Computer Networks 2 | CN2

Summary

CONTENTS

1. Preface 6
2. Routing 6
2.1. Administrative Distance (AD) 6
3. Interior Gateway Protocols (IGP) 6
3.1. Open Shortest Path First (OSPF) 7
3.1.1. SPF calculation 7
3.1.2. Network hierarchy 7
3.1.3. Router classification 7
3.1.4. Network types 8
3.1.5. Virtual links 10
3.1.6. Passive Interfaces 10
3.1.7. Link State Advertisement (LSA) Types 10
3.1.8. Route types 13
3.1.9. Area types 13
3.1.10. Flooding 14
3.1.11. Packet format 15
3.1.12. Neighbor states 18
3.1.13. Sub-Protocols 19
3.1.14. Routing computation and Equal-Cost MultiPath 19
3.1.15. Synchronizing the LSDB 21
3.1.16. Extending OSPF 22
3.2. Intermediate System to Intermediate System (IS-IS) 22
3.2.1. Connectionless Network Service (CLNS) 22
3.2.2. Network Service Access Point (NSAP) 24
3.2.3. Similarities between IS-IS and OSPF 24
3.2.4. Advantages of IS-IS over OSPF 24
3.2.5. Integrated IS-IS 25
3.2.6. Router types 26
3.2.7. Adjacencies 27
3.2.8. PDU 29
3.2.9. Hello process 30
3.2.10. Routing 30
4. Border Gateway Protocol (BGP) 32
4.1. Comparison to IGP 32
4.2. Internet Route Aggregation 32
4.3. Autonomous Systems (AS) 32
4.3.1. Autonomous System Numbers (ASN) 32
4.4. Sessions 33
4.4.1. Internal BGP (iBGP) 33
4.4.2. External BGP (eBGP) 35
4.4.3. Combining iBGP and eBGP 35

4.4.4. Multihop Sessions . . . 35
4.5. Messages . . . 36
4.5.1. OPEN . . . 36
4.5.2. KEEPALIVE . . . 36
4.5.3. NOTIFICATION . . . 36
4.5.4. UPDATE . . . 37
4.6. Path Attributes . . . 37
4.7. Path Calculation . . . 37
4.7.1. Route Selection . . . 38
4.8. Loop prevention . . . 38
4.9. Network statements . . . 38
4.10. Route filtering and manipulation . . . 39
4.10.1. Path Announcement . . . 39
4.11. Communities . . . 39
4.12. Connectivity options . . . 39
4.12.1. Single-Homed without BGP . . . 39
4.12.2. Single-Homed with BGP . . . 39
4.12.3. Dual-Homed . . . 40
4.12.4. Multi-Homed . . . 40
4.12.5. Dual Multi-Homed . . . 40
4.12.6. Traffic engineering . . . 40
5. The Internet . . . 41
5.1. Structure . . . 41
5.2. Public IP Address Assignment . . . 41
5.3. Peering vs. Transit . . . 42
5.4. Routing Policies . . . 42
5.5. Internet Exchange Point (IXP) . . . 43
5.6. Routing Security . . . 43
5.6.1. Resource Public Key Infrastructure (RPKI) . . . 43
5.6.2. BGP Monitoring . . . 45
6. Unicast . . . 45
7. Broadcast . . . 45
7.1. Layer 2 - Ethernet Broadcast Frames . . . 45
7.2. Layer 3 - IP Broadcast Packets . . . 45
8. Multicast . . . 46
8.1. Layer 2 - MAC Multicast . . . 46
8.2. Layer 3 - IP Multicast . . . 46
8.2.1. Benefits . . . 46
8.2.2. Inner workings . . . 46
8.2.3. Source . . . 46
8.2.4. Receiver . . . 46
8.2.5. Addressing . . . 47
8.2.6. Internet Group Management Protocol (IGMP) . . . 48
8.3. Protocol Independent Multicast (PIM) . . . 49
8.3.1. PIM Dense Mode . . . 49
8.3.2. PIM Sparse Mode . . . 51
8.3.3. PIM Sparse-Dense Mode . . . 52
8.4. Multicast routing concepts . . . 52

8.4.1. RPF Check . . . 53
8.4.2. The (*,G) multicast routing table entry . . . 53
8.4.3. The (S,G) multicast routing table entry . . . 53
9. Network Design . . . 53
9.1. Availability . . . 54
9.1.1. Reasons for downtime . . . 55
9.1.2. Redundancy . . . 55
9.2. Scalability . . . 56
9.3. Topology . . . 56
9.4. Campus . . . 57
9.4.1. Hierarchical . . . 58
9.4.2. First Hop Redundancy . . . 59
9.4.3. Software Defined Access . . . 60
9.5. Data Center . . . 60
9.5.1. Data Center Tiers . . . 61
9.5.2. Traffic patterns . . . 61
9.5.3. Logical Topology . . . 61
9.5.4. Three-Tier Data Center Architecture . . . 62
9.5.5. Leaf Spine Architecture . . . 62
9.5.6. Top of Rack (ToR) . . . 63
9.5.7. End of Row (EoR) . . . 63
9.5.8. Comparison . . . 64
10. WAN . . . 64
10.1. Private WAN . . . 64
10.1.1. Leased Lines . . . 64
10.1.2. Circuit Switching . . . 65
10.1.3. Packet Switching . . . 65
10.1.4. Connection-oriented vs Connectionless . . . 66
10.2. Public WAN . . . 66
10.2.1. VPN Types . . . 66
10.2.2. Common VPN Protocols . . . 66
10.2.3. Software-Defined WAN (SDWAN) . . . 67
10.3. Multi Protocol Label Switching (MPLS) . . . 67
10.3.1. Router Types . . . 67
10.3.2. Header . . . 68
10.3.3. RIB/FIB and LIB/LFIB . . . 69
10.3.4. Control plane . . . 70
10.3.5. Data plane . . . 70
10.3.6. Label Operations . . . 71
10.3.7. L3 VPN . . . 71
10.3.8. L2 VPN . . . 72
10.4. Internet VPN vs. MPLS VPN . . . 73
11. Overlay Technologies . . . 73
11.1. Generic Routing Encapsulation (GRE) . . . 73
11.2. Segment Routing (SR) . . . 73
11.2.1. Segments . . . 73
11.2.2. Data Plane . . . 74
11.2.3. Operations . . . 74

11.2.4. Control Plane . . . 74
11.2.5. Segment Significance . . . 74
11.2.6. Segment Routing Control Plane Types . . . 74
11.2.7. SR-MPLS . . . 76
12. VXLAN - EVPN . . . 76
12.1. Virtual eXtensible Local Area Network (VXLAN) . . . 77
12.1.1. Issues of traditional L2 Networks . . . 77
12.1.2. Overlay vs Underlay . . . 77
12.1.3. VXLAN Network Identifier (VNI) . . . 77
12.1.4. VXLAN Tunnel Endpoint (VTEP) . . . 77
12.1.5. Frame Format . . . 78
12.1.6. Control plane . . . 78
12.2. Ethernet Virtual Private Network (EVPN) . . . 79
12.2.1. Layer 2 . . . 79
12.2.2. Layer 3 . . . 81
13. Content Delivery Networks (CDN) . . . 83
13.1. Caching . . . 83
13.2. Components . . . 83
13.2.1. Origin Server . . . 83
13.2.2. Edge, Cache and CDN Servers . . . 83
13.2.3. DNS Infrastructure . . . 83
13.3. Key Benefits . . . 83
13.4. Request routing techniques . . . 83
13.4.1. DNS-Based Geo-Routing . . . 83
13.4.2. Anycast and BGP . . . 84
13.5. Caching, Cache-Control and HTTP Headers . . . 84
14. Quality of Service (QoS) . . . 85
14.1. Requirements . . . 85
14.2. Types . . . 85
14.3. Network performance . . . 85
14.3.1. Four main metrics . . . 85
14.3.2. Latency / Delay . . . 85
14.3.3. Jitter (Delay Variation) . . . 86
14.3.4. Packet Loss . . . 86
14.3.5. Bandwidth / Throughput . . . 86
14.3.6. Types of traffic . . . 86
14.3.7. Where queuing takes place . . . 86
14.4. Queuing algorithms . . . 87
14.4.1. First In First Out (FIFO) . . . 87
14.4.2. Priority queuing . . . 87
14.4.3. Round-robin . . . 87
14.4.4. Weighted Round-robin . . . 88
14.4.5. Class-Based Weighed Fair Queuing (CBWFQ) . . . 88
14.4.6. Low Latency Queuing (LLQ) . . . 88
14.5. Queue management . . . 88
14.5.1. Tail dropping . . . 88
14.5.2. TCP . . . 88
14.5.3. Random Early Detection (RED) . . . 89

14.5.4. Weighted Random Early Detection (WRED) . . . 89
14.5.5. Policing . . . 89
14.5.6. Shaping . . . 89
14.6. QoS models . . . 89
14.6.1. Best effort . . . 89
14.6.2. Integrated Services (IntServ) . . . 90
14.6.3. Differential Services (DiffServ) . . . 90
14.7. QoS classification . . . 90
14.7.1. Different Elements . . . 90
14.7.2. Marking . . . 91
14.7.3. Class models . . . 92
14.7.4. Network-Based Application Recognition (NBAR) . . . 92
Bibliography . . . 93

1. PREFACE

Exam is wicked as hell, good luck everyone <3

2. ROUTING

FOSS ftw

2.1. ADMINISTRATIVE DISTANCE (AD)

When multiple sources exist for routing information, such as static routes and BGP, a Cisco router uses the concept of administrative distances to prefer one routing source to the others. The protocol with the lowest administrative distance wins. The accepted best route is then installed in the routing table.

The administrative distances of the most common routing protocols are shown in the table to the right.

When running `show ip route`, the output is structured as follows

`[Protocol] [Type] [Net] [AD/Cost] [Via] [Updated] [If]`

The first number in the brackets (eg. `[160/...]`) is the administrative distance of the information source; the second number (eg. `[.../5]`) is the metric for the route.

Protocol	*AD*
RIP v1 and v2	120
EIGRP Internal	90
EIGRP External	170
OSPF	110
Integrated ISIS	115
BGP Internal	200
BGP External	20
Static to Next Hop	1
Static to Interface	1
Connected	0

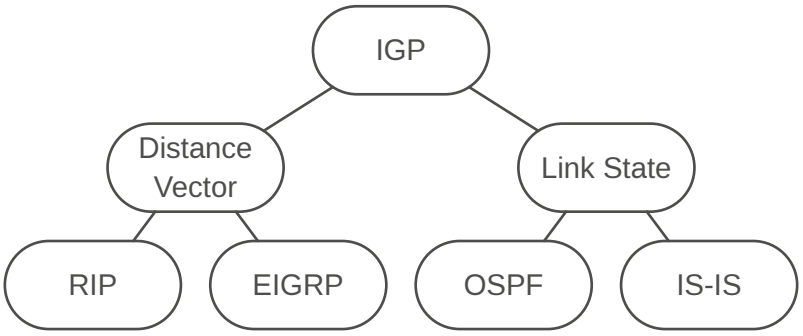
***Example** 1:*

```
Router# show ip route
Codes: I - IGRP derived, R - RIP derived, O - OSPF derived
       C - connected, S - static, E - EGP derived, B - BGP derived
       * - candidate default route, IA - OSPF inter area route
       E1 - OSPF external type 1 route, E2 - OSPF external type 2
       route
Gateway of last resort is 131.119.254.240 to network 129.140.0.0
O E2 150.150.0.0 [160/5] via 131.119.254.6, 0:01:00, Ethernet2
E    192.67.131.0 [200/128] via 131.119.254.244, 0:02:22, Ethernet2
O E2 192.68.132.0 [160/5] via 131.119.254.6, 0:00:59, Ethernet2
O E2 130.130.0.0 [160/5] via 131.119.254.6, 0:00:59, Ethernet2
E    128.128.0.0 [200/128] via 131.119.254.244, 0:02:22, Ethernet2
E    129.129.0.0 [200/129] via 131.119.254.240, 0:02:22, Ethernet2
```

3. INTERIOR GATEWAY PROTOCOLS (IGP)

Establish the global connectivity between routers, within an AS.

Link–state routing protocols use a two-layer area hierarchy composed of one backbone area and multiple regular areas which have to be connected to the backbone area.



3.1. OPEN SHORTEST PATH FIRST (OSPF)

OSPF is an instance of a link state protocol designed for intra-domain routing in an IP network. OSPF gathers link state information from available routers and constructs a topology map of the network. The version of OSPF used in IPv4 networks is known as OSPF version 2 (**OSPFv2**). OSPF for IPv6 networks is known as **OSPFv3**.

3.1.1. SPF calculation

Every time there is a change in the network topology, OSPF needs to reevaluate its shortest path calculations.

- For each **intra-area** topology change, routers **must rerun SPF**.
- An **inter-area** topology change **do not trigger the SPF recalculation**.
  - The router determines the best paths for interarea routes based on the calculation of the best path towards the ABR.
  - The changes that are described in type 3 LSAs do not influence how the router reaches the ABR.

  ⇒ SPF recalculation is not needed.

3.1.2. Network hierarchy

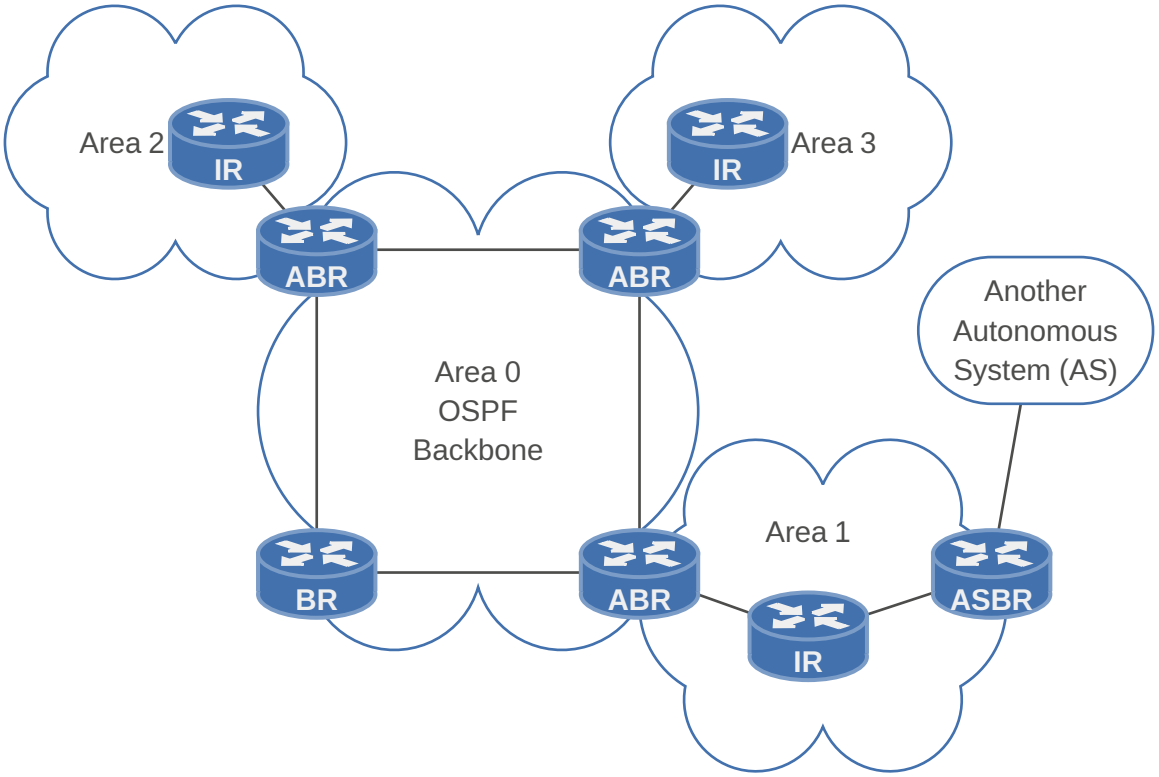
OSPF provides the functionality to divide an intra-domain network into sub-domains (***areas***). Areas are identified through a 32-bit area field. Area ID 0 is the same as `0.0.0.0`. Every intra-domain must have a core area with area **ID 0** (***backbone area***). All other areas connected to the backbone area are referred to as ***low-level areas***. The **backbone area is in charge of summarizing the topology** of one area to another area and vice versa.

3.1.3. Router classification

The routers are classified into four different types according to RFC 2328

Term	*Definition*
Internal Routers (IR)	A router with **all directly connected networks belonging to the same area**. These routers run a **single copy** of the basic routing algorithm.
Area Border Routers (ABR)	A router that **attaches to multiple areas**. Area border routers run multiple copies of the basic algorithm, **one copy for each attached area**. Area border routers condense the topological information of their attached areas for distribution to the backbone. The backbone in turn distributes the information to the other areas.
Backbone Routers (BR)	A router that **has an interface to the backbone area**. This includes all routers that interface to more than one area (i.e., area border routers). However, backbone routers do not have to be area border routers. Routers with all interfaces connecting to the backbone area are supported.

Term	*Definition*
AS Boundary Routers (ASBR)	A router that **exchanges routing information with routers belonging to other Autonomous Systems**. Such a router advertises AS external routing information throughout the Autonomous System. The paths to each AS boundary router are known by every router in the AS. This classification is completely independent of the previous classifications: AS boundary routers may be internal or area border routers, and may or may not participate in the backbone.



3.1.4. Network types

OSPF is designed to address four different types of networks:

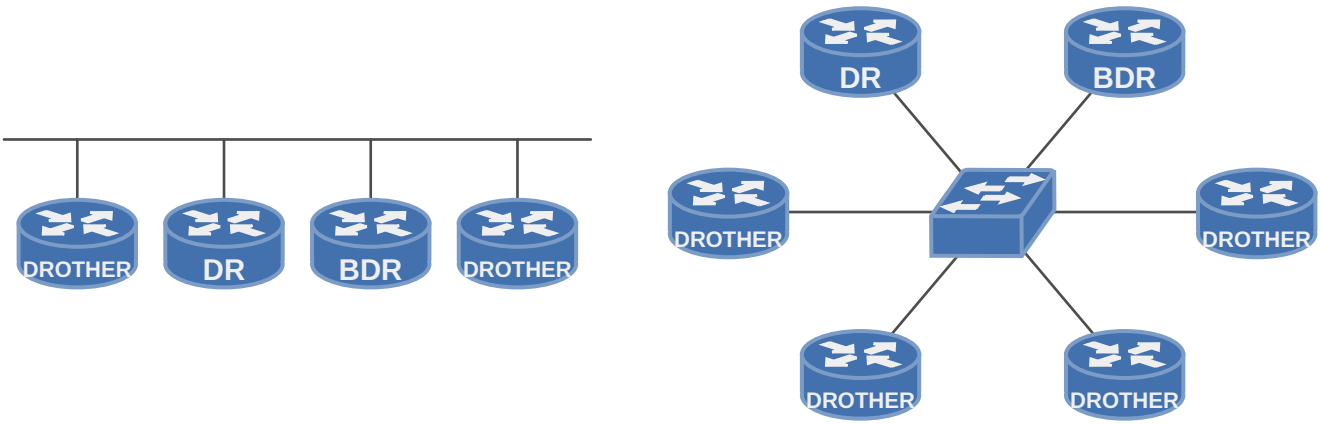
3.1.4.1. Point-to-point networks

Point-to-point networks refer to connecting a pair of routers directly by an interface/link.



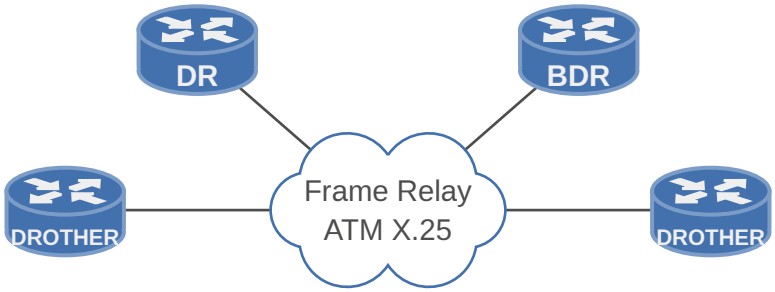
3.1.4.2. Broadcast networks

Broadcast networks are multi-access where all routers in a broadcast network can receive a single transmitted packet. In such networks, a router is elected as a Designated Router (DR) and another as a Backup Designated Router (BDR).



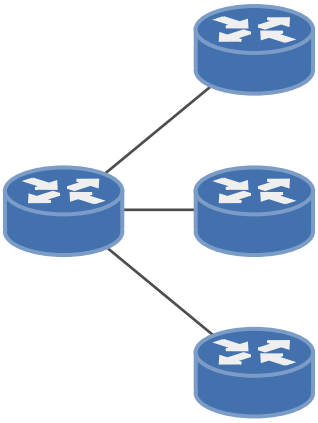
3.1.4.3. Non-broadcast multi-access networks (NBMA)

Non-broadcast multi-access networks (NBMA) are networks where more than two routers may be connected without broadcast capability. Such networks require an extra configuration to emulate the operation of OSPF on a broadcast network. Like broadcast networks, NBMA networks elect a DR and a BDR.



3.1.4.4. Point-to-multipoint networks

Point-to-multipoint networks are also non-broadcast networks much like NBMA networks. However, OSPF's mode of operation is different and is similar to point-to-point links.



3.1.4.5. Optimization on Non-Point-to-Point Networks

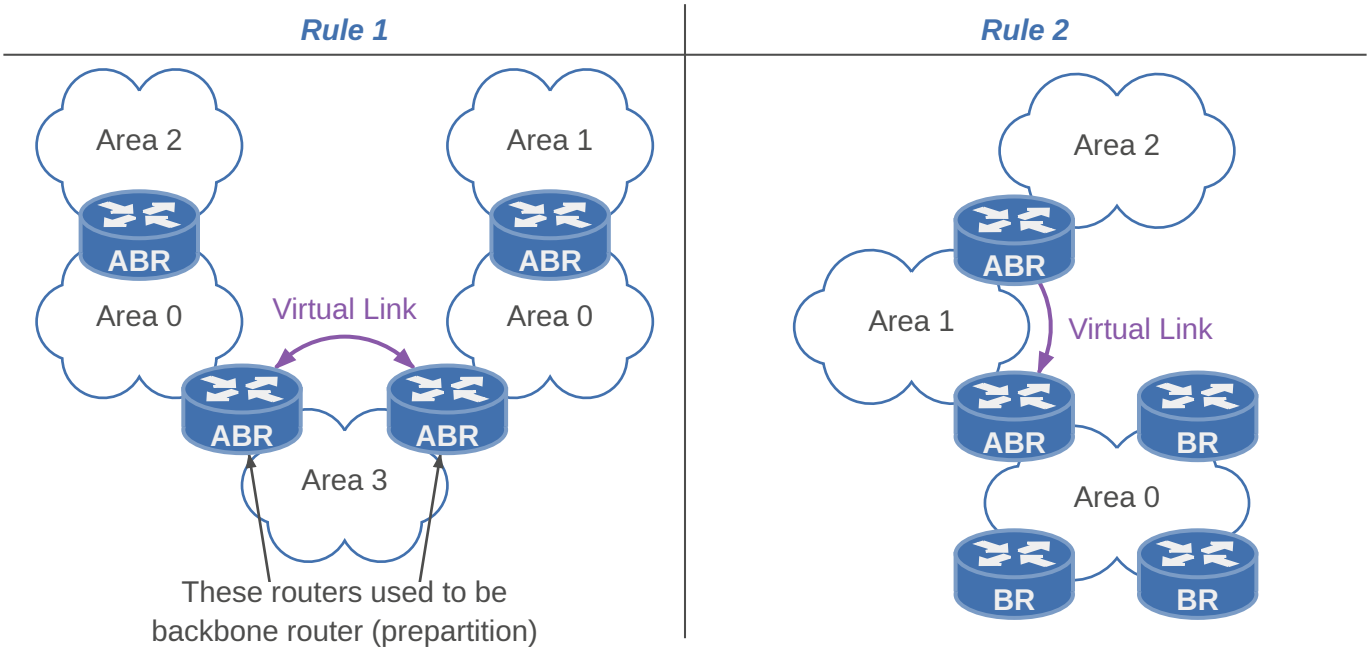
– Designated Router (DR) and Backup Designated Router (BDR) based on priority or router ID
– DR performs the LSA forwarding and LSDB synchronization tasks on behalf of all routers on the broadcast domain
– Each router establishes a FULL adjacency with the DR and the BDR by using the IPv4 multicast address `224.0.0.6`
– The BDR performs the DR tasks only if the DR fails.

3.1.5. Virtual links

– Virtual links cannot go through more than one area.
– Virtual links can only run through standard non-backbone areas. (not over stubby areas for example)

OSPF Design Rule 1: **Area 0 has to be contiguous.** For example, if a backbone is partitioned into two parts due to a link failure, virtual links are used. In such a case, virtual links are tunnelled through a non-backbone area.

OSPF Design Rule 2: **A non-backbone area has to be connected to the backbone area.** Virtual links are used to connect an area to the backbone using a non-backbone (transit) area. Virtual links are configured between two Area Border Routers.



3.1.6. Passive Interfaces

The passive interface is used on interfaces where the router is not expected to form any OSPF neighbor adjacency. On a passive interface, the router **stops sending and receiving OSPF Hello packets**.

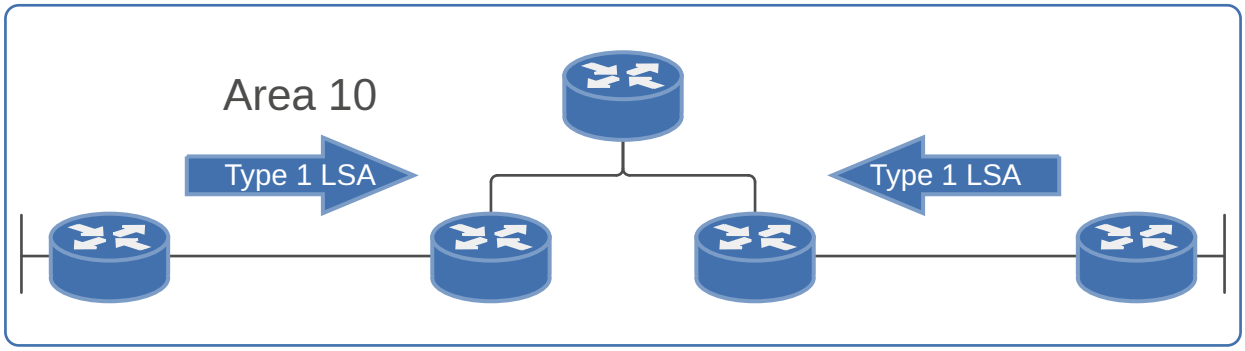
3.1.7. Link State Advertisement (LSA) Types

OSPF floods routing information such as link state advertisements. The scope of flooding of OSPF packets depends on the LSA types. The six different LSA types are:

3.1.7.1. Router LSA (Type 1)

Every router generates a Router LSA that **lists all the routers' outgoing interfaces**. For each interface, the **state and cost of the link** are included. Such LSAs are generated for **point-to-point links**.

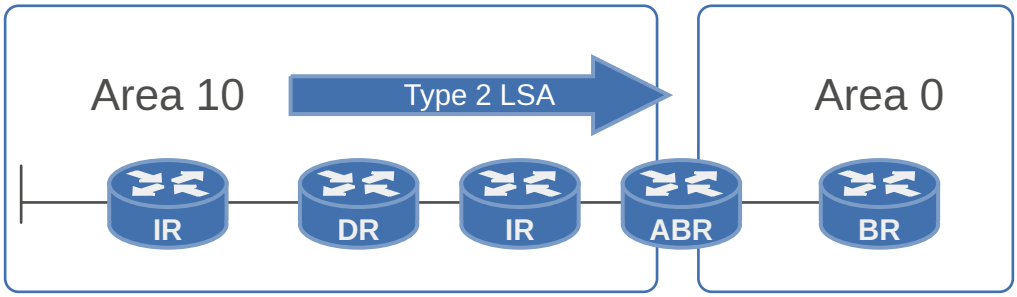
– Type: All routers in Area
– Scope: Flooding of Router LSAs is restricted to the area where they originate. [AREA]



3.1.7.2. Network LSA (Type 2)

Network LSAs are applicable in broadcast and non-broadcast networks where they are **generated by the DR**. A Network LSA **represents a LAN**. All attached routers and the DR are listed in the Network LSA.

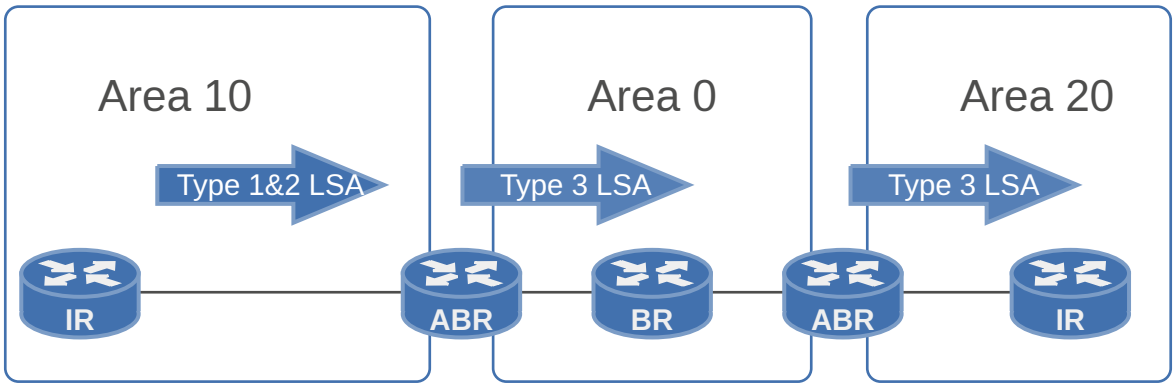
– Type: Only DRs
– Scope: Flooding of Network LSAs is also restricted to the area where they originate. [AREA]



3.1.7.3. Network Summary LSA (Type 3)

Area Border Routers (ABR) generate Network Summary LSAs that are used for **advertising destinations outside an area**.

– Type: Only ABRs
– Scope: Flooded in all the areas that are not totally stubby. [AREA]

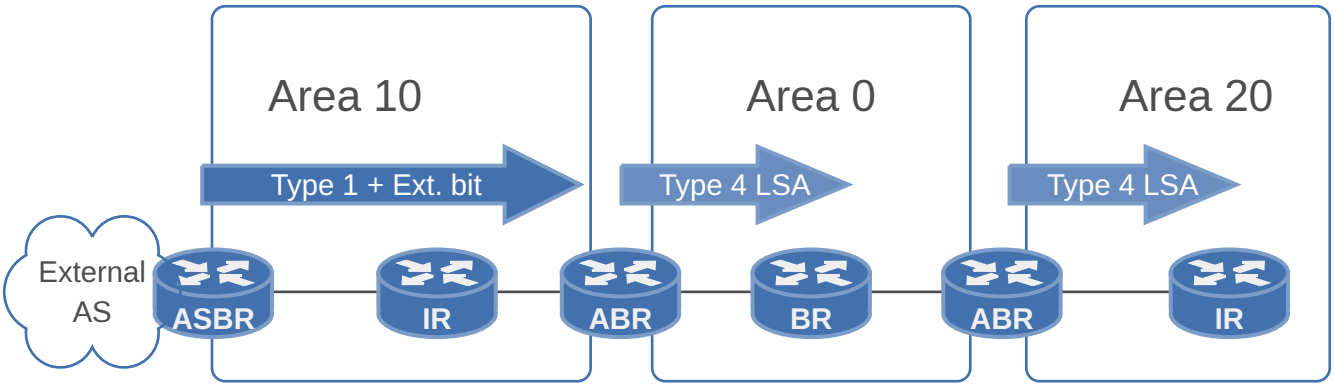


3.1.7.4. ASBR Summary LSA (Type 4)

Identifies the ASBR and provides a route to the ASBR. All traffic that is destined to an external AS requires routing table knowledge of the ASBR that originated the external routes.

The **ASBR sends a type 1 router LSA** with a bit (known as the ***external bit***) that is set to identify itself as an ASBR. When the ABR (identified with the border bit in the router LSA) receives this type 1 LSA, the **ABR builds a type 4 LSA** and floods it to the subsequent areas.

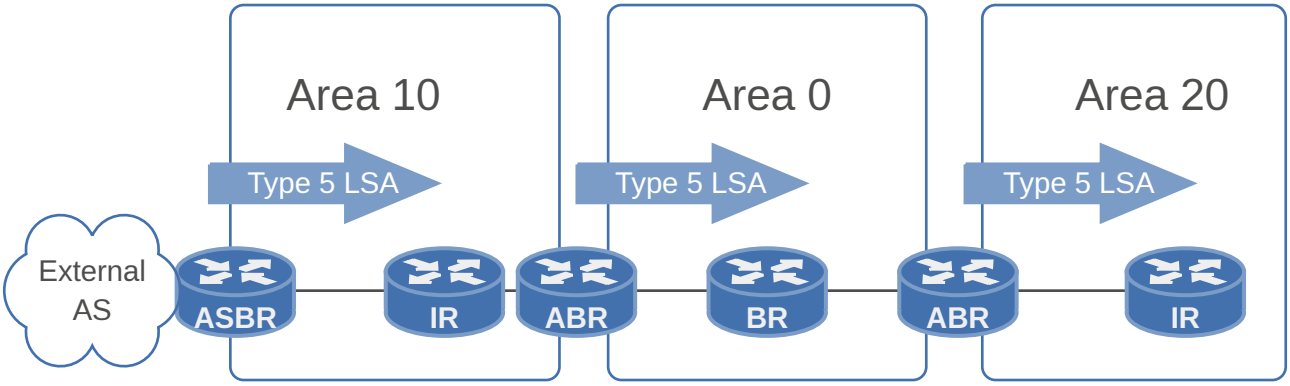
– Type: Only ABRs
– Scope: [AREA]



3.1.7.5. AS External LSA (Type 5)

AS External LSAs are **generated by ASBRs** and propagate the external networks within the OSPF domain. Destinations external to an OSPF AS are advertised using AS external LSAs.

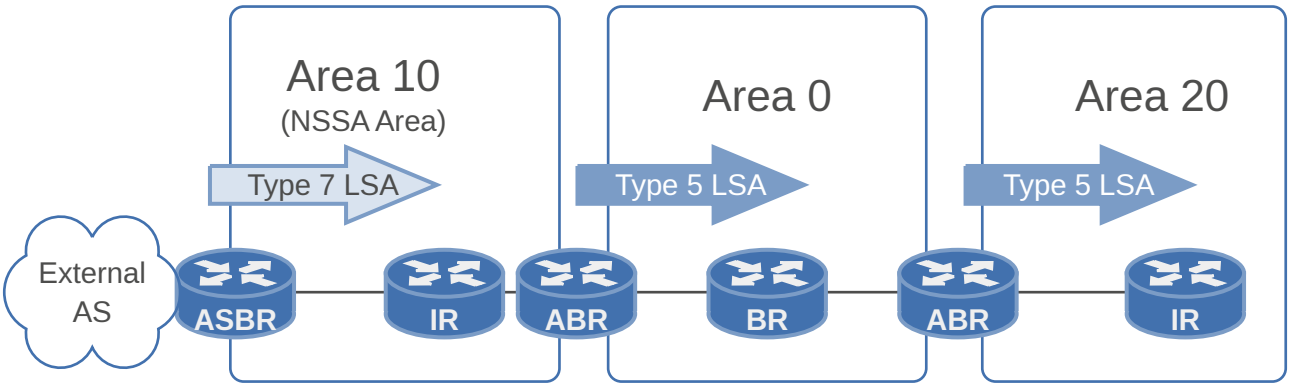
– Type: ASBR and ABR
– Scope: AS external LSAs are flooded in all the areas that are neither stub nor totally stubby. [DOMAIN]



3.1.7.6. External LSA (Type 7)

Also contain external networks within the OSPF domain. **NSSA areas do not allow type 5 external LSAs**.

– Type: ASBRs
– Scope: [AREA]

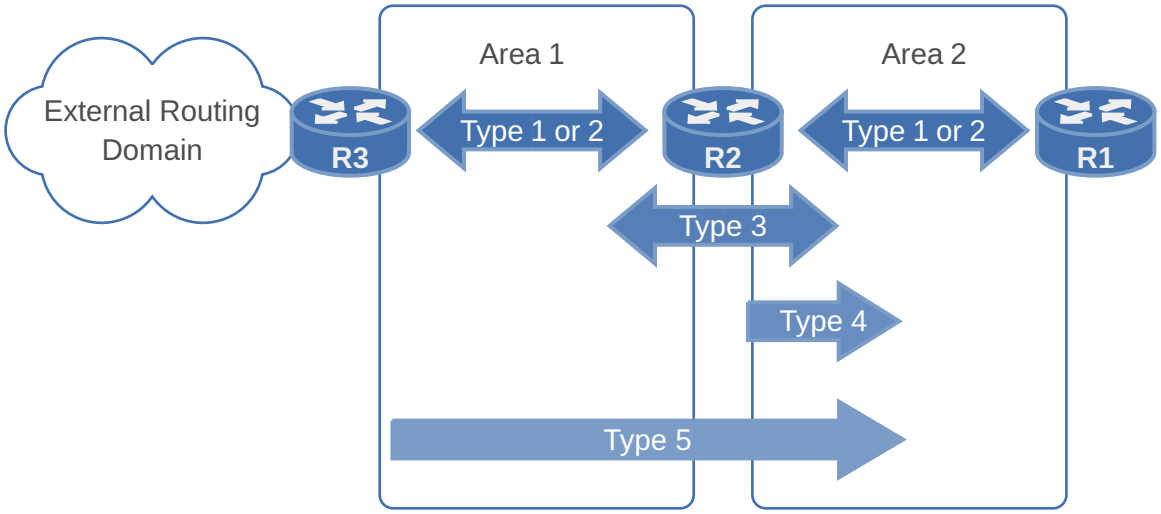


3.1.7.7. Summary

OSPF Accepted LSAs per Area Type

	LSA 1	*LSA 2*	*LSA 3*	*LSA 4*	*LSA 5*	*LSA 7*
Backbone Area	✓	✓	✓	✓	✓	×
Non-Backbone Area	✓	✓	✓	✓	✓	×

	LSA 1	LSA 2	LSA 3	LSA 4	LSA 5	LSA 7
Stub Area	✓	✓	✓	×	×	×
Totally Stubby Area	✓	✓	×	×	×	×
Not-So-Stubby Area	✓	✓	✓	×	×	✓
Totally Not-So-Stubby Area	✓	✓	×	×	×	✓



3.1.8. Route types

Term	Definition
Intra-area routes	Are originated and learned in the same local area **(O)**
Inter-area routes	Originate in other areas and are inserted into the local area to which your router belongs **(O IA)**
External routes	**(O E1 or O E2)**

3.1.9. Area types

Term	Definition
Backbone Area	– Has to be connected to all of the other areas. – Must always be contiguous – Generally, end users are not found within a backbone area
Stub Area	– Eliminates all external routes (type 5) – Creates a default route – Cannot contain ASBRs
Standart Area	– All routes present
Totally Stubby Area	– Eliminates all external routes (type 5) – Eliminates all inter-area routes (type 3) – Creates a default route – Cannot contain ASBRs

Term	*Definition*
Not so Stubby Area (NSSA)	– Allows injection of external routes into a stub area – Eliminates all external routes* (type 5) – No default route – Creates own area type 7 LSA * advertises redistributed external routes as LSA type 7 in the area itself. When traversing to a different OSPF area, it transforms them to a regular type 5 LSA
Totally Not so Stubby Area (Totally NSSA)	– Eliminates all external routes* (type 5) – Eliminates all inter-area routes (type 3) – Creates a default route – Creates own area type 7 LSA * advertises redistributed external routes as LSA type 7 in the area itself. When traversing to a different OSPF area, it transforms them to a regular type 5 LSA

3.1.9.1. Stub areas and stub networks

– Stubby = Eliminates all external routes (type 5)
– Totally stubby = Additionally eliminates all inter-area routes (type 3)

3.1.10. Flooding

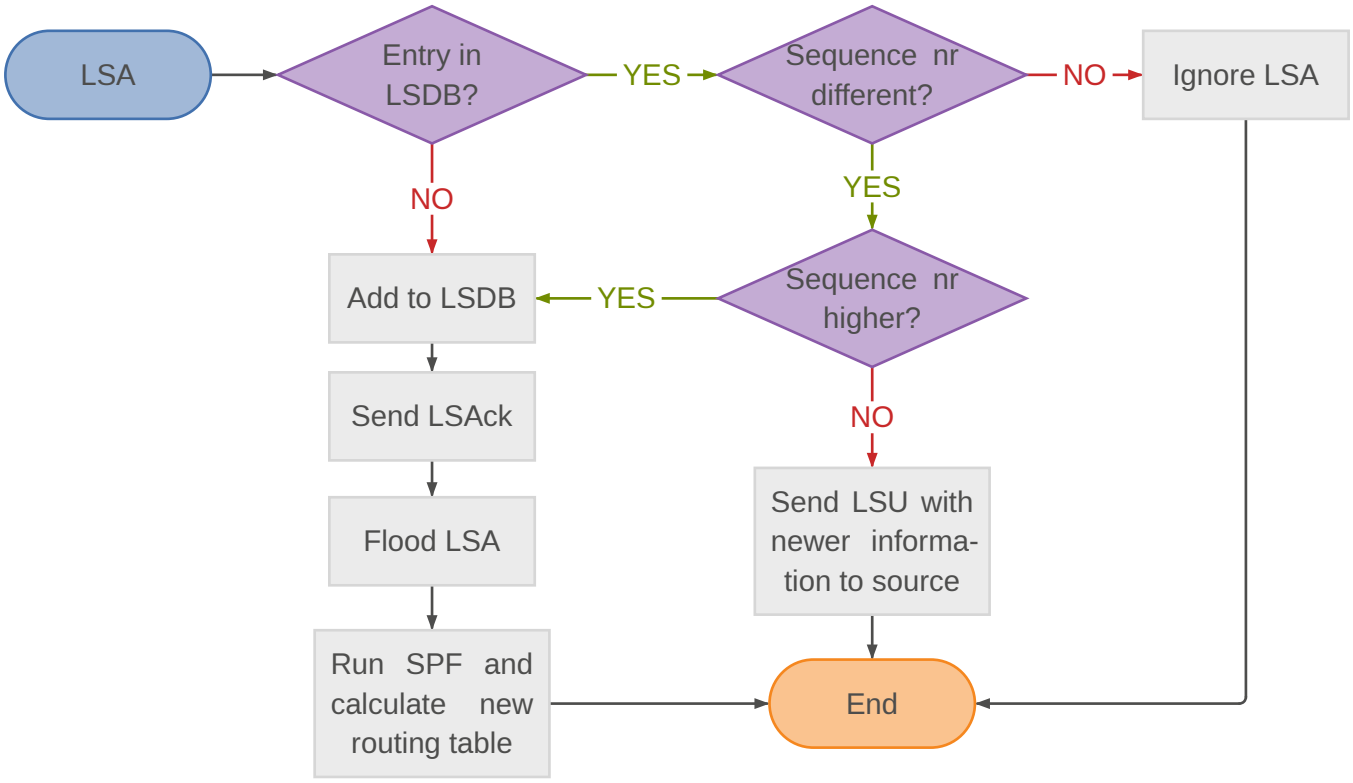
OSPF sits directly on top of IP in the ***TCP/IP stack*** by using the IP protocol number 89. OSPF packets use the **multicast destination MAC address `224.0.0.5`**. OSPF is required to provide its own reliable mechanism, instead of being able to use a reliable transport protocol such as TCP. OSPF addresses reliable delivery of packets through use of either an implicit or explicit acknowledgment.

Implicit acknowledgment: A duplicate of the LSA as an update is sent back to the router from which it received the update.

Explicit acknowledgment: The receiving router sends a link state acknowledgment packet on receiving a link state update.

Since a router may not receive acknowledgment from its neighbor to whom it sent a link state update message, a router is required to **track a link state retransmission list** of outstanding updates.

– An **LSA is retransmitted**, always as unicast, on a periodic basis **until an acknowledgment is received**, or the adjacency is no longer available.
– A router **floods all its LSAs every 30 minutes**, regardless of whether the content of the LSA such as the metric value has changed. Hence, the Link State Database (LSDB) is always synchronized between all routers in an area



3.1.11. Packet format

OSPF has 5 packet types:

Packet Type	*Function*	*Transmission Mode*	*Address*
Hello	Discovery/Maintain	Usually Multicast	`244.0.0.5`*
DBD	Database Summary	Unicast	Neighbor IP
LSR	Request LSA	Unicast	Neighbor IP
LSU	Link State Update	Multicast/Unicast	`244.0.0.5`/.6* or Unicast
LSAck	Acknowledgement	Multicast/Unicast	`244.0.0.5`/.6* or Unicast

* `244.0.0.5` for all routers, `244.0.0.6` for DR/BDR

3.1.11.1. Hello Packet (Hello)

The primary purpose of the hello packet is to **establish and maintain adjacencies.** The hello packet is also used in the **election process of the Designated Router and Backup Designated Router** in broadcast networks. Moreover, it is used for negotiating optional capabilities.

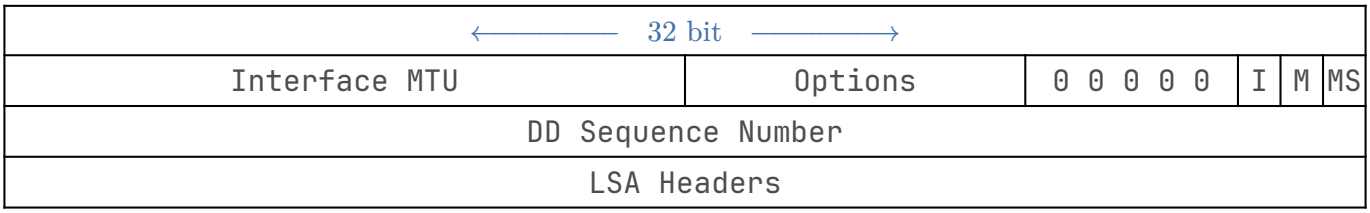
⟵ 32 bit ⟶		
Network Mask		
Hello Interval	Options	Priority
Router Dead Interval		
Designated Router		
Backup Designated Router		
Neighbors (4 bytes each)		

Field	*Definition*
Network Mask (32 bit)	Address mask of the router interface from which this packet is sent.
Hello Interval (16 bit)	Time difference in seconds between any two hello packets. The sending and the receiving routers are required to maintain the same value. Otherwise, a neighbor relationship is not established. For point-to-point and broadcast networks, the default value is `10s`, for other network types it's `30s`.
Options (8 bit)	Allow compatibility with a neighboring router to be checked.
Priority (8 bit)	Used when electing the DR and the BDR.
Router Dead Interval (32 bit)	Length of time in which a router will declare a neighbor to be dead if it does not receive a hello packet. Needs to be larger than the hello interval. The neighbors also need to agree on the value of this parameter. A routing packet that is received and does not match this field on a receiving router's interface is dropped. The default value is four times the default value for the hello interval (`40s` and `120s`).
Designated Router (32 bit)	DR/BDR field lists the IP address of the interface of the DR/BDR on the network, but not its router ID. If the DR/BDR field is `0.0.0.0`, this means that there is no DR/BDR.
Neighbors (4 bytes each)	This field is repeated for each router from which the originating router has received a valid Hello recently, meaning in the past Router Dead Interval.

3.1.11.2. Database Description Packet (DBD)

The database description packet contains a **summary of all the LSAs** (not the entire LSAs) that the neighboring router has in its LSDB. It includes:

– Link-state type
– Address of advertising router
– Link-cost
– Sequence number



Field	*Definition*
Interface MTU (16 bit)	Indicates the size of the largest transmission unit the interface can handle without fragmentation.
Options (8 bit)	Consist of several bit-level fields. The most interesting one is the E-bit which is set when the attached area is capable of processing AS-external-LSAs.
I (1 bit)	I-bit (initial-bit) is initialized to 1 for the initial packet that starts a database description session; for other packets for the same session, this field is set to 0.

Field	*Definition*
M (1 bit)	M-bit (more-bit) is used to indicate that this packet is not the last packet for the database description session by setting it to 1; the last packet for this session is set to 0.
MS (1 bit)	MS-bit (master-slave bit) is used to indicate that the originator is the master by setting this field to 1, while the slave sets this field to 0.
DD Sequence Number (32 bit)	Used for incrementing the sequence numbers of packets from the side of the master during a database description session. The master sets the initial value for the sequence number.
LSA Headers (32 bit)	Lists headers of the link state advertisements in the originator's link state database.

3.1.11.3. Link State Request Packet (LSR)

– Typically triggered after DBD
– Requests specific LSAs from neighbors (unicast)

The link state request packet is used for **pulling information**. Once the database description has been received from a neighbor, a router knows which LSAs are not in its LSDB and will request the entire missing LSAs from that neighbor. The fields are repeated for each unique entry:

⟵ 32 bit ⟶
Link State Type
Link State ID
Advertising Router

Field	*Definition*
Link State Type (32 bit)	Identifies a link state type such as a router or network.
Link State ID (32 bit)	Dictated by the link state type.
Advertising Router (32 bit)	Address of the router that has generated this LSA.

3.1.11.4. Link State Update Packet (LSU)

This packet is the **answer to a Link State Request** Packet. It contains the first field to be the number of LSAs followed by information on LSAs that match the LSA packet format. A link state update packet can contain one or more LSAs. It ensures that all routers have same view.

– Implicit acknowledgement
– Flooding of LSAs (multicast)
– Sending LSA responses to LSRs (unicast)

⟵ 32 bit ⟶
Number of LSAs
LSAs
...
LSAs

3.1.11.5. Link State Acknowledgement Packet (LSAck)

– Explicit acknowledgement

– Make LSA flooding reliable (multicast)
– Acknowledging direct LSU (unicast)

Each newly received LSA must be acknowledged. This is usually done by sending Link State Acknowledgment packets. However, acknowledgments can also be accomplished implicitly by sending Link State Update packets.

Many acknowledgments may be **grouped together** into a single Link State Acknowledgment packet. Such a packet is sent back out the interface which received the LSAs.

A Link State Acknowledgment Packet contains a regular OSPF header with the type field set to 5 and a set of one or more LSA headers as payload.

<table>
<tr><td colspan="3">← 32 bit →</td></tr>
<tr><td>Version</td><td>Type</td><td>Packet Length</td></tr>
<tr><td colspan="3">Router ID</td></tr>
<tr><td colspan="3">Area ID</td></tr>
<tr><td colspan="2">CheckSum</td><td>AuType</td></tr>
<tr><td colspan="3">Authentication</td></tr>
<tr><td colspan="3">Authentication</td></tr>
<tr><td colspan="3">LSA Headers...</td></tr>
</table>

3.1.12. Neighbor states

1) ***Down:*** Initial state of a neighbor conversation when no “Hellos” packets have been received from the neighbour. If a router doesn’t receive a hello packet from a neighbor within the RouterDeadInterval time, then neighbour state changes from full to down.
2) ***Init:*** The router has received a Hello from the neighbor but has not yet seen its own router ID in the neighbor Hello packet.
3) ***2-Way:*** The router has seen its own router ID in the Hello packet received from the neighbor. The DR and BDR election is taking place if necessary.
4) ***Exstart:*** Neighboring routers establish a master/slave relationship and determine the initial database descriptor (DBD) sequence number to use while exchanging DBD packets.
5) ***Exchange:*** The routers exchange DBD packets, which describe their entire link-state database.
6) ***Loading:*** routers exchange full Link State information based on DataBase Descriptor (DBD) provided by neighbors, the OSPF router sends Link State Request (LSR) and receives Link State Update (LSU) containing all Link State Advertisements (LSAs).
7) ***Full:*** Normal operating state that indicates everything is functioning normally. In this state, routers are fully adjacent with each other and all the router and network Link State Advertisements (LSAs) are exchanged and the routers’ databases are fully synchronized.

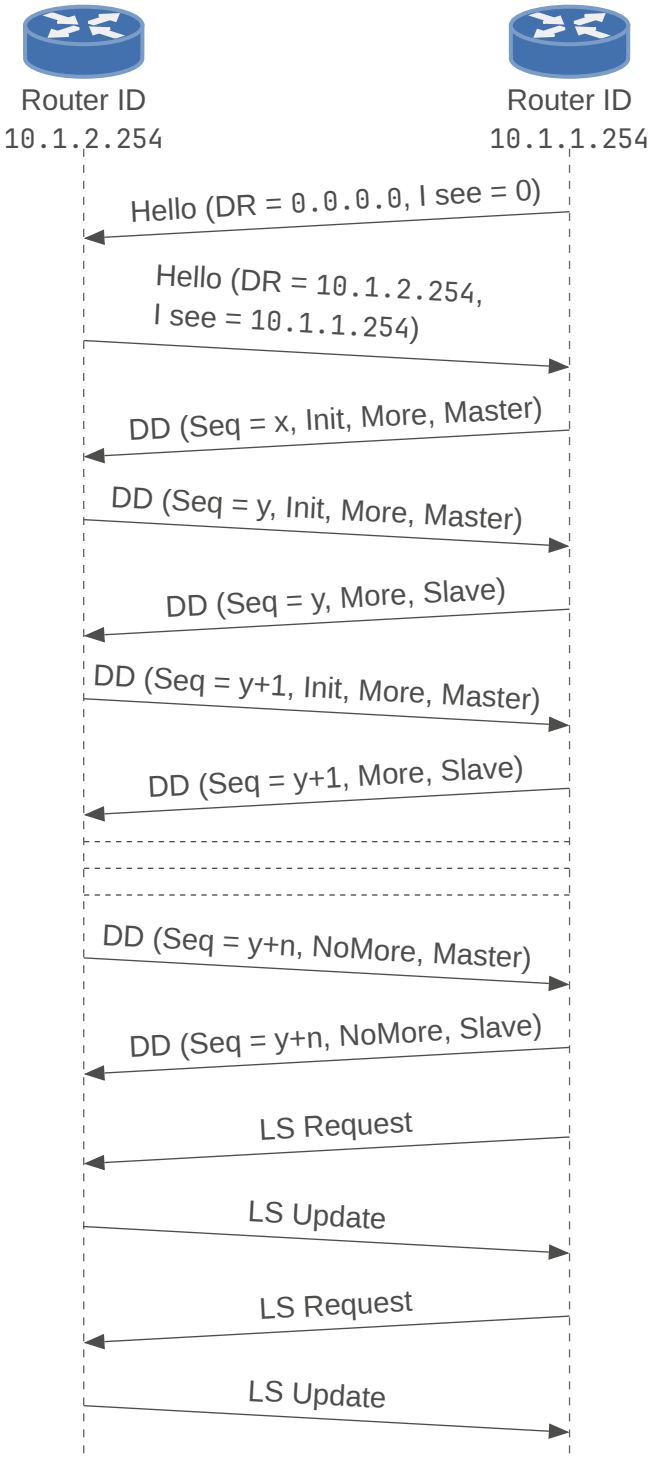
3.1.13. Sub-Protocols

3.1.13.1. Hello Protocol

During initialization/activation, the hello protocol is used for **neighbor discovery** as well as to **agree on several parameters** before two routers become neighbors.

– When using the hello protocol, **logical adjacencies are established** for point-to-point, point-to-multi-point, and virtual link networks.
– For broadcast and NBMA networks, not all routers become logically adjacent. The hello protocol is used for **electing Designated Routers and Backup Designated Routers**.

After initialization, for all network types the hello protocol is used for **keeping alive connectivity** which ensures bidirectional communication between neighbors. If the keep alive hello messages are not received within a certain time interval that was agreed upon during initialization, the link/connectivity between the routers is assumed to be not available.



3.1.13.2. Database Synchronization Protocol

Beyond basic initialization to discover neighbors, two adjacent routers need to build adjacencies. A complete link state advertisement of all links in the database of each router can be exchanged, but a special database description process is used to optimize this step. During the database description phase, only **headers of link state advertisements are exchanged**. Headers serve as adequate information to check if one side has the latest LSA. Since such a synchronization process may require exchange of header information about many LSAs, the database synchronization process allows for such exchanges to be **split into multiple chunks**.

These chunks are communicated using database description packets by indicating whether a chunk is an **initial packet (using I-bit)**, or a **continuation/more packet or last packet (using M-bit)**. One side needs to serve as a **master (MS-bit)** while the other side serves as a **slave**. The neighbor with the lower router ID becomes the slave.

3.1.14. Routing computation and Equal-Cost MultiPath

Default costs:

Bandwidth (b/s)	*Cost*
128K	781

Bandwidth (b/s)	*Cost*
10M	10
100M	1

LSAs type 1 and 2 are flooded throughout an area. This allows every router in an area to build link state databases with identical topological information.

– **Shortest path computation** based on Dijkstra's algorithm is performed **at each router for every known destination** based on the directional graph determined from the link state database.
– The link cost used for each link is the metric value advertised in the link state advertisement packet. The value can be between 1 and 65'535
– Dijkstra-based shortest path computation using link state information is applied only within an area. **For routing updates between areas**, information from one area is summarized using **Summary LSAs** without providing detailed link information.

The next hop is extracted from the shortest path computation to update the routing table and subsequently, the forwarding table.

– Routing table entries are for destinations identified through hosts or subnets or simply IP prefixes with CIDR notation, not in terms of end routers.
– Because of CIDR, multiple similar route entries are possible, eg. `10.1.64.0/24` vs `10.1.64.0/18`. To select the route preferred by an arriving packet, OSPF uses a best route selection process (**most specific match**).
– In case there are multiple paths available after this step, the second step selects the route where an **intra-area path is given preference** over an inter-area path, which in turn gives preference over external paths for routes learned externally.

3.1.14.1. Equal-cost multipath (ECMP)

ECMP means that if two paths have the same lowest cost, then the outgoing link (next hop) for both can be listed in the routing table so that traffic can be equally split. The original Dijkstra's algorithm generates only one shortest path even if multiple shortest paths are available. To capture multiple shortest paths, where available, Dijkstra's algorithm is slightly modified.

The router implementation handles the ECMP path selection on a per-flow basis rather than on a per-packet basis. The ECMP path selection is based on the hash of certain fields of the IP packet without having to maintain states at routers.

3.1.14.2. Route selection

When using OSPF routing hierarchy, the following rules apply:

– If the **source and destination** addresses of a packet reside within the **same area, intra-area routing is used**. Intra-area routes in OSPF are described by router (type 1) and network (type 2) LSAs. When displayed in the OSPF routing table, these types of intra-area routes are designated with an O.
– If the source and destination addresses of a packet reside within **different areas, but are still within the AS, inter-area routing is used**. These types of routes are described by network summary (type 3) LSAs. When routing packets between two nonbackbone areas, the backbone is used. This means that inter-area routing has pieces of intra-area routing along its path, for example:
  1) An intra-area path is used from the source router to the area border router.
  2) The backbone is then used from the source area to the destination area.
  3) An intra-area path is used from the destination area's area border router to the destination.

When you put these three routes together, you have an inter-area route. Of course, the SPF algorithm calculates the lowest cost between these two points. When displayed in the OSPF routing table, these types of routes are indicated with an O IA.

– If the destination address of a packet resides **outside the AS, external routing is used**. External routing information are injected into OSPF through redistribution from another routing protocol. The AS boundary routers (ASBRs) flood the external route information throughout the AS. Every router receives this information, with the exception of stub areas. The types of external routes used in OSPF are as follows:
  – E1 routes: E1 route's costs are the sum of internal and external (remote AS) OSPF metrics. If a packet is destined for another AS, an E1 route takes the remote AS metric and adds all internal OSPF costs. They are identified by the E1 designation within the OSPF routing table.
  – E2 routes: E2 routes are the default external routes for OSPF. They do not add the internal OSPF metrics. Multiple routes to the same destination use the following order of preference: intra-area, inter-area, E1, and E2.

OSPF will first look at the "type of path" to decide and secondly look at the metric. On equal types path cost will decide by the preferred path list for OSPF:

1) Intra-Area (O)
2) Inter-Area (O IA)
3) External Type 1 (E1)
4) NSSA Type 1 (N1)
5) External Type 2 (E2)
6) NSSA Type 2 (N2)

3.1.14.3. Route summarization

Route summarization helps solve two major challenges

– Large routing tables
– Frequent LSA flooding throughout the autonomous system

With route summarization, the ABRs or ASBRs consolidate multiple routes into a single advertisement

– Route summarization requires a good addressing plan
– Subnets in areas should be assigned contiguously to ensure that these addresses can be summarized into a minimal number of summary addresses

Summarization is only allowed on ASBRs and ABRs.

ABRs	***ASBRs***
Summarize type 3 LSAs	Summarize type 5 LSAs
An internal summary route is generated if at least one subnet within the area falls in the summary address range The summarized route metric is equal to the lowest cost of all the subnets within the summary address range	Summarization of external routes can be done on the ASBR for re-distributed routes before injecting them into the OSPF domain

3.1.15. Synchronizing the LSDB

1) A DROTHER router notices a change in a link state and sends an LSU packet (which includes the updated LSA entry) to the OSPF DR at multicast address `224.0.0.6`
2) The DR acknowledges receipt of the change and floods the LSU to others on the multiaccess network using the OSPF multicast address `224.0.0.5`
3) After receiving the LSU, each router responds to the DR with an LSAck

4) The router updates its LSDB using the LSU that includes the changed LSA Non-DR exchange their databases only with the DR

3.1.16. Extending OSPF

– Classical OSPF is not easy to extend to add new features
  – They require the creation of a new LSA
  – OSPF version 2 was developed exclusively for IPv4
  – RFC 7684 introduces Opaque LSAs

3.2. INTERMEDIATE SYSTEM TO INTERMEDIATE SYSTEM (IS-IS)

– Widely used (especially in ISP networks / as an intra-domain routing protocol)
– Fast convergence
– Equal Cost Multipath (ECMP) Load Balancing
– IS-IS supports different protocol suites RFC 1195
– Originally developed for ISO OSI environments (CLNS) but integrated IS-IS can be used to support pure-IP or dual environments. In modern IP-only environments:
  – There are no CLNP-based user applications.
  – The routing process is the primary user of the underlying CLNS mechanisms.
  – The only ISO packets typically observed are ES-IS and IS-IS control messages.
  – CLNP node-based addresses are still used to identify routers

Term	*Definition*
ES	End-host devices are called End Systems (ES)
IS	Routers are called Intermediate Systems (IS)
IIH	IS-IS Hello Packets
TLV	Type Length Value

3.2.1. Connectionless Network Service (CLNS)

Different to the TCP/IP suite, the OSI architecture has a strict distinction between services and protocols. ***Services*** are defined as the **functions provided by one layer to the layer above it**, while ***protocols*** are the **specific implementations of these services** In the OSI model, each layer provides services to the layer above it and relies on services from the layer below it.

	OSI Model	*TCP/IP Model*
L3 Service	CLNS (Connectionless Network Service)	No separate name
Service Type	Connectionless	Connectionless
Data-Plane Protocol	CLNP (Connectionless Network Protocol)	IP (Internet Protocol)
Control-Plane Protocol	IS-IS / ES-IS	OSPF / IS-IS / BGP / etc.
Addressing	NSAP (Network Service Access Point)	IP Address

3.2.1.1. Protocol Suite

The connectionless network service defined by CLNS is realized and supported within the ISO architecture by several protocols, including:

CLNP: The network-layer data protocol

ES-IS: The host-to-router discovery protocol

IS-IS: The router-to-router routing protocol

CLNP, ES-IS, and IS-IS are specified as separate network layer protocols, coexisting at Layer 3 of the OSI reference model.

3.2.1.2. Connectionless Network Protocol (CLNP)

The CLNP is the **OSI equivalent of the IP**.

– Both are **connectionless**.
– Both provide **best-effort delivery**.
– Both **rely on separate routing protocols for path calculation**.

CLNP operates at Layer 3 of the OSI model and provides connectionless, best-effort packet delivery between systems.

CLNP provides network-layer services to ISO transport protocols, rather than to TCP and UDP as in the TCP/IP architecture.

At the data-link layer, CLNP packets are identified by the Ethernet protocol type: 0xFE FE.

3.2.1.3. End System to Intermediate System (ES-IS)

Operates between hosts (End Systems) and routers (Intermediate Systems) in an ISO CLNS environment. Its primary function is **adjacency and reachability discovery** within a shared network segment (for example, a LAN). ES-IS automates the exchange of addressing and presence information between connected systems.

The protocol operates using two message types:

ESH: End System Hello – transmitted by hosts

ISH: Intermediate System Hello – transmitted by routers

These messages allow systems to discover neighboring devices and their network layer addresses. From a functional perspective, ES-IS can be loosely compared to the combined roles of:

– ARP (address resolution)
– ICMP (reachability signaling)
– DHCP (host configuration assistance)

When IS-IS is configured on certain router platforms, ES-IS functionality operates automatically in the background to support adjacency formation.

3.2.1.4. Intermediate System to Intermediate System (IS-IS)

IS-IS is a link-state routing protocol operating between routers (Intermediate Systems). Originally developed for routing CLNP traffic within ISO CLNS networks, IS-IS **dynamically exchanges topology and reachability information** between routers. It **builds a link-state database and computes shortest paths** using the SPF algorithm.

IS-IS operates in conjunction with ES-IS:

– ES-IS supports neighbor discovery.
– IS-IS establishes and maintains router adjacencies.
– IS-IS distributes routing information across the routing domain.

On multi-access networks, routers learn the data-link addresses (for example, MAC addresses, also referred to as SNPAs Subnetwork Points of Attachment) of adjacent systems and store this information in the adjacency database.

3.2.2. Network Service Access Point (NSAP)

An NSAP address identifies a network-layer entity within the OSI architecture. It is the **CLNS equivalent of an IP address**, although its structure differs significantly.

An NSAP address:

– Is variable in length (up to 20 bytes)
– Is hierarchically structured
– Identifies a system (node) rather than a specific interface
– Is used by routing protocols such as IS-IS

IDP		*DSP*		
AFI	**IDI**	**High-Order DSP**	**System ID**	**NSEL**
Variable-Length Area Address			6 Bytes	1 Byte

AFI (Authority and Format Identifier): Indicates the format of the NSAP address and the authority that assigned it.

IDI (Initial Domain Identifier): Variable length, identifies the administrative domain or organization responsible for the address.

DFI (Domain Specific Part Format Identifier): Specifies the format of the domain-specific part of the address.

DSP (Domain Specific Part): Variable length, contains the hierarchical structure of the address, which can include area identifiers and system identifiers.

3.2.3. Similarities between IS-IS and OSPF

– Standardized
– Link-state protocol
– Similar sync mechanism
– Use Dijkstra SPF algorithm
– Similar update/flooding process
– Quick convergence

3.2.4. Advantages of IS-IS over OSPF

– Detects a failure faster
– Simpler than OSPF
– Well-positioned for IPv6
– Scalability
  – Less “chatty”
  – TLVs instead of new LSPs
– IS-IS operates directly over the data link layer
  – No technical need for IP to establish adjacencies
  – Security: immune to remote IP-based attack vectors. packets are directly encapsulated over the data link and are not carried in IP packets or even CLNP packets. Therefore, to maliciously disrupt the IS-IS routing environment, an attacker has to be physically attached to a router in the IS-IS network
– Multiple protocols supported, but treated as metadata attributes

ISIS considered to be more scalable and better suited for large and complex networks. OSPF might struggle with very large networks, especially in one single area.

– IS-IS: groups updates into one LSP
– OSPF: many small LSA updates

3.2.5. Integrated IS-IS

Although IS-IS was not originally designed for IP routing, it was later extended to support IP networks. This extension is commonly referred to as ***Integrated IS-IS***. In modern IP-only environments:

– There are no CLNP-based user applications.
– The routing process is the primary user of the underlying CLNS mechanisms.
– The only ISO packets typically observed are ES-IS and IS-IS control messages.

3.2.5.1. Addressing

$$\underbrace{49.0011}_{\text{Area ID (11)}}.\underbrace{0000.0000.0003}_{\text{System ID (R3)}}.\underbrace{00}_{\text{NSEL}}$$

The following list consists of requirements and caveats that must be followed to define NSAP for IS-IS routing in general and particularly on Cisco routers:

– Each node in an IS-IS routing area must have a **unique SysID**
– The SysID of all nodes in an IS-IS routing domain must be of the **same length**
– The length of the SysID is 6 bytes (fixed) on Cisco routers

You can use one of the LAN MAC addresses on a router as its SysID, essentially embedding a MAC address (a Layer 2 address) in the NSAP. Another popular way to define unique SysIDs is by padding a dotted-decimal loopback IP address with zeros to transform it into a 12-digit address, which can then be easily rearranged to represent a 6-byte SysID in hexadecimal, by regrouping the digits in fours and separating them with dots.

***Example** 2:*

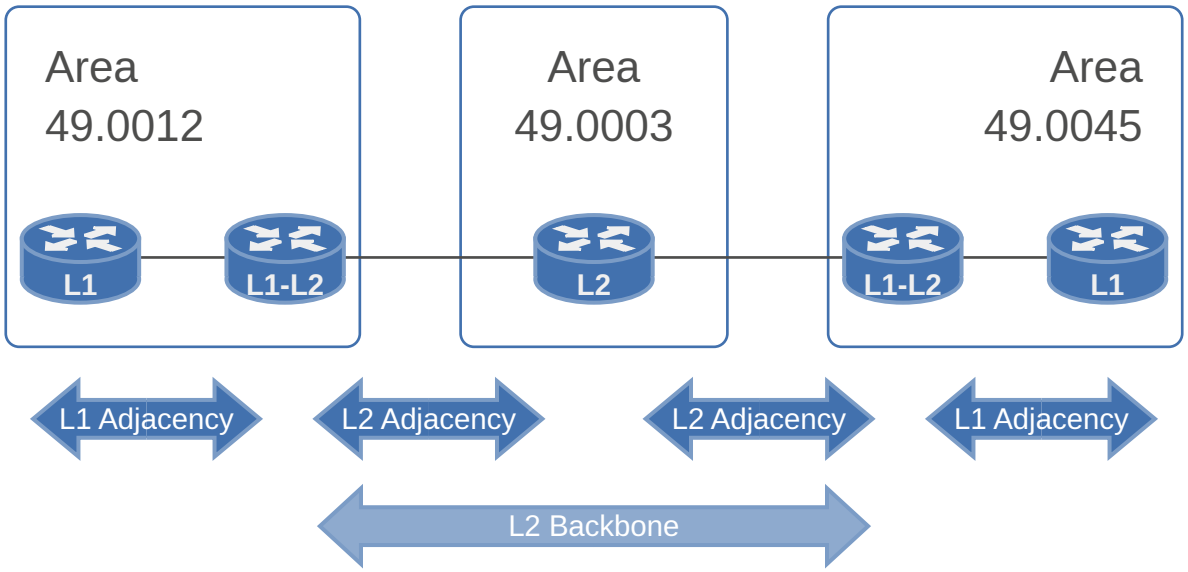
```
# Interface Loopback 0: IP address 192.168.1.24

Router(config)# router isis
Router(config-router)# net 49.0001.1921.6800.1024.00
```

3.2.5.2. Network Entity Title (NET)

– Identifies talking to the router itself (not to a specific application)
– NSAP address with an NSEL (NSAP Selector) of 0
– Included in LSP header
– Area part starts with 49
  – Authority and Format Identifier
  – Stands for private/local address

3.2.6. Router types



3.2.6.1. Level 1 router (L1)

– Knows the topology only of its own area
– All L1 routers have same LSPDB within area

L1 routers form **adjacencies only with other L1 routers** that belong to the same area. During the hello process, the routers verify that their Area IDs match. **If the Area IDs differ, no L1 adjacency is established**.

After adjacency establishment, L1 routers exchange Level-1 Link-State Packets (L1 LSPs). In order to exchange routing information, IS-IS uses LSPs (Link State Packet) which is similar to OSPF's LSAs. These LSPs contain:

– Information about directly connected neighbors within the area
– Reachable IP prefixes within the area
– Associated metrics

Through reliable flooding, each L1 router distributes its LSPs to all other routers in the same area so that all L1 routers have a consistent view of the area topology.

Based on the synchronized L1 LSDB, each router independently runs the SPF algorithm to compute optimal paths to all destinations within the area.

3.2.6.2. Level 2 router (L2)

– Knows about other areas
– All L2 routers have same LSPDB

An L2 router operates at the inter-area level and is responsible for routing **between different IS-IS areas**.

After adjacency establishment, Level-2 routers exchange Level-2 Link-State Packets (L2 LSPs). These LSPs contain:

– Information about neighboring Level-2 routers
– Reachable prefixes from their attached areas
– Associated metrics

Through reliable flooding, all Level-2 routers build a synchronized Level-2 LSDB that represents the inter-area topology.

Based on the synchronized Level-2 LSDB, each router independently runs a separate Level-2 SPF calculation.

3.2.6.3. Level 1/Level 2 router (L1-L2)

– Has a Level 1 link-state database for intra-area routing and a Level 2 link-state database for interarea routing
– L1-L2 routers maintain a separate L1 and L2 LSPDB

A L1-L2 router operates **simultaneously at both routing levels** and acts as the border router between an area and the Level-2 backbone, enabling communication between different IS-IS areas.

The router participates **independently in both flooding domains**, Level-1 LSPs within its local area and Level-2 LSPs across the backbone.

A L1-L2 router also runs two independent SPF calculations for both areas.

3.2.7. Adjacencies

An adjacency must be in an **up state** for a router to send or process received LSPs:

– A Level 1 adjacency is formed when the area addresses match unless configured otherwise.
– A Level 2 adjacency is formed alongside the Level 1 unless the router is configured to be Level 1-only.
– If no matching areas exist between the configuration of the local router and the area addresses information in the received hello, only a Level 2 adjacency is formed.
– If the transmitting router is configured for Level 2-only, the receiving router must be capable of forming a Level 2 adjacency. Otherwise, no adjacency forms.

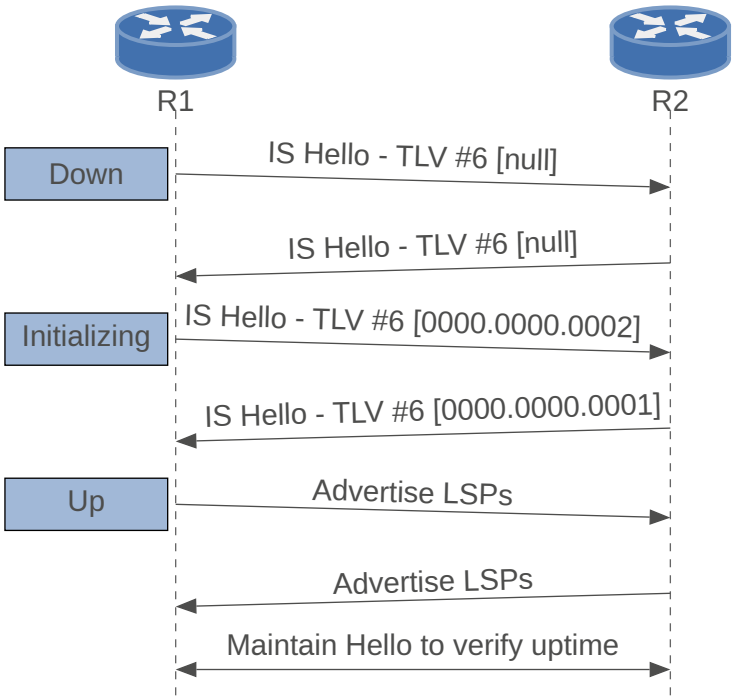
When designing IS-IS networks, always remember that the **backbone must be contiguous**. In other words, a Level 1-only router should never be inserted between any two Level 2 routers (Level 2-only or Level 1-2).

3.2.7.1. Point-to-point

IS-IS adjacencies on point-to-point links are **initialized by receipt of ISHs** through the ES-IS protocol. This is followed by the **exchange of point-to-point IIHs**. In the default mode of operation, **IIHs are padded to the MTU** size of the outgoing interface. Routers match the size of IIHs received to their local MTUs to ensure that they can handle the largest possible packets from their neighbors before completing an adjacency.

– An unnecessary pseudonode LSP is not included in the LSPDB of all routers in that level.
– CSNPs are not continuously flooded into a segment
– CSNPs are sent only once during start

3.2.7.2. Adjacency states for broadcast network segments



3.2.7.3. Multiaccess

The process of building adjacencies is **not triggered by receipt of ISHs**. A router sends IIHs on broadcast interfaces as soon as the interface is enabled.

Routers include the **MAC addresses of all neighbors** on the LAN that they have received hellos from, allowing for a simple mechanism to confirm two-way communication.

– Two-way communication is confirmed when subsequent hellos received contain the receiving router's MAC address (SNPA) in an IS Neighbors TLV field.
– Otherwise, communication between the nodes is deemed one-way, and the adjacency stays at the initialized state.

The broadcast medium is modeled as a node, called the pseudonode. The pseudonode role is played by an elected DIS.

Multicast Addresses:

– All L1 ISs: `01-80-C2-00-00-14`
– All L2 ISs: `01-80-C2-00-00-15`

3.2.7.3.1. DIS election

– Designated Intermediate System
– Similar to the DR in OSPF Protocol
– Responsibility
  – Creating and updating pseudonode LSPs
  – Flooding LSPs over the LAN
– L1 and L2 DIS may not be the same router
– Selection of the DIS
  – The highest priority. Configurable priority from 0 to 127.
  – Highest SNPA (Subnetwork Point of Attachment)
    – Highest MAC address of router's interface

3.2.7.3.2. Pseudonodes

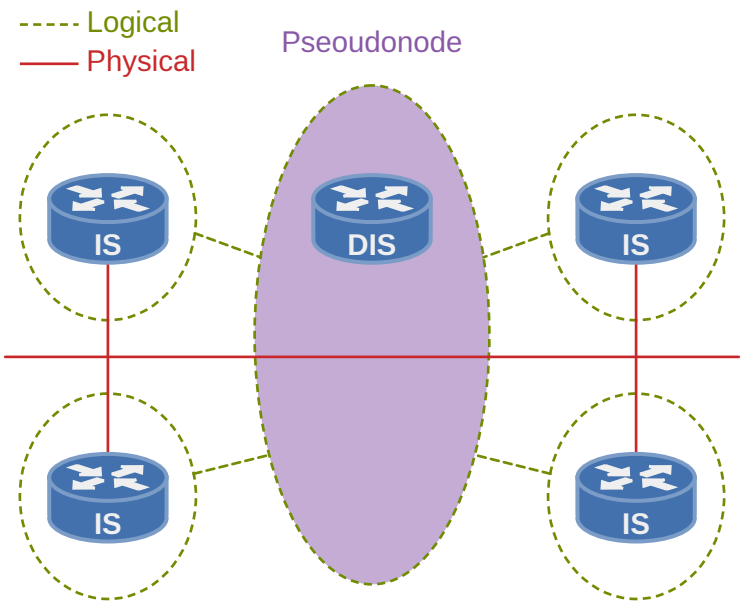
To minimize the complexity of managing multiple adjacencies on multiaccess media, such as LANs, while enforcing efficient LSP flooding to minimize bandwidth consumption, IS-IS models multiaccess links as nodes, referred to as ***pseudonodes***.

As the name implies, this is a virtual node, whose role is played by an elected DIS for the LAN. Separate DISs are elected for Level 1 and Level 2 routing.

– Election of the DIS is based on the **highest interface priority**, with the **highest SNPA address** (MAC address) breaking ties.
– The default interface priority on Cisco routers is 64.

The responsibilities of LAN Level 1 and Level 2 DISs include the following:

– Generating pseudonode link-state packets to **report links to all systems** on the LAN
– **Carrying out flooding** over the LAN for the corresponding routing level



Despite the critical role of the DIS in LSP flooding, **no backup DIS** is elected for either Level 1 or Level 2. If the current DIS fails, **another router is immediately elected** to play the role.

An elected router is **not guaranteed to remain the DIS if a new router with a higher priority shows up** on the LAN. Any eligible router at the time of connecting to the LAN immediately takes over the DIS role, assuming the pseudonode functionality.

3.2.7.4. Passive interfaces

Passive interfaces provide a method of **advertising network prefixes** into IS-IS, while **preventing an adjacency from forming** on that interface.

A passive interface **does not send out IS-IS traffic** and **will not process any received IS-IS packets**.

3.2.8. PDU

Each PDU has two packet types, one for each level.

Each type of IS-IS packet is made up of:

– A **header** with the common fields shared by all IS-IS packets
– A number of optional variable-length fields containing **specific routing-related information (Type, Length, and Value (TLV))** which has become a synonym for variable-length fields.

Enhancements to the original IS-IS protocol are normally achieved through the introduction of new TLV fields. A key strength of the IS-IS protocol design lies in the **ease of extension through the introduction of new TLVs** rather than new packet types.

3.2.8.1. Hello Packet (Hello)

Used to **establish adjacencies** between IS-IS neighbors. Once the neighbors are discovered, hello packets act as **keepalive messages** to maintain the adjacency. Additionally to the L1 and L2 types, Point-to-point hello packets exist.

1) discover
2) build
3) maintain

IS-IS neighbor adjacencies: Sent periodically

3.2.8.2. Link-State Packet (LSP)

Used to **distribute and exchange routing information** between IS-IS nodes. An IS-IS router floods an LSP throughout an area to identify its adjacencies and their states, path costs as well as reachable address prefixes.

– Carries associated networks (IPv4/IPv6) as metadata TLVs
– Sequenced to prevent duplication
– Unicast on Point-to-point Links, Multicast on broadcast media

3.2.8.3. Sequence number Packets (SNPs)

Used to control distribution of linkstate packets, providing mechanisms for **synchronization of the distributed Link-State** databases on the routers in an IS-IS routing area.

3.2.8.3.1. Complete Sequence Number PDU (CSNP)

Describe **summary of LSPs in the LSDB**. Similar to DBD in OSPF.

3.2.8.3.2. Partial Sequence Number PDU PSNP

Used to **request and acknowledge missing pieces** of link-state information. Like OSPF LSR and LSAck in one packet.

3.2.9. Hello process

Routers periodically send hello packets to adjacent peers, every hello interval. On Cisco routers, the default value of the hello interval is:

– 10s for ordinary routers
– 3.3s for the DIS on a multi-access link

IS-IS uses the concept of hello multiplier to determine how many hello packets can be missed from an adjacent neighbor before declaring it “dead”.

– The maximum time-lapse allowed between receipt of two consecutive hello packets received is referred to as the ***holdtime***.
– The holdtime is defined as the product of the hello interval and the hello multiplier.

3.2.10. Routing

3.2.10.1. Level 1 routing

Level 1 routing is routing within an area

– L1 routers follow a default route to the closest L1/L2 router.
  – L1-L2 routers do not advertise L2 routes into the L1 area

  - Default Route to the closest L1/L2 router
  - An IS-IS L1 area is equivalent to an OSPF totally stubby area.

3.2.10.2. Level 2 routing

Level 2 routing is routing between different areas

- L1-L2 routers inject L1 prefixes into the L2 topology.
  - Routes from the L1 level are advertised to the L2 topology populating the L1 topology metric into the L2 link-state packet (LSP) metric.
- L2 routers flag connectivity to the backbone to Level 1 routers by setting the attached bit in their Level 1 LSP, which is flooded throughout the area.

3.2.10.3. Interface metrics

Narrow metric:

- 6-bit field (value between 1 and 63)
- IS-IS assigns a default metric of 10 to all interfaces regardless of the interface bandwidth
  - A 1-Mbps link uses the same path metric as a 10-Gbps link by default

Wide metric:

- 24-bit field
- It should be used for large networks
  - The narrow-style metric can accommodate only 64 metric values, which is typically insufficient in modern networks

3.2.10.4. Path selection route types

IS-IS best-path selection uses the following processing order, identifying the route with the lowest path metric for each stage

Intra-area routes (L1): Routes that are learned from another router within the same level and area address

Inter-area routes (L2): Routes that are learned from another L2 router that came from an L1 router or from an L2 router from a different area address

External routes are no longer treated as a separate category for path selection; they are integrated based on their redistribution level and metric.

3.2.10.5. Route leaking

Even though the selected default router might be the closest in the area, it might not be the best exit out of the area when the overall cost to the destination is considered. There is a **possibility of suboptimal path selection**, which can be **corrected by *route-leaking***.

- Route-leaking is a technique that **redistributes the L2 level routes into the L1 level**
- Route leaking uses a **restrictive route map** or route policy to control which routes are leaked
- Set the ***Up/Down bit*** to mark routes leaked from Level 2 to Level 1, **preventing routing loops** by ensuring they aren't readvertised back into the backbone

3.2.10.6. IS-IS summarization

Because all routers within a level must maintain an identical copy of the LSPDB, **summarization occurs when routers enter an IS-IS level**, such as

- L1 routes entering the L2 backbone
- L2 routes leaking into the L1 backbone
- Redistribution of routes into an area

The default metric for the summary range is the smallest metric associated with any matching network prefix

You configure only the network that needs to have a different route and on the L1/L2 router that is the more optimal BR (not the default BR).

4. BORDER GATEWAY PROTOCOL (BGP)

RFC 1654 defines the Border Gateway Protocol as an EGP standardized **path-vector** routing protocol that provides scalability, flexibility, and network stability.

BGP does not advertise incremental updates or refresh network advertisements like OSPF or ISIS would – it prefers stability within the network. A flapping link could potentially result in the re-computation for thousands of routes.

4.1. COMPARISON TO IGP

IGP	*BGP*
Neighbors typically discovered using multicast packets on the connected subnets	Neighbor IP address is explicitly configured and may not be on common subnet
Does not use TCP	Uses a TCP connection between neighbors (port 179)
Advertises prefix/length	Advertises prefix/length, called Network Layer Reachability Information (NLRI)
Advertises metric information	Advertises a variety of path attributes (PA) that BGP uses instead of a metric to choose the best path
Emphasis on fast convergence to the truly most efficient route	Emphasis on scalability; might not always choose the most efficient route
Link-state logic	Path-vector logic

4.2. INTERNET ROUTE AGGREGATION

Problem:
- Increasing Internet Routing Table
  - If there are many small routes in routing table

Idea/Solution/Mitigation:
- Route Summarization
  - Allocate consecutive addresses in a single route by geography and ISP

4.3. AUTONOMOUS SYSTEMS (AS)

An AS is a network under **same administrative domain** using one or more IGPs. An IGP is not required within an AS, and iBGP could be used, however, it would not scale well. Routing and security policies are under the control of one service provider or of one company.

4.3.1. Autonomous System Numbers (ASN)

Organizations requiring connectivity to the internet must obtain an ASN. They were originally 2 bytes with 65'535 ASNs. This limited range was exhausted rather quickly, prompting the expansion of the ASN range

to 4 bytes in RFC 4893 , resulting in 4'294'967'295 ASNs, being backward compatibile with ASN 23456 (ASN_TRANS).

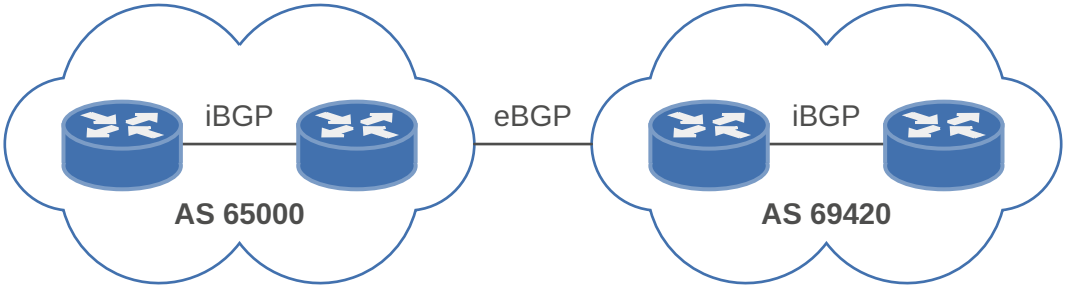
Two blocks of **private ASNs** are available to any organization. These can be used as long as the companies do not exchange them on the internet (similar to the private IPv4 addresses specified in RFC 1918). They are defined in RFC 6996 (Autonomous System Reservation for Private Use):

– 16-bit range: 64'512 – 65'534
– 32-bit range: 4'200'000'000 – 4'294'967'294

Note that RFC 7300 (Reservation of Last Autonomous System Numbers) define 65'535 (last 16-bit ASN) and 4'294'967'295 (last 32-bit ASN) as **reserved**, but not explicitly for private use.

4.4. SESSIONS

A BGP session refers to the established adjacency between two BGP routers. BGP sessions are always **point-to-point** and are categorized into two types, iBGP and eBGP.



4.4.1. Internal BGP (iBGP)

BGP that are peering **within the same AS**. iBGP sessions are considered more secure, and some of BGP's **security measures are lowered** in comparison to eBGP sessions. iBGP prefixes are assigned an **AD of 200** upon being installed into the router's RIB.

– AS-Path not modified
– Next-hop not modified

The need for BGP within an AS typically occurs when **transit connectivity** is provided between autonomous systems.

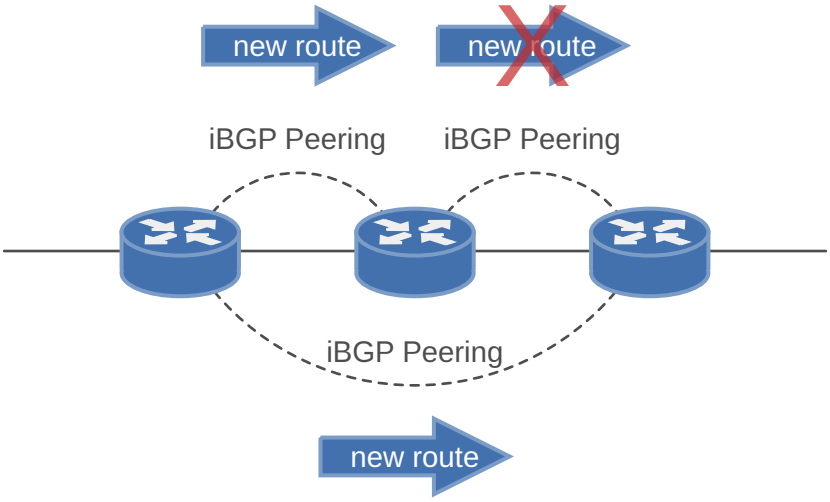
Advertising the full BGP table into an IGP is not a viable solution for the following reasons:

– **Scalability**: In January 2024, the internet had 943 000+ IPv4 networks, and it's still growing. IGPs do not scale to such a high number of routes.
– **Path Attributes**: IGP protocols do not know about BGP path attributes. Only BGP is capable of maintaining the path attribute as the prefix is advertised from one edge of the AS to the other edge.

4.4.1.1. Full Mesh Requirement

iBGP peers do not prepend their ASN to the AS_PATH because the NLRIs would fail the validity check (because it's the same ASN) and would not install the prefix into the IP routing table.

No other method exists to detect loops with iBGP sessions, and RFC 4271 prohibits the advertising of NLRI received from an iBGP peer to another iBGP peer (split horizon). It also states that all BGP routers within a single AS **must be fully meshed** to provide a complete loop-free routing table and prevent traffic ***blackholing***.



4.4.1.2. Peering via Loopback Addresses

BGP sessions are sourced by the outbound interface toward the BGP peers IP address by default.

It is preferable to configure the BGP neighbours to establish a session between their loopback addresses. The **loopback interface is virtual and always stays up**. In the event of link failure, the **session remains intact** if the IGP finds another path to the loopback address.

4.4.1.3. Scalability

The inability for BGP to advertise a prefix learned from one iBGP peer to another can lead to scalability issues within an AS: Let n be the number of iBGP speakers. There are $n(n-1)/2$ sessions required. In asymptotic notation, we would categorize this as $O(n^2)$.

4.4.1.3.1. Route Reflectors

RFC 1966 introduces the concept of route reflection, which allows an iBGP speaker to **advertise routes** learned from one iBGP peer to other iBGP peers. The router performing this function is called a ***route reflector*** (RR).

An RR forms iBGP sessions with two types of peers:

Client peers: iBGP peers configured as clients of the RR. The RR reflects routes between these peers.

Non-client peers: regular iBGP peers of the RR that are not configured as clients. These peers behave like normal iBGP speakers and are expected to maintain a full mesh among themselves.

The following rules govern route reflection:

1) If a RR receives a NLRI from a non-client peer, the RR advertises the NLRI to all clients. It does not advertise the NLRI to other non-client peers.
2) If a RR receives a NLRI from a client, the RR advertises the NLRI to all non-client peers and to all other clients (except the originating client).
3) If a RR receives a NLRI from an eBGP peer, the RR advertises the NLRI to all clients and all non-client peers, subject to normal BGP rules.

Only route reflectors need to be aware of this modified advertisement behaviour. Route-reflector **clients require no special configuration** beyond establishing the iBGP session with the RR. By introducing route reflectors, the **requirement for a full iBGP mesh can be relaxed**: each client only needs to peer with the RR to receive the routes from the rest of the AS.

4.4.2. External BGP (eBGP)

Sessions established with eBGP routers that are in **different ASes**. eBGP prefixes are assigned an **AD of 20** upon being installed into the router's RIB.

– Each eBGP device modifies the ***AS-Path*** attribute with its own AS.
– Each eBGP device modifies the ***next-hop*** attribute

4.4.2.1. Comparison to iBGP

– **TTL on BGP packets is set to one** by default. BGP packets drop in transit if a multihop BGP session is attempted.
– The advertising router modifies the BGP next-hop to the IP address sourcing the BGP connection.
– The advertising router prepends its ASN to the existing AS_PATH. The receiving router verifies that the AS_PATH does not contain an ASN that matches the local routers. BGP discards the NLRI if it fails the AS_PATH loop prevention check.

4.4.3. Combining iBGP and eBGP

Combining eBGP sessions with iBGP sessions can cause problems. The most common issue involves the failure of the next-hop accessibility. iBGP peers do not modify the next-hop address if the NLRI has a next-hop address other than `0.0.0.0`.

The **next-hop address must be resolvable in the global RIB** for it to be valid and advertised to other BGP peers.

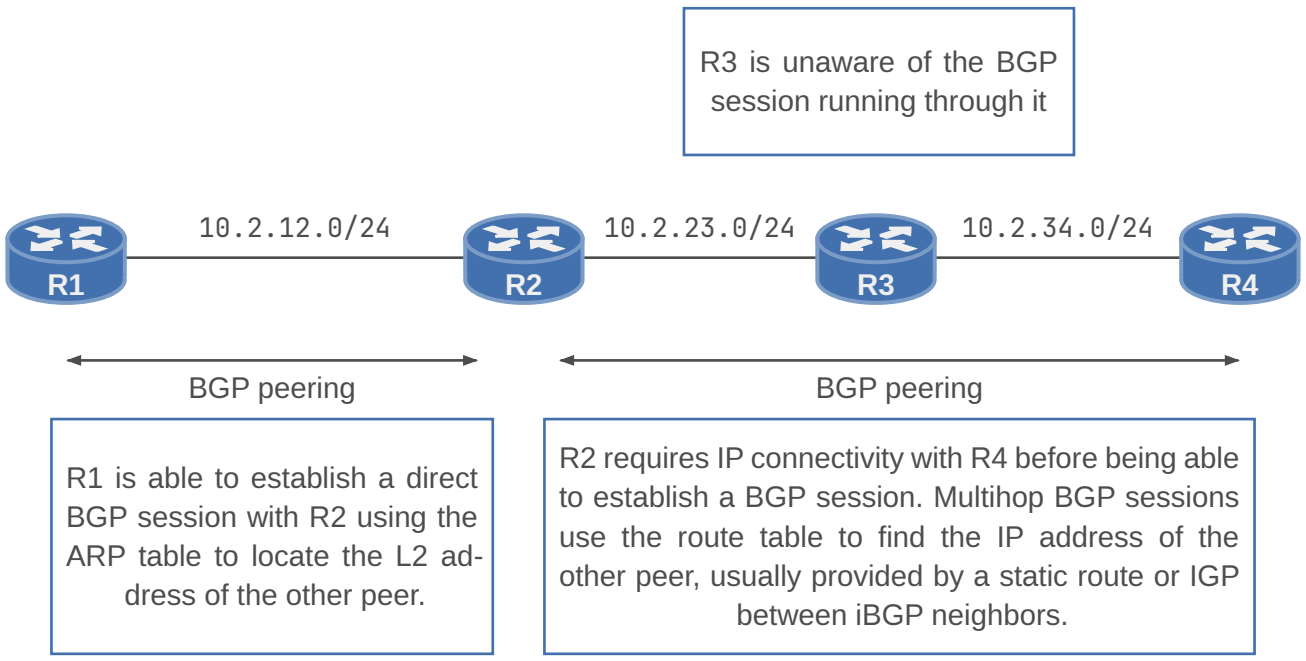
4.4.3.1. Next hop behavior

Problem: When paired with an eBGP neighbor the next-hop is passed to the iBGP neighbor but the **iBGP neighbor is not able to reach the next-hop**.

To avoid this issue, the next-hop IP address can be **modified**. NHOP is a BGP attribute that can also be manipulated. Configuring the ***next-hop-self*** feature modifies the next-hop address in all external NLRIs using the IP address of the BGP neighbour.

4.4.4. Multihop Sessions

TCP allows for handling of fragmentation, sequencing, and reliability (acknowledgement and retransmission) of communication packets. While BGP can form neighbour adjacencies that are directly connected, it can also form adjacencies that are multiple hops away. **Multihop sessions require that the routers use an underlying route** installed in the RIB (static or from any routing protocol) to **establish the TCP session with the remote endpoint**.



4.5. MESSAGES

BGP utilizes 4 different types of messages to establish, maintain and tear down BGP peers: OPEN, KEEPALIVE, UPDATE and NOTIFICATION.

4.5.1. OPEN

The OPEN message is used to **establish a BGP adjacency**. It contains the BGP version number, ASN of the originating router, hold time, the BGP identifier, and other optional parameters that describe the session capabilities.

Hold Time: Defines **how many seconds to wait before declaring a BGP neighbour unreachable**. Upon receiving an UPDATE or KEEPALIVE, the hold timer **resets to the initial value**. If the hold timer for a neighbour reaches zero, the BGP session to it is torn down: Routes from the timed-out neighbour are removed, and an appropriate route withdraw message is sent to other neighbours. When establishing a BGP session, **both routers propose a hold time** in their OPEN message, where the **smaller value is chosen**. For Cisco routers, the default hold time is 180s.

BGP Identifier: The BGP Router ID (RID, or simply BGP Identifier) is a unique 32-bit number that **identifies the BGP process on that router**. It can be used as a **loop prevention mechanism** for routers advertised within an autonomous system. It can be set manually or is assigned dynamically.

Router IDs **typically represent an IPv4 address** that resides on the router, such as a loopback address. Any valid IPv4 address can be used, including ones not configured on the router. Configuring a static BGP RID is considered a best practice, as it provides consistency and stability.

4.5.2. KEEPALIVE

A BGP process does not rely on the TCP connection state to ensure that its neighbours are still alive. KEEPALIVE messages are exchanged every third of the hold timer agreed upon between the two BGP routers. Cisco devices have a default hold time of 180 s, so the default KEEPALIVE interval is 60 s.

4.5.3. NOTIFICATION

A NOTIFICATION message is sent **when an error is detected** with the BGP session, such as a hold timer expiring, neighbour capabilities change, or when a BGP session reset is requested. It **causes the BGP connection to close**.

4.5.4. UPDATE

An UPDATE message **advertises feasible routes**, **withdraws previously advertised routes**, or does both. It contains Network Layer Reachability Information (NLRI), which includes the prefix, along with associated BGP path attributes when advertising those prefixes. Withdrawn NLRIs include only the prefix. An UPDATE message **can also act as a KEEPALIVE** message to reduce obsolete traffic.

A BGP update message is composed of:

- A list of routes to explicitly **withdraw**.
- The **attributes** associated with the new prefixes being advertised in this update.
  - The attributes include AS path, MED, community, and many others.
- The new **prefixes**
  - The routes include both a network and a mask.

4.6. PATH ATTRIBUTES

- Per RFC 4271 , **well-known** attributes must be recognized by all BGP implementations.
- **Optional** attributes do not have to be recognized by all BGP implementations.

BGP attributes can be classified in 4 categories:

Well-known mandatory: attributes must be included with every BGP update.

Well-known discretionary: attributes are not required in every BGP update.

Optional transitive: attributes stay with the route advertisement from AS to AS

Optional non-transitive: attributes should be removed if not understood and should not be forwarded to other ASs

Well-known Mandatory	*Well-known Discretionary*	*Optional Transitive*	*Optional Non-transitive*
– AS Path – Origin – Next Hop	– Atomic Aggregate – Local Preference	– Community – Aggregator	– MED – Weight – Cluster ID – Originator ID – Cluster List

NLRI (Network Layer Reachability Information) is the format used to **represent the prefixes a BGP speaker advertises to its peers** (informing them of networks reachable via specific paths). It consists of a network prefix, prefix length, and any BGP prefix attributes for that specific route.

4.7. PATH CALCULATION

A route advertisement consists of the NLRI and its path attributes. A BGP router may learn multiple paths to the same destination network. The attributes associated with each path influence the desirability of the route when the router selects the best path. A BGP router **advertises only its selected best path** to its peers.

Within the BGP table, **all learned routes and their associated path attributes are maintained**, and the best path is calculated. The selected **best path is then installed in the router's routing table** (RIB). If the best path becomes unavailable, the router can evaluate the remaining known paths to quickly select a new best path.

BGP recalculates the best path for a prefix when one of the following events occurs:

– A change in **reachability** of the BGP next hop.
– **Failure of an interface** connected to an eBGP peer.
– A change in **redistributed routes**.
– When receiving **new paths for the same prefix**.

Some router configurations modify the BGP attributes to influence inbound traffic or outbound traffic. BGP path attributes can be modified upon receipt or advertisement to influence routing in the local AS or neighbouring AS. A basic rule for traffic engineering with BGP is that **modifications in outbound routing policies influence inbound traffic, and modifications to inbound routing policies influence outbound traffic**.

4.7.1. Route Selection

BGP installs the **first received path as the best path automatically**. When additional paths are received, the newer paths are compared against the current best path. If there is a tie, then processing continues onto the next step, until the best path winner is identified.

BGP uses 11 steps to determine the best path:

1) Prefer highest weight (local to router) **CISCO specific - do not use!**
2) Prefer highest local preference (global within AS)
3) Prefer routes that the router originated
4) Prefer shorter AS paths (only length is compared)
5) Prefer lowest origin code (IGP < EGP < Incomplete)
6) Prefer lowest MED (also called metric)
7) Prefer external (EBGP) paths over internal (IBGP)
8) For IBGP paths, prefer path through closest IGP neighbor
9) For EBGP paths, prefer oldest (most stable) path
10) Prefer paths from router with the lower BGP router-ID
11) Prefer the path that comes from the lowest neighbor address

4.8. LOOP PREVENTION

As a path-vector routing protocol, BGP **does not contain a complete topology** of a network (as opposed to link-state routing protocols). BGP behaves similar to distance vector protocols to ensure a path is loop free.

The BGP attribute ***AS_PATH*** is a well-known mandatory attribute that includes a **complete listing of all the ASNs** that the prefix advertisement has traversed from its source AS. **If AS-Path includes the router's ASN, then it is ignored**.

4.9. NETWORK STATEMENTS

BGP Network statements **identify a specific network prefix to be installed** into the BGP table. After configuring a BGP network statement, the BGP process searches the global RIB for an exact network prefix match. The network prefix can be a connected network, secondary connected network, or any route from a routing protocol. After verifying that the network statement matches a prefix in the global RIB, the prefix is installed into the BGP table.

The following BGP Origin path attribute is set depending on the RIB prefix type:

Connected Network: The next-hop BGP attribute is set to `0.0.0.0`, the origin attribute is set to i (IGP), and for Cisco devices, the BGP weight is set to 32'768

Static Route or Routing Protocol: The next-hop BGP attribute is set to the next-hop IP address in the RIB, the origin attribute is set to i (IGP), the BGP weight is set to 32'768 and the MED is set to the IGP metric.

4.10. ROUTE FILTERING AND MANIPULATION

Route filtering is a method to **select routes to receive from (import) or advertise to (export) neighbouring routers**. This feature can be used to manipulate traffic flows, reduce memory utilization, or to improve security.

It is common for ISPs to deploy route filters on BGP peerings to customers: They want to ensure that only the customer's routes are allowed over the peering link, preventing the customer from accidentally becoming a transit AS on the internet.

Filtering of routes within BGP is accomplished with ***filter-lists***, ***prefix-lists***, or ***route-maps*** on Cisco IOS.

In Cisco IOS, regular expressions can be used in show commands and AS path accesslists to match BGP prefixes based on the information contained in their AS path.

4.10.1. Path Announcement

ASs announce paths to destination addresses, data flows back to the opposite direction.

4.11. COMMUNITIES

BGP communities provide **additional capability for tagging routes** and for modifying BGP routing policy on upstream and downstream routers.

Communities can be appended, removed, or modified selectively on each attribute as the route travels from router to router. They are an **optional transitive** BGP attribute that can traverse from autonomous system to autonomous system. A BGP community is a 32-bit number that can be included with a route. It can be displayed as a full 32-bit number (0 − 4'294'967'295) or as two 16-bit numbers (0 − 65'535:0 − 65'535), commonly referred to as new-format.

4.12. CONNECTIVITY OPTIONS

Single/Dual: denotes how many **links** there are

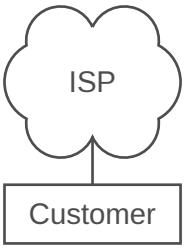
Multi-Homed/-Homed: denotes how many **ISPs** are connected

4.12.1. Single-Homed without BGP

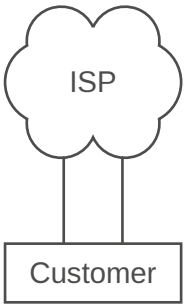
– The customer doesn't use BGP.
– Static default route on the customer side to reach outside networks
– Specific static route on the ISP side to reach the customer IP address prefix

4.12.2. Single-Homed with BGP

– The customer uses BGP
– Changes in the customer topology will then be sent to the provider.
– Provider may redistribute the changes to the internet.

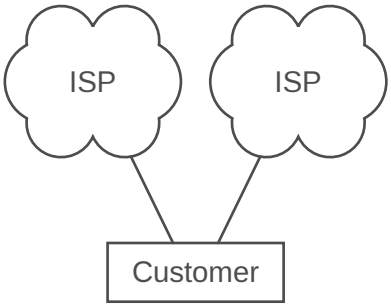


4.12.3. Dual-Homed



- One or two customer router(s)
- Customer optimally runs BGP between routers
- ISP announces default route
- Usually used in a primary/backup design
  - Local Preference to steer traffic
  - First Hop Redundancy Protocols used to ensure correct traffic routing
    - e.g. Hot Standby Router Protocol (HSRP)
- Load-Sharing possible

4.12.4. Multi-Homed



Used in an active/active design

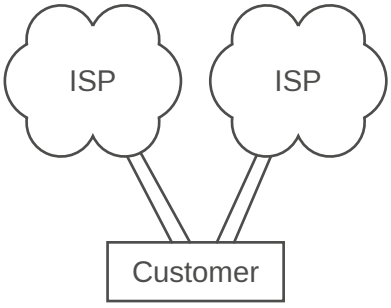
- By receiving the full routing table, the path through the internet can be optimized

A customer AS never wants to be transit

- Only advertise routes originating in own AS to ISPs

4.12.5. Dual Multi-Homed

Provides the most redundancy (ISPs as well as links).



4.12.6. Traffic engineering

4.12.6.1. Outbound

How can you influence how traffic is leaving the network?

Local Preference: Highest Local Preference wins ⇒ Decrease Local Preference on backup path

4.12.6.2. Inbound

How can you influence how traffic is entering the network?

MED (Multi-Exit Discriminator): Lowest MED wins ⇒ Increase MED on backup path

AS-Path prepending: Shortest AS-Path wins ⇒ Add AS several times on backup path

4.12.6.3. Caveats

An AS has direct control over egress traffic but lacks absolute control over ingress paths.

Example: an ISP's Local Preference settings will take over and effectively ignore any MED or AS_PATH attributes.

4.12.6.4. Aggregate

Used in Multi-Homed systems.

- Customer prefers Primary provider

– Using Alternate only as backup
– Primary provider advertises aggregated networks
– Alternate provider advertises individual network
– Remote autonomous systems prefer longest-match prefix
– Result: Traffic toward the customer flows through Alternate provider
– Solution: Don’t use Provider-aggregate public IP address

5. THE INTERNET

5.1. STRUCTURE

The internet is a network of networks.

– Internet is an interconnection of 10’000s autonomous service providers and customers.
– There is no central co-ordination for the management of interconnections, services and tariffs.

Who controls the internet?

– The control over paths is completely distributed. It is all based on trust.

Assumption:

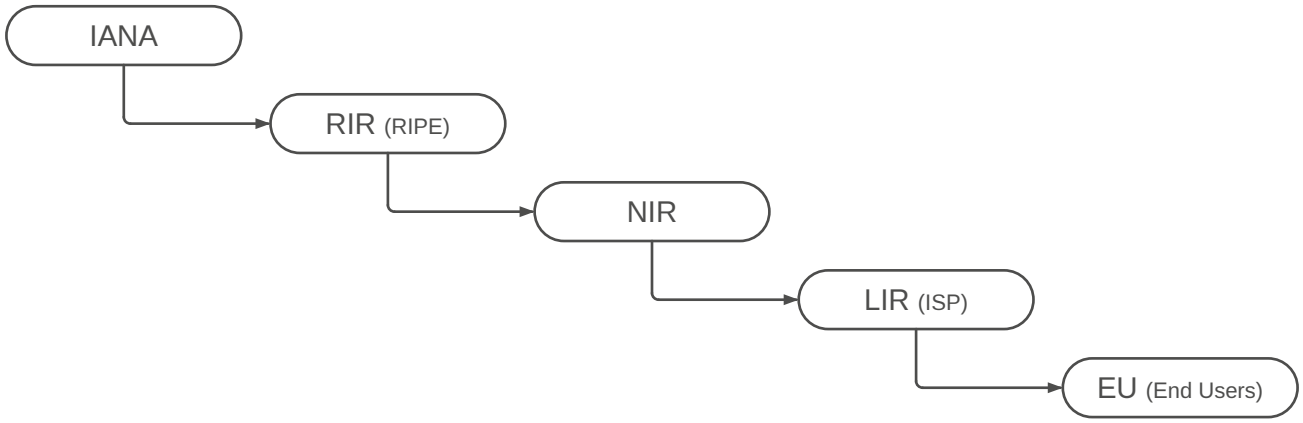
– The Internet was based on a well-ordered provider client hierarchy.

Reality:

– Unordered subset of interconnects
– Driven by business requirements underpinned by performance
– Non-disclosure and bilateral agreements
– Peering is now considered a corporate asset & legal concern

5.2. PUBLIC IP ADDRESS ASSIGNMENT

Term	*Definition*
ICANN	Internet Corporation for Assigned Names and Numbers
IANA	Internet Assigned Numbers Authority
IR	Internet Registry
RIR	Regional Internet Registry
NIR	National Internet Registry
LIR	Local Internet Registry



1) ICANN and IANA group public addresses by major geographic region.
2) IANA allocates those address ranges to Regional Internet Registries (RIR).
3) Each RIR further subdivides the address space by allocating public address ranges to National Internet Registries (NIR) or Local Internet Registries (LIR). (ISPs are typically LIRs.)
4) Each type of Internet Registry (IR) can assign a further subdivided range of addresses to the end-user organization to use.

5.3. PEERING VS. TRANSIT

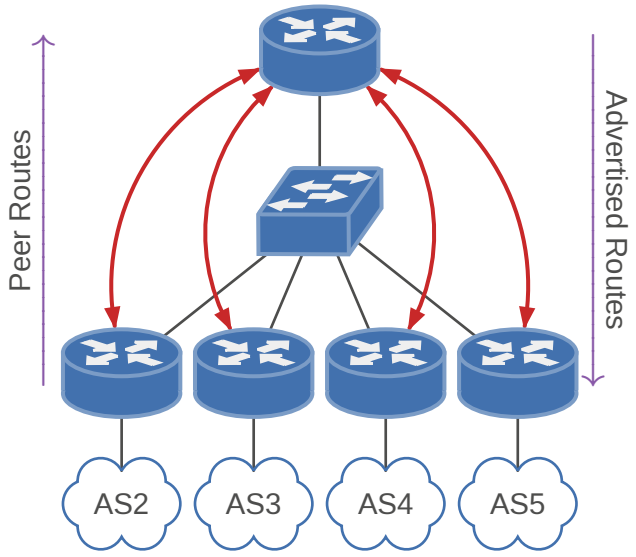
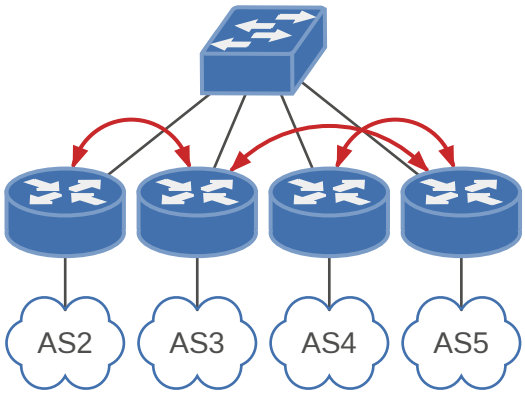
The nature of the linking between these ISPs is governed by a series of agreements known as peering arrangements.

Term	***Definition***
Transit	Business relationship where **one ISP provides reachability to all destinations** in it's routing table to its customers. – Transit fees, usually paid by a smaller ISP to a larger
Peering	Business relationship where ISPs **provide to each other reachability to each pre-defined portions** of their routing table – Peers are equals and pass traffic from one to another without worrying about payments. – They treat smaller ISPs as just another customer.

5.4. ROUTING POLICIES

– Each ISP has a unified routing policy framework
– The decision on which routes to advertise and which routes to accept is determined by ***routing policy***.
  – Routes, or prefixes, not only need to be advertised to another AS, but need to be accepted.
– ISPs do not provide free transit services and generally are either peers or customers of other ISPs.
  – Unless "arrangements" are made, transit ISPs will routinely block transit

5.5. INTERNET EXCHANGE POINT (IXP)

Public IXP	*Private IXP*
Member will generally peer with a route server. – Generally, over a common switched infrastructure The route server announces the members routes to all peers – Generally, the Route Server (RS) will inject its AS into the AS_PATH – The NEXT_HOP is preserved and keeps the RS out of the data path Lack of policy control – Single legal contract to manage	Members will peer on a one-to-one basis – Generally, over a common switched infrastructure - Public peering – Can be over private interconnects – Private peering Peers implement policy towards each other – One BGP session per neighbour – Multiple legal contracts to manage
	

5.6. ROUTING SECURITY

– Internet Routing Registry (IRR)
  – unreliable

5.6.1. Resource Public Key Infrastructure (RPKI)

RFC 8210 (The Resource Public Key Infrastructure (RPKI) to Router Protocol, Version 1)

RFC 6480 (An Infrastructure to Support Secure Internet Routing)

RPKI is a robust security framework for verifying the association between resource holder and Internet resource. Helps to secure Internet routing by validating routes and thus preventing the following:

Route hijacking: A prefix originated by an AS without authorization with malicious intent

Route leakage: A prefix that is mistakenly originated by an AS which does not own it

“Is this AS number (ASN) authorized to announce this IP range?”

– RPKI-capable routers can fetch the validated Resource Origin Authorizations (ROA) data set from a validated cache

VALID	*INVALID*	*NOT FOUND / UNKOWN*
Indicates that the prefix and ASN pair have been found in the database	Indicates that the prefix is found, but: – ASN received did not match – The prefix length is longer than the maximum length	Indicates that the prefix does not match any in the database

5.6.1.1. Trust Anchors (TA)

A TA is a certificate authority (CA) in RPKI terms. The five Regional Internet Registries (RIR) are the TAs and have following responsibilities:

1) Provide the infrastructure so that resource holders can sign their prefixes and ASNs.
2) Provide a public list so that others can verify these prefixes and ASNs.

Resource certificates are based on the X.509 V3 certificate format defined in RFC 5280 and extended by RFC 3779 , which binds a list of resources (IP, ASN) to the subject of the certificate.

X.509 certificates are typically used for authenticating either an individual or, for example, a website. In RPKI, certificates **do not include identity information**, as their only **purpose is to transfer the right to use Internet number resources**.

5.6.1.2. Route Origin Authorization (ROA)

An object cryptographically signed with a public key, containing three items:

1) The authorized AS number
2) The prefix that this AS is allowed to originate
3) The maximum prefix length (maxLength)

– Specifies which ASNs are authorized to originate certain IP prefixes, enabling routers to verify the legitimacy of BGP announcements
– A signed digital object that contains a list of addresses, prefixes and one AS number
– Created by a prefix holder to authorize an AS number to originate one or more specific route advertisements
– An ROA is valid if, the associated certificate can be validated up to the TA of the of the corresponding RIR e.g. (RIPE, APNIC, etc.)

Example 3: **Sample ROA**

Attribute	*Value*
Prefix originated	203.176.189.0
Maximum prefix length	/24
Origin ASN	AS17821

Maximum prefix length specifies the maximum length of the IP address prefix that the AS is authorised to advertise

5.6.1.3. RPKI Validators

– RPKI Validators also called Relying Party Software
– Run independently by an organization
– Synchronizes and validates resource records
– Preloaded with trust anchor locators (TALs) for all RIRs.

- Retrieves data with a specific protocol RFC 8182
  - Follows chain of trust top to bottom
- Validates cryptographic signatures on all objects.
  - Outputs a list of Validated ROA Payloads (VRPs).
- Periodically retrieves updates

Currently, RPKI only provides origin validation. While BGPsec path validation is a desirable characteristic and standardised in RFC 8205 , real-world deployment may prove limited for the foreseeable future. However, RPKI origin validation functionality addresses a large portion of the problem surface.

5.6.2. BGP Monitoring

BGP monitoring is a process that helps network operators detect and troubleshoot issues in their routing infrastructure. By understanding and analyzing BGP data, operators can optimize network performance, minimize downtime, and maintain the overall health of their networks.

6. UNICAST

Unicast	*Broadcast*	*Multicast*
Source sends n unicast datagrams, one for each receiver	Source sends a datagram for each network (summarizing)	Routers actively participate in multicast, making copies as needed
Only each individual target receive the packets	“Not receivers” still receive packets	“Not receivers” do not receive packets

7. BROADCAST

7.1. LAYER 2 - ETHERNET BROADCAST FRAMES

At Layer 2, the all-hosts broadcast MAC address is `ffff.ffff.ffff`. Switches must replicate such frames when forwarding them so that all devices within the VLAN receive the traffic; this process is known as flooding. Example: ARP request.

Switches forward broadcast traffic out of every interface in the same VLAN that is in a forwarding state, except for the port on which the frame was received.

7.2. LAYER 3 - IP BROADCAST PACKETS

A ***local broadcast*** (`255.255.255.255`) is restricted to the local network segment and is **never forwarded** by routers. The router accepts it, but only for the local interface. This is the default behavior on routers from all major vendors. If routers were to forward local broadcast packets, they would have to send them out of every interface with IP enabled, because the destination address `255.255.255.255` is considered reachable through all IP-enabled interfaces on the router.

In contrast, a ***directed broadcast*** (e.g., `192.168.1.255`) is the broadcast address for a specific subnet and can be routed to that subnet by a router. This works for local networks (e.g., `10.1.1.255`) and for remote networks (e.g., `10.3.3.255`).

This distinction highlights the fundamental difference between bridging traffic at Layer 2 and routing traffic at Layer 3.

8. MULTICAST

Broadcast traffic is confined to a single broadcast domain and is received by all devices within that segment, regardless of whether they require the data. In contrast, ***multicast*** is designed for **efficient** group communication across potentially **multiple network segments**, where traffic is delivered only to **explicitly interested receivers**. Furthermore, multicast traffic **can be routed through the network**, while broadcast traffic is typically not forwarded by routers and therefore remains limited to a single Layer 2 domain.

8.1. LAYER 2 - MAC MULTICAST

While Layer 3 multicast enables packets to be routed across multiple network segments, the final distribution to end hosts is performed at Layer 2 using multicast MAC addresses. This distinction ensures that multicast traffic reaches the correct networks while avoiding unnecessary flooding within each segment. Without Layer 2 multicast mechanisms, switches would treat multicast traffic similarly to broadcast traffic, leading to inefficient use of bandwidth and increased processing overhead on end devices.

8.2. LAYER 3 - IP MULTICAST

UDP-based

One-to-many	*Many-to-many*
– Music-on-hold services – Sensor updates	– Stock exchanges – Group chat applications

8.2.1. Benefits

- Enhanced Efficiency
  - Controls network traffic and reduces server and CPU loads
- Optimized Performance
  - Eliminates traffic redundancy
- Distributed Applications
  - Makes multipoint applications possible

8.2.2. Inner workings

Best effort delivery	*No congestion avoidance*	*Duplicated packets*
– Drops are to be expected – multicast applications should not expect reliable delivery of data and should be designed accordingly	– Lack of TCP windowing and “slow-start” mechanisms can result in network congestion – multicast applications should attempt to detect and avoid congestion conditions	– Out-of-order delivery: Some protocol mechanisms may also result in out-of-order delivery of packets – multicast applications should be designed to expect occasional duplicate packet

8.2.3. Source

- Sends packets to multicast group IP address
- All group members will receive multicast traffic
- Source doesn’t have to be a member of group

8.2.4. Receiver

- Any multicast capable device

– Express interest in particular multicast group
– Receive traffic destinated to Multicast group IP

8.2.5. Addressing

Type	*Address range*	*Details*
Link-local multicast	`224.0.0.0/24` (to `224.0.0.255`)	– Used for control protocols (e.g., routing protocols, IGMP) – Packets are not forwarded by routers and remain within the local network segment – Transmitted with TTL = 1
Internetwork	`224.0.1.0/24` (to `224.0.1.255`)	– Used for protocol control that must be forwarded through the Internet (eg. NTP)
SSM (Source-Specific Multicast)	`232.0.0.0/8` (to `232.255.255.255`)	– Used when both source and group are known RFC 4607
Globally scoped multicast	`235.0.0.0` - `238.255.255.255`)	– Can be routed across networks – Used for general multicast applications RFC 5771
Administratively scoped addresses	`239.0.0.0/8` (to `239.255.255.255`)	– Intended for private multicast domains – Similar to private IPv4 addresses RFC 2365

8.2.5.1. Link-local multicast

The link-local multicast range has special significance, as traffic in this range is handled differently from other multicast traffic. Packets sent to these addresses are not routed and **always remain within the local network segment**.

At Layer 2, link-local multicast traffic is typically **flooded to all ports**, similar to broadcast traffic. However, unlike broadcast, multicast frames are only processed by hosts that support multicast and listen to the corresponding multicast MAC addresses.

Switch behavior for this range is defined such that mechanisms like IGMP snooping **do not restrict forwarding**, ensuring that essential control traffic is always delivered RFC 4541 .

8.2.5.2. IPv4 MAC address mapping

There is a specific reserved range (25bits) for Multicast MAC: **0100.5E**00.0000 to **0100.5E**7F.FFFF. A multicast MAC consists of the reserved range and parts of the IP address.

Multicast IPv4 addresses belong to the Class D range and always begin with the binary prefix 1110:

0b1110 0000 - 0b1110 1111 (0d224 - 0d239)

This corresponds to the address range:

`224.0.0.0` – `239.255.255.255`

The lower-order 23 bits of the IP address are mapped to the lower-order 23 of the IP multicast address. 5 bits are variable.

***Example** 4:*

Multicast IP address 1110 11111 1111111 00000001 00000100 (239.255.1.4) gets the multicast MAC address 01-00-5e-7f-01-04.

Limitations: Since only the last 23 of the 32 bits of the IP multicast address are used in the MAC address, and 4 bits are reserved as a fixed prefix, 5 bits (bits 5-9) are lost during the mapping process. Meaning, out of all of the possible addresses $2^5 = 32$ could get the same address assigned.

***Example** 5:*

Any multicast address 1110xxxx x0000001 00000001 000000001 gets the same multicast MAC address 01-00-5E-01-01-01.

8.2.5.3. IPv6 MAC address mapping

IPv6 follows the same schema. The IPv6 multicast address range is **ff00::/8** (the first 8 bits are fixed). The corresponding Ethernet multicast MAC address range is **3333**.0000.0000 – **3333**.FFFF.FFFF (the first 16 bits are fixed).

***Example** 6:*

IPv6 multicast address ff02::1 maps to the multicast MAC address 33:33:00:00:00:01

Mapping is performed by taking the lower 32 bits of the IPv6 multicast address and inserting them into the lower 32 bits of the Ethernet multicast MAC address, resulting in a many-to-1 (2^{88}-to-1) mapping between IPv6 multicast addresses and Ethernet multicast MAC addresses.

8.2.6. Internet Group Management Protocol (IGMP)

IGMP is the protocol used to manage group subscriptions for IPv4 multicast. On the router, IGMP tracks multicast group memberships on each segment. The operation can be summarized as follows:

– The router sends ***query messages*** to **discover hosts** that are members of a multicast group.
– Hosts send ***membership report messages*** to **indicate interest in joining or leaving** a multicast group, and also respond to router queries with report messages.

The selection of which IGMP version to run on your network depends on the operating systems and behavior of the multicast application. There are three IGMP versions: 1, 2, and 3. Each of these has unique characteristics.

8.2.6.1. IGMPv1

IGMPv1 offers a basic query-and-response mechanism to determine which multicast streams should be sent to a particular network segment.

IGMPv1 lacks an explicit leave-signaling mechanism. When a host silently leaves a group, the router continues forwarding the multicast stream until the group membership timer expires. To maintain the group state, the router sends periodic ***Membership Queries*** to the **All-Hosts** address (224.0.0.1) **every 60 seconds**. If no Membership Reports are received after several consecutive query cycles, the router removes the group from the interface and prunes the stream.

8.2.6.2. IGMPv2

One of the most significant improvements of IGMPv2 over IGMPv1 was the addition of the **leave process**. A host using IGMPv2 can send a ***leave-group message*** to the router indicating that it is no longer interested

in receiving a particular multicast stream. This operation eliminates a significant amount of unneeded multicast traffic by not having to wait for the group to time out.

IGMPv2 added the capability of ***group queries***. This feature allows the router to send a message to the hosts belonging to a **specific multicast group**. Every host on the subnet is no longer subjected to receiving a multicast message.

8.2.6.3. IGMPv3

The most significant addition in IGMPv3 is support for **source filtering**. In IGMPv1 and IGMPv2, a host could not specify the source of a multicast stream. Source filtering allows a host to signal membership using include or exclude source lists, indicating from which senders it wants or does not want to receive traffic. This provides finer control and can improve security at the application level.

IGMPv3 enables hosts to signal source-specific multicast membership, allowing ***PIM Source-Specific Multicast*** (SSM) to be used for IP multicast routing. In this context, IGMPv3 hosts send membership reports to the multicast address **`224.0.0.22` (all IGMPv3 routers)**, replacing the earlier `224.0.0.2`.

8.2.6.4. IGMP Snooping

IGMP Snooping is a **Layer 2 switch feature** that listens for IGMP conversations between hosts and routers to intelligently map multicast traffic **only to the ports that have requested it**, preventing it from flooding the entire VLAN like a broadcast.

When IGMP snooping is not enabled on a VLAN, any multicast will be treated as broadcast and will be sent to all end-hosts connected to the VLAN.
When IGMP snooping is enabled on a VLAN, a Layer 2 switch listens for IGMP messages. During the snooping process, the switch learns which end-hosts want to receive which groups and builds an ***IGMP snooping table***. The switch is now using that table to forward the multicast traffic to the port that requested it. When an end-host leaves a group, the switch will also read the IGMP Leave message and will update the IGMP snooping table accordingly.

8.3. PROTOCOL INDEPENDENT MULTICAST (PIM)

PIM is the most widely used multicast routing protocol. PIM does not build its own routing table; instead, it **relies entirely on the existing unicast routing table** to make forwarding decisions.

This means that PIM can operate with **any underlying unicast routing protocol**, such as static routing, OSPF, or IS-IS. In contrast, older multicast routing protocols like DVMRP or MOSPF maintain their own separate routing information.

PIM uses the concept of the ***Reverse Path Forwarding (RPF)*** check to ensure that multicast **traffic follows the correct path** through the network.

PIM supports three different operating modes:

– “PIM Dense Mode” (Page 49)
– “PIM Sparse Mode” (Page 51)
– “PIM Sparse-Dense Mode” (Page 52)

8.3.1. PIM Dense Mode

PIM dense mode is a multicast routing protocol that floods multicast traffic across all network links until it receives prune messages from routers not interested in receiving the traffic. This is called a “push” model where we flood multicast traffic everywhere and then prune it when it’s not needed.

8.3.1.1. Dense-Mode mechanism

The PIM dense mode operation can be summarized in four steps:

Flooding: The multicast source starts sending traffic, the multicast packets are forwarded to **all network links within the multicast-enabled domain**. This flooding behavior ensures that the multicast traffic reaches all parts of the network, regardless of whether there are interested receivers or not.

Distribution Tree: In PIM Dense Mode, the distribution tree is implicitly formed through the **flooding and pruning process**.

Multicast packets are initially flooded throughout the entire network, creating a tree rooted at the source based on **Reverse Path Forwarding (RPF)**. Routers without interested receivers then prune unnecessary branches, resulting in a **source-based distribution tree**.

Prune Messages: Routers that have no interested receivers for a particular multicast group can send **prune messages** upstream towards the source, indicating that they no longer wish to receive traffic for that group. These prune messages propagate back towards the source along the distribution tree, informing upstream routers to stop forwarding multicast traffic for the pruned group on the corresponding interfaces.

State Maintenance: Routers maintain state information about active sources and receivers for each multicast group. This state includes information which interfaces want traffic to be forwarded and which interfaces have sent prune messages to stop.

8.3.1.2. IGMP and the Querier

Although PIM dense mode handles multicast routing between routers, it relies on IGMP to learn about receivers on directly connected networks.

On each local network segment, one router is elected as the ***IGMP querier***. This router periodically sends ***IGMP query messages*** to **discover interested receivers**. Hosts respond with ***IGMP membership reports*** for the **multicast groups they wish to join**. These IGMP messages are not only used by routers, but are also essential for switches running IGMP snooping. **IGMP snooping relies on the presence of a querier** to observe query and report messages in order to build multicast forwarding tables. If no IGMP querier is present, switches with IGMP snooping enabled may not learn any group memberships and can therefore drop multicast traffic. In contrast, **if IGMP snooping is disabled**, multicast traffic is flooded similarly to broadcast, and **no querier is required**.

Based on the received reports, the router determines whether there are active receivers for a given multicast group on an interface. If no host responds, the router assumes that no receivers are present and can prune that interface from the multicast distribution tree.

8.3.1.2.1. Role in Dense Mode

Even though dense mode uses a flood-and-prune approach, IGMP is essential to:

– Detect **whether receivers exist** on a local network
– **Prevent unnecessary multicast traffic** on access links
– **Trigger prune behavior** when no receivers are present

Thus, IGMP complements PIM dense mode by providing receiver awareness at the network edge, enabling more efficient multicast forwarding.

8.3.1.3. Challenges

PIM dense mode suffers from **inefficient bandwidth utilization and resource allocation**, particularly in large networks. The flooding nature of dense mode can result in unnecessary **congestion**, especially in scenarios where only a fraction of devices require multicast traffic. Dense mode is **better suited for smaller networks** or environments where multicast traffic is universally sought.

8.3.2. PIM Sparse Mode

Protocol independent multicast sparse-mode (PIM-SM) is a protocol that is used to route multicast packets in the network more efficiently. It interacts with IGMP to recognize networks, that have members of a multicast group and with the routing table to forward the traffic using the desired path. This is called a “pull” model, because multicast traffic is only forwarded on request.

The basic mechanism of PIM-SM can be summarized as follows:

– Receivers join a multicast group by sending **join messages** toward a ***Rendezvous Point (RP)*** in ASM, forming a shared distribution tree. In SSM, receivers instead join directly toward the source, creating a source-specific distribution tree.
– Routers forward multicast traffic only on interfaces for which they have received explicit join messages from downstream routers.

8.3.2.1. Any Source Multicast (ASM)

When IGMPv1 or IGMPv2 is used, the multicast **source is unknown to receivers**, as they can only join a group `(*,G)`. IGMPv3 allows receivers to specify a source `(S,G)`, which is why it is typically used for Source-Specific Multicast (SSM), although it can also operate in ASM.

Any-Source Multicast (ASM) is a **many-to-many** communication model in which any sender can transmit data to a multicast group, and receivers do not need to know the sender in advance. From the receiver’s perspective, the goal is simply to receive traffic sent to a specific multicast group, denoted as `G`, regardless of the source. This is expressed using the notation `(*,G)`, where `*` represents any source. In ASM, the network is responsible for automatically discovering active sources and delivering their traffic to interested receivers.

8.3.2.2. Rendezvous Point (RP)

If a receiver sends an `IGMP Join (*,G)` to the first-hop router, the first-hop router forwards a PIM Join `(*,G)` hop-by-hop towards the RP.

The RP acts as the **meeting place for sources and receivers**. A source initially sends its traffic to the RP using a **PIM Register tunnel**.

For this process to work, every router in the network must know the location of the RP. This is achieved through a mapping that assigns each multicast group to a specific RP. The mapping can either be configured statically on all routers or learned dynamically via mechanisms such as the ***Bootstrap Router (BSR)*** (RFC 5059) or ***Auto-RP*** (Cisco-proprietary BS).

However, the RP is **only required during the initial phase** of multicast communication. Once the receiver learns about the active source via the RP, the last-hop router can determine the optimal path to the source using the unicast routing table. It then builds a **shortest-path tree** (SPT) directly towards the source. As a result, multicast traffic no longer needs to traverse the RP and instead follows the most efficient path from source to receiver.

8.3.2.2.1. IPv6

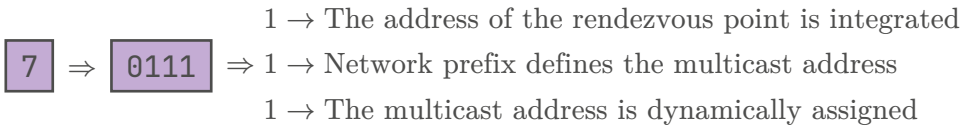
An advantage of IPv6 multicast compared to IPv4 multicast is that the Rendezvous Point address can be included in the multicast address.

Below is illustrated how an IPv6 address of a Rendezvous Point is included in an IPv6 multicast address. The flags are set to 0111. This means that the Network Prefix defines the IPv6 address of the Rendezvous Point and the multicast address is dynamically assigned. With the Network Prefix and the RPaddr field, the address of the Rendezvous Point can be calculated.

The advantage of this method is that the administrator only has to configure the multicast address.



Type	Flags	Scope	Rsvd	RpAddr	Plen	Network Prefix	GroupID
FF	7	8	0	1	40	2001:DB8:12:0	1111:2222
8 Bits	4 Bits	4 Bits	4 Bits	4 Bits	8 Bits	64 Bits	32 Bits



8.3.2.3. Reverse Path Forwarding (RPF)

To ensure that multicast packets follow correct and **loop-free paths**, routers rely on the concept of Reverse Path Forwarding (RPF).

RPF is a mechanism that verifies whether a multicast packet has arrived on the correct interface. A router performs an RPF check by consulting its **unicast routing table** and determining the **interface it would use** to reach the source (or RP). If the multicast packet arrives on that interface, it is accepted and forwarded; otherwise, it is discarded. This ensures efficient forwarding and prevents routing loops.

8.3.2.4. Source Specific Multicast (SSM)

In SSM, the receiver knows the exact source from which it wants to receive multicast traffic. This simplifies the multicast process, as a **shortest-path tree can be built directly toward the source without the need for a RP**.

The receiver subscribes to a channel using IGMPv3, which provides the **first-hop router** with both the source IP address and the multicast group address. As a result, PIM can immediately construct a source-specific tree `(S,G)`.

In SSM, no shared trees `(*,G)` are used. Multicast state is built exclusively as shortest-path trees toward the source. An SSM channel is therefore identified by an `(S,G)` pair, where `S` is the source address and `G` is the group address.

IANA has reserved the IPv4 address range of `232.0.0.0/8` for PIM SSM. It is recommended to allocate SSM multicast groups using that range.

PIM-SSM requires that the successful establishment of an `(S,G)` forwarding path from the source `S` to any receiver(s) depends on hop-by-hop forwarding of the explicit join request from the receiver toward the source. The receivers send an **explicit join** to the source because they have the source IP address in their join message with the multicast address of the group. PIM-SSM **leverages the unicast routing topology to maintain the loop-free behaviour**.

The receivers can receive traffic only from designated `(S,G)` channels to which they are subscribed, which is in contrast to ASM, where receivers need not know the IP addresses of sources from which they receive their traffic.

8.3.3. PIM Sparse-Dense Mode

In PIM sparse-dense mode we can use sparse or dense mode for each multicast group.

8.4. MULTICAST ROUTING CONCEPTS

Unicast routing protocols like OSPF are forward-looking, the routing information or routes stored in the routing database provide information on the destination.

The opposite is true for multicast routing. The objective is to send multicast messages away from the source towards all branches or segments of the network interested in receiving those messages. Messages must always flow away from the source, and never back on a segment from where the transmission originated. This means rather than tracking only destinations, multicast routers must also track the location of sources, the inverse of unicast routing. This method is called reverse path forwarding (RPF).

8.4.1. RPF Check

Even though multicast uses the exact inverse logic of unicast routing protocols, you can leverage the information obtained by those protocols for multicast forwarding.

In the case of a Shared Tree with a Rendezvous Point, the RPF check is carried out against the Rendezvous Point, because it is the one that knows the source.

8.4.2. The (*,G) multicast routing table entry

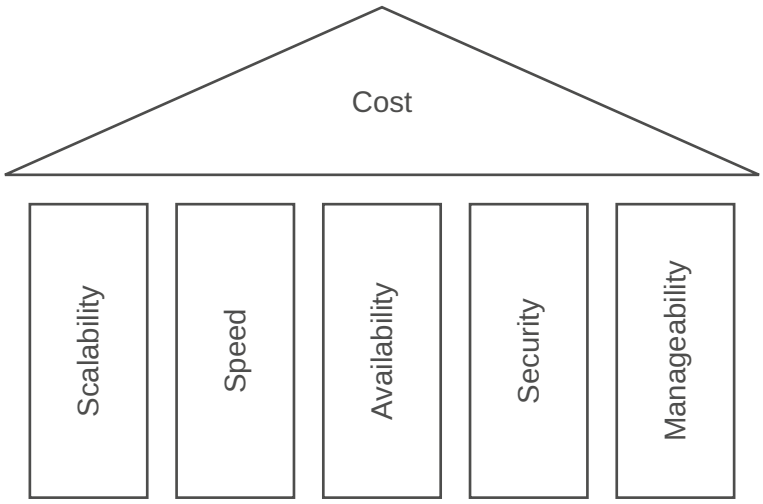
IGMP hosts sends an IGMP membership report also called ***IGMP Join***. The router adds the `(*,G)` entry to the ***mroute table***.

8.4.3. The (S,G) multicast routing table entry

In order to build a tree with an `(S,G)` the router needs to receive an `(S,G)` join or `(S,G)` membership report from hosts via IGMP.

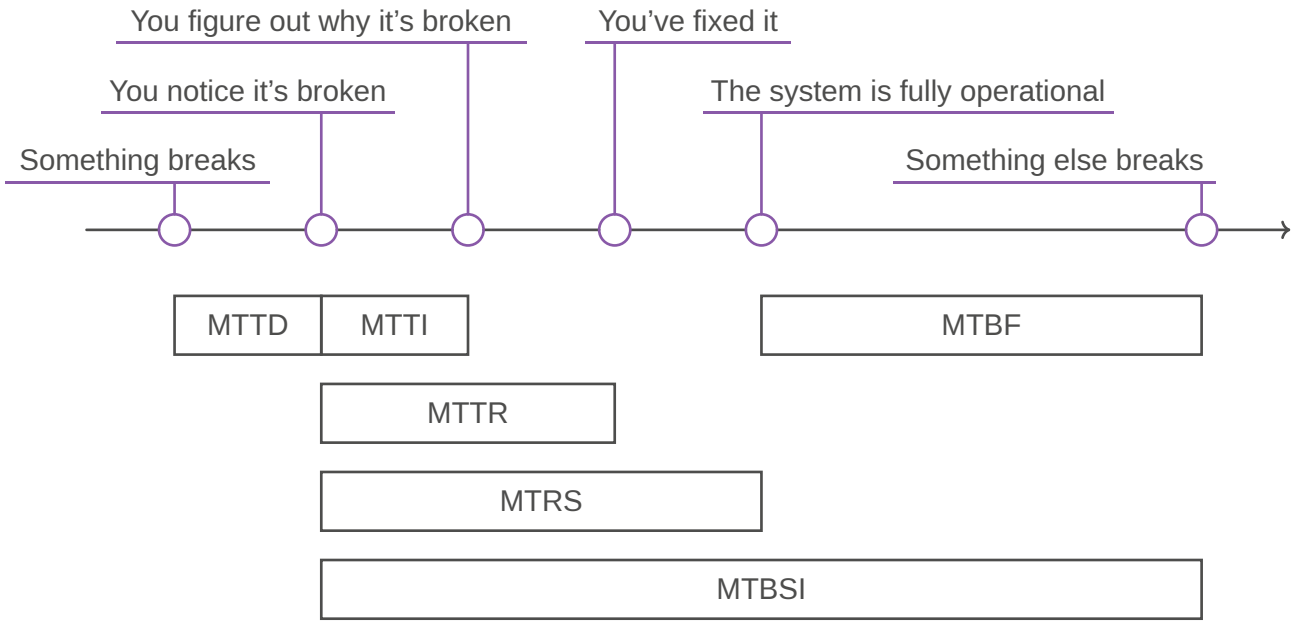
After a source for a group is known by the router, it adds the `(S,G)` to the multicast routing table.

9. NETWORK DESIGN



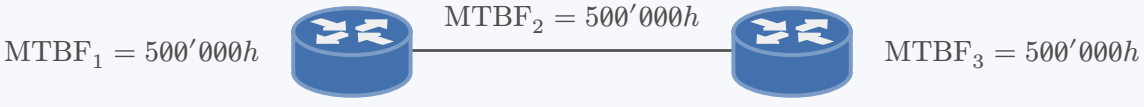
RFC 1925

9.1. AVAILABILITY



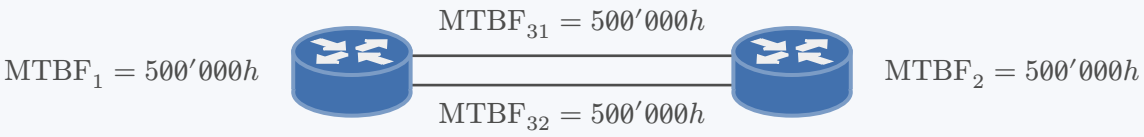
Term	*Definition*
MTBF	Mean Time Between Failures
MTTR	Mean Time To Repair (Resolve)
MTTD	Mean Time To Detect
MTTI	Mean Time To Identify
MTRS	Mean Time To Restore Service
MTBSI	Mean Time Between Service Incidents
Availability	$\frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$
MTBF combined	$\left(\sum_{n=1} \frac{1}{\text{MTBF}_n} \right)^{-1}$
MTBF parallel	$\sum_{n=1} \frac{\text{MTBF}_n}{n}$

Example 7: **MTBF combined**



$$\text{MTBF}_{\text{combined}} = \frac{1}{\frac{1}{\text{MTBF}_1} + \frac{1}{\text{MTBF}_2} + \frac{1}{\text{MTBF}_3}} = \frac{1}{\frac{1}{500000} + \frac{1}{500000} + \frac{1}{500000}} \approx 166'666h$$

***Example** 8:* **MTBF parallel**



$$\mathrm{MTBF}_{\text{parallel}} = \frac{\mathrm{MTBF}_{31}}{1} + \frac{\mathrm{MTBF}_{32}}{2} = \frac{500000}{1} + \frac{500000}{2} = 750'000h$$

9.1.1. Reasons for downtime

Hardware failure: Faults, config changes, network congestion, …

55%

Human error: Device mismanagement, accidental deletions, …

22%

Software failure: Failure to upgrade or patch software, security attacks, …

18%

Natural disaster: Floods, storms, earthquakes, …

5%

9.1.2. Redundancy

Redundancy is a tradeoff

- Adds complexity
  - Source of errors
  - Network Addressing
  - Routing / Traffic Flow
  - Which redundancy mechanism is appropriate?

Adding links (paths) increases the MTBF

- However, it also increases the MTTR
  - Increasing parallelism increases routing complexity
  - Thus, increasing convergence time
  - Convergence time is directly tied to the MTTR

Network Design is not trivial

Consider Backup Paths vs. Load Balancing

- Backup Paths:
  - Duplicate devices/links on the primary path
  - Build extra link to provide redundancy
  - Questions arise:
  - How much capacity does the backup link need?
  - How quickly will the network begin to use the backup path?
- Load Balancing:
  - ECMP
  - Ether- / Port-Channel

9.2. SCALABILITY

Things to consider when planning for expansion:

- What is the bandwidth need and future growth?
- How many more sites will be added in the next years?
- How many more users?
- How many more servers?
- How many more tenants (customers)?

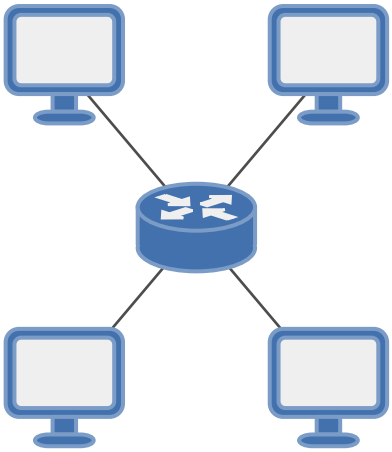
Scalability constraints

- Broadcasts
- Limitations
  - Addresses
- Separations of applications/tenants/customers

9.3. TOPOLOGY

A topology describes how a network is connected.

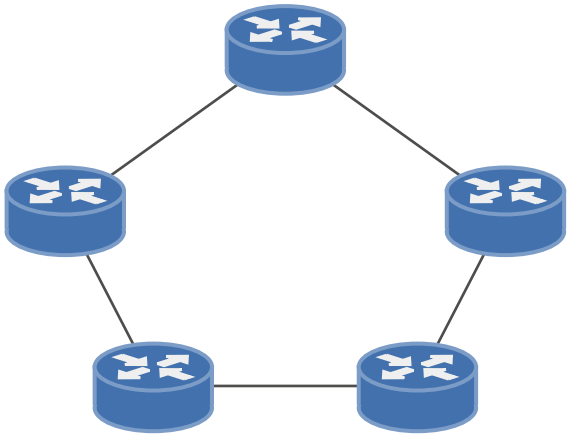
Star



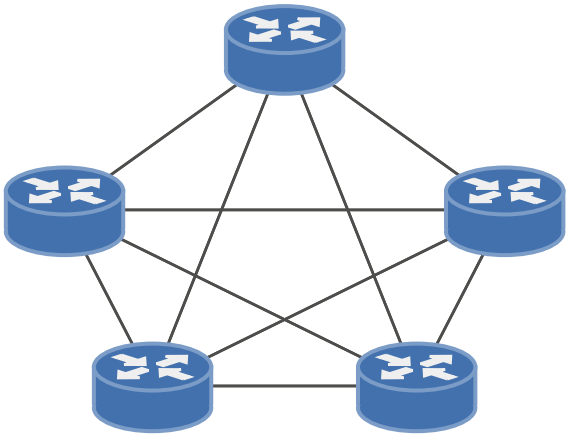
Bus



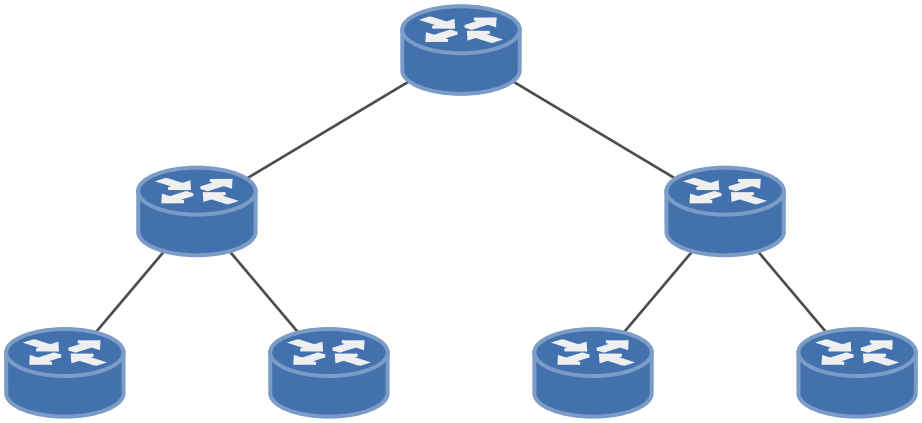
Ring



Full Mesh



Tree



Hierarchical	*Flat*
Eg. Tree	Eg. Ring

Divide and conquer
– Each layer has clearly defined tasks
– Allow summarization
  – Addressing
  – Traffic / Capacity Planning
 	
Improve Scalability
– Think about adding components
  – For Capacity
  – For More Ports
	– Small networks
	– Any-to-Any communication
	  – MPLS VPN
	  – LAN

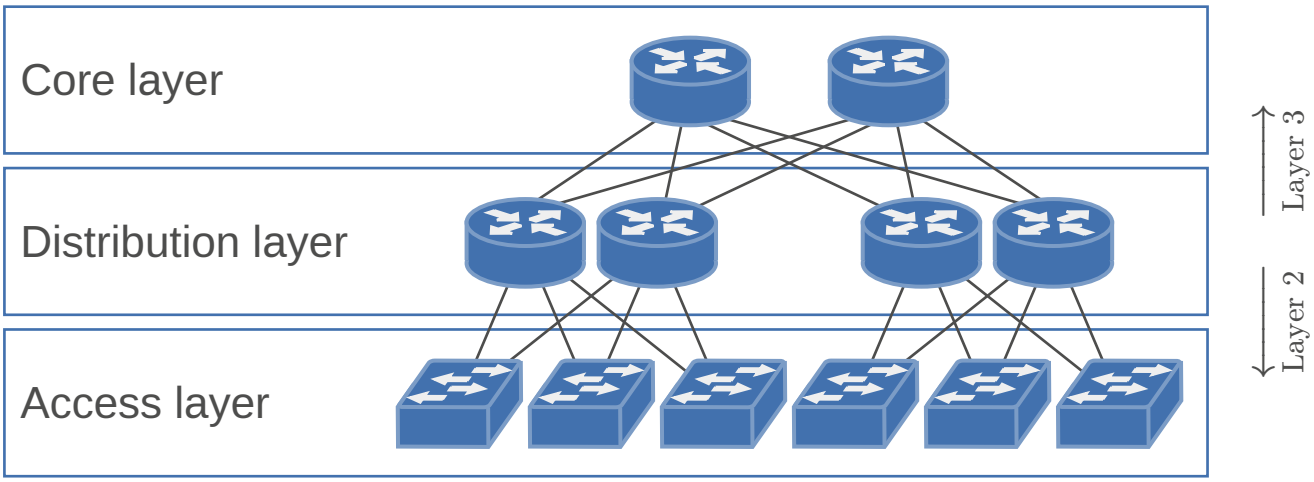
9.4. CAMPUS

– Enterprise network with hundreds/thousands of user
– More than one LAN (Local Area Network)
– Limited geographical area
  – connects multiple buildings
– Connected via Ethernet (and Wireless)
– One company owns the hardware

Traditional / Hierarchical Design Model	*Emerging Technologies / Fabric*

– Most common design
– Typically, consists of three layers
  – Core
  – Distribution
  – Access
	– Newer technologies
	– Underlay/Overlay networks
	  – Software-Defined Networking (SDN), Software-Defined Access (SDA)
	  – EVPN
	  – Segment Routing (SR)
	  – (MPLS)

9.4.1. Hierarchical



- Each layer has specific functions/capabilities
  - Simplifies network design, deployment and management
- Design elements can be changed easily
  - Adds scalability/modularity
  - Use the same “building blocks”
- Changes in the network affect a small subset of the network

9.4.1.1. Core

- Backbone: connects different distribution layer switches
  - Large networks
  - Geographically reasons e.g. several buildings
  - Simplifies design, otherwise full-mesh between distribution layer
- Simple but fast
- Only layer 3
  - Drives resiliency and stability
- Design criteria:
  - Scalability
  - Capacity
  - Redundancy

9.4.1.2. Distribution

- Aggregate data from multiple access switches and connect to the core
- Design Simplification
  - Scalability
  - Smaller Fault Domains
  - Redundancy
- Usually, Layer 3 Boundaries
  - Load Balancing
  - Inter-VLAN Routing
  - Optimizations: Route summarization and fast convergence, loop protection
  - Security Policies
  - QoS enforcement

9.4.1.3. Access

- Connects user devices/end-points to network
- High port density but low cost
- Power over Ethernet (PoE)

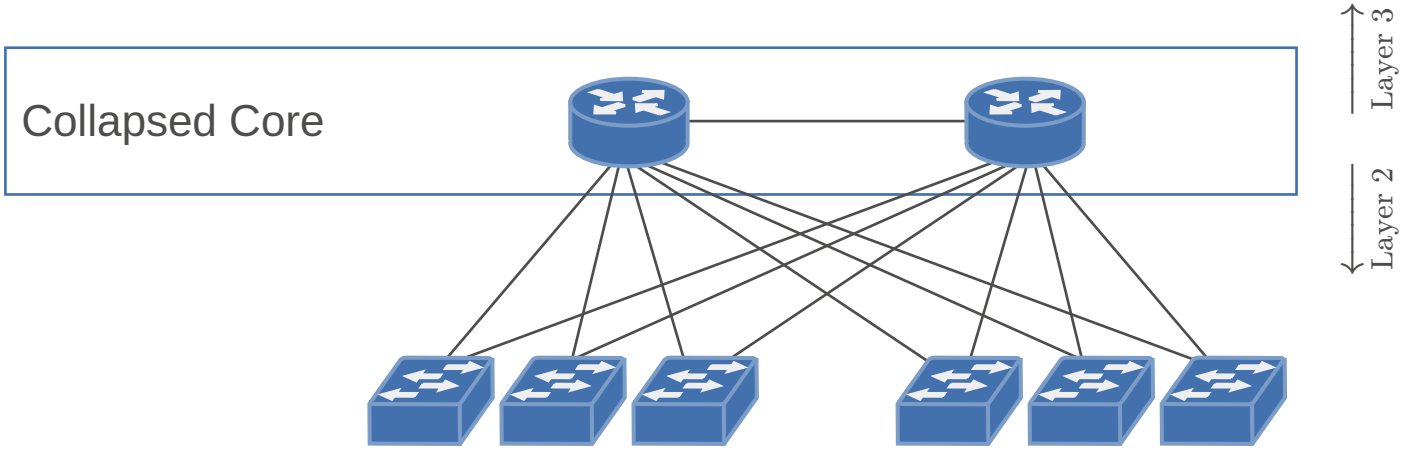
- Network Access Control
- QoS classification
- L2 features
  - VLAN
  - STP
  - IGMP snooping
  - DHCP snooping
  - Etc.

9.4.1.3.1. Simplified access

Switch Stacking: Merging multiple physical switches into one large logical switch

- Create a switch stack at the distribution layer
- Easier management
  - No FHRP
  - Single Point of Managmenet
- No loop
- Scalability
  - Add switches to stack easily

9.4.1.4. Collapsed core



- Dual Role:
  - Core and distribution combined
- Access Layer Aggregation
- Service Connectivity e.g.
  - External Services: WAN/Internet
  - WLAN
- Reduces complexity
- Limited Scalability

9.4.2. First Hop Redundancy

- Provide a resilient default gateway/first hop address to end-stations
- Different First Hop Redundancy Protocols (FHRP)
- Leverage Timers for Fast Failover
- Optimize Timers for Smooth Transitions
  - Goal: As few blackholed traffic as possible

9.4.2.1. Virtual Router Redundancy Protocol (VRRP)

- A group of routers function as one virtual router by sharing one virtual IP address (and each one its own MAC address)
- One (master) router performs packet forwarding for local hosts
- The rest of the routers act as “backup” in case the master router fails
- Backup routers stay idle as far as packet forwarding from the client side is concerned
- IETF Standard RFC 3768
- VRRP if you need multivendor interoperability

9.4.2.2. Hot Standby Router Protocol (HSRP)

- A group of routers function as one virtual router by sharing one virtual IP address and one virtual MAC address
- One (active) router performs packet forwarding for local hosts
- The rest of the routers provide “hot standby” in case the active router fails
- Standby routers stay idle as far as packet forwarding from the client side is concerned
- **Cisco proprietary**

9.4.2.3. Gateway Load Balancing Protocol (GLBP)

- All the benefits of HSRP plus load balancing of default gateway
  - Utilizes all available bandwidth
- A group of routers function as one virtual router by sharing one virtual IP address but using multiple virtual MAC addresses for traffic forwarding
  - Active Virtual Gateway (AVG) responds to ARP
  - Active Virtual Forwarder (AVF) is used for forwarding
  - AVG sends virtual MAC addresses of AVFs
- **Cisco proprietary**

9.4.3. Software Defined Access

- “One controller to rule them all”
- Example: Overlay Protocols
  - VXLAN / LISP
  - VXLAN / EVPN
- Decoupling Layer 2 / Layer 3
- Anycast Gateway
- Automation
- Simplification

9.4.3.1. Ethernet VPN (EVPN)

A technology for carrying layer 2 Ethernet traffic as a virtual private network using wide area network protocols. EVPN technologies include Ethernet over Multiprotocol Label Switching (MPLS) and Ethernet over Virtual Extensible LAN (VXLAN).

- Anycast Gateway

9.5. DATA CENTER

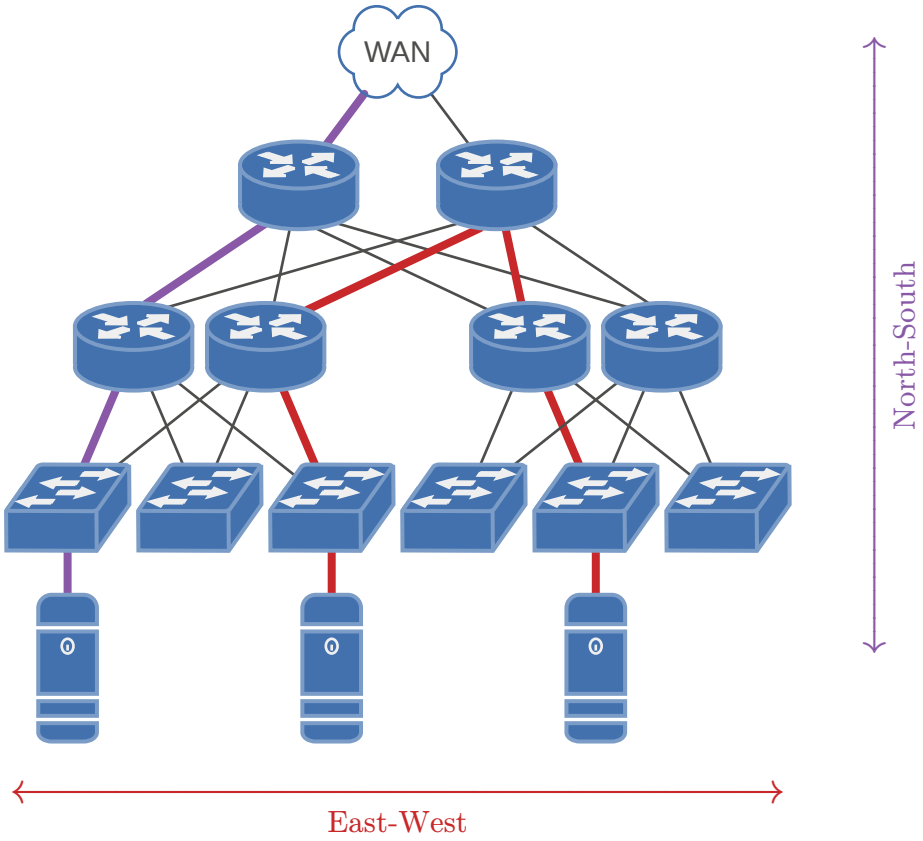
- Predominant East-West traffic
- Campus Requirements apply
- Additional Requirements
  - Agility
  - Multitenancy
  - Scalability

– Depends on the data center type
  – Enterprise Data Center differs from Cloud Data Center
– IP network
– Storage network

9.5.1. Data Center Tiers

Parameters	*Tier 1*	*Tier 2*	*Tier 3*	*Tier 4*
Uptime guarantee	99.671%	99.741%	99.982%	99.995%
Downtime per year	<28.8 hours	<22 hours	<1.6 hours	<26.3 minutes
Price	$	$$	$$$	$$$$
Compartmentalization	No	No	No	Yes

9.5.2. Traffic patterns



North-South traffic	*East-West traffic*
– Traffic traveling between different network scopes (e.g., internal and external) – Entering or Leaving Data Center – Client-Server traffic	– Traffic that stays within the same network scope – Does not leave data center(s) – Traffic between servers – Traffic between server and storage

9.5.3. Logical Topology

Web Tier
↓
Application Tier

↓
Database Tier

9.5.4. Three-Tier Data Center Architecture

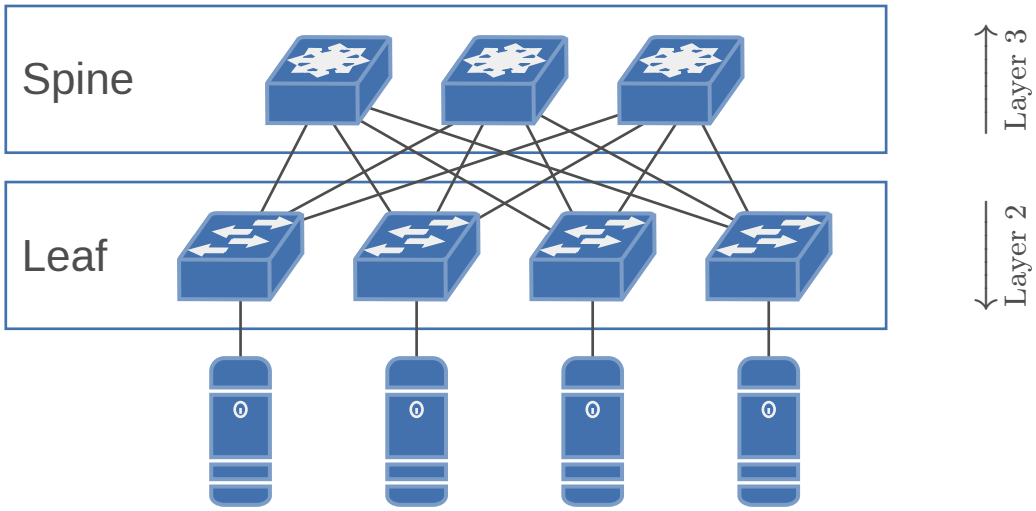
- Similar to Access-Distribution-Core “Hierarchical” (Page 58)
  - Distribution is called Aggregation
- Also called Fat Tree Data Center Network
- Server connect to Access layer

Problems:

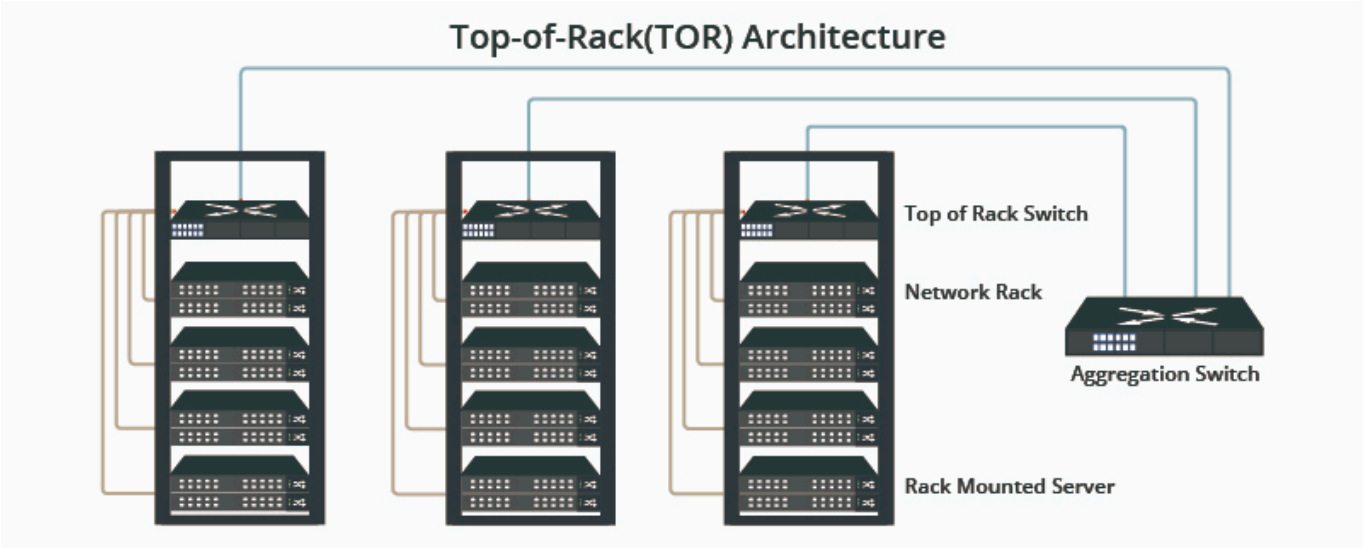
- Server Virtualization increased East-West traffic
  - VMs / Containers
- East-West traffic goes over all layers
- Three-Tier Data Center designed for North-South traffic
  - Failed to adapt to modern workloads
- Nowadays we need Layer 2 connectivity

9.5.5. Leaf Spine Architecture

- Sometimes called two-tier architecture
- Solution for modern workloads
- Brings a lot of advantages
  - Scalability – easy expansion
  - Reduce latency
  - Load-Balancing through ECMP
  - Fault Tolerance
  - No L2 problem thanks to Fabric
    - EVPN / VXLAN
    - Or similar technologies

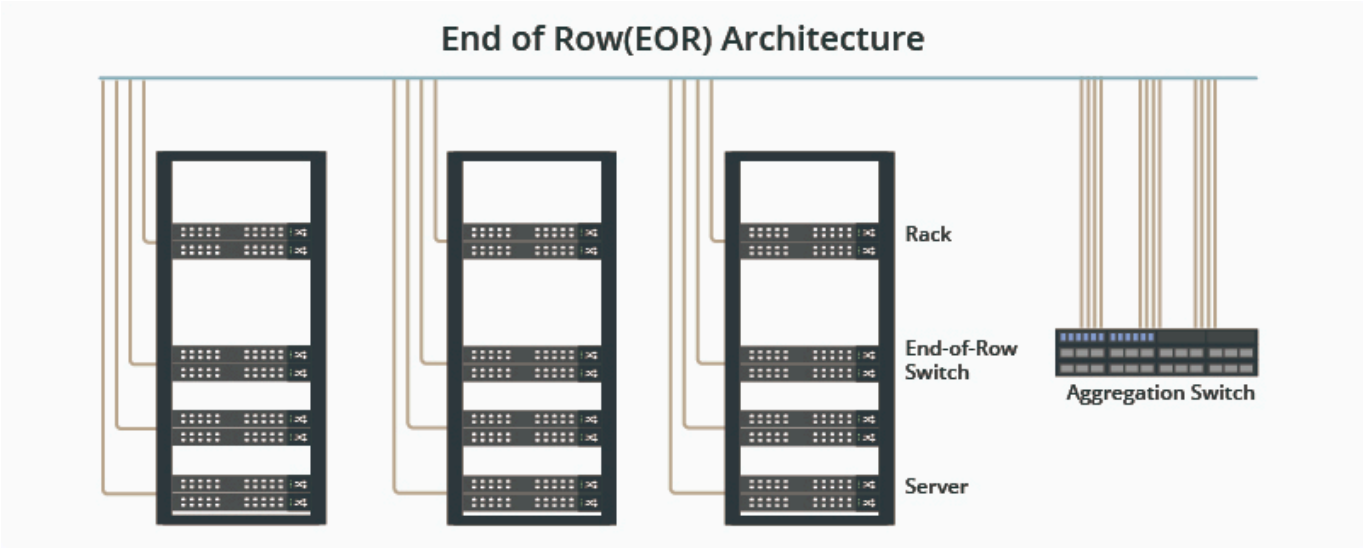


9.5.6. Top of Rack (ToR)



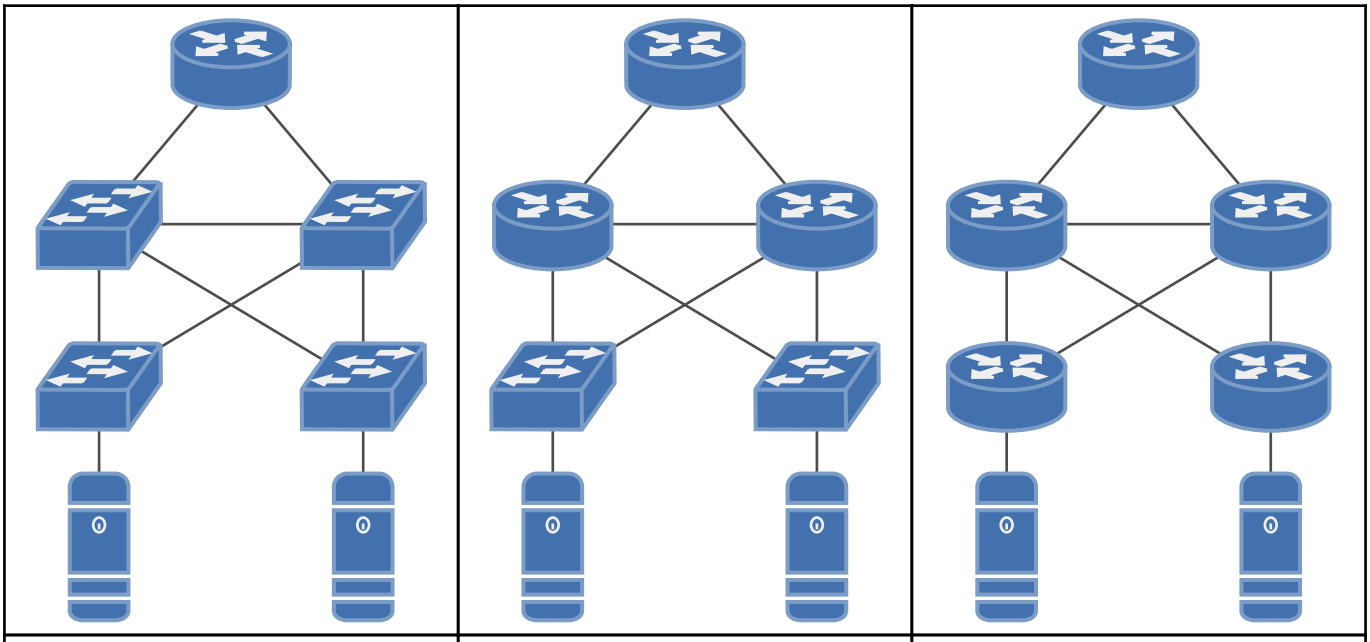
[1]

9.5.7. End of Row (EoR)



[2]

9.5.8. Comparison

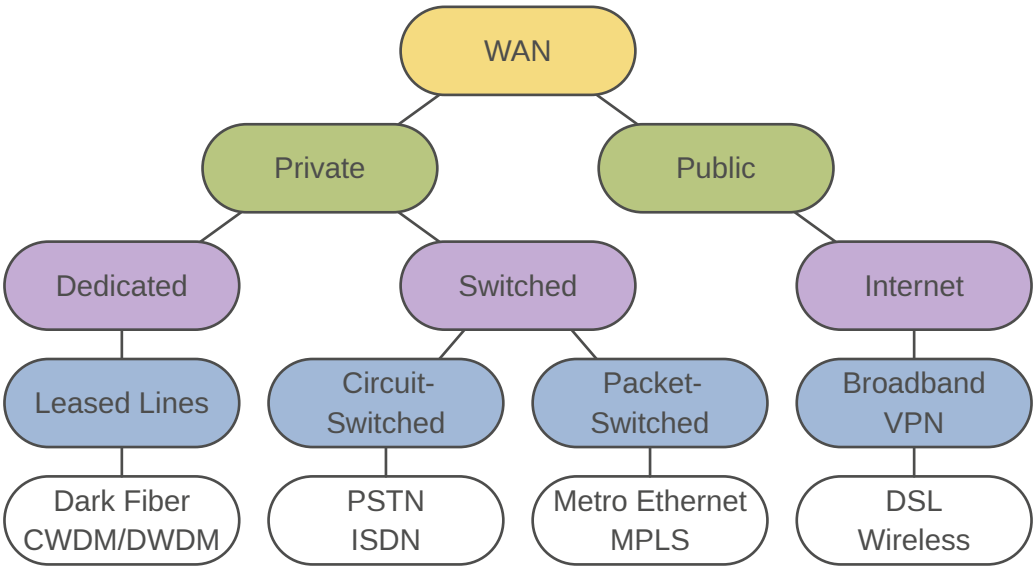


Problem: STP		Problem: Different networks

10. WAN

Wide-area networks (WANs) are used to connect remote LANs.

WAN link connection options:



10.1. PRIVATE WAN

10.1.1. Leased Lines

Point-to-point lines leased from a service provider.

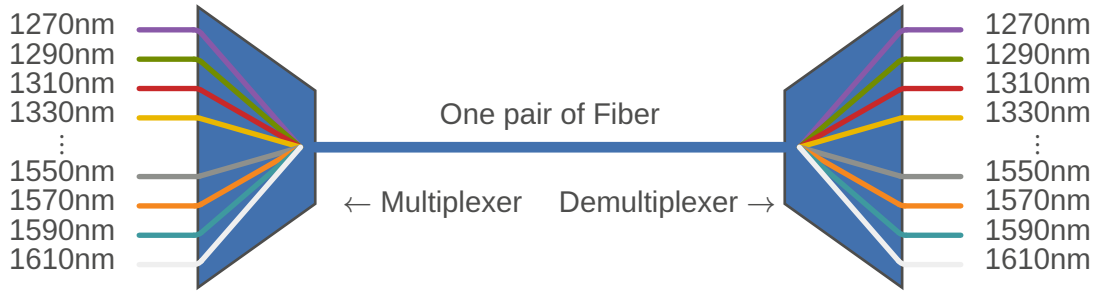
– The organization pays a monthly lease fee to a service provider to use the line
– The Layer 2 protocol is usually Ethernet
– Sometimes called private circuit

10.1.1.1. Dark Fiber

Physical fiber leased from a service provider. Extremely expensive and very difficult to get because Service Provider prefers to run services and sell Lambdas over it.

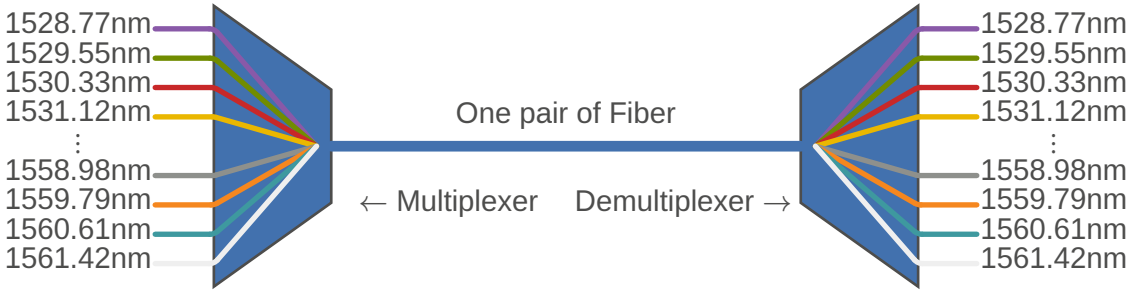
10.1.1.2. Coarse wavelength division multiplexing (CWDM)

– 16 CWDM Lambdas can be transmitted over one physical optic fiber
– 1270nm - 1610nm with 20nm of interval
– Maximum distance 120km
– MUX are passive equipments (only optics, no electronics)
– Cheaper solution in comparison with DWDM



10.1.1.3. Dense wavelength division multiplexing (DWDM)

– Can multiplex more than 80 different channels (wavelengths) of data onto a single fiber
– Assigns incoming optical signals to specific wavelengths of light
– Wavelength of 1528nm - 1563nm with an interval of 0,8nm
– Maximum distance 1000km
– Each channel is capable of carrying a 10Gb/s multiplexed signal
– Used in all modern submarine communications cable systems



10.1.2. Circuit Switching

Dynamically **establishes a dedicated circuit** (or channel or virtual connection) for voice or data between a sender and a receiver using a signaling protocol.

10.1.2.1. Integrated Services Digital Network (ISDN)

ISDN changes the internal connections of the public switched telephone network (PSTN)

– *Analog signals changed to time-division multiplexed (TDM) digital signals*
– *TDM allows two or more signals, or bit streams, to be transferred as sub channels*
– *ISDN is a legacy technology that has been replaced by high-speed Digital Subscriber Line (DSL) and other ethernet services*

10.1.3. Packet Switching

A packet-switched network (PSN) **splits traffic data into packets** that are routed over a shared network.

– Packet-switching networks **do not require a circuit** to be established

– The switches determine the forwarding of the packets based on the addressing information in each packet

10.1.4. Connection-oriented vs Connectionless

Connection-oriented systems	*Connectionless systems*
The network predetermines the route for a packet, and each packet has to carry an identifier	Full addressing information must be carried in each packet
ATM, Frame-Relay	MPLS, Internet, Metro Ethernet

10.1.4.1. Connection-oriented Systems

LEGACY!

Frame-Relay: PVCs (Permanent Virtual circuits) support data rates up to 4 Mb/s

ATM (asynchronous transfer mode): ATM VCs support link speeds up to 622 Mb/s

10.1.4.2. Connectionless systems

10.1.4.2.1. Metroethernet

Ethernet was originally developed to be a LAN access technology
– Ethernet standards IEEE 1000BASE-SX supports fiber-optic cable lengths of 550m
– Ethernet standards IEEE 1000BASE-LX supports fiber-optic cable lengths of 5km
– Ethernet standards IEEE 1000BASE-ZX supports cable lengths up to 70km
– The distance are even extended further thanks to Ethernet over MPLS (EoMPLS) and VPLS (Virtual Private Lan Service)

10.2. PUBLIC WAN

10.2.1. VPN Types

Type	*Properties*
Site-to-Site	– Fixed locations – Devices usually locked to IP Address
Remote Access	– Changing locations – Devices not locked to IP Address

10.2.2. Common VPN Protocols

Protocol	*Key Characteristics*	*Flexibility*
PPTP	Basic VPN protocol based on PPP. Specification lacks built-in encryption/authentication . Relies on PPP tunneling for security.	Limited built-in security features.
IPSec IKEv2	Part of the IPSec protocol suite. Standardized in RFC 7296 . De facto standard for secure internet communication. Provides confidentiality, authentication, and integrity.	Highly flexible and widely compatible.
OpenVPN	Open-source VPN protocol developed by OpenVPN Technologies. Not based on standards (RFC). Uses custom security protocol and SSL/TLS for key exchange. Provides full confidentiality, authentication, and integrity.	Highly flexible and configurable.

Protocol	Key Characteristics	Flexibility
WireGuard	Extremely fast VPN protocol with very little overhead and state-of-the-art cryptography. Developed by WireGuard. Potential for simpler, more secure, more efficient, and easier to use VPN.	Modern design aims for simplicity and ease of use.

10.2.3. Software-Defined WAN (SDWAN)

A software-defined wide area network (SD-WAN) connects local area networks (LANs) across large distances using controlling software that works with a variety of networking hardware.

10.3. MULTI PROTOCOL LABEL SWITCHING (MPLS)

RFC 4364 RFC 5036 RFC 3031

MPLS is **multiprotocol** and supports i.a.

– L3 payloads (IPv4, IPv6)
– L2 payloads (Ethernet, ATM Frame Relay, PPP and HDLC)

MPLS switching is based on **labels** instead of IP network addresses

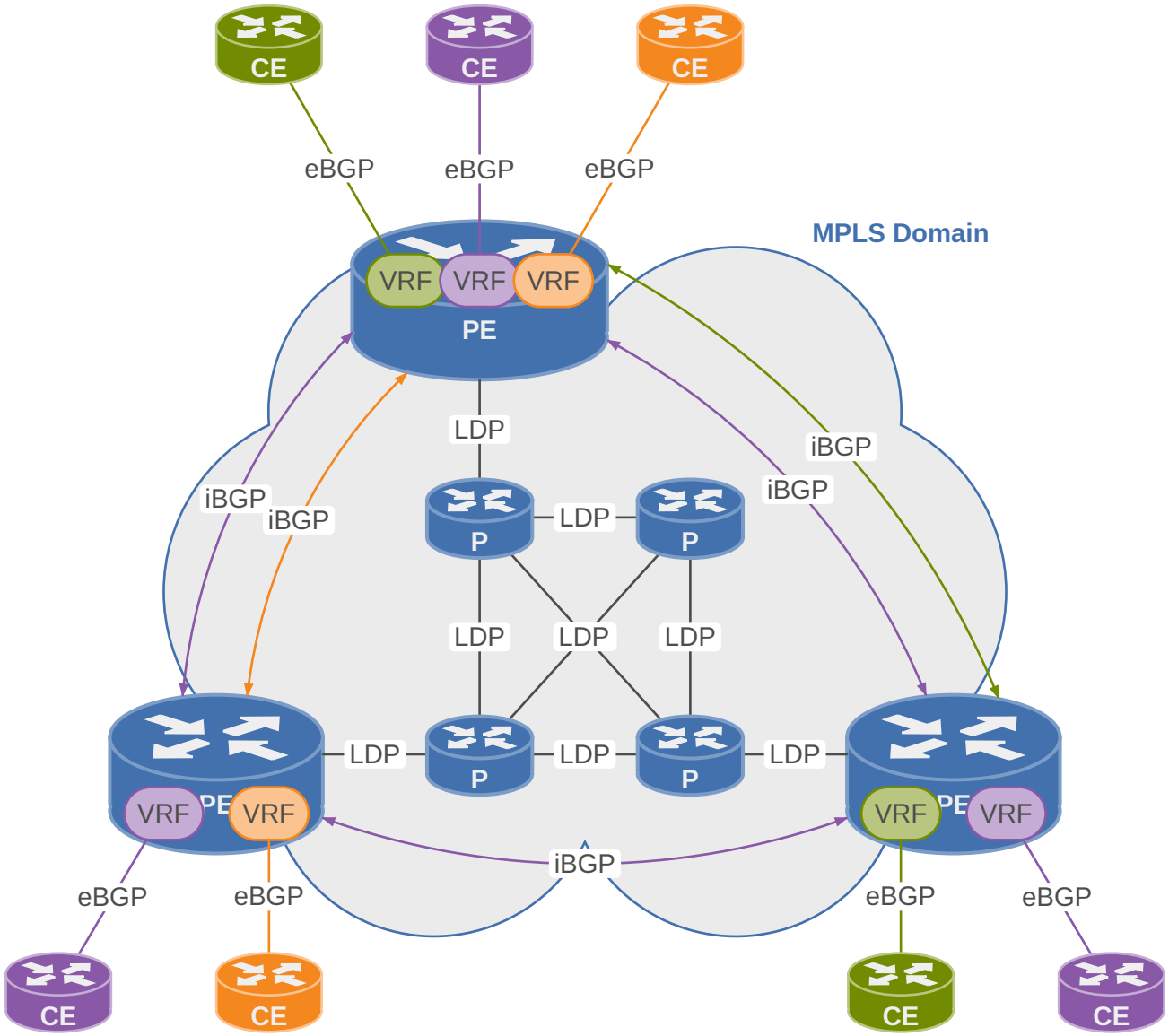
– Label Switched paths are built between distant routers (PEs)
– Only the PEs route IPv4 and IPv6 packets

Router type	Definition
LSR	Label Switched Router: Forwards labeled packets
Edge LSR	– Labels IP packets and forwards them into the MPLS domain – Removes labels and forwards IP packets out of the MPLS domain

– On ingress, a label is assigned and imposed by the IP routing process
– LSRs in the core swap labels based on the contents of the LFIB
– On egress, the label is removed and a routing lookup is used to forward the packet

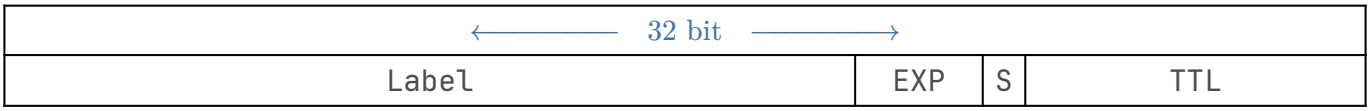
10.3.1. Router Types

Router type	Definition
P (Provider)	– Does not have a direct link to a CE router – Switches MPLS-labeled packets → Label Switched Router (LSR) – Runs an IGP and LDP
PE (Provider Edge)	– Shares a link with at least one CE router – Imposes and removes MPLS labels – Provides iBGP and VRF tables – Runs an IGP, LDP and MP-BGP
CE (Customer Edge)	– Has no knowledge of MPLS protocols and does not send any labeled packets but is directly connected to an MPLS router (PE) – Connects customer network to MPLS network



10.3.2. Header

– A new ether type (0x88 47) is used for a MPLS packet
– When entering the MPLS network, the L3 packet is encapsulated with an MPLS header



Field	*Definition*
Label (20 bit)	Label used for switching
EXP (3 bit)	Experimental field. Allows for QoS (Quality of Service) marking
S (1 bit)	Bottom-of-stack indicator. Indicates the presence of an additional MPLS label
TTL (8 bit)	Equal to IP TTL

With MPLS VPN, there will be a stack of two MPLS headers

– The top label is used for the Label switched path (LSP)
– The second (inner) label identifies the VPN

10.3.2.1. TTL

With MPLS TTL propagation, a traceroute command would receive ICMP Time Exceeded messages from each of the routers. However, many service providers do not want hosts outside the MPLS network to have visibility into the MPLS network.

It is possible to **disable MPLS TTL propagation**, the ingress PE then sets the MPLS header's TTL field to 255, and the egress PE leaves the original IP header's TTL field unchanged. As a result, the entire MPLS network appears to be a single router hop from a TTL perspective.

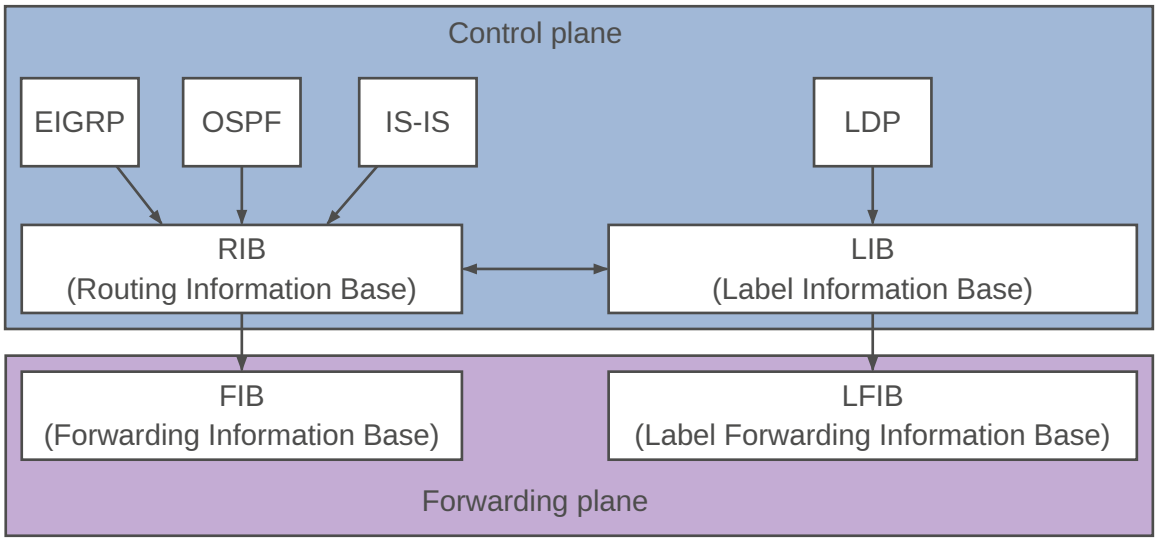
Router Type	*Behavior*
Ingress PE routers	After an ingress PE router **decrements the IP TTL field**, it pushes a label into an unlabeled packet and then **copies the packet's IP TTL field into the new MPLS header**'s TTL field.
P routers	When a P router swaps a label, the router **decrements the MPLS header's TTL** field and always **ignores the IP header**'s TTL field.
Egress PE routers	After an egress PE router **decrements the MPLS TTL field**, it pops the final MPLS header and then **copies the MPLS TTL field into the IP header** TTL field.

10.3.3. RIB/FIB and LIB/LFIB

Term	*Definition*
RIB	Raw reachability information about available networks, gathered with routing protocols such as IS-IS, OSPF or BGP, is stored in the Routing Information Base (RIB). The RIB is the **superset of the FIB**. The RIB may contain duplicate entries for the **same network prefix with different costs and from different protocols**.
FIB	The Forwarding Information Base (FIB) is built by an algorithm which **picks the best entries per network prefix from the RIB** and installs the selected entry in the FIB. The **data plane** makes forwarding decisions based on longest prefix matching (LPM) on the entries in the FIB.
LIB	When we use LDP, we locally **generate a label for each prefix that we can find in the RIB**, except for BGP prefixes. This information is then added to the Label Information Base (LIB). The LIB is the equivalent to the RIB for MPLS labels.
LFIB	The information in the LIB is used to build the Label Forwarding Information Base (LFIB). When the router has to forward a packet with an MPLS label on it, it will use the LFIB for forwarding decisions. The LFIB contains a **subset of the entries in the LIB based on the best LSP** (Label Switch Path). The LFIB is the equivalent to the FIB for MPLS labels.

The Label Information Base (LIB) is built using the global routing table. That means that the packets flow over the same path as they would have if MPLS was not used.

MPLS unicast IP forwarding is a key piece of the L3 VPN solution. It is used for the connectivity in the backbone. MPLS requires the use of control plane protocols (e.g. OSPF and LDP) to learn labels, correlate those labels to particular destination prefixes, and build the correct forwarding tables.



10.3.4. Control plane

Responsible for building the tables and making the routing decisions that will be used by the data plane.

MPLS unicast IP forwarding uses an IGP, like OSPF and one MPLS-specific control plane protocol: LDP.

10.3.4.1. Label Distribution Protocol (LDP)

Advertises labels for each prefix listed in the IP routing table. **Triggered by a new IP route** appearing in the unicast IP routing table. Upon learning a new route, the MPLS router allocates a label called a local label.

LDP establishes a session by performing the following
1) On all interfaces that are enabled for MPLS **hello messages are periodically sent**.
2) MPLS enabled routers **respond to received hello messages by attempting to establish a session** with the source of the hello messages.

– **UDP is used for hello** messages. It is targeted at “all routers on this subnet” multicast address (`224.0.0.2`)
– **TCP is used to establish the session**
– Both TCP and UDP use well-known LDP port number 646

10.3.5. Data plane

High performance component of a network device **responsible for the forwarding of packets**. Does forwarding decisions based on the information in the FIB provided by the control plane.

10.3.5.1. Routing decisions

MPLS routers **rely on the routing protocol’s decision** about the best route and can thus take advantage of the routing protocol’s loop prevention features and react to the routing protocol’s choice for new routes when convergence occurs.

In short, an LSR makes the following decisions:
– For each route in the routing table, find the corresponding label information in the LIB, based on the **outgoing interface and next-hop router** listed in the route.
– Adds the corresponding label information to the FIB and LFIB.

10.3.5.2. Label Allocation

Label allocation and distribution in an MPLS network follows these steps:

1) IP routing protocols build the IP routing table
2) Each LSR assigns a label to every destination in the IP routing table independently

3) LSRs announce their assigned labels to all other LSRs
4) Every LSR builds its LIB, LFIB, and FIB data structures based on received labels

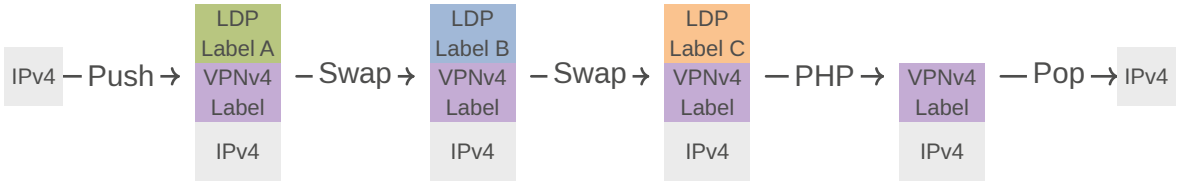
10.3.6. Label Operations

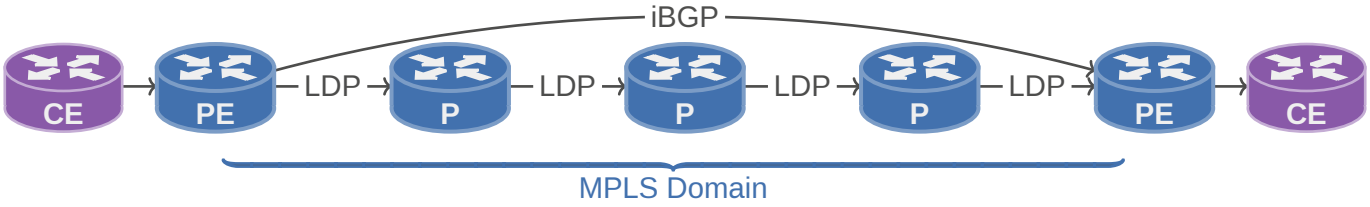
Label imposition (Push): By ingress PE router; classify and label packets

Label swapping or switching: By P router; forward packets using labels

Label disposition (Pop): By egress PE router; remove label and forward original packet to destination CE

Penultimate Hop Popping (PHP): A label is removed on the router before the last hop within an MPLS domain.





10.3.7. L3 VPN

MPLS L3 VPNs use **MP-BGP** to overcome some challenges when connecting an IP network to a large number of customer IP networks, including the issue of dealing with **duplicate IP address spaces** with many customers. The MPLS L3 VPN RFCs define the concept of using **multiple routing tables**, called ***VRF*** tables, which separate customer routes to avoid the duplicate address range issue.

10.3.7.1. Virtual Routing and Forwarding (VRF) Table

The use of separate tables solves the problems of preventing one customer's packets from leaking into another customer's network because of overlapping prefixes. Typically, routers need at least **one VRF for each customer** attached to that particular router. Each VRF has three main components:

– An IP routing table (RIB)
– A CEF FIB, populated based on that VRF's RIB
– A separate instance or process of the routing protocol used to exchange routes with the CEs that need to be supported by the VRF.

10.3.7.2. Multi-Protocol BGP (MP-BGP)

RDs allow BGP to advertise and **distinguish between duplicate IPv4 prefixes**. Each NLRI (prefix) is advertised as the traditional IPv4 prefix and adds another number (the RD) that uniquely identifies the route. The new NLRI format, called ***VPNv4***, consists of the following two parts:
– A 64-bit RD
– A 32-bit IPv4 prefix

10.3.7.3. Route Distinguishers (RD)

The RD should follow the following possible formats:
– 2-byte-integer:4-byte-integer
– 4-byte-integer:2-byte-integer
– 4-byte-dotted-decimal:2-byte-integer

In all three cases:
– The first value (before the colon) should be either an **ASN** or an **IPv4 address**.
– The second value, after the colon, can be any value you want.

10.3.7.4. Route Targets (RT)

MPLS uses Route Targets (RTs) to control **which routes a PE router imports into which VRFs**. When a PE router exports a route into BGP, it stamps the route with an RT value. When a remote PE receives that route, it checks whether any local VRF is configured to import that RT. If a match is found, the route is installed into that VRF; if not, the route is discarded. This mechanism ensures strict traffic separation between customers. PEs advertise RTs in BGP Updates as **BGP Extended Community path attributes**. RT values follow the same basic format as RD values. RT's are not unique. A VRF can have multiple RT's to import/export routes.

10.3.7.5. Configuration

– Creating each VRF, RD, and RT, plus associating the customer-facing PE interfaces with the correct VRF.
– Configuring the routing protocol (IGP, BGP or static routes) between PE and CE.
– Configuring mutual redistribution between the PE-CE routing protocol and BGP. (**This step is not necessary if the PE-CE protocol is eBGP.**)
– Configuring MP-BGP between PEs.

10.3.7.6. BGP Route Reflector

Between PE routers: customer routes exchanged via BGP (using Route Reflectors)

10.3.7.7. PE-CE routing protocols

PE-CE connectivity allows for flexible route exchange between the customer and the provider through the use of static routes, eBGP, or IGPs

10.3.8. L2 VPN

– It is a layer 2 service (also called Carrier Ethernet)
– The service providers typically have no participation in the enterprise Layer 3 WAN routing.
– The Service provider forwards the traffic of any given customer based on its Layer 2 information (Ethernet MAC addresses)
– Virtual Private Wire Service (VPWS): Also known as E-Line in Metro Ethernet forum terminologies, which provide a point-to-point connectivity model = Pseudowire (PW)
– Virtual Private LAN Service (VPLS): Also known as E-LAN in Metro Ethernet forum terminologies, which provide a multipoint or any-to-any connectivity model.
– Those techniques are gradually being replaced by Ethernet VPN (EVPN) that can provide Layer 2 and Layer 3 services

10.3.8.1. Virtual Private Wire Service (VPWS) = Pseudowire (PW)

– Because it's a Layer 2 service, the network is completely transparent to the CE routers
– No MAC learning at the PE level.

10.3.8.2. Virtual Private LAN Service (VPLS)

– Virtual Private LAN Service emulates a LAN segment across the MPLS backbone.
– It provides multipoint Layer 2 connectivity between remote sites.
– MAC learning at the PE level. Frames are forwarded based on the destination MAC address

10.4. INTERNET VPN VS. MPLS VPN

Internet VPN	*MPLS VPN*
– Public Internet-based – L3 only – Low cost – Throughput and latency concerns (best effort) – Infrastructure is managed by many ISPs. No single responsibility.	– Provided by single SP: controls all access and backbone facilities – L2 or L3 – With QoS control – Infrastructure owned by one organization – Enables SLAs to be delivered/enforced – Privacy without encryption

11. OVERLAY TECHNOLOGIES

With overlay networking, we create an ***overlay network***, which is a virtual network on top of the ***underlay network***, which is the physical network. We use **encapsulation** to wrap the original data packet or frame in a new packet or frame with an overlay header. The job of the underlay network is to get packets or frames from A to B. It is a simple network whose main job is transportation.

11.1. GENERIC ROUTING ENCAPSULATION (GRE)

RFC 2784

– Encapsulate data packets
– Unencrypted
– Set up direct point-to-point connection
– Simplifying connections between separate networks
– Enables usage of not supported protocols

11.2. SEGMENT ROUTING (SR)

RFC 8402

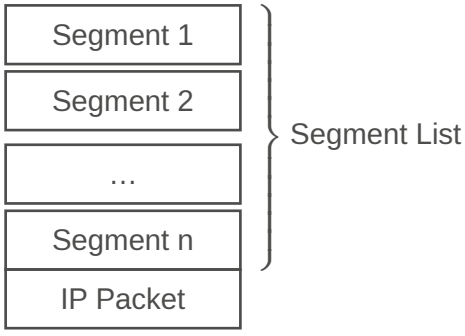
– Source routing paradigm
– Packet steering according list of instructions added in packet at source node
– Nodes execute instructions found in packet header
  – Intermediate nodes don’t maintain per-flow state information
  – State is in the packet
  – Source node controls traffic steering

11.2.1. Segments

Instruction: Segment

SID: Segment Identifier

Ordered list of instructions: Segment List



Represents any kind of instruction
– Topological:
  – Forwarding traffic on shortest ECMP1 path to destination
  – Forwarding traffic through a specific interface
– Service-based
  – Deliver packet to specific service and process it there

11.2.2. Data Plane

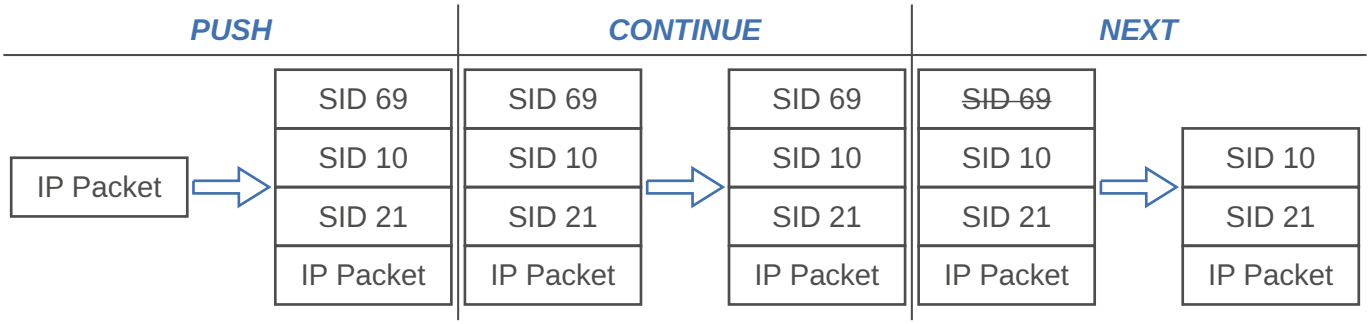
– Segment Routing with the MPLS Data Plane: RFC 8660
– Segment Routing over IPv6: RFC 9602

11.2.3. Operations

PUSH: insert segments to packet header and set active segment

CONTINUE: the active segment is not finished and remains active

NEXT: the active segment is completed - activate next segment in SID List



11.2.4. Control Plane

– No new control plane implementation
– Segments/Instructions exchange via ISIS, OSPF, BGP
  – ISIS Extension RFC 8667
  – OSPF Extensions RFC 8665 , RFC 8666
  – BGP Extensions RFC 9086
– Simplification: Removes unnecessary protocols e.g. LDP, RSVP-TE

11.2.5. Segment Significance

11.2.5.1. Global Segments

– All SR-enables nodes in SR Domain support these instructions
– Each node installs these segments in forwarding table

Example: forward packet according to shortest path to Node1

11.2.5.2. Local Segments

– Only originating node supports these instructions
– Therefore, only originating node writes them into forwarding table
– Other nodes have to know about their existence and meaning

Example: forward packet on interface to Node2

11.2.6. Segment Routing Control Plane Types

IGP Segments	*BGP Segments*
– IGP Prefix segment – IGP Node segment – IGP Anycast segment – IGP Adjacency segment – Layer-2 Adjacency-SID – Group Adjacency-SID	– BGP Prefix segment – BGP Anycast segment – BGP peer segment

11.2.6.1. IGP Prefix Segment

“steer traffic along the ECMP-aware shortest path to the prefix associated with this segment.”

– Global Segment
– Associated with an IGP prefix
– Usually, a multi-hop path
– Each node installs the forwarding entry for each Prefix-SID

11.2.6.2. IGP Node Segment

– Sub-type of IGP Prefix segment
– Advertised with N-Flag
– Associated with a prefix that identifies a specific node.
– Normally set on the loopback0
  – There is no difference in forwarding behavior between a Prefix-SID and a Node-SID.

11.2.6.3. IGP Anycast Segment

“steer traffic along the ECMP-aware shortest path towards the closest node of the anycast set.”

– Sub-type of prefix segment
– Associated with anycast prefix
– Group of nodes
  – Load Balancing (ECMP)
  – High-Availability
  – “steer traffic via given region (e.g. Central Europe), given router function (e.g. any spine in this data center).”

11.2.6.4. IGP Adjacency segment

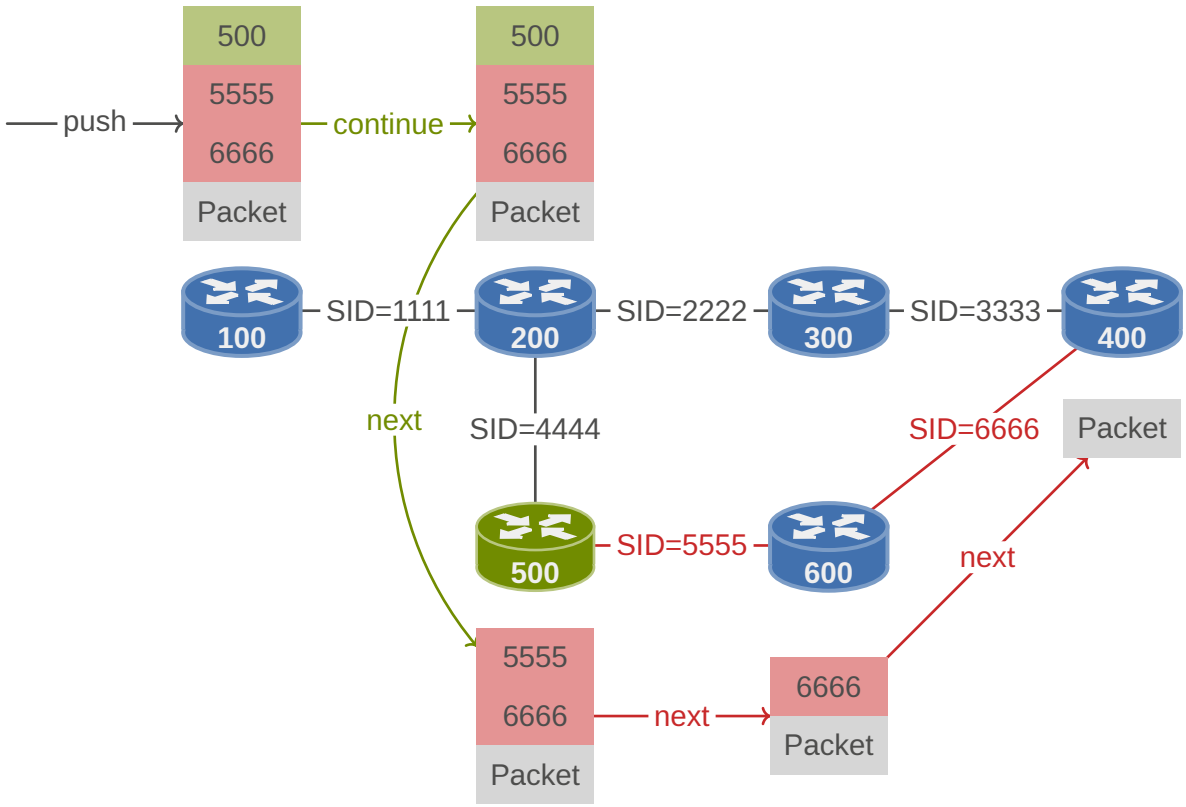
“steer traffic out of the link of the adjacency associated with this segment”

– Associated with unidirectional adjacency
– Local Segment
– The traffic steered on the link regardless of the shortest path routing.

11.2.6.5. Combining Segments

Different types of Segments can be combined in such end-to-end path, such as different types of IGP segments and BGP segments.

Adjacency vs **Node** segment:



11.2.7. SR-MPLS

Segment Routing with the MPLS Data Plane

- SR-MPLS re-uses MPLS data plane without any change
- Segments are represented by MPLS labels
  - Segment List is stack of MPLS label
  - Segment to process (Active Segment) on top
- Segments are distributed over IGP (or BGP)
  - No need for LDP anymore
  - Interoperability with LDP possible (mapping server)
- IPv4 & IPv6 address families

11.2.7.1. Global vs. Local

Global	*Local*
– Defined in the SR Global Block (SRGB) – Recommended: use identical SRGBs	– Defined in SR Local Block (SRLB) – Local property of an SR node

11.2.7.2. Operations

Segment list operation	*MPLS label stack operation*
PUSH	PUSH
CONTINUE	SWAP
NEXT	POP

12. VXLAN - EVPN

Goal: Establish a VPN connection that supports both layer 2 and layer 3 traffic.

12.1. VIRTUAL EXTENSIBLE LOCAL AREA NETWORK (VXLAN)

A tunneling protocol that tunnels Ethernet (layer 2) traffic over an IP (layer 3) network.

12.1.1. Issues of traditional L2 Networks

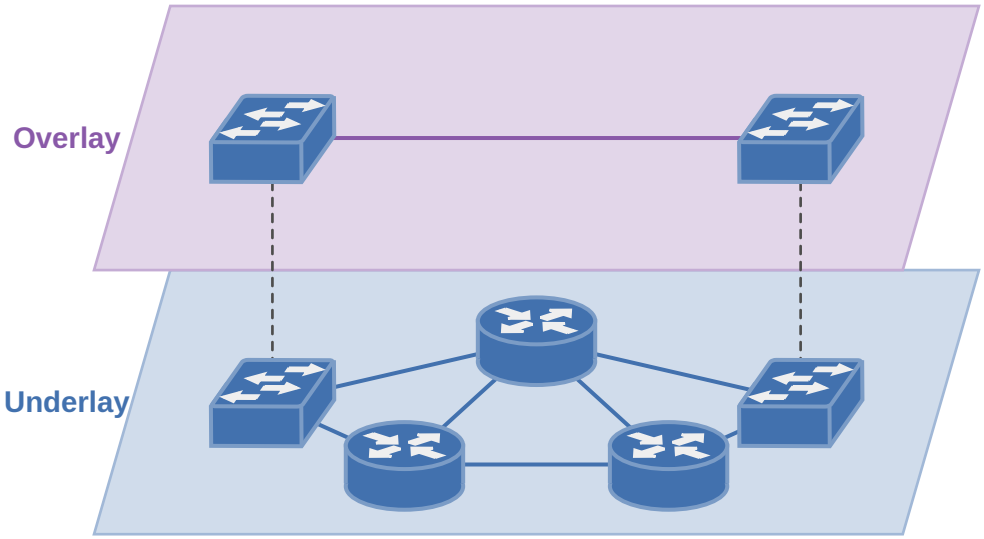
Spanning tree: It blocks any redundant links to avoid loops. Blocking links means we pay for links we can't use.

Limited amount of VLANs: The VLAN ID is 12-bit, which means we can create 4094 VLANs (0 and 4095 are reserved).

Large MAC address tables: Because of server virtualization, the number of addresses in the MAC address tables of our switches has grown exponentially. Before server virtualization, a switch only had to learn one MAC address per switchport. With server virtualization, we run many virtual machines (VM) or containers on a single physical server. Each VM has a virtual NIC and a virtual MAC address.

12.1.2. Overlay vs Underlay

VXLAN uses an overlay and underlay network:



An overlay network is a **virtual network** that runs on top of a physical underlay network. With VXLAN, the overlay is a layer 2 Ethernet network. The underlay network is a layer 3 IP network. Another name for the underlay network is a transport network. The overlay and underlay network are independent.

Underlay network: is simple; its only job is to get packets from A to B. When we use layer 3, we can use an IGP like OSPF or EIGRP and load balance traffic on redundant links.

Overlay network: is virtual and requires an underlay network, but whatever changes you make in the overlay network won't affect the underlay network. You can add and remove links in the underlay network, and as long as your routing protocol can reach the destination, your overlay network will remain unchanged.

12.1.3. VXLAN Network Identifier (VNI)

The VNI identifies the VXLAN and has a similar function as the VLAN ID for regular VLANs. We use 24 bits for the VNI, which means we can create 2^{24} (about 16 million) VXLANs.

12.1.4. VXLAN Tunnel Endpoint (VTEP)

The VTEP is the device that's responsible for encapsulating and deencapsulating layer 2 traffic. This device is the connection between the overlay and the underlay network. The VTEP comes in two forms:

– Software (host-based)
– Hardware (gateway)

12.1.4.1. Software VTEP

A software VTEP is located on a Hypervisor such as VMWare ESXI, **KVM** or **Proxmox**. These virtualization platforms use virtual switches, and some of them support VXLAN.

The VXLAN tunnels are between the virtual switches of the hypervisors. The underlay network is unaware of VXLAN.

12.1.4.2. Hardware VTEP

A hardware VTEP is located on a router, switch, or firewall which supports VXLAN. We also call a hardware VTEP a ***VXLAN gateway*** because it combines a regular VLAN and VXLAN segment into a single layer 2 domain. Some switches have VXLAN support with **ASICs**, offering **better VXLAN performance** than a software VTEP.

The VXLAN tunnels are between the physical switches. The devices that connect to the physical switches are unaware of VXLAN.

12.1.4.3. VTEP Interfaces

Each VTEP has two interfaces types:

VTEP IP interface: Connects the VTEP to the underlay network with a unique IP address. This interface **encapsulates and de-encapsulates** Ethernet frames.

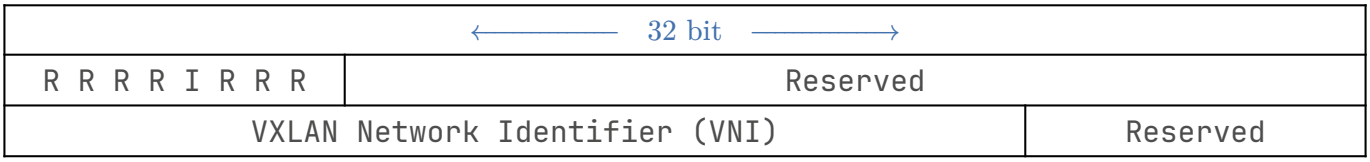
VNI interface: A virtual interface that **keeps network traffic separated** on the physical interface. Similar to an SVI interface.

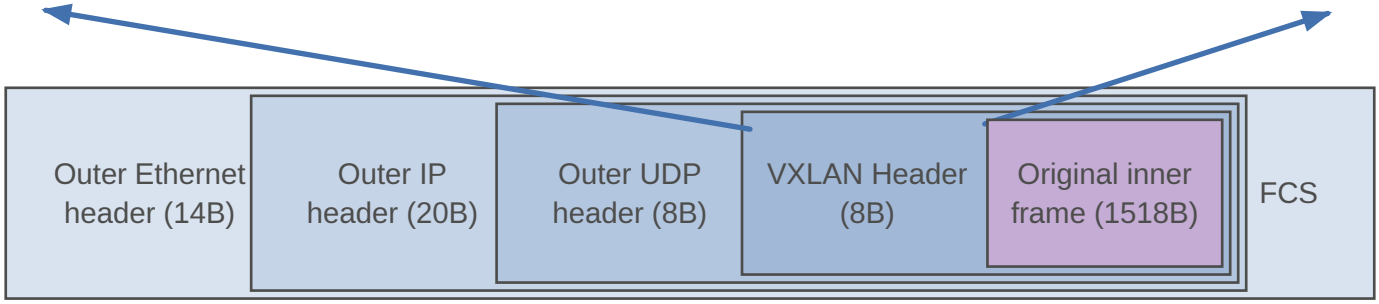
A VTEP can have multiple VNI interfaces, but they associate with the **same VTEP IP interface**.

12.1.5. Frame Format

When a VTEP encapsulates an Ethernet frame, it adds a VXLAN header. In this header, we find the VNI and some flags. RFC 7348

The official UDP port number for VXLAN is 4789. However, it's possible that you also run into UDP port 8472. When VXLAN was first implemented in Linux, there was no official port number yet, and many vendors used port 8472.





12.1.6. Control plane

Term	*Definition*
BUM Traffic	Broadcast, Unknown Unicast, Multicast

Term	Definition
Ingress replication	Sending a copy of multi-destination traffic to all participating VTEPs (Remote VTEPs statically configured)

With VXLAN, each VTEP has a VXLAN mapping (forwarding) table that maps a destination MAC address to a remote VTEP IP address. How do VTEP devices learn MAC addresses? There are different control plane solutions. Cisco supports these four options:

VXLAN with static unicast VXLAN tunnels: The VXLAN mapping table is manually configured which means that the IP addresses of all the peer VTEP need to be configured manually. It works, but it doesn't scale well. Additionally BUM traffic becomes a significant issue when scaling up the network size.

VXLAN with multicast underlay: The VXLAN mapping table is no more manually configured but instead interested VTEPs can join the Multicast group of a VNI to receive related traffic. Additionally BUM traffic can be handled in an elegant way as no more replication is required.

VXLAN with MP-BGP EVPN: VXLAN enhanced with proactive learning of MAC addresses using the iBGP protocol as control plane.

VXLAN with LISP: Proprietary solution by Cisco

12.1.6.1. Flood and learn / Data plane learning

1) ARP Request (BUM Traffic): Originating host broadcasts an ARP request, which constitutes BUM traffic, into the VNI.
2) Multicast Flooding: The local VTEP receives the ARP request and must deliver it to all VTEPs participating in the same VNI, as well as to all locally attached ports in that VNI. It forwards the frame to the designated multicast group associated with the VNI.
3) VTEP Learning: Upon receiving the multicast frame, each remote VTEP inspects the encapsulated ARP request and caches the source IP-to-MAC mapping for future use. Each VTEP then floods the inner frame to all local ports belonging to the VNI.
4) ARP Reply (Unicast Traffic): The target host recognizes its IP address and sends a unicast ARP reply back toward the originating host. All other recipients silently discard the request.
5) Reply Encapsulation: The responding VTEP encapsulates the ARP reply and forwards it directly to the originating VTEP, whose address was cached during step VTEP Learning.
6) Session Establishment: The originating VTEP decapsulates the reply and delivers it to the host. With the destination MAC now known, both hosts proceed with normal unicast communication. If the cache expires, the flooding process begins again.

12.2. ETHERNET VIRTUAL PRIVATE NETWORK (EVPN)

EVPN is a control plane technology used for building Layer 2 and Layer 3 VPNs over a common IP or MPLS network infrastructure. EVPN leverages existing tunneling protocols such as VXLAN, MPLS, or even GRE for data plane encapsulation, depending on the specific deployment requirements and network architecture.

12.2.1. Layer 2

EVPN Layer 2 combines the flexibility of Ethernet with the efficiency of routing protocols, making it a preferred solution for modern network architectures.

12.2.1.1. Key Features

Layer 2 Bridging Across Networks: EVPN allows devices in the same subnet to communicate as if they were on the same physical Layer 2 network, even when separated by a Layer 3 network.

Control Plane Efficiency: Unlike traditional Layer 2 networks that rely on flooding for MAC learning, EVPN uses BGP to distribute MAC address information in the control plane, improving scalability and reducing broadcast traffic.

Overlay and Encapsulation: EVPN often works with VXLAN as the data plane, encapsulating Layer 2 Ethernet frames within Layer 3 UDP packets. This enables the creation of virtual Layer 2 networks (or subnets) over a physical Layer 3 infrastructure.

Multi-Tenancy Support: EVPN supports multi-tenancy by segmenting traffic using VXLAN Network Identifiers (VNIs), similar to VLANs but with greater scalability.

Redundancy and Load Balancing: It supports multipath forwarding and redundancy through all-active multihoming, ensuring traffic continuity even during link or device failures.

12.2.1.2. Applications

Data Center Interconnections: Widely used in multi-tenant data centers to connect different locations or segments while maintaining Layer 2 connectivity.

WAN Connectivity: Facilitates seamless communication between geographically distributed sites by extending Layer 2 domains over WANs.

Scalable Network Fabrics: Ideal for creating scalable network overlays in large enterprises or cloud environments.

12.2.1.3. BGP Control plane

In EVPN, the customer MAC addresses are learned in the data plane over links connecting customer edge (leaf) to the provider edge (spine) devices. The MAC addresses are then distributed over the MPLS or IP core network using BGP with an MPLS label or route distinguisher identifying the service instance.

12.2.1.3.1. EVPN NLRI

EVPN defines a BGP NLRI that advertises different route types and route attributes. The EVPN NLRI is carried in BGP using BGP multiprotocol extensions. Each BGP UPDATE message carries two labels that identify exactly what kind of information it contains:

Address Family Identifier (AFI): Identifies the broad category of information being carried. For example, IPv4, IPv6, or — as in the case of EVPN — Layer-2 VPN, since EVPN operates at the Ethernet level rather than the IP level.

Subsequent Address Family Identifier (SAFI): Narrows down the specific type within that category. Since multiple technologies exist within the Layer-2 VPN family (such as VPLS and EVPN), the SAFI distinguishes between them.

Together, the AFI and SAFI form a pair that every BGP router inspects upon receiving an UPDATE message to decide whether it can process the contents. If a router does not support the advertised AFI/SAFI combination, it silently drops the message and does not propagate it to its neighbors.

12.2.1.3.2. BGP EVPN Autodiscovery Support on Route Reflector

EVPN Autodiscovery supports BGP route reflectors. A BGP route reflector can be used to reflect BGP EVPN prefixes without EVPN being explicitly configured on the route reflector. The route reflector does not participate in autodiscovery; that is, no pseudowires are set up between the route reflector and the PE devices. A route reflector reflects EVPN prefixes to other PE devices so that these PE devices do not need to have a full mesh of BGP sessions. The network administrator configures only the BGP EVPN address family on a route reflector.

BGP uses the Layer 2 VPN (L2VPN) RIB to store endpoint provisioning information, which is updated each time any Layer 2 VFI is configured. The prefix and path information is stored in the L2VPN database,

which allows BGP to make decisions about the best path. When BGP distributes the endpoint provisioning information in an update message to all its BGP neighbors, this endpoint information is used to configure a pseudowire mesh to support L2VPN-based services.

12.2.2. Layer 3

While EVPN L2 allows for the extension of Ethernet services across different locations, it often struggles in real-world situations where routing is necessary. By adding L3 features, EVPN improves data routing efficiency and enables smooth connections across various networks.

12.2.2.1. EVPN Route Types

1 - Ethernet Auto-Discovery Route: Used for networkwide messages related to Ethernet a todiscovery, essential for multihomed CE devices. For more information refer to RFC 7432 .

2 - MAC/IP advertisement Route: Used to advertise MAC and IP addresses within EVPN for control plane learning of ESI MAC addresses, facilitating communication within a VLAN. For more information refer to RFC 7432 .

3 - Inclusive Multicast Route: Establishes paths for BUM traffic within VLANs across PE devices, ensuring efficient multicast communication. For more information refer to RFC 7432 .

4 - Ethernet Segment Route: Facilitates multihoming of CE devices by allowing PE devices on the same Ethernet segment to discover each other through the Ethernet segment identifier (ESI). For more information refer to RFC 7432 .

5 - IP Prefix Route: Advertises IP prefixes for inter-subnet connectivity across data centers. For more information refer to RFC 9136 .

6 - Selective Multicast Ethernet Tag Route: Distributes Host's or VM's intent to receive Multicast traffic for a certain Multicast Group `(,G)` or Source-Group combination `(S,G)`. For more information refer to RFC 9251 .

7 - IGMP Join Synch Route: Synchronizes IGMP Join messages in an EVPN multihome environment. For more information refer to RFC 9251 .

8 - IGMP Leave Synch Route: Synchronizes IGMP Leave messages in an EVPN multihome environment. For more information refer to RFC 9251 .

12.2.2.2. Virtual Routing and Forwarding (VRF)

VRF is a technology used to create multiple instances of a routing table within a single physical router or switch. Each VRF instance maintains its own routing table, forwarding information, and interfaces, effectively creating separate routing domains within the same device. The main achievements through VRF in EVPN are:

Isolation of Customer Traffic: Each customer or tenant in an EVPN deployment can be assigned a separate VRF instance. This ensures that traffic from one customer's network is isolated from traffic belonging to other customers, even though they may be using the same physical network infrastructure.

Independent Routing: Each VRF maintains its own routing table, allowing customers to imple- ment their own routing policies and configurations without impacting other customers. This provides flexibility and autonomy for each customer to manage their network according to their specific requirements.

Security and Privacy: VRF provides a level of security and privacy by segregating traffic between different customers. This prevents unauthorized access or interception of traffic between different tenants within the EVPN network.

Scalability: VRF allows service providers to scale their EVPN deployments by accommodating multiple customers or tenants on the same physical network infrastructure. Each VRF in- stance operates independently, enabling the network to support a large number of customers while maintaining efficient routing and forwarding.

12.2.2.3. Integrated Routing and Bridging (IRB)

Without IRB, inter-subnet/inter-vlan traffic must traverse external routers or centralized gateways for routing, leading to “traffic tromboning”. Bridging (L2) and routing (L3) are then also handled separately, requiring more complex configurations and potentially creating bottlenecks at centralized routing points.

RFC 9135

There are two modes of operation for IRB:

Symmertric IRB: Uses EVPN as a Layer-2 and Layer-3 VPN overlay, with distributed inter-subnet traffic routed at any VTEP, ingress and egress. In Ethernet frames sent through the underlay network the L3 VNI is set in the VXLAN header which allows the identification of the tenant VRF used for routing decision on the destination VTEP.

Asymmetric IRB: Uses EVPN purely as a Layer-2 VPN overlay, with intersubnet traffic routed only at the ingress VTEP. As a result, ingress VTEP performs both routing and bridging, while the egress VTEP performs only bridging. In Ethernet frames sent through the underlay network the L2 VNI of the target network is set in the VXLAN header which allows direct bridging on the destination VTEP.

12.2.2.3.1. Symmetric IRB

With the symmetric IRB routing model, the VTEPs do routing and bridging on both the ingress and egress sides of the VXLAN tunnel. As a result, VTEPs can do inter-subnet routing for the same tenant virtual routing and forwarding (VRF) instance with the same VNI in both directions. The VTEPs use a dedicated Layer 3 traffic VNI in both directions for each tenant VRF instance.

You configure an extra VLAN with an IRB interface, mapped to a VNI, for each tenant L3 VRF instance. That VNI is the Layer 3 transit VNI between VTEPs for the tenant VXLAN traffic.

This model requires that the network has established Layer 3 connectivity between all source and destination VTEPs for EVPN type 2 routing. You configure EVPN Type 5 routing in the tenant VRF instance to provide the Layer 3 connectivity.

12.2.2.3.2. Asymmetric IRB

With the asymmetric model, leaf devices serving as VTEPs both route and bridge to initiate the VXLAN tunnel (tunnel ingress). However, when exiting the VXLAN tunnel (tunnel egress), the VTEPs can only bridge the traffic to the destination VLAN.

With this model, VXLAN traffic must use the destination VNI in each direction. The source VTEP always routes the traffic to the destination VLAN and sends it using the destination VNI. When the traffic arrives at the destination VTEP, that device forwards the traffic on the destination VLAN.

This model requires you to configure all source and destination VLANs and their corresponding VNIs on each leaf device, even if a leaf doesn’t host traffic for some of those VLANs. As a result, this model can have scaling issues when the network has a large number of VLANs. However, when you have fewer VLANs, this model can have lower latency over the symmetric model. Configuration is also simpler than with the symmetric model.

12.2.2.4. Distributed Anycast Gateway (DAG)

DAG is a default gateway addressing mechanism in a BGP EVPN fabric. The feature enables the use of the same gateway IP and MAC address across all the devices in an EVPN over MPLS network with IRB. This ensures that every device functions as the default gateway for the workloads directly connected to it.

The feature facilitates flexible workload placement, host mobility, and optimal traffic forwarding across the BGP EVPN fabric.

13. CONTENT DELIVERY NETWORKS (CDN)

A Content Delivery Network (CDN) is a geographically distributed system of servers designed to deliver content efficiently to users. Rather than fetching content from a centralized origin server, a CDN delivers content from strategically placed edge servers that are closer to the end user.

13.1. CACHING

Caching is a technique used to temporarily store copies of data in locations that allow for faster access. CDNs provide caching on the Networking layer.

13.2. COMPONENTS

13.2.1. Origin Server

The origin server holds the original version of the content that is to be distributed. It is usually located in a central data center and serves as the single source of truth for all CDN edge nodes.

13.2.2. Edge, Cache and CDN Servers

These are geographically distributed servers, often called edge nodes or points-of-presence (PoPs), that store and serve copies of content closer to the end users. Each edge node typically caches content on demand and uses algorithms to manage what is stored and when it should be refreshed or evicted.

13.2.3. DNS Infrastructure

The Domain Name System plays a key role in how requests are directed within a CDN, but its function depends on the specific request routing technique in use. In many setups, the DNS helps guide clients to an appropriate edge server by returning different IP addresses based on internal logic, such as geography, current load, or availability.

13.3. KEY BENEFITS

Latency Reduction: Serving content from nearby servers reduces round-trip time.

Availability: If one node fails, another can take over (failover and redundancy).

Scalability: CDNs handle spikes in traffic by load-balancing requests.

Cost Optimization: Offloading traffic from origin servers reduces backend and transit costs.

DDoS Protection: Distributed presence helps absorb attacks at the edge.

Load Reduction on Global Internet Infrastructure: By delivering content from servers closer to the user, CDNs reduce the amount of redundant data that must traverse long-distance networks.

13.4. REQUEST ROUTING TECHNIQUES

When a user requests content from a service that is distributed across a CDN, it must decide which of its many edge servers should handle the request. This process is known as request routing.

13.4.1. DNS-Based Geo-Routing

One of the most widely used techniques to direct users to the nearest or most optimal CDN node is DNS-based request routing, often referred to as DNS-based Geo-Routing. When using geo-routing, all the CDN

edge servers have their own unique IP address. When a client uses the Domain Name System to look up the IP address to a specific URL, the responding authoritative DNS server for the requested Domain can respond with an IP address of one of the edge servers it thinks is most suited for this client. The main decision factor is the geographical estimation based on the source IP address of the resolver. To achieve this, DNS servers rely on commercial or public GeoIP databases such as MaxMind or IP2Location, which map IP prefixes to geographical regions. In addition to geography, other factors such as current server load, health status, network latency, or even complex business rules can also influence the decision process. It's important to note that the DNS server's decision is based on the resolver's location, not necessarily the end user's.

13.4.1.1. EDNS(0) and Client Subnet Extension

To improve location accuracy in DNS-based geo-routing, many resolvers support the EDNS(0) Client Subnet (ECS) extension, specified in RFC 7871 . With ECS, the resolver includes part of the original client's IP address in the DNS query sent to the authoritative server. This allows the DNS server to make decisions based on the actual user location rather than just the resolver's location. Only a truncated subnet (typically /24 for IPv4) is sent, preserving a balance between accuracy and privacy. When supported by both a resolver and an authoritative server, ECS leads to better routing outcomes and lower latency for the end user.

13.4.2. Anycast and BGP

The main difference to geo-routing is that all edge servers are assigned the same IP address. Thanks to BGP, each of these locations (typically CDN edge servers) advertises that shared IP to its neighboring ASes. This way the "normal" decision-making algorithms (based on attributes like AS-path length, local preference, MED, etc.) can be used to automatically route traffic to what is considered the nearest CDN server. So whenever an end user sends a packet to this shared IP address, the network layer (guided by BGP path selection rules) determines which of the advertising nodes is "closest."

Pro	*Contra*
– Fault tolerance	– Suboptimal routing decisions – CDN operators have limited control over traffic distribution – Instability in routing tables, commonly known as route flapping, can cause traffic to oscillate between nodes

13.5. CACHING, CACHE-CONTROL AND HTTP HEADERS

In the context of the World Wide Web, caching behavior is largely governed by HTTP header fields. These headers instruct browsers and intermediary caches on how to store, validate, and reuse responses.

Term	*Definition*
Cache-Control	HTTP Header that holds directives (instructions) in both requests and responses that control caching in browsers and shared caches (e.g., Proxies, CDNs).
max-age	***Response:*** Indicates that the response remains fresh until N seconds after the response is generated ***Request:*** Indicates that the client allows a stored response that is generated on the origin server within N seconds
Age	HTTP response header that indicates the time in seconds for which an object was in a proxy cache.
Expires	HTTP response header that contains the date/time after which the response is considered expired in the context of HTTP caching.

Term	Definition
ETag	HTTP response header that acts as an identifier for a specific version of a resource. It lets caches be more efficient and save bandwidth, as a web server does not need to resend a full response if the content has not changed.
Last-Modified	HTTP response header contains a date and time when the origin server believes the resource was last modified. It is less accurate than an ETag for determining file contents, but can be used as a fallback mechanism if ETags are unavailable.

14. QUALITY OF SERVICE (QOS)

QoS is the set of mechanisms that control and prioritize network traffic to ensure ***predictable performance*** and ***appropriate resource allocation*** for different applications.

– Anything that can affect end user's perception on network service quality falls into the scope of QoS
– If we say that a network has QoS or provides QoS, it means that the network is able to meet the need of the end users' applications in a satisfactory way, with a good ***Quality of Experience (QoE)***

14.1. REQUIREMENTS

RFC 2386
– Multiple paths rather than just the IGP shortest path
– Knowledge of network resource availability
– Technique to mark that a traffic should get a specific treatment
– Route Pinning: Mechanism to keep a flow path fixed for a duration
  – We don't want to switch paths as soon as a "better" path is found
  – Otherwise routing oscillations can happen → shift back and forth between alternate paths

14.2. TYPES

Term	Definition
Routing-based QoS (Path Selection)	– QoS-aware routing – Choose best path/link – "Which path do packets take?"
Node-based QoS (Packet Treatment)	– Queuing & scheduling – Congestion management – Policing & shaping – "What happens to packets at each hop?"

14.3. NETWORK PERFORMANCE

14.3.1. Four main metrics

– Latency (ms)
– Jitter (ms)
– Packet Loss (%)
– Bandwidth (Gbit/s)

14.3.2. Latency / Delay

End-to-end delay: complete time for packets going from source to destination

One-way network delay: the time the first bit of the packet is put on the wire at the source reference point to the time the last bit of the packet is received at the receiver reference point RFC 2679 .

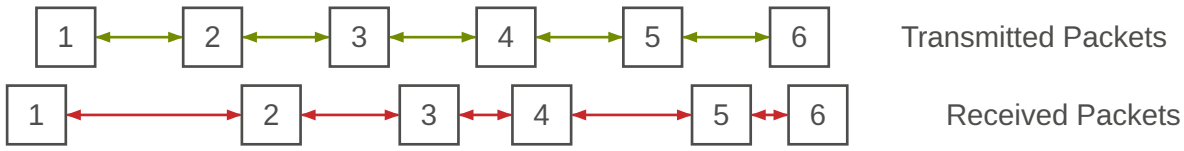
Transmission delay: the time it takes to transmit a packet into the wire. Transmission delay is insignificant for high-speed links.

Packet processing delay: the time it takes to process a packet at a network device, for example, queueing, table lookup, etc.

Propagation delay: the time it takes for the signal to travel over the distance.

14.3.3. Jitter (Delay Variation)

– Difference between the highest delay measured between A and B and the lowest delay measured between A and B.
– Due to the variable delays, the packets are received with nonuniform gaps between consecutive packets.
– Jitter does not impact the TCP-based applications as much from a user experience perspective and is relevant only for time-sensitive applications with a continuous data stream.



14.3.4. Packet Loss

If the number of packets to be queued continues to increase, the memory within the device fills up and packets are dropped.

Packet Loss Ratio (PLR): percentage of packet lost while traveling from the source to the destination.

14.3.5. Bandwidth / Throughput

– Maximum transferrable data rate
– Measured in bits per second (Mbps, Gbps, Tbps).
– Shared resources between users/applications
– Fundamental Limit

14.3.6. Types of traffic

– Different requirements regarding
  – Delay/latency
  – Jitter
  – Packet drop
  – Bandwidth
– Different Patterns, e.g.
  – Bursty
  – greedy

Voice	*Video*	*Data*
– Smooth	– Bursty	– Smooth/Bursty
– Benign	– Greedy	– Benign/Greedy
– Drop sensitive	– Drop sensitive	– Drop insensitive
– Delay sensitive	– Delay sensitive	– Delay insensitive
– UDP priority	– UDP priority	– TCP retransmits

14.3.7. Where queuing takes place

– Incoming buffer
  – Usually those buffers are only transit buffers
  – Cannot be filled up in case of a CPU overload: leads to drops
– Outgoing buffer
  – Filled up with packets waiting for transmission
  – Queue management is focused on the outgoing buffer

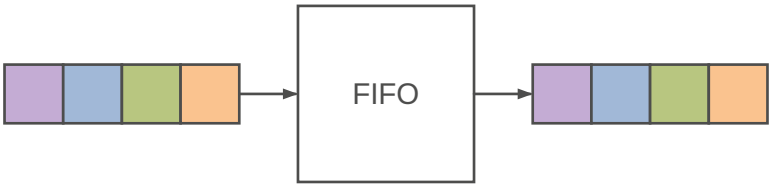
14.4. QUEUING ALGORITHMS

Incoming buffer: Usually those buffers are only transit buffers. Cannot be filled up in case of a CPU overload: leads to drops

Outgoing buffer: Filled up with packets waiting for transmission. Queue management is focused on the outgoing buffer

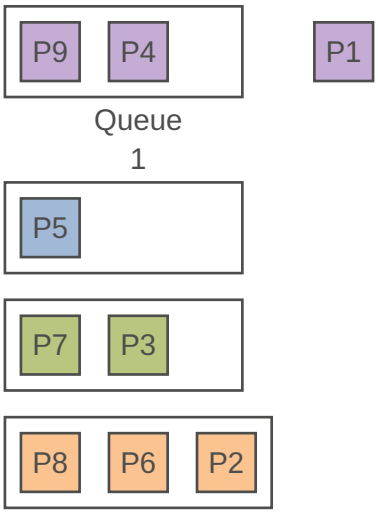
14.4.1. First In First Out (FIFO)

– No concept of priority or classes of traffic and consequently, makes no decision about packet priority.
– Fastest method of queuing, is effective for large links that have little delay and minimal congestion.



14.4.2. Priority queuing

– Uses multiple queues
– Allows prioritization
– Always empties first queue before going the the next queue.
  – Empty Queue 1
  – If queue 1 is empty, then dispatch one packet from Queue 2
  – If both Queue 1 and Queue 2 are empty, then dispatch one packet from Queue 3.
– Queue 2,3 and 4 may “starve”



14.4.3. Round-robin

– Uses multiple queues
– No prioritization
– Dispatches one packet from each queue in each round:
  – One packet from Queue 1
  – One packet from Queue 2
  – One packet from Queue 3
  – One packet from Queue 4
  – Then repeat
– Important traffic gets no prioritization

14.4.4. Weighted Round-robin

– Assign a weight to each queue.
– Dispatches packets from each queue proportionally to the assigned weight:
  – Dispatch up to four from Queue 1
  – Dispatch up to two from Queue 2
  – Dispatch one from Queue 3
  – Dispatch one from Queue 4
  – Go back to Queue 1
– Also called Weighted Fair Queuing (WFQ)

14.4.5. Class-Based Weighed Fair Queuing (CBWFQ)

– Extends the standard WFQ functionality to provide support for user-defined traffic classes.
– A queue for a class can be customized with:
  – The maximum bandwidth
  – A queue limit (maximum number of packets allowed to accumulate in the queue for the class)
  – The maximum percent of the bandwidth
– The bandwidth assigned to the packets of a class determines the order in which packets are sent.

14.4.6. Low Latency Queuing (LLQ)

– LLQ provides strict priority queuing for CBWFQ.
– LLQ allows delay-sensitive data such as voice to be sent first and have strict priority

14.5. QUEUE MANAGEMENT

14.5.1. Tail dropping

– Tail drop occurs when a packet needs to be added to a queue, but the queue is full.
– There is no differentiated drop mechanism and therefore premium traffic is dropped in the same way as best-effort traffic.

14.5.2. TCP

14.5.2.1. Congestion Control

Sender increase sending rate until packet loss (congestion) occurs, then decrease sending rate on loss event

14.5.2.2. Global synchronization

When congestion occurs, dropping affects most of the TCP sessions, which simultaneously back off and then restart again. This causes inefficient link utilization at the congestion point

1) Multiple TCP sessions start at different times
2) TCP window sizes are increased
3) Tail drops cause many packets of many sessions to be dropped at the same time
4) TCP sessions restart at the same time (synchronized)

14.5.2.3. Starvation

All buffers are temporarily seized by aggressive flows and normal TCP flows experience buffer starvation.
– Happens when there is a combination of TCP and UDP flows during a period of congestion
– After tail drops begin
  – TCP flows slow down simultaneously
  – UDP flows do not slow down
– UDP traffic starts filling up the queues and leaves little or no room for TCP packets.

14.5.3. Random Early Detection (RED)

– Address network congestion in a responsive rather than reactive manner.
– Goal: provide congestion avoidance by controlling the average queue size.
– Randomly Drop Packets in the queue
– Dropped TCP segments cause TCP sessions to reduce their window sizes.
– Avoidance of global synchronization
– Introduce fairness to reduce bias against bursty traffic
– Works well with TCP flows

14.5.4. Weighted Random Early Detection (WRED)

– Drops packets selectively based on the IP precedence, DSCP or EXP
– lower priority class will start random drops before higher priority class.
– When the queue is below the minimum threshold, there are no drops.
– As the queue fills up to the maximum threshold, a small percentage of packets are dropped. (mark probability denominator determines the percentage)
– When the maximum threshold is passed, all packets are dropped.

14.5.5. Policing

The process of discarding packets (by a dropper) within a traffic stream in accordance with the state of a corresponding meter enforcing a traffic profile. RFC 2475

Policing is applied to inbound traffic on an interface.

14.5.6. Shaping

The process of delaying packets within a traffic stream to cause it to conform to some defined traffic profile. RFC 2475

Shaping is applied on outgoing traffic.

14.6. QOS MODELS

Term	*Definition*
Best-Effort	This is the default queuing model for interfaces. All packets are treated in the same way. There is no QoS.
Integrated Services (IntServ)	IntServ provides a way to deliver the end-to-end QoS that real-time applications require by explicitly managing network resources to provide QoS to specific user packet streams, sometimes called microflows.
Differentiated Services (DiffServ)	DiffServ uses a soft QoS approach that depends on network devices that are set up to service multiple classes of traffic each with varying QoS requirements. Although there is no QoS guarantee, the DiffServ model is more cost- effective and scalable than IntServ.

14.6.1. Best effort

– The internet is a best effort network.
– According to the Net Neutrality principle, the ISPs are not allowed to differentiate traffic on the internet.

Pro	*Contra*
– Scalable, only limited by bandwidth limits – No special QoS mechanisms required – Easiest and quickest model to deploy	– No guarantees of delivery – No preferential treatment

14.6.2. Integrated Services (IntServ)

End-to-End QoS: Guaranteed QoS for specific application streams.

Connection-Oriented: Inspired by telephony.

Application-Driven: App requests service before sending data.

Traffic Profile: App informs the network of traffic characteristics.

RSVP (Resource Reservation Protocol): Signals the need for QoS along the path.

Admission Control: Blocks data if resources cannot be reserved.

14.6.3. Differential Services (DiffServ)

– Simple and scalable mechanism for classifying and managing network traffic and providing QoS guarantees on modern IP networks.
– DiffServ can provide an “almost guaranteed” QoS while still being cost-effective and scalable.
– DiffServ uses a “soft QoS” approach.
– DiffServ divides network traffic into classes based on business requirements.
– Each of the classes can then be assigned a different level of service.
– On network interfaces, a policy is configured to enforce the queuing mechanism with multiple classes of traffic

14.6.3.1. DSCP (marking)

– A 6-bit value in the IP header
– Used to classify packets

Example: “This packet is high priority”

14.6.3.2. PHB (behavior)

– Defines what routers actually do
– Examples:
  – Put packet in priority queue
  – Drop earlier/later
  – Forward faster/slower

DSCP determines PHB

These are standardized behaviors:

CS (Class Selector): Compatibility with old IP Precedence. Simple priority levels

AF (Assured Forwarding): Multiple classes + drop priorities. Used for controlled, differentiated service

EF (Expedited Forwarding): Very high priority. Low latency, low jitter (e.g., VoIP)”

14.7. QOS CLASSIFICATION

Class 1 (Real Time): Voice, Video

Class 2 (Mission Critical): Database, ERP

Class 3 (Best Effort): Web, P2P

14.7.1. Different Elements

14.7.1.1. DSCP (Differentiated Services Code Point)

What is it?: The 6-bit marking/label in the IP header.

Purpose: To classify packets.

Values/Types: 64 possible values (0-63).

Behavior: Interpreted by routers to determine PHB.

Analogy: The sticker on the package.

14.7.1.2. PHB (Per-Hop Behavior)

What is it?: The forwarding treatment/action a router applies.

Purpose: To define how classified packets are handled at each hop.

Values/Types: Various behaviors defined (e.g., CS, AF, EF, Best Effort).

Behavior: The actual forwarding treatment (queuing, dropping, etc.).

Analogy: The handling instructions for the sticker.

14.7.1.3. CS (Class Selector) PHB

What is it?: A specific PHB for backward compatibility with IP Precedence.

Purpose: To provide backward compatibility with IP Precedence and offer tiered service.

Values/Types: Uses DSCP values where the first 3 bits match IP Precedence and the last 3 are 0 (e.g., CS0-CS7, mapping to IPP 0-7).

Behavior: Treatment generally follows the legacy IP Precedence priority levels.

Analogy: “Legacy Priority Mail” handling, respecting older priority levels.

14.7.1.4. AF (Assured Forwarding) PHB Group

What is it?: A group of PHBs offering different service levels.

Purpose: For traffic needing assured delivery with different priority levels and drop probabilities.

Values/Types: 4 classes (AF1x to AF4x), 3 drop precedences per class (AFn1 to AFn3).

Behavior: Provides different levels of forwarding assurance and manages congestion using drop precedence within classes.

Analogy: Different tiers of “Priority Mail” handling with varying drop likelihoods.

14.7.1.5. EF (Expedited Forwarding) PHB

What is it?: A specific, high-priority PHB.

Purpose: For low-loss, low-latency, low-jitter traffic (e.g., VoIP).

Values/Types: Typically signaled by DSCP 46 (101110).

Behavior: Receives highest priority, minimal queuing delay, low loss, jitter, and latency.

Analogy: “Super Express Priority” handling.

14.7.2. Marking

Layer 2: 802.1p byts (3 bits) in dot1q header

Layer 3: TOS byte: Differentiated Services Code Point (DSCP) (6 bits) + Precedence (2 bits)

14.7.3. Class models

4-Class Model	*8-Class Model*	*12-Class Model*
Real Time	Voice	Voice
	Interactive Video	Real-Time Interactive
		Multimedia Conferencing
	Streaming Video	Broadcast Video
		Multimedia Streaming
Signaling or Control	Signaling	Signaling
	Network Control	Network Control
		Network Management
Critical Data	Critical Data	Transactional Data
		Bulk Data
Best Effort	Best Effort	Best Effort
	Scavenger	Scavenger

14.7.4. Network-Based Application Recognition (NBAR)

– NBAR is a classification engine that recognizes and classifies a wide variety of protocols and applications, including web-based and other difficult-to-classify applications and protocols that use dynamic TCP/UDP port assignments.
– Classifies traffic based on application signature
– The generic term is Deep Packet Inspection (DPI) engine
– NBAR is the **Cisco proprietary** approach of a DPI engine

BIBLIOGRAPHY

[1] George, “Top-of-Rack image.” Accessed: Apr. 03, 2026. [Online]. Available: https://www.fs.com/blog/top-of-rack-and-end-of-row-whats-the-difference-2934.html

[2] George, “End-of-Row image.” Accessed: Apr. 03, 2026. [Online]. Available: https://www.fs.com/blog/top-of-rack-and-end-of-row-whats-the-difference-2934.html