Nicolas Caluori und Joëlle Schönenberger | (L)Ostschweizer Fachhochschule

OSPF (IGP)

Equal Cost Multipath (ECMP) load balancing, fast convergence, widely used. Uses IP Port 89 on Layer 3. Uses 224.0.0.5 (all routers), 224.0.0.6 (all DR,BDR)
DR Election: Highest interface priority → if tie, highest Router ID. Priority 0 = ineligible for DR/BDR. No preemption. BDR = same rules, second best candidate

Router Operation

1. Establish neighbor adjacencies and exchange LSAs
2. Build the Link-State Database (LSDB)
3. Run Dijkstra's SPF algorithm:
   - **Intra-area change:** SPF recalculation
   - **Inter-area change:** No SPF needed — ABR handles updates
4. Build the routing table from SPF results

Areas

- AS split into **areas** (sub-domains), each with a 32-bit Area ID (e.g., *0.0.0.0* = Area 0)
- **Backbone Area (Area 0):** Core of the OSPF domain (must exist) Must connect to all other areas (directly or via virtual links), Must be contiguous (no disjointed segments), Should not contain end-user networks
- **Non-Backbone Areas:** Connect end users and local resources, All inter-area traffic must transit the backbone

Router Types

- **Area Border Router (ABR):** Connects two or more OSPF areas, must have 1 interface in backbone. 1 OSPF DB per Area
- **Internal Router:** All interfaces belong to the same area (non-backbone)
- **Backbone Router:** At least one interface in Area 0, Includes ABRs and routers internal to the backbone
- **AS Boundary Router (ASBR):** Connects to external AS/Network (e.g., BGP), Advertises external routes into OSPF, Can be in backbone or non-backbone area

Design Rules

- **Rule 1:** Backbone (Area 0) must be contiguous — no partitions allowed
- **Rule 2:** Every non-backbone area must connect to Area 0

LSA Type Overview

- **Type 1 – Router-LSA:** Sent by all routers, lists directly connected links (so all outgoing interfaces with state and cost) (intra-area only)
- **Type 2 – Network-LSA:** Sent by DR, lists all routers and DR in broadcast/multi-access network (intra-area only)
- **Type 3 – Summary-LSA:** Sent by ABRs, advertises networks from other areas, flooded in all the areas that are not "totally stubby"(inter-area)
- **Type 4 – ASBR-Summary:** Sent by ABRs, advertises path to ASBRs (inter-area)
- **Type 5 – AS-External-LSA:** Advertises external routes (e.g., BGP); flooded to all non-stub areas -> only normal areas
- **Type 7 – NSSA External:** Like Type 5 but used inside NSSA; converted to Type 5 by ABR

Area Types ({allowed LSA Types})

Standard (1,2,3,4,5) Normal

Stub Area (1,2,3)
- Blocks external LSAs (Type 5)
- ABR injects a default route (0.0.0.0)
- Supports LSA Types 1–3

Totally Stubby Area (1,2):
- Blocks external (Type 5) and summary (Type 3/4)
- Only allows default route from ABR

Not-So-Stubby Area – NSSA (1,2,3,7):
- Like stub area, but allows one ASBR inside
- External routes use LSA Type 7 (converted to Type 5 by ABR)

Totally NSSA (1,2,7):
- Like NSSA but blocks Type 3 so gets default route from ABR

Packet Types

- **Type 1 – Hello:** Used to discover, maintain, and verify neighbors; forms adjacencies. Also for election of DR,BDR in broadcast networks. Contains network mask(of sending routers' interface), Hello interval(p2p,broadcast: default=10s), Options, Priority(for election), Router dead interval(default=40s), DR/BDR IP, Neighbors.
- **Type 2 – Database Description (DBD/DD):** Exchange summaries of LSAs during adjacency formation (headers only). Contains Interface Max. MTU, Options, I/M/MS bits (Initial, More, Master-slave bit), DD Sequence num, LSA Header
- **Type 3 – Link State Request (LSR):** Sent when a router needs specific LSAs listed in the DBD. Contains Link State Type (router/network), Link State ID, Advertising Router (sender address)
- **Type 4 – Link State Update (LSU):** Used to flood new or updated LSAs. Contains Number of LSAs, full LSAs information
- **Type 5 – Link State Acknowledgment (LSAck):** Confirms receipt of LSAs to ensure reliable flooding. For this you send LSAck or implicitly by sending LSU with same info back. Many acks may be grouped together to a single LSAck.

Sub-Protocols

- **Hello Protocol:**
  - Used for neighbor discovery and parameter negotiation.
  - Maintains logical adjacencies on P2P, P2MP, and virtual links.
  - Elects DR/BDR on broadcast and NBMA networks.
  - Continuously sends hello packets to maintain bidirectional connectivity; failure to receive = neighbor down (in agreed router dead interval at initialization)
- **Database Sync Protocol:**
  - Syncs LSDB using Database Description (DBD) packets with only LSA headers.
  - Uses I-bit (initial), M-bit (more), and MS-bit (master/slave).
  - **ExStart:** Bi-dir comm; highest Router-ID = master. Determine initial seq nr
  - **Exchange:** Exchange of DBD packets (LSA headers).
  - **Loading:** Missing LSAs are requested.
  - **Full:** Databases fully synchronized.

OSPF Routing and ECMP

- Each router runs Dijkstra per area; link cost = metric from LSAs (1–65535)
- OSPF perfers more specific match (CIDR) and if then still multiple: intra-area > inter-area > external
- Routes added to RIB/FIB based on computed next hops
- **ECMP:** Modified Dijkstra supports Equal-Cost MultiPath if multiple paths have same cost → routes added with multiple next-hops for load balancing

OSPF Route Selection

- **Intra-Area (O):** Source and dest in same area; routes from Type 1 and 2 LSAs
- **Inter-Area (O IA):** Source and dest in different areas within same AS; via Type 3 LSAs through backbone
- **External (E1/E2):** Dest outside AS; info injected by ASBR via redistribution
  - **E1:** Total = external + internal OSPF cost
  - **E2:** Only external cost (default)
- **Preference order:** More specific route > Intra-area > Inter-area > E1 > E2
- **Cost calculation:** Cost = Reference Bandwidth(default 100 Mbps)/Interface Bandwidth

IS-IS (IGP)

CLNS - Connectionless Network Services

- **CLNS:** ISO Layer 3 datagram service; supports **CLNP**, **ES-IS**, **IS-IS**.
- **CLNP:** Connectionless Network Protocol, similar to IP, used in ISO stack (Ethernet Type 0xFEFE).
- **IS-IS:** Link-state routing protocol (Layer 3); forms adjacencies with ES-IS; designed for CLNP but extended (Integrated IS-IS) to support IP.
- **Integrated IS-IS:** Allows IP routing with IS-IS; used widely by service providers even without CLNP.

NET Addressing (type of NSAP (Network Service Access Point) address)

- **NET = Network Entity Title:** Unique router identifier in IS-IS
- **Format:** *49.AAAA.BBBB.BBBB.BBBB.00*
- **49.AAAA...:** Area ID (variable length)
- **BBBB.BBBB.BBBB:** System ID (usually 6 bytes = unique router ID)
- **00:** N-Selector (NSEL) (always *00* for routers)
- **System ID:** Must be unique per router (often based on lo-IP/MAC)
- Example: *49.0001.1921.6800.1024.00* based on IP 192.168.1.24
- **Area ID:** Used for routing hierarchy (like OSPF areas)

IS-IS Packet Types (all in L1 or L2)

- **IIH (Hello):** Builds and maintains adjacencies; includes system ID, holding time, prio
  - Built from 3 functions: **discover, build, maintain**
  - **Interval:** 10s (default); DIS sends every 3.3s on LANs
  - **Multiplier:** Missed Hello limit → Holdtime = Interval × Multiplier(default=3)
  - **Multicast:**
    - *01-80-C2-00-00-14* (AllL1ISs)
    - *01-80-C2-00-00-15* (AllL2ISs)
- **LSP (Link State PDU):** Contains topology info including prefixes with costs; flooded throughout the area -> similar to OSPF LSA Type 1
- **CSNP (Complete SNP):** Sent by DIS; lists all known LSPs (used for database sync)
- **PSNP (Partial SNP):** Used to request missing LSPs or acknowledge received LSPs
- **IS-IS Packet Structure:** Common header + TLVs

Router Types

- **L1 Router:** Only within one area; no inter-area routing
- **L2 Router:** Backbone router; routes between areas
- **L1/L2 Router:** Acts as both; separates databases, redistributes between levels

Broadcast/Multi-Access Links: DIS – Designated IS

- Required on broadcast links (no DIS on p2p)
- Sends periodic **CSNPs** to ensure DB sync, creates **pseudonode LSP**
- No backup DIS in IS-IS

DIS Election

- 1. Highest interface priority (0–127) Ciso default = 64
- 2. Highest SNPA (MAC-Address!)
- **Preemption:** Enabled – higher prio router automatically takes ver the DIS Role

Point-to-Point links (No DIS)

- **CSNP:** Sent once at adjacency startup
- **LSP:** Advertises topology changes (link-state info)
- **PSNP:** Acknowledges received LSPs or requests missing ones

Path Selection

Path Selection Order (in IS-IS) - Lower Metric better:

1. L1 intra-area routes
2. L2 intra-area routes
3. Leaked L2→L1 (internal metric)
4. L1 external (external metric)
5. L2 external (external metric)
6. Leaked L2→L1 (external metric)

Level 1 Routing

- Intra-area routing only (like OSPF intra-area)
- L1 routers use the closest L1/L2 router for inter-area traffic
- **L1/L2 routers:**
  - Do **not** advertise L2 routes into the L1 area (unless route leaking is active)
  - Set **Attached bit** to signal L2 connectivity to backbone
- L1 routers install a **default route** to nearest L1/L2
- L1 area like OSPF Totally Stubby Area
- **Distribution Bit:** Set to 'up' (1) on L2→L1 leaks; blocks re-advertisement L1→L2.
- **Route-Leaking** injects a more specific route into L1 to improve routing

Level 2 Routing

- Routing between areas (inter-area)
- L1/L2 routers inject L1 routes into L2 topology
- L1 routes are redistributed into L2 with L1 metric preserved in L2 LSP

IS-IS vs OSPF

Feature	IS-IS	OSPF
Layer	L2 (CLNS)	L3 (IP, proto 89)
Encapsulation	No IP, uses TLVs	IP packets
Hello Type	IIH	Hello packet
Area Model	L1/L2	Backbone + Areas
Metric	Cost (default 10)	Cost (bandwidth)
Router ID	System ID (6B)	32-bit Router ID
Adj. Types	L1, L2, L1/L2	DR/BDR, P2P
LSDB	Per level (L1/L2)	Per area
Scaling	Large-scale ISP core	Enterprise/campus
Routing Info	TLVs (flexible)	Fixed LSA types

BGP (EGP)

Config

- **next-hop-self** fixing iBGP: neighbor (neighbor IP) next-hop-self (when overriding)

BGP Sessions

- Point-to-point adjacencies between BGP routers
- **iBGP:** Between routers in the same AS, AD=200, more trusted (lower security overhead)
- **eBGP:** Between routers in different ASes, AD=20, stricter policy enforcement

Autonomous System Numbers (ASN)

- Unique ID for each AS; required for Internet routing with BGP
- Private ranges:
  - 64 512–65 534 (legacy 16-bit)
  - 4 200 000 000–4 294 967 294 (32-bit)

BGP Peering / Neighbors

- Two routers with a BGP TCP session (port 179) are called peers or neighbors
- Each BGP router is a **BGP speaker**
- BGP exchanges routing info between ASes (loop-free, policy-based)
- Supports CIDR, route aggregation; decisions based on policies/rules

Path Attributes

- Used for route control and policy enforcement
- **Well-known mandatory:** Always present (e.g., AS-Path, Origin, Next Hop)
- **Well-known discretionary:** Optional but recognized by all (e.g., Local Pref, Atomic Aggregate)
- **Optional transitive:** Passed between ASes (e.g., Community, Aggregator)
- **Optional non-transitive:** Not passed across ASes (e.g., MED, Weight, Originator ID, Cluster ID/List)
- **NLRI:** Routing table info: prefix, prefix length, and associated path attributes

Loop Prevention

- BGP uses **AS-Path** (list of ASNs) to detect loops
- If a router sees its own ASN in a received route, it discards it
- **AS-Override:** Allows reuse of same ASN across different customer sites (e.g., Swisscom); rewrites ASN to avoid loop detection

BGP Messages

- **OPEN:** Establishes session; includes version, ASN, Hold Time, BGP Identifier, optional params
  - **Hold Time:** Heartbeat in seconds (default 180s, Cisco), reset by KEEPALIVE/UPDATE; 0 = session down
  - **BGP Identifier:** 32-bit Router-ID, manually set or highest loopback/active IP; used for loop prevention
- **KEEPALIVE:** Sent every 1/3 of Hold Time (default 180s); ensures neighbor liveness (BGP doesn't rely on TCP keepalive)
- **UPDATE:** Advertises new routes, withdraws old ones, or both; includes NLRI (prefix + path attributes); can act as KEEPALIVE
- **NOTIFICATION:** Sent on session error (e.g. hold timer expired); terminates session immediately

BGP Network Statements

- **Purpose:** Advertise specific prefixes to BGP peers (does not activate interfaces)
- Prefix must exist exactly in the RIB (from static, connected, or learned route)
- Attributes (e.g., origin, next-hop, MED) depend on how the route exists in RIB
- BGP advertises only the best path for a prefix to peers, even if multiple exist

Best Path Calculation

- BGP maintains all received paths per prefix but advertises only the best one
- Best path is installed in RIB; recalculated on:
  - Next-hop reachability change
  - Interface failure to eBGP peer
  - Redistribution change
  - New/withdrawn path received
- **Influence:**
  - Outbound BGP policy → inbound traffic behavior
  - Inbound BGP policy → outbound traffic behavior

BGP Best Path Selection (in order):

1. Prefer highest **Weight** (Cisco-specific, local to router)
2. Prefer highest **Local Preference** (global within AS)
3. Prefer routes **originated by the router** (only small i in path, NH 0.0.0.0)
4. Prefer **shorter AS path** (only length is compared)
5. Prefer **lowest origin type:** IGP < EGP < Incomplete (I on origin)
6. Prefer **lowest MED** (Multi-Exit Discriminator) (also called metric)
7. Prefer **external (EBGP)** over **internal (IBGP)**
8. For iBGP: prefer path with **lowest IGP metric** to next-hop
9. For eBGP: Prefer **oldest** (more stable) path
10. Prefer **lowest BGP router ID**
11. Prefer path from **lowest neighbor IP address**

Route Filtering

- Filters control which routes are received/advertised
- Used for security, traffic shaping, memory optimization
- Tools: **prefix-list** (IP), **filter-list** (AS-path), **route-map** (flexible match/set)

BGP Communities

- 32-bit optional, transitive tag (e.g. *ASN:value, 65000:100*)
- Used to mark routes for policy control across ASes
- Can be added, modified, or removed at each hop

iBGP Scalability

- iBGP does not re-advertise routes between iBGP peers → full mesh required (cause no loop prevention)
- Session count = $n(n-1)/2$ (e.g., 5 routers = 10 sessions, 10 = 45)

Route Reflectors (RR)

- Solves iBGP full-mesh scaling by allowing selective route reflection
- Clients only peer with RR; unaware they're clients
- **RR Rules:**
  1. From **non-client** → advertise to **clients only**
  2. From **client** → advertise to **all** (clients & non-clients)
  3. From **eBGP peer** → advertise to **all** (clients & non-clients)
- Only the RR needs special config - clients remain unaware of route reflection. This eliminates the need for full iBGP mesh.

Peering vs. Transit

- **Transit:** ISP provides full reachability (paid relationship)
- **Peering:** ISPs exchange selected routes; equal relationship, usually unpaid
- **AS Path Filtering:** to avoid getting transit: ip as-path access-list 10 permit $\hat{}\$$

Internet Exchange Point (IXP)

- Facility where networks exchange traffic via BGP peering
- Reduces transit costs, latency, and offloads upstream links

Public Peering

- Members peer via shared switch fabric and a **route server**
- Route server distributes routes but stays out of data path (NEXT_HOP unchanged)
- Minimal policy control; one BGP session to route server
- Simplified setup (one legal contract)

Private Peering

- Direct BGP sessions between two parties (1:1)
- May use public or private interconnects
- Full policy control per neighbor
- Requires one session and legal contract per peer

Examples: Equinix, SwissIX (non-profit)

Enterprise Connectivity Options

- **Single-Homed:** One ISP, one link (BGP or static); simple but no redundancy
- **Dual-Homed:** One ISP, two links (or routers); redundancy within same provider
- **Multihomed:** Multiple ISPs; improved redundancy and routing control, but avoid being a transit — advertise only customer-owned prefixes
- **Dual-Multihomed:** Multihomed, but two links per ISP

Traffic Engineering (TE)

- **Outbound TE (Local Pref):** Set higher local pref to prefer exit path; affects outbound traffic; highest wins
- **Inbound TE (MED):** Signal entry preference with MED; lowest wins; only works if peer honors it
- **Inbound TE (AS-Path Prepending):** Add own ASN multiple times on backup path; shortest AS-path wins
- **TE Limitation:** AS controls outbound (e.g. local pref); inbound control limited — ISPs may ignore MED
- **TE with Aggregate:** Prefer primary ISP with summarized routes; advertise specific prefixes on backup for failover
- **Aggregate Impact:** Longest-match wins → specific prefixes may steer traffic to alternate ISP; avoid provider-owned aggregates

RPKI – Resource Public Key Infrastructure

- Framework to validate that a prefix is legitimately originated by a specific ASN
- Prevents route hijacking and accidental mis-originations (route leaks)
- RPKI validates origin only — not the AS-path

Key Components

- **Trust Anchors (TAs):** Root CAs of the 5 RIRs; issue certs for resource holders
- **ROA – Route Origin Authorization:** Digitally signed object that authorizes ASN to announce a prefix (contains AS, prefix that AS can originate, max prefix length)
- **RPKI Validators:** Software that downloads, verifies, and stores Validated ROA Payloads (VRPs)

RPKI Validation States

- **Valid:** Prefix + ASN match ROA
- **Invalid:** Prefix found, but ASN mismatch or prefix too long
- **Not Found / Unknown:** No matching ROA

Validator Details

- Use TALs (Trust Anchor Locators) to fetch data from RIRs
- Validate cryptographic signatures (via X.509 certs with RFC 3779)
- Outputs VRPs; invalid objects are discarded
- Update frequency: >=24h (recommended), 30–60min (practice)
- RRDP (RFC 8182) replacing rsync (uses HTTPS)

BGP Monitoring

Event tracking, BGP hijack detection, Route leak detection, RPKI status check, Reachability tracking, AS path change tracking, AS path visualization

Network Design

Pillars: Scalability, Speed, Availability, Security, Manageability -> overall Cost

Availability Concepts (most important requirement)

- **MTBF:** Mean Time Between Failures, **MTTR:** Mean Time to Repair
- $A = \frac{MTBF}{MTBF+MTTR}$ MTBF combined: $\sum \frac{1}{\frac{1}{MTBF_n}}$ parallel: $MTBF \cdot \frac{3}{2}$
- Lower MTTR + higher MTBF = more availability

Redundancy

- Adds reliability, decreases MTBF but increases MTTR and complexity
- Balance: resilience vs. manageability
- **Backup Paths:** Duplicate devices/links on primary path, build extra links for redundancy, consider backup link capacity, consider failover speed
- **Load Balancing:** ECMP, EtherChannel, Port-Channel

Hierarchical Design

- **Access:** Connect end devices, high port count, port security, L2, QoS marking
- **Distribution:** L3, policy control, HSRP/VRRP, loop protection, small fault domain
- **Core:** High-speed backbone, L3 only, no policy, scalable/redundant, no security
- **Collapsed Core:** Combines Core + Distribution (small/medium networks)

Flat Design

- Any-to-any; small networks, MPLS/LAN setups
- Lacks scalability and control

Fabric Design

- Modern design using **Underlay/Overlay**
- Underlay = transport (e.g., IP, MPLS)
- Overlay = logical virtual topology (e.g., EVPN)

Enterprise Campus

- 100s/1000s of users, multiple buildings, one physical location
- Multiple interconnected LANs, connected via Ethernet and Wireless

FHRP – First Hop Redundancy Protocols

- Ensures default gateway is always reachable, PCs can only have one
- Enables fast failover during router failure
- **VRRP – Virtual Router Redundancy Protocol (Multivendor):** Shared virtual IP, real MACs per router, Master router handles forwarding
- **HSRP – Hot Standby Router Protocol (Cisco):** Shared virtual IP + virtual MAC, One active, others in standby -> faster
- **GLBP – Gateway Load Balancing Protocol (Cisco):** Load balancing + redundancy, Shared virtual IP + multiple virtual MACs, Roles: **AVG:** Answers ARP and sends virtual MAC addresses of AVFs **AVF:** Forwards traffic

Data Center Designs

- **ToR - Top of Rack:** switches per rack, less cabling, easy expansions/exchanges per 'rack', scalable glass fiber, ideal for high service density (full racks). But more switches, more ports, more L2 Srv-2-Srv traffic, more STP to be managed
- **EoR - End of Row:** 1 switch per row, less switches, higher utilization of ports, switches all at one place, better L2 availability between racks. But more cabling

Traffic Directions

- **North–South:** Between external networks (client-server, in/out DC)
- **East–West:** Within DC (e.g., server-server, storage)

Three-Tier DC

- Access – Aggregation – Core (like Hierarchical)
- Optimized for North–South traffic
- Not ideal for East–West communication

Leaf-Spine Architecture

- Two-tier: Leaf switches (access) connect to Spines (core)
- High performance, low latency
- Scalable, ideal for East–West traffic

Multicast

Use Cases: 1-to-Many: Streaming, software updates, music-on-hold **Many-to-Many:** Gaming, VR, stock data, group chat

Benefits: Efficient bandwidth, lower server/CPU load, no redundancy, supports multipoint apps

Properties: UDP-based (no delivery guarantee, congestion control, or ordering) Apps must handle drops, duplicates, out-of-order packets

Source: Sends to group IP; doesn't need to join **Receiver:** Must explicitly join group to receive traffic

Multicast Address Ranges

- **224.0.0.0 – 224.0.0.255:** Link-local, TTL = 1 (not forwarded by routers)
- **224.0.1.0 – 224.0.1.255:** Reserved by IANA, routable
- **232.0.0.0 – 232.255.255.255:** Source-Specific Multicast (SSM)
- **239.0.0.0 – 239.255.255.255:** Administratively scoped (private multicast space)

Broadcast Basics

- **L2 (Bridging):** MAC ffff.ffff.ffff; switches flood to all ports in VLAN
- **L3 (Routing):**
  - 255.255.255.255: local broadcast, not routed
  - Directed broadcast (e.g. 10.1.1.255): can be routed if enabled
- L3 routes between subnets; L2 floods within same subnet

L2 Multicast: MAC Mapping

- **Step 1 – Get IP:** Example multicast IP address: *239.5.5.5*
- **Step 2 – Convert to Binary:** *239.5.5.5* = 11101111.00000101.00000101.00000101 → Take only the last 23 bits: 00000101.00000101.00000101
- **Step 3 – Map to MAC Prefix:** Use fixed MAC multicast prefix: 0100.5E → Map last 23 bits to: 0100.5E.05.05.05
- **Step 4 – Final MAC Address:** 0100.5E05.0505

IGMP – Internet Group Management Protocol (Host to first-hop-router)

Purpose: Manages group membership for IPv4 multicast on each segment

- **IGMPv1**
  - Basic join via query-response mechanism
  - No way for a host to leave a group explicitly
  - Router sends general membership queries every 60s to *224.0.0.1*
  - If no report is received, router removes group after timeout
  - Receiver has no knowledge of the multicast source
- **IGMPv2**
  - Adds *Leave Group* message (faster pruning of unused traffic)
  - General queries to *224.0.0.1* every 125s 'Any hosts interested in any groups?'
  - Supports *Group-Specific Queries* e.g. when someone leaves group ('anyone still interested in group xy?'), reducing broadcast overhead
  - Still source-agnostic: receivers don't know who the source is
- **IGMPv3**
  - Adds *source filtering* (Include/Exclude lists)
  - Enables Source-Specific Multicast (SSM) – receiver requests traffic only from selected source(s), no need for Rendezvous Point (RP) anymore
  - Adds support for application-level access control and filtering
  - Can also be used in ASM (Any Source Multicast), but mainly with SSM

IGMP Snooping

- Without snooping: multicast = broadcast on VLAN
- With snooping: switch listens to IGMP messages and builds a forwarding table
- Default: snooping is enabled; switch needs IGMP Query to operate

Reverse Path Forwarding (RPF)

- **RPF Check:** To avoid loops, verifies that a multicast packet arrives on the interface that a unicast packet destined for the multicast source would be forwarded out of.

Rendezvous Point (RP)

- Used in Shared Tree (**,G*) setups with PIM-SM
- RP acts as the common meeting point for sources and receivers
- RPF check is performed toward the RP (not the source)
- Once the source is known, routers may switch to a Source Tree (*S,G*)

Shared Tree vs. Source-Based Tree

Shared Tree (*,G):
- IGMP host sends a membership report (IGMP Join)
- Router adds *(*,G)* entry to multicast routing table
- * means 'any source' — source is unknown/unspecified

Source-Based Tree (S,G):
- Built when router receives an *(S,G)* join/report from IGMP host
- *S* = known multicast source; *G* = group
- Router adds *(S,G)* to mroute table once source is known

PIM – Protocol Independent Multicast (only between routers)

- Relies entirely on the unicast routing table (RIB) for multicast forwarding decisions
- Protocol-independent: works with static routes, OSPF, IS-IS, etc.

PIM-DM – Dense Mode (no RP)

Push Model: Floods multicast traffic to all interfaces; then prunes where no receivers exist.

- **1. Flooding:** Source sends traffic → forwarded out all multicast-enabled links using unicast RIB
- **2. Distribution Tree:** Initially includes entire network (shared tree rooted at source)
- **3. Prune Messages:** Routers without interested receivers send prunes upstream to remove themselves from the tree
- **4. State Maintenance:** Routers track source, receivers, interfaces to forward/prune per group

PIM-SM – Sparse Mode (ASM, SSM)

Pull Model: Multicast traffic is only sent where requested. Works with IGMP to detect interested receivers and uses unicast routing for forwarding.

- **1. Join/Prune:** Routers send explicit join/prune messages to request or stop receiving multicast for group (G) to other routers
- **2. Forwarding:** Routers only forward multicast packets for group (G) on interfaces from which explicit joins were received

ASM – Any Source Multicast (How it works with RP)

- Works with **IGMPv1** or **IGMPv2** → receiver does not know the source
- Receiver sends *IGMP Join (*,G)* to its first-hop router
- First-hop router forwards *PIM Join (*,G)* hop-by-hop toward the Rendezvous Point (RP)
- RP acts as a common meeting point for sources and receivers
- Sources send multicast traffic to the RP via a *PIM Register* tunnel
- All routers in the multicast domain must know the RP location

SSM – Source Specific Multicast

- Receiver subscribes using **IGMPv3**, providing both source (S) and group (G) to the first-hop router
- No Rendezvous Point (RP) required → **PIM-SSM builds only (S,G)** Shortest Path Trees (SPT)
- **No shared tree (*,G)** model used; SSM is source-directed
- IANA reserved *232.0.0.0/8* for SSM in IPv4
- Join messages are forwarded hop-by-hop toward the source to establish forwarding path
- Uses unicast routing table (RPF) to maintain loop-free delivery

PIM Sparse-Dense Mode

- **PIM Sparse Mode:** Pull model – multicast traffic is forwarded only on request
- **PIM Dense Mode:** Push model – traffic is flooded everywhere, then pruned
- **Sparse-Dense Mode:** Supports both modes per multicast group; choice depends on RP availability

PIM-DM vs. PIM-SM

PIM is protocol-independent: it relies on the unicast routing table for RPF checks and to forward joins toward the source or Rendezvous Point (RP).

- **PIM Dense Mode (DM)**
- "Push"model
- Floods multicast traffic throughout the network
- Prunes back where traffic unwanted

- **PIM Sparse Mode (SM)**
- "Pull"model
- Traffic sent only on request
- Requires explicit *Join* messages

Usage Recommendation:
- **Dense Mode:** Best for small or tightly scoped networks where most devices need multicast
- **Sparse Mode:** Preferred for large-scale or distributed environments where multicast receivers are few or spread out

VXLAN

Issues of L2: STP, Max amount of VLANs (4094), Large MAC Address tables

- **VXLAN (Virtual Extensible LAN):** Tunnels Ethernet (Layer 2) over IP using MAC-in-UDP encapsulation (Port 4789). For flexible and scalable network segmentation.
- **VNID (VXLAN Network Identifier):** 24-bit identifier (up to 16 million segments) that defines the VXLAN broadcast domain.
- **VTEP (Virtual Tunnel Endpoint):** Device (switch, router, or host) responsible for encapsulating/de-encapsulating VXLAN traffic.
- **NVE (Network Virtual Interface):** Logical interface on a VTEP used for VXLAN tunnel operations.

Tunnel

- VXLAN establishes IP tunnels between VTEPs to extend Layer 2 networks across Layer 3 boundaries.
- VXLAN enables both L2 and L3 VPN functionality in overlay networks.
- VXLAN traffic is encapsulated in UDP (default port: *4789*).

Frame Format

- Ethernet frame → VXLAN Header → UDP → Outer IP Header.
- The VXLAN header contains the 24-bit VNID and flags.
- Outer headers allow Layer 2 frames to traverse IP underlay networks.

Virtual Network Identifier (VNI)

- 24-bit VXLAN Network Identifier uniquely defines VXLAN segments.
- Replaces traditional VLAN IDs (12-bit), enabling 16 million logical segments.
- Used by VTEPs to map traffic into corresponding Layer 2 domains.

VXLAN Tunnel Endpoint (VTEP)

- Connects the overlay (VXLAN) and underlay (IP) networks.
- **Types:**
  - **Software VTEP:** Located on hypervisors using virtual switches.
  - **Hardware VTEP:** Located on routers/switches with ASICs for performance.
- **Interfaces:**
  - **VTEP IP Interface:** Connects to the underlay network and handles encapsulation.
  - **VNI Interface:** Virtual interface per segment (like SVI); handles segregation of Layer 2 domains.

MAC Address learning

- **On control plane:** happens proactively, **on data plane:** ad-hoc with flooding
- Each VTEP maintains a VXLAN mapping table linking destination MAC addresses to remote VTEP IPs.
- **Learning via ARP:**
  1. Host H1 sends ARP request, switches learn H1's MAC.
  2. ARP request is flooded to H2.
  3. H2 responds; switches learn H2's MAC.
- **Learning Methods:**
  - **Static VXLAN:** Manual MAC-to-VTEP mappings. Doesn't scale well; BUM traffic is inefficient.
  - **Multicast VXLAN:** VTEPs join multicast groups per VNI. Scales better, offloads BUM replication. 20+ VTEPs = there is too much traffic, doesn't scale well
  - **MP-BGP EVPN:** Modern solution using BGP as control plane. Dynamically learns MAC/IP info.

EVPN

Overcome flood-and-learn limitations, doesn't rely on data plane learning, utilizes robust control plane MP-BGP, works with different encapsulation techniques (VXLAN, MPLS), excellent scalability, l2 and L3 Support.

MP-BGP EVPN (Multiprotocol BGP for Ethernet VPN)

- Enables protocol-based VTEP discovery and host reachability via control-plane learning
- Reduces flooding by replacing data-plane learning
- Extends BGP with multiprotocol capabilities (AFI/SAFI)
- Uses *MP_REACH_NLRI* and *MP_UNREACH_NLRI* for route advertisement and withdrawal

EVPN Route Types

Type 2 – Host Advertisement: Advertises host MAC (mandatory), optionally IP, along with L2VNI and optionally L3VNI. Used for MAC learning, ARP suppression, and host mobility. Sent when host connects to VTEP.

Type 5 – Subnet Advertisement: Advertises IP prefix + prefix length with L3VNI. Used for inter-subnet routing. VTEP redistributes connected/static/dynamic IP routes. Additional attributes: L3VNI, extended communities.

Host Deletion & Move

Host Deletion: When a host detaches, its ARP (default: 1500s) and MAC entry (default: 1800s) time out on the VTEP. Upon aging, the VTEP withdraws the host's MAC/L2VNI and IP/L3VNI advertisements.

Host Move: When a host moves to a new VTEP, the new VTEP advertises updated reachability with a higher move sequence number. The old VTEP withdraws its entry, completing the migration.

Route Distinguisher (RD) vs. Route Target (RT)

- **Route Distinguisher (RD):** Uniquely identifies VPN routes — allows same IP prefix to be used in different VPNs. Can be IPv4 or ASN *Used to make routes unique in BGP (VPNv4/v6)*. Forms VPNv4 NLRI: *RD:IPv4 prefix*
- **Route Target (RT):** Controls route import/export between VRFs. *Used as extended BGP community.*
- **How RTs Work:**
  - A route is tagged with an RT when advertised by BGP.
  - Other VRFs import the route if the RT matches their import policy.
  - Allows overlapping or shared connectivity between tenants (e.g., shared services).
- **Format:** Typically in the form *ASN:nn* or *IP:nn*, e.g., `65000:100`, `1:10`
- **Multiple RTs can be used:** A route can have multiple RTs for flexible policies (e.g., one RT for VPN, another for shared services)

EVPN (Ethernet VPN) – L2

Key Features

- L2 bridging across L3 networks
- **BGP Control Plane:** Distributes MAC info (no flooding)
- **VXLAN Overlay:** Encapsulates L2 in L3 UDP (data plane)
- **Multi-Tenancy:** via VNI segmentation
- **Redundancy:** All-active multihoming, ECMP, fast convergence

Use Cases

- Multi-tenant datacenter interconnects (DCI)
- Extending L2 over WAN between remote sites
- Scalable, segmented L2 fabrics

BGP Control Plane

- PEs learn MACs from local CEs (data plane)
- MACs advertised via BGP (control plane)
- Uses Route Distinguishers and MPLS labels
- Remote PEs update L2 RIB/FIB with MAC and next-hop info
- Enables seamless L2 across IP/MPLS backbone

EVPN NLRI

- EVPN uses MP-BGP with specific AFI/SAFI
- Supports multiple route types and attributes
- Unsupported routes are dropped by BGP

Autodiscovery via Route Reflectors

- Route Reflector (RR) avoids full-mesh iBGP
- RR reflects EVPN routes to other PEs
- RR doesn't participate in EVPN or pseudowires
- RR needs only *address-family l2vpn evpn*
- L2VPN RIB stores endpoint/VFI info for control plane
- *BGP_UPDATE* from spines contain *ORIGINATOR_ID* (origin leaf)

Host Detection

- Host connects to VTEP → MAC learned locally
- VTEP advertises MAC + L2VNI via BGP EVPN
- MAC learning follows normal Ethernet semantics

Ingress Replication (IR)

- BUM traffic, when Multicast underlay network is not used, handle multi-destination traffic (ARP → unicast)

Early ARP Termination (ARP Suppression)

- Avoids flooding ARP requests
- VTEP queries control plane for MAC/IP/VNI mapping
- If known → direct unicast (no broadcast)

Silent Host Flow (Fallback)

- If IP/MAC unknown → ARP sent via **ingress replication**
- Replicated ARP request goes to remote VTEPs
- Only correct host responds → update reflected to all VTEPs
- Future traffic uses updated BGP mapping

VRF – Virtual Routing and Forwarding

- Multiple isolated routing tables on one device
- Each tenant = one VRF → traffic isolation
- Supports independent policies per tenant
- Key for scaling and multi-customer separation

IRB – Integrated Routing and Bridging

- Enables inter-VLAN routing inside EVPN
- Avoids central gateway → no 'traffic tromboning'
- Two modes: Symmetric and Asymmetric

Symmetric IRB (L2 + L3)

- Routing/bridging on ingress + egress VTEPs
- Uses L3 Transit VNI (same in both directions), One L3 VNI per VRF (Tenant)
- Scales well; clean separation of MAC and IP

Asymmetric IRB (L2)

- Routing only on ingress, bridging on egress
- VXLAN uses destination VNI in both directions
- One L2 VNI per VLAN/Subnet
- Simple config, but requires all VLANs/VNIs on all VTEPs

Distributed Anycast Gateway (DAG)

- Same gateway IP+MAC on all VTEPs
- Enables local default gateway for hosts
- Supports mobility + optimal forwarding

L3 Host Detection

- Host sends ARP/ND to local VTEP
- VTEP learns MAC/L2VNI and IP/L3VNI
- Info is advertised in EVPN (control plane)

MPLS

- Label Switched Path (LSP) → pre-determined path across MPLS network
- advantage eBGP between PE-CE: No mutual redistribution, same routing process
- encrypt traffic flowing over MPLS L3VPN backbone? yes (e.g. bank)
- Unicast Reverse Path Forwarding (uRPF): checks source of each packet & verifies that source is in routing table
- control plane (e.g. OSPF) → to learn labels
- iBGP used to exchange NLRI (RD, RT, IPv4 Prefix, NextHop &VPN Label) between PE
- imp-null = networks are directly connected, no more label switching

WAN

Connects remote LANs via SPs for data/voice/video; key needs: bandwidth, control, design, resilience, mgmt. **Requirements:**

- **Bandwidth:** App needs, peak usage, reserve for VoIP
- **Control/Security:** Trust provider? No full control
- **Availability:** Redundancy, SLA for failures
- **Mgmt:** Inband vs out-of-band

Private WAN

- **Point-to-Point:** Leased L2 line (Ethernet); monthly fee; private circuit
- **Dark Fiber:** Physical fiber lease; costly; ISPs prefer selling lambdas
- **Connection-oriented:** Predefined path, packets carry IDs (ATM, Frame Relay)
- **Connectionless:** No setup; full address in each packet (Ethernet, MPLS VPN)

Terminology

- **CE - Customer Edge:** no knowledge of MPLS, no labels; connected to PE
- **PE - Provider Edge:** connected to CE; runs iBGP and LDP; uses VRFs
- **P - Provider or LSR(Label Switch Router):** inside MPLS VPN, no CE connection; forwards labels

Databases, Planes

- **RIB (Routing IB (Information Base)):** Learned prefixes from routing protocols
- **FIB (Forwarding IB):** Built from RIB; only best routes for forwarding
- **LIB (Label IB):** All label mappings; 1 label per prefix
- **LFIB (Label Forwarding IB):** Built from LIB; used for actual forwarding decisions
- (L)FIB only contains currently best LSP (decision: Routing Protocol)
- **Control Plane:** Builds routing/label tables (RIB, LIB)
- **Data Plane:** Forwards packets (FIB, LFIB); pushes/swaps/pops labels (see below)

MPLS Header

4-byte header before IP:

- **Label (20b)** – actual MPLS label
- **EXP (3b)** – QoS/CoS, Now called Traffic Class (TC)
- **S-bit (1b)** – bottom of label stack indicator, 1 = True = last label before IP header
- **TTL (8b)** – time-to-live (eq toual IP TTL)

TTL MPLS

- ingress PE router decrements IP TTL field & copies packet's IP TTL field into new MPLS TTL
- P routers decrements MPLS TTL
- egress PE router decrements MPLS TTL, pops final MPLS header, copies IP TTL
- traceroute receive ICMP Time Exceeded, Provider doesn't want to expose MPLS network to fix: disable MPLS TTL propagation (on PE), PE set MPLS TTL = 255, egress leaves PE original IP TTL unchanged
- = MPLS network appears as single router hop from TTL perspective

Label Distribution Protocol (LDP) - Control Plane

- Distributes labels to neighbors using control plane
- **Hello messages:** Sent via UDP (Port 646) to 224.0.0.2 to discover neighbors
- TCP (Port 646) connection is used to exchange label bindings (prefix to local label)
- Routers advertise all local bindings after TCP session is up
- Label mapping used to build LIB → LFIB
- LDP router ID must be reachable (via routing table)
- Each router manages local labels independently

MPLS L3 Data Plane

- VPN traffic uses **2 labels** (stacked):
- **Outer label:** Transport label (LDP); identifies LSP between ingress/egress PE
- **Inner label:** VPN label (MP-BGP); identifies customer VRF
- **Push:** Ingress PE; classify and label packets
- **Swap:** P router; replaces label, forwards based on new label
- **Pop:** Egress PE removes label; sends original packet to CE
- **Penultimate Hop Popping:** MPLS feature, penultimate router removes the outer MPLS label before forwarding to egress PE, default enabled, ISPs disable it

VRF Tables

- VRF = Virtual Routing and Forwarding table (Virtual router inside a PE. Maintains isolated RIB + FIB per customer.)
- Stores separate routing info per customer (VPN isolation)
- Exists per MPLS-aware PE router; one per attached customer
- Contains: RIB, FIB, and separate routing process per CE

VPNv4

- 64-bit RD + 32-bit IPv4 = 96-bit VPNv4 prefix
- transferring VPNv4 between PE router → Multiprotocol iBGP (MP-iBGP)

Overlay Technologies

Modern Provider Network

- **MPLS:** Label-based forwarding (fast, scalable)
- **LDP:** Distributes labels for MPLS paths
- **IGP:** Underlay routing (e.g., OSPF, IS-IS)
- **MP-BGP:** Extends BGP to carry VPNv4/v6, EVPN routes

Drawbacks of Traditional Networks

- **Control Plane:** LDP/RSVP-TE adds complexity
- **Scalability:** Per-flow/path state limits growth; LSP and signaling overhead increase rapidly
- **OAM:**
  - **Troubleshooting:** Traceroute less useful in MPLS; labels hide topology
  - **Traffic Eng.:** LDP lacks TE; relies only on IGP cost
- **Fast Reroute:** Limited coverage; microloops possible

Segment Routing (SR)

- **Source routing:** Sender defines full path using Segment List (Segment = Instruction)
- **SID = Segment Identifier:** Each SID = 1 instruction (e.g., forward via ECMP, specific iface, or to a service)
- **State in packet:** No per-flow state in network; intermediate nodes follow SID instructions
- **No new control plane:** Uses existing protocols (OSPF, IS-IS, BGP) with extensions; no LDP or RSVP-TE needed
- **Segment List:** Ordered SID list carried in packet header; defines full route
- **Simple but powerful:** Enables TE, fast reroute, policy routing

Segment List Operations

- **Push:** Insert SIDs into packet; set active SID (top of list)
- **Continue:** Active SID not yet completed; keep processing it
- **Next:** Current SID completed; activate next SID in list

Segment Significance

- **Global Segments:** Known and supported by all SR nodes in the domain, Installed in forwarding tables across the network (e.g. "Forward packet according to shortest path to Node1")
- **Local Segments:** Defined and installed only on originating node, Not forwarded by others, but must be understood network-wide (e.g. "Forward packet on interface to Node2")
- **Global segments** are defined in the SR Global Block (SRGB) and should be consistent across all nodes; **local segments** are defined in the SR Local Block (SRLB) and are specific to the local SR node.

SR Control Plane Segment Types

- **IGP Prefix Segment:** Global SID tied to IGP prefix (multi-hop); all nodes install forwarding entries
- **IGP Node Segment:** Global SID for a specific node (shortest-path forwarding)
- **IGP Anycast Segment:** Global SID for a group of nodes; traffic sent to nearest
- **IGP Adjacency Segment:** Local SID; direct link to neighbor
- **L2 Adjacency SID:** Local SID for Layer-2 segment (e.g., Ethernet link)

Combining Segments:

- End-to-end paths can mix IGP and BGP segments
- Traffic to BGP Anycast → more ECMP in data centers

SR-MPLS

- Reuses existing MPLS data plane — no hardware change needed
- Segments = MPLS labels; Segment List = label stack (top = active)
- Segments distributed via IGP/BGP; no LDP required (interoperable if needed)
- Supports both IPv4 and IPv6 networks

Benefits of Segment Routing

Benefits: Simplification (removes protocols, simple operations, admin and mgmt), enhanced Traffig eng. (Delay, Bandwidth, Packet Loss, TE metric, Controller, Source-Node), Seamless deployment, Robust, Network Innovation (zB Container Networking)

- **Source Routing:** Balances distributed intelligence with centralized optimization
- **TI-LFA:** Fast reroute technique; protects against link/node failure with microloop avoidance and no pre-calculation dependency
- **Traffic Engineering (TE):** Optimizes network performance by analyzing and controlling data flow to reduce congestion and improve QoS
- **Service Function Chaining (SFC):** Chains SDN services in order; automates traffic between VNFs and optimizes routing for performance

QoS

Internet is best effort: no guarantees, no QoS; all traffic treated equally (net neutrality); simple, scalable, but no delivery/order assurance or prioritization

QoE & Route Pinning

- **QoE – Quality of Experience:** Perceived service quality from user perspective
- **Route Pinning:** Keeps flow on a fixed path to prevent oscillation (don't switch immediately to "better"path)

Network Performance Metrics

- **Latency / Delay [ms]:** Time for packets to travel src → dest (Voip < 150ms)
- **End-to-End Delay:** Total time sender to receiver
- **One-Way Delay:** From first bit sent to last bit received
- **Delay Components:** Transmission delay (time to push onto link), Processing delay (lookup, queuing), Propagation delay (physical travel time)
- **Jitter [ms]:** Variation in delay between packets, caused by re-routing/queuing (Voip<30ms), Calc: no queue - queued delay
- **Throughput:** Rate of successfully delivered data
- **Packet Loss [%]:** Dropped packets due to congestion or errors (Voip < 1%)
- **Bandwidth [Gbit/s]:** Maximum transfer capacity of a link

Queuing Algorithms

- **FIFO (First-In First-Out):** Basic, no prioritization
- **Priority Queuing (PQ):** Multiple queues, serve highest first; others may starve
- **Round-Robin:** One packet per queue in turn (fair, but ignores priority)
- **Weighted Fair Queuing (WFQ):** Round-Robin with weights, e.g., 2 packets from Q1, 4 from Q2
- **Class-Based WFQ (CBWFQ):** WFQ with user-defined classes, queue limits, max bandwidth guaranteed or max % of bandwidth (logical queues based on IP Precedence only)
- **Low Latency Queuing (LLQ):** Adds strict priority queue (priority class) to CBWFQ for delay-sensitive traffic (e.g. voice) (based on IP Precedence, DSCP, src, port, protocol...)

Queue Management

- **Tail Drop:** Drops packets when queue full; huge interruption of traffic → same as no connectivity
- **TCP Global Sync:** Many TCP flows back off and restart simultaniously → link underutilization
- **TCP Starvation:** TCP slows down after drops, UDP doesn't → queues filled with UDP, TCP squeezed out
- **RED:** Random early drops before full queue to prevent global sync and TCP collapse. Dropped TCP segments cause TCP sessions to reduce their windows sizes
- **WRED:** RED + DSCP/EXP-based drop logic, prioritizes higher-marked traffic
- **DSCP / EXP:** DSCP (6-bit in IP header) marks packets for QoS; used in DiffServ for classifying traffic. EXP (3-bit in MPLS label) serves same purpose within MPLS networks; often mapped from DSCP.

Policing vs. Shaping

- **Policing (Inbound mostly):** Drops packets that exceed configured rate limits
- **Shaping (Outbound):** Buffers packets to smooth traffic bursts and conform to profile

QoS Models

- **Best Effort:** No guarantees, all traffic treated equally (follows Internet neutrality)
- **Integrated Services (IntServ):** End-to-end QoS, per-flow resource reservation, precise but not scalable (uses RSVP)
- **Differentiated Services (DiffServ):** Class-based, scalable approach using marking (e.g.,), no hard guarantees

Traffic Marking

- **L3 Marking:** ToS byte → DSCP (6 bits) + IP Precedence (3 bits)
- **L2 Marking:** Dot1q header → 802.1p CoS bits

Modular QoS CLI (MQC)

- **Class Map:** Define traffic classes (e.g., match voice or video)
- **Policy Map:** Define actions for each class (e.g., limit, shape, priority)
- **Service Policy:** Apply policies to interfaces or directions (in/out)

CDN

- **Origin Server:** Central content source (original files), usually in a datacenter
- **Edge / CDN Server (POP - Point of Presence):** Geographically distributed, caches content
- **DNS Infrastructure:** Directs users to optimal edge server (e.g. via Geo-Routing)

Key Benefits

- **Latency Reduction:** Nearby edge servers reduce round-trip time
- **Availability:** Failover and redundancy in case of node failure
- **Scalability:** Handles traffic spikes via load balancing
- **Cost Optimization:** Reduces backend and transit load on origin
- **DDoS Protection:** Edge servers absorb attacks -> not all traffic on one server
- **Global Load Reduction:** Less long-distance traffic across the Internet

Request Routing Techniques

- Decides which edge server should serve a client request
- Goal: Best performance (e.g. proximity, load, responsiveness)

DNS-Based Geo-Routing

- Each edge has a unique IP
- DNS server picks closest/optimal edge server based on:
  - Resolver IP location (not user!)
  - GeoIP DBs (MaxMind, IP2Location), load, latency, business rules
- Limitation: DNS Resolver != user location -> can cause wrong choice

EDNS(0) and Client Subnet Extension (ECS)

- Resolver includes part of client IP in DNS request (e.g. /24 subnet)
- Authoritative DNS makes better decision based on actual client region
- Improves accuracy without revealing full IP

Anycast with BGP

- Same IP (e.g. 7.7.7.7) advertised from multiple locations
- BGP routing decides which path is "best" (AS-path, local pref, etc.)
- No DNS logic or per-client decision — pure BGP convergence
- **Pros:** Fast failover, simple, no app logic needed
- **Cons:** Less control, BGP != best latency, route flapping risk

HTTP Caching & Headers

- Caching is controlled via HTTP headers between clients, proxies, and servers
- **Cache-Control:** Main directive (*no-cache, no-store, max-age, must-revalidate*, etc.)
- **Expires:** Absolute expiration time (older method, replaced by *Cache-Control*)
- **ETag:** Validator tag (version/hash), used with *If-None-Match*
- **Last-Modified:** Timestamp used with *If-Modified-Since* for revalidation
- **Age:** Time (in seconds) since response was fetched from origin
- **Validation:** Client uses *ETag* or *Last-Modified*; server returns 304 if unchanged

Other Infos

AD: Inter-protocol choice (e.g., OSPF vs RIP) → lower wins.
Cost/Metric: Intra-protocol choice (e.g., OSPF path A vs B) → lower wins.
Routing Preference Order (across protocols):

1. Most specific prefix
2. Lowest Administrative Distance
3. Static default route

Administrative Distances (Smallest Administrative Distance wins)

Protocol	Distance
Connected	0
Static (Interface)	1
Static (Next Hop)	1
BGP External	20
EIGRP Internal	90
OSPF	110
ISIS	115
RIP v1/v2	120
EIGRP External	170
BGP Internal	200

EVPN BGP Routing Table Infos

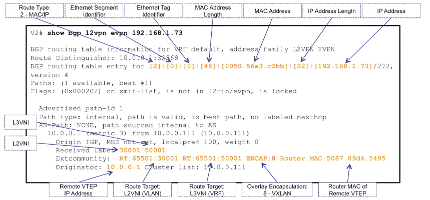
* = Would not be there if it was L2 VNI BGP Routing Table

- **Route Distinguisher:** 172.16.255.101:32777
- **Route Type:** 2
- **MAC Address Length:** 48
- **MAC Address:** 5254.00f8.29a8
- ***IP Address Length:** 32
- ***IP Address:** 10.10.0.100
- **L2 VNI:** 30010
- ***L3 VNI:** 50000
- **Remote VTEP IP Address:** 172.16.254.101
- **L2 Route Target:** 1:10
- ***L3 Route Target:** 65000:50000

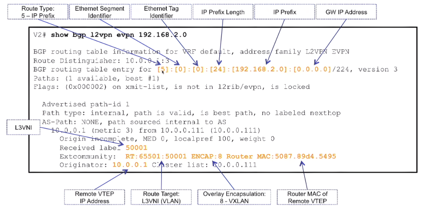
—

```
leaf-03# show bgp l2vpn evpn 10.10.0.100
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 172.16.255.101:32777
BGP routing table entry for
[2]:[0]:[0]:[48]:[5254.00f8.29a8]:[32]:[10.10.0.100]/272, version 19897
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn,
          is not in HW
Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
Imported to 2 destination(s)
AS-Path: NONE, path sourced internal to AS
172.16.254.101 (metric 81) from 172.16.255.1 (172.16.255.1)
Origin IGP, MED not set, localpref 100, weight 0
Received label 30010 50000
Extcommunity: RT:1:10 RT:65000:50000 ENCAP:8 Router MAC:5254.00ca.69ae
Originator: 172.16.255.101 Cluster list: 172.16.255.1
```

Route Type 2:



Route Type 5:



Prüfung Vorjahr

Network design

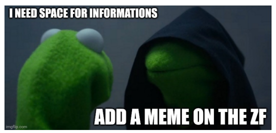
- **3-tier campus network**: Default Gateway (D), QoS marking (A), STP Root Port (A), HSRP, VRRP or GLBP (D), "Simple" (C), OSPF Totally Stub Area (D), High availability (C)
- **Campus Design**: used to reduce size of L2 domain: EVPN, MPLS

Rest

- **MP_REACH_NLRI:** Next hop, MAC Address
- VXLAN is a data plane technology which encapsulates Ethernet frames in UDP datagrams to tunnel layer 2 frames over a layer 3 network.
- The underlay network is unaware of VXLAN devices that connect to the physical switches are unaware of VXLAN.
- A route distinguisher is used to uniquely identify a route in combination with the destination prefix.

El Memez

Always space for memez ;)



Calc