

Advanced Networking

Routing:

RIP, OSPF, Hierarchical routing, BGP

Renato Lo Cigno

Renato.LoCigno@disi.unitn.it

Routing Algorithms: One or Many?

- Is there a single routing protocol in the Internet?
- How can different protocols and algorithms coexist
 - Homogeneous results
 - Risk of inconsistent routing
- Complexity of routing algorithms/protocols
 - Can they scale?
 - There is a tradeoff between traffic and computation?
- Hierarchical routing
- Policy routing: what is it, why not "performance"?



RIP - History

- * Late 1960s: Distance Vector protocols were used in the ARPANET
- * Mid-1970s: XNS (Xerox Network system) routing protocol is the precursor of RIP in IP (and Novell's IPX RIP and Apple's routing protocol)
- * 1982: Release of routing software for BSD Unix
- * 1988: RIPv1 (RFC 1058)
 - classful routing
- * 1993: RIPv2 (RFC 1388)
 - adds subnet masks with each route entry
 - allows classless routing
- * 1998: Current version of RIPv2 (RFC 2453)



RIP at a glance

- A simple intradomain protocol
- Straightforward implementation of Distance Vector Routing...
 - Distributed version of Bellman-Ford (DBF)
- ...with well known issues
 - slow convergence
 - works with limited network size
- Strengths
 - simple to implement
 - simple management
 - widespread use

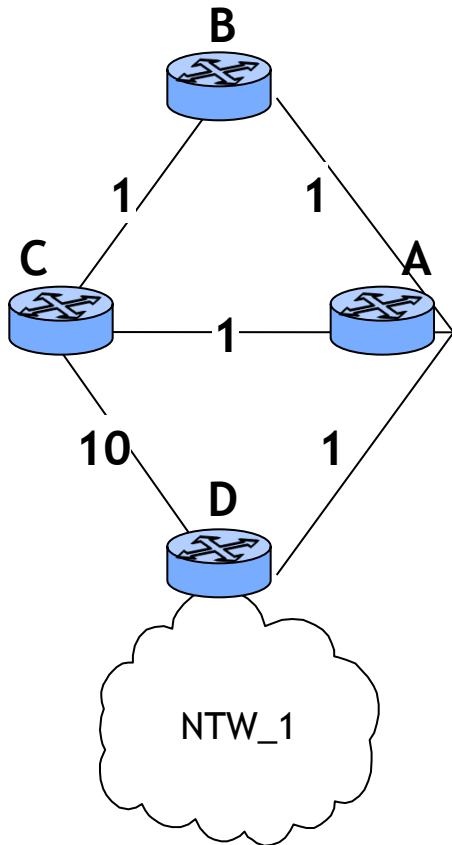


RIP at a glance

- Metric based on hop count
 - maximum hop count is 15, with “16” equal to “ ∞ ”
 - imposed to limit the convergence time
 - the network administrator can also assign values higher than 1 to a single hop
- Each router advertises its distance vector every 30 seconds (or whenever its routing table changes) to all of its neighbors
 - RIP uses UDP, port 520, for sending messages
- Changes are propagated across network
- Routes are timeout (set to 16) after 3 minutes if they are not updated



Recall: “counting to infinity” problem



Router A		
Dest	Next	Metric
NTW_1	D	2

Router B		
Dest	Next	Metric
NTW_1	A	3

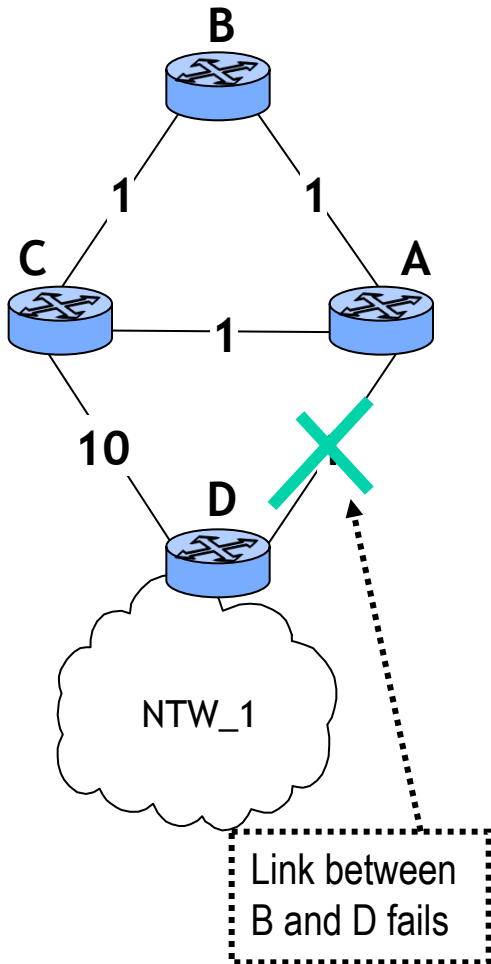
Router C		
Dest	Next	Metric
NTW_1	A	3

Router D		
Dest	Next	Metric
NTW_1	dir	1

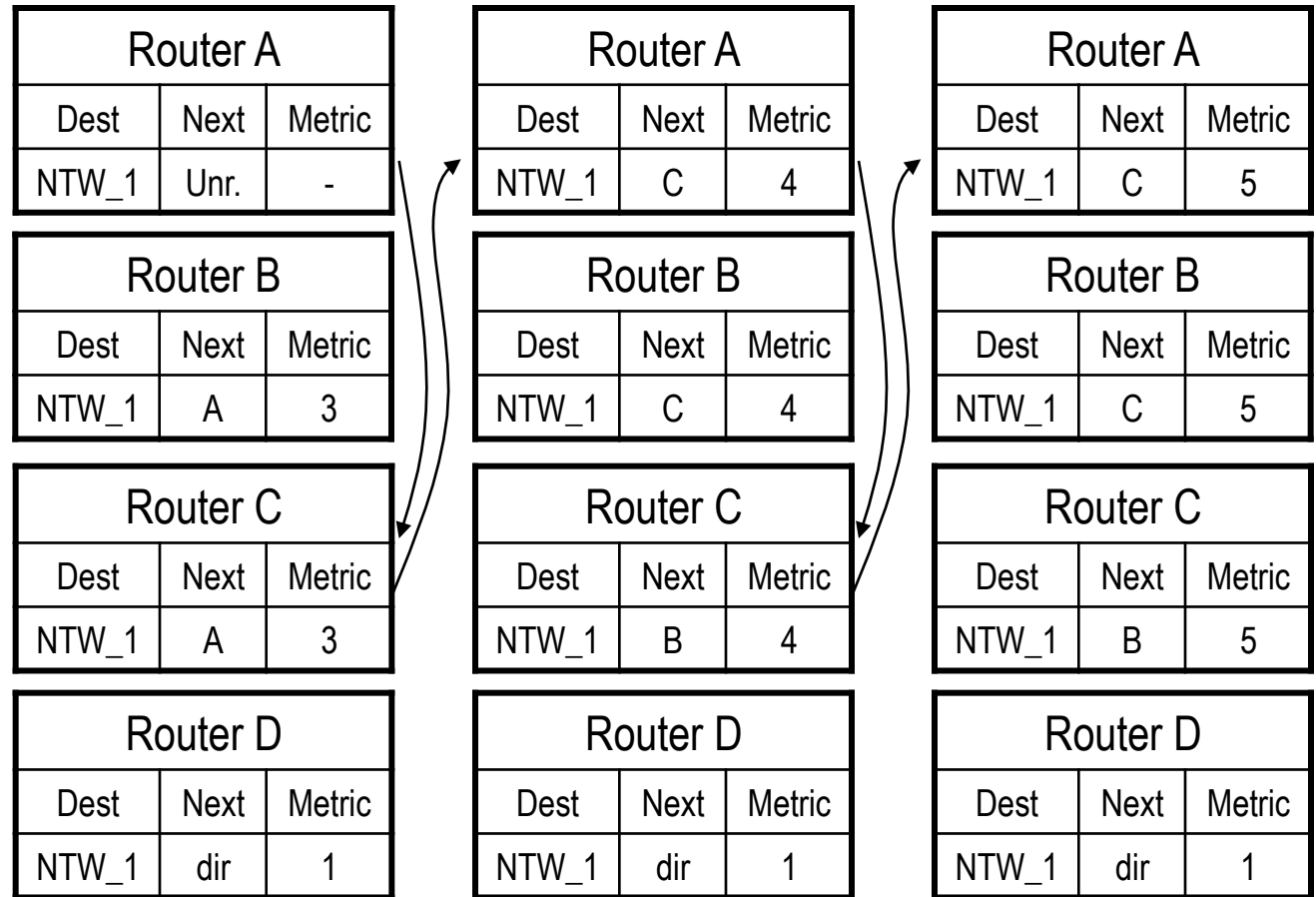
- Consider the entries in each routing table for network NTW_1
- Router D is directly connected to NTW_1



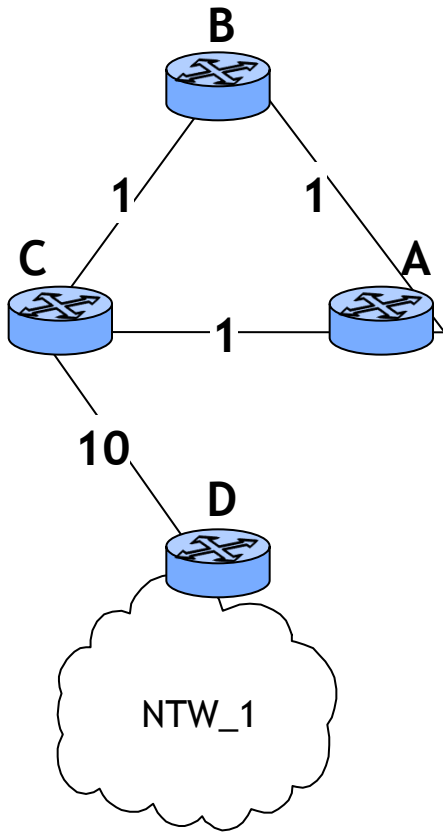
Recall: “counting to infinity” problem (2)



time



Recall: “counting to infinity” problem (3)



time



Router A		
Dest	Next	Metric
NTW_1	C	11

Router A		
Dest	Next	Metric
NTW_1	C	12

Router B		
Dest	Next	Metric
NTW_1	C	11

Router B		
Dest	Next	Metric
NTW_1	C	12

...

Router C		
Dest	Next	Metric
NTW_1	B	11

Router C		
Dest	Next	Metric
NTW_1	D	11

Router D		
Dest	Next	Metric
NTW_1	dir	1

Router D		
Dest	Next	Metric
NTW_1	dir	1



RIP: solution to “counting to infinity”

- Maximum number of hops bounded to 15
 - this limits the convergence time
- Split Horizon
 - simple
 - each node *omits* routes learned from one neighbor in update sent to that neighbor
 - with poisoned reverse
 - each node *include* routes learned from one neighbor in update sent to that neighbor, setting their metrics to infinity
 - drawback: routing message size greater than simple Split Horizon



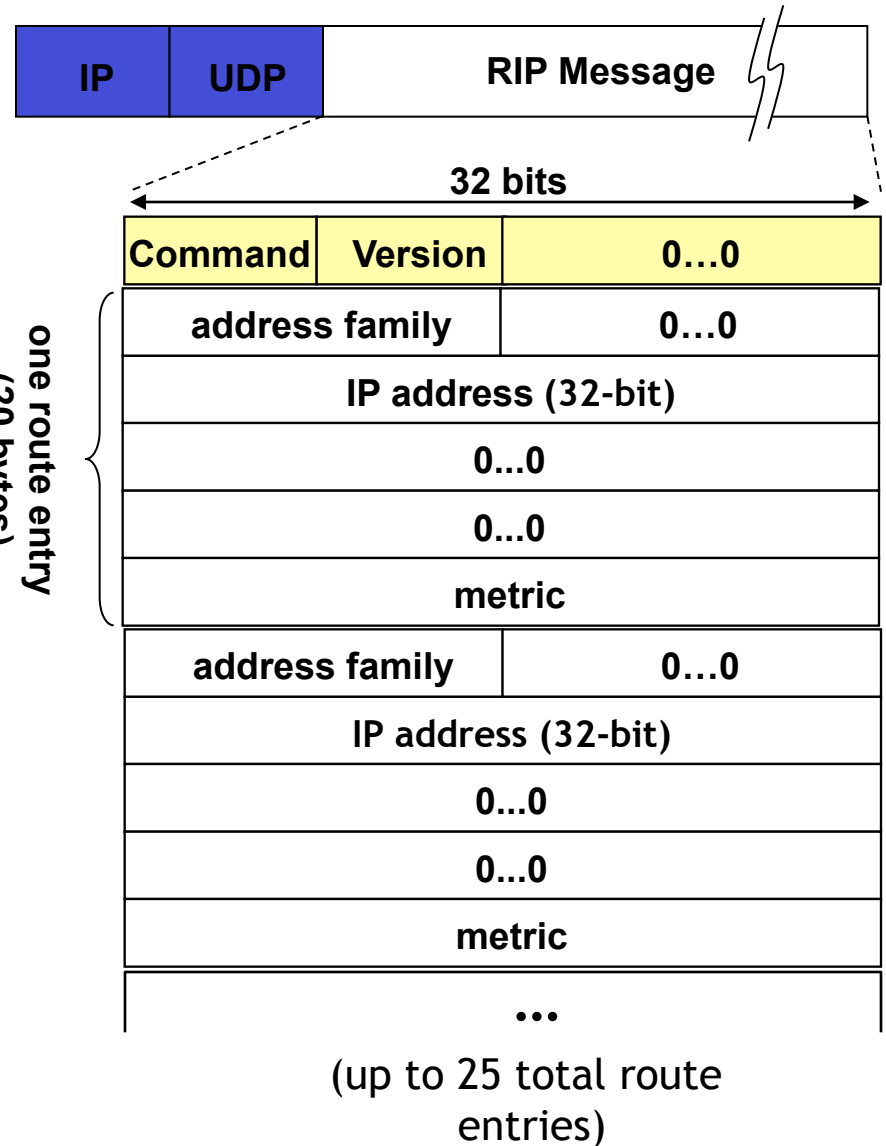
RIP: solution to “counting to infinity” (cont’ d)

- Triggered updates: nodes send messages as soon as they notice a change in their routing tables
 - only routes that has changed are sent
 - faster reaction...
 - ...but more resources are used (bandwidth, processing)
 - cascade of triggered updates
 - superposition with regular updates



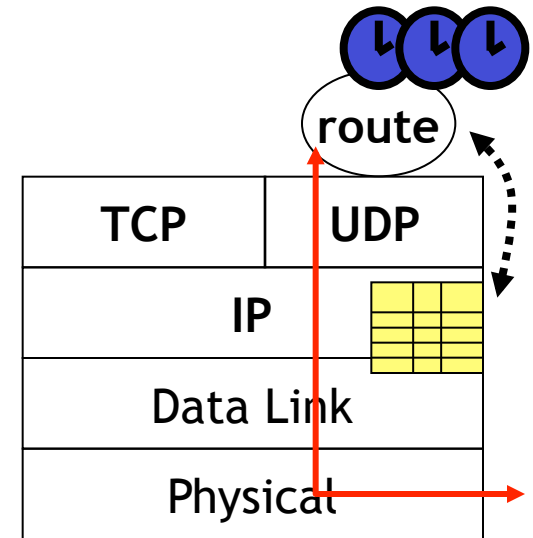
RIP-1: Message Format

- Command: 1=request
2=response
 - Updates are replies whether asked for or not
 - Initializing node broadcasts request
 - Requests are replied to immediately
- Version: 1
- Address family: 2 for IP
- IP address: non-zero network portion, zero host portion
 - Identifies particular network
- Metric
 - Path distance from this router to network
 - Typically 1, so metric is hop count



RIP procedures: introduction

- RIP routing tables are managed by application-level process
 - e.g., *routed* on UNIX machines
- Advertisements are sent in UDP packets (port 520)
- RIP maintains 3 different timers to support its operations
 - Periodic update timer (25-30 sec)
 - used to sent out update messages
 - Invalid timer (180 sec)
 - If update for a particular entry is not received for 180 sec, route is invalidated
 - Garbage collection timer (120 sec)
 - An invalid route in marked, not immediately deleted
 - For next 120 s. the router advertises this route with distance infinity



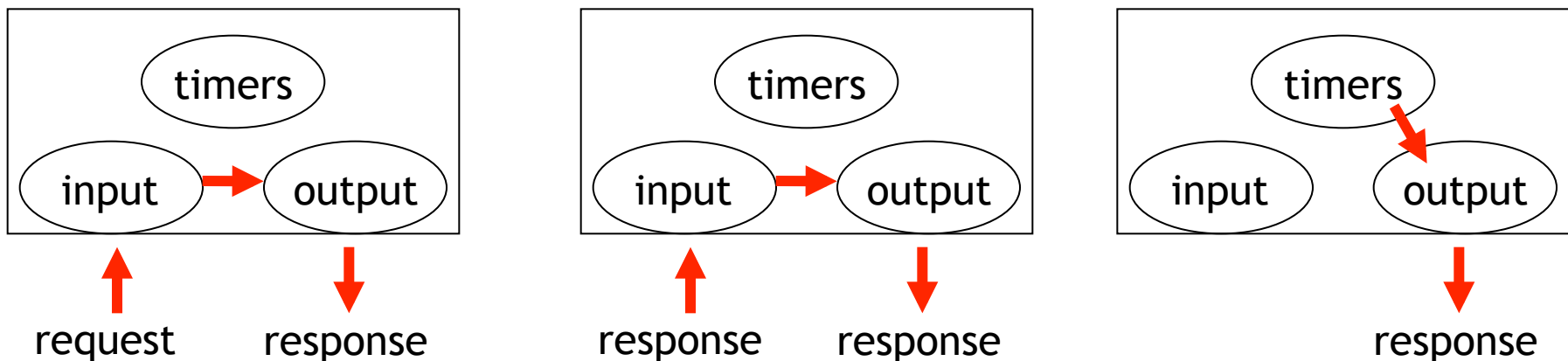
RIP procedures: input processing

- *Request Messages*
 - they may arrive from routers which have just come up
 - action: the router responds directly to the requestor's address and port
 - request is processed entry by entry
- *Response Messages*
 - they may arrive from routers that perform regular updates, triggered updates or respond to a specific query
 - action: the router updates its routing table
 - in case of new route or changed routes, the router starts a triggered update procedure



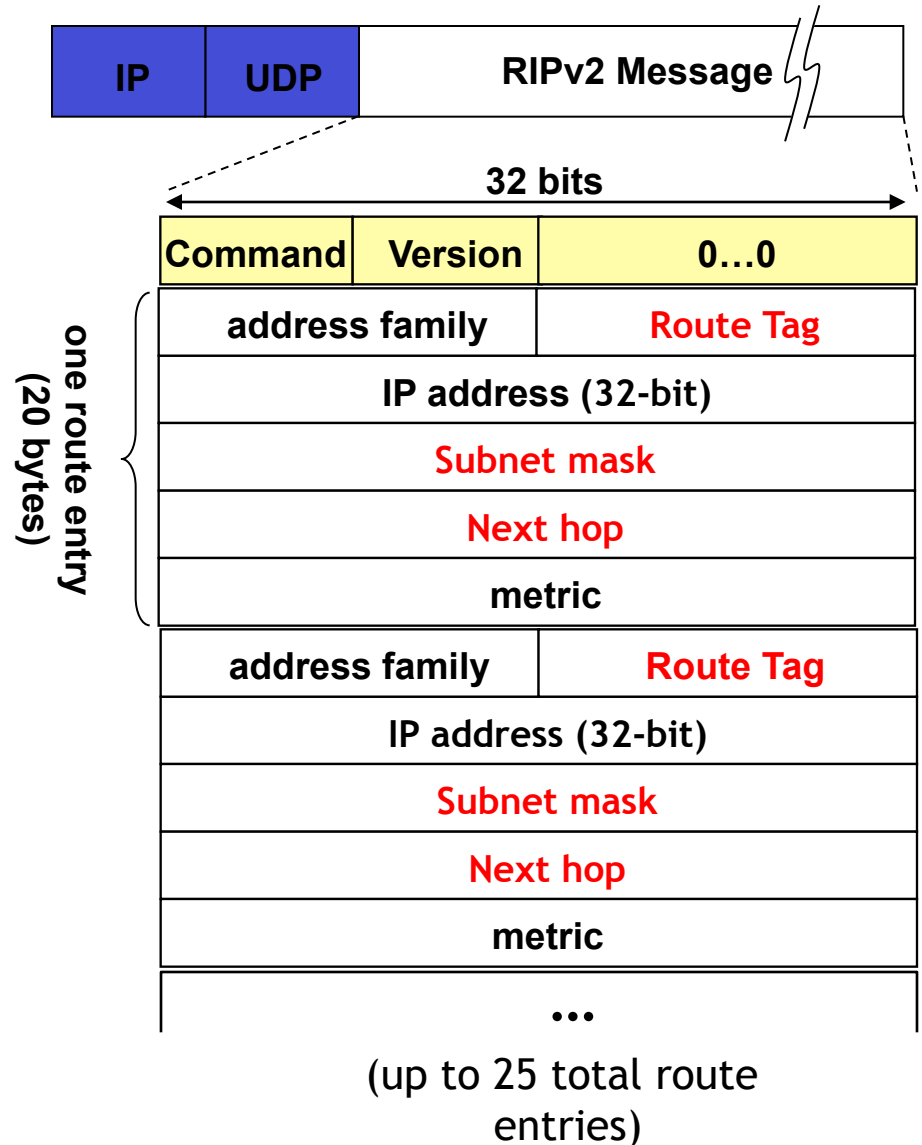
RIP procedures: output processing

- Output are generated
 - when the router comes up in the network
 - if required by the input processing procedures
 - by regular routing update
- Action: the router generates the messages according to the commands received
 - the messages contain entries from the routing table



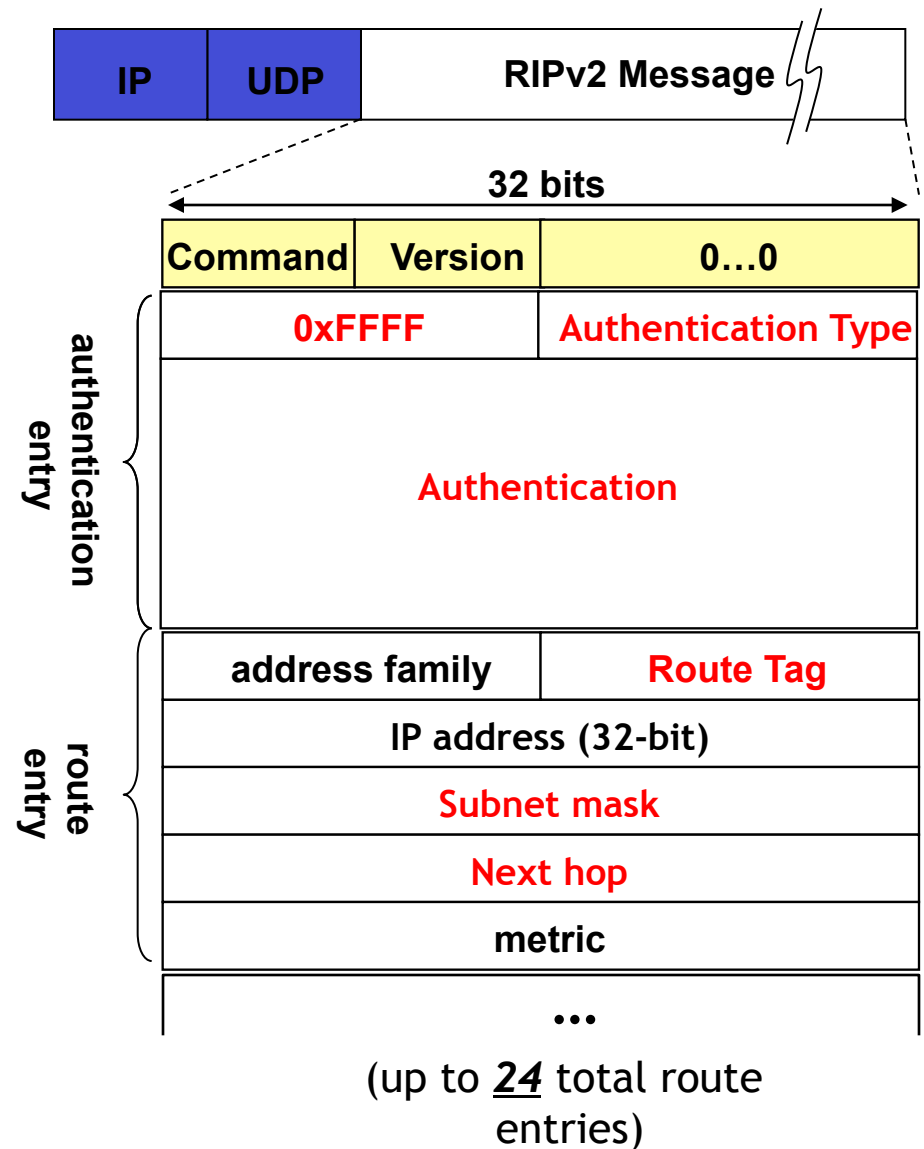
RIPv2: Message Format

- Version: 2
- Route Tag: used to carry information from *other routing protocols*
 - e.g., autonomous system number
- Subnet mask for IP address
- Next hop
 - identifies a better next-hop address on the same subnet than the advertising router, if one exists (otherwise 0....0)



RIPv2: authentication

- Any host sending packets on UDP port 520 would be considered a router
- Malicious users can inject fake routing entries
- With authentication, only authorized router can send Rip packets
 - Authentication type
 - password
 - MD5
 - Authentication
 - plain text password
 - MD5 hash



RIPv2: other aspects

- Explicit use of subnets
- Interoperability
 - RIPv1 and RIPv2 can be present in the same network since RIPv1 simply ignores fields not known
 - RIPv2 responds to RIPv1 Request with a RIPv1 Response
- Multicast
 - instead of broadcasting RIP messages, RIPv2 uses multicast address 224.0.0.9



RIP limitations: the cost of simplicity

- Destinations with metric more than 15 are unreachable
 - If larger metric allowed, convergence becomes lengthy
- Simple metric leads to sub-optimal routing tables
 - Packets sent over slower links
- Accept RIP updates from any device (if no security is implemented)
 - Misconfigured device can disrupt entire configuration



RIP Was the first ... but ...

- Why is RIP not enough to manage the Internet?
- Can Link-State protocols perform better?
 - OSPF
 - MOSPF (no MRIP exists!!)
- Inter-AS routing requires an entirely different approach ... if not for else for the sake of competition!



Non-RIP, DV Protocols: EXAMPLE IGRP (Interior Gateway Routing Protocol)

- CISCO proprietary; builds on RIP (mid 80' s)
- Distance Vector, like RIP
- several cost metrics (delay, bandwidth, reliability, **load** etc.)
- uses TCP to exchange routing updates
- routing tables exchanged only when costs change
- Loop free routing achieved by using a Distributed Updating Alg. (DUAL) based on *diffused computation*
- In DUAL, after a distance increase, the routing table is *frozen* until all affected nodes have learned of the change (cfr. split horizon in RIP)



Open Shortest Path First (OSPF)

- RIP limited in large internets
- OSPF is often preferred interior routing protocol for TCP/IP based internets
- Uses link state routing
- Floods the messages to all routers in the AS (area)



OSPF “advanced” features (not in RIP)

- Security: all OSPF messages are authenticated (to prevent malicious intrusion);
 - TCP or Unicast in general connections used sometimes
- Multiple same-cost paths allowed
 - only one path in RIP
- For each link, multiple cost metrics for different TOS (eg, satellite link cost set “low” for best effort; high for real time)
- Integrated uni- and multicast support: Multicast (MOSPF) uses same topology data base as OSPF
- Hierarchical OSPF in large domains



Link State Routing

- When initialized, router determines link cost on each interface
- Router advertises these costs to all other routers in topology
- Router monitors its costs
 - When changes occurs, costs are re-advertised
- Each router constructs topology and calculates shortest path to each destination network
- No distributed version of routing algorithm
- Can use any algorithm
 - Dijkstra is recommended and normally used
 - All routers in AS must use same algorithm

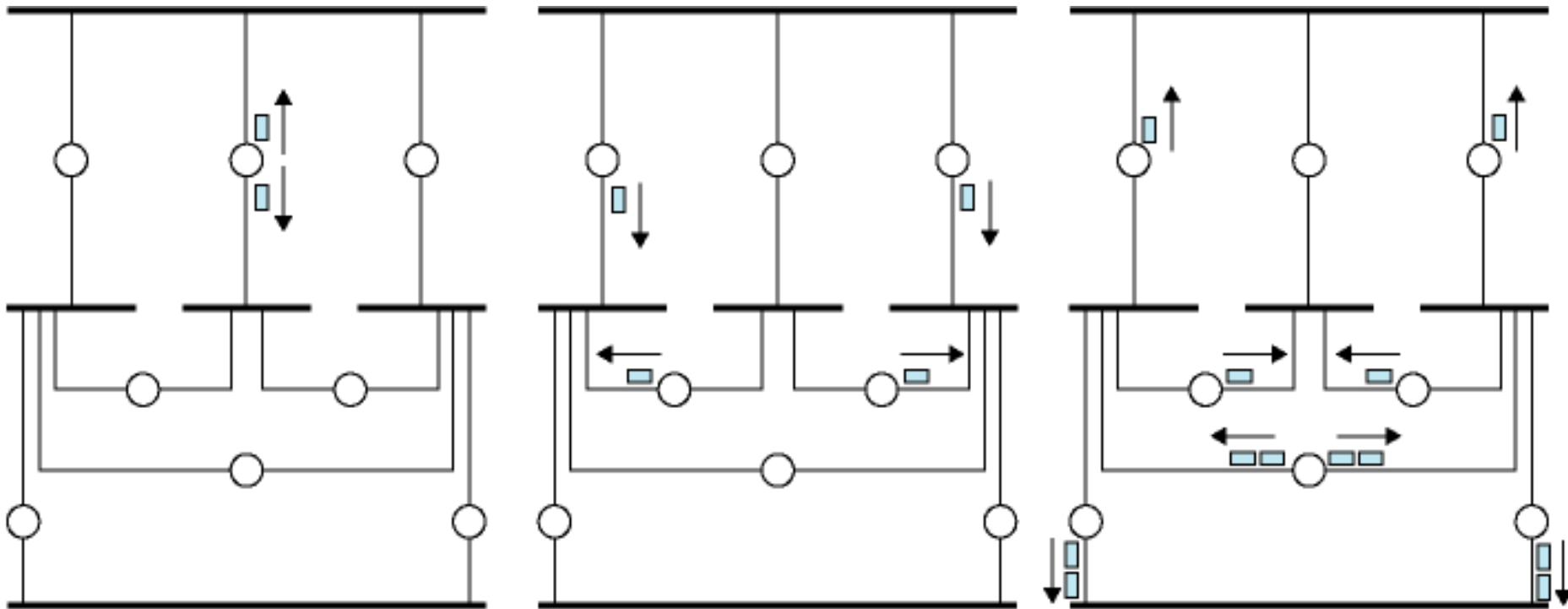


Flooding

- Packet sent by source router to every neighbor
- Incoming packet resent to all outgoing links except source link
- Duplicate packets already transmitted are discarded
 - Prevent incessant retransmission
- All possible routes tried so packet will get through if route exists
 - Highly robust
- At least one packet follows minimum delay route
 - Reach all routers quickly
- All nodes connected to source are visited
 - All routers get information to build routing table
- High traffic load



Flooding Example



(a) First hop

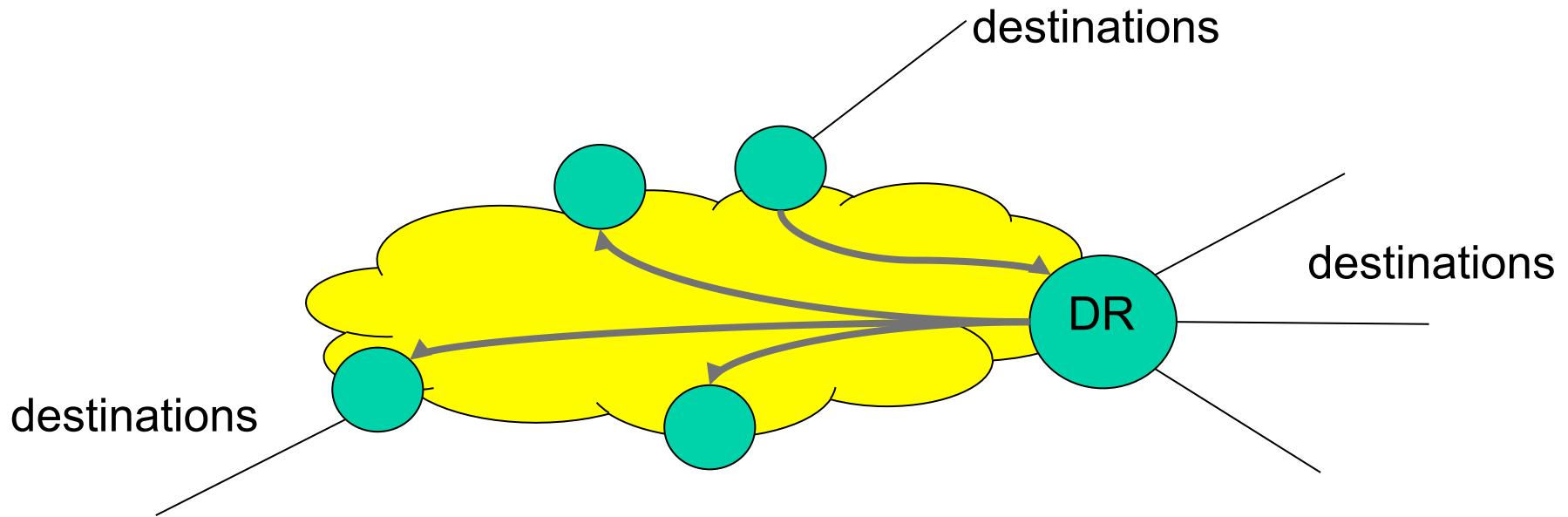
(b) Second hop

(c) Third hop



Alternative to flooding

- Designated Router (DR) election (with backup-DRB)
- Used on broadcast domains
- Link-State updates are sent to DR/DRB only which diffuse to all others (unicast confirmed communications)



OSPF Overview

- Router maintains descriptions of state of local links
- Transmits updated state information to all routers it knows about (flooding)
- Router receiving update must acknowledge
 - Lots of traffic generated
- Each router maintains database
 - Directed graph

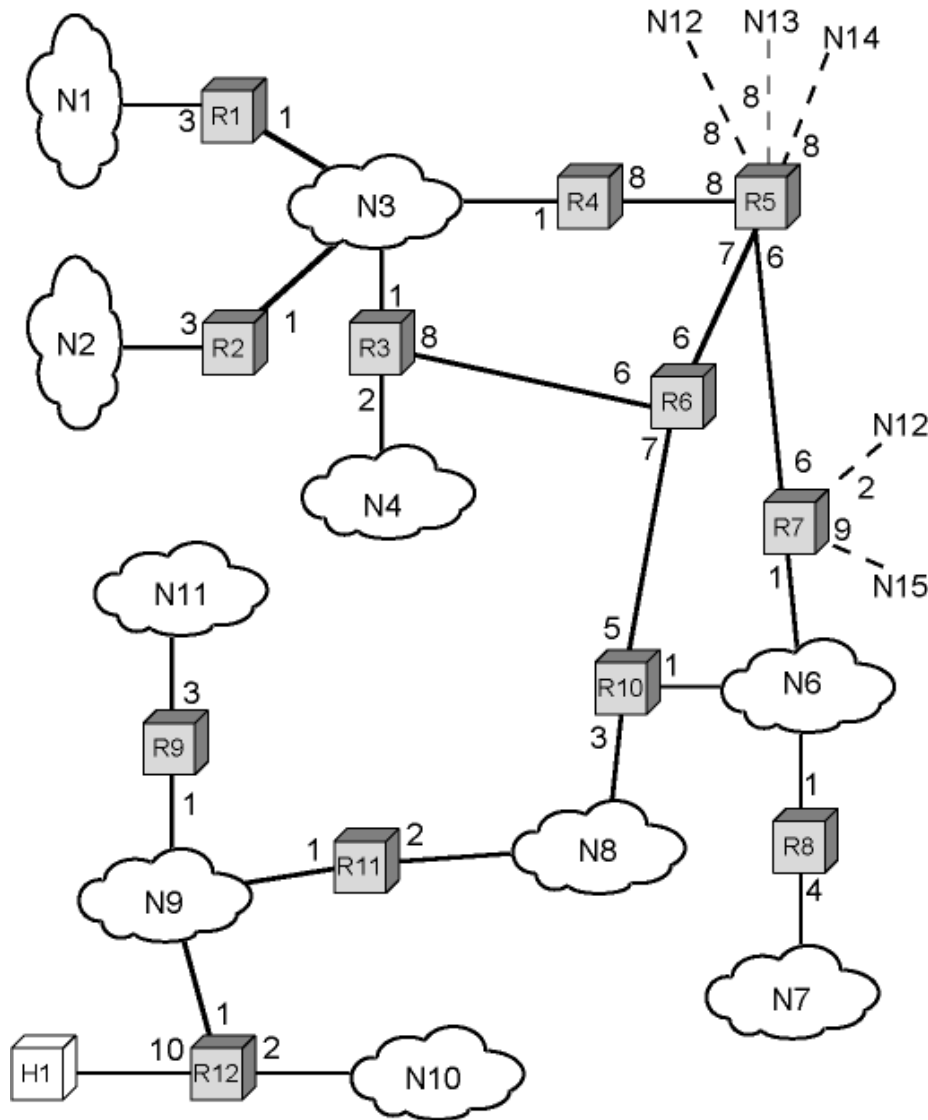


Router Database Graph

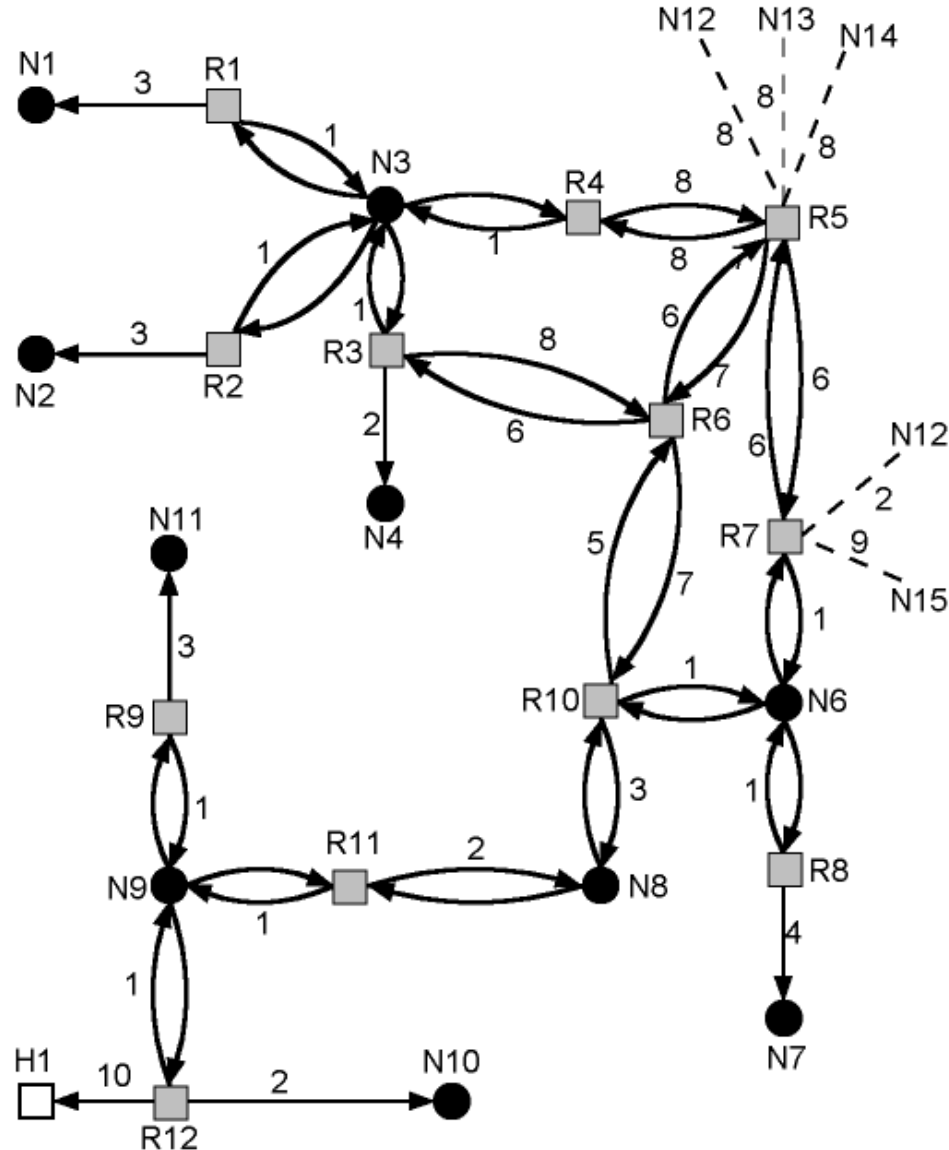
- Vertices
 - Router
 - Network
 - Transit
 - Stub
- Edges
 - Connecting two routers
 - Connecting router to network
- Built using link state information from other routers



Sample Autonomous System



Directed Graph of Autonomous System in previous slide

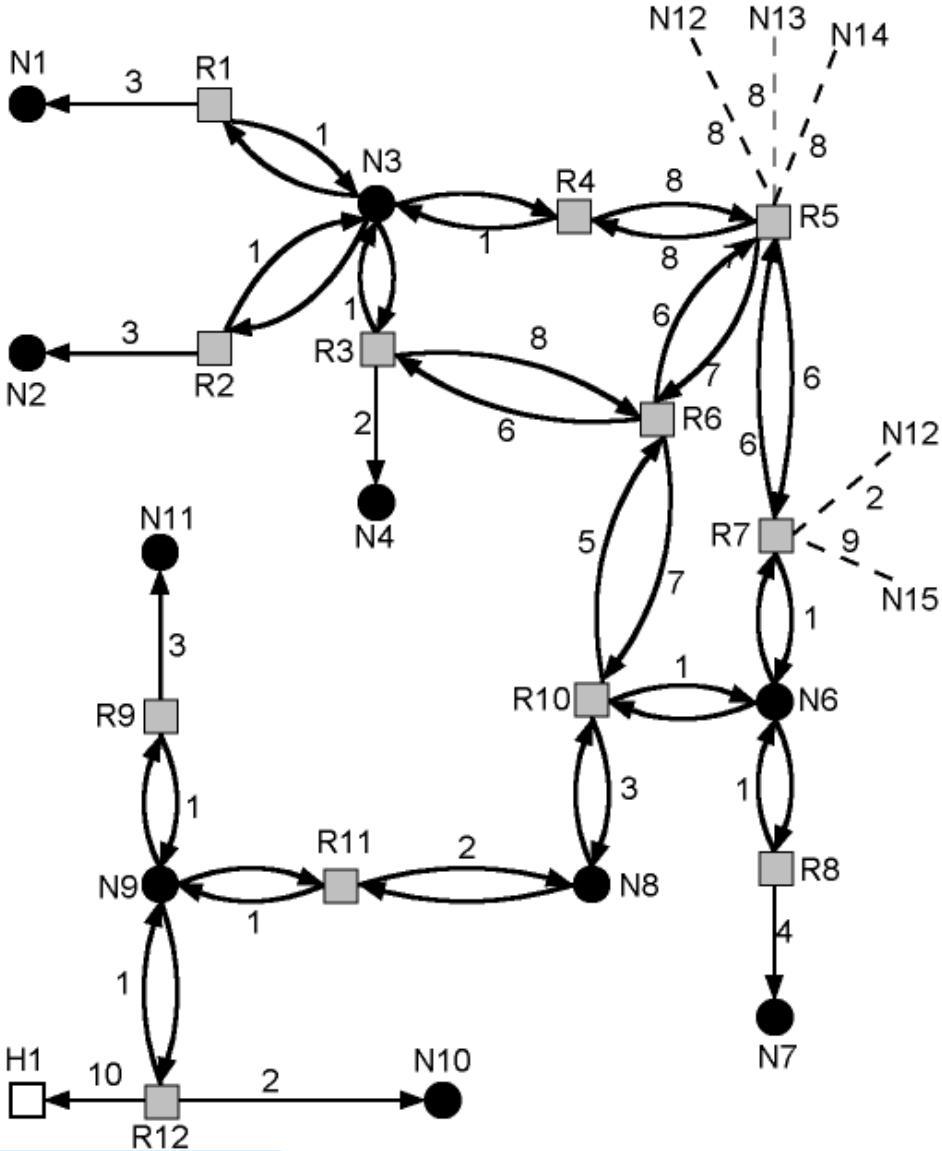


Link Costs

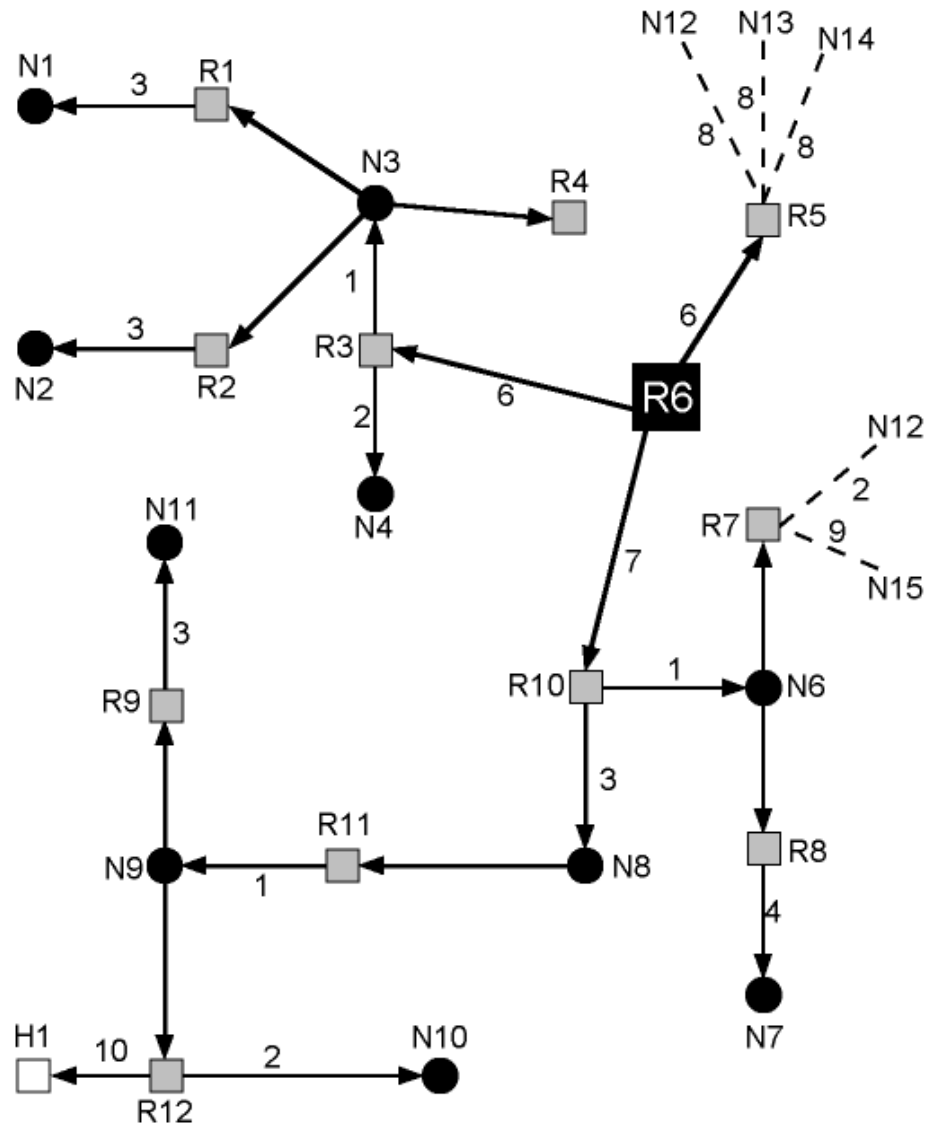
- Cost of each hop in each direction is called routing metric
- OSPF provides flexible metric scheme based on type of service (TOS)
 - Normal (TOS 0)
 - Minimize monetary cost (TOS 2)
 - Maximize reliability (TOS 4)
 - Maximize throughput (TOS 8)
 - Minimize delay (TOS 16)
- Each router can generate 5 spanning trees (and 5 routing tables) - AS decision!



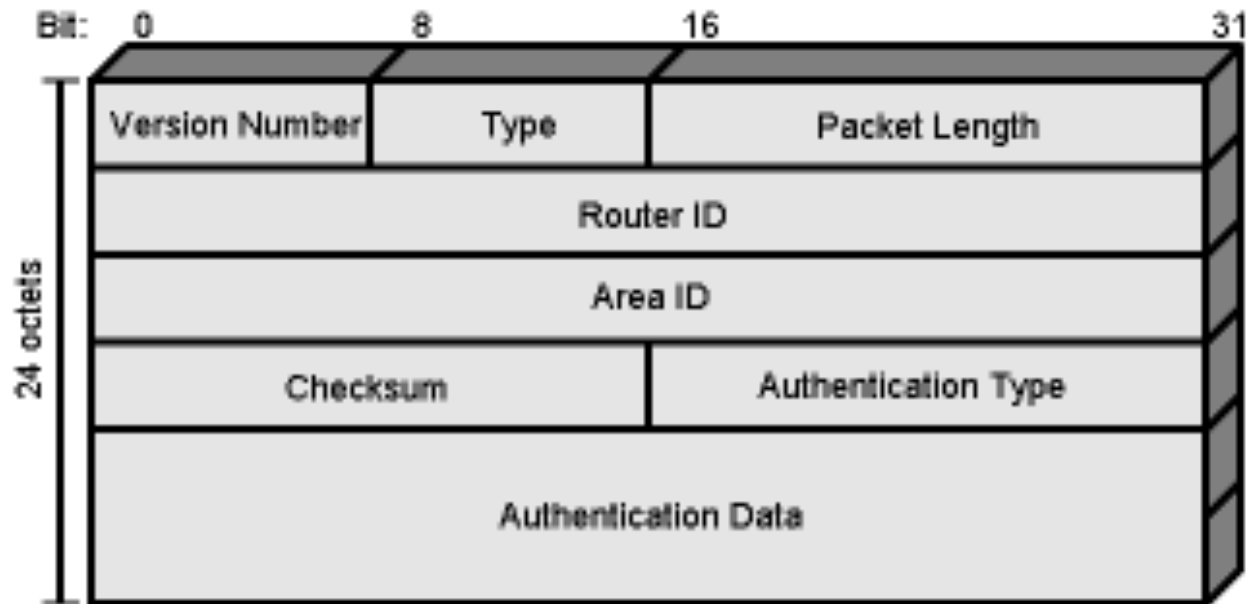
What is the SP for Router 6?



The Tree for Router R6



OSPF Packet Header



Packet Format Notes

- Version number: 2 is current
- Type: one of 5, see next slide
- Packet length: in octets including header
- Router id: this packet's source, 32 bit
- Area id: Area to which source router belongs
- Authentication type:
 - Null
 - Simple password
 - Encryption
- Authentication data: used by authentication procedure



OSPF Packet Types

1. Hello: used in neighbor discovery
2. Database description: Defines set of link state information present in each router's database
3. Link state request
4. Link state update
5. Link state acknowledgement

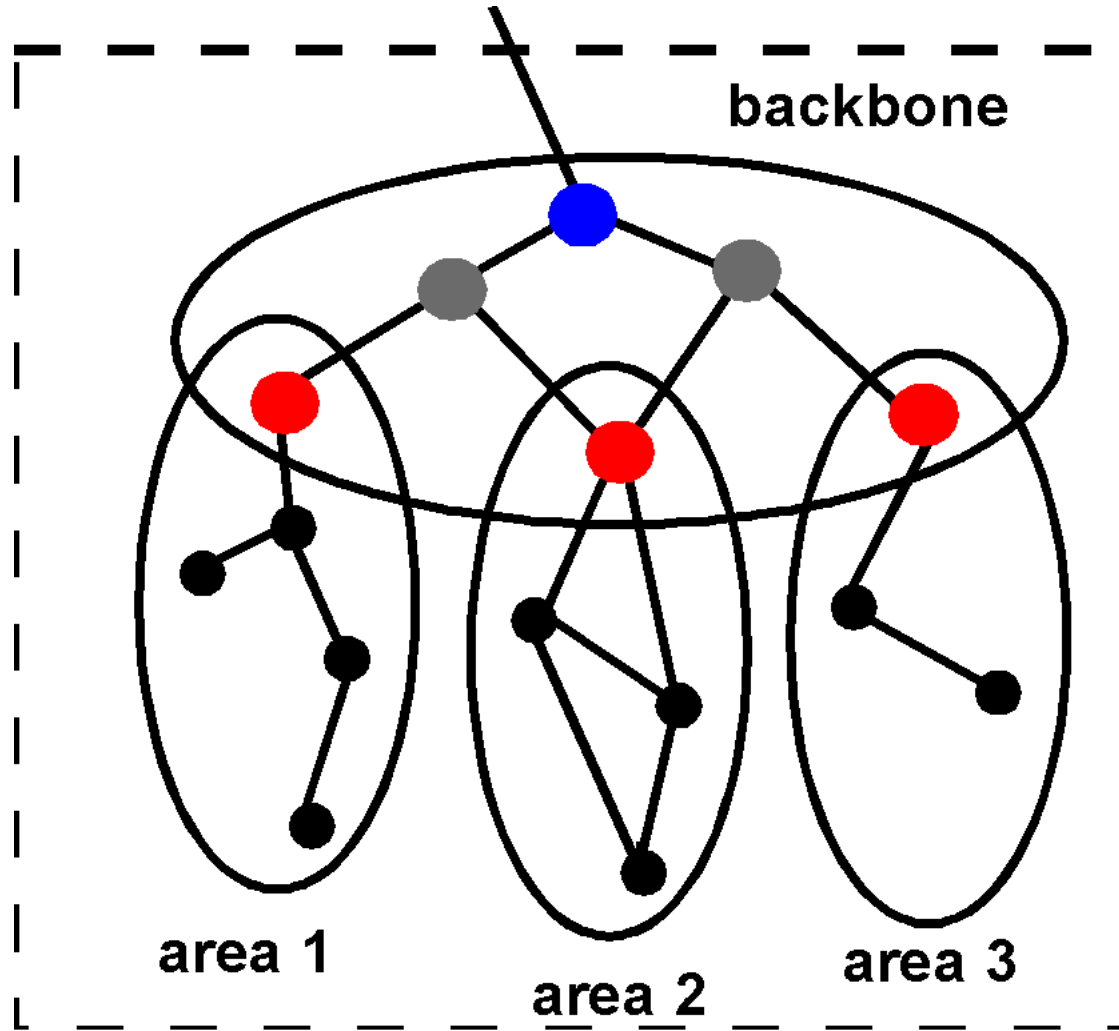


Areas

- Make large internets more manageable
- Configure as backbone and multiple areas
- Area - Collection of contiguous networks and hosts plus routers connected to any included network
- Backbone - contiguous collection of networks not contained in any area, their attached routers and routers belonging to multiple areas



Hierarchical OSPF



Operation of Areas

- Each area runs a separate copy of the link state algorithm
 - Topological database and graph of just that area
 - Link state information broadcast to other routers in area
 - Reduces traffic
 - Intra-area routing relies solely on local link state information



Inter-Area Routing

- Path consists of three legs
 - Within source area
 - Intra-area
 - Through backbone
 - Has properties of an area
 - Uses link state routing algorithm for inter-area routing
 - Within destination area
 - Intra-area

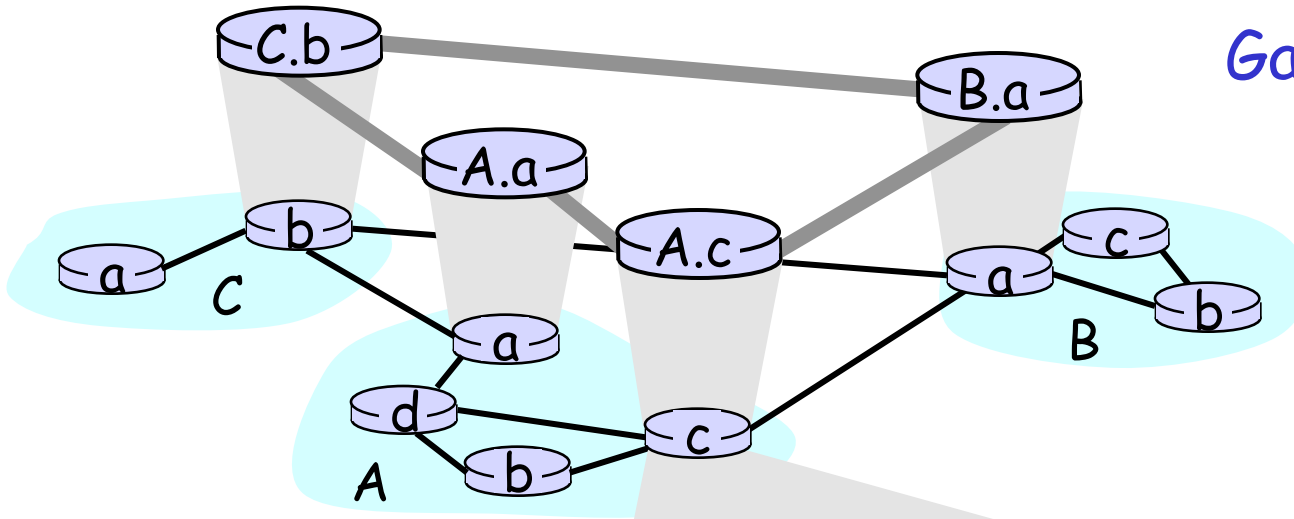


Hierarchical OSPF

- Two level hierarchy: local area and backbone
- Link state advertisements do not leave respective areas
- Nodes in each area have detailed area topology; they only know direction (shortest path) to networks in other areas
- **Area Border routers** “summarize” distances to networks in the area and advertise them to other Area Border routers
- **Backbone routers** run an OSPF routing alg limited to the backbone
- **Boundary routers** connect to other ASs



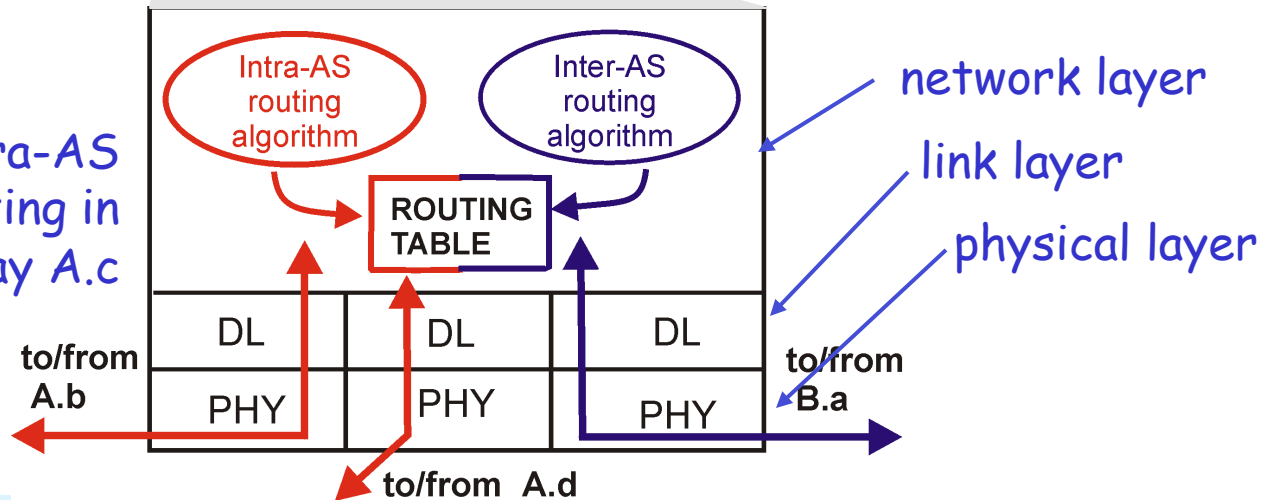
Intra-AS and Inter-AS routing



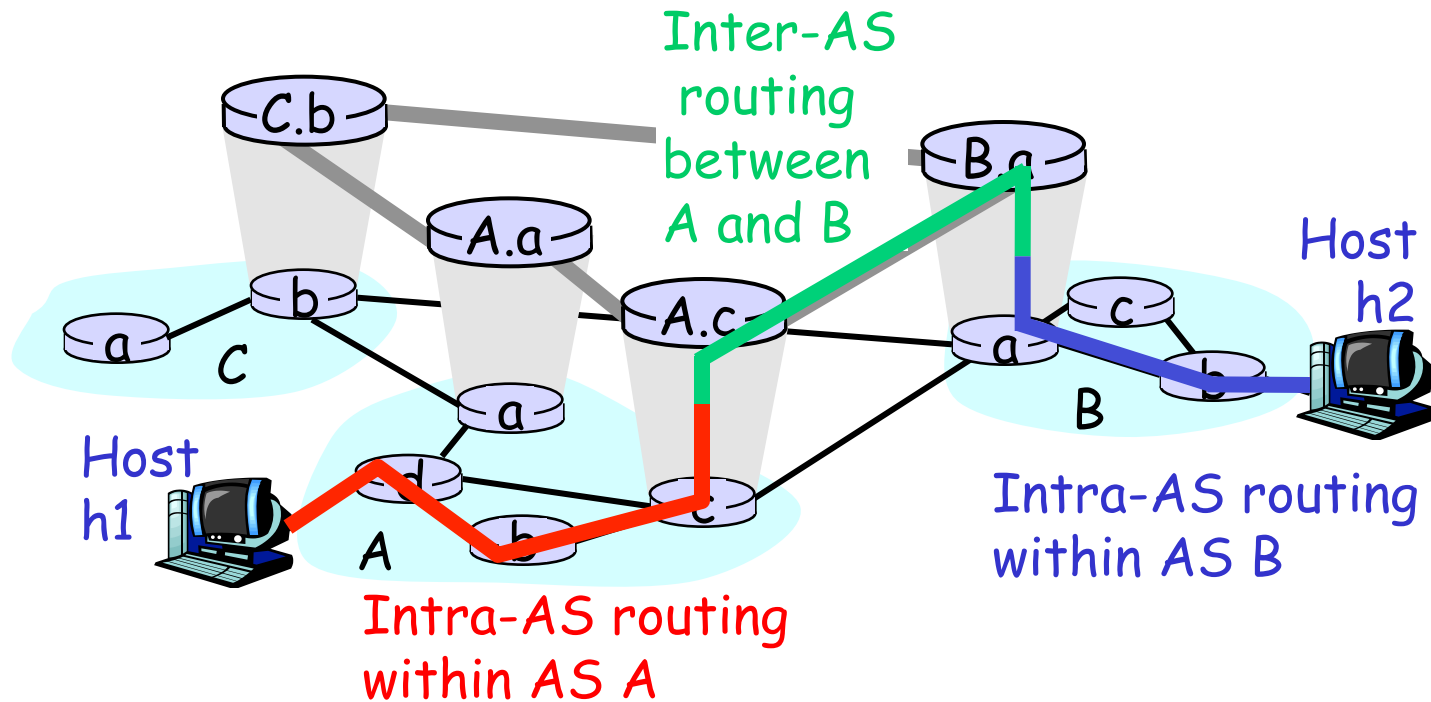
Gateways:

- perform inter-AS routing amongst themselves
- perform intra-AS routing with other routers in their AS

inter-AS, intra-AS routing in gateway A.c



Intra-AS and Inter-AS routing



- We'll examine specific inter-AS and intra-AS Internet routing protocols shortly



Inter-AS routing

- BGP (Border Gateway Protocol): the de facto standard
- **Path Vector** protocol - an extension of Distance Vector
- Each Border Gateway broadcast to neighbors (peers) the entire path (ie, sequence of AS' s) to destination
- For example, Gwy X may store the following path to destination Z:

$$\text{Path (X,Z)} = X, Y_1, Y_2, Y_3, \dots, Z$$



- Now, suppose Gwy X send its path to peer Gwy W
- Gwy W may or may not select the path offered by Gwy X, because of cost, policy or loop prevention reasons
- If Gwy W selects the path advertised by Gwy X, then:

$$\text{Path (W,Z)} = w, \text{Path (X,Z)}$$

Note: path selection based not so much on cost (eg, # of AS hops), but mostly on administrative and policy issues (eg, do not route packets of competitor's AS)



Why different Intra- and Inter-AS routing ?

- **Policy:** Inter is concerned with policies (which provider we must select/avoid, etc). Intra is contained in a single organization, so, no policy decisions necessary
- **Scale:** Inter provides an extra level of routing table size and routing update traffic reduction above the Intra layer
- **Performance:** Intra is focused on performance metrics; needs to keep costs low. In Inter it is difficult to propagate performance metrics efficiently (latency, privacy etc). Besides, policy related information is more meaningful.

We need **BOTH!**



Border Gateway Protocol (BGP)

- Allows routers (gateways) in different ASs to exchange routing information
- Messages sent over TCP
 - Messages in next slide
- Three functional procedures
 - Neighbor acquisition
 - Neighbor reachability
 - Network reachability



BGP Messages

- Open
 - Start neighbor relationship with another router
- Update
 - Transmit information about single route
 - List multiple routes to be withdrawn
- Keepalive
 - Acknowledge open message
 - Periodically confirm neighbor relationship
- Notification
 - Send when error condition detected
 - Used for closing connections too



Neighbor Acquisition

- Neighbors attach to same subnetwork
- If in different ASs routers may wish to exchange information
- Neighbor acquisition is when two neighboring routers agree to exchange routing information regularly
 - Needed because one router may not wish to take part
- One router sends request, the other acknowledges
 - Knowledge of existence of other routers and need to exchange information established at configuration time or by active intervention

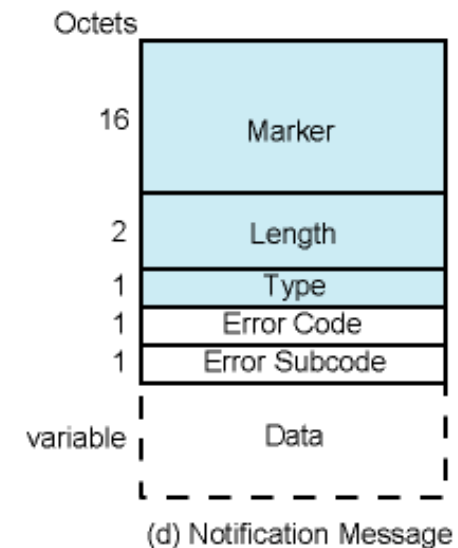
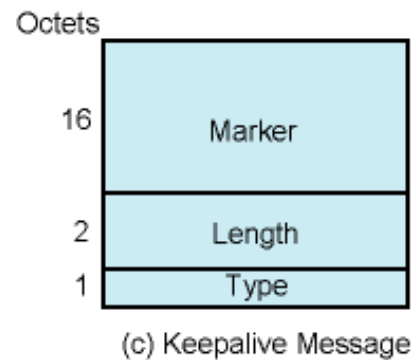
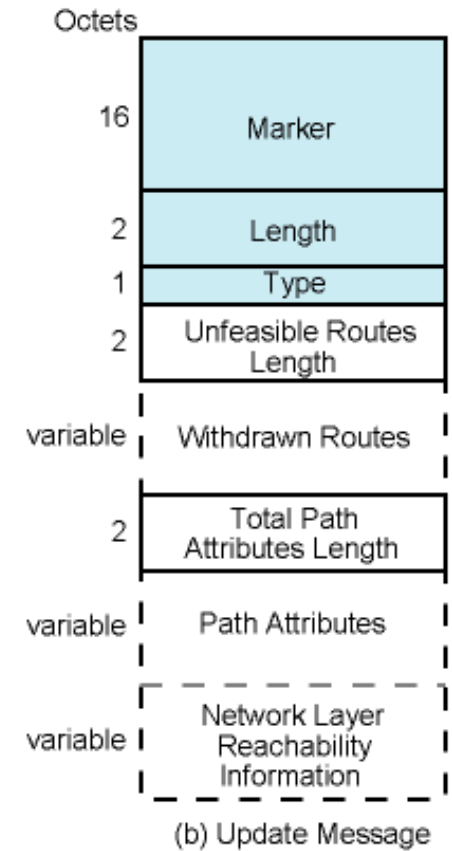
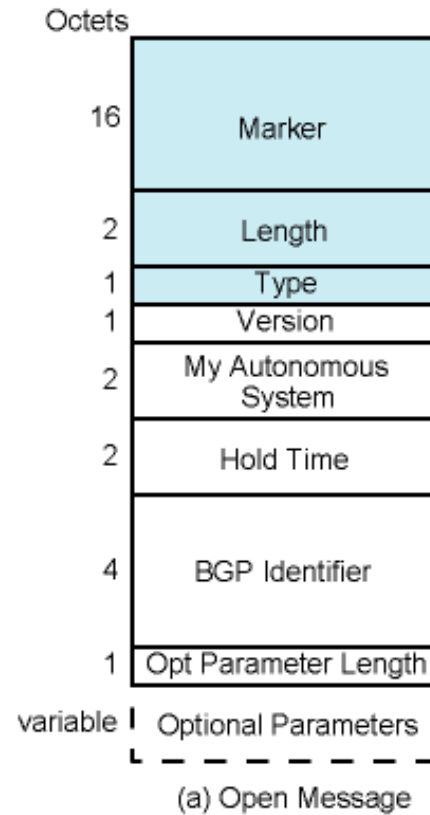


Neighbor Reachability

- Periodic issue of keepalive messages
- Between all routers that are neighbors
- Each router keeps database of subnetworks it can reach and preferred route
- When change is made, router issues update message (to neighbors only)
- All BGP routers build up and maintain routing information



BGP Message Formats



Neighbor Acquisition Detail

- Router opens TCP connection with neighbor
- Sends open message
 - Identifies sender's AS and gives IP address
 - Includes Hold Time
 - As proposed by sender
- If recipient prepared to open neighbor relationship
 - Calculate hold time
 - \min [own hold time, received hold time]
 - Max time between keepalive/update messages
 - Reply with keepalive



Keepalive Detail

- Header only
 - Enough to prevent hold time expiring
 - If hold time expires a topology change is triggered
-
- 'Marker' is a field that used for authentication purposes



Update Detail

- Information about single route through internet
 - Information to be added to database of any recipient router
 - Network layer reachability information (NLRI)
 - List of network portions of IP addresses of subnets reached by this route
 - Total path attributes length field
 - Path attributes field (next slide)
- List of previously advertised routes being withdrawn
- May contain both



Path Attributes Field

- Origin
 - Interior (e.g. OSPF) or exterior (BGP) protocol
- AS_Path
 - ASs traversed for this route
- Next_Hop
 - IP address of boarder router for next hop
- Multi_Exit_disc
 - Information about routers internal to AS
- Local_Pref
 - Tell other routers within AS degree of preference
- Atomic_Aggregate, Aggregator
 - Uses subnet addresses in tree view of network to reduce information needed in NLRI



Withdrawal of Route(s)

- Route identified by IP address of destination subnetwork(s)
- May be issued because subnets are not reachable or because policies have changed



Notification Message

- Error notification
- Message header error
 - Includes authentication and syntax errors
- Open message error
 - Syntax errors and option not recognised
 - Proposed hold time unacceptable
- Update message error
 - Syntax and validity errors
- Hold time expired
- Finite state machine error
- Cease
 - Close connection in absence of any other error



BGP Routing Information Exchange

- R1 constructs routing table for AS1 using OSPF
- R1 issues update message to R5 (in AS2)
 - AS_Path: identity of AS1
 - Next_Hop: IP address of R1
 - NLRI: List of all subnets in AS1
- Suppose R5 has neighbor relationship with R9 in AS3
- R5 forwards information from R1 to R9 in update message
 - AS_Path: list of ids {AS2,AS1}
 - Next_Hop: IP address of R5
 - NLRI: All subnets in AS1
- R9 decides if this is preferred route and forwards to neighbors



Routing Domain Confederations

- Set of connected AS
- Appear to outside world as single AS
 - Recursive
- Effective scaling

