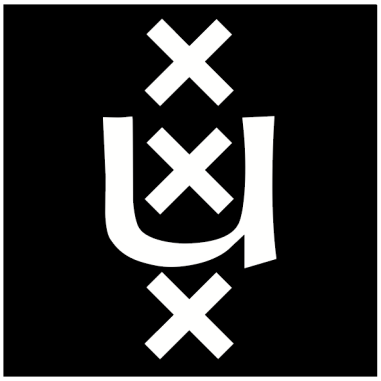
University of Amsterdam

InterNetworking and Routing For Dummies



Xavier Torrent Gorjón

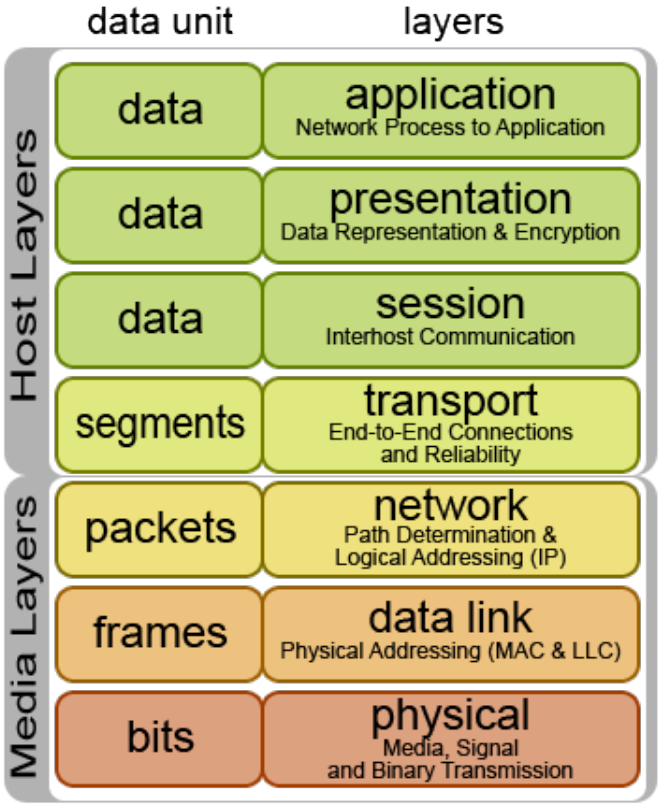
Xavier.TorrentGorjon@os3.nl

March 19, 2015

Contents

1 Overview

1.1 OSI Model



1.2 Interfaces and Protocols

Interfaces Interfaces connect different layers on the same computer. Uses Protocol Data Units (PDU).

Protocols Protocols are used to communicate between parties data on a specific layer. Uses Service Data Units (SDU) inside Service Access Points (SAP).

1.3 Encapsulation and Multiplexing

Encapsulation When data units go down one level in the layer model, headers are added to add information regarding the current layer.

Multiplexing Multiple protocols can coexist on the same layer. However,

when going down the layer model, these protocols should be treated equally. For example, TCP and UDP are multiplexed down at the IP level, and demultiplexed back when reading the information of IP packets.[1]

1.4 ES Models: Strong vs Weak

Strong ES Model Hosts suppress packets with a destination address that references another of its interfaces.

Weak ES Model Hosts accept packets that match with one of its interfaces addresses, even if it does not receive it on that interface.

1.5 IP Addressing (IPv4)

- 32-bit addresses
- Decimal-dotted notation (a.b.c.d, 0 =< a,b,c,d =< 255).
- Special addresses:

  0.0.0.0 IP address unknown.

  127.0.0.1 Loopback address.

  Host part all 0 Subnet identifier.

  Host part all 1 Directed broadcast.

  255.255.255.255 Local subnet broadcast.

- Private addresses:

  10.0.0.0/8

  172.16.0.0/12

  192.168.0.0/16

  169.254.0.0/16

1.6 Subnetting

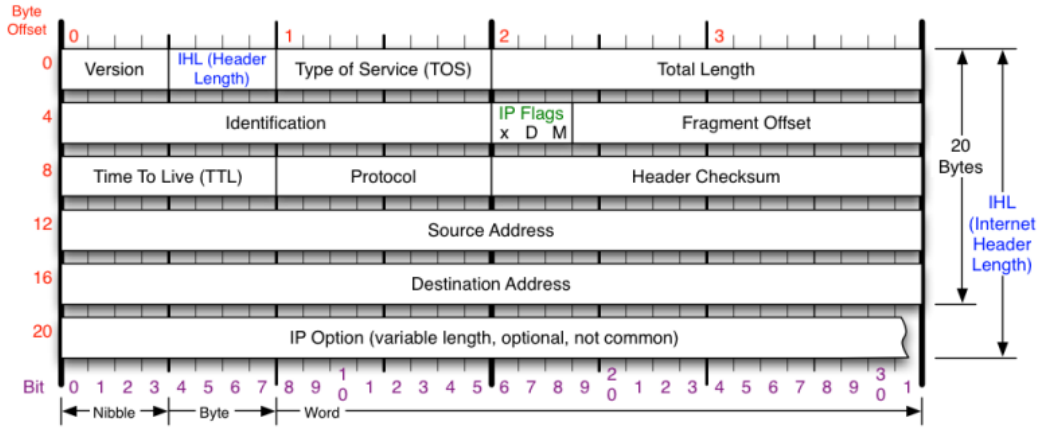
- Originally classful subnetting (subnets in A/B/C ranges, with 24, 16 and 8 bits of network addresses respectively; D range for multicast and an unused E range).[2]
- Classless Inter-Domain Routing (CIDR), with network masks to mark the difference between network address and host address. Routing done by selecting most specific match.

[1] `http://www.tcpipguide.com/free/t_TCPIPProcessesMultiplexingandClientServerApplicati-2.htm`

[2] `http://en.wikipedia.org/wiki/Classful_network#Introduction_of_address_classes`

- Variable Length Subnet Masks (VLSM) to use different subnets that do not have the requirement of having the same size. Add the possibility of subnets inside subnets. This was not possible in RIPv1.
- A "link" is defined as the topological area in which a packet with $TTL = 1$ can be delivered (aka. not being forwarded).
- A "subnet" is the topological area in which the interfaces receive the same network prefix.

1.7 IP Packet Format



Version

Version of IP Protocol. 4 and 6 are valid. This diagram represents version 4 structure only.

Header Length

Number of 32-bit words in TCP header, minimum value of 5. Multiply by 4 to get byte count.

Protocol

IP Protocol ID. Including (but not limited to):

1 ICMP	17 UDP	57 SKIP
2 IGMP	47 GRE	88 EIGRP
6 TCP	50 ESP	89 OSPF
9 IGRP	51 AH	115 L2TP

Total Length

Total length of IP datagram, or IP fragment if fragmented. Measured in Bytes.

Fragment Offset

Fragment offset from start of IP datagram. Measured in 8 byte (2 words, 64 bits) increments. If IP datagram is fragmented, fragment size (Total Length) must be a multiple of 8 bytes.

Header Checksum

Checksum of entire IP header

IP Flags

x D M

x 0x80 reserved (evil bit)
D 0x40 Do Not Fragment
M 0x20 More Fragments follow

RFC 791

Please refer to RFC 791 for the complete Internet Protocol (IP) Specification.

2 CLEAN: Calculating, Legacy, Endianness, Addressing, Networks

2.1 Calculating: Counting

Counting Process that starts with $n = 0$ as the initial count. Every counted object is labeled with the actual n value, and n is updated to $n = n + 1$. Process ends when all objects have been counted.

2.2 Legacy

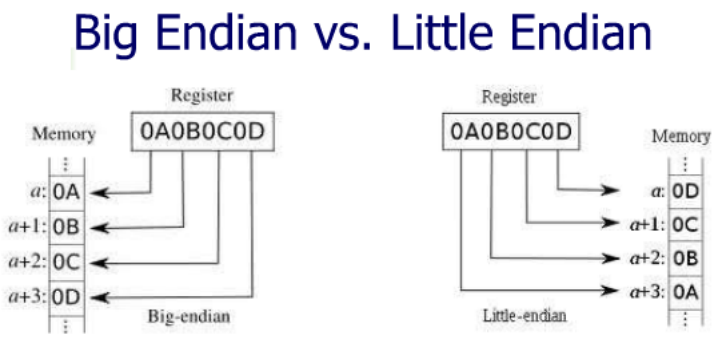
- Everybody knows what Karst thinks of legacy.

2.3 2-adic vs Binary

2-adic	Binary	2-adic to base-10	Binary to base-10
1	0	1	0
2	1	2	1
11	00	3	0
12	01	4	1
21	10	5	2
22	11	6	3
111	000	7	0

- The whole point is that binary resets at every range increase.

2.4 Big-endian and Little-endian



2.5 Addressing

- `http://www.exploringbinary.com/binary-converter/`

3 IPv6

3.1 Rationale

- $4x$ address space size increase = 2^{96} address number increase.
- Headers have a fixed size of 40 bytes. Supports extended headers for additional functionality.
- NATs no longer needed due the vast amount of addresses.

3.2 Addressing

- 128-bit addresses.
- 8 blocks of 4 nibbles ($8x4x4 = 128$ bits)
- Consecutive blocks of all-zeroes can be replaced by :: once.
- No broadcasts, no subnet masks.
- `http://www.iana.nl/assignments/ipv6-address-space/ipv6-address-space.xhtml`

Reserved addresses

::/8	Special-purpose
100::/8	Special-purpose
2000::/3	Global unicast
fc00::/7	Unique local unicast
fe80::/10	Link-local (Link-scoped) unicast
ff00::/8	Multicast

Special-purpose addresses

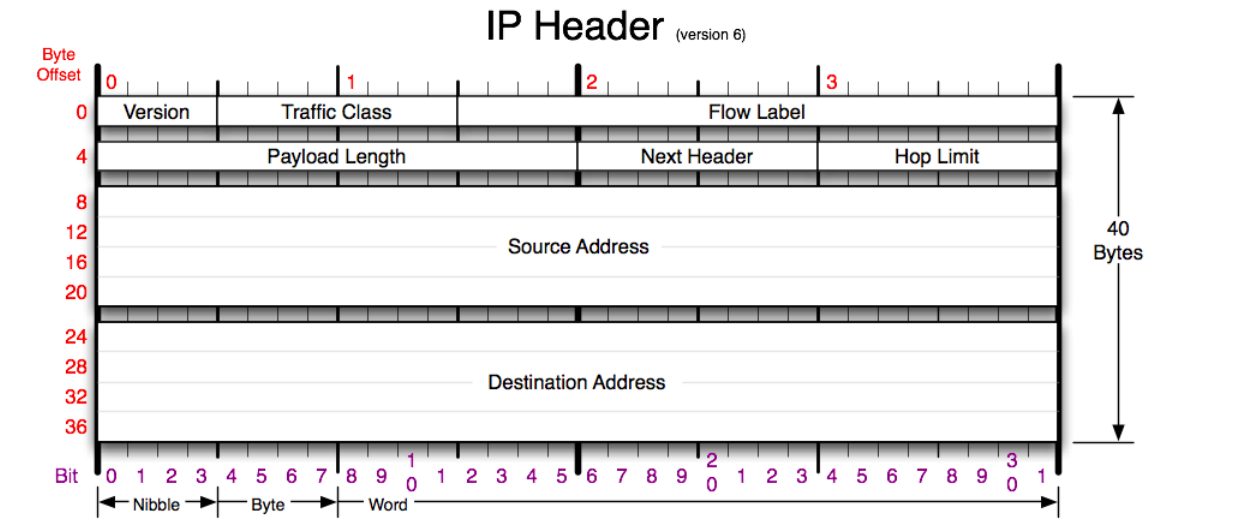
::/128	Unspecified address
::1/128	Localhost address
::a.b.c.d/128(from ::/96)	IPv4-compatible addresses
::ffff:a.b.c.d/128 (from ::ffff:0:0/96)	IPv4-mapped addresses
64:ff9b::/96	Well-known prefix
100::/64	Discard-only address block

3.3 Neighbour Discovery Protocol (NDP)

- IPv6 does not use ARP. Uses ICMPv6 instead.
- ICMPv6 types for NDP:

133	Router Solicitation
134	Router Advertisement
135	Neighbor Solicitation
136	Neighbor Advertisement
137	Redirect Message

3.4 IPv6 Header



Version

Version of IP Protocol. 4 and 6 are valid. This diagram represents version 6 structure only.

Traffic Class

8 bit traffic class field.

Flow Label

20 bit flow label.

Payload Length

16-bit unsigned integer. Length of the IPv6 payload, i.e., the rest of the packet following this IPv6 header, in octets. Any extension headers are considered part of the payload.

Source Address

128-bit address of the originator of the packet.

Next Header

8-bit selector. Identifies the type of header immediately following the IPv6 header. Uses the same values as the IPv4 Protocol field.

Destination Address

128-bit address of the intended recipient of the packet (possibly not the ultimate recipient, if a Routing header is present).

Hop Limit

8-bit unsigned integer. Decremented by 1 by each node that forwards the packet. The packet is discarded if Hop Limit is decremented to zero.

RFC 2460

Please refer to RFC 2460 for the complete Internet Protocol version 6 (IPv6) Specification.

4 Layer 2: Bridging and Switching

4.1 Layers 1 and 2

Layer 1 Repeaters, hubs. Same collision domain. Same link segment.

Layer 2 Bridges, switches. Same collision domain. Same link segment.

4.2 Layer 2: MAC and LLC

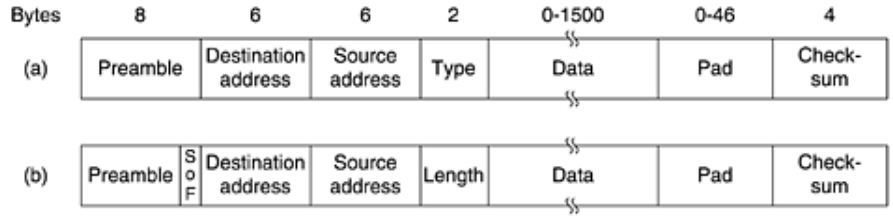
MAC Media Access Control. Work from IEEE 802.3[3].

CSMA/CD Carrier Sense Multiple Access With Collision Detection. Ethernet is the most common.

1. Is my frame ready for transmission? If yes, move to 2.
2. Is medium idle? If not, wait until it becomes ready.
3. Start transmitting.
4. Did a collision occur? If so, go to collision detected procedure.
5. Reset retransmission counters and end frame transmission.

LLC Logical Link Control. Multiplexing mechanisms, flow control and error management. Interface between MAC and Network Layer.

4.3 Frame Formats



DIX Ethernet DEC-Intel-Xerox initial frame structure[4].

8bytes Preamble Each one containing 10101010. Used for synchronization.

6bytes Destination Address

6bytes Source Address

2bytes Type Indicator of the used transport protocol.

0-1500bytes Payload

[3] http://en.wikipedia.org/wiki/IEEE_802.3
[4] http://www.epubbud.com/read.php?g=5HEKFDZU&two=1&tocp=38

0-46bytes Pad Valid Ethernet frames have, at least, 64 bytes (not counting Preamble!). If a frame is less than that (Payload ¡ 46bytes), a pad is added to it. This is done to ease collision detection.

4bytes Checksum CRC of all the frame fields.

802.3 Ethernet Changes from DIX:

7bytes Preamble + 1byte Start of Frame Same as before but changing last byte to have compatibility with 802.4 and 802.5 (Token Bus and Token Ring).

2bytes type -¿ 2bytes length IEEE tried to change the purpose of the field (and move "type" information to *inside* the Payload), but some people did not change. Rule of thumb: if its value is over 1500, it is a Length field, otherwise it is a Type field.

Changes on the Payload There are 8 additional bytes of 'metadata' on the payload that are used mostly for nothing, effectively reducing the MTU from 1500 to 1492.

4.4 MAC Addresses

MAC48 Physical, obsolete.

EUI48 Virtual, including physical.

EUI64 Extended.

OUI First 24bits of the MAC address. Identify who issued the device.

4.5 EUI48 to EUI64

[0:23bits]:FF:FE:[24:–bits]. When used to generate IPv6 address, the 7th most significative bit is swapped.

4.6 Ethernet Types

0x0800 IP

0x0806 ARP

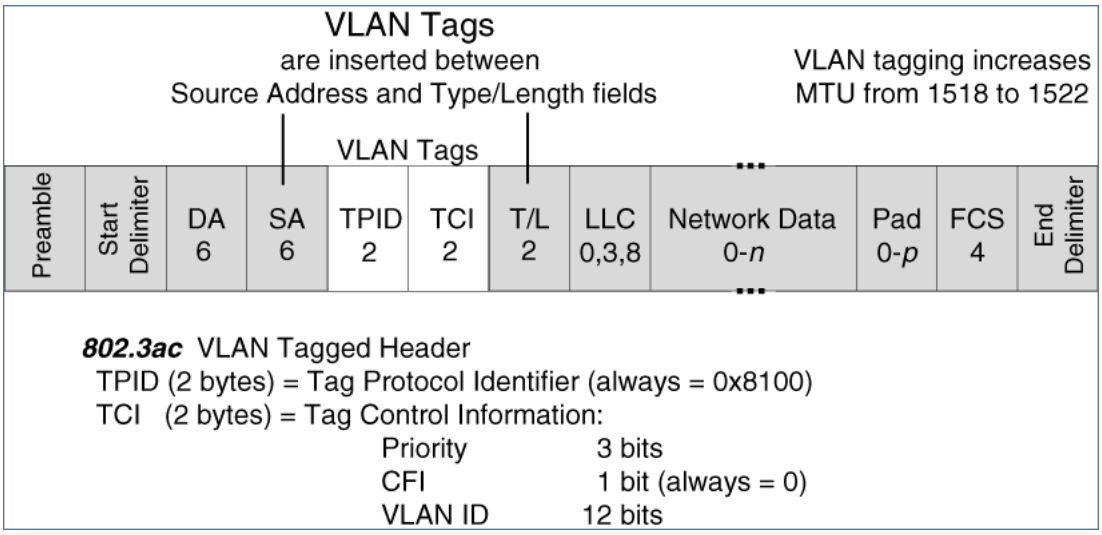
0x86DD IPv6

4.7 Bridges and Switches

Transparent Bridges Use Store-and-Forward. Copy data from one port to another (or multiple ports). They can learn (and remember) where other devices are when they receive messages.

Switches are synonyms of Bridges Usually refer to bridges with multiple interfaces.

4.8 VLANs (802.1Q-2011)



4.9 Layered Extensions

TO-DO

4.10 PBB-TE, TRILL, SPB

TO-DO

5 STP Protocol

5.1 Goals and properties

1. Eliminate edges (connections) until there are no possible loops.
2. After performing the algorithm, graph turns into a tree.
3. Topology changes cause changes on the tree.
4. Protocol works by electing a Root Node.

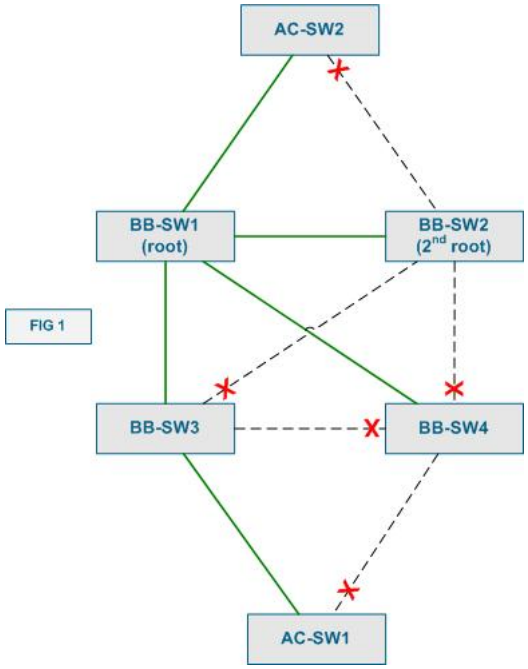
5.2 Configuration Messages

ID based on a variable priority and its MAC address.

Root Node Elected as the node with the lowest ID on the system.

Root ports on each Node Chosen by[5]:

1. Lower advertised Root ID.
2. Lower advertised cost to Root.
3. Lower transmitting bridge ID.
4. Lower port ID.



[5] https://www.youtube.com/watch?v=iB7BxtZVy3c

5.3 Timing Parameters

Hello Time Time between two configuration messages.

Max Age Parameter to discard messages that are too old.

Forward Delay Half of the delay before transitioning from blocking to forwarding.

1. Can be understood as two different waiting times. During that waiting, it does not forward packets.
2. First waiting: Listen for neighbours (other bridges).
3. Second waiting: Learn the location of MAC addresses.

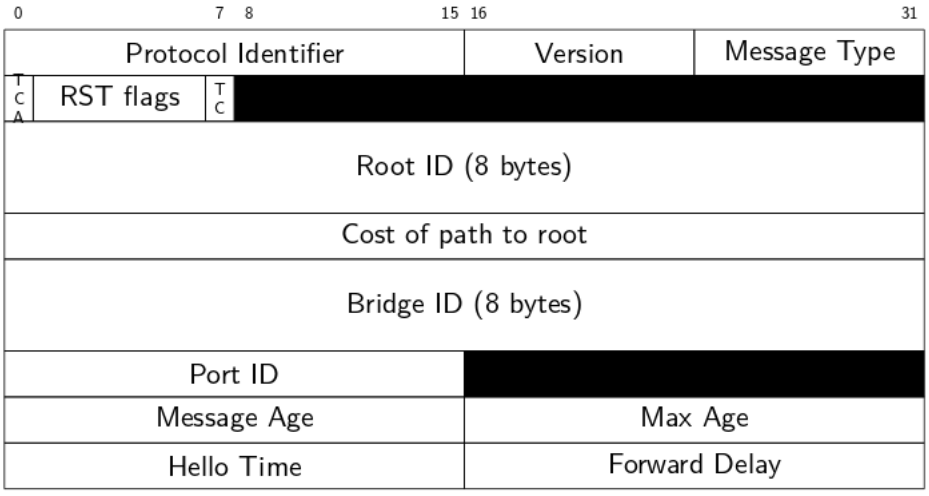
5.4 Topology Change Mechanism

Memory Bridges remember where other bridges are located.

Stable Topology When the topology is stable, bridges have a long caching time.

Topology Changes When a topology change is detected, the bridge detecting the change (and subsequent ones) sends a Topology Change Notification on his Root Port. When the message reaches the Root Bridge, it sets up its Topology Change flag. This causes other bridges to switch to the short caching delay.

5.5 Bridge Protocol Data Unit (BPDU)



Protocol Identifier All zeroes.

Version 0 for STP, 2 for RST

Message Type 0 for Hello, 2 for RST, 128 for TCN

Flags TCA (Topology Change Ack), [Proposal, Agreement, ...], TC (Topology Change)

Root ID Root Bridge.

Cost to Root Cost to the Root Bridge.

Bridge ID Bridge transmitting.

Port ID Port used on the transmitting bridge.

Message Age Age of the BPDU information.

Max Age Tipically 20 seconds (min 6).

Hello Time Tipically 2 seconds.

Forward Delay Tipically 15 seconds (min 4).

5.6 STP Enhancements

Rapid Spanning Tree (RST) Changes on the BPDUs to make the algorithm faster.

STP on VLANs Global STP for all VLANs. Individual STP on each one of them.

MSTP Divide LAN into regions. Run STP on each one of them.

6 Routing

6.1 Basics

Direct Routing Added automatically by $ifconfig$ on unix systems.

Global Routing Done by using routing tables.

Netstat/Route flags Flags of the route command:

- **G** Needs gateway / directly connected.
- **H** Route to Host / Route to Network
- **S** Static route (manually added) / Dynamic route (added by a protocol)

ARP Address Resolution Protocol.

IP Replacess $ifconfig$, *route* and *arp*.

Route Selection Most specific entry on the routing table first. Use default if there is no match.

6.2 Internet Routing

Autonomous Systems (AS) An AS is a connected group of one or more IP prefixes that has a single and well-defined routing policy. Each AS is managed by one or more operators and hosts a collection of routers and networks.

Edge Routers Edge Routers are used to communicate between different ASs, using an External Gateway Protocol such as BGP4 (which holds a monopoly in these type of communications).

Internal Routers Internal Routers use a different set of Internal Gateway Protocols (IGP) to communicate between themselves, such as RIP, OSPF or IS-IS.

6.3 Dynamic Routing Mechanisms

Distance Vector Routing (RIP) Bellman-Ford.

Link State Routing (OSPF) Dijkstra

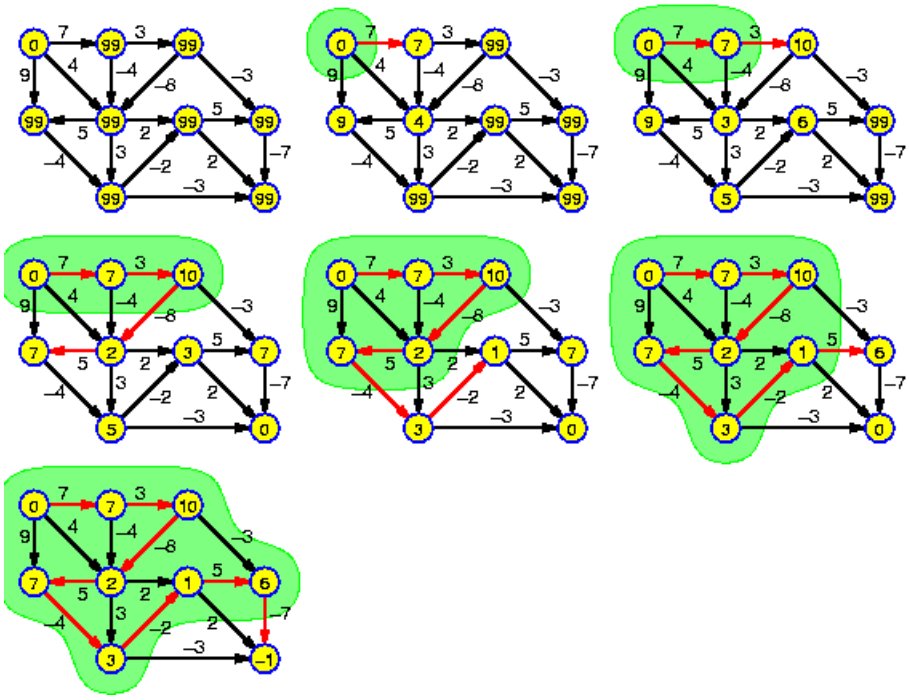
Path Vector Routing (BGP)

7 Algorithms

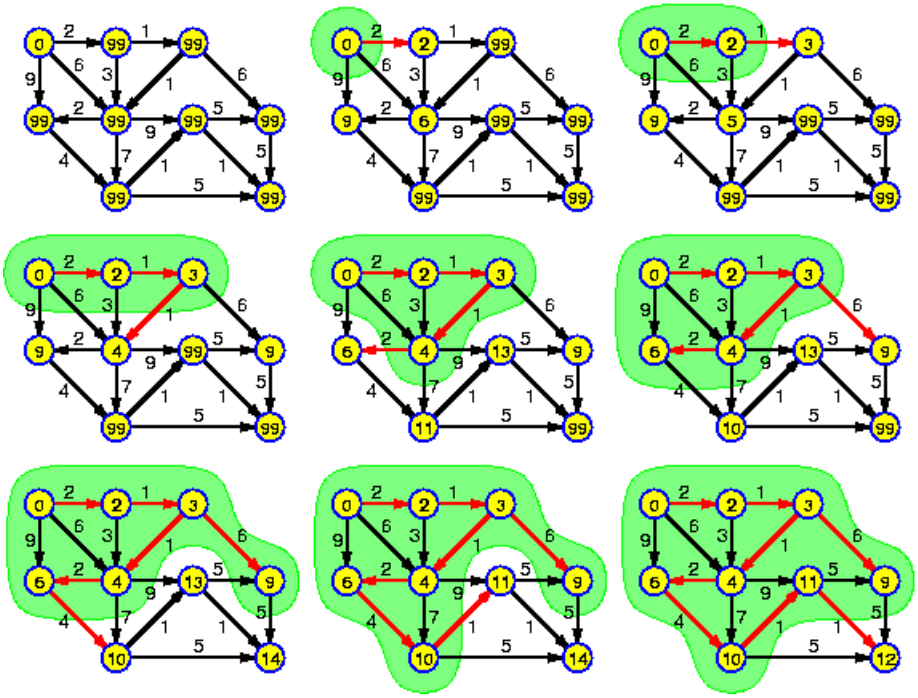
7.1 Counting to Infinity Remarks

1. In Computer Science, infinite is finite.
2. To avoid issues, do not advertise back to the advertiser (Split Horizon).
3. Poisoned Reverse: Announce back infinite cost to advertiser.

7.2 Bellman-Ford



7.3 Shortest Path Tree (Dijkstra)



7.4 Minimum Spanning Tree (Prim, Kruskal)

Prim Algorithm[6]:

1. Pick random node N and add it to the tree T.
2. From all the vertex on the current tree T, pick V_i with the lowest cost. Add the connected node to the tree.
3. Repeat 2 until all nodes are present on the tree.

Kruskal Algorithm[7]:

1. Pick the lowest vertex on the graph and create a tree with the two nodes connected.
2. From the rest of the graph, pick the vertex with the lowest value that does not generate a loop. This will create new trees or expand the current ones.
3. Repeat 2 until all nodes are present on a single.

[6] https://www.youtube.com/watch?v=cplfcGZmX7I
[7] https://www.youtube.com/watch?v=71UQH7Pr9kU

8 Distance Vector Protocol- RIP

8.1 RIP Version 1

8.1.1 Basics

1. Based on the Bellman-Ford algorithm.
2. Keep a table of the routes to destinations as distance (metric), gateway (next hop).
3. Periodically send table to neighbours.
4. Update table with incoming information (regularly can only get better, unless TC announced by root).
5. Sent updates as soon as a route changes (later addition).
6. Supports one level depth subnet masks. It differentiates updates from the same and another. networks.
7. Infinity = 16 hops.

8.1.2 Timers

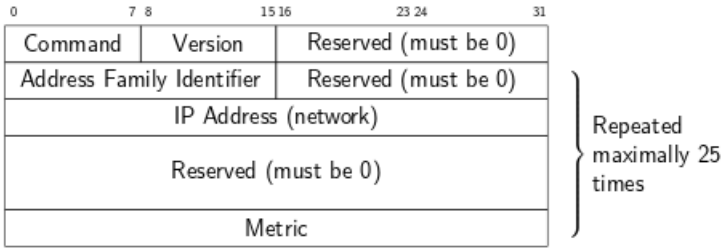
Update timer updates are sent every 30 seconds.

Invalid/Timeout timer routes time out after 180 seconds.

Flush timer routes disappear after 240 seconds.

Hold-down timer Prevent incorporating possibly bad routing information which might be present in a network that didnt converge yet (180 seconds).

8.1.3 Packet Format



1. Requests broadcasted to 255.255.255.255 or to a directed broadcast address. Answers as unicast.
2. Uses UPD port 520.

3. UDP packet 512 bytes - 8 bytes UDP header: 504 bytes. 20 bytes/route = max 25 route updates/packet.

8.2 Protocol Extensions

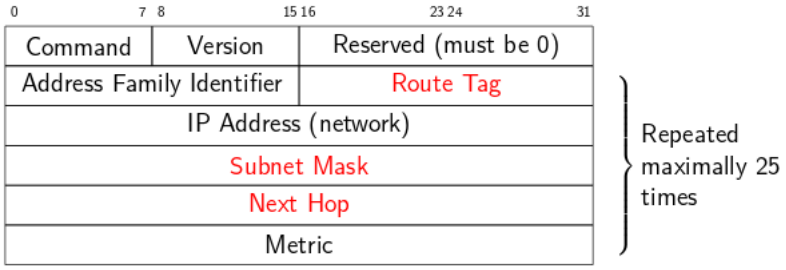
8.2.1 Interior Gateway Routing Protocol (IGRP)

1. Runs directly on top of IP (protocol 9).
2. Larger notion of infinity (100-255).
3. Can handle up to four parallel paths.
4. Uses three types of network routes.
5. Metric includes, besides hop count, other information delay, bandwidth, reliability and load.

8.2.2 Enhanced Interior Gateway Routing Protocol (EIGRP)

1. Runs directly on top of IP (protocol 88).
2. Keeps state of neighbours.
3. Remembers *all* paths.

8.3 RIP Version 2



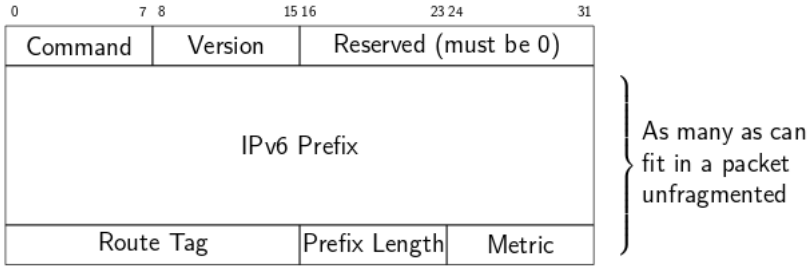
Route Tag Identification of route origin. Differentiates internally and externally generated routes. Can be used for authentication.

IP Address Destination Network. RIPv2 uses multicast group 224.0.0.9, not broadcast.

Subnet Mask CIDR support.

Next Hop Gateway (if different from advertising router).

8.4 RIPng



1. Runs on UPD (Port 521).
2. Packets can be as large as MTU.
3. Used in IPv6. Multicast address FF02::9.

9 Link State Routing - OSPF

9.1 Basics

1. Based on the Dijkstra algorithm.
2. Used instead of distance vector algorithms on more complex topologies.
3. Faster convergence time than distance vector algorithms.

9.2 Link State Packets

1. LSP represent the state of a Router and its links to the rest of the network.
2. Enough for point-to-point links.
3. Broadcast Networks and NBMA networks are represented by virtual nodes.

9.3 Non-broadcast Networks

NBMA Non-Broadcast Multiple Access

1. Full-mesh connectivity, but not permanently.
2. Connectivity by Designated Routers (DR).

Point-to-multipoint

1. Subset of all point-to-point links.
2. No full-mesh.
3. No DR elected.

9.4 LSP Generation

1. Announcements every 30 minutes (RIP was 30 seconds)
2. Announcements trigger as soon as changes are detected (link up/down, change on link cost).
3. Use smart flooding to detect duplicates (propagation looks like a tree).

9.5 LSP Problems

LSP arriving out of order Solutions: Timestamps and Sequence numbers. Timestamps require clock synchronization. Sequence numbers must be large enough to guarantee ordering, and the value should always be increased when forwarding the LSP.

9.6 OSPF Properties

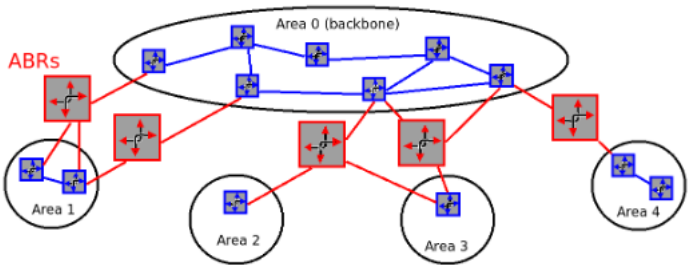
1. Supports hierarchical routing and subnets.
2. Efficient multicast for flooding.
3. Uses metrics based on cost on each interface.
4. Supports virtual links, load balancing, unnumbered interfaces and authentication.
5. Built on top of IP (protocol 89).
6. LSP are named LSA (Link State Advertisement).
7. OSPF uses multicast address 224.0.0.6 (FF02::6) for DR and multicast address 224.0.0.5 (FF02::5) for all routers.

9.7 Parameters and LSAs

1. Timing parameters must be the same for all OSPF neighbours.
2. If routing table overflows, external routers are dropped first to generate space.
3. LSAs must be acknowledged, and can be queued for transmission. They must time out at about the same time.

9.8 Network representation

1. A DR and a Backup Designated Router (BDR) are elected on each multi-access network with Hello packets. The (B)DR represents the network as a virtual node and acts on its behalf. Routers can have custom priorities for the election.
2. Hierarchical Routing:



9.9 Router Types

Backbone Router At least one interface is inside in the Backbone (Area 0).

Internal Router All interfaces are in the same Area.

Area Border Router (ABR) Has one interface in the Backbone and one or more in other areas.

Autonomous System Boundary Router (ASBR) Participates in external routing protocols.

9.10 Stubby Areas

Stubby Area Area in which no external routing information is injected by the ABRs.

Totally Stubby Area Stubby Area in which not even inter-area summaries are injected.

Not-So-Stubby Area Stubby Area in which external routing information can be generated and propagated locally.

9.11 OSPF Packet Format

10 Path Vector Routing - BGP