Network-Layer

IPv4

Network Adress Translation (NAT): Used to translate private IP addr. to global allowing for specific IP addr. to be reused (priv. Network) NAT runs on a device connected to two networks usually an internal network connected by a router to the Internet
Static NAT: Mapping unregistered (private) IP addr. to a registered on a static one-to-one basis. Useful when device needs to be accessible from outside the network
Dynamic NAT: Maps unregistered IP to registered IP from a group of registered IPs dynamically, meaning that public IP of device can change.
Overloading NAT, PAT (Port & Address Translation): form of Dynamic NAT that maps Multiple private IPs to a single public IP by using different Ports. Also known as PAT. A packet from a host leaves the network with the public IP of the router and a dynamically assigned Port Nr. Router holds Record of mappings in NAT-Table. When Receiving a packet the mapping is reversed.
Overlapping NAT: Used if two hosts share the same IP address. Maintains a Lookup Table of duplicate addresses. Intercepts traffic and replaces these IPs with a unique registered IP
Inside Local, outside Local, inside Global, outside Global

IPv4

Classful Unicast Networks:

Class	Address	Subnet	Range	
A	0.0.0.0	/8	10.0.0.0/8 – 10.255.255.255/8	1 Class A
B	128.0.0.0	/16	172.16.0.0/12 – 172.31.255.255/12	16 Class B
C	192.0.0.0	/24	192.168.0.0 – 192.168.255.255	256 Class C
D (multicast)	224.0.0.0	N/A	Not routed; implemented because of IPv4	
E (reserved)	240.0.0.0	N/A	Addr. shortage. Need to be replaced by public	

Private Networks (RFC 1918)

Special Reserved Addresses:
127.0.0.1/8 – Loopback, Class A, Address for communicatin with one-self
169.254.0.0/16 – Link local addr., Class B, Broadcast Domain wide network

IPv4 Header: Fields:

Version: decimal value 4 = 0100b; **IHL:** Internet Header Length, tells Length of header in 4 Byte increments; Min 20 Bytes, at least 5 (5·4)=20, max 15 → max 60 Bytes (15·4=60). **Type of service:** DS field (Differentiated Services), used for Quality of Service (QoS), defines packet priority; **Total Length:** Entire size of packet (header + data) in Bytes, min 20B max 65535B; **Identification:** contains fragmentation ID, shows to which packet the fragment belongs to; **Flags:** First bit always 0, Second bit called 'Dont fragment' bit if set, packet sent without fragmenting it, can be rejected because of this. **Fragment Offset:** Specifies Byte position of the fragmentation, 65528 (2^13 − 1); **TTL:** when packet passes through router TTL is decremented by 1; if TTL reaches 0, packet gets dropped; **Protocol:** Contais a value which stands for the used transportation Protocol (Most common: UDP, TCP, OSPF, ICMP); **Options:** used for network testing, debugging, security ect. field is usually empty; **Padding:** Because the Options field is variable in size the padding field is used to bring the Header to multiple of 32 bits.

Version	IHL	Type of Service	Total Length
Identification		Flags	Fragment Offset
Time to Live	Protocol	Header Checksum	
Source Address			
Destination Address			
Options ←|→ Padding			

IPv6

IPv6 Header: Fields:

Version: 4 bit inter Protocol version nr=6; **Traffic Class:** 8 bit traffic Class. Differentiated Services (Quality of Service QoS) comparable to TOS in IPv4.
Flow Label: 20 bit, Each source chooses its own flow label values; routers use Source Address and Flow Label to identify flows. Flow label value = 0 = no special QoS is requested (common case today).
Payload Length: 16 bit unsigned integer. Rest of the packet following IPv6 header in octets, but without 40 Byte header in calculation comparable to Total Length in IPv4; Max Payload 64 Kbytes;
Next header: 8 bit selector: identifies type of header immediately following the IPv6 Header. comparable to Protocol field in IPv4 and uses the same values as the IPv4 Protocol field [RFC 1700] is either is either the Upper Layer (TCP/UDP layer 4) or the Extension Header
Hop Limit: 8 bit unsigned integer. Decides Lifetime of package. If TTL=0, ICMP Packet is sent to sender. comparable to TTL in IPv4.
Source Addr: 128 bit of originator of the packet. **Destination Addr:** 128 bit Addr. of the intended recipient of packet. (possibly not the ultimate recipient if a routing header is present.)
Extension Headers: 0: Hop-by-Hop option; 43: Routing, 44: Fragment, 50: Encapsulating Security Payload, 51: Authentication Header (AH), 59: no next Header, 60: Destination option, 62: Mobility Header.
Extension headers important for the router come at the beginning

Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
Source Address			
Destination Address			

IPv6 Changes:

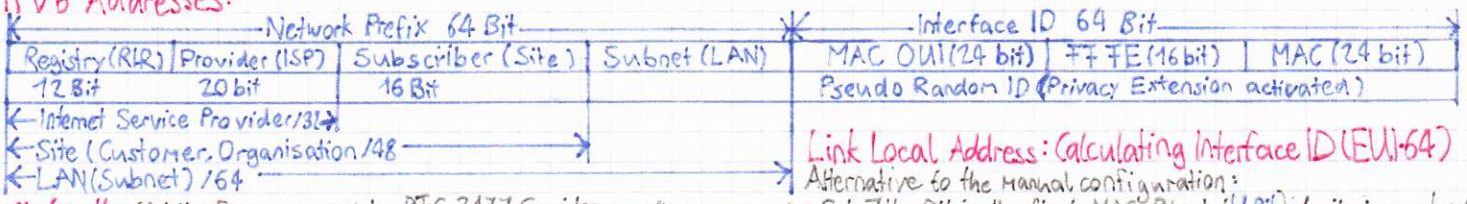
- 40 Bytes IPv6 Header. Bigger than IPv4 Header (20–60 Bytes). But strongly simplified on the contrary to the IPv4 Header
- Enhancement of the Addr. range from IPv4 with 2^32 (4,3·10^9) to IPv6 with 2^128 (3.4·10^38) adresses. Growth by the factor 2^96
- Address: 8 Blocks with 4 Hexadecimal digits. Each block is 16 Bit. Total of 128 bit
- Simplification of the header information, this decreases the burden on the router
  - → Removed Fields: Fragment Offset (Segmentation was removed), Header Checksum (Layer 3 Errors nearly impossible), Traffic Class instead of Type of Service, Next Header Field instead of Protocol Field (Specifies which protocol is used on higher Layer)
- Segmentation was removed. Only possible with extension headers. Oversized MTUs get Discarded and Error Message is sent
- Stateless automatic configuration of IPv6 Addresses. In many cases DHCP is obsolete
- Usage for mobile Devices easier (Mobile IP)
- Encryption and Authentication check built in. (Implementation of IPSec as a part of IPv6)
- Support of Quality of Service and Multicast
- No NAT needed anymore because there are more than enough IP addresses

Mechanism to prevent fragmentation: MTU Path Discovery

IPv6

Network Layer

IPv6 Addresses:

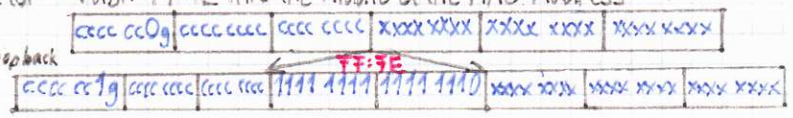


Prefix length: 64 bits: Recommended by RFC 3177. Consistency makes management easy. Must for SLAAC/MSFT DHCPv6; More than 64 bits: Enables more host per broadcast domain. Considered bad practice. 64 bits offers more space for hosts than the media can support efficiently.; Less than 64 bits: Address space conservation. Special Cases: /126 /127 valid for p2p (RFC6164) /128 loopback complicates Address Management

Link Local Address: Calculating Interface ID (EUI-64)

Alternative to the manual configuration:

- Set 7th Bit in the first MAC-Block (U Bit): to its inverted value
- Push FF:FE into the middle of the MAC-Address



Address-types: IPv6 Interfaces have multiple IP addresses — 1 or many global Unicast Adresses and 1 link Local Address.

Unicast:

Link-local:	FE80::/10	Inside of closed Network, don't get routed, Pendant to MAC-Address, For Next Hop-Calculation in routing, Cannot Communicate with Internet
Unique Local:	FC00::/7-FDFF/7	Only in local network, Only within 1 Routing Domain Routed
Loopback	::1/128	
Global Unicast:	2000::/3	Assigned by IANA, e.g. for ARIN, RIPE; Components in Graphic above; Host-Address Assigned with EUI 64, Pseudo Random, DHCP, Manually; Configuration: EUI 64, SLAAC + DHCP (for e.g. DNS) Stateful Autoconfig or DHCP

Multicast: FF00::/8

Addressing of all Interfaces that belong to a Multicast group. Assigned by IANA. First 32 bit relevant. The FF-Prefix follow 4 bits for Flags and 4 bits for the Scope: Node local (1), Link Local (2), Site Local (5), Organisation local (8), Global (8)

Node local	FF01:	Packet never leave Interface, Loopback
Link Local	FF02:	Don't get Routed
	FF02::1	All nodes/hosts on a segment
	FF02::2	All routers on a segment
FF02::5/	FF02::6	OSPFv3
	FF02::9	RIP (v2)
	FF02::A	EIGRP
	FF02::D	All Protocol-Independent-Multicast (PIM) routers
	FF02::1:2	DHCP Relay Agents / DHCP-Server

Solicited-Node Multicast

FF02::1:FF... (... lowest 24 bits of Link local Adr.)

For each configurated Unicast and Multicast address a repeating Solicited-Node-Multicast Address exists. Is a Multicast Address that is on a IPv6 Address of a interface. Usage: Replacement of ARP, Duplicate Address Detection (DAD), Neighbor Detection Protocol (NDP)
Composed of Prefix + lower 24 bits from Unicast Link Local Adr.
Example: Link Local is FE80::200:CFF:AA3A:8B18 → Solicited-Node Multicast is FF02::1:FF3A:8B18
An IPv6 Multicast FF02 gets converted to 33:33

Anycast:

Address that can be assigned to more than one interface.. Multiple Devices can have the same Anycast Address.
Usage: Load Balancing, DNS Rootserver, Content Delivery Network

Special:

2002::	Adresses for 6to4 Tunneling mechanisms
2001:db8::/32	For documentation
::	Unspecified, used as a placeholder

IPv4 Pendant

169.254.0.0/16 private/local Addresses RFC 1918
Loopback address 127.0.0.1
Public IPv4 Address

Subnet Broadcast-Address Not existing
224.0.0.5, 224.0.0.6 224.0.0.9

No Multicast Address

Protocols ICMPv6 and Neighbor-Discovery-Protocol:

Neighbor Discovery: Nodes can find their respective neighbors; Finding out Link-Layer-Address of neighbors; finding Routers; Maintainig Neighbor Reachability Information (NUD) **Router Discovery:** With Router Discovery the default GW can be found
Replacement of ARP

ICMP Package Groups: Group 1: (Messages between Router and IPv6 Devices): ICMP Type Field 133 = Router Solicitation (RS) Host requests advertisements by sending an RS to all-Router Multicast Address FF02:2 Sent automatically when station starts up.
ICMP Type Field 134 = Router Advertisements (RA): Routers advertise themselves in the network and send information for IP-Autoconfiguration. Routers periodically send these packages to the All-Nodes Multicast Address FF02:1.
Group 2: (Messages between IPv6-Devices): ICMP Type Field 135 = Neighbor Solicitation (NS): Host searches for all Neighboring Hosts and their MAC-Address by sending a packet to the All-nodes Multicast Address FF02:1. In IPv4 the Broadcast address was used. Can also be used to discover Duplicate Address with Duplicate Address Detection (DAD)
ICMP Type Field 136 = Neighbor Solicitation Advertisement (NSA): Is the response to the NS and contains MAC Addr. of the Host.
Group 3: (Redirects): Redirect: Is used to see whether the packets could have a better first Hop. Can occur if there are multiple routers in the network.

IPv6

Complementary

Internet Communication: Link Local Address: First address that IPv6 Device has to configure. Prefix FE80:: and Host Part Random (Privacy Extension) or EUI-64. Check if address in use with DAD by sending Neighbor Solicitation request.

Global Unicast Address: Routable IP Addr. required to reach Internet. To obtain sends Router Solicitation request from link local Addr. to mcast FF02::2. Stateless Address Auto Configuration (SLAAC): Used to generate IP Addresses. After Router Solicitation all Routers respond with a Router Advertisement (RA) which contain valid IPv6 Global Unicast Address and Flags. For each RA with A-Flag host generates IPv6 Addr (Process called SLAAC); Information for DNS can be queried from DHCPv6 Server when RA has O-Flag set or otherwise usage of IPv6 RA options for DNS Conf (less common); M-Flag (Managed bit) indicates that SLAAC is not used → Global Unicast Address queried from DHCPv6 stateful); L-Flag = IPv6 Prefix is "On-Link" = All Hosts in same subnet / LAN. If L-Flag is not set all traffic sent over Router; Prevents Neighbor Discovery Spoofing.

Extension Headers: Next Header-chain, Last Next Header indicate upper layer.

Routing Extension Header: List of routers a package needs to pass. type 0: IPs listed in EH, copied into Header by each Router. type 2: used in mobile IP.

Fragment Extension Header: If package > MTU: Min MTU = 1280 bytes. If needed to fragment EH added with Fields: Offset (13 bit); More (1 bit); Id (32 bit).

Prefix Delegation: DHCP for Network Prefixes. DHCP PD can manage Prefixes for ISP.

Mobile IP: Needed for VOIP & VPN, etc.; Mobile Client registers new IP with Home Agent who relays packets to client because they are still sent to client's Home Address. Client can later make new binding with server.

Mcast → RP Address: start with FF7: NetPrefix [:plen] :: RPAdr (1)

Migration from IPv4 to IPv6: Best Approach IPv6 only; how difficult

Dual Stack: Support both: Most Common but most complex requirements. ⊕ Conceptually Simple ⊖ Each Device must support both (cost) ⊖ 2x Routing protocols ⇒ 2x resources & config ⊖ No cross communication

Tunneling: Bypass Legacy Infrastructure to bridge IPv6 only Backbone like DSL / Mobile comm. networks: ⊕ Cheap (No investments) ⊖ Complicated & Security

Translation: Similar to NAT but protocol changed by capable Device (Router) ⊕ Transparent ⊖ Simple ⊕ Cheap ⊖ Leads to problems w. end-to-end Appl (e.g. VoIP)

Rollout: 1. IPv6 Readiness audit (network & Apps) 2. Get ISP Independent address space → /48 per DC preferably large block (BGP traffic engineering) 3. Get IP connectivity for all your ISPs 4. Pilot IPv6 (in non-critical part of DMZ) 5. Enterprise-wide rollout (takes some time)

Roadblocks (to be expected): Network HW (Switches) might lack some features (RA Guard, ND Inspection, etc.); DC equipment might be Problem! Applications will be Nightmare; IPv4 only Clients

Solutions (some at least): IPv6 connectivity to Internet; Slim layer of IPv6 in DMZ; SSL VPN over IPv6

IPv6 Readiness: Common Issues in Application that do not support IPv6: server does not listen to IPv6 Addr.; Server application does not work with IPv6; Often Resolved through tunneling strategies NAT64 or IPv6-to-IPv4; has own drawback: Client-address based access control (blacklisting)

Security

Many Security issues of IPv4 + some more; Good Practice: reduce Layer 2 networks as much as possible e.g. by setting up p2p connections to the first-hop Router / layer 3 Switch or by setting L-Flag in RAs to 0;

Duplicate Address Detection Denial: responding to every DAD → No Available Addresses to devices

Neighbor Advertisement Spoofing: Similar to ARP-Spoofing in IPv4; Attacker responds to ICMPv6 Neighbor Solicitation "Who uses address X", pretending to be X → Traffic interception & Denial of Service; No easy Solution

Router Advertisement Spoofing: Respond to Router Solicitation with highest Priority effectively redirecting all traffic through Attacker; Configure Hostports / RA Guard

DHCPv6 Reply Attack: Respond to DHCP Discovery Requests, can configure clients to e.g. use attacker-controlled DNS Server. Denial-of-Service by sending invalid IPv6 Addr. Solution: Block DHCPv6 response port on all interfaces which are not DHCP Interfaces

Neighbor Discovery Cache Flooding: Attack not feasible in most IPv4 Netw. bc lack of addresses. With IPv6 No lack of Addresses. Attacker pretends to be many devices & fills up ND Table (Flooding) Solution: Per interface ND Table reduces impact

Dual Stack Exposure: two attack vectors; avoid existing barriers with other protocol

Tunneling Related Exposure: Tunnels render Firewall useless; Solution: block tunnels

User Authentication & Logging: Impossible with SLAAC → Use Stateful DHCP and M-Flag

Layer 2 Defense Scenario: Layer 2 as Bcast Medium → Host-Routing

IP Sec (defined as Extension Header):

Guarantees: Confidentiality: With ESP payload gets encrypted

Integrity: Hash transmitted as Integrity Check Value (ICV)

Authentication: Sender and Receiver authenticated by Authentication Header

Anti-replay: Sequence Nr. makes replay attacks impossible

Authentication Header (AH): Offers authentication & Integrity: ICV hash includes Src & Dst → Traversing NATs not possible; ICV hash excludes TTL & Header Checksum

Transport Mode: AH header between IP Header & ICMP/TCP/UDP Header

Fields: Next Header, Length, Security Parameter Index (SPI) (32 bit, identifies flow) Sequence (prevents Replay Attacks), Integrity Check Value (ICV)

[IP Header | AH Header | TCP Header | Data]

← Authenticated →

Tunnel Mode: Adds New IP Header; used for privately addressed IP Packets if Tunneled over Internet

[New IP Header | AH Header | IP Header | TCP Header | Data]

← Authenticated →

Encapsulation Security Payload (ESP): Encrypts IP traffic; Must for VPN connecting sites

Transport mode: Original IP Header used with inserted ESP → IP Header Cleartext

[IP Header | ESP Header | TCP Header | Data | ESP Trailer | ESP Auth Trailer]

← Authenticated → ← encrypted →

Tunnel Mode: New IP Header used → Original encrypted. Useful for site-to-site VPN

[New IP Header | ESP Header | IP Header | TCP Header | Data | ESP Trailer | ESP Auth Trailer]

← Authenticated → ← encrypted →

AH & ESP Combination:

Transport Mode:

[IP Header | AH Header | ESP Header | TCP Header | Data | ESP Trailer | ESP Auth Trailer]

← authenticated → ← encrypted →

Tunnel Mode:

[New IP Header | AH Header | ESP Header | IP Header | TCP Header | Data | ESP Trailer | ESP Auth. Trail]

← authenticated → ← encrypted →

IPSec Tunnel: Internet Key Exchange (IKE): Establishes Security Association (SA) bt. Peers

IKEv1 Phase 1: (meta data tunnel) 1) Negotiate Hash Algorithm; Authentication; Diffie-Hellman group; TTL; Encap Algorithm 2) Diffie Hellman Key Exchange 3) Authenticate peer (PSK or Cert); Modes: aggressive mode (3 msg) (id not authenticated) cleartext Main Mode (slower) (6 msg) (id encrypted

Phase 2: Param Negotiation through Phase 1 Tunnel; Neg. Protocol (AH/ESP); modes; Enc. Alg.; TTL; DH (optional, used for perfect forward secrecy (PFS)

IKEv2 Only 1 mode for Phase 1 & No quick mode for phase 2; Total of 4 msg.

Improvements: less Bandwidth required; Support EAP Auth; Support for NAT built in Keep-alive for Tunnels

IP Multicast

Overview

Unicast: Src sends N unicast Datagrams, one per receiver (10 clients = 10 unicast packets → Source requires a lot of HW power)

Broadcast: Each Subnet has one bcast addr. (last IP of Subnet); Source sends one bcast per subnet. Each client in subnet receives bcast, also if not interested.

Multicast: Source sends mcast only once to a mcast addr. Routers duplicate the packet and forward them to next Routers/receivers. The not interested do not receive the packet. to specify "interest" and receive the packets clients have to join mcast group

Applications and Services: Mcast used for one-to-many or many-to-many communication

One-to-many: Music-on-hold-services, Sensor updates, SW distribution, Radio, IPTV

Many-to-many: Stock exchanges, Multiplayer Games, Group chat applications

Benefits: Enhanced Efficiency (Controls network traffic and reduces Server/CPU loads); Optimized Performance (Eliminates traffic Redundancy); Distributed applications (Multipoint)

UDP Based: Best effort delivery, no congestion avoidance, Some mcast protocol mechanisms result in occasional generation of duplicates, out-of-order delivery

Concept: Mcast source: sends message to mcast group IP (doesn't need to be member)

Mcast receiver: any mcast-enabled device that joins a mcast group.

Multicast Layer 2: Internet Group Management Protocol (IGMP)

IGMP is the protocol used to manage group subscription for IPv4 mcast

IGMP Membership Report: Msg that a host sends to join a Mcast Group- Sent to 224.0.0.22 (all Routers) which then builds a IGMP group membership table.

IGMPv1: Has No mechanism to leave group, if host has closed mcast app Router doesn't know and continues sending. Host quietly leaves group.

IGMP General Query Message: Used to detect host that left group by Router. Sent 3 times every 60s. Host supposed to refresh group interest by sending Membership report as response. If 3 times (3min) nothing received → Group times out.

IGMPv2: Can Send IGMP leave-group message: Sent to Router if mcast 224.0.0.2 to indicate leave. Router responds with a Message: **IGMP Group Specific Query Message**: To make sure, nobody else is listening to the mcast stream attached to same Router, Expects Membership report, if no report is received by Router within 3s the Group times out.

IGMPv3: Receiver not only specifies group but also which source the streams should Include or Exclude with a list → No Rendezvous Points (RP) needed anymore

IGMPv3 enabled Routers join group 224.0.0.22 to which Hosts send member reports With IGMPv3 of L2, PIM SSM can be used as an IP Multicast Protocol

IGMP Snooping: Without Snooping Switches handle Mcast like Bcast → repl. icate packet to all ports except incoming. With Snooping Switches recognize & eval. Membership Reports & Group Leaves and maintain a IGMP Snooping Table according to which they distribute the Mcast packets. **Disabling IGMP Snooping**: If Snooping is enabled but over Mcast communicating devices are not using Layer 3 Mcast or IGMP all their Mcast traffic is dropped bc. they do not send IGMP Joins

Protocol Independent Multicast (PIM)

PIM: L3 Protocol for Mcast, Requires IGMP Running in L2 & IGMPv3 for Source Specific Multicast

PIM is Protocol Independent because works with any underlying unicast Routing Protocol; Uses Routing Table to send join toward Src/RP; Unicast for RPF-Checks

PIM Dense Mode: Labor use, Uses Push-Model: Traffic flooded through network: (S,G) State created in every router in the network

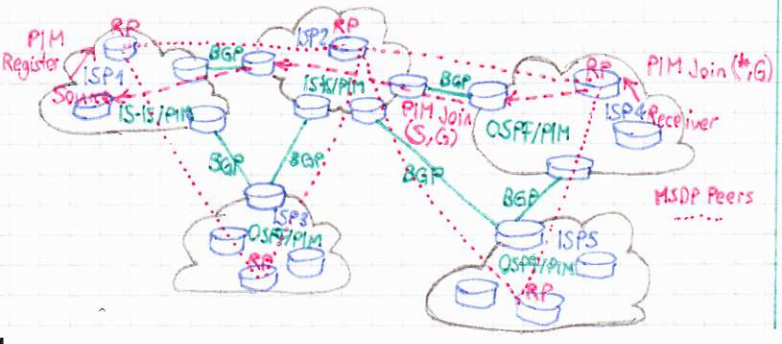
Pruned back where unwanted: Router sends Prune Messages: When no directly connected hosts receive the mcast traffic, When no downstream routers interested in receiving the mcast traffic; When traffic received on a non-RPF Interface

After prune process mcast stream is optimized and doesn't flood network anymore; Flood & Prune Process repeats every 3m; (S,G) State still everywhere

Multicast over the Internet

Idea: ISP1 (e.g Swisscom) has a Source and ISP4 (e.g Cablecom) has a Receiver Every network has its own routing protocols, doesn't matter which. Every network own RP For this reserved IP address Range 233.0.0.0/8. Globally Scoped, Statically assigned. 16 bit ASN mapped to 233.X.Y.0/24 (e.g. AS 3303 mapped to 233.12.231.0/24).

Multicast Source Discovery Protocol (MSDP): Protocol used to connect RPs in different Autonomous Systems and allow inter-area mcast to work.



Multicast Addressing

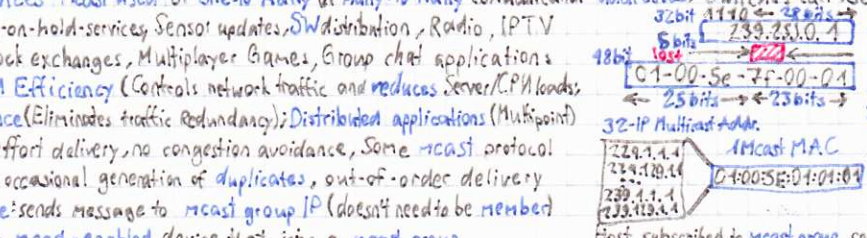
Reserved Multicast Addresses: Link-local addresses: 224.0.0.0 - 224.0.0.255 Transmitted with TTL = 1. Examples: .1 (All Hosts). 2 (All Routers). 13 (PIMv2 Routers). 22 IGMP

IANA Reserved Addresses: 224.0.1.0 - 224.0.1.255; Example: .1 Network Time Protocol

Source Specific Multicast Addresses: 232.0.0.0 - 232.255.255.255 SSM range

Administratively scoped Addresses: 239.0.0.0 - 239.255.255.255 ← Most; Private Address range, Similar to RFC 1918 unicast address. For private Networks

Multicast MAC Address Mapping: L3 Mcast addresses get translated to L2 addresses. Switches can use mcast MAC to distinguish frames & forward efficiently

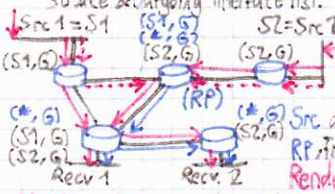


Low-order 23 bits of IP mcast addr are mapped to low-order 23 bits of MAC mcast addr. First 4 bits of mcast IP always 1110b → 28 individual bits left → 23 bits mapped → 5 bits lost → $2^5 = 32$ Mcast IPs mapped to 1 Mcast MAC. Example: Any Mcast IP 1110 xxxx x000 0001 0000 0001 0000 0001 gets mapped to same Mcast MAC 01:00:5E:01:01:01;

Host subscribed to mcast group could act Data they are "not interested"; Rare in reality

Multicast Layer 3

When Router receives IGMP Request, he builds a mcast routing table (mroute), so he knows where to forward, mroute contains receiver, source & Outgoing Interface list. Two different types of trees:



Source Tree (S,G): S = Source; G = Group

Shared Tree (*,G): * = All Sources, G = Group Used when Source unknown (IGMP v1/v2)

Src doesn't know Receiver → Packet follows (S,G)-table to the RP, from there packet are sent by RP with (*,G) to receivers

Rendezvous Point (RP): where Src Stream & Request of Recv. meet.

Multicast Forwarding: Mcast routing builds mcast-tree in direction of source (backwards from Unicast Routing); Mcast concerned where packet is from, Unicast concerned where packet is going; Mcast routing uses RPF for loop-prevention

Reverse Path Forwarding (RPF): Router forwards mcast packets only if received on RPF Interface. Mcast Src addr checked against unicast Routing table. E.g. a packet is sent to a mcast dst. IP Has a Src which is routed in routing table → has to be incoming on the Interface which the Src is routed, else dropped

Determines Interface & upstream Router in the direction of the source in which PIM joins are sent. Interface becomes "Incoming" or RPF-Interface.

Tree built hop-by-hop from Receiver to Source (or RP) with join-messages following the unicast routing table. Mcast data flows down the tree.

Equal-Cost Paths: Can't use both; **Tie-breaker**: Use highest Next-hop IP Addr.

PIM Sparse Mode

Productive environment: Uses Pull-Model → Traffic only sent to where is requested: Explicit Join behaviour

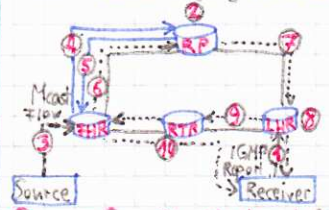
Basic mechanism: Routers receive explicit join/prune messages from neighboring routers that have downstream group members; Router forwards data addressed to a mcast group (G) only onto interfaces where explicit joins have been received

Any Source Multicast (ASM): IGMPv1/2 (Host only asks for mcast group)

RP for source discovery: Receivers "joined" to Shared Tree (Rooted at the RP)

Steps of ASM:

1. Receiver sends IGMP Membership Report to the last-hop-router (LHR). LHR reviews its config for RP info for group and sends request to RP with PIM control Protocol
2. (*,G) update used to provide the RP info of an active receiver for the mcast group
3. Source sends mcast stream to the first-hop-router (FHR) 4. FHR, configured with PIM Sparse Mode & RP info, encapsulates mcast stream in unicast Register Message to RP
5. If already Receiver exists, RP sends (S,G) PIM Join to FHR and Register Stop Msg. to FHR.
6. After receiving Register Stop, FHR sends mcast stream to RP as (S,G) stream 7. Based on interested receiver outgoing interface list populated, (S,G) flow sent to LHR via PIM Interfaces



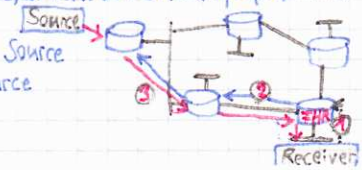
8. LHR checks unicast routing table to reach Source
9. If check verifies alternate path is better based on unicast protocol, (S,G) Join is sent from LHR to Source
10. Source sends mcast flow to LHR via new (S,G) state and is created in mcast routing table. When data flow migrated to Shortest Path tree (S,G), the mcast LHR sends RP Prune Message to the RP

Source Specific Multicast (SSM)

SSM: IGMPv3 (Host asks for mcast group and source) Host for source discovery; No RP; Reserved IPv4 address range of 232.0.0.0/8 for PIM SSM leverages the unicast routing topology to maintain the loop-free behaviour

Steps of SSM:

1. Receiver sends an (S,G) IGMPv3 membership report to the first-hop Router
2. first-hop Router sends (S,G) join to Source
3. Traffic flows on shortest path from Source to the Receiver



Network-Layer

Link State Routing

Link State Routing (OSPF, IS-IS): Each router must know the whole network topology; each router creates his own topology-DB (Map) in which all LSAs he received from other routers are saved; ⊕ Link state protocols are more reliable, easier to debug & use less Bandwidth than Distance Vector Protocols; ⊖ They require more memory space & more Processing Power. They burden the network more because of LSA flooding; in order to tell all nodes about the network topology & to find the best path, 3 steps are required: 1. Hello Protocol, 2. LSAs, 3. Path calculation (shortest Path first algorithm)

Hello Protocol: In the first step the router finds out his neighbors & exchanges the info until their DBs are identical. Keep-alive Messages (every 10s) ensure that a neighbor is still up. If neighbor doesn't answer for 30-40s its marked as down. Negotiates parameters such as the election of a designated router.

IS sent as IP Multicast: 224.0.0.5 (point-to-point, point-to multipoint) or Unicast packets in Non-broadcast mode.

Link State Advertisments (LSA): Mechanism to distribute knowledge gained from the Hello Protocol to all other routers. Sends updates to other nodes and assures that only the youngest info is accepted.

- Whole network flooded with the logical sight of the routers (router ID, neighbor ID, cost)
- Each topology change invokes LSAs to inform the whole network of the topology change.
- When all routers have synced their DB, flooding is stopped

↳ if a router gets started he needs to know its last sequence number, that is needed to ensure network security, for this he asks neighboring routers for his last sequence Nr.

↳ With the sequence Nr. a router can recognize whether an info that he has received is up to date or if there is more recent info

↳ If the Maximal sequence number is reached, the router starts at 0. This would mean that preceeding updates are 'older'. But the Max is 2^{32} • updateInterval(10s) = 1361y (30Min)

Open Shortest Path First (OSPF): fast convergence time, scalable; OSPFv2 for IPv4, OSPFv3 for IPv6/4; Uses Hello protocol for neighborhood, LSA, Dijkstra (building neighborhood with hello, LSA exchange via LSA flooding, building Link-state-DB according to LSAs, shortest path first execution, building the routing table); Classless routing with VLSM, sends SubnetMask with updates; authentication with MD5, Cleartext, SHA; When OSPF recognizes a topology change it distributes a complete update with all Link-state infos; All routers of an OSPF-Area need to have identical & synced Linkstate info in their DBs. Selective updates: If the OSPF network is converged and a new link is up, only this is sent as an update to the neighbors; Metric: Cumulative Cost of all outgoing interfaces from source to destination, Bandwidth, Delay

Uses Multicast and Unicast instead of Broadcasts.

IPv4: Multicast 224.0.0.5 to all OSPF routers, Multicast 224.0.0.6 to DR/BDR Router. MAC: 01-00-5e-00-00-05

IPv6: Multicast FF02::5 to all OSPFv3 routers, FF02::6 to all DR/BDR Router;

Bundles Network into Areas: smaller logical subnets. This holds the networks stable (From 5000 routers unstable) → faster convergence, smaller routing tables, SPF-Algorithm is faster.

2 Routing types:

- Inter-Area Routing: each Inter-Area-Traffic has to go through Area 0
- Intra-Area Routing: Connects enduser and ressources

Area Border Router (ABR): ABR-Routers have at least one interface connected to Area 0. If this is physically impossible it can be done Logically with a OSPF virtual Link.

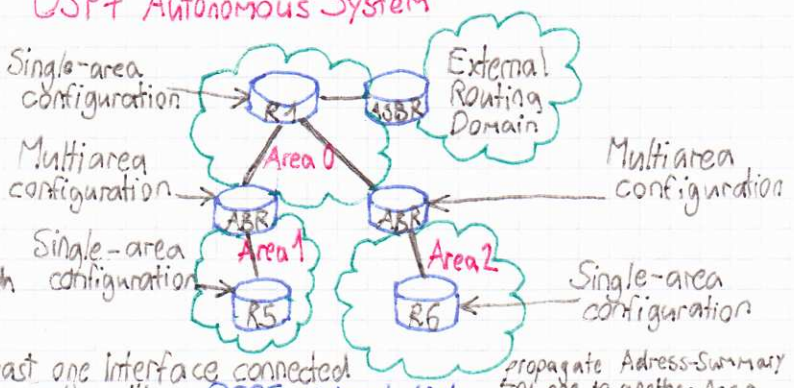
Autonomous system Boundary Router (ASBR): ASBR Routers have one Interface in the OSPF Domain and another Interface in any other Routing Protocol Domain (RIP, EIGRP).

Shortest-Path-First-Algorithm (Dijkstra):
After all tables were shared over the network, the cost can be calculated. This is done with the Dijkstra Algorithm. A tree with minimal length is calculated. The aim is to find the shortest possible path for each router. A table is used with three columns: Candidates, Cost from the Candidates to the Root, Nodes within the tree

1. → First the Router itself into the column "Nodes within tree" with the cost of 0.
2. → All direct neighbors, of the router that was added to the tree most recently, are added to the "Candidate List" column from the Link-state-DB in the form of a triple: <ROUTER_ID>, <NEIGHBOR_ID>, <COST>. This describes the link of last added router to his direct neighbor with the cost of the direction.
3. → The Candidate List is now moved into the column "Cost from Root to Candidate" but with the cost from the Root (Initial) Router to the newly added Candidate. Now Duplicates get checked and all but the way with the lowest cost get removed. The Candidate with the lowest cost from root now gets added into the Tree
4. → All remaining Candidates in the "Cost from Root" column now get moved into the "Candidate" column
5. → The Iteration starts back at Step 2. The most recently added is now the Candidate that was added to the Tree in Step 3.
6. → When only one Candidate is left, this one is also added to Tree

The Algorithm is invoked each Time the Topology changes.

OSPF Autonomous System

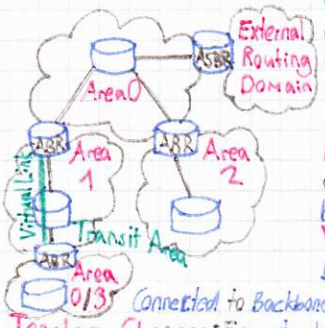


Open Shortest Path First (OSPF)

Overview

Link State Protocol; Designed for intra-domain routing; Used Version for IPv4 = 2; for IPv6 Version = 3; Operation: Establish neighbor adjacencies & LSA Exchange; Build Link-State DB; Execute SPF (Dijkstra Algorithm); Build Routing table
Encapsulation: On Transport Layer (L4); encap by L3 → IP with Protocol Nr. 89;
Router ID: OSPF Router ID is a 32bit value expressed as IPv4 Address; At least one IP on an interface in up/up state must be configured to be choosable;
Choosing Process: Happens at OSPF Initialization; chosen from following source in this order: 1. Use Router-ID specified in "router-id <address>" command; 2. Use highest IPv4 Addr. of all active loopback interfaces; 3. Use highest IPv4 Addr. among all active non-loopback interfaces; Change of OSPF Router ID requires OSPF Reset with "clear ip ospf" or Router reload
Packages: Use mcast 224.0.0.5; Reliability without TCP with
Implicit ACK (Duplicate of received LSA sent back to sender of ACK)
Explicit ACK (Receiver send LSACk when receiving LSU as response);
Retransmission List: Router tracks state of sent LSA in Retransmission List, from list periodic retransmission of LSU as Unicast until Acknowledged or adjacency gone;

Hierarchical Structure



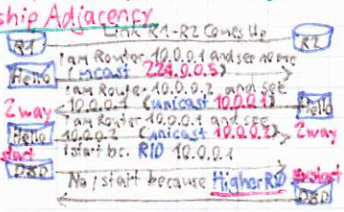
OSFF has functionality to divide intra-domain into sub-domains called OSPF Areas. Identified by 32bit Area ID
Backbone Area: Connects non-backbone Areas; not intended for use by client devices; Rules: Always Area 0; must be contiguous; All Non-backbone areas must be connected;
Non-Backbone Area: Usually set up according functional or geographic groupings. Any Inter-Area traffic must go through backbone; Indirect Connections to Backbone possible with:
Virtual Links: Used when unpractical / impossible to have Backbone contiguous or Non-backbone area cannot be Connected to Backbone directly; Only through one Area; Only through classical Areas
Topology Changes: Every topology change leads to SPF recalculation → Area below 50 Routers
Area Border Router (ABR): At least one Interface connected to Area 0 (or Virtual Link); Have to maintain separate OSPF DB for each connected Area
Autonomous System Border Router (ASBR): One Interface in OSPF Domain & other Routing Protocol dom.
Internal Router: Router only in Non-Backbone Area Backbone Router: Router in Area 0 (not exclusive)

Hello Packet: discover, build, maintain neighborship adjacencies, DR/BDR election; negotiation of capabilities with options field.
[Net Mask, Hello Interval, options, priority, Dead Interval, DR, BDR, neighbors]
Database Description (DBD) Packet: Describes LSDB; Used for Comparison between Routers → ensure sync; Only LSA Summaries / headers
[MTU, Options, DD Sequence Nr, LSA Headers] Options: E-bit: AS-external-LSA processing; I-bit: Initial; M-bit: !last; MS: Master(1) / Slave(0)
Link State Request (LSR) Packet: Request Missing LSAs
[LSType, LSID, Advertising Router), ...]
Link State Update (LSU) Packet: For flooding LSAs & Response to LSR
[Nr of LSAs, [LSA], ...]
Link State Advertisement (LSA): Flood every 30min; 60min → Link Dead
[LSAge, LSType, link State ID, Advertising Router, LS Seq. nr, Checksum, Len, ...] (sent to DR/BDR with 224.0.0.6)

Timers: Timers must match for entire segment: Hello Interval: Freaency of Hello Packets (Cisco default = 10s); Dead Interval: Before taking down neighborship adjacency (Auto: 4x Hello)
Designated / Backup Designated Router (DR/BDR): Act as central Point for exchanging OSPF info on same segment; Information Exchange only with DR/BDR by Routers → Improves Performance
Election Process: on multiaccess networks; Based on router-priority in OSPF Hello → Router ID
Triggered only on startup / (B)DR goes offline → not when Router with higher-Priority comes online

LSA Types:
T1, Router: Describe State of connected links; Flooding restricted to origin area
T2, Network: DR describes Network segment Flooding restricted to origin area
T3, Summary: ABR summarizes LSAs to advertise outside of Area Flooding: All, but TSA
T4, ASBR-Summary: Sent to describe ASBR by ABR; ASBR sends T1+E; ABR summ to T4;
T5, AS External LSA: ASBR creates → propagates through domain Flooding: All, but SA & TSA
T5 Ext. Summarization possible on ASBR

Flow: 1: Establishing Neighborship Adjacency
DOWN: No hellos have been received
INIT: Hello from neighbor: No two-way conf.
2WAY: Hello with own RID received.
EXSTART: (Exchange start): After DR/BDR election, Establish Master/Slave Rel.: Determine Initial DBD Sequence Number
Stuck in EXSTART: MTU Mismatch → "ip ospf neigh."



2: Building Link State Database (LSDB)
EXCHANGE: DBD Compare
LOADING: Request/Receive missing
FULL: Identical LSDB

3: Execute SPF Algorithm
All Routers now execute SPF (see Dijkstra)

4: Build Routing Table: Installing Shortest Paths from SPF in routing table

Routing: Cost = Reference Bandwidth / Interface Bandwidth
Reference Bandwidth: Default = 100 Mb/s; Must be adjusted
Preference: Intra-area, Inter-area, E1, NSSA 1; E2, NSSA 2;
Intra-area (O): Cost to Destination within Area
Inter-area (O IA): Cost to ABR + adv. cost of summarized
External (O E1): Cost to ASBR + external Metric (default 20)
External (O E2): Cost = External cost (ignoring internal cost) (default Ext. Route.)
External (O N1 / O N2): External E1/E2 propagated to Totally NSSA / NSSA
Equal Cost Multipath: Requires modified Dijkstra, Load Balancing based on src/dst (per stream)
Summarization: at ASBR/ABR → smaller routing table, fewer LSA Floodings
continous network numbers in areas improve summarization; Only Smallest Cost propagated
ABRs summarize T3 LSAs before injecting into Area 0; Locally created route to 0 → avoid loops
ASBR: summarize T5 LSAs before redistributing; Locally created route to virtual addr. Null0

Area Type determination:
Classical Area: Accepts T3, T4, T5
Stub Area: Accepts T3; Filters T4, T5; Ext. Netw. reach T3 default route
Totally not so Stubby Area: Filters T3, T4, T5; ASBR in Area
Inter / External Netw. reach T3 default route is only T3 LSA
Not So Stubby Area: Accepts T3; Filters T4, T5, ASBR in Area
No default route generated by ABR
Totally Stub Area: Filters T3, T4, T5; No ASBR
Inter / External Netw. reach T3 default route is only T3 LSA

Stub: Blocks T5 (SA) Area Types:
No external routes (O E1, O E2)
default route to ABR
No external routing domain in Area

Totally Blocks T3, T4, T5 (TSA)
No inter, external (O IA, O E1, O E2)
default route to ABR
No external routing domain in Area

Not So Stubby (NSSA)
Stub + Export ASBR
T7 converted to T5 at ABR
No external routes (O E1, O E2)
Can have external routing domain

Totally Not So Stubby (TNSSA)
TSA + NSSA
No inter, external (O AI, O E1, O E2)
Can have external routing domain
default route to ABR

Network Types:
Broadcast Multi-Access Networks: Is a network where multiple Routers are connected to same bcast domain, and optimization through DR/BDR Election happens
Non-broadcast Multi-Access Networks (NBMA): Network where more than two Routers are connected without bcast ability. These networks require extra configuration and emulate broadcast network. Also DR/BDR get elected
Point-to-Point Links: If only two routers exist in bcast domain no DR/BDR election happens
Point-to-Multipoint Network: Are also non-bcast networks but act like point-to-point and not like NBMAs
Passive Interfaces: For Security & Performance standpoint all Interfaces behind which no further Routers expected shall be marked passive
On Passive Interfaces routers don't send / receive LSAs → Attackers can't send Malicious LSAs

Intermediate System to Intermediate System (IS-IS)

Overview:

Link-State Protocol: Designed for use as a dynamic routing protocol for routing Connectionless Network Protocol packets in the ISO CLNS environment

Security: Not IP based → Packets don't get routed. Attacker must be directly connected

Efficiency: 1000 Routers per Area, IS-IS Updates grouped as one LSP. Prefixes installed in more efficient way, less CPU intensive than OSPF; Default Timers → Faster Failure Detection

Connectionless Network Service (CLNS): Supported by L3 Protocols that coexist

Connectionless Network Protocol (CLNP): Similar to IP; Alternative till 2000. Ethernet Prot. type 0x FEFE

Provides Network Services for ISO transport protocols (Not UDP/TCP); Identified in L2 by

End System-to-Intermediate System (ES-IS): Protocol running between hosts & Routers

Automates information exchange, facilitates adjacency; Counterpart to ICMP, ARP & DHCP

Intermediate System-to-Intermediate System (IS-IS): Exchange Routing Information; Obtains data-link addresses (usually MAC, referred as subnetwork points of attachment (SNPA) and stores them in adjacency DB; Capable of Building RIB for IPv4/IPv6; Used by SPs on Internet Backbone

Packages: Each IS-IS packet contains Header with common fields, several optional variable-length fields containing specific routing related information (Type, Length & Value (TLV)) used to extend IS-IS easily by introduction of new TLV fields → Key Strength of IS-IS

Hello Packets (IIH): Establish & Maintain Adjacencies; Subcategories: LAN Level 1 hello packets; LAN Level 2 hello packets; Point-to-Point hello packet; Hello Interval: IS-IS sends packets to peers in hello interval. Default Value: 10s (Non DIS); 3,3s (DIS) (default 3)

Hello Multiplier: Determines when declare Router dead (Holdtime) = Hello Interval * Hello Multiplier

Link-State Packets (LSP): Communication of Link-state information; Subcategories: Level 1 LSP, Level 2 LSP; Acknowledged on point-to-point network with PSNP

Sequence Number Packets (xSNP): Control Distribution of LSP & Request of LSP Providing mechanism for Synchronization of Link-State Databases (LSDB)

Complete Sequence Number Packets (CSNP): Description/Complete list of LSP; equivalent OSPF DBD

Partial Sequence Number Packets (PSNP): Acknowledge & Request missing LSPs

Adjacencies: must be in up state for a router to send/process LSPs

Level 1 Adjacency: Formed when area addresses match unless configured otherwise

Level 2 Adjacency: Formed alongside Level 1 adj. unless Router configured Level 1 only

If areas don't match → Only Level 2 adj.; If transmitting Router Level 2 only, Receiving Router must be Level 2 capable → Otherwise no adj.; Backbone must be contiguous

Steps: 1. R1 sends out Level 1 hello: Active Neighbor TLV not included 2. R2 receives R1's IIH and Verifies Parameters: Authentication type & Credentials must match (if any); SystemID must be unique within area; Ensure protocols supported (OSI, IP, both) match; Interfaces must share common subnet, MTU matches; 3. R2 creates new entry in neighbor table and sets state to initializing for adjacency type (L1 or L2) 4. R2 advertises L1 IIH hello with R1's MAC (Active Neighbor TLV); Completes 3-Way Handshake 5. R1 & R2 set neighbor state to up for the adj. then advertise LSPs to eachother 6. When LSDBs Synchronized, Both Run SPF Dijkstra and install routes to RIB

LAN Adjacencies: Area 1, Area 2, L1, L2

Point-to-Point Adjacencies: Initialized by receipt of Intermediate System Hello (ISH) through the ES-IS Protocol. Ensures MTU match by Padding. Padding can be limited/config

Multi-Access Media Adjacencies: on broast Interfaces ISH not utilized. IIH are broasted as link goes up; Subsequent Hellos confirm Two-way state → Otherwise stays initialized

Designated Intermediate System (DIS): Minimizes Complexity of mgmt of multiple adj. Pseudo nodes = virtual node, who's role played by elected DIS

Election: Based on Highest Interface prio with highest SNPA as Tiebreaker

Default Priority (Cisco) = 64 (better); Separate DIS Role for Level 1 / Level 2

Responsibilities: Generating Pseudonode link-state packets to report links to all systems Carrying out flooding over LAN (CSNPs) for corresponding routing level

Backup DIS: Not Existant; Unnecessary → bc. short hello interval allows fast failure detection and subsequent replacement; Higher Priority DIS immediately overtakes DIS Role

MAC Level Mcast Addresses
Level 2 Adj: end with 15
Level 1 Adj: end with 14

Hierarchy:

IS-IS Supports 2 levels of hierarchy: Level 2 = Backbone (contiguous); Level 1 = Other

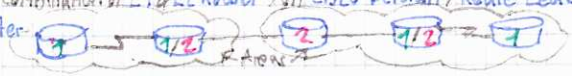
Plus hierarchical segmentation in Areas to contain LSP flooding & reduce overhead

Level 1 Router: knows local Area; Uses Level 1-2 Router as Default Gateway with Route-

By default L1-only Routers receive only Default Route from L1-2 Router but more specific a leaking

Level 2 Router: knows Routes to Areas; Signals Backbone connection with "attached" bit in L1 LSP, which is flooded through Area

Level 1-2 Router: Combination of L1 & L2 Router; On Cisco Default; Route Leaking done at L1-2 Router



Network Service Access Point (NSAP) OSI Terminology for a network layer address. Provide Foundation for routing datagrams in a CLNP network. IS-IS functions Designed around NSAP; IP Router require NSAP for IS-IS

Components: Area ID: <Private AFI 49>. <4hex Area ID> From 0 to 65,535

Sys ID: Unique per Node; Size match within Domain; On Cisco default 6 byte long ID Use of MAC addr of router (embedding Layer 2 addr into Layer 3) with many Interfaces must decide Or Use Loopback IP Address by rearranging IP in 4 digit groups and splitting with 0s

N-Selector (NSEL): Set to 0x00 (for Router) Marks Special NSAP type → Network Entity Title (NET)

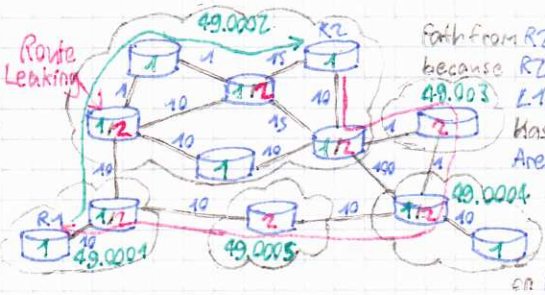
Routing: Uses the Dijkstra Shortest Path First algorithm

Supports multiple protocols by building topology first & then overlaying protocols like IPv4/6 both L1/L2 installed to Routing Tbl.

Route Preference: L1 preferred over L2 Path; Lowest metric preferred.

Route Leaking: Level-1 Routers are unaware of end-to-end metrics of interarea routes and use a Default GW for interarea routing → They send Traffic to their nearest Level-1-2 Router which causes Suboptimal Interarea Routing

Route Leaking solves Issue by allowing Level 1-2 Routers to leak specific L2 interarea routes into Level 1 of their local Area



Path from R2 to R1 is suboptimal because R2 sends Traffic to nearest L1-2 Router and this one has the nearest path to Area of R1 via other Areas 3, 4, 5 with Metric 22 & via own Area metric 35. Route Leaking on L1-2 Router that indicates direct connection solves issue

Default Interface Priority = 10

Border Gateway Protocol (BGP)

Overview:

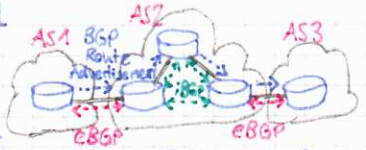
Path-Vector Protocol: External Gateway Protocol (EGP) → connects different AS in the Internet; Provides: Scalability, flexibility and network stability.
Router ID (RID): 32 bit unique number: Identifies BGP Router; Use Loopback IP
Communities: 32 bit optional transitive → can traverse AS. Can be appended, removed, modified; Provides additional fields for tagging → Modifies routing policies
Network Layer Reachability Information (NLRI): Routing Updates (Prefix)
Network Statement: Identifies netw. prefix installed into BGP table; Origin Path Attr: Connected Network: next-hop 0.0.0.0; origin: IGP; BGP weight = 32768; Depends on RIB prefix type
Static Route or Routing Protocol: next-hop = from RIB; origin: I (IGP); weight: 32768; MED = IGP Metric
Redistribution Statement: Like Network Statement but origin type = "incomplete"
Autonomous System Number (ASN): originally 2 bytes extended to 4 bytes
Managed: by IANA & Assigned to RIRs; RIR assigns to ISPs and Organizations with Provider Independant (PI) address Space; In EMEA (Europe) RIR is called RIPE
Private ASN: like RFC1918 spaces: 64512-65535 & 4'200'000'000 - 4'294'967'294
Loop Prevention: Check AS_PATH for own ASN → Discard of advertisement

Messages:

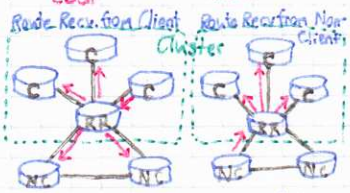
OPEN: init session, negotiate capabilities [BGP Version, ASN, holdtime, RID, etc.]
Holdtime: reset on UPDATE/KEEPALIVE; 180s
KEEPALIVE: every 1/3 of holdtime
UPDATE: advertises new routes & withdraws Include NLRIs; Composed of: withdrawal list, attributes of new prefix [AS_PATH, MED, ...] & new prefix
NOTIFICATION: sent on error, closes session holdtime expiration, capability change, reset required

Session: (adjacency between routers)

always Point-to-Point, not necessarily with direct neighbor
Discovery: Manual
Multihop Session: Sessions with no direct link
Requires: IP connectivity e.g. IGP like OSPF, Static
Best Practice: use Loopback as BGP interface & RID = Loopback (eBGP: set TTL to 2; iBGP: resistant to internal link failure if alternative path)
External BGP (eBGP): BGP between two AS
TTL = 1 by default → Multihop = drop → Set TTL
Changes Next-hop to own interface IP; Appends AS to AS_PATH → Does Loop Prevention
Internal BGP (iBGP): Connects BGP Routers in AS
TTL = 255; AS_PATH never modified; Next-hop per default not modified



Route Reflection: iBGP must be fully meshed
Router called Route Reflector (RR) forwards all routes
iBGP only connects to RR; Rules:
#1 NLRI from non-client → only forward to client
#2 NLRI from client → forward to all
#3 NLRI from eBGP → forward to all



Clients: Not fully meshed Non Clients: Fully meshed

Enterprise Connectivity Options

Single-Homed

Customer Uses BGP:
Changes in topology sent to provider
Provider May redistr. Changes to Internet
Customer doesn't use BGP:
Default route to reach outside networks
Specific Route on ISP side to reach customer
Easiest: Static Routing

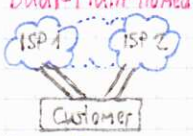
Dual-Homed

Used in Primary / Backup Design
For outbound routing: ISP runs BGP between Core & Access; Only Advertises Default route
For Inbound Routing: Use BGP to make ISP learn customer routes out
Local Pref on Customer side decides primary / backup
Characteristic: Single Point of failure due to ISP failure

Multi-Homed

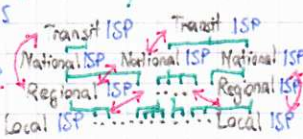
Used in Active/Active Design
By receiving full routing table, path through Internet can be optimized
Customer NEVER wants to be transit: Filters on eBGP Routers to only advertise out to ISP1/2 routes originated in own AS
Most Redundancy

Dual-Multi-Homed



Dual Homed Bidirectional:
Identical Local Pref required on ISP & Customer Side
iBGP between Customer Routers
eBGP between ISP & Customer

The Internet



Network of Networks; over 80'000 ASs, Unordered set of interconnection
Transit: ISP provides reachability to other ISP. Share routing tables; usually smaller ISP pays bigger ISP
Peering: Business Relationship where ISPs provide to each other reachability to each predefined portions of routing table, No fees
Routing Policies: Determine which routes to advertise / accept
Internet Exchange Point (IXP): Network hubs to peer or exchange traffic
Public IXP: Route Server exchanges infos between all parties; Adds itself to AS_PATH, but not nexthop; Lack of Policy control
Private IXP: Members peer bilaterally → better granularity; - more legal contracts

Routing:

BGP table contains all routes, best installed in RIB, route fails: next best path installed in RIB (first automatically best); Only best path advertised to neigh.
Recalculation on: BGP next-hop reachibility change; Failure of Interface connected to eBGP peer; Redistribution change: Reception of new paths for route
Route Selection Process: Attributes processed in order listed
- highest weight (local to router) (0-65535) (default: 32768)
- highest local preference (global within AS)
- Routes that router originated
- Shorter AS_PATHs (only length (quantity of AS listed) is compared)
- Lowest Origin Code (IGP < EGP < Incomplete)
- Lowest MED (also called metric)
- Prefer external (eBGP) paths over internal (iBGP)
- For iBGP paths, prefer path through closest IGP Neighbor
- For eBGP paths, prefer oldest (→ most stable) path
- Route with lowest neighbor BGP RID value
- Paths from router with lower BGP RID

Route Manipulation:

Manipulate inbound traffic to shape flow
Filtering: e.g. with REGEX → filter AS_PATH; Selectively advert./receive routes
- * any; ^$ locally originated routes; ^100_ learned from AS100; _100$ originated from AS100; _100_ Any instance of AS 100; ^[0-9]+$ directly connected ASs; '_' = \[a-zA-Z0-9]

Traffic Engineering: No real inbound traffic control; Possible to influence (if not overruled by sending party
Local Preference: Influence how traffic leaves network; Highest Pref routes traffic
MED: Influence how traffic enters network; Lowest MED routes all traffic
AS_PATH Prepending: Add local AS multiple times to AS_PATH → Unattractive
Aggregate: Specific routes are preferred if one ISP summarizes route but other doesn't → traffic flows through latter
Next-Hop Self Manipulation: Nexthop Access Issue; iBGP set next-hop Attribute to own address when advertising;
'router bgp <AS> \n neigh <IP> next-hop-self'

Path Attributes

Well-known mandatory: must be included with every Prefix advertisement
Well-known discretionary: may or may not included
Optional transitive: attributes stay with route adv. AS to AS
Optional nontransitive: cannot be shared from AS to AS
Additional Info
AS-Override: Overrides loop-prevention; Replaces AS number of originating router with AS of sending BGP router; → Multisite MPLS networks with same ASN
eBGP Multihop minimum value: must be at least 2 e.g. for loopback peering

Wide-Area-Network (WAN)

(L) = Legacy

Purpose: Connects remote LANs. Service Providers (SP) provide interconnection links

Requirements: Bandwidth • Control/Security • Network Design • Availability/Resilience • Network Mgmt.

Private-Dedicated-Leased Lines: Point-to-Point Lines. Monthly lease fee.
⊕ Guaranteed Bandwidth ⊖ Very Expensive

Variants:
- Dark Fiber: Physical Fiber Leased from SP; Excessive; Difficult to get
- Coarse Wavelength Division Multiplexing (CWDM): 16 Lambdas; 120km; Employment of MUX passive optical; ⊕ Cheaper than DWDM ⊖ Less Performant
- Dense Wavelength Division Multiplexing (DWDM): 80 Channels with 10 Gb/s; 1000km; Electronic active Device ⊖ Very Expensive; Modern submarine Communication

Private-Switched-Circuit Switching (L): Dedicated Circuits (virtual conn.) for data before transmis

Integrated Services Digital Network (ISDN): Changes analog Public Switched Telephone Net (PSTN) to digital Time-division Multiplexed (TDM); Allows 2 or more signals as subchannels on one channel

High Bandwidth Fiber Standards: American ANSI: Synchronous Optical Networking (SONET) European ETSI and itu: Synchronous Digital Hierarchy (SDH) represent transmission rates

Terminology: Local loop (last-mile): Copper or fiber connects Customer Premises Equipment (CPE) to Central Office (CO). Data Communication Equipment (DCE) devices that produce data [Cust→SP] Data Terminal Equipment (DTE) Pass Cust. Data over WAN. Demarcation Point: indicates Responsibility change

Public-Internet-Broadband IPsec VPN: Site interconnectivity requires encrypted traffic established by

Options: 1: Site-to-Site VPN with Point-to-Point VPN tunnels: IPSec tunnels; Connects LANs; ⊖ traffic overhead; No VoIP
2: Remote-Access-VPN: IPsec tunnel between PC and GW/Router; Used by Employee to connect to Company Net. 3: Dynamic Multipoint VPN (DMVPN): Cisco proprietary: Static Hub-to-Spoke tunnels to HQ; Spoke-to-Spoke (on demand) between Branches

Private-Switched-Packet Switching (L): Packet-switched networks (PSN) split traffic data into packet. Not require circuits. Switches determine route based on addressing info

Connection-oriented systems: Predetermined route; only ID carried. Examples: Frame-relay; ATM

connectionless systems: Full addressing info carried; Switches/Routers evaluate route based on it

Metroethernet: Metropolitan Area Network (MAN). Commonly connects subscribers to SP but also can connect offices. MPLS: Removes OSI Layer dependencies; Exists as L2 and L3 VPN

WAN Topology Choices:

Multiprotocol Label Switching (MPLS)

Basics:
- Carries L3 payloads (IPv4, IPv6) and L2 payloads (Ethernet, ATM, Frame Relay, PPP, HDLC); • Based on Label for path instead of IP addr; • Was SP technology but now used in many Enterprises

Dataplane: Responsible for forwarding; no intelligence; does what Controlplane says

Cisco Express Forwarding (CEF): Cisco name for FIB and Adjacency Table. The Adj. Table contains pre-created data-link header routers copy in front of packet; ⊕ improves Performance

FIB & LFIB are used to forward MPLS Packets. Contain Label Info; Outgoing Int; Next-Hop

Forwarding Information Base (FIB): Used for unlabeled incoming Packets; Cisco IOS matches packet Dest. IP Addr to best Prefix in FIB and forwards based on that entry

Label Forwarding Information Base (LFIB): For labeled incoming Packets; Cisco IOS compares Label of incoming packet to LFIB list and forwards on that entry

MPLS Header & Label [Label (20 bit) | EXP (3b) | S (1b) | TTL (8b)] 4 Byte

MPLS Experimental allows mark for QoS; S-bit: (Bottom of stack Indicator) MPLS uses Multiple Headers stacked; TTL: used to identify Loops; Method for Traceroute

MPLS TTL Propagation: Routers ignore IP TTL bc. of encap. Can be disabled

the MPLS TTL is set to 255 and the MPLS Network appears as 1 Router

Ingress Provider Edge (PE) routers: Decrement IP TTL and copy it to MPLS TTL

P Routers: Swaps Label and decrements only MPLS TTL ignores IP TTL

Egress Provider Edge (PE) routers: Decrements MPLS TTL, pops final MPLS header and then copies MPLS TTL into IP header TTL field.

MPLS Layer 2 VPN: Layer 2 service also called Carrier Ethernet. SPs typically have no participation in Layer 3 WAN routing and forward data based on its L2 Information

Virtual Private Wire Service (VPWS) = Pseudo wire (PW): known as E-Line in Metro-Ethernet forums; Provides Point-to-Point connectivity model = PW

If Router connected to VPWS, an IGP needs to be deployed to forward between sites

Network completely transparent to CE router; **No** MAC learning at PE level

Virtual Private LAN Service (VPLS): known as E-LAN in Metro-Ethernet Forums Provides Multipoint any-to-any connectivity model; • Emulates LAN segment across MPLS-Backbone; MAC-learning at PE level; Frames forwarded based on dest. MAC

Broadcast, Multicast and unknown unicast: If dest Addr unknown → frame flooded to all Ports

Loop-prevention: Uses Split-Horizon: received frames from MPLS tunnel IPN from core never back

If there is a full mesh of PWs, PEs do not need to relay packets back to the core.

Cisco Dynamic Multipoint VPN (DMVPN): Combinations of different technologies: Multipoint GRE (mGRE) tunnels, Next-Hop Resolution Protocol (NHRP), IPSec Encry. VPN tunnel can share multiple endpoints that all use tunnel; With DMVPN each Spoke has a static connection to the Hub and On-Demand tunnels to other Spokes

Generic Routing Encapsulation (GRE): Developed by Cisco but standardized by Internet Engineering Task Force (IETF); Protocol number 47; Provides capability to tunnel any unicast or multicast packet; Two endpoints of GRE tunnels act as if connected directly.

MPLS Layer 3 VPN

Configuration: 4 steps; order not important: 1) Creation of each VRF, RD, RT and associating customer-facing PE interfaces with correct VRF; Each VRF needs IPv4 address family; Also one VPNv4 address family for PE connectivity; 2) Configuration of the routing protocol (IGP, BGP or Static) between PE and CE; 3) Config of mutual redistribution between PE-CE Routing protocol and MP-BGP. Only if PE-CE Protocol not eBGP (bc. MP-BGP support external BGP Routes per default)
4) Configuration of MP-BGP between PEs;

Penultimate (1-less than ultimate) Hop Popping (PHP): Does what name implies

Enabled per Default on Cisco. The Penultimate (Next-to-last) Hop has the responsibility to Pop the outer MPLS label; The PE-router has more responsible processes than P routers, With PHP the label lookup process for PE can be cut in half, which reduces the load for PE routers. For P router it doesn't matter whether he pops or swaps label, hence the Performance of a MPLS L3VPN increases.

Control Plane: Builds Tables; makes Routing decisions; Relies on Control Plane Protocols to populate FIB & LFIB with Labels; Unicast IP forwarding with IGP (OSPF, IS-IS etc) and MPLS specific

Label Distribution Protocol (LDP): Advertises labels for each prefix in IP routing tbl. LDP used to send msg. containing IP prefixes and corresponding labels. Adv. triggered by new Route

Hello feature to discover LDP Neighbors and to determine IP of Neighbor to which a TCP connection gets established. LDP Hellos multicasted to 224.0.0.2 using UDP 646 Optionally contains transport address = IP the MPLS Router wants to use. After Neighbor-Discovery through Hellos a TCP 646 connection formed. Over TCP bindings (prefixes Advertised

Label Information Base (LIB): Labels and related info stored in LIB which are used to forward packets. Best label and outgoing interface are populated into the FIB & LFIB. LIB contains all Labels known to Router and FIB & LFIB contain only Label for the currently used best

Label Switched Path (LSP) segment.

MPLS Layer 3 VPN

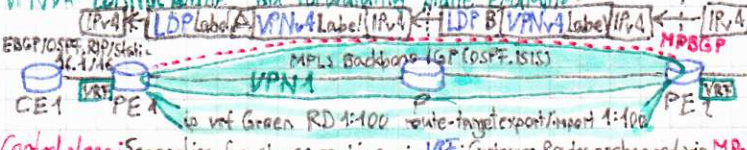
MPLS Layer 3 VPN: Provide privacy in L2 WAN services without encryption

Multiprotocol-Border-Gateway-Protocol (MP-BGP) overcomes challenges of connecting multiple LANs, e.g. Duplicate IP-Address spaces. RFC defines use of multiple routing-table instances called Virtual Routing and Forwarding (VRF) to avoid the issue

Internet VPN: • Public Internet-based **v.s. MPLS VPN:** provided by single SP; L2 or L3
Ubiquitous connectivity; • Layer 3 only; Low cost • QoS control; • Enables delivery of SLA; • Infrastructure
Throughput/Latency = Best effort; Managed by ISP; owned by one Organisation; • Privacy without encryption

Terminology: **Customer Edge (CE):** Router, no knowledge of MPLS; Directly connected to MPLS PE
Provider Edge (PE): Connects CE (IP) and P (MPLS); Distributes VPN info through MP-BGP to other PEs with VPN-V4 addresses, extended community, label
Provider (P): MPLS router with no direct link to a CE router. Only forwards labeled packets
Virtual Routing and Forwarding (VRF): Concept of virtual Router realized with distinct routing tables (VRF), that store routes of different customer VPNs separately. At least one VRF required for each customer attached to routers; VRF exists inside single MPLS-aware router
VRF Components: RIB, per VRF, a CEF FIB populated by RIB; Separate instance of IGP to CE-Router
MP-BGP and Route Distinguishers (RD): RD allows BGP to advertise and distinguish between duplicate IPv4-Prefixes. RDs are attached to the advertisement of each Network Layer Reachability Information (NLRI) (prefix) to uniquely identify routes. New NLRI Format VPN-V4 consists of: 64 bit RD and 32 bit IPv4 prefix. Following possible formats:
Type 0: = 2 byte-type: 2-byte-ASN: 4 byte-SP assigned; Type 1: = 2 byte-type: 4-byte-IP v4: 2 byte-ass SP
Type 2: = 2 byte-type: 4-byte-ASN: 2 byte-SP assigned;
Route Target (RT): RTs are a 8-byte field, VRF parameter, which are unique values to define import/export rules for VPNv4 routes. PEs advertise RTs in BGP Updates as BGP Extended Community path attributes. RTs placed in a export list are attached to every advert. These attached RTs on receive get compared with the import list, only on match route gets imported

VPNV4 control plane and forwarding plane example



Control plane: Separation of customer routing via VRF; Customer Routes exchanged via MP-iBGP by PE
Forwarding plane: Separation of customer traffic via VPNv4 label (also LDP-Header); used by receiving PE to identify VRF Table
Processing Steps: 1) CE2 forwards IPv4-packet to PE2 via IGP; 2) Lookup of the VRF Routing table for the egress PE; 3) PE2 pushes VPNv4 label (learned v. MP-BGP) to packet
4) PE2 pushes outer IGP label (learned via LDP) to packet and forwards to next-hop Router
5) P-router swaps outer IGP label and forwards to PE1; according to LFIB with CEF
6) PE1 pops first outer IGP label and the VPNv4 label and forwards IPv4 packet to CE1

Network Virtualization (NV)

Reason & Advantages: Segmentation & Security between Customers; Address reuse (in different VLANs, VRFs, MPLS-VPNs etc.); Higher capacity utilization (sharing physical Ressources); Higher Flexibility (adaptable Config)

Technologies: NV gives Physical Network ability to support multiple virtual networks. Allows better Utilization of network resources and maintains Hierarchy, Virtualizes devices, data paths and services

Link/Device Virtualization: Virtualize links/devices into multiple logical from one physical and the opposite: to make one logical link/device out of multiple physical (clustering)

Virtual Links: **Purpose:** Link Virtualization to segment traffic and Clustering avoid STP enabling Loops!

Multiple Logical from physical: Single-Hop link: Done by adding address field (802.1q, MPLS Label, etc.)

VLAN Trunking [802.1q]: Trunking VLANs over physical link allows complex logical topologies. 4096 VLANs on same physical cable because of 12-bit VLAN ID Field inserted into Ethernet-Header

Multiple-hop link: For multiple hop links use of tunnels (GRE, VPN, VXLAN, etc.) is required

Single Logical Link from multiple physical: Usually done with LACP or proprietary EtherChannel/PortAgg.

Link Aggregation Control Protocol (LACP) [802.1AX, 802.1aq, 802.3ad]: Standardized by the IEEE. Bundles up to 8 physical links in one logical. Supports automatic formation of connections by having multiple modes: (On, Off, Active, Passive); Channels only formed: both On/both Active/one Active and one passive. Automatic formations negotiated by LACP packets multicast MAC 01-80-c2-00-00-02

Multichassis EtherChannel (MEC): Advanced EtherChannel technology; link aggregation to two separate physical switches. Only Supported by VSS and allows VSS to appear as single logical device

Network Function Virtualization (NFV)

: to virtualize network functions and Physical Devices as Virtual Machines (VMs). NFV can run on specialized HW or common servers/hypervisors. Opens the field to other vendors and open-source solutions (e.g. OpenSwitch)

⊕ Run network services on third-party HW; • Increase flexibility and agility (Duplicate Networks and VMs for testing; Reduce time-to-market and time-to-deployment, reduce maintenance cost); • Reduce need for spare parts; • Reduce number of redundant components (Simplification); Cost benefits (Pay-as-you-go/grow licensing)

⊖ New technology; Increased complexity; Organizational changes required; May not provide same level of security; • Higher energy consumption than specialized Hardware

Challenges: **Performance** challenges (Dedicated HW always faster); **Network** challenges (service insertion and stitching?); **Orchestration** challenges (Mgmt of Services/VMs); **Operational** challenges (Redundancy, Auto-scaling); **Product delivery** challenges (licensing cost (Recurring costs instead of one-time investments));

Virtual Context: • Single physical device; • Multiple virtual contexts • Separate Mgmt plane; • Shared resources; • Tied to physical device; • Not flexible

vs.

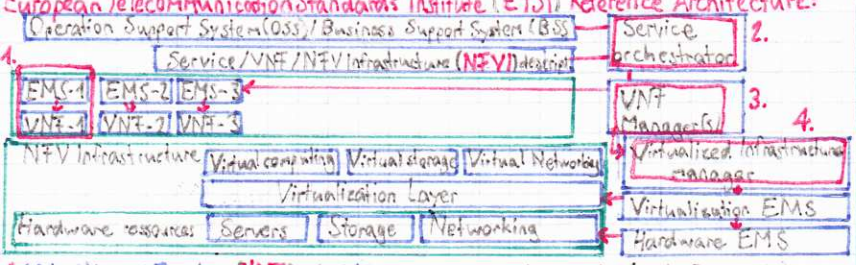
Virtual Appliance: • Network appliance in VM or Container format → **Deployable on any hypervisor/cloud platform** • VM appliances operate above network virtualization technologies (Virtual Ethernet NICs); • Virtual appliances run on any transport provided by hypervisor (VLAN, VXLAN, etc.)

NFV Use-Cases/Examples:

Virtualized Network Functions: Switches, Routers, Firewalls, Load-Balancers, **Customer Premises Equipment (CPE)**

Networking as a Service (NaaS): Routing as a Service (RaaS), VPN as a Service, Server Load Balancing as a Service, Web/Email Security as a Service, Intrusion Detection/Prevention as a Service, Application Monitoring as a Service, Unified Communication-as-a-Service

European Telecommunication Standards Institute (ETSI) Reference Architecture:



1. Virtual Network Functions (VNF) is the VM that contains the network function (Router, Firewall etc.). Element Management System (EMS) configures & monitors the network function.
2. Service Orchestrator manages/coordinates the creation of end-to-end Network Service that involves VNFs and/or VIMs from different domains, uses DB to describe available Network Services, Link Types etc. Tells the VNF Manager to start the needed VMs and to connect them in the right way
3. VNF Manager manages lifecycles of VNFs. Creates, maintains & terminates VNF instances installed on the VMs. Responsible for Fault, Configuration, Accounting, Performance (FCAP) and Security Management of VNFs. Scales VNFs up/down → scaling CPU usage up/down
4. Virtualized Infrastructure Manager (VIM) creates, maintains & terminates VMs from physical resources in NFVI domain. Resource allocation is complex bc. a lot of requirements/constraints need to be met at the same time. Allocation and release dynamic, must scale up/out automatically

Use-Case: Distributed Soft Switches: Widely spread in Datacenters; Switches required to automatically move ports of specific VM if it is moved. Achieved with **Distributed virtual switch** architecture with centralized control plane; technology similar to Stackable Switches

Use-Case: Datacenter Virtualized Services: Virtualization of complete three-tier application stack implemented with NFV. Whole Network Infrastructure for App. Virtual → Hybrid cloud migration

Use-Case: Enterprise Remote Site: Companies with many remote sites require different physical devices (Switch/Router/Firewall). With NFV run on 1 device or maybe two for resilience ⊕ E.g. replaced/Enhanced w/o going on-site ⊖ Single Point of Failure, Security Considerations

Use-Case: VPN for Hybrid Cloud: Companies that have Datacenter and add a Cloud Segment but don't trust Provider can encapsulate traffic on virtual Router running in the Cloud connecting cloud to site in Enterprise WAN. Especial Proprietary vRouter software possible

Use-Case (SP): Lightweight 4over6: SPs (e.g. german Telekom) reduce network complexity by having IPv6 access network, which supports IPv4 over tunneling. SP orchestrates the IPv4-tunnels that have VM-based tunnel terminations on the SP end with an orchestration Software to maintain the Mapping state.

Virtual Chassis: Device Partitioning: L2 (VLAN): Associates to one or more interfaces and privatizes an L2 interface. Each VLAN has MAC tbl. and STP-instance. L3 (VRF): Associates with ≥1 interfaces and privatizes L3 interface. Each VRF has own RIB and Routing Process (EIGRP, OSPF, BGP)

Virtual Device Context (VDC): Is when 3 plane virtualized: **Control plane:** Routing Processes, generation of IGP packets etc.; **Dataplane & I/O:** Multiplexing N streams of data traffic; **Management plane:** Ability to manage Virtual Contexts separate VDC can be multiple contexts on physical device → ⊖ low resiliency & little flexibility/mobility

Stacking multiple Devices: Redundant Logical Switching Architectures have the problem of overheads in Management, resiliency and performance (STP disabling Inter) **Solution: Stack Wise** (Access-Layer) / **Virtual Switching System** (VSS, Distribution-Lay)

⊕ If port number higher than available, switch can be expanded with additional chassis • Fast link (backplane) between chassis, better than connection of mult. independent switches. • One Active Control plane; • Only one switch must be configured; • One distributed forwarding Tbl.

Stack Wise: Connects mult. switches over **ring-topology**. Managed by **master-switch**, master controls centralized funtions, builds & propagates L3 FIB, manages/propagates config file (Running Config) to stack, controls console, upgrades Stack. **Slaves** receive config on boot-up.

Virtual Switching System (VSS): Consists of switches defined as member of virtual switch domain resulting in single logical switch, with single control plane and mult forwarding planes. ⊕ Simplifies config and mgmt, single point of mgmt., minimization of STP-reliance, non-loop design → reduced complexity, increased forwarding capacity, resilience

Unified Control Plane: VSS has an supervisor in each chassis with inter-chassis **Stateful Switchover (SSO)** with one in ACTIVE and one in HOT_STANDBY mode both run in synchronized mode → Same boot-env, running-config, protocol state, line card status. ACTIVE Supervisor manages control plane functions (Route Processor (RP) and hardware control. Supervisors connected via **Virtual Switch Link (VSL)** → faster

Overlay Networks:

Combination of virtualization functionalities. Virtual overlay over the physical topology. Routers & Switches mostly match with **underlay** links often get tunneled over different routers. Overlays manufactured for any use-case, complexity of physical network doesn't affect Logical Network. Overlays promote rapid deployment and operation of new network functionality w/o risk physical

Reason: Segmentation of **Underlay** & **Overlay** that have different **requirements:** Robust Underlay/Fabric need to be a high capacity resilient fabric, that is programmable and manageable with intelligent packet handling. Flexible Overlay Virtual Network which usually gets built/configured automatically (Software-Defined Overlay), enable address scaling/stretching, virtual topologies as well as mobility.

Types of Overlays: **L2 Overlay:** Acts as huge Switch/L2 Domain and com. devices can communicate over L2 without routing necessary; • Emulate LAN segment; • Transport Ethernet Frames; • Exposure to open L2 flooding; Useful in emulation of physical topologies (When Protocol used does not work with IP packets

L3 Overlay: Most prominent is MPLS. Looks for end-user like big router that employs tunnels to the Edges of the L3 Overlay → automatical encryption/security

Network Overlays: Logical topology built within overlay is distinct from that used by the underlying network infrastructure; **Edge Devices:** Router/switch endpoints; • Protocols for resiliency/loops; Traditional VPNs; • VPLS/LISP; • Inter-DC (DCI)

Build Methods: Map and Encap Approach: Intelligence can map physical to logical Endpoints. e.g. MAC is mapped to logical address registered in table. LISP is a technology that differentiates physical and logical and can map accordingly

Translation Approach: if mapping is impossible the Translation Approach is required. Works via **MAC-address rewriting** (MAC-Network Address Translation)

Host Overlays: Edge devices: Virtual Endpoints only; Single admin domain, VXLAN (Connection between VM Servers), NVGRE, STT, Geneve

Hybrid Overlays: Edge devices: Physical & Virtual; Resiliency & Scale; Open Standard.

Underlay Networks:

Network Infrastructure underlying Overlay Connectivity between Physical Devices via L3 → maximum stability and performance; Via tuning of routing protocols convergence time can be minimalized Via Equal Cost Multipath (ECMP) routing (OSPF, IS-IS, ...) performance optimal

Properties: Stable; Static, Scalable, fast-reconverging, excellent traffic Load Balance

NFV Use-Cases

Use-Case: Virtual CE: SPs must send End-Users dedicated expensive Routers, each time functionalities change they need to be replaced. vCE requires minimum on-site HW. ⊕ Upgrade & Scale

Use-Case: Virtual PE: Normally numerous Customers on every PE router. With vPE can optionally be reduced to one customer per PE. ⊕ Scale-Up/Out trivial

Use-Case: Per Customer Service: SP can build unique Service chains for each Customer (e.g. a Firewall no longer requires physical Device → deployed with one click

Use-Case: Control Plane functionalities: One of first virtualized Functionalities was the iBGP Route Reflector; Since it only requires an instance of the Routing Protocol and no Packet Forwarding; Was evident to virtualize

Software Defined Networks (SDN)

SDN is a new approach at Networking. Enables high scale, rapid network & service provisioning and management. Opportunity for Network engineers to lose repetitives

SDN Deployments: Traditional: Not an SDN deployment, but a normal network; But if question is "which are the three types of SDN deployments" Eg.: OSPF is it

Pure SDN (Software-only SDN): Outsources control plane intelligence from network devices to external controller, which programs the switches with their forwarding table. Very interesting from research perspective. Best if environment 100% virtual from same vendor. Examples: Openflow; VMWare NSX

Hybrid SDN (Hardware-based SDN): Has management entity, takes over configuration and provision, Keeps concept of Data-learning/forwarding Eg.: SD-Access; Cisco ACI

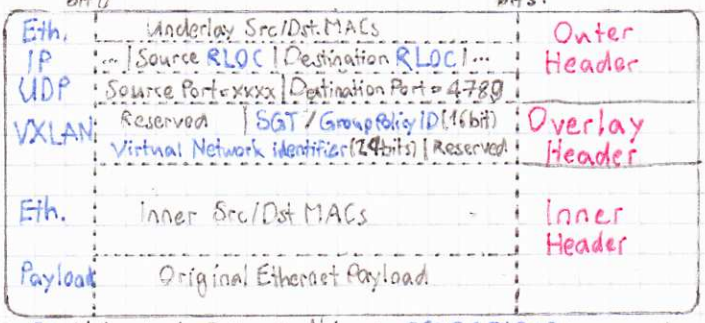
VXLAN: Offers method of data encapsulation, with underlying L3 Protocol to use funtionalities as Equal Cost Multipath (ECMP) and avoid Loop and STP easily

UDP-Encapsulation: VXLAN is a User Datagram Protocol (UDP)-based encapsulation. By encapsulating actual traffic in UDP, intermediate nodes don't have to be VXLAN compatible

Inner-Header: Ethernet L2 Frame with src/dst. MAC, together with the IP packet contained

VXLAN-Header: In between, transports two crucial pieces of metadata: Virtual Network Identifier (VNI) always present. Traditional VXLAN only contains the VNI but for SDN extensions VXLAN GPO/GPE contain additional field for lower granular Segmentation group-based with Scalable Group Tag/Group Policy ID

Outer Header: Standard UDP format encapsulated by IP but with Routing Locators (RLOC)s instead of Src/Dst IP that identify the node that Ecapsulates/Decapsulates



Intermediate System to Intermediate System (IS-IS): Predominant for SP-Networks; In Enterprise Networks OSPF & EIGRP been used more will change in future. IS-IS has powerful attributes: Simplified design, protocol-independent nature, and extensibility via type-length-value (TLV). Used for Cisco SDN

Location/ID Separation Protocol (LISP): operates as Overlay Control Plane for user and device reachability, in the Cisco SDA

basic functionality: Mapping IP-Address to a Destination, Solving "Where is?" question for Host/Service as DNS does for "Who is?" by mapping name to IP

Mapping System: contains Info on all hosts & Devices attached to network noting their IP-Address, MAC-Address, and Location. Acts as Single Source of truth. Always contains Correct & Actualized data. Location defined as node : behind which an end-device is located behind. DB built on Switch/Router/Dedicated Server

Registration: Mapping System DB populated by connected LISP speaking Switch/Router

Querying: If LISP node has end-user traffic to deliver to end-device it queries the LISP Mapping System and determines which RLOC represents node/dst. is behind

Terminology:

LISP Tunnel Router (XTR) (Edge/Internal Border): Registers Endpoint ID (EID)s with MS/MR; Edge Devices Encap/Decap; Ingress (entry point)/Egress (ITR/ETR)

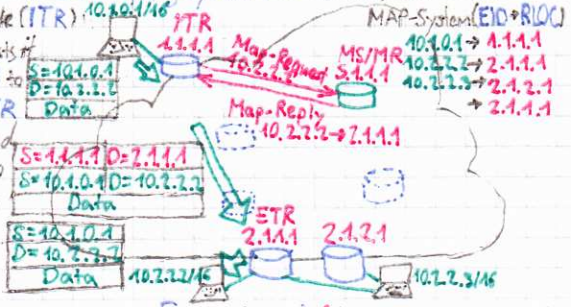
LISP Proxy Tunnel Router (PXTR) (External Border): Provides Default-GW when no mapping exists; Connects between LISP & Non-LISP domains; Ingress/Egress (PITR/PETR)

LISP Map-Resolver/Map Server (MS/MR): Device hosts LISP Mapping system DB and handles all registrations & queries → sends out replies accordingly

Map-Registration: The RLOCs (ITRs & ETRs) register subnets and hosts that are behind them into the LISP Mapping System DB as they come online

Map-Request: If a node (ITR) wants to communicate to hosts it queries the LISP MS/MR to determine how to reach ETR hence the hosts behind

Map-Reply: Response to the Map-Request, indicates which RLOC/set of RLOCs host or subnet is behind



LISP Operation Modes:

Proxy Mode: MS/MR responds directly to the ITR, by sending a Map-Reply on behalf of ETR which simplifies the transaction. Mode used for SD-Access

Non-Proxy Mode: Instead of responding directly, the MS/MR forwards the Map-Request to the ETR which sends the Map-Reply. Example described in Roaming Operation

Example (Proxy):
1. Where is 10.2.2.2; packet sent via Underlay to the ITR (RLOC 1.1.1.1)
2. Map-Request for 10.2.2.2
3. Map-Reply 10.2.2.2 lies behind RLOC 2.1.1.1
4. Cache entry for received Map-Reply
5. VXLAN tunnel Encapsulation
6. Decapsulation by ETR (RLOC 2.1.1.1)
7. Packet sent via underlay to end-host

Scalable Group Tags (SGT): used for application of network policies. Examples of Network policies are Access Control Lists (ACL) or QoS policies. Nowadays Networks us IP-Adrs to apply policies, but not intended for this use. IP overloaded with meaning already. Another mean frequently used for application are VLANs

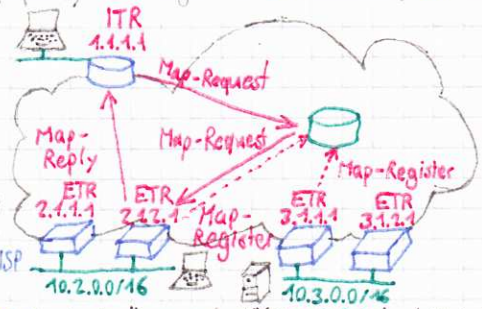
Group-Based Policies: SGT is small 16-bit tag assigned to users and devices as they authenticate and connect into network. Assignment can be simple & static for Example: each port of switch assigns connected devices with a specific tag or each AP SSID is assigned a tag. Assignment can also be more sophisticated & dynamic for Example: Based on dynamic validation of username/pw or other device credentials such as device type, that decide in which group a end-device ends up.

Authentication, Authorization and Accounting (AAA): AAA-Servers such as Cisco Identity Services Engine (ISE) are used to manage the assignment of tags/groups. Tags and the group-interactions that are permitted or denied are defined. Server dynamically distributes the policies on to the network devices, which apply policies locally for enforcement

Host Mobility: Average user nowadays has 2 to 3 devices all with Wi-Fi equipped and used in mobile fashion; Routing Protocols assume devices are in fixed Locations; no more true

Host Scale: Moving Devices want to retain their IP when roaming from AP to AP because IP-base Phonecalls cannot handle IP-Address changes. LISP Mapping can handle this task easily by just changing RLOC. LISP only sends info on request in contrary to routing protocols that flood data to ensure nodes have a consistent Tbl. This exhausts HW resources, especially in lower-edge devices. LISP is scalable, IGPs not

Roaming Operation: for Host-Mobility to work there must be a Roaming Process. figure to the right represents this process and at the same time also illustrates the Non-Proxy mode of LISP



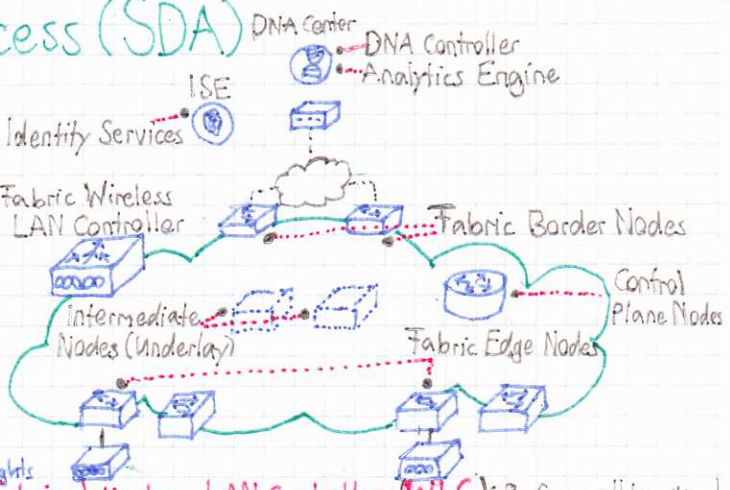
Roaming Steps:
1. Initially the Host 10.2.0.1 is in its "home subnet" (10.2.0.0/16) which is the subnet normally defined as attached to the 2.1.2.1 switch.
2. As host 10.2.0.1 comes up the switch with RLOC 2.1.2.1 registers the host at the LISP MS/MR. If anyone now wants to know 10.2.0.1's location the MS/MR can forward the Map-Request to RLOC 2.1.2.1 who responds directly bc. Non-Proxy
3. Now the host 10.2.0.1 roams over to RLOC 3.1.1.1. Host will keep IP-Addr but now it's IP does not match the destination subnet anymore
4. Roamed-to RLOC detects newly attached host and sends a Map-Register to the MS/MR, which now updates its mapping table with 10.2.0.1/32
5. The MS/MR now sends an update to the roamed-from RLOC to indicate that the host has roamed and is now behind the RLOC 3.1.1.1
6. The Roamed-from RLOC 2.1.2.1 now updates its tables to indicate the roaming and caches the location of the roamed-to RLOC

Stale Cache: Caching of Replies & Roaming updates reduces the load on the MS/MR, because previously queried destinations can be satisfied out of the local cache, avoiding unnecessary lookups. Problem of cache are stale cache entries that point to wrong (old) destinations. If an ITR with a stale cache entry sends traffic to wrong ETR, the ETR responds with a Solicit Map-Request (SMR) msg → Tells ITR that cache is out-of date → enforces re-query

Scalable Group Access Control List (SGACL): Security policies based on SGTs, are typically enforced on egress using SGACLs. Beneficial Property: Not necessary to insert policies for every possible egress destination of every ingress point in the network → Aids ability scale policy enforcement

Transport: SGT Exchange Protocol (SXP) exists for environments unable to carry SGT end-to-end in dataplane. Is a TCP-based protocol used to map individual IP-Addresses to SGTs. Better Solution is to transport the SGT inline, which is a traditional network means every device must be SGT capable. With VXLAN not necessary, due to encapsulation only end-devices must support SGT.

Software Defined Access (SDA)



Device Roles:

Digital Network Architecture (DNA) Center: Provides platform from which an SD-Access solution is designed, provisioned, monitored and maintained. Also provides Analytics platform functionality used to absorb telemetry and provide actionable insight base on this network data.

Workflow: four-step workflow model is employed by DNA-Center.
- **Design**: Common attributes such as major network capabilities (DHCP and DNS Services, IP host pools, etc.) are created aswell: network locations/floors & maps
- **Policy**: Allows creation of fabric-wide policies, such as creation of Virtual Networks, Scalable Groups and contracts that define access rights.
- **Provision**: Preceding two steps come together, and the actual network fabrics are created. devices are assigned to various roles within those fabrics, and device configurations are created & pushed-out to devices that make up Underlay & Overlay.
- **Assurance**: Allows for definition and gathering of telemetry from the underlying network, with the goal of analyzing the vast amount of data and to refine actionable insights

Identity Services Engine (ISE): Allows for the Policy and Segmentation capabilities of SD-Access to be based around endpoint and group identity, which allows network-based policies to be decoupled from IP-Addressing.

Control Plane Nodes (= LISP MS/MR): One or more network devices implement the LISP MS/MR functionality. Keep track of Location of attached end-devices

Border Nodes (= LISP PxTR): One or more elements that connect the SD-Access fabric to the outside. Provide reachability in and out the fabric environment and translate between fabric constructs such as virtual networks and SGTs and the corresponding constructs outside of the fabric environment.

Edge Nodes (= LISP xTR): Network elements that attach end-devices into the SD-Access fabric. Serve as the Ingress & Egress nodes and are responsible for device connectivity & Policy enforcement. In LISP Edge nodes are referred to as RLOC: and the endpoints are identified by Endpoint IDs (EID: MAC/IP.

Fabric Wireless LAN Controller (WLC): Performs all traditional tasks of a WLC such as: controlling APs, managing the Radio Frequency (RF) space used by APs, authentication of wireless clients and managing wireless Roaming. Also Fabric WLC responsible for provision of a constantly up-to-date map which EIDs sit behind which RLOC

Intermediate Nodes: Network infrastructure devices that sit inbetween of SD-Access fabric nodes. In most deployments, intermediate nodes consist of distribution or Core switches. Provisioned purely in the network Underlay and do not participate in VXLAN-based overlay directly. Forward packets based on outer IP-Address in the VXLAN headers UDP encapsulation. Do Not have to be SGT-Capable. Selected for their speed, robustness and ability to be automated by Cisco DNA Center for this role.

Fabric Capabilities:

Stretched Subnets: In traditional networks requires use of wide-spanning Layer 2 VLANs that place entire network domain into a single failure domain. SD-Access addressed this problem by the use of Distributed Anycast Default Gateway functionality, which ensures that the Default-GW is always on the first hop switch and can always be reached over same virtual MAC adress for this Default-GW. Any IP host deployed in an SD-Access fabric is per default stretched

IP Host Pools: Are the IP Subnets which are deployed in an SD-Access fabric and are created within the Cisco DNA Center. Span over whole network and are rolled-out onto all networks Edge switches, that act as local default GW for connected devices. Endpoints attaching to the SD-Access fabric are mapped into appropriate IP host pools either statically, or dynamically based on user/device authentication

Segmentation and Policies in the Overlay:

Any device connecting to fabric-based network such as SD-Access get authenticated (static / dynamic). According to authentication, they are assigned to SG linked with their role, aswell as being mapped into an associated VN. With VXLAN both tags are carried end-to-end across the network, allow Micro/Macro segm

Virtual Network (VN): Macro-segmentation is done using VNs. VNs are created within the Cisco DNA Center and are mapped to VRFs, which provide complete address separation, and are carried across network fabric as Virtual Network IDs (VNIs/VNIDs) mapped into the VXLAN dataplane header. Every fabric deployment contains a "default" VN, into which all devices are mapped per default. Also a Infrastructure VN (INFRA_VN) exists in every fabric deployment and contains any network infrastructure devices e.g. APs & extended node switches

Default-deny policy: Between VNs a default-deny policy is implemented. Means per-default VNs are completely separated and no VN-intercommunication possible

Scalable Group (SG): Micro-segmentation offered in a secure form by SGs for access control. Created within Cisco DNA Center & ISE. SGs are carried across the fabric network as SGT mapped into the VXLAN dataplane header. Scalable Group ACLs (SGACL) are used for egress traffic filtering and Access control with a SD-Access deployment, enforced at fabric edge and/or fabric border positions. Group-based policies are define using either a whitelist model or blacklist model.

Whitelist model (Default for Macro-segmentation / VN): All traffic denied between entities unless it is explicitly permitted (whitelisted)

Blacklist model (Default for Micro-segmentation / SG): All traffic allowed between entities unless it is explicitly blocked (blacklisted)

Software Defined (SD)-WAN:

Management and Orchestration Plane:

vBond Orchestrator: orchestrates the control and management plane and is the first point of authentication. Enables zero-touch provisioning, which means that new network devices can be connected and require no configuration. The vBond Orchestrator sits in the internet, requires public IP address, and facilitates NAT traversal

vManage: Place where one can connect to configure and manage the whole network. Policies and Templates can be created, software-updates initiated and configurations managed. Graphical User Interface with role-based access control (RBAC).

Third-Party Automation: vManage has a REST-API that allows Third-Party Applications control over all aspects of vManage administration. It is a secure HTTPS interface with standard REST methods GET, PUT, POST, DELETE.

vAnalytics: not really part of Management plane, but provides a strategic overview that can be used to manage a network. vManage receives telemetry information from all devices from the Bidirectional Forwarding Detection (BFD), which provides information about jitter, latency and packet loss, and organizes those into better interpretable graphs. Helps with detection of anomalies in the network in order to respond to potential failures quickly, and provides forecasting info over data-extrapolation or AI, that helps a network admin to predict & prevent

Control Plane:

vSmart Controller: Receives routing updates of the nodes and manages the routing table, that he distributes to the vEdge Routers. He facilitates fabric discovery and implements control plane Policies. This Controller must always be available, hence highly Resilient. In normal SD-Access WAN there are 3 vSmart Controllers running at the same time.

Data Plane:

vEdge Router: Data plane are the Devices located in the branches and receive the routing information from the Control Plane, the WAN Edge routers. These devices are called vEdge Routers and use traditional routing protocols such as OSPF, BGP, and VRRP

Software Defined Wide Area Network (SD-WAN)

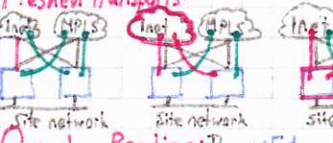
Overlay Management Protocol (OMP): No suitable Protocol for SD-WAN before → Cisco created OMP. TCP-based extensible control Plane protocol runs between vEdge Routers and vSmart Controllers inside a permanent TLS/DTLS connection with SHA256 and AES-256-GCM. The vSmart Controllers form a full mesh of OMP peers, whereas vEdge routers are not required to peer with all vSmart Controllers.

Service Side: Is where customers are connected. VPNs that are like VRFs can be created that provide a segmentation of network traffic for the different customers. VPNs 1-511 are used for this cause.

Transport Locator (TLOC): Are the Identifications of the IPSec Tunnel Endpoints, TLOC denoted by a Triple consisting of the system IP Address of the OMP Speaker, a color, and an encapsulation type. The different connections that exist between the network nodes go over at least 2 different networks to provide resiliency. Each color represents a different network, and each node per default has a full meshed connectivity to all network nodes.

TLOC Extensions - Path redundancy: vEdge routers use Bidirectional Forwarding Detection (BFD) to easily detect failures in IPSec tunnels. BFD sends hello packets that time out when the traffic is rerouted over the secondary transport network. If vEdge router is connected only to a given transport, two vEdge Routers establish a connection to each other over the TLOC Extensions to have access to a secondary transport network. This ensures the circuit-, transport- and router failure protection even though the transports are not meshed.

Meshed transports:

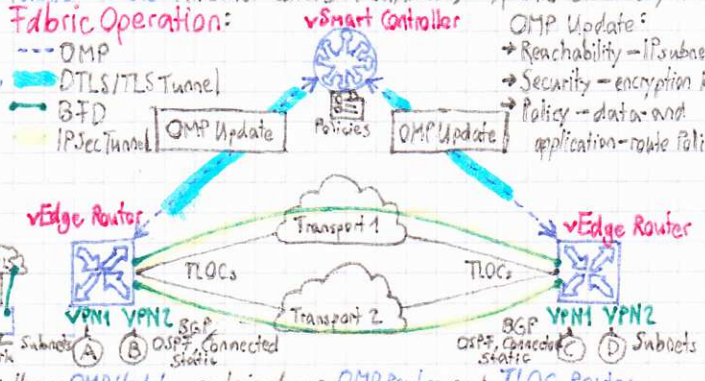


TLOC Extensions:

Transport side: Is the underlay network, always VPN 0 in the Cisco SD-WAN solution. Traffic is always tunneled and encrypted, and has at least 2 transports because the network shall not be dependent on one.

TLOC Colors: Specific Colors that are used are either categorized as private (private network) or public (over Internet). Tunnels can only connect equal colors, and if a private network is routed into the Internet the best practice is to restrict any unequal color combinations.
Private Colors: mpls, private 1-6, metro-ethernet
Public Colors: All other Colors: (red, blue, ..., public-ethernet, etc)

Fabric Operation:



Overlay Routing: The vEdge sends OMP Updates to the vSmart Controller. In those OMP Updates contained are OMP Routes and TLOC Routes.

TLOC Routes: Identify Transport locations, locations in the overlay network that connect to physical transport, such as the point at which a WAN interface connects to a carrier. Also advertised to a vSmart Controller. Attributes: Site-ID, Encap SPI, Encap-Authentication, Encap-Encryption, Public IP, Public Port, Private IP, Private Port, Tag, Weight

OMP Routes (vRoutes): Responsible for the ability of the vSmart Contr. to learn the topology of the overlay network and services available in the network. Most important Attributes: TLOC (next-hop); Site-ID; Label; VPN-ID (from 0 to 65530); Tag; Preference; Originator System IP; Origin Protocol; Origin Metric.
OMP best path Algorithm: 1. Next-hop TLOC is reachable; 2. Prefer vEdge-sourced route over vSmart-sourced route; 3. Prefer OMP-Route with lower admin-distance. 4. Prefer OMP-Route with higher route-preference; 5. Prefer OMP-Route with higher TLOC-preference; 6. Prefer lowest origin: Connected, NAT, static, EBGP, OSPF Intra, OSPF Inter, OSPF External, iBGP, Unknown/Unset; 7. Prefer route from higher Router-ID (System-IP); 8 Prefer highest TLOC private IP Address

Service-Side (Customer-side) implementation: Either OSPF or BGP can be used for service-side routing. Routes of both can be redistributed into the OMP network.

Loop Prevention: done with Down-bit (DN) in OSPF, Extended Community Site of Origin (SoO) in BGP

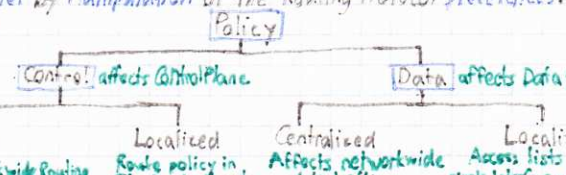
OMP automatical Redistribution works for following types, that are learnt locally or from its routing peers: Connected, Static, OSPF intra-area, OSPF interarea. Some types require Explicit configuration for Redistribution: BGP and OSPF External routes.

Site Redundancy and High Availability:

Routed Site Design: On L3 the redundant vEdge connections are easily managed with a routing Protocol such as OSPF/BGP that provide loop-prevention, load balancing (ECMP) and Bidirectional Redistribution to OMP. Also, one vEdge could be configured as primary and one as failover by manipulation of the Routing Protocol preferences.

Bridged Site Design: For smaller Sites/branches that have no routing protocol running a L2-Solution exists. Virtual Router Redundancy Protocol (VRRP) runs between vEdge routers and defines one active and one Standby gateway. In case of failover, the new active VRRP vEdge sends out gratuitous ARP to update the ARP tables on the hosts and the MAC table on the intermediate L2 switches.

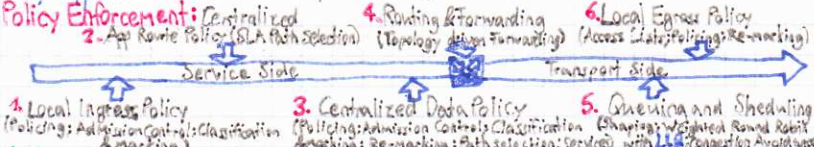
Policies:



Control Policies: Powerful tool, can change routing behaviour for entire SD-WAN fabric.
• Service chaining; • Traffic Engineering (Problem: Prefer the main data center (DC) over the data recovery (DR)-DC. If main DC fails → Reroute to DR-DC; Solution: Control Policy to influence TLOC priority); • Extranet VPNs; • Service & path affinity; Arbitrary VPN topologies (Problem: Must provide various VPNs with different connectivity, based of applications being serviced: VPN1 (CRM System = Hub and Spoke); VPN2 (Voice = full mesh); Solution: VPN1 (vSmart advertises DC prefixes to spokes only and denies everything else); VPN2 (no filter; advert. all prefixes to every node); Hierarchical Policy (Problem: BFD usually fully meshed → scales badly. By limiting BFD sessions to intraregions and between hubs → better; Solution: Overwriting TLOCs of subareas with a single TLOC → TLOC Summarization

Policy Definition: For configuration of vSmart policies three building blocks required: Groups of Interest (Prefixes, Sites, TLOC, VPN, Colors, SLAs) → What should be influenced; Policy Definition (Control Policies, AAR Policies, Data Policies) → Definition of policy to be implemented; Policy Application → Where is the Policy Attached?

Policy Enforcement:



Internet Breakouts:

Direct Internet Access (DIA): Normally remote sites need access to internet via DC, which is enforced bc. of Security Considerations. With SD-WAN multiple internet connections can be established over same or different ISPs and selected and traffic at no risk can then be routed into the Internet directly.
Example: If a remote site wants to provide a Guest Wi-Fi without placing its infrastructure at risk, by isolating guest users from corporate users. Can be done by deploying a data Policy in the guest VPN with a NAT. The Policy force matches all traffic in guest Wi-Fi VPN through a locally defined NAT on the transport-side interface.

Centralized Policies: Affect networkwide routing/data traffic. Applied onto the vSmart Controller by the vManage; Types: • Centralized Control Policy (Overlay Routing); • Centralized Data Policy (Overlay Data plane); Centralized App-Aware Policy (Application-SLA); VPN Membership (Overlay Routing & Segmentation)

Localized Policies: Affect only locally. Applied onto vEdge routers by the vManage; Types: • Localized Control Policy (OSPF & BGP); • Localized Data Policy (QoS, Mirror, and ACL)

Data Policies: look at traffic and parts of IP packets, such as Src.Adr/Dest.Adr, Ports, Protocol Nr. and DSCP values. Services: • Service Chaining (based on traffic characteristics); • NAT; Traffic policing & counting; Transport Selection and traffic engineering (Problem: Must send critical applications over MPLS transport and non-critical applications over Internet transport; Solution: Deploy data policy to set transport for relevant traffic (Bidirectionally set up a local TLOC for desired traffic, override OMP routing decision. Fallback on overlay routing on fail) • QoS classification;

Inbound Policies: Determine which routes are installed in local routing DB of the vSmart Controller

Policy Attachment: Can be applied to OMP Updates on the outgoing or incoming line of the vSmart. If applied on vSmart → Centralized Policy; vEdge → Localized Policy; on vEdge application on Tunnel side or Service side. Distribution handled by vManage which sends NETCONF/YANG packets to either the vSmart Controller or the vEdge Router

Application Aware Routing (AAR): Killer Criterion; Reason for companies to switch to SD-WAN. vEdge routers perform path liveliness & quality measurements via BFD. AAR can use info for routing depending on App requirements and current network performance. Only Service-side traffic considered and Max 4 SLA classes per vEdge possible. Uses as many tunnels as available with load balancing; if no tunnel matches SLA next best is used.
AAR Config: 1. Define SLA params (Jitter, Delay, Packet loss); 2. Define Application list; 3. Define sites, prefixes and VPNs; 4. Define AAR Policy; 5. Apply AAR Policy to selected sites

Quality of Service (QoS)

Description / Goal:
Primary Goal → provide priority (including bandwidth, jitter control and latency) to selected types of traffic. QoS = good user perception, can be compromised by speed mismatches in the network which create bottlenecks. Routers have incoming buffers, only for transit → can not be filled up by a CPU overload. outgoing buffers get filled up with packets waiting for transmission. They require a queue management that can prioritize packets to avoid congestion of time critical ones

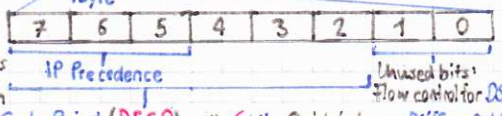
Problems: **Low Throughput:** Bandwidth assigned to given data stream too low for service to useful
Dropped Packets: If router under heavy load, might decide or is forced to drop packets. TCP Application retransmit but delays occur. Packet Loss Ratio (PLR) is percentage of packets lost src to dst.
Errors: Packets sometimes get corrupted due to noise / interference, especially in Wifi → Retransmission
Latency: time interval or delay when system component is waiting for other system component to do something. Total time it takes data packet to travel from Node to Node
Delay: Queuing packets can cause delays because new packet not processed before previous are finished. Types: End-to-end delay = time needed for destination Application to get packet generated by source Application; Network delay (one-way delay) = time difference between first bit put on wire and last bit of packet is received (Sub-Types: Transmission delay = time it takes to transmit packet into wire; Packet Processing delay = time to process packet at network device; Propagation delay = time of signal to travel
Jitter: Packets of same stream can reach dest. with different delays. This Variation in delay known as jitter. Only relevant for time-sensitive applications with continous data stream
Out-of-order Delivery: Packets can take different routes and therefore can arrive out of order. TCP reorders packets on the network layer, but other protocols require higher layers

QoS Models: **Best-Effort:** Default queuing model for interfaces. All packets are treated same way. No QoS happening. Internet is a best-effort network → must be bc. Net-neutrality
Integrated Services (IntServ): Provides way to deliver end-to-end QoS that real-time applications require by explicitly managing network ressources to provide QoS to specific user packet streams, sometimes called microflows. fine-grained QoS System.
Differentiated Services (DiffServ): Soft QoS Approach, depends on network devices that are set up to service multiple classes of traffic, each with different QoS requirements Provides "almost guaranteed" QoS while cost-effective & scalable. Coarse-grained QoS System.

QoS Classification: QoS with DiffServ requires different classes of traffic that can be individually defined, but best-practice is to use [RFC 4594] Marking recommendations which define type of content & exact markers used with specific traffic type. Number of classes can be different and different mapping strategies exist

Marking at Layer 2: DA | SA | VLAN Tag | Etype | Data | FCS
Marking done in VLAN [802.1q]
Header. 3 bits reserved for User Priority [802.1p] — TPID | User Priority | CFI | VLAN ID

Marking at Layer 3: Version | length | TOS | LEN | ID | Offset | TTL | Proto | FCS | IP-SA | IP-DA | Data
Marking happens in Type-Of Service (TOS) field of the IP Header with 8 bits. Contains IP Precedence overlapping with the Differentiated Services Code Point (DSCP) with 6 bits. Outdated new DiffServ field (ECN)



Network-Based Application Recognition (NBAR): Classification Engine that recognizes and classifies wide variety of protocols & applications, including web-based and other difficult-to-classify applications and protocols that use dynamic TCP/UDP port assignment. Traffic classified based on application signatures using deep packet inspection at L7.

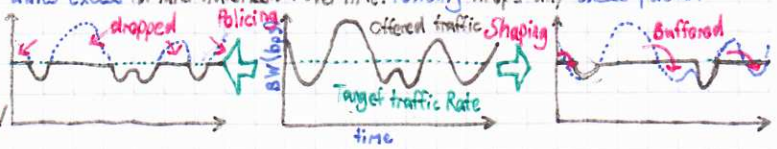
QoS Configuration: three main components:
1. **Define Classes of Traffic:** Each class of traffic defined using a class-map
   - Definition Example: class [class-name]
     class-map match-{all (all matches must be fulfilled) | any (any match fulfilled)} [class-name]
   - Matching Example: match dscp [dscp-value,...] \n match precedence [precedence-value,...] match cos [CoS-value,...] \n match access-group [access list] \n match protocol [protocol/application-name] (NBAR)
   - Marking Example (used when QoS marking of applications not trusted): set ip dscp [dscp-value] \n set precedence [precedence-value] \n set cos [CoS-value]
2. **Define QoS Policies for Classes:**
   - Definition Example: policy-map [policy-name]
   - Algorithms Example: priority percent 30 (Allocates 30% of interface BW) bandwidth {remaining} percent 5 (remaining optional → else of total BW) \n fair-queue
   - Policing Example: Police cir 10'000'000 bc 50'000 conform-action transmit exceed-action drop (Means maximally 1 Mbit/s can be consumed with a maximum burst of 50'000 bits. allowed. Burst size is the measurement period during which the avg rate is measured and calculated. All exceeding traffic is dropped
   - Shaping Example: Shape average 50'000'000 (Means only 50 MBit/s can be consumed on average. When traffic exceeds the allocated 50 MBit/s the traffic is going to be buffered and resent later.
3. **Apply a Service Policy:**
   - Incoming Buffer: QoS classification, marking and policing
     Example: interface [interface] \n service-policy input [policy-map-name]
   - Outgoing Buffer: LLQ, CBWFQ, policing and shaping
     Example: interface [interface] \n service-policy output [policy-map-name]

Types of Traffic: Different types of traffic have different sensitivities to problems on left-side
Voice Traffic: Very sensitive to delays, jitter and dropped packets. Voice packets are not retransmitted if lost; Requirements: Latency under 150 ms; Jitter under 30 ms; Packet-Loss under 1%; Bandwidth at least 30 Kb/s
Video Traffic: Very sensitive to dropped packets. Has high volume of data Requirements: Latency under 400 ms; Jitter under 50 ms; Packet-loss under 1%; Bandwidth at least 384 Kb/s; Must be a series of equally spaced packets. With a Dejitter Buffer this can be accomplished. Before Video gets streamed the minimum playout Buffer must be full, only then buffer starts sending packets constantly for playback
Data Traffic: insensitive to drops & delays. Ensured by TCP by re-sending & re-ordering

Queuing algorithms: **First in First Out (FIFO):** No concept of priority or classes of traffic. Fastest queuing method; effective for large links with little delay and min. congestion
Priority Queuing (PQ): Uses multiple queues and allows prioritization. Packets assigned to respecting queues according to priority. Highest priority queue always worked on first, and only if empty next lower queue gets emptied. Starvation of lower queues if higher always
Round-Robin (RR): Uses multiple queues without prioritization. Each Round one packet of each queue
Weighted Round-Robin / Weighted Fair Queuing (WFQ): Uses multiple queues, with a priority / weight defined for each. WFQ Scheduler goes from queue to queue in circles and sends from current queue until empty or as many packets as queue weight is
☺ No Starvation can happen because Scheduler goes to lower priority based on packet-count
Class-Based Weighted Fair Queuing (CBWFQ): Extends standard WFQ functionality to provide support for user-defined traffic classes. Customization with max-bandwidth (BW) queue limit (max packets in queue) and max-bandwidth-percentage. BW determines packet order
Low Latency Queuing (LLQ): Provides strict priority queuing for CBWFQ. Allows delay-sensitive data such as voice to be sent first → strict priority. Is a PQ where first queue always has priority and second queue is queued with CBWFQ.

Queue Management: if queues are full different queue management strategies ensure that packets still can be prioritized even though buffers are full
Tail dropping: If a queue is full there is no way to buffer other incoming packets, they get dropped. No differentiated drop mechanism → Premium traffic dropped same as best-effort
TCP global synchronization: Normally TCP sessions start at different times, but as TCP-windows are increased, tail drop may cause packets of multiple session to be dropped at the same time, which results in a TCP Restart at the same time
TCP Starvation: Happens when TCP and UDP flows are present during congestion. If TCP global synchronization happens → TCP Sessions slow down → UDP fills queues → TCP packets dropped
Random Early Detection (RED): Provides congestion avoidance by controlling average queue size. Random drops of packets cause TCP to resize windows while avoiding TCP Synchronization. Doesn't work well with Voice or other non-TCP Applications
Weighted Random Early Detection (WRED): Packets dropped selectively based on the IP precedence, DSCP or EXP. Lower priority classes start random drops before higher priority. If queue below minimum threshold → no drops. If queue fills up to max-threshold, a certain percentage of packets are dropped. Percentage determined by Mark Probability Denominator
Example: class TCP_PORT443 \n random-detect \n random-detect precedence 5 30 90 4
For Precedence 5 WRED drops 1 in every 4 packets when queue length = 40. Amount of packet drops increases linear between the min (30) and the max (40) threshold. at queue-length 35 half a packet in every 4 packets dropped → 1 out of 8 packets

Shaping vs. Policing: Traffic Shaping buffers excess packets in a queue and re-schedules excess for later transmission over time. Policing drops any excess packet



Campus Network Design

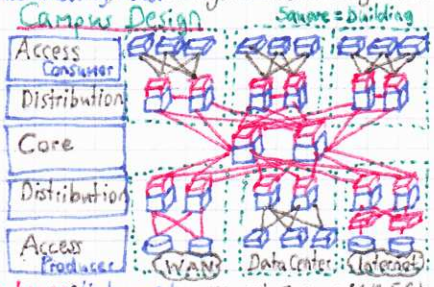
Technical Requirements:
Scalability, planning for expansion (BW need, future growth) Constraints (Broadcasts, flat networks, addressing limitations); • Availability (most important); • Performance; • Security; • Manageability • Usability; • Adaptability; • Affordability

Network Design: **Flat Design:** Used where only few hosts (e.g. up to 100) and when any-to-any communication is mandatory (LANs, any-to-any WANs (MPLS VPN), between Servers)

Hierarchical Design: Divide & Conquer mentality, split up network in different layers, which allow easier fault isolation & troubleshooting; Used in large scale networks (e.g. Tel.net)

Offers: • Hierarchy (Each layer has role); • Modular Topology (Building Blocks); • Easy to grow, understand & troubleshoot • Small Fault domains (clear demarcation) • Promotes load balancing & redundancy • Promotes deterministic traffic patterns • Leveraging L2 & L3 strength • Uses L3 routing for load balancing, fast convergence, scalability & control

Campus Design



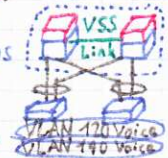
Access Layer: **Building Access Layer:** High speed access ports for users (1/2/5 Gbps) High Density but low cost; Network Access-Control (NAC); Dynamic provide correct VLAN; QoS, trust boundary / Classification; IP Multicast, broadcast supression, IGMP Snooping; Uplink-Services (All links actively forwarding, fast uplink failover, Bandwidth (BW) scalability.

Access Switch Requirements: Port Security based on MAC / certificates; Host in correct VLAN; Certain Hosts (IP Tel. / WLAN AP) need 802.1q trunk; DHCP; QoS marking; DHCP Snooping / Root Guard

Design within a building block: **L2 Access - No VLAN Spanning → V-Topology**
• Load balance on uplinks; Redundancy on L3; HSRP for load balancing and first-hop redundancy; No need for L2 link between Distr. Switches; STP Root and HSRP Primary tuning; Root Guard / BPDU-Guard; Used for STP Avoidance → requires Star Topology between Access and Distribution Layer; Point-to-Point L3 between Distr.

L2 Access - VLAN Spanning Required → Y-Topology
• Layer 2 link between Distribution Switches creates loop • STP blocks 50% of links, has high convergence time • STP Root and HSRP primary or GLBP and STP port cost tuning to load-balance on uplinks; Root-Guard or BPDU-Guard

Routed Access and Virtual Switching Systems (VSS): Used for environments that need VLAN Spanning; VSS turns two Distr. Layer Switches into one logical → No Loops EtherChannel from Access Switch to VSS Distr. Switch. Default GW on Access Switches → load balancing.



Distribution Layer: **Building Distribution Layer:** Aggregates wiring closets (access layer) and uplinks to Core; Layer 3 Boundaries: (Protects Core from high density peering and problems in Access Layer; Aggregates VLANs and provides inter-VLAN Routing; Routing to other buildings); First-Hop redundancy; Route summarization, fast convergence, redundant path load sharing; Availability, load Balancing, QoS; Implement Security Policies;

Distribution Switch Requirements: Implements the "intelligence" in Campus Design • terminate L2 domains (i.e. first-hop, i.e. Default GW for clients), that possible failures within a broadcast domain can't spread over the building blocks; • first-hop redundancy (HSRP, VRRP, GLBP); IP filter (access-lists) to allow/forbid traffic between VLANs; Fast-convergence for STP; Summarization of IP-addresses → Routing Tbls stay small

First Hop Redundancy: Used to provide a resilient default GW / first hop address to end-stations; VRRP, HSRP & GLBP provide millisecond timers and excellent convergence performance; Preempt timers tuning to avoid black-holed traffic

First Hop Redundancy with Virtual Router Redundancy Protocol (VRRP) (open standard)
Group of Routers functions as one virtual Router by sharing virtual IP, Each Router keeps own MAC; Master (VRRP Active) performs all packet forwarding; Rest of Routers (VRRP Backup) stay idle; In case of failover the virtual IP is transferred to other Router. Client receive new MAC Address via ARP request

First Hop Redundancy with Hot Standby Router Protocol (HSRP) (Cisco only)
Group of Routers functions as virtual Router with virtual IP and virtual MAC Because of virtual Mac not ARP Request required on failover → faster than VRRP HSRP Active Router performs packet forwarding, rest of Routers (HSRP Hot-Standby) stay idle

First Hop Redundancy & Load Balancing with Gateway Load Balancing Protocol (GLBP)
Group of Routers functions as virtual Router by sharing virtual IP but using multiple virtual MAC adresses for forwarding; Allows traffic from single common subnet to go through multiple redundant GWs using single virtual IP Address Routers return on ARP-Requests alternately their MAC → per device load balancing

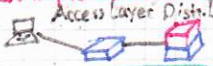
Availability Calculation:

$$\text{Availability} = \frac{MTBF}{MTBF + MTTR}$$

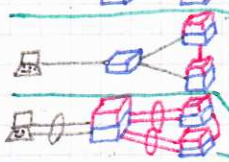
$$MTBF_{1..n} = \frac{1}{\frac{1}{MTBF_1} + \dots + \frac{1}{MTBF_n}}$$

Mean Time Between Failure (MTBF): time between a network becoming operational and the next failure; For Equipment, Individual Links, human actors
Mean Time To Repair (MTTR): Higher Network complexity results in higher MTTR
Redundancy: Increasing resiliency increases MTBF, but also the MTTR because of more complex routing; Thereby the convergence times are increased which are directly tied to the MTTR.

Example Calculation: First MTBF calculation of whole system



$\frac{1}{10k} + \frac{1}{180k} + \frac{1}{100k} + \frac{1}{320k} = 8426h \rightarrow A = \frac{8426h}{8426h+4h} = 99.952\%$



Calculation of Distr. Switch & Access-Distr Link Connection of Combination in parallel results $\frac{1}{100k} + \frac{1}{320k} \approx 76k$
in a 50% increase $\frac{1}{MTBF}\left(1+\frac{1}{2}+\frac{1}{3}\right) = 76k \cdot 1.5 \approx 1.14k$
This in Series with Access Switch & PC to Access Link = Total MTBF
$\frac{1}{1.14k} + \frac{1}{180k} + \frac{1}{10k} \approx 8748h$

Access Switch: 180k h MTBF
Distr. Switch: 320k h MTBF
Link PC to Access: 10k h MTBF
Link Access-Distr.: 100k h MTBF
MTTR = 4h

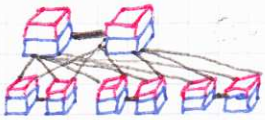
Access-Distr Link doubled = increase of 50% $100k \cdot 1.5 = 150k$
Combination of Access-Distr with Distr. Switch increase by 50% bc. double ← $\frac{1}{\frac{1}{150k} + \frac{1}{320k}} = 102kh$
PC to Access Link doubled = 15k $\frac{1}{10k} + \frac{1}{10k} = 5kh$
All together $\frac{1}{\frac{1}{153k} + \frac{1}{180k} + \frac{1}{15k}} = 12kh$ $\frac{12kh}{12kh+4h} = 99.96\%$

Increase Overall Availability Most by exchange of weakest link

Core Layer: **Building Core Layer** (it's all about speed): Backbone of the network: Connects Network building blocks; Redundancy, Extremely fast convergence, QoS, Redundant links but no STP; No complexity and No Services

Split Layer 2 Core: Cheap bc. switches used instead of routers; Routing Protocols on Distr. Routers install 2 equal-cost paths → near instantaneous convergence; Scaling Limitations

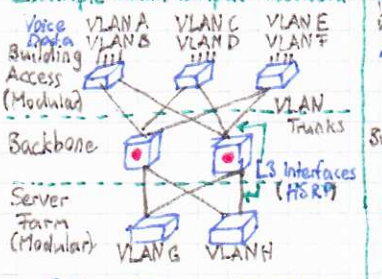
Redundant Layer 3 Core: Expensive bc. Routers used; Each Link in own /30 Subnet; No L3 link required between Distr. Switches. No STP in Core - all routed; Load Balancing
If Summarization enabled, L3 Link between Distr. Switches Required.



Core Layer: Easier adding modules; Fewer links in core; Easier BW upgrade Routing Protocol peering reduced; Equal cost Layer 3 Links → best convergence
vs. No Core Layer: Fully meshed Distr layer; Physical cabling requirement; Routing Complexity

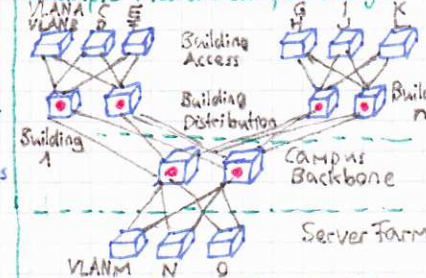
Best Practices: Layer 3 Routing Protocols: Typically deployed in distr. to core and core to core interconnections; Used to quickly re-route around failed node/link while providing load balancing; Build triangles not Squares for deterministic convergence (Does not require re-convergence on Link / Box failure);
Limit OSPF & EIGRP Peering (only peer on links intended to use in transit);
Insure redundant L3 paths to avoid black-holes)
Summarize Distribution to Core to limit EIGRP query diameter / OSPF LSA propagation

Example: Small Campus Network



Collapse Core and Distribution into Backbone.

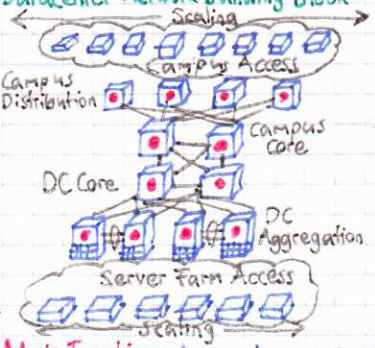
Example: Medium Campus Design



Distinct Access, Distribution & Core

Datacenter Network Design

Datacenter Network Building Block:



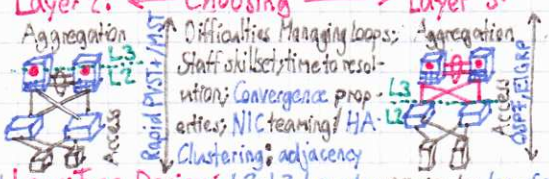
Access Layer Design:

Layer 2: ⊕ Plug-n-Play (No user config required to build forwarding DB; Simple to Support teaming or L2 Multicast clusters; Easy traffic segmentation with VLANs; Very fast movement of end-station adresses
⊖ MAC addr. consumption; BPDU generation is CPU intensive with increasing number of VLANs; VLAN sprawl causes flooding and bcast to propagate unnecessarily; 50% of links blocking; Miscoring can cause L2 Loops

Layer 3: ⊕ Availability & Scaling; L3 routed topologies reduce consumption of L2 tables via route summarization; Minimizes bcast domains → high level of stability; Meet server stability requirements or isolate particular Appl. Envir. Creates smaller Failure Domains → ↑ stability; All uplinks are available, no blocking (ECMP Max); Fast Uplink Convergence: failover and failback, no ARP-Table to rebuild for Aggregation Switches
⊖ L2 adjacency limited to access pairs (clustering & NIC Teaming limited); IP Address Space management more difficult than L2 Access; Migrating to L3 Access IP changes difficult; Requires services to be kept at each access Layer pair

Main Functions: Aggregate servers redundantly (NIC Teaming); Server clustering Broadcast containment; L2 topologies consist of looped & loop-free models
VLANs are extended across inter-switch link trunk for looped access
VLANs are not extended across inter-switch link trunk for loop-free access
Looped Design: VLANs extended between aggregation switches, creating Looped topology; STP to prevent loops (PVST+, MST); Redundant paths but one blocking; VLANs may be load balanced across access layer uplinks; inter-switch link utilization must be considered as may be used to reach active services

Layer 2: ← Choosing → **Layer 3:**



Loop-Free Design: L2-L3 boundary varies by loop-free model used: U / Inverted U
Implication considerations with service modules, L2 adjacency, and single attached servers

L2 Looped Triangle Topologies

Supports VLAN spanning; Resiliency with Dual Homing & STP; Quick convergence with RSTP Supports Stateful Services in Aggregation Layer; Widely Used

L2 Looped Square

Supports VLAN spanning; Resiliency w. Dual Homing & STP; Quick convergence w. RSTP; Stateful Services in Agg Layer; Higher Density Access Layers;

Looped Design Benefits: Services (like firewall & load balancing) easily deployed at Aggregation Layer and shared across multiple Access Layer switches; VLANs primarily contained between Access switch pairs but can be extended/spanned to other access switches to Support: NIC Teaming, Clustering L2 Adjacency

L2 Loop-Free U Topologies

VLANs contained in switch pairs (no spanning); No STP blocking; all uplinks active; Some implications for certain service modules (Sync via Access)

Loop-Free Benefits: STP is enabled but no port in blocking state, so all links are forwarding

L2 Loop-Free Inverted U Topologies

Supports VLAN spanning
No STP blocking; all uplinks active
Access switch uplink failure blackholes single attached servers
Supports all service modules

Inverted U Drawback: Black-holing
Distributed Etherchannel reduces chances
NIC Teaming improves resiliency; Inter-Switch link scaling consideration using active-standby

Aggregation Layer Design:

Main Functions: Aggregating Sessions leaving & entering Datacenter; Must be capable of supporting many 10 GigE and GigE interconnections; Provides services (e.g. load balancing, firewall, SSL offloading) to servers across Access Layer
Aggregation Switches carry the workload of STP Processing and Default GW Redundancy Protocol Processing; Most critical Layer in Datacenter because port density, Oversubscription values, CPU processing, and service modules introduce unique implication into overall design.

Active-Standby Service Design: Services: Application control engine, Firewall service module, SSL Module; Under Utilizes access Layer Uplinks; Under Utilizes service modules and switch fabrics; Services only Active on one switch of Aggregation Switch Pair; No Load Balancing Possible

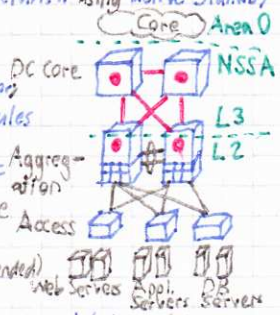
Active-Active Service Design: Services: Application control engine, SLB + SSL + FW, 4G/8G/16G switch fabric options, Active-Standby distribution per context, Firewall service module; Permits Uplink Load Balancing
Increases overall service performance

STP, HSRP and Service Context Alignement: Align server access to primary components to aggregation layer: VLAN Root, Primary HSRP Instance, Active service context; Provides more predictable design; Simplifies troubleshooting
More efficient traffic flow; Decreases chance of flow ping-pong across inter-switch

Scaling Aggregation Layer: Aggregation modules provide: STP Scaling, HSRP scaling, Access Layer Density, 10 GigE uplinks, Application service scaling, SLB/Firewall, Fault domain sizing

Core Layer Design:

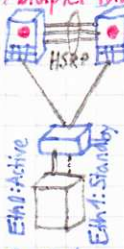
Main Functions: High speed packet switching backplane for all flows going in and out of Data center; Provides connectivity to multiple aggregation modules and resilient L3 routed fabric with no single point of failure; Runs an interior Routing Protocol, such as OSPF or EIGRP, and load balances traffic between the Campus Core and the Aggregation layers; L2/3 Boundaries (L2 extensions through core not recommended)



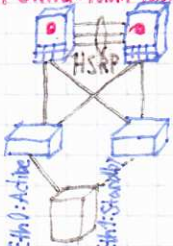
Routing Protocol Design:
NSSA helps to limit LSA Propagation, but permits route distribution; Advertise default into NSSA, summarize routes out; OSPF default reference both ways is 100M, use auto-cost-reference-bandwidth set to 10G value; VLANs on 10GE trunks have OSPF cost = 1G; Loopback interfaces simplify trouble shooting; Use passive-network default; Use authentication → avoids undesired adjacencies; Timers Shortest path first = 1/1 hello/dead = 1/3

NIC Teaming Configuration:

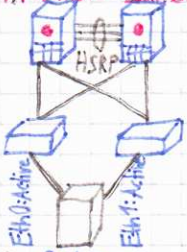
Feature which provides Network redundancy & load balance traffic. At least 2 Network Interface Cards (NIC)s required. Groups multiple NICs into one single logical NIC which helps when NIC fails to load or network fails. (TLB)
Adapter Fault Tolerance (AFT) | Switch Fault Tolerance (SFT) | Transmit Load Balancing



On Failover:
Src MAC Eth1 = Src MAC Eth0
IP Addr Eth1 = IP Addr Eth0
Failover between two network adapter connected to same switch



On Failover:
Src MAC Eth1 = Eth0
IP Address Eth1 = Eth0
Failover between two network adapter when each connected to separate switch



One Port Receives, All Ports Transmit.
Incorporates Fault Tolerance. One IP Addr. Multiple MACs.

Link Aggregation Protocol (LACP):

Features: Max bundled ports 1 to 8; LACP Packets sent with Multicast group MAC Addr. 01:80:c2:00:00:02; During LACP detection period packets transmitted every second;

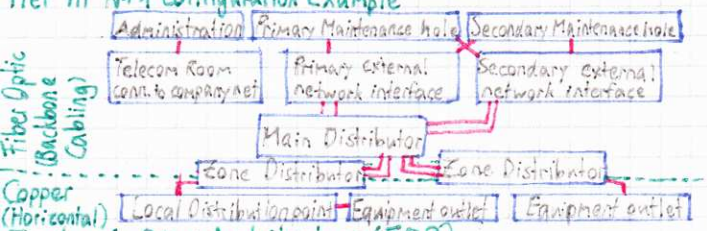
Further Description on Network Virtualization (Top-left)

Datacenter Network Design

Tier Classifications

	Tier I	Tier II	Tier III	Tier IV
Site Availability:	99.671%	99.749%	99.982%	99.995%
Redundancy:	N	N+1	N+1	2(N+1)
Yearly Downtime:	28h 50min 22s	22h 7s	1h 34m 40s	26m 17s

Tier III N+1 Configuration Example

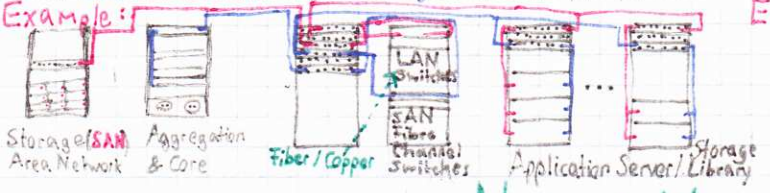


End-of-Row Architecture (EOR)

Each server is rewired to connect to a common aggregation switch directly. Racks lined up side-by-side in a row and a rack at either side for providing network connect. EOR treats entire row like one pod → Upgrades or issues in network affect entire row

Advantages: less switches → less costs for devices & ext. maintenance; higher Port utilization, especially if not all racks full; Switches located at one location → simplifies maintenance & handling; Better L2 Availability between racks; **Drawbacks:** higher cabling cost (additional patchpanel)

Example:



Network Redundancies

N: Stands for baseline capacity. A Facility with N redundancy has everything it needs to operate as designed, but has no redundancies to accomodate equipment failure

N+1: Means that a facility has the capacity needed to run at full IT load, with an additional component to account for failure or maintenance. N+1 redundancy standards typically require require extra unit for every four needed. If 12 Cooling units needed an N+1 facility would have 15

2N: System is fully redundant, with a independent, mirrored system that can fully take over operational needs should the the first system go offline for any reason. These systems can easily be maintained because one system can be shut down and mirrored can overtake

2(N+1): Provides a completely paralleled backup system along with additional components to account for failure & maintenance in each system.

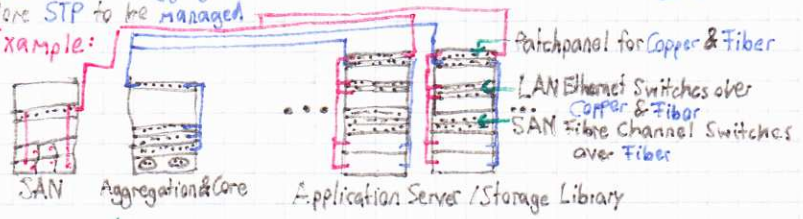
Top-of-Rack Architecture (TOR)

Servers connect to Ethernet switches installed inside the Rack → all copper cabling for servers stay within rack. Ethernet switch links rack to the DC Network with fiber running directly from rack to a common aggregation area. Rack managed individually → Upgrades or issues in network only affect servers within one rack.

Advantages: less cabling → less cabling costs; easy expansion/changes per rack; Scalable glass fiber infrastructure (40G & 100G); Ideal for high density (full racks)

Drawbacks: More switches → higher costs for devices & maintenance; More Ports needed in aggregation; More L2 server-to-server traffic in aggregation More STP to be managed

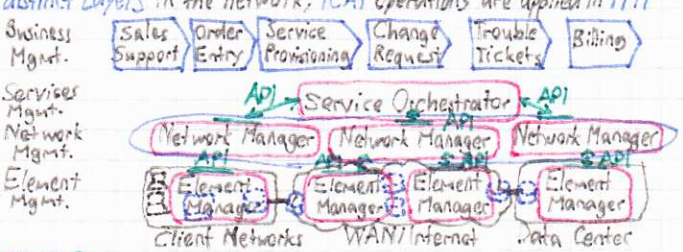
Example:



Network Management

Includes:
Network Provisioning; Network Monitoring; Failure Alerting; Backup/Restoration; Logging

Telecom Management Framework (TMF): Tries to split the distinct Layers in the network; FCAP operations are applied in TMF



FCAP Implementation

Fault Management Implementation: 1. Fault Detection: Enable logging on devices and send to central syslog server, Set log level correctly; Configure snmp traps that send errors to central mgmt server; Implement test robots (application robot/Application response monitoring); Error Correlation (Intelligent Syslog Servers such as Splunk®); Integrate snmp-queries on monitoring system (ELK); 2. Alerting Processes: Monitoring via Alerting Servers; Error Message filtering; Alarm Acknowledgement 3. Fault Isolation: Syslog servers simplify fault isolation; Graphical Mgmt Tools (Nagios, Observium, Icinga, etc.); Tools (Ping, Traceroute) Traffic analysis with IP SLA; Overload Detection (Netflow); Physical property Analysis with WMI, Sniffer if all else fails; 4. Fault Correction: Workarounds required: Reboot, Routing Configuration, Resource Assignement

Accounting Management Implementation: 1. Definition of Chargeable Services: Used Bandwidth or data volume, Number of Transactions, Prioritization of Application/Traffic, Resource consumption 2. Measurement: Network monitoring tools (snmp based; PRTG, Observium) Performance measurement on servers; 3. Report/Invoice Generation: barriers against abuse required (late notice)

Configuration Management Examination Questions: Is there a authorization system for configuration changes (no impersonal pws!)? Is there a device inventory including age, software release, etc.? Are configurations stored centrally? Are there Backups of them? What is process for bringing new equipment. Pre-provisioning? Automation? Are changes logged? Are protocols created? What is the life cycle process? (Operation times before replacement) How are security updates, patches, new releases rolled out?

FCAP:

Fault Management: Process of locating, diagnosing and correcting problems. Increases network reliability and effectiveness, increases productivity of network users.
Components: Fault Detection, Fault Correction, Fault Isolation, Network recovery, Alarm generation, Alarm handling, Alarm filtering, Alarm correlation, Diagnostic test, Error logging, Error handling, Error statistics

Configuration Management: Process of obtaining data from the network and other systems and using that data to manage setup of all network devices; Allows rapid access to config info, facilitates remote config and provisioning, up-to-date Inventory, Config Versions
Components: Resource Initialization, Network provisioning, Auto-discovery, Backup & Restore Resource shutdown, Change Management, Pre-provisioning, Inventory/asset Management, Remote Configuration, Software distribution, Job initiation, Job tracking

Accounting Management: Measuring usage of network resources by users in order to establish metrics, check quotas, determine costs, and bill users. Measures and reports accounting info based on groups & users, administers cost of network (internal) verification of third-party billing for usage
Components: Track Service, Resource usage, Accounting limits, Combine costs for multiple resources, Set quotas for usage, Audits, Fraud reporting

Performance Management: Ensuring that data network remains accessible and as uncongested as possible. Reduces network overcrowding and inaccessibility, provides a consistent level of service to users, determine utilization trends to proactively resolve problems
Components: Performance data collection, Performance data analysis, Utilization- & error rates, Consistant Performance level, Capacity planning, Maintaining historical logs

Security Management: Protecting sensitive info on devices attached to a data network by controlling access points to that info. Builds network user confidence, secures sensitive info from internal & external sources, protects network functionality from attacks
Components: Selective resource access, Access logs, Security alarms, Event reporting User access right checking, Compliance, Security related information distribution

Network Management

Documentation

Element Management Layer: Devices: Routers, Switches, Firewalls etc. (Vendor, Model, Interfaces, RAM, SW-Release, etc.); Geographical information of Devices (Country, City, Building, Room Nr., Rack Nr. etc.)); Server Farms (Load Balancers, Firewalls, Storage, etc.) Connection between devices: WAN: Link Speed, ToS, SP); LAN: L2 technology (e.g. 1000B-T); Encapsulation (e.g. trunk) L1: Copper / Fiber cable → Most errors happen in L1

Wiring and Media: L1 Errors: cable defect, wrong patching, flapping links, wrong power budget

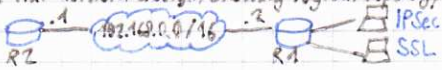
Cable Plan: Includes: Single-Mode fibers, Multi-Mode fiber, Shielded Twisted Pair copper (STP) + cable length, Unshielded Twisted Pair (UTP) + cable length, Coaxial Cable, Microwave / Laser / Infrared / Radio Links / WLAN Access Points

Example:

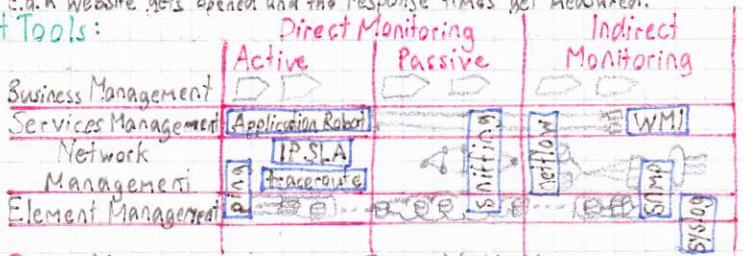
Network Management Layer: Which networks exist?: Routed-Networks; MPLS-VPN, VRF, VXLAN; L2 Networks; Wireless-Networks (AP, Point-to-Point); Internet Access; VPN-Remote

For Troubleshooting L3 Map Required: Including IP network and subnets; Showing hierarchical or flat network design, Showing logical topology (Star, mesh, etc.)

Example:



Management Tools:



Syslog: Standard Logging mechanism in Linux / Unix; Device send acc. of a internally recognized event a Log-Message to a Syslog-Server; For the determination of the significance there are 8 (0-7) severities; To identify the sending instance there is a Facility (5-bit) in the Log-Message; Uses UDP-Port 514; No Security Measures; No Acknowledgement, Log-Messages collected centrally with sender IP; Timestamp; Usage: Defined Severities of events trigger alarms; Error reproduction with ↑

Netflow: Aim: Monitoring of Dataflows in network.

Monitoring Agent (Router): Creates Probes by observing data packets in network; Recognizes identical packets; Creates Flow-Entry in netflow cache; sends Cache-entry to a Monitoring Collector: Collects netflow data of multiple probes; Saves Data in a DB; Examination (Reports, Visualization);

Naming: Caching and Exporting can be a huge burden for a Router

Windows Management Instrumentation (WMI): Developed by Microsoft → and in reality mostly limited to MS-Products; Comparable to snmp; Combined with Scripts can be a powerful tool for a MS-Server-Administrator

Sniffing: Taps: Collecting network traffic by splitting the physical transmission; Real Taps: Each direction sent to separate Port; Aggregator Tap: Both directions → one port

Monitoring Port on Switch: Switch Copies traffic of specified port to another port

⊕ No additional Device required ⊖ No guarantee all packets copied → In case of Overload

Ping: Used to check availability of device / url; Sends packets to specified Destination

Traceroute: Like Ping but lists each Node that decreased TTL → Each node in path

Cisco IP SLA: Part of Cisco IOS software, and allows analyzation of IP service levels for IP Applications and services by using active traffic monitoring for measuring network performance; Simulates network data and IP Services;

Application Robot (Application Response Monitoring): OSI-Layers 4-7 get tested; Active: Measurement Station acts as Application Client; Typical and representative Processes get started; Response times get measured;

Passive: Network packets between End-devices (typically Client-Server) get recorded; With Analysis-Software the partial processes of a complex flow evaluated

Measurement: Very fast very complex; require deep-insight into flows; Requires Real-time data; Problem: how data of different partial processes can be combined

Action: Services Mgmt: Emulation of Clients, Monitoring of Server Software

Element Mgmt: Sniffing & Analyzation of Layer 4-7 packets

Opnet ACE Analyst: Isolate application message flows; Visualize application performance; Drill down into application messages; Understand dependencies impacting performance

Event Monitoring

Happens on Service-, Network- and Element-Management Layers

Direct Information Gathering:

Active: Monitoring runs Service to monitor itself. Monitoring-Station acts as Web-Client

Passive: Monitoring taps into Service to monitor transparent. Taps allow passive traffic sniffing; SPAN-Port-Switch copies traffic to other port, conn. to Monitoring

Indirect Information Gathering:

Most common IT-Monitoring. Monitored Device collects metrics into device DB, which decides what is monitored. DB-Types: SNMP-MIB, Netflow, WMI Repository

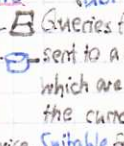
Event Monitoring: Properties of interest get defined; If they happen during normal business they get measured; Monitoring does not interfere → transparent

Procedure in practice: Measured metric saved in DB-variable; Monitoring Service queries DB-var regularly and saves it. Data gets Visualized and Presented

Pull Principle: (UDP Port 161)

Queries for an attribute sent to a device (Request) which are answered with the current value (Response)

Push Principle: (UDP Port 162)



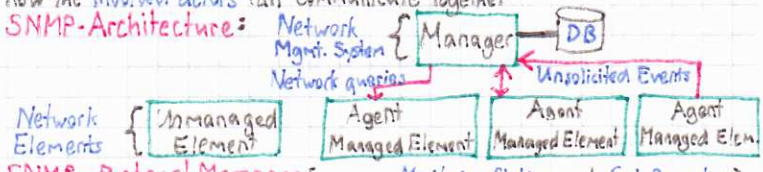
Device sends attribute in SNMP-Trap / Netflow msg to Monitoring Station

Managed Device Suitable for monitoring tasks Managed Device

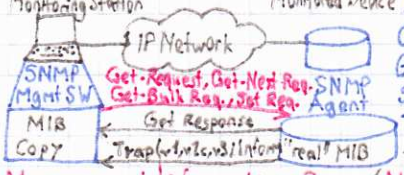
Active Process Monitoring: As a part of Monitoring a Process get actively executed, resembling to the real Process, Attributes defined get measured; E.g. a Website gets opened and the response times get measured.

Simple Network Management Protocol (SNMP): Consist of 3 Actors: 1. Management Station (SNMP-SW on a station from which the Management (Control & Monitoring) gets done); 2. Management Agent (SNMP-SW on Monitored device); 3. Hierarchical DB (Management Information Base); and a Protocol how the involved actors can communicate together

SNMP-Architecture:



SNMP-Protocol Messages: Monitoring Station sends Get-Request → Monitored Device Agent responds with Get-Response;



Get-next Request → Agent sends exactly next val

Get Bulk Request → Agent sends multiple values

Set Request: Config of Agent in MIB

Trap (unacknowledged), Inform (acknowledged) → Agent sends value on Event (Alarm-info)

Management Information Base (MIB): Hierachically structured; Each object uniquely identified; MIB-"Tree" split into partial parts, some assigned by the IETF and some Vendor administered; Each Node represented by descrip. and number; Root-Node = ISO (1); Object Identifier (OID) used as identification scheme, is a ordered sequence of non-negative integers left to right containing at least 2 elements; Each element separated by a point represents a Node of another Hierarchy level: 1.3.6.1.2.1.4 = ISO.Organization.DOD.Internet.etc.

Central Task for Monitoring-Responsible: Loading & Actualizing of all MIBs

CLI / Terminal Session

Drawbacks CLI: Configuration Robots need to send CLI Commands in right order; Commands & Output different depending on SW/HW

No error-handling; CLIs not transaction oriented → No Rollbacks

Drawbacks Config-Files: Configs only saved / loaded as whole; Config files represent IOS Commands → depending on SW / HW;

Abstraction: Explicit mapping from network design to desired state and device config (Separation of infrastructure state & service state); Independent from HW / Vendor / SW Release

Easier to: Validate config compliance, compare current state with desired state, identify mismatches, change device configs