# Practical Cisco Certified Design Expert v3.0

CCDE

https://t.me/learningnets

# Practical Cisco Certified Design Expert v3.0

Orhan Ergun

CCDE #2014:17
CCIE #26567

https://t.me/learningnets

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

https://t.me/learningnets

In order to get CCDE number, you need to Pass CCDE Written Pass CCDE Practical

A CCDE Written pass score required to schedule the CCDE Practical exam
CCDE Written valid for three years

https://t.me/learningnets

CCDE Lab

Passing CCDE Written results in a Cisco Certified Specialist certification + badge

https://t.me/learningnets

- CCDE Practical Exam details

  - Four scenario over 8 Hours
  - 1 hour lunch break after first 4 hours
  - Two scenarios every 4 hours
  - Each scenario can be Maximum 2 hours – If you finish a scenario earlier, remaining time won't be added to the next scenario
  - Business Challenges – Background info, emails, text and diagram
  - There is no configuration for any vendor equipment

CCDE Lab

https://t.me/learningnets

- Vendor agnostic exam, but still some technologies relevant to Cisco (erg. HSRP, GLBP, EIGRP, DMVPN,GETVPN)

- Written exam in Pearson Vue, Practical exam in Cisco CCIE Lab Locations

- Reading intensive exam. Must skim through some material in the scenario.

https://t.me/learningnets

- During the exam, pen and paper will be provided, also you will be able to use the computer to take notes.

- Unlike CCIE exams, Internet is not open, so you can’t take any help, you are alone there!

- You can go to the washroom, don’t worry ☺️

75-80%

https://t.me/learningnets

- CCIE is not required nor a prerequisite
- No configuration skills tested
- Analyze and translate business requirements into solutions
- Lab up technologies if it helps understanding them
- Knowing how things work in the field is a plus

75-80%

https://t.me/learningnets

Keep it high-level , you don’t need LLD

- No device specific details required

- Know at least what below devices do:

  - Firewalls, next-gen firewalls
- Routers
- Switches
- Load balancers • etc

75-80%

https://t.me/learningnets

- Take time to analyze the current environment
- Do not make assumptions,
- Don’t make over Engineering!
- All necessary information is provided during the scenario!

75-80%

https://t.me/learningnets

## HLD topics in CCDE Exam!

- Topology decisions
- Identifying protocols/features
- Build implementation/migration plans
- Adjusting network architecture/design to support business changes
- Merging networks (acquisitions)
- Divestiture (selling off)
- Resiliency/redundancy levels

75-80%

https://t.me/learningnets

# What are New in CCDE v3 Exam?

- Practical Exam will be in the Cisco CCIE Lab locations
- Scheduling done via CCIE/CCDE portal
  - Registration opens 90 days prior to exam date
- Every year, six CCDE exams are expected. (Previously 4, even sometimes 3)
- Exam score will be provided within 48 hours.

75-80%

https://t.me/learningnets

# What are New in CCDE v3 Exam?

- Introduction of Core and Area of Expertise modules

- • Core module covers technologies all candidates must know
- • Enterprise technologies (no Data Center/Service Provider)
  • Core module is vendor-agnostic

- • Area of Expertise module covers specific technology areas
  • More detailed knowledge expected
  • Cisco-specific technologies may appear in Area of Expertise module

75-80%

https://t.me/learningnets

## What are New in CCDE v3 Exam?

Area of Expertise options:

Large-Scale Networks
On-prem and Cloud Services
Workforce Mobility

75-80%

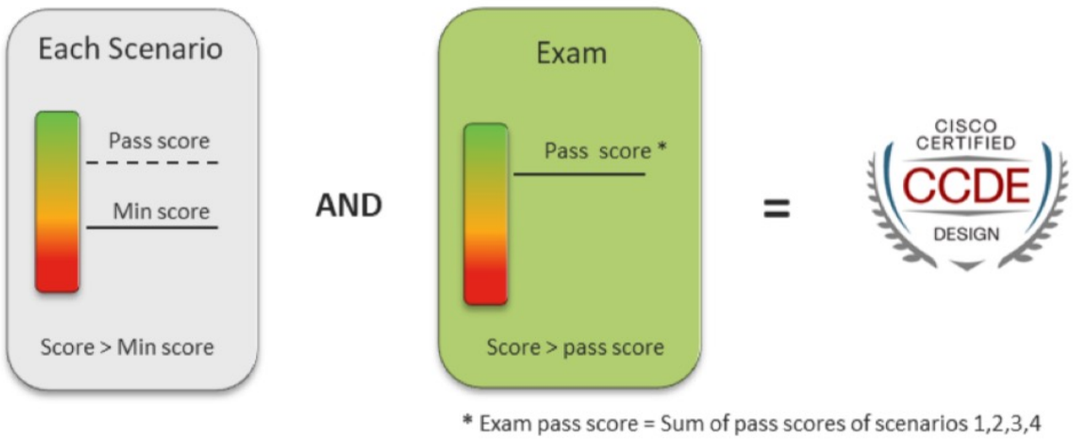
https://t.me/learningnets

# What are New in CCDE v3 Exam?

| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 |
|---|---|---|---|---|
| Exam Topics | Unified Exam Topics | | | |
| Technology List | Core | | | Area of Expertise |
| Format | Scenario-based | | | |
| Total Exam Time | 8 hours | | | |
| Time/scenario | Max 2 hours | Max 2 hours | Max 2 hours | Max 2 hours |
| Backwards navigation | Disabled | | | |
| Point values | Hidden | | | |

https://t.me/learningnets

# What are New in CCDE v3 Exam?




* Exam pass score = Sum of pass scores of scenarios 1,2,3,4

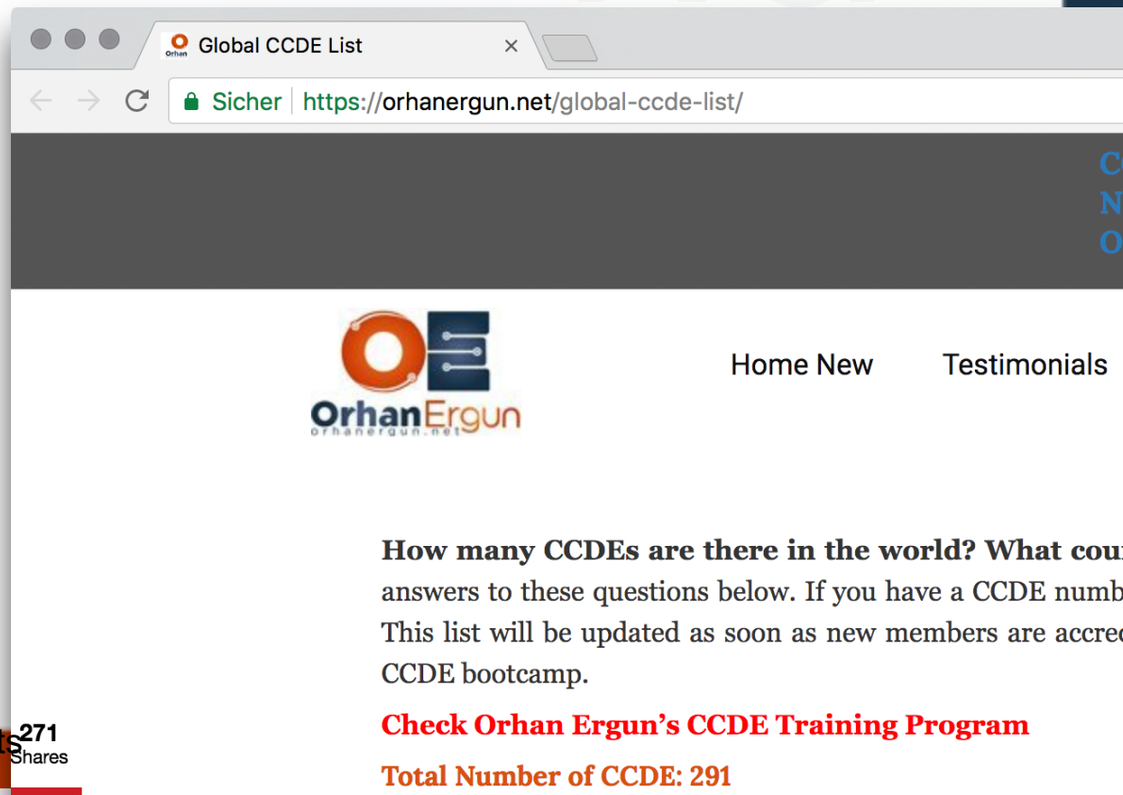
If candidates score less than the overall aggregated pass score of all scenarios OR score less than the minimum score set on 1 (or more) of the individual scenarios, they will fail the exam.

Upon passing the exam, candidates will not be provided any additional score details.

Each scenario has both a minimum score and a pass score.
Both minimum and pass scores are
set on a per-scenario basis, based upon exam difficulty level.

https://t.me/learningnets

- Less than 500 CCDE in the world as of 2021 , Most of them in U.S and as a company, work for Cisco
- https://orhanergun.net/global-ccde-list
- CCDE exam cancelled on May 2017 due to leak
- Almost 30 people lost their certificates as they involved in cheating as per Cisco
- New scenarios have been added to the exam since the cancellation, new topics have been introduced with the new scenarios
- All new topics will be covered in the course as well




https://t.me/learningnets
271 Shares

# CCDE Task Domains

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

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

https://t.me/learningnets

# Merge and Divest

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

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

merge div&est

https://t.me/learningnets

# Merge and Divest

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

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

merger & divest

https://t.me/learningnets

# Adding Technologies

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

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

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

https://t.me/learningnets

# Adding Technologies

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

  - What are the alternatives ?

  - Which additional information do you need to deploy this technology?

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

https://t.me/learningnets

# Replacing Technologies

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

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

- What is the potential impact to overall network?

https://t.me/learningnets

# Replacing Technologies

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

  - What will be the migration steps?
  - Is there a budget constraint?
  - Does this new technology require a learning curve ?

https://t.me/learningnets

# Scaling

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

https://t.me/learningnets

# Scaling

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

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

https://t.me/learningnets

# CCDE Written Recommended Reading List

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

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

These resources are useful for the CCDE Practical exam as well

https://t.me/learningnets

# Course Outline

## Layer 2 Technologies

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

outline

https://t.me/learningnets

# Course Outline

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

outline

https://t.me/learningnets

# Course Outline

## Transport Technologies

- **Broadband Basics**
- **The Fixed Broadband Services**

XDSL
FTTH
Cable Broadband
BPL
Fixed Satellite
Wireless

- **The Mobile Services:**

3G
LTE
Access Networks and Broadband Services Summary

outline

https://t.me/learningnets

# Transport Technologies

- **Transport Systems:**

Fiber Optic
Microwave
SONET/SDH
WDM
CWDM
DWDM
Dark Fiber
IRU
Leasing Fiber
Terrestrial Fiber Systems
Submarine Cable Systems
Cable Landing Stations
Beach Manhole
Transport Systems Summary

outline

https://t.me/learningnets

# Course Outline

## Layer 3 Routing

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

outline

https://t.me/learningnets

# Course Outline

## IS-IS

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

outline

https://t.me/learningnets

# Course Outline

## EIGRP

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

outline

https://t.me/learningnets

# BGP Course Outline

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

outline

https://t.me/learningnets

# BGP Course Outline

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

https://t.me/learningnets

# BGP Course Outline

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

outline

https://t.me/learningnets

# Course Outline

## MPLS and Applications

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

outline

https://t.me/learningnets

## Course Outline

# MPLS and Applications

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

outline

https://t.me/learningnets

# Course Outline

## MPLS and Applications

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

outline

https://t.me/learningnets

# VPN Design

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

outline

https://t.me/learningnets

# Course Outline

## VPN Design

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

outline

https://t.me/learningnets

# Course Outline

## SD-WAN

- Why SD-WAN?
- Traditional WAN vs. SD-WAN
- MEF SD-WAN
- SD-WAN Components (Architecture – Controller, Edges, Orchestration)
- Dynamic Path Selection
- WAN Optimization
- De-duplication
- Forward Error Correction
- Quality of Service
- ZTP
- Cloud On--Ramp

outline

https://t.me/learningnets

## Course Outline

# QOS Design

- IP QoS (RFC 4594)
- MPLS QoS
- Diffserv MPLS Tunneling Models – Uniform, Short Pipe and Pipe Models and Comparison
- Diffserv-Aware QoS
- QoS in the CCDE Exam
- Summary
- Bonus Materials

outline

https://t.me/learningnets

## Security Design

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

outline

https://t.me/learningnets

## Course Outline

### Multicast Design

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

outline

https://t.me/learningnets

## Course Outline

# IPV6 Design

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

https://t.me/learningnets

Course Outline

# Wireless Design and Architecture

## Wireless Basics

- Wireless Terms and Terminologies
- Types of Wireless Networks
- Wireless Standards

## WLAN Architecture

- Autonomous WLAN architecture
- Centralized WLAN architecture
- Distributed WLAN architecture

## WLAN Design

- Coverage Design
- Roaming Design
- Channel Design
- Capacity Design

outline

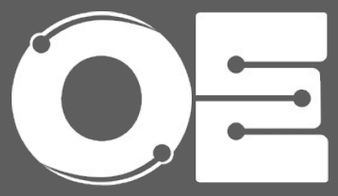
https://t.me/learningnets

## Evolving Technologies

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

outline

https://t.me/learningnets



# Transport Technologies

https://t.me/learningnets

# Agenda

- **Broadband Basics**

- **The Fixed Broadband Services**

XDSL
FTTX
Cable Broadband
BPL
Fixed Satellite
Wireless

- **The Mobile Services:**

3G
LTE
Access Networks and Broadband Services Summary

https://t.me/learningnets

# Agenda

- **Transport Systems:**

Fiber Optic
Microwave
SONET/SDH
WDM
CWDM
DWDM
Dark Fiber
IRU
Leasing Fiber
Terrestrial Fiber Systems
Submarine Cable Systems
Cable Landing Stations
Beach Manhole
Transport Systems Summary

https://t.me/learningnets

In the telecommunication world, the operators provide Internet, Voice, Video, Cable TV, Satellite TV and/or Internet, VPN, IPTV, Cloud, Hosting and many other services to their customers.

https://t.me/learningnets

Today, most of the operators provide more than one service to increase their revenues. Almost every operator provides an Internet service as of 2019.

These operators can be any of the following types:

- Landline Phone Companies
- Cable Companies
- Cellular Phone Companies
- Satellite Companies

https://t.me/learningnets

Service Providers provide different types of services. Some Service Providers provide broadband services, while others provide mobile services, cloud services, edge computing, VPN, Internet or hosting services

Companies generally provide more than one type of broadband access to their customers to increase their revenue

https://t.me/learningnets

In this module, we will look at different types of fixed and mobile based broadband services such as XDSL, FTTX, Cable Broadband, Fixed and Mobile Satellite, Wireless Internet Service and Mobile Broadband LTE (Long Term Evolution)

https://t.me/learningnets

Broadband signals compared to narrowband signals have much more band in the frequency spectrum

Thus, the term is referred as broadband

https://t.me/learningnets

Higher band in the frequency spectrum allows faster data communication.

Early Dial-up modems (over telephone lines) worked based on narrowband, thus they only provided voice communication and slow data speeds such as 56kpbs

https://t.me/learningnets

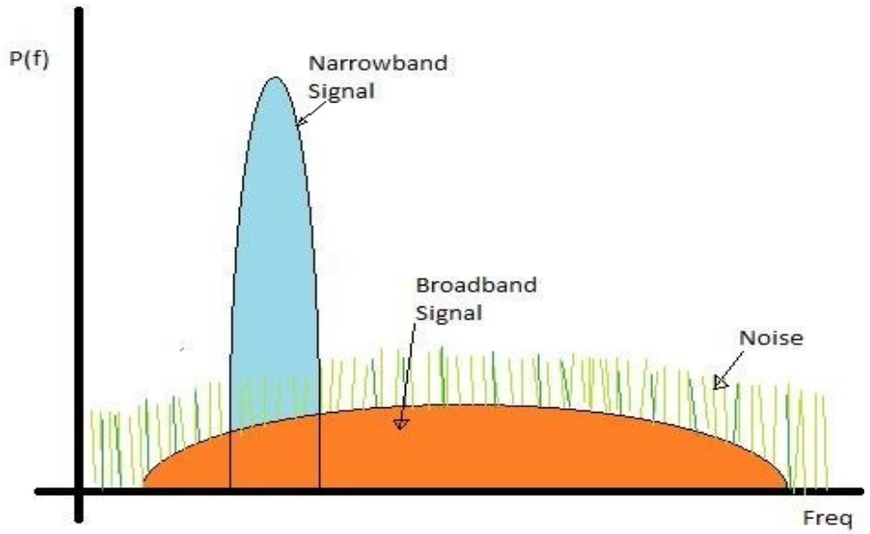
Broadband allows much higher data speeds such as 1Gbps or even more

DSL, FTTX, Cable Broadband, 3G, 4G Mobile Broadband are the examples of broadband technologies

These technologies will be covered in detail later in this module

https://t.me/learningnets

Carrier and ISP are very commonly used terms in the industry. There is a difference between them.




- Carrier is the company that owns the phone lines and maintains them.

- Service Provider is the company that is responsible for making sure that the services (such as voice, VPN and Internet Service) are functioning properly.

https://t.me/learningnets

Sometimes the Service Provider owns its hardware that provides the technical directions to the carrier, other times they manage the carrier’s hardware.

https://t.me/learningnets

The difference between Carrier and Service Provider is similar to FedEx and EBay.

One is bringing the service to you; while the other is selling those goods to you.

Broadband Technologies can be categorized as Fixed or Mobile Broadband Technologies.

https://t.me/learningnets

The Fixed Services provided in this chapter are:

- XDSL
- FTTH
- Cable Broadband
- BPL
- Fixed Satellite
- Wireless

The Mobile Services provided in this chapter are:

- 3G
- LTE

https://t.me/learningnets

# Fixed Broadband Service Technologies

## DSL

Dial-up modems were using analog transmission which were limiting the bandwidth around 56kbps

DSL uses digital transmission between telephony exchange (Central Office in U.S) and the customer modem

DSL provides broadband communication over ordinary copper telephone lines

https://t.me/learningnets

# Fixed Broadband Service Technologies

DSL services were initiated by the telephone companies which needed to provide higher bandwidth than 56kbps dialup, because their competitors such as Cable TV and Satellite companies begun providing 10Mbps and 50Mbps respectively.

https://t.me/learningnets

# Fixed Broadband Service Technologies

The Upstream and Downstream speed can be different, which is then called Asymmetric DSL.

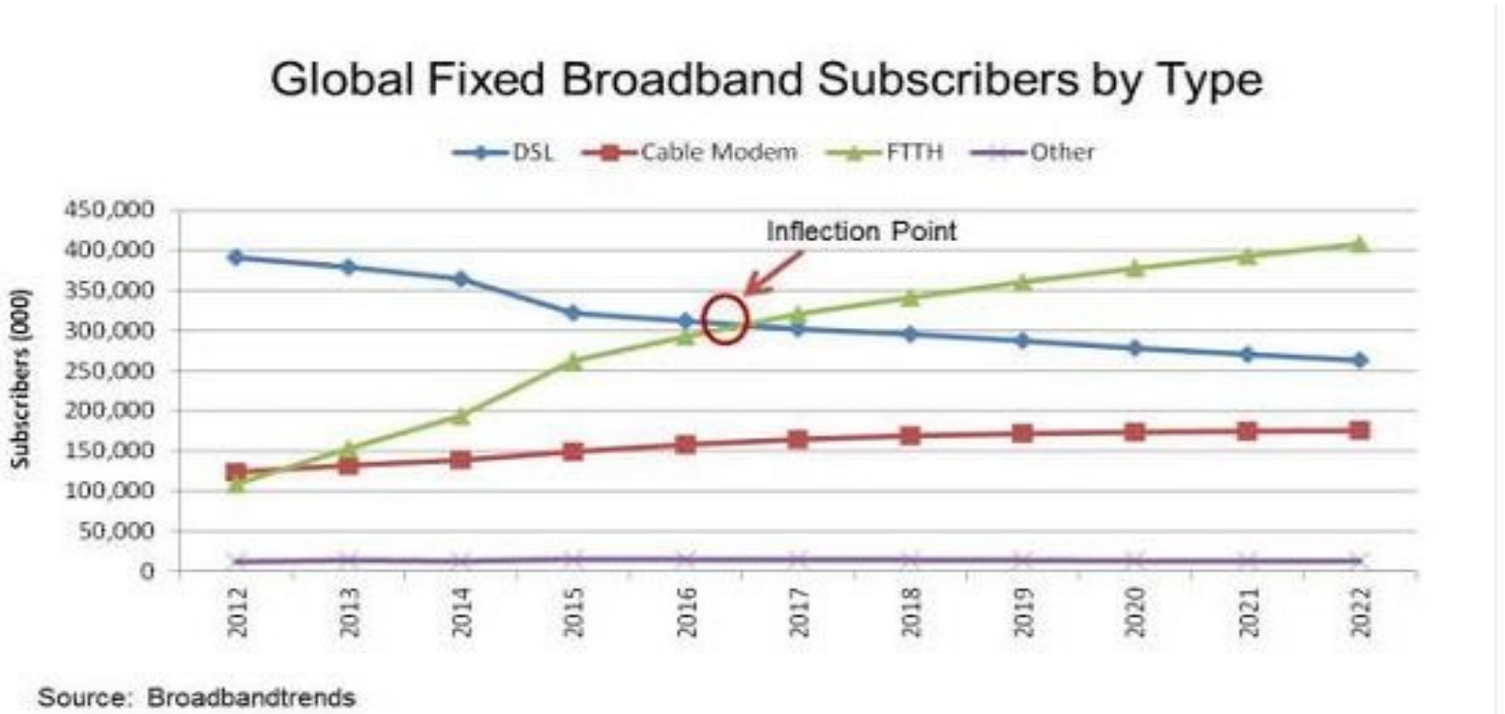
DSL is also known as Digital Subscriber Loop.

DSL was the most commonly deployed fixed broadband technology until recently.

https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

The actual data rate depends on several factors such as the modulation method used, the number of sub channels used, distance between CPE and DSLAM and also the quality of the copper wire which is the main factor causing a noise on the connection.




https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

Noise is a major bandwidth limiting factor for DSL connections

Bandwidth decreases because of copper signal attenuation and interference of surrounding signal (these two factors are simply called as noise)

Noise is strongly dependent on the distance between CPE and DSLAM and also the copper quality

https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

Copper quality can be described as a combination of wire age, wire radius, number of junction points and isolation of cables

For an example; a 5-year-old 0,5mm radius copper wire with 2-3 junction points can carry higher speeds to longer distances than a 20-year-old 0,4mm radius copper wire with 5-6 junctions

https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

DSLAM is the physical DSL modem termination equipment and located in the telephony exchange or street cabinet of the Service Provider

https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

The functionality of it is similar to CMTS in Cable broadband or eNodeB in Mobile broadband, as we will see later in this section

DSLAM which stands for Digital Subscriber Line Access Multiplexer is used to aggregate multiple DSL customers and send the traffic to the IP Backbone and vice versa.

https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

DSLAMs are placed very close to subscriber locations to provide faster speeds in recent years.

Old DSLAMs were very large, since they were located in Telecom Exchange and terminated 1000s of subscribers,

while new DSLAMs are pizza-box size and are located at the street cabinets.

https://t.me/learningnets

# DSL vs. other fixed access method Penetration in the world

DSL provides voice and Internet service over the same copper cable

DSL distances can go up to 5.5 km without a repeater

There are many different DSL implementations such as ADSL, ADSL2, ADSL2+, VDSL, VDSL2 and G.fast

https://t.me/learningnets

# DSL Technologies Speed Range

| Family | ITU | Name | Ratified | Max.Speed |
|---|---|---|---|---|
| ADSL | G.992.1 | G.dmt | 1999 | 7 Mbps down, 800 kbps up |
| ADSL2 | G.992.3 | G.dmt.bis | 2002 | 8 Mb/s down, 1 Mbps up |
| ADSL2plus | G.992.5 | ADSL2plus | 2003 | 24 Mbps down, 1 Mbps up |
| ADSL2-RE | G.992.3 | Reach Extended | 2003 | 8 Mbps down 1 Mbps up |
| SHDSL | G.991.2 | G.SHDSL | 2001 | 5.6 Mbps up/down |
| VDSL | G.993.1 | Very-high-data-rate DSL | 2004 | 55 Mbps down, 15 Mbps up |
| VDSL2 | G.993.2 | Very-high-data-rate DSL 2 | 2005 | 100 Mbps up/down |

https://t.me/learningnets

# Higher capacity, lower distance!

Newer technologies use enhanced modulation techniques and also use higher frequency,thus they can provide higher data speeds (bandwidth)

Higher frequency signals travel shorter distances compared to lower frequency signals, thus VDSL2 compared to ADSL2+ provides more bandwidth but requires much less copper distance from the modem to the DSLAM.

https://t.me/learningnets

# Higher capacity, lower distance!

VDSL can be used to provide speeds up to 100Mbps over short distances

but the main problem is interoperability due to lack of standardization

VDSL2 is a standard based mechanism which provides higher capacity compared to VDSL

https://t.me/learningnets

# Higher capacity, lower distance!

As of 2019, DSL speeds can go up to 10Gbps in theory with the XG.fast technology which works over existing copper telephone lines

https://t.me/learningnets

# Higher capacity, lower distance!

Nokia has achieved a connection speed of 5Gbps—about 625MB/sec—over 70 meters of conventional twisted-pair copper telephone wire,and 8Gbps over 30 meters

The trial used a relatively new digital subscriber line (DSL) protocol called XG.fast (aka G.fast2)

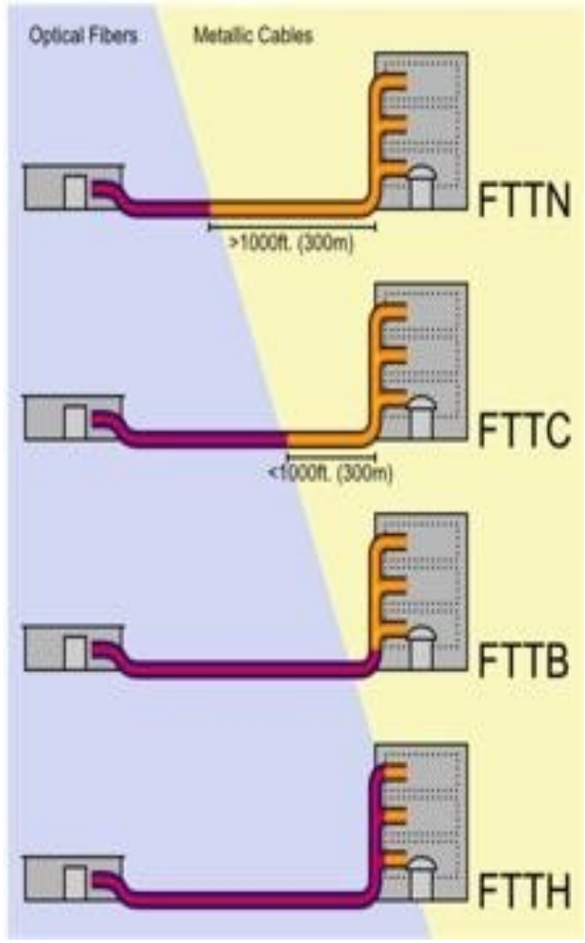
https://t.me/learningnets

# FTTX

Fiber to the X (FTTx) is a generic term for any broadband network architecture using optical fiber to provide all or part of the local loop used for last mile telecommunications

FTTN (Fiber to the Node), FTTC (Fiber to the Cabinet or Curb)

FTTB (Fiber to the Building) and FTTH (Fiber to the Home or Premise) are the FTTX deployment models.




https://t.me/learningnets

# FTTX

With FTTN and FTTC, fiber is laid up to the cabinet/node and copper wire connection between the street cabinet to the destination will complete the connection

With FTTB, FTTP and FTTH, fiber is laid up to the building, premise or home. Premise can be a home, apartment, condos, small businesses etc. Single mode fiber is used in all FTTx architectures.

https://t.me/learningnets

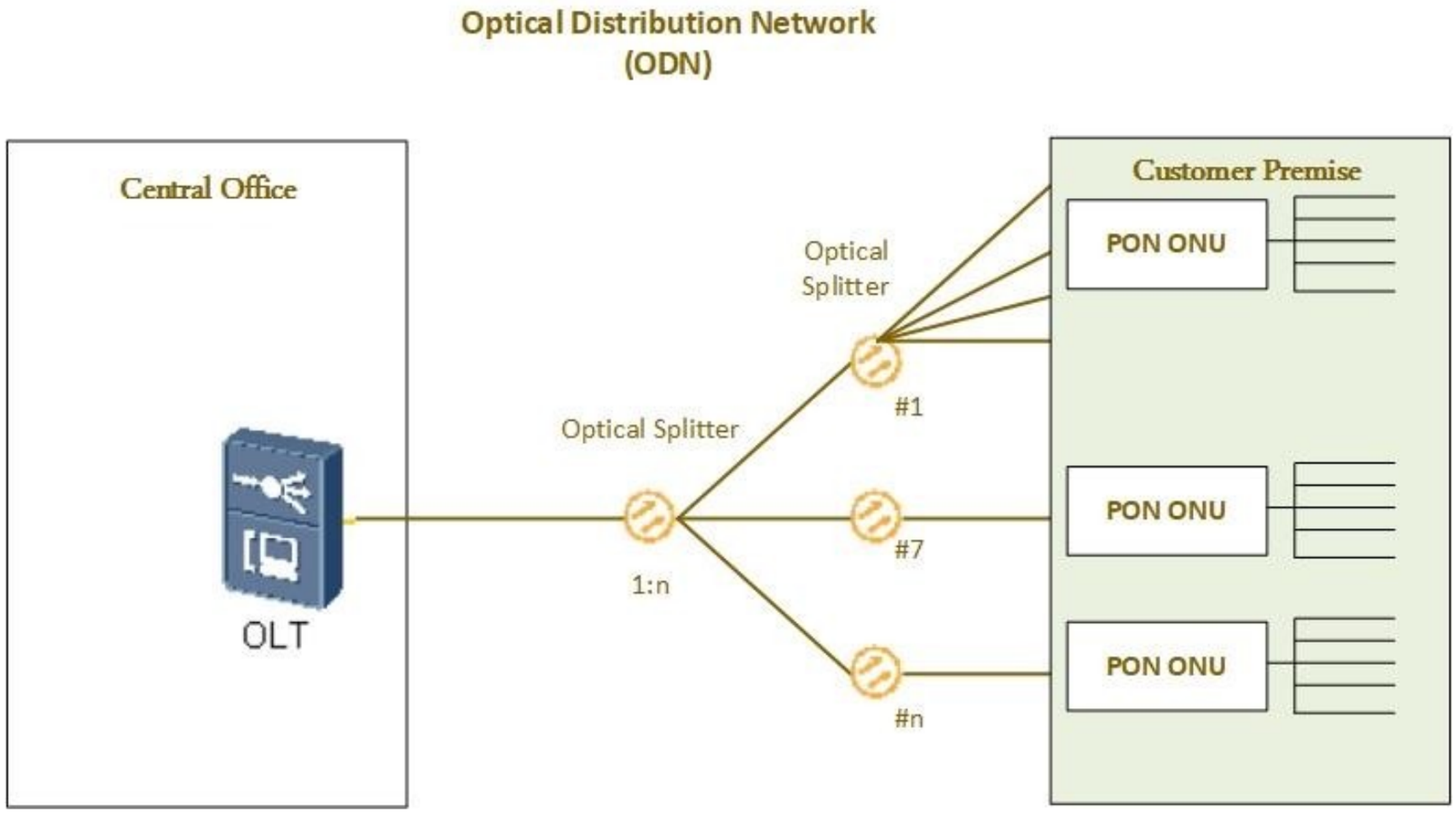
# FTTX

From the Operator’s central office (Telephone Exchange) to the destination (home users, other ISPs or enterprises), fiber is distributed over ODN (Optical Distribution Network)

ODN (Optical Distribution Network) is the physical path for optical transmission between OLT (Optical Line Terminal) and ONU (Optical Network Unit) equipment

https://t.me/learningnets

# FTTX




https://t.me/learningnets

# What is ONU in PON (Passive Optical Network)

Optical Network Unit (ONU) is a generic term denoting a device that terminates any one of the distributed (leaf) endpoints of an Optical Distribution Network (ODN), which runs a PON protocol

https://t.me/learningnets

# What is ONU in PON (Passive Optical Network)

In the case of a multi-dwelling unit (MDU) or multi-tenant unit (MTU), a multi-subscriber ONU typically resides in the basement or a wiring closet (FTTB case) and has FE/GE/Ethernet over native Ethernet link or over xDSL (typically VDSL) connectivity with each CPE at the subscriber premises.

https://t.me/learningnets

# What is ONU in PON (Passive Optical Network)

In the case where fiber is terminated outside the premises (neighborhood or curb side) on an ONT/ONU, the last-leg-premises connections could be via existing or new copper, with xDSL as the physical layer (typically VDSL)

In this case, the ONU effectively is a "PON-fed DSLAM"

https://t.me/learningnets

# OLT – ONU and ONT

OLT is the device which terminates the subscribers' fibers, located in the Service Provider network

OLT has two directions: upstream (getting different types of data and voice traffic from users) and downstream, which is getting data, voice and video traffic from metro network or from a long-haul network and sending it to all ONT modules on the ODN

https://t.me/learningnets

# OLT – ONU and ONT

ONT is the end user device which terminates the fiber, located at the user side (Apartments, Building, and Street Cabinet). ONT is also known as ONU

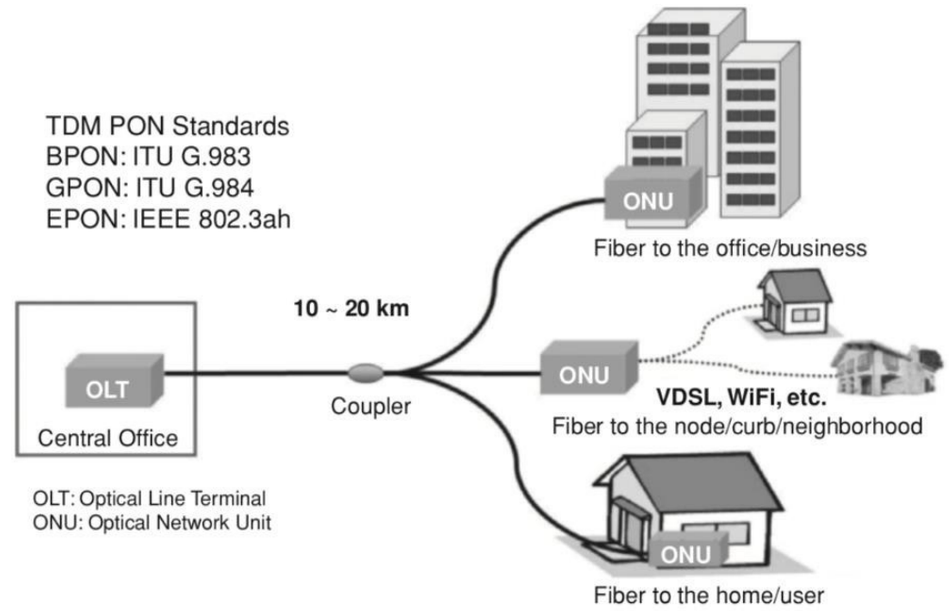
ONT is an ITU term (International Telecommunication Union)

https://t.me/learningnets

# OLT – ONU and ONT

ONU (Optical Network Unit) is an IEEE term; there is a slight difference though

In Europe, ONT term is used mostly, not ONU




Depends on the geographical location, these two terms might be used interchangeably

https://t.me/learningnets

# P2P and P2MP ODN (Optical Distribution Network)

Optical Distribution Network (ODN) can be setup as P2P (Point to Point) or P2MP (Point to Multipoint)

P2P ODN has single fiber per customer, between OLT and ONT

P2MP ODN shares the fiber cable between many subscribers

https://t.me/learningnets

# P2P and P2MP ODN (Optical Distribution Network)

P2MP ODN can be setup as AON (Active Optical Network) or PON (Passive Optical Network)

The most popular P2MP architecture is PON (Passive Optical Network)

https://t.me/learningnets

# Active and Passive Optical Networking for FTTx

P2P ODN (Optical Distribution Network) can be provided to large Enterprises or other Service Providers

It is a dedicated fiber per customer

https://t.me/learningnets

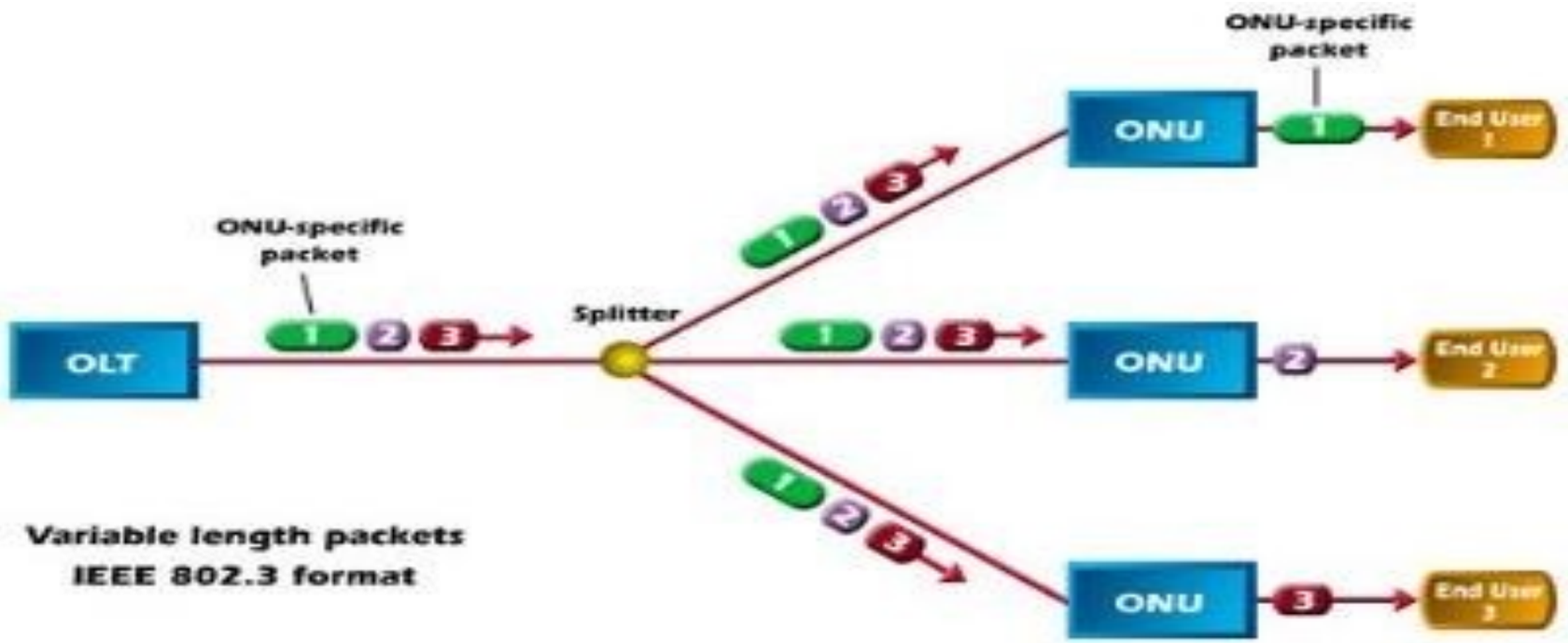
# Active and Passive Optical Networking for FTTx

Shared fiber ODN infrastructure with AON (Active Optical Networks) consists of equipment's which require electricity (Switches, Routers, amplifiers, repeaters)

AON can reach much higher distance compared to PON, but almost all residential FTTx deployments are done with PON (Passive Optical Network)

https://t.me/learningnets

# Optical Splitter and GPON




https://t.me/learningnets

# Optical Splitter and GPON

Optical Splitter is a passive equipment (Doesn’t require electrical power).

It broadcasts all the packets to the end users.

Split ratio on the splitter can be up to 1/128.

**PON Splitter**

**PON Transceiver**

https://t.me/learningnets

# Optical Splitter and GPON

GPON is one of the most popular PON standards. OLT sends the packets with ONU-ID,

Thus, although the packets are broadcasted by splitter, the correct ONU accepts its own packets (Similar to MAC learning in Ethernet Switches)

Splitters can be connected to the two different OLTs for the redundancy

Although this provides higher availability, it increases the cost of PON deployment

https://t.me/learningnets

# Optical Splitter and GPON

The most common splitters deployed in a PON system are 1: N or 2: N splitter ratio, where N is the number of output ports

Generally, the 1: N splitters are deployed in Star/Hub and Spoke topologies, while 2: N splitters are deployed in ring topologies to provide physical network redundancy

https://t.me/learningnets

# Optical Splitter and GPON

Using optical splitters in PON allows the service provider to conserve fibers in the backbone, essentially using one fiber to feed as many as 128 end users

A typical split ratio in a PON application is 1:32, which means one incoming fiber splits into 32 outputs

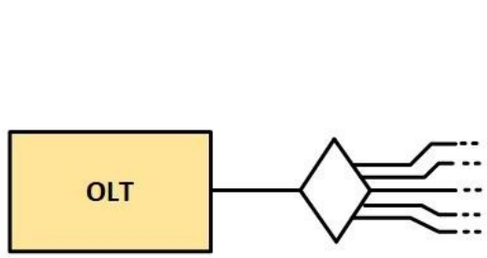
And the qualified fiber optic signal can be transmitted over 20 km

https://t.me/learningnets

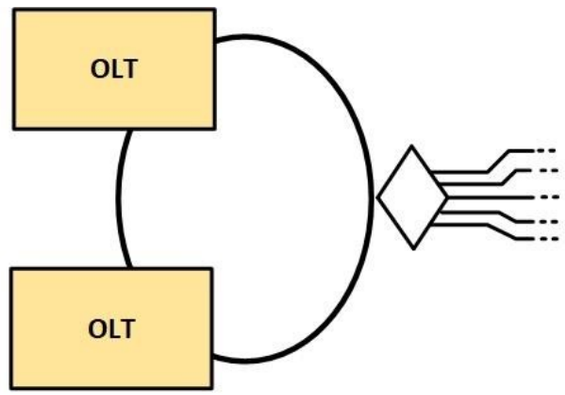
# Optical Splitter and GPON

If the distance between the OLT and ONT is small (in 5 km), you can consider about 1:64 or even 1:128

With higher split ratios, the PON network has both advantages and disadvantages




Hub and Spoke Topology




Ring Topology

https://t.me/learningnets

# Optical Splitter and GPON

Fiber optic splitters with higher split ratios can share the OLT optics and electronic costs as well as the feeder fiber costs and potential new installation costs

In addition, larger split ratio allows more flexibility and fiber management at the Headend

At the same time, higher split ratio splitters reduce bandwidth per ONU (optical network unit)

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

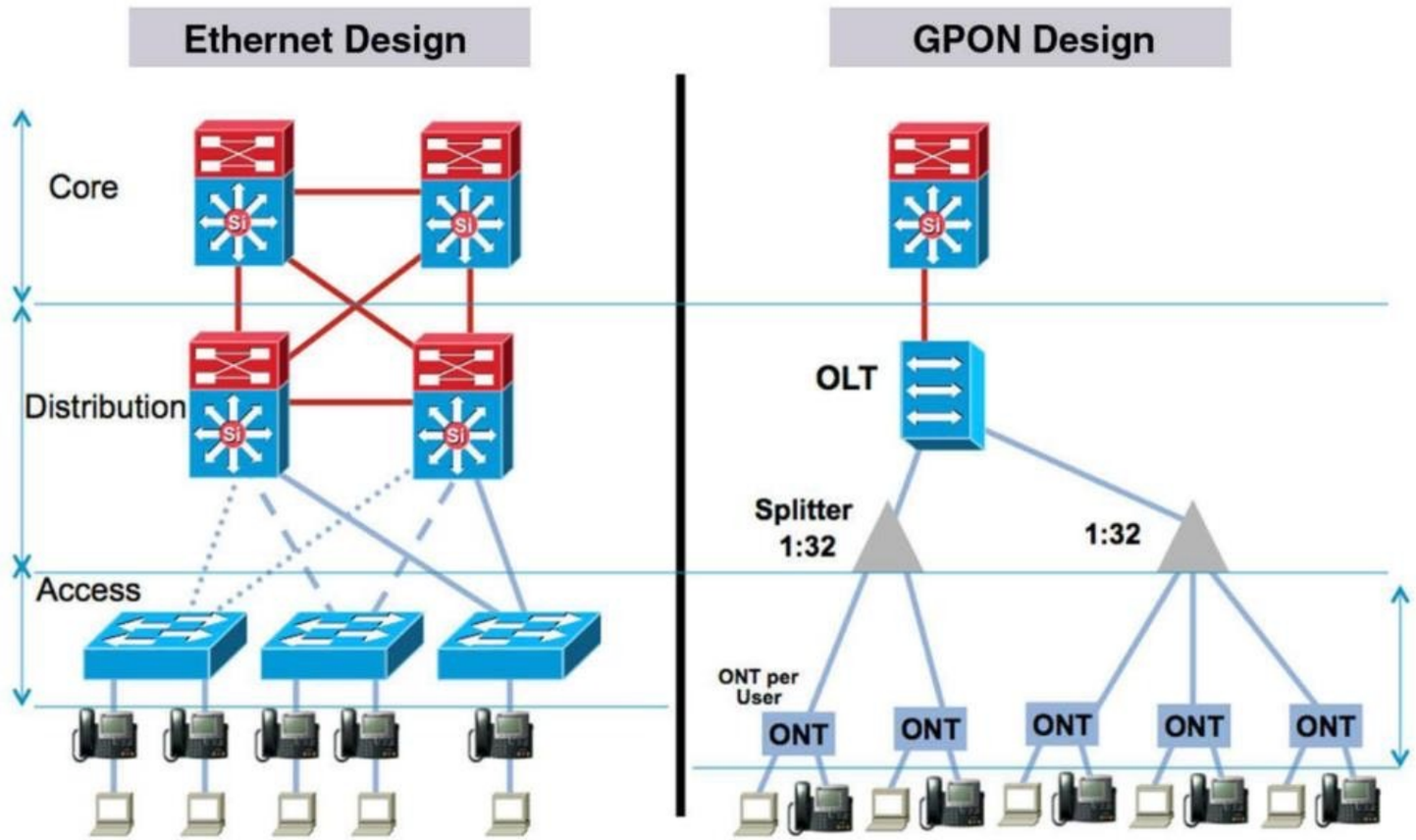
GPON (Gigabit Passive Optical Network) is used to reduce the number of active switching nodes in the network design

Network Design Best practice in Campus networks and many Datacenter networks (Not Massively Scale Datacenters) use Three-Tier: Access, Distribution and Core network design

Although the design decision depends on the scalability requirements in the Campus and DC, two layer: Access and Collapsed Distribution/Core design can be used

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture




https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

In Three-tier Traditional campus networks, there are active Ethernet devices used in each tier

Active means, nodes require electricity. Active Ethernet switches forward traffic based on forwarding rules

If it's a Layer 2 network, traffic is forwarded based on Layer 2 information, if it is a Layer 3 design, traffic is forwarded based on routing protocol information

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

GPON in the campus network replaces traditional three-tier design with two-tier optical network, by removing the Active access and distribution layer Ethernet switches with the ONT, Splitter and OLT devices

Although ONT requires power, the power requirement of ONT, compared to Active Ethernet switch is much less and the Splitter doesn’t require power at all

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

Splitter is purely a non- Active device

Analysis show that using GPON in the Campus network, instead of Active devices, reduces power requirement significantly

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

Many capabilities which are provided by Active Ethernet switches, such as Vlan awareness, Security features, Quality of service, Multicast, Redundancy, loop prevention etc. are provided with GPON design as well

In design, there is no best solution. Every solution has its advantages and the disadvantages

This is also true for the comparison between GPON vs. Active Ethernet

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

So far, it has been given that GPON has many advantages

On the other side, GPON has its bandwidth limitation as a technology

2.5Gbps Download, 1.25Gbps Upload limitation

Although with the newer generation of PON solution, it can provide more Download and Upload bandwidth

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

When we compare GPON with Traditional Active Ethernet, it is true that the Download bandwidth is 2.5Gbps and Upload bandwidth is 1.25Gbps, which is less than what traditional Ethernet can provide

Depending on the Split ratio on the Splitter, 2.5Gbps bandwidth might be shared with 32, 64 or even with 128 different end points

https://t.me/learningnets

# GPON vs. Traditional Ethernet Architecture

Thus, it is true to say that when there is more bandwidth requirement per end point, Active Ethernet architecture can provide more bandwidth, probably with a better cost

Cost analysis should be made carefully, as there might be different Fiber optic cable and transceiver requirements for each solution

https://t.me/learningnets

# Cable Broadband

For providing a broadband Internet service over Cable TV network, the Cable Broadband/Internet requires cable modem at the customer premise and CMTS (Cable Modem Termination System) at the cable provider facility

These facilities are cable television headend (It is similar to Telco CO (Central Office))

https://t.me/learningnets

# Cable Broadband

In DSL, bandwidth is dedicated on the access network per user

In PON (FTTx), the bandwidth is shared on the access network between the end users

In cable broadband the bandwidth is shared between the end users, similar to PON

https://t.me/learningnets

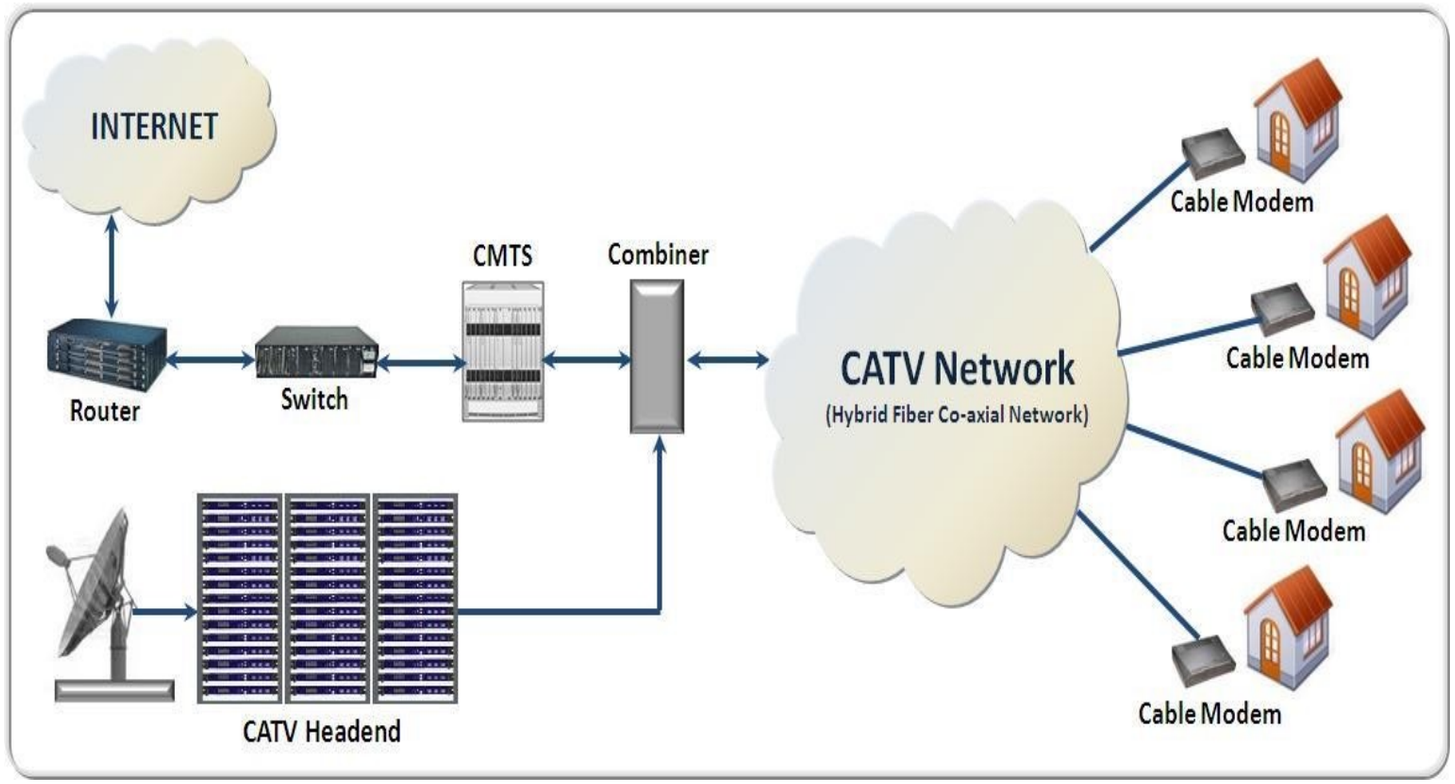
# Cable Broadband

CMTS and cable modem is connected either via coaxial cable or HFC

Coaxial cable used by cable TV allows broadband communication by transmitting of several channels using distinct frequencies

https://t.me/learningnets

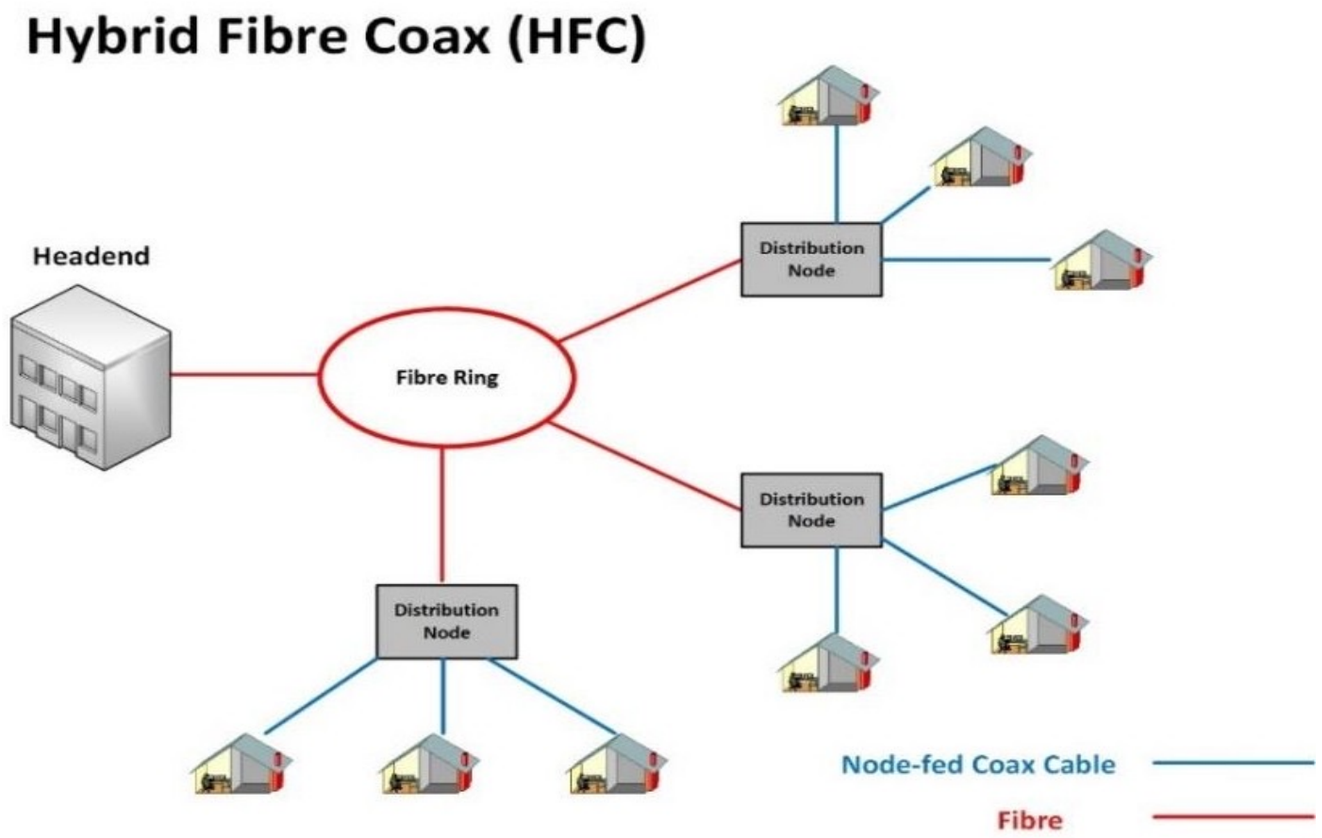
# Cable Broadband




https://t.me/learningnets

# HFC System (Hybrid Fiber Coax)

Most of the Cable Broadband systems use HFC which is Hybrid Fiber-Coaxial system




https://t.me/learningnets

# HFC (Hybrid Fiber-Coaxial) and Speed Dependency

Since with Cable Broadband, users share the access network bandwidth, the actual transfer rate achieved using cable TV networks is related to the number of users connected to the optical node at the same time,

Also, since all users will not be accessing the Internet at the same time (Statistical multiplexing)

More users means lower bandwidth available to each individual user

https://t.me/learningnets

# DOCSIS

The most common system used by cable TV companies to offer Internet access is called DOCSIS (Data Over Cable Service Interface Specification)

It was founded by CableLabs (Non-Profit Organization)

DOCSIS defines interface requirements for cable modems involved in high speed data delivery over HFC network

https://t.me/learningnets

# DOCSIS and the other Broadband Standards

DSL systems use ADSL, VDSL etc. for CPE to communicate with network nodes (DSLAM/MSAN). FTTX networks use GPON, EPON, WDM-PON between CPE (ONU in this case) and OLT

Cable Broadband uses DOCSIS for all communication standards between Cable modem and CMTS (Cable Modem Termination System)

https://t.me/learningnets

# DOCSIS and the other Broadband Standards

Cable Modem receives speed, IP address and time configuration parameters through CMTS from DHCP, TFTP and TOD (Time of Day) servers

The TFTP config file, sets the user broadband speeds, thus it is very crucial in cable broadband architecture to assign bandwidth and speed to the user (Some users try to upload their own tftp config file to change their broadband speed)

https://t.me/learningnets

# CMTS (Cable Modem Termination System)

CMTS is a provider edge system which connects the RF cable plant side to the provider IP core network

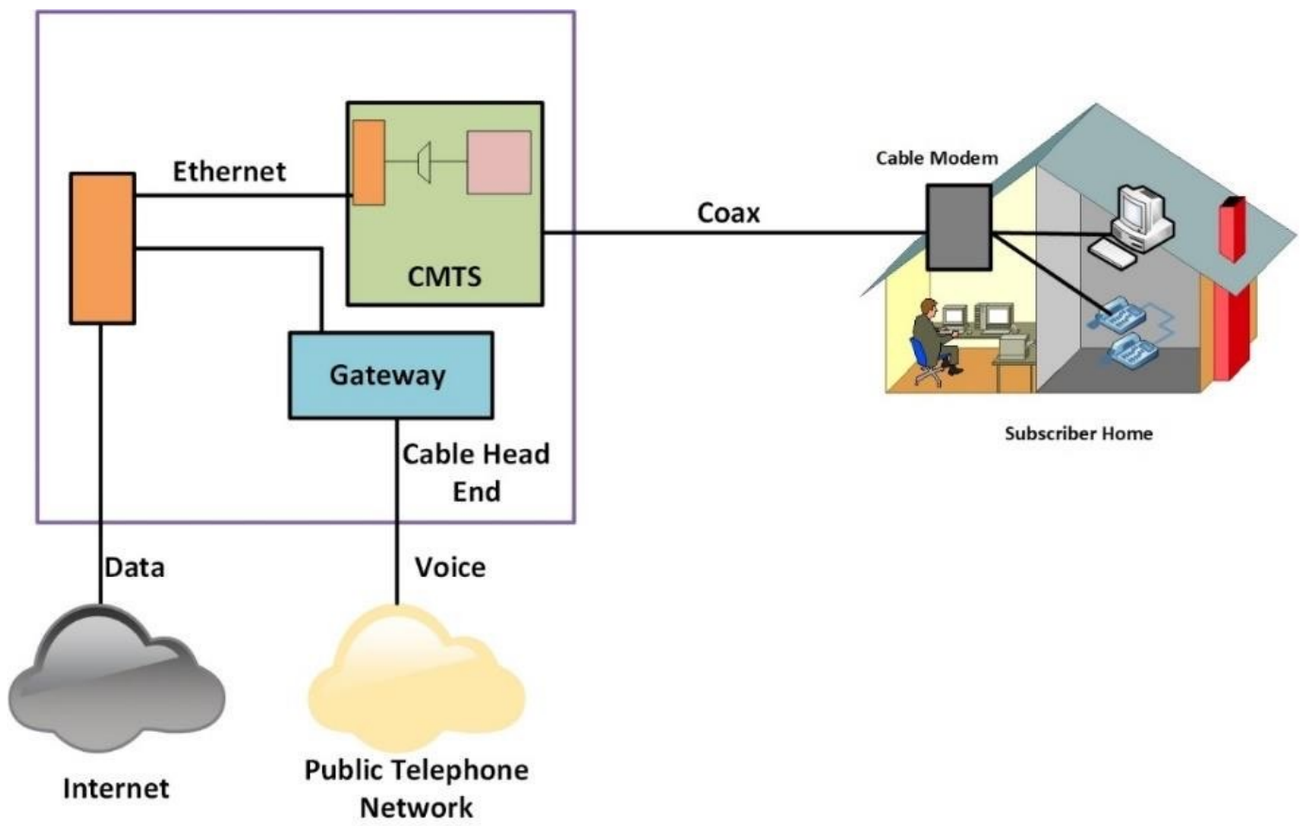
CMTS allows Cable Operator to offer broadband and other IP based services to the end subscribers connected to their cable network, including Voice and Video

CMTS functionality is similar to DSLAM in DSL and OLT in FTTx

https://t.me/learningnets

# CMTS (Cable Modem Termination System)

Cable Broadband service is mostly used in United States. It is used in some European countries as well, but more and more FTTX deployment is seen in Middle East, Europe and Africa




https://t.me/learningnets

# Fixed Wireless Service

WISPs provide a fixed wireless Internet service

This means that it relies upon a direct, line-of- sight connection from the access point to the roof of your home

Access point in WISP environment is commonly known as Base Station (BS)

https://t.me/learningnets

# Fixed Wireless Service

WISP shouldn’t be confused with mobile wireless technologies such as satellite and mobile broadband

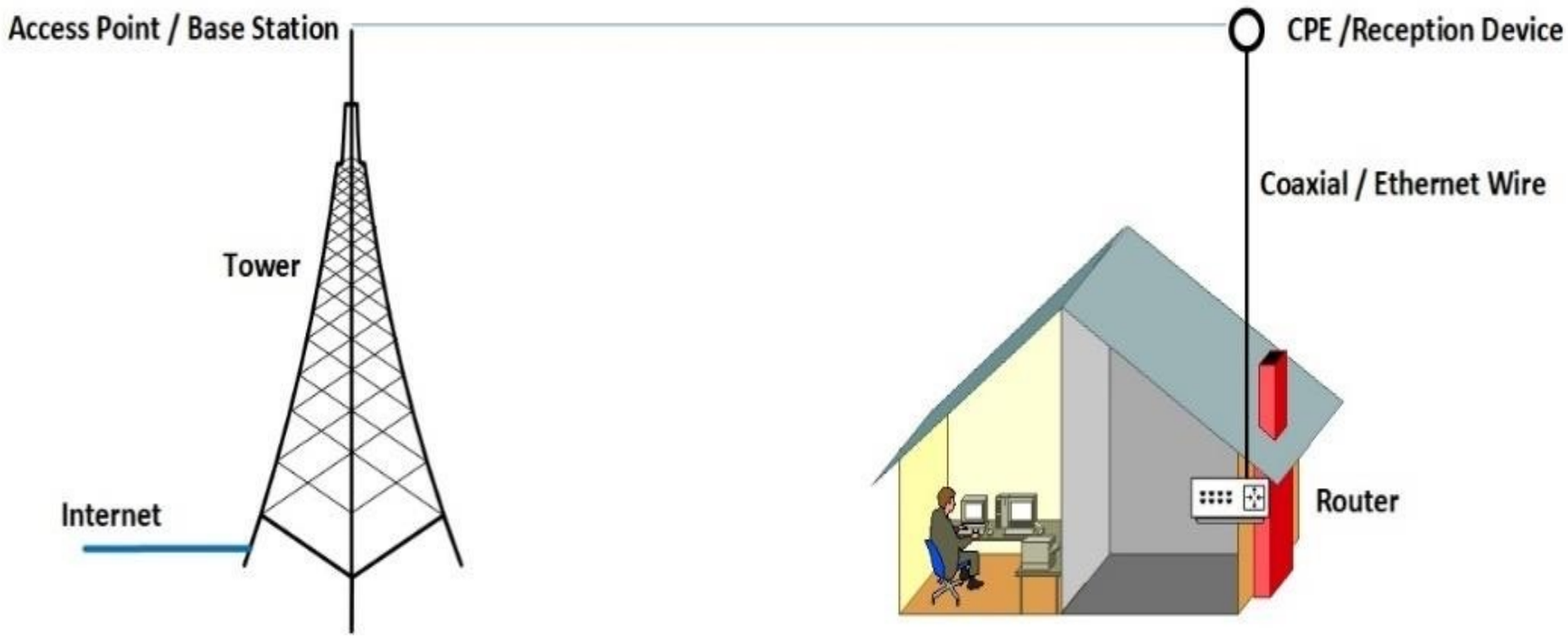
Satellite can provide Fixed or Mobile options and will be explained in the next section

WISP provides a fixed service which means end user devices are stationary (Fixed WISP service) or within very close diameter (Hot Spot WISP service)

https://t.me/learningnets

# Fixed Wireless Service

There are three main components in WISP networks, these are Base Stations (Access Point), Client Premises Equipment (CPE) and Backhaul network




https://t.me/learningnets

# Fixed Wireless Service

Base Stations are the equipment used to distribute the wireless from a single point, mostly located on a roof, water tower or tall buildings in order to transmit signal over obstacles such as trees and buildings

Base Stations allow for transmission of wireless signals at a range of anywhere between a couple of hundred meters up to distances of around 20km+ depending on the base station equipment used and environmental factors such as interference

https://t.me/learningnets

# Fixed Wireless Service

On the receiving end, at the house or office that was mentioned earlier, you will need a device called a CPE

This is in order to receive and transmit data wirelessly to and from the Base Station

These come in many shapes and sizes but the type and size of CPE mainly depends on the distance, signal strength and overall performance required

https://t.me/learningnets

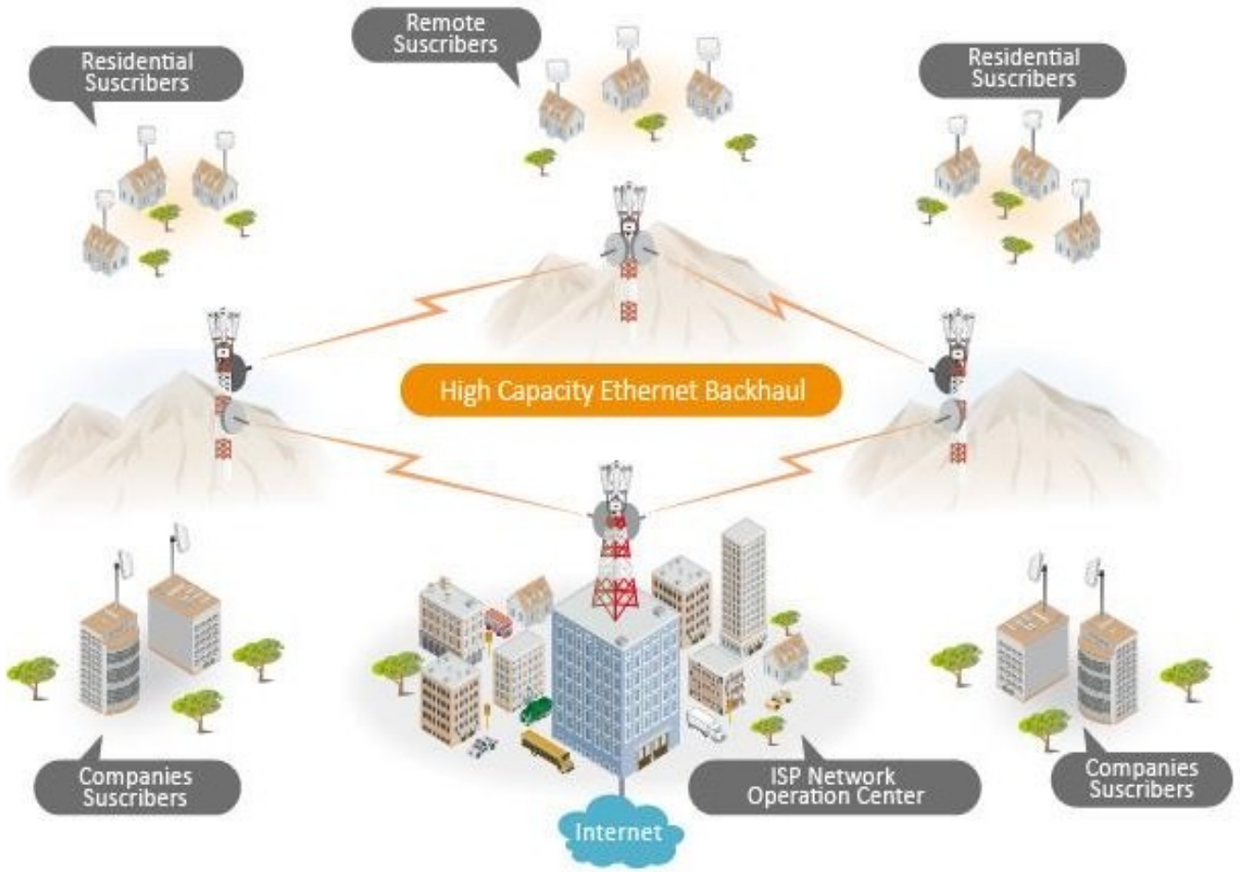
# Fixed Wireless Service

https://t.me/learningnets

# Fixed Wireless Service

WISP provides a fixed wireless solution; so, the customer antenna or dish is fixed and the customer cannot use their broadband service outside of their home

WISP is different than a Mobile/Cellular Operator as the Mobile Operator provides an Internet service that can also be used outside of homes.




In the figure a, Typical WISP setup and backhaul infrastructure are shown

https://t.me/learningnets

# Fixed Wireless Service

Backhaul speed depends on bandwidth between the tower and the central hub location

WISP can be an only choice in rural areas for high speed broadband as bringing fiber or other technologies are economically not profitable in those areas

CPE (Customer Premise Equipment) is shown in the figure here. Generally, WISP CPE is located at the roof of the house

https://t.me/learningnets

# Fixed Wireless Service

In order to provide Internet access to end users, WISP needs a circuit to the Internet

This can be provided by terrestrial services such as fiber or microwave point-to-point circuits, connecting your WISP location to a telecom provider, which will deliver the user traffic to the Internet

https://t.me/learningnets

# Fixed Wireless Service

In many of the places that new WISPs are being established, these backhaul options are not available or are very expensive

Often, satellite broadband as we will see in the next section is the only reliable option to connect the community to the Internet backbone

https://t.me/learningnets

# Different WISP Service Models

There are three types of WISP services in general

These are Hot-spot WISP, Standard WISP and Hybrid WISP

So far, we have been describing the Standard WISP setup

https://t.me/learningnets

# Different WISP Service Models

**Hot Spot WISP:** A directional or Omni-directional access point is placed on a building or tower near a target location with subscribers such as a village, an entertainment venue, stadium, airport, etc.

It delivers a powerful WiFi signal, so that those within the distance (typically up to 500 meters) can receive WiFi and if authorized, connect to the network

https://t.me/learningnets

# Different WISP Service Models

**This service option;** leverages the standard WiFi built into most smartphones and laptops

By using repeaters or point-to-point bridges, additional towers can be installed, linked together, and more area can be covered with a WiFi signal

Subscribers of this service are typically the individual users

https://t.me/learningnets

# Different WISP Service Models

**Standard WISP:** Similarly, an access point/base station is installed on the tower, but a different wireless technology, P2MP (Point to Multi-Point) is used, supporting much longer distances – several kilometers

In this case, small antennas or CPE's (Customer Premise Equipment) are installed at residences and offices

https://t.me/learningnets

# Different WISP Service Models

CPE provides an Ethernet connection to which a switch or WiFi router can be connected to it, so that those within the building can connect using WiFi, or standard wired Ethernet

With Standard WISP service, subscribers are generally businesses or residences

https://t.me/learningnets

# Different WISP Service Models

**Hybrid WISP:** A hybrid WISP design might use the standard WISP solution which includes CPE devices, to distribute the network to longer distances, where Hot Spot services might be used for users with their own WiFi devices

In this case, subscribers might be businesses and residents, as well as individual subscribers who use the Hot Spot services

https://t.me/learningnets

# WISP Billing Models

Following are the two basic WISP billing models:

**Unlimited services billing:** The subscriber typically signs up for a particular plan with a fixed MRC monthly recurring cost

This model is attractive to businesses that want a fixed price every month, and have no concerns about running out of capacity

https://t.me/learningnets

# WISP Billing Models

**Usage based billing:** These services are designed for residential clients and individual subscribers.

A plan will typically include some data rate along with a data cap (quota)

Data cap can be described as how much traffic the user is allowed to pass over the network within a time period, such as a month

The WISP needs to be able to help the subscriber select a plan that will meet their needs

https://t.me/learningnets

# Satellite Broadband

Satellite communication enables the connection for most of the rural areas and underserved areas (even in 2021) to the Internet

In this section, Satellite broadband system component and the challenges which Satellite broadband systems encounter, will be explained

https://t.me/learningnets

# Satellite Broadband

In addition to broadband communication, Satellite systems can provide weather forecasts, GPS, or satellite TV broadcasting

Other common satellite applications, such as extending cellular coverage, connecting ATM machines, and restoring communication infrastructure quickly among others, provides benefits to end users, governments and the companies

https://t.me/learningnets

# Satellite Broadband

Satellite communication networks are radio communication networks where a communication link is facilitated between two earth stations by using satellite, which serves as an in-space signal repeater for the transmitting earth station

A satellite communication link can be made in one of two directions: earth-to-space, known as the uplink; and space-to-earth, known as the downlink.

https://t.me/learningnets

# Satellite Broadband

Residential satellite broadband is generally provided through Geostationary (GEO) satellites

The Main problem with satellite communication is latency (delay), which is not acceptable for some applications such as online gaming or financial trading

Latency also decreases the overall throughput of the communication, so as an example, web browsing might seem very slow

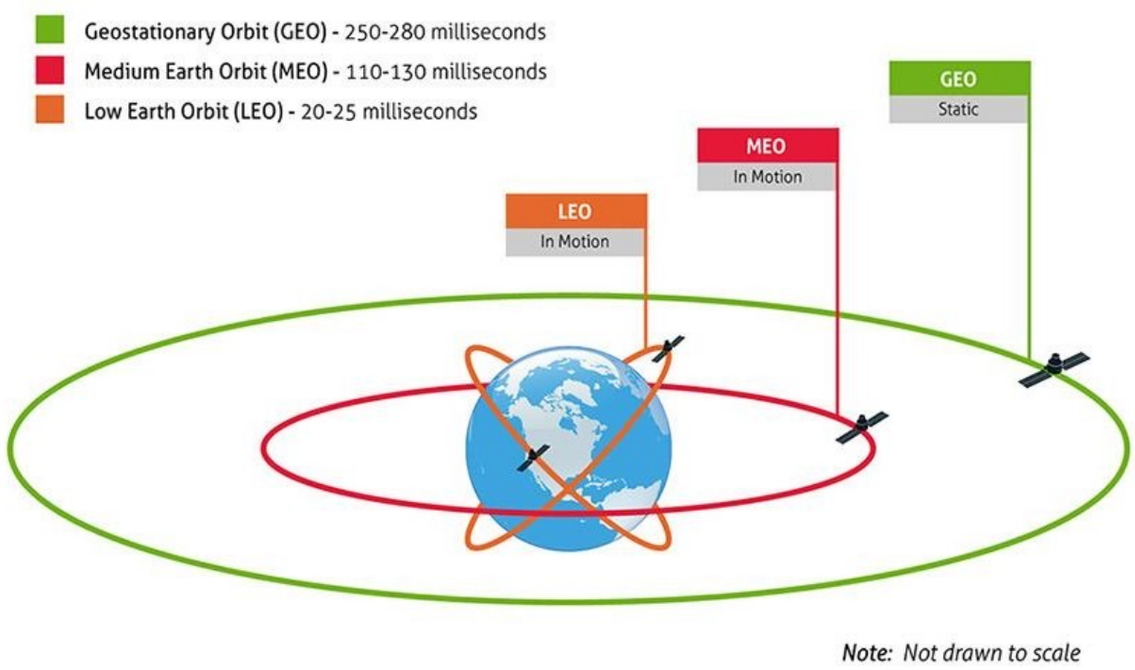
https://t.me/learningnets

# Satellite Broadband

In the figure below, three different Orbits are shown

LEO, MEO and GEO based satellites are categorized based on their distance from earth

Based on the distance between Earth and the Satellite, the Satellite can be positioned in one of these three orbits




https://t.me/learningnets

# Satellite Broadband

LEO based systems have the lowest latency

On the other side, GEO based systems have the highest latency

The reason is that LEO (Low Earth Orbit) has the closest Orbit to the Earth, while GEO (Geostationary Earth Orbit) has the farthest Orbit to the Earth

https://t.me/learningnets

# Satellite Broadband

Since latency is proportional to distance, closer orbit satellite system (LEO) has the lowest latency

While broadband requirements have traditionally been met through technologies like fiber, copper, microwave and 2G/3G/4G, satellites can now deliver connectivity with similar performance, including multi-gigabit speeds and low latency

https://t.me/learningnets

# Satellite Broadband

Common thought by many network engineers and consumers about satellite is that it is slow

This is mainly because most of the satellite systems historically were deployed at the GEO orbit to cover the entire world with very few numbers of satellites

https://t.me/learningnets

# Satellite Broadband

But there are currently more MEO (Medium Earth Orbit) and LEO based communication satellite systems, operating in many different satellite bands, providing Broadband Internet access or VPN services to the end users and to the commercial business

https://t.me/learningnets

# Satellite System Components

**VSAT** stands for very-small-aperture terminal which is the customer's site antenna LNB. Indoor modem is a satellite modem which transfers data from the computer to the Antenna. Satellite systems (technologies) are either GEO, MEO or LEO based satellites

**Ground Stations** are also called as Satellite Hub or Teleport and control all communications over the satellite link between the user and the Internet

https://t.me/learningnets

# Satellite System Components

VSAT dishes are commonly used in GEO based communication satellite systems

They are deployed at the residential or business side to receive satellite signals, focus the received energy and transmit the signals to the satellite modem which is located in the indoor location

This figure shows a satellite modem, or commonly referred as Satellite CPE

These modems are located indoor, receive signals and transmit the signals to the Wireless Access Points

These modems can also be manufactured with wireless capability

https://t.me/learningnets

# Satellite System Components

Satellite Earth station or commonly called Ground Station as well, is shown in the figure below

Satellite signal is sent from Earth Station to the real satellite in the sky and vice versa

Earth station is connected to the Internet and user Internet request arrives to the Earth station

https://t.me/learningnets

# Satellite System Components

In the figure below, the actual satellite is shown

Some satellites are 10 meters high, but some satellites are just couple centimeters long

Depending on the purpose, there are many different types of Satellite systems

https://t.me/learningnets

# Satellite System Components

In this module, we are only talking about Communication Satellites which can provide us Broadband access connectivity

There is an O3B satellite system which is located in the MEO orbit and works in KA band (Higher frequency, high rain fade) just couple meters tall

https://t.me/learningnets

# Satellite System Components

In general, newer communication satellites are smaller compared to older satellite systems, such as ones who were launched 10 – 15 years ago

After understanding the Satellite Internet components, let’s have a look at how it works and what is the actual path when users request content from Internet

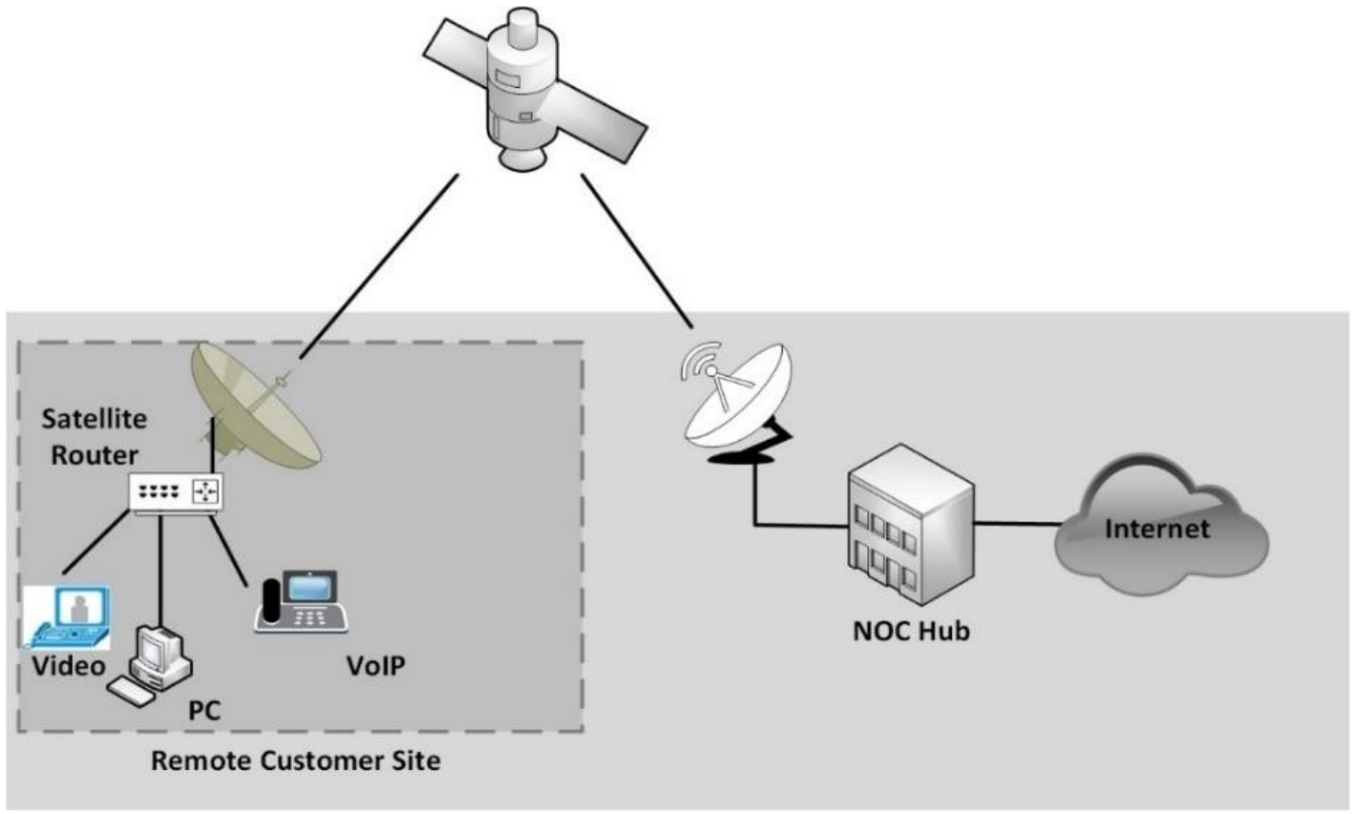
After that, we will also discuss the latency issue with Satellite Internet

https://t.me/learningnets

# How Satellite Internet works?

The orbiting satellite transmits and receives its information to a location on Earth called the Network Operations Center (NOC)

NOC is connected to the Internet so all communications made from the customer location (satellite dish) to the orbiting satellite will flow through the NOC before its reached to the Internet and the return traffic from the Internet to the user will follow the same path




https://t.me/learningnets

# How Satellite Internet works?

Data over satellite travels at the speed of light and Light speed is 186,300 miles per second

The orbiting satellite is 22,300 miles above earth (This is true for the GEO based satellite)

The data must travel this distance 4 times:

1. Computer to satellite
2. Satellite to NOC/Internet
3. NOC/Internet to satellite
4. Satellite to computer

https://t.me/learningnets

# How Satellite Internet works?

This adds a lot of time to the communication. This time is called "Latency or Delay" and it is almost 500 milliseconds

This may not be seen so much, but some applications like financial and real-time gaming don’t like latency

Who wants to pull a trigger, and wait half a second for the gun to go off?

https://t.me/learningnets

# How Satellite Internet works?

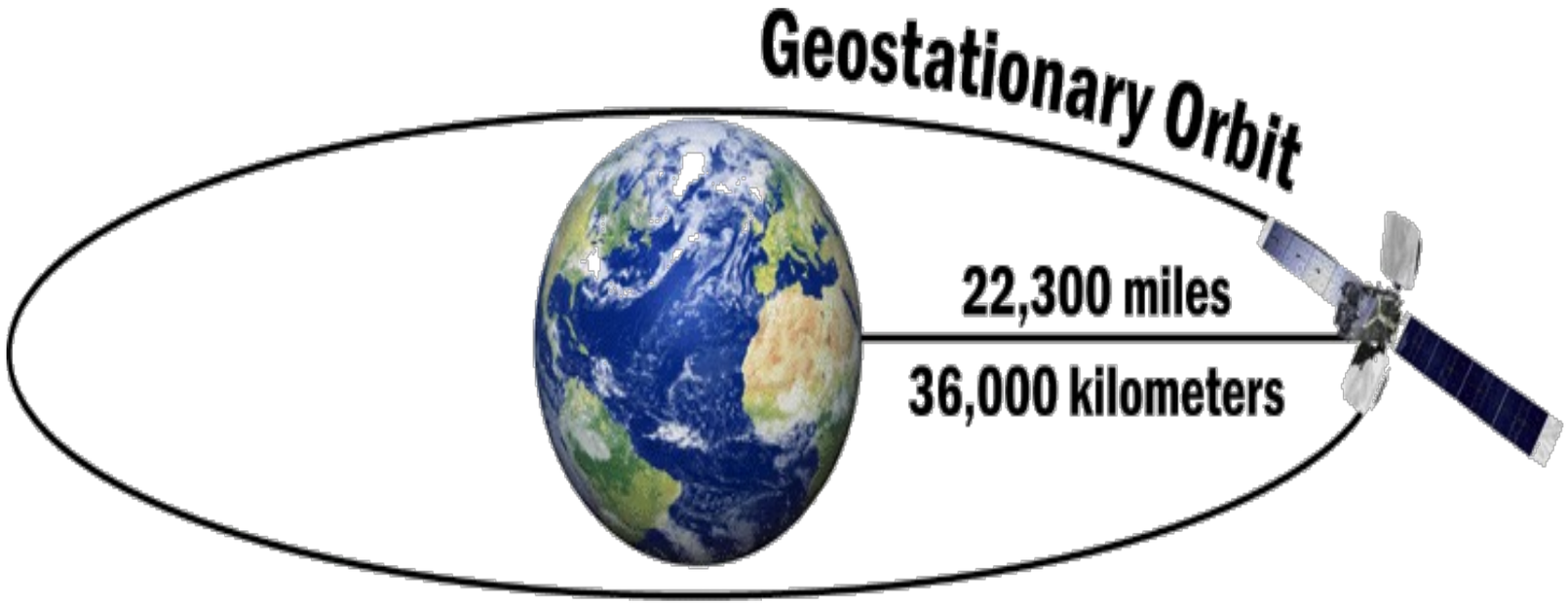
But, latency is related to which orbit the satellite is positioned

Let’s have a look at different Satellite Orbits to understand the satellite latency and its effect to the communication

https://t.me/learningnets

# Geostationary (GEO) Satellites

Geostationary satellites are earth-orbiting about 22,300 miles (35,800 Kilometers) directly above the equator

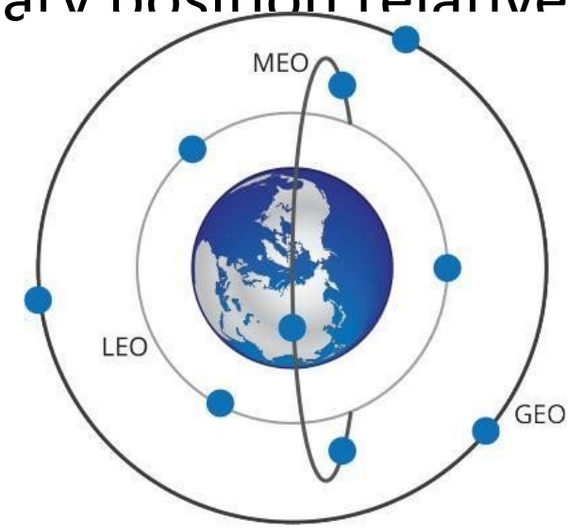



https://t.me/learningnets

# Geostationary (GEO) Satellites

They travel in the same direction as the rotation of the Earth

This gives the satellites the ability to stay in one stationary position relative to the Earth




Communication satellites and weather satellites are often given geostationary orbits, so that the satellite antennas that communicate with them do not have to move to track them, so it can be pointed permanently at the position in the sky where they stay

https://t.me/learningnets

# Geostationary (GEO) Satellites

The latency in GEO Satellites is very high compared to MEO and LEO Satellites

The geostationary orbit is useful for communication applications, because ground based antennas, which must be directed toward the satellite, can operate effectively without the need for expensive equipment to track the satellite’s motion

https://t.me/learningnets

# Geostationary (GEO) Satellites

There are hundreds of GEO satellites in orbit today, delivering services ranging from weather and mapping data to distribution of digital video-on-demand, streaming, and satellite TV channels globally

The higher orbit of GEO based satellite means greater signal power loss during transmission, when compared to lower orbit

https://t.me/learningnets

# Medium Earth Orbit Satellites

MEO is the region of space around the Earth above low Earth orbit and below geostationary orbit

Historically, MEO constellations have been used for GPS and navigation applications, but in the past five years, MEO satellites have been deployed to provide broadband connectivity to service providers, government agencies and enterprises

https://t.me/learningnets

# Medium Earth Orbit Satellites

Current applications include delivering 4G LTE and broadband to rural, remote, and underserved areas where laying fiber is either impossible or not cost effective – such as cruise or commercial ships, offshore drilling platforms, backhaul for cell towers, and military sites, among others

https://t.me/learningnets

# Medium Earth Orbit Satellites

In addition, Service Providers are using managed data services from these MEO satellites to quickly restore connectivity in regions where the service has been lost due to undersea cable cuts or where major storms have occurred

MEO satellite constellations can cover the majority of Earth with about eight satellites

https://t.me/learningnets

# Medium Earth Orbit Satellites

Because MEO satellites are not stationary, a constellation of satellites is required to provide continuous service

This means that antennas on the ground need to track the satellite across the sky, which requires ground infrastructure which is more complex compared to GEO based satellites

https://t.me/learningnets

# Low Earth Orbit (LEO) Satellites

Unlike geostationary satellites, low and medium Earth orbit satellites do not stay in a fixed position in the sky

Consequently, ground based antennas cannot be easily locked into communication with any one specific satellite

https://t.me/learningnets

# Low Earth Orbit (LEO) Satellites

Low Earth orbit satellites, as their name implies, orbit much closer to earth

LEOs tend to be smaller in size compared to GEO satellites, but require more LEO satellites to orbit together at one time to be effective

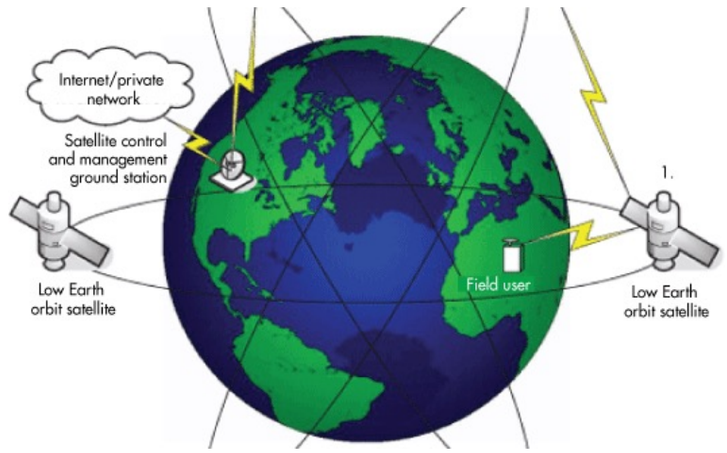
Lower orbits tend to have lower latency for time-critical services because of the closer distance to earth

https://t.me/learningnets

# Low Earth Orbit (LEO) Satellites

It’s important to reiterate that many LEO satellites must work together to offer sufficient coverage to a given location

Although many LEOs are required, they require less power to operate because they are closer to earth




https://t.me/learningnets

# Low Earth Orbit (LEO) Satellites

Choosing to go with more satellites in the LEO orbit on less power, or using fewer larger satellites in GEO, is the biggest decision to make here

Due to the high number of satellites required in LEO constellations, LEO satellites systems are expected to be high initial manufacturing and launch costs and more expensive ground hardware compared to GEO

https://t.me/learningnets

# Satellite Operating Bands

Before sharing the different frequencies (bands) which are used in satellite communication, it is important to understand the phenomena called ‘rain fade’

Rain fade is an interruption of wireless communication signals as a result of rain, snow, or ice, and losses which are especially prevalent at frequencies above 11 GHz

https://t.me/learningnets

# Satellite Operating Bands

Satellite communications use microwave frequencies, which require direct line of sight between the receiving and transmission equipment

https://t.me/learningnets

# Satellite Operating Bands

The following frequency ranges are commonly used:

**C Band (4-8GHz):** These lower frequencies have longer wavelengths and require larger dishes (1.8-2.4m) for reception but are not affected by "rain fade”

Larger dish size requires more expensive equipment compared to Ku and Ka bands. The C-band frequency range has one significant problem

https://t.me/learningnets

# Satellite Operating Bands

It is the frequency region assigned to terrestrial microwave radio communication systems

There are an emerging number of these microwave systems located all over the world and they carry a large volume of commercial communications

Consequently, the VSAT locations are needed to be restricted in order to prevent interference with the terrestrial microwave communication systems

https://t.me/learningnets

# Satellite Operating Bands

C-band was the first band that was used for satellite communication systems

However, when the band became overloaded (due to the same frequency being used by terrestrial microwave links), satellites were built for the next available frequency band, the Ku-band

https://t.me/learningnets

# Satellite Operating Bands

**Ku band (12-18GHz):** A shorter wavelength permits smaller dishes. The Ku-band frequency range is allocated to be exclusively used by satellite communication systems, thereby eliminating the problem of interference with microwave systems

Due to higher power levels at new satellites, Ku-band allows for significantly smaller earth station antennas and RF units to be installed at the VSAT location

Ku-band is typically used for broadcasting and Internet communication

https://t.me/learningnets

# Satellite Operating Bands

**Ka band (26.5-40GHz):** Ka-Band is a relatively new frequency band for satellite broadband and will provide additional transmission capacity

Ka-band has several advantages, with perhaps the most significant as having the 2-3 GHz increase in bandwidth, which is double available bandwidth than in Ku band and five times more than C band

https://t.me/learningnets

# Satellite Operating Bands

Due to the smaller wavelength, Ka-band components are typically smaller, leading to smaller antennas on the same-sized platform

But its sensitivity to rain fade makes it difficult for rainy regions and will support the use of small antennas (<1.2m)

https://t.me/learningnets

# Satellite Operating Bands

It is good to mention about recent ideas in communication satellite systems

Space-X, Google, Facebook, and many other Internet giants have low earth orbit or low altitude satellite communication system plans

https://t.me/learningnets

# Satellite Operating Bands

Elon Musk’s company SpaceX, has been granted permission by the US Federal Communications Commission (FCC) to set up a vast network of thousands of low Earth orbit communication satellites

His Satellite project name is Starlink

https://t.me/learningnets

# Satellite Operating Bands

Starlink will mostly be targeting to high-frequency traders at big banks, who might be willing to fork out large sums for dedicated, faster connections

Initially, Starlink will consist of 4425 satellites orbiting between 1100 and 1300 kilometers up, a greater number of active satellites compared to what are currently in orbit

https://t.me/learningnets

# Satellite Operating Bands

When sending a network traffic via Starlink, a ground station will begin by using radio waves to talk to a satellite above it

Once in space, the message will be fired from satellite to satellite using lasers until it is above its destination

https://t.me/learningnets

# Satellite Operating Bands

From there, it will be beamed down to the right ground station using radio waves again

Between distant places, this will allow messages to be sent about twice as fast as through the optical fibers on Earth that currently connect the Internet, despite having to travel to space and back

This is because the speed of the signal in glass is slower than it is through space

https://t.me/learningnets

# Satellite Operating Bands

For most people, the regular Internet is fast enough already

But for certain applications such as high-frequency trading or online gaming, speed/latency is the most important aspect of communication

https://t.me/learningnets

# Mobile Service Technologies

Mobile broadband also referred to as WWAN (Wireless Wide Area Network), is a general term used to describe high-speed Internet access from mobile providers for portable devices

https://t.me/learningnets

# Mobile Service Technologies

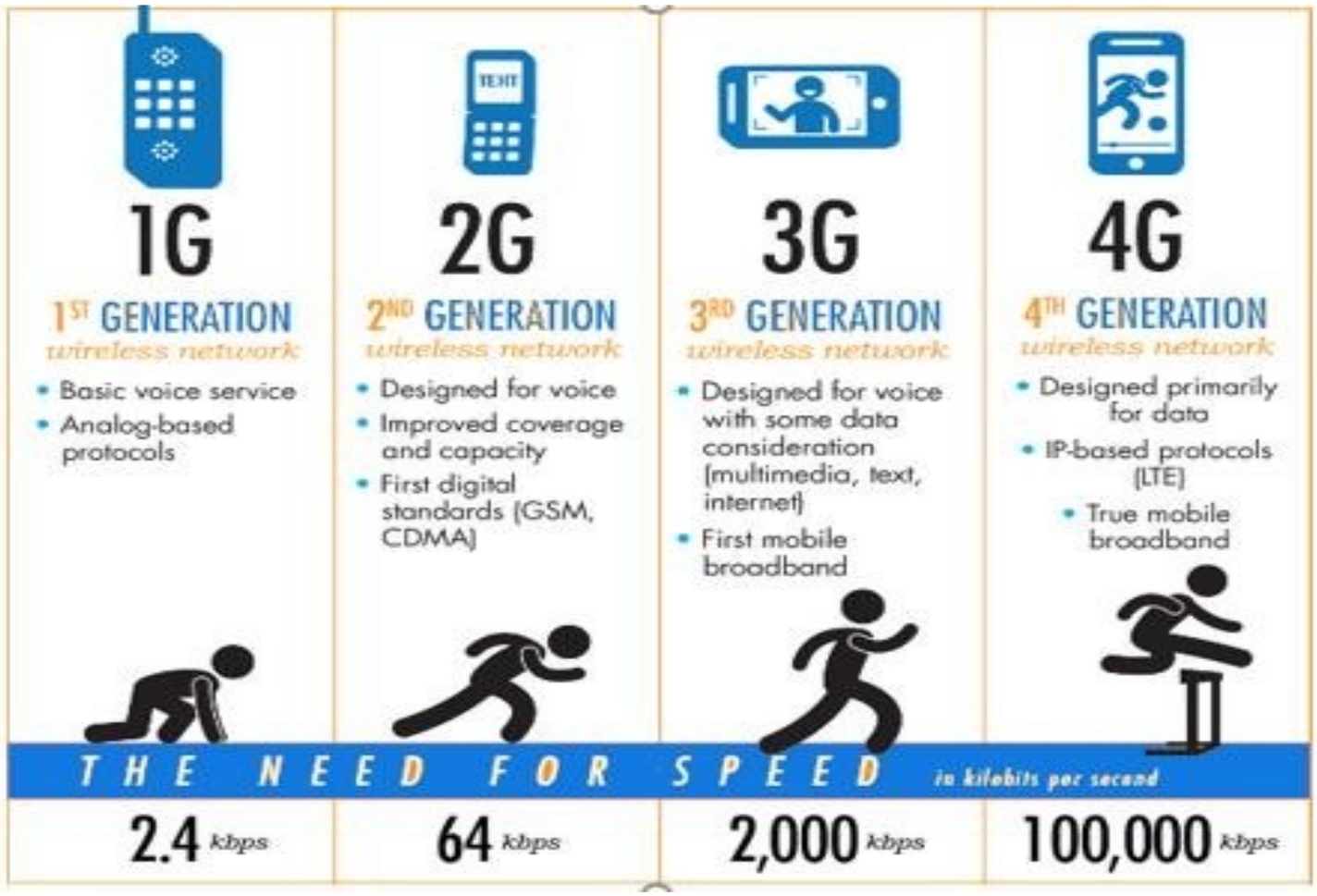
Just as wireless networking evolved from 802.11b to 802.11n with faster speeds and other improved performance and security features, mobile broadband performance continues to evolve, with having so many players in this growing field

Access network infrastructure links the backbone network to the customers

https://t.me/learningnets

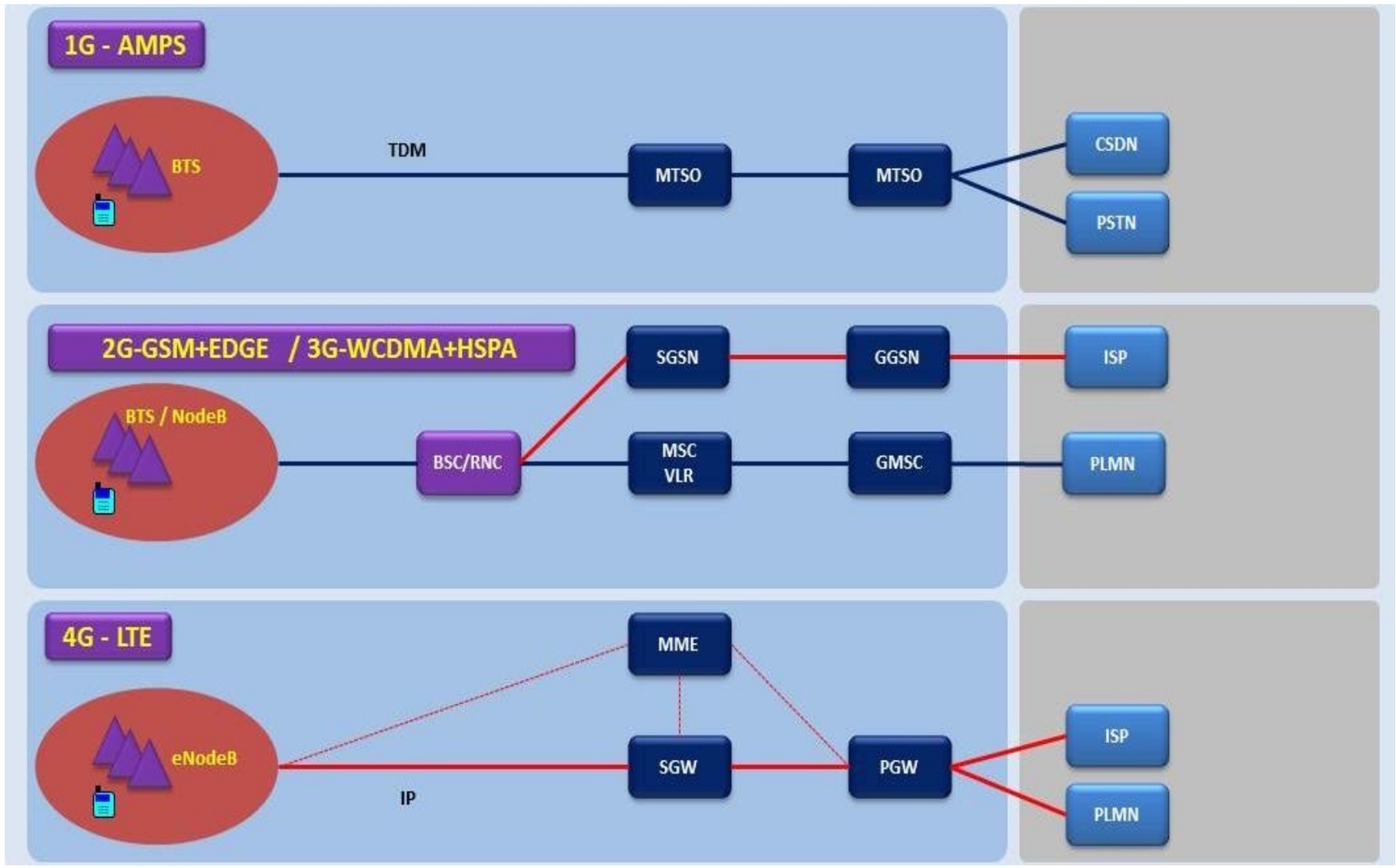
# Mobile Service Technologies

Started from 1G as 2.4kbps, Mobile broadband as of 2018 can provide more than 100Mbps.




https://t.me/learningnets

# Mobile Service Technologies




https://t.me/learningnets

# LTE Architecture – Terms

UE is a User Equipment. SIM Card + Mobile device together is called UE

eNodeB is Enhanced Node B, it is an enhanced base station because radio controller is located in the same place with the Base Station, this is different from 3G architecture

https://t.me/learningnets

# LTE Architecture – Terms

MME is the mobility management entity. MME is the key control node for LTE access network

It is responsible for tracking and paging procedures including retransmissions, and also for idle mode of User Equipment (UE)

https://t.me/learningnets

# LTE

Serving GW is the gateway which terminates the interface towards E-UTRAN

For each UE associated with the EPS, at given point of time, there is a single Serving GW

https://t.me/learningnets

# LTE

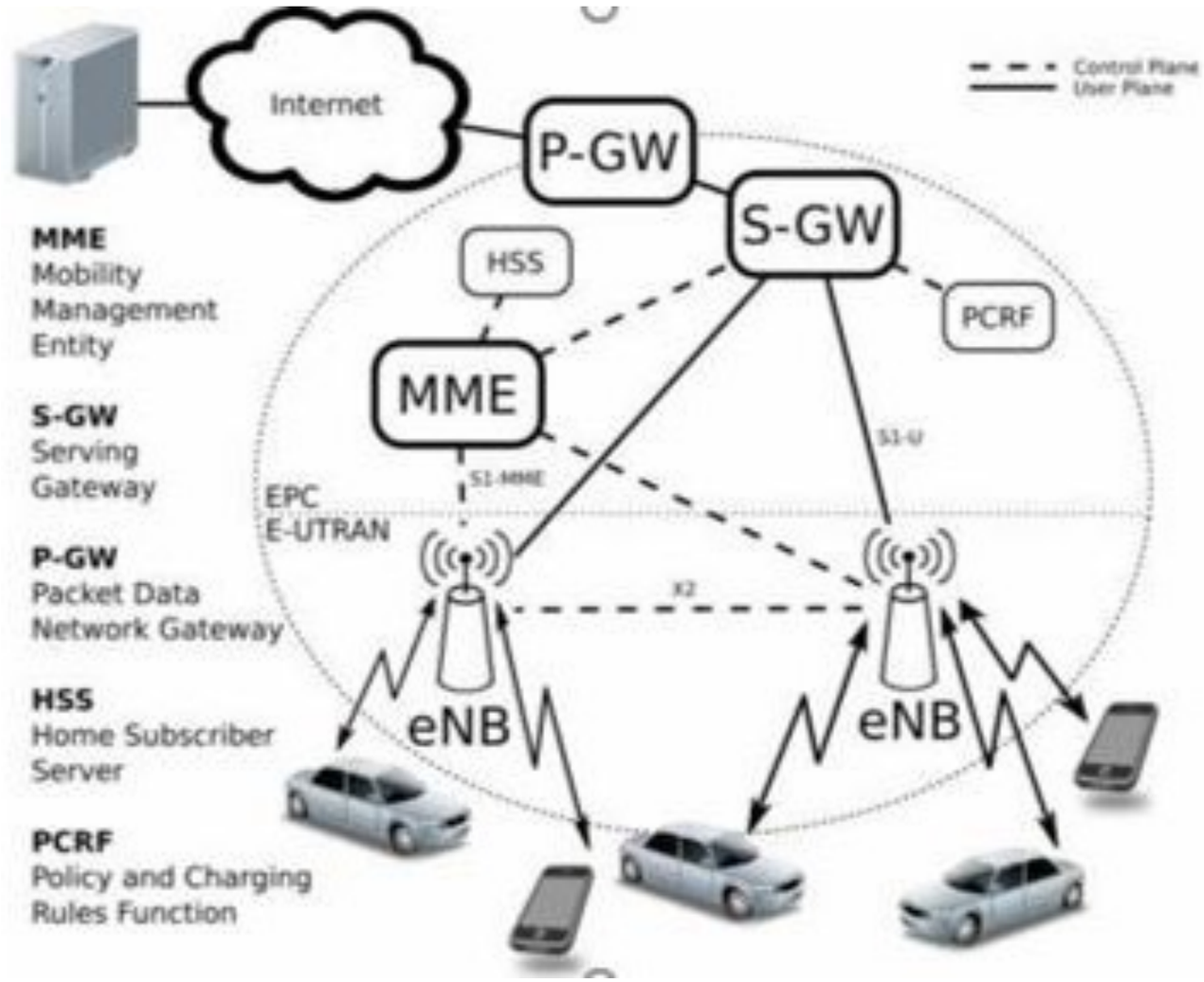
SGW is responsible for handovers with neighboring eNodeB's, also for data transfer in terms of all packets across user plane which is known as data plane by many network engineers

PGW is the gateway between the Radio Access Network and the IP network

It handles the User IP address allocation as well

https://t.me/learningnets

# LTE




https://t.me/learningnets

# LTE

From the classical/traditional network engineering point of view, PGW is the mobile networking node which is connected to the IP Core/backbone network

Mobile broadband comes with data cap. This limits the usability in home networks for using HD quality videos

https://t.me/learningnets

# LTE

Ultra HD video consumes so much bandwidth

Although the cost of having frequency and deploying small cells to provide enough bandwidth puts pressure on mobile carrier, broadband penetration of mobile is increasing very rapidly

https://t.me/learningnets

## 4 Key Design Principles of Mobile Networks

Not all Mobile network systems are designed in Cellular based, but as of 2018, almost all Mobile network system designs are based on Cellular architecture

Let me explain what would be the other deployment option for mobile networks, other than cell based systems and then we will highlight the 4 main characteristics of cell based mobile networks

https://t.me/learningnets

# 4 Key Design Principles of Mobile Networks

With the previous cellular systems design, mobile network operators used to place their radio transmitters at the tallest buildings in the area which they wanted to provide coverage

Single, very high-power transmitters were used to cover very large geographic areas

https://t.me/learningnets

# LTE

Design principles of cell based mobile phone networks, are as follow:

1. Cellular Based Mobile Networks: With the cell based telephone systems, so many low- power, small coverage area transmitters are used instead of a single, powerful, monolithic transmitter to cover a wide area

https://t.me/learningnets

# LTE

2. The second design principle which is frequency reuse, takes into account the fact that cellular telephony, like all radio-based services, has been allocated a limited number of frequencies by the regulators

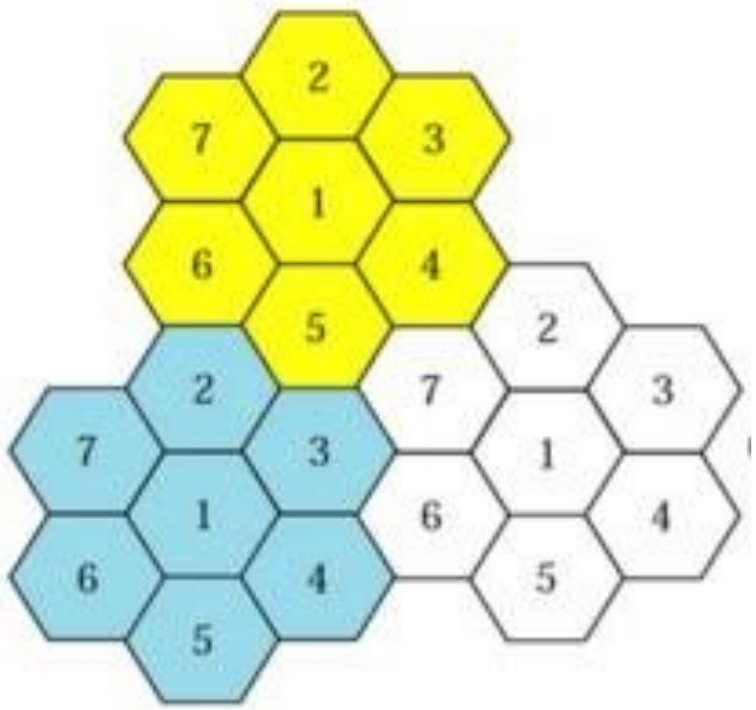
Frequency reuse is commonly used by the wireless Internet Service Providers as they have serious frequency limitation issues

By using geographically small, low-power cells, frequencies can be reused by non-adjacent cells

https://t.me/learningnets

# LTE

a. Reuse factor of 4




b. Reuse factor of 7

https://t.me/learningnets

3. Cell Splitting: When usage areas become saturated by excessive usage, cells can be split

This is the third design principle of cellular based mobile phone systems

https://t.me/learningnets

# LTE

When the network operator realizes that the number of callers are refused because of congestion reaching a critical level, they can split the original cell into two cells, by installing a second site within the same geographical area that uses a different set of non- adjacent frequencies

This has an extended impact; as the cells become smaller, the total number of calls that the aggregate can support increases dramatically

https://t.me/learningnets

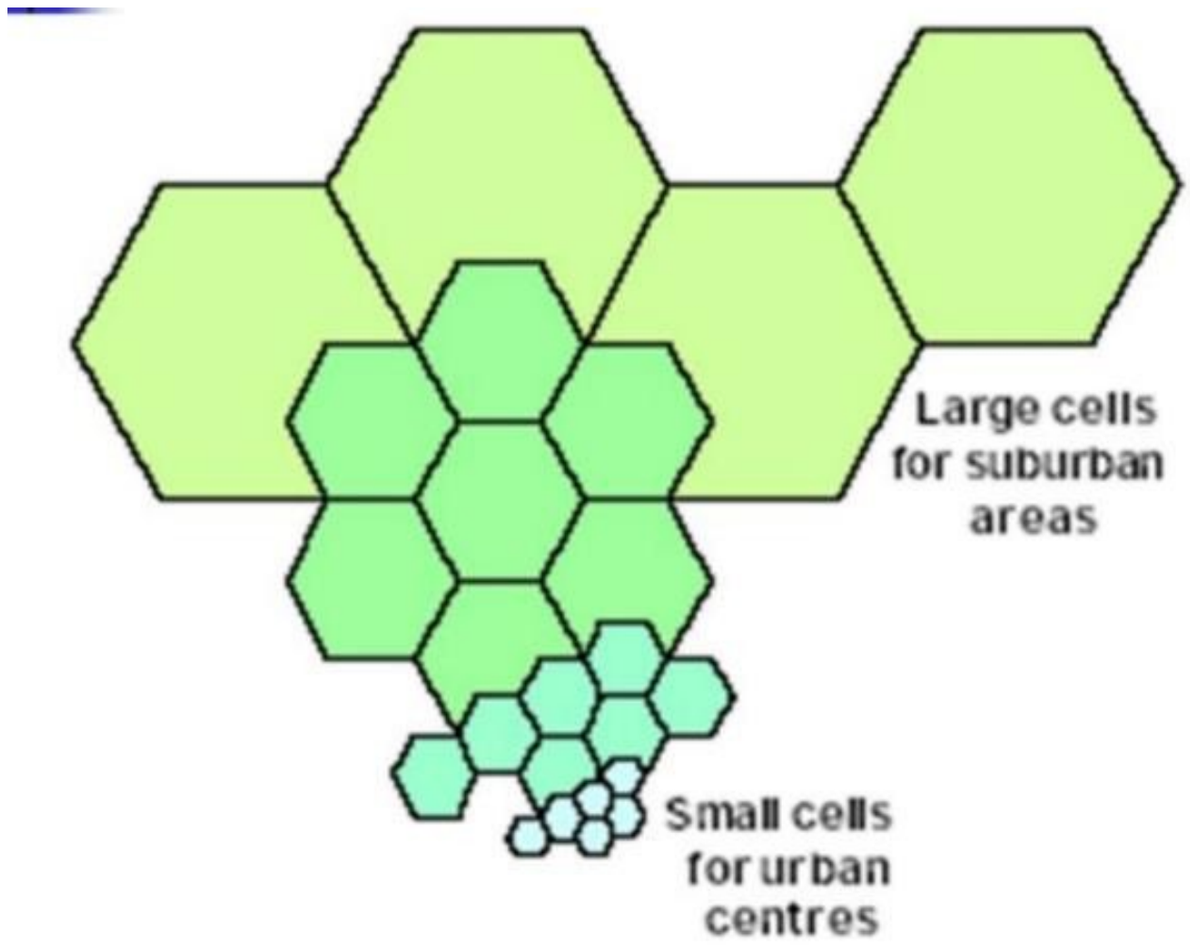
# LTE

The smaller the cell, the larger the total number of simultaneous calls that can be accommodated

Of course, this also causes the cost of deployment to increase exponentially

https://t.me/learningnets

# LTE




https://t.me/learningnets

# LTE

4. The Fourth and last design principle of cell based phone systems is cell-to-cell handover, which simply means that the cellular network has the ability to track a call as the user moves across a cell

When the signal strength detected by the current cell is perceived by the system to be weaker than the signal strength of the second cell, the call is handed over to the second cell

https://t.me/learningnets

# LTE

Ideally, the user doesn’t sense the handover

Cell handover and other cellular system capabilities are under the central control of a mobile telephone switching office (MTSO, sometimes called as MSO – Mobile Switching Office or MSC – Mobile Switching Center)

https://t.me/learningnets

So far, we have seen a brief overview of the architecture of Service Provider Access types and showed various different types of Broadband Services

Then various broadband access technologies were discussed and the major features of DSL, HFC, PON, Wi-Fi, WiMAX, Cellular and Satellite networks were presented in detail

https://t.me/learningnets

DSL can offer data rates over 10 Mb/s within a short distance. Efforts to develop next-generation DSL focus on increasing data rates and transmission distance

With DSL, voice and data can be supported by a single phone line

In the past decade, DSL has become one of the dominant broadband access technologies worldwide

However, TV broadcasting and IP TV are still a technical challenge for DSL networks

With using VDSL and advanced data compressing techniques, transferring video over IP may be offered in the near future

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

Traditionally, Cable TV companies offered analog TV broadcasting. With the development of many protocols and technologies such as DOCSIS, HFC, CMTS, Cable Modem, etc. Broadband Internet access is provided to subscribers

Cable Access Providers have added VoIP and digital TV over HFC networks. Further development has extended the cable plant beyond 1 GHz, and higher data rates are the main focus of HFC networks

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

As optical fibers can provide essentially unlimited bandwidth, PON (Passive Optical Network) is considered the most promising wired access technology.

Most of the FTTx deployment is based on PON today

https://t.me/learningnets

# Access Networks and Broadband Services Summary

In addition to wired broadband solutions, many wireless technologies have been developed to provide broadband Internet access, such as free-space optical communications, Wi-Fi, WiMAX, cellular and satellite networks

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

Free-space optical communications can support gigabit per second data rates and a transmission distance of a few kilometers, but the atmospheric effects impose severe constraints for free-space optical communications

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

Wi-Fi is widely used for wireless local area networks with transmission distances up to a few hundred meters

With a mesh topology, Wi-Fi networks can support extended reach and broadband Internet access

WiMAX can support wireless access over 10 km, but it requires higher transmitted power and the data rate is lower than in Wi-Fi networks

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

Cellular networks are used primarily for mobile voice communication. With digital encoding technologies, data transmission service is added to the Cellular networks

Currently 5G is developing in many countries and commercial deployment is expected in 2020

https://t.me/learningnets

## Access Networks and Broadband Services
## Summary

Satellite networks are used primarily for direct video distribution, but data service over satellite is covering a large geographical area

The main disadvantage of satellite communication in the past was limited data rates

Many new satellite systems have been launched in recent years and it is possible to receive 10s of Mb/s of Broadband access over satellite with very low latency due to LEO (Low Earth Orbit) and MEO (Medium Earth Orbit) satellite systems

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

The advantages of wireless technologies are mobility, scalability, low cost, and ease of deployment

Except for free-space optical communication, other wireless technologies use RF or microwave frequencies

Network capacity and reliability with Wireless are much lower compared to wired access networks

https://t.me/learningnets

# Access Networks and Broadband Services
## Summary

Emerging multimedia applications demand broadband access networks, and a number of wired and wireless broadband access technologies have been developed over the past decade

By solving the last mile bottleneck problem, it will be possible to use a variety of new applications

https://t.me/learningnets

# Fiber Optic

Fiber-Optic cables carry information between two places using entirely optical (light-based) technology

A fiber-optic cable is made up of incredibly thin strands of glass or plastic known as optical fibers; one cable can have as few as two strands or as many as several hundreds

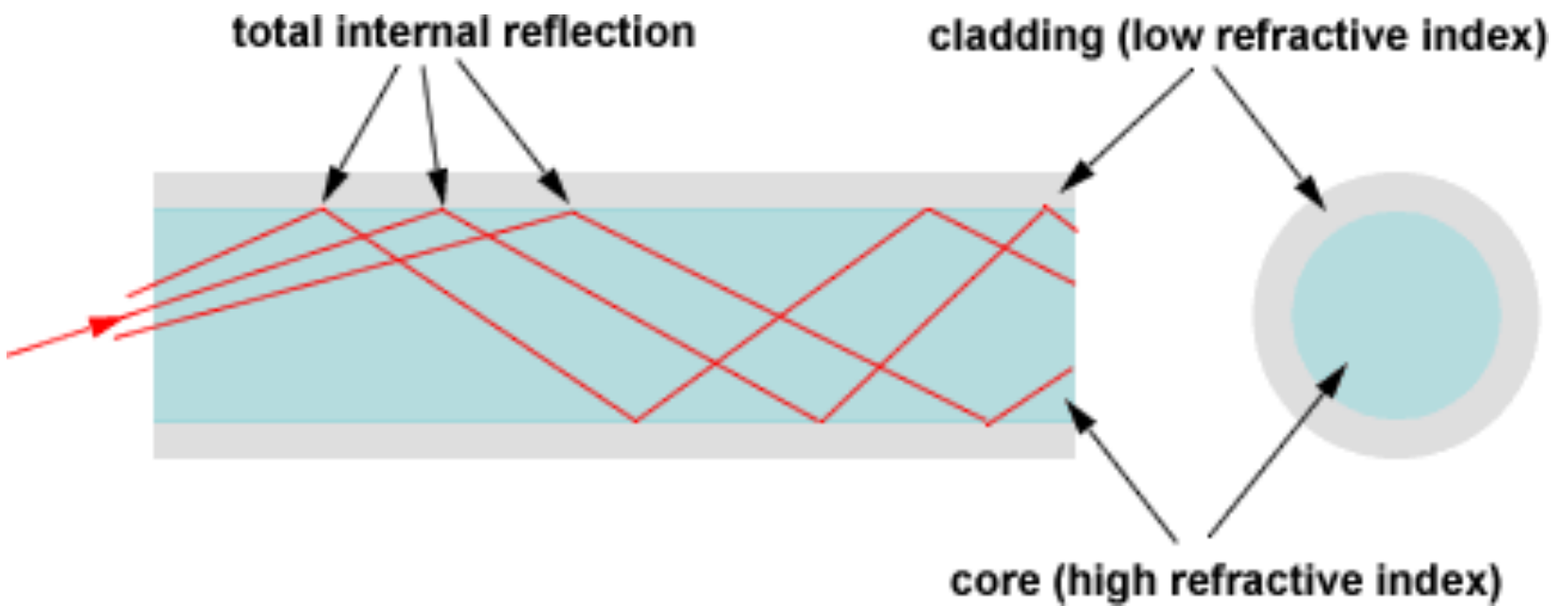
Each strand is less than a tenth as thick as a human hair and can carry something like 25,000 telephone calls, so an entire fiber-optic cable can easily carry several million calls

https://t.me/learningnets

# Total Internal Reflection

Fiber technology works based on total internal reflection

The basic functional structure of an optical fiber consists of an outer protective cladding and an inner core through which light pulses travel




https://t.me/learningnets

# Total Internal Reflection

The difference in refractive index of the cladding and the core allows total internal reflection in the same way as happens at an air-water boundary, as shown in the figure below

Thus, fiber optics use total internal reflection

https://t.me/learningnets

# Fiber Optic Cable Installation

Fiber can be installed in three ways:

- Aerial fiber optic cable installation
- Direct burying
- Sub marine fiber optic cable installation

https://t.me/learningnets

# Fiber Optic Cable Installation

https://t.me/learningnets

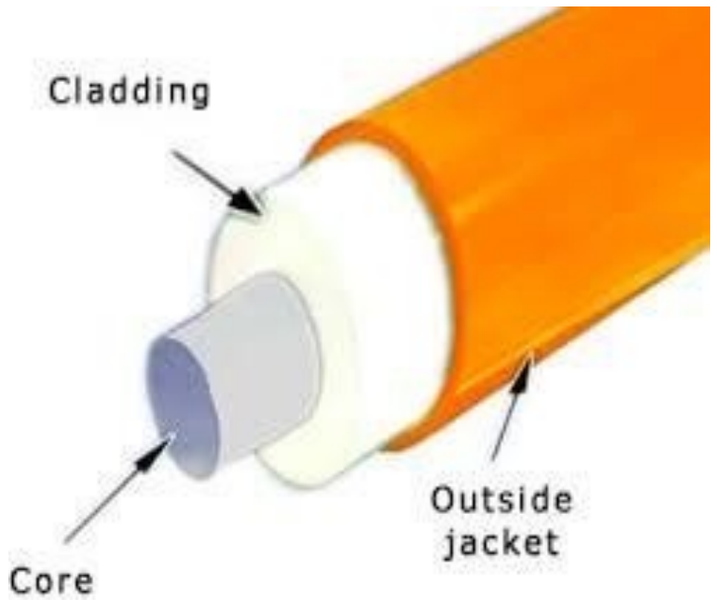
# Fiber Optic Cable Types

## Single Mode Fiber

The simplest type of optical fiber is called single-mode

It has a very thin core about 5-10 microns (millionths of a meter) in diameter

In a single-mode fiber, all signals travel straight down the middle without bouncing off the edges. Cable TV, Internet, and telephone signals are generally carried by single-mode fibers




https://t.me/learningnets

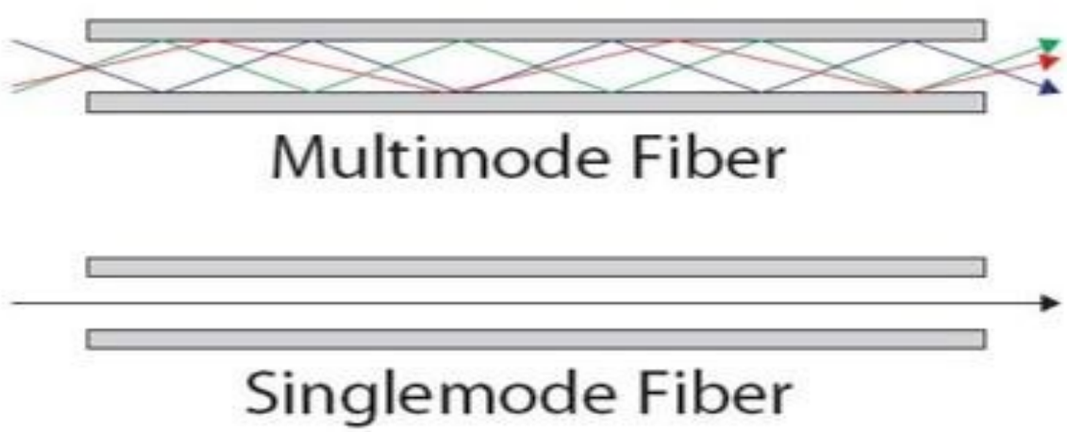
## Fiber Optic Cable Types

# Multimode Fiber

Each optical fiber in a multi-mode cable is about 10 times bigger than one in a single-mode cable

This means light beams can travel through the core by following a variety of different paths, in other words, in multiple different modes

Multi-mode cables can send information only over relatively short distances




https://t.me/learningnets

# Microwave

Within the broader spectrum of radio frequency (RF) communications, point-to-point communications are usually carried out using microwave frequencies between 1 GHz and 100 GHz along line-of-sight paths called links

These frequencies and their propagation characteristics allow the transmission of vast amounts of data between remote communication sites without the need to lay cables between them

https://t.me/learningnets

# Microwave

The first practical application of microwaves in a communication system took place more than 80 years ago

In the 1930s, an experimental microwave transmission system was used to connect United Kingdom with France—bridging the English Channel without cables

https://t.me/learningnets

# Microwave

Microwave is commonly used in the “Backhaul” part of the networks

When Microwave is used for backhaul purpose in Mobile Operator networks, it is called Microwave backhaul

Before we explain Microwave Backhaul further, it is important to share the definition of “Backhaul”

https://t.me/learningnets

# Microwave

The term backhaul is often used in telecommunications and refers to transmitting a signal from a remote site of the network to another site, usually a central one

Backhaul usually implies a high- capacity line, meaning high-speed lines capable of transmitting high bandwidth at very fast speeds

Carrying the data, voice or video traffic from Cell Site (Access Network in the Mobile Operators) to Core network is known as Backhauling the traffic.

https://t.me/learningnets

# Microwave

Many Mobile Operators transitioned their networks from 2G/3G radio access to 4G access technologies

These networks needed to deliver far more bandwidth to support richer applications while demand for backhaul capacity has been growing since the 2G era

https://t.me/learningnets

# Microwave

Microwave systems have emerged as a good way to deliver capacity with a better pricing compared to fiber option

The mobile backhaul infrastructure carries voice and data communication traffic from the Access network (cell sites) to the Core network

https://t.me/learningnets

# Microwave

Most cell sites are connected to the Core network via 100Mbps to 1Gbps as of 2019 in the Mobile Operator networks

Microwave has been used by many Mobile Operators for years, but when it comes to speed comparison, which one is faster?

https://t.me/learningnets

# Microwave

Question is not asking about the bandwidth capacity, it is the speed comparison

When it comes to bandwidth, Fiber is an obvious answer

But let’s have a look at speed

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

As it is mentioned above, most of the Mobile (Cellular) networks use microwave to connect their cell towers to backhaul networks

For mobile operators, a reason using microwave is not the speed which microwave provides

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

It is used to connect their remote sites (Rural areas), because microwave is a faster and cheaper deployment option compared to fiber

(Reliability and security is a different aspect which is out of the scope of this book)

When more capacity (bandwidth) is required, fiber becomes more economical

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

The actual cost of fiber deployment is the laboring cost

For example, digging a trench and laying the fiber, getting the required permissions from the land owners and from the municipalities

You have to dig a trench that’s hundreds (or thousands) of kilometers long, or lease access to ducts that have already been laid by infrastructure companies

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

Geography of the land is very important for fiber deployments

For example, when faced with a mountain or river, do you go straight across at great expense, or do you make a diversion to the nearest bridge or tunnel?

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

Combine all of these factors and you'll understand why most of the world's terrestrial fiber networks are deployed alongside existing roads and railways

Let's go back to our initial discussion. Microwave or Fiber, which one is faster?

In other words, which one has lower latency? (More bandwidth means more people can send a data over the link but faster means lower latency)

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

A radio signal travels through air at just under the speed of light (299,700km per second)

Through a glass or plastic fiber, where light has to bounce along the refractive index rather than travel in a straight line, the speed of light is reduced to around 200,000km/s

This fact is known by transport network engineers but not by most of the network engineers

Why is this important? Let’s say we have two buildings with 200km distance between them

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

With microwave (radio signal through air), latency between the two buildings would be 1,064 milliseconds or 1064 microsecond

With fiber, the latency between the buildings would be 1.594 millisecond or 1594 microsecond. But really, is 500 microsecond difference critical? Yes, in fact for some business, it is a lot

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

Have you ever heard HFT (High Frequency Trading) networks? In HFT business, end to end latency is called tick-to-trade latency

Tick is market data and trade is a buy/sell or any other type of order (Tick data flows from the stock exchanges to the HFT Company, Trade data flows from the HFT to stock exchanges)

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

The whole business relies on having a faster network than any other company in the market

Having an information from the exchanges, carrying it through the network to your servers, processing it and taking an action based on “algo” (Algorithm but in HFT environment there are tons of jargon) all happens under less than 100 microseconds

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

In fact, some HFT vendors claim that they provide tick-to-trade latency as less than 50 microseconds

Transport medium such as microwave and fiber contribute to propagation and serialization delays

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

Propagation delay is what I explained above, how long it would take to carry data between point A and B. Microwave is the winner in our comparison

Serialization delay means, how long it would take to place a data into a wire from the networking device

It is a known fact that 10G fiber has much lower serialization delay than 1G fiber

https://t.me/learningnets

# Microwave or Fiber, which one is faster?

To be more precise, for 128- byte data, 1G (Gigabit Ethernet) serialization delay is 1 microsecond, 10G serialization delay is 0.1 microsecond

It seems that this is not a large difference, but you should consider that the delay would be accumulated at every port

In general, companies that desire to deploy low-latency infrastructures would implement 10 Gigabit Ethernet wherever possible

Especially in the HFT (High Frequency Trading) business

https://t.me/learningnets

# SDH/SONET

Synchronous Optical Networking (SONET) and Synchronous Digital Hierarchy (SDH) are standardized protocols that transfer multiple digital bit streams synchronously over optical fiber

https://t.me/learningnets

# SDH/SONET

SONET and SDH are essentially the same, SONET is used mostly in U.S and Canada, SDH is used in rest of the world

They were designed to support real time applications; thus it was made for carrying circuit mode transport such as E1, T1, DS3

Today, both SONET and SDH can transport ATM, Ethernet and IP as their client protocol

https://t.me/learningnets

# SDH/SONET

| Synchronous Digital Hierarchy (SDH) Level | SONET | Speed (Mbps) |
|---|---|---|
| STM-0 | STS-1 (OC-1) | 51.84 |
| STM-1 | STS-3 (OC-3) | 155.52 |
| STM-2 | STS-6 (OC-6) | 311.04 |
| STM-3 | STS-9 (OC-9) | 466.56 |
| STM-4 | STS-12 (OC-12) | 622.08 |
| STM-8 | STS-24 (OC-24) | 1244.16 |
| STM-16 | STS-48 (OC-48) | 2488.32 |
| STM-32 | STS-96 (OC-96) | 4976.64 |
| STM-64 | STS-192 (OC-192) | 9953.28 |
| STM-256 | STS-768 (OC-768) | 39,813.12 |

https://t.me/learningnets

# SONET/SDH is Mostly Obsolete Now!

With SONET/SDH, the maximum speed is 40Gbps; the reason is that the development of these technologies has stopped. Both SONET/SDH are now obsolete technologies

There are multiple reasons, but mainly these technologies don’t carry multiple wavelengths over single optical fiber and cost of the equipment is too much compared to Ethernet

https://t.me/learningnets

# SONET/SDH is Mostly Obsolete Now!

Today, Ethernet is ubiquitous and used in LAN, MAN and WAN, thus mass production of the Ethernet equipment made Ethernet products so cheap

With SDH you can only transmit a single wavelength per fiber and that is typically either 1310nm (short-haul) or 1550nm (long-haul)

This is also referred to as black & white optics

https://t.me/learningnets

# SONET/SDH is Mostly Obsolete Now!

SDH does not scale very well and begins with the basic block of an STM-1 and scales up by a factor of 4 (i.e. STM-4, STM-16, STM-64, and STM-256)

SDH was deployed in metro areas and can have a maximum reach of approximately 80km using 1550nm optics

SONET/SDH can be carried over WDM but not vice versa

https://t.me/learningnets

# WDM

WDM stands for Wavelength division multiplexing

There are two different WDM mechanisms, DWDM and CDWM. WDM is considered Layer 0 by some transmission people and SONET/SDH as Layer 1

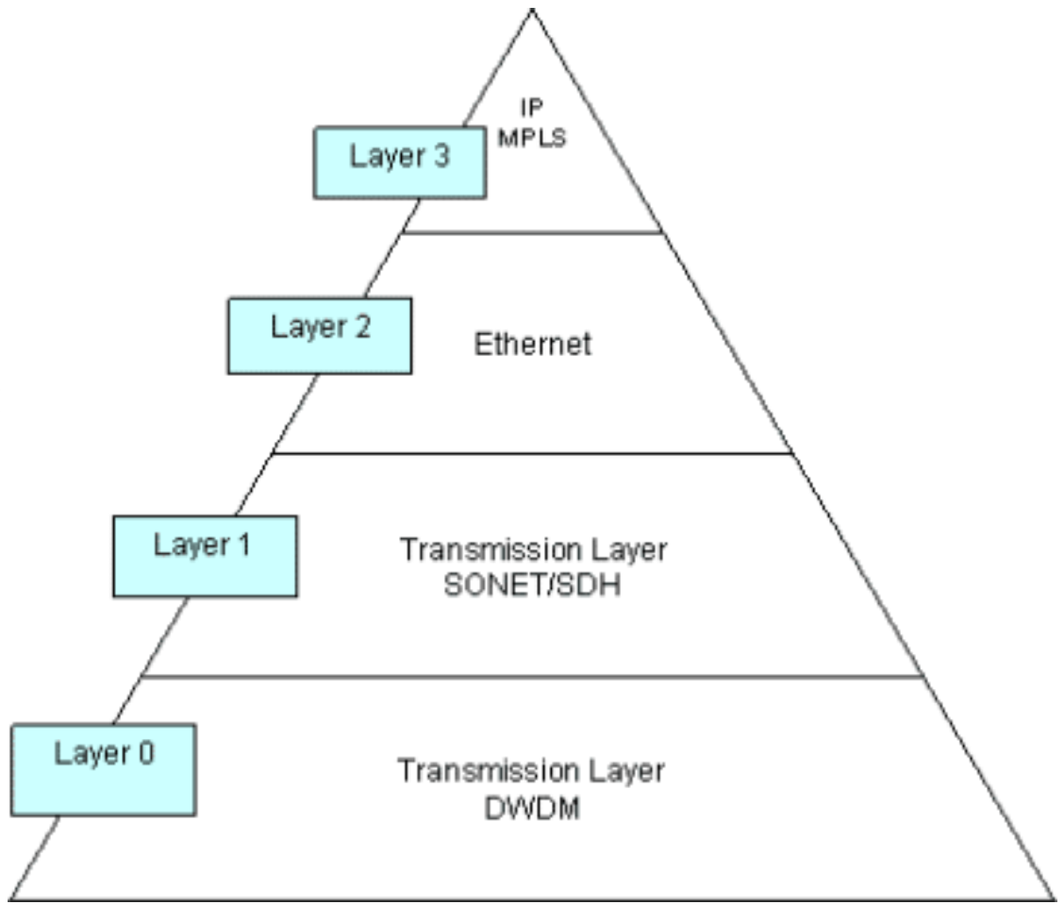
Some transmission people consider WDM as Layer 1 and SONET/SDH as Layer 1.5, but keep in mind that SONET/SDH can be a client of WDM but not vice versa

https://t.me/learningnets

# WDM

DWDM allows you to transmit multiple wavelengths on a single fiber

and as the name suggests, it utilizes wave division multiplexing




https://t.me/learningnets

# WDM

DWDM can be deployed in ultra-long haul, long-haul, regional, and metro areas

DWDM also uses amplification to reach long distances and can be an expensive technology to deploy

https://t.me/learningnets

# WDM

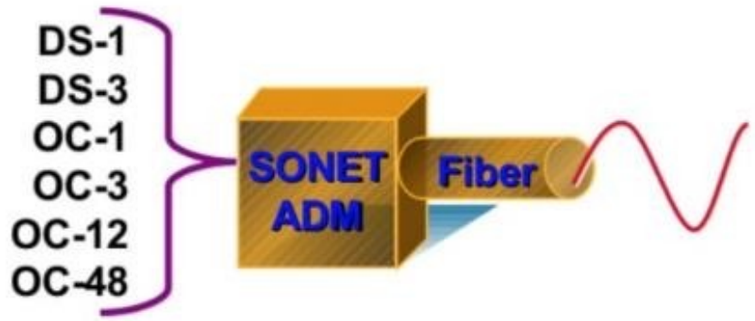
On the other hand, CWDM which is a metro technology, multiplexes a maximum of 8 wavelengths per fiber and does not require amplification (i.e. much cheaper than DWDM)

WDM is used to transport SDH/Ethernet/IP between regions or cities or to aggregate traffic where large bandwidth is required

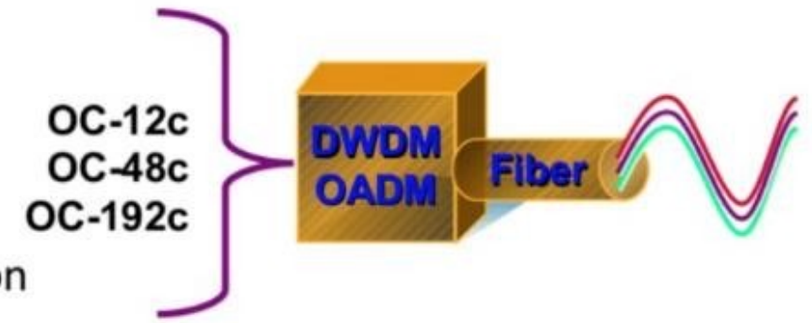
https://t.me/learningnets

# WDM

- TDM (SONET/SDH)
  - –Take sync and async signals and multiplex them to a single higher optical bit rate
  - –E/O or O/E/O conversion




- WDM
  - –Take multiple optical signals and multiplex them onto a single fiber
  - –No signal format conversion

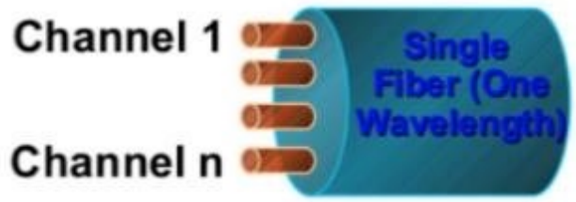


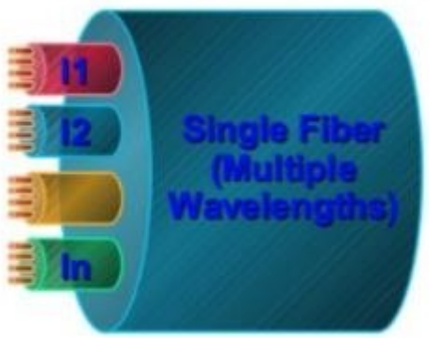

https://t.me/learningnets

# WDM

## TDM vs. WDM

- Time division multiplexing
  - –Single wavelength per fiber
  - –Multiple channels per fiber
  - –4 OC-3 channels in OC-12
  - –4 OC-12 channels in OC-48
  - –16 OC-3 channels in OC-48
- Wave division multiplexing
  - –Multiple wavelengths per fiber
  - –4, 16, 32, 64 wavelengths per fiber
  - –Multiple channels per wavelength



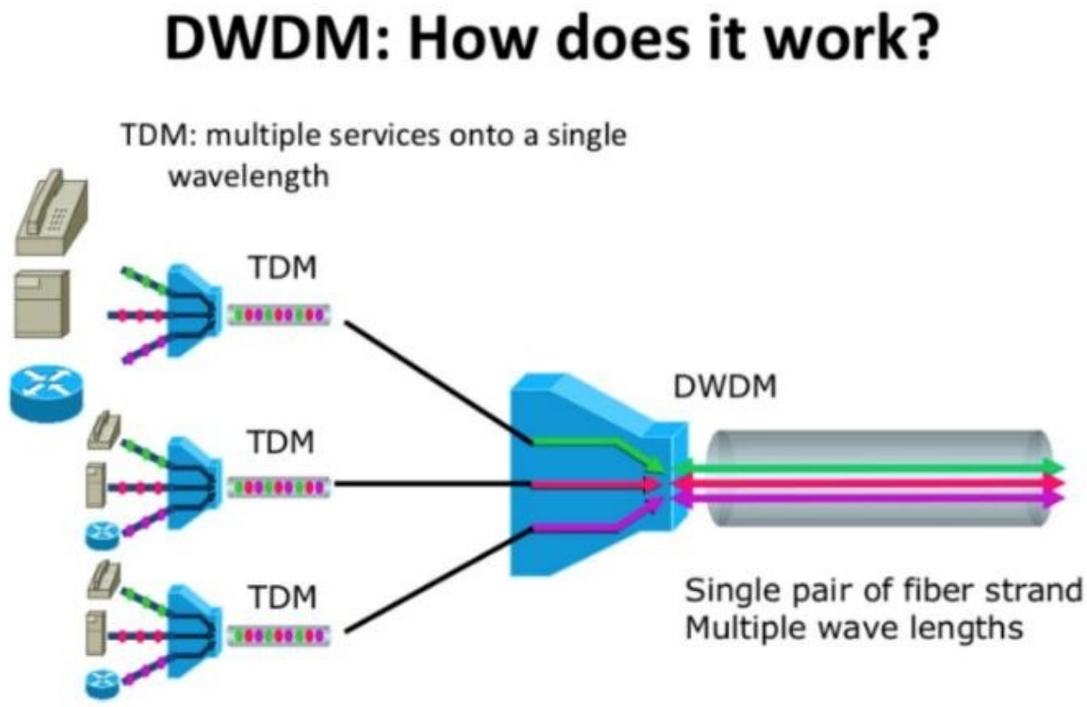




https://t.me/learningnets

# DWDM

Dense Wavelength Division Multiplexing (DWDM) is an optical multiplexing technology used to increase bandwidth over existing fiber networks

DWDM works by combining and transmitting multiple signals simultaneously at different wavelengths on the same fiber

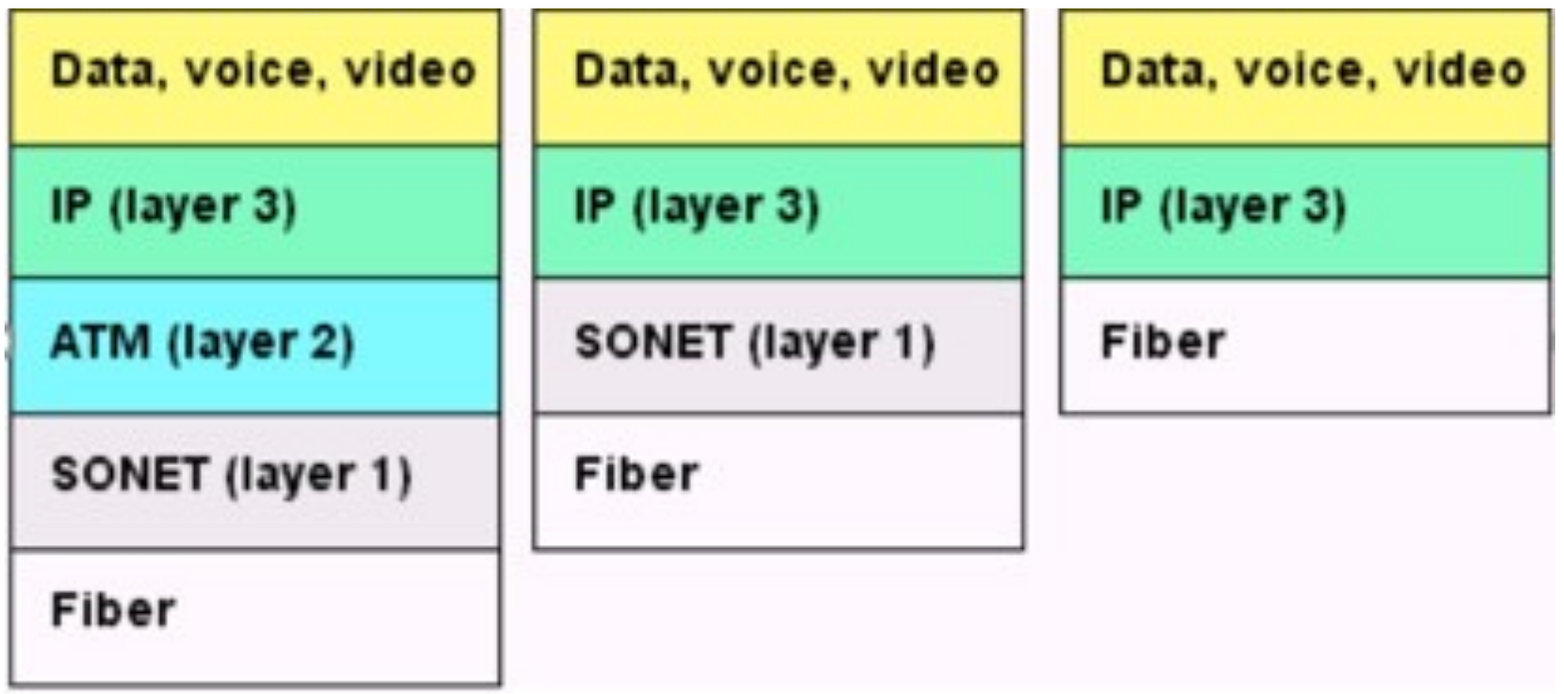



https://t.me/learningnets

# IP Transport Evolution on Wide Area Network

The evolution started over ATM, then with POS, and today IPoDWDM (IP over DWDM) is very common

Less number of layers provides lower OPEX and CAPEX, also provides better management and less complexity




https://t.me/learningnets

# Dark Fiber

Dark fiber is the fiber optic infrastructure that’s not yet in use by a service carrier or a provider

It is an installation of cables currently lying dormant, ready to be connected at some time in the future

The name “dark” from dark fiber comes from the technology that the literally lacks the lasers that send light through the cables

https://t.me/learningnets

# Purchasing and Leasing Capacity on Fiber Links

Service Providers can build their own fiber network or buy it from other providers which are generally called Carriers

Buying a fiber can be in two ways, leasing or IRU (Indefeasible rights of use) based

https://t.me/learningnets

# Indefeasible right of use (IRU)

IRU is a permanent contractual agreement that cannot be undone, between the owners of a cable and a customer of that cable system. Indefeasible means “not capable of being voided or undone”

The customer purchases the right to use a certain amount of capacity of the fiber system for a specified number of years

https://t.me/learningnets

# Indefeasible right of use (IRU)

Service Providers use the transport network of other carriers. This is very common; in fact, even the biggest networks use other carrier’s transport/transmission infrastructure, especially outside of their main location

https://t.me/learningnets

# Indefeasible right of use (IRU)

For example, an operator might provide services mainly in U.S, but wants to extend its network to Europe. Instead of setting up a fully-fledged telecom environment in Europe to provide a service, (for example to Business and Residential customers), one option is to use local carrier networks in Europe

https://t.me/learningnets

# Indefeasible right of use (IRU)

IRU is a permanent contractual agreement that cannot be undone between the owners of a cable and a customer of that cable system

Cable is mostly a fiber cable as fiber can carry more data than any other type of media

Customers who purchase IRU, can lease the capacity to other companies

https://t.me/learningnets

# Indefeasible right of use (IRU)

**Let me give you an analogy:**

Think of it in this way, if you are renting an apartment, you sign a contract with the Landlord as a tenant

You cannot rent that apartment to someone else

This is similar to leasing. But if you are the landlord, you can rent it to anyone you want

This is an example for Indefeasible Right of Use based agreement

https://t.me/learningnets

# IRU vs. Leasing a Fiber

IRU contracts are almost always long term, such as 10 to 20 years (Cable lifetime is generally considered as 25 years)

Leased fiber doesn’t have to be a long-term contract

IRU period can start from 10 years. The currently most common version of IRU contract is 15 years

https://t.me/learningnets

# IRU vs. Leasing a Fiber

The most common leased service is IPLC which is International Private Leased Circuit

IPLC can be half circuit or full circuit. IPLC unlike IRU doesn’t dictate the buyers to pay the cost of fiber upfront

https://t.me/learningnets

# IRU vs. Leasing a Fiber

IPLC is not a prepaid service

Leasing is very flexible but IRU can be very cost effective

IRU gives the purchaser the right to use some capacity on telecommunications cable system, including the right to lease that capacity to someone else

https://t.me/learningnets

# Should smaller companies purchase an IRU based fiber?

Smaller companies that need a leased line between 2 different countries, for example between London and New York do not buy an IRU

They receive capacity from a telecommunications company

https://t.me/learningnets

# Should smaller companies purchase an IRU based fiber?

Those telecom companies may also have received a larger amount of capacity from another company (and so on), until at the end of the chain of contracts there is a company that has an IRU, or wholly owns a cable system

https://t.me/learningnets

# Carrying Network Traffic between Countries

How Internet Traffic is carried between the Countries? Internet traffic is carried between the countries mostly in two ways; terrestrial and submarine cables

https://t.me/learningnets

# Carrying Network Traffic between Countries

Satellite can be used as backup, but limited capacity and latency makes it unattractive and as a last resort, for example U.S FCC (Federal Communication Commission) indicates that satellites account for just 0.37 percent of all U.S. international capacity

Terrestrial cables are used between the landlocked countries in general

https://t.me/learningnets

# Terrestrial Fiber Optic Cables

There are 49 landlocked countries (According to Wikipedia) in the World, largest is the Kazakhstan and the smallest landlocked country is Vatican

Terrestrial cables are only used on land and a cross border agreement is signed between the countries

https://t.me/learningnets

# Terrestrial Fiber Optic Cables

Some terrestrial cables pass several countries

Terrestrial cables carry much higher capacity compared to submarine cables

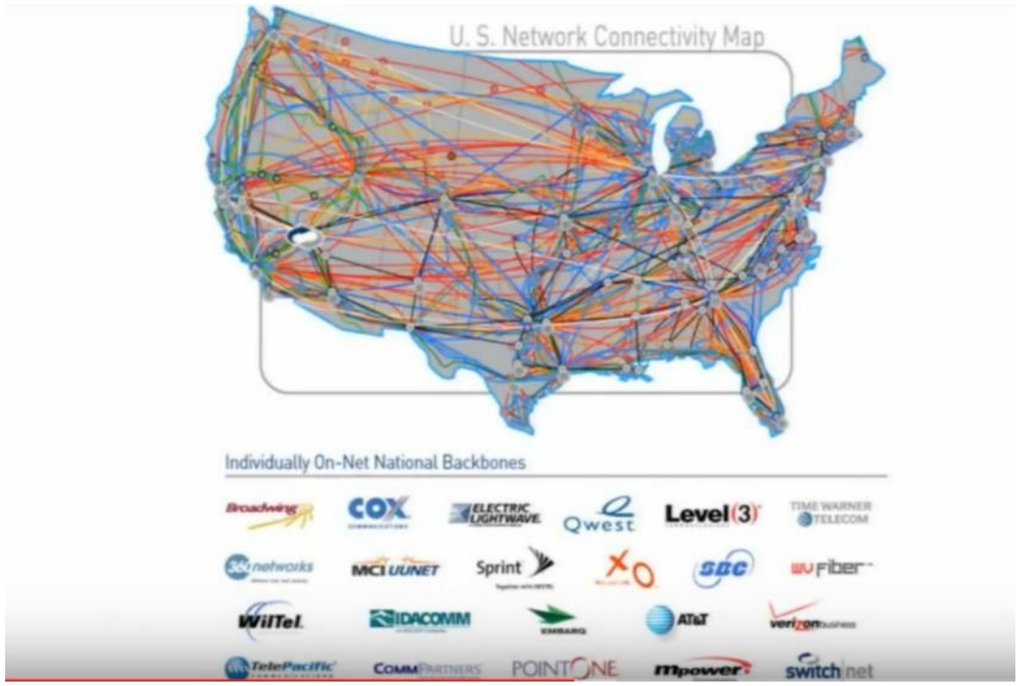
Terrestrial cables are not only used for inter-country communication but commonly used inside the country for Metro Access and inter-city connections as well

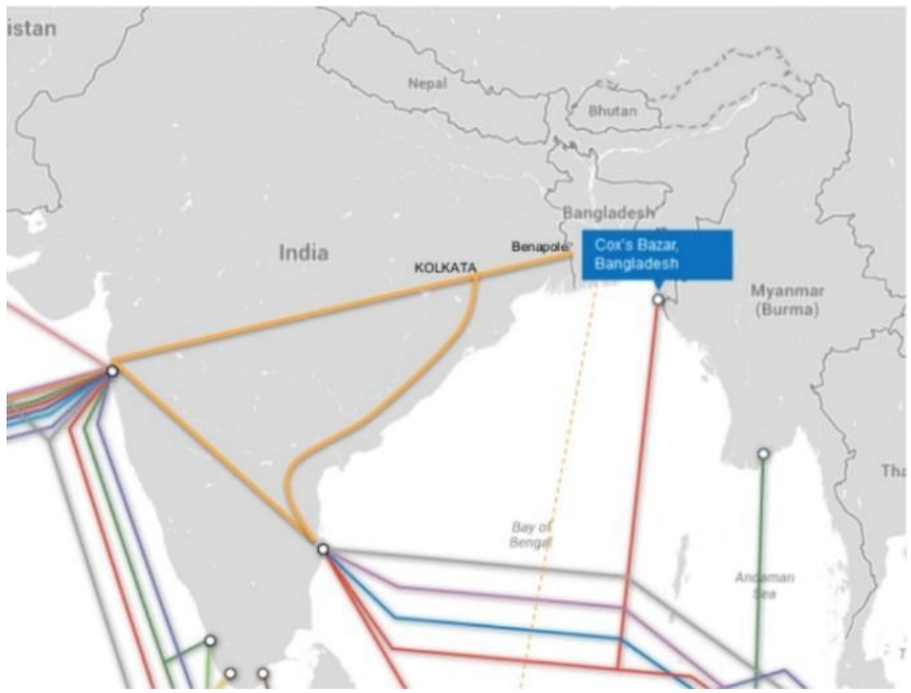
https://t.me/learningnets

# Terrestrial Fiber Optic Cables

Terrestrial fiber cables can connect countries even when they have access to submarine cables

For example, Bangladesh operators are using newly deployed Bangladesh to India route, in addition to SEA-ME-WE 4 and SEA-ME-WE-5 Submarine cable systems




https://t.me/learningnets

# Submarine Fiber Optic Cable Systems

It is also called as subsea or undersea cable as well

Sub marine cables use fiber optic infrastructure

There are approximately 450 submarine cables in service as of 2018

Cables are typically as wide as garden hose but of course actual components which carries the light is a diameter of a human hair

https://t.me/learningnets

# Submarine Fiber Optic Cable Systems

Cables are laid in deep sea directly on the ocean floor, but nearer to the shore

Cables are buried under the seabed for protection, that’s why people don’t see the cables when they go to the beach

https://t.me/learningnets

# Submarine Fiber Optic Cable Systems

As of 2022, more than 1.2 million kilometers of sub marine cables are laid in the world

Some submarine cables are quite short, like the 131 kilometer CeltixConnect cable, while other submarine cables are quite long such as 20.000 km Asia America Gateway Cable

https://t.me/learningnets

# Major route concept in sub marine fiber optic cable

Undersea cables are built between locations that have something “important to communicate.”

Europe, Asia, and Latin America all have large amounts of data to send and receive from North America

This includes Internet backbone operators ensuring that emails and phone calls are connected and also content providers who need to link their massive data centers to each other

This explains why you see so many cables along these major routes

https://t.me/learningnets

# Major route concept in sub marine fiber optic cable

For example, there’s just not much data that needs to go directly between Australia and South America

If that situation were to change, you can be sure someone would build a new cable between those two locations

https://t.me/learningnets

# Who builds sub marine fiber cables?

Cables were traditionally owned by telecom carriers who would form a consortium of all parties interested in using the cable

In the late 1990s, an influx of entrepreneurial companies built lots of private cables and sold off the capacity to users

https://t.me/learningnets

# Who builds sub marine fiber cables?

Both the consortium and private cable models still exist today, but one of the biggest changes in the past few years is the type of companies involved in building cables

https://t.me/learningnets

# Who builds sub marine fiber cables?

Content Providers such as Google, Facebook, Microsoft, and Amazon are major investors in new cables

The amount of capacity deployed by private network operators – like these content providers – has outpaced Internet backbone operators in recent years

Faced with the prospect of ongoing massive bandwidth growth, owning new submarine cables makes sense for these companies

https://t.me/learningnets

# Who uses submarine cables?

Users of submarine cable capacity include a wide range of types

Telecom carriers, mobile operators, multinational corporations, governments, content providers, and research institutions all rely on submarine cables to send data around the world

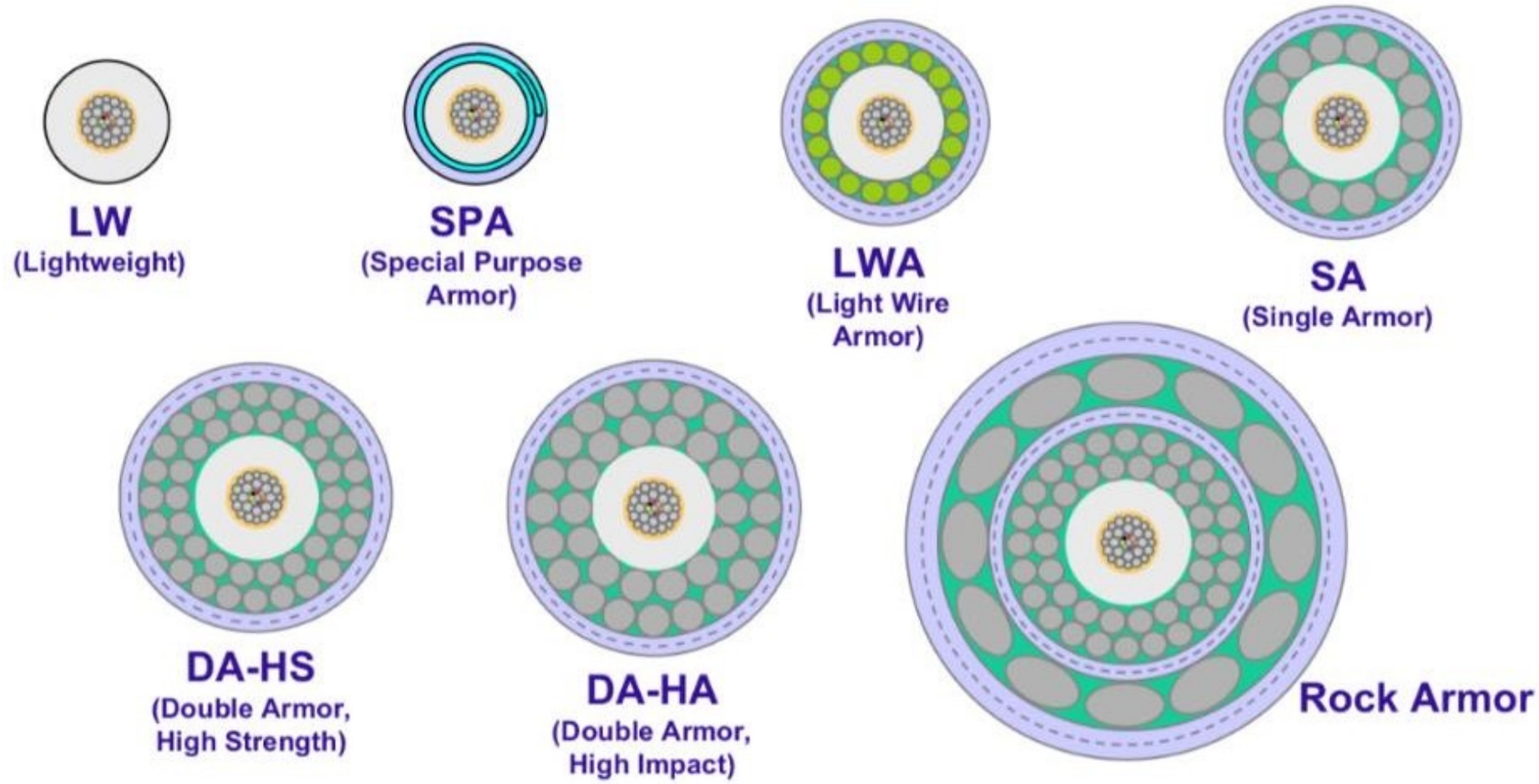
https://t.me/learningnets

# Who uses submarine cables?

Ultimately, anyone accessing the Internet, regardless of the device they are using, has the potential to use submarine cables

https://t.me/learningnets

# Submarine Cable Types




https://t.me/learningnets

# Submarine Cable Types

| Submarine Cable | Typical Application |
|---|---|
| Double Armored (DA) | Beach Manhole to 400 meter depth |
| Single Armored (SA) | 400 meter 900 meter depth |
| Light Weight (LW) | After 2000 meter depth |
| Special Application (SPA) | Between 900 meter to 2000 meter depth |
| Light-Wire Armored (LWA) | Buried |

https://t.me/learningnets

# What are the capacity of the Sub Marine Cables?

Cable capacity varies a lot and depends on the owners and the applications of the cable owners

https://t.me/learningnets

# What are the capacity of the Sub Marine Cables?

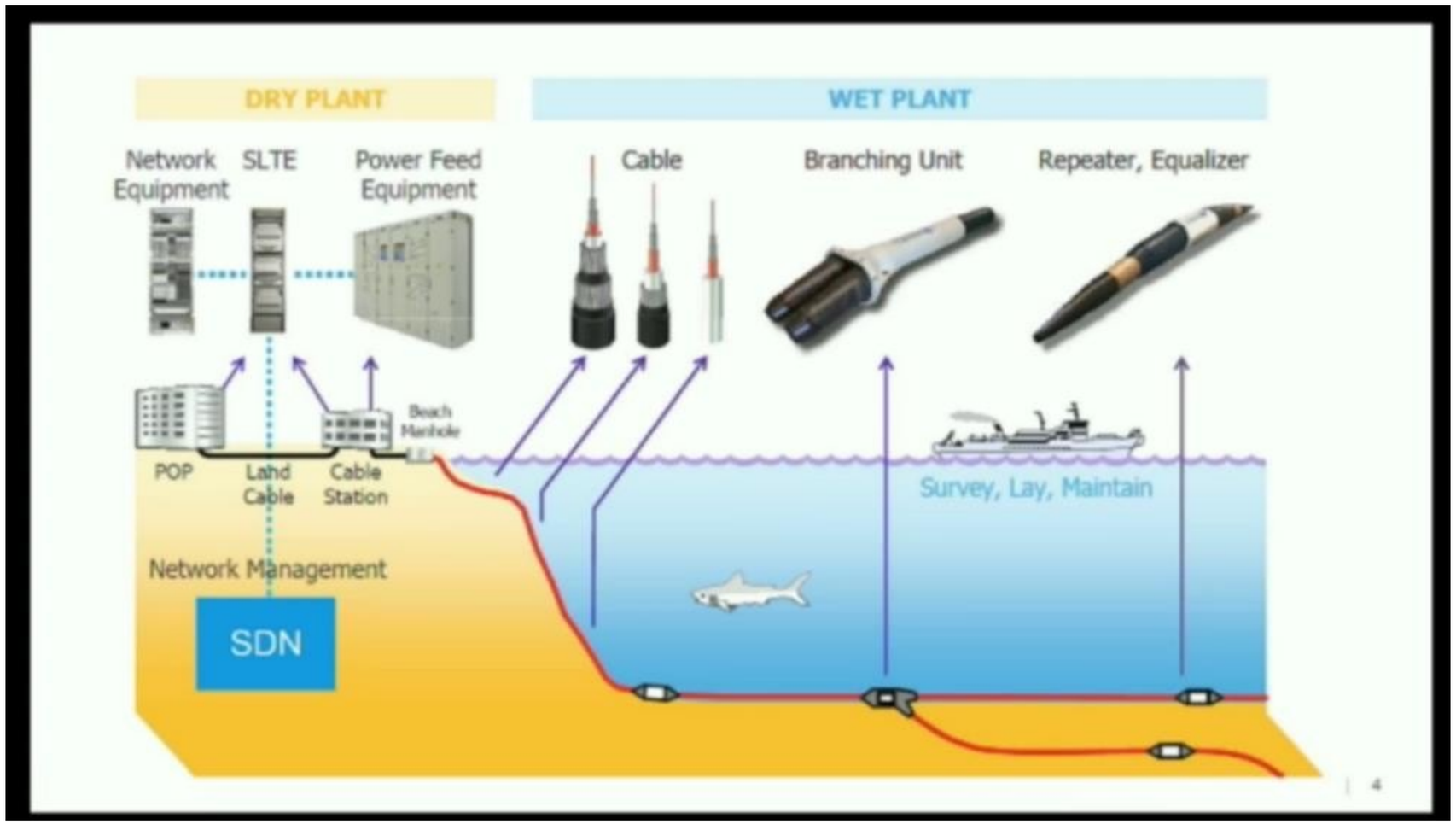
Some submarine cables are capable of carrying 160Tbps. There are two principal ways of measuring a cable's capacity; Potential capacity and Lit capacity.

- Potential capacity is the total amount of capacity that would be possible if the cable's owner installed all available equipment at the ends of the cable. This is the metric most often cited in the press.

- Lit capacity is the amount of capacity that is actually running over a cable. Cable owners rarely purchase and install the transmission equipment to fully realize a cable's potential from day one. Because these equipment's are expensive, owners instead prefer to upgrade their cable gradually, as customer demand dictates.

https://t.me/learningnets

# What are the capacity of the Sub Marine Cables?

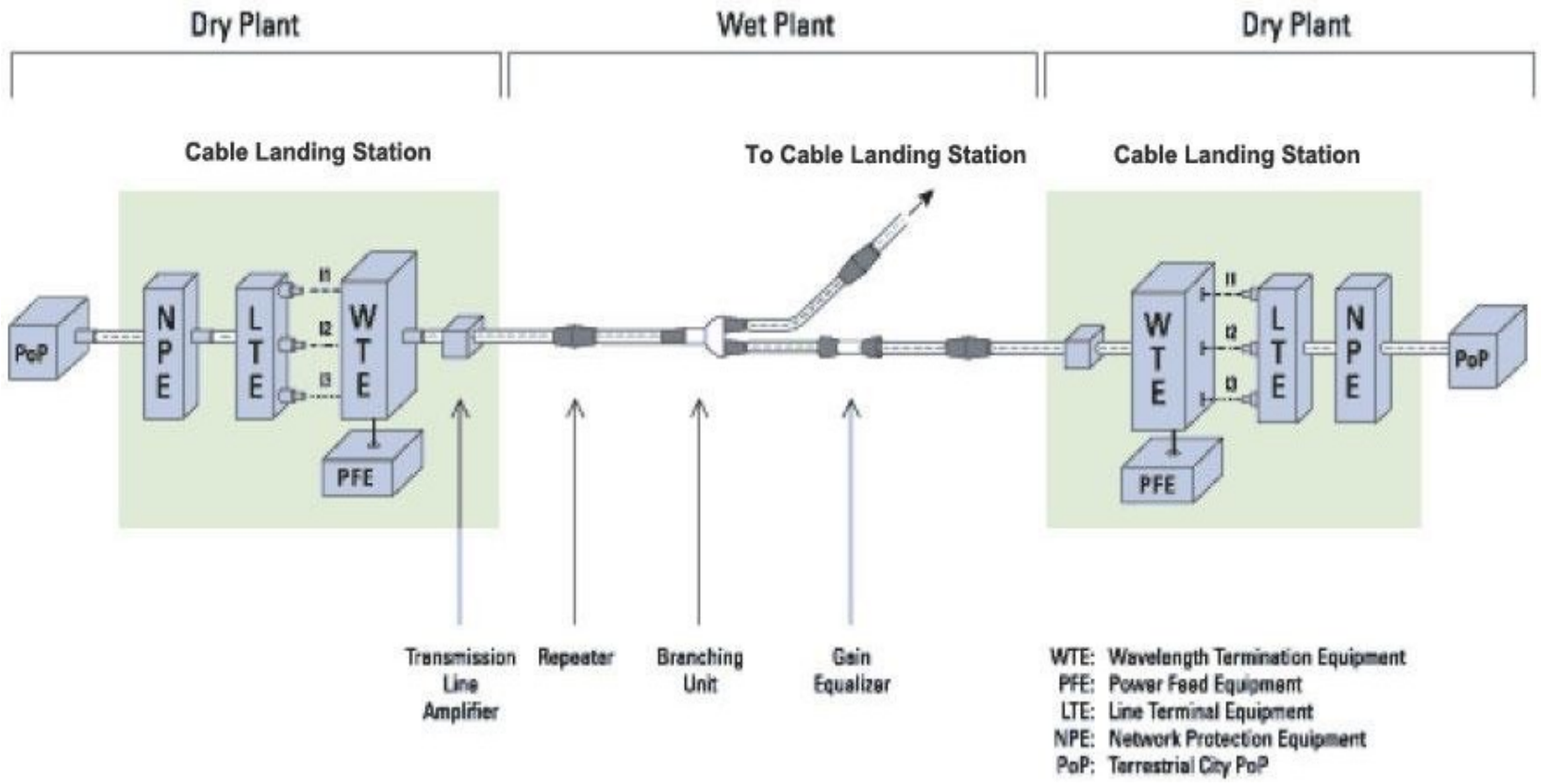



https://t.me/learningnets

# Cable Landing Point

It is a cable station for submarine cable system, also called landing station

Cable Landing Station is one important component of a submarine cable system which comprises of Wet Plant and Dry Plant




https://t.me/learningnets

# Cable Landing Point

The PFE and the SLTE of a submarine cable system are installed at the cable landing station. The cable landing station is usually carefully chosen to be in the following areas:

1. Areas that have little marine traffic to minimize the risk of cables being damaged by ship anchors and trawler operations
2. Sandy sea-floors so that the cable can be buried to minimize the chance of damage

https://t.me/learningnets

# Cable Landing Point

Cable Landing Station typically terminates undersea signals and connects the submarine cable with the domestic network

They are most often very close to beach. Their construction & features are very similar to other telecom offices, such as Central Offices

Large terminal stations cost $10-$15M

Very small stations can be less than $5M

https://t.me/learningnets

# Cable Landing Point

Some Cable Landing Station requirements can be listed as follows:

- Raised floor
- DC -48v power, with battery backup for at least 1 hour. Diesel Engine emergency backup
- HVAC to maintain room temperature between 22 - 24C
- Fire/smoke detection, with connection to emergency/control center
24-hour access for maintenance and repair
- Transmission Equipment
- Colocation for backhaul

https://t.me/learningnets

# Beach manhole

The precise point of connection is known as “the beach manhole,” a small underground place where the undersea cable is connected to dry land

https://t.me/learningnets

# Beach manhole

Cable is laid between Beach Manhole and the Landing Stations

Worldwide Submarine/Undersea cable map might be beneficial for some readers, you can find it on https://www.submarinecablemap.com/

https://t.me/learningnets

# Transport Technologies Module Summary

In this module, many transport network components which are used in the Service Provider networks were explained

Fiber and Microwave, as the important building blocks of almost all transmission systems were introduced. Different Transport network systems such as SONET/SDH, CWDM and DWDM were mentioned

https://t.me/learningnets

# Transport Technologies Module
# Summary

Traffic between the countries and the continents have been carried over Sub Marine/Undersea Fiber Optic Cable systems for decades

Detail architecture of these systems such as Dry and Wet Plans and the nodes, also the components of each Sub Marine system were explained

https://t.me/learningnets

# Transport Technologies Module Summary

In this module, different transport capacity purchasing options were explained as well

IRU, Dark Fiber and Leasing options for having transport network capacity were shared

https://t.me/learningnets

# LAYER 2 TECHNOLOGIES

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

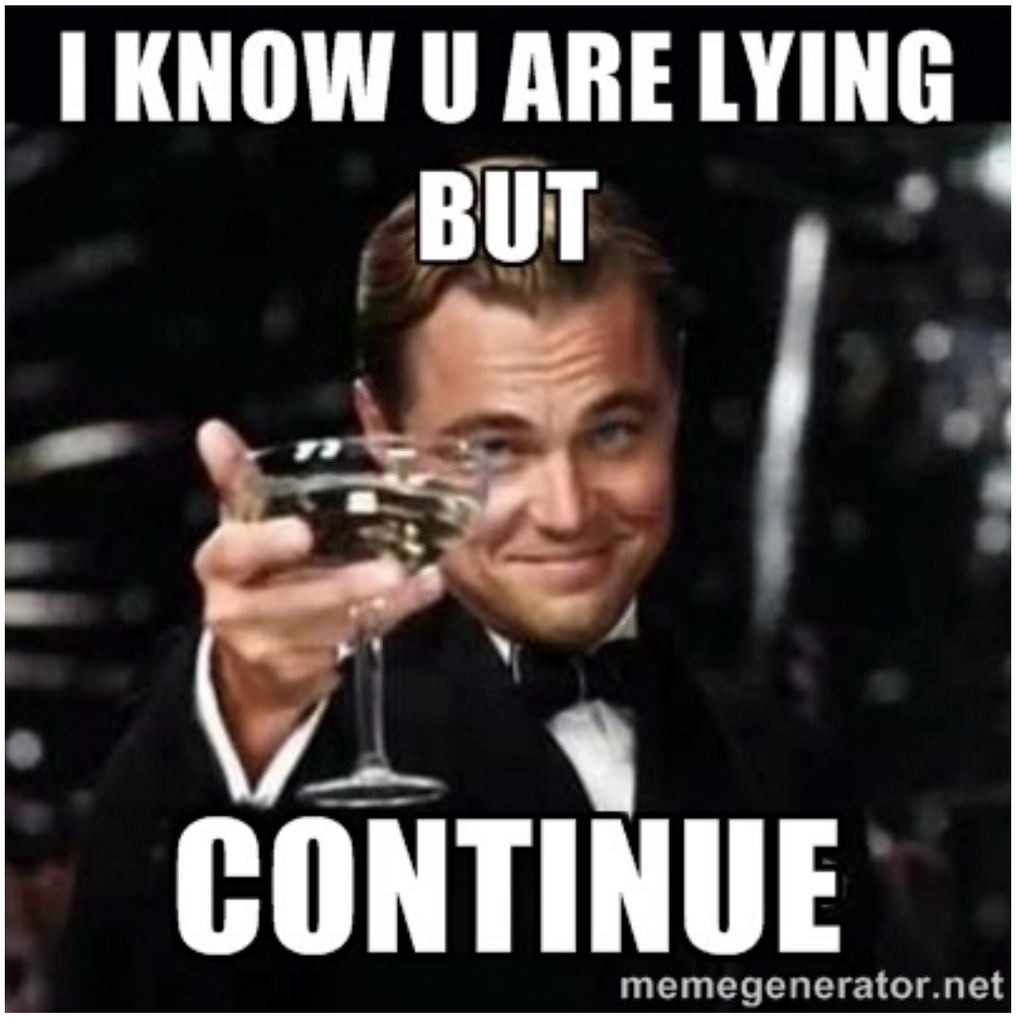
# Ethernet and it's control planes

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

https://t.me/learningnets

# Ethernet and it’s control planes

- Is Spanning Tree really the most common control plane?




https://t.me/learningnets

# Ethernet and it's control planes

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

https://t.me/learningnets

# Control and Data Plane

- Control plane refers to all the functions and processes that determine which path to use in the network. Routing protocols, spanning tree, LDP etc. are examples.

  - Control plane packets are destined to or locally originated by the networking devices, such as routers, switches itself.

- Data plane refers to all the functions and processes that forward packets/frames from one interface to another.

https://t.me/learningnets

# Control and Data Plane

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

https://t.me/learningnets

# Control and Data Plane

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

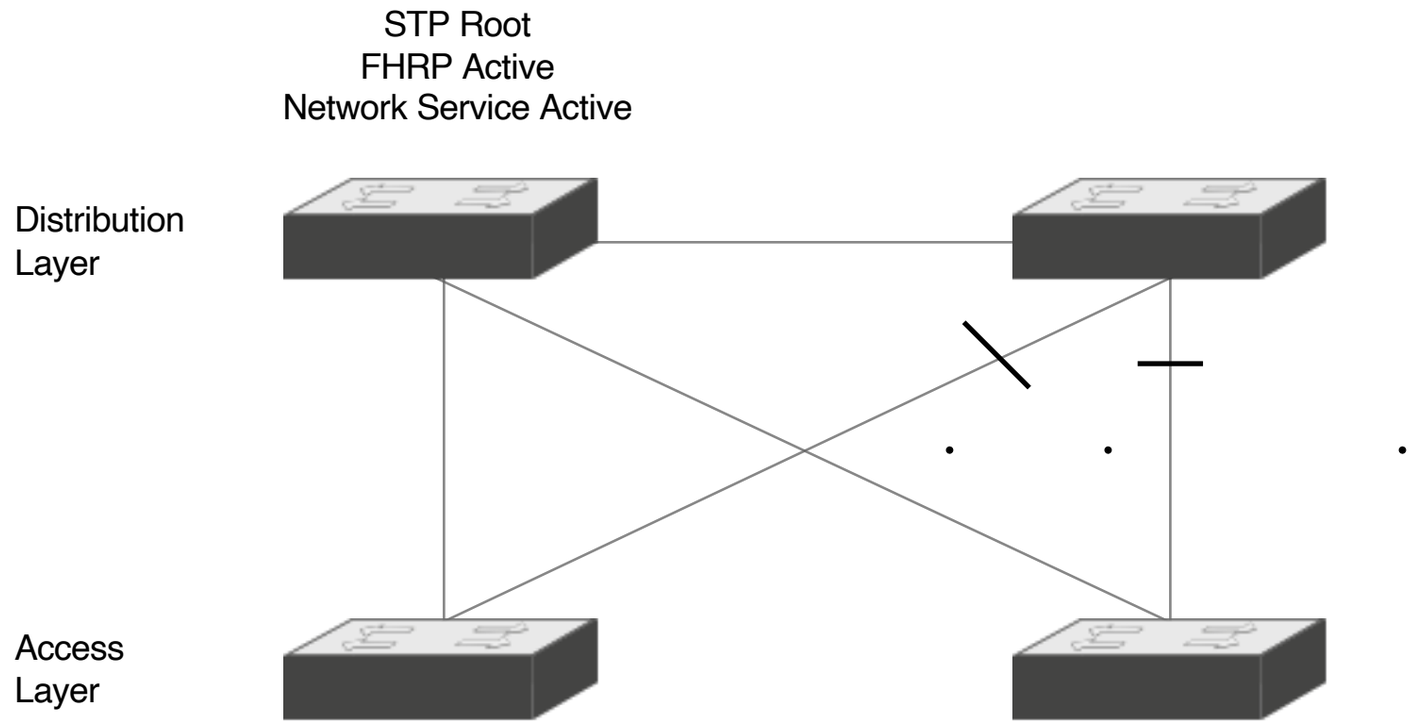
  - Data plane learning is used in Layer 2 in general, networking devices examine the packets/frames and learn MAC to interface binding.

pla
ne

https://t.me/learningnets

# Vlan and Flow Based Load Balancing

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

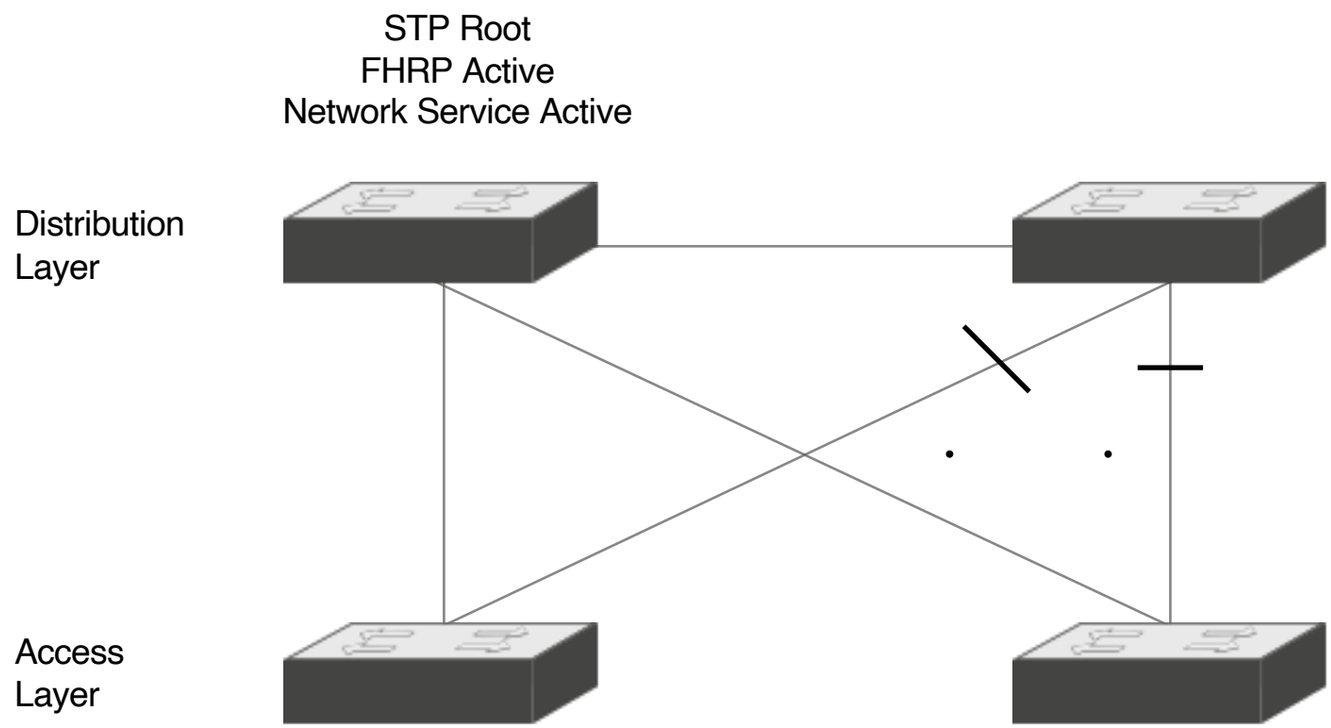



Triangle Spanning Tree Topology

https://t.me/learningnets

# Vlan and Flow Based Load Balancing

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




Triangle Spanning Tree Topology

https://t.me/learningnets

# Vlan and Flow Based Load Balancing

- In VPLS and EVPN, Vlan and flow based load balancing concepts will be revisited

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

https://t.me/learningnets

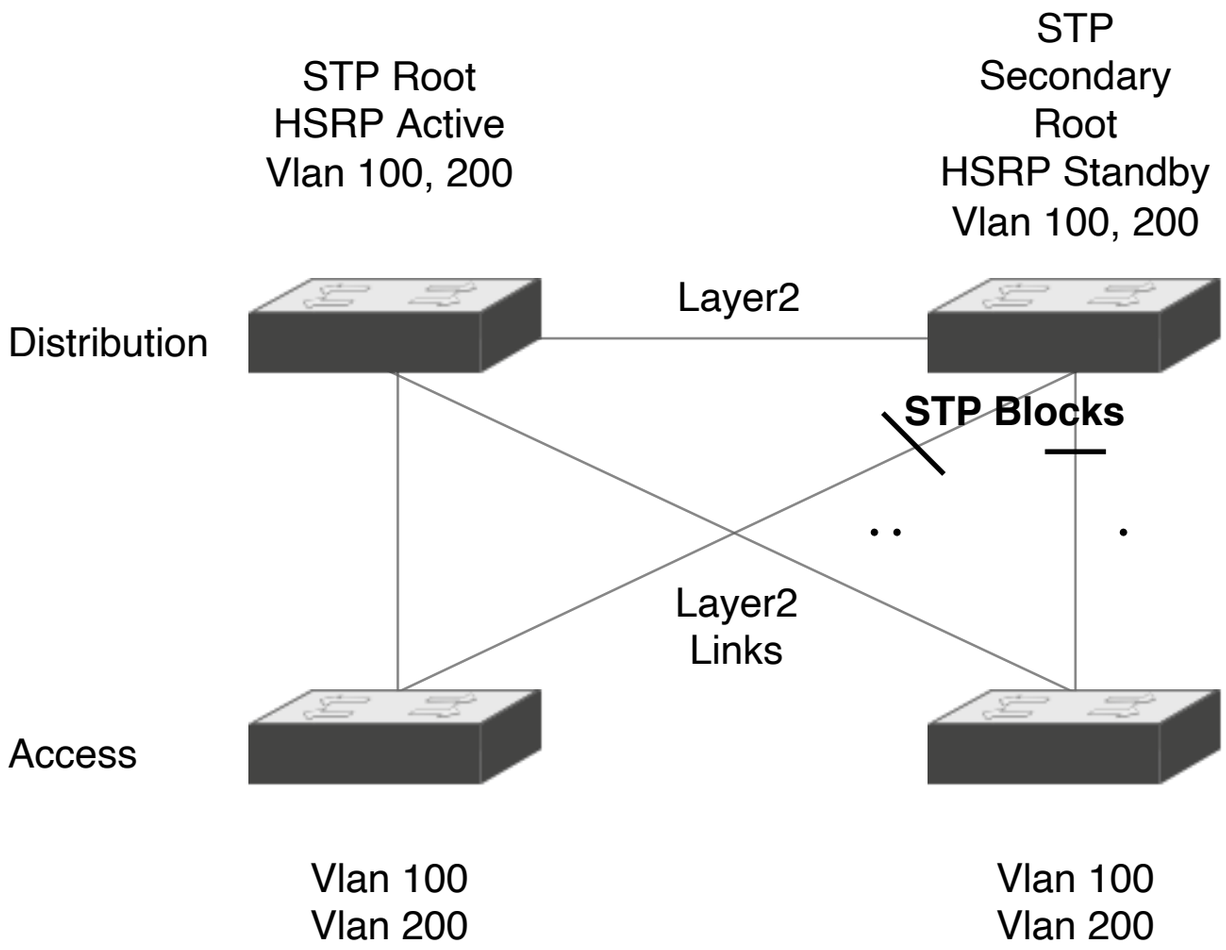
# Vlan and Flow Based Load Balancing

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

https://t.me/learningnets

# Spanning Tree

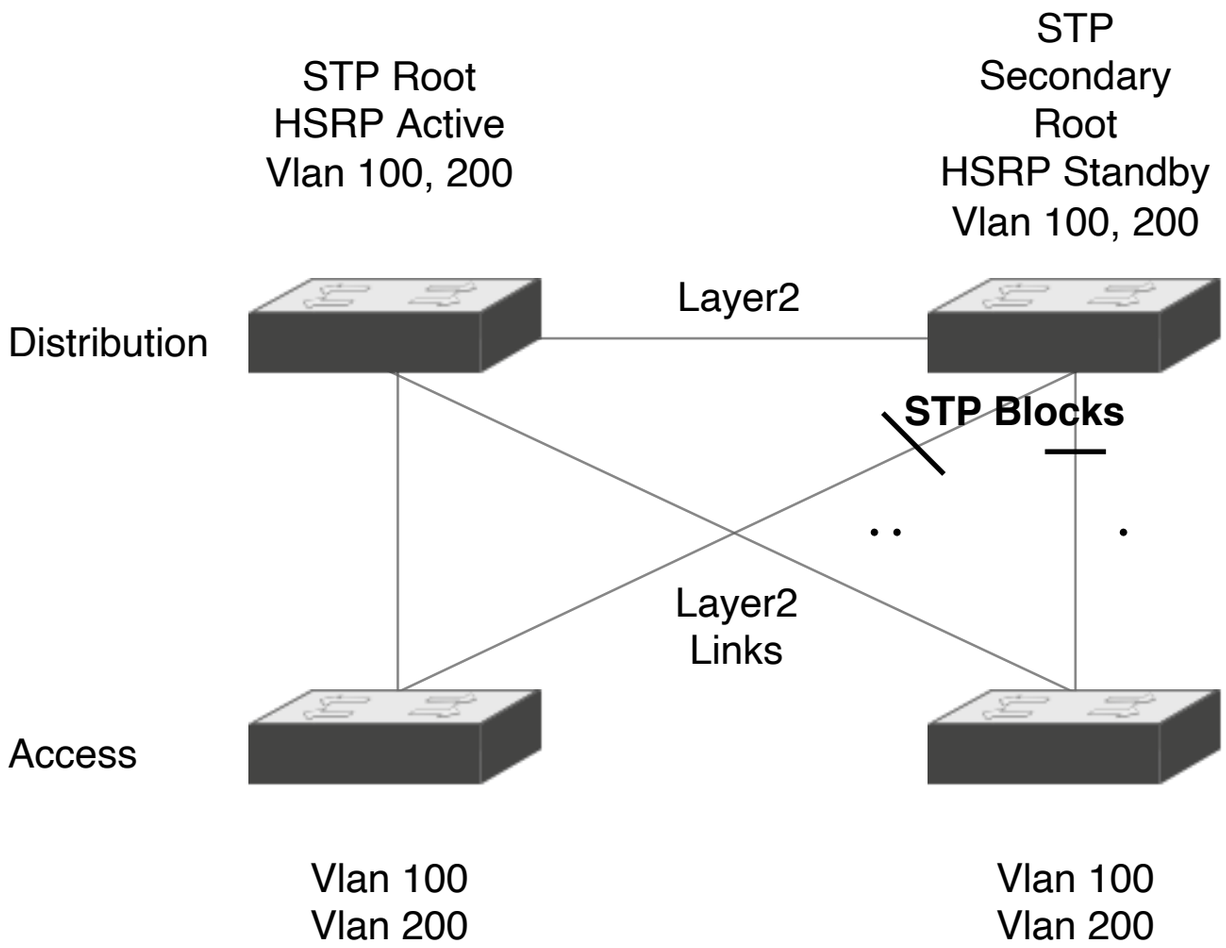
- Spanning tree is a control plane mechanism for Ethernet. It is used to create a layer 2 topology (A tree) by placing the root switch on top of the tree.




https://t.me/learningnets

# Spanning Tree

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




https://t.me/learningnets

# Spanning Tree

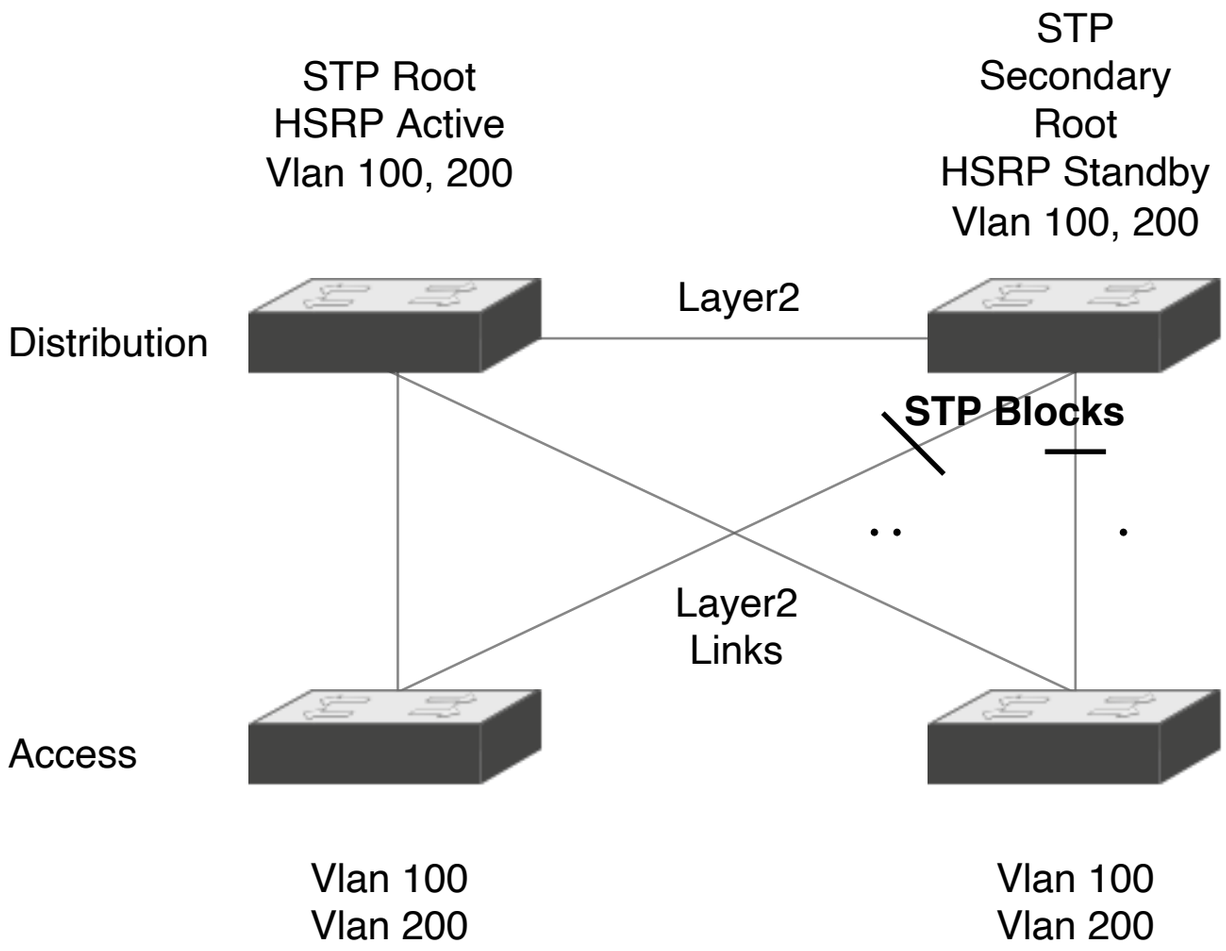
- Loop has to be mitigated but blocking links don’t allow using all the links actively. Spanning tree doesn’t provide multipathing.

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

https://t.me/learningnets

# Spanning Tree Root Switch Selection

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

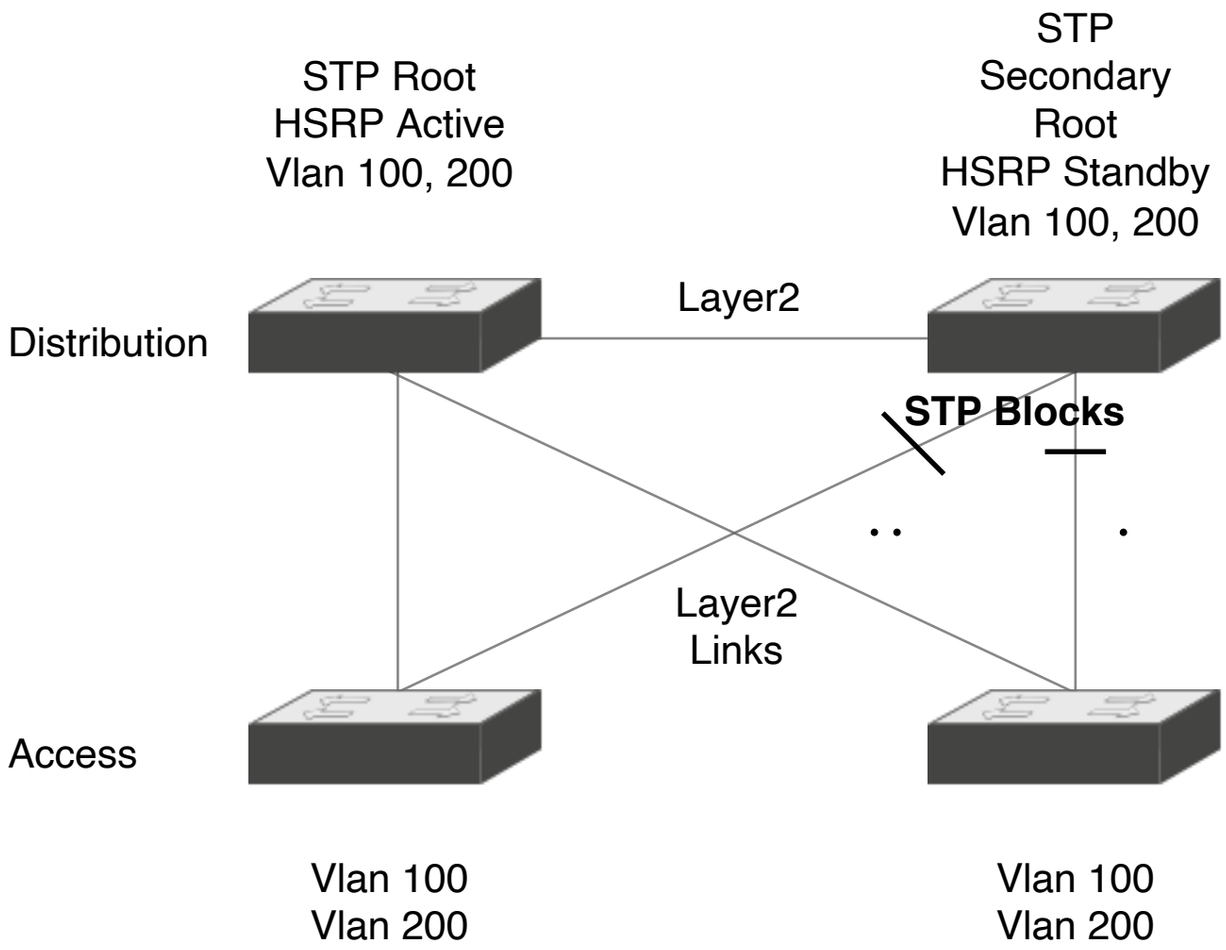



https://t.me/learningnets

# Spanning Tree Root Switch Selection

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

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




https://t.me/learningnets

# Spanning Tree and Load Balancing Capability

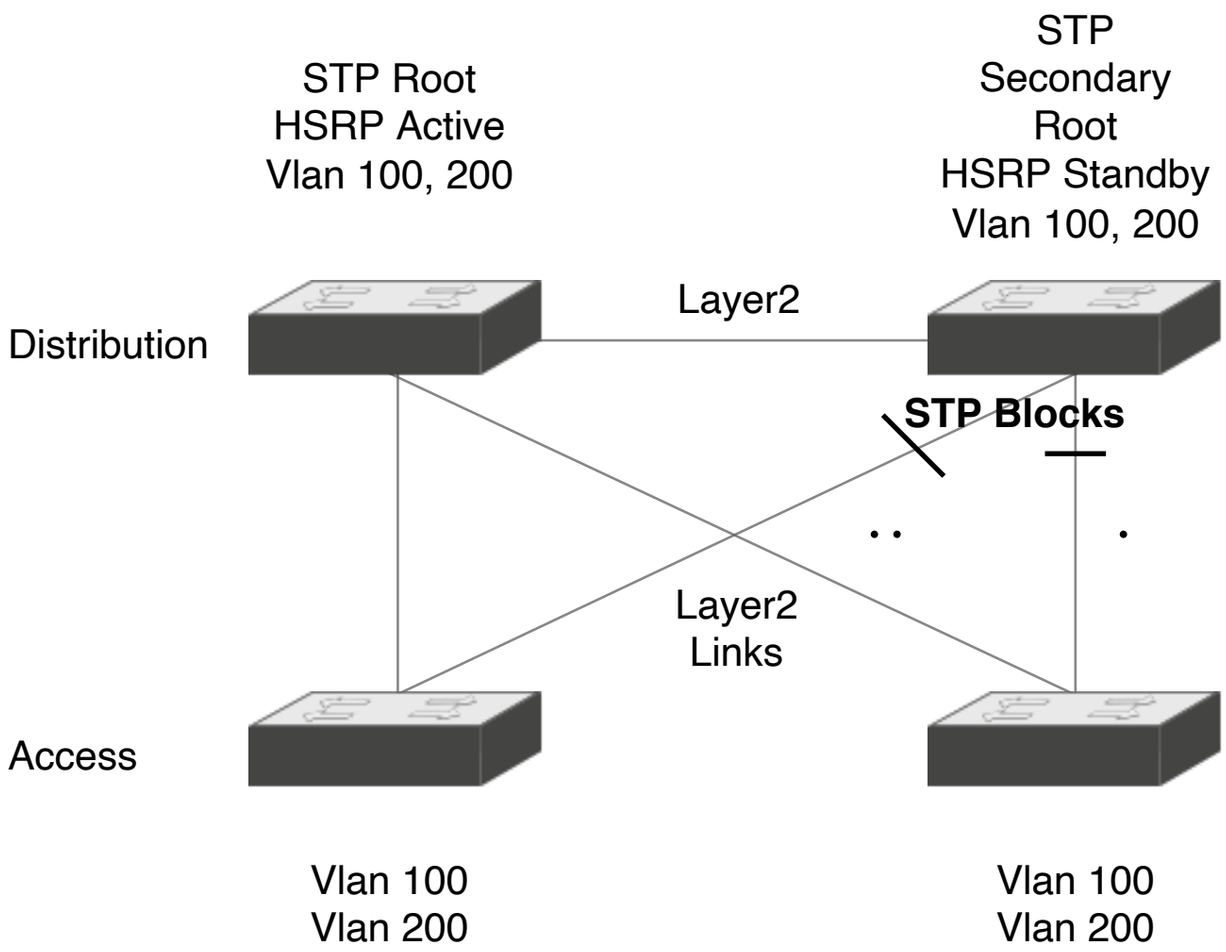
- Two common load balancing techniques are in the Layer 2 networks; Vlan based and Flow based load balancing
  - Spanning tree doesn't allow the flow based load balancing
- Some implementations of spanning tree allows only Vlan based load balancing. Some of them allow only active-standby redundancy

https://t.me/learningnets

# Spanning Tree and Load Balancing Capability

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

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




https://t.me/learningnets

# Spanning Tree and Load Balancing Capability

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

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

https://t.me/learningnets

# Spanning Tree Toolkit

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

https://t.me/learningnets

# Spanning Tree Toolkit

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

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

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

https://t.me/learningnets

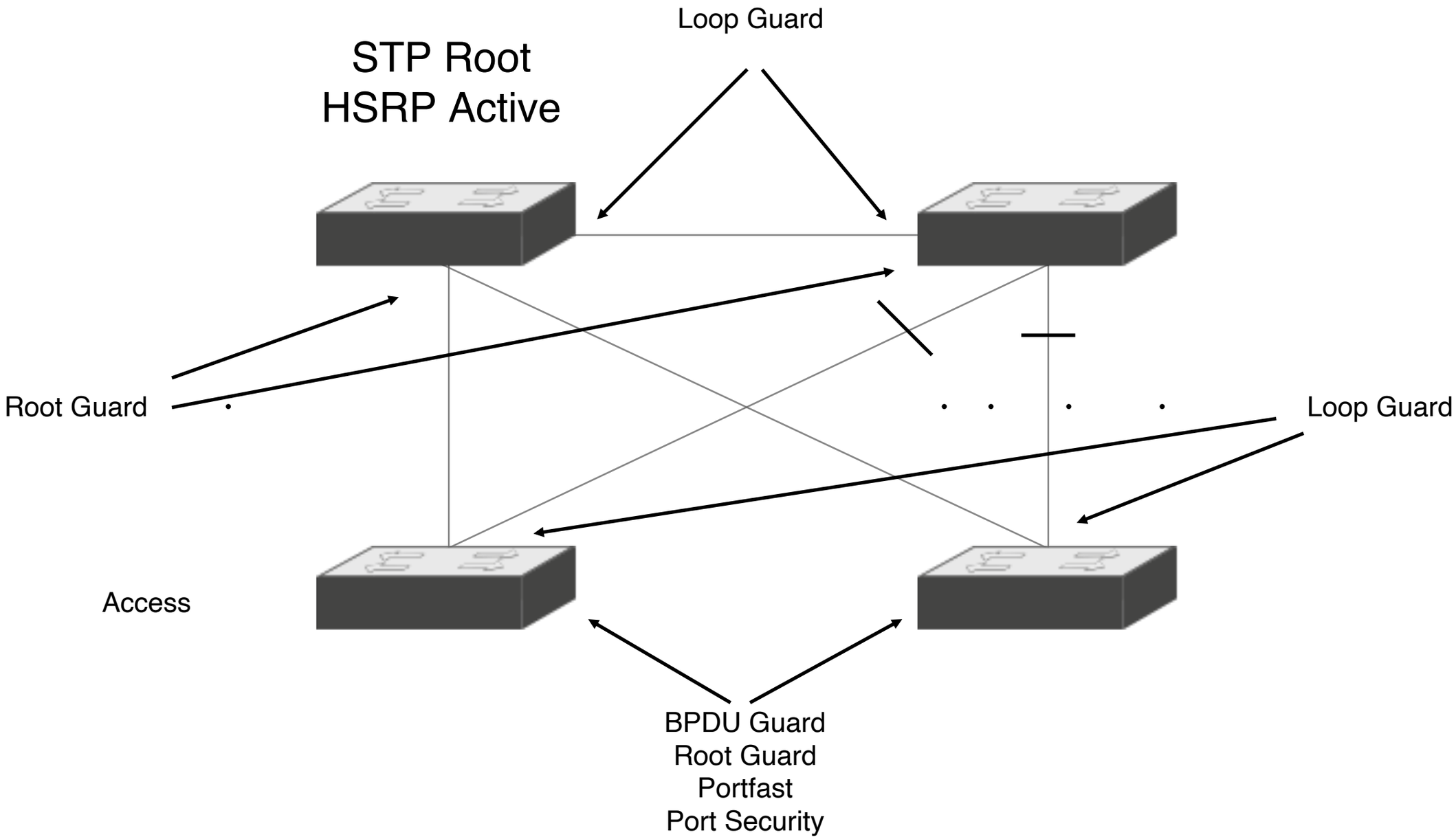
# Spanning Tree Toolkit

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

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

https://t.me/learningnets

# Spanning Tree Toolkit Placement




https://t.me/learningnets

# Spanning Tree Toolkit Placement

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

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

https://t.me/learningnets

# Spanning Tree Toolkit Placement

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

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

https://t.me/learningnets

# Spanning Tree Toolkit Placement

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

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

https://t.me/learningnets

# Spanning Tree Toolkit Placement

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

https://t.me/learningnets

# Spanning Tree Best Practices

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

https://t.me/learningnets

# Spanning Tree Best Practices

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

https://t.me/learningnets

# Odd – Even Vlan Load Balancing

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

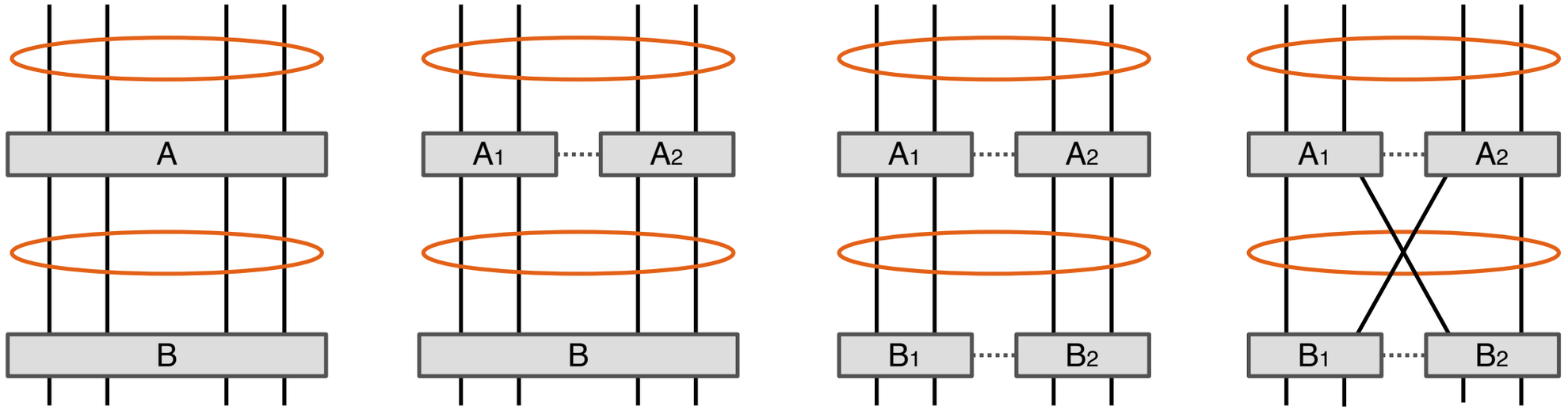
https://t.me/learningnets

# LAG, MC-LAG, VSS and VPC

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

https://t.me/learningnets

# LAG, MC-LAG, VSS and VPC




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

https://t.me/learningnets

- Caveats:

- Flows are mapped to the physical link by the hashing algorithm of the network device in per flow load balancing.

  - Per packet load balancing can cause reordering problem at the destination due to jitter.

- Each Individual flow can reach up to physical link speed, for example if flow generates 1,5 Gbps traffic but it is mapped to 1 Gbps physical link, capacity need to be increased.

https://t.me/learningnets

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

https://t.me/learningnets

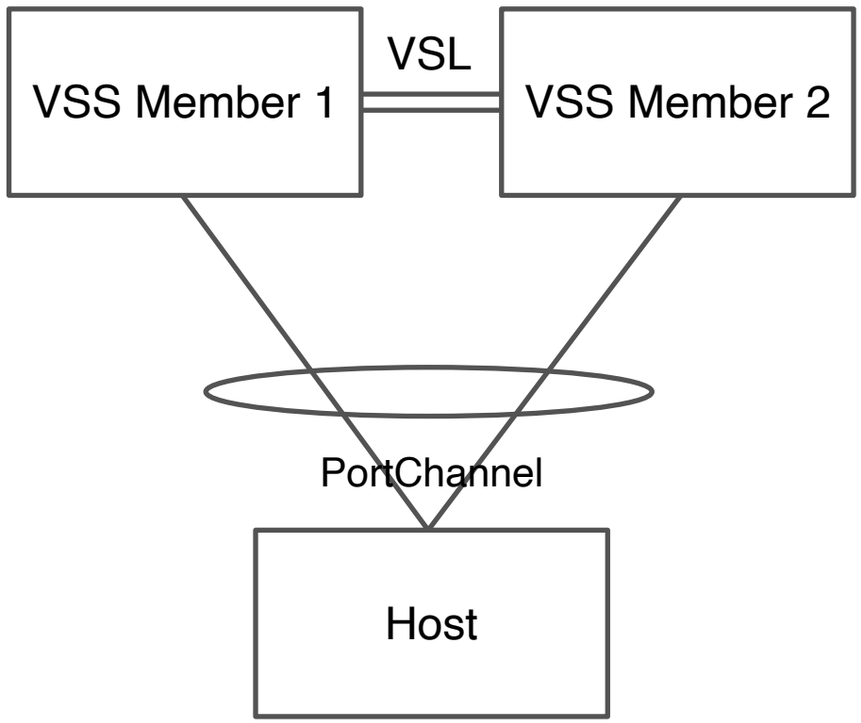
# VSS and VPC

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

https://t.me/learningnets

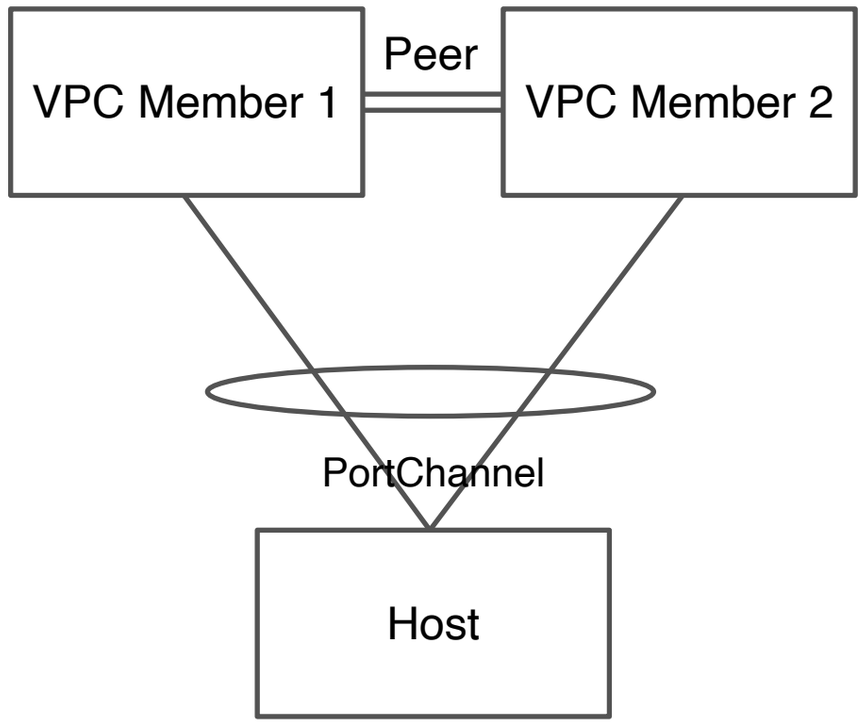
# VSS and VPC

Single Control Panel




VSS Client

Independant Control Panel



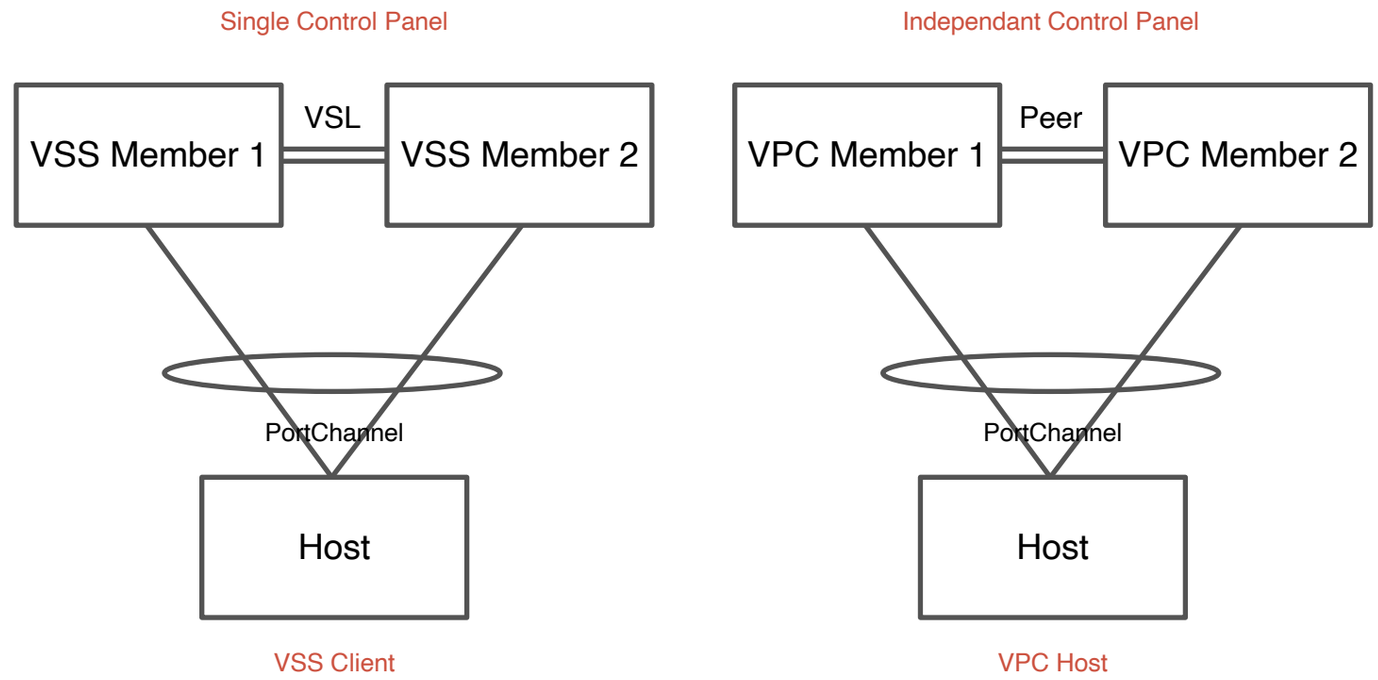

VPC Host

https://t.me/learningnets

# VSS and VPC

- VSS and VPC enabled switches use Cisco Preparatory protocol between to send control messages such as system IDs

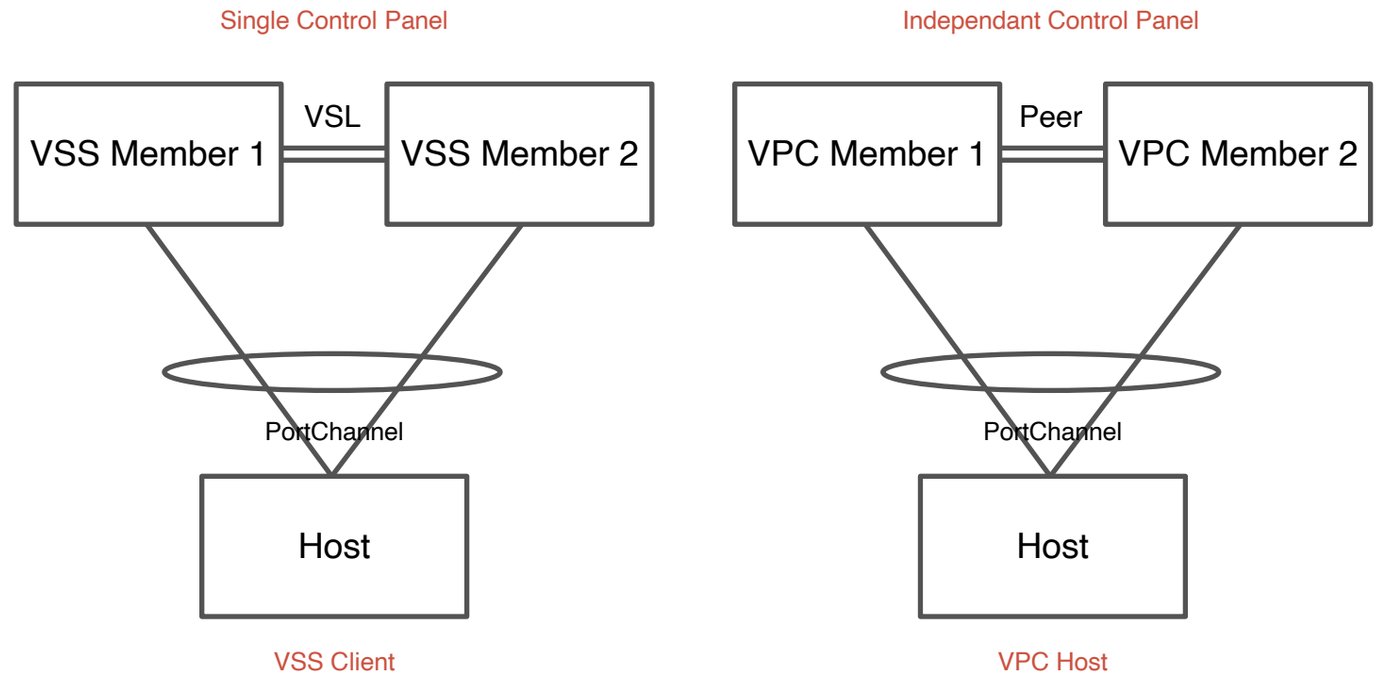
- Downstream switch sees two VSS or VPC switches as one switch




https://t.me/learningnets

# VSS and VPC

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




https://t.me/learningnets

# VSS and VPC

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

https://t.me/learningnets

# VLAN, VTP and TRUNKING Best Practices

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

https://t.me/learningnets

# VLAN, VTP and TRUNKING Best Practices

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

https://t.me/learningnets

# VLAN, VTP and TRUNKING Best Practices

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

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

https://t.me/learningnets

# First Hop Redundancy Protocols

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

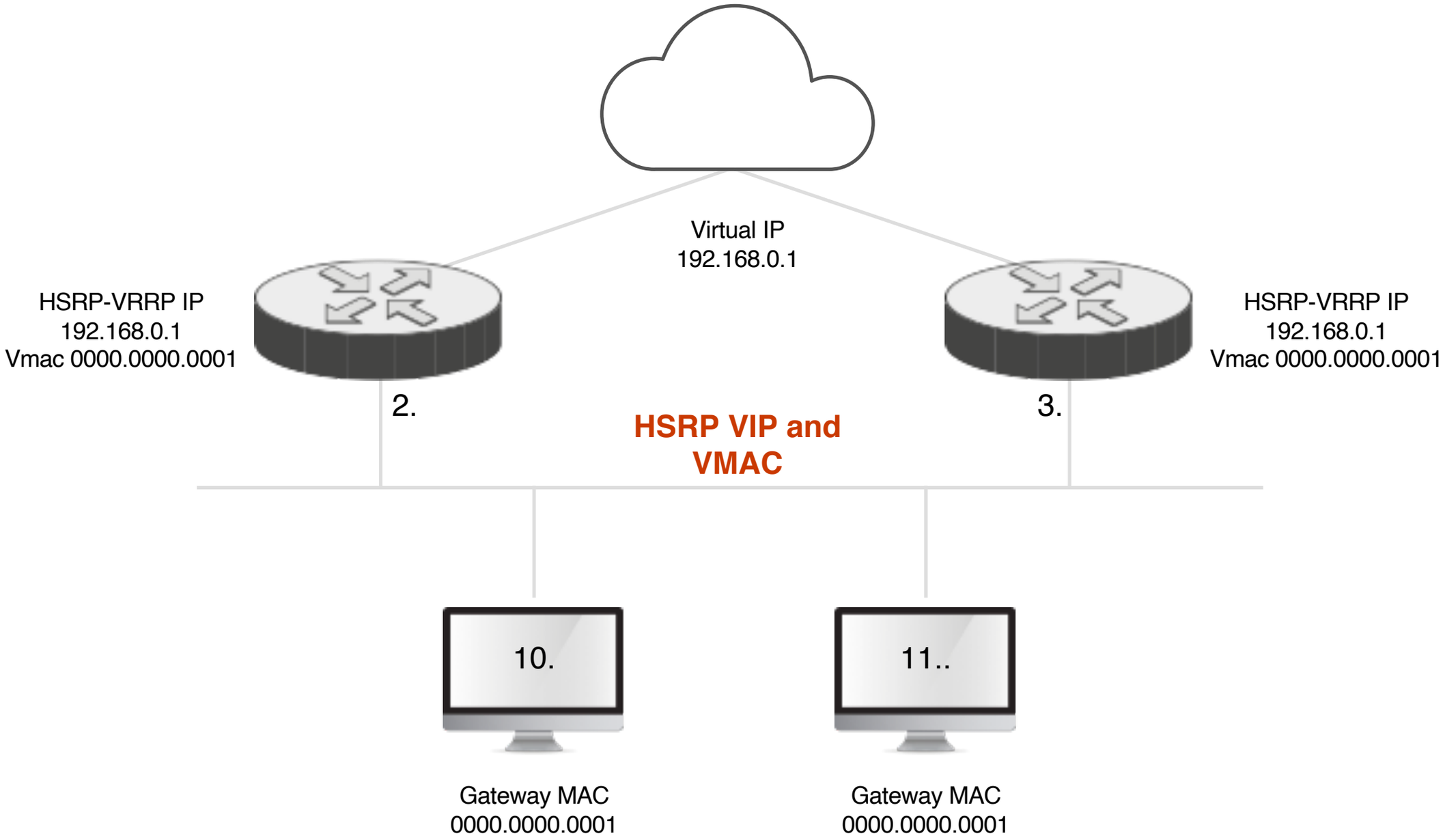
https://t.me/learningnets

# First Hop Redundancy Protocols

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

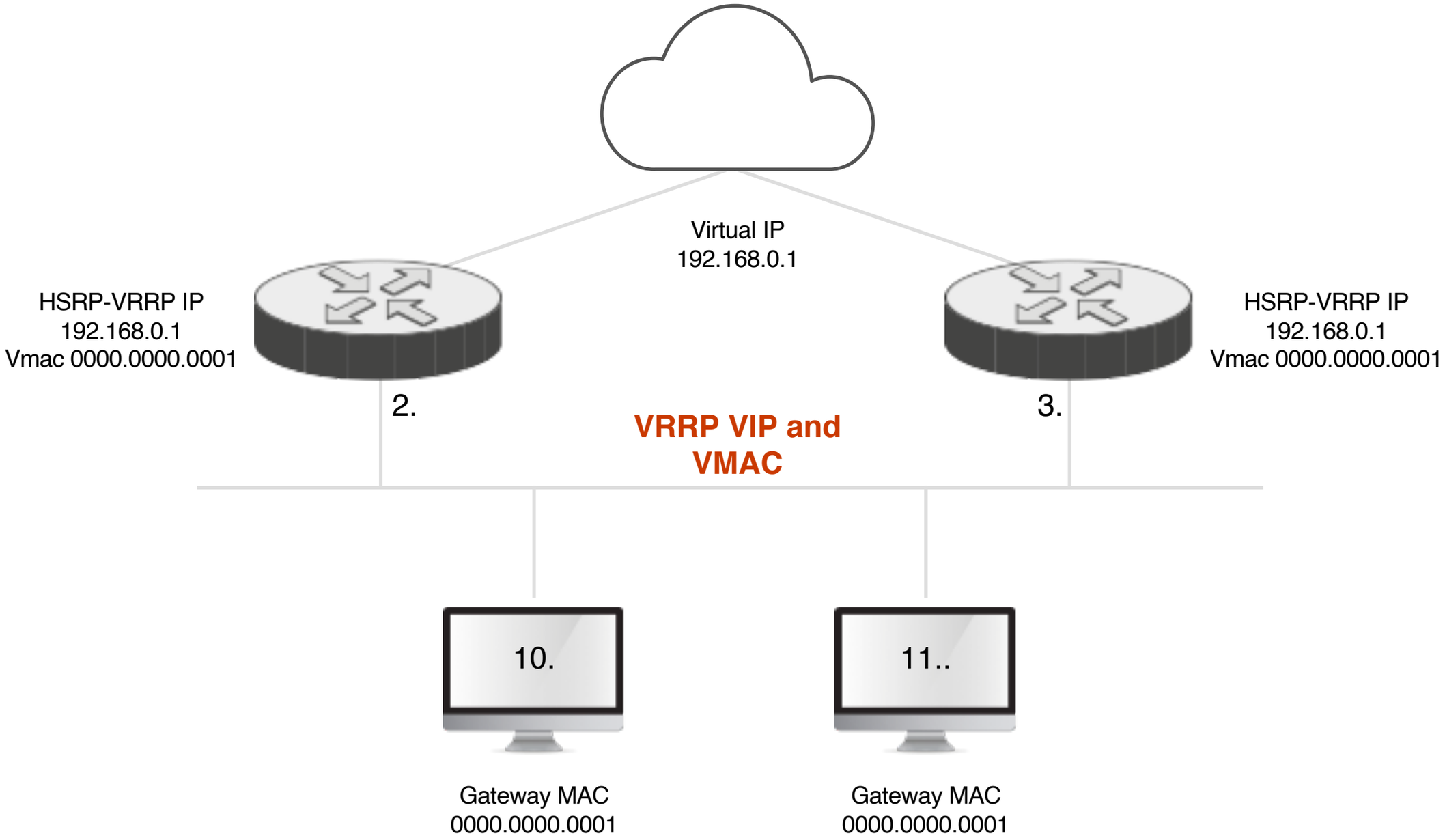
https://t.me/learningnets

# HSRP




https://t.me/learningnets

# VRRP




https://t.me/learningnets

# GLBP

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

https://t.me/learningnets

# GLBP

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

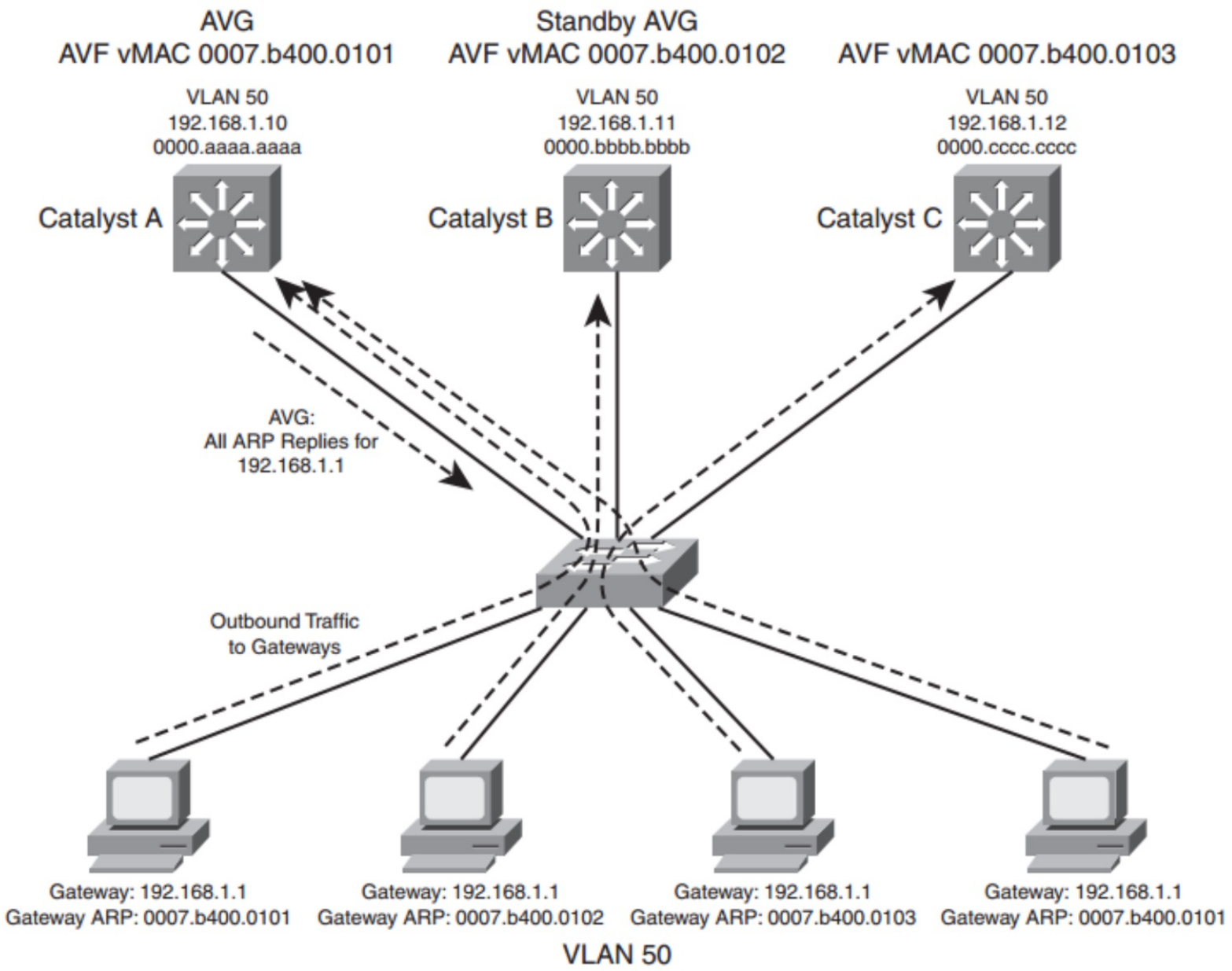
https://t.me/learningnets

# GLBP

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

https://t.me/learningnets

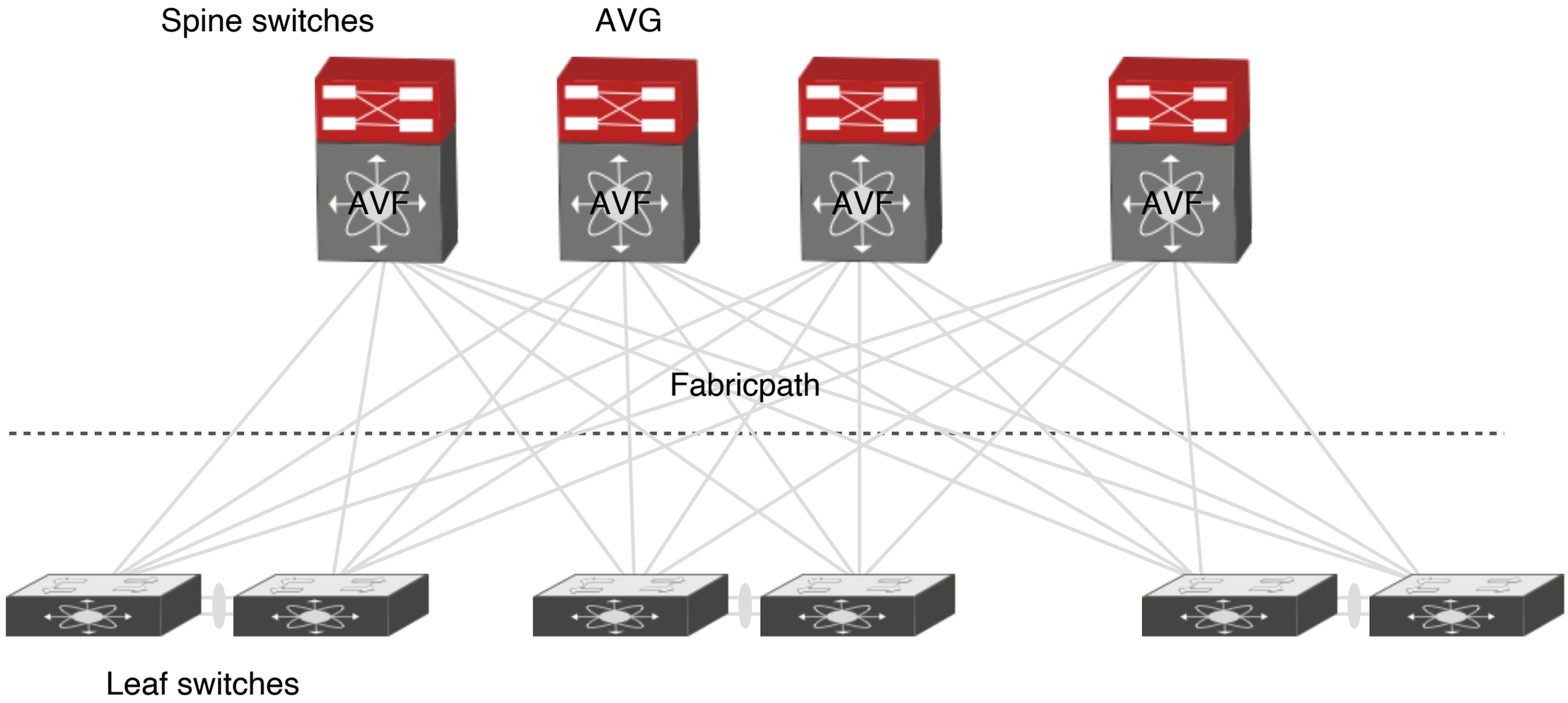
# GLBP Operation – How it works




https://t.me/learningnets

# GLBP on Leaf and Spine Topology

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




https://t.me/learningnets

# GLBP Operation - How it works

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

glb
p

https://t.me/learningnets

# HSRP - VRRP - GLBP Comparison

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

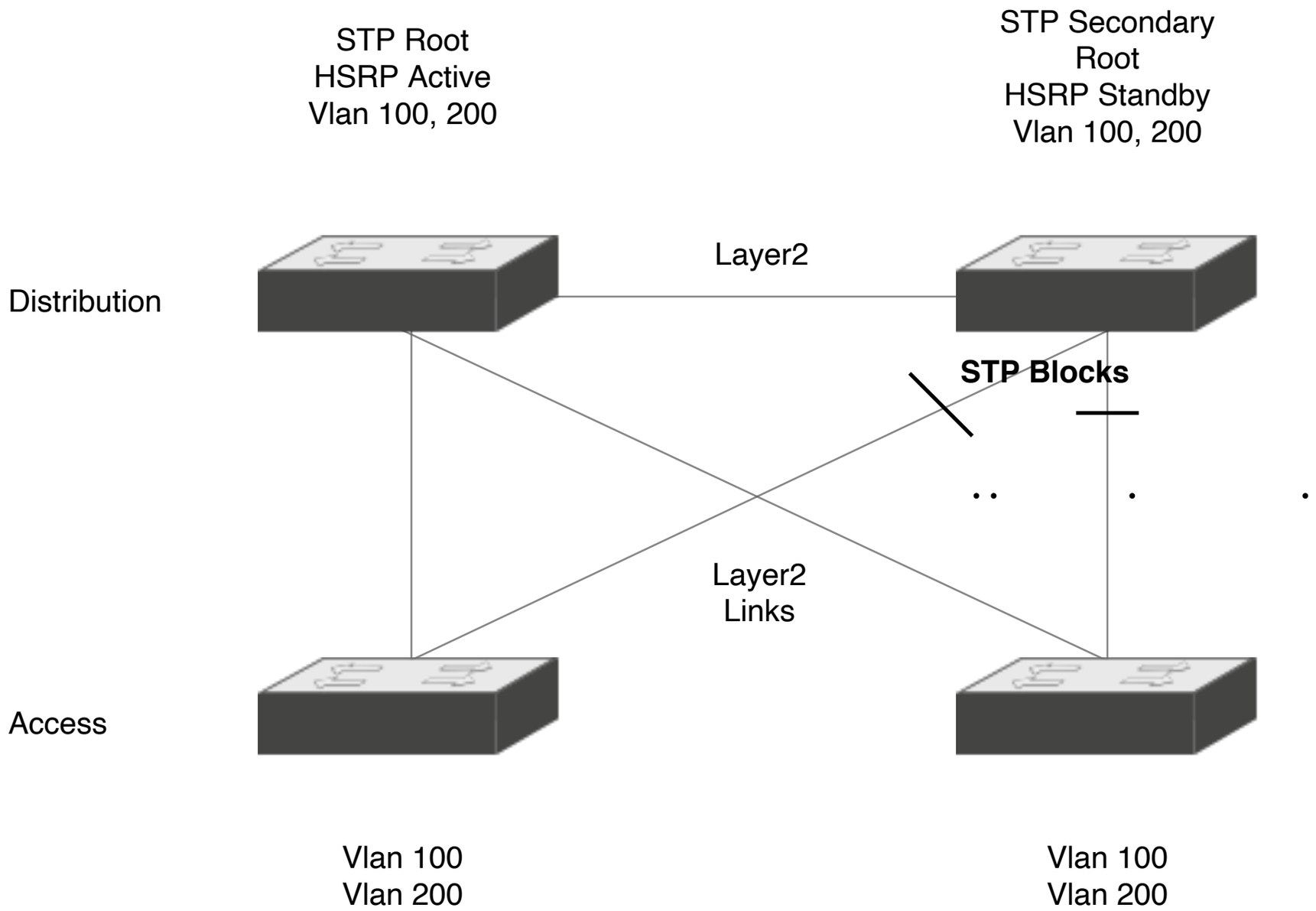
https://t.me/learningnets

# Layer 2 and Layer 3 Interaction

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

https://t.me/learningnets

# Layer 2 and Layer 3 Interaction




https://t.me/learningnets

# Layer 2 and Layer 3 Interaction

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

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

https://t.me/learningnets

# Layer 2 and Layer 3 Interaction

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

https://t.me/learningnets

# Layer 2 and Layer 3 Interaction Summary

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

https://t.me/learningnets

# Layer 2 Traffic Engineering

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

https://t.me/learningnets

# Layer 2 Traffic Engineering

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

https://t.me/learningnets

# Layer 2 Access Design

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

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

https://t.me/learningnets

# Layer 2 Access Design

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

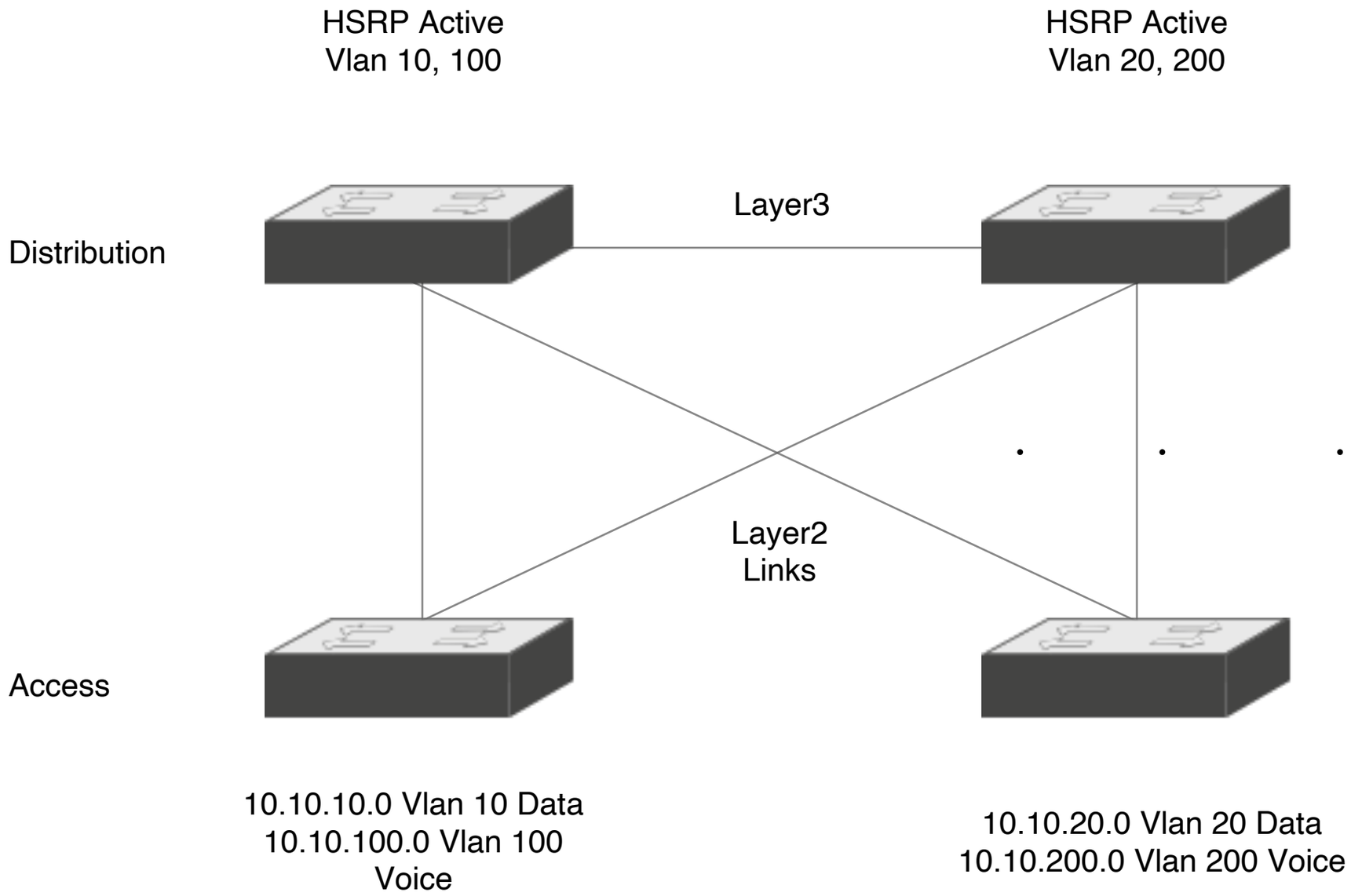
https://t.me/learningnets

# Layer 2 Access Design

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

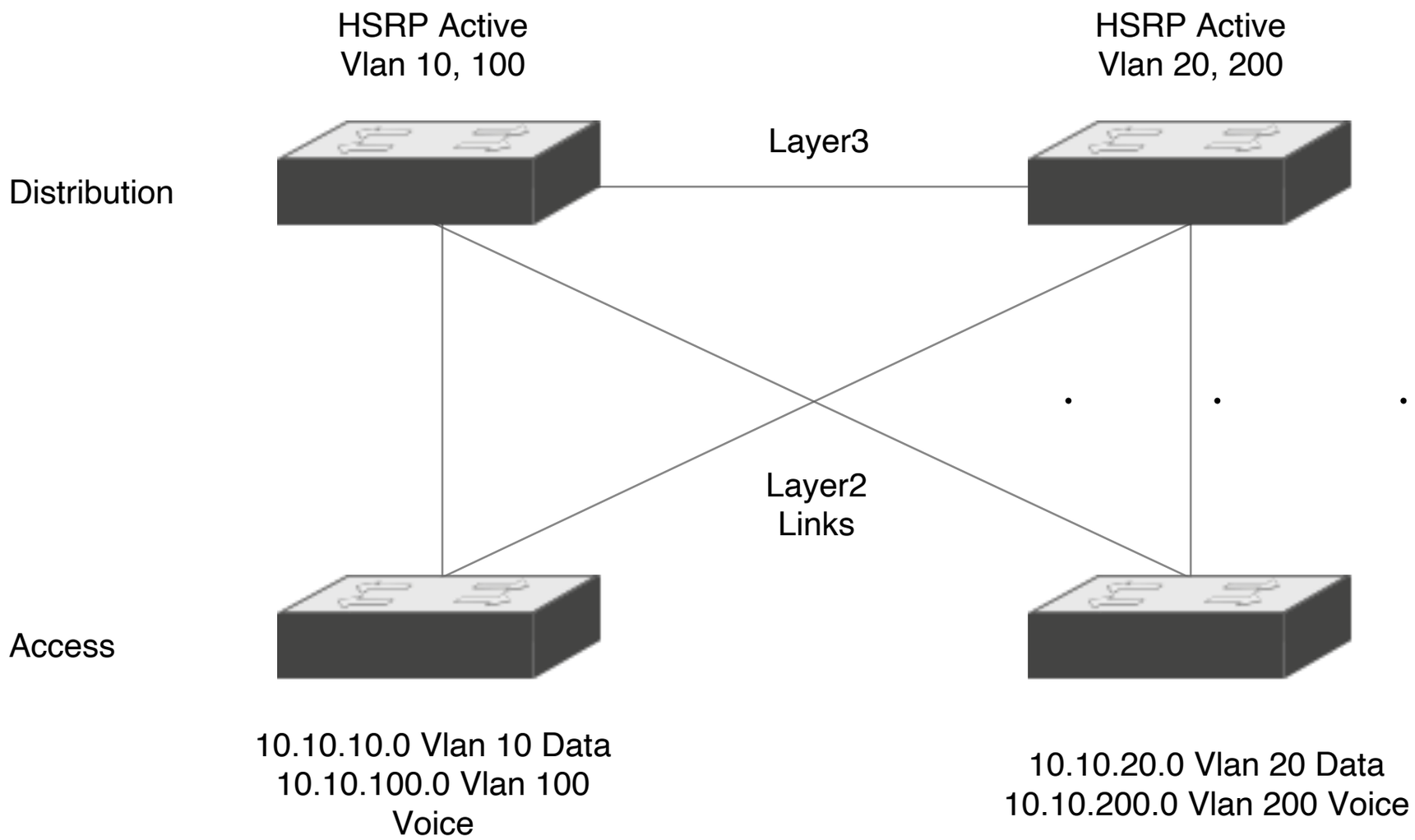
https://t.me/learningnets

# Layer 2 Loop Free Topology




https://t.me/learningnets

# Layer 2 Loop Free Topology




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

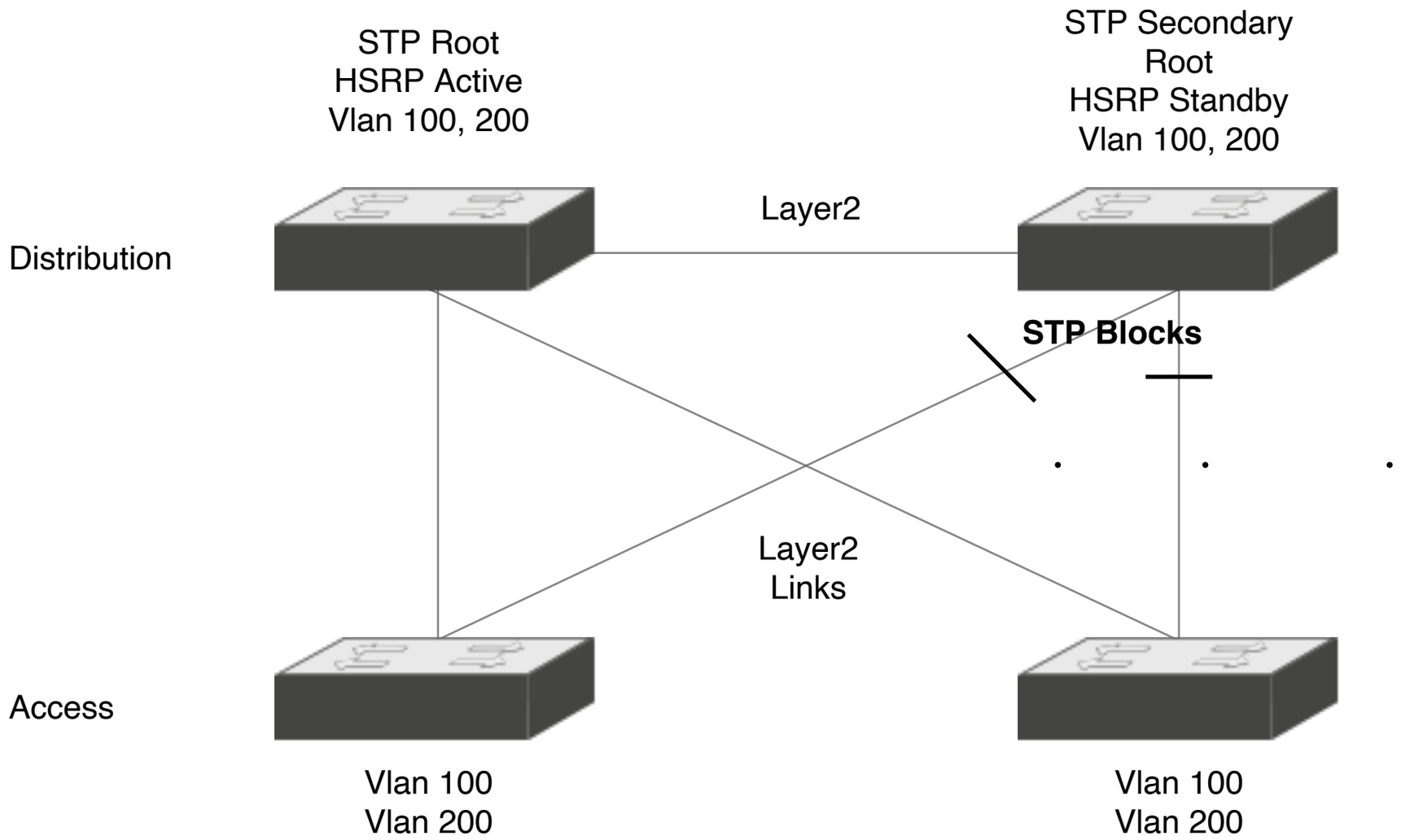
https://t.me/learningnets

# Layer 2 Loop Free Topology

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

https://t.me/learningnets

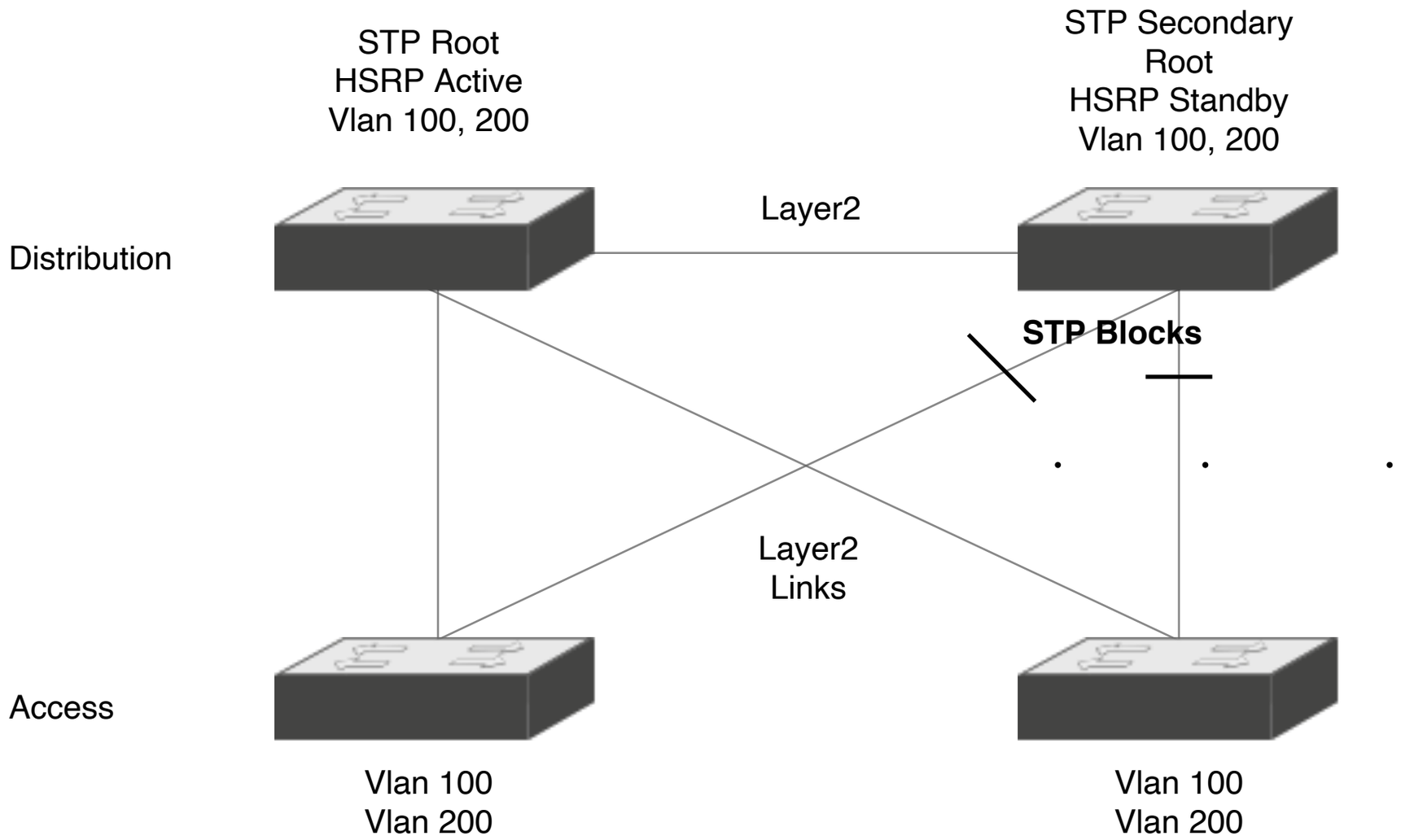
# Layer 2 Looped Topology




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

https://t.me/learningnets

# Layer 2 Looped Topology

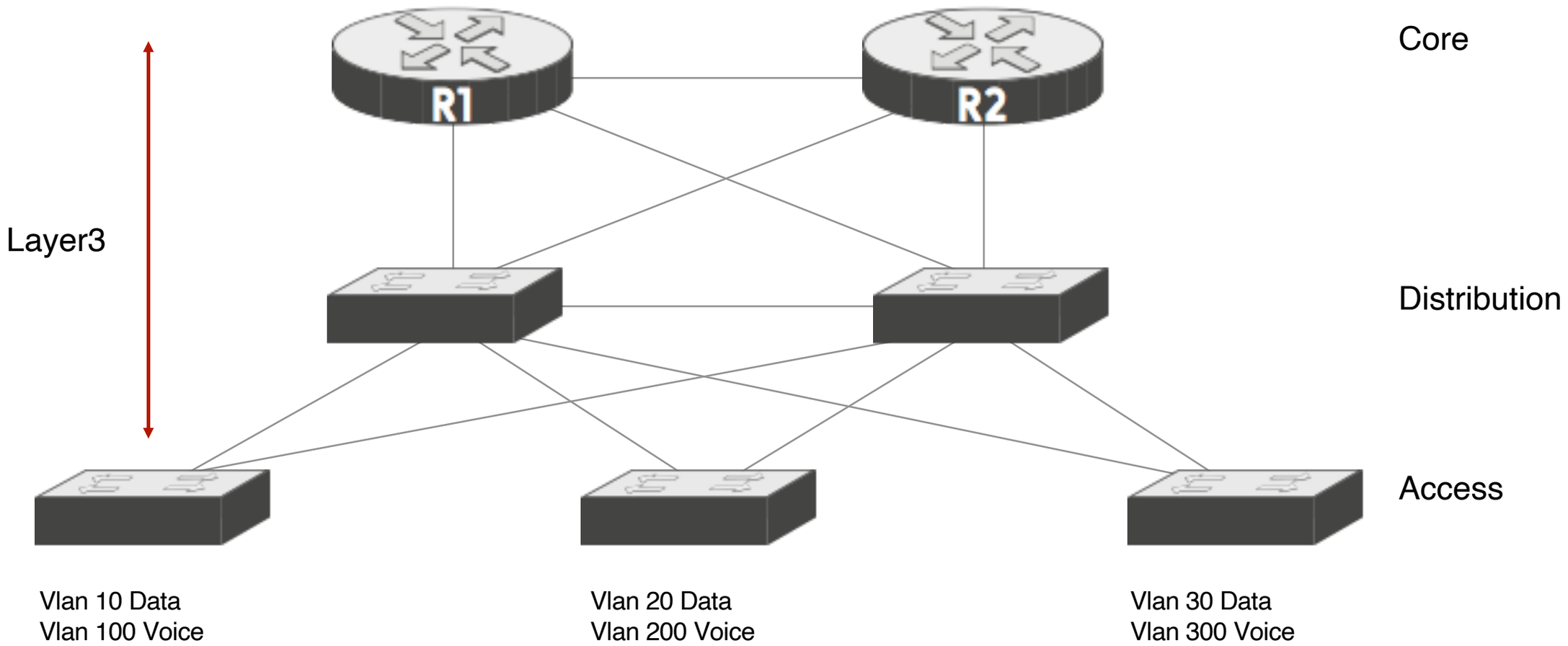



- Same Vlan can be used on every access switch.
- Bi-sectional bandwidth usage is low

https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

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

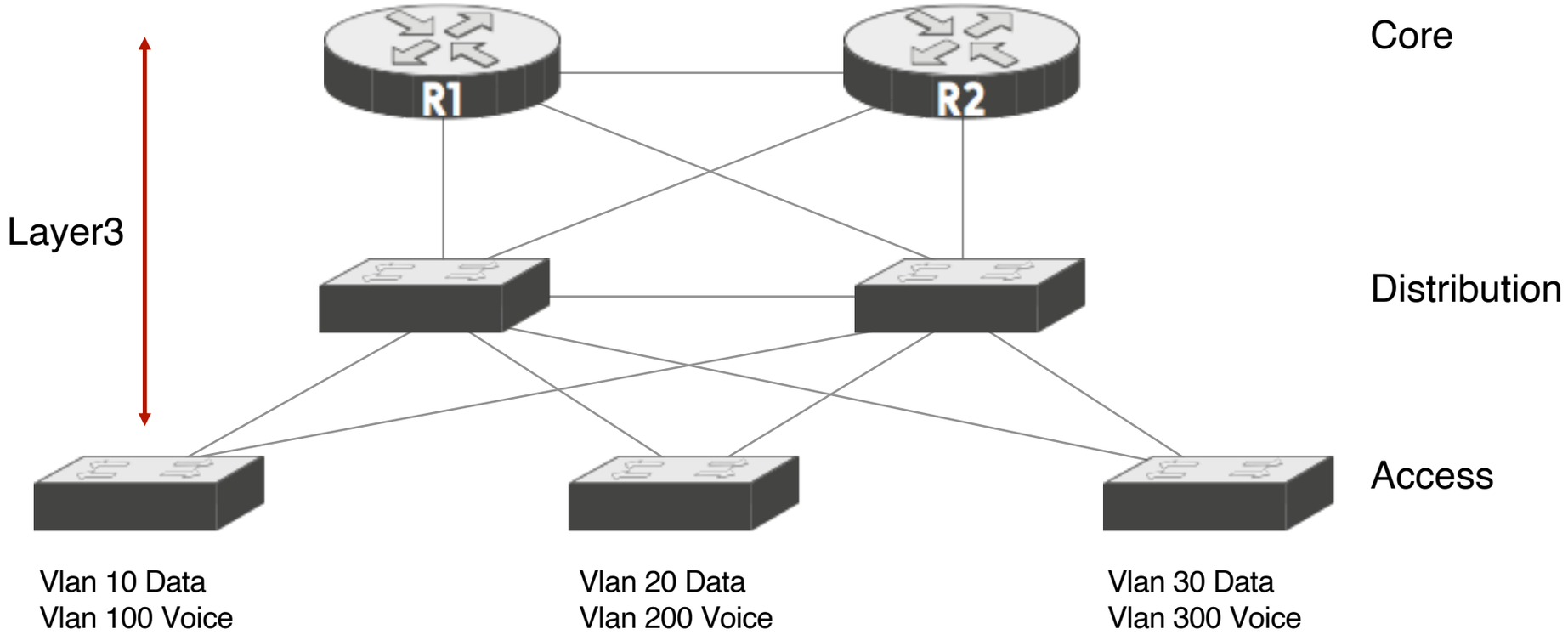



Unique Voice and Data Vlans

https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

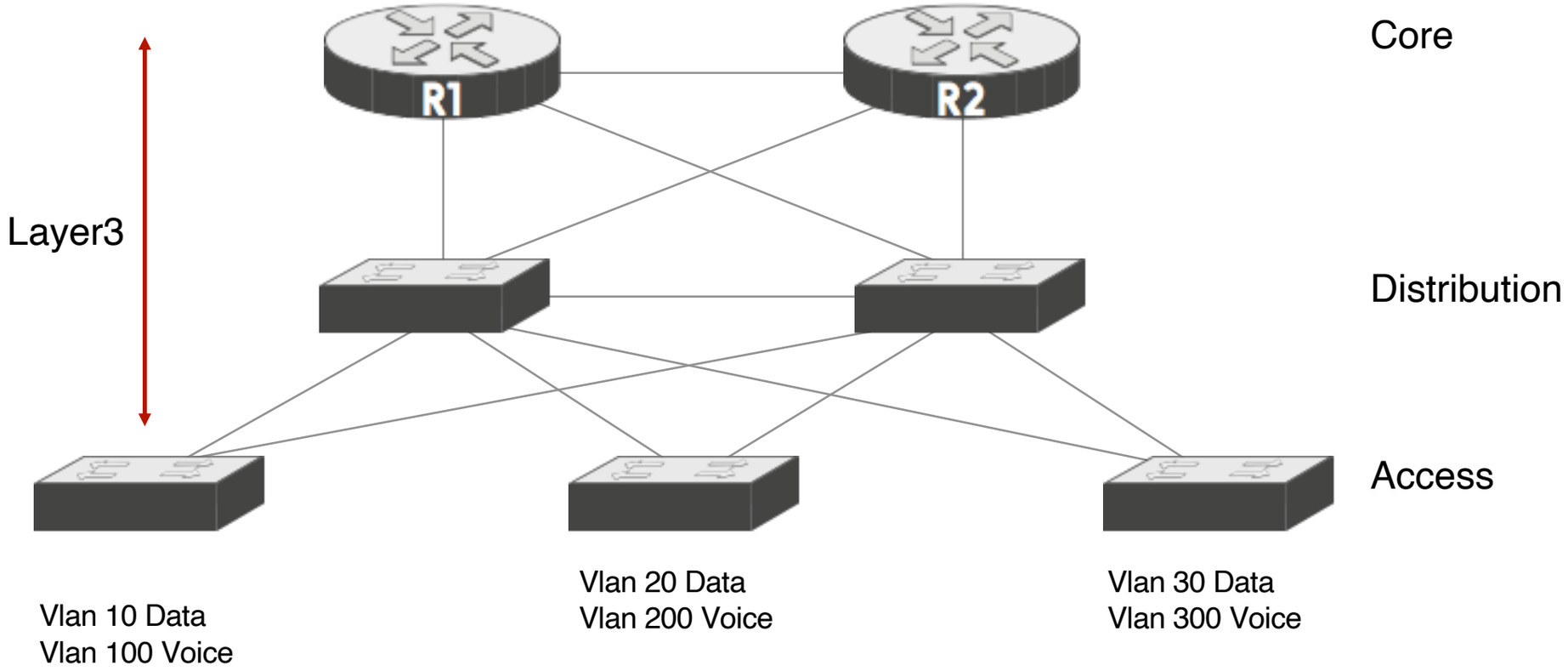
- No need to have any first hop redundancy protocol since the access layer switch is the first hop gateway.
- There is no spanning tree in this design as there is no layer 2 link between the network devices




https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

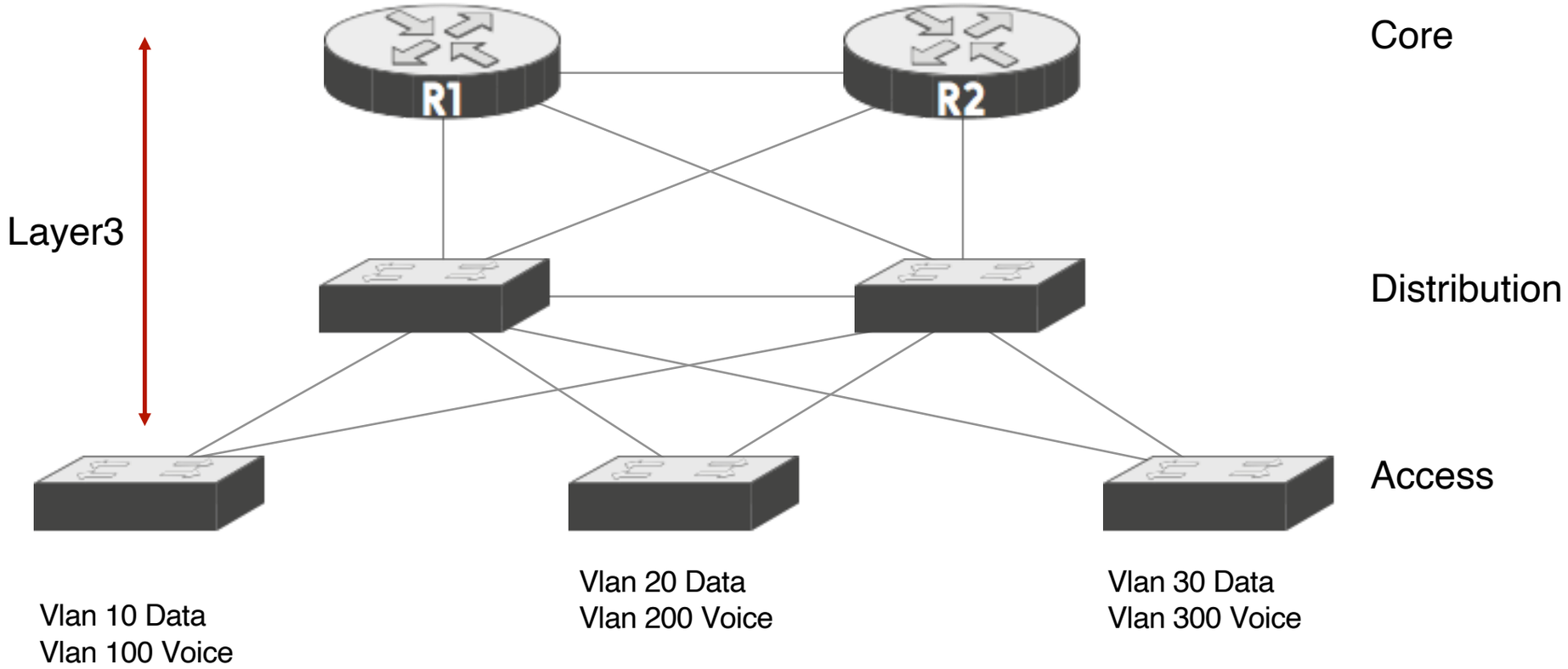
- We can take advantage of fast convergence since we can use any IGP protocols between access and distribution layer and we can tune it, of course tuning protocol convergence time comes with its cost.




https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

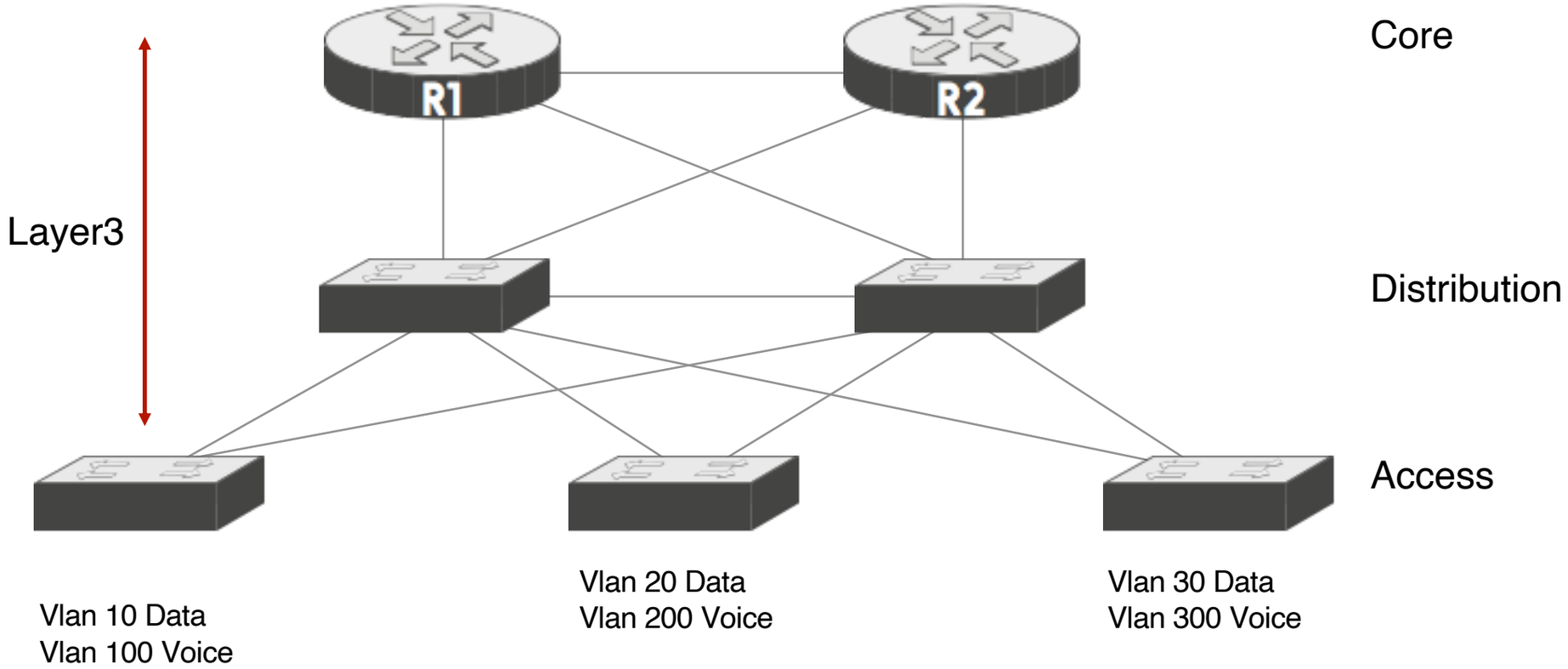
- Tuning routing protocol for faster convergence, may impact overall stability of the network. You might have false positive.




https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

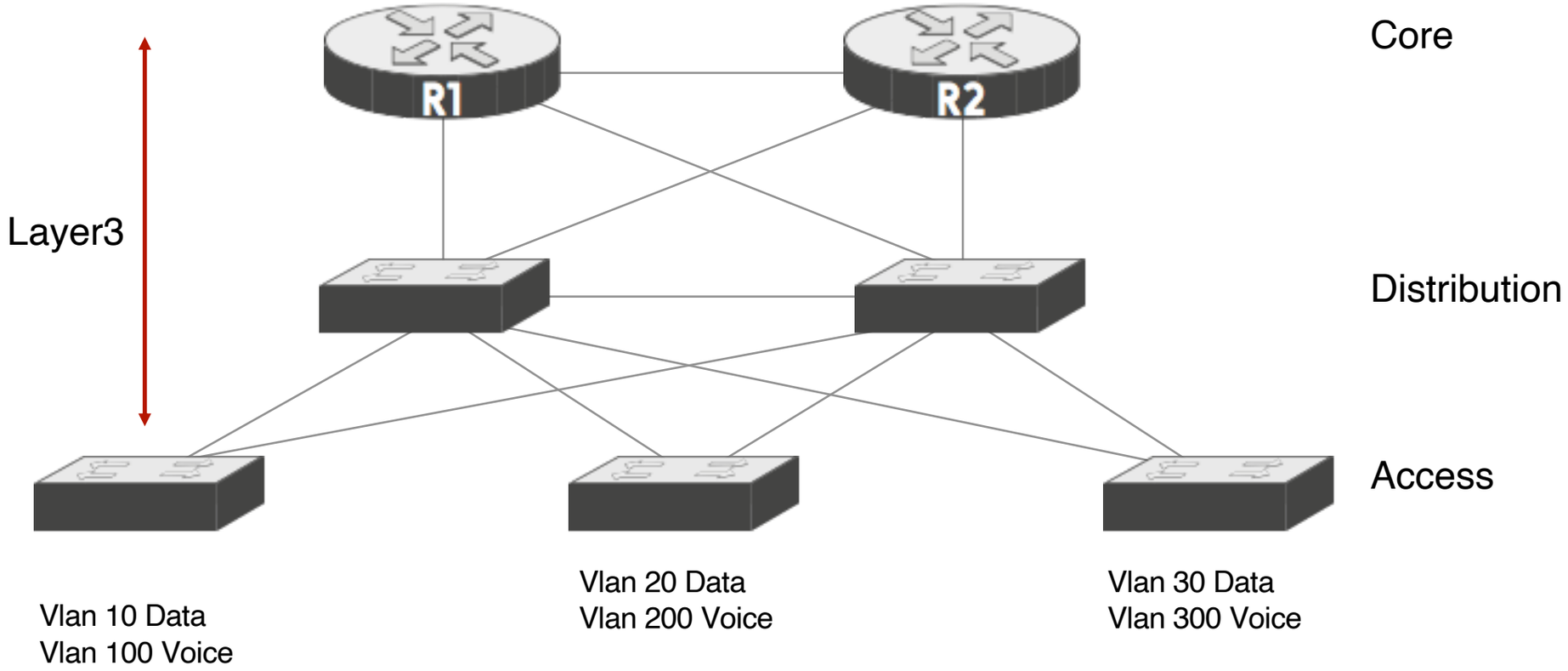
- Also configuration will be much more complex.




https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

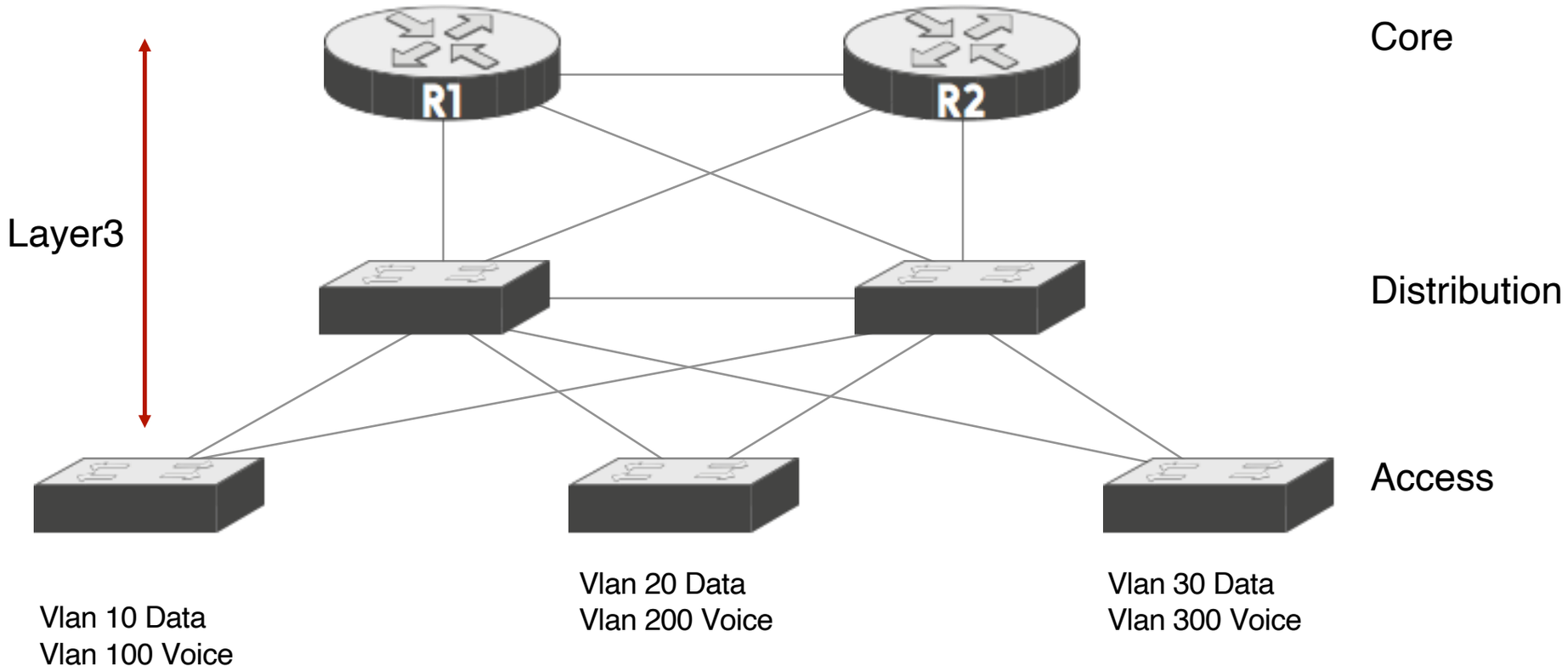
- Although, there is no spanning tree anymore, still you may want to protect user site loop by enabling spanning tree at the edge towards user.




https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

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




https://t.me/learningnets

# Layer 3 Access Design/Routed Access Design

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

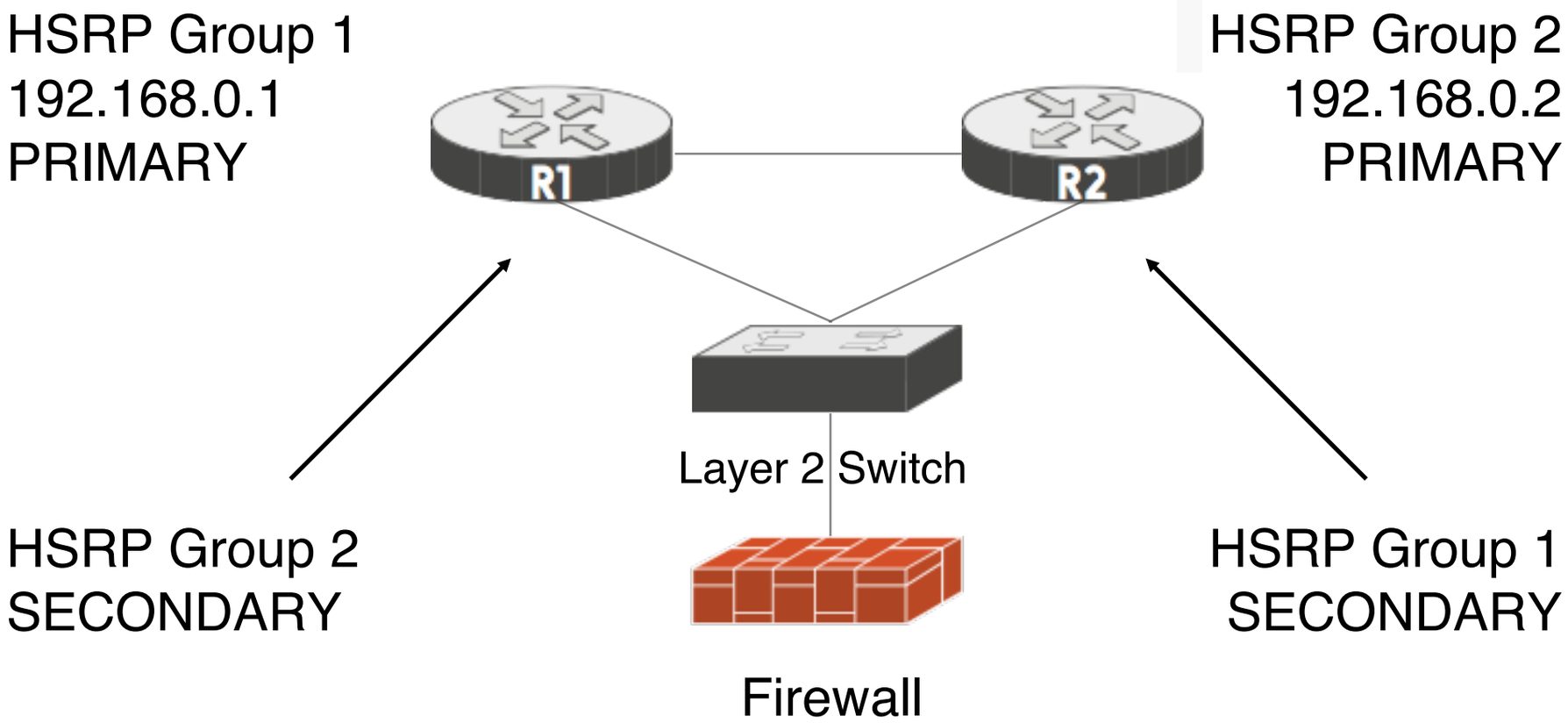
- Host based overlays such as VXLAN, NvGRE, STT and Geneve will be explained in the VPN Design Course.

https://t.me/learningnets

# LAYER 2 CASE STUDIES

https://t.me/learningnets

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




https://t.me/learningnets

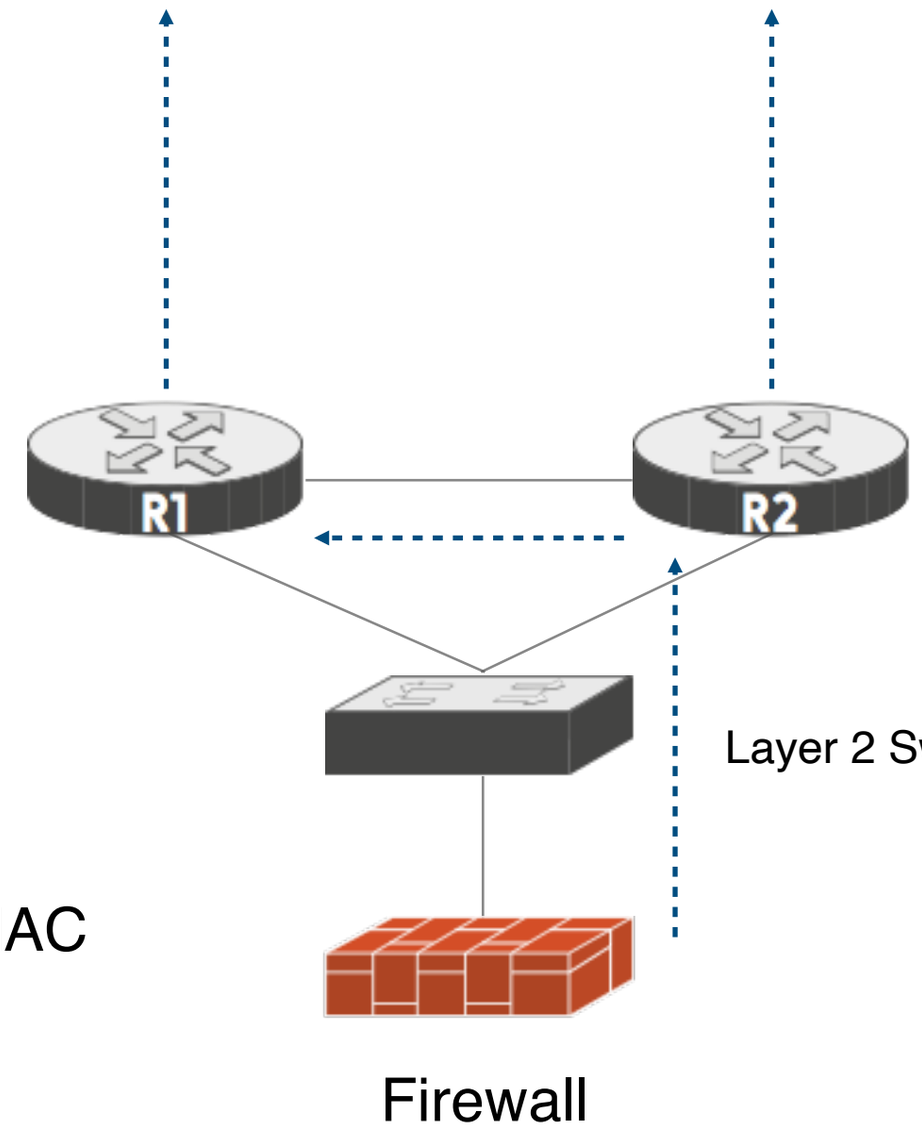
- Create two HSRP Groups both routers; each router is active for one of the HSRP Groups egress from firewall: Static routes on FW to HRSP Group; BGP handles outbound forwarding.

https://t.me/learningnets

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

https://t.me/learningnets

Virtual MAC
0000.0000.0001




Virtual MAC
0000.0000.0002

Default Gateway MAC
0000.0000.0002

https://t.me/learningnets

- What about Gateway Load Balancing Protocol (GLBP)?

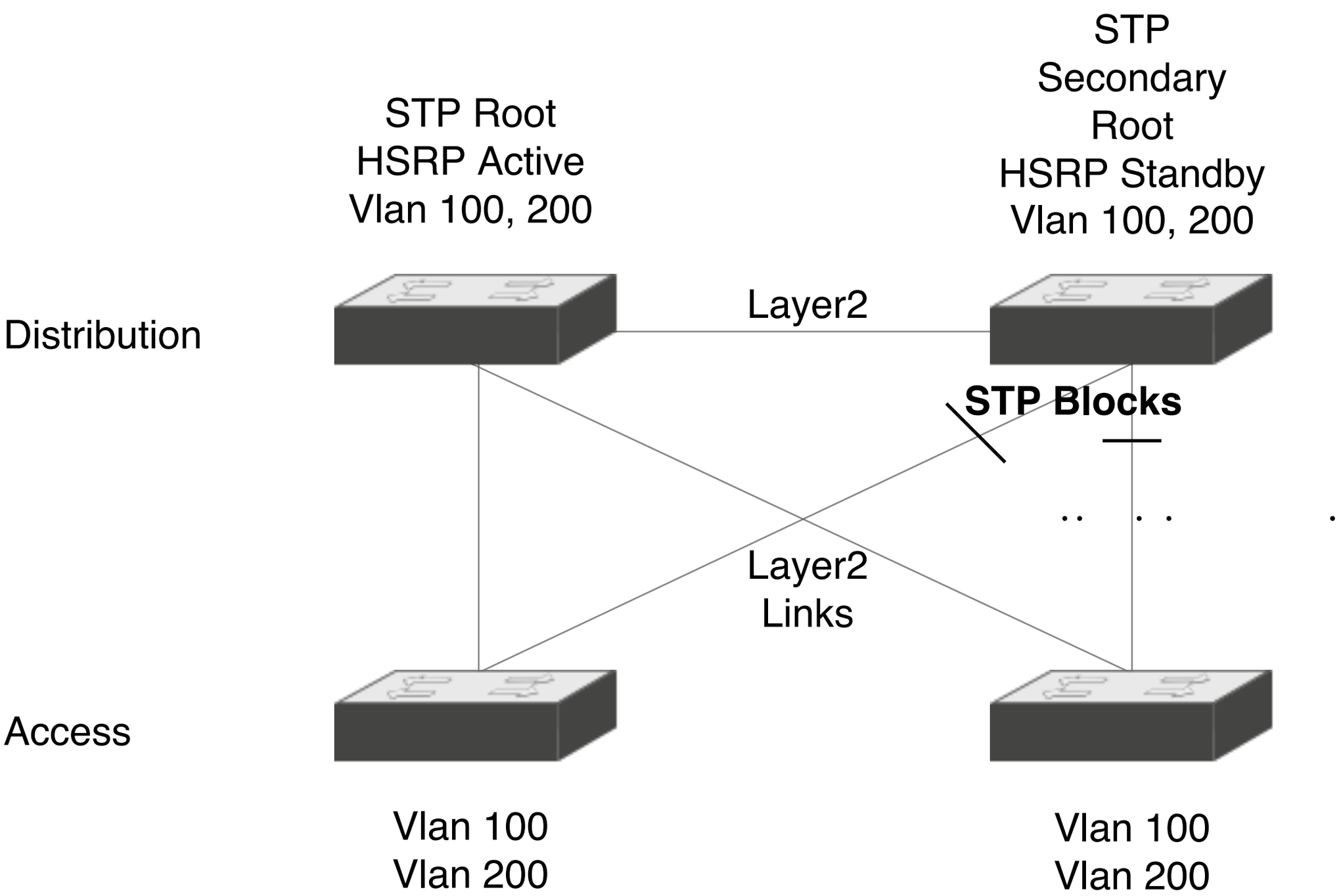
GLBP

- 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.

https://t.me/learningnets

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

https://t.me/learningnets




https://t.me/learningnets

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

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

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

https://t.me/learningnets

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

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

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

https://t.me/learningnets

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

https://t.me/learningnets

- Some users from the particular vlan use one distribution switch as an active default gateway and other users within the same vlan use previously standby switch as an active switch.

  - In this way you can use both distribution switches as active-active and you can utilize all the links in the layer 2 networks but supporting this configuration instead of using GLBP is more complex from the design point of view.

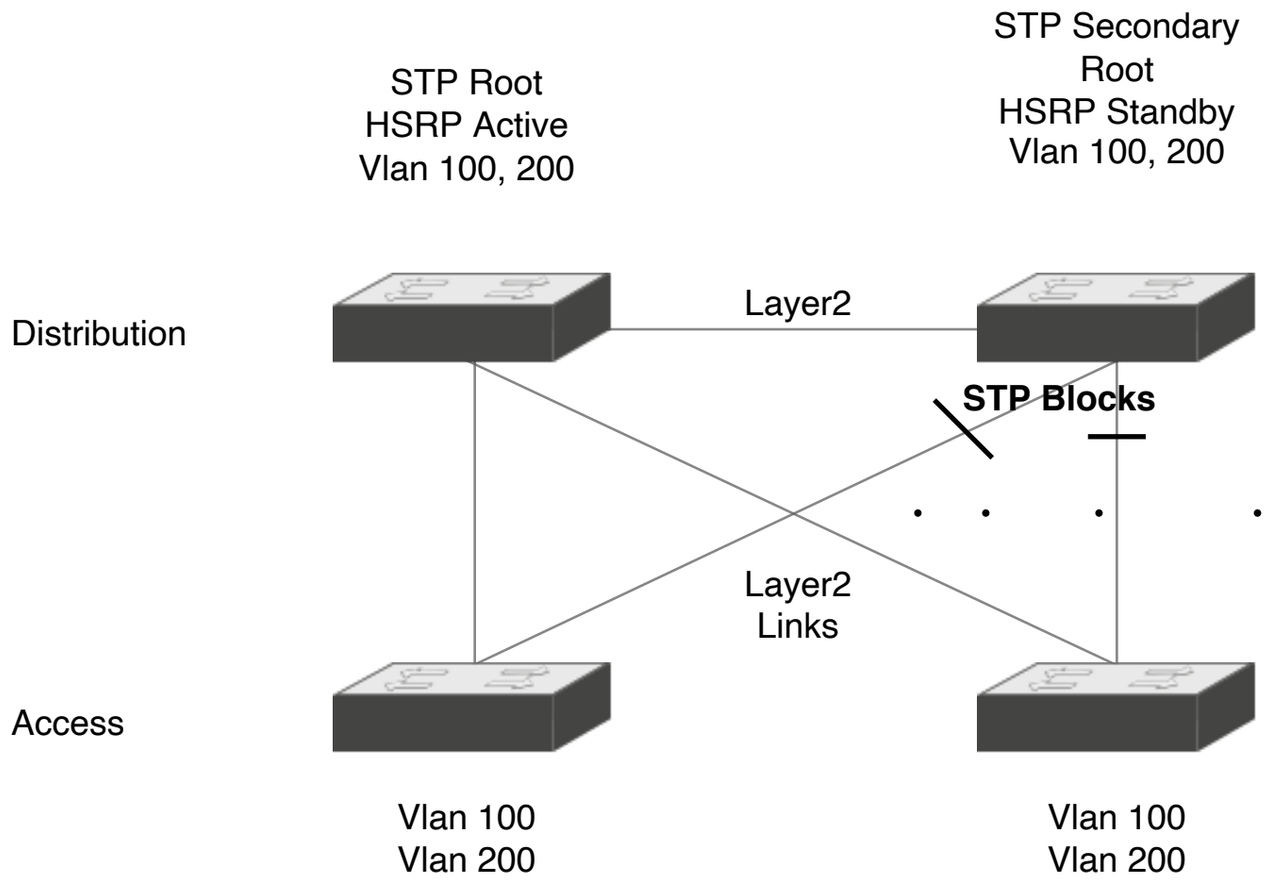
- If you want both right and left distribution switches to be used active-active for the same Vlan, let's say Vlan 100, then you need to use GLBP.

https://t.me/learningnets

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

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

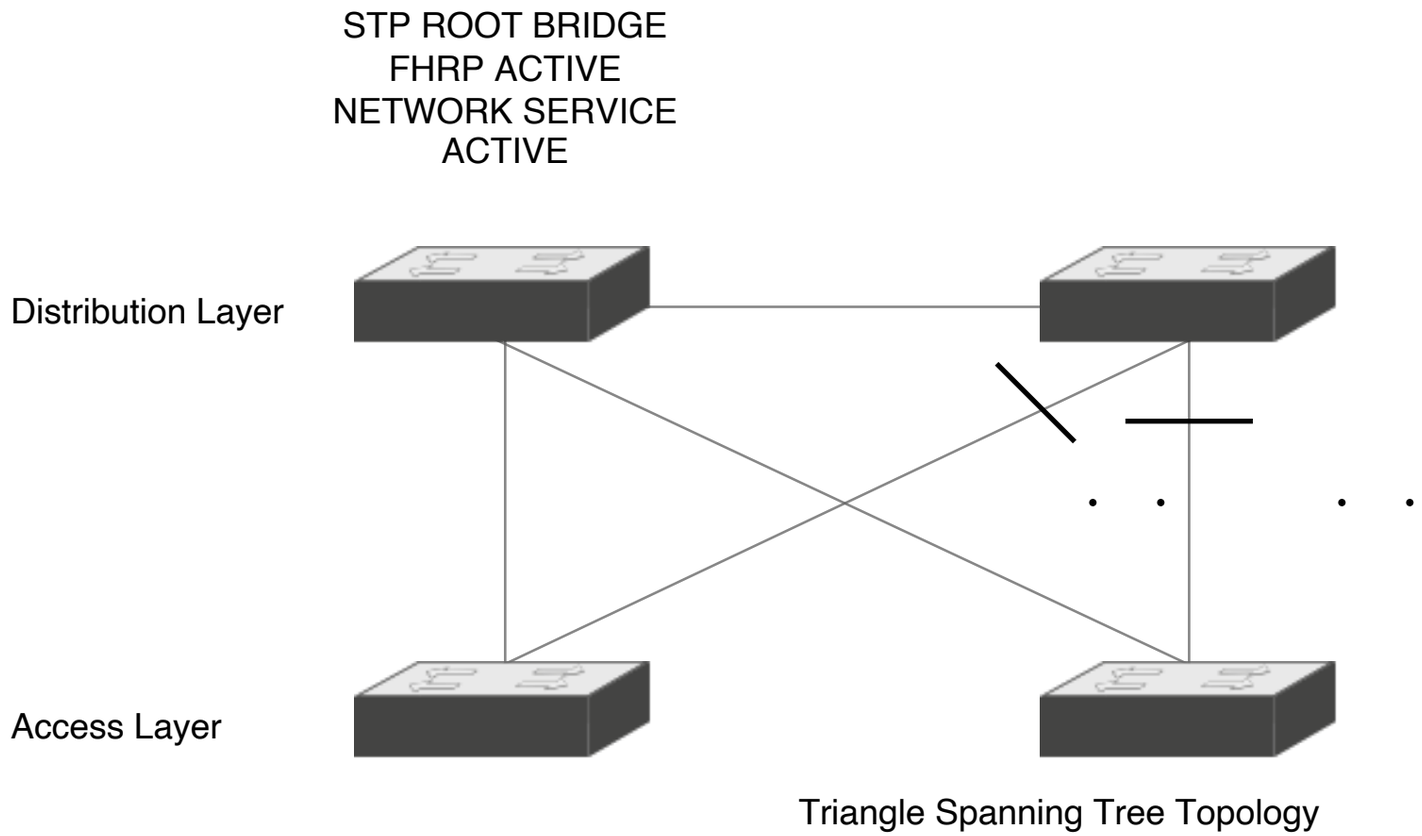
https://t.me/learningnets




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

https://t.me/learningnets

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




Triangle Spanning Tree Topology

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

https://t.me/learningnets

# QUESTION

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

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

https://t.me/learningnets

# SOLUTION

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

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

https://t.me/learningnets

- With distribution layer, any access layer switches can be upgraded smoothly. Also functions are shared between the access and distribution layer devices.

  - Access layer provides edge functions such as filtering, client access, QoS, security features.

- Distribution layer responsible from the route and traffic/speed aggregation.

https://t.me/learningnets

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

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

---

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

https://t.me/learningnets

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

https://t.me/learningnets

# First Hop Redundancy Protocol Damage

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

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

https://t.me/learningnets

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

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

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

https://t.me/learningnets

- But with GLBP, in a given Vlan, there can be more than 2 active gateways. That's why clients traffic can be divided among the active gateways.

  - If failure happens in a GLBP enabled network, only some of the clients traffic in a given vlan is effected. If there is two active gateways, only half of them will be affected.

- Thus for the requirements given in the question, GLBP is the best choice.

https://t.me/learningnets

## Port-Security comes to the rescue

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

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

https://t.me/learningnets

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

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

https://t.me/learningnets

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

https://t.me/learningnets

- But in the requirement of this case study it is clearly told that bpdu is not generated. In this case Port-security helps.

  - Port-security doesn’t allow two mac addresses to be shown on the two ports of a given switch. If it happens, port-security feature shutdown the ports.

- That’s why it is one of the best practices to enable port-security not only as security feature but spanning tree feature as well.

https://t.me/learningnets

# Layer 2 Looped Design Use Case

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

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

https://t.me/learningnets

- In the data centers hosts (Specifically Virtual Machines) can move between access switches. Vlans should be spread on those switches.

  - Also it is very common in the campus environment where WLAN is used commonly on every access switches.

- In these environments where layer 2 needs to be extended on many access switches, layer 2 looped design is the only design option.

https://t.me/learningnets

# NETWORK DESIGN TOOLS & THE BEST PRACTICES

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

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

https://t.me/learningnets

# Reliability

- Within the reasonable amount of time, which depends on the application type and architecture, delivering the legitimate packets from source to destination.
    - This time is known as delay or latency and it is one of the packet delivery parameters. Consistency of delay known as jitter and it is very important for some type of applications such as voice and video, jitter is our second delivery parameters

reli
abil
ity

https://t.me/learningnets

# Reliability

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

reli
abil
ity

https://t.me/learningnets

# Effect of packet loss on IPTV Service

https://t.me/learningnets

# Reliability

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

reli
abil
ity

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

https://t.me/learningnets

# Reliability

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

reli
abil
ity

https://t.me/learningnets

# Reliability

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

reli
abil
ity

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

https://t.me/learningnets

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

---

https://t.me/learningnets

# Reliability

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

reli
abil
ity

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

https://t.me/learningnets

# Resiliency

- Resiliency is how the network behaves once the failure happen. Is that highly available, will it convergence and when?

  - Resilience can be considered as combination of high availability and convergence.

- In order any network design to be resilient, it should be redundant and converge fast enough to avoid application timeout.

https://t.me/learningnets

# Resiliency

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

https://t.me/learningnets

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

---

https://t.me/learningnets

# Fast Convergence & Reroute

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

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

https://t.me/learningnets

# Fast Convergence & Reroute

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

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

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

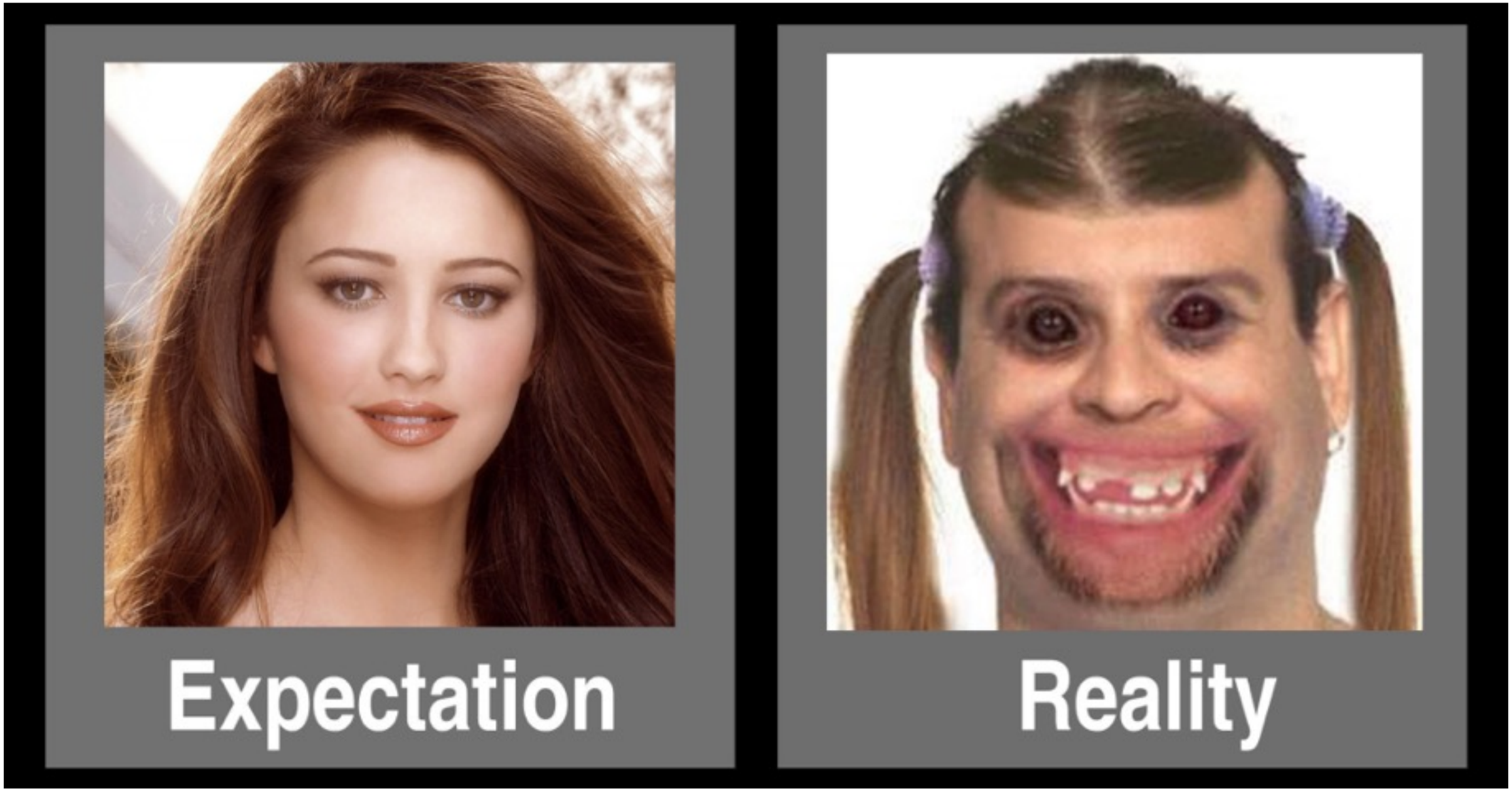
https://t.me/learningnets

# Fast Convergence & Reroute

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

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

https://t.me/learningnets




https://t.me/learningnets

# Fast Convergence & Reroute

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

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

https://t.me/learningnets

# Scalability

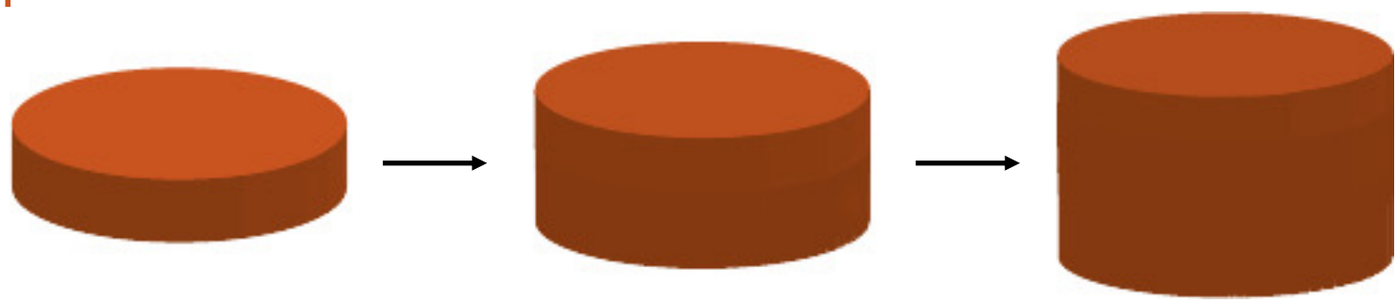
- Scalability is the ability to change, modify or remove the part of entire system without having a huge impact on the overall design.

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

https://t.me/learningnets

# Scale up & Scale out

Scale up



Scale out

https://t.me/learningnets

# Scalability

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

https://t.me/learningnets

# Scalability

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

https://t.me/learningnets

# Cost-Opex & Capex

- Cost is generally afterthought in network design. But most of the time it is very important constraint in the projects. If we breakdown the components of cost we have OPEX and CAPEX.

  - OPEX refers to operational expenses such as support, maintenance, labor, bandwidth and utilities.

- Creating a complex network design may show off your technical knowledge but it can also cause unnecessary complexity making it harder to build, maintain, operate and manage the network.

https://t.me/learningnets

https://t.me/learningnets

# Cost-Opex & Capex

- A well- designed network reduces OpEx through improved network uptime (which in turn can avoid or reduce penalties related to outages), higher user productivity, ease of operations, and energy savings.

  - Consider creating the simplest solution that meets the business requirements.

- In the CCDE exam, which solution is best from the OPEX point of view type of questions are common!

https://t.me/learningnets

# Cost-Opex & Capex

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

cost
ope
xan
dca
pex

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

https://t.me/learningnets

# TCO - Total cost of ownership

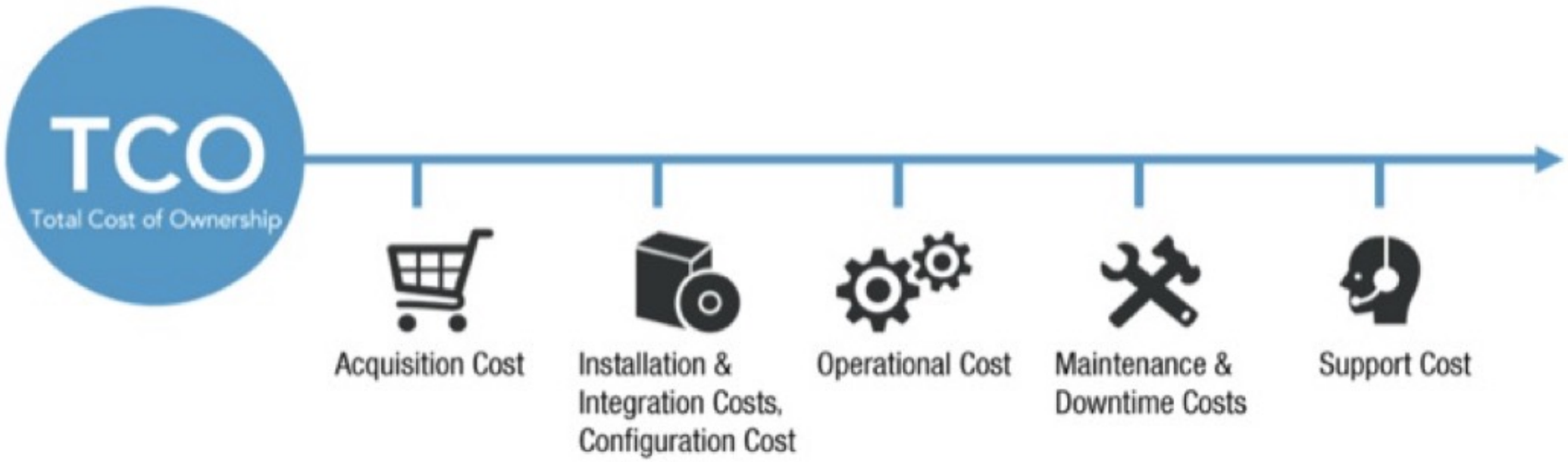
- Last metric in the COST constraint is TCO (Total cost of ownership)

cost
opex
capex

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

https://t.me/learningnets

# TCO - Total cost of ownership




https://t.me/learningnets

# Flexibility

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

flex
ibili
ty

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

https://t.me/learningnets

# Flexibility

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

flexibility

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

https://t.me/learningnets

# Modularity

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

modularity

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

https://t.me/learningnets

# Modularity

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

https://t.me/learningnets

# Modularity

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

https://t.me/learningnets

# Modularity

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

https://t.me/learningnets

# Modularity

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

https://t.me/learningnets

# Modularity

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

https://t.me/learningnets

# Design Considerations For Network Mergers & Acquisitions

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

https://t.me/learningnets

# Design Considerations For Network Mergers & Acquisitions

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

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

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

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

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

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Where will be the first place in the network for the merger?

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

- Which type of security infrastructure will merged network support?

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

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

https://t.me/learningnets

# Key Points For Any Type Of Network Mergers & Acquisitions Projects

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

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

https://t.me/learningnets

# High Availability

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

https://t.me/learningnets

# High Availability

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

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

https://t.me/learningnets

# High Availability

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

https://t.me/learningnets

# Convergence

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

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

https://t.me/learningnets

# Convergence

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

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

https://t.me/learningnets

# Convergence

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

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

https://t.me/learningnets

# Convergence

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

https://t.me/learningnets

# Some Routing Best Practices

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

https://t.me/learningnets

# Some Routing Best Practices

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

https://t.me/learningnets

# Some Routing Best Practices

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

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

https://t.me/learningnets

# Load Balancing

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

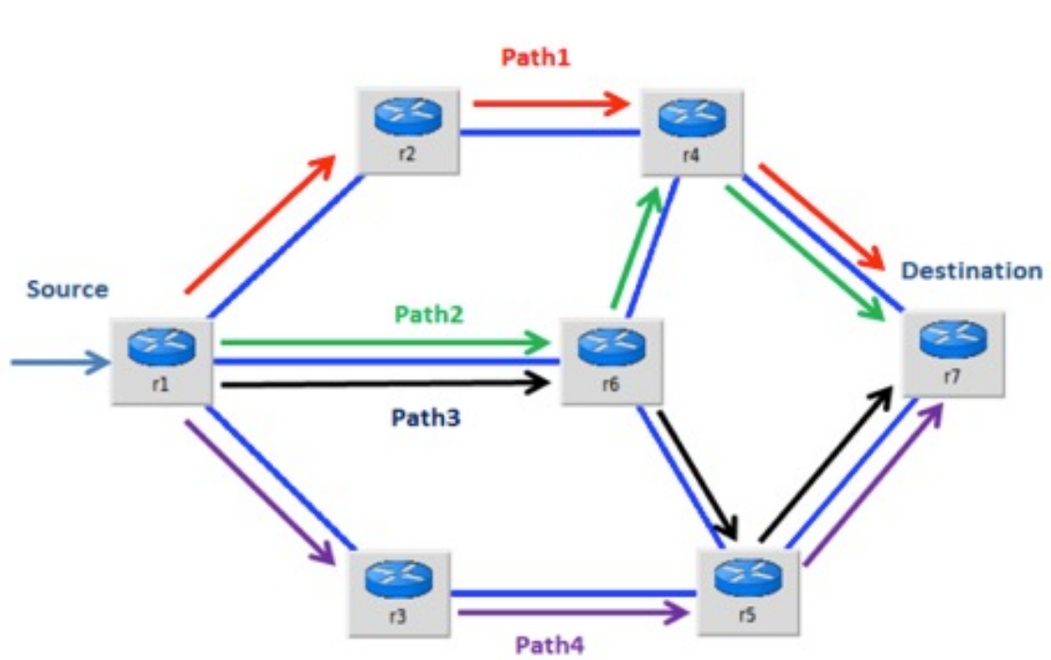
load balancing

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

https://t.me/learningnets

# ECMP (Equal Cost Multipath)

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




load balancing

https://t.me/learningnets

# Redistribution

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

redi
stri
buti
on

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

https://t.me/learningnets

# Redistribution

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

redistribution

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

https://t.me/learningnets

# Redistribution

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

redistribution

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

https://t.me/learningnets

# Redistribution

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

redistribution

https://t.me/learningnets

# Optimal Routing

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

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

https://t.me/learningnets

# Optimal Routing

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

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

https://t.me/learningnets

# Optimal Routing

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

optimal

https://t.me/learningnets

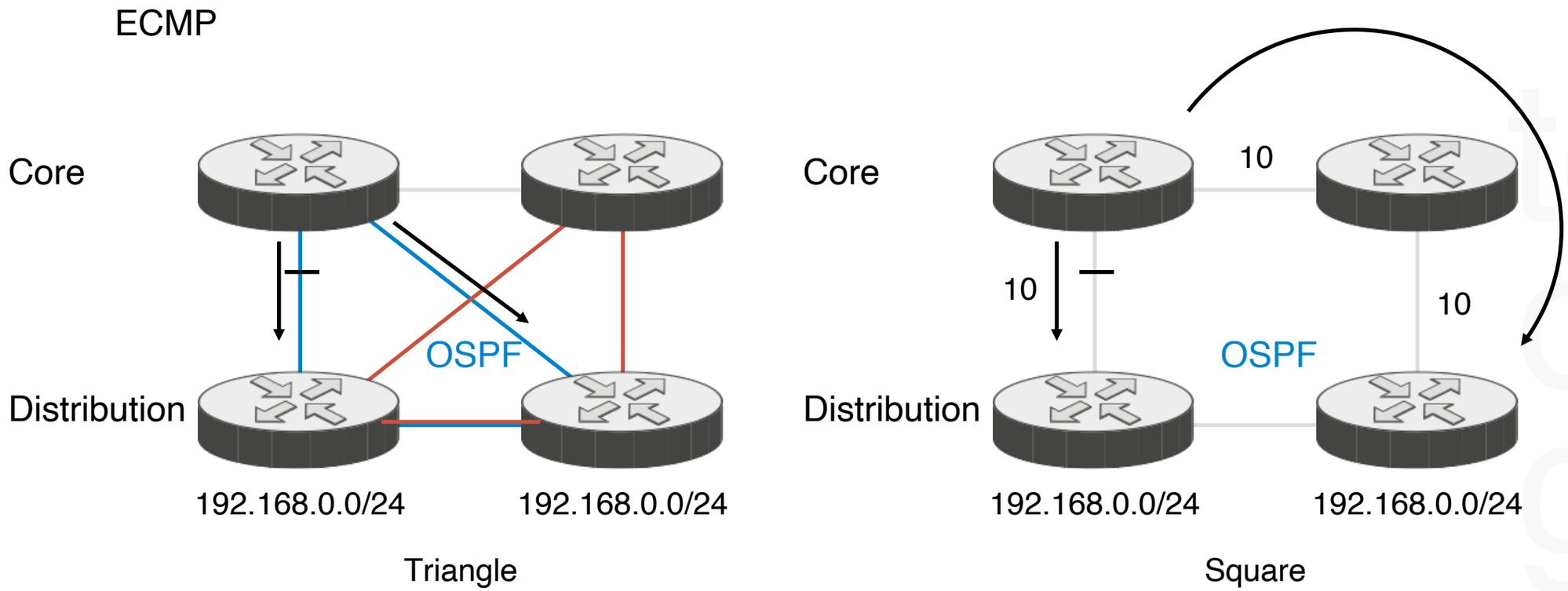
# Network Topologies

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

https://t.me/learningnets

# Network Topologies

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




https://t.me/learningnets

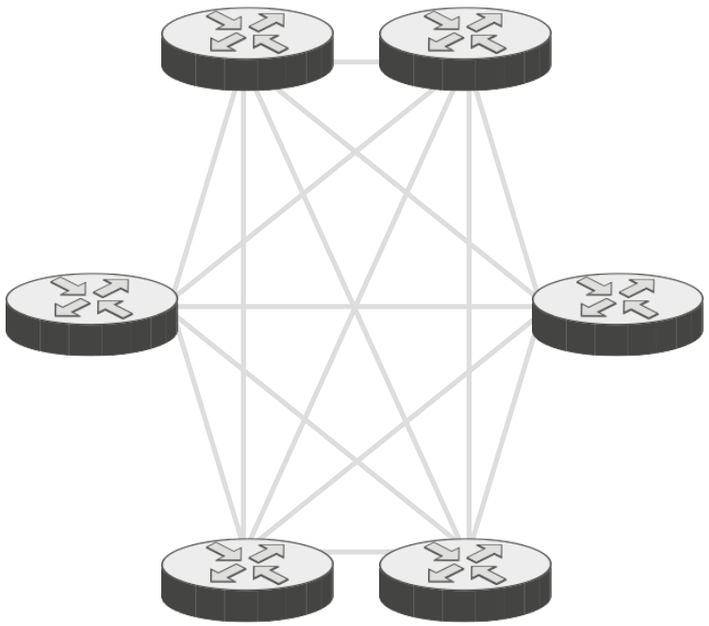
# Network Topologies

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

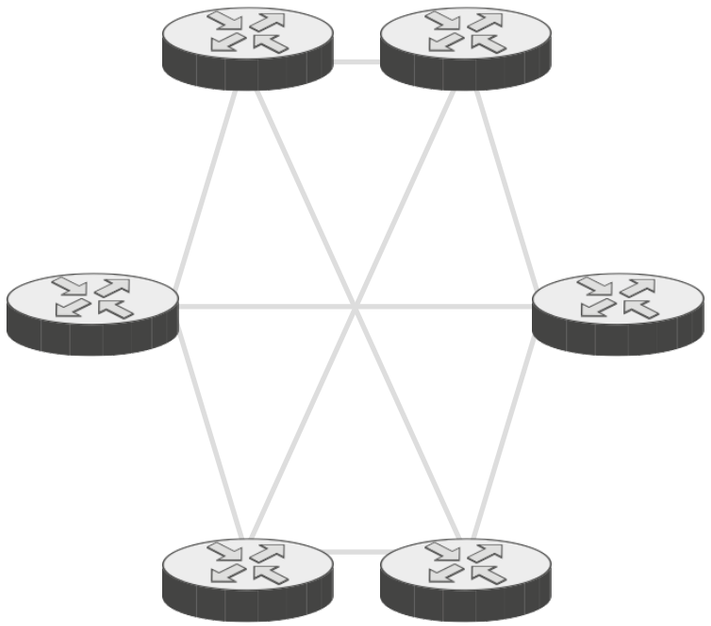
  - Simply adding some links and creating partial-mesh topology instead of ring provides a more optimal path, better resource usage, and faster convergence time in case of link or node failure.

https://t.me/learningnets

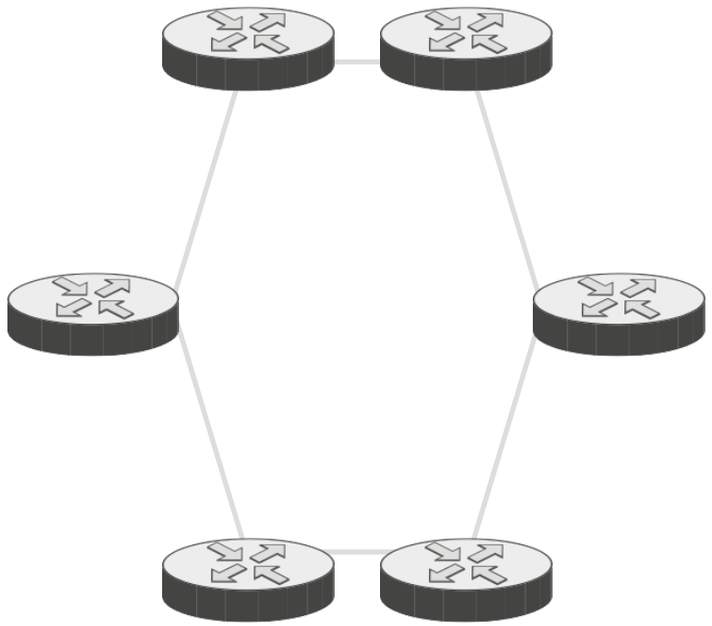
# Common Access & WAN Topologies



Full Mesh Topology



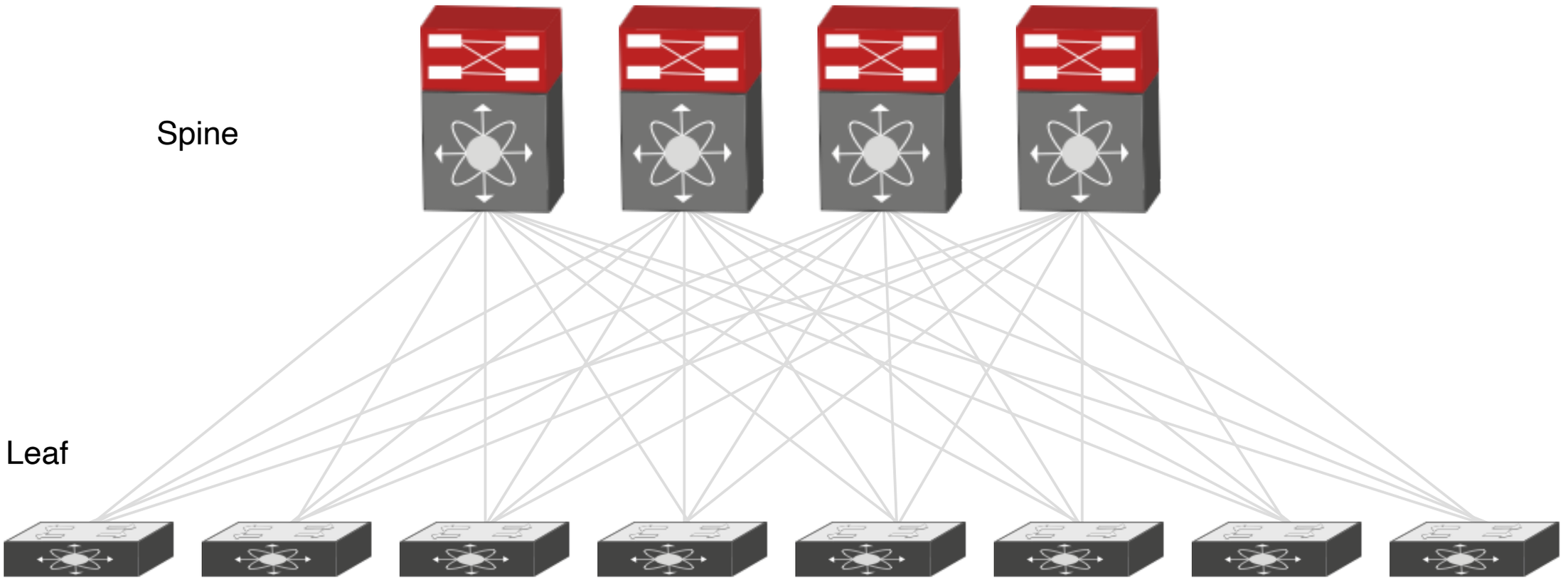
Partial Mesh Topology



Ring Topology

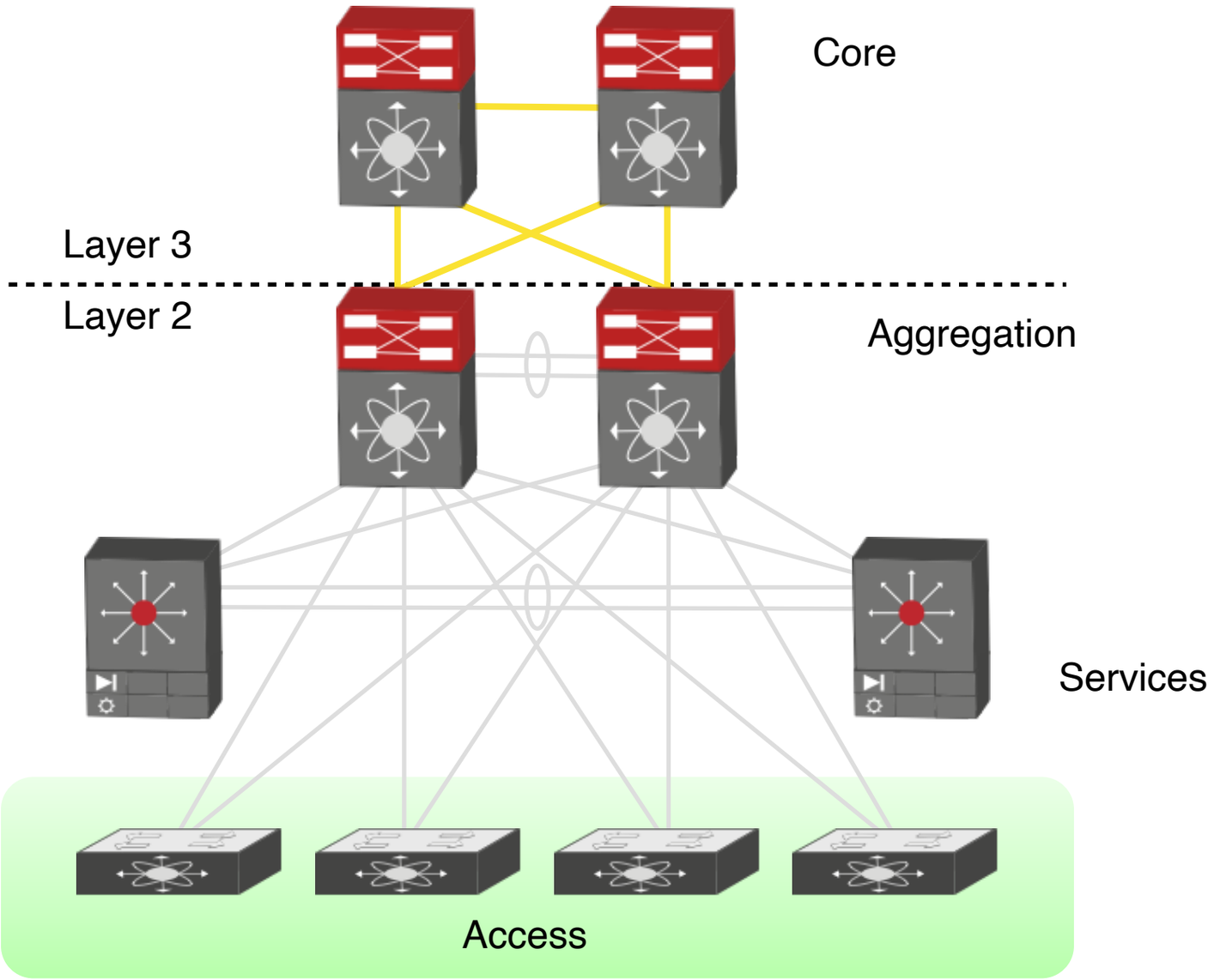
https://t.me/learningnets

# DC Topologies Clos




https://t.me/learningnets

# DC Topologies Three Tiers




https://t.me/learningnets

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

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

https://t.me/learningnets

# Security

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

https://t.me/learningnets

# Security

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

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

https://t.me/learningnets

- Always use BGP hardening features on the Internet Gateways.
  - BGP Ingress filtering provides protection against route leaking and hijacking.

https://t.me/learningnets

# Security

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

https://t.me/learningnets

## Security

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

security

https://t.me/learningnets

## Simplicity & Complexity

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

https://t.me/learningnets

# Push & Pull Based Control Plane Protocols

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

https://t.me/learningnets

## Simplicity & Complexity

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

https://t.me/learningnets

# Simplicity & Complexity

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

https://t.me/learningnets

# Simplicity & Complexity

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

https://t.me/learningnets

# Simplicity & Complexity

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

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

https://t.me/learningnets

# Simplicity & Complexity

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

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

https://t.me/learningnets

# Simplicity & Complexity

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

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

https://t.me/learningnets

# Simplicity & Complexity

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

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

https://t.me/learningnets

# NETWORK DESIGN TOOLS & THE BEST PRACTICES

QUIZ

https://t.me/learningnets

## Question 1

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

Which below statement is true for the below figure?




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

B. IS-IS would be a better option

C. More links provide better resiliency

D. More links provide faster convergence

E. More links provide better security

https://t.me/learningnets

# Answer 1

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

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

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

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

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

https://t.me/learningnets

## Question 2

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

A. BFD

B. Routing fast hellos

C. Loopguard

D. SPF Timers

E. BGP Scanner time

https://t.me/learningnets

# Answer 2

A. BFD

B. Routing fast hellos

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

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

https://t.me/learningnets

## Question 3

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

A. Static Routing

B. OSPF

C. IS-IS

D. EIGRP

E. BGP

F. RIP

https://t.me/learningnets

# Answer 3

A. Static Routing

B. OSPF

C. IS-IS

D. EIGRP

E. BGP

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

RIPv2 on the other hand supports BFD.

https://t.me/learningnets

## Question 4

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

A. Each module can be designed independently from each other

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

C. Each module can have a separate routing protocol

D. Each module can have different security policy

https://t.me/learningnets

# Answer 4

A. Each module can be designed independently from each other

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

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

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

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

https://t.me/learningnets

## Question 5

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

A. Predictability increases security

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

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

D. Modular network design reduces deployment time

https://t.me/learningnets

# Answer 5

A. Predictability increases security

D. Modular network design reduces deployment time

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

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

https://t.me/learningnets

## Question 6

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

A. Deploying Spanning Tree

B. Deploying Fast Reroute

C. Implementing Route tags

D. Only one way redistribution is enough

https://t.me/learningnets

## Answer 6

C. Implementing Route tags

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

https://t.me/learningnets

## Question 7

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

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

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

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

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

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

https://t.me/learningnets

# Answer 7

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

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

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

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

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

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

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

https://t.me/learningnets

## Question 8

What is the key benefit of hierarchical network design?

A. less Broadcast traffic

B. Increased flexibility and modularity

C. Increased security

D. Increased availability

https://t.me/learningnets

## Answer 8

B.Increased flexibility and modularity

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

https://t.me/learningnets

## Question 9

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

A. Convergence will be slower

B. Sub optimal routing may occur

C. Traffic blackholing may occur

D. Routing table size grows

https://t.me/learningnets

# Answer 9

B.Sub optimal routing may occur

C.Traffic blackholing may occur

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

https://t.me/learningnets

## Question 10

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

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

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

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

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

https://t.me/learningnets

# Answer 10

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

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

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

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

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

https://t.me/learningnets

## Question 11

Which routing protocol supports unequal cost multi path routing?

A. OSPF

B. IS-IS

C. EIGRP

D. RIPv2

https://t.me/learningnets

## Answer 11

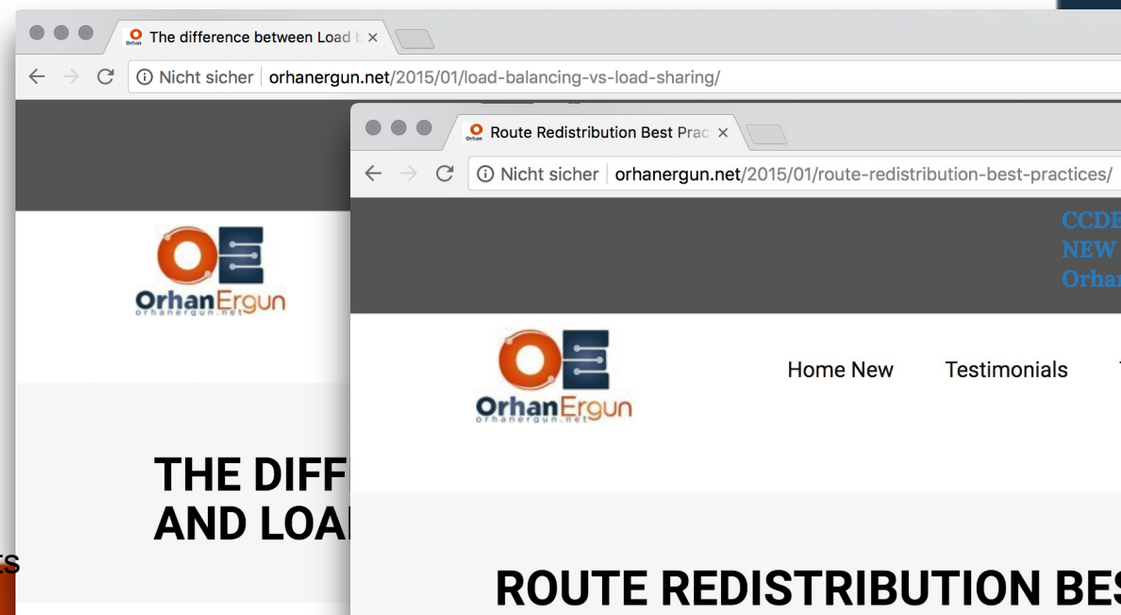
B.EIGRP

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

https://t.me/learningnets

# Extra Study Resources

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




https://t.me/learningnets

# OSPF

## Open Shortest Path First

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

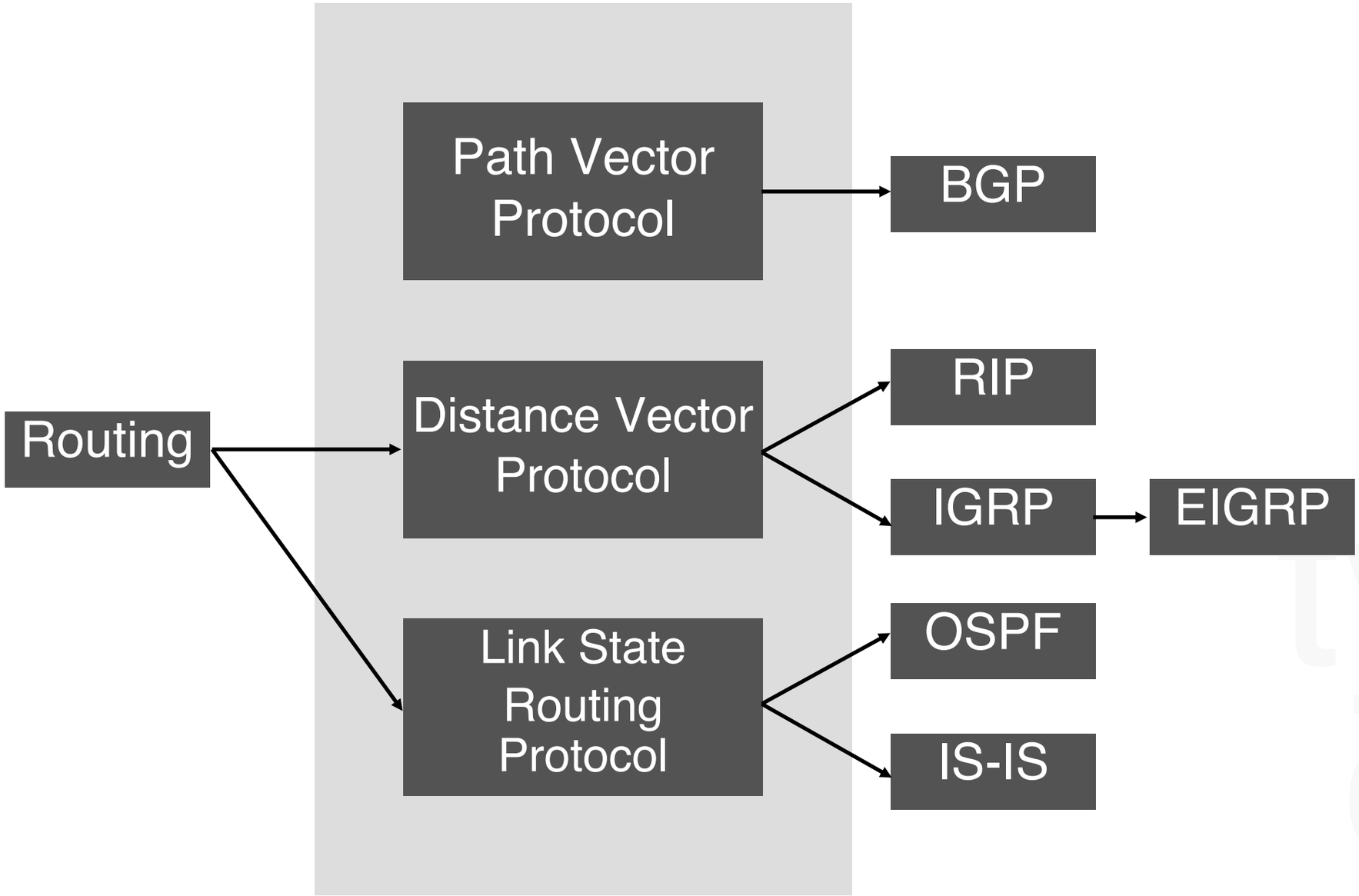
# Theory

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

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

https://t.me/learningnets

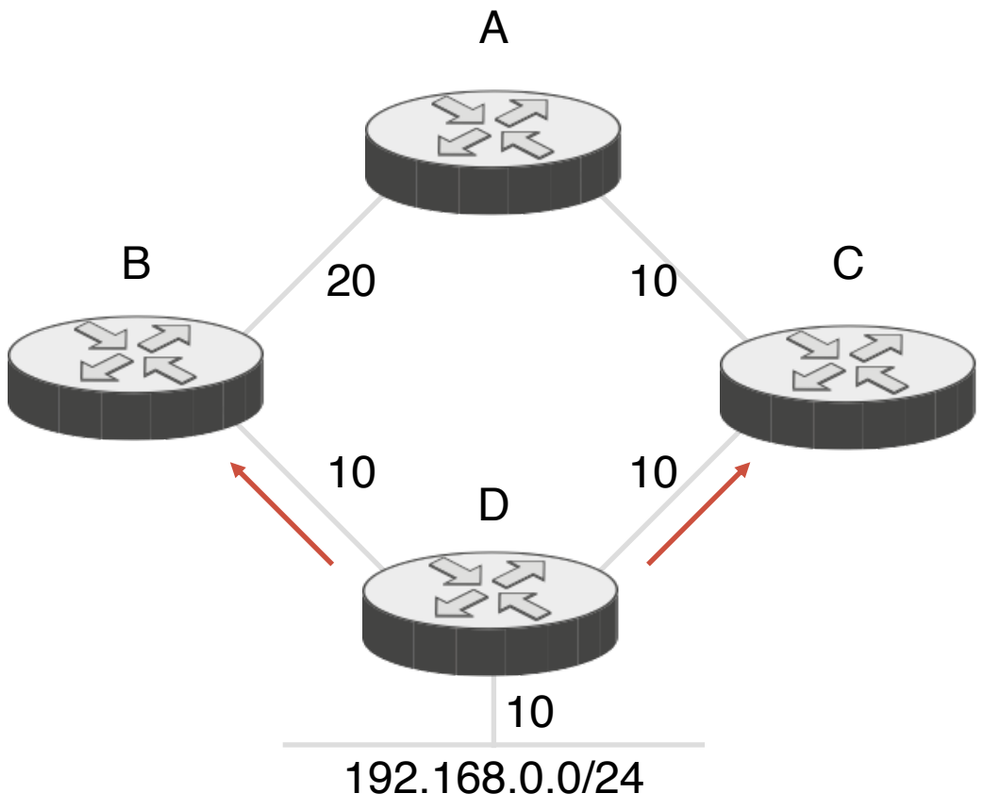
# 3 Types of Routing Protocols




https://t.me/learningnets

# LSA Flooding

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



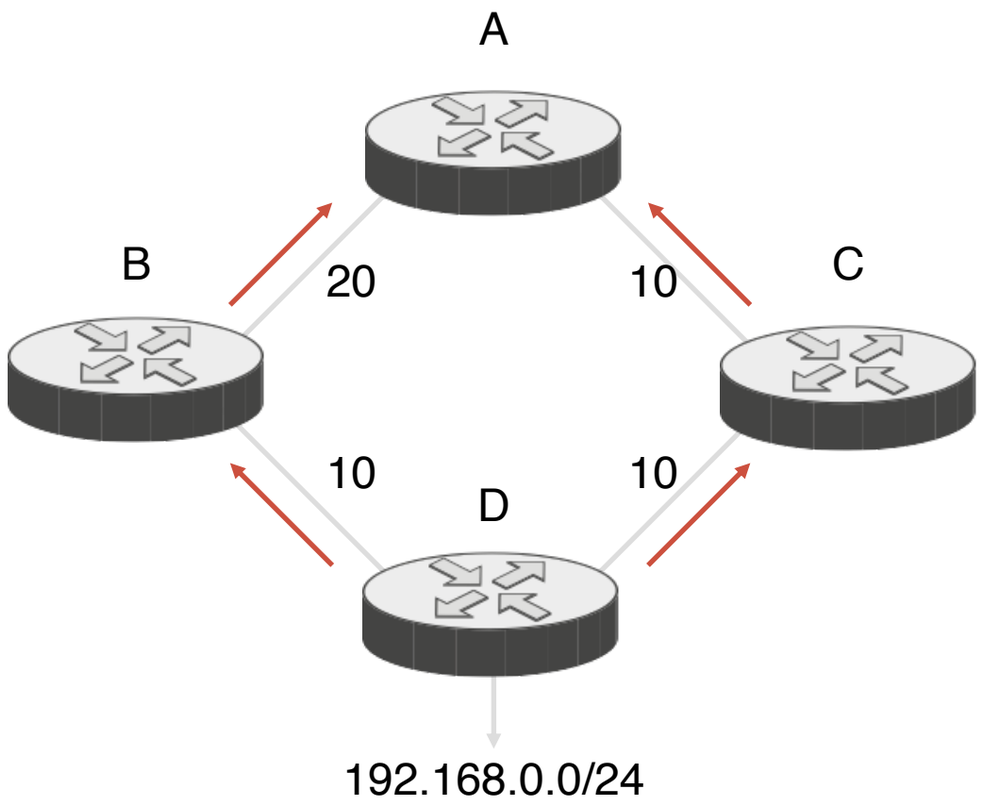

- D determines that it is connected to 192.168.0.0/24 with metric 10
- Connected to B with metric 10
- Connected to B with metric 10

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

lsa

https://t.me/learningnets

# LSA Flooding




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

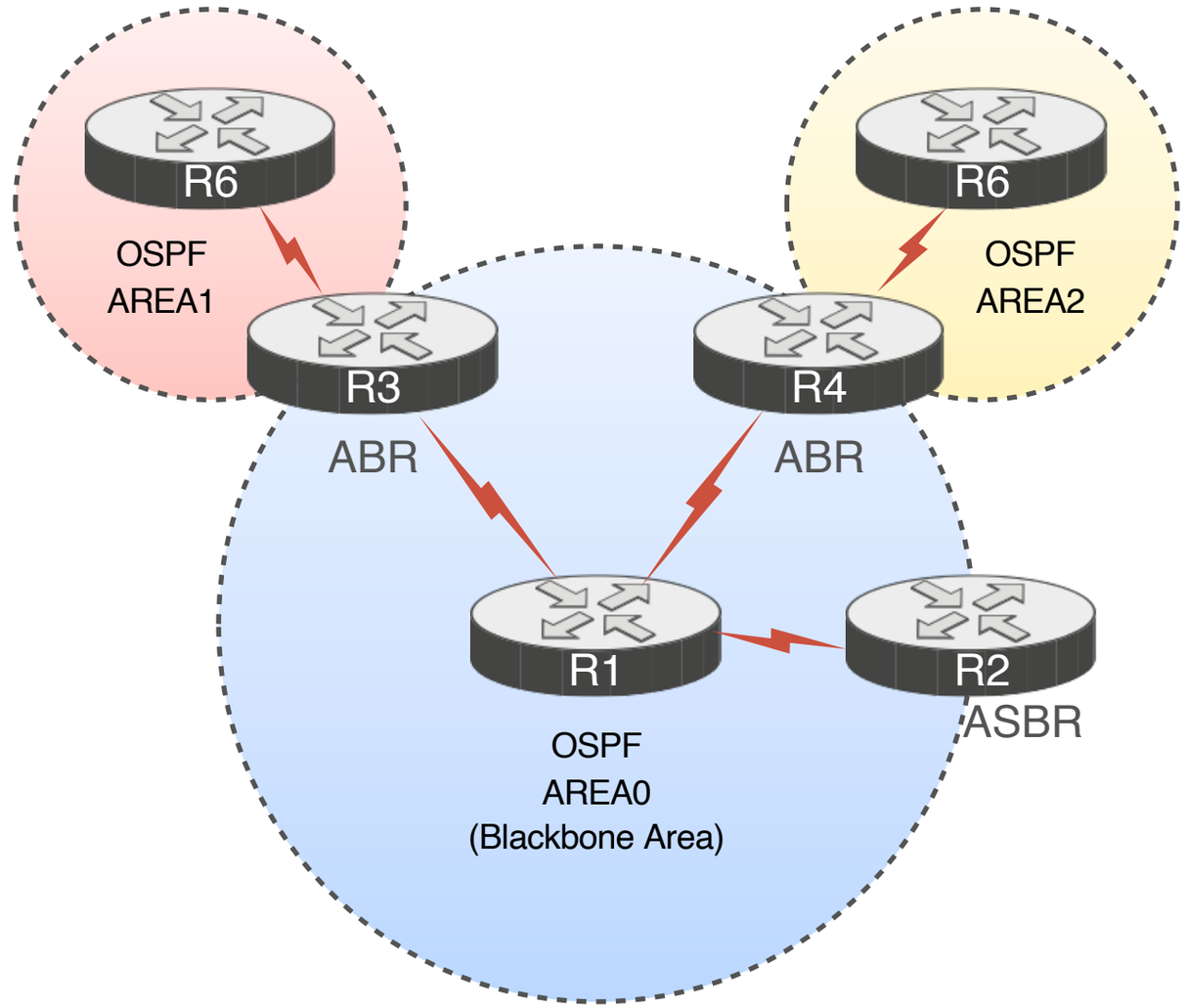
https://t.me/learningnets

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

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

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

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




https://t.me/learningnets

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

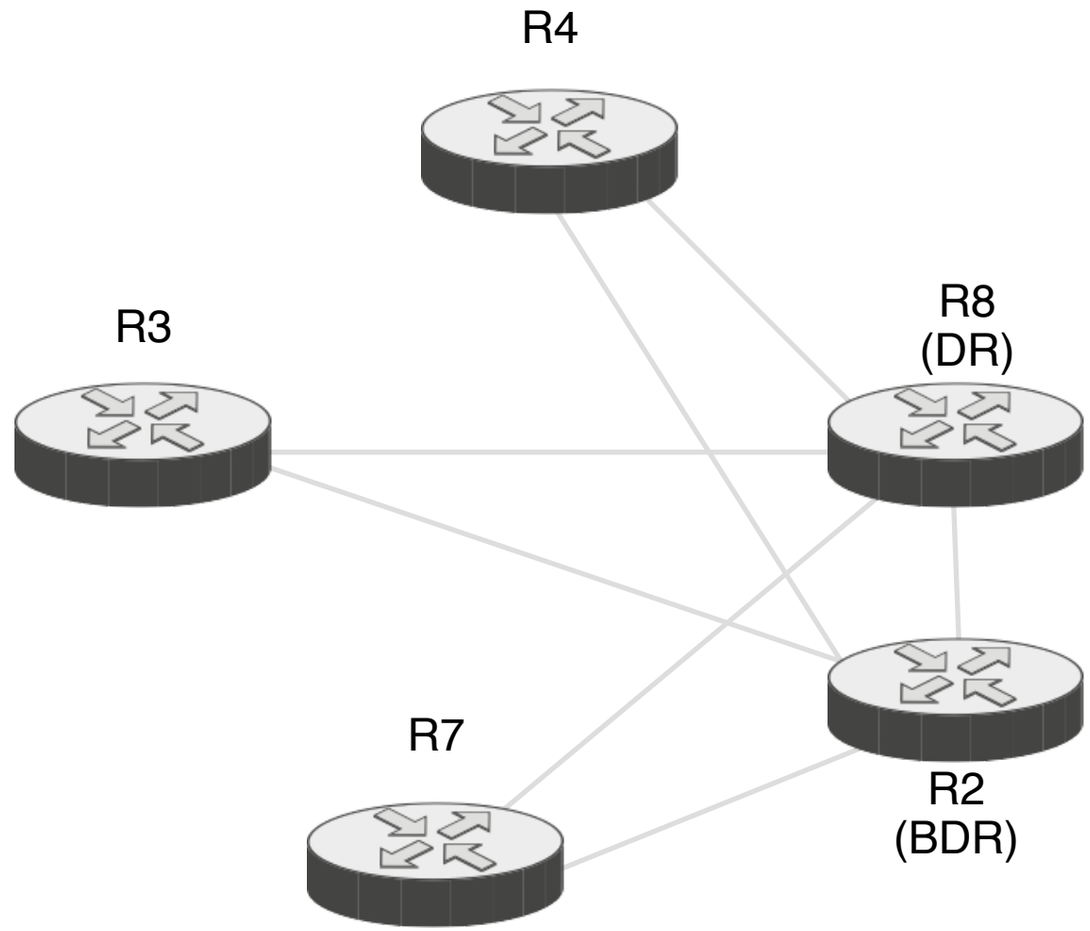
  - OSPF, unlike EIGRP and IS-IS works differently on different media. On broadcast network DR (Designated Router) creates a pseudo node to avoid unnecessary flooding.

https://t.me/learningnets

# OSPF DR & BDR

- DR creates type 2 LSA (Network LSA) to inform the connected routers on the broadcast network.

- Highest priority OSPF router on the broadcast segment wins the Designated Router (DR) election. If priorities are the same then highest router ID wins the DR election. On every broadcast segment there can be only 1 DR.

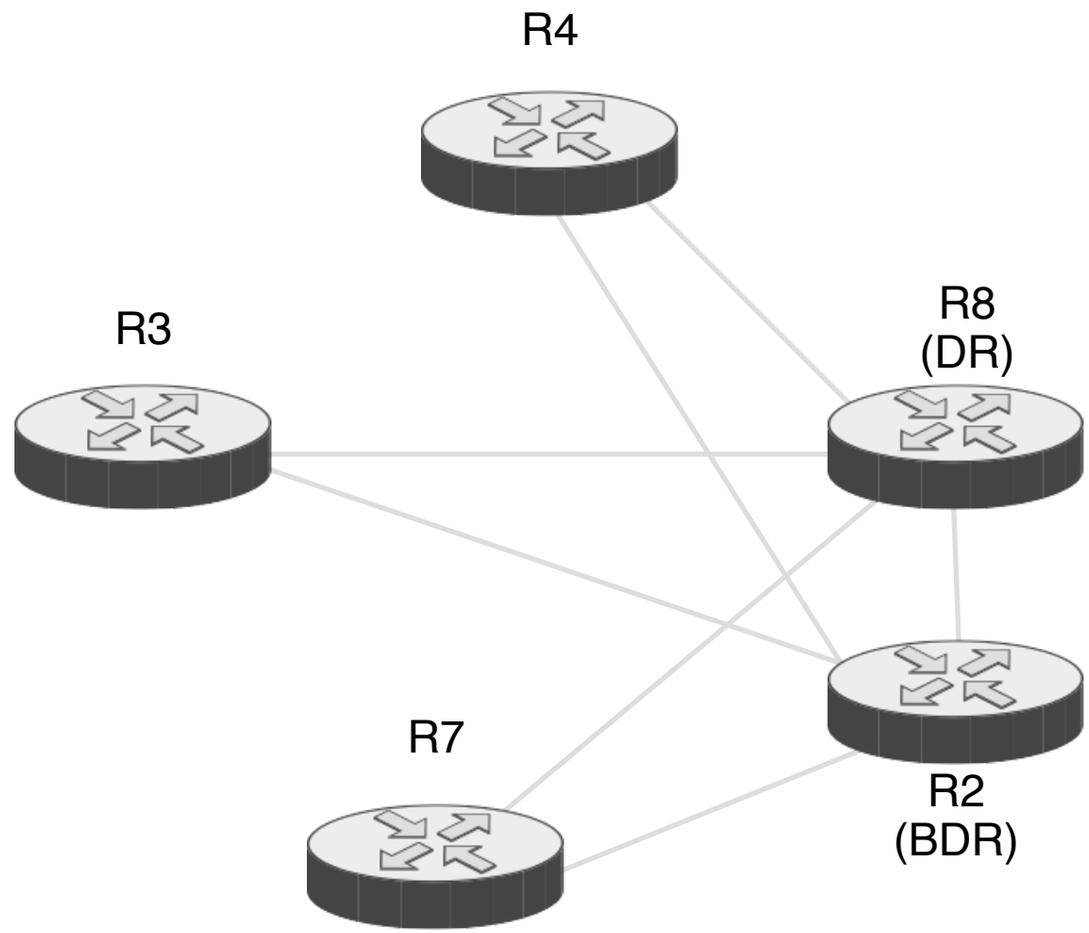



https://t.me/learningnets

# OSPF DR & BDR

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

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




https://t.me/learningnets

# DR & BDR Question

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

Direct back to back connection

https://t.me/learningnets

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

https://t.me/learningnets

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

https://t.me/learningnets

# 5 Critical LSAs for OSPF Design

| | Description |
|---|---|
| 1.Router | • Router information<br>• Connections to other routers<br>• Connections to links (link states) |
| 2.Pseudonode | • Pseudonode information<br>• Connections to routers<br>• 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)<br>• 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) |

https://t.me/learningnets

# OSPFv2 all LSA Types

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

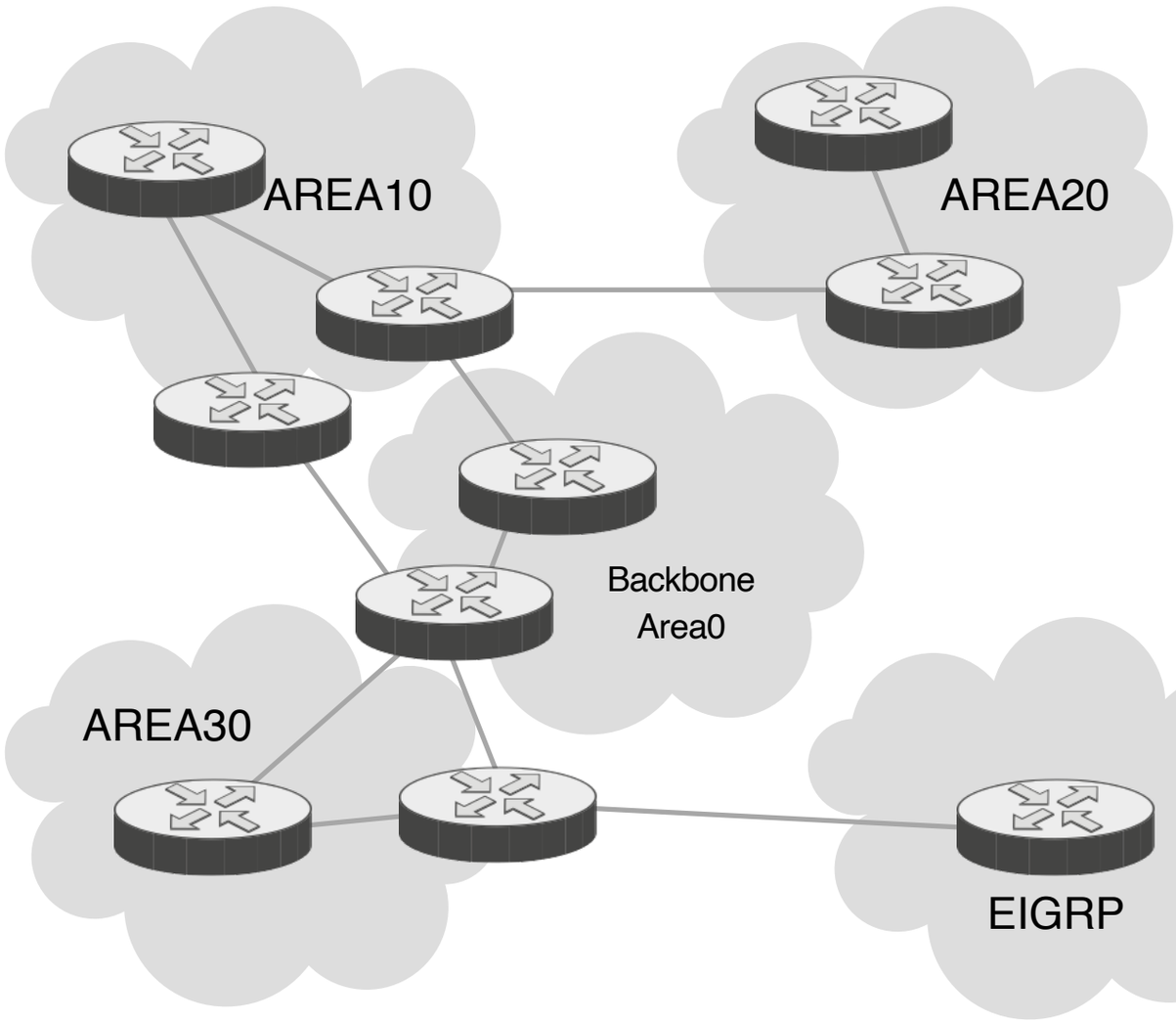
https://t.me/learningnets

# OSPF Area Types

| Area | Restriction |
|---|---|
| Normal | None |
| Stub | No type 5 LSA/No External 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 |

https://t.me/learningnets

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




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

https://t.me/learningnets

# OSPF Fast Convergence

- Network convergence is the between the failure event and the recovery.
    - Through the path all the routers process the event and update their routing and forwarding table.
- Thus; there are 4 steps for convergence in general:
    1. Failure Detection
    2. Failure Propagation
    3. Processing the new information
    4. Routing and Forwarding table update

https://t.me/learningnets

# OSPF Fast Convergence

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

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

https://t.me/learningnets

# Convergence Tools

- DETECTION

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

- PROPAGATION

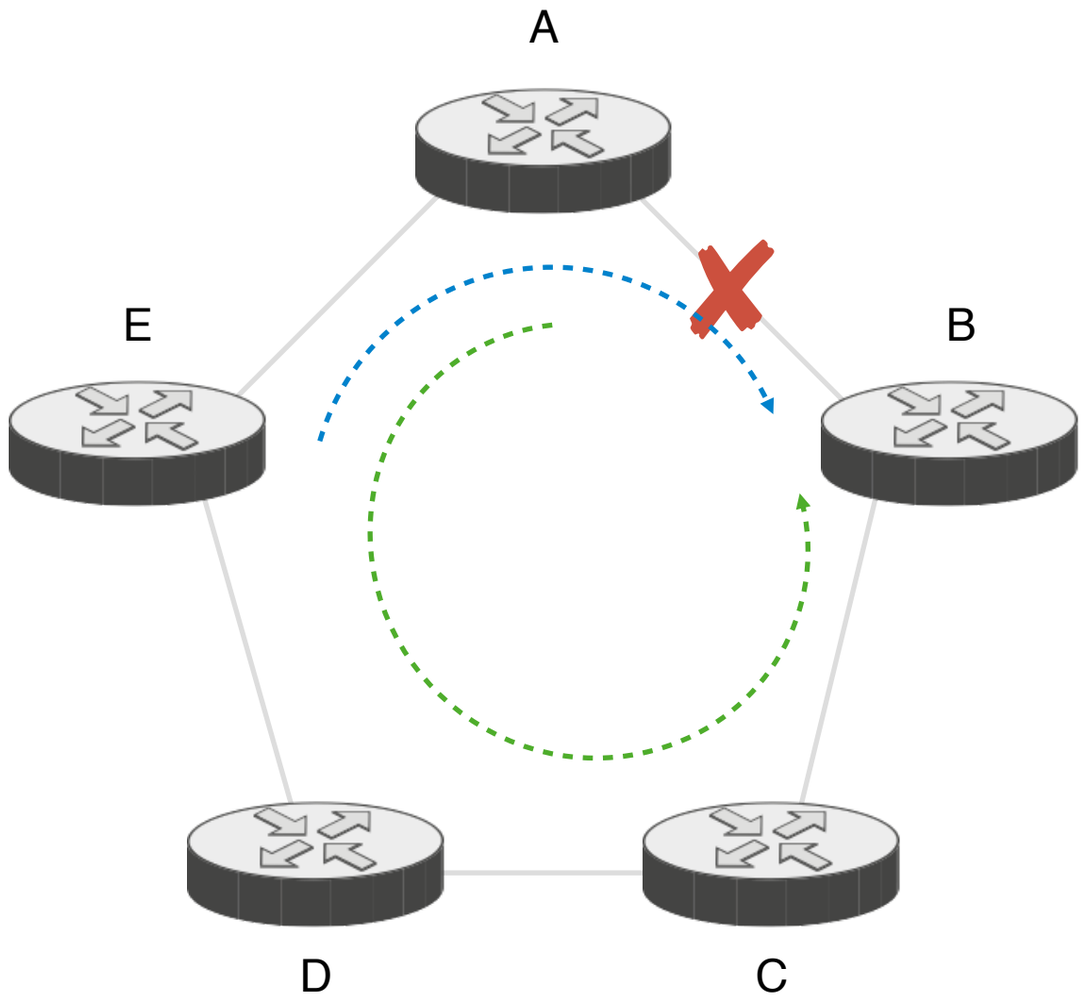
  Interface event dampening
  LSA Pacing

- PROCESSING

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

https://t.me/learningnets

# Convergence & Micro-loop




https://t.me/learningnets

# OSPF Scalability and Multi Area Design

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

https://t.me/learningnets

# OSPF Scalability and Multi Area Design

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

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

https://t.me/learningnets

# OSPF Scalability and Multi Area Design

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

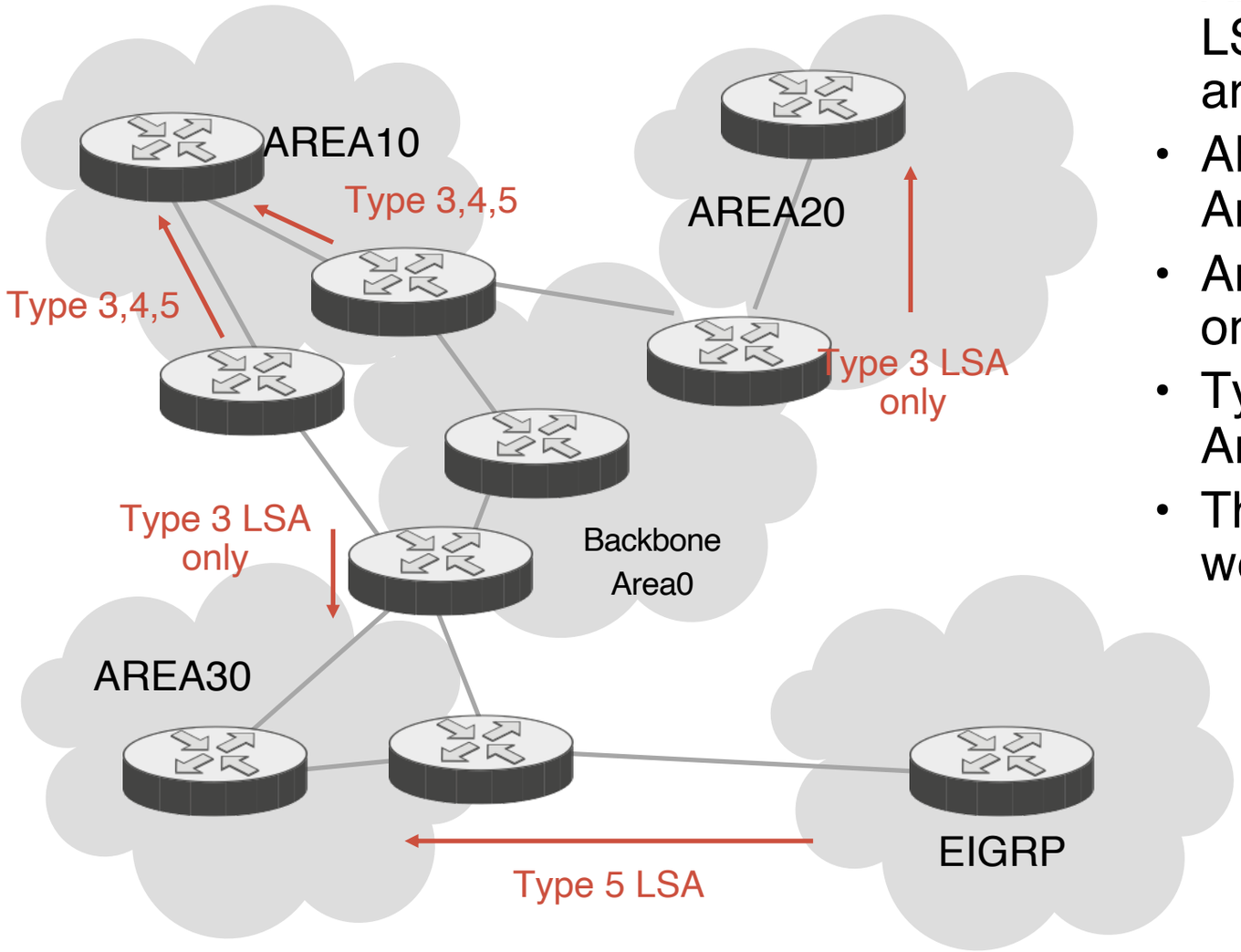
https://t.me/learningnets

# Multi Area OSPF – Fault Isolation

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

https://t.me/learningnets

# What is the problem with the below area design?




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

https://t.me/learningnets

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

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

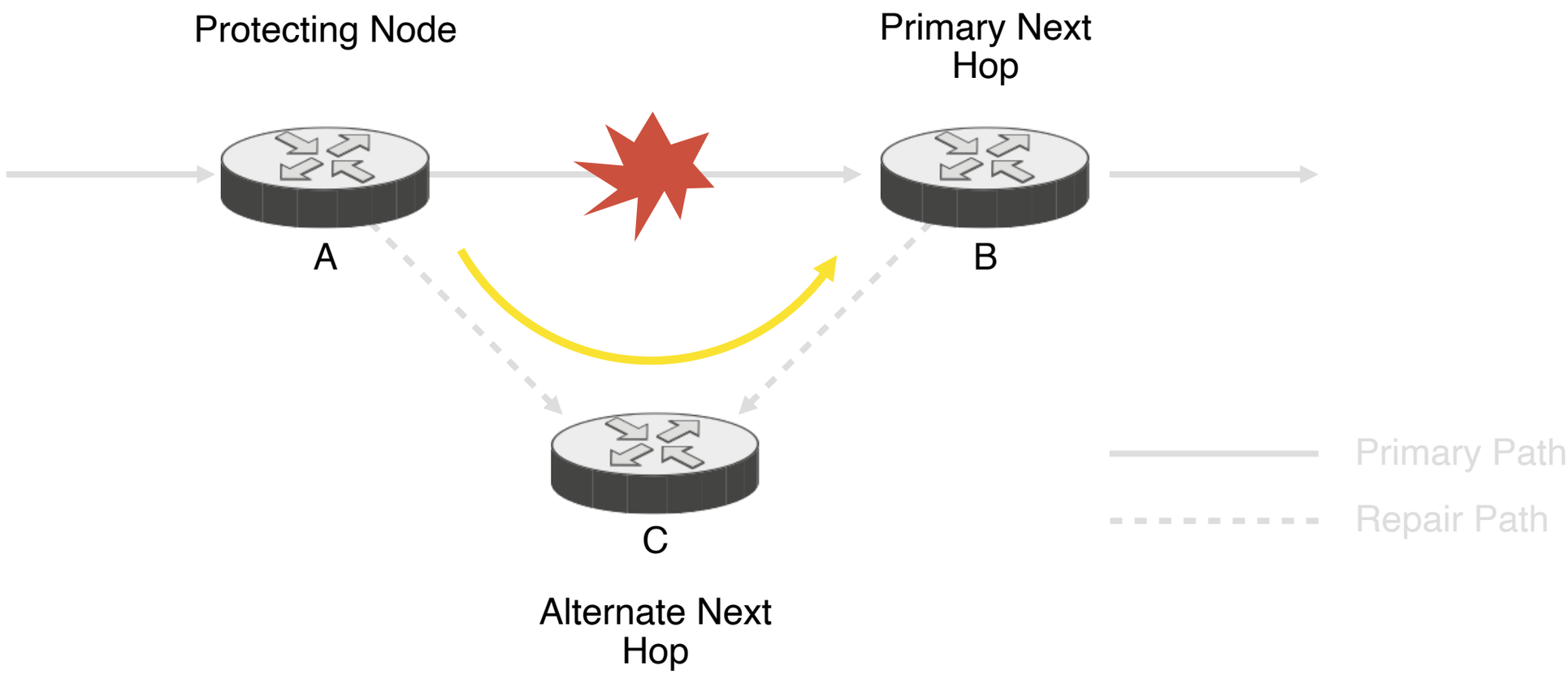
https://t.me/learningnets

# Fast reroute with OSPF

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

https://t.me/learningnets

# Convergence & Micro-loop




https://t.me/learningnets

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

https://t.me/learningnets

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

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

https://t.me/learningnets

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

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

https://t.me/learningnets

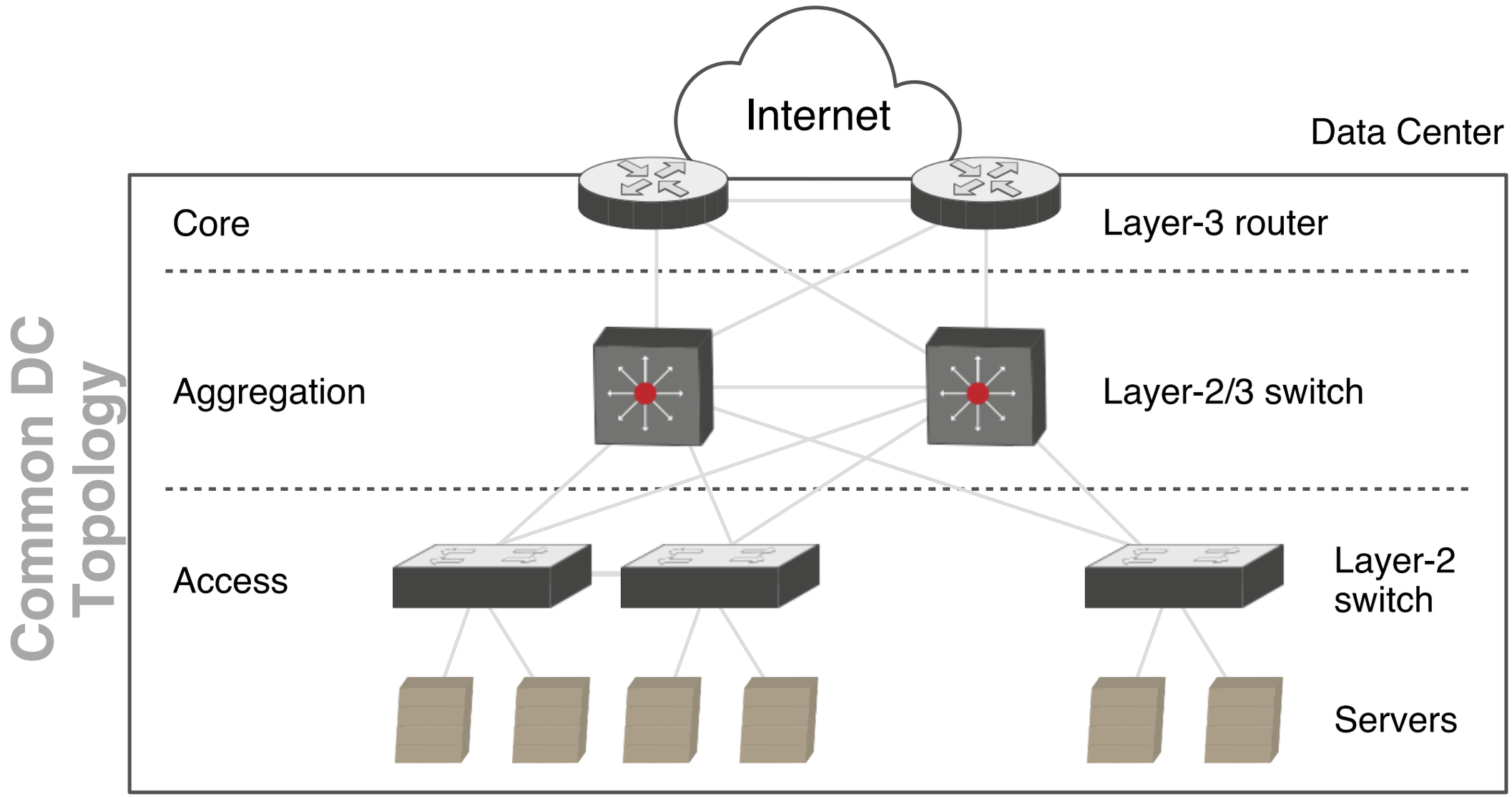
# Overlay Technologies and OSPF (GRE, MGRE, DMVPN, GETVPN, LISP)

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

https://t.me/learningnets

# OSPF in the Datacenter

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

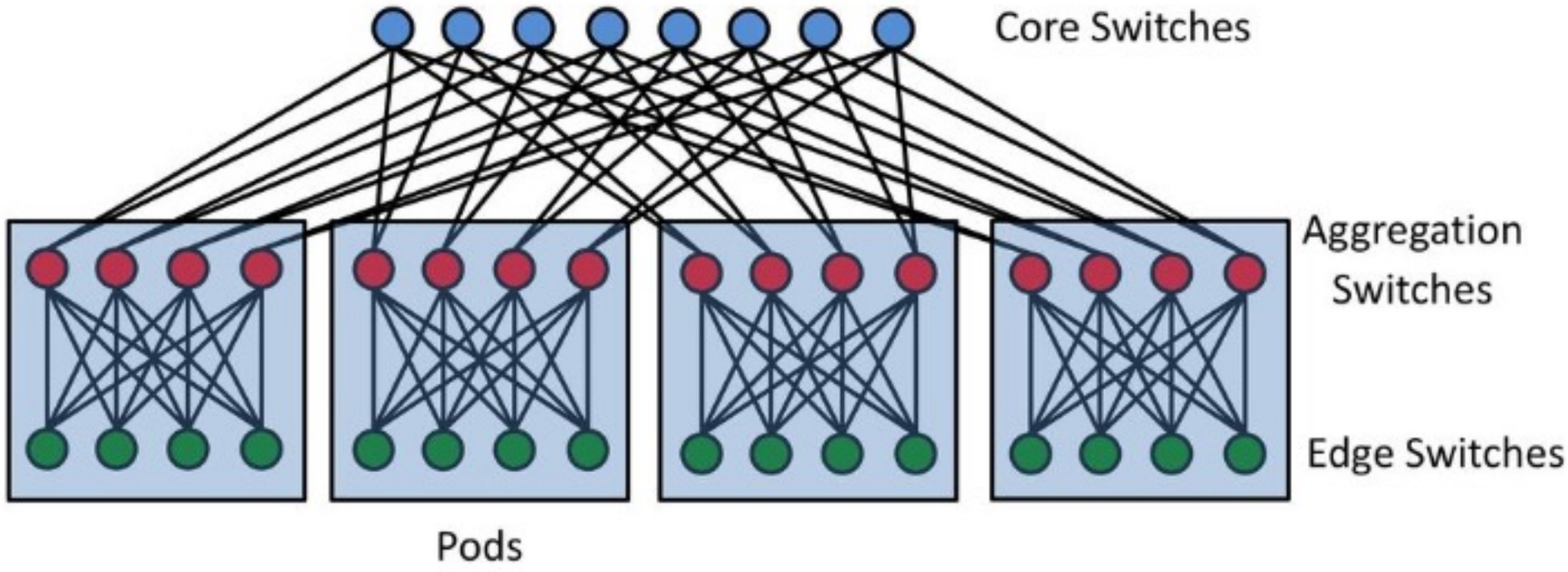



https://t.me/learningnets

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

https://t.me/learningnets

# 3 stage CLOS topology




https://t.me/learningnets

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

https://t.me/learningnets

- The Fabric SHOULD support 250k routes @ 5k fabric nodes with convergence time below 250ms.

- The Fabric SHOULD support 500k routes @ 7.5k fabric nodes with convergence time below 500ms.

- The Fabric SHOULD support 1M routes @ 10k fabric nodes with convergence time below 1s.

https://t.me/learningnets

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

https://t.me/learningnets

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

https://t.me/learningnets

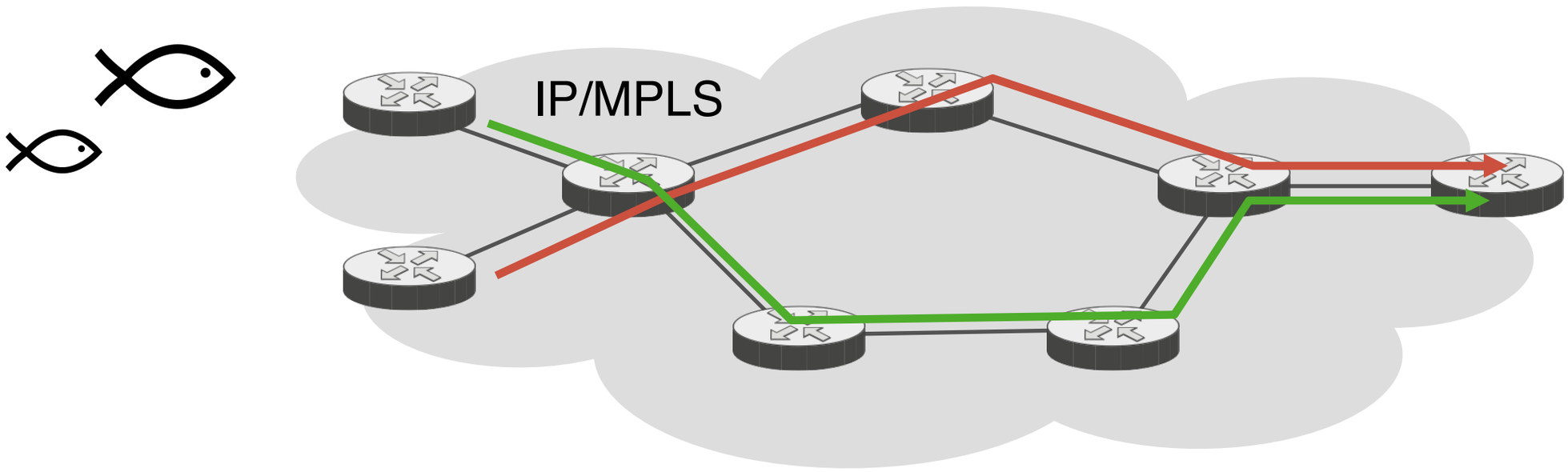
# OSPF in the Service Provider Networks

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

https://t.me/learningnets

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

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




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

https://t.me/learningnets

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

https://t.me/learningnets

# OSPF Design Best Practices

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

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

https://t.me/learningnets

# OSPF Design Best Practices

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

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

https://t.me/learningnets

# OSPF Design Best Practices

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

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

https://t.me/learningnets

# OSPF Design Best Practices

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

https://t.me/learningnets

# OSPF Design Best Practices

- If PE loopback mask is /24, OSPF advertises it as /32 but LDP assigns a label for /24, since there is a mismatch between two control plane protocols (LDP and OSPF), packet is dropped. Because OSPF advertises loopback interfaces as /32. They should follow each other.

  - Either OSPF network type should be point to point to advertise loopback as /32 so routing table and LDP is same, or use /32 loopback subnet mask.

https://t.me/learningnets

# OSPF Design Best Practices

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

https://t.me/learningnets

# OSPF Design Best Practices

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

https://t.me/learningnets

# OSPF Design Best Practices

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

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

https://t.me/learningnets

# OSPF Design Best Practices

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

https://t.me/learningnets

# OSPF Design Best Practices

- OSPF needs an IP address for adjacency, IS-IS doesn't require an IP Address for neighborship, remote attack to the IS-IS is hard if not impossible, thus IS-IS is seen more secure compare to OSPF.

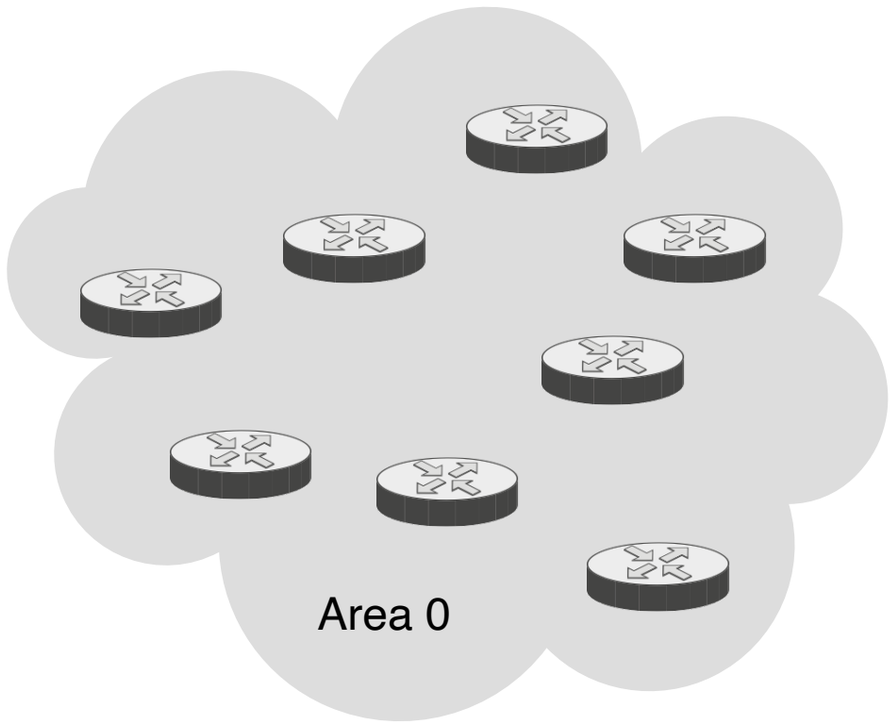
  - OSPF provides MPLS TE supports, similar to IS-IS, but distance vector protocols don't.

- OSPF is a good protocols for those who look Enterprise level and standard base protocol.

https://t.me/learningnets

# OSPF Frequently Asked Questions

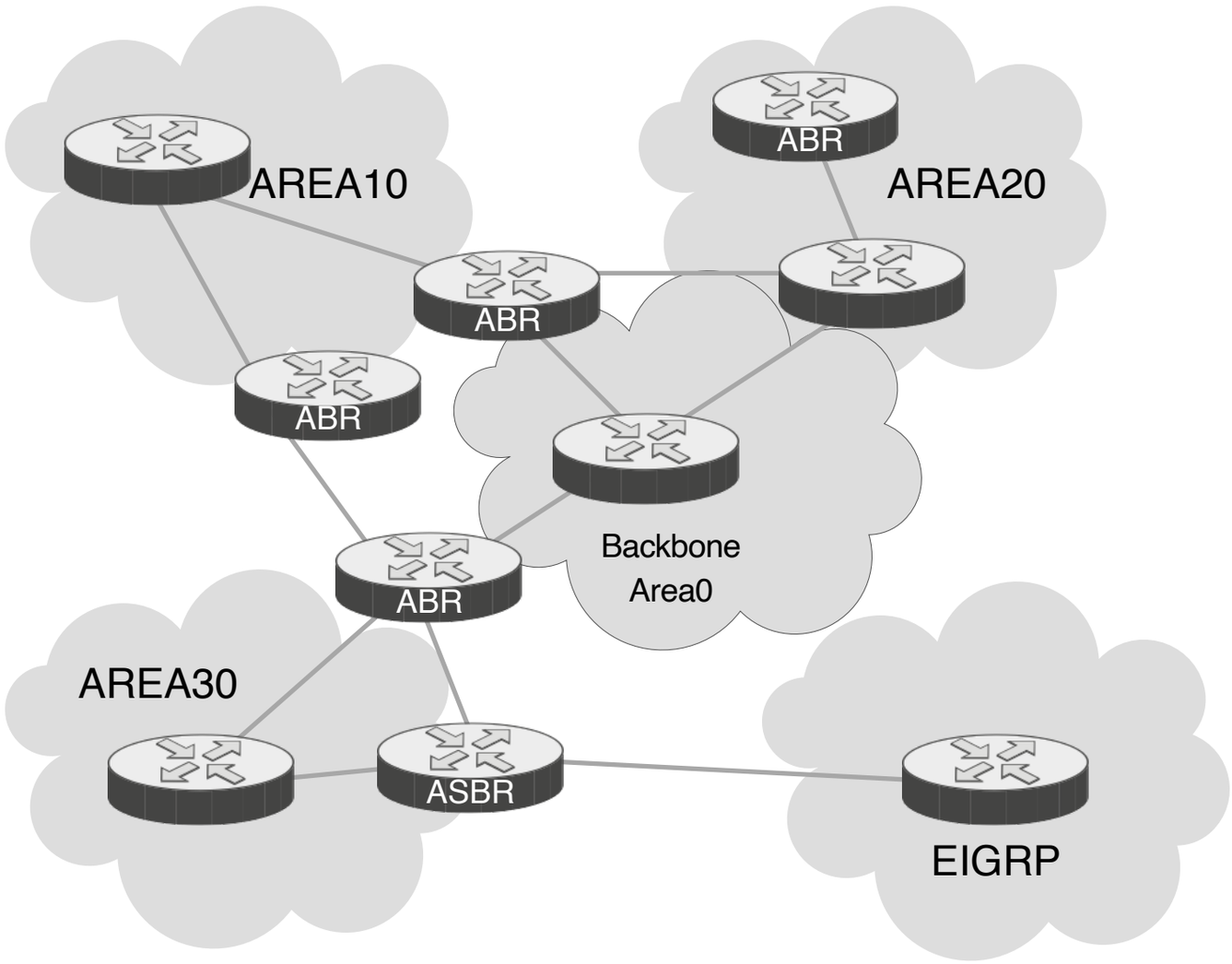
- How many routers should be in one OSPF Area ?




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

https://t.me/learningnets

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

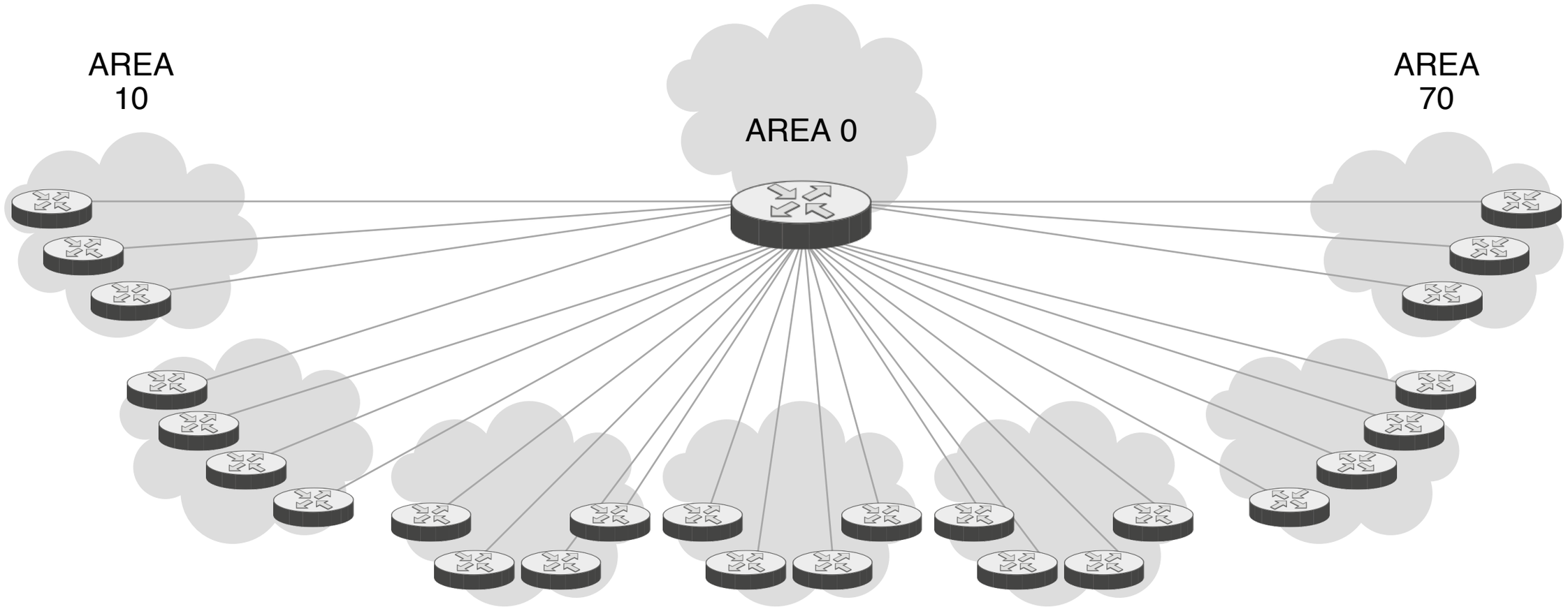



https://t.me/learningnets

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

https://t.me/learningnets

# How many OSPF area is suitable per OSPF ABR?




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

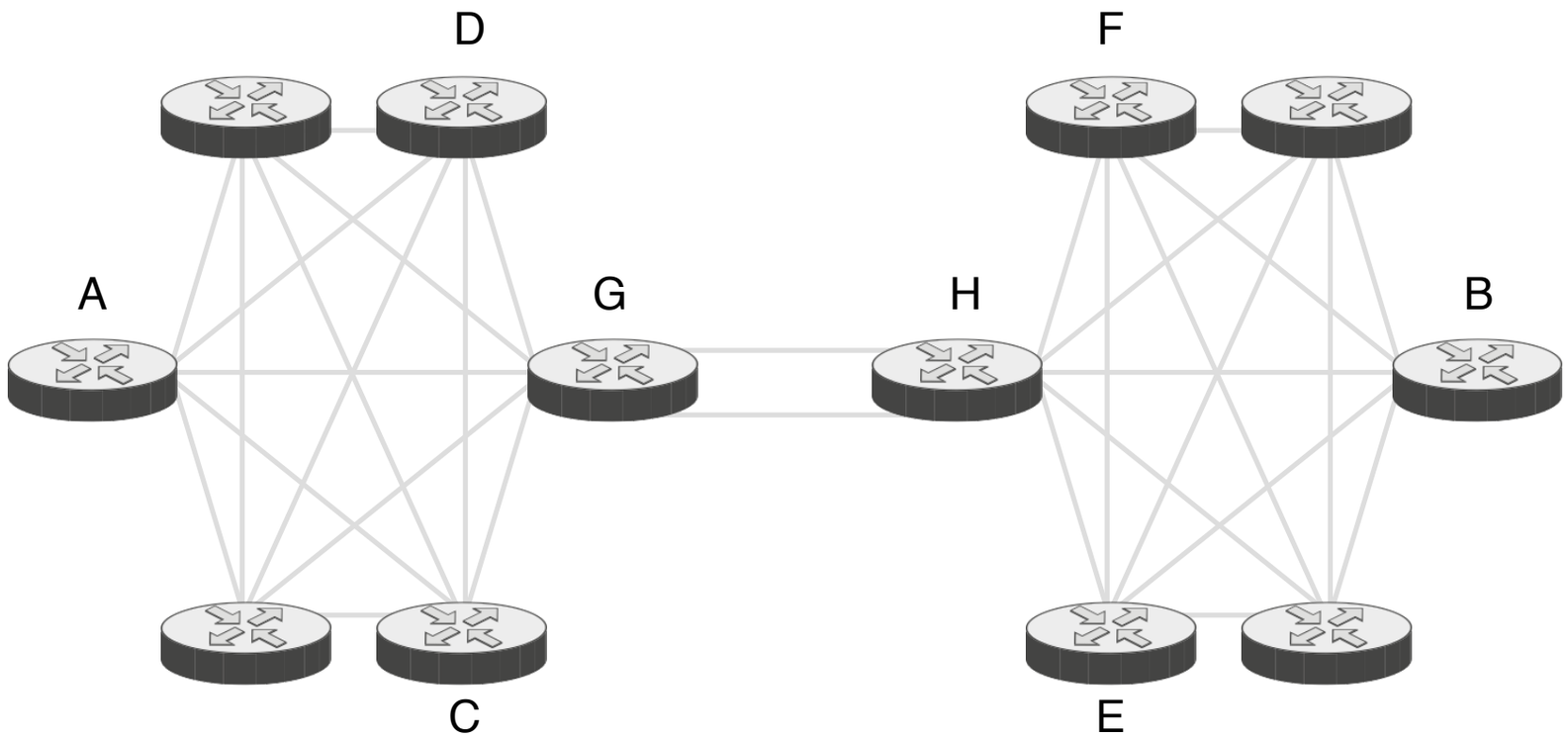
https://t.me/learningnets

# OSPF CASE STUDIES

https://t.me/learningnets

# ABR Placement

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




https://t.me/learningnets

# ABR Placement

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

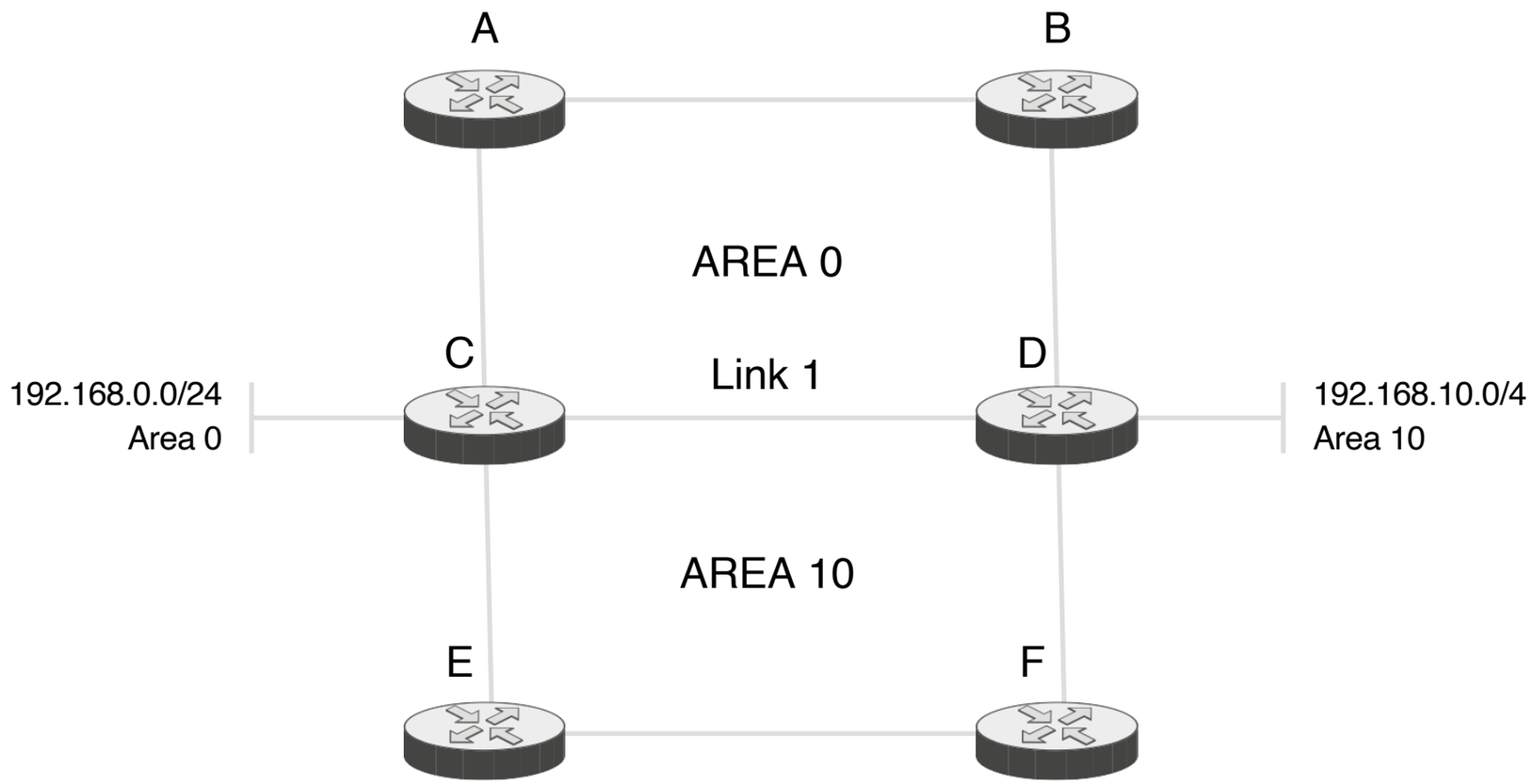
    - If we use Router G or Router H as an ABR, we will have only 32 paths max  (5x6) +2 between Router A and B, this will greatly reduce the load on the resources, reduces the overall complexity thus makes troubleshooting easier.

- Put ABR always a place where you can separate the complex topologies.

https://t.me/learningnets

# Multi-Area OSPF Adjacency

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




https://t.me/learningnets

# Multi-Area OSPF Adjacency

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

https://t.me/learningnets

# Multi-Area OSPF Adjacency

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

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

https://t.me/learningnets

# Multi-Area OSPF Adjacency

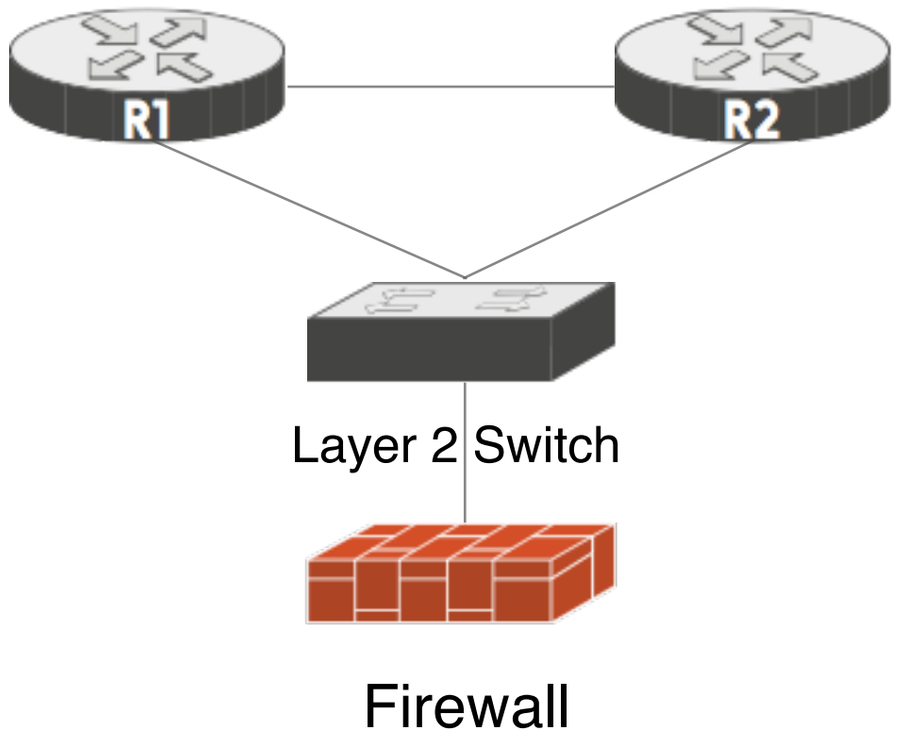
- Over 1 OSPF adjacency multiple area can be allowed with the RFC 5185.
    - Below is a sample configuration from the Cisco device which supports RFC 5185.

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

https://t.me/learningnets

# NSSA at the Internet Edge

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




https://t.me/learningnets

# NSSA at the Internet Edge

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

https://t.me/learningnets

# NSSA at the Internet Edge

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

https://t.me/learningnets

# OSPF and BGP Interaction

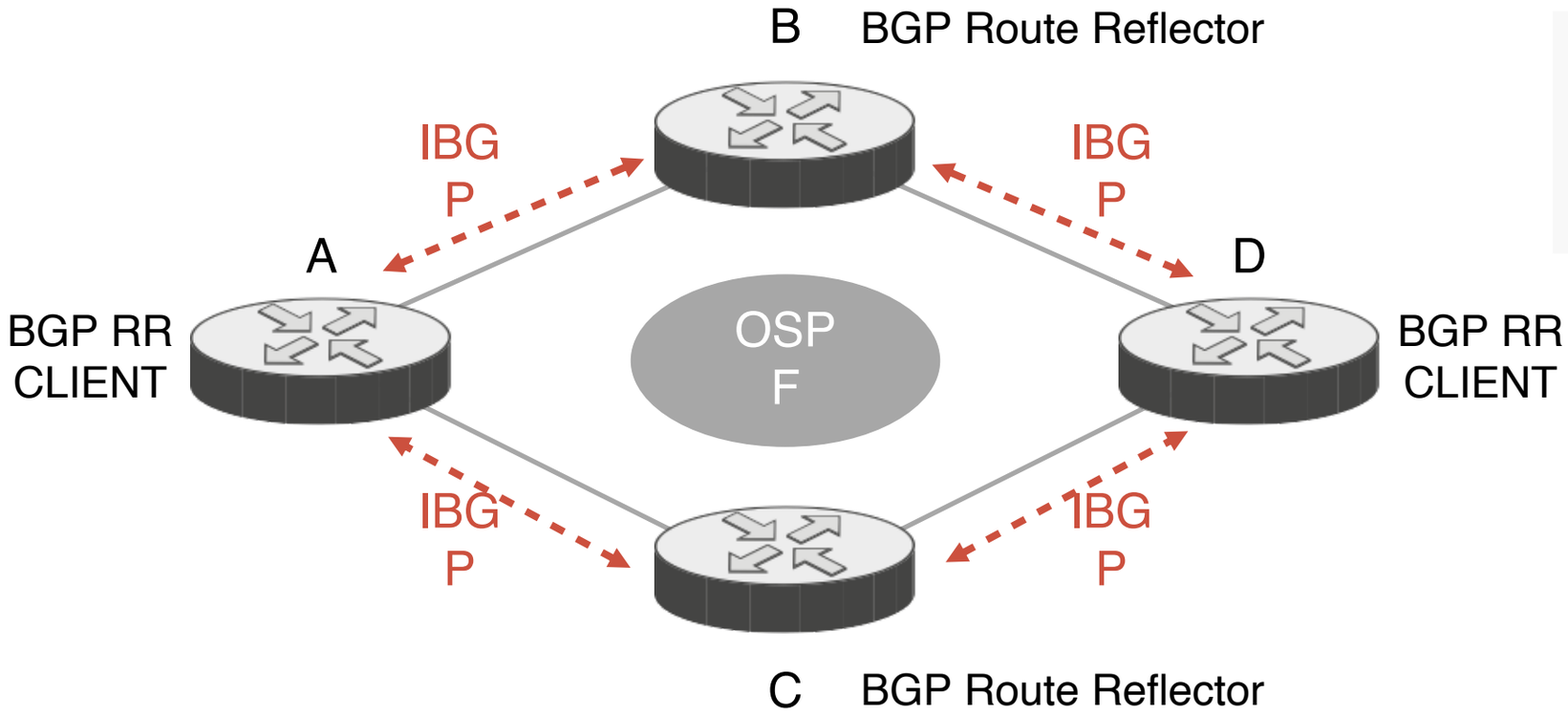
- OSPF is running as an IGP protocol in the below network. Also there is no MPLS in the core and all routers run BGP.

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

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

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

What would be your design recommendation ?




https://t.me/learningnets

# OSPF and BGP Interaction

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

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

https://t.me/learningnets

# OSPF and BGP Interaction

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

https://t.me/learningnets

# OSPF and BGP Interaction

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

https://t.me/learningnets

# Case Study Key Point

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

https://t.me/learningnets

# OSPF in the CCDE Exam

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

https://t.me/learningnets

# Summary

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

https://t.me/learningnets

# OSPF

## Open Shortest Path First

QUIZ

https://t.me/learningnets

## Question 1

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

A. 50

B. 100

C. 250

D. Less than 50

E. It depends

https://t.me/learningnets

## Answer 1

E. It depends

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

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

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

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

https://t.me/learningnets

## Question 2

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

A. Provides Scalability

B. Allow Multi-Area OSPF design

C. Provides fast convergence

D. Provides High Availability

E. Better Traffic Engineering

https://t.me/learningnets

# Answer 2

A.Provides Scalability

B.Allow Multi-Area OSPF design

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

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

https://t.me/learningnets

## Question 3

What does topology information mean in OSPF?

A. IP addresses of the directly connected interface.

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

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

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

https://t.me/learningnets

## Answer 3

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

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

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

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

https://t.me/learningnets

## Question 4

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

A. They are used for high availability.

B. They are used for easier troubleshooting.

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

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

https://t.me/learningnets

## Answer 4

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

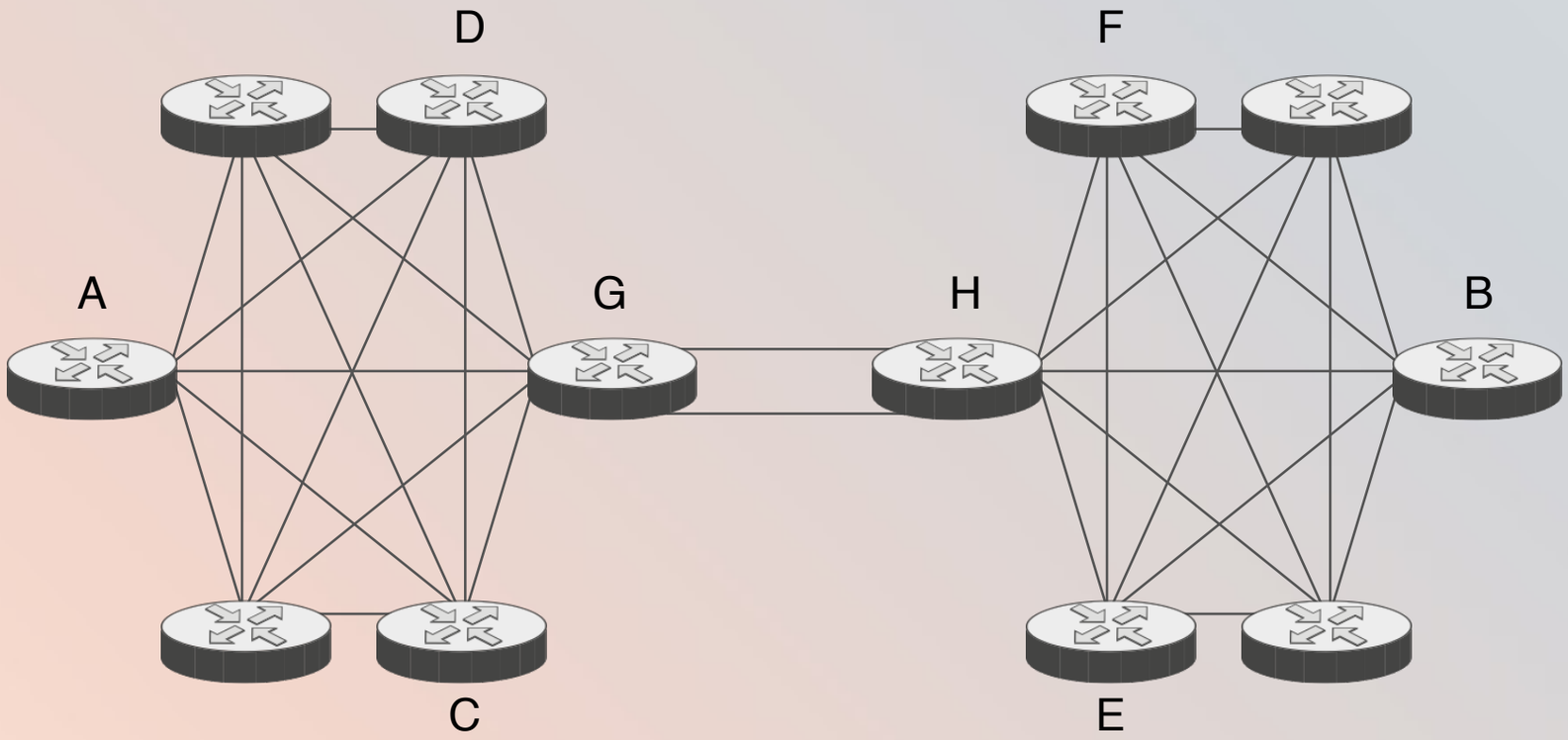
OSPF areas are used mainly for scalability. Having smaller domain means, keeping topology information in an area and not sending between the areas. More than one area doesn't provide high availability and doesn't make troubleshooting easier.

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

https://t.me/learningnets

# Question 5

Which router in the below topology should be an ABR?




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

https://t.me/learningnets

# Answer 5

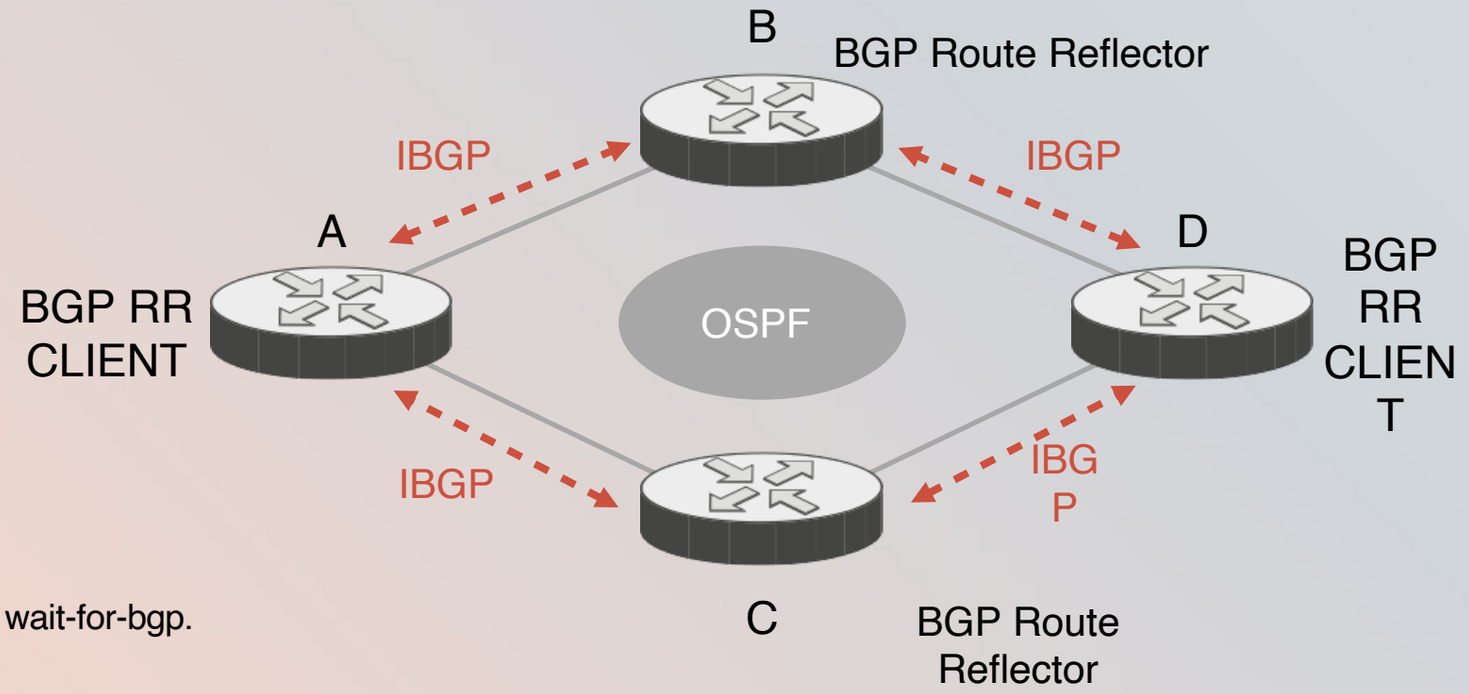
A. G or H

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

https://t.me/learningnets

# Question 6

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




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

B. OSPF prefix-list.

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

D. IGP LDP synchronization.

https://t.me/learningnets

# Answer 6

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

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

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

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

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

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

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

https://t.me/learningnets

## Question 7

How many levels in OSPF hierarchy used ?

A. One

B. Two

C. Three

D. As many as possible

https://t.me/learningnets

## Answer 7

B. Two

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

https://t.me/learningnets

## Question 8

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

A. It slows down the convergence.

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

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

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

E. It prevents topology information between OSPF areas.

https://t.me/learningnets

# Answer 8

A.It slows down the convergence.

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

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

E.It prevents topology information between OSPF areas.

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

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

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

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

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

https://t.me/learningnets

## Question 9

Why Designated Router is used in OSPF network?

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

B. It is used to create topology information

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

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

https://t.me/learningnets

# Answer 9

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

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

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

https://t.me/learningnets

## Question 10

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

A. OSPF Fast Reroute.

B. OSPF Multi Area Design.

C. IGP LDP Synchronization.

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

https://t.me/learningnets

## Answer 10

C. IGP LDP Synchronization.

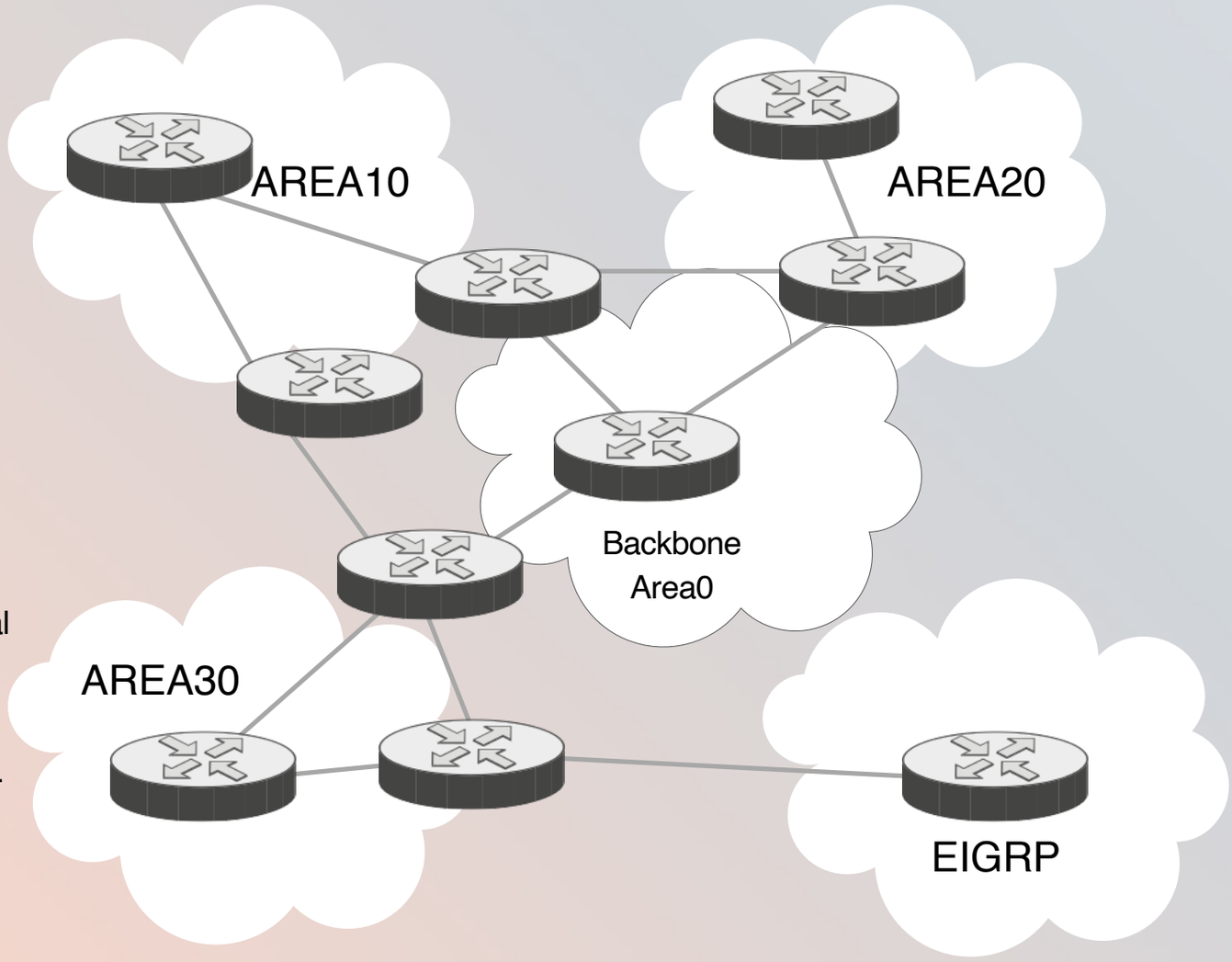
The problem occurs when link or node fails when OSPF and LDP is used together. It also occurs when IS-IS and LDP is together and the IG-LDP synchronization provides a label for the IGP pre xes in the Label database, otherwise since IGP converge rst and then LDP, packets would be blackholed.

Chicken and egg problem is solved and blackholing is avoided.

https://t.me/learningnets

## Question 11

Which below option is correct for the given topology?




A. Area 20 has to be Stub area.

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

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

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

https://t.me/learningnets

## Answer 11

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

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

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

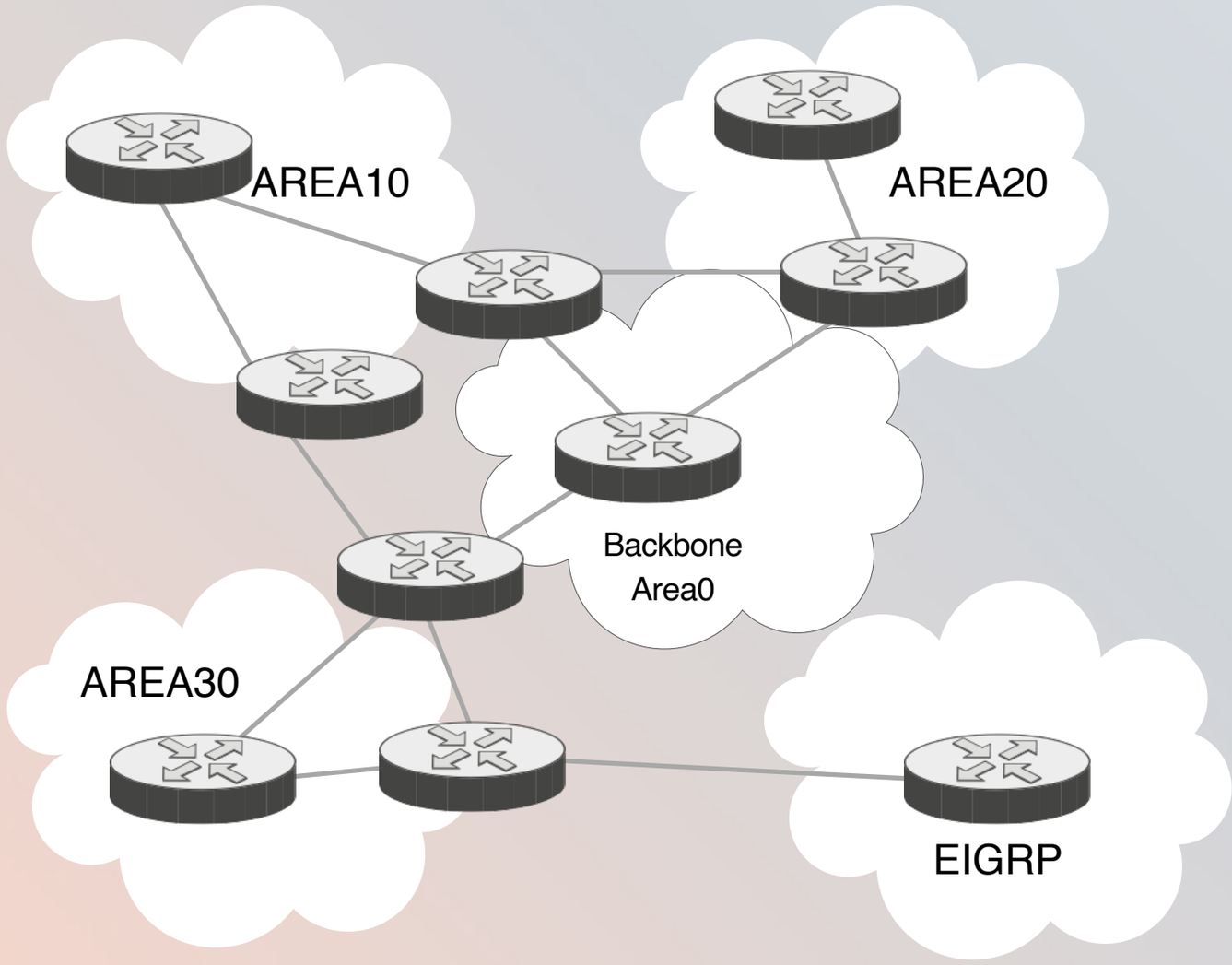
ABR of Area 20 doesn't have to DR. In fact, DR and ABR shouldn't be the same router. Since both operations are resource intensive and separating these two ask is a best practice.

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

https://t.me/learningnets

## Question 12

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




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

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

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

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

https://t.me/learningnets

## Answer 12

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

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

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

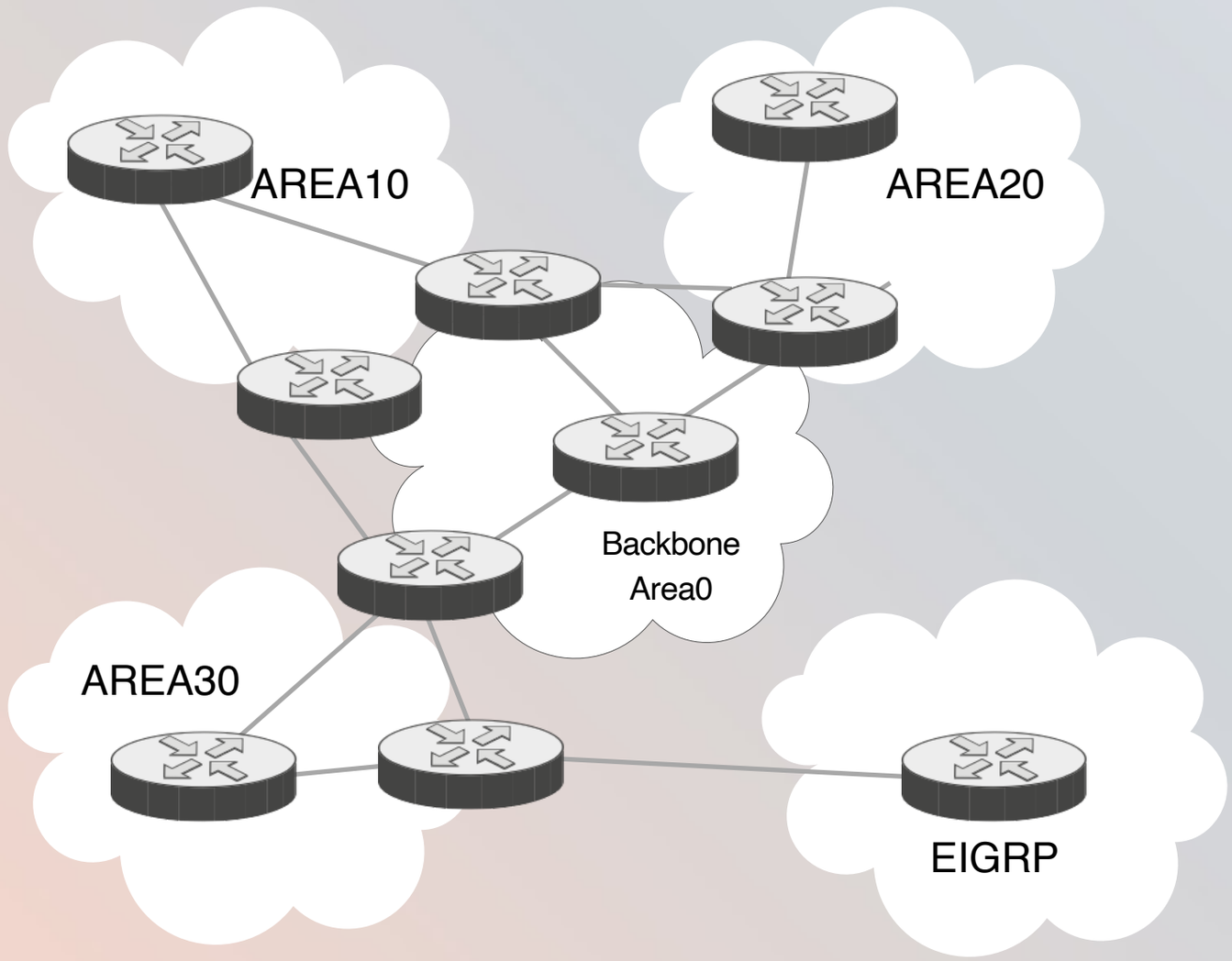
EIGRP preFIxes will be allowed and they will be seen as Type 7 LSA in the Area 30.

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

https://t.me/learningnets

## Question 13

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




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

B. Area 10 will not have EIGRP prefixes.

C. Area 10 cannot reach to EIGRP prefixes.

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

https://t.me/learningnets

# Answer 13

B. Area 10 will not have EIGRP prefixes.

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

https://t.me/learningnets

## Question 14

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

A. Full Mesh

B. Partial Mesh

C. Hub and Spoke

D. Ring

https://t.me/learningnets

## Answer 14

C. Hub and Spoke

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

This brings scalability into OSPF.

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

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

https://t.me/learningnets

## Question 15

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

A. To carry the customer prefixes.

B. Reachability between the MPLS VPN endpoints.

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

D. LDP requires OSPF as an IGP.

https://t.me/learningnets

## Answer 15

B. Reachability between the MPLS VPN endpoints.

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

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

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

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

https://t.me/learningnets

## Question 16

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

A. OSPF max-metric

B. OSPF prefer-primary path

C. OSPF sham-link

D. Passive-interface

E. Virtual link

https://t.me/learningnets

## Answer 16

C OSPF sham-link

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

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

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

https://t.me/learningnets

## Question 17

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

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

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

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

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

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

https://t.me/learningnets

# Answer 17

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

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

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

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

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

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

Changing IGP metric can affect BGP best path.

https://t.me/learningnets

## Question 18

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

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

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

C. Future Multi Area design migration can be easier.

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

https://t.me/learningnets

# Answer 18

C Future Multi Area design migration can be easier.

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

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

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

https://t.me/learningnets

## Question 19

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

A. Type 1

B. Type 2

C. Type 3

D. Type 4

E. Type 5

https://t.me/learningnets

# Answer 19

C.Type 3

D.Type 4

E.Type 5

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

https://t.me/learningnets

## Question 20

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

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

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

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

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

https://t.me/learningnets

# Answer 20

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

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

https://t.me/learningnets

## Question 21

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

A. One

B. Two

C. Three

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

https://t.me/learningnets

## Answer 21

B. Two

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

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

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

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

https://t.me/learningnets

## Question 22

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

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

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

C. Increase the security of the routing domain.

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

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

https://t.me/learningnets

# Answer 22

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

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

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

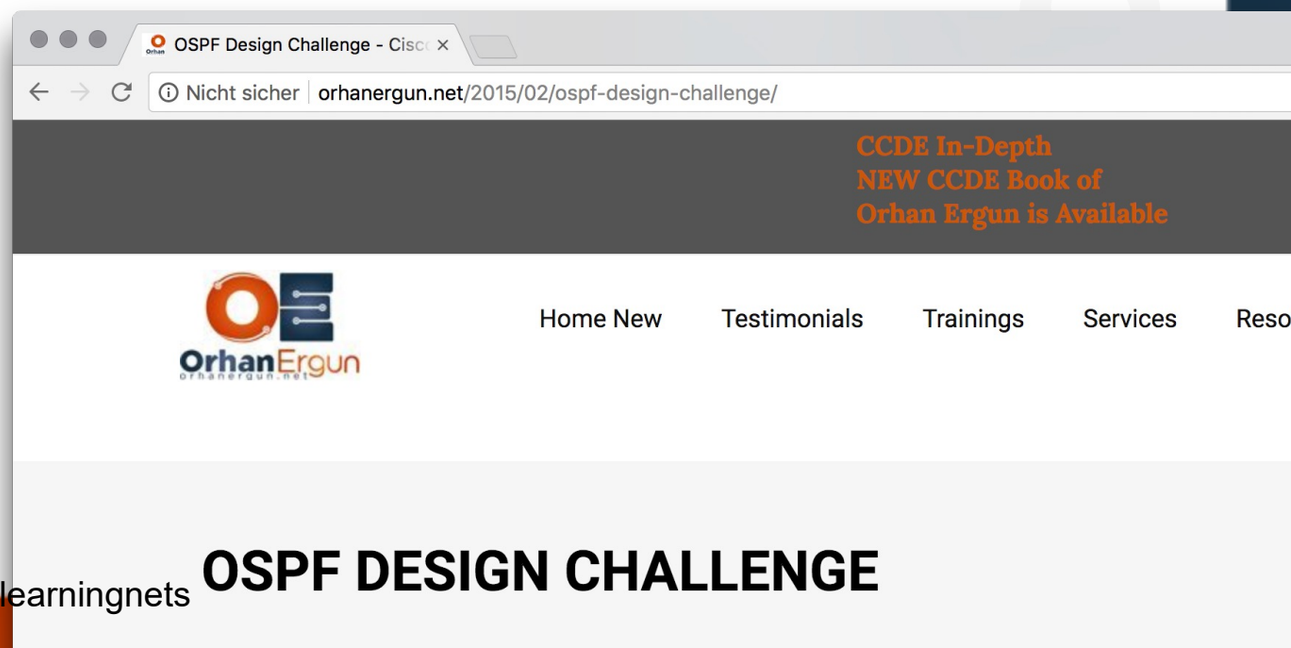
Availability and security doesn't increase with route summarization. But topology change affects is de nitely reduced.

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

https://t.me/learningnets

# Extra Study Resources

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




https://t.me/learningnets

# IS-IS

## Intermediate System to Intermediate System

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

# Theory

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

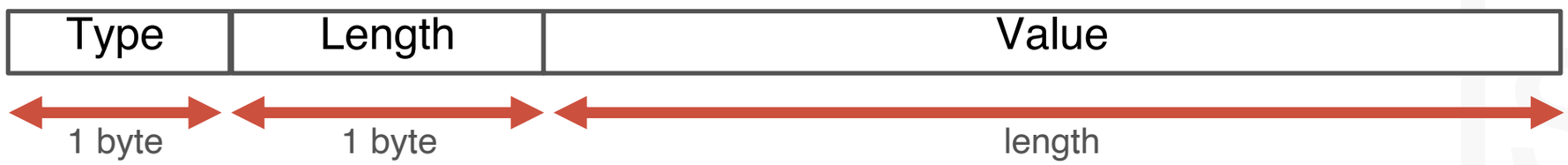
- Commonly used in Service Providers, Datacenter (as an underlay) and some large Enterprise networks.

https://t.me/learningnets

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

- IS-IS TLV Codes - Specified in RFC 1195

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



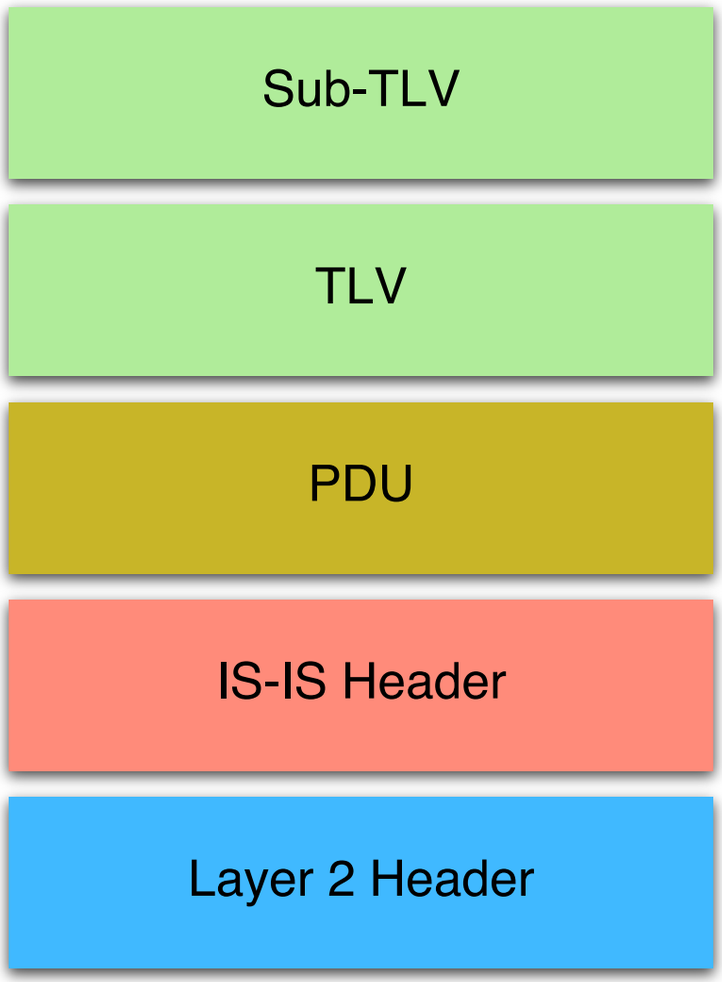

https://t.me/learningnets

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

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

https://t.me/learningnets

# TLVs are carried in LSP




https://t.me/learningnets

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

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

theory

https://t.me/learningnets

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

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

theory

https://t.me/learningnets

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

https://t.me/learningnets

# Example

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

https://t.me/learningnets

# IS-IS vs. OSPF Terminology

## IS-IS & OSPF Terminology

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

https://t.me/learningnets

## IS-IS & OSPF Terminology

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

https://t.me/learningnets

# IS-IS Fast Convergence

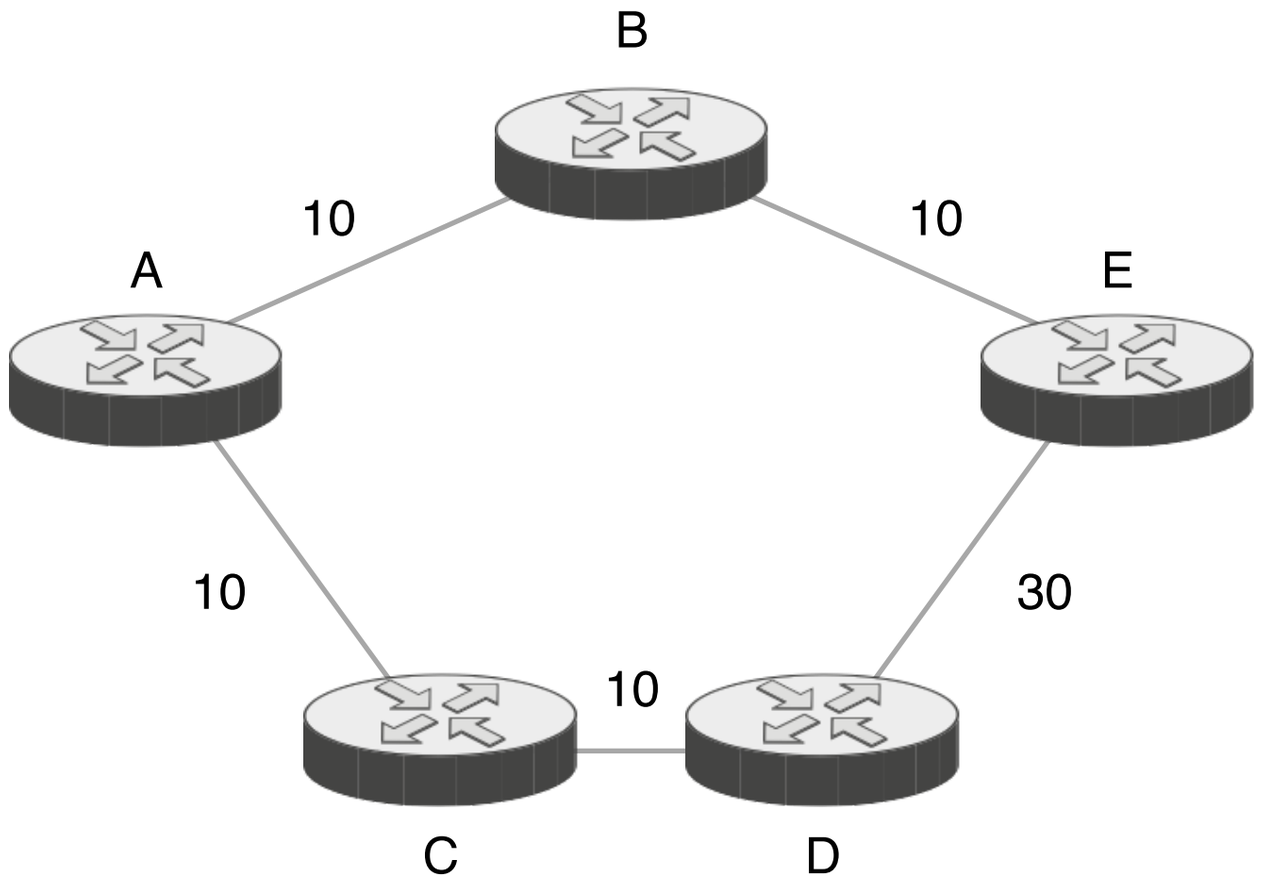
- Similar convergence characteristic with OSPF, SPF,LSA Throttling timers, LSA Pacing,  Processing and Propagation timers are tunable in IS-IS as well

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

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

https://t.me/learningnets

# Convergence and Micro-loop




https://t.me/learningnets

# Fast Reroute with IS-IS

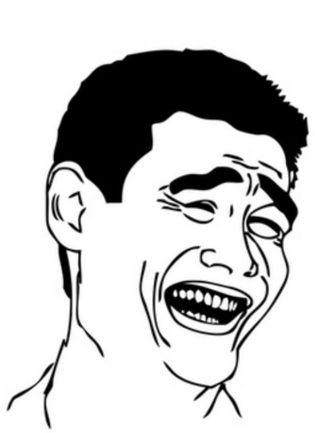
- IS-IS similar to OSPF, supports MPLS Traffic Engineering Fast Reroute, IP Traffic Engineering (LFA, Remote LFA) and also supports Segment Routing Fast Reroute.

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

https://t.me/learningnets

# IS-IS Scalability

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

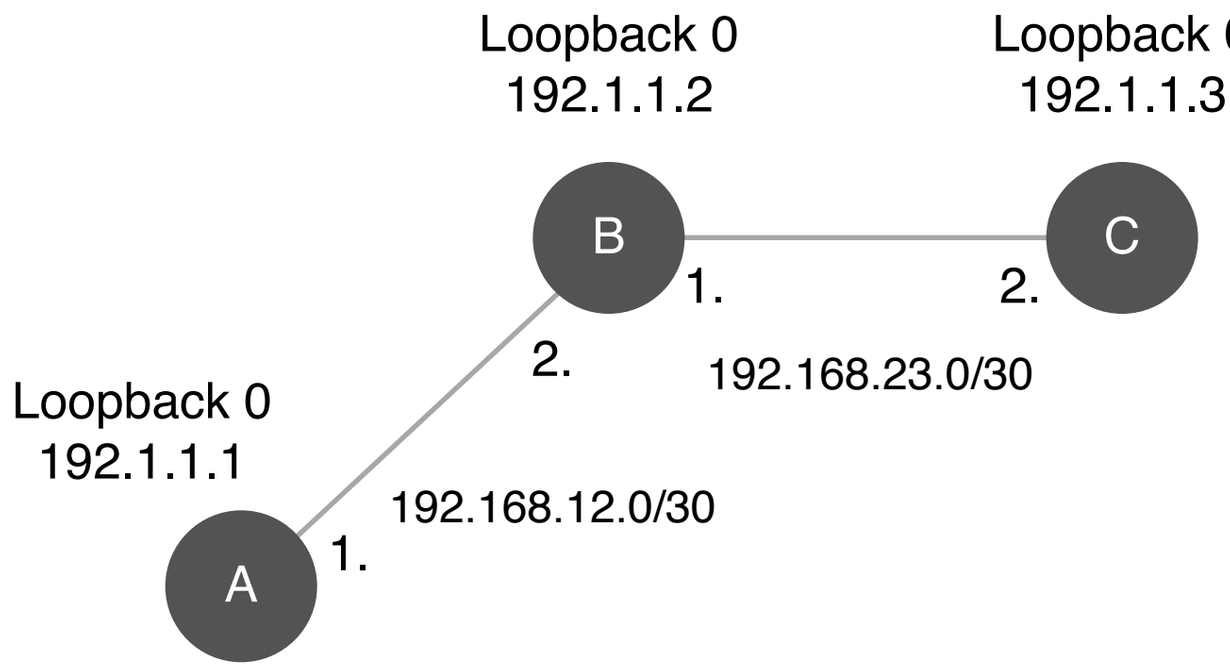


ONE LSP

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

https://t.me/learningnets

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




https://t.me/learningnets

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

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

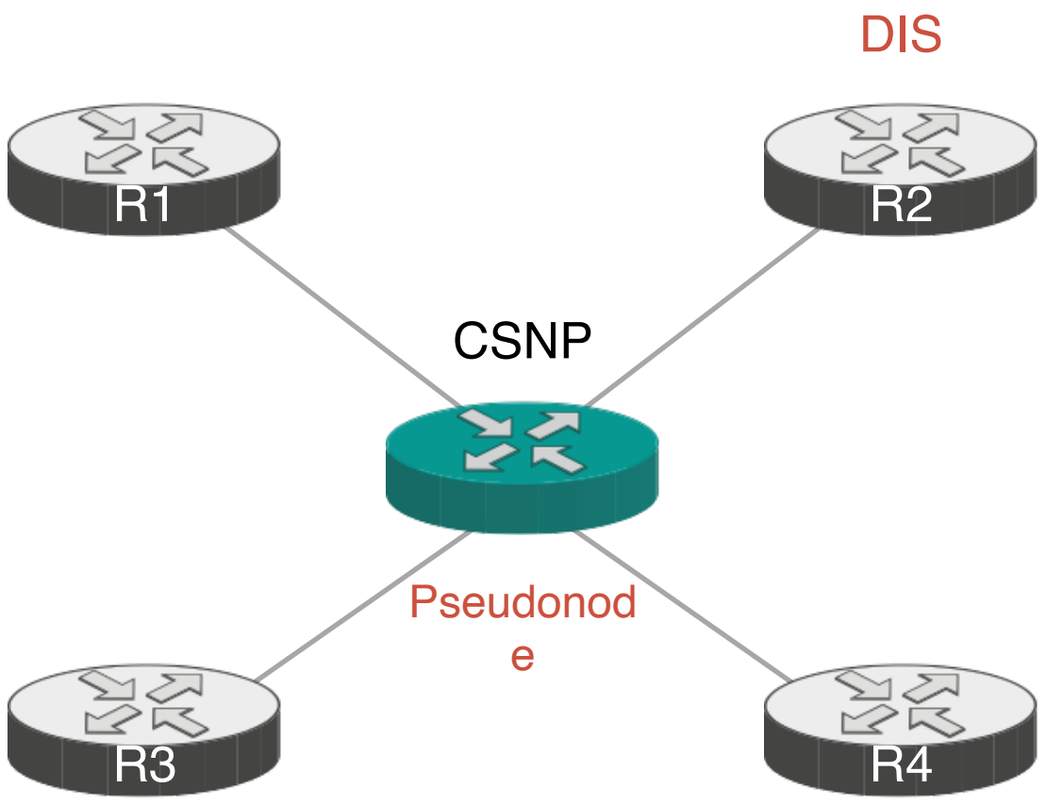
https://t.me/learningnets

# IS-IS Flooding in LAN

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

https://t.me/learningnets

# IS-IS Flooding in LAN




https://t.me/learningnets

# IS-IS Flooding in LAN

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

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

https://t.me/learningnets

# IS-IS Flooding in LAN

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

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

https://t.me/learningnets

# IS-IS Flooding in LAN

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

- DIS concept in Broadcast link provides scalability.

https://t.me/learningnets

# IS-IS Flooding in LAN

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

https://t.me/learningnets

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

https://t.me/learningnets

# Multi Level IS-IS Design

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

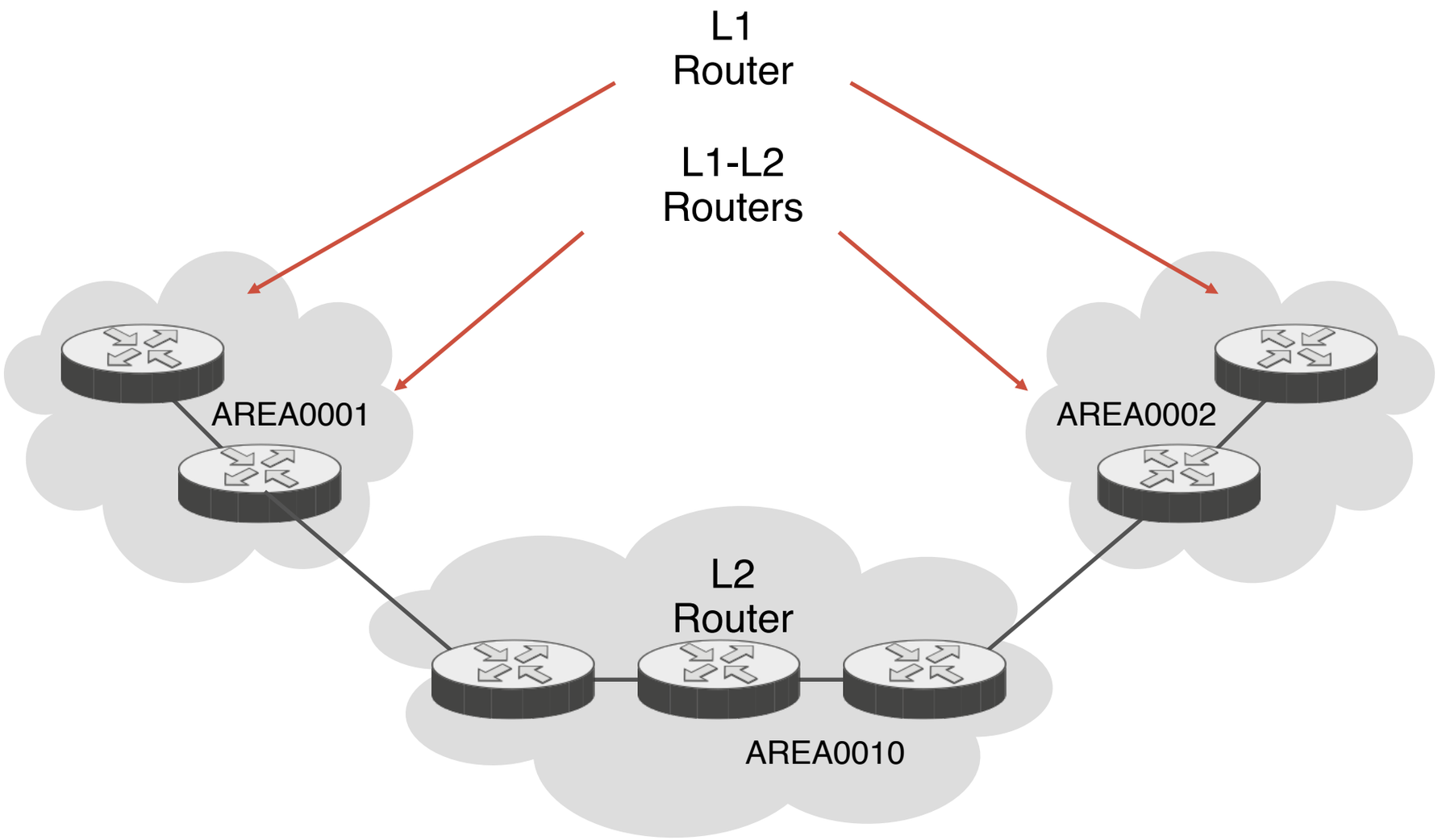
https://t.me/learningnets

# Multi Level IS-IS Design

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

https://t.me/learningnets

# IS-IS Router Types




https://t.me/learningnets

# IS-IS Router Types

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

https://t.me/learningnets

# IS-IS Router Types

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

https://t.me/learningnets

# IS-IS Router Types

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

https://t.me/learningnets

# ATT (Attached) Bit and Default Route

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

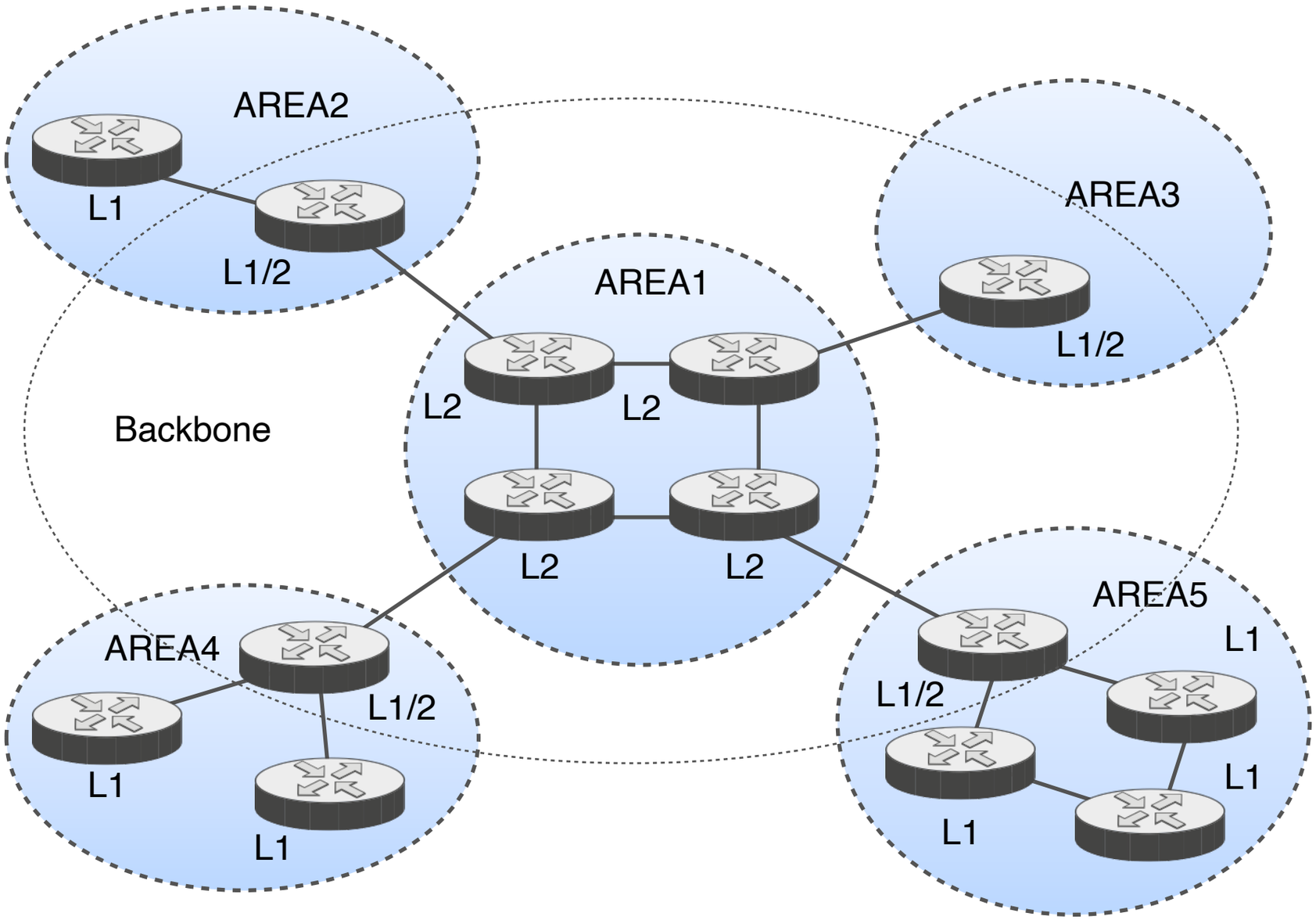
https://t.me/learningnets

# ATT (Attached) Bit and Default Route

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

https://t.me/learningnets

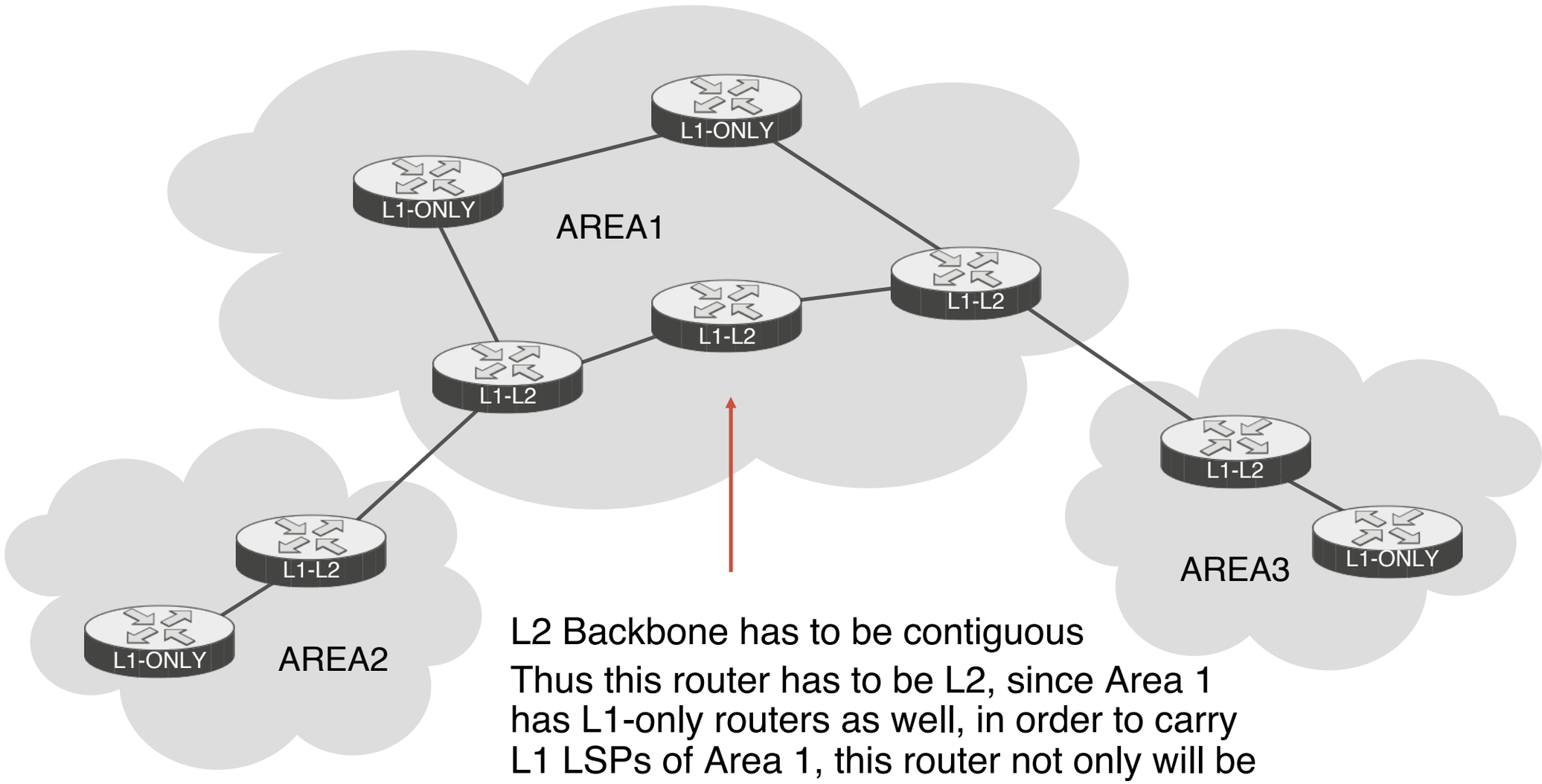
# ATT (Attached) Bit and Default Route




https://t.me/learningnets

# Hierarchy Levels

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




L2 Backbone has to be contiguous

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

https://t.me/learningnets

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

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

https://t.me/learningnets

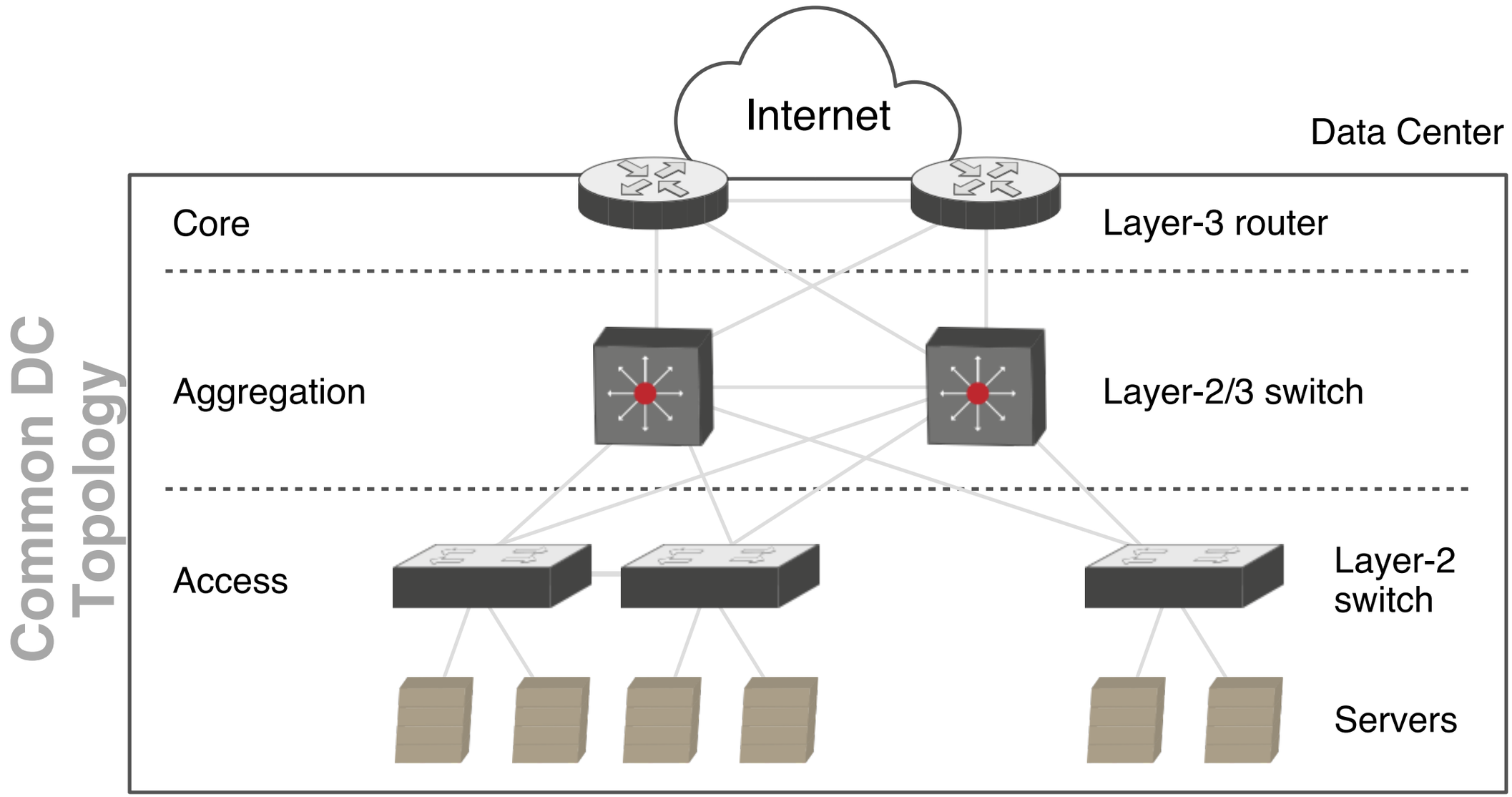
# Overlay Technologies and IS-IS (GRE, MGRE, DMVPN, GETVPN, LISP)

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

https://t.me/learningnets

# IS-IS in the Datacenter

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

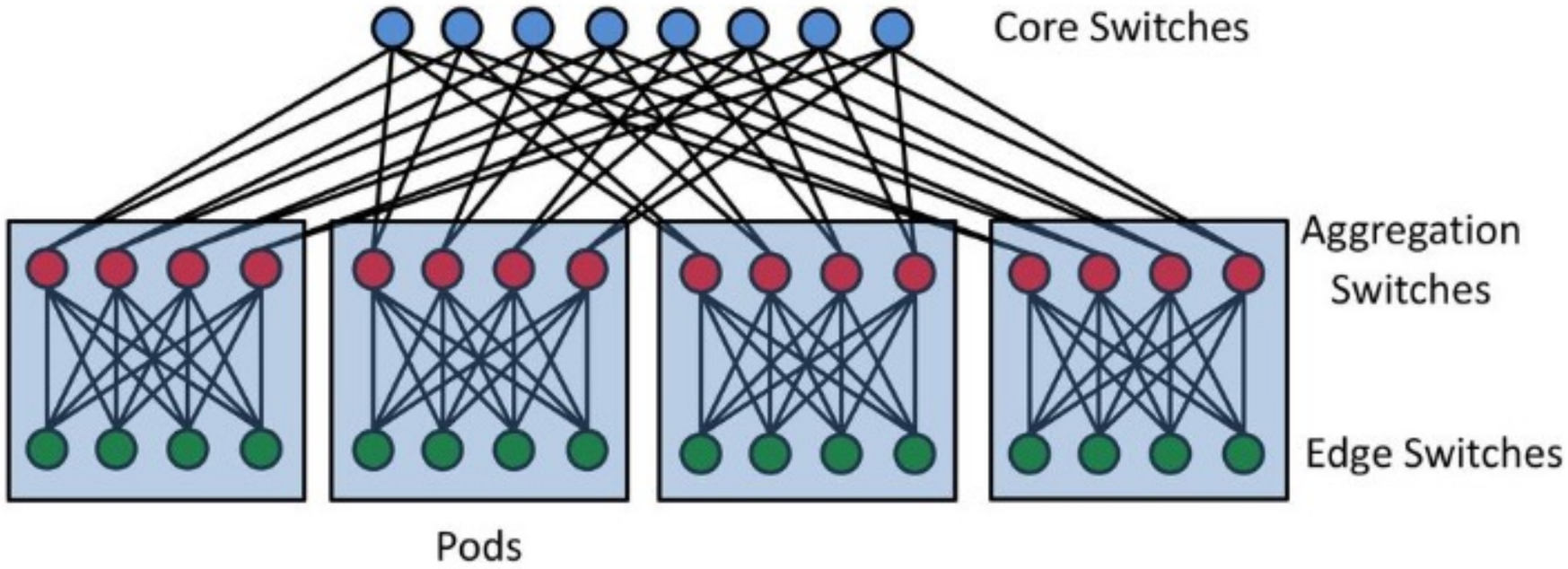



https://t.me/learningnets

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

https://t.me/learningnets

# 3 stage CLOS topology




https://t.me/learningnets

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

https://t.me/learningnets

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

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

https://t.me/learningnets

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

https://t.me/learningnets

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

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

https://t.me/learningnets

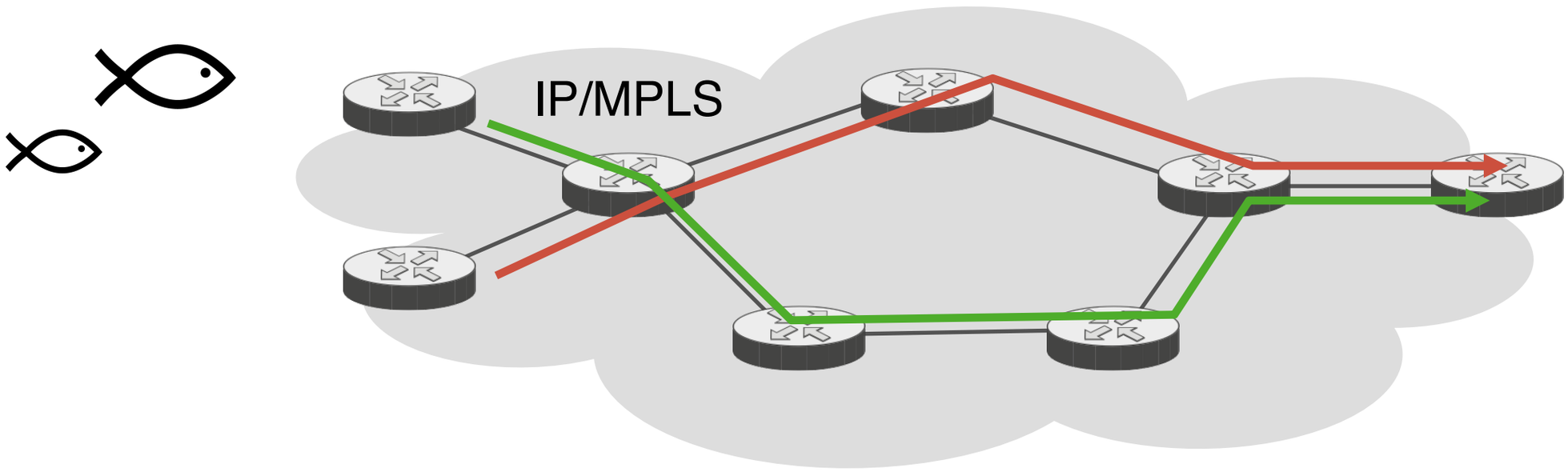
# IS-IS in the Service Provider Networks

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

https://t.me/learningnets

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

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




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

https://t.me/learningnets

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

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

https://t.me/learningnets

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

https://t.me/learningnets

# IS-IS Design Best Practices

- Unless there is a valid reason, don't deploy Multi Level IS-IS, keep the design simple, it provides better convergence, less configuration on the L1L2 nodes and optimal traffic flow.

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

best practices

https://t.me/learningnets

# IS-IS Design Best Practices

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

best practices

https://t.me/learningnets

# IS-IS Design Best Practices

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

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

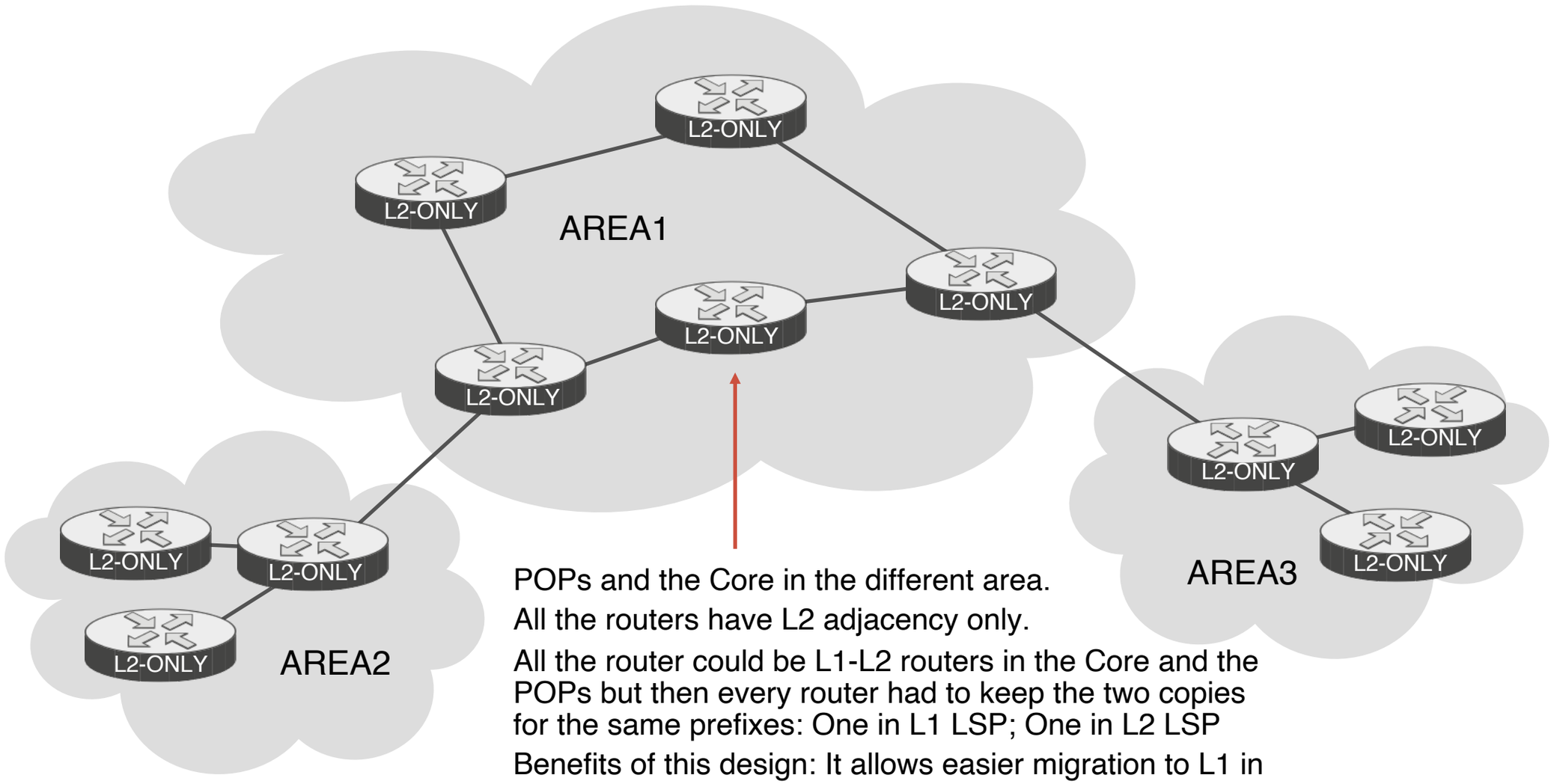
- If you start green field network design with L1-L2 then all the routers have to keep two databases for every prefix. So it is resource intensive without additional benefit

best practices

https://t.me/learningnets

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

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




POPs and the Core in the different area.

All the routers have L2 adjacency only.

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

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

https://t.me/learningnets

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

https://t.me/learningnets

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

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

https://t.me/learningnets

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

https://t.me/learningnets

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

https://t.me/learningnets

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

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

https://t.me/learningnets

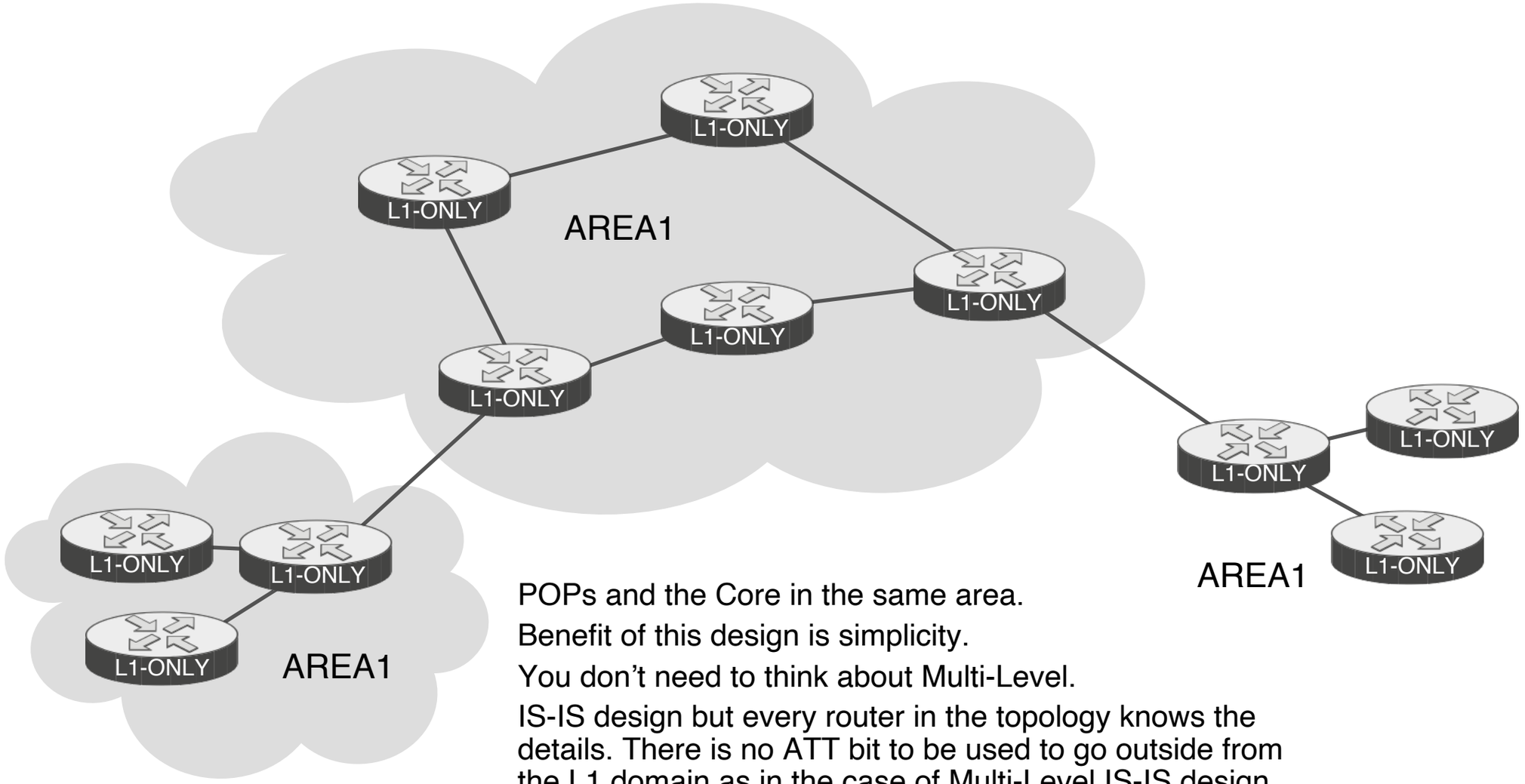
# Design Examples with IS-IS Areas and Levels

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

https://t.me/learningnets

# L1 in the POP and Core

- Area Design L1 in the POP and Core.




POPs and the Core in the same area.

Benefit of this design is simplicity.

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

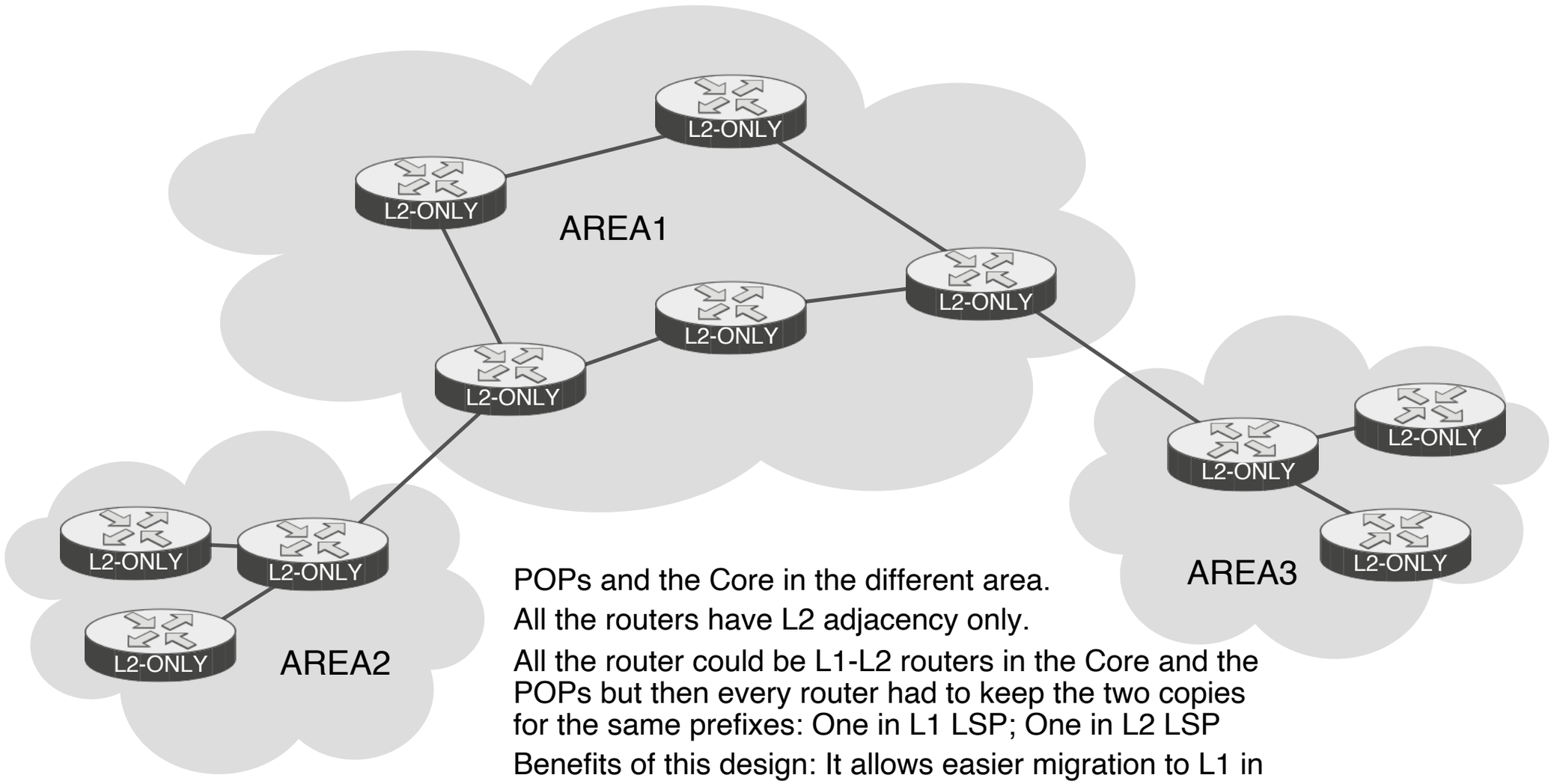
IS-IS design but every router in the topology knows the details. There is no ATT bit to be used to go outside from the L1 domain as in the case of Multi-Level IS-IS design.

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

https://t.me/learningnets

# L2 in the POP and Core

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




POPs and the Core in the different area.

All the routers have L2 adjacency only.

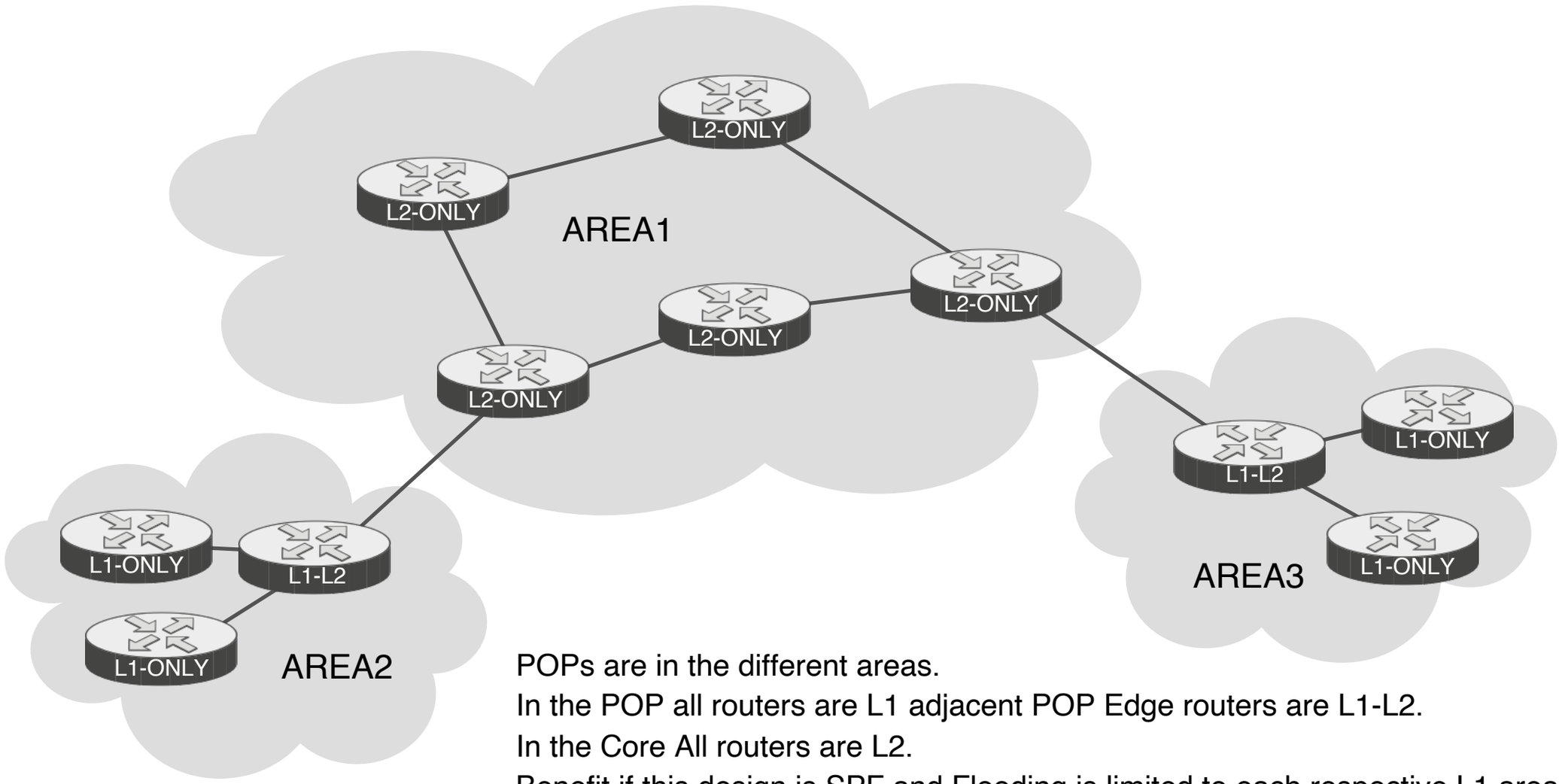
All the router could be L1-L2 routers in the Core and the POPs but then every router had to keep the two copies for the same prefixes: One in L1 LSP; One in L2 LSP

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

https://t.me/learningnets

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

- L1 Only POPs L2 Only Core.

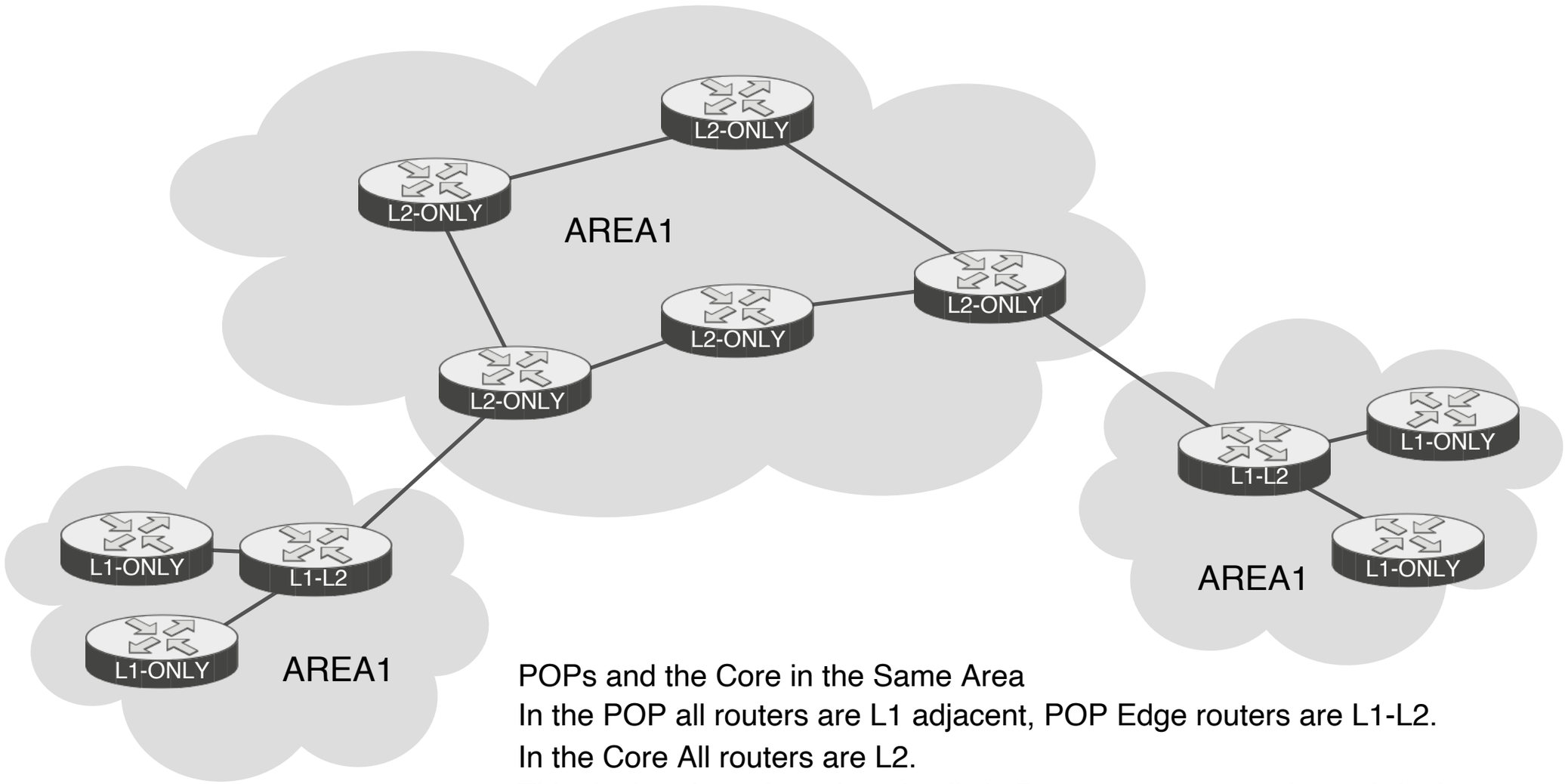



POPs are in the different areas.
In the POP all routers are L1 adjacent POP Edge routers are L1-L2.
In the Core All routers are L2.
Benefit if this design is SPF and Flooding is limited to each respective L1 areas.
L1 routers only know the ATT bit for outside of their areas.
If scalability is the business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

https://t.me/learningnets

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

- L1 Only POPs L2 Only Core.




POPs and the Core in the Same Area

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

In the Core All routers are L2.

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

https://t.me/learningnets

# IS-IS - BGP Interaction and Overload Bit

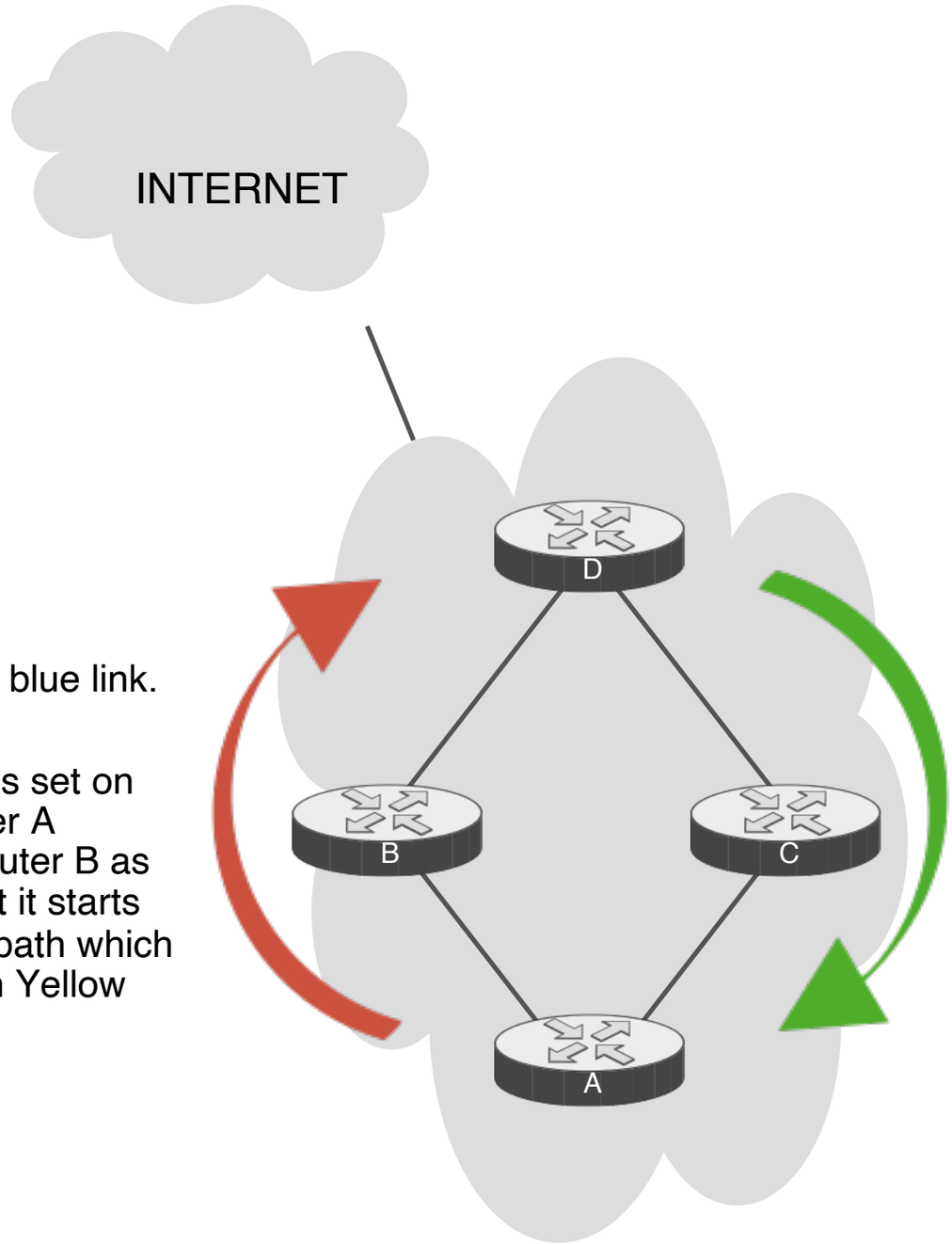
- Network design, especially large scale network design is all about managing interaction and understand the tradeoffs. IS-IS as an IGP provides an underlay infrastructure for BGP, MPLS overlays.

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

https://t.me/learningnets

# IS-IS - BGP Interaction

- Interaction with BGP




Primary path is blue link.

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

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

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

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

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

https://t.me/learningnets

# IS-IS and MPLS Interaction

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

https://t.me/learningnets

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

https://t.me/learningnets

## Fortunately there are at least three ways to fix it.

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

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

https://t.me/learningnets

## Fortunately there are at least three ways to fix it.

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

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

https://t.me/learningnets

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

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

https://t.me/learningnets

# Fast Service Restoration with IS-IS and LDP

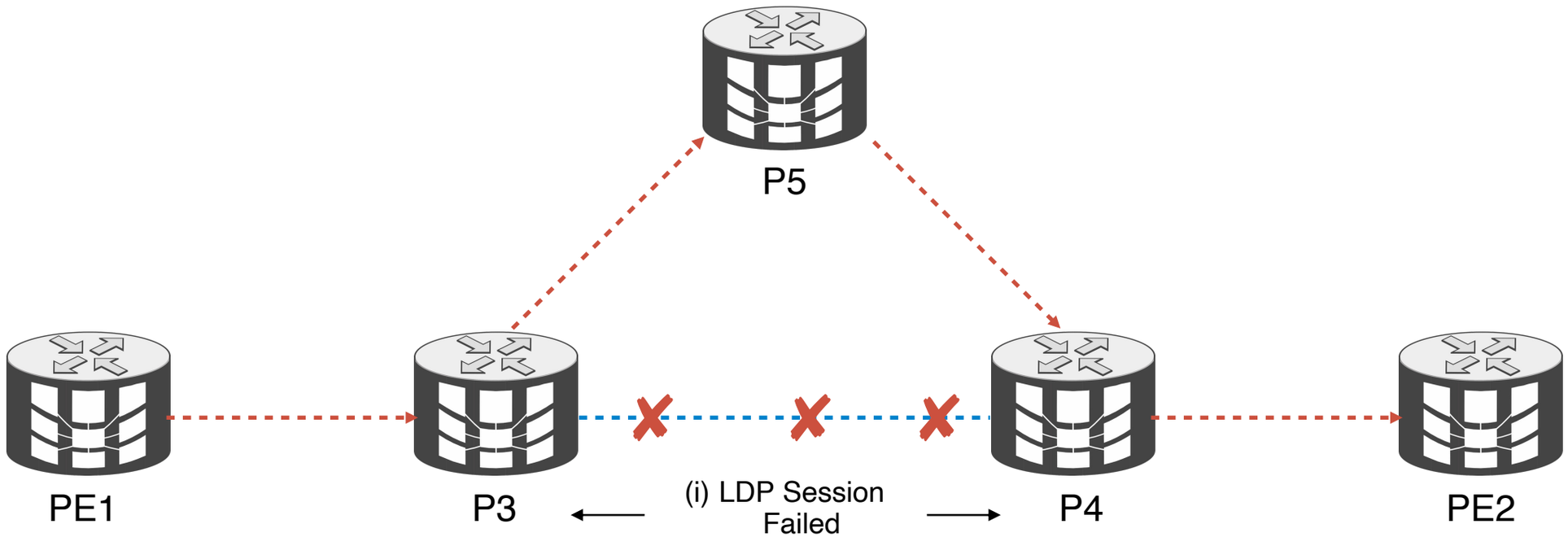
IGP - LDP Sync

- When there is link up event or LDP session down event, IGP advertises max/metric and indicates that traffic should follow the alternate links.

- This solution is called, cost-out procedure as well and very similar to IGP-BGP synchronization.

- Problem happens because LDP converge after IGP when link comes up, this creates black holing.

https://t.me/learningnets




(ii)IGP Assign Max metric for this link

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

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

https://t.me/learningnets

# Fast Service Restoration with IS-IS and LDP

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

rest
orati
on

https://t.me/learningnets

# Fast Service Restoration with IS-IS and LDP

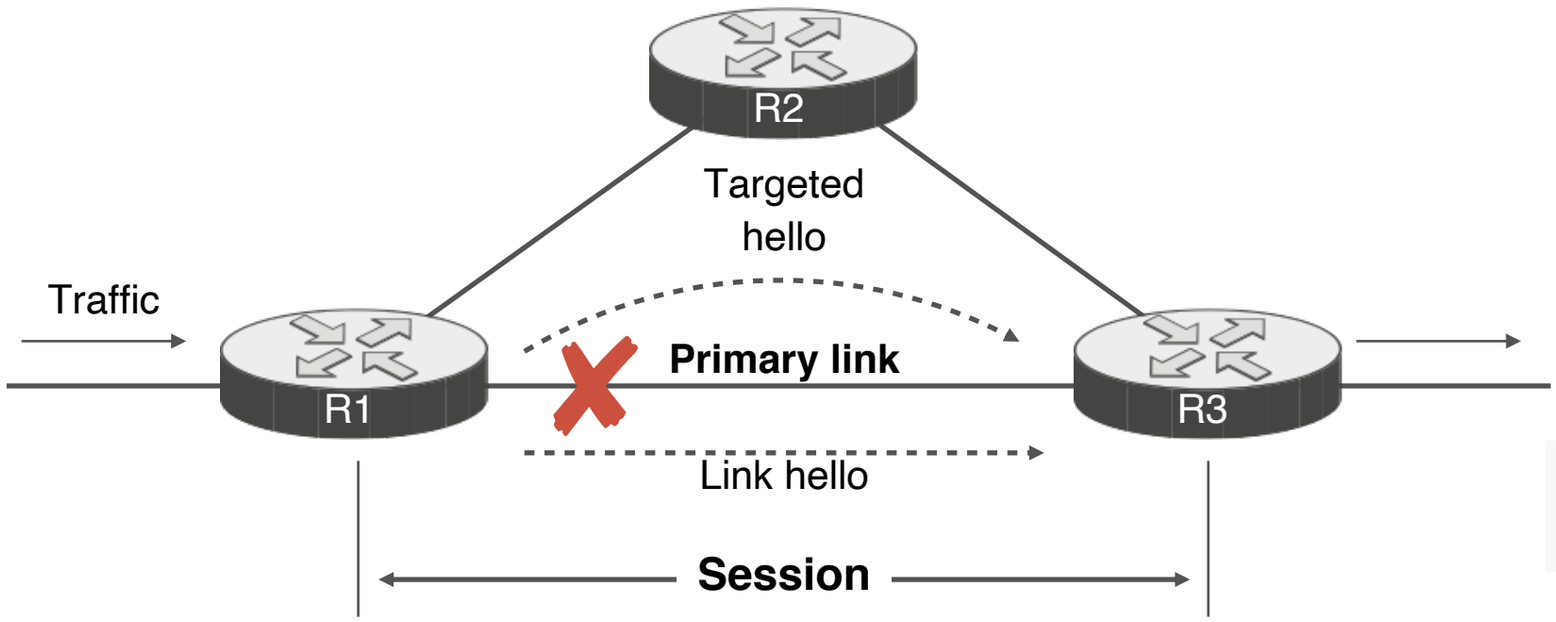
## LDP Session Protection

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

https://t.me/learningnets

# Fast Service Restoration with IS-IS and LDP

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




https://t.me/learningnets

# Fast Service Restoration with IS-IS and LDP

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

rest
orati
on

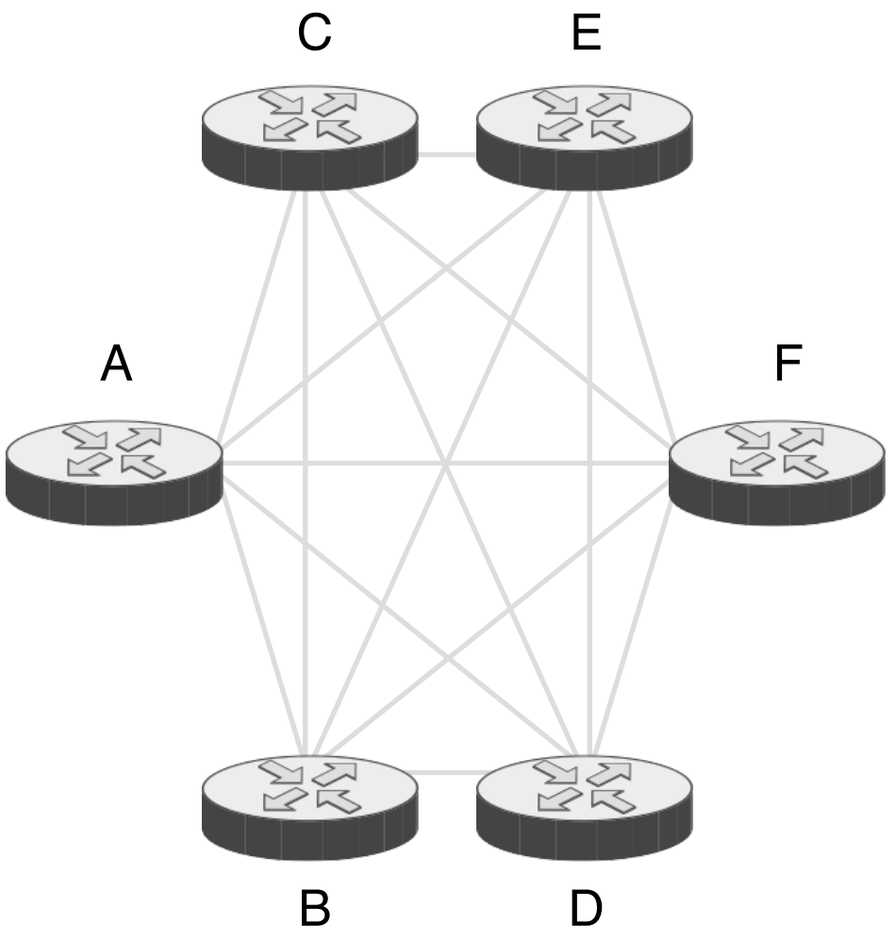
https://t.me/learningnets

# Fast Service Restoration with IS-IS and LDP

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

https://t.me/learningnets

# IS-IS in a Full Mesh Topology




Full Mesh Topology

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

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

https://t.me/learningnets

# IS-IS in a Full Mesh Topology

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

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

https://t.me/learningnets

# IS-IS Advantages and the Disadvantages

disad vanta ges

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

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

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

adva ntag es

https://t.me/learningnets

# IS-IS Advantages and the Disadvantages

disad vanta ges

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

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

adva ntag es

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

https://t.me/learningnets

# IS-IS Advantages and the Disadvantages

disad
vanta
ges

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

adva
ntag
es

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

https://t.me/learningnets

# IS-IS CASE STUDIES

https://t.me/learningnets

# OSPF to IS-IS Migration

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

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

migration

https://t.me/learningnets

# OSPF to IS-IS Migration

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

- 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.

https://t.me/learningnets

# OSPF to IS-IS Migration

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

- 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.

https://t.me/learningnets

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

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

mig
rati
on

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

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

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

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

2. Deploy IS-IS over entire network

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

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

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

2. Deploy IS-IS over entire network

- Deploy both IPv4 and IPv6.

- Deploy IS-IS passive interface at the edge links, these links should be carried in IBGP. Also prefix suppression can be used to carry infrastructure links in IBGP but these are not a requirement of Fastnet.

- Make sure the IS-IS LSDB is consistent with OSPF routing table

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

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

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

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

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

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

migration

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

5. Remove OSPF from entire network.

- All the OSPF processes can be removed.

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

https://t.me/learningnets

# Detail OSPF to IS-IS Migration Steps

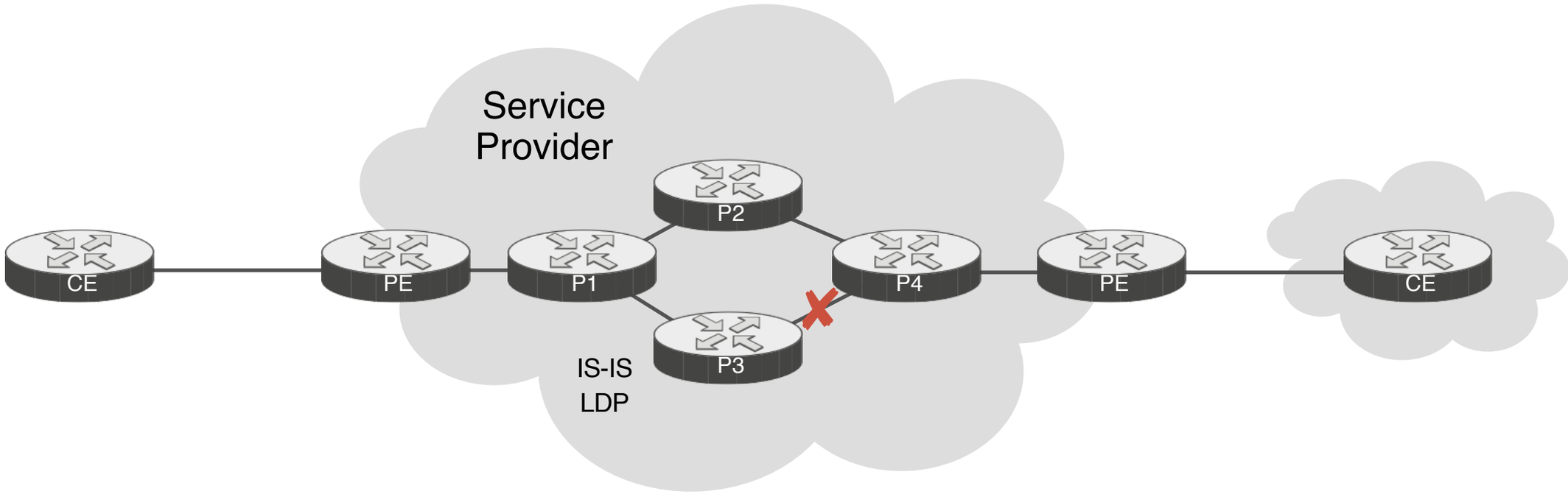
6. Verify IGP and Related operation.

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

Enjoy the party!

https://t.me/learningnets

# IS-IS and MPLS Interaction - Avoiding Black-holing




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

https://t.me/learningnets

# QUESTIONS

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

https://t.me/learningnets

# ANSWER 1

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

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

https://t.me/learningnets

# ANSWER 2

- It doesn't matter which IS-IS Level (L1 or L2 ) is used to provide a solution for this problem.

  - Here the question is seeking if you know the solution already.

- This type of questions will be asked in the CCDE Practical exam and the task domain will be Analyzing the design.

https://t.me/learningnets

# ANSWER 3

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

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

https://t.me/learningnets

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

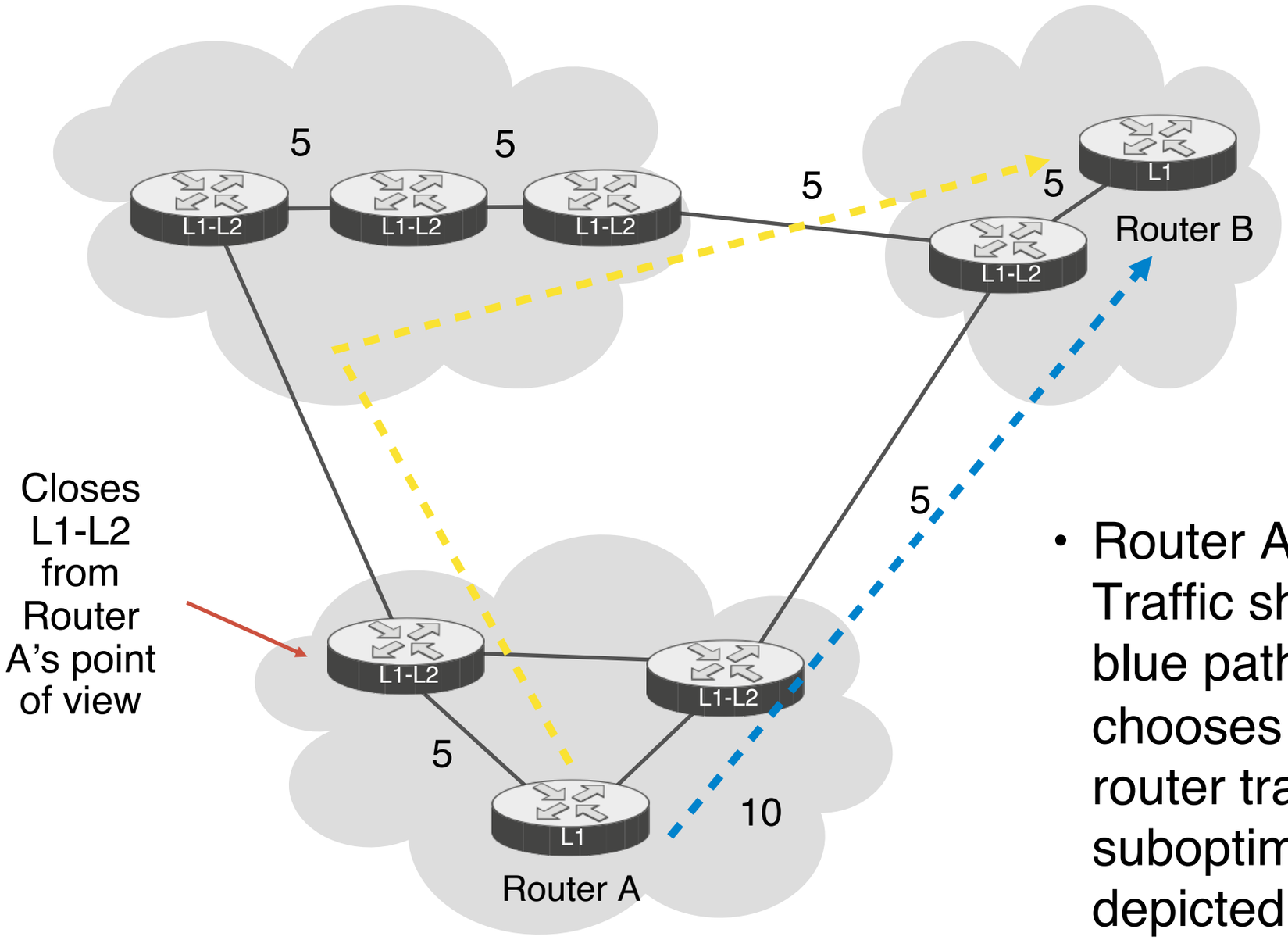
https://t.me/learningnets

# Suboptimal Routing with Multi Level IS-IS Design

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

https://t.me/learningnets

# Suboptimal Routing with Multi Level IS-IS Design




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

https://t.me/learningnets

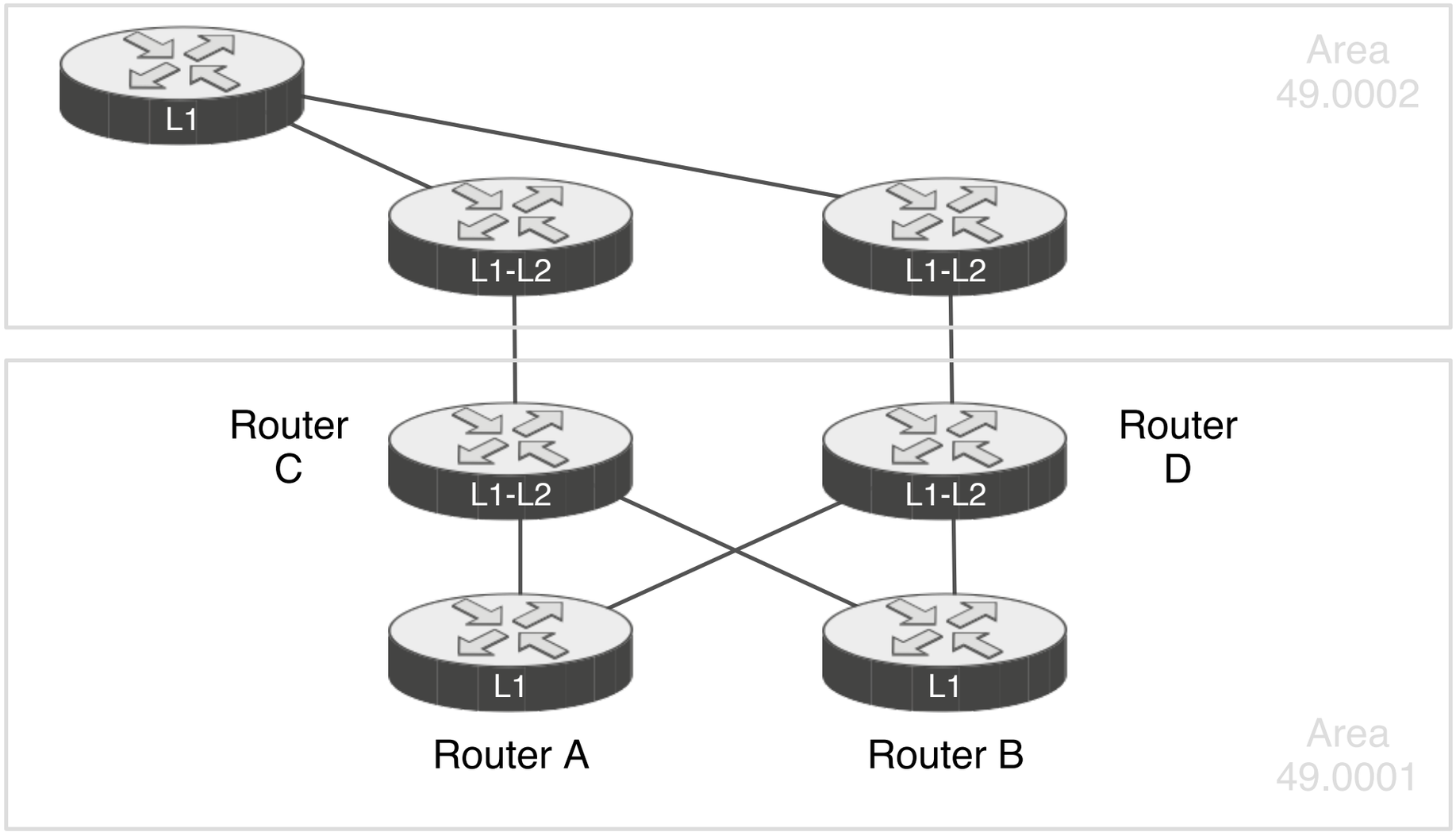
# IS-IS Level Engineering

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

https://t.me/learningnets

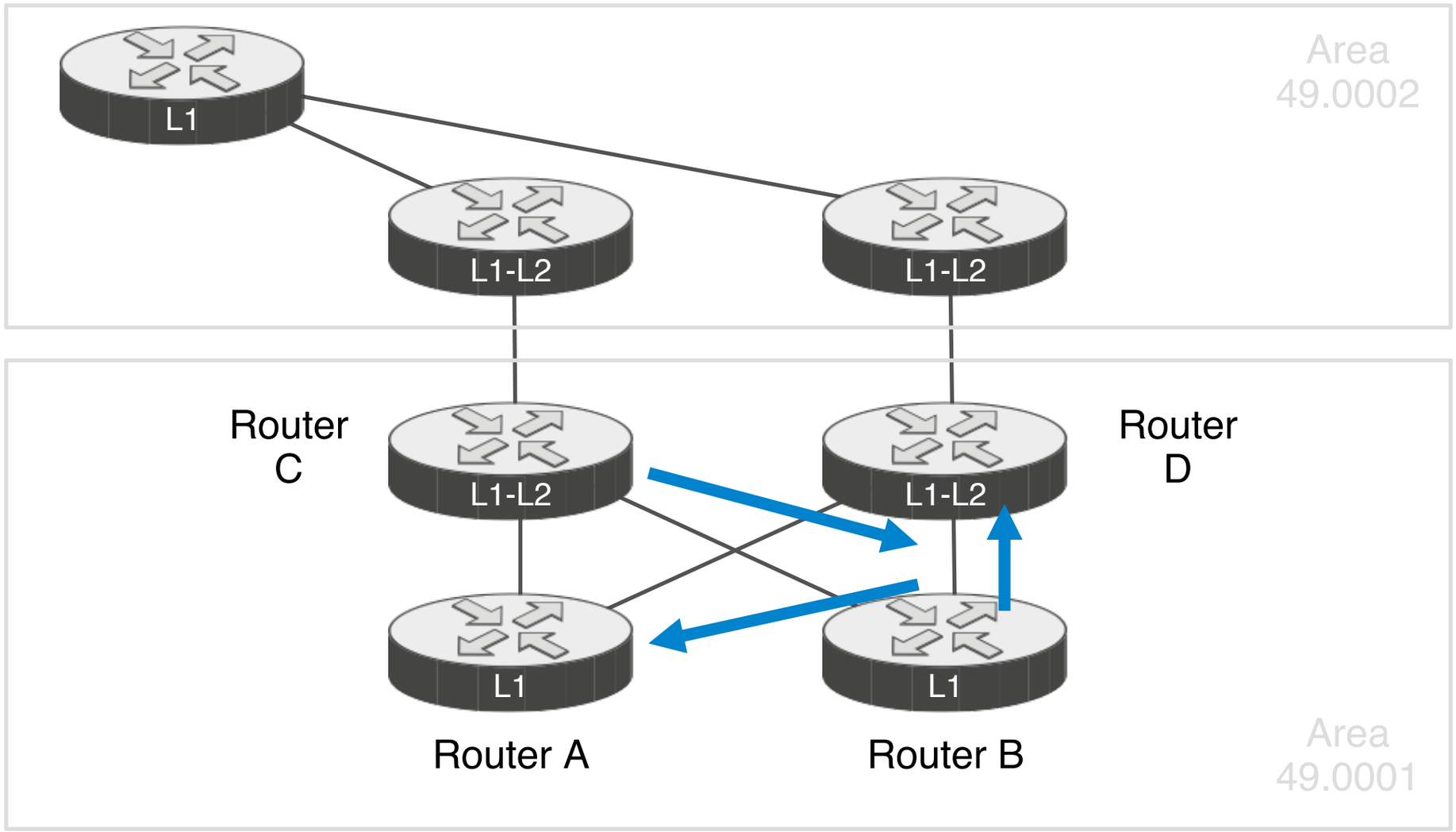
# IS-IS Level Engineering

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




https://t.me/learningnets

# IS-IS Level Engineering



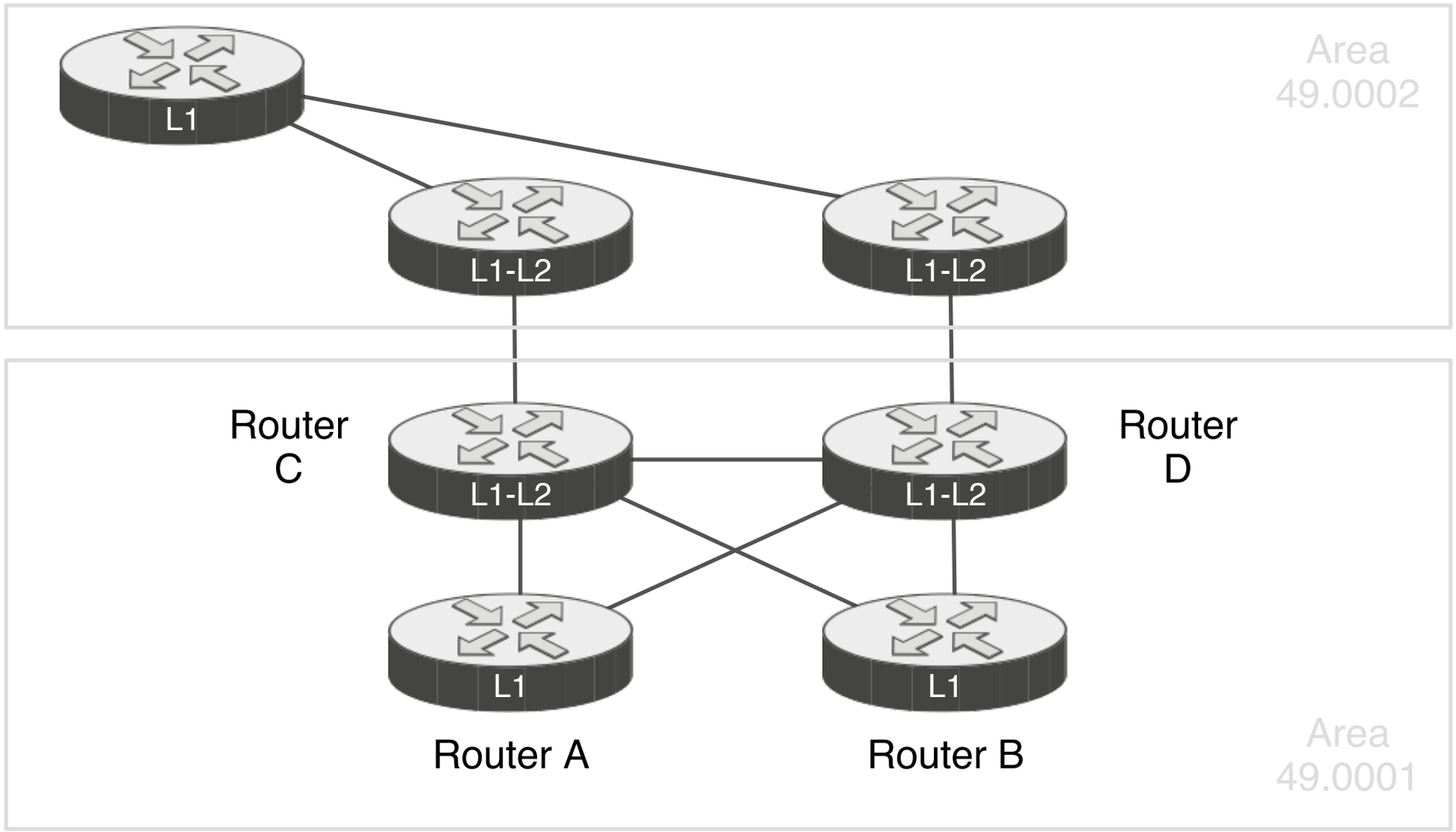

https://t.me/learningnets

# IS-IS Level Engineering

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

https://t.me/learningnets

# IS-IS Level Engineering




https://t.me/learningnets

# IS-IS Level Engineering

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

https://t.me/learningnets

# IS-IS Level Engineering

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

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

https://t.me/learningnets

# IS-IS in the CCDE Exam

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

https://t.me/learningnets

# Summary

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

https://t.me/learningnets

# IS-IS

## Intermediate System to Intermediate System

QUIZ

https://t.me/learningnets

## Question 1

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

A. Backbone Area

B. Stub Area

C. Totally Stub Area

D. Totally NSSA Area

https://t.me/learningnets

## Answer 1

D. Totally NNSA Area

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

https://t.me/learningnets

## Question 2

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

A. Tuned IS-IS LSP Timers

B. BFD

C. Tuned IS-IS SPF Timers

D. IS-IS Hello Timers

https://t.me/learningnets

## Answer 2

B-BFD

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

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

https://t.me/learningnets

## Question 3

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

A. In order to prevent routing loop.

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

C. In order to prevent routing oscillation.

D. For fast convergence.

https://t.me/learningnets

## Answer 3

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

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

IGP should wait BGP before staring to accept network traffic.

https://t.me/learningnets

## Question 4

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

A. ISPF

B. Partial SPF

C. Exponential Back Off

D. LSP Throttiling

E. SPF Throttling

https://t.me/learningnets

# Answer 4

C. Exponential Back Off

D. LSA Throttiling

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

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

https://t.me/learningnets

## Question 5

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

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

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

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

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

https://t.me/learningnets

# Answer 5

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

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

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

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

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

https://t.me/learningnets

## Question 6

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

A. One

B. Two

C. Three

D. As many as possible

https://t.me/learningnets

## Answer 6

B. Two

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

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

https://t.me/learningnets

## Question 7

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

A. Route tag should be used.

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

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

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

https://t.me/learningnets

# Answer 7

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

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

https://t.me/learningnets

## Question 8

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

A. Elect a DIS and Backup DIS.

B. Use IS-IS Mesh Group.

C. Use DR and BDR.

D. Deploy Multi Level IS-IS design.

https://t.me/learningnets

## Answer 8

B. Use IS-IS Mesh Group.

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

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

https://t.me/learningnets

## Question 9

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

A. 100 IS-IS LSP

B. 3 IS-IS LSP

C. 300 IS-IS LSP

D. 1 IS-IS LSP

https://t.me/learningnets

## Answer 9

D.1 IS-IS LSP

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

https://t.me/learningnets

## Question 10

Which below statements are correct for IS-IS design?

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

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

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

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

https://t.me/learningnets

# Answer 10

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

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

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

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

https://t.me/learningnets

# IS-IS – Study Resources:

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

https://t.me/learningnets

# EIGRP

## Enhanced Interior Gateway Routing Protocol

https://t.me/learningnets

# Agenda

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

https://t.me/learningnets

# Theory

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

https://t.me/learningnets

# EIGRP RFC

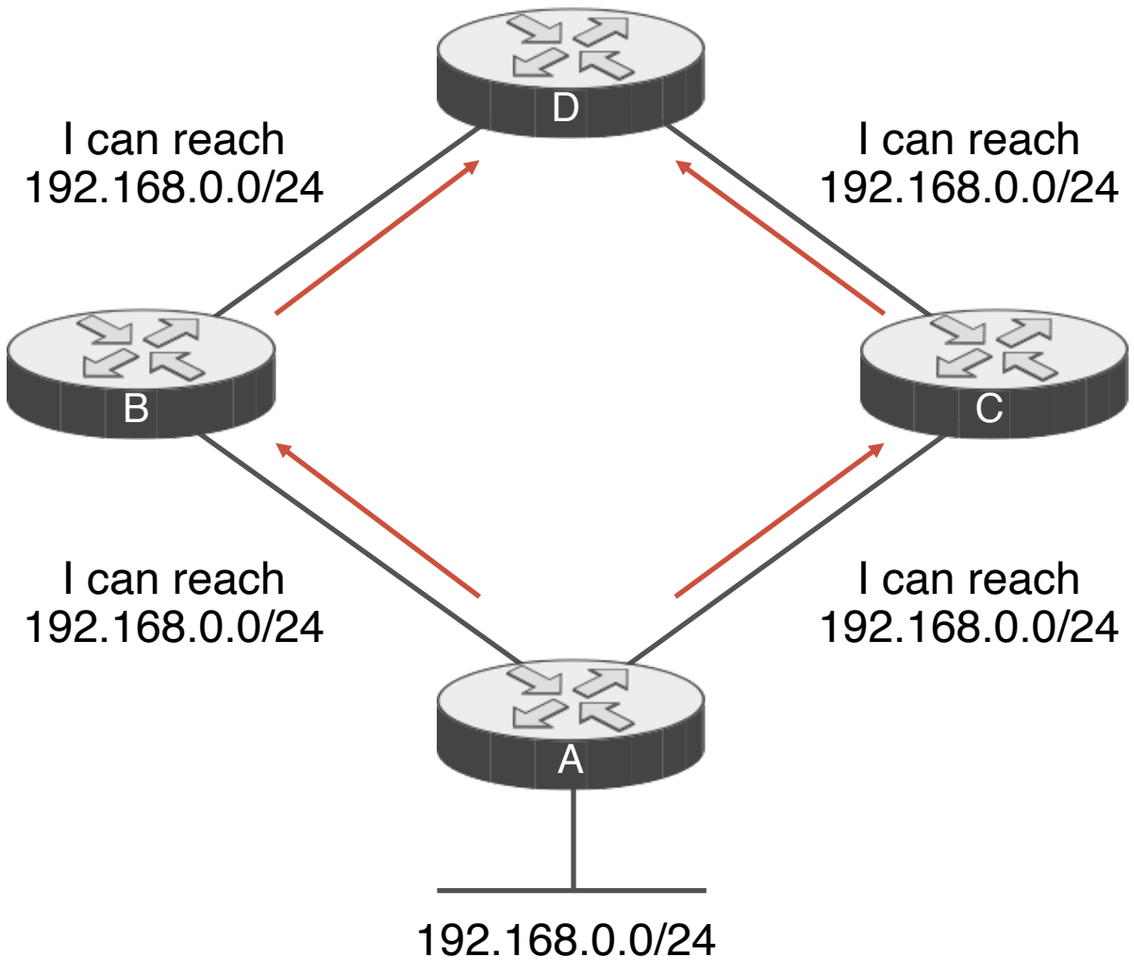
- EIGRP is Cisco preparatory protocol, May 2016, RFC (Informational Track) 7868 published.

  - In EIGRP RFC, EIGRP Stub feature has been not shared.

- Vendor interoperability is an issue with EIGRP, when you want to deploy different vendor equipment, you may have an issue.

https://t.me/learningnets

# EIGRP Routers don’t have topology information of the network




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

EIGRP only knows prefix and next hop information.

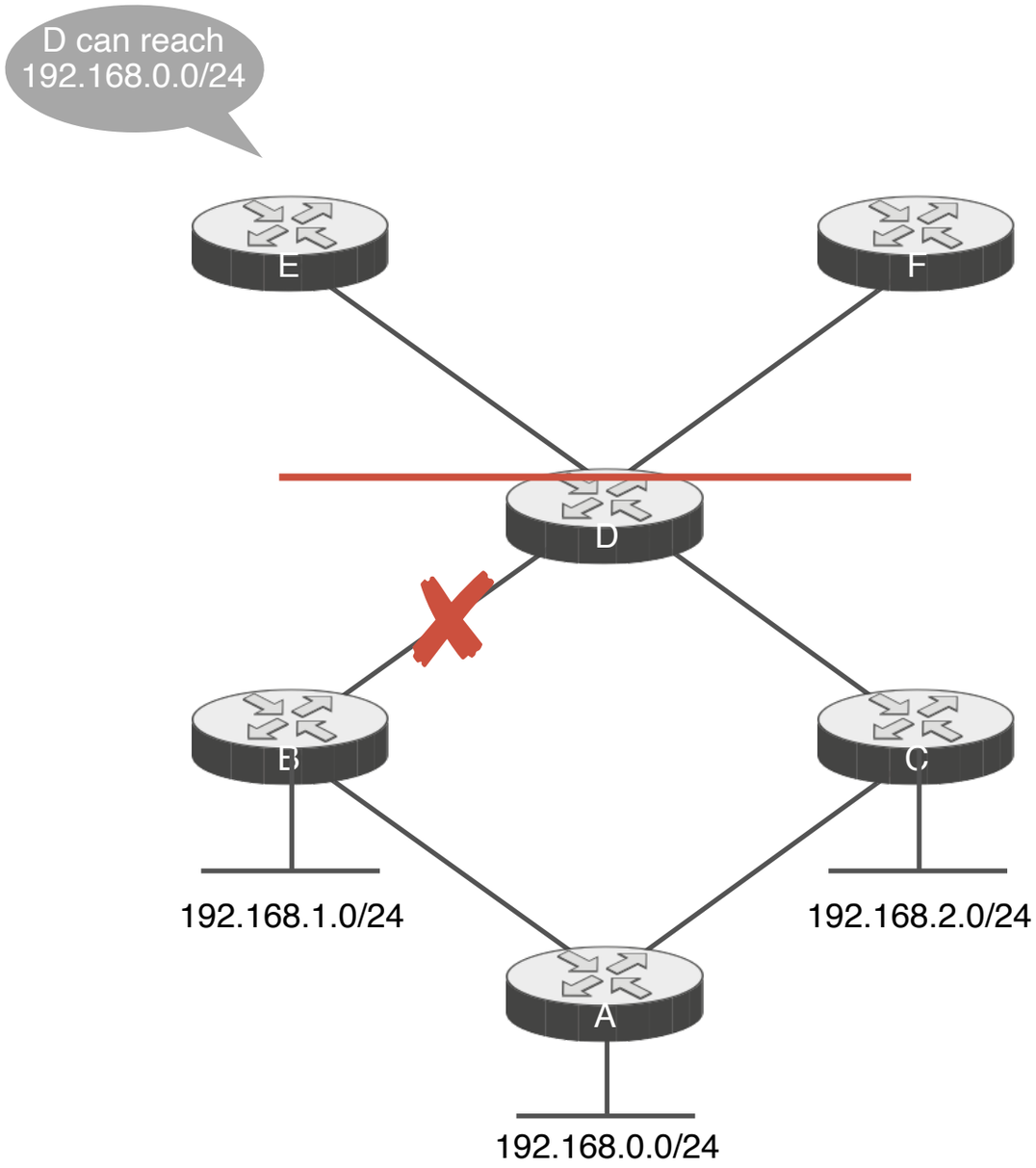
In this topology A advertises that it has 192.168.0.0/24 network.

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

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

https://t.me/learningnets

# EIGRP Routers don’t have topology information of the network




For E and F Topology information is hidden here.

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

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

https://t.me/learningnets

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

https://t.me/learningnets

# EIGRP Terminology

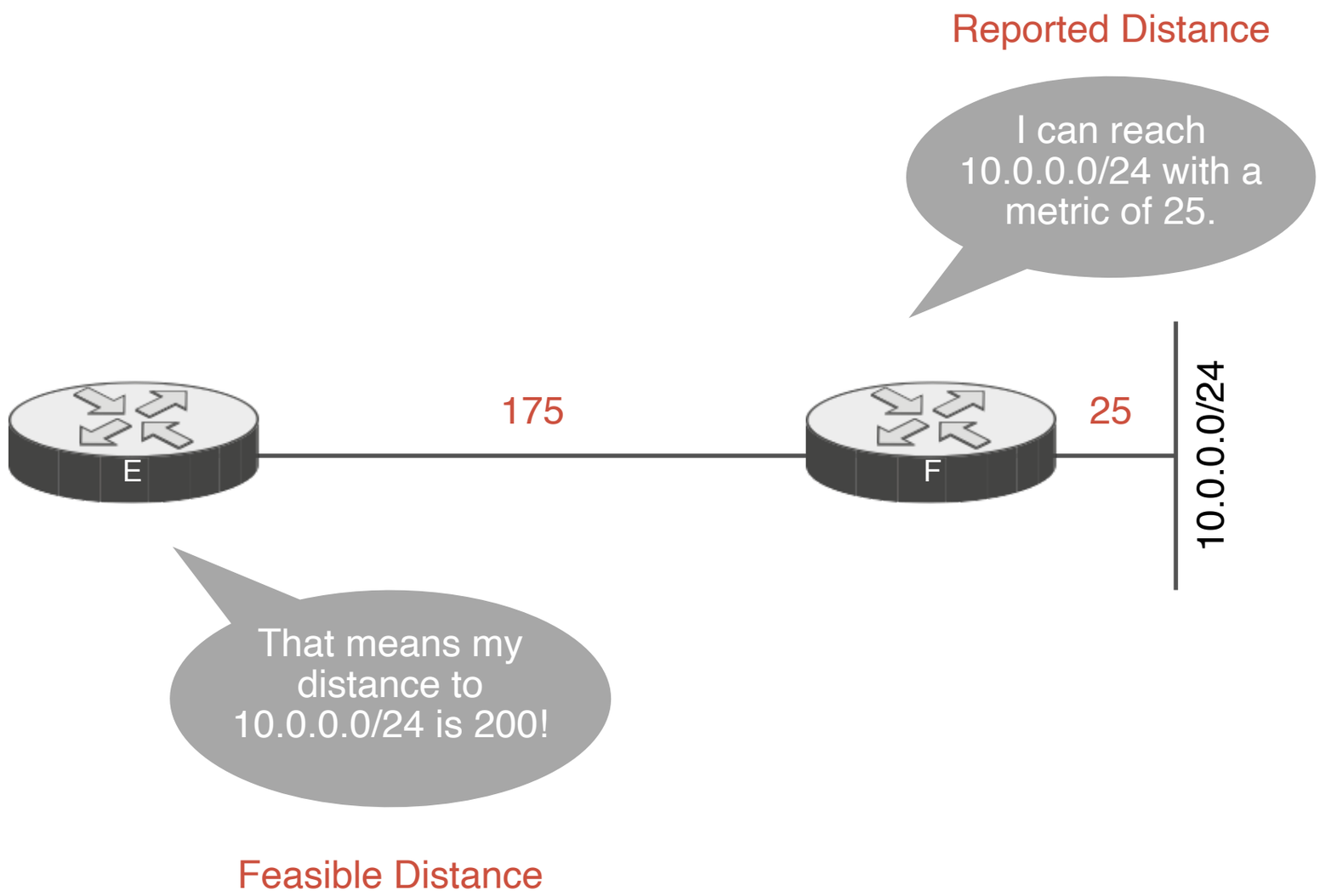
- Feasible distance is the best path, primary path Successor is the next-hop router for the route.

  - Each EIGRP router advertise its routes with a metric called Reported distance, also known as advertised distance.

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

https://t.me/learningnets

# Feasible Distance and Reported Distance




https://t.me/learningnets

- Feasible successors are placed in EIGRP topology table.
  - Reported distance is the feasible distance of the neighboring router.

https://t.me/learningnets

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

feasible

https://t.me/learningnets

# EIGRP Metric Calculation

- Classical EIGRP uses 5 K Values for metric calculation.
  - K1 through K5

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

me
tric

https://t.me/learningnets

# EIGRP Metric Calculation

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

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

metric

https://t.me/learningnets

# EIGRP Metric Calculation

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

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

me
tric

https://t.me/learningnets

# EIGRP Metric Calculation

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

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

https://t.me/learningnets

# EIGRP Fast Convergence and Fast Reroute

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

https://t.me/learningnets

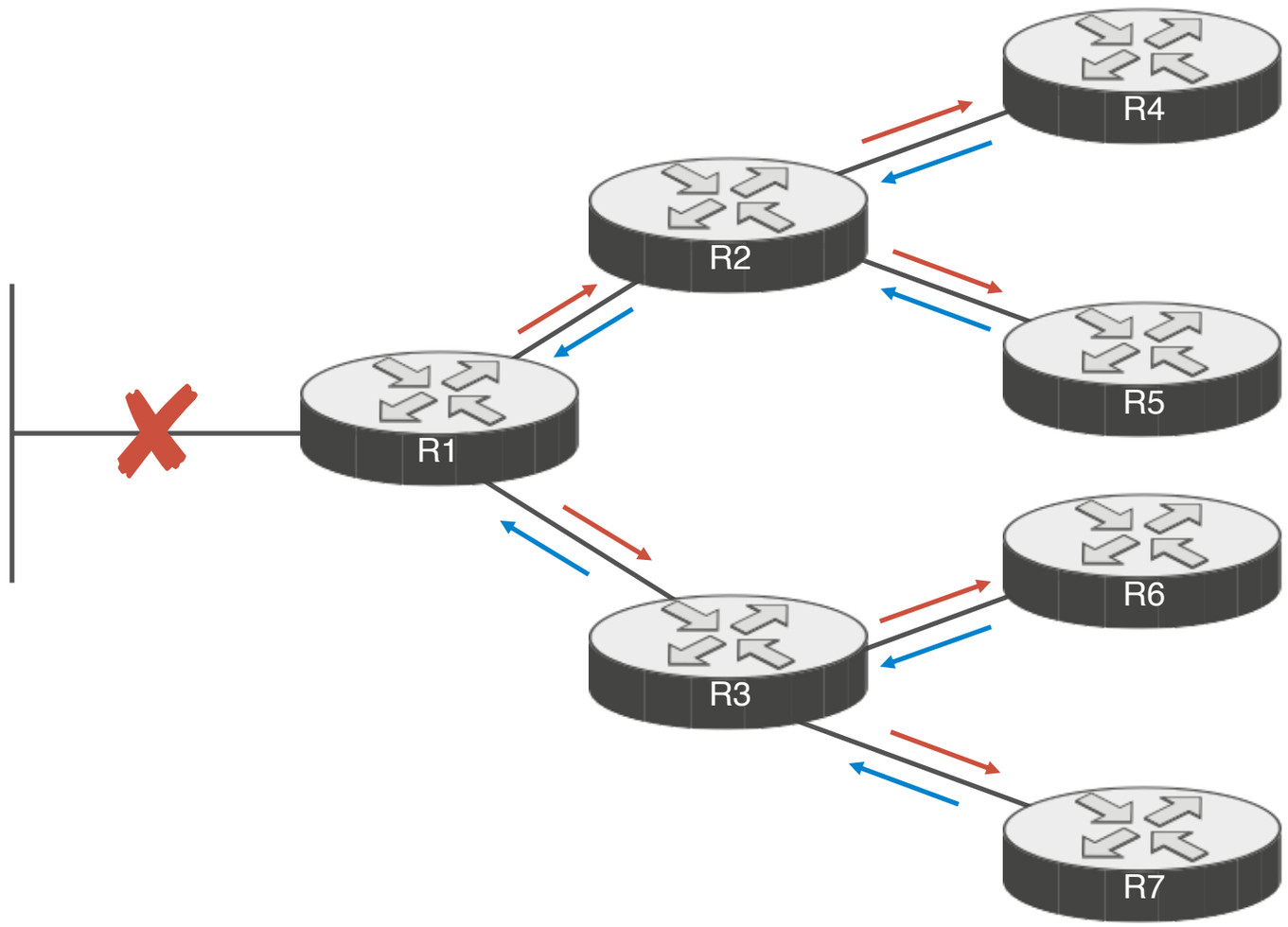
# EIGRP Fast Convergence and Fast Reroute

- If there is a feasible successor, without running DUAL algorithm, backup route is taken from the EIGRP topology table and placed into the RIB and FIB.

- If there is no feasible successor (backup route), prefixes are marked as active and EIGRP query is sent to all EIGRP neighbors of the node.

https://t.me/learningnets

# EIGRP Query is sent to all EIGRP neighbors




https://t.me/learningnets

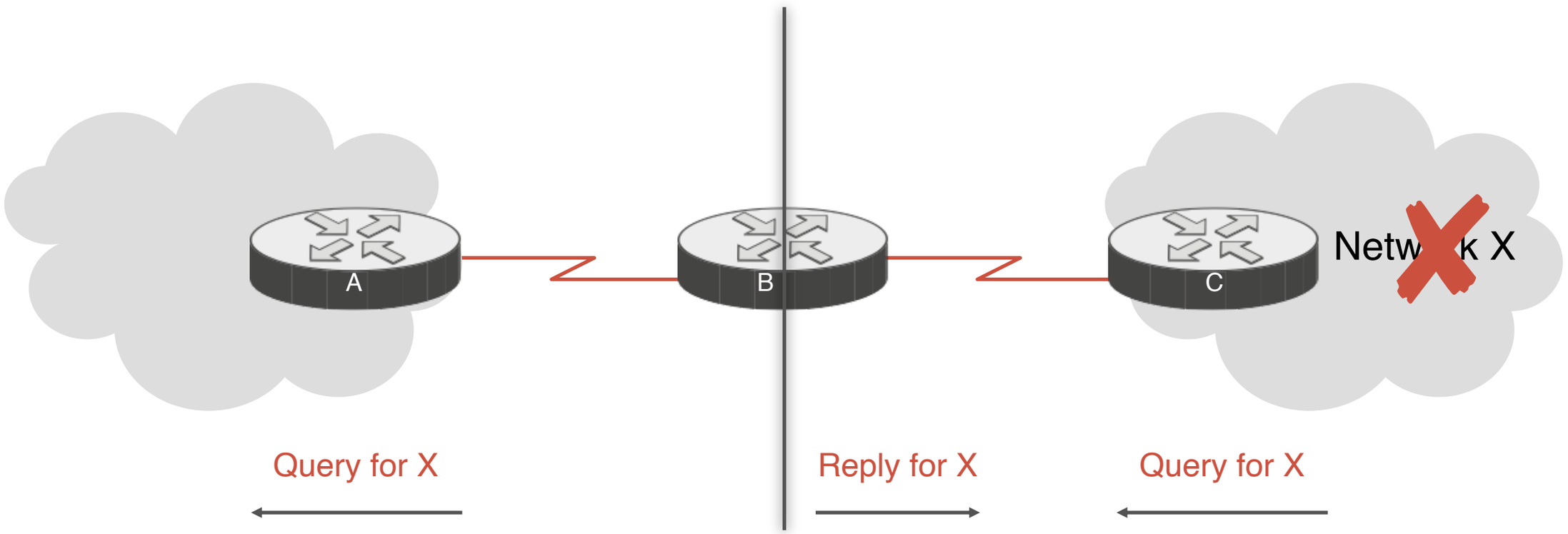
# EIGRP Query is sent to all EIGRP neighbors

- The router will have to get all of the replies from its neighbors before the router calculates the new path.
  - If any neighbor fails to reply the query within three minutes, this route is stuck in active (SIA), and the router resets the neighbor which fails to reply.
- Solution is to limit query range which is also known as EIGRP query domain!

https://t.me/learningnets

# EIGRP Query Scope in multi AS scenario

EIGRP queries are not bounded by AS Boundary.
Queries from AS 1 will be propagated to AS 2 in below topology




https://t.me/learningnets

# EIGP Convergence

- EIGRP Feasible successor helps for fast convergence.
  - For fast failure detection, BFD can be used.
- EIGRP supports IP FRR as well, in this case, feasible successor routes are not only placed in EIGRP topology table, but also in RIB and FIB as a standby route.

https://t.me/learningnets

# EIGRP Convergence

- Having EIGRP Feasible Successor (Backup Route) is important for the fast convergence, but how feasible successor is calculated?
  - Answer is, backup route should satisfy the feasibility condition!
- What is EIGRP feasibility condition?

https://t.me/learningnets

# EIGRP Feasibility Condition

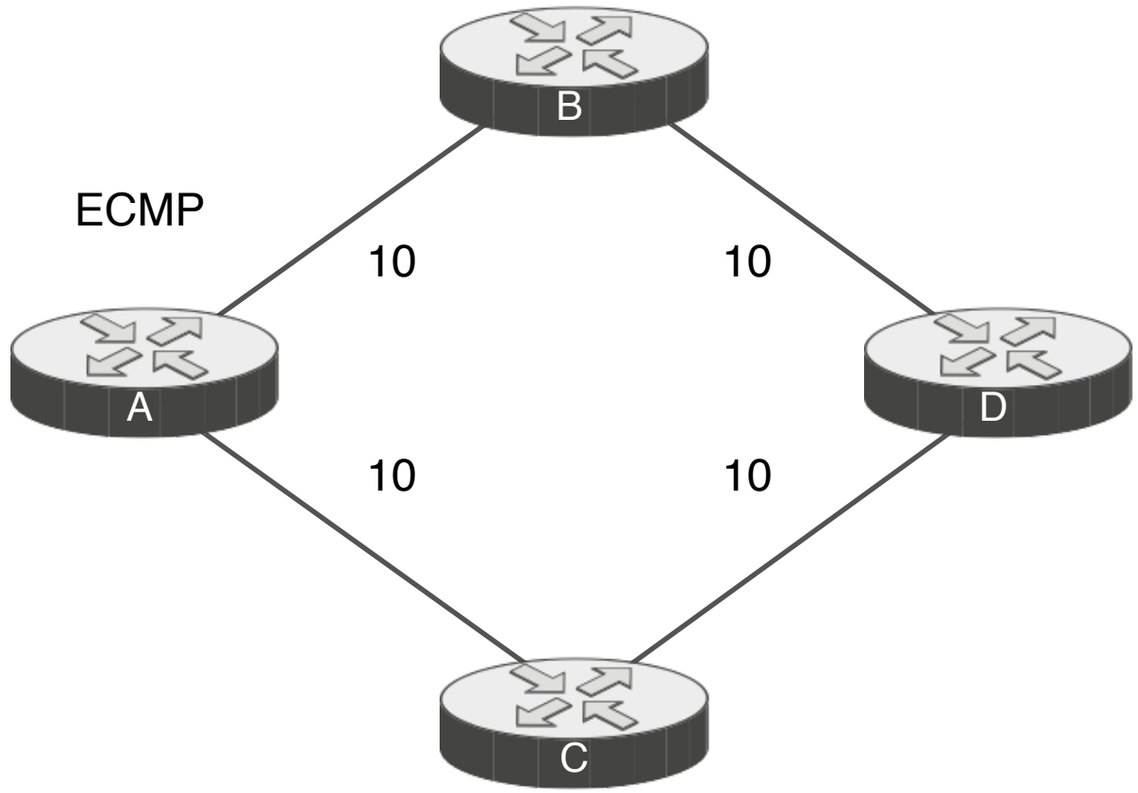
- If node A'S  next-hop router's reported distance is less than the feasible distance of node A than A's backup router is loop free. And it can be used as a backup and placed in the topology table.

feasibility

- In order the path to be placed in the EIGRP topology table, it has to satisfy the EIGRP feasibility condition.

https://t.me/learningnets

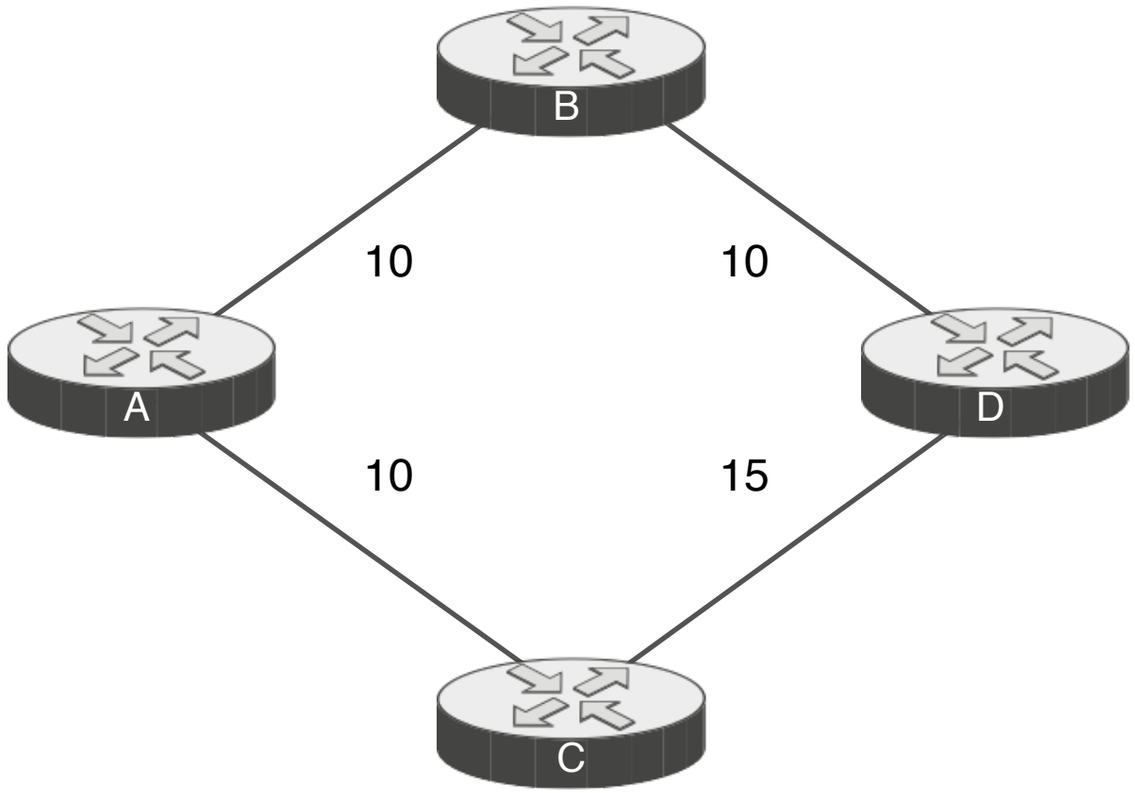
# EIGRP Feasibility Condition




From the Router A's point of view, Router B and Router C is the Equal Cost routers, so both ABD and ACD path can be used. Router A installs both Router B and Router C in not only EIGRP topology table but also in the routing table.

https://t.me/learningnets

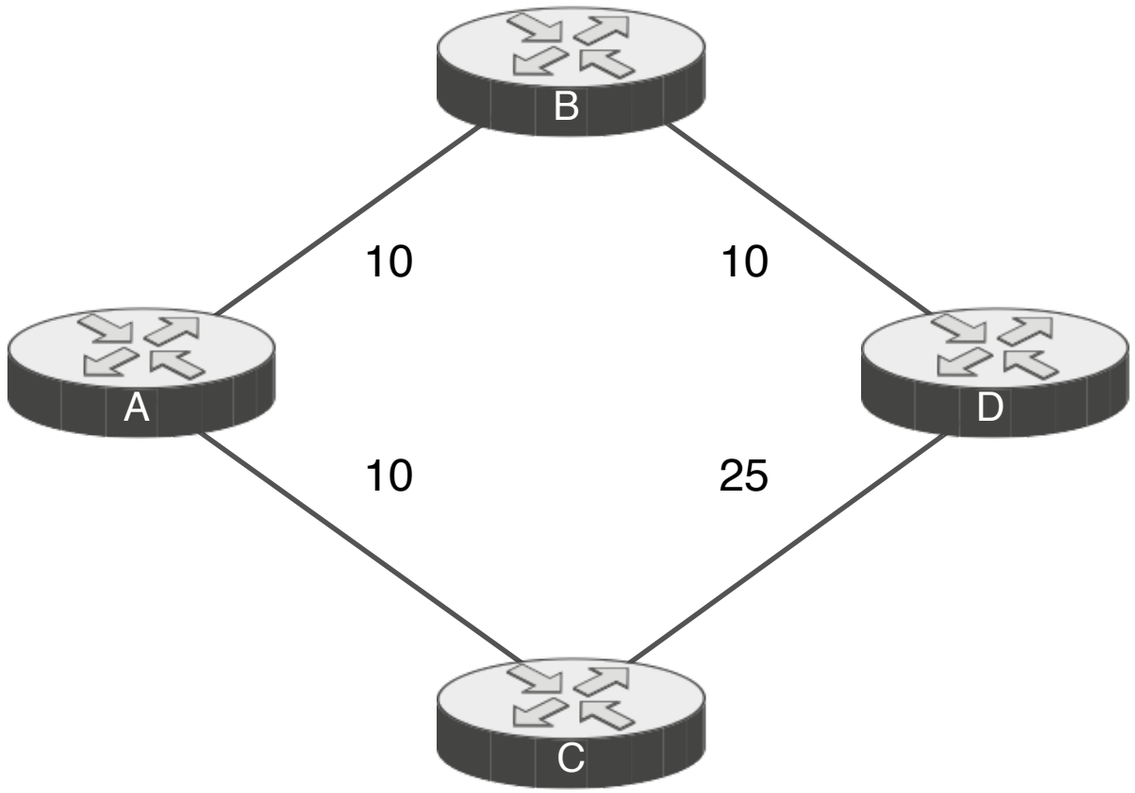
# EIGRP Feasibility Condition




Router C to Router D link cost is 15. In order to satisfy feasibility condition for the Router A, Router  C -  Router D link cost should be smaller than, A- B – D total cost. Since 15 < 10 + 10, Router C can be used as a backup router by the Router A to Reach Router D

https://t.me/learningnets

# EIGRP Feasibility Condition

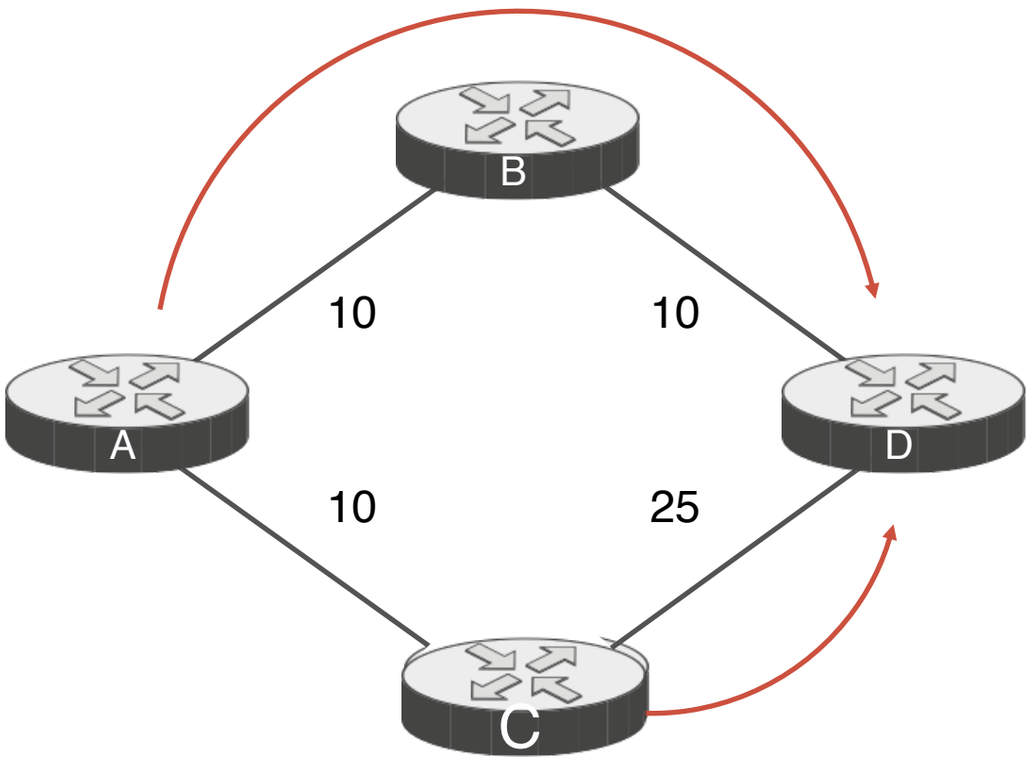



Router C to Router D link cost is 25. In order to satisfy feasibility condition for the Router A, Router C - Router D link cost should be smaller than, A- B – D total cost Since 25 > 10(A-B) + 10(B-D), Router C can not be used as a backup router by the Router A to Reach Router D. What if C-D would be 20?

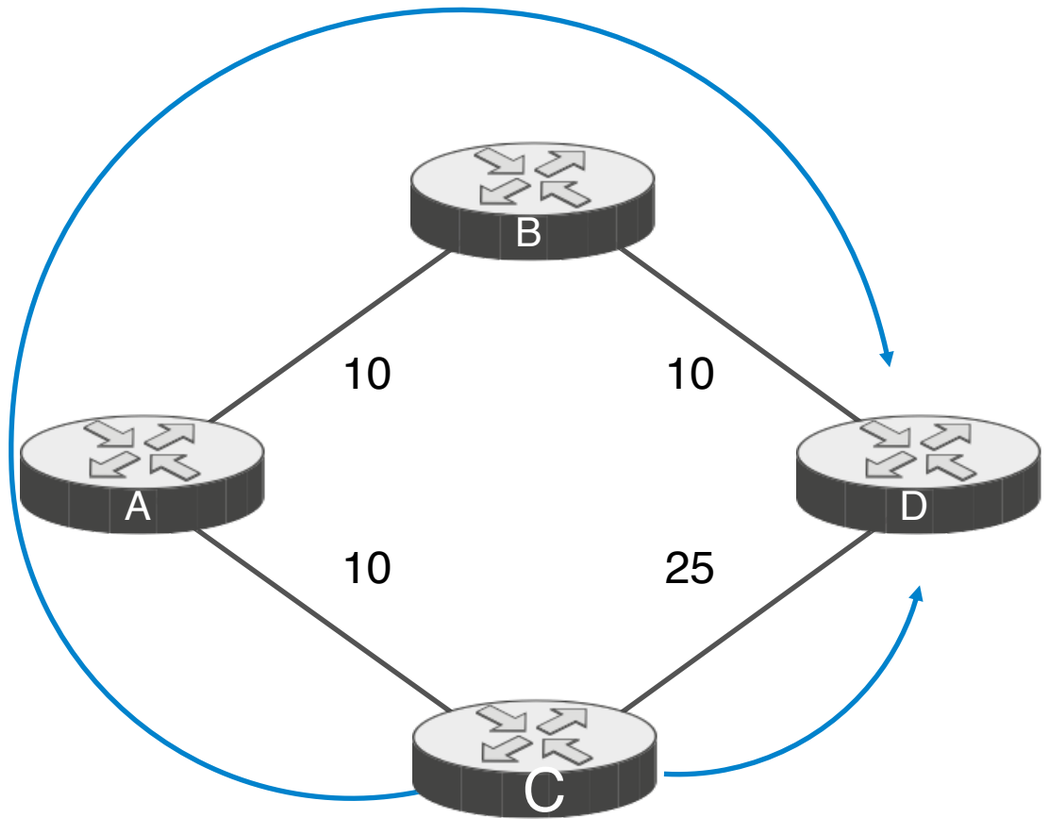
https://t.me/learningnets

# Difference between EIGRP FS and EIGRP LFA Calculation

## EIGRP FS




## EIGRP LFA




https://t.me/learningnets

# EIGRP Scalability

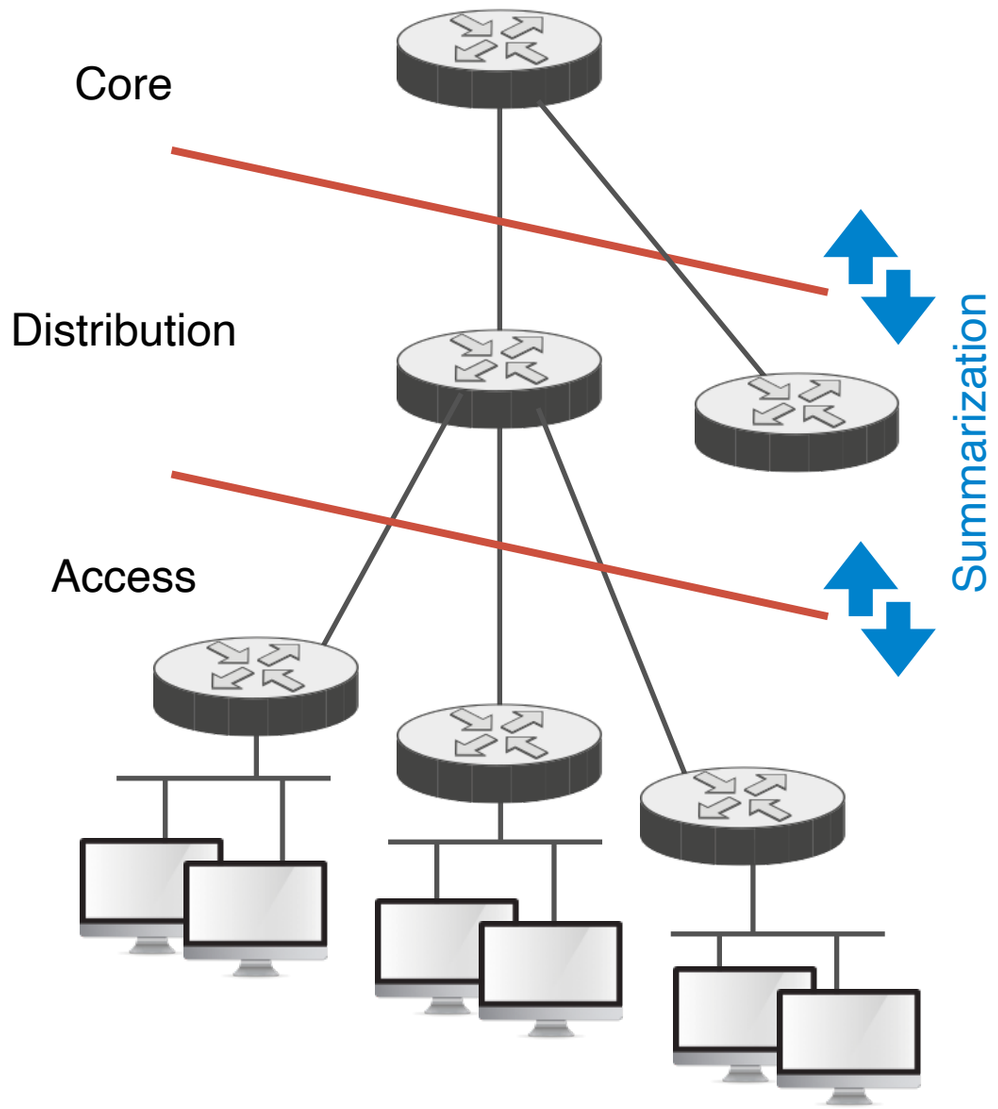
- EIGRP supports manual summarization in any interface at any router in the network, because there is no topology information.

scalability

- Unlike OSPF and IS-IS, EIGRP is not limited with two tier hierarchy, summarization can be done at multiple points.

https://t.me/learningnets

# EIGRP Summarization can be done at multiple points, such as both Core and Distribution




https://t.me/learningnets

# EIGRP Summarization can be done at multiple points, such as both Core and Distribution

•For EIGRP, limiting query domain is very important for scalability.

- Query domain can be kept small by introducing summarization and filtering.

summarization

https://t.me/learningnets

# EIGRP Summary and Filters

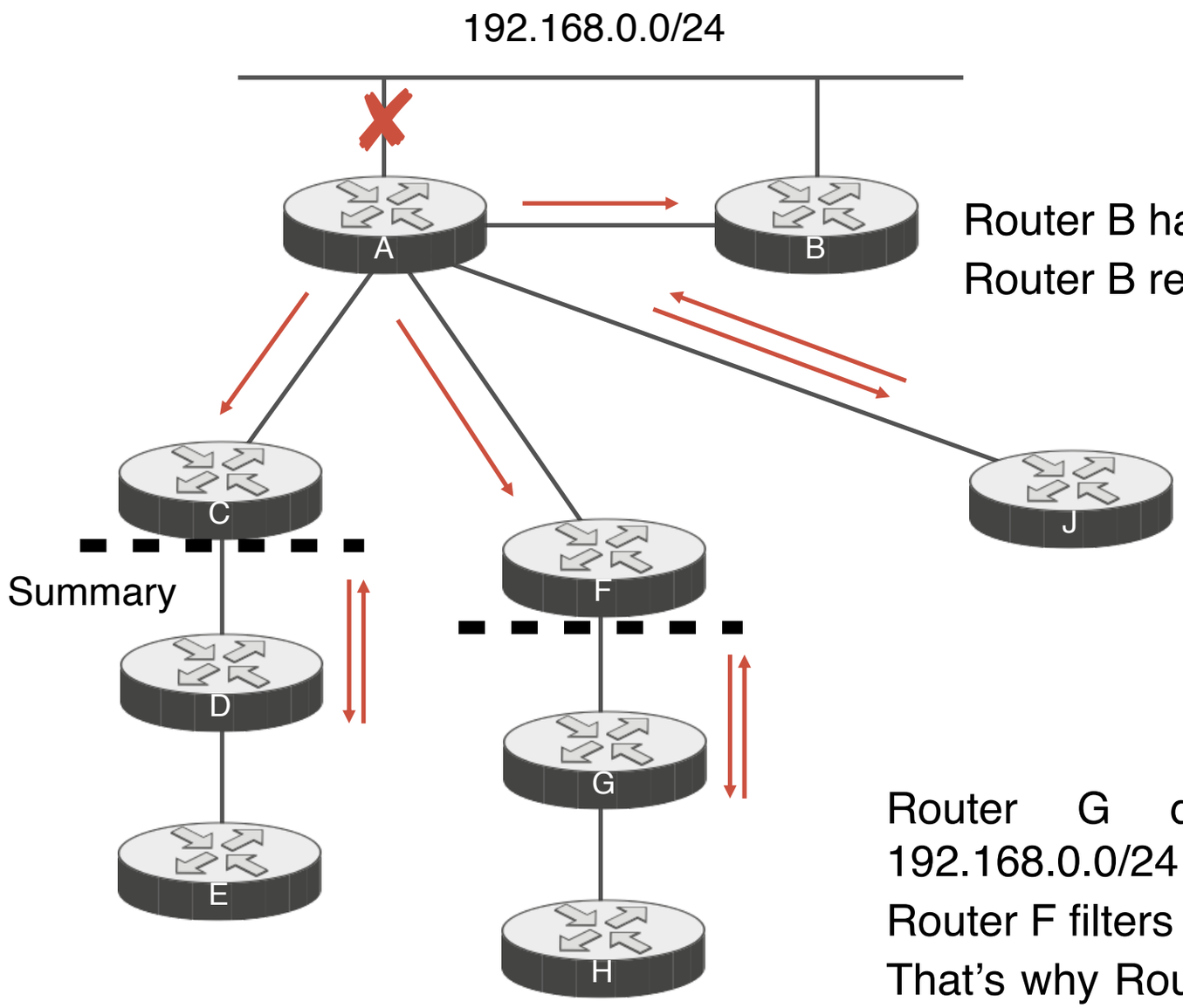
- Summary and Filters reduces the query domain to one hop beyond.
  - Summary and Route filters help for scalability but the problem, nodes still send and receive the EIGRP query in case link or node fails.
- If you want to stop receiving EIGRP query, then EIGRP stub feature should be enabled.

filt
ers

https://t.me/learningnets

When EIGRP node loses the Connection to the prefixes. If there is no feasible successor installed in EIGRP topology database.

Router is marked as active and EIGRP query is sent to every neighbor.




Router B has an alternate path.
Router B replies immediately.

Router J doesn't have any EIGRP neighbor. It replies the Query immediately.

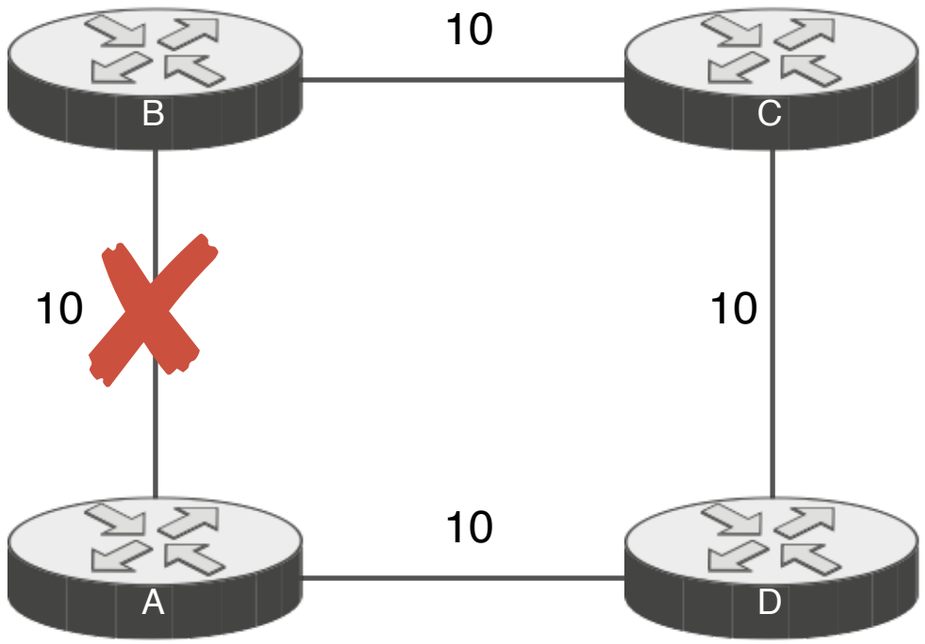
Router D doesn't know the 192.168.0.0/24 network. Router C sends a summary 192.168.0.0/16.
That's why it replies without asking to Router E.

Router G doesn't know the 192.168.0.0/24 network.
Router F filters the 192.168.0.0/24
That's why Router G replies without asking to Router H.

https://t.me/learningnets

# Query Domain Range in Ring Topology

- EIGRP is best in Hub and Spoke topology but not good at Rings. See the below examples.

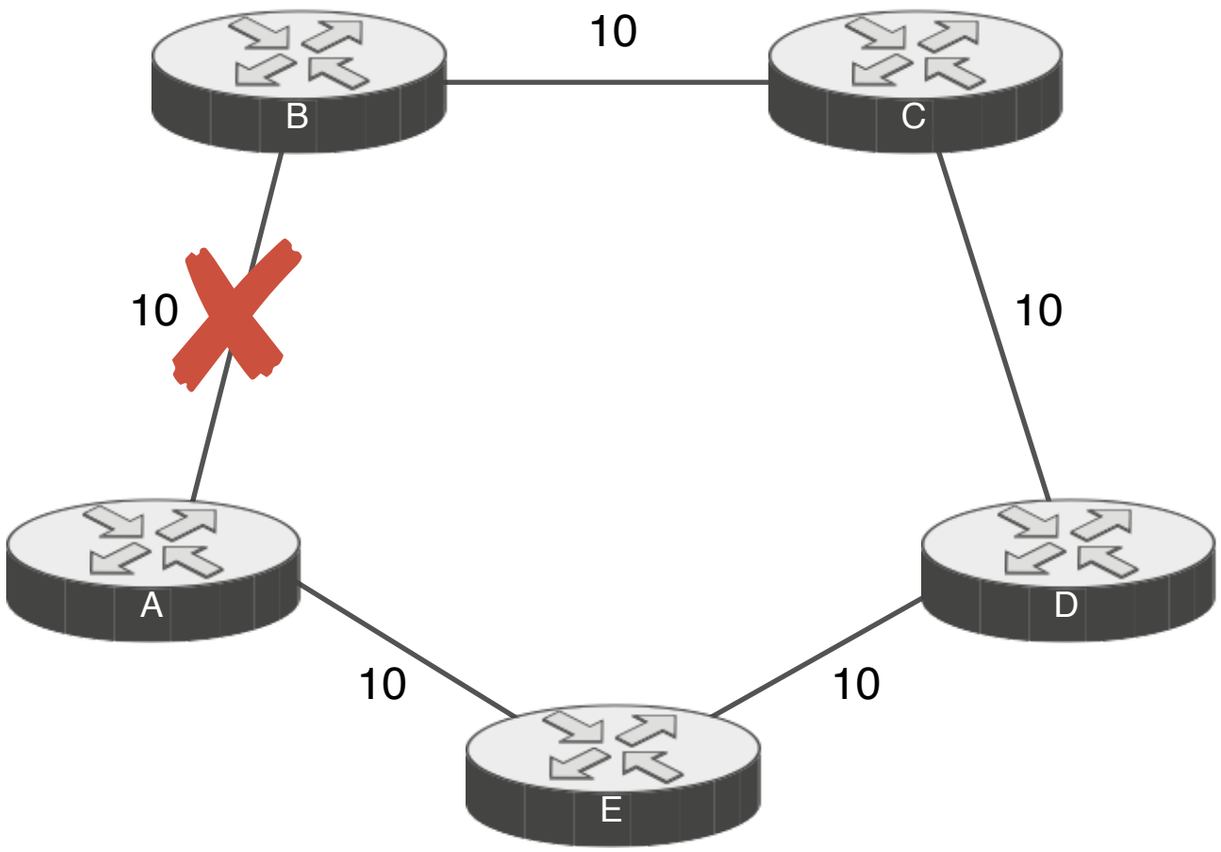



in this network, if A-B link fails, Router A sends a query to Router D, Router D sends a query to Router C.

Siince Router C is not using Router D to reach Router B, query stops at Router C and Router C replies that it has an Alternate path.

https://t.me/learningnets

# Query Domain Range in Ring Topology




in this network, if A-B link fails, Router A sends a query to Router E, Router E sends a query to Router D and Router D sends a query to Router C.

Query range in this network is three hops.

Ring topology is very challenging for all IGP protocols including EIGRP.

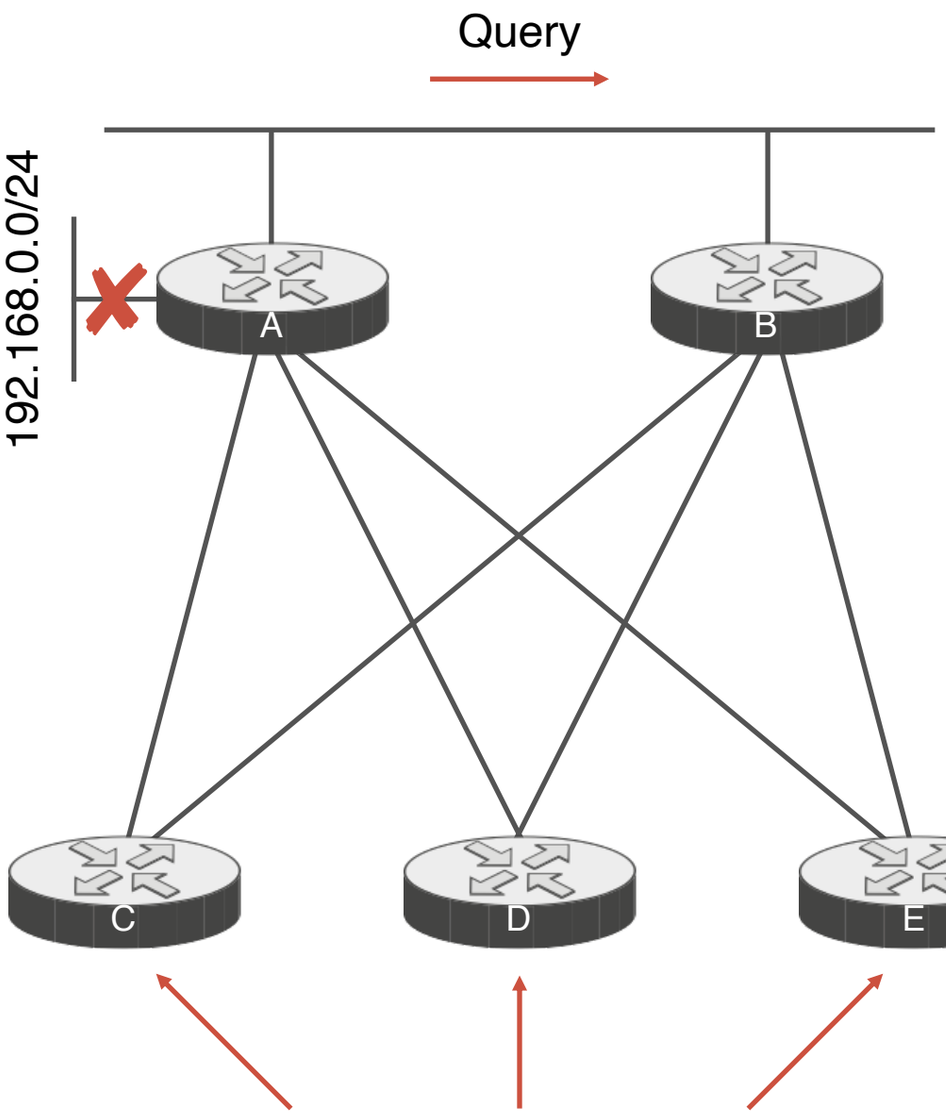
If this topology would be triangle instead of ring EIGRP could find a feasible successor and convergence would be very fast.

https://t.me/learningnets

- In Hub and Spoke, Spokes should be a STUB. This is similar behavior of BGP transit AS. In order to prevent being Transit AS as a customer, Customer only send his AS path to the upstream and filters all the other AS.

https://t.me/learningnets

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




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

EIGRP STUB

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

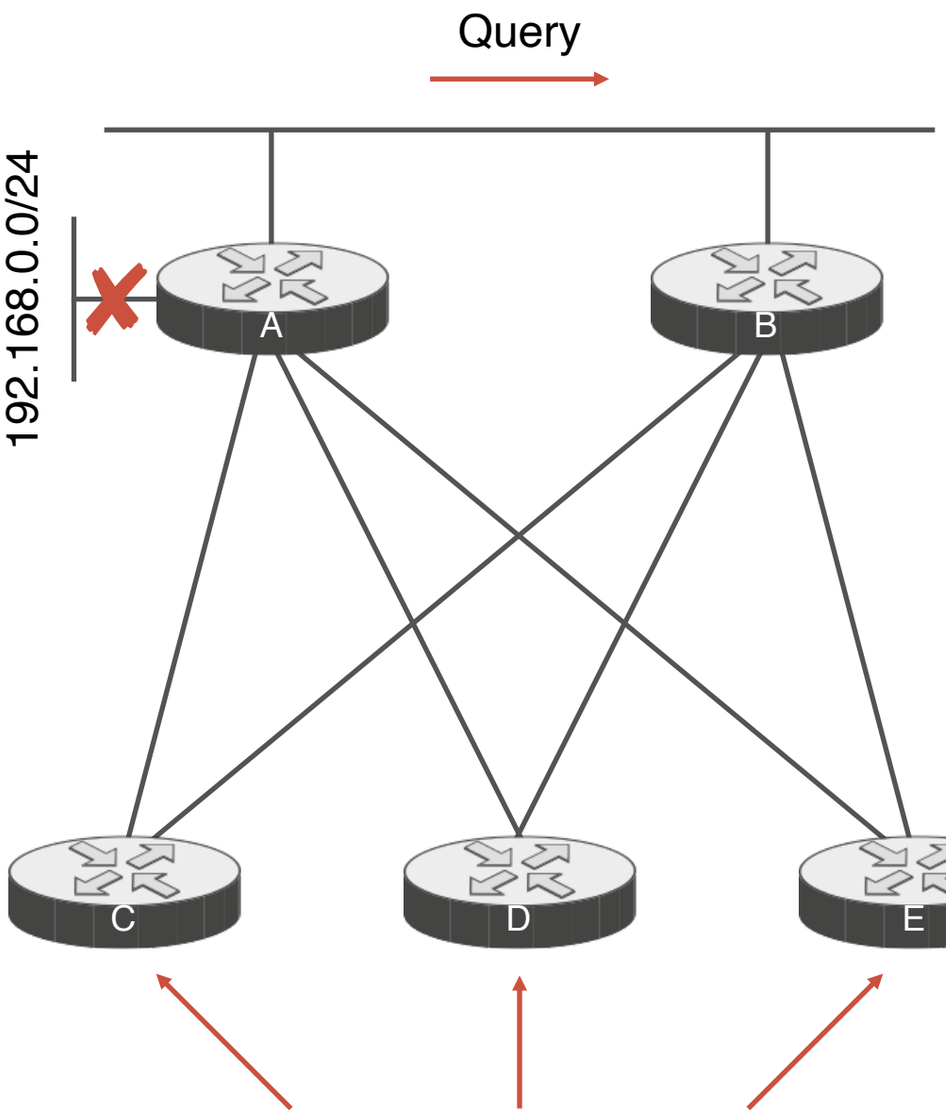
Also Hub site doesn't even send an EIGRP query if 192.168.0.0/24 network fails as in the picture Router A sends a query only to Router B. This helps for the convergence, provide faster convergence.

If EIGRP Stub wouldn't be enabled but filtering or summarization is enabled at the Hub sites, spokes sites still would receive a query and they would process.

This might create a resource problem on the Hub for large scale Hub and Spoke deployment.

https://t.me/learningnets

# Access network always should be configured as stub




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

EIGRP is the most stable IGP protocol in Hub and Spoke Topologies.

Spoke sites have two connections for redundancy not to be used as transit between Hub sites.

Having a direct link between the Hub sites is important.

Otherwise Spoke sites might be used as transit nodes.

https://t.me/learningnets

# EIGRP Scalability Summary

- Limiting query domain is important for scalability.
  - Summary, Filtering and most importantly EIGRP Stub feature limits the EIGRP query range/scope.
- EIGRP allows manual summarization at each hop, any interface at any router, no need an ABR, L1L2 router as in the case with OSPF and IS-IS.

summary

https://t.me/learningnets

# EIGRP Scalability Summary

- Ring topology is hardest topology for EIGRP scalability.
  - Installing more Feasible Successors increase resource usage but provides fast convergence.

summary

https://t.me/learningnets

# Overlay Technologies and EIGRP (GRE, MGRE, DMVPN, GETVPN, LISP)

- EIGRP can work on top of many overlay technologies. EIGRP can be used as an Underlay for these Overlays (VPNs) as well

  - GRE, MGRE, DMVPN, GETVPN and LISP can be used to create overlay/VPN in the networks.

- EIGRP can be used for these overlay mechanisms as an underlay infrastructure routing protocol.

https://t.me/learningnets

# Overlay Technologies and EIGRP (GRE, MGRE, DMVPN, GETVPN, LISP)

- EIGRP works over GRE, MGRE and DMVPN.

- EIGRP doesn't work over GETVPN and LISP, because both are tunnelless VPN mechanisms, routing protocols can be an underlay for them but not an overlay.

ove
rlay

https://t.me/learningnets

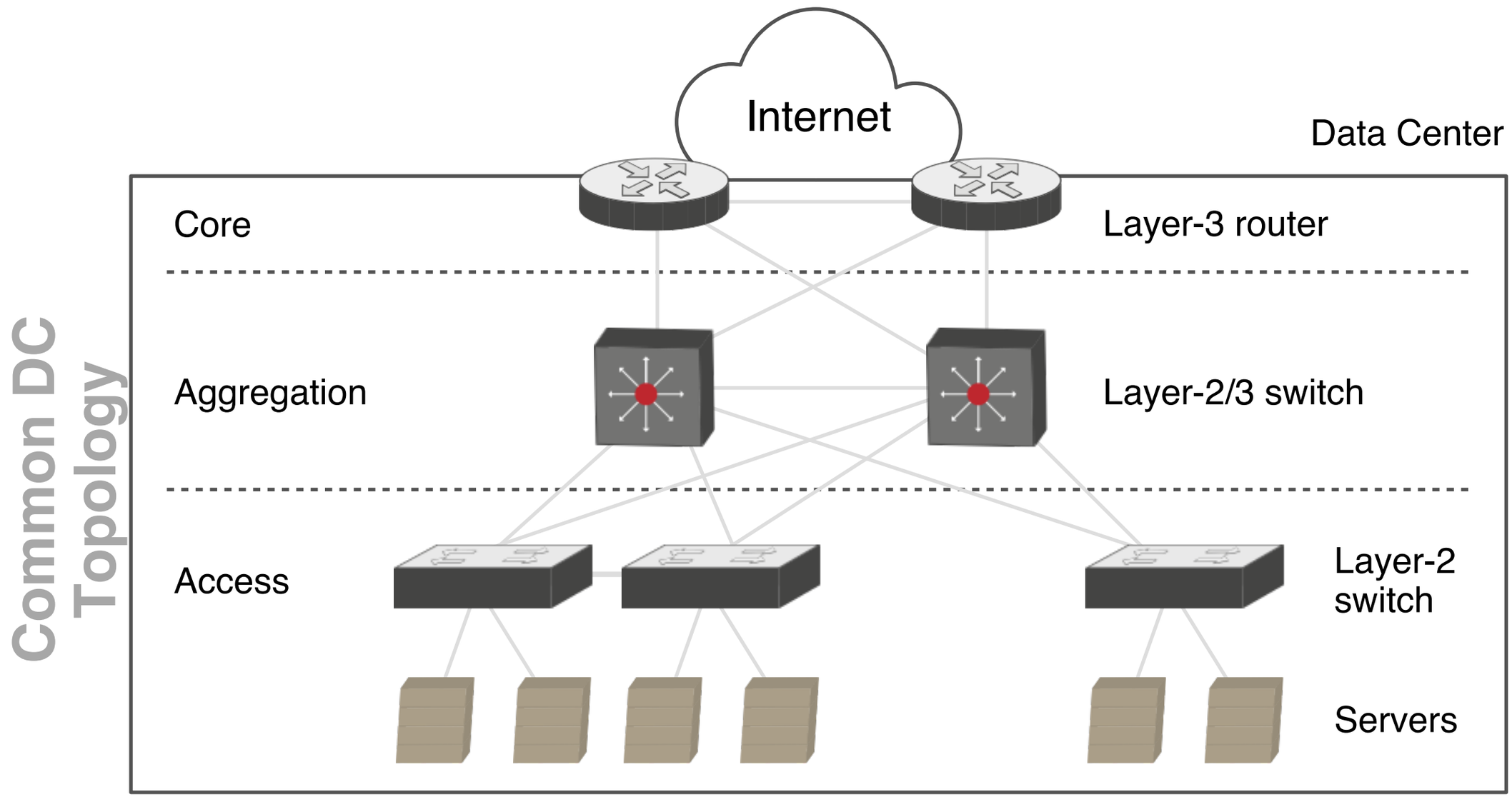
# Overlay Technologies and EIGRP (GRE, MGRE, DMVPN, GETVPN, LISP)

- EIGRP is one of the best mechanisms which work over DMVPN in large scale designs.
  - EIGRP with GRE is not scalable for large scale deployment but scaling limitation comes from GRE, it is not the EIGRP problem, MGRE provides scalability with EIGRP even in large scale deployment.
- EIGRP can work as a PE-CE protocol in MPLS L3 VPN deployment which we will talk in the MPLS lesson.

https://t.me/learningnets

# EIGRP in the Datacenter

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




https://t.me/learningnets

# EIGRP in the Datacenter

- Also EIGRP can be used as a Datacenter Fabric Protocol, but this is not common.

- Datacenters are very densely connected networks, from Core towards access, summarization would be highly necessary as Access switches (TOR in CLOS topology) don’t need to have full routing table and they cannot due to resource limit.

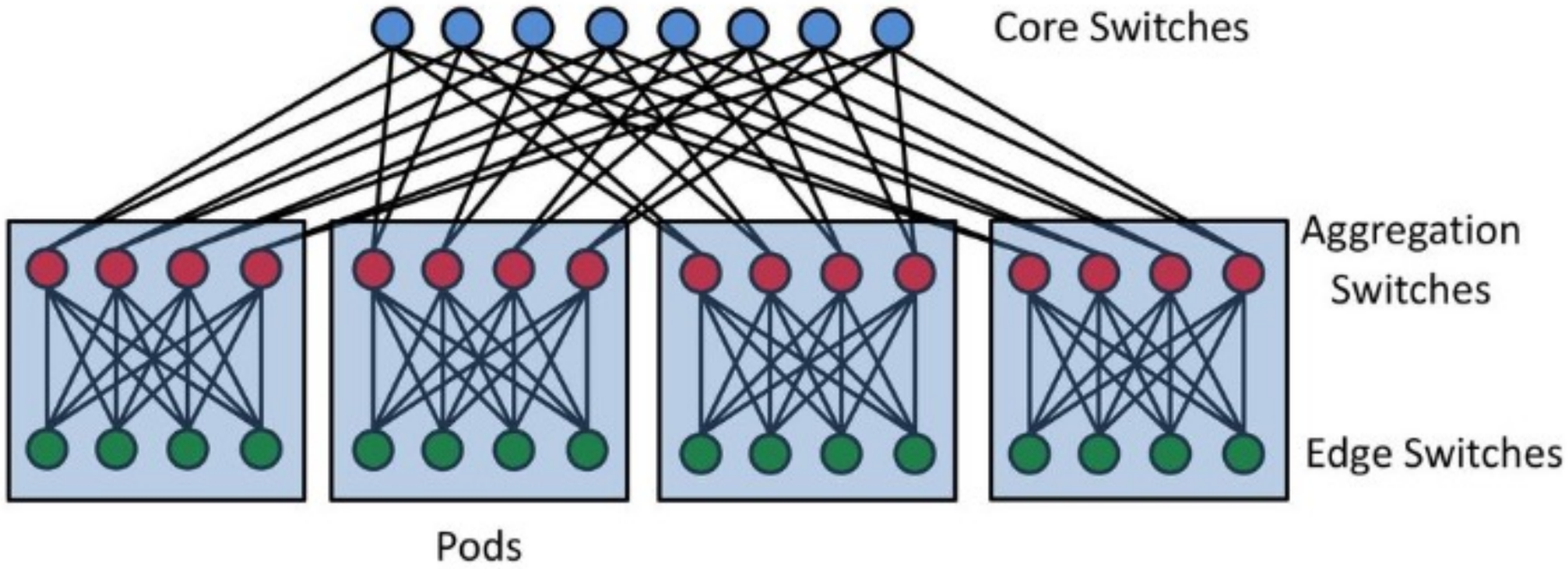
https://t.me/learningnets

# EIGRP in the Datacenter

- Large scale Datacenters mainly use CLOS (Leaf and Spine) topology, depends on scale, multi stage CLOS topologies are used

https://t.me/learningnets

# 3 stage CLOS topology




https://t.me/learningnets

- To create DC Fabric, EIGRP can be used in some Datacenters as a layer 3 fabric protocol, but as it is stated before, EIGRP is not common in the Datacenters as a L3 fabric protocol.

- EIGRP is used mainly on the WAN, classical example EIGRP over DMVPN.

https://t.me/learningnets

# EIGRP in the Service Provider Networks

- EIGRP is very uncommon in the Service Provider networks.
  - Due to lack of MPLS Traffic Engineering Support and multi-vendor interoperability issues, Service Providers use Standard based protocols.
- Service Providers are heterogeneous environment (Many vendors are used in the network) interoperability between the network equipments is critical.

https://t.me/learningnets

# EIGRP Design Best Practices

- Enable summarization in large scale deployment, it will provide scalability.

- Limit the EIGRP query scope, it will help for fast convergence, availability, troubleshooting and scalability.

https://t.me/learningnets

# EIGRP Design Best Practices

- Enable EIGRP Stub feature at the Spokes site in Hub and Spoke deployment, EIGRP stub routers will not be used as a transit node, it will help for fast convergence and scalability as well.

https://t.me/learningnets

# EIGRP Design Best Practices

- When you need to do traffic engineering, change delay attribute, not bandwidth, delay is cumulative, changing delay mat not have an impact for the other paths, also bandwidth is used for QoS, RSVP and other applications as well.

best practices

https://t.me/learningnets

# EIGRP Design Best Practices

- Don't lower the EIGRP fast hellos for faster failure detection, use BFD instead.

https://t.me/learningnets

# EIGRP Advantages and the Disadvantages

- EIGRP is considered more flexible than OSPF and IS-IS because it supports manual summarization in any interface at any router in the network.

disa
dva
ntag
es

- EIGRP, in addition to ECMP, support UCMP as well.

adva
ntag
es

https://t.me/learningnets

# EIGRP Advantages and the Disadvantages

- EIGRP supports Stub feature which protects Spoke routers to be a transit router, this is not easily possible with OSPF.

- EIGRP is better suitable in large scale Hub and Spoke design than OSPF and IS-IS.

disadvantages
advantages
https://t.me/learningnets

# EIGRP Advantages and the Disadvantages

- There is no topology information, end to end path may not be optimal.

- EIGRP is Cisco preparatory, vendor interoperability may be an issue.

- With Feasible Successor support, it converges faster than OSPF and IS-IS when there is no ECMP with them.

disad
vanta
ges

adva
ntag
es

https://t.me/learningnets

# EIGRP Advantages and the Disadvantages

•There is no topology information, EIGRP cannot be used in distributed or centralized traffic engineering.

disa
dva
ntag
es

- Well known protocol by many network engineers and commonly deployed in many Enterprise networks.

adva
ntag
es

https://t.me/learningnets

# EIGRP CASE STUDIES

https://t.me/learningnets

# MPLS VPN and DMVPN with EIGRP

- Enterprise company wants to use MPLS L3 VPN (Right one) as its primary path between their Remote office and the Datacenter.
  - Customer uses EIGRP and EIGRP AS 100 for the Local Area Network inside the office.
- They want to use their DMVPN network as a backup path.
  - Customer runs EIGRP AS 200 over DMVPN.

case study

https://t.me/learningnets

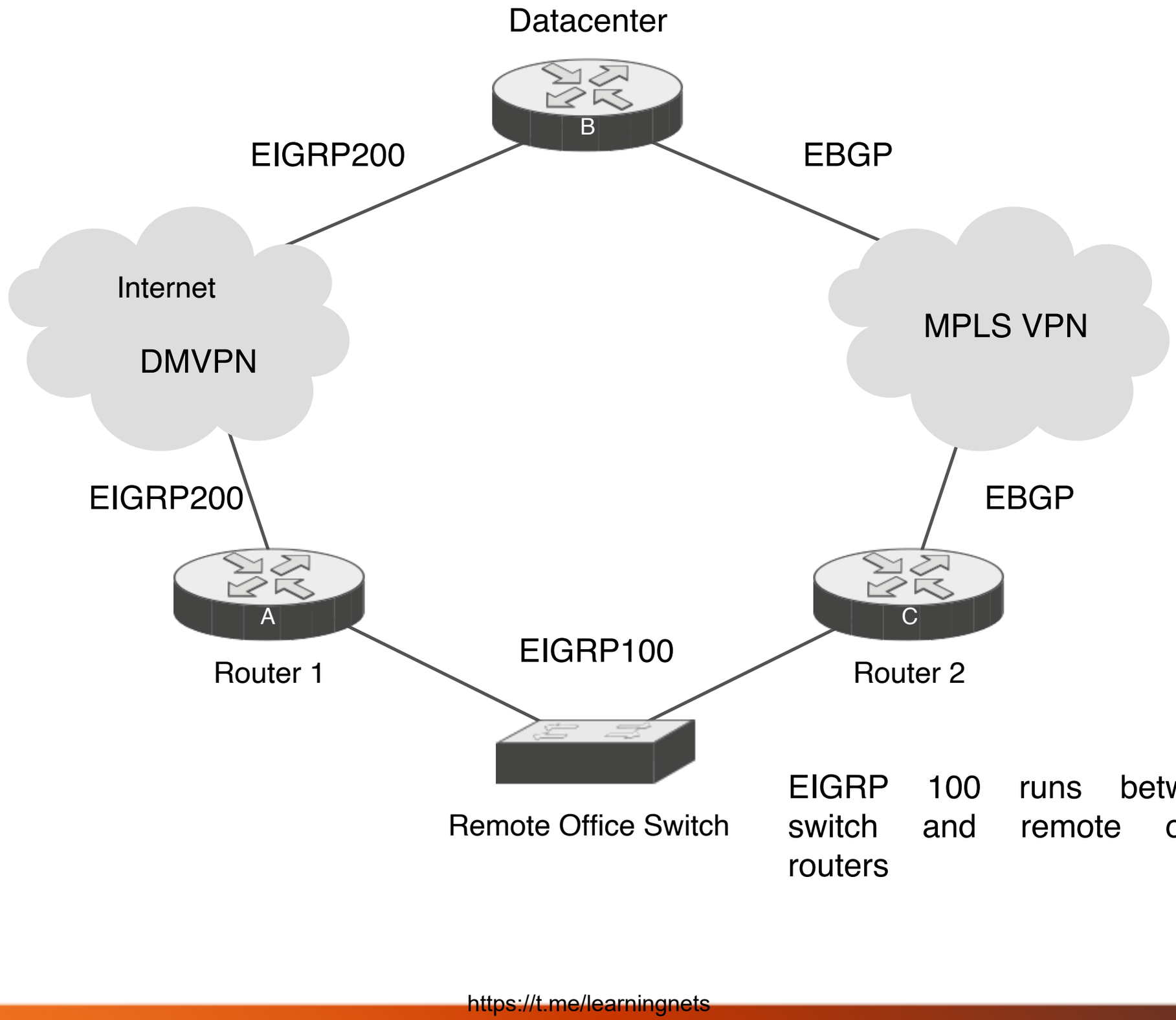
# MPLS VPN and DMVPN with EIGRP

- Service Provider doesn't support EIGRP as a PE-CE protocol, only static routing and BGP.
  - Customer selected to use BGP instead of static routing since cost community attribute might be used to carry EIGRP metric over the MP- BGP session of service provider.
- Redistribution is needed on the R2 between EIGRP and BGP (Two ways).
  - Since customer uses different EIGRP AS numbers for the LAN and DMVPN networks, redistribution is need on R1 too.

case study

https://t.me/learningnets

# Network Topology




EIGRP 100 runs between switch and remote office routers

https://t.me/learningnets

# Question 1

- Should customer use same EIGRP AS (Autonomous System) on the DMVPN network and the their office LAN? What is the problem with that design?

https://t.me/learningnets

## Answer 1: No. They shouldn't.

- Since Customer requirement is to use MPLS VPN as primary path, if the customer runs same EIGRP AS on Local Area Network and over DMVPN, EIGRP routes is seen as internal from DMVPN but external from MPLS VPN.
  - Internal EIGRP is preferred over external because of Admin Distance, customer should use different AS numbers.
- DMVPN could be used as a primary path for some delay, jitter, loss insensitive traffic but customer didn't specify that.

https://t.me/learningnets

# Question 2

- When company changed the EIGRP AS on DMVPN and started to use different EIGRP AS on the DMVPN and the LAN, which path is used between the Remote Offices and the datacenter?

https://t.me/learningnets

# Answer 2

- Since redistribution is done on R1 and R2, remote switch and datacenter devices see the routes both from DMVPN and BGP as EIGRP external. Then the metric is compared. If the metric (Bandwidth and Delay in EIGRP) is the same, both path can be used (Equal Cost Multipath-ECMP).

https://t.me/learningnets

# Question 3

- Does result fits for the customer traffic requirement?

https://t.me/learningnets

# Answer 3

- You should first remember what was the customer expectation for the links, they want to use MPLS VPN for all their application as primary path.

- That's why answer is Yes, it satisfy the customer requirement. Because if customer uses different EIGRP AS on LAN and DMVPN, with just metric adjustment, MPLS VPN path can be used as primary with just the metric arrangement.

https://t.me/learningnets

# Question 4

- What happens when the primary MPLS VPN link goes down?

https://t.me/learningnets

# Answer 4

- Traffic from remote office to the datacenter goes through Switch - R1 - DMVPN path. From the datacenter, since those will not be known through MPLS VPN when it fails, only DMVPN link is used.

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

https://t.me/learningnets

# Question 5

- What happens when failed MPLS VPN link comes back?

https://t.me/learningnets

# Answer 5

- R2 receives the datacenter prefixes over MPLS VPN path via EBGP, also from R1 via EIGRP. Once the link comes back, datacenter prefixes will still be received via DMVPN and MPLS VPN and appears on the office switch as an EIGRP external.

- Since metric was arranged previously to make MPLS VPN path as primary, no further action is required.

https://t.me/learningnets

# Answer 5

- This is tricky part. If it is Cisco switches or from other vendor which uses BGP weight attribute into consideration for the best path selection, then redistributed prefixes weight would be higher than the prefixes which are received through MPLS VPN so R2 uses Switch-R1 DMVPN path which violates the customer expectation.

https://t.me/learningnets

# EIGRP Stub and Caveats

- Haleo is a fictitious Enterprise company which uses DMVPN Wide Area Network technology over the Internet.
  - They have only 1 datacenter and the WAN routers which connect remote branches to the datacenter and the campus network are terminating many remote branches.
- For the redundancy and capacity purpose Haleo is using two DMVPN hub routers.
  - They know that EIGRP is best IGP for the large scale DMVPN design but recently they had some issues with their WAN network.

case study

https://t.me/learningnets

# Question 1

- Which additional information do you need from Haleo to address their issue?

1. Datacenter network topology
2. Which routing protocol Haleo is using
3. Encryption method of Haleo
4. Routing configuration
5. Wan Network Topology

https://t.me/learningnets

# Answer 1

- Routing configuration: We know that Haleo is using EIGRP but we don't know if they have EIGRP Stub on the edges , whether split horizon is disabled on the Hub, so we need to know their EIGRP features.

A

- Wan network topology: We need Wan network topology since we don't know how many routers in the branches, how they are connected, background information only mention that Haleo is using two hub routers.

https://t.me/learningnets

# Answer 1

- Lets look at the options

  - Datacenter Network topology: We don't need this information since in the background information, we are told that problem on the Wan network.

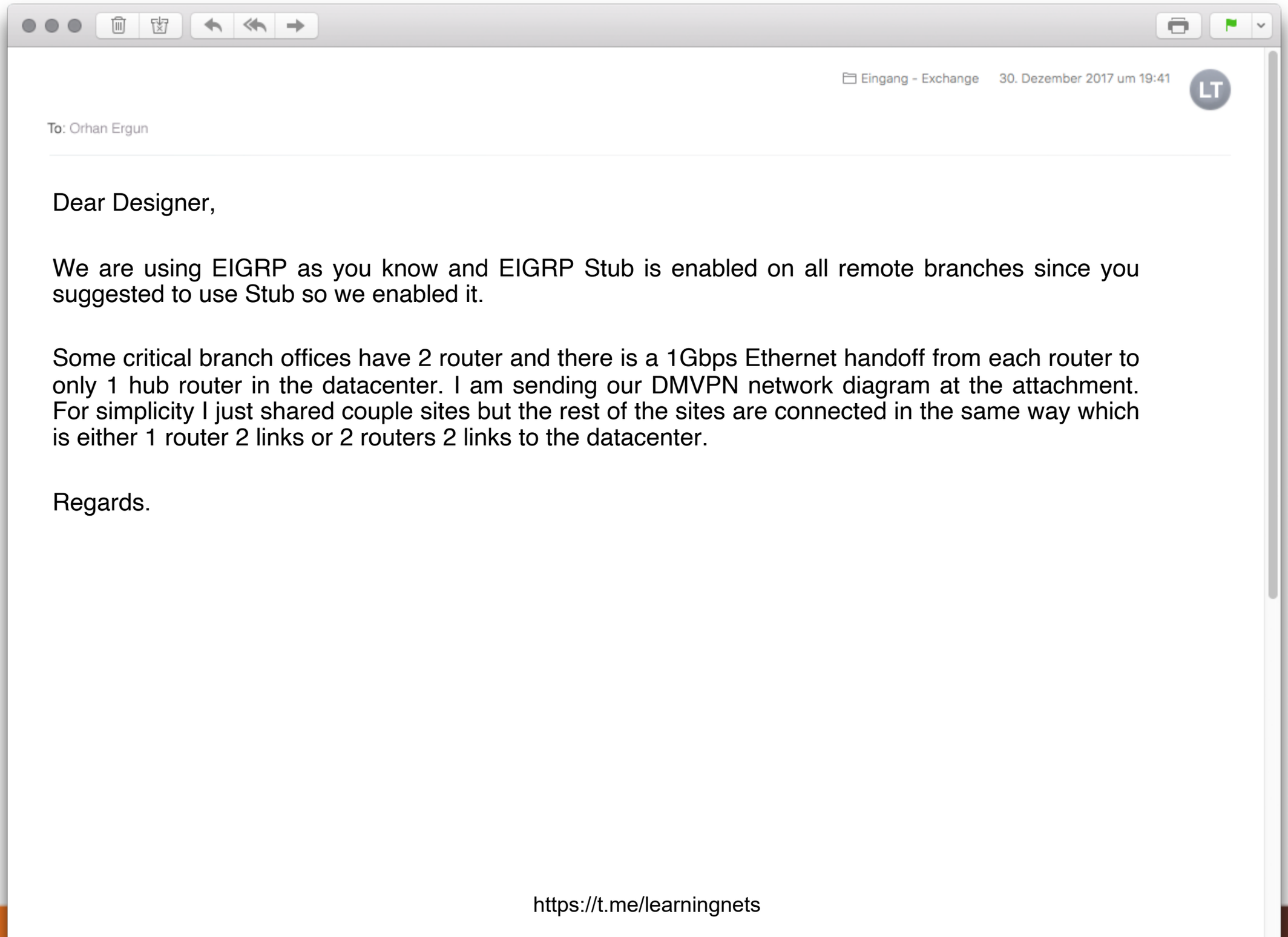
- Which routing protocol Haleo is using: We don't need it as well since in the background information it is already stated as EIGRP.

A

  - Encryption method: Nothing said about encryption, we can't assume that they are using encryption over DMVPN network since DMVPN doesn't require encryption.

https://t.me/learningnets

# Answer 1



Eingang - Exchange 30. Dezember 2017 um 19:41

LT

To: Orhan Ergun

Dear Designer,

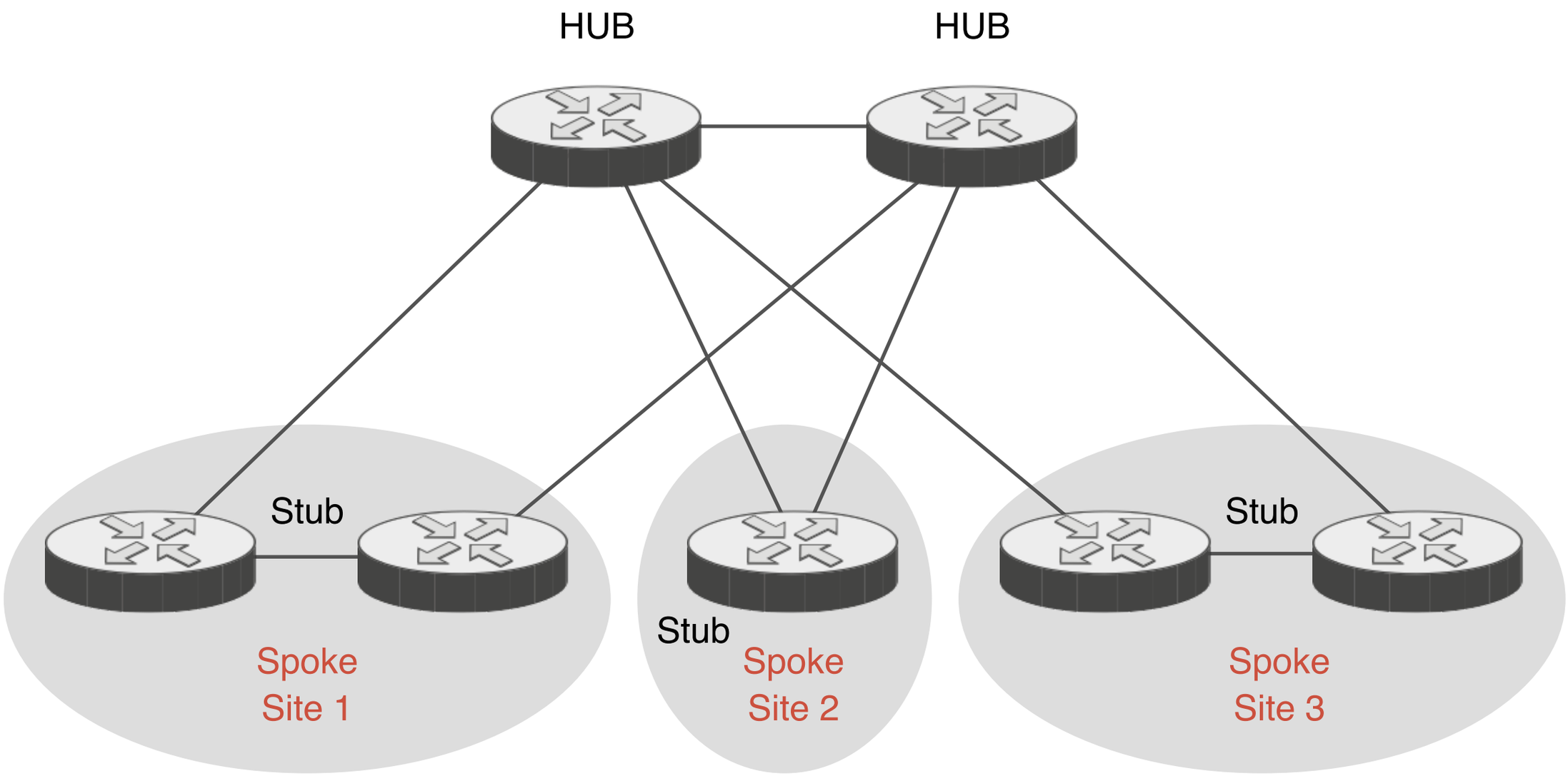
We are using EIGRP as you know and EIGRP Stub is enabled on all remote branches since you suggested to use Stub so we enabled it.

Some critical branch offices have 2 router and there is a 1Gbps Ethernet handoff from each router to only 1 hub router in the datacenter. I am sending our DMVPN network diagram at the attachment. For simplicity I just shared couple sites but the rest of the sites are connected in the same way which is either 1 router 2 links or 2 routers 2 links to the datacenter.

Regards.

https://t.me/learningnets

# Customer Topology

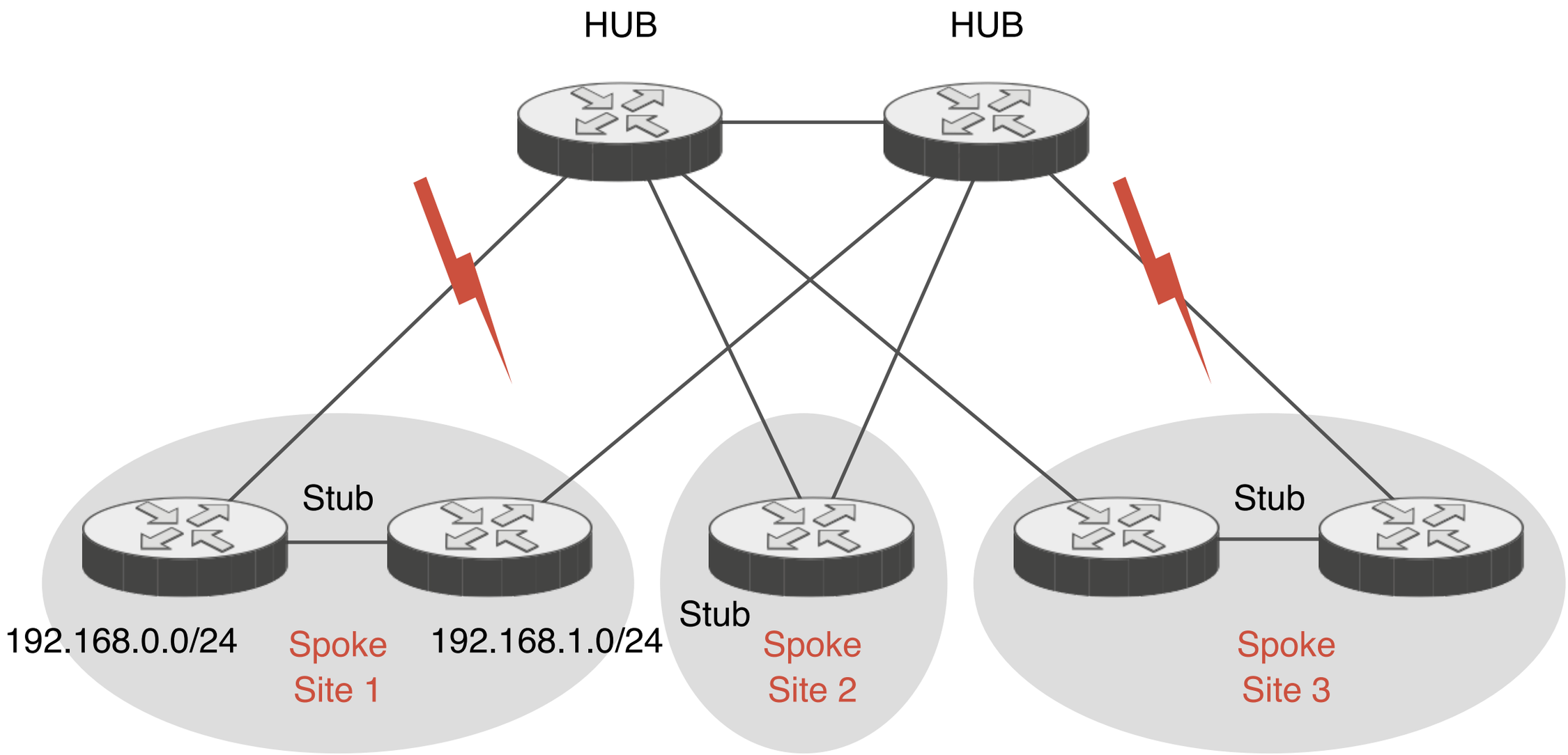



https://t.me/learningnets

# Question 2

- Based on the provided information what might be the problem? How it can be solved?

https://t.me/learningnets




https://t.me/learningnets

# Answer 2

- Since the spoke routers are running as EIGRP stub, they don't send the prefixes which are learned from each other to the Hubs.

  - That's why If the link between Hub and the Spoke sites which have two routers fails, router is isolated from the rest of the network.

- Spokes in the Spoke site 1 send their network to each other. So 192.168.0.0/24 and 192.168.1.0/24 learned by both spokes but since they are EIGRP Stub, they don't send the learned routes to the hub.

https://t.me/learningnets

# Answer 2

- If the Hub and Spoke link is failed in the Spoke Site 1, 192.168.0.0/24 network will not be reachable anymore.
  - Same think for the Spoke Site 3 since that site also has 2 routers and EIGRP stub is enabled.
- The solution is to enable EIGRP Stub leaking. In DMVPN it is good to send summary or default route to the Spokes by the Hubs.
  - Spoke should send the routes which they learn from each other to the Hub and also should send the routes which they learn from Hub to the each other. In this way sites which have more than 1 router which has EIGRP Stub configuration doesn't have an issue in case of any failure.

https://t.me/learningnets

# Race Condition and Routing Loop

- Enterprise company is using EIGRP on their network. They have MPLS VPN and also DMVPN network for their WAN network.
  - MPLS service is layer 3 VPN, managed by the Service Provider.
- Company has 30ms end to end latency SLA for their MPLS VPN network and all their application's fits to his service level agreement.
  - That's why they want to use MPLS VPN for their primary path and in case it fails, DMVPN should be used as backup.

case study

https://t.me/learningnets

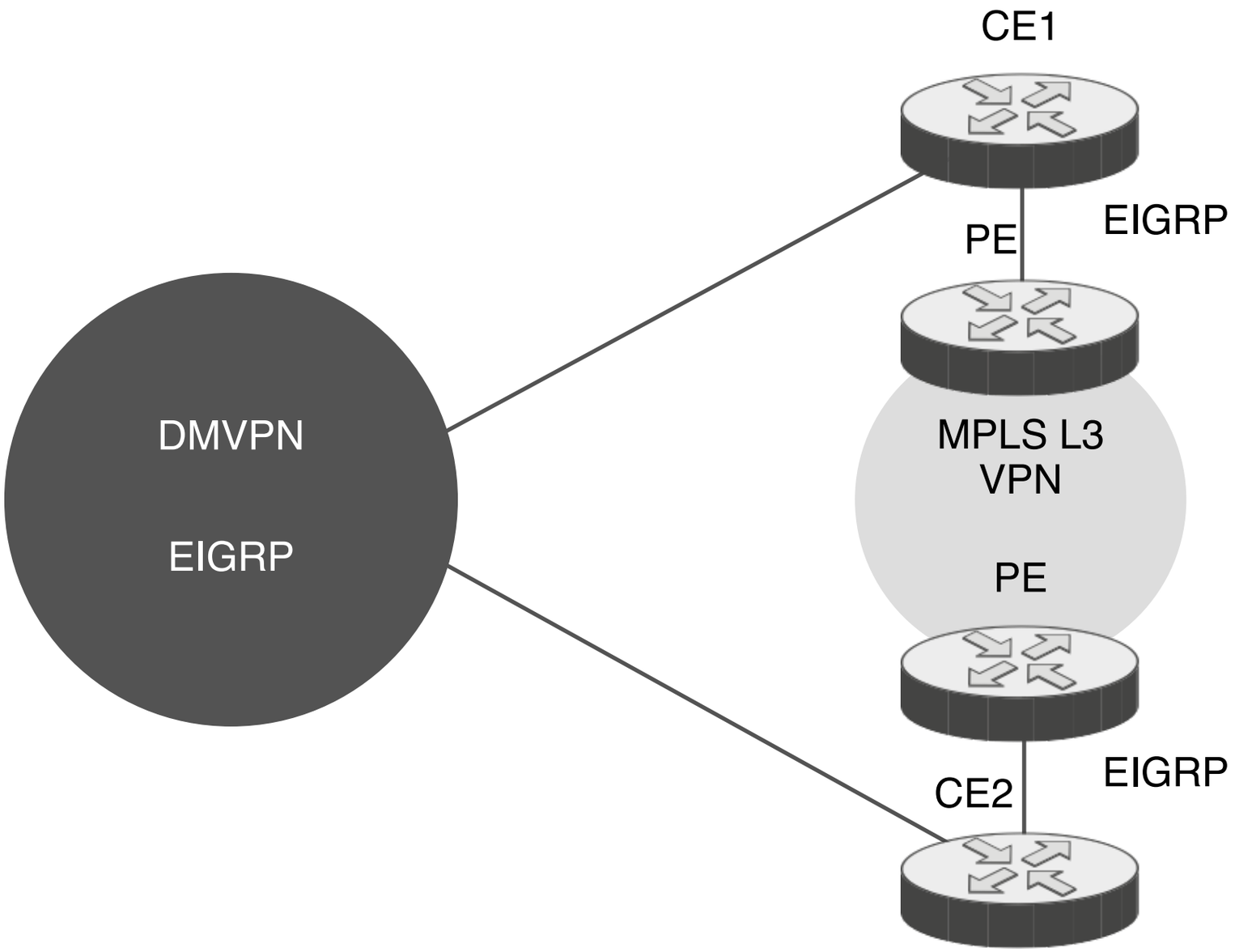
# Race Condition and Routing Loop

- Recently they had an outage and when they researched, they realized that the convergence of their EIGRP was too higher than their expectation.
  - They are asking your help to address the issue and provide a solution.
- They send their topology at the attachment.

case study

https://t.me/learningnets

# Network Topology



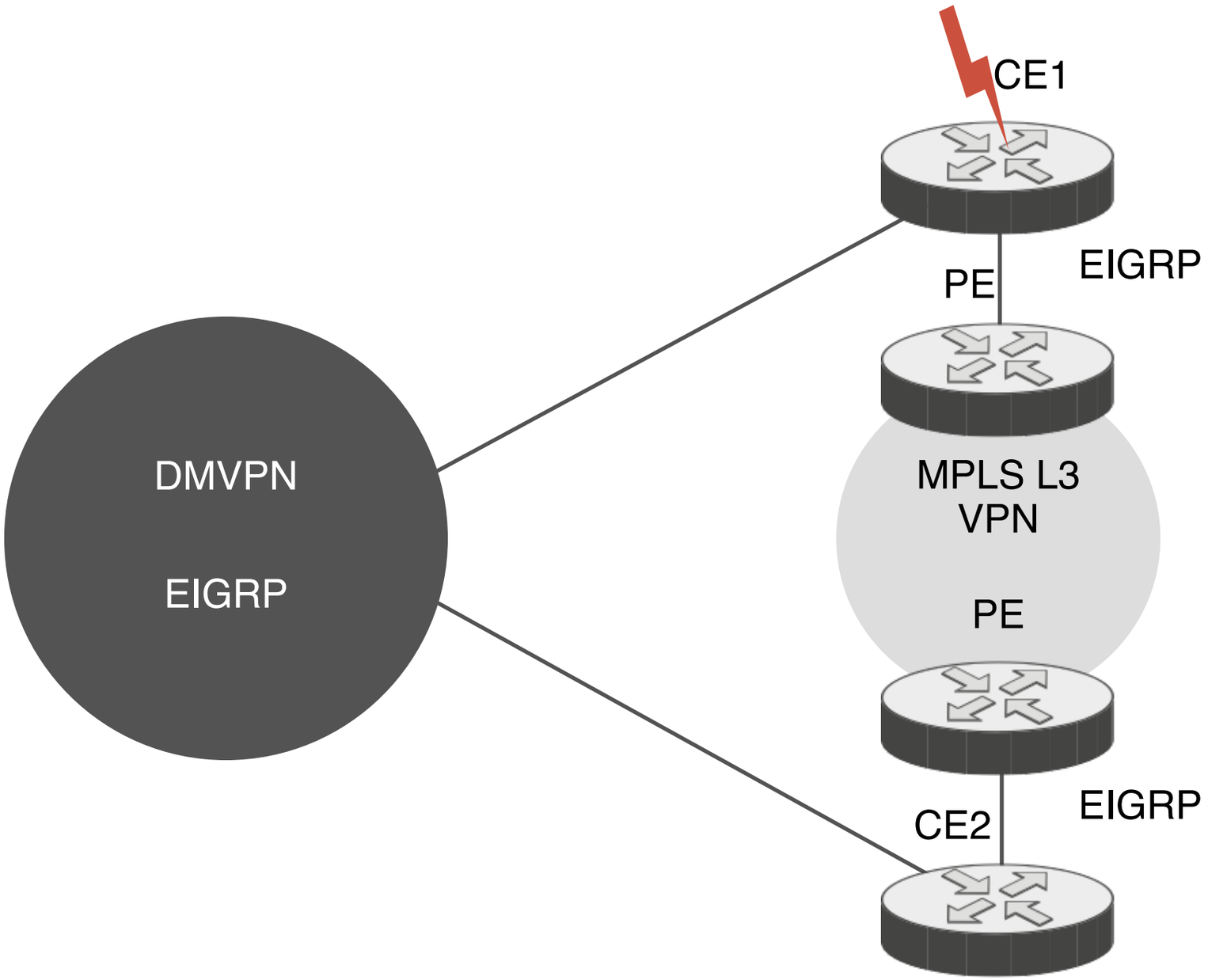

https://t.me/learningnets

# Race Condition and Routing Loop

- EIGRP because it is a distance vector protocol, suffers from two problems.
  - In case of failure, you will see either suboptimal routing or black hole situation, both due to the EIGRP race condition.
- In the topology below if the loopback interface of CE1 fails, depending on the timing of EIGRP and MP-BGP update, you might have routing loop.

case study

https://t.me/learningnets




https://t.me/learningnets

- CE1 sends an EIGRP queries to PE1 and CE2 asking it's loopback prefix. Network as we know from the scenario is using MPLS VPN as primary path.

  - That's why PE1 prefers the prefix via CE1, since it learns the prefix from the CE1 it assumes there are no alternate path.

- PE 1 doesn't send queries further and reply with prefix unreachable to CE1.

  - When PE1 stops learning CE'1 loopbacks from EIGRP, it removes it from its BGP table as well (EIGRP to MP-BGP redistribution).

case study

https://t.me/learningnets

- When PE1 removes the prefix from its BGP routing table it sends a BGP withdrawal to the PE2.

  - PE1 sent previously an EIGRP query to the CE2, and CE2 would propagate it.

- Here the EIGRP Race condition can be a big issue.

  - Depending on the arrival of EIGRP query and the BGP Withdrawal message to the BGP, persistent routing loop occurs.

case study

https://t.me/learningnets

- If BGP withdrawal over the MPLS VPN backbone via MP-BGP comes faster than EIGRP query through CE1 - CE2 and eventually to the PE2, everything is fine.

  - Because PE2 removes the prefix from BGP table and stops redistributing into the EIGRP (MP-BGP to EIGRP Redistribution).

- But if EIGRP query from the CE2 comes faster than BGP withdrawal to the PE2.

case study

https://t.me/learningnets

- Since PE2 has a path to the CE1 over the MPLS VPN, it replies to the CE2 that It has an alternate path, this reply goes up to the originator which is CE1 back and CE1 sends it to the PE1, in return PE1 sends it to the PE2.

  - CE1 now thinks that CE2 has an alternate path for its own loopback.

- Meanwhile PE2 receives the BGP withdraw which was previously sent by PE1.

  - PE2 sends an EIGRP query to CE2 so CE2 again things that it is not reachable so it sends a EIGRP query to CE1.

case study

https://t.me/learningnets

# EIGRP in the CCDE Exam

- EIGRP and VPN interaction, such as EIGRP, MPLS L3 VPN, DMVPN case study.
  - EIGRP and other routing protocols interaction, redistribution, extending EIGRP to the OSPF or IS-IS domains or vice-versa
- EIGRP Stub and other scalability features.

https://t.me/learningnets

# Summary

- EIGRP Theory – Distance Vector, No-topology, terminology
- EIGRP Fast convergence and Fast Reroute
- EIGRP Scalability – Query Scope, Summary, Filter and Stub
- EIGRP in the Datacenter and Service Provider Networks
- EIGRP Advantages – Disadvantages – Summarization at every hop, scalability, Cisco preparatory

https://t.me/learningnets

# EIGRP

## Enhanced Interior Gateway Routing Protocol

QUIZ

https://t.me/learningnets

## Question 1

Which below technology provides similar functionality with EIGRP Feasible Successor?

A. ISPF

B. Partial SPF

C. Loop Free Alternate Fast Reroute

D. OSPF Stub Areas

E. IS-IS Level 1 Domain

https://t.me/learningnets

## Answer 1

C. Loop Free Alternate Fast Reroute

Although EIGRP convergence was not explained in the EIGRP chapter, it is important to mention here. EIGRP Feasible successor is the backup path, which satisfies the feasibility condition.

Which means it should satisfy the EIGRP's loop free backup path condition.

There is no ISPF, Partial SPF, PRC or SPF in EIGRP. These algorithms are used in link state protocols.

Answer of this question is LFA FRR, which is one of the IP Fast Reroute mechanisms. IP FRR mechanisms will be explained in the MPLS traffic engineering section later in the book.

https://t.me/learningnets

## Question 2

How many levels of hierarchy is supported in EIGRP?

A. One

B. Two

C. Three

D. Unlimited

https://t.me/learningnets

# Answer 2

D.Unlimited

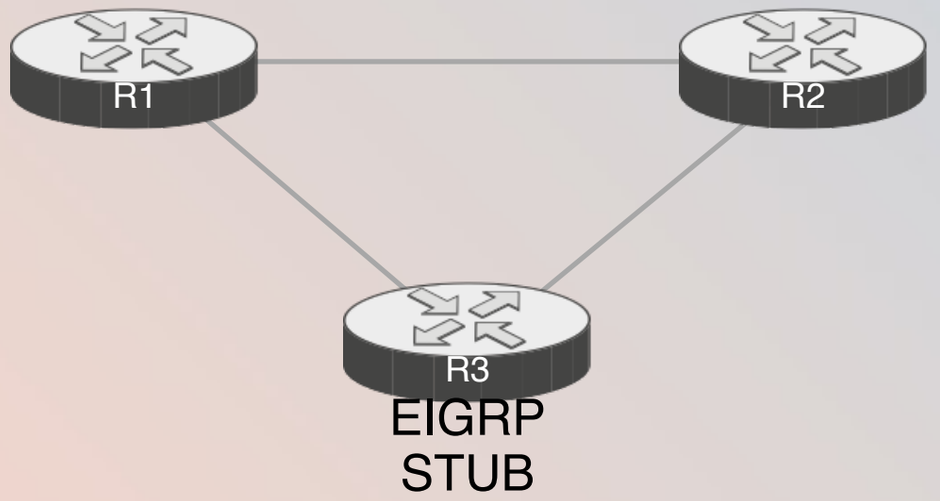
Unlike OSPF and IS-IS, there is no limit in EIGRP. OSPF and IS-IS support two levels of hierarchy as it was explained earlier.

There is no topology information in EIGRP and summarization can be done anywhere in EIGRP. Unlimited level of hierarchy is possible with EIGRP, that's why answer of this question is D.

https://t.me/learningnets

## Question 3

In the below topology R3 is configured as EIGRP Stub. If the link between R1 and R2 fails, which below statements are true for the below topology? (Choose Two)




A. Since R3 is configured as EIGRP Stub, R3 is not used as transit for the traffic between R1 and R2.

B. R1 can reach the networks behind R2 through R3.

C. R3 has all the R1 and R2 prefixes in its routing table.

D. R3 doesn't have R1 and R2 prefixes.

E. R2 can reach the networks behind R1 through R3.

https://t.me/learningnets

# Answer 3

A.Since R3 is configured as EIGRP Stub, R3 is not used as transit for the traffic between R1 and R2.

C.R3 has all the R1 and R2 prefixes in its routing table.

As it was explained in the EIGRP Stub section in this chapter, when node is configured as EIGRP stub, it is not used as transit node anymore.

Question is asking when the R1 and R2 link fails, whether R3 will be transit node. No, it will not be transit node.

Which mean, R1 cannot reach R2 through R3 and R2 cannot reach R1 through R3. That's why B and E are incorrect.

R3 has all the prefixes of R1 and R2 even if it is configured as EIGRP Stub.

That's why the correct answer of this question is A and C.

https://t.me/learningnets

## Question 4

Which below option is considered as loop free path in EIGRP?

A. If reported distance is less than feasible distance.

B. If reported distance is same as the feasible distance.

C. If reported distance is higher than feasible distance.

D. If administrative distance is higher than feasible distance.

https://t.me/learningnets

# Answer 4

A. If reported distance is less than feasible distance.

In order a path to be chosen as loop free alternate which means satisfy the EIGRP feasibility condition as it was explain in the EIGRP chapter of the book, reported distance has to be less than feasible distance. That’s why the answer of the question is A.

https://t.me/learningnets

## Question 5

What happens if the backup path satisfies the feasibility condition? (Choose Two)

A. It is placed in link state database.

B. It is advertised to the neighbors.

C. It is placed in the topology table.

D. It can be used as unequal cost path.

E. It is placed in the routing table.

https://t.me/learningnets

## Answer 5

C. It is placed in the topology table.

D. It can be used as unequal cost path.

EIGRP database is called Topology database. Link state database is used in link state protocols.

If backup path satisfies feasibility condition, it is placed in topology table, not in routing table. If it would be best path (successor) or equal cost path, it would be placed in routing table. But since question says, backup path, it is only placed in EIGRP topology database.

Since it is not the best path, it is not advertised to the neighbors.

With ‘variance’ command, it can be used as unequal cost path and can be placed in the routing table.

https://t.me/learningnets

## Question 6

Which below statements are true for EIGRP Summarization? (Choose Two)

A. EIGRP Auto-summarization is on by default for all the Internal and External routes.

B. EIGRP Route summarization can reduce the query domain which helps for convergence.

C. EIGRP Route Summarization can reduce the query domain which can prevent Stuck in Active problem.

D. Summarization cannot be done at each hop in EIGRP.

https://t.me/learningnets

# Answer 6

B. EIGRP Route summarization can reduce the query domain which helps for convergence.

C. EIGRP Route Summarization can reduce the query domain which can prevent Stuck in Active problem.

Summarization can be done at each hop in EIGRP. This is different than OSPF and IS-IS. Auto-Summarization is not enabled for all the routes by default in EIGRP. Summarization helps to reduce query domain boundary, which in turn help for convergence, SIA problem, troubleshooting and so on.

https://t.me/learningnets

## Question 7

Which below statement is true for EIGRP queries? (Choose Two)

A. EIGRP queries always send.

B. Limiting EIGRP query domain helps for scalability.

C. If summarization is configured, EIGRP query is not sent.

D. If filtering is configured, EIGRP query is not sent.

E. If EIGRP Stub is configured, EIGRP query is not sent.

https://t.me/learningnets

# Answer 7

B. Limiting EIGRP query domain helps for scalability.

E. If EIGRP Stub is configured, EIGRP query is not sent.

If EIGRP Stub is configured, as it was explained before, EIGRP query is not sent. With summarization and filtering still EIGRP query is sent. EIGRP query domain size affects scalability. If the query domain size is reduced, scalability increases.

https://t.me/learningnets

## Question 8

Why passive interface should be enabled on the access/customer ports?

A. To prevent injecting the customer prefixes to the network.

B. To reduce the size of the routing table.

C. For the fast convergence.

D. For higher availability.

https://t.me/learningnets

## Answer 8

A. To prevent injecting the customer prefixes to the network.

Passive interface should be used on all hosts, access and customer ports. Otherwise security attack can happen and prefixes can be injected into the routing domain. It doesn't provide faster convergence. And the reason to disable routing protocols on the customer/access ports is not to reduce routing table size.

https://t.me/learningnets

## Question 9

If the path in the network will be changed by changing the EIGRP attribute, which below statement would you recommend as a network designer?

A. Bandwidth should be changed.

B. Delay should be changed.

C. Reliability should be changed.

D. PBR should be configured.

https://t.me/learningnets

## Answer 9

B. Delay should be changed.

PBR is not an EIGRP attribute. Reliability is not used for EIGRP path selection. Bandwidth and Delay attributes are used for EIGRP path selection and metric is calculated based on these two parameters.

But, since bandwidth can be used by many applications such as QoS, RSVP-TE and so on it should be changed, otherwise other things in the network can change too.

Also since the minimum bandwidth is used for path calculation, changing bandwidth can affect entire network design. Not only the path, which we want.

On the other hand, delay is additive and changing it can only affect the path, which we want.

https://t.me/learningnets

## Question 10

When EIGRP is used as MPLS VPN PE-CE routing protocol, which below mechanism helps for loop prevention even if there is a backdoor link?

A. Up/Down bit

B. Sham link

C. Site of Origin

D. Split Horizon

https://t.me/learningnets

## Answer 10

C. Site of Origin

EIGRP Site of Origin is used to prevent loop even if there is a backdoor link. Backdoor link causes race condition in MPLS VPN topologies and it can create sub optimal routing and routing loop.

It will be explained in the MPLS VPN section in the MPLS chapter in detail.

https://t.me/learningnets

# EIGRP – Study Resources:

- Books :
- http://www.amazon.com/EIGRP-Network-Design-Solutions-Definitive/dp/1578701651/ref=sr_1_1?ie=UTF8&qid=1436565482&sr=8-1&keywords=eigrp
- Videos :
- Ciscolive Session – BRKRST -2336
- Podcast :
- http://packetpushers.net/show-144-open-eigrp-with-russ-white-ciscos-donnie-savage/
- Articles :
- http://www.cisco.com/c/en/us/td/docs/ios/12_0s/feature/guide/eigrpstb.html
- http://www.cisco.com/c/en/us/td/docs/ios xml/ios/iproute_eigrp/configuration/xe-3s/ire-xe-3s-book/ire- ipfrr.html

https://t.me/learningnets

# BGP

## Border Gateway Protocol

https://t.me/learningnets

# BGP Course Outline

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

https://t.me/learningnets

# BGP Course Outline

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

https://t.me/learningnets

# BGP Course Outline

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

https://t.me/learningnets

# BGP Course Outline

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

outline

https://t.me/learningnets

# BGP Course Outline

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

outline

https://t.me/learningnets

## Why BGP ?

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

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

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

https://t.me/learningnets

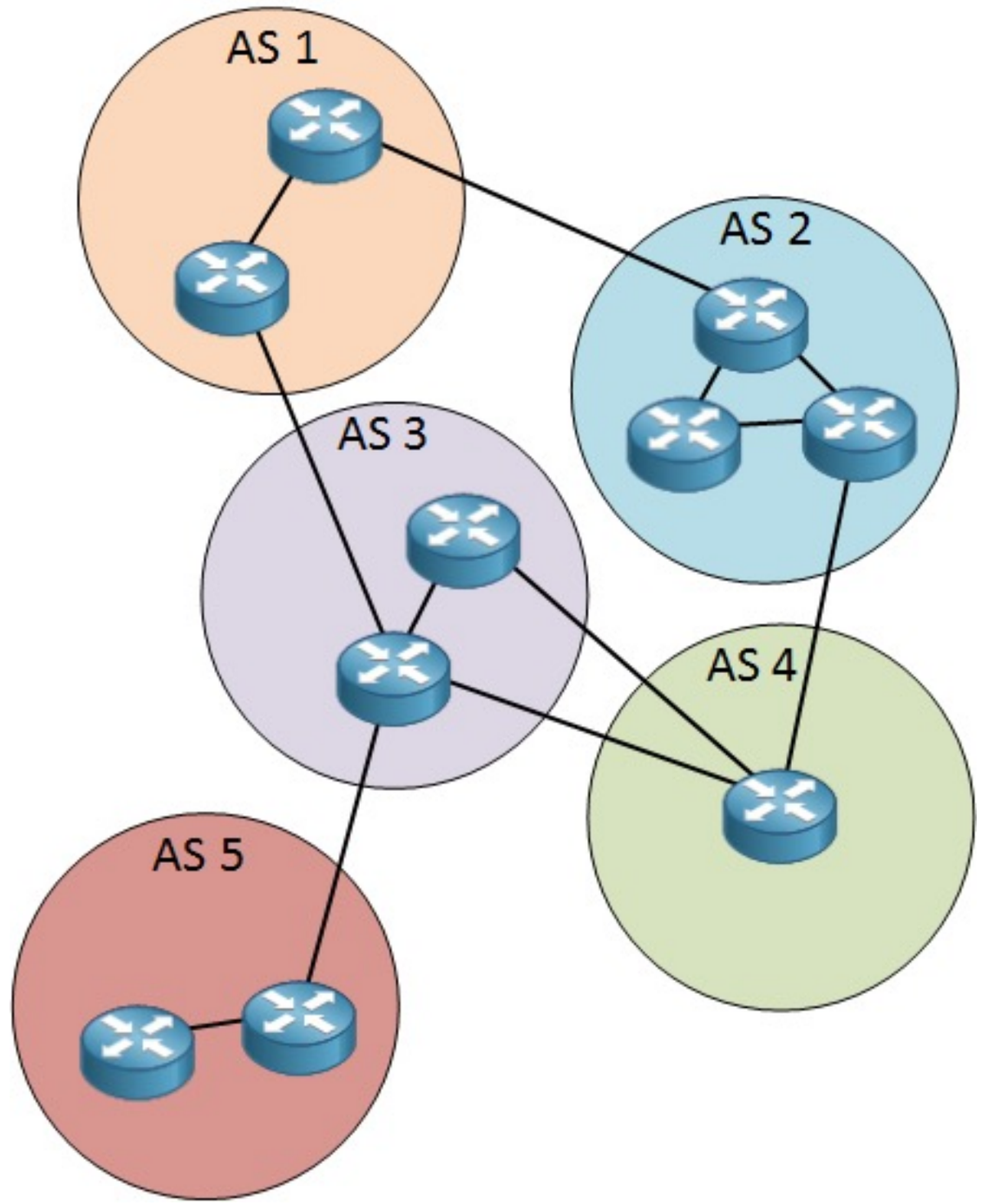
# BGP – Border Gateway Protocol Basics

## Autonomous System

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

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

https://t.me/learningnets




https://t.me/learningnets

# BGP – Border Gateway Protocol Basics

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

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

https://t.me/learningnets

# BGP Best Path Selection

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

- BGP is a policy based protocols which provides IP based Traffic Engineering inside an Autonomous System. In fact IGPs don't support traffic engineering like the BGP does

https://t.me/learningnets

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

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

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

https://t.me/learningnets

# BGP best path selection steps

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

https://t.me/learningnets

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

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

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

https://t.me/learningnets

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

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

https://t.me/learningnets

# BMP - BGP Monitoring Protocol

- BMP is defined in RFC 7854

- It is used to monitor BGP sessions

- Until BMP, information about BGP session was received with CLI which can be a CPU intensive

- BMP is an automated way of collecting the BGP data from the routers

https://t.me/learningnets

# BMP - BGP Monitoring Protocol

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

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

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

https://t.me/learningnets

# BMP - BGP Monitoring Protocol

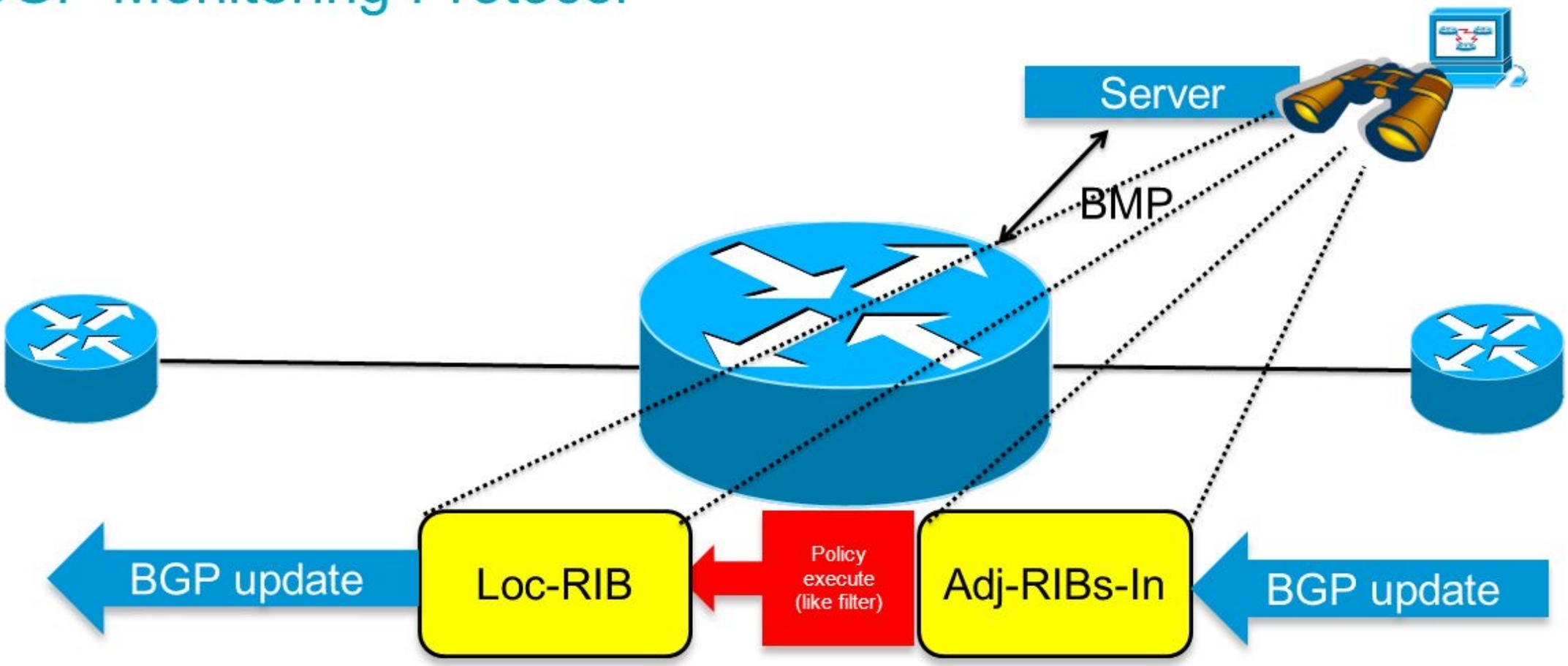
- BMP provides an access to the Adjacency-RIB-In database of router

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

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

https://t.me/learningnets

# BGP Monitoring Protocol




© 2013-2014 Cisco and/or its affiliates. All rights reserved.



https://t.me/learningnets

# BMP - BGP Monitoring Protocol

- BMP operates over TCP

- When a TCP connection is established, BMP messages are being sent from the BMP client to a BMP collector

- No BMP message is ever sent from the collector to the client

https://t.me/learningnets

# What you cannot do with BMP?

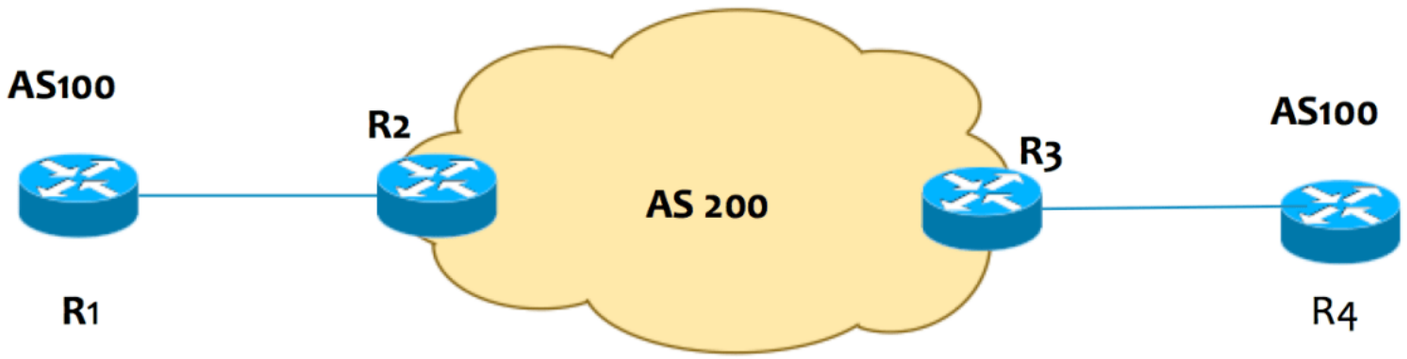
- You can’t monitor outgoing routes (Adj-RIB-Out)
- You can’t monitor the best BGP routes (Loc-RIB)
- You can’t monitor why prefixes were rejected (Post-Policy routes)
- You can’t see why routes didn’t win the best path selection
- There are two drafts in IETF as BMP extension, one is for Outgoing routes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

https://t.me/learningnets

# EBGP

- EBGP is used between two different Autonomous Systems, loop prevention in EBGP is done by the AS Path attribute, that's why it is a mandatory BGP attribute

- If BGP node sees its own AS Path in the incoming BGP update message, BGP message is rejected




https://t.me/learningnets

# EBGP

## EBGP Traffic Engineering

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

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

- In all these cases , customer may want to optimize their incoming and outgoing traffic

https://t.me/learningnets

## EBGP Traffic Engineering

- Network traffic flows in two directions ; Incoming and outgoing

- Incoming traffic engineering refers receiving traffic into the Local Autonomous System from one of the many available paths or receiving specific application/services traffic from any path

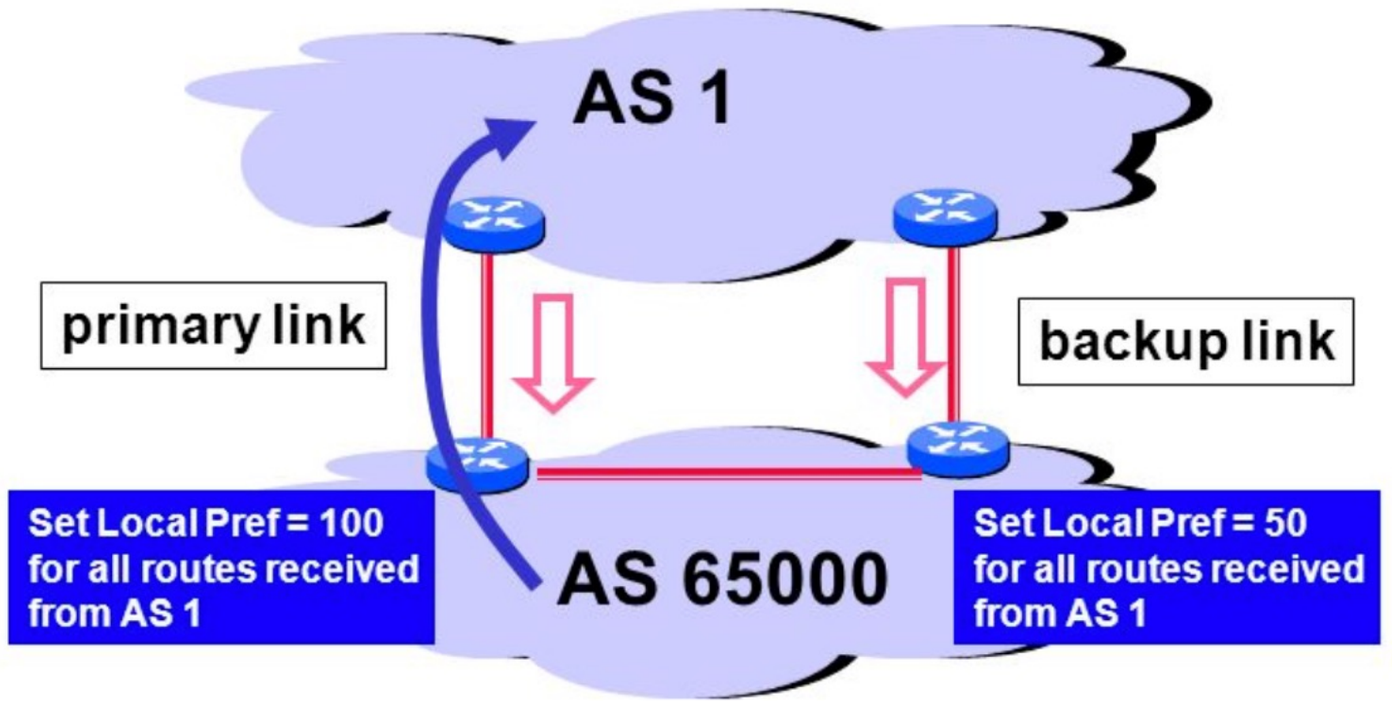
https://t.me/learningnets

# EBGP Traffic Engineering

- Outbound Traffic Engineering: Refers sending the traffic from Local AS to the other Autonomous Systems from one of the many paths or sending specific application/services traffic to other AS from any path

- For the BGP outgoing traffic, commonly,  local preference attribute is used

https://t.me/learningnets




AS 65000 has two paths to AS1, by increasing Local Preference on one of the links, AS 65000 sends all outbound traffic from the AS over that path

https://t.me/learningnets

# EBGP Outbound Traffic Engineering

- BGP weight attribute can be used for the outgoing traffic engineering as well but don’t forget that it is local to the router which mean is not propagated between the IBGP neighbors and it is Cisco preparatory, no vendor interoperability

https://t.me/learningnets

# GAO – Rexford Model

- Service Providers chooses to send the traffic for their customer prefixes over the customer link first, then peering links, lastly through upstream ISP. Because they want to utilize the customer link as much as possible to charge more money

- Utilizing Customer link can generate revenue , Settlement Free Peering Link or Upstream ISP link usage cost money, thus technical solution provided by GAO and Rexford is used to realize this business requirement

https://t.me/learningnets

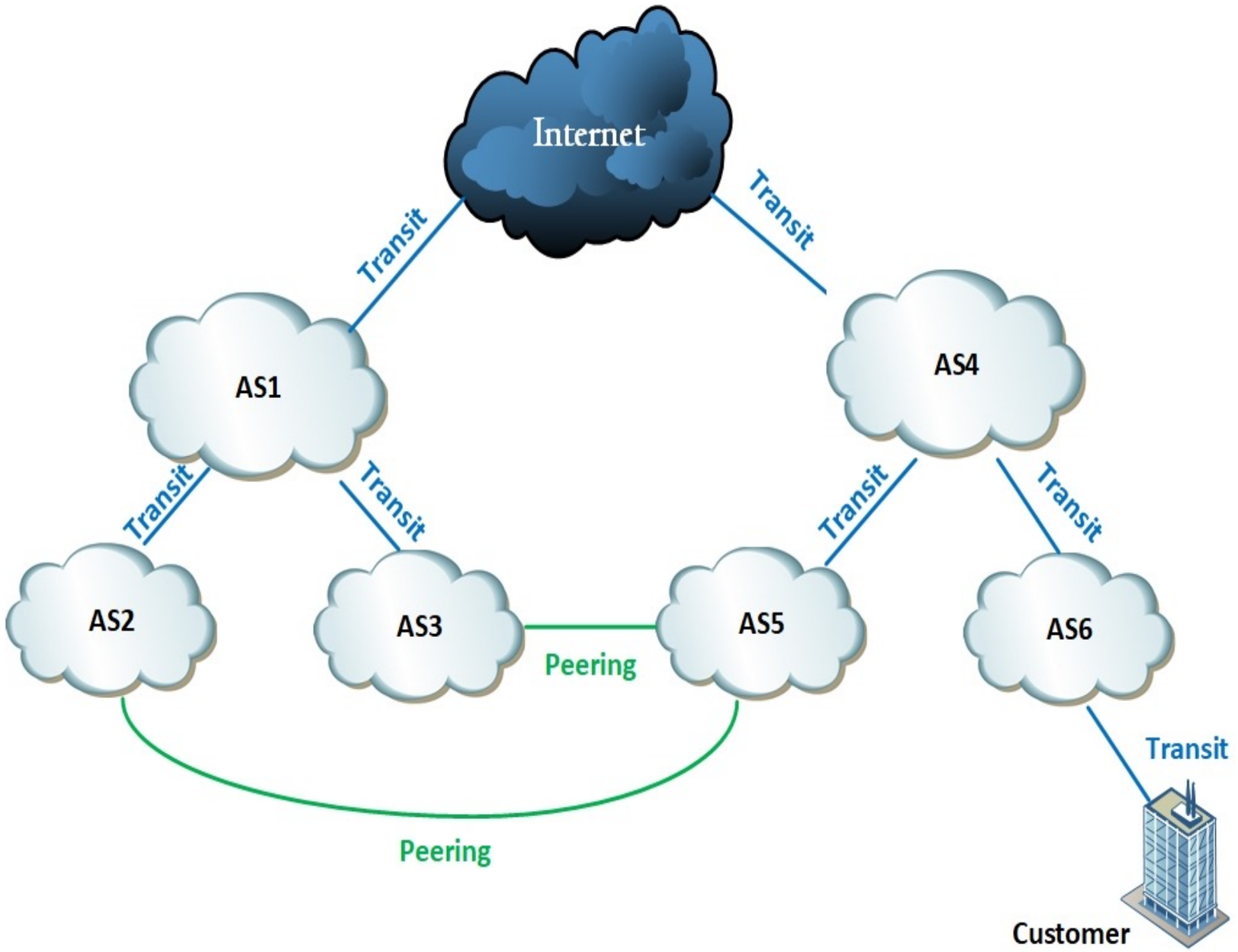
# GAO – Rexford Model

- Service Providers may implement Local Preference attribute to achieve this

- Basic local preference policy could be; Local Preference 100 towards Customer, local Preference 90 towards peering link and Local Preference 80 towards upstream ISP

https://t.me/learningnets

# GAO – Rexford Model




https://t.me/learningnets

## EBGP Inbound Traffic Engineering

- **BGP Inbound traffic engineering can be achieved in multiple ways:**

1. MED ( BGP External metric attribute )
2. AS-Path prepending
3. BGP Community attribute

https://t.me/learningnets

## EBGP Inbound Traffic Engineering with MED

- Creating an inter domain policy with the MED attribute is not a good practice

- MED attribute is used between two Autonomous system. If the same prefix is coming from two different AS to the 3rd AS, although you can use always-compare MED feature, it is not good practice to enable this feature since it can cause BGP MED Oscillation problem

https://t.me/learningnets

## Don’t compare BGP MED when the prefixes are received from two different AS !

- As per RFC 4451 – BGP MED Considerations : BGP speakers often derive MED values by obtaining the IGP metric associated with reaching a given BGP NEXT_HOP within the local AS. This allows MEDs to reasonably reflect IGP topologies when advertising routes to peers. While this is fine when comparing MEDs between multiple paths learned from a single AS, it can result in potentially "weighted" decisions when comparing MEDs between different autonomous systems.

- This is most typically the case when the autonomous systems use different mechanisms to derive IGP metrics or BGP MEDs, or when they perhaps even use different IGP protocols with vastly contrasting metric spaces (e.g., OSPF vs. traditional metric space in IS-IS)

https://t.me/learningnets

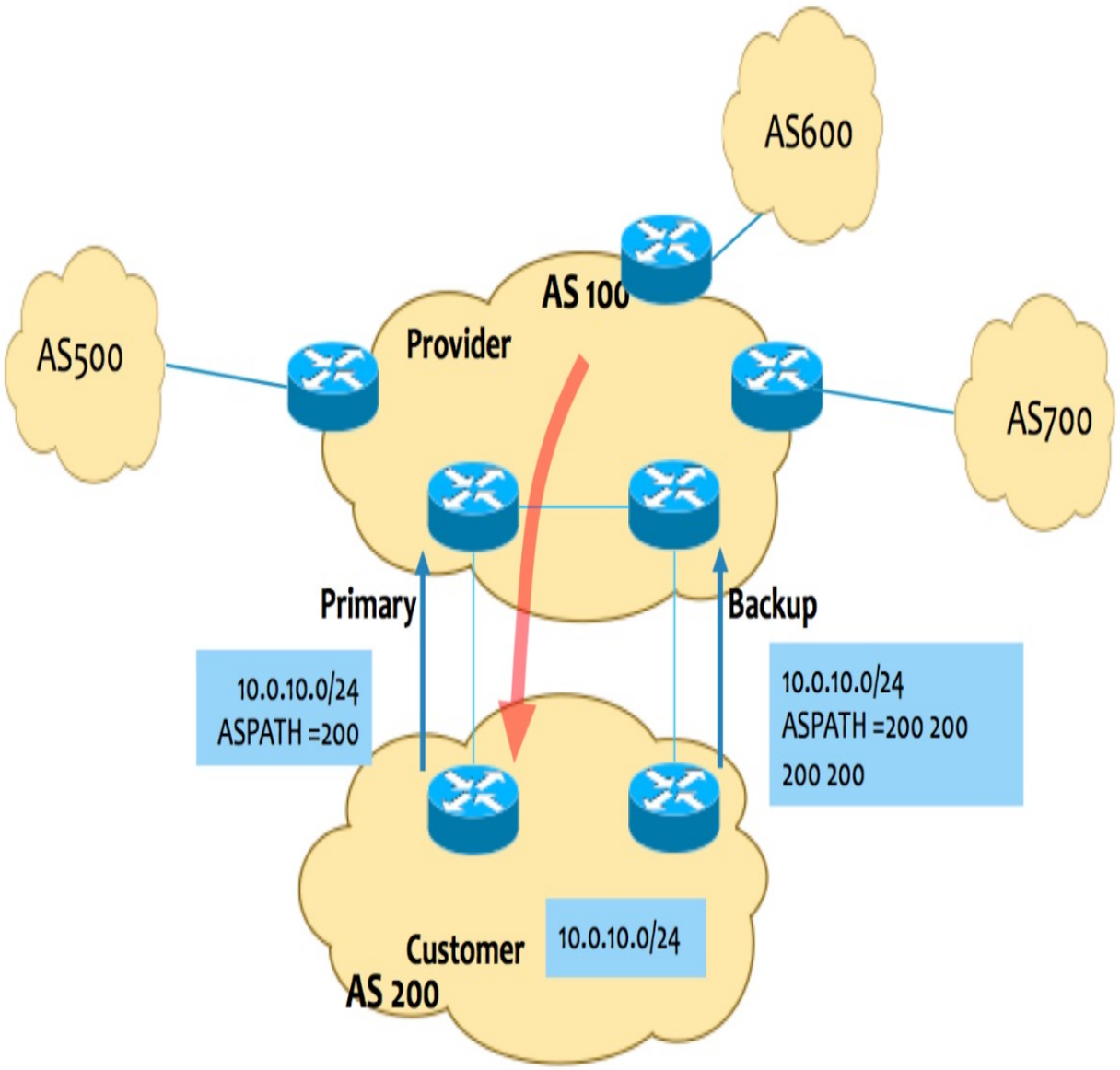
## EBGP Inbound Traffic Engineering with AS-Path Prepending

- BGP As-path is a mandatory BGP attribute which has to be sent in every BGP message. BGP as-path prepending is one of the BGP traffic engineering methods

- BGP As-path prepending is used to influence inbound traffic to the company. BGP As-path prepending is used in active-standby link scenarios. When there are two BGP neighborship which prefix will be advertised, one link for set of prefixes or maybe all the prefixes can be used as backup. In this case, one way to achieve this setup is using BGP AS-path prepending.

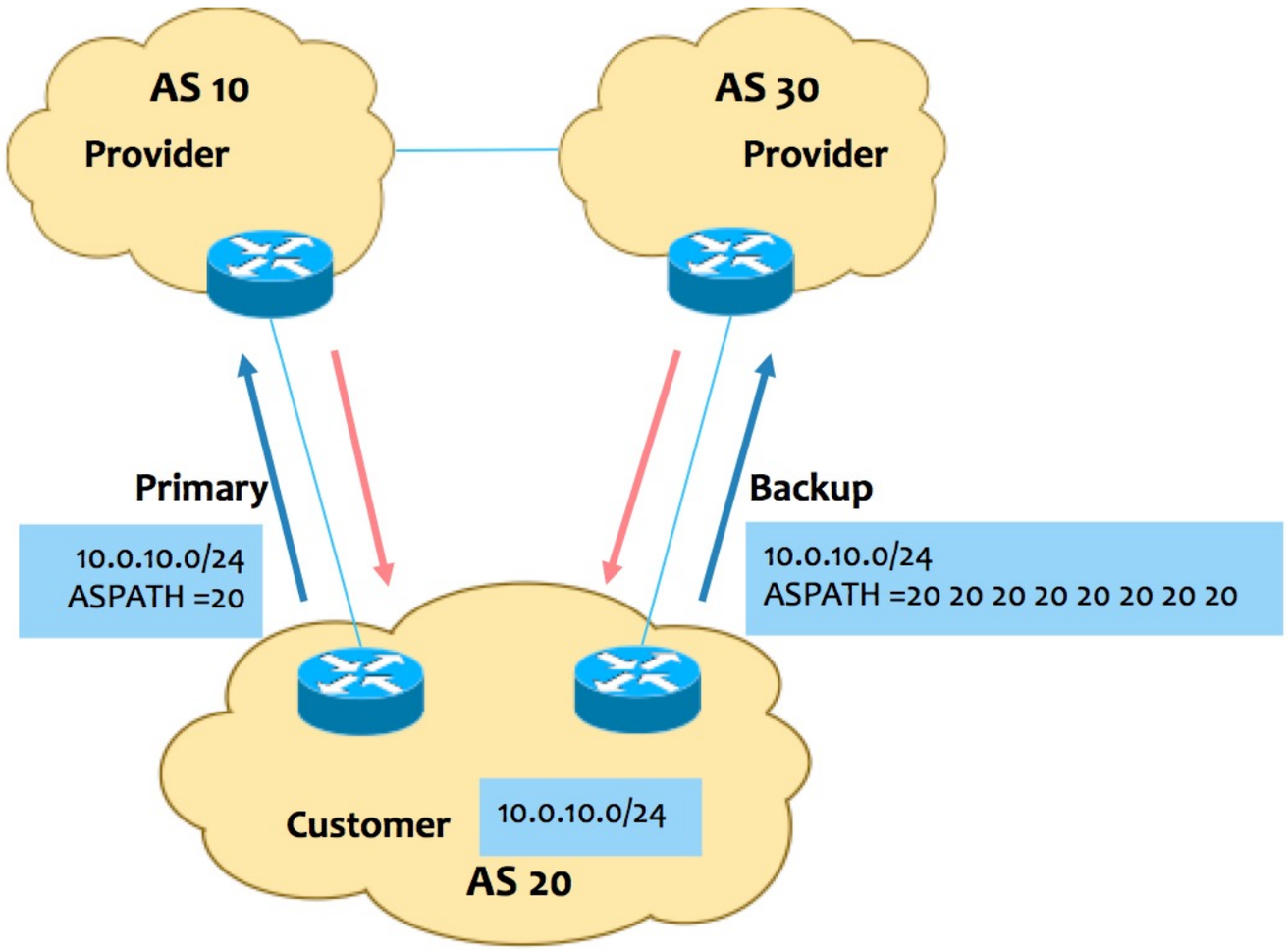
https://t.me/learningnets

# EBGP Inbound Traffic Engineering with AS-Path Prepending




Customer AS 200 wants to use of the links as backup. 10.0.10.0/24 prefix is sent via backup link with the 3 prepend. Thus AS path is seen through the backup link by the upstream service provider which is AS 100 as ‘ 200 200 200 200 ‘

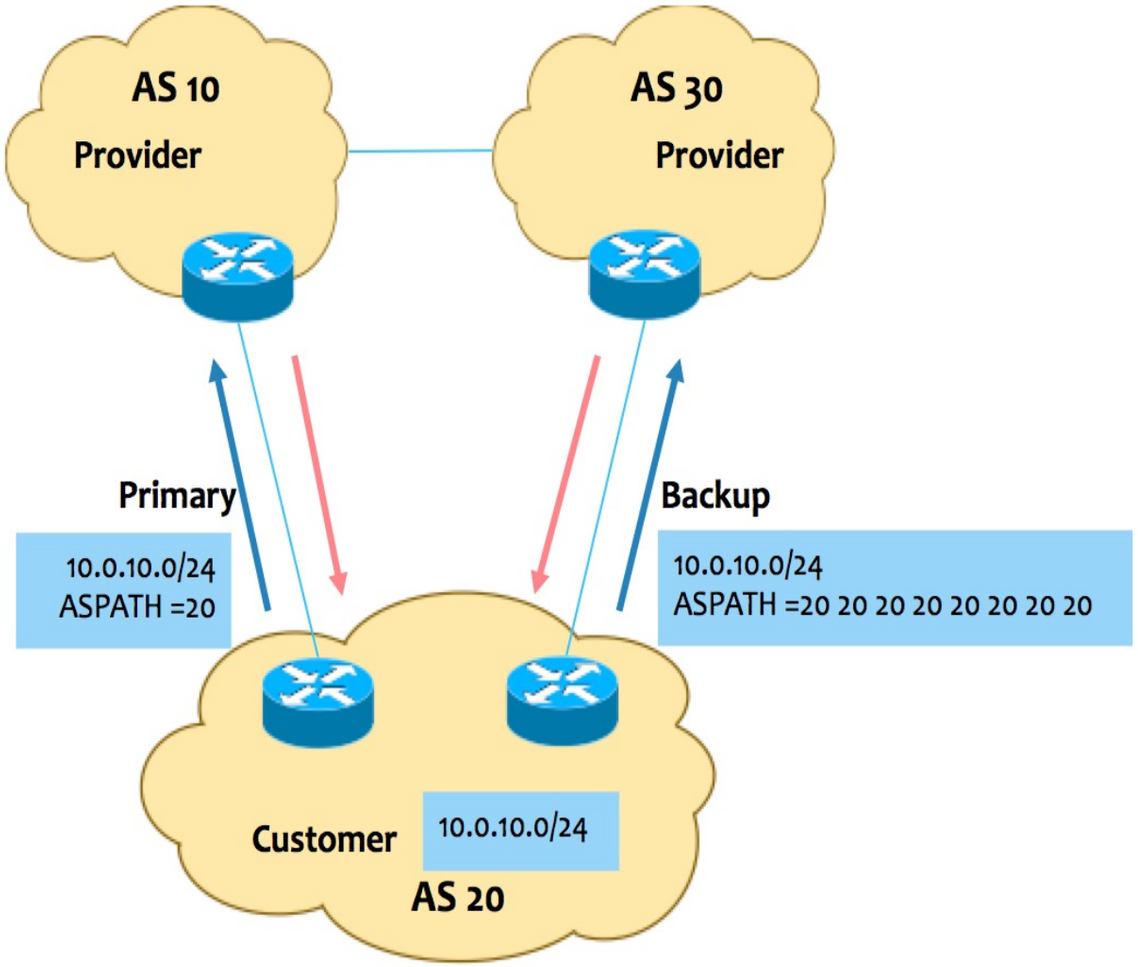
https://t.me/learningnets




There are some challenges with BGP as-path prepending when it is used in multi-homed BGP setup

- Customer AS 20 is connected to two Service Providers. Customer is sending 10.0.10.0/24 prefix to both ISP
- They are advertising this prefix to their upstream ISPs and also each other through BGP peering
- AS 30 wants to be used as backup. Thus Customer is sending the 10.0.10.0/24 prefix towards AS30 with As-path prepends. Customer prepends its own AS path with 7 more AS
- You might think that link from AS 30 won't be used anymore so it will be used as backup. But that's not totally true !

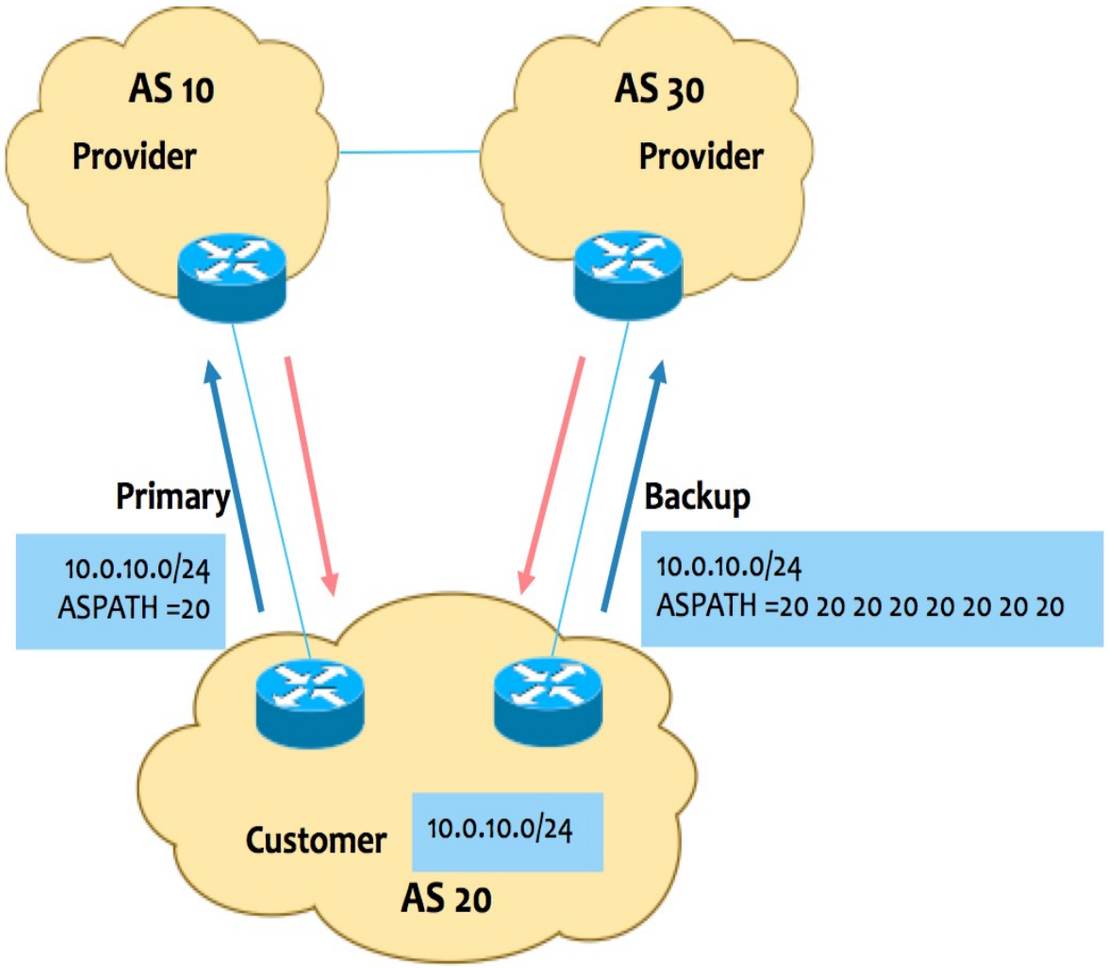
https://t.me/learningnets




- Traffic from their upstream ISPs will go to the AS 10 because all the other ASes over Internet will see the advertisement from AS 30 with lots of prepends. So far so good

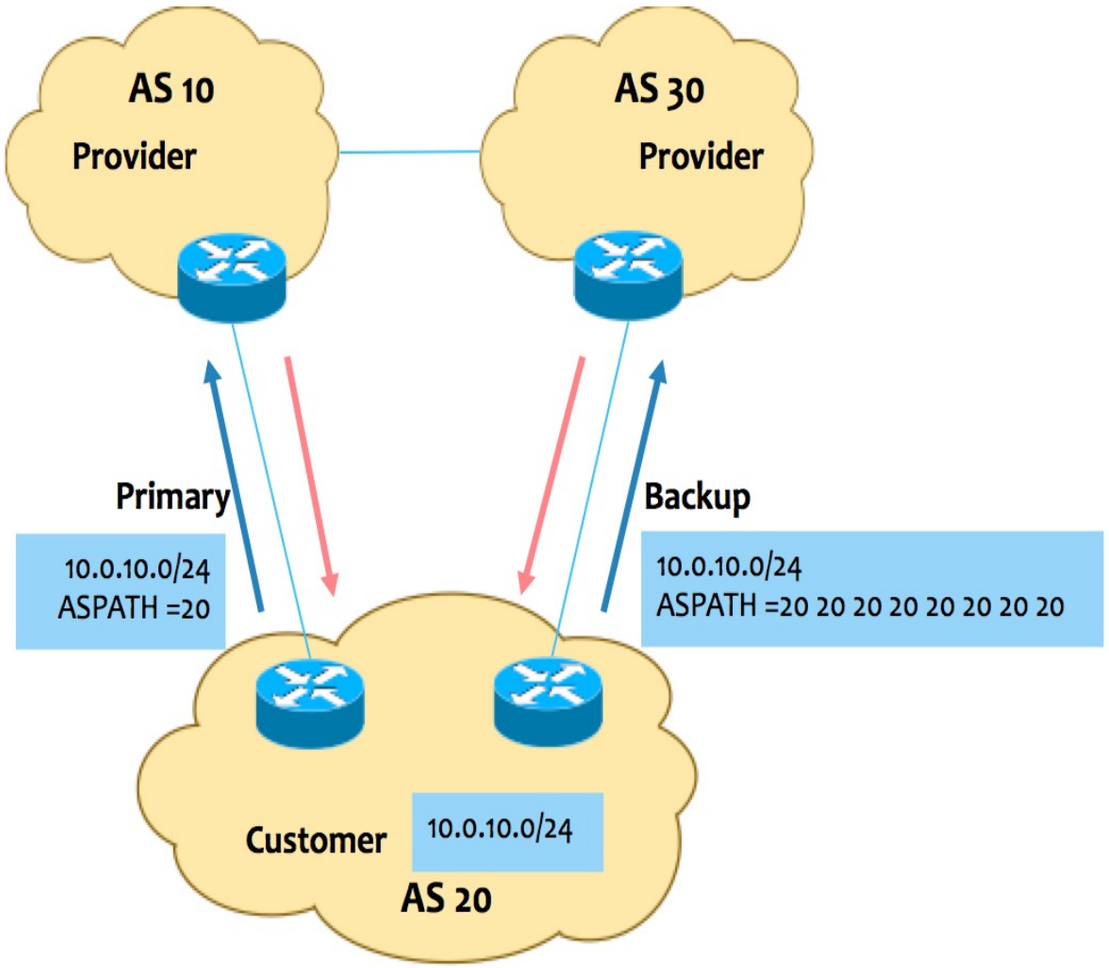
- But all the customers of AS 30 will still send the traffic for 10.0.10.0/24 prefix over the link which wants to be used as backup, although AS 30 learns 10.0.10.0/24 prefix over BGP peering link with AS 10 as well, its upstream providers as well.

https://t.me/learningnets



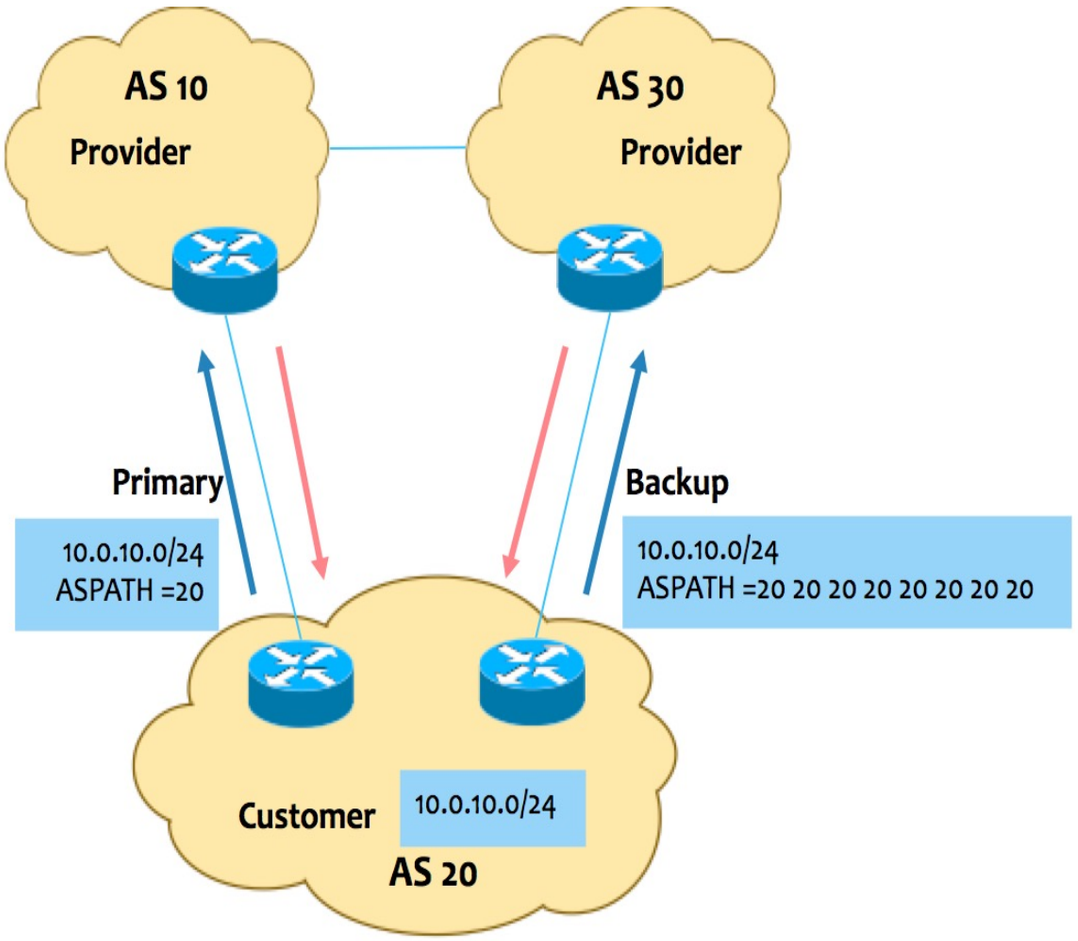

- Service Providers chooses to send the traffic for their customer prefixes over the customer link first, then peering links, lastly through upstream ISP. Because they want to utilize the customer link as much as possible to charge more money

https://t.me/learningnets




- Service Providers implement Local Preference attribute to achieve this. Basic local .preference policy could be; Local Preference 100 towards Customer, local Preference 90 towards peering link and Local Preference 80 towards upstream ISP

https://t.me/learningnets

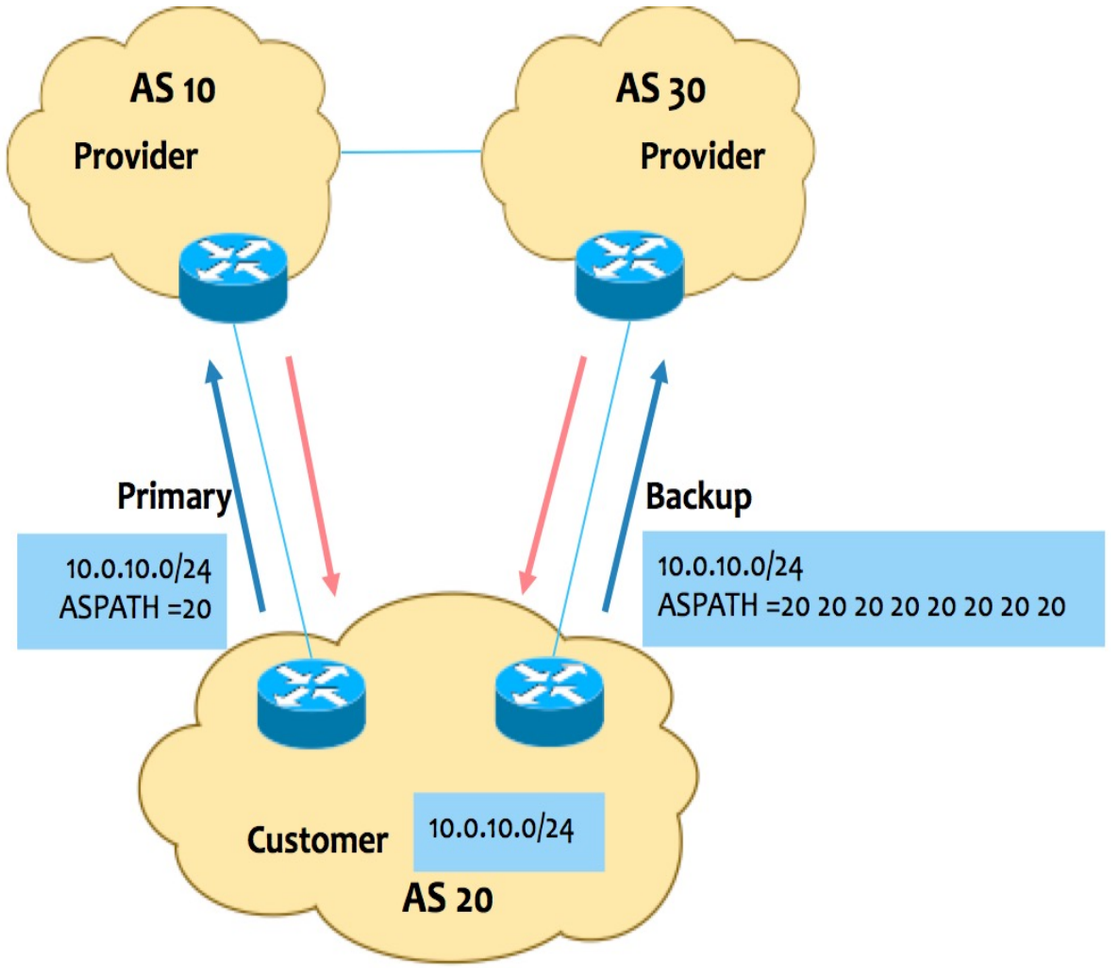



- Customer of AS 30 would still use customer link for 10.0.10.0/24 prefix although customer wants that link to be used as backup
- AS 20 is sending that prefix with AS-path-prepends but service provider implements Local Preference for that prefix

- Since Local Preference attribute is more important in the BGP best path selection process, if the traffic comes to any of the BGP routers of AS 30, it is sent through customer link. Not through BGP peering link with AS 10 or any upstream provider of AS 30

**This problem can be solved with BGP community**

https://t.me/learningnets




Instead of prepending AS,BGP community attribute technique should be used instead of prepending AS, if the topology is multi homed BGP!

- AS 20 sends 10.0.10.0/24 prefix with the BGP community which changes local preference value of AS 30, link between customer and AS 30 is not used anymore.

- As an example AS 20 could send the community as 30:70 which reduces the Local Preference to 70 for the AS 20 prefixes over the customer BGP session, AS 30 would start to use BGP peer link to reach to 10.0.10.0/24 prefix

https://t.me/learningnets

- Community attribute is sent over the BGP session between BGP Peers. Upon receiving the prefixes BGP peer can take an action for their predefined communities

- ISPs publish their supported community attribute values. For example they can say that if my customer send the prefixes with the attached 5000:110 community I will apply Local preference 110 towards that circuit.

(Level 3 (Tier 1 ISP) Community Values and Corresponding Local Preference)

https://t.me/learningnets

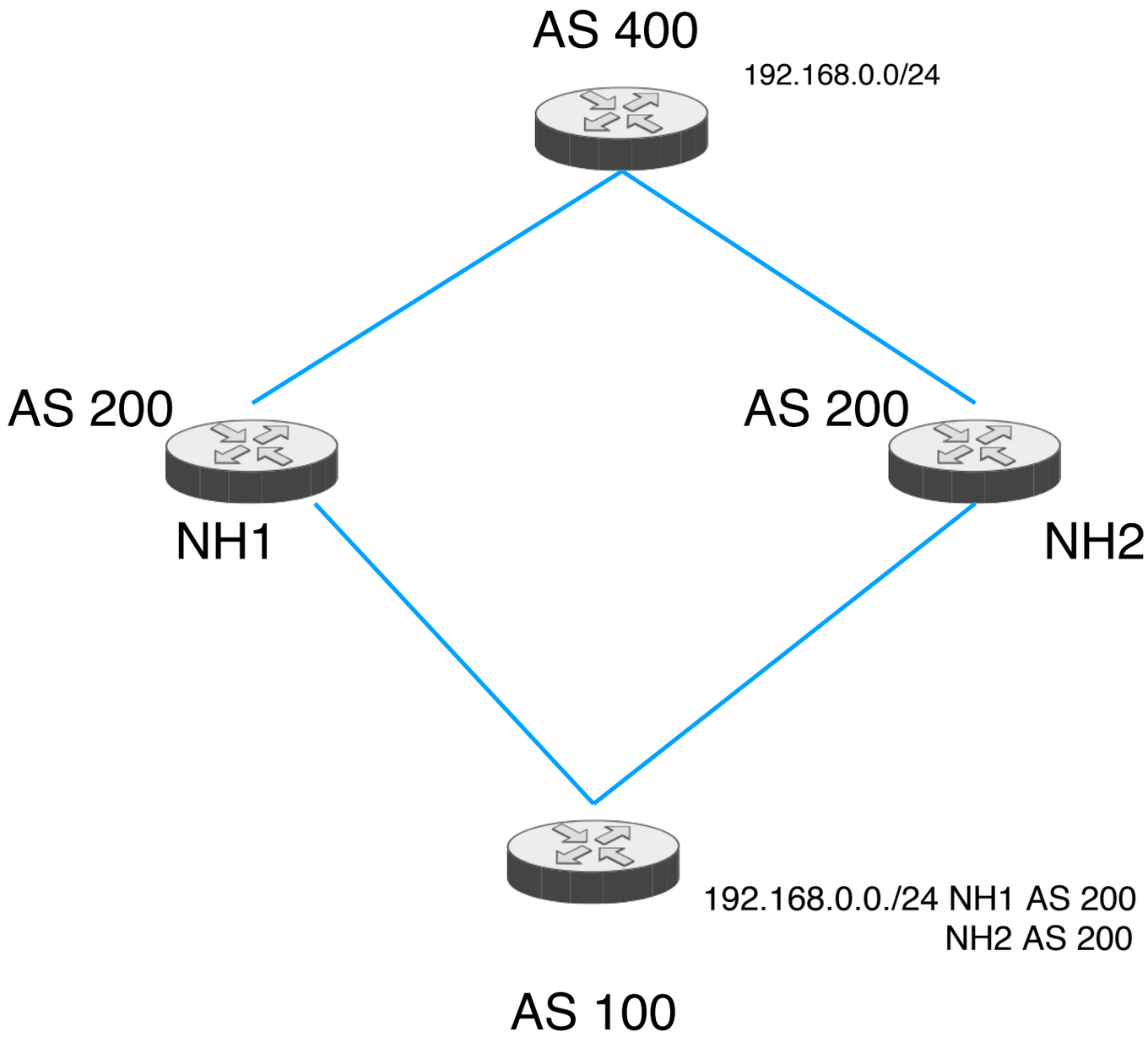
# BGP Multipath

- BGP by default installs only single path in IBGP and EBGP deployment
- If prefixes is learned via multiple path, BGP supports multipath for IBGP , EBGP or across both IBGP and EBGP via EIBGP Multipath feature
- Multipath feature should be enabled manually

https://t.me/learningnets

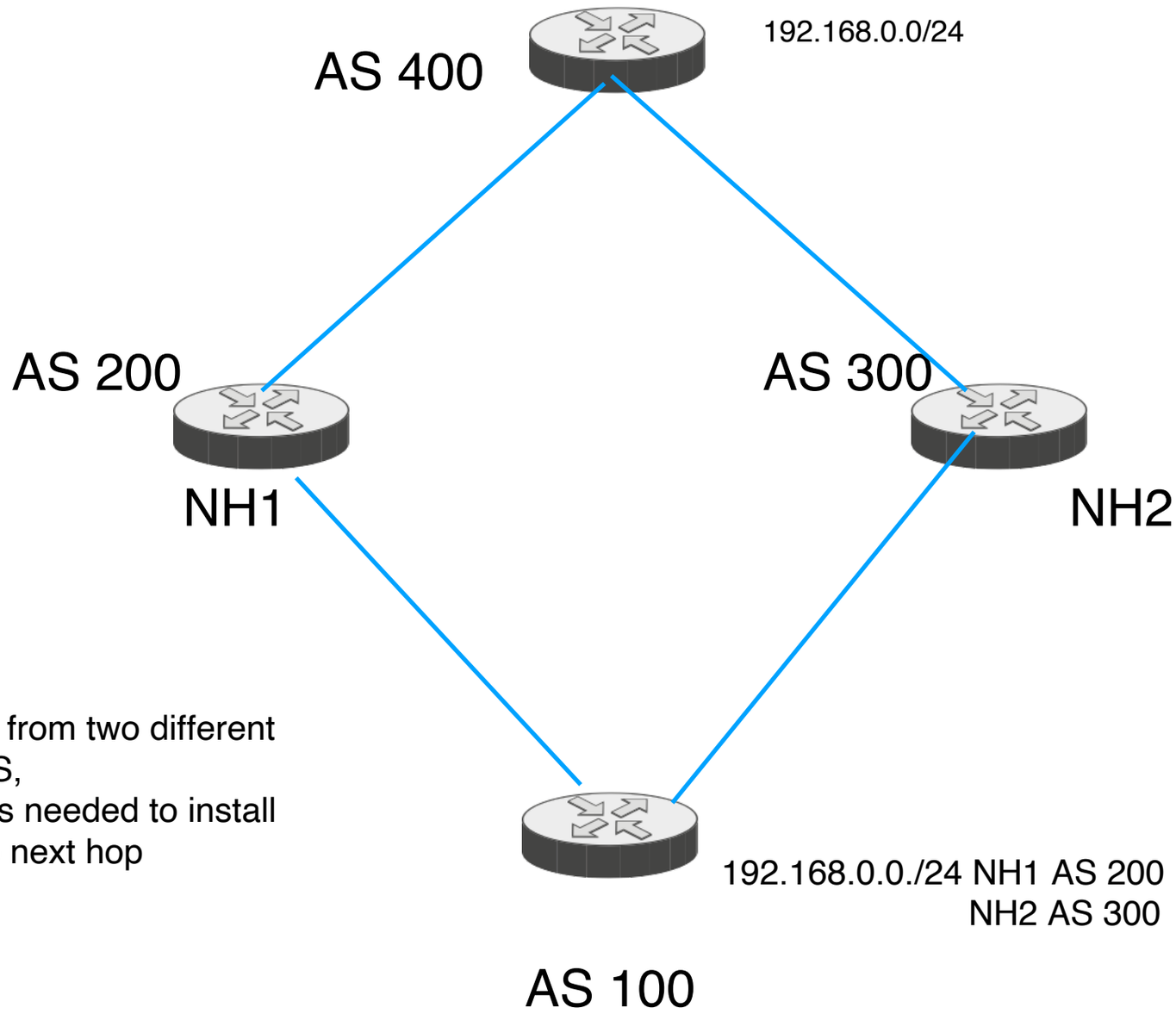
# EBGP Multipath

EBGP Multipath works by default only if next hops from the single EBGP AS




https://t.me/learningnets

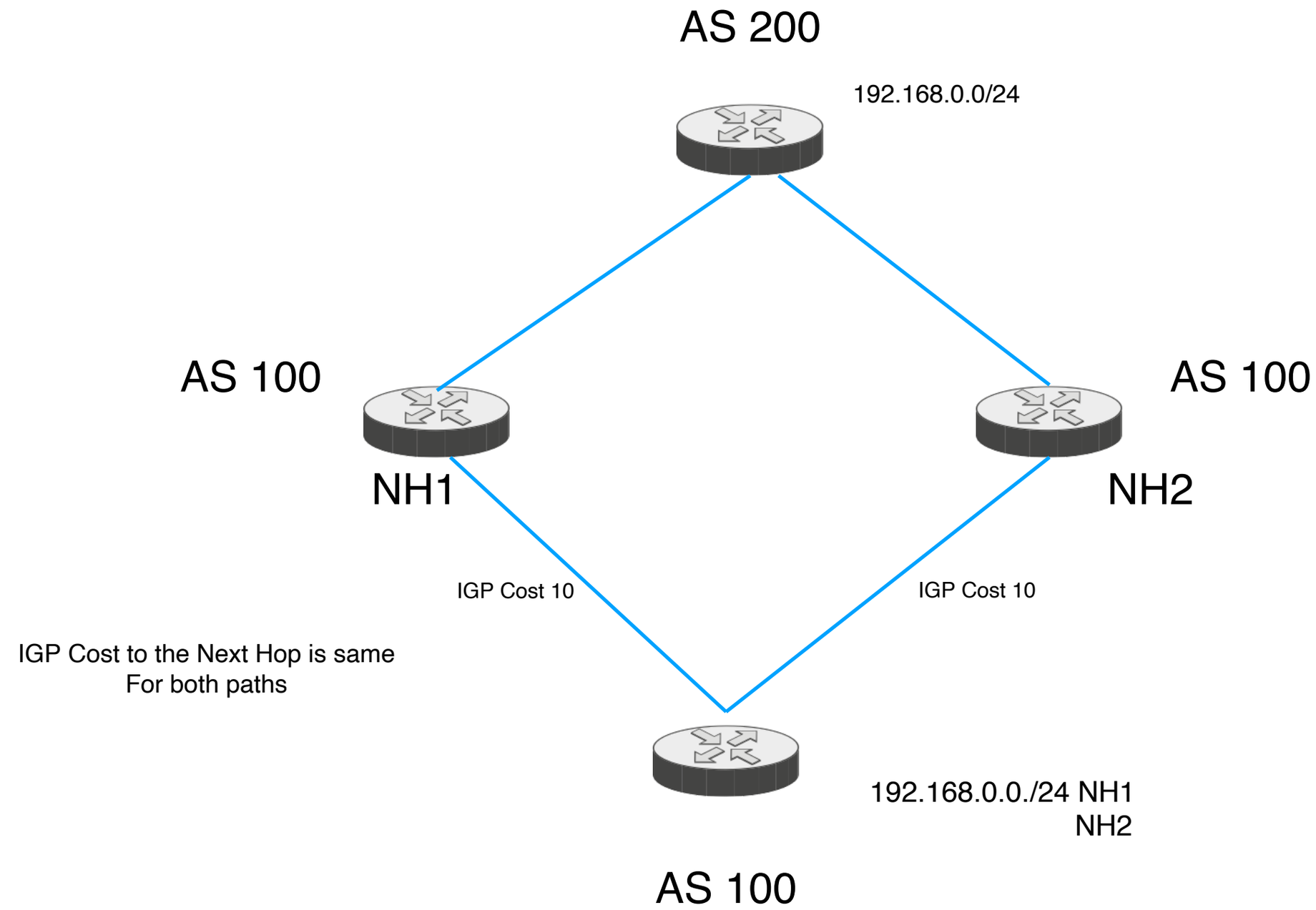
# Two different AS requires ‘ multipath-relax ‘




When prefix is advertised from two different EBGP AS,
as-path relax command is needed to install the prefix via two next hop

https://t.me/learningnets

# IBGP Multipath




https://t.me/learningnets

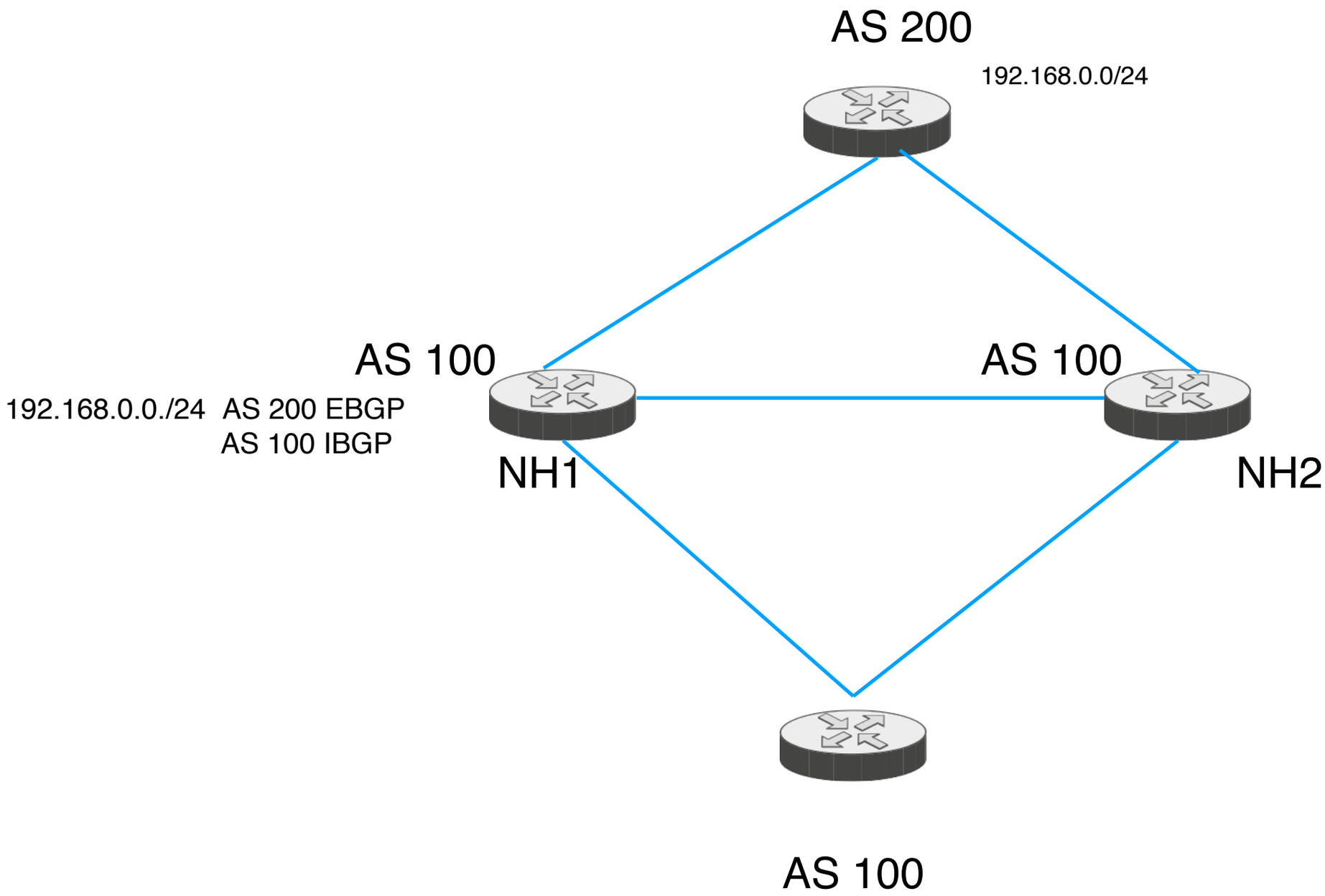
# EIBGP Multipath

- BGP Best path selection algorithm prefers EBGP paths over IBGP paths

- This prevents having both IBGP and EBGP prefixes to be installed in the routing table at the same time

- EIBGP multipath feature allows same prefix to be installed both with IBGP and EBGP next hops

https://t.me/learningnets




https://t.me/learningnets

# EIBGP Multipath

- EIBGP Multipath feature can create routing loop , that is why important to understand it from design point of view

- It is typically used in MPLS L3 VPN deployments

https://t.me/learningnets

# Inter-domain Routing

- To understand BGP peering, first we must understand how network is connected to each other on the Internet

- The Internet is a collection of many individual networks, which interconnect with each other under the common framework of ensuring global reachability between any two points

https://t.me/learningnets

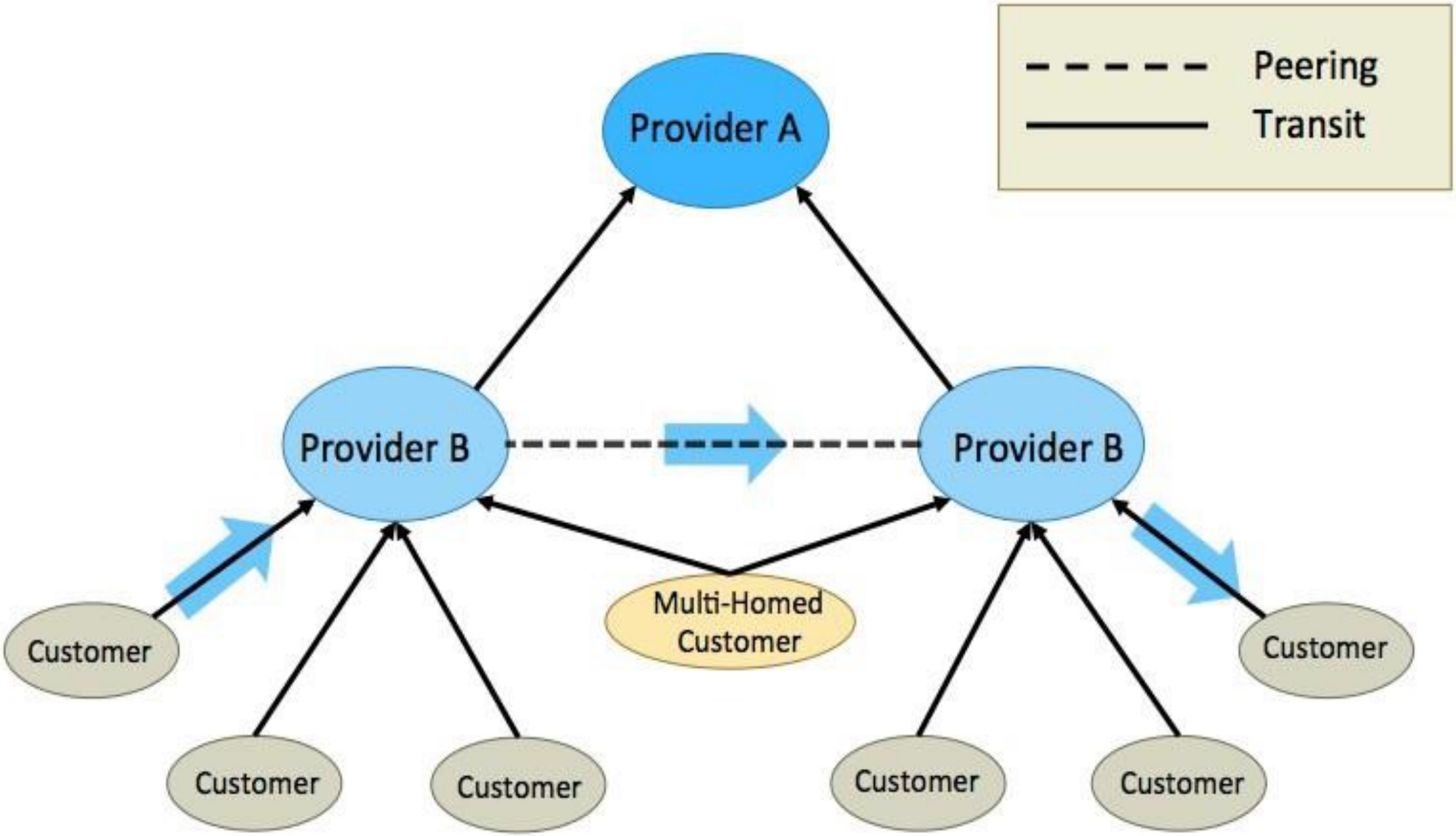
## There are 3 primary relationships for this interconnection

- **Provider** – Typically someone you pay money to, who has the responsibility of routing your packets to/from the entire Internet

- **Customer** – Typically someone who pays you money, with the expectation that you will route their packets to/from the entire Internet

- **Peers** – Two networks who get together and agree to exchange traffic between each others' networks, typically for free. There are two types of peering in general , Private and Public which will be explained later

https://t.me/learningnets




https://t.me/learningnets

# Settlement Free Peering

- Peering is a BGP session between the two Routers. When different companies have Peering with each other, they exchange network traffic over the peering session

https://t.me/learningnets

## There are three reason to have BGP peering on Internet:

- Company wants to receive an Internet service

- Company wants to sell an Internet service

- Two companies exchange their customer prefixes and exchange network traffic but don’t pay each other, which is called Settlement Free Peering

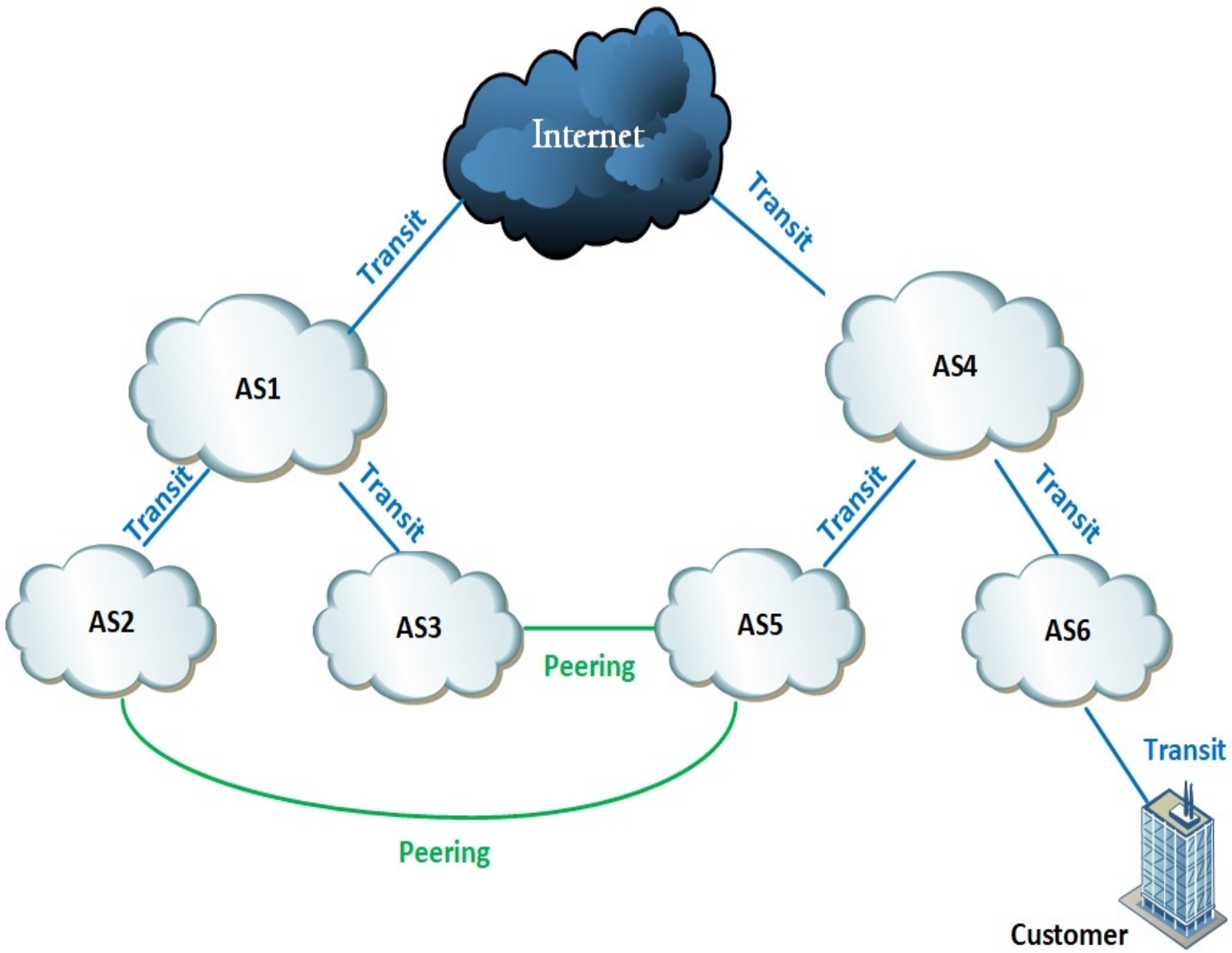
https://t.me/learningnets

- Settlement Free Peering is also referred as Settlement Free Interconnection and here onwards, to make it short, SFI term will be used

- SFI is an agreement between different Service Providers. It is an EBGP neighborship between different Service Providers to send BGP traffic between them without paying upstream Service Provider

https://t.me/learningnets

# Business relationship between the networks.




https://t.me/learningnets

# Private BGP peering

- Private Peering is a direct interconnection between two networks, using a dedicated transport service or fiber. It is also known as bilateral peering in the industry. May also be called a Private Network Interconnect, or PNI.

- Inside a datacenter this is usually a dark-fiber cross-connect. May also be a Telco-delivered circuit as well.

- If there is big amount of traffic between two networks, private peering makes more sense than public peering

https://t.me/learningnets

# Private Peering

- Private peering can be setup inside Internet Exchange Point as well

- Larger companies generally use Private peering rather than Public peering since they want to select who they are going to be peer with and the amount of traffic between them are large, they don’t want to exchange traffic with everyone by joining to the Public Peering

https://t.me/learningnets

# Public BGP Peering

- Typically, public peering is done at the Internet Exchange Point. BGP Route servers are used in public peering to improve scalability.

- Multilateral Peering is commonly adopted in Public Peering

https://t.me/learningnets

# BGP Route Server

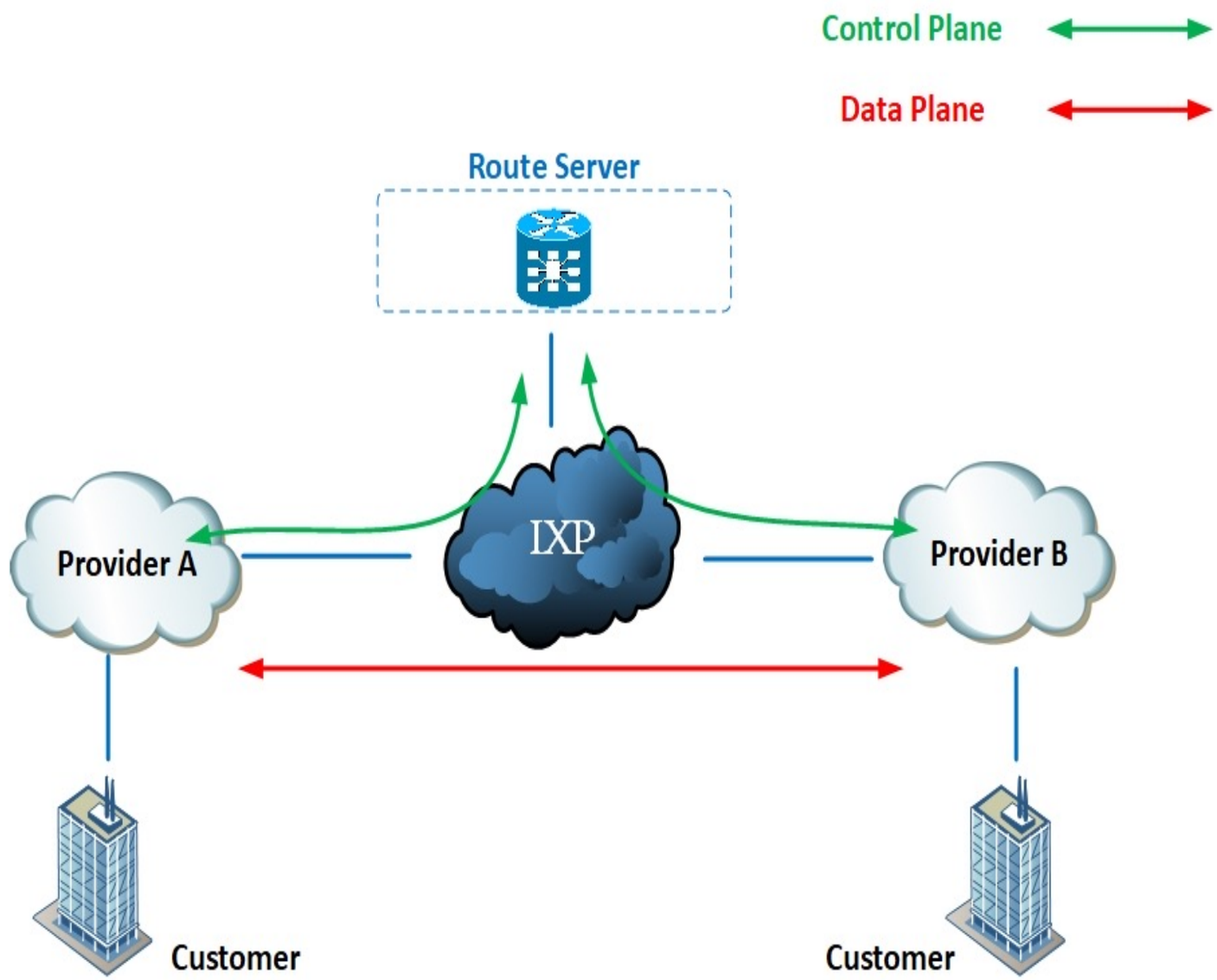
- BGP Route Server is used at the Internet Exchange Point to simplify BGP Peering process. Instead of managing, maintaining hundreds of Peering sessions in large Internet Exchange Point, BGP Route Server is used

- Every BGP speaking router has a BGP session with BGP Route Server

- Route Server doesn’t change the BGP Attributes, although the type of BGP Peering session is EBGP

https://t.me/learningnets

# BGP Route Server




https://t.me/learningnets

# BGP Route Server

- BGP Route Server doesn't change the next-hop to itself, thus it is used only as a Control Plane device, not a Data Plane. Which means, actual traffic is passed between the companies that participant to the Public Peering Internet Exchange Point

- It is very similar to BGP Router Reflector which is used in IBGP topologies. The difference is, BGP Route Server is used in EBGP

https://t.me/learningnets

# Bilateral Peering

- When two networks negotiate with each other and establish a peering session directly, this is called Bilateral Peering. Generally done when there is a big amount of traffic between two networks

- Also Tier 1 Operators just do Bilateral peering as they don’t want to peer with anyone other than other Tier 1 Operators. Rest of the companies are their potential customers, not their peers

https://t.me/learningnets

# Multilateral Peering

- Bilateral peering offers the most control, but some networks with very open peering policies may wish to simplify the process, and simply “connect with everyone”. To help facilitate this, many Exchange Points offer “multilateral peering exchanges”, or an “MLPE”

- An MLPE is typically an exchange point that offers “route-server”, allowing a member to establish a single BGP session and receive routes from every other member connected to the MLPE

https://t.me/learningnets

# Multilateral Peering

- Effectively, connecting to the MLPE is the same as agreeing to automatically peer with everyone else connected to the MLPE, without requiring the configuration of a BGP session for every peer

- Basically, Public Peering and MLPE is almost the same thing and used mostly interchangeably

https://t.me/learningnets

# Looking Glass

- It is a server commonly deploy by an IXP to provide a view to the prefixes in the specific IXP

- It gives publicly available information so any network owner can check the available prefixes in the IXP before they decide to join to that particular IXP

https://t.me/learningnets

# Looking Glass

- There are many publicly available Looking Glasses in the World. They are configured as read-only so person who wants to check the particular looking glass cannot change the BGP routing information

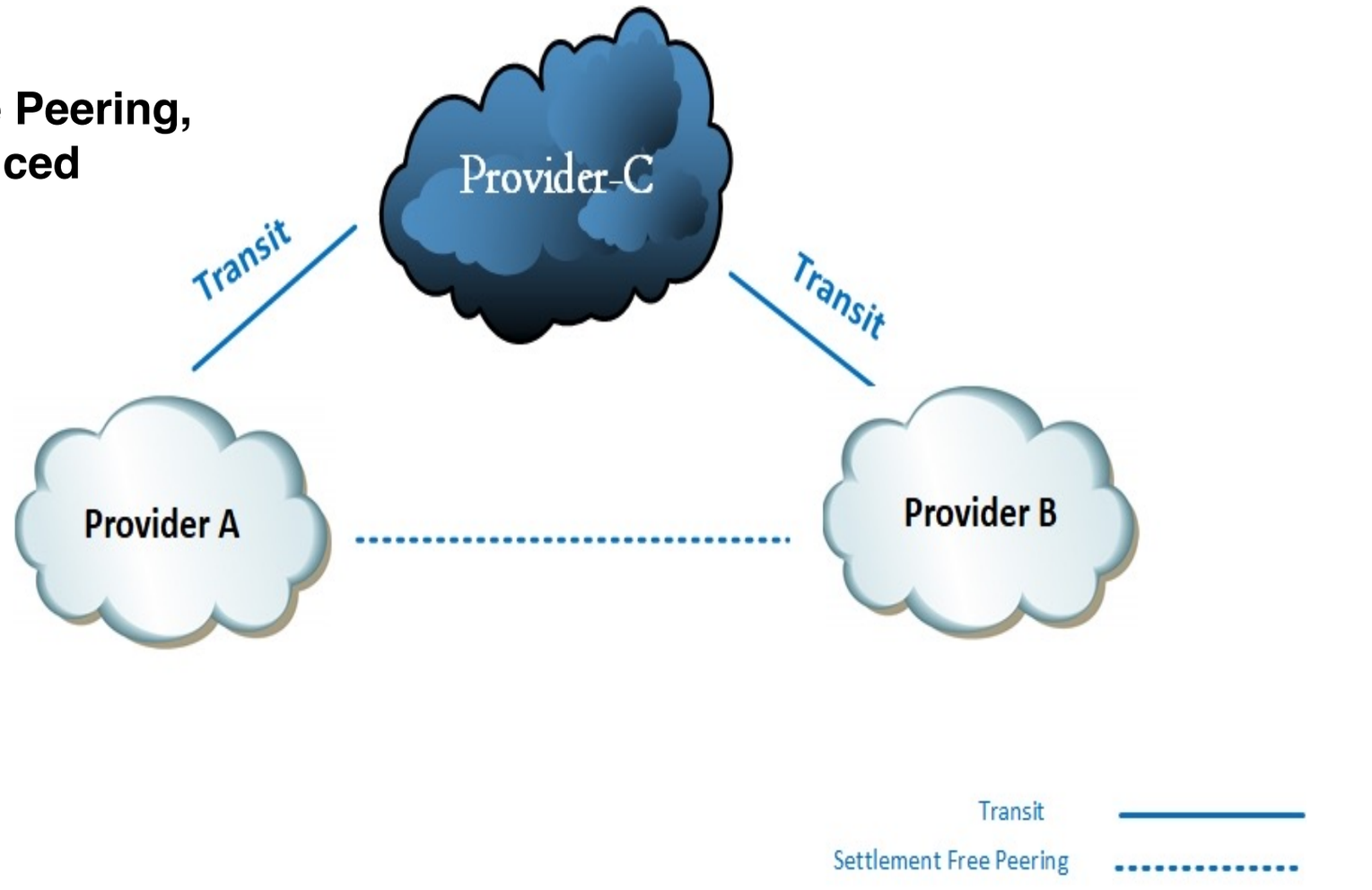
- From http://www.bgplookingglass.com you can see the Looking Glass database

https://t.me/learningnets

# Benefits of Settlement Free Peering

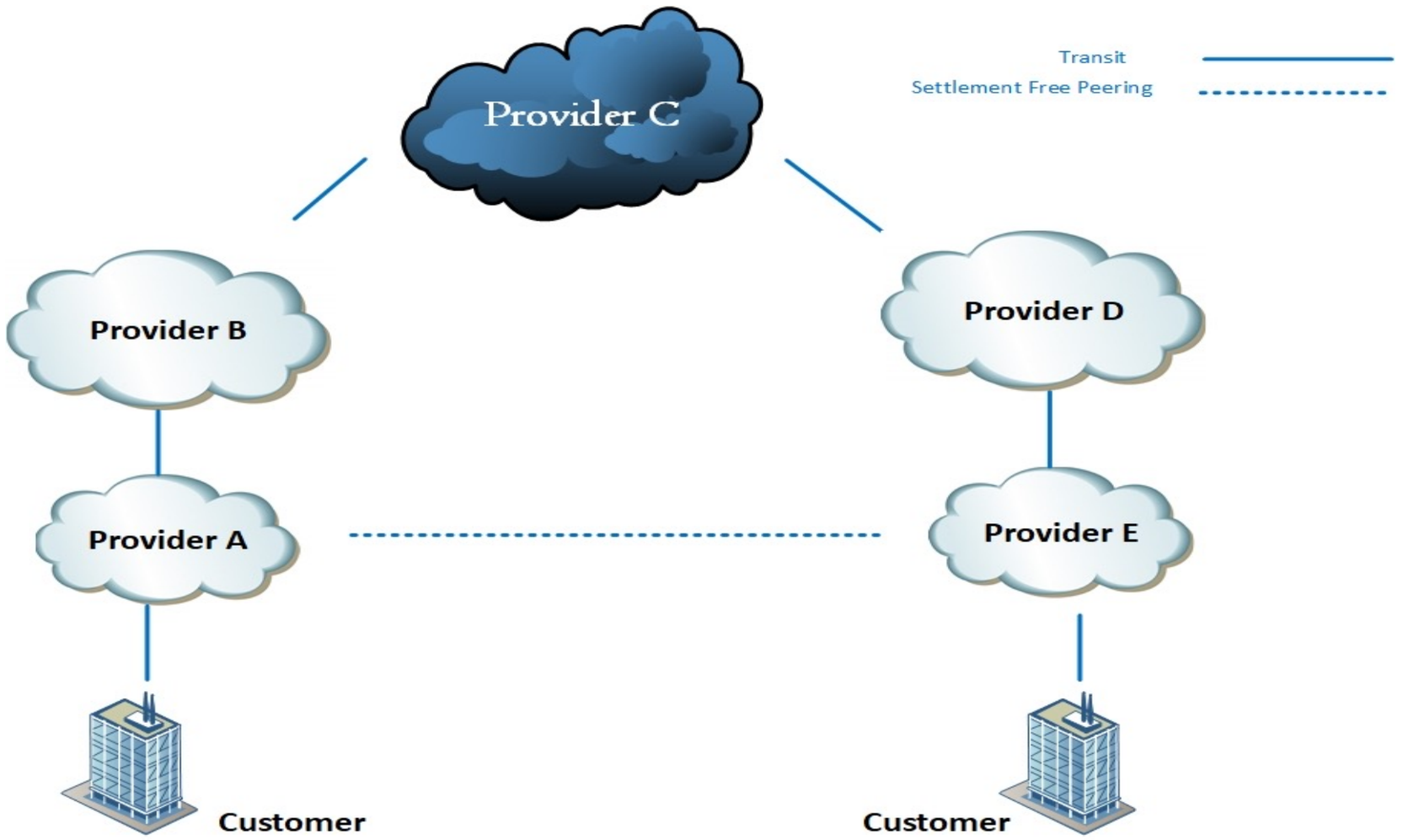
- Reduced operating cost: A transit provider is not being paid to deliver some portion of your traffic. Peering traffic is free!

**Through Settlement Free Peering, Transit Cost is reduced**




https://t.me/learningnets

Improved routing: By directly connecting with another network with whom you exchange traffic, you are eliminating a middle-man and potential failure point




https://t.me/learningnets

# Settlement Free Peering Benefits

- Distribution of traffic: By distributing traffic over interconnections with many different networks, the ability to scale is potentially improved

- Almost every country has Internet Exchange Point where Service Providers, Content Networks, CDNs, Enterprises, Mobile Operators, Carriers, TLD (Top Level Domains) and Root DNS Servers can meet

https://t.me/learningnets

# What is IXP (Internet Exchange Point)?

- A layer 2 network where multiple network entities meet, for the purposes of interconnection and exchanging traffic with one another

- Internet Exchange Points start with a single Layer 2 switch at one location. Networks peer with each other in this facility

- When the number of participant grows, more switches are added at that location and more locations are added to the IXP itself. For example, AMS-IX in Netherlands have many places, inside many Datacenters and each Datacenter they have more than one switch for the Settlement Free Interconnection

https://t.me/learningnets

# What is IXP (Internet Exchange Point)?

- Often referred to as an Internet Exchange (IX), or “public peering”
- Today most Exchange Points are Ethernet based LANs, where all members sharing a common broadcast domain, and each member is given a single IP per router out of a common IP block (such as a /24)
- Typically done at the carrier neutral datacenters and many cases Datacenter owners provide racks for the peering fabric for free

https://t.me/learningnets

# Why Networks Peer at the IXP?

- An Exchange Point acts as a common gathering point, where networks who want to peer can find each other

- A network new to peering will typically go to an exchange point as their first step, and be able to so many other like-minded networks interested in peering with them

- The more members an exchange point has, the more attractive it becomes to new members looking to interconnect with the most other networks. This is called as “critical mass”

https://t.me/learningnets

# Where are the Internet Exchange Points?

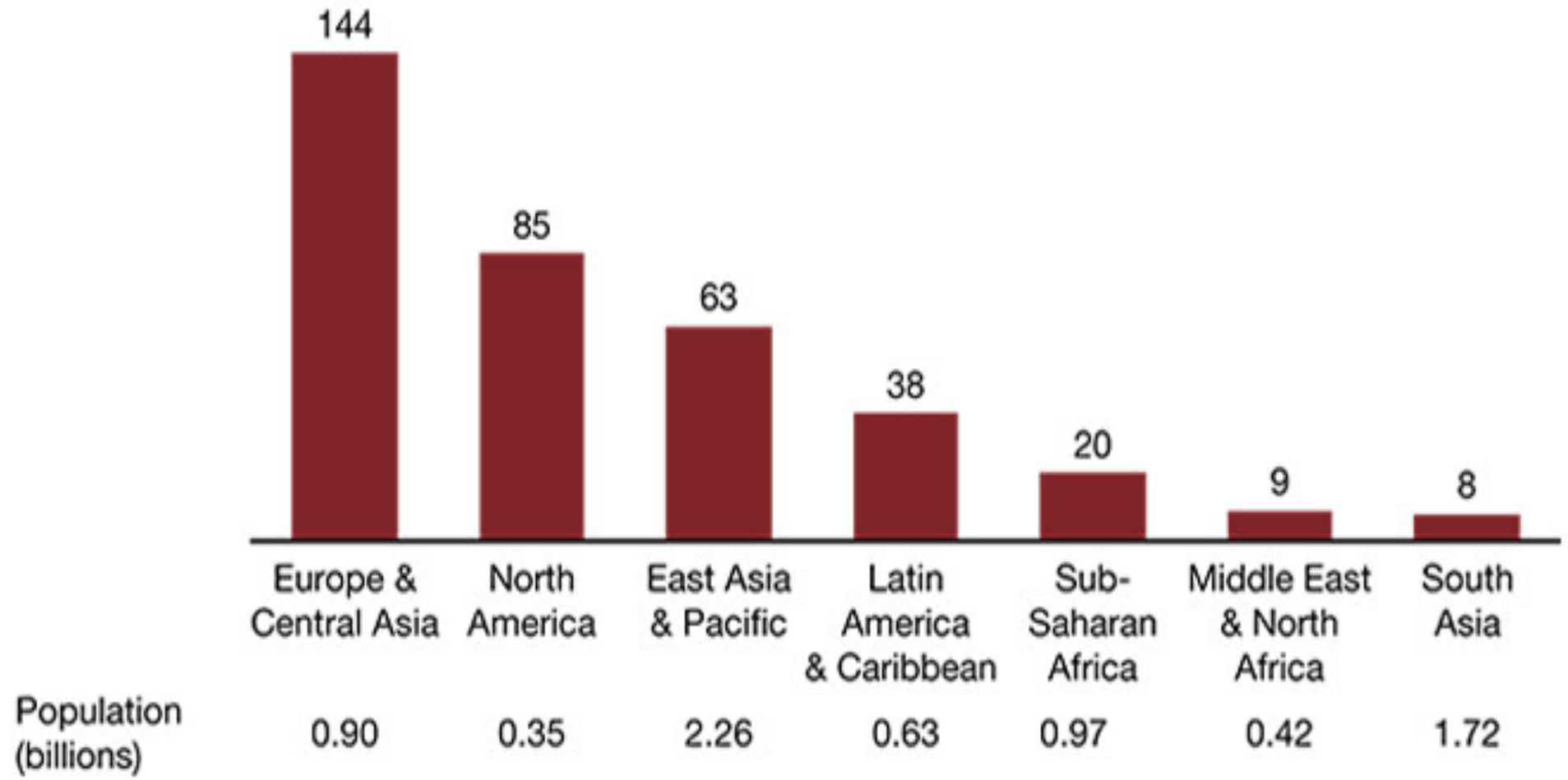
- Most of the IXPs in the World in Europe. There are many IXP in North America as well

- IXPs in the Europe work mainly based on Membership model, IXPs in the U.S work based on Commercial model. There are exceptions in each case though

- Most European IXPs grew from non-commercial ventures, such as research organizations. Most African IXPs were established by ISP Associations and Universities

https://t.me/learningnets

# Number of IXPs in the World




Source: ITU; Packet Clearing House; The World Bank, World Development Indicators

© 2016 PwC. All rights reserved.

https://t.me/learningnets

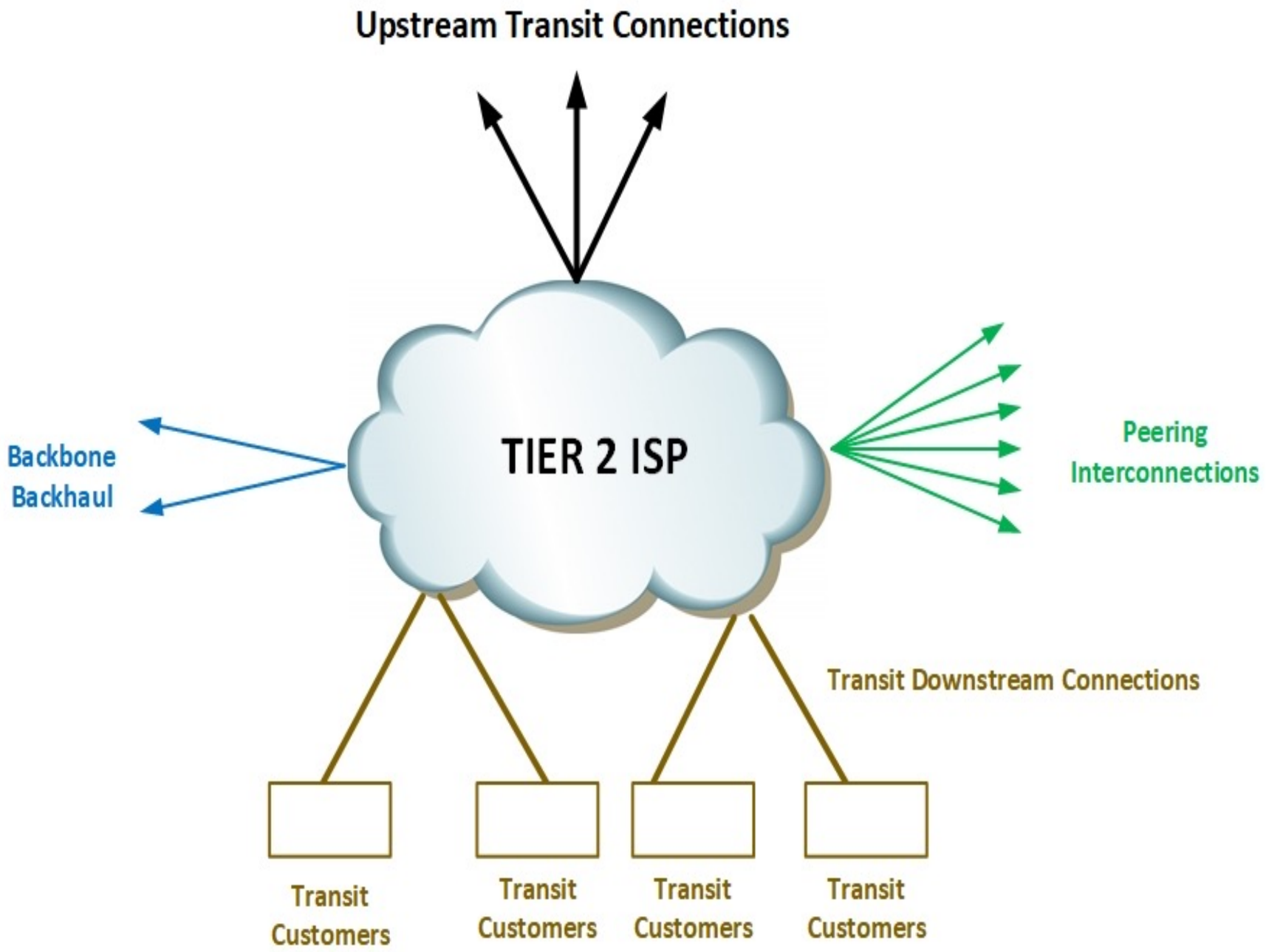
# ISP Tiers – Tier 1, Tier 2, Tier 3 ISP

- Tier 1 Service Provider is a network which does not purchase transit from any other network and peers with every other Tier 1 network to maintain global reachability

- Tier 2 Service Provider is a network with transit customers and some peering, but still buys full transit to reach some portion of the Internet

https://t.me/learningnets

# Tier 2 ISP and its Connections



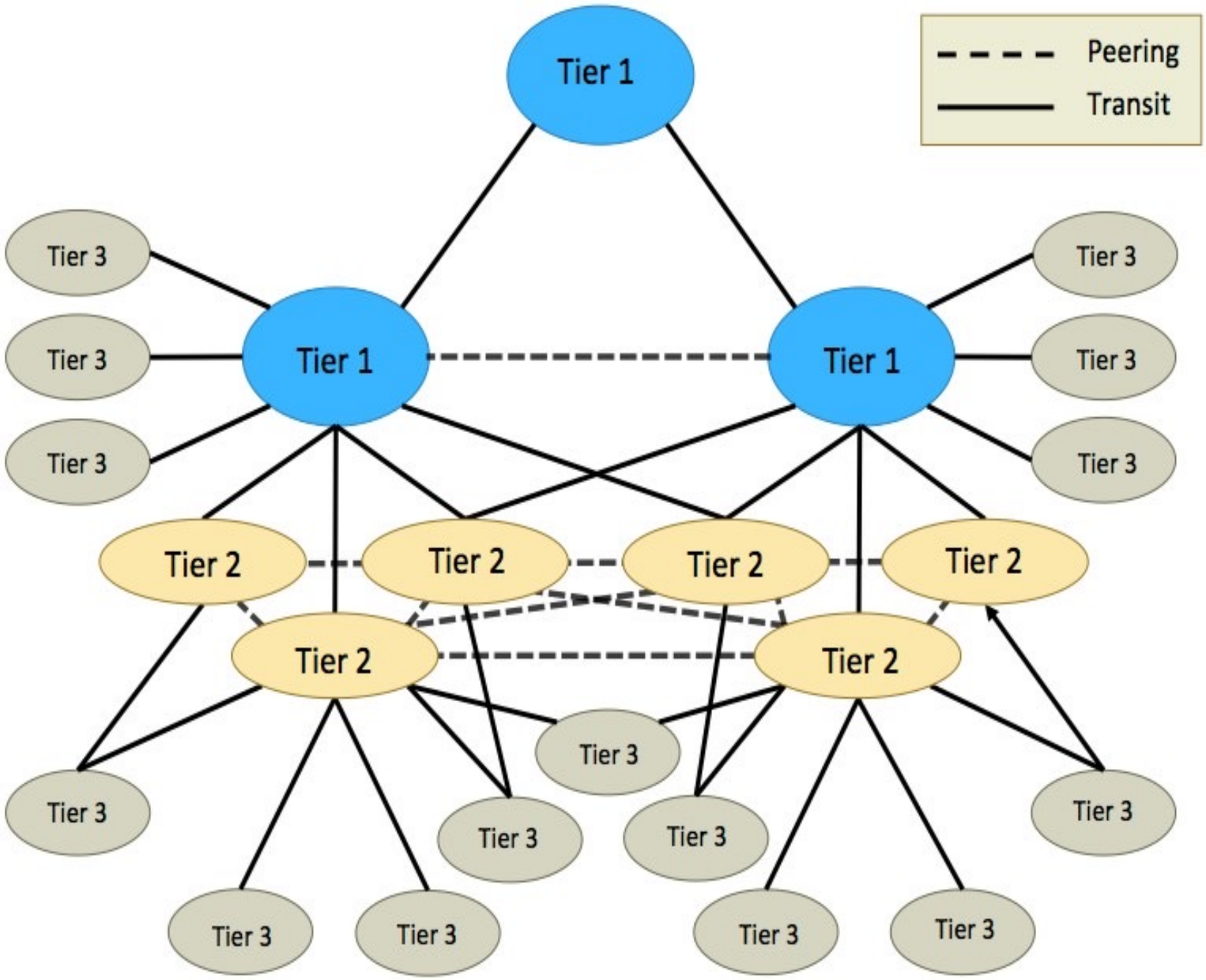

https://t.me/learningnets

# Tier 3 ISP

- Tier 3 Service Provider is considered as stub network. They are generally considered as local/access ISPs. They don’t sell any IP Transit service to anyone. Sometimes, Tier 3 ISP definition is used to describe Enterprise, SMB or End Users

https://t.me/learningnets

# Tier 1 , Tier 2 and Tier 3 ISP Relationship




https://t.me/learningnets

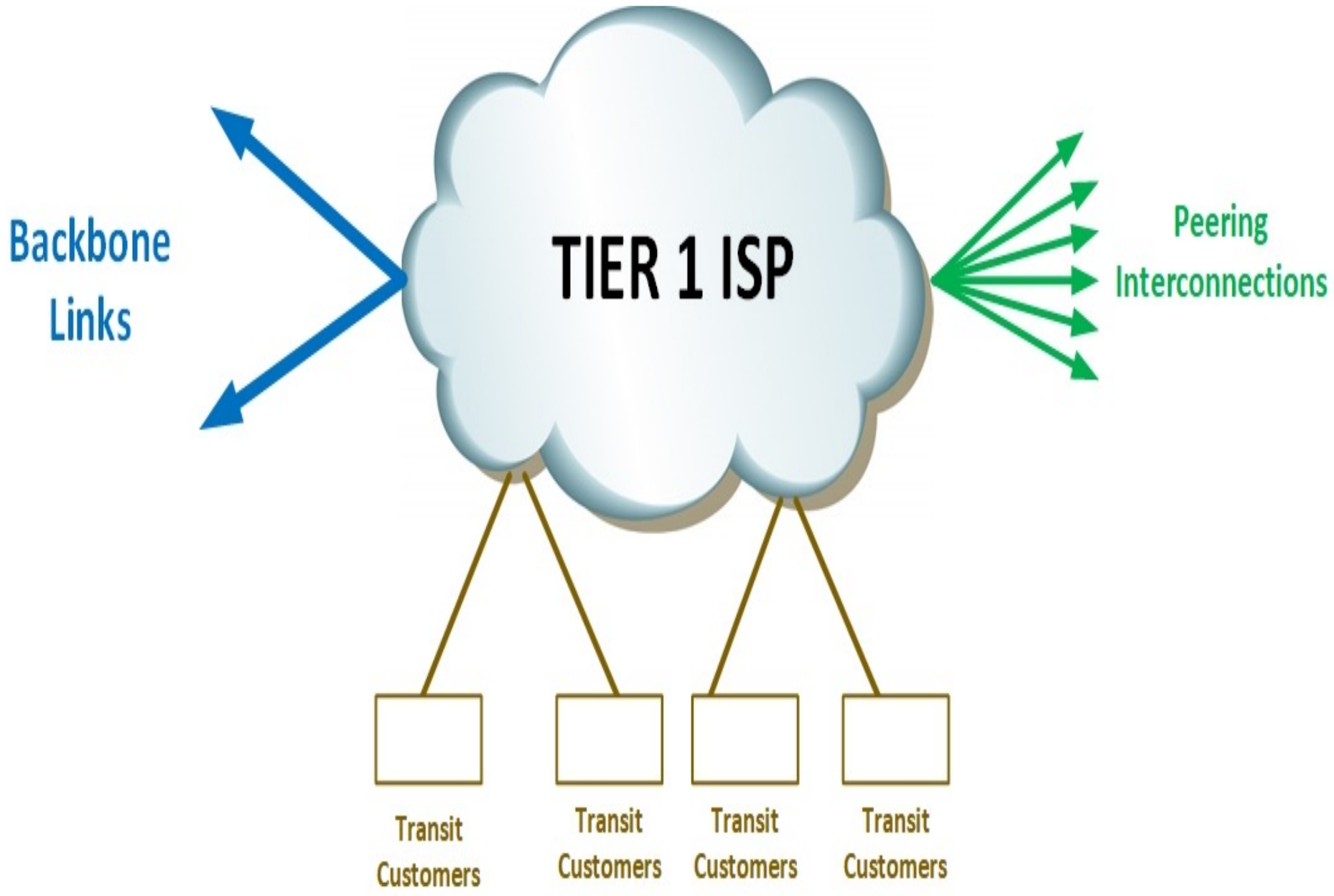
# Tier 1 Service Providers

- A Tier 1 ISP is an ISP that has access to the entire Internet Region solely via its settlement free peering relationship

- Tier 1 ISPs only peer with other Tier 1 ISPs and sometimes with CDN and the Search Engines. They don’t have any Transit ISP but they are the top tier ISP

https://t.me/learningnets

# Tier 1 ISP and its Connections




https://t.me/learningnets

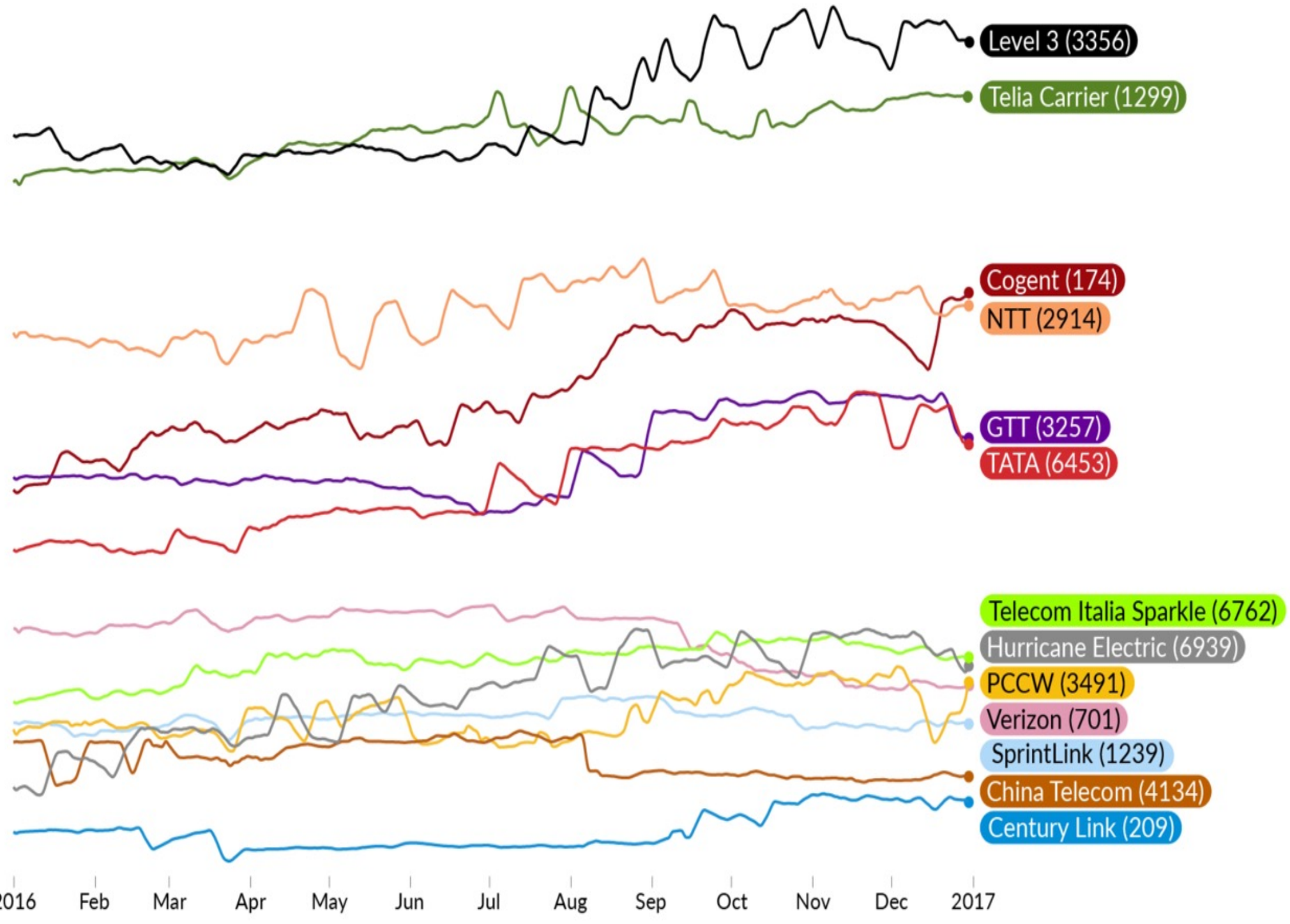
# Who are Global Tier 1 ISPs in the World?

- As of 2016, there are 13 Tier 1 ISP which don’t have any transit provider

- Baker’s Dozen is considered as Tier 1 ISP List and every year list is updated with the ISP ranking. List is provided by measuring the Transit IP Space of each ISP

https://t.me/learningnets

# 2016 Baker's Dozen Tier 1 ISP Rankings




https://t.me/learningnets

# IP Transit

- IP Transit is the service of allowing traffic from another network to cross or "transit" the provider's network, usually used to connect a smaller Internet Service Provider to the rest of the Internet

- It's also known as Internet Transit.  ISPs simply connect their network to their Transit Provider and pay the Transit Provider, which will do the rest

https://t.me/learningnets

# Selling an IP Transit Service in the IXP

- Selling an IP Transit Service in the IXP is common

- Many big Service Providers such as Tier 1 or Regional Tier 2 Providers join an IXP as they see IXP is not only peering point but a location where they can sell their IP Transit Services

https://t.me/learningnets

# Selling an IP Transit Service in the IXP

- Although some IXPs don’t allow selling or buying an IP Transit, there is no real control mechanism which can prevent this situation

- When companies have a peering, they still receive an IP Transit Service (Except Tier 1s) and they use IP Transit as backup connection

https://t.me/learningnets

# BGP Soft Reconfiguration and Route Refresh

- BGP is a policy based protocol and we use inbound and outbound filters with the attributes. BGP updates are kept in many different places in the router.

- BGP RIB which is routing table of BGP , RIB which is a router's general routing table created by all the routing protocols , FIB which is a forwarding table which is data plane

https://t.me/learningnets

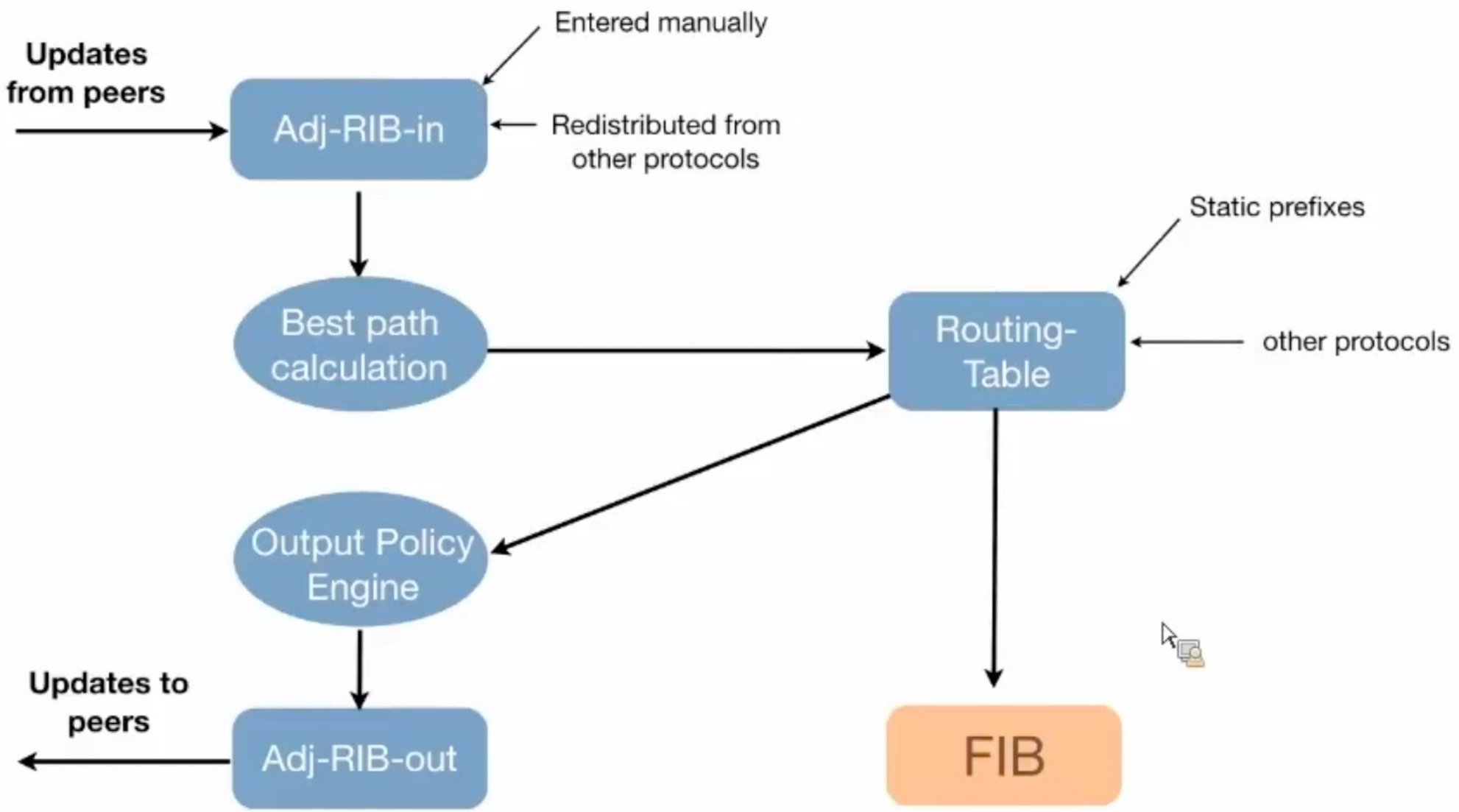
# BGP Soft Reconfiguration and Route Refresh

- In addition to BGP RIB, there are BGP use adjacency RIB-IN and RIB-OUT databases in the Routers

- All the prefixes from the remote BGP neighbor is placed in the BGP RIB-IN database first.

https://t.me/learningnets

# BGP Soft Reconfiguration and Route Refresh




https://t.me/learningnets

- Then inbound filter is applied , if we want to allow them, then prefix is taken into BGP RIB database

- If we enable BGP Soft Reconfiguration Inbound , we keep received prefixes in the BGP RIB-IN database, if it is not enabled, we ignore them

https://t.me/learningnets

# BGP Soft Reconfiguration and Route Refresh

- That's why if BGP soft reconfiguration inbound is enabled, even if you filter the prefixes after receiving from the neighboring BGP device , you can still reach them for maybe troubleshooting purposes

- It helps you to verify whether your filter is working correctly

https://t.me/learningnets

# BGP Soft Reconfiguration and Route Refresh

- But obviously this is memory intensive since you keep those prefixes in BGP RIB-IN database in addition to BGP RIB database

- BGP Route refresh works in a different way to accomplish the same task. Still filter is applied for the incoming or outgoing prefixes

https://t.me/learningnets

- With Route Refresh, you don’t keep the prefixes in the separate databases

- You either take them into BGP RIB database or ignore entirely after filtering

- Thus memory consumption is more efficient

https://t.me/learningnets

# BGP Soft Reconfiguration and Route Refresh

- Don’t forget that Router Memories are expensive

https://t.me/learningnets

# IBGP

- IBGP is used inside an Autonomous system. In order to prevent routing loop, IBGP requires BGP nodes to have full mesh interconnections among them.

- This rule is not required in EBGP because routing loop prevention is done by checking the AS number in the AS path in EBGP. In IBGP, AS number is not sent between the BGP neighbors

- Full mesh IBGP sessions may create configuration complexity and resource problem due to high number of BGP sessions in large scale BGP deployment

https://t.me/learningnets

# IBGP

- Route reflectors and confederations can be used to reduce the sessions on each router. Number of sessions and configuration can be reduced by the route reflectors and confederations but they both have important design considerations

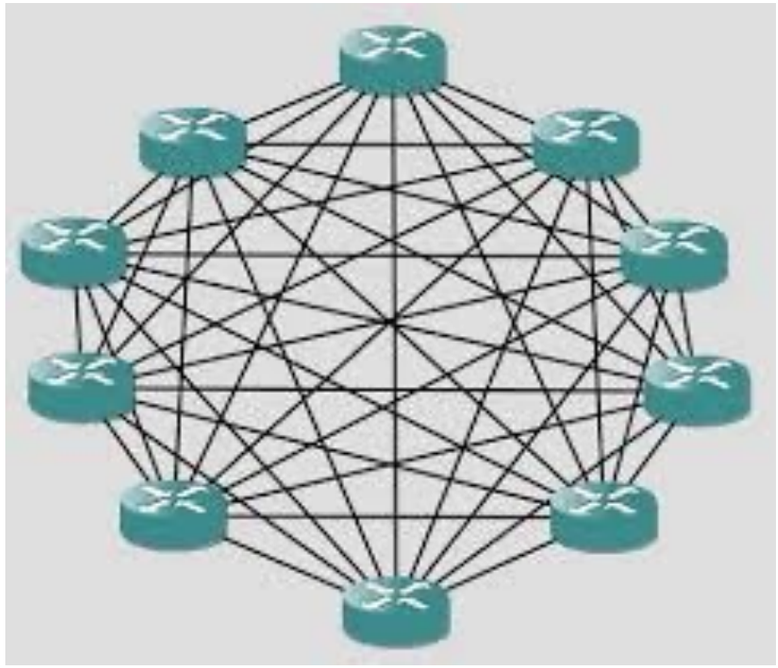
- Confederations divide the autonomous system to smaller sub-Autonomous systems

- Confederations give the ability to have ebgp rules between Sub-ASes. Also inside each Sub-AS, different IGP can be used. Also merging company’s scenarios is easier with Confederation than Route Reflectors

https://t.me/learningnets

# BGP Route Reflectors

- It is used to avoid Full Mesh connectivity requirement in IBGP



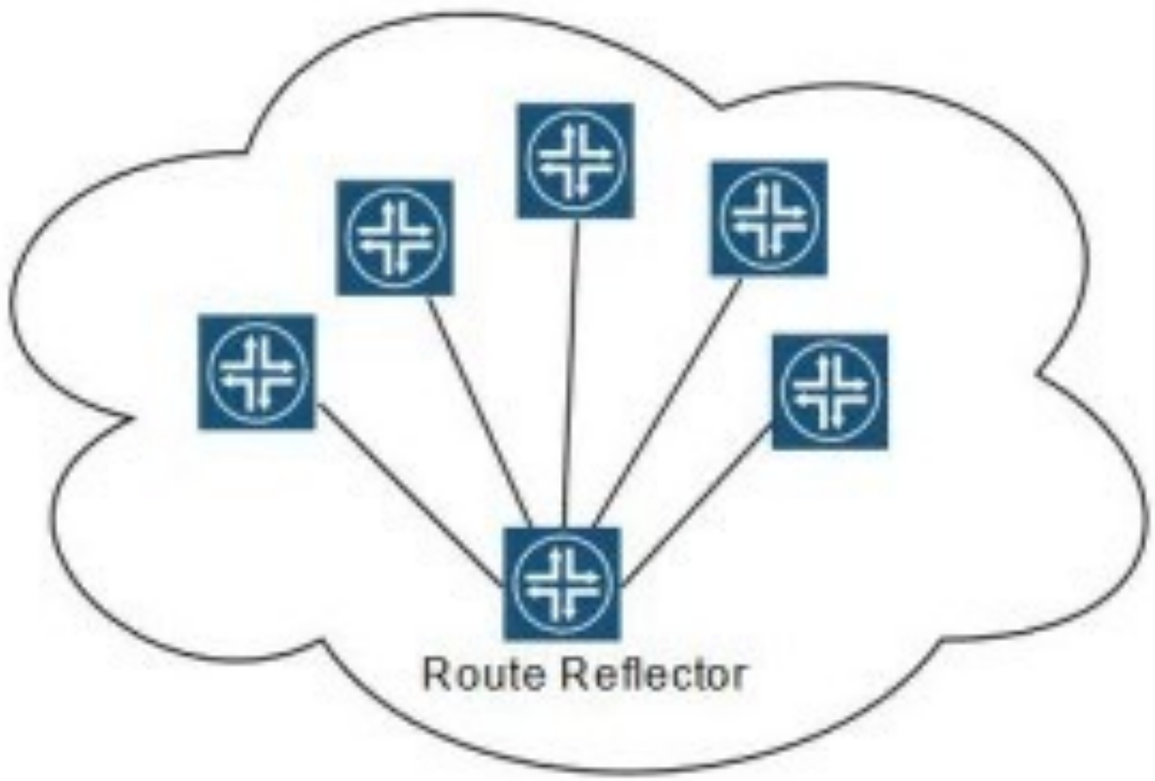
- There are many design caveats when BGP RR is used, most important ones are BGP RR can create sub optimal routing, increases convergence time, reduces redundancy

https://t.me/learningnets

# BGP RR Creates Logical Hub and Spoke Topology

- It creates a logical Hub and Spoke Topology. Each BGP RR Client has a BGP session with only BGP RR, not with each other.

- Thus Full Mesh IBGP topology become Hub and Spoke IBGP Topology with BGP RR




https://t.me/learningnets

# BGP RR Path Selection and Distribution

- Route reflectors choose the best path to the exit point based on their perspective, and not the client's perspective.

- A path to the exit point of the network for a certain prefix can be optimal for the Route Reflector based on its lowest IGP metric to the exit point, but this might not be true from the client's perspective.

- Route Reflectors only advertise one path as their best path for a prefix and don't advertise any other paths to their clients.

https://t.me/learningnets

# BGP RR Sub-optimal Routing

- With this Route Reflectors behavior which removes additional BGP advertisements to the control plane of its clients, an issue of suboptimal routing will occur for Route Reflector clients.

- This is because the Route Reflector client will not have all the available routes and it cannot compare the IGP metric of every path in order to determine the shortest path.

https://t.me/learningnets

# BGP RR Sub-optimal Routing

- Sub optimality in reflecting the path from the RR to the clients usually happens when the Route Reflector is not topologically near its clients. This sub optimality is more seen when RRs are not in the forwarding path, especially in virtual RR’s that are completely out of path.

https://t.me/learningnets

# BGP RR Works Based On

- Route selection is based on their point of view
- RRs propagate only the best path over their sessions by hiding other paths.
- This might not be the best path according to the client's point of view.
- RRs run best path algorithm and advertise only one update to their clients, which may result in suboptimal routing.
- RR's are usually deployed based on exit points in the network

https://t.me/learningnets

# BGP RR vs. Regular BGP Speakers Best Path Selection

- Route Reflectors use the same best-path selection process as normal BGP speakers do. When receiving the same prefix coming from multiple peers, the tiebreaker decision process is done:
- Highest LOCAL_PREF
- Locally originated via network/aggregate or redistributed via IGP.
- Shortest AS-Path
- Lowest origin type
- Lowest MED
- eBGP paths over iBGP
- Path with the lowest IGP metric to the BGP next hop

https://t.me/learningnets

- If all the steps before the 7th step are equal, then step 7 will be the deciding factor for the best path for the Route Reflector. So, the preferred path will be the lowest IGP metric to the BGP next hop.

- By default, Route Reflector's only advertise the best path to their clients, so in case of the tiebreaker explained above, the traffic will be send to the exit point with the lowest cost/shortest path possible.

https://t.me/learningnets

# In-band vs. Out-band BGP RR for Optimal Routing

- In-band Route Reflectors usually have better view of the IGP topology of the network than out-of-band Route Reflectors, so they can advertise better optimal best paths to their clients

- This due to proximity to the RR Clients, with inbound BGP RR, RR is topologically deployed closer to the RR Clients

https://t.me/learningnets

# BGP Route Reflector Design Options

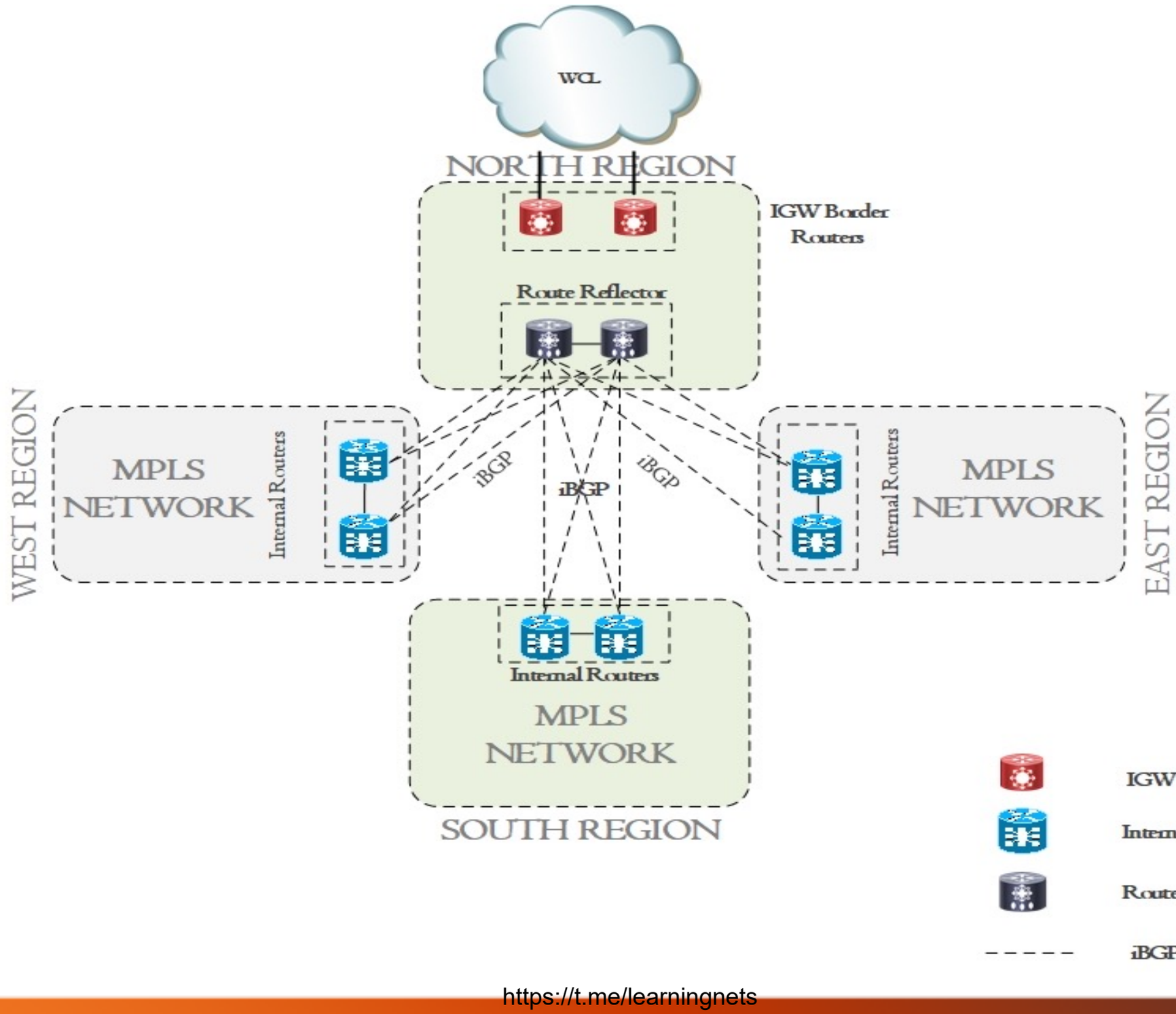
- BGP Route Reflectors can be deployed in a distributed or centralized way

- Distributed BGP Route Reflectors provide optimal routing compare to Centralize design

- Centralized BGP Route Reflectors can be used together with BGP ORR (Optimal Route Reflection) to provide optimal routing for RR Clients

https://t.me/learningnets

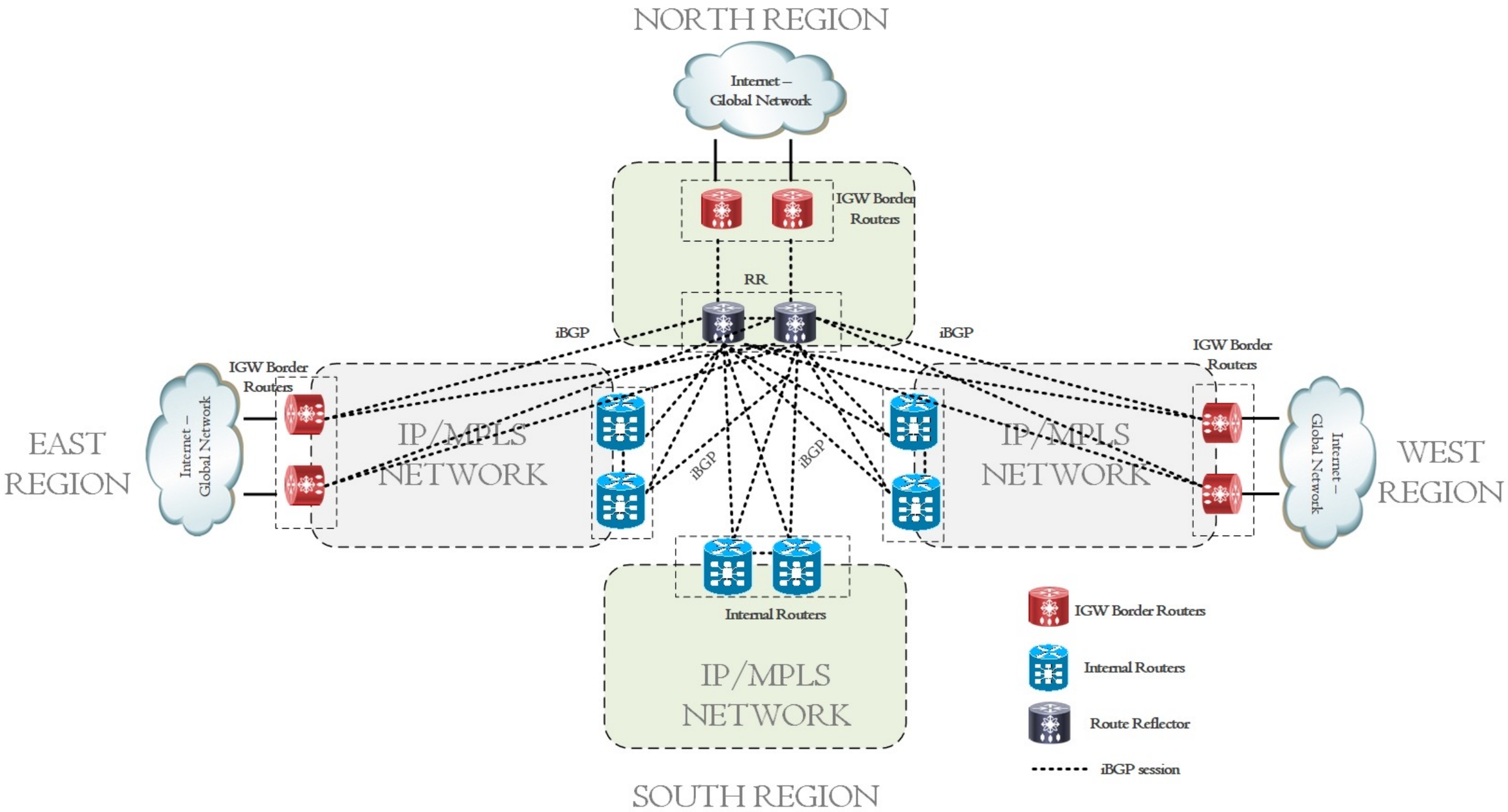
# Centralized BGP RR Design




https://t.me/learningnets

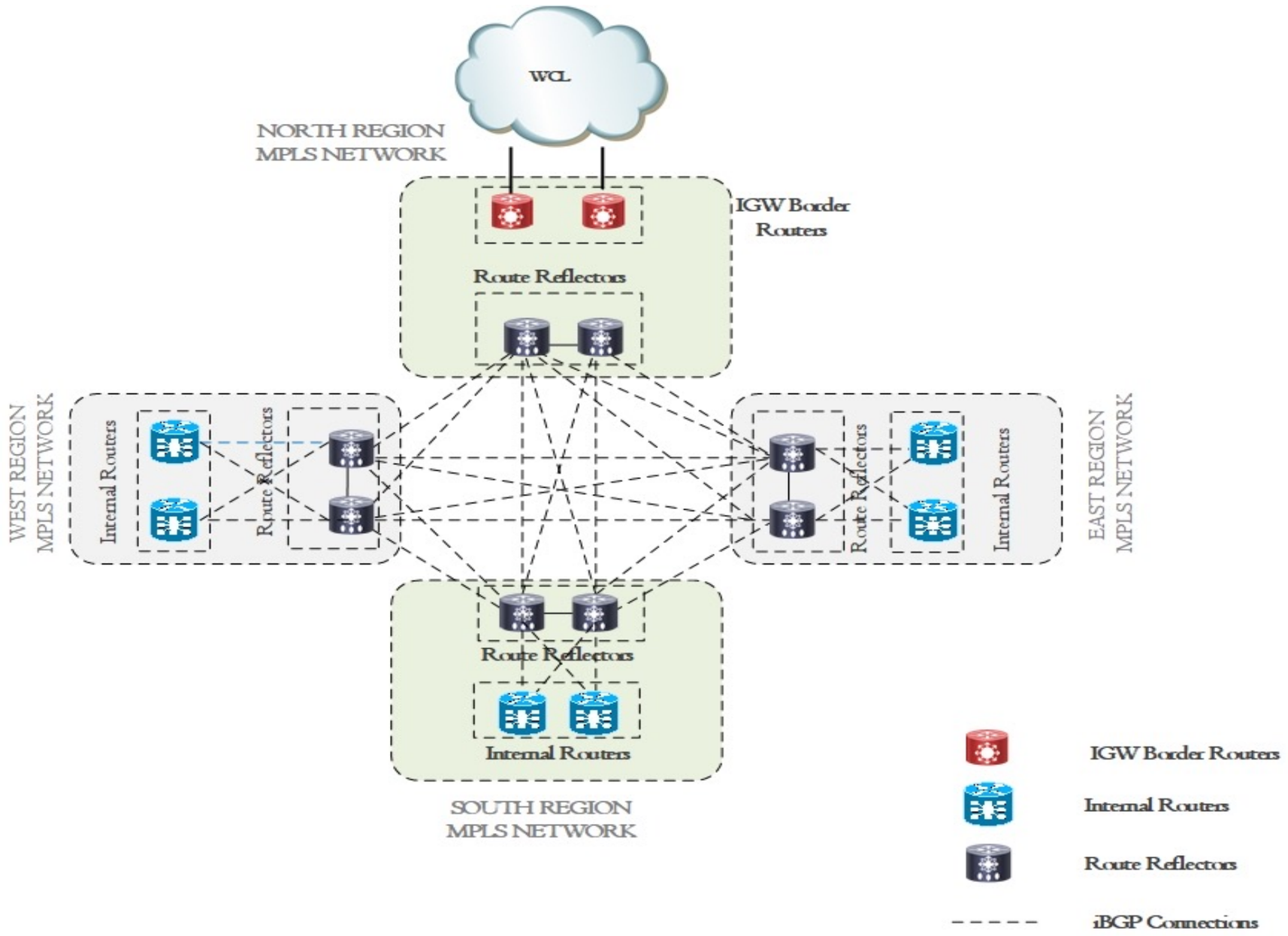
# Centralized BGP RR Design

In Centralized RR Design, even though there are Internet exit from East and West Region as well, all Internal routers use North Region Internet connectivity as RR's point of view North Region Internet Gateways have lowest IGP cost

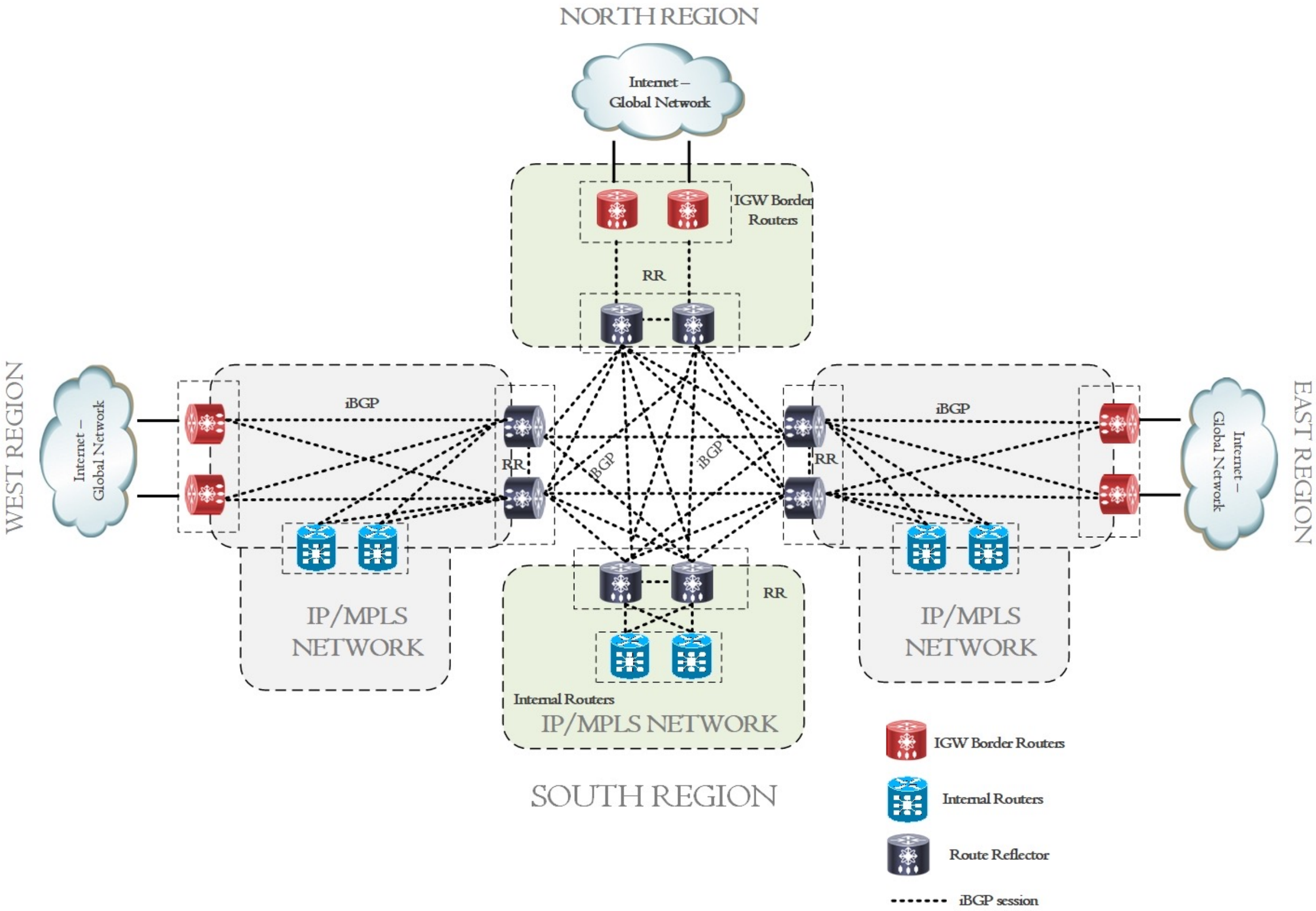



https://t.me/learningnets

# Distributed BGP RR Design




https://t.me/learningnets




https://t.me/learningnets

# BGP Route Reflector Cluster

- Route reflectors create a hub and spoke topology from the control plane standpoint. RR is the hub and the clients are the spokes.

- RRs and RR Clients form a cluster.

- We should have more than one RR in a cluster for redundancy and load sharing

https://t.me/learningnets

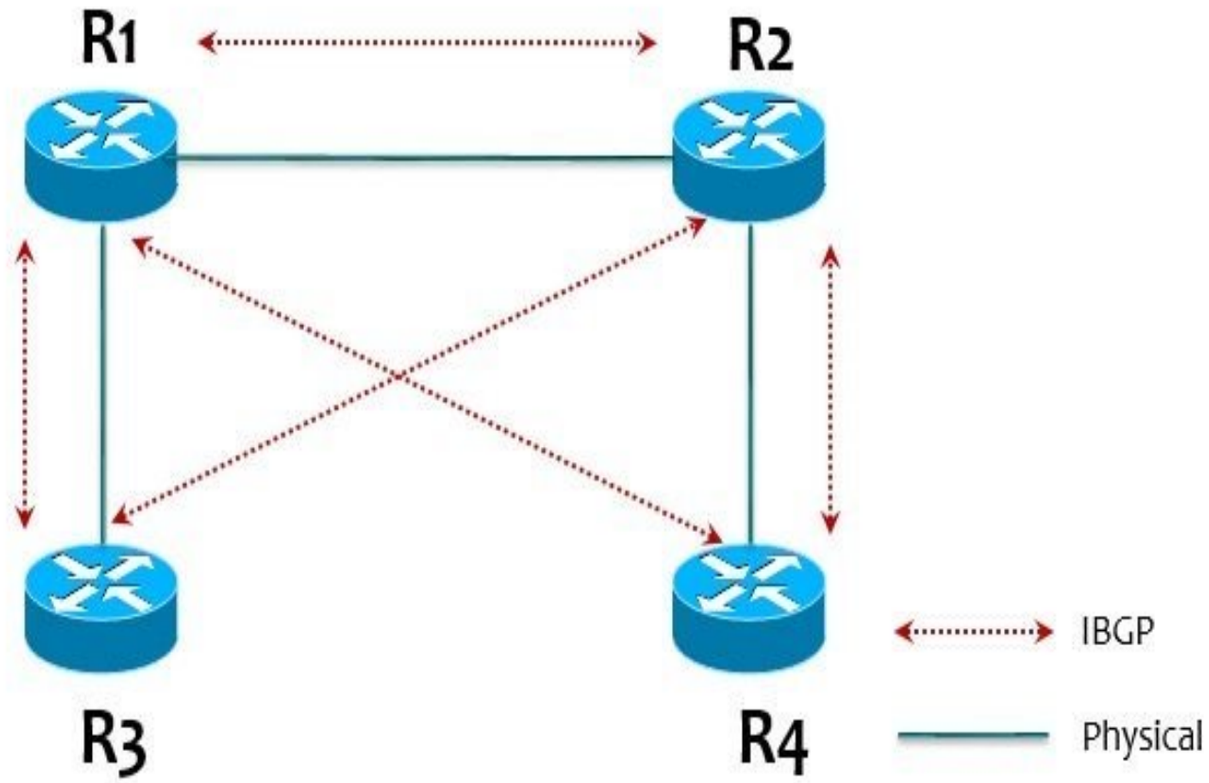
# BGP RR Cluster

Assume that we use both route reflectors as cluster ID 1.1.1.1 which is R1's router ID.

R1 and R2 receive routes from R4.
R1 and R2 receive routes from R3.

Both R1 and R2 as route reflectors appends 1.1.1.1 as cluster ID attributes that they send to each other.
However, since they use same cluster, they discard the routes of each other.

That's why, if RRs use the same cluster ID, RR clients have to connect to both RRs.




https://t.me/learningnets

- BGP Route reflector cluster is the collection of BGP Route reflector and Route reflector clients

https://t.me/learningnets

- The RR uses Cluster ID for the loop prevention RR Clients don’t know which cluster they belong to

- Using same BGP Cluster ID is good for resource consumption but bad for fast convergence

https://t.me/learningnets

# Changing the BGP Route Reflector Behavior via BGP Shadow RR and BGP Add-Path

- If you want to send more than one best path by the BGP Route Reflectors  for multi pathing or fast reroute purpose then below are the approaches.

- Unique RD per VRF per PE. Unique Route Distinguisher is used per VRF per PE. No need, Add-Path, Shadow RRs, Diverse Paths. But only applicable  in  MPLS  VPNs.

- BGP Add-Path

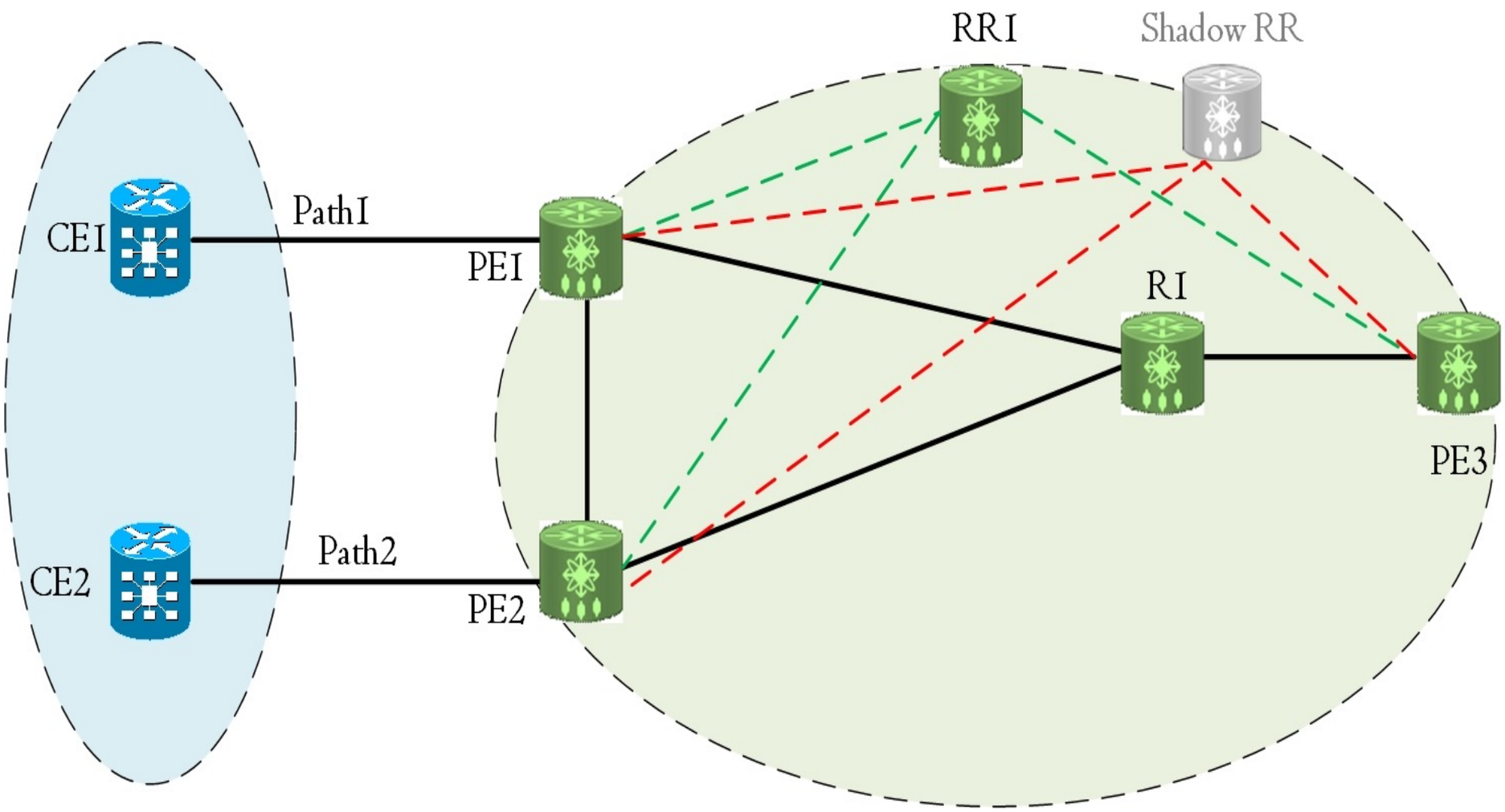
- BGP Shadow Route Reflectors

- BGP Shadow Sessions

https://t.me/learningnets

# SHADOW ROUTE REFLECTORS

**Shadow Route reflectors; you have two Route reflectors, one route reflector sends best path, second one calculate the second best and sends the second best  path.**




https://t.me/learningnets

# BGP SHADOW ROUTE REFLECTORS

- Shadow Route reflector deployments don’t require MPLS in the network

- Shadow RR is using a different RR node to advertise the second best path, Shadow Session approach is the same as Shadow RR, except no need to have a separate box , using the same box but maybe having a virtual context on the RR node for the Shadow sessions

https://t.me/learningnets

# BGP SHADOW ROUTE REFLECTORS

With Shadow Route Reflector second IBGP session is created between Shadow RR and PE3. Over the second IBGP session, second best is sent. This session is called shadow Route reflector sessions.

https://t.me/learningnets

# 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

https://t.me/learningnets

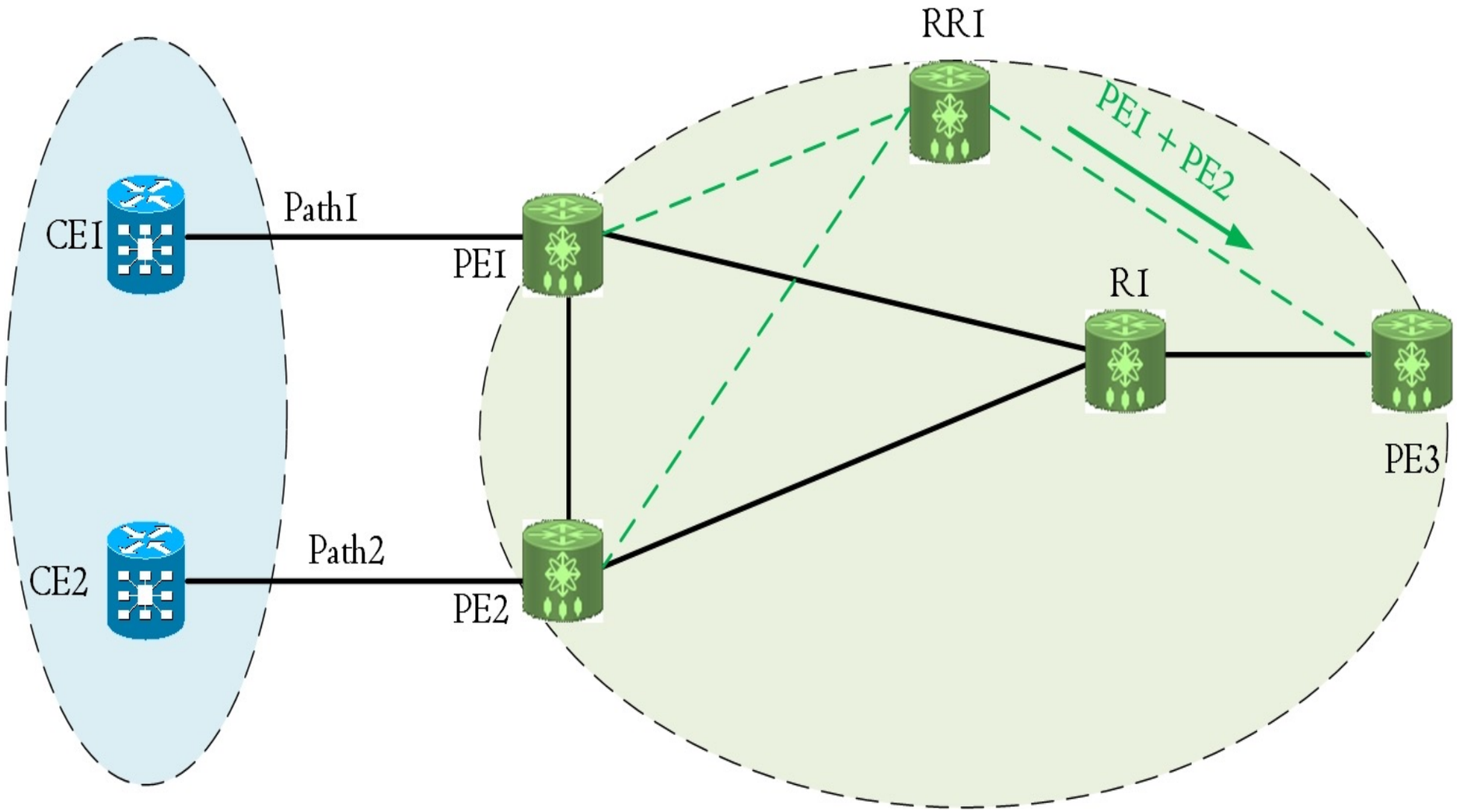
# BGP Add-Path

- In IBGP, if multiple paths are sent over the same BGP session, last one is kept by the receiving BGP speaker, because for the first one implicit withdrawn is sent, if route is completely gone then BGP Explicit withdrawn is sent

- With Add-Path, withdrawn is not sent thus receiving BGP router keeps all the paths and can make a best path selection based on its own view

https://t.me/learningnets

# BGP Add-Path




https://t.me/learningnets

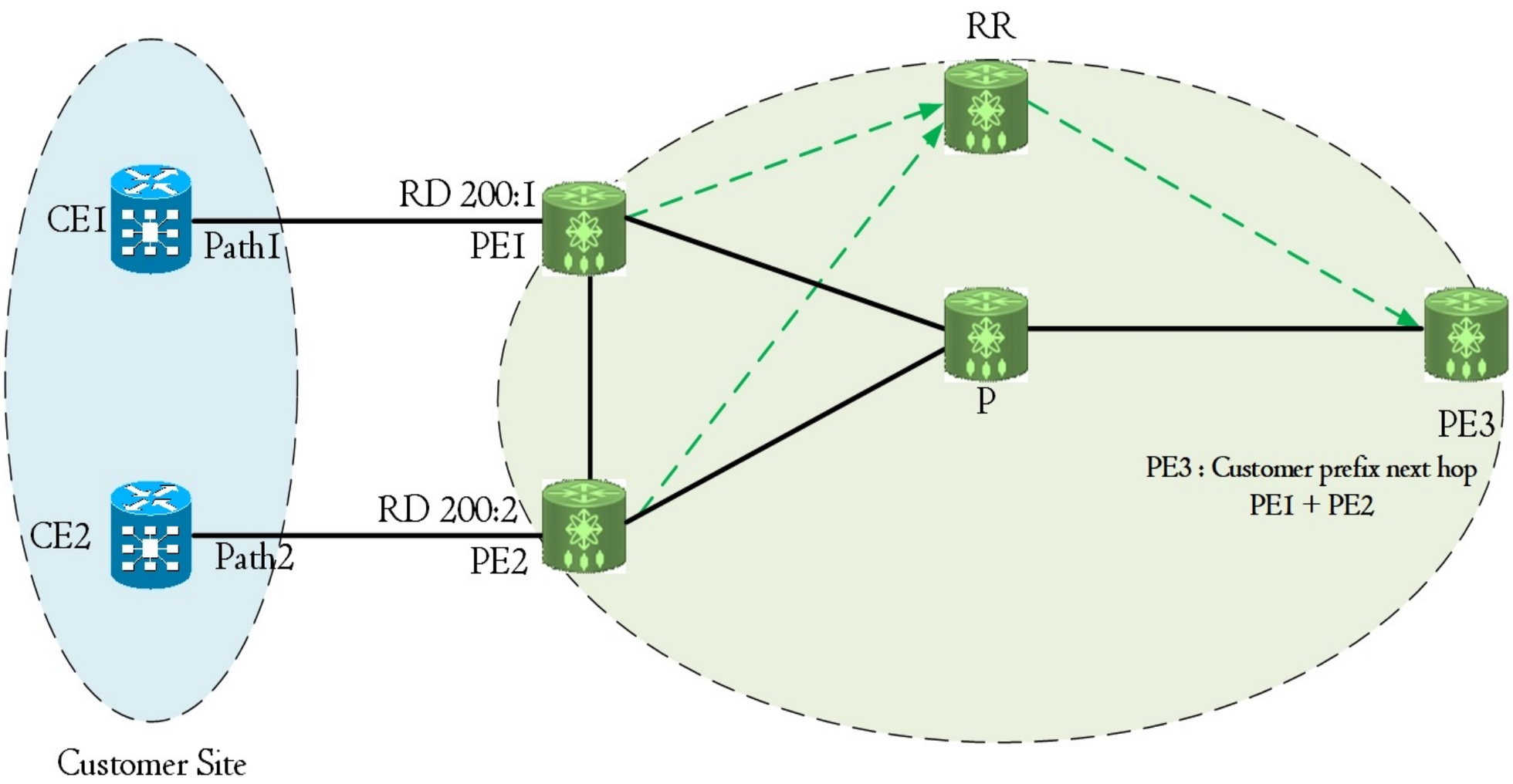
# BGP Route Reflector behavior in MPLS VPN Using Unique RD per VRF per PE Approach

- When RR is used in MPLS VPN, unique RD (Different RD) is configured on the PE routers to advertise unique VPN prefixes

- When VPN RR receives the prefixes, since there will be unique VPN prefix for the same customer prefix, RR doesn’t perform best path selection, it reflects both prefixes with their own next hops

https://t.me/learningnets

# BGP Route Reflector behavior in MPLS VPN Using Unique RD per VRF per PE Approach




https://t.me/learningnets

# Comparison between BGP Add-path, Shadow RR, Shadow Sessions and Unique RD per VRF per PE

| Design Concern | BGP Add Path | BGP Shadow RR | BGP Shadow Sessions | MPLS Unique RD per PE per VRF |
|---|---|---|---|---|
| Best in MPLS | No | No | No | Yes |
| How many IBGP Session between RR and RR-Client | One IBGP session, Path IDs are different for different next-hop | One session per Route reflectos. If there is only one more Shadow RR which sends second best path, two IBGP sessions on the RR Client, one for each RR | One session per next-hop. Only one RR but multiple separate IBGP session is required between RR and RR Client | One IBGP session between VPN RR and RR Client, different RDs make the same IP prefixes unique |
| Resource Requirement | Best | Worst, requires separate RR and IBGP session per next-hop | Better than Shadow RR because doesn't requre separate Route reflector, worse than ADD path because require extra IBGP session per next-hop | Same as Add-path, doesn't require extra IBGP session or Route Reflector |
| Migration of existing Route Reflectors | Very hard,all Route Reflectors and clients need to be upgraded to support Add-path | Easy, only Route Reflector code needs to be upgraded | Easy,only Route Reflector code needs to be upgraded | Easiest because there is no upgrade on any device. Only unique/separate Route Distinguisher needs to be configured on the Pes per VRF |
| Standard Protocol | Yes IETF Standard | Yes IETF Standard | Yes IETF Standard | Yes IETF Standard |
| Stuff Experince | Not well known | Not well known | Not well known | Known |
| Troubleshooting | Hard,default behaviour of BGP which is advertising only one best path is changing. Operation stuff needs to learn new troubleshooting skill | Easy | Easy | Easy |
| IPv6 Support | Yes | Yes | Yes | Yes |
| Provisioning | Easy, only one IBGP session between Route reflector and the client | Hard, one IBGP session per next-hop | Hard, one IBGP session per next-hop | Easiest,only the consideration is to have unique RD per VRF per PE |

https://t.me/learningnets

# BGP RR Benefits

- **Main benefit of BGP Route Reflector is Scalability**

- BGP Route Reflector reduces the total number of BGP sessions in the network and also reduces the number of BGP session per router

- BGP RR simplifies the configuration of the BGP routers

https://t.me/learningnets

# BGP RR Benefits

- Route Reflector hides the available paths. This is benefit for some networks, problem for the others

- BGP RR provides RBAC Opportunity, Low level engineers can maintain the RR Client and only Advanced level engineers can touch the Route Reflectors

https://t.me/learningnets

# BGP RR Problems

- It hides the path. In the previous slide, we mentioned it is as benefit. If resource utilization is concern, it is benefit, other than that it is a problem for suboptimal routing and other requirements

- BGP RR prevents Fast Reroute which might be requirements for some networks

- BGP RR increases Control Plane Convergence time

https://t.me/learningnets

# BGP RR Problems

- BGP RR can create sub optimal routing

- BGP RR can be a single point of failure if it is not designed correctly

https://t.me/learningnets

# BGP Add-Path and BGP ORR Requirement

- Reducing routing churns via oscillation, faster convergence, better load sharing and availability are some advantages of BGP Add-Path

- Improved Path Diversity is another benefit from this solution, which can bring effective BGP level load and fast connectivity restoration (ex. BGP PIC - Prefix Independent Convergence for faster convergence-FRR)

https://t.me/learningnets

# Memory consumption on the edge devices with BGP Add-Path

- By expanding the network to more exit point peering connections, which can result in getting same routes from more peers (especially when receiving full routing tables)

- More paths and lots of updates are advertised to clients, so the number of BGP announcements will increase for Route Reflector clients, which might lead to significant memory problems on the edge devices.

https://t.me/learningnets

- Introducing a large number of BGP states to all routers will create a lot of entry on the Route Reflector clients BGP Table. Some clients might not support Add-Path, others that support, might not have enough capacity.

https://t.me/learningnets

# How can optimal routing with BGP can be guaranteed?

- Add-path is a BGP capability, which mean it needs to be agreed between RR and RR Client. Upgrading both RR and RR Client might take so much time to migrate the BGP Software to one which supports BGP Add-Path feature as there might be so many Edge device in the network that acts as RR Client

- If all available next hops won’t be advertise how can optimality can be guaranteed?

- ANSWER is BGP ORR (Optimal Route Reflection)

https://t.me/learningnets

# BGP Optimal Route Reflection – BGP ORR

- Optimal Route Reflection is a IETF Draft but there are many vendor implementation as of 2019

- Based on this solution, the RR will do the optimal path selection based on each client's point of view. It runs SPF calculation with their clients as the root of the tree and calculates the cost to the BGP next-hop based on this view

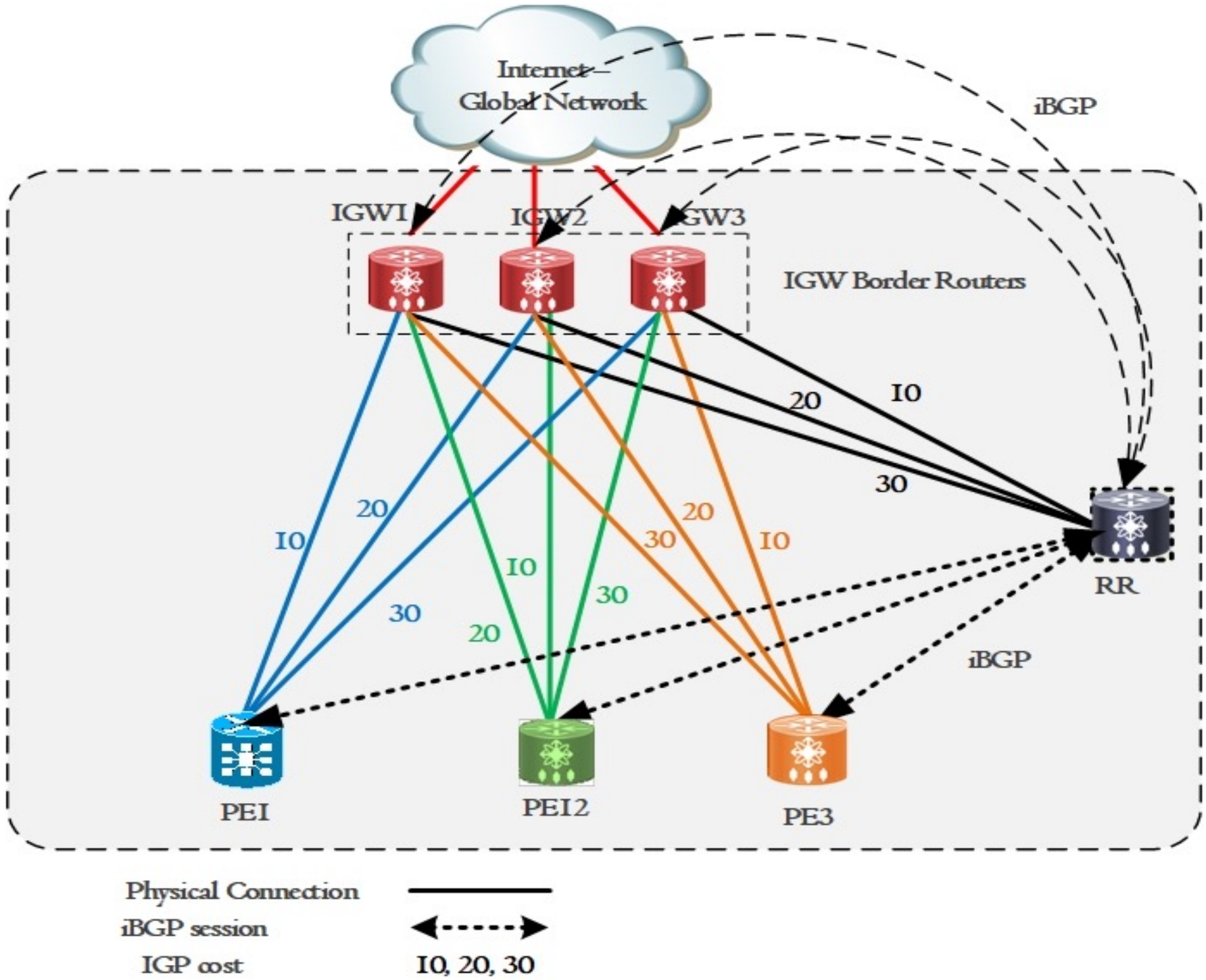
https://t.me/learningnets

# BGP Optimal Route Reflection – BGP ORR

- From the logical point of view, the Route Reflector position is virtualized, making it independent of its RR-Clients

- With ORR Route Reflectors location would be independent from the selection process of the best-path. Each ingress BGP border router can have a different exit point to the transit providers, for the same prefix for example

https://t.me/learningnets




https://t.me/learningnets

# Requirements for BGP ORR

- Link-state routing protocol is required in the network for the Route Reflectors to have a complete view of the network topology based on the IGP perspective. No changes are required to be done by the clients

- ORR is applicable only when BGP path selection algorithm is based on IGP metric to BGP next hop, so the path will be the lowest metric for getting the Internet traffic out of the network as soon as possible

https://t.me/learningnets

# BGP ORR is not an Alternative but Complementary to the BGP Add-Path

- This solution is not an alternative to BGP Add-Path or other methods for Path Diversity, though it is an alternative to provide optimal routing

- It can be used together to improve the quality of multiple advertisements, to propagate the route that can be the best path. Also, it can add resiliency and faster re-convergence for the network. For example, by receiving 4 paths from exit point peers across the network, it will choose the best path plus the 3 other paths based on the IGP cost. So, it’s a true way to add resiliency through add-path

https://t.me/learningnets

# How BGP ORR Works?

- With ORR, at the 1st step, the topology data is acquired via ISIS, OSPF, or BGP-LS. The Route Reflector will then have the entire IGP Topology, so it can run its own computations (SPF) with the client as the root. There could be as many rSPF (Reverse SPF) run based on the number of RR clients, which can increase the CPU load on the RR

- So, a separate RIB for each of the clients/groups of clients is kept by the RR. BGP NLRI and next-hop changes trigger ORR SPF calculations. Based on each next-hop change, the SPF calculation is triggered on the Route Reflector

https://t.me/learningnets

# How BGP ORR Works?

- The Route Reflectors should have complete IGP view of the network topology for ORR, so a link-state routing protocol is required to be used in the network. OSPF/IS-IS can be used to build the IGP topology information

- IGP is great for link state distribution within a routing domain or an autonomous system but for link state distribution across routing domains EGP is required. BGP-LS provides such capability at high scale by carrying the link state information from IGP protocols as part of BGP protocol messages

https://t.me/learningnets

# How BGP ORR Works?

- Route Reflectors keeps track of which route it has sent to each client, so it can resend a new route based on changes in the network topology (BGP/IGP changes reachability). The Route Reflector function is 1 process per route but the ORR function is 1 process per route per client router

- ORR brings the flexibility to place the Route Reflector anywhere in the topology, which provides Hot Potato Routing, supports resiliency via ORR Groups, requires no support from clients and finally brings much better output when is used with ADD-PATH

https://t.me/learningnets

# Different types of ORR (Optimal Route Reflection) Deployments

1. Optimal BGP path selection based on client IGP perspective

2. Optimal BGP Path Selection Based on Policy

https://t.me/learningnets

# 1. Optimal BGP path selection based on client IGP perspective

- Optimal BGP path selection is done Based on the Client's IGP Perspective, and not the RR's IGP perspective. To reduce the SPF calculation overhead on the RR, Optimization such as partial and incremental SPF can be used

https://t.me/learningnets

# 2. Optimal BGP Path Selection Based on Policy

- This solution is based on User Defined Policy. The clients will always send traffic to a specific exit point of the network regardless of how the topology looks like

- For example, one of the Policy methods can be using for the customers who pay more and gets SLA (Can be classified and marked with BGP Communities), so the traffic can be sent to particular Internet region and particular Transit Operator, instead of doing Hot Potato routing

https://t.me/learningnets

# Same or Different BGP RR for Different Services (Different BGP Address Families)

- For the different address families, different set of Route reflectors can be used, this avoids fate sharing

***For example if IPv4 RR is attacked, VPN customers may not be impacted if different sets of RR is used***

https://t.me/learningnets

# Same or Different BGP RR for Different Services (Different BGP Address Families)

- If you are using VPN Route reflectors , you can use multiple Route reflectors for different prefixes if scalability is a concern

- Based on Route Targets, we can use Route Reflector Group-1 to serve odd Route Target values, Route Reflector Group-2 to serve even Route target values

https://t.me/learningnets

# Route Target (RT) Constraints

- If you are using VPN Route reflectors , you can use multiple Route reflectors for different prefixes if scalability is a concern

- Based on Route Targets, we can use Route Reflector Group-1 to serve odd Route Target values, Route Reflector Group-2 to serve even Route target values

https://t.me/learningnets

# Route Target (RT) Constraints

- In this solution PEs send all the RT values to both Route Reflector Groups. They receive and process all the prefixes but based on odd/even ownership they filter the unwanted ones. But processing the prefixes which will be filtered anyway is not efficient way

- Instead Route Target Constraints should be deployed so Route reflectors can signal to the PEs which are route reflector clients their desired Route Target values

https://t.me/learningnets

# BGP Confederations

- RFC 5065 describes the use of Autonomous System Confederations for BGP

- BGP confederations help with this scalability issue by allowing the engineer to subdivide the autonomous system into smaller sub-autonomous systems

- There are generally two design methods when considering BGP confederations

https://t.me/learningnets

# Different BGP Confederation Designs

- Same IGP (OSPF , IS-IS , EIGRP etc.) in each Sub-AS
- Different IGP in the sub-AS
- There are pros and cons of each method as usual
- Implementing BGP confederation significantly reduces the total number of BGP sessions

https://t.me/learningnets

# Different BGP Confederation Designs

- Implementing BGP confederations involves quite a change to BGP configurations and the architecture itself, adding more complexity to achieve stable and scalable BGP design

- Migrating a network to a BGP confederation will be disruptive. Routers that are part of a sub-AS will need to change their BGP configuration to use the sub-AS instead of the real AS numbers

https://t.me/learningnets

# How BGP Confederation Works

- BGP routers within a sub-AS peering are IBGP peers
- BGP routers in different sub-AS are EBGP peers which means that the AS number is prepended when an update travels between the sub-AS
- If a router has to send an update towards its IBGP neighbor within a sub-AS, it will not change the AS_PATH attribute
- BGP between the sub-ASs is called as intra-confederation EBGP

https://t.me/learningnets

# BGP Confederation Route Preference – Best Path Selection

- EBGP routes that are exchanged between the sub-ASs are also known as confederation external routes, which are preferred over IBGP routes when it comes to best path selection

- If BGP has to choose between two paths to the same destination, one path leading inside the sub-AS, and another outside the sub-AS but within confederation, it will choose the external path – towards the neighboring sub-AS.

- If it has to choose between confederation EBGP route and EBGP route that leads outside the confederation, BGP will choose the second one

https://t.me/learningnets

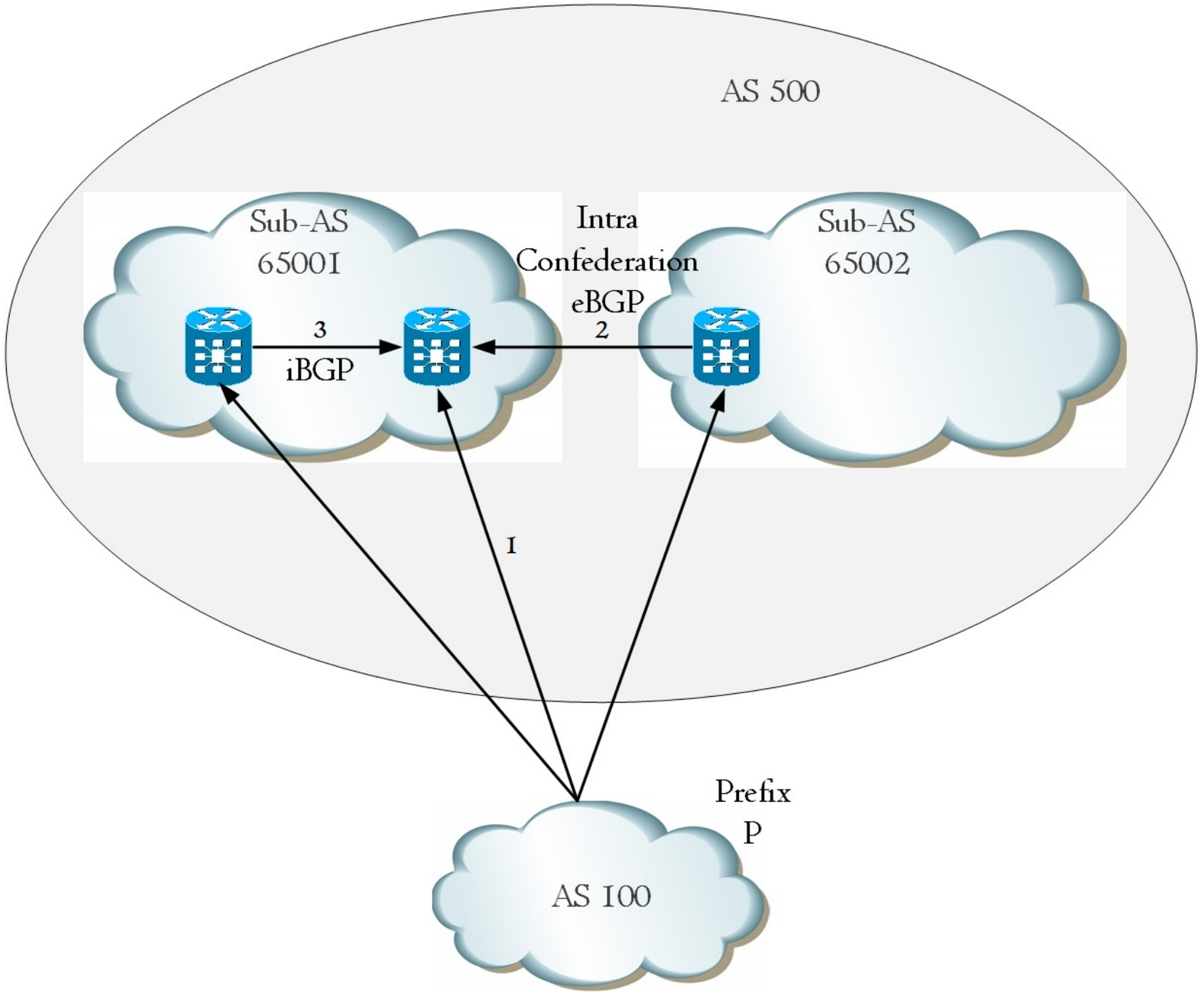
# BGP Confederation Route Preference – Best Path Selection

If same prefix learned over real EBGP , intra-confederation EBGP and IBGP within sub-AS, preference will be:

1. Real EBGP Connection (Confederation AS- ID)
2. Intra-Confederation Connection
3. IBGP Connection (Route is learned from an IBGP neighbor within sub-AS)

https://t.me/learningnets

# BGP Confederation Route Preference – Best Path Selection




https://t.me/learningnets

# How BGP Confederation Works

- Between Confederation EBGP and Real EBGP, there are some differences

- With BGP confederation EBGP session, MED, local preference and the next-hop are sent unmodified, this is similar to how IBGP works, but AS-Path attribute is changed

https://t.me/learningnets

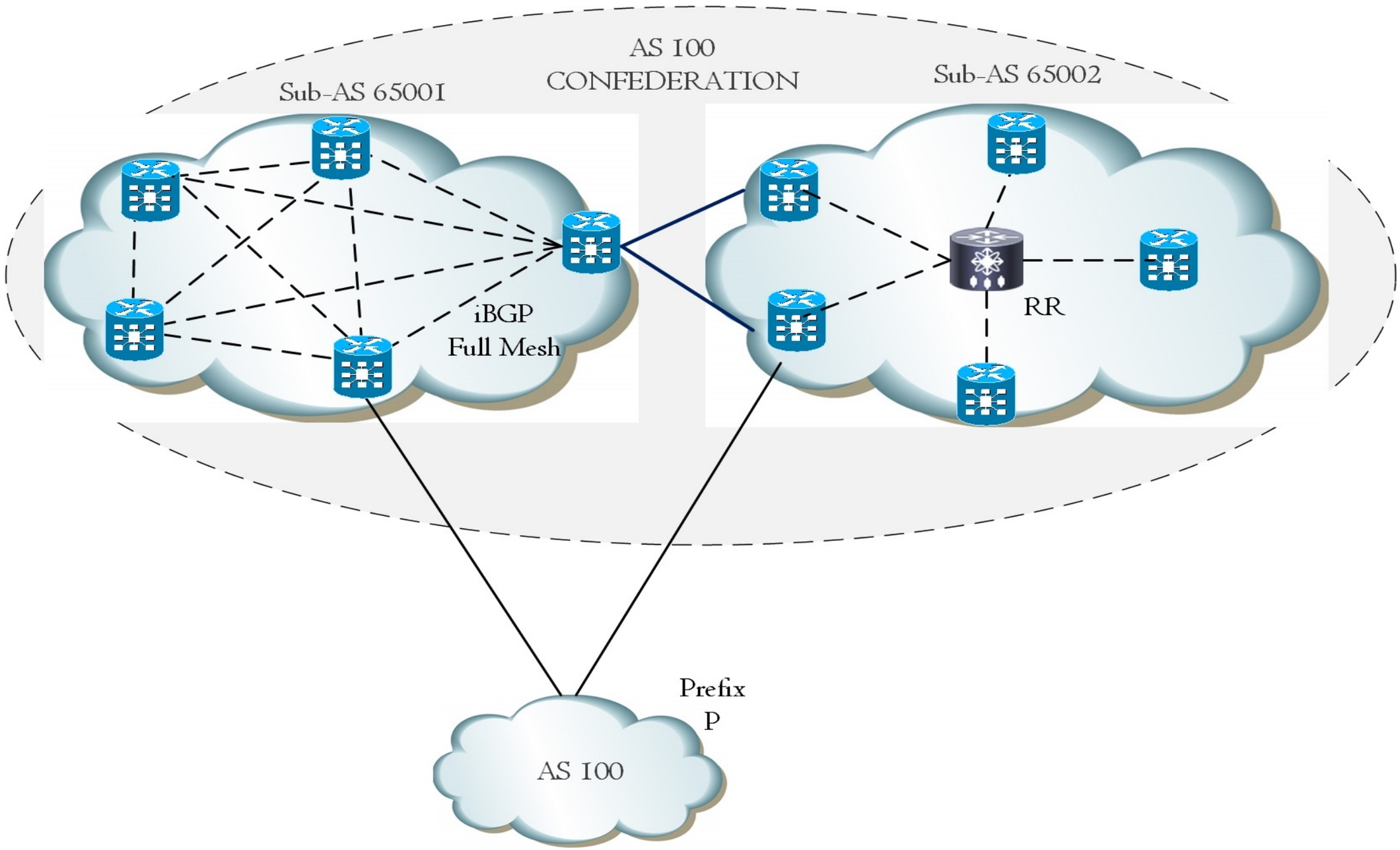
# How BGP Confederation Works

- Confederation sub-ASs exchange routing information as if they are using IBGP, and the only attribute that changes is AS_PATH. In other words, EBGP behaves like IBGP when implemented inside a confederation

- Because the next-hop is sent unchanged, either an IGP needs to run across the entire confederation or the border routers need to set the next-hop to themselves

https://t.me/learningnets

# BGP Confederation




https://t.me/learningnets

# How BGP Confederation Works

- To make it appear as one AS to all real EBGP peers, the sub-AS in the AS path need to be stripped when sending updates to its peers

- One of the advantages of running a confederation is that a policy can be applied for the sub-AS which does not apply for the entire real AS

- For example prefixes can be sent between the BGP peers within sub-AS with the NO_EXPORT_SUBCONFED community and policy can be distributed within the sub-AS but not outside the sub-AS

https://t.me/learningnets

# BGP Confederation Routing Loop Avoidance

- An EBGP connection between sub-ASs also serves as kind of a loop-avoidance mechanism

- If in the previous topology route is learned from AS 64512 is advertised somehow back to AS 64512, routing update is not accepted by the Originator AS

- This is done with an AS_CONFED_SEQ parameter inside AS-Path Attribute

https://t.me/learningnets

# Routing loop avoidance in BGP Confederation

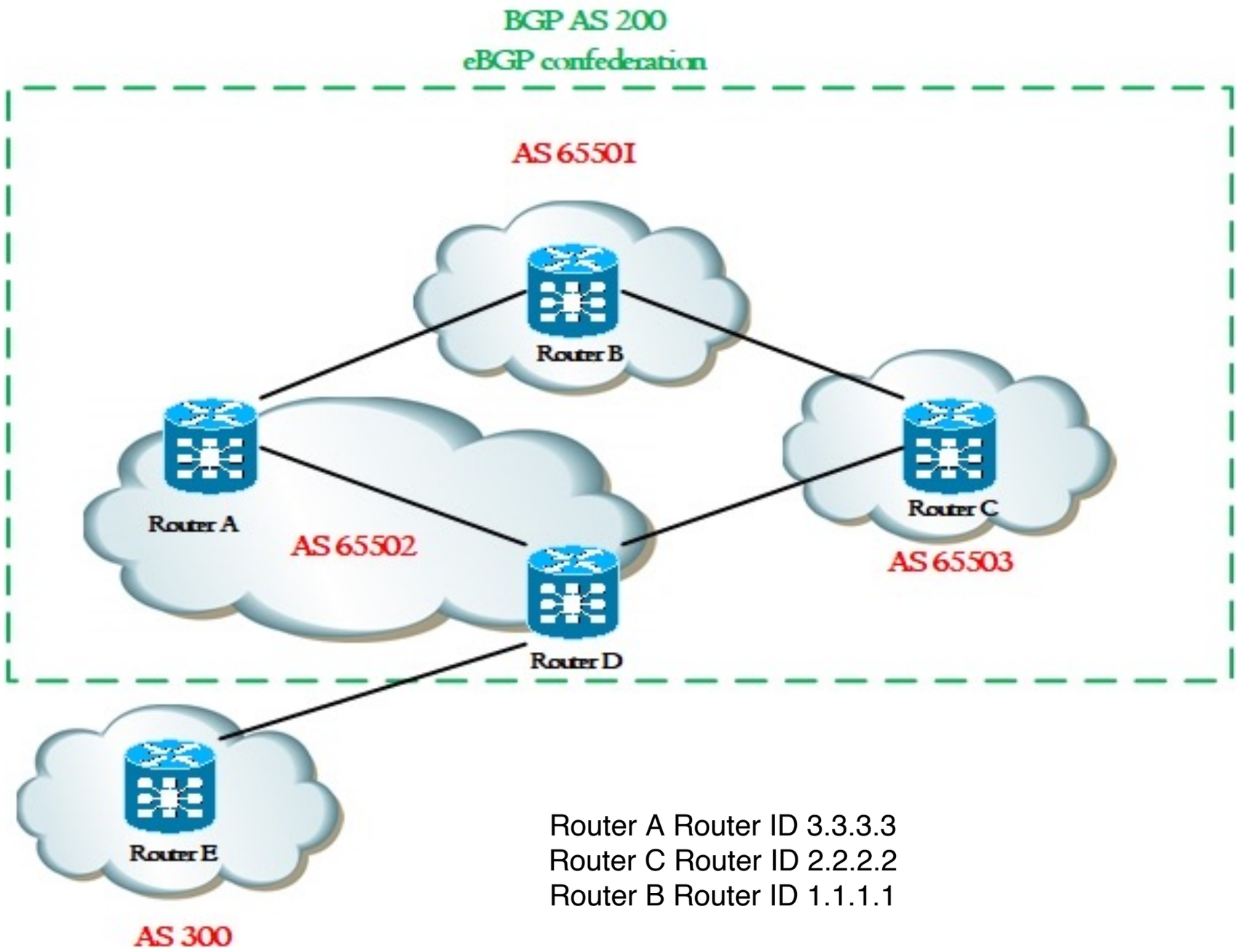
- Based on RFC 5065 section
  When comparing routes using AS_PATH length, CONFED_SEQUENCE and CONFED_SETs SHOULD NOT be counted

- BGP is using the AS_CONFED_SEQ portion of the AS path attribute for routing loop control inside the confederation. However, it's not being used as criteria for BGP path selection inside the confederation

- Let’s have a look at next page for the example

https://t.me/learningnets

In the picture below all other BGP attributes are identical, BGP chose the path it received from the router with the lowest BGP router ID.  Since Router C has a lower router ID (2.2.2.2) than Router **A** (3.3.3.3), Router B choses the path through Router C as best, even though as-path length is longer through the Router C for the destination at AS 300 ( 2 AS vs. 1)




Router A Router ID 3.3.3.3
Router C Router ID 2.2.2.2
Router B Router ID 1.1.1.1

https://t.me/learningnets

# When to choose BGP Confederation instead of BGP RR?

- The main difference between BGP RR and Confederation is that a confederation may contain different IGPs, adding more flexibility to scaling your network

- Therefore, choosing a Confederation over BGP RR would be more appropriate in case your IGP is exceeding its scalability limit and becomes unmanageable, and you would like to manage many independent ASs, each of which may run a different IGP

https://t.me/learningnets

# Full mesh IBGP to RR Migration

## Drivers

- Number of IBGP sessions are too much
- Some devices cannot manage receiving the route from multiple next hops
- Less number of devices to touch when New BGP routers are added to the network

https://t.me/learningnets

# Full mesh IBGP to RR Migration

## Preparation Steps:

- Verify that BGP next hops are reachable via IGP ( Example Loopbacks)
- You may want to have OOB (Out of Band) Management access
- Schedule the migration during maintenance windows

https://t.me/learningnets

# Full mesh IBGP to RR Migration

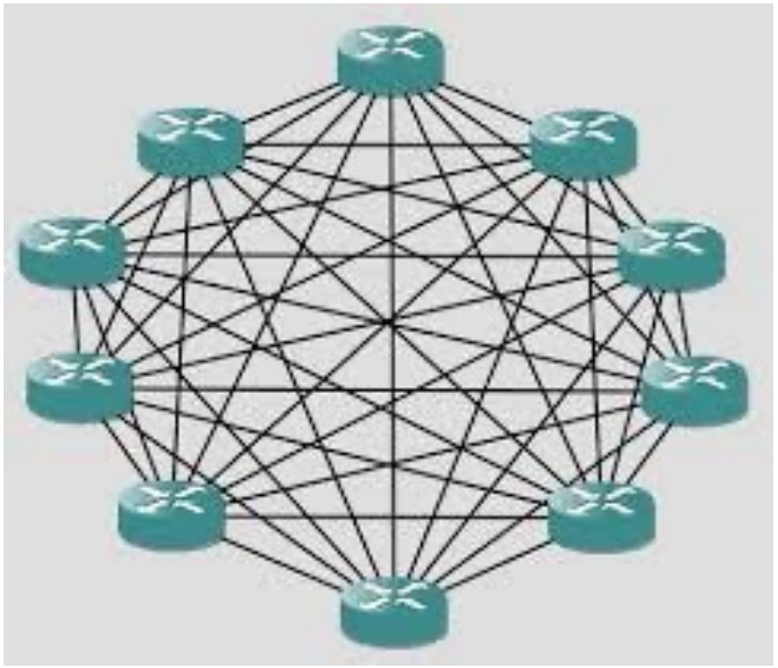
## Preparation Steps:

- Migrate one POP location during one maintenance window to minimize downtime risk
- Have a proper rollback plan and execute if things go wrong in time

https://t.me/learningnets

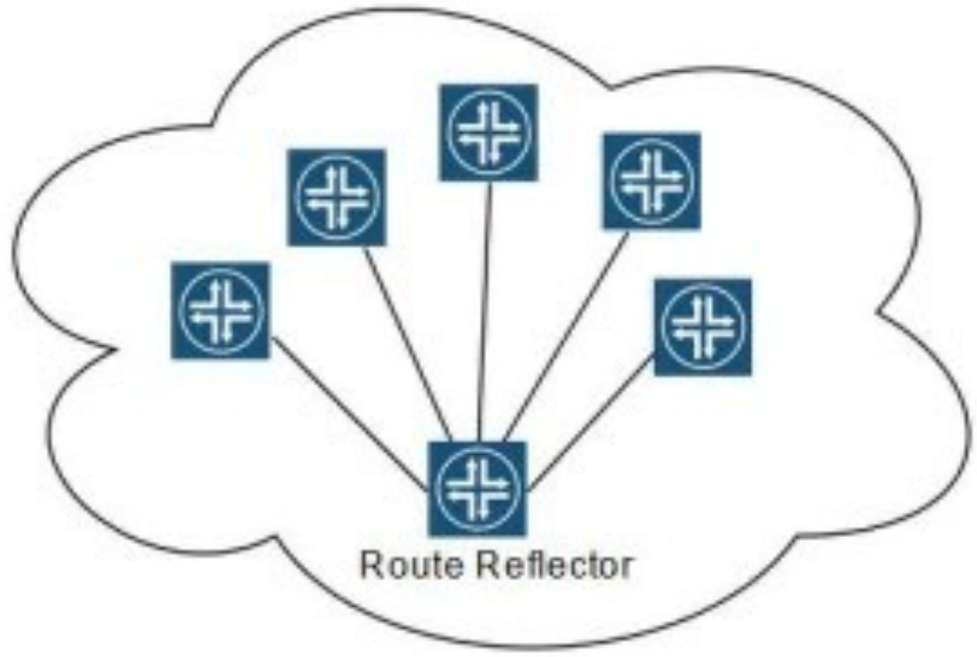
# Full mesh IBGP to RR Migration

BEFORE



(Full Mesh IBGP)

AFTER




(BGP Route Reflector)

https://t.me/learningnets

# Full mesh IBGP to RR Migration

## Migration Procedure

- Migration depends on if new routers will be deployed as RR or some of the existing routers will be migrated to become RR

- If new routers are added as an RR, downtime can be further minimized

https://t.me/learningnets

# Full mesh IBGP to RR Migration

- Let’s examine what happens when existing routers are used as an RR

- In this case, Core Routers are selected as an RR, they are migrated first

- Make sure there are two Core routers to minimize downtime

https://t.me/learningnets

# Full mesh IBGP to RR Migration

- First migrate the Core routers that will become Route Reflectors and then migrate RR client routers one at a time

- In order to do this, one of the two Core routers advertise into the IGP that it shouldn’t be used as a Transit (OSPF Max-metric , IS-IS Overload)

https://t.me/learningnets

# Full mesh IBGP to RR Migration

- When all edge BGP routers (RR Clients) are migrated to both of the RRs, their full mesh IBGP sessions should be removed

- After removing the Full Mesh IBGP sessions, BGP reachability to the prefixes should be verified again

https://t.me/learningnets

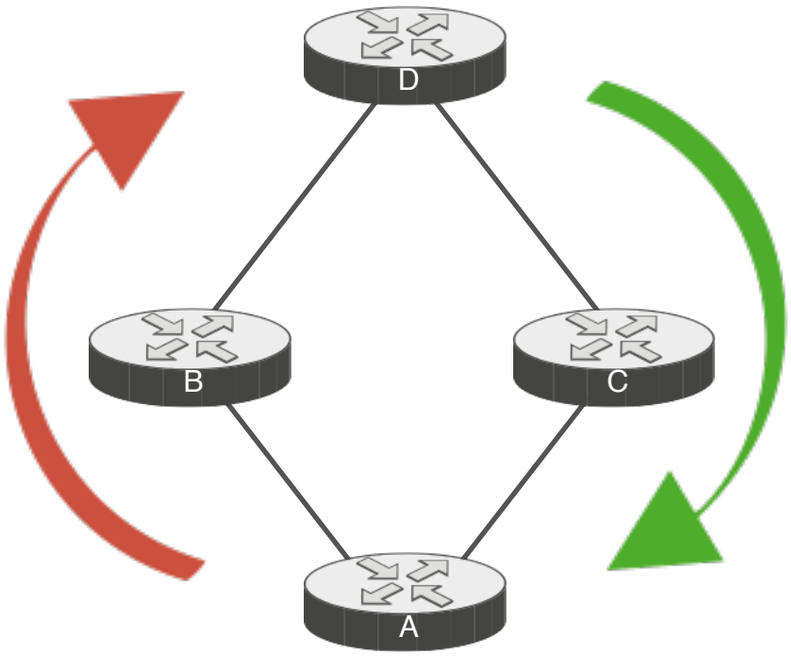
# BGP – IGP Interactions

- When a router needs to be restarted or maybe having a resource issue, operator may want to remove it from the network path

- In order to do this without losing packet, a router signal it’s IGP neighbor that they shouldn’t send the traffic towards it anymore

https://t.me/learningnets

# BGP – IGP Interactions




IS-IS Overload Bit
OSPF Max-Metric Router LSA
If BGP as an IGP , BGP Graceful Shutdown Community
EIGRP Stub Feature

Node B or C removal can create packet loss
if nodes don’t signal their neighbors with increased IGP metric

https://t.me/learningnets

# BGP – MPLS Interactions

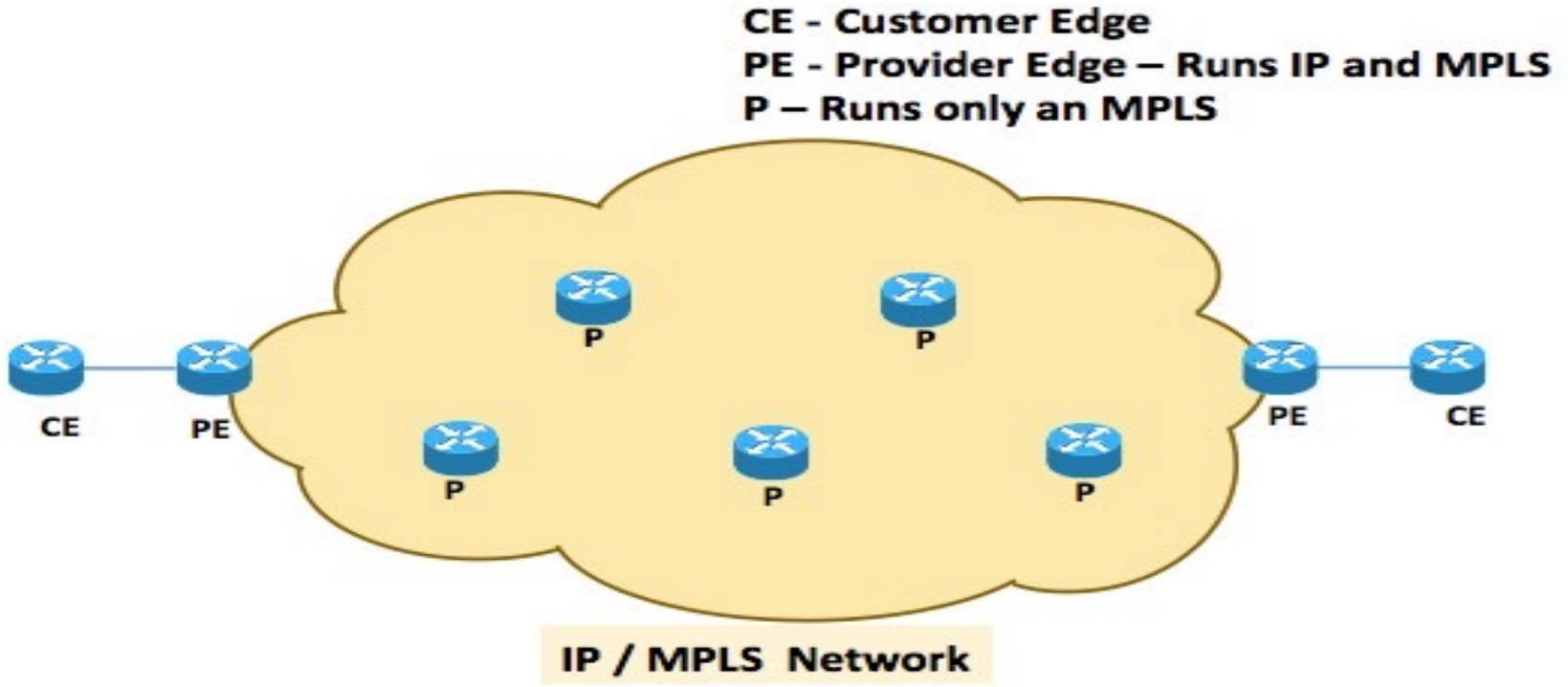
- When BGP and MPLS is used together, generally it is used for VPN services

- MPLS removes the requirement of having BGP in the Mid/Core routers

- This phenomena is known in design as ‘ BGP Free Core ‘ design

https://t.me/learningnets

# BGP – MPLS Interactions




**BGP Only needs to run at the PE routers in the above topology**

https://t.me/learningnets

# BGP Labeled Unicast (BGP LU)

- BGP LU (Labeled Unicast) is used in the multi domain networks to connect the domains or in a single domain to advertise MPLS labels for the BGP next hops

- It was specified first in RFC 3107 ‘ Carrying Label Information in BGP-4 ‘ and then updated with RFC 8277

https://t.me/learningnets

# BGP Labeled Unicast (BGP LU)

- Practical deployment cases with BGP LU is Inter-AS Option C , Carrier Supporting Carrier and Seamless/Unified MPLS Scenarios

- BGP LU is used both as an Intra AS (With Seamless MPLS) and Inter-AS Routing Protocol (Inter-AS Option C , CSC)

https://t.me/learningnets

# BGP Labeled Unicast (BGP LU)

- With BGP LU ,  BGP sends the IPv4 prefix + Label (Address Family Identifier (AFI) 1 and Subsequent Address Family Identifier (SAFI) 4) , which is different than sending a Label for the VPN prefix (Ex: MPLS L3 VPN)

- We will discuss Seamless MPLS , Inter AS MPLS VPNs and Carrier Supporting Carrier (CSC) in the MPLS lesson

https://t.me/learningnets

# BGP LS (Link State and TE Information Distribution using BGP

- RSVP-TE have been providing resource allocation and provide an LSP with the distributed path computation algorithm (CSPF) since decades

- In order to overcome Bin Packing , Dead Lock or Network wide optimal traffic engineering, centralized controllers have been used for a long time

- RFC 7752 specifies the details of **North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP**

https://t.me/learningnets

# BGP LS (Link State and TE Information Distribution using BGP

- PCE (Path Computation Element) is an SDN controller which provides optimal path computation in Multi Area and Multi AS (Autonomous System) deployments

- It requires Link State and Traffic Engineering attributes such as Link coloring , SRLG , reserved bandwidth etc., from the network

- Link state IGP protocols (OSPF, IS-IS) can be used for this purpose but they are considered chatty and non-scalable

https://t.me/learningnets

# BGP LS (Link State and TE Information Distribution using BGP

- BGP LS is used to distribute Link state information and traffic engineering attributes from the network nodes to the Centralized TE controller

https://t.me/learningnets

# BGP EPE (Egress Peer Engineering)

- Monetary cost , latency and packet loss are important parameters for the Quality of User Experience and Traffic Engineering can be done by optimizing any of the above parameters

- BGP NLRIs don’t provide the information about the cost, latency or loss of the path or exit point for the destinations

https://t.me/learningnets

# BGP EPE (Egress Peer Engineering)

- The data-plane interconnection link (NNI) and control-plane (eBGP) direct connection between two ASs allows Internet traffic to travel between the two, usually as part of a formal agreement called *peering*

- This peering can be settlement free based or settlement based (Ex : IP Transit)

https://t.me/learningnets

# BGP EPE (Egress Peer Engineering)

- The selection of the best exit link for a given destination prefix selection and the enforcement of this selection on a network is not simple task. This is because the decision for one prefix might impact other traffic by changing the utilization of the NNI link and potentially leading to overload

- This is an end to end traffic engineering requirement !

https://t.me/learningnets

# Traditional EPE and the Limitations

- Traditionally, SPs use a policy to manipulate the BGP attributes contained in NLRIs received from a peer. This policy-based manipulation is usually performed on the Egress ASBR, but sometimes also on a route reflector (RR) and the Ingress ASBR

https://t.me/learningnets

# Traditional EPE and the Limitations

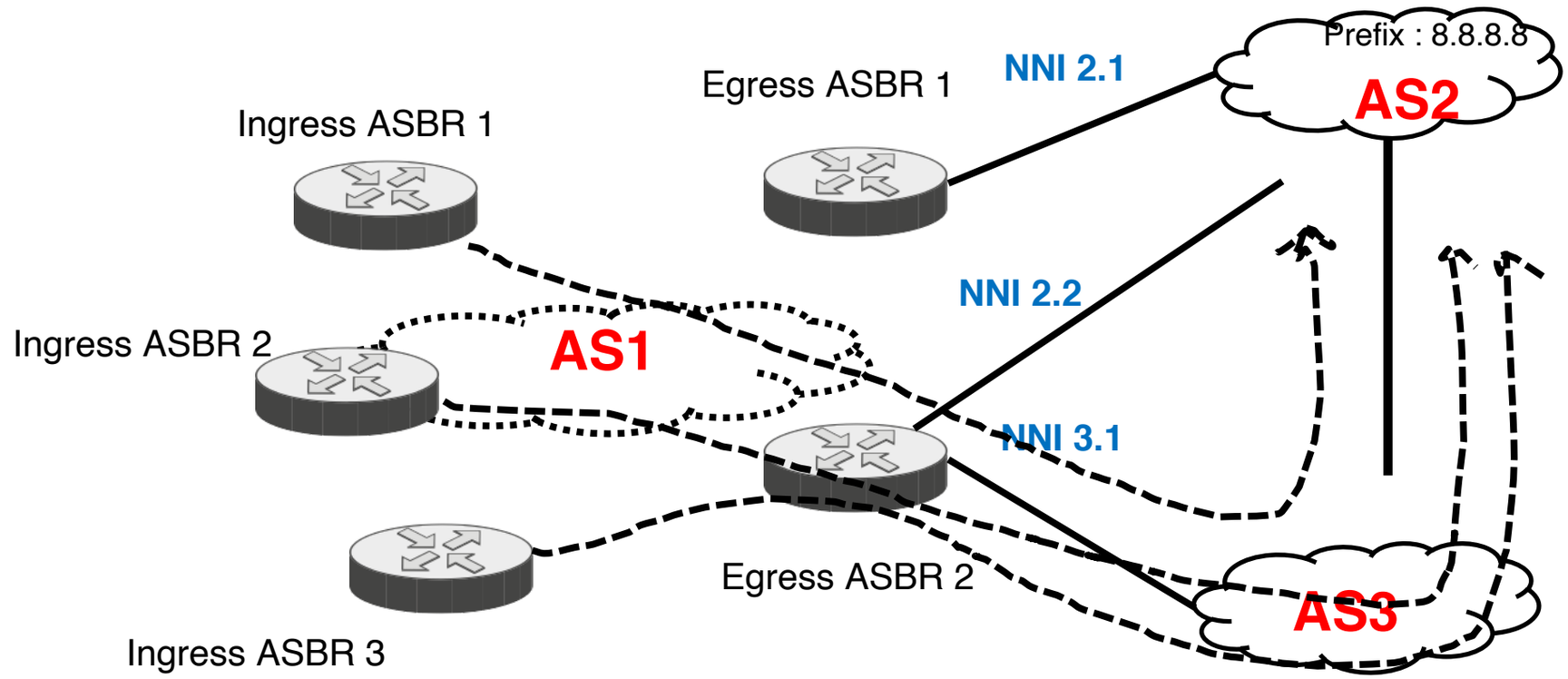
- This traditional technique provides some level of flexibility and control on how traffic leaves the SP and AS

- However, it is also limited by the BGP path selection algorithm and the fact that the results apply to all traffic for given prefix, regardless of the traffic's origin (Doesn't matter which Ingress ASBR sends the traffic)

https://t.me/learningnets

# Traditional EPE Example

**Assume 8.8.8.8 is in AS 2 , AS2 and AS3 are peer**
**AS1 learns the prefix from both AS2 and AS3**
**If on Egress ASBR2 Local prefix is higher for 8.8.8.8 towards AS3, NNI3.1 link is used by every Ingress ASBR**




https://t.me/learningnets

# Modern – Better way of EPE

- If EPE had the ability to distribute traffic among several egress links based not only on destination address, but also by considering the ingress ASBR (or ingress port etc.), this would provide much finer granularity and also bandwidth management

- This would be especially true if EPE were combined with traffic statistics and centralized optimization

https://t.me/learningnets

# Modern – Better way of EPE

- The EPE solution should direct traffic for a given prefix that enters the network on a particular Ingress ASBR to a particular egress NNI (Egress Link) on a particular Egress ASBR

https://t.me/learningnets

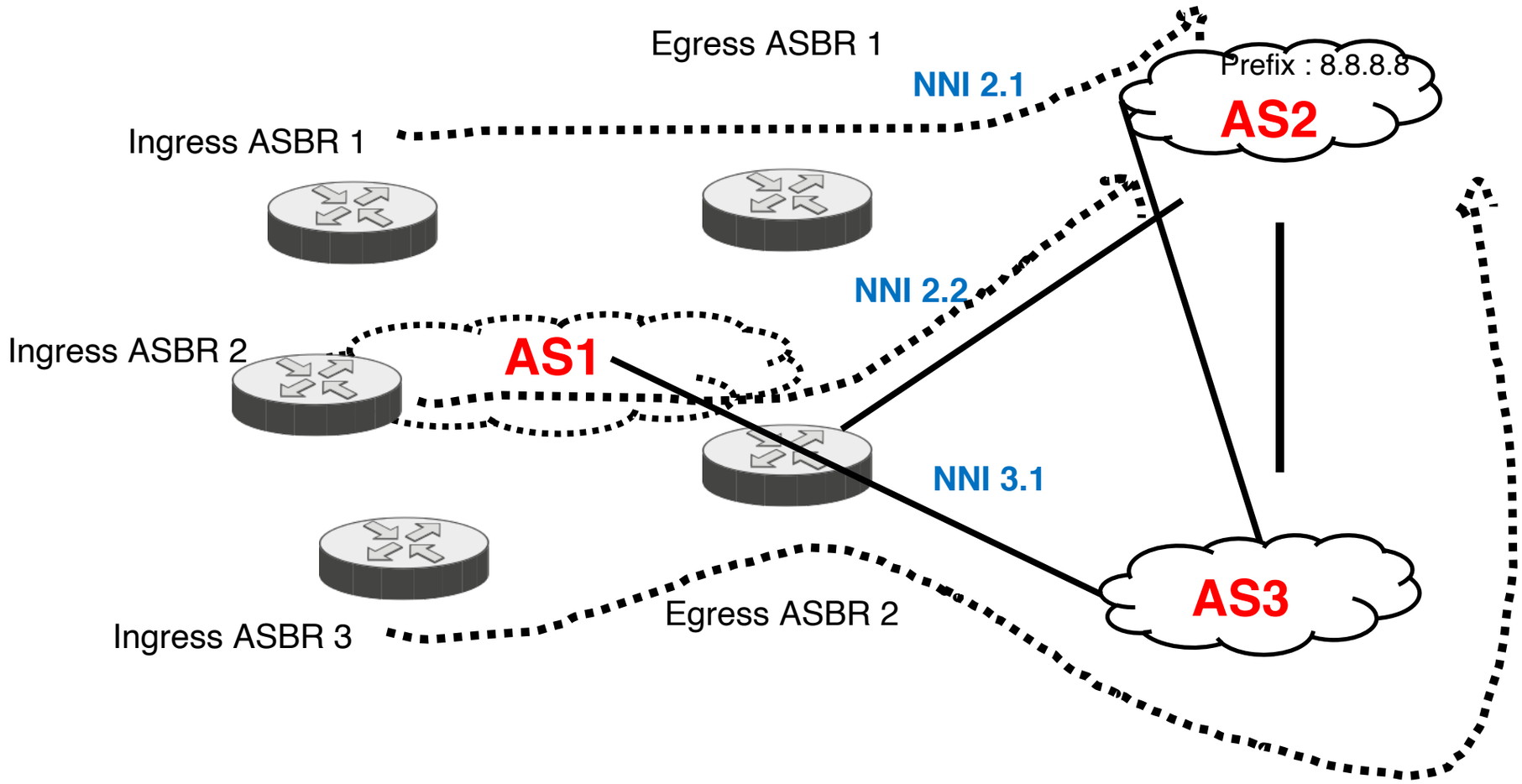
# Modern – Better way of EPE

Better Policy for utilization of both internal and inter-domain resources (NNI Bandwidth etc.) can be as follows for AS 1

- [Ingress ASBR1, 8.8.8.8 ]  to NNI 2.1
- [Ingress ASBR2, 8.8.8.8 ] to NNI 2.2
- [Ingress ASBR3, 8.8.8.8 ] to NNI 3.1

https://t.me/learningnets

# Modern – Better way of EPE




https://t.me/learningnets

# Modern EPE Requirements

- The utilization data for each of NNI. This data is provided by traditional or modern telemetry infrastructure (for example, SNMP interface statistics)

- The reachability information for destination IP prefixes This information is provided by eBGP advertisement from peer ASs

https://t.me/learningnets

# Modern EPE Requirements

- The fine-grain partitioning of egress traffic into “flows” with information about the traffic volume carried by each

- An Egress Peering Engineering (EPE) controller that executes some logic to map these “flows” to NNI in a globally optimal way (Distributed TE cannot provide global optimality, Bin Packing , Dead-Lock are known problems)

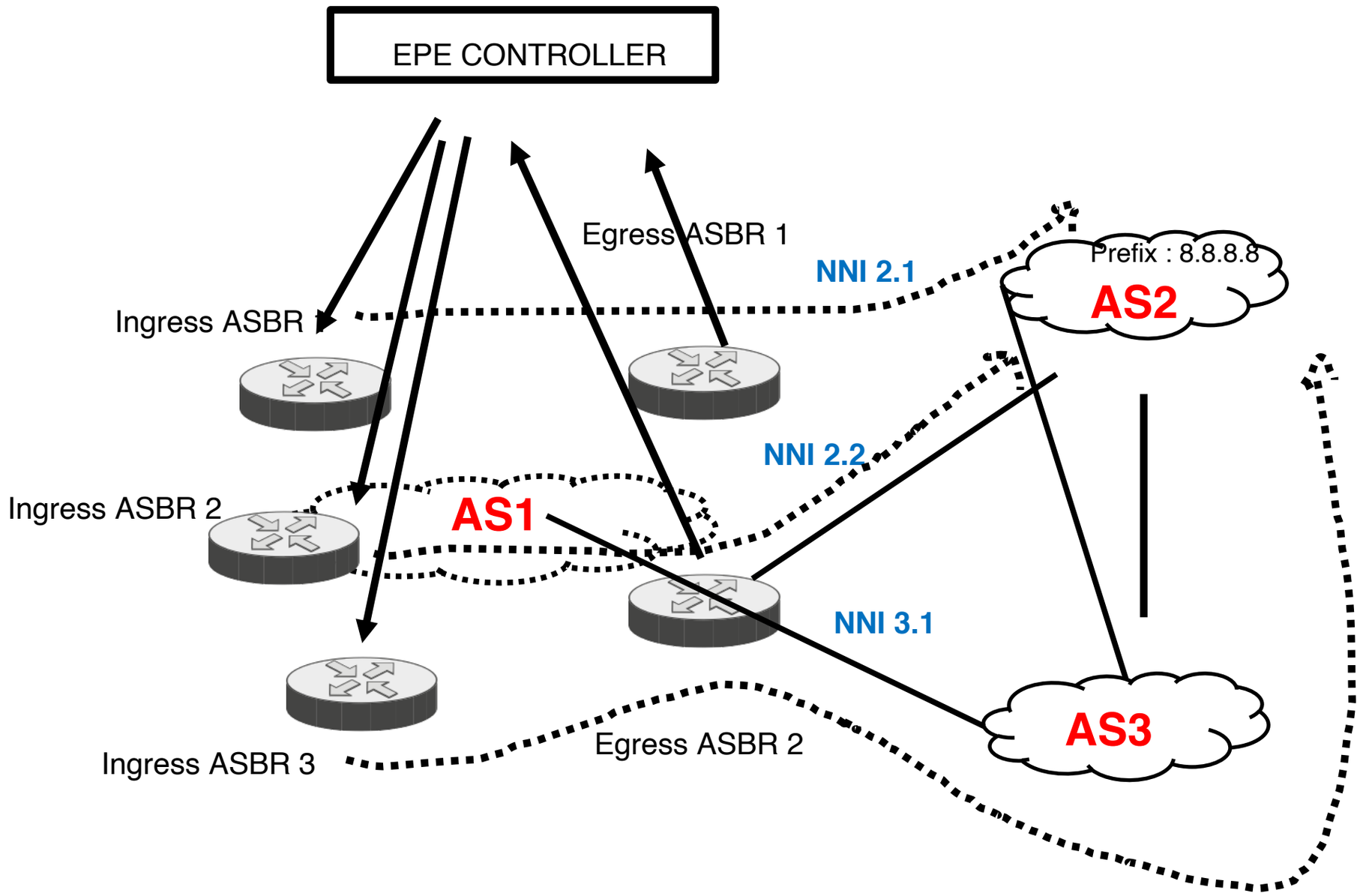
https://t.me/learningnets

# Modern EPE Requirements

- A network infrastructure that allows forwarding traffic from an ingress AS Border Router (ASBR) on a customer service provider network to the designated egress NNI, as determined by the EPE controller

https://t.me/learningnets

# Modern – Better way of EPE




https://t.me/learningnets

# BGP RTBH (Remotely Triggered Blackholing)

- Remotely triggered blackholing is used for DDOS prevention for a long time

- DDOS attacks have an economical impact

- According to NBC News article, More than 40% of DDOS Attacks cost $1 million per day

https://t.me/learningnets

# BGP RTBH (Remotely Triggered Blackholing)

- Remote Triggered Blackhole is a technique which is used to mitigate DDOS attack dynamically

- Before RTBH, customer used to call Operator when there is an attack, Operator NOC engineer used to connect to the attacked network, trace the source of the attack, place the filters accordingly and attack goes away

https://t.me/learningnets

# BGP RTBH (Remotely Triggered Blackholing)

- Manual operation is open to configuration mistakes, cannot scale in large networks and between the attack and the required action, services stay down

https://t.me/learningnets

# BGP RTBH (Remotely Triggered Blackholing)

There are two types of RTBH

- Destination based RTBH
- Source based RTBH

https://t.me/learningnets

# Destination BGP RTBH (Remotely Triggered Blackholing)

- First RTBH idea was Destination based RTBH

- With this technique, SP and the Customer agree on the discard community

https://t.me/learningnets

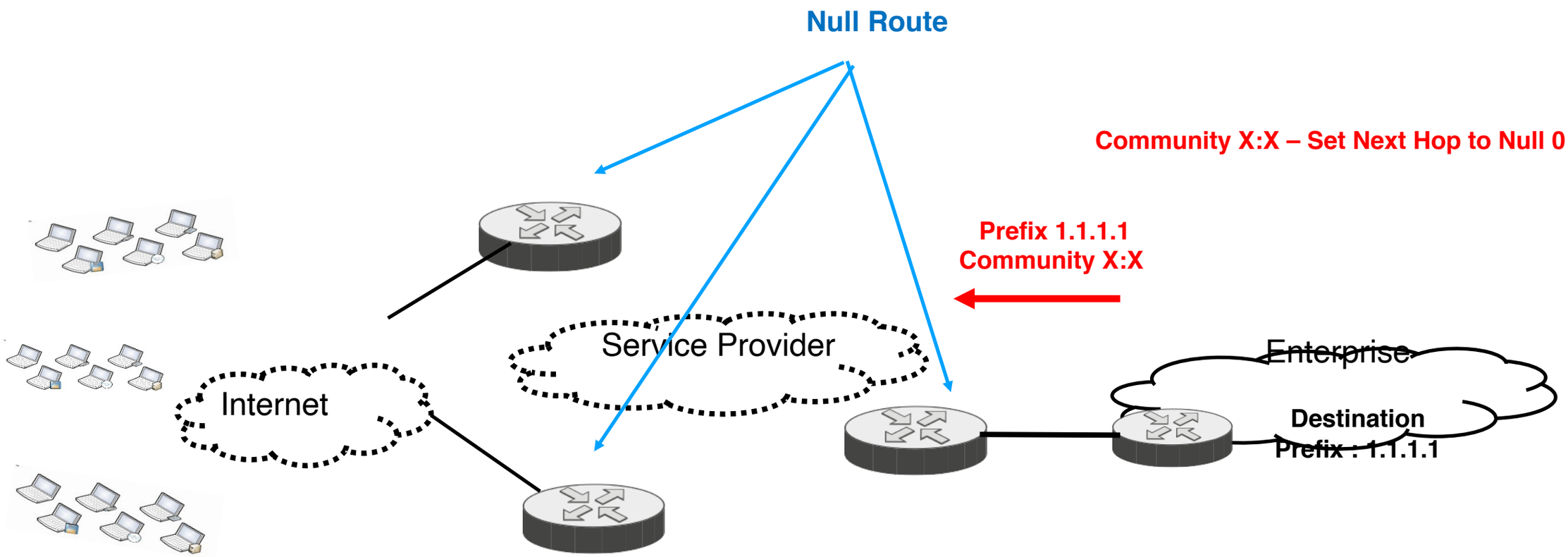
# Destination BGP RTBH (Remotely Triggered Blackholing)

- When there is an attack to the server, victim (customer) send the server prefix with the previously agreed community value

- When SP receives the update with that community, action is set to next hop to null, so packet is dropped before reaching to the customer link

https://t.me/learningnets

# Destination BGP RTBH (Remotely Triggered Blackholing)




https://t.me/learningnets

# Destination BGP RTBH (Remotely Triggered Blackholing)

- Problem with this attack, server will not be reachable from the legitimate sources too

- Attack is completed but at least the other services might stay up

- Also customer might change the IP address of the attacked server in DNS, which might take time to propagate this though

https://t.me/learningnets

# Destination BGP RTBH (Remotely Triggered Blackholing)

- RFC 3882 covers Destination based RTBH

- Better than manual processing

- Requires pre-configuration of null route on all edge routers in the SP network

https://t.me/learningnets

# Source BGP RTBH (Remotely Triggered Blackholing)

- RFC 5635 brings the idea of Source RTBH

- Instead of customer specifying the attacked system IP address to the SP, customer calls SP that they are under attack

- By combining uRPF and discard route (null route) configuration, based on the attack source, DDOS is mitigated (In theory)

https://t.me/learningnets

# uRPF and S/RTBH

- uRPF is Unicast Reverse Path Forwarding (Not Filtering☺) is defined in RFC 3704

- It is designed to limit the impact of DDOS attacks, by denying traffic with spoofed addresses access to the network

https://t.me/learningnets

# uRPF and S/RTBH

- Routers make their forwarding decisions based on Destination IP Address

- With uRPF, router looks at the Source IP address as well

- There are four types of uRPF : Strict,  Loose , Feasible Path and VRF mode

https://t.me/learningnets

# Strict and Loose Mode uRPF

- Routers look at Source IP and then the Routing Table

- If source is reachable via the input interface , then it is forwarded, otherwise packet is dropped , this mode is called Strict mode uRPF

- If source is reachable via any route in the routing table then it is forwarded, otherwise packet is dropped, this mode is called Loose mode uRPF

https://t.me/learningnets

# Strict and Loose Mode uRPF

- uRPF Strict mode is generally used for Single Homed Customers

- For Multihomed customers, traffic can come from different interfaces, thus uRPF Loose mode should be used

- For Source Based RTBH, Source addresses are injected into the BGP and with uRPF check traffic is sent to the blackhole (Null)

https://t.me/learningnets

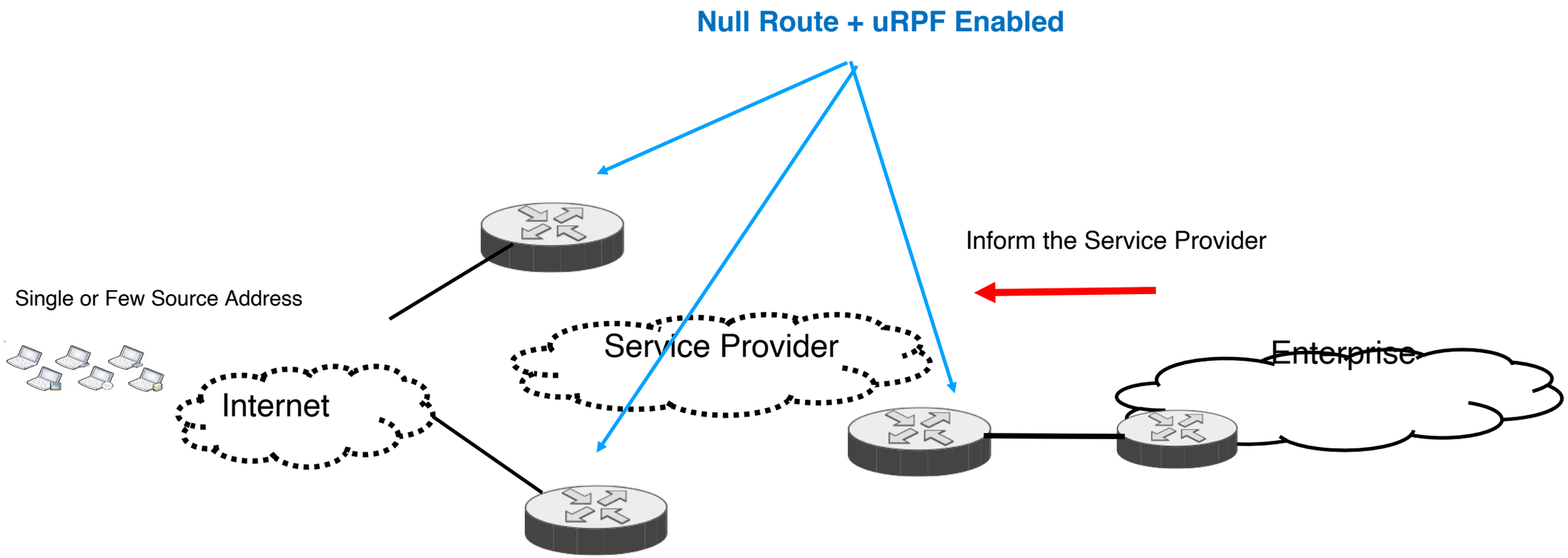
# Strict and Loose Mode uRPF

- BGP Injector Machine is used to quickly detect source addresses and placed into BGP

- Exa BGP is commonly used as an Injector tool

https://t.me/learningnets

# Source BGP RTBH (Remotely Triggered Blackholing)




https://t.me/learningnets

# Source BGP RTBH (Remotely Triggered Blackholing)

- Advantage of Source RTBH over Destination RTBH is customer's (victim) destination address is still usable

- This method only useful if source is single or few addresses

https://t.me/learningnets

# BGP Flowspec

- Defined in RFC 5575

- New AFI and SAFI is defined for BGP Flowspec

- AFI/SAFI : 1/133 :  Unicast Traffic Filtering Applications

https://t.me/learningnets

# BGP Flowspec

- Flow routes are automatically validated against unicast routing information or via routing policy

- When validated, firewall filter is created based on match and action criteria

https://t.me/learningnets

# BGP Flowspec

- Instead of blackholing entire IP address, there are many fields in the flow which can be used for filtering

https://t.me/learningnets

# BGP Flowspec Matching Fields

- Type 1: Destination Prefix
- Type 2 : Source Prefix
- Type 3 : IP Protocol
- Type 4 : Source or Destination Port
- Type 5 : Destination Port
- Type 6 : Source Port

https://t.me/learningnets

# BGP Flowspec Matching Fields

- Type 7 : ICMP Type
- Type 8 : ICMP Code
- Type 9 : TCP Flags
- Type 10 : Packet Length
- Type 11 : DSCP
- Type 12 : Fragment Encoding

https://t.me/learningnets

# BGP Flowspec Actions

- After identifying the flow, any of the below actions can be taken:

- Rate limit to 0 (Drop the traffic)
- Remarking
- Redirect to VRF (Route Target)
- Sampling the Traffic (Monitoring)

https://t.me/learningnets

# BGP Flowspec Vendor Support

- BGP Flowspec is supported by Commercial vendors and the Open Source implementation

- Commercial vendors , Arbor Peakflow , Juniper DDOS Secure , Alcatel Lucent , JunIper JUNOS , Cisco on ASR and CSR

https://t.me/learningnets

# BGP Flowspec and other DDOS Mitigation Approaches

- BGP Flowspec, compare to other methods such as manual filtering , Destination or Source based RTBH , is more granular

- BGP Flowspec can filter the traffic based on many other fields in the IP Packet , still provides same level of automation as RTBH

- Automation means, distributing the filtering policy rules to the edge nodes in the network, based on the matching criteria

https://t.me/learningnets

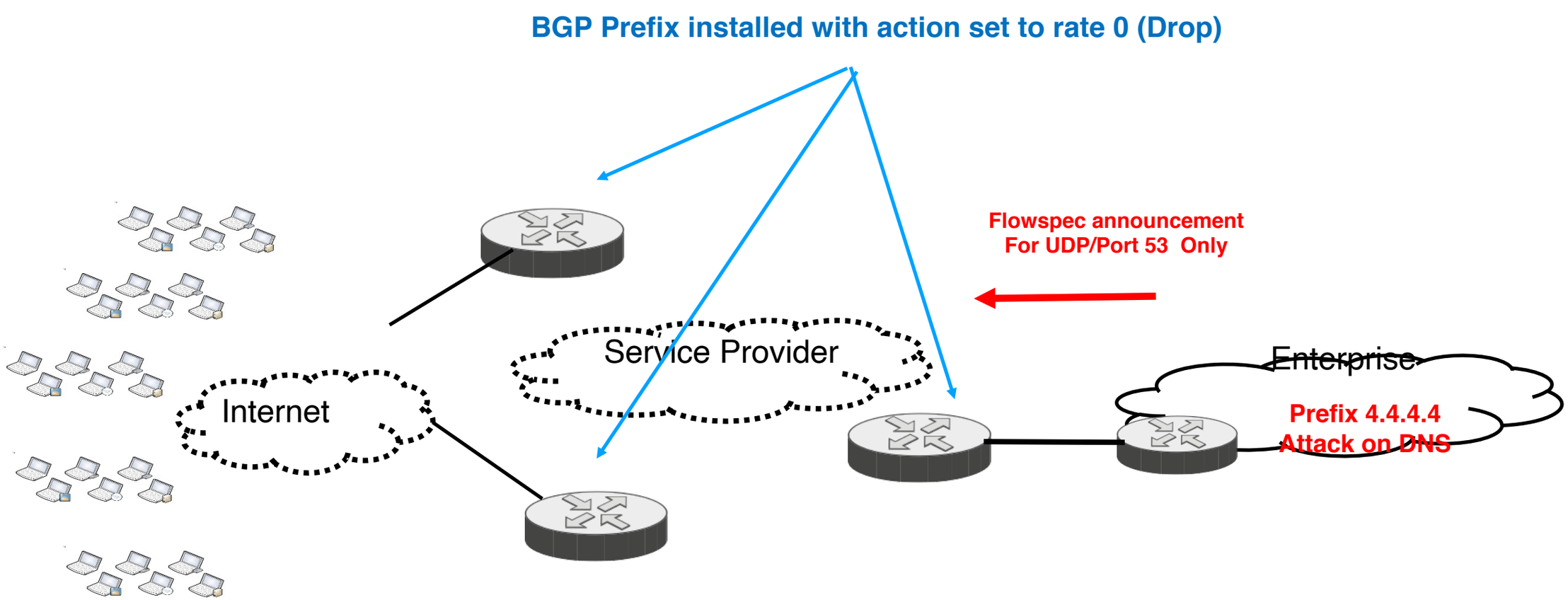
# How BGP Flowspec Works?

- Service Provider allows Customer to advertise Flowspec routes

- When attack starts towards customer (victim) IP address, customer initiates the filter , not just entire IP address but based on the attack type (DNS attack , NTP attack etc.)

https://t.me/learningnets

# BGP Flowspec Example




https://t.me/learningnets

# Who initiates Flowspec filters, SP or Customers?

- With BGP Flowspec, customer can initiate the filtering rules by advertising flowspec routes , saying to the SP that filter UDP port 53 for their specific IP address

- This is not mandatory though. If SP doesn’t provide this option, maybe because they don’t trust their customer expertise , SP may want customer to call them or through customer portal etc. when they are under attack and SP can initiate Flowspec filters , this option requires manual operation thus can take more time to stop attack

https://t.me/learningnets

# BGP Session Culling

- BGP session culling mitigates the negative impact of maintenance activities on IP networks, specifically on the IXP (Internet Exchange Points)

- RFC 8327 published for the BGP Session Culling

https://t.me/learningnets

# BGP Session Culling

- The approach is to ensure BGP-4 sessions that will be affected by maintenance are forcefully torn down before the actual maintenance activities start on the lower layers, such as physical layer

https://t.me/learningnets

# BGP Session Culling

- BGP Session Culling minimizes the amount of disruption that lower- layer network maintenance activities cause, by making BGP speakers preemptively converge onto alternative paths while the lower-layer network's forwarding plane remains fully operational

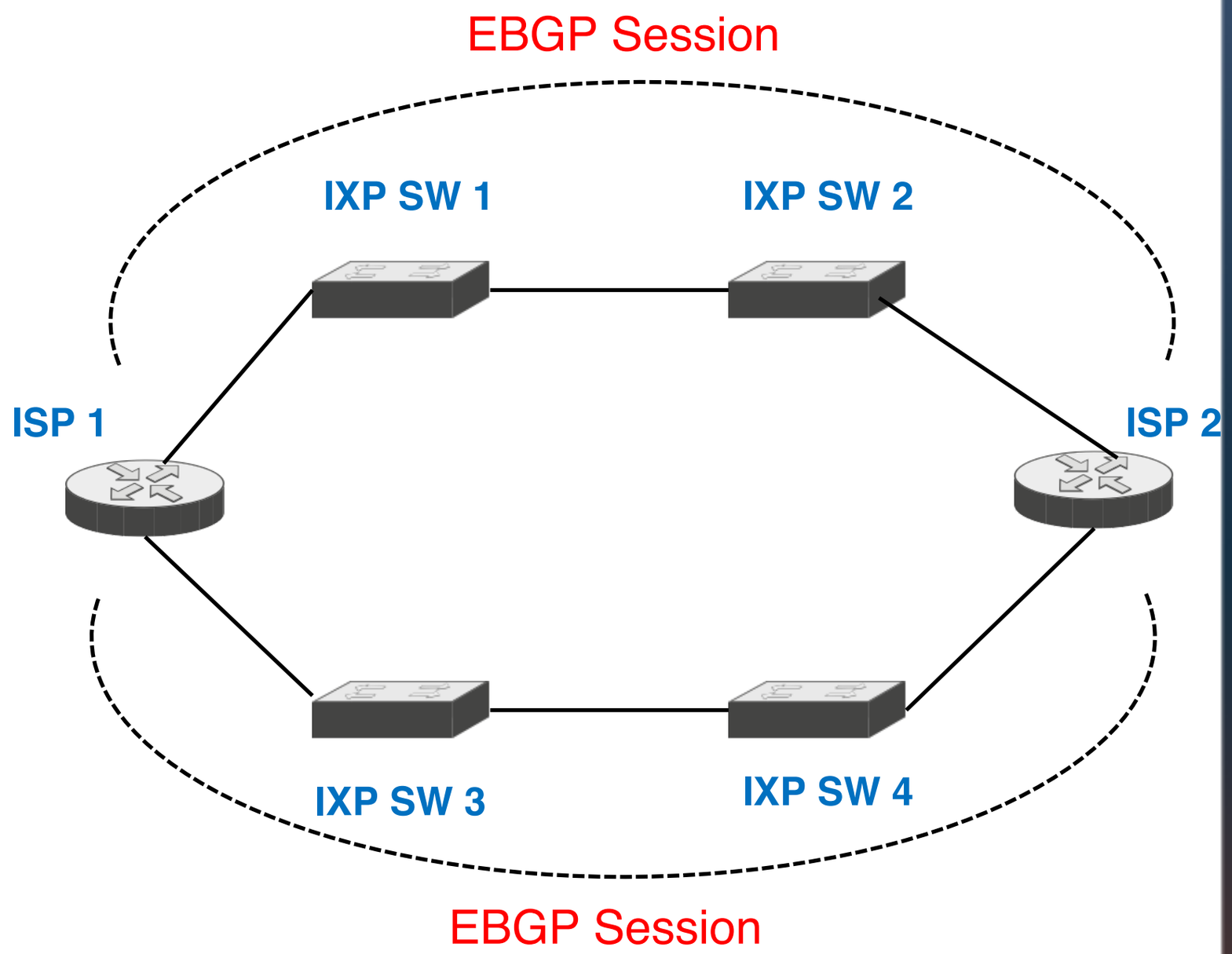
https://t.me/learningnets

# BGP Session Culling

There are two paths between ISP1 and ISP2
Two EBGP sessions are created between two operators

BGP Session Culling is basically a Layer4 Access list (ACL) applied on the IXP Layer2 ports before maintenance. The ACL blocks traffic to and from IXP subnet, BGP port TCP 179, allowing all other traffic

The ACL causes that BGP Hold timer expires, BGP sessions are culled down and end-user traffic can be rerouted over alternative paths. Afterwards maintenance is commenced.




https://t.me/learningnets

# BGP Session Culling Alternatives

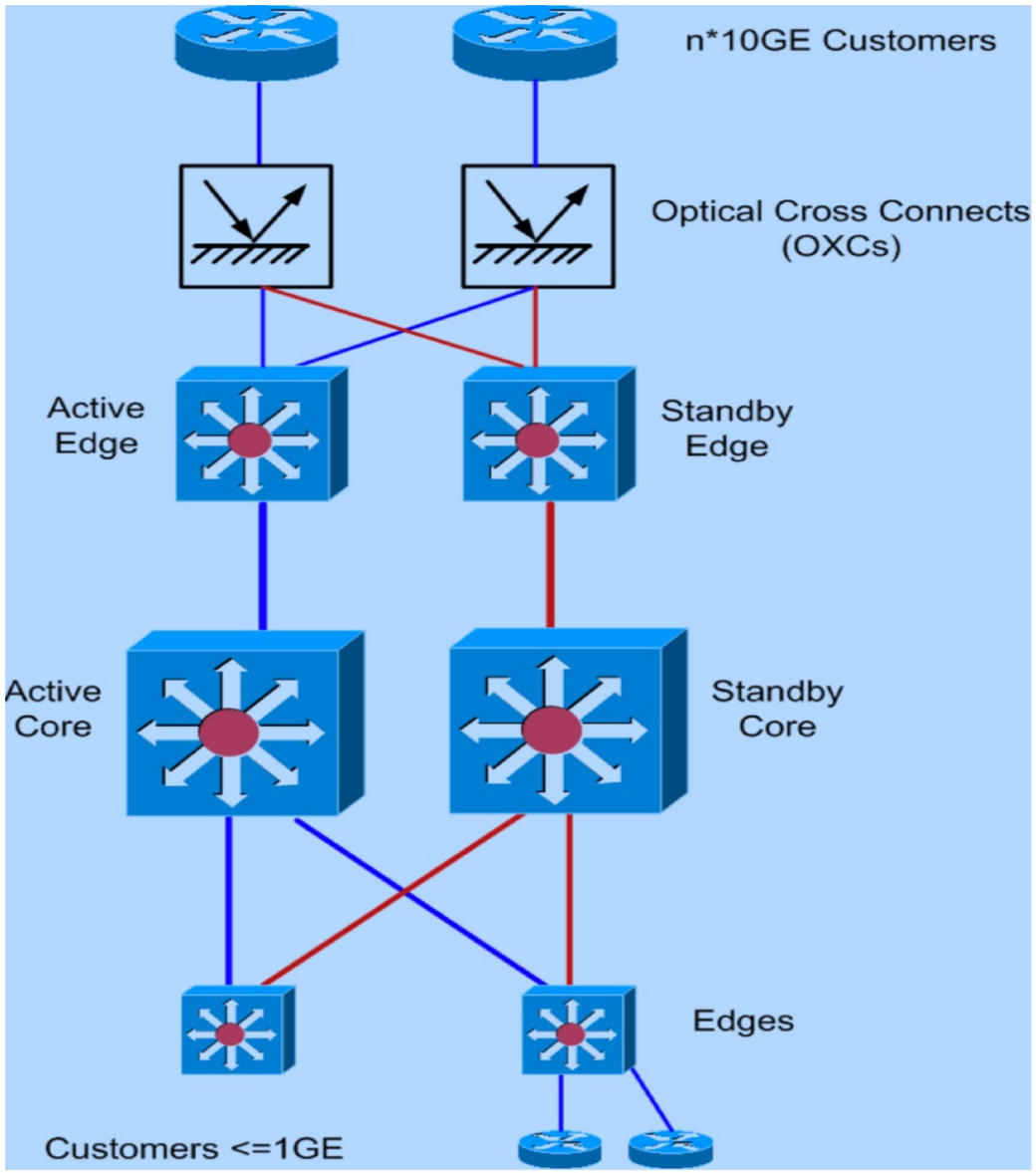
BGP Session Culling is described as Involuntary BGP Session Teardown technique in the RFC 8327. Lower layer network operator , such as IXP , cannot access to the IXP member routers , thus cannot bring the BGP session down which would help to avoid dataplane traffic loss

Alternative to Involuntary BGP Session Teardown, BGP Session can be teardown voluntarily. In this case, ISP shutdown the BGP session, IXP operator monitors the dataplane and when the traffic reaches to a minimum, IXP operator starts maintenance on the switches

Some IXPs use Optical Switch and terminate their IXP members to the Optical Cross Connect (OXC) Switch which is connected to the Fabric Switches

https://t.me/learningnets

- AMS-IX  (Amsterdam Internet Exchange) uses Optical Cross Connect Switch to have redundancy. It helps to redirect to the second plane when there is a maintenance activity on the first plane




**AMS-IX Customer to Fabric Connections**

https://t.me/learningnets

# BGP Session Culling Summary

- BGP session culling gains popularity and is applied at more and more IXPs
- It mitigates a negative impact of maintenance activities while requiring no input from the ISPs.
- It is a IETF standard
- Alternative method :  Voluntary BGP Session Teardown, Optical Cross Connect (OXC) Switches
- Session Culling is a Involuntary BGP Session teardown technique

https://t.me/learningnets

# BGP Graceful Restart, BGP Graceful Shutdown and BGP Administrative Shutdown Communication

- RFC 4724 specifies BGP Graceful Restart procedure
- RFC 8326 specifies BGP Graceful Shutdown
- RFC 8203 specifies BGP Administrative Shutdown Communication
- All three are separate mechanisms, provide different functionalities

https://t.me/learningnets

# BGP Graceful Restart

- GR ensures normal forwarding of data during the restart of routing protocols to prevent interruption of key services

- Graceful Restart is available today for OSPF, ISIS, EIGRP, LDP and BGP. Standards are defined for OSPF, ISIS, BGP and LDP to ensure vendor interoperability

- GR is usually used when the active route processor (RP) fails because of a software or hardware error, or used by an administrator to perform the master/slave switchover

- GR is known as NSF (Non Stop Forwarding)

https://t.me/learningnets

# BGP Graceful Restart

- During BGP peer relationship establishment, devices negotiate GR capabilities by sending supported GR capabilities to each other

- Dual processor systems which support Stateful Switch Over (SSO) or In-Service Software Upgrades (ISSU) can continue to forward traffic while restarting the control plane on the second processor

https://t.me/learningnets

# BGP Graceful Restart

- Usually, when BGP on a router restarts, all the BGP peers detect that the session went down and then came up. This "down/up" transition results in a "routing flap" and causes BGP route re-computation, generation of BGP routing updates, and unnecessary churn to the forwarding tables

https://t.me/learningnets

# BGP Graceful Restart

- The Graceful Restart is a BGP capability which is used by a BGP speaker to indicate its ability to preserve its forwarding state during BGP restart

- While control plane restarting, routers can forward the traffic, routes are marked as stale and removed after BGP session is re-established

https://t.me/learningnets

# BGP Graceful Shutdown

- BGP Graceful Shutdown is a well known BGP community

- BGP Graceful Shutdown is used for the planned maintenance activities

- The well-known community allows implementers to provide an automated graceful shutdown mechanism that does not require any router reconfiguration at maintenance time

https://t.me/learningnets

# BGP Graceful Shutdown

- Loss comes from transient lack of reachability during BGP convergence that follows the shutdown of an EBGP peering session between two Autonomous System Border Routers (ASBRs)

- Graceful Shutdown is used when GR is not applicable, for example during the maintenance forwarding/data plane might be impacted

https://t.me/learningnets

# BGP Graceful Shutdown

- Graceful Shutdown can be applied to reduce or avoid packet loss for outbound and inbound traffic flows initially forwarded along the peering link to be shut down

https://t.me/learningnets

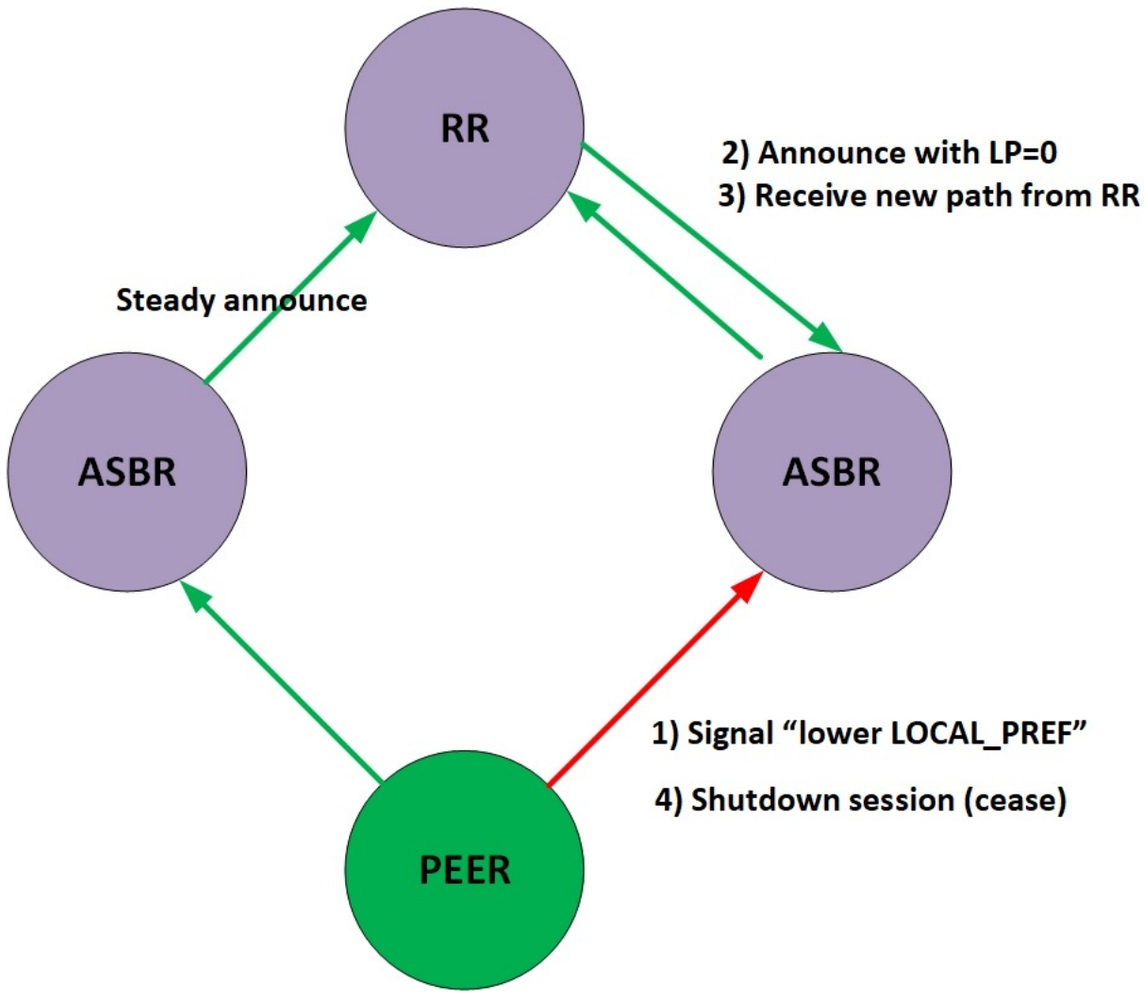
# BGP Graceful Shutdown

- In both Autonomous Systems (ASes), Graceful Shutdown trigger rerouting to alternate paths if they exist within the AS while allowing the use of the old path until alternate ones are learned. This ensures that routers always have a valid route available during the convergence process

https://t.me/learningnets

# BGP Graceful Shutdown

## Graceful Shutdown triggers “Path Hunting”




- Initiated by the operator on the router before maintenance by sending the GRACEFUL_SHUTDOWN well-known community (65535:0 as per IANA)
- Receiving EBGP peer sets LOCAL_PREFERENCE to 0 and selects paths to route traffic away from the initiator, (similar to setting overload in ISIS)
- When BGP session goes down, minimizes impact to traffic because alternate paths have already been installed.

https://t.me/learningnets

# BGP Administrative Shutdown Communication

- BGP Shutdown Communication is specified in RFC 8203

**According to RFC 8203:**

- BGP Administrative Shutdown Communication enhances the BGP Cease NOTIFICATION message "Administrative Shutdown" and "Administrative Reset" subcodes for operators to transmit a short freeform message to describe why a BGP session was shutdown or reset

https://t.me/learningnets

# BGP Administrative Shutdown Communication

- Operators, before the BGP session shutdown, inform the BGP peers with BGP Administrative Shutdown Communication procedure

https://t.me/learningnets

# BGP Administrative Shutdown Communication

- If a BGP speaker decides to terminate its session with a BGP neighbor, and it sends a NOTIFICATION message with the Error Code "Cease" and Error Subcode "Administrative Shutdown" or "Administrative Reset", it MAY include an UTF-8 encoded string

- They can send each other ‘ happy face ‘  emoji ☺

https://t.me/learningnets

# BGP Administrative Shutdown Communication

- Operators are encouraged to use the Shutdown Communication to inform their peers of the reason for the shutdown of the BGP session and include out-of-band reference materials

https://t.me/learningnets

# BGP Administrative Shutdown Communication

- An example of a useful Shutdown Communication would be: "[TICKET-1-1438367390] software upgrade; back in 2 hours"

- "[TICKET-1-1438367390]" is a ticket reference with significance to both the sender and receiver, followed by a brief human-readable message regarding the reason for the BGP session shutdown followed an indication about the length of the maintenance. The receiver can now use the string 'TICKET-1-1438367390' to search in their email archive to find more details

https://t.me/learningnets

# AIGP – Accumulated IGP Metric Attribute for BGP

- AIGP is specified in RFC 7311
- IGPs are designed to run within a single administrative domain and they make path-selection decision based on metric value
- BGP is an inter AS domain routing protocol and there is no inter AS metric which can be used for end to end shortest path selection
- AIGP is an Optional and Non-Transitive BGP Attribute

https://t.me/learningnets

# AIGP – Accumulated IGP Metric Attribute for BGP

- BGP is designed to provide routing over a large number of independent ASs with limited or no coordination among respective administrations. BGP does not use metrics in the path selection decisions

https://t.me/learningnets

# AIGP – Accumulated IGP Metric Attribute for BGP

- The accumulated IGP (AIGP) metric attribute for BGP enables deployment in which a single administration can run several contiguous BGP ASs.

- Such deployments allow BGP to make routing decisions based on the IGP metric

- In such networks, it is possible for BGP to select paths based on metrics as is done by IGPs.

- In this case, BGP chooses the shortest path between two nodes, even though the nodes might be in two different ASs

https://t.me/learningnets

# AIGP – Accumulated IGP Metric Attribute for BGP

- AIGP impacts the BGP best-route decision process

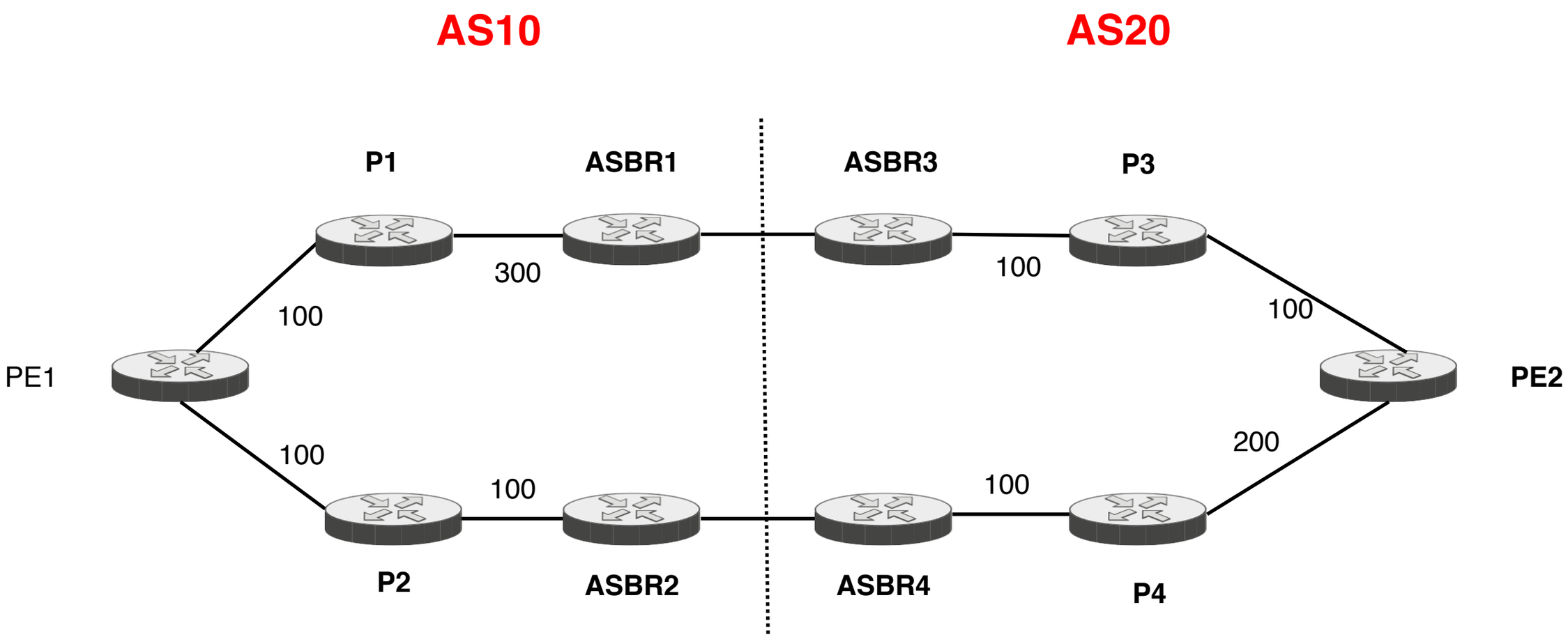
- The AIGP attribute preference rule is applied after the local-preference rule.

- The AIGP distance is compared to break a tie in the BGP best path selection

https://t.me/learningnets

# How AIGP Works

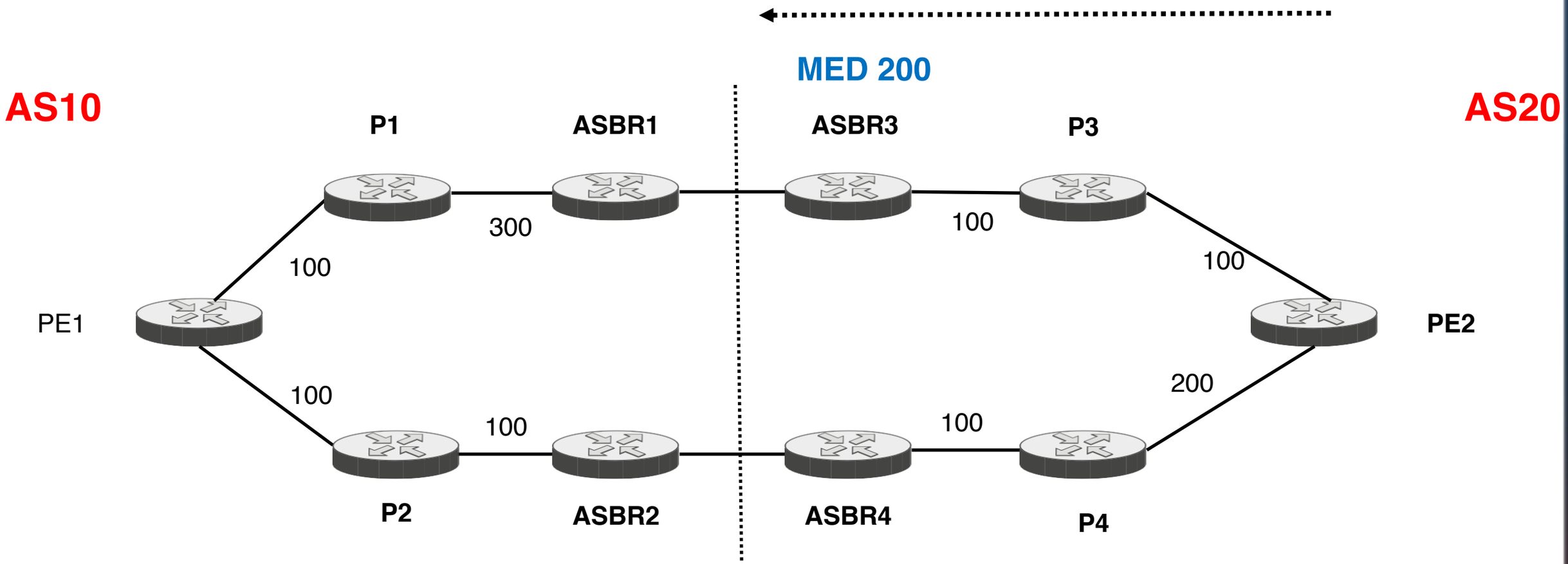
**Bottom Path End to End IGP Cost is PE1 to ASBR2 + ASBR 4 to PE2 = 500**
**Top Path End to End IGP Cost is PE1 to ASBR 1 + ASBR3 to PE2 = 600**
**Thus better path is selected by PE to reach PE 2 if AIGP is enabled**




https://t.me/learningnets

# BGP MED vs. AIGP Metric Attributes

**Top Path AIGP Metric is 600 , BGP MED is 200**
**Bottom Path AIGP Metric is 500 , BGP MED is 300**
**If AIGP wouldn't be used, Top Path would be selected as Best ,**
**that is not end to end optimal path**




https://t.me/learningnets

# EBGP Default Route Propagation Behavior without Policies– RFC 8212

- This is specified in RFC 8212 and updates RFC 4271

- Behavior of EBGP route propagation without Import and Export Route Policy created many problems on Internet

- With RFC 8212 routes are neither imported nor exported unless specifically enabled by configuration

https://t.me/learningnets

# EBGP Default Route Propagation Behavior without Policies– RFC 8212

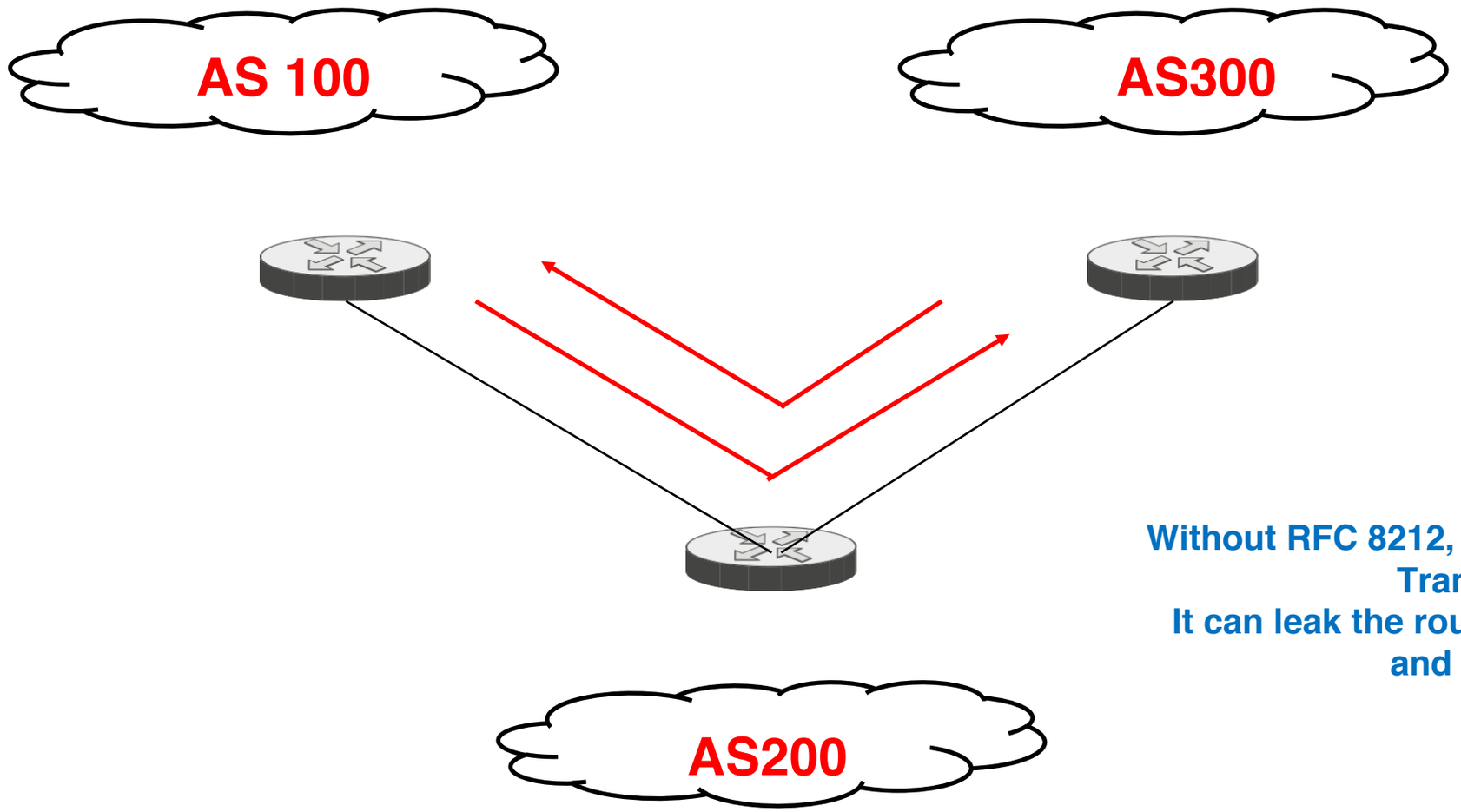
- Many deployed BGP speakers send and accept any and all route announcements between their BGP neighbors by default

- This behavior results with Route Leaks which will be covered in next topic

- Route Leak in general resulting in routing of traffic through an unexpected path

https://t.me/learningnets

# EBGP Default Route Propagation Behavior without Policies– RFC 8212




**Without RFC 8212, AS 200 can become a Transit AS**
**It can leak the routes between AS 100 and AS 300**

https://t.me/learningnets

# EBGP Default Route Propagation Behavior without Policies– RFC 8212

- As the Internet has become more densely interconnected, the risk of a misbehaving BGP speaker poses significant risks to Internet routing (Route Leak , Configuration mistakes etc.)

- RFC 8212 intends to improve this situation by requiring the explicit configuration of both BGP Import and Export Policies for any External BGP (EBGP) session such as customers, peers, or confederation boundaries for all enabled address families

https://t.me/learningnets

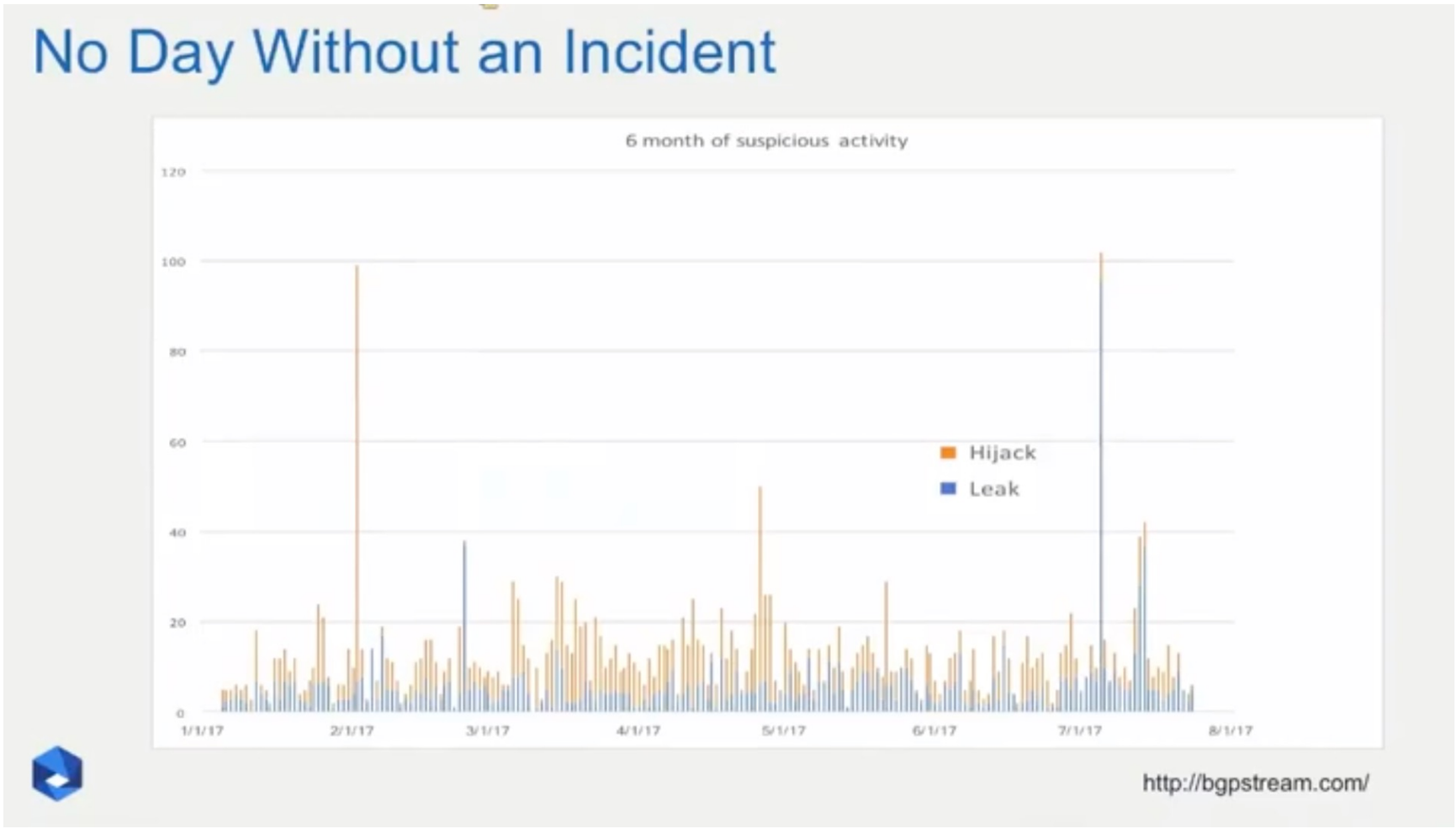
# BGP Information Security

- Border Gateway Protocol (BGP) is based entirely on trust between networks
- No built-in validation that updates are legitimate
- The chain of trust spans whole Internet , continents
- Lack of reliable data creates big issues !

https://t.me/learningnets

# BGP Information Security

- We will deal with BGP Route Leaks, Different Types of BGP Hijackings and IP Address Spoofing

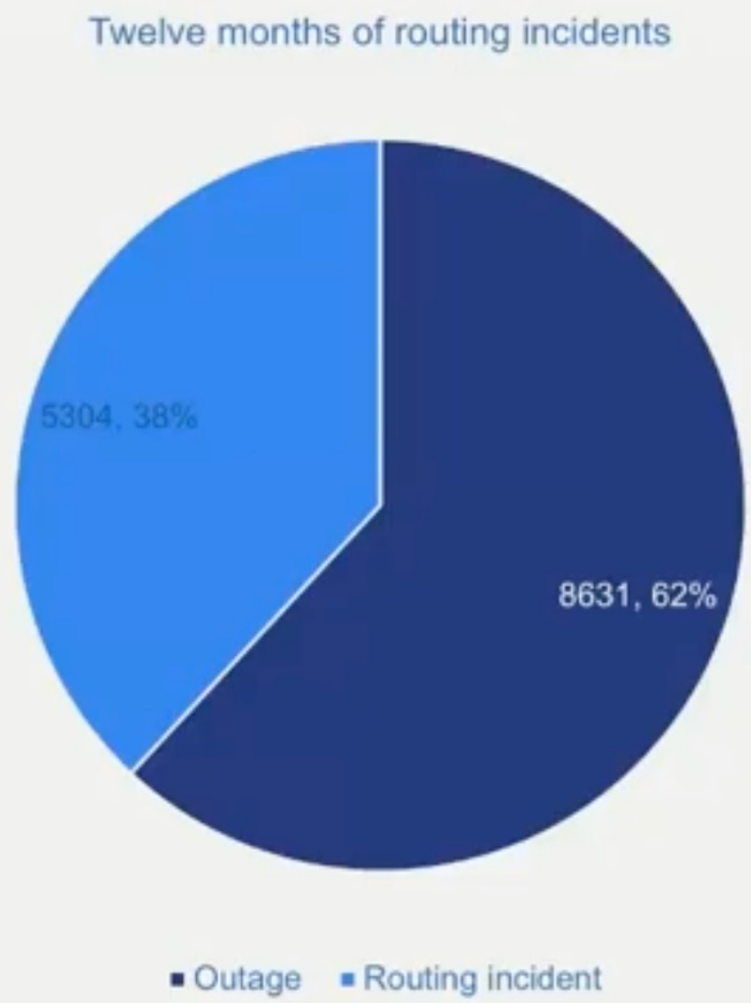



https://t.me/learningnets

# BGP Information Security

## Global stats

- 13,935 total incidents (either outages or attacks like route leaks and hijacks)
- Over 10% of all Autonomous Systems on the Internet were affected
- 3,106 Autonomous Systems were a victim of at least one routing incident
- 1,546 networks caused at least one incident




https://t.me/learningnets

# BGP Information Security

We will be talking about many tools:

- Prefix Filtering , Distribute Lists, AS-PATH Filtering
- RPKI ROA, Validation, IRR Toolset, BGPQ3
- BGPSEC

https://t.me/learningnets

# BGP Information Security

Biggest Problems with current approaches for BGP Security:

- Not enough deployment
- Lack of reliable data

https://t.me/learningnets

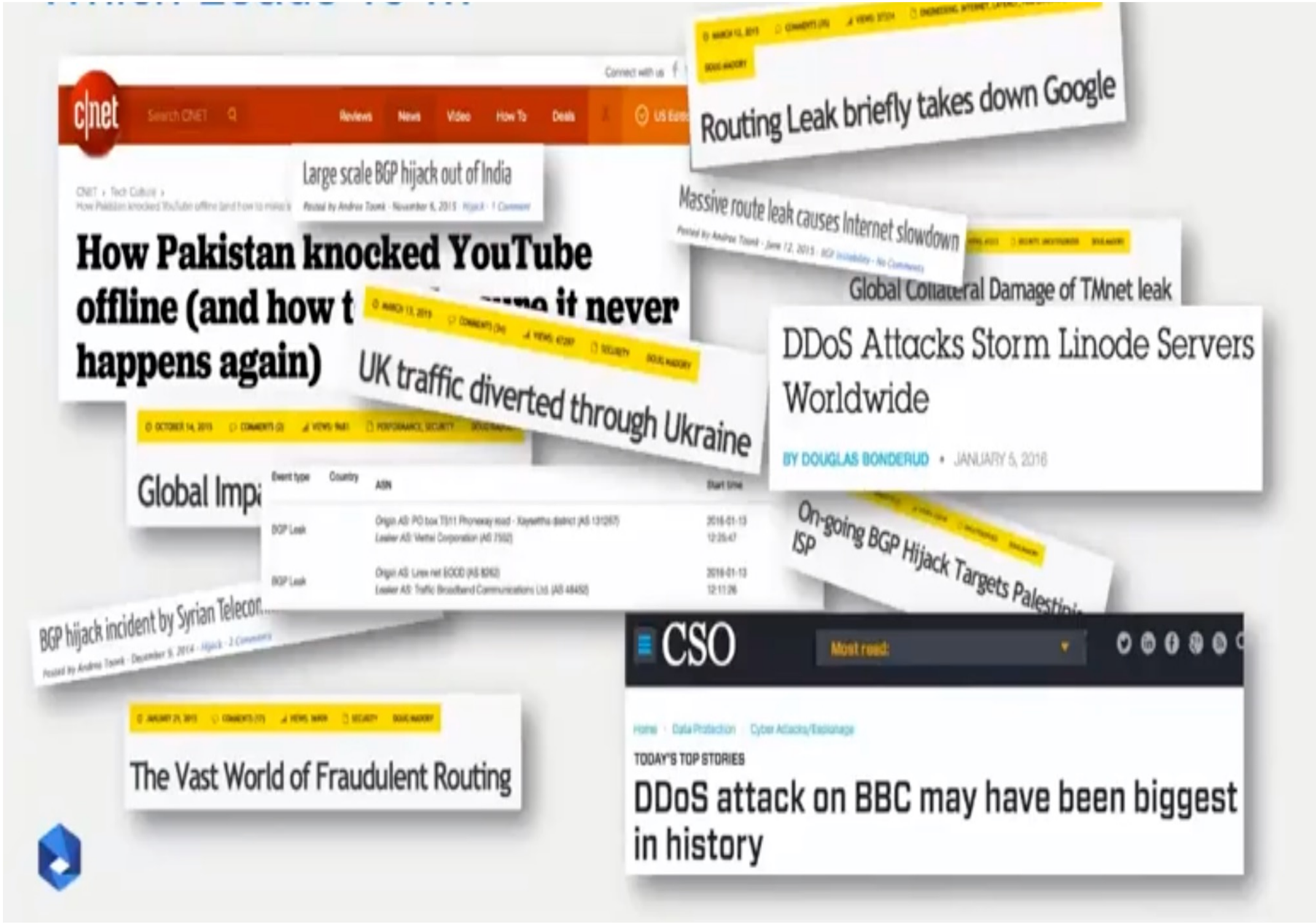
# Routing Incidents Cause Real World Problems

- Insecure routing is one of the most common paths for malicious threats

- Attacks can take anywhere from hours to months to even recognize

- Inadvertent error can take entire countries offline, while attackers steal money or data (Current Bitcoin Attack)

https://t.me/learningnets




https://t.me/learningnets

# We will deal with BGP Route Leaks, Different Types of BGP Hijacking and IP Address Spoofing

## The Threats: What's Happening?

| Event | Explanation | Repercussions | Solution |
|---|---|---|---|
| **Prefix/Route Hijacking** | A network operator or attacker impersonates another network operator, pretending that a server or network is their client. | Packets are forwarded to the wrong place, and can cause Denial of Service (DoS) attacks or traffic interception. | Stronger filtering policies |
| **Route Leak** | A network operator with multiple upstream providers (often due to accidental misconfiguration) announces to one upstream provider that is has a route to a destination through the other upstream provider. | Can be used for traffic inspection and reconnaissance. | Stronger filtering policies |
| **IP Address Spoofing** | Someone creates IP packets with a false source IP address to hide the identity of the sender or to impersonate another computing system. | The root cause of reflection DDoS attacks | Source address validation |

https://t.me/learningnets

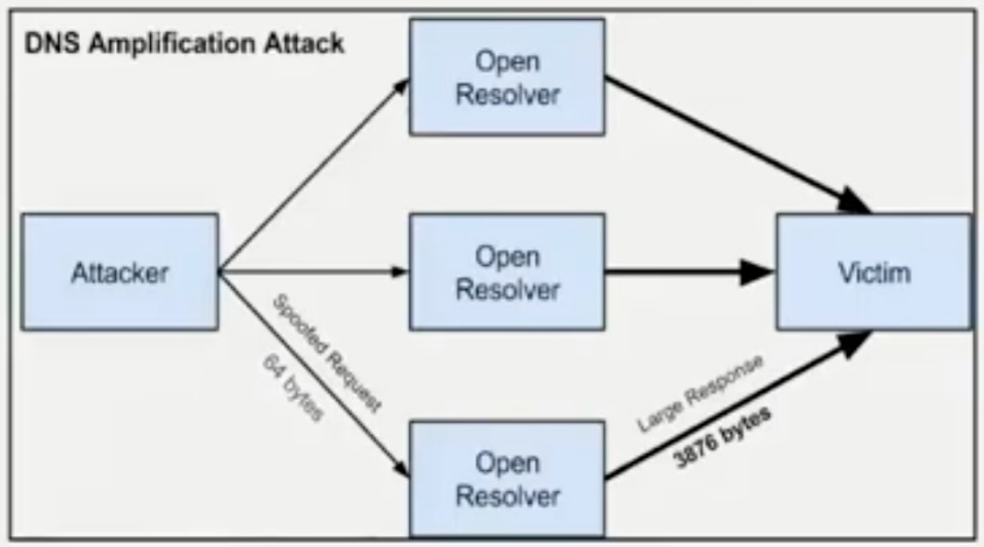
# IP Address Spoofing

## IP Address Spoofing

**IP address spoofing** is used to hide the true identity of the server or to impersonate another server. This technique can be used to amplify an attack.

**Example:** DNS amplification attack. By sending multiple spoofed requests to different DNS resolvers, an attacker can prompt many responses from the DNS resolver to be sent to a target, while only using one system to attack.

**Fix:** Source address validation: systems for source address validation can help tell if the end users and customer networks have correct source IP addresses (combined with filtering).




https://t.me/learningnets

# BGP Information Security

- BGP Information Security deals with the anything related with the network traffic, it doesn’t deal with the BGP Transport Security, such as TCP AO , MD5 Authentication etc.

https://t.me/learningnets

# BGP Information Security – Route Leaks

- BGP Route Leak is dangerous for the network traffic in many ways

- BGP Route Leak happens mostly due to miss-configuration but some BGP Route Leak Types is considered as an Attack

https://t.me/learningnets

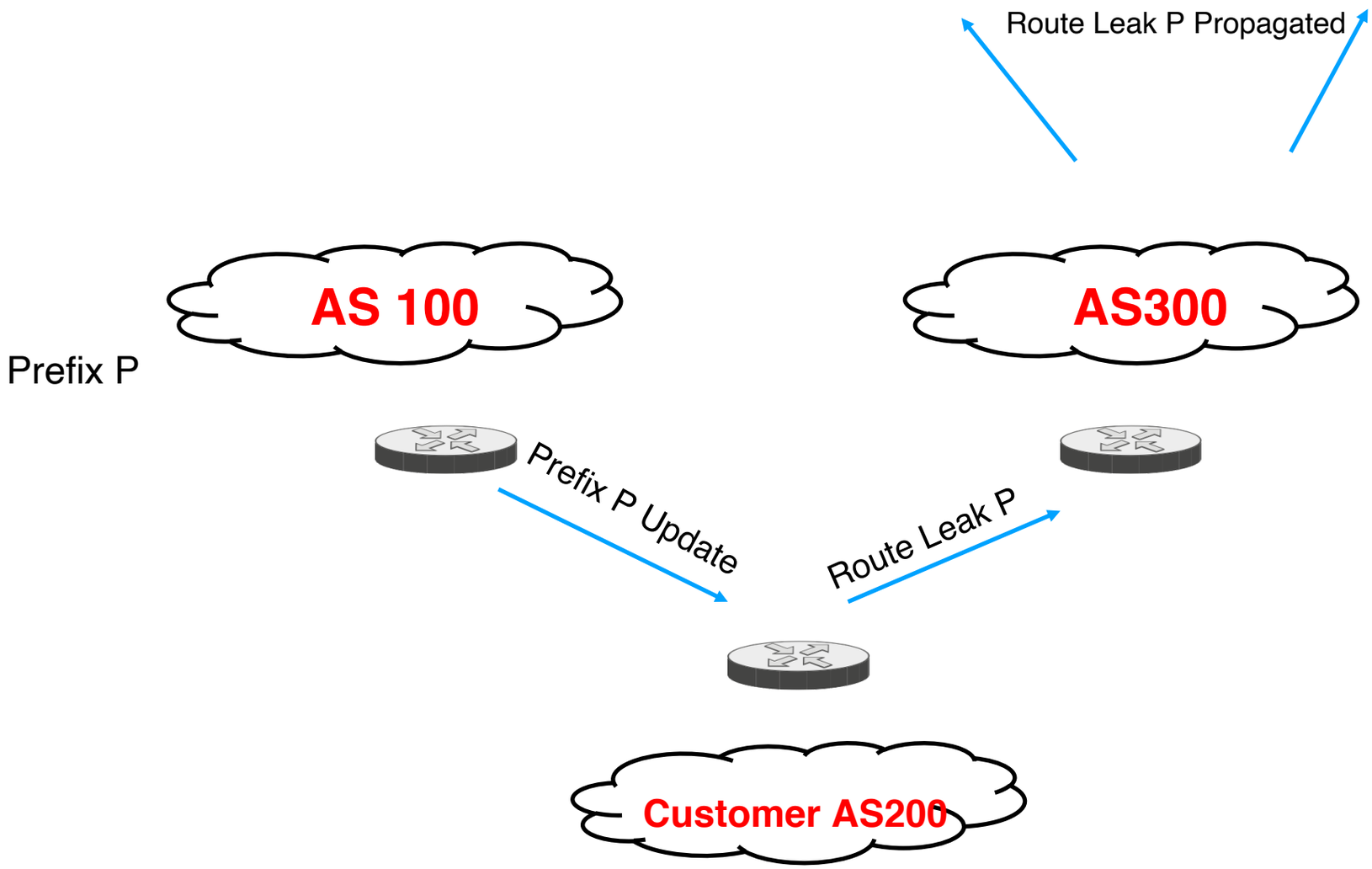
# BGP Information Security – Route Leaks

- It can create blackhole , extra latency , packet loss , thus will have bad effect for the customer experience

- The result of a route leak can be redirection of traffic through an unintended path that may enable eavesdropping or traffic analysis, or simply blackhole network traffic as there is no enough capacity on the network which leaked the prefixes

https://t.me/learningnets

# BGP Information Security – Route Leaks




https://t.me/learningnets

# BGP Information Security – Route Leaks

- RFC 7908 highlights the Problem Definition and Classification of BGP Route Leaks

- Formal definition of BGP Route Leak is the propagation of routing announcement(s) beyond their intended scope. That is, an announcement from an Autonomous System (AS) of a learned BGP route to another AS is in violation of the intended policies of the receiver, the sender, and/or one of the ASes along the AS path

https://t.me/learningnets

# BGP Route Leak Types

Based on RFC 7908, several Route Leak Type is defined

1. Hairpin Turn with Full Prefix
2. Lateral ISP-ISP-ISP Leak
3. Leak of Transit Provider Prefixes to Peer
4. Leak of Peer Prefixes to Transit Provider
5. Prefix Re-origination with Data Path to Legitimate Origin
6. Accidental Leak of Internal Prefixes and More-Specific Prefixes

https://t.me/learningnets

# BGP Route Leak - Hairpin Turn with Full Prefix

- A multihomed AS learns a route from one upstream ISP and simply propagates it to another upstream ISP

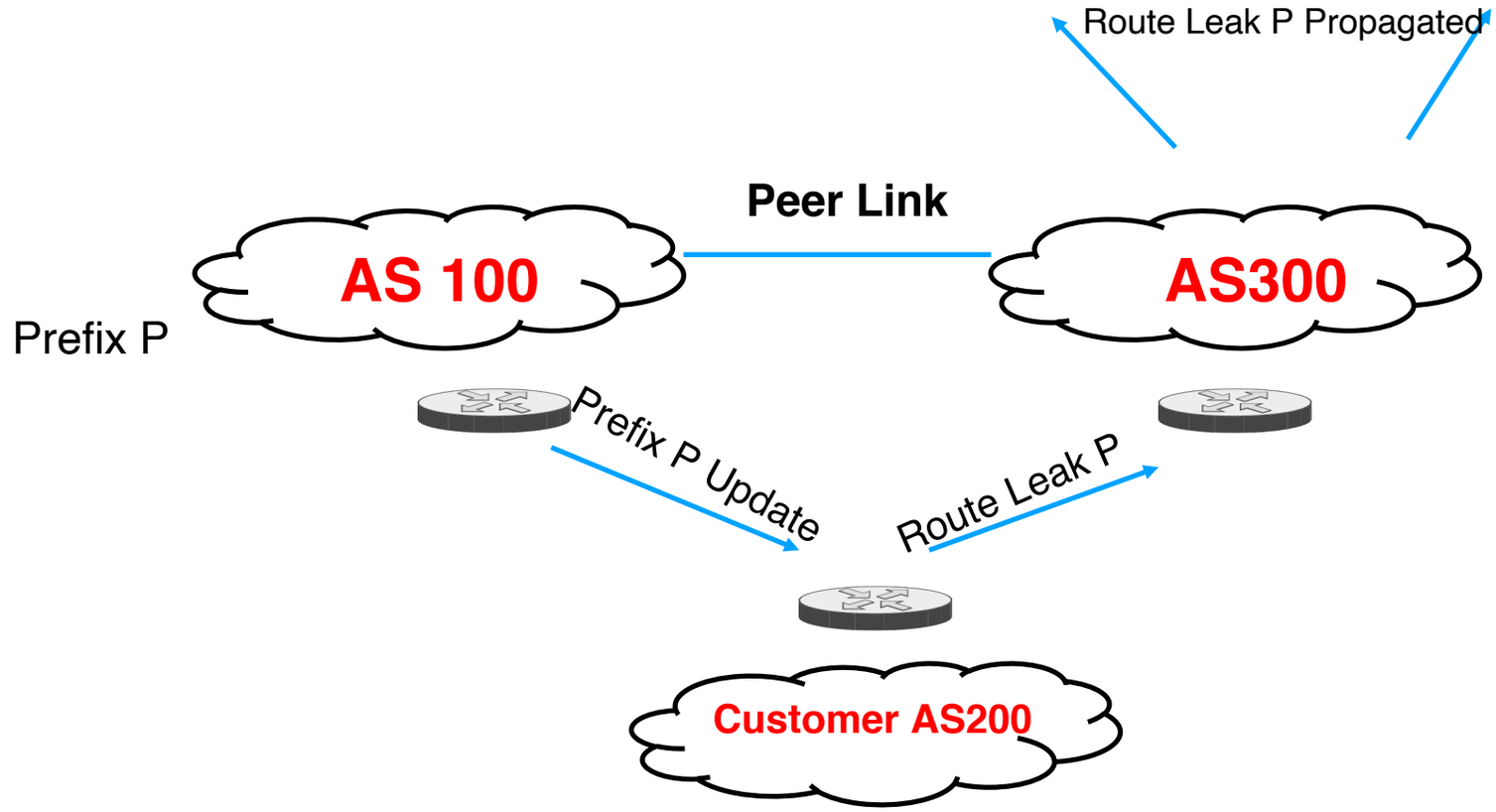
- It should be noted that leaks of this type are often accidental (not malicious)

- The leak often succeeds (the leaked update is accepted and propagated) because the second ISP prefers customer announcement over peer announcement of the same prefix

https://t.me/learningnets

# BGP Route Leak - Type 1 - Hairpin Turn with Full Prefix

**AS 300 Receives Prefix P from, both Customer AS 200 and Peer AS 100 As 300 Prefers AS 200 because based on GAO-Rexford Model, Customer is preferred to Peer announcements**




https://t.me/learningnets

# BGP Route Leak - Type 2 – Lateral ISP-ISP-ISP Leak

- The term "lateral" here is synonymous with "non-transit" or "peer-to-peer”

- This type of route leak typically occurs when, for example, three sequential ISP peers (ISP-A, ISP-B, and ISP-C) are involved, and ISP-B receives a route from ISP-A and in turn leaks it to ISP-C

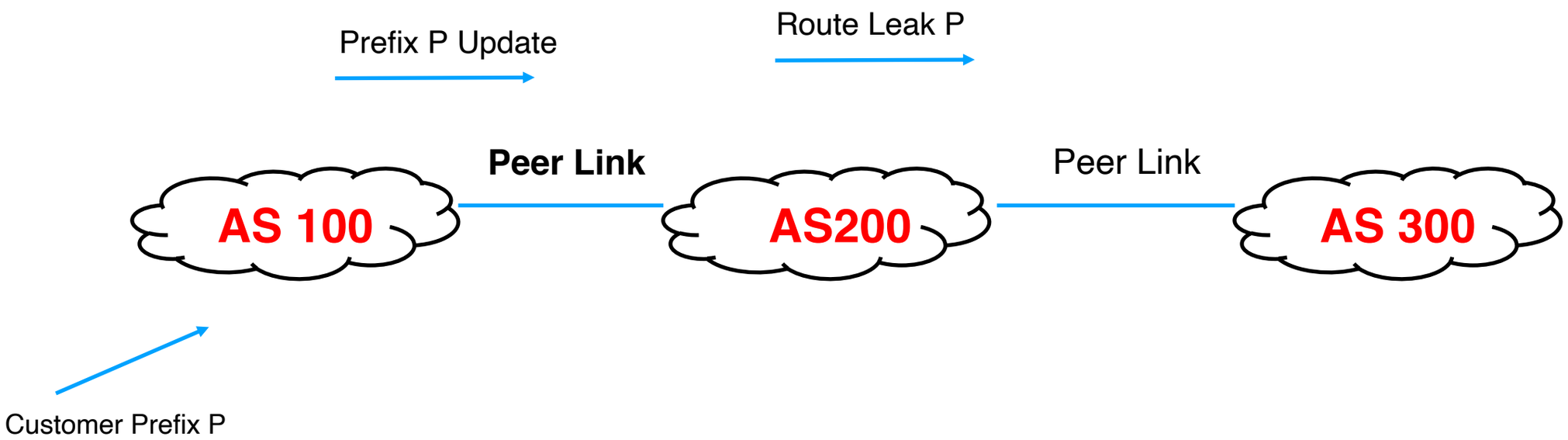
https://t.me/learningnets

# BGP Route Leak - Type 2 – Lateral ISP-ISP-ISP Leak

- The typical routing policy between laterally (i.e., non-transit) peering ISPs is that they should only propagate to each other their respective customer prefixes

https://t.me/learningnets

# BGP Route Leak - Type 2 – Lateral ISP-ISP-ISP Leak




**AS 200 shouldn’t send Customer Prefix of AS 100 to AS 300, It is a Route Leak**

https://t.me/learningnets

# BGP Route Leak – Type 3 – Leak of Transit Provider Prefixes to Peer

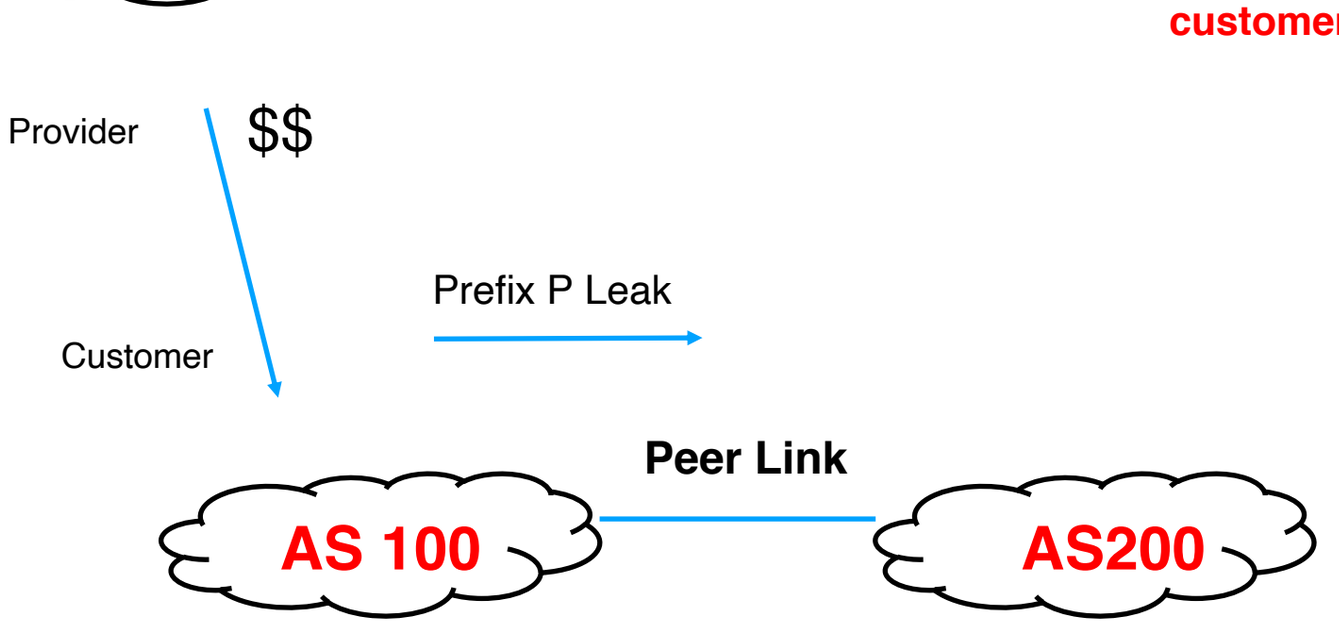
- This type of route leak occurs when an offending AS leaks routes learned from its transit provider to a lateral (i.e., non-transit) peer

https://t.me/learningnets

# BGP Route Leak – Type 3 – Leak of Transit Provider Prefixes to Peer

**AS 300**

**AS 100 shouldn't send the prefixes which are learned from upstream Transit AS to its Peers Transit prefixes should be sent to the customers**




https://t.me/learningnets

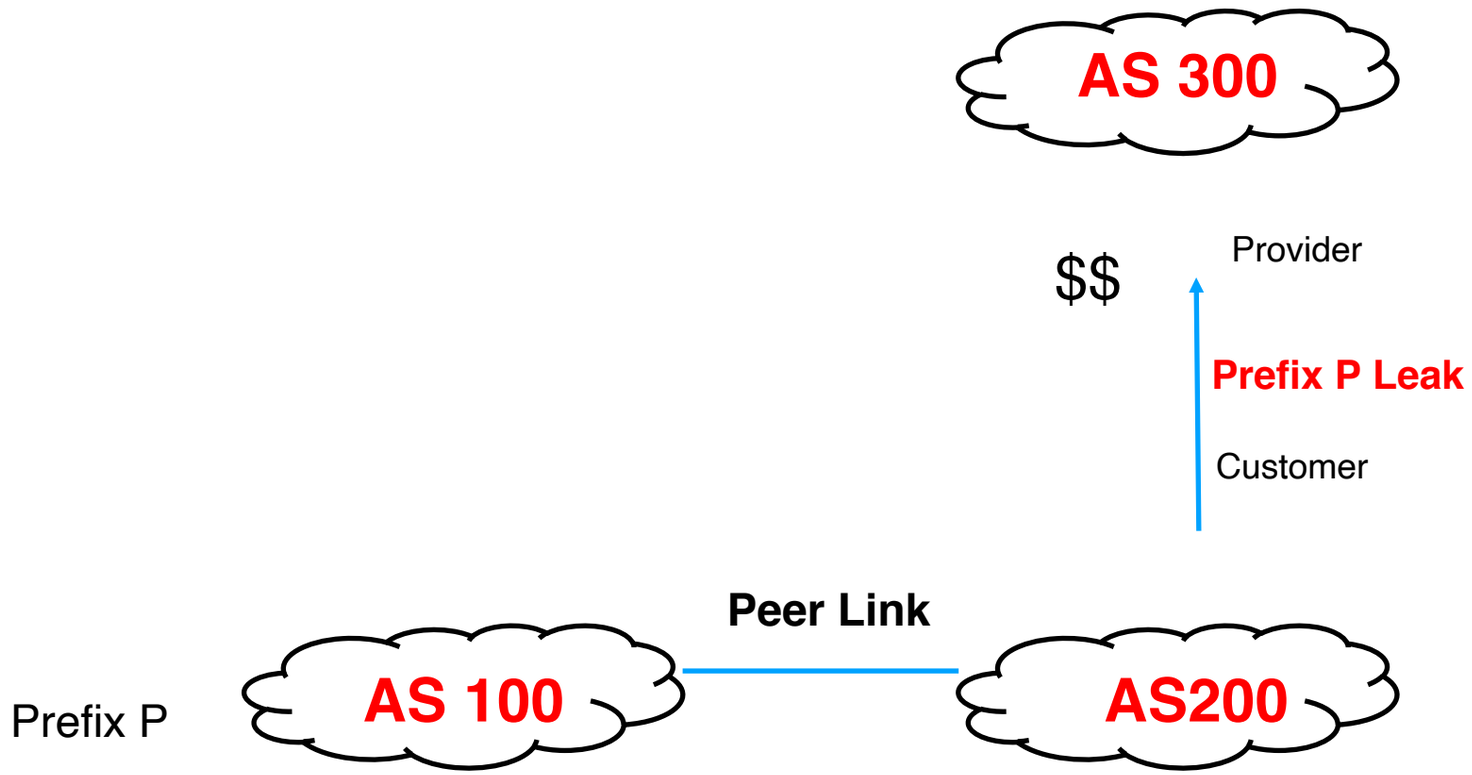
# BGP Route Leak - Type 4 - Leak of Peer Prefixes to Transit Provider

- This type of route leak occurs when an offending AS leaks routes learned from a lateral (i.e., non-transit) peer to its (the AS's) own transit provider

https://t.me/learningnets

# BGP Route Leak - Type 4 - Leak of Peer Prefixes to Transit Provider

AS 200 shouldn't send the Prefixes which are learned from its Settlement Free Peer to its own Transit Provider




https://t.me/learningnets

# BGP Route Leak - Type 4 - Leak of Peer Prefixes to Transit Provider

- Type 4 BGP Route Leak incidents commonly seen on the Internet

- Some Examples are :

- Axcelx-Hibernia route leak of Amazon Web Services (AWS) prefixes causing disruption of AWS and a variety of services that run on AWS

- Hathway-Airtel route leak of 336 Google prefixes causing widespread interruption of Google services in Europe and Asia

- Moratel-PCCW route leak of Google prefixes causing Google's services to go offline

https://t.me/learningnets

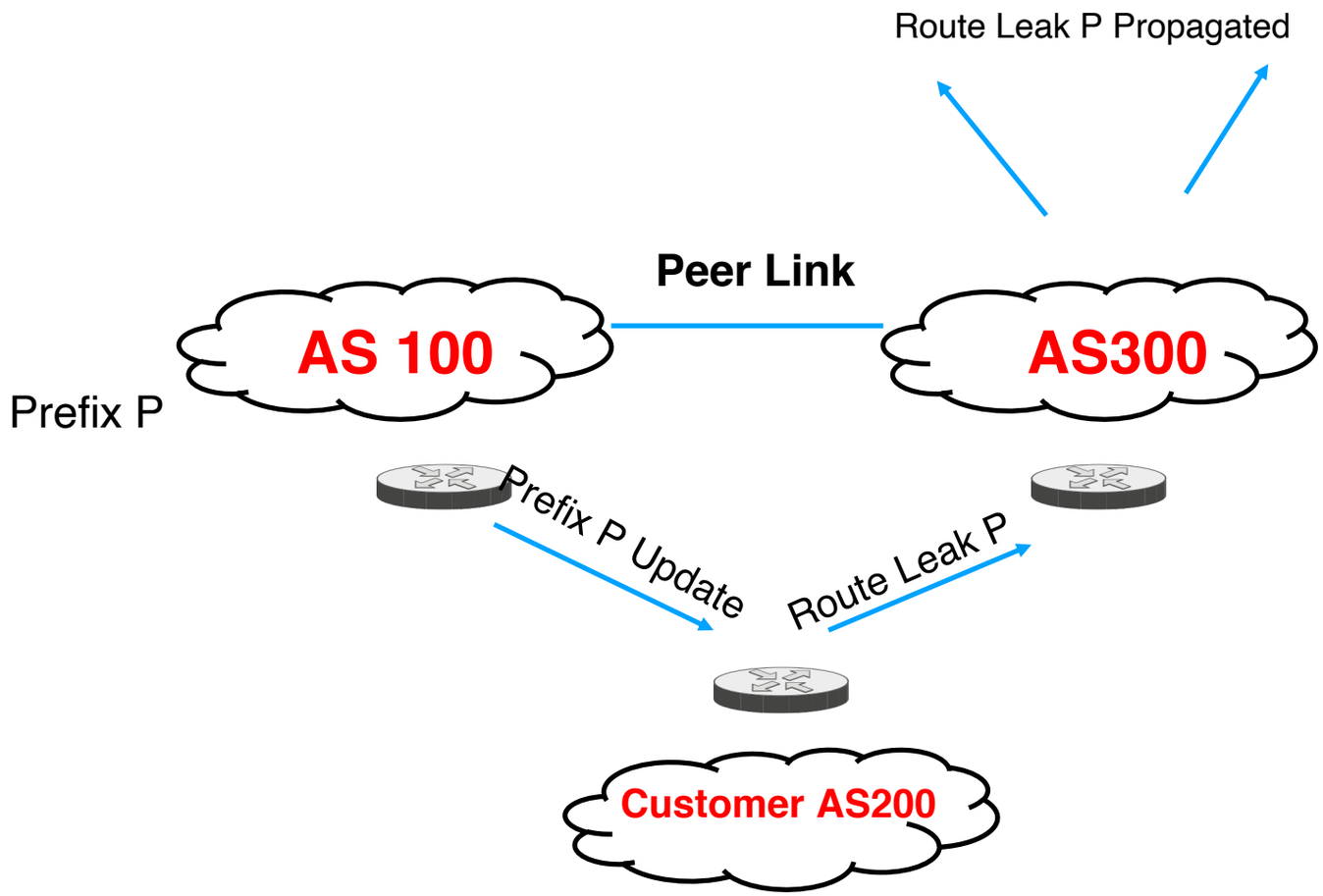
# BGP Route Leak - Type 5: Prefix Re-origination with Data Path to Legitimate Origin

- A multihomed AS learns a route from one upstream ISP and announces the prefix to another upstream ISP as if it is being originated by it (i.e., strips the received AS path and re-originates the prefix). This can be called re-origination or mis-origination

https://t.me/learningnets

# BGP Route Leak - Type 5: Prefix Re-origination with Data Path to Legitimate Origin

**Similar to Type 1 Route Leak but in Type 5, Customer AS 200 strips the AS 100 which is the Originator AS from the AS Path**




https://t.me/learningnets

# BGP Route Leak - Type 6: Accidental Leak of Internal Prefixes and More-Specific Prefixes

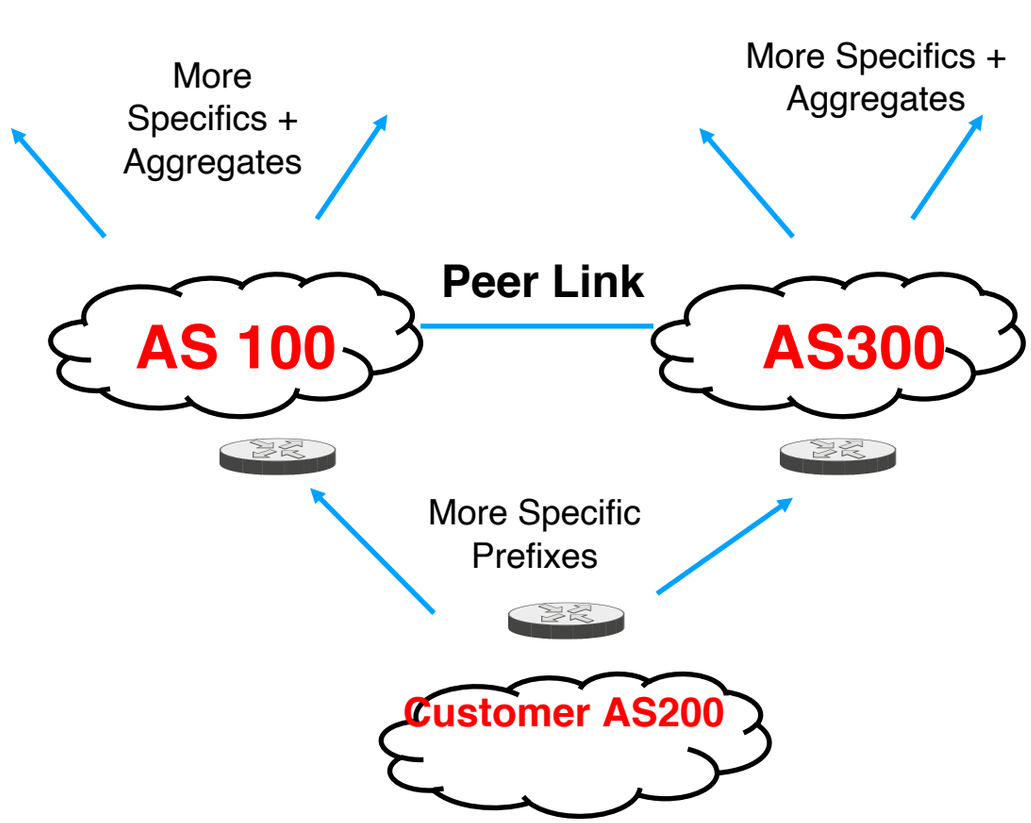
- An offending AS simply leaks its internal prefixes to one or more of its transit-provider ASes and/or ISP peers. The leaked internal prefixes are often more-specific prefixes subsumed by an already announced, less-specific prefix

- Partly because of this route leak , 512k incident happened in August 2014

https://t.me/learningnets

# BGP Route Leak - Type 6: Accidental Leak of Internal Prefixes and More-Specific Prefixes

**Transit ISPs accept More Specific announcements from downstream AS and advertises the more specifics to DFZ**




https://t.me/learningnets

# BGP Route Leak - Type 6: Accidental Leak of Internal Prefixes and More-Specific Prefixes

- Internal Prefix Route leak generally short lived route leaks

- It disappears after short amount of time

- In August 2014, 512k incident happened this will be explained next!

https://t.me/learningnets

# 512K Incident, Type 6 BGP Route Leak Contribution to it!

- The 12th August 2014 was widely reported as a day when the Internet collapsed

- What was happening was that the Internet's growth had just exceeded the default configuration limits of certain models of network switching equipment

- 0800 UTC on that day, when the Internet was flooded with 22,000 new prefixes, which were withdrawn very rapidly thereafter

https://t.me/learningnets

# 512K Incident, Type 6 BGP Route Leak Contribution to it!

- All these routes shared a common origin, AS 701, and were all more specifics of already announced aggregate routes

- The announcements were short-lived, and were withdrawn soon after their announcement

https://t.me/learningnets

# 512K Incident, Type 6 BGP Route Leak Contribution to it!

- This was a Type 6 Route Leak by AS 701, announced 22k more specific prefixes to the Internet. This amount of extra prefixes suddenly exceeds the 512000 prefixes in the Default Free Zone

- This triggered some BGP session restarts due to tripping some maximum prefix threshold values that was installed by network operators

https://t.me/learningnets

# Preventing BGP Route Leaks

- Because BGP is founded on trust and thus insecure, it can be extremely difficult to quickly resolve a route leak affecting your network, as you'll need to convince other networks to choose the legitimate route over the incorrect one

- While you won't have complete control in a route leak situation, you do have some options to fight with ongoing route leaks

https://t.me/learningnets

# Preventing BGP Route Leaks

- IRR Route Objects, RPKI ROAs and BGPSEC all contributed to prevent Route Leak

- Depends on the Type of Route Leaks and percentage of the implementation and whether upstream ISP deployed these methods, IRR, RPKI and even BGPSEC may not be effective

- Most people deploy Static Prefix Filters but when the network gets larger, maintaining prefix list becomes easily cumbersome

https://t.me/learningnets

# BGP Hijacking

- Hijacking occurs when an attacker claims to own a prefix or sub-prefix that belongs to another AS causing redirection of routes from the AS to the attacker

- Attackers hijack prefixes to produce different malicious activities. For example, the hijacker can blackhole all traffic to the victim causing a DoS for that network

- In another scenario, the attacker becomes a man-in-the-middle, intercepting the traffic without affecting victim reachability

https://t.me/learningnets

# BGP Hijacking

- Phishing attacks can also be done by hijacking a prefix through redirecting traffic to an incorrect destination

- Additionally, the attacker can use stolen IP addresses to send spam

- There are examples of each of these with BGP Hijacks

https://t.me/learningnets

# Some BGP Hijacking Incidents

## Hijacking for Monetary Gain

- Between October 2013 – May 2014 , Canadian Hijacker ISP originated more specifics (Sub-Prefix Hijack) to stole Bitcoin
- Hijacked AWS, LeaseWeb and couple other OTT Subnets
- It was towards IX Peers at TORIX (Toronto IX)
- $83,000 was stolen

https://t.me/learningnets

# Some BGP Hijacking Incidents

## Hijacking for Censorship

- March 28 – 30 Election in Turkey
- Turk Telekom brought up DNS servers and redirected DNS traffic
- Man in the Middle affected Google, Level3, OpenDNS etc.

https://t.me/learningnets

# Some BGP Hijacking Incidents

## Hijacking for Spamming

- Using un-announced address space to send spam
- Find unused space & announce for few hours and send spam
- Rinse and Repeat
- This technique is known as IP Squatting

https://t.me/learningnets

# Types of BGP Hijacks

**BGP Hijacking are classified into 4 Subtypes:**

- Hijacking a prefix
- Hijacking a prefix and its AS
- Hijacking a sub-prefix
- Hijacking a sub-prefix and its AS

https://t.me/learningnets

# BGP Prefix Hijacking

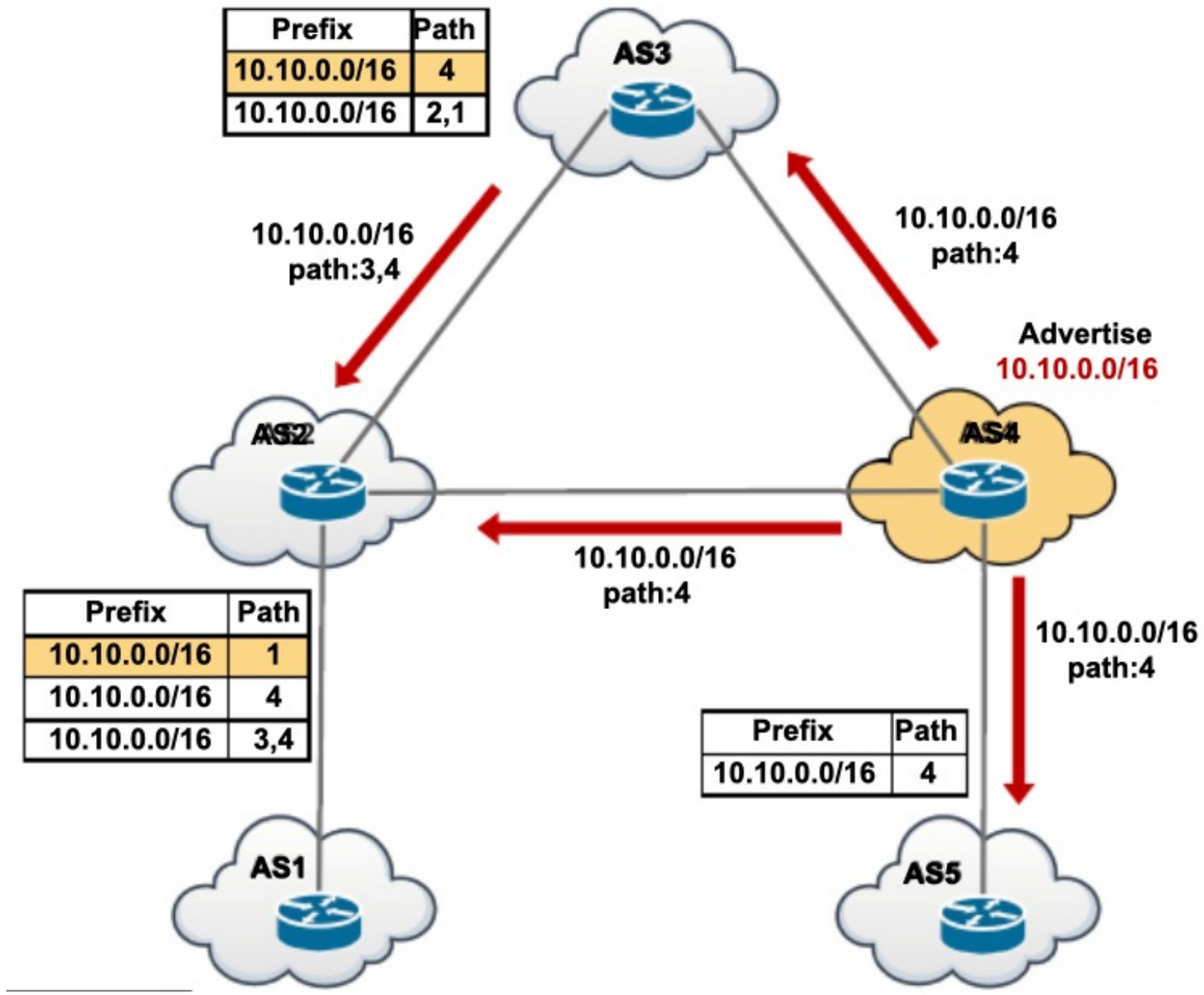
- In this type of hijack, an attacker configures its BGP router to announce a prefix belonging to another AS

- BGP allows any BGP speaker to announce any route regardless of whether the route actually exists or not

- Attacker’s neighbors will adopt it as a new route

https://t.me/learningnets

AS4 sends hijacked prefix 10.10.0.0/16
Legitimate Origin is AS1
AS3 and AS5 believes that AS4 owns the prefix
AS2 doesn't change the best path since both AS1 and AS4 has same AS Path length




https://t.me/learningnets

# BGP Prefix Hijacking

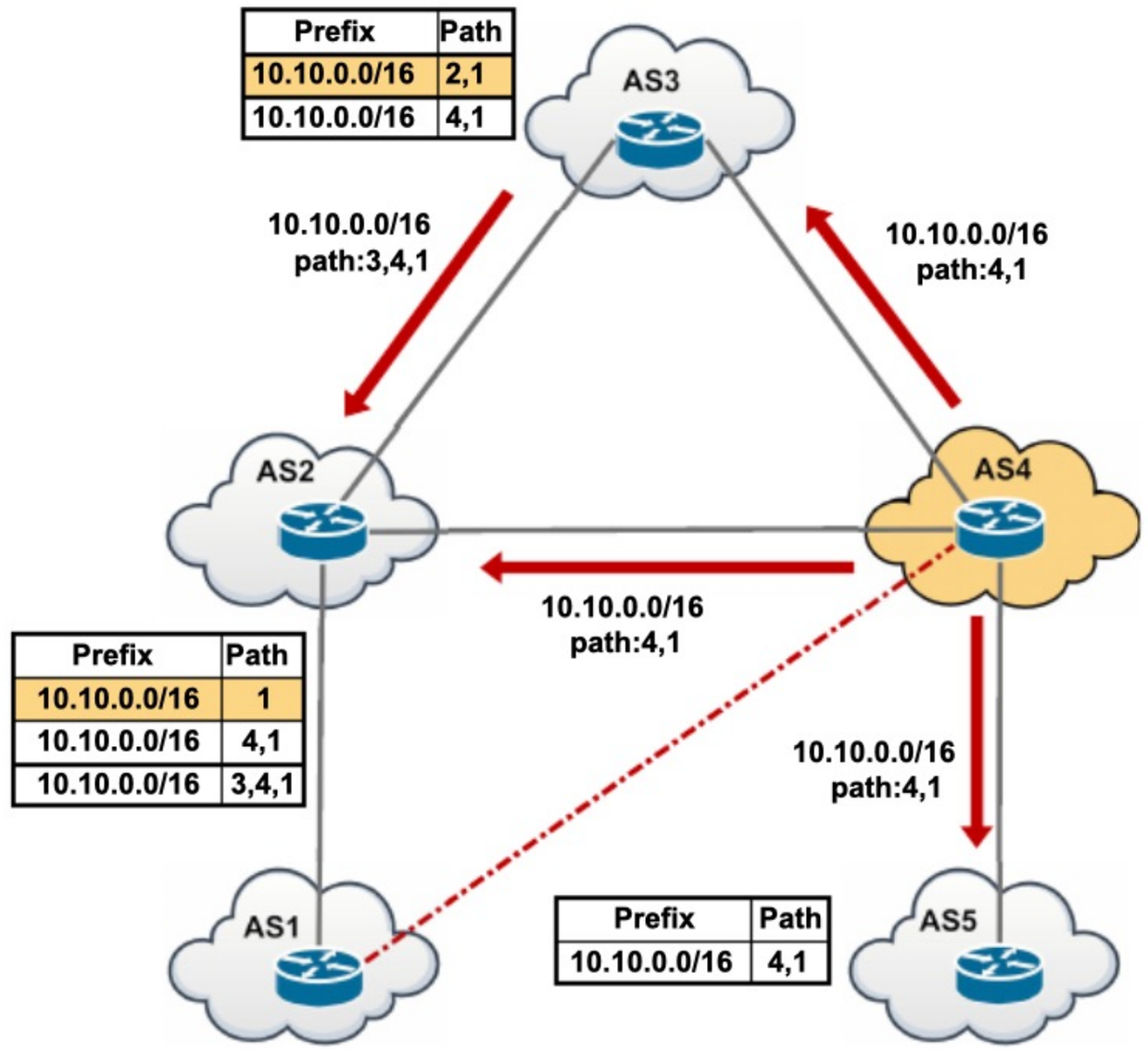
- BGP Prefix Hijacking is commonly known as BGP Exact Prefix Hijacking (Exactly same prefix length with the victim AS is announced)

- Since shortest route paths are typically preferred, only a part of the Internet that is closer to the hijacker (in number of AS-hops) switches to route paths towards the hijacker

https://t.me/learningnets

# Prefix and its AS Hijack

**Origin validation can stop Prefix Hijack but not this one, because**
**With Prefix and its AS Hijack, attacker manipulates the path as it is connected to the Origin**
**AS4 sends an announcements to the AS 2 , 3 and 5 as [4, 1]**
**AS2 and AS3 doesn't use AS4, but AS5 still does**




https://t.me/learningnets

# BGP Sub-prefix Hijacking

- With BGP Sub-prefix Hijacking, an attacker announces a sub-prefix that belongs to a victim AS

- BGP selects the most specific address or longest address match. For example, a BGP router will select a specific address such as 10.10.0.0/24 over a more general address such as 10.10.0.0/16

https://t.me/learningnets

# BGP Sub-prefix Hijacking

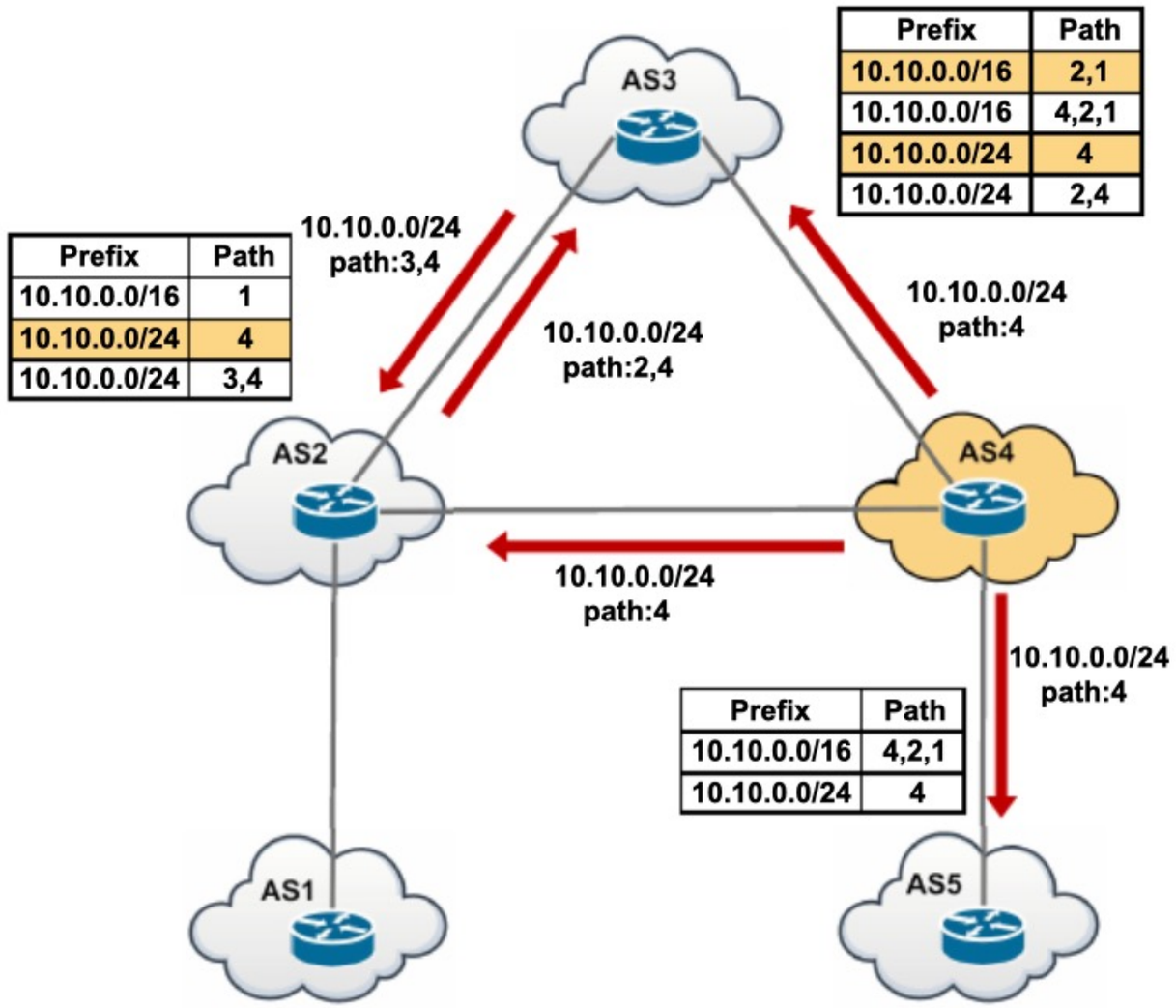
- This is the most widely propagated type of hijacking since all ASes between the attacker and the victim are affected

- this type of hijacking can be globally propagated when there is no other advertisement or filtering for this route

https://t.me/learningnets

# BGP Sub-prefix Hijacking

AS4 announces a prefix 10.10.0.0/24 which is a part of the prefix 10.10.0.0/16 owned by AS1
All ASes in this example chooses AS 4 as an Origin AS




https://t.me/learningnets

# Famous Sub-Prefix Hijack

- 2008 Pakistan Telecom and Youtube
- Pakistan Telecom wanted to block access to Youtube by announcing /24 of Youtube's /22 blocks unintentionally
- Good percentage of Youtube traffic was redirected to Pakistan Telecom
- In 2010 Chona Telecom leaked 50k prefixes, 2017 Russian SP leaked 80 important prefixes

https://t.me/learningnets

# BGP Sub-Prefix and its AS Hijack

- Attacker Announces attacker announces a fake path to a subnet of a target prefix

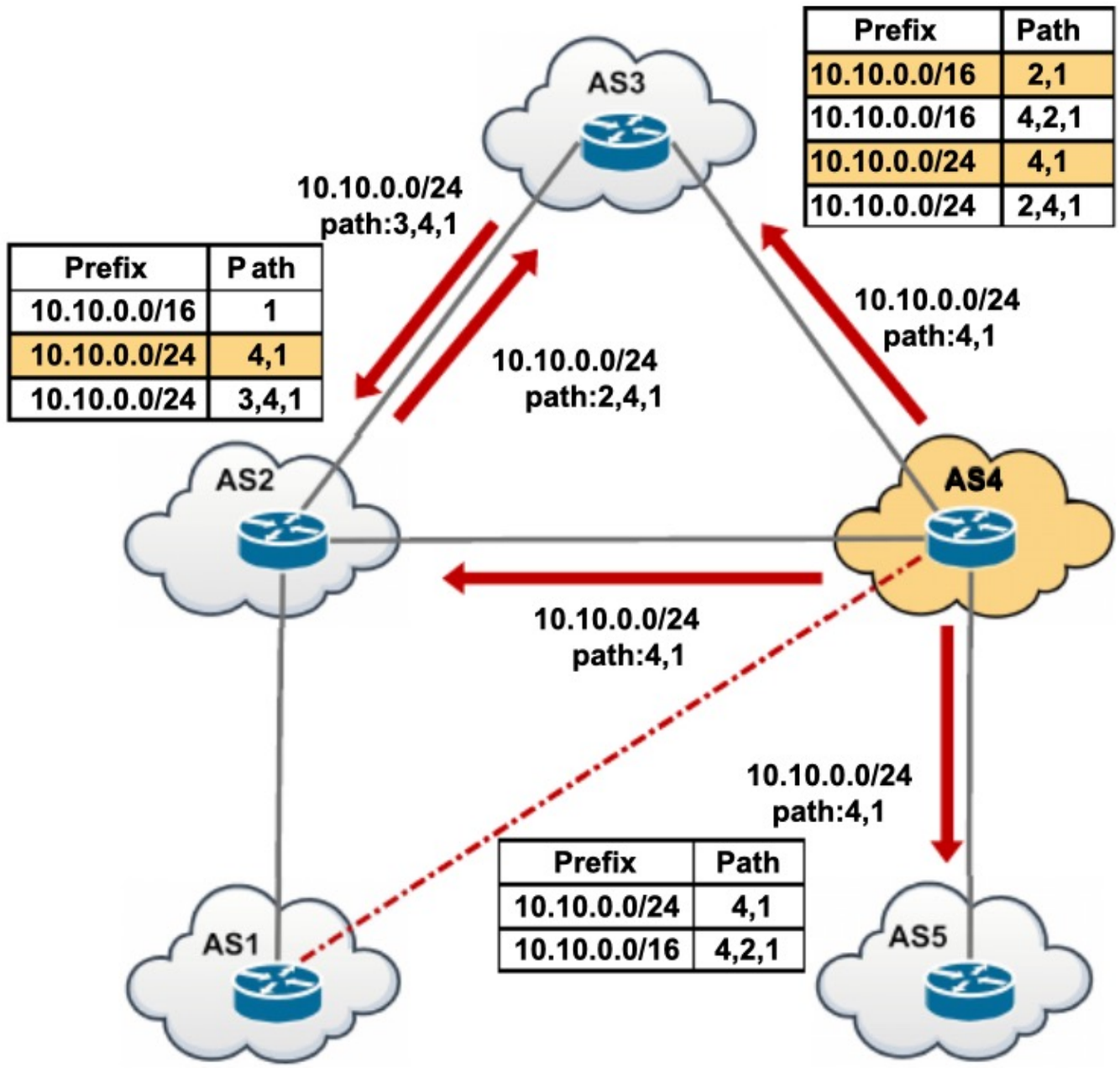
- Using a fake path with sub-prefix hijack represents a critical challenge for detection as the attacker does not claim to own a full prefix length

- This attack type is considered as most difficult to detect

https://t.me/learningnets

# BGP Sub-Prefix and its AS Hijack

**AS 2 , 3, and 5 believe that AS4 is the Origin AS for the prefix 10.10.0.0/24**




https://t.me/learningnets

# What is Traffic Interception?

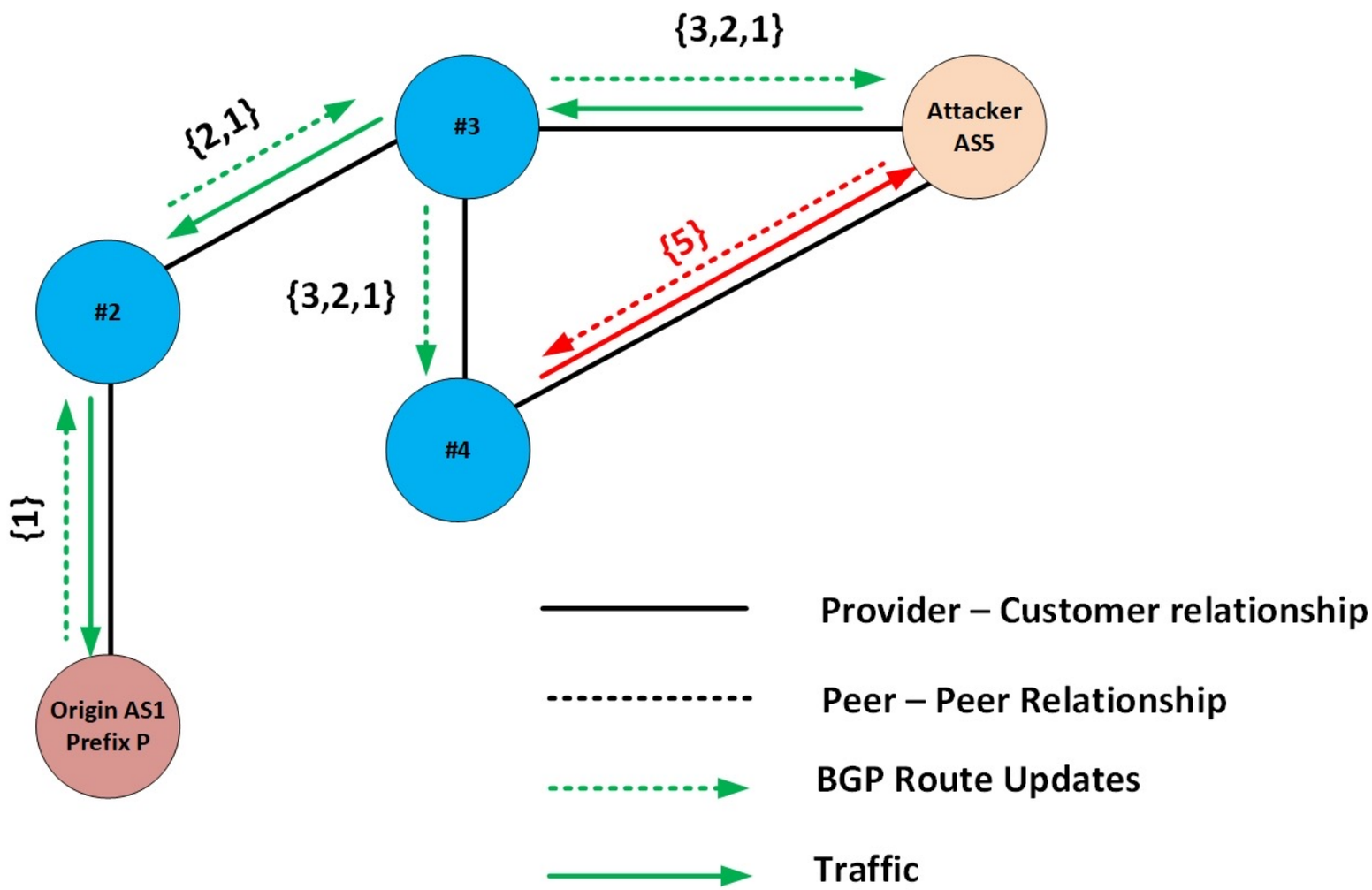
- After hijacking the prefix, malicious AS can also forward the hijacked traffic back to real destination and this type of attack is called Traffic Interception

- As the traffic reach to the destination, connectivity is not disrupted and interception is transparent to the victim Ases

- This type of attack can lead to a Man-in-the-middle attack which allows a malicious AS to eavesdrop or modify the traffic

https://t.me/learningnets

# What is Traffic Interception?




https://t.me/learningnets

# How to detect and mitigate Route Leaks, Hijacks and Interceptions?

- IP hijacking can be prevented to some extent by means of filters and hijack detection systems

- Announcements by customer ASes and peer ASes in which prefixes are out of the allocated range can be filtered

- If route filters at the links between providers and their customers are properly configured in order to prevent customer ASes from advertising the routes for the prefixes which do not belong to them

https://t.me/learningnets

# Challenges with Prefix Filters to mitigate Hijacking

**<u>According to following reasons, Prefix Filters is insufficient and difficult:</u>**

- To install ingress filters, it is not always possible for providers to know which prefixes are assigned to which customers. If customers have multiple providers, they may have different address prefixes from different providers

https://t.me/learningnets

# Challenges with BGP Prefix Filters to mitigate Hijacking

- And enforcing ingress filters in peering edges is also difficult as it is not knowable that peer ASes allocate which addresses to their customers

- Even if route filters are installed in ingress points, when there is one provider that does not practice route filtering, IP hijacking becomes possible

- Because of these reasons, building Filters by asking to the customer is not practical, instead IRR and RPKI repositories are used to generate filters, in the next section, Alternatives to the Filter will be explained

https://t.me/learningnets

# Alternate methods to build BGP Prefix Filters

- In addition to Filters, IRR , RPKI and BGPSEC will help to mitigate/stop Hijacks , Route Leaks and Interception

- IRR and RPKI have same goals which is Origin AS Validation, BGPSEC provides both Origin and Path Validation

- IRR and RPKI databases can be used to automatically generate prefix filters with the help of some tools

https://t.me/learningnets

# Alternate methods to build BGP Prefix Filters

- There are some tools designed to work with IRR policies to automatically generate ingress and eggress filters by parsing aut-num object, commonly used tools are IRRToolset, BGPQ3 and IRRPT

- Network operator register their announcements in the form of ROA objects and they are used by other operators either to generate filters or to validate announcements using more advanced techniques such as RPKI-to-Router protocol

https://t.me/learningnets

# Important to know about IRR

- IRR, RPKI , BGPSEC and Filters are not mutually exclusive. In a given network  all four approaches can be used. In fact having Filter and RPKI at the same time are seen as common deployment as of 2019

https://t.me/learningnets

# Three Source Database for Routing Security

- There are three databases and repositories which should be used by the network operators to document routing policy and maintain contact information.

- These are IRR, RPKI and PeeringDB

- You can publish your contact information, create route objects and Routing Registry through these systems

- You will learn all three in the next sections

https://t.me/learningnets

# Three Source Database for Routing Security

- Most of the Hijack cases can be solved with Origin Validation

- IRR Route Objects and RPKI ROAs are used to provide Origin Validation

- Cases which cannot be solved with Origin Validation might require Path Validation and there are couple solutions that provides Path Validation for BGP routes

https://t.me/learningnets

# What is Origin Validation?

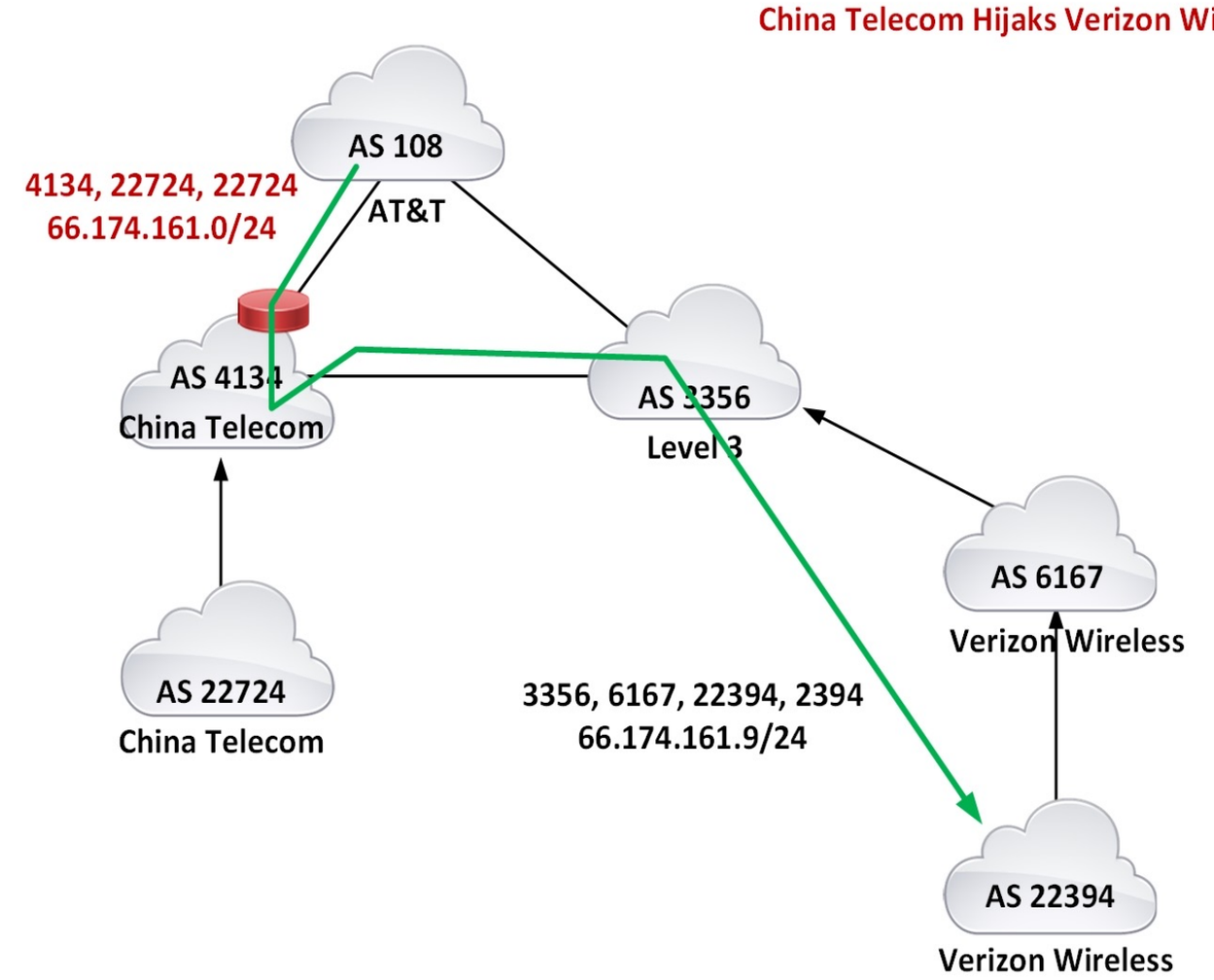
- Origin Validation is a reliable way of telling whether a BGP Route Announcement is authorized by the legitimate holder of the address space

- RPKI ROAs and IRR Route Objects are used for Origin Validation

https://t.me/learningnets

# Origin Validation Example

In this example China Telecom hijacked Verizon Wireless 66.174.161.0/24
If Origin Validation technique would be used, ATT wouldn't prefer China Telecom (Because of shorter AS path) because China Telecom is not authorized to originate that prefix




https://t.me/learningnets

# Origin Validation Example

- Current Origin Validation mechanism is RPKI (Resource Public Key Infrastructure)

- Hijack in previous example could be prevented if Victim (Verizon Wireless) would generate RPKI ROA and if AT&T would verify the ROA and prefer cryptographically signed path

- In the next section,  RPKI will be explained and how it provides Origin Validation will be shown

https://t.me/learningnets

# RPKI – Resource Public Key Infrastructure

- Resource Public Key Infrastructure (RPKI) is a specialized PKI that aims to improve the security of the Internet routing system, specifically the BGP

- It does this through the issuing of X.509-based resource certificates to holders of IP addresses and AS numbers in order to prove assignment of these resources

- These certificates are issued to Local Internet Registries (LIRs) by one of the five Regional Internet Registries (RIRs)

https://t.me/learningnets

# RPKI – Resource Public Key Infrastructure

- APNIC, ARIN, LACNIC, AFRINIC and RIPE are the five RIR who are responsible for allocation and assignment of these resources in their service regions

- As of 2019, 15% - 20% of the prefixes in DFZ have ROAs

- Compare to Route Objects in IRR, this is very small percentage

https://t.me/learningnets

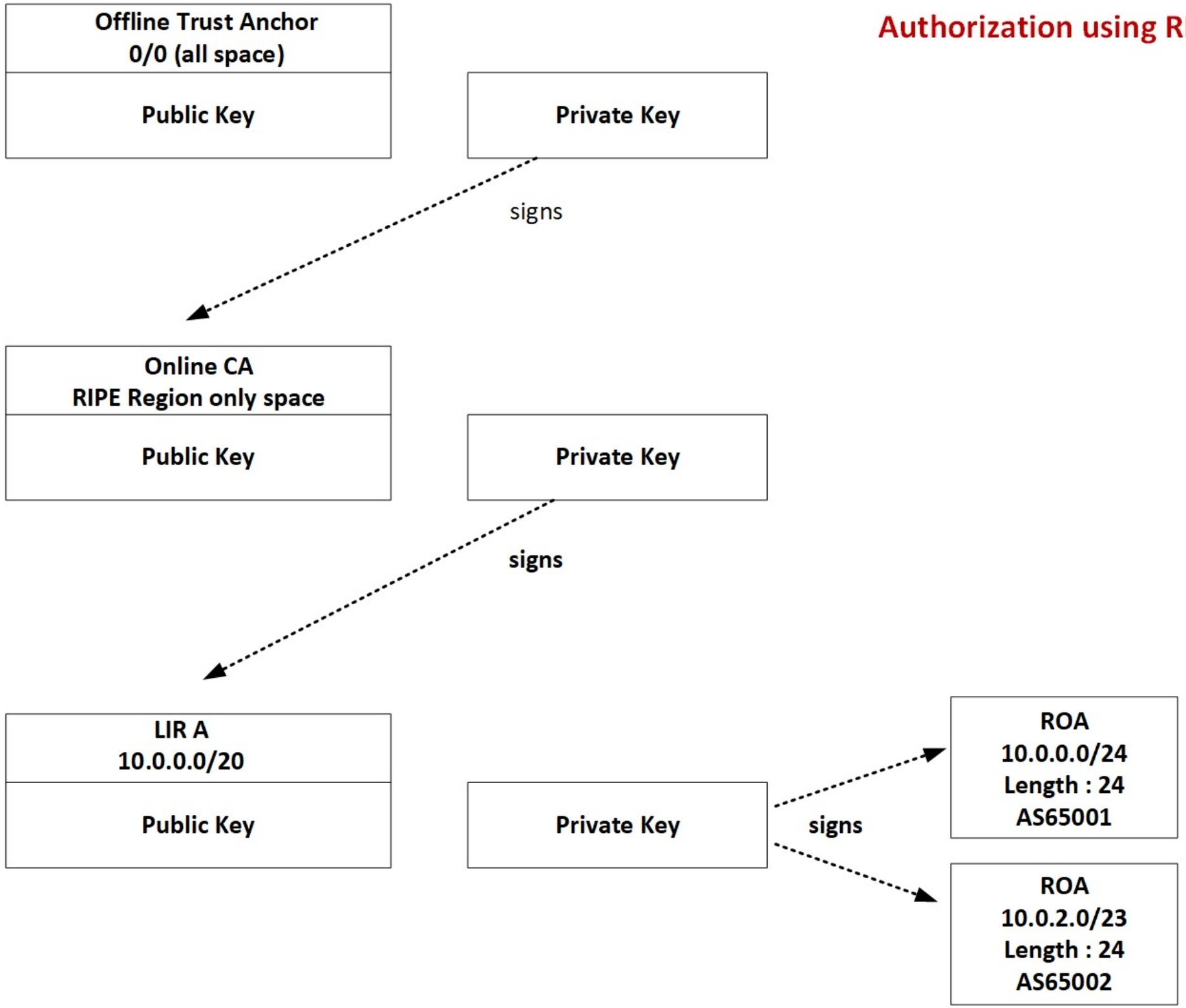
# RPKI – Resource Public Key Infrastructure

- Resource certificates allow LIRs (ISPs)  to generate Route Origin Authorizations (ROAs) which attest to which networks (specifically AS numbers) are authorized to originate which ranges of IP addresses

- This then allows other networks to determine whether route announcements are valid and should therefore be accepted, thus reducing the likelihood of fake routes being propagated across the Internet

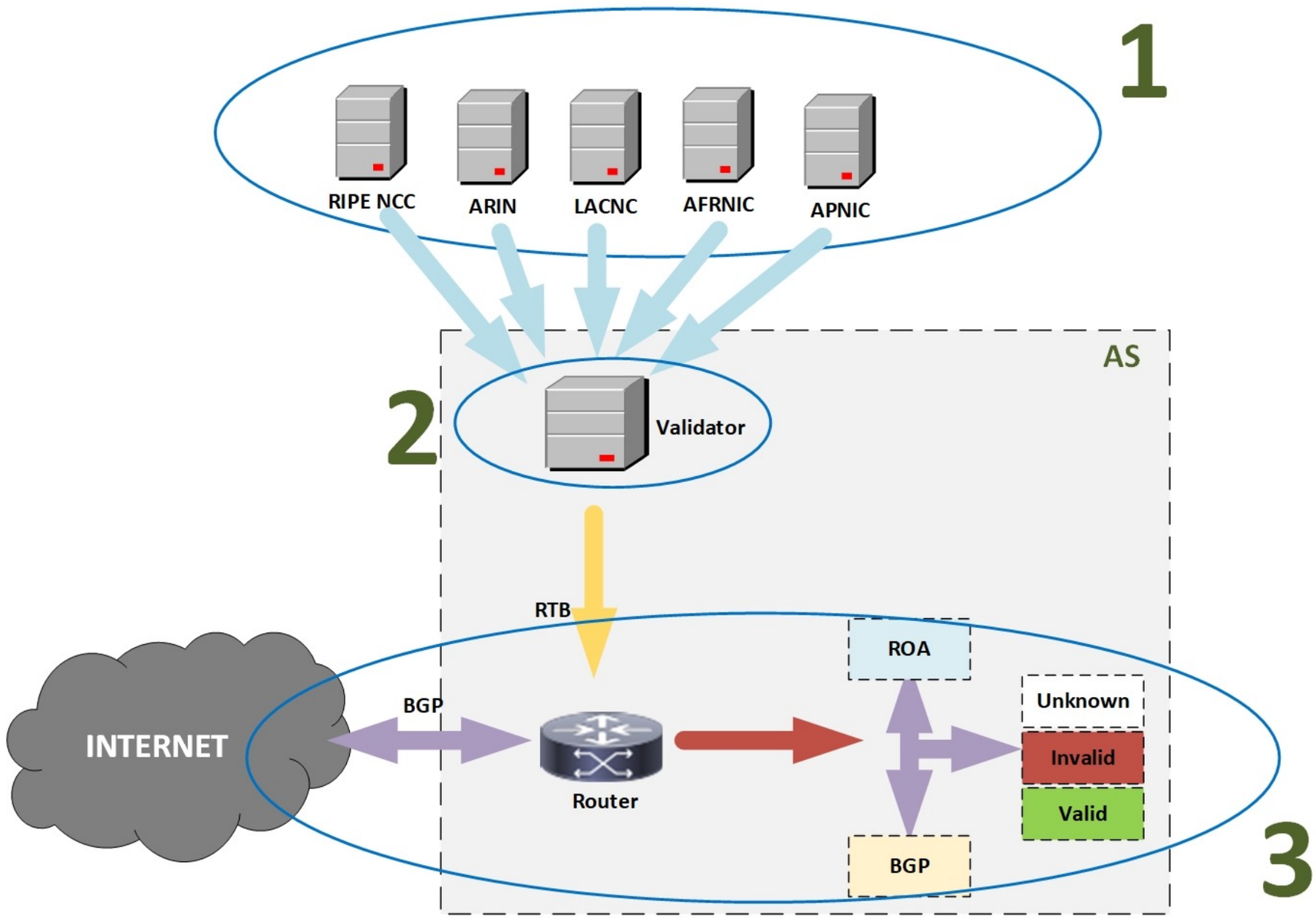
https://t.me/learningnets

# Authorization using RPKI




https://t.me/learningnets

# RPKI Overview




https://t.me/learningnets

# RPKI – Resource Public Key Infrastructure

- Each RIR acts as a CA and trust anchor for the resources assigned within their service regions

https://t.me/learningnets

# Configuring RPKI on Router Example

```
route-map rpki-loc-pref permit 10
 match rpki invalid
 set local-preference 90
!
route-map rpki-loc-pref permit 20
 match rpki not-found
 set local-preference 100
!
route-map rpki-loc-pref permit 30
 match rpki valid
 set local-preference 110
```

```
router bgp 64500
 bgp log-neighbor-changes
 bgp rpki server tcp 10.1.1.6 port 8282 refresh 5
 network 192.0.2.0
 neighbor 10.1.1.2 remote-as 64510
 neighbor 10.1.1.2 route-map rpki-loc-pref in
```

https://t.me/learningnets

# BGP Hijacks is solved with RPKI

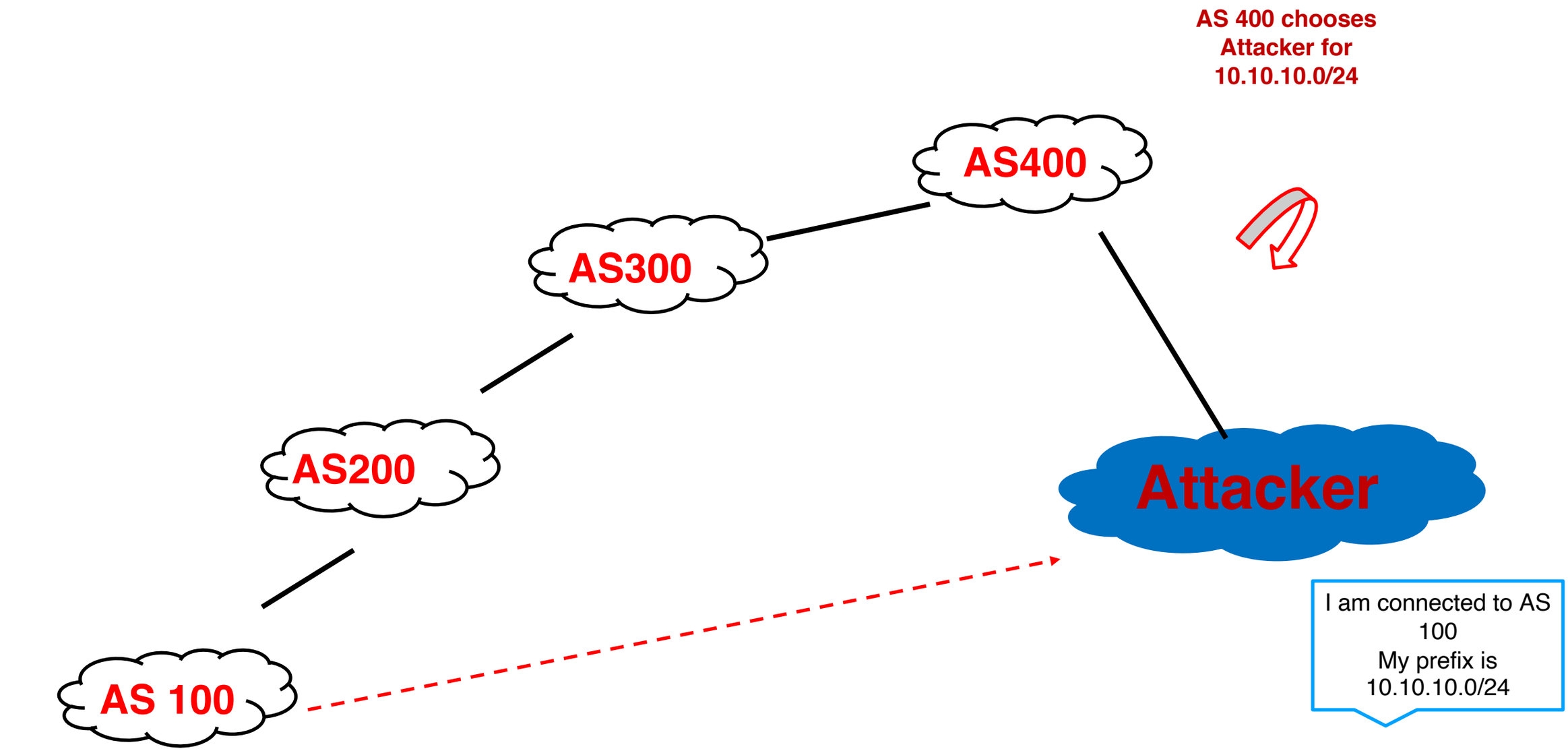
- Exact Prefix Hijacks and Sub-Prefix Hijacks can be solved with RPKI but still attacker can launch a path-shortening attack as we will see in the next example

- AS-Path Shortening is an advertising the prefix with the legitimate Origin AS but shorten it, so other ASes can see insecure path is the best due to shorter AS-Path

- Generally AS-Path Shortening is considered as Malicious attack, not as ‘Misconfiguration’

https://t.me/learningnets

# Path-Shortening Attack cannot be solved with RPKI




https://t.me/learningnets

# RPKI without BGPSEC?

- RPKI provides Origin Validation, not Path Validation

- If BGPSEC is not available, lack of path validation can be resolved with densely peering (Google and Akamai has more than 130 peering facilities in common)

- This prevents Path shortening attacks which cannot be solved with RPKI, thus densely peering is considered an advantage on top of RPKI

https://t.me/learningnets

# Do everyone really need an RPKI?

- Because of centralization of the Web, if a few largest companies deploy RPKI, millions of people benefit from RPKI

- These are large content providers (OTTs) such as Google, Akamai, Cloudflare, Amazon, Facebook, Microsoft etc.

https://t.me/learningnets

# RPKI Summary

- It is available since 2011
- RPKI is a security framework for verifying the association between resource holders and their Internet resources
- Attached digital certificates to network resources upon request that lists all resources held by the member
- -- AS Numbers
- --IP Addresses
- Operators associate these two resources through ROA
- Provides Origin Validation, not Path Validation
- Offline cryptography , no Router resource issue

**Last but not least RPKI doesn’t prevent Route Leaks**

https://t.me/learningnets

# IRR – Internet Routing Registry

- IRR is a public database of Internet route object. IRRs are used for determining and sharing route and other related information used for configuring routers

- Using IRR, networks exchange their routing policies with each other

- If the Regional Internet Registry (RIR) in your region operates an IRR it should be used to document the network routing policy and related route announcements

https://t.me/learningnets

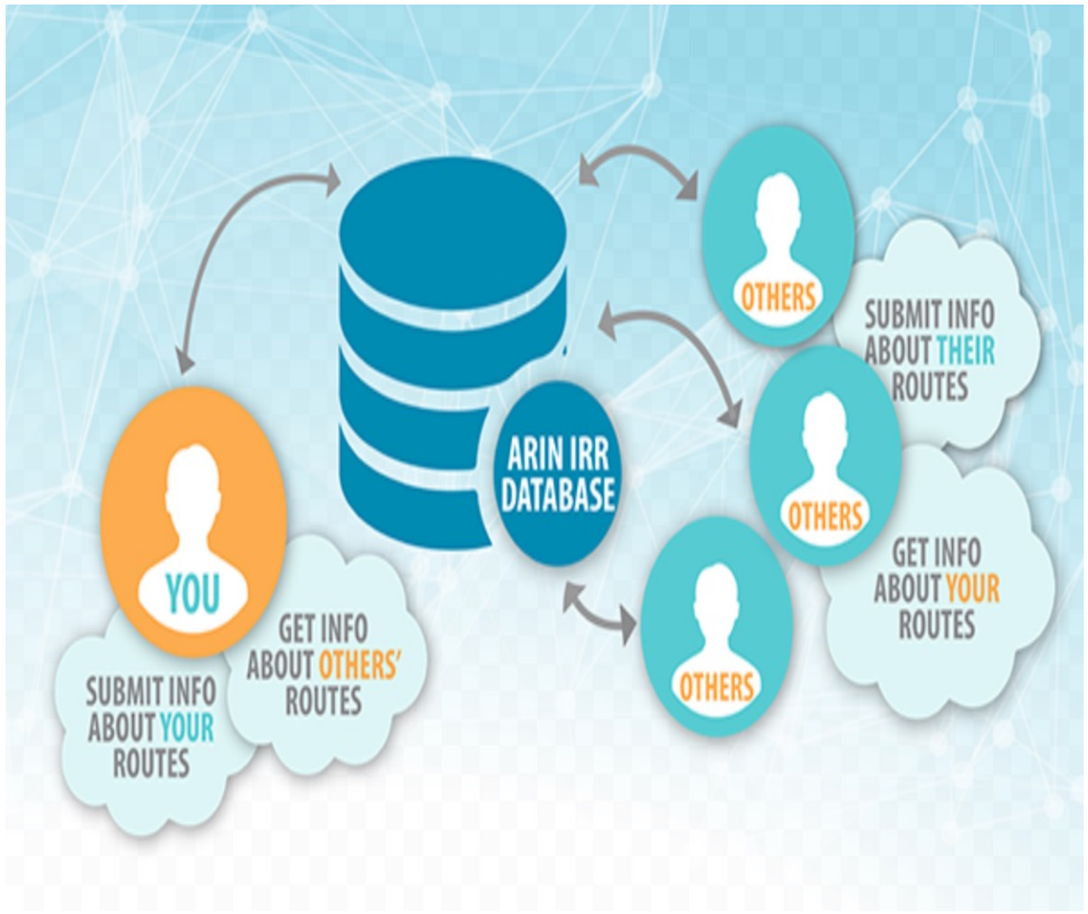
# Example of BGP Routing Policy

- Who are my BGP peers ? Customers, peers, upstream
- What routes are originated by each neighbor?
- What routes are Imported from each neighbor?
- What routes are exported to each neighbor?
- What routes are preferred when multiple routes exist?
- What to do if no route exists?

https://t.me/learningnets

# IRR – Internet Routing Registry

- Route Objects are created in the IRR database




- Route Objects consist of IP Prefix , AS Number which announces the prefix and Origin of the Registry (Example RIPE NCC)

https://t.me/learningnets

# IRR – Internet Routing Registry

- When you want to create a Route Object, you should be the Prefix Holder at least, some IRRs require ASN owner to approve Route Object creation (Ex: RIPE NCC)

- Difference between Route Objects in IRR and the RPKI ROA is, ROA digitally signs the prefixes and the Maximum length of the prefix can be specified with ROAs, not with the Route Objects

https://t.me/learningnets

# IRR – Internet Routing Registry

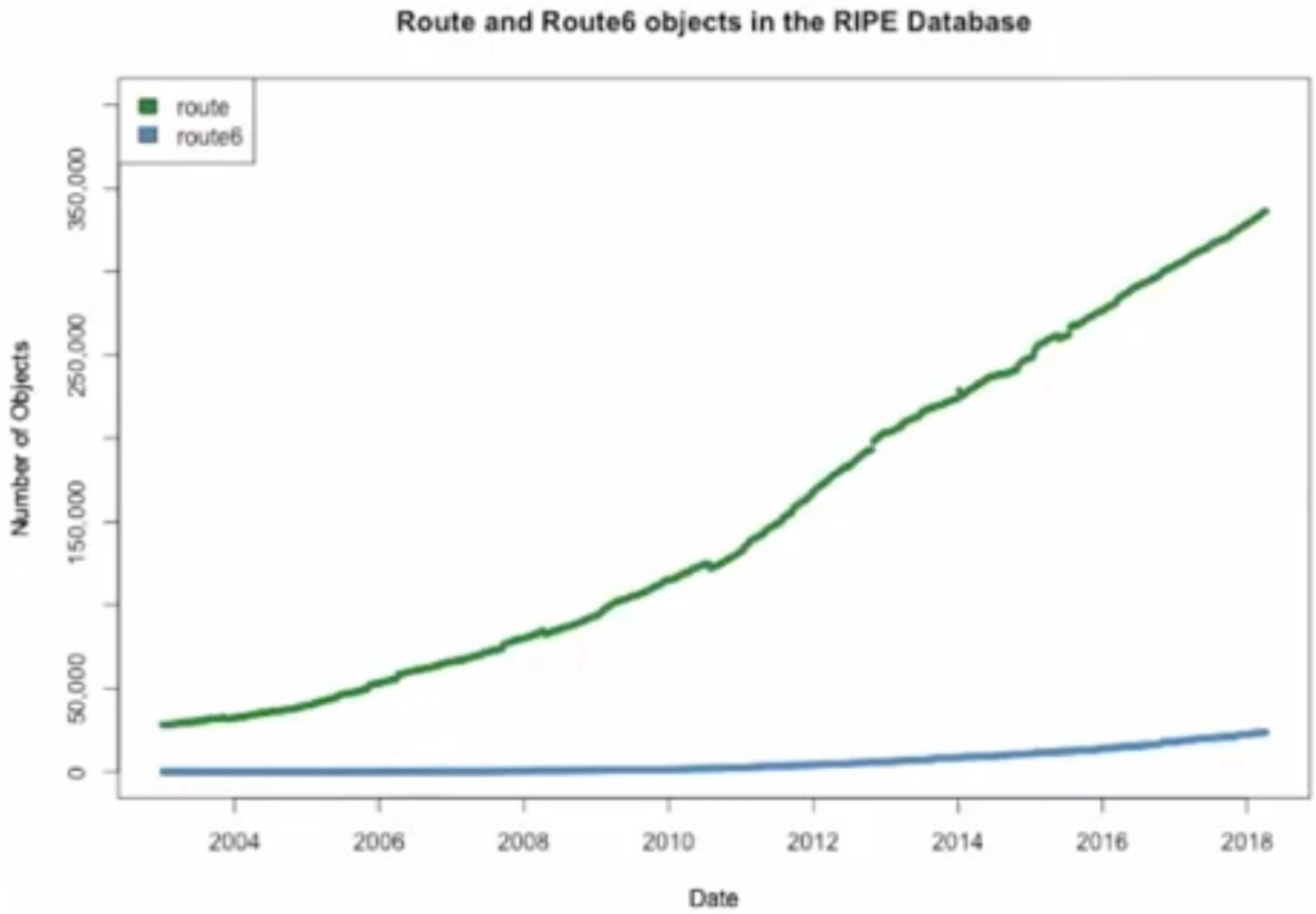
- There are many IRR (34 as of 2019), most widely used ones are RIPE Database and RADB

- **Based on Internet Society below IRR should be used for given regions:**

| Region | Preferred IRR | Alternative IRR |
|---|---|---|
| America | ARIN | RADb/NTTCOM |
| Africa | AFRINIC | RADb/NTTCOM |
| Asia Pacific | APNIC | RADb/NTTCOM |
| Europe | RIPE NCC | RADb/NTTCOM |
| Latin America and Caribbean | RADb | NTTCOM |

https://t.me/learningnets

# Number of Route and Route 6 Objects in RIPE Database




https://t.me/learningnets

# IRR – Internet Routing Registry

- IRR supposed to verify the holdership, which mean they need to verify whether the ASNs which announce the prefixes are valid owner of the prefixes

https://t.me/learningnets

# IRR – Internet Routing Registry

- Unfortunately lots of IRR don’t verify the holdership

- This create legitimacy issue as people can create fake announcements through IRR databases

https://t.me/learningnets

# How IRR Works?

- Routing policy information is documented in RPSL (Routing Policy Specification Language)

- Information related to an Internet resource (AS number, customer cone ,routes etc.) , or supporting functions are contained within RPSL objects and stored in IRR Database

- There are many Route Objects in IRR but we will cover three types of Objects defined in IRR Database

https://t.me/learningnets

# How IRR Works?

## IRR Objects:

- AUT-NUM
- Route/Route6
- AS-SET

- In addition to these objects, other IRR objects are; inet-rtr, peering-set, rtr-set, filter-set. Each object has its own purpose. Together they express routing poliies

https://t.me/learningnets

# IRR – Aut-Num Object

- Aut-num objects contain the registration details of the Autonomous System Number by the RIR

```
aut-num:        AS64500
descr:          Provider 64500
remarks:        ++ Customers ++
mp-import:      from AS64501 accept
                AS64501
mp-export:      to AS64501 announce ANY
mp-import:      from AS64502 accept
                AS64502
mp-export:      to AS64502 announce ANY
remarks:        ++ Peers ++
mp-import:      from AS64511 accept
                AS64511:AS-A::
mp-export:      to AS64511 announce
                64500:AS-ALL
remarks:        ++ Transit ++
mp-import:      from AS64510 accept ANY
                except FLTR-BOGONs
mp-export:      to AS64510 announce
AS64500:AS-ALL
```

Here is an example of a basic aut-num object registered by AS64500.

https://t.me/learningnets

# IRR – Route/Route6 Object

- Route/route6 objects contain routing information for IPv4/IPv6 address space resources. They show what routes that an AS originates

```
route:          2001:db8:1000::/36
descr:          Provider 64500
origin:         AS64500
mnt-by:         MAINT-AS64500
created:        2012-10-27T12:14:23Z
last-modified:  2016-02-27T12:33:15Z
source:         RIPE
```

The route6 object above shows that AS64500 is allowed to announce the address prefix /36.

https://t.me/learningnets

# IRR- AS-SET Object

- AS-SET Object in IRR is used to describe your network's customer-cone which is a set of ASNs that are owned by your customers

```
as-set:          AS64500:AS-CUSTOMERS
descr:           AS64500 regional customers
members:         AS64501, AS64502
tech-c:          EXAMPLE1-AP
admin-c:         EXAMPLE2-AP
mnt-by:          MAINT-AS64500
last-modified:   2008-09-04T06:40:26Z
source:          APNIC
```

This example contains all of AS64500's customers which include **AS64501** and **AS64502**.

https://t.me/learningnets

# PeeringDB

- PeeringDB is an Open Source database for networks to share their peering information and other relevant information amongst each other

- Networks are responsible for maintaining their own records

- A PeeringDB record allows you to consolidate your network information in a single location, and look up information about other networks

https://t.me/learningnets

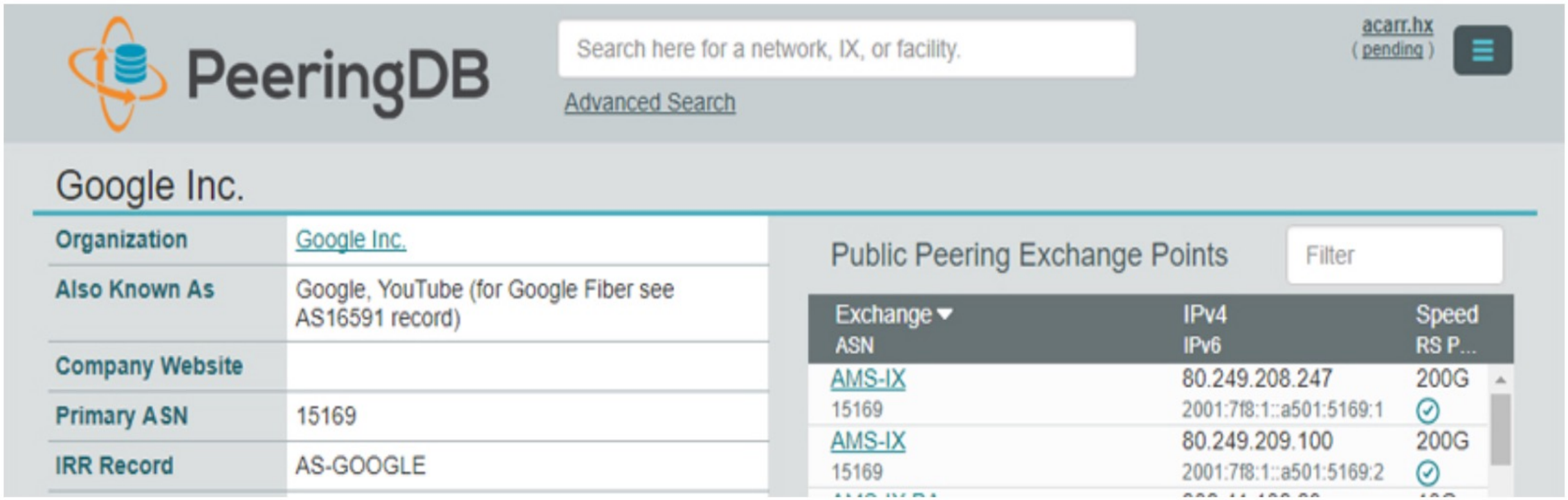
# PeeringDB

- PeeringDB records are used to supplement routing information stored in RPKI and IRR repositories

- PeeringDB allows you to publish information to let other networks know about your network, lets other network know how to contact you, and it is the first place when deciding where and whom to peer with

https://t.me/learningnets

# PeeringDB




In order to use PeeringDB, you must register an account. After registering an account, you should request affiliation to your organization, or ASN to unlock full access to data (for example, non-public network contact information).

The example above shows the PeeringDB record for Google Inc. You can see information such as the name of its IRR record, primary ASN, and public peering exchange points.

https://t.me/learningnets

# BGPSEC - Cryptographic Path Validation

- The community has considered a number of solutions that can eliminate the attacks that can be launched against the RPKI

- Building on the RPKI's guarantees that a BGP route has an authorized origin AS, BGPSEC also provides **path validation**

- BGPSEC is an IETF Standard, RFC 8205

https://t.me/learningnets

# BGPSEC - Cryptographic Path Validation

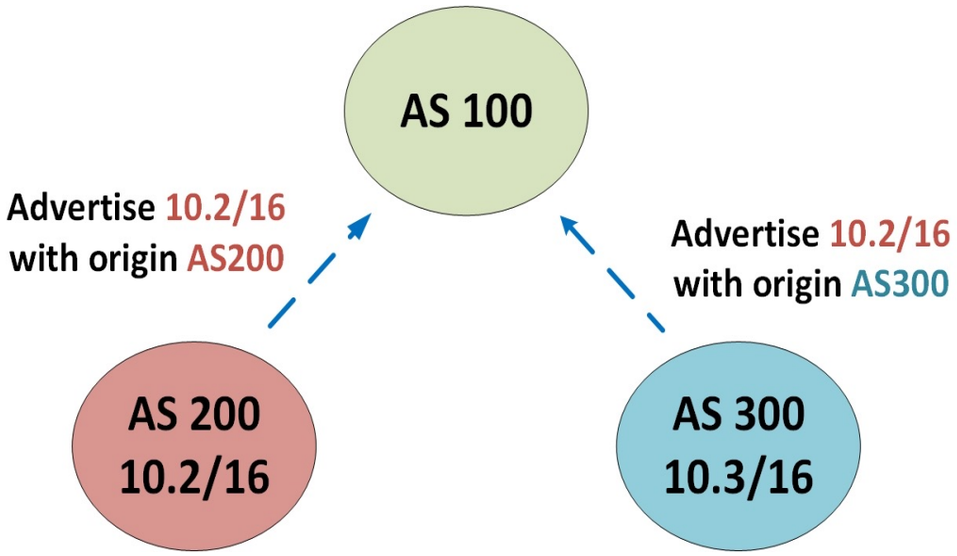
- BGPSEC builds on RPKI by adding cryptographic signatures to BGP messages

- It requires each AS to digitally sign each of its BGP messages

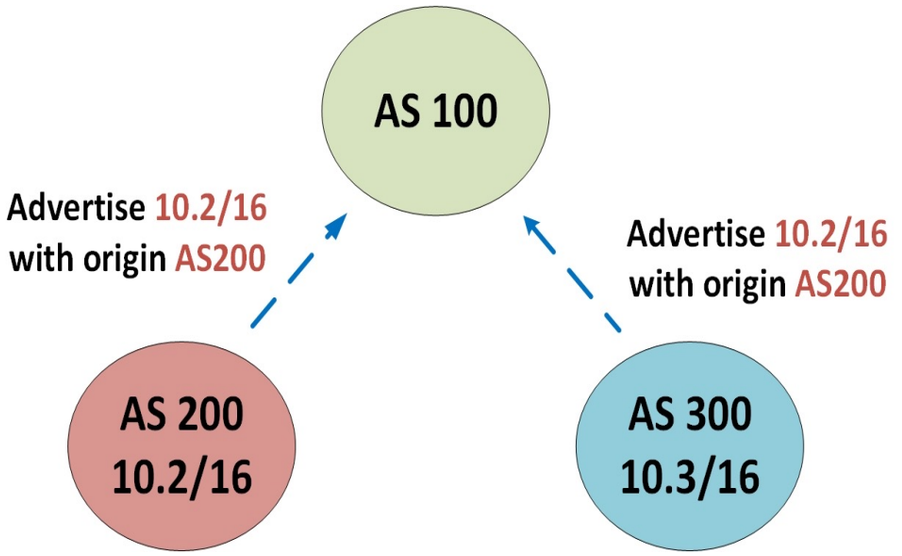
- The signature on a BGPSEC message covers (1) the prefix and AS-level path; (2) the AS number of the AS receiving the BGPSEC message; and includes (3) all the signed messages received from the previous ASes on the path

https://t.me/learningnets

# BGPSEC solves some issues which RPKI cannot solve




**Origin Validation - RPKI**




**Path Validation - BGPSEC**

https://t.me/learningnets

# BGPSEC Challenges

- RPKI requires an Offline Cryptography but BGPSEC is an online cryptographic protocol; routers must cryptographically sign and verify every BGP message they send

- This high computational overhead, which could require routers to be upgraded with crypto hardware accelerators, could slow down BGPSEC deployment

https://t.me/learningnets

# BGPSEC Challenges

- AS cannot validate the correctness of an AS-level path (and therefore filter bogus routes) unless all the ASes on the path have applied their signatures to the message

- This means the security benefits of BGPSEC apply only after every AS on the path has deployed BGPSEC

https://t.me/learningnets

# Protocol Downgrade Attack

If BGPSEC is partially deployed, it can cause Protocol Downgrade attack, which is selecting insecure short path over secure long path due to BGP Policies of the companies (GAO-Rexford)

https://t.me/learningnets

# Protocol Downgrade Attack

- BGPSEC is changing the way of how BGP works

- It replaces as-path attribute with bgpsec-path attribute

- If there is a partial BGPSEC deployment (current status of Internet), BGPSEC and Legacy routers need to communicate with each other

https://t.me/learningnets

# Protocol Downgrade Attack

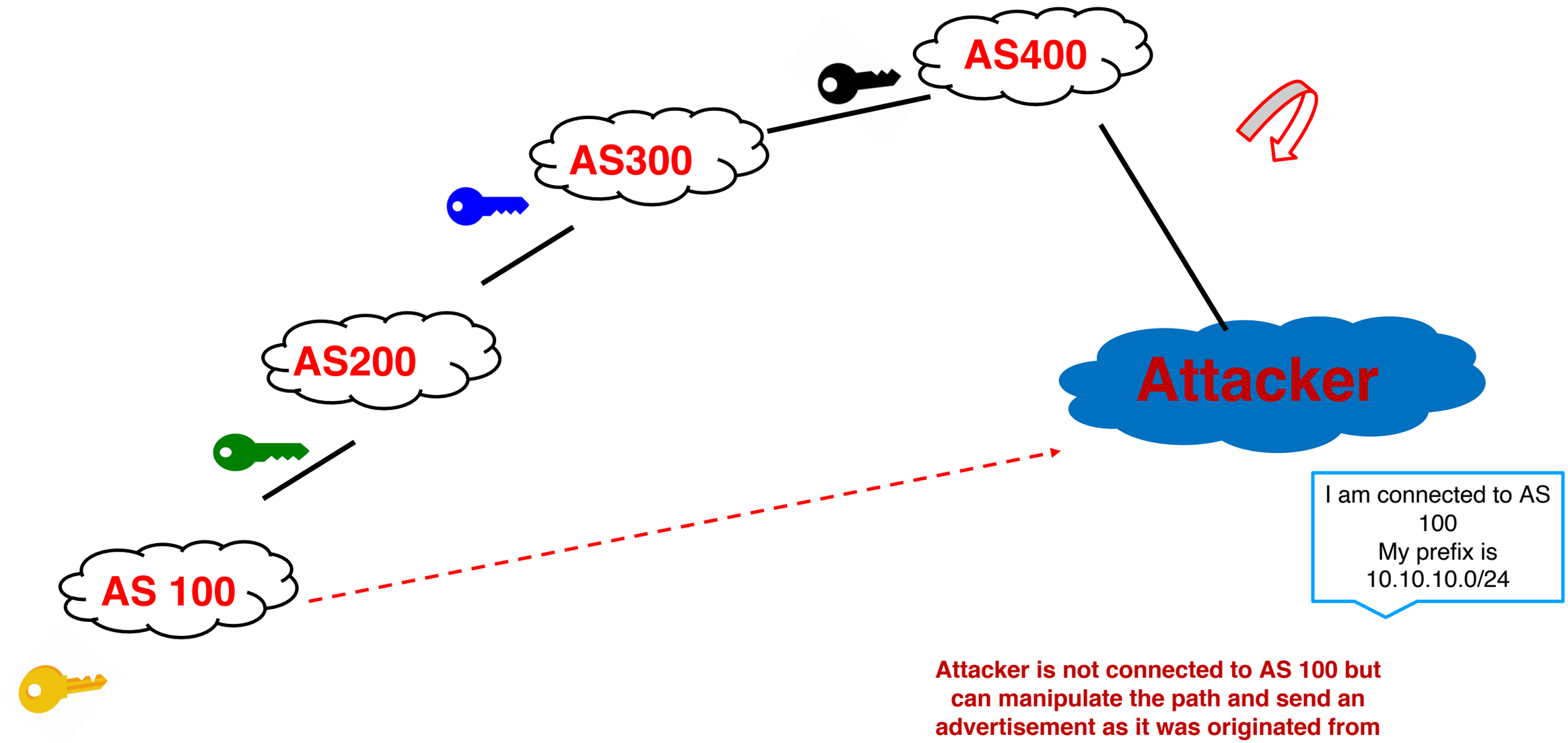
- To communicate with Legacy routers, BGPSEC speaking routers must receive and send ‘insecure’ routes

- Last but not least, if an AS considers AS-Path length is selected before Secure route via BGPSEC, short insecure path is preferred to longer secure paths , let’s have a look at the example

https://t.me/learningnets

Only Attacker is not secure , even if all ASes deploy BGPSEC, still 10.10.10.0/24 is chosen through Attacker AS, as other ASes consider AS-path length is preferred to secure path in BGP best path selection decision

AS 400 chooses Attacker AS to reach 10.10.10.0/24




Attacker is not connected to AS 100 but can manipulate the path and send an advertisement as it was originated from AS 100

Prefix 10.10.10.0/24

https://t.me/learningnets

# BGPSEC Summary

- It is an IETF Standard
- Almost no deployment yet
- Provides Origin and Path Validation
- Requires all ASes to deploy, not much partial deployment benefit
- Protocol downgrade attacks are possible
- Requires BGP messages to change, router hardware to change, so much changes..

https://t.me/learningnets

# BGP in the Datacenter

- BGP is used on the Internet facing routers which are mostly placed in the Datacenter but this is not the topic of BGP for the Internet (Inter-domain routing)

- BGP can be used as an IGP in the Datacenter and we will cover it

- RFC 7938 – Use of BGP for Routing in Large Scale Datacenters cover the usage of BGP in DC networks

https://t.me/learningnets

# BGP in the Datacenter – Why not other IGPs?

- OSPF and IS-IS were considered and used by some of the Web-scale companies initially

- Lack of Multiprotocol support by OSPF , lack of good Open source implementation for link state protocols and most importantly flooding and blast radius impact of link state protocols were the biggest factors of using BGP as IGP in the Datacenters

https://t.me/learningnets

# Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Massively Scale Datacenters or commonly known as Massively Scale, Hyper Scale or Warehouse-Scale Datacenters can have 100s of thousands of servers

## These datacenters require :

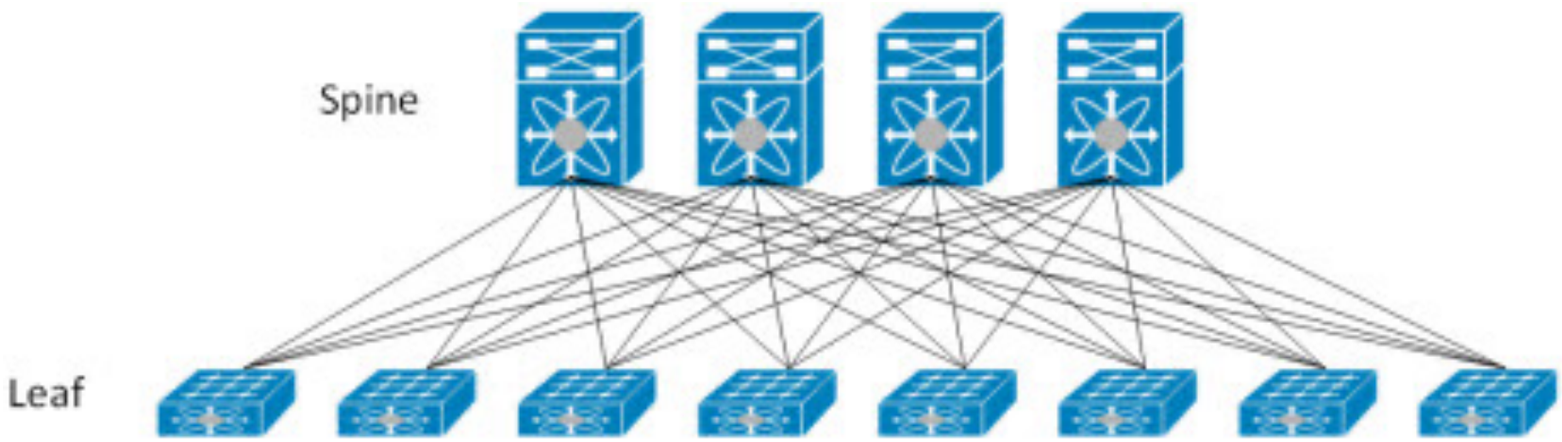
- Select a topology that can be scaled "horizontally" by adding more links and network devices of the same type without requiring upgrades to the network elements themselves

- This requires CLOS topology

https://t.me/learningnets

# CLOS Topology

- In the CLOS topology, there are Leaf and Spine Switches
- There is no shortcut links between Leaf or Spine switches




https://t.me/learningnets

# Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Narrow set of software features/protocols supported by a multitude of networking equipment vendors. (Open Source Implementation is preferred , Exa BGP , FRRouting etc.)

- Routing protocol that has a simple implementation in terms of programming code complexity and ease of operational support (State Machine of BGP vs. IGP protocols)

https://t.me/learningnets

# Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Minimize the failure domain of equipment or protocol issues as much as possible (IGP Flooding scope, Periodic database refresh vs. BGP Incremental Update)

- Allow for some traffic engineering, preferably via explicit control of the routing prefix next hop using built-in protocol mechanics (BGP 3rd part next hop allows some level of Traffic Engineering)

https://t.me/learningnets

# Routing Protocol and Topology Requirements in the Large Scale Datacenters based on RFC 7938

- Also in the hyper scale DC requirements, having protocol synchronization (Inter-dependency between protocol) is not wanted, thus having a BGP as single protocol is an important parameter to have simple and less OPEX design

https://t.me/learningnets

# Counter Arguments for using BGP in the Datacenter as Routing Protocol

- BGP is perceived as a "WAN-only protocol“ and not often considered for enterprise or data center applications

- BGP is believed to have a "much slower" routing convergence compared to IGPs

- BGP is perceived to require significant configuration overhead and does not support neighbor auto-discovery

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is for WAN

- BGP is perceived as a "WAN-only protocol“ and not often considered for enterprise or data center applications

- This is partly through, thus there are some tweaks for BGP to use in the Datacenter as Routing Protocol

- In the WAN networks, expectation from BGP is stability, in the Datacenter Stability is important but rapid notification and convergence is more important

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is Slow

- BGP is believed to have a "much slower" routing convergence compared to IGPs, timers can be tuned, BGP can converge much faster than conventional thoughts

**For Fast BGP convergence there are in general three timers**

- Minimum Route Advertisement Interval
- Keepalive and Hold Timers

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is Slow MRAI Timer

- BGP has MRAI per neighbor

- Events within this minimum interval window are collected together and sent at one shot when the minimum interval expires

- This is essential for the most stable code, but it also helps prevent unnecessary processing in the event of multiple updates within a short duration such as link flaps

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is Slow MRAI Timer

- The default value for this interval is 30 seconds for EBGP peers, and 0 seconds for IBGP peers. However, waiting 30 seconds between updates is not necessary for a densely connected network such as CLOS topology

- 0 is the more appropriate choice for MRAI timers in EBGP in the DC, because we're not dealing with routers across administrative domains when we deploy EBGP in the Datacenter

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is Slow Keepalive and Hold Timer

- By default, the keepalive timer is 60 seconds and the hold timer is 180 seconds

- This means that a node sends a keepalive message for a session every minute. If the peer does not see a single keepalive message for three minutes, it declares the session dead

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is Slow Keepalive and Hold Timer

- By default, for EBGP sessions for which the peer is a single routing hop away, if the link fails, this is detected and the session is reset immediately

- What the keepalive and hold timers do is to catch any software errors while the link is up but has become one-way due to an error, such as in cabling issue

https://t.me/learningnets

# Counter Arguments for BGP in DC – It is Slow Keepalive and Hold Timer

- Some operators enable Bidirectional Forwarding Detection (BFD) for sub-second detection of errors due to cable issues. However, to catch errors in the BGP process itself, you need to adjust these timers

- Inside the data center, three minutes is too much. Recommended values configured inside the data center are 3 seconds for keepalive and 9 seconds for the hold timer

https://t.me/learningnets

# Counter Arguments for BGP in DC – Lack of BGP Neighbor Auto-Discovery

- There are two IETF draft now for BGP neighbor auto discovery

1. BGP LLDP Peer Discovery
2. BGP Neighbor Discovery Draft

https://t.me/learningnets

# Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- In BGP, neighbor adjacency is configured manually, putting a neighbor IP address and ASN into the BGP

- Neighbor Auto Discovery is a desired future to reduce OPEX when BGP is used inside the Datacenter

https://t.me/learningnets

# Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- This IETF draft can change the behavior of manual adjacency setup, allowing the BGP adjacency on the point-to-point links to be established automatically, using the LLDP protocol

- LLDP is an Industry standard

https://t.me/learningnets

# Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- It is intended to replace proprietary Cisco Discovery Protocol (CDP)

- With LLDP, network devices such as switches operating on Layer 2 (Link layer) of OSI model can collect upper layer information of neighboring device, such as IP address, OS version etc.

https://t.me/learningnets

# Counter Arguments for BGP in DC – BGP LLDP Peer Discovery

- BGP peering discovery using LLDP could be used in large Layer 3 data centers where eBGP is being used as a single routing protocol

- Deployment of BGP with enabled BGP peering discovery using LLDP in large data centers in the future would significantly lower the BGP configuration overhead

https://t.me/learningnets

# Counter Arguments for BGP in DC – BGP Neighbor Discovery Draft

- This approach suggests the way of BGP hello message

- Instead of using other protocols such as LLDP, the draft introduces a new BGP Hello message

https://t.me/learningnets

# Counter Arguments for BGP in DC – BGP Neighbor Discovery Draft

- The message is sent periodically on the interfaces where BGP neighbor auto-discovery is enabled to the multicast IP address using UDP port 179

- The hello message contains ASN of the sender along with IP address, router id etc

https://t.me/learningnets

# BGP Path Hunting

- When BGP is used in the DC, based on the AS allocation, it might suffer from the BGP Path Hunting behavior

- BGP Path Hunting will slow down the convergence when there is a failure in the network, link or node failure

- Let’s have a look at the details of BGP Path Hunting before start talking about BGP ASN allocation

https://t.me/learningnets

# BGP Path Hunting

- Without topology information a Router does not know the physical link state of every other node in the network, it doesn't know whether the route is truly gone (because the node at the end went down itself) or is reachable via some other path

- That's why a Router proceeds to hunt down reachability to the destination via all its other available paths. This is called path hunting

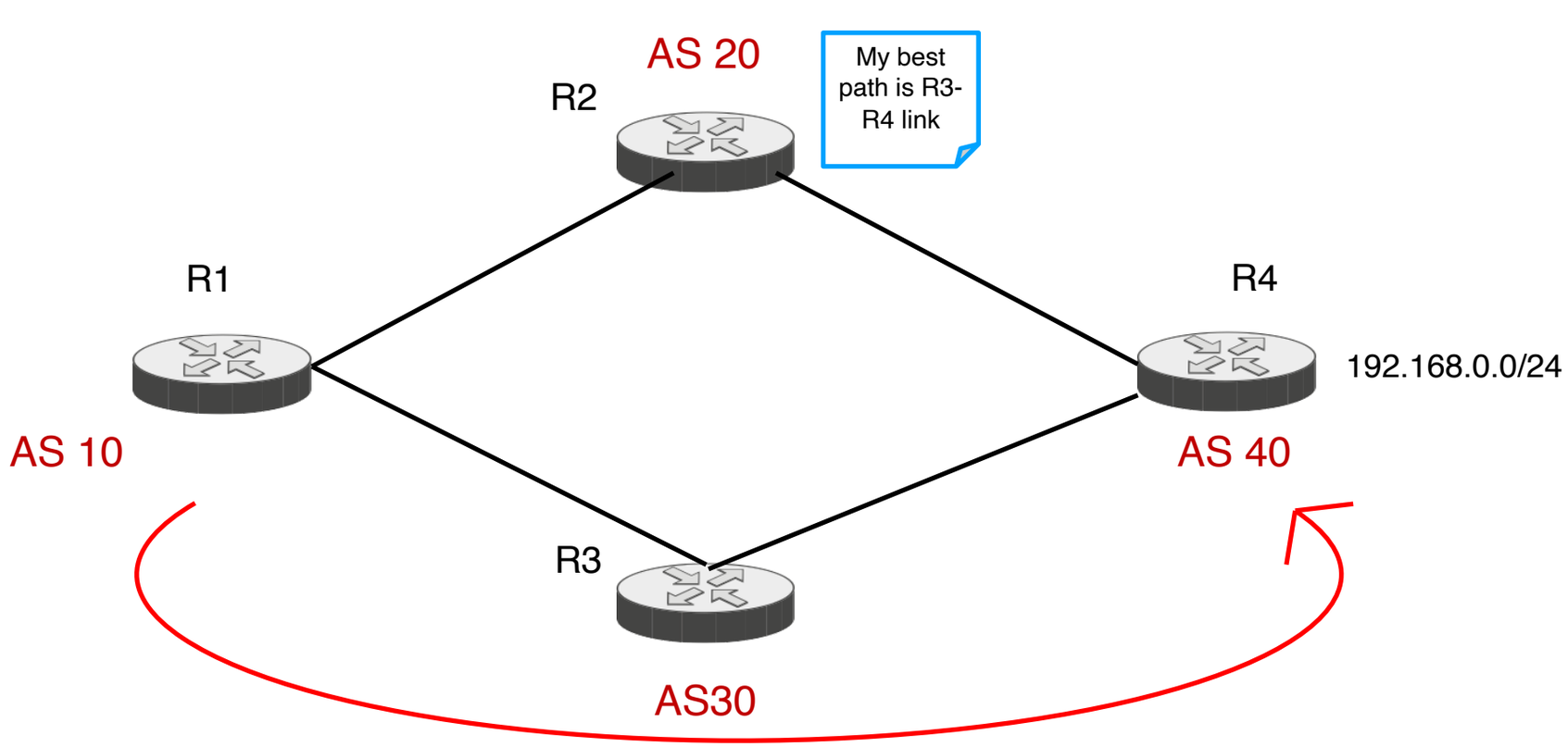
https://t.me/learningnets

# BGP Path Hunting

Let's assume R1 selected best path to 192.168.0.0/24 as R3

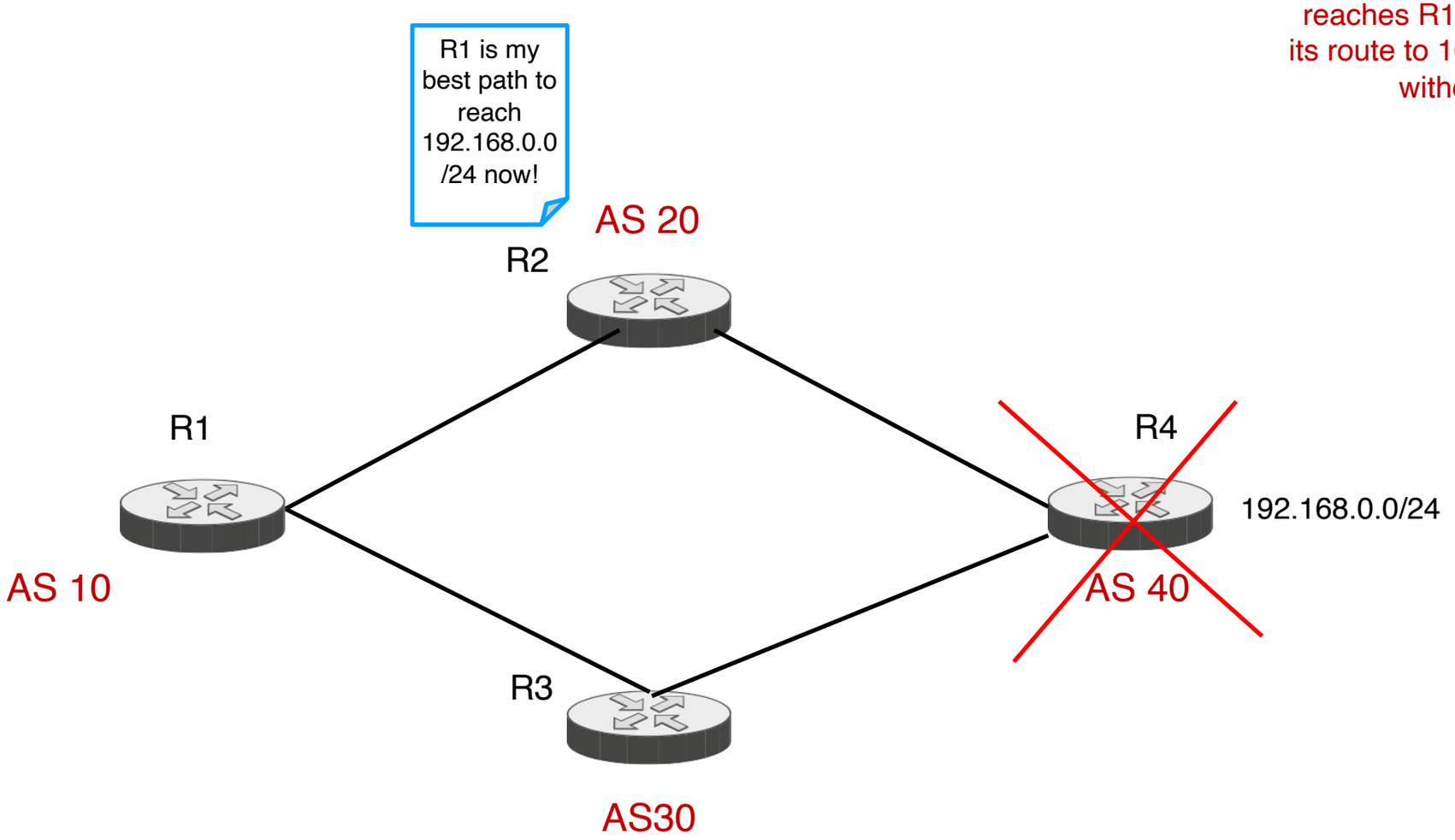
R1 advertises [R1 R3 R4] As-Path to R2

R2 accepts the advertisement but doesn't use it , as R2 has shorter path to 192.168.0.0/24




https://t.me/learningnets

Now, when the Router R4 fails, R2 loses its best path to 10.1.1.1, and so it re-computes its best path via R1, AS_PATH [R1, R3, R4] and sends this message to R1.
R2 also sends a route withdrawal message for 10.1.1.1 to R1 When R3's withdrawal to route 10.1.1.1 reaches R1, R1 also withdraws its route to 10.1.1.1 and sends its withdrawal to R2




https://t.me/learningnets

# BGP Path Hunting

- EBGP AS number allocation will trigger path hunting when there is a failure to the destination

- Path Hunting will slow down the convergence which is not good for the Datacenter BGP

- We will look at next how ASN allocation should happen to avoid BGP Path Hunting behavior when EBGP is used inside the Datacenter

https://t.me/learningnets

# ASN Numbering Schema when EBGP is used inside Datacenter

- All ToR (Top of Rack)  switches (Sometimes referred as Tier 3 switches) are assigned their own ASN , unique ASN

- Leaf Switches (Tier 2 Switches) inside a pod have a same ASN, but leaves in different pod have a unique ASN.POD sometimes referred as Cluster

- Spines share a common ASN, Spine switches sometimes referred as Tier 1 devices

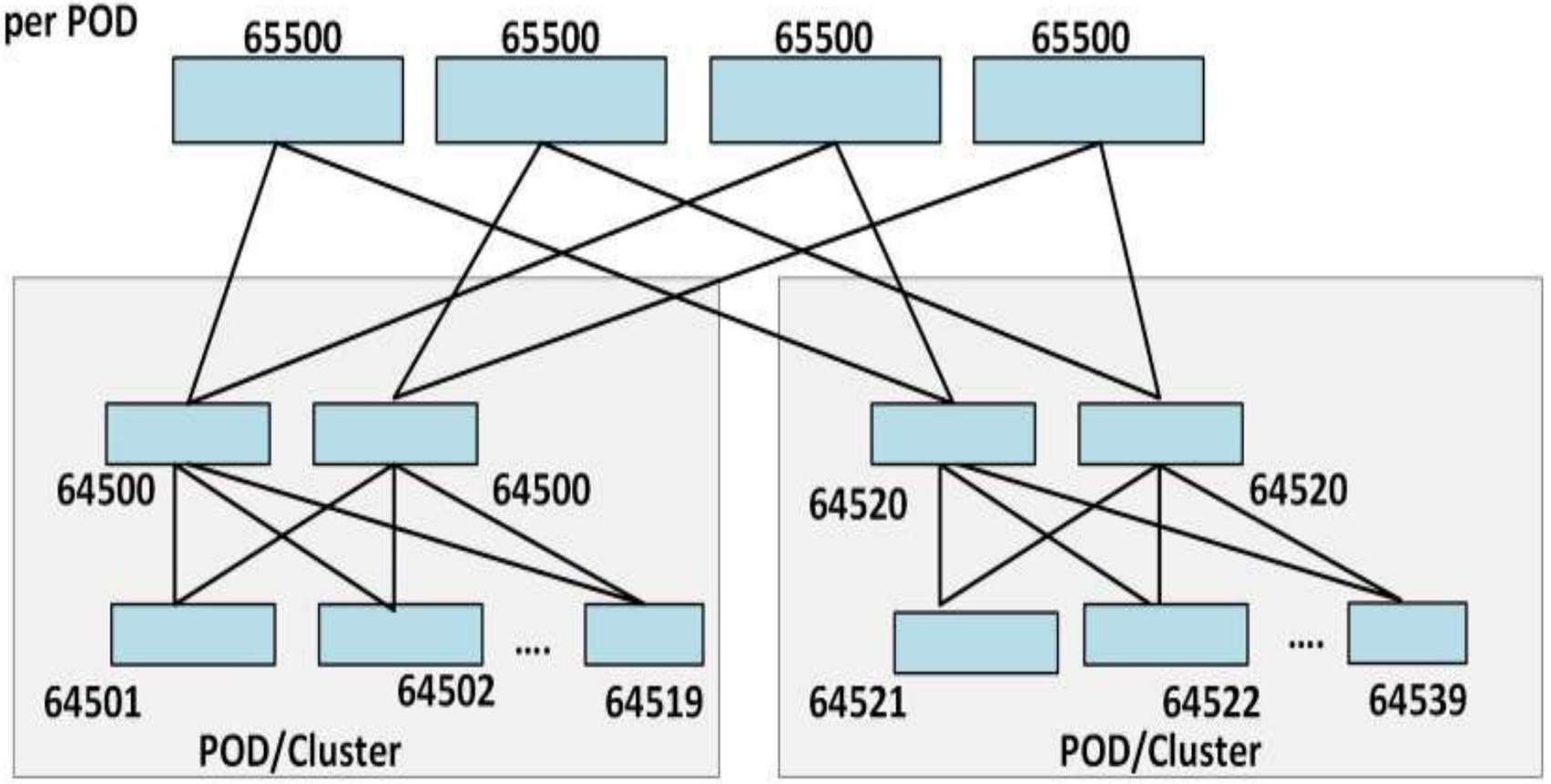
https://t.me/learningnets

# Recommended BGP ASN Allocation Schema for 3 Tier CLOS Networks

Unique AS per ToR

Same AS on Leaf switches per POD

Same AS on all SPINES




https://t.me/learningnets

# Possible Problems with BGP ASN Allocation Schema

- If 2 byte ASN space is used, Private ASN range is recommended as there might be mistake to leak Public ASN to the Peers or Transit

- Since Operator knows most of the public AS numbers of the large companies, separating AS number for DC usage is better for troubleshooting

https://t.me/learningnets

# Possible Problems with BGP ASN Allocation Schema

- 2 byte ASN space Private range is limited to 1023 AS

- There might be much more TOR switches in the DC than 1023 AS

- Two Options : Either use 4 byte ASN or assign same AS numbers on different POD/Cluster’s TOR switches

https://t.me/learningnets

# Possible Problems with BGP ASN Allocation Schema

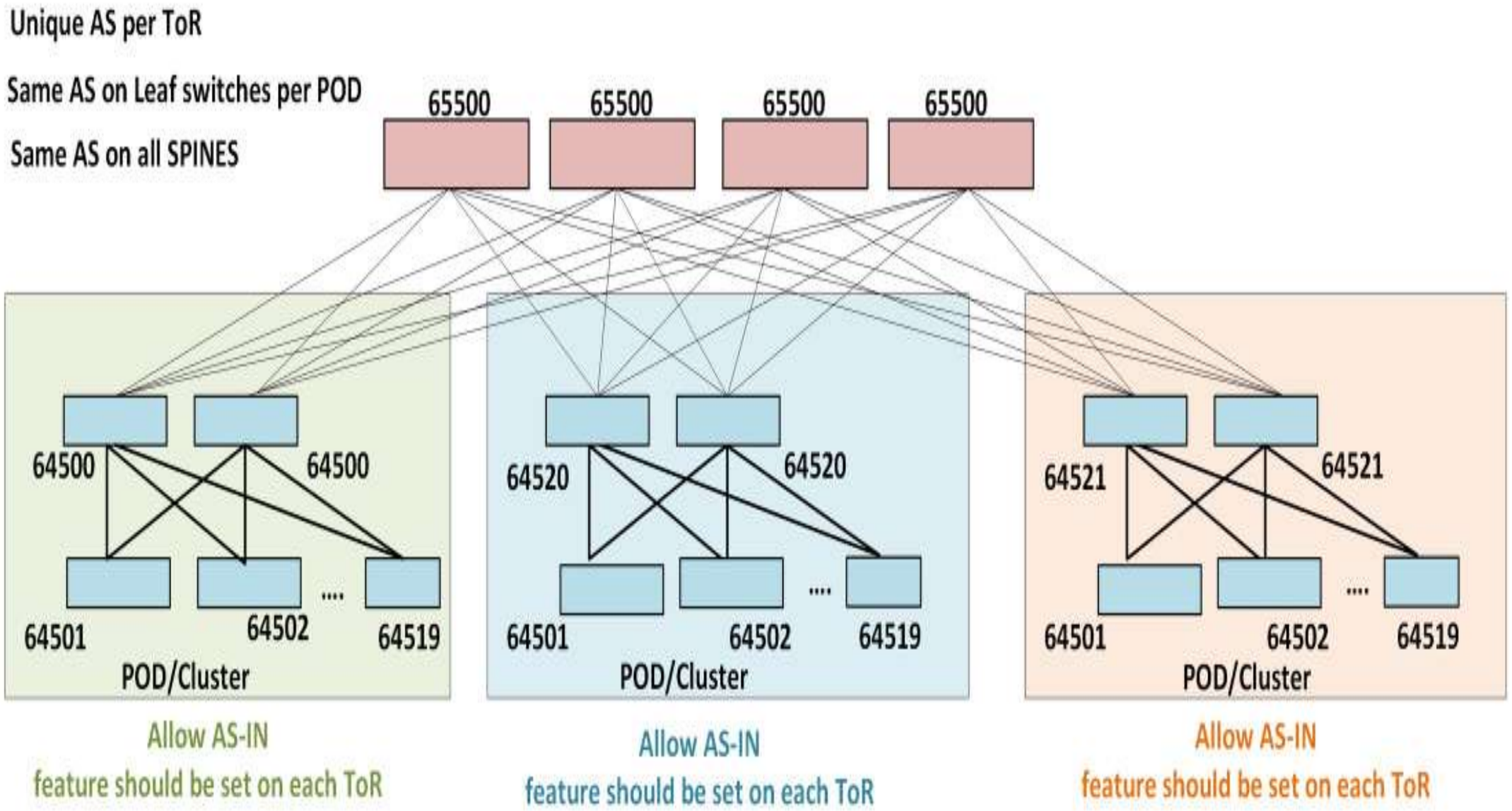
- 4 byte ASN is not still supported by some BGP implementations thus it might limit the vendor selection of DC equipment

- Also using 4 bye ASNs puts additional protocol complexity

- If same AS numbers are used in the different Cluster TOR switches, when there is a traffic between separate Cluster TOR, Allow-as in feature is required

https://t.me/learningnets

# Recommended BGP ASN Allocation when Private 2 Byte Private ASN is used




https://t.me/learningnets

# BGP in the Wide Area Network

- BGP is designed first for the Wide Area Network and today it is mostly still used for the Wide Area Network

- When it is first invented, it was for the IPv4 Unicast address family but today BGP supports 20 different address families which are mostly used for the WAN use cases

- There are some Datacenters (Hyper-scale/MSDC) which use BGP in their Datacenter as IGP

https://t.me/learningnets

# BGP in the Wide Area Network

- BGP is used for the Inter-domain routing mainly

- It is the only protocol which is used in the Global Internet

- It is the most scalable routing protocol and for the Inter-domain routing scalability is must as there are 750k prefixes in the Global routing table which is commonly known as Default Free Zone (DFZ)

https://t.me/learningnets

# BGP in the Wide Area Network

- Stability in the Default Free Zone is important since the failure impacts so many networks and so many routers in the Internet

- Policy is the key function of BGP which companies in the Internet express their business intent with BGP

https://t.me/learningnets

# BGP in the Wide Area Network

- BGP is used in the WAN network not just for the Internet but for VPNs (MPLS Layer 2 and Layer 3 VPNs) commonly as well

- When it is used in the WAN, stability and multi-protocol capability is very important along with the Traffic engineering and reliability

- In the Datacenter, instead of Stability and multi-protocol capability , Fast convergence , scalability , policy control are the key parameters expected from BGP

https://t.me/learningnets

# BGP PIC – Prefix Independent Convergence

- BGP PIC is used to make the re-convergence after topology change independent of the number of BGP prefixes

- BGP PIC is an evolution of Forwarding Plane to have BGP Fast Convergence

https://t.me/learningnets

# Three Different FIB Architecture

- BGP PIC has two modes , BGP PIC Core and BGP PIC Edge

- BGP PIC Core provides fast convergence in case of Core Link or Node Failure, IGP converges is important for BGP PIC Core

- BGP PIC Edge provides fast convergence in case of Edge Link or Node Failure

https://t.me/learningnets

# Three Different FIB Architecture

- For us to understand both BGP PIC Core and Edge and how they provide Fast Convergence for the failure scenarios, let's have a look at different FIB architectures in the routers

- These are Flat FIB , Hierarchical FIB and Generalized FIB

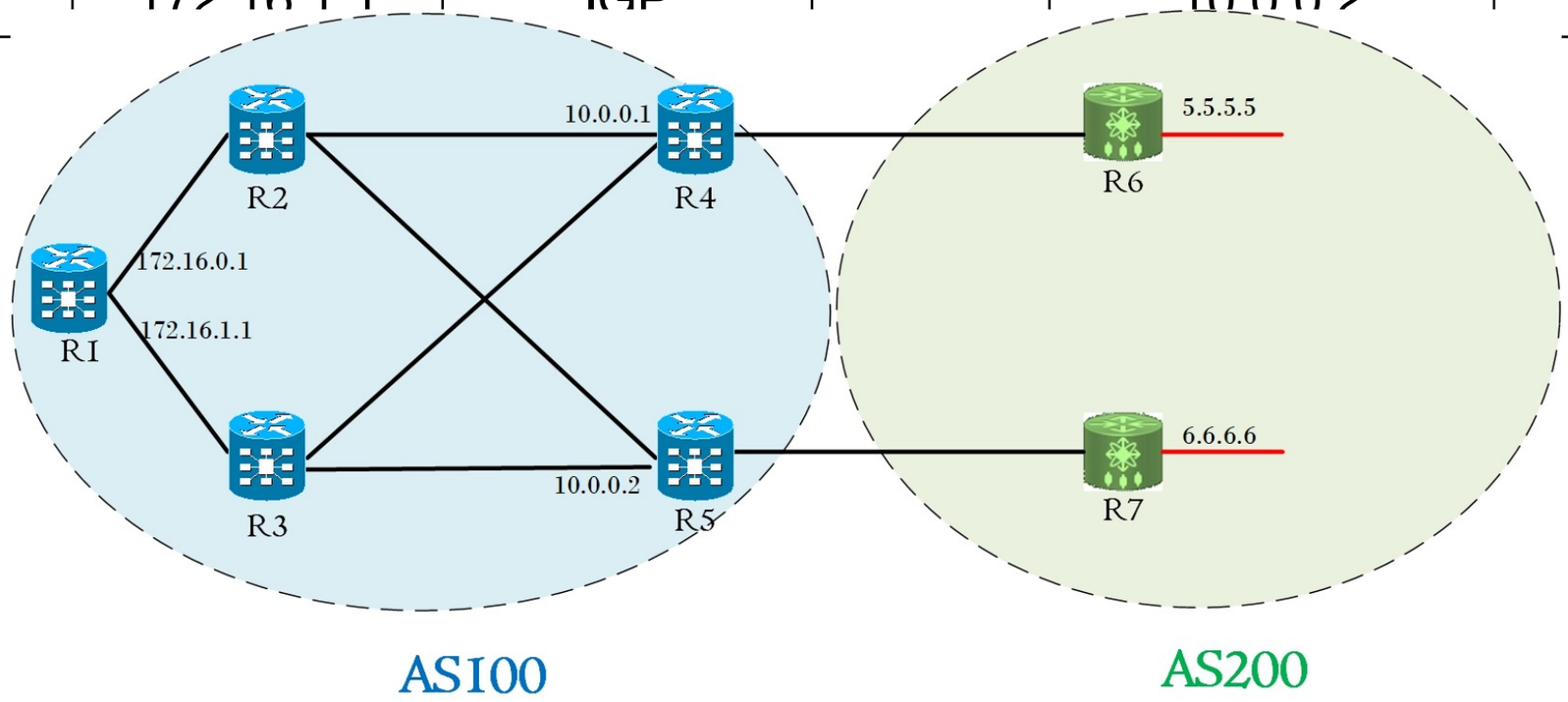
https://t.me/learningnets

R1's RIB

| Prefix | Next Hop or Outgoing Interface | Protocol |
|---|---|---|
| 5.5.5.5 | 10.0.0.1 | BGP |
| 6.6.6.6 | 10.0.0.2 | BGP |
| 10.0.0.1 | 172.16.0.1 | IGP |
| 10.0.0.1 | 172.16.1.1 | IGP |
| 10.0.0.2 | 172.16.0.1 | IGP |
| 10.0.0.2 | 172.16.1.1 | IGP |

R1's FIB

| Prefix | Next Hop or Outgoing Interface |
|---|---|
| 5.5.5.5 | 172.16.0.1 |
| 6.6.6.6 | 172.16.1.1 |
| 10.0.0.1 | 172.16.0.1 |
| 10.0.0.1 | 172.16.1.1 |
| 10.0.0.2 | 172.16.0.1 |
| 10.0.0.2 | 172.16.1.1 |



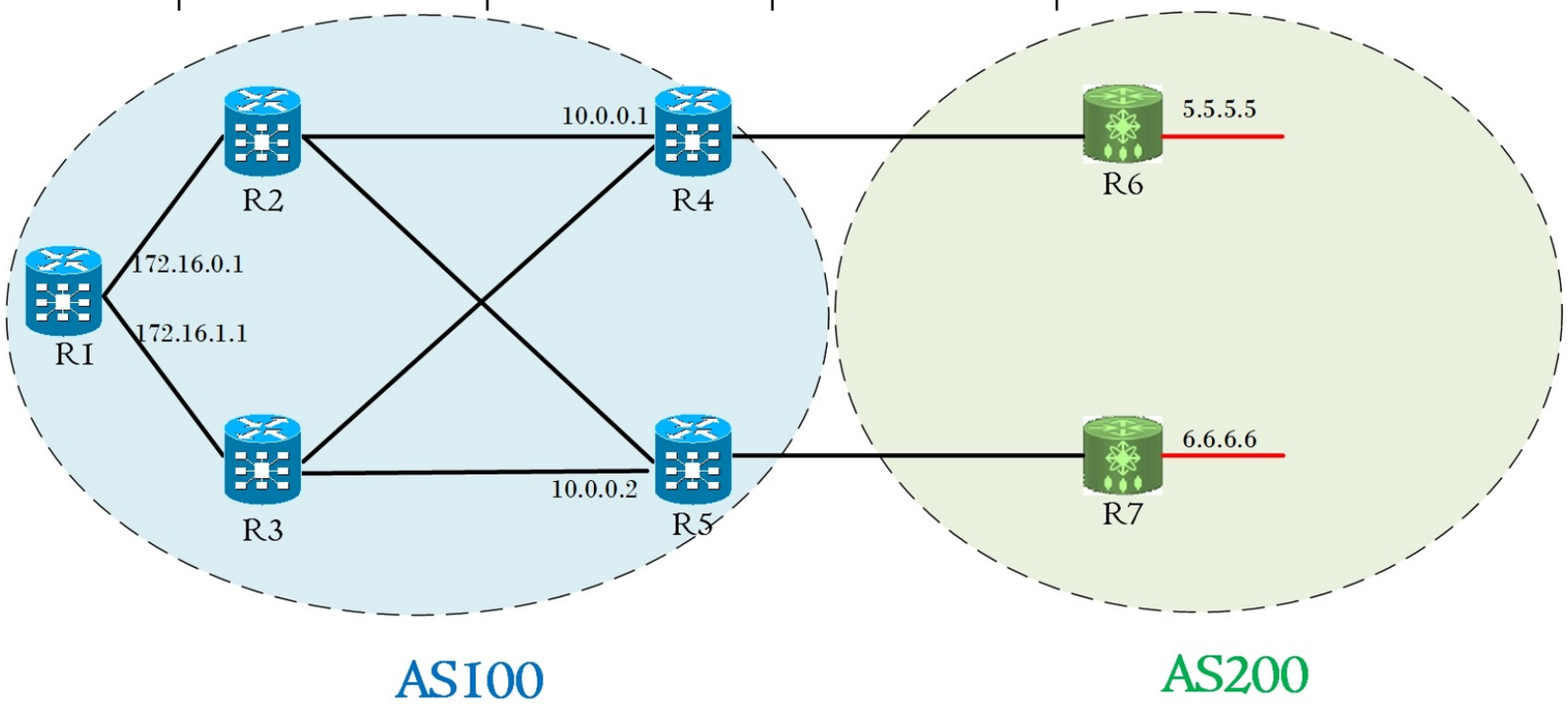

https://t.me/learningnets

R1's RIB

| Prefix | Next Hop or Outgoing Interface | Protocol |
|---|---|---|
| 5.5.5.5 | 10.0.0.1 | BGP |
| 6.6.6.6 | 10.0.0.2 | BGP |
| 10.0.0.1 | 172.16.0.1 | IGP |
| 10.0.0.1 | 172.16.1.1 | IGP |

R1's FIB

| Prefix | Next Hop or Outgoing Interface |
|---|---|
| 5.5.5.5 | 10.0.0.1 |
| 6.6.6.6 | 10.0.0.2 |
| 10.0.0.1 | 172.16.0.1 |
| 10.0.0.1 | 172.16.1.1 |
| 10.0.0.2 | 172.16.0.1 |
| 10.0.0.2 | 172.16.1.1 |

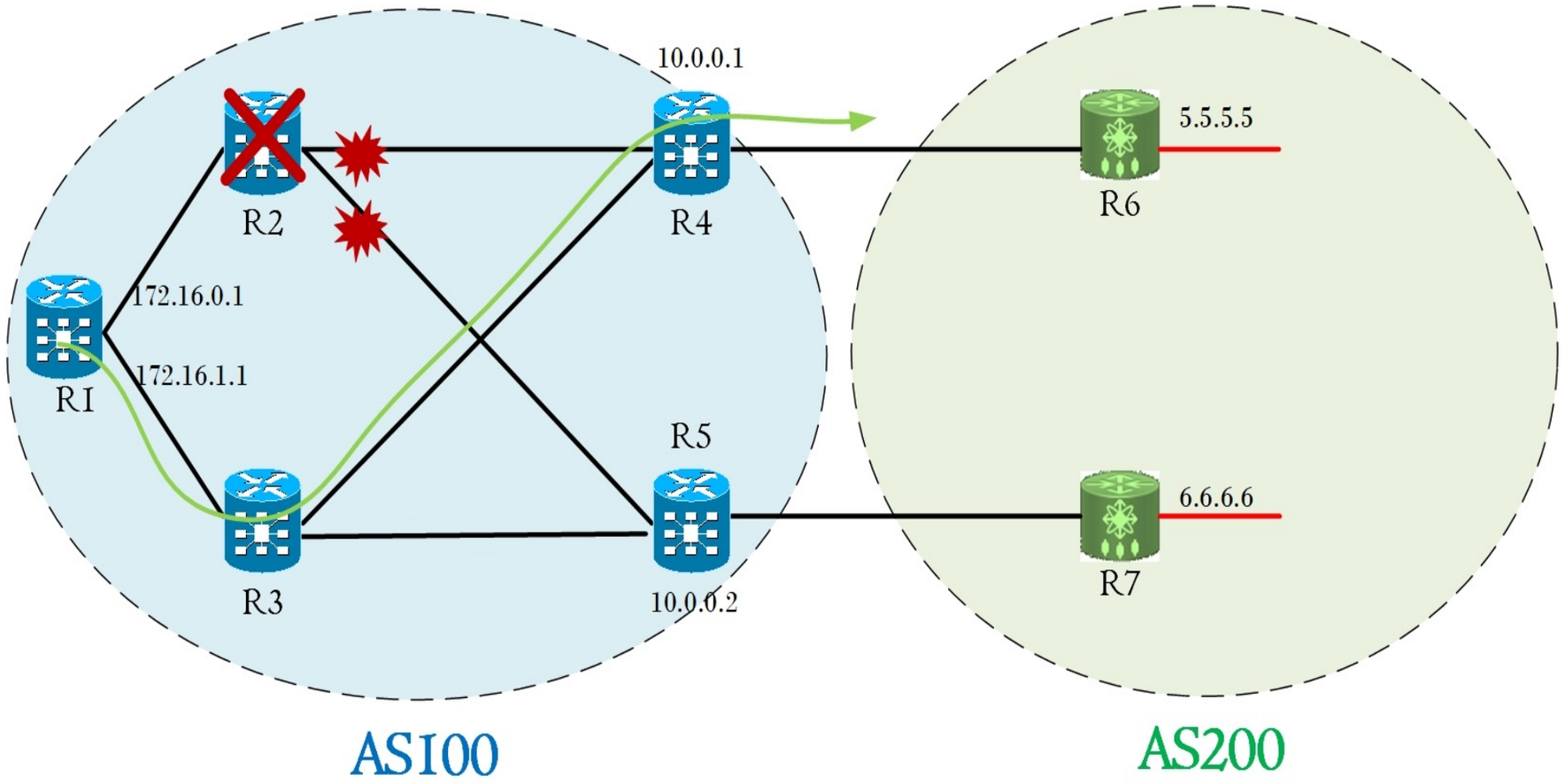



With Hierarchical FIB
R1's FIB has recursive route for BGP Next hop
Thus both RIB and FIB has Recursion

https://t.me/learningnets

# Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE

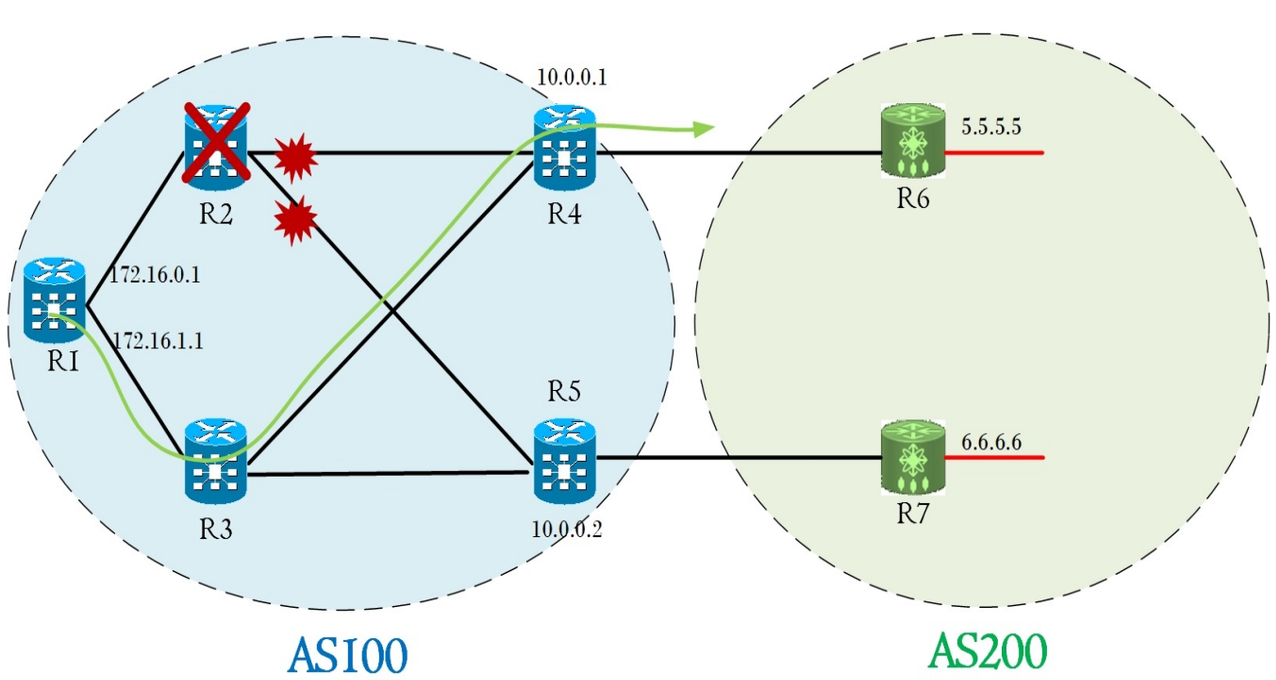
- Hierarchical FIB helps for the BGP PIC Core, when there is a Core Link or Device Failure, BGP next hop doesn't change, FIB just changes the IGP next hop if there is one available




https://t.me/learningnets

# Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE

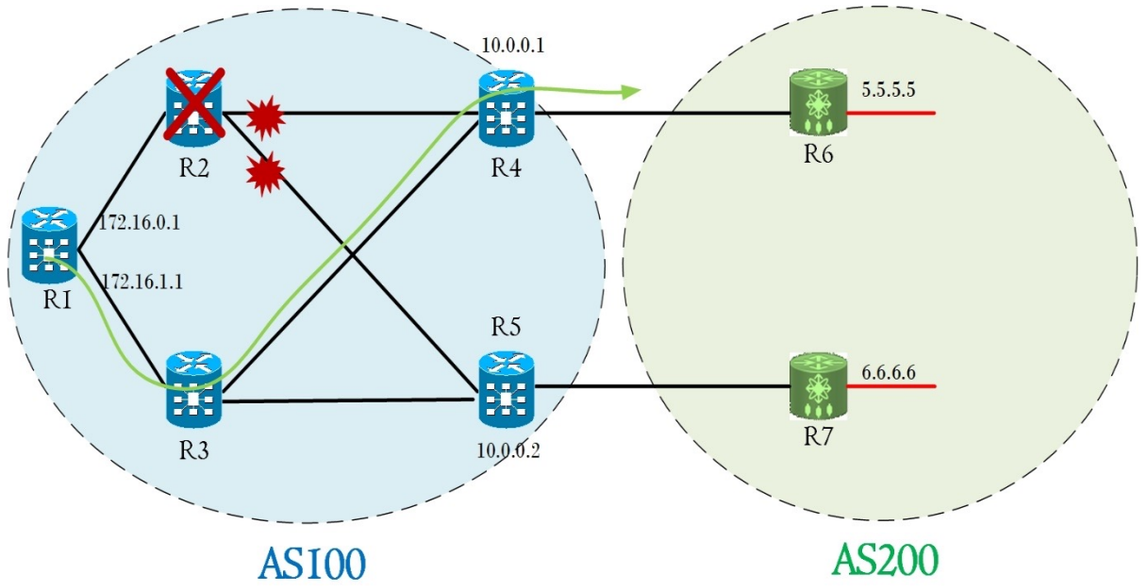
5.5.5.5 and 6.6.6.6 as a BGP Next Hop stays up, FIB just changes the IGP next hop which is used to reach BGP Next Hop




| Prefix | Next Hop or Outgoing Interface |
|---|---|
| 5.5.5.5 | 10.0.0.1 |
| 6.6.6.6 | 10.0.0.2 |
| ~~10.0.0.1~~ | ~~172.16.0.1~~ |
| 10.0.0.1 | 172.16.1.1 |
| 10.0.0.2 | 172.16.0.1 |
| 10.0.0.2 | 172.16.1.1 |

https://t.me/learningnets

# Three Different FIB Architecture – Hierarchical FIB Architecture/BGP PIC CORE




- Hierarchical FIB is only helpful for BGP PIC Core not for BGP PIC Edge

- For BGP PIC Core, R1 doesn't need to have two BGP next hop for external prefix

- For BGP PIC Edge , two BGP Next hop for external prefix is needed

https://t.me/learningnets

# Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

- Generalized FIB Architecture relies on a concept, sharing BGP Path List for the external prefixes and IGP Path List for the BGP Next Hops

- External BGP Prefixes will share the multiple BGP Next hops and BGP next hop will

https://t.me/learningnets

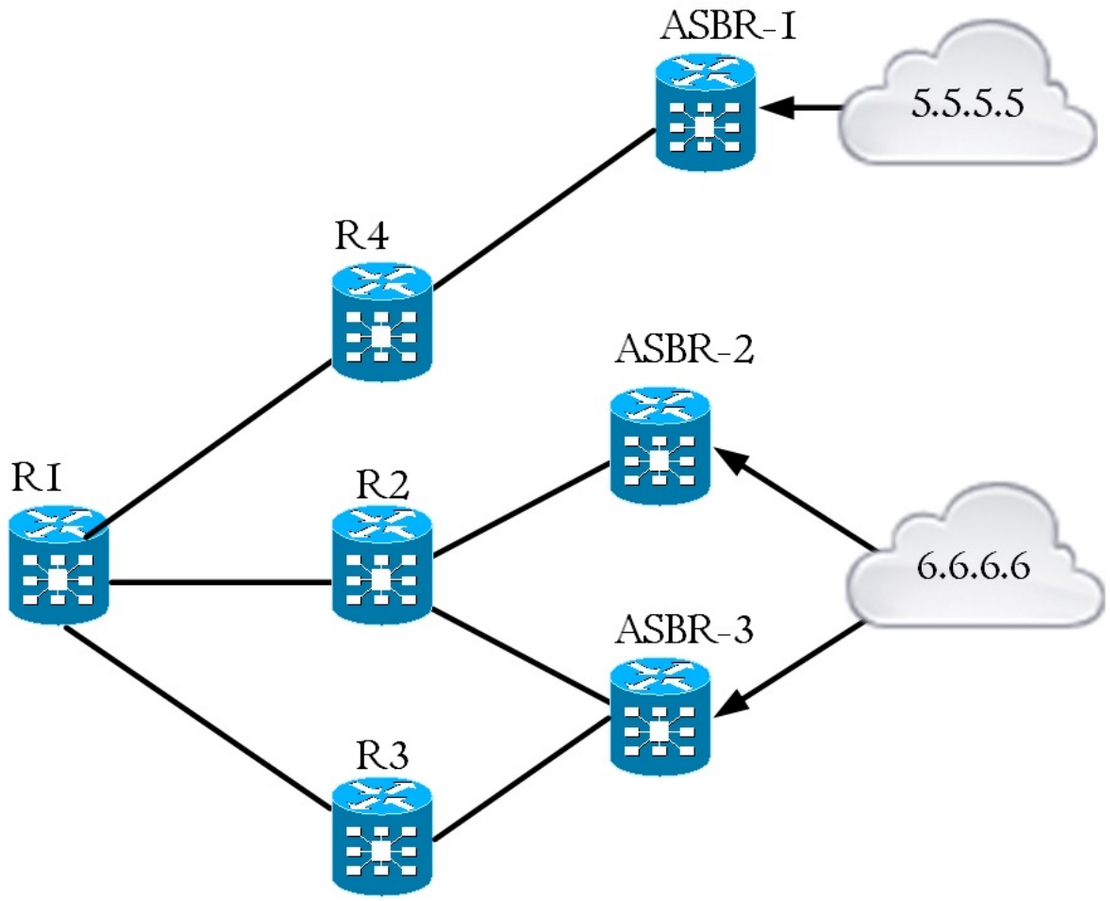
# Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

- Generalized FIB is an evolution of Hierarchical FIB and helps for the Core and Edge link or node failure situations

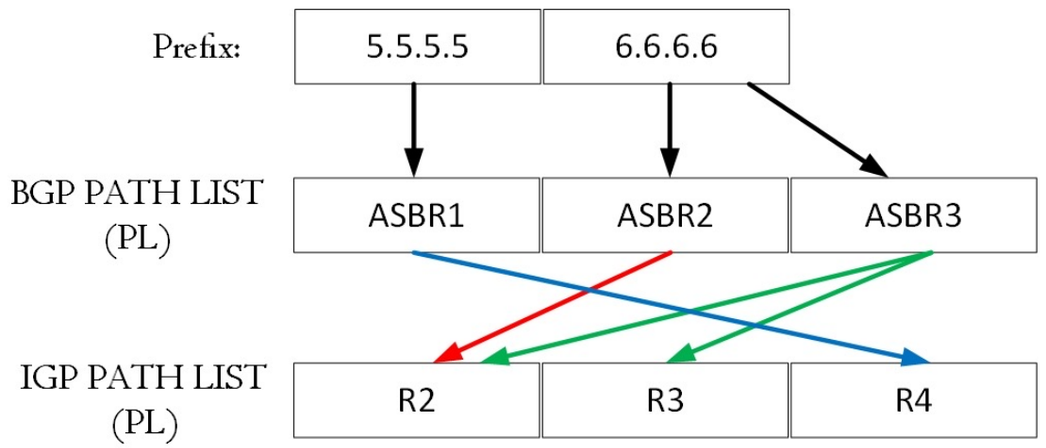
- Many vendors provide Generalized FIB hardware

https://t.me/learningnets

# Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

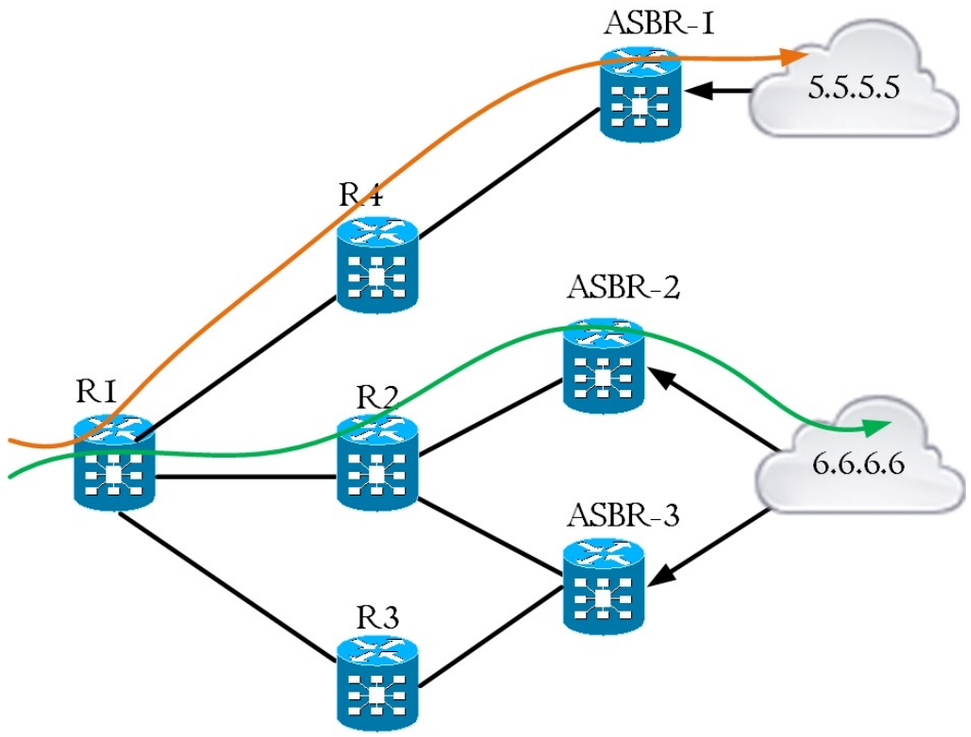



Let's assume R1 selects ASBR 2 as the next hop to reach 6.6.6.6
Thus in the IGP BGP PL, ASBR 2 will be Primary path and ASBR 3 will be secondary path

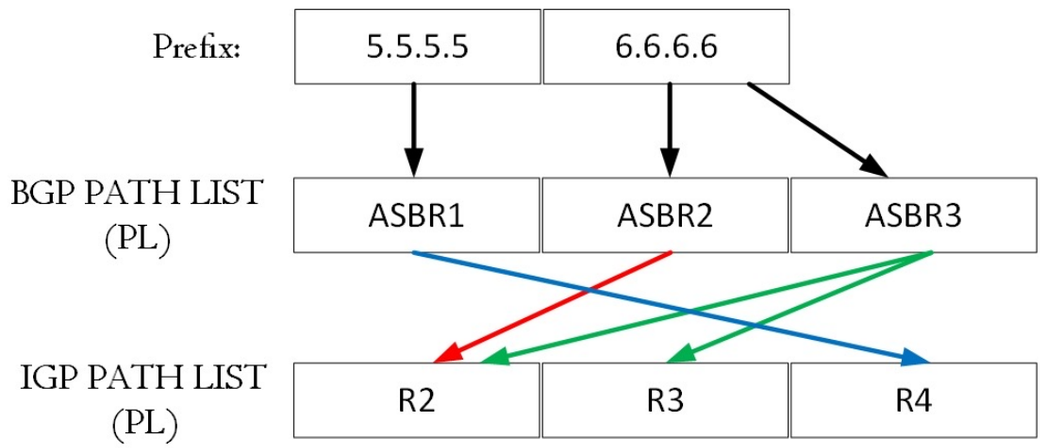



https://t.me/learningnets

# Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE



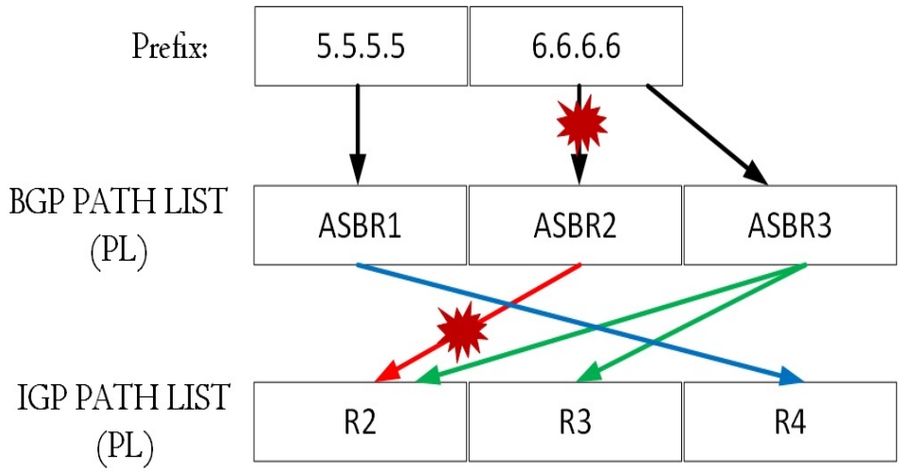

Let's assume R1 selects ASBR 2 as the next hop to reach 6.6.6.6
Thus in the IGP BGP PL, ASBR 2 will be Primary path and ASBR 3 will be secondary path

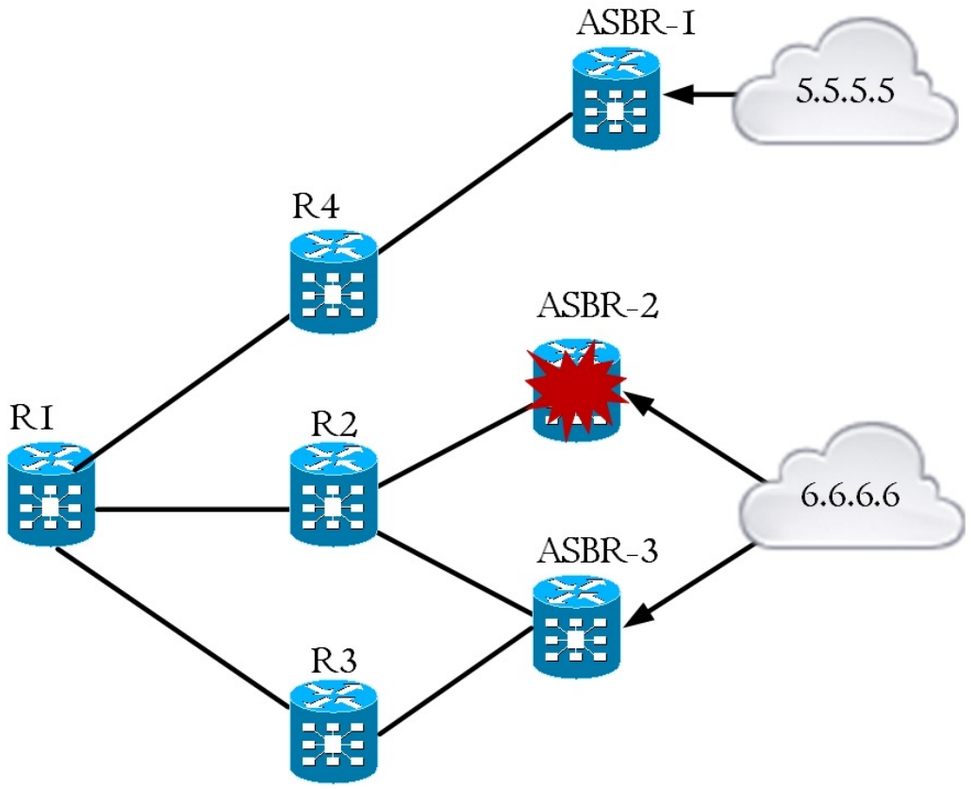


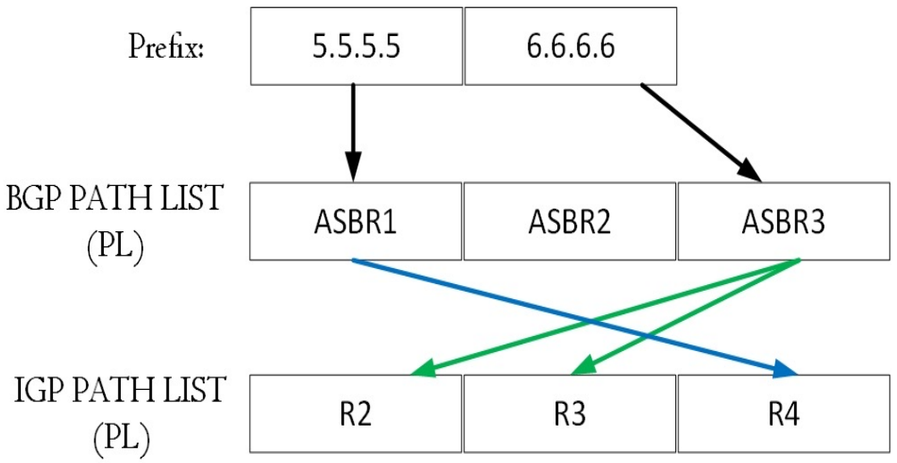

https://t.me/learningnets

# Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

If ASBR 2 fails, IGP detects the failure and R2 is removed from IGP Path List and FIB Backwalking process will start removing any BGP Next hop from the BGP Path List if there is no available next hop in the IGP Path List pointing for it



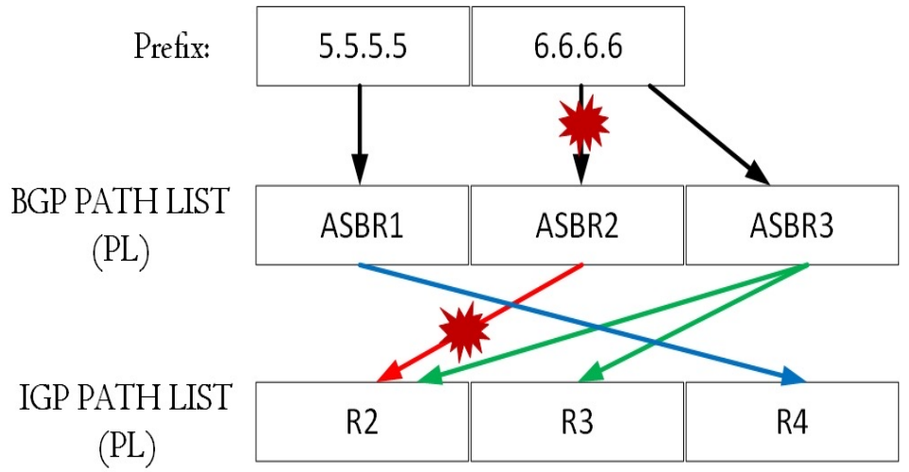



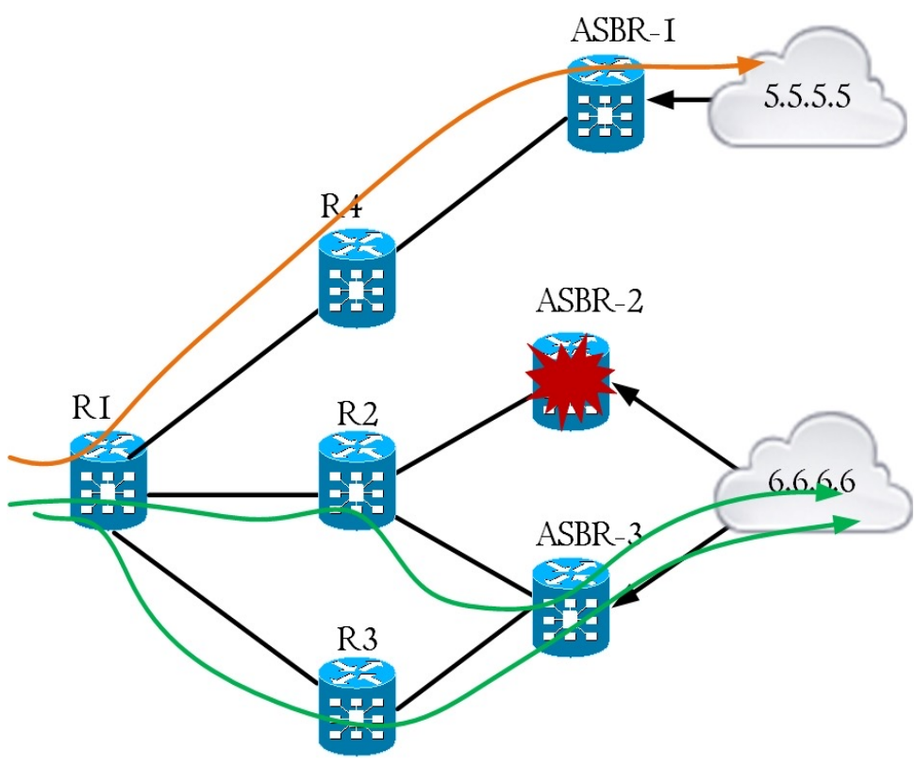



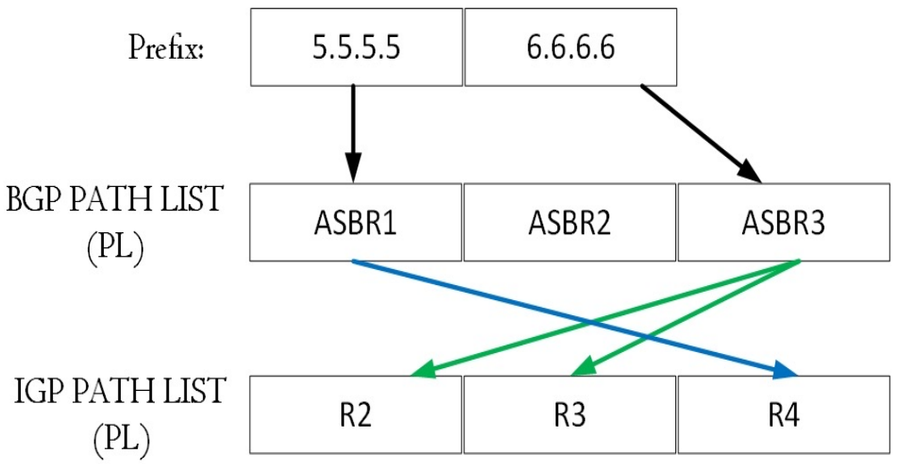

https://t.me/learningnets

# Three Different FIB Architecture – Generalized FIB Architecture/BGP PIC CORE and EDGE

ASRB3 becomes best path immediately after ASBR 2 is removed from the BGP Path List.
Backwalking process to find whether there is any BGP next hop for the external prefixes in the BGP Path List takes only couple milliseconds




https://t.me/learningnets

# BGP PIC Core and Edge Summary

- BGP PIC Core and Edge provides fast convergence incase of Core and Edge link or node failure situations

- Flat FIB points direct Layer 2 adjacency not to the BGP next hop or BGP Path List for the external prefixes, thus incase of failure, all the prefixes need to be refreshed with the new next hop which can take so much time

- Hierarchical and Generalized FIB helps for the BGP PIC Core and Edge scenarios

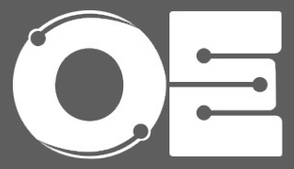
https://t.me/learningnets

# BGP PIC Core and Edge Summary

- If there is no alternate IGP next hop, BGP PIC Core is useless, if there is no alternate BGP next hop, BGP PIC Edge is useless as well

- BGP PIC provides fast dataplane convergence , control plane convergence will follow after that

https://t.me/learningnets



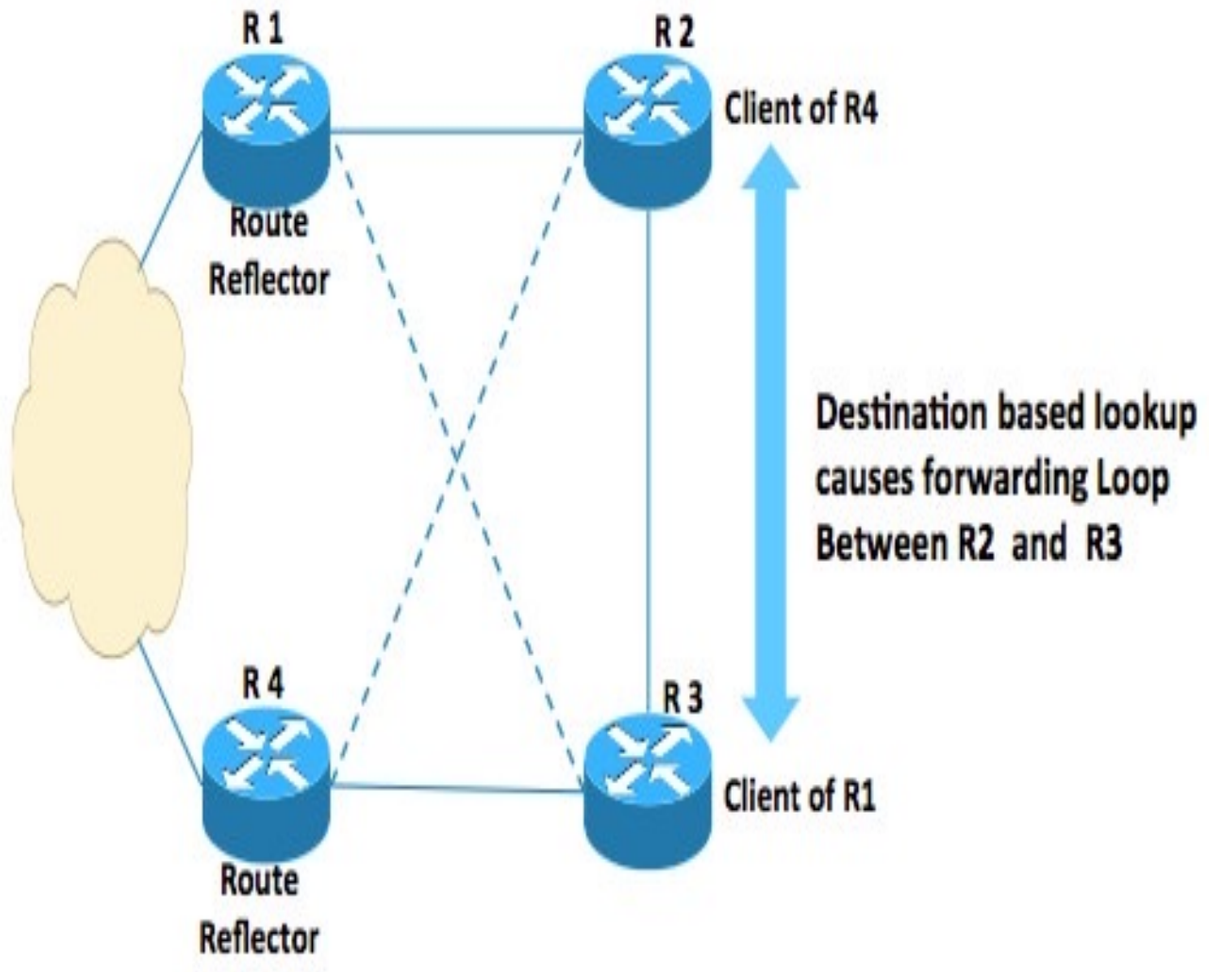
# BGP Case Studies

## Border Gateway Protocol

https://t.me/learningnets

# BGP Routing Loop with IP Route Reflector

In the diagram below; R2 is route reflector client of R4, R3 is route reflector client of R1.




1. MPLS or any tunneling mechanism is not enabled. What is the problem with this design ?

2. Would you have the problem if MPLS is enabled ?

https://t.me/learningnets

- 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

https://t.me/learningnets

- 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

https://t.me/learningnets

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

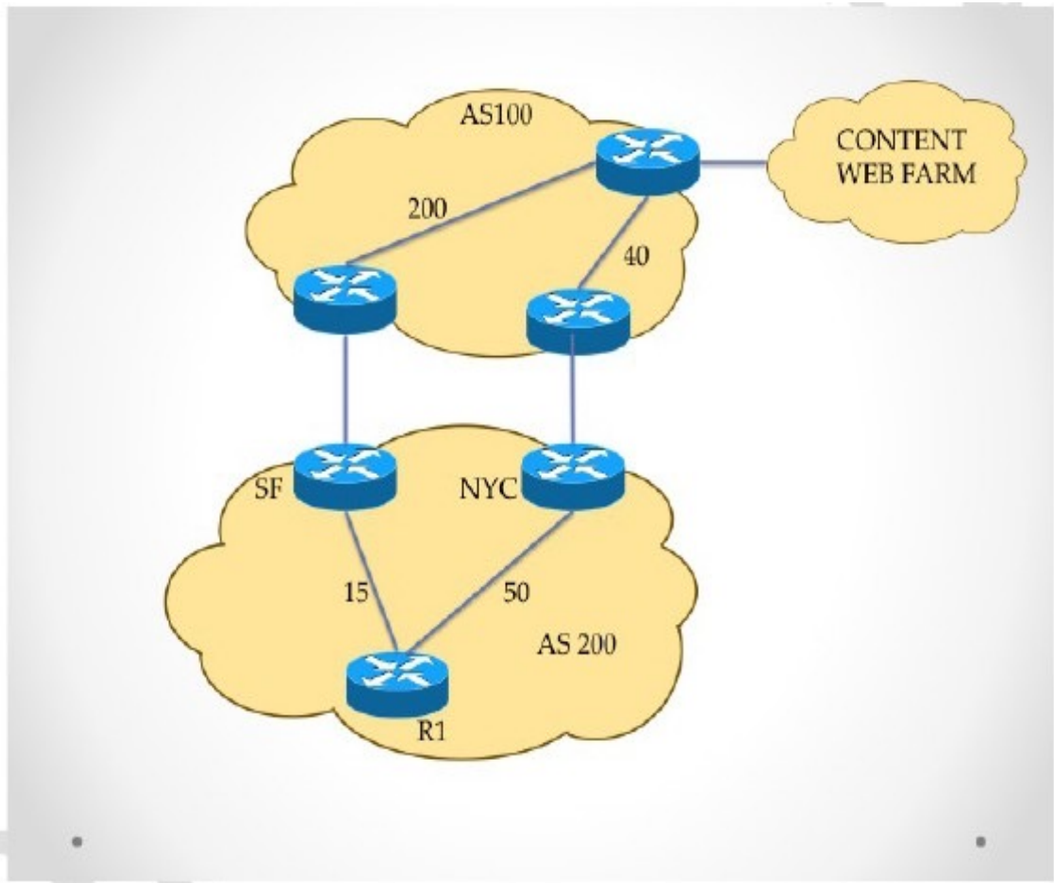
If MPLS would be enabled, we wouldn’t have the same behavior since when R2 do the destination lookup for the prefix A, it will find the next hop R4 but in order to reach to R4, it would push the transport label

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

https://t.me/learningnets

# BGP Hot Potato Routing

- AS200 is a customer service provider of AS100 transit Service provider. Customers of AS200 is trying to reach a web page located behind AS100. AS200 is not implementing any special BGP policy. What would be the ingress and egress traffic for AS 200 ?



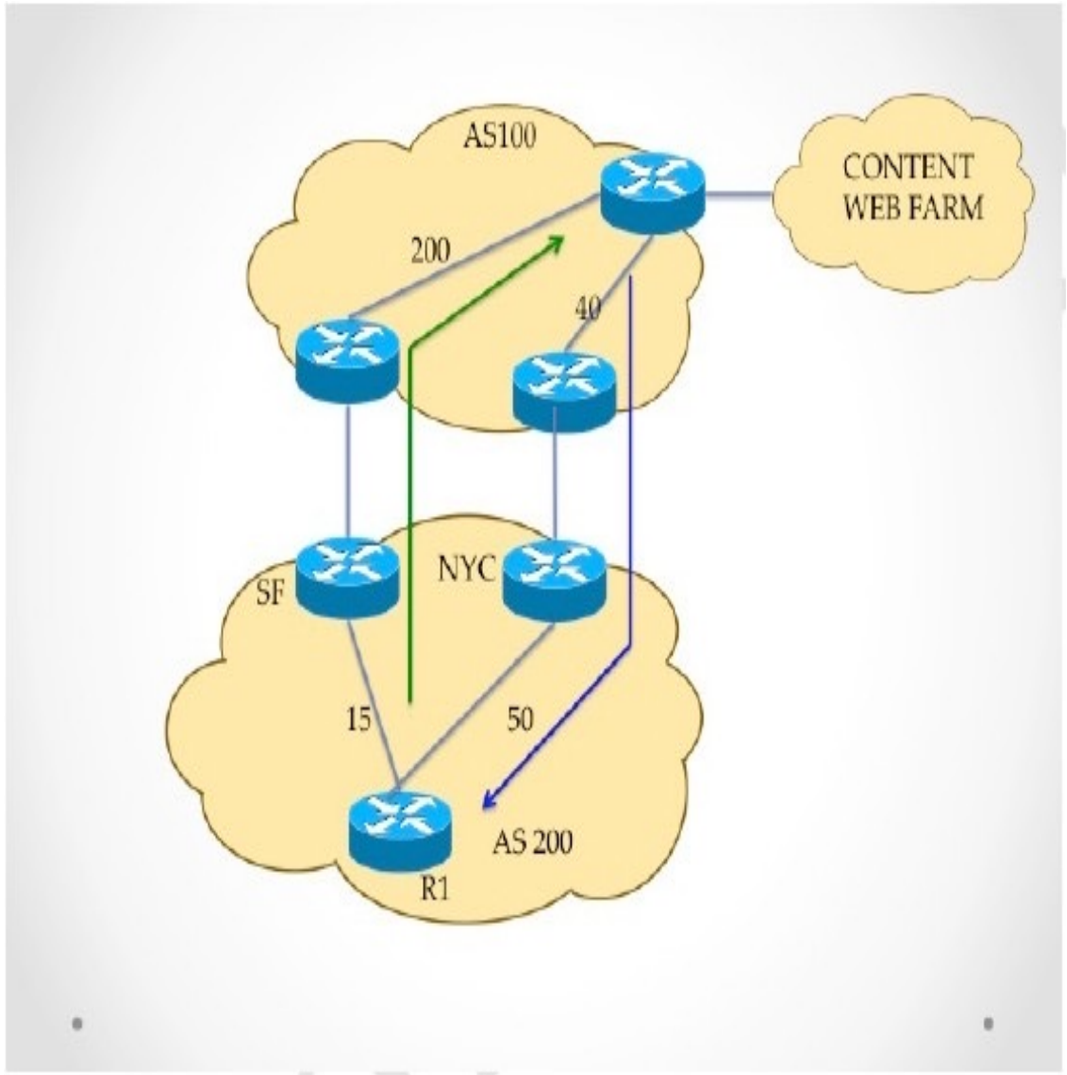

https://t.me/learningnets

- Topology picture depicts the AS 100 and AS 200 connections. They have a BGP peer ( Customer- Transit ) relationship on two locations. San Francisco and New York

- IGP distances are shown in the diagram. Since there is no any special BGP policy ( Local pref, MED, AS-Path is the same , Origin and so on ) , Hot Potato rule  will apply so egress path will be chosen from AS 200 and AS100 based on IGP distances

https://t.me/learningnets

- Egress traffic from AS 200 is the green arrow in the below diagram, since SF path is shorter IGP distance. Ingress traffic to AS200 from AS 100 is the blue arrow, since NYC connection from AS100 shorter IGP distance (40 vs. 200)



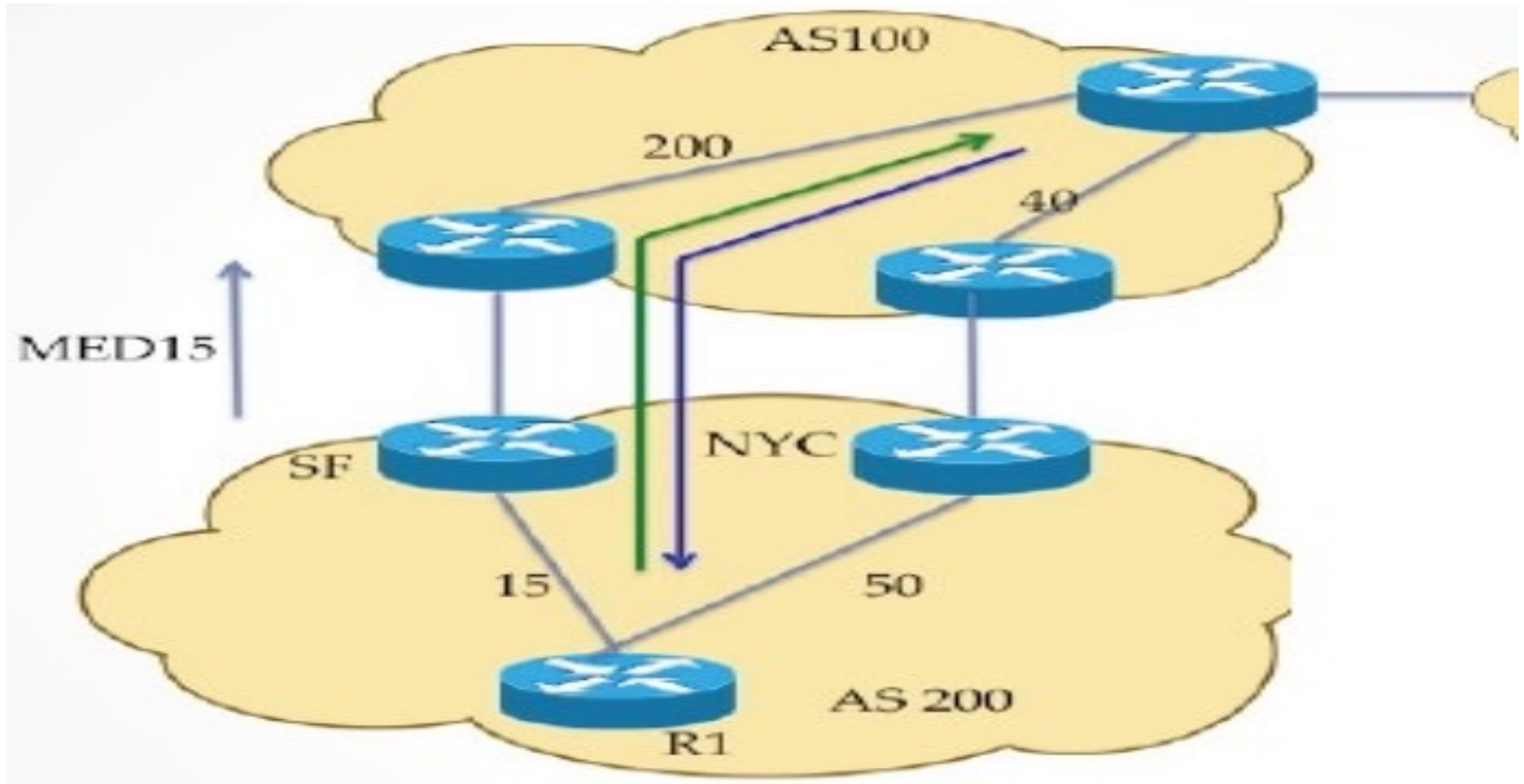

https://t.me/learningnets

- 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.

https://t.me/learningnets

**That's why AS200 is sending its prefixes from SF with lower MED than NYC as depicted in the below diagram**

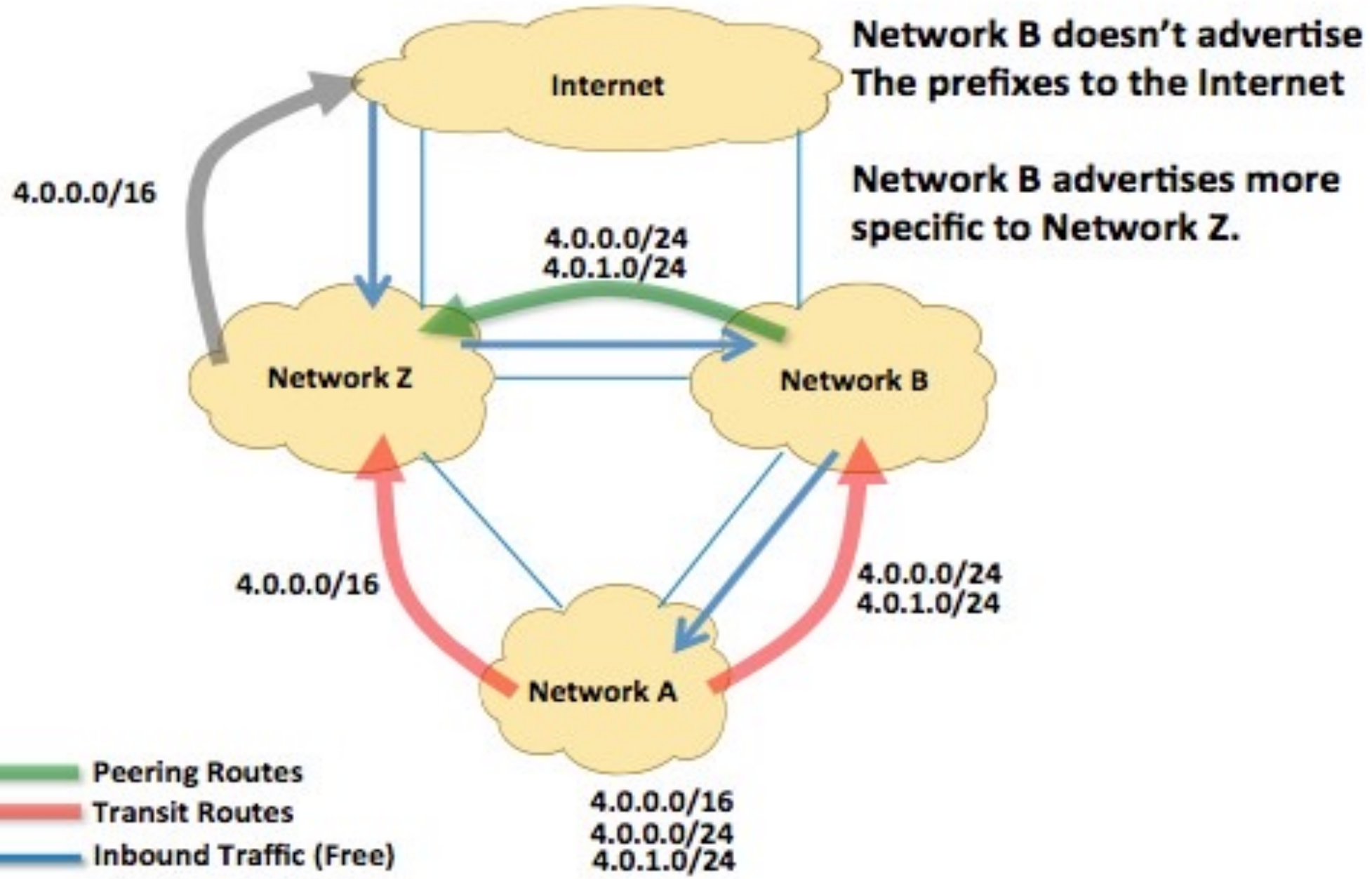



https://t.me/learningnets

# How to use Transit service for Free with BGP Jack Move

- Network A is a customer of Network Z, Network B is a peer of Network Z.
- Network A becomes transit customer of Network B.
- Network A announces 4.0.0.0/16 aggregate to Network Z and more specific prefixes,4.0.0.0/24 and 4.0.1.0/24 to Network B. Network B sends more specific to its peer Z.
- Network Z only announces the aggregate to the world. What is the impact of this design ?
- How can it be fixed ?

https://t.me/learningnets




https://t.me/learningnets

- 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

https://t.me/learningnets

- 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

https://t.me/learningnets

- 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

https://t.me/learningnets

# BGP Unintended Behaviors (aka BGP Wedgies)

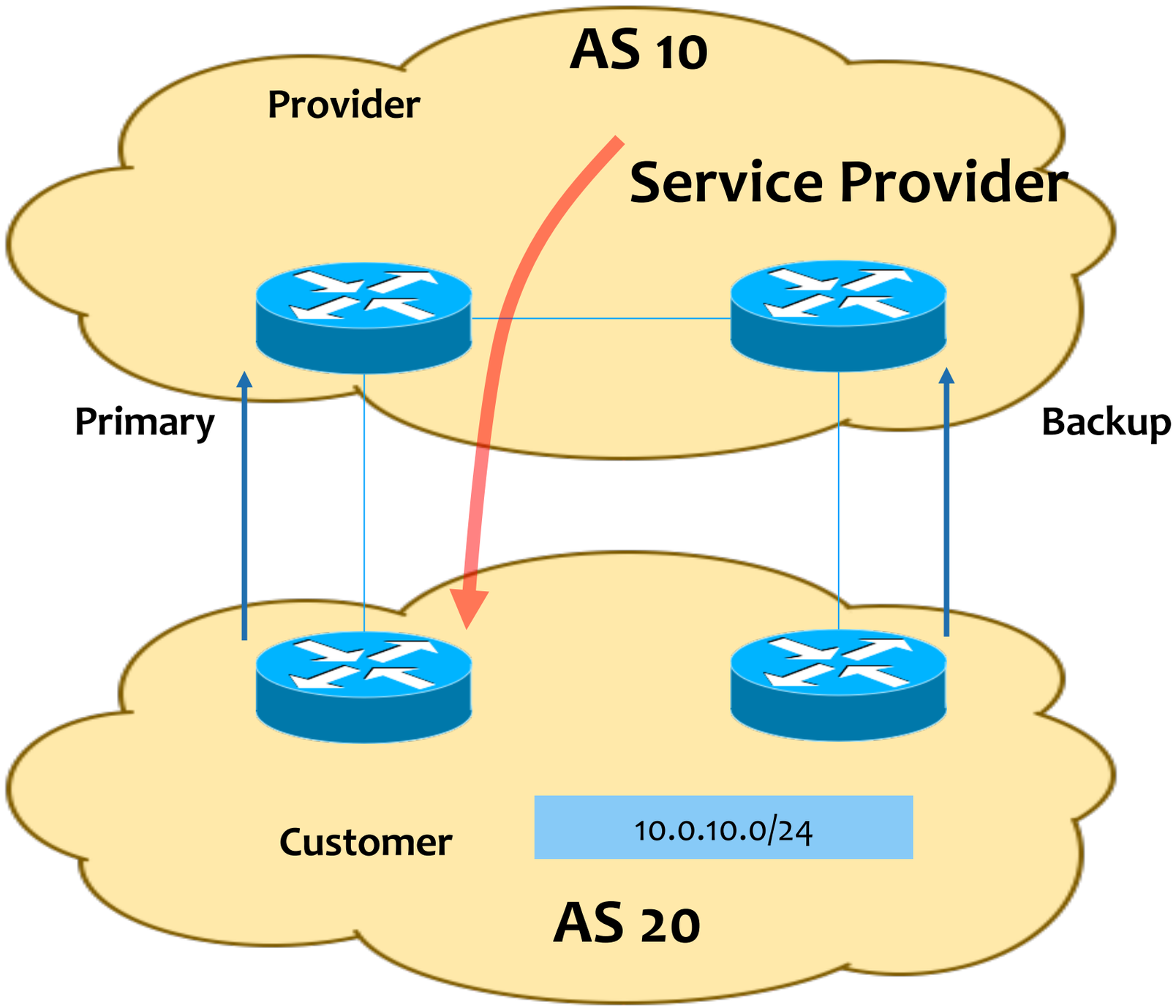
- Customer is running a BGP session with 1 service provider, they are considering to receive a transit service from the second Service Provider as well though

- Customer is using their own AS number which is AS20

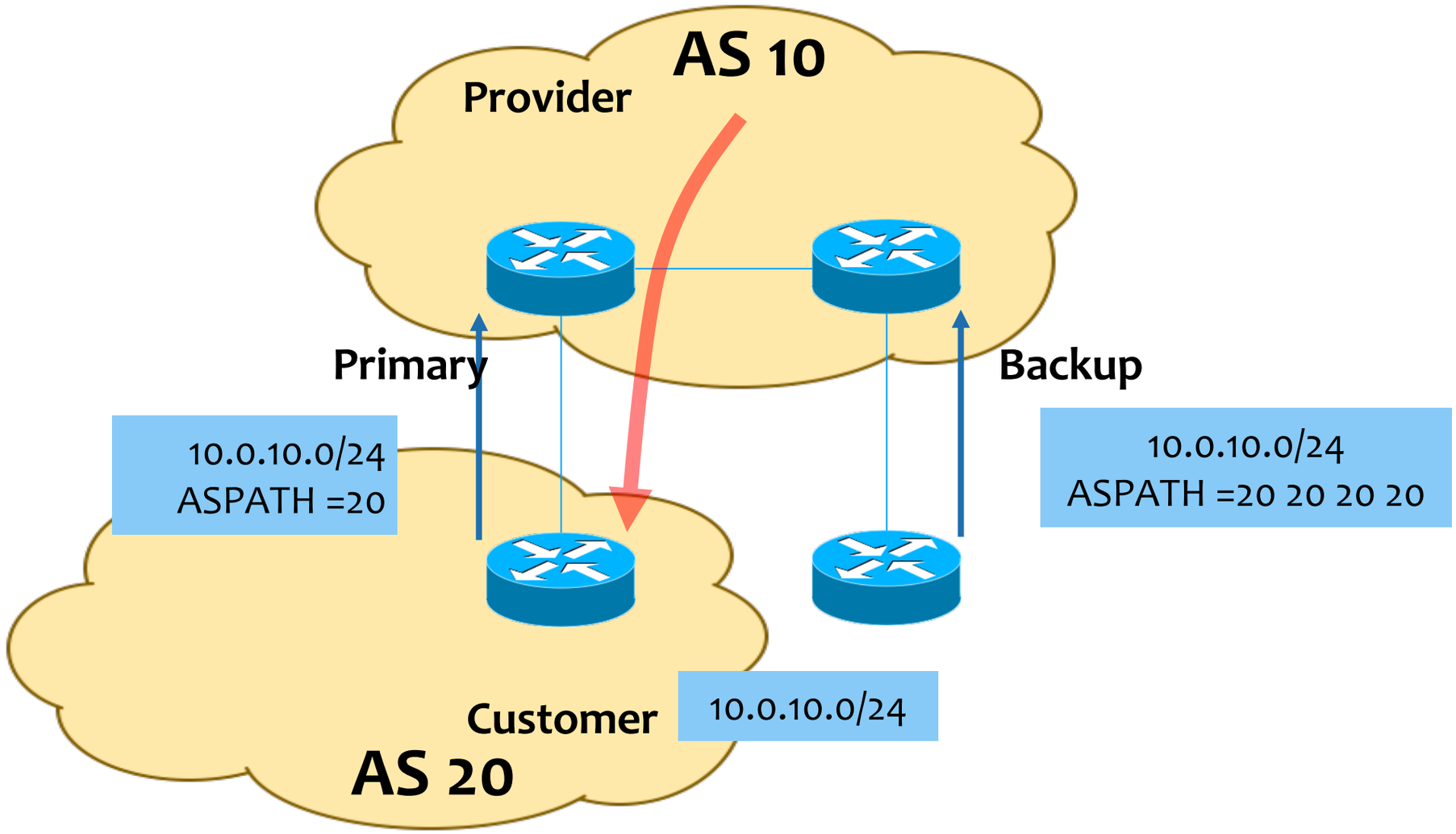
- They have 2 connections to their service provider and as it seems in the topology left path will be used as primary for their incoming traffic.

**Question 1: How can you achieve this ?**

https://t.me/learningnets




https://t.me/learningnets



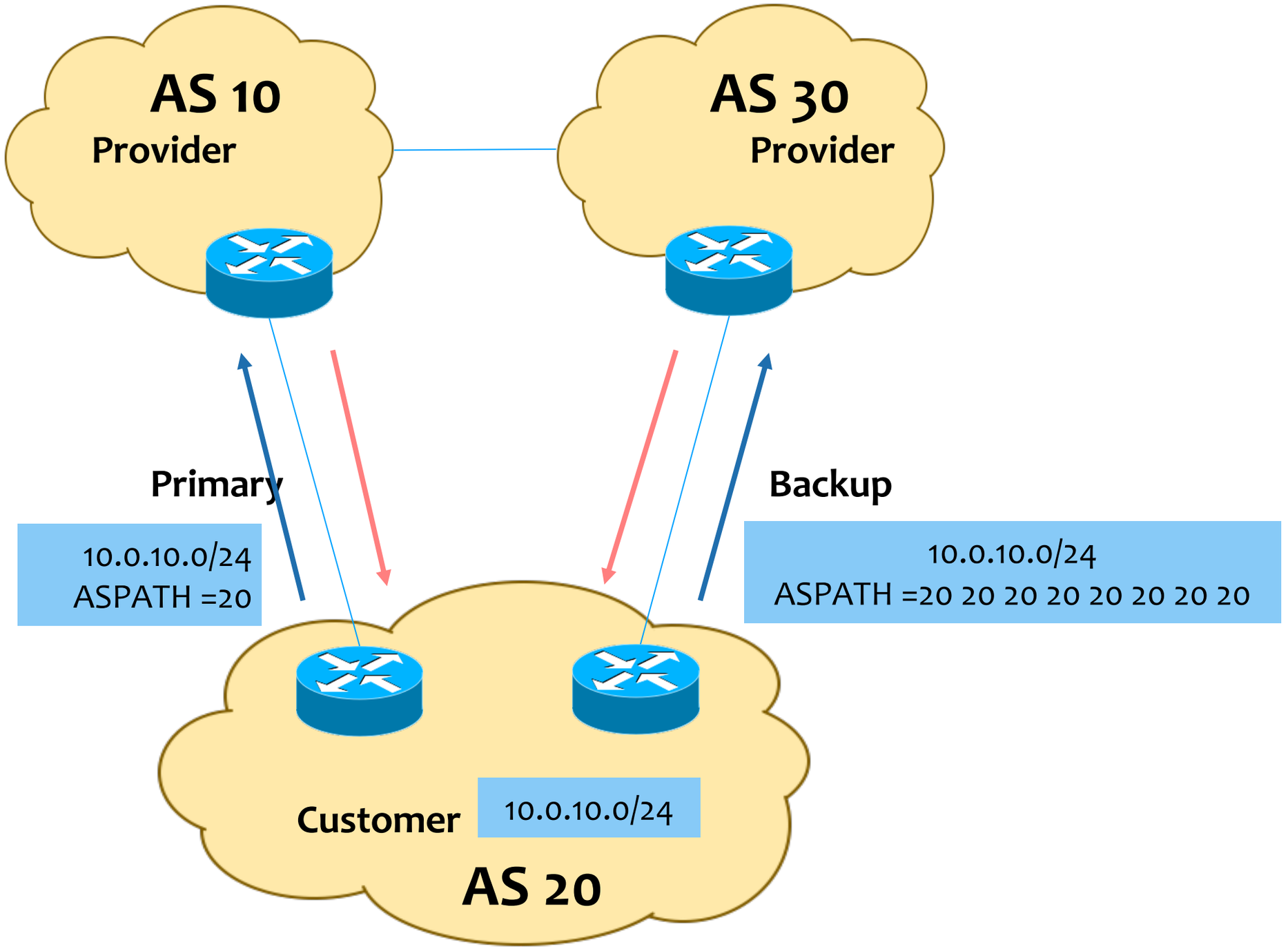

**Prepending will (usually) force inbound traffic from AS 10 to take primary link**

https://t.me/learningnets

- Customer purchased a new link from the second service provider which uses AS number 30 and decommissioned one of it’s link from the old service provider

- They want to use second service provider link as backup link. They learned from the early experience the as-path prepending trick

https://t.me/learningnets




https://t.me/learningnets

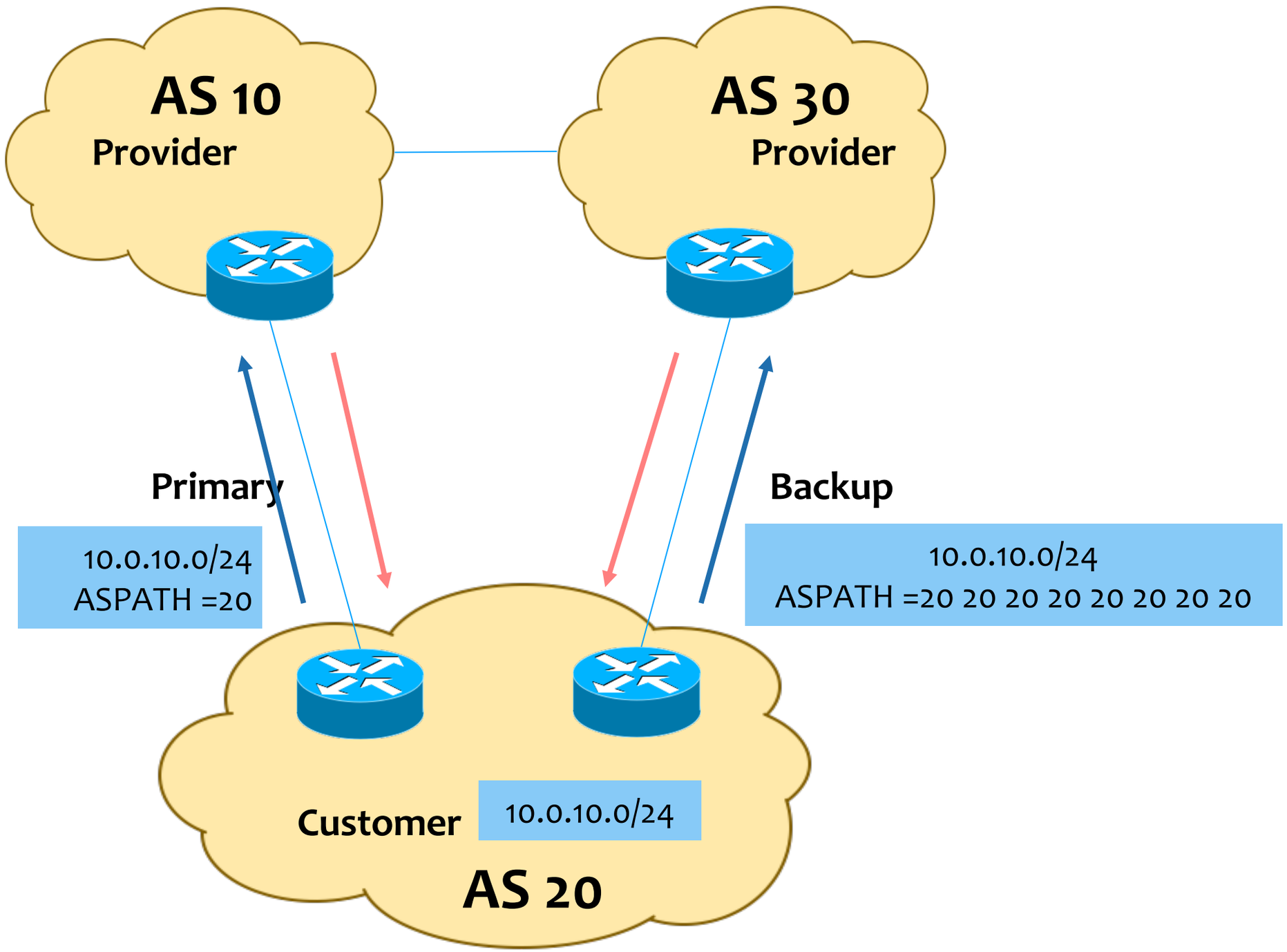
- Question 2: Is there a problem with this design ?

- Yes

- No

Question 3 : What is the problem, how you can solve it ?

https://t.me/learningnets



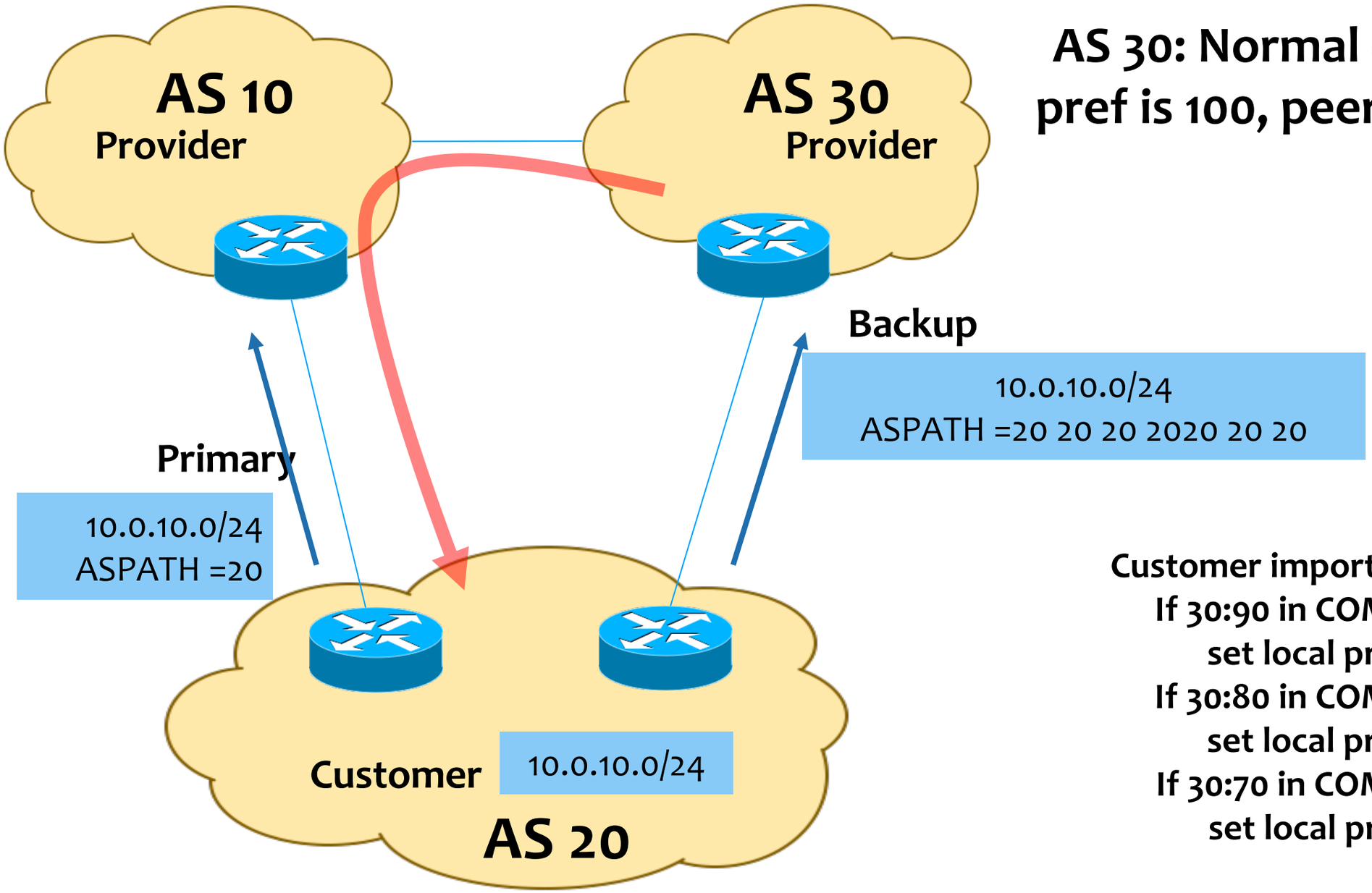

https://t.me/learningnets

- There is a problem with the design since customer wants to use second service provider as backup. AS-Path prepending in this way is often used as a form of  load Balancing

- BUT AS 30 will send traffic on “Backup” link, because it prefers customer routes due to higher local preference Service providers use on the customer link than the peer link and local preference is considered before ASPATH Length, so as-path prepending is not effected in this design

- Solution is to use communities

https://t.me/learningnets




**AS 30: Normal customer local pref is 100, peer local pref is 90**

**Customer import policy at AS 30:**
**If 30:90 in COMMUNITY then**
**set local preference to 90**
**If 30:80 in COMMUNITY then**
**set local preference to 80**
**If 30:70 in COMMUNITY then**
**set local preference to 70**

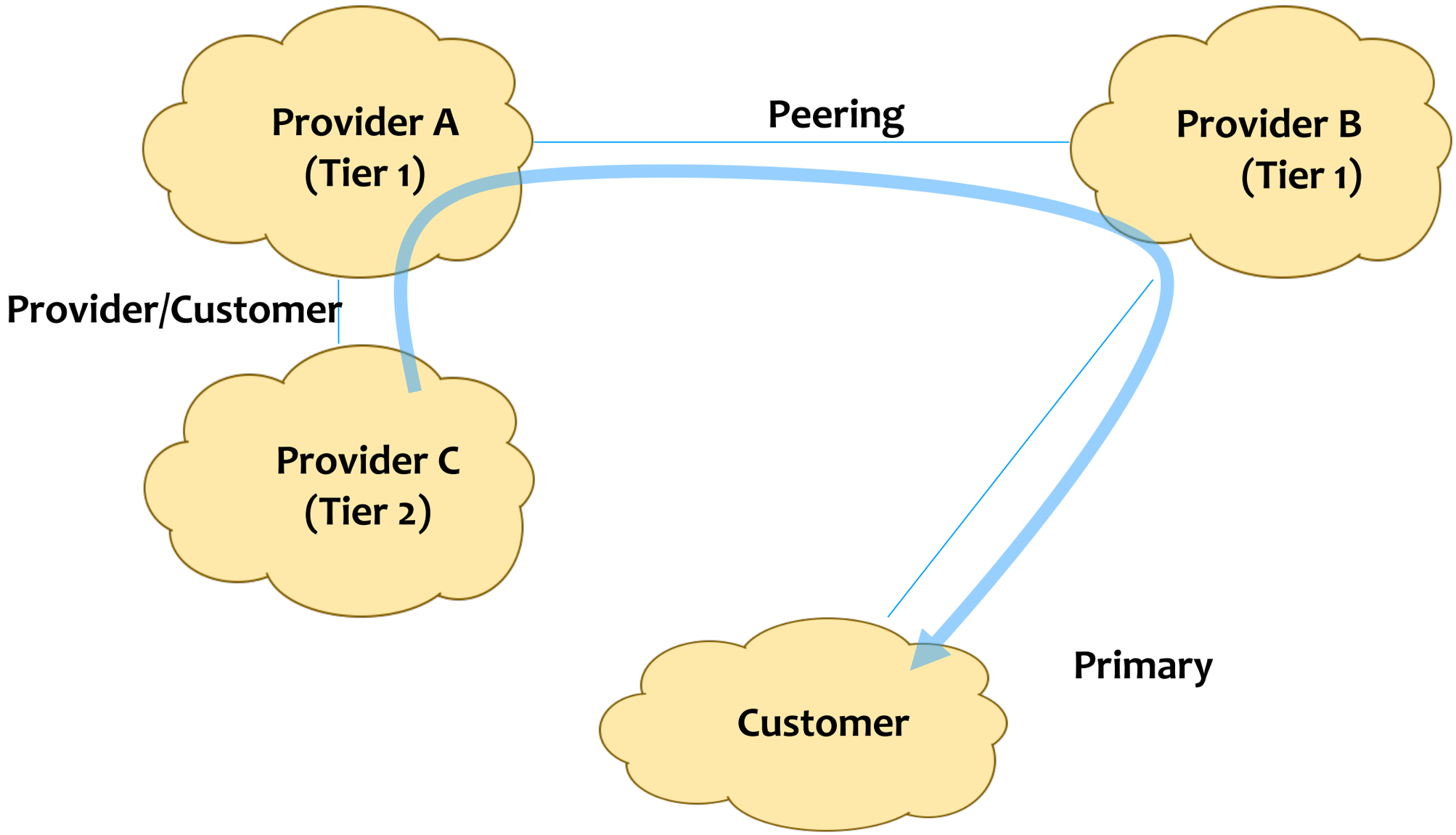
**GAO – REXFORD MODEL**

https://t.me/learningnets

- Question 5:  What if customer uses second service provider link as primary and the old provider secondary and the second provider peering connection as depicted in the below topology ?

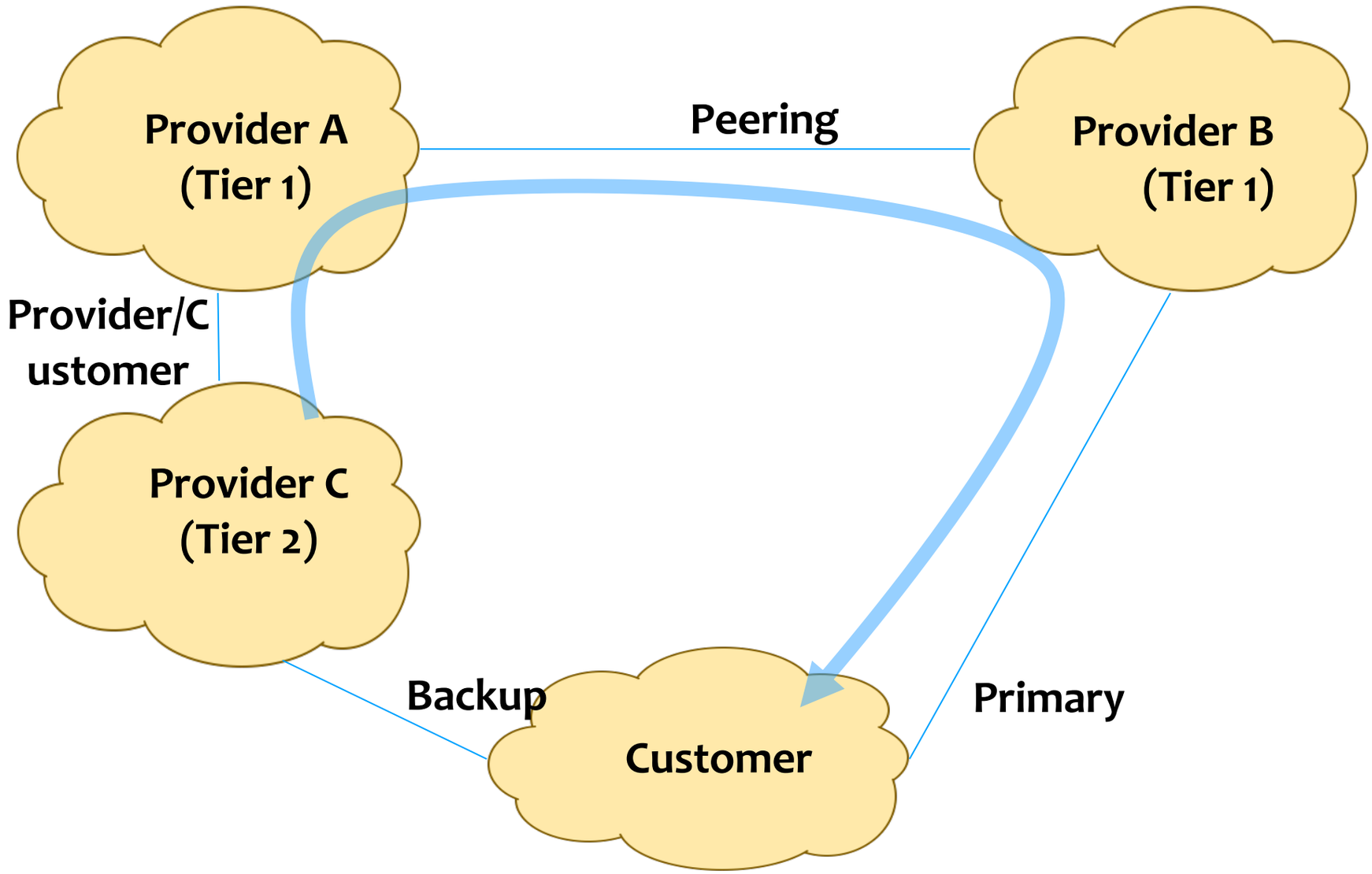
- Does community help ?

https://t.me/learningnets




https://t.me/learningnets

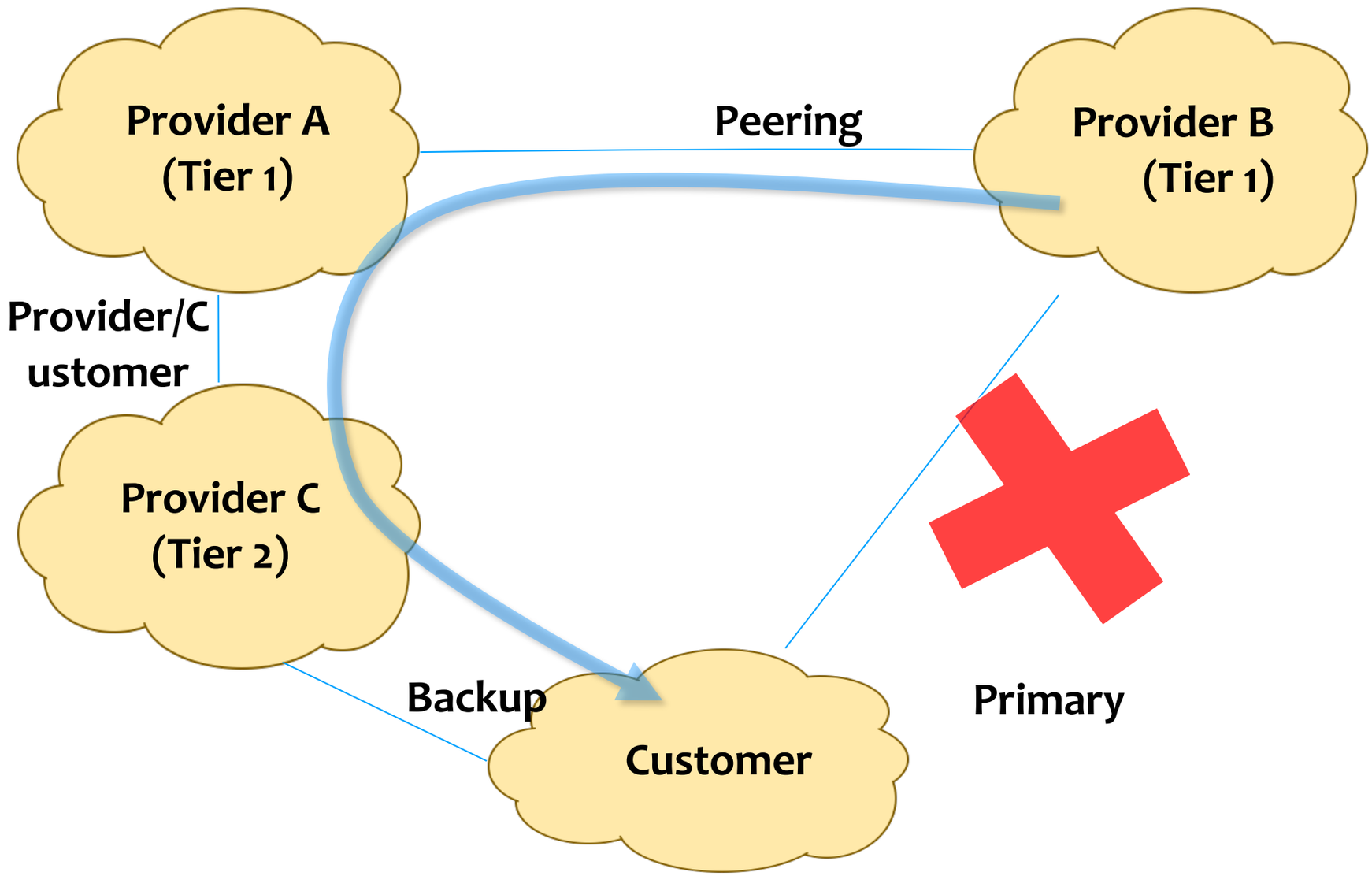
## Customer sends community to Provider C, in order to use Provider B as backup




https://t.me/learningnets

- Question 6: What happens if Primary link fails ?

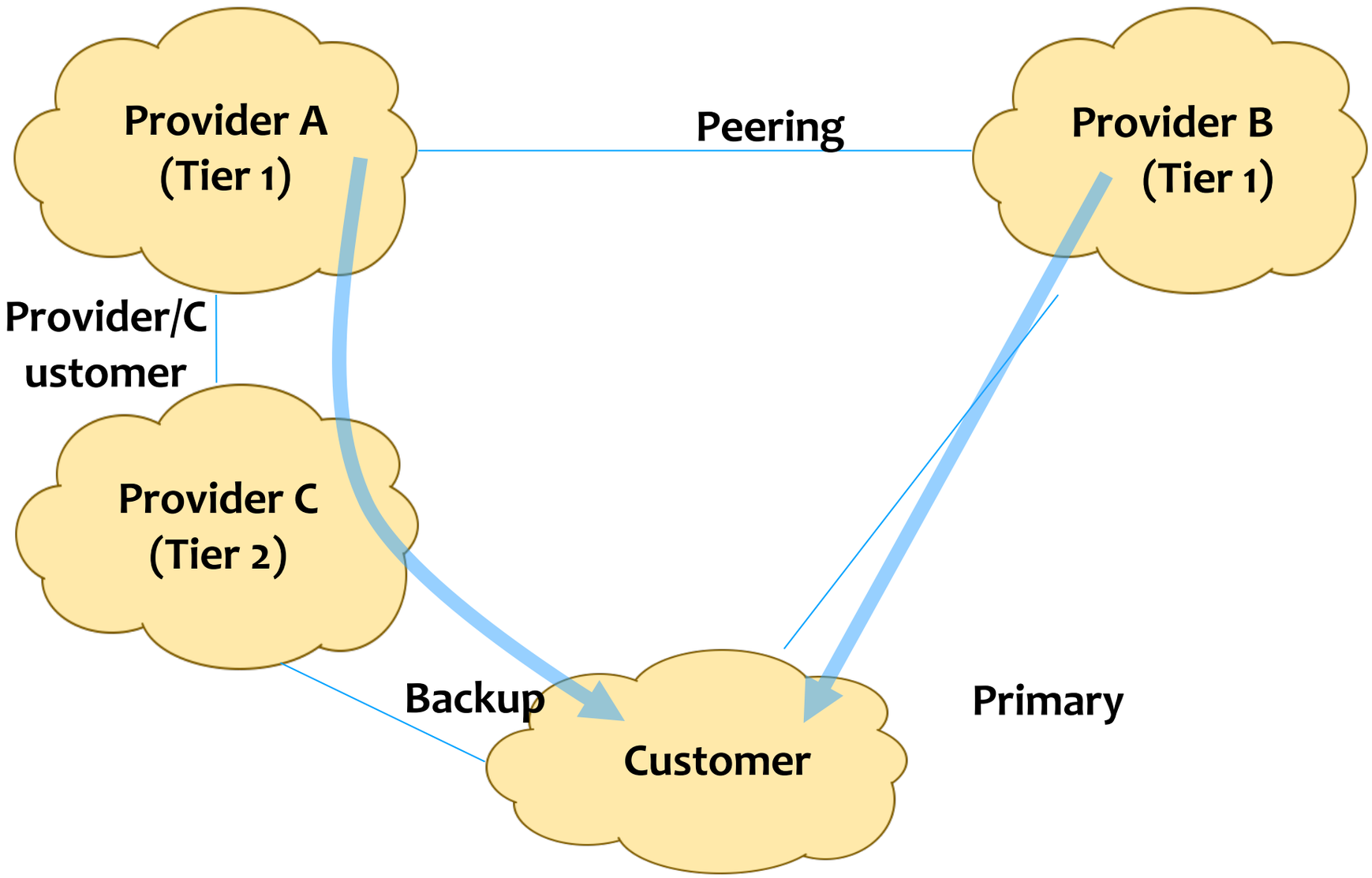
https://t.me/learningnets




https://t.me/learningnets

- Question 7: What happens when the primary link comes back ?

https://t.me/learningnets




https://t.me/learningnets

- When primary link comes back, both path is used for the incoming traffic anymore

- Because Provider A continue to choose to send Provider C since the community attribute is sent by Customer to Provider C, not to Provider A

- Solution to fix it, either Provider C will send a Provider A for its customer a community attribute, or Backup BGP link will be resetted when primary link comes back

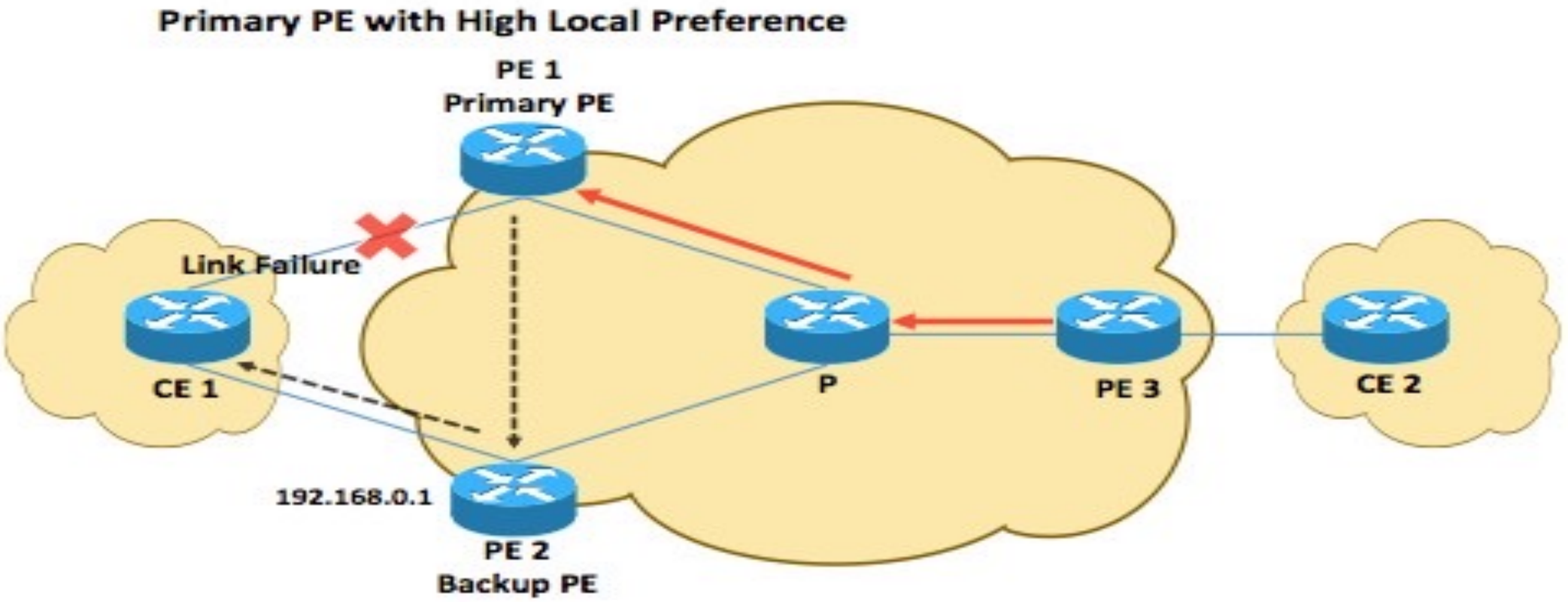
https://t.me/learningnets

# BGP Best External

- Question : Where is BGP Best External feature used? What is the benefit of using it ?

https://t.me/learningnets

# BGP Best External




Without Best External PE3 cannot learn the standby/backup path.
Best External is used for Active/Standby customer
Link topologies on the Service Provider.
Customer doesn't need to run BGP Best External

https://t.me/learningnets

**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

https://t.me/learningnets

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

https://t.me/learningnets

# BGP Route Reflector Clusters

- Question : Customer wants to use two BGP Route reflector for the redundancy but they don’t know the design best practices whether they should use same or different BGP Route Reflector Cluster ID ? Can you help them ?

- Yes
- No

https://t.me/learningnets

**Should you use different or same Cluster IDs if you have more than one RR in BGP design?**

- Almost always use same RR. With different cluster IDs on RR, you will accept and keep the prefixes on the RR. Those prefixes will never be used. But with same cluster ID, prefixes will not be accepted since the ID is the same, this will reduce the resource consumption.

Do you need to install all the prefixes into the RIB and FIB ?

NO ! If RR is in the data path !.

Is there any Exception?

Yes, for example Seamless/Unified MPLS scenario

https://t.me/learningnets

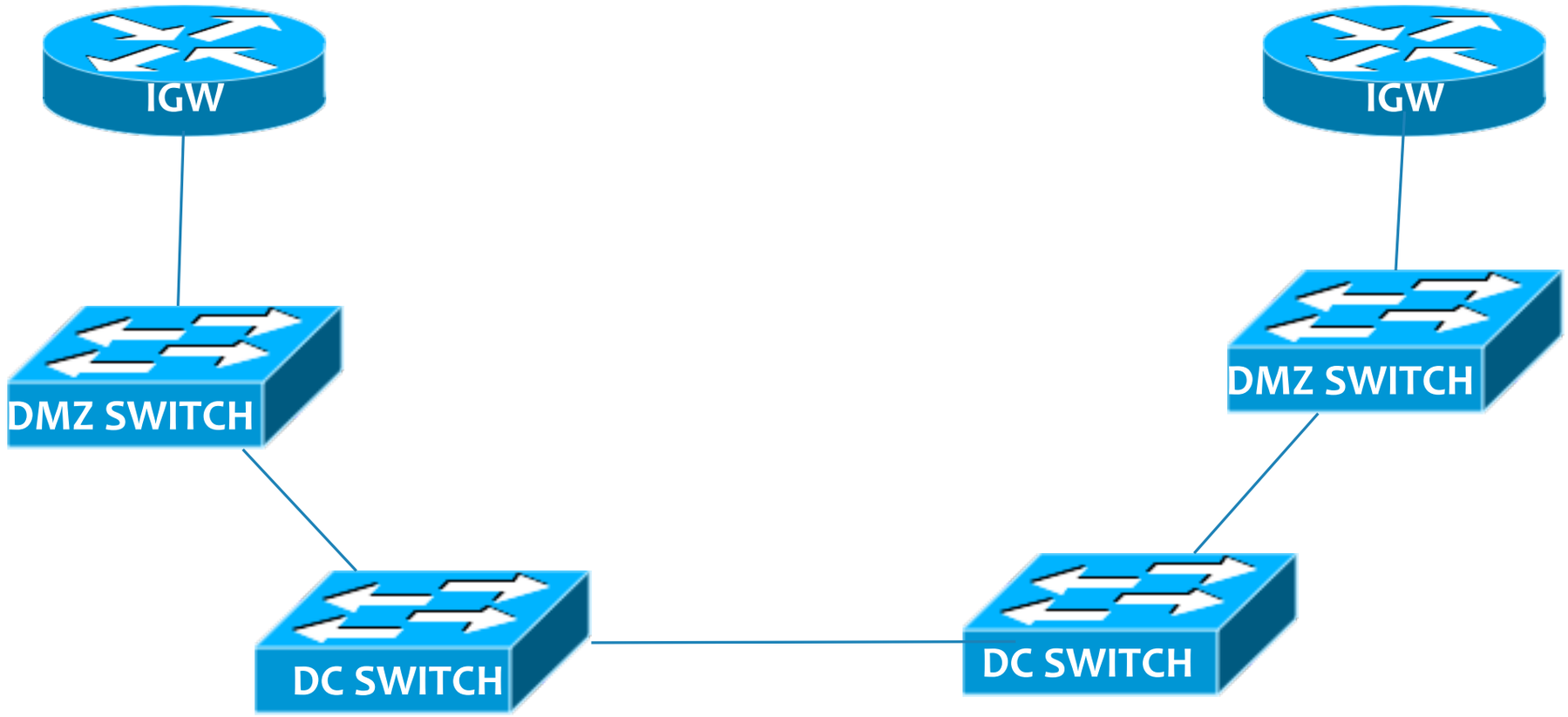
# IBGP over Physical and Tunnel Connections

- Enterprise company has two datacenters. They have 200 remote offices and all the locations access to the internet from the Datacenters

- They recently had an outage on the internet circuit and all the Internet sessions from the remote offices which uses that link wad dropped

- What are the solutions to prevent the session failure in case of a link failure on the Internet Gateways of this company ?

https://t.me/learningnets

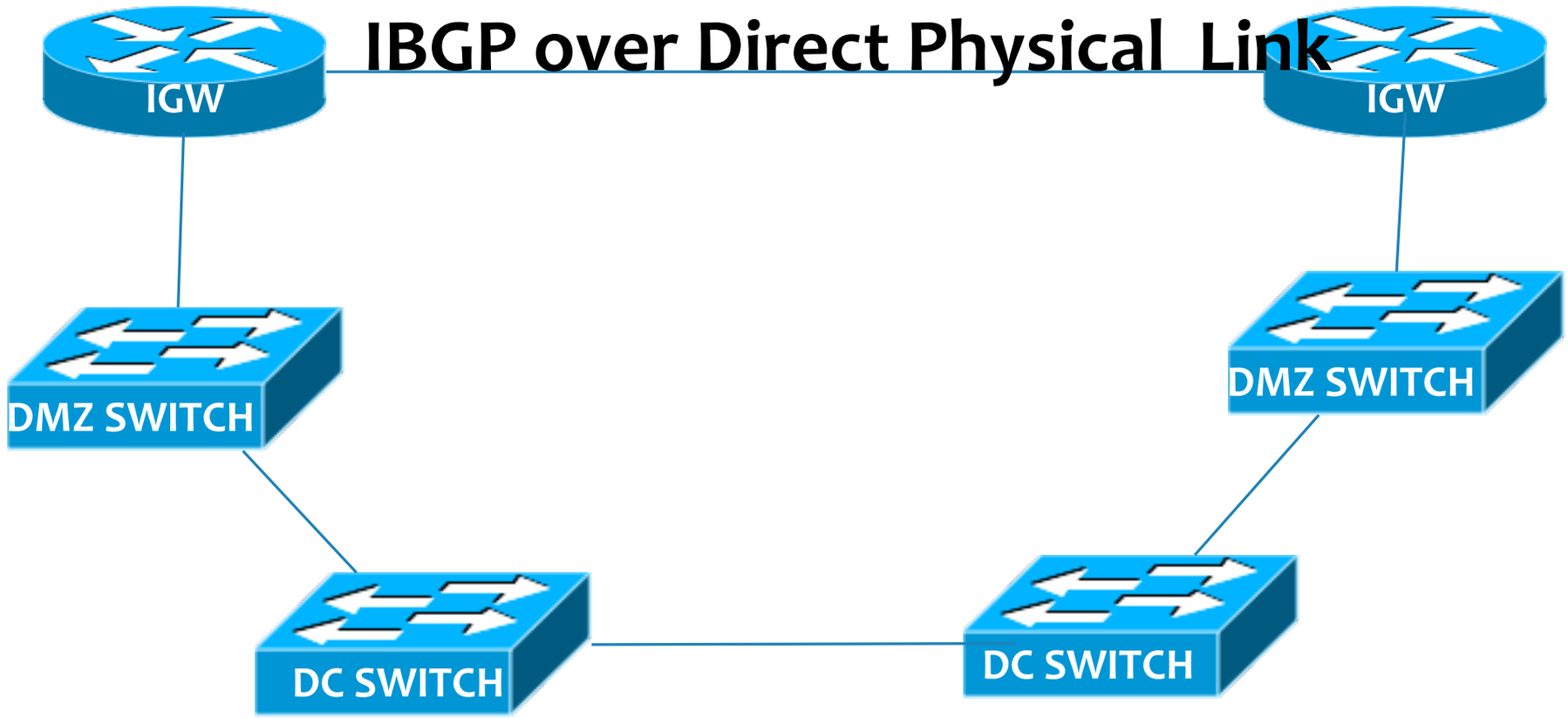



DATACENTER 1

DATACENTER 2

https://t.me/learningnets

**Solution 1 – Best option but can be costly. Budget might be concern, also deployment might take longer compare to other solutions**

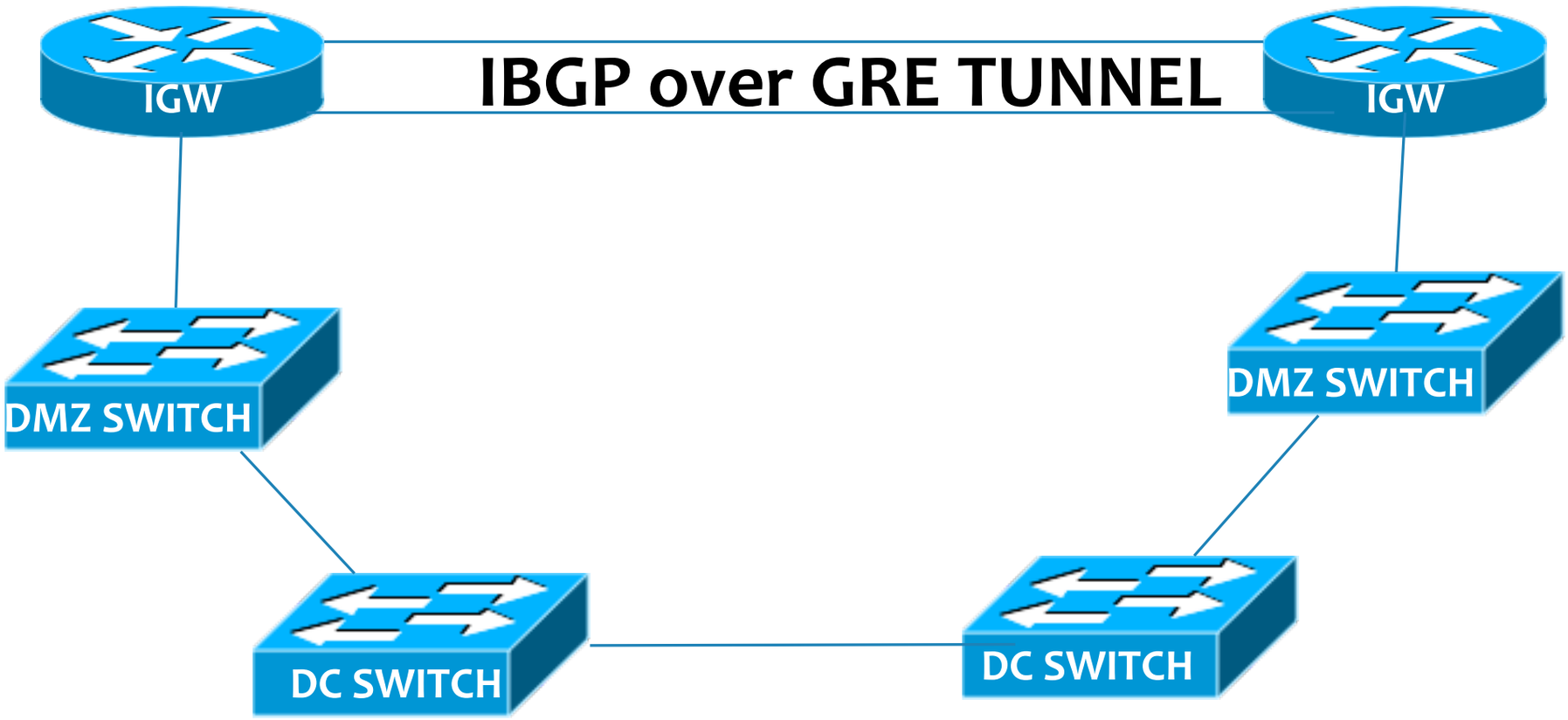



**DATACENTER 1**

**DATACENTER 2**

https://t.me/learningnets

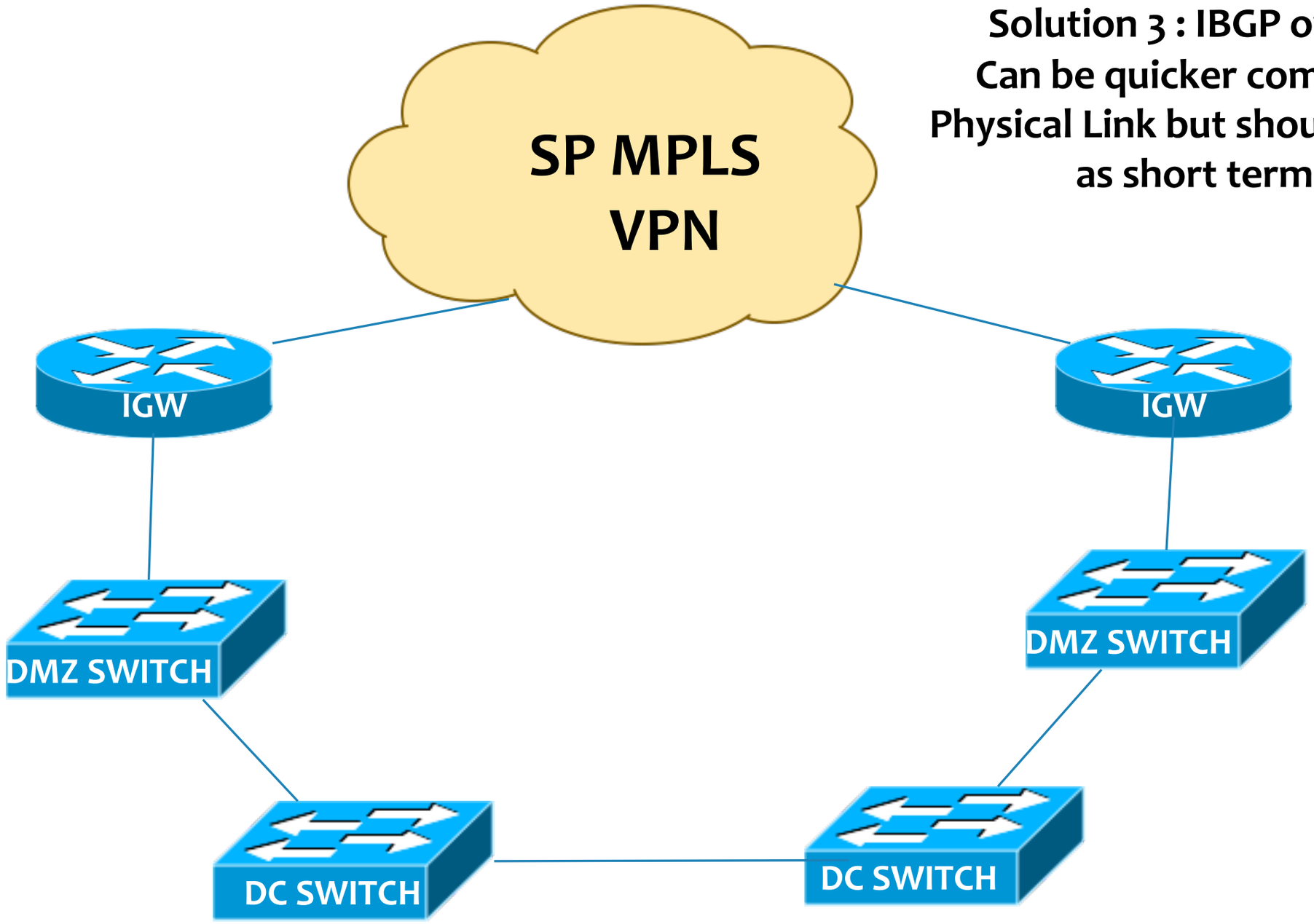
**Solution 2 – Fastest option , don't require Service Provider Interaction.**
**It should be used as a short term solution**




**DATACENTER 1** **DATACENTER 2**

https://t.me/learningnets

**Solution 3 : IBGP over MPLS VPN**
**Can be quicker compare to Direct Physical Link but should be considered as short term solution**




**DATACENTER 1**

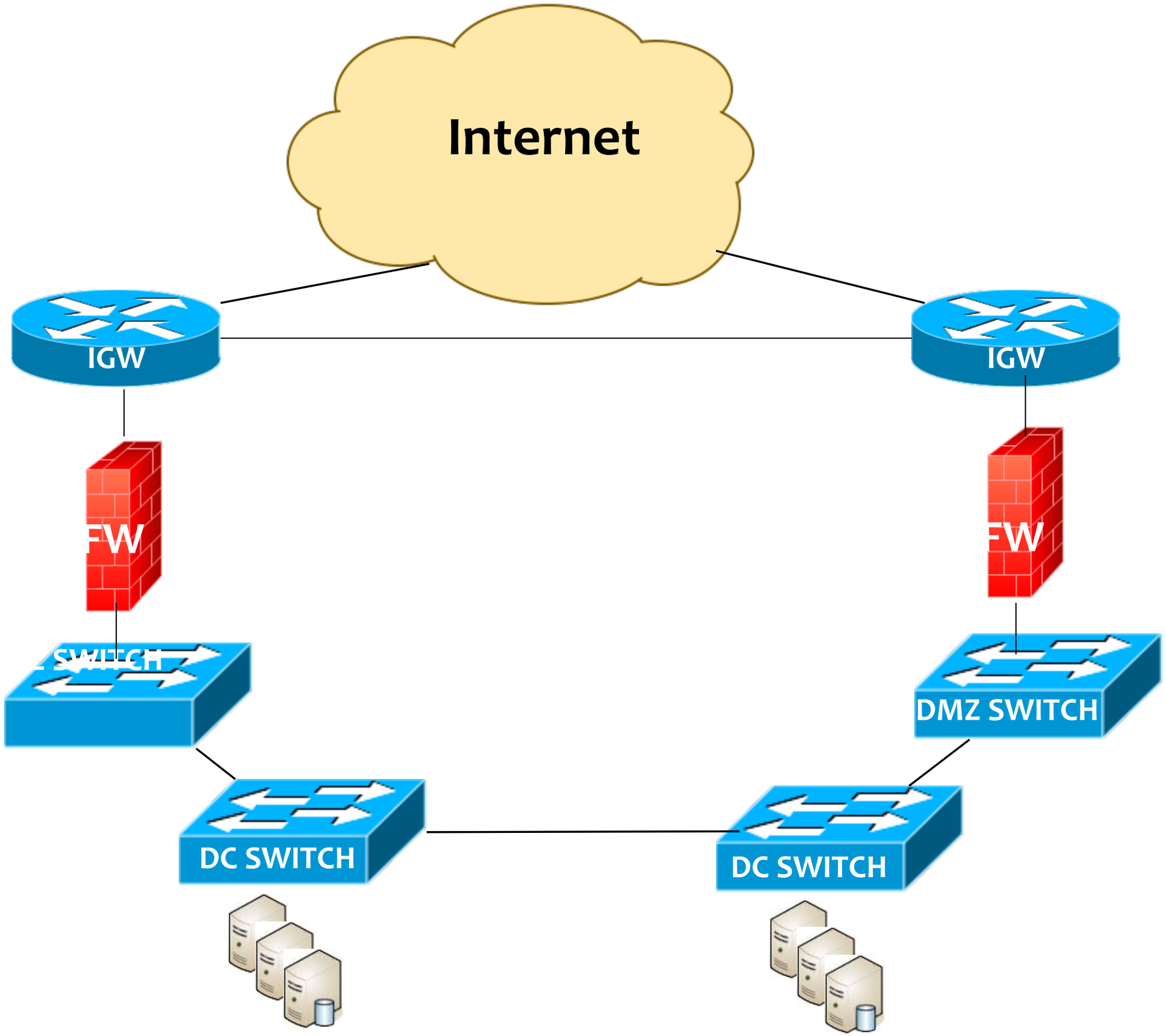
**DATACENTER 2**

https://t.me/learningnets

# Enterprise Edge Design with BGP

- U.S based Enterprise e-commerce company is designing a new network. They have two datacenters and both datacenters will host many servers.
- There are 1000km between the two datacenters.
- Their networking team knows that for the best user performance traffic should be symmetric between servers and the users/clients.
- In addition to datacenter interconnect link they have direct physical connection
- Based on the below topology what might be the issue ?

https://t.me/learningnets




https://t.me/learningnets

- It is already given in the requirements that traffic from DC1 should come back to DC1 directly. Asymmetric traffic cause firewall to drop all the traffic

- So if the users are accessing to DC1 servers it should go back from the DC1. Classical design for this, servers uses DC switch as default gateway. DC switches receive default route redistributed to their IGP from BGP by the IGW. And IGP cost is used to reach to the closest IGW by the DC switches

- Incoming traffic can be a problem whenever there is a stateful device in the path

- In the above topology if traffic comes to DC1 it has to go back from DC1 and vice versa, it is not only for asymmetric flow on the Firewalls, Load Balancers and so on but also important to avoid hairpin

- If traffic destined to DC1 comes to DC2, it has to go through direct physical internet link to the DC1, this adds additional latency and consume unnecessary bandwidth

https://t.me/learningnets

- Question 2: How does company achieve symmetric traffic flow so they don’t have any traffic drop or performance issue ?

- They can split their public IP space to half and advertise specifics from each datacenters and summary from both datacenters as a backup in case first DC IGW link or node fails

- Imagine they have /23 address space, they can dive 2x/24 and advertise each /24 from local datacenters only and /23 from both datacenters. Since their upstream SP will profer longest match routing over any other BGP attribute, traffic returns to the location where it is originated

https://t.me/learningnets

# BGP vs. IGP Comparison

| orhanergun.net | OSPF | IS-IS | EIGRP | BGP |
|---|---|---|---|---|
| Scalablability | 2 tier hierarchy , less scalable | 2 tiers hierarchy , less scalable | Support many tiers and scalable | Most scalable routing protocol |
| Working on Full Mesh | Works well with mesh group | Works well with mesh group | Works very poorly, and there is no mesh group | Works very poorly, but RR removes the requirement |
| Working on a Ring Topology | Its okay | Its okay | Not good if ring is big due to query domain | Good with Route Reflector |
| Working on Hub and Spoke | Works poorly, require a lot of tuning | Works bad requires tuning | Works very well. It requires minimum tuning | IBGP works very well with Route Reflector |
| Fast Reroute Support | Yes - IP FRR | Yes - IP FRR | Yes - IP FRR and Feasible Successor | Requires BGP PIC + NHT + Best external + Add-Path |
| Suitable on WAN | Yes | Yes | Yes | Yes, but in very large scale or when policy is needed |
| Suitable on Datacenter | DCs are full mesh. So, No | DCs are full mesh so No | DCs are full mesh so no | Yes, in large scale DC and it is not uncommon |
| Suitable on Internet Edge | No it is designed as an IGP | No it is designed as an IGP | No, it is designed as an IGP | Yes, it is designed to be an Inter domain protocol |
| Standard Protocol | Yes IETF Standard | Yes IETF Standard | No, there is a draft but lack of Stub feature | Yes, IETF Standar |
| Stuff Experince | Very well known | Not well known | Well known | Not well known |
| Overlay Tunnel Support | Yes | Doesn't support IP tunnels | Yes | Yes |
| MPLS Traffic Engineering Support | Yes with CSPF | Yes, with CSPF | No | No |
| Security | Less secure | More secure since it is on layer2 | Less secure | Secure since it runs on TCP |
| Suitable as Enterprise IGP | Yes | No, it lacks Ipsec | Yes | Not exactly, very large scale networks only |
| Suitable as Service Provider IGP | Yes | Definitely | No, it doesn't support Traffic Engineering | Maybe in the datacenter but not as an IGP |
| Complexity | Easy | Easy | Easy | Complex |
| Policy Support | Good | Good | Not so Good | Very good |
| Resource Requirement | SPF requires more processing power | SPF requires more processing power | DUAL doesn't need much power | Requires a lot of RAM and decent CPU |
| Extendibility | Not good | Good, thanks to TLV support | Good, thanks to TLV support | Very good, it supports 20 + address families |
| IPv6 Support | Yes | Yes | Yes | Yes |
| Default Convergece | Slow | Slow | Fast with Feasible Successor | Very slow |
| Training Cost | Cheap | Cheap | Cheap | Moderate |
| Troubleshooting | Easy | Very easy | Easy | Moderate |
| Routing Loop | Good protection | Good protection | Open to race condition | Good protection |

https://t.me/learningnets

# BGP in the CCDE Exam

- IBGP will be main focus for the CCDE Exam
- BGP Route Reflectors and the different BGP Route Reflector Design Options
- BGP Confederation and using multiple Sub AS for the multinational providers and so on.
- Multi Protocol BGP should be understood very well

https://t.me/learningnets

# BGP in the CCDE Exam

- Different NLRI, especially IPv6, IPv6 and VPN address families with BGP should be understood

- BGP AS Migration needs to be understood, changing one AS with another one , maybe in a merger or acquisition scenario with Local AS feature

- BGP Traffic Engineering with BGP Path Attributes. Local Pref, MED, As-Path Prepend and Community etc.

https://t.me/learningnets

# BGP Summary

- BGP Use Cases , Theory
- EBGP Loop Prevention , EBGP Traffic engineering , BGP Path Attributes
- EBGP – Inter-domain routing, IXP , Peering , IP Transit Business
- Service Provider Business models – GAO- Rexford model and default SP BGP policy

https://t.me/learningnets

# BGP Summary

- IBGP , Full Mesh IBGP, Route Reflectors, Confederation and their design options

- BGP Add-Path , Shadow RR , Unique RD per VRF per PE

- BGP Egress Peer Engineering

- BGP RTBH and BGP Flowspec

https://t.me/learningnets

# BGP Summary

- BGP Information Security , Route Leaks , BGP Hijacks, Route Origin Validation and Path Validation
- IRR, Peeringdb, RPKI and BGPSEC
- BGP in the Datacenter – IBGP vs. EBGP , EBGP ASN Allocation, BGP Path Hunting
- BGP PIC – Flat , Hierarchical and Generalized FIB Architecture , PIC Edge and PIC Core
- BGP vs. IGP Comparison

https://t.me/learningnets

# BGP Quiz !!

## Questions and the Answers

https://t.me/learningnets

# BGP Quiz - Question 1

**Which of the below option is the reason to run IBGP? (Choose Two)**

A. It is used for the reachability between PE devices in MPLS network

B. It is used to carry EBGP prefixes inside an Autonomous System

C. It is used with Route Reflectors for the scalability reason in large scale networks

D. It is used to prevent failures outside your network from impacting your internal network operation

https://t.me/learningnets

# BGP Quiz – Answer 1

One of the correct answers of this question is to carry EBGP prefixes inside an Autonomous system.

IGP is used for the reachability between PE devices in an MPLS network.
Option C is valid but not the correct answer, because; question is asking the reasons, not the best practices.

Option D is one of the correct answers as well because with IBGP, internal network is protected from the outside failures by separating the local failure domains.
That’s why; answers of this question are B and D.

https://t.me/learningnets

# BGP Quiz - Question 2

**Which of the below options are true for the BGP Route Reflectors? (Choose Three)**

A. Route Reflectors provide scalability in large scale network design

B. Route Reflectors hide the available paths

C. Route Reflectors selects and advertise only the best path to Route Reflector clients

D. Route Reflectors can be placed anywhere in the IP backbone as an IPv4 RR

https://t.me/learningnets

# BGP Quiz - Answer 2

Route reflectors are used to improve scalability of the BGP design in large-scale deployments.

Route reflectors hide the available path information by selecting and advertising only the best path to the clients.

Thus the correct answer of this question is A, B and C.

Option D is wrong because, Route Reflectors should follow the physical topology in an IP backbone, it cannot be placed everywhere, careful planning is required. Otherwise forwarding loop occurs as it was explained in one of the case studies in the BGP chapter.

https://t.me/learningnets

# BGP Quiz - Question 3

**Which below attributes are commonly used for BGP path manipulation? (Choose Three)**

A. Local Preference

B. Origin

C. As-Path

D. Community

E. Weight

https://t.me/learningnets

# BGP Quiz – Answer 3

Origin is not used commonly for the BGP path manipulation. Weight is Cisco preparatory and it is only local to the routers. It shouldn't be used for path manipulation.

BGP path manipulation was explained in detail in BGP chapter.

Answer of this question is A, C and D.

https://t.me/learningnets

# BGP Quiz - Question 4

**Which of the below options is used in the Public Internet Exchange Points to reduce configuration overhead on the BGP devices?**

A. BGP Route Reflectors

B. BGP Prefix Lists

C. BGP Route Servers

D. BGP Map Servers

https://t.me/learningnets

# BGP Quiz – Answer 4

There is nothing called BGP Map Servers. In the Public Internet Exchange points BGP Route Servers are used to reduce configuration overhead.

They improve scalability. Very similar to Route Reflectors but Route Reflectors are used in IBGP, not in the Public Exchange Points. That’s why answer of this question is C.

https://t.me/learningnets

# BGP Quiz - Question 5

**Which below options are true for the BGP Confederation? (Choose Three)**

A. It is done by creating Sub-Autonomous system

B. It is easier to migrate from full-mesh IBGP, compare to BGP Route Reflectors

C. Between Sub Autonomous Systems mostly EBGP rules apply

D. Compare to BGP Route Reflector design, it is less commonly deployed in the networks

https://t.me/learningnets

# BGP Quiz - Answer 5

From the migration point of view, Full mesh IBGP to BGP Confederation is harder, compare to BGP Route Reflectors. Thus Option B is invalid.

All the other options are correct thus the answer of this question is A, C and D

https://t.me/learningnets

# BGP Quiz - Question 6

**Which below option is used for inbound BGP path manipulation? (Choose Three)**

A. Local Preference

B. MED

C. As-Path prepending

D. Community

E. Hot Potato Routing

https://t.me/learningnets

# BGP Quiz - Answer 6

Hot Potato Routing and Local Preference are used for Outbound BGP Path manipulation as explained in the BGP chapter in detail.

MED should be used if there is only one upstream ISP but still it is used for inbound path manipulation. AS-Path prepending and the communities are used for the multihoming connections as well.

That’s why; answer of this question is B, C and D.

https://t.me/learningnets

# BGP Quiz - Question 7

**What does MP-BGP (Multi Protocol BGP) mean?**

A. BGP implementation which can converge less than a second

B. BGP implementation which is used in Service Provider networks

C. BGP implementation which can carry multiple BGP Address Families

D. BGP implementation which is used in Enterprise Networks

https://t.me/learningnets

# BGP Quiz - Answer 7

MP-BGP (Multi Protocol BGP) is the BGP implementation, which can carry multiple Address Families. BGP in 2016, can carry more than 20 different Address Families such as IPv4 Unicast, IPv6 Unicast, IPv4 Multicast, L2 VPN, L3VPN, Flowspec and so on.

That’s; why; answer of this question is C.

https://t.me/learningnets

# BGP Quiz - Question 8

- **What does Hot Potato Routing mean?**

A. Sending the traffic to the most optimum exit for the neighboring AS

B. Sending the traffic to the closest exit to the neighboring AS

C. By coordinating with the neighboring AS, sending traffic to the closest exit point

D. It is the other name of BGP Multipath

https://t.me/learningnets

# BGP Quiz - Answer 8

Hot Potato Routing means, sending the traffic to the closest exist point from the Local Autonomous system to the neighboring Autonomous System by taking the IGP metric into consideration

There is no coordination between the Autonomous System in Hot Potato Routing definition. But Coordination with the Hot Potato Routing case study was provided in the BGP chapter.

That’s why; answer of this question is B

https://t.me/learningnets

# BGP Quiz - Question 9

**Fictitious Service Provider is considering providing an availability SLA for their MPLS VPN customers. They want to provide sub second convergence in case link or node failure scenarios.**
**What would you suggest to this company to achieve their goal? (Choose Two)**

A. Implementing BFD

B. Implementing BGP PIC Core and Edge

C. Implementing BGP Route Reflectors

D. Implementing IGP FRR

https://t.me/learningnets

# BGP Quiz - Answer 9

They should implement BGP PIC features to protect BGP from the link or node failure. Especially Edge node failures, even if MPLS Traffic Engineering or IP FRR deployed, couldn’t be recovered in sub second.

Since BGP PIC convergence is mostly depends on IGP convergence as well, deploying IGP FRR (Fat Reroute) provides a necessary infrastructure for the BGP PIC.

They should be deployed together. BFD is just a failure detection mechanism. IGP Convergence is depends on many other parameters tuning

That’s why; answer of this question is B and D

https://t.me/learningnets

# BGP Quiz - Question 10

- **With which below options, internal BGP speaker can receive more than one best path even if BGP Route Reflectors are deployed? (Choose Three)**

A. BGP Shadow RR

B. BGP Shadow Sessions

C. BGP Add-path

D. BGP Confederation

E. BGP Multipath

https://t.me/learningnets

# BGP Quiz - Answer 10

Shadow Sessions, Shadow RR and BGP Add-path design provides more than best path to the internal BGP speaker even if BGP Route Reflectors are deployed.

BGP Multipath requires more than one best path and all the path attributes to be the same. Thus it requires one of the above mechanisms. BGP Confederation doesn’t provide this functionality.

That’s why; answer of this question is A, B and C.

https://t.me/learningnets

# BGP Quiz - Question 11

**Which below option is recommended to send more than one best path to the VPN PEs in the MPLS VPN deployment if VPN Route Reflectors are deployed?**

A. BGP Add-path

B. BGP Shadow RR

C. BGP Full Mesh

D. Unique RD per VRF per PE

https://t.me/learningnets

# BGP Quiz - Answer 11

BGP Add-path, BGP Shadow RR and Sessions deployments are suitable for the IP backbones.

If there is an MPLS backbone, configuring unique RD per VRF per PE is best and recommended design option since there is no software or hardware upgrade, no additional BGP sessions and so on.

That’s why the answer of this question is D

https://t.me/learningnets

# BGP Quiz - Question 12

**What are the reasons to send more than one BGP best path in IP and MPLS deployment? (Choose Four)**

A. BGP Multipath

B. BGP Fast Reroute

C. BGP Multihop

D. Preventing Routing Oscillation

E. Optimal BGP routing

https://t.me/learningnets

# BGP Quiz - Answer 12

There are many reasons to send more than one BGP best path in both IP and MPLS deployments.

These are; avoiding routing oscillations, BGP Multipathing, Fast convergence/Fast Reroute and Optimal Routing.

Sometimes for the optimal routing, just sending more than one BGP best path is not enough but may require all available paths though.

That’s why, answer of this question is A, B, D and E

https://t.me/learningnets

# BGP Quiz - Question 13

**What is the drawback of sending more than one BGP best path in BGP?**

A. More resource usage

B. Sub Optimal Routing

C. Slower Convergence

D. Security Risk

https://t.me/learningnets

# BGP Quiz - Answer 13

Sending more than one BGP best path requires more memory, CPU, network bandwidth, thus more resource usage in the network.

As a rule of thumb, whenever more information is sent, it consumes more resource, may provide optimal routing, better high availability, better convergence.

All other options are wrong, except Option A

https://t.me/learningnets

# BGP Quiz - Question 14

**What below options are the advantages of Full Mesh IBGP design compare to BGP Route Reflector design? (Choose Four)**

A. It can provide more optimal routing compare to Route Reflector design

B. It can provide faster routing convergence compare to Route Reflector design

C. It provides better resource usage compare to Route Reflector design

D. It can provide better protection against route churn

E. Multipath information is difficult to propagate in a route reflector topologies

https://t.me/learningnets

# BGP Quiz - Answer 14

Although there are advantages of using BGP Route Reflectors, there are many drawbacks as well. Probably it is more harmful than deploying Full Mesh IBGP if the requirement is optimal routing, faster convergence and avoiding route churns.

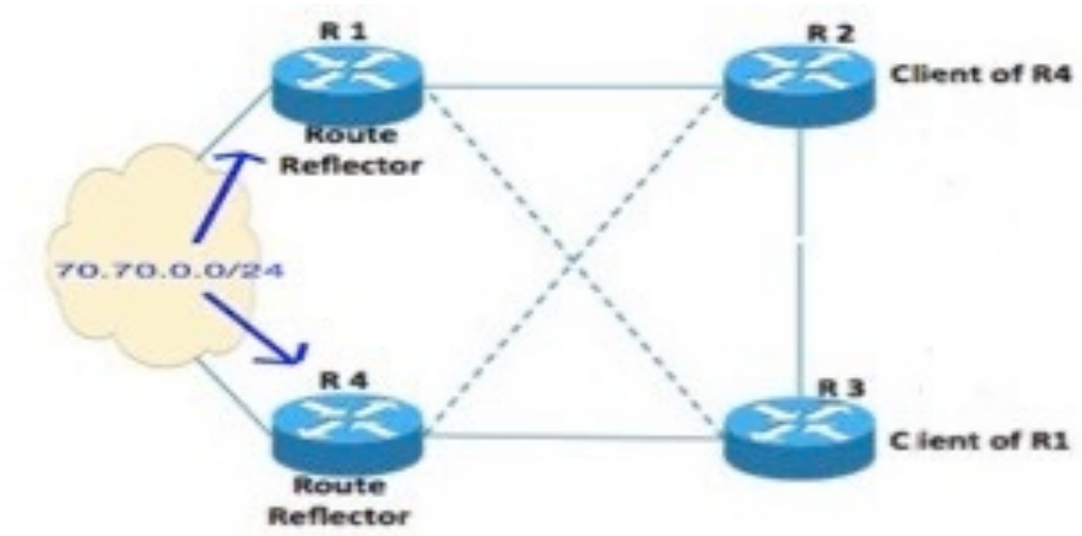
Sending multiple paths is difficult since it requires Shadow Sessions, RR or Add-path deployments in Route Reflector topologies.

Full Mesh IBGP design consumes more device and network resources and requires more configurations on the devices compare to Route Reflector design. That’s why the answer of this question is A, B, D and E

https://t.me/learningnets

# BGP Quiz - Question 15

In the below topology IP backbone is shown. R2 is the RR client of R4 and R3 is the RR client of R1.




What is the next hop of R2 and R3 for the 70.70.0.0/24 prefix?

- A. R1 is the next hop of R2, R4 is the next hop of R3
- B. R1 is the next hop of R3, R4 is the next hop of R2
- C. R2 is the next hop of R3, R3 is the next hop of R2
- D. R4 is the next hop of both R2 and R3

https://t.me/learningnets

# BGP Quiz - Answer 15

Since it is given as IP backbone, IP destination based lookup is done for the BGP prefixes.
Sine BGP prefixes require recursion and IGP next hop needs to be found for the BGP prefixes, R2's and R3's IGP next hops for the BGP prefixes should be found.

On R2, For the BGP next hop of 70.70.0.0/24 BGP prefix is R4. R2 can only reach R4 through R3.

Thus, R2's IGP next hop is R3. It applies for the R3.
R2's IGP next hop is R3 and R3's IGP next hop is R2. That's why the answer of this question is C.

https://t.me/learningnets

# BGP Quiz - Answer 15

Please note that in this topology BGP Route Reflectors don’t follow the physical topology, which is against to BGP Route Reflector design requirement in IP networks.

That’s why, in this design between R2 and R3, routing loop occurs.

Correct design is R2 should be the Route Reflector client of R1 and R3 should be the Route Reflector client of R4

https://t.me/learningnets

# BGP Quiz - Question 16

**What can be the problem with BGP design in the Enterprise if there are more than one datacenter?**

A. Convergence is very slow

B. Asymmetric routing issues if there are stateful devices

C. Route Reflector deployment is harder compare to SP deployment

D. Traffic flow cannot be optimized

https://t.me/learningnets

# BGP Quiz - Answer 16

All the options are wrong except Option B.

Asymmetric can be a problem in Enterprise design, which has stateful devices as it was explained in the BGP chapter. Because stateful devices require symmetric routing for the flow information and firewalls, load balancers, IDS/IPS are common elements at the Internet edge or within the datacenters in Enterprise design.

In the Service Providers, CGN (LSN) is deployed to overcome IPv4 exhaustion problem. These nodes also require symmetric routing.
Answer of this question is B

https://t.me/learningnets

# BGP Quiz - Question 17

**Which below option is true for the VPN Route Reflectors in MPLS deployments? (Choose Two)**

A. It can be deployed in centralize place

B. It doesn’t have to follow physical topology, can have more flexible placement compare to IP Route Reflectors

C. It is best practice to use VPN Route Reflectors for the IP Route Reflectors as well

D. It always provides most optimal path to the Route Reflector clients

https://t.me/learningnets

# BGP Quiz - Answer 17

VPN Route reflector can be deployed in the centralized placed and they have more flexible placement advantage compare to the IP Route Reflector.
The reason is there is no IP destination based lookup in the MPLS networks.

Thus there is no layer 3 routing loop problem as in the case of IP Route Reflector which was explained in the Answer 15.

https://t.me/learningnets

# BGP Quiz - Answer 17

It is not best practice to deploy IP and VPN services on the same node. Reason will be explained in Answer 18.
VPN RR, similar to IP RR, cannot always provide most optimal path to their clients. Because they selects the BGP best path from their point view, not from their clients point of view.

That’s why the answer of this question is A and B

https://t.me/learningnets

# BGP Quiz - Question 18

**What can be the problem with using IP and VPN Route Reflector on the same device? (Choose Two)**

A. Attack for the Internet service can affect VPN Customers

B. Attack for the VPN service can affect Internet Customers

C. Scalability of the Route Reflectors are reduced

D. They have to participate in the IGP process

https://t.me/learningnets

# BGP Quiz - Answer 18

When a Route Reflector is used for more than one service, it is called Multi Service Route Reflector. The problem of using Internet and VPN services on the same BGP Route Reflector is Fate Sharing

Internet based attacks can affect VPN customers and any problem on the VPN service users affect Internet customers. Also in case of failure, all the customers fail.

https://t.me/learningnets

# BGP Quiz - Answer 18

Thus using a separate BGP Route Reflector per service is a best practice. Using Multi Service RRs don’t reduce the scalability. And when using multi service RRs, they still don’t have to participate in IGP process

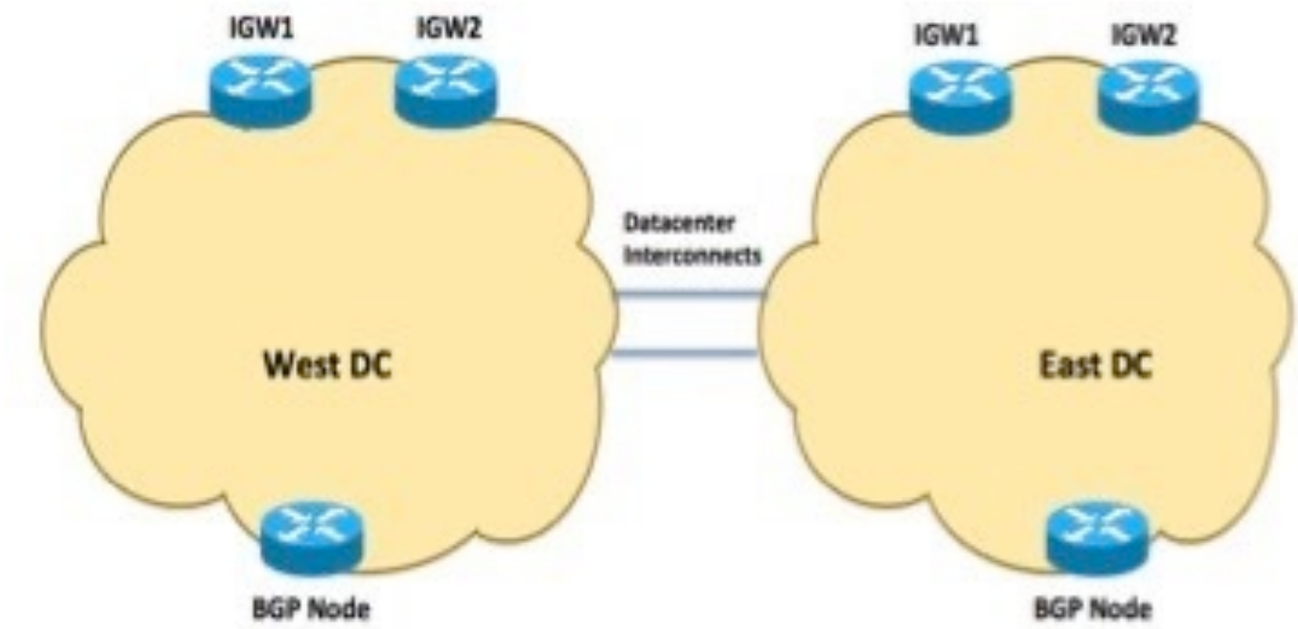
They can be designed as inline RR that participates IGP process in specific design such as Seamless MPLS.
Answer of this question is A and B

https://t.me/learningnets

# BGP Quiz - Question 19

In the below topology there are two datacenters of the Service Provider. If the requirement were to provide closest exit for the Route Reflector clients, in which datacenter would you deploy the Route Reflectors?

A. In West DC
B. In East DC
C. Doesn’t matter the placement
D. Both in East and West DC




https://t.me/learningnets

# BGP Quiz - Answer 19

Route Reflectors should be placed in both East and West DC. Otherwise Route Reflector would choose the best path from their point of view and would send the best path to the Route Reflector Clients from their best path.

If RR would be placed in West DC, all BGP RR Clients in East DC would choose the West DC IGW (Internet Gateways) as exit point and vice versa.

Thus the correct answer of this question is D

https://t.me/learningnets

# BGP Quiz - Question 20

**Which below options are true for the BGP PIC deployment? (Choose Two)**

A. BGP PIC can provide sub second convergence even if there are millions of prefixes in the routing table

B. BGP edge devices don’t have to receive more than one best path for BGP PIC Edge to work

C. BGP PIC Edge can protect both from Edge link and Node failure

D. BGP PIC has to work with BGP Add-Path

https://t.me/learningnets

# BGP Quiz - Answer 20

BGP edge nodes have to receive more than one best path for BGP PIC Edge operation. This was explained in the BGP chapter in detail. BGP Add-Path is one of the mechanisms, which is used to send multiple paths even RR is deployed in the network.

But BGP Add-Path is not mandatory for BGP PIC.

BGP PIC Edge can protect from both Edge link and node failures and can provide sub second convergence even if there are millions of prefixes.

That’s why the correct answer of this question is A and C

https://t.me/learningnets

# BGP Quiz - Question 21

**Which below option provide Route Origin Validation ?**

A. IRR

B. Peeringdb

C. RPKI

D. BGPSEC

E. Flowspec

https://t.me/learningnets

# BGP Quiz - Answer 21

IRR is used to provide information for filtering on the Internet facing routers but doesn’t provide Origin validation

Flowspec is not a validation mechanism, it provides protection for DDOS attack

BGPSEC provides path validation

RPKI is the only correct answer, it provides Origin Validation which mean, by creating ROAs, networks can validate the prefixes if they are coming from correct origin AS

https://t.me/learningnets

# BGP Quiz - Question 22

**What are the important consideration when EBGP is used in the Massively Scale Datacenters?**

A. BGP Path Hunting

B. ASN allocation schema

C. Whether there is a BGP Route Server

D. Traffic Engineering

E. Add-path or Shadow RR deployment

https://t.me/learningnets

# BGP Quiz - Answer 22

Add-path or Shadow RR are not a concern for EBGP deployments, question is asking EBGP deployment

Also in the Datacenter BGP Route Server is not used, it is used in the IXP networks

In the MSDC environments, ASN numbering schema is important to advertise routes , for example allow-as in might be required, or due to different ASN design BGP Path hunting can be a problem

https://t.me/learningnets

# BGP Quiz - Answer 22

Also traffic engineering with BGP is one of the main reasons why BGP is used inside the Massively Scale Data Centers. Between different TOR and Leaf or Leaf and Spine switches, for different type of applications or different type of traffic class, such as Elephant and Mice flows, DC operators provide different paths for different applications and the services.

Answer of this question is A. B and D

https://t.me/learningnets

# BGP – Extra Study Resources

**Books :**

- http://www.amazon.com/BGP-Design-Implementation-Randy-Zhang/dp/1587051095/ref=sr_1_1?ie=UTF8&qid=1436564612&sr=8-1&keywords=bgp+design+and+implementation

- **Videos** :

- https://www.nanog.org/meetings/nanog38/presentations/dragnet.mp4
  https://www.youtube.com/watch?v=txtiNFyvWjQ

https://t.me/learningnets

- **Articles :**

- https://www.nanog.org/meetings/nanog51/presentations/Sunday/NANOG51.Talk3.peering-nanog51.pdf

- http://ripe61.ripe.net/presentations/150-ripe-bgp-diverse-paths.pdf

- http://orhanergun.net/2015/05/bgp-pic-prefix-independent-convergence/

- http://orhanergun.net/2015/01/bgp-route-flap-dampening/

- https://www.nanog.org/meetings/nanog48/presentations/Tuesday/Raszuk_To_AddPaths_N48.pdf

https://t.me/learningnets

- 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/

https://t.me/learningnets

# Multicast Theory

- If the requirement is to send the flow in real time to multiple receivers, then the efficient way of this is Multicast

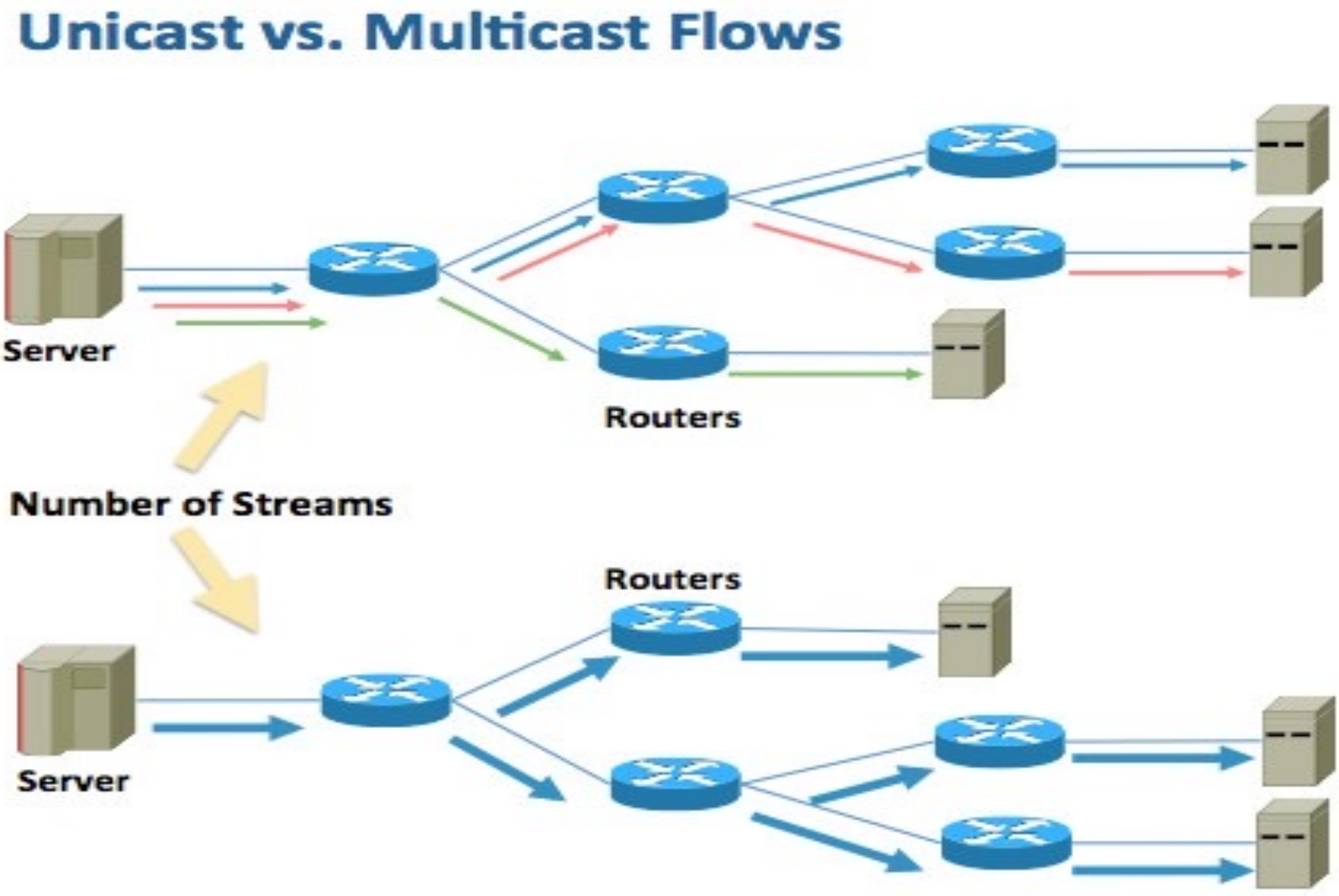
- Multicast is a 30 years old protocol and still many people are struggling to understand it

- Probably the biggest problem with the understanding Multicast is Source Based Routing

- IP unicast routing works based on destination based routing, IP multicast routing works based on source based routing

- Tree in IP unicast routing is created from the source towards destination, in IP Multicast routing from destination(Receiver) towards source (Sender)

https://t.me/learningnets




https://t.me/learningnets

- 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.

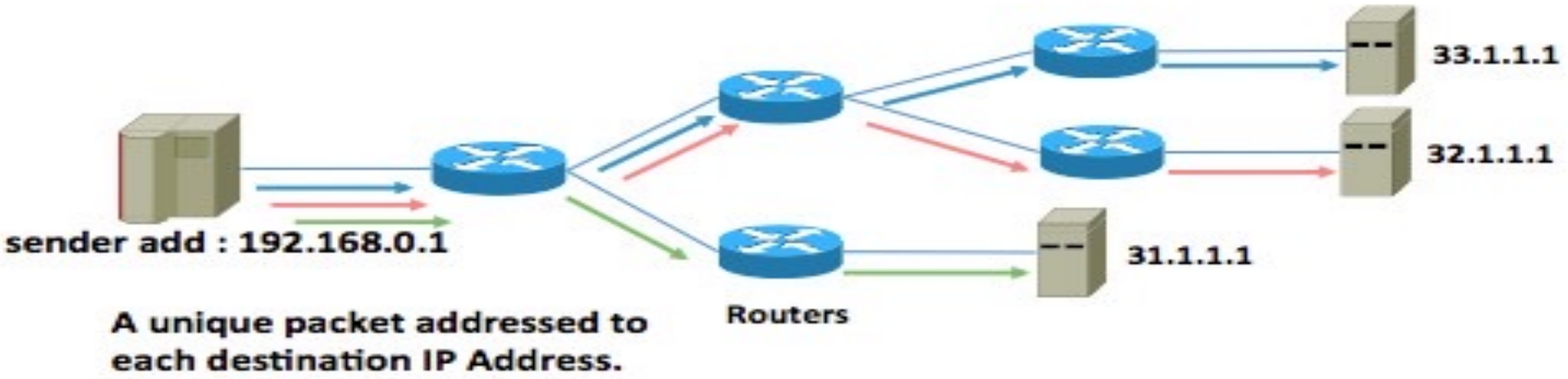
- Multicast works on UDP, not TCP. That’s why there is no error control, congestion avoidance, it is purely best effort

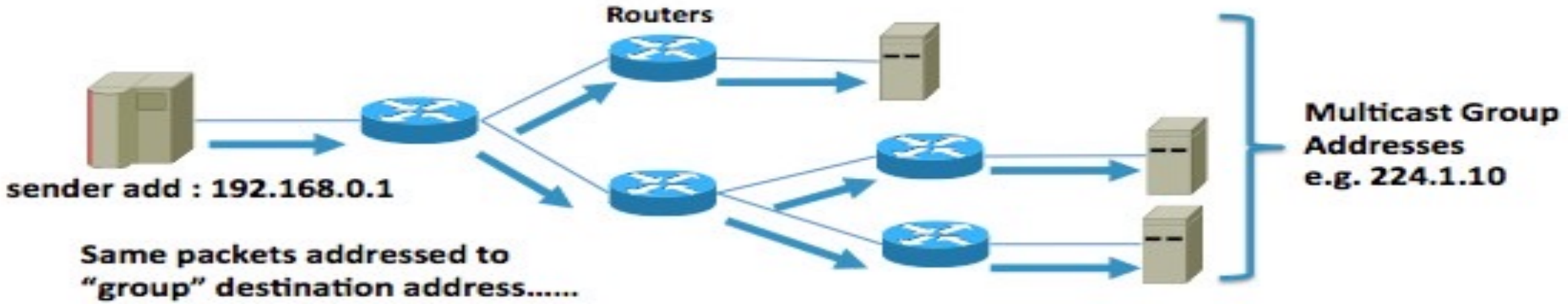
https://t.me/learningnets

- Receiver can receive duplicate Multicast traffic in some situations. SPT switchover is the example where duplicate traffic delivery occurs. During a SPT switchover, multicast traffic is received both from Shared tree and Shortest path tree.
- This is one of the inefficiency of Multicast

https://t.me/learningnets

# Unicast and Multicast Addressing




- Source addresses in Multicast always unicast address. Multicast address is a class D address range. 224/4. You should never see Multicast Class D address as a source Multicast address.

https://t.me/learningnets

- Source address can never be class D multicast group address. Separate multicast routing table is maintained for the multicast trees. Source don’t need to join any group, they just send the traffic.

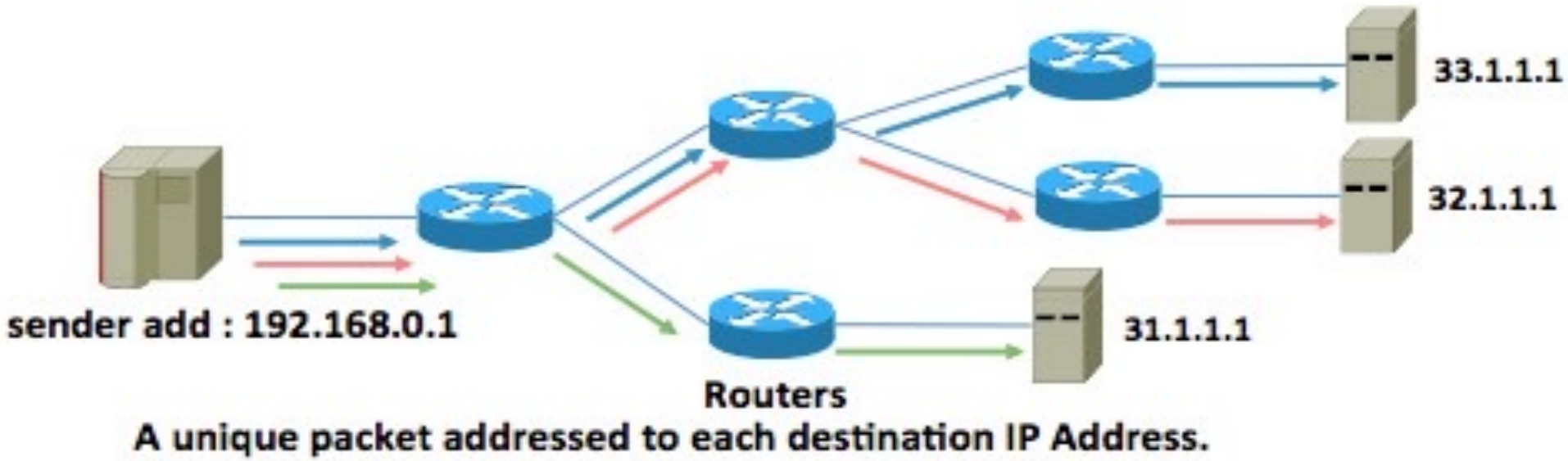
- Multicast routing protocols (DVMRP, PIM) is used to build the trees. Tree is built hop by hop from the receivers to the source. DVMRP was the first Multicast Routing Protocol and it is depreciated and not used anymore

- Source (Sender) is the root of the tree in shortest path tree.

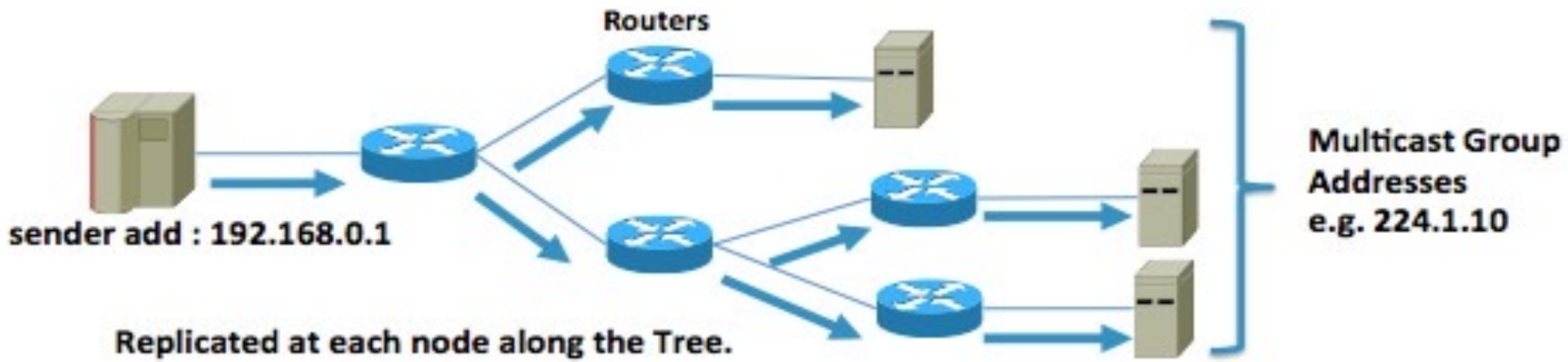
- Rendezvous Point is the root of the Shared Tree (Rendezvous Point is used in ASM and PIM Bidir)

https://t.me/learningnets

# Unicast and Multicast Addressing




A unique packet addressed to each destination IP Address.




Replicated at each node along the Tree.

https://t.me/learningnets

- 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 ; 224.0.0.0 – 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.

https://t.me/learningnets

- 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

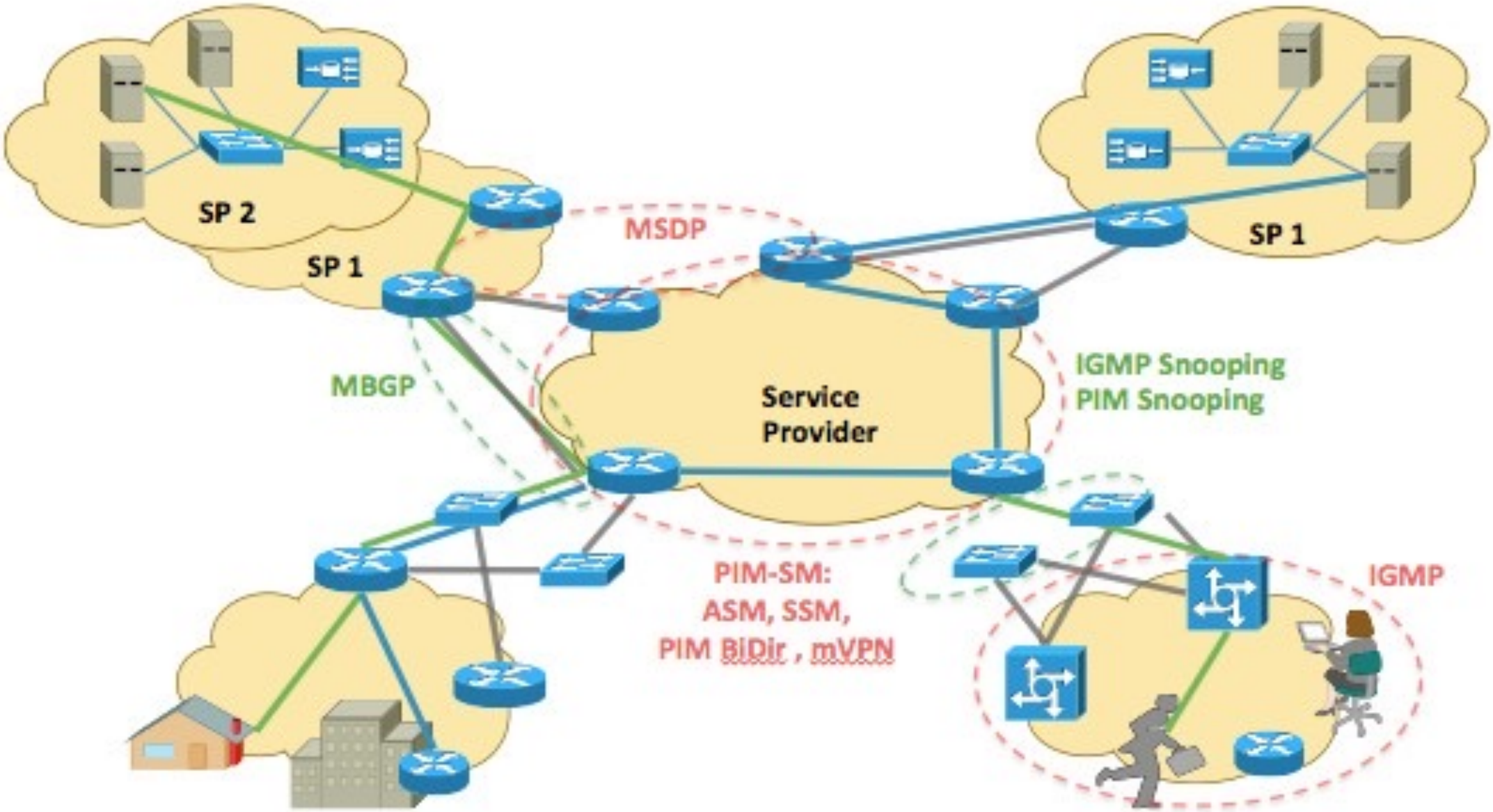
✓ 239.1.1.1

✓ 239.129.1.1

- All above address uses same multicast MAC address, which is 0100.5e01.0101

https://t.me/learningnets

# We will talk about below Multicast Protocols




https://t.me/learningnets

- Receiver uses IGMP to communicate with the router. IGMP v2 is used in PIM ASM (Any Source Multicast) and only IGMPv3 can be used with PIM SSM (Source Specific Multicast)

- IGMP is a host to router protocol. After first hop router you don’t see IGMP messages. IGMP Membership report which is also known as IGMP join is sent by the host application to the first hop router, then first hop router creates a PIM Join towards the root of the tree

- Receiver sends an IGMP Membership Report to the first hop router which in return sends PIM Join to the root of the tree.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

# Multicast Distribution Trees

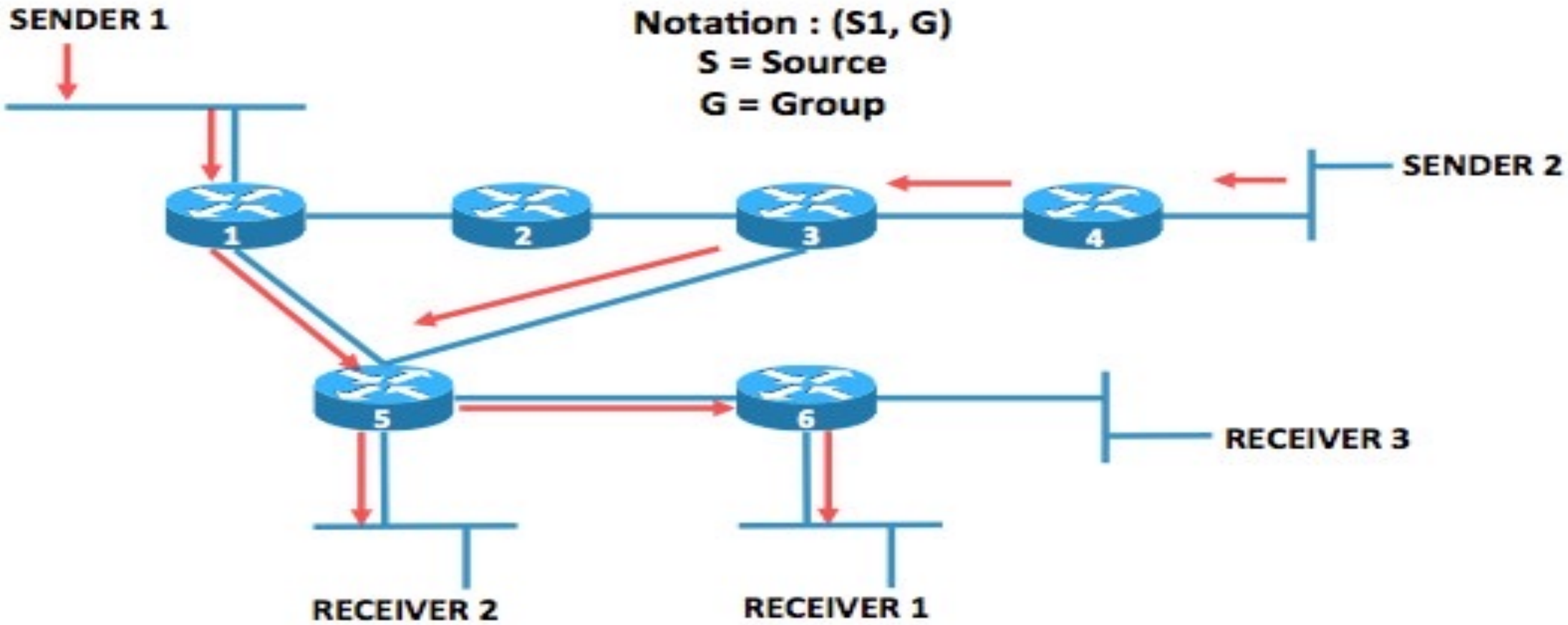
- Shortest path tree or Source tree is created between the source and the receiver in PIM SSM.

- Shortest path tree is created between the source and the Rendezvous point in PIM ASM before the SPT (Shortest Path Tree) switchover.

https://t.me/learningnets

# Short Path or Source Distribution Tree

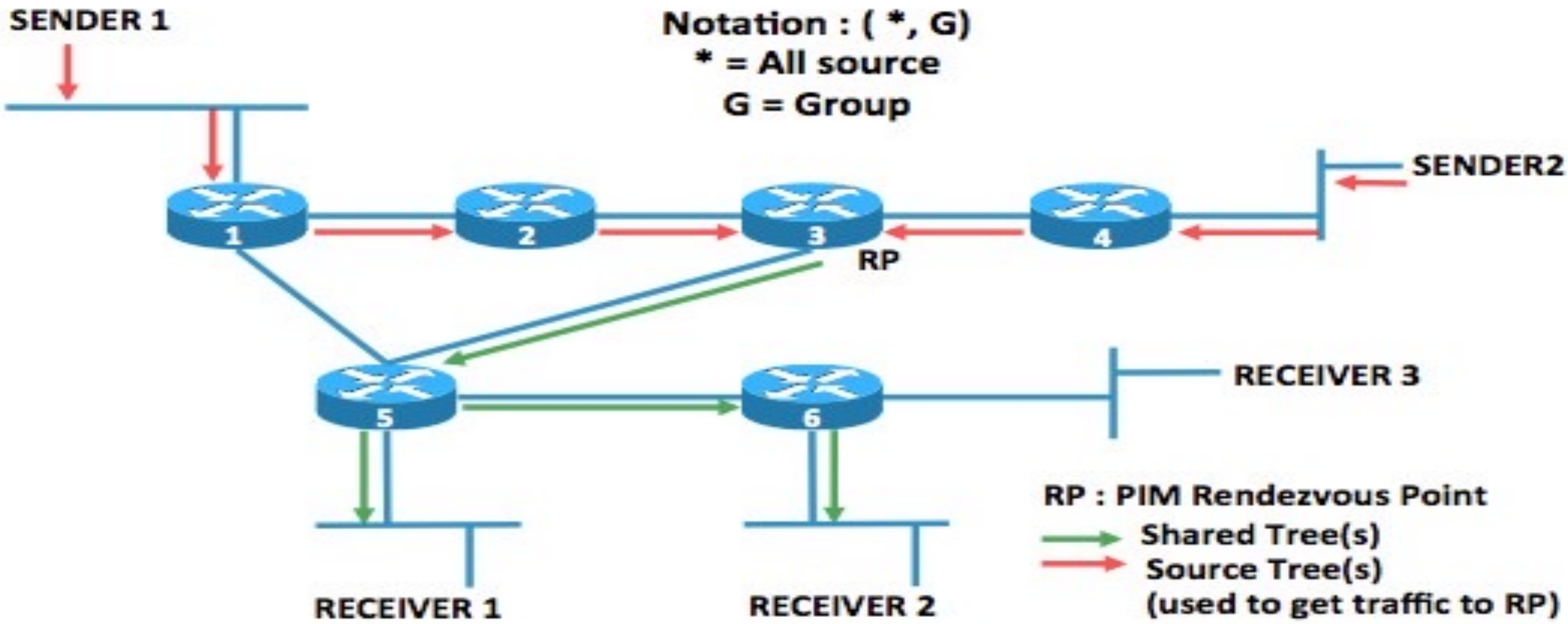
**There is no Rendezvous Point , No Rendezvous Point Engineering !**




- Shared tree is created between the Rendezvous Point and the receiver only in PIM ASM, still there is shortest path tree between the Source and Rendezvous point in PIM ASM.

https://t.me/learningnets

# Shared Tree




- Shortest path tree uses more memory but provides optimal path from the source to all receivers. That's why minimizes the delay.

- Shared tree uses less memory because you don't have separate multicast state for each source for the given multicast group address but may create suboptimal routing for some receivers. That's why shared tree may introduce extra delay.

https://t.me/learningnets

## PIM Protocol Independent Multicast

- PIM is a multicast routing protocol. Two PIM mode, PIM sparse and PIM dense mode.

## PIM Dense Mode:

- PIM dense mode is a flood and prune based (pushing) protocol. It consumes extra resources (Bandwidth, CPU and memory). Even though there is no intended receivers, multicast traffic is sent everywhere. Then if there are no receivers, routers prune the traffic. But flood and prune mechanism is repeated every so often.

https://t.me/learningnets

## PIM Sparse Mode

- Most commonly used PIM mode is the PIM Sparse Mode. Works based on pull mechanism unlike PIM Dense mode which is push based mechanism.
- Receivers in PIM Sparse mode request to join to a specific multicast group. Tree is created from the receivers up to the root.
- Root is a sender if the tree is shortest/source based, it is a rendezvous point if the tree is shared tree.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

# PIM SSM – Source Specific Multicast

- PIM SSM is a source specific multicast. Don’t require rendezvous point. Because sources are known by the receivers. Receivers create a shortest path tree towards the source Root of the tree is the sender.
- If SSM is enabled, only (S,G) entries are seen in the Multicast Routing table.

https://t.me/learningnets

- IANA reserves 232/8 Class D Multicast Address range for the PIM SSM but it’s not must to use these address range though.

- Require IGMPv3 at the receiver or IGMPv2 to v3 mapping on the last hop routers.

- Source specific multicast is most suitable for the one to many applications. IPTV is an example of one to many applications.

https://t.me/learningnets

## PIM Bidir – Bidirectional PIM

- PIM-Bidir is suitable to many to many multicast application such as trading floors application where all senders at the same time a receivers.

- Only uses shared tree Thus only (*,G) entries are seen in the multicast routing table.

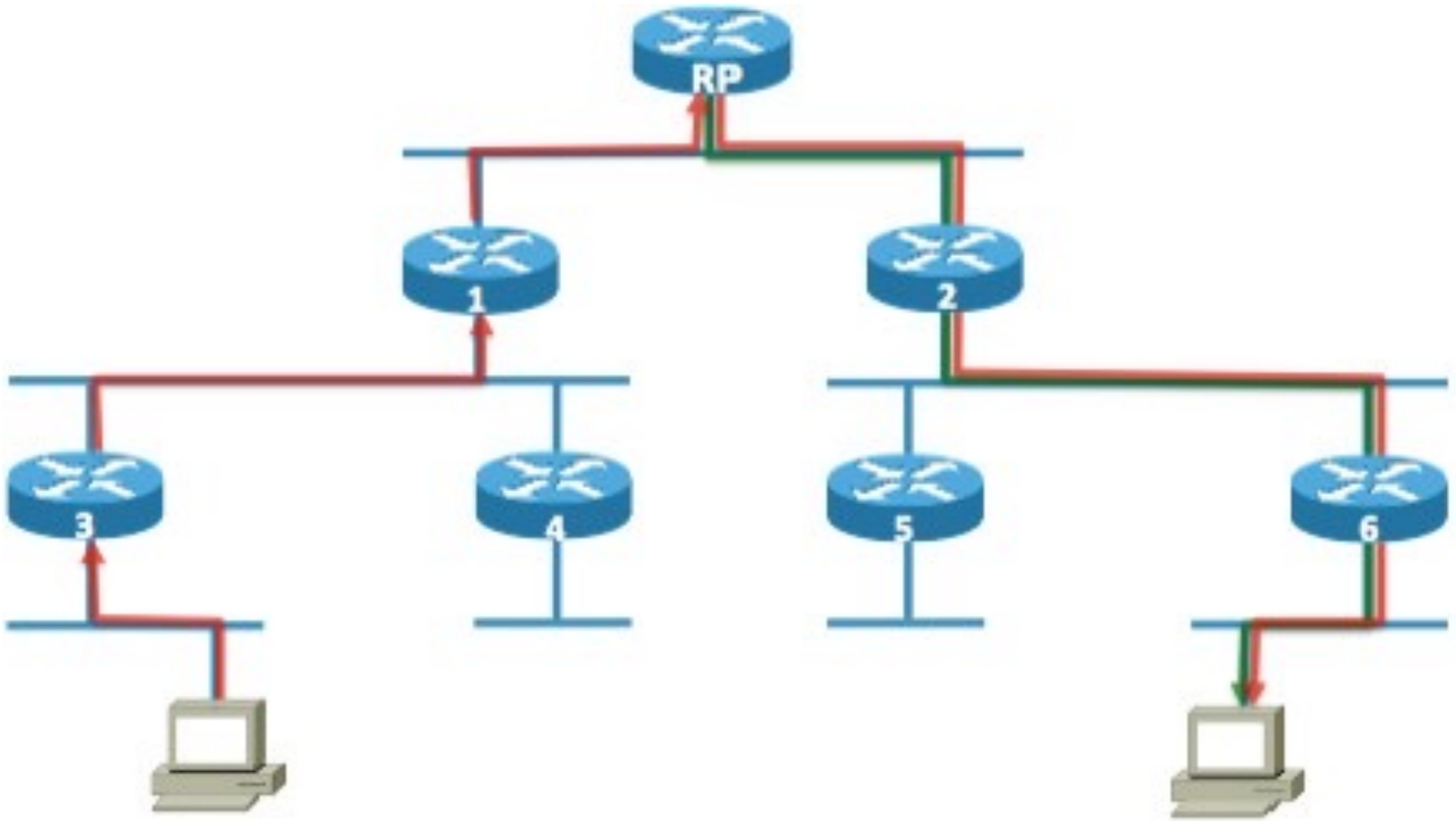
https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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).

https://t.me/learningnets




❖Arriving Causes Router and RP to Create (G) State

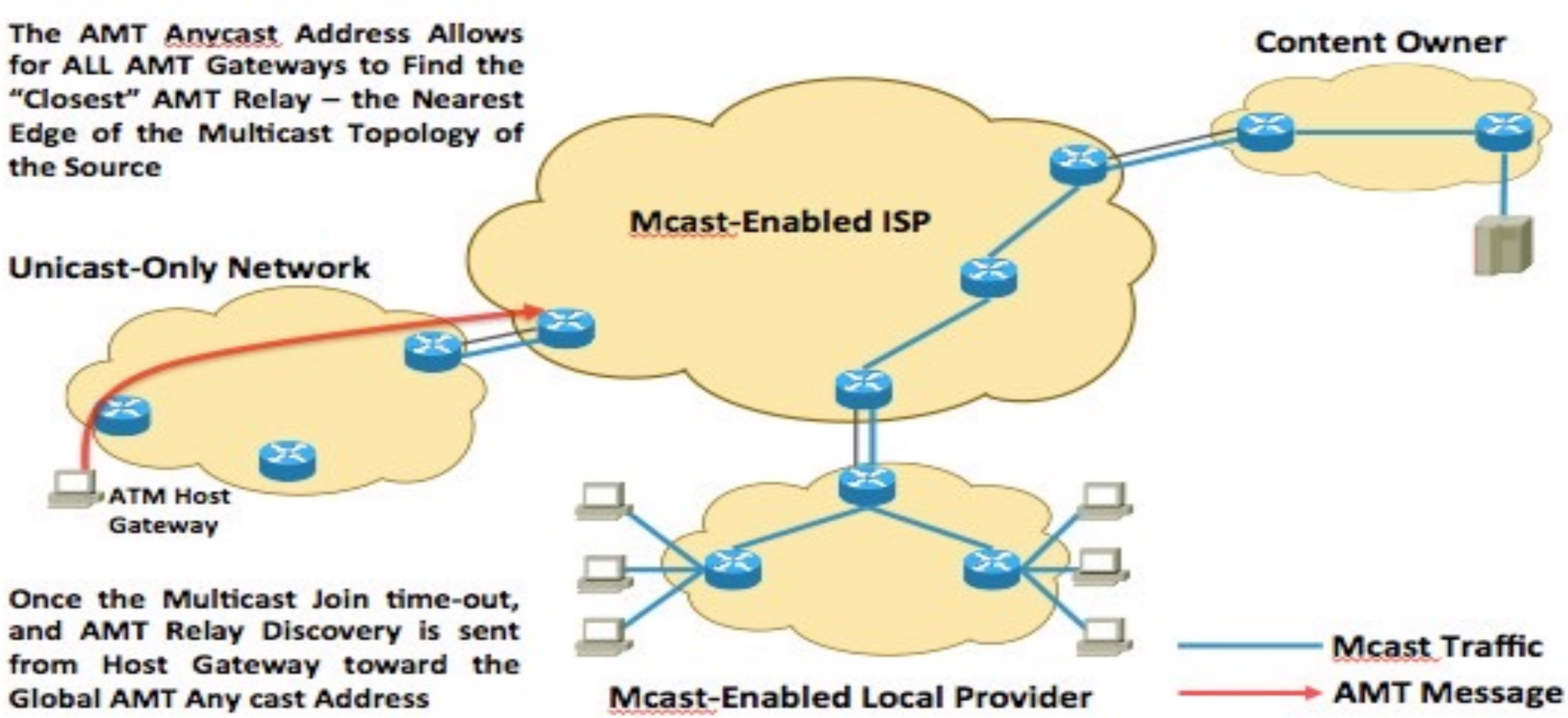
https://t.me/learningnets

## Multicast Case Study

- Terrano is a European based manufacturer company.
- Users of Terrano wants to watch the stream from the content provider which has a peering to Terrano's Service Provider.
- But the problem is Terrano doesn't have Multicast in the network.

https://t.me/learningnets

**What solution would Terrano use without enabling IP multicast in its network to improve user satisfaction by allowing Multicast stream from the Content provider?**




https://t.me/learningnets

- 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.

https://t.me/learningnets

- Terrano can receive multicast content at this point. Because end to end tree is built.

- AMT Host gateway feature is implemented on the receiver PC.

- Ideally  AMT (Automatic Multicast Tunneling) Host Gateway feature should be provided by the first hop routers of Terrano

https://t.me/learningnets

# Multicast – Study Resources

❖ **Books** :

❖ http://www.amazon.com/Developing-IP-Multicast-Networks-I/dp/1578700779/ref=sr_1_1?ie=UTF8&qid=1436564509&sr=8-1&keywords=ip+multicast

❖ **Videos** :

❖ Ciscolive Session – BRKIPM – 1261 – Speaker Beau Williamson

❖ **Podcast** :

❖ http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/multicast-enterprise/whitepaper_c11- 474791.html

❖ http://packetpushers.net/community-show-multicast-design-deployment-considerations-beau-williamson- orhan-ergun/

https://t.me/learningnets

❖ **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

https://t.me/learningnets

# IP and MLS

# Quality of Service

https://t.me/learningnets

# Agenda

- What is Quality of Service
- Different QoS Approaches
- Quality of Service Toolset
- Congestion Management Tools
  1. Classification and Marking
  2. Policing and Shaping
  3. Queuing and Different Queuing Techniques
- Congestion Avoidance Tools – RED, WRED, AQM
- Buffer Bloat
- QoS Design Best Practices
- QoS Design Strategies
- QoS in Different Places in Networks
  1. QoS in Wide Area Networks
  2. QoS in Campus Networks
  3. QoS in SP Networks
  4. QoS in MPLS VPN Networks
  5. QoS in DMVPN and GETVPN Networks
- Diffserv Aware MPLS-TE Quality of Service
- QoS Frequently Asked Questions
- QoS Summary

https://t.me/learningnets

# 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

- QoS is the ability of the network to give a higher preference to one piece of data over another

https://t.me/learningnets

# Quality of Service (QoS)

- Providing Network resources to satisfy application demands

- Managing fairness!

https://t.me/learningnets

# Different QoS Approaches

**<u>Two Quality Of Service approaches have been defined by the standard organizations:</u>**

- Intserv (Integrated Services) and Diffserv (Differentiated Services)

https://t.me/learningnets

# Different QoS Approaches

- Intserv requires each and every flow to request a bandwidth from the network and network would reserve the required bandwidth for the user during a conversation

https://t.me/learningnets

# Different QoS Approaches

- 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

https://t.me/learningnets

# Different QoS Approaches

- 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

https://t.me/learningnets

# Different QoS Approaches

- It is scalable approach compare to the Intserv Quality of Service model

https://t.me/learningnets

# Different QoS Approaches

| Design Requirement | Integrated Services | Diffentiated Services |
|---|---|---|
| What is it? | QoS Architecture that specifies the elements to guarantee QoS on the Networks | QoS Architecture that specifies a simple and scalable mechanim for managing network traffic to provide QoS on the Networks |
| Scalablability | Low, each flow requires reservation on each and every nework hop | Good scalability as the flows are aggregated in an application classes and different application classes get different treatment, not per flow |
| Signalling | RSVP | Not Applicable, No reservation/no state entry |
| Known as | It is known as Intserv and Hard QoS | It is known as Diffserv and Soft QoS |
| Widely Deployed | No, not deployed | Yes, widely deployed |
| Complexity | Complex since it requires randevous point, Anycast RP for the Randevous point redundancy, Randevous point engineering for the optimal multicast routing | Easy, it requires source information only. There is no Randevous Point in Source Specific Multicast, no RP Engineering , no Anycast RP |
| Resource Requirement | Too much, each flow requires bandwidth reservation | Resource reservation is not applicable |

https://t.me/learningnets

# QoS Toolset

- For Diffserv purpose, there are many tools to categorize the traffic, also traffic may need to be prioritize or punish in case of congestion, based on business and application requirements

https://t.me/learningnets

# QoS Toolset – Classification and Marking

- First action that we have to perform in overall QoS design, is to identify/categorize one packet or packet flow is different from another, this is called classification

- An action that organizes packets into different traffic types, to which different policies can then be applied

- Classification of packets can happen without marking

https://t.me/learningnets

# QoS Toolset – Classification and Marking

- Writes a value into the packet header

- When it is done once, you don’t need to do the Deep Packet inspection at every hop again, so you can reduce te performance impact

https://t.me/learningnets

# QoS Toolset – Classification and Marking

## Classification can be done on:

- Layer 1 – such as ingress physical or sub-interface
- Layer 2 – IEEE 802.1Q/p COS bits
- Layer 3 – TOS Byte/DSCP
- Layer 4 – TCP/UDP Ports
- Layer 7 – Application Based, for example using NBAR

https://t.me/learningnets

# QoS Toolset – Classification and Marking

- Layer 1 classification is not scalable, it is hard to manage

- Layer 2 classification and marking may not be end to end, if packet travels from Ethernet to Non-Ethernet, marking values get lost

- Layer 3 classification and marking is end to end, thus recommended option

https://t.me/learningnets

# QoS Toolset – Classification and Marking

- If more granular classification and marking is necessary, Layer 4 and Layer 7 might be needed

- Layer 7 would be necessary if there are applications tunnelled through HTTP, so they look like same, using port 80, but you may need to threat the applications differently

https://t.me/learningnets

# QoS Toolset – Classification and Marking

## Marking can be done on:

- Layer 2 – IEEE 802.1Q/p COS
- Layer 2.5 – MPLS EXP
- Layer 3 – IP DSCP
- Internal in Router – QoS Group

https://t.me/learningnets

# QoS Toolset – Policing and Shaping

- QoS policing and shaping mechanisms are used to identify traffic violation and make responses. Policing and shaping adopt the same algorithms for identifying traffic violation, but they make different responses.

https://t.me/learningnets

# QoS Toolset – Policing and Shaping

- The policing mechanism checks traffic in real time, and takes immediate actions according to the settings when it discovers violation

- For example, the policing mechanism can identify if the traffic payload exceeds the defined traffic flow rate, and then decide to re-mark or drop the excessive parts

- It can control the traffic of both inbound and outbound directions

https://t.me/learningnets

# QoS Toolset – Policing and Shaping

- The shaping mechanism works together with queuing mechanism

https://t.me/learningnets

# QoS Toolset – Policing and Shaping

- Policing perform checks for traffic violations agains a configured rate and take immediate actions – such as dropping or remarking

- Policing don’t delay the traffic

- Policing can be done at Data plane and control plane

https://t.me/learningnets

# QoS Toolset – Policing and Shaping

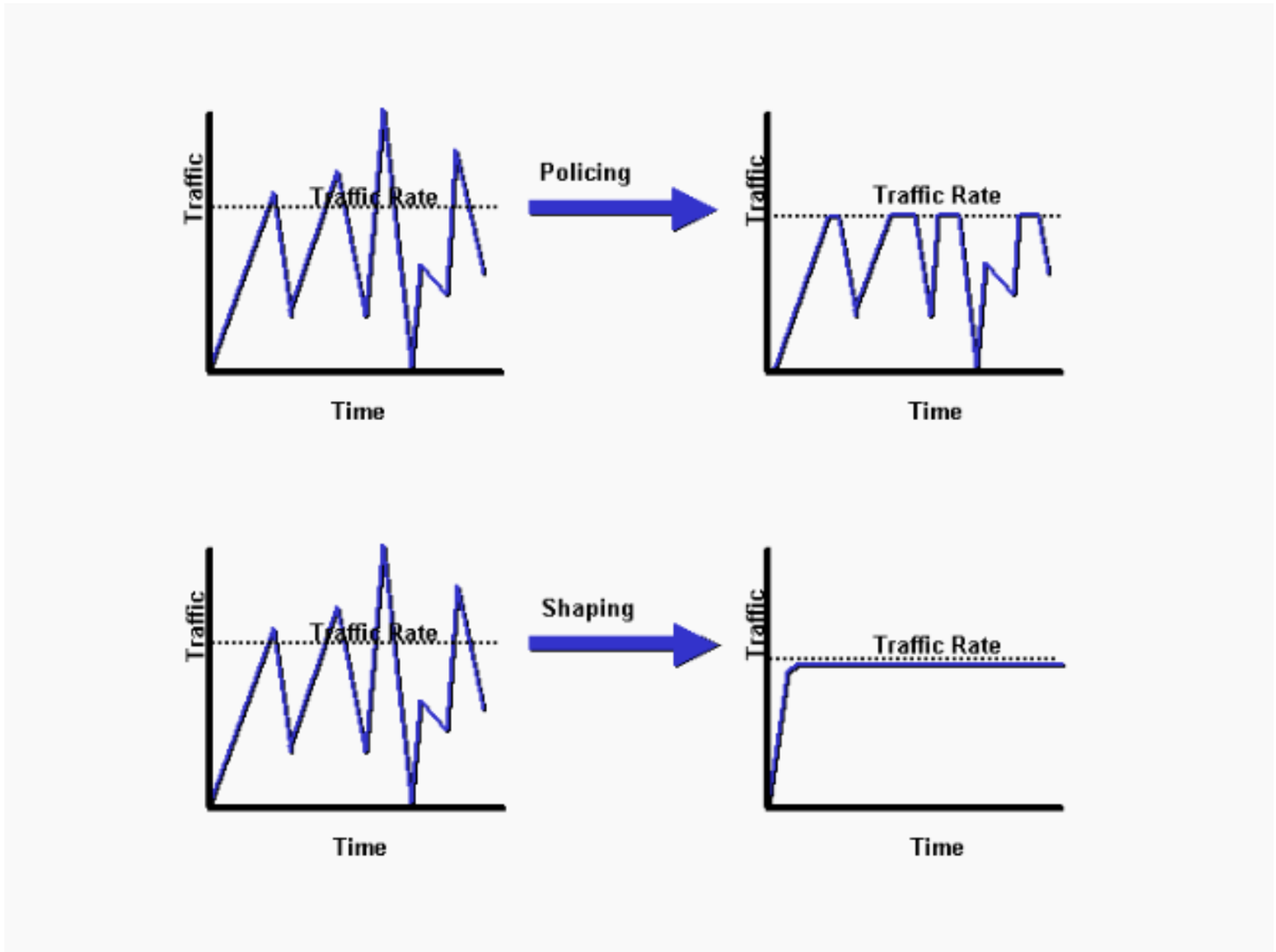
- Shaping smoot out the bursty traffic for helping infrastructure to deal with them more eaily

- Smoot out the peaks of traffic arrival, so that it never exceeds the configured rate

- If the offered traffic momentarily spikes above the contracted rate, the excess traffic is buffered and delayed, until the offered traffic once agains goes below the configured rate

https://t.me/learningnets

# QoS Toolset – Policing and Shaping




https://t.me/learningnets

# QoS Best Practices

- Always classify and mark applications as close their source as possible

https://t.me/learningnets

# QoS Best Practices

- 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

- Also, there is no standard for COS and MPLS EXP, but there is for IP DSCP (RFC 4594), thus when deploying the policy based on DSCP, we can have a better chance to have end to end consistent policy

https://t.me/learningnets

# QoS Best Practices

- Implement QoS always at the hardware if it is possible to avoid performance impact

- Switches support QoS in the the hardware, so for example in the Campus and Datacenter, classify and mark the traffic at the switches

https://t.me/learningnets

# QoS Best Practices

- Not just classification, but policing also should be done to avoid performance impact of software QoS

- Police unwanted traffic flows as close to their sources as possible

https://t.me/learningnets

- If traffic exceed the CIR or PIR and if it will be markdown, follows standard-based marking values. For example if  the application class for the conforming traffic is AF31 , exceeding traffic should be markdown with AF32 and violating traffic should be markdown with AF33.

- Enable queuing policies at every node where the potential for congestion exist.

https://t.me/learningnets

- QoS design should support minimum 3 classes; EF (Expedited Forwarding), DF (Default Forwarding/Best Effort) and AF(Assured Forwarding)

- If company policy allows YouTube, gaming and other non-business applications, Scavenger class is created and CS1 PHB is implemented. CS1 is defined as less than best effort service in the standard RFC.

https://t.me/learningnets

- On AF queues, DSCP-based WRED should be enabled. Otherwise TCP synchronization occurs. WRED allows the packet to be dropped randomly and DSCP functionality provides packet to be dropped based on their priority.

https://t.me/learningnets

# RFC 4594

| Application Class | Per-Hop Behavior | Admission Control | Queuing & Dropping | Application Examples |
|---|---|---|---|---|
| VoIP Telephony | EF | Required | Priority Queue (PQ) | Cisco IP Phones (G.711, G.729) |
| Broadcast Video | CS5 | Required | (Optional) PQ | Cisco IP Video Surveillance / Cisco Enterprise |
| Realtime Interactive | CS4 | Required | (Optional) PQ | Cisco TelePresence |
| timedia Conferencing | AF4 | Required | BW Queue + DSCP WRED | Cisco Jabber, Cisco WebEx |
| ultimedia Streaming | AF3 | Recommended | BW Queue + DSCP WRED | Cisco Digital Media System (VoDs) |
| Network Control | CS6 | | BW Queue | EIGRP, OSPF, BGP, HSRP, IKE |
| Signaling | CS3 | | BW Queue | SCCP, SIP, H.323 |
| / Admin / Mgmt (OAM) | CS2 | | BW Queue | SNMP, SSH, Syslog |
| Transactional Data | AF2 | | BW Queue + DSCP WRED | ERP Apps, CRM Apps, Database Apps |
| Bulk Data | AF1 | | BW Queue + DSCP WRED | E-mail, FTP, Backup Apps, Content Distributio |
| Best Effort | DF | | Default Queue + RED | Default Class |
| Scavenger | CS1 | | Min BW Queue (Deferential) | YouTube, iTunes, BitTorent, Xbox Live |

https://t.me/learningnets

**Voice QoS Requirements**

- Voice traffic should be marked to DSCP EF per the QoS Baseline and RFC 3246 , 4594
- 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.

https://t.me/learningnets

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

https://t.me/learningnets

## Video Best Practices:

- Streaming Video (whether unicast or multicast) should be marked to DSCP CS4 as designated by the QoS Baseline.

- Loss should be no more than 5 %.

- Latency should be no more than 4–5 seconds (depending on video application buffering capabilities).

- There are no significant jitter requirements.

- Guaranteed bandwidth (CBWFQ) requirements depend on the encoding format and rate of the video stream.

- Streaming video is typically unidirectional and, therefore, Branch routers may

- not require provisioning for Streaming Video traffic on their WAN/VPN edges (in the direction of Branch-to-Campus).

https://t.me/learningnets

# ➢Data Applications QoS Requirements

- ❖Best Effort Data
- ❖Bulk Data
- ❖Transactional/Interactive Data

https://t.me/learningnets

# 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.

https://t.me/learningnets

## 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

https://t.me/learningnets

- 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.

https://t.me/learningnets

# 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.

https://t.me/learningnets

- 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.

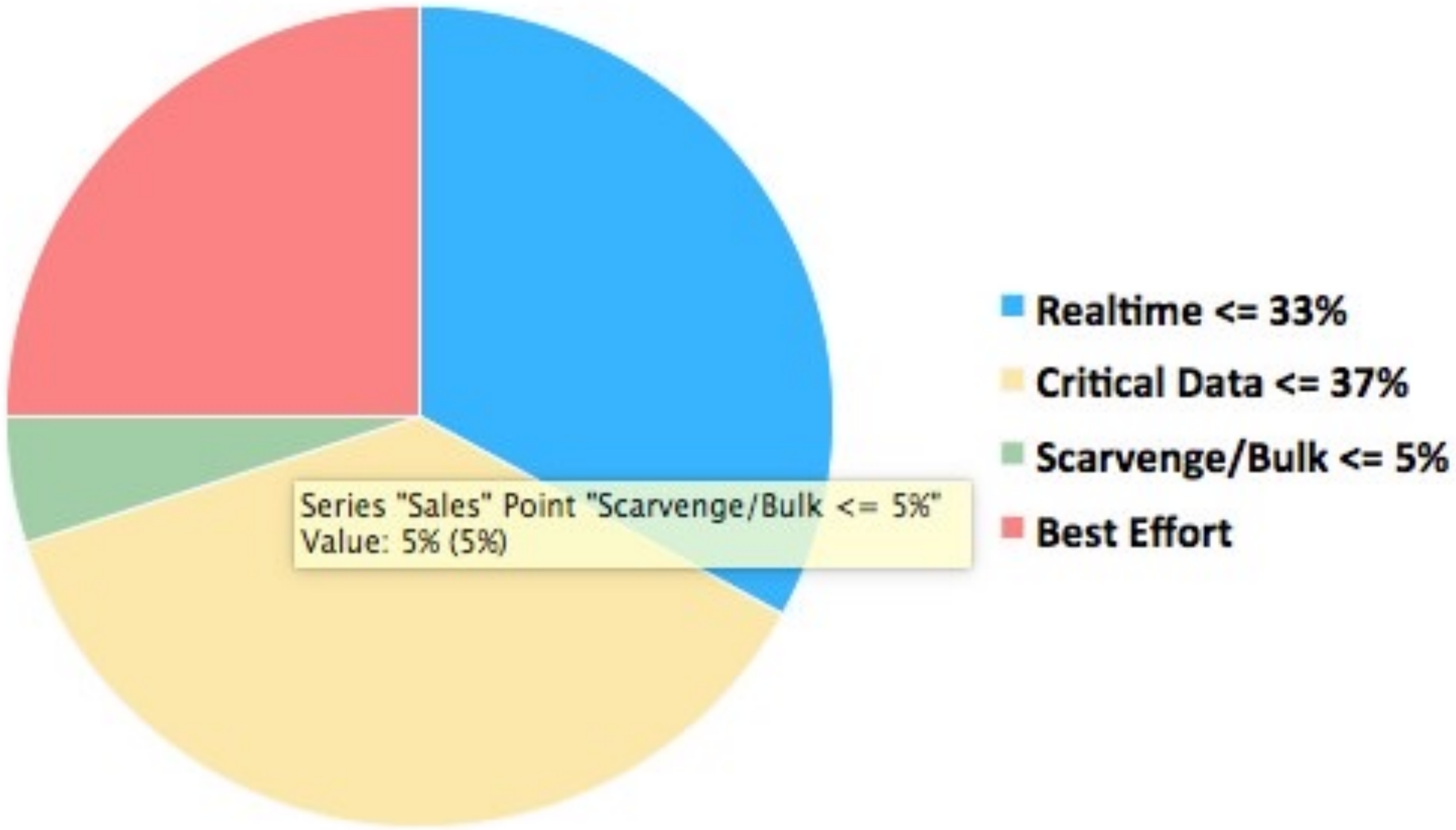
- Transaction is a foreground operation; the user waits for the operation to complete before proceeding.

https://t.me/learningnets

- 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).

https://t.me/learningnets

- 


https://t.me/learningnets

# QOS Study Resources

- **Books :**

- http://www.amazon.com/End---End-QoS-Network- Design/dp/1587143690/ref=sr_1_1?ie=UTF8&qid=1436564258&sr=8-1&keywords=end+to+end+qos+network+design

- **Videos :**

- **Ciscolive Session – BRKCRS -2501**

- **https://www.youtube.com/watch?v=6UJZBeK_JCs**

https://t.me/learningnets

## ❖ 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

https://t.me/learningnets

IPv6

# Internet Protocol Version 6

https://t.me/learningnets

# IPv6  Module Outline

Business Drivers – Why IPv6

IPv6 Addressing
IPv6 Address Plan

IPv4 and IPv6 Transition Mechanisms

1. Dual Stack
2. Tunneling Mechanisms
3. Translation

IPv6 Dual Stack

- Challenges with Dual Stack
- Requirements with Dual Stack
- Happy Eyeballs

https://t.me/learningnets

# IPv6  Module Outline

IPv6 Tunneling

Manuel Tunnels/Configured Tunnels
 IP in IP and GRE Tunnels

Semi Automatic Tunnels
- Tunnel Brokers
- Elements of a Tunnel Brokers – TB/TS
- Hurricane Electric Tunnel Broker Service

Automatic Tunnels
1. 6 to 4
2. 6rd – IPv6 Rapid Deployment
3. ISATAP
4. TEREDO
5. DS-Lite – Dual Stack Lite

https://t.me/learningnets

# IPv6  Module Outline

IPv6 Routing Protocols

1. OSPFv3 and comparison with OSPFv2
2. IS-IS and comparison with OSPF
3. EIGRP

IPv6 in BGP

https://t.me/learningnets

# IPv6 Business Drivers

- **IPv4 address exhaustion**
- **Business Continuity (E-Commerce, Content Providers, Content Delivery Networks)**
- **Easier Network Mergers and Acquisitions  (No overlap, NAT etc.)**

https://t.me/learningnets

## IPv6 Addressing

- 24 September 2015 ARIN said “There is no more IPv4!” As of 2021, none of the five RIRs has IPv4

- IPv7 , IPv8 and IPv9 are already been deprecated. IPv10 is known as NDN – Named Data Networking – Van Jacobsen

- In IPv6, number of hosts is not important as there is 64 bits for the host portion, number of network which needs to be aggregated is an important consideration

https://t.me/learningnets

# IPv6 Addressing

- IPv6 addresses are 128 bits long - 32 hexadecimal characters
- Hexadecimal is widely used in computing
- Hex is a base 16 numerical system
- Every Hex character is knows as **Nibble**
- Total 8 groups , groups are known as words or quads
- Each group is 16 bits and separated by “:”

| Binary | Hex | Decimal |
|---|---|---|
| 0000 | 0 | 0 |
| 0001 | 1 | 1 |
| 0010 | 2 | 2 |
| 0011 | 3 | 3 |
| 0100 | 4 | 4 |
| 0101 | 5 | 5 |
| 0110 | 6 | 6 |
| 0111 | 7 | 7 |
| 1000 | 8 | 8 |
| 1001 | 9 | 9 |
| 1010 | A | 10 |
| 1011 | B | 11 |
| 1100 | C | 12 |
| 1101 | D | 13 |
| 1110 | E | 14 |
| 1111 | F | 15 |

https://t.me/learningnets

# IPv6 Addressing

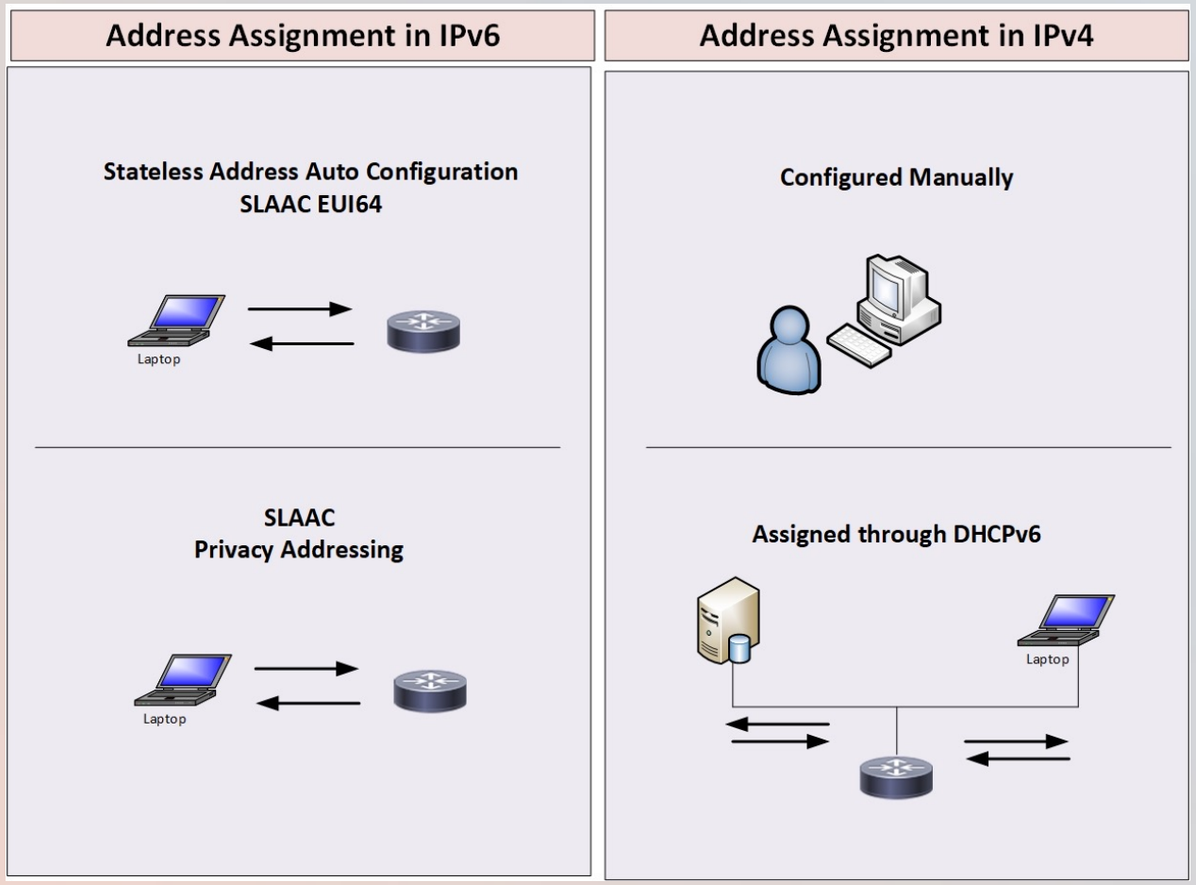
- 128 bit address has two portion. Network and Host
- Network (Sometimes called as Topology portion) is 64 bits
- Host Portion is 64 bits as well

2001:0db8:1010:aaaa:0000:0000:0000:0001

https://t.me/learningnets

# IPv6 Address Assignment to the Hosts

- IPv6 Hosts can have an address manually or automatically similar to IPv4, there are some differences though




https://t.me/learningnets

## IPv6 Address Assignment to the Hosts

- Although there are two methods in IPv4 for obtaining an IP address, IPv6 has three methods

- Static configuration is basically the same in both protocols, although less relevant for IPv6, because of the length of the address

- DHCP is also there for both protocols, and IPv6 DHCPv6 is described in RFC 3315

https://t.me/learningnets

# IPv6 Address Assignment – SLAAC

- The new method that IPv6 introduces is called Stateless Address Autoconfiguration(SLAAC), and described in RFC 4862

- SLAAC works by combining part of the address from an interface's gateway, learned via Router Advertisements(RAs), and an interface's layer 2 address with "ff:fe" adding in the middle of it

- So, router announces it's own prefixes, host take prefix part and interface ID part is generated by using MAC address of the host (Different with Privacy Extensions, it will be explained)

https://t.me/learningnets

# IPv6 Address Assignment – SLAAC

- In practice, this generally means that using SLAAC an interface's address will be composed of; the first 64 bits of its gateway's address, plus the higher 24 bits of its Ethernet MAC, plus 16 hardcoded bits of "ff:fe", plus the lower 24 bits of its MAC address

- IPv6 addresses are 128 bits, and 64 + 24 + 16 + 24 = 128

https://t.me/learningnets

# IPv6 Address Assignment – SLAAC

- As an example, let's say an interface's gateway had the address 2001:db8::1, and the interface's MAC address was 01:23:45:67:89:ab

This would result in an IPv6 address of 2001:db8::123:45ff:fe67:89ab

- It's easier to understand if you write it out in long form :

2001:db80:0000:0000:0000:0000:0000:0001 Gateway 0123:4567:89ab MAC 0123:45ff:fe67:89ab MAC after padding

2001:db80:0000:0000:0123:45ff:fe67:89ab Final IP address

https://t.me/learningnets

# IPv6 Address Assignment – SLAAC Privacy Extension

- If your operator is using SLAAC then you can be tracked using the last 48 bits of your IPv6 address because it’s unique

- Websites that you visit can see that you have a new IP address, because you have roamed to a new operator, they can also see that the last 48 bits of the address stay the same every time.

- Hence, every website you visit will know your device regardless of what network you’re on. You can be easily tracked across providers

https://t.me/learningnets

# IPv6 Address Assignment – SLAAC Privacy Extension

- This problem was resolved by issuing an update to the SLAAC protocol called “Privacy Extensions for Stateless Address Autoconfiguration in IPv6”, which is defined in RFC 4941

- Basically, your computer’s Ethernet interface no longer uses its MAC to fill in the last 48 of its IPv6 address. Instead it picks a series of bits randomly, and fills in the last 48 bits with the random bits

https://t.me/learningnets

# IPv6 Address Assignment – SLAAC Privacy Extension

- But sometimes tracking might be required internally in the inside network , for example for troubleshooting or logging purposes

- Thus the recommendation for the Privacy extension is, use it for the external communication but not for internal network

https://t.me/learningnets

## IPv6 Address Assignment – When to use DHCPv6?

- SLAAC can be used in production environment, but many organizations don't like unpredictability, particularly with Privacy Extension

- DHCPv6 can be used when there is a requirement for predictability

https://t.me/learningnets

# IPv6 Address Assignment – When to use DHCPv6?

- In fact, different place in the network have different requirement

- SLAAC can be used for the customer address assignment in the SP but same organization generally use Static address assignment or DHCPv6 for servers in their Datacenter

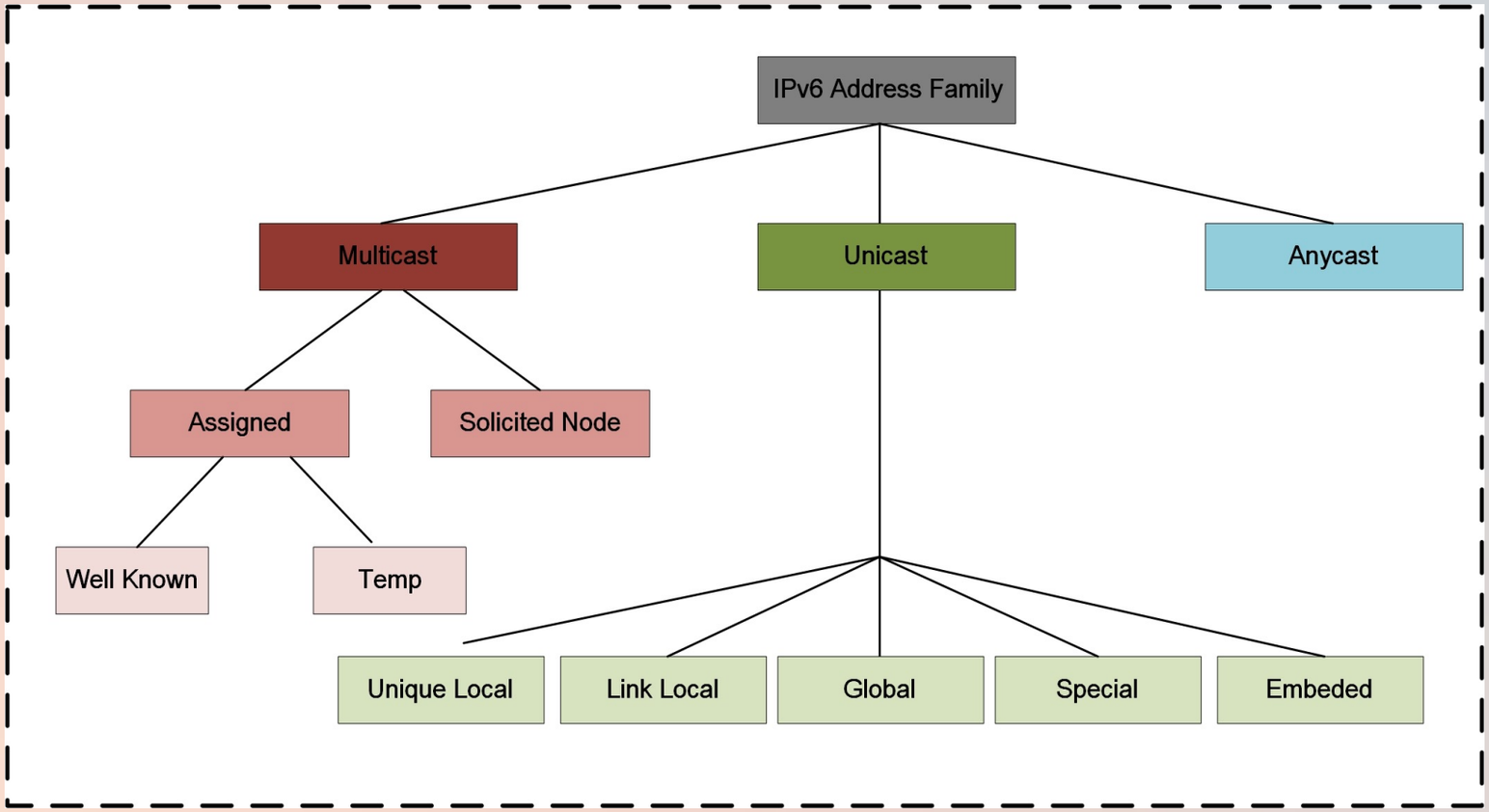
https://t.me/learningnets

# IPv6 Address Assignment – When to use DHCPv6?

If DHCPv6 will be used, network needs to stop announcing SLAAC prefixes and M bit needs to be set in the Router Advertisement messages (RA)

https://t.me/learningnets

# IPv6 Addressing




**There is no Broadcast Addressing in IPv6**

https://t.me/learningnets

# IPv6 Addressing

- With IPv4, host portion of the IP address is used to create subnet, this is not the case with IPv6

- Every IPv6 interface has Link Local IPv6 address (fe80::/10). All interfaces on specific router can have identical LLA (Link Local Address) Ex : FE80::10.10.10.10

- If router will talk to outside world, then interface should have at least two addresses, link local and GUA (Global Unicast Address)

https://t.me/learningnets

# Unicast IPv6 Address Types

- Three types of Unicast IPv6 Address Types:

1. Link-Local – Non-routable , exists on single layer 2 domain

fe80::/10

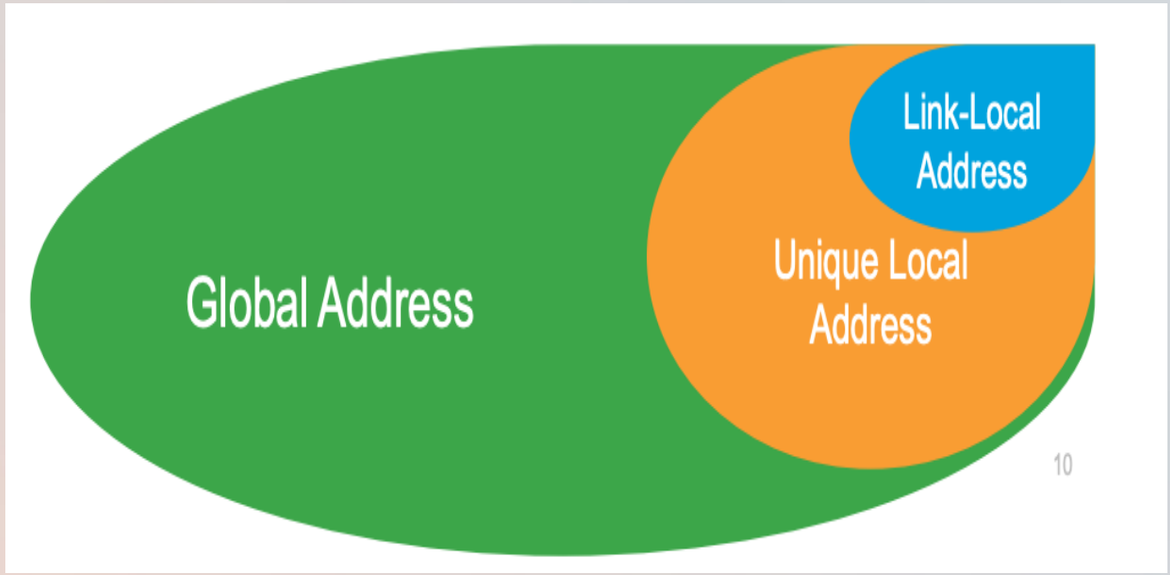
2. Unique-Local : Routable within AS

fc00::/7

3. Global Unicast Address – Routable across Inter domain

2000::/3

(Range is between 2000 – 3fff – First 2 byte)




https://t.me/learningnets

# Unicast IPv6 Address Types

- In IPv6, every bit position (called as Nibble), can have 16 different option, 0 to F

- Global unicast address for example is assigned as 2000::/3 mean, 2000 – 3fff for the first 16 bits, which mean only 2 bits are used of the first nibble

- Thus, only 1/8 of the entire IPv6 address space is currently allocated by IANA

https://t.me/learningnets

# IPv6 Address Recommendations

- First 64 bits are fixed of the Link Local Address

- Interface identifier/Host Portion can be modified, Vlan number, IPv4 addresses can be encoded into IPv6 Link Local Address for easier troubleshooting

- IGP routing can run by just using Link Local Address, but in this case management of those interfaces cannot be possible since Link Local Address is not routable

https://t.me/learningnets

# IPv6 Address Recommendations

- ULA – Unique Local Addressing is not routable in DFZ but inside the AS, similar to RFC 1918 address space

- General recommendation is to not use ULA

- If it is used, for the IPv6 Internet destinations, local network needs an NPT (Network Prefix Translation – from ULA to Global Unicast Address)

https://t.me/learningnets

# IPv6 Address Recommendations

- General recommendation to assign IP address for internal network is Global Unicast Address

- There are vast number of prefixes in Global Unicast Address space

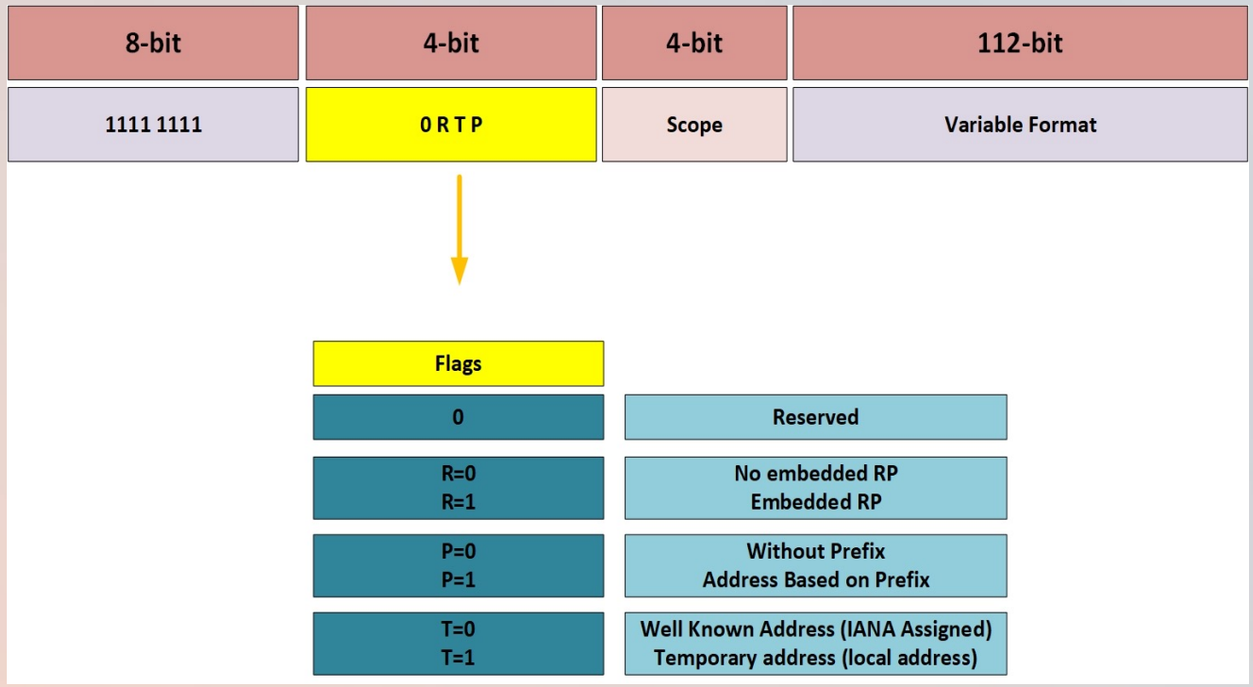
- No need to manage NAT etc, just one address space for internal and external purposes

https://t.me/learningnets

# Multicast IPv6 Address

- IPv6 Multicast Address was defined in the RFC  4291 (It specifies Solicited-Node Multicast Address as well, you will see next)

- IPv6 Multicast uses ff00::/8 range




https://t.me/learningnets

# Solicited-Node Multicast Address

- Every Unicast address must build corresponding Solicited-Node Multicast Address

- Solicited-Node Multicast provides the functionality for Neighbor Discovery in IPv6

- So, it serves the purpose of IPv4 ARP and also used for Duplicate Address Detection

https://t.me/learningnets

# Solicited-Node Multicast Address

- There is no broadcast in IPv6, instead IPv6 uses Multicast

- A device comes up and sends an ICMPv6 neighbor solicitation message to check if anyone else is already using the address it wants to use

- The source for this messages is an unspecified address (::) and the destination address is the solicited node multicast address of the unicast address being checked for duplicates

https://t.me/learningnets

# Solicited-Node Multicast Address

- Second use case of Solicited-Node Multicast Address is to learn Layer 2 address of IPv6 Unicast address

- When a Router attempts to learn a layer 2 address of the remote Router, it sends NS (Neighbor Solicitation) message to the remote router's solicited node multicast address which is generated from the remote router's unicast address

https://t.me/learningnets

# Well Known IPv6 Multicast Addresses

FF02 is Well Known Multicast Address , and the operation with it Is limited to single link.
Routing protocols and link operations use it

| Address | Scope | Meaning |
|---|---|---|
| ff02::1 | Link Local | All Nodes |
| ff02::2 | Link Local | All Routers |
| ff02::5 | Link Local | OSPF v3 Routers |
| ff02::6 | Link Local | OSPF v3 DR Routers |
| ff02::9 | Link Local | RIPng |
| ff02::A | Link Local | EIGRP |

https://t.me/learningnets

# Special Use IPv6 Addresses

- RFC 5156 defines Special Use IPv6 Addresses :

- Default Route ::/0

- Loopback ::1

- 6to4 Auto Tunnel 2002::/16

- Documentation Prefix 2001:0db8::/32 (I used in many examples in this document as well)

- ::/128 Unspecified address (Similar to APIPA address in IPv4) used in Duplicated Address Detection

https://t.me/learningnets

# IPv6 Header – Some fields

- IPv4 Protocol field is replaced with IPv6 Next Header Field
- IPv4 Header is 20 byte, IPv6 Header is 40 byte
- IPv4 TTL is IPv6 Hop Limit
- IPv6 Minimum MTU and Maximum MTU

https://t.me/learningnets

# IPv6 Fragmentation

- In IPv6 Routers don’t perform fragmentation, Routers participate (help) for fragmentation

- Path MTU Discovery is done by the Hosts, not the Routers in IPv6.

- For PMTUD to work, ICMPv6 Type 2 Packet Too Big (PTB) shouldn’t be prevented on the path from receiver to sender

- Otherwise MTU should be set to Minimum IPv6 MTU which is 1280 bytes

https://t.me/learningnets

# IPv6 Fragmentation

- Minimum MTU 1280 means, when IPv6 packet with this size is sent from sender to receiver, none of the device on the path should never fragment the packet

- If ICMP Type 2 PTB reaches from Receiver to Sender, it tells sender what value it should set the packet so packet can be sent between source and the destination

https://t.me/learningnets

# IPv6 Address Planning

https://t.me/learningnets

# IPv6 Address Plan

- We need to analyze the entire network; we need to know each Place In the Network where IPv6 will be deployed

- How many locations , countries, regions , cities , buildings etc.

- Which services, applications and the systems will be connected in each location

- For SP and Enterprises services and the number of locations are different (Fixed networks, mobile networks, datacenters , POP locations , branch offices, campus locations etc.)

https://t.me/learningnets

# IPv6 Address Plan

IPv6 addresses can be received from Upstream ISP or one of the 5 different RIRs (LACNIC, AFRINIC, APNIC, ARIN , RIPE)

https://t.me/learningnets

# IPv6 Address Plan

- Some networks get IPv6 address from one of the 5 RIRs and use it in different regions, some networks get address from each RIR in each region, second approach is usually seen as better for traffic engineering but dealing with different RIRs can create  more operational challenge

https://t.me/learningnets

# IPv6 Address Plan – Getting IPv6 space from RIR

- You can become a member of RIR and get your own allocation – Membership is available to all organizations who are operating a network

- RIRs list their address allocation policies on their website

- Minimum allocation is a /32 for Service Providers, if you need more than that you need to prove that you will have more than 65k /48 assignments

https://t.me/learningnets

# IPv6 Address Plan – Getting IPv6 space from Upstream ISP

- Enterprises usually receive IPv6 address space from their Upstream ISPs

- They receive /48 from upstream

- If they will multihome , then better to receive from their IPv6 address space from RIR directly

https://t.me/learningnets

# IPv6 Address Plan – PI vs PA Discussion

- PI (Provider Independent) or PA (Provider Assigned/Aggregatable)?

- PI space is better for organizations who want to multihome to different SPs

- PA is okay if you are single homed or plan to do NAT for IPv6 in the network which is not recommended

https://t.me/learningnets

# IPv6 Address Plan – PI vs PA Discussion

If Enterprise will terminate their business with one of their multihomed upstream ISPs, PA is challenging as Enterprise need to give their IPv6 address space back and re-address their network , thus many organizations are going down the PI path

https://t.me/learningnets

# IPv6 Address Plan Fundamentals

- In general, we don’t worry about the number of hosts in IPv6 address planning , we have 2 to 64 IPv6 addresses for hosts

- Enterprises usually get /48 as a prefix length or multiple /48 based on their scalability requirements

- Service Providers should get /32 as minimum or multiple of them based on their scale

https://t.me/learningnets

# IPv6 Address Plan Fundamentals

- We need to have a methodology for writing an IPv6 Addressing Plan

- In General there are 5 rules to write an efficient, stable , scalable and secure IPv6 address plan

- Simplicity , Embedding Information , Reserving some addresses, Summarization and Involvement of other teams

https://t.me/learningnets

# IPv6 Address Plan – Simplicity

- IPv6 address plan should be simple, we don’t want to explain our addressing schema days or weeks

- Templates should be used as much as possible

https://t.me/learningnets

# IPv6 Address Plan – Simplicity

- Use nibble boundary, A nibble boundary means : subnetting address space based on the address numbering, each number in IPv6 represents 4 bits = 1 nibble , which means that IPv6 addressing can be done on 4-bit boundaries

https://t.me/learningnets

# IPv6 Address Plan – Simplicity

- Nibble boundary example:

- 2001:cafe:0:10::/61 is :

2001:cafe:0000:0010:0000:0000:0000:0000 to

2001:cafe:0000:0017:ffff:ffff:ffff:ffff

With /61, the address blocks don’t use the entire nibble range , BUT,

https://t.me/learningnets

# IPv6 Address Plan – Simplicity

- Nibble boundary example:

- 2001:cafe:0:10::/60 is :

2001:cafe:0000:0010:0000:0000:0000:0000 to 2001:cafe:0000:001F:ffff:ffff:ffff:ffff

With /60, this subnet uses the entire nibble range (0 to F)

This makes the numbering plan for IPv6 simpler, for example Service Provider can have a particular meaning for their particular POP infrastructure addressing

https://t.me/learningnets

# IPv6 Address Plan – Encoding some information

- Encoding/Embedding some information helps for operation people as well as help for easier troubleshooting

- Location information , country , city , Vlan number can be encoded in Link Local and Global Addresses

https://t.me/learningnets

# IPv6 Address Plan – Reserving some IP Space

- Designing networks for future growth is essential , same is true for IPv6 address plan

- New locations can be added, company can grow organic or inorganically, thus for better summarization , reserving some IP address in advance is clever decision

- You can reserve /64 but assign /126 or /127 for P2P addresses for example

https://t.me/learningnets

# IPv6 Address Plan – Summarizable

- Summarization reduces routing table size
- It reduces number of information to deal with
- It reduces convergence time in case of failure
- It reduces troubleshooting time thus increases high availability of the network
- Thus summarizable addresses will help for better network design

https://t.me/learningnets

# IPv6 Address Plan – Involvement of Other teams

- For example, application and the security teams need to be involved in IPv6 deployment and Address planning

- Based on the IPv6 addresses assigned for each location, security team need to deploy accurate security policies

https://t.me/learningnets

# IPv6 Address Plan – Different PIN Addressing Suggestions

- Different Place in the Network requires special addressing requirements

- P2P links , Loopback addresses , Host/End User Addresses , Buildings , Campus Locations , Services (Broadband, Internet) etc.

https://t.me/learningnets

# IPv6 Address Plan  Example – Different PIN Addressing Suggestions – ISP Network

- ISPs and Enterprise Networks have different requirements as the services, number of sites etc. are different, let's look at ISP Address Plan suggestions based on their PINs and the services

- All ISPs should get minimum /32 from their RIR (Larger ones and more serious ones get /29)

- Loopback Address: Numbering all loopbacks out of one /64 block and assign /128 per loopback

https://t.me/learningnets

# IPv6 Address Plan – Different PIN Addressing Suggestions –ISP Network

## Infrastructure Addressing:

- /48 per region, /48 allows 65k subnets
- /48 for the whole backbone
- General recommendation for the Infrastructure addressing is, summarize between the sites

https://t.me/learningnets

# IPv6 Address Plan – Different PIN Addressing Suggestions –ISP Network

- LAN subnets : /64 per LAN

- P2P links should be assigned either /64 or /127  (RFC 6164)

- You can see Link local addresses are recommended for the P2P infrastructure link by some documents , it is good for having less address (No Global IPv6 address required in this case) but link local address are not routable so NOC cannot reach to these addresses

https://t.me/learningnets

# IPv6 Address Plan – Different PIN Addressing Suggestions –ISP Network

- NOC – Network Operation Center is considered as part of the network infrastructure, usually /60 within the infrastructure /48 block is seen as enough

- Shared/Critical services such as DNS , SMTP , POP3/IMAP etc. also considered as part of the infrastructure and one /64 is enough for these addresses

https://t.me/learningnets

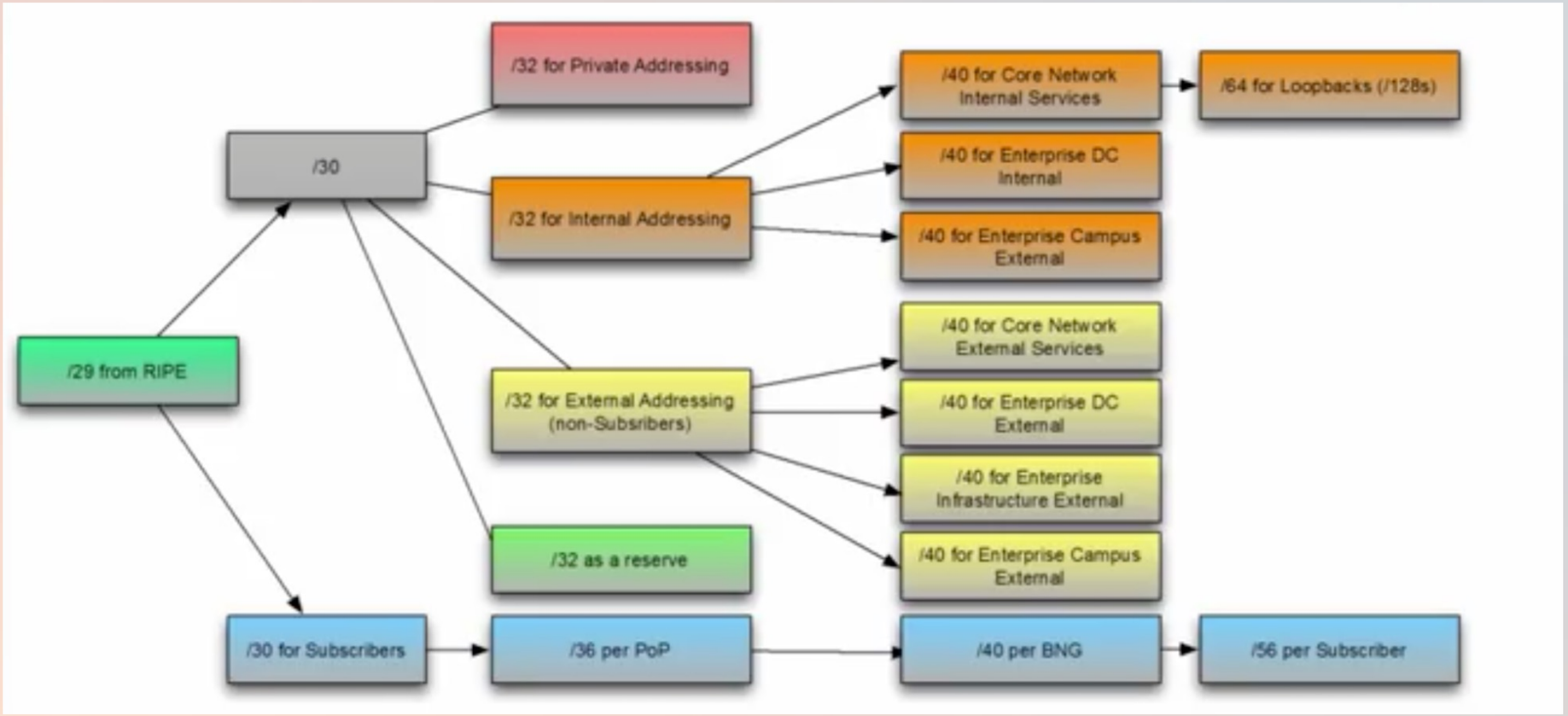
# IPv6 Address Plan – Different PIN Addressing Suggestions –ISP Network

- ISP to Customer Link : /48 should be allocated for this, between the POPs /48 should be divided

- And customer should get one /48 for large scale networks, /56 for small scale and if it is LAN , then /64

https://t.me/learningnets

# ISP IPv4 Address Assignment Example




https://t.me/learningnets

# IPv6 Transition Mechanisms

https://t.me/learningnets

# IPv6 Transition Mechanisms

- Vast majority of the content is working on IPv4 as of 2021. How IPv6 users can connect to the IPv4 world and How IPv4 users can reach to the IPv6 content

- This is accomplished with the IPv6 transition technologies

https://t.me/learningnets

# IPv6 Transition Mechanisms

- Probably the IPv6 transition technologies is a misleading term. Because; IPv4 infrastructure is not removed with these technologies

- Thus, probably the IPv6 integration or co-existence mechanisms are the better terms

- But still throughout this course I will be using IPv6 transition technologies

https://t.me/learningnets

# IPv6 Transition Mechanisms

- If the underlay transport is MPLS; best methods are 6PE and 6VPE

- If the underlay transport is IP; dual stack, tunneling and translation are the options

- Depends on the company, their customer requirements and many other factors, one method might be better than other

https://t.me/learningnets

# IPv6 Transition Mechanisms

There are three types of IPv6 Transition Mthods:

1. Dual Stack
• IPv6 + IPv4

The entire infrastructure is running both IPv4 and IPv6

2. Tunnels
- IPV6 - IPv4 – IPv6
- IPv4 – IPv6 – IPv4

Two IPv6 islands communicate over IPv4 part of the network or two IPv4 islands communicate over IPv6 part of the network

3. Translation
- IPv6 – IPv4 (NAT64)
- Private IPv4 – Public IPv4 (NAT44)

With translation, IPv6 only device can communicate with IPv4 only device. But they think that they communicate with the same version of device

https://t.me/learningnets

# IPv6 Transition Mechanisms – Dual-Stack

- Many people state that IPv6 Dual Stack is the best transition method.

- Is Really Dual Stack best deployment method ?

- Many people would recommend it, before we try to answer this question, let’s understand how Dual-Stack works, what are the advantages and disadvantages, what are the challenges etc.

https://t.me/learningnets

# IPv6 Transition Mechanisms – Dual-Stack

- Dual Stack is Native IPv6 and IPv4 Service, first was defined in RFC 2893

- Having IPv6 and IPv4 at the Hosts, network, operation/support tools, content and the application

- IPv4 applications use the IPv4 stack, and IPv6 applications use the IPv6 stack

https://t.me/learningnets

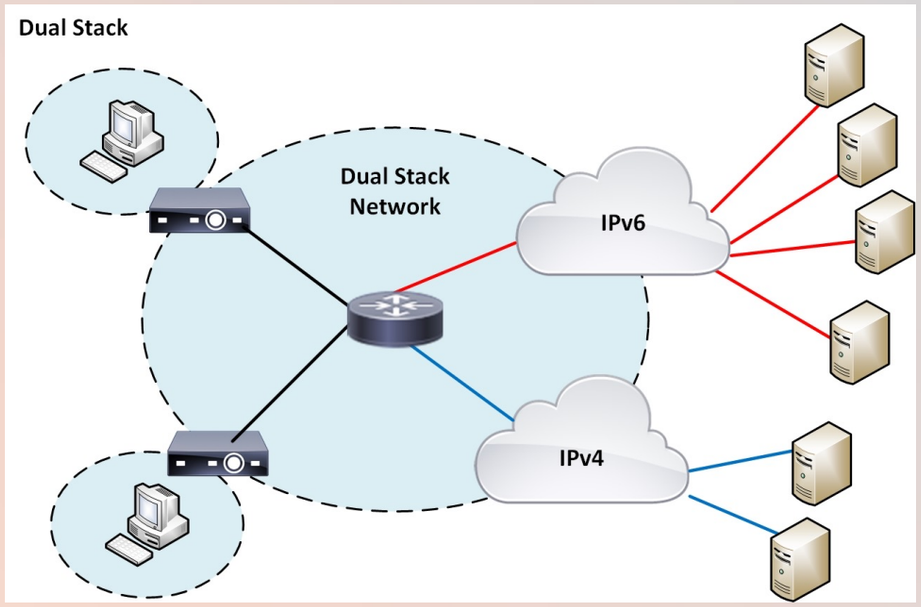
# IPv6 Transition Mechanisms – Dual-Stack

- Routing protocols handle both IPv4 and IPv6

- Since entire network will have both IPv4 and IPv6, when it is needed IPv4 can be removed without causing down time

https://t.me/learningnets

# IPv6 Transition Mechanisms – Dual-Stack




Network, Applications, Services, CPE and Access Networks needs to run Both IPv4 and IPv6

https://t.me/learningnets

# IPv6 Transition Mechanisms – Dual-Stack

- Dual Stack is considered as Simplest solution , without any tunneling and translation mechanism (Most deployments will need translation, we will discuss)

- Every interface speaks both IPv4 and IPv6

- Communication is driven by DNS
  - If destination address in A record, communication is done via IPv4
  - If destination address in AAAA record, communication is done via IPv6
  - If both A and AAAA records are replied by DNS, then IPv6 is preferred

https://t.me/learningnets

# Happy Eyeballs

- Happy Eyeballs has two versions, Version 1 defined in RFC 6555 , version 2 defined in RFC 8305

- Happy Eyeballs tries to prefer IPv6 connections to IPv4 connections, but will use an IPv4 connection if IPv6 isn’t working fast enough

- It is important to select IP address family

- Overall goal with the Happy Eyeballs is to improve end user performance in Dual Stack environment

https://t.me/learningnets

# Happy Eyeballs

- When client receives both AAAA and A response from DNS server, before Happy Eyeballs, browsers preferred IPv6 over IPv4 and tried to setup TCP connection over IPv6

- If the IPv6 routing , IPv6 Server have a problem or simply IPv6 is filtered on the path, browser waits TCP to timeout which can take a minute (Very bad user experience thus users usually turn off IPv6)

https://t.me/learningnets

# Happy Eyeballs

- If there is second IPv6 address in DNS AAAA response, and if it has a problem as well, browser has to wait a minute more

- With Happy Eyeballs, if IPv6 connection somehow has a problem, failing back to IPv4 happens between 200 to 400 ms

https://t.me/learningnets

# Happy Eyeballs

- With Happy Eyeballs, when there is a broken IPv6 path, failing back to IPv4 is much faster, RFC recommends 200 – 400 ms.

- Without Happy Eyeballs, first response when there is a bro

- ken IPv4 connectivity is to turn off IPv6 , which doesn’t help for IPv6 adoption

- Thus, Happy Eyeballs helps for transitioning to IPv6!

https://t.me/learningnets

# Happy Eyeballs

| Device | DNS query sending style | IPv6 broken, time until fallback to IPv4 | | | Comments |
|---|---|---|---|---|---|
| | Dual-stack destination | Black hole | No route | Address unreachable | |
| Symbian^3 on Nokia N8 (11.012) | **A** first and used if possible. **AAAA** if no IPv4. | N/A | N/A | N/A | Symbian^3 prefers IPv4 hence tested fallback scenarios are N/A. The DNS query order is a configuration parameter. |
| Windows 7 Starter Edition on HP *IE 8.0.7600 & Google Chrome 8.0.552.224 & Safari 5.0.2* | **A** and after reply **AAAA**. Uses IPv6 if both available. | ~**21**s | ~**21**s (after 3 SYN & ICMPv6 errors) | ~**21**s (after 3 SYN & ICMPv6 errors) | Same initial delay with those browsers. **The 21 seconds is TCP timeout after 3rd SYN failed.** |
| iOS4 4.2.1 on Apple iPhone4 *Safari* | **A** first and **AAAA** immediately after. Uses IPv6 if both available. | **No fallback** | ~**4**s (After 5 SYN & ICMPv6) | ~**4**s (After 5 SYN & ICMPv6) | Lucky observation: waits ~350 ms for AAAA to arrive after A is received before going for IPv4 |
| Apple OS/X 10.6.6 on iMac *Safari 5.0.3* *Firefox 3.6.13* | **A** first and **AAAA** immediately after. Uses IPv6 if both available. | ~**75**s | ~**4**s (After 5 SYN & ICMPv6) | ~**4**s (After 5 SYN & ICMPv6) **Firefox:** no fallback at all! | Special note that Firefox did not fallback on address unreachable error. |
| Android 2.3.1 on Samsung Nexus S *Native browser* | **AAAA** and after reply **A**. Uses IPv6 if both available. | ~**21**s | ~**0**s (acts on first ICMPv6) | ~**0**s (acts on first ICMPv6) | **The 21 seconds is TCP timeout after 3rd SYN failed.** |
| Maemo5 IPv6 enabled version on Nokia N900 *Firefox & native* | **AAAA** and after reply **A**. Uses IPv6 if both available. | ~**189**s | ~**0**s (acts on first ICMPv6) | ~**0**s (acts on first ICMPv6) | 189s is after 6th SYN failed. Kernel: 2.6.28-based |
| Ubuntu 10.04 /10.10 on "PC" *Firefox 3.6.13* | **AAAA** and after reply **A**. Uses IPv6 if both available. | ~**21**s | ~**0**s (acts on first ICMPv6) | ~**0**s (acts on first ICMPv6) | Note: immediate fallback to IPv4 happens also during complex page load (i.e. minimizes damage when IPv6 is always preferred) Kernel (10.04): 2.6.32-27, (10.10): 2.6.35-24 |



https://t.me/learningnets

# Happy Eyeballs

- Both Operating Systems and the Browsers implement this algorithm

- Microsoft, Linux, Android , Apple IOS etc. comes with Happy Eyeballs

- Apple wrote Happy Eyeballs v2 RFC 8305

https://t.me/learningnets

# Dual-Stack Requirements

- Require sufficient amount of IPv4 addresses (Shouldn’t limit the growth of iPv6 deployment)

- If company doesn’t have an IPv4 address, they may need to purchase an IPv4 address from market

https://t.me/learningnets

# Dual-Stack Requirements

- CPEs should support both IPv4 and IPv6

- Dual Stack also requires hardware and infrastructure which support IPv6

https://t.me/learningnets

# Problems with Dual Stack

- How we can have Dual Stack if we don’t have enough IPv4 addresses?
  - Solution to this is CGN (Thus earlier I said that we will discuss)
- Having Dual Stack requires extra staff training
- Deploying both IPv6 and IPv4 requires extra device resources (RIB and FIB memory and CPU)

https://t.me/learningnets

## Solution to IPv4 depletion with Dual Stack

- Many Service Provider that want to deploy Dual Stack has IPv4 depletion problem

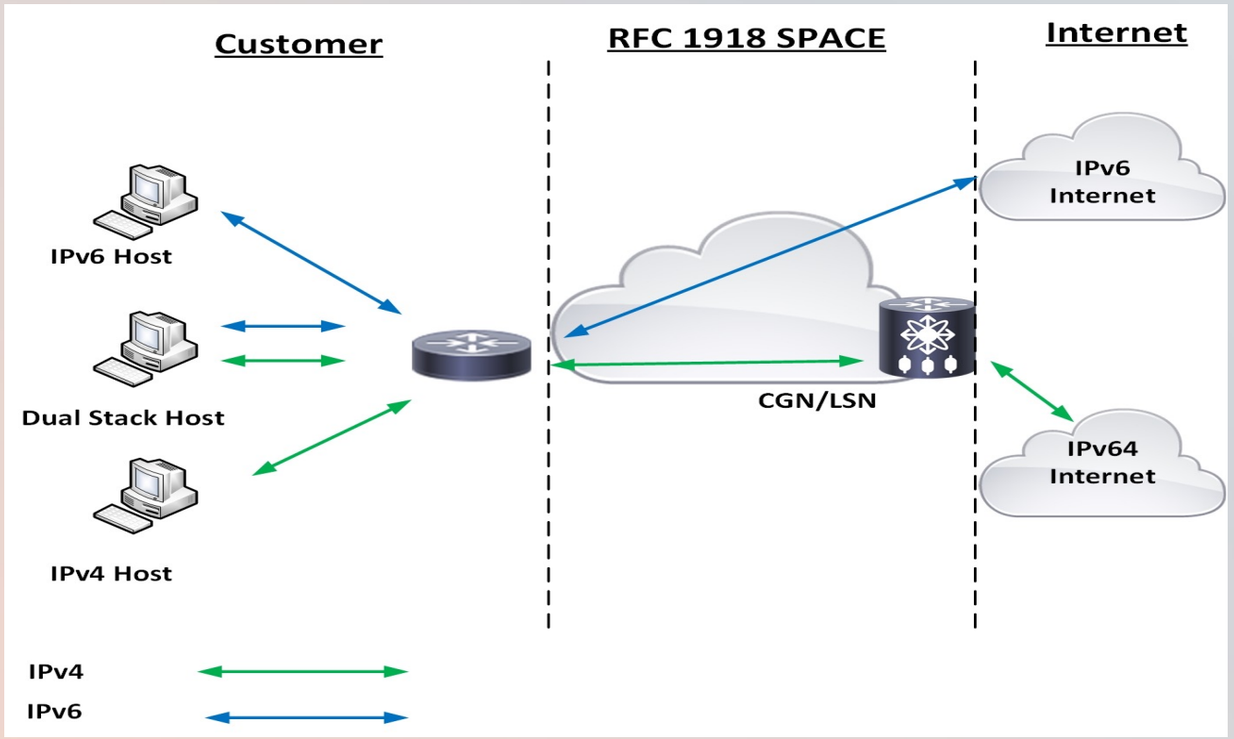
- Thus, in real life we see IPv6 deployment in the network and CGN for IPv4, this is called Dual Stack with SP NAT (CGN) deployment

https://t.me/learningnets

# Solution to IPv4 depletion in Dual Stack Design

- IPv6 is available at the Host side

- But This solution requires CGN since Service Provider doesn't have enough IPv4 Public address and they don't purchase an IPv4 public address from market




**Dual-Stack with CGN**

https://t.me/learningnets

## Solution to IPv4 depletion in Dual Stack Design

- Advantage of this design is company can have IPv6 everywhere

- Another advantage is amount of IPv4 Public address is low since there is NAT

- Disadvantage of this design is there is NAT, so all the possible problems of NAT (Will be discussed in detail) is applicable with this design option

https://t.me/learningnets

# IPv6 Transition Mechanisms – Tunnels

- There are three different type of Tunnels :

  1. Manuel Tunnels
  2. Semi Automatic Tunnels
  3. Automatic Tunnels

- With Tunnels, two IPv6 islands communicate over IPv4 part of the network or two IPv4 islands communicate over IPv6 part of the network
  - IPV6 - IPv4 – IPv6
  - IPv4 – IPv6 – IPv4

https://t.me/learningnets

# IPv6 Transition Mechanisms – Tunnels

- All tunneling mechanisms require that the endpoints of the tunnel run both IPv4 and IPv6 protocol stacks, that is, endpoints must run in dual-stack mode

- The dual-stack routers run both IPv4 and IPv6 protocols simultaneously and thus can interoperate directly with both IPv4 and IPv6 end systems and routers

https://t.me/learningnets

# IPv6 Transition Mechanisms – Tunnels

- It is possible to protect the IPv6 traffic over IPv4 tunnels using IPv4 IPSEC, by applying a crypto map to both the tunnel interface to encrypt outgoing traffic, and to the physical interface to decrypt the traffic flowing

- Protecting tunnels in this way may negatively impact performance

- Last but not least, NAT is not allowed along the path of any tunneling mechanism

https://t.me/learningnets

# Manuel/Configured IPv6 Tunnels

- Manuel Tunnels for IPv6 were explained in RFC 4213 (Basic Transition Mechanisms for IPv6 Hosts). It is known as Configured IPv6 Tunnels

- Tunnel source and the destination manually need to be configured

- A technique for establishing point-to-point tunnels by encapsulating IPv6 packets within IPv4 headers to carry them over IPv4 routing infrastructures.

https://t.me/learningnets

# Manuel/Configured IPv6 Tunnels

- Requires Dual Stack Tunnel End Points, both IPv4 and IPv6 addresses are configured at each end of the Tunnel

- When there are so many sites, Manuel tunnels are not considered scalable

- 6PE and 6VPE are considered as Manuel/Configured Tunnels, both of these concepts will be explained later in detail

https://t.me/learningnets

# Manuel/Configured IPv6 Tunnels

**According to RC 4213, Tunneling can be used in a variety of ways:**

- Router-to-Router. IPv6/IPv4 routers interconnected by an IPv4 infrastructure can tunnel IPv6 packets between themselves. In this case, the tunnel spans one segment of the end-to-end path that the IPv6 packet take

- Host-to-Router. IPv6/IPv4 hosts can tunnel IPv6 packets to an intermediary IPv6/IPv4 router that is reachable via an IPv4 infrastructure. This type of tunnel spans the first segment of the packet's end-to-end path

https://t.me/learningnets

# Manual/Configured IPv6 Tunnels

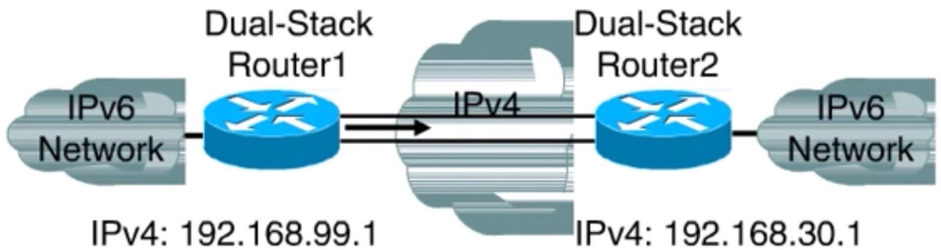
- Host-to-Host. IPv6/IPv4 hosts that are interconnected by an IPv4 infrastructure can tunnel IPv6 packets between themselves. In this case, the tunnel spans the entire end-to-end path that the packet takes

- Router-to-Host. IPv6/IPv4 routers can tunnel IPv6 packets to their final destination IPv6/IPv4 host. This tunnel spans only the last segment of the end-to-end path

https://t.me/learningnets

# Manuel IPv6 Tunnels - How it works

## Manually Configured Tunnel (RFC4213)




```
router1#

interface Tunnel0
 ipv6 address 2001:db8:c18:1::3/64
 tunnel source 192.168.99.1
 tunnel destination 192.168.30.1
 tunnel mode ipv6ip
```

```
router2#

interface Tunnel0
 ipv6 address 2001:db8:c18:1::2/64
 tunnel source 192.168.30.1
 tunnel destination 192.168.99.1
 tunnel mode ipv6ip
```

- Manually Configured tunnels require:
  - Dual stack end points
  - Both IPv4 and IPv6 addresses configured at each end

https://t.me/learningnets

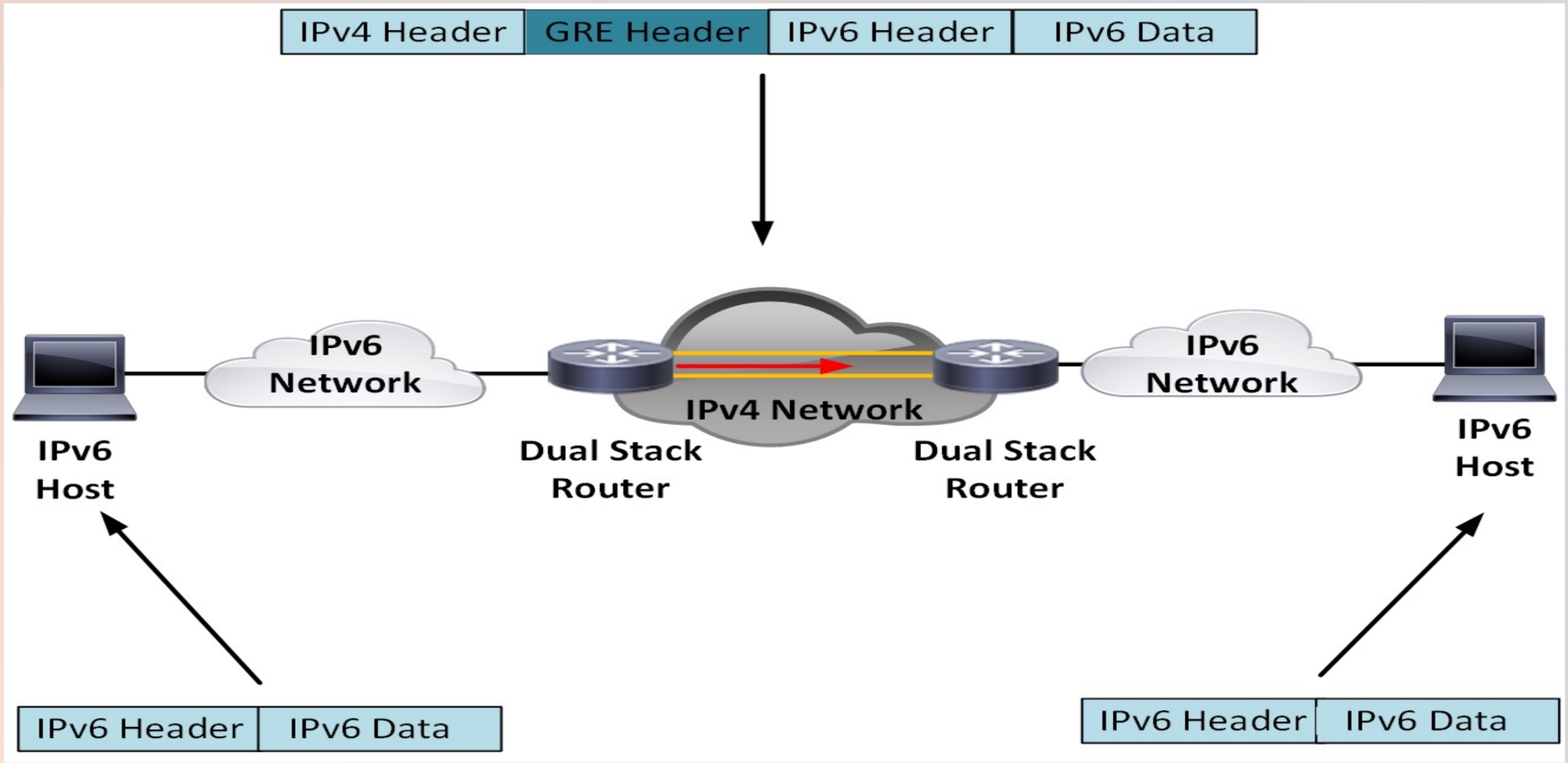
# Manuel IPv6 Tunnels – GRE Tunnel

- GRE Tunnel is another Manuel/Configured Tunneling mechanism which can be used for transporting IPv6 packets over IPv4 infrastructure or vice versa

- The IPv6 over IPv4 GRE tunnel uses the standard GRE tunneling technique that is designed to provide the services necessary to implement any standard point-to-point encapsulation scheme

https://t.me/learningnets

# Manuel IPv6 Tunnels – GRE Tunnel




https://t.me/learningnets

# Semi Automatic IPv6 Tunnels – Tunnel Brokers

- With Semi-Automatic Tunnels, tunnel destination address is automatically learned , not manually configured

- Tunnel Brokers are Semi-Automatic Tunnels, explained in RFC 3053

- At the host side, tunnel can be initiated by the PC or Router which can serve to entire LAN

https://t.me/learningnets

## Semi Automatic IPv6 Tunnels – Tunnel Brokers

- The Tunnel Broker idea is an alternative approach based on the provision of dedicated servers, called Tunnel Brokers, to automatically manage tunnel requests coming from the users

- Tunnel brokers can be seen as virtual IPv6 ISPs, providing IPv6 connectivity to users already connected to the IPv4 Internet

https://t.me/learningnets

# Semi Automatic IPv6 Tunnels – Tunnel Brokers

- The IPv6 Tunnel Broker provides an automatic configuration service for IPv6 over IPv4 tunnels to users connected to the IPv4 Internet

- IPv4 connectivity between the user and the Service Provider is required

- Tunnel Broker can be considered as ‘Authoritative Server’ which provides the IP addresses and the parameters of the actual Tunnel endpoints to the IPv6 end users (Hosts or Routers)

https://t.me/learningnets

# Elements of a Tunnel Broker

**Tunnel Broker (TB)**

- The TB is the place where the user connects to register and activate tunnels. The TB manages tunnel creation, modification and deletion on behalf of the user

- For scalability reasons the tunnel broker can share the load of network side tunnel end-points among several tunnel servers

https://t.me/learningnets

# Elements of a Tunnel Broker

**Tunnel Broker (TB)**

- It sends configuration orders to the relevant tunnel server whenever a tunnel has to be created, modified or deleted

- The TB may also register the user IPv6 address and name in the DNS

https://t.me/learningnets

# Elements of a Tunnel Broker

**Tunnel server (TS)**

- A Tunnel Server is a dual-stack (IPv4 & IPv6) router connected to the global Internet
- Upon receipt of a configuration order coming from the TB, it creates, modifies or deletes the server side of each tunnel
- It may also maintain usage statistics for every active tunnel

https://t.me/learningnets

# How Tunnel Broker Mechanisms Works

- The client of the Tunnel Broker service is a dual-stack IPv6 node (host or router) which is connected to the IPv4 Internet

- Approaching the TB, the client should be asked first of all to provide its identity and credentials so that proper user authentication, authorization and (optionally) accounting can be carried out (e.g., relying on existing AAA facilities such as RADIUS)

- This means that the client and the TB have to share a pre-configured or automatically established security association to be used to prevent unauthorized use of the service

https://t.me/learningnets

# How Tunnel Broker Mechanisms Works

- **When the host request to create IPv6 Tunnel, The Tunnel Broker manages the client requests as follows:**

- Tunnel Broker first designates (e.g., according to some load sharing criteria defined by the TB administrator) a Tunnel Server to be used as the actual tunnel end-point at the network side

- It chooses the IPv6 prefix to be allocated to the client; the prefix length can be anything between 0 and 128, most common values being 48 (site prefix), 64 (subnet prefix) or 128 (host prefix)

https://t.me/learningnets

## How Tunnel Broker Mechanisms Works

- It fixes a lifetime for the tunnel; - it automatically registers in the DNS the global IPv6 addresses assigned to the tunnel end-points

- It configures the server side of the tunnel

https://t.me/learningnets

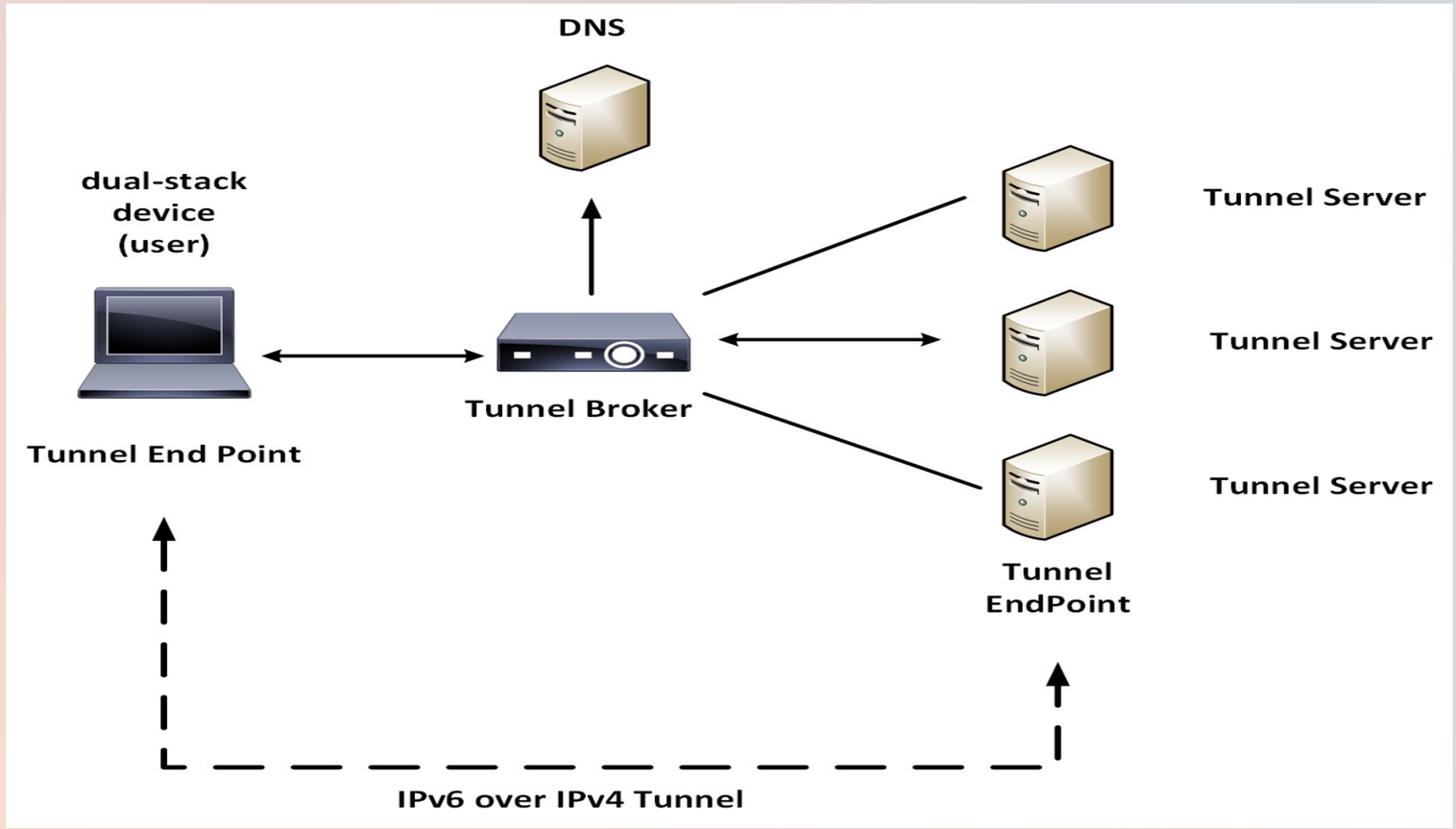
## How Tunnel Broker Mechanisms Works

- It notifies the relevant configuration information to the client, including tunnel parameters and DNS names

After the above configuration steps have been carried out (including the configuration of the client), the IPv6 over IPv4 tunnel between the client host/router and the selected TS is up and working, thus allowing the tunnel broker user to get access to the IPv6 Internet

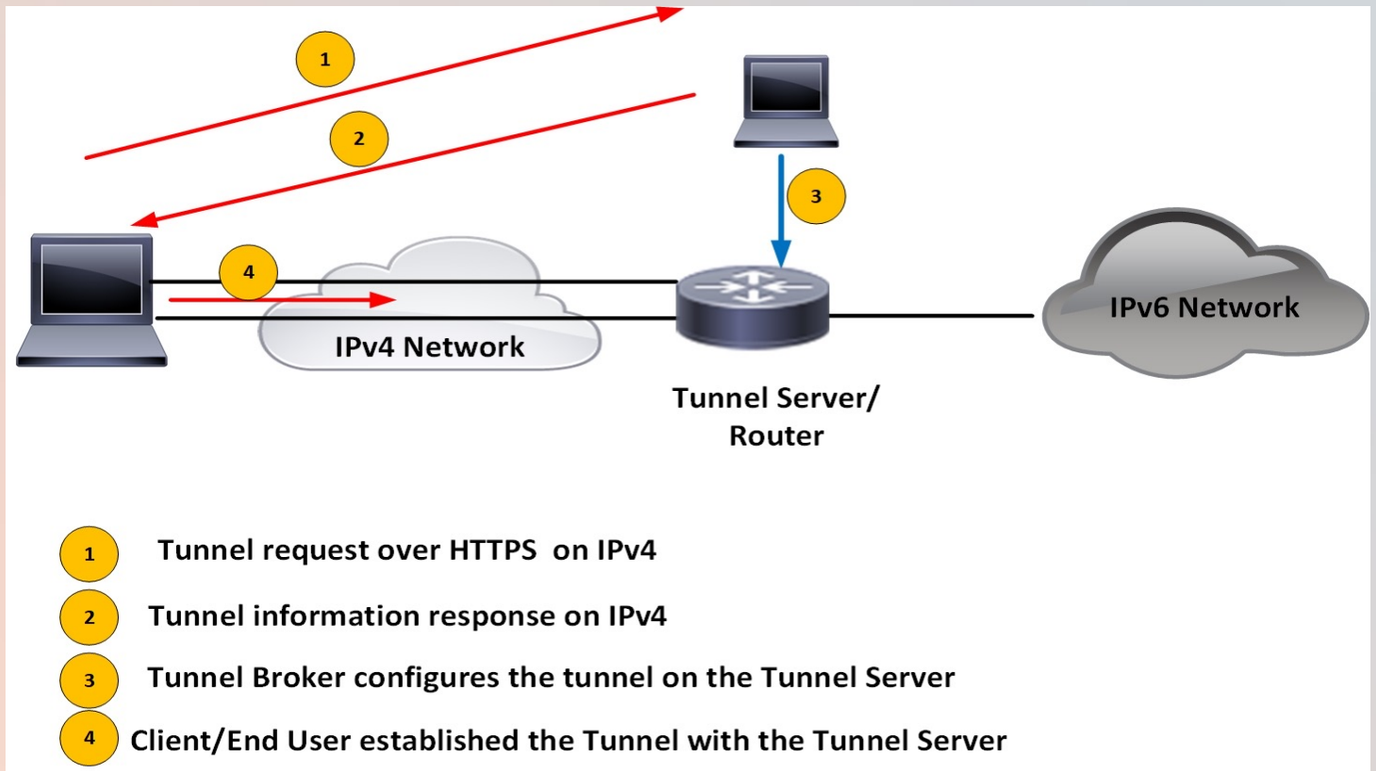
https://t.me/learningnets

# How Tunnel Broker Works




https://t.me/learningnets

# How Tunnel Broker Works




**Tunnel Broker**

https://t.me/learningnets

## Who provides IPv6 Address to the Tunnel Broker Service End Users?

- The IPv6 addresses assigned to both sides of each tunnel must be global IPv6 addresses belonging to the IPv6 addressing space managed by the TB

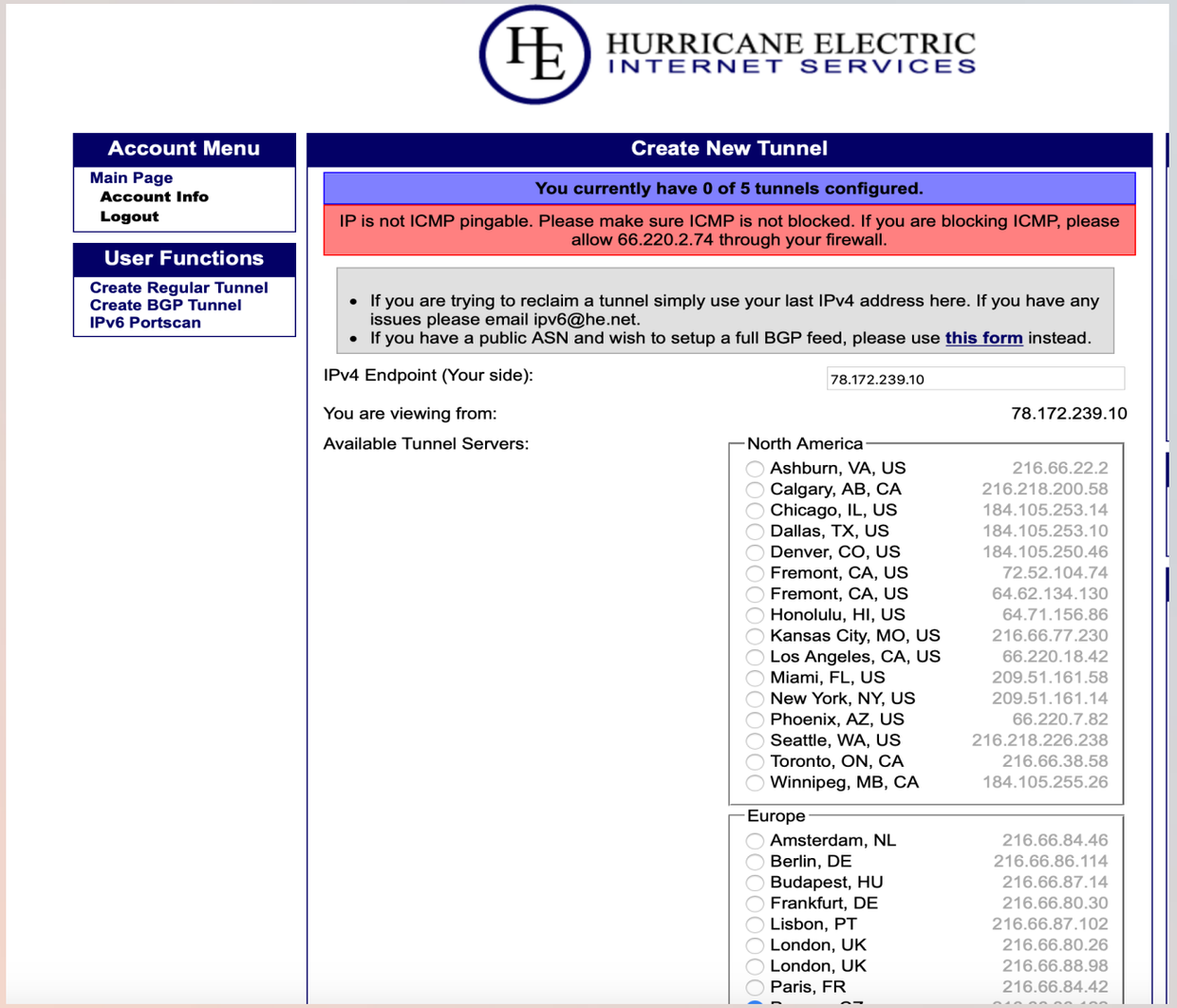
- You can find public FREE Ipv6 Tunnel Broker services from http://www.sixxs.net/ and http://tunnelbroker.net/

https://t.me/learningnets

# Tunnel Broker – Hurricane Electric

- I created an account on HE to have IPv6 tunnel from my PC to HE Tunnel Servers as you can see from this figure

- I am a IPv6 requesting client, HE is a Tunnel Broker, communication between us was a Https portal of Hurricane Electric




https://t.me/learningnets

# Automatic Tunnels for IPv6 Transition

- With Automatic Tunnels, Tunnel endpoints must be derived automatically , thus provides scalability in design

- IPv4 endpoint addresses are embedded in IPv6 address

- Automatic tunnels are generally suitable for temporary tunnels, transient connectivity : site to site or host to host VPNs

https://t.me/learningnets

# 6to4 Automatic Tunnel

- RFC 3056 is explained 6to4 Automatic Tunnels as  “Connection of IPv6 Domains via IPv4 Clouds”

- Tunnel IPv4 endpoint addresses are embedded in the IPv6 address with this tunneling mechanism thus it is a stateless tunneling mechanism

- 6to4 Tunnels is used to connect two IPv6 islands over IPv4 network

https://t.me/learningnets

## 6to4 Automatic Tunnel

- It creates a block of IPv6 addresses from a locally configured IPv4 address by embedding that IPv4 address to the prefix 2002::/16, resulting in a /48 IPv6 prefix

- IPv6 packets are encapsulated by adding an IPv4 header with the Protocol field set to 41

https://t.me/learningnets

## 6to4 Automatic Tunnel

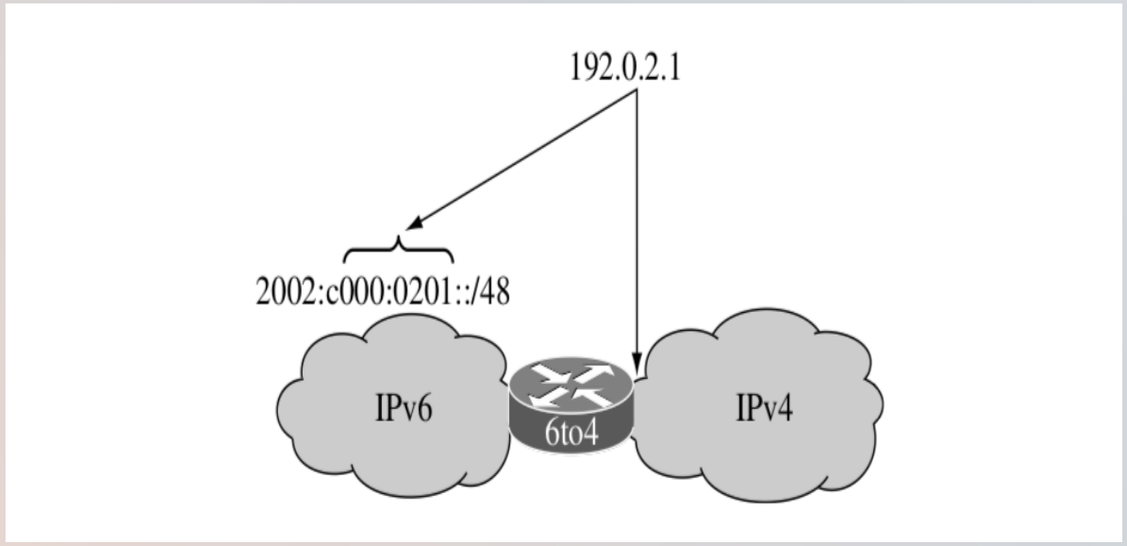
- The 6to4 IPv6 address space is built by the 2002::/16 prefix reserved for the 6to4 mechanism, followed by the 32 bits of the IPv4 external address of the border router of the site, giving the site a /48 prefix

https://t.me/learningnets

# 6to4 Automatic Tunnel

- The border router has an external IPv4 address (192.0.2.1). The IPv6 site behind the border router uses 2002:c000:0201::/48 to number its whole network.

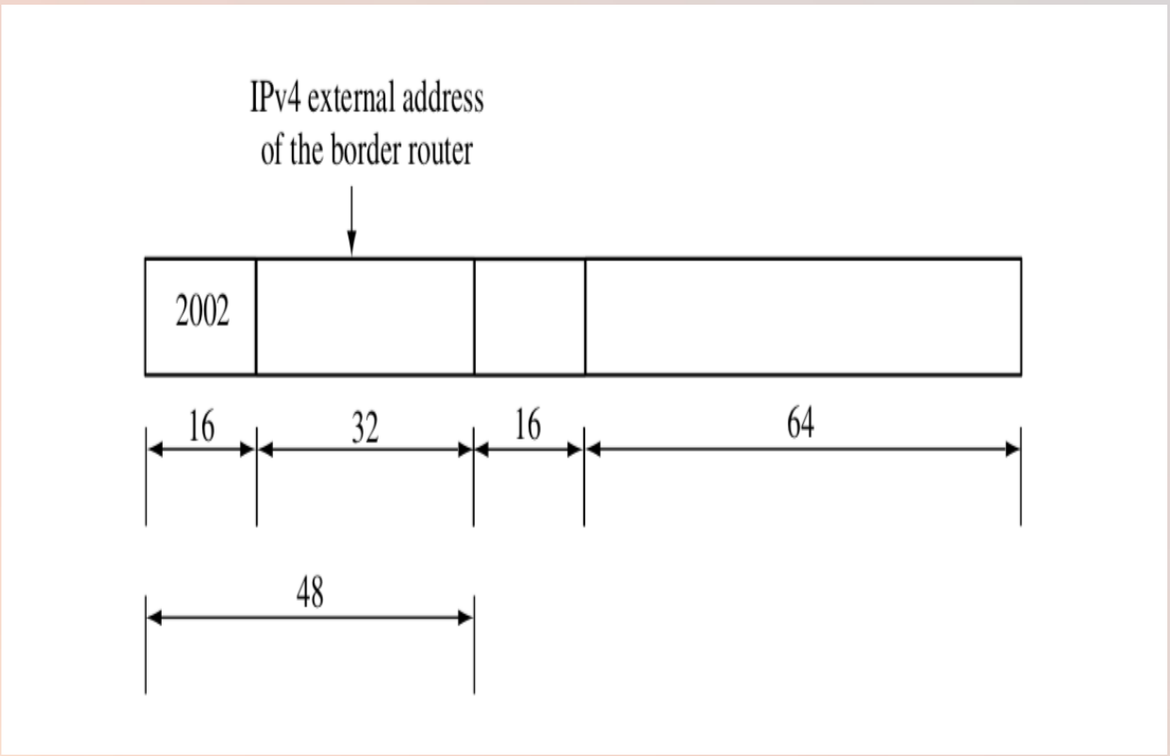
- The address space is based on 2002:<ipv4 external address in hex>::/48, where the IPv4 address is the border router external IPv4 address (192.0.2.1), represented in hexadecimal as c000:0201




6to4 site IP address space is based on the border router IPv4 address

https://t.me/learningnets

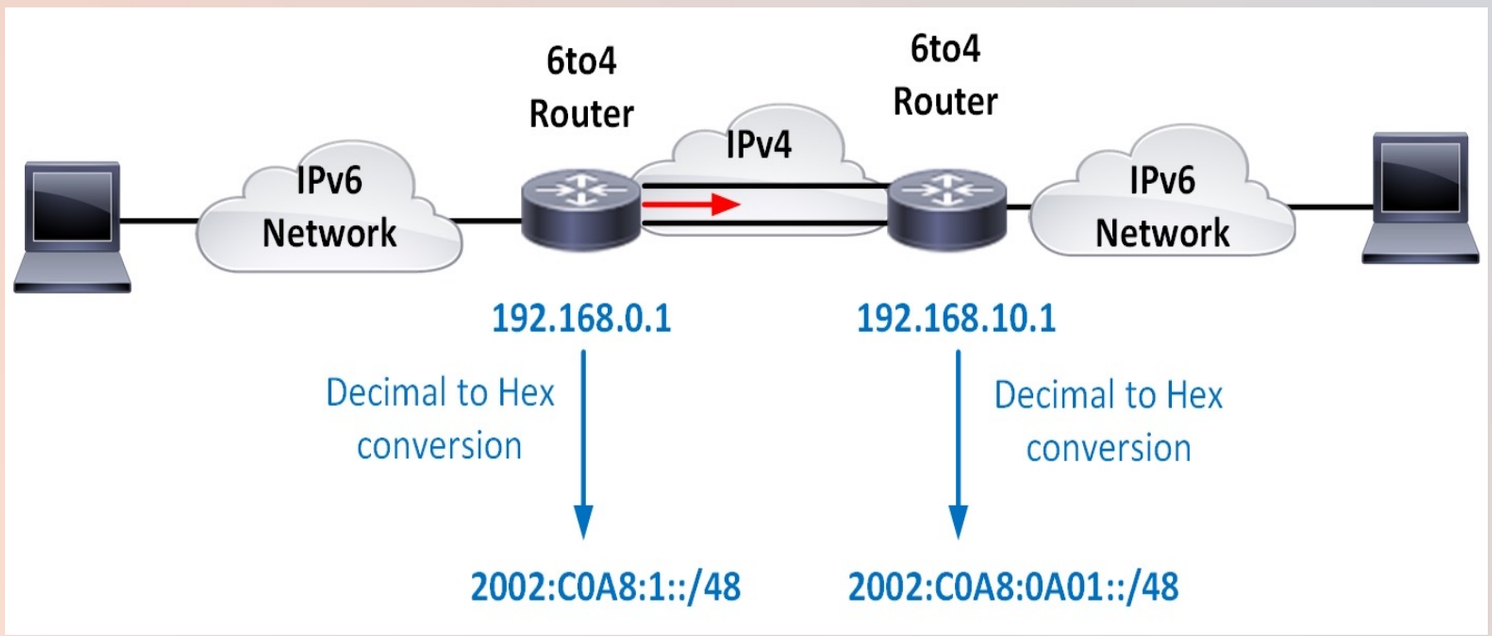
# 6to4 Automatic Tunnel




- The 6to4 mechanism needs to be only implemented in border routers. Hosts inside the IPv6 site do not need to support 6to4

https://t.me/learningnets

# 6to4 Automatic Tunnel




- 2002::/16 is allocated for 6to4 tunnels
- Border Router IPv4 address is Public as well, not RFC 1918

https://t.me/learningnets

# 6to4 Deployment/Configuration

- Below is the 6to4 Tunnel Configuration on the Cisco device

Router#

Interface loopback 0

ip address 192.168.0.1  255.255.255.0

ipv6 address 2002:C0A8:1:1::./64 eui-64

Interface Tunnel 0

tunnel source Loopback 0

tunnel mode ipv6ip 6to4

ipv6 route 2002::/16 Tunnel 0

https://t.me/learningnets

# 6to4 Tunnel Border Relay

- If host A sends a packet to C with a non-6to4 address such as 3ffe:b00:1::1, the R1 router, the border router of A's site, does not know where to route the packet since the destination address is not a 6to4 address. R1 needs a 6to4 relay to the non-6to4 IPv6 Internet

- A 6to4 relay is a 6to4 border router that has connectivity to the rest of the IPv6 networks

- It is used as a transit for the other 6to4 sites to reach the non-6to4 IPv6 networks

https://t.me/learningnets

# 6to4 Tunnel Border Relay

- To enable the 6to4 relay, the 6to4 relay router is a 6to4 router with a default route to the IPv6 Internet

- The 6to4 relay should contain some ingress filtering

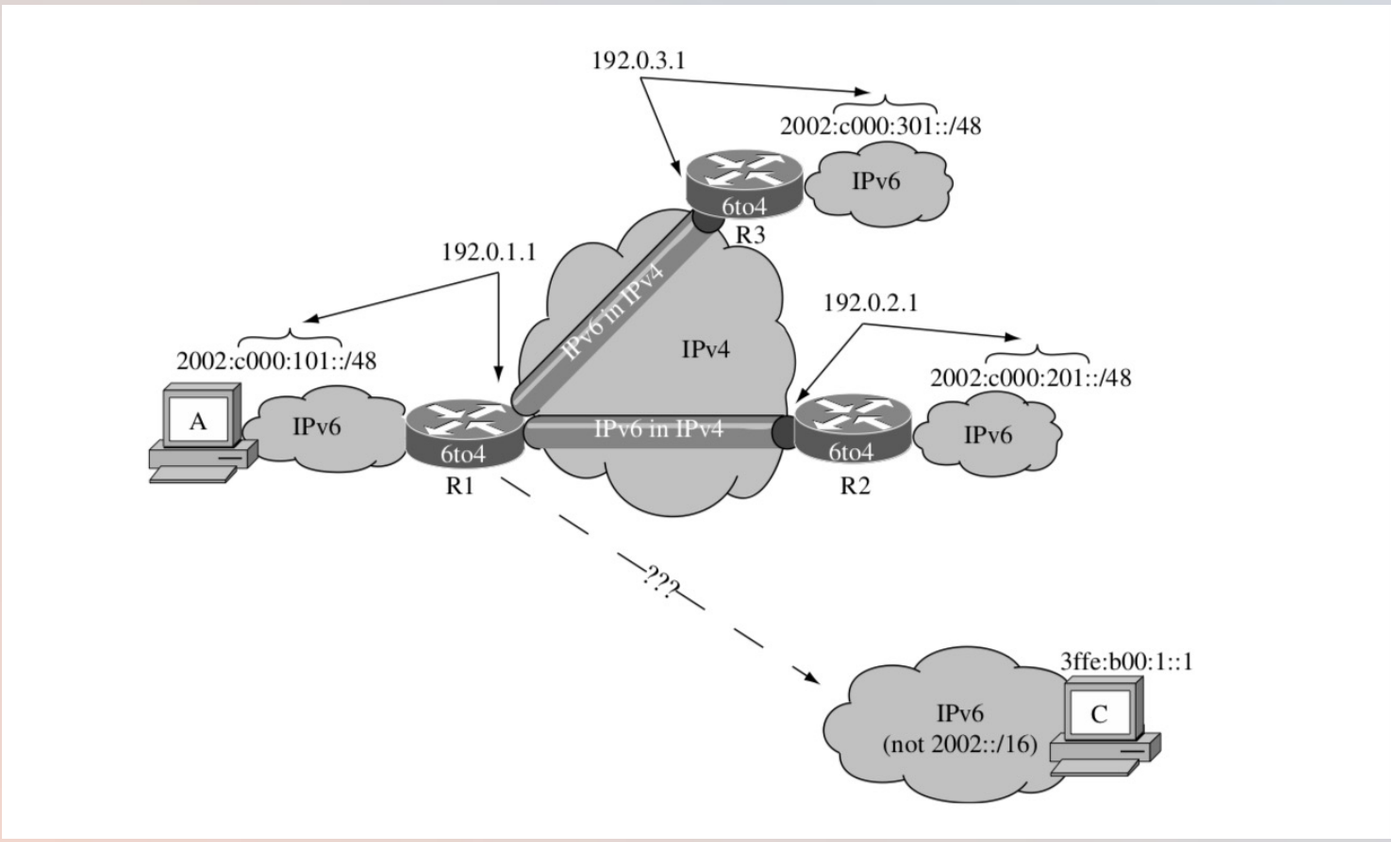
- A 6to4 site that is using a 6to4 relay installs in the 6to4 border router an IPv6 default route pointing to the 6to4 address of the relay

- 6to4 relay routers do not require specific features to act as 6to4 relays, just a static route entry

https://t.me/learningnets

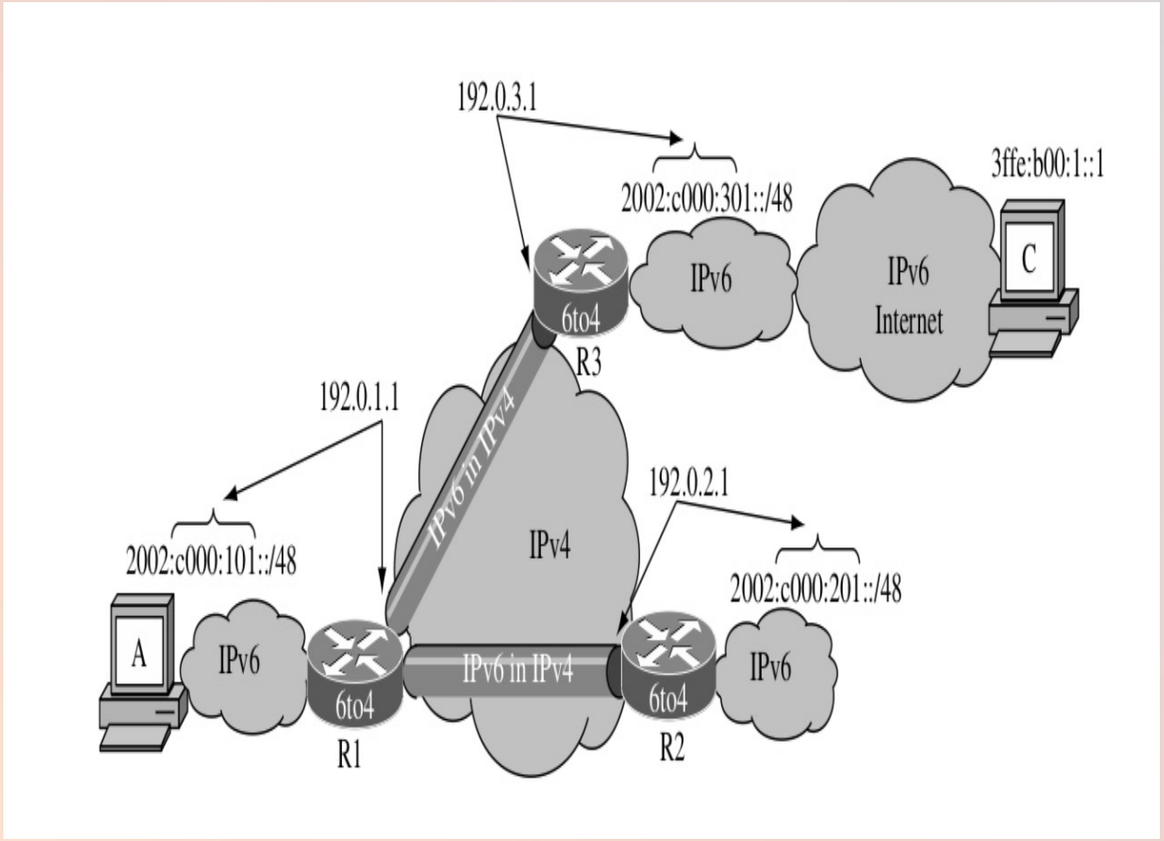
# 6to4 Tunnel Border Relay – Without Relay

- Without Border Relay, 6to4 site cannot communicate with non-6to4 site




https://t.me/learningnets

# 6to4 Tunnel Border Relay – With Relay




In the left figure, A sends a packet to C. R3 is the 6to4 relay for R1 and is connected to the non-6to4 IPv6 Internet.

When R1 wants to forward the packet to a non-6to4 destination address (3ffe:b00:1::1), it cannot build an automatic tunnel to some other 6to4 router, since it cannot extract the IPv4 address from the IPv6 destination address. If R1 has a default route to go through R3 via the 6to4 mechanism, then R1 encapsulates the IPv6 packet to R3 and R3 decapsulates it and forwards it to the IPv6 network

https://t.me/learningnets

# 6to4 Tunnel Summary

- 6to4 is an Automatic Tunneling mechanism

- Assign a IPv6 prefix to the attached network

- The 6to4 mechanism uses the 2002::/16 prefix. Any 6to4 site has the following address space: 2002:<ipv4 external address in hex>::/48. Only the border router has to support 6to4

- 2002::/16 address is assigned to the customer network, not an address from Global Unique Address of the company, this is the limitation of 6to4  tunneling mechanism

https://t.me/learningnets

# 6to4 Tunnel Summary

- Border Routers run dual stack and support 6to4
- A 6to4 capable border router cannot use 6to4 if it is not assigned a public IPv4 address.
- All 6to4 sites are reachable through their IPv4 border router address
- All 6to4 sites have a 6to4 relay, statically configured on the site border router, to transit the non-6to4 IPv6 traffic
- Hosts do not need to support or know about 6to4

https://t.me/learningnets

# 6rd – IPv6 Rapid Deployment

- RFC 5569 is written for IPv6 Rapid Deployment on IPv4 Infrastructures

- 6rd is an extension of 6to4 tunnels

- It is an Automatic tunneling mechanism (IPv6 over IPv4 network)

https://t.me/learningnets

# 6rd – IPv6 Rapid Deployment

- With 6to4 tunnels, 2002::/16 reserved prefix need to be used as an IPv6 prefix

- With 6rd, this restriction is removed, IPv6 prefix can be used from the company's local address block (Global Unicast Block)

- 6rd consists of two main hardware components, the CE (Customer Equipment) router and the BR (Border Relay) router

https://t.me/learningnets

# 6rd - CE (Customer Edge) Router

- The CE router is positioned at the edge of the service provider IPv4 access infrastructure
- Provides IPv6 connectivity to this end user's network
- The native IPv6 traffic coming from the end user hosts is encapsulated in IPv4 by the CE router and tunneled to the Border Relay router or directly to other CE routers in the same 6rd domain
- Conversely, encapsulated 6rd traffic received from the Internet through the Border Relay router and 6rd traffic from other CE routers will be de-capsulated and forwarded to the end-user nodes

https://t.me/learningnets

# 6rd - Border Relay Router

- The BR router provides connectivity between the CE routers and the IPv6 Internet

- Both the CE and BR routers are dual-stack devices, and the devices between the BR and CE routers can be IPv4 only

https://t.me/learningnets

# 6rd - Border Relay Router

- At the CE router, if the packet IPv6 destination address matches the locally configured 6rd prefix, the packet is considered to be part of the local 6rd domain and needs to be forwarded to another CE router

- In such a case, the IPv4 address embedded in the IPv6 destination address is used as the IPv4 destination address of the 6rd tunnel, and the local WAN interface IPv4 address is used as the source address for the 6rd tunnel, which is an IPv6 packet directly encapsulated in IPv4

https://t.me/learningnets

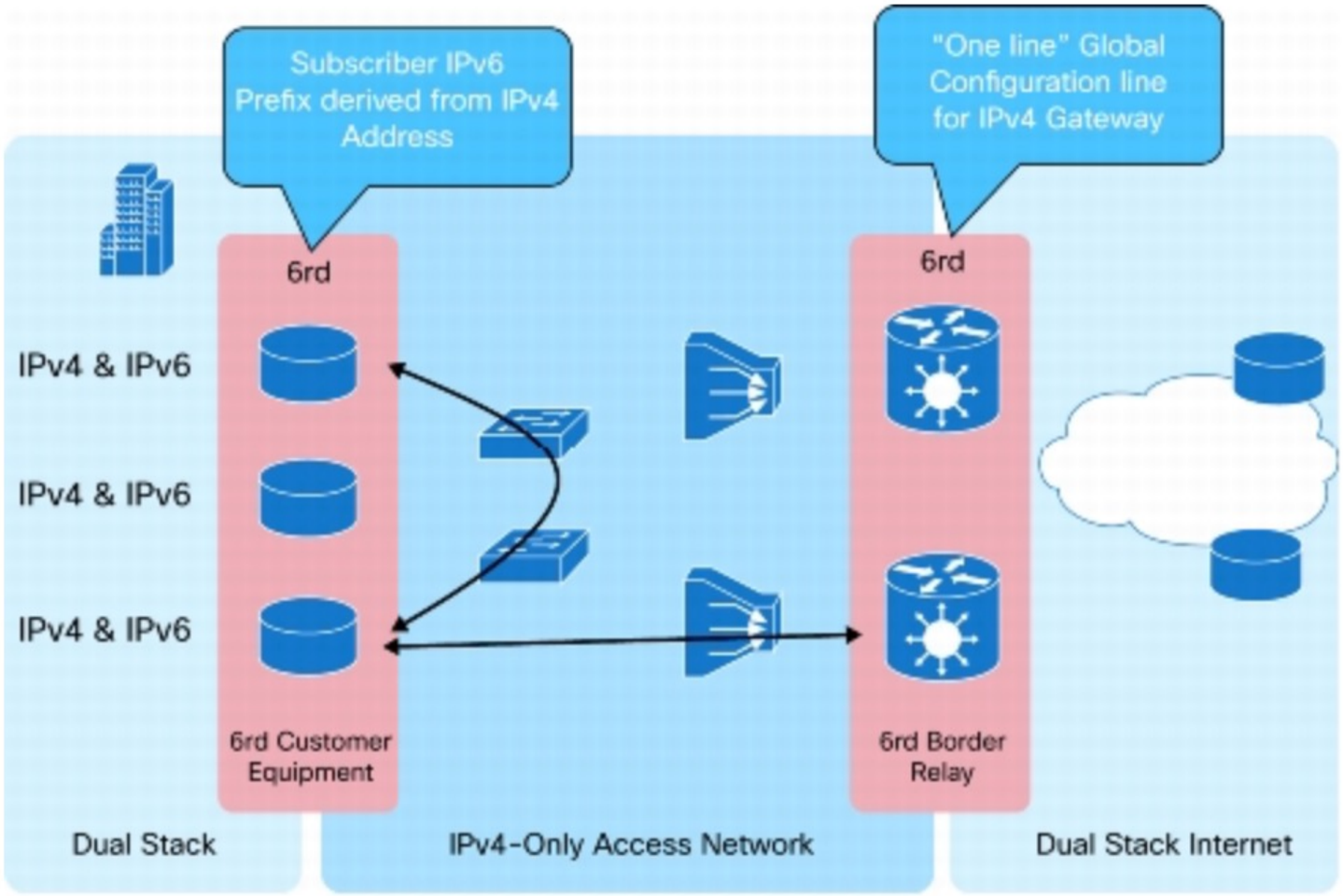
# 6rd - Border Relay Router

- If the IPv6 destination address does not match the locally configured 6rd prefix-in other words, if the packet does not belong to the local 6rd domain-the packet will be tunneled to the BR router by a 6rd tunnel

- In this case, the locally configured BR IPv4 address on the CE router is used as the destination address for the encapsulated packet

https://t.me/learningnets

# 6rd - Border Relay Router




6rd consists of two main hardware components, the CE (Customer Equipment) rou

https://t.me/learningnets

# 6rd – IPv6 Rapid Deployment

- Similar to 6to4 IPv6 tunnels, 6rd is a standard based stateless tunneling mechanism

- Since 6rd is Stateless, packets don’t have to go through the same BR (Border Relay) router

- Border Relay is required if the destination IPv6 address is outside the company network

https://t.me/learningnets

# 6rd – IPv6 Rapid Deployment

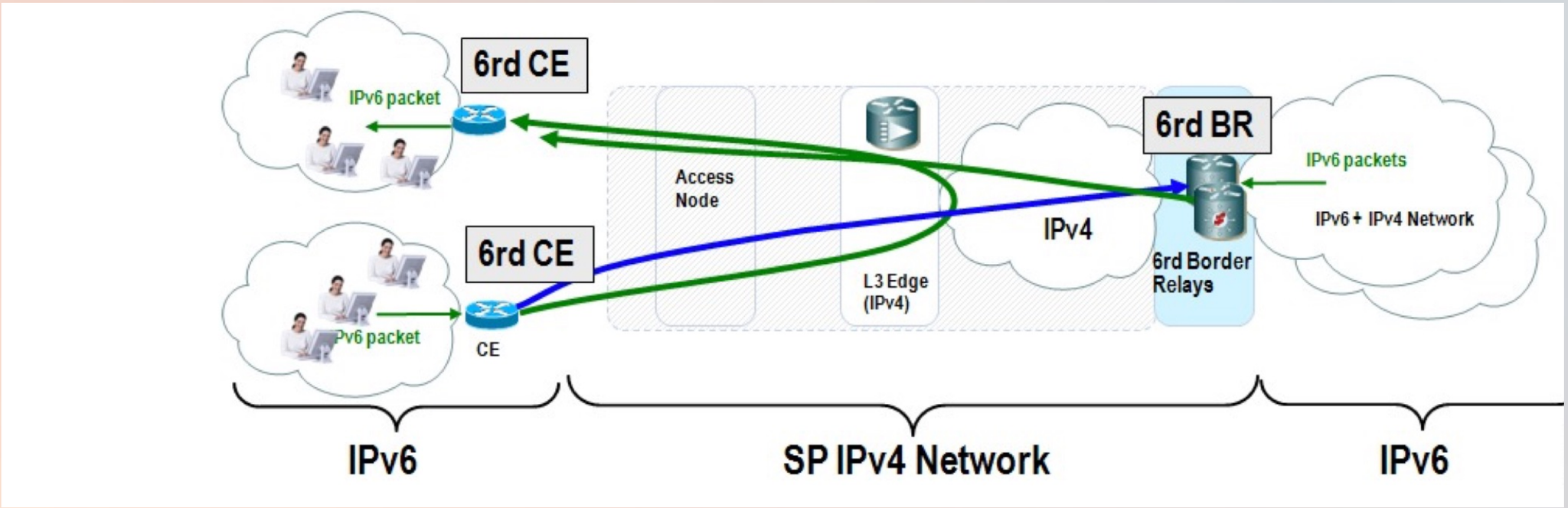
If two IPv6 customer sites talk to each other, traffic doesn't go through Border Relay Routers

- For HA and Load Balancing, more than one Border Relay is used

- Each border relay router needs to be configured with the same IPv4 address (Anycast) so that CE routers are routed to the closest border relay

https://t.me/learningnets

# 6rd – IPv6 Rapid Deployment Traffic Flow Within Domain and Outside Domain




https://t.me/learningnets

# 6rd Summary

- Extension of 6to4 Tunnel

- Used as a stateless tunneling mechanism as IPv6 Transition mechanism

- IPv6 networks can communicate over IPv4 network

- Removed the requirement of having 2002::/16 IPv6 prefix for tunneling, instead company local IPv6 prefix can be used for tunneling

https://t.me/learningnets

# ISATAP - Intra-Site Automatic Tunnel Addressing Protocol

- It is a host to router or host to host Automatic tunneling mechanism, host can tunnel IPv6 traffic across IPv4 network

- Defined in RFC5214, so it is a standard based mechanism

- ISATAP enables unicast communication between IPv6/IPv4 hosts across the IPv4-only Intranet and Internet (If ISATAP router has an access to the IPv6 Internet), it doesn’t support Multicast

https://t.me/learningnets

# ISATAP - Intra-Site Automatic Tunnel Addressing Protocol

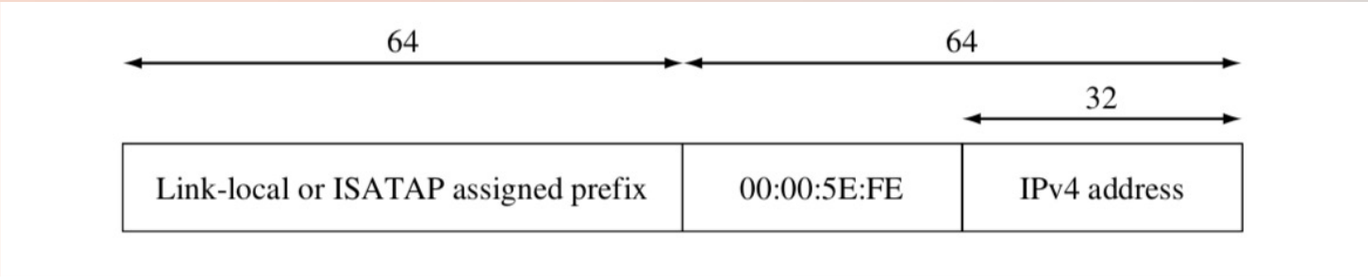
- Embeds the IPv4 address of the node in the last 32 bits of the interface identifier part of its IPv6 address

- It utilizes DNS to determine what prefix it is to assign and what gateway to use

https://t.me/learningnets

# ISATAP - Intra-Site Automatic Tunnel Addressing Protocol

- Below figure shows the format of an ISATAP address. The first 32 bits of the interface identifier are ‘00:00:5E:FE’, reserved by IANA for ISATAP, and define the ISATAP interface identifier




https://t.me/learningnets

# ISATAP - Intra-Site Automatic Tunnel Addressing Protocol

- ISATAP creates a virtual link over a full IPv4 site, thus enabling ISATAP on a host/router creates on them a virtual interface with an IPv6 address

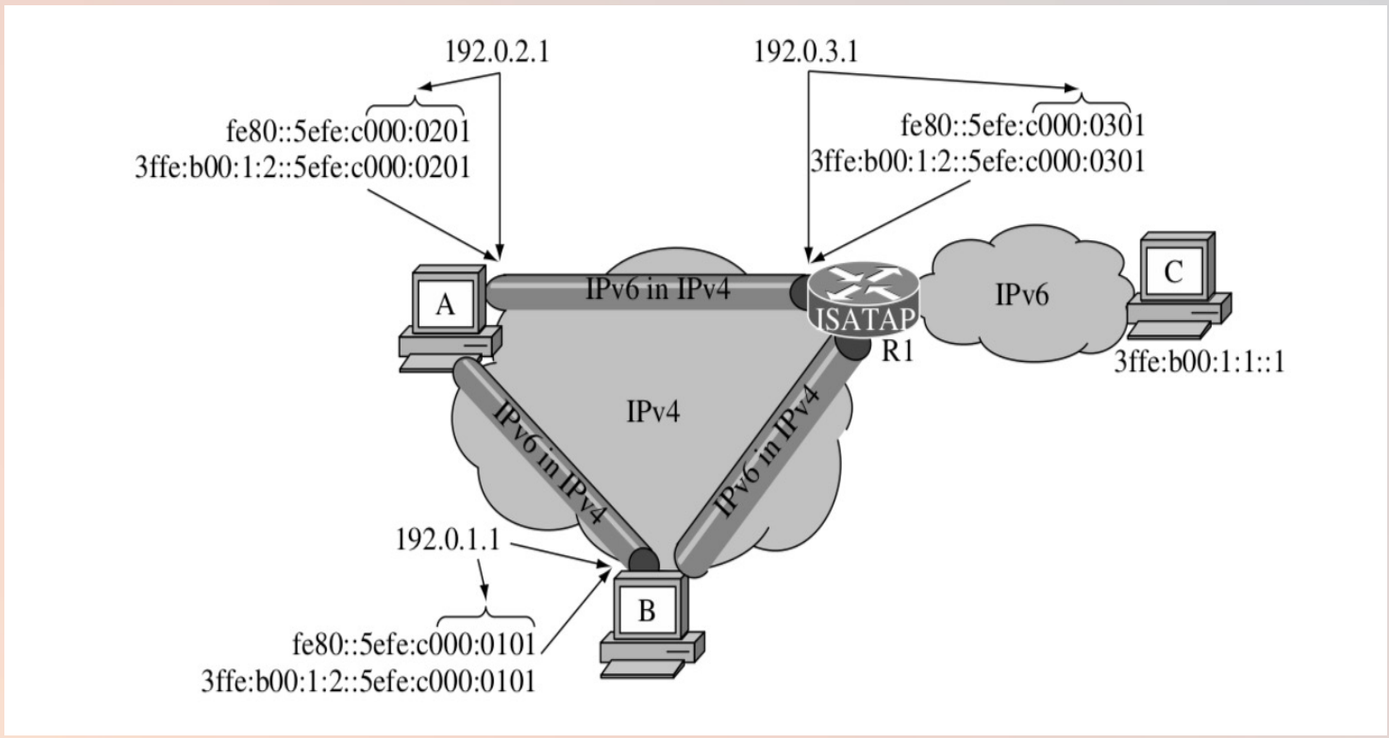
- ISATAP implementation in an organization is designed to take your entire IPv4 network, and make it one big IPv6 logical link. You don't "subnet" ISATAP networks. So, all of your IPv4 becomes one large IPv6 subnet as far as ISATAP is concerned

- Link-local address  (fe80::/64), unique-local, or global addresses can be used as IPv6 ISATAP addresses

- Hosts and the Router which will have IPv6 over IPv4 tunneling, needs to have Dual-Stack and ISATAP to be enabled

https://t.me/learningnets

# ISATAP – Link Local and Global ISATAP Addressing

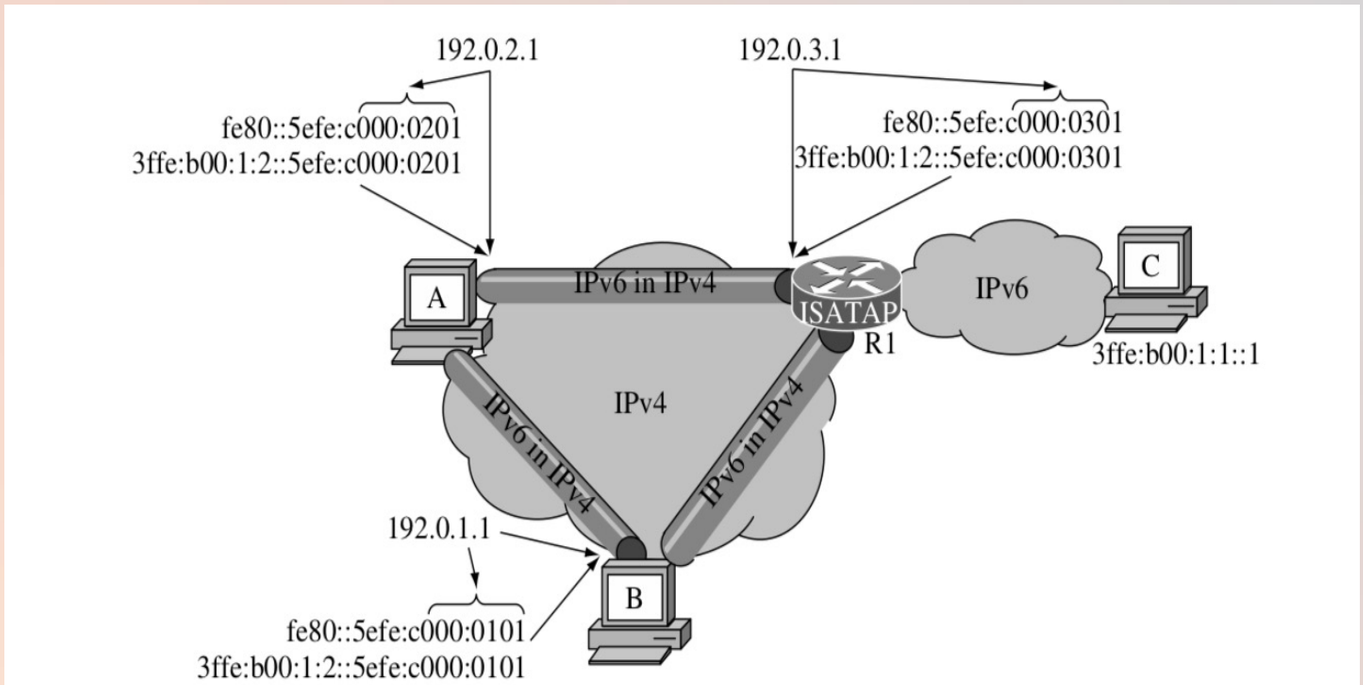



Host A, host B and router R1 are all dual-stack and have an enabled implementation of ISATAP

Host C is IPv6 without ISATAP. The network manager uses the 3ffe:b00:1::/48 for its site. It assigns 3ffe:b00:1:2::/64 to the ISATAP virtual link over the IPv4 network

https://t.me/learningnets

# ISATAP – Link Local and Global ISATAP Addressing




Host A has the 192.0.2.1 IPv4 address and creates its link-local address based on the ISATAP format: fe80::5efe:c000:0201, computed using the ISATAP 32 bit interface identifier (0000:5efe) and the hexadecimal representation (c000:0201) of its IPv4 address (192.0.2.1).

Host B and router R1 do the same respectively, which builds a virtual link , all are link local neighbor of each other

https://t.me/learningnets

# ISATAP Router

- In ISATAP, when hosts are dual-stack and ISATAP enabled, they directly communicate via IPv6 with each other without ISATAP router

- ISATAP router is needed when there is ISATAP host communicate with non-ISATAP enabled IPv6 host and IPv6 Internet

- ISATAP Router is needed for publishing a prefix for ISATAP IPv6 auto-addressing on the hosts

https://t.me/learningnets

# ISATAP Router

- When a host auto-configures it's ISATAP address, it first contacts the ISATAP Router to learn it's prefix

- The ISATAP Router is statically configured in all ISATAP nodes or hosts can discover the ISATAP router via DNS

https://t.me/learningnets

# ISATAP Summary

- ISATAP is used within a site, or within one administrative domain

- ISATAP creates a virtual link over a potentially wide IPv4 network. Any broadcast-like packet, such as one sent to ff01::1, on the link will create a potentially large number of packets on the network. So the more ISATAP nodes that are deployed, the less it is scalable

- ISATAP does not traverse NAT

- ISATAP doesn’t support Multicast, so you cannot run routing protocols over ISATAP tunnels

https://t.me/learningnets

# TEREDO

- Teredo enables nodes to tunnel IPv6 over IPv4 through NATs , similar to 6to4 and 6rd, it is an auto tunnel IPv6 over IPv4 protocol

- It provides automatic tunneling that allows IPv6/IPv4 (Dual stack) hosts to establish IPv6 connectivity with each other across the IPv4 Internet even when IPv4 network address translation (NAT) devices need to be traversed

- Because of this capability, Teredo is considered more suitable than 6to4 for small office/home office (SOHO) environments that use NATs to hide their private IPv4 addresses from the Internet

https://t.me/learningnets

# TEREDO

- Similar to 6to4, Teredo uses it's own address prefix , which is 2001:0::/32

- If the local network is behind an IPv4 NAT, and the NAT gateway does not support 6to4, Teredo can be used by the host to reach IPv6 network

https://t.me/learningnets

# TEREDO

- Teredo encapsulates IPv6 packets over UDP over IPv4

- Teredo makes a lot of assumptions about the NAT behavior. This is based on experimental results from trying different vendors implementations at specific points in time

https://t.me/learningnets

# TEREDO

- Not only were all vendors not tried, but there is no guarantee that the observed behaviors will remain in the subsequent releases of those NAT products

- Teredo might not function because of some new or non-observed NAT behavior

https://t.me/learningnets

# DS-LITE

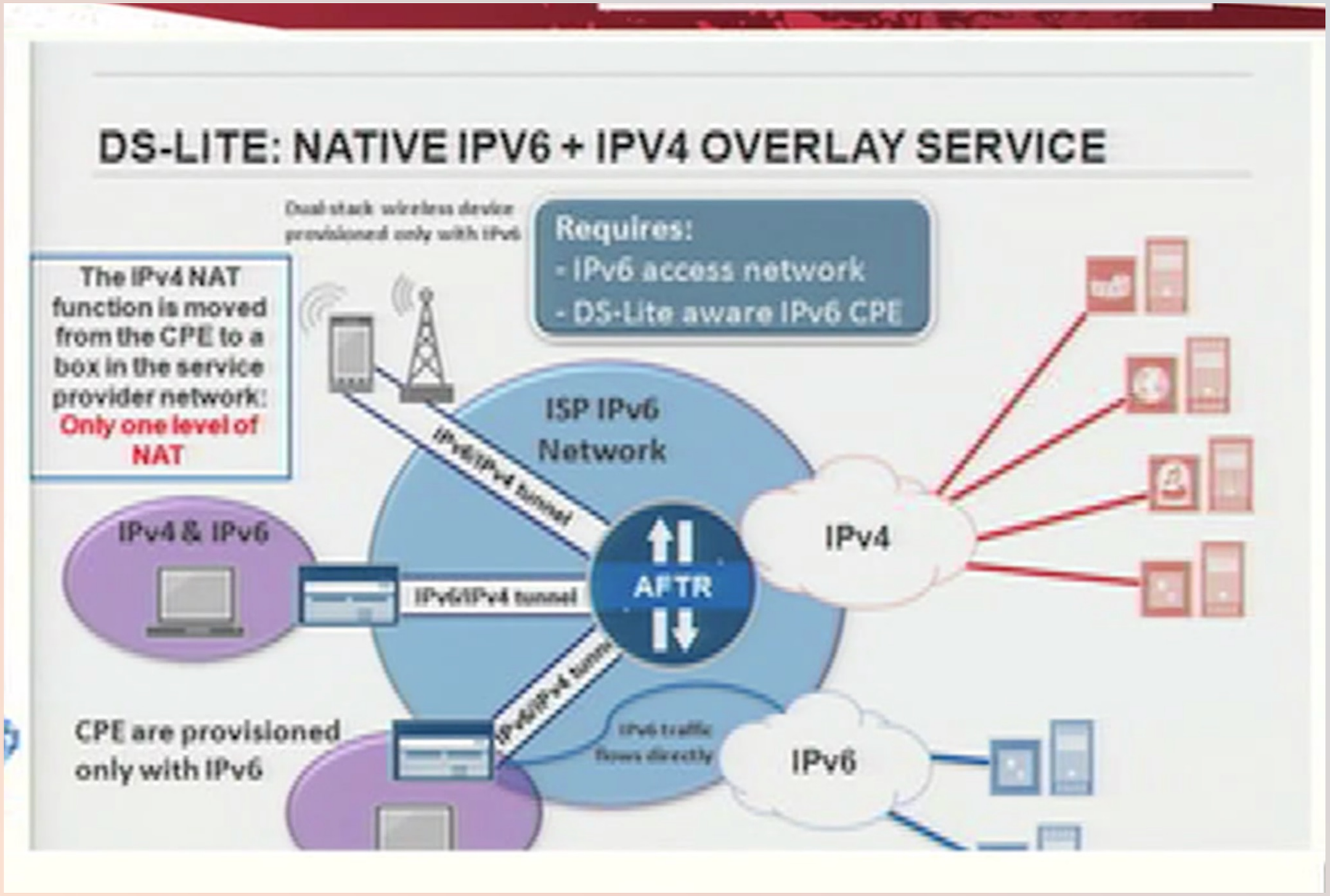
- It is an IPv4 overlay on top of an IPv6 infrastructure

- Similar to 6rd, DS-Lite is stateless , if there is NAT in the network for CGN purpose, that part is only stateful

- Access network needs to be IPv6 and CPE needs to be DS-Lite aware IPv6 node

https://t.me/learningnets

# DS-LITE




https://t.me/learningnets

# DS-LITE

- When Service Provider use DS-LITE, link between the provider and the customer (access network) is IPv6, this is considered as an advantage because network doesn’t have to run dual-stack

- Dual stack requires more resources on the device and management of it is more costly compare to IPv6 only infrastructure

- With DS-LITE only one layer of NAT is necessary , this is for connecting IPv6 only host to IPv4 content or IPv4 private addresses to be NATed to public address

https://t.me/learningnets

# DS-LITE

- It requires CPE to support DS-LITE, this is considered as disadvantage as there might be millions of CPEs and legacy devices may not be upgradable to support DS-LITE

- Some Operators deployed DS-LITE already

https://t.me/learningnets

# IPv6 Routing Protocols

https://t.me/learningnets

# IPV6 Routing Protocols

- When IPv6 routing is enabled together with IPv4 routing protocol, we need to make sure sufficient device resources on the networking devices

- IPv6 Routing Protocols (General Purpose routing protocols at least – Not RPL etc.), they are very similar to their IPv4 counterpart

- There will be some differences which will be covered next

https://t.me/learningnets

# IPv4 and IPv4 Routing Protocols Side by Side

| IPv4 Routing Protocols | IPv6 Routing Protocols |
|---|---|
| RIPv2 | RIPnG – Next Generation |
| OSPFv2 | OSPFv3 |
| ISIS | ISIS for IPv6 |
| EIGRP | EIGRP for IPv6 |
| BGPv4 | MP-BGP |
| PIM | PIM for IPv6 |

https://t.me/learningnets

# OSPF for IPv6 (OSPFv3)

- Operates very similar to OSPFv2 , both are link state protocols also other things are similar such as the LSA flooding rules, the LSA aging mechanisms, and the interface types (broadcast, point-to-point, point-to-multipoint, among others)

- Both OSPFv2 and OSPFv3 have two level of hierarchy (Backbone and Non-Backbone Areas)

- OSPFv2 only supports IPv4 but OSPFv3 supports both IPv4 and IPv6

https://t.me/learningnets

# OSPF for IPv6 (OSPFv3)

- But in OSPFv3, topology and reachability information are carried in different LSA

- Thus, adding a loopback interface for example doesn’t trigger full SPF run as it doesn’t change the topology of the network

- New LSA Type defined for OSPFv3

https://t.me/learningnets

## OSPF for IPv6 (OSPFv3)

- If you make a simple change, like changing the IP address on one of your routers then the topology itself doesn't change

- In OSPFv2, a new type 1 LSA and perhaps a type 2 LSA have to be flooded. Other routers that receive the new LSA(s) have to recalculate the SPT even though the topology did not change

- In OSPFv3, they changed this by creating a *separation* between prefixes and the topology

https://t.me/learningnets

## OSPF for IPv6 (OSPFv3)

There is no prefix information in LSA type 1 and 2, you only find topology information in these LSAs, you don't find any IPv6 prefixes in them!

- Prefixes are now advertised in type 9 LSAs and the link-local addresses that are used for next hops are advertised in type 8 LSA

- Type 8 LSAs are only flooded on the local link, type 9 LSAs are flooded within the area.

https://t.me/learningnets

- Type 1 and Type 2 LSA repurposed , Type 8 and Type 9 LSA are added in OSPFv3 (Link and Intra-Area Prefix Respectively)

- Type 8 LSA is link local only, Type 9 is Area wide

- LS Types indicates Scope as well (E is domain wide, 0x4005)

| OSPFv3 LSAs | | OSPFv2 LSAs | |
|---|---|---|---|
| **LS Type** | **Name** | **Type** | **Name** |
| 0x2001 | Router LSA | 1 | Router LSA |
| 0x2002 | Network LSA | 2 | Network LSA |
| 0x2003 | Inter-Area Prefix LSA | 3 | Network Summary LSA |
| 0x2004 | Inter-Area Router LSA | 4 | ASBR Summary LSA |
| 0x4005 | AS-External LSA | 5 | AS-External LSA |
| 0x2006 | Group Membership LSA | 6 | Group Membership LSA |
| 0x2007 | Type-7 LSA | 7 | NSSA External LSA |
| 0x0008 | Link LSA | | *No Corresponding LSA* |
| 0x2009 | Intra-Area Prefix LSA | | *No Corresponding LSA* |

https://t.me/learningnets

# OSPF for IPv6 (OSPFv3)

- **Inter-Area Prefix LSA:**

  These LSAs are IPv6 equivalent of IPv4's Type-3 Summary LSAs. These LSAs are originated by the ABR to specify IPv6 prefixes that belong to other areas. A separate LSA is originated for each address prefix

- **Inter-Area Router LSA:**

  These LSAs are IPv6 equivalent of IPv4's Type-4 Summary LSAs. Originated by the ABR, the Inter-Area Router LSA describes the route to the ASBR. Each LSA describes a route to a single router

| OSPFv3 LSAs | | OSPFv2 LSAs | |
|---|---|---|---|
| **LS Type** | **Name** | **Type** | **Name** |
| 0x2001 | Router LSA | 1 | Router LSA |
| 0x2002 | Network LSA | 2 | Network LSA |
| 0x2003 | Inter-Area Prefix LSA | 3 | Network Summary LSA |
| 0x2004 | Inter-Area Router LSA | 4 | ASBR Summary LSA |
| 0x4005 | AS-External LSA | 5 | AS-External LSA |
| 0x2006 | Group Membership LSA | 6 | Group Membership LSA |
| 0x2007 | Type-7 LSA | 7 | NSSA External LSA |
| 0x0008 | Link LSA | | *No Corresponding LSA* |
| 0x2009 | Intra-Area Prefix LSA | | *No Corresponding LSA* |

https://t.me/learningnets

# OSPF for IPv6 (OSPFv3)

- An OSPFv2 router forms adjacencies using its configured IPv4 interface address

- OSPFv3, however, makes use of IPv6's link-local address scope (FE80::/10). All OSPFv3 adjacencies are formed using link-local addresses

https://t.me/learningnets

# OSPF for IPv6 (OSPFv3)

- Neighboring routers are referred to not by IP address, but by OSPF ID, demonstrating OSPFv3's fundamental separation of the SPF tree and IP addressing

- OSPFv3 router IDs *are not IPv4 addresses*; they are merely unique 32-bit identifiers expressed in the familiar dotted-decimal notation

https://t.me/learningnets

# OSPF for IPv6 (OSPFv3)

- Unknown LSA Type Handling : OSPFv2 routers simple discard LSAs of an unknown type. OSPFv3 LSAs may be discarded, or optionally stored and flooded as though they were understood

https://t.me/learningnets

# OSPF v2 vs. OSPFv3

| Design Concern | OSPFv2 | OSPFv3 |
|---|---|---|
| Scalability | Good | Better since Router and Network LSA doesn't contain prefix information but only topology information |
| Working on Full Mesh | Works well with mesh group | Works well with mesh group |
| Working on Hub and Spoke | Works poorly, require a lot of tuning | Works bad requires tuning |
| Fast Reroute Support | Yes - IP FRR | Yes - IP FRR but limited platform support |
| Suitable on WAN | Yes | Yes |
| Suitable on Datacenter | DCs are full mesh. So, Not well | DCs are full mesh so Not well |
| Suitable on Internet Edge | No it is designed as an IGP | No it is designed as an IGP |
| Standard Protocol | Yes IETF Standard | Yes IETF Standard |
| New LSAs | None | Links LSA (Type 8 ) is used for adjaceny formation and link local scope only, Inter-Area-Prefix LSA (Type9) which is one of the biggest enhancement since it is used to carry prefix information only,inside an area |
| LSA Types | Router(Type1),Network(Type2),Summary(Type3),ASBR External(Type4),AS External(Type5),NSSA(Type7) | Router(Type1),Network(Type2),Inter-Area Prefix(Type3),Inter-Area Router(Type4),AS External(Type5),NSSA(Type7),Link LSA(Type8),Intra-Area-Prefix-LSA(Type9) |
| Transport | Multicast, 224.0.0.5 and 224.0.0.6 | Same but with IPv6 addresses. Multicast. FF02::5 and FF02::6 |
| Reachability info handeling | Inside an Area, Router and Network LSA carries the reachability information,between areas reachability info is carried in Summary(Type3) LSA | Inside an area reachability information is carried in Intra Area Prefix LSA (Type9) which is new LSA type, inter area prefixes are still carried in Type 3 LSA but name is changed as Inter-Area prefix LSA |
| Topology info handeling | Inside an Area Router and Network LSA carries the topology information,topology info is not carried beyond an area | Same.Inside an Area Router and Network LSA carries the topology information,topology info is not carried beyond an area |
| Stuff Experince | Very well known | Not well known, especially topology and reachability information handling,Multi Area Adjaceny and new LSA types should be understood better |
| Overlay Tunnel Support | Yes it supports | Yes it supports |
| MPLS Traffic Engineering Support | Yes with CSPF or external controller | Yes, with CSPF or external controller |
| Security | MD5 Authentication | Authentication is removed since it runs on top of IPv6, IPv6 supports IPSEC and Authentication, this simplifies the OSPF header |
| Suitable as Enterprise IGP | Yes | Yes |
| Suitable as Service Provider IGP | Yes | Definitely |
| Complexity | Easy | Moderate |
| Resource Requirement | Full SPF runs on prefix or topology change so it is worse than OSPFv3 | If topology doesn't change, full SPF is not needed. Prefix information is carried in new LSA, not in Router LSA anymore |
| IPv6 Support | No | Yes |
| IPv4 Support | Yes | Yes |
| Default Convergece | Slow | Even slower if multiple address families are used |
| Troubleshooting | Easy | Harder,requires understanding IPv6 addressing, after that it is same packet types, LSA, LSU, DBD |
| Routing Loop | Inter area prefixes should be received from ABR,all non-backbone areas should be connected to the backbone area | Same as OSPFv2. Inter area prefixes should be received from ABR,all non-backbone areas should be connected to the backbone area |

https://t.me/learningnets

# IS-IS for IPv6

- RFC5308 adds IPv6 address family support to IS-IS

- IS-IS can be deployed for IPv4 and IPv6 as Single Topology and Multi Topology

- With Single Topology IS-IS, IPv4 and IPv6 shares same link state topology and same SPF calculation for both IPv4 and IPv6

- Separate link state topology for IPv4 and IPv6 IS-IS in Multi Topology Routing IS-IS , Separate SPF Calculation for IPv4 and IPv6 in MTR IS-IS

https://t.me/learningnets

# IS-IS for IPv6

- Single Topology better for resource usage but Multi Topology might be good for resilience based on the network topology

- In order to have IPv6 support in IS-IS, unlike OSPF , we don’t need to have separate version of the protocol, instead new TLVs (Type, Length, Value Encoding) are simply added

- Thus, IS-IS is considered as extendible protocol when it is compared with other link state routing protocol which is OSPF

https://t.me/learningnets

# IS-IS for IPv6

**2 Type/Length/Values (TLV) added to support IPv6 routing in IS-IS:**

1. IPv6 Reachability TLV (0xEC) :  Describes network reachability such as IPv6 routing prefix, metric information and some option bits

2. IPv6 Interface Address TLV (0xE8) :  Contains a 128 bit address  For Hello PDUs, must contain the link-local address (FE80::/10). For LSP, must only contain the non link-local address

https://t.me/learningnets

# EIGRP for IPv6

- EIGRP is a Cisco propriety protocol which supports IP routing, there is an Informational RFC 7868 , no stub feature

- EIGRP, similar to IS-IS, has a TLV based encoding schema, thus, it is considered as extendible protocol

- EIGRP has been extended to support IPv6 routing , new TLVs have been added

https://t.me/learningnets

# EIGRP for IPv6

- IPv6_REQUEST_TYPE
- IPv6_METRIC_TYPE
- IPv6_EXTERIOR_TYPE are the new TLVs.
- Hellos are sourced from the link-local address and destined to FF02::A (all EIGRP routers).
- This means that neighbors do not have to share the same global prefix (with the exception of explicitly specified neighbors where traffic is unicasted)

https://t.me/learningnets

# EIGRP for IPv6

- Automatic summarization is disabled by default for IPv6 (unlike IPv4)

- No split-horizon in the case of EIGRP for IPv6 (because IPv6 supports multiple prefixes per interface)

- No need to have new version of the protocol to have IPv6 with EIGRP, similar to IS-IS

https://t.me/learningnets

# IPv6 in BGP

- BGP4 is used for IPv6, IPv4 and many address family

- When BGP is used any address family other than IPv4 Unicast, we call it MP-BGP (Multi Protocol BGP)

- Single protocol (BGPv4) is used to carry both IPv4 and IPv6

https://t.me/learningnets

# IPv6 in BGP

- IPv6 behavior is similar as the IPv4 behavior

- IPv4 routes can be exchanged over an IPv6 TCP session and vice versa

- There may be two next-hop addresses in the next-hop attribute

https://t.me/learningnets

# IPv6 in BGP

Address Family Information (AFI) for IPv6:

- AFI = 2 (RFC 1700)
- Sub-AFI = 1 Unicast
- Sub-AFI = 2 (Multicast for RPF check)
- Sub-AFI = 3 for both Unicast and Multicast
- Sub-AFI = 4 Label
- Sub-AFI= 128 VPN

https://t.me/learningnets

# IPv6 in BGP

**To make BGP-4 available for other network layer protocols, RFC2858 defines multi-protocol extensions for BGP-4:**

1. Enables BGP-4 to carry information of other protocols (e.g MPLS,IPv6)
2. New BGP-4 optional and non-transitive attributes:
    1. MP_REACH_NLRI
    2. MP_UNREACH_NLRI
3. Protocol independent NEXT_HOP attribute
4. Protocol independent NLRI attribute

https://t.me/learningnets

## IPv6 in BGP

# New optional and non-transitive BGP attributes:

- MP_REACH_NLRI (Attribute code: 14)

  Carry the set of reachable destinations together with the next-hop information to be used for forwarding to these destinations

- MP_UNREACH_NLRI (Attribute code: 15)

  Carry the set of unreachable destinations

https://t.me/learningnets

# IPv6 in BGP - IPv6 over IPv4 or IPv4 over IPv6 Sessions

- BGP routers that support IPv6 allow BGP sessions to be set up using IPv6 addresses. MP-BGP speakers tell their neighbors which AFI+SAFI combinations they want to use in the OPEN message at the beginning of a BGP session

- This can lead to the situation where IPv6 routing information is exchanged over IPv4, or IPv4 routing information is exchanged over IPv6

https://t.me/learningnets

## IPv6 in BGP - IPv6 over IPv4 or IPv4 over IPv6 Sessions

- Actually, nothing wrong with that, but it can create problem:

How does the router know what IPv6 next hop address to include in its updates towards an IPv4 neighbor?

To avoid this situation, it’s considered best practice to only exchange IPv4 prefixes over an IPv4 external BGP (EBGP) session and only exchange IPv6 prefixes over an IPv6 EBGP session.

https://t.me/learningnets

# IPv6 in BGP - IPv6 over IPv4 or IPv4 over IPv6 Sessions

- However, internal BGP (IBGP) doesn't update the next hop address, so there are no problems exchanging both IPv4 and IPv6 prefixes over the same IBGP session

- So most networks use the existing IPv4 IBGP sessions to exchange IPv6 prefixes rather than set up a whole new set of IPv6 IBGP sessions

https://t.me/learningnets

# IPv6 in BGP - IPv6 over IPv4 or IPv4 over IPv6 Sessions

- The only downside of this approach is that if then something bad happens to IPv4, the IPv4 IBGP sessions go down and IPv6 is also affected

- If IPv6 had its own IBGP sessions, it may have continued to operate independently from IPv4

https://t.me/learningnets

# VPN Design

In VPN Design Course below technologies will be covered

- WAN VPN Technologies
- Datacenter VPN Technologies

WAN VPN Technologies

1. GRE
2. mGRE
3. IPSEC
4. DMVPN
5. GETVPN
6. LISP

https://t.me/learningnets

Datacenter VPN Technologies

1. EoMPLS
2. VPLS
3. OTV
4. E-VPN
5. PBB-EVPN
6. VxLAN

https://t.me/learningnets

# VPN Theory

- Virtual Private Networks extend the private network over another company's network. It can be setup over another private network such as MPLS or public network such as Internet

- All VPN technologies add extra byte to the packet or frame which increases the overall MTU so the network links should be accommodated to handle bigger MTU

https://t.me/learningnets

# VPN Theory

- VPN works based on encapsulation and decapsulation. For example in GRE, mGRE or DMVPN encapsulate IP packets into another IP packet and VPLS or EVPN encapsulates Layer 2 frame into an MPLS packets

https://t.me/learningnets

- Some VPN technologies support to run routing protocol on top, some doesn’t support. In order to support routing over tunnel, tunnel endpoints should be aware from each other.

For example MPLS Traffic Engineer tunnels don’t support routing protocols to run over since the LSPs are unidirectional which mean Head-end and Tail-end routers are not associated

All WAN technologies except IPSEC, GETVPN and LISP in our list supports routing protocols to run over.

https://t.me/learningnets

# GRE

- Manual point to point tunnels
- Supports routing protocols to run over
- IPv4 and IPv6 can be transported over GRE
- Non-IP protocols such as IPX,SNA etc. can be carried over GRE tunnel as well

https://t.me/learningnets

# GRE

- If there are too many sites which need to communicate with each other, GRE is not scalable. But in Hub and Spoke topologies it can be used since whenever new site is added, only new site and hub should be revisited

- Even though in Hub and Spoke topologies, the configuration can be too long on the Hub site

https://t.me/learningnets

- GRE add 24 bytes to the IP Packet. 4 byte GRE header and 20 bytes new IP header, this increases MTU size of the IP packet

- GRE doesn’t come by default with encryption so in order to encrypt the packet, IPSEC should be enabled over GRE tunnel

https://t.me/learningnets

- Classical use cases of GRE tunnel is over Internet with IPSEC, VRF-lite to carry different VPN information separately, IPv6 tunneling over IPv4 transport

- GRE is used mostly together with IPSEC to support the traffic which is not supported by IPSEC by default. For example IPSEC tunnels don’t support Multicast by default but together with GRE, GRE over IPSEC supports multicast as well.

https://t.me/learningnets

# mGRE – Multipoint GRE

- Allows multiple destinations such as multiple spoke sites to be grouped into a single multipoint interface

- mGRE is a multipoint bidirectional GRE tunnel

https://t.me/learningnets

# mGRE – Multipoint GRE

- Uses only 1 IP subnet so routing table of the routers are reduced which is good for the convergence, troubleshooting and device performance

- Remote endpoint address is not configured that’s why it requires additional mechanisms for tunnel end point discovery. These additional mechanisms can be BGP or NHRP. When it is used with NHRP, solution is called DMVPN which we will see later in detail

https://t.me/learningnets

- Supports IPSEC and routing protocols to run on top
- Supports IPv6 and non-IP protocols
- Don’t require manually configured GRE tunnels that’s why configuration complexity is reduced greatly
- Suitable for full mesh topology such as spoke to spoke tunnels

https://t.me/learningnets

- Use cases are:

- MPLS VPN over mGRE: in this case single mGRE interface is enough to carry multiple VPN information.

- Another use case is VRF Lite over mGRE which requires 1:1 VRF to mGRE interface. So, each VRF requires separate mGRE interface. DMVPN uses this concept.

- mGRE adds 28 bytes of overhead because of the additional 4-byte Key field  which is not typically included in the GRE header when using a point-to-point GRE tunnel

https://t.me/learningnets

# IPSEC

- Provides secure transmission of packets at the IP layer (Not Layer2)

- Packets are authenticated and encrypted

- Provides site to site and remote access VPN topologies

https://t.me/learningnets

# IPSEC

- IPSEC has two modes. Transport mode and tunnel mode.

- The difference between the two is that in transport mode; only the IP payload is encrypted, in tunnel mode ; the entire IP datagram(including header) is encrypted.

- IPSEC is point to point tunneling mechanism, thus in large scale implementation it can not scale.

https://t.me/learningnets

- It is mainly used to provide encryption for the data over the Public Internet.

This data shouldn’t be real-time, packet loss, jitter and delay sensitive since there is no QoS SLA over the Public Internet, it is best effort

- IP Multicast is not supported over IPSEC VPNs.

https://t.me/learningnets

- QoS is supported through TOS byte preservation

- Routing protocols don’t run over IPSEC VPNs, it requires additional GRE encapsulation (GRE over IPSEC)

- IPSEC VPN is a standard protocol which supported by every vendor

https://t.me/learningnets

# DMVPN

- Point to Multipoint and Multipoint to Multipoint Tunneling technology

- Works based on two standard technologies; NHRP and mGRE (Multipoint GRE)

https://t.me/learningnets

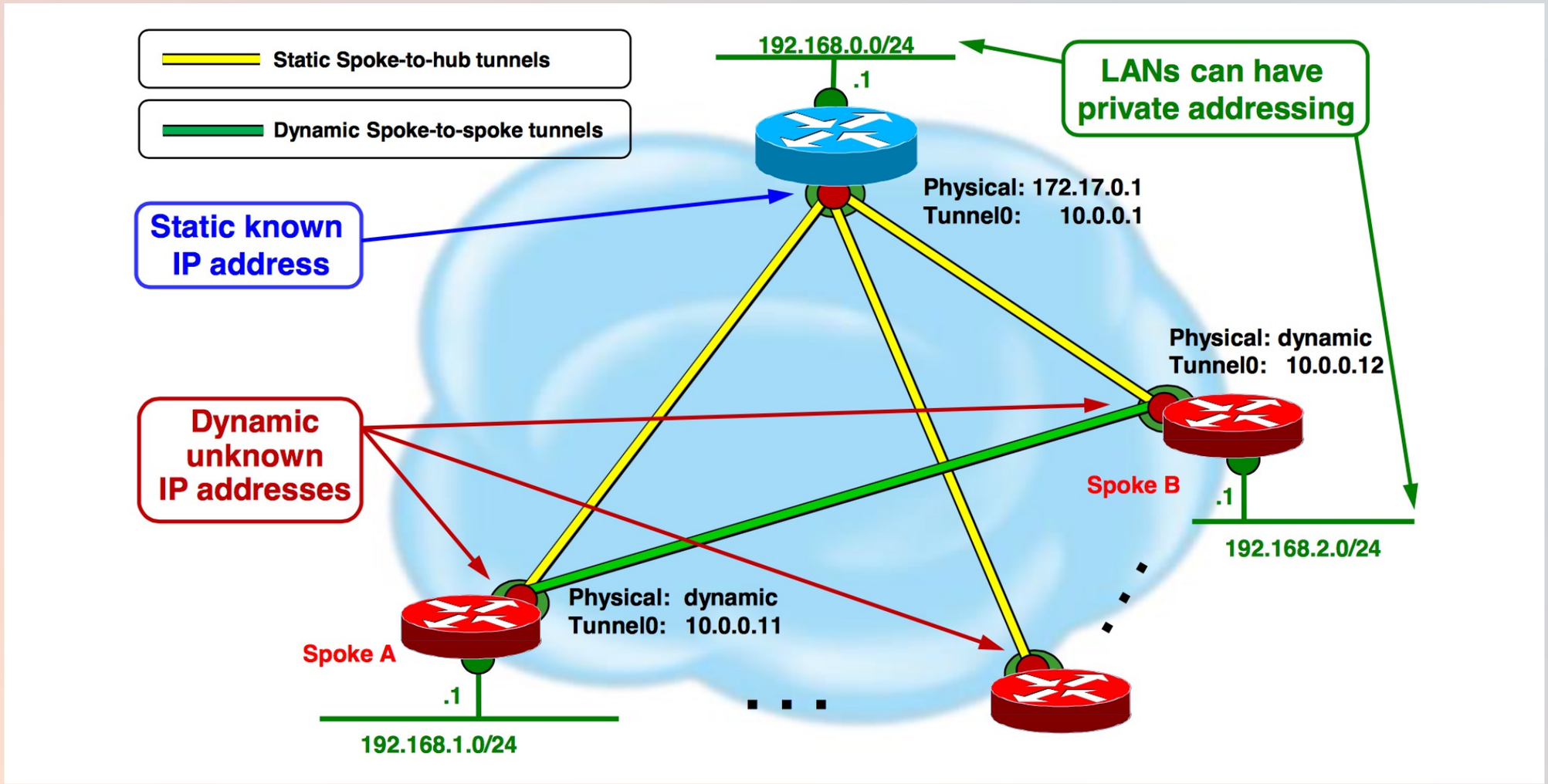
# DMVPN

- NHRP is used to map NBMA/Underlay address to the Tunnel/Overlay address

- mGRE tunnel interface doesn’t require manual tunnel destination configuration

- Local address/LAN subnets are advertised over the overlay tunnels through the routing protocols

https://t.me/learningnets




https://t.me/learningnets

- Multicast is not natively supported, but ingress replication is done at the HUB, thus Multicast support of DMVPN can create scalability problem in Large scale design

- Per tunnel QoS is supported, this prevent hub to overutilized the spoke bandwidth

https://t.me/learningnets

- All routing protocols except IS-IS can run over DMVPN

- IS-IS can not run since DMVPN only support IP protocols and IS-IS works based on Layer 2

https://t.me/learningnets

- Invented by Cisco but other vendors provide the capability with different name, Cisco’s implementation is not compatible with the other vendors

- IPSEC is optional and can be implemented as native point to point IPSEC or GETVPN (Group key)

- DMVPN can carry IPv4 and IPv6 unicast and multicast packets over the overlay tunnels

https://t.me/learningnets

- DMVPN also can work over IPv4 and IPv6 private and public infrastructure transport such as Internet or MPLS VPN

- There are three phases of DMVPN, Phase 1, Phase 2 and Phase 3

- In all the DMVPN Phases, overlay routing protocol control packets pass through the HUB. There is no spoke-to-spoke control plane traffic.

https://t.me/learningnets

# DMVPN Phase -1

- Spokes use Point to Point GRE but Hub uses a multipoint GRE tunnel

- mGRE tunnel interface simplifies configuration greatly on the Hub

- Spokes have to specify the tunnel destination as Hub since they run P2P GRE tunnel not mGRE in DMVPN Phase 1

https://t.me/learningnets

# DMVPN Phase -1

- Summarization is allowed from the Hub down to spokes. The Hub can only send default route as well ,since the remote spokes next hop is not preserved by the Hub

- Hub changes the IGP next hop to itself hence spoke to spoke traffic passes through the Hub

- Spoke to spoke tunnel cannot be created, that’s why DMVPN phase 1 doesn’t provide full mesh connectivity

https://t.me/learningnets

- DMVPN Phase 2

- Spoke to spoke dynamic on demand tunnels are first introduced in Phase 2.

- In contrast to Phase 1, mGRE (Multipoint GRE, not Multicast) interface is used in Phase 2 on the Spokes

- Thus, spokes don’t require tunnel destination configuration under the tunnel interface and tunnel mode is configured as ” Multipoint Gre”.

- Spoke to spoke traffic doesn’t have to go through the HUB. Spokes can trigger on demand tunnel between them.

https://t.me/learningnets

- The biggest disadvantage of Phase 2 is, each spoke has to have all the Remote – LAN subnets of each other since the Hub, preserves the next hop addresses of the spokes.

- Thus, spokes have to have a reachability to the tunnel addresses of each other.

- This disallows the summarization or default routing from Hub down to the spokes.

- This is a serious design limitation for the large scale DMVPN networks.

https://t.me/learningnets

- For the distance vector protocols, Split horizon needs to be disabled and “no next-hop self” should be enabled on the HUB.

https://t.me/learningnets

# DMVPN Phase 3

- Spoke to spoke dynamic tunnels are allowed

- Spokes don’t have to have the next hop of each other’s private address for the local subnets.

- An NHRP redirect message is sent to the spokes to trigger spoke to spoke tunnels. Hub provides the Public/NBMA addresses of the spokes to each other.

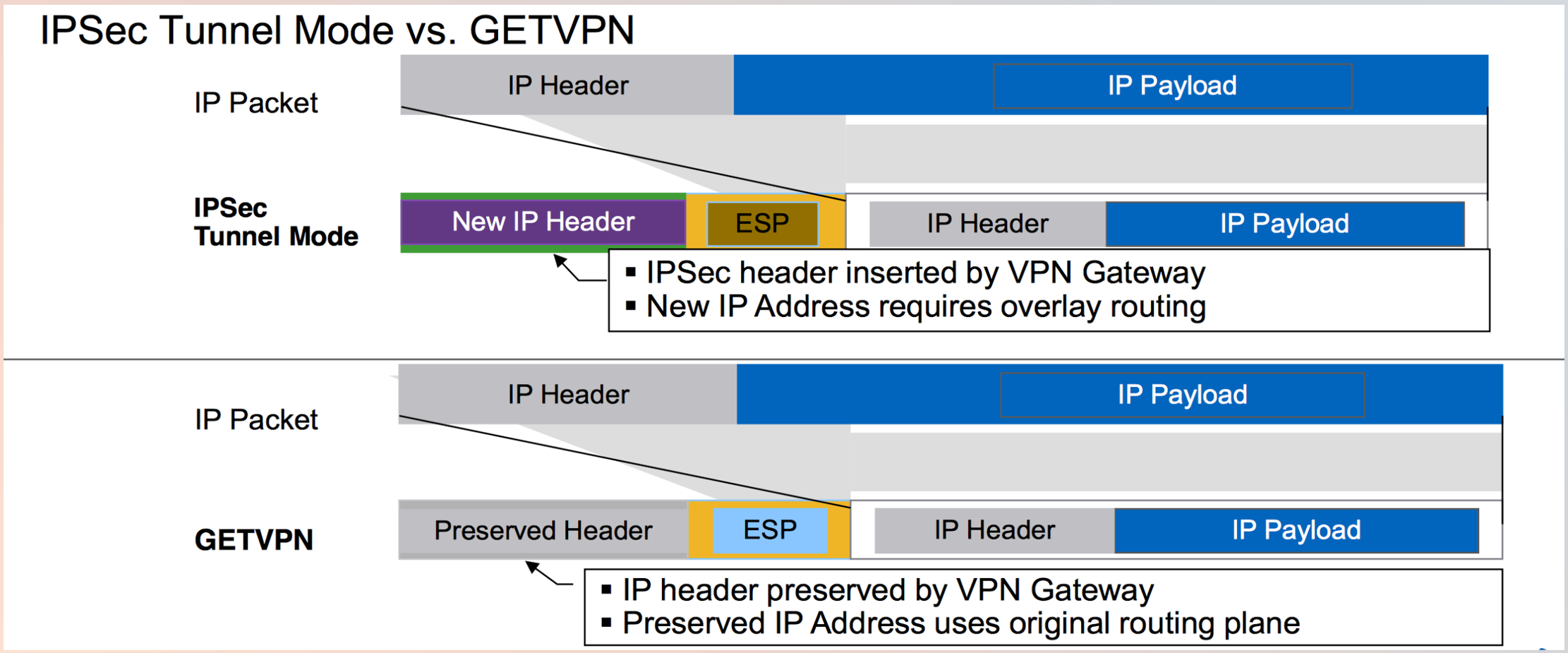
https://t.me/learningnets

# DMVPN Phase 3

- Since the next hop in the routing table of the spokes is HUB's tunnel address, spokes don't have to have the specific next hop information of each other.
- This allows summarization and default routing in Phase 3.
- Hub can send a summary or just a default route down to the spokes.
- Hence, Phase 3 is extremely scalable

https://t.me/learningnets

# GETVPN




https://t.me/learningnets

- It can run on top of any routing protocol

- It doesn’t support overlay routing protocol since there is no tunnel to run routing protocol over

- Excellent Multicast support compared to DMVPN since there is no replication but it uses native multicast

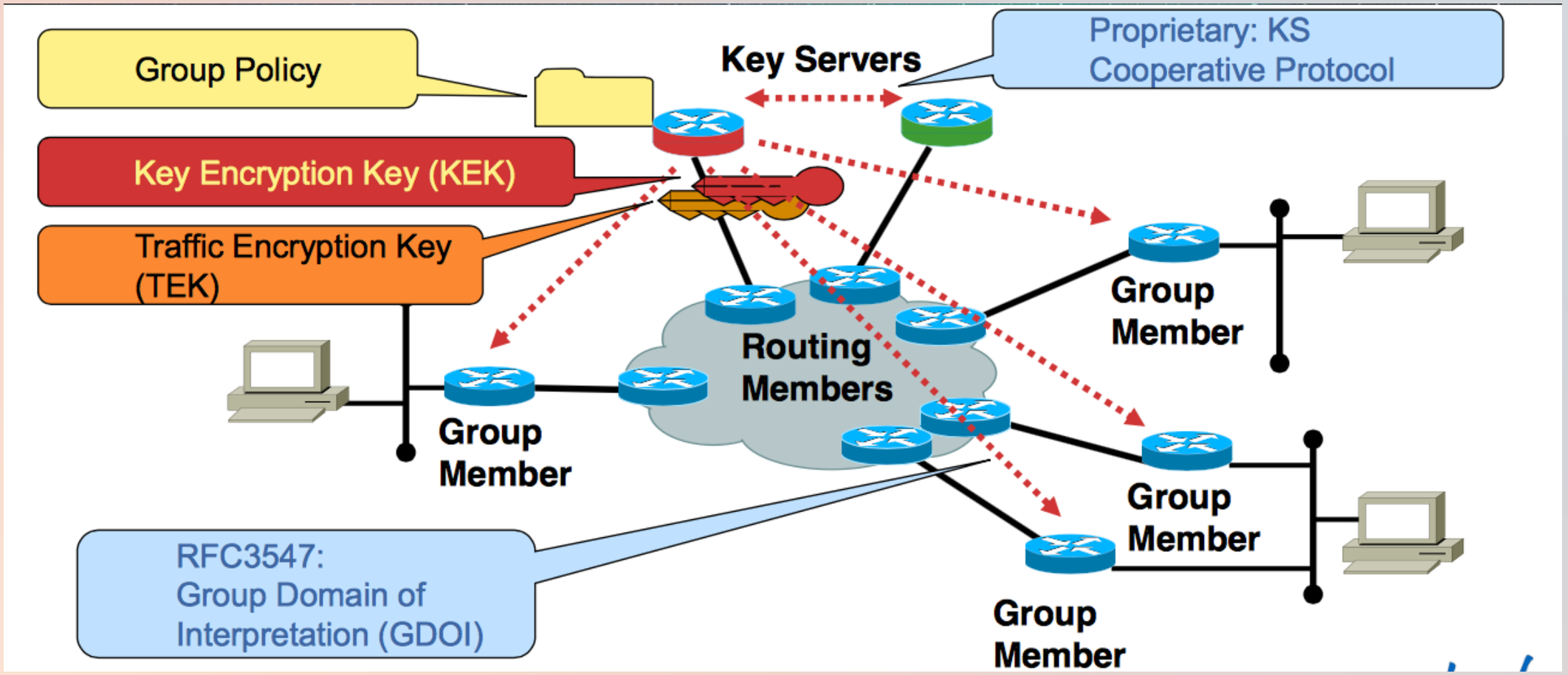
https://t.me/learningnets

- Doesn’t support Non-IP Protocols

- Uses Group Key mechanism instead of Point to Point Keys which are used in DMVPN by default, thus from the security scalability point of view, GETVPN is much scalable technology compare to the DMVPN

https://t.me/learningnets

- Group Members (GMs) “register” via GDOI with the Key Server (KS)
  - KS authenticates & authorizes the GMs
  - KS returns a set of IPSEC SAs for the GMs to use
  - GMs exchange encrypted traffic using the group keys
  - Traffic uses IPSec Tunnel Mode with  IP address preservation




https://t.me/learningnets

| orhanergun.net | DMVPN | GETVPN |
|---|---|---|
| Scalability | Scalable | Much more scalable than DMVPN |
| Working on Full Mesh topology | Permanent hub and spoke tunnels and on demand spoke to spoke tunnels, it works but limited scalability | It works perfectly if the underlying routing architecture is full mesh topology, GET VPN needs underlay routing |
| Working on Hub and Spoke | Works very well | Works very well |
| Suitable on private WAN | Yes | Yes |
| Suitable over Public Internet | Yes | No. GETVPN cannot run over Public Internet because of IP header preservation |
| End point discovery | To setup the Mgre tunnels uses underlay routing, for the private address discovery uses NHRP (Next hop Resolution Protocol) | It uses underlay routing to create VPN, there is no overlay tunnels |
| Tunnel Requirement | Yes, it uses Mgre(Multi Point GRE) tunnels to create overlays | It is tunnelles VPN, uses underlaying routing to encrypt the data between endpoints |
| Standard Protocol | No,Cisco proprietary | No,Cisco proprietary but Juniper also supports the same idea with Group VPN feature |
| Stuff Experince | Not well known | Not well known |
| Overlay Routing Protocol Support | Except IS-IS other routing protocols are supported, IS-IS runs on top of Layer 2 but only IP protocols can run over DMVPN | It is tunnelles VPN so routing protocols cannot run on top of GETVPN but it requires underlaying routing protocols to setup the communication |
| Required Protocols | NHRP and Mgre | GDOI and ESP |
| QoS Support | Good, can support per tunnel QoS which uses shaping on the DMVPN Hub to protect capacity and SLA | Good, it uses underlaying network's QoS architecture,in addtion to queueing,shaping at the GET VPN Group Members to protect SLA is enabled |
| Multicast Support | Multicast over the tunnel is handled at the DMVPN Hub. Hub replicates multicast traffic which is not efficiend | Native multicast support.Multicast replication is done in the network, doesn't need Hub device to replicate. Multicast MDTs (Source , Shared) are used in the traditional way, so multicast handling of GETVPN is much better than DMVPN |
| Security | Point to Point IPSEC SA | Multipoint to Multipoint IPSEC SA |
| Resource Requirement | More | Less |
| IPv6 Support | Yes,it can be setup over IPv6 transport or it can carry IPv6 payload. So IPv6 over DMVPN and DMVPN over IPv6 both are possible | Yes |
| Default Convergence | Slow | Fast |
| Can run over other VPN ? | DMVPN is already tunneled VPN technology so only routing is enough, it doesn't make sense to run tunnel over tunnel | GETVPN can run over DMVPN since GETVPN is tunnelless technology, use case of GETVPN over DMVPN is to carry private addressing over Internet. Most common use case of GETVPN is over MPLS VPN or VPLS since both VPN technologies are full mesh by default and GET VPN provides very good scalability for encryption |

https://t.me/learningnets

# LISP-Locator Identity Separation Protocol

- Provides IP mobility with the pull based approach. Tunnel end points are known by the centralized database and the routers pull the information from the database when it is needed.

- IP in IP encapsulation mechanism which uses underlaying routing infrastructure, thus, can provide any to any connectivity

https://t.me/learningnets

# LISP-Locator Identity Separation Protocol

- There is also attempt to provide MAC in IP encapsulation

- It is an Experimental RFC (RFC 6830) but invented by Cisco.

https://t.me/learningnets

- Similar to DNS infrastructure. Mapping Database keeps the host to gateway mapping information.

- Host space is called EID (Endpoint Identifier) and its gateway is called RLOC (Routing Locator)

- Most interesting LISP use cases are VM Mobility in the Datacenter, BGP Traffic Engineering at the Internet Edge and IPv6 transition (IPv6 over IPv4 or IPv4 over IPv6 infrastructure)

- From the security point of view MR/MS is a target point

https://t.me/learningnets

# LISP Device Roles

## ITR – Ingress Tunnel Router

- Receives packets from site-facing interfaces

- Encapsulates to remote LISP sites, or natively forward to non-LISP sites

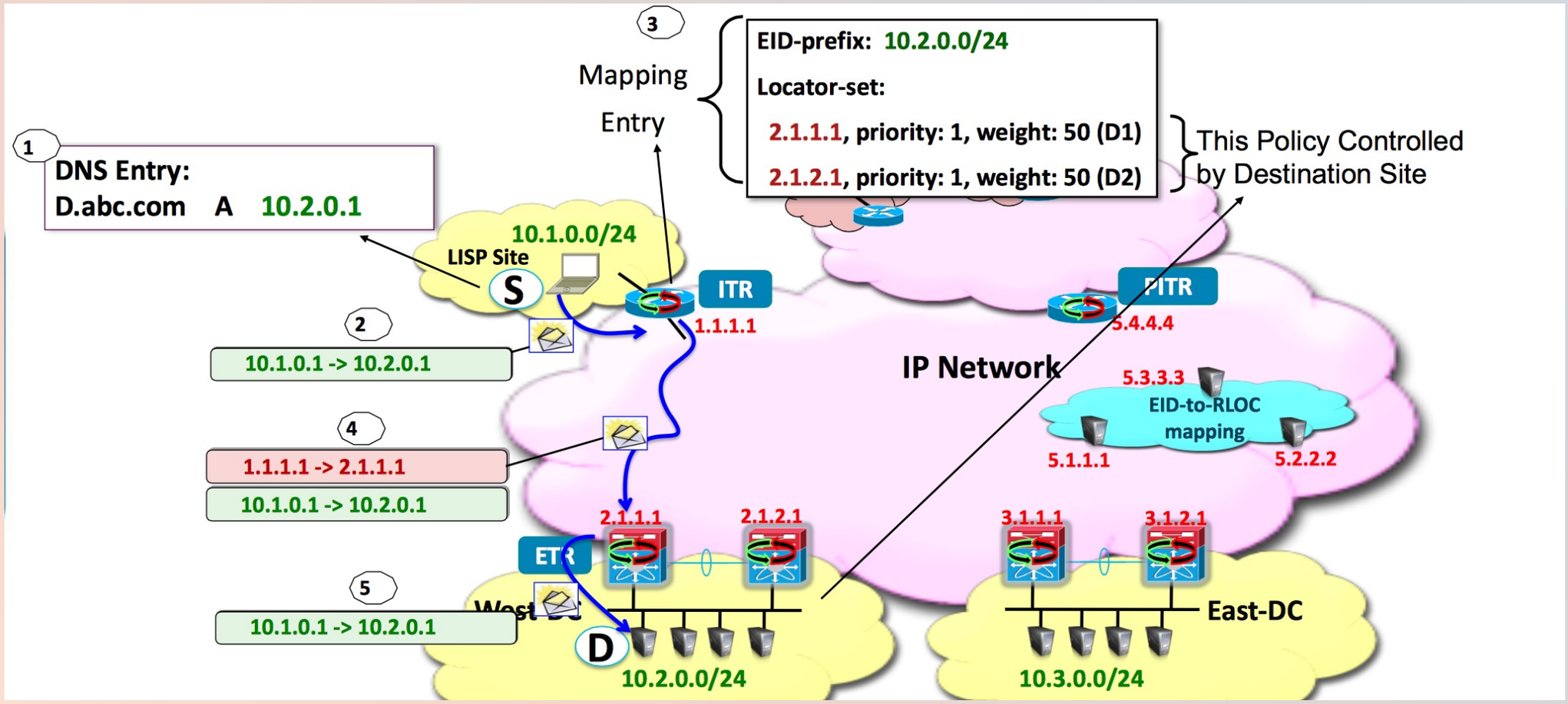
## ETR – Egress Tunnel Router

- Receives packets from core-facing interfaces
- De-cap and deliver packets to local EIDs at the site

- EID to RLOC Mapping Database

- Contains RLOC to EID mappings

https://t.me/learningnets




https://t.me/learningnets

# SD-WAN

https://t.me/learningnets

# SD-WAN Agenda

- Why SD-WAN?
- Traditional WAN vs. SD-WAN
- MEF SD-WAN
- SD-WAN Components (Architecture – Controller, Edges, Orchestration)
- Dynamic Path Selection
- WAN Optimization

https://t.me/learningnets

# SD-WAN Agenda

- De-duplication
- Forward Error Correction
- Quality of Service
- ZTP
- Cloud On--Ramp

https://t.me/learningnets

# SD-WAN Software Defined Wide Area Network – What is SD-WAN ?

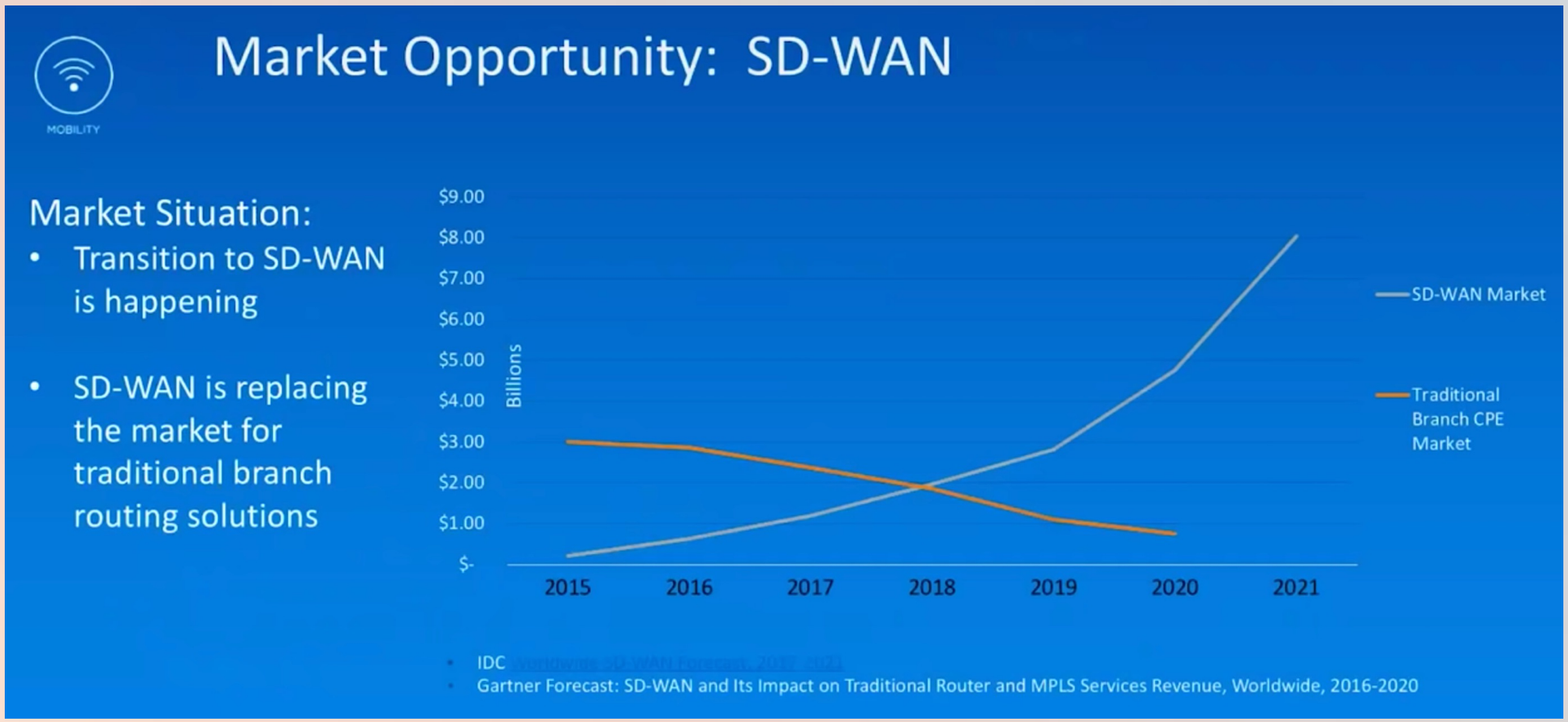
- SD-WAN is an acronym for software-defined networking in a wide area network (WAN)

- SD-WAN simplifies the management and operation of a WAN by decoupling the networking hardware from its control mechanism

- This concept is similar to how software-defined networking implements virtualization technology to improve data center management and operation

https://t.me/learningnets

# SD-WAN Market and Estimated Market Size




https://t.me/learningnets

# SD-WAN – Why SD-WAN ?

- A key reason of SD-WAN is to allow companies to build higher-performance WANs using lower-cost and commercially available Internet access, enabling businesses to partially or wholly replace more expensive private WAN connection technologies such as MPLS

https://t.me/learningnets

# Lots of buzzwords! Which one is SD-WAN?

Application based routing

Multipath routing

Quality of Experience

Centralized orchestration

Visibility and Analytics

Automation

Encrypted tunneling

Access independent

ZeroTouchDeployment

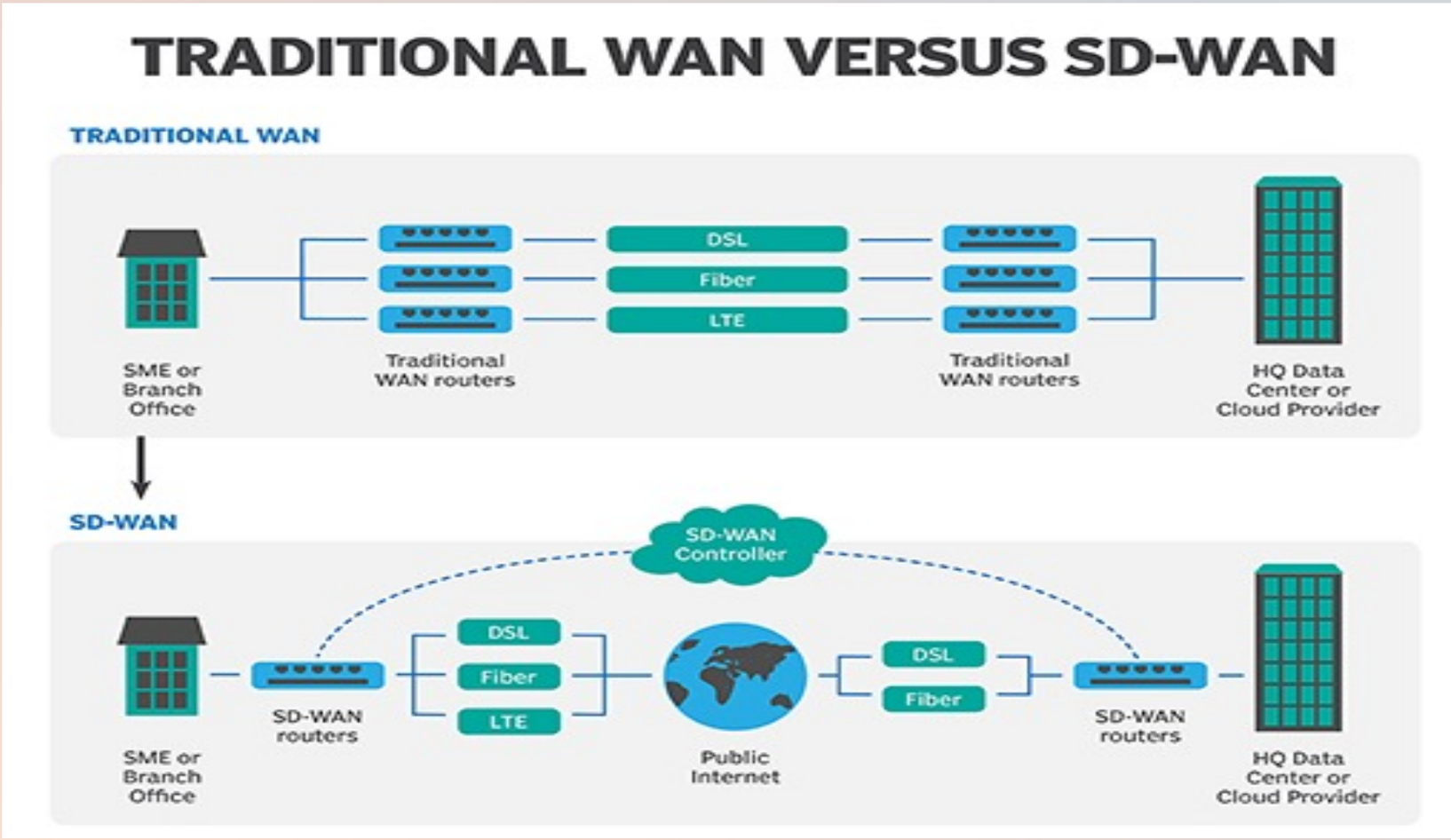
Security

Reduce # HW on site

Cloud integration

https://t.me/learningnets

# Traditional WAN vs. SD-WAN




https://t.me/learningnets

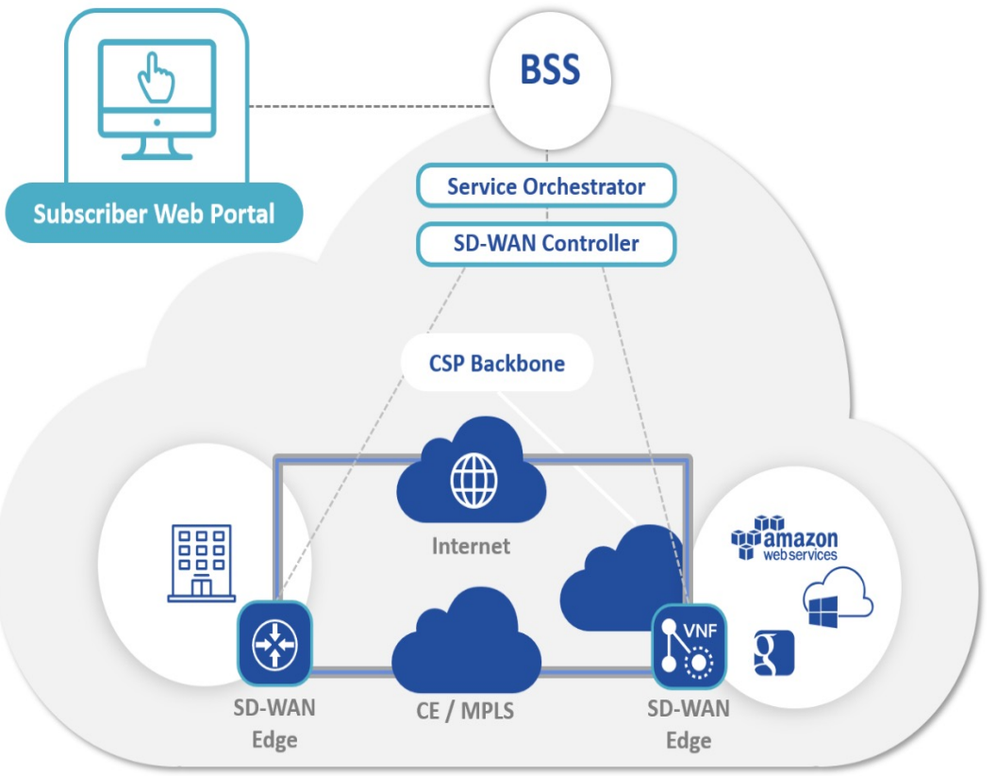
# Why SD-WAN – MEF Definition

- MEF’s SD-WAN service definition standard describes requirements for an application-aware, over-the-top WAN connectivity service that uses policies to determine how application flows are directed over multiple underlay networks regardless of the underlay technologies or service providers who deliver them

https://t.me/learningnets

# MEF SD-WAN Service Definition

## SD-WAN Service Constructs




**SD-WAN Edge**

Physical or virtual

**SD-WAN Controller**

Centralized management of SD-WAN edges & gateways

**Service Orchestrator**

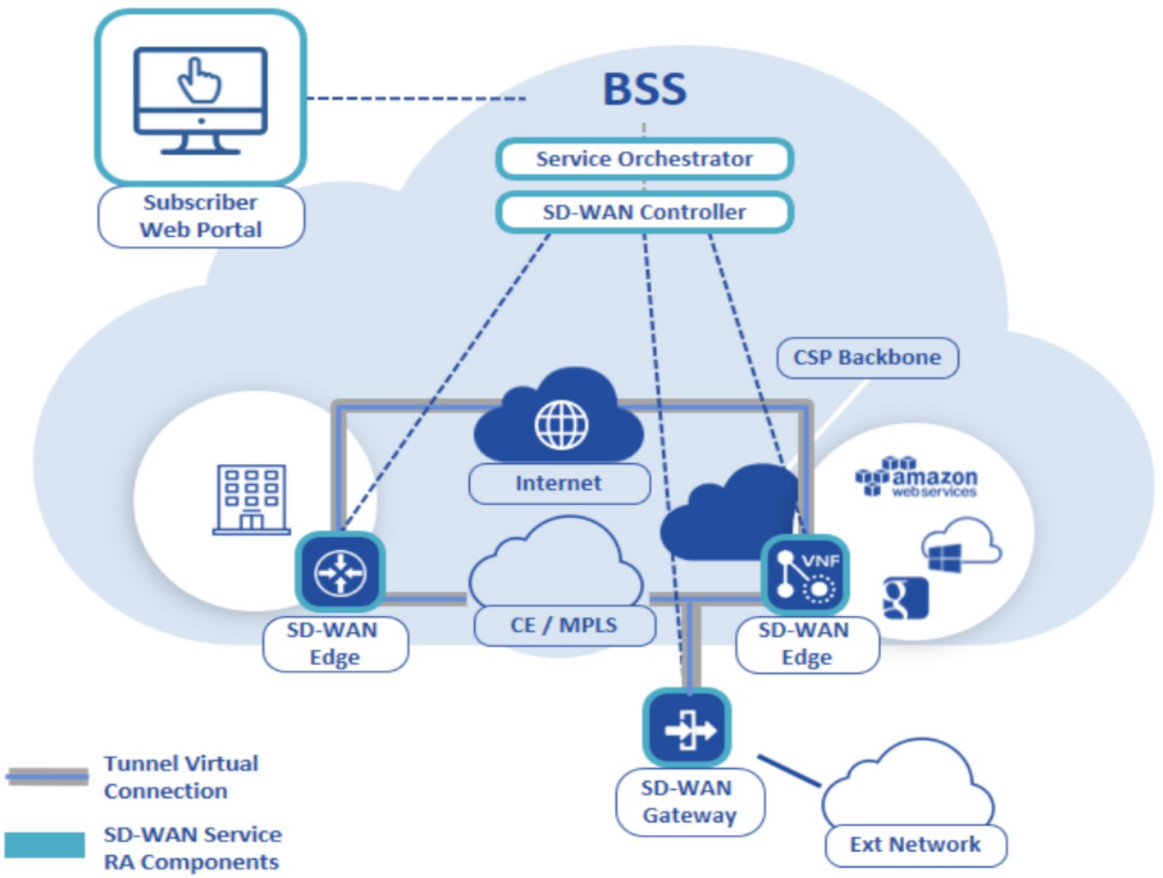
Lifecycle Service Orchestration of SD-WAN and other services

**Subscriber Web Portal**

Subscriber service ordering and modification

https://t.me/learningnets

# SD-WAN Components – Common Architectural Elements




SD-WAN Service Components. Source: MEF

**SD-WAN 5 Common Elements:**

**Edge**
**Controller**
**Gateway**
**Orchestrator**
**Web Portal**

https://t.me/learningnets

# SD-WAN Components – Common Architectural Elements- SD-WAN Edge

- SD-WAN Edge, which could either be formed by a physical device, such as a CPE, or by a virtual CPE that's based on a virtual network function (VNF)

- The SD-WAN Edge performs several critical functions

Source: cisco.com

https://t.me/learningnets

# SD-WAN Components – Common Architectural Elements- SD-WAN Edge

- It acts as the security-policy enforcer and conducts WAN optimization tasks that include data deduplication, compression, and packet buffering

- It also creates and removes encrypted tunnels over underlay networks, whether it’s a wired or wireless connection

https://t.me/learningnets

# SD-WAN Components – Common Architectural Elements- SD-WAN Edge

- Since SD-WAN Edges often connect to public Internet WANs, they would include, at a minimum, some NAT and firewall capabilities

- Most of the time actual cost of the deployment comes from SD-WAN Edges as there are much more Edge device in the network than Centralized Core Devices (Gateways , Controllers , Orchestrator)

https://t.me/learningnets

# SD-WAN Controller

- SD-WAN Controller centralizes management to the SD-WAN Edge and to the SD-WAN Gateway

- The SD-WAN Controller provides physical or virtual device management for all SD-WAN Edges and SD- WAN Gateways associated with the controller

- This includes, but is not limited to, configuration and activation, IP address management, and pushing down policies onto SD-WAN Edges and SD-WAN Gateways

https://t.me/learningnets

# SD-WAN Controller

- The SD-WAN controller maintains connections to all SD-WAN Edges and SD-WAN Gateways to identify the operational state of SD- WAN tunnels across different WANs and retrieve QoS performance metrics for each SD-WAN tunnel

- These metrics are used by the Orchestrator

https://t.me/learningnets

# SD-WAN Orchestrator

- The Service Orchestrator provides the service management of the SD-WAN service lifecycle including service fulfillment, performance, control, assurance, usage, analytics, security and policy

- For example, the Service Orchestrator is responsible for configuring the end-to-end SD- WAN managed service between SD-WAN Edges and SD-WAN Gateways over one or more underlay WANs, e.g., Internet and MPLS, setting up application-based forwarding over WANs based on security, QoS or business or intent-based policies

https://t.me/learningnets

# SD-WAN Gateway

- The SD-WAN Gateway is a special case of an SD- WAN Edge that also enables sites interconnected via the SD-WAN to connect to other sites interconnected via alternative VPN technologies, e.g., CE or MPLS VPNs

https://t.me/learningnets

# SD-WAN Gateway

- There are two ways to deliver an SD-WAN service to sites connected via another VPN service.

- One way requires an SD- WAN Edge to be placed at each subscriber site connected to the VPN service so SD-WAN tunnels can be created over the VPN

https://t.me/learningnets

# SD-WAN Gateway

- Another way is to use an SD-WAN Gateway

- In this scenario, an SD-WAN Gateway initiates and terminates the SD-WAN tunnels like an SD-WAN Edge and initiates and terminates VPN connections to and from sites interconnected by the VPN

- This approach enables sites interconnected via SD-WAN and other VPN technology domains to intercommunicate

https://t.me/learningnets

# SD-WAN Gateway

- This approach does not require SD-WAN Edges to be placed at each VPN site to achieve interconnectivity

- However, SD-WAN service capabilities such as application-based traffic forwarding over multiple WANs or QoS and Security policy management will not be available at the MPLS VPN sites because they do not have SD-WAN Edges which perform these functions

https://t.me/learningnets

# SD-WAN Web Portal

- Subscriber Web Portal is added to the enterprise's existing managed services portal

- It works in conjunction with the service orchestrator to monitor the SD-WAN as a service

- The MSP or CSP typically integrates the Subscriber Web Portal for the SD-WAN managed service into their existing customer portal used for other managed services

https://t.me/learningnets

# SD-WAN Key Characteristics

1. The ability to support multiple connection types, such as MPLS, Last Mile Fiber Optical Network or through high speed cellular networks e.g. 4G LTE and 5G wireless technologies

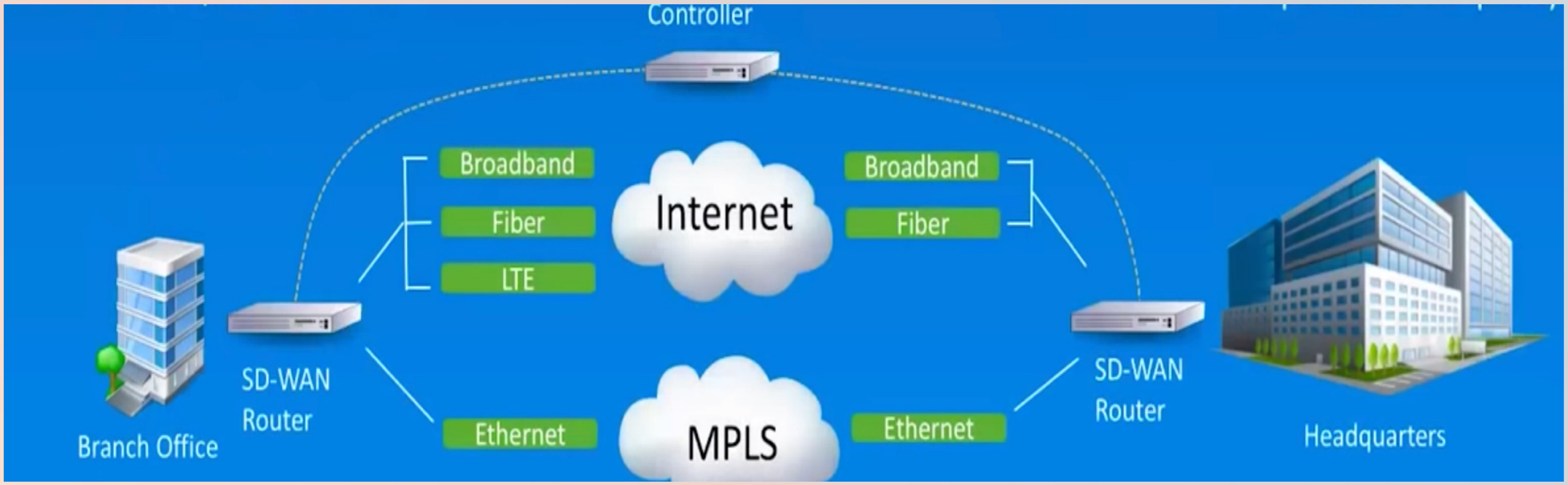
2. The ability to do dynamic application aware path selection, for load sharing and resiliency purposes

3. A simple interface that is easy to configure and manage

4. The ability to support VPNs, and third-party services such as WAN optimization controllers, firewalls and web gateways

https://t.me/learningnets

# SD-WAN – Different Transport Mechanisms




https://t.me/learningnets

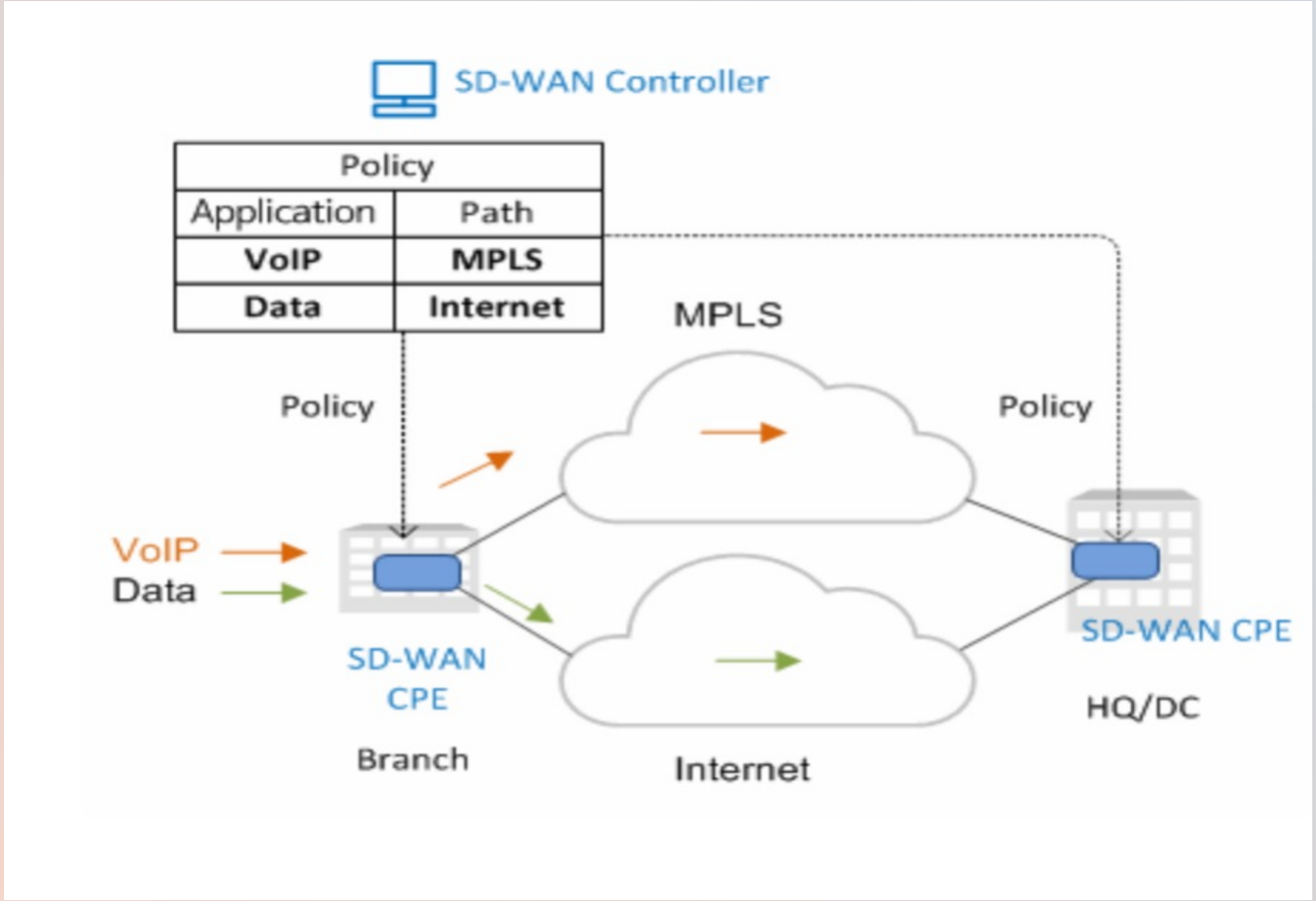
# SD-WAN – Dynamic Path Selection

- This feature ensures traffic uses the best path depending on the business need, such as mission-critical and delay-sensitive applications

https://t.me/learningnets

# SD-WAN – Dynamic Path Selection

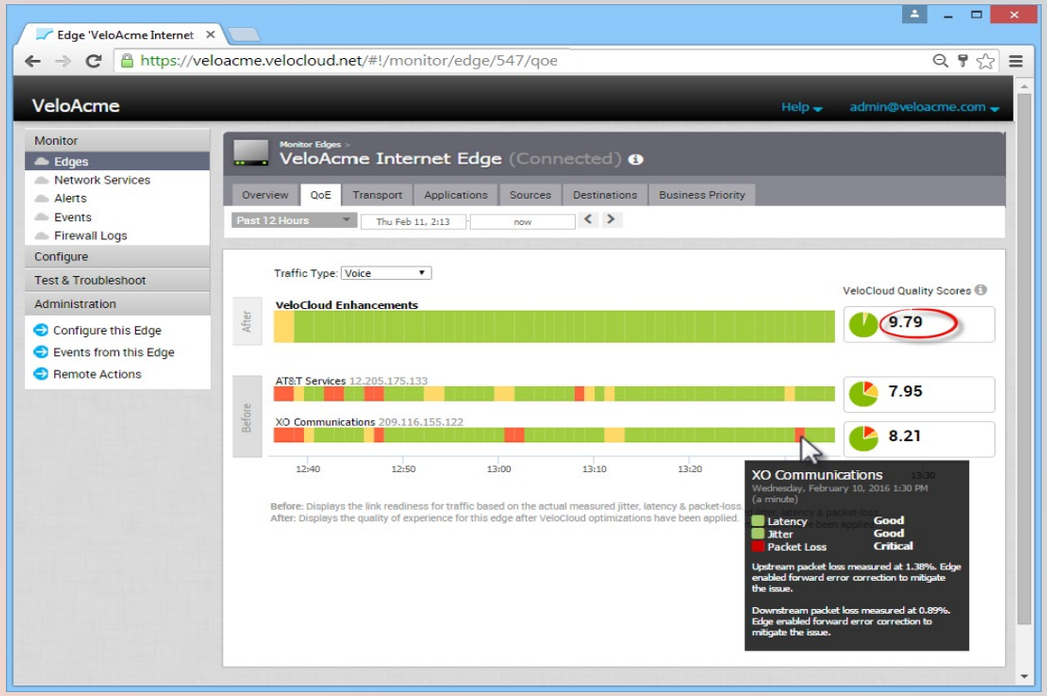
- SD-WAN solution requires a path selection/control solution that allows for each application to dynamically switch their paths in real-time, in response to network conditions, rather than sticking to one particular underlay!




https://t.me/learningnets

# SD-WAN – Simpler Management Compare to Legacy WAN

- GUI provides simpler management, reduced troubleshooting time, mass deployment and update , centralized monitoring and so on.




Source : VeloCloud SD-WAN

https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

- WAN optimization accelerates application traffic by overcoming latency and reducing the amount of data traversing the WAN by applying techniques like, deduplication, compression and caching to dramatically increase the amount of available bandwidth

https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

- Most SD-WAN implementations offer a way to encrypt your branch-to-branch corporate traffic using IPSEC which protects the data in transit

- Because most SD-WAN vendors offer IPsec, it’s common to think that SD-WANs are inherently secure

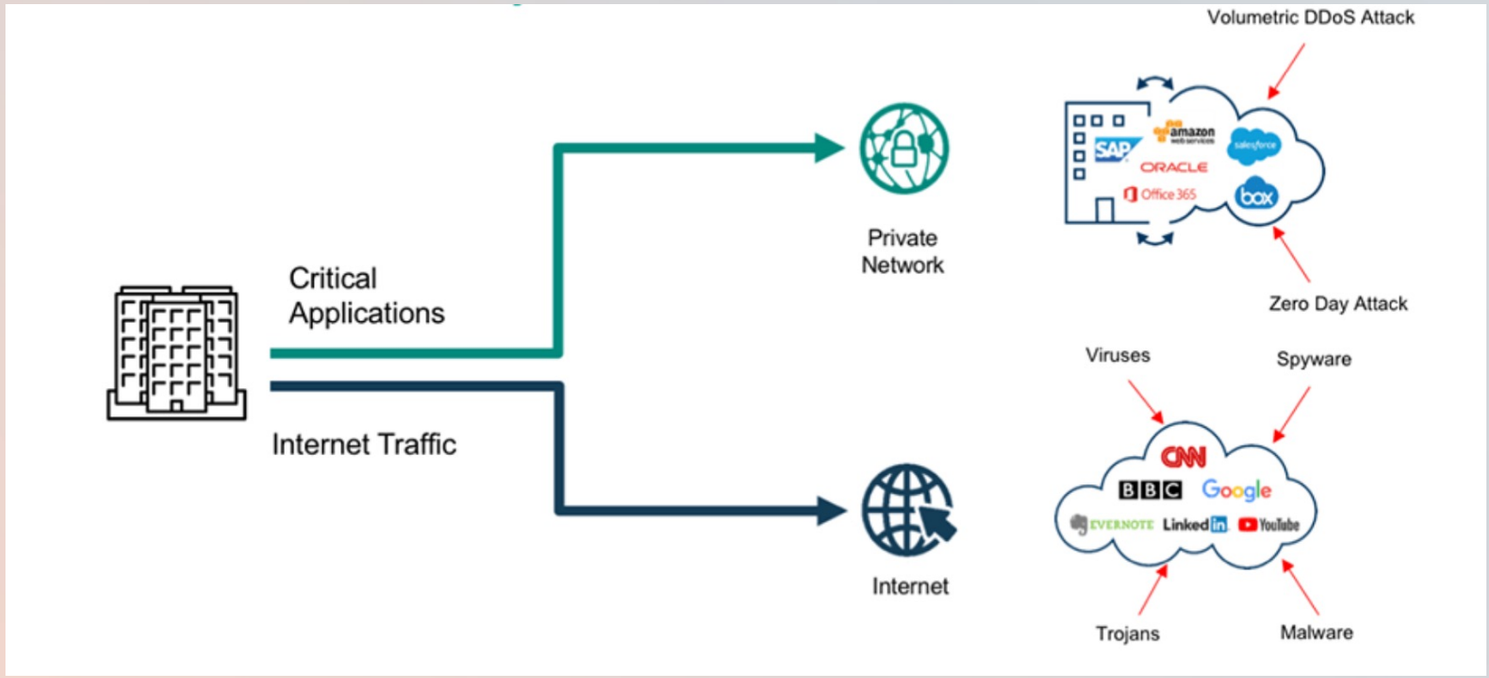
https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

- It’s true that IPsec handles protecting the data as it traverses the network

- But it has no impact on DDOS protection, man-in-the-middle and malware for direct branch-to-cloud traffic

- Centralized security control should be re thought when it comes to SD-WAN security




Source : ARYAKA

https://t.me/learningnets

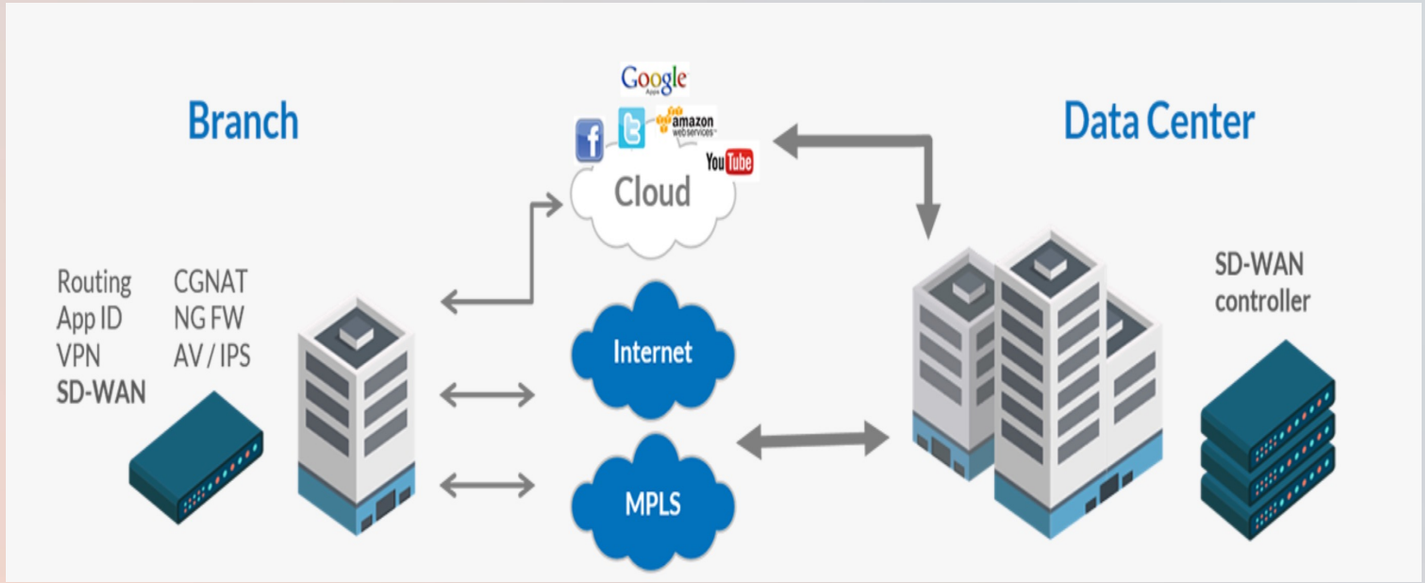
# SD-WAN – Wan Optimization – Security and Other Services

- For example, you still need stateful firewall capabilities between the public Internet and your WAN edge device to grant or deny access

https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

- Most NGFWs also comes with a variety of UTM functions, including intrusion detection and prevention (IDS/IPS), quarantining or otherwise deflecting detected malware, and web filtering, which knows about risky Internet sites and prevents your users from visiting them




Source : VERSA

https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

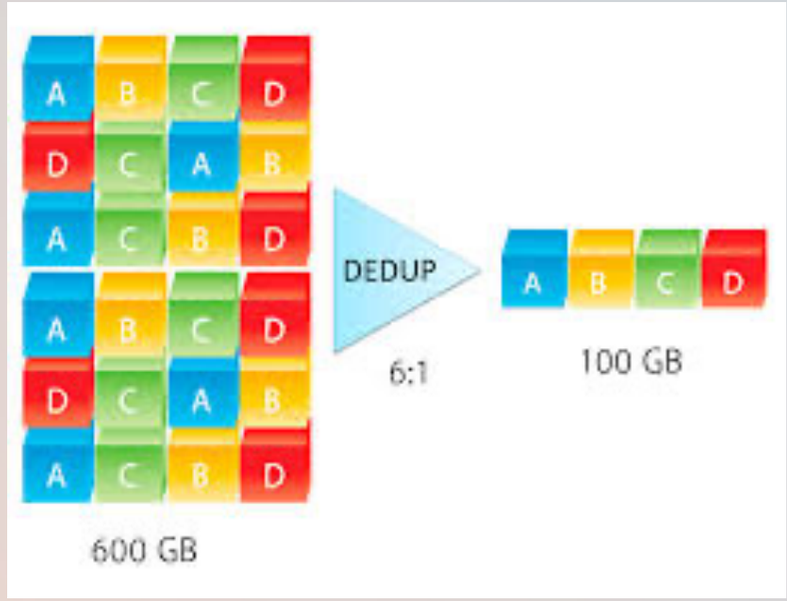
- Since every branch constitutes a WAN edge with exposure to the Internet, you may need all these capabilities at each branch site!

https://t.me/learningnets

# SD-WAN – Wan Optimization – Deduplication and Compression

- Deduplication analyzes blocks of data, looking for repetition

- It replaces multiple copies of data with references to a single, compressed copy, thereby reducing the amount of capacity needed




https://t.me/learningnets

# SD-WAN – Wan Optimization – Deduplication and Compression

- Data Deduplication (dedupe) provides storage savings by eliminating redundant blocks of data

- Storage capacity reduction is accomplished only when there is redundancy in the data set

- This means the data set must be comprised of multiple identical files or files that contain a portion of data that is identical to the content found in other files

https://t.me/learningnets

# SD-WAN – Wan Optimization – Deduplication and Compression

- Data compression reduces the number of bits required to represent the information

- Compressing large files into smaller bits allows users to store more data and also it makes data transmission much quicker and easier

- Compressed data must be decompressed so that the original data can be extracted and the amount a document is compressed is measured by something called the compression ratio

https://t.me/learningnets

# SD-WAN – Wan Optimization – Deduplication and Compression

- Unlike deduplication, compression is not concerned with whether a second copy of the same block exists, it simply wants to store the most efficient block on the storage

- Examples of common file level compression that we use in our day-to-day lives include MP3 audio and JPG image files

https://t.me/learningnets

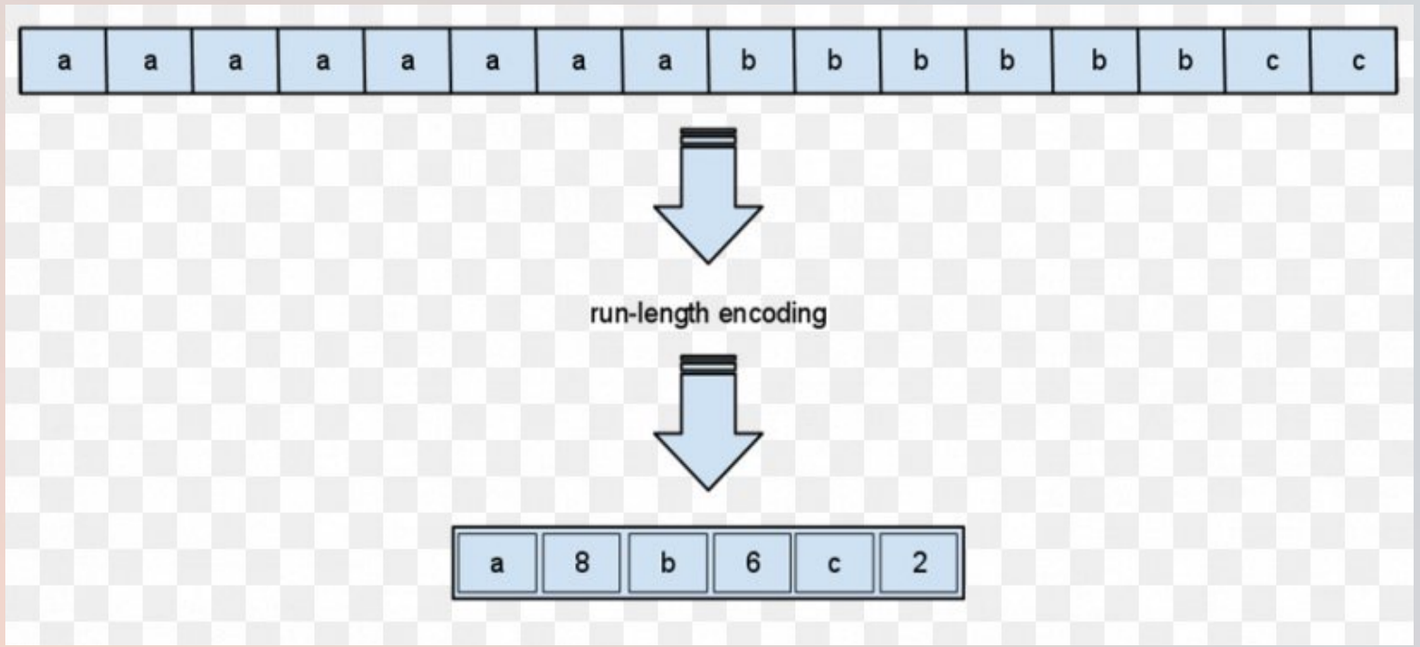
# SD-WAN – Wan Optimization – How Data Compression Work?

- Entropy Encoding is one of the techniques for compression
- You might start with a string like:  AABCABBCABACBAAACBCCAABAAACBAA
- You might notice that some letters appear more than others - A appears about 2x as much as B and C, and the other letters don't appear at all
- Using that information, you can choose an encoding that represents the characters in the string with less information, e.g., A may be encoded using binary 0, while B and C are assigned 10 and 11respectively. If you were originally using 8 bits per character, that is a big savings

https://t.me/learningnets

# SD-WAN – Wan Optimization – How Data Compression Work?

- Another encoding schema can be Run-length encoding




https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

- Packet loss occurs when network congestion or problems in the physical infrastructure cause packets to be lost during transmission

- It’s expressed as a percentage of packets

https://t.me/learningnets

# SD-WAN – Wan Optimization – Security and Other Services

- Packet loss is addressed by some WAN optimization appliances using forward error correction (FEC) that allows receiving stations to automatically regenerate lost packets without requiring transmission

- Let’s have a look at Forward Error Correction

https://t.me/learningnets

# SD-WAN - Forward Error Correction

- For some applications it is necessary to have good error protection

- Sometimes, it will be impossible for the receiver to communicate back with the sender to check for errors in the received packages

https://t.me/learningnets

# SD-WAN - Forward Error Correction

- Some algorithms are made for this kind of situation as for example in a multiple receiver communication

- They use a forward error correction, which is based on the addition of redundant bits over the bit stream of data

https://t.me/learningnets

# SD-WAN - Forward Error Correction

- A simple example of forward error correction is *(3,1) repetition code.* In this example, each bit of data is sent three times and the value or meaning of the message is decided upon majority vote. The most frequently sent bit is assumed to be the value of the message (see table below)

| Triplet received | Interpreted as |
|---|---|
| 000 | 0 (error free) |
| 001 | 0 |
| 010 | 0 |
| 100 | 0 |
| 111 | 1 (error free) |
| 110 | 1 |
| 101 | 1 |
| 011 | 1 |

https://t.me/learningnets

# SD-WAN - Good to have capabilities with SD-WAN

- Some of these features might be good to have for some companies and must have for others depends on the application requirements and the constraints

- Quality of Service , Zero Touch Deployment , Global Coverage , Vendor POC support , Cloud Enablement

https://t.me/learningnets

# SD-WAN - Quality of Service

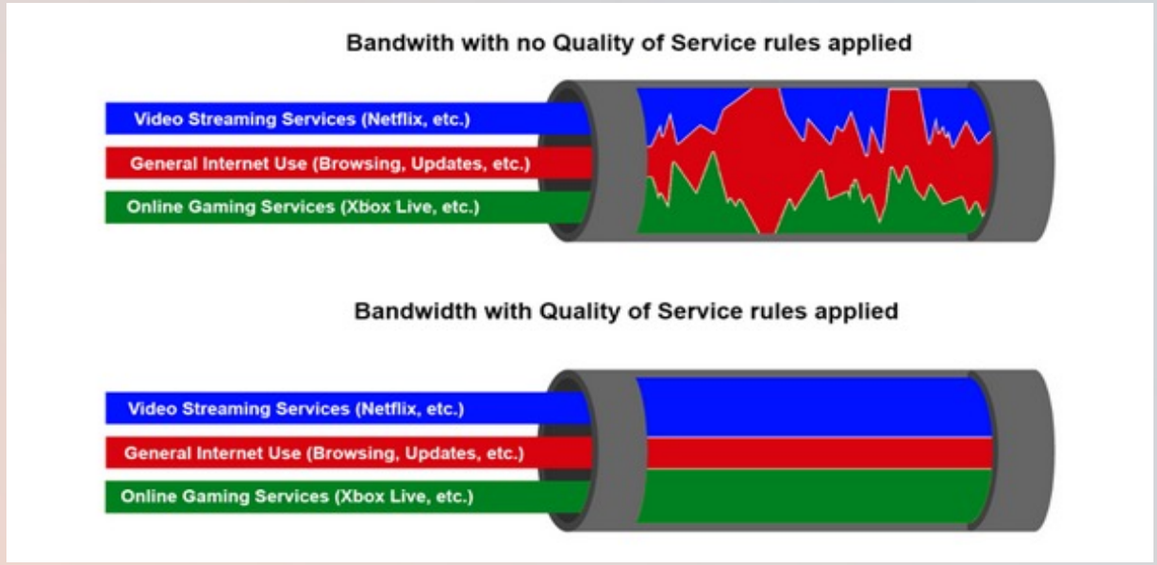
- Internet connectivity is one of the cheapest and most widely available bandwidth options

- However, when it comes to building a corporate wide area network (WAN), Internet connectivity is still not seen as a reliable medium for important business data

https://t.me/learningnets

# SD-WAN - Quality of Service

- Quality of service (QoS) refers to the ability of a network to provide higher levels of service using traffic prioritization and control mechanisms




Source : WanDynamics

https://t.me/learningnets

# SD-WAN - Quality of Service

- Some SD-WAN vendors market their Forward Error Correction (FEC) and Dynamic Path Selection/Control features as QOS but they are not QOS mechanisms

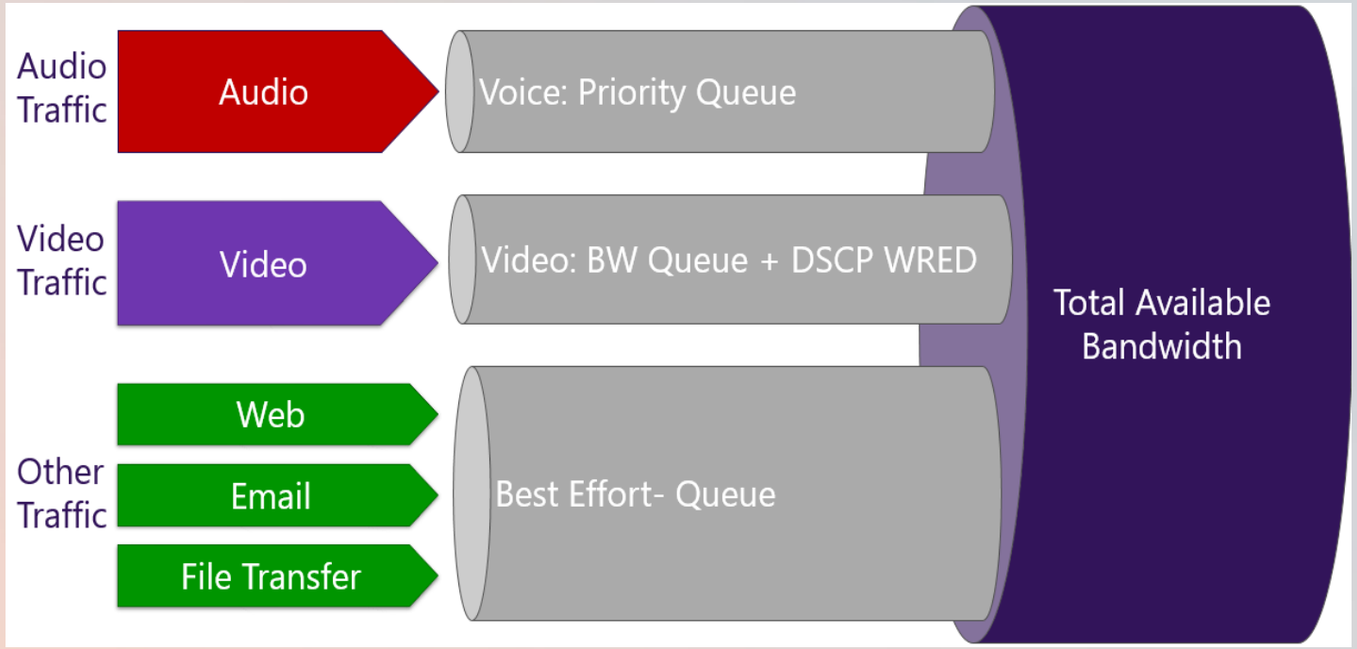
- Although these features improve the network performance, they shouldn’t marketed as QOS features!

- Some SD-WAN vendors support Traffic Shaping , Rate Limiting , Policying as QoS features as well

https://t.me/learningnets

# SD-WAN - Quality of Service

- QoS simply prioritization some traffic and punishing others!




https://t.me/learningnets

# SD-WAN - Zero-touch Deployment/Provisioning

- With this capability, IT teams can bring up services without the need to interact with physical equipment, resulting in fast and efficient deployment of services

- ZTP can be found in switches, wireless access points, SD-WAN nodes, NFV-platforms , firewalls and many other networking devices

- Not all ZTP implementations are truly ‘Zero Touch’ though, so sometimes you will also come across terms like ‘minimal touch provisioning’ or ‘one touch provisioning’

https://t.me/learningnets

# SD-WAN - Zero-touch Deployment/Provisioning

**<u>Below steps usually included in a ZTP process</u>**

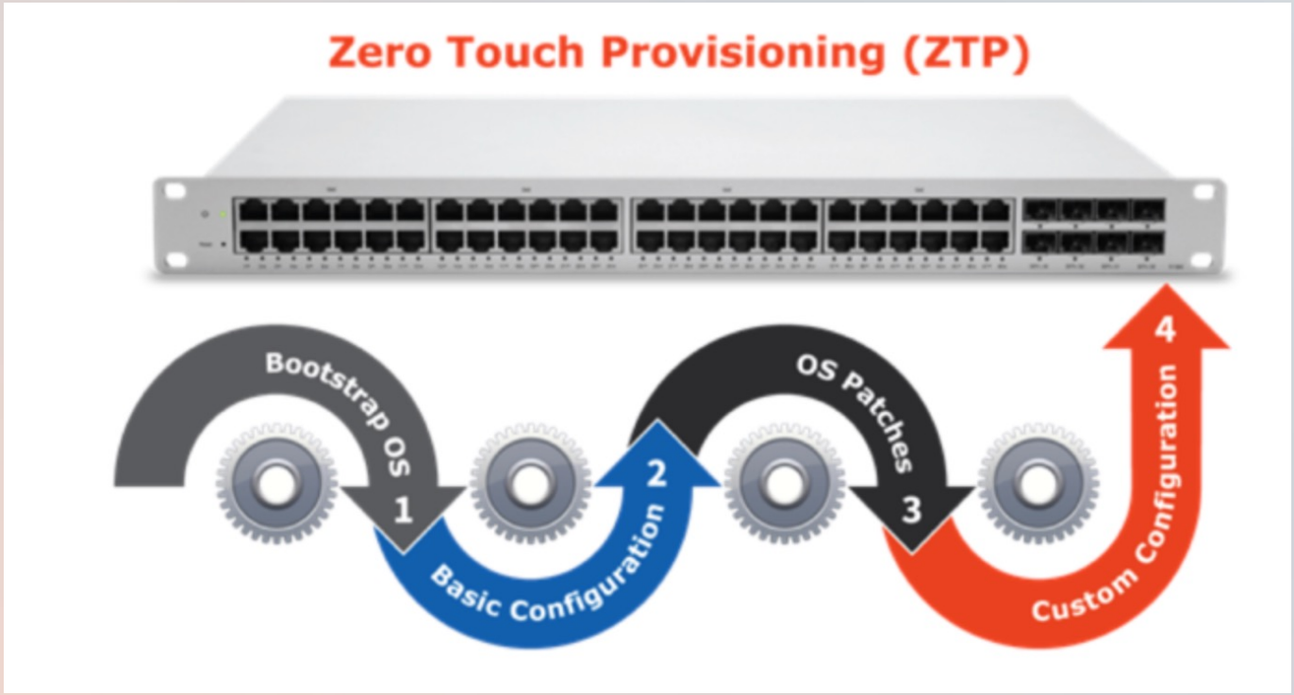
1. Making the device reachable
2. Firmware upgrade
3. Base configuration (DNS, NTP, RADIUS...)
4. Specific configuration (VLANs, interfaces, routing protocols...)

https://t.me/learningnets

# SD-WAN How ZTP Works?

- As the name Zero Touch Provisioning would suggest, the goal is to install a networking appliance somewhere without someone needing to configure it locally

- A new or replacement device can be sent to a site, physically installed and powered up by a locally present employee without IT skills




https://t.me/learningnets

# SD-WAN How ZTP Works?

- Most ZTP implementations are DHCP driven
- Some vendors use ZTP-USB
- Some vendors gives ability to execute scripts (Python or Bash)
- Some vendors retrieves a configuration file via HTTP or TFTP
- Internet connectivity is required in the first place

https://t.me/learningnets

# SD-WAN How ZTP Works?

- More and more vendors are offering a cloud service to support the configuration and ZTP process

- Cisco Meraki, Riverbed, Citrix and Juniper Networks are among those

https://t.me/learningnets

# SD-WAN How ZTP Works?

- All that is required is registering the serial numbers of the devices purchased and the vendor will ensure the devices are correctly registered and visible under your management portal account

- The device can then be fully configured and managed via the cloud

https://t.me/learningnets

# SD-WAN - Global Coverage

- If your business requires international connectivity, you may need to analyze the provider's point-of-presence (POP) coverage to understand the effect on application performance

- Certain providers and vendors operate a significant global network presence that includes specific POPs for both private and internet traffic

https://t.me/learningnets

# SD-WAN - Global Coverage

- SD-WAN features are focused on application performance, but latency and jitter challenges can arise when deploying international services

https://t.me/learningnets

# SD-WAN Vendor POC Support

- The proof of concept for SD-WAN is an excellent way to understand and verify the capability of an SD-WAN offering

- Some vendors offer demo hardware for a period of time, often with presales resources to assist with the configuration

https://t.me/learningnets

# SD-WAN Cloud Connection

- Some SD-WAN products have the ability to program “cloud breakout” based on applications, allowing direct access to trusted sites (like SalesForce.com), while tunneling traffic to unknown sites to either cloud-based or centrally-based inspection services

- This ensures improved productivity, minimizes unnecessary inspection of trusted traffic and provides better security than traditional hub-spoke MPLS solutions

https://t.me/learningnets

# SD-WAN Cloud Connection - SAAS

- Enterprises today face major user experience problems for SaaS applications because of networking problems

- The centralized Internet exit architecture can be inefficient and results in poor SaaS performance

- And branch sites are running out of capacity to handle Internet traffic which is a concern because more than 50% of branch traffic is destined to the cloud

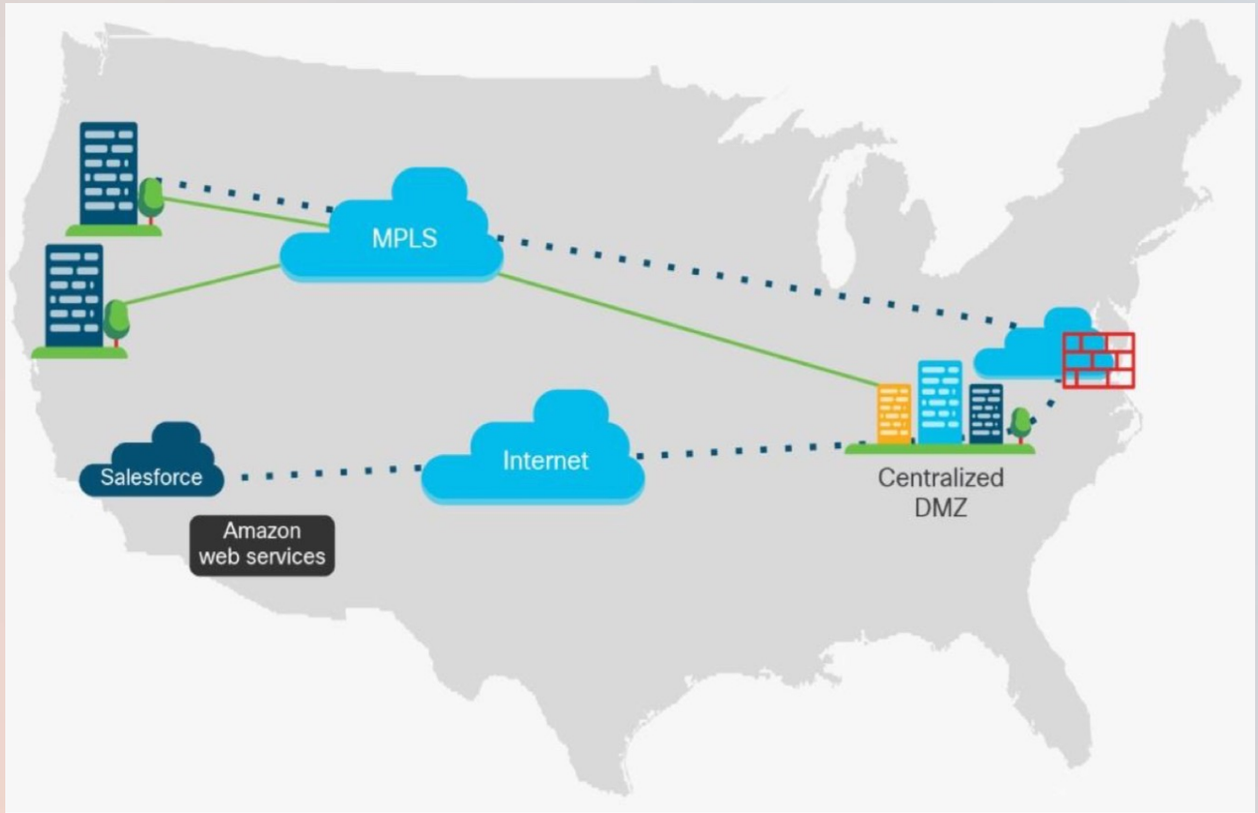
https://t.me/learningnets

# SD-WAN Cloud Connection- SAAS

- Common network designs consolidates application and service controls at centralized DMZs and the data centers

- As a result, enterprise traffic destined for the Internet or public clouds must be backhauled through a centralized DMZ facility

- This causes the traffic to trombone or hairpin, creating an inefficient route that increases the distance between the user and the application



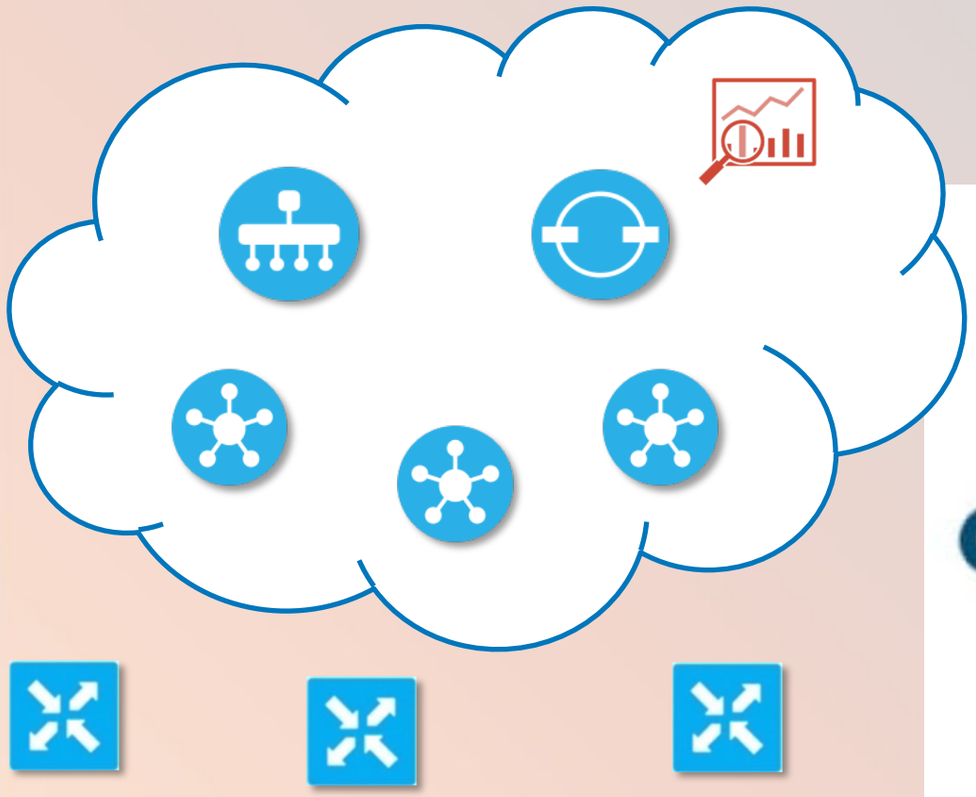

https://t.me/learningnets

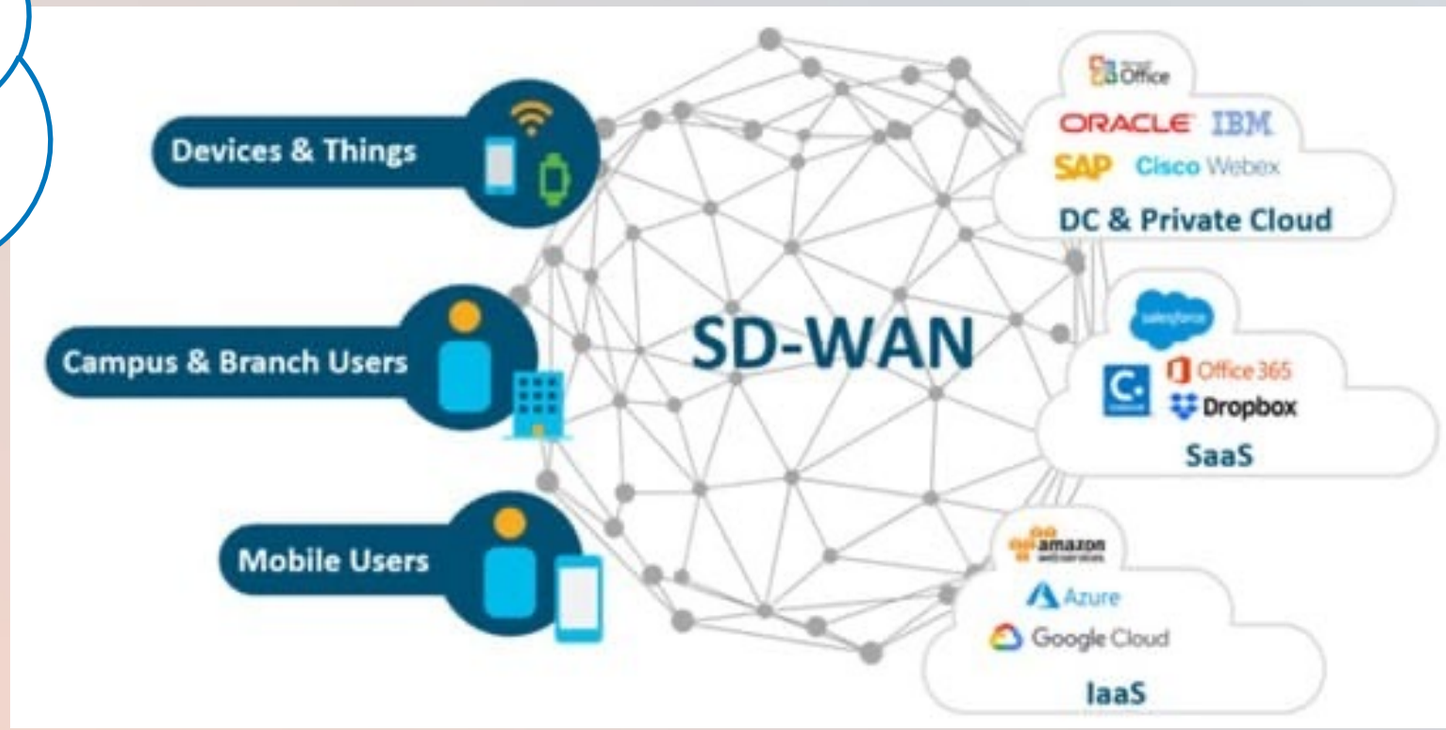
# Traditional WAN vs SD-WAN Comparison

| Traditional WAN | V.S. | SD-WAN |
|---|---|---|
| **Months** to provision<br><br>**On-site technician** required to set up new sites<br><br>**Manual** configuration of routers and gateways | **Deployment** | **Hours** to provision<br><br>**Zero-touch configuration**<br><br>**Automatically** deploy traffic rules |
| **Separate** management systems for diverse connections and carriers<br><br>**High IT expertise** required for networks changes and maintenance | **Management** | **Unified** user interface to set up / manage diverse connections<br><br>**Dynamically self-propagate** configuration changes through network |
| **Limited visibility** into network conditions | **Visibility** | **Full visibility** into real-time network conditions |
| **No flexibility** for temporary connections<br><br>**Cannot quickly respond** to changing bandwidth needs | **Flexibility** | **Instantly set up or delete connections** based on real-time demand<br><br>**Adjust bandwidth quickly** as needed |
| **Not optimized** for connectivity to public clouds | **Cloud connectivity** | **Connect in minutes** to multiple clouds and SaaS |
| **Unpredictable public Internet quality** can lead to unacceptable latency and packet loss<br><br>**Private / MPLS circuits are expensive** and typically require long term contracts | **Cost** | **Achieve balance between performance and cost** using quality of service (QoS) protocols to differentiate application priorities and<br><br>**Pay as you grow** |

https://t.me/learningnets

# Cisco SD-WAN






https://t.me/learningnets

# Why SD-WAN?

- Traditional WAN is not capable to handle today's application and WAN requirements

- Adoption of Services like SaaS (Software as a Service) & IaaS (Infrastructure as a Service)

https://t.me/learningnets

# Why SD-WAN?

- Exposing an enterprise to the internet can introduce threat and compliance issues
- Limited Application Visibility and understanding to networks
- Expensive WAN Circuits with limited features

https://t.me/learningnets

# Value Proposition by SD-WAN Solution

- Increasing bandwidth through the activation of idle backup links and dynamic load-balancing

- Delivering faster cloud access by enabling direct internet access at the branch

https://t.me/learningnets

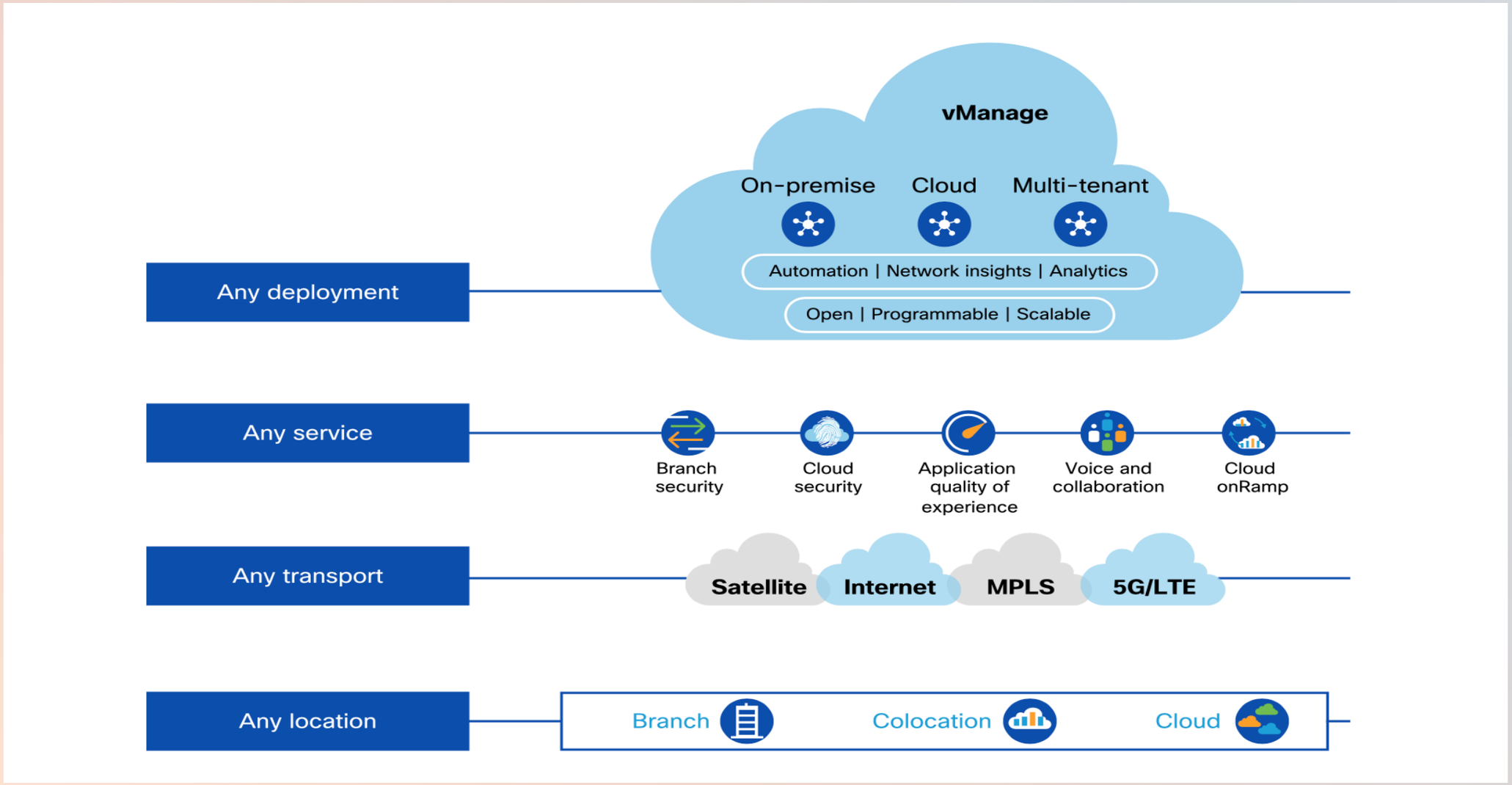
# Value Proposition by SD-WAN Solution

- Reducing operational and management costs through centralized management that was commonly cloud-based

- Lowering WAN costs through the use of cheaper internet or LTE connectivity as an alternative to MPLS

https://t.me/learningnets

# Cisco SD-WAN Cloud Scale Architecture



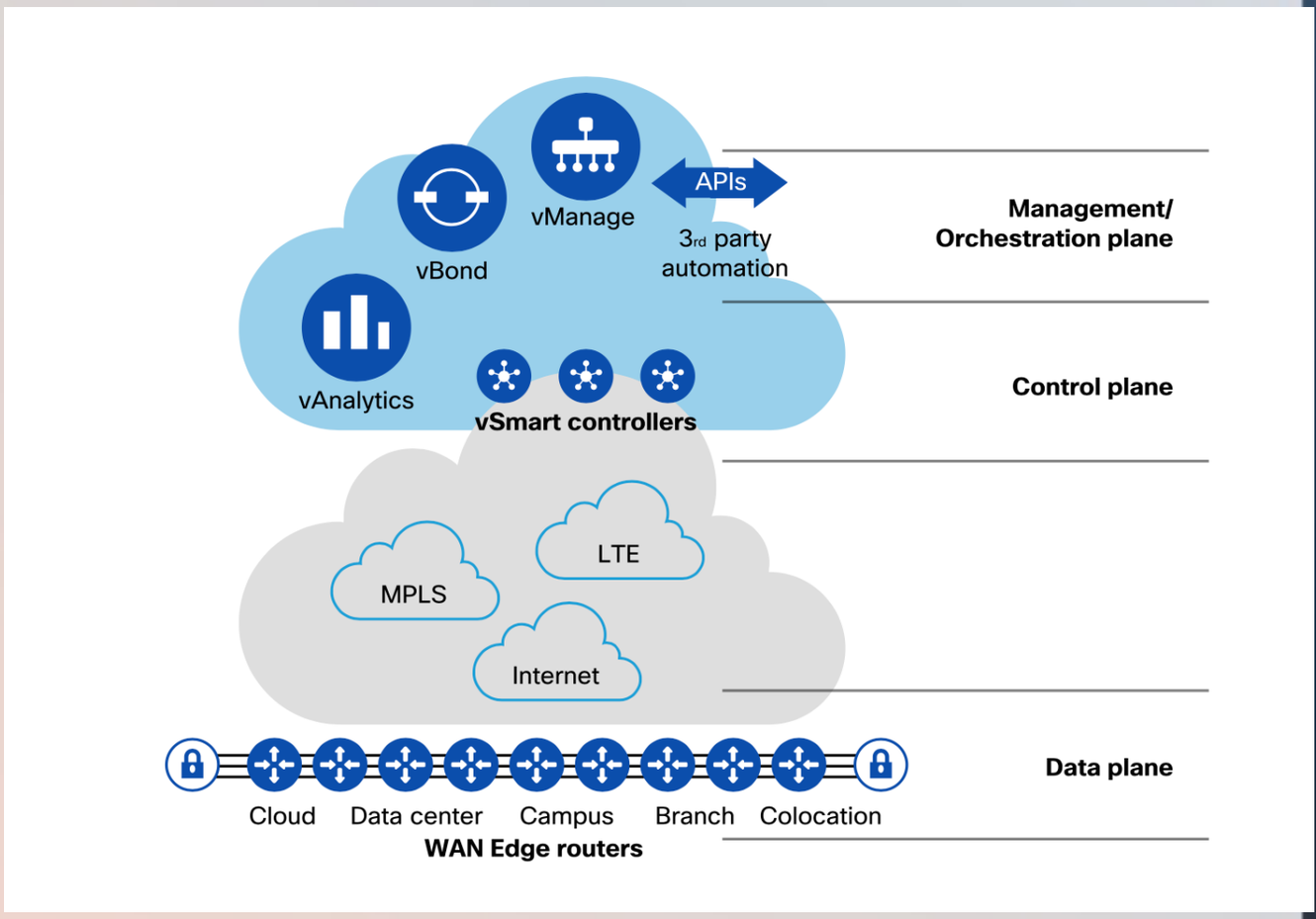

https://t.me/learningnets

# Component and Architecture of Cisco SD-WAN

- The Cisco SD-WAN solution is a cloud-delivered Wide Area Network (WAN) overlay architecture that extends the principles of software-defined networking (SDN) into the WAN

- Cisco SD-WAN is broken into four parts:

  - Data Plane
  - Control Plane
  - Management Plane
  - Orchestration Plane




https://t.me/learningnets

# Component and Architecture of Cisco SD-WAN

- Cisco SD-WAN consist of four main components:
  - vManage
  - vBond
  - vSmart
  - vEdge or WAN Edge

https://t.me/learningnets

# vManage

- In the Management Plane, vManage represent the user interface of the solution

- Network administrators and operators perform configuration, provisioning, troubleshooting, and monitoring activity here

- vManage provides single pane of glass for management of all the SD-WAN components

https://t.me/learningnets

# vBond

- Cisco vBond resides in Orchestration Plane

- The vBond controller is largely responsible for the Zero-Touch Provisioning process as well as first-line authentication, control/management information distribution, and facilitating

https://t.me/learningnets

# vBond

- Network Address Translation (NAT) traversal

- vBond is responsible for onboarding the device into the SD-WAN fabric

https://t.me/learningnets

# Cisco vSmart

- Cisco vSmart is the "brain" of the solution and exists within the control plane

- As policies are created on vManage, vSmart is the component responsible for enforcing these policies centrally

- Routing Decisions are made by vSmart only and it act like a BGP ’RR’ for propagating routes using OMP (Overlay Management Protocol)

https://t.me/learningnets

# Cisco WAN Edge or vEdge

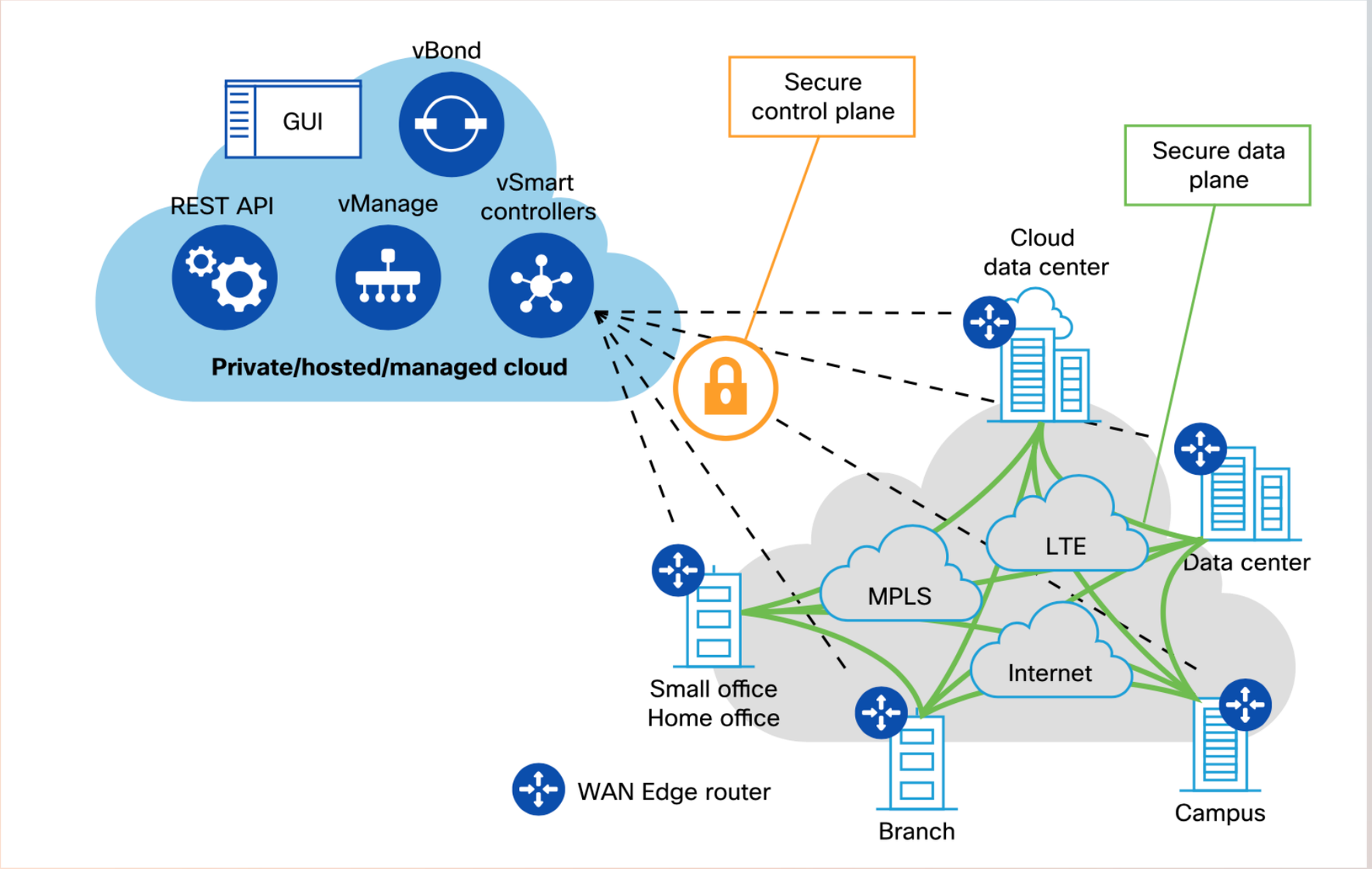
- WAN Edge comes under the Data Plane and used for forwarding

- Cisco WAN Edge routers come in multiple forms, virtual and physical

- The WAN Edge routers form Internet Protocol Security (IPSec) tunnels with each other to form the SD-WAN overlay

https://t.me/learningnets

# Cisco SD-WAN Fabric Components



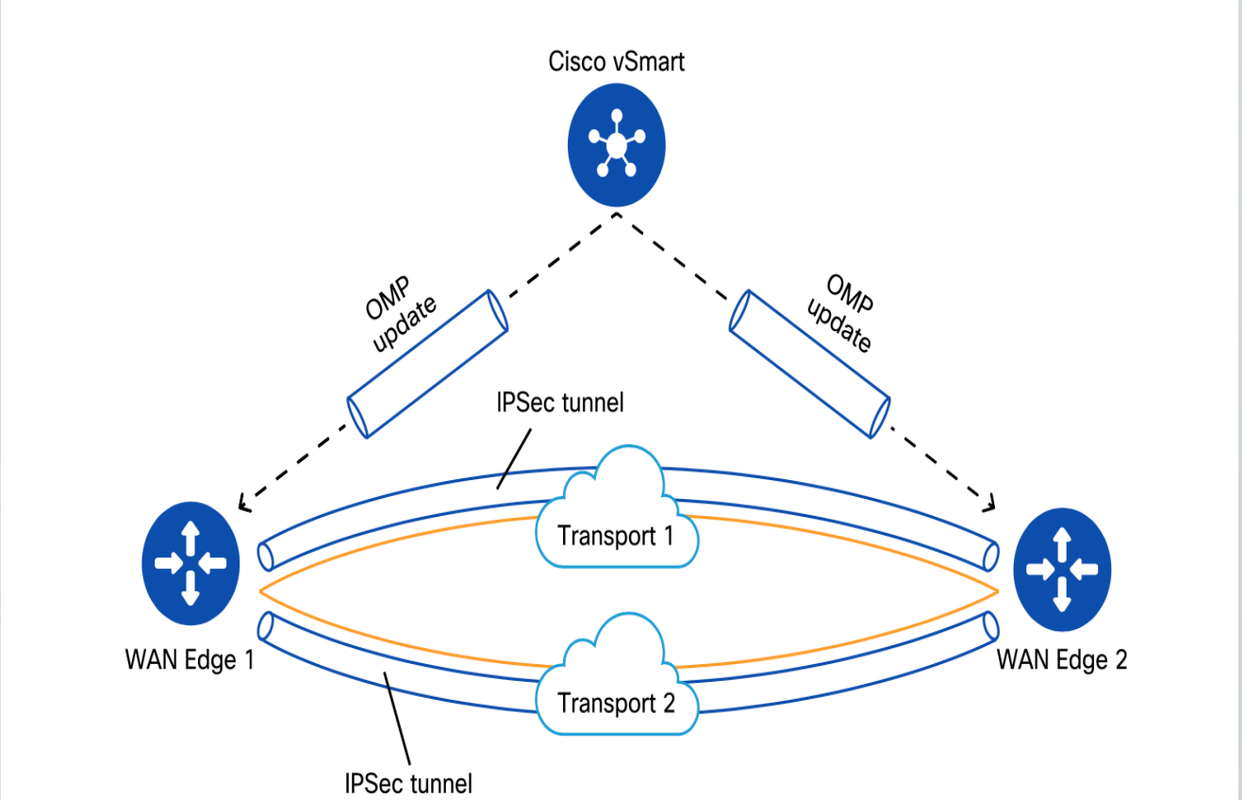

https://t.me/learningnets

# Cisco SD-WAN Viptela Control Plane

https://t.me/learningnets

# OMP (Overlay Management Protocol)

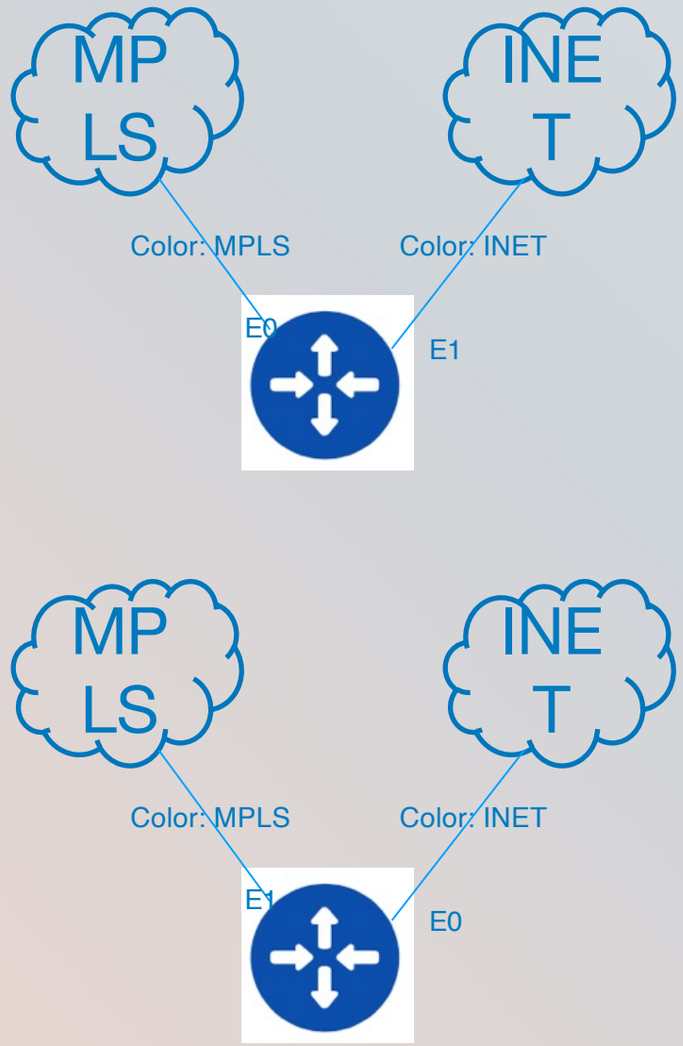
- OMP Works as BGP in Traditional Routing
- Exchanges Routing Information via vSmart
- vSmart acts like BGP RR
- All WAN Edge builds up the OMP Session with vSmart only




https://t.me/learningnets

# TLOC (Transport Locator)

- TLOC is Transport Locator use to identify transport path
- TLOC replaces the next-hop information
- One WAN Edge can be configured up to 8 colors
- TLOC Entry consist of following items:
  - System IP
  - Color
  - Encapsulation Protocol
    - GRE
    - IPSEC




https://t.me/learningnets

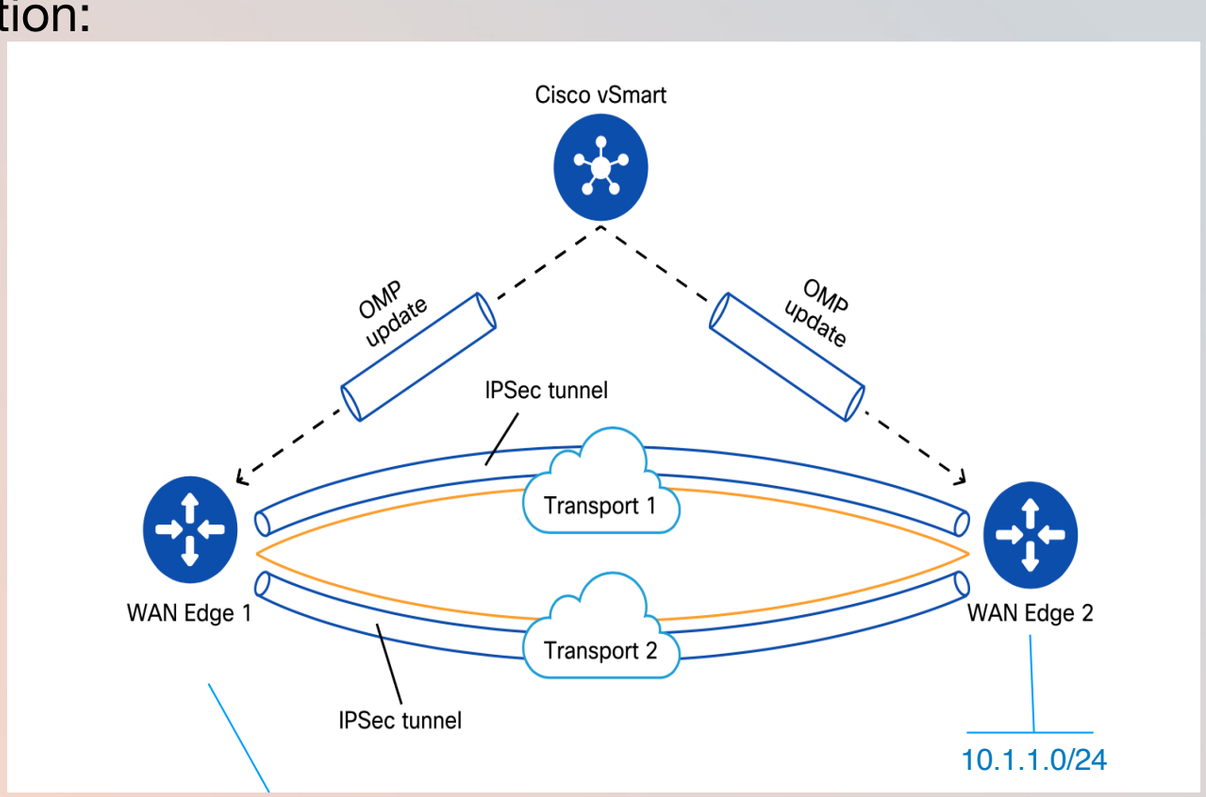
# Types of Route

- OMP Route
- TLOC Route
- Service Route

https://t.me/learningnets

# vRoute or OMP Route

- OMP Route is used to provide the destination prefix reachability along with the TLOC Information
- It is shared via vSmart in OMP update
- OMP Route Contains the below information:
  - Destination Prefix
  - TLOC
  - Prefix Attributes




https://t.me/learningnets

# TLOC Route

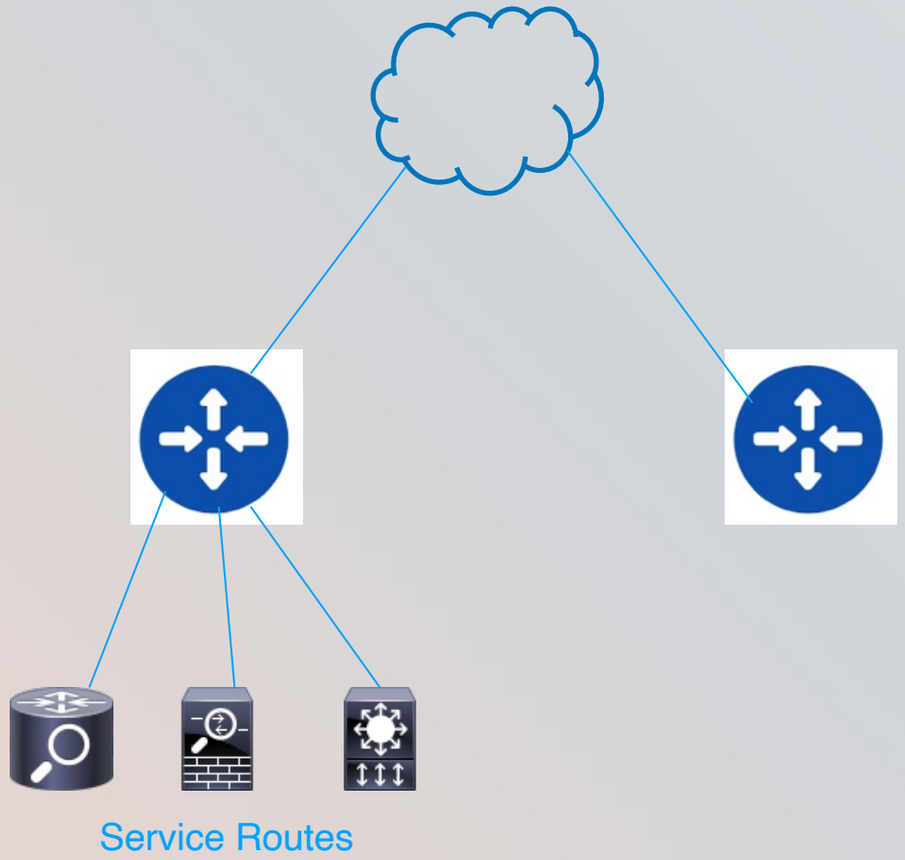
- It Provides the further TLOC reachability information to the WAN Edge
- WAN Edge does recursive lookup here, first it looks for OMP Route to get the TLOC information based on the destination prefix, and then look for TLOC route to get the destination WAN edge reachability based on the TLOC information received from OMP Route
- Once the WAN Edge have both OMP Route and TLOC Route, then only it will establish the IPSEC VPN with respective WAN Edge
- TLOC Route contains the following information:
  - TLOC
  - WAN IP
  - Attributes

https://t.me/learningnets

# Service Route

- It contains the information about the reachability of any Service Device (Firewall, Load Balancer or IPS/IDS)

- The WAN Edge which is connected to any of these device will be generating the Service Route



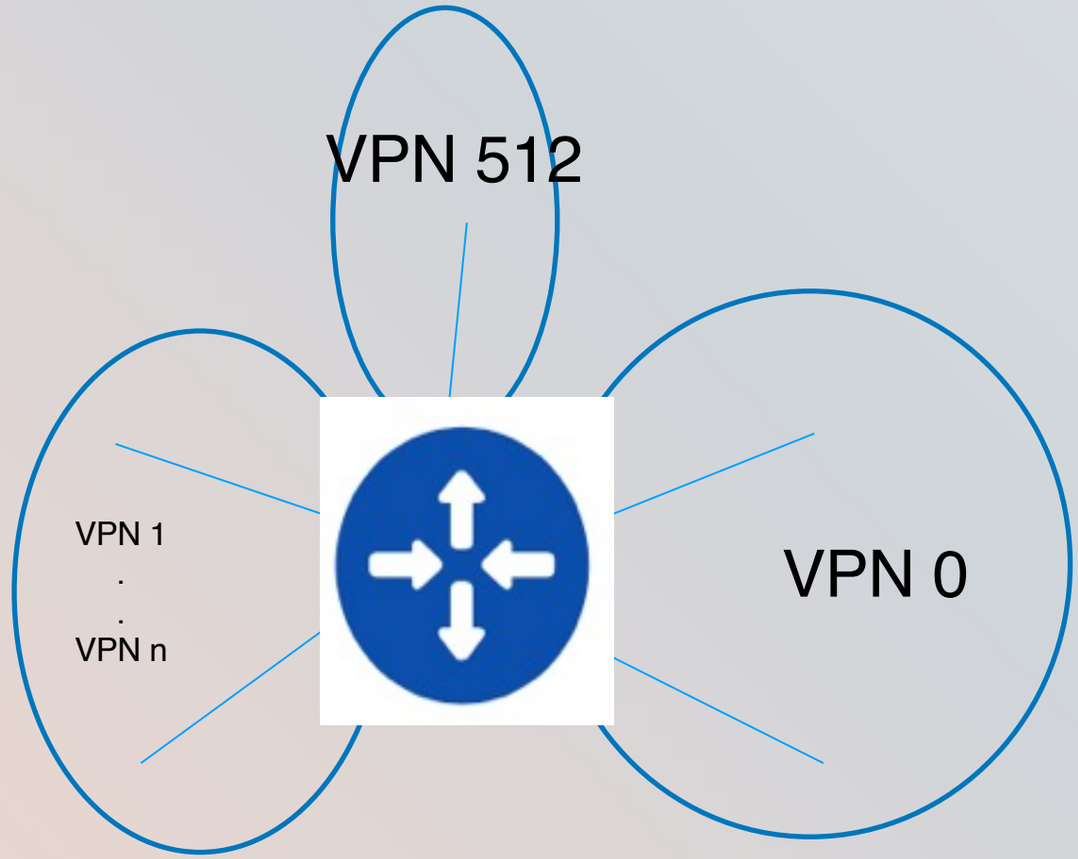

https://t.me/learningnets

# Cisco SD-WAN Viptela Data Plane

https://t.me/learningnets

# VPN

- VPN Plays a major role in the SD-WAN data Plane operations
- VPN is alternative of VRF on Viptela Hardware
- All the VPN can be configured only with number not name unlike VRF
- Each VPN have its own routing table.
- There are two types of VPN configure by default:
  - VPN0
  - VPN512




https://t.me/learningnets

# VPN 512

- By Default Configured VPN
- Used for Out of Band Management only
- Equivalent to default VRF Management
- Only Used for Management Traffic

https://t.me/learningnets

# VPN0

- VPN0 is known as Transport VPN
- Used for Initiate and terminate IPSEC VPN
- All the Transport Interfaces should be configured under VPN0
- Used as front door VRF for terminating VPN Traffic
- System IP is part of VPN0 by default
- Used for making communication to other controllers
- Secondly can be used for management and control traffic as well

https://t.me/learningnets

# VPN1 to VPNn

- All the VPN from VPN1 to the limit of VPN's are known as Service VPN

- Used as regular VRF's on the WAN Edge devices for communicating to LAN Side services

https://t.me/learningnets

# VPN1 to VPNn

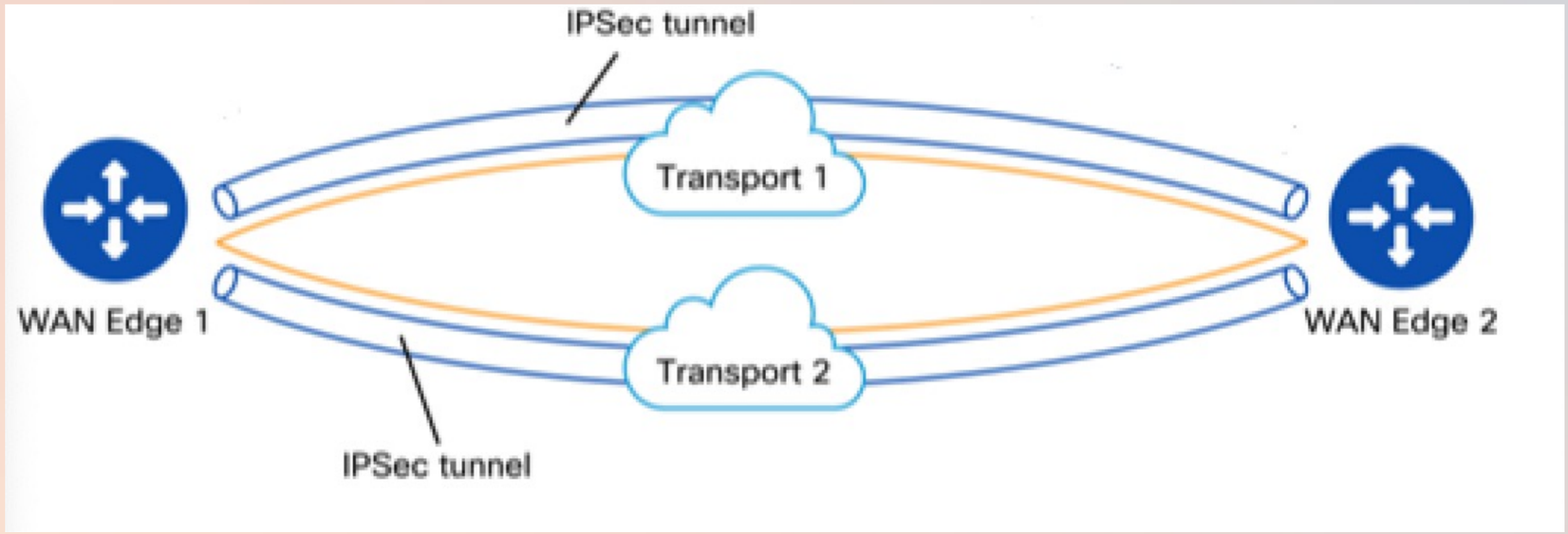
- All the traffic originated from Service VPN, will be having VPN tag on it and sent through the same IPSEC tunnel between the WAN Edge devices

- VPN Tag will be used on the destination WAN edge for landing the traffic into appropriate VPN

https://t.me/learningnets

# IPSEC VPN

- IPSEC tunnel will be configured between all the WAN Edges in full mesh manner by default
- One IPSEC tunnel will be used to carry multiple VPN traffic
- By Default using AES Encryption type




https://t.me/learningnets

# IPSEC Key Exchange

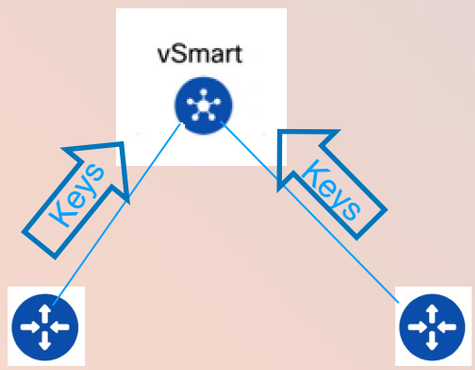
- For scalability purpose, SD-WAN is not using IKE (Either IKEv1 or IKEv2) for key exchange between the WAN Edges

- WAN Edges will be using there DTLS/SSL secure channel with vSmart for key exchange process

https://t.me/learningnets

# IPSEC Key Exchange

- All the WAN Edges will share there keys along with OMP update to vSmart and vSmart will be responsible for exchanging the keys




https://t.me/learningnets

# Bidirectional Forwarding Detection (BFD)

- BFD Probes will sent by all the the WAN Edges to other WAN Edges through all the transports

- BFD probes send over the IPSEC tunnel

- BFD probes makes sure the TLOC reachability, if BFD Probes fails, the that TLOC will be considered as invalid

https://t.me/learningnets

# Bidirectional Forwarding Detection (BFD)

- BFD probes also be used for checking the circuit quality, by checking the following parameters:
  - Drop Counts
  - Jitter
  - Latency
  - Bandwidth

https://t.me/learningnets

# Cisco SD-WAN Viptela Management Plane

https://t.me/learningnets

# Device Templates

- vManage is used for management plane and pushing configuration to the WAN Edges

- Device templates is the combination of multiple feature templates

https://t.me/learningnets

# Device Templates

- Feature templates are used to enable specific global configuration to the WAN edge device

- The other half of the configuration is policy other than templates

https://t.me/learningnets

# Type of Policy in SD-WAN

- Centralized Control Policy
- Centralized Data Policy
- Localized Control Policy
- Localized Data Policy

https://t.me/learningnets

# Centralized Control Policy

- It’s a Central policy and defined on vSmart

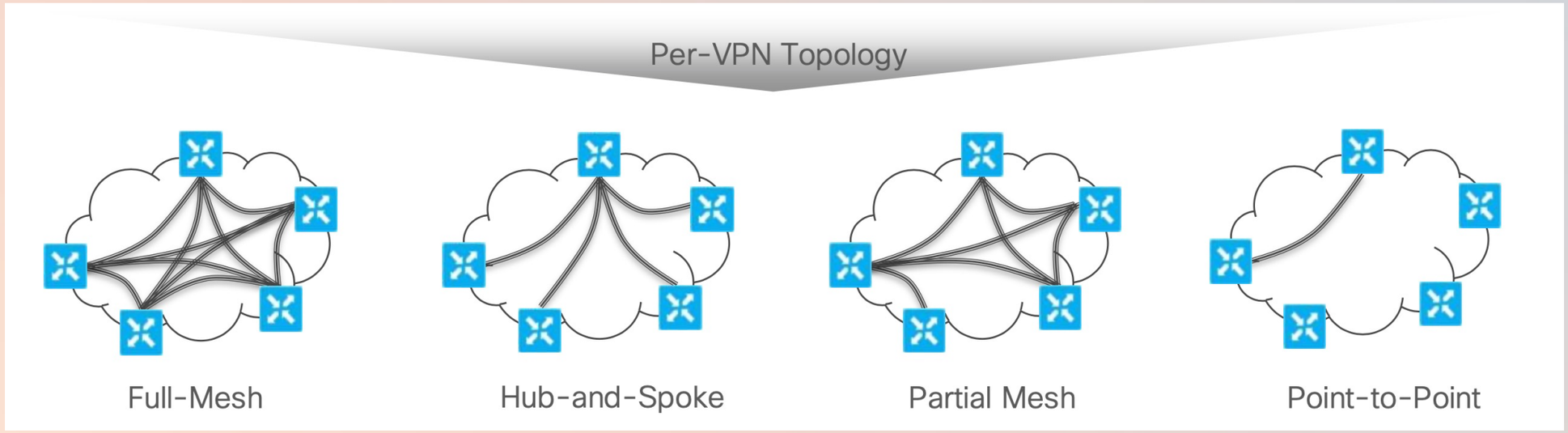
- It defines how routing (Route learning and Advertisement ) will take place in SD-WAN through vSmart

https://t.me/learningnets

# Centralized Control Policy

- Similar like Route-map in traditional routing

- It also defines the topology between the WAN edges (by default it's full mesh)




https://t.me/learningnets

# Centralized Data Policy

- This policy is used to match the data packet based on IP and port
- This can be used as ACL in traditional routing
- This policy is managed by vSmart and downloaded to WAN Edge

https://t.me/learningnets

# Centralized Data Policy

- This use to match the QoS Parameters

- This policy also used to change the next-hop of the packet

- This can be applied to Transport VPN as well as Service VPN incoming packets

https://t.me/learningnets

# Localized Control Policy

- This policy locally stored on WAN Edge router

- This is used for manipulating  or filtering the routing information

https://t.me/learningnets

# Localized Control Policy

- This mainly used for routing information configured on service side VPN

- It works as filter list or route-map for the routing protocol (OSPF or BGP) configured on service VPN (Interface)

https://t.me/learningnets

# Localized Data Policy

- This policy is similar like Centralized Data Policy but stored locally on the WAN Edge

- This policy also used for matching data packet but mainly used for QOS deployment

- This policy can be applied in individual interface unlike the Centralized Data Policy which is applied on per VPN basis

https://t.me/learningnets

# MPLS

## Theory

- If the requirement is to have scalable VPN solution which can provide fast reroute traffic protection, then only choice is MPLS

- MPLS is a protocol independent transport mechanism. It can carry layer 2 and layer 3 payloads. Packet forwarding decision is made solely on the label without the need to examine the packet itself.

- MPLS interacts as an overlay with IGP and BGP in many ways. For example in Multi-level IS-IS design, level 1 domain breaks end to end LSP.

- In OSPF and EIGRP summarization creates the same problem and we mentioned about this problem in respective chapters.

https://t.me/learningnets

- MPLS is a transport mechanism.
- It supports label stacking which allows many applications to run over MPLS. MPLS application is the topic of this session
- All application needs additional label
- Most of the MPLS operation requires minimum 2 labels
- In MPLS layer 3 VPN, Transport and BGP label
- In Layer 2 VPN, Transport and VC Label
- Labels can be assigned by the Standard porotocol for Transport can be assigned by 4 protocols currently. LDP, RSVP, BGP and Segment Routing. For the Service MPLS Label , SR and RSVP are not used

https://t.me/learningnets

# MPLS Applications

Important MPLS Applications/Services are below. In this chapter all of them will be explained.

- Layer 2 MPLS VPN
- Layer 3 MPLS VPN
- Inter AS MPLS VPN
- Carrier Supporting Carrier
- MPLS Traffic Engineering with RSVP
- Seamless MPLS/Unified MPLS

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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).

https://t.me/learningnets

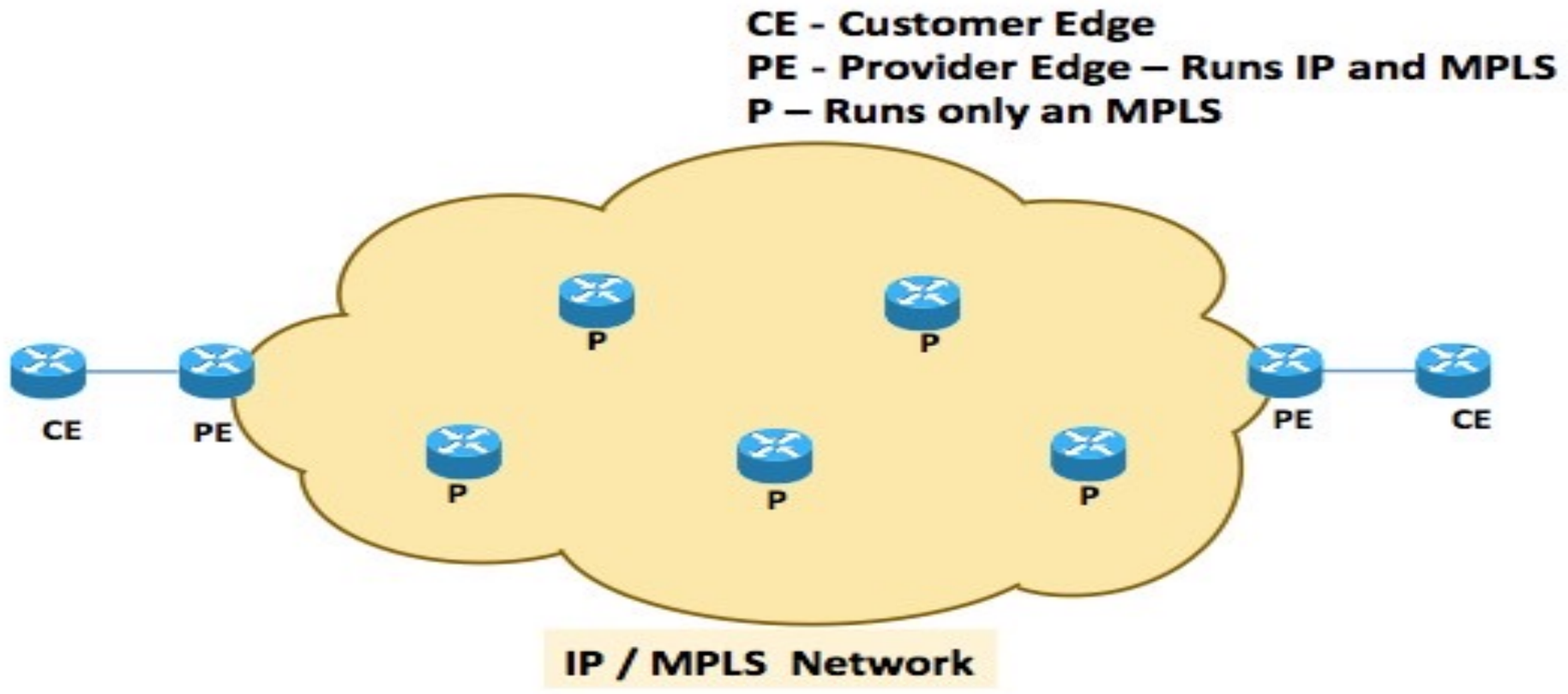
- Both VPWS and VPLS can be accomplished in two ways.

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- PE is Provider Edge device. In MPLS networks, all the intelligence is at the edge. Core is kept as simple as possible. KISS principle in network design comes from the ‘ Intelligent Edge, Dummy Core ‘ idea.

- P is the Provider device and only have a connection to the P devices. P device doesn’t have a connection to the customer network.

- PE device looks at the incoming frame or packet and identify which egress PE device is used for transport. Second lookup is made to determine the egress interface on the egress device.

https://t.me/learningnets

- 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.

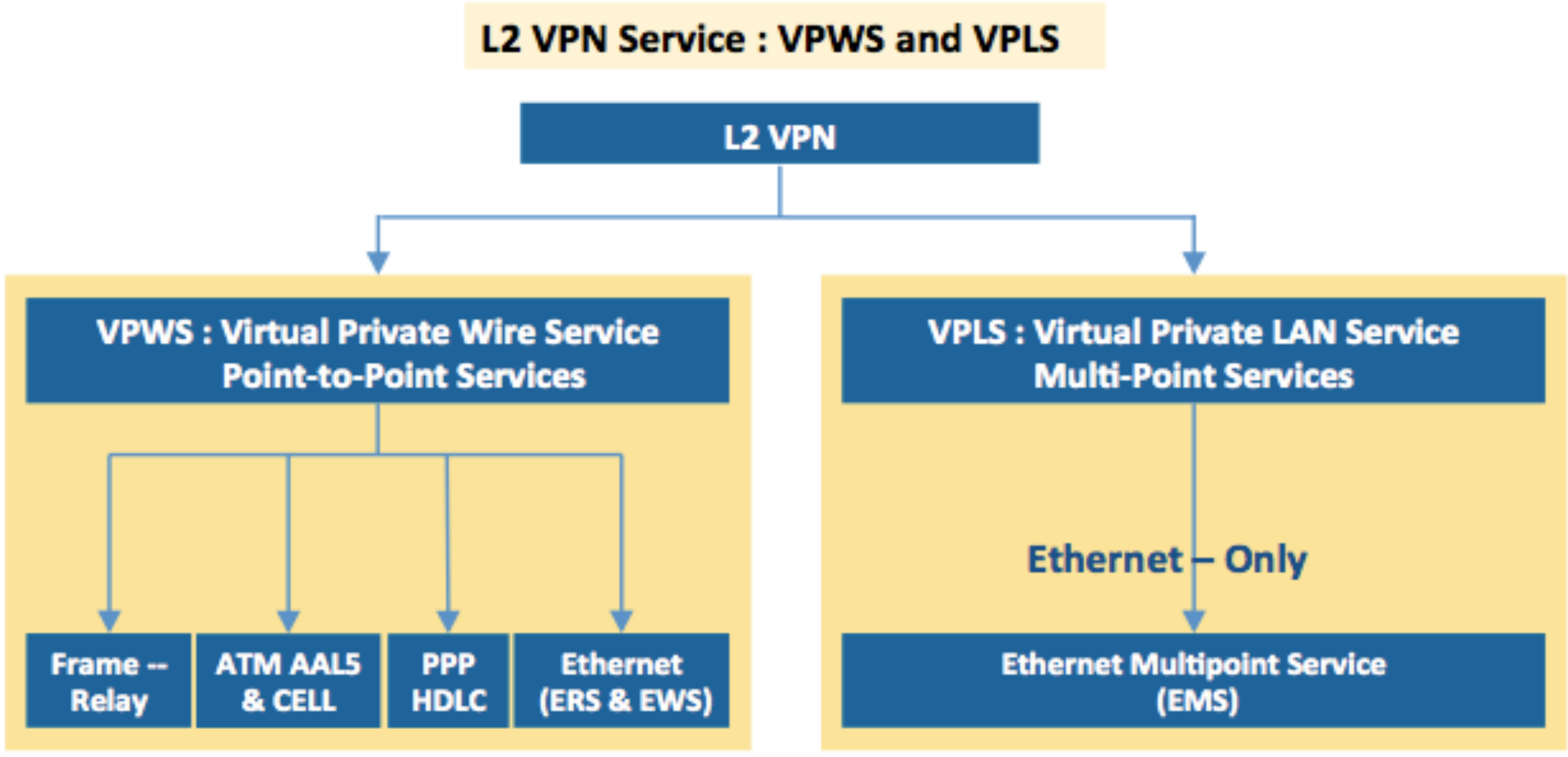
- Inner label is called pseudo wire or VC label in MPLS layer 2 VPN and used to identify the individual pseudo wire on the egress PE.

https://t.me/learningnets

- 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.

https://t.me/learningnets

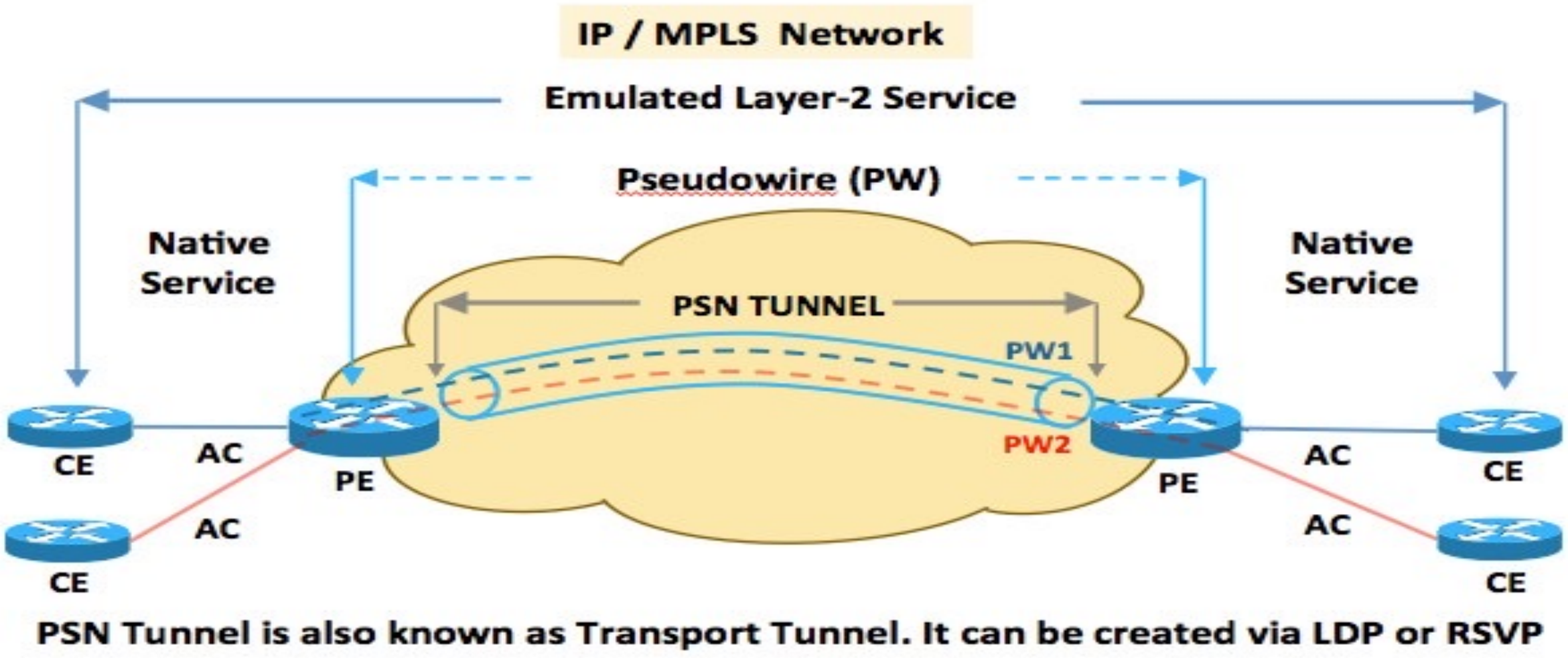



- VPWS can carry almost all layer 2 payloads as can be seen from the above diagram. VPLS can carry Ethernet only. Although there is an attempt in the IETF for the other layer2 payloads over VPLS as well.

https://t.me/learningnets

- 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.

https://t.me/learningnets



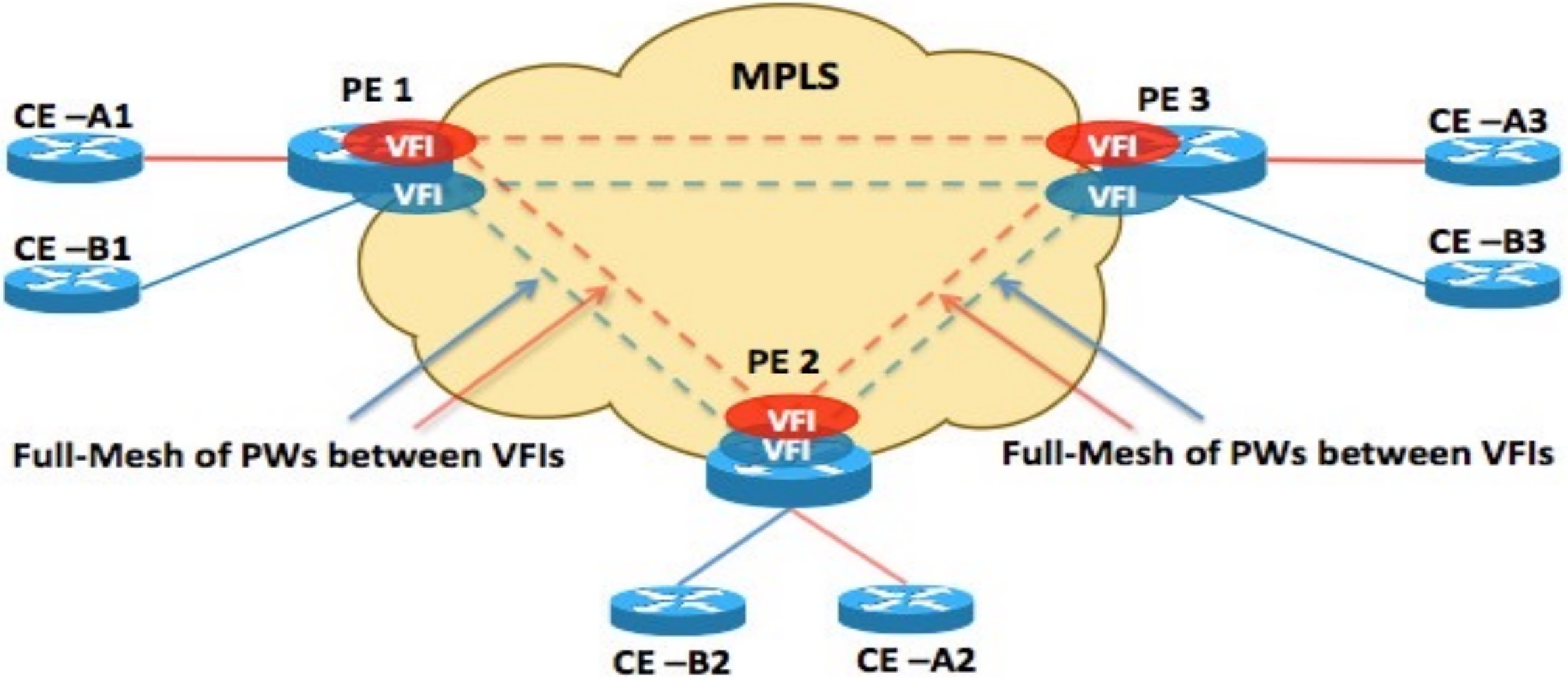

- There is only one egress point for the VPWS service which is the other end of the pseudo wire, thus PE device doesn't have to keep Mac Address to PW binding. PE device doesn't have to learn MAC address of the customer. This provides scalability.

https://t.me/learningnets

- 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

https://t.me/learningnets

## In order to create a VPLS , full mesh of Pseudowires is necessary

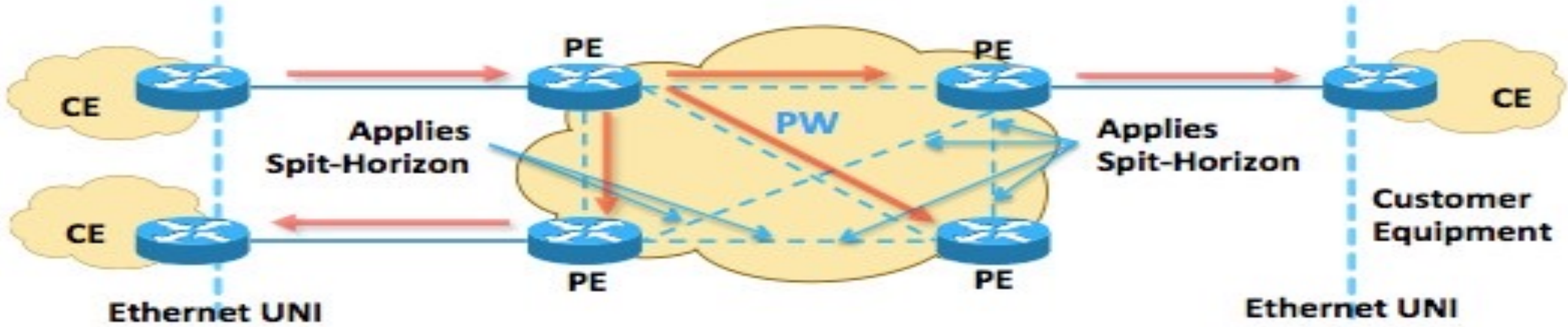


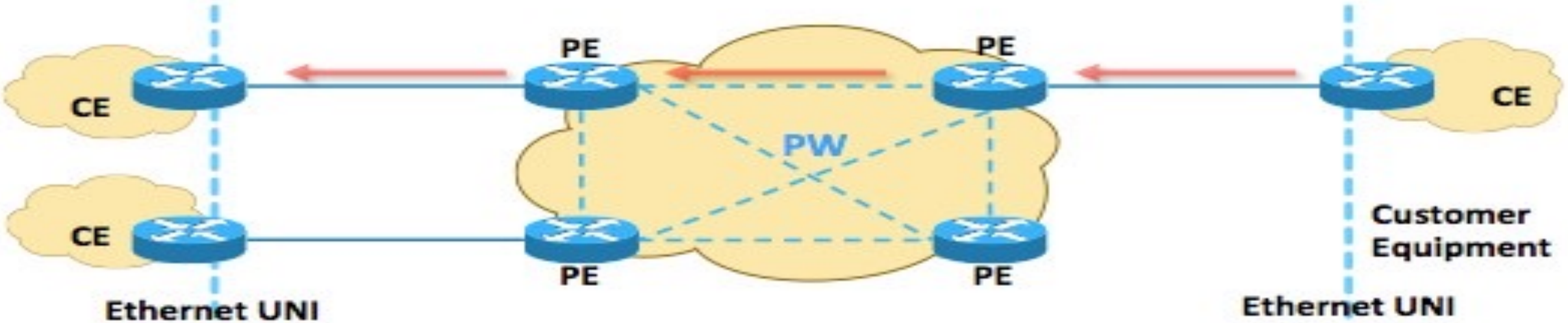

- VFI, also known as VSI is a virtual forwarding/switching instance is an equivalent of VRF in Layer3 VPN. It is used to identify the VPLS domain per customer.

https://t.me/learningnets

- 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

https://t.me/learningnets




- 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.

https://t.me/learningnets

- 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.

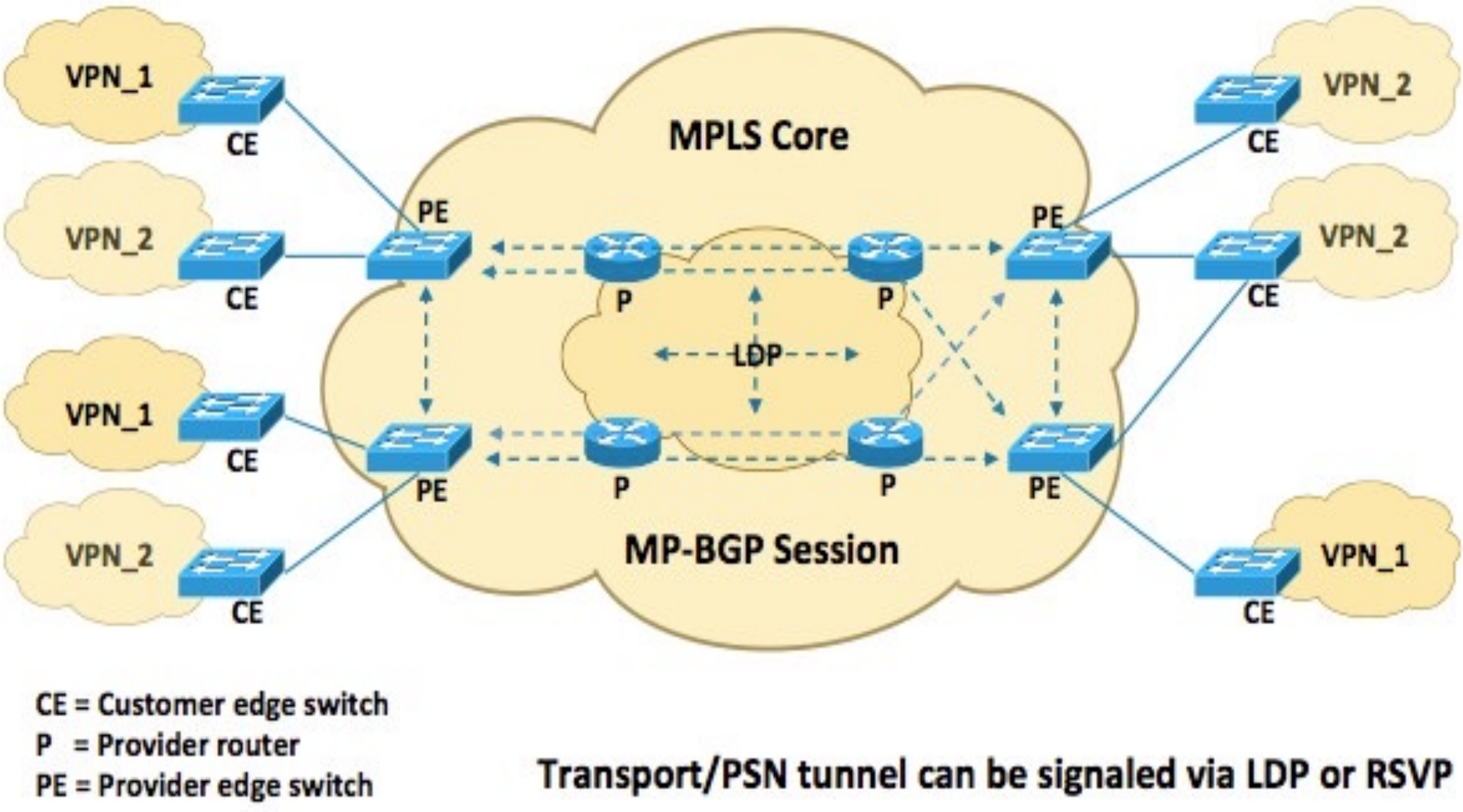
- Number of PE which needs to join the VPLS instance which is also known as VSI (Virtual Switch Instance) might be too high.

https://t.me/learningnets

- 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.

https://t.me/learningnets

# Layer 3 MPLS VPN




https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- Inner label which is also known as BGP label provides VPN label information with the help of MPBGP (Multiprotocol BGP). This label allows data plane separation. Customer traffic is kept separated over common MPLS network with the VPN label.

- MP-BGP session is created between PE devices only. P devices don’t run BGP in MPLS environment. This is known as BGP Free Core design.

https://t.me/learningnets

- 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.

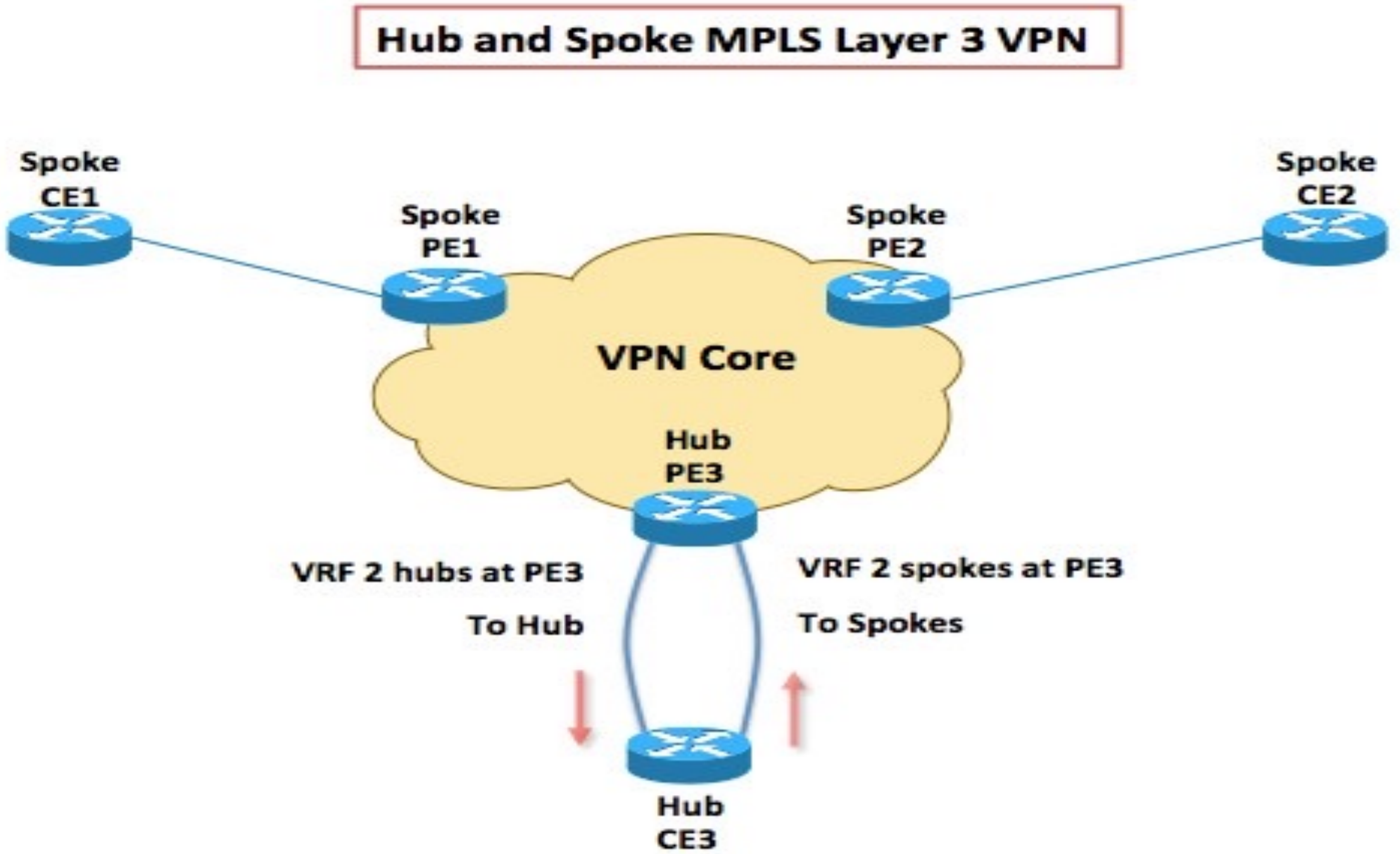
- Route Target is an extended community attribute is used to import and export VPN prefixes to and from VRF. Export route Target is used to advertise prefixes from VRF to MP-BGP, Import Route Target is used to receive the VPN prefixes from MP-BGP into customer VRF.

https://t.me/learningnets

- 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.

https://t.me/learningnets

# Hub and Spoke MPLS VPNs




https://t.me/learningnets

# MPLS VPN Resiliency

- Customers for the increased resiliency may want to have two MPLS connections from the different service providers. Primary and secondary VPNs are same type of VPN in general, so if the primary is Layer2 VPN , since this is the operational model which customer wants to operate, secondary link from the other provider also is chosen as layer2 VPN.

https://t.me/learningnets

- If Layer3 VPN is received from one service provider, second link from the different provider is also received as Layer3 VPN.

- Of course, neither MPLS Layer2 VPN nor Layer 3 VPN doesn’t have to have MPLS VPN as a backup, but the Internet or any other transport can be a backup for the customer.

- Look at the comparison chart for single vs dual MPLS VPN design, for choosing Single vs. Dual Providers.

https://t.me/learningnets

# Inter AS MPLS VPN

- Customer might be connected to two different service providers. This might be redundancy purpose or maybe Service Provider may not have a POP in some of its customer locations.

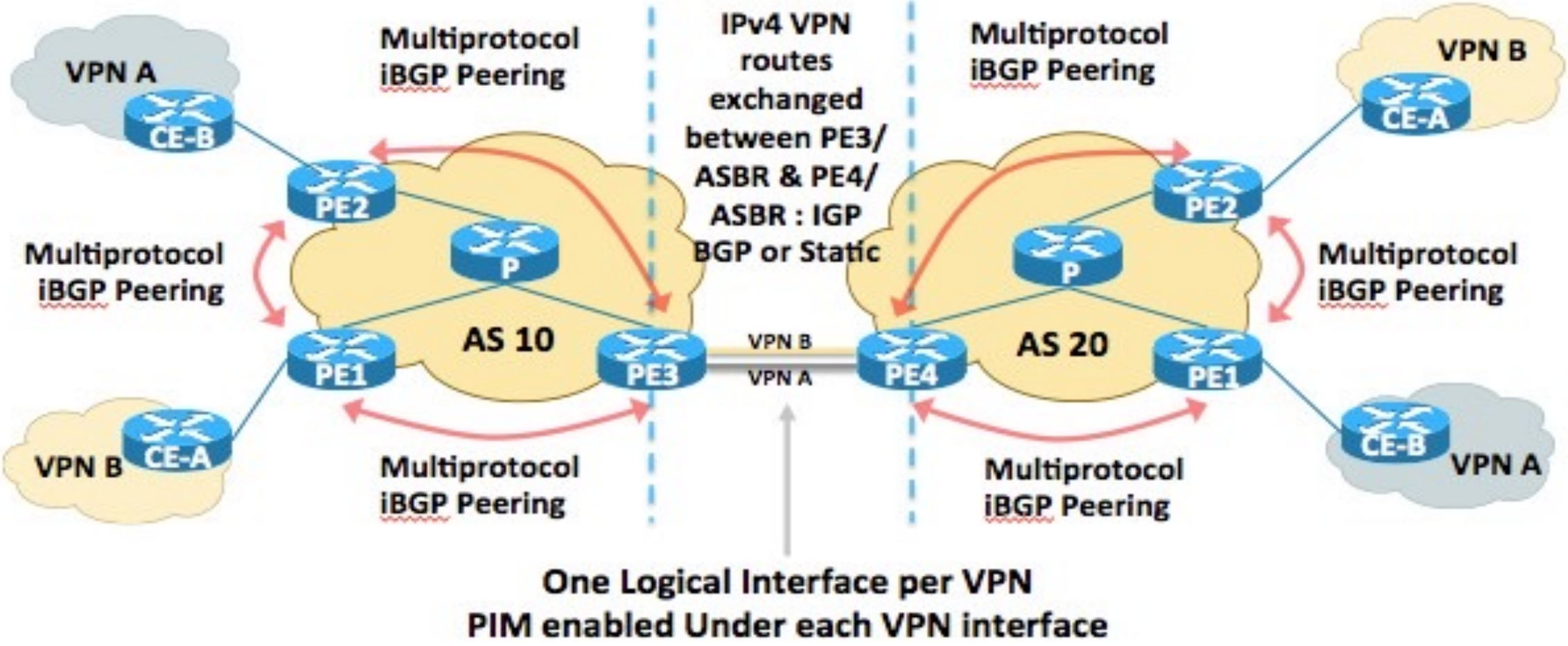
- This requires VPN agreement between the Service Provider to support their customers end-to-end MPLS VPN deployment

- There are 3 model defined in RFC 4364 for Inter AS MPLS VPNs. Inter AS Option A. It is also known as 10A

https://t.me/learningnets

## Inter AS MPLS VPN Option A

**Instead of Full Mesh MP-IBGP Peering inside an AS, VPN Route Reflector can be used.**
**Inter-AS Option A requires all customer VRFs and VPN routes on the ASBRs.**




https://t.me/learningnets

- 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.

https://t.me/learningnets

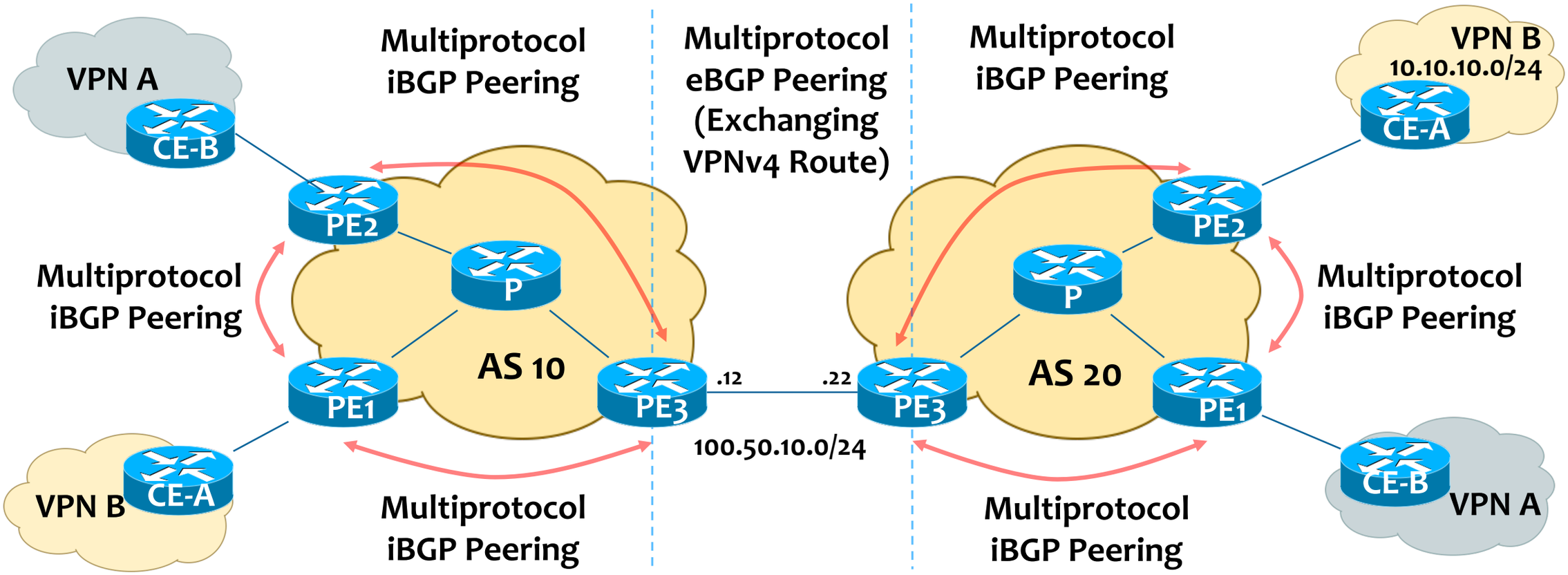
- This maybe complex from the configuration point of view than running BGP for every customer. Routes from MPBGP is redistributed to IGP and vice versa on the ASBRs. ASBR keeps both MPBGP information in the BGP table and so in the in the routing table. Inter AS Option B removes this restriction.

https://t.me/learningnets

- 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.

https://t.me/learningnets




# Inter AS MPLS VPN Option B

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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

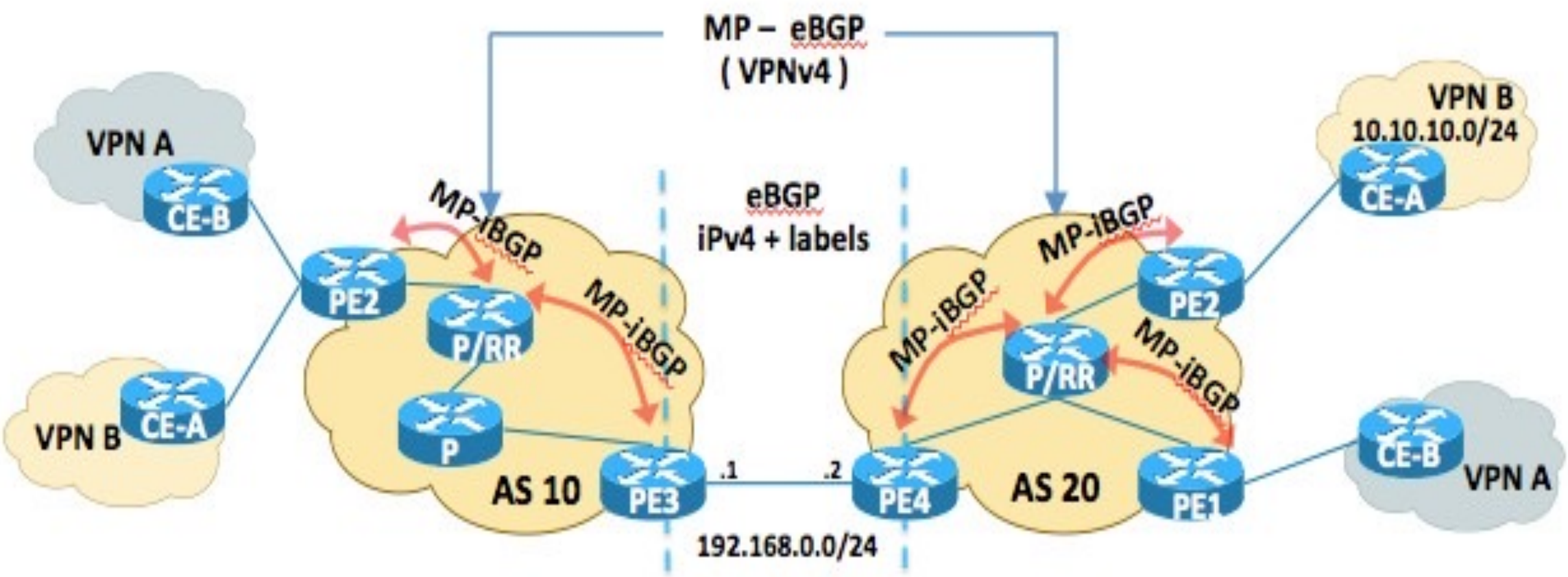
https://t.me/learningnets

# INTER AS MPLS VPN OPTION C

**Inter AS MPLS VPN Option C**

**Instead of Full Mesh MP-IBGP Peering inside an AS, VPN Route Reflector can be used.**

**VPNv4 between Route Reflectors, RFC3107 between ASBRs in Inter-AS Option C**




https://t.me/learningnets

# MPLS Traffic Engineering

- If the requirement is one of the below then the MPLS Traffic Engineering is a good choice

- SLA – Admission Control via RSVP

- Traffic Engineering – Strategic or Tactical

- Traffic Protection – RSVP –TE extension

- IP and MPLS is a destination based routing. With MPLS Traffic Engineering, source routing is achieved.

https://t.me/learningnets

# More Traffic Engineering Drivers

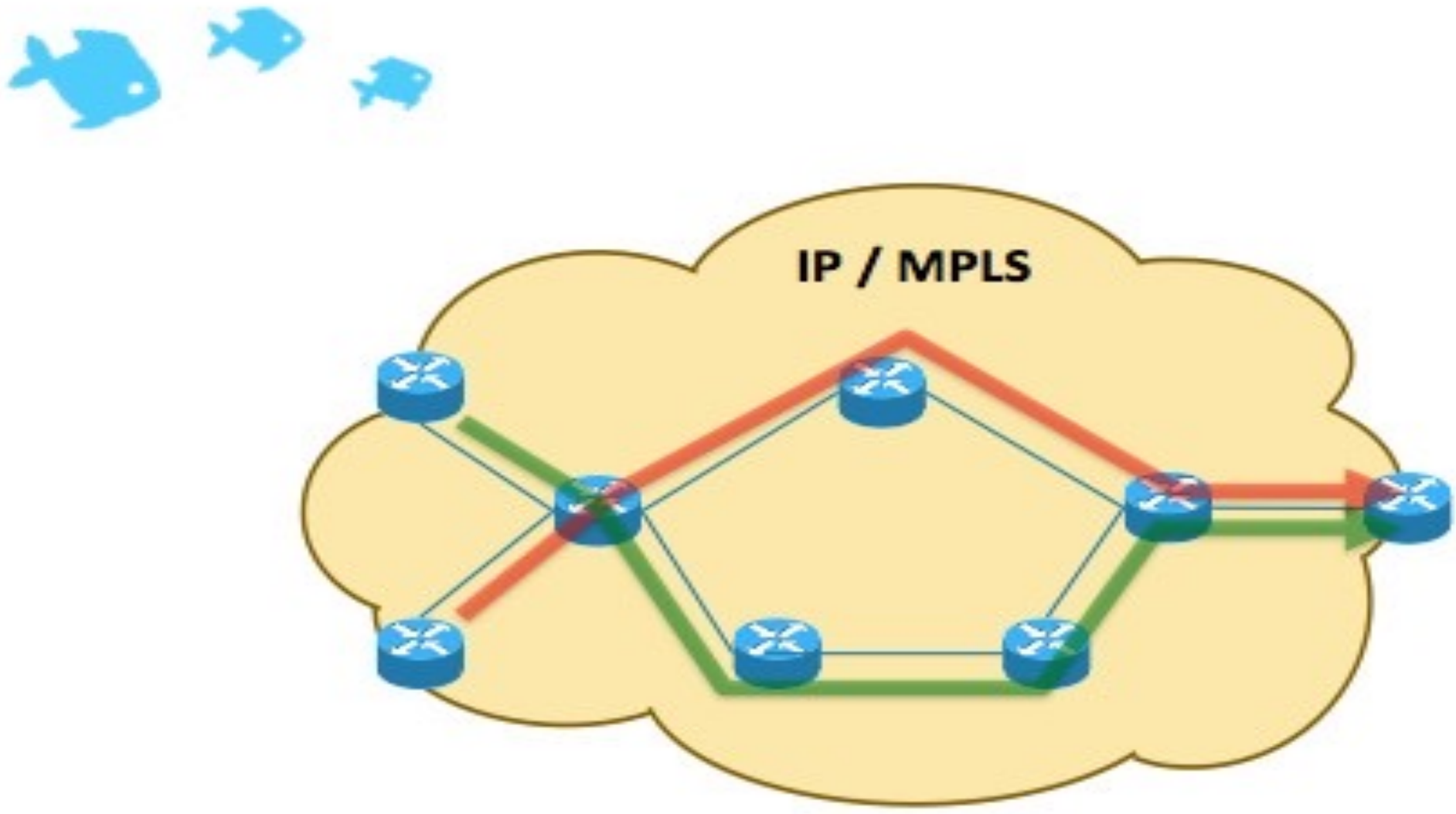
- Resource Optimization
- Disjointness
- Regulatory mandates
- Application Performance

https://t.me/learningnets

# MPLS Traffic Engineering

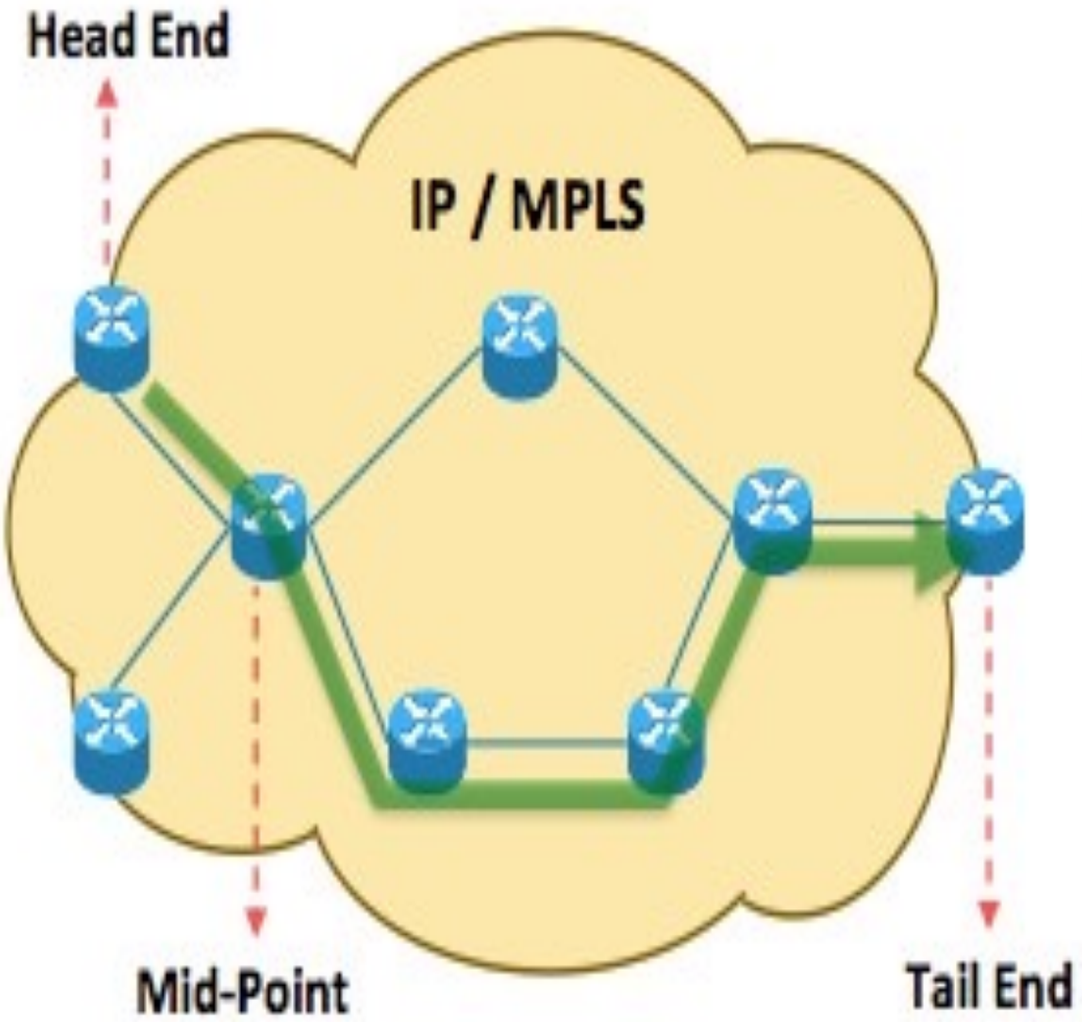
- MPLS Traffic Engineering allows explicit routing.
- From Head-End, entire path can be calculated and signaled.
- Below is a classical fish diagram. If IP/MPLS would be enabled only, destination based shortest path routing would force top path to be used.
- And bottom path would never be used.
- Only if top path fails, after topology convergence, bottom path could be used. MPLS Traffic Engineering may provide optimal traffic usage.

https://t.me/learningnets




**Classical Fish Diagram of MPLS Traffic Engineering.**
**Without MPLS TE, IGP protocols always chooses shortest path.**
**Source routing is not possible with IGP protocols**

https://t.me/learningnets




**Link Information Distribution**

-- ISIS –TE

-- OSPF-TE

**Path Calculation (CSPF)**

**Path Setup (RSVP-TE)**

**Forwarding Traffic down Tunnel**

-- Auto-route announce

-- Static route

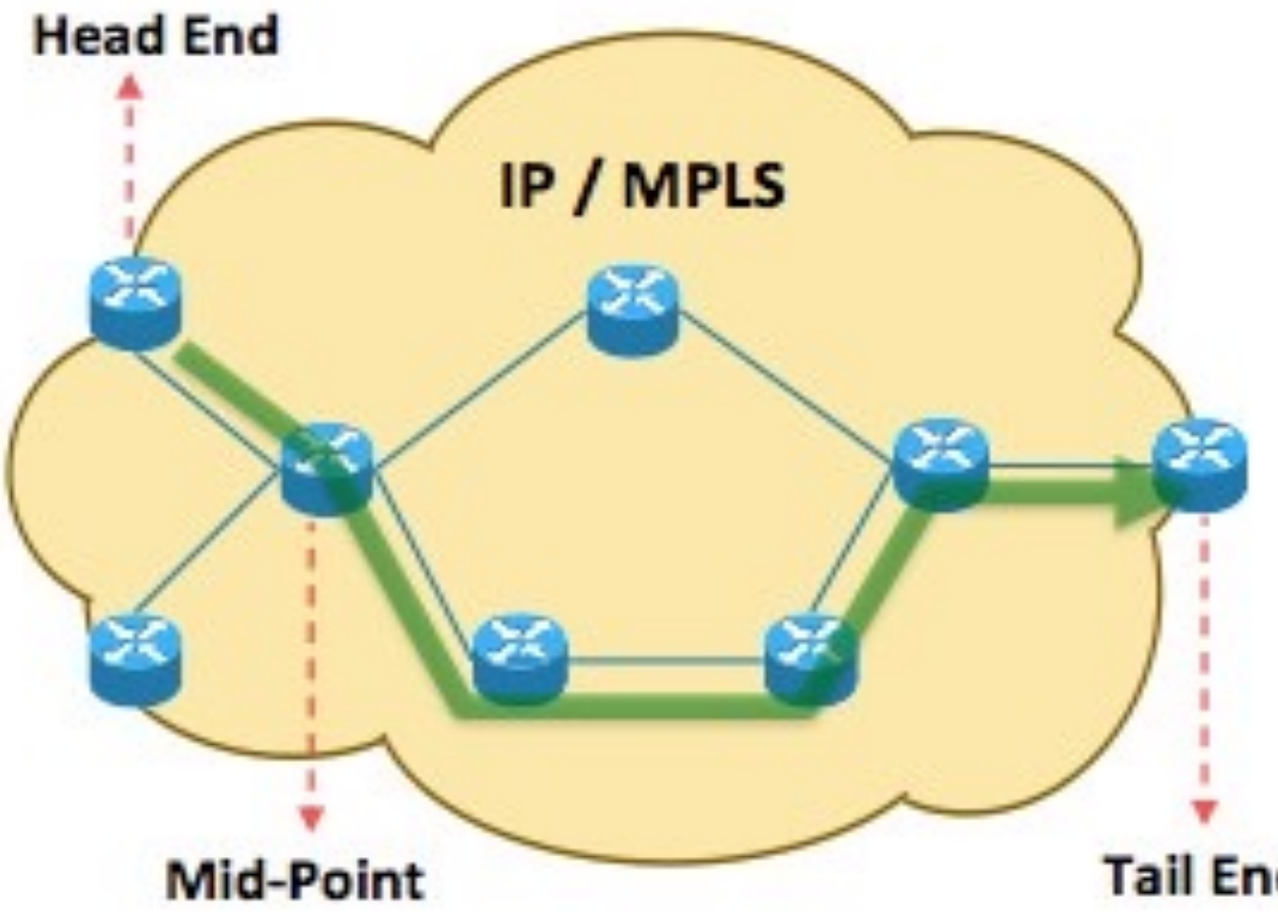
-- CBTS

-- PBR

-- Forwarding Adjacency

-- Pseudowire Tunnel selection

https://t.me/learningnets




**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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- Second step is topology calculation. Topology can be calculated either in distributed manner with CSPF, or as a centralized through NMS, controller etc. Below picture shows how path is calculated for the given constraints.

- MPLS Traffic Engineering allows constraint-based routing. Bandwidth, SRLG, Administrative group can be a constraint.

https://t.me/learningnets

- 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).

https://t.me/learningnets

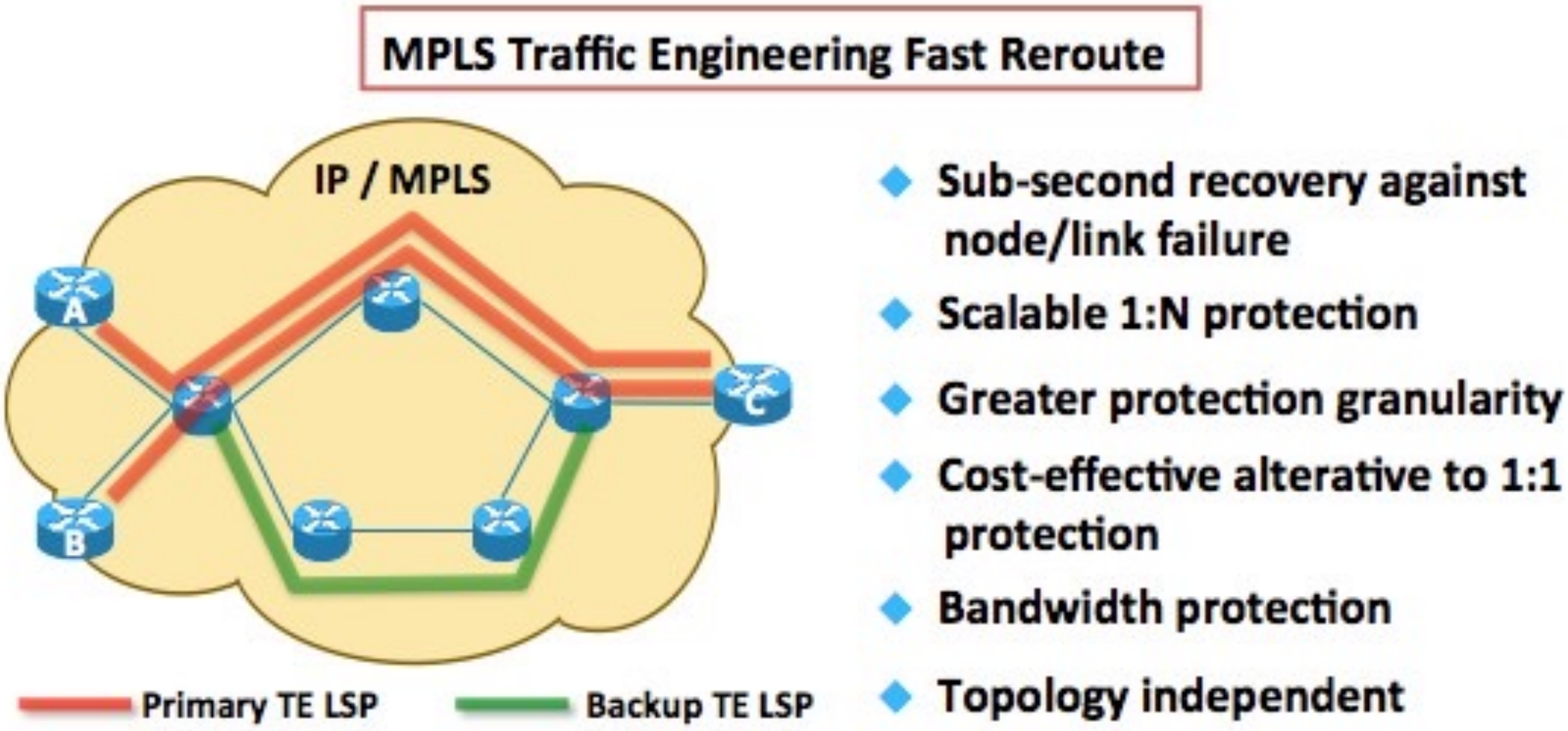
**Link Attributes for MPLS Traffic Engineering Database:**

| Link Attributes for MPLS Traffic Engineering Database | |
|---|---|
| **Additional link Characteristics** | • Interface Address<br>• Neighbor Address<br>• Physical bandwidth<br>• Maximum reservable bandwidth<br>• Unreserved Bandwidth (at eight priorities)<br>• TE metric<br>• Administrative group (attribute flags) |
| **ISDIS or OSPF flood link information** | |
| **All TE nodes build a TE topology database** | |
| **Not required if usingoffDline path computation** | |

https://t.me/learningnets

## MPLS Traffic Engineering Fast Reroute

Fast reroute is a local protection mechanism.




---

The router reacts to a failure is called Point of Local Repair.

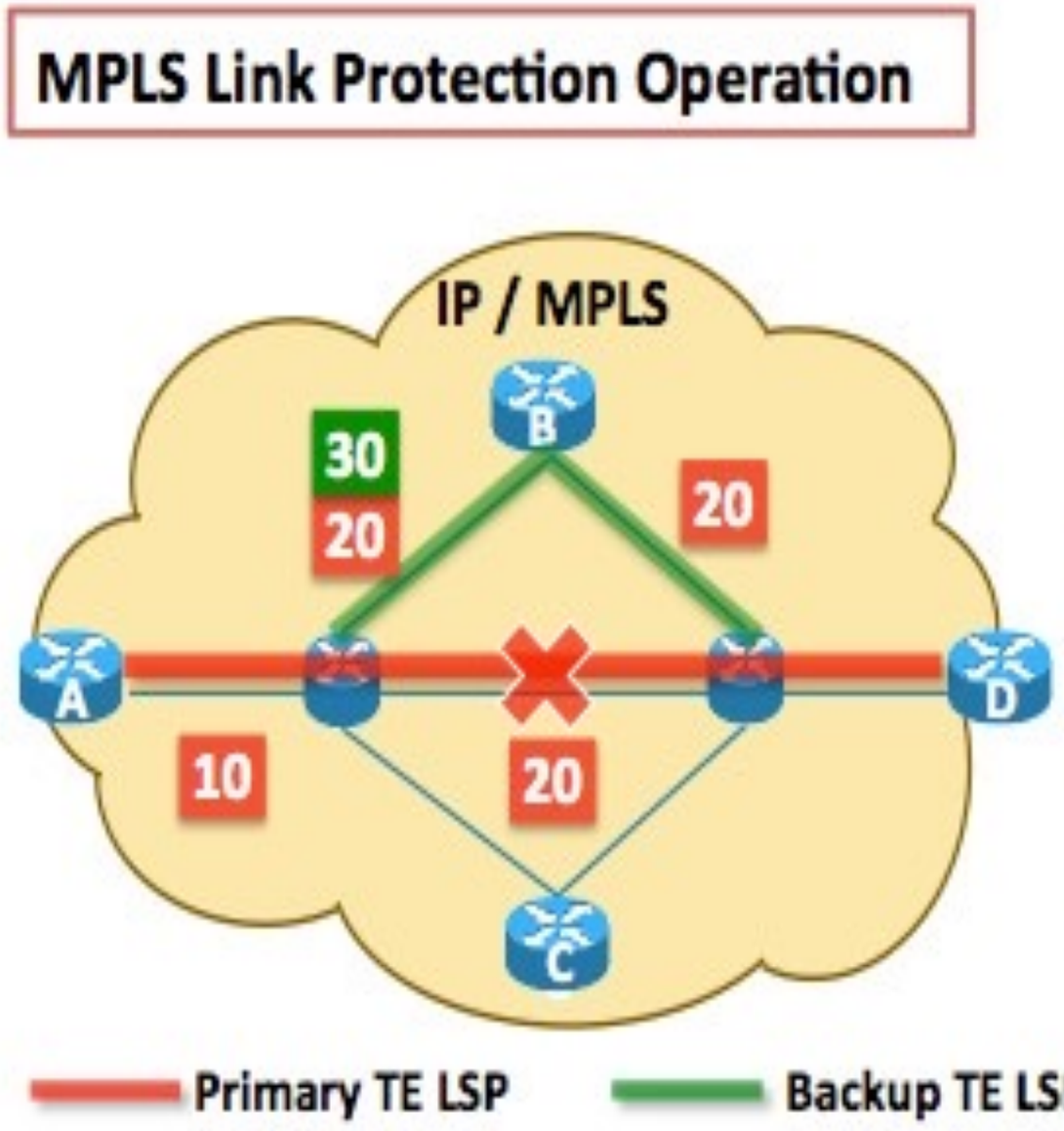
https://t.me/learningnets

- Whenever failure happens, traffic continues through alternate path since MPLS Traffic Engineering Fast reroute is a data plane protection mechanism and backup path can be used as soon as failure is detected

- Control plane still converges. If control plane finds more optimal path than TE FRR backup LSP, then new more optimal primary path is signaled in an MBB (Make Before Break) manner.

https://t.me/learningnets

# MPLS Traffic Engineering Fast Reroute Link Protection




Requires pre-signaled next-hop (NHOP) backup tunnel

Point of local Repair (PLR) swaps the topmost label and pushes backup label

Backup terminates on Merge Point (MP) where traffic rejoin primary LSP

Restoration time expected under ~ 50 ms because failure is local

https://t.me/learningnets

## 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.

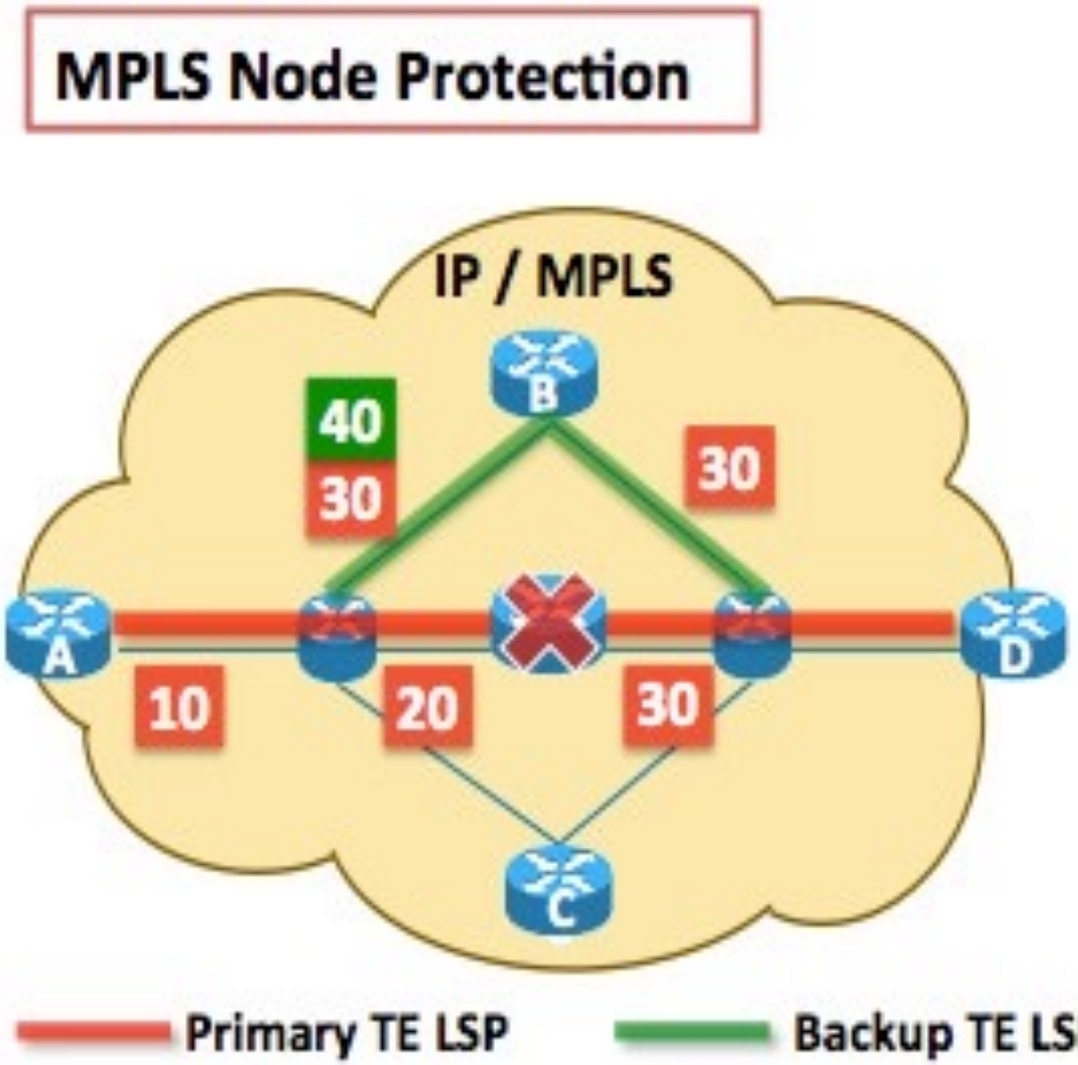
- MPLS traffic engineering fast reroute can cover all the failure scenarios. An IP Fast reroute technology such as LFA (Loop Free Alternate) requires highly mesh topologies to find an alternate path, which will be programmed in the data plane.

https://t.me/learningnets

- 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.

https://t.me/learningnets

# MPLS Traffic Engineering Fast Reroute Node Protection




Requires pre-signaled next-next- hop (NNHOP) backup tunnel

Point of local Repair (PLR) swaps the next-next hop label and pushes backup label

Backup terminates on Merge Point (MP) where traffic rejoin primary LSP

Restoration time depends on failure detection time

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

➢ Please explain the traffic flow between two Head Quarters of Orefe ?

➢ What might the possible problems ? How can those be avoided ?

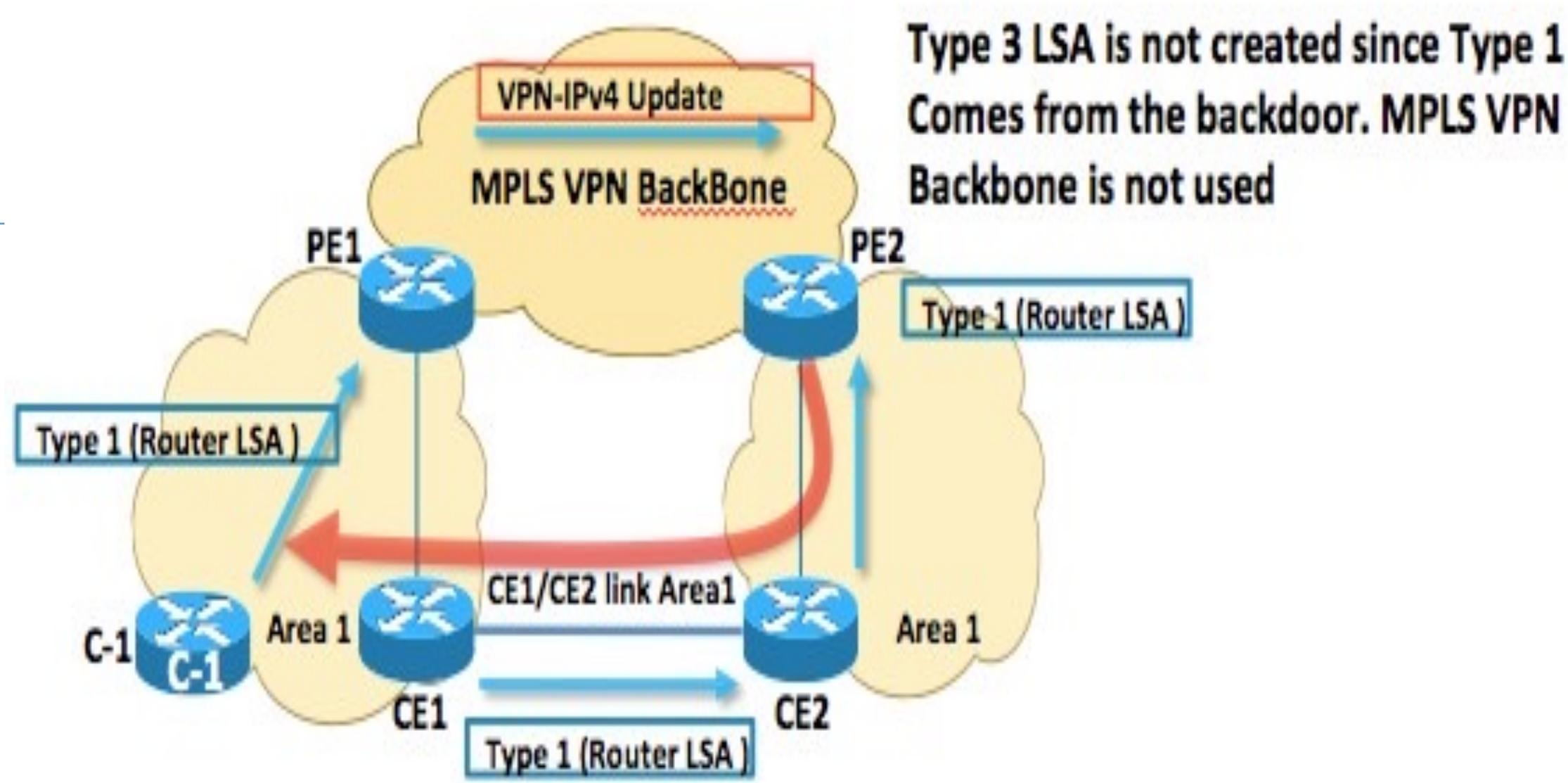
❖ Both Head Quarters are in the OSPF Area1 , Backdoor VPN link is in Area 1 as well. Topology is shown below.

❖ 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.

https://t.me/learningnets

**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.**



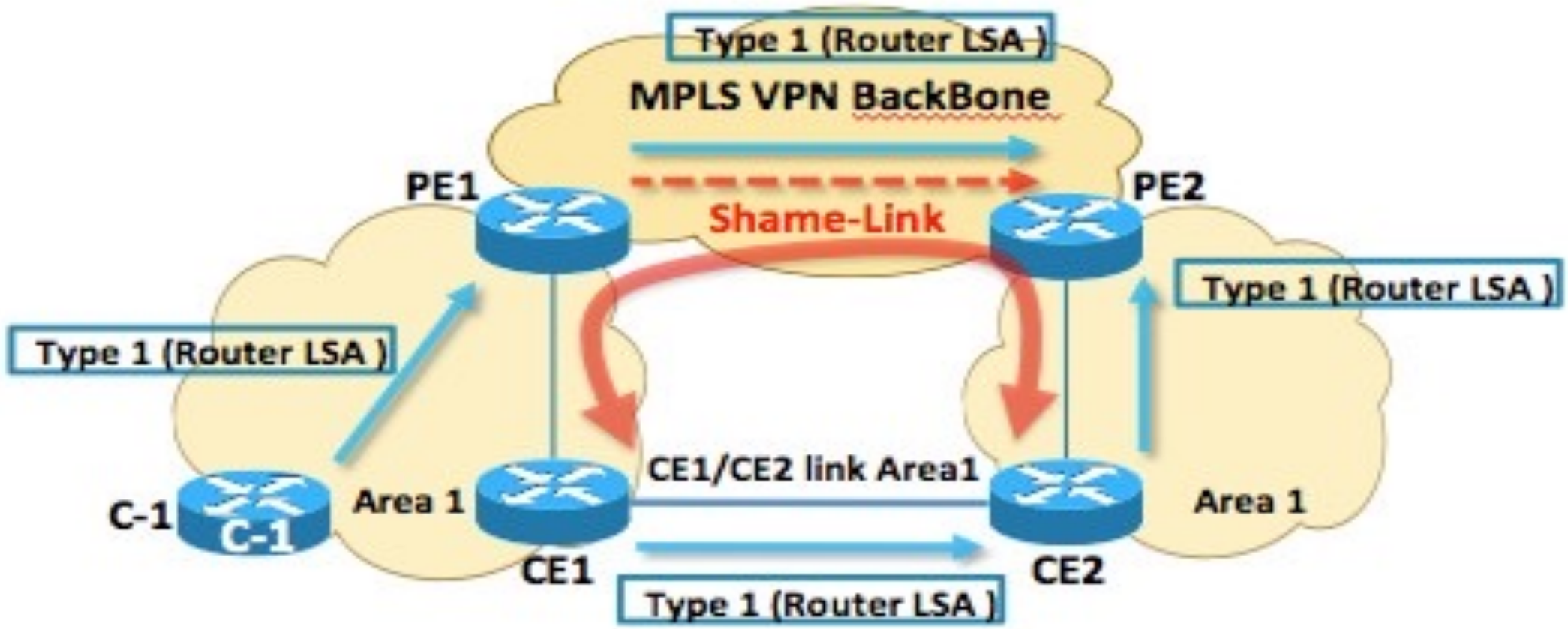

https://t.me/learningnets

- 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.

https://t.me/learningnets




**With only Metric manipulation now,**
**MPLS Backbone can be made preferable**

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

# MPLS Case Study 2

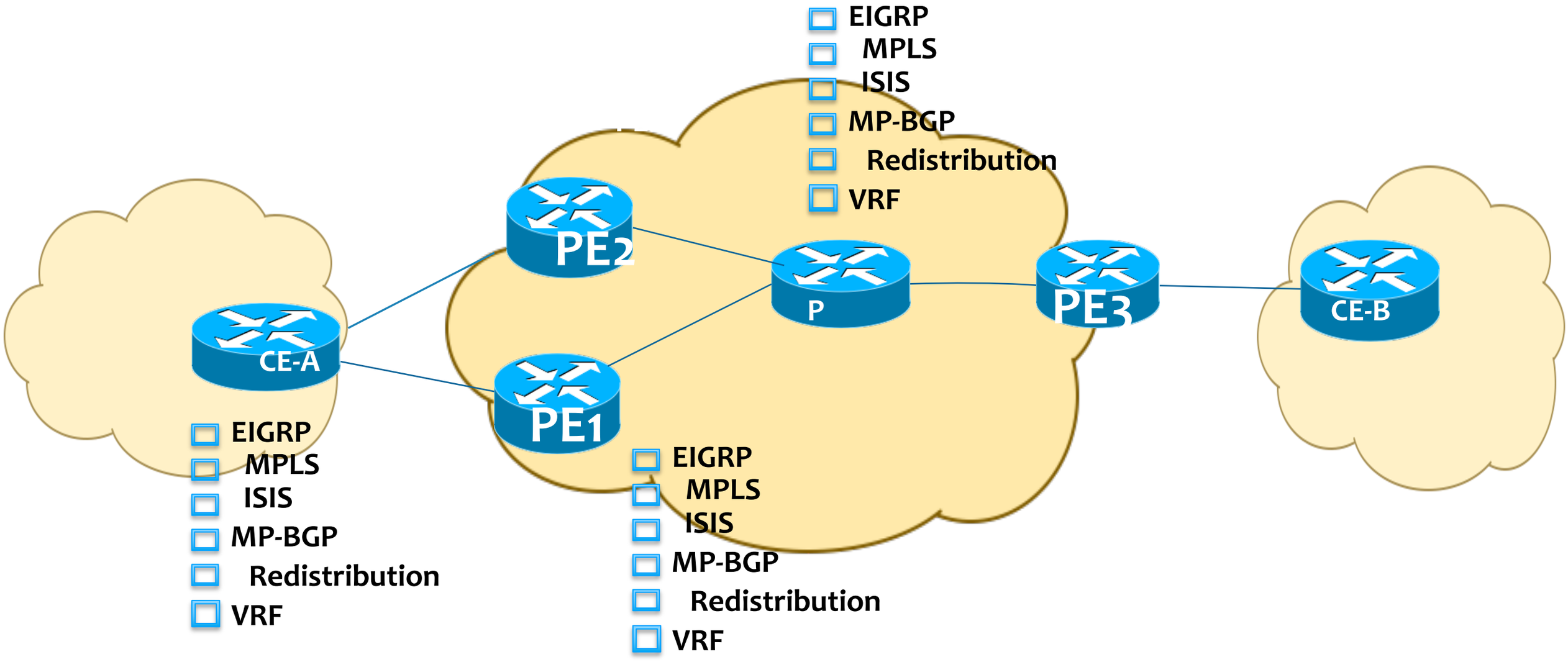
- Enterprise Company is using MPLS Layer 3 VPN for their Wide Area Network connections.

- Their PE-CE IGP protocol is EIGRP

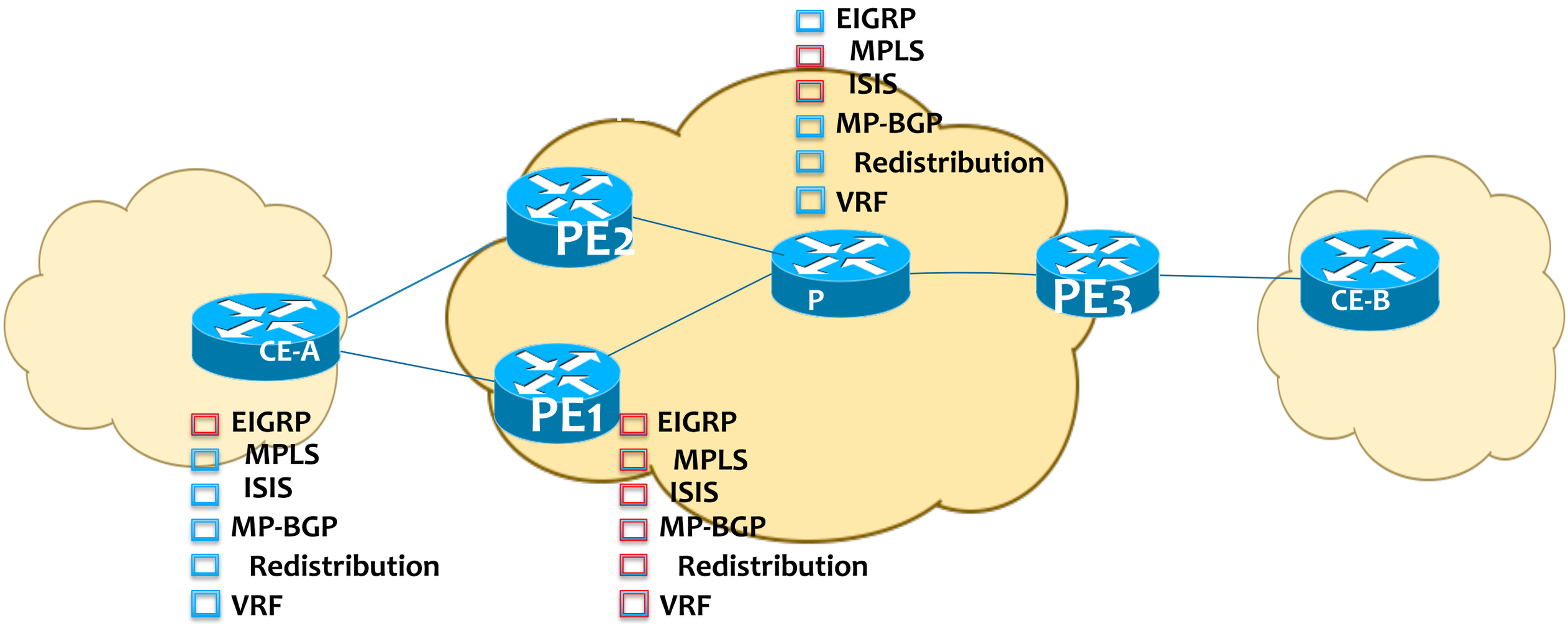
- Service Provider’s of the company is using IS-IS as an internal infrastructure IGP in their backbone and LDP for the label distribution

Please can you select the all required protocols from the check box near to the shown routers ?

https://t.me/learningnets




https://t.me/learningnets




https://t.me/learningnets

**MPLS Case Study – 3**

- Maynet is a fictitious service provider. They have MPLS on their core network. They provide MPLS layer 2 and layer 3 VPN services to their business customers.

- In Access and Aggregation network Maynet doesn't run MPLS but they are also considering enabling MPLS towards Aggregation first and finally to the access networks.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- Fast reroute backup LSP can protect multiple primary LSP, thus in the MPLS Traffic Engineering chapter, it is showed as 1:N protection.

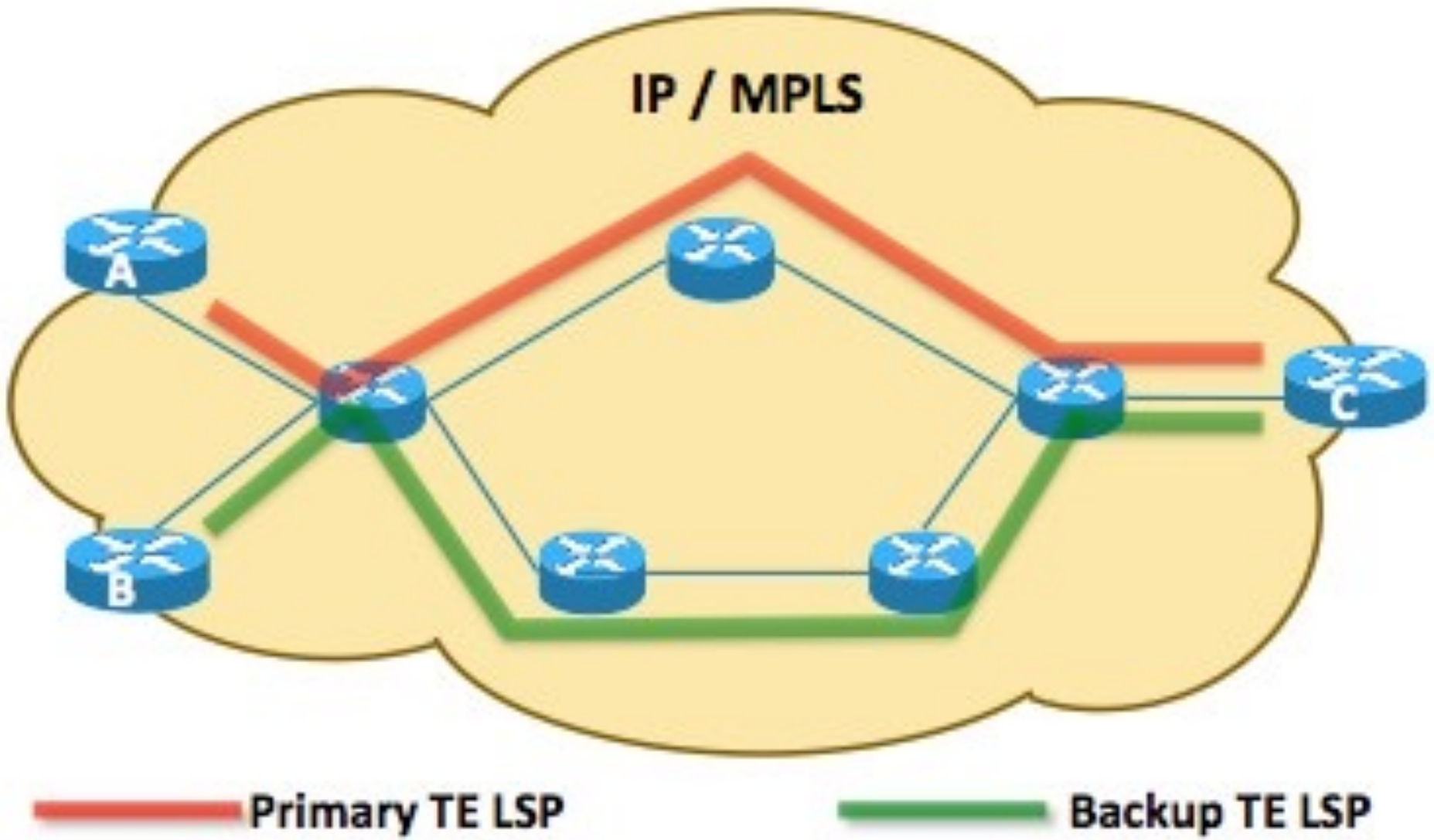
- In contrast, path protection is a 1:1 protection schema where the one backup LSP only protects one primary LSP.

- There are two drawback of path protection.

https://t.me/learningnets

- 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.

https://t.me/learningnets




https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

**MPLS Case Study – 4**

- Smallnet is an ISP which provides Broadband and business internet to their customer. Biggercom is a transit service provider of Smallnet which provides layer 3 IP connectivity between Smallnet sites.

- Smallnet wants to carry all its customers' prefixes through BGP over Biggercom infrastructure. Biggercom doesn't want to carry more than a 1000 prefixes of Smallnet in the Smallnet VRF.

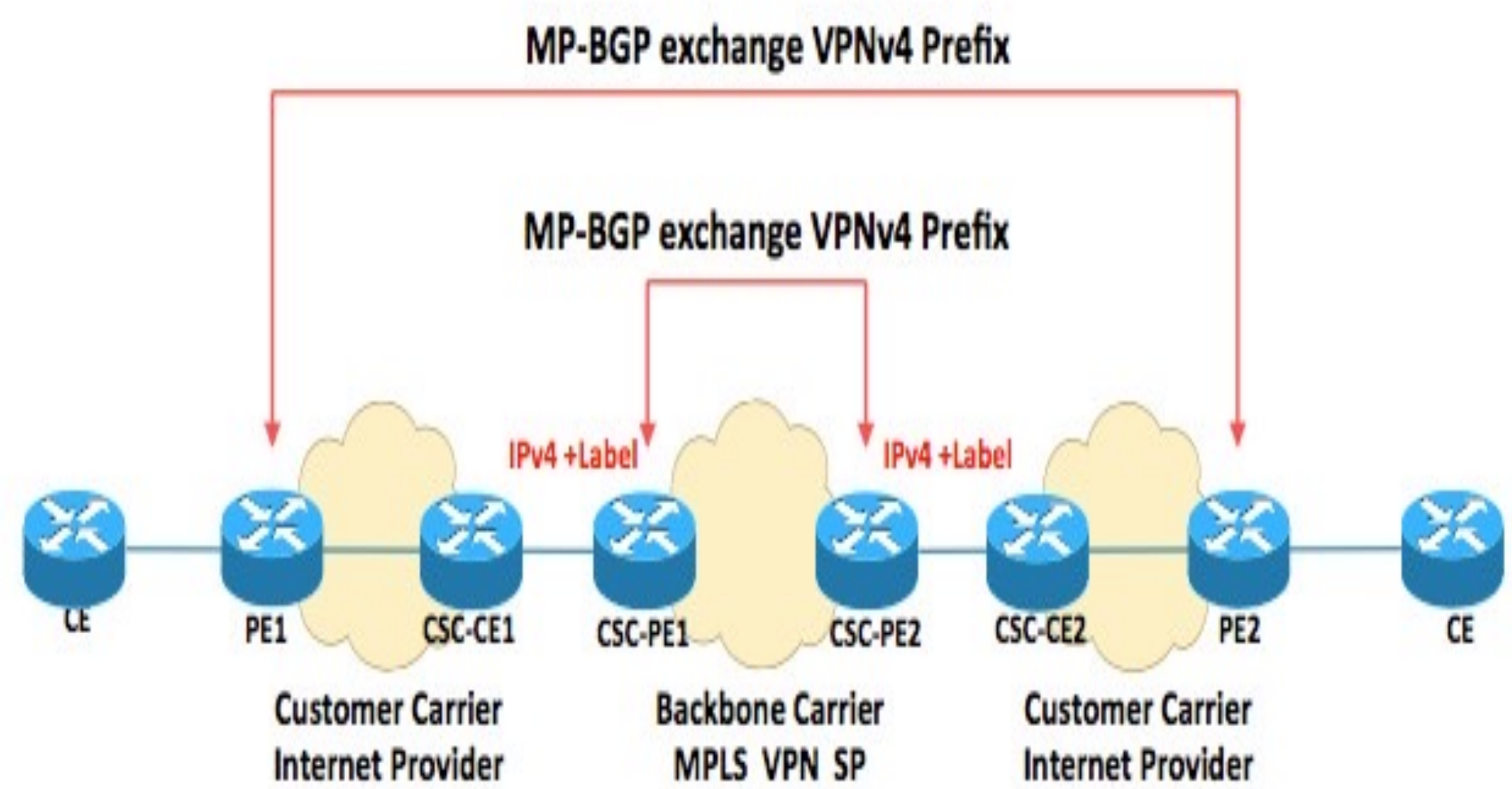
https://t.me/learningnets

❖Smallnet has around 3200 customer prefixes.

❖Provide a scalable solution for Biggercom and please explain the drawback of this design.

❖Since the requirements are ; Biggercom provides IP connectivity , doesn’t carry more than 1000 prefixes of Smallnet, and this is achieved through Carrier Supporting Carrier architecture.

https://t.me/learningnets


MP-BGP exchange VPNv4 Prefix
MP-BGP exchange VPNv4 Prefix
IPv4 +Label
IPv4 +Label
CE
PE1
CSC-CE1
CSC-PE1
CSC-PE2
CSC-CE2
PE2
CE
Customer Carrier
Internet Provider
Backbone Carrier
MPLS VPN SP
Customer Carrier
Internet Provider

❖ In the above picture, Carrier Supporting Carrier architecture is shown.

❖ In the Carrier supporting Carrier terminology, there is a Customer and Backbone Carrier. In our case study, Smallnet is a Customer Carrier, Biggercom is a Backbone Carrier.

❖ There is no customer VRF at the Smallnet network.

https://t.me/learningnets

- 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.

https://t.me/learningnets

❖ 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.

https://t.me/learningnets

- 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.

https://t.me/learningnets

**MPLS Case Study – 5**

- Orko is an Enterprise company which has a store in 7 countries throughout Middle East. Head Quarter and Main Datacenter of Orko is located in Dubai.

- 65 stores of Orko, all connected to datacenter in Dubai via primary MPLS L3 VPN link. Availability of Orko is important so secondary connections to the datacenter is provided via DMVPN over the Internet

https://t.me/learningnets

- 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.

https://t.me/learningnets

❖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.

https://t.me/learningnets

➢ 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 ?

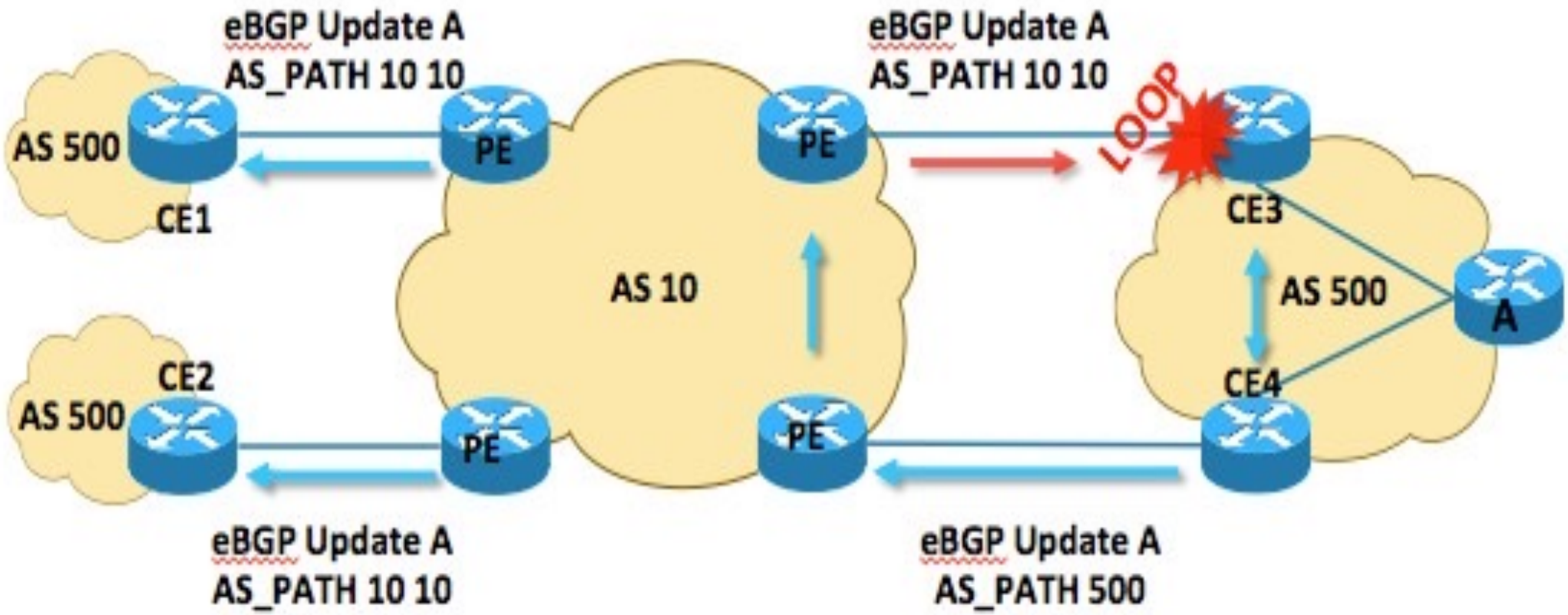
❖ Since Orko is running BGP everywhere and it uses unique AS number, BGP loop prevention mechanism doesn’t allow the BGP prefixes with the same AS in the AS-path.

https://t.me/learningnets

- 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.

https://t.me/learningnets




❖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.

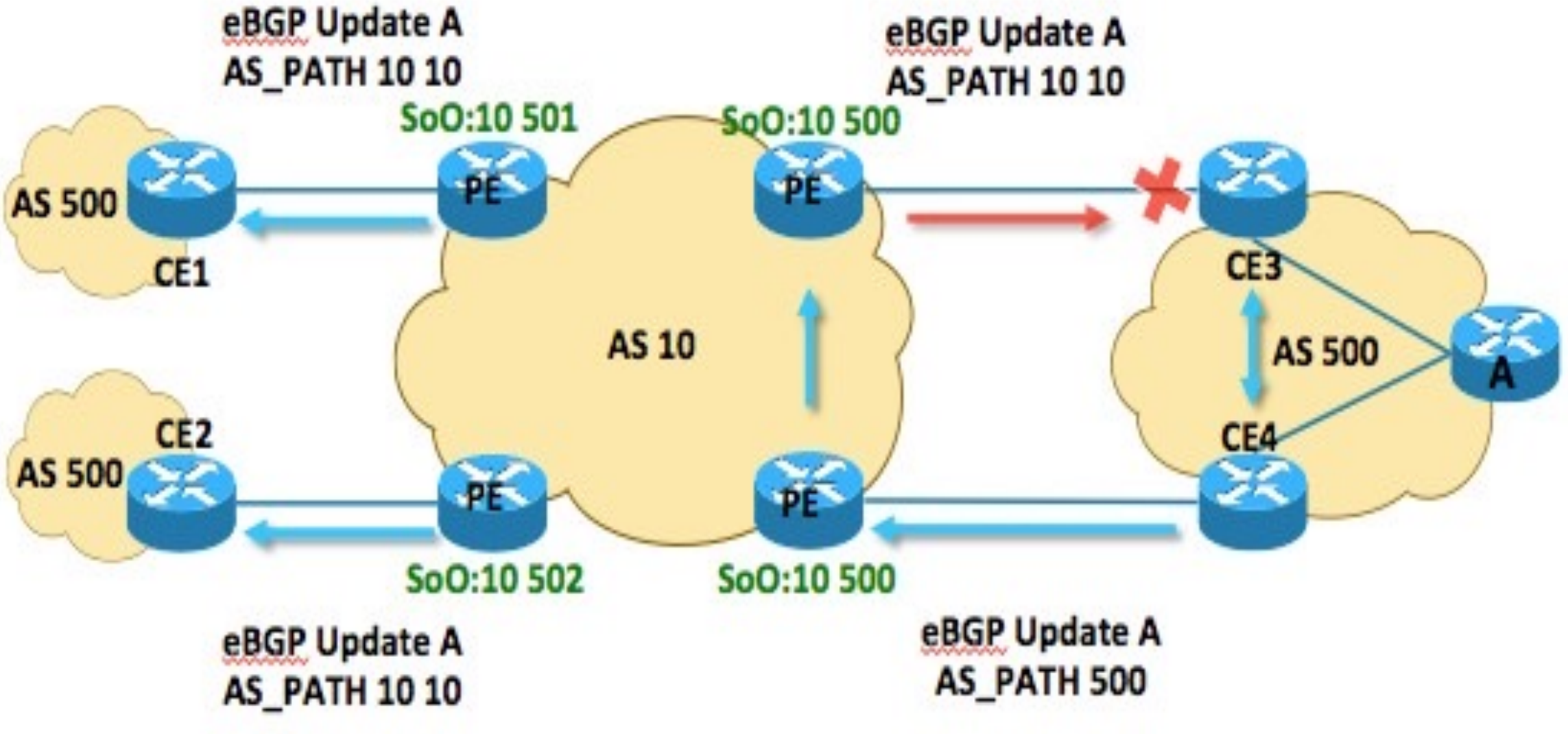
https://t.me/learningnets

- 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

https://t.me/learningnets




- 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

https://t.me/learningnets

# MPLS Case Study -6

Question 1 :

- Two Service Providers company in order to support their customers in other regions decided to create VPN neighborship between them.

- Number of VPN service between them is expected to be too much and their security team definitely disagree to share their internal routing information with other provider.

- Which VPN solution would be the best for this agreement ?

https://t.me/learningnets

- Since the number of expected VPN customers are too much, among the available Inter-AS MPLS VPN options, Option B and Option C is the suitable ones.

- But Option C requires internal routing information such as PE and VPN RR addresses to be leaked between the Service Providers.

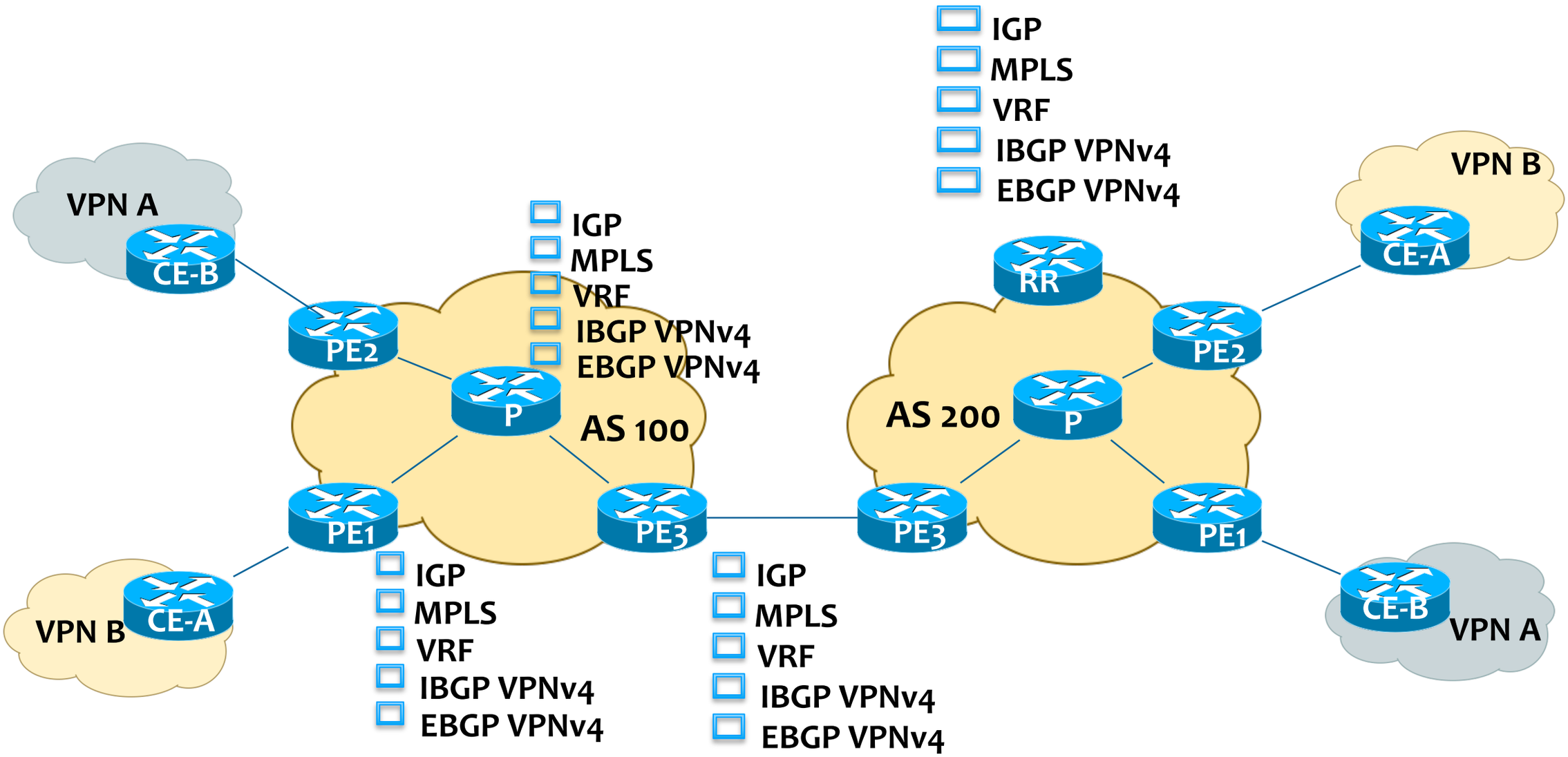
- That’s why the solution based on these requirement is Inter- AS MPLS Option B

https://t.me/learningnets

Question 2 :

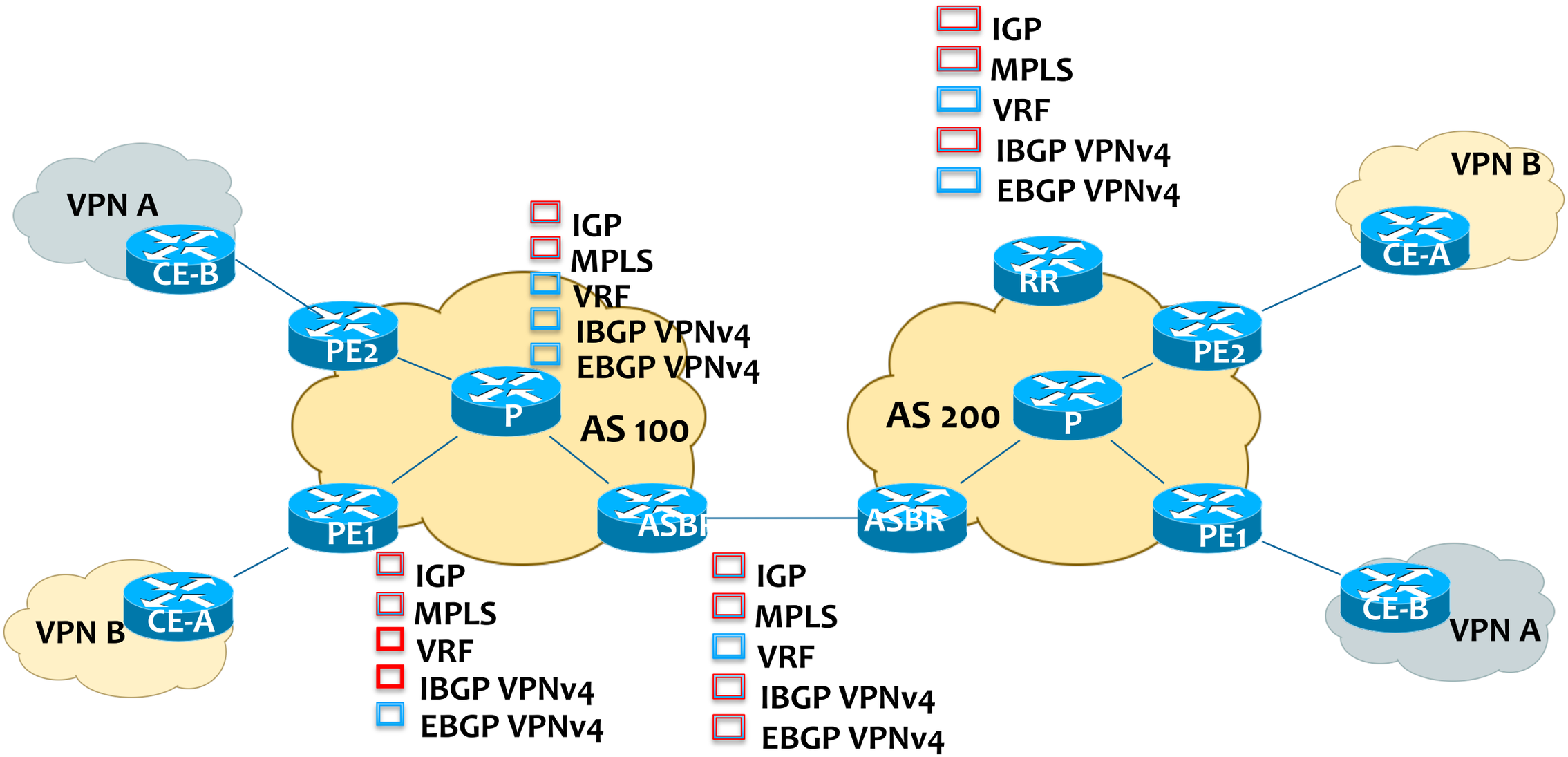
Based on the provided simplified network topologies below of the two Service Providers, please select the protocols which need to be used on the devices which have a check-box next to them.

https://t.me/learningnets




# Inter AS MPLS VPN Option B

https://t.me/learningnets




# Inter AS MPLS VPN Option B

https://t.me/learningnets

# MPLS Case Study 7

- Enterprise company has 6 datacenters. Between the datacenter they have non-IP clustering heartbeat traffic

- They don’t want to implement any vendor specific solution between the datacenters

- Their Service Provider is able to provide MPLS services.

https://t.me/learningnets

Question 1 : Which Datacenter Interconnect solution is most appropriate for this company and why ?

1. OTV
2. LISP
3. EoMPLS
4. TRILL
5. Fabricpath
6. VPLS

https://t.me/learningnets

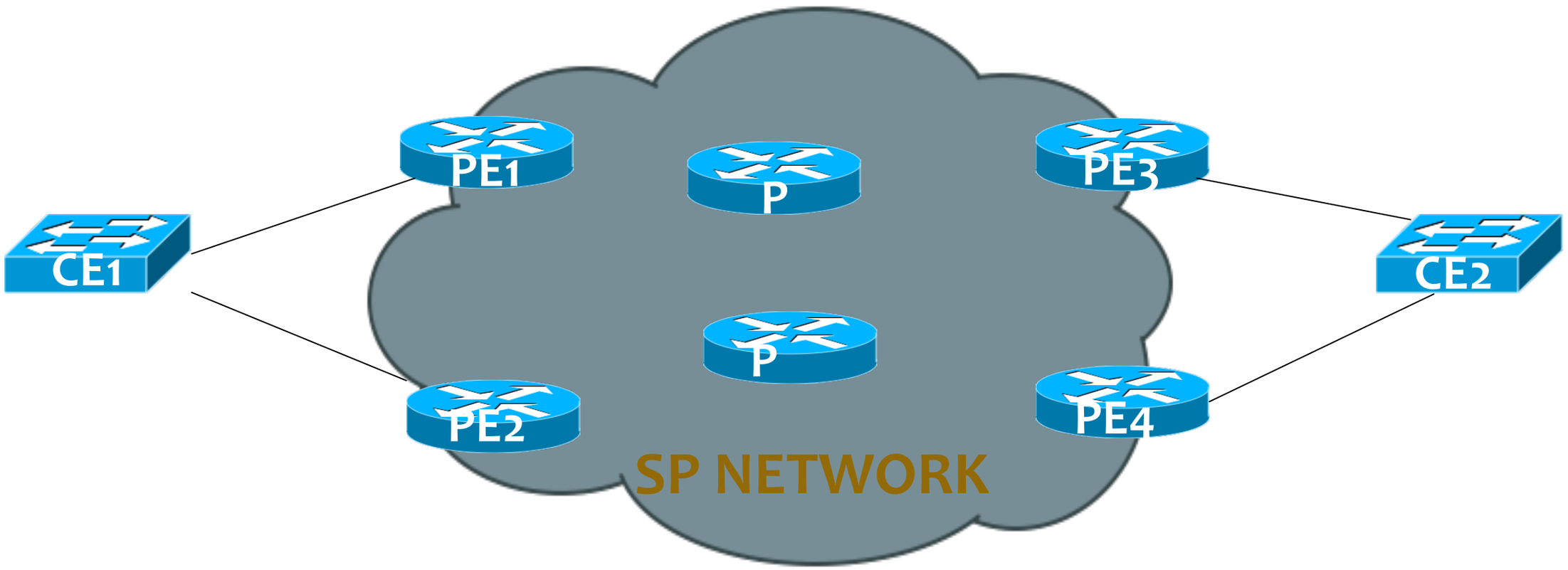
Answer 1 : Company is looking for standard based layer 2 DCI solution. We understand that they are looking for Layer 2 extension since they have an applications which requires non-IP heartbeat.

- Since OTV and Fabricpath Cisco specific solutions, they can not be chosen. Also Fabricpath is not recommended to be used as DCI solution
- LISP is not a L2 extension protocol
- EoMPLS could be used but since company has a lot of datacenters, it is not scalable.
- TRILL is not recommended to be used as DCI solution
- Best option based on the given requirement is VPLS

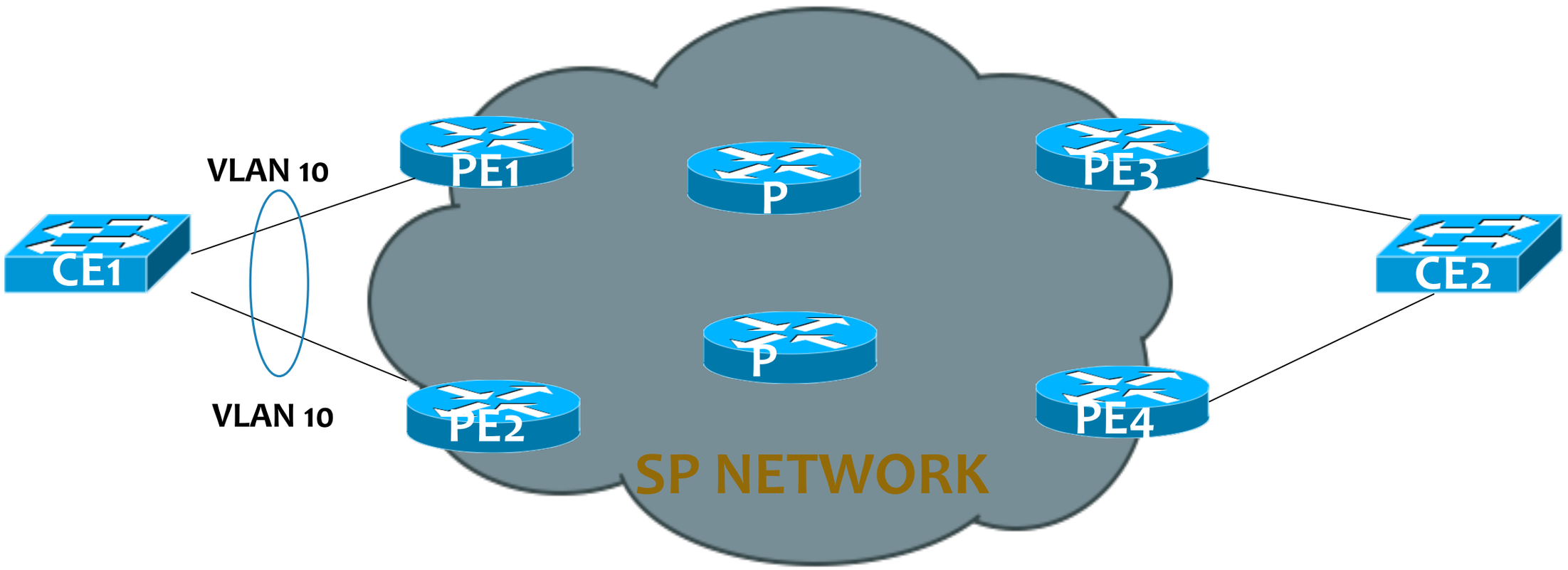
https://t.me/learningnets

- Question 2 : Company sent their topology as it is shown below. They are asking if there is any solution to minimize the effect for specific Vlans in case their DC interconnect switch and the Service Provider link goes down ?

https://t.me/learningnets




https://t.me/learningnets




https://t.me/learningnets

If the requirement is minimum impact for the specific Vlans, we need flow based load
balancing.

Since same vlan but different flows can be carried over the both links with flow based
load balancing, in case one link fails of the bundle, some flows are affected but no
the entire Vlan. This minimizes the affect for the specific Vlans in case of a link
failure.

But unfortunately VPLS can not provide flow based load balancing, since the MAC
address learning is done through data plane.

MAC addresses are advertised through control plane with only via E-VPN
technology

https://t.me/learningnets

over the BGP and only E-VPN can provide flow based load balancing among all the

# MPLS Case Study 8

- Enterprise company is using legacy frame relay circuits and they admit that they are using very old and limited technology and wants to upgrade their network
- They decided to receive an MPLS VPN service from their regional Service Provider.
- For many years company has been managing their WAN networks themselves and they want to continue to manage themselves.
- Service Provider presented Ethernet over MPLS, VPLS and MPLS Layer 3 VPN offerings and they are very confused now.
- Can you help them to understand the MPLS VPN offerings advantage and disadvantages by filling the below chart and recommend a best service which fulfills their requirements

https://t.me/learningnets

# MPLS QUIZ

➢ **Question 1: What was the initial purpose of MPLS?**

- ❖ Avoid IP destination based lookup and increase the performance of the routers
- ❖ MPLS VPNs
- ❖ MPLS Traffic Engineering
- ❖ Alternative to the link state routing protocols
- ❖ Virtualization

https://t.me/learningnets

➢ **Question 2: Which below options are the characteristics of MPLS layer 2 VPNs?**

- ❖ MPLS layer 2 VPN allows carrying of Layer 2 information over service provider backbone.
- ❖ Layer 2 VPN can provide point to point type of connectivity between the customer sites.
- ❖ It is used to carry layer 3 information of the customers over the Service Providers
- ❖ It is used for datacenter interconnect
- ❖ Layer 2 VPN can provide point to multi point type of connectivity between the customer sites.

https://t.me/learningnets

➢ **Question 3: Which below options are used in the MPLS header?**

❖ 20 bits MPLS label space.

❖ Link cost

❖ 12bit TTL field.

❖ 3bit EXP bit for the QoS.

❖ Protocol number

https://t.me/learningnets

➢ **Question 4: Which option below can be used as a PE-CE routing protocol in MPLS Layer 3 VPN?**

- ISIS
- BGP
- PIM
- HSRP
- OSPF
- Static Route

https://t.me/learningnets

➢ **Question 5: What are the two possible options to create MPLS layer 2 VPN Pseudowire?**

- ❖ Martini Draft, LDP signalled Pseudowires.
- ❖ Segment Routing
- ❖ BGP EVPN
- ❖ Rosen GRE Draft
- ❖ Kompella Draft, BGP signalled Pseudowires.

https://t.me/learningnets

➢ **Question 6: Why Sham-link is used in the context of OSPF PE-CE deployment?**

❖ In order to have type 1 LSA from the service provider edge router.

❖ In order to have type 3 LSA from the service provider edge router

❖ In order to have type 5 LSA from the service provider edge router

❖ In order to have type 1 LSA in the service provider core

❖ In order to have type 3 LSA in the service provider core

https://t.me/learningnets

➢ **Question 7: Which below options are the results of having MPLS in the network?**

❖ BGP Free Core

❖ Hiding service specific information ( customer prefixes etc ) from the core.

❖ More scalable network

❖ Faster convergence

❖ Better security

https://t.me/learningnets

➢ **Question 8: If customer is looking to carry layer 2 traffic with the encryption, which below options can be chosen?**

- ❖ VPLS
- ❖ EoMPLS
- ❖ GET VPN
- ❖ GRE over IPSEC
- ❖ IPSEC
- ❖ L2tpv3

https://t.me/learningnets

➢ **Question 9: Which below options are correct for the Inter as MPLS VPN Option A?**

- ❖ It provides most flexible QoS deployment compare to other Inter-AS MPLS VPN options
- ❖ It is least secure Inter-AS option
- ❖ It is most scalable Inter-AS option
- ❖ It requires MPLS between the Autonomous Systems
- ❖ BGP+Label (RFC3107) is used between two Autonomous Systems

https://t.me/learningnets

➢ **Question 10: Which below terms are used to define a label which provides reachability from one PE to another PE in MPLS networks?**

❖ Topmost Label.

❖ Transport Label.

❖ Outer Label.

❖ VC Label

❖ VPN Label

❖ Inner Label

https://t.me/learningnets

➢ **Question 11: Which options provide Control Plane MAC address advertisement for MPLS layer 2 VPNs?**

❖ EVPN

❖ VPLS

❖ EoMPLS

❖ BGP L3VPN

❖ PBB EVPN

❖ VXLAN EVPN

https://t.me/learningnets

# 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

https://t.me/learningnets

## ❖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

https://t.me/learningnets

## ❖ Articles:

- http://orhanergun.net/2015/02/carrier-supporting-carrier-csc/
  http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/WAN_and_MAN/L3VPNCon.html
  http://orhanergun.net/2015/06/advanced-carrier-supporting-carrier-design/
  http://d2zmdbbm9feqrf.cloudfront.net/2013/usa/pdf/BRKMPL-2100.pdf
  https://routingfreak.wordpress.com/tag/h-vpls/
- http://blog.ine.com/2010/08/16/scaling-mpls-networks/
  http://www.networkcomputing.com/networking/mpls-traffic-engineering-guide-success-and-alternatives/d/d-id/1269268
- https://www.ietf.org/proceedings/49/slides/ppvpn-11.pdf
  http://searchtelecom.techtarget.com/tip/Making-the-case-for-Layer-2-and-Layer-3-VPNs
  http://www.huawei.com/au/static/HW-076762.pdf
  http://www.cisco.com/c/en/us/td/docs/ios/12_0s/feature/guide/fsldpsyn.html

https://t.me/learningnets

# Wireless Design Fundamentals

https://t.me/learningnets

# Wireless Design Fundamentals Course Agenda

## Wireless Basics

- Wireless Terms and Terminologies
- Types of Wireless Networks
- Wireless Standards

## WLAN Architecture

- Autonomous WLAN architecture
- Centralized WLAN architecture
- Distributed WLAN architecture

## WLAN Design

- Coverage Design
- Roaming Design
- Channel Design
- Capacity Design

https://t.me/learningnets

# Wireless Terms and the Terminologies

Before we start Wireless Design and Architecture, we will have a look at many Wireless Terms and Terminologies and different types of Wireless LAN technologies in high level

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Antenna:

It is a component of a device dedicated to sending and/or receiving wireless signals

Any device meant to send or receive wireless signals must have one or more antennas either externally and visibly, or internal antennas, which are usually not visible

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Airtime:

Airtime is a measure of the wireless capacity in a given area and is measured in percentages

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Access Point:

Device that has one or more wireless radios that allow other devices to connect wirelessly to connect to the Network

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Beamforming:

It is a technology that shapes/directs the wireless signal towards a wireless client/device to deliver a stronger signal

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Band Steering:

- A form of client steering, band steering is functionality that allows the wireless access point(s) to determine which frequency band a device should be connected to based on what the device supports and which frequency band will provide the better performance under the circumstances

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Captive Portal:

- In wireless networking, a captive portal is a process running on an AP that can intercept and redirect clients who have associated to a web page where they must agree to terms of service, provide a password, or even purchase access

- These are common in hotels, airports, guest networks, and other locations that offer Internet access but want to charge a fee and/or restrict it to authorized users

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Coverage:

- Wireless coverage is a term used to describe how far wireless signals reach at a satisfactory strength

- The best way to measure the wireless coverage in a given area, is to use a tool known as a heatmapper

https://t.me/learningnets

# Wireless Terms and the Terminologies

## DBM:

- Wi-Fi signal strength is measured in dBm, which is short for decibel-milliwatt

It is provided as a negative value, and under normal circumstances, signal strength close to a wireless access point will be around -30 dBm

https://t.me/learningnets

# Wireless Terms and the Terminologies

## DFS:

- DFS is short for Dynamic Frequency Selection
- The term usually describes functionality that allows a wireless channel to prioritize radar signals
- This means that a DFS-enabled access point must change the channel if detecting that a radar is sending signals using the channel it is currently on

https://t.me/learningnets

# Wireless Terms and the Terminologies

**Frequency band:**

- A frequency band is a section of a lager frequency area. Wi-Fi uses two frequency bands: 2.4 GHz and 5 GHz

- 2.4 GHz is the frequency band used by older wireless technologies (802.11 b/g/n), which in practice means the slowest connections

- Because of the long reach of its signals, this frequency band is also the most vulnerable to interference from other Wi-Fi networks, as well as other wireless equipment (Bluetooth, baby monitors), and microwave ovens

https://t.me/learningnets

# Wireless Terms and the Terminologies

**Frequency band:**

- 5 GHz is the frequency band used by newer wireless standards such as 802.11ac, as well as some of the older (802.11 a/n), in practice the fastest connections

- It is known to provide the wireless signals with the shortest reach, but the highest speeds/best performance, which is why it is the preferred frequency band for new laptops and (most) wireless device

- Because of the shorter reach of the signals on this frequency band, multiple access points are often necessary to provide satisfactory coverage in a home

https://t.me/learningnets

# Wireless Terms and the Terminologies

## ISM Band:

- The Industrial, Scientific, and Medical frequency bands are unlicensed bands used by a variety of devices for wireless connectivity

- In the 2.4 GHz ISM band, 802.11b and g network devices, Bluetooth devices, NFC devices, baby monitors, and microwave ovens all compete for bandwidth

https://t.me/learningnets

# Wireless Terms and the Terminologies

**Interference:**

- Interference is disruptions or noise that destroys the wireless signals and consumes airtime in an area

- Sources of interference can be neighboring networks, multiple networks in the same house, microwaves, baby monitors, and wireless devices that use Bluetooth

- Today, interference is primarily an issue on the 2.4 GHz frequency band

https://t.me/learningnets

# Wireless Terms and the Terminologies

**Channel:**

- A given area of a frequency band
- The rules and regulations for which channels are permitted for private wireless communication vary between countries and regions
- For example, channel 13 on the 2.4 GHz band is commonly used in Europe, but not allowed in the United States

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Channel:

- When many wireless networks in the same area transmit on the same channel, it can cause a lot of interference and have devastating effects on performance

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Channel Bonding:

- In 802.11b and g networks, multiple channels can be combined to obtain greater throughput when all access points and clients can support it

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Clients:

Computers, mobile phones, tablets, and all other devices that need to connect to the wireless network are wireless clients

https://t.me/learningnets

# Wireless Terms and the Terminologies

**Radio:**

- Radio waves are the most common form of wireless communication

- All Wi-Fi and Bluetooth devices contain radio transmitters and receivers, connected to internal or external antenna

- A device that can transmit on multiple frequency bands usually has one radio for each frequency band

https://t.me/learningnets

# Wireless Terms and the Terminologies

## Roaming:

- In a wireless network with multiple APs, a client that is moving from the coverage area provided by one AP to that provided by another is roaming

- It must disassociate from the first AP before it can associate to the next AP

https://t.me/learningnets

# Wireless Terms and the Terminologies

## RSSI:

- Received Signal Strength Indication, how strong the signal is as received by the antenna of a wireless device

https://t.me/learningnets

# Wireless Terms and the Terminologies

**SSID:**

- The network name that wireless access points transmit so that you can find the network and connect to it

- An SSID can be shared between several access points on the same wireless network

https://t.me/learningnets

# Wireless Terms and the Terminologies

## WLAN:

- Wireless Local Area Network, a formal generic term for wireless LANs using IEEE 802.11 standards

https://t.me/learningnets

# Types of Wireless Networks

- **Wireless LAN**
- **Wireless MAN**
- **Wireless PAN**
- **Wireless WAN**

https://t.me/learningnets

# Types of Wireless Networks

## WLAN

- Wireless LAN (WLAN) technology provides internet access within a building or a limited outdoor area

- First used within offices and homes, WLAN technology is now also used in stores and restaurants

https://t.me/learningnets

# Types of Wireless Networks

## Wireless MAN

- Wireless metropolitan area networks have been installed in cities worldwide to provide access for people outside an office or home network
- These networks cover a wider area than office or home networks, but the principles are the same
- APs are located on the sides of buildings or on telephone poles throughout the coverage area
- APs are connected to the internet via a wired network and broadcast a wireless signal throughout the area

https://t.me/learningnets

# Types of Wireless Networks

## Wireless PAN

- Wireless personal area networks cover a very limited area -- typically a maximum of 100 meters for most applications -- using protocols like Bluetooth and Zigbee

- Bluetooth enables hands-free phone calls, connects a phone to earpieces or transmits signals between smart devices

- Zigbee is used in many IOT networks

https://t.me/learningnets

# Types of Wireless Networks

## Wireless WAN

- Wireless WANs use cellular technology to provide access outside the range of a wireless LAN or metropolitan network

- These networks enable users to make phone calls to others connecting either through a wireless WAN or a wired telephone system

- Users can also connect to the internet to access websites or server-based applications

https://t.me/learningnets

# Types of Wireless Networks

| | Wireless LAN (WLAN) | Wireless MAN (WMAN) | Wireless PAN (WPAN) | Wireless WAN (WWAN) |
|---|---|---|---|---|
| TYPE OF NETWORK | Local area network | Metropolitan area network | Personal area network | Wide area network |
| GOAL | Provide internet access within a building or limited outdoor area | Provide access outside office and home networks, typically regional | Transmit signals between devices in limited areas, typically 100 meters | Provide access outside the range of WLANs and WMANs |
| CONNECTIVITY | Cellular | IEEE 802.16 WiMax | Bluetooth, Zigbee and infrared | LTE |

https://t.me/learningnets

# Wireless Standards

- Wireless standards are a set of services and protocols that dictate how your Wi-Fi network (and other data transmission networks) acts

- The most common set of wireless standards you will encounter is the IEEE 802.11 Wireless LAN (WLAN) & Mesh

- The IEEE updates the 802.11 Wi-Fi standard every few years

https://t.me/learningnets

# Wireless Standards

- The 802.11 wireless networking protocols have gone through several iterations, each one surpassing the previous version in capability and speed

https://t.me/learningnets

# Wireless Standards

- Not all equipment functions with every version, so it's important to know which protocol version your equipment uses

- In general, newer equipment supports the most recent of the protocols, and older equipment may not

https://t.me/learningnets

# Wireless Standards

- Typically, equipment supports multiple protocols

- For example, equipment tagged 802.11ac/n/g is compatible with three protocols

https://t.me/learningnets

# 802.1b Protocol

802.11b was the first wireless standard to be widely adopted in homes and businesses

Its introduction is credited with the rise in popularity of hotspots and staying connected during travel

Equipment using 802.11b was comparatively inexpensive and built into many laptops

https://t.me/learningnets

# 802.1a Protocol

- The 802.11a standard operates in a different frequency range

- By broadcasting in the 5 GHz range exclusively, 802.11a devices run into less competition and interference from household devices

- 802.11a is capable of transmission speeds up to 54 Mbps like the 802.11g standard

https://t.me/learningnets

# 802.1g Protocol

- The 802.11g standard improves on 802.11b
- It uses the same crowded 2.4 GHz shared by other common household wireless devices, but 802.11g is faster and capable of transmission speeds up to 54 Mbps
- Equipment designed for 802.11g still communicates with 802.11b equipment
- However, mixing the two standards is not usually recommended

https://t.me/learningnets

# 802.1n Protocol (Wi-Fi 4)

- 802.11n uses multiple-input/multiple-output (MIMO) technology and a wider radio frequency channel than its predecessors

- It increases a wireless local area network (WLAN) speed and improves reliability

- Operating at 600 Mbps, it offers 10 times the speed of 802.11g and uses both the 2.4 GHz and 5 GHz bands

https://t.me/learningnets

# 802.1n Protocol (Wi-Fi 4)

- Introduced in 2009, this version had slow initial adoption

- 802.11n operates on both 2.4GHz and 5GHz, as well as supporting multi-channel usage

- Each channel offers a maximum data rate of 150Mbps, which means the standard's maximum data rate is 600Mbps

https://t.me/learningnets

# 802.1ac Protocol (Wi-Fi 5)

- 802.11ac, also known as Wi-Fi 5, added Dual Band support.
- It can use the 2.4 GHz band and the 5 GHz band at the same time
- 802.11ac is approximately three times faster than 802.11n
- This protocol provides support for eight streams, up from four in 802.11n
- 802.11ac uses only the 5 GHz band

https://t.me/learningnets

# 802.1ax Protocol (Wi-Fi 6)

- The most recent release of the 802.11 protocols is 802.11ax, also called Wi-Fi 6

- It accommodates a growing number of devices and applications by increasing network efficiency to meet mobile and IoT devices

https://t.me/learningnets

# 802.1ax Protocol (Wi-Fi 6)

- 802.11ax delivers greater efficiency and security than previous versions of the protocol

- Its theoretical maximum transfer speed is about 10 Gbps—30 percent faster than Wi-Fi 5

- 802.11ax is backward compatible with the other protocol versions

https://t.me/learningnets

# Wi-Fi Standards Comparison

| IEEE Standard | 802.11a | 802.11b | 802.11g | 802.11n | 802.11ac | 802.11ax |
|---|---|---|---|---|---|---|
| Year Released | 1999 | 1999 | 2003 | 2009 | 2014 | 2019 |
| Frequency | 5Ghz | 2.4GHz | 2.4GHz | 2.4Ghz & 5GHz | 2.4Ghz & 5GHz | 2.4Ghz & 5GHz |
| Maximum Data Rate | 54Mbps | 11Mbps | 54Mbps | 600Mbps | 1.3Gbps | 10-12Gbps |

https://t.me/learningnets

# WLAN Architecture

## There are three different WLAN Architecture:

- Autonomous WLAN architecture
- Centralized WLAN architecture
- Distributed WLAN architecture

https://t.me/learningnets

# WLAN Architecture - Autonomous WLAN Architecture

- For many years, the conventional access point was a standalone WLAN device where control, data and management planes of operation existed and operated on the edge of the network architecture

- These APs are often referred to as fat APs or standalone APs

- However, the most common industry term for the traditional access point is autonomous AP

https://t.me/learningnets

# WLAN Architecture - Autonomous WLAN Architecture

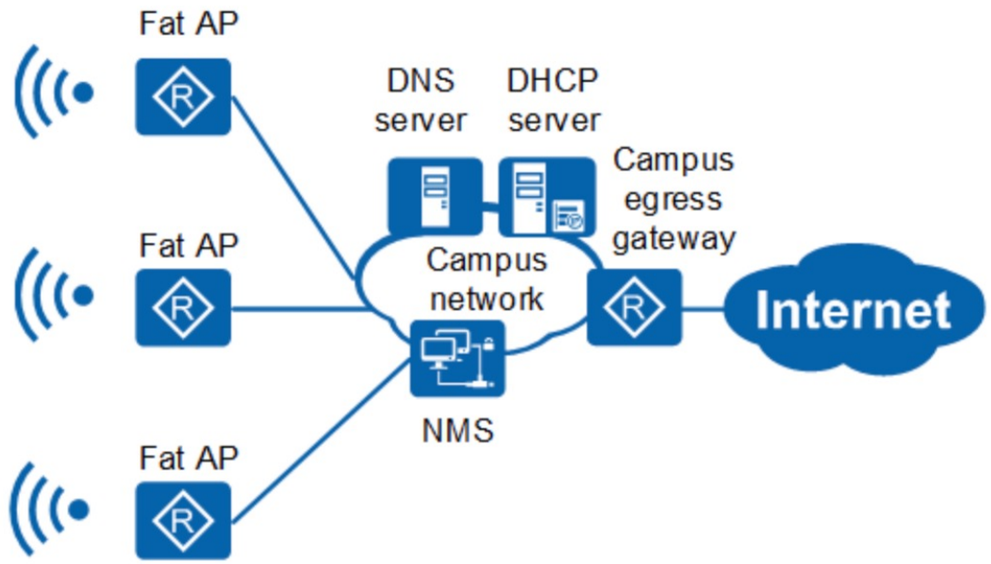
- All configuration settings exist in the autonomous access point itself, and therefore, the management plane resides individually in each autonomous AP

- All encryption and decryption mechanisms and MAC layer mechanisms also operate within the autonomous AP

https://t.me/learningnets

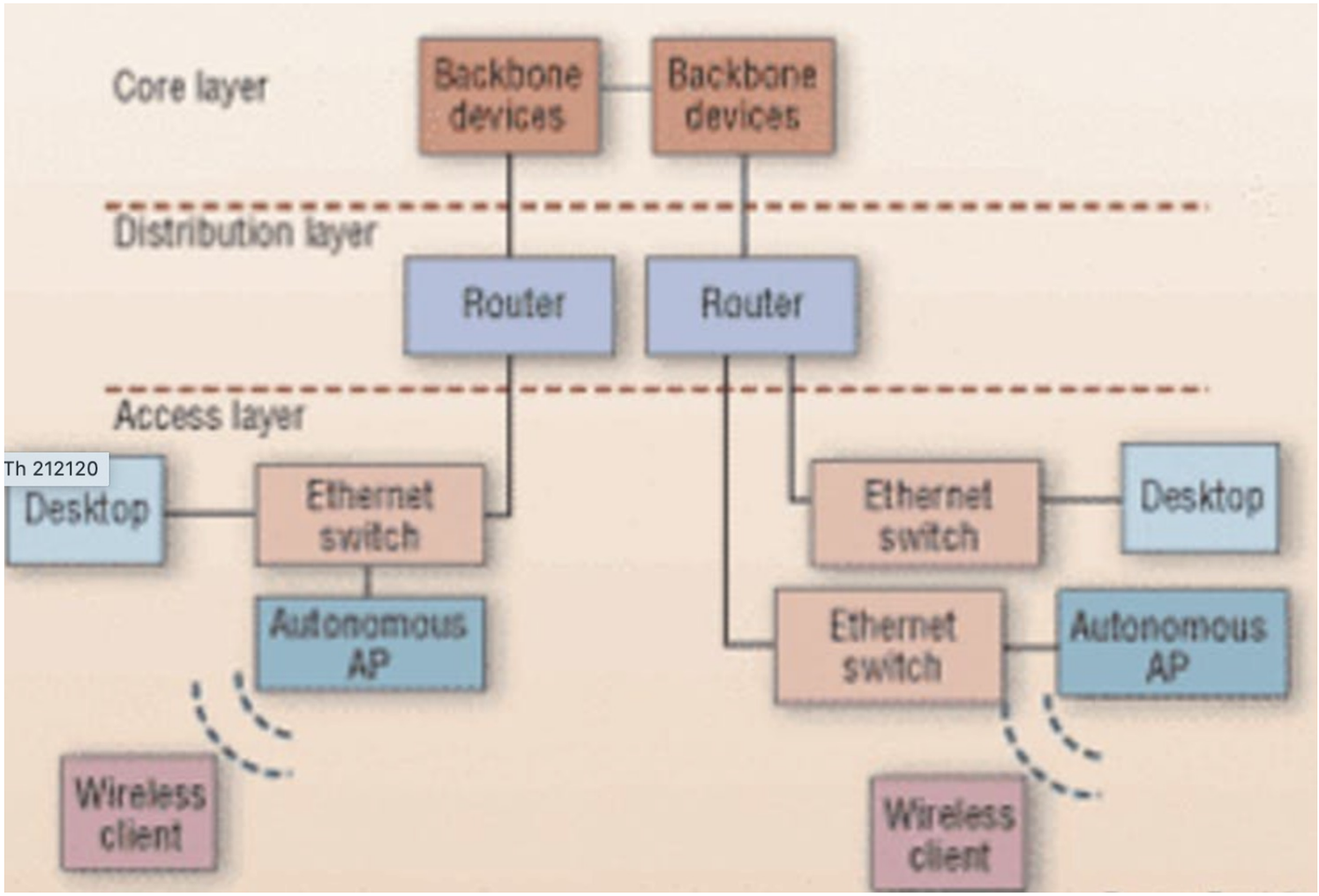
# WLAN Architecture - Autonomous WLAN Architecture

- The data plane also resides in each autonomous AP because all user traffic is forwarded locally by each individual access point




https://t.me/learningnets

# WLAN Architecture - Autonomous WLAN Architecture




https://t.me/learningnets

# WLAN Architecture - Autonomous WLAN Architecture

- Access points operate as layer 2 devices; however, they still need a layer 3 address for connectivity to an IP network

- The BVI is the management interface of an AP

- An autonomous access point typically encompasses both the 802.11 protocol stack and the 802.3 protocol stack

https://t.me/learningnets

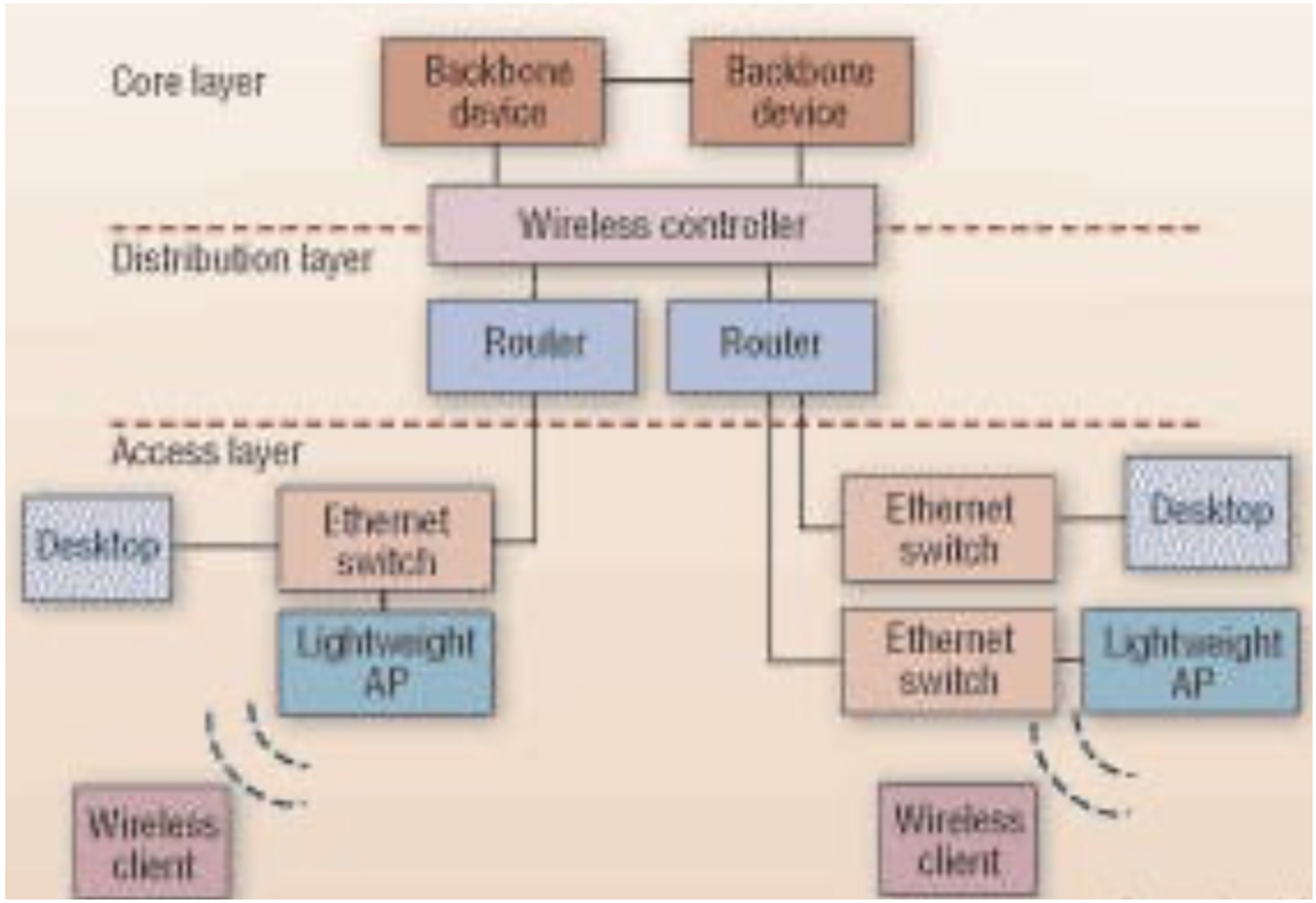
# WLAN Architecture - Centralized WLAN Architecture

- This model uses a central WLAN controller that resides in the core of the network

- In the centralized WLAN architecture, autonomous APs have been replaced with controller-based access points, also known as lightweight APs or thin APs

https://t.me/learningnets

# WLAN Architecture - Centralized WLAN Architecture




https://t.me/learningnets

# WLAN Architecture - Centralized WLAN Architecture

- Beginning in 2002, many WLAN vendors decided to move to a WLAN controller model where all three logical planes of operation would reside inside the controller

- In a centralized WLAN architecture, the three logical planes exist in a WLAN controller

https://t.me/learningnets

# WLAN Architecture - Centralized WLAN Architecture

- **Management Plane:** Access points are configured and managed from the WLAN controller

- **Control Plane:** Adaptive RF, load balancing, roaming handoffs, and other mechanisms exist in the WLAN controller

- **Data Plane:** The WLAN controller exists as a data distribution point for user traffic

https://t.me/learningnets

# WLAN Architecture - Centralized WLAN Architecture

- Access points tunnel all user traffic to a central controller

- The encryption and decryption capabilities might reside in the centralized WLAN controller or may still be handled by the controller-based APs, depending on the vendor

https://t.me/learningnets

# Centralized WLAN Architecture - WLAN Controller

**WLAN Controller has several functions:**

- AP Management
- WLAN Management
- User Management
- Device Monitoring
- VLANs
- Security Support
- Captive Portal
- Adaptive RF Spectrum Management
- Layer 3 Roaming Support

https://t.me/learningnets

# Centralized WLAN Architecture - WLAN Controller

**There are two types of data-forwarding methods when using WLAN controllers:**

- Centralized Data Forwarding
- Distributed Data Forwarding

https://t.me/learningnets

# Centralized WLAN Architecture - WLAN Controller

- **Centralized Data Forwarding:** Where all data is forwarded from the AP to the WLAN controller for processing

- It may be used in many cases, especially when the WLAN controller manages encryption and decryption or applies security and QoS policies

https://t.me/learningnets

# Centralized WLAN Architecture - WLAN Controller

- **Distributed Data Forwarding:** Where the AP performs data forwarding locally may be used in situations where it is advantageous to perform forwarding at the edge and to avoid a central location in the network for all data, which may require significant processor and memory capacity at the controller

https://t.me/learningnets

# Distributed WLAN Architecture

- A recent trend has been to move away from the centralized WLAN controller architecture toward a distributed architecture

- Some WLAN vendors, such as Aerohive Networks, have designed their entire WLAN system around a distributed architecture

https://t.me/learningnets

# Distributed WLAN Architecture

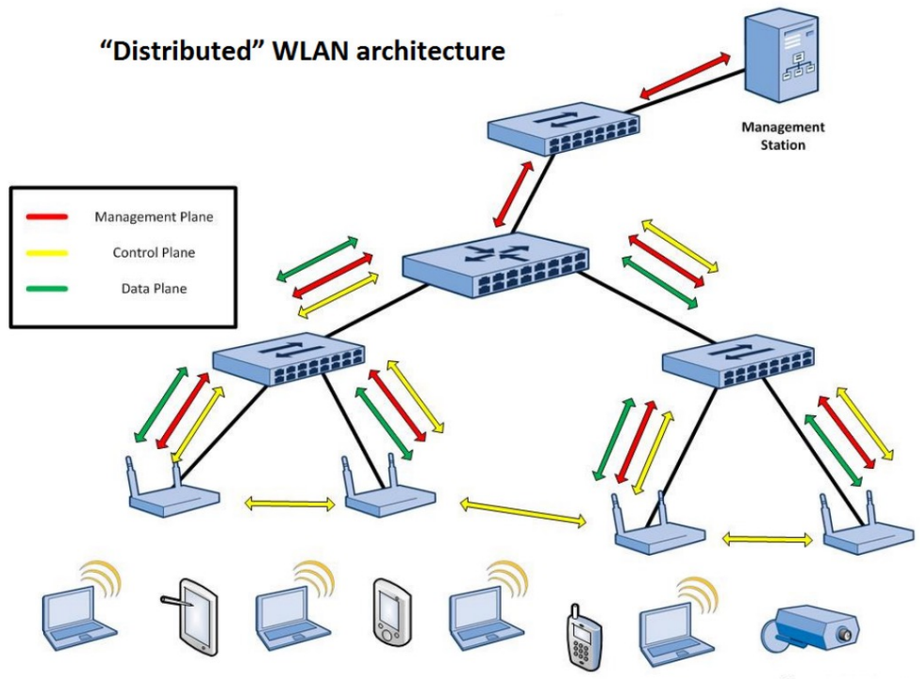
- Some of the WLAN controller vendors now also offer a distributed WLAN architecture solution, in addition to their controller-based solution

- In these systems, cooperative access points are used, and control plane mechanisms are enabled in the system with inter-AP communication via cooperative protocols

https://t.me/learningnets

# Distributed WLAN Architecture

- A distributed WLAN architecture combines multiple access points with a suite of cooperative protocols, without requiring a WLAN controller




Source: arubanetworks

https://t.me/learningnets

# Distributed WLAN Architecture

- Distributed WLAN architectures are modeled after traditional routing and switching design models, in that the network nodes provide independent distributed intelligence but work together as a system to cooperatively provide control mechanisms

https://t.me/learningnets

# Summary of WLAN Architectures

- Ultimately, the choice between an autonomous and a lightweight WLAN architecture is scale, OPEX and TCO

- Autonomous APs can become difficult to manage if deployed in large numbers

- There are no defined rules as to how large a deployment must be before it can be considered too large to manage.

https://t.me/learningnets

# Summary of WLAN Architectures

- A Centralize architecture could translate into lower operational costs because network policies, security settings, and radio-frequency power settings can be managed from a single device

- This architecture saves time versus managing individual APs, and it reduces human error

https://t.me/learningnets

# Summary of WLAN Architectures

- The amount of intelligence desired is a factor when choosing a WLAN Architecture

- While basic coverage can be obtained with a few autonomous APs, advanced self-healing and optimization can be seamlessly performed with a Centralized Architecture

- Additionally, wireless VoIP, asset tracking, and other intelligent applications are best served by a Centralized WLAN solution because of the centralized management

https://t.me/learningnets

# WLAN Design

**We will be covering below WLAN Design topics:**

- Coverage Design
- Roaming Design
- Channel Design
- Capacity Design

https://t.me/learningnets

# Coverage Design

- When designing a WLAN, probably the first thing that comes to mind will always be the coverage area or zone from which Wi-Fi clients can communicate

- The primary coverage goals for any WLAN are to provide high data rate connectivity for connected clients and to provide for seamless roaming

https://t.me/learningnets

# Coverage Design

- A common mistake is to design a WLAN based solely on an access point's capabilities

- The exact opposite should be considered during the design phase

https://t.me/learningnets

# Coverage Design

- A proper WLAN coverage design should be based on the perspective of the Wi-Fi clients

- Therefore, a quality received signal for the client is needed to provide high data rate connectivity

https://t.me/learningnets

# Coverage Design – Received Signal

- So what exactly is considered a quality received signal? As shown in the table, depending on the proximity between an AP and a Wi-Fi client, an 802.11 radio might receive an incoming signal anywhere between –30 dBm and the noise floor

| Quality | dBm | mW |
|---|---|---|
| Very Strong | –30 dBm | 1/1,000th of 1 milliwatt |
| Very Strong | –40 dBm | 1/10,000th of 1 milliwatt |
| Very Strong | –50 dBm | 1/100,000th of 1 milliwatt |
| Very Strong | –60 dBm | 1 millionth of 1 milliwatt |
| Strong | –70 dBm | 1 ten-millionth of 1 milliwatt |
| Fair | –80 dBm | 1 hundred-millionth of 1 milliwatt |
| Weak | –90 dBm | 1 billionth of 1 milliwatt |
| Very Weak | –95 dBm | Noise floor |

https://t.me/learningnets

# Coverage Design – Received Signal

- When designing for coverage, the normal recommended best practice is to provide for a –70 dBm or stronger received signal that is well above the noise floor

- In other words, a received signal of –70 dBm or higher is considered to be a quality received signal

https://t.me/learningnets

# Coverage Design – Received Signal

- A received signal of –70 dBm or higher usually guarantees that a client radio will use one of the highest data rates that the client is capable of!

https://t.me/learningnets

# Coverage Design - Signal-to-Noise Ratio (SNR)

- Another reason for planning for –70 dBm coverage is because the received signal of –70 dBm is usually well above the noise floor

- Signal-to-noise ratio (SNR) is an important value because if the background noise is too close to the received signal or the received signal level is too low, data can be corrupted

https://t.me/learningnets

# Coverage Design – Background Noise

- Before explaining SNR in detail, Possible Causes of Wireless Background Noise:

- Many sources can be a cause of background noise in a wireless network, for example, signals originating in an unlicensed frequency band may interfere with the intended frequencies of a network

https://t.me/learningnets

# Coverage Design – Background Noise

- Since the 2.4 GHz band used by 802.11b and 802.11g includes frequencies that are used by many different types of devices such as radios, microwave ovens, cordless telephones and Bluetooth devices etc. noise can be transmitted either accidentally or deliberately

https://t.me/learningnets

# Coverage Design – Background Noise

- Also, unwanted signals coming from distanced channels may act as noise for a given wireless network

- Furthermore, noise may be generated by wireless networks themselves if their signals are corrupted by the environment and become unusable

- A common cause of this is multi-path reception, where a signal reaches the receiver by different routes, i.e., if it is reflected off a concrete wall, and interferes with itself

https://t.me/learningnets

# Coverage Design - Signal-to-Noise Ratio (SNR)

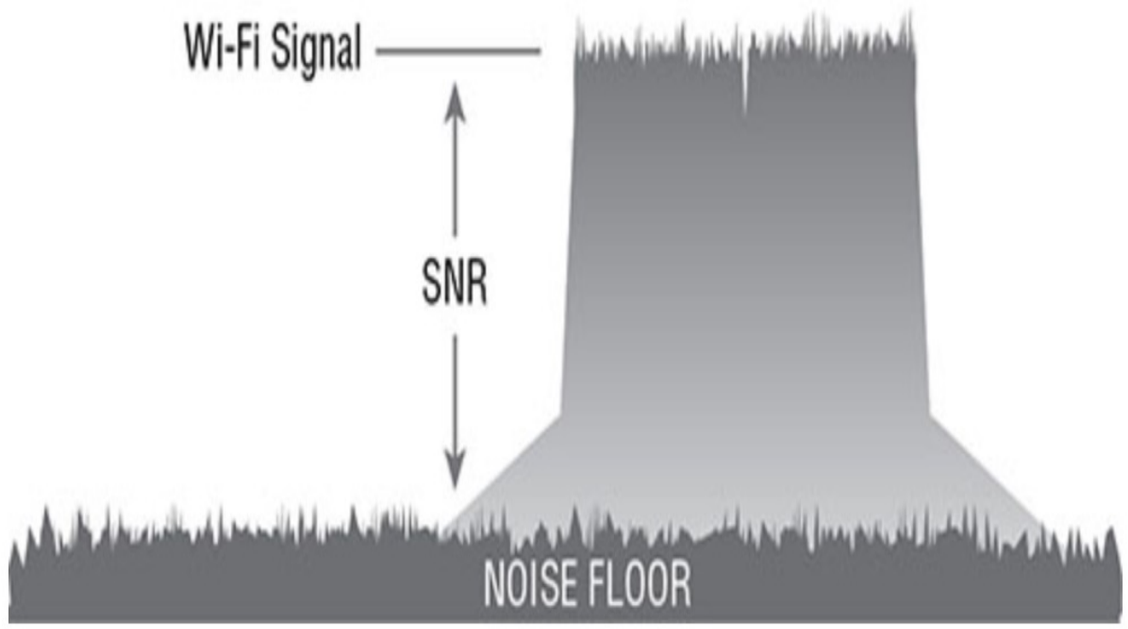
- The SNR is not actually a ratio; it is simply the difference in decibels between the received signal and the background noise (noise floor) measured in dBs

- If an 802.11 radio receives a signal of –70 dBm and the noise floor is measured at –95 dBm, the difference between the received signal and the background noise is 25 dB. So, the SNR is 25 dB

https://t.me/learningnets

# Coverage Design - Signal-to-Noise Ratio (SNR)

- In most instances, a received signal of –70 dBm will be 20 dB or higher above the noise floor. In most environments, a –70 dBm signal ensures high rate connectivity, and the 20 dB SNR ensures data integrity




https://t.me/learningnets

# Coverage Design - Signal-to-Noise Ratio (SNR)

- Data transmissions can become corrupted with a very low SNR

- If the amplitude of the noise floor is too close to the amplitude of the received signal, data corruption will occur and result in layer 2 retransmissions

- An SNR of 25 dB or greater is considered good signal quality, and an SNR of 10 dB or lower is considered poor signal quality

https://t.me/learningnets

# Coverage Design - Signal-to-Noise Ratio (SNR)

An SNR of below 10 dB will likely result in data corruption and retransmission rates as high as 50 percent

- To ensure that frames are not corrupted due to a low SNR, most Wireless vendors recommend a minimum SNR of 20 dB for data WLANs and a minimum SNR of 25 dB for WLANs that require voice-grade communications

https://t.me/learningnets

# Coverage Design - VoWifi

- VoWiFi communications are more susceptible to layer 2 retransmissions than other types of application traffic

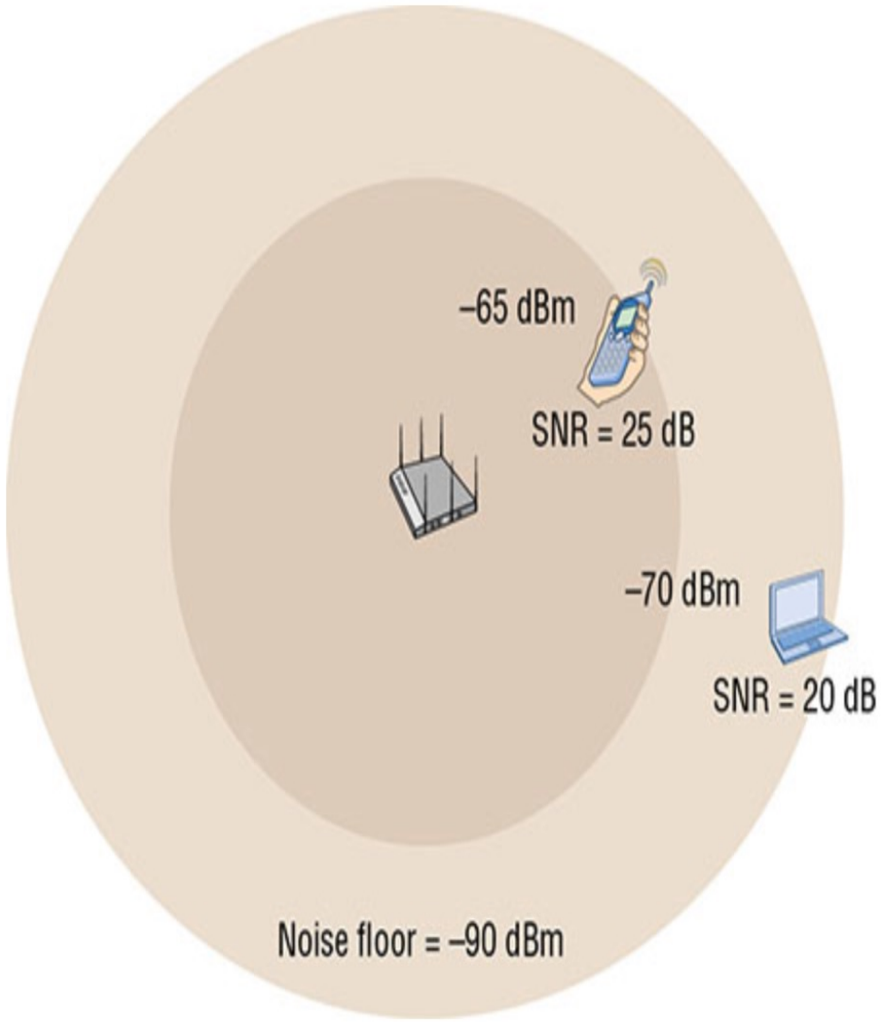
- Therefore, when you are designing for voice-grade WLANs, a –65 dBm or stronger signal is recommended so that the received signal is higher above the noise floor

https://t.me/learningnets

# Coverage Design - VoWifi

- As shown in the right figure, even if the noise floor were a very high –90 dBm, the SNR of a –65 dBm received signal for a VoWiFi client would still be 25 dB

- Always check the recommendations of the manufacturer of the VoWiFi client




https://t.me/learningnets

# Coverage Design - VoWifi

- One VoWiFi vendor may state that a –67 dBm signal is sufficient, whereas another vendor may suggest an SNR as high as 28 dB

- When you are designing for voice, SNR is the most important RF metric

https://t.me/learningnets

# Coverage Design – Dynamic Rate Switching

- Will a client device be able to communicate with an AP if the signal drops below –70 dBm?

- The answer is yes, because most client devices can still decode an 802.11 preamble from received signals that are as low as only 4 dB above the noise floor

- As mobile client radios move away from an access point, they will shift down to lower-bandwidth capabilities by using a process known as dynamic rate switching (DRS)

https://t.me/learningnets

# Coverage Design – Dynamic Rate Switching

- Data rate transmissions between the access point and the client stations will shift down or up, depending on the quality of the signal between the two radios

https://t.me/learningnets

# Coverage Design – Dynamic Rate Switching

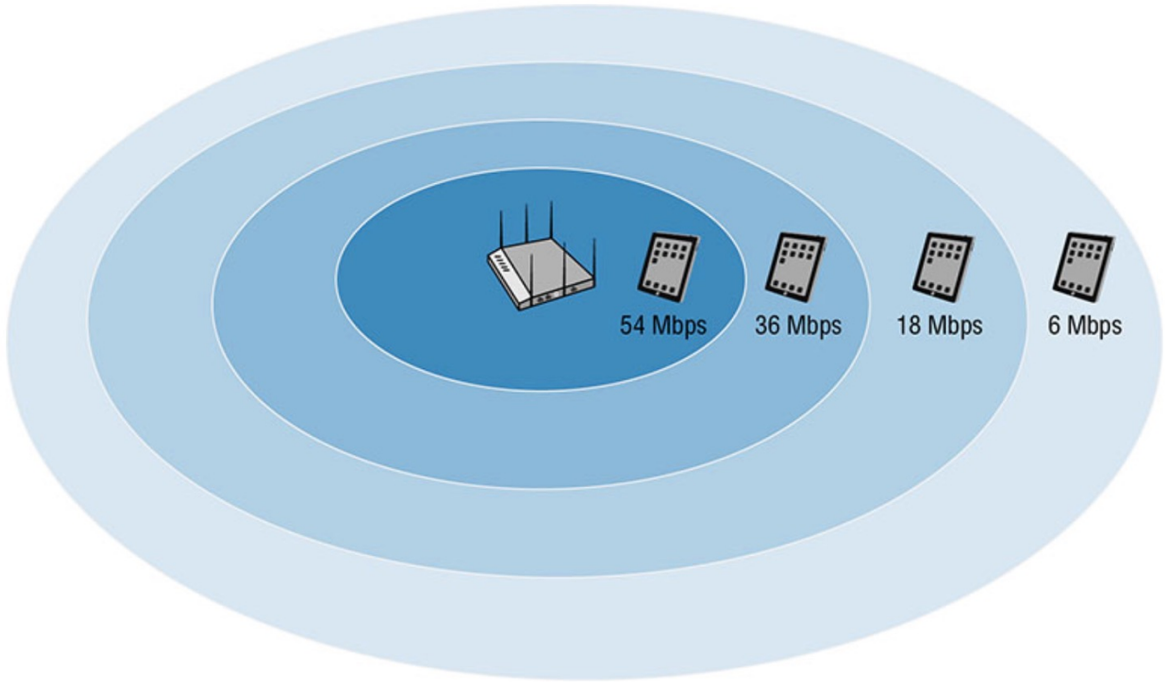
- There is a direct correlation between signal quality and distance from the AP

- As mobile client stations move farther away from an access point, both the AP and the client will shift down to lower rates that require a less complex modulation and coding scheme (MCS)

https://t.me/learningnets

# Coverage Design – Dynamic Rate Switching

- In the right figure, an 802.11a/g client might connect at 54 Mbps when receiving a –70 dBm signal, but it might shift to transmitting with a lower data rate of 6 Mbps if the signal is much weaker




https://t.me/learningnets

# Coverage Design – Dynamic Rate Switching

- Dynamic rate switching (DRS) is also referred to as dynamic rate shifting, dynamic rate selection, adaptive rate selection, and automatic rate selection

- All these terms refer to a method of speed fallback on a Wi-Fi radio receiver (Rx) as the incoming signal strength and quality from the transmitting Wi-Fi radio decreases

https://t.me/learningnets

# Coverage Design – Dynamic Rate Switching

- The objective of DRS is upshifting and downshifting for rate optimization and improved performance

- From the client's perspective, the lower data rates will provide larger concentric zones of coverage than the higher data rates

https://t.me/learningnets

# Coverage Design – Transmit Power

- A big factor that will affect both WLAN coverage and roaming is the transmit power of the access points

- Although most indoor APs may have full transmit power settings as high as 100 mW, they should rarely be deployed at full power

- This extends the effective range of the access point; however, designing WLANs strictly for range is an outdated concept

https://t.me/learningnets

# Coverage Design – Transmit Power

- WLAN capacity design and reduction of airtime consumption are important design considerations as well!

- APs at maximum transmit power will result in oversized coverage and not meet your capacity needs

https://t.me/learningnets

# Coverage Design – Transmit Power

- Access points deployed at full transmit power in an indoor environment will also increase the chance of co-channel interference, which can result in unnecessary medium contention overhead

- APs at full power also increase the chance of sticky clients which negatively impact roaming

https://t.me/learningnets

# Coverage Design – Transmit Power

- Typical indoor WLAN deployments are designed with the APs set at about one-fourth to one-third maximum transmit power

- Higher user density environments may require that the AP transmit power be set at the lowest setting of 1 mW

https://t.me/learningnets

# Coverage Design – Transmit Power

- Another consideration is the transmit power of the **CLIENTS**!
- One heavily debated topic is the concept of a balance power link between an AP and a client
- In simpler words, the transmit power settings between an AP and a client are the same
- Very often WLAN clients transmit at higher power levels than indoor access points

https://t.me/learningnets

# Coverage Design – Transmit Power

- The transmit power of many indoor APs may be 10 mW or less due to high-density design needs

- However, most clients, such as smartphones and tablets, may transmit at fixed power of 15 mW or 20 mW

- Because clients often transmit at a higher power than the APs and because clients are mobile, co-channel interference (CCI) is often caused by a power mismatch

https://t.me/learningnets

# Roaming Design

- Roaming is the method by which client stations move between RF coverage cells in a seamless manner

- Client stations switch communications through different access points

- Seamless communications for client stations moving between the coverage zones is vital for uninterrupted mobility

https://t.me/learningnets

# Roaming Design

- Roaming is one of the most common issues you will need to troubleshoot in WLAN

- Roaming problems are usually caused by poor network design

https://t.me/learningnets

# Roaming Design

- Client stations, not the access point, decide whether or not the client roams between Access Points

- Some vendors may involve the access point or WLAN controller in the roaming decision, but ultimately the client station initiates the roaming process with a reassociation request frame

https://t.me/learningnets

# Roaming Design

- The method by which a client station decides to roam is a set of proprietary rules determined by the manufacturer of the 802.11 radio, usually defined by a roaming trigger threshold

https://t.me/learningnets

# Roaming Design

- Roaming thresholds usually involve signal strength, SNR, and bit-error rate

- As the client station communicates on the network, it continues to look for other access points via probing and listening on other channels and will hear received signals from other APs

https://t.me/learningnets

# Roaming Design

- The most important variable will always be received signal strength

- As the received signal from the original AP gets weaker and a station hears a stronger signal from another known access point, the station will initiate the roaming process

https://t.me/learningnets

# Roaming Design

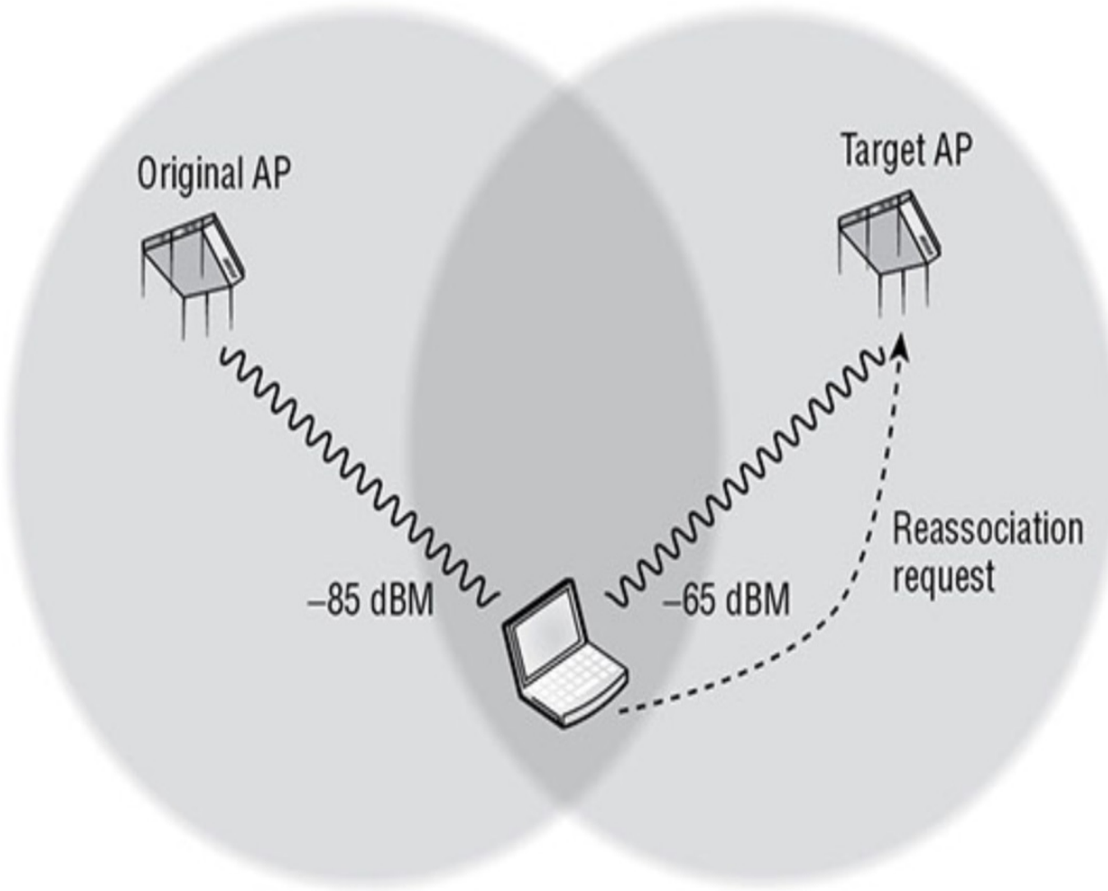
- However, other metrics, such as SNR, error rates, and retransmissions, may also have a part in what triggers a client to roam

- SNR is a metric used by some WLAN clients to trigger roaming events as well as dynamic rate switching

https://t.me/learningnets

# Roaming Design

- As shown in the right figure, as the client station moves away from the original access point with which it is associated and the signal drops below a predetermined threshold, the client station will attempt to connect to a new target access point that has a stronger signal




https://t.me/learningnets

# Roaming Design – Layer 3 Roaming

- One major consideration when designing a WLAN is what happens when client stations roam across layer 3 boundaries

- Wi-Fi operates at layer 2, and roaming is essentially a layer 2 process

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming

- The roaming is seamless at layer 2, but user VLANs are tied to different subnets on either side of the router

- As a result, the client station will lose layer 3 connectivity and must acquire a new IP address

https://t.me/learningnets

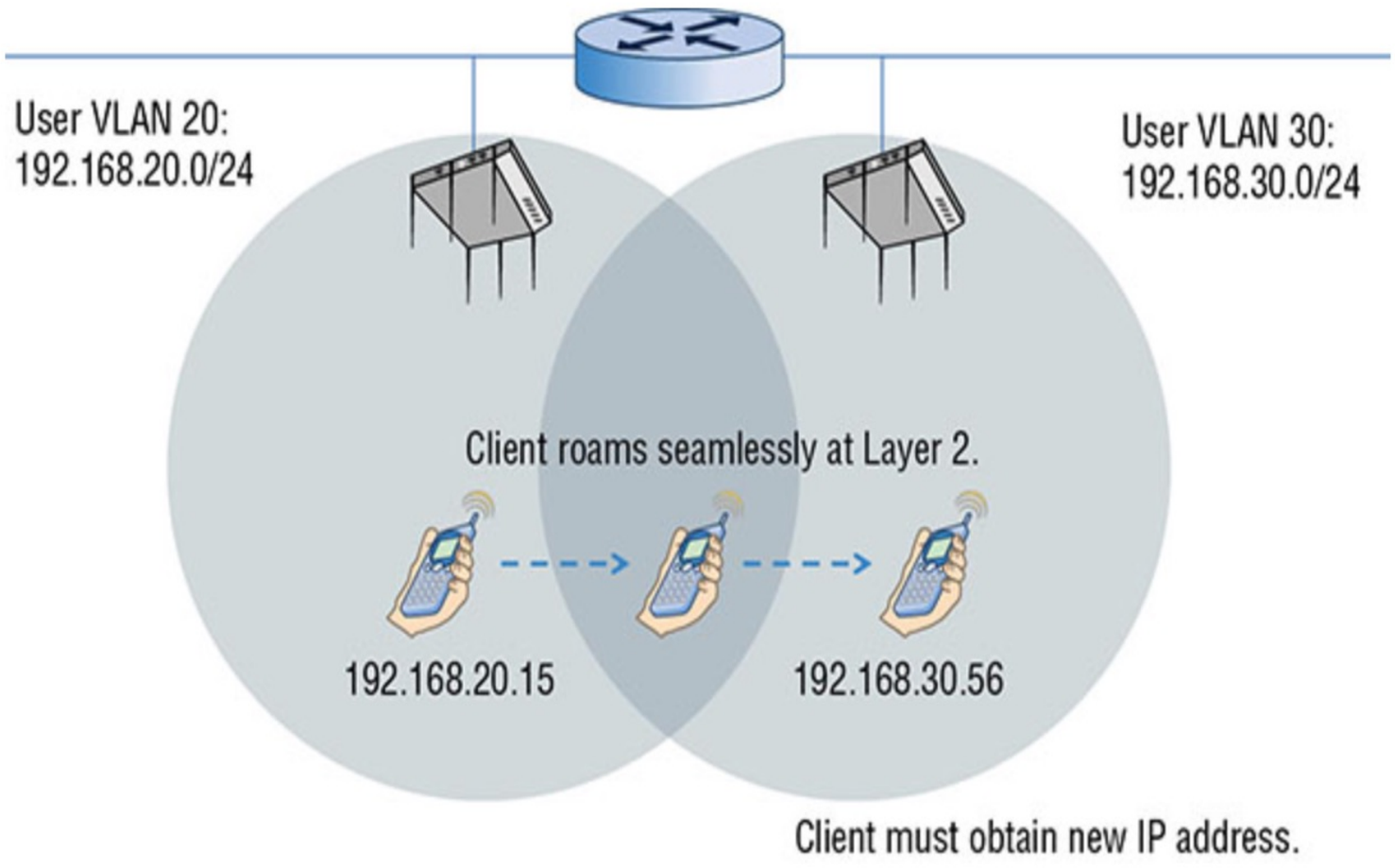
# Roaming Design – Layer 3 Roaming

- Any connection-oriented applications that are running when the client re-establishes layer 3 connectivity will have to be restarted

- For example, a VoIP phone conversation would disconnect in this scenario, and the call would have to be reestablished

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming




https://t.me/learningnets

# Roaming Design – Layer 3 Roaming

- Because 802.11 wireless networks are usually integrated into pre-existing wired topologies, crossing layer 3 boundaries is often a necessity, especially in large deployments

- The only way to maintain upper-layer communications when crossing layer 3 subnets is to provide a layer 3 roaming solution that is based on the Mobile IP standard

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- Mobile IP is an Internet Engineering Task Force (IETF) standard protocol that allows mobile device users to move from one layer 3 network to another while maintaining their original IP address

- Mobile IP is defined in RFC 5944

https://t.me/learningnets

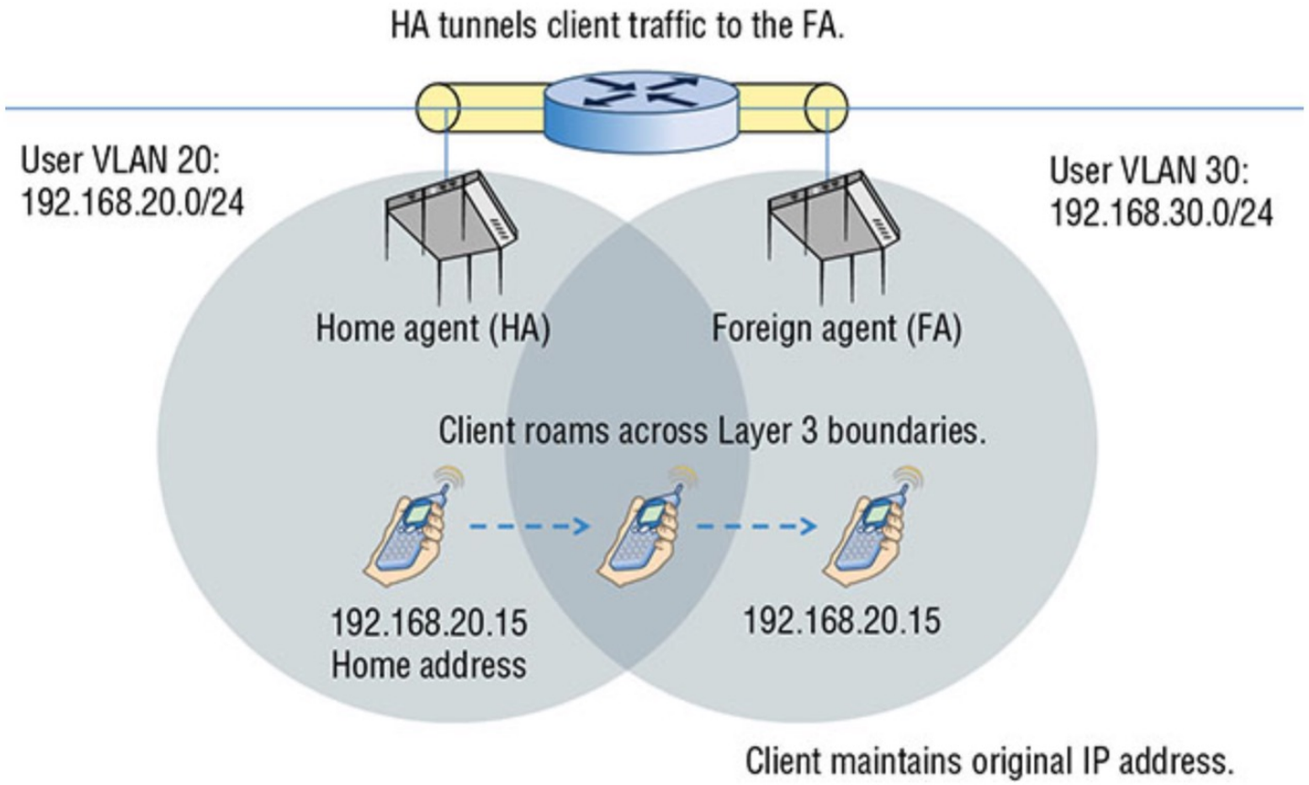
# Roaming Design – Layer 3 Roaming – Mobile IP

- Layer 3 roaming solutions based on Mobile IP use some type of tunneling method and IP header encapsulation to allow packets to traverse between separate layer 3 domains, with the goal of maintaining upper-layer communications

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- Most WLAN vendors now support some form of layer 3 roaming solution, as shown in the right figure




https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- A mobile client receives an IP address, also known as a home address on a home network

- The mobile client must register its home address with a device called a home agent (HA)

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- Client's original associated access point serves as the home agent

- The home agent is a single point of contact for a client when it roams across layer 3 boundaries

- The HA shares client MAC/IP database information in a table, called a home agent table (HAT) with another device, called the foreign agent (FA)

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- In this example, the foreign agent is another access point that handles all Mobile IP communications with the home agent on behalf of the client

- The foreign agent's IP address is known as the care-of address

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- When the client roams across layer 3 boundaries, the client is roaming to a foreign network where the FA resides
- The FA uses the HAT tables to locate the HA of the mobile client station
- The FA contacts the HA and sets up a Mobile IP tunnel

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- Any traffic that is sent to the client's home address is intercepted by the HA and sent through the Mobile IP tunnel to the FA

- The FA then delivers the tunneled traffic to the client, and the client is able to maintain connectivity using the original home address

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- In our example, the Mobile IP tunnel is between two APs on opposite sides of a router

- If the user VLANs exist at the edge of the network, tunneling of user traffic occurs between access points that assume the roles of HA and FA

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming – Mobile IP

- In a multiple WLAN controller environment, an IP tunnel is created between controllers that are deployed in different routed boundaries with different user VLANs

- One controller functions as the home agent, while another controller functions as the foreign agent

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming

- Although maintaining upper-layer connectivity is possible with these layer 3 roaming solutions, increased latency is sometimes an issue

- Additionally, layer 3 roaming may not be a requirement for your network

https://t.me/learningnets

# Roaming Design – Layer 3 Roaming

- Less complex infrastructure often uses a simpler layer 2 design

- Larger enterprise networks often have multiple user and management VLANs linked to multiple subnets; therefore, a layer 3 roaming solution will be required

https://t.me/learningnets

# Channel Design

- Another key component of WLAN design is the selection of the proper channels to be used among multiple APs in the same location

- A proper channel pattern or channel reuse design is needed to guarantee seamless roaming as well as to prevent two types of interference that are a result of improper channel design

https://t.me/learningnets

# Channel Design - Adjacent Channel Interference

- Adjacent channel interference (ACI) is a degradation of performance resulting from overlapping frequency space that occurs due to an improper channel reuse design

- In the WLAN industry, an adjacent channel is considered to be the next or previous numbered channel

- For example, channel 3 is adjacent to channel 2

https://t.me/learningnets

# Channel Design - Adjacent Channel Interference

- When designing a wireless LAN, you need overlapping coverage cells in order to provide for roaming

- However, the overlapping cells should not have overlapping frequencies, and in the United States only channels 1, 6, and 11 should be used in the 2.4 GHz ISM band to get the most available, non-overlapping channels

https://t.me/learningnets

# Channel Design - Adjacent Channel Interference

- Overlapping coverage cells with overlapping frequencies result what is known as adjacent channel interference

- If overlapping coverage cells also have frequency overlap from adjacent channels, the transmitted frames will become corrupted, the receivers will not send ACKs, and layer 2 retransmissions will significantly increase

https://t.me/learningnets

# Channel Design - 2.4 GHz Channel Reuse

- To avoid adjacent channel interference, a channel reuse design is necessary

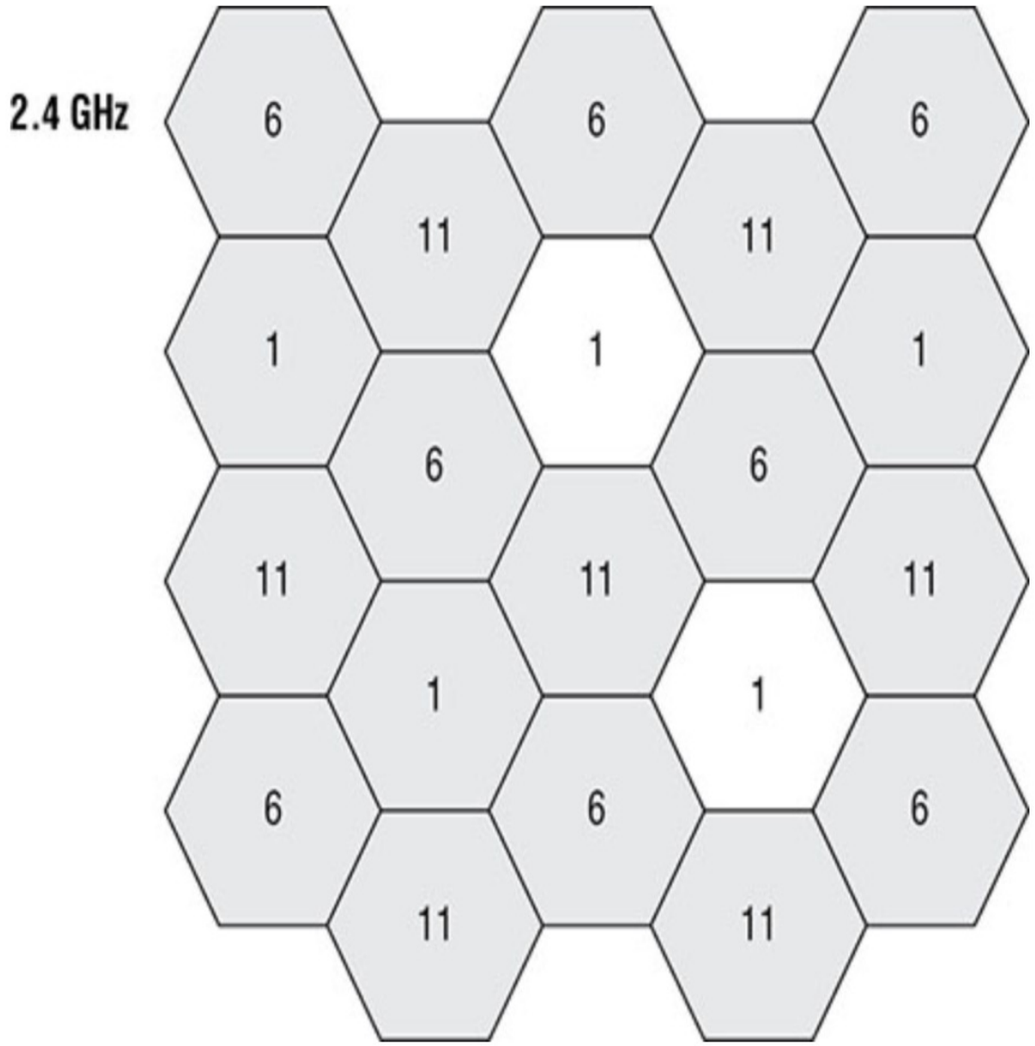
- Once again, overlapping RF coverage cells are needed for roaming, but overlapping frequencies must be avoided

- The only three channels that meet these criteria in the 2.4 GHz ISM band are channels 1, 6, and 11 in the United States

https://t.me/learningnets

# Channel Design - 2.4 GHz Channel Reuse

- APs in the 2.4 GHz band, therefore, should always be placed in a channel reuse pattern similar to the one pictured in the right figure




- Any WLAN channel reuse pattern that uses three or more channels is sometimes referred to as a multiple-channel architecture (MCA)

https://t.me/learningnets

# Channel Design - Co-Channel Interference

- Another of the most common mistakes many businesses make when first deploying a WLAN is to configure multiple access points all on the same channel

- If all the APs are on the same channel, unnecessary medium contention overhead occurs

- CSMA/CA dictates half-duplex communications, and only one radio can transmit on the same channel at any given time

https://t.me/learningnets

# Channel Design - Co-Channel Interference

- If an AP on channel 1 is transmitting, all nearby access points and clients on the same channel within hearing range will defer transmissions

- The result is that throughput is adversely affected:
  - Nearby APs and clients have to wait much longer to transmit because they have to take their turn
  - The unnecessary medium contention overhead that occurs because all the APs are on the same channel is called co-channel interference (CCI)

https://t.me/learningnets

# Channel Design - Co-Channel Interference

- Co-channel interference is almost impossible to prevent in the 2.4 GHz band because only three channels are available for a reuse pattern

- Does an RF signal just stop at the edge of a coverage cell designed for –70 dBm coverage?

  The answer is no, the RF signal continues to propagate, and the signal can be heard by other 802.11 radios at a great distance

https://t.me/learningnets

# Channel Design - Co-Channel Interference

- An 802.11 radio will defer transmissions if it hears the PHY layer preamble transmissions of any other 802.11 radio at a signal detect (SD) threshold just four decibels (dB) or greater above the noise floor

- Any radio that hears another radio on the same channel will defer, which results in medium contention overhead and delay

https://t.me/learningnets

# Channel Design - 5 GHz Channel Reuse

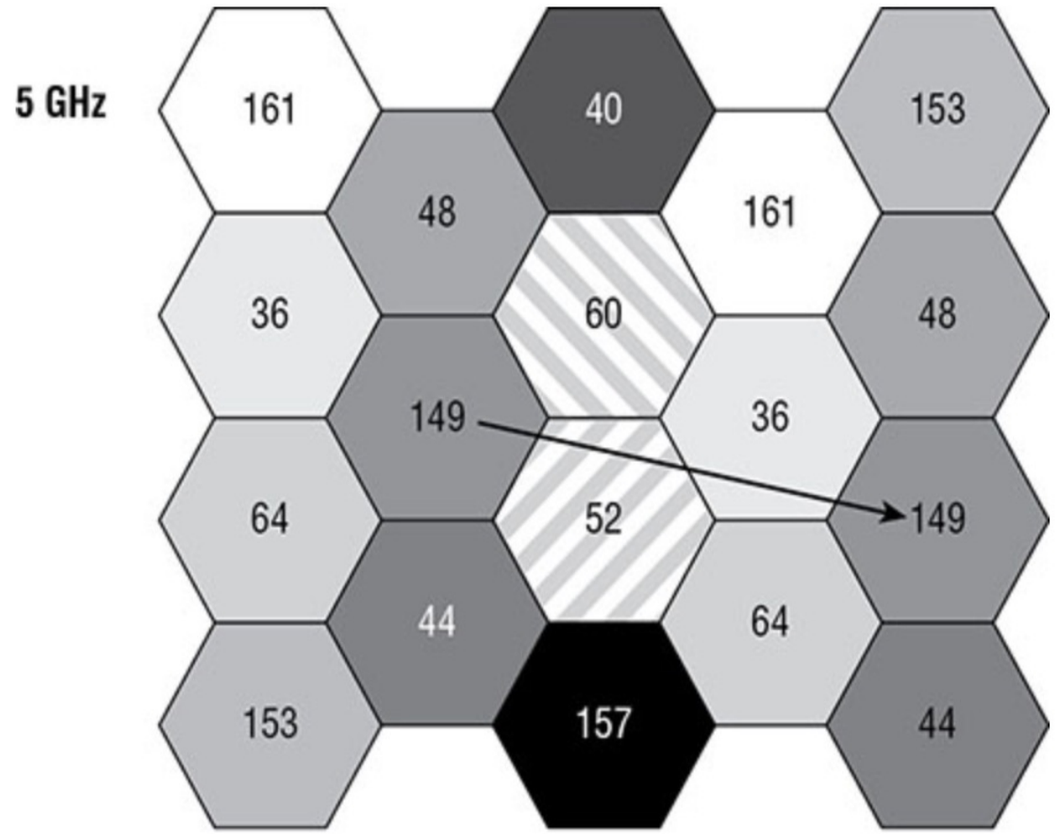
- Channel reuse patterns should also be used in the 5 GHz frequency bands

- If all the 5 GHz channels are legally available for transmissions, a total of 25 channels may be available for a channel reuse pattern at 5 GHz

https://t.me/learningnets

# Channel Design - 5 GHz Channel Reuse

- Depending on the region, and other considerations, 8 channels, 12 channels, 17 channels, 22 channels, or other combinations may be used for 5 GHz channel reuse patterns




https://t.me/learningnets

# Channel Design - 5 GHz Channel Reuse

- Whenever possible, use as many channels as possible in 5 GHz to reduce CCI

- The more channels that are used, the greater the chance that CCI can be prevented, including co-channel interference that originates from client devices

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- 802.11n technology introduced the capability of bonding two 20 MHz channels to create a larger 40 MHz channel

- Channel bonding effectively doubles the frequency bandwidth, meaning double the data rates that can be available to 802.11n/ac radios

- Many WLAN access point vendors require that channel bonding be manually enabled because there is the potential for channel bonding to negatively impact the performance of the WLAN

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- One of the advantages of using 5 GHz instead of 2.4 GHz is that there are many more 5 GHz 20 MHz channels that can be used in a reuse pattern

- Only three 20 MHz channels can be used in 2.4 GHz. The problem with using only three 20 MHz channels is that there will always be some amount of co-channel interference even though these channels are non-overlapping

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- Therefore, a certain amount of medium contention overhead always exists at 2.4 GHz simply because there are not enough channels and frequency space

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- Medium contention overhead due to APs on the same 20 MHz channel can be almost completely avoided in 5 GHz because there are more channels

- A 5 GHz channel reuse plan of eight or more 20 MHz channels will greatly decrease co-channel interference and medium contention overhead

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- If only eight 20 MHz channels are being used, then a four channel 40 MHz channel reuse pattern exists in 5GHz

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- Although the bandwidth is doubled for the 802.11n/ac radios, there will be an increase of medium contention overhead because there are only four 40 MHz channels, and access points and clients on the same 40 MHz channel will likely hear each other

- The medium contention overhead may have a negative impact and decreases any gains in performance that the extra bandwidth might provide

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- Another problem with channel bonding is that it usually will result in a higher noise floor of about 3 dB

- If the noise floor is 3 dB higher, then the SNR is 3 dB lower, which means that the radios may shift down to lower MCS rates and therefore lower modulation data rates

- In many cases, this offsets some of the bandwidth gains that the 40 MHz frequency space provides

https://t.me/learningnets

# Channel Design - 40 MHz Channel Design

- So, should you use channel bonding or not? If four or fewer 40 MHz channels are available, you might not want to turn on channel bonding, especially if the 5 GHz radios are transmitting at a higher power level

- If the majority of the WLAN clients do not support channel bonding, don’t use channel bonding

https://t.me/learningnets

# Channel Design - Static Channels and Transmit Power vs. Adaptive RF (RRM)

- Probably the most debated topic when it comes to WLAN design is whether to use static channel and power settings for APs versus using adaptive channel and power settings

- Radio resource management (RRM) is an industry standard term used to describe the automatic and adaptive power and channel configuration of access points

https://t.me/learningnets

# Channel Design - Static Channels and Transmit Power vs. Adaptive RF (RRM)

- WLAN vendor APs can dynamically change their configuration based on accumulated RF information gathered from the access point radios

- Based on the accumulated RF information, the access points adjust their power and channel settings, adaptively changing the RF coverage cells

https://t.me/learningnets

# Channel Design - Static Channels and Transmit Power vs. Adaptive RF (RRM)

- Radio resource management is also referred to as adaptive RF

- When implemented, RRM provides automatic cell sizing and automatic monitoring and optimization of the RF environment, which can best be described as a self-organizing wireless LAN

https://t.me/learningnets

# Channel Design - Static Channels and Transmit Power vs. Adaptive RF (RRM)

- Adaptive RF capabilities are turned on by default on most WLAN vendor AP

- The algorithms for RRM constantly improve year after year.

- The majority of commercial WLAN customers use RRM because it is easy to deploy

https://t.me/learningnets

# Channel Design - Static Channels and Transmit Power vs. Adaptive RF (RRM)

- RRM is usually the preferred method in enterprise deployments with thousands of APs

- However, careful consideration should be given to using static channel and power settings in complex RF environments

https://t.me/learningnets

# Channel Design – Single Channel Architecture

- In Single Channel Architecture, the client stations see transmissions on only a single channel with one SSID (logical WLAN identifier) and one BSSID (layer 2 identifier)

- From the perspective of the client station, only one access point exists

https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- In this type of WLAN architecture, all access points in the network can be deployed on one channel in 2.4 GHz or 5 GHz frequency bands

https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- Uplink and downlink transmissions are coordinated by a WLAN controller on a single 802.11 channel in such a manner that the effects of co-channel interference are minimized

https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- Client stations believe they are connected to only a single access point, although they may be roaming across multiple physical Access Points

https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- As earlier said, clients make the roaming decisions

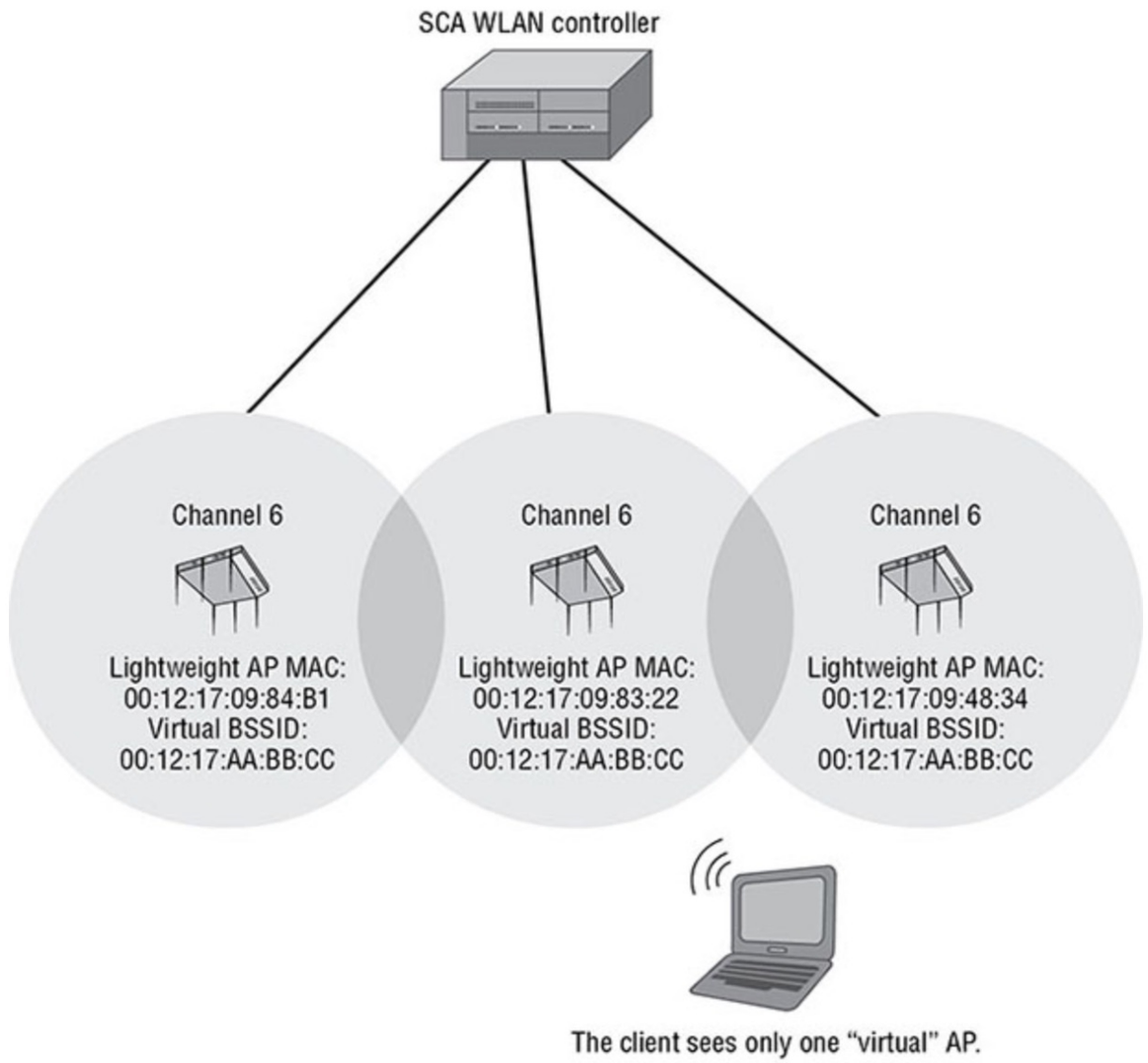
- In a single-channel architecture (SCA) system, the clients think they are associated to only one AP, so they never initiate a layer 2 roaming exchange. All the roaming handoffs are handled by a central WLAN controller

https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- The main advantage is that clients experience zero handoff time, and the latency issues associated with roaming times are resolved

- The virtual AP used by SCA solutions is potentially an excellent marriage for VoWiFi phones




https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- We know that client stations make the roaming decision in an MCA environment

- However, client stations do not know that they roam in an SCA environment

https://t.me/learningnets

# Channel Design - Single-Channel Architecture

- The clients must still be mobile and transfer layer 2 communications between physical access points

- All the client-roaming mechanisms are now handled back on the WLAN controller, and client-side roaming decisions have been eliminated in Single Channel Architecture

https://t.me/learningnets

# Capacity Design

- When a wireless network is designed, two concepts that typically compete: **Capacity and Range**

https://t.me/learningnets

# Capacity Design

- In the early days of wireless networks, it was common to install an access point with the power set to the maximum level to provide the largest coverage area possible

https://t.me/learningnets

# Capacity Design

- This was typically acceptable because there were very few wireless devices

- Also, access points were very expensive, so companies tried to provide the most coverage while using the fewest access points

https://t.me/learningnets

# Capacity Design

- Most WLANs are designed with a primary focus on client capacity needs
- This does not mean that coverage design is now ignored
- We still plan for a –70 dBm or better received signal, high SNR, seamless roaming, and a proper channel reuse pattern

https://t.me/learningnets

# Capacity Design

- Important to know that, how you design for coverage will also impact capacity needs

- As it is said before, APs configured for full transmit power are no longer ideal

https://t.me/learningnets

# Capacity Design

- Adjusting the AP transmit power to limit the effective coverage area is known as cell sizing and is one of the most common methods of meeting client capacity needs

- Typical indoor WLAN deployments are designed with the APs set at about one-fourth to one-third transmit power

https://t.me/learningnets

# Capacity Design

- Higher user and client density environments may require that the AP transmit power be set at the lowest setting of 1 mW

- In other words, more APs are needed to meet capacity needs, and therefore AP transmit power will need to be lowered

https://t.me/learningnets

# Capacity Design

- Limiting the transmit power of APs also helps reduce CCI caused by APs, which has a direct impact on performance

https://t.me/learningnets

# Capacity Design  - High Density

- The terms high density (HD) and very high density (VHD) are often used when discussing capacity design and planning for a WLAN

https://t.me/learningnets

# Capacity Design - High Density

- Different WLAN engineers have different opinions as to what constitutes a high-density WLAN, however, due to the huge number of client devices, most WLANs should be considered high density by default

https://t.me/learningnets

# Capacity Design - High Density

- The average person may want to connect to an enterprise WLAN with as many as three or four Wi-Fi devices

- Obviously, the density of client devices also depends on the number of users

https://t.me/learningnets

# Capacity Design - High Density

- Most high-density environments consist of multiple areas where roaming is also a top priority

- APs are deployed in many different rooms with walls that will often contribute to different levels of attenuation

https://t.me/learningnets

# Capacity Design  - Very High Density

- Any WLAN environment that has a tremendous amount of people in a single open area is often referred to as a very high-density (VHD) WLAN

- Prime examples include auditoriums, gymnasiums, cafeterias, etc.

https://t.me/learningnets

# Capacity Design  - Very High Density

- Most VHD environments do not have walls that comes with attenuation
- All the APs likely hear each other within the open space
- Design for a very high-density WLAN is quite complex and different from standard high-density environments with walls

https://t.me/learningnets

# Capacity Design - Ultra High Density

- An ultra high-density WLAN is defined as an environment with tens of thousands of users and devices all within the same space

- The best examples of an ultra high-density WLAN are stadiums and sports arenas

https://t.me/learningnets

# Capacity Design - How many clients can connect to an AP?

- There are simply too many variables to always give the same answer for any WLAN vendor's AP

- The default settings of an enterprise WLAN radio might allow as many as 100–250 client connections

https://t.me/learningnets

# Capacity Design  - How many clients can connect to an AP?

- Since most enterprise APs are dual-frequency with both a 2.4 GHz and 5 GHz radio, theoretically 200–500 clients could associate to the radios of a single AP

https://t.me/learningnets

# Capacity Design  - How many clients can connect to an AP?

- Although more than a hundred devices might be able to connect to an AP radio, these numbers are not realistic for active devices due to the nature of the half-duplex shared medium

https://t.me/learningnets

# Capacity Design - How many clients can connect to an AP?

- The performance needs of this many client devices will not be met and the user experience will be miserable

- The perception will be that the Wi-Fi is “slow” If the access point is using 802.11n/ac radios with 20 MHz channels, a good rule of thumb is that each radio could support 35–50 active devices for average use

https://t.me/learningnets

# Capacity Design – Common Design Questions

- Below questions should be asked while designing WLAN for capacity

1. What type of applications will be used on the WLAN?
2. How many users and devices are expected?
3. What types of client devices are connecting to the WLAN?

https://t.me/learningnets

# Capacity Design - What type of applications will be used on the WLAN?

- As previously stated, 35–50 active Wi-Fi devices per radio, communicating through a dual-frequency 802.11n/ac access point, with average application use, such as web browsing and email, is realistic

| Application | Required Throughput |
|---|---|
| Email/web browsing | 500 Kbps to 1 Mbps |
| Printing | 1 Mbps |
| SD video streaming | 1 Mbps to 1.5 Mbps |
| HD video streaming | 2 Mbps to 5 Mbps |

https://t.me/learningnets

# Capacity Design - What type of applications will be used on the WLAN?

- However, bandwidth-intensive applications, such as high-definition video streaming, will have an impact

- Different applications require different amounts of throughput

| Application | Required Throughput |
|---|---|
| Email/web browsing | 500 Kbps to 1 Mbps |
| Printing | 1 Mbps |
| SD video streaming | 1 Mbps to 1.5 Mbps |
| HD video streaming | 2 Mbps to 5 Mbps |

https://t.me/learningnets

# Capacity Design - How many users and devices are expected?

**Three important questions need to be asked with regard to users**

1. First, how many users currently need wireless access and how many Wi-Fi devices will they be using?

2. Second, how many users and devices may need wireless access in the future?

3. Where are the users?

These first two questions will help you to begin adequately planning for a good ratio of devices per access point while allowing for future growth

https://t.me/learningnets

# Capacity Design - How many users and devices are expected?

- The third question of great significance is, where are the users? Sit down with network management and indicate on the floor plan of the building any areas of high user density

- For example, one company might have offices with only 1 or 2 people per room, whereas another company might have 30 or more people in a common area separated by cubicle walls

https://t.me/learningnets

# Capacity Design - How many users and devices are expected?

- Other examples of areas with high user density are call centers, classrooms, and lecture halls

https://t.me/learningnets

# Capacity Design - How many users and devices are expected?

- You should always plan to conduct a validation survey when the users are present, not during off-hours

https://t.me/learningnets

# Capacity Design - How many users and devices are expected?

- A high concentration of human bodies can attenuate the RF signal because of absorption

https://t.me/learningnets

# Capacity Design - What types of client devices are connecting to the WLAN?

- Always remember that all client devices are not equal. Many client devices consume more airtime due to lesser MIMO capabilities

- For example, an older 802.11n tablet with a 1×1:1 MIMO radio transmitting on a 20 MHz channel can achieve a data rate of 65 Mbps with TCP throughput of 30 Mbps to 40 Mbps

https://t.me/learningnets

# Capacity Design - What types of client devices are connecting to the WLAN?

- An 802.11n tablet with a 2×2:2 MIMO radio transmitting on a 20 MHz channel might achieve a data rate of 130 Mbps with TCP throughput of 60 Mbps to 70 Mbps

- Many laptops also have 3×3:3 MIMO capabilities and thus are capable of higher data rates

- The bulk of newer smartphones and tablets are now 2×2:2 MIMO capable

https://t.me/learningnets

# Capacity Design - What types of client devices are connecting to the WLAN?

- The point is that devices with less MIMO capabilities consume more airtime and therefore affect the aggregate performance of any WLAN

- An AP can service more 2×2:2 MIMO clients efficiently as opposed to legacy 1×1:1 MIMO clients, which operate at lower data rates

https://t.me/learningnets

# Capacity Design – How many AP per room?

- The location where APs are physically mounted also needs to be taken into consideration based on capacity needs

- Some large areas of a building may only have two or three users that require Wi-Fi access

- Conversely, other areas, such as an auditorium may have hundreds of users and devices that need Wi-Fi connectivity

https://t.me/learningnets

# Capacity Design – How many AP per room?

- In education environment, due to capacity requirements, it has become commonplace to deploy one AP per room

- Please note that one AP per classroom maybe entirely unnecessary

- One AP per every two or three classrooms may be sufficient to meet capacity needs

https://t.me/learningnets

# Capacity Design – How many AP per room?

- How many APs are needed depends on the capacity requirements as well as customer environment

- Do you need to deploy one AP in every room?

  It depends on the number of devices, the type of devices, and the application traffic

https://t.me/learningnets

# Capacity Design - Band Steering

- The unlicensed 5 GHz frequency spectrum offers many advantages over the unlicensed 2.4 GHz frequency spectrum for Wi-Fi communications

- The 5 GHz U-NII bands offer a wider range of frequency space and many more channels

https://t.me/learningnets

# Capacity Design - Band Steering

- A proper 5 GHz channel reuse pattern using multiple channels will greatly decrease medium contention overhead caused by co-channel interference

- In the 2.4 GHz band, there is always medium contention overhead due to CCI simply because there are only three channels

https://t.me/learningnets

# Capacity Design - Band Steering

- Another consideration of the 2.4 GHz band is that, in addition to this band being used for WLAN networking, it is heavily used by many other types of devices, including microwave ovens, baby monitors, cordless telephones, and video cameras

https://t.me/learningnets

# Capacity Design - Band Steering

- With all of these different devices operating in the same frequency range, there is much more RF interference and a much higher noise floor than in the 5 GHz bands

https://t.me/learningnets

# Capacity Design - Band Steering

- So, if the use of the 5 GHz bands will provide better throughput and performance, how can we encourage the clients to use this band?

- For starters, it is the client that decides which AP and which band to connect to, typically based on the strongest signal that it hears for the SSID that it wants to connect to

https://t.me/learningnets

# Capacity Design - Band Steering

- Most access points have both 2.4 GHz and 5 GHz radios in them, with both of them advertising the same SSIDs

- Since the 5 GHz signals naturally attenuate more than the 2.4 GHz signals, it is likely that the client radio will identify the 2.4 GHz radio as having a stronger signal and connect to it by default

https://t.me/learningnets

# Capacity Design - Band Steering

- In many environments, the client would be capable of making a strong and fast connection with either of the AP's radios but will choose the 2.4 GHz signal because it is the strongest

https://t.me/learningnets

# Capacity Design - Band Steering

- A technology known as band steering has been developed to try to encourage dual-band client radios to connect to a 5 GHz AP radio instead of a 2.4 GHz AP radio

https://t.me/learningnets

# Capacity Design – How Band Steering works?

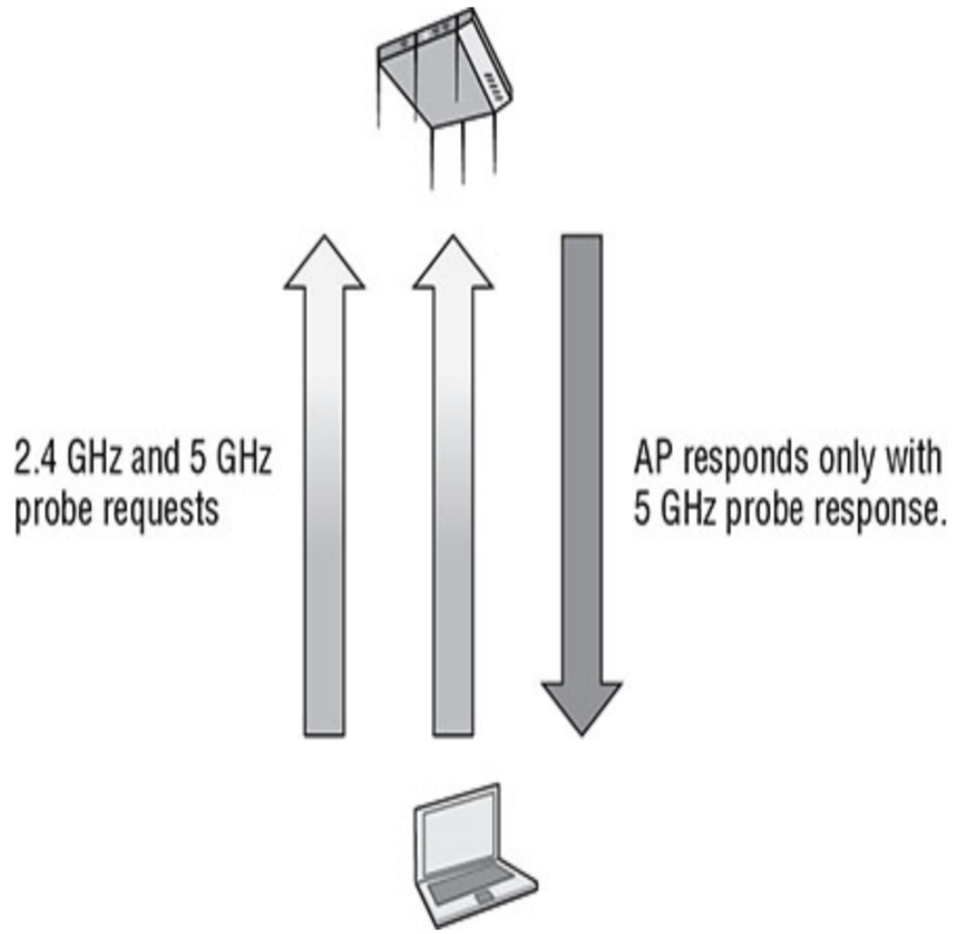
- When a dual-frequency client first starts up, it will transmit probe requests on both the 2.4 and 5 GHz bands, looking for an AP

- When a dual-frequency AP hears probe requests on both bands originating from the same client, the AP knows that the client is capable of operating in the 5 GHz band

https://t.me/learningnets

# Capacity Design – How Band Steering works?

- As depicted in the right figure, the AP will then try to steer the client to the 5 GHz band by responding to the client using only 5 GHz transmissions




https://t.me/learningnets

# Capacity Design – Load Balancing

- Load balancing the clients between access points, ensures that a single AP is not overloaded with too many clients, and that the total client population can be served by numerous APs, with the final result being better performance

https://t.me/learningnets

# Capacity Design – Load Balancing

- When a client wants to connect to an AP, the client will send an association request frame to the AP

- If an AP is already overloaded with too many clients, the AP will ignore the association response of the client

- Assumption is that the client will then send another association request to another nearby AP with a lesser client load

- Over time, the client associations will be fairly balanced across multiple APs

https://t.me/learningnets

# Capacity Design – Load Balancing

- The client load information will obviously have to be shared among the access points

- Load balancing is a control plane mechanism that can exist either in a distributed architecture where all the APs communicate with each other, or within a centralized architecture that utilizes a WLAN controller

https://t.me/learningnets

# Capacity Design - Airtime Consumption

- Designing for client device capacity has now become the norm

- WLAN design practices now dictate that you design to minimize airtime consumption, which is directly related to capacity design

https://t.me/learningnets

# Capacity Design - Airtime Consumption

- Wi-Fi is a half-duplex RF medium and only one radio can transmit on a channel at any given time

https://t.me/learningnets

# Capacity Design - Airtime Consumption

- Whenever a radio has won a transmission opportunity, the radio owns the available airtime until it finishes transmitting

- Yes, every radio needs to be able to transmit and deliver data; however, there are some simple WLAN design best practices that can minimize unnecessary airtime consumption

https://t.me/learningnets

# Capacity Design - Airtime Consumption

- Co-channel interference (CCI) is the top cause of unnecessary airtime consumption that can be minimized with proper WLAN design best practices

- Designing for –70 dBm coverage and high SNR also ensures that client devices will transmit 802.11 data frames at high data rates, based on the client radio's capabilities

https://t.me/learningnets

# Capacity Design - Airtime Consumption

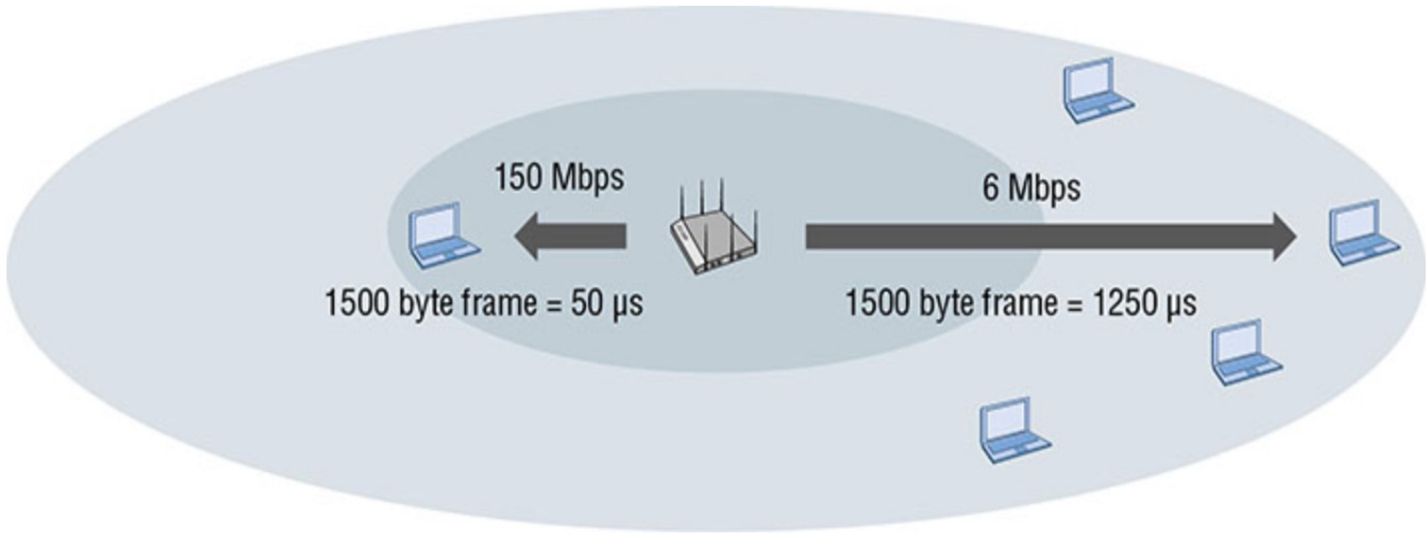
- So, what are some other WLAN design best practices that can reduce airtime consumption?

- One of the best ways to cut down on airtime consumption is to disable some of the lower data rates on an AP

https://t.me/learningnets

# Capacity Design - Airtime Consumption

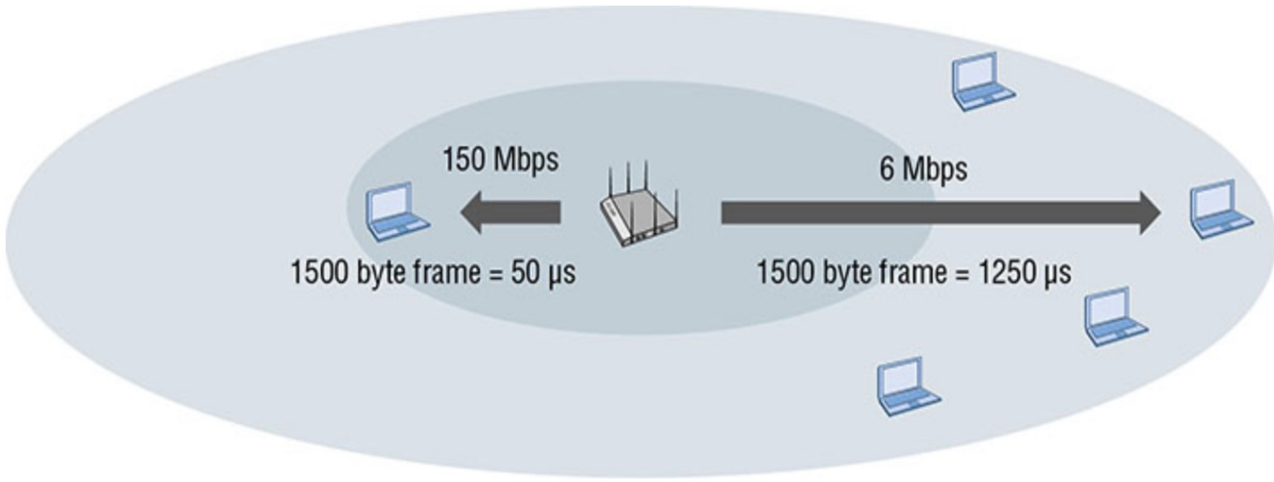
- Figure  depicts an AP communicating with multiple client stations at 6 Mbps while communicating with one client using a 150 Mbps data rate




https://t.me/learningnets

# Capacity Design - Airtime Consumption

- When 802.11 radios transmit at very low data rates, such as 6 Mbps or even lower, they effectively cause medium contention overhead for higher data rate transmitters due to the long wait time

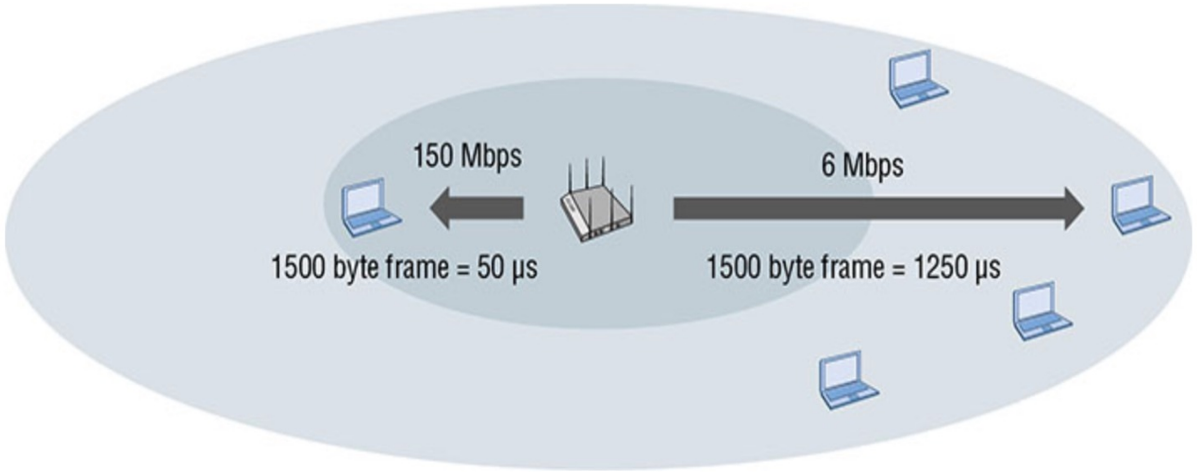



https://t.me/learningnets

# Capacity Design - Airtime Consumption

- A radio transmitting a 1,500-byte data frame at 150 Mbps might occupy the medium for 50 microseconds

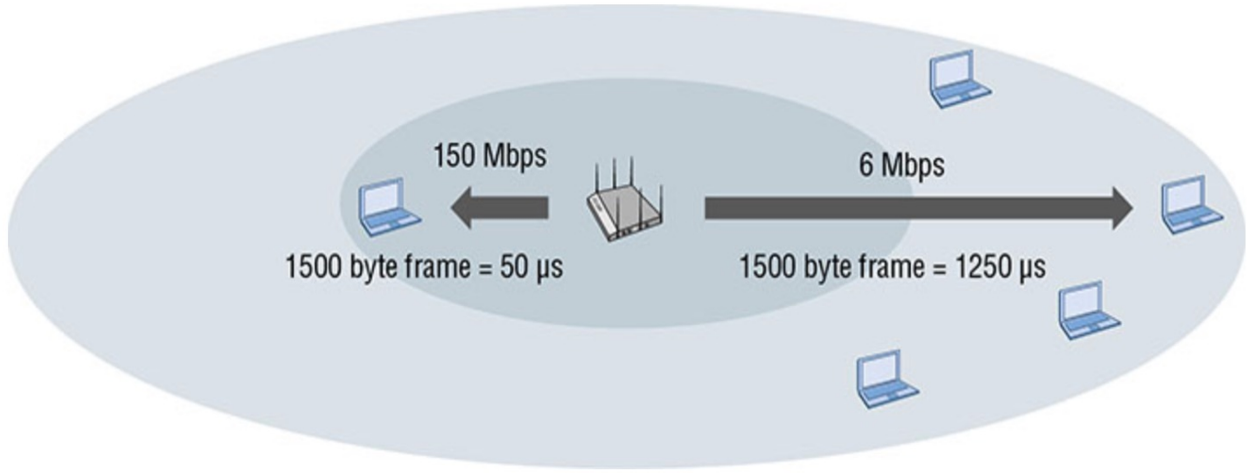
- Another radio transmitting at 6 Mbps may take 1,250 microseconds to deliver the same 1,500 bytes



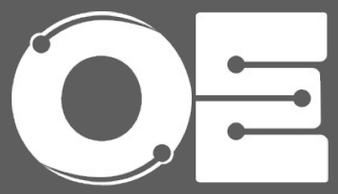

https://t.me/learningnets

# Capacity Design - Airtime Consumption

- In other words, the same data payload consumes 25 times more airtime when being delivered at the lower data rate




https://t.me/learningnets



# Area of Expertise – Large Scale Networks

https://t.me/learningnets

# Agenda

- Segment Routing Basics
- SR Segment Types
- SR Fast Reroute
- SR – Traffic Engineering
- SR – LDP Interworking
- EVPN
- LSN – Large Scale NAT/CGN
- mVPN
- Automation – Cloosed Loop Automation and Lifecycle Management
- uRPF, RTBH, QoS Intserv and Diffserv and IP Multicast Routing were explained in their modules

https://t.me/learningnets

# Segment Routing Theory

- Segment Routing (SR) is a protocol, designed to forward data packets on a network based on source routes

- Segment Routing divides a network path into several segments and assigns a segment ID (SID) to each segment and forwarding node

- The segments and nodes are sequentially arranged into a segment list to form a forwarding path

https://t.me/learningnets

# Segment Routing Theory

- Segment Routing is divided into two types, based on the forwarding plane

- Segment Routing MPLS (SR MPLS for short) is based on the MPLS forwarding plane, whereas Segment Routing IPv6 (SRv6 for short) is based on the IPv6 forwarding plane

https://t.me/learningnets

# Segment Routing Theory

- It is an alternate to LDP and RSVP which means without using the LDP and RSVP we still can generate and distribute the transport labels and steer the traffic without using the RSVP signaling

https://t.me/learningnets

# Segment Routing Theory

- Segment routing is based on label switching but for label generation and distribution LDP and RSVP is not used. Segment Routing is an extension to an IGP (OSPF/ISIS)

- Labels are called segments in Segment Routing MPLS

https://t.me/learningnets

# Global and Local Segments

The term Global means, ”within the SR Domain”

The SR Domain is the set of nodes participating in the SR network

https://t.me/learningnets

# Global Segment

- All SR enabled nodes in the SR Domain support the instruction that is associated with a Global Segment and each node in the SR domain installs the instruction of a Global Segment in its forwarding table

- An example of a Global Segment is a segment with instruction ‘forward packet according to shortest path to destination Node A’

https://t.me/learningnets

# Local Segment

- Only the node that originates a Local Segment supports the instruction that is associated with that Local Segment and only the node that originated the Local Segment installs the associated instruction in its forwarding table

- An example of a Local Segment is a segment advertised by Node A with the instruction ‘forward packet on interface to Node B’

https://t.me/learningnets

# Local Segment

- The local keyword of a Local Segment doesn't mean that nodes in the network don't know about other nodes' Local Segments

- In order to be able to use Local Segments, other nodes need to know about their existence and meaning

https://t.me/learningnets

# Segment Routing Global Block

- Segment Routing Global Block (SRGB) is the range of labels reserved for segment routing

- SRGB is a local property of a segment routing node

https://t.me/learningnets

# Segment Routing Global Block

- In MPLS, architecture, SRGB is the set of local labels reserved for assigning labels to global segments like Node-SIDs originated by a router

- In IOS-XR default SRGB range is from 16000-23999 and Dynamic label range is 24000-1048575

https://t.me/learningnets

# Segment Routing Global Block

- Every SR node advertises its SRGB as label base and label range

- So, in the case of IOS-XR it will be Base=16000 and range = 8000 via IGP

https://t.me/learningnets

# Segment Routing Global Block

- It is possible that different nodes can have different SRGB values either because the operator configured or due to vendor implementation differences and hence the absolute Node-SID value associated with an IGP Prefix Segment can change from node to node

https://t.me/learningnets

# Segment Routing Global Block

- There is a reserved SR (SRGB) block at every node and allocations of labels for Prefix/Node-SID value happens from that block

https://t.me/learningnets

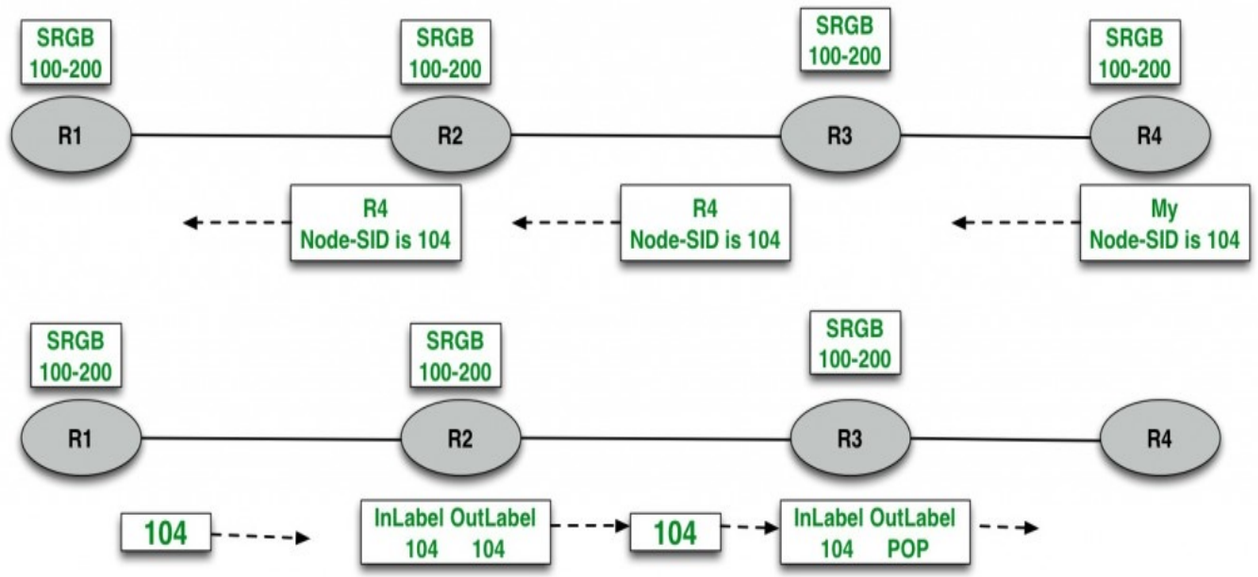
# Segment Routing Global Block

- In an SR domain if SRGB range of all the nodes is same then all the SR nodes will install the same SID/label value for a node or prefix

https://t.me/learningnets

# Segment Routing Global Block

- In this figure, I can assign an absolute value of 104 to R4 and R4 advertises that via an IGP, When R3 receives this information, It reserves 104 from its SRGB block for R4




- This information gets further relayed to R2 via IGP

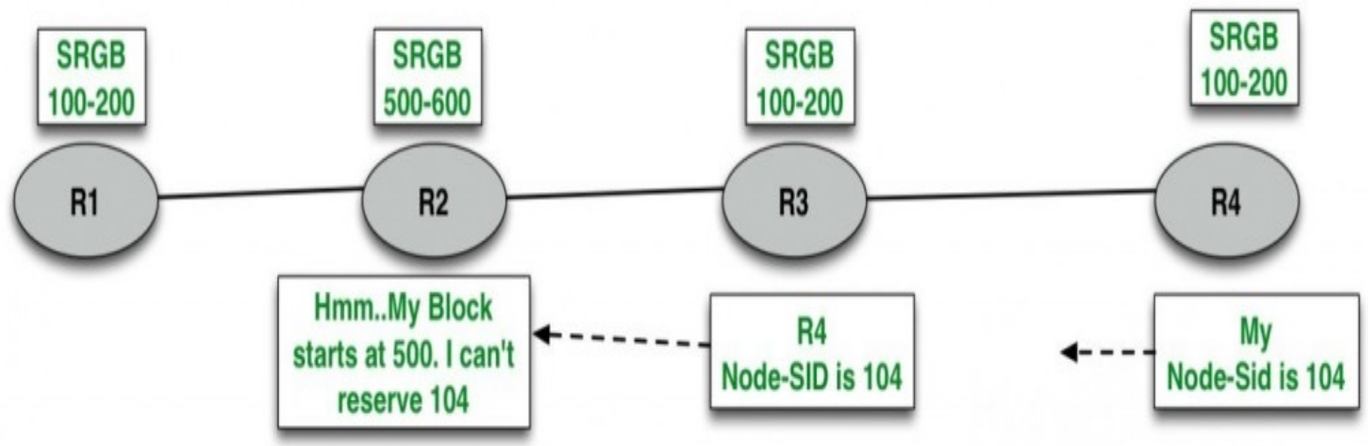
https://t.me/learningnets

# Segment Routing Global Block

- But what if the SRGB are different for different nodes

- In this figure If I assign an absolute value of 104 as Node-SID to R4 and when R4 advertises this via an IGP, R2 can't not assign 104 as it's outside of its SRGB range

- So how do we solve this situation?




In order to avoid a situation like this we have an option to assign a Unique Index value rather than an Absolute value to a Node-SID.

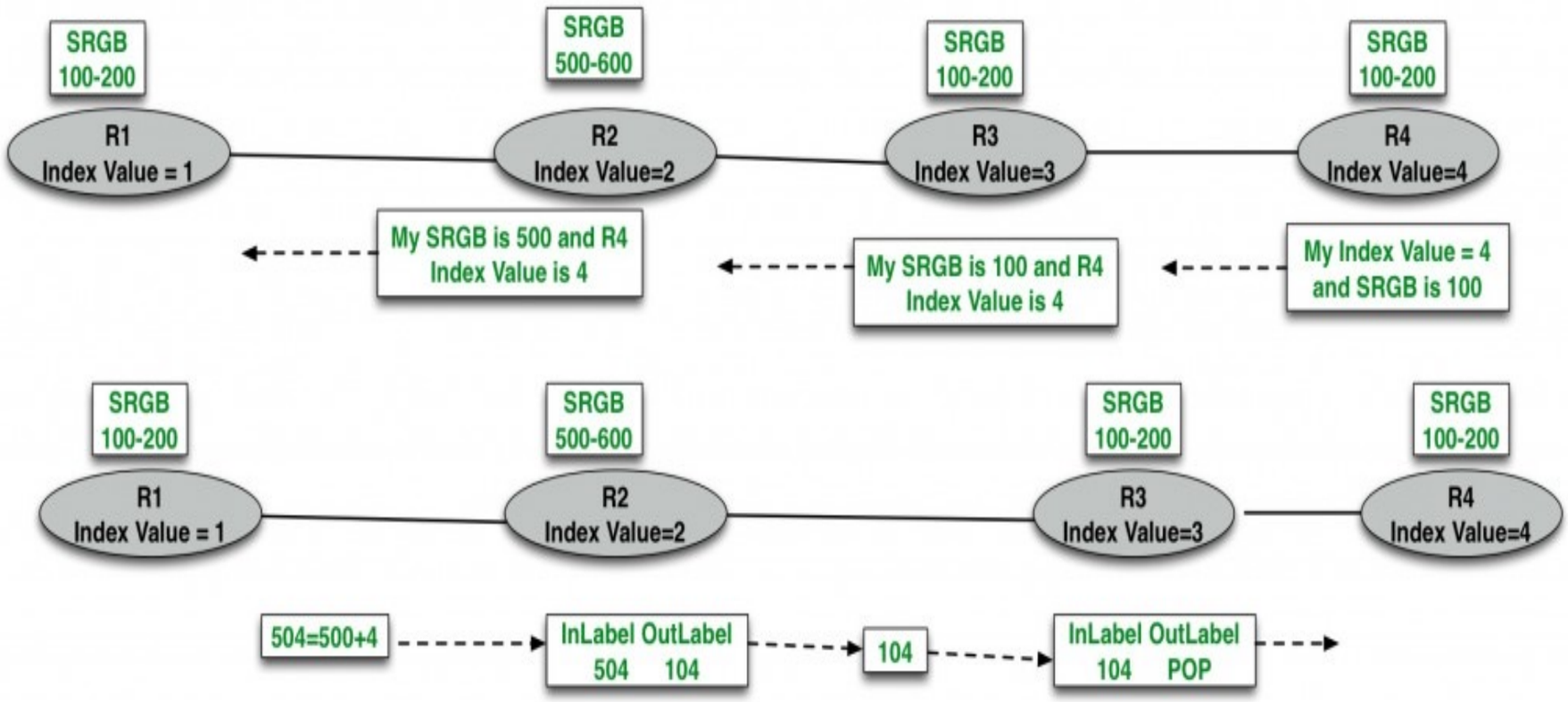
https://t.me/learningnets

# Segment Routing Global Block

- In this figure though, The operator has configured Unique Index values to each SR Node

- When R4 advertises its Index Value to R3, then R3 programs 104 as a local label

- R3 further advertises Index value of R4 to R2, R2 assigns its local label 504 (500+4) based on the below formula




https://t.me/learningnets

# Segment Routing Global Block

- Now, as you can see that the label operation looks very similar to LDP when operating in the independent label distribution mode with the difference that the label value used to forward a packet to each downstream router is computed by the upstream router based on the advertised prefix SID index using the formula Local Label (Prefix SID) = Start label + SID Index

https://t.me/learningnets

# Type of Segments?

- Segment or SID (Segment ID) is distributed with with IGP and BGP

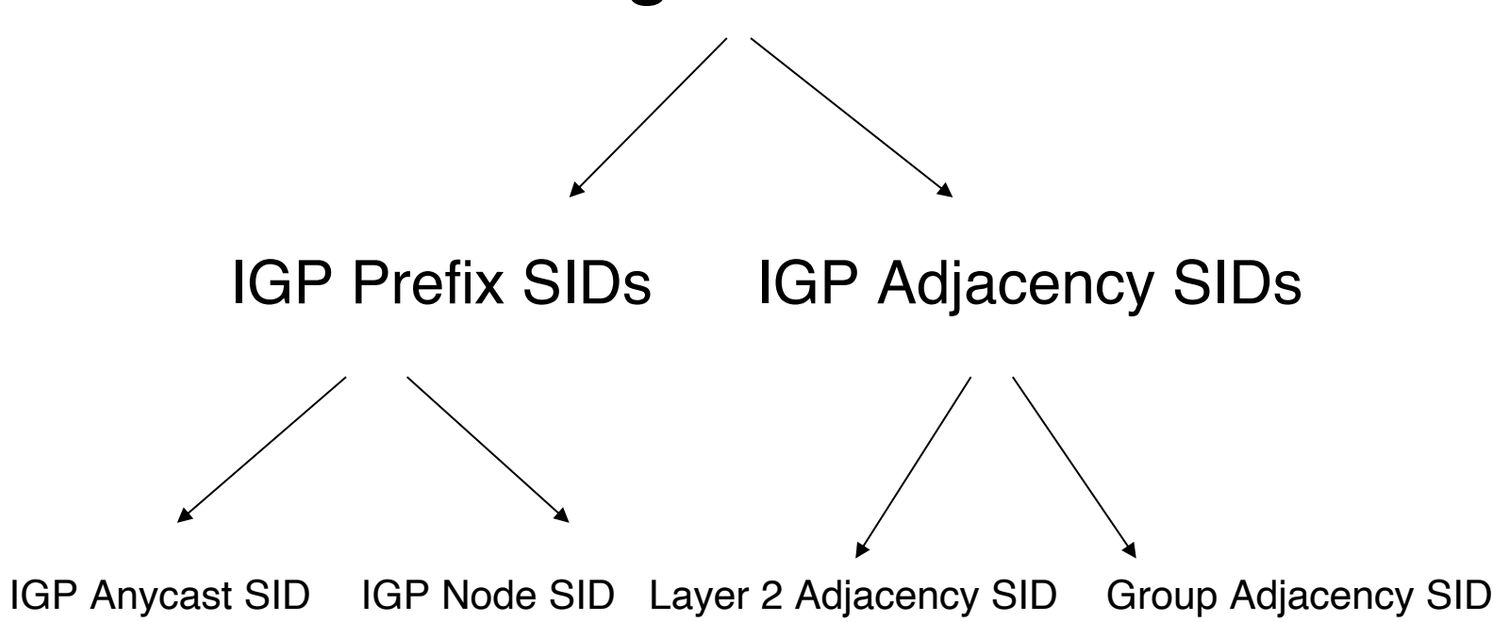
- Segments distributed by IGP are called IGP Segments and those distributed by BGP are called BGP Segments

- There will be more segment type in the future so these may not be the exhustive list

https://t.me/learningnets

# IGP Segments

- Interior gateway protocol (IGP) distributes two types of segments: Prefix segments and Adjacency segments




https://t.me/learningnets

# IGP Segments

- Each router (node) and each link (adjacency) has an associated segment identifier (SID)

- A prefix SID is associated with an IP prefix

- The prefix SID is manually configured from the segment routing global block (SRGB) range of labels, and is distributed by IS-IS or OSPF

https://t.me/learningnets

# IGP Segments

- There are two types of IGP Prefix Segments, these are Anycast Segment and Node Segment

https://t.me/learningnets

# IGP Segments

- A node SID is a special sub-type of prefix SID that identifies a specific node

- It is configured under the loopback interface with the loopback address of the node as the prefix and it must be globally unique

https://t.me/learningnets

# IGP Segments

- An adjacency segment is identified by a dynamic label called an adjacency SID, which represents a specific adjacency, such as egress interface, to a neighboring router

- The adjacency SID is distributed by IS-IS or OSPF

- The adjacency segment steers the traffic to a specific adjacency, and it is locally unique

https://t.me/learningnets

# IGP Segments

- IGP Adjacency Segment comes with the two types as well, these are Layer 2 Adjacency SID and Group Adjacency SID

- Now, let’s have a look at these SIDs in more detail

https://t.me/learningnets

# IGP - Node/Prefix-SID

- The first major type of SID is Node-SID or Prefix-SID

- A Node-SID represents a Node and has a global significance, something like a loopback of a router

https://t.me/learningnets

# IGP - Node/Prefix-SID

- Like an operator assigns a loopbacks to their routers, it's expected that the Node-SID value will be assigned to every node

- The assigned value can be an absolute or Index value and must be globally unique

https://t.me/learningnets

# IGP- Node/Prefix-SID

- All the nodes, switch packets towards Node-SID via the shortest path

- From global significance we mean here that every node in the SR domain will install an MPLS label in the forwarding plane representing Node-SID

https://t.me/learningnets

# IGP- Node/Prefix-SID

- How is this different then every router installing a label for /32 loopback in the case of MPLS+LDP domain?

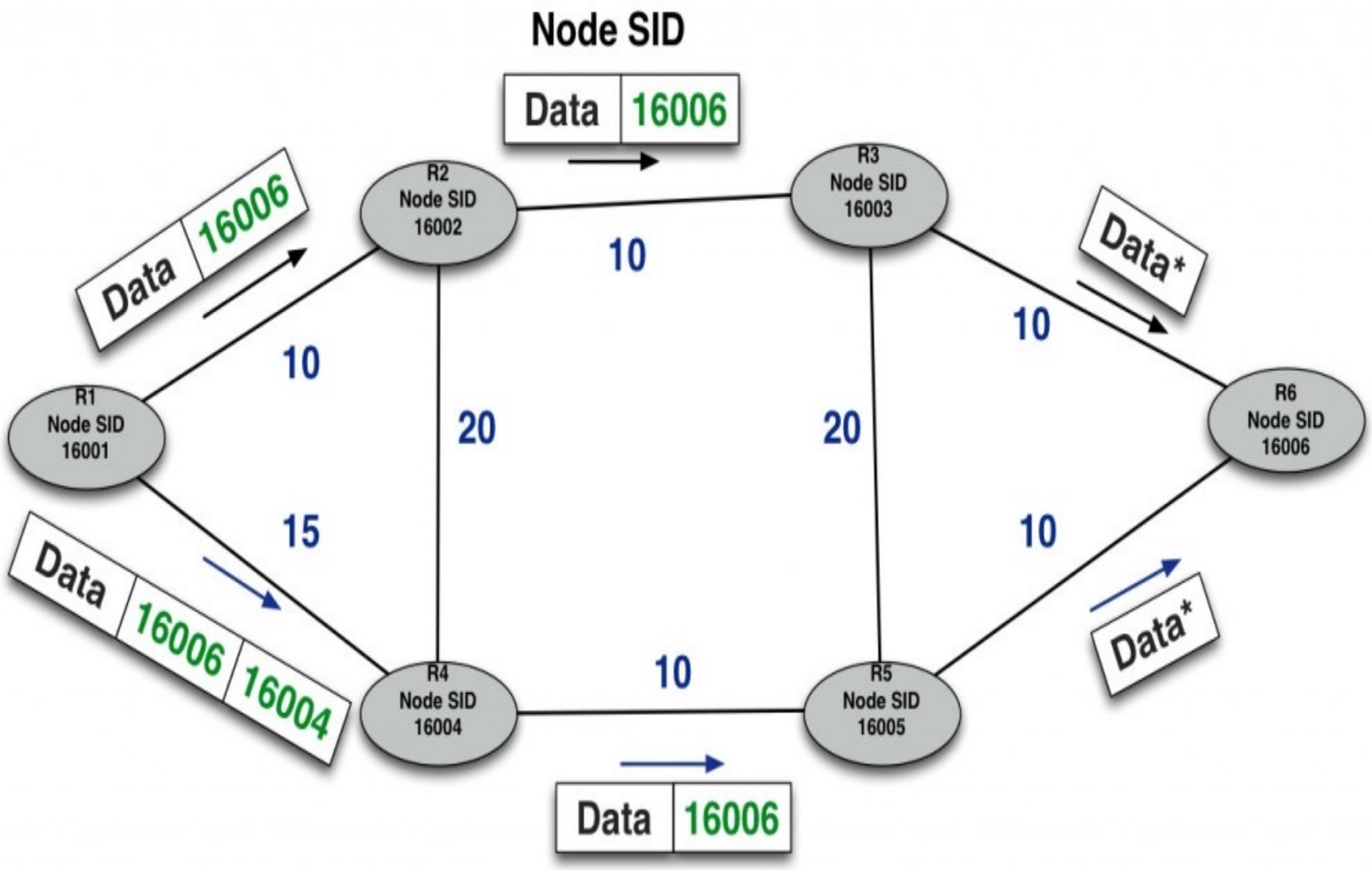
- Let’s wait till the end of this section to see if there is any difference

https://t.me/learningnets

# IGP- Node/Prefix-SID

- In this figure, Router has been assigned a unique Node-SID value

- IGP metrics are highlighted in Blue




https://t.me/learningnets

# IGP Anycast SID

- Assigning a unicast IP address on more than one node in the domain makes that unicast prefix an anycast prefix

- An IGP Anycast Segment (Anycast SID) is a special type of IGP Prefix Segment that is associated with an anycast prefix

- This SID is not identifying a specific node, rather a group of nodes

https://t.me/learningnets

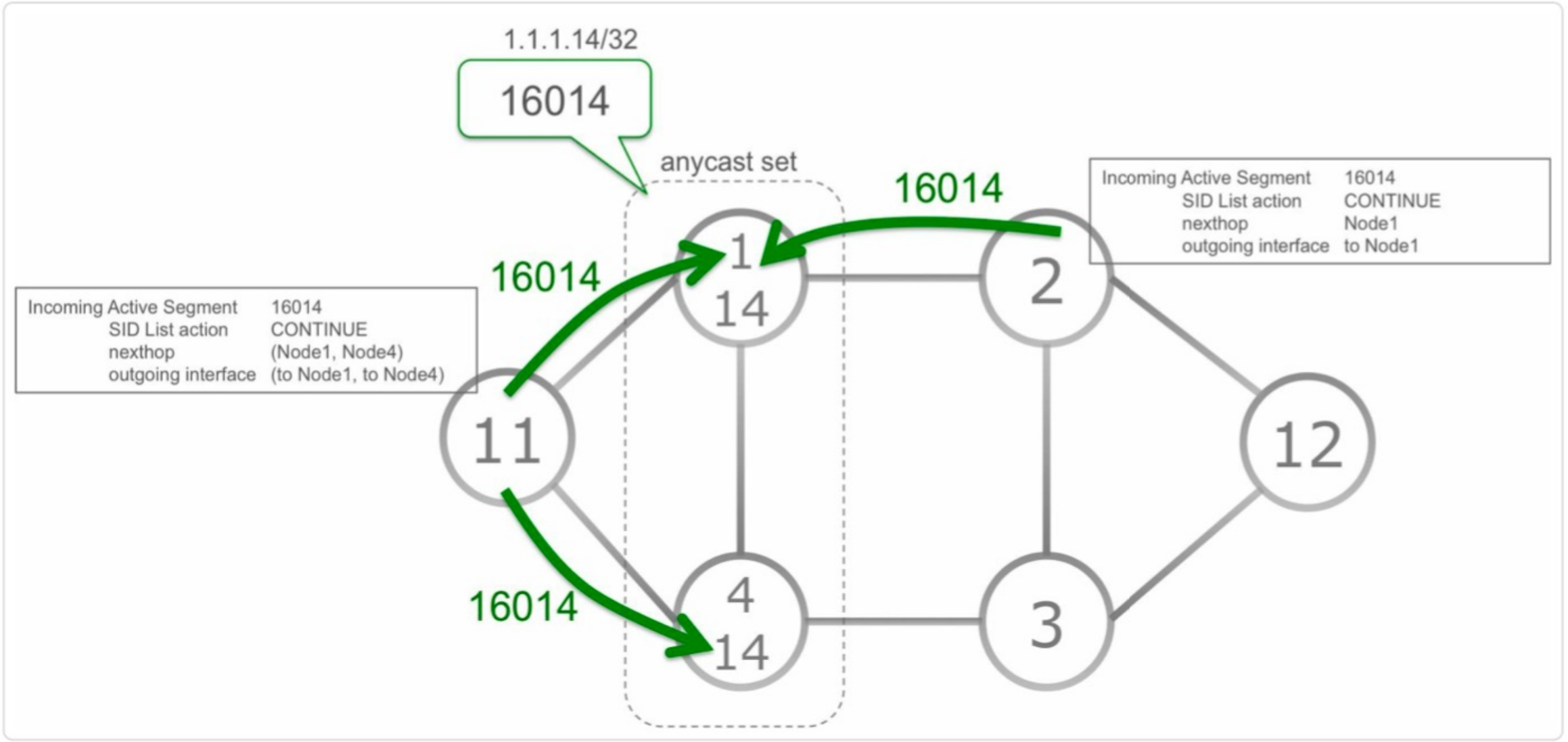
# IGP Anycast SID

- That's why Anycast SID is Prefix SID but not a Node SID

- In Node SID, N Flag is set, with Anycast SID, N Flag is not set with the advertisement

https://t.me/learningnets

# IGP Anycast SID




https://t.me/learningnets

# IGP Anycast SID

- Anycast SID is ECMP Aware

- Provides High Availability and Load Balancing

- Traffic Engineering based on High Level Intent (Send the traffic to the Backbone Routers in Paris, send the traffic to the Spine Switches in the East DC etc.)

https://t.me/learningnets

# IGP- Adjacency-SID

- The other most important SID is the Adjacency SID. They usually represent a link and have a local significance

- What I mean by local significance is that the remote SR node, unlike for Node-SID doesn’t install a state in the forwarding plane for an Adjacency SID

https://t.me/learningnets

# IGP - Adjacency-SID

- Only the directly connected Nodes program the forwarding plane for Adjacency SIDs

- They are absolute values (Not Index) and allocated dynamically by the router from the dynamic label range, which is outside of SRGB block

https://t.me/learningnets

# IGP - Adjacency-SID

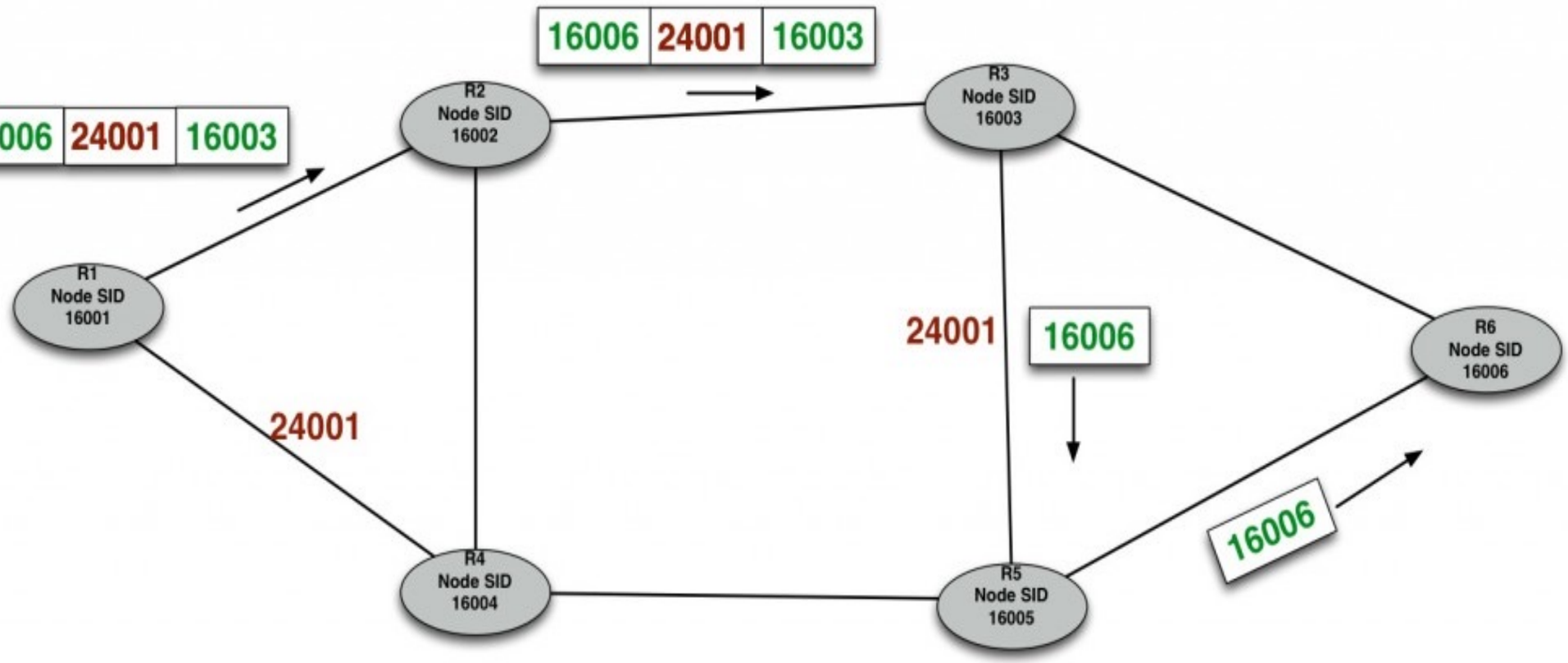
- So in the case of IOS-XR they will be allocated from the Dynamic label range i.e. 24000-1048575

- Since Adjacency SID's has local significance they don't have to be unique in the SR domain

https://t.me/learningnets

# IGP - Adjacency-SID

- In this figure, 24001 Adjacency SID is present between R1- R4 and R3-R5




https://t.me/learningnets

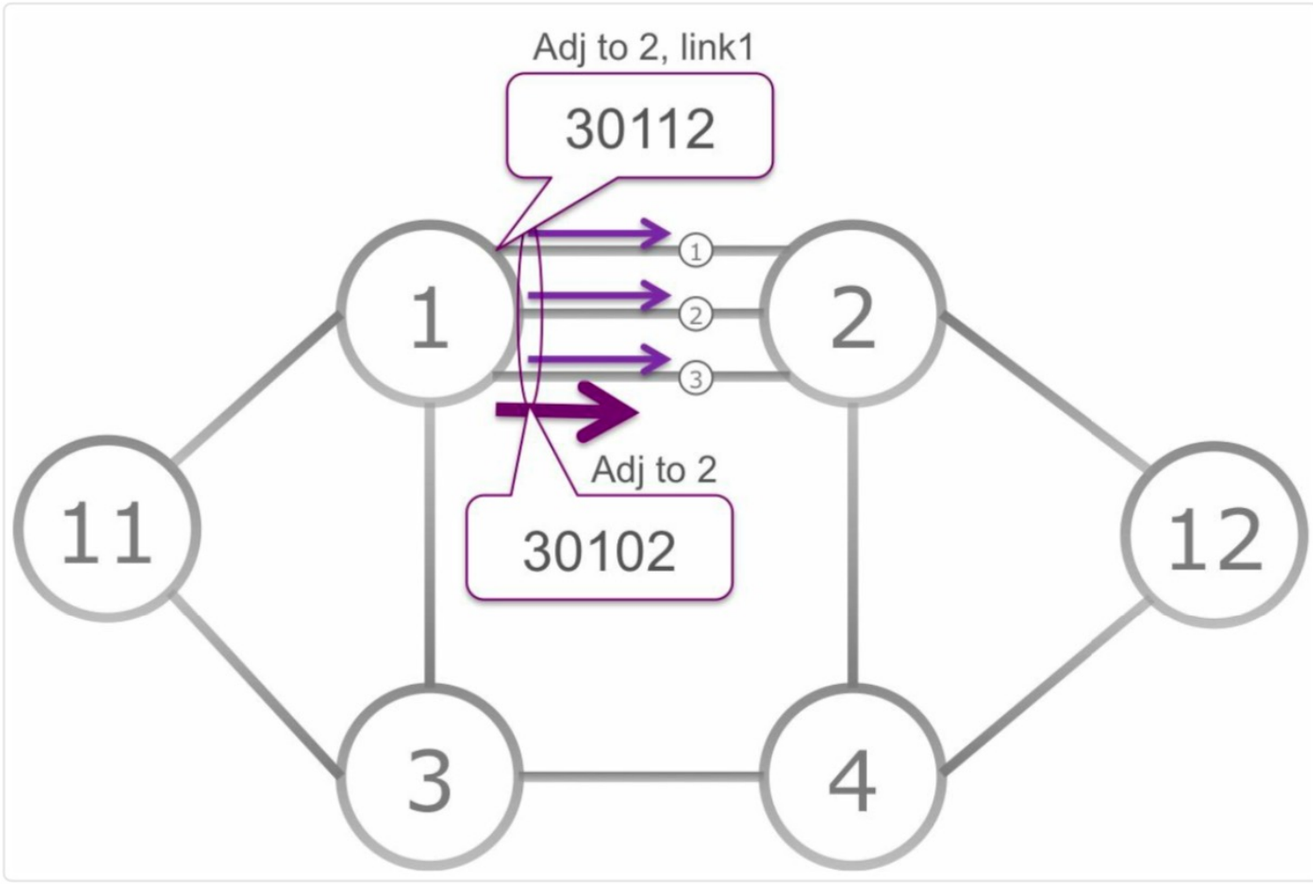
# Layer 2 Adjacency SID

- This is one of the sub-Adjacency SID type
- Link bundles are widely used. It combines multiple Layer2 interfaces in a single Layer 3 interface
- Traffic, normally is load balanced over all its member links
- If there is a requirement to steer traffic over specific indivudial path links, then a Layer 2 Adjacency SID is allocated for each bundle member

https://t.me/learningnets

# Layer 2 Adjacency SID

- Node 1 allocates 30102 as the Layer 3 Adjacency SID for the adjacency of Node 1 towards Node 2




- Node 1 allocates L2 Adj-SID for each bundle member

- For example 30112 for Link 1 in the bundle

30102 is used on Node 1 to load balanced over all 3 links but 30112 is used to steer the traffic over only Link1

https://t.me/learningnets

# IGP - Adjacency-SID

## Conclusions:

- Adjacency SID provides Explicit Routing advantage, this can be seen as advantage, but at the same time disadvantage because usually more Segments is used in the network with them

- Make sure SID Depth is not a problem!

https://t.me/learningnets

# IGP - Adjacency-SID

## According to inventor of Segment Routing:

- “We try to rarely use Adj-SIDs. They don’t support ECMP and they only express one single link hop.We prefer expressing an explicit path with Prefix-SIDs. We get shorter SID lists and we spread the demand over all the ECMP paths of the Prefix-SIDs”

https://t.me/learningnets

# BGP Segments

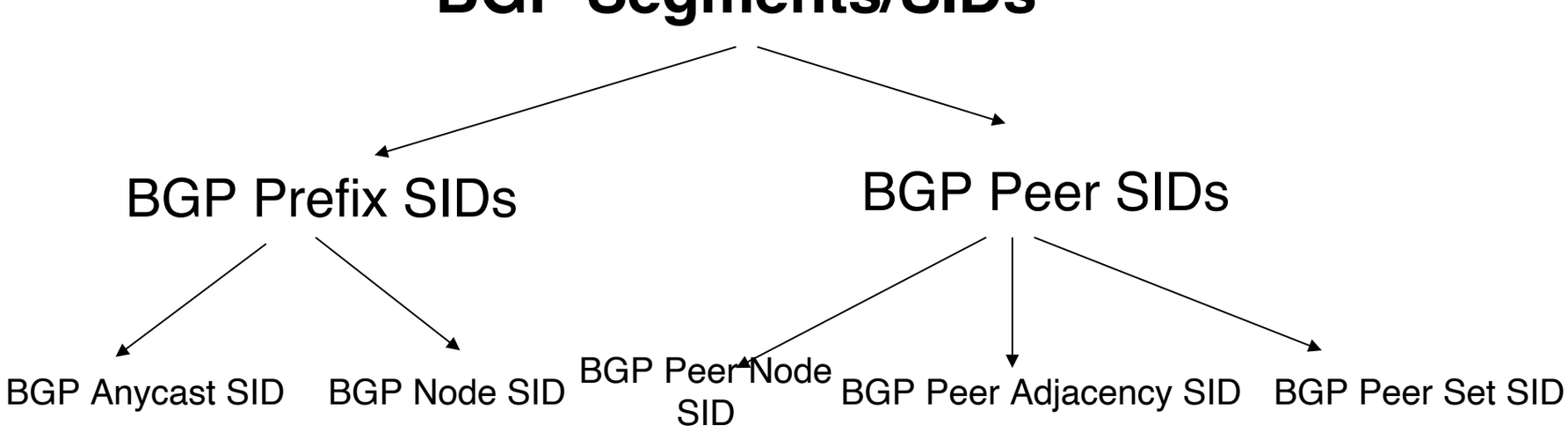
- BGP can be used as SR Control Plane and can distribute SIDs in the network

- BGP as in RFC 7938 and RFC 3107 , RFC 8277 cases, used in the networks for many reasons

- When BGP is used for these reasosn, BGP distribute the SIDs, not IGP

https://t.me/learningnets

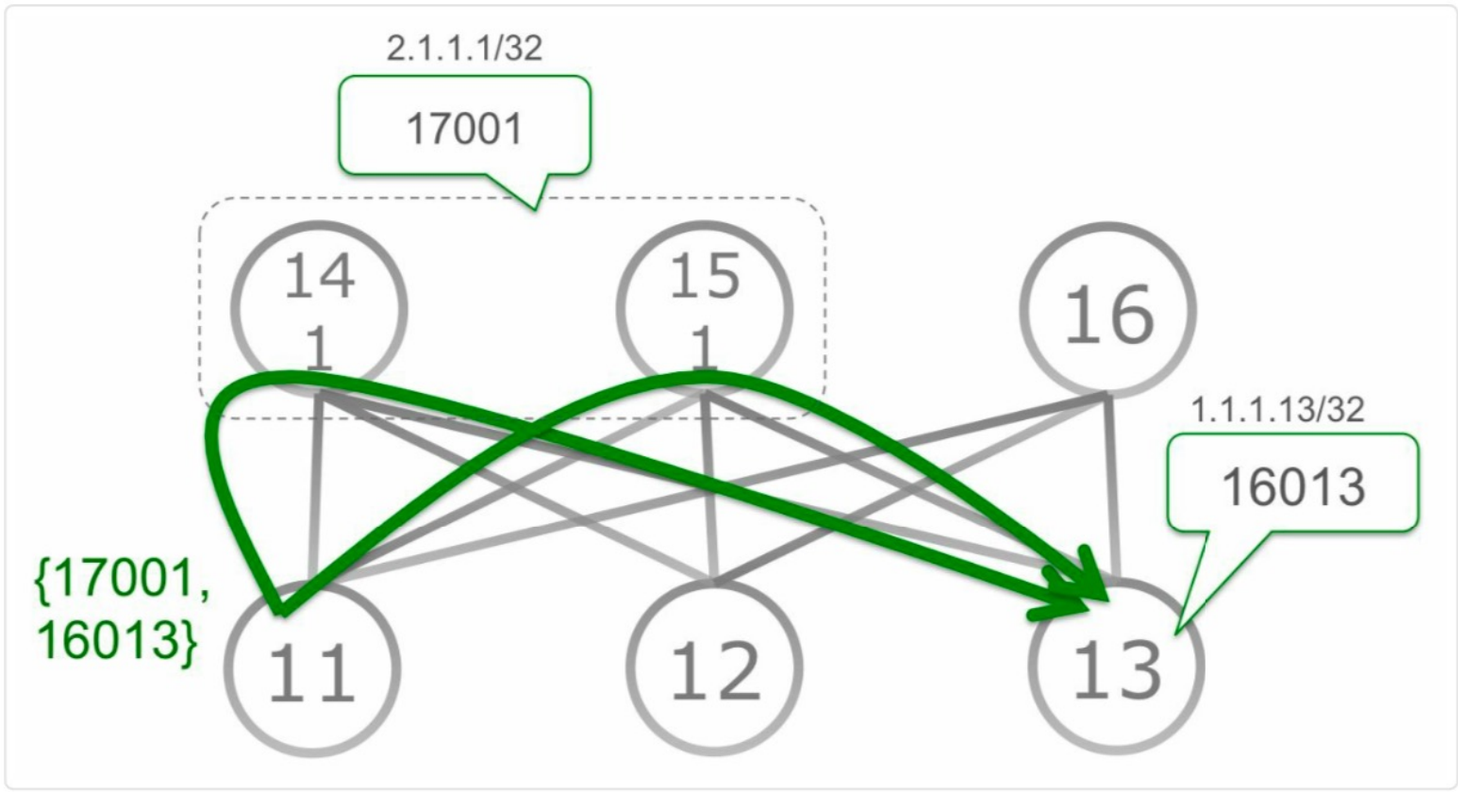
# BGP Segments




https://t.me/learningnets

# BGP Anycast SID

- Leaf and Spine/CLOS
- If I want to exclude Spine 16, send a traffic to Leaf 13, by only using Spine 14 and 15, allocate BGP Anycast SID 17001 for them




https://t.me/learningnets

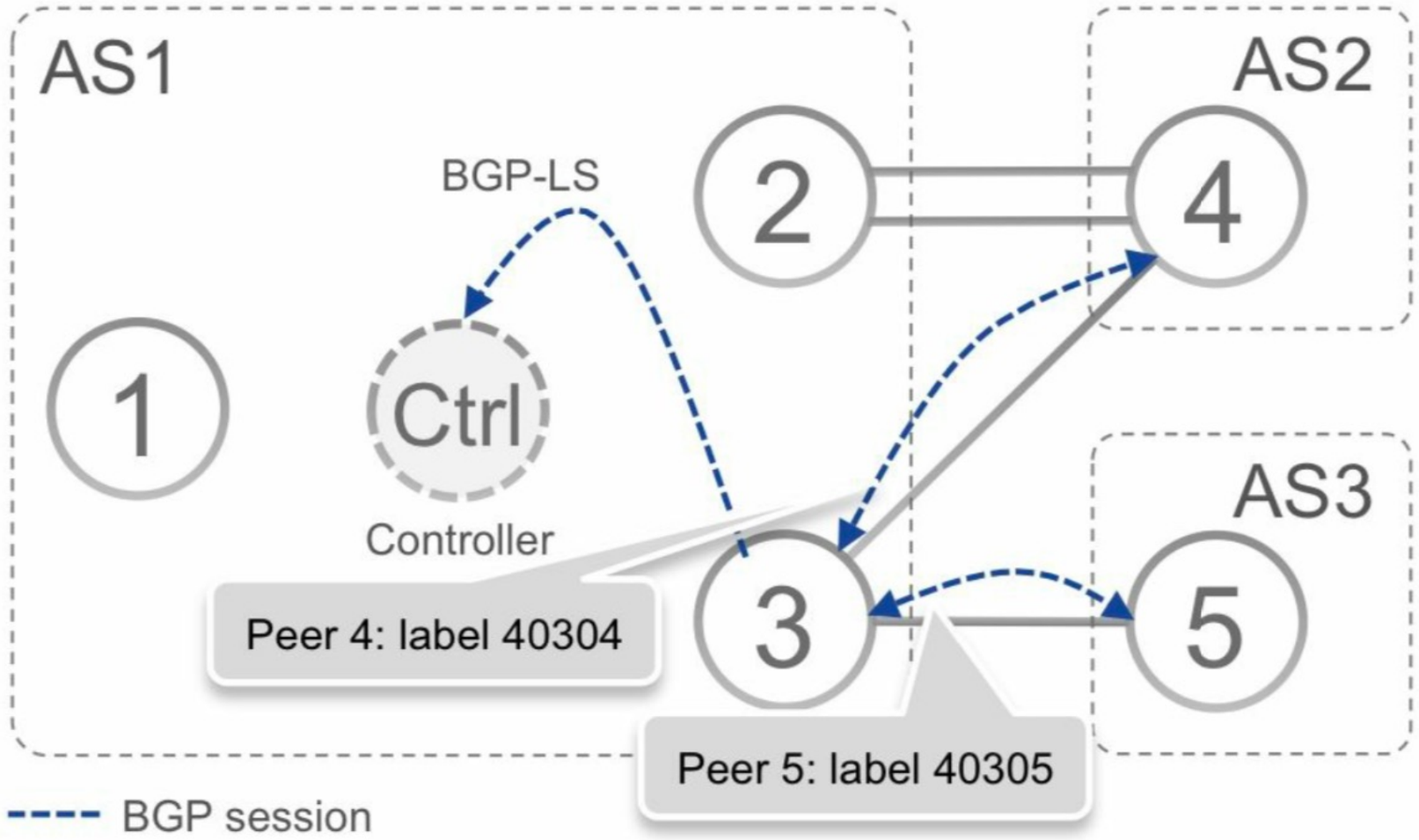
# BGP Peer Segments

- BGP Peer SID is associated with BGP Peering Sessions to a specific neighbor or a set of neighbors
- These are Local Segments
- Signalled via BGP-LS address family

https://t.me/learningnets

# BGP Peer Node Segment

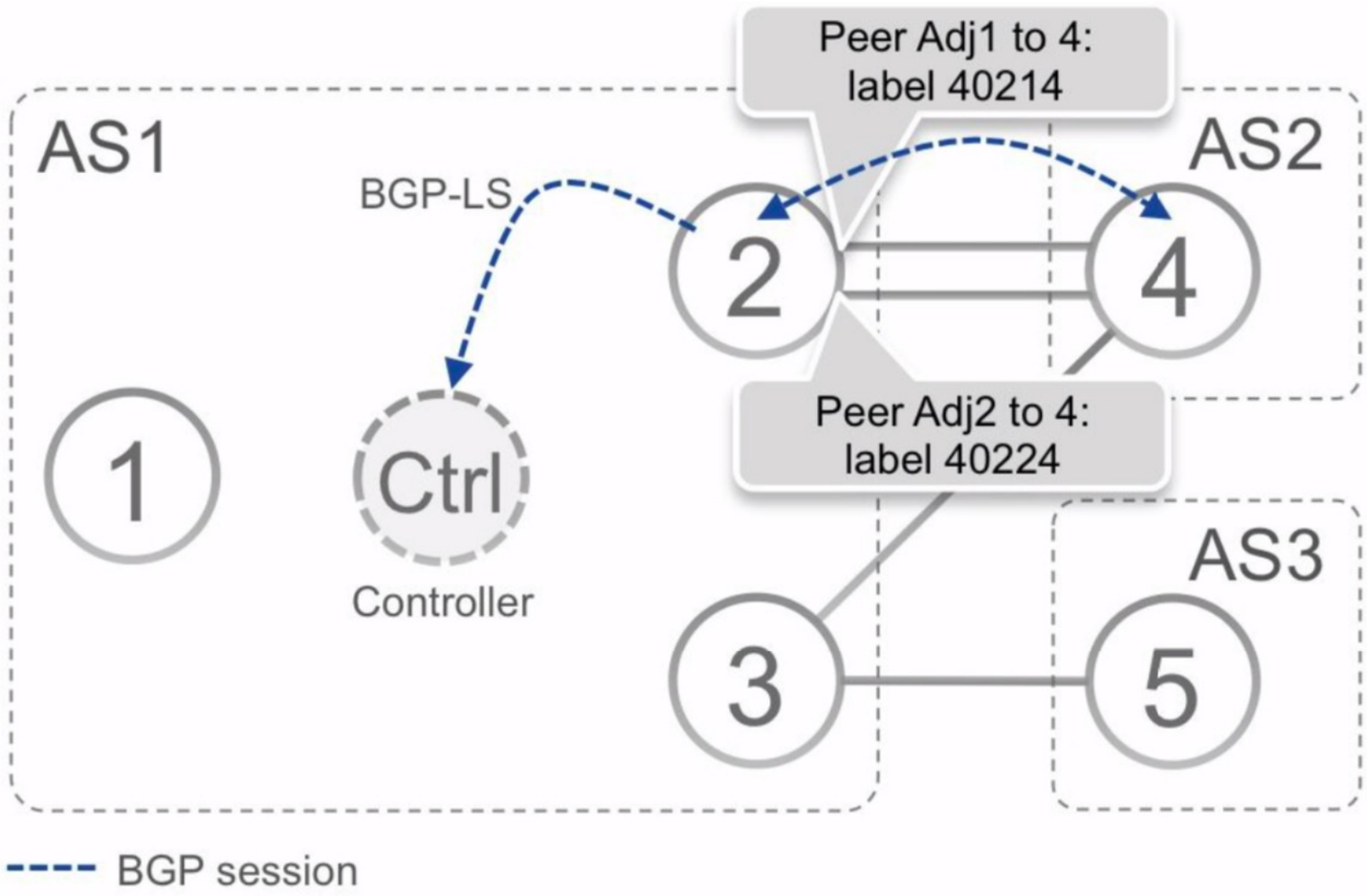
- Steer traffic to the specific BGP peer node via ECMP multipath towards that peer router




https://t.me/learningnets

# BGP Peer Adjacency Segment

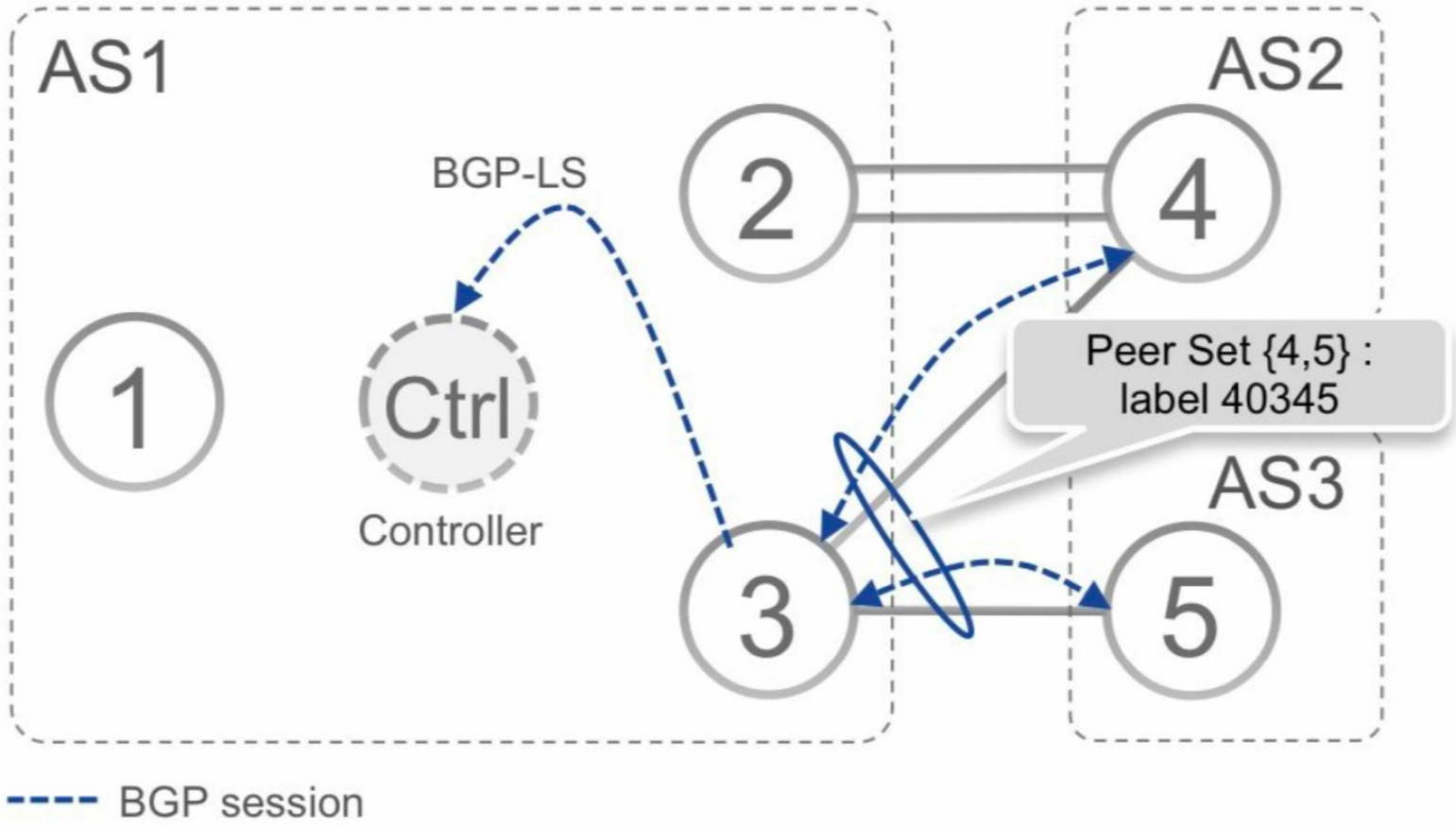
- Steer traffic to the specific BGP peer node via the specified interface towards that peer router




https://t.me/learningnets

# BGP Peer Set Segment

- Steer traffic to the specific BGP peer nodes that are members of the peer group via ECMP BGP Multiplath




https://t.me/learningnets

# SR/LDP Interworking

- One important concept to know is SR/LDP interworking

- For a technology to be successful, it’s very important that it doesn’t require a forklift upgrade and provide a migration/interoperability with current systems

https://t.me/learningnets

# SR/LDP Interworking

- If an operator is already running an LDP network and wants to implement SR but not every node is SR capable which could be due to various reasons like those nodes are from different vendors and don't support SR yet, or they may not be running the SR capable operating system

- SR/LDP interworking could be an option for them!

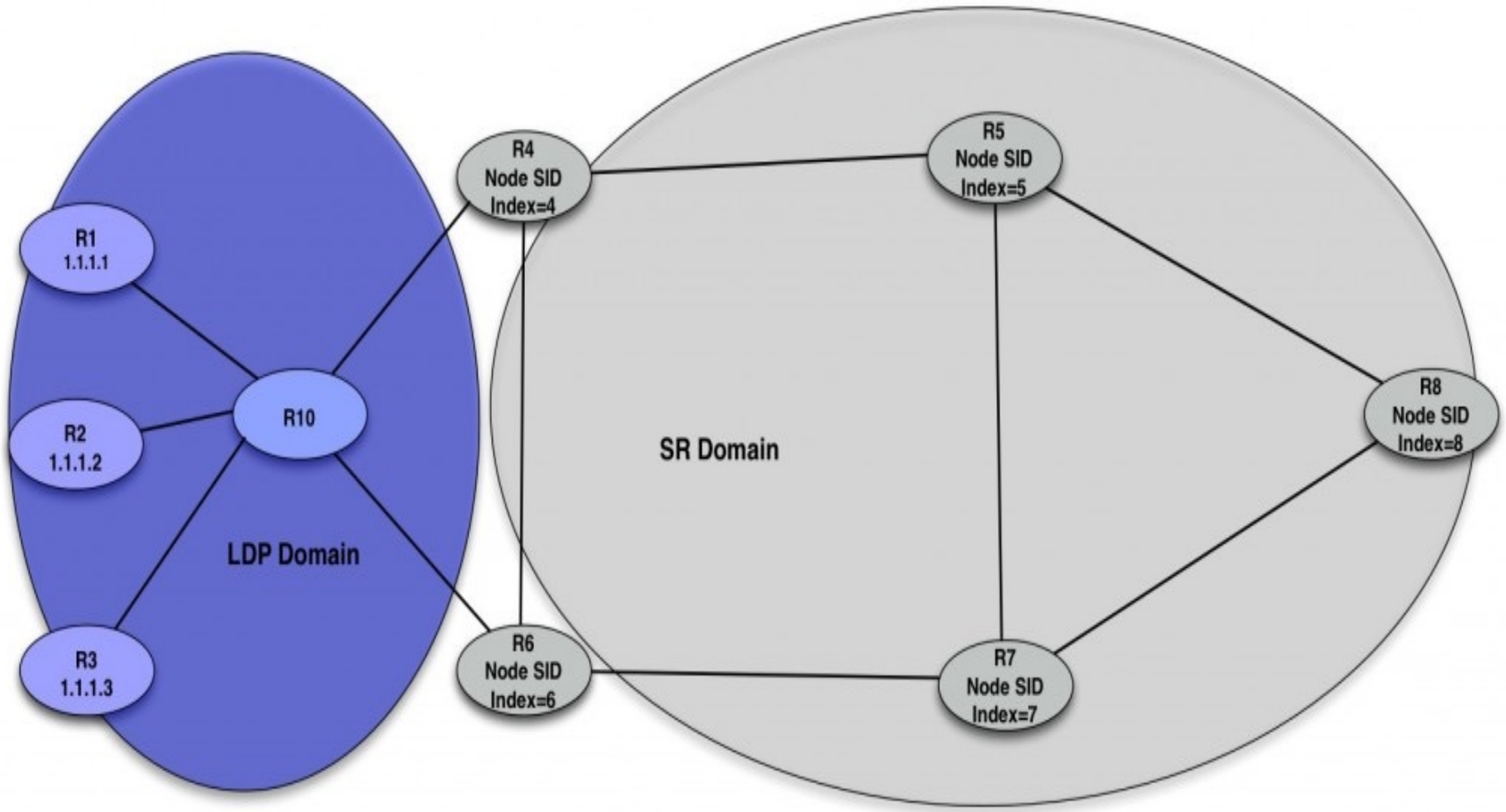
https://t.me/learningnets

# SR/LDP Interworking

- In this topology, R1, R2, R3 and R10 (In Blue) are only running LDP and R4-to-R8 Routers are SR capable Routers

- R4 and R6 will be running both SR and LDP as they are on the border between SR and LDP

- The whole network is running single IGP let's say IS-IS




https://t.me/learningnets

# SR/LDP Interworking

## Label Mapping from LDP to SR Domain:

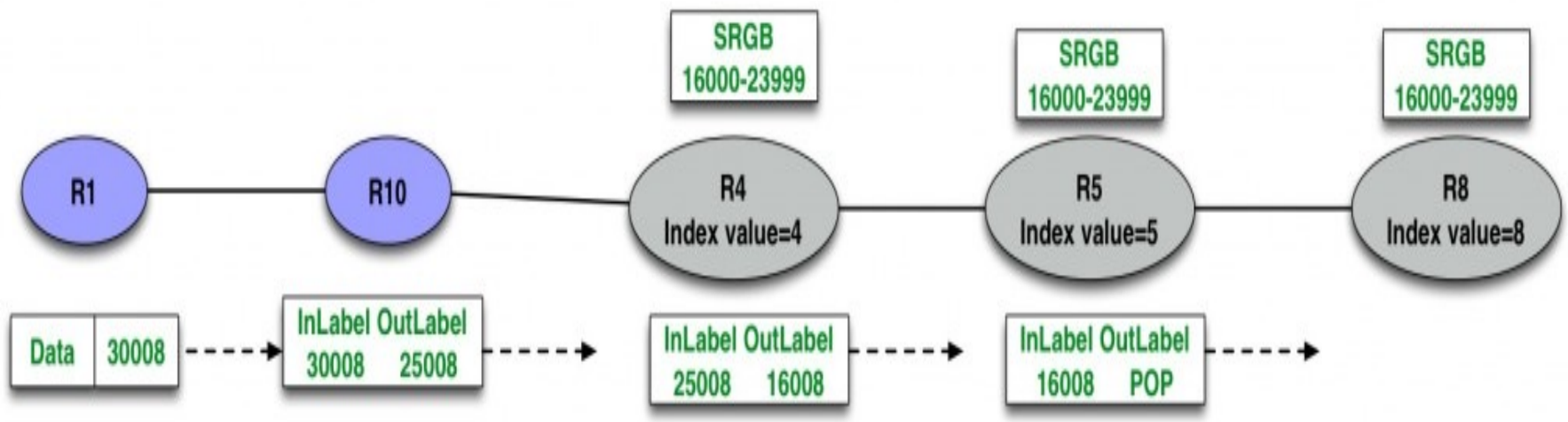
- So, if a router from LDP only domain wants to communicate with SR domain, it's pretty straightforward. Let's say R1 wants to send a packet to R8

- Since R4 and R6 are running LDP, they will allocate labels corresponding to the SR Nodes including R8 and advertise it to other LDP neighbors

https://t.me/learningnets

# SR/LDP Interworking

- When R1 sends the packet with label advertised by the LDP to R10, R10 swaps to the label towards either R4 or R6 (based on shortest path and in the example it's R4)




- R4 upon receiving the LDP label swaps to the SR label corresponding to the R8

https://t.me/learningnets

# SR/LDP Interworking

## Label Mapping from SR to LDP Domain:

- If R8 which is SR only node, wants to send a packet to LDP only nodes then we have a problem

- Since R8 doesn’t know the Node-SID corresponds to R1 it can’t send a packet to R1 (or other LDP nodes)

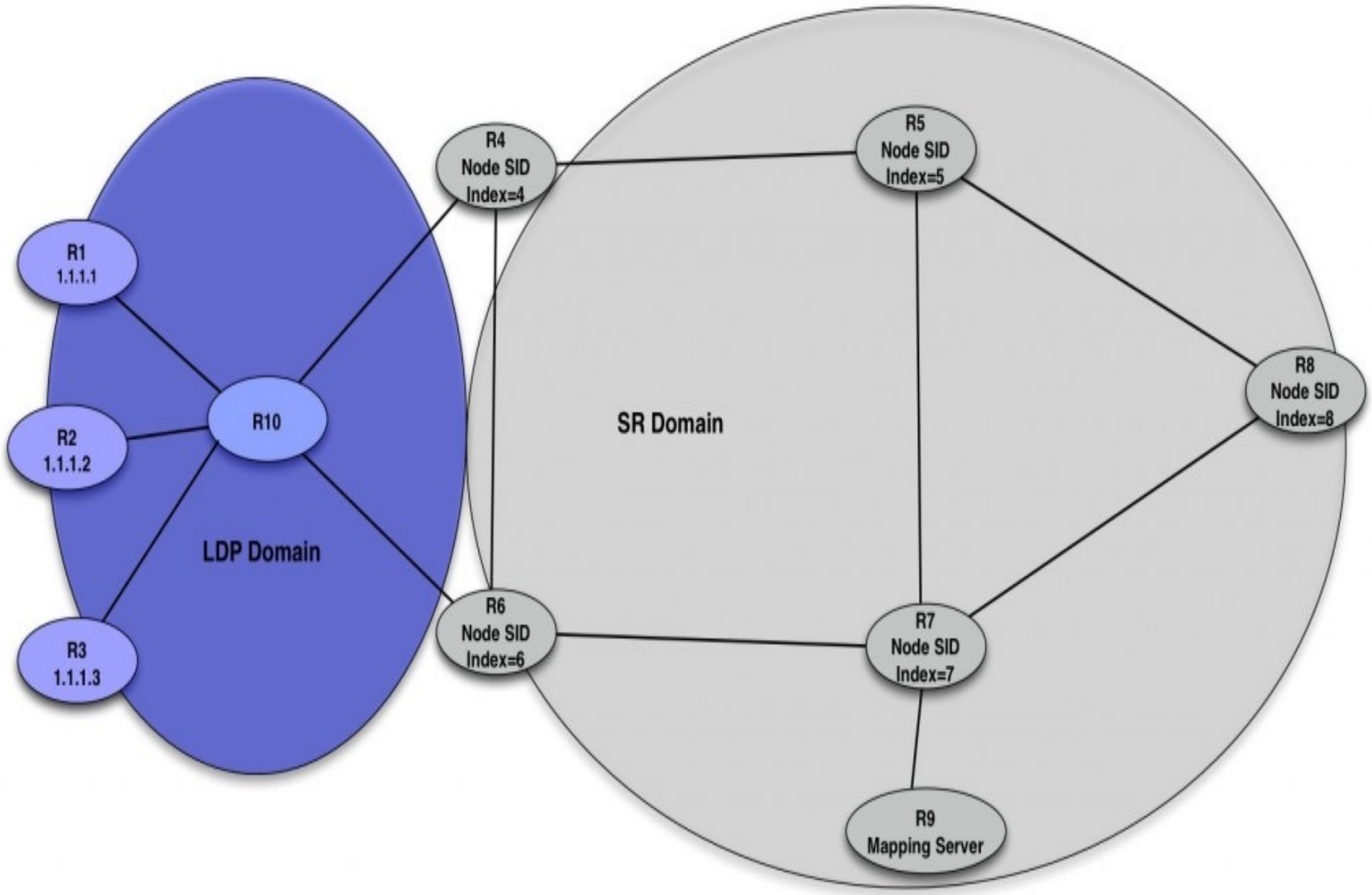
- In order to solve this problem, Mapping server functionality was created

https://t.me/learningnets

# SR/LDP Interworking

- In this topology, there is separate mapping server node for one reason that it doesn't have to be " inline",  as long as its participating in the IGP domain, otherwise any SR node can be a mapping server and doesn't have to have a separate box

- But obviously it will be a crucial box as it failure means the failure of communication between SR and LDP domain




https://t.me/learningnets

# SR/LDP Interworking

- On the mapping server, we define the Node-SID and their corresponding label value mapping something like this:

| Prefix | Index Value | Range |
|---|---|---|
| 1.1.1.1/32 | 2001 | 3 |

https://t.me/learningnets

# SR/LDP Interworking

- Where 1.1.1.1 /32 is the starting address of the loopback, 2001 is the starting label Index Value and range is 3. Mapping server will advertise this mapping to all the SR Nodes via IGP

- Assuming SRGB range is 16000-23999 then Node-SID values derived by SR Nodes for LDP routers will be

  - **1.1.1.1 /32 Node-SID label value is = 18001 (SRGB  Base 16000+2001)**

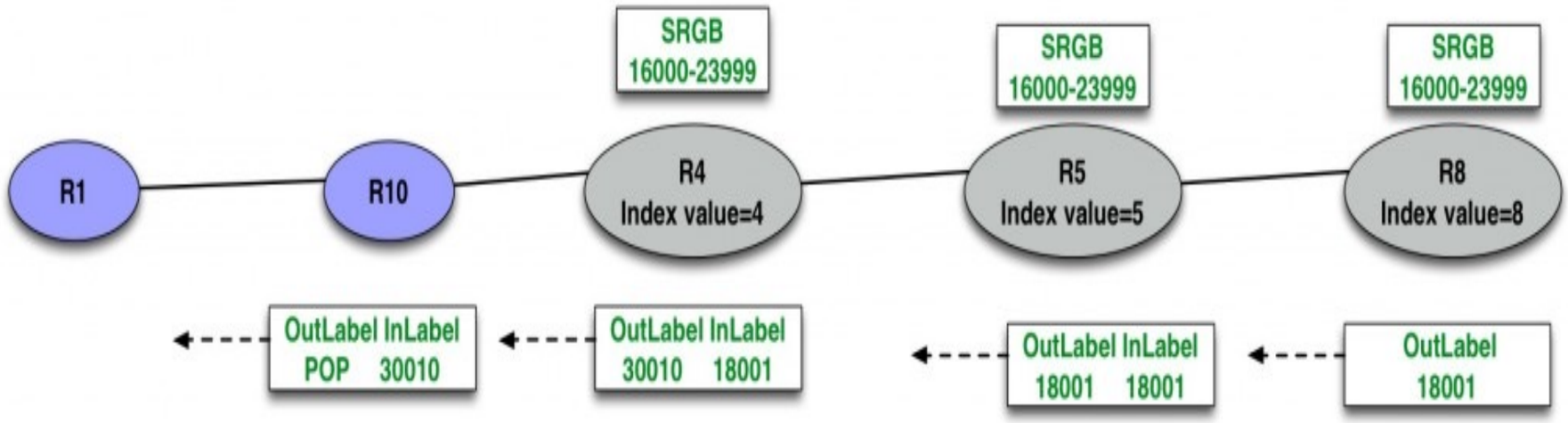
  - **1.1.1.2 /32 Node-SID label value is = 18002 (SRGB Base 16000+2002)**

  - **1.1.1.3 /32 Node-SID label value is = 18003 (SRGB Base 16000+2003)**

https://t.me/learningnets

# SR/LDP Interworking

- This way now SR Nodes will know what Node-SID value to use if they want to reach an LDP only Node




- In this topology, Once the packet reaches at R4 which is on the border of SR and LDP, it will swap the label from the SR-to-LDP Label which was advertised by R10 for R1

https://t.me/learningnets

# Segment Routing and MPLS VPN

- What are the differences between MPLS VPN and Segment Routing

- In fact, this is a wrong question, because Segment Routing provides transport, MPLS VPN provides Service

https://t.me/learningnets

# Segment Routing and MPLS VPN

- They are not mutually exclusive, but they are complementary

- In order to create end to end MPLS VPN service, Segment routing and MPLS VPN have to be provisioned in the network together

- So, Segment Routing is Underlay, MPLS VPN is Overlay

https://t.me/learningnets

# Segment Routing and MPLS VPN

- MPLS VPN can be setup without Segment Routing

- Instead of Segment Routing as a Transport mechanism, LDP, RSVP,BGP or plane IP can be a Transport for the VPNs

https://t.me/learningnets

# Segment Routing and MPLS VPN

- Please note that Segment Routing eliminates to use LDP for the transport label operation only

- If you setup MPLS layer 2 VPN service and use LDP for PW signaling , Segment Routing and Targeted LDP are used as two control plane protocols to setup MPLS L2 VPN

https://t.me/learningnets

# Segment Routing Fast Reroute – TI-LFA

- Segment Routing Fast Reroute – Traffic Engineering with Segment Routing uses LFA mechanism to provide 50 msec fast reroute capability

- LFA and rLFA (remote-LFA) cannot provide 100% coverage due to topology constraints but TI-LFA and RSVP-TE FRR can provide 100% coverage for any topology

https://t.me/learningnets

# Segment Routing Fast Reroute – TI-LFA

- Compare to LFA and rLFA provides better Fast Reroute coverage, but before we talk about TI-LFA, let's have a look at LFA and rLFA

- Also, let's understand some terms, such as P Space, Extended P Space, Q Space and PQ Node

https://t.me/learningnets

# LFA – Loop Free Alternate

- Traditionally MPLS FRR was the only option available to achieve sub-50ms recovery time

- Deploying MPLS FRR isn’t that straightforward

https://t.me/learningnets

# LFA – Loop Free Alternate

- IP FRR provide similar fast recovery methods in LDP based networks, but a lot simpler to implement from a complexity perspective

https://t.me/learningnets

# LFA – Loop Free Alternate

- Idea behind LFA is similar to MPLS FRR i.e. it pre-installs the backup next-hop into the forwarding plane

- LFA's doesn't introduce any protocol extensions and can be implemented on a per router basis, which makes it a very attractive option

https://t.me/learningnets

# LFA – Loop Free Alternate

- A route that doesn’t cause forwarding loop is called as Loop Free Alternate path. The router calculates these loop free alternate paths in advance and program them in the FIB

- Link State protocols are needed for LFA (OSPF or IS-IS)

https://t.me/learningnets

# LFA – Loop Free Alternate

- RFC 5286 offers a method for calculating LFA based on route inequalities

- Let’s have a look at these inequalities, one for link protection and another for node protection

https://t.me/learningnets

# LFA – Loop Free Alternate

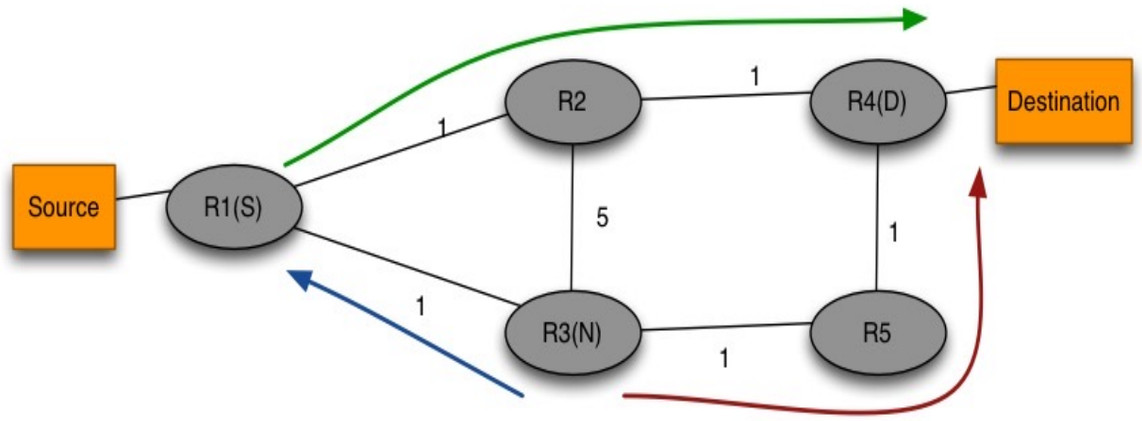
- **<u>LFA – Link Protection Inequality</u>**

Path is loop-free because N's best path is not through local router." Traffic sent to backup next hop is not sent back to S.

- Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D)

- Distance_opt is used to indicate the shortest distance from X and Y.
  S is used to indicate the calculating router.
  N is a neighbor of S.

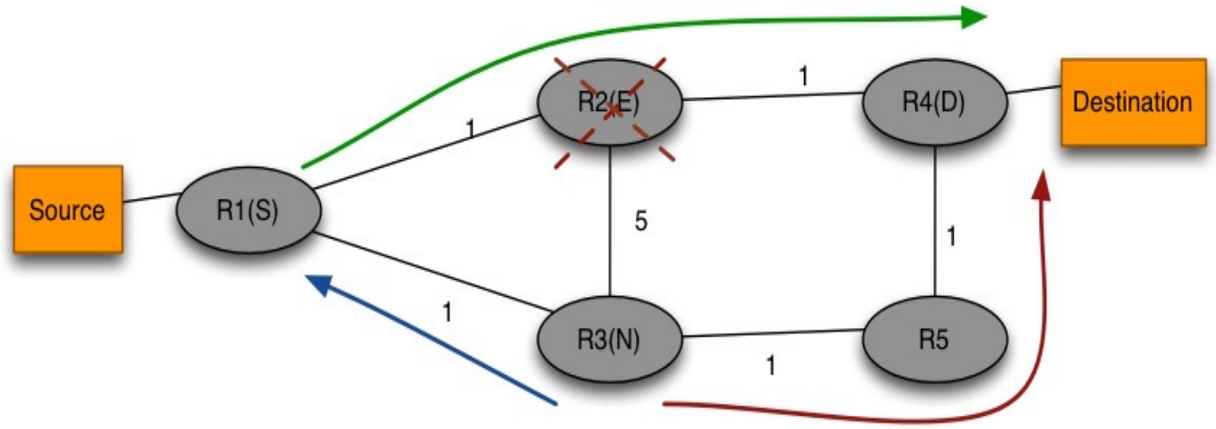
- D is the destination under consideration.




https://t.me/learningnets

# LFA – Loop Free Alternate

- **<u>LFA Inequality Node Protection</u>**




- D(N,D) < D(N,E) + D(E,D)

- “N’s path to D must not go through E”

https://t.me/learningnets

# LFA – Loop Free Alternate

## Per-Link and Per-Prefix LFA:

- **Per-Link:** Per link LFA just looks at the primary link and based on inequality condition#1 it determines if there is a backup path to that link.

- This backup path will carry all the destinations

- Per link LFA is less CPU intensive compared to the Per Prefix as rSPF (Reverse SPF) runs per neighbor

https://t.me/learningnets

# LFA – Loop Free Alternate

## Per-Link and Per-Prefix LFA:

- **Per-Prefix:** Per Prefix determines the backup path for each prefix which means you can have different backup paths for different prefixes

- In general coverage of Per-Prefix will be more than Per-Link but it’s also more CPU intensive

https://t.me/learningnets

# Remote LFA

- Simple LFA doesn’t always provide full coverage and it is topology dependent

- If topology is highly meshed, chance of finding LFA increases

https://t.me/learningnets

# Remote LFA

- With rLFA, idea is  we find a router which is more than one hop away from the calculating router (PLR), from which traffic will be forwarded to the destination without traversing the failed link and somehow we tunnel the packet to that router

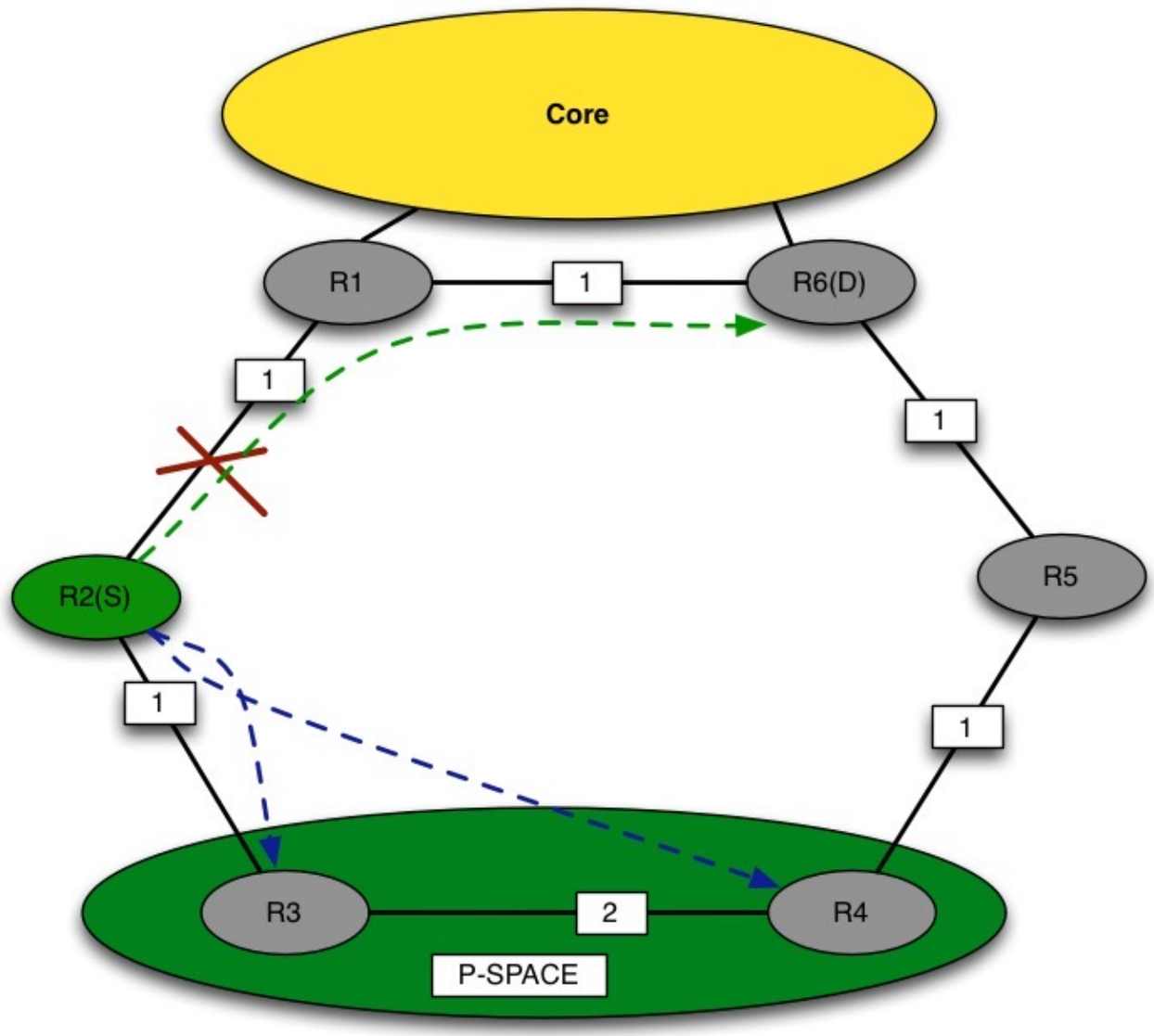
- Let’s understand the P Space, Q Space , Extended P-Space and PQ Node terminologies

https://t.me/learningnets

# Remote LFA – What is P-Space?

- P-Space is set of routers that R2 (S) can reach without using the R2 – R1 link

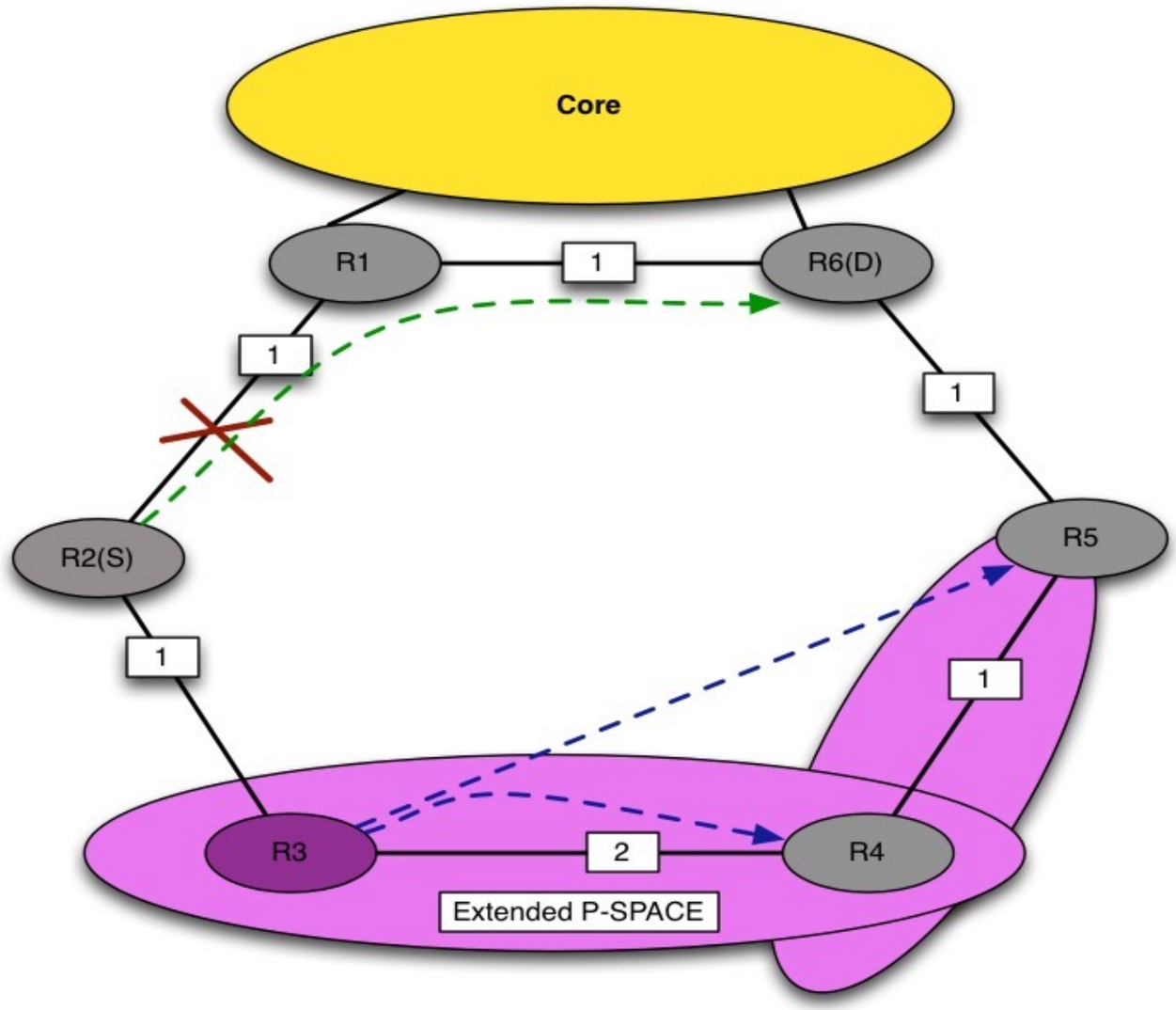
- So R3 and R4 are the P-Space Nodes




https://t.me/learningnets

# Remote LFA – Extended P-Space

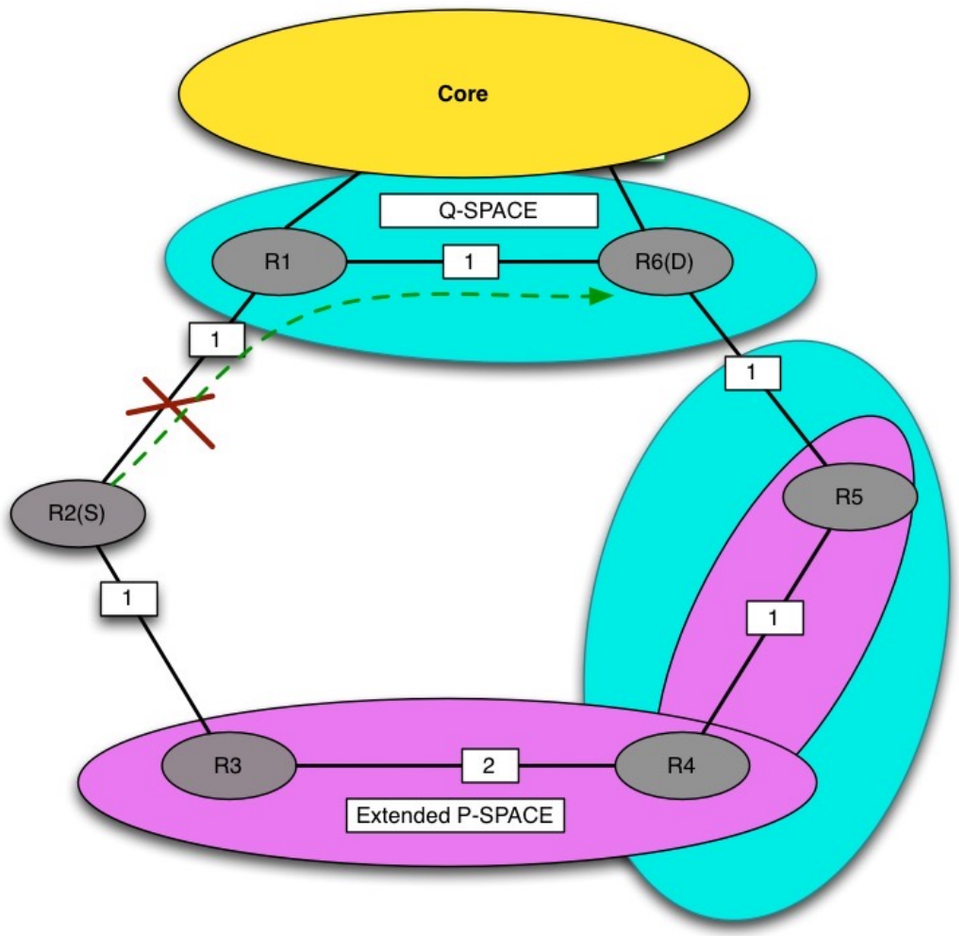
- Extended P-Space “contains the routers that are R2 (S)’s direct neighbor, i.e. R3 can reach without using the R2 (S) -R1 link which is R4 and R5 node




https://t.me/learningnets

# Remote LFA – Q-Space and PQ Nodes

- Q-Space “contains the routers that normally reach R6 (D) without using the R2 (S) -R1 link which  is R1, R5 and R4 nodes




A router that is in both Extended P-Space and Q-Space is a PQ-node

https://t.me/learningnets

# Remote LFA - Summary

- Compare to regular LFA, Remote LFA increases FRR coverage

- It still doesn’t provide 100% FRR coverage in every topology

- It requires targeted LDP session to PQ Node

https://t.me/learningnets

# TI-LFA (Topology Independent LFA)

- If the requirement is 100% Coverage

- Post optimal covergence path

- Less protocol and eliminating targeted LDP sessions

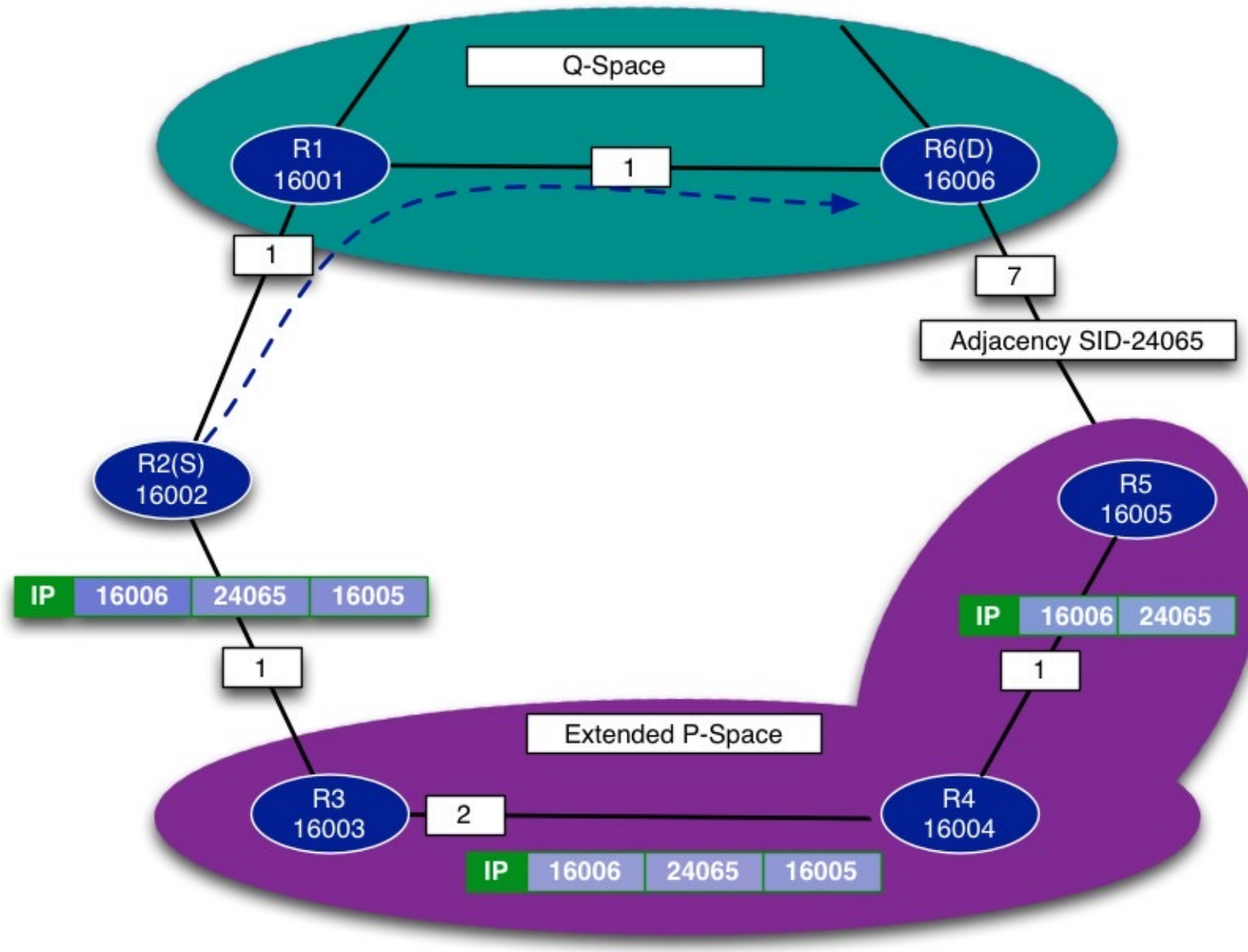
Then TI-LFA is the only choice

https://t.me/learningnets

# TI-LFA (Topology Independent LFA)

- In this topology, there is no overlapping router between Extended P-Space and Q Space, so there is no PQ node.

- Remote LFA doesn't work




https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- Quite often it is desirable to be able to route traffic over a specified path that is different than the default highest bandwidth or shortest path through the network, which is where traditional resource reservation protocol (RSVP) Traffic Engineering was intended to be deployed

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

## RSVP-TE Disadvantages:

- RSVP-TE still has a very low adoption rate, due to its difficult implementation

- It requires too much manual configuration

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

## RSVP-TE Disadvantages:

- It is complex to maintain with its elaborate tunnel configurations

- It relies on soft-state signaling between endpoints, which reduces the ability to scale

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

## RSVP-TE

- • Little deployment and many issues

- • Not scalable
  – Core states in On2 – No inter-domain

- • Complex configuration – Tunnel interfaces

- • Complex steering – PBR, autoroute

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- Simple, Automated and Scalable
  – No core state: state in the packet header

- – No tunnel interface: “SR Policy”
  – No head-end a-priori configuration: on-demand policy instantiation

  – No head-end a-priori steering: automated steering

- • Multi-Domain
  – SR PCE for compute
  – Binding-SID (BSID) for scale

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- SRTE uses a “Policy” (SID-list) to steer traffic through the network

- Policy path can be computed by router or controller

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- SID list pushed by head-end

- Rest of network executes instructions embedded in the SID list

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- SR brings simplification over LDP is pretty straight forward as:

1. You have one less protocol to run for label distribution. Labels are advertised with IGP

2. SR bring the global label concept which makes troubleshooting easier

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)
# How it works?

- Similar to RSVP-TE, in SR-TE, a Head-end needs to first build a TED, run a CSPF to find a path from Point A to B, Once the path is found then stack the labels accordingly on the packets

- The difference is that there is no need to signal the path by the head-end towards the tail-end as we know that the state is in the packet (labels) which means Midpoints and Tail-end doesn’t know about it (hence no state)

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- How Traffic Engineering Database is built with SR?
- Similar to RSVP based TE, In a SR domain, SR-Nodes flood their various attribute in IGP which allows a Head-End to build its TED. Some of the attributes which are specifically relevant in SR are:
- Prefix-SID
- Adjacency SID’s
- SRGB
- Index value
- Things like Link affinities (aka : Color / SRLG’s)
- Node capabilities like max label stack supported by a Node (SR Data plane Capability), SR Algorithm support.

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- A Head-end node will collect all the above information from IGP (IS-IS or OSPF) to build its TED

- Note: In a Centralized Controller based model, a Controller will learn all this info through BGP-LS to build its TED

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

## CSPF- Constraint based SPF

- Once the TED is built, A Head-end node can run its CSPF to find a path from Point A to B. Constraints could be like

- The lowest latency path from Point A to B

    - Similar to RSVP-TE, we can have two metrics. One representing link bandwidth usually by IGP cost and TE metric representing latency

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

## CSPF- Constraint based SPF

- Avoiding or including certain links. (Link affinities).
- Bandwidth cannot be reserved as a constraint with Distributed Traffic Engineering
- Bandwidth as a constraint can be handled by a Central PCE controller

https://t.me/learningnets

# SR-TE (Segment Routing Traffic Engineering)

- In SR, a midpoint node can tell you how much link bandwidth is available (like link-speed as that's a pretty much stationary information)

- But there is no way to keep track of ongoing tunnels + Bandwidth requested for each tunnel without bringing the state at the midpoint which is contrary to SR goal

- Thus, bandwidth cannot be a constraint in a distributed architecture

https://t.me/learningnets

# SR Policy

- An SR Policy is uniquely identified by a tuple
  - (head-end, color, end-point)

- Head-end: where the SR Policy is implemented
- Color: a numerical value to differentiate multiple SRTE Policies between the same pair of nodes
- End-point: the destination of the SR Policy (Tail-end)

https://t.me/learningnets

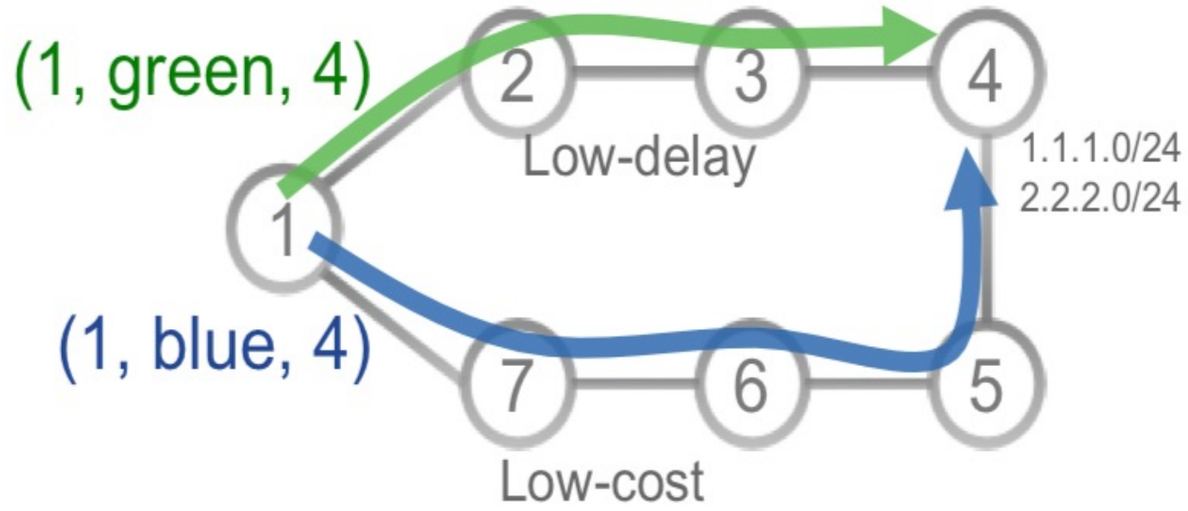
# SR Policy

## Example:

- – Low-cost="blue", Low-delay="green"

- – steer traffic to 1.1.1.0/24 via Node4 into Low-cost SR Policy (1, blue, 4)

- – steer traffic to 2.2.2.0/24 via Node4 into Low-delay SR Policy (1, green, 4)




https://t.me/learningnets

# SR Policy – Dynamic Path

- A dynamic path expresses an optimization objective and a set of constraints

- The head-end computes a solution to the optimization problem as a SID-list or a set of SID-lists

- When the head-end does not have enough topological information (e.g. multi-domain problem), the head-end may delegate the computation to a PCE

https://t.me/learningnets

# SR Policy – Explicit Path

- An explicit path is an explicitly specified SID-list or set of SID- lists

https://t.me/learningnets

# SR Policy

## SR Policy – configuration example

On Node1:

```
segment-routing
 traffic-eng
  policy POLICY1
   color 20 end-point ipv4 1.1.1.4
   binding-sid mpls 1000
   candidate-paths
1   preference 100
     dynamic
      metric type te
      constraints
       affinity
        exclude-any color red
    !
2   preference 200
     explicit segment-list SIDLIST1
  !
segment-list name SIDLIST1
 index 10 mpls label 16002
 index 20 mpls label 30203
 index 30 mpls label 16004
```

Path preference 100

Dynamic path

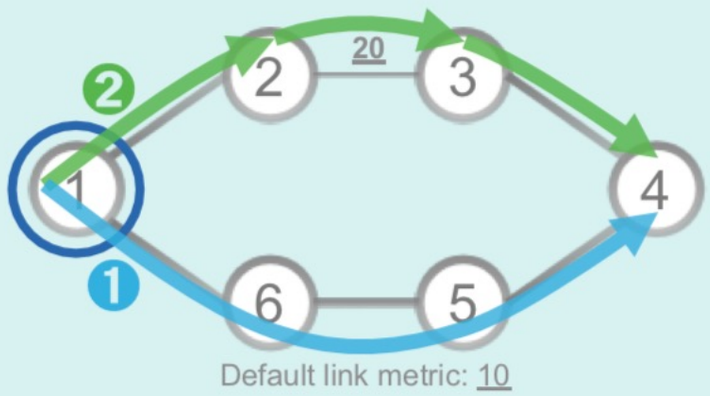
Opt. Obj.: TE metric

Constraint

Path preference 200

Explicit SID-list1

SID-list1




Default link metric: 10

```
segment-routing
 traffic-eng
  affinity-map
   color red bit-position 0
```

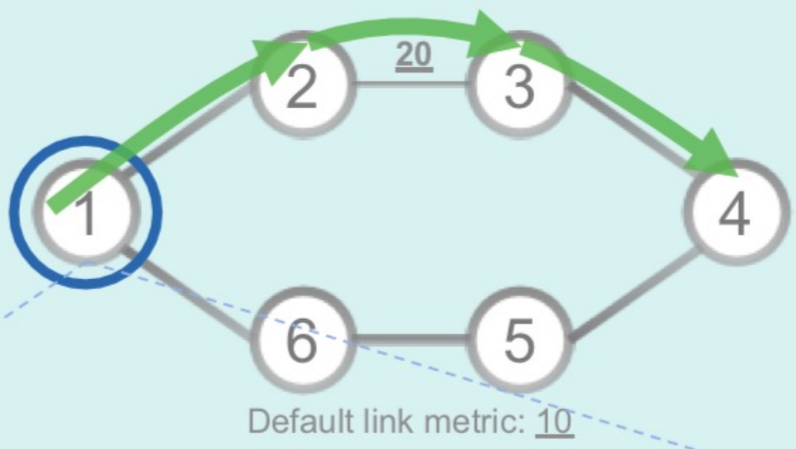
https://t.me/learningnets

# SR Policy

On Node1:

```
segment-routing
 traffic-eng
  policy POLICY1
   color 20 end-point ipv4 1.1.1.4
   binding-sid mpls 1000
   candidate-paths
    preference 100
     dynamic
      metric type te
     constraints
      affinity
       exclude-any red
    !
    preference 200
     explicit segment-list SIDLIST1
!
segment-list name SIDLIST1
 index 10 mpls label 16002
 index 20 mpls label 30203
 index 30 mpls label 16004
```

Selected Path:
- Valid Path
- Highest Pref value




**FIB @ head-end Node1**
Incoming label: **1000**
Action: pop and push <**16002**, **30203**, **16004**>

https://t.me/learningnets

# ODN – On-Demand Next Hop

- With ODN, We don’t create SR Policy, we create template and when there is a match, policy is created automatically, not manually

- A service head-end automatically instantiates an SR Policy to a BGP next-hop when required (on-demand)

https://t.me/learningnets

# ODN – On-Demand Next Hop

**On-demand SR Policy**

- • Configure an SR Policy template for each color for which on- demand SR Policy instantiation is desired

```
segment-routing
 traffic-eng
  on-demand color 10
   dynamic
    metric type igp
  !
  on-demand color 20
   dynamic
    metric type delay
```

SR Policy template High-BW (color 10)

SR Policy template low-delay (color 20)

- • An example with two color templates configured: – color 10 for high bandwidth (optimize IGP metric) – color 20 for low-delay (optimize link-delay metric)

•If an SR Policy template exists for color C, then a route with color C can trigger an on-demand SR Policy

https://t.me/learningnets

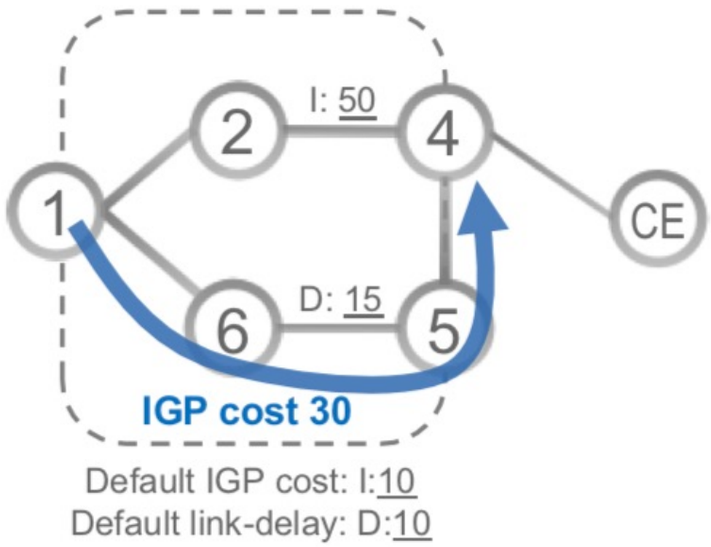
# Automated Steering

- Service traffic is automatically steered on the right SLA path – Steered into an SR Policy based on color and next-hop of the service route

- Independent from ODN: SR Policy can already exist or be instantiated on-demand (ODN) when receiving the service route update

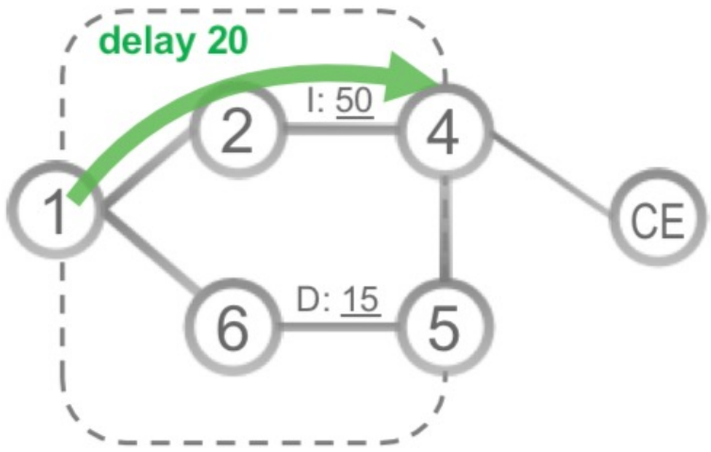
https://t.me/learningnets

# On-Demand SR- Policy Work-Flow Example




Default IGP cost: I:10
Default link-delay: D:10

**Basic VPN should use lowest cost underlay path**




Default IGP cost: I:10
Default link-delay: D:10

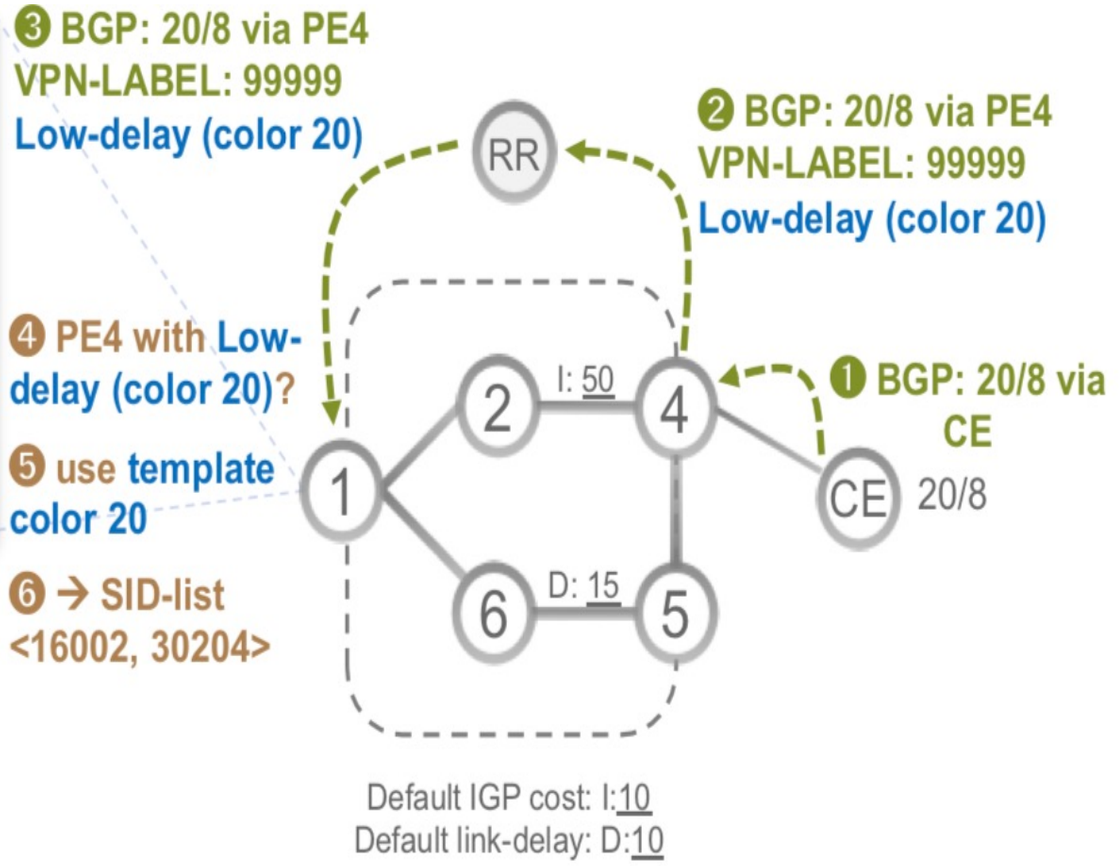
**Premium VPN should use lowest delay path**

https://t.me/learningnets

# On-Demand SR- Policy Work-Flow Example

```
router bgp 1
 neighbor 1.1.1.10
  address-family vpnv4 unicast
!
segment-routing
 traffic-eng
  on-demand color 20
   dynamic
    metric
     type delay
```

SR Policy template
Low-delay (color 20)




no route-policy required!

https://t.me/learningnets

# Binding SID (B-SID)

- The Binding-SID is fundamental to SR, it provides scaling, network opacity and service independence

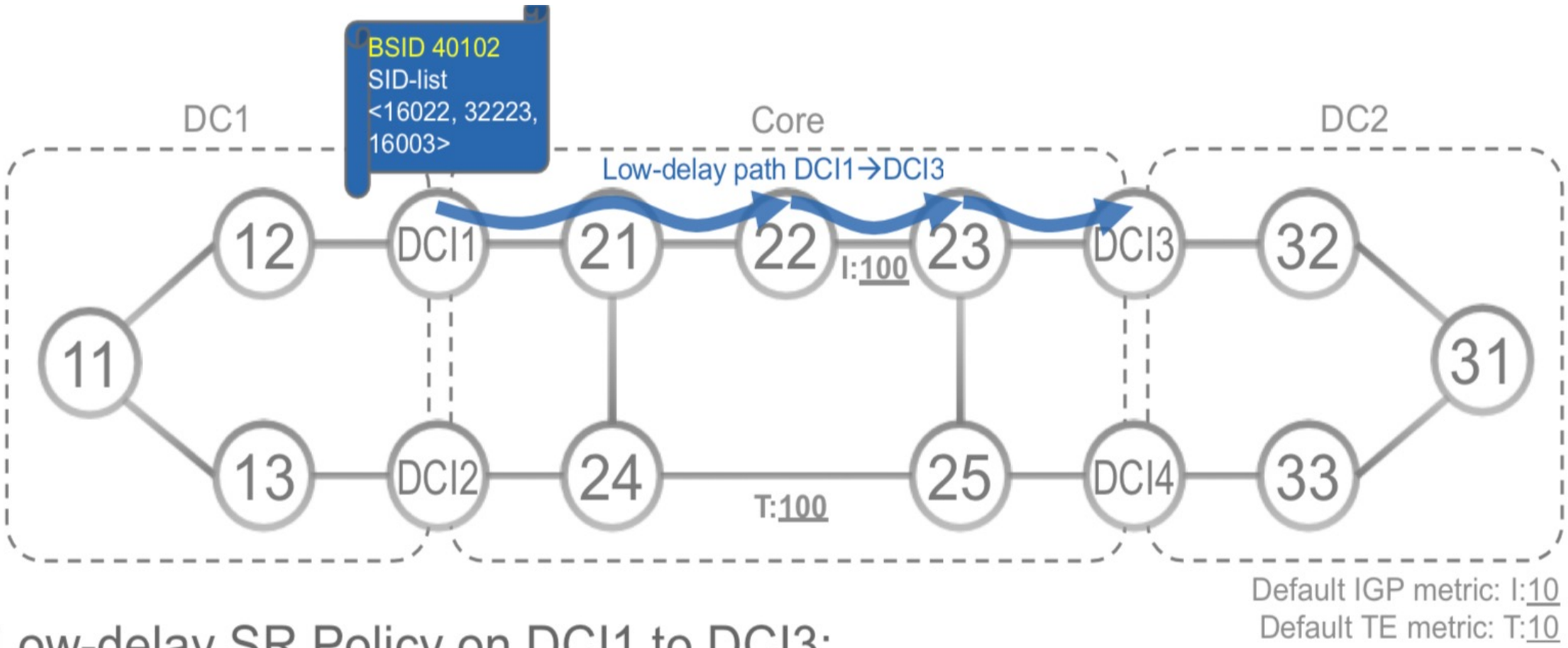
- Use of BSID decreases the number of segments imposed by the source

- BSID acts as a stable anchor point that isolates one domain from the churn of another domain

- BSID provides opacity and independence between domains

https://t.me/learningnets

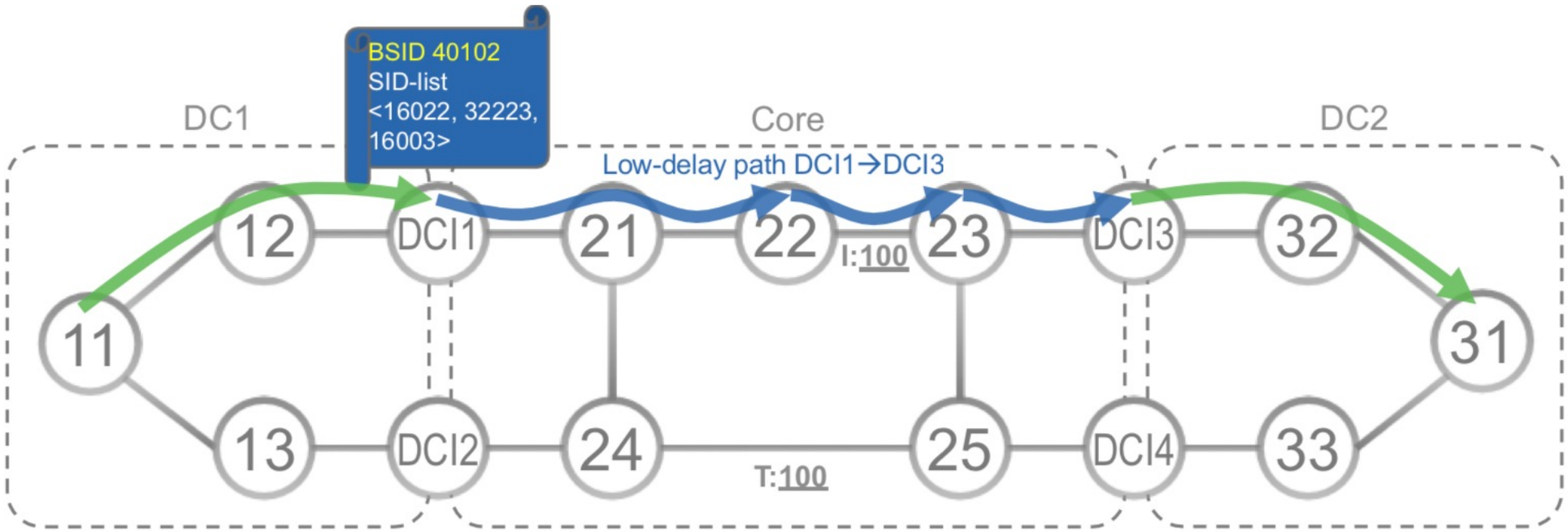
# Binding-SID Illustration




- Low-delay SR Policy on DCI1 to DCI3:
  - BSID: 40102
  - SID-list <16022, 32223, 16003>

https://t.me/learningnets

# BSID - Reduced SID-List Size

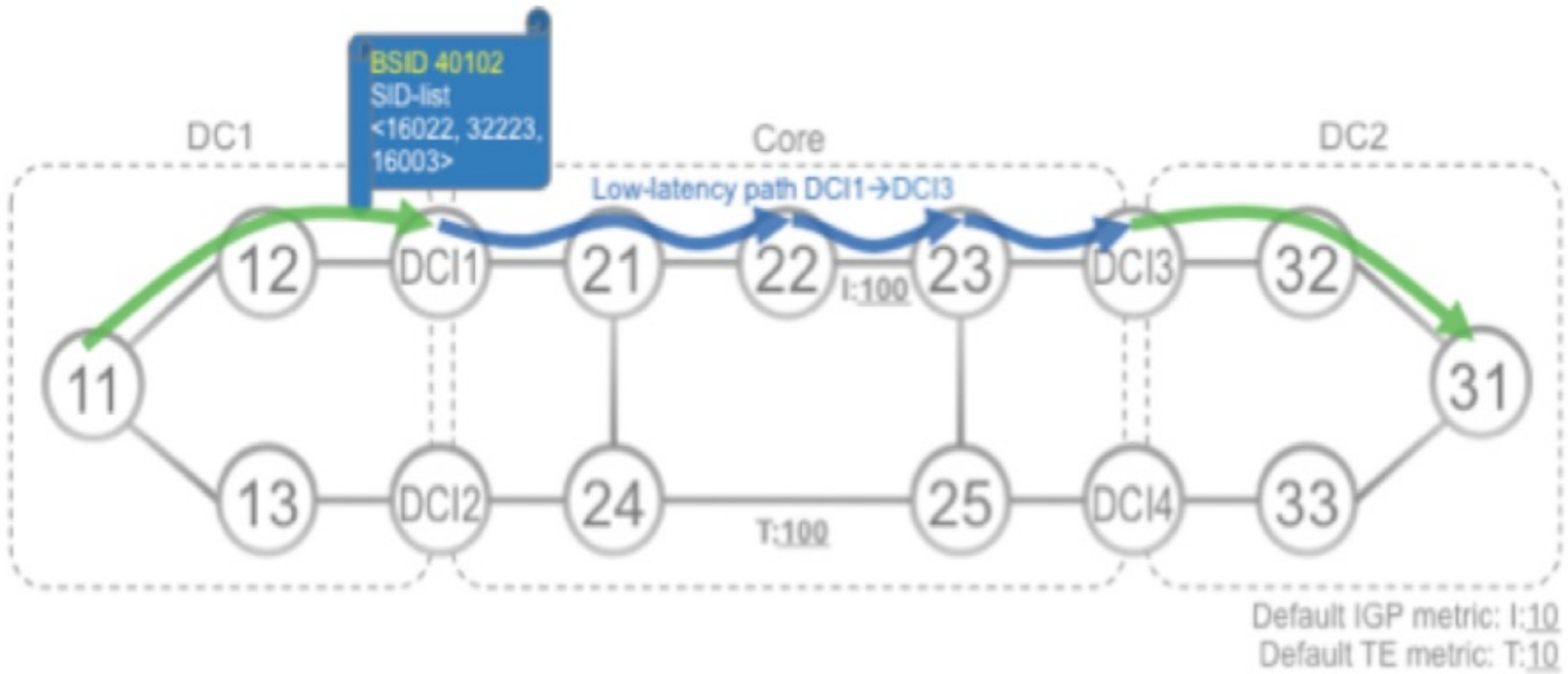



- Low-delay SR Policy from Node11 to Node31:
  - Without intermediate core SR Policy: <16001, 16022, 32223, 16003, 16031>
  - With intermediate core SR Policy: <16001, 40102, 16031>

https://t.me/learningnets

# BSID - Stable Anchor Point

- When the Core domain's topology changes, the BSID of the intermediate SR Policy on DCI1 does not change

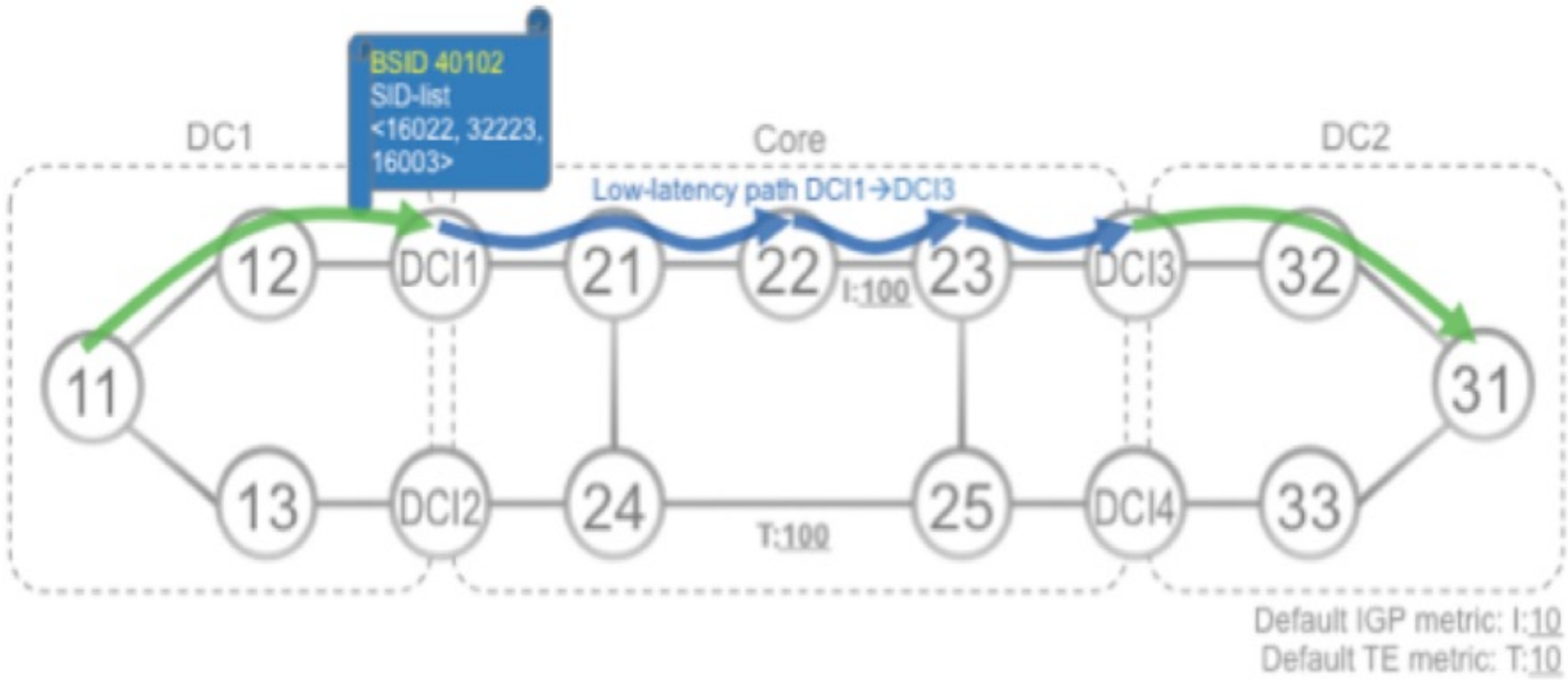



- SR Policy on Node11 does not change and Node11 is shielded from the churn in domain DC1

https://t.me/learningnets

# BSID - Opacity and Independence

- The administrative authority of the Core domain does not want to share information about its topology
- B-SID keeps network and service opaque
- Node11 does not know the details of how the Core domain provides the low-delay service




https://t.me/learningnets

# EVPN

- EVPN is considered as next generation VPLS, but it is much more than that, as you will see in this lesson

- 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

https://t.me/learningnets

# EVPN

- In VPLS, 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

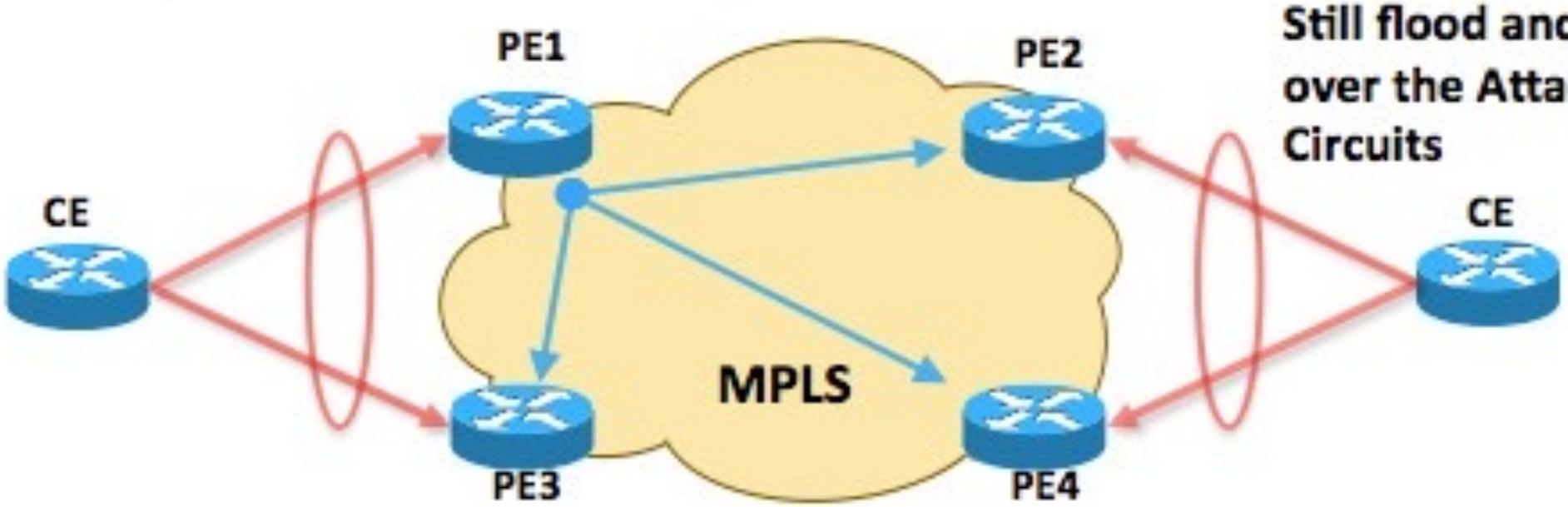
- EVPN can support active active flow -based load balancing so same Vlan can be used on both PE device actively. This provides faster convergence in customer link, PE link or node failure scenarios

https://t.me/learningnets

# EVPN

Data Plane address learning from Access.

Control Plane address advertisement and learning over Core.




Still flood and learn over the Attachment Circuits

BGP MAC adv.
Route EVPN NLRI
MAC M1 via PE1.

https://t.me/learningnets

# EVPN

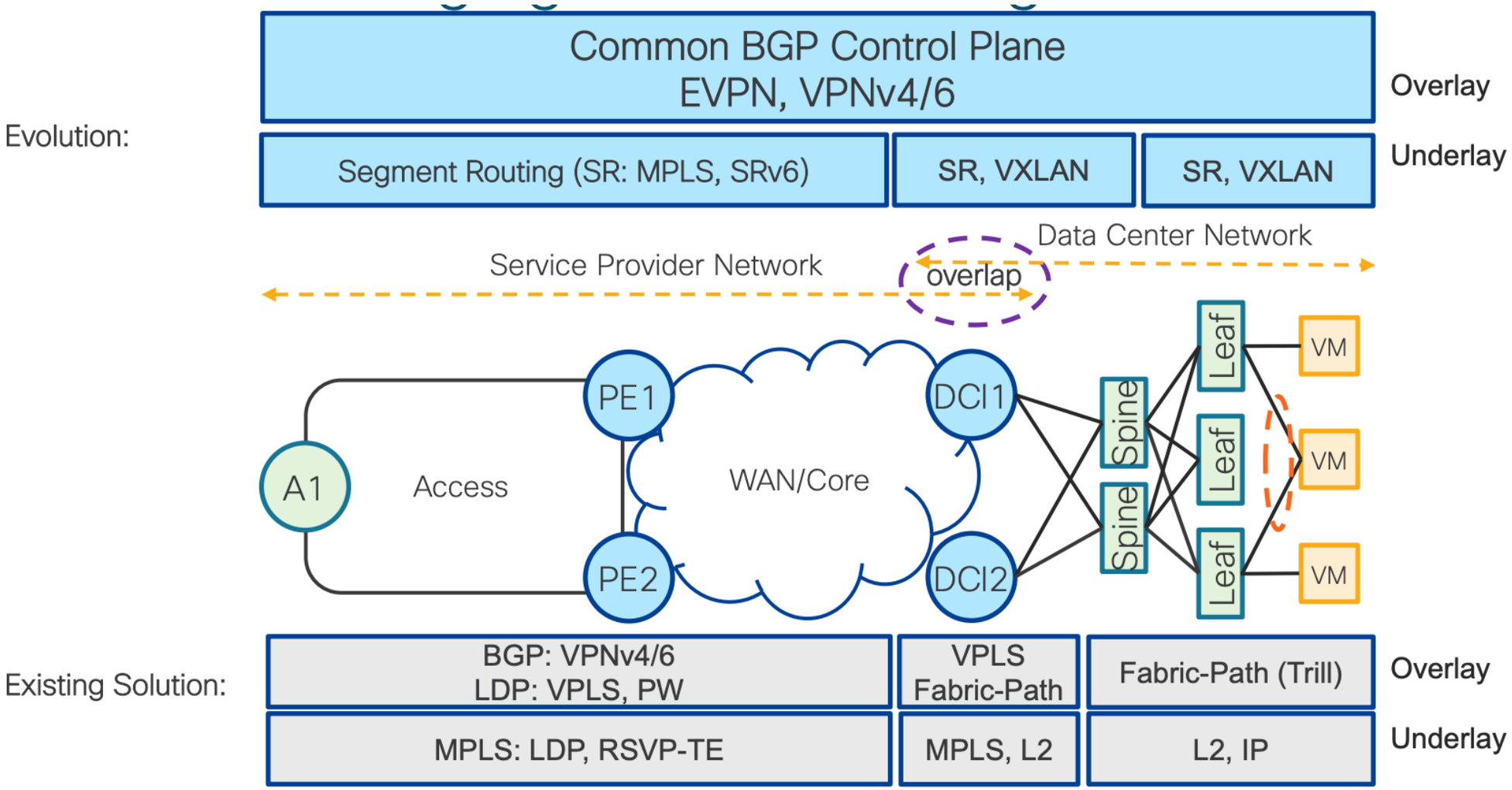
- Customer MAC addresses are advertised over the MP-BGP (Multiprotocol BGP) control plane between PEs

- There is no data plane MAC learning over the core network in EVPN (VPLS it would be Dataplane learning)

- But Customer MAC addresses from the attachment circuit is still learned through the data plane

https://t.me/learningnets

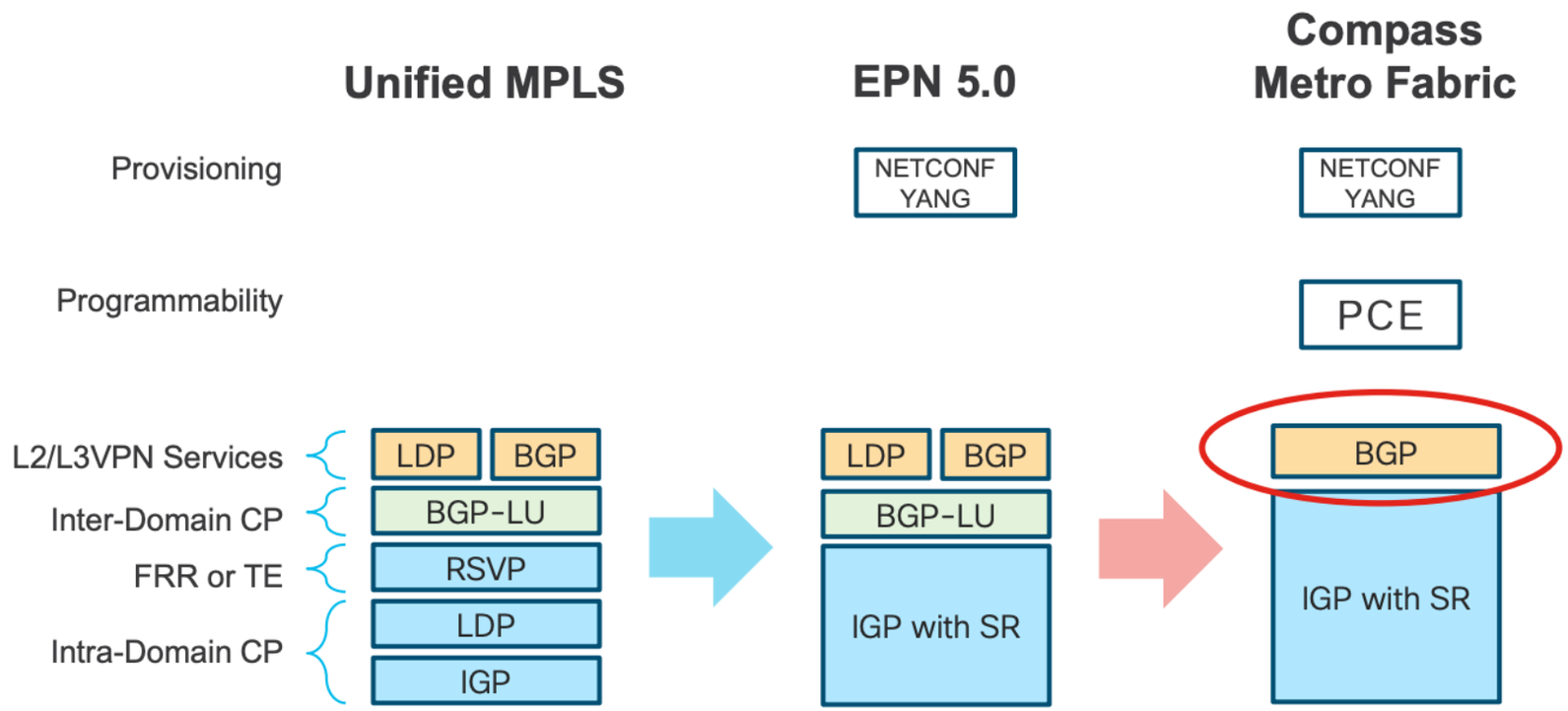
# EVPN




https://t.me/learningnets

# EVPN as a single Service Level Protocol

## Service Provider Network – Simplification Journey

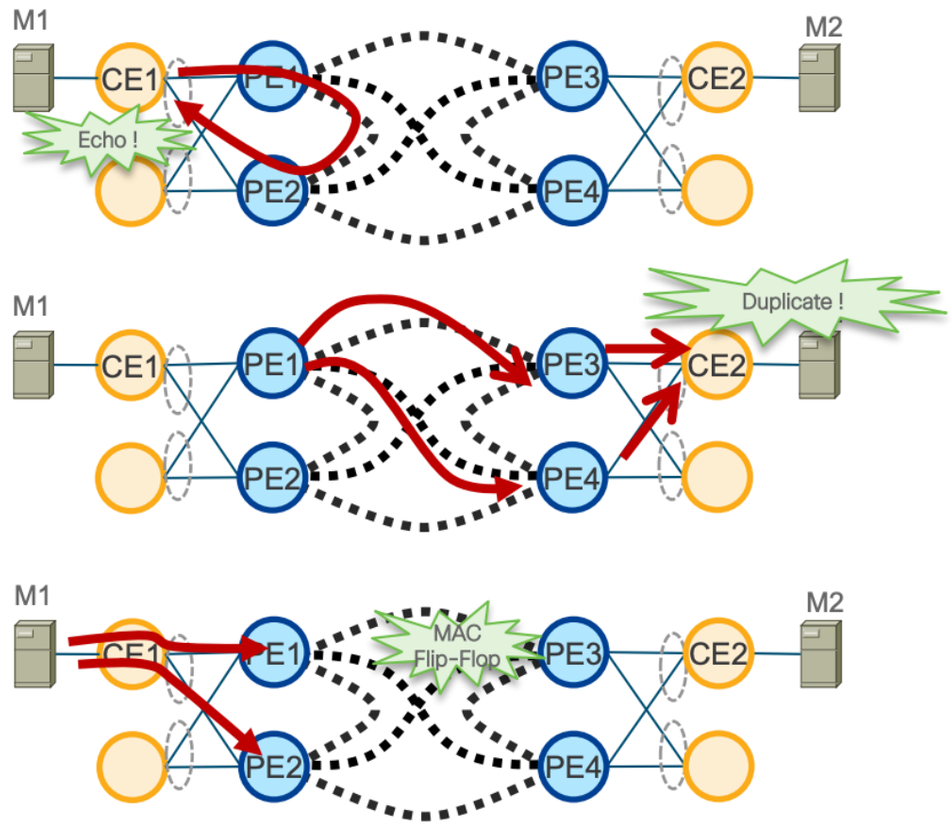



https://t.me/learningnets

# EVPN advantages compare to MP2MP VPLS

## Next-Generation Solutions for L2VPN
### Solving VPLS challenges for per-flow Redundancy

- Existing VPLS solutions do not offer an All-Active per-flow redundancy
- Looping of Traffic Flooded from PE
- Duplicate Frames from Floods from the Core
- MAC Flip-Flopping over Pseudowire
  - E.g. Port-Channel Load-Balancing does not produce a consistent hash-value for a frame with the same source MAC (e.g. non MAC based Hash-Schemes)




https://t.me/learningnets

# EVPN Advantages

## EVPN Advantages:

**Integrated Services**

- Integrated Layer 2 and Layer 3 VPN services
- L3VPN-like principles and operational experience for scalability and control

**Network Efficiency**

- All-active Multi-homing & PE load-balancing (ECMP)
- Fast convergence (link, node, MAC moves)
- Control-Place (BGP) learning. PWs are no longer used.
- Optimized Broadcast, Unknown-unicast, Multicast traffic delivery

**Service Flexibility**

- Choice of MPLS, VxLAN or SRv6 data plane encapsulation
- Support existing and new services types (E-LAN, E-Line, E-TREE)
- Peer PE auto-discovery. Redundancy group auto-sensing
- Fully support IPv4 and IPv6 in the data plane and control plane

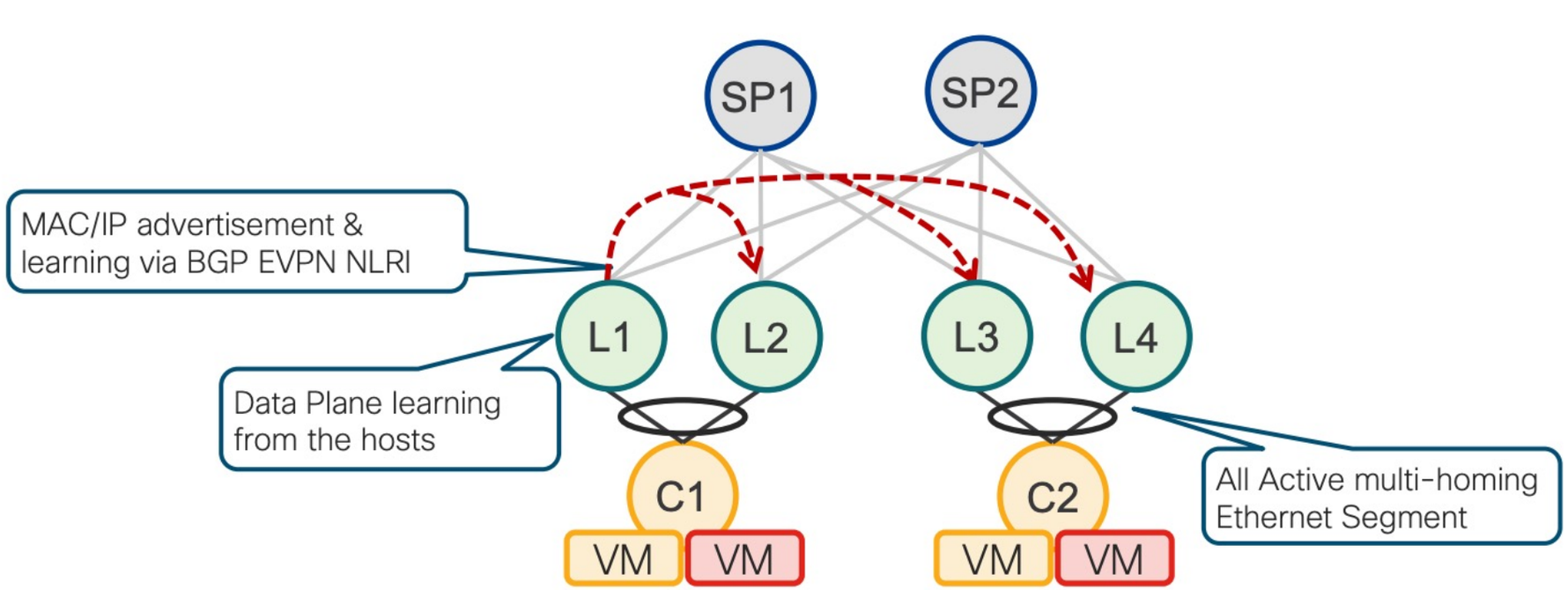
**Investment Protection**

- Open-Standard and Multi-vendor support

https://t.me/learningnets

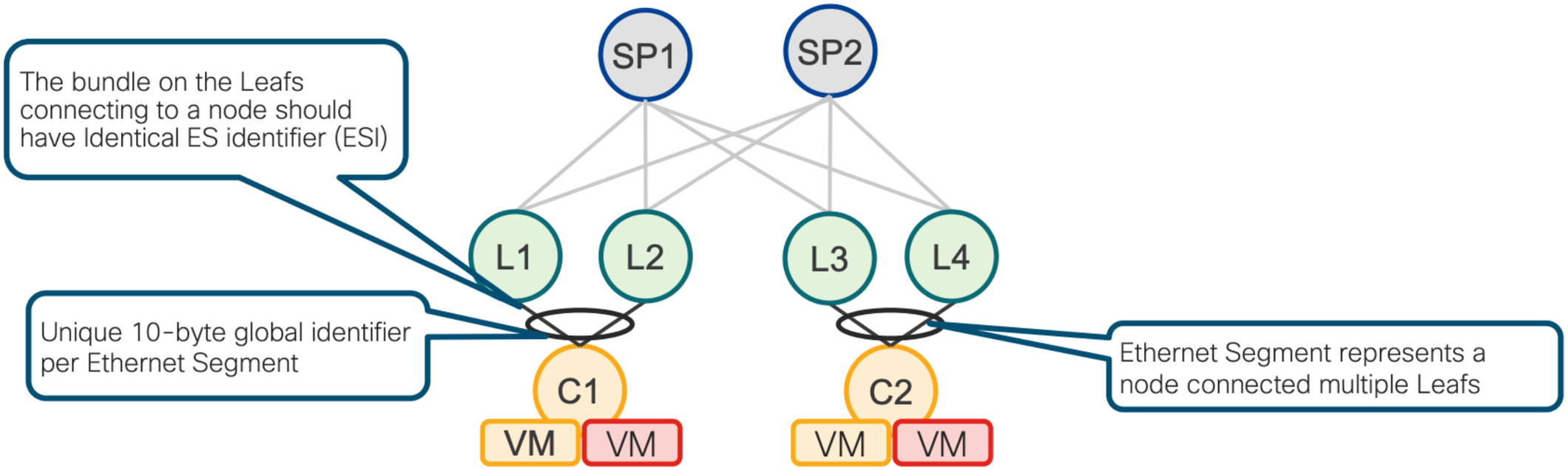
# EVPN Host Learning

- Leafs run Multi-Protocol BGP to advertise & learn MAC/IP addresses over the Network Fabric
- MAC/IP addresses are advertised to rest of Leafs




https://t.me/learningnets

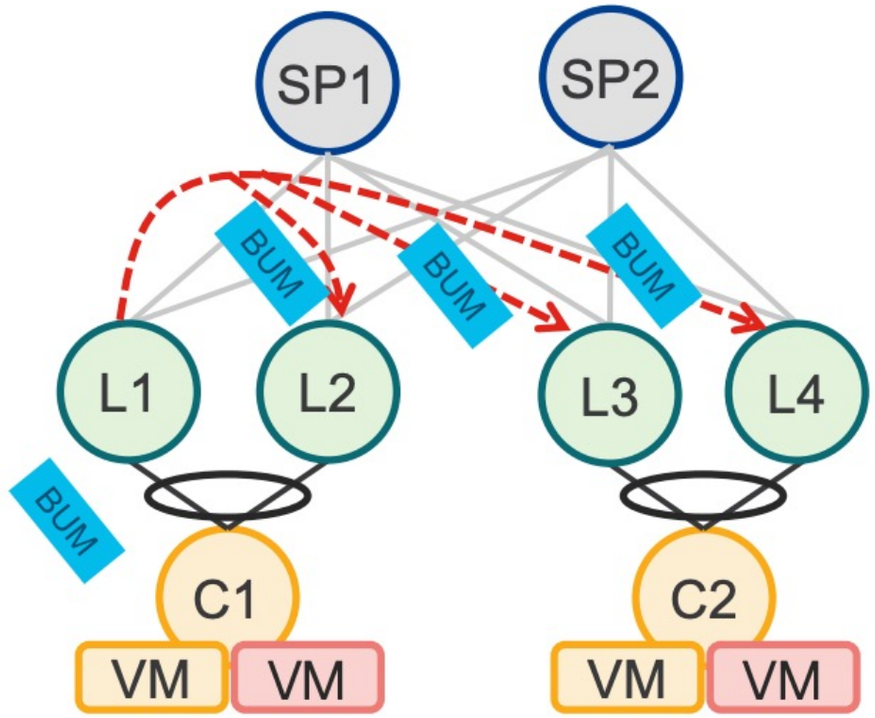
# Ethernet Segment for Multihoming




https://t.me/learningnets

# EVPN BUM Handling (EVPN Route Type 3- Inclusive Multicast Route)

- EVPN works with IR – Ingress Replication or Native Multicast




- BUM traffic is flooded over the fabric and flood list is created with EVPN Type 3 route

Inclusive Multicast Route 0x3 • Usage: • Multicast tunnels used to transport Broadcast, Multicast and Unknown Unicast frames (BUM)

https://t.me/learningnets

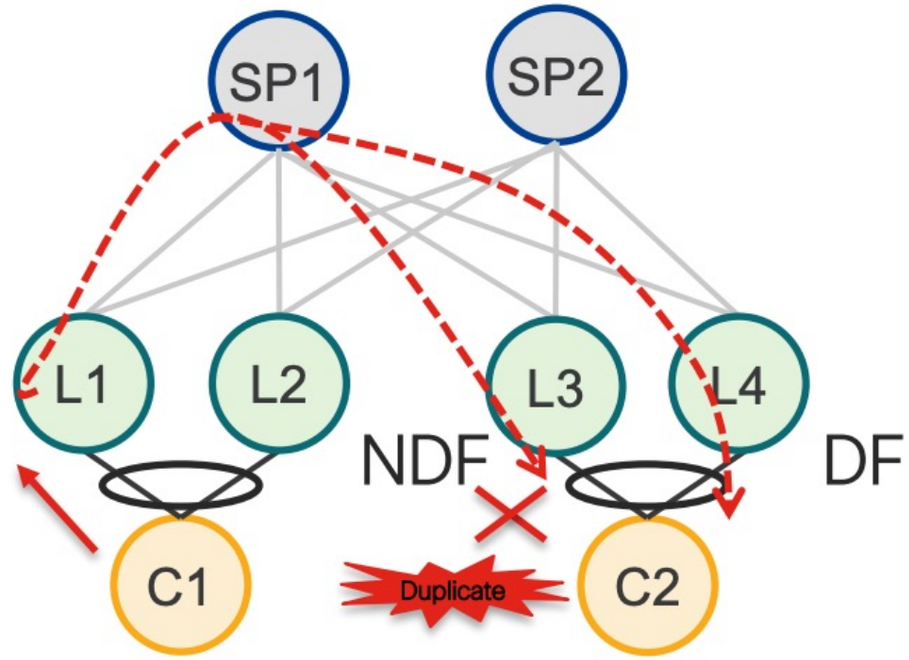
# EVPN Designated Forwarder Election (EVPN Type-4 Ethernet Segment Route)

- BUM Traffic cause duplicate packets at the Multihomed sites

EVPN Type 4 – Ethernet Segment Route is used for:

1. Auto-discovery of multi-homed Ethernet Segments

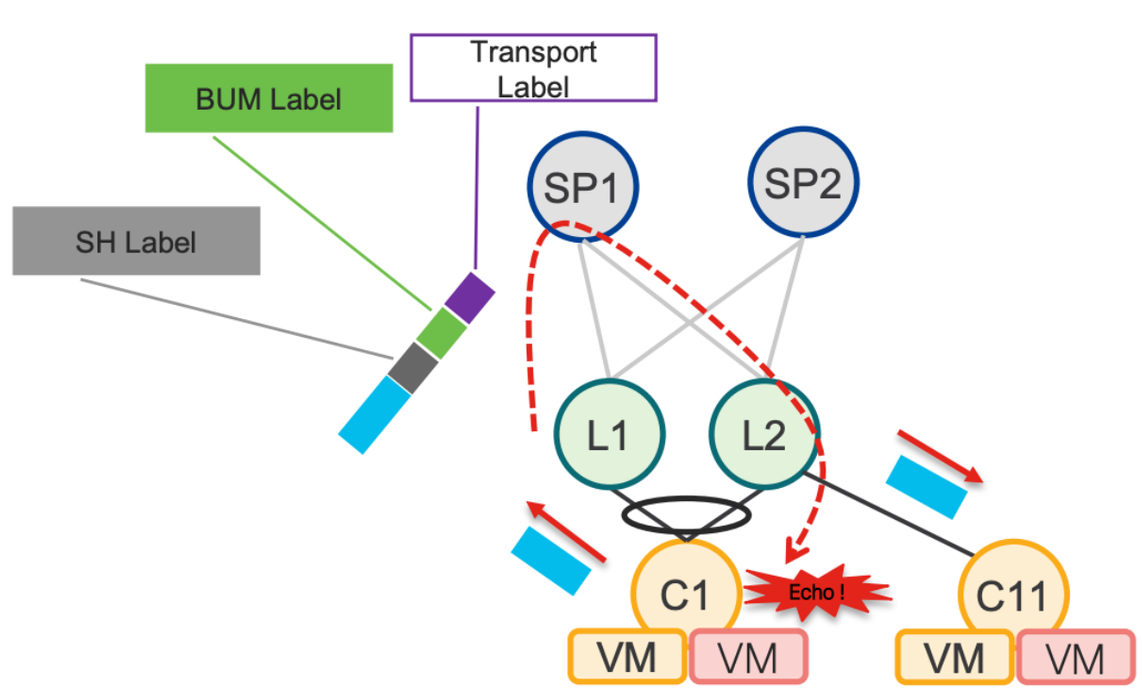
2. Designated Forwarder election




https://t.me/learningnets

# EVPN Split Horizon (EVPN Type—1 Ethernet Auto-Discovery Route)

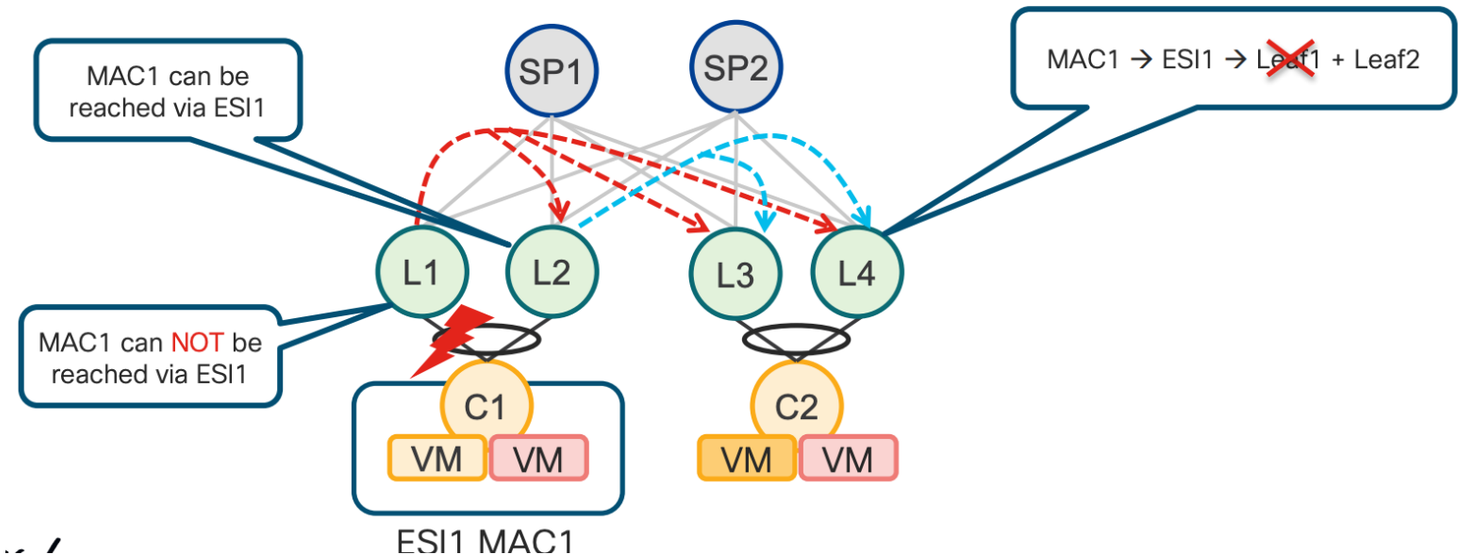
- How to prevent flooded traffic from echoing back to a multi-homed Ethernet Segment?




https://t.me/learningnets

# EVPN Mac Mass-Withdraw (Type-1 Ethernet AD Route)

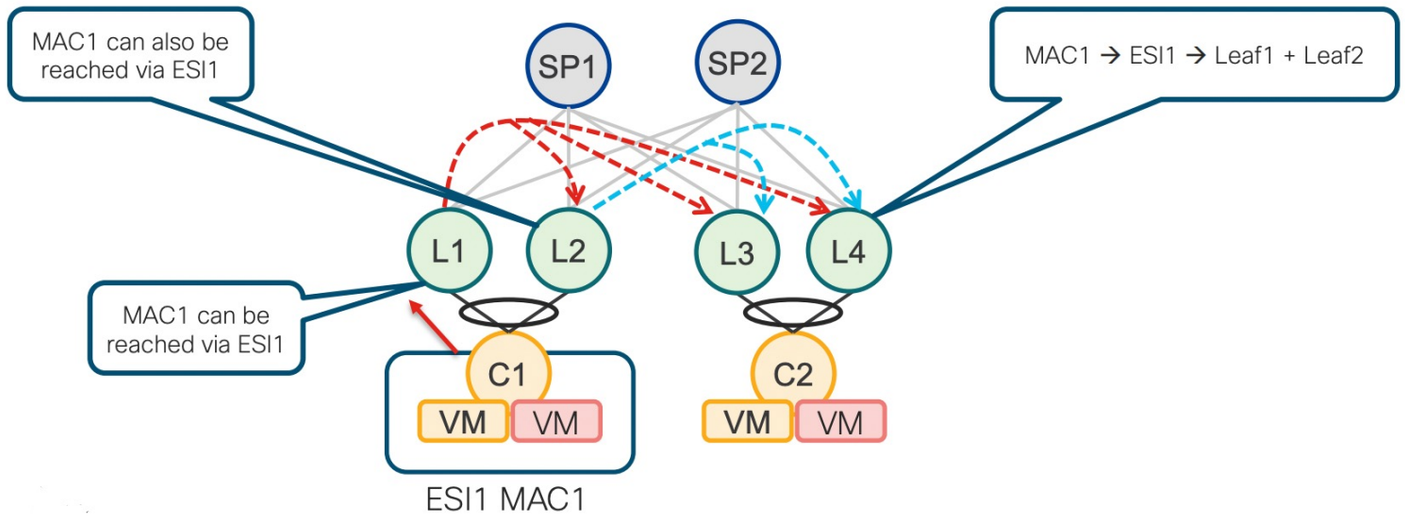
- How to inform other Leafs of a failure affecting many MAC addresses quickly while the control-plane re-converges?




https://t.me/learningnets

# EVPN Aliasing (Type-1 Ethernet AD Route)

- How to load-balance traffic towards a multi-homed device across multiple Leafs when MAC addresses are learnt by only a single Leaf?




- ESI to Leaf Mapping table is created on each device via EVPN Type-1 Route

https://t.me/learningnets

# EVPN – Type 2 Route (MAC Advertisement Route)

- Most fundamental EVPN Route Type

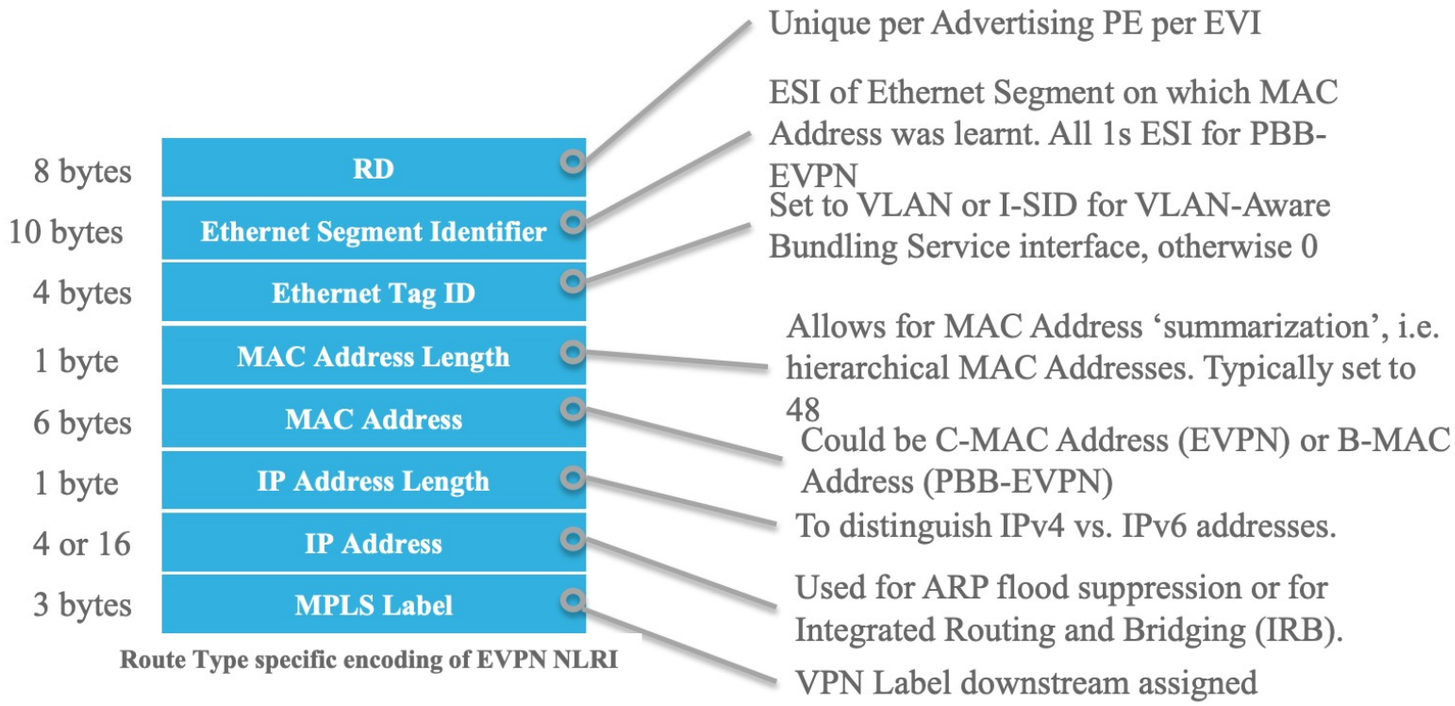
- MAC addresses are advertised with Type 2 routes

- For ARP, MAC + IP, as an optional can be advertised with this route type as well

https://t.me/learningnets

# EVPN – Type 2 Route (MAC Advertisement Route)

- With this route type, only host route (/32 or /128) is advertised



Route Type specific encoding of EVPN NLRI

https://t.me/learningnets

# EVPN – Type 5 Prefix Advertisement Route

- You might think that, IP address information is already sent in EVPN Route Type 2, then why we have Type 5 route?

- Type 2 Routes are just host route, if subnet need to be advertised then Type 5 is necessary

- For example, default route to the extrenal networks, require Type 5 routes

https://t.me/learningnets

# IPv6 Translation Techniques

- Translation techniques allow IPv6 only devices to communicate with IPv4 only devices (NAT64, 464XLAT)

- Also, they are used to extend the life of IPv4 (NAT – CGN/LSN)

https://t.me/learningnets

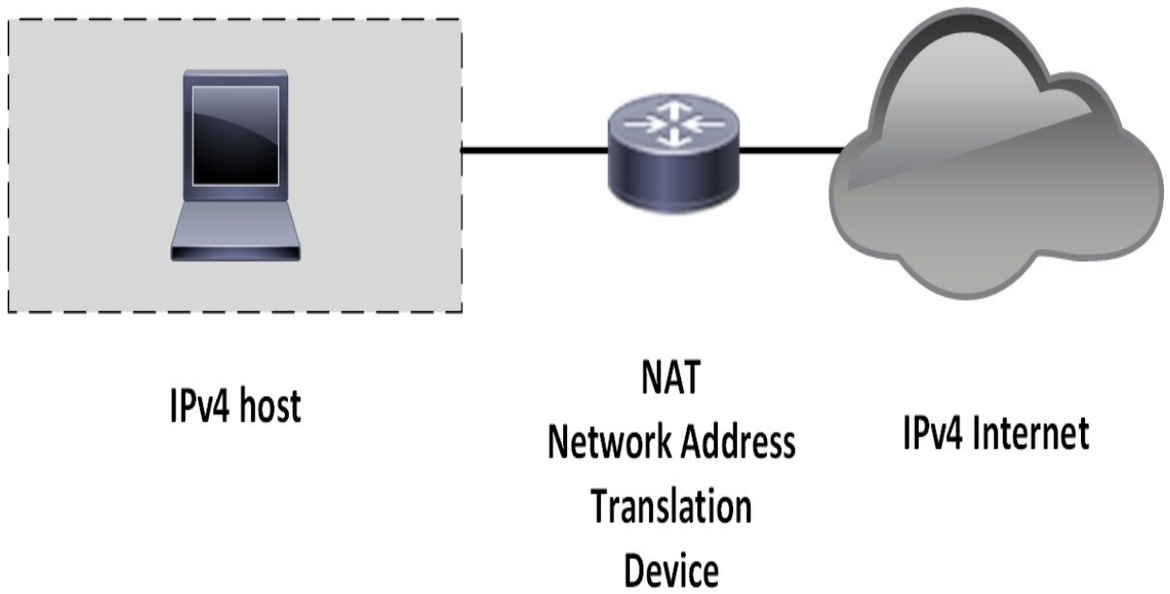
# IPv6 Translation Techniques

- NAT 44
- NAT64 and DNS 64
- CGN (aka LSN)
- 464XLAT

https://t.me/learningnets

# NAT44

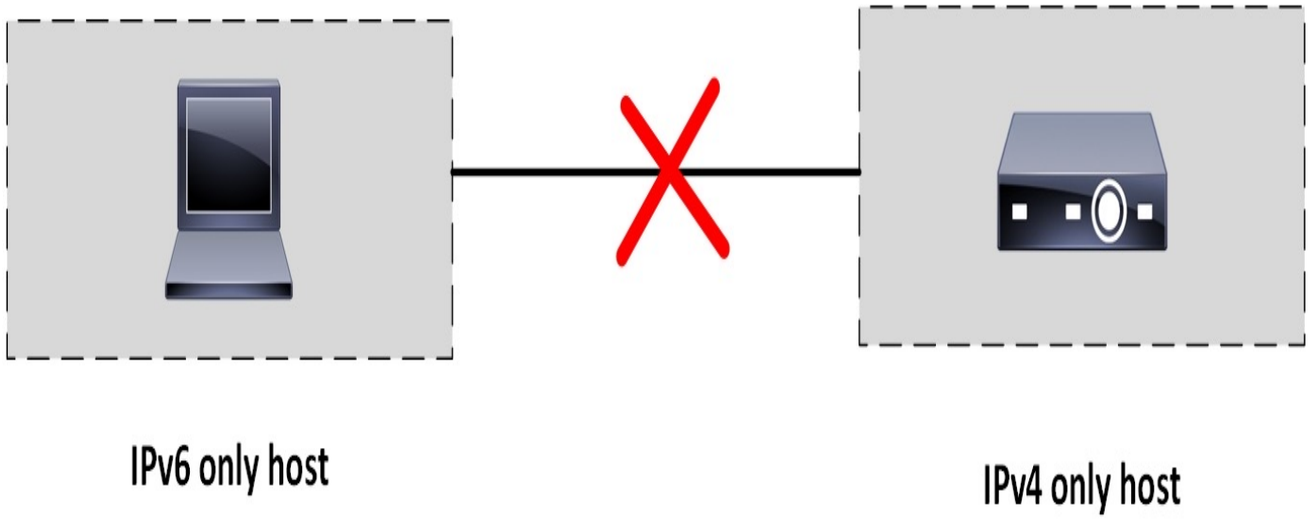
- Regular NAT 44 is an operation to translate Private IPv4 address to a Public IPv4 address




https://t.me/learningnets

# NAT64

- IPv6 only host cannot communicate directly to IPv4 host



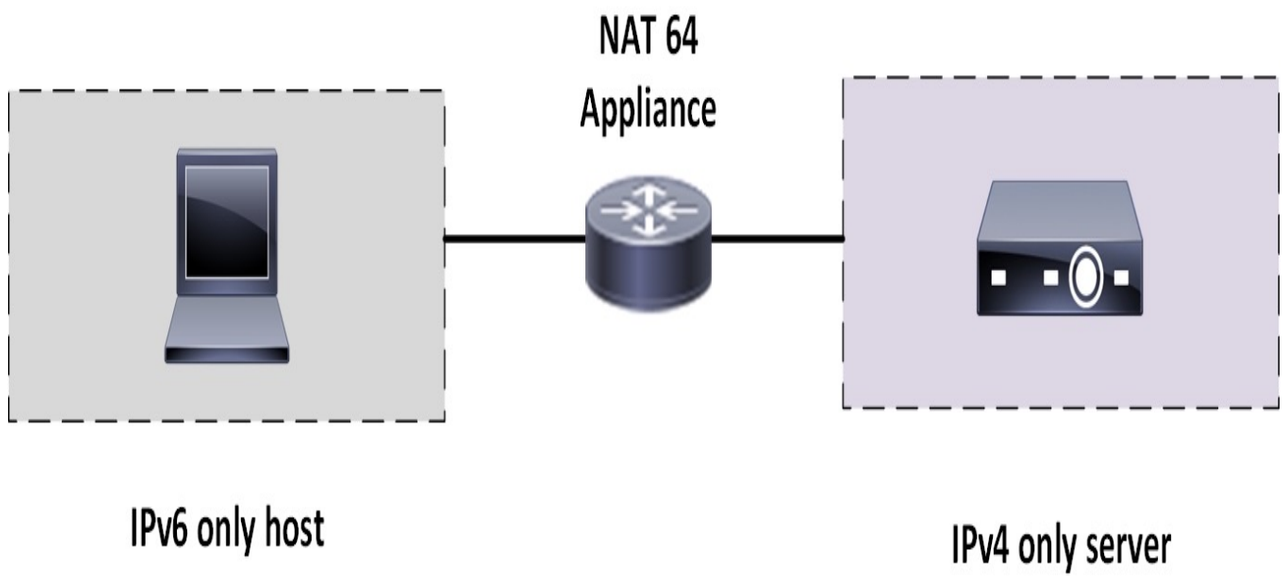

- For that reason, several transition mechanisms developed

- Typical use case for NAT64 is IPv6 only Greenfield networks

https://t.me/learningnets

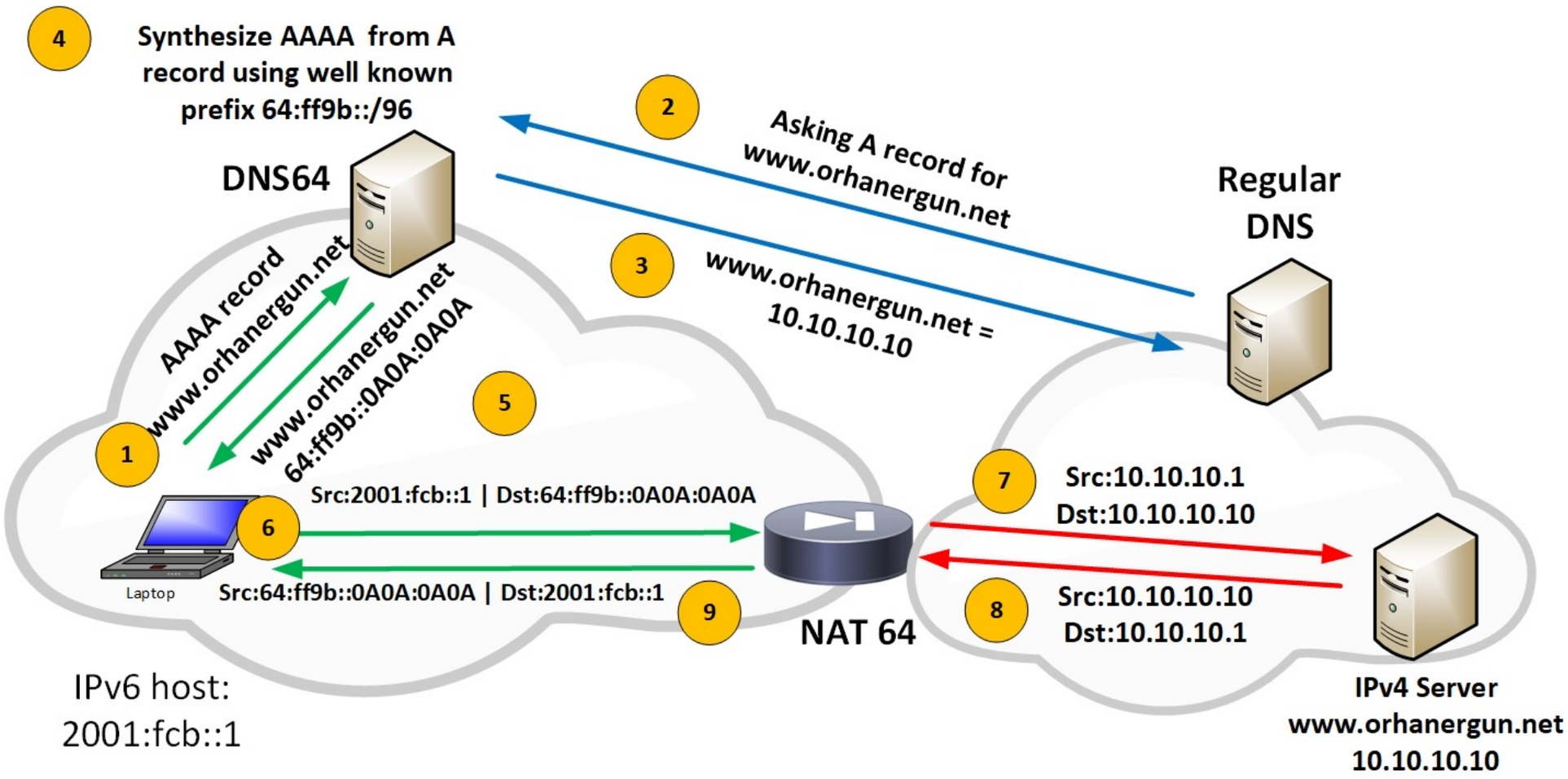
# NAT64

- In the case of NAT64, user doesn't have an IPv4, It only has an IPv6 address
- NAT64 makes possible IPv6 only hosts to talk to IPv4 only server for example




https://t.me/learningnets

# NAT64 – How it works




https://t.me/learningnets

# NAT 44 and NAT64 Difference

- Translation of IP address into another IP address is known as Network Address Translation (NAT44 , NAT-PT)

- Translation of IP address from one address family into another address family is known as AFT (Address Family Translation), example is NAT 64, NAT46

https://t.me/learningnets

# CGN – Carrier Grade NAT

- CGN is commonly known as LSN (Large Scale NAT) in the operator community

- In CGN, IP addresses and the transport ports (TCP and UDP) are shared among the users

- CGN/LSN is commonly referred as NAT 444

- NAT444 , having two layer of NAT, first at the customer side, second layer of NAT in the SP network

https://t.me/learningnets

# CGN – Carrier Grade NAT

- CGN moves Public IPv4 address pool from the customer edge to more centralized location

- It is Address + Port Translation solution (NAPT)

- Can be accomplished in multiple ways

1. NAT 444 (Two layer of NAT44)
2. NAT 464
3. DS-LITE (It is a tunneling solution but mostly used together with CGN)

https://t.me/learningnets

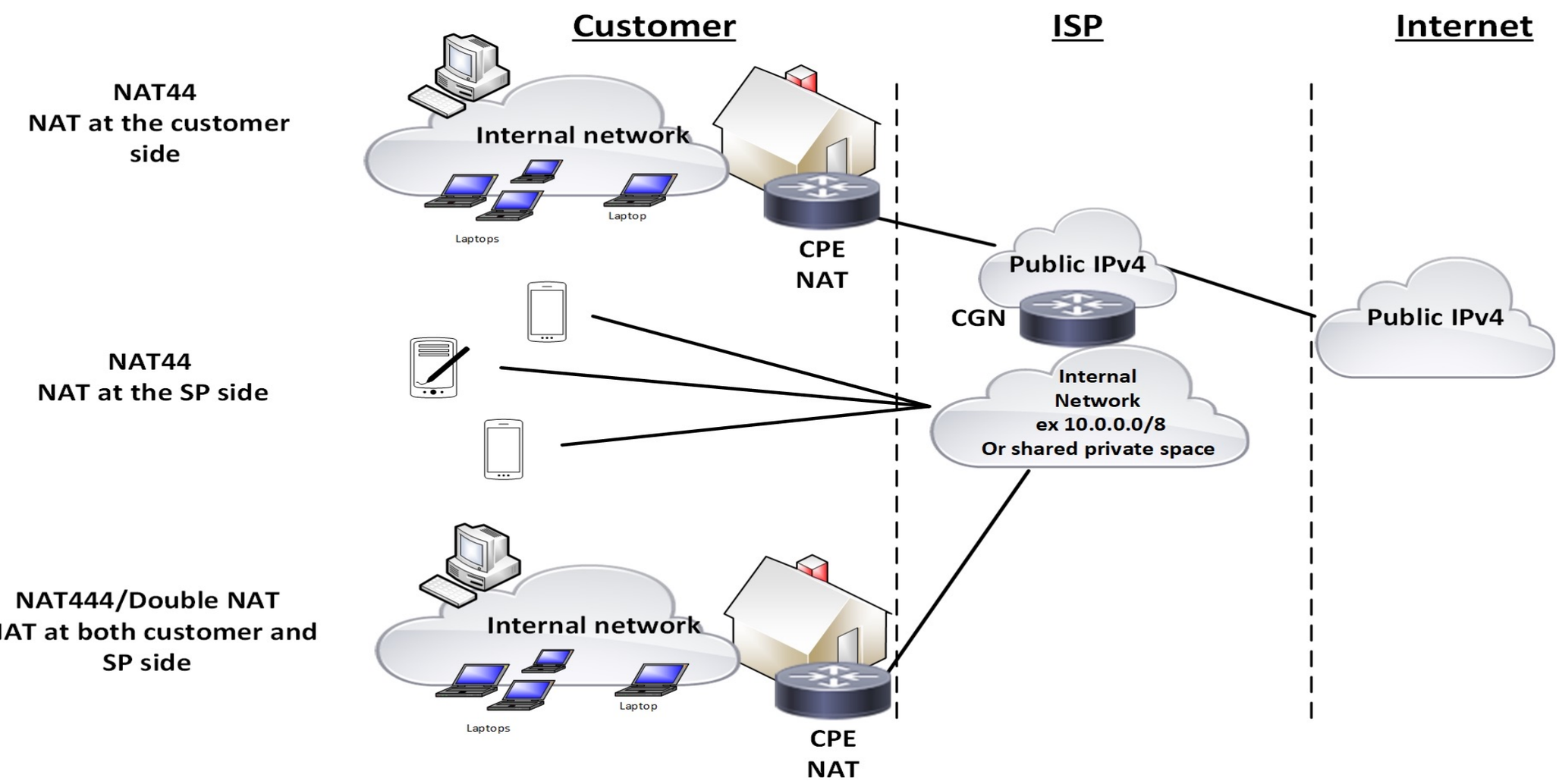
# CGN – Carrier Grade NAT

- Difference between Customer NAT (Residential NAT) and SP NAT (CGN, LSN) is, with Residential NAT, single public IPv4 address represent one household, with SP NAT (CGN,LSN), single public IPv4 address is shared across multiple households

- With Residential NAT, 16-bit port space(65000 TCP and UDP ports) is for single household but with SP NAT, 16-bit port space of the IP address is shared among multiple households

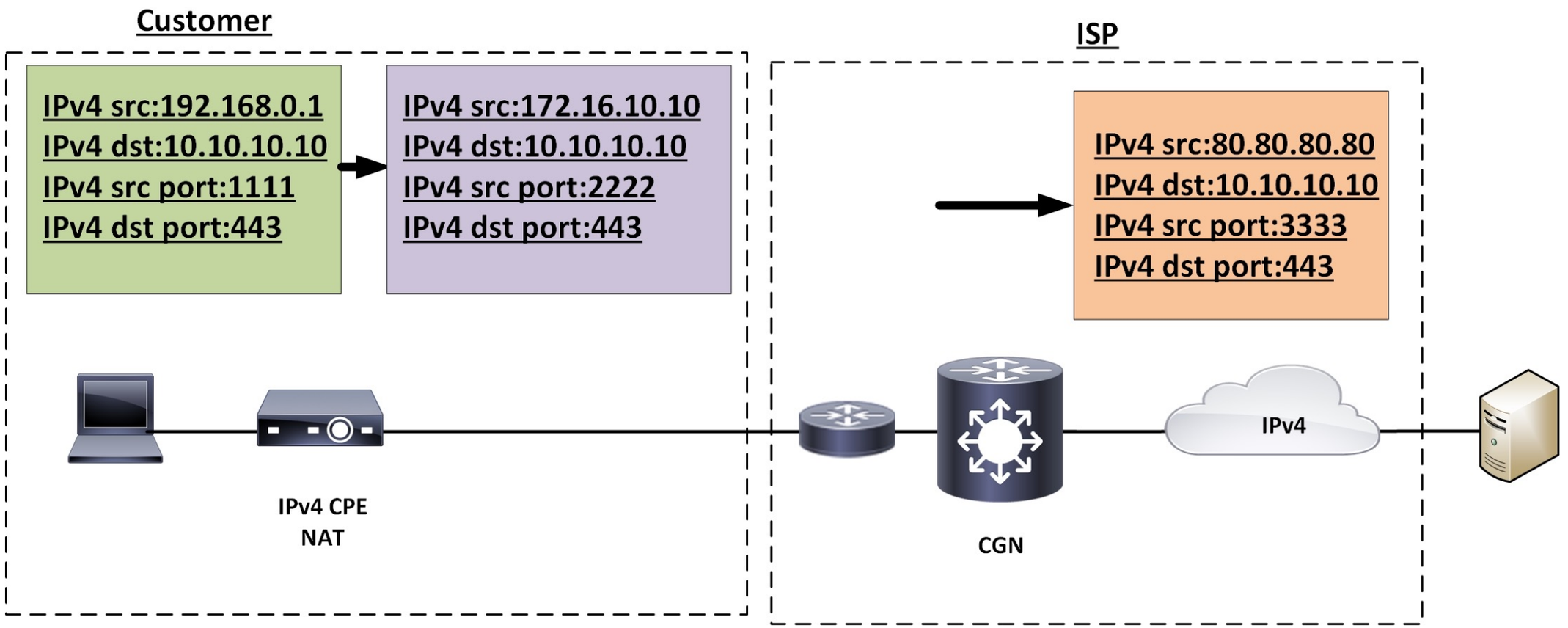
https://t.me/learningnets

# How NAT (Network Address Translation) Wor*ks*




https://t.me/learningnets

# CGN – How CGN Work




https://t.me/learningnets

# CGN – Carrier Grade NAT, LSN – Large Scale NAT Deployment Options

- CGN can be deployed either as Inline or Offline

- Inline CGN deployment is more common in Enterprise and Residential networks as network traffic pass through the NAT box

https://t.me/learningnets

# CGN – Carrier Grade NAT, LSN – Large Scale NAT Deployment Options

- Offline CGN removes the NAT from the primary data path and utilizes source routing mechanisms to send the traffic to the NAT boxes

- Offline CGN is more common deployment model in the SP networks

https://t.me/learningnets

# CGN Advantages

- It is well known NAT , two times NAT operation , customer and SP side, no IPv6 learning curve

- CPE – Customer NAT doesn’t need to change

- CPE doesn’t need to support IPv6

https://t.me/learningnets

# CGN – Sharing Addresses Problems

- CGN is an IP address sharing solution, many users share the same Public IP address, there are problems with it

  - Some applications break , applications which can work with single Layer of NAT may not work with two layers of NAT

https://t.me/learningnets

# CGN – Sharing Addresses Problems

- Sharing addresses makes operations/troubleshooting harder

How many ports should be assigned to each user? It is called Port Spray

Many websites open 80-100 TCP connection (Newspapers), some apps open hundreds of sessions (Google Map etc.)

https://t.me/learningnets

# CGN – Sharing Address Problems

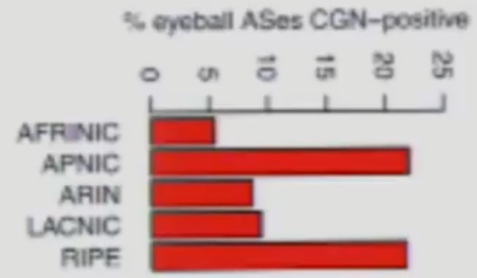
- Intense logging will be needed for the Lawful intercept
- Traceability of users behind CGN
- CGN in forwarding path (Inline deployment) becomes single point of failure
- Offline CGN deployment requires source routing which creates unnecessary complexity
- CGN IP address getting blacklisted due to address sharing (Not every user is innocent)

https://t.me/learningnets

# CGN Current Deployment in SP networks

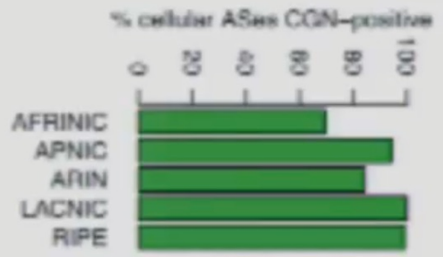
## Eyeball Networks (Non-Cellular)

- CGN-positive: **17.1%**
  - particularly in the European and Asia-Pacific Region




## Cellular Networks

- CGN-positive: **94%**
  - CGN is the norm for cellular




https://t.me/learningnets

# Multicast VPNs

- A BGP/MPLS IP VPN service that supports multicast is known as a "Multicast VPN" or "MVPN”

- RFC 4364 specifies Unicast IP/MPLS VPN Service

- RFC 6513 specifies Multicast in MPLS/BGP IP VPNs

https://t.me/learningnets

# Multicast VPNs

- In a "BGP/MPLS IP VPN" (RFC 4364) unicast routing of VPN packets is achieved without the need to keep any per-VPN state in the core of the SP's network (the "P routers")

- Routing information from a particular VPN is maintained only by the PE routers that attach directly to sites of that VPN

https://t.me/learningnets

# Multicast VPNs

- Customer data travels through the P routers in tunnels from one PE to another (usually MPLS Label Switched Paths, LSPs), so to support the VPN service the P routers only need to have routes to the PE routers

https://t.me/learningnets

# Multicast VPNs

- The PE-to-PE routing is optimal, but the amount of associated state in the P routers depends only on the number of PEs, not on the number of VPNs

https://t.me/learningnets

# Multicast VPNs

## Scalability vs Optimality Trade-off

- However, in order to provide optimal multicast routing for a particular multicast flow, the P routers through which that flow travels have to hold state that is specific to that flow

https://t.me/learningnets

# Multicast VPNs

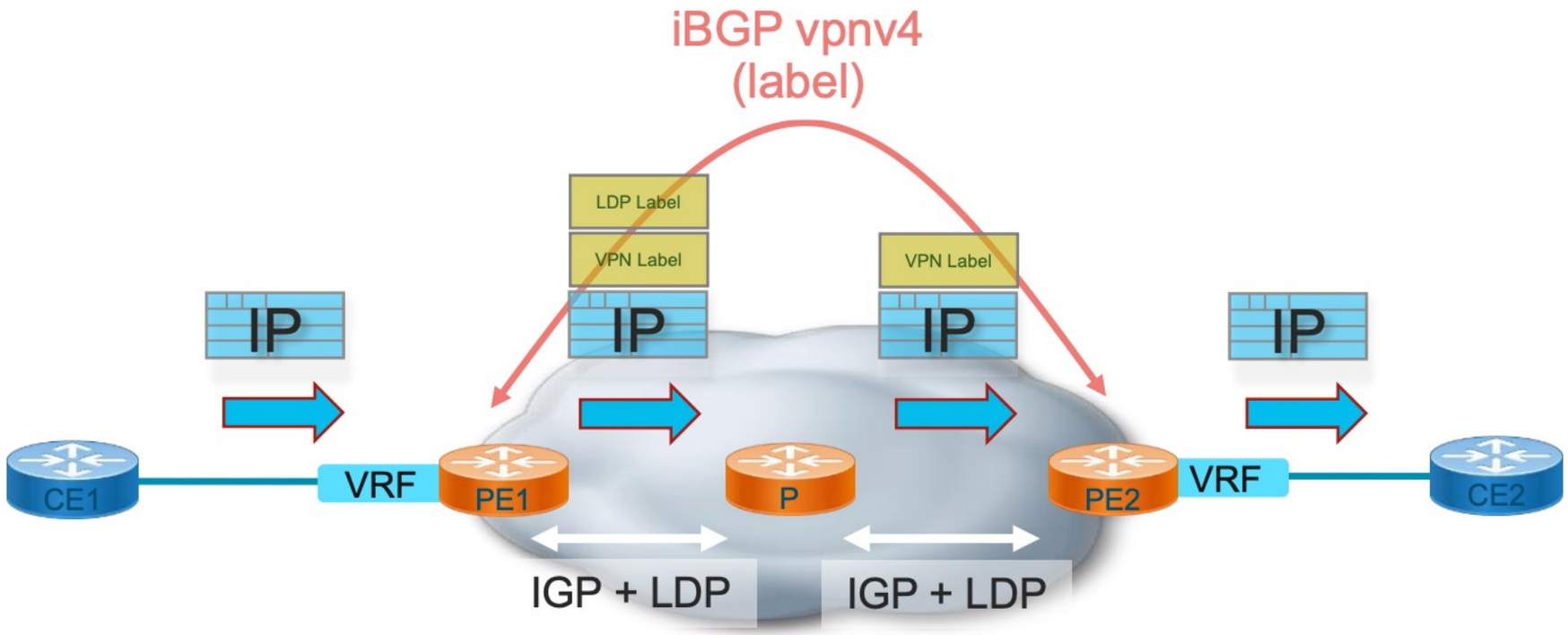
- A multicast flow is identified by the (source, group) tuple where the source is the IP address of the sender and the group is the IP multicast group address of the destination

- Scalability would be poor if the amount of state in the P routers were proportional to the number of multicast flows in the VPNs

- Therefore, when supporting multicast service for a BGP/MPLS IP VPN, the optimality of the multicast routing must be traded off against the scalability of the P routers

https://t.me/learningnets

# Unicast Control Plane has to work for Multicast to work!




Multicast uses Unicast for RPF, thus Unicast network has to work for it!

https://t.me/learningnets

# Multicast VPN Architecture

**There are two basic models for mVPN deployment: Classic mVPN and NG mVPN**

- The Classic mVPN version (PIM/GRE) is based on the Draft Rosen and requires PIM in the Core of the provider network to be able to support mVPN traffic

- This requirement introduces new protocols (PIM) and encapsulation mechanisms (mGRE) in the control and data planes

- As we will discuss later, this model has a poor scalability and flexibility

- NG mVPN model (BGP/MPLS) uses the same protocols used in MPLS L3 VPN unicast service (RFC 2547, RFC 4364 BGP MPLS VPNs) but with some extensions to the control and data plane protocols as necessary to be able to support multicast traffic

- This model retains the good scalability and flexibility of the unicast VPN service

https://t.me/learningnets

# Multicast VPNs

- We will first look at different types of MDTs, which mean Underlay transport network, after that all other required protocols and the options for Multicast VPNs will be explained

- MDTs can be created by any Core-Tree Protocols (PIM, mLDP, RSVP, Ingress Replication) as we will discuss later in detail

https://t.me/learningnets

# MDT – Multicast Distribution Tree

- A path between the Multicast Sender/Source to the Receivers is called Multicast Distribution Tree

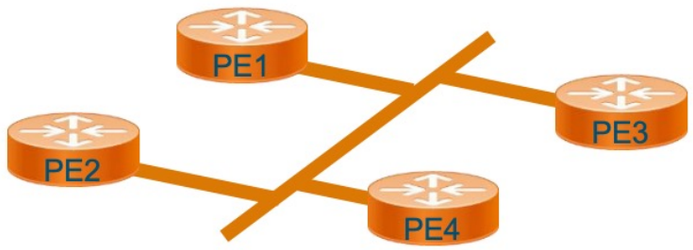
- There are 3 different type of MDT : Default MDT, Data MDT and Partitioned MDT

- MDT is called by Cisco, IETF terminology for MDT is PMSI (Provider Multicast Service Interface)

https://t.me/learningnets

# MDT – Multicast Distribution Tree

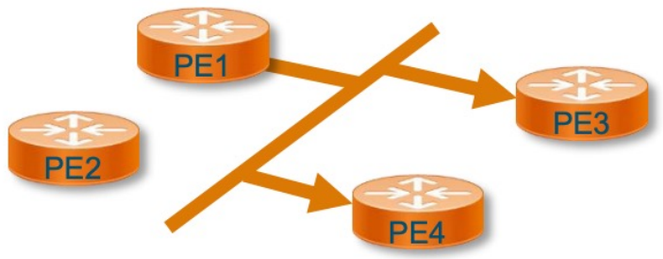
## Default MDT




- Connects all PEs
- Bidirectional
- Always present

Multi-Directional Inclusive PMSI
MI-PMSI

## Data MDT

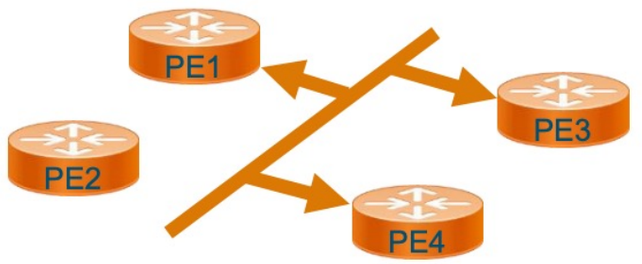



- Connects subset of PEs
- Unidirectional
- On-demand

Selective PMSI
S-PMSI

MDT = Multicast Distribution Tree
PMSI = Provider Multicast Service Interface

## Partitioned MDT

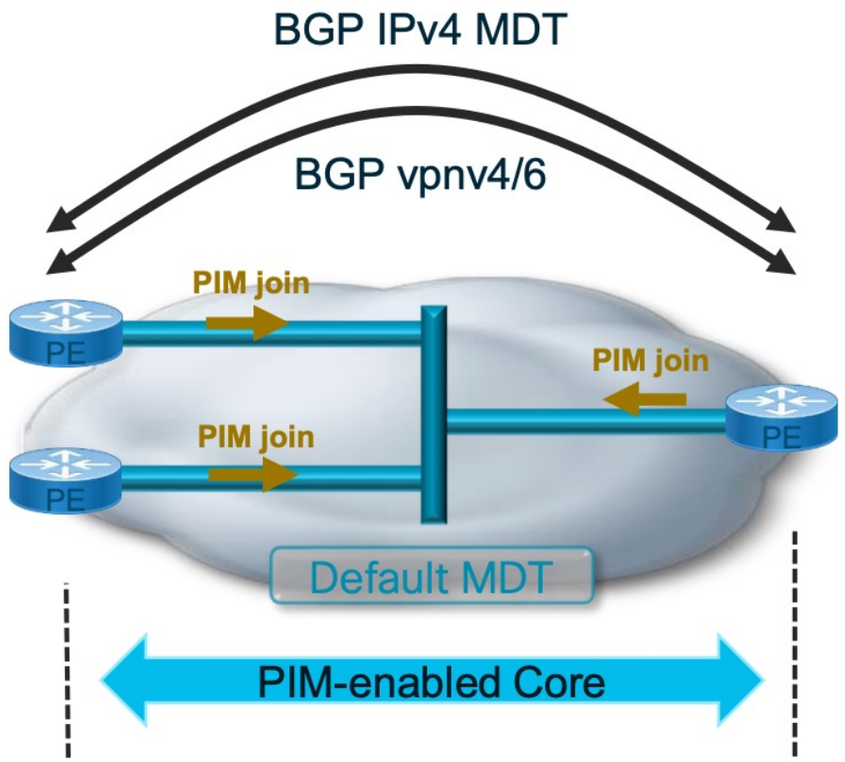



- Connects subset of PEs
- Uni- or Bidirectional
- On-demand

Multidirectional Selective PMSI
MS-PMSI

https://t.me/learningnets

# MDT – Default MDT

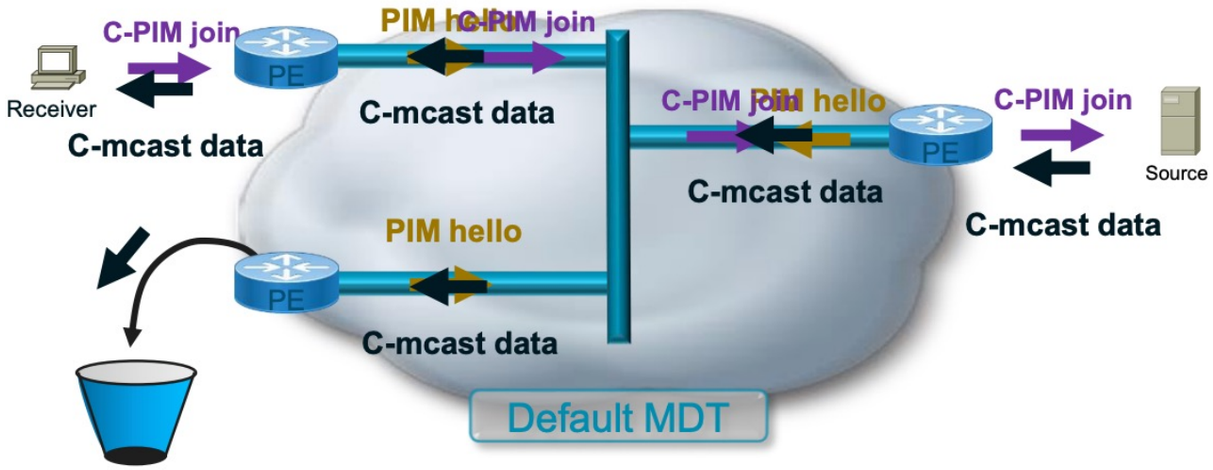



1. Unicast routing (RPF)
2. Build Default MDT
   - 2.1 Advertise PE as member of mVPN by BGP IPv4 MDT
   - 2.2 PE routers join Default MDT PIM tree

Default MDT connects all mVPN PE routers and carries all PIM signaling and all mcast traffic by default

https://t.me/learningnets

# MDT – Default MDT

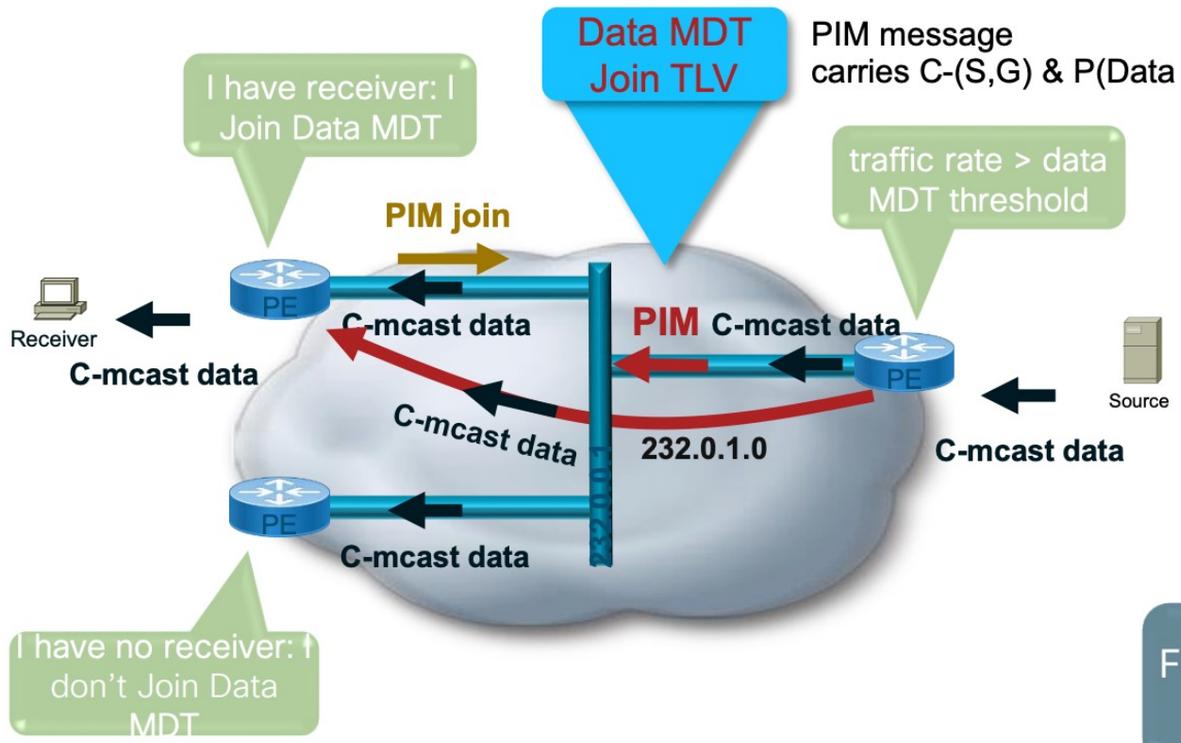



1. C-PIM neighbors across Default MDT
2. C-PIM signaling across Default MDT
3. C-multicast data

One caveat of Default MDT: drop mcast traffic on egress PE if no receiver

https://t.me/learningnets

# MDT – Data MDT




1. Signal Data MDT
2. PEs with receiver Join Data MDT
3. Ingress PE offloads traffic after 3 seconds from Default MDT onto Data MDT

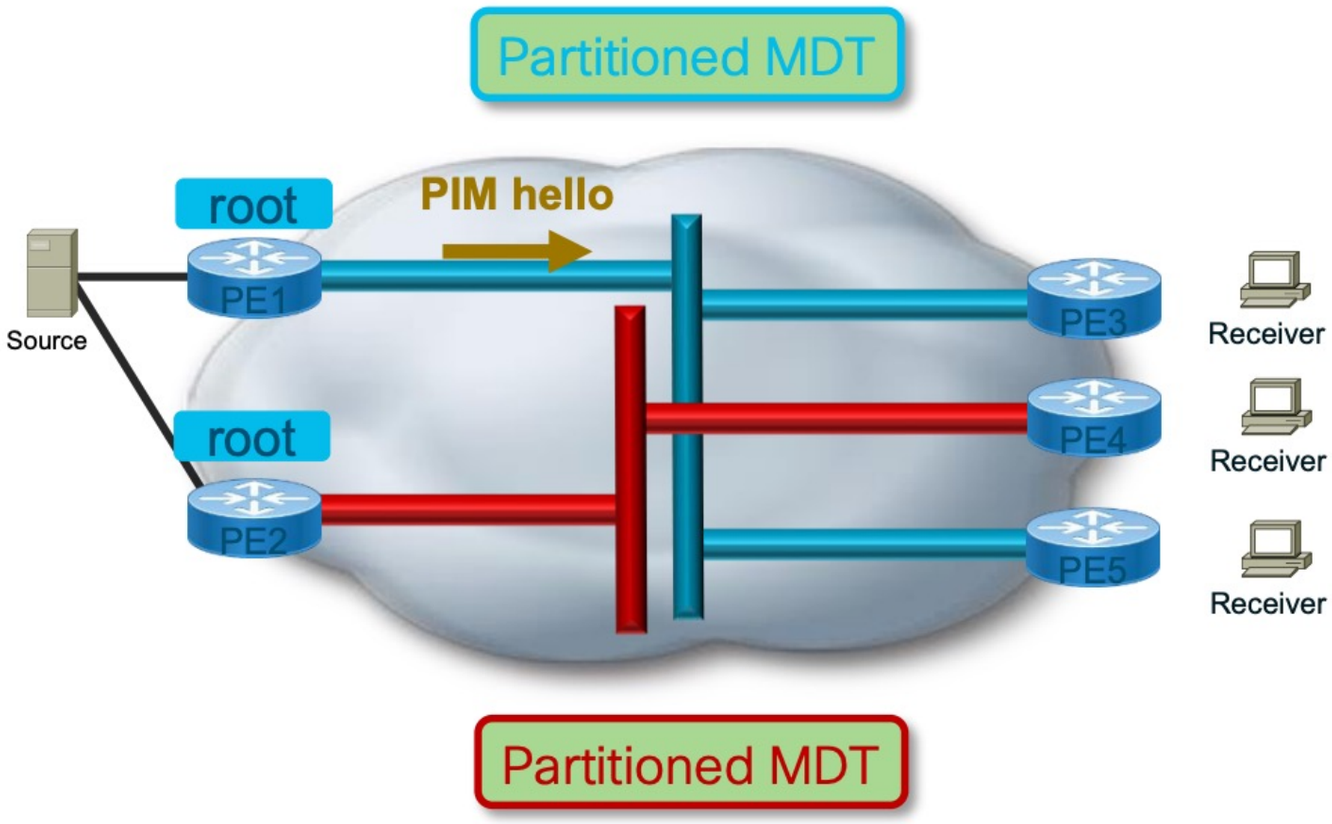
For high rate C-mcast traffic: go onto Data MDT

https://t.me/learningnets

# MDT – Partitoned MDT

- Bidirectional Selective PMSI (Thus subset of Pes)

- BGP AD is necessary for the Partitoned MDT



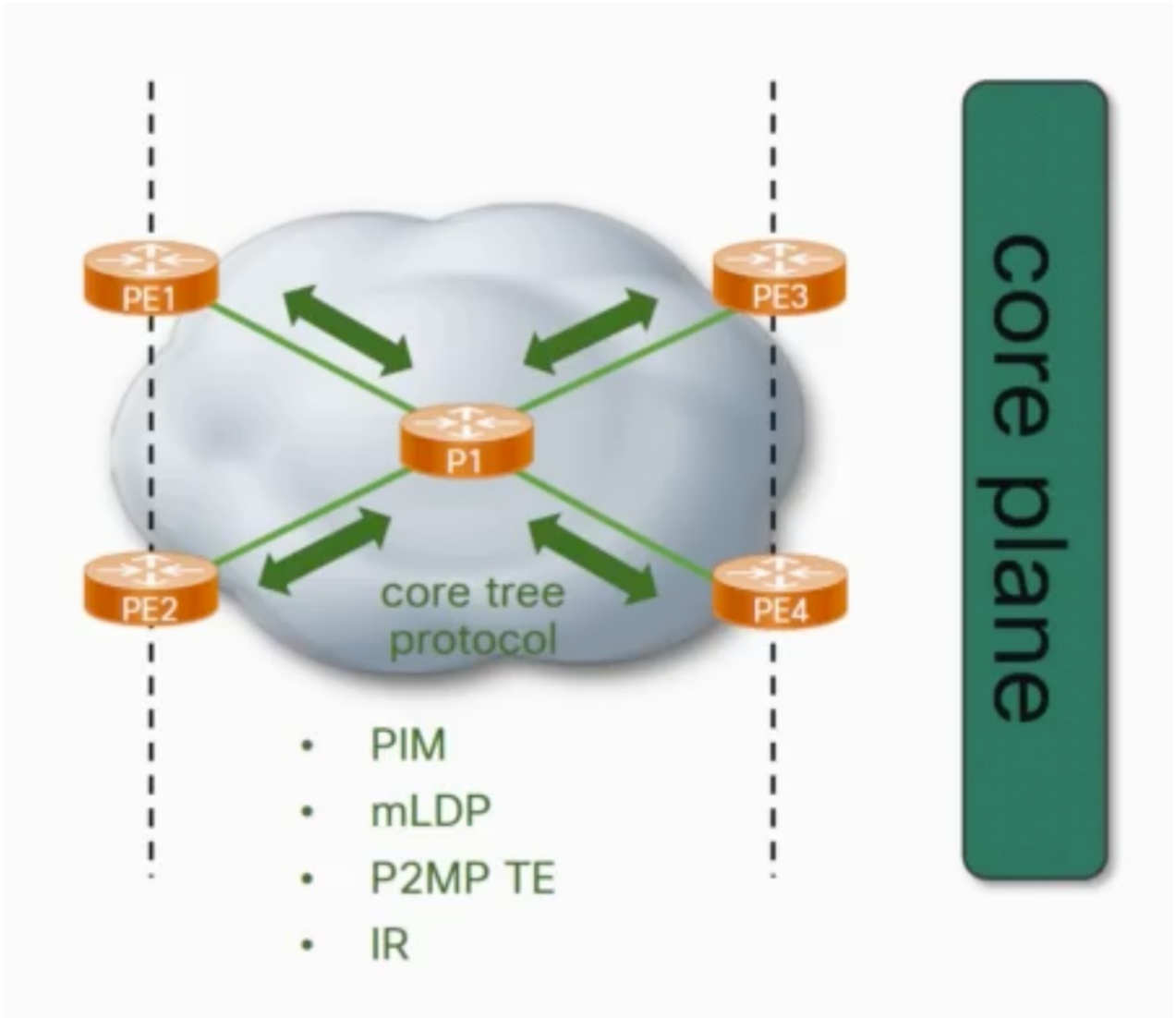

https://t.me/learningnets

# mVPN Required Protocols (All Necessary Components)

https://t.me/learningnets

# mVPN – Planes

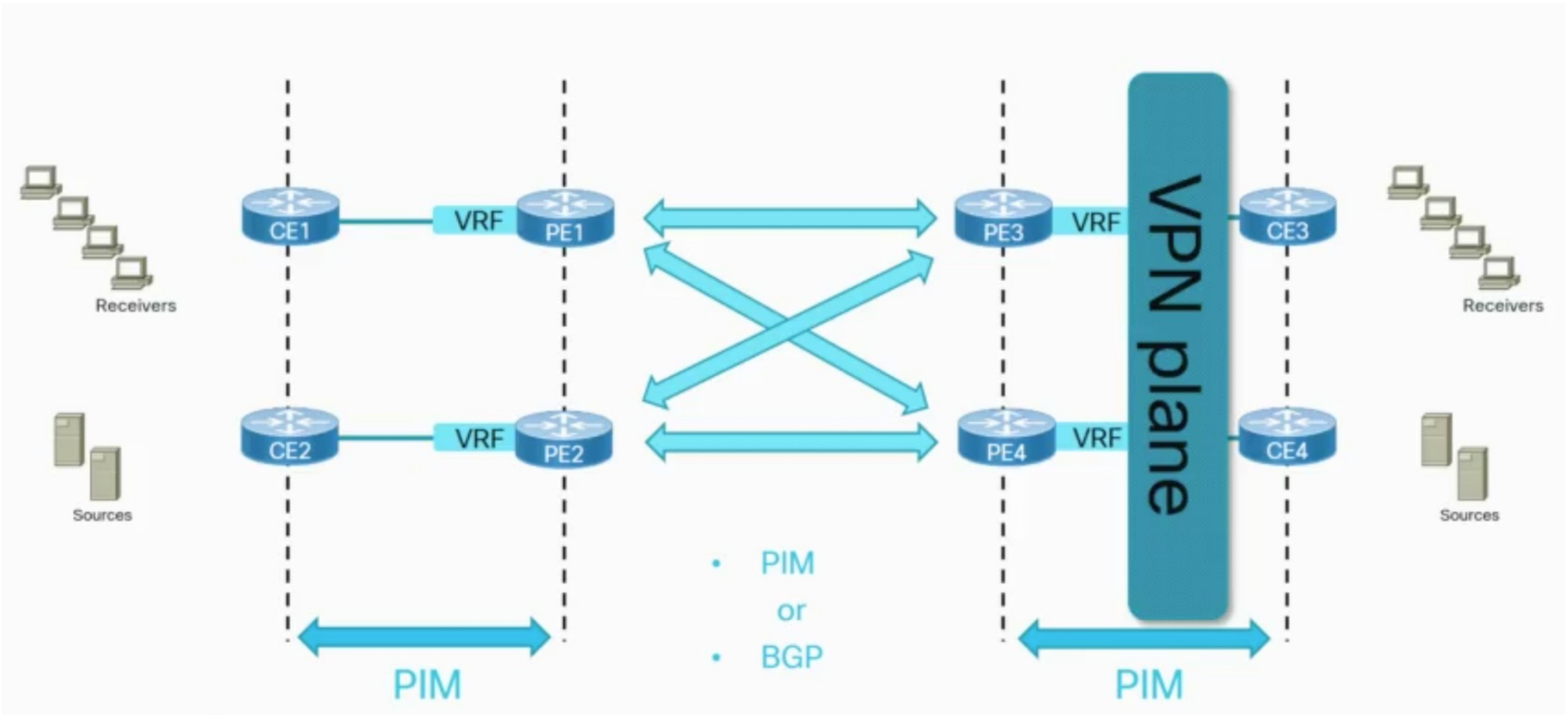
- Underlay can be built by 4 different methods
- Classical method is PIM, but it can be mLDP, P2MP RSVP or IR (Ingress Replication)




https://t.me/learningnets

# mVPN – Planes

- Overlay can be Static or Dynamically via PIM or BGP
- So, VPN Plane is Overlay and it can be BGP if scale is needed or PIM, for the classical mVPNs




https://t.me/learningnets

# mVPN Core Tree Protocols

- mVPN Core Tree Protocols are used to signal MDTs, which are underlays

- Classical mVPN have been using PIM for years, Label Switch Multicast protocols are mLDP and P2MP RSVP-TE

https://t.me/learningnets

# Core Tree Protocols – PIM

- PIM/IP Multicast as we know it, encapsulation is GRE, not MPLS needed

- Well-known protocol

- (*,G) and (S,G) / PIM Sparse Mode, PIM SSM, PIM BiDir

https://t.me/learningnets

# Core Tree Protocols – PIM

- It is used for Default Model / Rosen

- Replication of multicast on the core routers

- Classical mVPN uses PIM in the Underlay for Core Tree Signalling Protocol, as well as in the Overlay for the Customer Multicast Signalling Protocol

https://t.me/learningnets

# Core Tree Protocols – mLDP

- Before we talk about mLDP and other Multicast MPLS based NG-MVPN protocols, maybe we should talk about why LSM (Label Switch Multicast)

https://t.me/learningnets

# Why LSM (Label Switch Multicast)?

## Classic mVPN

- Only GRE encapsulation
- Only PIM in core
- Default MDT / Data MDT

## LSM / NG mVPN

- Leverage MPLS encapsulation
  Share control and data plane with unicast
- Leverage new core tree protocols
  mLDP, P2MP TE, BIER
- Fast ReRoute (FRR)
- Leverage proven functionality:
  Default & Data MDT
- More flexible designs per VPN
- Manageability: no need to track Multicast Groups per VPN/Default/Data MDT

https://t.me/learningnets

# Core Tree Protocols – mLDP

- mLDP is not a periodic update protocol
- PIM is a periodic update protocol
- Multipoint LDP = LDP + extensions
- P2MP tree - Receiver driven – Root learned from routing
- MP2MP tree – Configuration driven – Root configured

https://t.me/learningnets

# Core Tree Protocols – mLDP

- Protection by MPLS TE or Loop-Free Alternate (LFA)

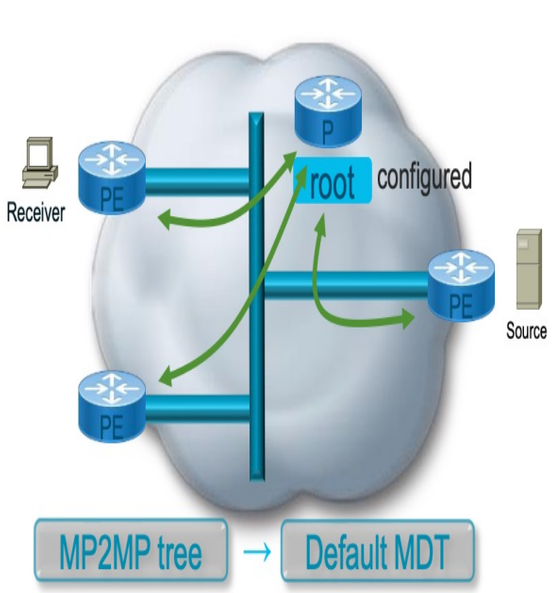
- No PHP – Top label identifies the tree

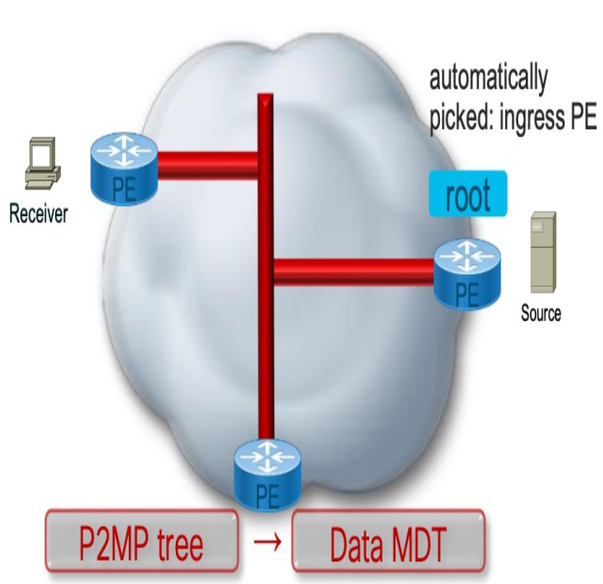
- Replication of Multicast on the core routers (Not Ingress Replication)

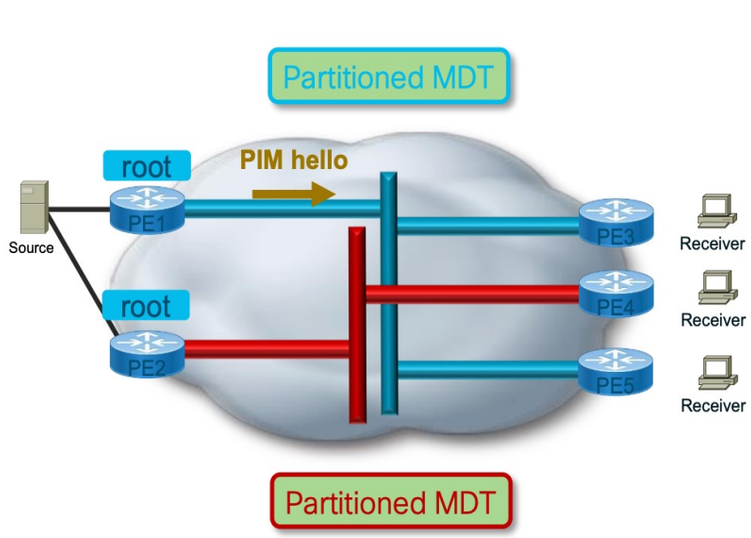
https://t.me/learningnets

# Core Tree Protocols – mLDP

- **With mLDP, Default MDT, Data MDT or Partitoned MDT can be created/signalled**




https://t.me/learningnets

# Core Tree Protocols – P2MP RSVP-TE

- RSVP is soft state protocol, so when scalability is requirement, it might introduce problem

- FRR is supported from day one!

- Bandwidth reservation is possible

https://t.me/learningnets

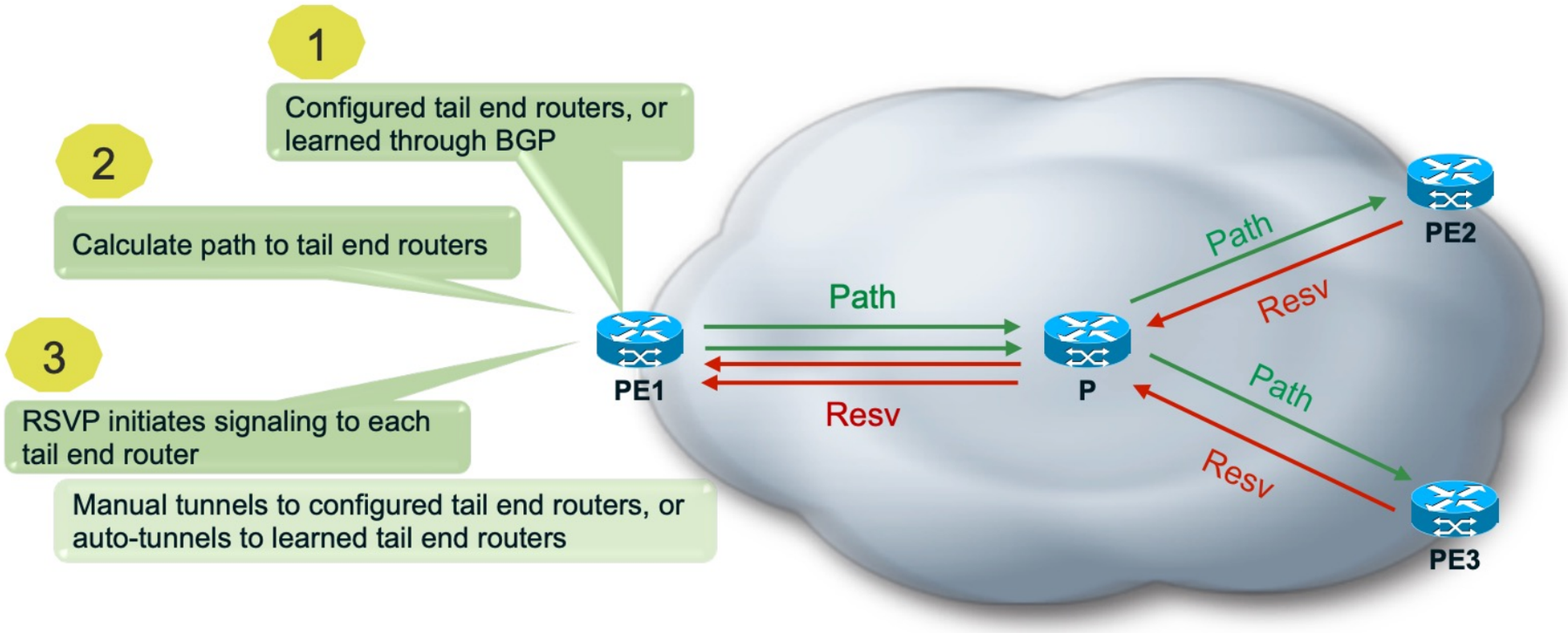
# Core Tree Protocols – P2MP RSVP-TE

- It can signal only P2MP LSPs, not MP2MP LSPs, thus for the Default MDT, there will be more control plane state entry with P2MP-TE compare to mLDP

- Explicit Routing is possible with it, not with mLDP

https://t.me/learningnets

# Core Tree Protocols – P2MP RSVP-TE




https://t.me/learningnets

# Core-Tree Protocols – IR (Ingress Replication)

- When we don’t use P2MP LSP in the Core

- It uses Unicast LSP

- Multicast doesn’t need to be enabled in the Network

https://t.me/learningnets

# Core-Tree Protocols – IR (Ingress Replication)

- If mVPN deployment is small and you don’t want to deploy Multicast in the Network core, it can be choosen

- It is known as Headend Replication as well

- Basically a tunneling Multicast Traffic over Unicast Network

https://t.me/learningnets

# Customer Multicast Signalling (Overlay Signalling) Protocols

- Entire purpose of mVPN is to carry Customer Multicast Traffic over Customer Layer 3 VPN Network

- We need to map Customer Multicast to Core-Tree protocol

- This mapping can be done as an Overlay Signalling or In-band Signalling approach

https://t.me/learningnets

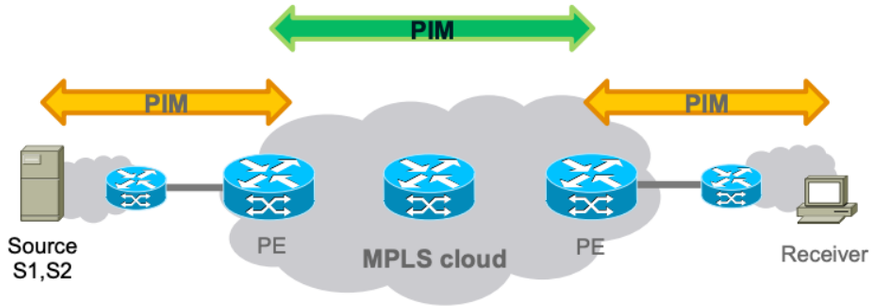
# Customer Multicast Signalling (Overlay Signalling) Protocols

- The purpose is to get the multicast signaling from PE-CE across the MPLS core

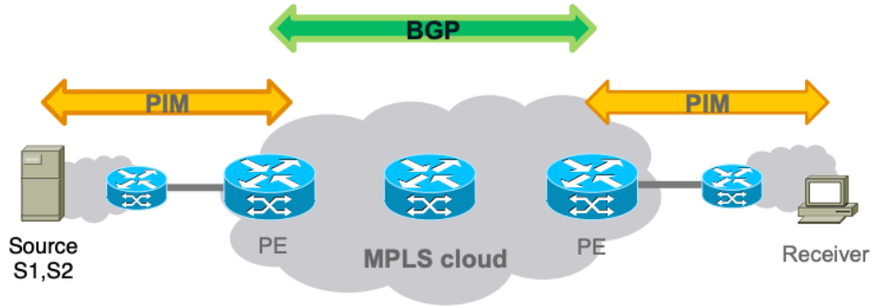
- The multicast signaling on PE-CE is still always PIM!

https://t.me/learningnets

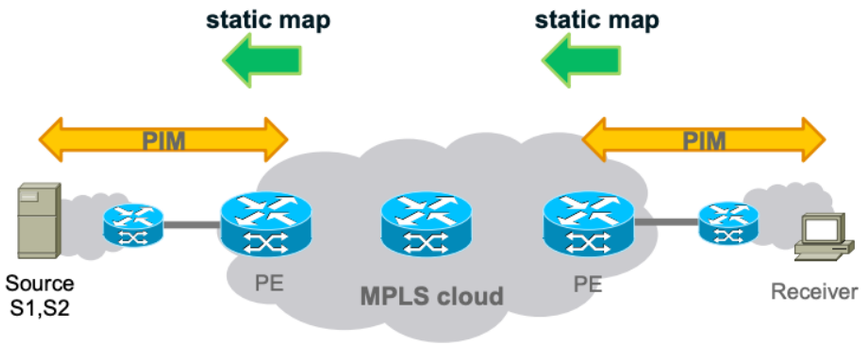
# Customer Multicast Signalling (Overlay Signalling) Protocols



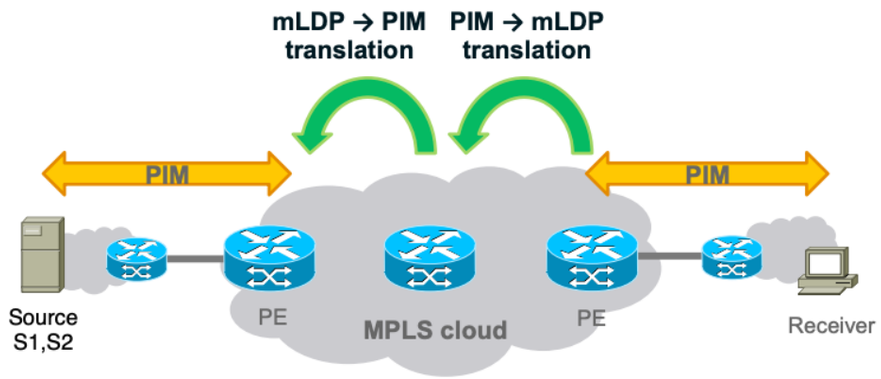

PIM in Overlay




BGP in Overlay




Static




In-band

https://t.me/learningnets

# Customer Multicast Signalling (Overlay Signalling) Protocols  PIM and Static Mapping

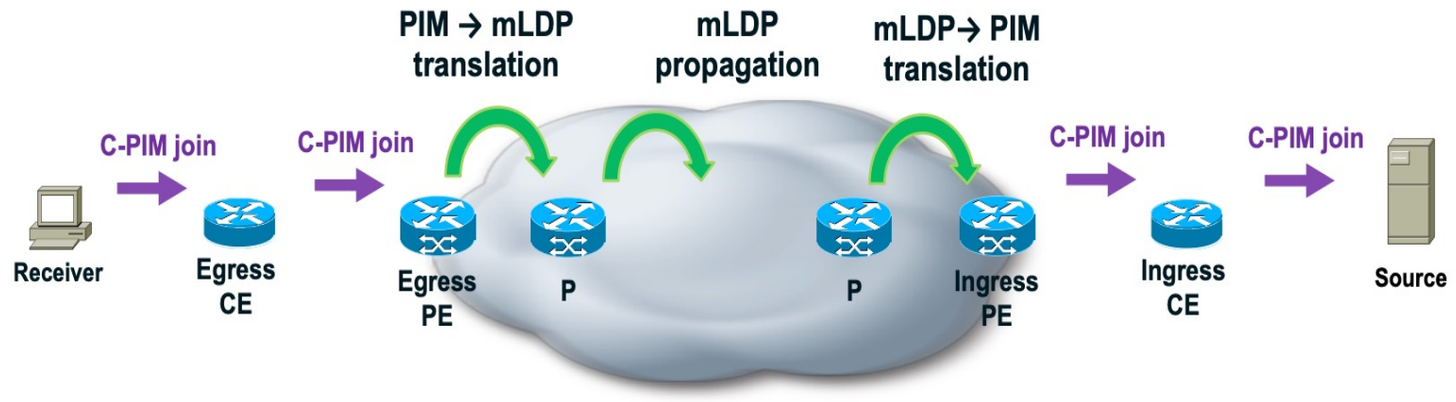
- PIM as Overlay Signalling is used to carry C-Multicast messages within Provider VRF context

- Static Mapping is possible only for P2MP TE LSP, not with mLDP

https://t.me/learningnets

# Customer Multicast Signalling (Overlay Signalling) Protocols - In-band Signalling

- It is only possible with mLDP
- No Overlay signalling protocol is necessary
- Customer PIM messages (Join/Prunes) are translated to mLDP LSP




Tradeoff is State vs Less Protocol

We don't have Overlay protocol but every Join/Prune state is carried in the mLDP in the Core Network as well

https://t.me/learningnets

# Customer Multicast Signalling (Overlay Signalling) Protocols  BGP Signalling

- BGP can be choosen as Overlay C-Mast Signalling Protocol for Scalability reason

- When BGP is used for Overlay Signalling, it can be used for two purposes

1. Auto Discovery
2. C-Mast Routes

https://t.me/learningnets

# Customer Multicast Signalling (Overlay Signalling) Protocols  BGP Signalling

- PIM is assumed to be the C-multicast protocol, i.e. the PE-CE multicast routing protocol
- When BGP is used for Multicast Routing, the C-Join of the C-multicast protocol is not sent over the VPN backbone.
- Instead, it is translated into a BGP C-multicast tree join route advertised through BGP.
- There are two types of C-multicast BGP routes:
  - C-multicast Shared Tree Join route
  - C-multicast Source Tree Join route

https://t.me/learningnets

# Customer Multicast Signalling (Overlay Signalling) Protocols  BGP Signalling

This BGP signaling (AD and C-mcast) uses address family “ipv4 mvpn”

| Route Type | Name | AD or C-mcast | Usage |
|---|---|---|---|
| 1 | Intra-AS I-PMSI A-D route | AD | Advertise member PE |
| 2 | Inter-AS I-PMSI A-D route | AD | Same as type 1, but for inter-AS |
| 3 | S-PMSI A-D route | AD | Advertised by Source PE to signal Data MDT |
| 4 | Leaf A-D route | AD | Advertised by Receiver-PE as a response to type 3 route with leaf info required flag set |
| 5 | Source Active A-D route | AD | Advertised by Source PE with active source to facilitate SPT switchover |
| 6 | Shared Tree Join route | C-mcast | Advertise (*,G) by Receiver PE |
| 7 | Source Tree Join route | C-mcast | Advertise (S,G) by Receiver PE |

https://t.me/learningnets

# Auto Discovery

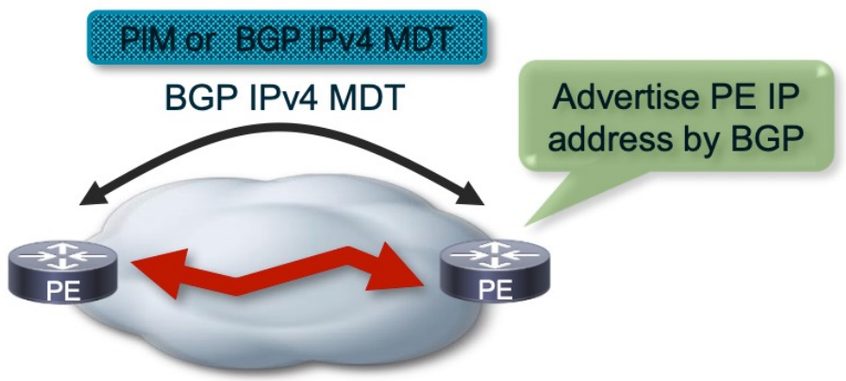
- When mVPN Options are choosen in Cisco's mVPN portfolio, one of the criteria is Auto-Discovery

- It simplifies the configuration, thus reduces OPEX but might introduce additional protocol complexity

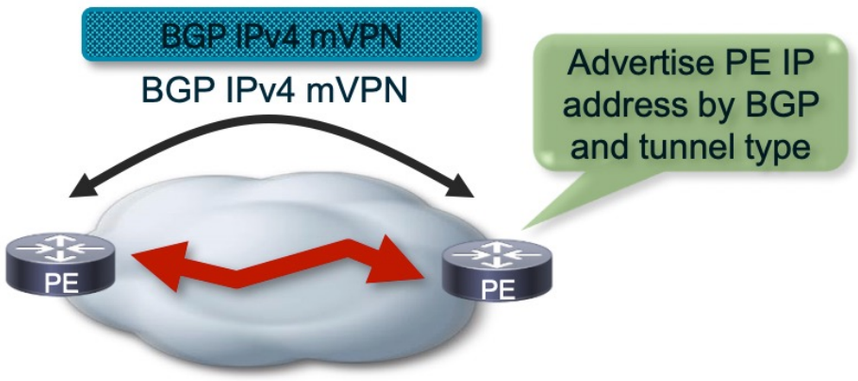
https://t.me/learningnets

# Auto Discovery

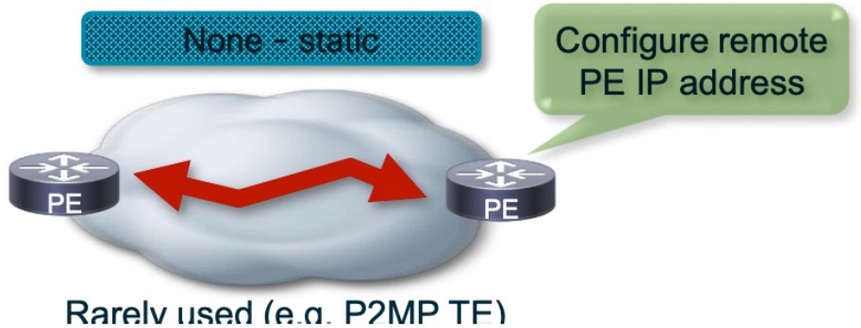
- The process of discovering all the PEs with members in a given mVPN
- In order to build the MDT




Used by classic mVPN




Used by NG mVPN




Rarely used (e.g. P2MP TE)

https://t.me/learningnets

# mVPN Profiles

- Cisco, based on Core Tree Protocol, Overlay Signalling, MDT Type, Encapsulation Type and Auto Discovery support, has 27 Options

- Cisco calls this option as mVPN Profiles

- There is a total of 27 profiles, numbered 0 to 26

https://t.me/learningnets

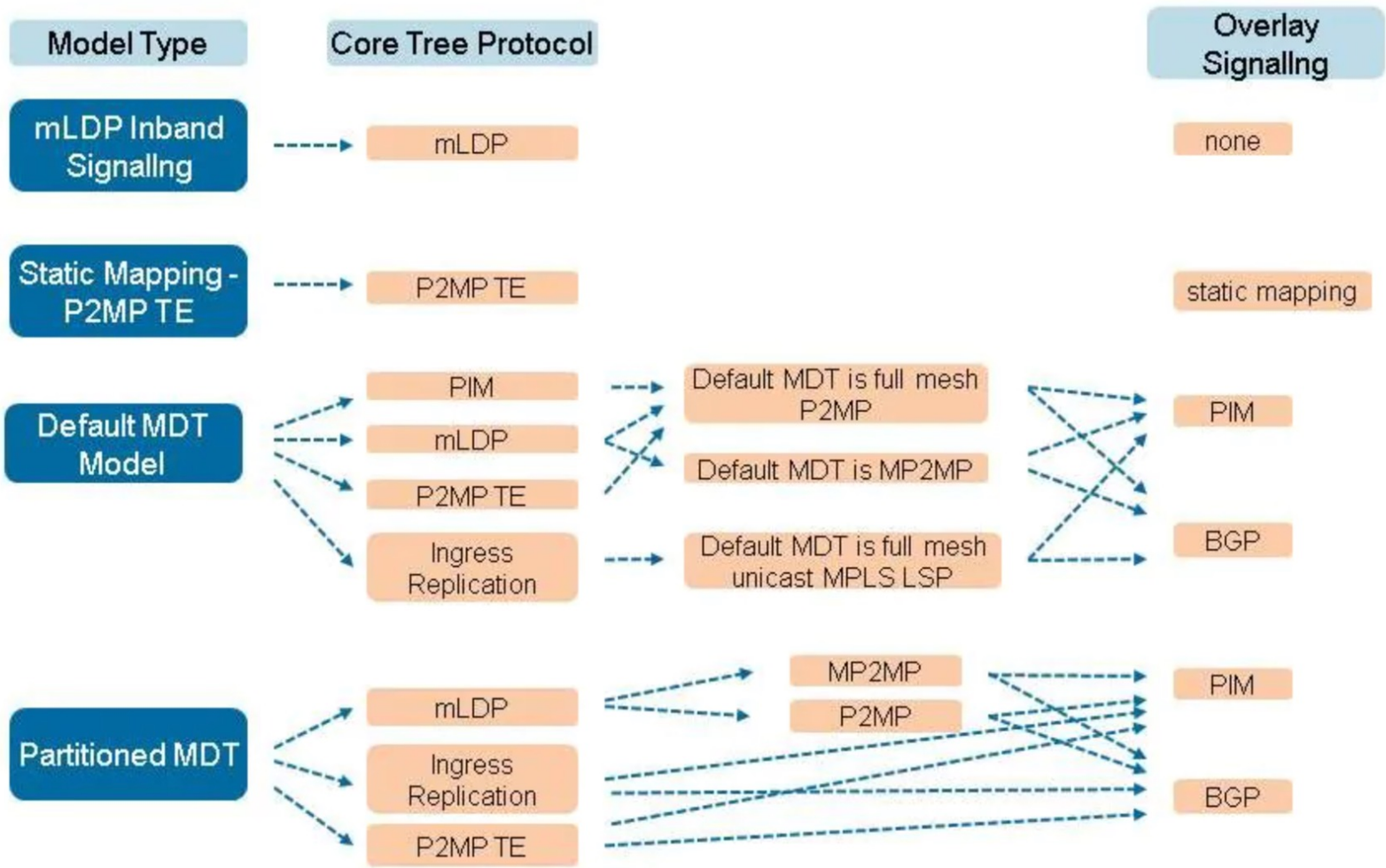
# mVPN Profiles

Profiles

Profile 0 Default MDT - GRE - PIM C-mcast Signaling
Profile 1 Default MDT - MLDP  MP2MP  PIM C-mcast Signaling
Profile 2 Partitioned MDT - MLDP MP2MP - PIM C-mcast Signaling
Profile 3 Default MDT - GRE - BGP-AD - PIM C-mcast Signaling
Profile 4 Partitioned MDT - MLDP MP2MP - BGP-AD - PIM C-mcast Signaling
Profile 5 Partitioned MDT - MLDP P2MP - BGP-AD - PIM C-mcast Signaling
Profile 6 VRF MLDP - In-Band Signaling
Profile 7 Global MLDP In-band Signaling
Profile 8 Global Static - P2MP-TE
Profile 9 Default MDT - MLDP - MP2MP - BGP-AD - PIM C-mcast Signaling
Profile 10 VRF Static - P2MP TE - BGP-AD
Profile 11 Default MDT - GRE - BGP-AD - BGP C-mcast Signaling
Profile 12 Default MDT -  MLDP - P2MP - BGP-AD - BGP C-mcast Signaling
Profile 13 Default MDT - MLDP - MP2MP - BGP-AD - BGP C-mcast Signaling
Profile 14 Partitioned MDT - MLDP P2MP - BGP-AD - BGP C-mast Signaling
Profile 15 Partitioned MDT - MLDP MP2MP - BGP-AD - BGP C-mast Signaling
Profile 16 Default MDT Static - P2MP TE - BGP-AD - BGP C-mcast Signaling
Profile 17 Default MDT - MLDP - P2MP - BGP-AD - PIM C-mcast Signaling
Profile 18 Default Static MDT - P2MP TE - BGP-AD - PIM C-mcast Signaling
Profile 19 Default MDT - IR - BGP-AD - PIM C-mcast Signaling
Profile 20 Default MDT - P2MP-TE - BGP-AD - PIM - C-mcast Signaling
Profile 21 Default MDT - IR - BGP-AD - BGP - C-mcast Signaling
Profile 22 Default MDT - P2MP-TE - BGP-AD BGP - C-mcast Signaling
Profile 23 Partitioned MDT - IR - BGP-AD - PIM C-mcast Signaling
Profile 24 Partitioned MDT - P2MP-TE - BGP-AD - PIM C-mcast Signaling
Profile 25 Partitioned MDT -  IR - BGP-AD - BGP C-mcast Signaling
Profile 26 Partitioned MDT - P2MP TE - BGP-AD - BGP C-mcast Signaling

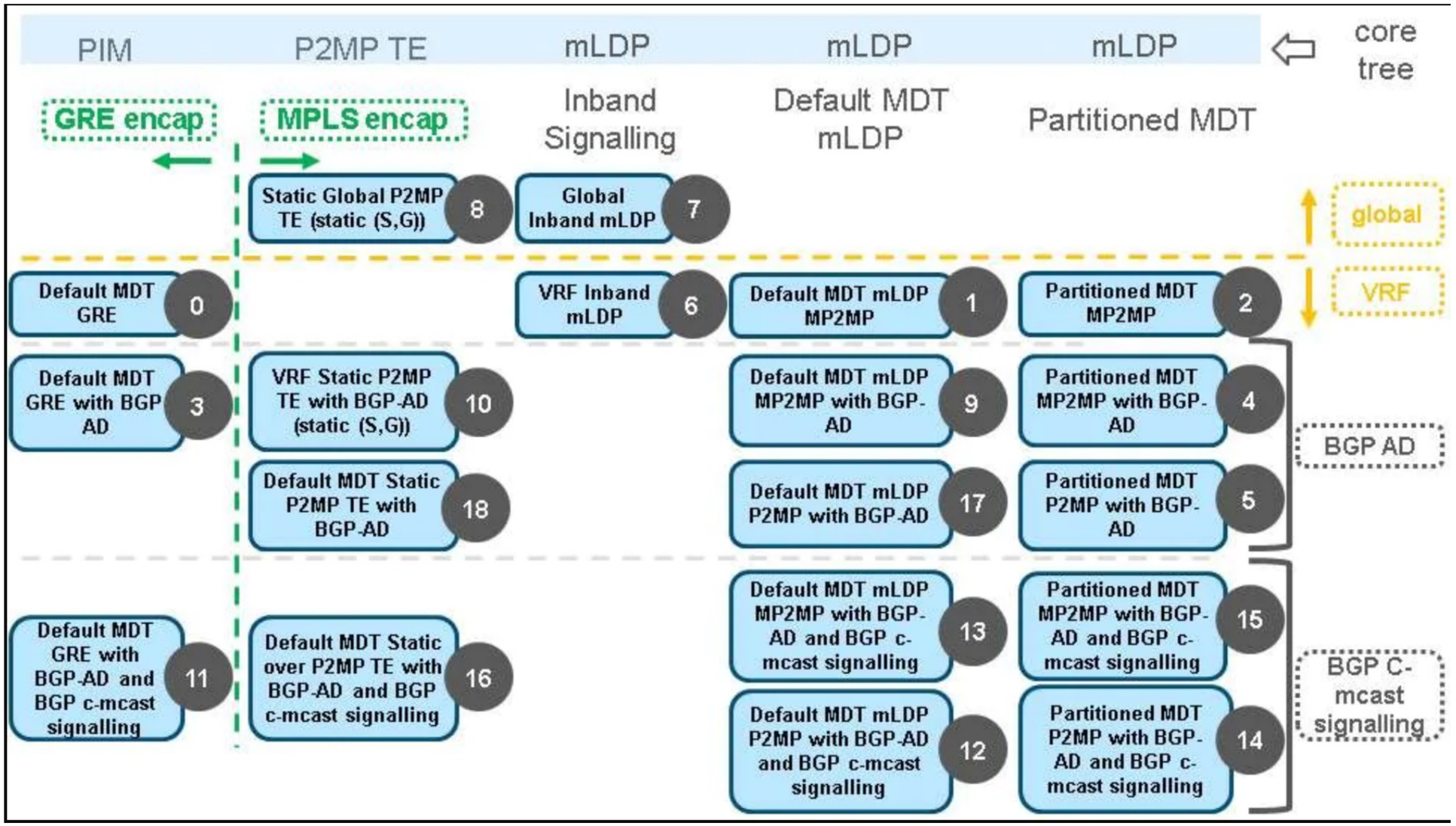
https://t.me/learningnets

# mVPN Profiles



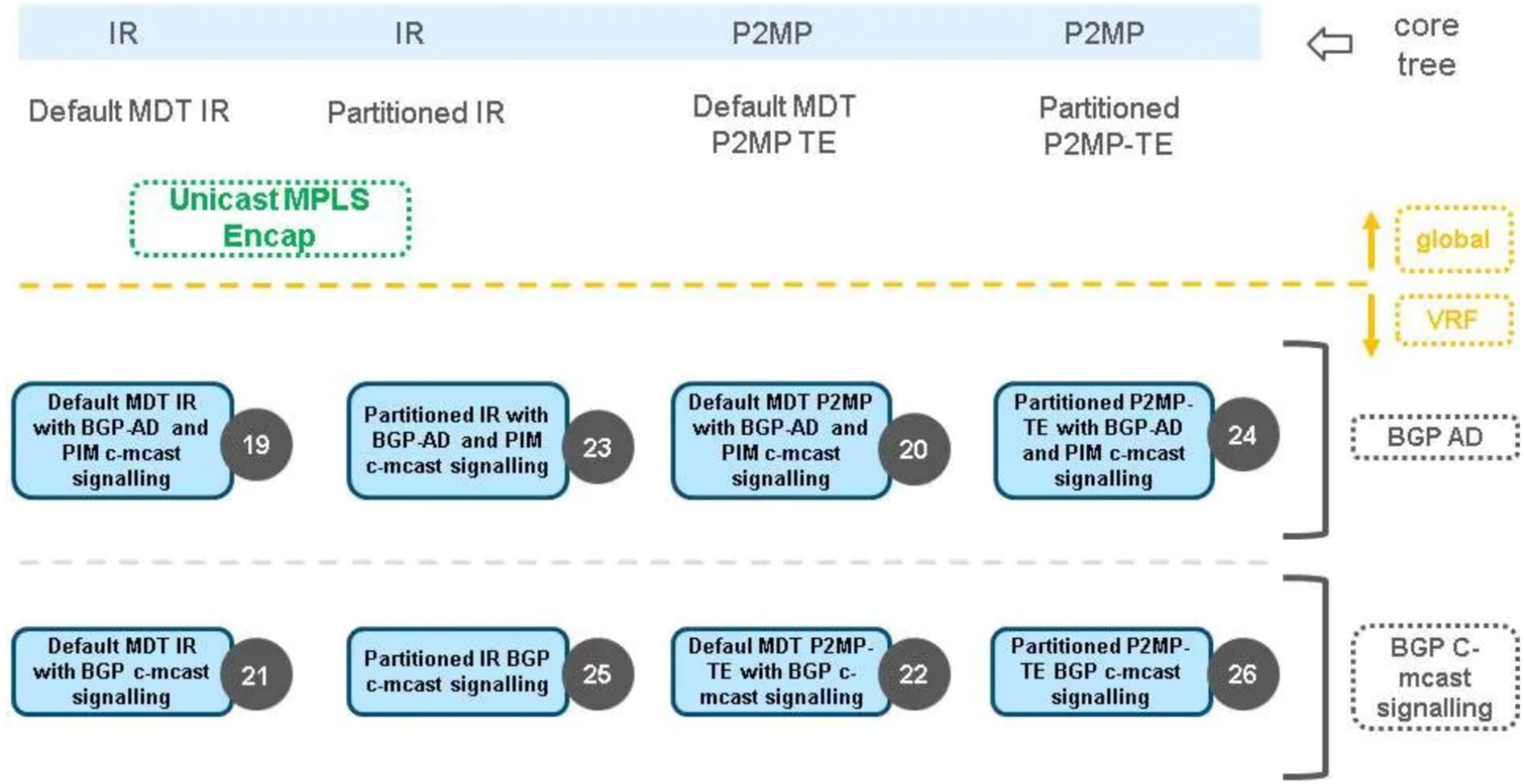

https://t.me/learningnets

# mVPN Profiles




https://t.me/learningnets

# mVPN Profiles




https://t.me/learningnets

# mVPN Core Tree Signalling Protocols Comparison

| PIM | mLDP | P2MP TE |
|---|---|---|
| Mature / Well known | Enhancement to existing protocol | Enhancement to existing protocol |
| Soft state (periodic refresh) | Hard state (no periodic updates) | Soft state (periodic refresh) |
| GRE enacpsulation | MPLS encapsulation | MPLS encapsulation |
| No Fast Restoration | Fast Restoration (povided by LFA or MPLS TE) | Fast Restoration |
| No bandwidth reservation | No bandwidth reservation | Bandwidth reservation |
| P2MP trees only | P2MP and MP2MP trees | P2MP trees only |
| Inter-as and CsC | Inter-as and CsC | Inter-as, but no CsC |
| High complexity | Medium complexity | High complexity |
| Medium core state<br>C-state present in core with Data MDT | Medium core state<br>C-state present in core with Data MDT<br>All C-state in core with in-band signaling | High core state |
| Follows unicast routing | Follows unicast routing | Allows explicit or bandwidth contraint routing |
| Suitable for all mcast applications | Suitable for all mcast applications<br>Best for many-to-many | Mostly suitable for video delivery<br>Best for few-to-many |

https://t.me/learningnets

# PE-PE Customer Signalling Protocols Comparison (C-Multicast) PIM vs BGP

| PIM | BGP |
| --- | --- |
| Older protocol, proven, well known | New enhancement to existing protocol |
| No changes needed<br>Complex with ASM, but well known | New procedures (troubleshooting!)<br>Complex for ASM |
| Soft state (periodic refresh) | Hard state (no periodic updates) |
| Info driven to specific PE router | Info driven to all PE routers |
| PIM adjacencies to all PE routers | BGP adjacencies to all PE routers<br>but likely only to RRs |
| Medium scalability | Very high scalability |

https://t.me/learningnets

# CCDE Practical Scenario – 1

https://t.me/learningnets

# Ornio: The Carpet Manufacturer and Seller

- Ornio is a company engaged in the production of high quality carpets, and it is one of the most well-known carpet brands in the field of carpet manufacturing. It produces Iranian, Indian, and Turkish carpets in its 80-production facilities.

https://t.me/learningnets

- Ornio was created as a workshop, but it has expanded its brand to many countries because of its excellent quality carpets and accessories. Today, Ornio boasts a turnover of more than $80 million, 70% of which is due to increasing exports and product offerings.

https://t.me/learningnets

- The group's headquarter is located in Turkey, one of the most famous districts in the carpet industry, and it includes an office and several production facilities.

- In addition, there are other operational offices abroad and a distribution network that includes branches in Dubai, Qatar, India, Egypt, and Iran, all of which constitute an organization of sales based on retailers and located in over 30 countries throughout the Middle East, Europe, and Unites States. In sum, it has 800 stores growing very fast.

https://t.me/learningnets

- Ornio's datacenter is located in Istanbul which is a most crowded city of Turkey. Currently, the company has only one datacenter, and it started to consider the second datacenter for disaster recovery purpose. The entire IT team is located in Turkey. Ornio handles its IT operations in remote offices through local service contractors.

https://t.me/learningnets

- All the applications of Ornio reside in the datacenter; there are no data or voice applications in the branch offices. Ornio uses IP telephony solution for its voice communication between the branches and the headquarter offices. Its call manager and voice gateways are located in the datacenter.

- For redundancy purpose, it uses two voice gateways for incoming and outgoing voice traffic in the datacenter. Also, It has subscriber and publisher call managers for optimal resource usage and high availability.

https://t.me/learningnets

- All the branch offices of Ornio are connected to one headquarter router via frame-relay circuit. Currently, there is only one connection from all the branches. There is no Frame Relay connection between the branch offices, so all the traffic between branch offices go through it's headquarter.

- Branch offices have their own Internet connection. Ornio is considering building a backup VPN connection over the Internet. Although it is acceptable for the company not to have encryption over the Frame Relay, it definitely wants to have an encryption over the Internet if it set up VPN connections over the Internet as a backup.

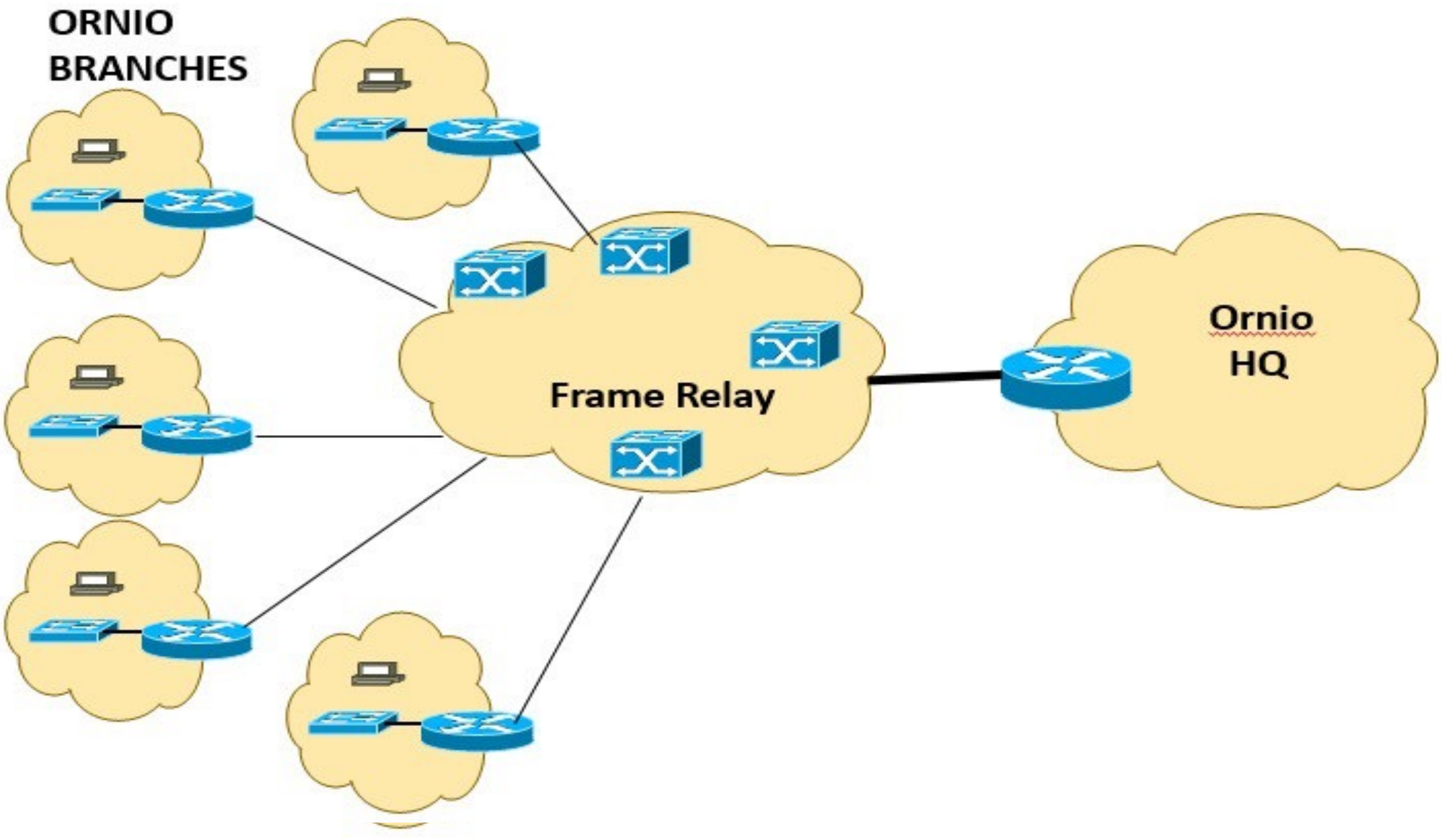
https://t.me/learningnets

- The company recently had many problems with its routing infrastructure and decided to outsource the control of its routing infrastructure to the service provider. Also, it wants a better quality of service for its voice traffic, and it wants to have direct communication between its branch offices. Although overall bandwidth required for the voice is low on the network, it is one of the highly critical applications for Ornio.

- Due to its increasing bandwidth need, Ornio has concerns with its Frame Relay WAN network. If Ornio wants to upgrade its WAN bandwidth, it should be immediate and flexible.

https://t.me/learningnets

# Ornio High Level Wide Area Network Diagram




https://t.me/learningnets

**Question 1: What is the concern of Ornio with its voice communication?**

A. It is not encrypted.

B. There is no loss, jitter, latency guarantee over Frame Relay.

C. Frame Relay is a legacy solution that doesn't offer higher rate bandwidth if it is needed.

D. Its Frame Relay network is based on hub and spoke. Voice communication is not optimal.

https://t.me/learningnets

## ➢ Question 2: Should Ornio change its Frame Relay WAN?

❑ Yes

❑ No

https://t.me/learningnets

➢ **Question 3: Which solution should Ornio choose for its Frame Relay WAN Replacement?**

1. IPSEC VPN over the Internet
2. IPSEC with Frame Relay
3. MPLS VPWS
4. VPLS
5. MPLS L3 VPN

https://t.me/learningnets

## ➢ Question 4: Why?

1. It can support encryption.
2. It can provide hub and spoke VPN for Ornio, so it does not need to change its existing infrastructure.
3. It provides full mesh communication pattern.
4. Ornio can outsource its WAN network to the service provider
5. It can provide better QoS control compare to layer 2 VPNs.

https://t.me/learningnets

## ➢ Question 5: Which additional information do you need from Ornio to migrate its WAN network?

1. Its QoS policy
2. Its application traffic pattern
3. Its routing information
4. Hardware capabilities
5. Does service provider have L3 VPN Offering in every site?

https://t.me/learningnets

# Email-1 is available

❖ **Email 1**

❖ **From**: Level Tokgoz (levent_tokgoz@ornio.com)

❖ **To**: Orhan Ergun (orhan@orhanergun.net)

❖ Subject: WAN Routing and service provider services

https://t.me/learningnets

Orhan,

- *We have several problems with our Wide Area Network (WAN) recently and decided to continue with MPLS Layer 3 VPN offering. We want to have very smooth migration, and we can't tolerate to lose connectivity to our branches.*

- *Our local service provider has its own POP location at all our branch office locations as well as at the Headquarter.*

- *We are currently using EIGRP as a routing protocol over the Frame Relay connections. We have only one router at our branches, which does layer 3 routing between local area network Vlans.*

https://t.me/learningnets

- *We want to have some policy control/traffic engineering capability on our Wide Area Network whenever we want without asking for any permission from the service provider.*

- *We know that with MPLS Layer 3 VPN, the service provider will control our most of our routing, but we still want to influence it as much as we can.*

- *This is a very big project for us, and we want to have full visibility.*

*Regards*

Levent Tokgoz
Network and Security Director
Ornio Group
Maslak/Instabul

https://t.me/learningnets

## ➢ Question 6: Should Ornio change its routing protocol?

❑ Yes
❑ No

https://t.me/learningnets

## ➢ Question 7: Which routing protocol should Ornio choose?

1. OSPF
2. IS-IS
3. RIPv2
4. Static Routing
5. BGP

https://t.me/learningnets

## ➢ Question 8: Why?

1. Not all service providers support EIGRP for MPLS Layer 3 VPN PE-CE protocol.

2. It is Cisco owned protocol, so Ornio doesn’t want to be restricted to a vendor.

3. If it wants to do traffic engineering over its WAN, Ornio has this ability without service provider communication.

4. It can carry more prefixes over BGP than any other routing protocol.

5. It can send non-IP traffic over BGP.

https://t.me/learningnets

➢ **Question 9: Should Ornio use the same autonomous system number on every location or unique autonomous system number for every location?**

A. Same AS
B. Unique AS

https://t.me/learningnets

## ➢ Question 10: Why?

1. Using unique AS per site limits the number of customer sites to number of available BGP AS.
2. Using unique AS may require allocation of AS numbers outside of private AS range.
3. Not all service providers use a unique AS per location.
4. It requires very careful attention to avoid AS collision.
5. Using the same AS is the common best practice.

https://t.me/learningnets

# Email-2 is available

❖ **Email 2**

❖ **From**: Leven Tokgoz (leven_tokgoz@ornio.com)

❖ **To**: Orhan Ergun (orhan@orhanergun.net)

❖ **Subject**: New WAN Routing Protocol

https://t.me/learningnets

***Hi, Orhan***

- *As per your recommendation, we are pleased to continue with the BGP as a new WAN routing protocol.*
- *But we want to know what can be the problem with BGP. If there is something from the service provider site can be done, please do not hesitate to inform us. We can accept to have some level of configuration in order for the solution to work. However, if the configuration is required on each site, we prefer the service provider to do it.*

*Warmest regards*

Levent Tokgoz
Network and Security Director
Ornio Group
Maslak/Instabul
26674

https://t.me/learningnets

➢ **Question 11: What is the problem of using the same BGP AS everywhere?**

1. It requires redistribution from VRF to MP-BGP on the PEs.

2. BGP Loop prevention mechanism rejects prefixes if the same AS number is seen in the AS path.

3. BGP Fast reroute mechanism cannot be implemented with the same AS number everywhere design.

4. Service provider should remove private AS numbers if it will be announce to internet.

5. No site unique characteristics can be identified from the AS path.

https://t.me/learningnets

## ➢ Question 12: How can the problem with the using same AS number on every location be solved?

1. Use unique AS per location.

2. Service provider can override the AS Path, so customer sites don't receive their AS number in the path.

3. Using MPLS in the service provider network.

4. By changing the BGP loop prevention, customer can allow the same AS path to be received.

5. It is not a problem with single homed site, unless Ornio multi-homed its site.

https://t.me/learningnets

➢ **Question 13: Is there any problem on the Service Provider network, if Service Provider configures AS-Override feature ?**

❑ Yes

❑ No

https://t.me/learningnets

➢ **Question 14: What is the problem with BGP AS override configuration at the service provider network?**

1. Service provider BGP configuration becomes more complex.

2. BGP routing loop at the multi-homed customer site.

3. Customer BGP configuration becomes more complex.

4. Some customer sites still reject the BGP prefixes due to its own AS number in the BGP update

5. It is not supported on the BGP route reflector

https://t.me/learningnets

- Question 15: **Is there any problem on the Ornio network, if Service Provider configures AS-Override feature ?**

❑ Yes

❑ No

https://t.me/learningnets

➢ **Question 16: Please provide the necessary steps for the Ornio's MPLS Layer 3 VPN migration.**

1 . Remove the old Frame Relay circuit from the transit site.

2. Establish a BGP over the MPLS circuit.

3. Arrange the routing protocol metric to choose MPLS over Frame Relay.

4. Establish a new circuit at the transit site.

5. Choose a transit site for the communication between migrated and non-migrated site.

6. Establish a new circuit at the remote site.

7. Remove the old Frame Relay circuit from the remote site

8. Enable Quality of Service and Monitoring for the new MPLS connection

https://t.me/learningnets

**Email-3 is available**

# ❖ Email 3

❖ **From:** Levent Tokgoz (levent tokgoz@ornio.com)

❖ **To**: Orhan Ergun (orhan@orhanergun.net)

❖ **Subject:** MPLS Layer 3 VPN Migration and Multicast Service

https://t.me/learningnets

❖ Hi Orhan,

❖ *It was a great achievement. Thanks for helping us for the MPLS migration.*

❖ *We have the below topology.*

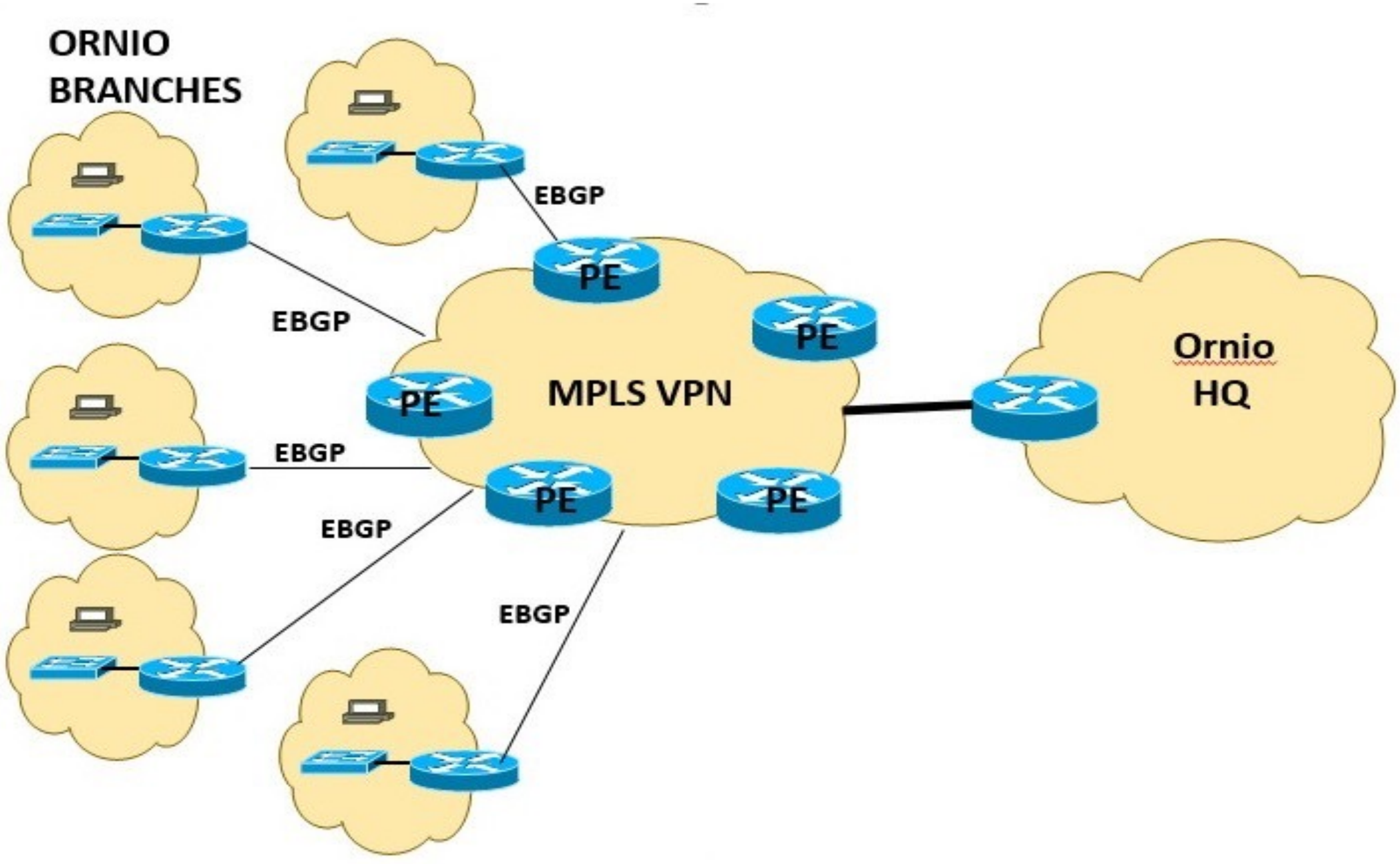
❖ *Although we don’t have the superior expertise on BGP, we followed your advice and chose BPG as a WAN protocol. Our service provider could offer EIGRP as WAN protocol, a protocol that we are very comfortable with. As you had indicated, BGP will provide us the traffic engineering capability.*

https://t.me/learningnets

- *However, we want to implement a new multicast application over our Wide Area Network.*

- *We will have 2 Media server located at the Datacenter and all the branch offices should receive the information at the same time. We don’t need to have any multicast stream from the branch offices towards Head Quarter.*

https://t.me/learningnets




https://t.me/learningnets

- *As per our discussion with the Service Provider, they are providing a Multicast service over the MPLS layer 3 VPN.*

- *We heard that Multicast PIM-BIDIR minimizes the resource requirement on the routers but we want to get your advice.*

- *Best regards*

- *Levent Tokgoz*

- *Network and Security Director*

- *Ornio Group*

- *Maslak/Instabul*

- *26674*

https://t.me/learningnets

| | PIM SSM | PIM ASM | PIM Bidir |
|---|---|---|---|
| Minimum Amount of State in the Router | | | |
| Most optimal Routing | | | |
| Requires Randevous Point | | | |
| Works with IGMPv2 | | | |
| Randevous Point Load Balancing | | | |

https://t.me/learningnets

## ➢ Question 18: Should Ornio use PIM-bidir between headquarter and the remote offices over the MPLS Layer 3 VPN?

❑ Yes

❑ No

https://t.me/learningnets

## ➢ Question 19: Why should Ornio not use PIM-bidir for its multicast traffic?

1. Its application is point to multipoint, so PIM SSM or ASM is more suitable.

2. Service provider cannot have Data MDT for their Rosen GRE multicast implementation.

3. Not every router supports PIM-bidir in Ornio's network.

4. Pim bidir requires much more configuration compared to PIM ASM and PIM SSM.

https://t.me/learningnets

# Email-4 is available

- **Email 4**

- **From:** Levent Tokgoz (levent_tokgoz@ornio.com)

- **To:** Orhan Ergun (orhan@orhanergun.net)

- **Subject:** Redundant remote sites

https://t.me/learningnets

*Hi, Orhan,*

*We will not choose PIM-bidir even though we have verified that our devices support it.*
*Thanks for the information.*
*We started to plan next year budget and we want to include the device and link redundancy for the critical branch offices.*
*Please recommend the solution and let me know if you need additional information.*

*Thanks*

Leven Tokgoz

Network and Security Director

Ornio Group
Maslak/Istanbul
26674

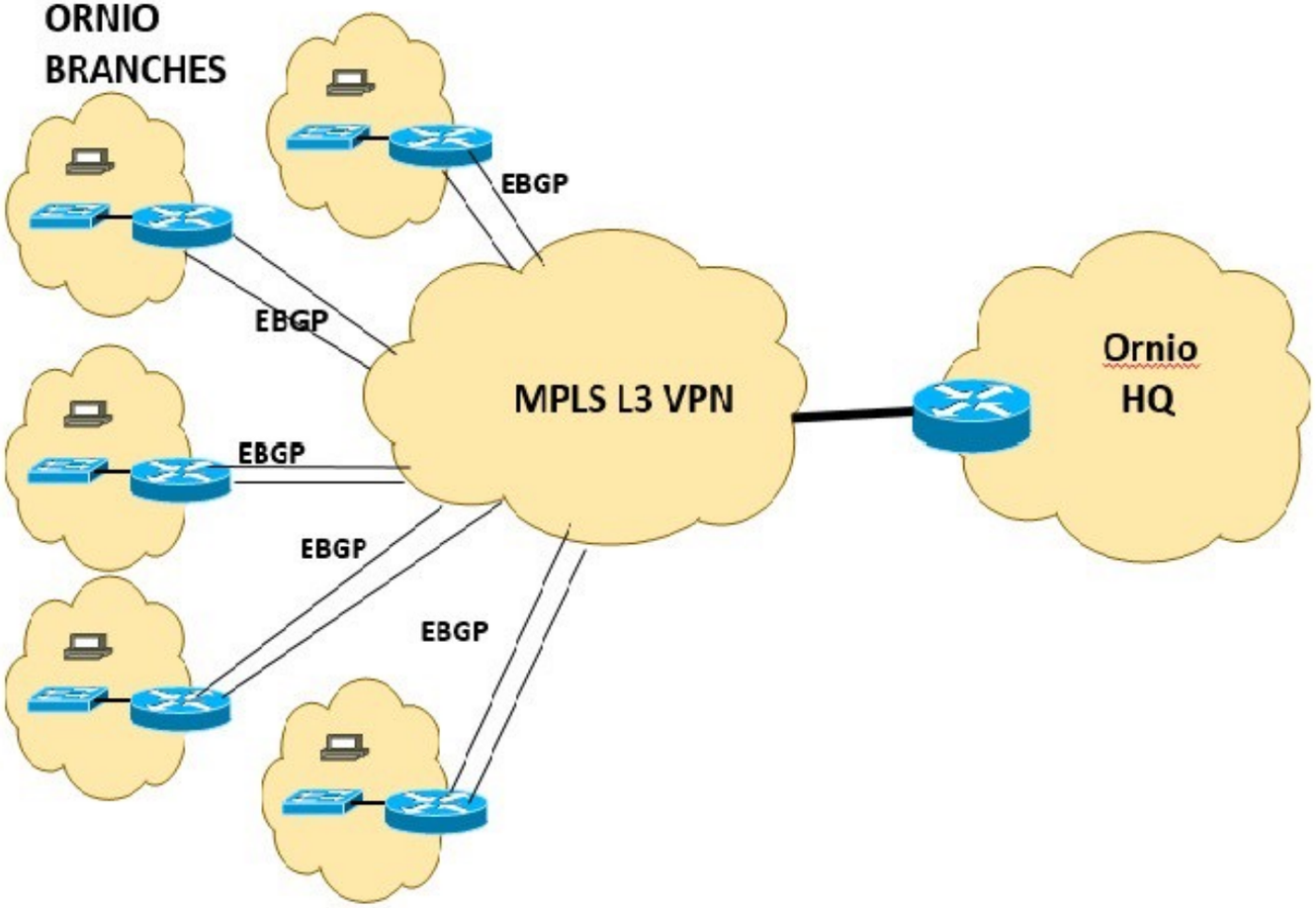
https://t.me/learningnets

## ➢ Question 20: Do you need additional information for Ornio’s future redundancy plan for the branch offices?

❑ Yes

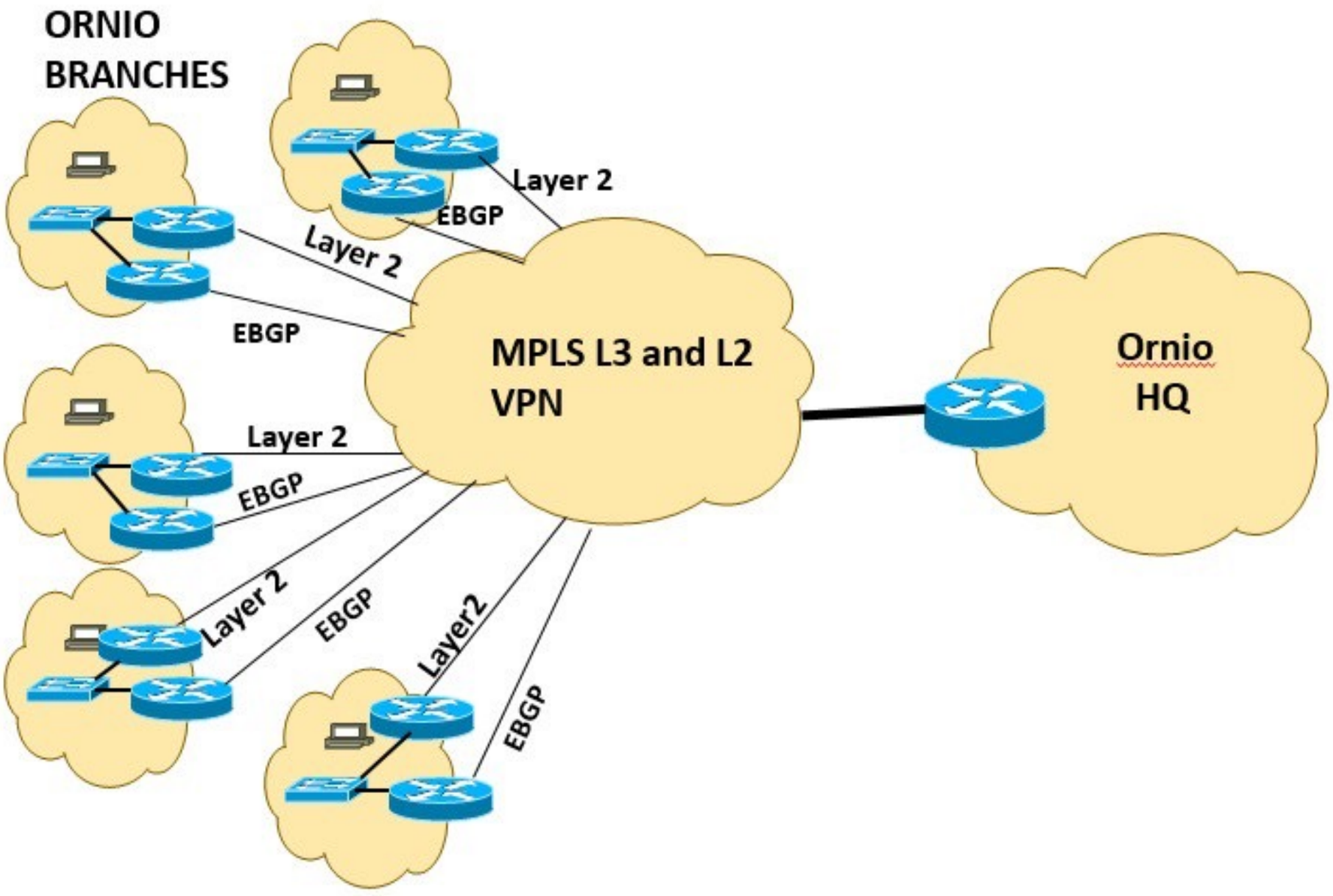
❑ No

https://t.me/learningnets



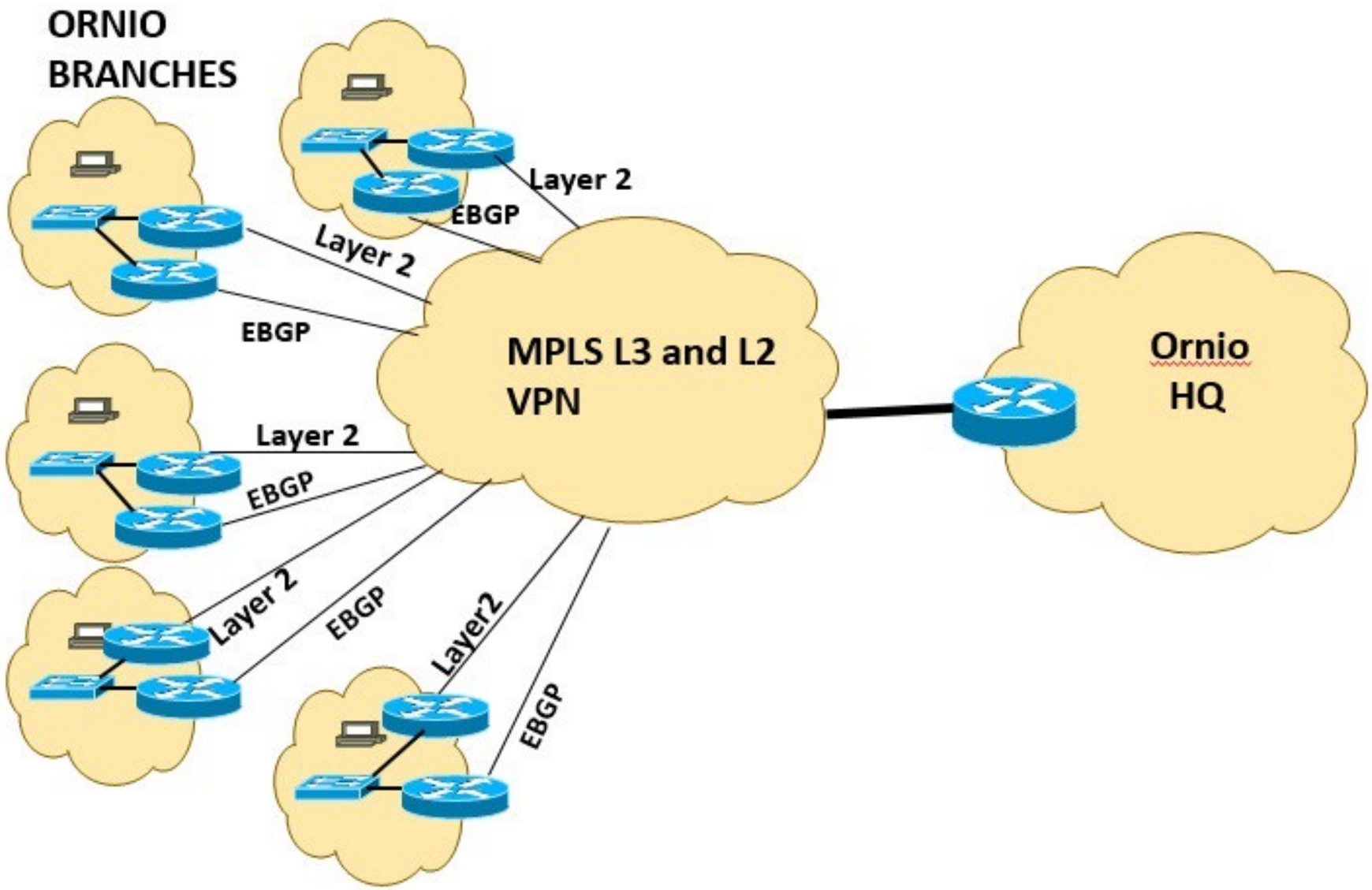

https://t.me/learningnets

# Topology A



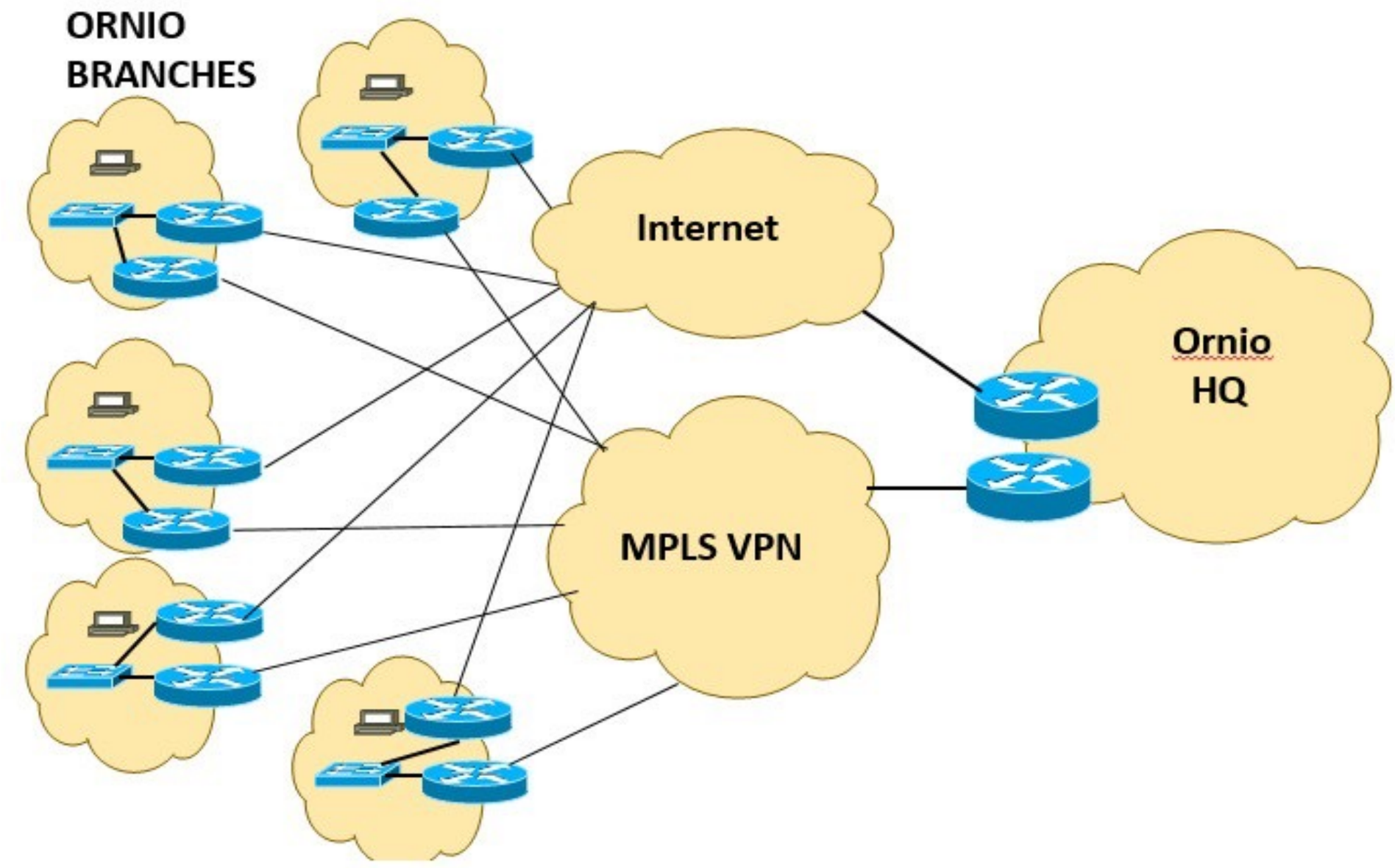

https://t.me/learningnets

# Topology B



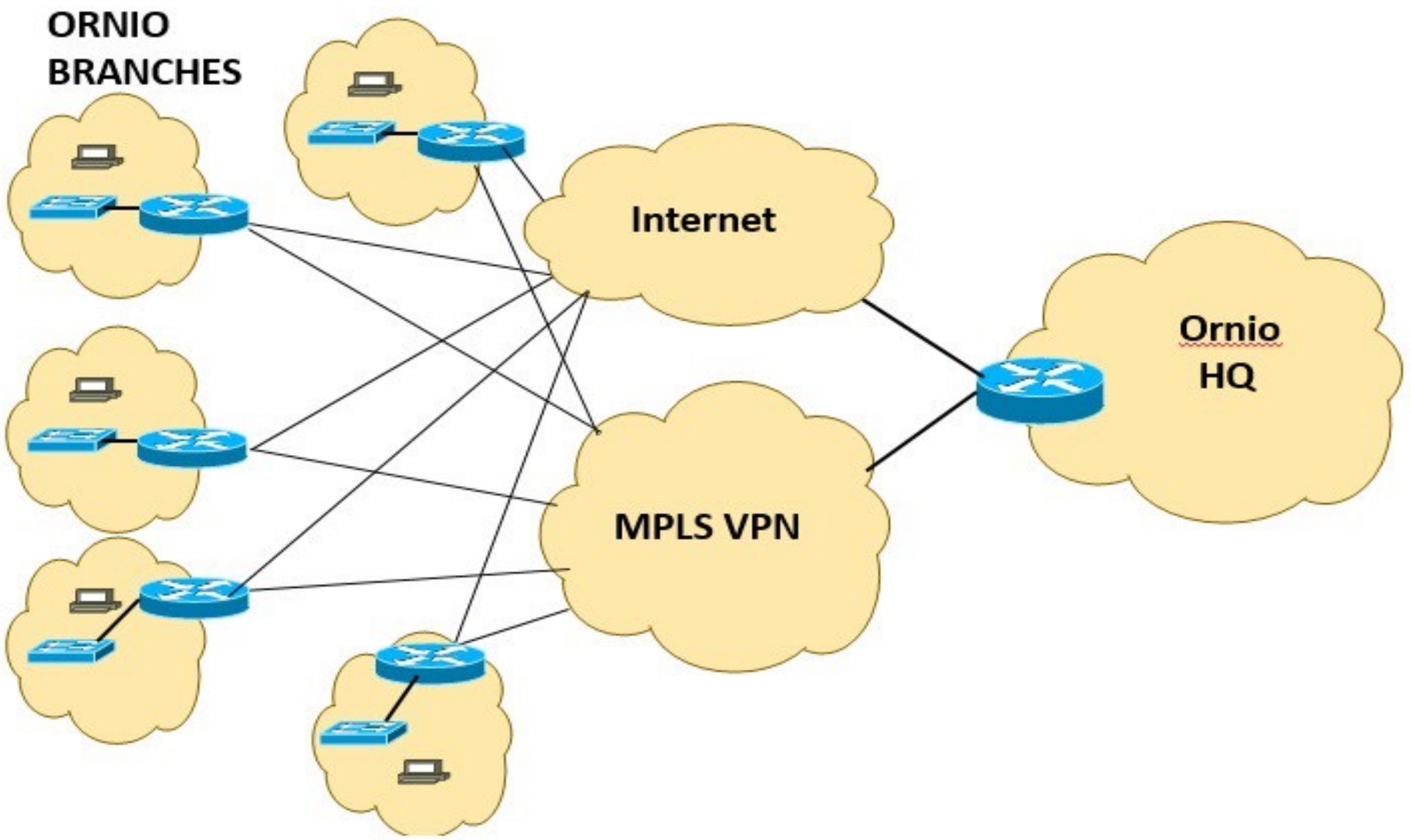

https://t.me/learningnets

# Topology C




https://t.me/learningnets

# Topology D




https://t.me/learningnets

# Answers of the Scenario -1

➢ **Answer 1**

1. Although this is true, by default Frame Relay is not encrypted. In addition, the question did not mention that it has a problem with encryption for voice traffic. That is why this option is not the correct answer.

2. We don’t know this. Service Provider might be offering those SLAs. The voice traffic is the main concern but not mentioned in the scenario. That is why this option is not the correct answer.

https://t.me/learningnets

3. This is definitely true. However, it is indicated clearly in the scenario that its voice traffic on the network is low. This can be a concern for bandwidth hungry applications, but the question is asking Ornio's concern with the voice

4. This is the correct answer. It is clearly mentioned in the scenario that Ornio wants to have direct communication, even if it has direct communication with Frame Relay (Full-mesh)

https://t.me/learningnets

Note:  Although control plane traffic for voice would go to the datacenter, they use IP Telephony. Also, Ornio's call managers are located in the central site/datacenter; hence, data plane traffic (RTP) would follow the direct path.

https://t.me/learningnets

## ➢ Answer 2

- ❖ Yes.
- ❖ If you would say No, you would lose some points here, but the scenario would allow you to continue with the other questions

https://t.me/learningnets

## Answer 3

Answer is MPLS Layer 3 VPN

The detailed explanation will be provided in Answer 4.

https://t.me/learningnets

## ➢ Answer 4

1. By default, MPLS Layer 3 VPN doesn’t provide encryption. Although it can be used in conjunction with IPSEC to provide encryption, Ornio doesn’t look for encryption if it is not over the Internet. This is not the correct answer

2. Yes. MPLS Layer 3 VPN can support Hub and Spoke topologies, although by default it provides full mesh connectivity. But Ornio is looking for direct connectivity between spoke sites (Branch Offices), so either VPLS or MPLS Layer 3 VPN provides this by default. It is very difficult to implement full mesh with the other options such as VPWS due to the number of offices and the number of required connectivity. It is not scalable. This is not the correct answer.

https://t.me/learningnets

3. Yes. This is true and what we are looking for. This is one of the correct answers. All the business requirements are met with this option.

4. Yes. With MPLS layer 3 VPN, customer can outsource its WAN network to the service provider. It is peer-to-peer VPN and the service provider controls the routing. Customer equipment has routing connectivity, including static routing, with the service provider. Ornio is exactly looking for this. Please note that the CE is not managed by the service provider, so even if the customer is managing it, the service provider controls the WAN routing. This is one of the correct answers.

https://t.me/learningnets

5. Yes. This is true but not the correct answer.Layer 3 VPN can provide better, flexible QoS control compared to the Layer 2 VPNs. With Layer 2 VPNs, ingress PE can have its operation based on Layer 2 markings, which is only 3 bits or 8 bits. But if it is Layer 3, then PE can do its operation over 6 bits DSCP values, which can provide 64 different values. Nothing is told for the Quality of service in the scenario so far.

❖ We have two correct answers, 3 and 4

https://t.me/learningnets

## ➢ Answer 5

1. At this point, we are just going to migrate, so Ornio's QoS policy will be considered later on. In order to migrate, we do not need QoS information from neither customer nor service provider site. This information will be required to map QoS classes between them.

2. Its existing and desired application traffic pattern is already given in the scenario. It has Hub and Spoke, but it wants to have full mesh connectivity.

3. Yes, definitely we would like to know this.

4. This is more of a CCDA concern, not for the CCDE exam.

5. Yes. This is important. We need to know the capability of the service provider. Can every branch office get this service? Do we need some sort of Inter-AS agreement after which QoS, multicast might be a concern.

https://t.me/learningnets

## ➢ Answer 6:

❖ Yes

## ➢ Answer 7:

❖ **BGP**

https://t.me/learningnets

➢ **Answer 8:**

❖1.Yes. This is true even though it is not the correct answer. Service providers generally support BGP and static routing as PE-CE routing protocol with their MPLS Layer 3 VPN offering. However, this is not the main concern of Ornio, as it can be understood from its first email.

❖2.This is true as well but not the correct answer. Although there is an attempt to make EIGRP an open protocol – if some critical features such as EIGRP Stub are not open – it cannot be truly open protocol. But since the Ornio's business requirement is not related with this, this is not the correct answer.

https://t.me/learningnets

- 3.Yes. As indicated in Email-1 as well, Ornio wants to have some degree of traffic engineering capability. Only BGP can influence the MPLS VPN traffic. This is the correct answer.

- 4.Yes, BGP can carry millions of prefixes, but Ornio doesn’t have such concerns.

- 5.Yes. BGP can carry non-IP traffic but MPLS VPN only carries IP traffic. This is not the correct answer.

- Only correct answer is Option 3

https://t.me/learningnets

## ➢ Answer 9:

- ❖ Same AS (Autonomous System)
- ❖ The reason will be explained in detail in Answer 10.

https://t.me/learningnets

## ➢ Answer 10:

❖ 1.This is one of the correct answers. Ornio has 800 sites and private BGP AS range (64512 - 65535) is limited to 1000 around. Since we know that Ornio’s operation is growing, in future it will have AS allocation problem. This is why the company should choose the same AS number everywhere.

https://t.me/learningnets

2. If Ornio chooses to deploy unique AS number after 1000 sites, it would need to use public AS numbers. Although this would be a big concern over the global internet, in the context of VPN, using it does not cause harm. This is another correct answer.

3. There is no such restriction. This is not the correct answer.

https://t.me/learningnets

4. Yes. This is one of the correct answers. If Ornio chooses to deploy unique AS number per location, then it has to be careful not to use the same AS number on more than one location.

5. Such practice is not common. This is not the correct answer.

1, 2 and 4 are the correct answers

https://t.me/learningnets

## ➢ Answer 11:

1. This is not true. If BGP is used as PE-CE protocol, there is no need to redistribute it on the PE. Also, this is not a concern of using the same AS number everywhere. If this would be true, it would be the concern of unique AS number per site as well.

https://t.me/learningnets

2. This is one of the correct answers. BGP works this way.

3. There is no such restriction. If BGP PIC for the fast reroute purpose will be enabled, it can be enabled both for the same and unique BGP AS allocation methods.

https://t.me/learningnets

4. This is true even though it is not the correct answer. This is because it is not the problem of using same BGP AS everywhere.

5. This is one of the correct answers. If the same AS is used on every site, then you cannot use AS path attribute for the traffic engineering and path preference.

https://t.me/learningnets

## ➢ Answer 12:

1. This is not the correct answer, since the decision is made. Ornio will use the same AS number everywhere for the above reasons.
2. This is the only correct answer. Service provider can override the customer AS number with its own AS number. At the receiving site, customer will not see similar AS in the accepted path
3. Using MPLS in the service provider is not relevant. It doesn’t create or solve any problem for this issue.

https://t.me/learningnets

4. This is true, but not the correct answer. This is because in the Email-2, customer stated that they don't want to configure a BGP Policy if the same AS creates an issue on every site. Hence, this is not the correct solution. It is against customer's requirements.

5. This is not true, it creates same problem for the single homed site as well.

https://t.me/learningnets

## ➢ Answer 13:

❖ Yes

❖ Detailed explanation will be provided in Answer 15.

https://t.me/learningnets

## ➢ Answer 14:

1. Service provider will add only one additional line command. It can still be considered more complex, but since the question does not state, choose, or apply, this is not the right answer.
2. Yes, multi-homed customer sites have a BGP loop once they are attached to the two PEs. This is the biggest problem and that is why this is the correct answer.

https://t.me/learningnets

3. With the BGP AS Override, nothing is configured at the customer site; thus, BGP AS Override does not create additional complexity at the customer site.

4. No. If all the PEs having a VPN for the customer sites allow BGP AS Override, it is not rejected.

5. BGP AS Override configuration is done on the service provider Edge Routers, not on the BGP route reflector.

https://t.me/learningnets

## ➢ Answer 15:

❖ No. Sometimes in the exam, they may not ask you to state the reasons. It is not asked here as well. However, for the curios people, since all the branch offices as well as headquarter site of Ornio has only one router in the MPLS VPN domain, there is no much concern. There is no dual-homed site.

https://t.me/learningnets

## ➢ Answer 16:

5, 4, 6, 2, 3, 7, 8, 1

https://t.me/learningnets

## ➢ Answer 17:

| | PIM SSM | PIM ASM | PIM Bidir |
|---|---|---|---|
| Minimum Amount of State in the Router | | | x |
| Most optimal Routing | x | | |
| Requires Randevous Point | | x | x |
| Works with IGMPv2 | | x | x |
| Randevous Point Load Balancing | | x | |

There is no rendezvous point load balancing in PIM-bidir. Phantom RP works as an active/standby manner.

https://t.me/learningnets

## ➢ Answer 18: No

## ➢ Answer 19:

1. This is the only correct answer. Ornio has a multicast application. The sender is at the company's headquarter, and all the receivers are at the branches. PIM SSM or PIM ASM is more suitable since shortest path tree will be created. PIM SSM wouldn't require RP at all, so it would be best for this design.

https://t.me/learningnets

2. This statement is definitely correct. Data MDT in Rosen GRE Multicast requires (S, G) multicast state. If customer deploys PIM-bidir for multicast, since there is no (S, G) state in PIM-bidir, data MDT cannot be created. Although this statement is correct, since the data MDT is a service provider concern, not the Ornio's, answer is not what we are looking for.

https://t.me/learningnets

3. We don’t know this. Even though we knew this information, this is not a design concern but more of an operational concern.

4. This is also an operational issue, although complexity is a design attribute. Generally, we need absolute measurement in CCDE exam. We cannot assume.

https://t.me/learningnets

## ➢ Answer 20:

❖ **No**

❖ Detailed answer will be provided in Answer 22

https://t.me/learningnets

## ➢ Answer 21:

- ❖ Answer is in option 4.
- ❖ In the scenario section, you are told that Ornio has local internet connection at remote sites already, and, in the future, they may want to use it.

https://t.me/learningnets

- There is small difference between Answer 4 and 5. In Answer 5, there is only one router at the branches. But in Email-4, you are told that Ornio is looking for device and link level redundancy. Thus, Answer 4 is fit for all the business requirements

https://t.me/learningnets