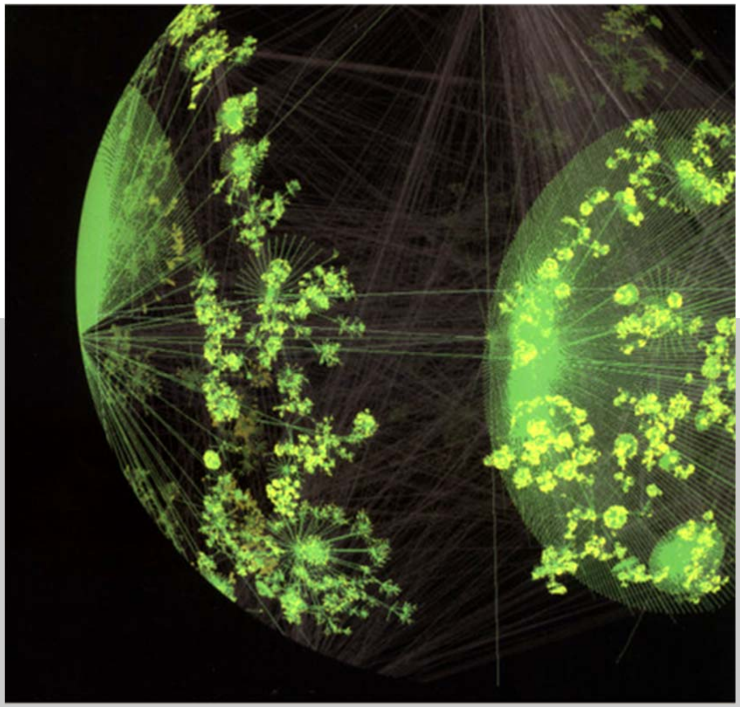


Chapter 4
Static and Dynamic Routing

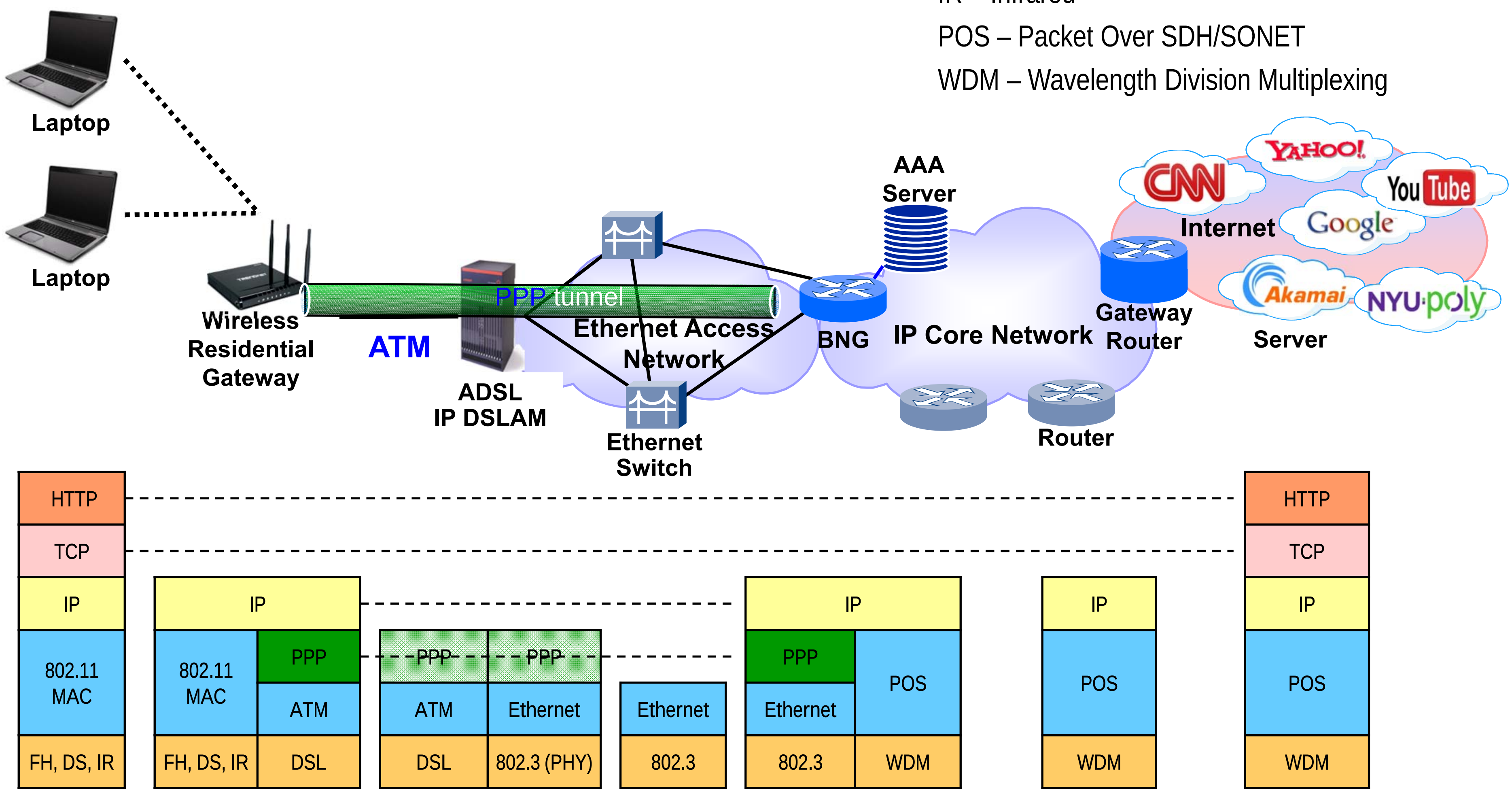
TCP/IP Essentials

A Lab-Based Approach

Spring 2017

IP Networking Example
– High Speed Internet Web Browsing

AAA – Authentication, Authorization, Accounting
BNG – Broadband Network Gateway
DSL – Digital Subscriber Line
DSLAM – DSL Access Multiplexer
DSSS – Direct Sequencing Spread Spectrum
FHSS – Frequency Hopping Spread Spectrum
IR – Infrared
POS – Packet Over SDH/SONET
WDM – Wavelength Division Multiplexing



Routing

Routing is to transfer packets from a source to a destination using network layer protocol information.

Two activities:

- Determine optimal routing paths
- Transport packets through an internetwork

Routing table

- Records optimal routes.
- Gets consulted when a forwarding decision is to be made.
- Can be set manually, updated by some ICMP messages, or by using dynamic routing protocols.

Packet Forwarding from Source to Destination

- Find out the IP address by DNS query for a given domain name of the destination
- If the destination is
  - in the same network (or subnet), send the packet directly to the destination
  - in a different network, a router is needed to forward the datagram
    - If no router available, drop the packet
- IP packets have to be encapsulated in a link layer frame (e.g., Ethernet frame)
  - A link layer frame can only be sent within the same network (or subnet)
  - The MAC address of the other end is required for sending the link layer frame
    - ARP

Communications in the Same Network/Subnet
What is "the Same Network/Subnet"?

- Host X wants to send packets to host Y
- What does the X know
  - X's IP address
  - X's subnet mask
  - Y's IP address
- Computation by X
  - X's network/subnet ID: (X's IP add) & (X's subnet mask)
  - Y's network/subnet ID: (Y's IP add) & (X's subnet mask)
  - If the above two results are the same, X believes that Y is in the same network/subnet
- If X and Y have different subnet masks, they may have different calculation results
  - Each calculates network/subnet ID by using its own subnet mask

Communications between Two Network Segments (in the Same Network)

- Two LAN segments connected by a bridge, host X in segment 1 and host Y in segment 2
- Assume that at the beginning,
  - the ARP tables of X and Y are empty
  - the bridge has correct entries for X and Y in its filtering database
- X tries to send an IP packet to Y
  - X broadcasts an ARP request
  - The bridge forwards the ARP request to segment 2 and all other connected segments
  - Y sends an ARP reply destined to X
  - The bridge forwards the ARP reply to segment 1
  - X sends out the Ethernet frame
  - The bridge forward the frame to segment 2
- In each packet, what are the values in the following fields?
  - IP: source IP address, destination IP address
  - ARP: sender IP address, target IP address
  - Ethernet for ARP: source Ethernet address, destination Ethernet address
  - Ethernet for IP: source Ethernet address, destination Ethernet address

Next-Hop Routing

- Direct delivery: send datagram directly through Layer 2 (Ethernet, …) when the source and the destination are on the same (sub)network.
- Indirect delivery: when the source and the destination are NOT on the same network
  - Need to send datagram through a router.
  - Consult the routing table to determine the next hop router.
  - Only ONE hop on the path is listed in the routing table.

Routing Table Entries

- Destination address – a specific host or network IP address
- Next hop address - the IP address of the next-hop router, or of a directly connected network.
- Flags:
  - U: route is up
  - G: route points toward a gateway (router); if this flag is not set, destination is directly connected
  - H: route points to a host, i.e., destination address is the complete host address; if this flag is not set, route is to a network and destination address is a netID or subnetID
  - D: route created by redirect
  - M: route modified by redirect
- Interface - the name of the interface for the next hop

Routing Table Lookup

To route each IP packet, the destination IP address is first extracted and then

- The network prefix gets calculated to determine whether the network prefix matches any directly connected network address so the packet can be delivered directly.
- If not direct delivery, a routing table lookup takes place in the following order named as the Longest-prefix-matching rule
  - Find matching host address
  - Find matching network address
  - Find default entry
- To keep the routing table small, network-specific entries and default router are often used

Destination	Gateway	Genmask	Flags	MSS	Window	irtt	Iface
128.238.4.0	0.0.0.0	255.255.255.0	U	40	0	0	eth0
127.0.0.0	0.0.0.0	255.0.0.0	U	40	0	0	lo
0.0.0.0	128.238.4.4	0.0.0.0	UG	40	0	0	eth0

Statically Setting IP Routing Tables

- Static Routing: set IP routing table without a routing protocol
- Use static routing when
  - The network is small
  - Only a single connection point to other networks
- Ways to set IP routing tables with static routing
  - By default when the interface is configured during bootstrap
    - e.g., using the Dynamic Host Configuration Protocol (DHCP)
  - Use route command from the system bootstrap file
  - Via ICMP redirect messages
  - Via ICMP router advertisement/router discovery messages

ICMP Redirect (RFC792)

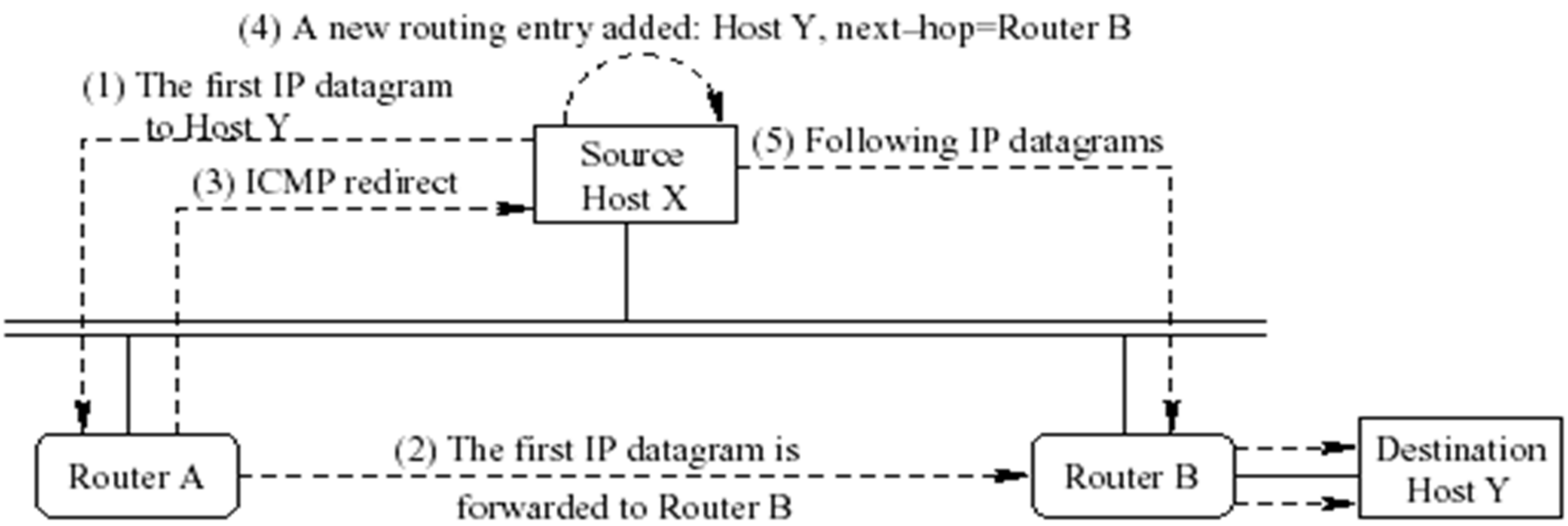
- A router sends an ICMP redirect error message to the sender if the datagram should have been sent to another router.
- Allows the host to update its routing table with a better path
  - When the host may start with a default router
  - When the network topology changes

0 — 7	8 — 15	16 — 31
type (5)	code (0 - 3)	checksum
Connected gateway IP address		
IP header (including options) plus the first 8 bytes of the original IP datagram payload		

- Codes
  - 0: Redirect Datagram for the Network (or subnet)
  - 1: Redirect Datagram for the Host
  - 2: Redirect Datagram for the Type of Service and Network
  - 3: Redirect Datagram for the Type of Service and Host

ICMP Redirect Example

- Host X uses Router A as its default router
- Host X sends a datagram destined to Host Y
- Router A looks up it routing table
  - Router B is the next-hop router
  - The datagram is sent out on the same interface it was received on
- Router A sends a ICMP redirect message to Host X
- Host X update the routing entry for Host Y, with a D flag



ICMP Router Discovery (RFC 1256)

Used to configure the default route for a host when it bootstraps

- After bootstrapping a host broadcasts an ICMP router solicitation message
- In response, each router sends an ICMP router advertisement message
- Also, routers periodically broadcast ICMP router advertisement
- A host chooses one or more of the advertised addresses as its default router

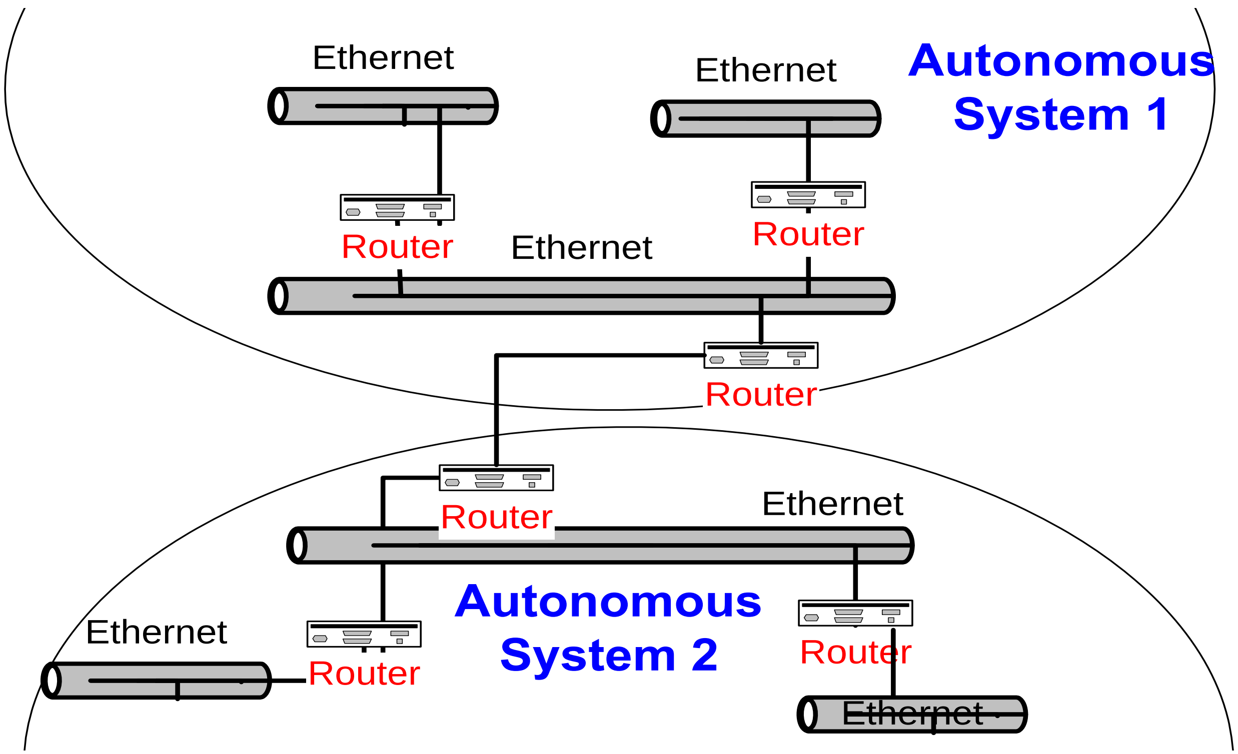
<table>
<tr><th>0   7</th><th>8   15</th><th>16   31</th></tr>
<tr><td>type (9)</td><td>code (0)</td><td>checksum</td></tr>
<tr><td>no. of addresses</td><td>address length (2)</td><td>lifetime</td></tr>
<tr><td colspan="3">router address [1]</td></tr>
<tr><td colspan="3">preference level [1]</td></tr>
<tr><td colspan="3">router address [2]</td></tr>
<tr><td colspan="3">preference level [2]</td></tr>
<tr><td colspan="3">… …</td></tr>
</table>

Dynamic Routing

- Routers communicate with each other
  - Using a routing protocol
  - Gain information about the network status and build their routing tables
- Dynamic routing is used to
  - Eliminate loops in paths, and
  - React to changes in the network topology

Autonomous Systems

- Internet is organized as a collection of Autonomous System (AS)
- An AS is a region of the Internet that is administered by a single entity.
  - e.g. an enterprise network or a campus network



Interdomain and Intradomain Routing

Routing is carried differently within an autonomous system and between autonomous system.

Intradomain Routing

- Routing within an AS
- Ignores the network outside the AS
- Protocols for intradomain routing are also called Interior Gateway Protocols or IGP's.
- Popular protocols are
  - RIP (simple, old)
  - OSPF (better)

Interdomain Routing

- Routing between AS's
- Assumes that the Internet consists of a collection of interconnected AS's
- Normally, there is at least one dedicated router (AS Boundary Router) in each AS that handles interdomain traffic.
- Protocols for interdomain routing are also called Exterior Gateway Protocols or EGP's.
- Routing protocols:
  - BGP (popular)

Routing Algorithms

- A routing algorithm forms the core of each dynamic routing protocol
- Use a "cost" metric to determine the optimal path to a destination
  - Path length, reliability, delay, bandwidth, load, communication cost
- Two types of routing algorithms
  - Distance Vector Routing
  - Link State Routing

Routing Algorithms (cont'd)

Goal: Given a network where each link, between two nodes i and j, is assigned a cost, find the path with the least cost between source node s and destination node d.

Parameters:

d_{ij}: Cost of link between node i and node j;

$d_{ij} = \infty$, if nodes i and j are not directly connected;

$d_{ii} = 0$.

N: Set of nodes in network.

Routing Algorithm Overview

Distance Vector

- Each node knows the distance (="cost") to its directly connected neighbors.
- Each node sends its neighbors a list of the current distances to all nodes in network.
- If all nodes eventually update their distances (including to those not directly connected), the routing tables get converged.

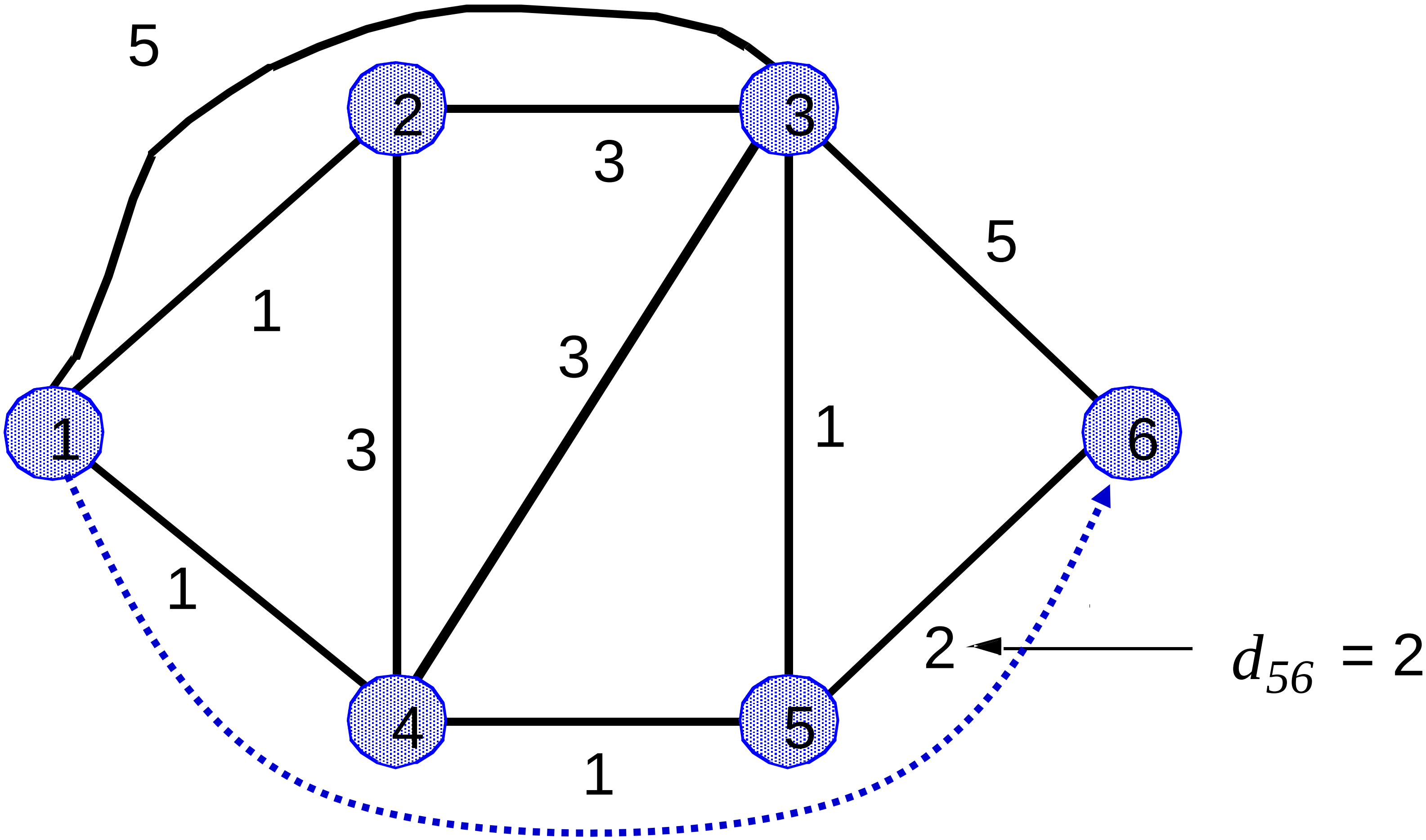
Link State

- Each node broadcasts distance information (i.e. link state) to ALL other nodes in the network
- All nodes in the same network have an identical database for the status of all links.
- Each router calculates the shortest path to all other destinations independently

Distance Vector Routing vs. Link State Routing

- Both work well in most circumstances
- Link state routing
  - Converges faster
  - Less prone to routing loops
- Distance vector routing
  - Requires less resources
  - Less cost to implement and support

Example



How does node 1 find the optimal path to node 6?

Distance Vector

- Each node maintains two tables:
  - Distance Table: Cost to each node via each outgoing link.
  - Routing Table: Minimum cost to each node and next hop node.
- Nodes exchange messages on the routing cost to each node (minimal info, only a part of the routing table)
- Reception of messages triggers recalculation of routing table

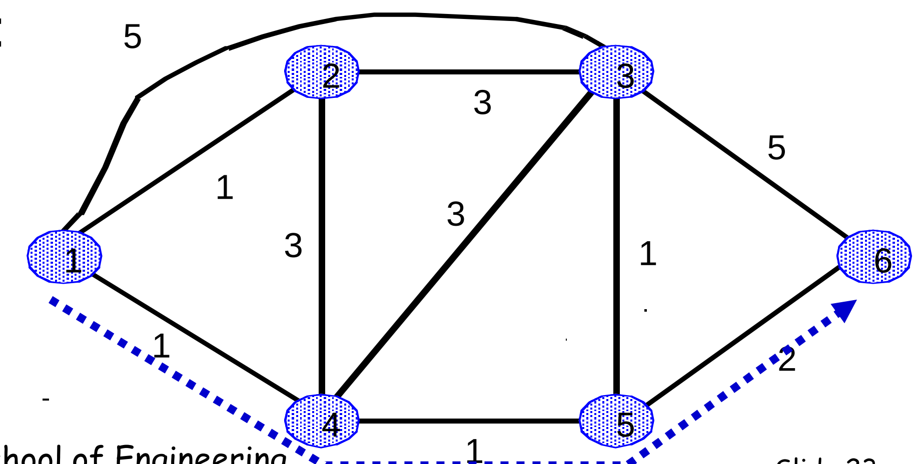
Distance Vector Solution Example

Show how the entries at node 1 change for node 6:

Time	Messages received about the dist. to node 6	Distance via			Routing	
		2	3	4	via	Cost
T = 0	No message received. Node 1 is not aware of Node 6.					
T = 1	Node 3 says the dist. is 5. Node 2 and 4 are not aware of Node 6.	∞	10	∞	3	10
T = 2	Node 2 says the dist. is 8; Node 3 and 4 both say the dist. is 3.	9	8	4	4	4
T = 3	Node 2 says the dist. is 6; Node 3 and 4 say the dist. is 3	7	8	4	4	4
T = 4	Node 2 says the dist. is 5; no change in the messages from Node 3 and 4	6	8	4	4	4

A routing table can be formed at node 1:

Destination	Next Hop	Cost
2	2	1
3	4	3
4	4	1
5	4	2
6	4	4



$D_x(y) = \min\{c(x,y) + D_y(y), c(x,z) + D_z(y)\}$
$= \min\{2+0 , 7+1\} = 2$

$D_x(z) = \min\{c(x,y) + D_y(z), c(x,z) + D_z(z)\}$
$= \min\{2+1 , 7+0\} = 3$

node x table

from \ cost to	x	y	z
x	0	2	7
y	∞	∞	∞
z	∞	∞	∞

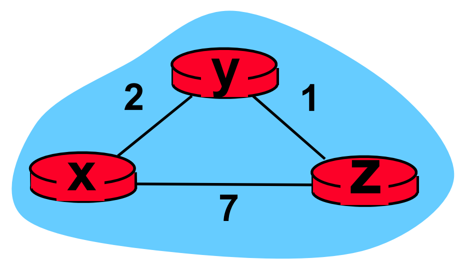
from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	7	1	0

node y table

from \ cost to	x	y	z
x		∞∞	∞
y	2	0	1
z	∞	∞	∞

node z table

from \ cost to	x	y	z
x	∞	∞	∞
y		∞∞∞	∞
z	7	1	0



time

*Slide reproduced from: <**C**omputer Networking – A **T**op **D**own **A**pproach>,*

$$D_x(y) = \min\{c(x,y) + D_y(y),\ c(x,z) + D_z(y)\} = \min\{2+0\ ,\ 7+1\} = 2$$

$$D_x(z) = \min\{c(x,y) + D_y(z),\ c(x,z) + D_z(z)\} = \min\{2+1\ ,\ 7+0\} = 3$$

node x table

from \ cost to	x	y	z
x	0	2	7
y	∞	∞	∞
z	∞	∞	∞

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	7	1	0

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	3	1	0

node y table

from \ cost to	x	y	z
x	∞	∞	∞
y	2	0	1
z	∞	∞	∞

from \ cost to	x	y	z
x	0	2	7
y	2	0	1
z	7	1	0

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	3	1	0

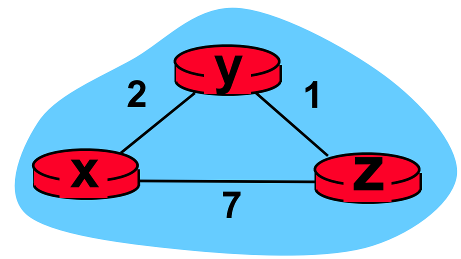
node z table

from \ cost to	x	y	z
x	∞	∞	∞
y	∞	∞	∞
z	7	1	0

from \ cost to	x	y	z
x	0	2	7
y	2	0	1
z	3	1	0

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	3	1	0

time



Discussion of Distance Vector Routing

- Entries of routing tables can change while a packet is being transmitted. This can lead to a single datagram visiting the same node more than once (Looping).
- If the period for updating the routing tables is too short, routing table entries can be changed before convergence (from the previous updates) is achieved.
  - Example: ARPANET used to have a Distance Vector algorithm with an update period of <1 sec. This resulted in instability of routing.
- Similar behave as using Bellman-Ford Algorithm which includes more hops in routing as the routing table gets updated

Link State Route Calculations

Calculate shortest path for node s

Dijkstra's Algorithm:

- s algorithm is executed at each (source) node
- N a set contains all nodes
- D_n cost of the least-cost path from node s to node n

$M = \{s\}$;

for each $n \notin M$

  $D_n = d_{sn}$;

while ($M \neq N$) do

  Find $w \notin M$ for which $D_w = min\{D_j : j \notin M\}$;

  Add w to M;

  for each $n \notin M$

    $D_n = min_w [\, D_n, D_w + d_{wn} \,]$;

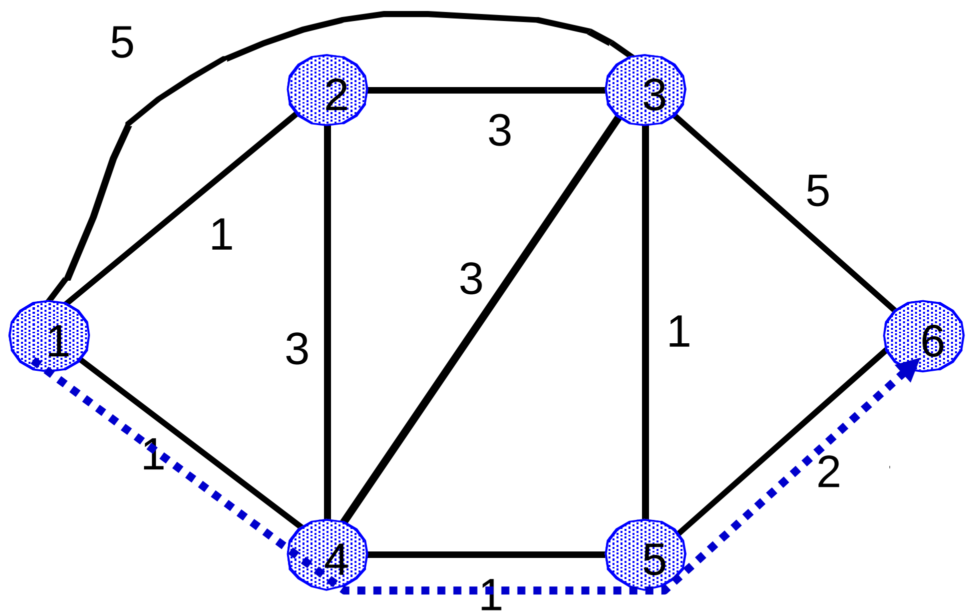
    Update route;

enddo

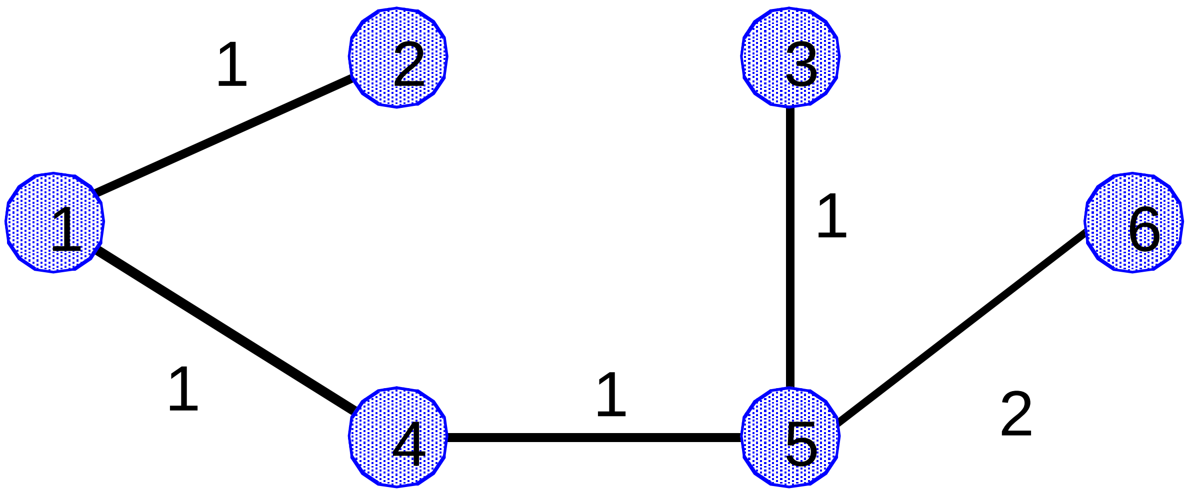
Link State Solution (at node 1)

Dijkstra's algorithm at node 1

	M	D1	D2	D3	D4	D5	D6
0	{1}	0	1	5	1	∞	∞
1	{1, 2}	0	1	4	1	∞	∞
2	{1, 2, 4}	0	1	4	1	2	∞
3	{1,2, 4, 5}	0	1	3	1	2	4
4	{1,2, 4, 5, 3}	0	1	3	1	2	4
5	{1, 2, 4, 5, 3, 6}	0	1	3	1	2	4



Resulting Routing Tree (at node 1)



The tree is translated into a routing table at node 1:

Destination	Next Hop	Cost
2	2	1
3	4	3
4	4	1
5	4	2
6	4	4

Discussion of Link State

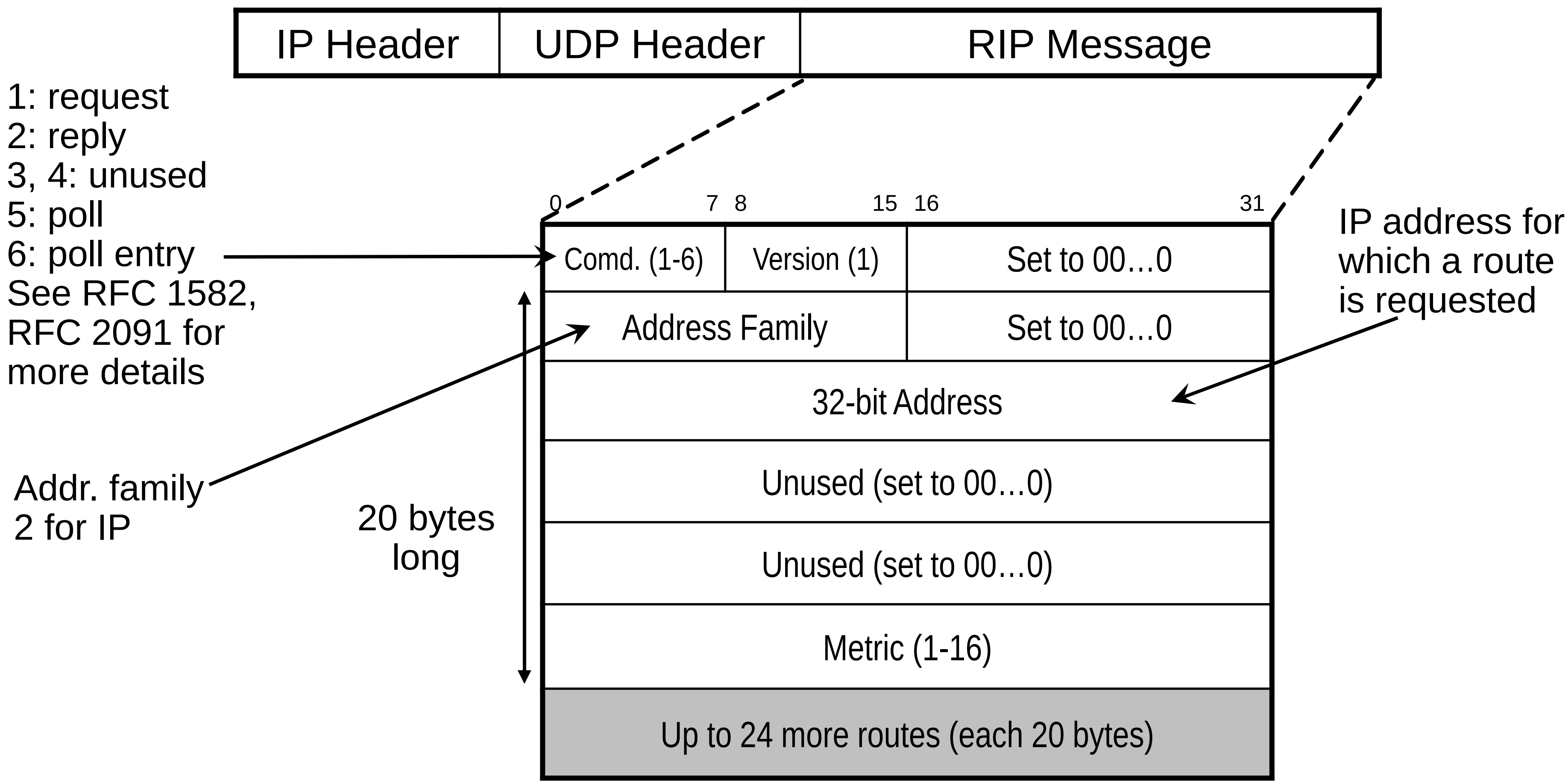
- Each node requires complete topology information.
- Link state information must be flooded to all nodes.
- Each node must maintain a global database.
- Convergence of the algorithm is guaranteed.

RIP - Routing Information Protocol

- A distance vector algorithm based protocol
- Uses path hop count as the routing metric
- Each link is assigned a hop-count value (typically 1)
- RIP routers maintain only the best route
- RIP-2 is the latest version; RIPng extends RIP-2 to support IPv6
- Each router sends routing update messages at regular intervals (default 30 sec.) and whenever the network topology changes
- Each router updates it routing table and send routing update messages to neighbors when receiving a routing message indicating a route change

RIP Packet Format

- RIP messages are encapsulated in UDP datagrams, port number 520



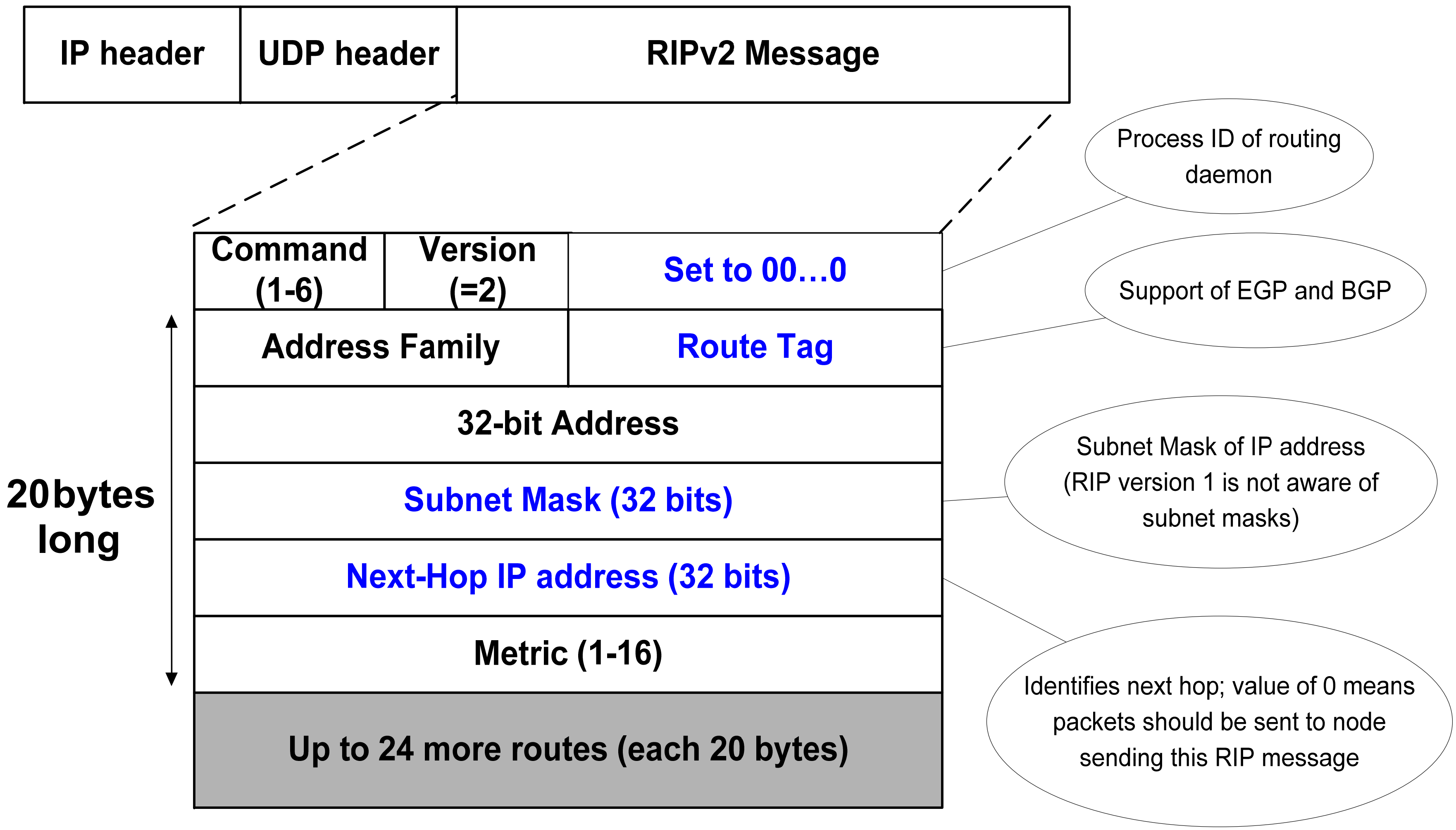
RIP Timers

- Route-update timer: clock the interval between periodic routing updates, generally set to 30 sec
- Route-invalid timer: make a route invalid if it is not updated over this period of time, default 180 sec
- Route-hold-down timer:
  - A route enters into a hold-down state when receiving an update packet indicating the route is unreachable → set route-hold-down timer
  - The timer specifies a interval during which routing information regarding better paths is suppressed,
  - The timer is at least 3 times the value of the update timer, default 180 sec
- Route-flush timer: the amount of time must pass before the route is removed from the routing table, default 240 sec

Routing with RIP

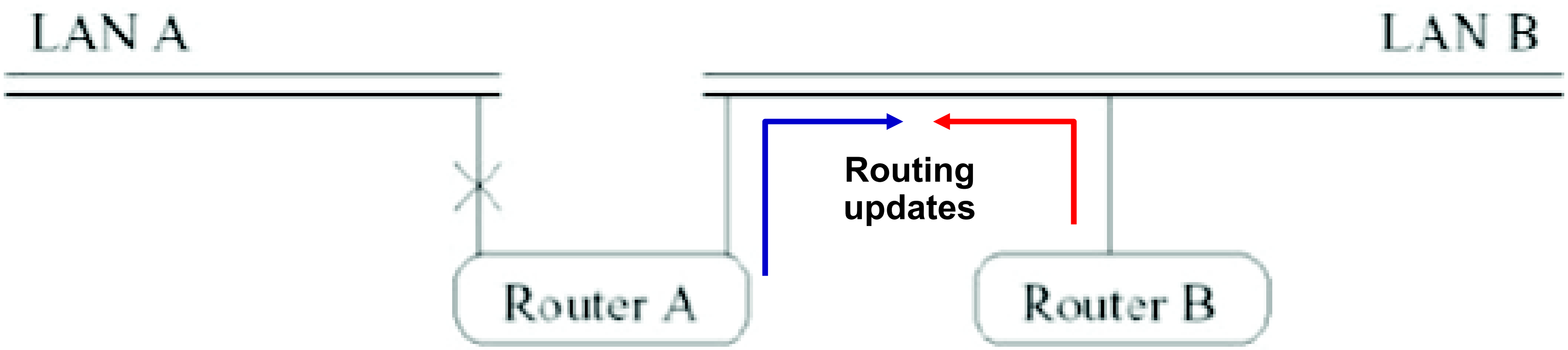
- RIP operation is supported in `routed` daemon with dedicated UDP port 520.
- Initialization: Broadcast a request packet (command = 1, address family = 0, metric=16) on the interfaces requesting current routing tables from routers.
- Request received: Routers that receive above request send their entire routing table.
- Response received: Update the routing table (see distance vector algorithm).
- Regular routing updates: Every 30 seconds, send all or part of the routing tables to every neighbor.
- Triggered Updates: Whenever the metric for a route changes, send data that has changed.

RIPv2



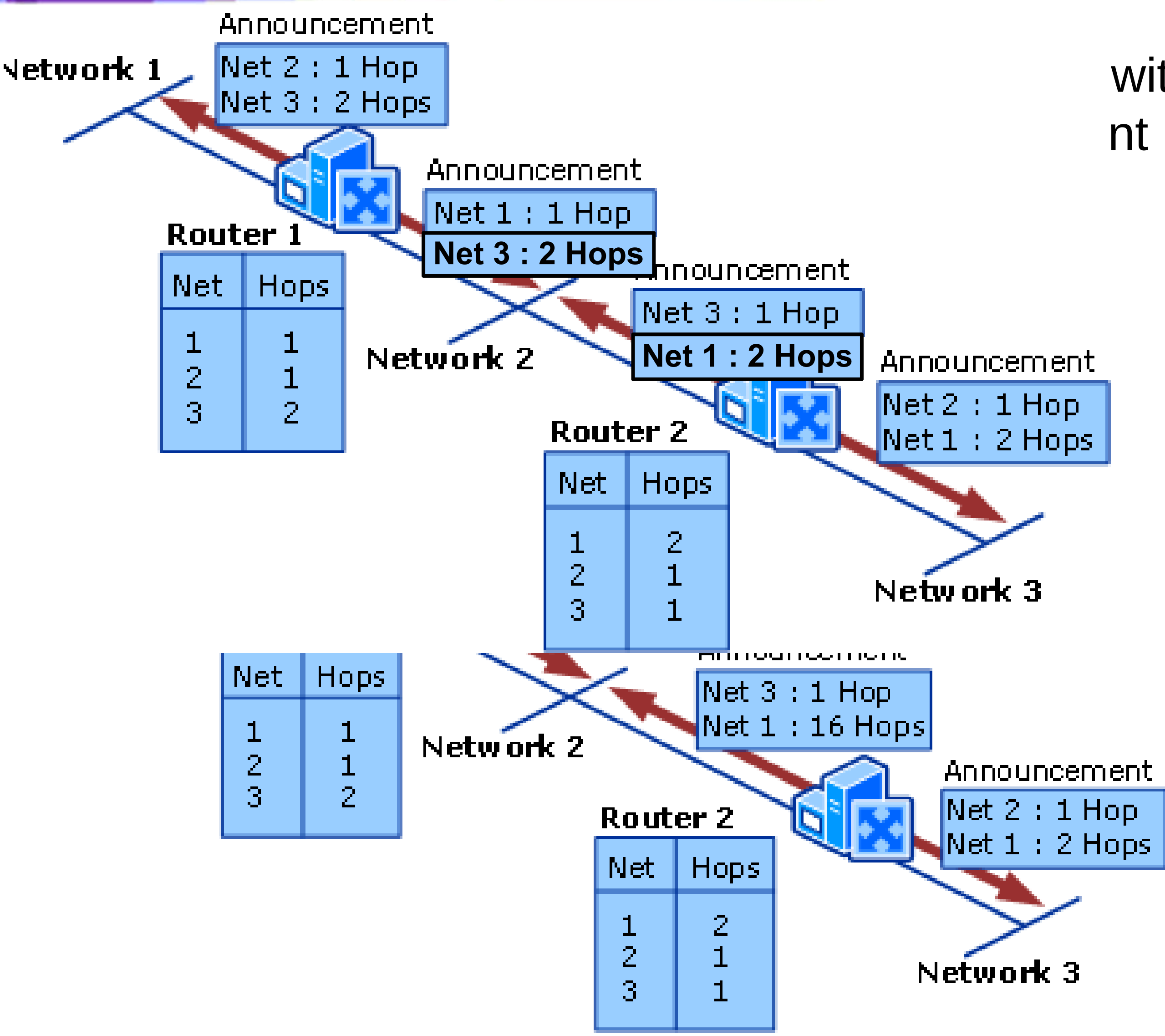
RIPv2 also supports multicast and provides authentication.

Count-to-Infinity Problem in RIP



- To resolve this problem, RIP uses a hop-count limit of 15
- When the path length reaches 16, consider it as unreachable
- Downside of the hop-count limit:
  - The size of the network is limited
  - Takes a long time for the routing tables to converge after a topology changes
- Use split horizon technique to improve the stability: information about a route is not allowed to be sent back in the direction from which it came

Split Horizon Announcement Process



with Poison Reverse
nt Process

Open Shortest Path First (OSPF)

- Open
  - Developed by IETF IGP working group, RFC2328
- SPF
  - Each router floods link-state information through its neighbors all to other routers
  - Based on the flooded link-state information, each router maintains a complete link-state database
  - Based on the link-state database, a routing table is constructed using SPF (e.g., Dijkstra's) algorithm
- Runs over IP directly, protocol number 89

OSPF Features

- Use flexible metrics instead of only hop count
- Supports variable-length subnetting
- Supports multiple routes
  - One for each IP Type of Service (ToS)
  - Allows load balancing among equal-cost paths
- Authenticates route exchanges
- Quick convergence
- Uses multicast rather than broadcast of its messages to reduce network load

OSPF Operations

- Each router sends OSPF Hello packets to neighbors after it is assured that its interfaces are functioning
- Each router also receives Hello packets from neighbors → let the router know that other routers are functional
- All routers periodically send Link State Advertisements (LSAs) to provide information on the link states, so that failed routers can be detected quickly
- By using the information in LSAs, a router
  - builds a topological database containing an overall picture of the area
  - Constructs a routing table using SPF (e.g., Dijkstra's) algorithm based on the link-state database

Hierarchical OSPF

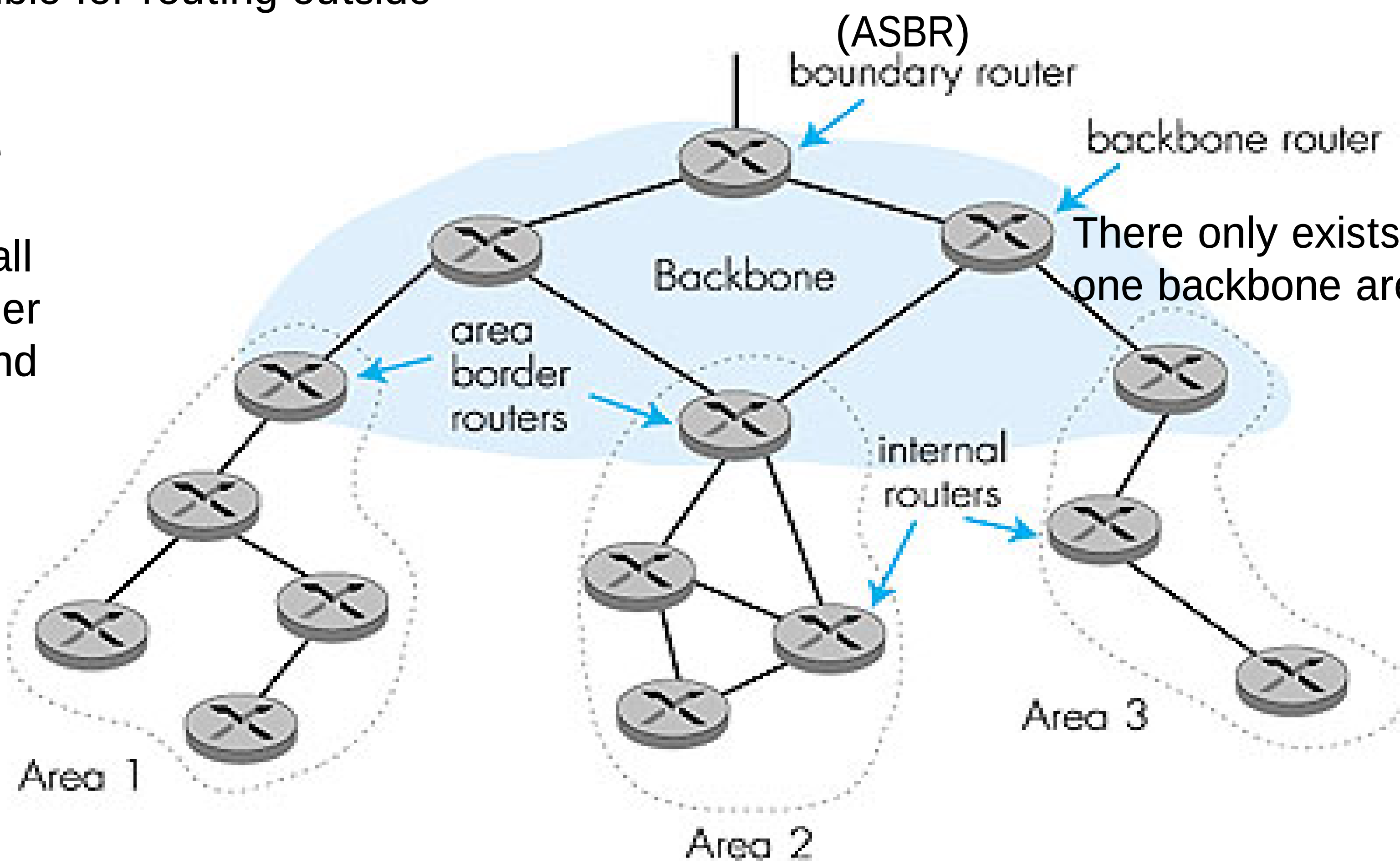
- An AS can be organized as two-level hierarchy under OSPF
  - AS is partitioned into self-contained areas
  - Areas are identified by a 32-bit area ID
  - Areas are interconnected by a backbone area with a reserved area ID 0.0.0.0
- Four types of routers
  - Internal router
  - Area border router
  - Backbone router
  - AS Boundary Router (ASBR)

Two Level Hierarchy OSPF AS

For each area, the border router is responsible for routing outside the area

Backbone area contains all area border routers and possibly others

There only exists one backbone area



OSPF Packets

- Five types of OSPF packets
  - Hello (type 1)
  - Database description (2)
  - Link-State Request(3) / Update(4) / Acknowledgement(5)
- OSPF common header

Version	Type (1-5)	Packet Length	
Router ID			
Area ID			
Checksum		Authentication Type	
Authentication			
Authentication			

OSPF Common Header Fields

- Version number: 2
- Type: type of OSPF packet
- Packet length: in bytes, includes OSPF header
- Router ID: 32-bit number assigned to each OSPF running router – uniquely identifies router within AS
- Area ID: any four-byte number (0.0.0.0 reserved for backbone area as area zero)
- Checksum: error detection
- Three Authentication related fields to authenticate OSPF packets

Classless Interdomain Routing (CIDR)

- Routing table are getting longer with the exponential growth of the Internet.
- CIDR uses Supernetting to summarize multiple routing entries into a smaller number of entries.
- CIDR is supported in almost all new routers.

CIDR – Type Address

- IP address in CIDR (Classless Inter-Domain Routing)
  - Not classified into classes
  - Two components of an IP address
    - Network prefix ranging from 13 to 27 bits – a Variable Length Subnet Mask (VLSM)
    - Host ID using the remaining bits → 19 to 5 bits
  - Slashed-notation

  A dotted-decimal IP address + "*/*" + *Number of bits used for the network prefix*
- Network address are assigned in a hierarchical manner.
- In the core network, routing entries for networks with the same higher level prefix, a CIDR block, can be summarized into one entry – i.e. supernetting for route aggregation
- The longest-prefix-matching rule is still used in table lookups.

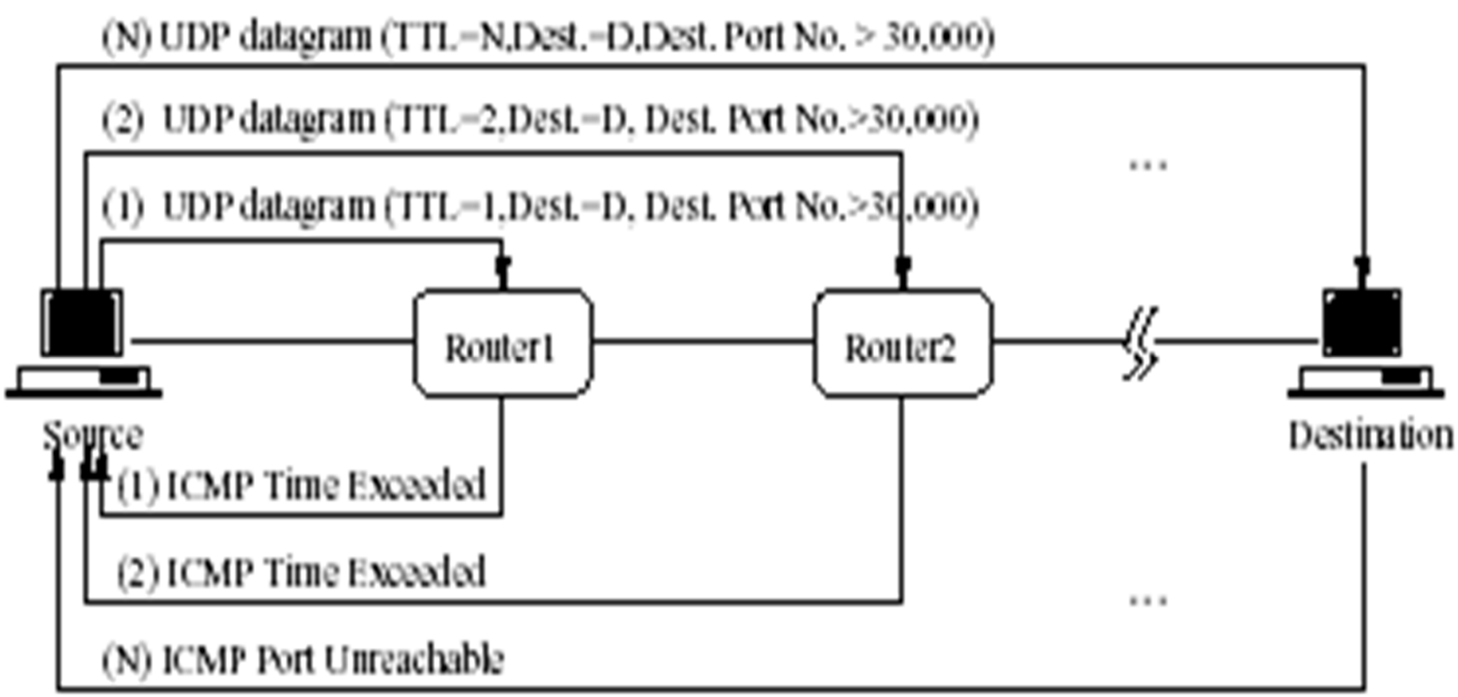
Private IP Address

- A Private Network is designed to be used mainly inside an organization
  - Intranet is a private network (LAN) that its access is limited to the users inside the organization
  - Extranet is also a private network (LAN) like the intranet but it allows some users outside the organization to access the network
- Blocks of IP addresses are assigned for private use
- Private IP addressed are not recognized globally
- Private IP addresses are used either in isolation or in connection with Network Address Translation (NAT) technique

Class	NetID	Block
A	10.0.0	1
B	172.16 to 172.31	16
C	192.168.0 to 192.168.255	256

Traceroute

- Help determine all the routers in an end-to-end path
- Use the Time-to-Live (TTL) field in the IP header and the ICMP protocol.
- Traceroute operation:



BGP- Border Gateway Protocol

- An interdomain routing protocol for routing between ASes
- Currently in version 4
- Note: In the context of BGP, a gateway is nothing else but an IP router that connects autonomous systems.
- Uses TCP port 179 to establish BGP session and exchange routing messages (active routes and incremental updates)
- BGP is based on distance vector protocol, but unlike in RIP, routing messages in BGP contain complete routes – Path Vector Routing
- Network administrators can specify routing policies

BGP Autonomous System Types

- BGP’s goal is to find any path (not an optimal one). Since the internals of each connected AS are never revealed, finding an optimal path is not feasible.
- For each AS, BGP distinguishes:
  - Local traffic: traffic carried within an AS that either originated in that same AS, or is intended to be delivered within that AS
  - Transit traffic: traffic that was generated outside that AS and is intended to be delivered outside the AS
- Three AS types:
  - Stub AS has connection to only one other AS, comparable to a cul-de-sac in our road analogy; only carry local traffic.
  - Multihomed AS has connection to two or more other ASes but does not carry transit traffic
  - Transit AS has connection to two or more other ASes and carries transit traffic

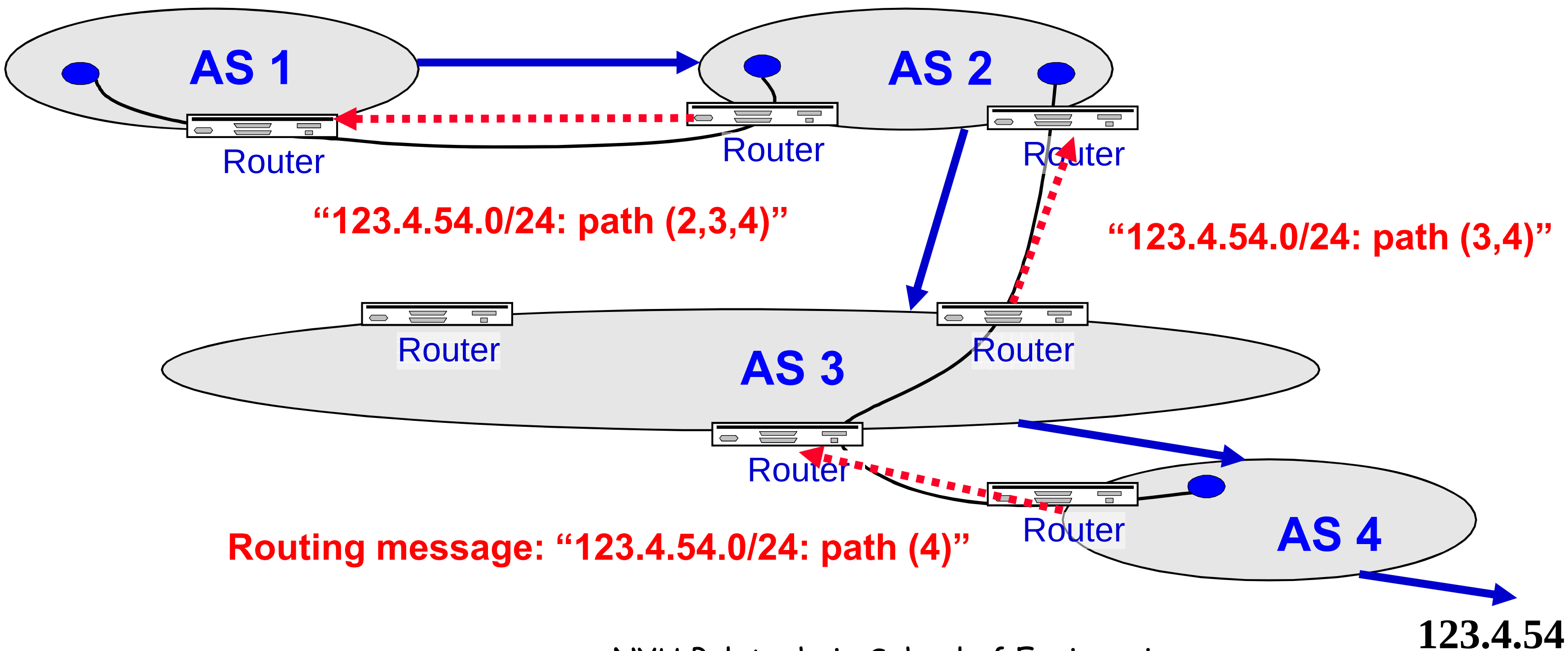
BGP Interdomain Routing

ASes exchange info about who they can reach

- IP prefix: block of destination IP addresses
- AS path: sequence of ASes along the path

Policies configured by each AS's operator

- Path selection: which of the paths to use?
- Path export: which neighbors to tell?



BGP Route Information Management
– Carried in Each BGP Gateway