CISCO CERTIFIED DESIGN EXPERT

PRACTICAL CISCO CERTIFIED DESIGN EXPERT V2.0



www.OrhanErgun.net

PRACTICAL

CISCO CERTIFIED DESIGN

EXPERT V2.0

www.OrhanErgun.net

Orhan Ergun

CCDE #2014:17

CCIE #26567



Study Guide for 2015 Cisco CCDE Practical Exam V2.0

Author: Orhan Ergun
Editor: Temitope Yinka Sodiq
Copyright© 2015 Orhan Ergun

No part of this book should be reproduced or transmitted in any form or by an means, electronic or mechanical, including photocopying, recording, or by an information storage and retrieval system without the written consent from Orhan Ergun, except for the inclusion of brief quotations in a book review.

Printed in the United State of America

First printed in November 2015

Warning and Disclaimer

This book is designed to provide information about Cisco CCDE Practical Exam. Every effort has been made to make this book as complete and accurate as possible, but no warranty or fitness is implied.

The authors and editors shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book.

The opinions express in this book belong to the author and the author alone.

Special thanks to Temitope Yinka Sodiq for editing and proof reading this book.

Contents

CONTENTS I
WHAT IS CCDE PRACTICAL? 1
CCDEV2 WORKBOOK RELEASE NOTES 1
COURSE OUTLINE 1
ABOUT THE COURSE 1
OVERVIEW OF LAYER 2 TECHNOLOGIES K 1
LAYER 2 TECHNOLOGIES SPANNING TREE 1
SPANNING TREE THEORY 1
SPANNING TREE TOOLKIT 1
SPANNING TREE BEST PRACTICES 1
FIRST HOP REDUNDANCY PROTOCOLS 1
FHRP CASE STUDY 1
SPANNING TREE – HSRP INTERACTION 1
LAYER 2: ACCESS DESIGN 1
LAYER 3: ACCESS DESIGN 1
GENERAL ROUTING KNOWLEDGE AND BEST PRACTICES 1
GENERAL ROUTING BEST PRACTICES – STUDY RESOURCES 1
OSPF 1
OSPF CASE STUDY 1
OSPF CASE STUDY – 2 1
OSPF – STUDY RESOURCES 1
ISKIS 1
IS-IS CASE STUDY 1
HIGH LEVEL MIGRATION PLAN FROM OSPF TO IS-IS FOR FASTNENT 1
ISKIS – STUDY RESOURCES 1
EIGRP 1
EIGRP CASE 1
EIRP – STUDY RESOURCES 1
BFP 1
EBGP 1
BGP PERRING 1
BENEFITS OF PERRING 1

www.orhanergun.net

BGP CASE STUDY – 1 1
BGP CASE STUDY – 2 1
BGP CASE STUDY – 3 1
BGP – STUDY RESOURCES 1
MULTICASE 1
MULTICASE CASE STUDY 1
MULTICASE – STUDY RESOURCES 1
QUALITY OF SERVICE (QOS) 1
QOS K STUDY RESOURCES 1
MPLS 1
LAYER 2 MPLS VPN 1
EVPN 1
LAYER 3 MPLS VPN 1
INTER AS MPLS VPN 1
MPLS TRAFFIC ENGINEERING 1
MPLS TRAFFIC ENGINEERING FAST REROUTE 1
MPLS CASE STUDY – 1 1
MPLS CASE STUDY – 2 1
MPLS CASE STUDY – 3 1
MPLS CASE STUDY – 4 1
MPLS – STUDY RESOURCES 1

What is CCDE Practical?

- Cisco Certified Design Expert Lab
- It is a network design exam
- Prerequisite : Student must pass CCDE Written exam
- Four scenario's over 8 Hours
- Two scenario every 4 hours
- Business Challenges
- There is no configuration for any vendor equipment
- Vendor agnostic, but still some technologies relevant to Cisco (erg. HSRP, GLBP, EIGRP, DMVPN).
- Exam score will be made available immediately after exam.
- Reading intensive. 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
- Exam score is provided based on these job tasks
- Exam is given every 3 months in the Pearson Professional Center's
- Passing score around 75 – 80 %

"There are lots of peer technical CCIE tracks, but CCDE was the only one that matched my strategic design role. It is unique, vendor-neutral course, based on universal design principles and technologies that can be applied to solving comprehensive business needs."

Tony Brown, Enterprise Systems Architect - Verizon

CCDEv2 Workbook Release Notes

Please check back periodically for release notes on workbook updates.

Changes by Date

July 10, 2015

- Multicast, MPLS, additional resources finished.

July 7, 2015

- OSPF, EIGRP, BGP finished.

July 2, 2015

- Initial workbook release

August 4, 2015

- Scenario added

August 6, 2015

- Second edition workbook release

Course Outline

Layer 2 Technologies % 10

- Spanning Tree
  CST, PVST+, RSTP, RPVST+, MST
- First Hop Redundancy Protocols
  HRSP, VRRP, GLBP
- Layer 2 and 3 access design

Layer 3 Routing %40

- OSPF
- IS-IS
- EIGRP
- BGP

MPLS and Applications %30

- Layer 2 MPLS VPN
- Layer 3 MPLS VPN
- Inter-AS VPNs
- Option A, B , C
- MPLS Traffic Engineering
- Carrier Supporting Carrier

Network Overlays - %10

- GRE
- DMVPN
- GETVPN
- L2TPv3

QOS - % 5

MULTICAST -% 5

About the Course

- Every Monday, Wednesday and Saturday from 8PM to 10:30PM with GMT+3 for one month.
- Sessions will be recorded; you can download them to watch later.
- Before each session, you will receive a meeting invite.
- Ask questions during the sessions and ask as much as you like!
- If the question is irrelevant with the topics or it will take much time to explain and send an email, I will reply each one of them offline.

Overview of Layer 2 Technologies K

- We will talk specifically about Ethernet and its legacy control plane, spanning t r e e .

- Ethernet has newer control and data plane with TRILL, SPB, PBB-TE and so on but CCDE doesn't cover these technologies

- Or some resiliency technologies such as G8032, REP in layer 2 are not covered in CCDE exam.

- HSRP, VRRP and GLBP will be covered in this part of the course.

- Also pros and cons of layer 2 and layer 3 access designs (Rooted Access) design will be explained.

- Don't forget, CCDE is a layer 3-infrastructure exam and layer 2 part is very minimal.

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

- Loop has to be mitigated but blocking links don't allow using all the links actively. Thus with spanning tree, flow based load balancing is not possible.

- Some implementations of spanning tree which we will see below allows only Vlan based load balancing. Some of them allow only active-standby redundancy.

- In flow-based load balancing, hosts in 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.

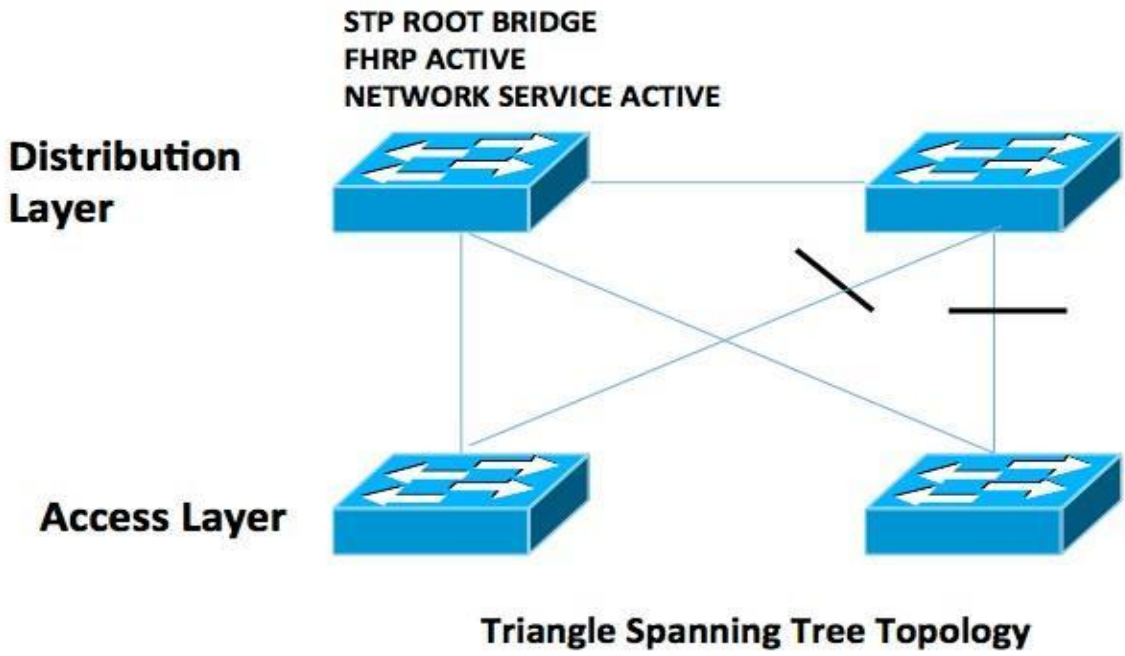


Fig.2: Access- Distribution Switches and Spanning Tree

- As soon as spanning tree detects a loop, it blocks a link to prevent the loop.

- CST 802.1d which is classical/legacy spanning tree; supports only one instance for all vlans. It doesn't support vlan load balancing.

- PVSTP+ Cisco's 802.1d implementation but supports one to one instance to vlan mapping.

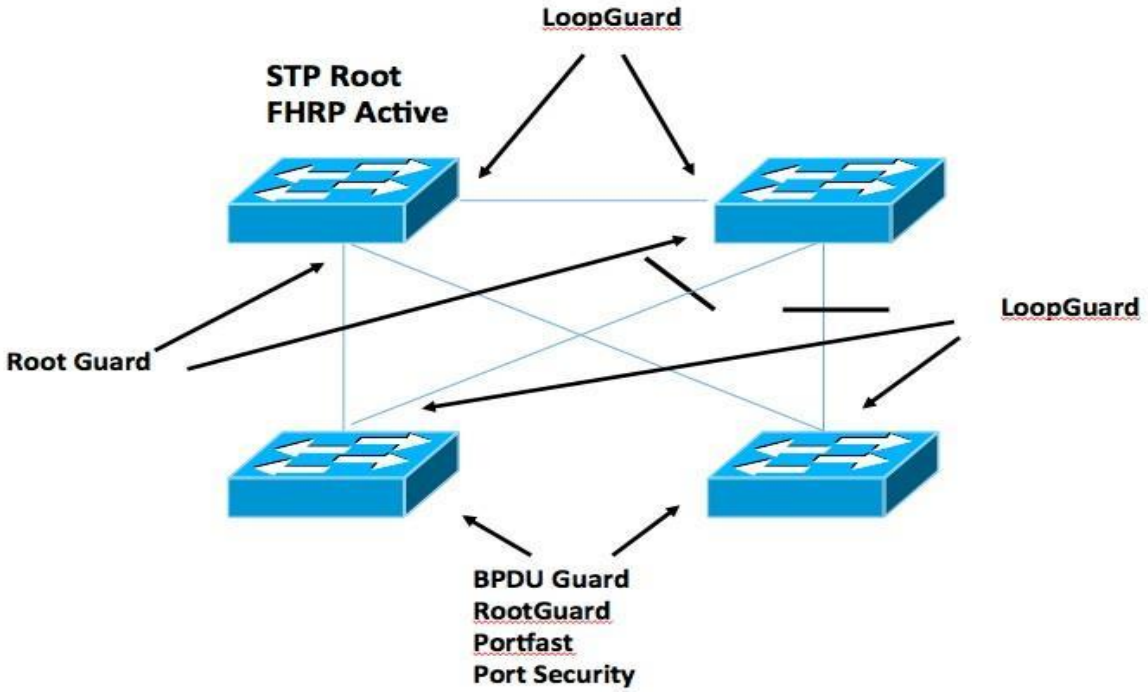
- Enhancements to PVSTP provide good optimatizations, but it has slow convergence compared to MST and RSTP and cannot scale as MST.

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 after link failure.
- BackboneFast—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.
- 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



- 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.

- 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 Best Practices

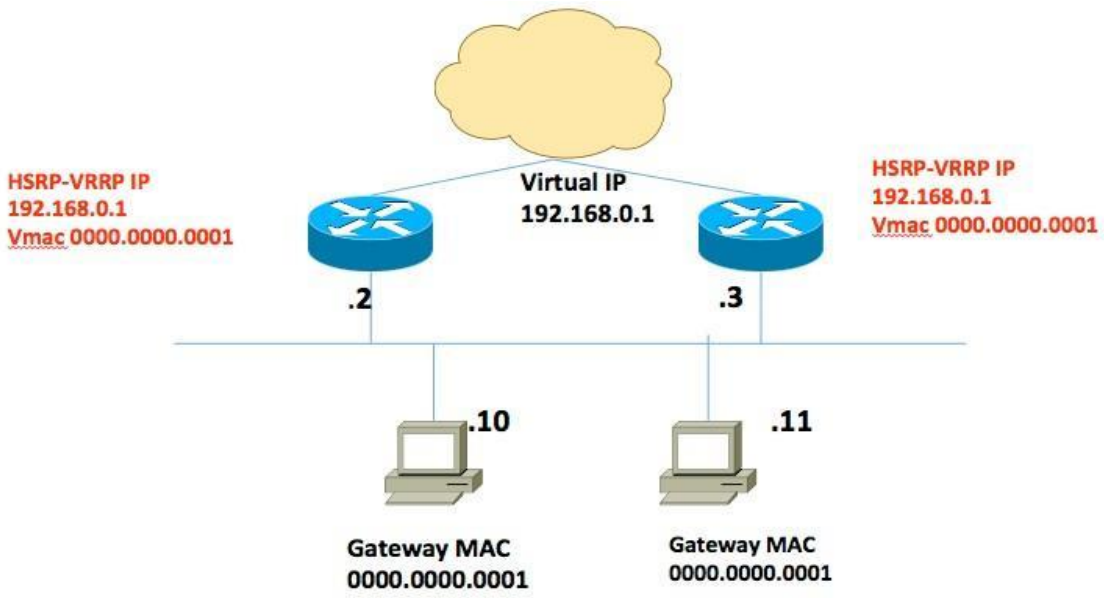
- Use RSTP or RPVST+ for fast convergence for direct and indirect failures.
- Use MST for scaling. If you have large scale vlan deployment and CPU is a concern, you can take advantage of grouping vlans to MST instance.
- Don’t use 802.1d, CST. If you will use standard base, use RST or MST.
- Take advantage of vlan load balancing; thus you can use your uplink capacity. It is called bisectional bandwidth as well.
- Vlan load balancing can be cumbersone, operationally hard but gives advantage of using all uplinks.
- 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.

- Three commonly used first hop redundancy protocols are HSRP, VRRP and GLBP.

Feature	HSRP	VRRP	GLBP
Transparent default gateway redundancy	Yes	Yes	Yes
Virtual IP address can also be a real address	No	Yes	No
IETF standard	No	Yes	No
Preempt is enabled by default	No	Yes	No
Multiple forwarding routers per group	No	No	Yes
Default Hello timer (seconds)	3	1	3

Figure-3 Basic FHRP Comparison

- HRSP and GLBP are the Cisco specific protocols but VRRP is IETF standard, so use VRRP if you need multivendor or interoperability.
- HSRP and VRRP use 1 Virtual IP and 1 Virtual MAC address for gateway functionality.



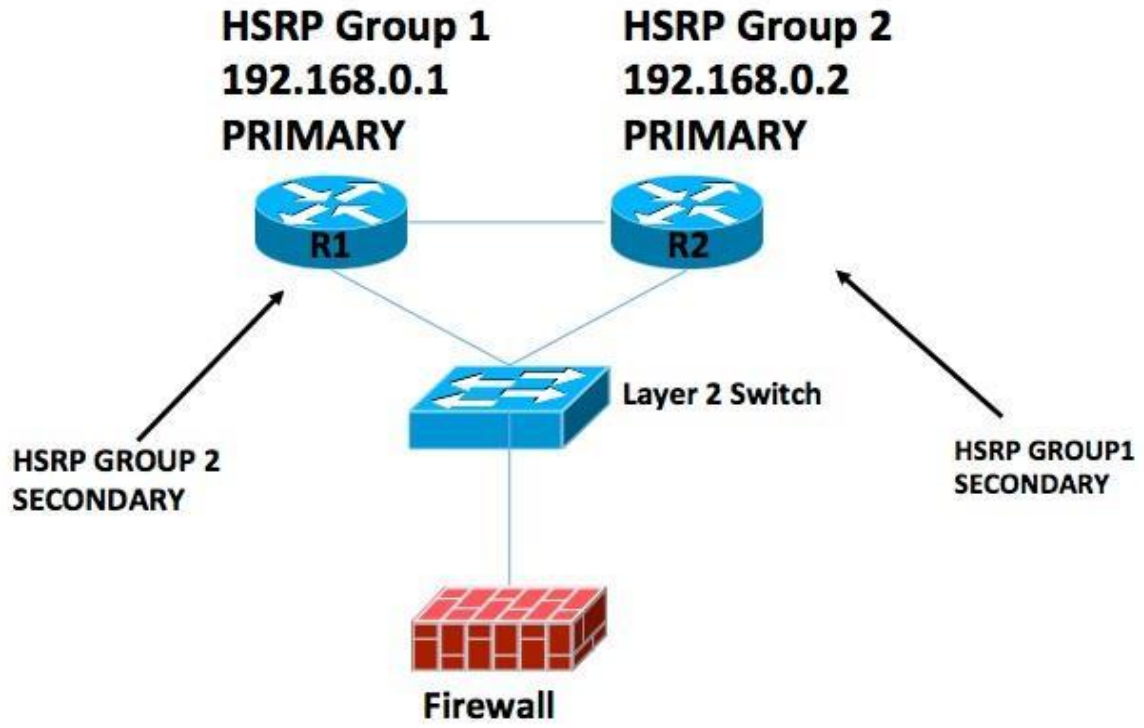
HSRP VIP-VMAC

- 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 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.

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

Let's look at the pictures below.



Egres Traffic Engineering Policy

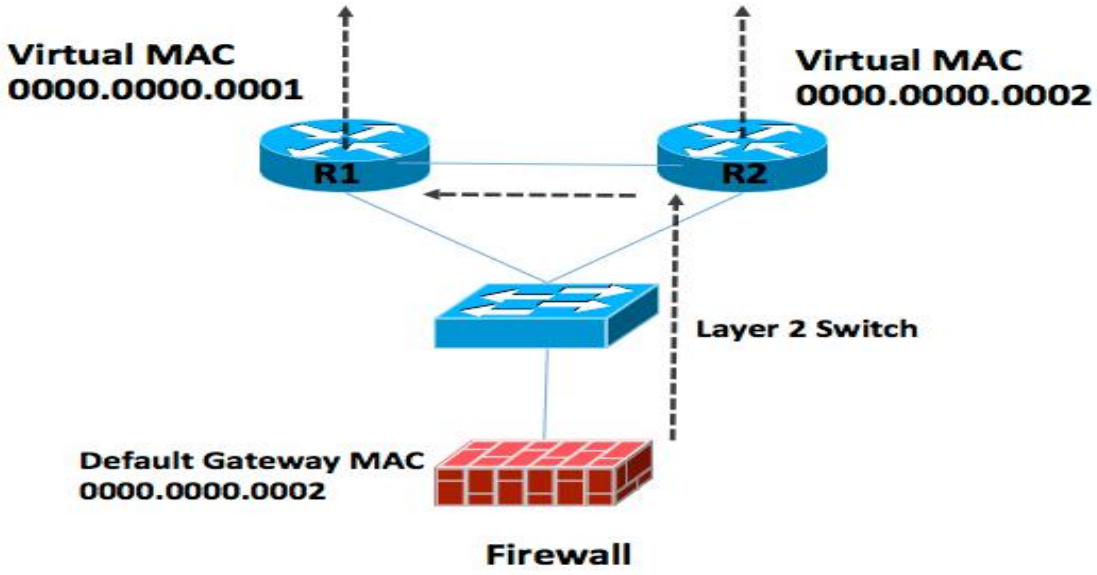
Create two HSRP Groups both routers; each router is active for one of the HSRP groups egress from firewall: Static routes on FW to HRSP group; BGP handles outbound forwarding.

On the firewall default route is pointed to the both internet gateways. (We divide the default route to the half actually.)

Firewall(config)#

First half of the default route is sent to the HRSP group 1 address route ouside 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



What about Gateway Load Balancing Protocol (GLBP)? 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 Internet Gateway is not used.

Spanning Tree – HSRP Interaction

- One important factor to take into account when tuning HRSP is its preemptive behavior. Prememption causes the HSRP is its preemptive behavior. Premption causes the primary HSRP peer to re-assume the primary role when it comes back online after a failure or maintenance event.

- 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.

- 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.

- The recommended best practice is to measure the system boot time, and set the HSRP preempt delay statement to 50 percent greater than this value. This ensures that the HSRP primary distribution node has established full connectivity to all parts of the network before HSRP preemption is allowed to occur

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 design.

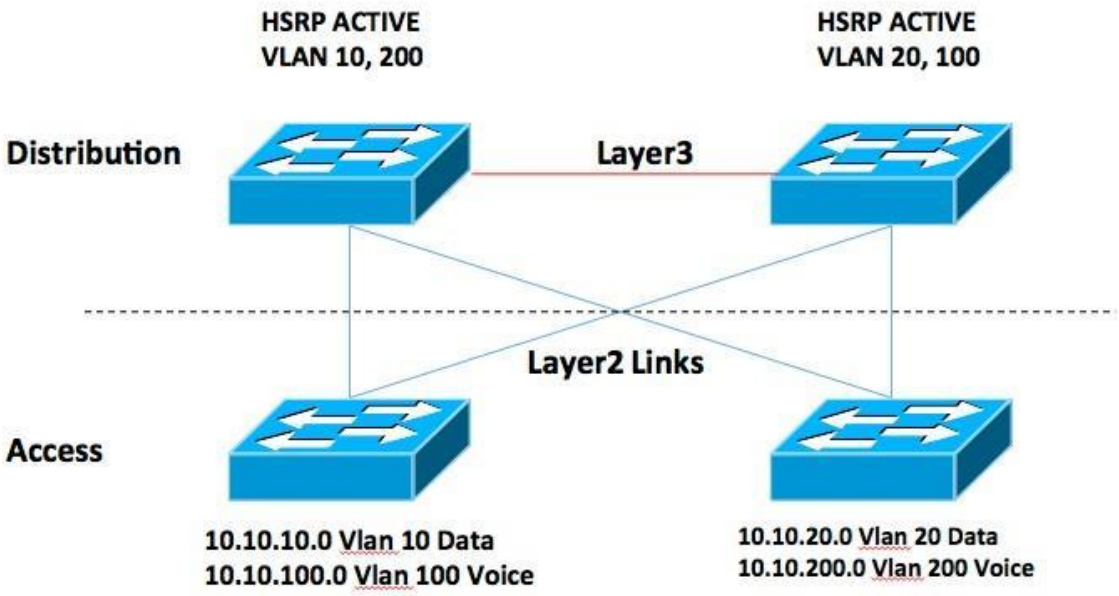
- In loop free design, the link between distribution layer switches is layer 3 thus there is no loop in the topology so spanning tree doesn't block any link.

- In looped design, the link between distribution layer switches is layer 2 so spanning tree will block one of the links to prevent loop.

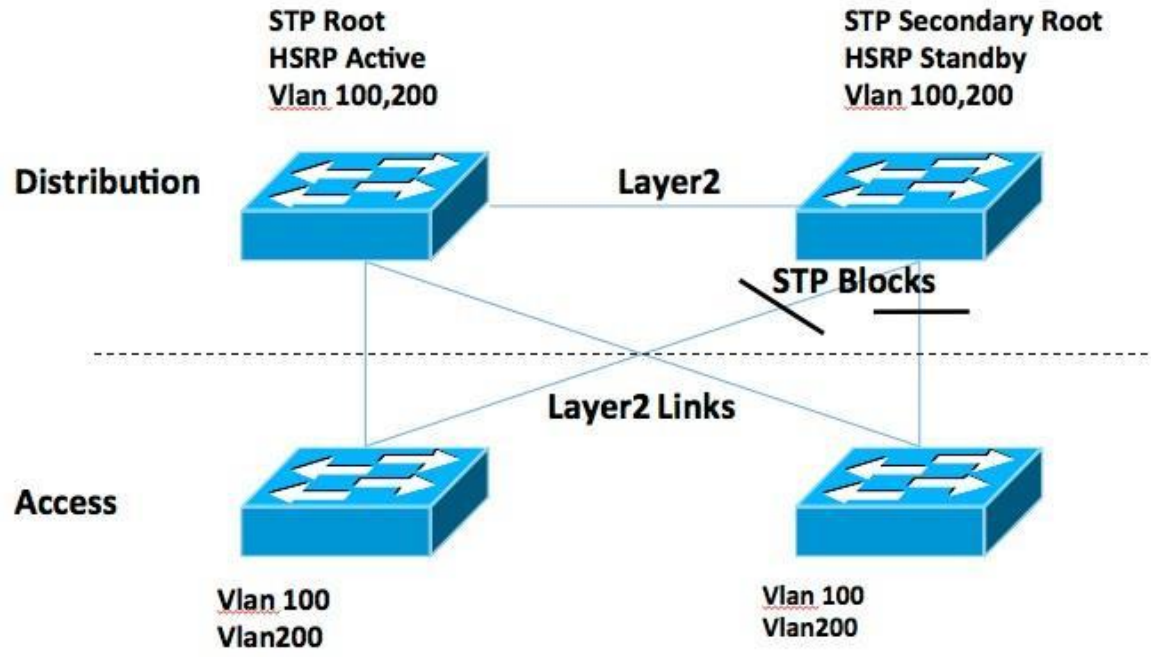
- Same vlan can be used on every access switch.

- We need to have FHRP since we want to have more than one distribution switch for redundancy

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



Layer 2 Loop free topology



Layer 2 Looped Topology

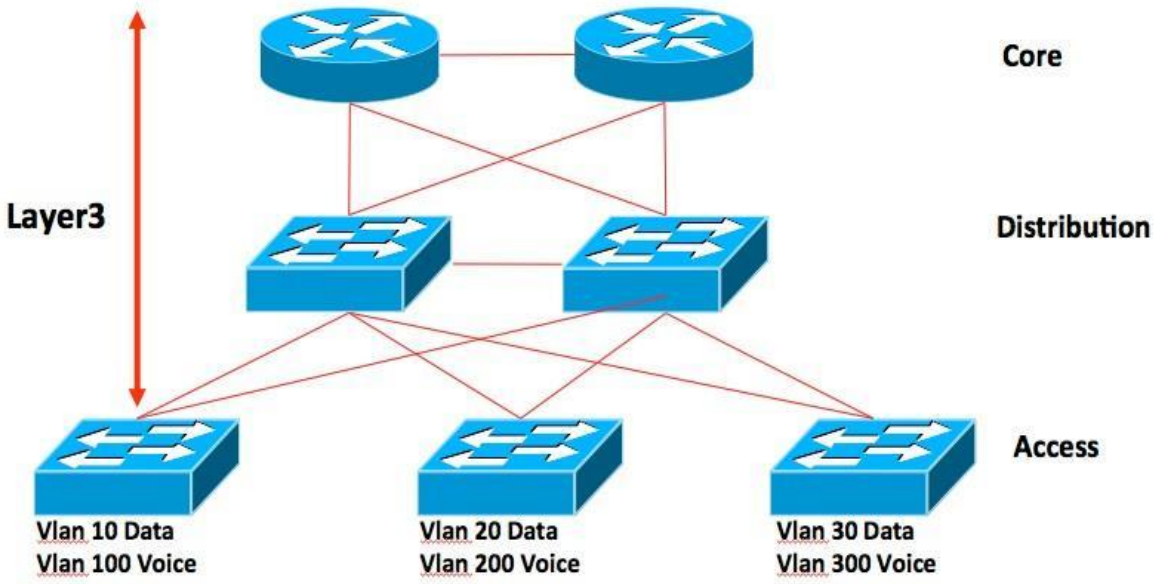
We want to align STP Root with FHRP active and if we have network services device such as Firewalls, we want to align active firewalls with STP and FHR

Question : Where would Layer2 looped design better from the design point of view ?

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

- 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

- No need to have any first hop redundancy protocol since the access layer switch is the first hop gateway.

- 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.

- There is no spanning tree anymore, at least between access and distribution layer, but still you may want to protect user site loop by enabling spanning tree at the edge.

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

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

Layer 2: Study Resources

Books :

http://www.amazon.com/Designing-Network-Architectures-Foundation-Learning/dp/1587142880/ref=sr_1_1?ie=UTF8&qid=1436567445&sr=8-1&keywords=CCDP+Arch

Videos

Ciscolive Session – BRKCRS – 2031 Ciscolive Session – BRKRST – 3363 Ciscolive Session – BRKCRS – 2468
https://www.youtube.com/watch?v=R75vN-frPhE

Articles :

http://www.pitt.edu/~dtipper/2011/COE.pdf

http://orhanergun.net/2015/05/common-networking-protocols-in-lan-wan-and-datacenter/

http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCInfra_6.pdf

http://blog.ine.com/2010/02/22/understanding-mstp/

http://blog.ine.com/wp-content/uploads/2011/11/understanding-stp-rstp-

convergence.pdf

https://www.cisco.com/web/ME/exposaudi2009/assets/docs/layer2_attacks_and_mitigation_t.pdf

General Routing Knowledge and Best Practices

- RIB – Routing Table, FIB – Forwarding Table

- Routing and Forwarding is not the same thing.

- Routing table keeps interface routes (Connected), Static routes and Dynamic routing protocols information

- Static route is a routing protocol

- In modern platform there is software and hardware forwarding information table.

- Cisco takes FIB and Adjacency table information for MAC rewrite and build CEF table

Load Balancing and Load Sharing is not the same thing. Understanding load sharing traffic count is critical for Unequal load sharing. Here is my article.

- OSPF and IS-IS can do the unequal cost load sharing with the help of MPLS-Traffic Engineering.

- There are some attempts in IETF for link-state protocol Unequal cost load balancing.

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

- Redistribution should be used in conjunction with the filtering mechanisms, route-map, distribute-list and so on.

- Be aware of routing loops.

- You don't redistribute between routing protocols directly, routes are installed in RIB and pull from the RIB to other protocol. So route should be in the RIB to be redistributed.

- If you can avoid, don't use redistribution. Managing it can be really complex.

- BGP doesn't have to converge slowly. Understand the data plane and control plane convergence difference.

- Route reflector is not necessarily a good idea

- You don't need to have multi-area OSPF or multi-level IS-IS

design, what business problem are you trying to solve?. Resiliency ? Opex ? Security ? Reliability ? Scalability ?

- Don't believe anyone blindly, people tend to believe whoever talks more, loudly. SDN doesn't have to be a good idea for example !

- Networks have to be complex. Robustness requires complexity. Imagine you have two routers, if you connect them through only one link, solution won't be robust for the failure. If you connect them via two links, it will be highly available, robust but more complex.

- In a design, the motto is ' Two is company, three is crowded '. You can't escape from the complexity but you should avoid unnecessary one.

- Also don't forget that things can be seen as simple but actually it might be very complex.



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.

You summarize the prefixes in one place of the network, you create sub optimality in another place.

You enable the new feature in one place of the network; such as IPv6 or Multicast and you open the network to a lot of new

security attack.

Those features were needed for robustness but they create fragility.

You may not see the impact immediately but it can be huge. This is known as Butterfly effect in design.

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

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? Or does it make life easier ? What is the purpose ?

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



General Routing Best Practices – Study Resources

Videos:

http://ripe61.ripe.net/archives/video/19/

Articles :

http://orhanergun.net/2015/01/route-redistribution-best-practices/

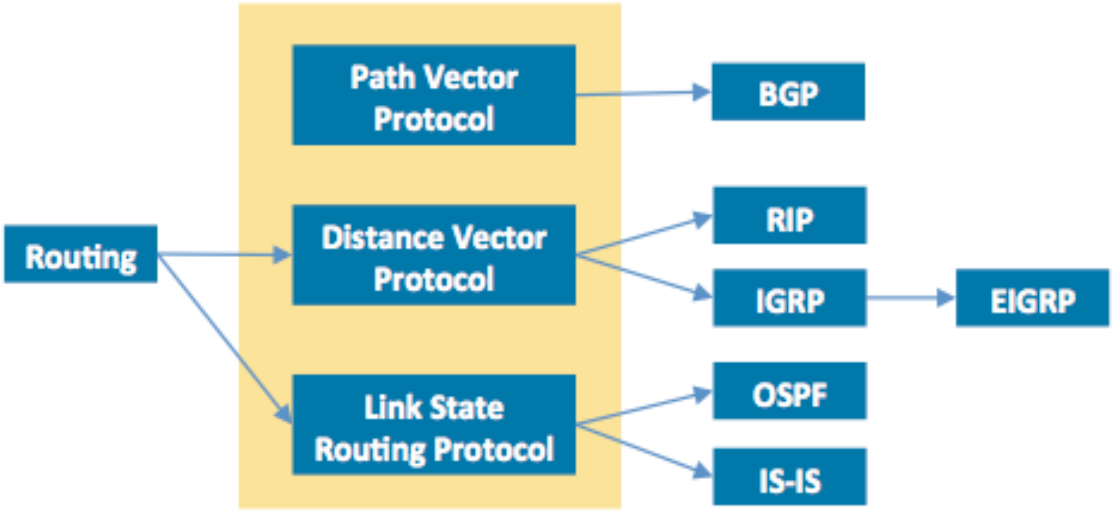
https://tools.ietf.org/html/draft-ietf-ospf-omp-02

https://www.ietf.org/rfc/rfc3439.txt

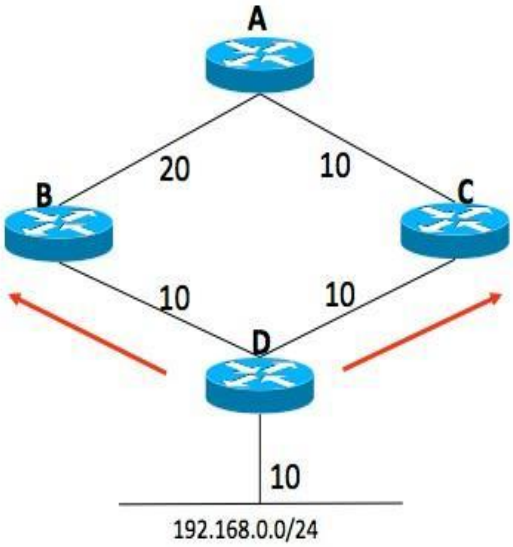
http://orhanergun.net/2015/06/if-the-system-lets-you-make-the-error-it-is-badly-designed/

http://orhanergun.net/2015/01/load-balancing-vs-load-sharing/

OSPF



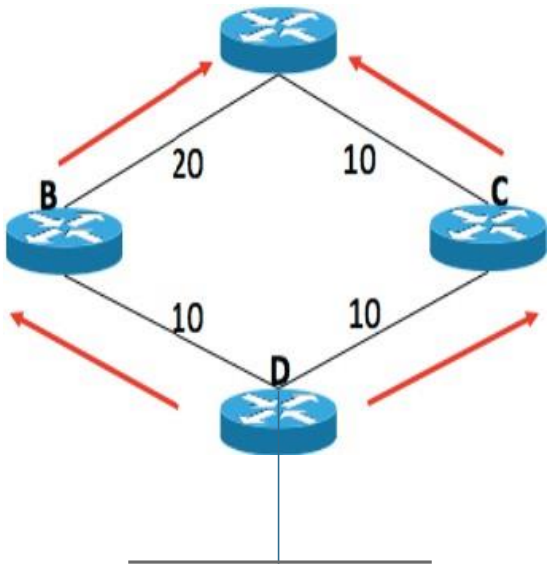
- OSPF is a link state routing protocol



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 C with metric 10

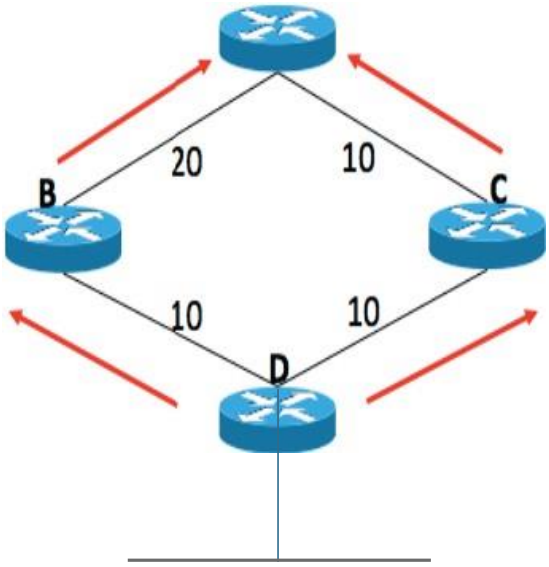
D advertises this information describing all of its links to Its neighbors B and C



192.168.0.0/24

This process of recording and re-transmittingIs 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).



192.168.0.0/24

Each router in the network uses this Information to build path tree to each destination in the network.

The J!Kill.i shortest path first (SPF) algorithm is used to build this tree.

Flooding of topology reachability information throughout the network seriously impacts the scaling of a network.

To resolve this network is broken into flooding domains/Areas in OSPF and Domain in IS-IS is used for hierarchy.

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

- In a particular area every routers have identical topology map. Every router knows which network behind which router and their metrics.

- Works differently on different media. On broadcast network DR creates a pseudo node to avoid unnecessary flooding

- Highest priority on the interface wins Designated Router (DR) election. If priorities are the same then highest router ID

- Unlike IS-IS, there is Backup Designated Router (BDR)
- There are 11 types of LSAs,5 of them are important for the routing design

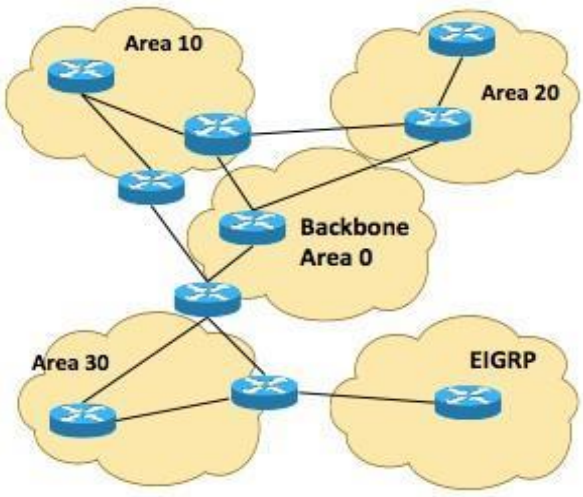
	Description
1. Router	• Router information • Connections to other routers • Connections to links(link states)
2. Pseudonode	• Pseudonode information • Connections to routers • Connections to broadcast link
3. Summary	• Destinations reachable within an area (flooding domain)
4. Border Router	• Cost to 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)

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

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

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



OSPF uses two level hierarchical model

Areas are use for scalability

Regular , Stub , Totally Stub , NSSA and Totally NSSA Areas

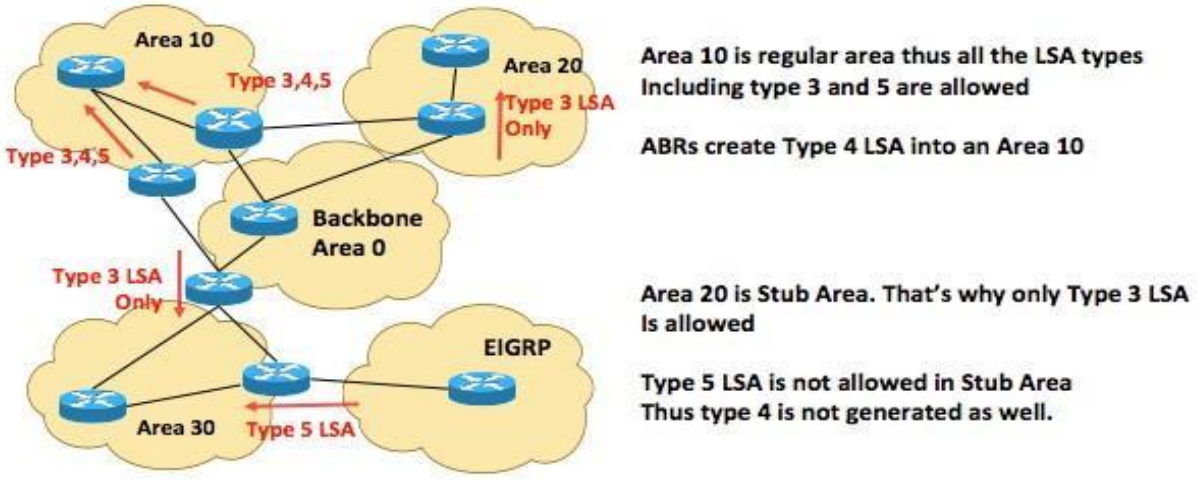
Router keeps separate link state database for each area which It belongs

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

- To reduce the impact of flooding, OSPF might use multi area concept



- ABR has a connection to more than one area

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

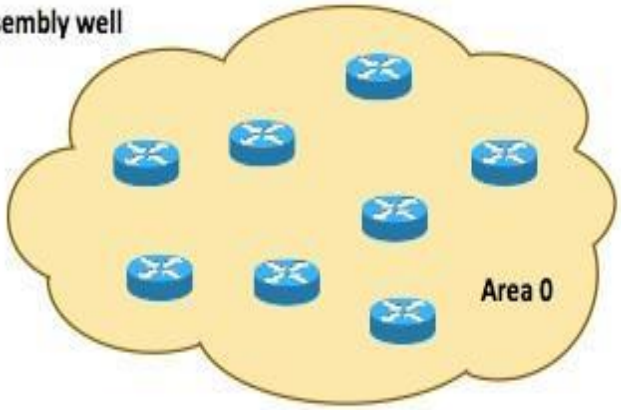
- OSPF interacts with many protocols in the network such as spanning tree, BGP, MPLS and so on. Understanding the impact of such an interaction is the first step for the robust network design.

How Many Router in an Area ?

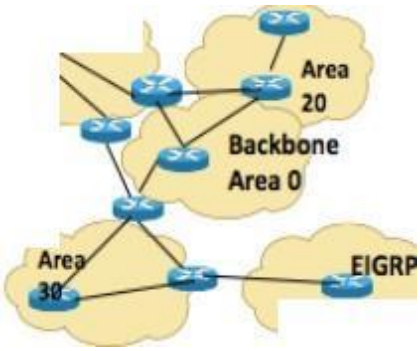
Number of neighbor is more important question which we should ask

Always try to keep router LSA under the MTU size to avoid fragmentation

Routers cannot deal with fragmentation and reassembly well



How Many ABRs per Area



In above diagram, there are 2 ABRs in Area 10. For the redundancy and optimal traffic flow, two is enough

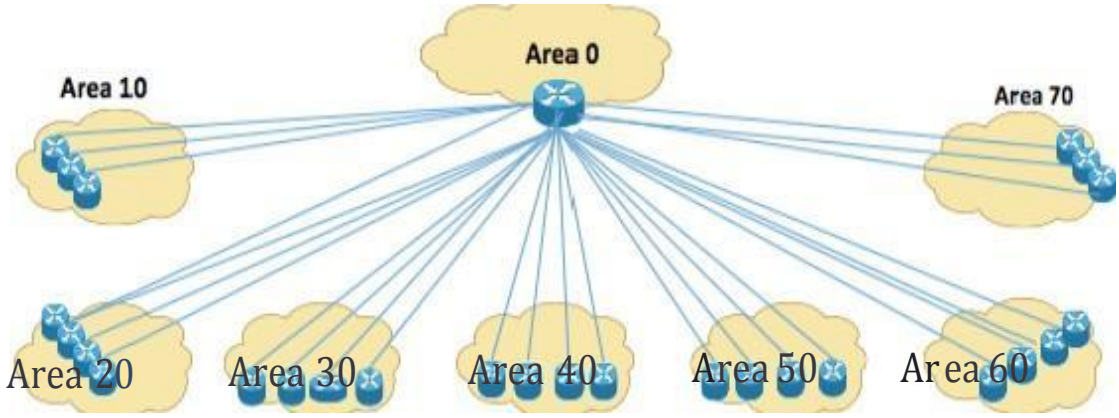
In network design, there is a word. • Two is company, here Is crowded.

More ABRs will create more Type 3 LSA replication within the

backbone and non-backbone areas

In large scale OSPF design, number of ABRs will have an huge Impact on number of prefixes

How Many Areas per ABR ?



More Areas per ABR might create a resource problem on the ABR

Much more T) lpe 3 LSA will be generated by the ABR. Between the Areas there will not be Type 1 or Type 2 LSA

- Aggregation removes reachability information and it can be done on either ABR or at ASBR

- Aggregation breaks the MPLS LSP, since LDP cannot have aggregated FEC unless the RFC 5283 – LDP Extension to Inter Area LSP is in use

- You don't need to configure summarization, it is automatic,

but you need to manually configure if you want aggregation.

FAST NETWORK 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:

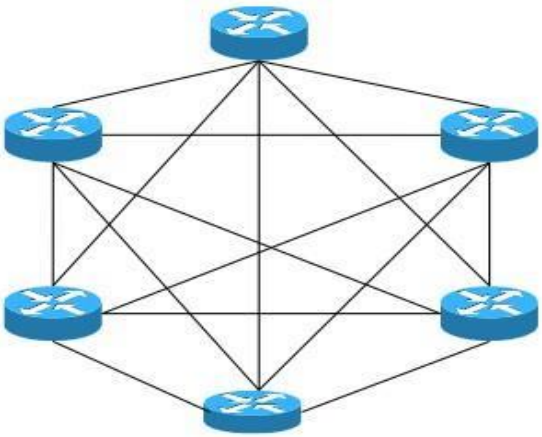
1. Failure Detection
2. Failure Propagation
3. Processing the new information
4. Routing and Forwarding table update

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	PROPAGATION	PROCESSING
Carrier Delays Debounce Timers Bidirectional Forwarding Detection – BFD Protocol Hello/Dead Timers	Interface event dampening LSA Pacing	Full, Partial and Incremental SPF MinLSA Arrival Interval LSA and SPF Throttling timers

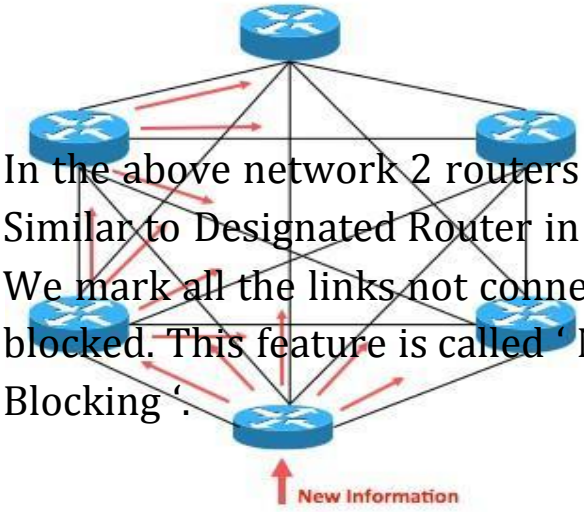


OSPF in FULL MESH TOPOLOGY

If there is only 2 routers totally 1 link
3 Routers 3 links
4 Routers 6 links
5 Routers 10 links
N Routers (N) (N-1)/2 links

Full Mesh Topologies are complex and OSPF needs to be tuned to work in Full Mesh topologies.

EIGRP works better in Full Mesh topologies than OSPF and IS-IS



Flooding in full mesh is a big concern, especially in large scale OSPF deployment

Each router receives at least one copy of the new information from each neighbor

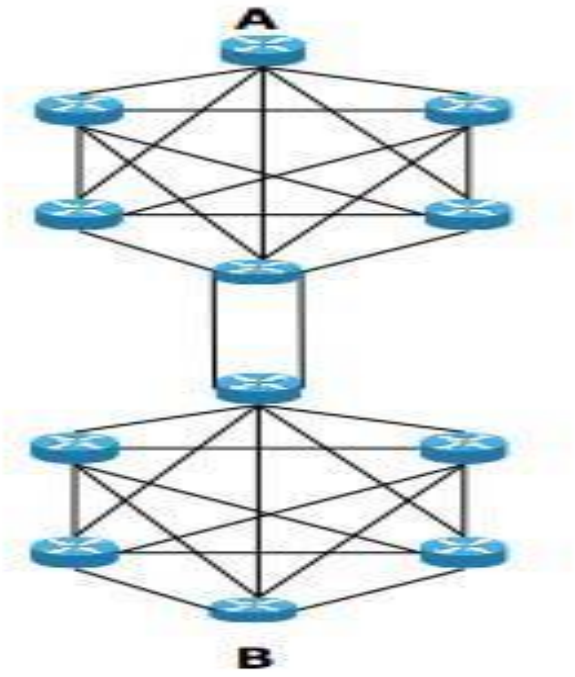
Mesh Group is one mechanism which we can use to reduce the amount of flooding in a full mesh

With Mesh Group, we designate couple routers in the topology to flood. Those routers will responsible from flooding event

In the above network 2 routers are designated as a flooder (Similar to Designated Router in Broadcast Networks !). We mark all the links not connected to those 2 routers as blocked. This feature is called ' Mesh Group ' or ' Flood Blocking '.

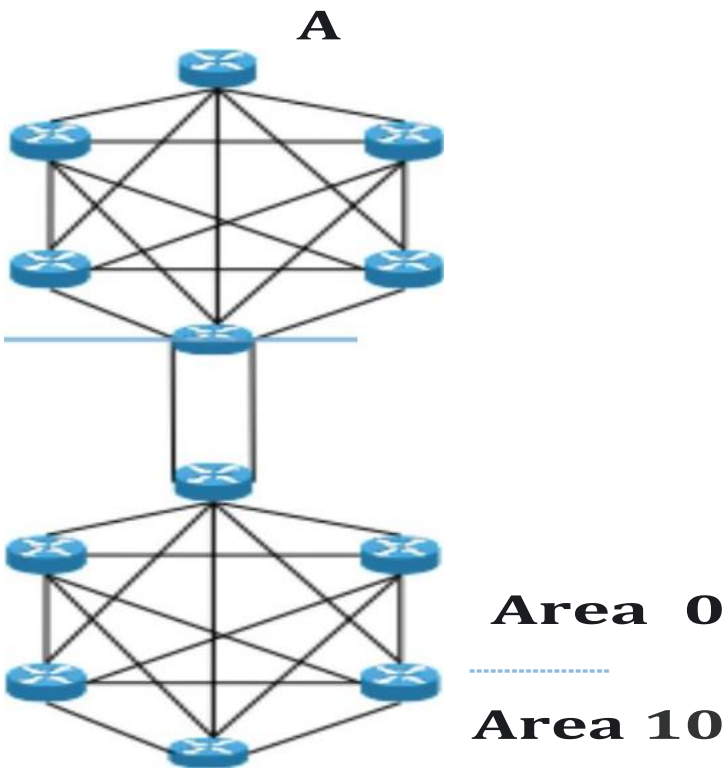
- Any topology posses a challenge for OSPF. OSPF needs to be adjusted to work in a particular topology which can create a configuration complexity. But still, it can deal with Ring topology better than distance vector.

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



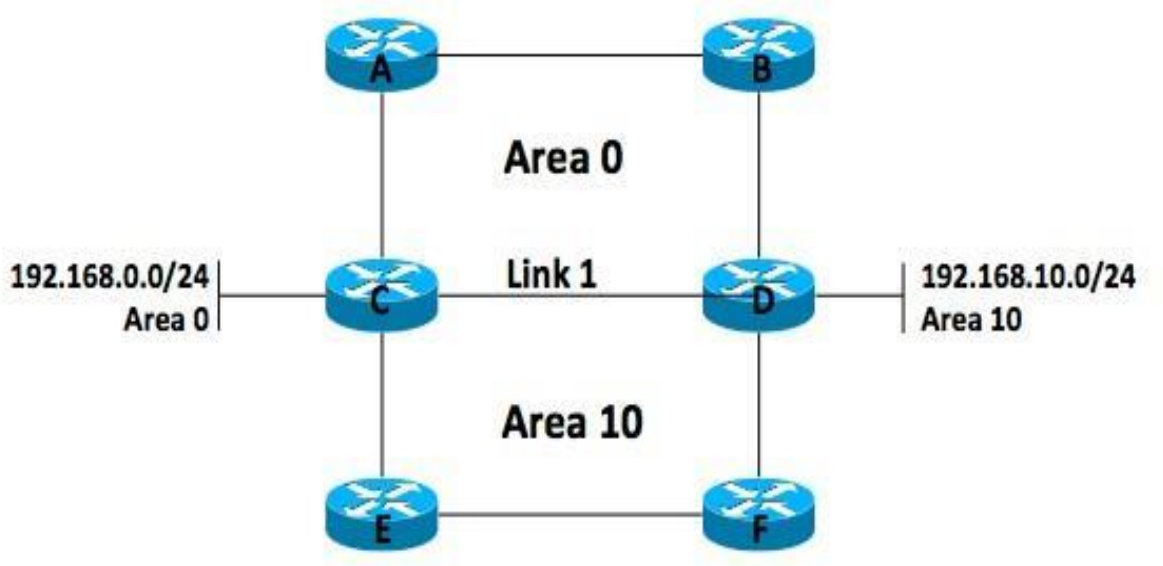
Between Router A and Router B there are 1800 different paths. (5x6) x 2 (5x6) If we would put all of them in a same area we would have flooding, convergence,
resource utilization, troubleshooting problems.

If we turn one of the middle routers as an ABR, we will have only 32 paths max. (5x6) +2


A
Area 0
Area 10

OSPF Case Study – 2

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 ? Why ? Is this a problem ? What is the possible solution.



- 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 1, router D will choose an path through B, A, and C to 192.168.0.0/24 with the same reason

This is suboptimal. Placing link into Area 1 and creating virtual link was the temporary solution. Real solution to this: RFC 5185 -OSPF Multi Area Adjacency.

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
```

OSPF – 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

http://blog.ine.com/wp-content/uploads/2011/01/Loop-Prevention-in-OSPF.pdf

http://orhanergun.net/2015/03/ospf-design-test/

ISKIS

- IS-IS is a link-state routing protocol, similar to OSPF.
- Commonly used in Service Providers and large Enterprise networks.
- Similar convergence characteristic with OSPF.
    - Excellent scalability
    - Flexibility in terms of tuning

Easily extensible with Type/Length/Value (TLV) extensions. This is the advantage of IS-IS over OSPF.

- 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.
- IPv6 Address Family support (RFC 2308)
- Multi-Topology support (RFC 5120)
- MPLS Traffic Engineering (RFC 3316)
- IS-IS is a Layer 2 protocol and is not encapsulated in IP

Each IS-IS router is identified with a Network Entity Title (NET)

- ISPs commonly choose addresses as follows:
- First 8 bits – pick a number (49 used in these examples)
  Next 16 bits – area
- Next 48 bits – router loopback address
- Final 8 bits – zero

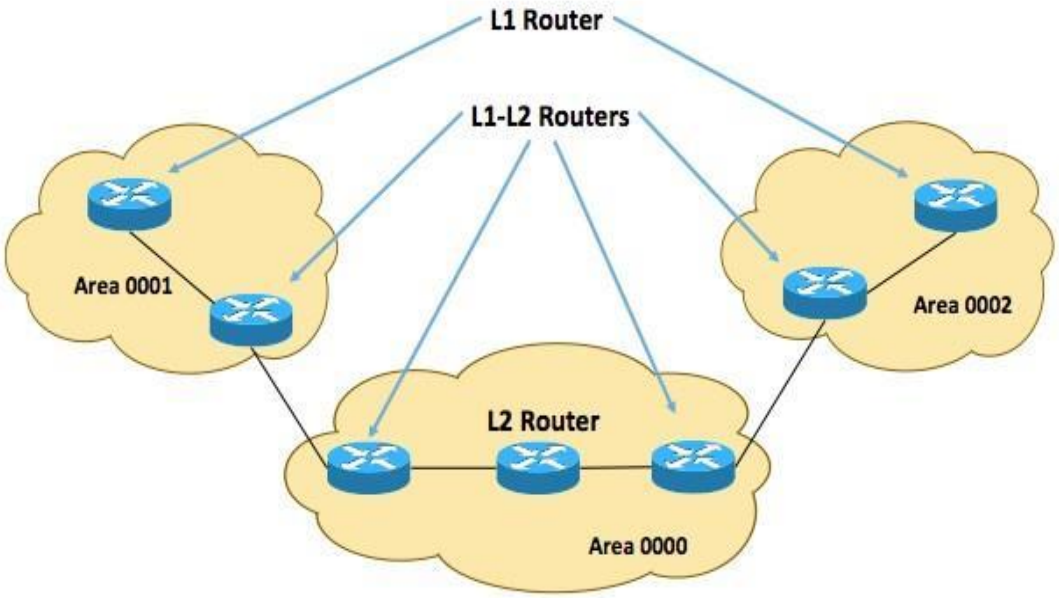
Example:

- NET: 49.0001.1921.6800.1001.00
- Router:192.168.1.1(loopback) in Area1

IS-IS and OSPF Terminology

OSPF	ISIS
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)

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



There is no backbone area in IS-IS as in the case of OSPF. There is only contiguous Level2 routers.

Three type of routers :

Level 1 Router :

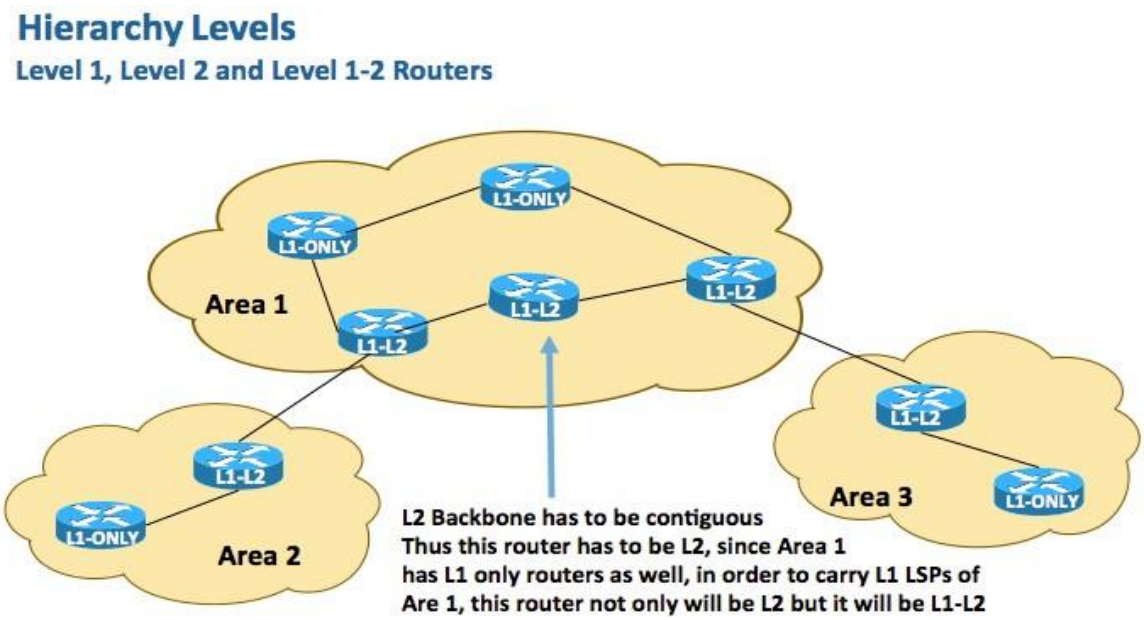
- Can only form adjacencies with Level 1 routers within the same area
- LSDB only carries intra area information

Level 2 Routers :

- Can form adjacencies in multiple areas
- Exchange information about the L2 area

Level 1-2 Routers

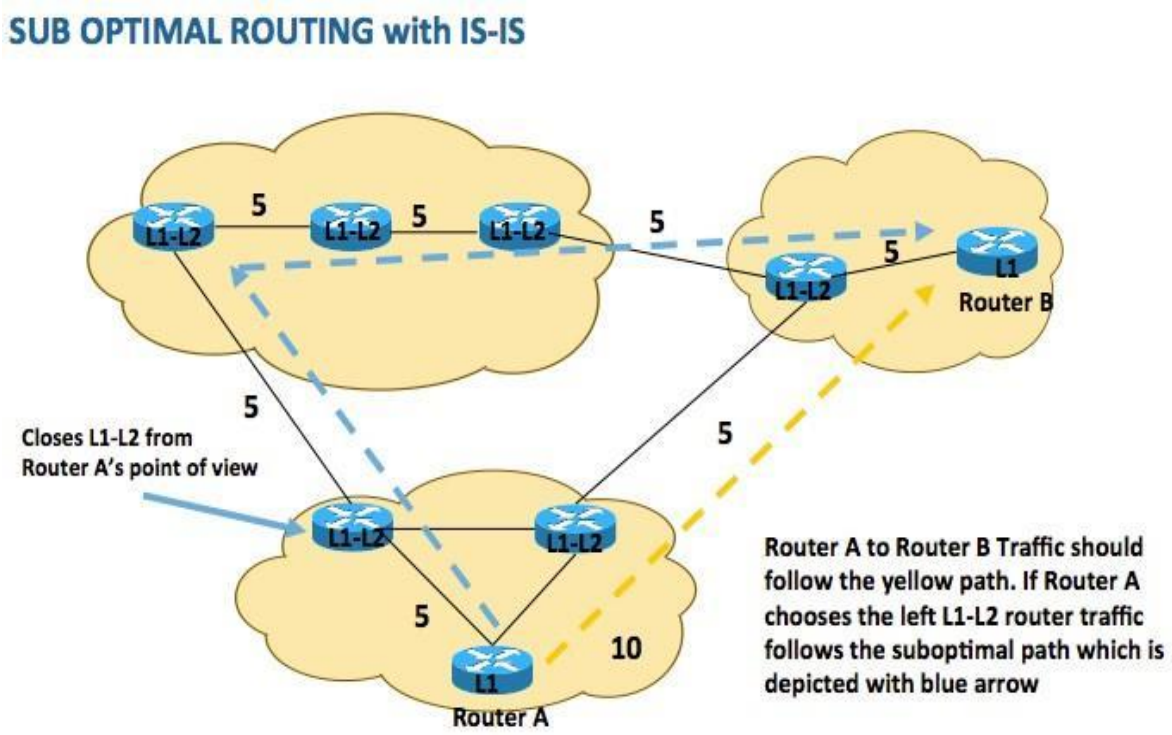
- These routers keep separate LSDB for each level, 1 for Level 1database, 1 for level 2 databases.
- These routers allow L1 routers to reach to other L1 in different area via the L2 topology
- Level 1 routers look at the ATT bit in L1 LSP of L1-L2 routers.
- And use it as a default route to the closes Level1-2 router in the area.



- In new IS-IS design, starting with L2-only is a best option. You can migrate to multi-level design easier. If you start with L1-only it will be hard. If you start with L1-L2 then all the routers have to keep two databases for every prefix.

- If you design multi level IS-IS and you have more than one

exit (L1-L2 routers) , you will more likely create a suboptimal routing. 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

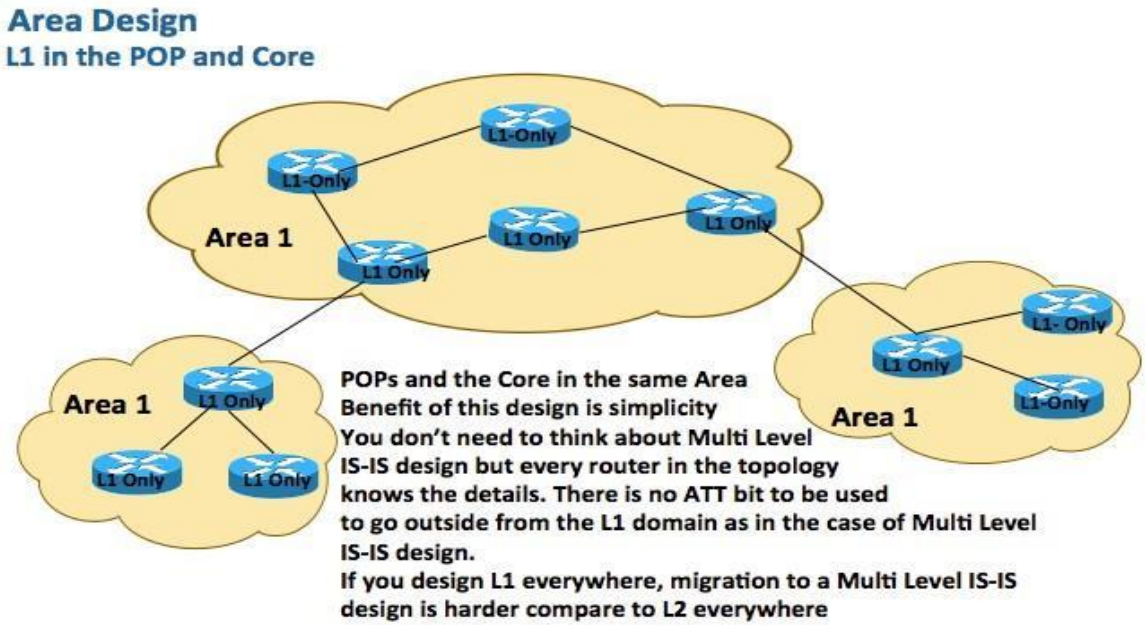


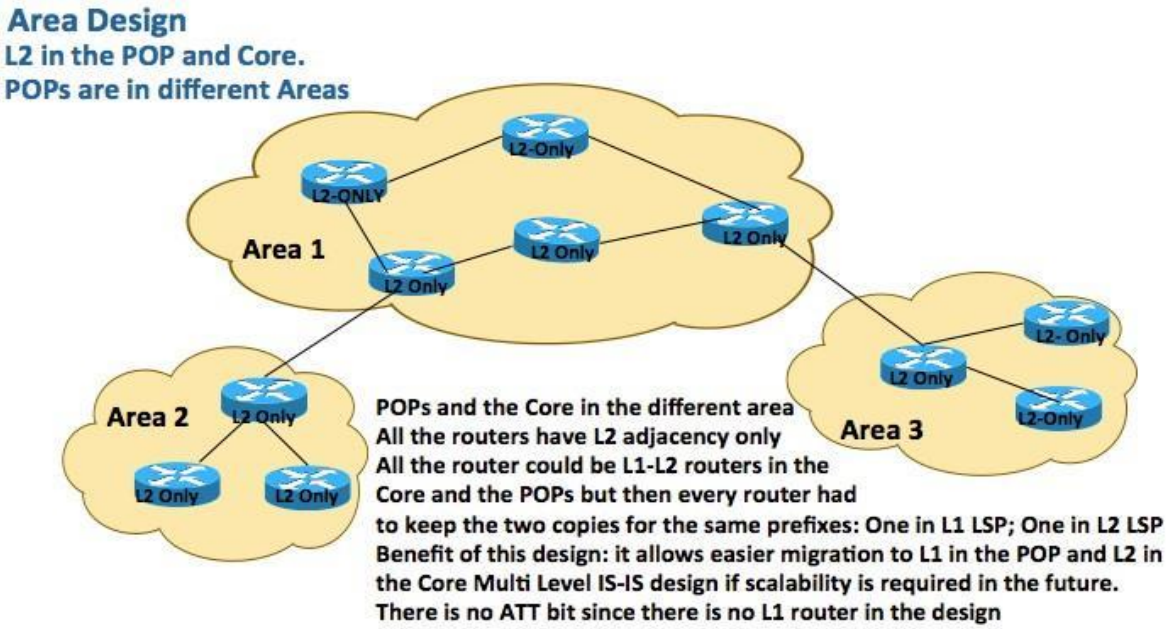
You can place edge and core in level 2 with same area

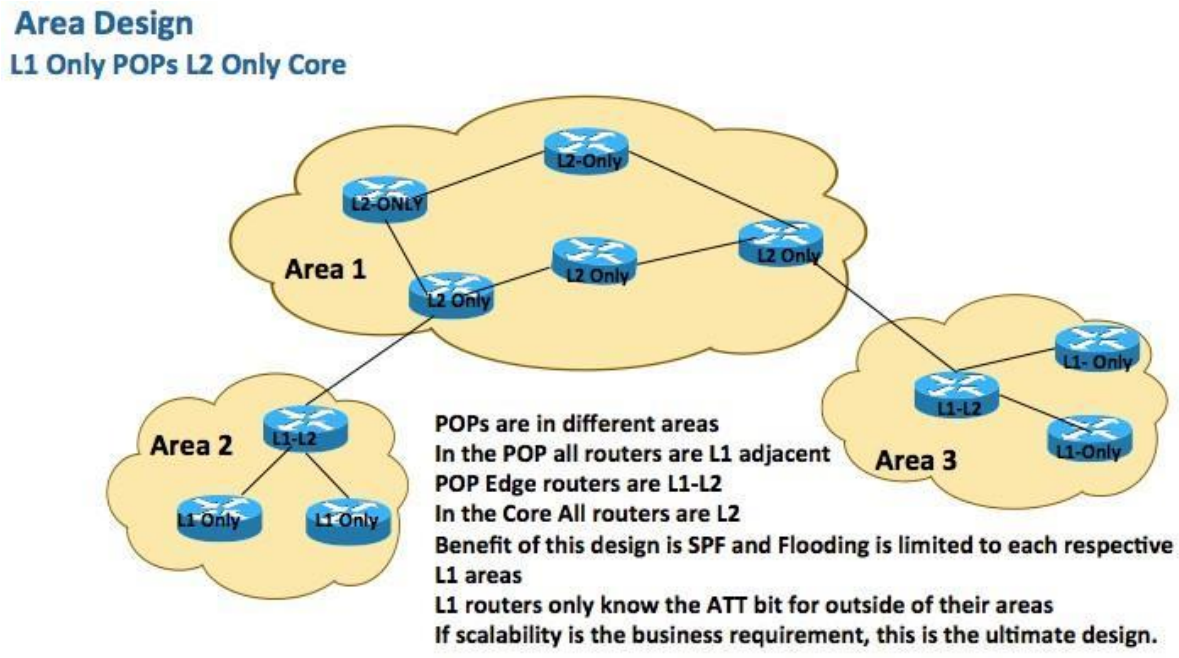
You can place edge and core in level 2 but in different areas which can make future multi-level migration easier.

Or as in the below picture; You can place the POPs in Level 1 areas and core in Level2.

This can create a suboptimal routing but provides excellent scalability.

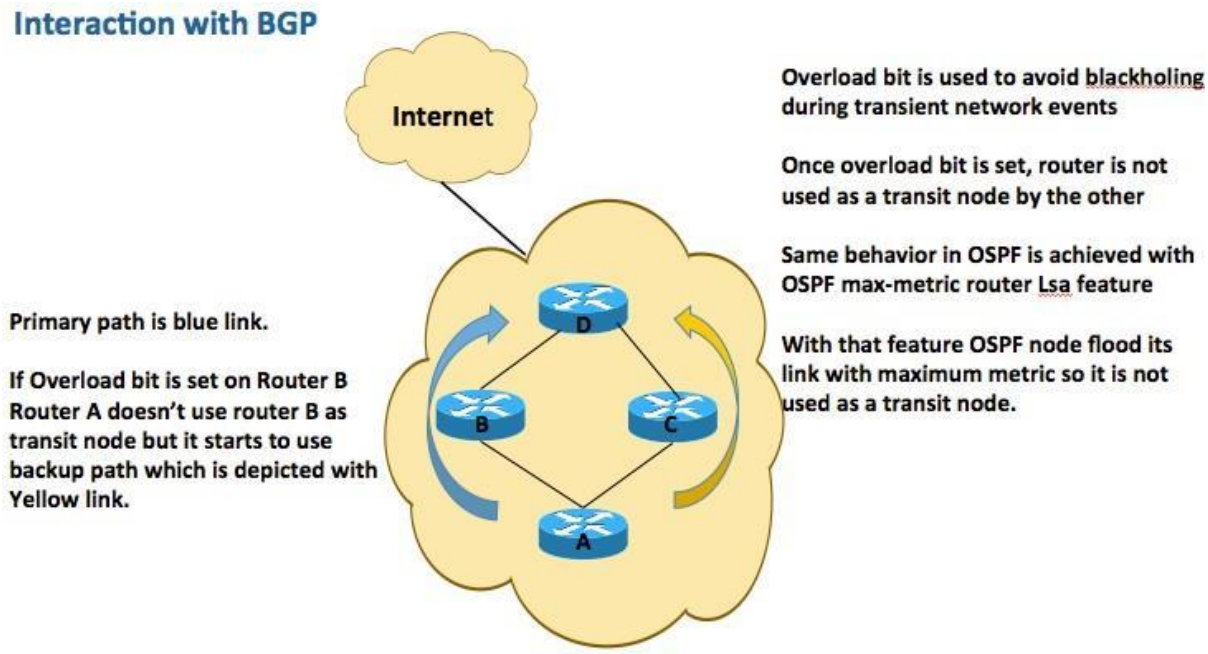


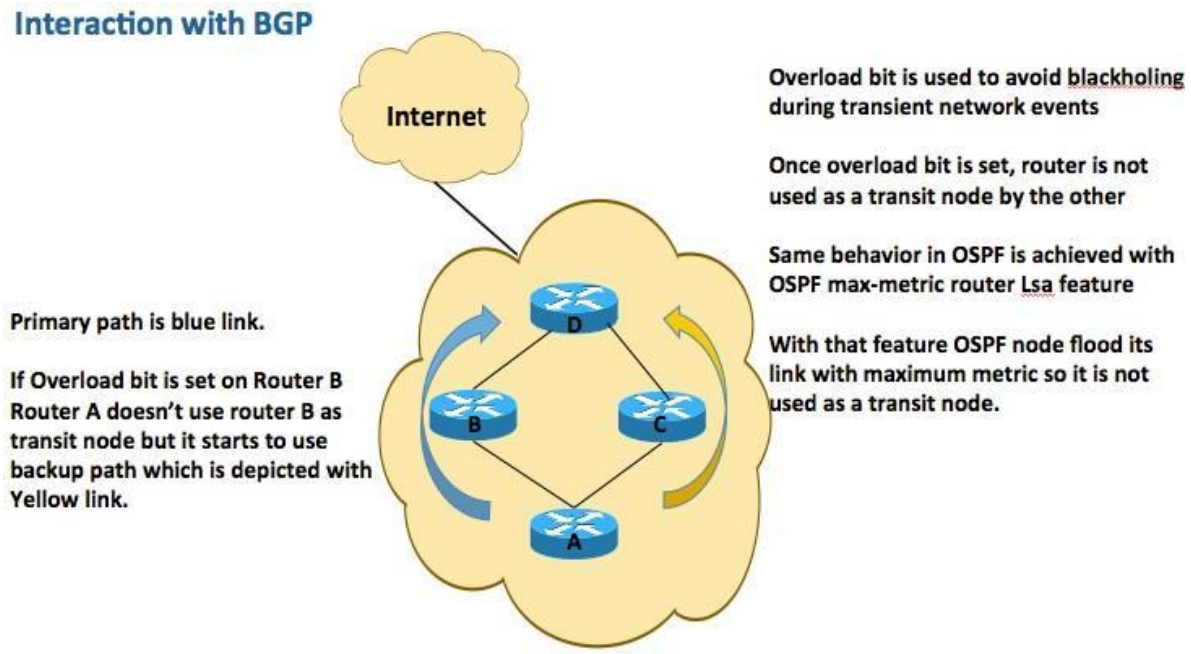




Network design, especially large scale network design is all about managing interaction and understand the tradeoffs. IS-IS as an IGP provides an underlay infrastructure for BGP, MPLS overlays.

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.





So, see below pictures for Interaction of IS-IS with MPLS. Multi-level IS-IS design breaks MPLS LSP.

At least three ways to fix it.

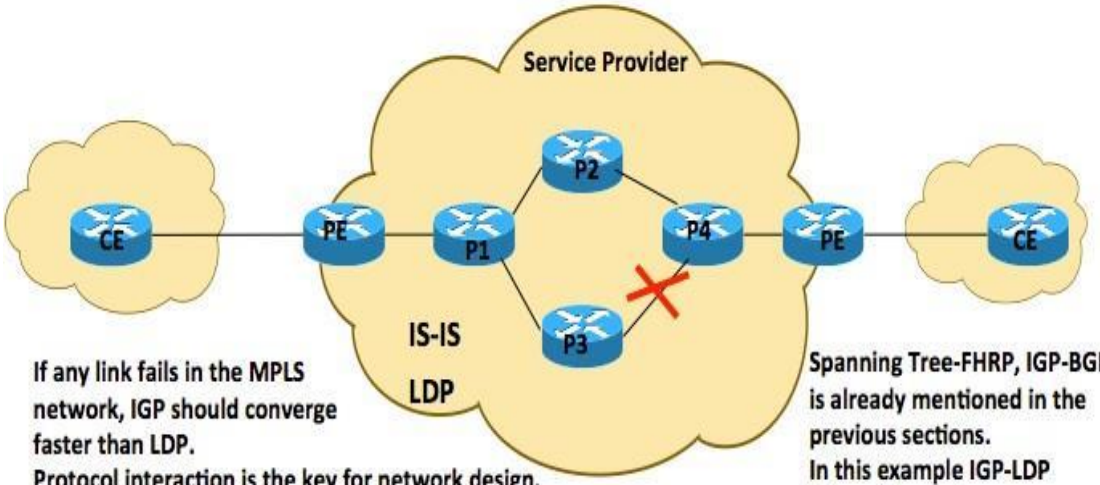
1.You can leak the loopback of PEs into L1 domain.

2. RFC 5185 allows to aggregated FECs. So you don't need /32 in the routing table to assign a label.

3. RFC 3107 , BGP +Label. Label is assigned by BGP. Seamless MPLS uses this concept.

4.

IGP – LDP SYNCHRONIZATION



If any link fails in the MPLS network, IGP should converge faster than LDP.
Protocol interaction is the key for network design.
They need to synchronize otherwise blackhole will occur.

Spanning Tree-FHRP, IGP-BGP is already mentioned in the previous sections.
In this example IGP-LDP interaction is shown.

LDP and RSVP both suffers from the same Problem.

This wouldn't be the case if separate protocol wouldn't be required for labeling.

Segment routing sends label in the IGP messages thus you don't need to worry about LDP-IGP Interaction with Segment Routing

IS-IS Case Study

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.

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.

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.

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

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 FASTNENT

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 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

2. Deploy IS-IS over entire network

   IS-IS admin distance is higher than OSPF in many platforms, leave as it is in this step.

   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.

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

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

4. Check OSPF Leftovers

   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 isis neighbor' so should be the same number of neighbors for both.

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

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.

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!

Books:

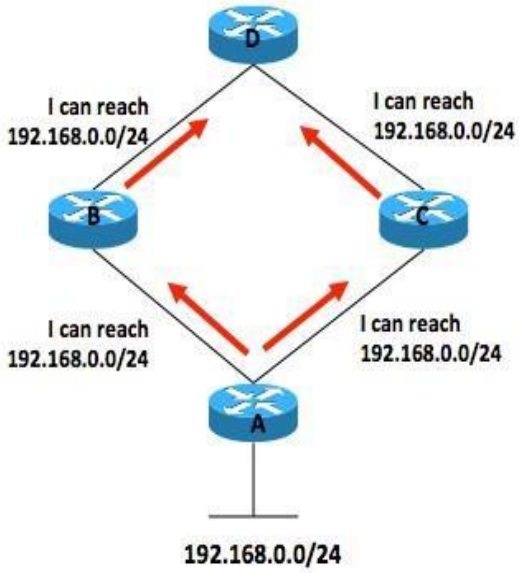
http://www.amazon.com/--Deployment-IP-Networks/dp/0201657724/ref=sr_1_1?ie=UTF8&qid=1436565940&sr=8-1&keywords=is-is+russ+white

Videos

Ciscolive Session – BRKRST – 2338

Podcast:

http://packetpushers.net/show-89-ospf-vs-is-is-smackdown-where-you-can-watch-their-eyes-reload/



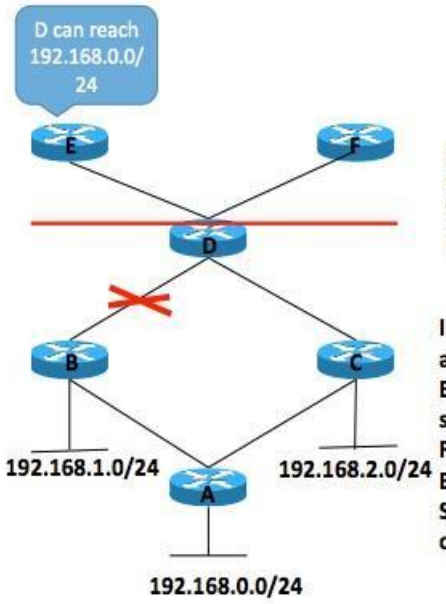
In Distance vector protocols topology information Is hidden beyond the next hop.

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 that 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.



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 for 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 topology don't keep the topology information of all the other nodes. Instead they trust what their neighbors tell them.

Feasible distance is the best path, primary path Successor is the next-hop router for the route

If my next-hop router's Reported distance is less than my feasible distance than my backup router is loop free alternate ! This is feasibility condition of EIGRP.

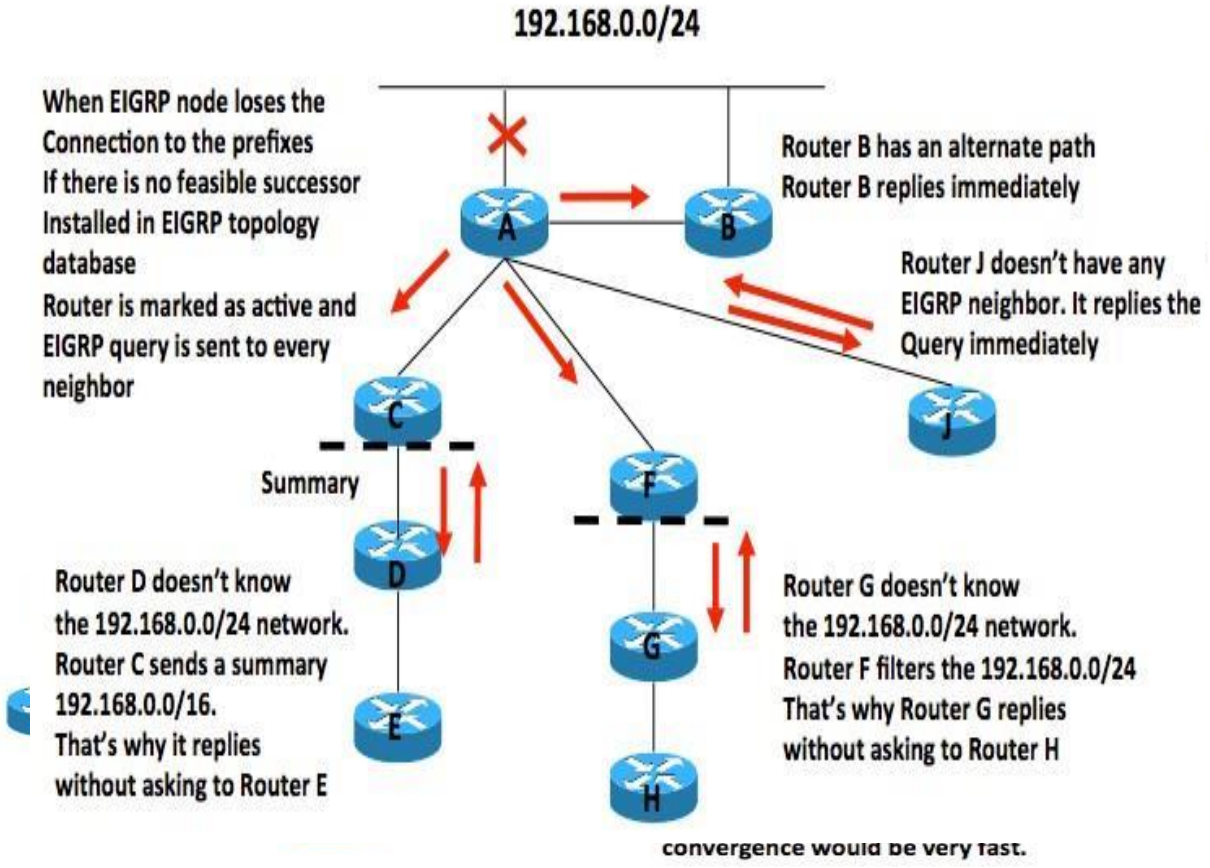
Feasible successors are the routers which satisfy the feasibility condition. These are the Backup routers

Feasible successors are placed in EIGRP topology table

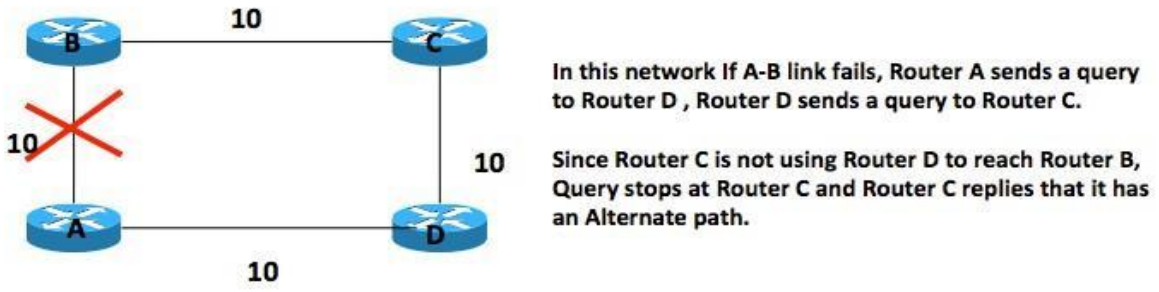
In dual algorithm there is only one Successor and all the others can be FS or looped.

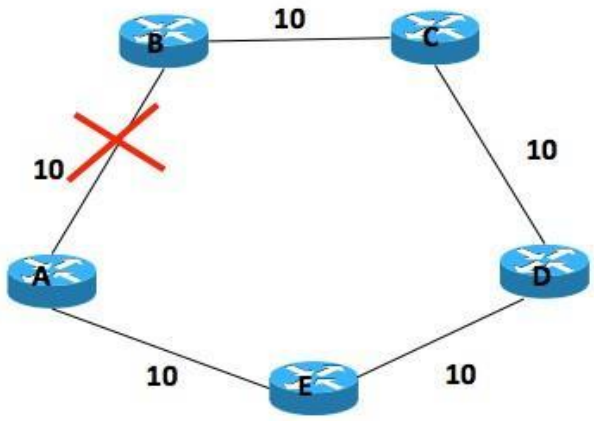
But Cisco's EIGRP implementation says , if two routers have equal FD (COST) then both are successor. So normally there is no ECMP in Dual Algorithm !

Summary and Filters reduces the query domain to one hop beyond.



EIGRP is best in Hub and Spoke and Mesh topologies, but not good at Rings. See the below examples.





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.

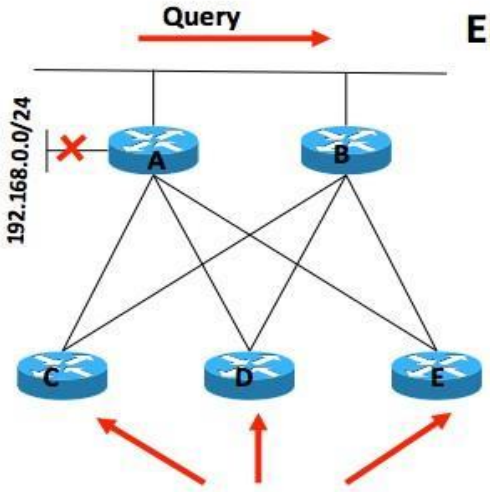
Query range in this network is three hops.

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 AS in BGP filter the AS-Paths between the providers.

EIGRP Stub allows to be not queried , router will not advertise the routes to peer if route is learned from another peer.



EIGRP STUB

When EIGRP Stub feature is enabled spoke sites are not used as transit site.

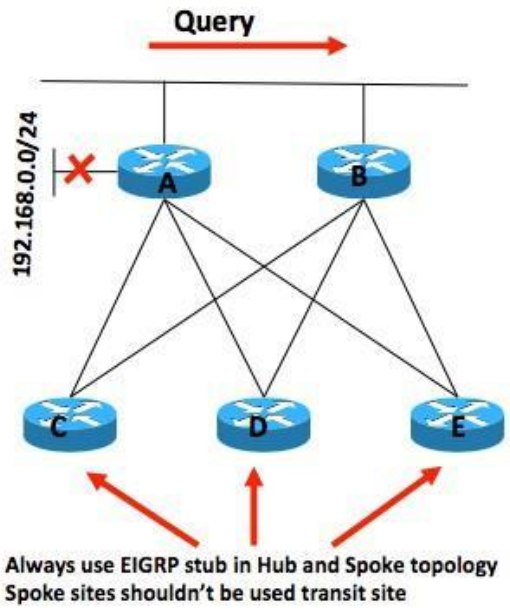
Also Hub site doesn't even send an EIGRP query if 192.168.0.0/24 network fails as in the picture Router A sends a query only to Router B. This helps for the convergence, provide faster convergence.

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.

Always use EIGRP stub in Hub and Spoke topology
Spoke sites shouldn't be used transit site

Access network always should be configured as stub



EIGRP is the most scalable IGP protocol inn 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.

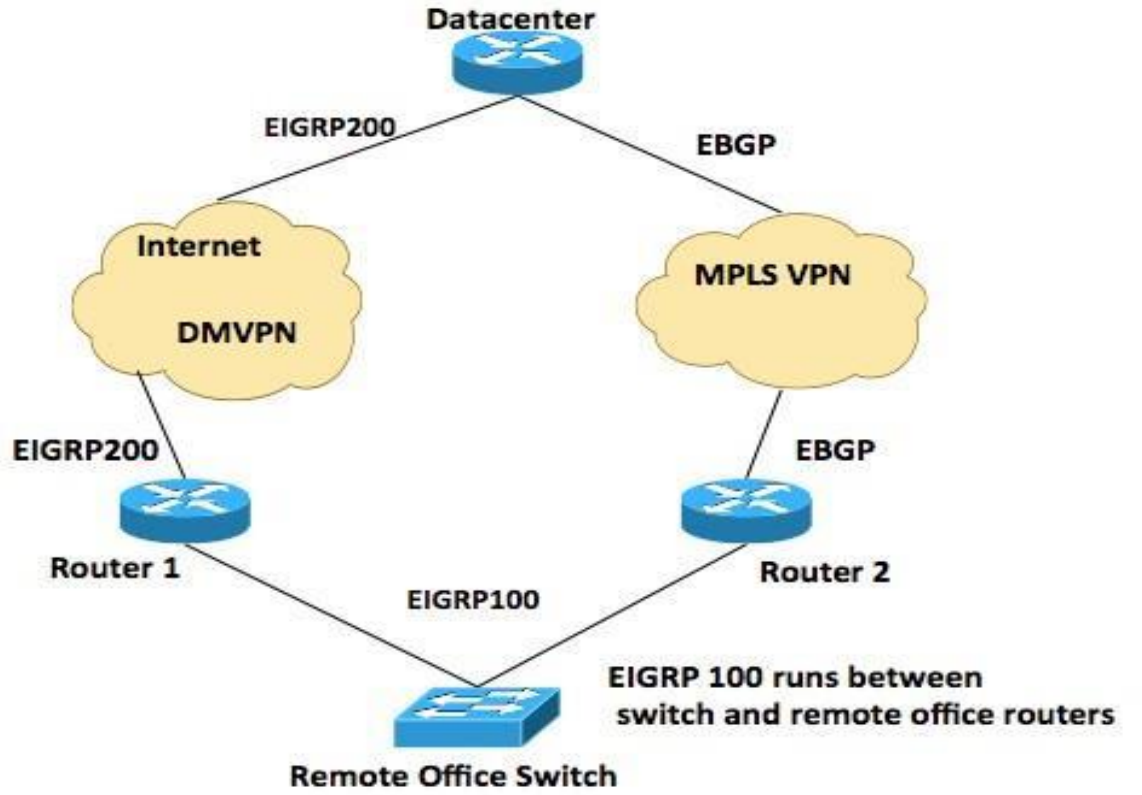
Otherwise Spoke sites might be used as transit nodes.

Summarization metric is received from a route which has a lowest metric In this case that route goes down, metric change so summarization effect to upstream will be lost

You can create a loopback interface within the summary address range with a lower metric than any other route in the summary but problem with this approach, if all the routes fail in that summary range but loopback stays, black hole occurs

Better answer to this particular problem is, within the EIGRP named mode, you can use summary-metric ,with that you statically states the metric you want to use.

EIGRP Case



- In the above topology diagram, customer wants to use MPLS L3 VPN (Right one) as its primary path between Remote office and the Datacenter.

- Customer uses EIGRP 100 for the Local Area Network inside the office.
- Customer runs EIGRP AS 200 over DMVPN.
- 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.

Question 1

Should customer use EIGRP the same AS on the DMVPN and the LAN ?

Answer : No it shouldn't. Since Customer requirement is to use MPLS VPN as primary path and nothing specified for specific application only use MPLS VPN and other should use DMVPN, 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.

Question 2

What is the path between remote office and the datacenter ?

Answer : 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 : Yes. Because if customer uses different EIGRP AS on LAN and DMVPN, with just metric adjustment, MPLS VPN path can be used as primary.

Question 4

What happens when the primary MPLS VPN link goes down ?

Answer : It depends. If you redistribute the data center prefixes which are received by R1 on R2, R2 sends the traffic towards switch and switch uses only R1.

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, only DMVPN link is used. So DMVPN link is used as primary when the failure happens.

Question 1

What happens when failed MPLS VPN link comes back ?

Answer: This is tricky part. R2 receives the datacenter prefixes over MPLS VPN path via EBGP, also from R1 via EIGRP . When R2 receives the prefixes from R1 as an EIGRP route those prefixes shouldn't be redistributed on R2 to send through MPLS VPN path.

If you don't redistribute them, 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.

If you redistribute them on R2, when the link comes back, R2 continues to use MPLS VPN path, so switch can do load sharing or with metric adjustment you can force to use MPLS as primary.

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.

EIRP – 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

http://blog.ine.com/2009/05/01/understanding-unequal-cost-load-balancing/

BFP

Scale extremely very large, robust. Runs over TCP and probably that's why it is considered as robust since TCP inherently reliable.

Multi-Protocol , many address family ! today almost a 20 different mechanism is carried over BGP. New AFI, SAFI allow this extensibility.

EBGP and IBGP are our main focus. If the BGP connection between the domains, so different Autonomous Systems then the connections is called EBGP (External BGP).

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

BGP Path Selection.

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

There might be vendor specific attributes such as Weight. Also there are some intermediary steps which is not used commonly. Below is the BGP best path selection criteria list as a designer we have to keep in mind.

- Local Preference
- As-Path D Origin D MED
- Prefer EBGP over IBGP
- Lowest IGP metric to the BGP next hop(Hot Potato)
- Multipath
- Lastly prefer lowest neighbor address. Let's go through details of EBGP.

EBGP

BGP Inbound traffic engineering can be achieved in multiple ways.

- MED (BGP External metric attribute)
- AS-Path prepending
- Community

Med is used between two ASes. It is not carried between multiple ISPs. It is not transitive attribute.

AS Path is mandatory and it is carried over entire internet , although some service providers can filter excessive prepending.

Community is sent over the BGP session then receiving ISP takes an action to prefer desired path. Generally it is implemented with Local Preference at the receiving side.

Very important design topic in BGP is BGP peering.

BGP Perring

To understand peering, first we must understand how networks connect to each other on the Internet.

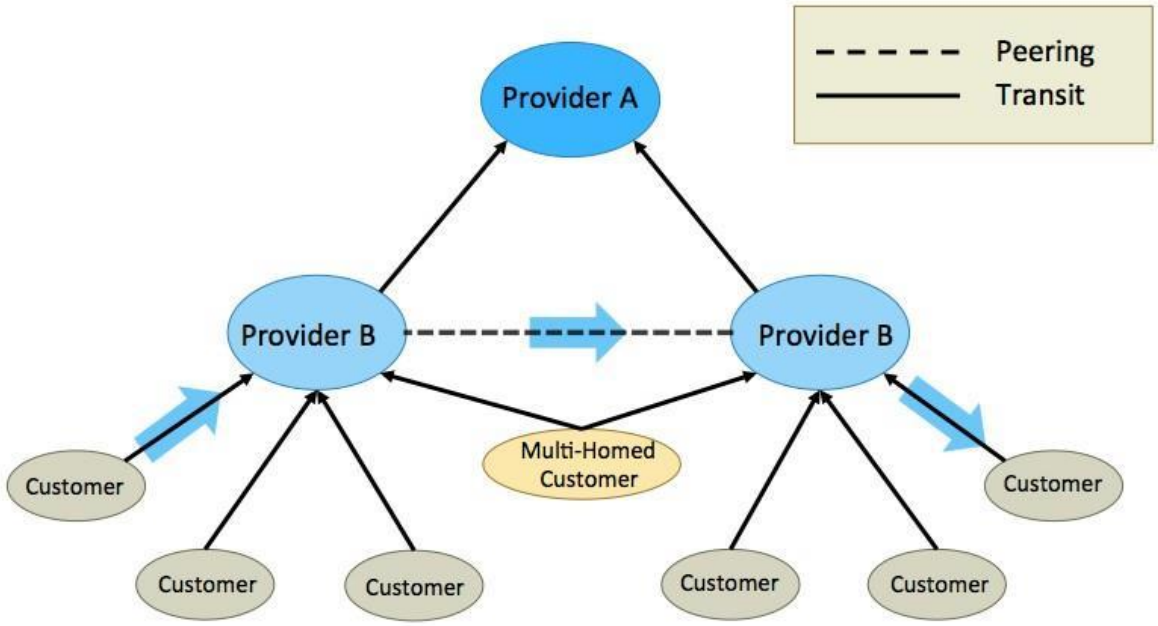
The Internet is a collection of many individual networks, which interconnect with each other under the common framework of ensuring global reachability between any two points.

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.

Peer – Two networks who get together and agree to exchange traffic between each others' networks, typically for free.



Benefits of Perring

Reduced operating cost : You are no longer paying a transit provider to deliver some portion of your traffic. Peering traffic is free , so this reduces your transit bills.

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

Distribution of traffic : By distributing traffic over interconnections with many different networks, you can potentially improve your ability to scale.

Two types of peering

:**Private peering** :
Private Peering is a direct interconnection between two networks, using a dedicated transport service or fiber.

It may also be called a Private Network Interconnect, or PNI.

- Inside a datacenter this is usually a dark fiber “cross-connect”.
- It may also be a Telco-delivered circuit as well.

Though these typically cost a lost of money, and are avoided

whenever possible. Often the cost of the interconnection itself is shared.

A common model is “I’ll buy this one, you can buy the next one”.

Public peering : It is done at the exchange point. Commonly referred as ‘ IX’. What are the considerations of public vs. private peering?

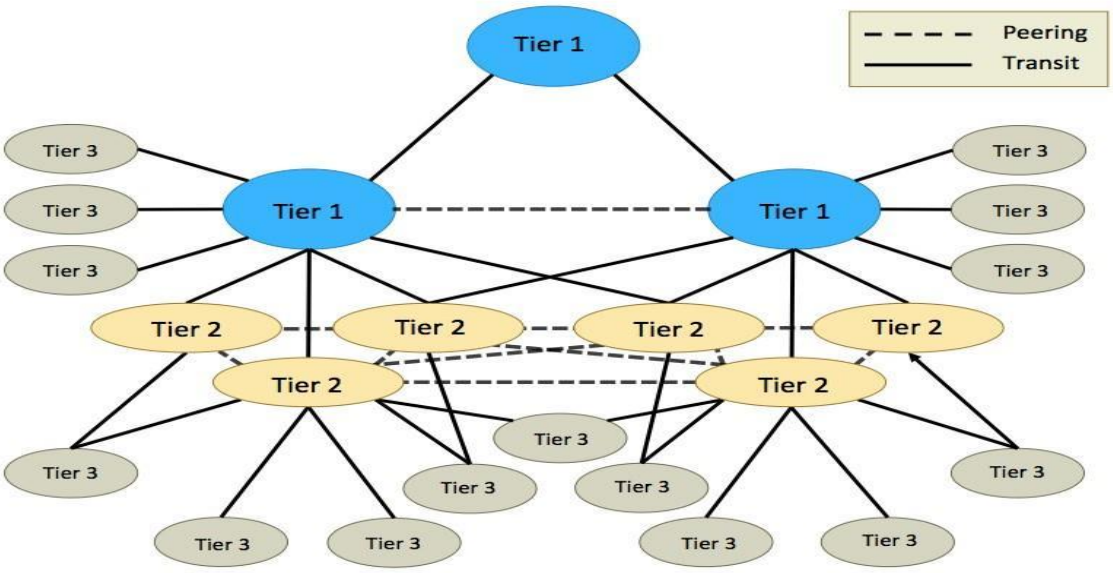
- An Exchange Point is typically the optimal choice for a network maintaining a large number of “small” interconnections.
- Trying to maintain private interconnections with dedicated physical links for every peer is often financially or logistically prohibitive.
    - For example, maintaining 100 GigE cross-connects to peer with 100 small peers would probably exceed the cost of an Exchange Point port
    - Not to mention the overhead of provisioning and maintaining the ports.
- But a Private Peer is typically the optimal choice for two networks exchanging a large volume of traffic.

For example, if two networks exchange 10Gbps of traffic with each other, it is probably cheaper and easier to provision a dedicated 10GE between them, rather than have them each pay for another 10GE exchange port.

Many networks maintain a mix of public and private peers.

When we talk about service provider network interconnections, mostly use ‘ Tier “ definition.

- Tier 1 – A network which does not purchase transit from any other network, and therefore peers with every other Tier 1 network to maintain global reachability.
- Tier 2 – A network with transit customers and some peering, but which still buys full transit to reach some portion of the Internet.
- Tier 3 – A stub network, typically without any transit customers, and without any peering relationships.

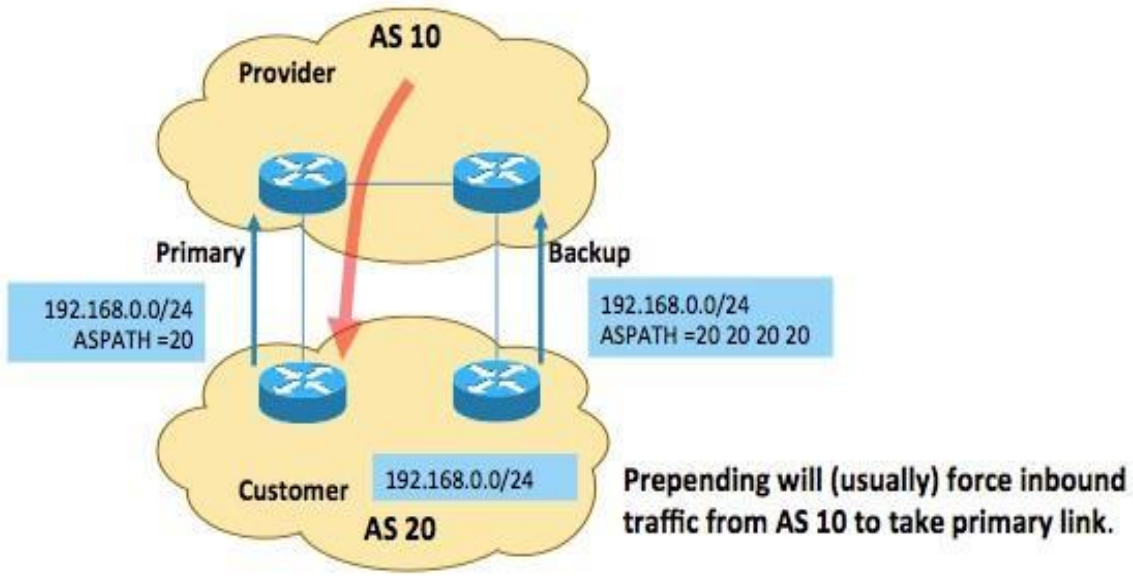


Service provider unless customer change their path preference with communities, chooses almost always Customer over Peering links vs. transit links.

For example, Local preference 100 towards customer , 90 towards peer, 80 towards transit provider. Higher local preference is preferred.

Over the global internet, intended policy is not always the outcome. See below examples.

Customer AS20 wants to use left link as primary path.



IBGP

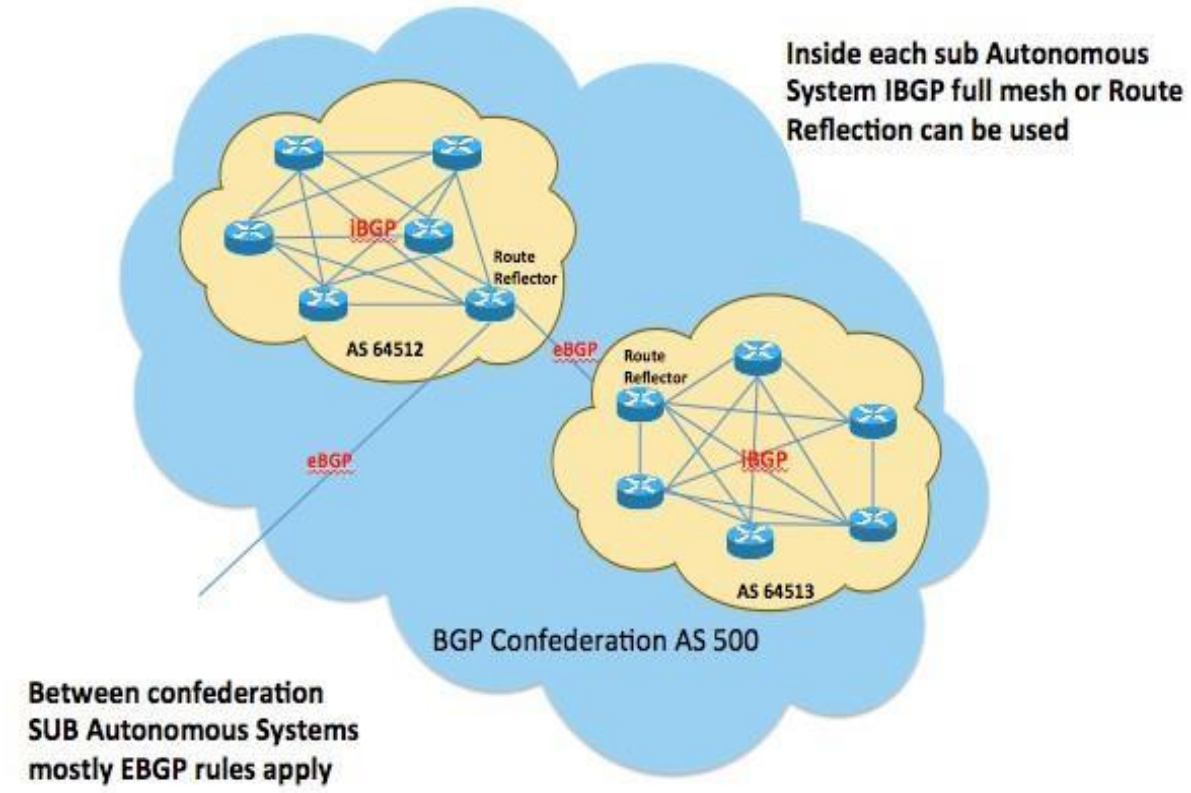
IBGP is used inside an Autonomous system. In order to prevent routing loop, iBGP requires BGP nodes to have full mesh interconnections among them.

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

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 Confederation

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

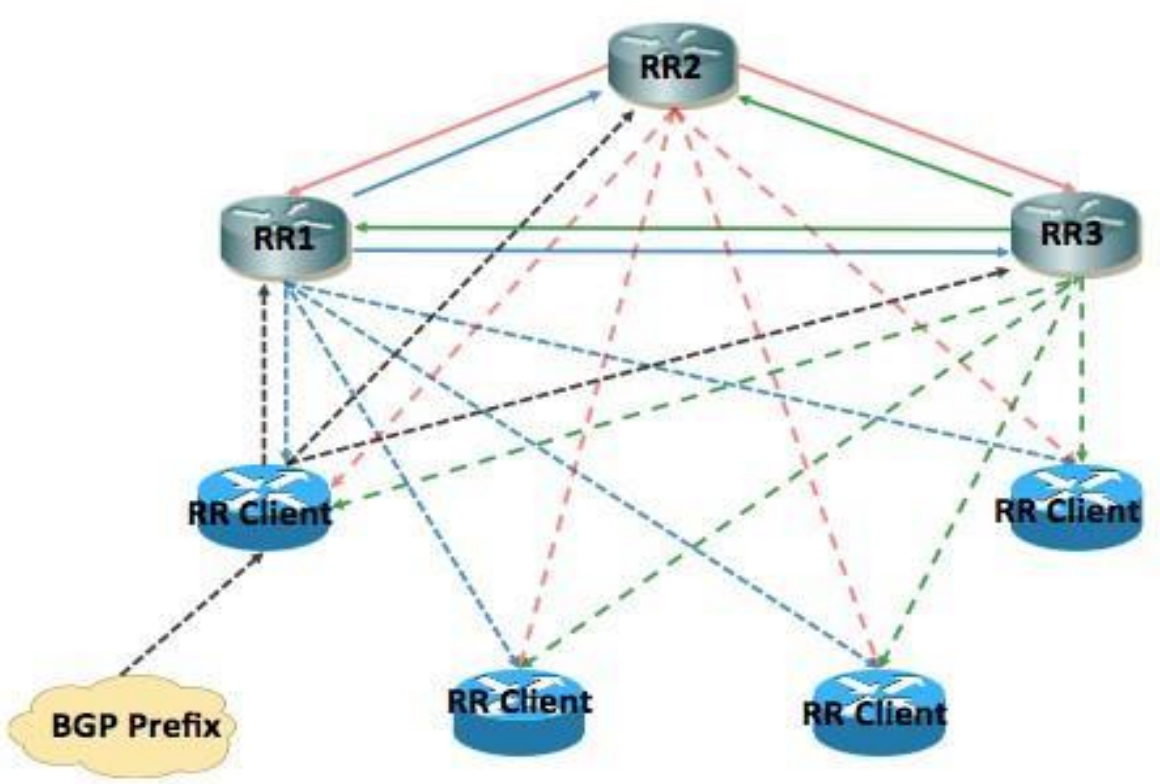
RRs and RR Clients form a cluster.

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

For the different address families, different set of Route reflectors can be used, this avoid fate sharing. For example if IPv4 RR is attacked, VPN customers may not be impacted if different sets of RR is used. RR can be a client of another RR. Hierarchical RR deployment is possible for very large scale

scenarios. It is like an EIGRP summarization at every tier.

Don’t forget that RR hides the path, you need additional mechanisms to advertise all or selected paths to the Route reflector clients if it is needed.



Prefix p/24 is sent from the RR client to the 3 of the RRs. Route reflector has full mesh among them. They send the prefixes to each other.

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

Cluster ID is used for the loop prevention in IBGP Route Reflection by the RR. RR Clients don't know which cluster they belong to.

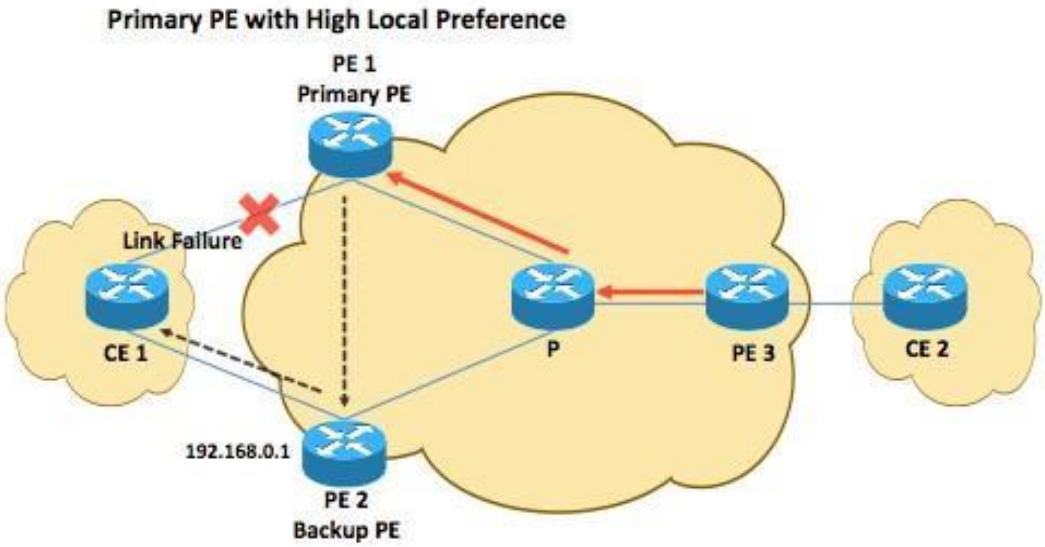
Should you use different or same Cluster IDs if you have more than one RR in BGP design?

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 !. Exception? Yes, for example Seamless/Unified MPLS scenario

BGP Best External:

It is used in Active Standby BGP link scenario to advertise backup path to the other bgp nodes.



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.

In the above picture, instead of P router, if we would have BGP Route reflector then PE3 wouldn't receive the backup path.

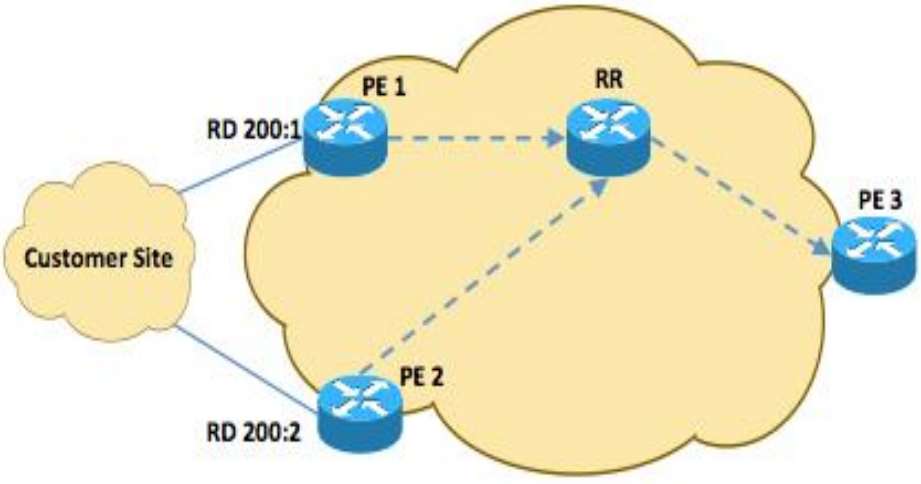
Because Route reflectors hide the paths, selects best path and advertise only the best path to the Route reflector clients.

But if you want to send additional paths for multi pathing or fast reroute purpose then below are the approaches.

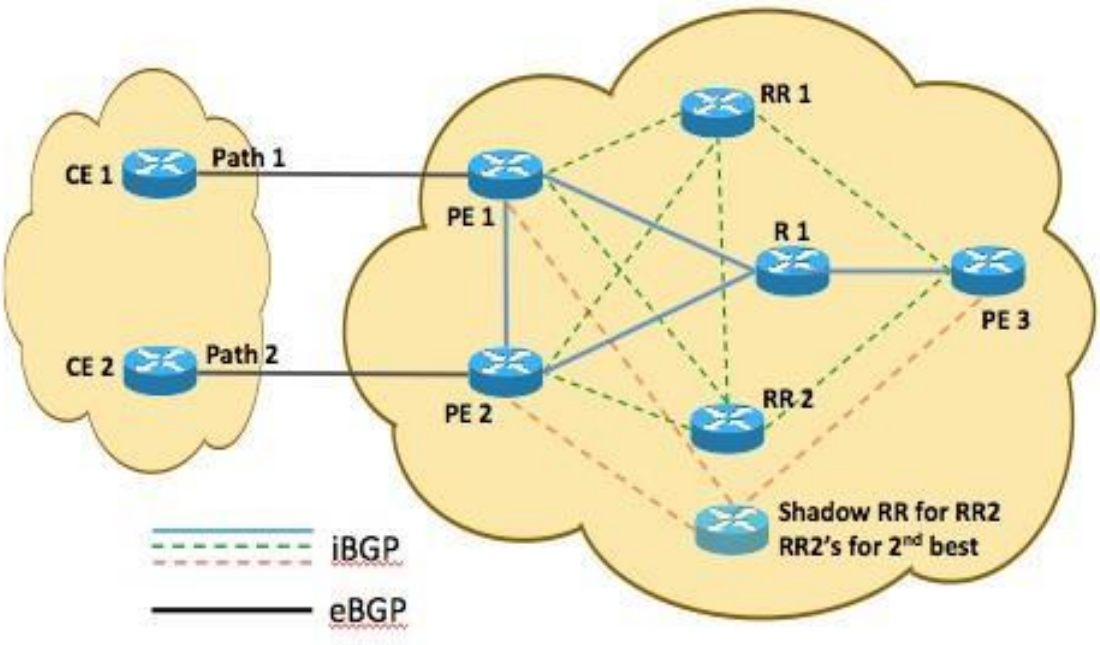
- Unique RD per VRF per PE. Unique Route Distinguisher is used per VRF per PE. No need, Add-Path, Shadow RRs, Diverse Paths. But only applicable in MPLS VPNs.

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



- 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

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 BGP, if multiple paths are sent over the same BGP session,

last one is kept since it seems as the latest update.

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.

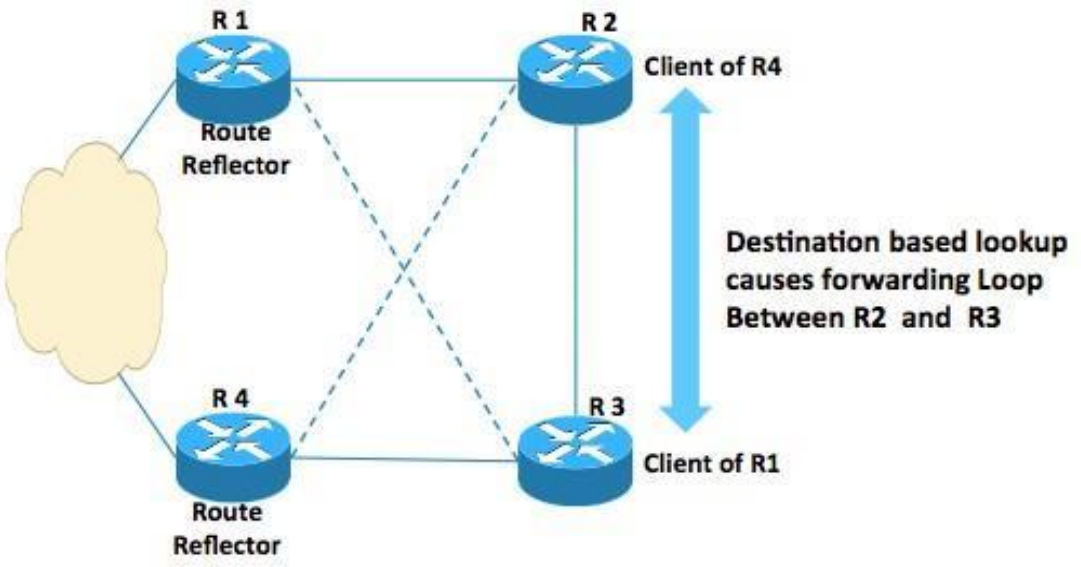
In this solution PEs send all the RT values to both Route Reflector Groups. They receive and process all the prefixes but based on odd/even ownership they filter the unwanted ones. But processing the prefixes which will be filtered anyway is not efficient way.

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.

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

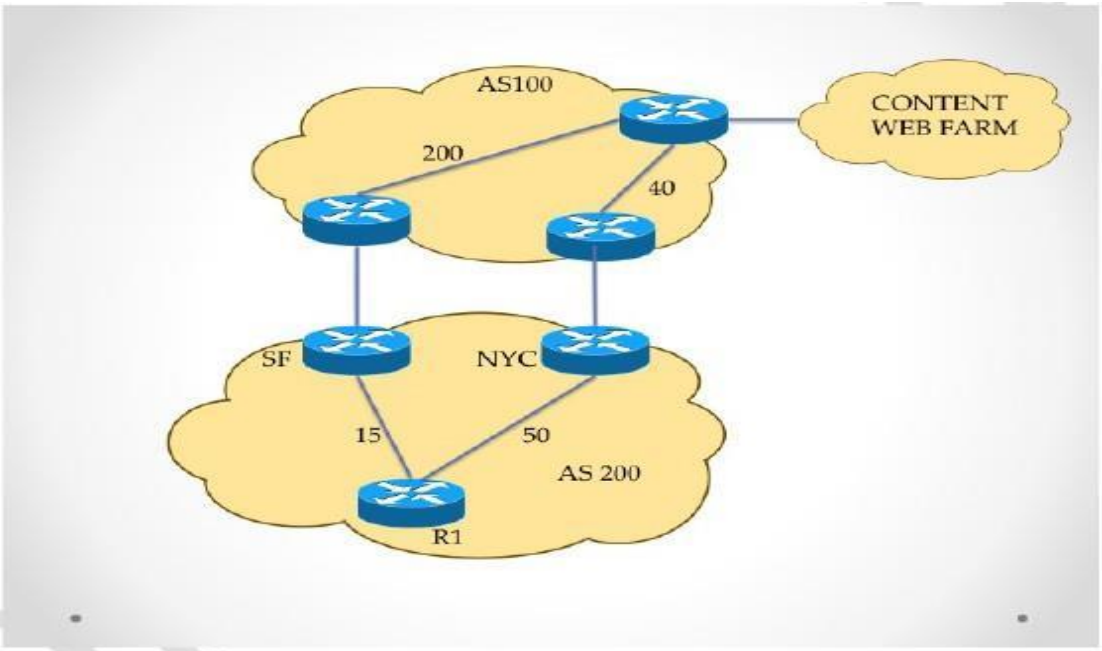
R2 will do the lookup for prefix A and send it to R2 , R3 will send it back. Packet will loop between R2 and R3.

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

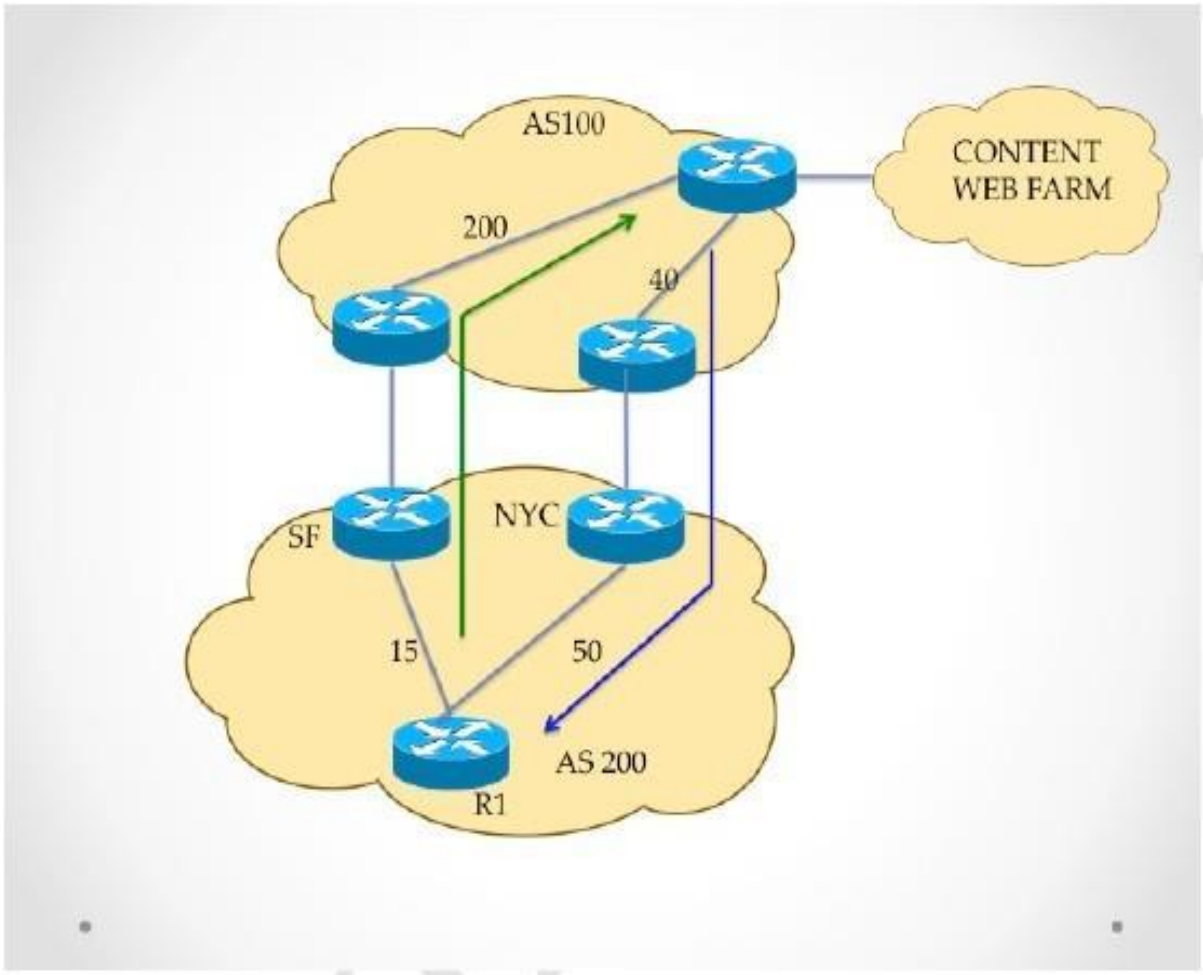
AS200 is a customer service provider of AS100 transit Service provider. Customers of AS200 is trying to reach a web page located behind AS100. AS200 is not implementing any special BGP policy. What would be the ingress and egress traffic for AS 200 ?



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.

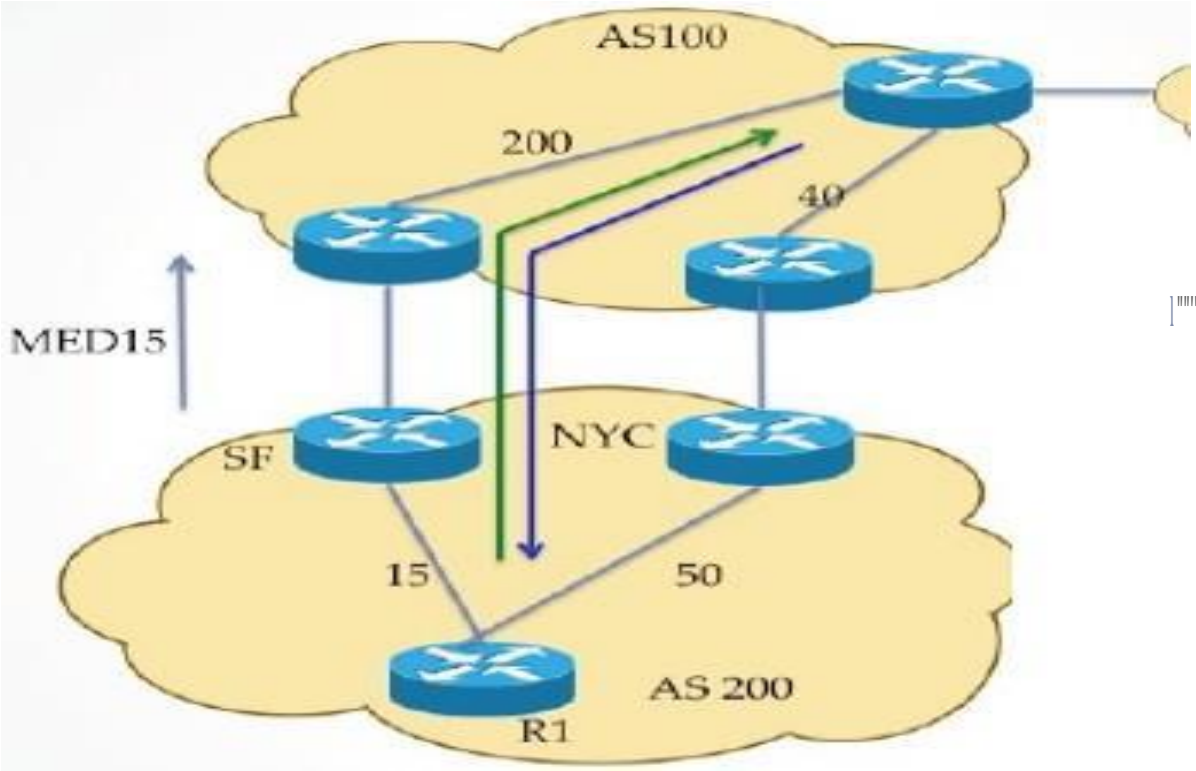
Eggress traffic from AS 200 is the green arrow in the below diagram, since SF path is shorter IGP distance. Ingress traffic to AS200 from AS 100 is the blue arrow, since NYC connection from AS100 shorter IGP distance (40 vs. 200)



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 ?

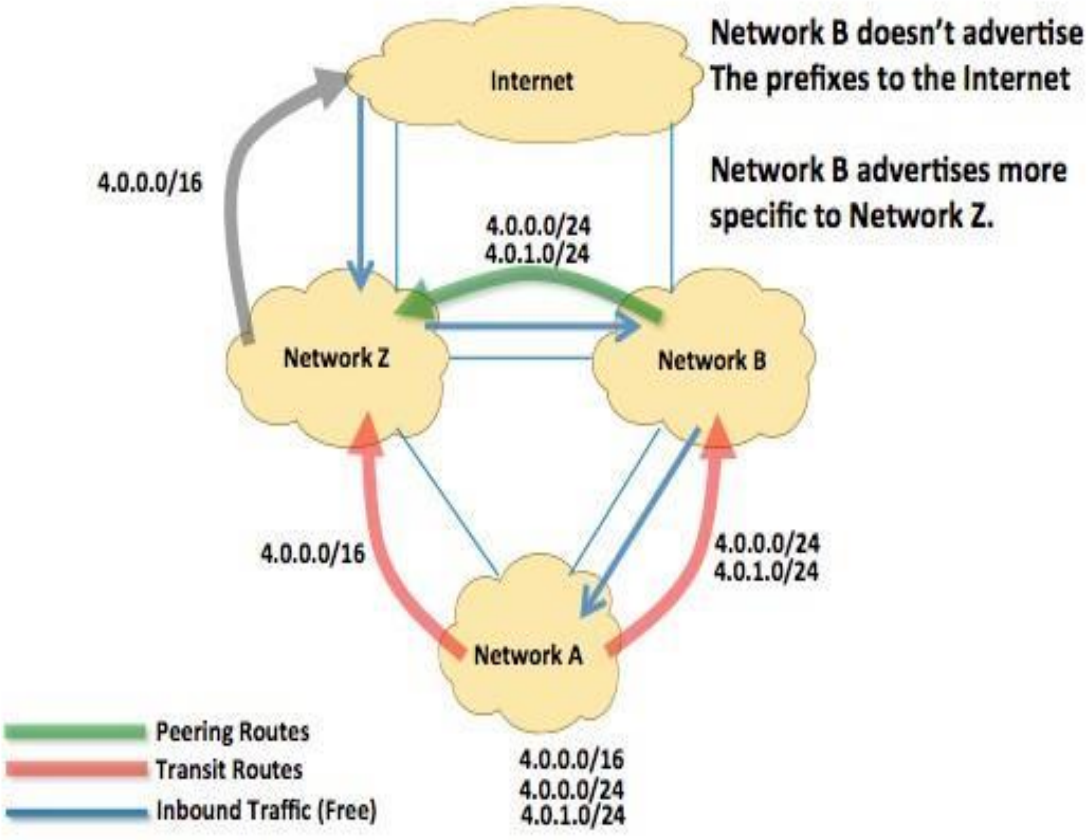
Customer AS200 should force AS100 for cold potato routing. By forcing for cold potato routing ,AS 100 has to carry the Web content traffic to the closest exit point to AS200, which is San Francisco.

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.

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://blog.ine.com/2010/11/22/understanding-bgp-convergence/#9

http://orhanergun.net/2015/05/bgp-pic-prefix-independent-convergence/

http://orhanergun.net/2015/03/bgp-design-quiz/

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/

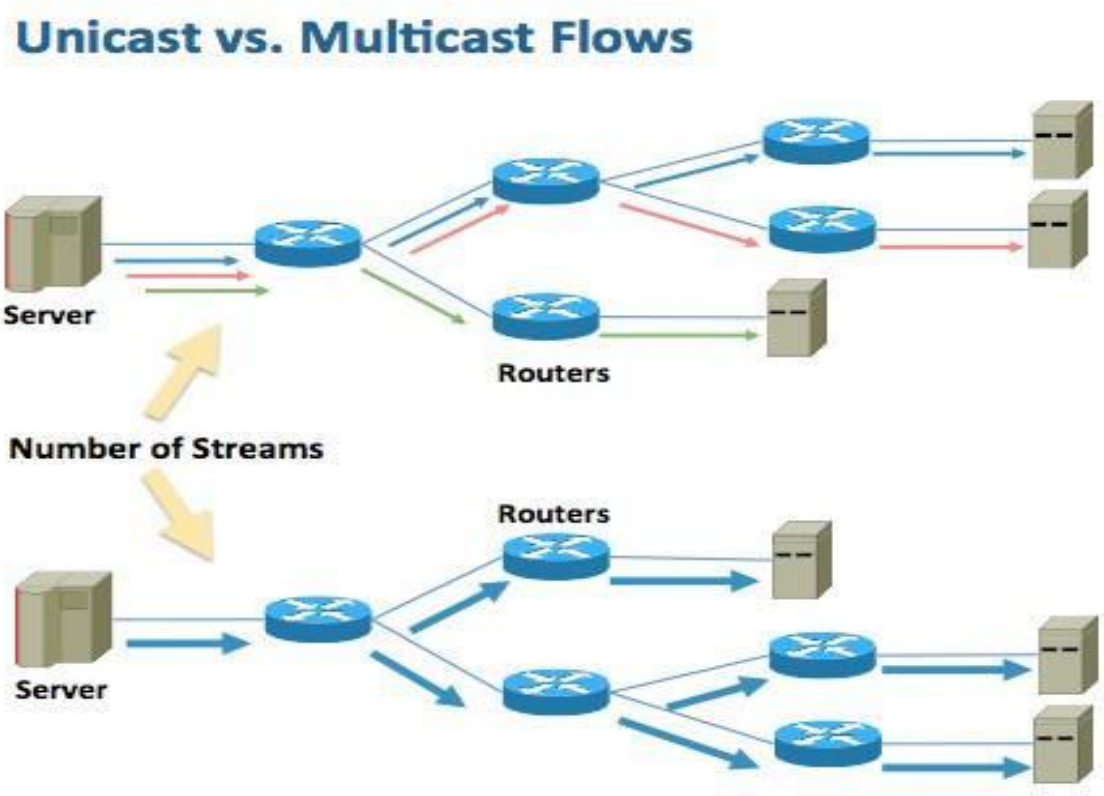
http://packetpushers.net/bgp-rr-design-part-2/

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

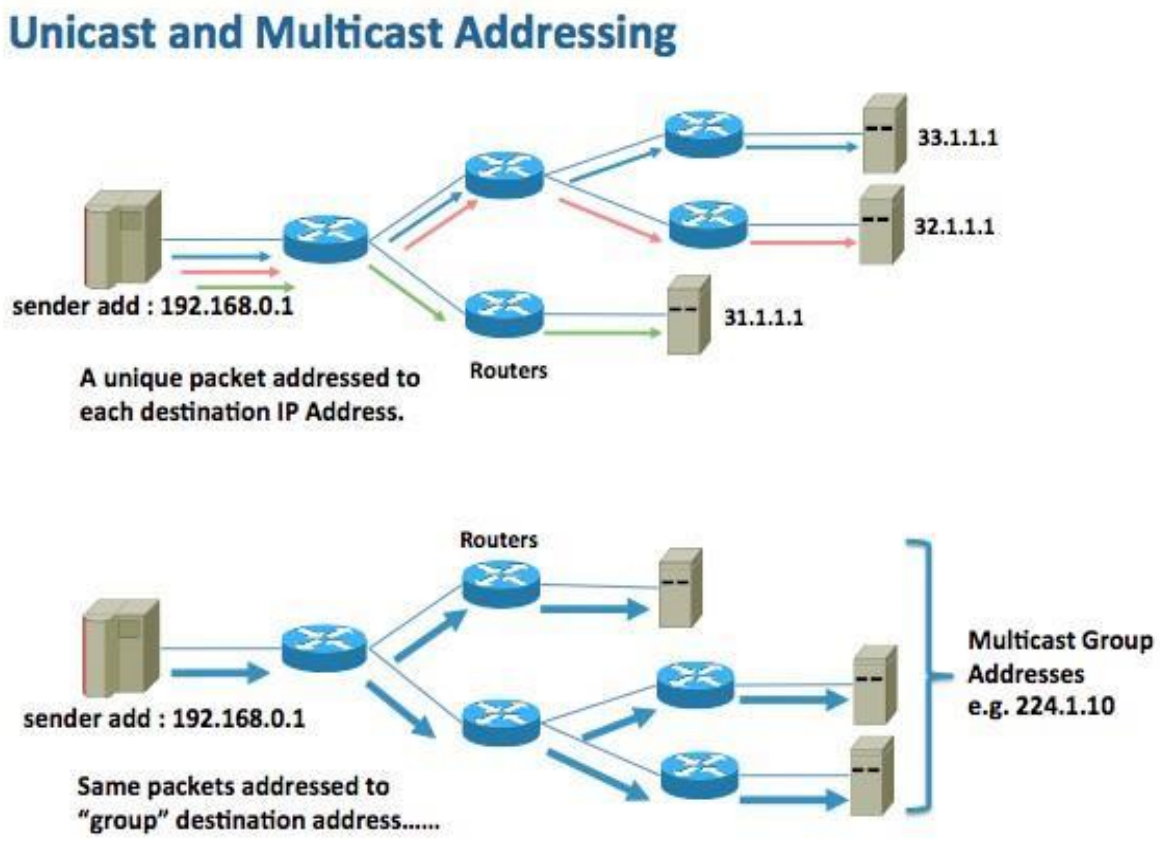
http://packetpushers.net/bgp-path-huntingexploration/



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.



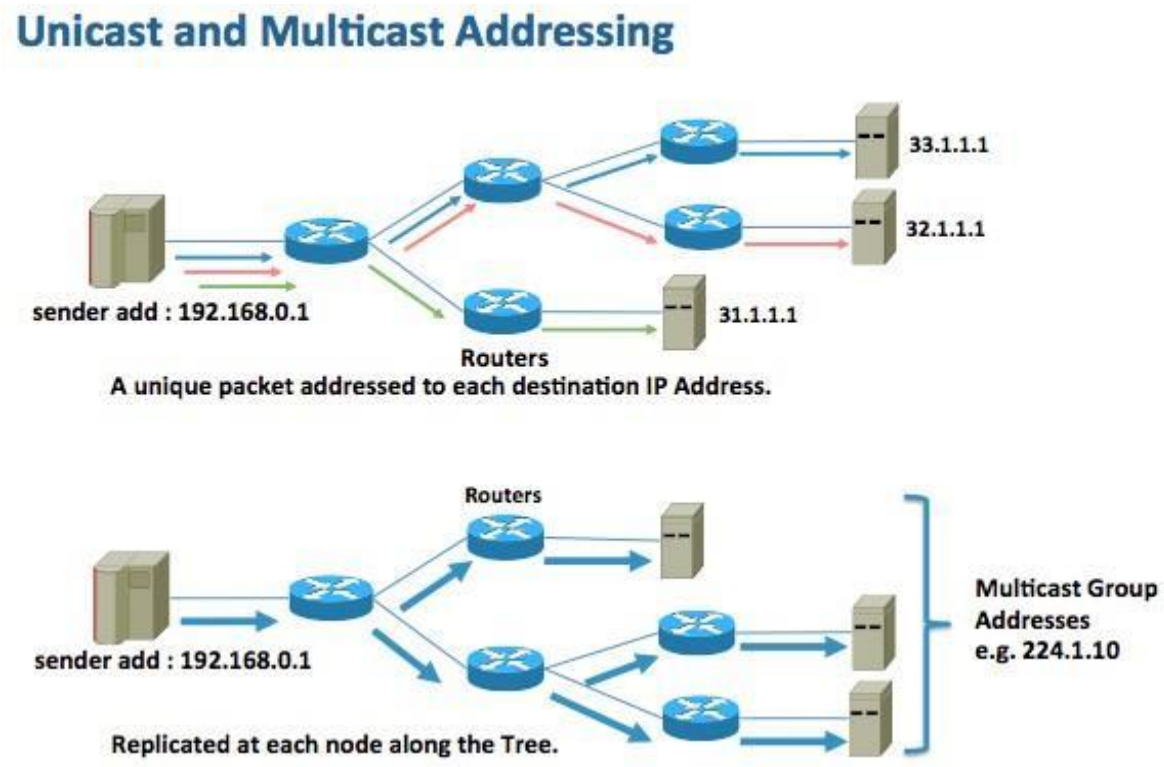
Source addresses in Multicast always unicast address. Multicast address is a class D address range. 224/4.

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.

Source is root of the tree in shortest path tree.

Rendezvous Point is the root of the Shared 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

.

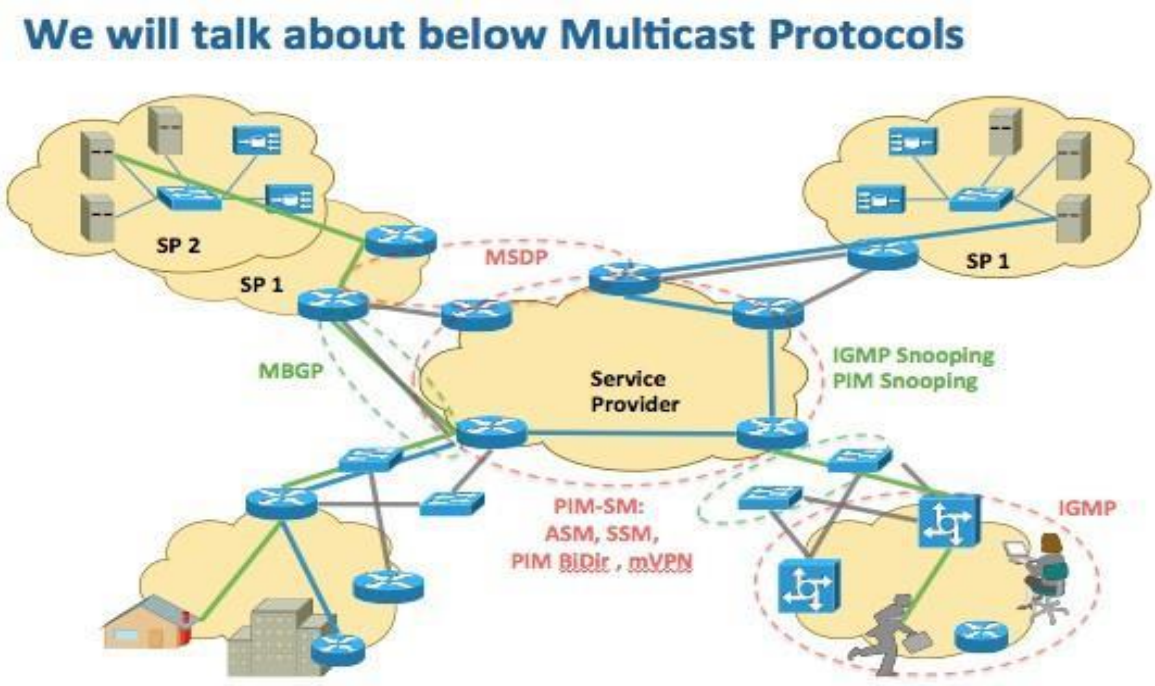
238.1.1.1

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



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).

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.

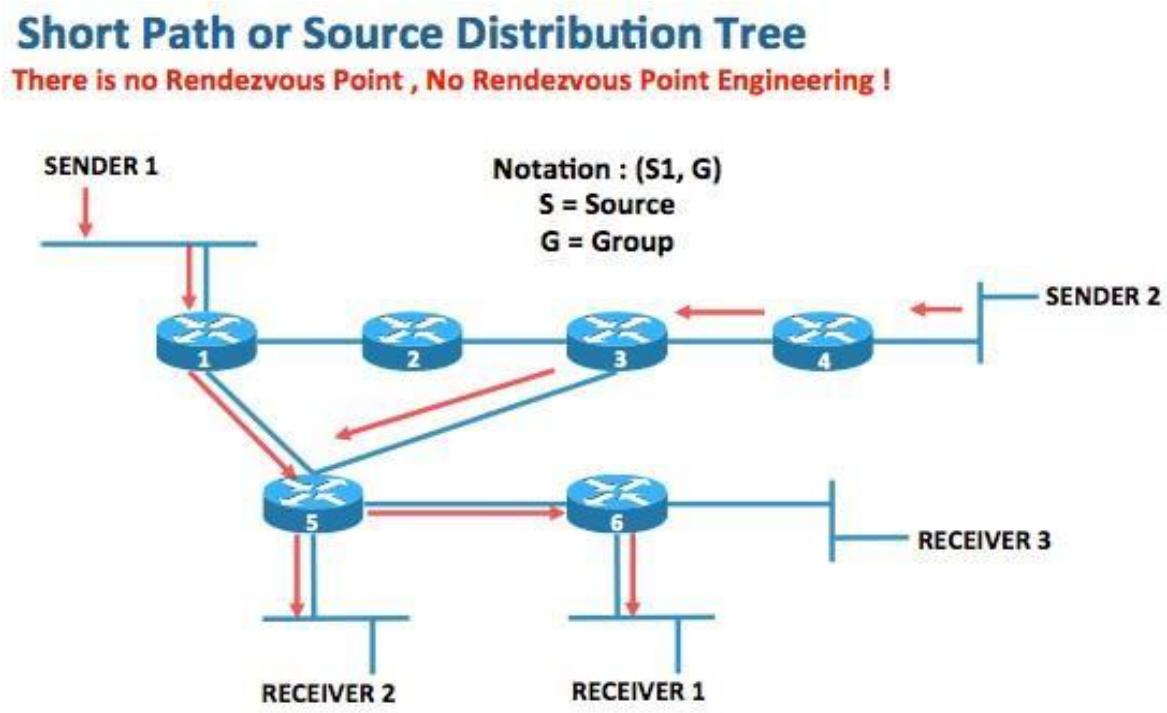
Switches with IGMPv2 has to track every multicast packet to see catch the control plane traffic. Every multicast data packet is inspected with IGMPv2, since there is no IGMPv2 router multicast group address. That's why in IGMPv2, IGMP snooping should be enabled on the hardware otherwise huge performance 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.

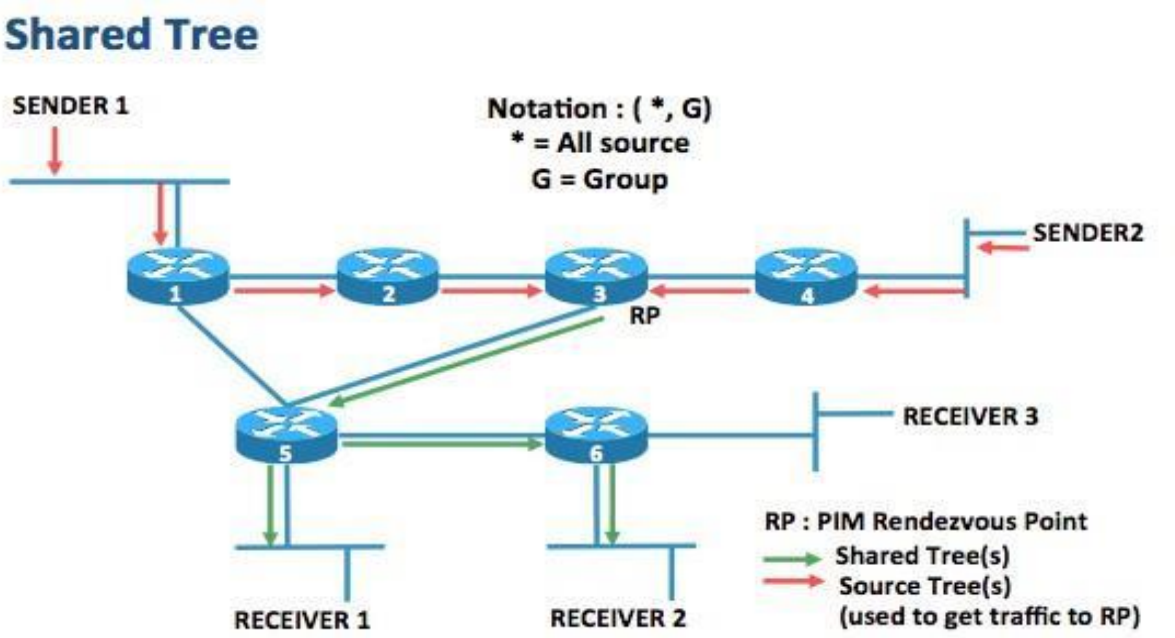
Multicast Distribution Trees

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.



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.



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 is a multicast routing protocol. Two PIM mode, PIM sparse and 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 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.

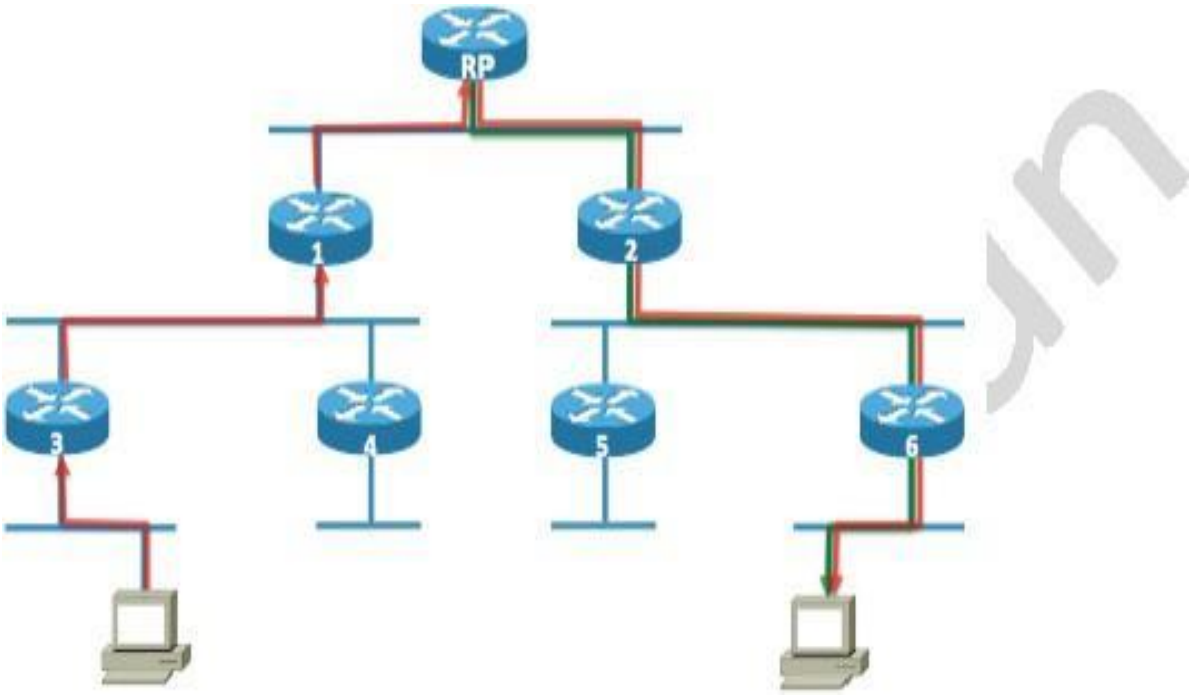
Rendezvous point information can be configured as Static on the every router, it can be learned through Auto-RP which is Cisco specific protocol or BSR which is IETF standard.

There is no Auto-RP anymore in IPv6 Multicast.

PIM SSM 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. Require IGMPv3 at the source or IGMPv2 to v3 mapping on the first hop routers. 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. 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

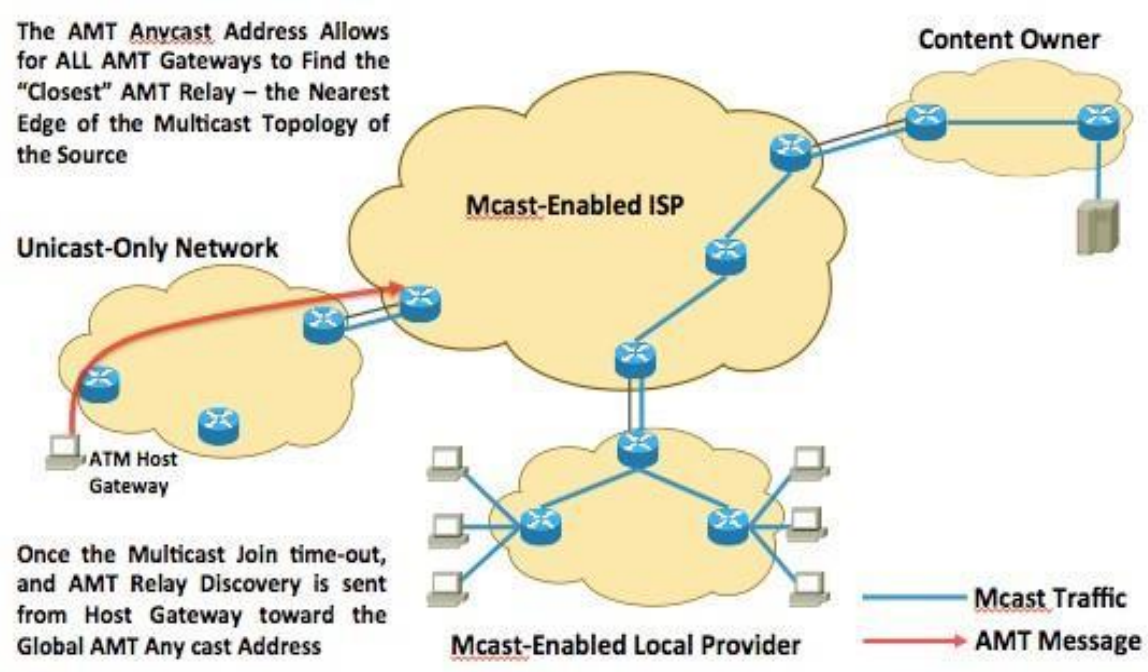
Multicase Case Study

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?

Solution can be provided with Automatic Multicast Tunneling (AMT) RFC 7450.



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 T

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.

Police unwanted traffic flows as close to their sources as possible.

Enable queuing policies at every node where the potential for congestion exist.

Application	Layer 3 Classification			Layer 2 CoSMPLS EXP
	IPP	P	DSCP	
P Routing	6	CS6	48	6
Voice	5	EF	46	5
Interactive Video	4	AF41	34	4
StreamingDVideo	4	CS4	32	4
LocallyDDefined MissionDCritical Data	3	YYYYYY	25	3
CallDSignals	3	AF31/CS3	26/24	3
Transactional Data	2	AF21	18	2
Network Management	2	CS21	16	2
Bulk Data	1	AF11	10	1
Scavenger	1	CS1	8	1
Best Effort	0	0	0	0

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

Streaming Video:

When addressing the QoS needs of Streaming Video traffic, the following guidelines are recommended:

- 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.

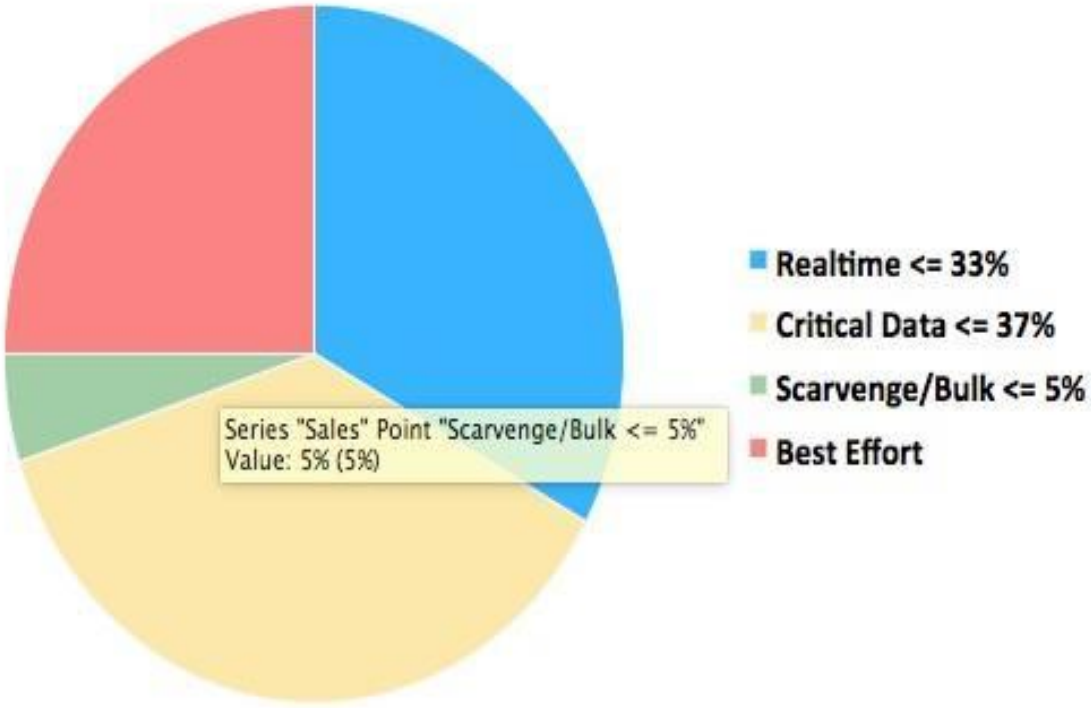
For example, with Transactional Data client-server applications such as SAP, PeopleSoft.

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).

Real time, Best Effort ,Critical Data and Scavenger Queuing Rule – 4 class QoS deployment


Realtime <= 33%
Critical Data <= 37%
Scarvenge/Bulk <= 5%
Best Effort
Series "Sales" Point "Scarvenge/Bulk <= 5%"
Value: 5% (5%)

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

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.

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 D MPLS Traffic Engineering D Seamless MPLS

Layer 2 MPLS VPN

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)

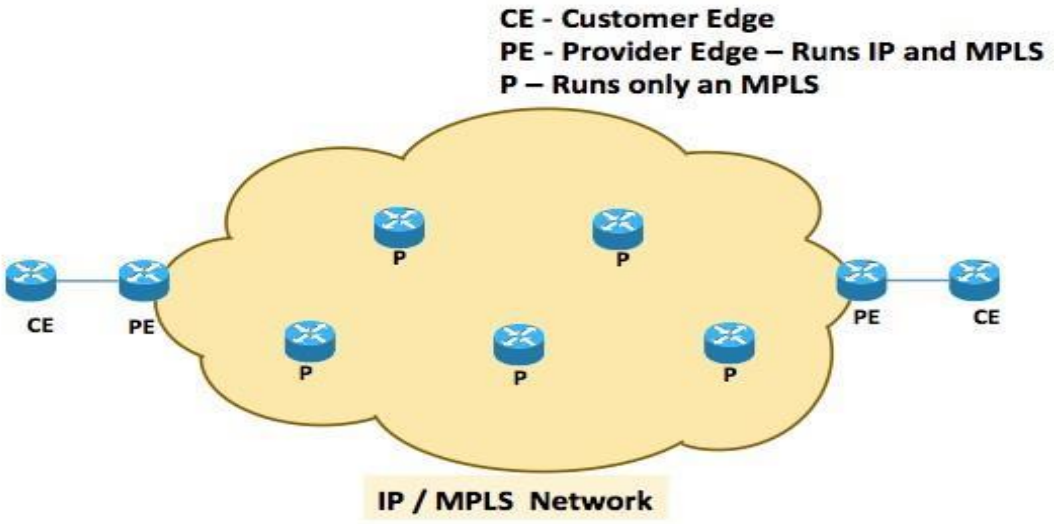
MPLS layer 2 VPN can be point to point which is called VPWS (Virtual Private Wire Service), multi point to multi point which is called VPLS (Virtual Private Lan service).

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 provider network.

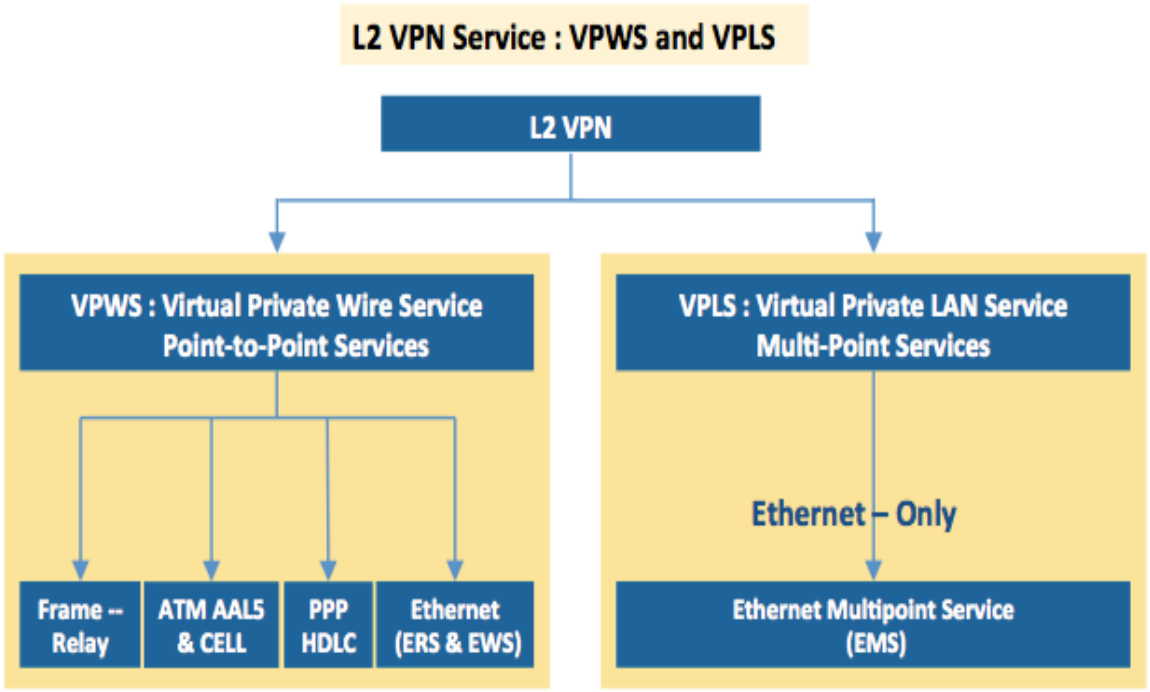
PE device looks at the incoming frame or packet and identify which egress PE device is used for transport. Second lookup is made to determine the egress interface on the egress device.

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.

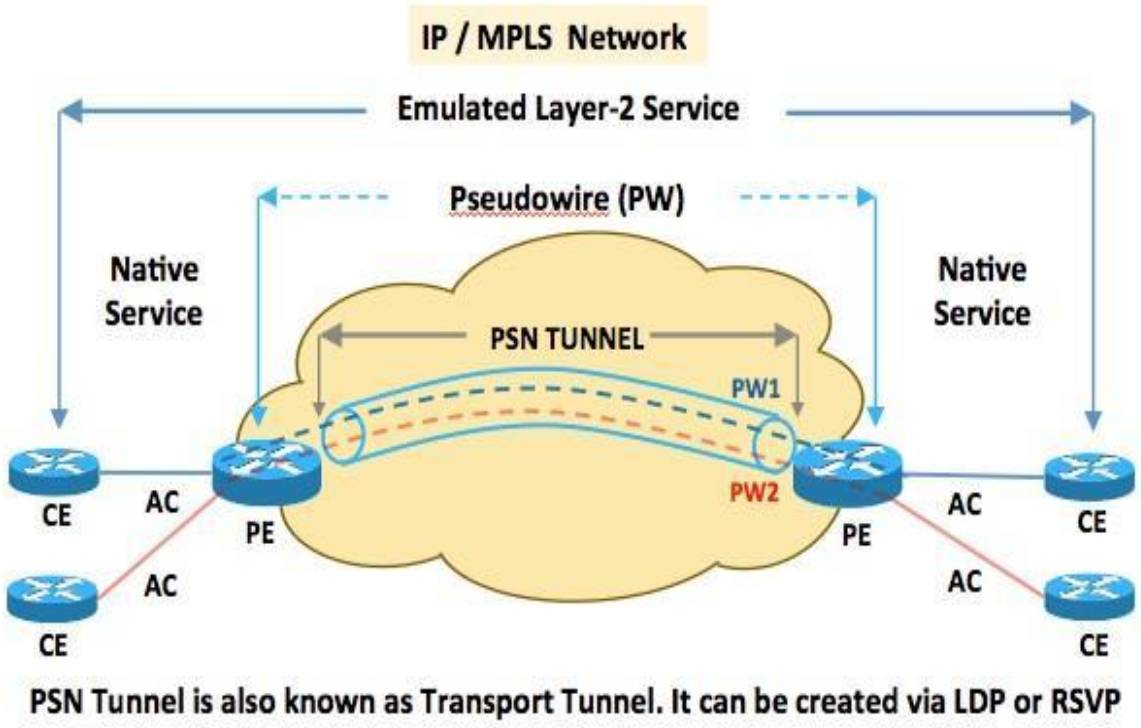
In Martini method; both transport and VC (Virtual Circuit) label is sent (signaled) via LDP (Label Distribution Protocol).Targeted LDP session is created between PEs.

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.



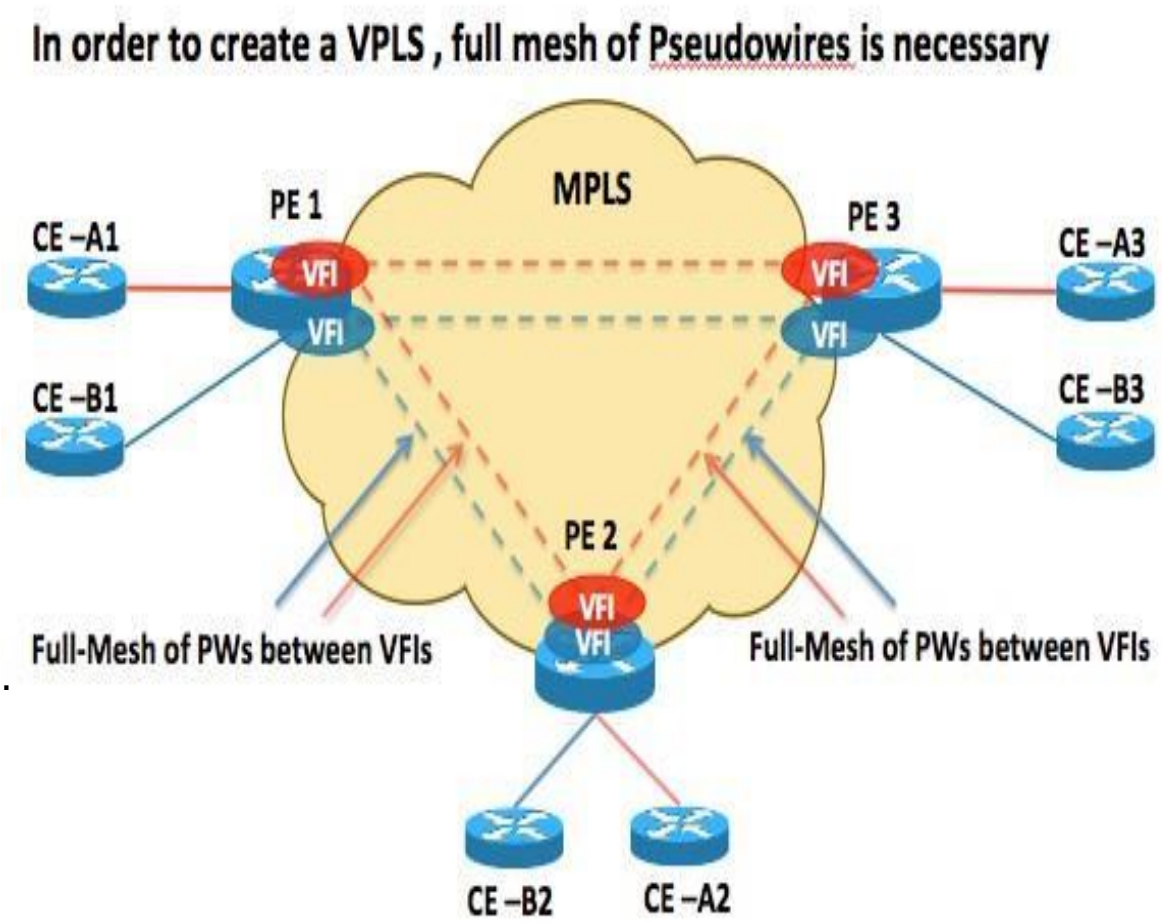
VPWS can carry almost all layer 2 payloads as can be seen from the above diagram. VPLS can carry Ethernet only. Although there is an attempt in the IETF for the other layer2 payloads over VPLS as well.

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.



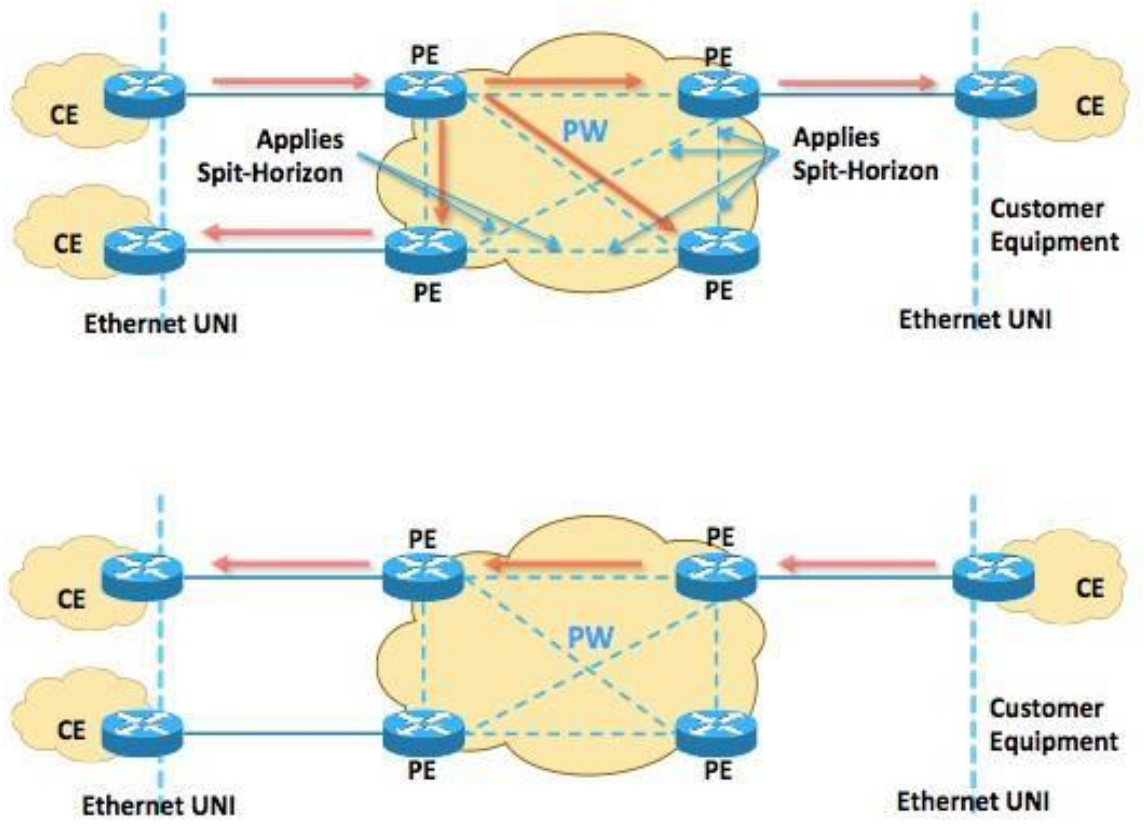
There is only one egress point for the VPWS service which is the other end of the pseudo wire, thus PE device doesn't have to keep Mac Address to PW binding. PE device doesn't have to learn MAC address of the customer. This provides scalability.

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



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.

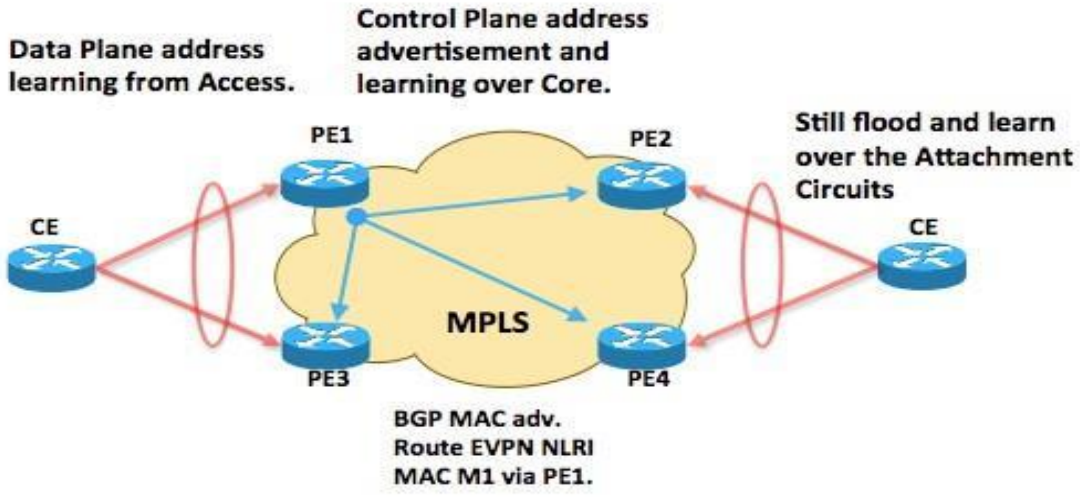
EVPN

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.

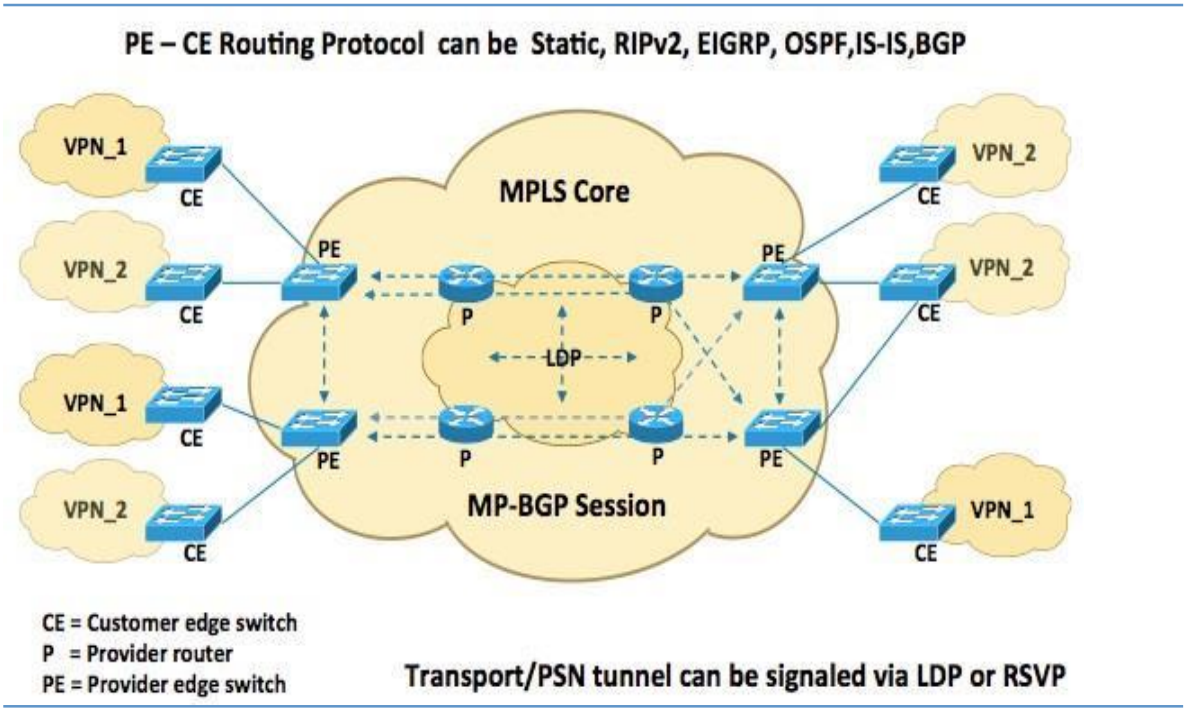
In VPLS, active active flow based load balancing is not possible.

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.



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.



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 provide 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.

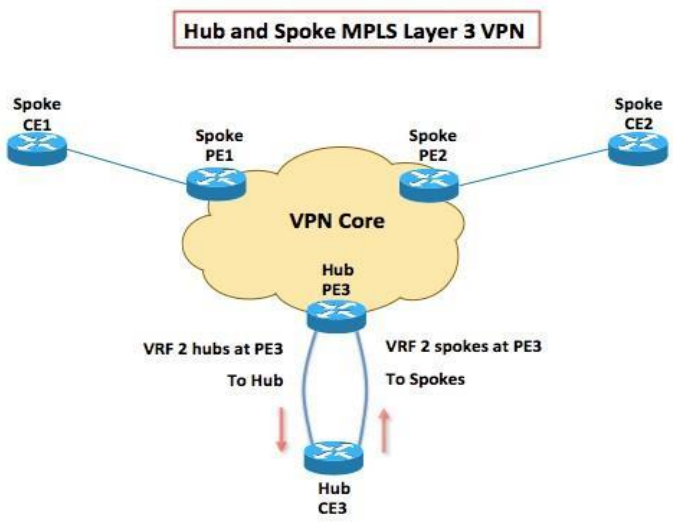
MP-BGP session is created between PE devices only. P devices don't run BGP in MPLS environment. This is known as BGP Free Core design.

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 Layer 3 VPN

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.

Below chart shows the selection criteria for choosing Single vs. Dual Providers.

Resiliency Drivers vs. Simplicity

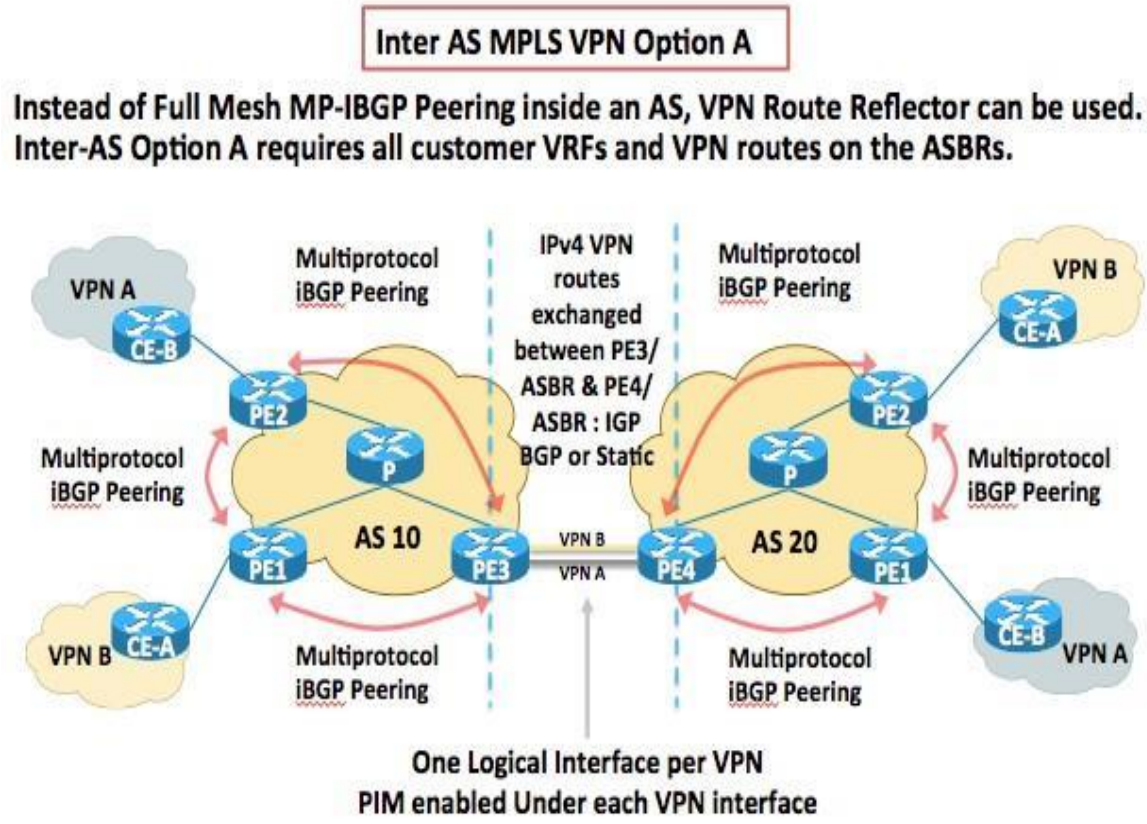
	Single Provider	Dual Providers
Pro:	• Common QoS support model • Only one vender to "tune" • Reduce number of circuits • Overall simpler design	• Separate fault domains • Diverse product offerings • Leverage vendors for better pricing
Con:	• Carrier failure could be catastrophic • Do not have another carrier to leverage	• Dual Vender Management overhead • Increase Bandwidth "Paying for bandwidth twice" • Increased overall design complexity • May be reduced to "common denominator "between carriers"

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 customer 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 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.

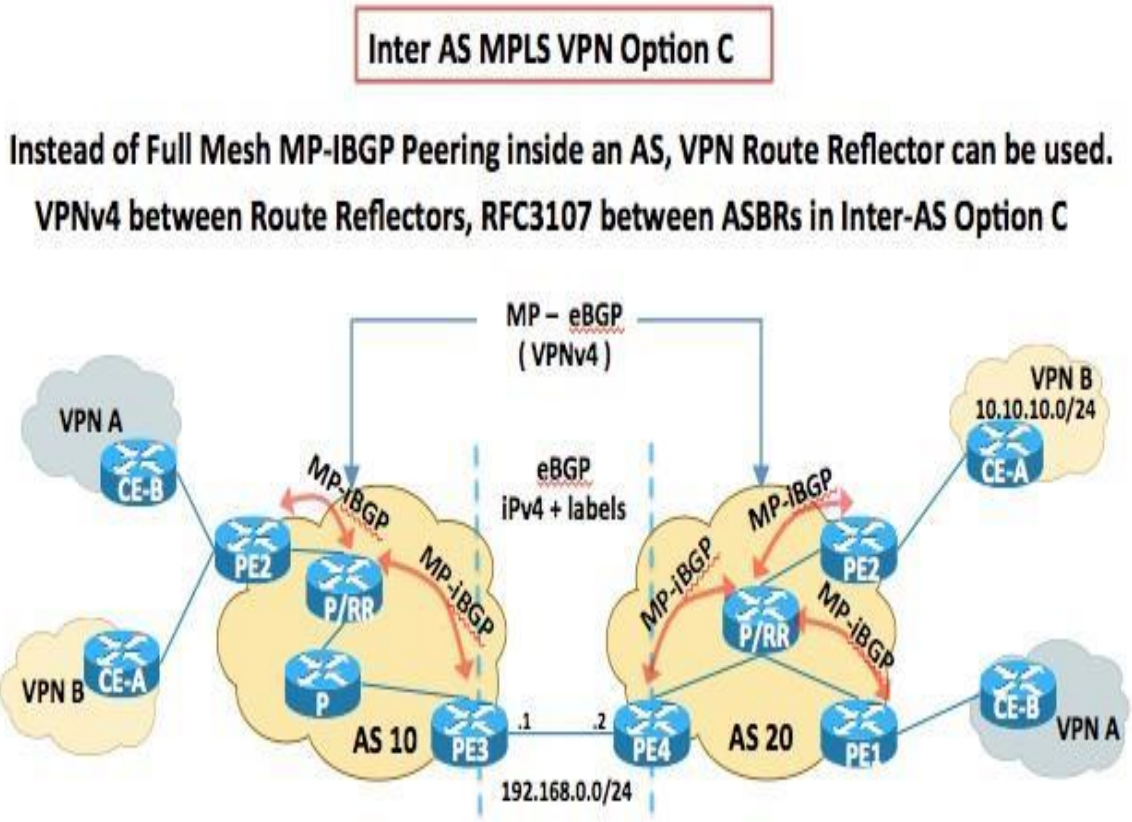
For each customer VPN, one logical or physical link is setup. Over the link, any routing protocol can run. But in order to carry end to end customer routing

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.



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.

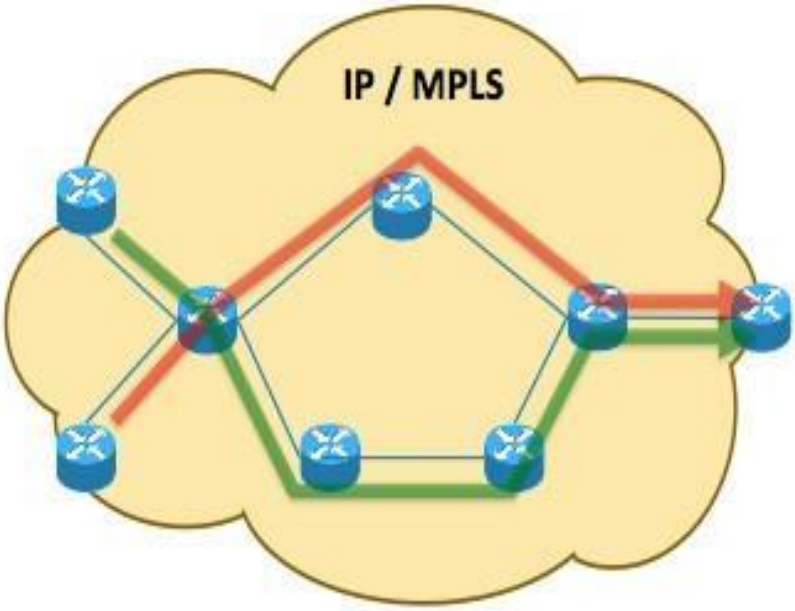
Whenever BGP next hop changes, new VPN label is assigned by the next hop device. In the Option B case, if BGP next hop self is enabled on the ASBR, ASBR allocates a new

IP and MPLS is a destination based routing. With MPLS Traffic Engineering, source routing is achieved.

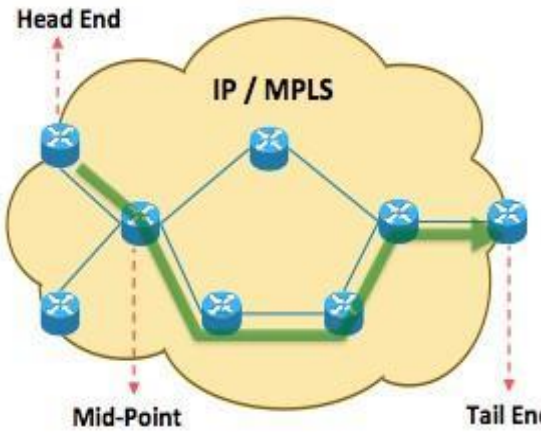
MPLS Traffic Engineering allows explicit routing. From Head-End, entire path can be calculated and signaled.
Below is a classical fish diagram. If IP/MPLS would be enabled only, destination based shortest path routing would force top path to be used.

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

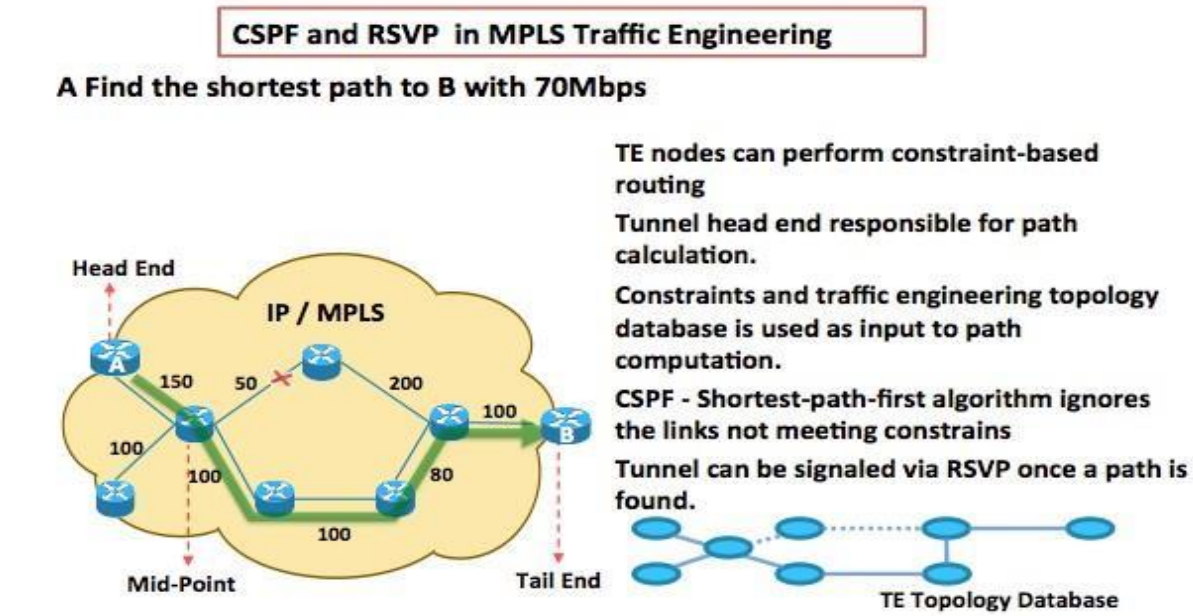


Link Information Distribution
- -- ISIS –TE
- -- OSPF-TE

Path Calculation (CSPF)

Path Setup (RSVP-TE)

Forwarding Traffic down Tunnel
- -- Auto-route announce
- -- Static route
- -- CBTS
- -- PBR
- -- Forwarding Adjacency
- -- Pseudowire Tunnel selection



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 CSCP, 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).

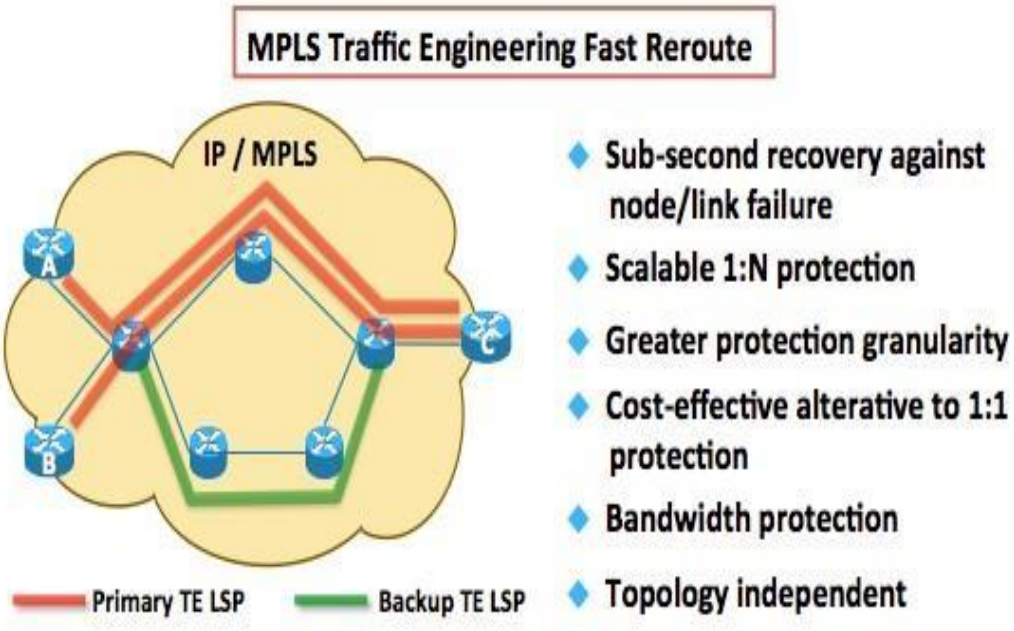
OSPF and IS-IS has been extended to carry additional link state attribute for better traffic engineer with mpls. In the below picture, additional link state information is shown. This information is not kept in LSDB but they are kept in TED (Traffic Engineering Database).

Link Attributes for MPLS Traffic Engineering Database	
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 using offDline path computation	

Link Attributes for MPLS Traffic Engineering Database

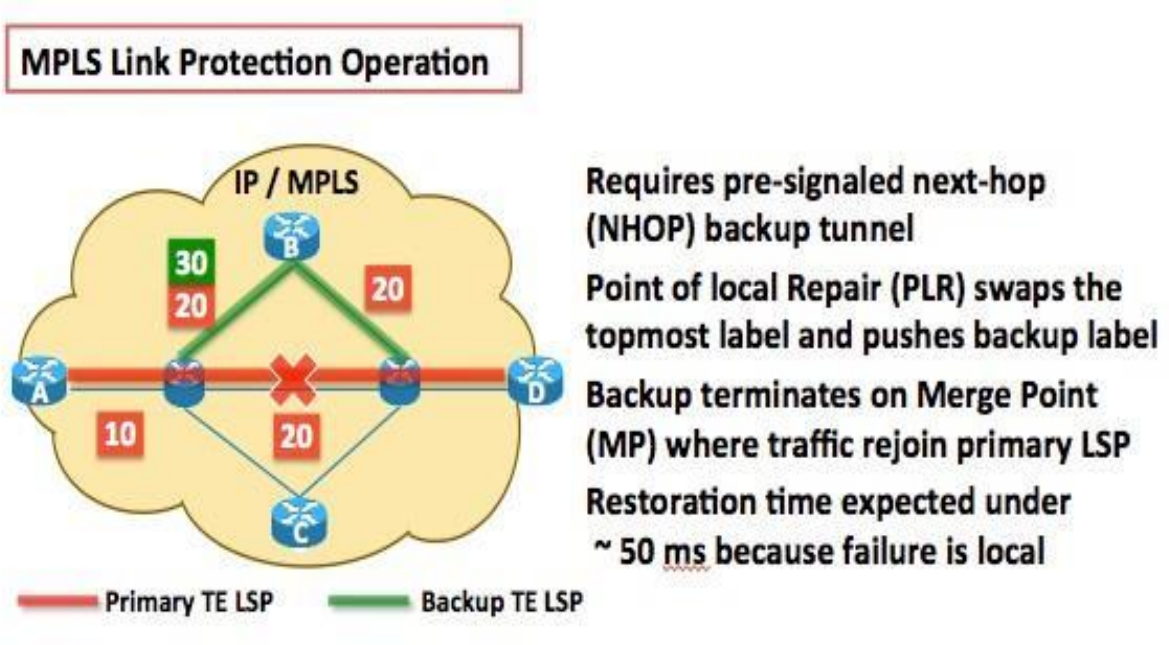
MPLS Traffic Engineering Fast Reroute

Fast reroute is a local protection mechanism.



The router reacts to a failure is called Point of Local Repair.

Whenever failure happens, MPLS Traffic Engineering Fast reroute is a data plane protection mechanism. 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.



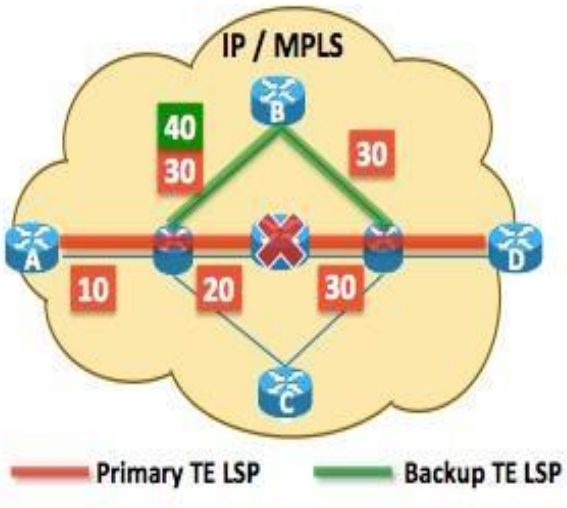
MPLS Traffic Engineering Fast Reroute Node Protection

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 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 ?

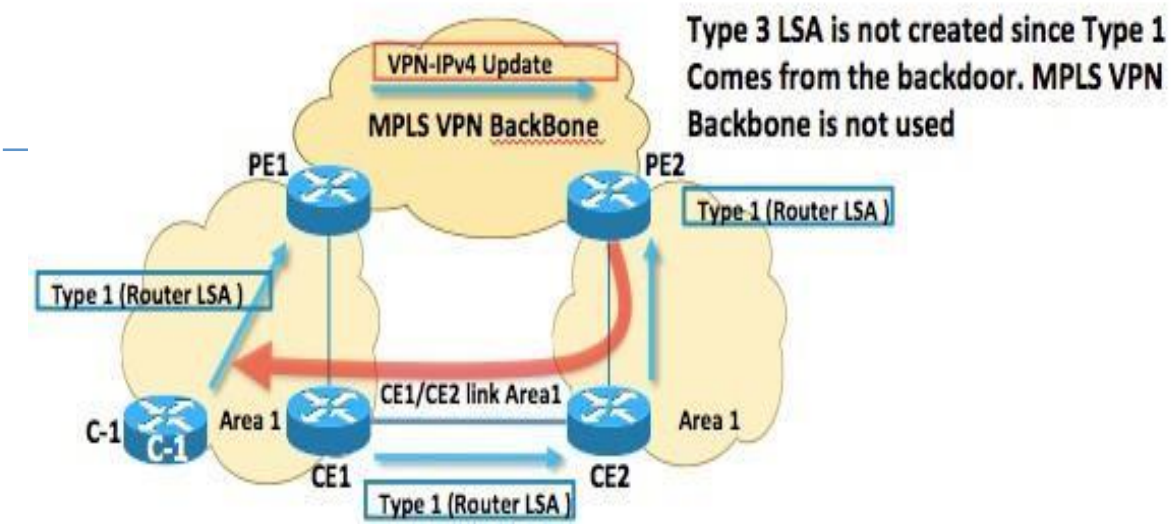
What might the possible problems ? How can
those be avoided ?

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.

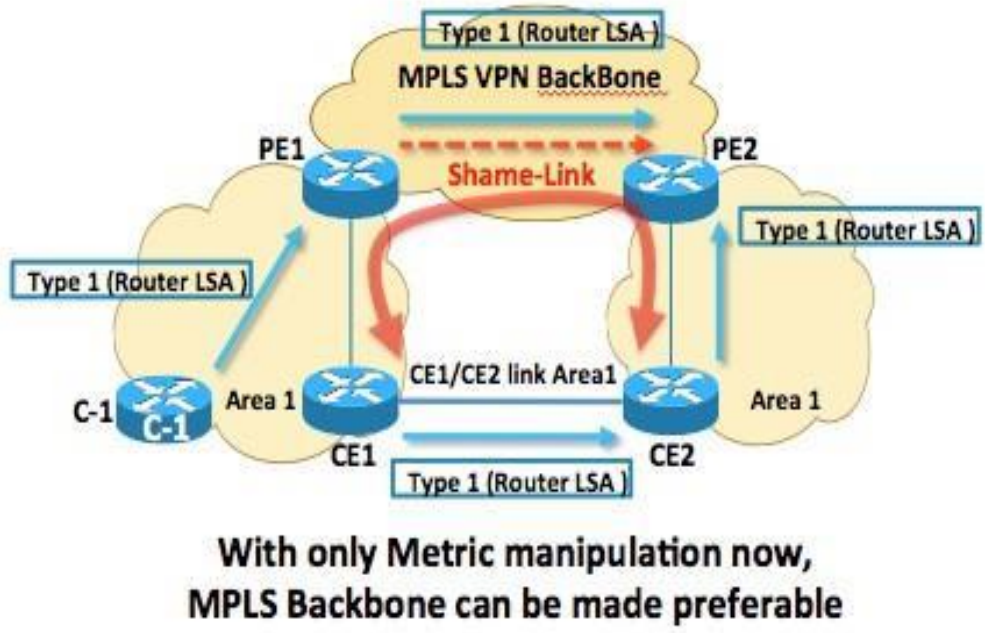
If domain ID is not set exclusively, then by default Process ID is used for domain ID.



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 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

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.

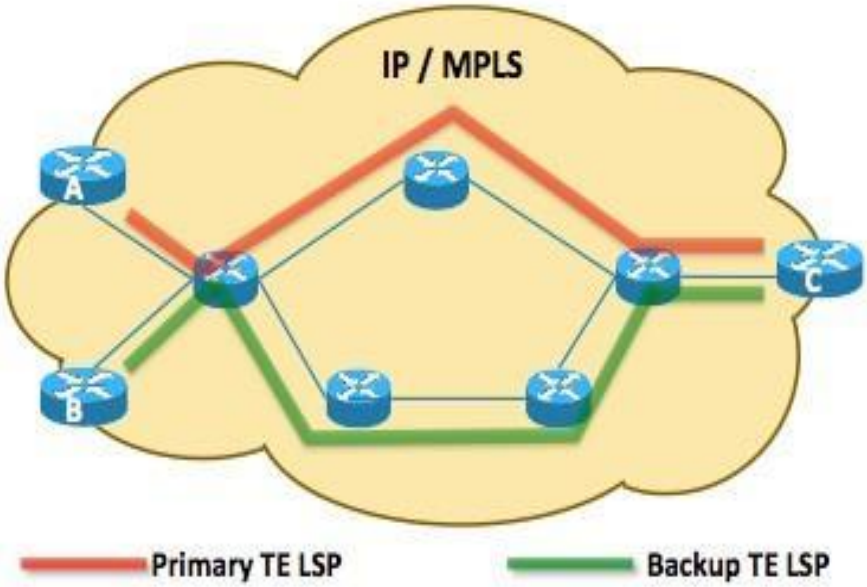
Fast reroute backup LSP can protect multiple primary LSP, thus in the MPLS Traffic Engineering chapter, it is showed as 1:N protection.

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 – 3

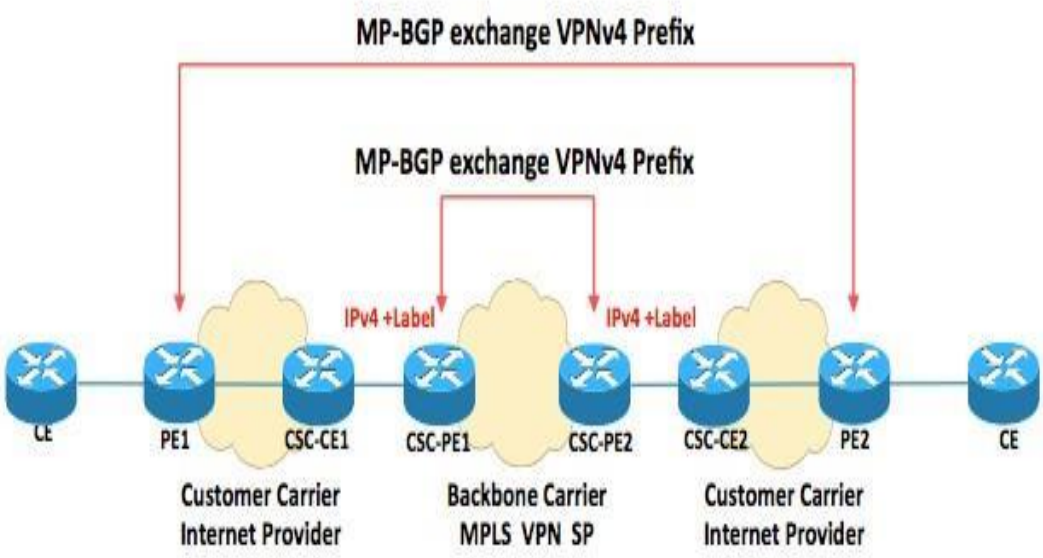
Smallnet is an ISP which provides Broadband and business internet to their customer. Biggercom is a transit service provider of Smallnet which provides layer 3 IP connectivity between Smallnet sites.

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 connectity , doesn't carry more than 1000 prefixes of Smallnet, and this is achieved through Carrier Supporting Carrier architecture.



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 – 4

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.

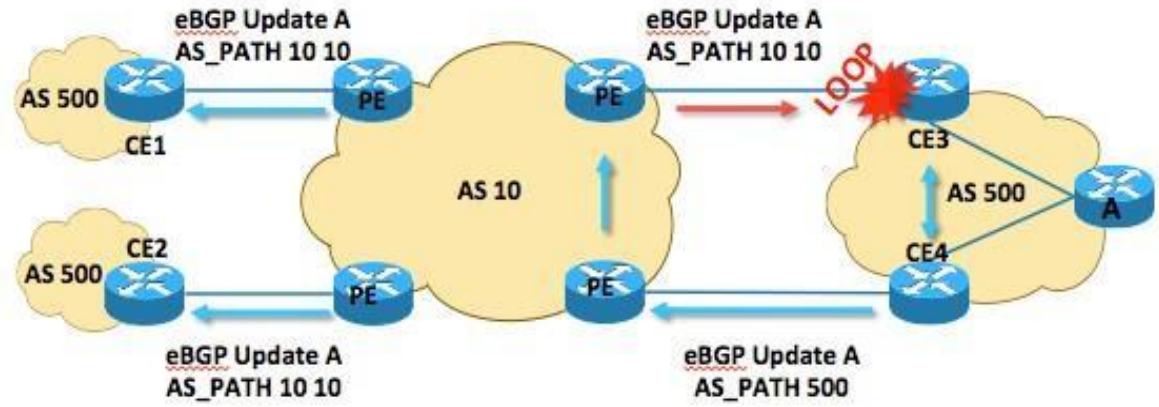
Would this solution work with BGP as a PE-CE routing protocol ? What can be done to make the solution work ?

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.

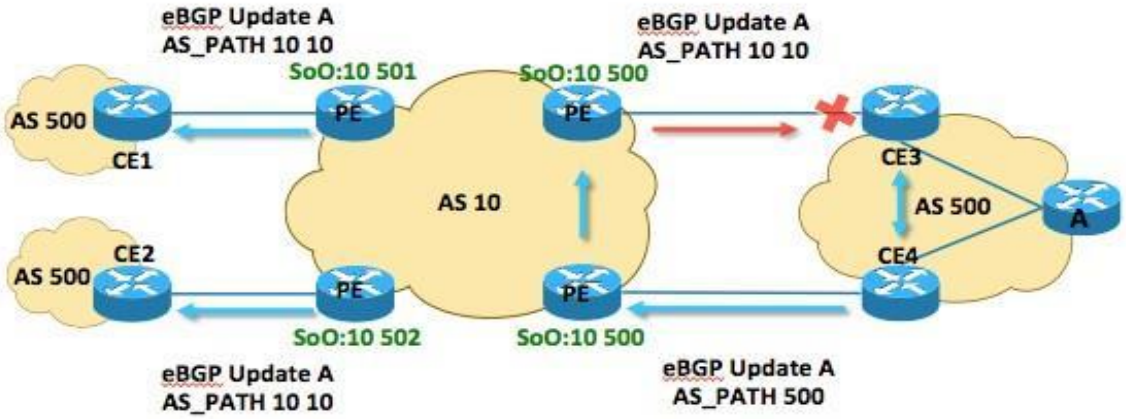


As it can be seen from the above topology, Site 3 which is the Orko's datacenter originates a BGP prefixes which is advertised to Service Provider PE device, PE3.

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.

Both PE1 and PE2 still advertise the prefixes from the Site3 (Datacenter) to their respective CEs because the configured SoO values don't match the attached SoO on the datacenter prefixes.

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
http://blog.ine.com/wp-content/uploads/2010/04/understanding-eigrp-soo-bgp-cost-community.pdf
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