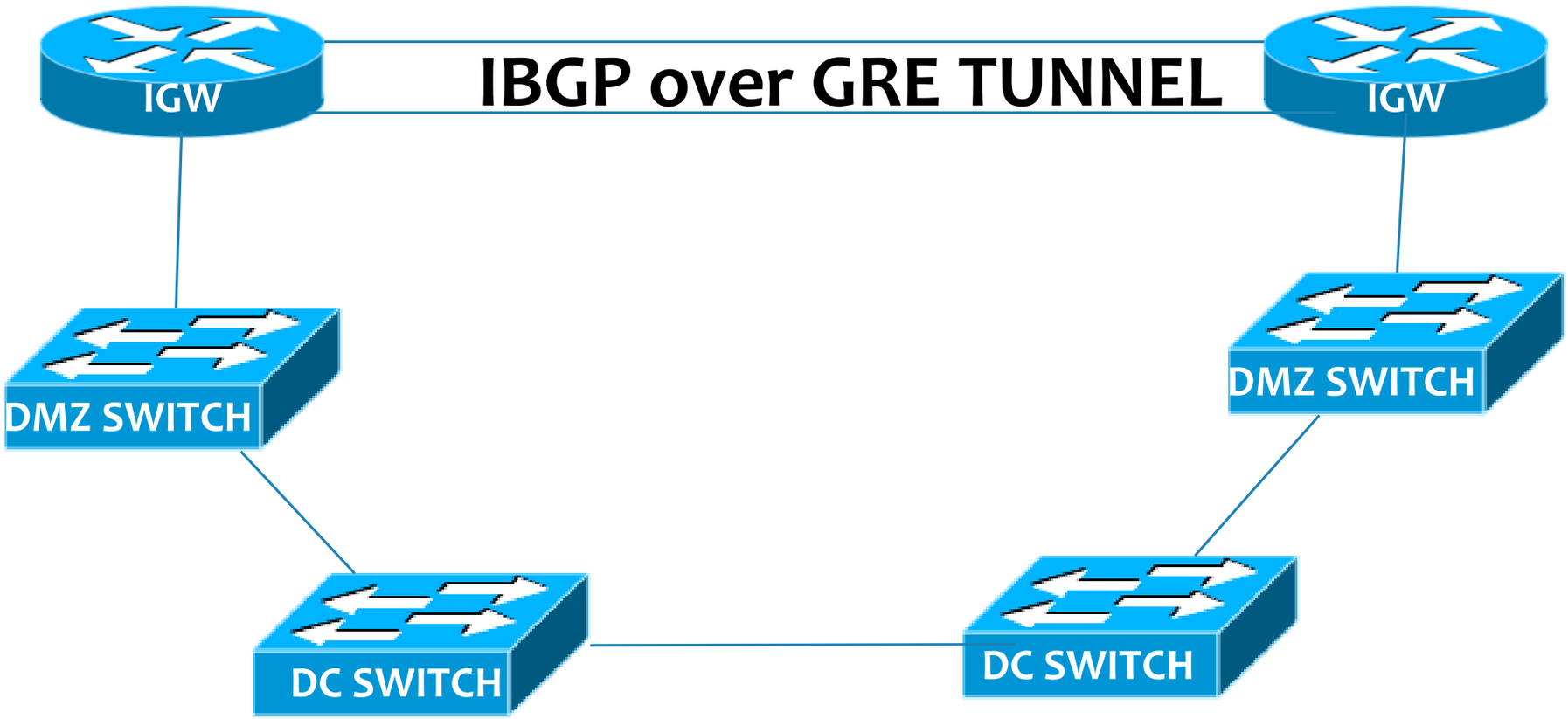
DATACENTER 1

DATACENTER 2

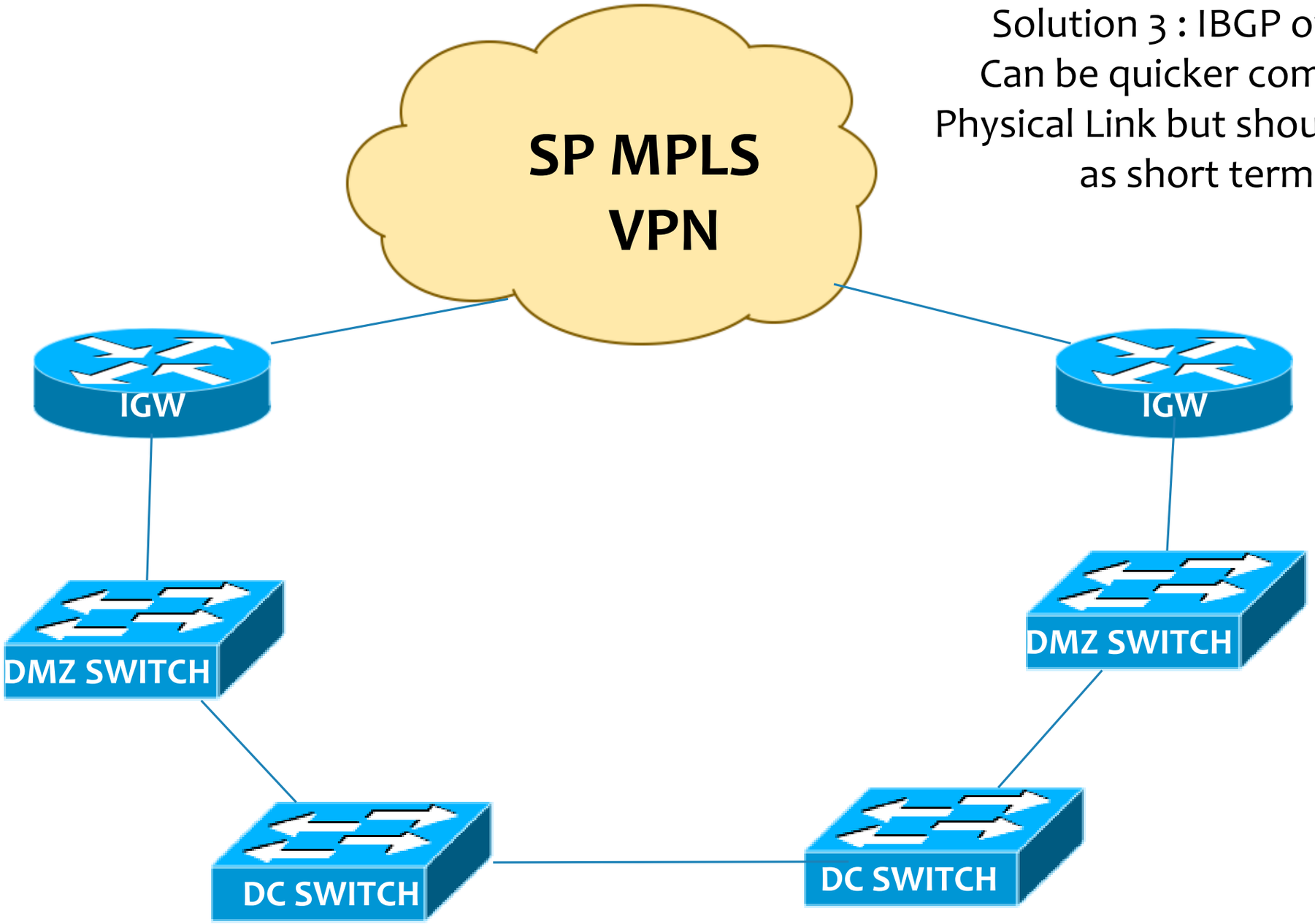
Solution 1 – Best option but can be costly. Budget might be concern, als
deployment might take longer compare to other solutions



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

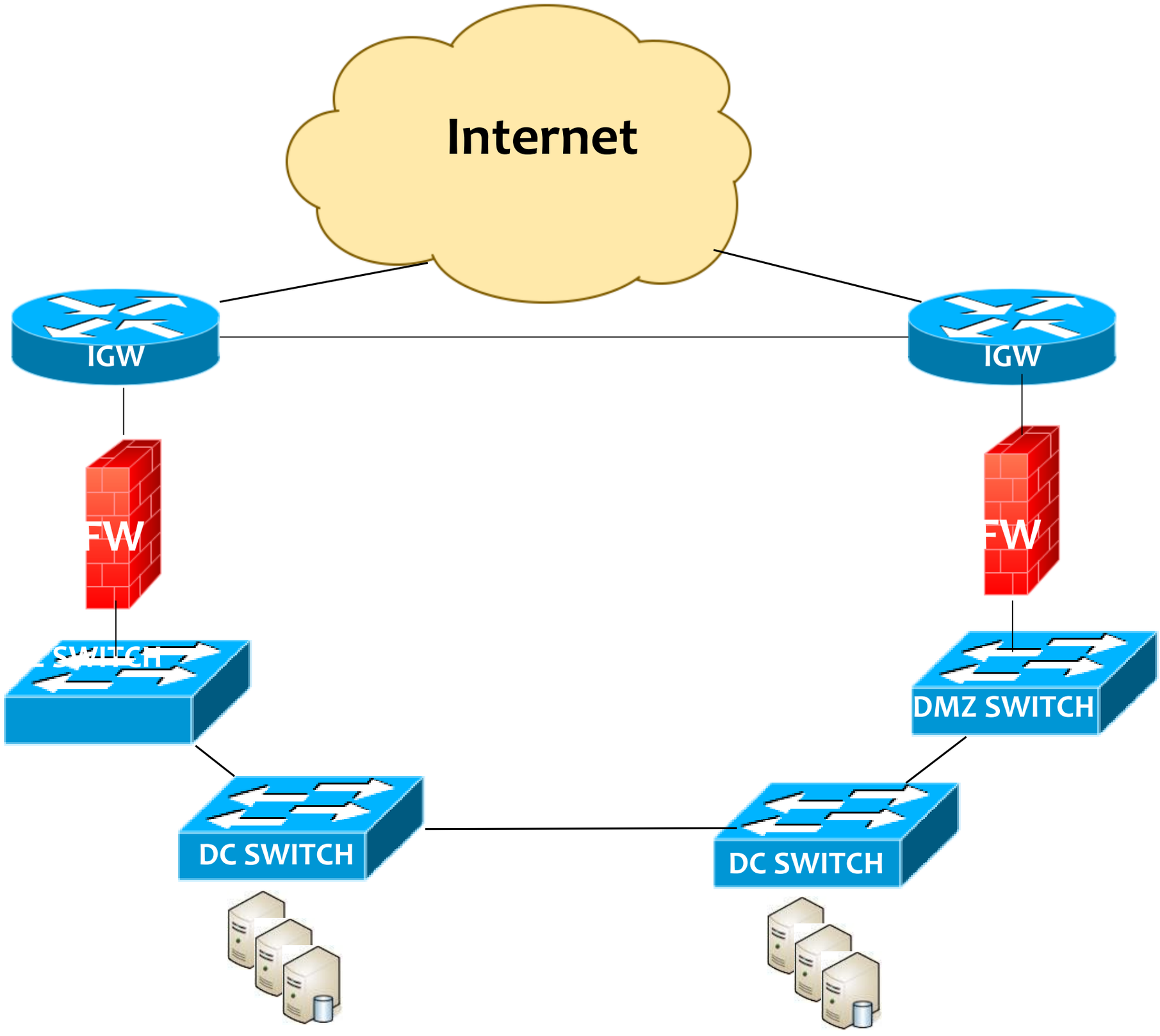


Solution 3 : IBGP over MPLS VPN
Can be quicker compare to Direct Physical Link but should be considered as short term solution



BGP Case Study 8

- 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 ?



- 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 – 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/

Multicast
Theory

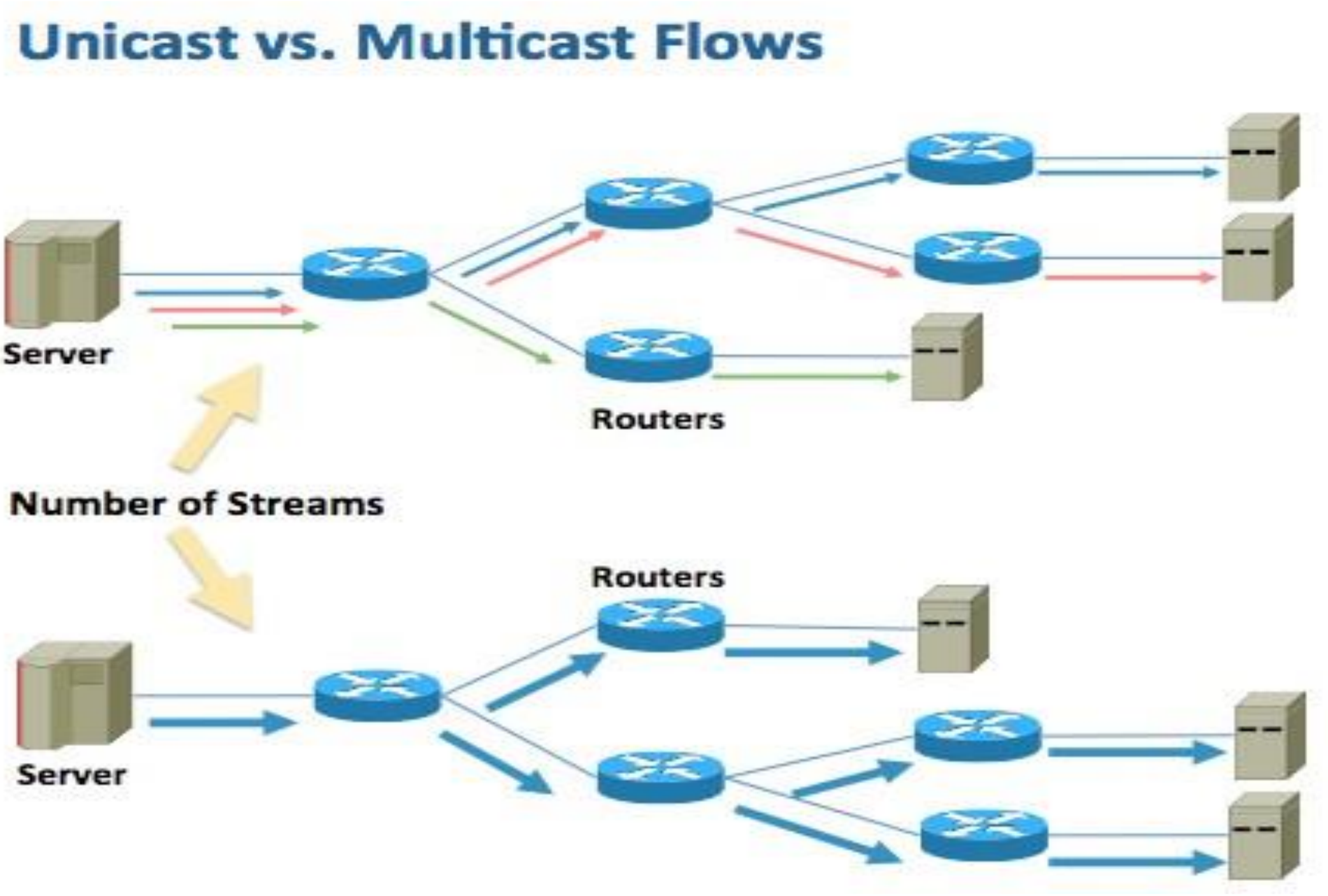
- If the requirement is to send the flow in real time to multiple receivers, then the efficient way of this is Multicast

- Multicast is a 30 years old protocol and still many people are struggling to understand it

- Probably the biggest problem with the understanding Multicast is Source Based Routing

- IP unicast routing works based on destination based routing, IP multicast routing works based on source based routing

- Tree in IP unicast routing is created from the source towards destination, in IP Multicast routing from destination(Receiver) towards source (Sender)

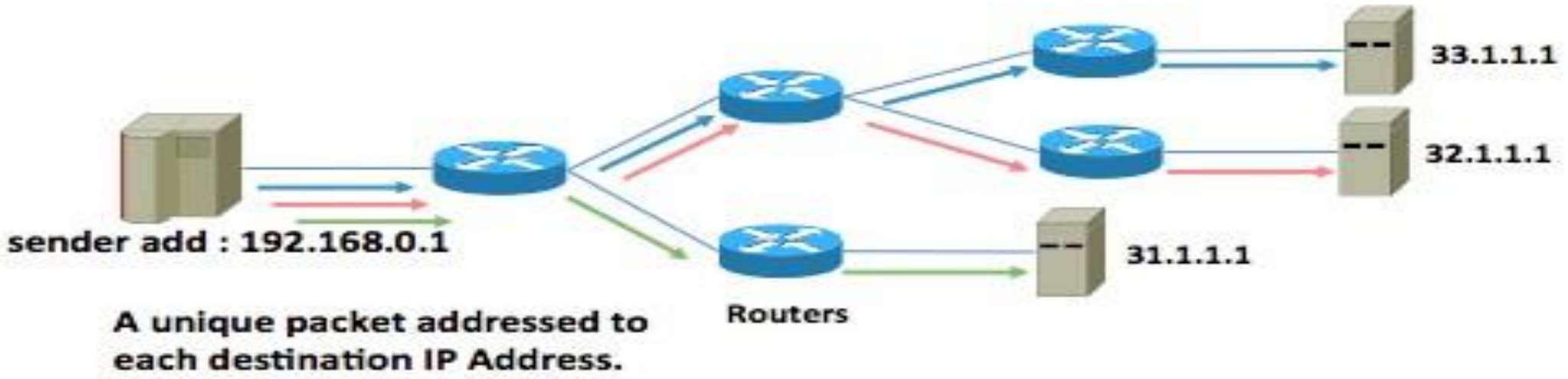


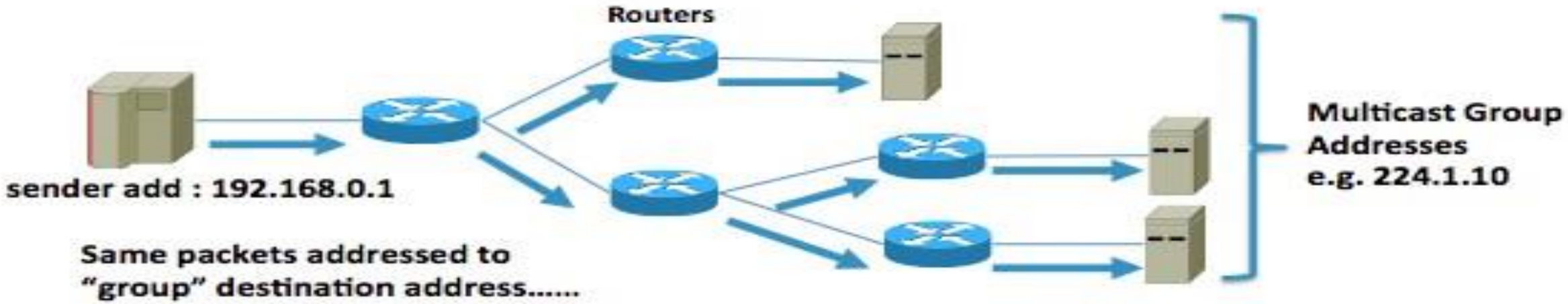
- Server has to send three copies of stream for three receivers in Unicast. Server just sends one copy and network replicates the traffic to their intended receivers in Multicast.

- Multicast works on UDP, not TCP. That’s why there is no error control, congestion avoidance, it is purely best effort

- Receiver can receive duplicate Multicast traffic in some situations. SPT switchover is the example where duplicate traffic delivery occurs. During a SPT switchover, multicast traffic is received both from Shared tree and Shortest path tree.
- This is one of the inefficiency of Multicast

Unicast and Multicast Addressing





- Source addresses in Multicast always unicast address. Multicast address is a class D address range. 224/4. You should never see Multicast Class D address as a source Multicast address.

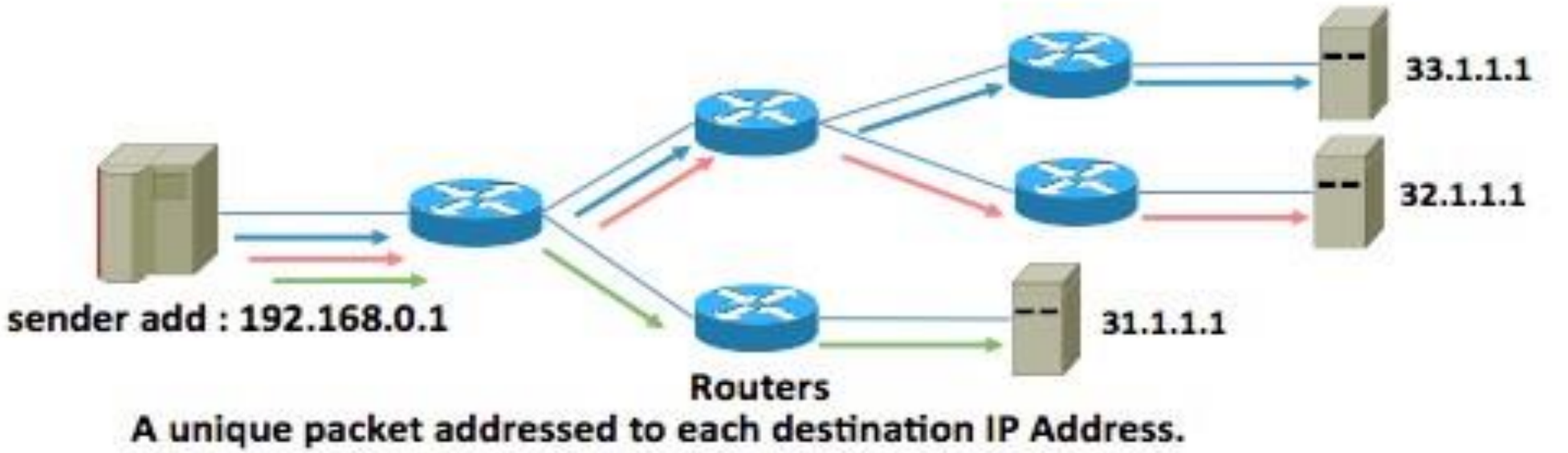
- Source address can never be class D multicast group address. Separate multicast routing table is maintained for the multicast trees. Source don’t need to join any group, they just send the traffic.

- Multicast routing protocols (DVMRP, PIM) is used to build the trees. Tree is built hop by hop from the receivers to the source. DVMRP was the first Multicast Routing Protocol and it is depreciated and not used anymore

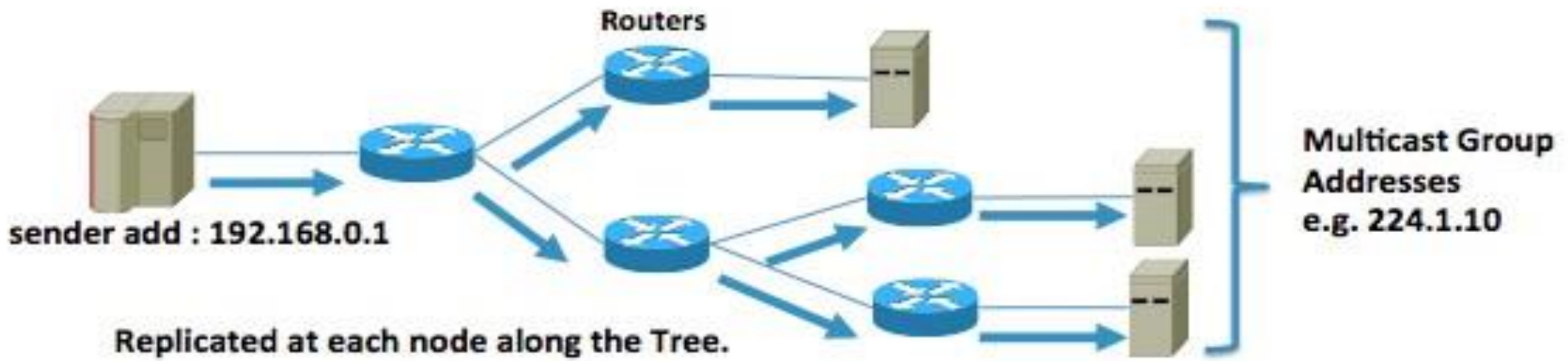
- Source (Sender) is the root of the tree in shortest path tree.

- Rendezvous Point is the root of the Shared Tree (Rendezvous Point is used in ASM and PIM Bidir)

Unicast and Multicast Addressing



A unique packet addressed to each destination IP Address.



Replicated at each node along the Tree.

- Link local addresses 224.0.0.0 -224.0.0.255.
- TTL of link local address multicast is 1. They are just used in the local link. OSPF, EIGRP etc uses addresses from this range.
- IANA reserved address scope ; 224.0.1.0 – 224.0.1.255
- This address range is used for the networking applications such as NTP, 224.0.1.1 TTL is greater than 1.

- Administratively scoped multicast addresses: 239.0.0.0 – 239.255.255.255. This address range is reserved to be used in the domain. Equivalent to RFC 1918 private address space. There is 32/1 Overlapping between IP Multicast IP and Mac Addresses.

✓ 224.1.1.1

✓ 224.129.1.1

✓ 225.1.1.1

✓ 238.1.1.1

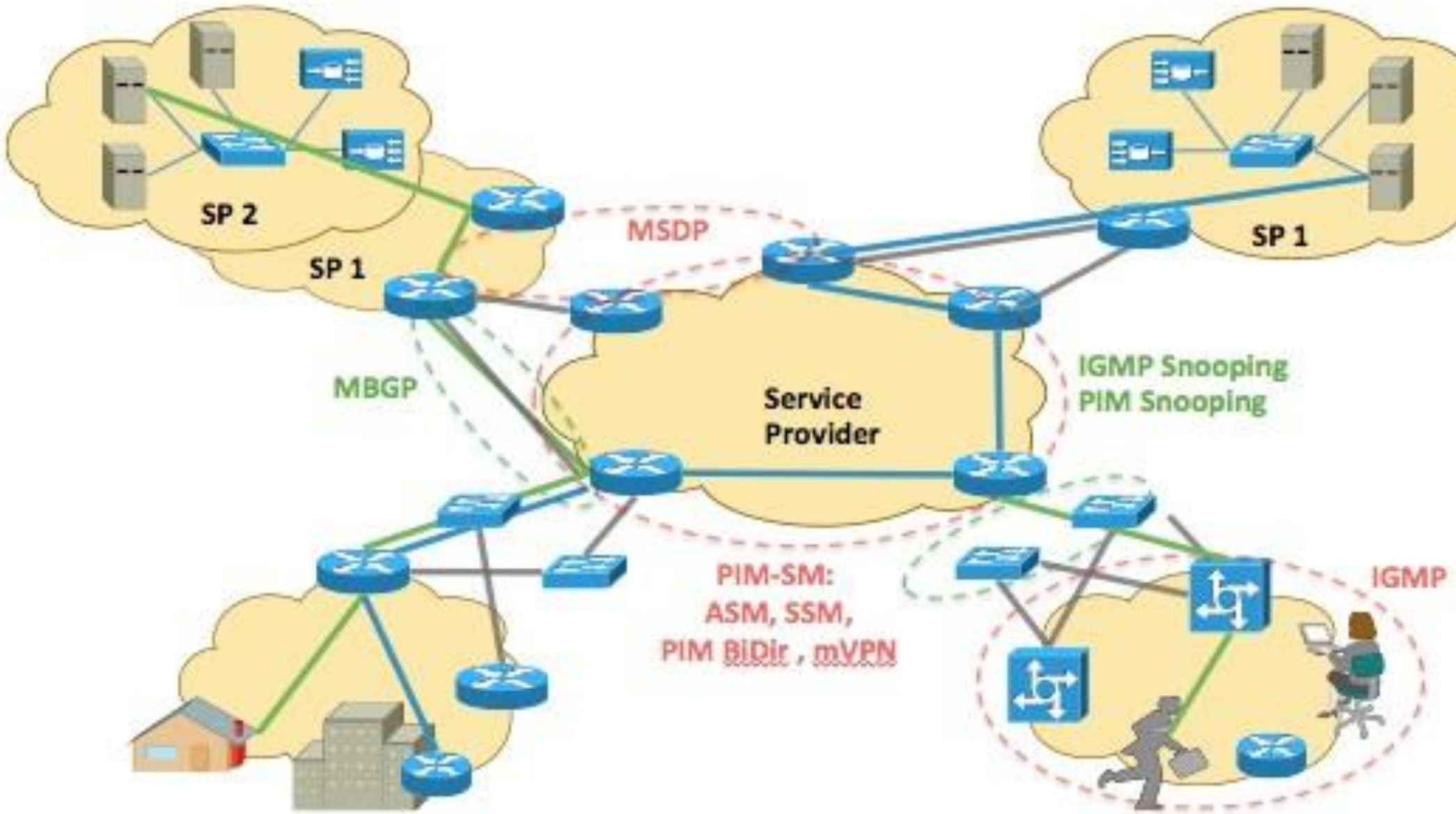
✓ 238.129.1.1

✓ 239.1.1.1

✓ 239.129.1.1

- All above address uses same multicast MAC address, which is 0100.5e01.0101

We will talk about below Multicast Protocols



- Receiver uses IGMP to communicate with the router. IGMP v2 is used in PIM ASM (Any Source Multicast) and only IGMPv3 can be used with PIM SSM (Source Specific Multicast)

- IGMP is a host to router protocol. After first hop router you don’t see IGMP messages. IGMP Membership report which is also known as IGMP join is sent by the host application to the first hop router, then first hop router creates a PIM Join towards the root of the tree

- Receiver sends an IGMP Membership Report to the first hop router which in return sends PIM Join to the root of the tree.

- By default switch sends Multicast traffic to the every ports. Even though un interested hosts receive the multicast packets. This is not efficient. That’s why IGMP snooping is used to help for multicast scaling.

- When IGMP snooping is enabled, switch tracks the multicast traffic and when the traffic comes from the upstream router, switch sends the traffic to the interested receivers.

- Switches with IGMPv2 has to track every multicast packet to see catch the control plane traffic. Every multicast data packet is inspected with IGMPv2, since there is no IGMPv2 router multicast group address. That’s why in IGMPv2, IGMP snooping should be enabled on the hardware otherwise huge performance impact can be seen.

- This problem is solved in IGMPv3.IGMPv3 uses special 224.0.0.22 Multicast group address. Switch only tracks these multicast control plane packets. That’s why,IGMPv3 provides scalability. Even in the software platform, IGMP snooping can be used if IGMPv3 is enabled.

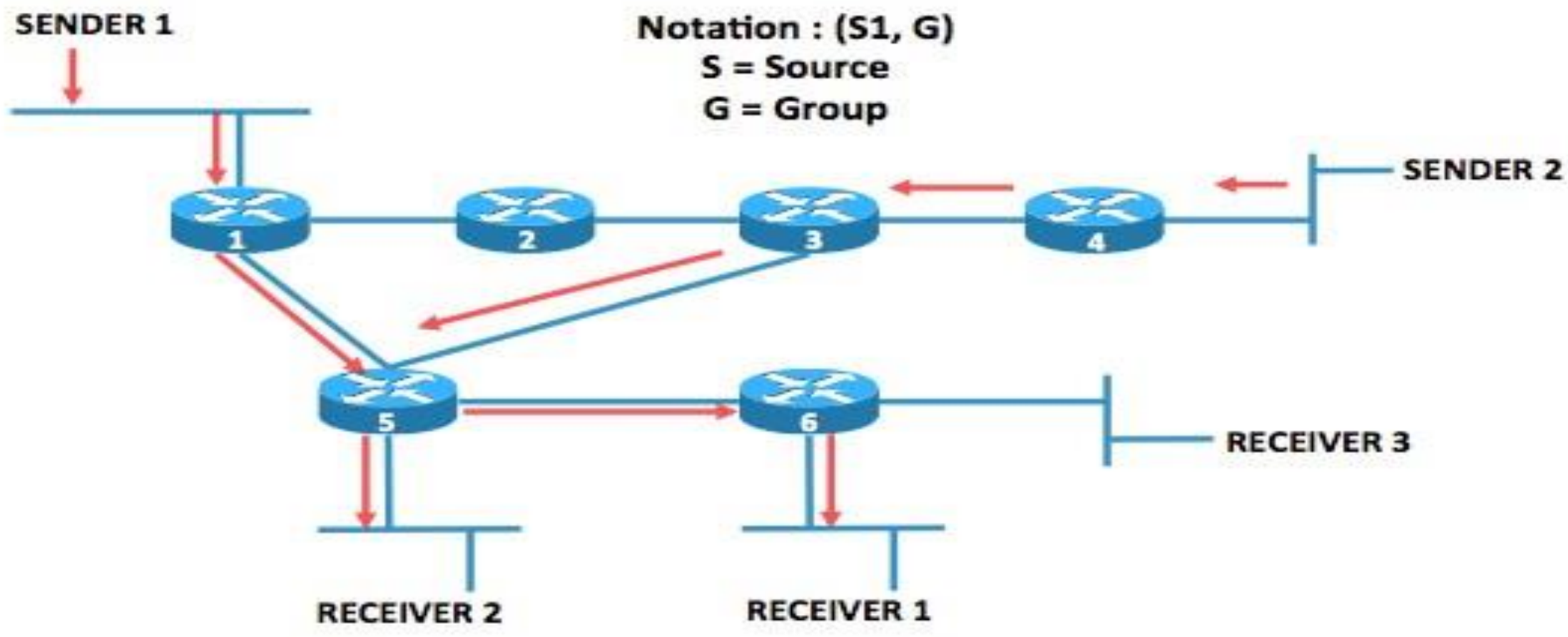
Multicast Distribution Trees

- Shortest path tree or Source tree is created between the source and the receiver in PIM SSM.

- Shortest path tree is created between the source and the Rendezvous point in PIM ASM before the SPT (Shortest Path Tree) switchover.

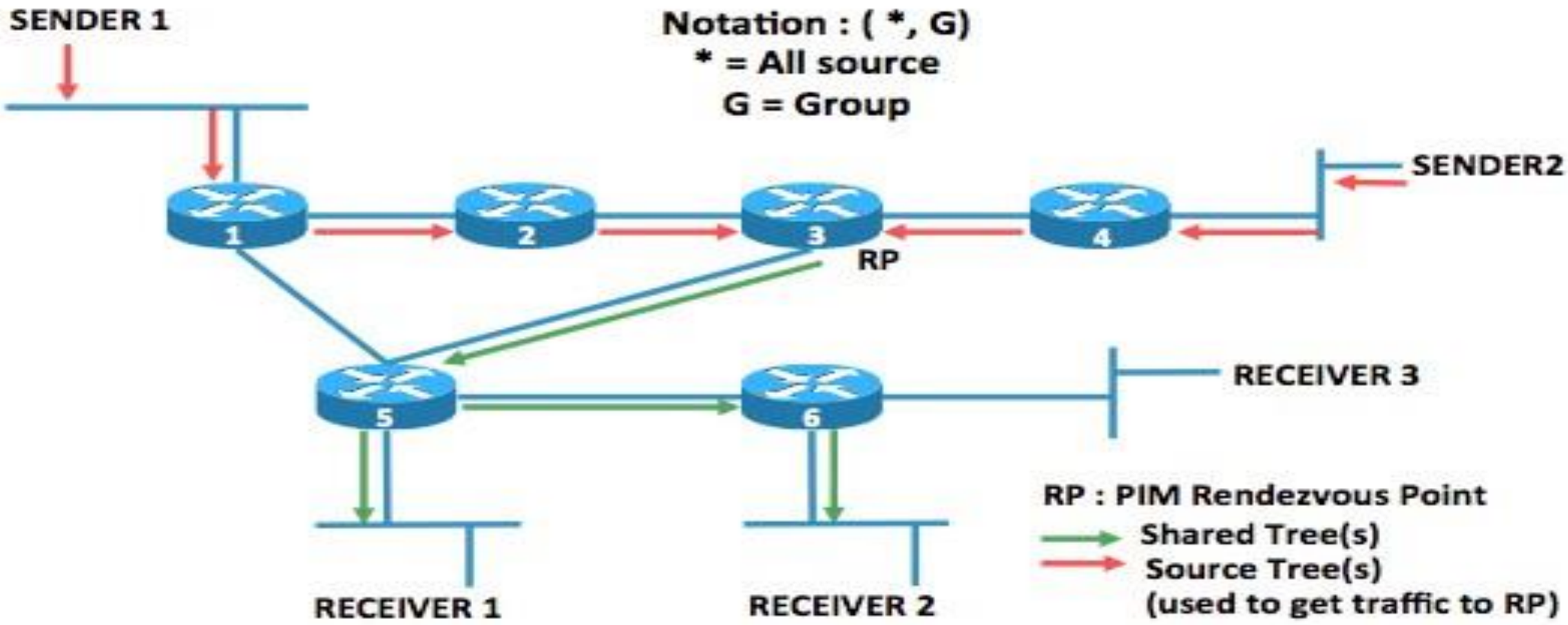
Short Path or Source Distribution Tree

There is no Rendezvous Point , No Rendezvous Point Engineering !



- Shared tree is created between the Rendezvous Point and the receiver only in PIM ASM, still there is shortest path tree between the Source and Rendezvous point in PIM ASM.

Shared Tree



- Shortest path tree uses more memory but provides optimal path from the source to all receivers. That's why minimizes the delay.

- Shared tree uses less memory because you don't have separate multicast state for each source for the given multicast group address but may create suboptimal routing for some receivers. That's why shared tree may introduce extra delay.

PIM Protocol Independent Multicast

- PIM is a multicast routing protocol. Two PIM mode, PIM sparse and PIM dense mode.

PIM Dense Mode:

- PIM dense mode is a flood and prune based (pushing) protocol. It consumes extra resources (Bandwidth, CPU and memory). Even though there is no intended receivers, multicast traffic is sent everywhere. Then if there are no receivers, routers prune the traffic. But flood and prune mechanism is repeated every so often.

PIM Sparse Mode

- Most commonly used PIM mode is the PIM Sparse Mode. Works based on pull mechanism unlike PIM Dense mode which is push based mechanism.
- Receivers in PIM Sparse mode request to join to a specific multicast group. Tree is created from the receivers up to the root.
- Root is a sender if the tree is shortest/source based, it is a rendezvous point if the tree is shared tree.

- PIM sparse mode can be implemented in three ways. PIM ASM (Any source multicast) , PIM SSM (Source Specific Multicast) and PIM Bidir (Bidirectional Multicast).

- If the source is known then there is no need for PIM ASM.

- If source is not known, there is a common element which is called Rendezvous Point in PIM ASM for source registration. All the sources are registered to the PIM Rendezvous Point. Receiver join information is sent to the Rendezvous point as well.

- It can be thought as dating service. Meets the receivers and senders (Source). Since PIM ASM requires Rendezvous Point and RP engineering, it is considered as hardest multicast routing protocol mode.

- The default behavior of PIM ASM is that routers with directly connected members will join the shortest path tree as soon as they detect a new multicast source.

- Rendezvous point information can be configured as Static on the every router, it can be learned through Auto-RP which is Cisco specific protocol or BSR which is IETF standard.

- There is no Auto-RP anymore in IPv6 Multicast.

PIM SSM – Source Specific Multicast

- PIM SSM is a source specific multicast. Don’t require rendezvous point. Because sources are known by the receivers. Receivers create a shortest path tree towards the source Root of the tree is the sender.
- If SSM is enabled, only (S,G) entries are seen in the Multicast Routing table.

- IANA reserves 232/8 Class D Multicast Address range for the PIM SSM but it’s not must to use these address range though.

- Require IGMPv3 at the source or IGMPv2 to v3 mapping on the first hop routers.

- Source specific multicast is most suitable for the one to many applications. IPTV is an example of one to many applications.

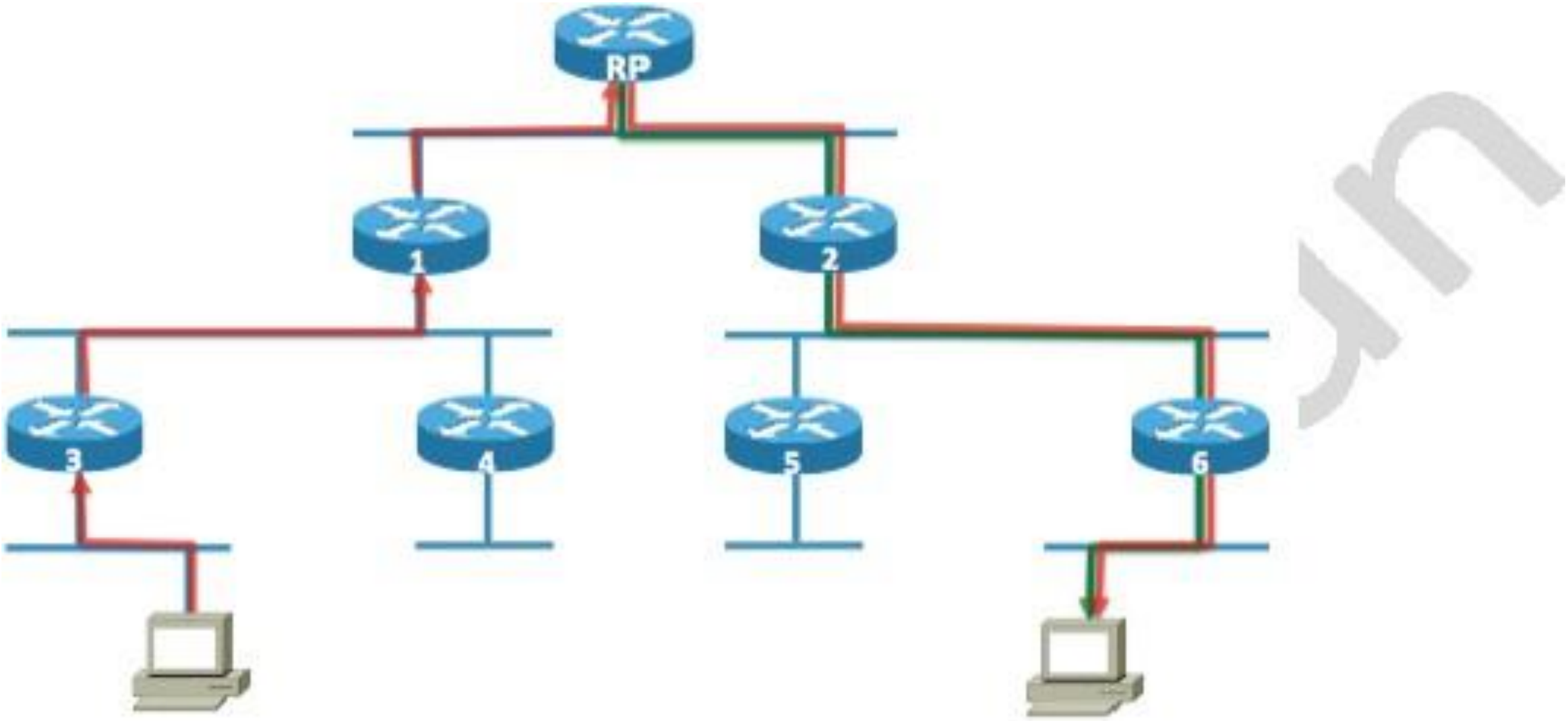
PIM Bidir – Bidirectional PIM

- PIM-Bidir is suitable to many to many multicast application such as trading floors application where all senders at the same time a receivers.

- Only uses shared tree Thus only (*,G) entries are seen in the multicast routing table.

- Traffic is brought up from the sources to the Rendezvous Point and then down to the receivers. Since there is only shared tree, PIM bidir uses less amount of state among the other PIM modes. Only (*,G) state is used.

- In PIM-bidir all trees rooted at the RP. In PIM ASM RFC check is used to prevent a routing loop, in PIM-Bidir in order to prevent a loop, Designated Forwarded is elected on every lin. Router with best path to the Rendezvous Point is selected as Designated Forwarder (DF).

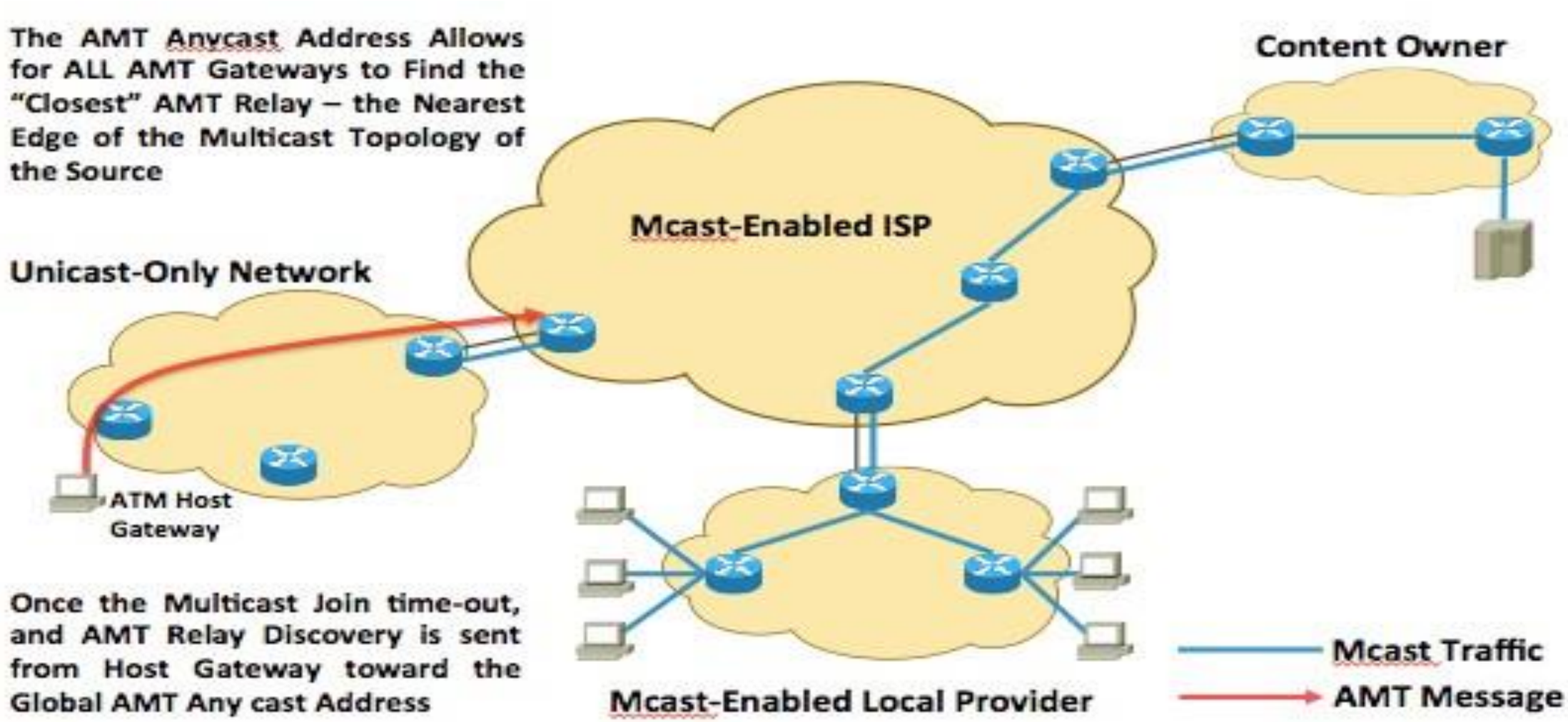


❖Arriving Causes Router and RP to Create (G) State

Multicast Case Study

- Terrano is a European based manufacturer company.
- Users of Terrano wants to watch the stream from the content provider which has a peering to Terrano’s Service Provider.
- But the problem is Terrano doesn’t have Multicast in the network.

What solution would Terrano use without enabling IP multicast in its network to improve user satisfaction by allowing Multicast stream from the Content provider?



- AMT discovery messages are sent to the Anycast address. Since in our case study, a service provider of Terrano supports multicast they can send the PIM

- (S,G) join to the Content Provider. Service Provider needs to have AMT Relay software on their router.

- AMT messages are unicast messages. Multicast traffic is encapsulated in Unicast packets.

- Terrano can receive multicast content at this point. Because end to end tree is built.

- AMT Host gateway feature is implemented on the receiver PC.

- Ideally AMT (Automatic Multicast Tunneling) Host Gateway feature should be provided by the first hop routers of Terrano

Multicast – Study Resources

❖**Books** :

❖http://www.amazon.com/Developing-IP-Multicast-Networks-I/dp/1578700779/ref=sr_1_1?ie=UTF8&qid=1436564509&sr=8-1&keywords=ip+multicast

❖**Videos** :

❖Ciscolive Session – BRKIPM – 1261 – Speaker Beau Williamson

❖**Podcast** :

❖http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/multicast-enterprise/whitepaper_c11- 474791.html

❖http://packetpushers.net/community-show-multicast-design-deployment-considerations-beau-williamson- orhan-ergun/

- **Articles** :

- https://tools.ietf.org/html/rfc7450

- http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ip-multicast/whitepaper_c11-508498.html

- https://www.juniper.net/techpubs/en_US/release-independent/nce/information-products/topic-collections/nce/bidirectional-pim/configuring-bidirectional-pim.pdf

- http://www.juniper.net/documentation/en_US/junos13.3/topics/concept/multicast-anycast-rp-mapping.html

- http://d2zmdbbm9feqrf.cloudfront.net/2015/usa/pdf/BRKIPM-1261.pdf

Quality of Service (QoS)

- Quality of service (QoS) is the overall performance of a telephony or computer network, particularly the performance seen by the users of the network.

- Two Quality Of Service approaches have been defined by the standard organizations. Namely Intserv (Integrated Services) and Diffserv (Differentiated Services).

- Intserv was demanding each and every flow to request a bandwidth from the network and network would reserve the required bandwidth for the user during a conversation.

- Think this is an on demand circuit switching, each flows of each user would be remembered by the network. This clearly would create a resource problem (CPU, Memory , Bandwidth) on the network thus never widely adopted.

- The second Quality of Service Approach is Diffserv (Differentiated Services) doesn’t require reservation but instead flows are aggregated and place into the classes.

- Then each and every node can be controlled by the network operator to treat differently for the aggregated flows.

- It is scalable approach compare to the Intserv Quality of Service model.

- Always classify and mark applications as close their source as possible

- Mark the packets with DSCP if it is possible. Because 802.1p bit get lost when the packet enter to the IP or MPLS domain, mapping is needed.

- Implement QoS always at the hardware if it is possible to avoid performance impact. . Switches support QoS in the the hardware, so for example in the Campus, classify and mark the traffic at the switches.

- Follows standard based DSCP markings to ensure interoperability

- Police unwanted traffic flows as close to their sources as possible.

- If traffic exceed the CIR or PIR and if it will be markdown, follows standard-based marking values. For example if the application class for the conforming traffic is AF31 , exceeding traffic should be markdown with AF32 and violating traffic should be markdown with AF33.

- Enable queuing policies at every node where the potential for congestion exist.

- QoS design should support minimum 3 classes; EF (Expedited Forwarding), DF (Default Forwarding/Best Effort) and AF(Assured Forwarding)

- If company policy allows YouTube, gaming and other non-business applications, Scavenger class is created and CS1 PHB is implemented. CS1 is defined as less than best effort service in the standard RFC.

- On AF queues, DSCP-based WRED should be enabled. Otherwise TCP synchronization occurs. WRED allows the packet to be dropped randomly and DSCP functionality provides packet to be dropped based on their priority.

RFC 4594

Application Class	Per-Hop Behavior	Admission Control	Queuing & Dropping	Application Examples
VoIP Telephony	EF	Required	Priority Queue (PQ)	Cisco IP Phones (G.711, G.729)
Broadcast Video	CS5	Required	(Optional) PQ	Cisco IP Video Surveillance / Cisco Enterprise T
Realtime Interactive	CS4	Required	(Optional) PQ	Cisco TelePresence
timedia Conferencing	AF4	Required	BW Queue + DSCP WRED	Cisco Jabber, Cisco WebEx
ultimedia Streaming	AF3	Recommended	BW Queue + DSCP WRED	Cisco Digital Media System (VoDs)
Network Control	CS6		BW Queue	EIGRP, OSPF, BGP, HSRP, IKE
Signaling	CS3		BW Queue	SCCP, SIP, H.323
/ Admin / Mgmt (OAM)	CS2		BW Queue	SNMP, SSH, Syslog
Transactional Data	AF2		BW Queue + DSCP WRED	ERP Apps, CRM Apps, Database Apps
Bulk Data	AF1		BW Queue + DSCP WRED	E-mail, FTP, Backup Apps, Content Distributio
Best Effort	DF		Default Queue + RED	Default Class
Scavenger	CS1		Min BW Queue (Deferential)	YouTube, iTunes, BitTorent, Xbox Live

Voice QoS Requirements

- Voice traffic should be marked to DSCP EF per the QoS Baseline and RFC 3246.
- Loss should be no more than 1 %.
- One-way Latency (mouth-to-ear) should be no more than 150 ms.
- Average one-way Jitter should be targeted under 30 ms.
- 21–320 kbps of guaranteed priority bandwidth is required per call (depending on the sampling rate, VoIP codec and Layer 2 media overhead).

Voice quality is directly affected by all three QoS quality factors: loss, latency and jitter.

Video QoS Requirements

In general we are interested in two type of video traffic. Interactive Video and Streaming Video. Interactive Video :

When provisioning for Interactive Video (IP Videoconferencing) traffic, the following guidelines are recommended:

- Interactive Video traffic should be marked to DSCP AF41; excess Interactive- Video traffic can be marked down by a policer to AF42 or AF43.
- Loss should be no more than 1 %.
- One-way Latency should be no more than 150 ms.
- Jitter should be no more than 30 ms.
- Overprovision Interactive Video queues by 20% to accommodate bursts

Video Best Practices:

- Streaming Video (whether unicast or multicast) should be marked to DSCP CS4 as designated by the QoS Baseline.

- Loss should be no more than 5 %.

- Latency should be no more than 4–5 seconds (depending on video application buffering capabilities).

- There are no significant jitter requirements.

- Guaranteed bandwidth (CBWFQ) requirements depend on the encoding format and rate of the video stream.

- Streaming video is typically unidirectional and, therefore, Branch routers may

- not require provisioning for Streaming Video traffic on their WAN/VPN edges (in the direction of Branch-to-Campus).

➢Data Applications QoS Requirements

- ❖Best Effort Data
- ❖Bulk Data
- ❖Transactional/Interactive Data

Best Effort Data

- The Best Effort class is the default class for all data traffic. An application will be removed from the default class only if it has been selected for preferential or deferential treatment.

- Best Effort traffic should be marked to DSCP 0. Adequate bandwidth should be assigned to the Best Effort class as a whole, because the majority of applications will default to this class; reserve at least 25 percent for Best Effort traffic.

Bulk Data

- The Bulk Data class is intended for applications that are relatively non- interactive and drop-insensitive and that typically span their operations over a long period of time as background occurrences. Such applications include the following:
- FTP
- E-mail
- Backup operations
- Database synchronizing or replicating operations
- Content distribution

- Any other type of background operation

- Bulk Data traffic should be marked to DSCP AF11; excess Bulk Data traffic can be marked down by a policer to AF12; violating bulk data traffic may be marked down further to AF13 (or dropped).

- Bulk Data traffic should have a moderate bandwidth guarantee, but should be constrained from dominating a link.

Transactional/Interactive Data

- The Transactional/Interactive Data class, also referred to simply as Transactional Data, is a combination to two similar types of applications: Transactional Data client-server applications and Interactive Messaging applications.

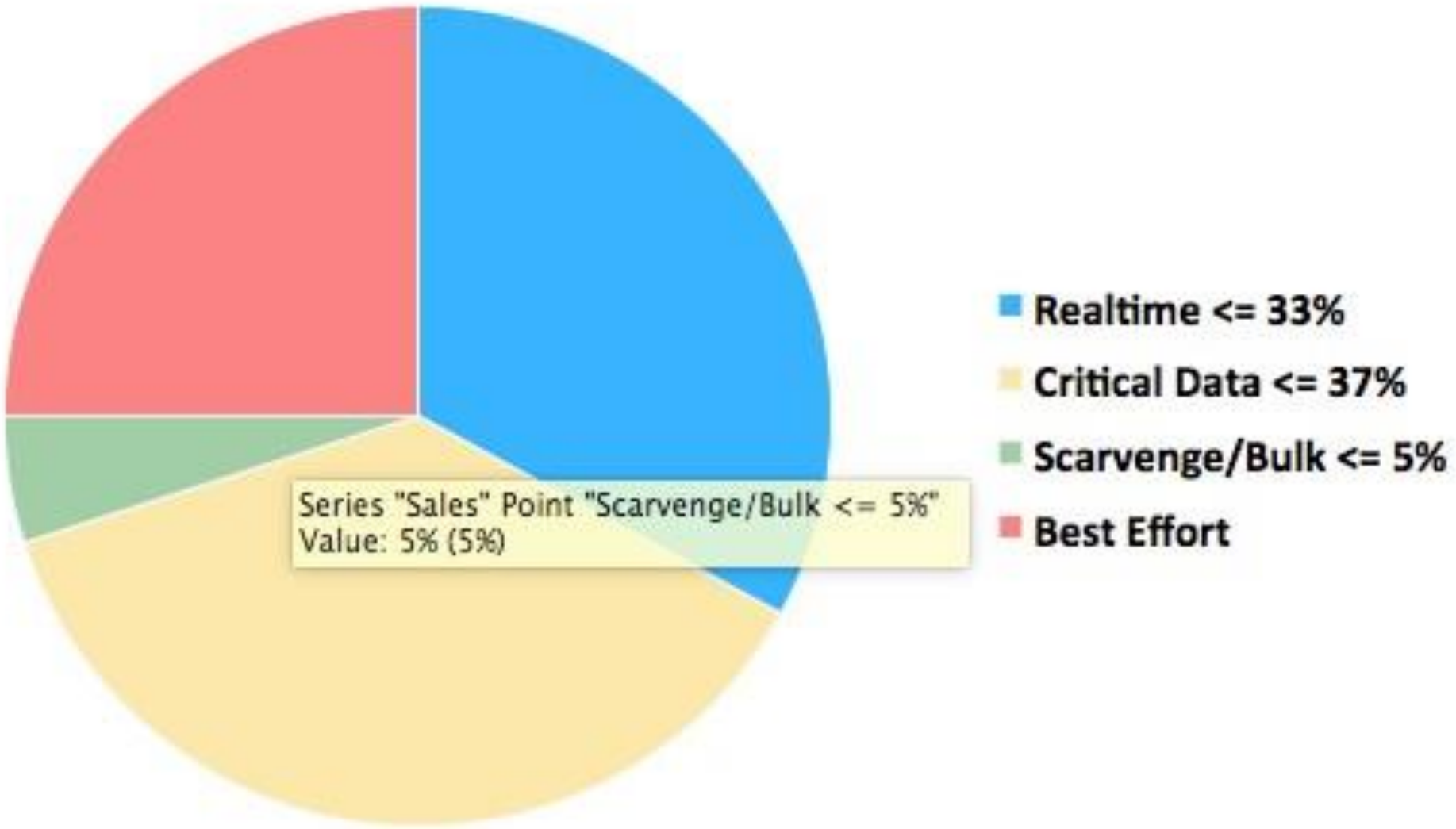
- The response time requirement separates Transactional Data client-server applications from generic client-server applications.

- For example, with Transactional Data client-server applications such as SAP, PeopleSoft.

- Transaction is a foreground operation; the user waits for the operation to complete before proceeding.

- E-mail is not considered a Transactional Data client-server application, as most e-mail operations occur in the background and users do not usually notice even several hundred millisecond delays in mail spool operations.

- Transactional Data traffic should be marked to DSCP AF21; excess Transactional Data traffic can be marked down by a policer to AF22; violating Transactional Data traffic can be marked down further to AF23 (or dropped).

- 

QOS Study Resources

- **Books :**

- http://www.amazon.com/End---End-QoS-Network- Design/dp/1587143690/ref=sr_1_1?ie=UTF8&qid=1436564258&sr=8-1&keywords=end+to+end+qos+network+design

- **Videos :**

- **Ciscolive Session – BRKCRS -2501**

- **https://www.youtube.com/watch?v=6UJZBeK_JCs**

❖ Articles :

- ❖ http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND/QoS-SRND- Book/QoSIntro.html
- ❖ http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Video/qosmrn.pdf
- ❖ http://orhanergun.net/2015/06/do-you-really-need-quality-of-service/
- ❖
- ❖ http://d2zmdbbm9feqrf.cloudfront.net/2013/usa/pdf/BRKCRS-2501.pdf
- ❖
- ❖ https://ripe65.ripe.net/presentations/67-2012-09-25-qos.pdf

VPN Design

In VPN Design Course below technologies will be covered

- WAN VPN Technologies
- Datacenter VPN Technologies

WAN VPN Technologies

1. GRE
2. mGRE
3. IPSEC
4. DMVPN
5. GETVPN
6. LISP

Datacenter VPN Technologies

1. EoMPLS
2. VPLS
3. OTV
4. E-VPN
5. PBB-EVPN
6. VxLAN

VPN Theory

- Virtual Private Networks extend the private network over another company’s network. It can be setup over another private network such as MPLS or public network such as Internet

- All VPN technologies add extra byte to the packet or frame which increases the overall MTU so the network links should be accommodated to handle bigger MTU

- VPN works based on encapsulation and decapsulation. For example in GRE, mGRE or DMVPN encapsulate IP packets into another IP packet and VPLS or EVPN encapsulates Layer 2 frame into an MPLS packets

- Some VPN technologies support to run routing protocol on top, some doesn’t support. In order to support routing over tunnel, tunnel endpoints should be aware from each other.

For example MPLS Traffic Engineer tunnels don’t support routing protocols to run over since the LSPs are unidirectional which mean Head-end and Tail-end routers are not associated

All WAN technologies except IPSEC and LISP in our list supports routing protocols to run over.

GRE

- Manual point to point tunnels
- Supports routing protocols to run over
- IPv4 and IPv6 can be transported over GRE
- Non-IP protocols such as IPX,SNA etc. can be carried over GRE tunnel as well
- If there are too many sites which need to communicate with each other, GRE is not scalable. But in Hub and Spoke topologies it can be used since whenever new site is added, only new site and hub should be revisited.
- Even though in Hub and Spoke topologies, the configuration can be too long on the Hub site

- GRE add 24 bytes to the IP Packet. 4 byte GRE header and 20 bytes new IP header, this increases MTU size of the IP packet

- GRE doesn’t come by default with encryption so in order to encrypt the packet, IPSEC should be enabled over GRE tunnel

- Classical use cases of GRE tunnel is over Internet with IPSEC, VRF-lite to carry different VPN information separately, IPv6 tunneling over IPv4 transport

- GRE is used mostly together with IPSEC to support the traffic which is not supported by IPSEC by default. For example IPSEC tunnels don’t support Multicast by default but together with GRE, GRE over IPSEC supports multicast as well.

mGRE – Multipoint GRE

- Allows multiple destinations such as multiple spoke sites to be grouped into a single multipoint interface

- mGRE is a multipoint bidirectional GRE tunnel

- Uses only 1 IP subnet so routing table of the routers are reduced which is good for the convergence, troubleshooting and device performance

- Remote endpoint address is not configured that’s why it requires additional mechanisms for tunnel end point discovery. These additional mechanisms can be BGP or NHRP. When it is used with NHRP, solution is called DMVPN which we will see later in detail

- Supports IPSEC and routing protocols to run on top
- Supports IPv6 and non-IP protocols
- Don’t require manually configured GRE tunnels that’s why configuration complexity is reduced greatly
- Suitable for full mesh topology such as spoke to spoke tunnels
- Use cases are:
- MPLS VPN over mGRE: in this case single mGRE interface is enough to carry multiple VPN information.
- Another use case is VRF Lite over mGRE which requires 1:1 VRF to mGRE interface. So each VRF requires separate mGRE interface. DMVPN uses this concept.
- mGRE adds 28 bytes of overhead because of the additional 4-byte Key field which is not typically included in the GRE header when using a point to point GRE tunnel

IPSEC

- Provides secure transmission of packets at the IP layer (Not Layer2)

- Packets are authenticated and encrypted

- Provides site to site and remote access VPN topologies

- IPSEC has two modes. Transport mode and tunnel mode.

- The difference between the two is that in transport mode; only the IP payload is encrypted, in tunnel mode ; the entire IP datagram(including header) is encrypted.

- IPSEC is point to point tunneling mechanism, thus in large scale implementation it can not scale.

- It is mainly used to provide encryption for the data over the Public Internet. This data shouldn’t be real-time, packet loss, jitter and delay sensitive since there is no QoS SLA over the Public Internet, it is best effort.

- IP Multicast is not supported over IPSEC VPNs.

- QoS is supported through TOS byte preservation

- Routing protocols don’t run over IPSEC VPNs, it requires additional GRE encapsulation (GRE over IPSEC)

- IPSEC VPN is a standard protocol which supported by every vendor

DMVPN

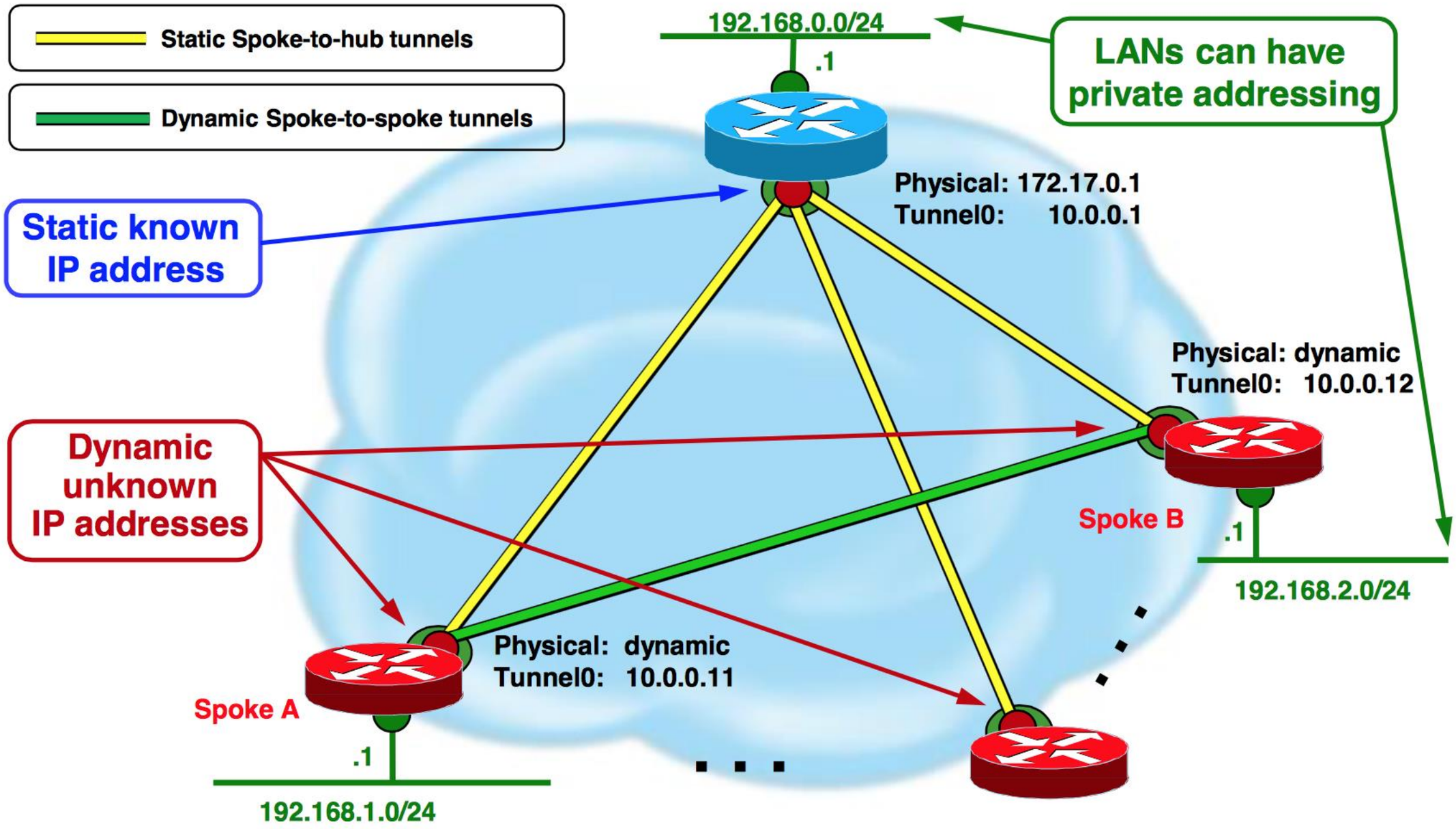
- Point to Multipoint and Multipoint to Multipoint Tunneling technology

- Works based on two standard technologies; NHRP and mGRE (Multipoint GRE)

- NHRP is used to map NBMA/Underlay address to the Tunnel/Overlay address.

- mGRE tunnel interface doesn’t require manual tunnel destination configuration.

- Local address/LAN subnets are advertised over the overlay tunnels through the routing protocols



- Multicast is not natively supported, but ingress replication is done at the HUB, thus Multicast support of DMVPN can create scalability problem in Large scale design

- Per tunnel QoS is supported, this prevent hub to overutilized the spoke bandwidth

- All routing protocols except IS-IS can run over DMVPN. IS-IS can not run since DMVPN only support IP protocols and IS-IS works based on Layer 2.

- Invented by Cisco but other vendors provide the capability with different name, Cisco’s implementation is not compatible with the other vendors.

- IPSEC is optional and can be implemented as native point to point IPSEC or GETVPN (Group key)

- DMVPN can carry IPv4 and IPv6 unicast and multicast packets over the overlay tunnels.

- DMVPN also can work over IPv4 and IPv6 private and public infrastructure transport such as Internet or MPLS VPN.

- There are three phases of DMVPN, Phase 1, Phase 2 and Phase 3

- In all the DMVPN Phases, overlay routing protocol control packets pass through the HUB. There is no spoke to spoke control plane traffic.

DMVPN Phase -1

- Spokes use Point to Point GRE but Hub uses a multipoint GRE tunnel.
- mGRE tunnel interface simplifies configuration greatly on the Hub.
- Spokes have to specify the tunnel destination as Hub since they run P2P GRE tunnel not mGRE in DMVPN Phase 1
- Summarization is allowed from the Hub down to spokes. The Hub can only send default route as well ,since the remote spokes next hop is not preserved by the Hub.
- Hub changes the IGP next hop to itself hence spoke to spoke traffic passes through the Hub.
- Spoke to spoke tunnel cannot be created, that’s why DMVPN phase 1 doesn’t provide full mesh connectivity

• DMVPN Phase 2

- Spoke to spoke dynamic on demand tunnels are first introduced in Phase 2.

- In contrast to Phase 1, mGRE (Multipoint GRE, not Multicast) interface is used in Phase 2 on the Spokes

- Thus, spokes don’t require tunnel destination configuration under the tunnel interface and tunnel mode is configured as ” Multipoint Gre”.

- Spoke to spoke traffic doesn’t have to go through the HUB. Spokes can trigger on demand tunnel between them.

- The biggest disadvantage of Phase 2 is, each spoke has to have all the Remote – LAN subnets of each other since the Hub, preserves the next hop addresses of the spokes.

- Thus, spokes have to have a reachability to the tunnel addresses of each other.

- This disallows the summarization or default routing from Hub down to the spokes.

- This is a serious design limitation for the large scale DMVPN networks.

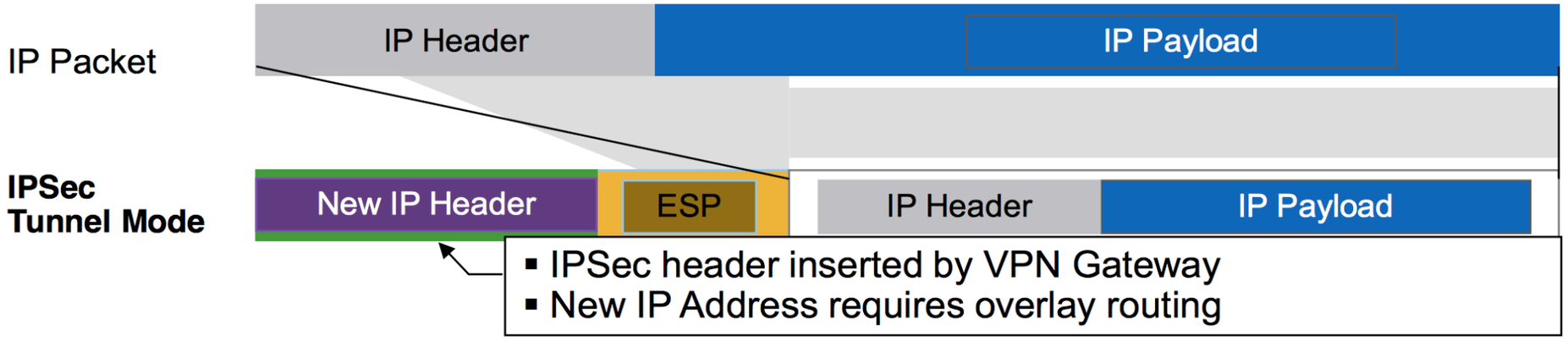
- For the distance vector protocols, Split horizon needs to be disabled and “no next-hop self” should be enabled on the HUB.

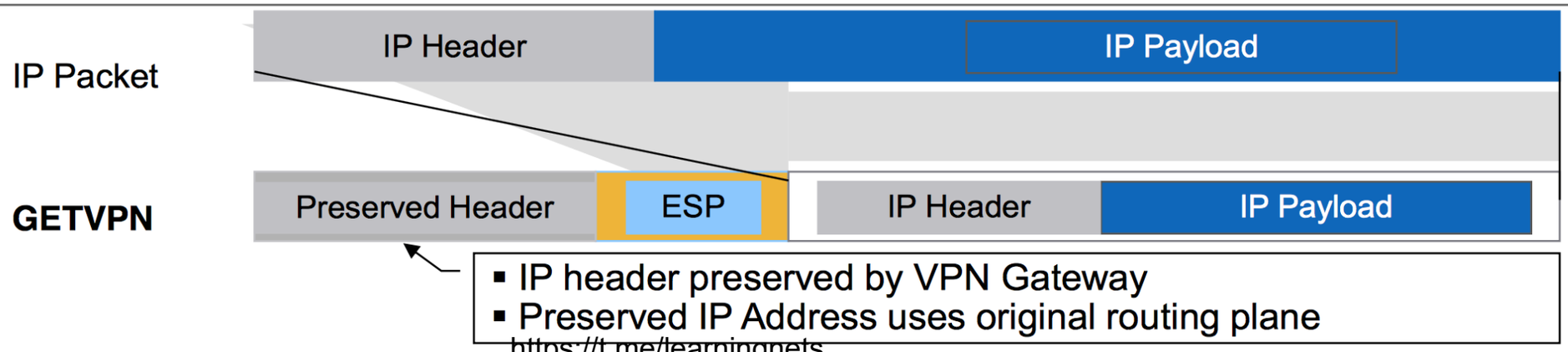
DMVPN Phase 3

- Spoke to spoke dynamic tunnels are allowed.
- Spokes don’t have to have the next hop of each other’s private address for the local subnets.
- An NHRP redirect message is sent to the spokes to trigger spoke to spoke tunnels. Hub provides the Public/NBMA addresses of the spokes to each other.
- Since the next hop in the routing table of the spokes is HUB’s tunnel address, spokes don’t have to have the specific next hop information of each other.
- This allows summarization and default routing in Phase 3.
- Hub can send a summary or just a default route down to the spokes.
- Hence, Phase 3 is extremely scalable

GETVPN

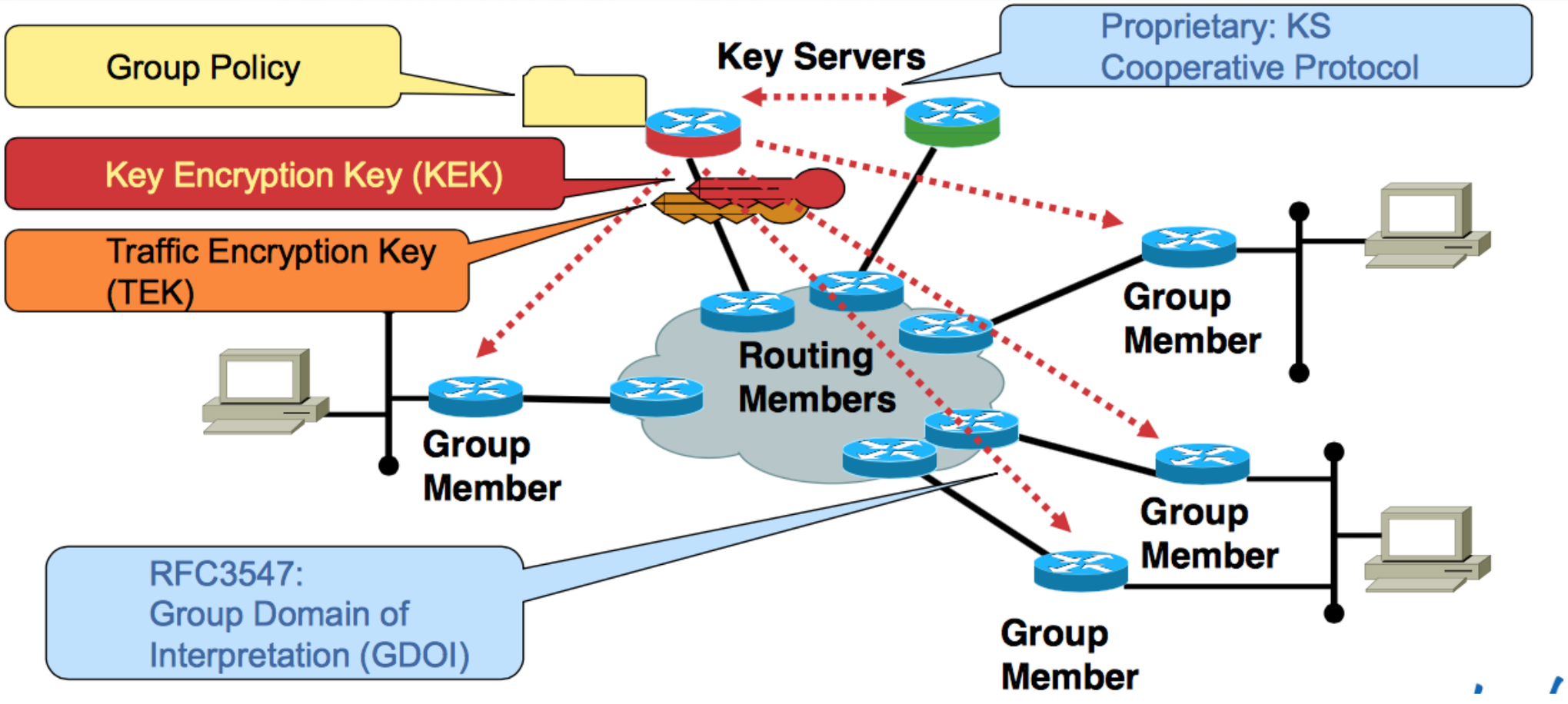
IPSec Tunnel Mode vs. GETVPN





- It can run on top of any routing protocol
- It doesn’t support overlay routing protocol since there is no tunnel to run routing protocol over
- Excellent Multicast support compare to DMVPN since there is no replication but it uses native multicast
- Doesn’t support Non-IP Protocols
- Uses Group Key mechanism instead of Point to Point Keys which are used in DMVPN by default, thus from the security scalability point of view, GETVPN is much scalable technology compare to the DMVPN

- **Group Members (GMs) "register" via GDOI with the Key Server (KS)**
  - **KS authenticates & authorizes the GMs**
  - **KS returns a set of IPSEC SAs for the GMs to use**
  - **GMs exchange encrypted traffic using the group keys**
- **Traffic uses IPSec Tunnel Mode with IP address preservation**



orhanergun.net	DMVPN	GETVPN
Scalability	Scalable	Much more scalable than DMVPN
Working on Full Mesh topology	Permanent hub and spoke tunnels and on demand spoke to spoke tunnels, it works but limited scalability	It works perfectly if the underlying routing architecture is full mesh topology, GET VPN needs underlay routing
Working on Hub and Spoke	Works very well	Works very well
Suitable on private WAN	Yes	Yes
Suitable over Public Internet	Yes	No. GETVPN cannot run over Public Internet because of IP header preservation
End point discovery	To setup the Mgre tunnels uses underlay routing, for the private address discovery uses NHRP (Next hop Resolution Protocol)	It uses underlay routing to create VPN, there is no overlay tunnels
Tunnel Requirement	Yes, it uses Mgre(Multi Point GRE) tunnels to create overlays	It is tunnelles VPN, uses underlaying routing to encrypt the data between endpoints
Standard Protocol	No,Cisco proprietary	No,Cisco proprietary but Juniper also supports the same idea with Group VPN feature
Stuff Experince	Not well known	Not well known
Overlay Routing Protocol Support	Except IS-IS other routing protocols are supported, IS-IS runs on top of Layer 2 but only IP protocols can run over DMVPN	It is tunnelles VPN so routing protocols cannot run on top of GETVPN but it requires underlaying routing protocols to setup the communication
Required Protocols	NHRP and Mgre	GDOI and ESP
QoS Support	Good, can support per tunnel QoS which uses shaping on the DMVPN Hub to protect capacity and SLA	Good, it uses underlaying network's QoS architecture,in addtion to queueing,shaping at the GET VPN Group Members to protect SLA is enabled
Multicast Support	Multicast over the tunnel is handled at the DMVPN Hub. Hub replicates multicast traffic which is not efficiend	Native multicast support.Multicast replication is done in the network, doesn't need Hub device to replicate. Multicast MDTs (Source , Shared) are used in the traditional way, so multicast handling of GETVPN is much better than DMVPN
Security	Point to Point IPSEC SA	Multipoint to Multipoint IPSEC SA
Resource Requirement	More	Less
IPv6 Support	Yes,it can be setup over IPv6 transport or it can carry IPv6 payload. So IPv6 over DMVPN and DMVPN over IPv6 both are possible	Yes
Default Convergence	Slow	Fast
Can run over other VPN ?	DMVPN is already tunneled VPN technology so only routing is enough, it doesn't make sense to run tunnel over tunnel	GETVPN can run over DMVPN since GETVPN is tunnelless technology, use case of GETVPN over DMVPN is to carry private addressing over Internet. Most common use case of GETVPN is over MPLS VPN or VPLS since both VPN technologies are full mesh by default and GET VPN provides very good scalability for encryption

LISP-Locator Identity Separation Protocol

- Provides IP mobility with the pull based approach. Tunnel end points are known by the centralized database and the routers pull the information from the database when it is needed.

- IP in IP encapsulation mechanism which uses underlaying routing infrastructure, thus, can provide any to any connectivity

- There is also attempt to provide MAC in IP encapsulation

- It is an Experimental RFC (RFC 6830) but invented by Cisco.

- Similar to DNS infrastructure. Mapping Database keeps the host to gateway mapping information.
- Host space is called EID (Endpoint Identifier) and its gateway is called RLOC (Routing Locator)
- Most interesting LISP use cases are VM Mobility in the Datacenter, BGP Traffic Engineering at the Internet Edge and IPv6 transition (IPv6 over IPv4 or IPv4 over IPv6 infrastructure)
- From the security point of view MR/MS is a target point

LISP Device Roles

- ITR – Ingress Tunnel Router

– Receives packets from site-facing interfaces

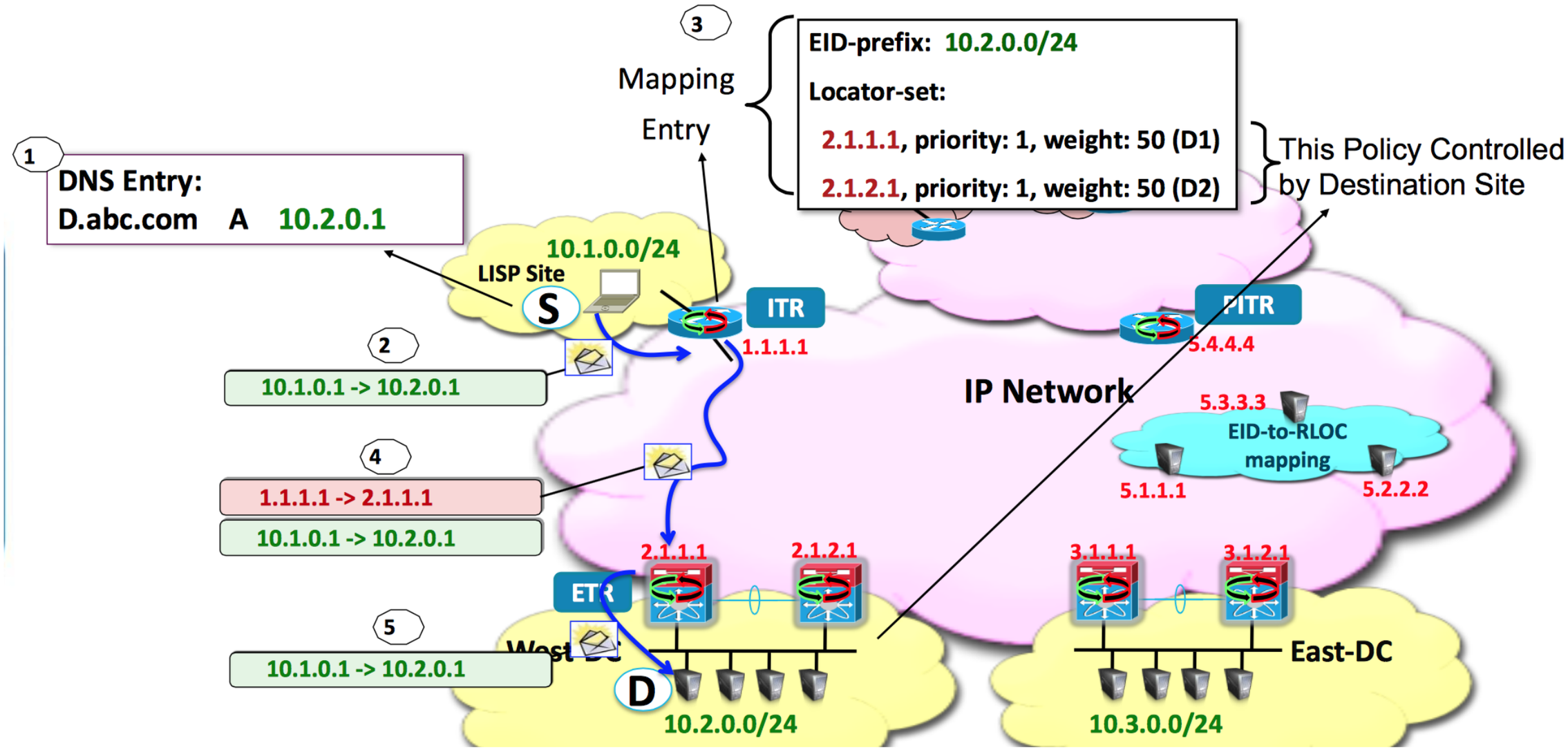
– Encapsulates to remote LISP sites, or natively forward to non-LISP sites

- ETR – Egress Tunnel Router

– Receives packets from core-facing interfaces
– De-cap and deliver packets to local EIDs at the site

- EID to RLOC Mapping Database

- Contains RLOC to EID mappings



MPLS

Theory

- If the requirement is to have scalable VPN solution which can provide fast reroute traffic protection, then only choice is MPLS

- MPLS is a protocol independent transport mechanism. It can carry layer 2 and layer 3 payloads. Packet forwarding decision is made solely on the label without the need to examine the packet itself.

- MPLS interacts as an overlay with IGP and BGP in many ways. For example in Multi-level IS-IS design, level 1 domain breaks end to end LSP.

- In OSPF and EIGRP summarization creates the same problem and we mentioned about this problem in respective chapters.

- MPLS is a transport mechanism.
- It supports label stacking which allows many applications to run over MPLS. MPLS application is the topic of this session
- All application needs additional label
- Most of the MPLS operation requires minimum 2 labels
- In MPLS layer 3 VPN, Transport and BGP label
- In Layer 2 VPN, Transport and VC Label
- Label can be assigned by 4 protocols currently. LDP, RSVP, BGP and Segment Routing

MPLS Applications

Important MPLS Applications/Services are below. In this chapter all of them will be explained.

- Layer 2 MPLS VP
- Layer 3 MPLS VPN
- Inter AS MPLS VPN
- Carrier Supporting Carrier
- MPLS Traffic Engineering
- Seamless MPLS

- Layer 2 frame is carried over the MPLS transport. Same MPLS infrastructure can have all above MPLS application/services at the same time. You can serve to Layer2 VPN customers, Layer3 VPN customers by having MPLS Traffic Engineering LSPs for SLA or FRR purpose.

- If you are extending MPLS towards the access domain of the backbone then you can have end to end MPLS backbone without the need the protocol translation.

- 2 different layer 2 VPN architectures provide similar services defined in MEF (Metro Ethernet Forum)

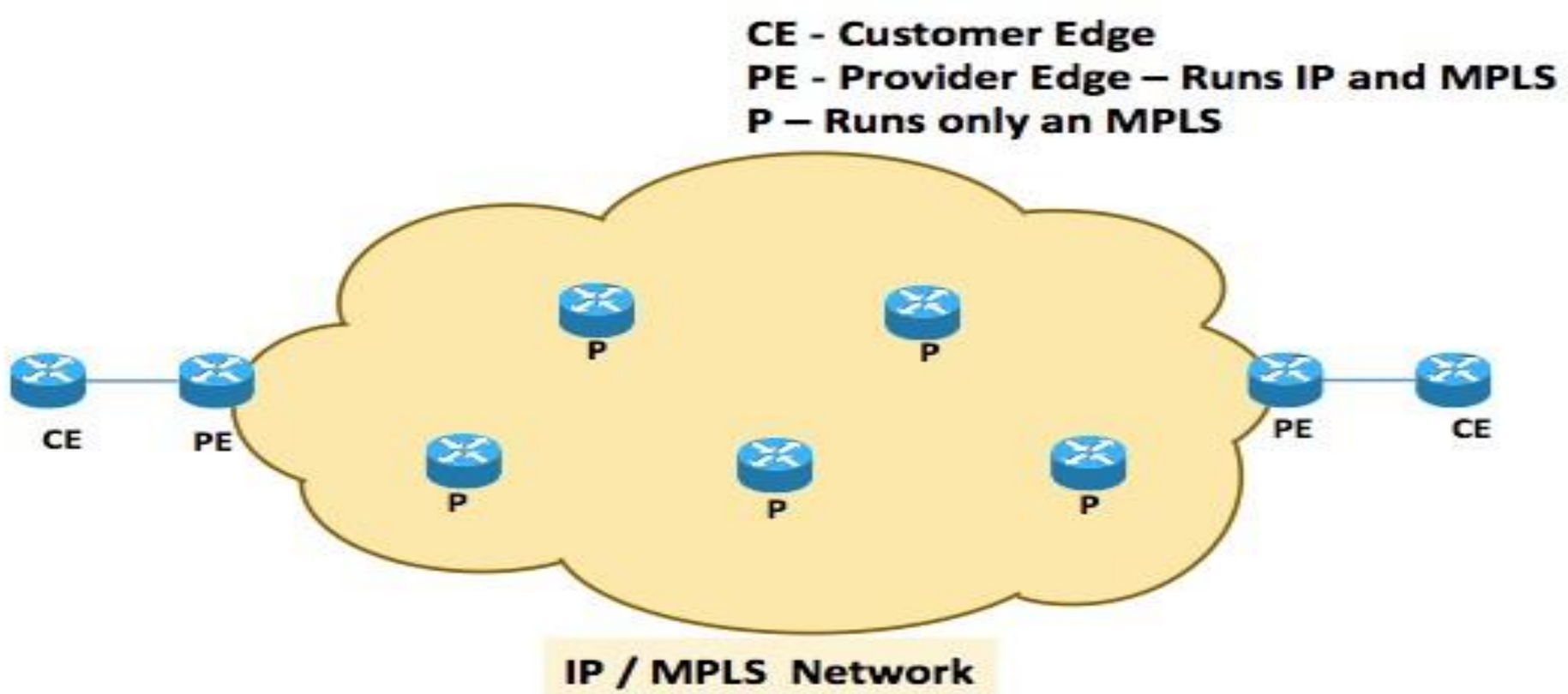
- MPLS layer 2 VPN can be point to point which is called VPWS (Virtual Private Wire Service), multi point to multi point which is called VPLS (Virtual Private Lan service).

- Both VPWS and VPLS can be accomplished in two ways.

- In both methods transport tunnel is created between the PE devices via LDP protocol.

- In Kompella method, pseudo wire is signaled via BGP which is one of the methods.

- In Martini method, pseudo wire is signaled via LDP (Label Distribution Protocol) which is the second method.



- CE is customer equipment which can be managed by Service Provider or Customer depending on SLA.

- PE is Provider Edge device. In MPLS networks, all the intelligence is at the edge. Core is kept as simple as possible. KISS principle in network design comes from the ‘ Intelligent Edge, Dummy Core ‘ idea.

- P is the Provider device and only have a connection to the P devices. P device doesn’t have a connection to the customer network.

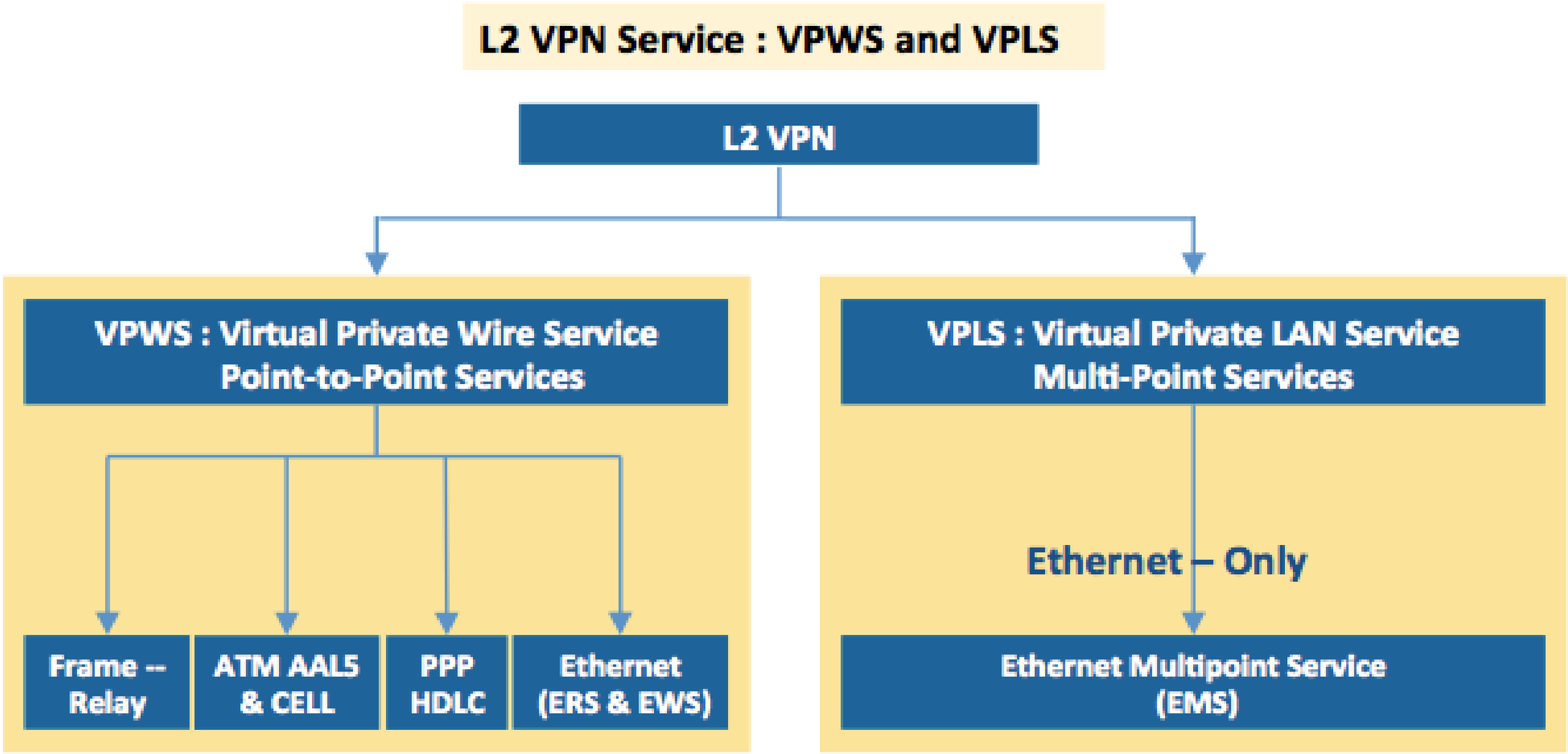
- PE device looks at the incoming frame or packet and identify which egress PE device is used for transport. Second lookup is made to determine the egress interface on the egress device.

- Packet gets two labels in both MPLS layer 2 and MPLS layer 3 VPN. Outer label which is also called topmost or transport label is used to reach to the egress device.

- Inner label is called pseudo wire or VC label in MPLS layer 2 VPN and used to identify the individual pseudo wire on the egress PE.

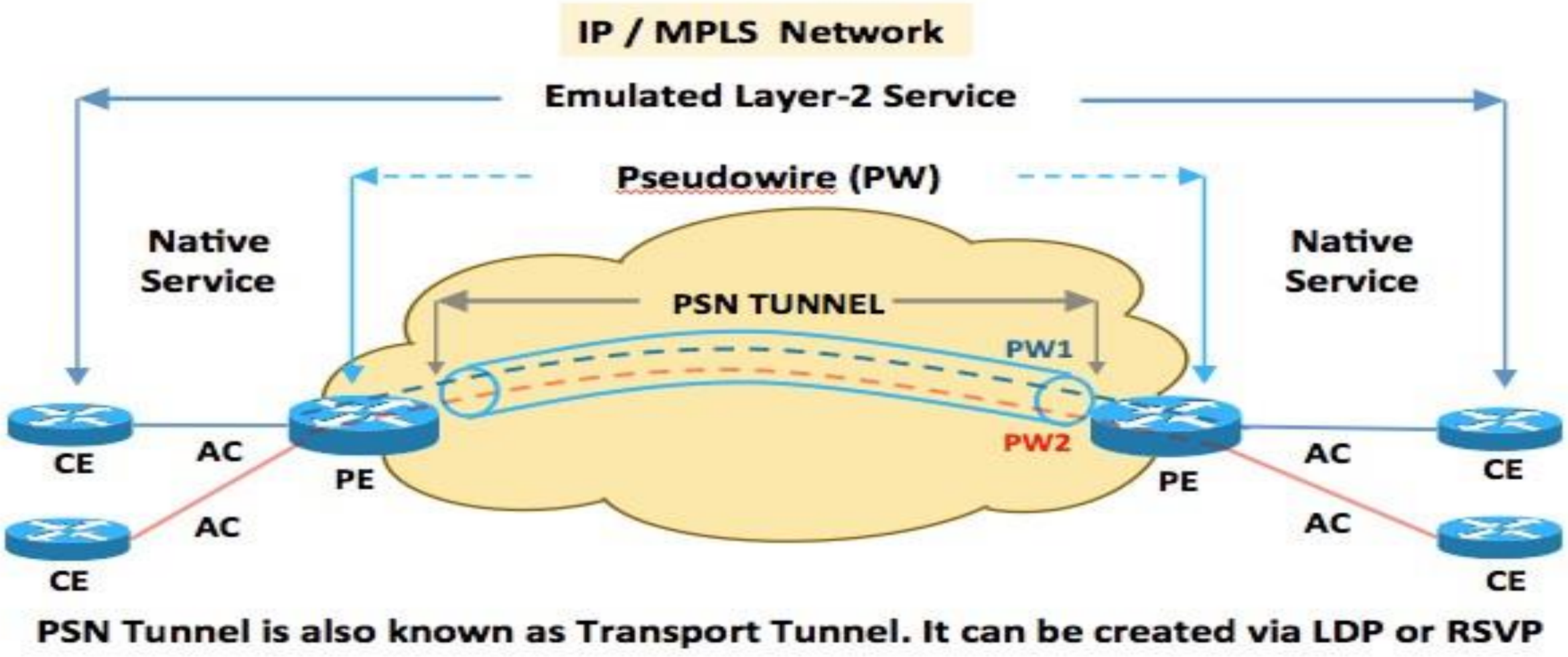
- In Martini method; both transport and VC (Virtual Circuit) label is sent (signaled) via LDP (Label Distribution Protocol).Targeted LDP session is created between PEs.

- In Kompella method; transport label is signaled via LDP, VC label is signaled via MP-BGP (Multiprotocol BGP).New address family is enabled if there is already BGP for other services.



- VPWS can carry almost all layer 2 payloads as can be seen from the above diagram. VPLS can carry Ethernet only. Although there is an attempt in the IETF for the other layer2 payloads over VPLS as well.

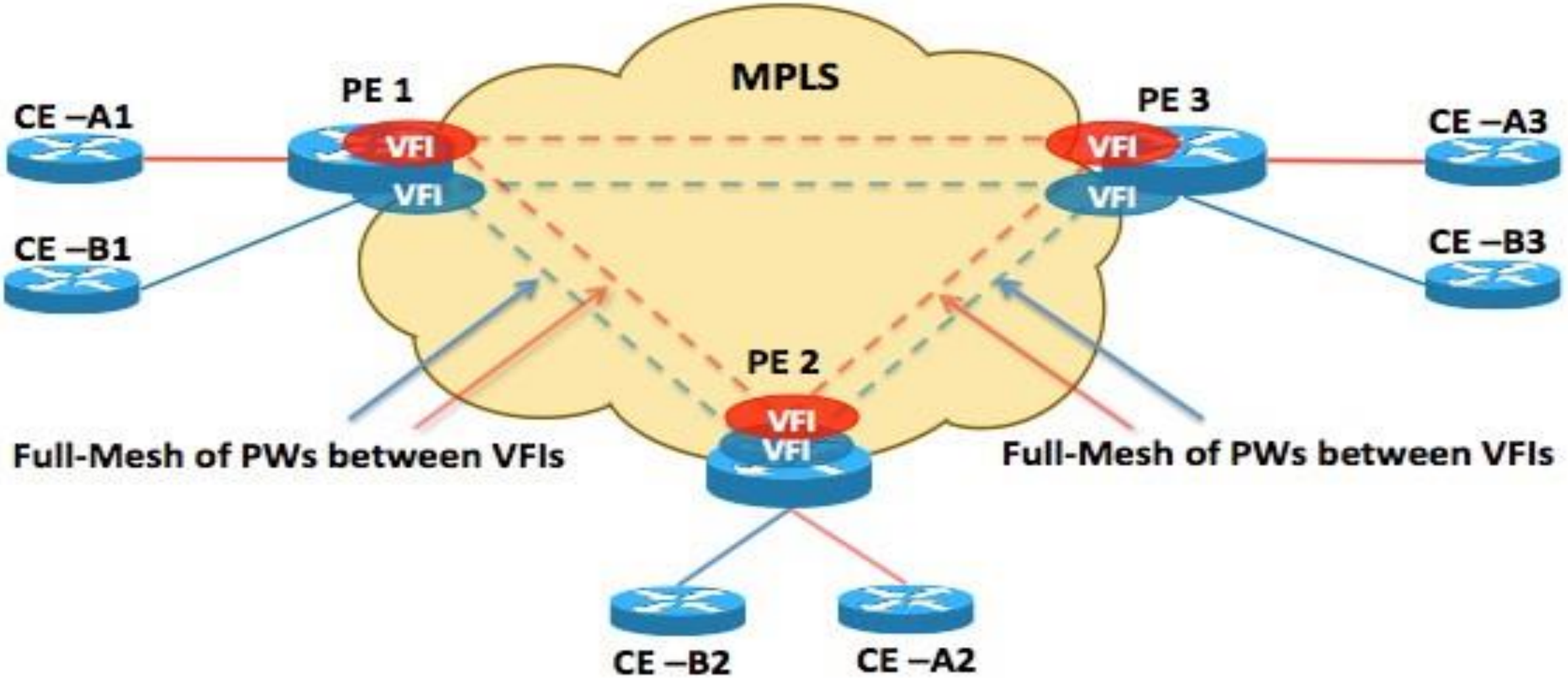
- In VPWS; PE devices learn only Vlan information if the VC type is Vlan, if the VC type is Ethernet then PE device doesn’t keep any state.



- There is only one egress point for the VPWS service which is the other end of the pseudo wire, thus PE device doesn't have to keep Mac Address to PW binding. PE device doesn't have to learn MAC address of the customer. This provides scalability.

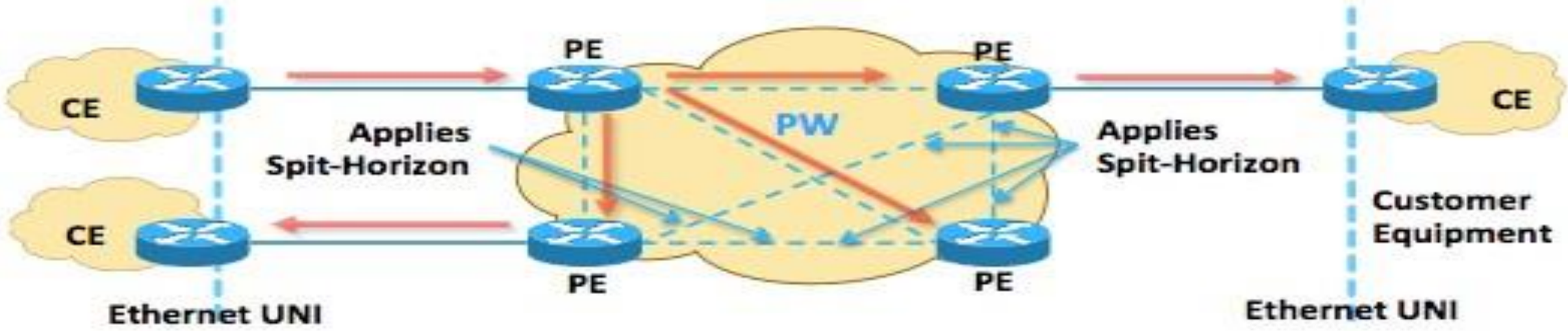
- But in VPLS, since Service Provider provides an emulated bridge service to the customer, MAC to PW binding is necessary. Destination might be any PE since the service is multipoint. PE devices keep MAC addresses of the Customer. Although there is PBB-VPLS which provides more scalability by eliminating learning of customer MAC addresses, it is not widely implemented

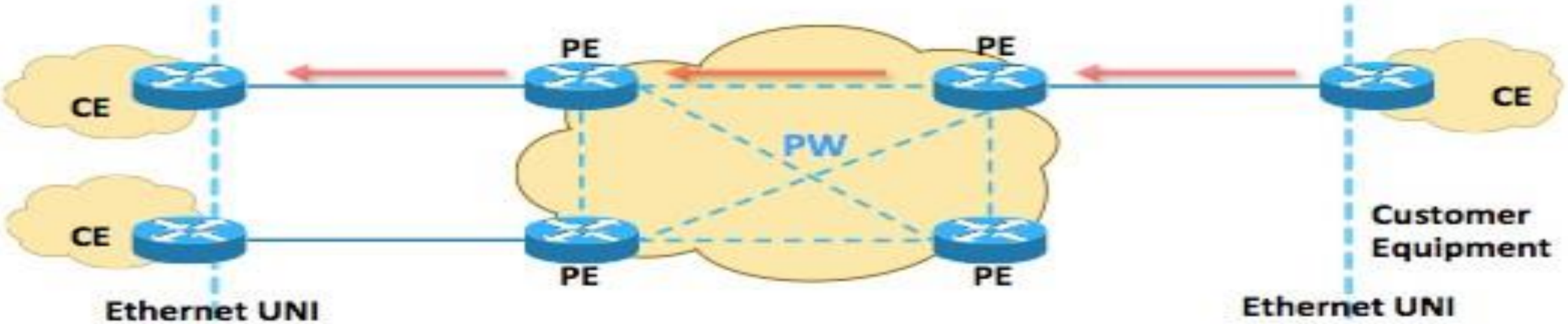
In order to create a VPLS , full mesh of Pseudowires is necessary



- VFI, also known as VSI is a virtual forwarding/switching instance is an equivalent of VRF in Layer3 VPN. It is used to identify the VPLS domain per customer.

- As it is depicted in the above topology, Point to Point pseudo wire is created between the PE devices for the VPWS (EoMPLS, point to point service). In order to have VPLS service full mesh of point to point pseudo wire is created between all the PEs which has a membership in the same VPN





- There is no Spanning tree in the Service Provider core network for loop avoidance in VPLS. Instead there is a split horizon rule in the core by default enabled.

- According the VPLS split horizon, if the customer frame is received from pseudo wire, it is not send back to another pseudo wire. Since there is full mesh pseudo wire connectivity among the all VPLS PE in a given customer VPN, full reachability between the sites is achieved.

- Number of PE which needs to join the VPLS instance which is also known as VSI (Virtual Switch Instance) might be too high.

- In this case auto discovery to learn the other PE in the same VPN is highly important from the scalability point of view. Auto discovery can be achieved in multiple ways. Radius server is one way but more common is BGP. Multi protocol BGP can carry VPLS membership information as well.

EVPN

- EVPN is a next generation VPLS. In VPLS customer mac addresses are learned through data plane. Source mac addresses are recorded based on source address from both AC (Attachment Circuit) and Pseudo wire.

- In VPLS, active active flow based load balancing is not possible.

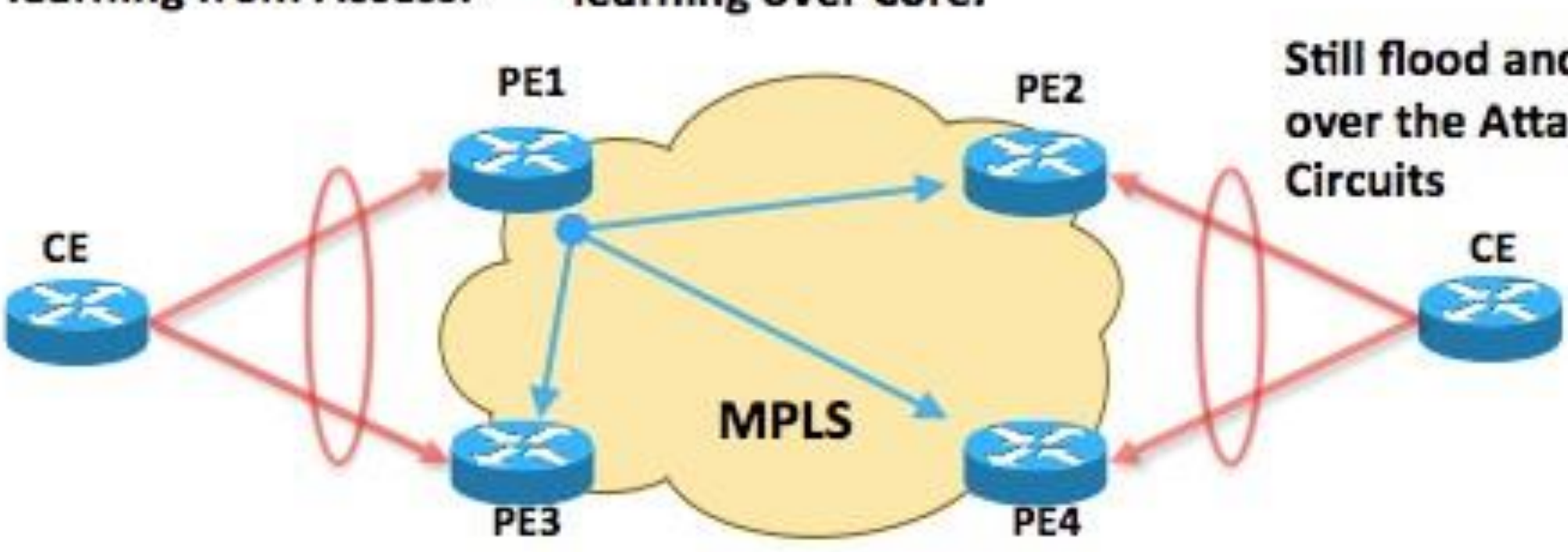
- Customer can be dual homed to the same or different PEs of Service Provider but either those links can be used as Active/Standby for all Vlans or Vlan based load balancing can be achieved.

- EVPN can support active active flow based load balancing so same Vlan can be used on both PE device actively. This provides faster convergence in customer link, PE link or node failure scenarios.

Data Plane address learning from Access.

Control Plane address advertisement and learning over Core.

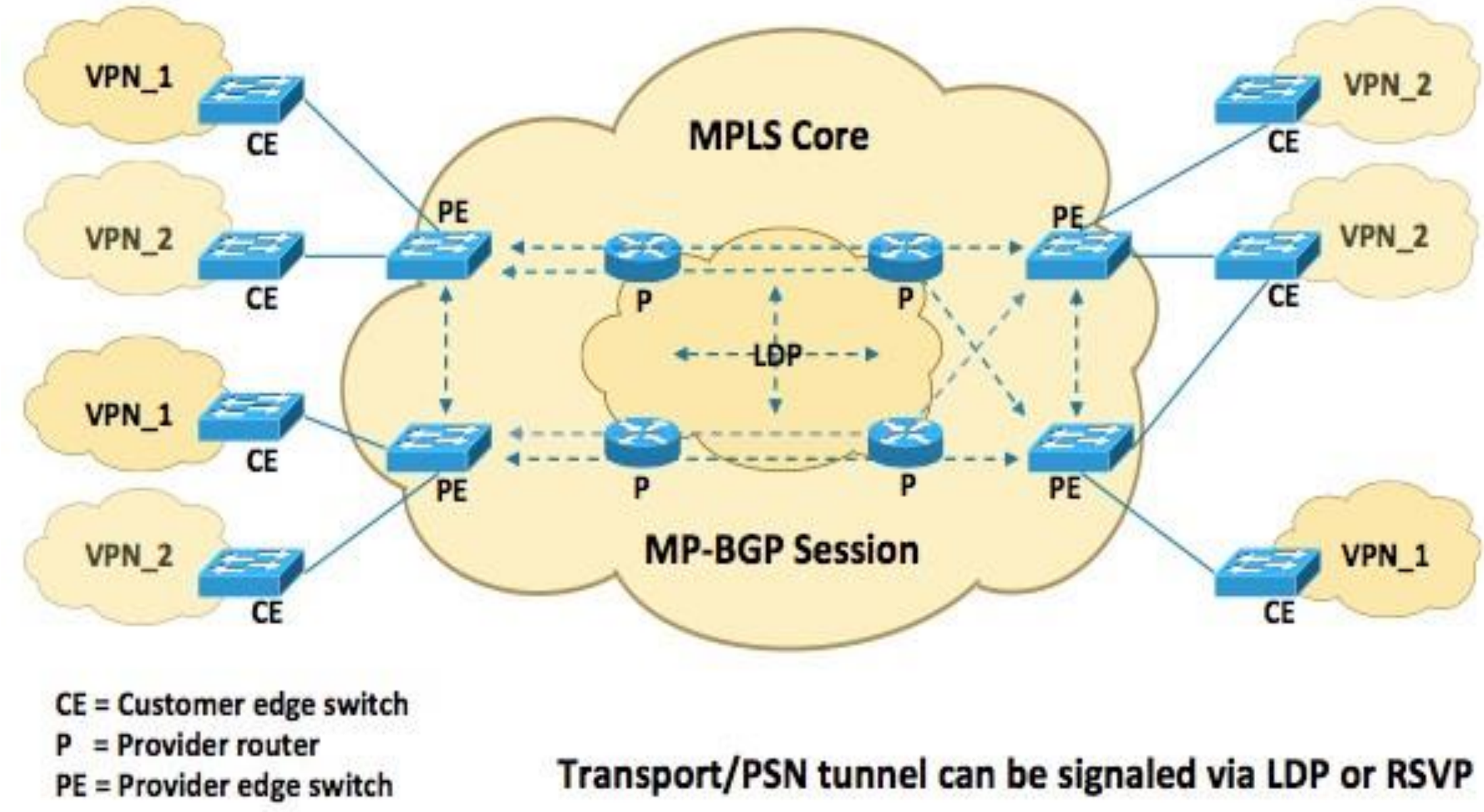
Still flood and learn over the Attachment Circuits



BGP MAC adv.
Route EVPN NLRI
MAC M1 via PE1.

- Customer MAC addresses are advertised over the MPBGP (Multiprotocol BGP) control plane. There is no data plane MAC learning over the core network in EVPN. But Customer MAC addresses from the attachment circuit is still learned through the data plane.

Layer 3 MPLS VPN



- Customer runs a routing protocol with the Service Provider to carry the IP information between the sites. As it is stated earlier, static routing is a routing protocol.

- CE devices can be managed by the Customer or Service Provider depending on the SLA.

- Service provider might provide additional services such as IPv6, QoS, and Multicast. By default IPv4 unicast service is provided by the Service Provider in MPLS Layer 3 VPN architecture.

- Transport tunnels can be created by LDP or RSVP. RSVP is extended to provide MPLS Traffic engineering service in MPLS networks.

- Inner label which is also known as BGP label provides VPN label information with the help of MPBGP (Multiprotocol BGP). This label allows data plane separation. Customer traffic is kept separated over common MPLS network with the VPN label.

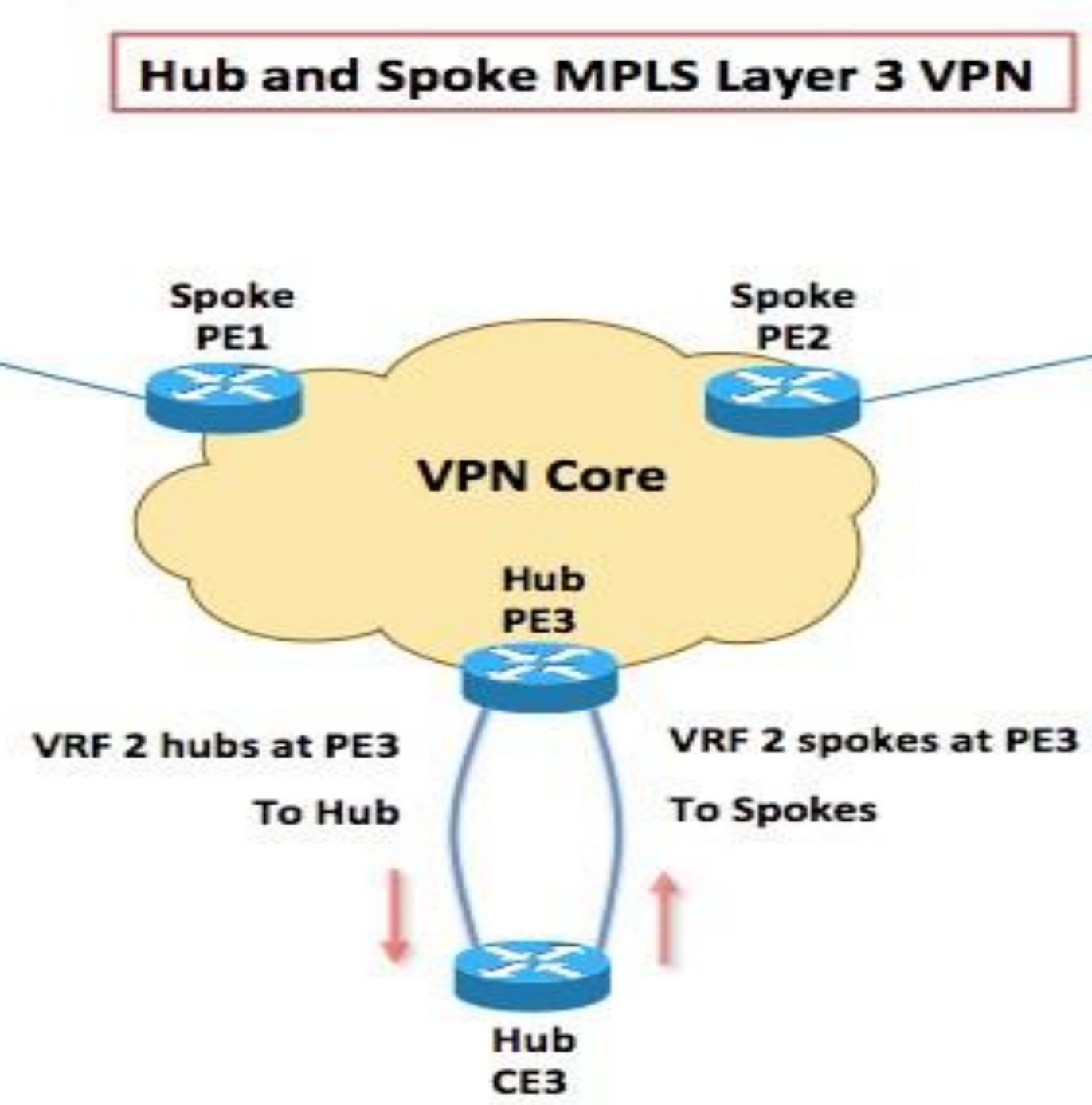
- MP-BGP session is created between PE devices only. P devices don’t run BGP in MPLS environment. This is known as BGP Free Core design.

- Route Distinguisher is a 64 bit value which is used to make the customer prefix unique throughout the Service Provider network. With RD (Route Distinguisher) different customers can use the same address space over the Service Provider backbone.

- Route Target is an extended community attribute is used to import and export VPN prefixes to and from VRF. Export route Target is used to advertise prefixes from VRF to MP-BGP, Import Route Target is used to receive the VPN prefixes from MP-BGP into customer VRF.

- MPLS Layer 3 VPN by default provides any to any connectivity (multi point to multipoint) between the VPN customer sites. But if customer wants to have Hub and Spoke topology, Route Target community can provide the flexibility.

Hub and Spoke MPLS VPNs



MPLS VPN Resiliency

- Customers for the increased resiliency may want to have two MPLS connections from the different service providers. Primary and secondary VPNs are same type of VPN in general, so if the primary is Layer2 VPN , since this is the operational model which customer wants to operate, secondary link from the other provider also is chosen as layer2 VPN.

- If Layer3 VPN is received from one service provider, second link from the different provider is also received as Layer3 VPN.

- Of course, neither MPLS Layer2 VPN nor Layer 3 VPN doesn’t have to have MPLS VPN as a backup, but the Internet or any other transport can be a backup for the customer.

- Look at the comparison chart for single vs dual MPLS VPN design, for choosing Single vs. Dual Providers.

Inter AS MPLS VPN

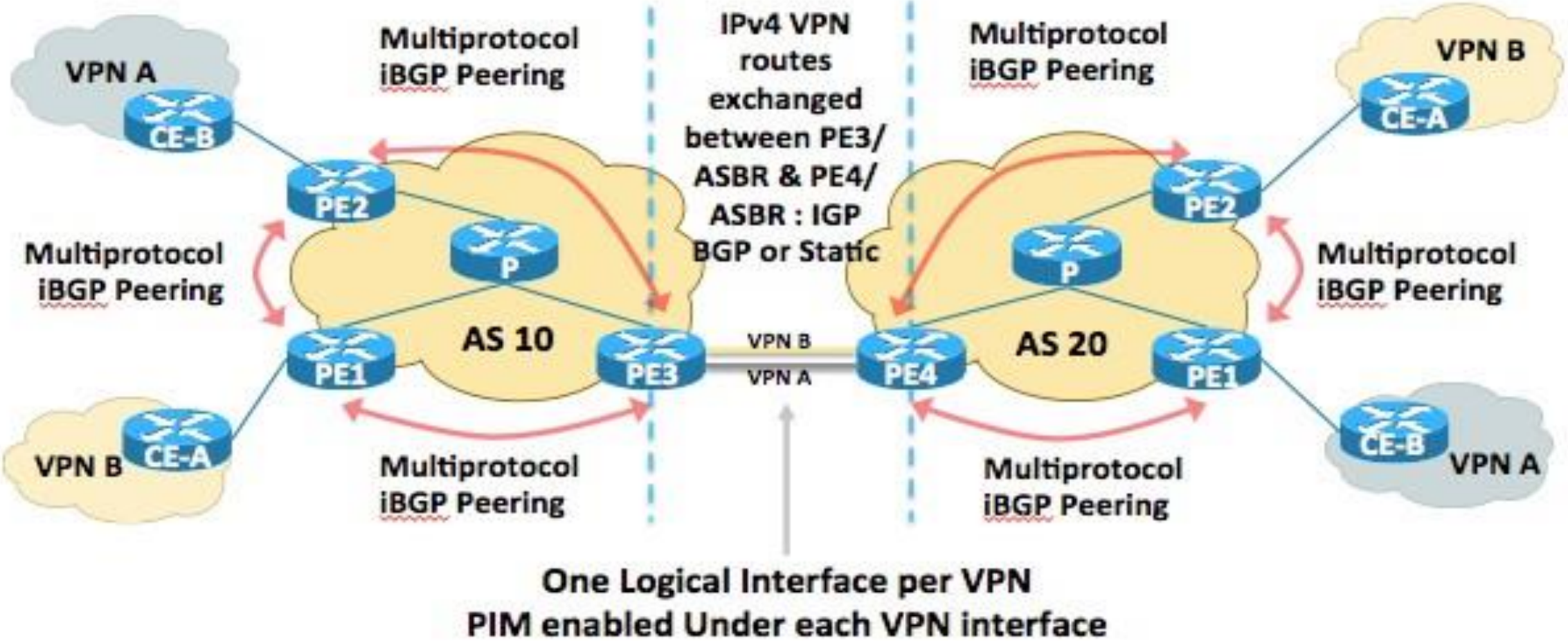
- Customer might be connected to two different service providers. This might be redundancy purpose or maybe Service Provider may not have a POP in some of its customer locations.

- This requires VPN agreement between the Service Provider to support their customers end-to-end MPLS VPN deployment

- There are 3 model defined in RFC 2547 for Inter AS MPLS VPNs. Inter AS Option A. It is also known as 10A

Inter AS MPLS VPN Option A

Instead of Full Mesh MP-IBGP Peering inside an AS, VPN Route Reflector can be used.
Inter-AS Option A requires all customer VRFs and VPN routes on the ASBRs.



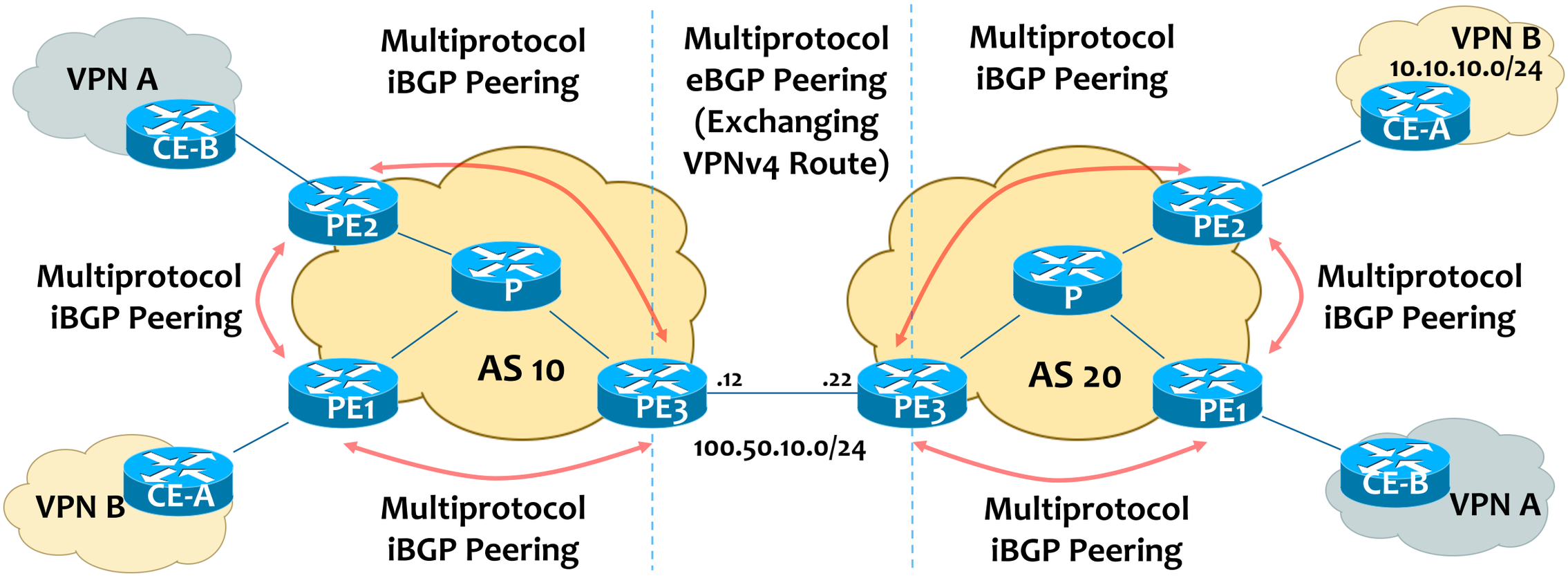
- Inter AS Option A is known as Back to Back VRF approach as well. Service Provider treats each other as customers. Between the service providers, there is no MPLS but only IP routes are advertised.

- For each customer VPN, one logical or physical link is setup. Over the link, any routing protocol can run. But in order to carry end to end customer routing attribute, it is ideal to run the same IGP at the customer edge and between ASBRs.

- This maybe complex from the configuration point of view than running BGP for every customer. Routes from MPBGP is redistributed to IGP and vice versa on the ASBRs. ASBR keeps both MPBGP information in the BGP table and so in the in the routing table. Inter AS Option B removes this restriction.

- Since there is no internal routing information shared between the Service Providers, Inter AS Option A is seen as most secure among all the Inter AS VPNs.

- But since there is huge operational work needs to be done especially on the ASBRs, it is the least scalable Inter AS MPLS VPN approach among all the others.



Inter AS MPLS VPN Option B

- In Inter AS Option B, there is no more one separate logical or physical link per customer VPN.

- Inter AS MPLS Option B don’t require VRF on the ASBR (Autonomous System Boundary Router).

- Instead MPBGP runs for the VPN address family and advertises a customer routes between ASBRs.

- In the both Service Provider network, ASBRs run internal VPNv4 either full mesh or between RRs so every participant PE for the customer VPN receives the prefixes.

- MPLS runs between the Service Providers.
- ASBRs don’t have to keep VRF for the customer. They just need to have VPN information for the customers. This information is advertised through Multiprotocol BGP inside the Service Provider Network.

- ASBRs set the next hop from IBGP to EBGP. But from EBGP next hop doesn’t change by default. Then either on the ASBR next hop self is enabled or ASBR to ASBR link is advertised into Service Provider global routing table to have BGP next hop reachability from the PEs.

- Compare to Inter AS MPLS Option A, Inter AS Option B is more scalable but if the ASBR to ASBR link is advertised, since there will be shared routing information between the providers, it is considered as less secure then Option A

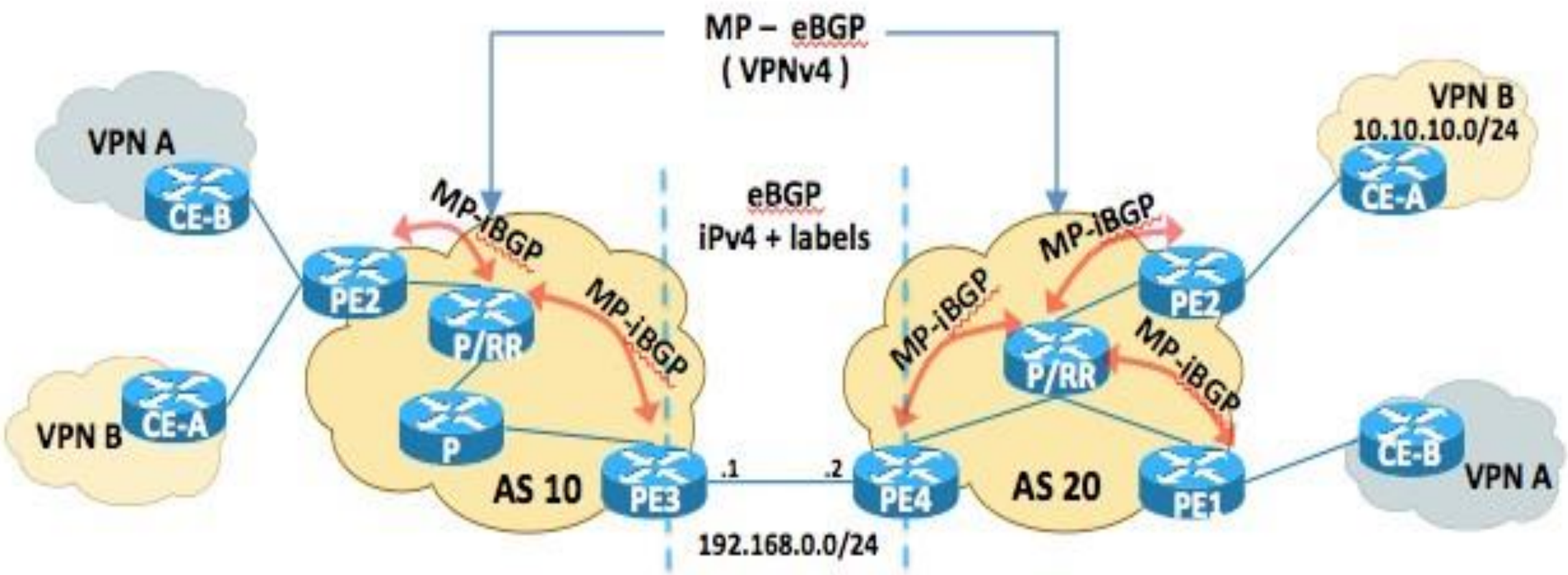
- Whenever BGP next hop changes, new VPN label is assigned by the next hop device. In the Option B case, if BGP next hop self is enabled on the ASBR, ASBR allocates a new

INTER AS MPLS VPN OPTION C

Inter AS MPLS VPN Option C

Instead of Full Mesh MP-IBGP Peering inside an AS, VPN Route Reflector can be used.

VPNv4 between Route Reflectors, RFC3107 between ASBRs in Inter-AS Option C



MPLS Traffic Engineering

- If the requirement is one of the below then the MPLS Traffic Engineering is a good choice

- SLA – Admission Control via RSVP

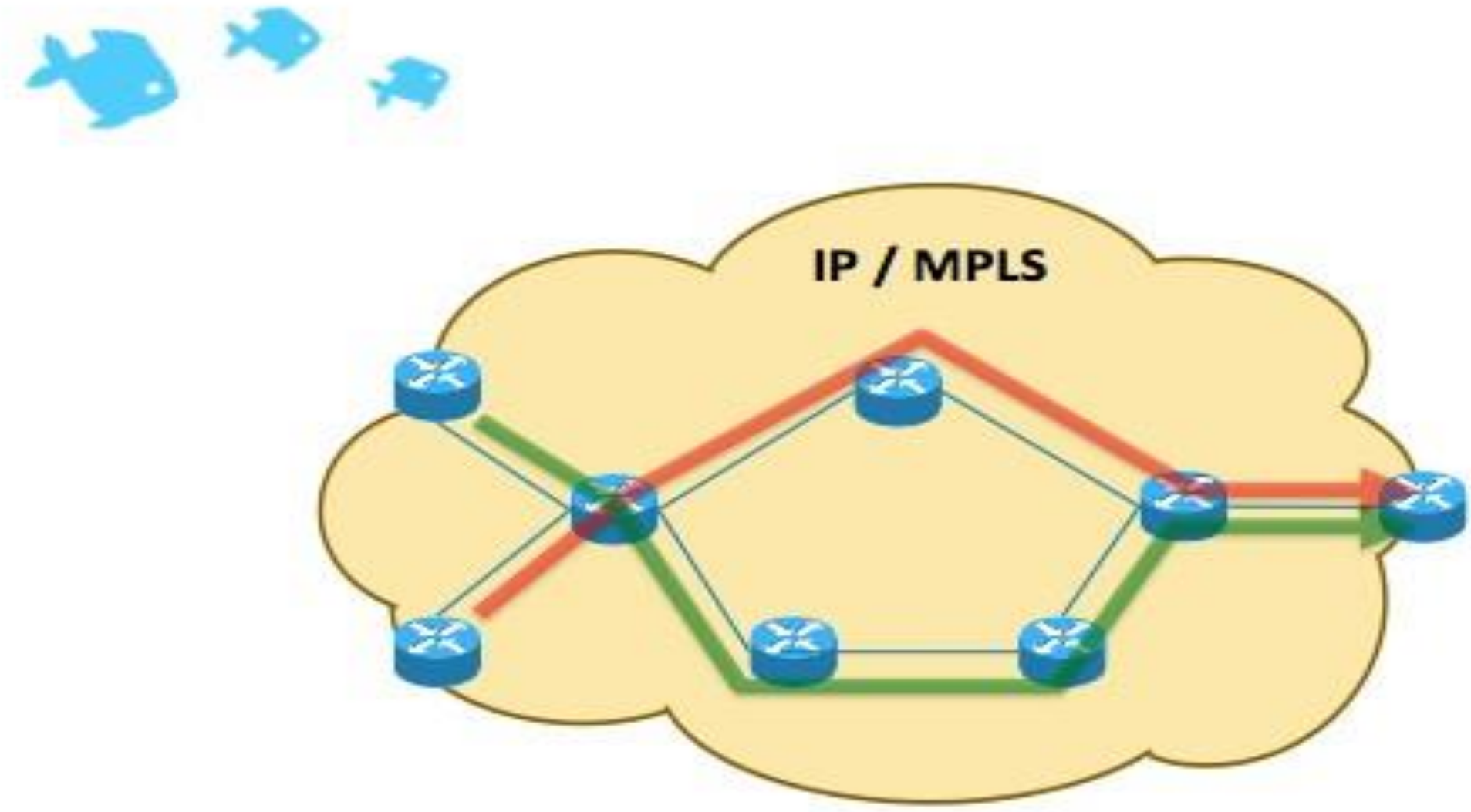
- Traffic Engineering – Strategic or Tactical

- Traffic Protection – RSVP –TE extension

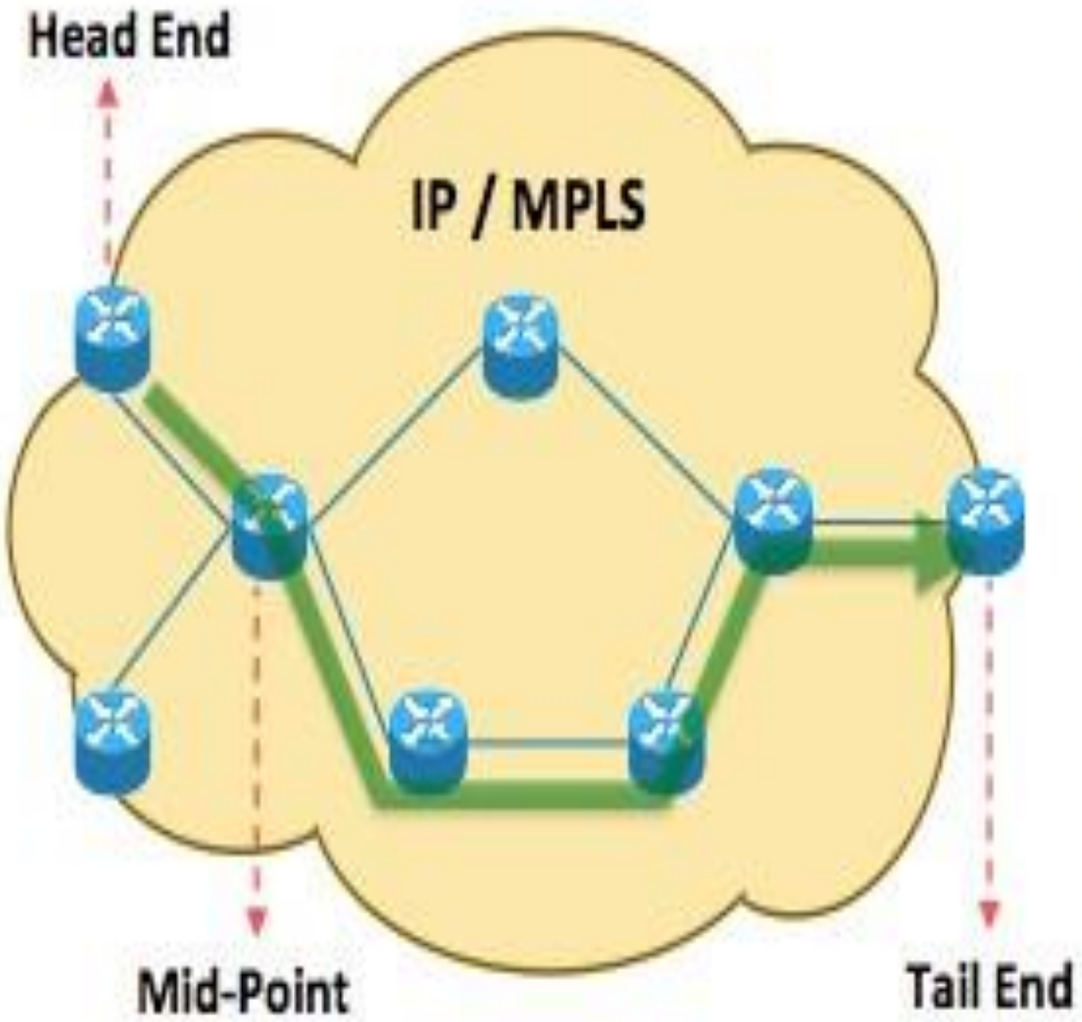
- IP and MPLS is a destination based routing. With MPLS Traffic Engineering, source routing is achieved.

MPLS Traffic Engineering

- MPLS Traffic Engineering allows explicit routing.
- From Head-End, entire path can be calculated and signaled.
- Below is a classical fish diagram. If IP/MPLS would be enabled only, destination based shortest path routing would force top path to be used.
- And bottom path would never be used.
- Only if top path fails, after topology convergence, bottom path could be used. MPLS Traffic Engineering may provide optimal traffic usage.



Classical Fish Diagram of MPLS Traffic Engineering.
Without MPLS TE, IGP protocols always chooses shortest path.
Source routing is not possible with IGP protocols



Link Information Distribution

-- ISIS –TE

-- OSPF-TE

Path Calculation (CSPF)

Path Setup (RSVP-TE)

Forwarding Traffic down Tunnel

-- Auto-route announce

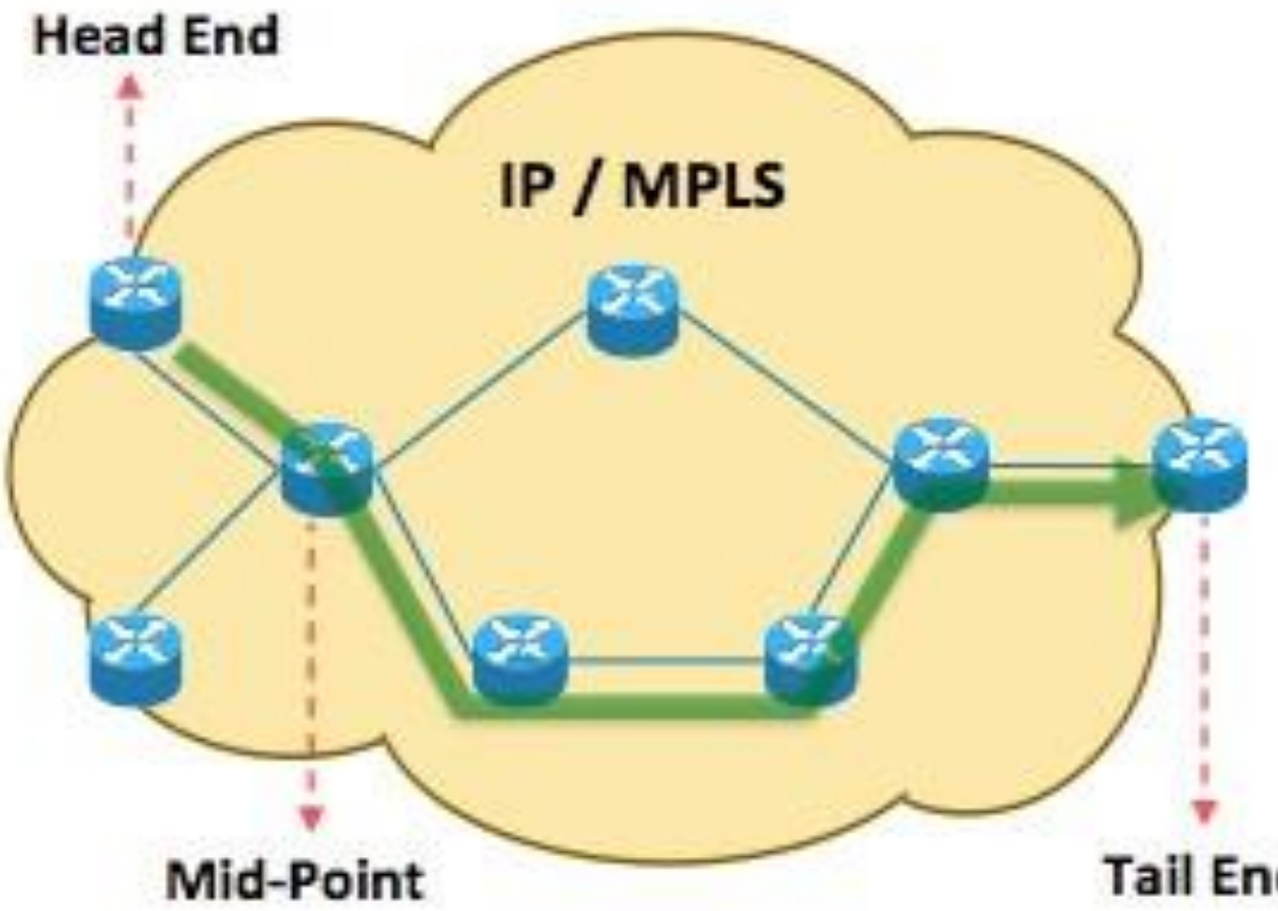
-- Static route

-- CBTS

-- PBR

-- Forwarding Adjacency

-- Pseudowire Tunnel selection



Link Information Distribution

-- ISIS –TE

-- OSPF-TE

Path Calculation (CSPF)

Path Setup (RSVP-TE)

Forwarding Traffic down Tunnel

-- Auto-route announce

-- Static route

-- CBTS

-- PBR

-- Forwarding Adjacency

-- Pseudowire Tunnel selection

- 4 necessary steps to calculate the paths and sending the traffic in MPLS Traffic Engineering

- 4 steps is necessary for creating an MPLS Traffic Engineering path and sending traffic down to that path.

- Information distribution. This can be done by the OSPF or IS-IS. It requires link state topology database. Only OSPF and IS-IS can provide.

- You may hear BGP-LS. BGP-LS is a BGP link state, new BGP address family which carries link state information over BGP. The purpose is even in multi-area, multi-level OSPF- IS-IS design; carry the topology information at the specific nodes from IGP to BGP by redistribution, then from BGP nodes to the BGP RR and from BGP RR to the SDN controller such as Stateful PCE.

- Second step is topology calculation. Topology can be calculated either in distributed manner with CSPF, or as a centralized through NMS, controller etc. Below picture shows how path is calculated for the given constraints.

- MPLS Traffic Engineering allows constraint-based routing. Bandwidth, SRLG, Administrative group can be a constraint.

- Traffic Engineering Database is created with the help of link-state routing protocols only. Input from link state database can be carried to offline tool to calculate ERO (end to end label switched path). Or calculation is done in a distributed manner by CSPF (Constrained based Shortest Path First).

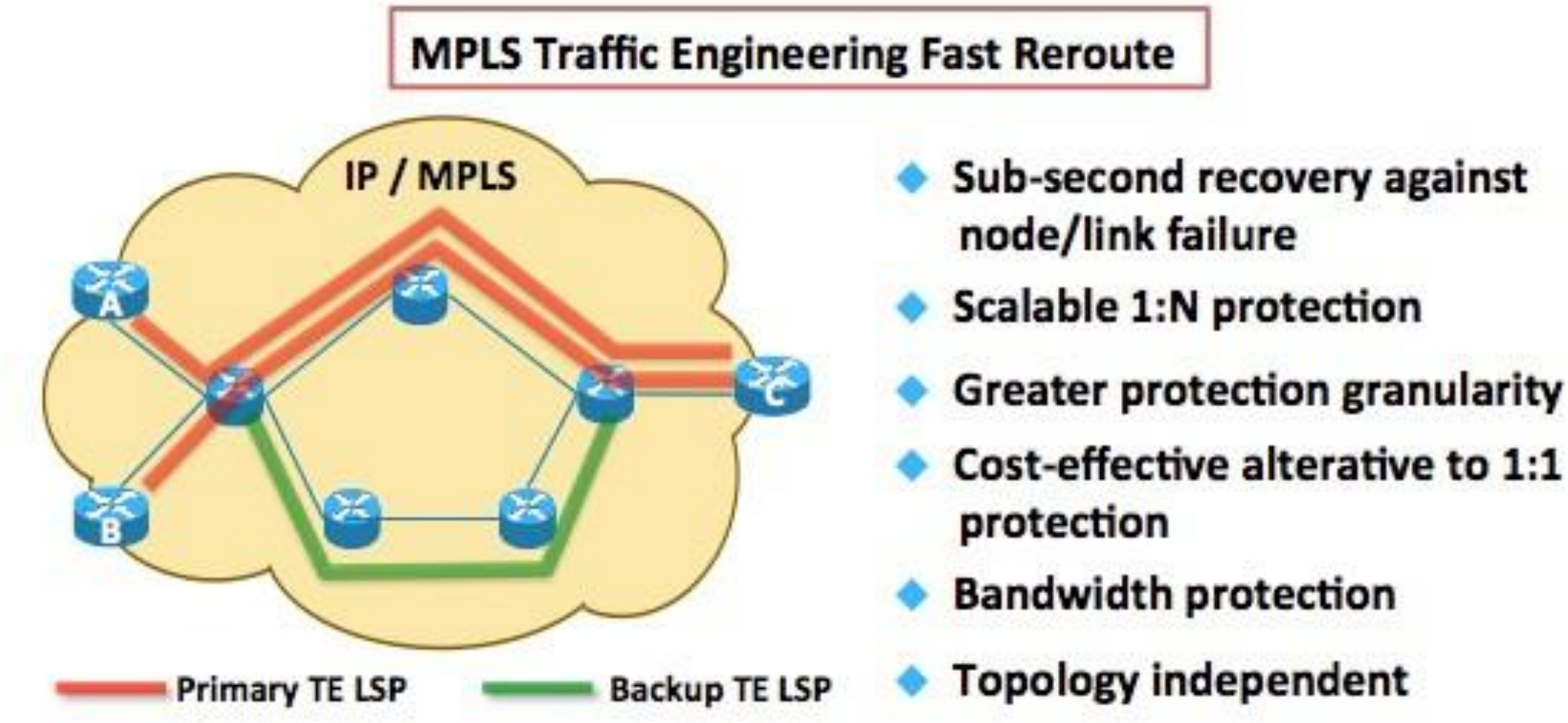
- OSPF and IS-IS has been extended to carry additional link state attribute for better traffic engineer with MPLS. In the below picture, additional link state information is shown. This information is not kept in LSDB but they are kept in TED (Traffic Engineering Database).

Link Attributes for MPLS Traffic Engineering Database:

Link Attributes for MPLS Traffic Engineering Database	
Additional link Characteristics	• Interface Address • Neighbor Address • Physical bandwidth • Maximum reservable bandwidth • Unreserved Bandwidth (at eight priorities) • TE metric • Administrative group (attribute flags)
ISDIS or OSPF flood link information	
All TE nodes build a TE topology database	
Not required if usingoffDline path computation	

MPLS Traffic Engineering Fast Reroute

Fast reroute is a local protection mechanism.

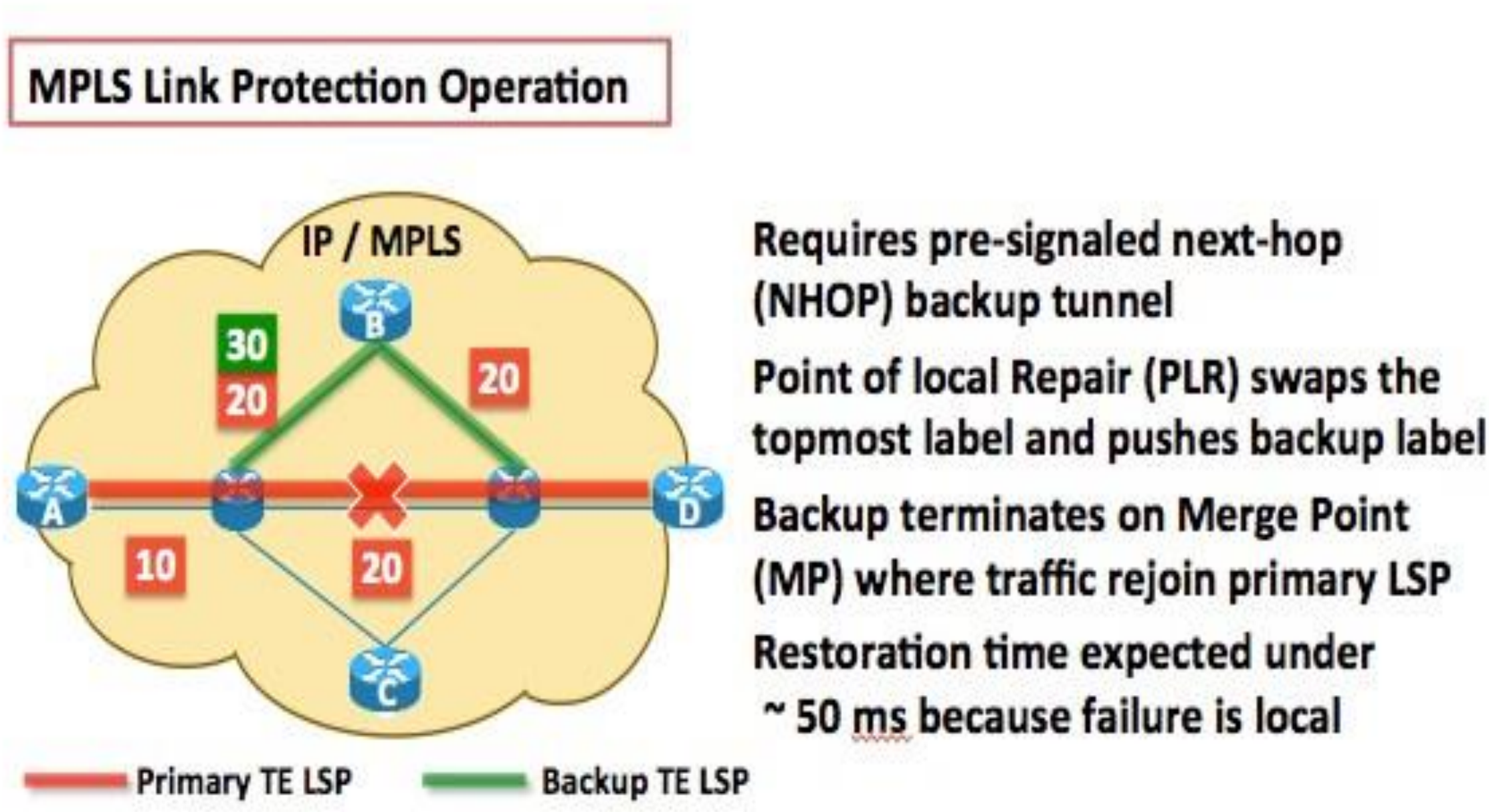


The router reacts to a failure is called Point of Local Repair.

- Whenever failure happens, traffic continues through alternate path since MPLS Traffic Engineering Fast reroute is a data plane protection mechanism and backup path can be used as soon as failure is detected

- Control plane still converges. If control plane finds more optimal path than TE FRR backup LSP, then new more optimal primary path is signaled in an MBB (Make Before Break) manner.

MPLS Traffic Engineering Fast Reroute Link Protection



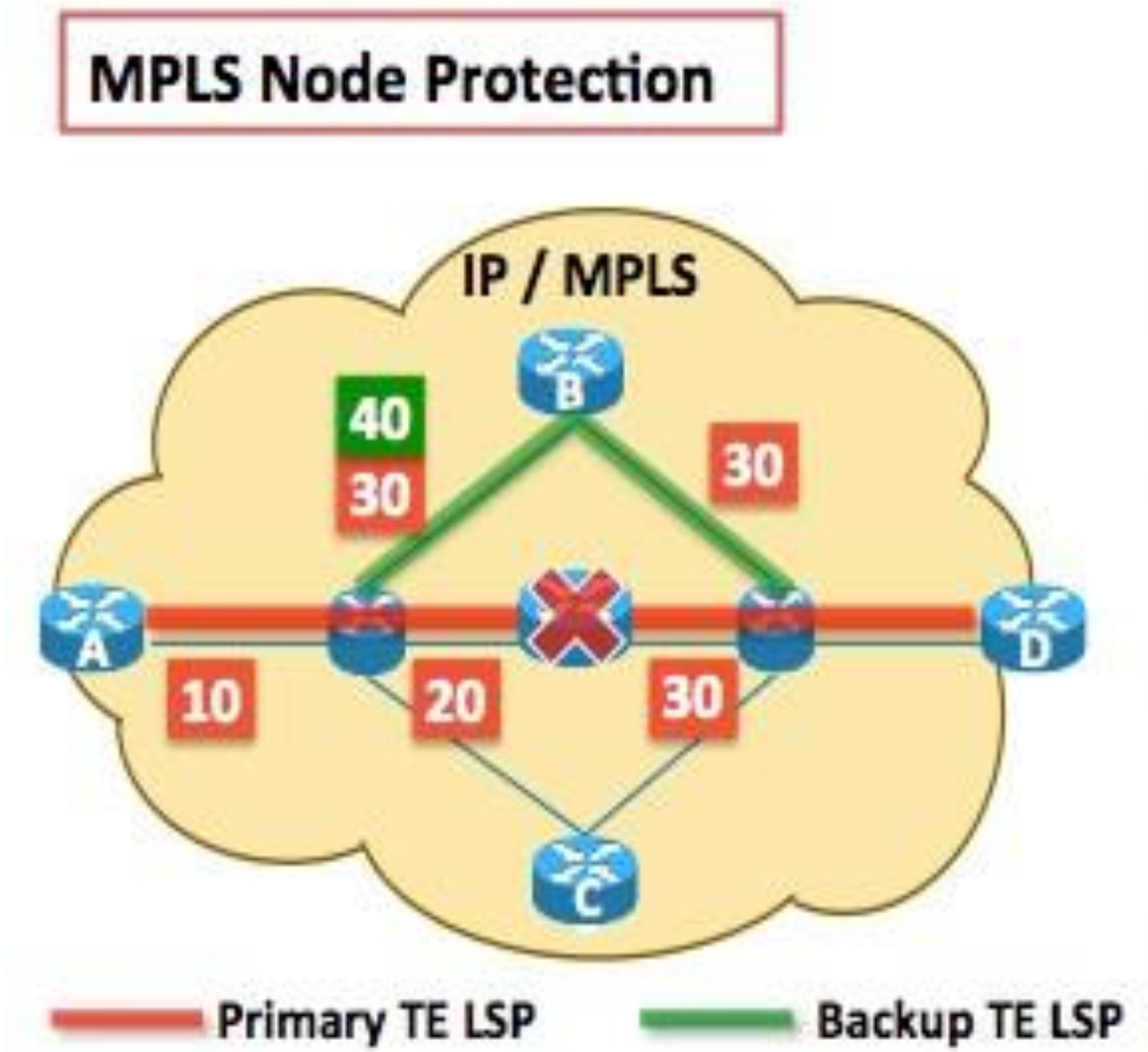
MPLS Traffic Engineering Fast Reroute Node Protection

- Most of the failure is a link failure in the networks. Node failure is less common compare the link failure. Thus many networks only enable link protection.

- MPLS traffic engineering fast reroute can cover all the failure scenarios. An IP Fast reroute technology such as LFA (Loop Free Alternate) requires highly mesh topologies to find an alternate path, which will be programmed in the data plane.

- If the topology is a ring, then LFA cannot even work. It requires a tunnel to the PQ node. Remote LFA is another IP fast reroute technology, which allows to be created a tunnel from the PLR to the PQ node.

MPLS Traffic Engineering Fast Reroute Node Protection



Requires pre-signaled next-next- hop (NNHOP) backup tunnel

Point of local Repair (PLR) swaps the next-next hop label and pushes backup label

Backup terminates on Merge Point (MP) where traffic rejoin primary LSP

Restoration time depends on failure detection time

MPLS Case Study – 1

- Orefe is a supermarket company, operated in Turkey. Most of their stores are in Istanbul but they have 46 stores operated in the cities close to Istanbul.

- They recently decided to upgrade their WAN (Wide Area Network). They have been using Frame Relay between the stores, HQ and their datacenters and due to limited traffic rate of frame relay, they want to continue with the cost effective alternative. Also the new solution should allow Orefe to have higher capacity when they need.

- After their discussions with their network designer, they decided to continue with the MPLS layer 3 VPN.

- Main reasons for Orefe to choose MPLS Layer 3 VPN is to handover the core network responsibility to the Service Provider. If they would choose Layer 2 service, their internal IGP which is OSPF would be extended over WAN as well and their networking team although has brilliant engineer, due to increase operational load, MPLS layer 3 VPN has been chosen by Orefe.

- Since they are using OSPF as their internal IGP in 2 Head Quarter, 3 Datacenter and 174 Branch Offices across the country, Orefe wants to have OSPF routing protocol with their service provider.

- Kelcomm is the fictitious service provider, which provides an MPLS VPN service to Orefe. Unlike other service provider, which only provides BGP and static routing to their MPLS Layer 3 VPN customers, Kelcomm agreed to run OSPF with Orefe.

- Orefe has a VPN link between its 2 Head Quarters. They will keep that link as an alternate to MPLS VPN. In case MPLS link fails, best effort VPN link over the Internet will be used as a backup.

➢ Please explain the traffic flow between two Head Quarters of Orefe ?

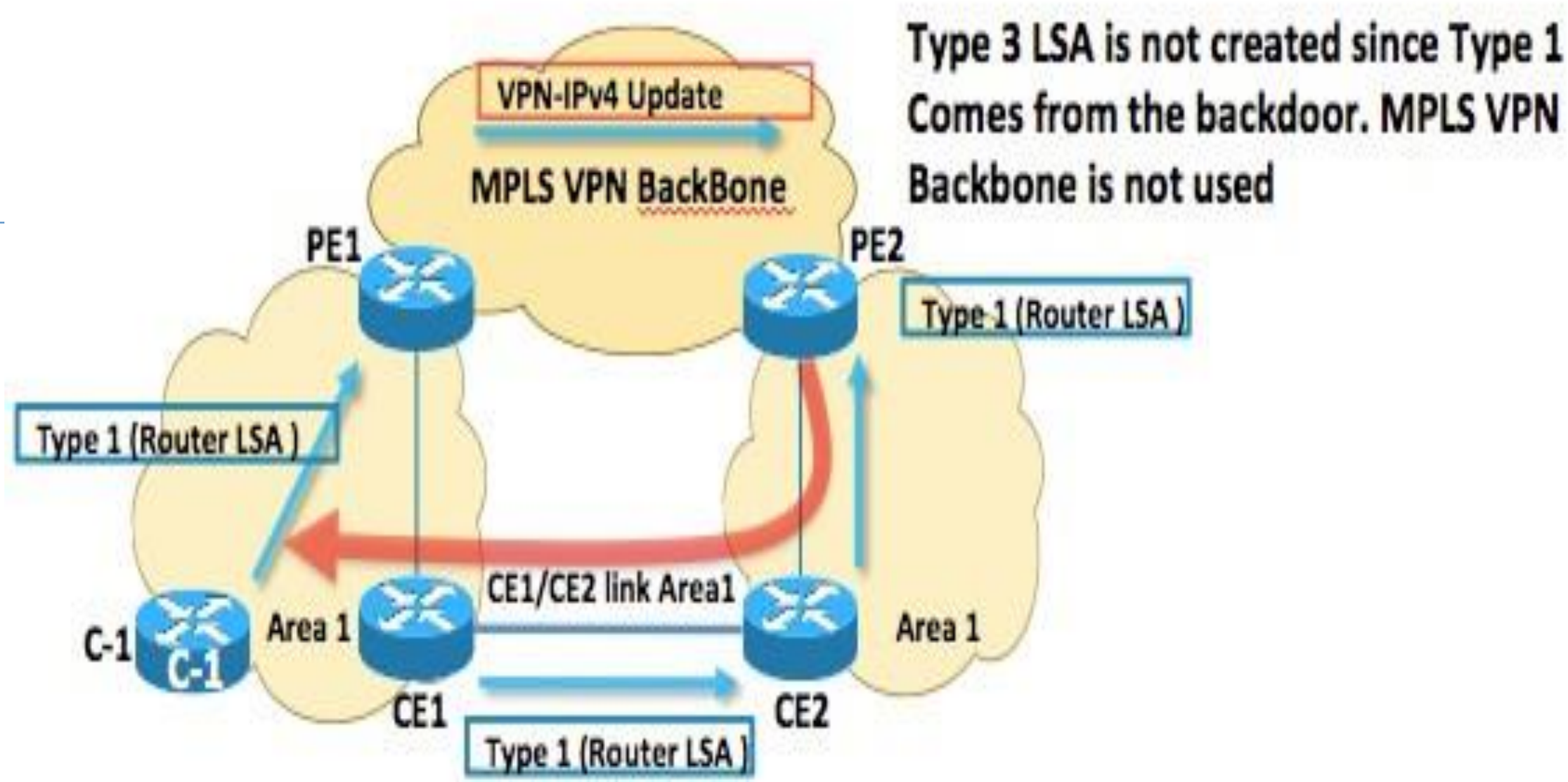
➢ What might the possible problems ? How can those be avoided ?

❖ Both Head Quarters are in the OSPF Area1 , Backdoor VPN link is in Area 1 as well. Topology is shown below.

❖ If OSPF is used as PE-CE protocol in MPLS Layer 3 VPN environment, the rule is, routes are received as type 3 LSAs over the MPLS backbone if the domain ID is the same.

If they are different, then routers are received as OSPF Type 5 LSAs.

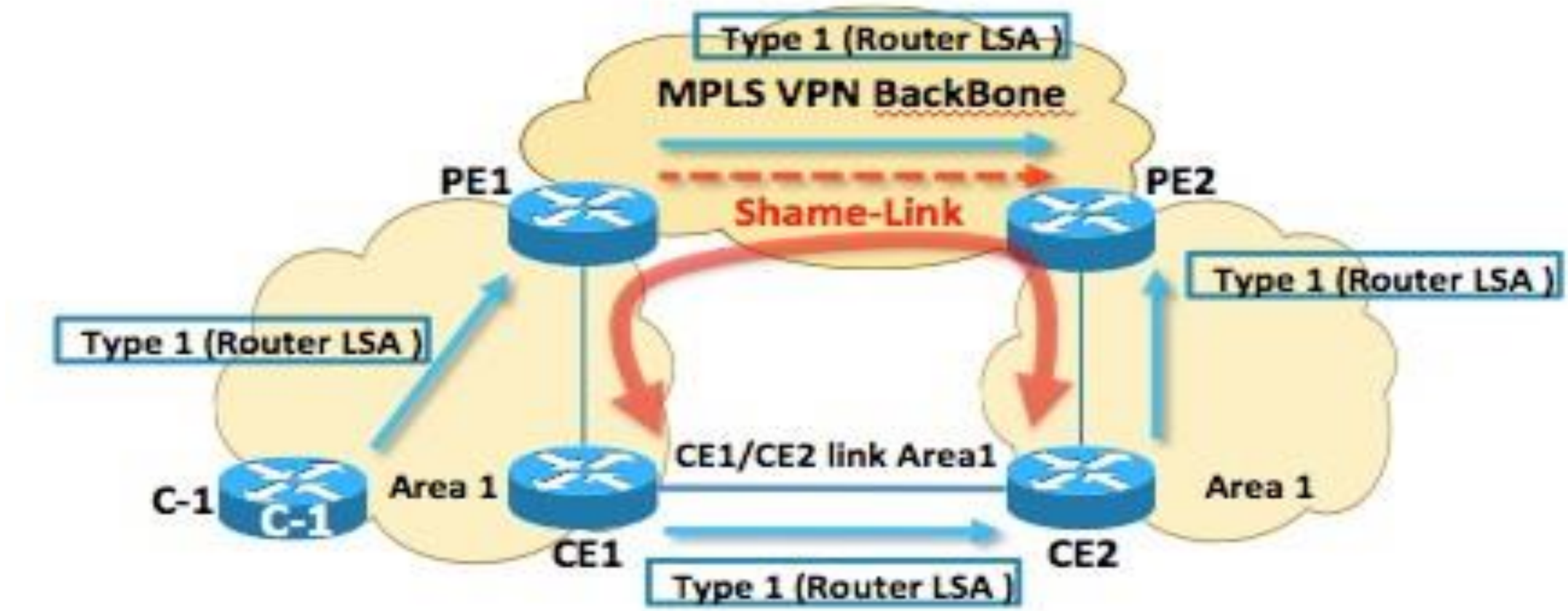
If domain ID is not set exclusively, then by default Process ID is used for domain ID.



- As it can be seen from the above picture, the backdoor VPN link (Best Effort- No Service Guarantee) is used as primarily. Customer doesn’t want that because they pay for guaranteed SLA so they want to use MPLS backbone as primary path.

- But OSPF sends prefixes over the backdoor link as type 1 LSA. When PE2 at the remote site receives the prefixes via Type 1 OSPF LSA, it doesn’t even generate Type 3 LSA to send down a CE2.

- Two approaches can help to fix this problem. One option is shown as below. OSPF Sham-link.



With only Metric manipulation now,
MPLS Backbone can be made preferable

- With the OSPF Sham-link, PE2 will send OSPF Type1 LSA towards CE2. And only with metric manipulation, MPLS backbone can be made preferable.

- Another approach would place the PE-CE link into Area0. For the Head Quarters, Orefe would have been put those links in Area 0 in the first place. If multi area design is required, then Orefe should place the Branch offices to be in non-backbone area.

- Once PE-CE links are placed in Area0, then backdoor link should be placed in different area. This makes CE1 and CE2 an ABR.

- They receive the prefixes over backdoor link as type 3, without Sham-link they receive also as Type 3 (Assume Domain-ID, Process ID matches between PEs), and then only with metric manipulation, MPLS backbone can be made preferable.

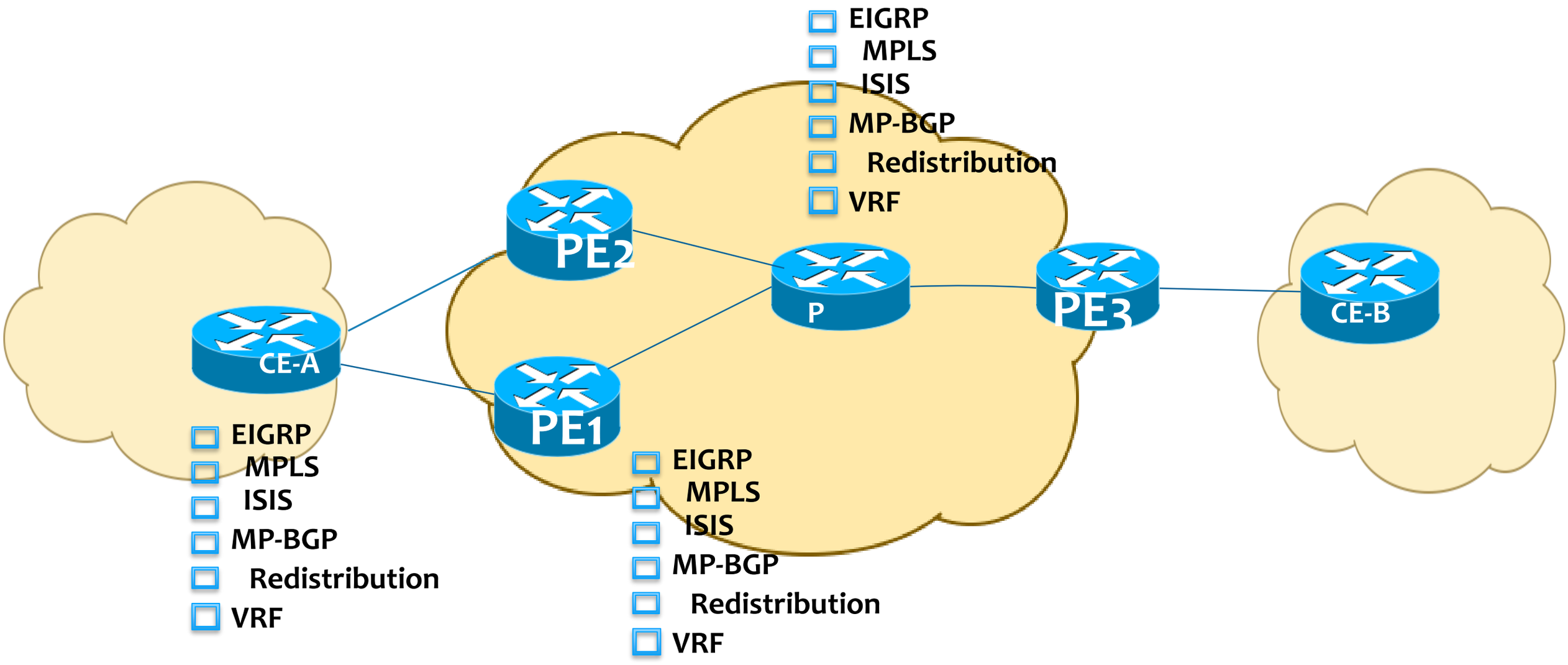
MPLS Case Study 2

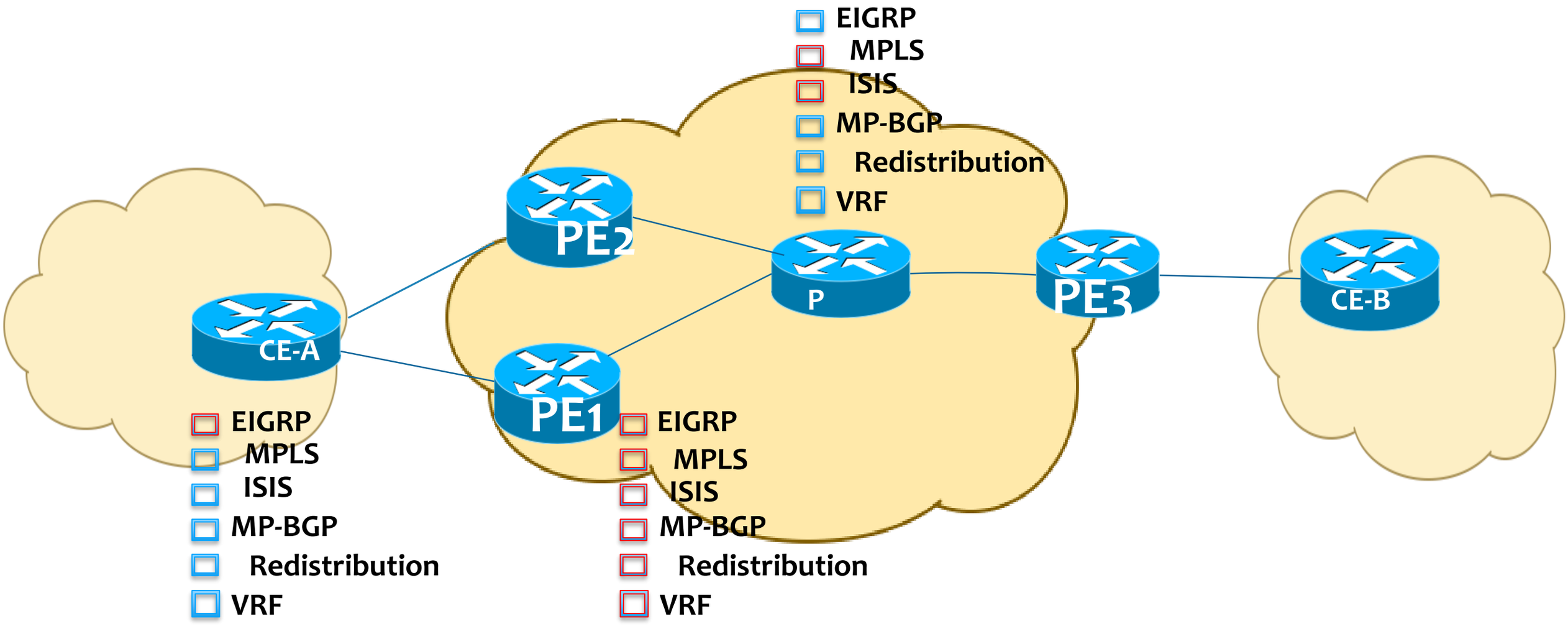
- Enterprise Company is using MPLS Layer 3 VPN for their Wide Area Network connections.

- Their PE-CE IGP protocol is EIGRP

- Service Provider’s of the company is using IS-IS as an internal infrastructure IGP in their backbone and LDP for the label distribution

Please can you select the all required protocols from the check box near to the shown routers ?





MPLS Case Study – 3

- Maynet is a fictitious service provider. They have MPLS on their core network. They provide MPLS layer 2 and layer 3 VPN services to their business customers.

- In Access and Aggregation network Maynet doesn’t run MPLS but they are also considering enabling MPLS towards Aggregation first and finally to the access networks.

- Recently they reconsidered the Core Network Availability and they decided to enable MPLS Fast Reroute between all edge devices in their core network.

- Although due to limited size of edge devices, full mesh RSVP-TE LSP is not a problem for Maynet, protection mechanism suggested by their transport team has serious concern.

- They would like understand your opinion about the issue thus they ask below questions.

- What is MPLS Traffic Engineering Path Protection?
- What are the pros and cons of having MPLS Path Protection ?
- Why transport department is suggesting MPLS TE FRR Path protection instead of local protection technologies?
- Please compare the two architectures and highlight the similarities and differences for Maynet to decide the final architecture.

- MPLS Traffic Engineering Fast Reroute is a local protection mechanism where the nodes local to the failure reacts to a failure.

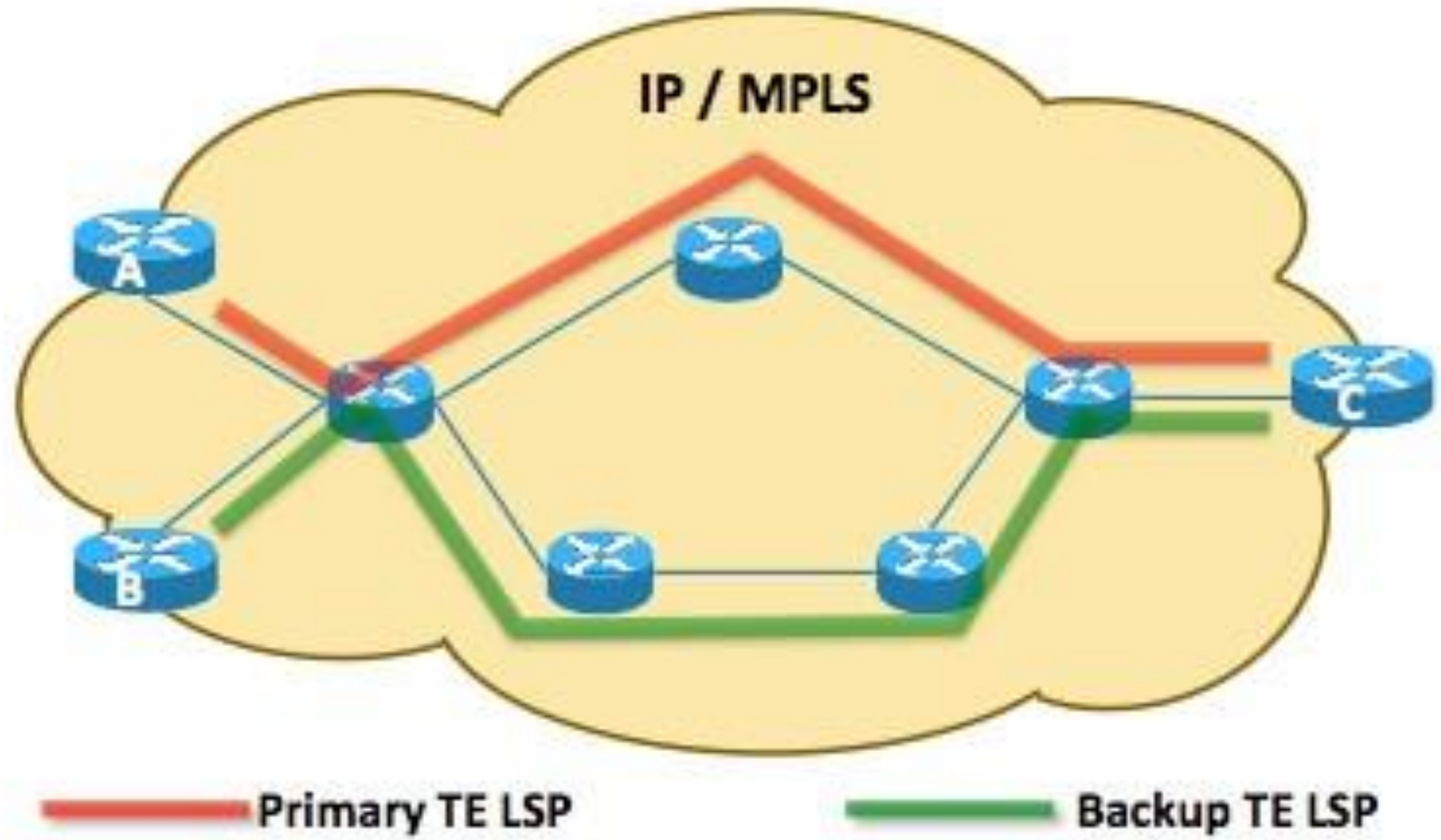
- Control plane convergence follows the data plane fast reroute protection and if more optimal path is found, new LSP is signaled in a MBB (Make Before Break) manner.

- Fast reroute backup LSP can protect multiple primary LSP, thus in the MPLS Traffic Engineering chapter, it is showed as 1:N protection.

- In contrast, path protection is a 1:1 protection schema where the one backup LSP only protects one primary LSP.

- There are two drawback of path protection.

- First one, backup LSP just waits an idle and only can carry the traffic if the primary LSP fails. So this is obviously conflict with the MPLS Traffic Engineering, since the whole idea behind MPLS Traffic engineering is optimize the traffic usage so cost saving.



- As it is depicted in the above picture, green path is a backup path and it cannot pass through any devices or links which primary LSP passes.

- The second biggest drawback of having MPLS Traffic Engineering path protection as opposed to Local protection with the link or node protection is the number of LSP.

- Since one backup LSP is created for each primary LSP, number of RSVP-TE LSP will be almost double compare to 1:N local protection mechanisms.

- In the transport networks, SONET/SDH, OTN, MPLS-TP all have linear protection schema which is very similar to MPLS Traffic Engineering Path Protection.

- That’s why if the decision is taken together with the Transport team, they suggest you to continue their operational model but at the end core network will have scalability and manageability problems.

- Last but not least, switching the alternate path in path protection might be slower than local protection mechanisms since the point of local repair (Node which should reach to a failure) may not be directly connected to a failure point.

- Thus failure has to be signaled to the Head End which might be many hops away from the failure point.

- In the above topology, even the router in the middle of a topology fails, failure has to be signaled to the R2 and R2 switchover to the backup (Green) LSP.

MPLS Case Study – 4

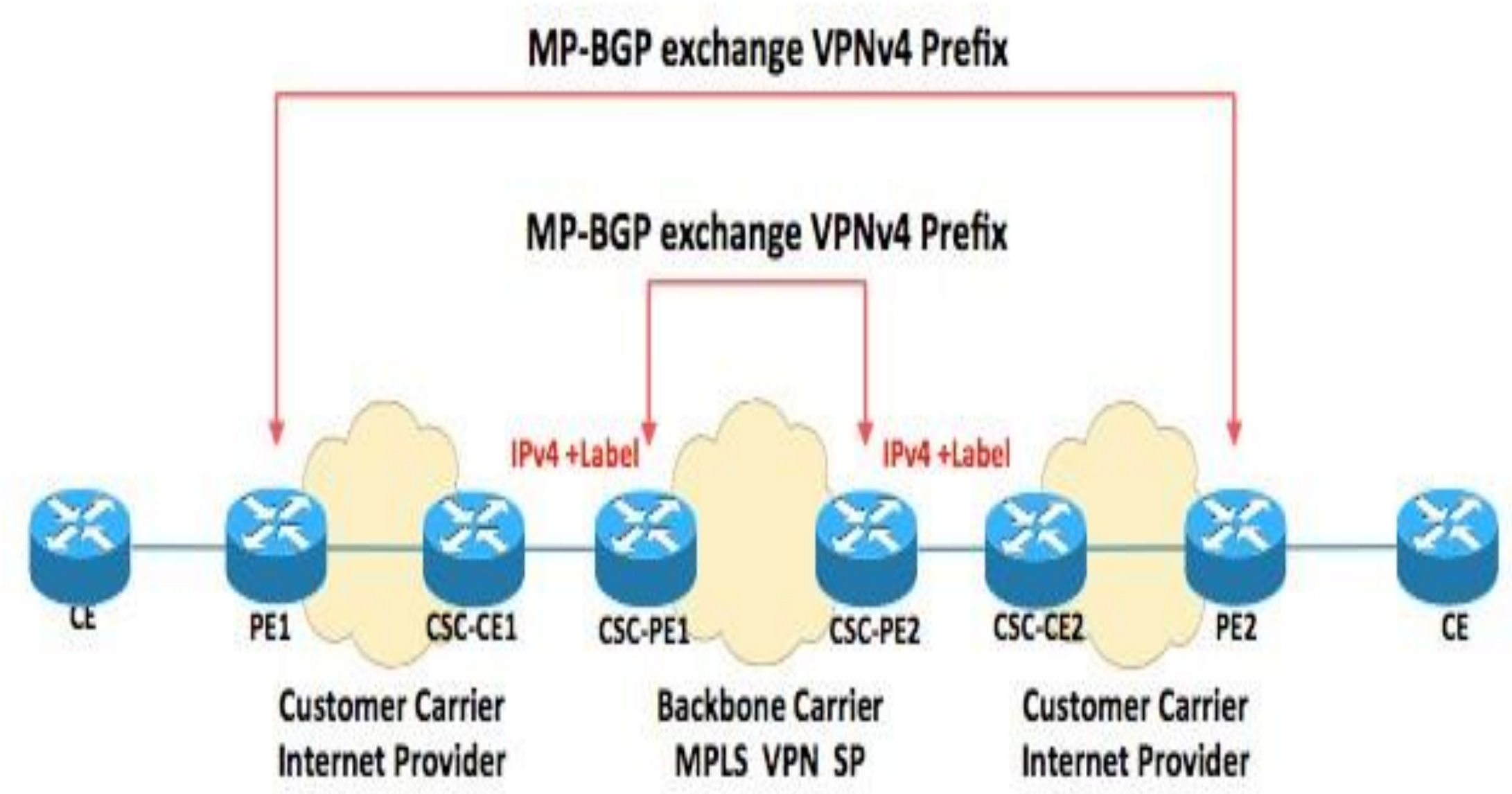
- Smallnet is an ISP which provides Broadband and business internet to their customer. Biggercom is a transit service provider of Smallnet which provides layer 3 IP connectivity between Smallnet sites.

- Smallnet wants to carry all its customers' prefixes through BGP over Biggercom infrastructure. Biggercom doesn't want to carry more than a 1000 prefixes of Smallnet in the Smallnet VRF.

- Smallnet has around 3200 customer prefixes.

- Provide a scalable solution for Biggercom and please explain the drawback of this design.

- Since the requirements are ; Biggercom provides IP connectivity , doesn’t carry more than 1000 prefixes of Smallnet, and this is achieved through Carrier Supporting Carrier architecture.


MP-BGP exchange VPNv4 Prefix
MP-BGP exchange VPNv4 Prefix
IPv4 +Label
IPv4 +Label
CE
PE1
CSC-CE1
CSC-PE1
CSC-PE2
CSC-CE2
PE2
CE
Customer Carrier
Internet Provider
Backbone Carrier
MPLS VPN SP
Customer Carrier
Internet Provider

- In the above picture, Carrier Supporting Carrier architecture is shown.

- In the Carrier supporting Carrier terminology, there is a Customer and Backbone Carrier. In our case study, Smallnet is a Customer Carrier, Biggercom is a Backbone Carrier.

- There is no customer VRF at the Smallnet network.

- Biggercom has different VRF for its individual customer and Smallnet is one of them.

- Smallnet has many Internet customer routes which have to be carried through backbone carrier network. BGP is used to carry large amount of Customer prefixes. If Customer demands full Internet routing table (At the time of this writing it is over 520K prefixes) then BGP already is the only way.

- Thus BGP session is created between Smallnet and Biggercom.

- Over the BGP session’s customer prefixes of Smallnet is NOT advertised. Instead, loopback interfaces of Smallnet Route Reflectors or PEs are advertised.

- IBGP session is created between the Smallnet Route Reflectors. And customer prefixes of Smallnet is advertised and received over this BGP session.

- One big design caveat for Carrier Supporting Carrier Architecture is, between the Customer Carrier and Backbone Carrier MPLS has to be enabled. So between Smallnet and Biggercom network, MPLS and BGP is enabled. The reason of MPLS is to hide the customer prefixes of Smallnet from the Biggercom.

- If MPLS wouldn’t be enabled on the link between Smallnet and Biggercom, Biggercom had to do IP destination lookup on the incoming IP packet which is a customer prefixes of Smallnet. Since Biggercom doesn’t have a clue about the customers of Smallnet, packet would be dropped.

MPLS Case Study – 5

- Orko is an Enterprise company which has a store in 7 countries throughout Middle East. Head Quarter and Main Datacenter of Orko is located in Dubai.

- 65 stores of Orko, all connected to datacenter in Dubai via primary MPLS L3 VPN link. Availability of Orko is important so secondary connections to the datacenter is provided via DMVPN over the Internet

- Orko is working with single service provider. MPLS and Internet circuit is terminated on the same router.

- In order to have a better policy control and scalability reason, Orko decided to run BGP with its service provider over the MPLS circuit.

- Orko doesn’t have Public ASN and Private AS , 500 is provided by its service provider. Orko uses unique AS number 500 on every locations, including its datacenter. In the datacenter, Orko has two MPLS circuit for the redundancy and they are

  terminated on the different routers.

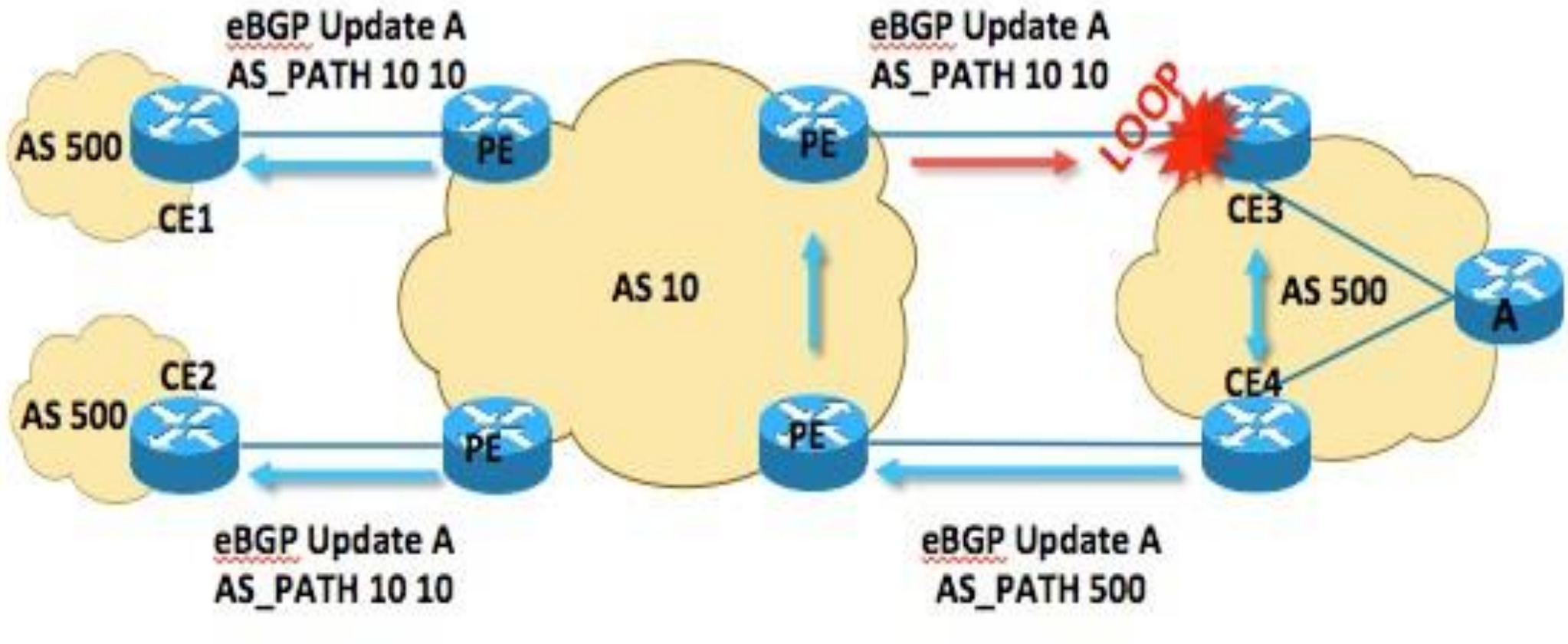
- Would this solution work with BGP as a PE-CE routing protocol ? What can be done to make the solution work ?

- What can be the possible risks and how they can be mitigated ?

  - Since Orko is running BGP everywhere and it uses unique AS number, BGP loop prevention mechanism doesn’t allow the BGP prefixes with the same AS in the AS-path.

- Solution wouldn’t work unless Service Provider implements AS-Override or Orko implements on the every router Allow-as command.

- Even though both solutions would allow the BGP prefixes of Orko in the BGP table of the routers, due to Multi homing in the datacenter, solution creates another problem which is BGP routing loop.

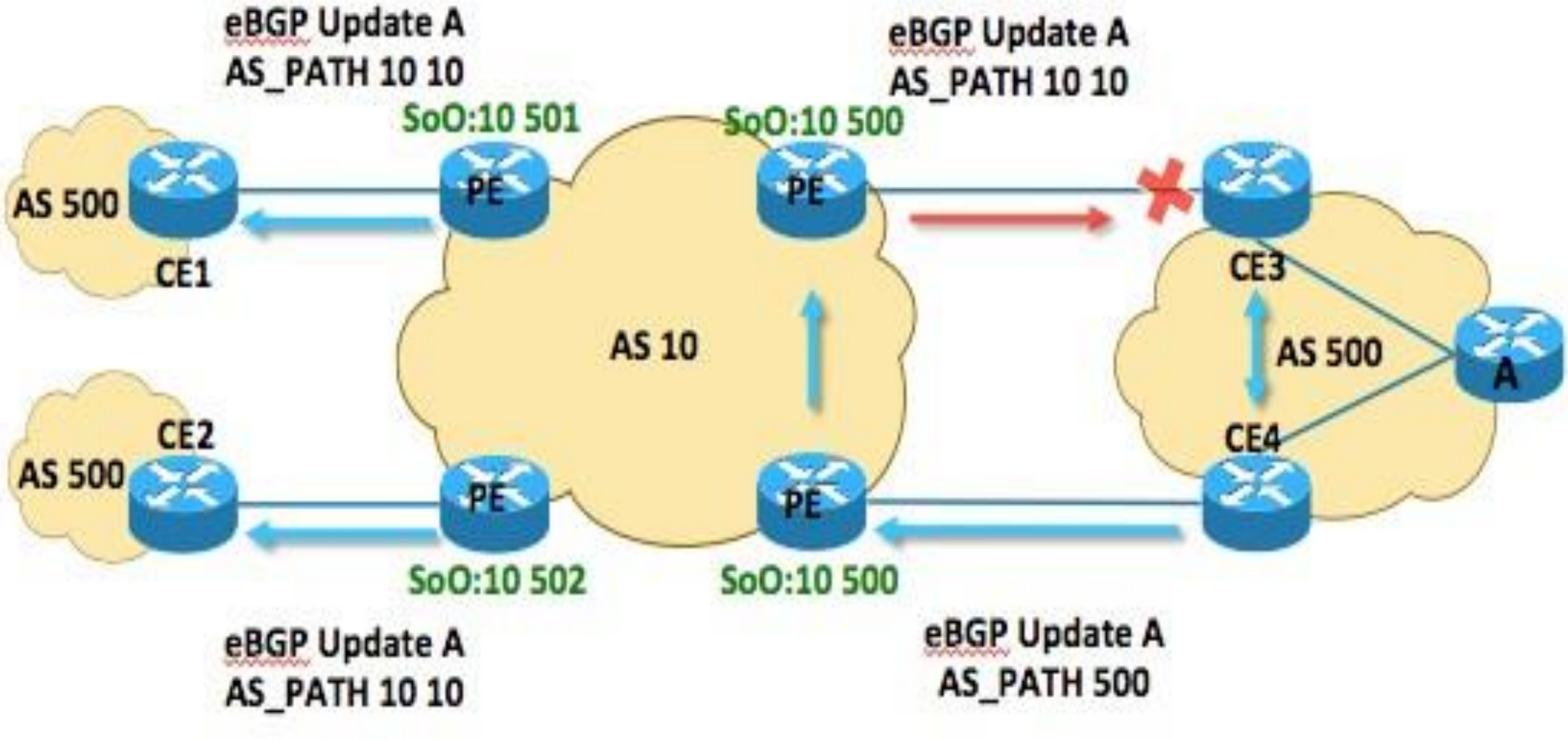


- As it can be seen from the above topology, Site 3 which is the Orko's datacenter originates a BGP prefixes which is advertised to Service Provider PE device, PE3.

- PE3 advertises this prefixes to PE4. Service Provider configures BGP AS Override on its PE toward Orko’s PE-CE link.

- But this creates a problem on the PE4 to CE3. Since prefixes come as “ AS 10, 10 “ , CE 3 would allow locally originated prefixes from the MPLS backbone , and this creates a BGP routing loop

- That’s why, if BGP AS override or Allow-as in is configured, it creates a routing loop at the multi homed site. One solution to this problem can be with BGP Site Of Origin



- SoO 10:500 is set on the PE3-CE3 and PE3-CE4 links. When the PE4 receive the prefixes from PE3, it doesn't advertise the prefixes to CE3.

- SoO 10:500 is set on the PE1-CE1 and SoO 10:500 is set on the PE2-CE2 links

MPLS Case Study -6

Question 1 :

- Two Service Providers company in order to support their customers in other regions decided to create VPN neighborship between them.

- Number of VPN service between them is expected to be too much and their security team definitely disagree to share their internal routing information with other provider.

- Which VPN solution would be the best for this agreement ?

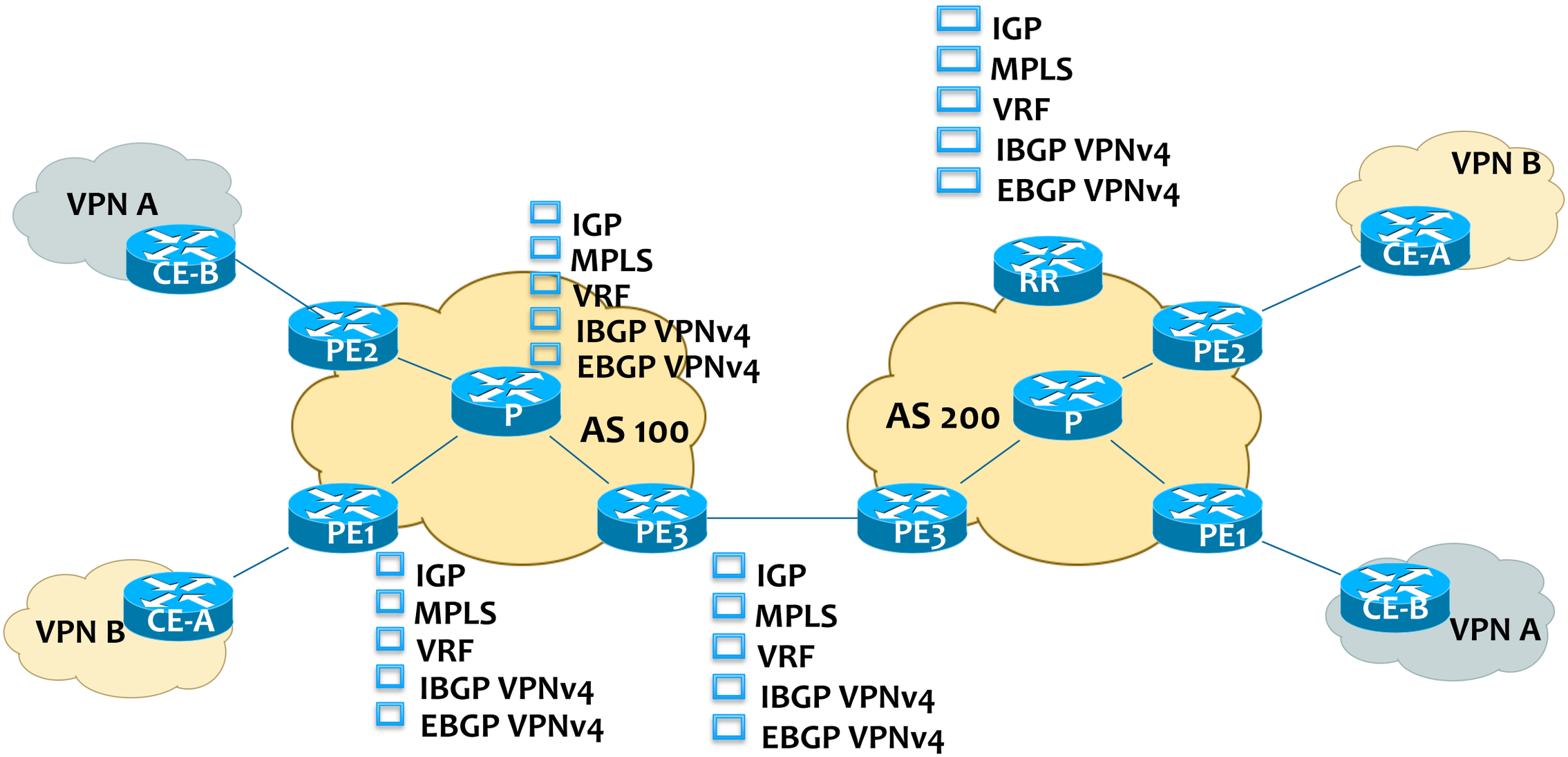
- Since the number of expected VPN customers are too much, among the available Inter-AS MPLS VPN options, Option B and Option C is the suitable ones.

- But Option C requires internal routing information such as PE and VPN RR addresses to be leaked between the Service Providers.

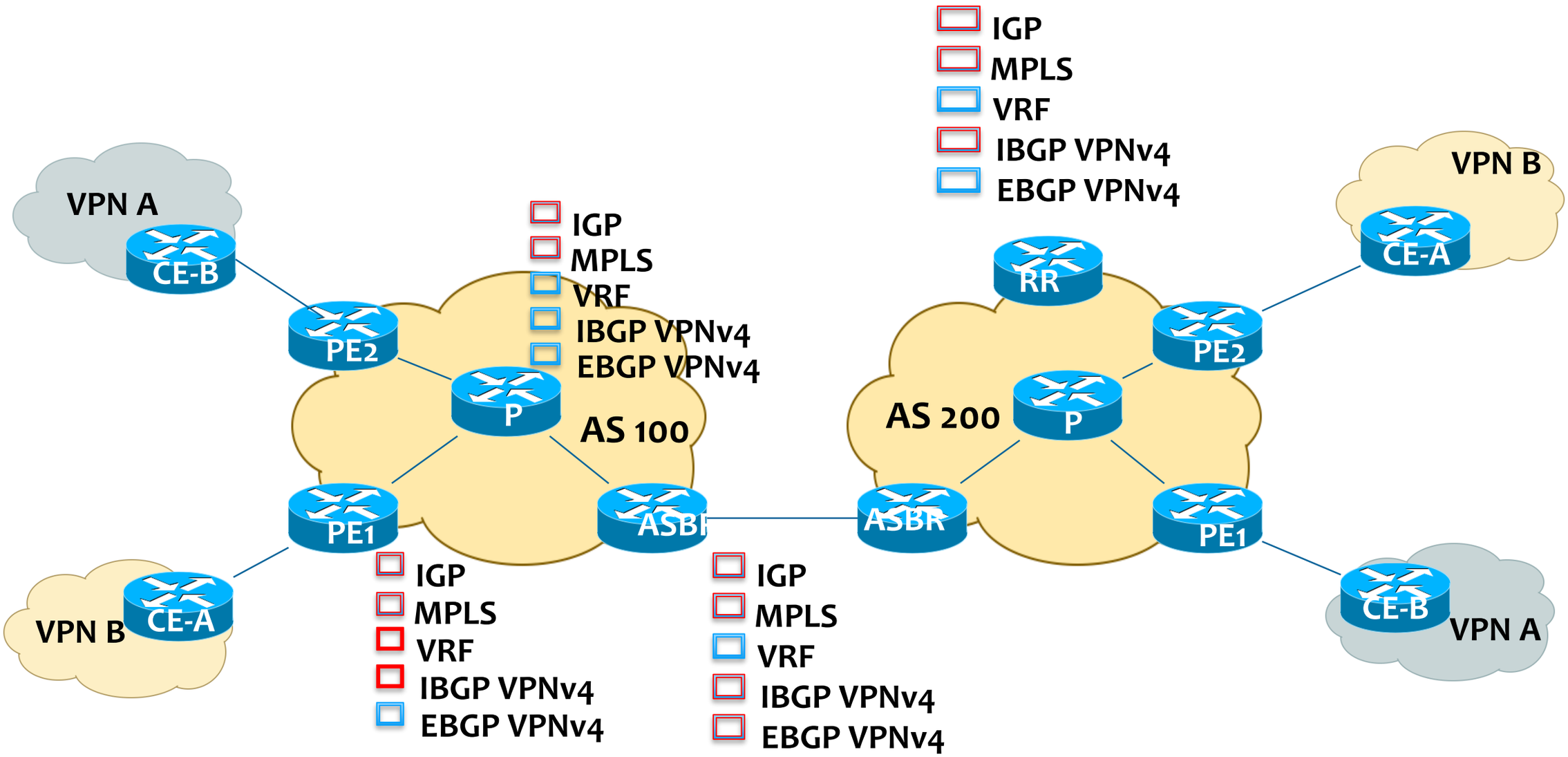
- That’s why the solution based on these requirement is Inter- AS MPLS Option B

Question 2 :

Based on the provided simplified network topologies below of the two Service Providers, please select the protocols which need to be used on the devices which have a check-box next to them.



Inter AS MPLS VPN Option B



Inter AS MPLS VPN Option B

MPLS Case Study 7

- Enterprise company has 6 datacenters. Between the datacenter they have non-IP clustering heartbeat traffic

- They don’t want to implement any vendor specific solution between the datacenters

- Their Service Provider is able to provide MPLS services.

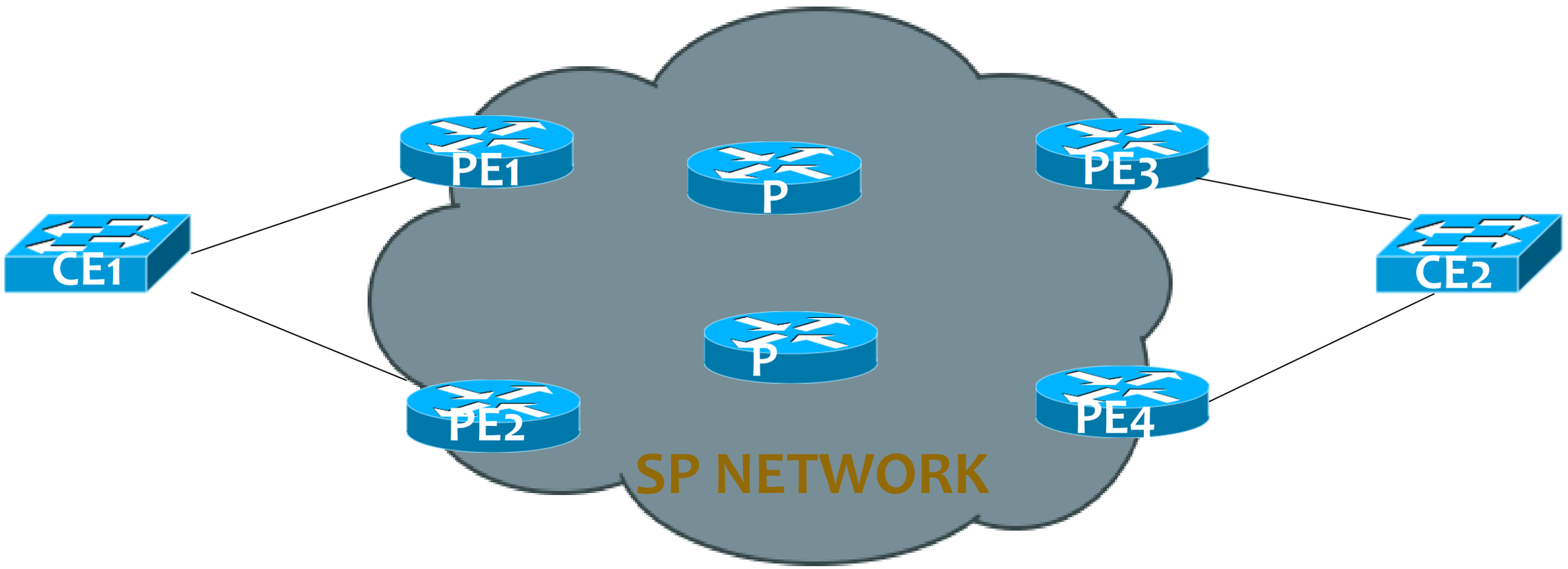
Question 1 : Which Datacenter Interconnect solution is most appropriate for this company and why ?

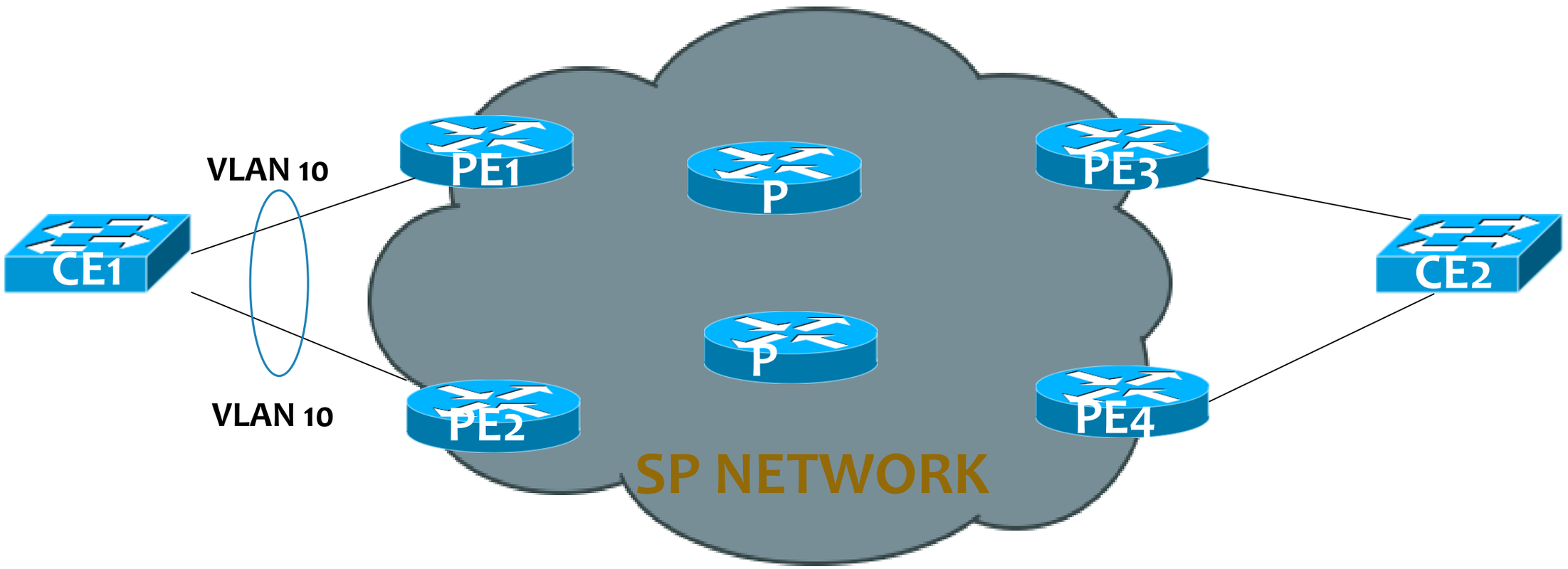
1. OTV
2. LISP
3. EoMPLS
4. TRILL
5. Fabricpath
6. VPLS

Answer 1 : Company is looking for standard based layer 2 DCI solution. We understand that they are looking for Layer 2 extension since they have an applications which requires non-IP heartbeat.

- Since OTV and Fabricpath Cisco specific solutions, they can not be chosen. Also Fabricpath is not recommended to be used as DCI solution
- LISP is not a L2 extension protocol
- EoMPLS could be used but since company has a lot of datacenters, it is not scalable.
- TRILL is not recommended to be used as DCI solution
- Best option based on the given requirement is VPLS

- Question 2 : Company sent their topology as it is shown below. They are asking if there is any solution to minimize the effect for specific Vlans in case their DC interconnect switch and the Service Provider link goes down ?





If the requirement is minimum impact for the specific Vlans, we need flow based load
balancing.

Since same vlan but different flows can be carried over the both links with flow based
load balancing, in case one link fails of the bundle, some flows are affected but no
the entire Vlan. This minimizes the affect for the specific Vlans in case of a link
failure.

But unfortunately VPLS can not provide flow based load balancing, since the MAC
address learning is done through data plane.

MAC addresses are advertised through control plane with only via E-VPN technology
over the BGP and only E-VPN can provide flow based load balancing among all the

MPLS Case Study 8

- Enterprise company is using legacy frame relay circuits and they admit that they are using very old and limited technology and wants to upgrade their network
- They decided to receive an MPLS VPN service from their regional Service Provider.
- For many years company has been managing their WAN networks themselves and they want to continue to manage themselves.
- Service Provider presented Ethernet over MPLS, VPLS and MPLS Layer 3 VPN offerings and they are very confused now.
- Can you help them to understand the MPLS VPN offerings advantage and disadvantages by filling the below chart and recommend a best service which fulfills their requirements

MPLS QUIZ

➢ **Question 1: What was the initial purpose of MPLS?**

- Avoid IP destination based lookup and increase the performance of the routers
- MPLS VPNs
- MPLS Traffic Engineering
- Alternative to the link state routing protocols
- Virtualization

➢ **Question 2: Which below options are the characteristics of MPLS layer 2 VPNs?**

- ❖ MPLS layer 2 VPN allows carrying of Layer 2 information over service provider backbone.
- ❖ Layer 2 VPN can provide point to point type of connectivity between the customer sites.
- ❖ It is used to carry layer 3 information of the customers over the Service Providers
- ❖ It is used for datacenter interconnect
- ❖ Layer 2 VPN can provide point to multi point type of connectivity between the customer sites.

➢ **Question 3: Which below options are used in the MPLS header?**

- ❖ 20 bits MPLS label space.
- ❖ Link cost
- ❖ 12bit TTL field.
- ❖ 3bit EXP bit for the QoS.
- ❖ Protocol number

➢ **Question 4: Which option below can be used as a PE-CE routing protocol in MPLS Layer 3 VPN?**

- ISIS
- BGP
- PIM
- HSRP
- OSPF
- Static Route

➢ **Question 5: What are the two possible options to create MPLS layer 2 VPN Pseudowire?**

- ❖ Martini Draft, LDP signalled Pseudowires.
- ❖ Segment Routing
- ❖ BGP EVPN
- ❖ Rosen GRE Draft
- ❖ Kompella Draft, BGP signalled Pseudowires.

➢**Question 6: Why Sham-link is used in the context of OSPF PE-CE deployment?**

❖In order to have type 1 LSA from the service provider edge router.

❖In order to have type 3 LSA from the service provider edge router

❖In order to have type 5 LSA from the service provider edge router

❖In order to have type 1 LSA in the service provider core

❖In order to have type 3 LSA in the service provider core

➢ **Question 7: Which below options are the results of having MPLS in the network?**

❖ BGP Free Core

❖ Hiding service specific information (customer prefixes etc) from the core.

❖ More scalable network

❖ Faster convergence

❖ Better security

➢ **Question 8: If customer is looking to carry layer 2 traffic with the encryption, which below options can be chosen?**

- ❖ VPLS
- ❖ EoMPLS
- ❖ GET VPN
- ❖ GRE over IPSEC
- ❖ IPSEC
- ❖ L2tpv3

➢ **Question 9: Which below options are correct for the Inter as MPLS VPN Option A?**

- ❖ It provides most flexible QoS deployment compare to other Inter-AS MPLS VPN options
- ❖ It is least secure Inter-AS option
- ❖ It is most scalable Inter-AS option
- ❖ It requires MPLS between the Autonomous Systems
- ❖ BGP+Label (RFC3107) is used between two Autonomous Systems

➢ **Question 10: Which below terms are used to define a label which provides reachability from one PE to another PE in MPLS networks?**

❖ Topmost Label.

❖ Transport Label.

❖ Outer Label.

❖ VC Label

❖ VPN Label

❖ Inner Label

➢ **Question 11: Which options provide Control Plane MAC address advertisement for MPLS layer 2 VPNs?**

❖ EVPN

❖ VPLS

❖ EoMPLS

❖ BGP L3VPN

❖ PBB EVPN

❖ VXLAN EVPN

MPLS – Study Resources

- **Books** :

- http://www.amazon.com/Definitive-Network-paperback-Networking-Technology/dp/1587142414/ref=sr_1_1?ie=UTF8&qid=1436563214&sr=8-1&keywords=definitive+mpls+network+designs

- http://www.amazon.com/MPLS-Enabled-Applications-Emerging-Developments-Technologies/dp/0470665459/ref=sr_1_1?ie=UTF8&qid=1436563734&sr=8-1&keywords=mpls+enabled+applications

- http://www.amazon.com/Network-Convergence-Applications-Generation-Architectures/dp/0123978777/ref=sr_1_1?ie=UTF8&qid=1436563938&sr=8-1&keywords=network+convergence

❖Videos :

- ❖Ciscolive Session – BRKRST – 2021 Ciscolive Session – BRKMPL – 2100

- ❖https://www.youtube.com/watch?v=DcBtot5u_Dk
  https://www.nanog.org/meetings/nanog37/presentations/mpls.mp4
  https://www.youtube.com/watch?v=p_Wmtyh4kSo
  https://www.nanog.org/meetings/nanog33/presentations/l2-vpn.mp4

❖ Articles:

- http://orhanergun.net/2015/02/carrier-supporting-carrier-csc/
  http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/WAN_and_MAN/L3VPNCon.html
  http://orhanergun.net/2015/06/advanced-carrier-supporting-carrier-design/
  http://d2zmdbbm9feqrf.cloudfront.net/2013/usa/pdf/BRKMPL-2100.pdf
  https://routingfreak.wordpress.com/tag/h-vpls/
- http://blog.ine.com/2010/08/16/scaling-mpls-networks/
  http://www.networkcomputing.com/networking/mpls-traffic-engineering-guide-success-and-alternatives/d/d-id/1269268
- https://www.ietf.org/proceedings/49/slides/ppvpn-11.pdf
  http://searchtelecom.techtarget.com/tip/Making-the-case-for-Layer-2-and-Layer-3-VPNs
  http://www.huawei.com/au/static/HW-076762.pdf
  http://www.cisco.com/c/en/us/td/docs/ios/12_0s/feature/guide/fsldpsyn.html

CCDE Practical Scenario – 1

Ornio: The Carpet Manufacturer and Seller

- Ornio is a company engaged in the production of high quality carpets, and it is one of the most well-known carpet brands in the field of carpet manufacturing. It produces Iranian, Indian, and Turkish carpets in its 80-production facilities.

- Ornio was created as a workshop, but it has expanded its brand to many countries because of its excellent quality carpets and accessories. Today, Ornio boasts a turnover of more than $80 million, 70% of which is due to increasing exports and product offerings.

- The group's headquarter is located in Turkey, one of the most famous districts in the carpet industry, and it includes an office and several production facilities.

- In addition, there are other operational offices abroad and a distribution network that includes branches in Dubai, Qatar, India, Egypt, and Iran, all of which constitute an organization of sales based on retailers and located in over 30 countries throughout the Middle East, Europe, and Unites States. In sum, it has 800 stores growing very fast.

- Ornio's datacenter is located in Istanbul which is a most crowded city of Turkey. Currently, the company has only one datacenter, and it started to consider the second datacenter for disaster recovery purpose. The entire IT team is located in Turkey. Ornio handles its IT operations in remote offices through local service contractors.

- All the applications of Ornio reside in the datacenter; there are no data or voice applications in the branch offices. Ornio uses IP telephony solution for its voice communication between the branches and the headquarter offices. Its call manager and voice gateways are located in the datacenter.

- For redundancy purpose, it uses two voice gateways for incoming and outgoing voice traffic in the datacenter. Also, It has subscriber and publisher call managers for optimal resource usage and high availability.

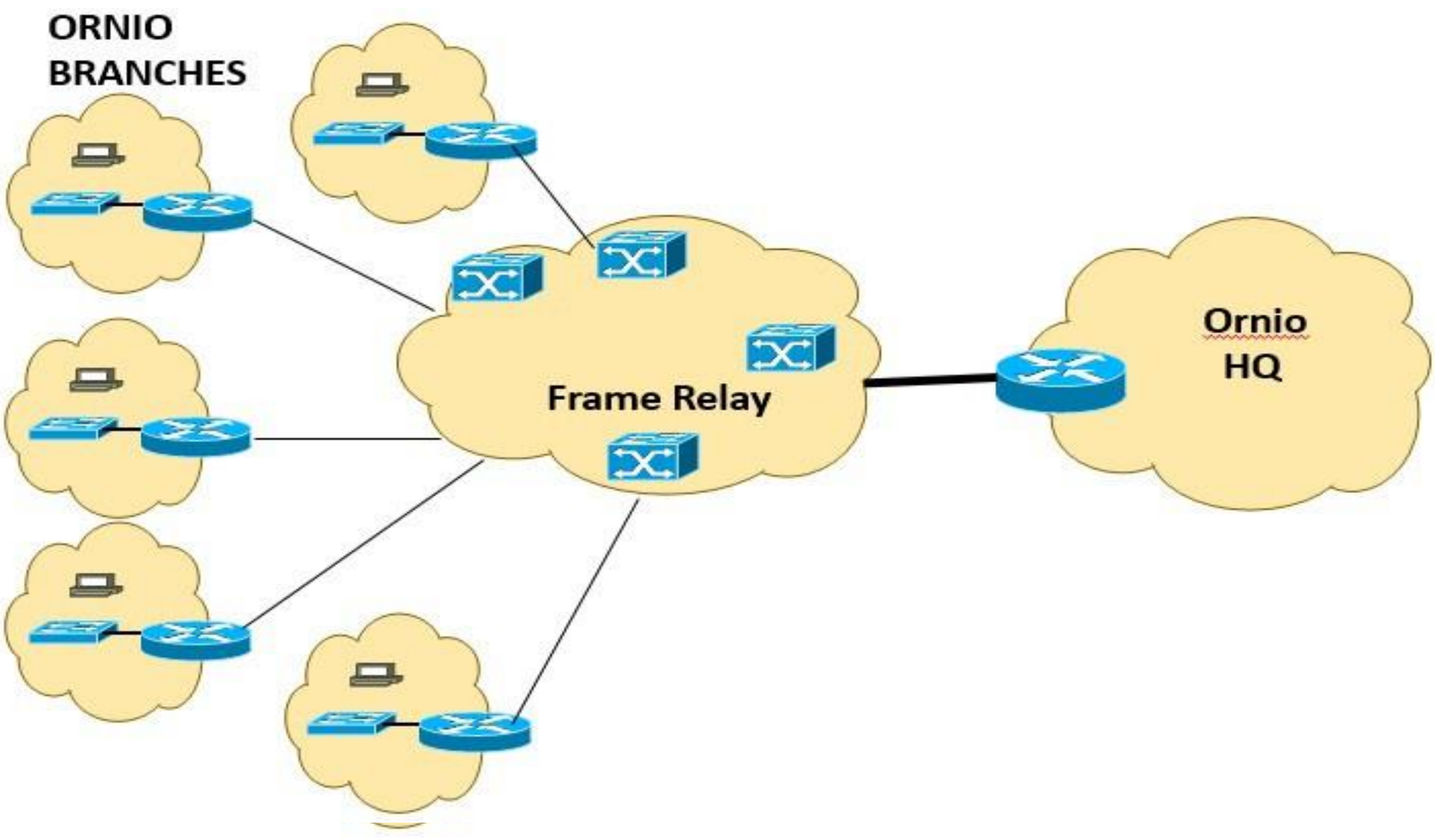
- All the branch offices of Ornio are connected to one headquarter router via frame-relay circuit. Currently, there is only one connection from all the branches. There is no Frame Relay connection between the branch offices, so all the traffic between branch offices go through it's headquarter.

- Branch offices have their own Internet connection. Ornio is considering building a backup VPN connection over the Internet. Although it is acceptable for the company not to have encryption over the Frame Relay, it definitely wants to have an encryption over the Internet if it set up VPN connections over the Internet as a backup.

- The company recently had many problems with its routing infrastructure and decided to outsource the control of its routing infrastructure to the service provider. Also, it wants a better quality of service for its voice traffic, and it wants to have direct communication between its branch offices. Although overall bandwidth required for the voice is low on the network, it is one of the highly critical applications for Ornio.

- Due to its increasing bandwidth need, Ornio has concerns with its Frame Relay WAN network. If Ornio wants to upgrade its WAN bandwidth, it should be immediate and flexible.

Ornio High Level Wide Area Network Diagram



➢ **Question 1: What is the concern of Ornio with its voice communication?**

A. It is not encrypted.

B. There is no loss, jitter, latency guarantee over Frame Relay.

C. Frame Relay is a legacy solution that doesn't offer higher rate bandwidth if it is needed.

D. Its Frame Relay network is based on hub and spoke. Voice communication is not optimal.

➢ Question 2: Should Ornio change its Frame Relay WAN?

❑ Yes

❑ No

➢ **Question 3: Which solution should Ornio choose for its Frame Relay WAN Replacement?**

1. IPSEC VPN over the Internet
2. IPSEC with Frame Relay
3. MPLS VPWS
4. VPLS
5. MPLS L3 VPN

➢ Question 4: Why?

1. It can support encryption.
2. It can provide hub and spoke VPN for Ornio, so it does not need to change its existing infrastructure.
3. It provides full mesh communication pattern.
4. Ornio can outsource its WAN network to the service provider
5. It can provide better QoS control compare to layer 2 VPNs.

➢ Question 5: Which additional information do you need from Ornio to migrate its WAN network?

1. Its QoS policy
2. Its application traffic pattern
3. Its routing information
4. Hardware capabilities
5. Does service provider have L3 VPN Offering in every site?

Email-1 is available

❖ **Email 1**

❖ **From**: Level Tokgoz (levent_tokgoz@ornio.com)

❖ **To**: Orhan Ergun (orhan@orhanergun.net)

❖ Subject: WAN Routing and service provider services

Orhan,

- *We have several problems with our Wide Area Network (WAN) recently and decided to continue with MPLS Layer 3 VPN offering. We want to have very smooth migration, and we can't tolerate to lose connectivity to our branches.*

- *Our local service provider has its own POP location at all our branch office locations as well as at the Headquarter.*

- *We are currently using EIGRP as a routing protocol over the Frame Relay connections. We have only one router at our branches, which does layer 3 routing between local area network Vlans.*

- *We want to have some policy control/traffic engineering capability on our Wide Area Network whenever we want without asking for any permission from the service provider.*

- *We know that with MPLS Layer 3 VPN, the service provider will control our most of our routing, but we still want to influence it as much as we can.*

- *This is a very big project for us, and we want to have full visibility.*

Regards

Levent Tokgoz
Network and Security Director
Ornio Group
Maslak/Instabul

➢ Question 6: Should Ornio change its routing protocol?

- ❑ Yes
- ❑ No

➢ **Question 7: Which routing protocol should Ornio choose?**

1. OSPF
2. IS-IS
3. RIPv2
4. Static Routing
5. BGP

➢ Question 8: Why?

1. Not all service providers support EIGRP for MPLS Layer 3 VPN PE-CE protocol.
2. It is Cisco owned protocol, so Ornio doesn’t want to be restricted to a vendor.
3. If it wants to do traffic engineering over its WAN, Ornio has this ability without service provider communication.
4. It can carry more prefixes over BGP than any other routing protocol.
5. It can send non-IP traffic over BGP.

➢**Question 9: Should Ornio use the same autonomous system number on every location or unique autonomous system number for every location?**

A. Same AS
B. Unique AS

➢Question 10: Why?

1. Using unique AS per site limits the number of customer sites to number of available BGP AS.
2. Using unique AS may require allocation of AS numbers outside of private AS range.
3. Not all service providers use a unique AS per location.
4. It requires very careful attention to avoid AS collision.
5. Using the same AS is the common best practice.

Email-2 is available

❖Email 2

❖**From**: Leven Tokgoz (leven_tokgoz@ornio.com)

❖**To**: Orhan Ergun (orhan@orhanergun.net)

❖**Subject**: New WAN Routing Protocol

Hi, Orhan

- *As per your recommendation, we are pleased to continue with the BGP as a new WAN routing protocol.*
- *But we want to know what can be the problem with BGP. If there is something from the service provider site can be done, please do not hesitate to inform us. We can accept to have some level of configuration in order for the solution to work. However, if the configuration is required on each site, we prefer the service provider to do it.*

Warmest regards

Levent Tokgoz
Network and Security Director
Ornio Group
Maslak/Instabul
26674

➢ **Question 11: What is the problem of using the same BGP AS everywhere?**

1. It requires redistribution from VRF to MP-BGP on the PEs.

2. BGP Loop prevention mechanism rejects prefixes if the same AS number is seen in the AS path.

3. BGP Fast reroute mechanism cannot be implemented with the same AS number everywhere design.

4. Service provider should remove private AS numbers if it will be announce to internet.

5. No site unique characteristics can be identified from the AS path.

➢**Question 12: How can the problem with the using same AS number on every location be solved?**

1. Use unique AS per location.

2. Service provider can override the AS Path, so customer sites don’t receive their AS number in the path.

3. Using MPLS in the service provider network.

4. By changing the BGP loop prevention, customer can allow the same AS path to be received.

5. It is not a problem with single homed site, unless Ornio multi-homed its site.

➢**Question 13: Is there any problem on the Service Provider network, if Service Provider configures AS-Override feature ?**

❑Yes

❑No

➢ **Question 14: What is the problem with BGP AS override configuration at the service provider network?**

1. Service provider BGP configuration becomes more complex.
2. BGP routing loop at the multi-homed customer site.
3. Customer BGP configuration becomes more complex.
4. Some customer sites still reject the BGP prefixes due to its own AS number in the BGP update
5. It is not supported on the BGP route reflector

- Question 15: **Is there any problem on the Ornio network, if Service Provider configures AS-Override feature ?**

❑ Yes

❑ No

➢ **Question 16: Please provide the necessary steps for the Ornio's MPLS Layer 3 VPN migration.**

1 . Remove the old Frame Relay circuit from the transit site.

2. Establish a BGP over the MPLS circuit.

3. Arrange the routing protocol metric to choose MPLS over Frame Relay.

4. Establish a new circuit at the transit site.

5. Choose a transit site for the communication between migrated and non-migrated site.

6. Establish a new circuit at the remote site.

7. Remove the old Frame Relay circuit from the remote site

8. Enable Quality of Service and Monitoring for the new MPLS connection

Email-3 is available

❖Email 3

❖**From:** Levent Tokgoz (levent tokgoz@ornio.com)

❖**To**: Orhan Ergun (orhan@orhanergun.net)

❖**Subject:** MPLS Layer 3 VPN Migration and Multicast Service

❖ Hi Orhan,

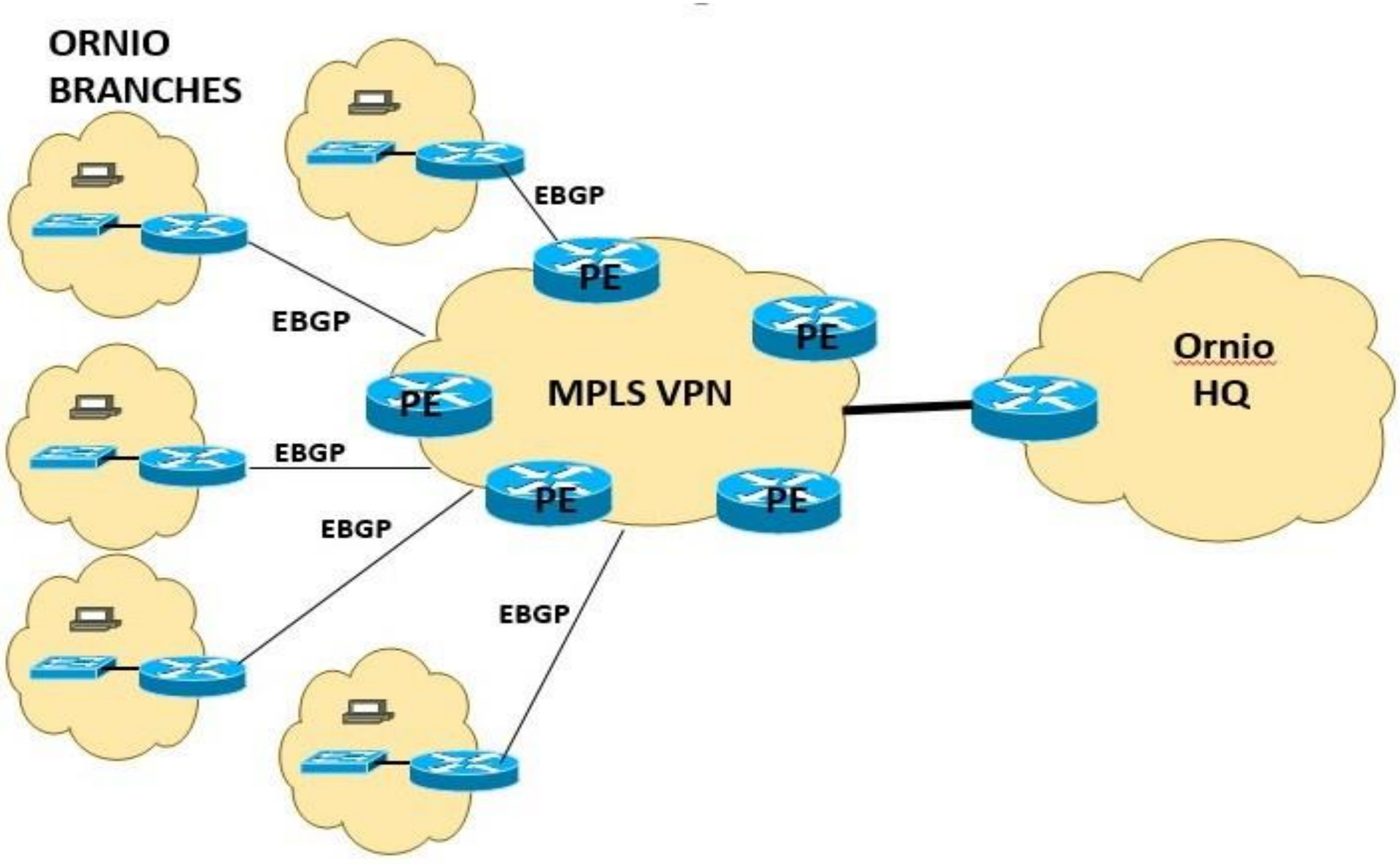
❖ *It was a great achievement. Thanks for helping us for the MPLS migration.*

❖ *We have the below topology.*

❖ *Although we don’t have the superior expertise on BGP, we followed your advice and chose BPG as a WAN protocol. Our service provider could offer EIGRP as WAN protocol, a protocol that we are very comfortable with. As you had indicated, BGP will provide us the traffic engineering capability.*

- *However, we want to implement a new multicast application over our Wide Area Network.*

- *We will have 2 Media server located at the Datacenter and all the branch offices should receive the information at the same time. We don’t need to have any multicast stream from the branch offices towards Head Quarter.*



- *As per our discussion with the Service Provider, they are providing a Multicast service over the MPLS layer 3 VPN.*

- *We heard that Multicast PIM-BIDIR minimizes the resource requirement on the routers but we want to get your advice.*

- *Best regards*

- *Levent Tokgoz*

- *Network and Security Director*

- *Ornio Group*

- *Maslak/Instabul*

- *26674*

	PIM SSM	PIM ASM	PIM Bidir
Minimum Amount of State in the Router			
Most optimal Routing			
Requires Randevous Point			
Works with IGMPv2			
Randevous Point Load Balancing			

➢ **Question 18: Should Ornio use PIM-bidir between headquarter and the remote offices over the MPLS Layer 3 VPN?**

❑ Yes

❑ No

➢ **Question 19: Why should Ornio not use PIM-bidir for its multicast traffic?**

1. Its application is point to multipoint, so PIM SSM or ASM is more suitable.

2. Service provider cannot have Data MDT for their Rosen GRE multicast implementation.

3. Not every router supports PIM-bidir in Ornio's network.

4. Pim bidir requires much more configuration compared to PIM ASM and PIM SSM.

Email-4 is available

- **Email 4**

- **From:** Levent Tokgoz (levent_tokgoz@ornio.com)

- **To:** Orhan Ergun (orhan@orhanergun.net)

- **Subject:** Redundant remote sites

Hi, Orhan,

We will not choose PIM-bidir even though we have verified that our devices support it.
Thanks for the information.
We started to plan next year budget and we want to include the device and link redundancy for the critical branch offices.
Please recommend the solution and let me know if you need additional information.

Thanks

Leven Tokgoz

Network and Security Director

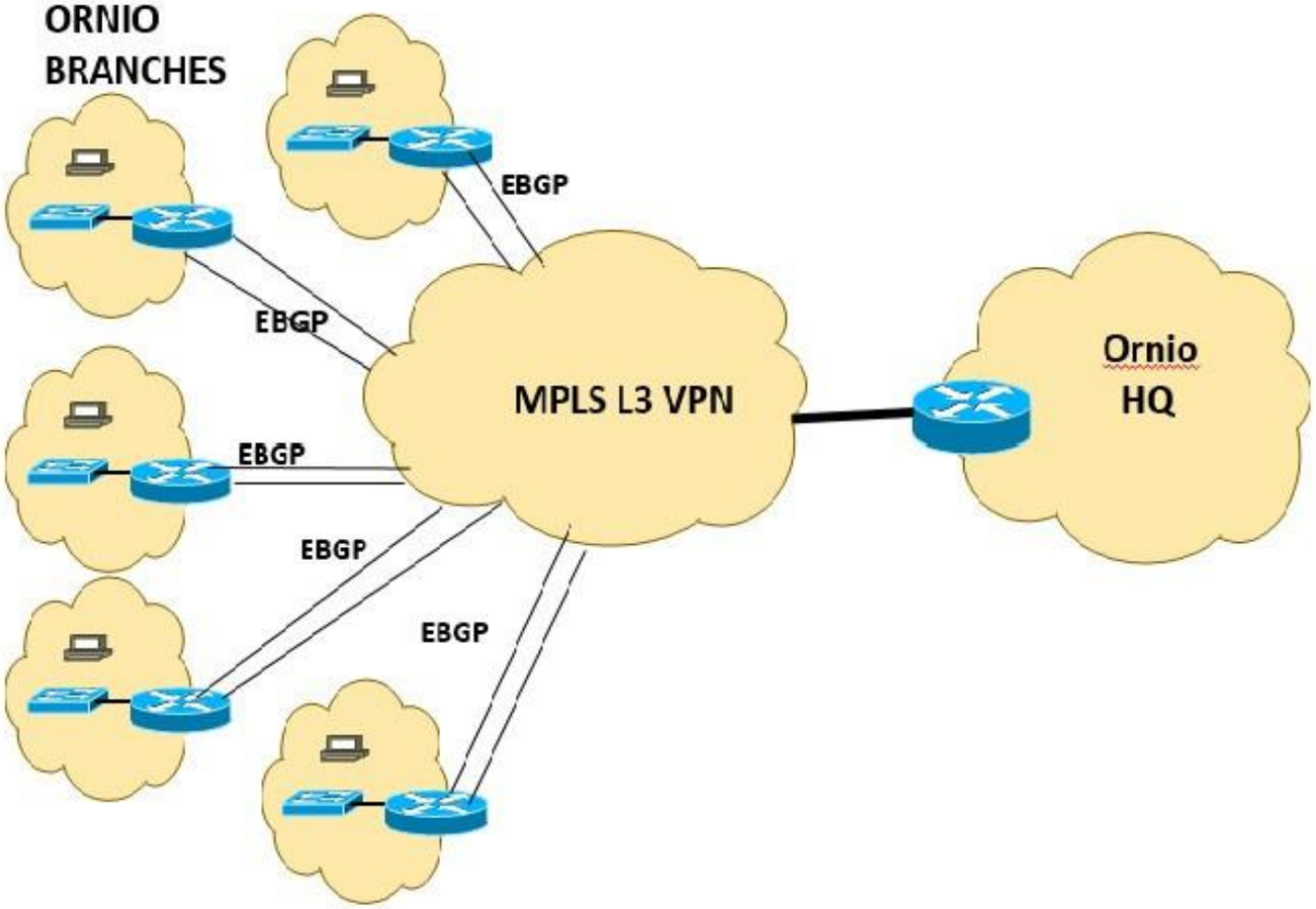
Ornio Group
Maslak/Istanbul
26674

➢ **Question 20: Do you need additional information for Ornio's future redundancy plan for the branch offices?**

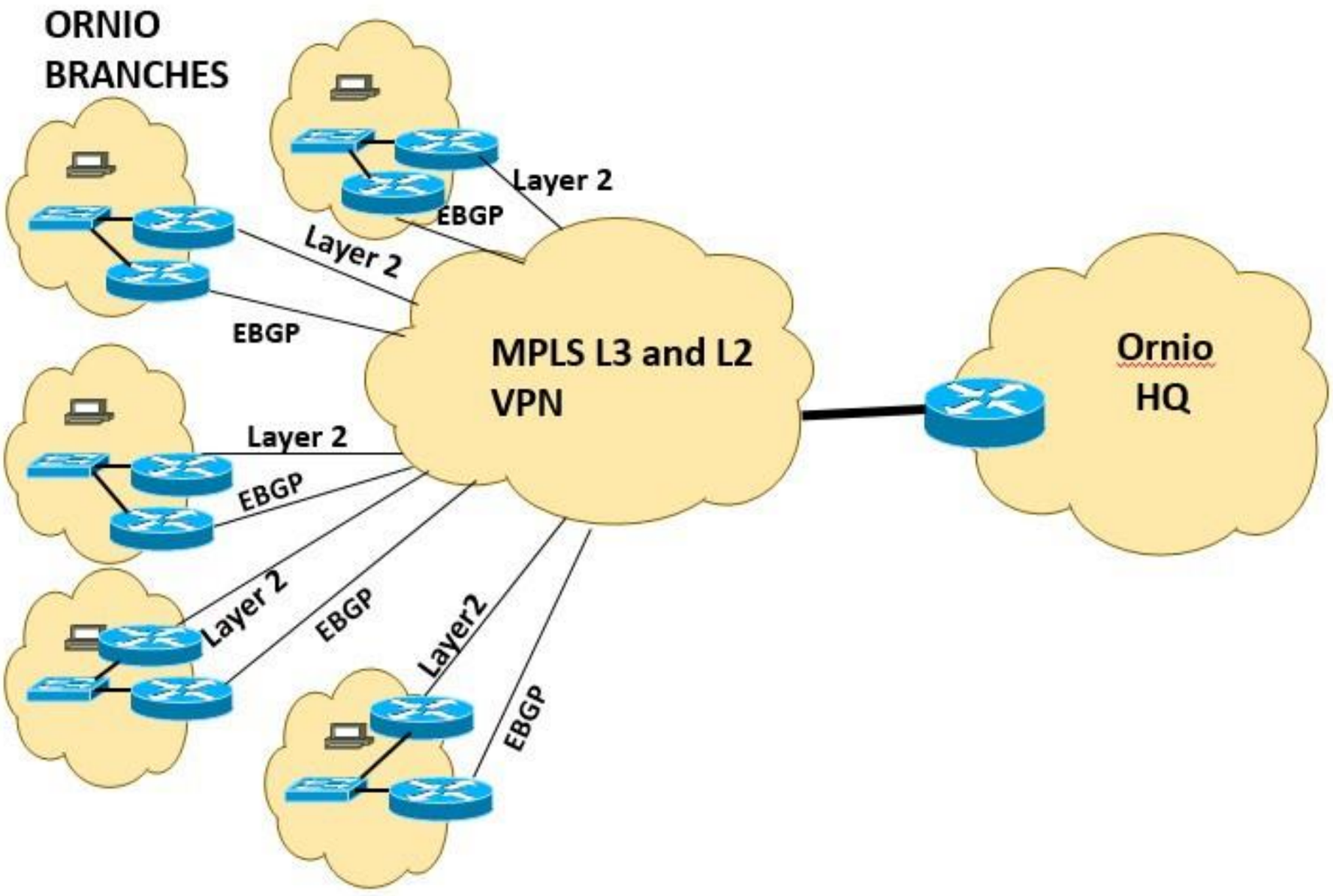
❑ Yes

❑ No

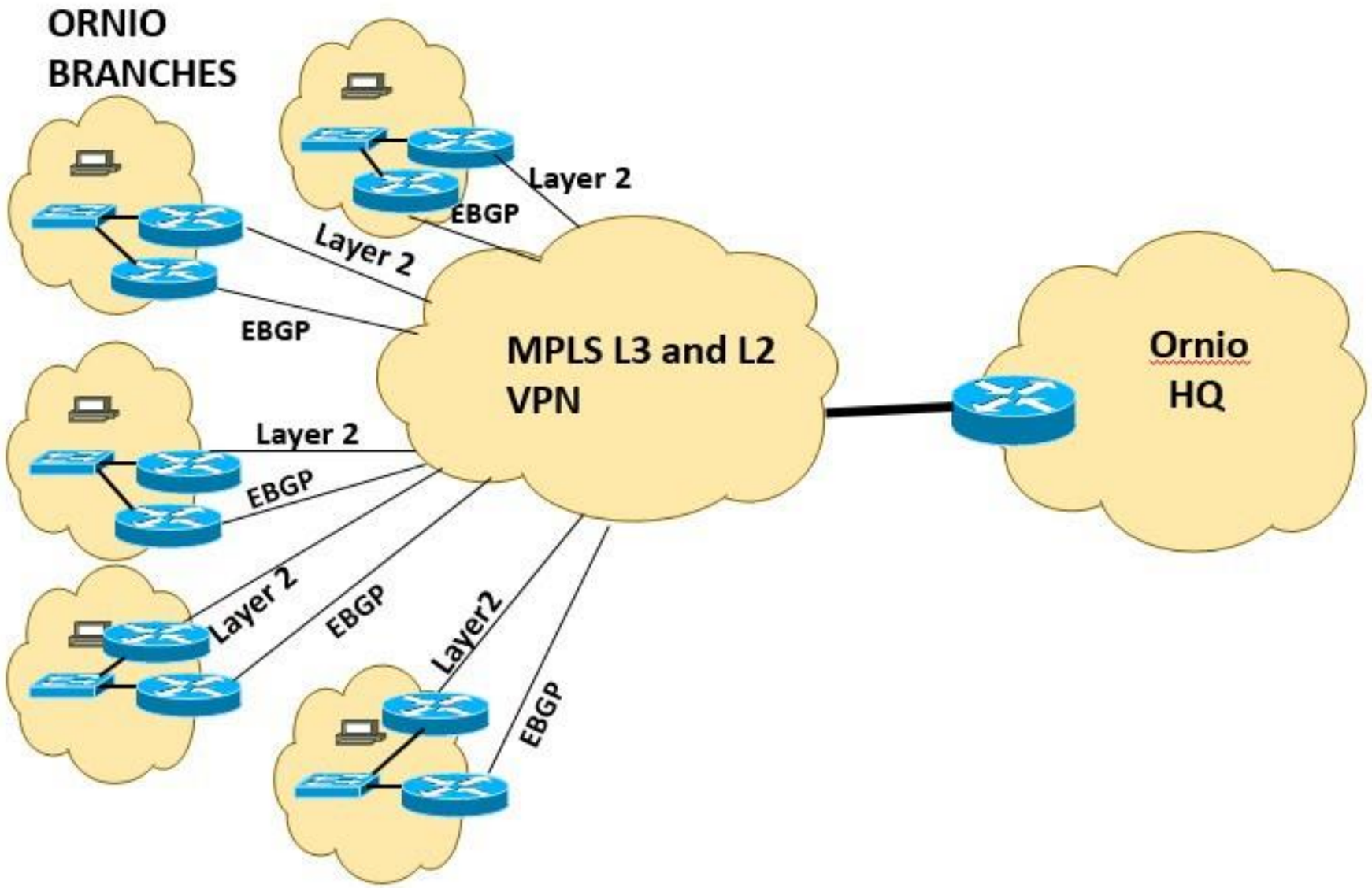
ou



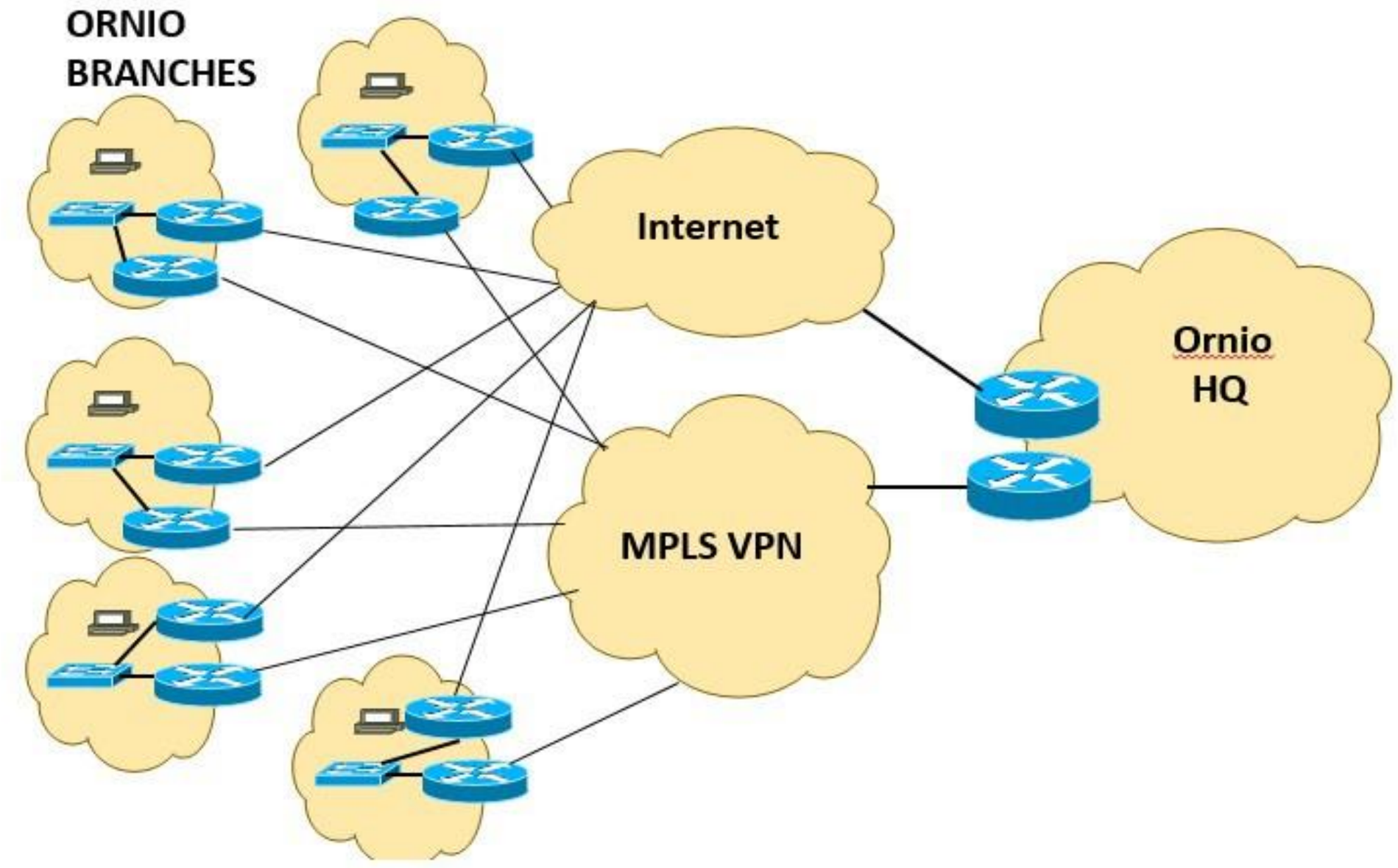
Topology A



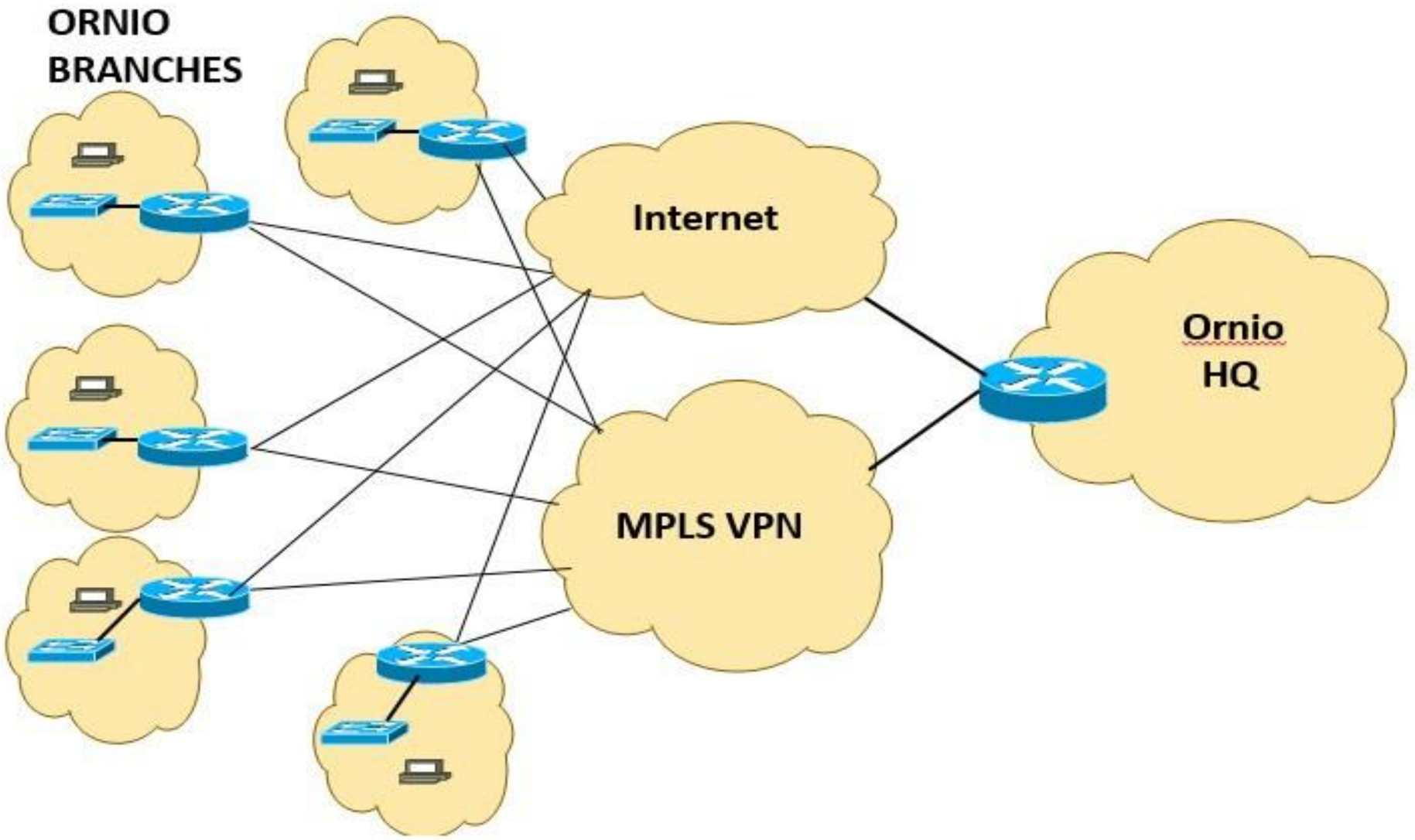
Topology B



Topology C



Topology D



Answers of the Scenario -1

➢ **Answer 1**

1. Although this is true, by default Frame Relay is not encrypted. In addition, the question did not mention that it has a problem with encryption for voice traffic. That is why this option is not the correct answer.

2. We don’t know this. Service Provider might be offering those SLAs. The voice traffic is the main concern but not mentioned in the scenario. That is why this option is not the correct answer.

3. This is definitely true. However, it is indicated clearly in the scenario that its voice traffic on the network is low. This can be a concern for bandwidth hungry applications, but the question is asking Ornio's concern with the voice

4. This is the correct answer. It is clearly mentioned in the scenario that Ornio wants to have direct communication, even if it has direct communication with Frame Relay (Full-mesh)

Note: Although control plane traffic for voice would go to the datacenter, they use IP Telephony. Also, Ornio’s call managers are located in the central site/datacenter; hence, data plane traffic (RTP) would follow the direct path.

➢ Answer 2

- ❖ Yes.
- ❖ If you would say No, you would lose some points here, but the scenario would allow you to continue with the other questions

Answer 3

Answer is MPLS Layer 3 VPN

The detailed explanation will be provided in Answer 4.

➢ Answer 4

1. By default, MPLS Layer 3 VPN doesn’t provide encryption. Although it can be used in conjunction with IPSEC to provide encryption, Ornio doesn’t look for encryption if it is not over the Internet. This is not the correct answer

2. Yes. MPLS Layer 3 VPN can support Hub and Spoke topologies, although by default it provides full mesh connectivity. But Ornio is looking for direct connectivity between spoke sites (Branch Offices), so either VPLS or MPLS Layer 3 VPN provides this by default. It is very difficult to implement full mesh with the other options such as VPWS due to the number of offices and the number of required connectivity. It is not scalable. This is not the correct answer.

3. Yes. This is true and what we are looking for. This is one of the correct answers. All the business requirements are met with this option.

4. Yes. With MPLS layer 3 VPN, customer can outsource its WAN network to the service provider. It is peer-to-peer VPN and the service provider controls the routing. Customer equipment has routing connectivity, including static routing, with the service provider. Ornio is exactly looking for this. Please note that the CE is not managed by the service provider, so even if the customer is managing it, the service provider controls the WAN routing. This is one of the correct answers.

5. Yes. This is true but not the correct answer.Layer 3 VPN can provide better, flexible QoS control compared to the Layer 2 VPNs. With Layer 2 VPNs, ingress PE can have its operation based on Layer 2 markings, which is only 3 bits or 8 bits. But if it is Layer 3, then PE can do its operation over 6 bits DSCP values, which can provide 64 different values. Nothing is told for the Quality of service in the scenario so far.

❖ We have two correct answers, 3 and 4

➢ Answer 5

1. At this point, we are just going to migrate, so Ornio's QoS policy will be considered later on. In order to migrate, we do not need QoS information from neither customer nor service provider site. This information will be required to map QoS classes between them.

2. Its existing and desired application traffic pattern is already given in the scenario. It has Hub and Spoke, but it wants to have full mesh connectivity.

3. Yes, definitely we would like to know this.

4. This is more of a CCDA concern, not for the CCDE exam.

5. Yes. This is important. We need to know the capability of the service provider. Can every branch office get this service? Do we need some sort of Inter-AS agreement after which QoS, multicast might be a concern.

➢ Answer 6:

❖ Yes

➢ Answer 7:

❖ **BGP**

➢ **Answer 8:**

❖ 1.Yes. This is true even though it is not the correct answer. Service providers generally support BGP and static routing as PE-CE routing protocol with their MPLS Layer 3 VPN offering. However, this is not the main concern of Ornio, as it can be understood from its first email.

❖ 2.This is true as well but not the correct answer. Although there is an attempt to make EIGRP an open protocol – if some critical features such as EIGRP Stub are not open – it cannot be truly open protocol. But since the Ornio's business requirement is not related with this, this is not the correct answer.

- 3.Yes. As indicated in Email-1 as well, Ornio wants to have some degree of traffic engineering capability. Only BGP can influence the MPLS VPN traffic. This is the correct answer.

- 4.Yes, BGP can carry millions of prefixes, but Ornio doesn’t have such concerns.

- 5.Yes. BGP can carry non-IP traffic but MPLS VPN only carries IP traffic. This is not the correct answer.

- Only correct answer is Option 3

➢ Answer 9:

- ❖ Same AS (Autonomous System)
- ❖ The reason will be explained in detail in Answer 10.

➢ Answer 10:

❖ 1.This is one of the correct answers. Ornio has 800 sites and private BGP AS range (64512 - 65535) is limited to 1000 around. Since we know that Ornio’s operation is growing, in future it will have AS allocation problem. This is why the company should choose the same AS number everywhere.

2. If Ornio chooses to deploy unique AS number after 1000 sites, it would need to use public AS numbers. Although this would be a big concern over the global internet, in the context of VPN, using it does not cause harm. This is another correct answer.

3. There is no such restriction. This is not the correct answer.

4. Yes. This is one of the correct answers. If Ornio chooses to deploy unique AS number per location, then it has to be careful not to use the same AS number on more than one location.

5. Such practice is not common. This is not the correct answer.

1, 2 and 4 are the correct answers

➢ Answer 11:

1. This is not true. If BGP is used as PE-CE protocol, there is no need to redistribute it on the PE. Also, this is not a concern of using the same AS number everywhere. If this would be true, it would be the concern of unique AS number per site as well.

2. This is one of the correct answers. BGP works this way.

3. There is no such restriction. If BGP PIC for the fast reroute purpose will be enabled, it can be enabled both for the same and unique BGP AS allocation methods.

4. This is true even though it is not the correct answer. This is because it is not the problem of using same BGP AS everywhere.

5. This is one of the correct answers. If the same AS is used on every site, then you cannot use AS path attribute for the traffic engineering and path preference.

➢ **Answer 12:**

1. This is not the correct answer, since the decision is made. Ornio will use the same AS number everywhere for the above reasons.
2. This is the only correct answer. Service provider can override the customer AS number with its own AS number. At the receiving site, customer will not see similar AS in the accepted path
3. Using MPLS in the service provider is not relevant. It doesn’t create or solve any problem for this issue.

4. This is true, but not the correct answer. This is because in the Email-2, customer stated that they don't want to configure a BGP Policy if the same AS creates an issue on every site. Hence, this is not the correct solution. It is against customer's requirements.

5. This is not true, it creates same problem for the single homed site as well.

➢ Answer 13:

❖ Yes

❖ Detailed explanation will be provided in Answer 15.

➢ Answer 14:

1. Service provider will add only one additional line command. It can still be considered more complex, but since the question does not state, choose, or apply, this is not the right answer.
2. Yes, multi-homed customer sites have a BGP loop once they are attached to the two PEs. This is the biggest problem and that is why this is the correct answer.

3. With the BGP AS Override, nothing is configured at the customer site; thus, BGP AS Override does not create additional complexity at the customer site.

4. No. If all the PEs having a VPN for the customer sites allow BGP AS Override, it is not rejected.

5. BGP AS Override configuration is done on the service provider Edge Routers, not on the BGP route reflector.

➢ Answer 15:

- ❖ No. Sometimes in the exam, they may not ask you to state the reasons. It is not asked here as well. However, for the curios people, since all the branch offices as well as headquarter site of Ornio has only one router in the MPLS VPN domain, there is no much concern. There is no dual-homed site.

➢ Answer 16:

5, 4, 6, 2, 3, 7, 8, 1

➢ Answer 17:

	PIM SSM	PIM ASM	PIM Bidir
Minimum Amount of State in the Router			x
Most optimal Routing	x		
Requires Randevous Point		x	x
Works with IGMPv2		x	x
Randevous Point Load Balancing		x	

There is no rendezvous point load balancing in PIM-bidir. Phantom RP works as an active/standby manner.

➢ Answer 18: No

➢ Answer 19:

1. This is the only correct answer. Ornio has a multicast application. The sender is at the company’s headquarter, and all the receivers are at the branches. PIM SSM or PIM ASM is more suitable since shortest path tree will be created. PIM SSM wouldn’t require RP at all, so it would be best for this design.

2. This statement is definitely correct. Data MDT in Rosen GRE Multicast requires (S, G) multicast state. If customer deploys PIM-bidir for multicast, since there is no (S, G) state in PIM-bidir, data MDT cannot be created. Although this statement is correct, since the data MDT is a service provider concern, not the Ornio's, answer is not what we are looking for.

3. We don’t know this. Even though we knew this information, this is not a design concern but more of an operational concern.

4. This is also an operational issue, although complexity is a design attribute. Generally, we need absolute measurement in CCDE exam. We cannot assume.

➢ Answer 20:

- **No**
- Detailed answer will be provided in Answer 22

➢ Answer 21:

- ❖ Answer is in option 4.
- ❖ In the scenario section, you are told that Ornio has local internet connection at remote sites already, and, in the future, they may want to use it.

- There is small difference between Answer 4 and 5. In Answer 5, there is only one router at the branches. But in Email-4, you are told that Ornio is looking for device and link level redundancy. Thus, Answer 4 is fit for all the business requirements