COMPUTER NETWORKS AND INTERNET PROTOCOLS

IP Routing – Introduction [Routing Table]

SOUMYA K GHOSH
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

SANDIP CHAKRABORTY
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Discussion Points

- IP addresses and Address allocation
- Packet routing/ forwarding
  - Routing/ Forwarding tables
  - Longest-prefix match forwarding

Ref: Computer Networking: A Top-Down Approach; Computer Networks; Lecture materials of Jennifer Rexford, Princeton, USA; TCP/IP Tutorials and Technical Overview, IBM Redbooks

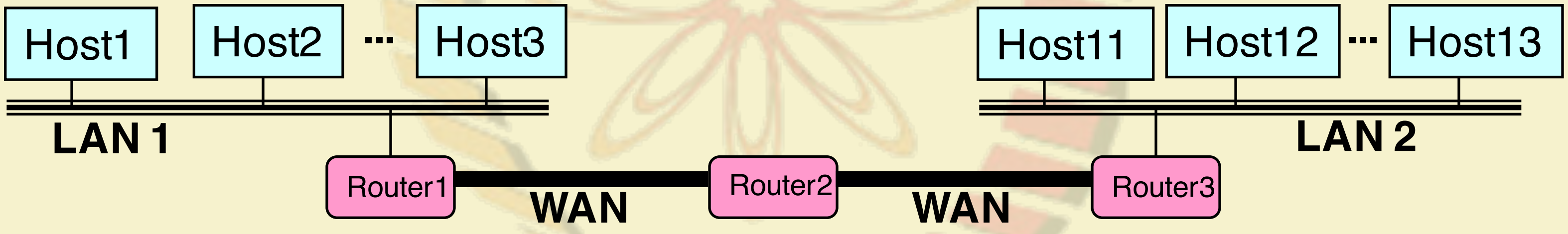
IP Address (IPv4)

- A unique 32-bit number
- Identifies an interface (on a host, on a router, …)
- Represented in dotted-quad notation

14	35	158	15
00001110	00100011	10011110	00001111

Inter-Networking

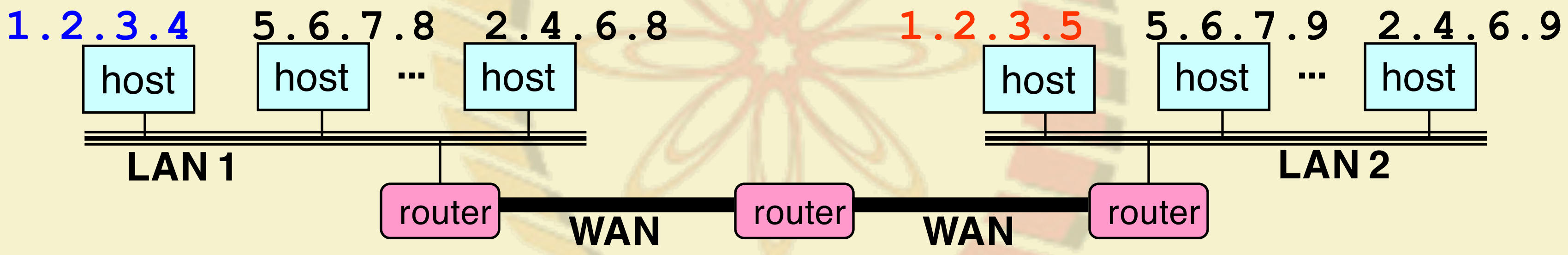
- The Internet is an "inter-network"
  - Used to connect *networks* together, not *hosts*
  - Needs a way to address a network (i.e., group of hosts)



LAN = Local Area Network
WAN = Wide Area Network

Scalability Issue

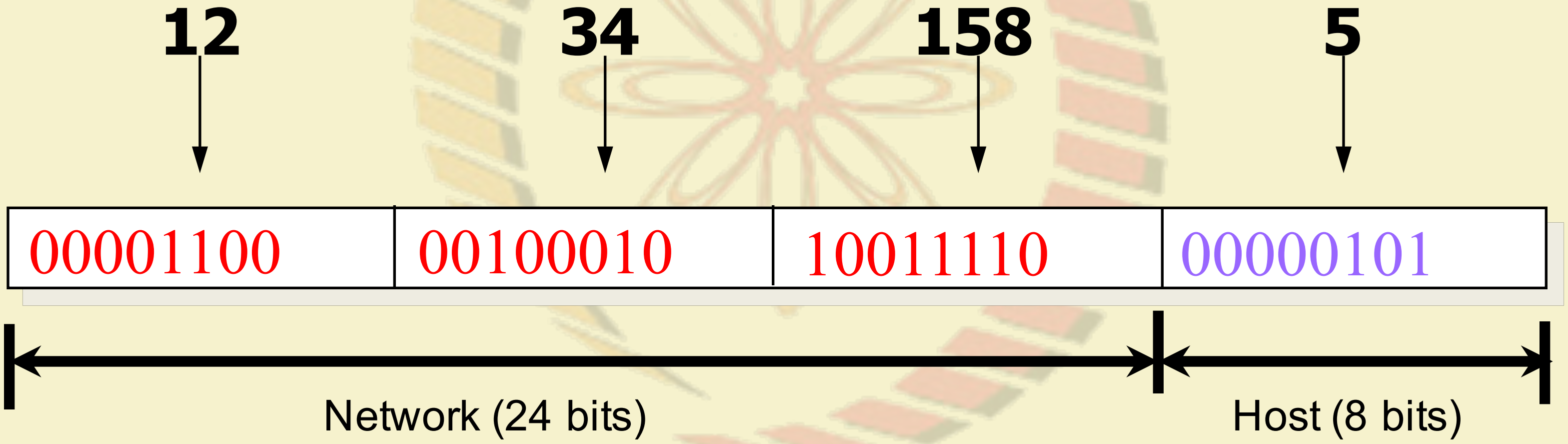
- Suppose hosts had arbitrary addresses, then
  - Every router would need a lot of information
  - To know how to direct packets toward the host



Hierarchical Addressing: IP Prefixes

- Divided into network and host portions (left and right)
- 12.34.158.0/24 is a 24-bit prefix with 2^8 addresses

Analogy with our Postal mail system ?

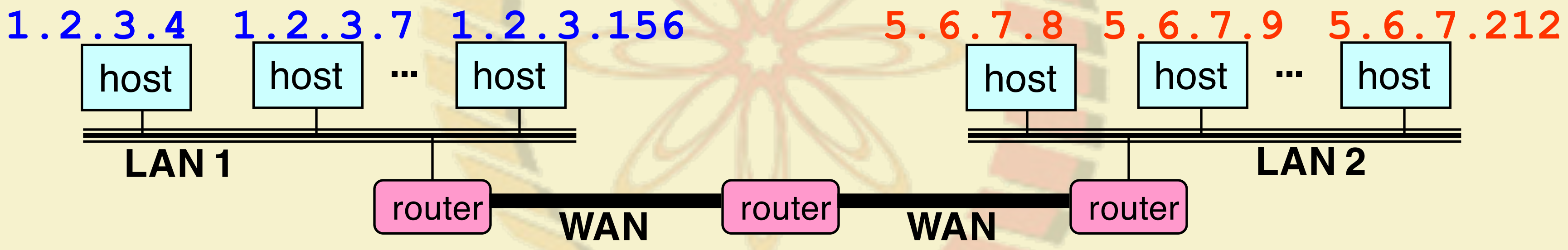


IP Address and a 24-bit Subnet Mask

IP Address	12	34	158	5
	00001100	00100010	10011110	00000101
	11111111	11111111	11111111	00000000
Mask	255	255	255	0

Scalability Improved

- Number related hosts from a common subnet
  - 1.2.3.0/24 on the left LAN
  - 5.6.7.0/24 on the right LAN

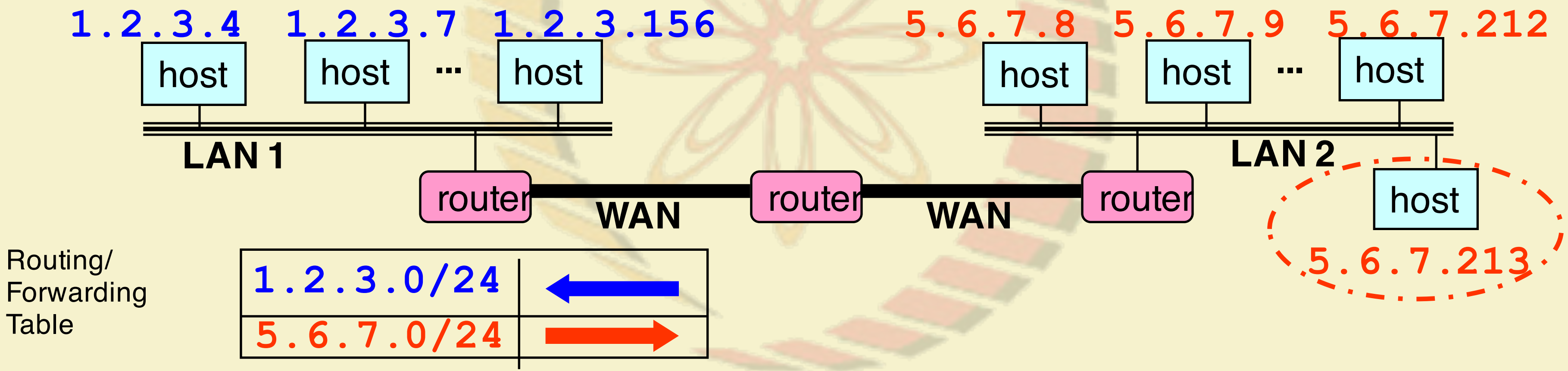


Routing/
Forwarding
Table

1.2.3.0/24	←
5.6.7.0/24	→

Easy to Add New Hosts

- No need to update the routers
  - E.g., adding a new host 5.6.7.213 on the right
  - Doesn't require adding a new forwarding-table entry



IP Address Allocation

Classful Addressing

- Fixed allocation sizes
  - Class A: 0*
    - Very large /8 blocks (e.g., Org. A has 20.0.0.0/8)
  - Class B: 10*
    - Large /16 blocks (e.g,. Org. B has 128.40.0.0/16)
  - Class C: 110*
    - Small /24 blocks (e.g., Org. C has 192.60.225.0/24)
  - Class D: 1110*
    - Multicast groups
  - Class E: 11110*
    - Reserved for future use

Classless Inter-Domain Routing (CIDR)

Use two 32-bit numbers to represent a network.
Network number = IP address + Mask

IP Address : 12.4.0.0   IP Mask: 255.254.0.0

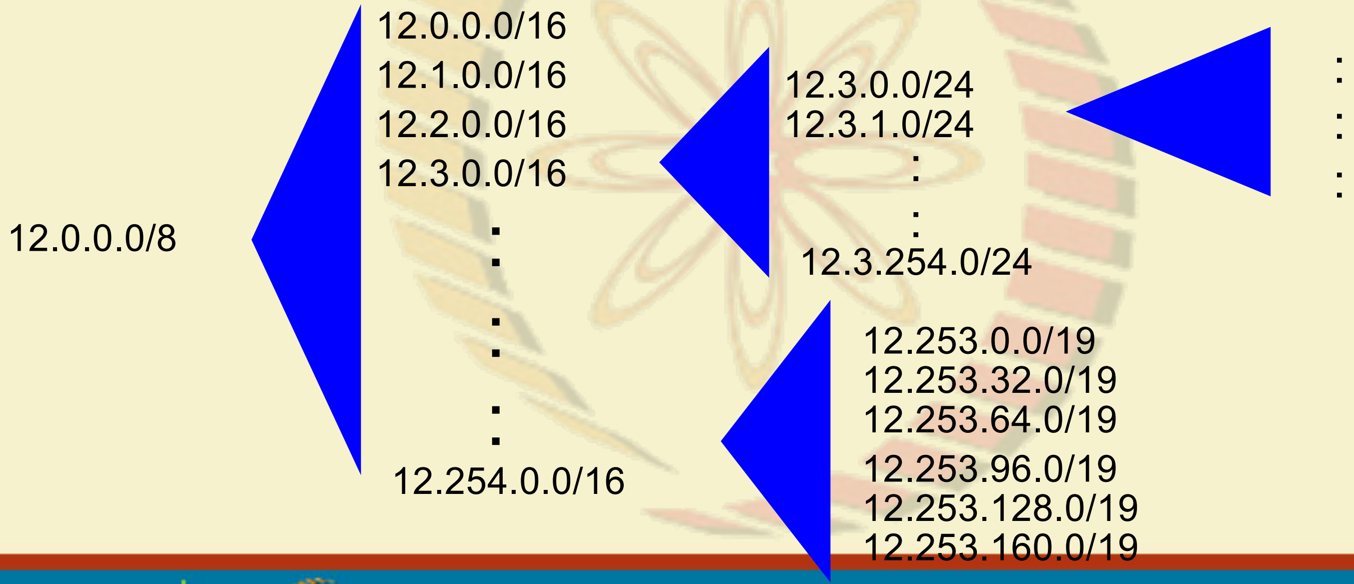
Address	00001100	00000100	00000000	00000000
Mask	11111111	11111110	00000000	00000000

|← Network Prefix →|← for hosts →|

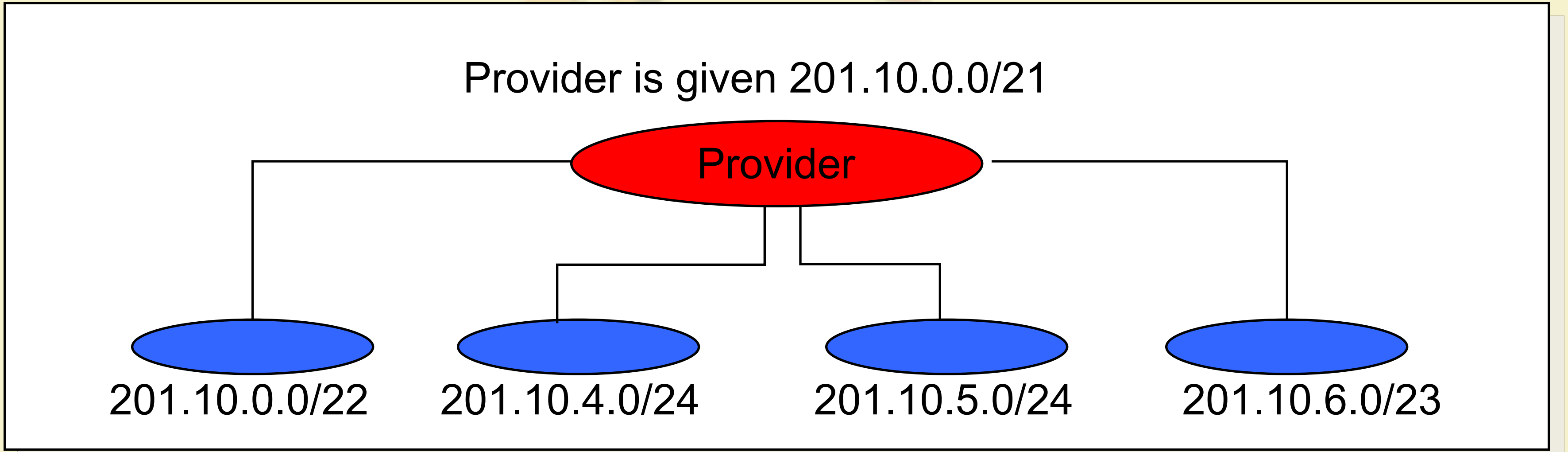
Represented as: 12.4.0.0/15

CIDR: Hierarchal Address Allocation

- Prefixes are key to Internet scalability
  - Address allocated in contiguous chunks (prefixes)
  - Routing protocols and packet forwarding based on prefixes

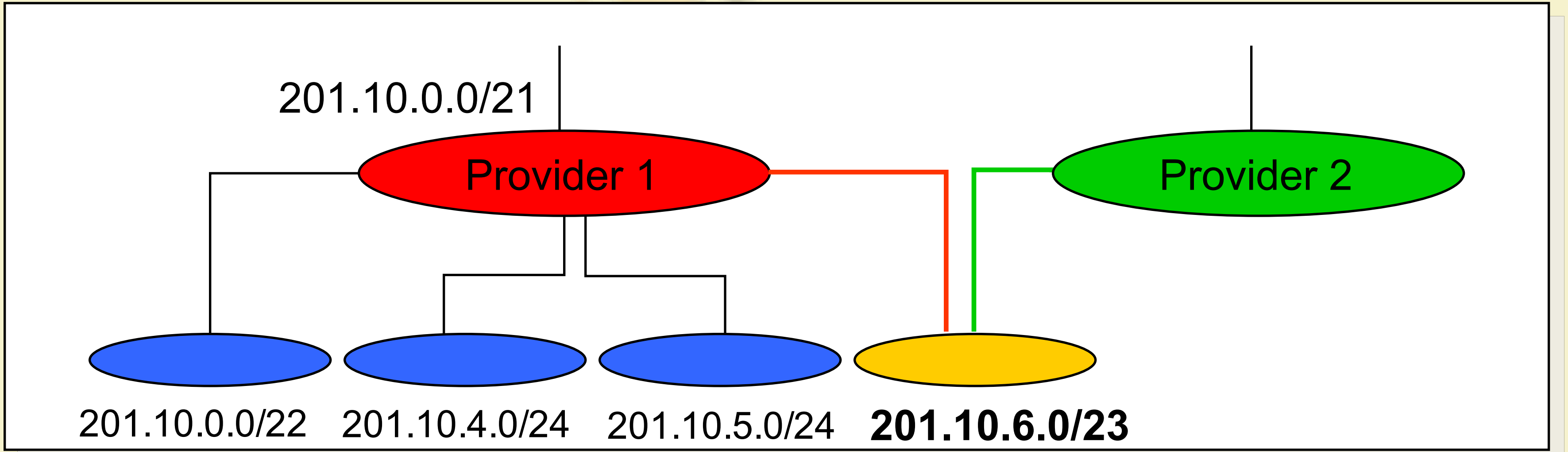


Scalability: Address Aggregation



Routers in the rest of the Internet just need to know how to reach 201.10.0.0/21. The provider can direct the IP packets to the appropriate customer.

Challenges in Aggregation



Multi-homed customer with 201.10.6.0/23 has two providers. Other parts of the Internet need to know how to reach these destinations through *both* providers.

Scalability Through Hierarchy

- Hierarchical addressing
  - Critical for scalable system
  - Don’t require everyone to know everyone else
  - Reduces amount of updating when something changes
- Non-uniform hierarchy
  - Useful for heterogeneous networks of different sizes
  - Initial class-based addressing was far too coarse
  - Classless Inter-Domain Routing (CIDR) helps

Obtaining a Block of Addresses

- Separation of control
  - Prefix: assigned *to* an institution
  - Addresses: assigned *by* the institution to their nodes
- Who assigns prefixes?
  - Internet Corporation for Assigned Names and Numbers (ICANN)
    - Allocates large address blocks to Regional Internet Registries
  - Regional Internet Registries (RIRs)
    - E.g., ARIN (American Registry for Internet Numbers)
    - Allocates address blocks within their regions
    - Allocated to Internet Service Providers and large institutions
  - Internet Service Providers (ISPs)
    - Allocate address blocks to their customers
    - Who may, in turn, allocate to their customers…

Ownership of an Address

- Address registries
  - Public record of address allocations
  - Internet Service Providers (ISPs) should update when giving addresses to customers
  - However, records are notoriously out-of-date
- Ways to query
  - UNIX: “whois –h whois.arin.net 128.112.136.35”
  - http://www.arin.net/whois/
  - http://whois.inregistry.in
  - …

Are 32-bit Addresses Enough?

- Not all that many unique addresses
  - 2^{32} = 4,294,967,296 (just over four billion)
  - Plus, some are reserved for special purposes
  - And, addresses are allocated in larger blocks
- Many devices need IP addresses
  - Computers, PDAs, routers, tanks, toasters, …
- Long-term solution: a larger address space
  - IPv6 has 128-bit addresses ($2^{128} = 3.403 \times 10^{38}$)
- Short-term solutions: with IPv4 address space
  - Private addresses
  - Network address translation (NAT)
  - Dynamically-assigned addresses (DHCP)

Hard Policy Questions

- How much address space per geographic region?
  - Equal amount per country?
  - Proportional to the population?
  - What about addresses already allocated?
- Address space portability?
  - Keep your address block when you change providers?
  - Pro: avoid having to renumber your equipment
  - Con: reduces the effectiveness of address aggregation
- Keeping the address registries up to date?
  - What about mergers and acquisitions?
  - Delegation of address blocks to customers?
  - As a result, the registries are horribly out of date

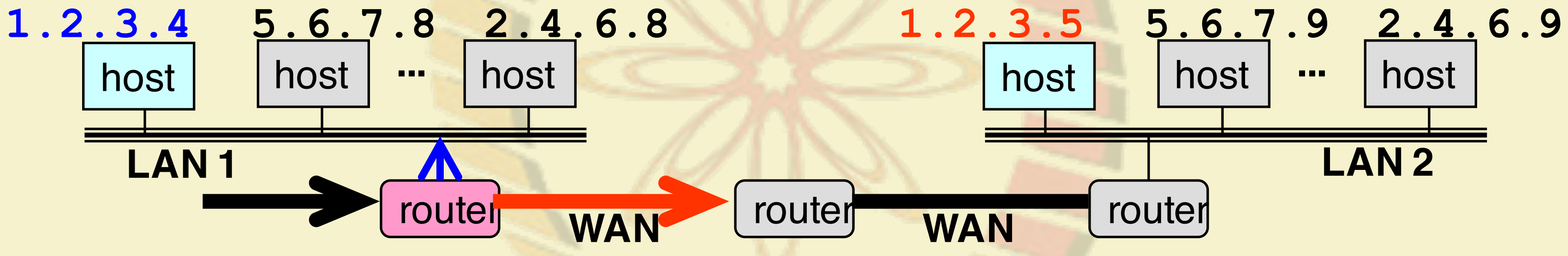
Packet Forwarding

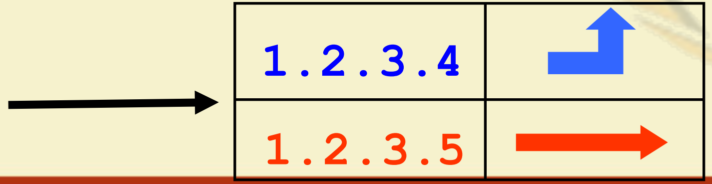
Hop-by-Hop Packet Forwarding

- Each router has a forwarding table
  - Maps destination addresses…
  - … to outgoing interfaces
- Upon receiving a packet
  - Inspect the destination IP address in the header
  - Index into the table
  - Determine the outgoing interface
  - Forward the packet out that interface
- Then, the next router in the path repeats
  - And the packet travels along the path to the destination

Separate Table Entries Per Address

- If a router had a forwarding entry per IP address
  - Match *destination address* of incoming packet
  - … to the *forwarding-table entry*
  - … to determine the *outgoing interface*

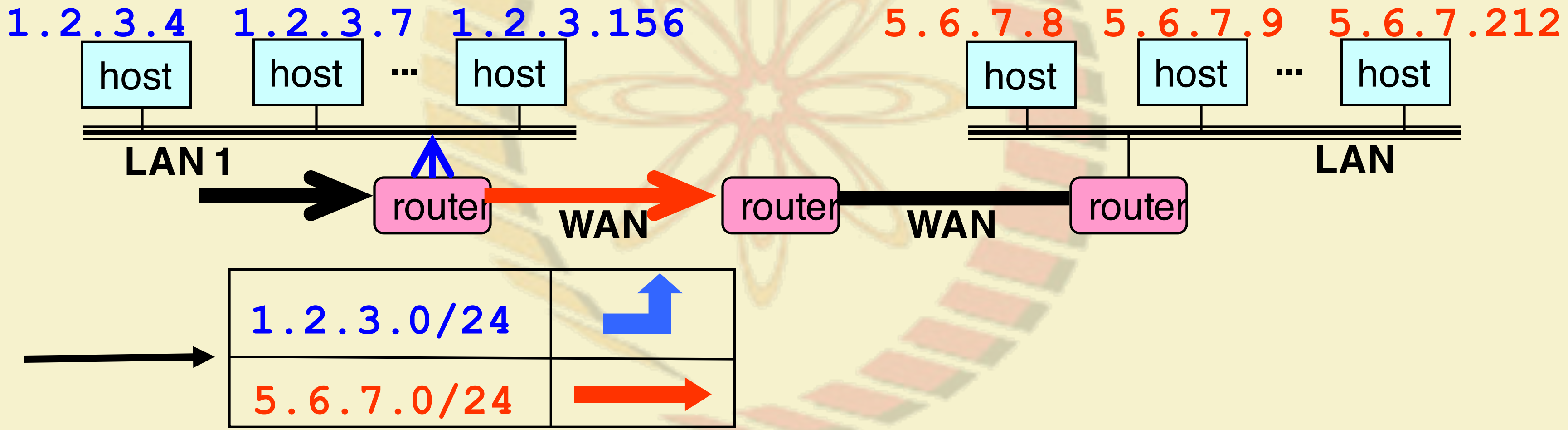




Routing/ Forwarding table

Separate Entry Per 24-bit Prefix

- If the router had an entry per 24-bit prefix
  - Look only at the top 24 bits of the destination address
  - Index into the table to determine the next-hop interface

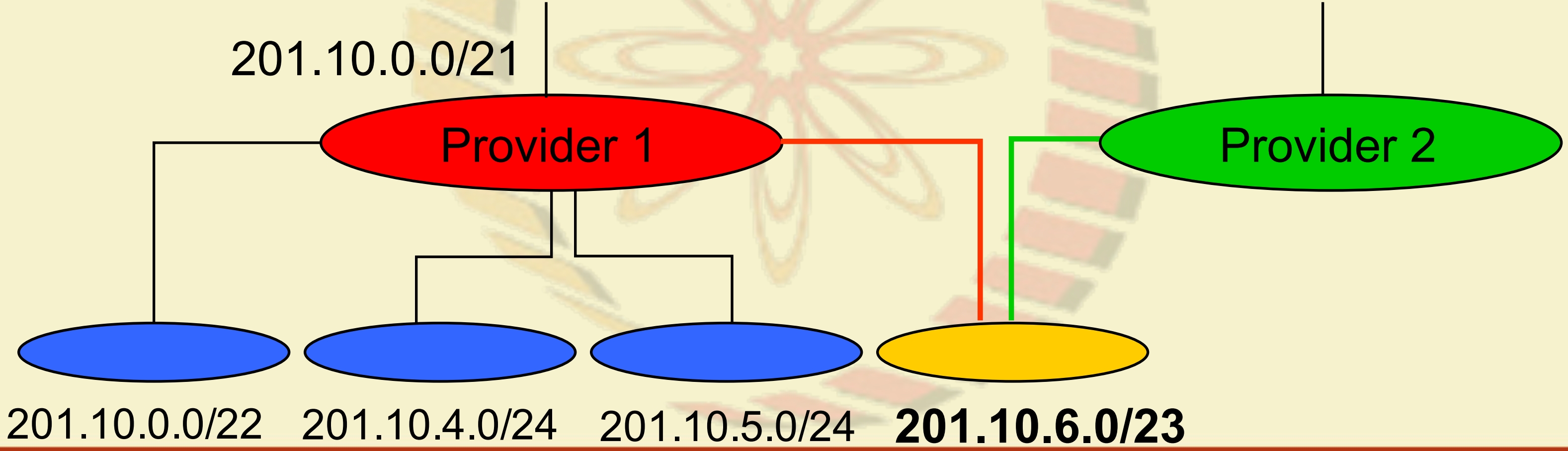


Separate Entry Classful Address

- If the router had an entry per classful prefix
  - Mixture of Class A, B, and C addresses
  - Depends on the first couple of bits of the destination
- Identify the mask automatically from the address
  - First bit of 0: class A address (/8)
  - First two bits of 10: class B address (/16)
  - First three bits of 110: class C address (/24)
- Then, look in the forwarding table for the match
  - E.g., 1.2.3.4 maps to 1.2.3.0/24
  - Then, look up the entry for 1.2.3.0/24
  - … to identify the outgoing interface

CIDR complicates Packet Forwarding

- CIDR – packet forwarding
  - CIDR allows efficient use of the limited address space
  - But, CIDR makes packet forwarding much harder
- Forwarding table may have many matches
  - E.g., table entries for 201.10.0.0/21 and 201.10.6.0/23
  - The IP address 201.10.6.17 would match *both*!



Longest Prefix Match (LPM) Forwarding

- Forwarding tables in IP routers
  - Maps each IP prefix to next-hop link(s)
- Destination-based forwarding
  - Packet has a destination address
  - Router identifies longest-matching prefix
  - Cute algorithmic problem: very fast lookups

Routing/ forwarding table

destination

201.10.6.17

4.0.0.0/8
4.83.128.0/17
201.10.0.0/21
201.10.6.0/23
126.255.103.0/24

outgoing link

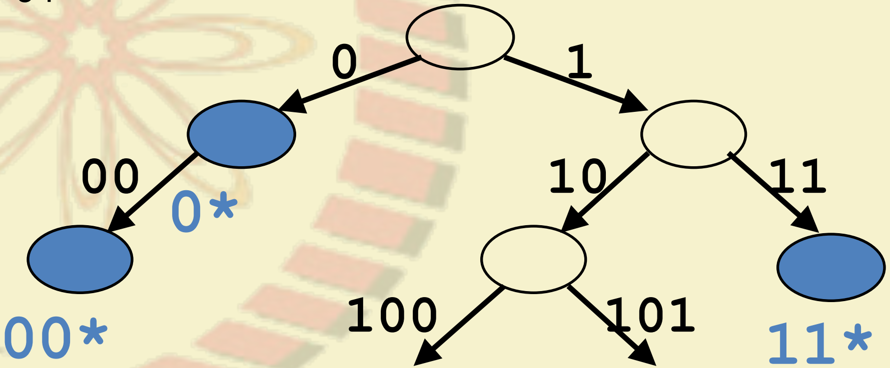
Serial0/0.1

LPM Algorithm

- Scan the forwarding table one entry at a time
  - See if the destination matches the entry
  - If so, check the size of the mask for the prefix
  - Keep track of the entry with longest-matching prefix
- Overhead is linear in size of the forwarding table
  - Table entries are quite large, 200000+
  - The router may have just a few nanoseconds before the next packet is arriving
- Need greater efficiency to keep up with *line rate*
  - Better algorithms
  - Hardware implementations

Patricia Tree

- Store the prefixes as a tree
  - One bit for each level of the tree
  - Some nodes correspond to valid prefixes which have next-hop interfaces in a table
- When a packet arrives
  - Traverse the tree based on the destination address
  - Stop upon reaching the longest matching prefix



Faster Lookups

- Patricia tree is faster than linear scan
  - Proportional to number of bits in the address
- Patricia tree can be made faster
  - Can make a k-ary tree
    - E.g., 4-ary tree with four children (00, 01, 10, and 11)
  - Faster lookup, though requires more space
- Can use special hardware
  - Content Addressable Memories (CAMs)
  - Allows look-ups on a key rather than flat address

Where do Forwarding Tables Come From?

- Routers have forwarding tables
  - Map prefix to outgoing link(s)
- Entries can be statically configured
  - E.g., “map 12.34.158.0/24 to Serial0/0.1”
- But, it is not adadtive
  - to failures
  - to new devices
  - to load balancing
- That is where other technologies come in…
  - Routing protocols, DHCP, and ARP (later in course)

Packet Forwarding by End Devices/ Hosts

- End host with single network interface
  - PC with an Ethernet link
  - Laptop with a wireless link
- Don't need to run a routing protocol
  - Packets to the host itself (e.g., 1.2.3.4/32)
    - Delivered locally
  - Packets to other hosts on the LAN (e.g., 1.2.3.0/24)
    - Sent out the interface
  - Packets to external hosts (e.g., 0.0.0.0/0)
    - Sent out interface to local gateway
- How this information is learned
  - Static setting of address, subnet mask, and gateway
  - Dynamic Host Configuration Protocol (DHCP)

Packet Delivery to the End Hosts?

- How does the last router reach the destination?
- Each interface has a persistent, global identifier
  - MAC (Media Access Control) address
  - Burned in to the adaptors Read-Only Memory (ROM)
  - Flat address structure (i.e., no hierarchy)
- Constructing an address resolution table
  - Mapping MAC address to/from IP address
  - Address Resolution Protocol (ARP)

Summary

- IP address
  - A 32-bit number
  - Allocated in prefixes
  - Non-uniform hierarchy for scalability and flexibility
- Packet forwarding
  - Based on IP prefixes
  - Longest-prefix-match forwarding

MIME Headers

- Located between the Email Header and Body
  - MIME-Version: 1.1
  - Content-Type: type/subtype
  - Content-Transfer-Encoding: encoding type
  - Content-Id: message id
  - Content-Description: textual explanation of non-textual contents

thank you!

COMPUTER NETWORKS AND INTERNET PROTOCOLS

IP Routing - I [Intra-domain routing]

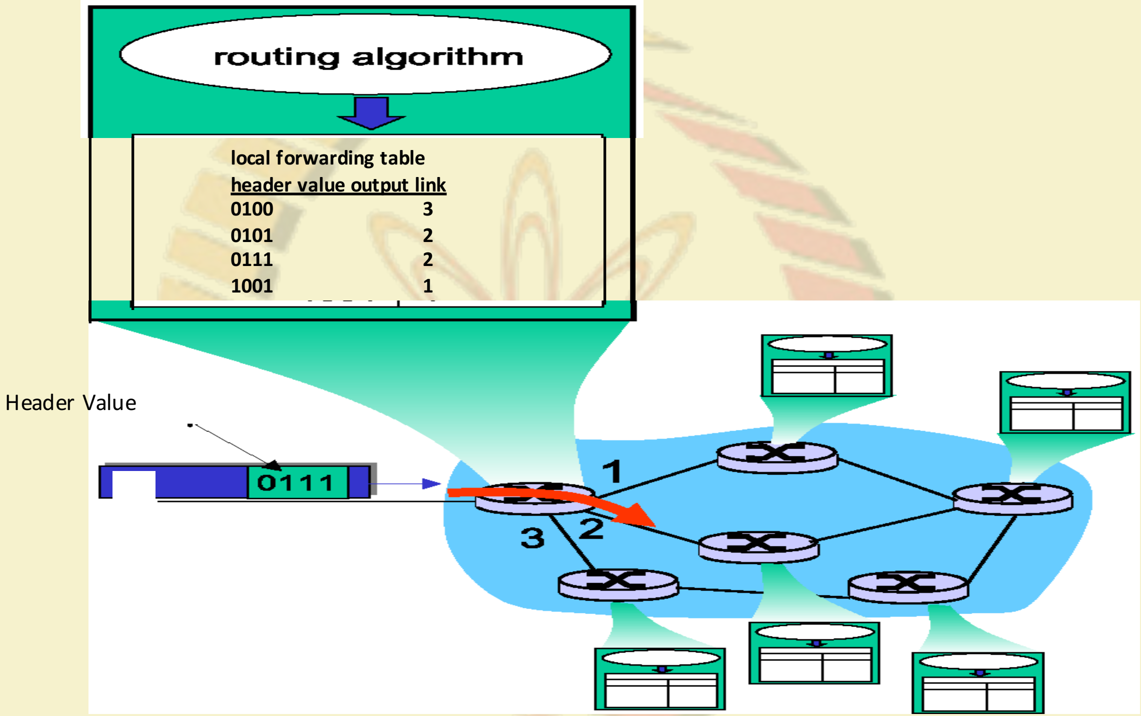
SOUMYA K GHOSH
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

SANDIP CHAKRABORTY
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Routing ?

- Forwarding – moving packets between ports
  - Look up destination address in forwarding table
  - Find out-port or <out-port, MAC Addr> pair
- Routing is process of populating forwarding table
  - Routers exchange messages about nets they can reach
  - Goal: Find optimal route for every destination

Routing...


routing algorithm
local forwarding table
header value output link
0100 3
0101 2
0111 2
1001 1
Header Value
0111
1
2
3

Stability

- Stable routes are often preferred over rapidly changing ones
- Reason 1: management
  - Hard to debug a problem if it's transient
- Reason 2: higher layer optimizations
  - E.g., TCP RTT estimation
  - Imagine alternating over 500ms and 50ms routes
- Choosing between optimality and stability!

Routing algorithms

- Global (centralized) vs. Decentralized
- Global: All routers have complete topology
- Decentralized: Only know neighbors and share information from them
- Intra-domain vs. Inter-domain routing
  - Intra : All routers under same administrative control
  - Inter : Decentralized, scale to Internet

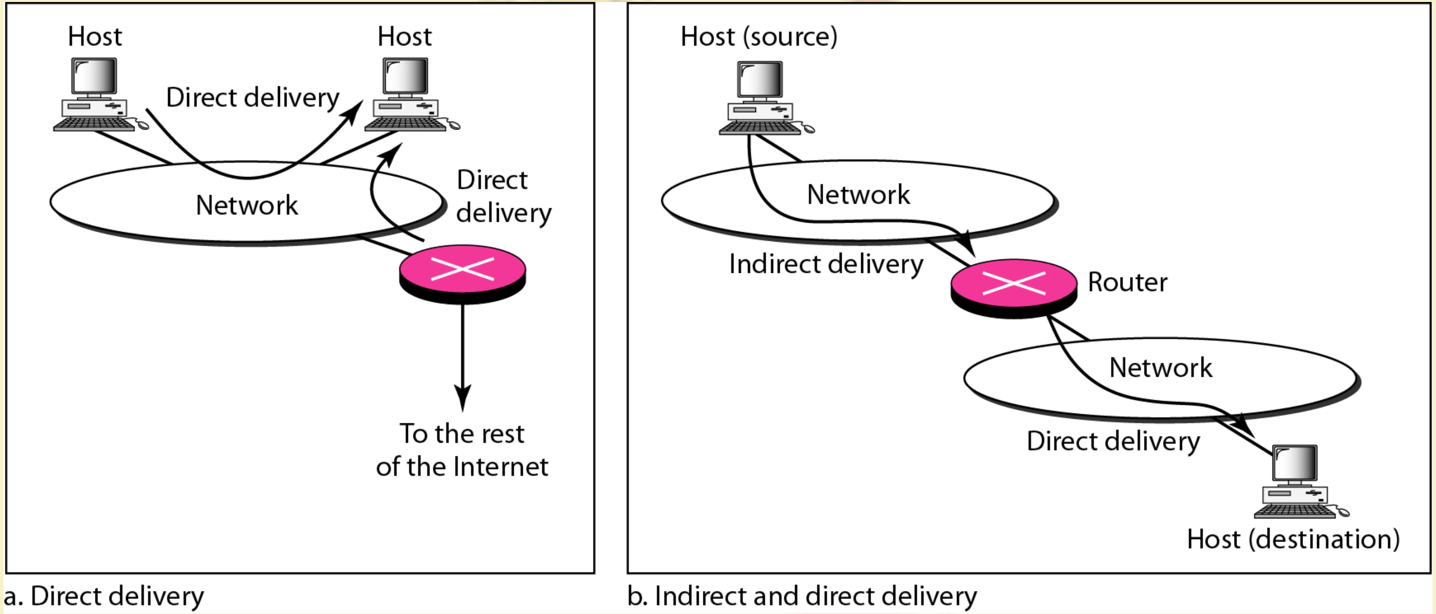
Optimality

- View network as a graph
- Assign cost to each edge
  - Can be based on latency, b/w, utilization, queue length, . . .
- Problem: Find lowest cost path between two nodes
  - Each node individually computes the cost

Scaling issues

- Every router must be able to forward based on any destination IP address
  - Given address, it needs to know ”next hop” (table)
  - Naıve: Have an entry for each address
  - There may be huge number of entries!
- Solution? - Entry covers range of addresses
  - Can’t do this if addresses are assigned randomly! (e.g., Ethernet addresses)
  - Address aggregation is important
  - Addresses allocation should be based on network structure

Routing/Forwarding Packets in the Internet



Ref: Data Communication and Networking by B.A. Forouzan

Route method vs. Next-hop method

a. Routing tables based on route

b. Routing tables based on next hop

Routing table for host A

Destination	Route
Host B	R1, R2, host B

Destination	Next hop
Host B	R1

Routing table for R1

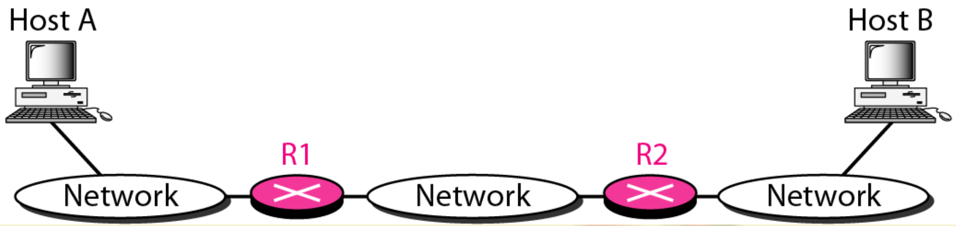
Destination	Route
Host B	R2, host B

Destination	Next hop
Host B	R2

Routing table for R2

Destination	Route
Host B	Host B

Destination	Next hop
Host B	---



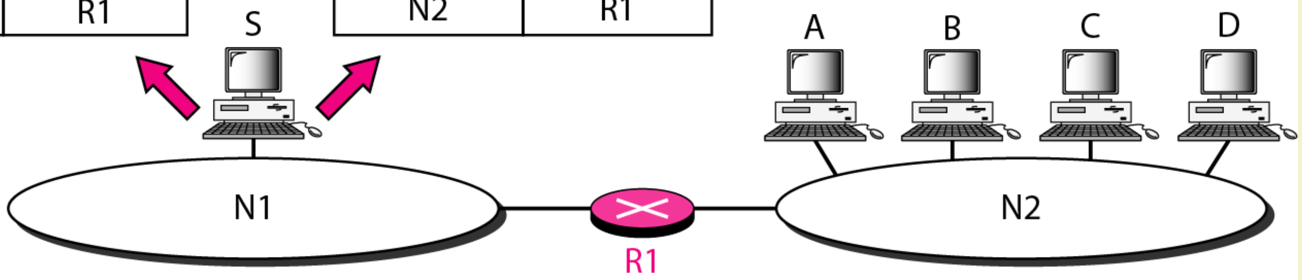
Host-specific vs. Network-specific method

Routing table for host S based on host-specific method

Destination	Next hop
A	R1
B	R1
C	R1
D	R1

Routing table for host S based on network-specific method

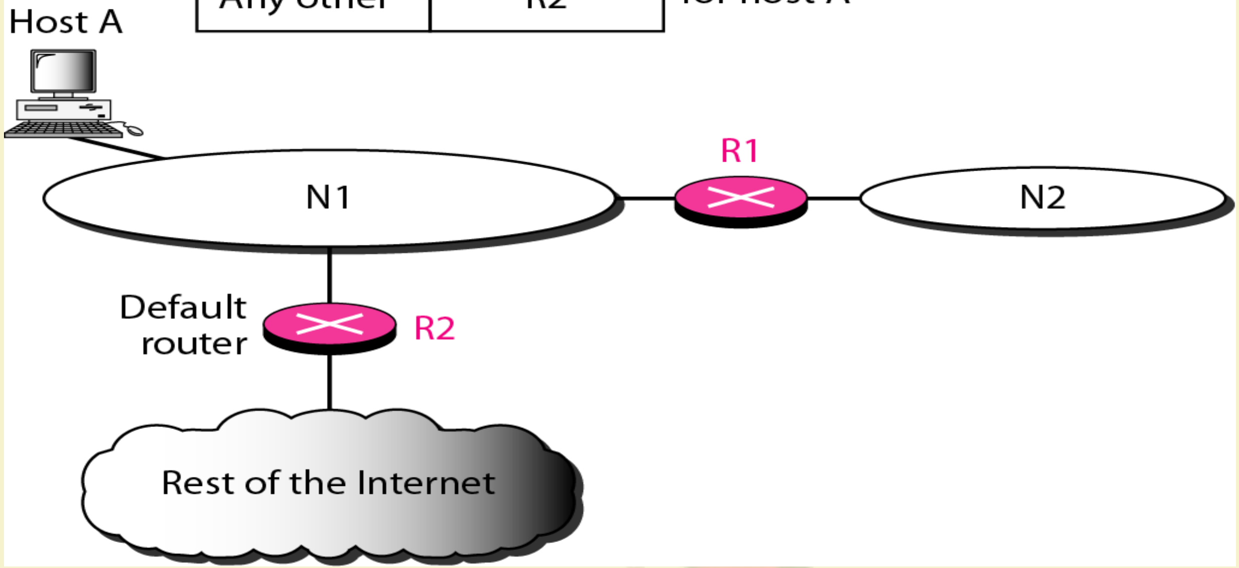
Destination	Next hop
N2	R1



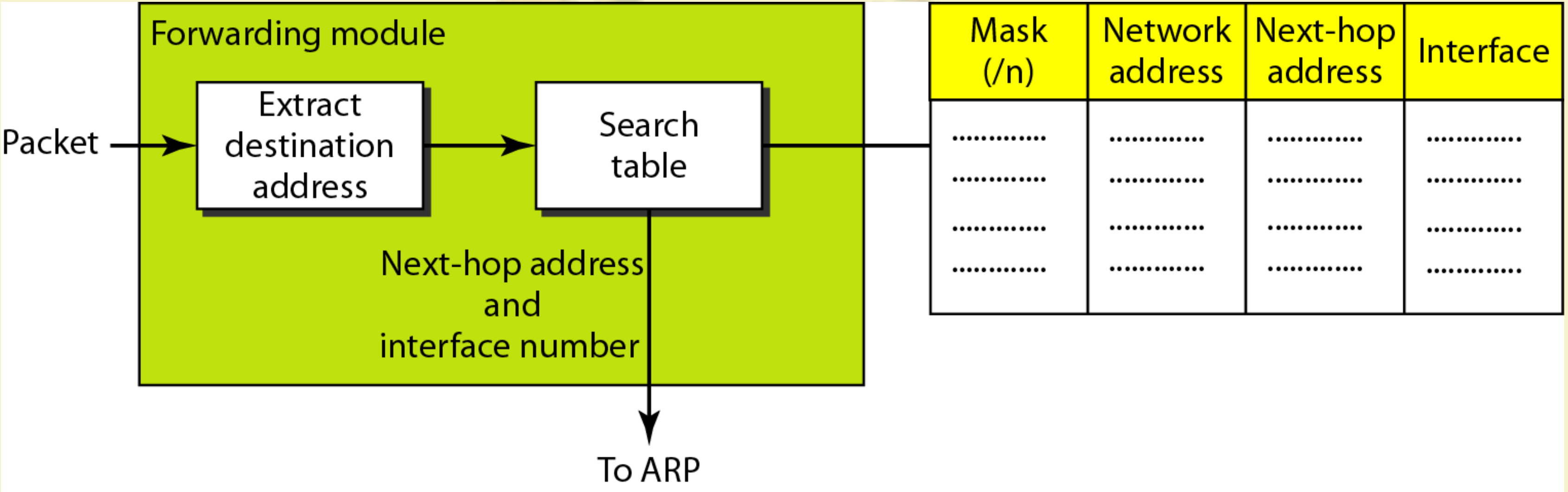
Default Route

Destination	Next hop
N2	R1
Any other	R2

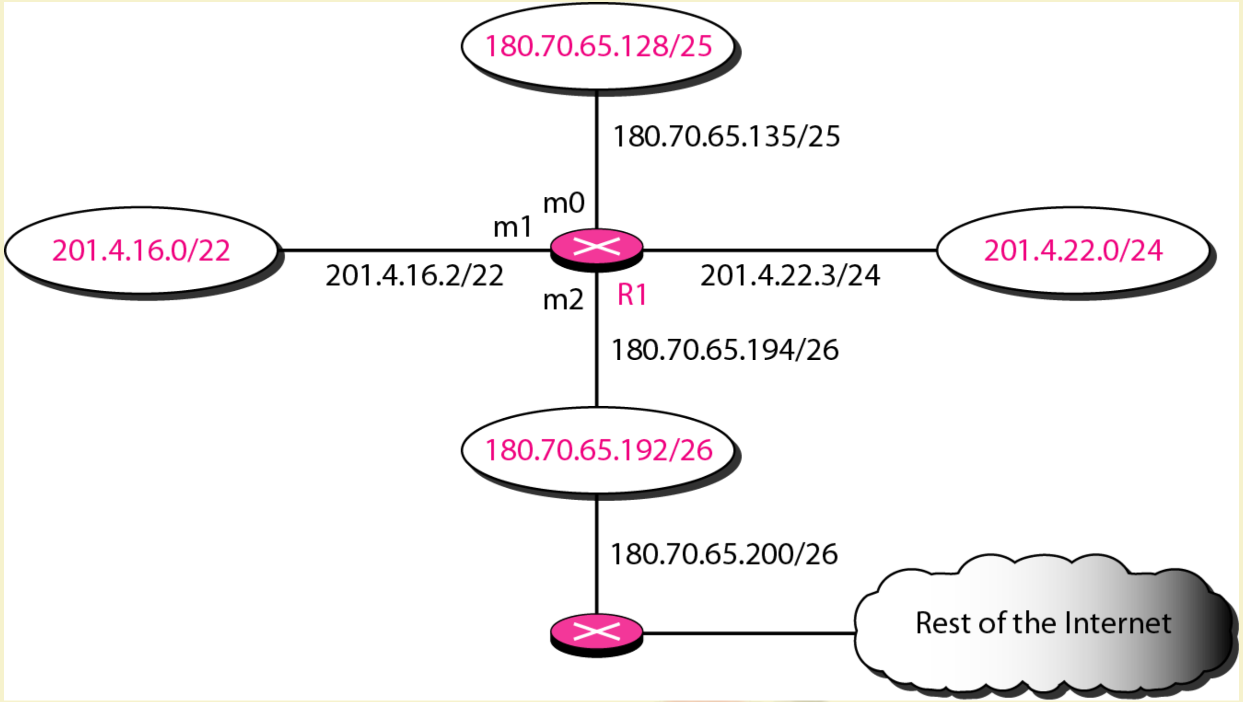
Routing table for host A



Basic Routing module in classless address



Example Routing Configuration



Routing table for router R1

Mask	*Network Address*	*Next Hop*	*Interface*
/26	180.70.65.192	—	m2
/25	180.70.65.128	—	m0
/24	201.4.22.0	—	m3
/22	201.4.16.0	....	m1
Any	Any	180.70.65.200	m2

Example: forwarding process if a packet arrives at R1 in with the destination address 180.70.65.140.

Router performs the following steps:

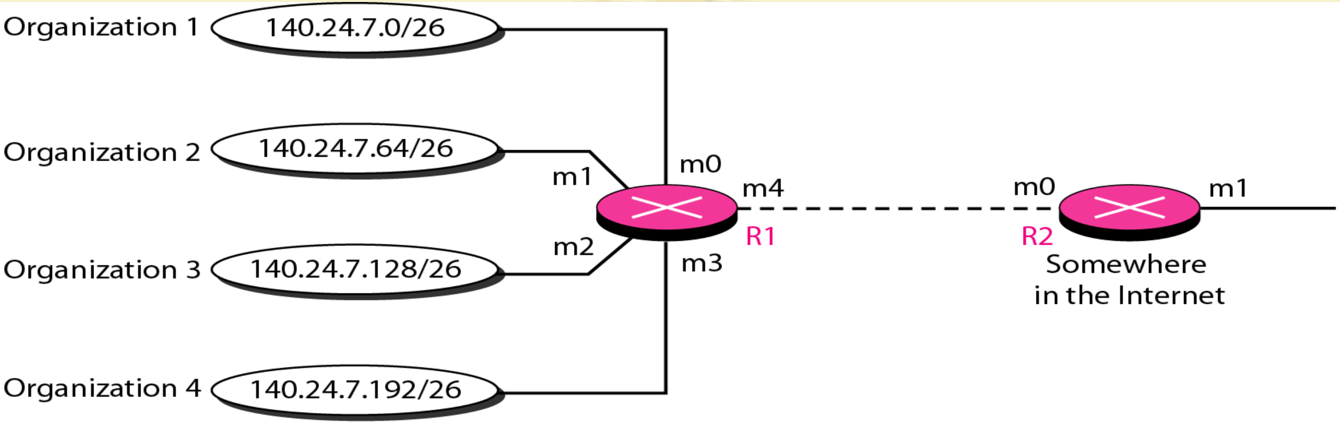
1. The first mask (/26) is applied to the destination address. The result is 180.70.65.128, which does not match the corresponding network address.
2. The second mask (/25) is applied to the destination address. The result is 180.70.65.128, which matches the corresponding network address. The next-hop address and the interface number m0 are passed to ARP for further processing.

Example: Show the forwarding process if a packet arrives at R1 with the destination address 201.4.22.35.

The router performs the following steps:

1. The first mask (/26) is applied to the destination address. The result is 201.4.22.0, which does not match the corresponding network address.
2. The second mask (/25) is applied to the destination address. The result is 201.4.22.0, which does not match the corresponding network address (row 2).
3. The third mask (/24) is applied to the destination address. The result is 201.4.22.0, which matches the corresponding network address. The destination address of the packet and the interface number m3 are passed to ARP.

Address aggregation



Mask	Network address	Next-hop address	Interface
/26	140.24.7.0	----------	m0
/26	140.24.7.64	----------	m1
/26	140.24.7.128	----------	m2
/26	140.24.7.192	----------	m3
/0	0.0.0.0	Default	m4

Routing table for R1

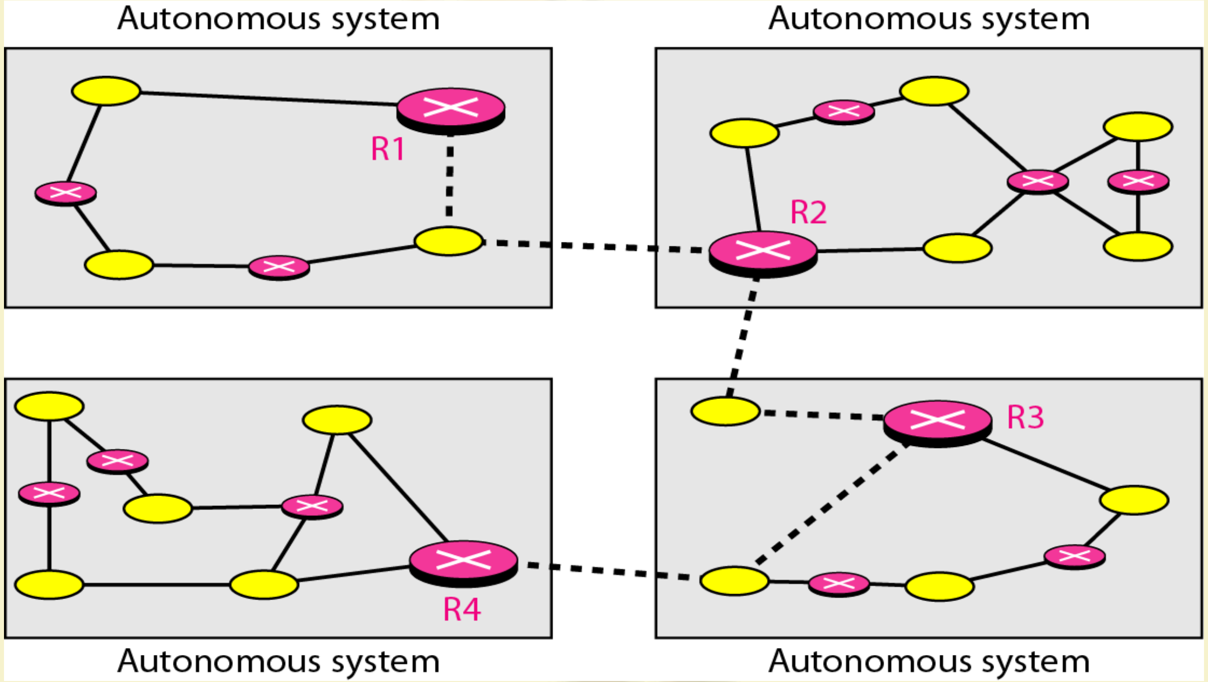
Mask	Network address	Next-hop address	Interface
/24	140.24.7.0	----------	m0
/0	0.0.0.0	Default	m1

Routing table for R2

Autonomous Systems (AS)

- Correspond to an administrative domain
  - Internet is not a single network
  - ASes reflect organization of the Internet
  - e.g., Stanford, large company, etc.
- Goals:
  - ASes want to choose their own local routing algorithm
  - ASes want to set policies about non-local routing
  - Each AS assigned unique 16-bit number

Autonomous systems



AS Traffic

- Local traffic – packets with src or dst in local AS
- Transit traffic – passes through an AS
- Stub AS -Connects to only a single other AS
- Multihomed AS
  - Connects to multiple ASes
  - Carries no transit traffic
- Transit AS - Connects to multiple ASes and carries transit traffic

Intra-domain Routing

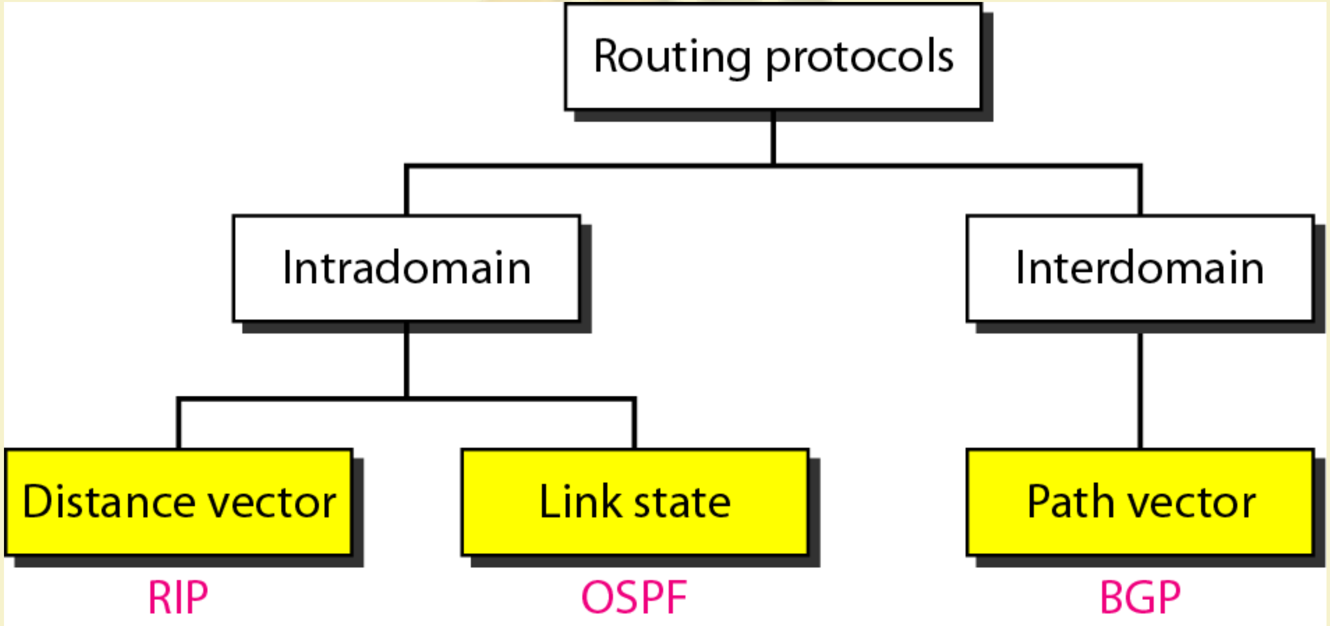
- Intra-domain routing: within an AS
- Single administrative control: *optimality* is important
  - Contrast with inter-AS routing, where policy dominates
  - Next lecture will cover inter-domain routing (BGP)

Intra-domain Routing Algorithms

- Distance vector
  - Require only local state (less overhead, smaller footprint)
  - Harder to debug
  - Can suffer from loops

- Link state
  - Have a global view of the network
  - Simpler to debug
  - Require global state

Popular routing protocols



Distance vector routing

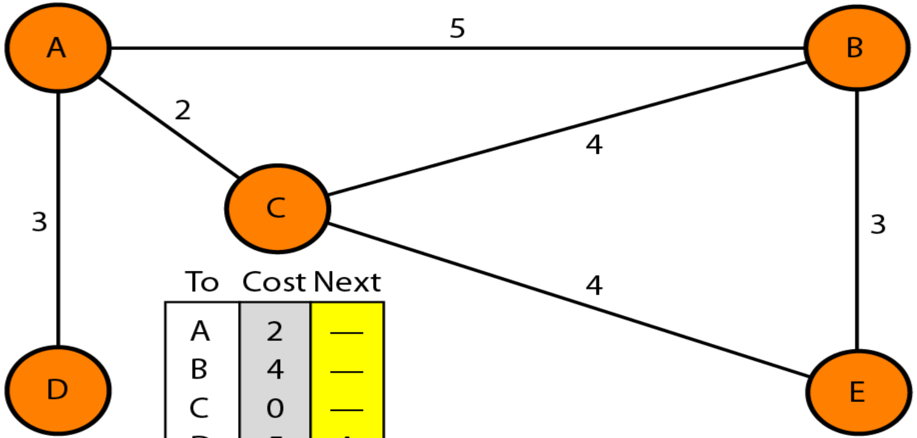
In distance vector routing, each node shares its routing table with its immediate neighbors periodically and when there is a change.

To	Cost	Next
A	0	—
B	5	—
C	2	—
D	3	—
E	6	C

A's table

To	Cost	Next
A	5	—
B	0	—
C	4	—
D	8	A
E	3	—

B's table



To	Cost	Next
A	3	—
B	8	A
C	5	A
D	0	—
E	9	A

D's table

To	Cost	Next
A	2	—
B	4	—
C	0	—
D	5	A
E	4	—

C's table

To	Cost	Next
A	6	C
B	3	—
C	4	—
D	9	C
E	0	—

E's table

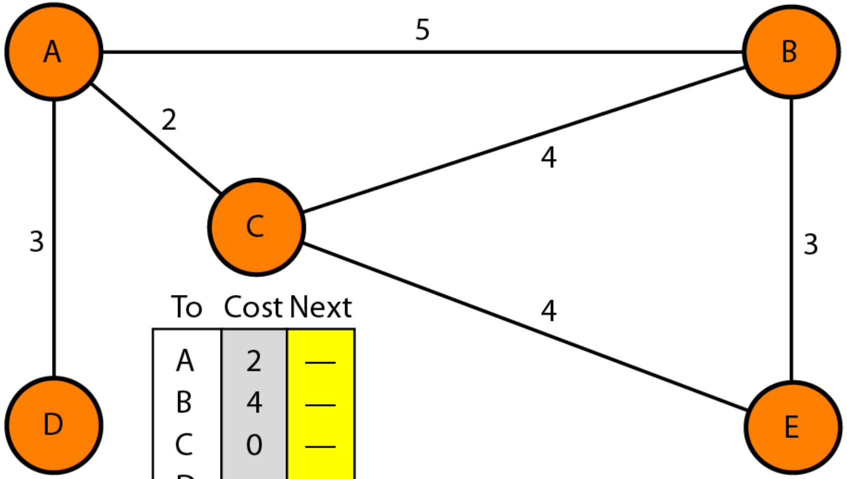
Distance Vector routing: Initialization of tables

To	Cost	Next
A	0	—
B	5	—
C	2	—
D	3	—
E	∞	

A's table

To	Cost	Next
A	5	—
B	0	—
C	4	—
D	∞	
E	3	—

B's table



To	Cost	Next
A	3	—
B	∞	
C	∞	
D	0	—
E	∞	

D's table

To	Cost	Next
A	2	—
B	4	—
C	0	—
D	∞	
E	4	—

C's table

To	Cost	Next
A	∞	
B	3	B
C	4	C
D	∞	
E	0	D

E's table

thank you!

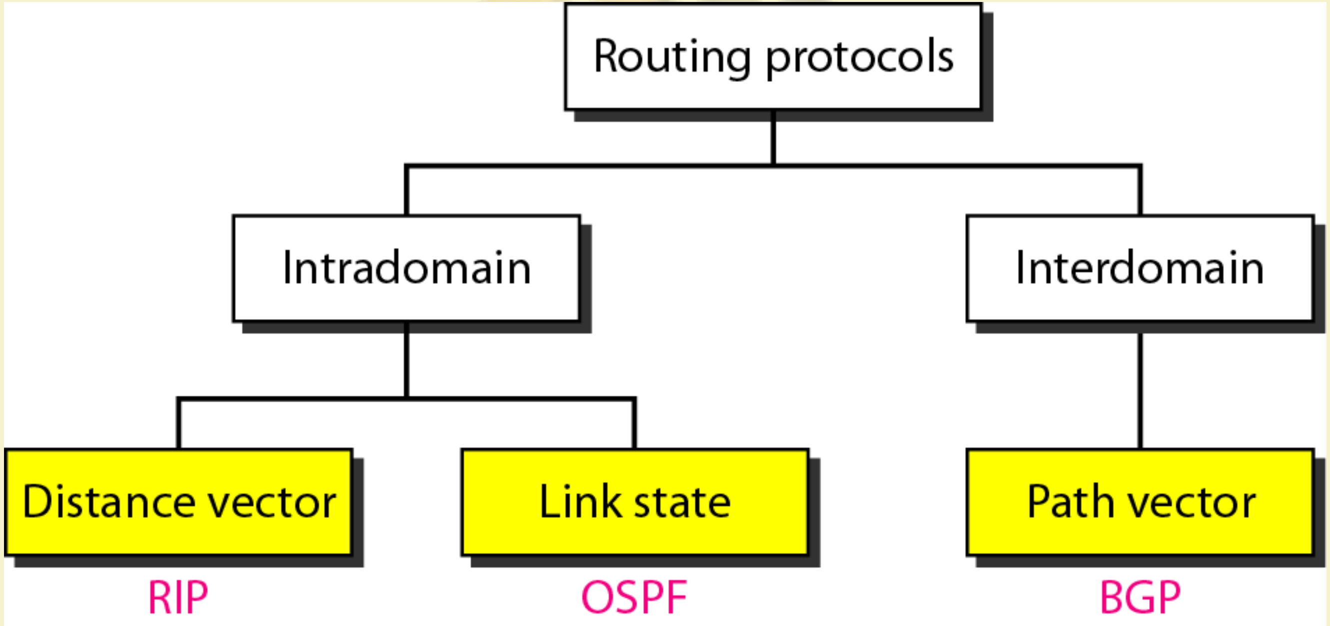
COMPUTER NETWORKS AND INTERNET PROTOCOLS

IP Routing - II [Intra-domain – Distance Vector, Link State]

SOUMYA K GHOSH
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

SANDIP CHAKRABORTY
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Dynamic Routing Protocols



Ref: *Data communications and networking by Behrouz A. Forouzan; TCP/IP Tutorials and Technical Overview, IBM Redbooks*

Routing Protocols

- **Static routing**: Static routing uses preprogrammed definitions representing paths through the network.
- **Dynamic routing**: Dynamic routing algorithms allow routers to automatically discover and maintain awareness of the paths through the network. The difference between these protocols is the way they discover and calculate new routes to destination networks.

Four broad categories:

- Distance vector protocols
- Link state protocols
- Path vector protocols
- Hybrid protocols

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

Routing Protocols

Distance vector protocols - Each router in the internetwork maintains the distance or cost from itself to its neighbors. The path represented by the smallest cost becomes the preferred path to reach the destination. This information is maintained in a distance vector table. The table is periodically advertised to each neighboring router. Each router processes these advertisements to determine the best paths through the network.

Link state protocols - Each router advertises a list of all directly connected network links and the associated cost of each link. This is performed through the exchange of link state advertisements (LSAs) with other routers in the network. Using these advertisements, each router creates a database detailing the current network topology. The topology database in each router is same.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

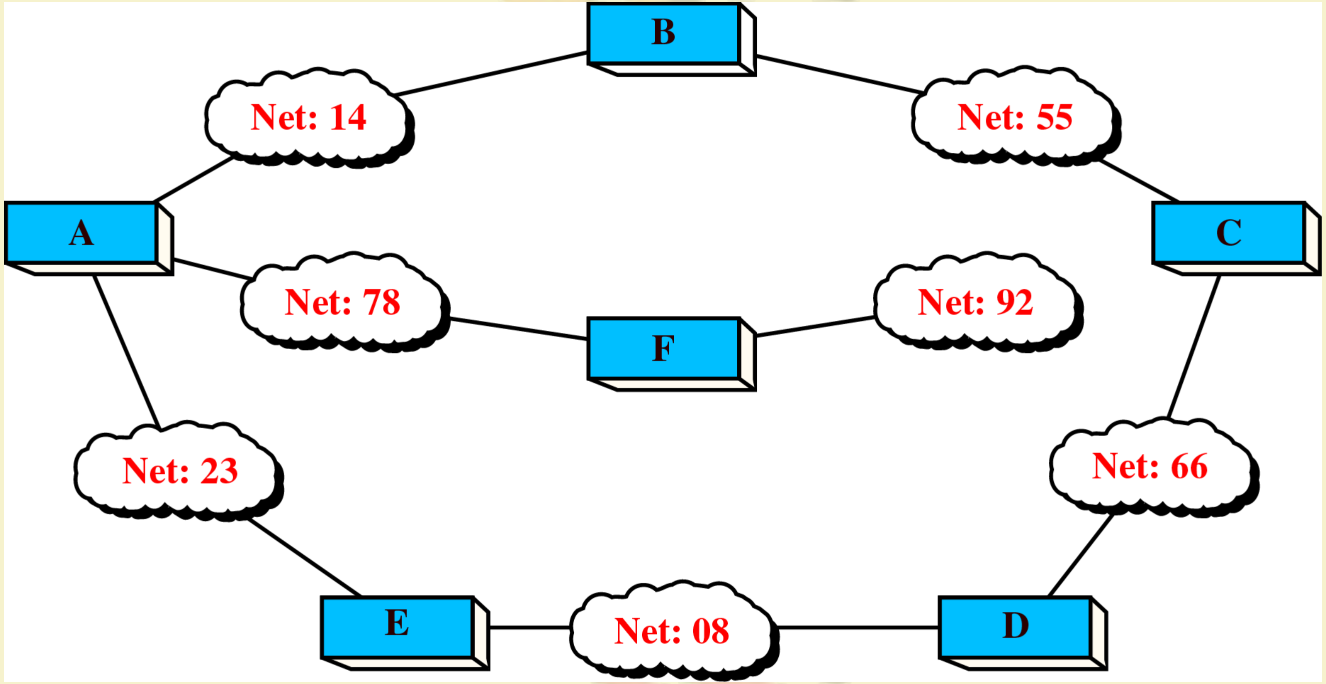
Routing Protocols

Path vector protocols - The path vector routing algorithm is somewhat similar to the distance vector algorithm. However, instead of advertising networks in terms of a destination and the distance to that destination, networks are advertised as destination addresses and path descriptions to reach those destinations.

Hybrid protocols - These protocols attempt to combine the positive attributes of both distance vector and link state protocols. Networks using hybrid protocols tend to converge more quickly than networks using distance vector protocols. Finally, these protocols potentially reduce the costs of link state updates and distance vector advertisements.

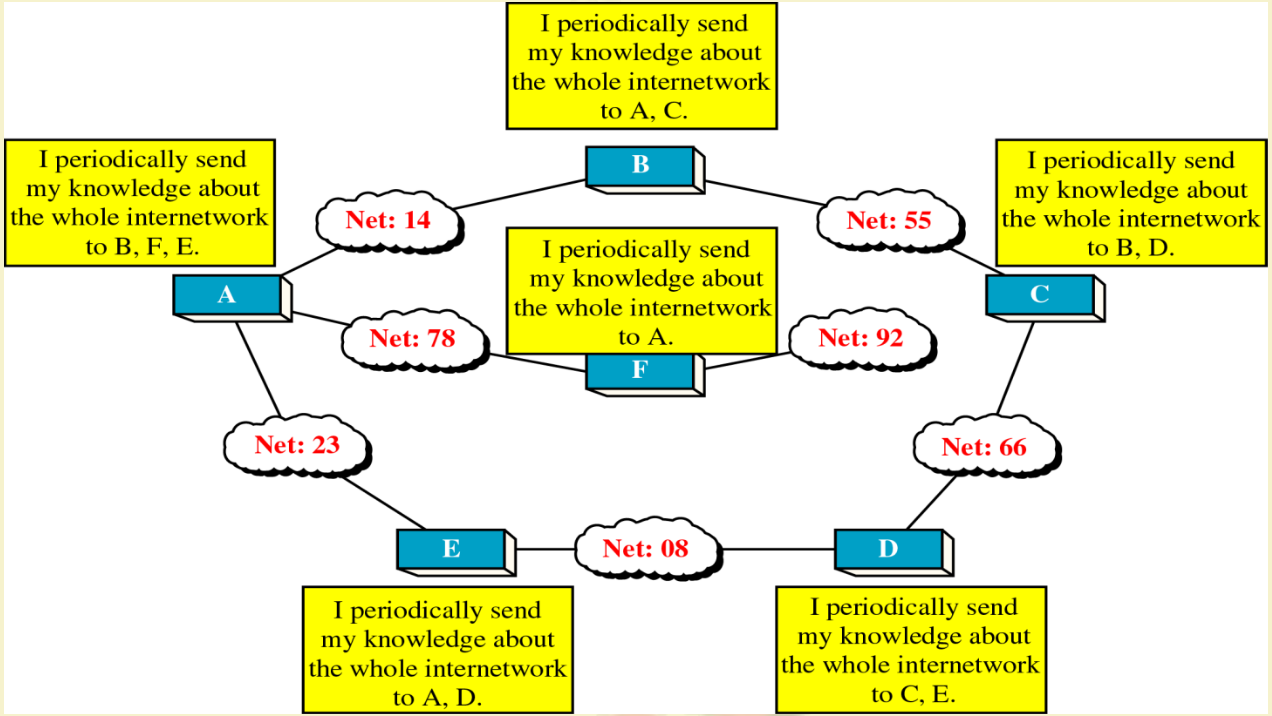
Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

Example of an Internet



Ref: Data communications and networking by Behrouz A. Forouzan;

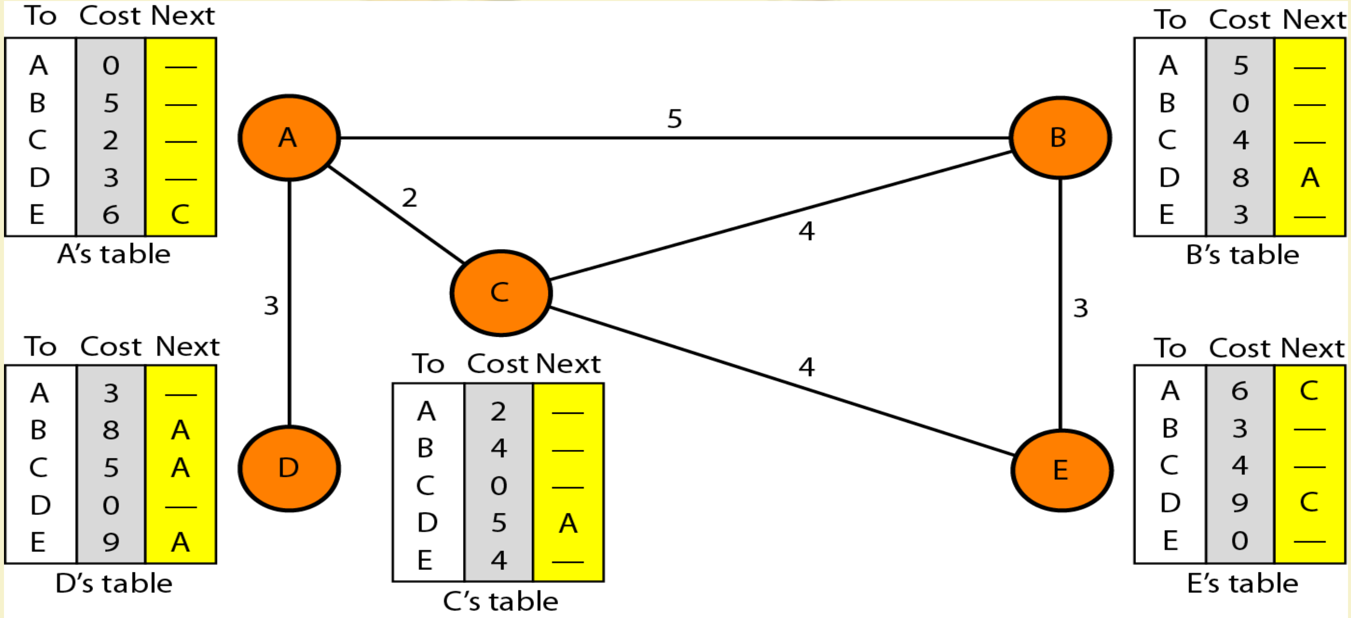
Distance Vector Routing - Concept



Ref: Data communications and networking by Behrouz A. Forouzan;

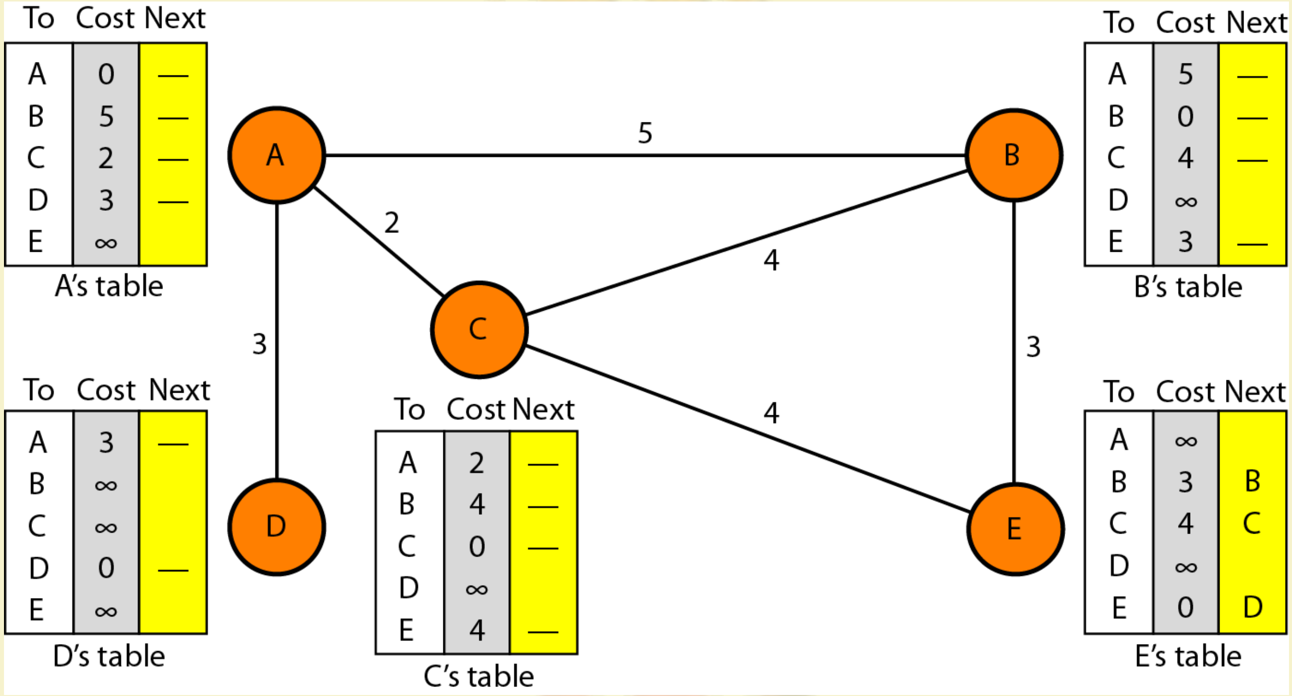
Distance Vector Routing

In distance vector routing, each node shares its routing table with its immediate neighbors periodically and when there is a change.



Ref: Data communications and networking by Behrouz A. Forouzan

Distance Vector Routing: Initialization of routing tables



Ref: Data communications and networking by Behrouz A. Forouzan

Distance Vector Routing – Updating Routing Table

To	Cost
A	2
B	4
C	0
D	∞
E	4

Received from C

To	Cost	Next
A	4	C
B	6	C
C	2	C
D	∞	C
E	6	C

A's modified table

Compare

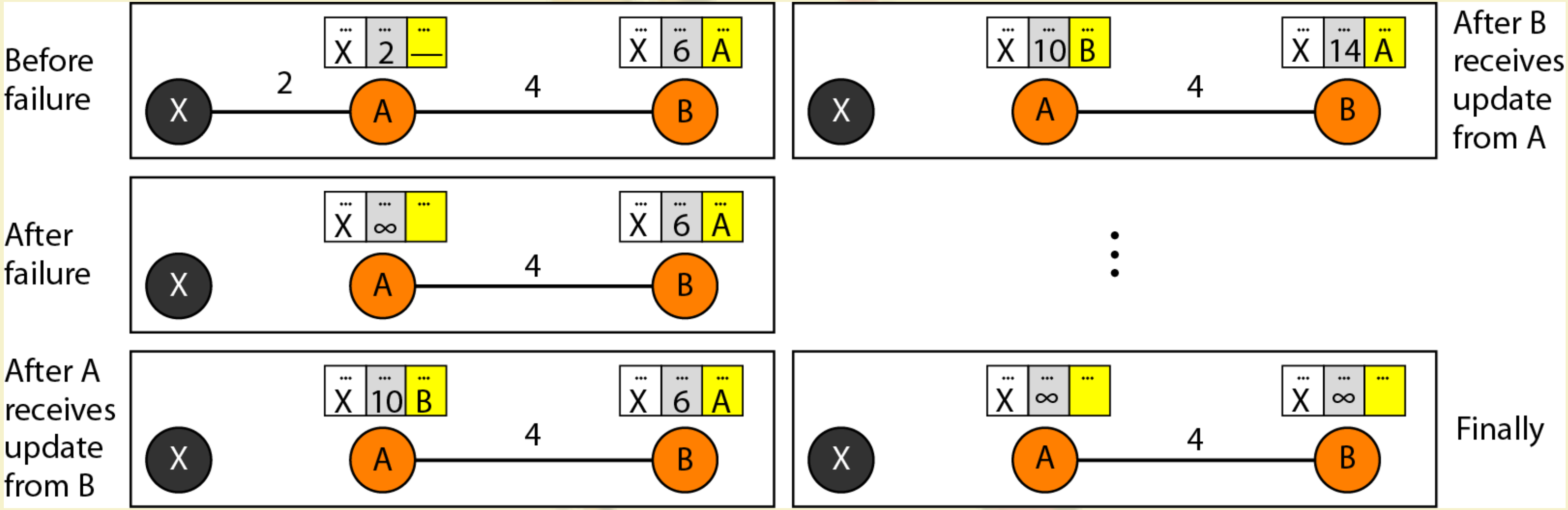
To	Cost	Next
A	0	—
B	5	—
C	2	—
D	3	—
E	∞	

A's old table

To	Cost	Next
A	0	—
B	5	—
C	2	—
D	3	—
E	6	C

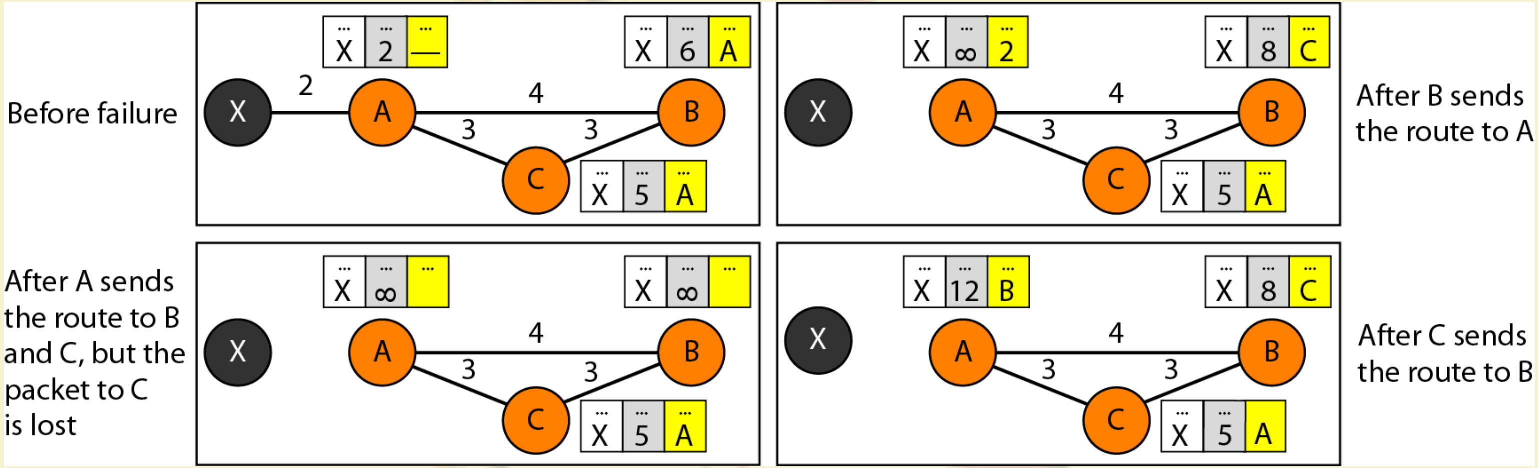
A's new table

Two-node Instability



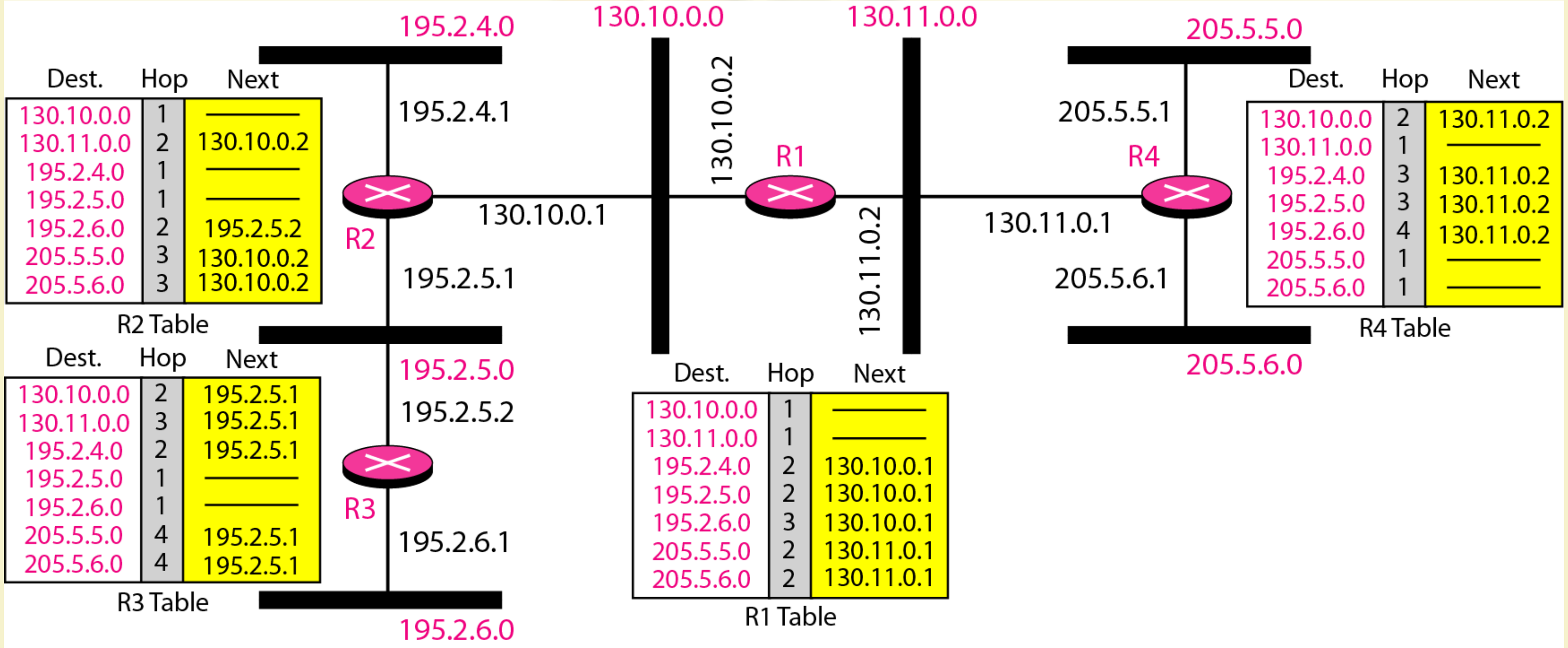
Ref: Data communications and networking by Behrouz A. Forouzan

Three-node Instability



Ref: Data communications and networking by Behrouz A. Forouzan

Example of a domain using RIP



Dest.	Hop	Next
130.10.0.0	1	———
130.11.0.0	2	130.10.0.2
195.2.4.0	1	———
195.2.5.0	1	———
195.2.6.0	2	195.2.5.2
205.5.5.0	3	130.10.0.2
205.5.6.0	3	130.10.0.2

R2 Table

Dest.	Hop	Next
130.10.0.0	2	195.2.5.1
130.11.0.0	3	195.2.5.1
195.2.4.0	2	195.2.5.1
195.2.5.0	1	———
195.2.6.0	1	———
205.5.5.0	4	195.2.5.1
205.5.6.0	4	195.2.5.1

R3 Table

Dest.	Hop	Next
130.10.0.0	1	———
130.11.0.0	1	———
195.2.4.0	2	130.10.0.1
195.2.5.0	2	130.10.0.1
195.2.6.0	3	130.10.0.1
205.5.5.0	2	130.11.0.1
205.5.6.0	2	130.11.0.1

R1 Table

Dest.	Hop	Next
130.10.0.0	2	130.11.0.2
130.11.0.0	1	———
195.2.4.0	3	130.11.0.2
195.2.5.0	3	130.11.0.2
195.2.6.0	4	130.11.0.2
205.5.5.0	1	———
205.5.6.0	1	———

R4 Table

Ref: Data communications and networking by Behrouz A. Forouzan

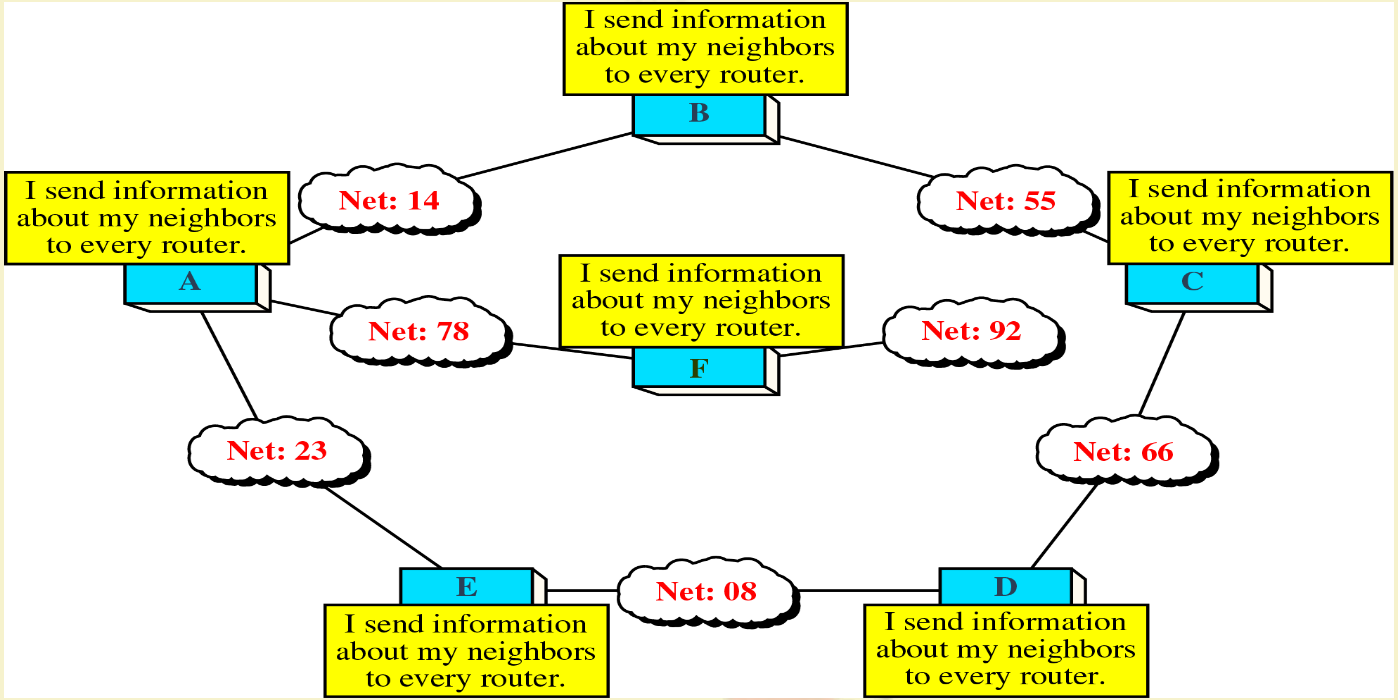
Link State Routing

A link state is the description of an interface on a router (e.g., IP address, subnet mask, network type) and its connectivity to neighboring routers. The collection of these link states forms a link state database. The routing algorithms use the principle of a *link state* to determine network topology.

Link state approach to determine network topology

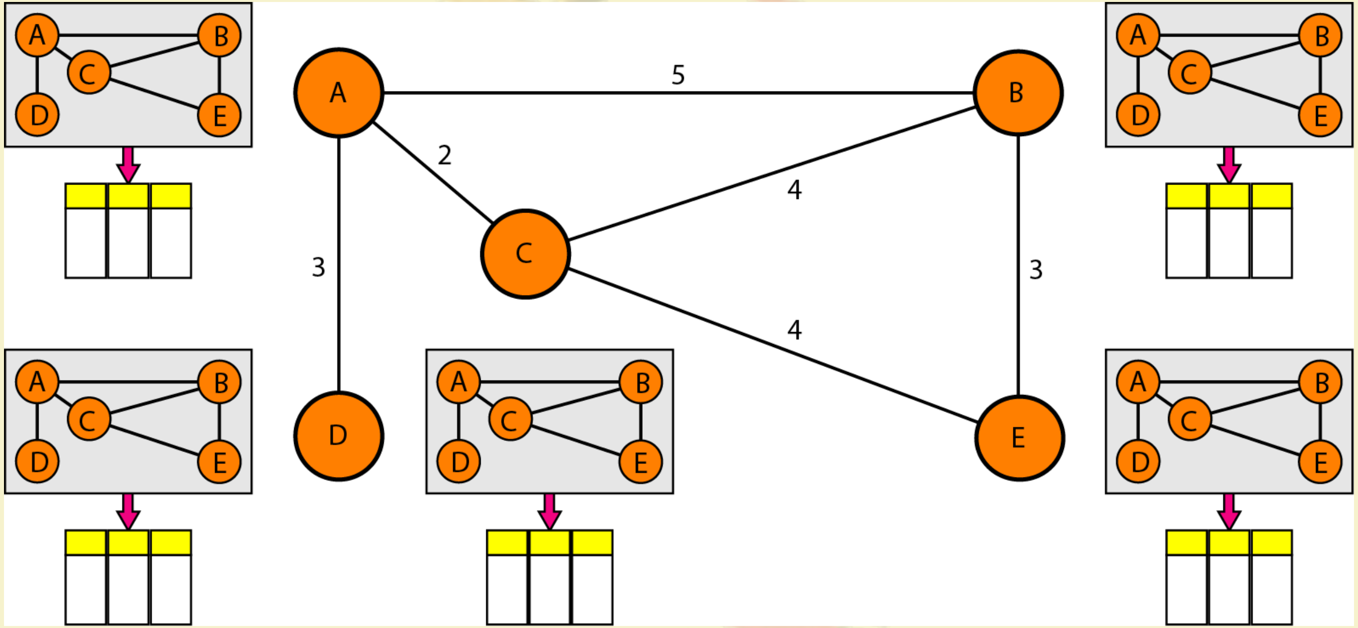
1. Each router identifies all other routing devices on the directly connected networks.
2. Each router advertises a list of all directly connected network links and the associated cost of each link; through the exchange of link state advertisements (LSAs) with other routers in the network.
3. Using these advertisements, each router creates a database detailing the current network topology. The topology database in each router is identical.
4. Each router uses the information in the topology database to compute the most desirable routes to each destination network. This information is used to update the IP routing table.

Link State Routing - Concept

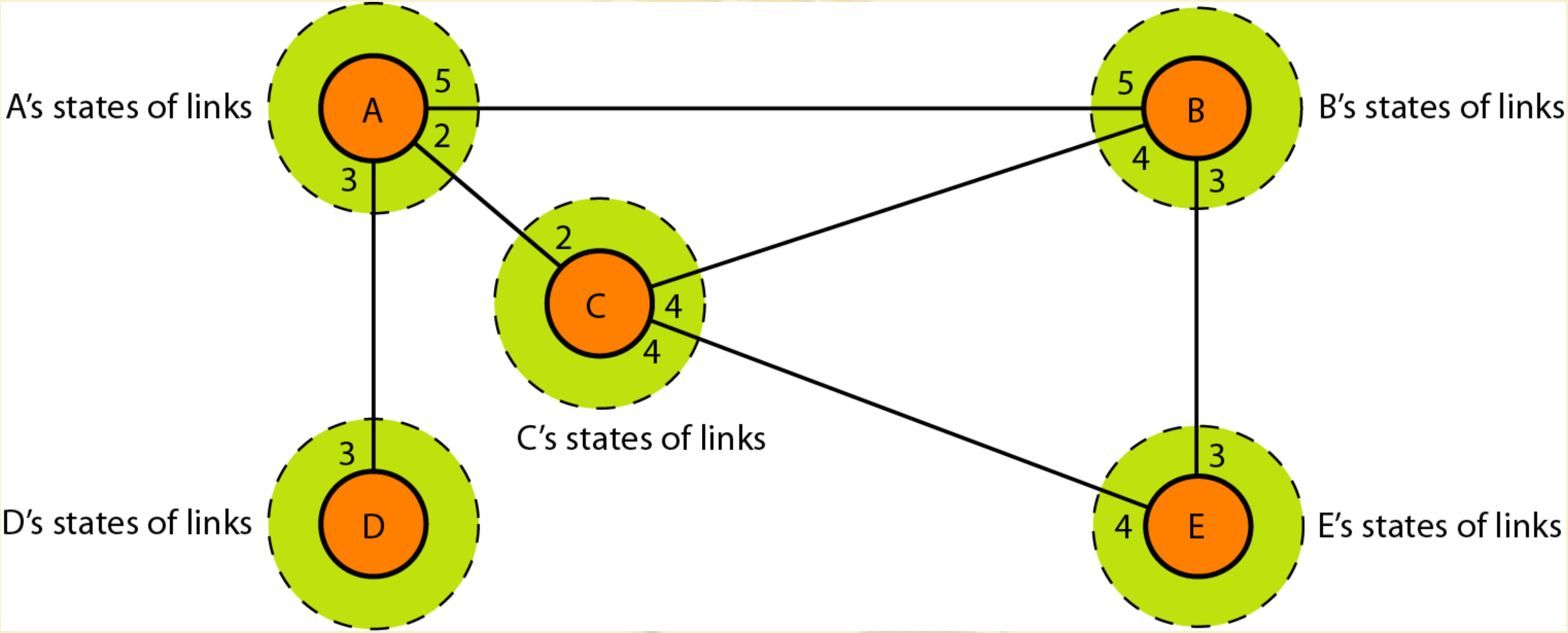


Ref: Data communications and networking by Behrouz A. Forouzan;

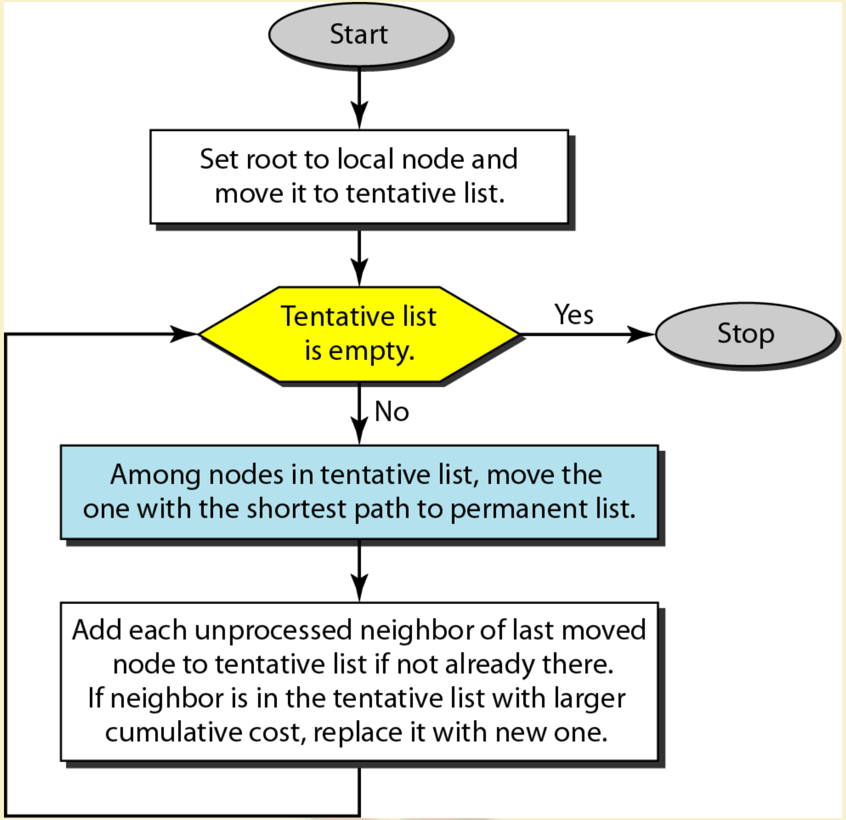
Link State Routing



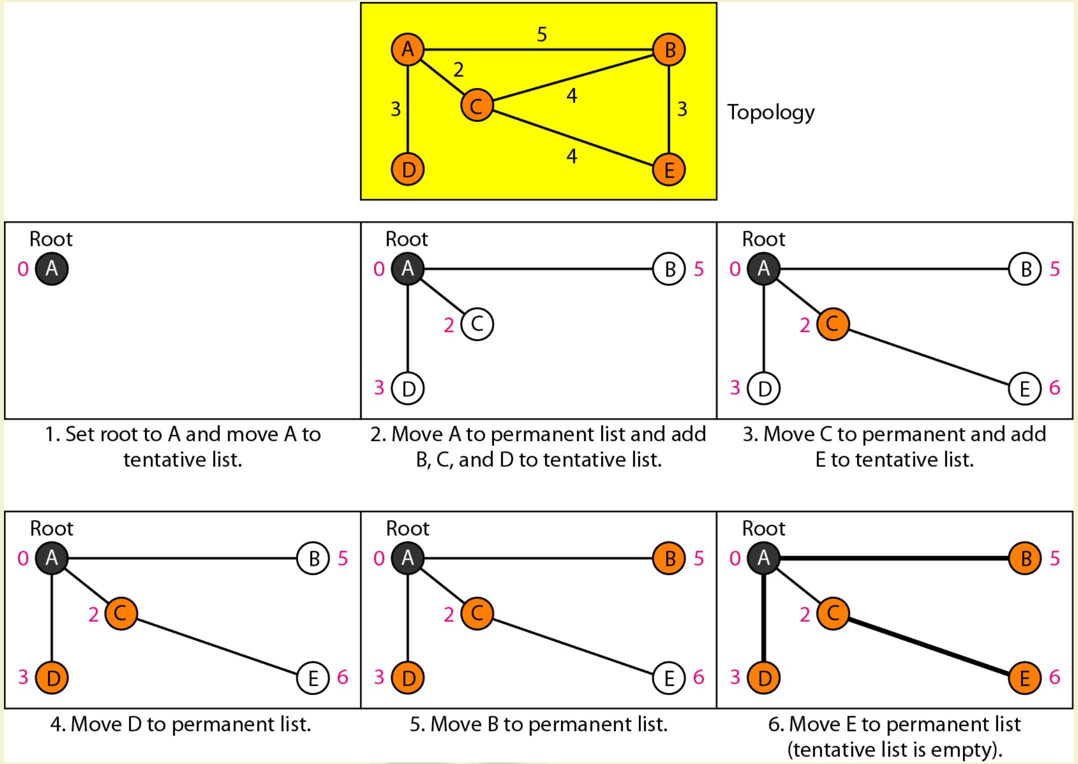
Link state knowledge



Dijkstra algorithm



Shortest Path tree



Node A - Routing table

Node	*Cost*	*Next Router*
A	0	—
B	5	—
C	2	—
D	3	—
E	6	C

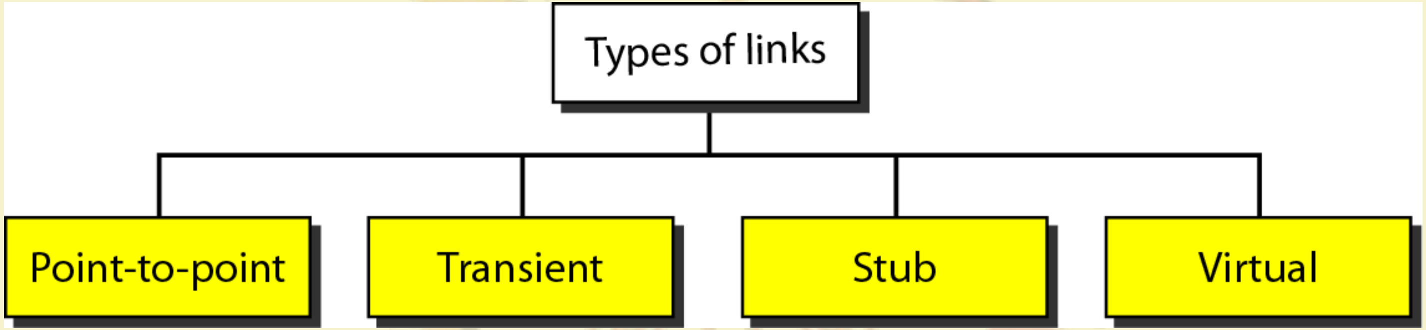
Link State Routing - Summary

1. Sharing knowledge about the neighbourhood
2. Sharing with every other router
3. Sharing when there is a change

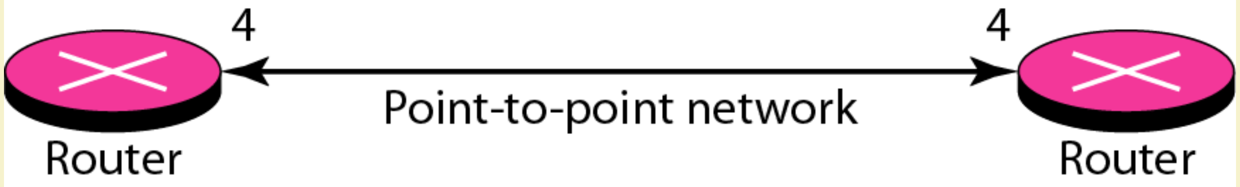
OSPF (Open Shortest Path First) uses Link State Routing to update the routing table.

thank you!

Types of links

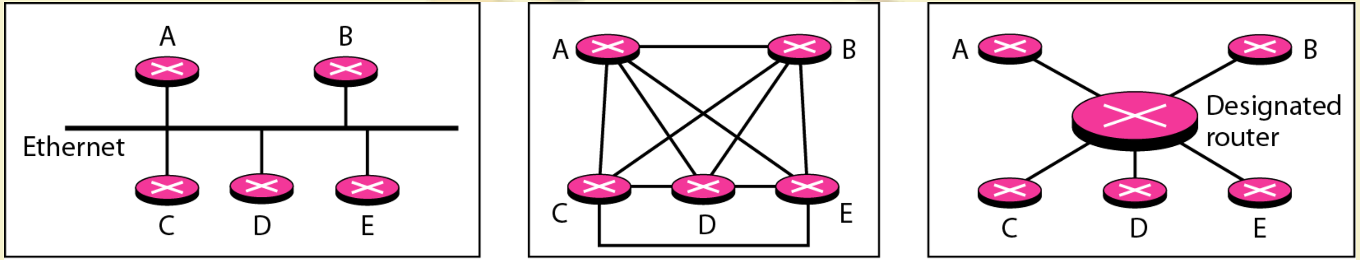


Point-to-point link



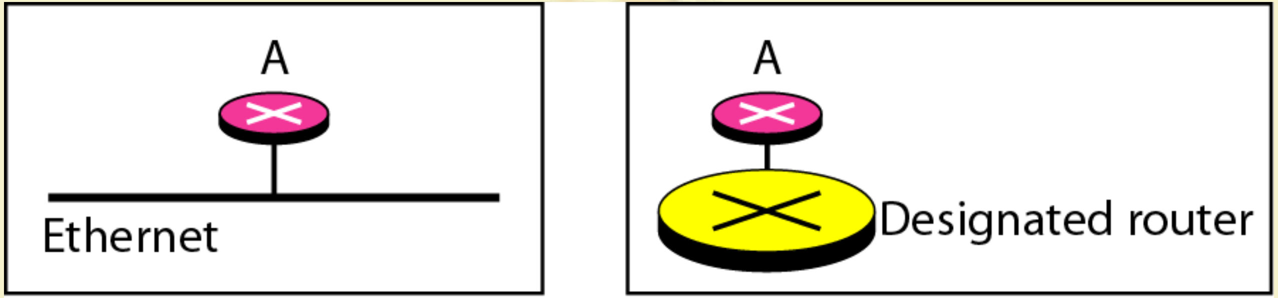
- Connects two routers without any other router(s) or host(s)

Transient link



a. Transient network
b. Unrealistic representation
c. Realistic representation

Stub link



a. Stub network b. Representation

Virtual link

Network administrator may create a virtual link between two routers, when the link between the routers are broken.

thank you!

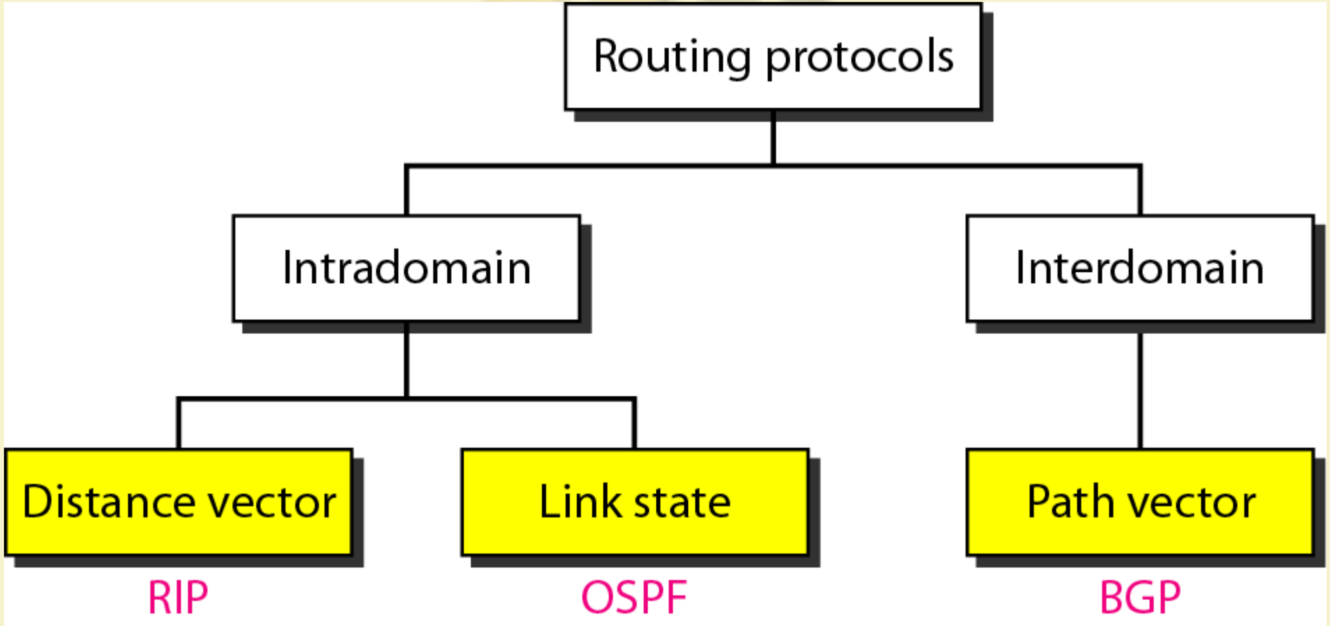
COMPUTER NETWORKS AND INTERNET PROTOCOLS

IP Routing - III [Autonomous System (AS), Path Vector Routing]

SOUMYA K GHOSH
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

SANDIP CHAKRABORTY
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

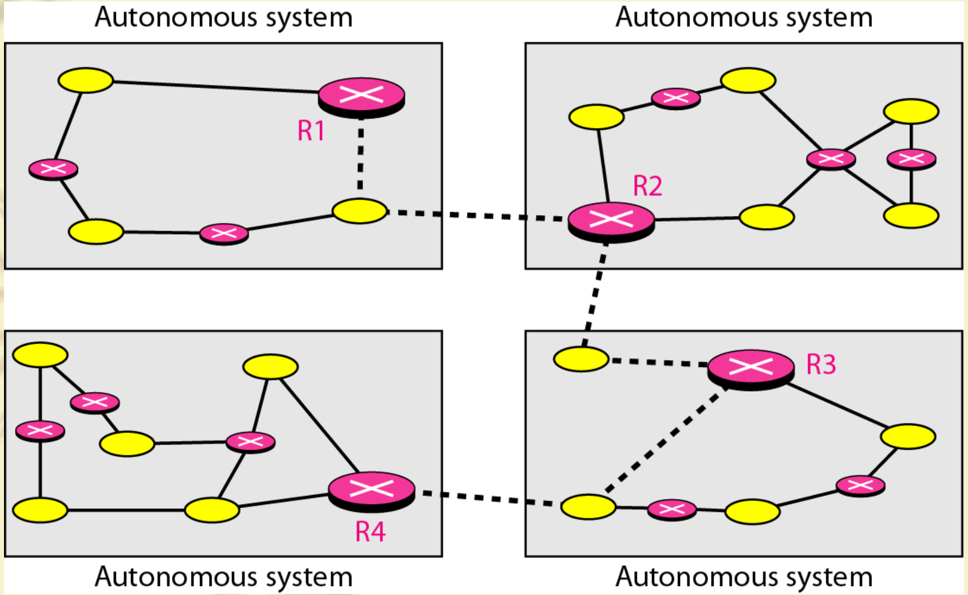
Dynamic Routing Protocols



Ref: Data communications and networking by Behrouz A. Forouzan; TCP/IP Tutorials and Technical Overview, IBM Redbooks

Autonomous System (AS)

AS is a logical portion of a larger IP network. An AS normally consists of an internetwork within an organization. It is administered by a single management authority. An AS can connect to other autonomous systems managed by the same organization or other public or private networks.



Ref: Data communications and networking by Behrouz A. Forouzan; TCP/IP Tutorials and Technical Overview, IBM Redbooks

Routing Protocols in AS

Two sets routing protocols are used – (i) to determine routing paths within an AS;
(ii) others are used to interconnect a set of autonomous systems:

Interior Gateway Protocols (IGPs): Interior Gateway Protocols allow routers to exchange information within an AS. Examples of these protocols are Open Short Path First (OSPF) and Routing Information Protocol (RIP).

Exterior Gateway Protocols (EGPs): Exterior Gateway Protocols allow the exchange of summary information between autonomous systems. An example of this type of routing protocol is Border Gateway Protocol (BGP).

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

Routing Protocols in AS

Two sets routing protocols are used – (i) to determine routing paths within an AS;
(ii) others are used to interconnect a set of autonomous systems:

Interior Gateway Protocols (IGPs): Interior Gateway Protocols allow routers to exchange information within an AS. Examples of these protocols are Open Short Path First (OSPF) and Routing Information Protocol (RIP).

Exterior Gateway Protocols (EGPs): Exterior Gateway Protocols allow the exchange of summary information between autonomous systems. An example of this type of routing protocol is Border Gateway Protocol (BGP).

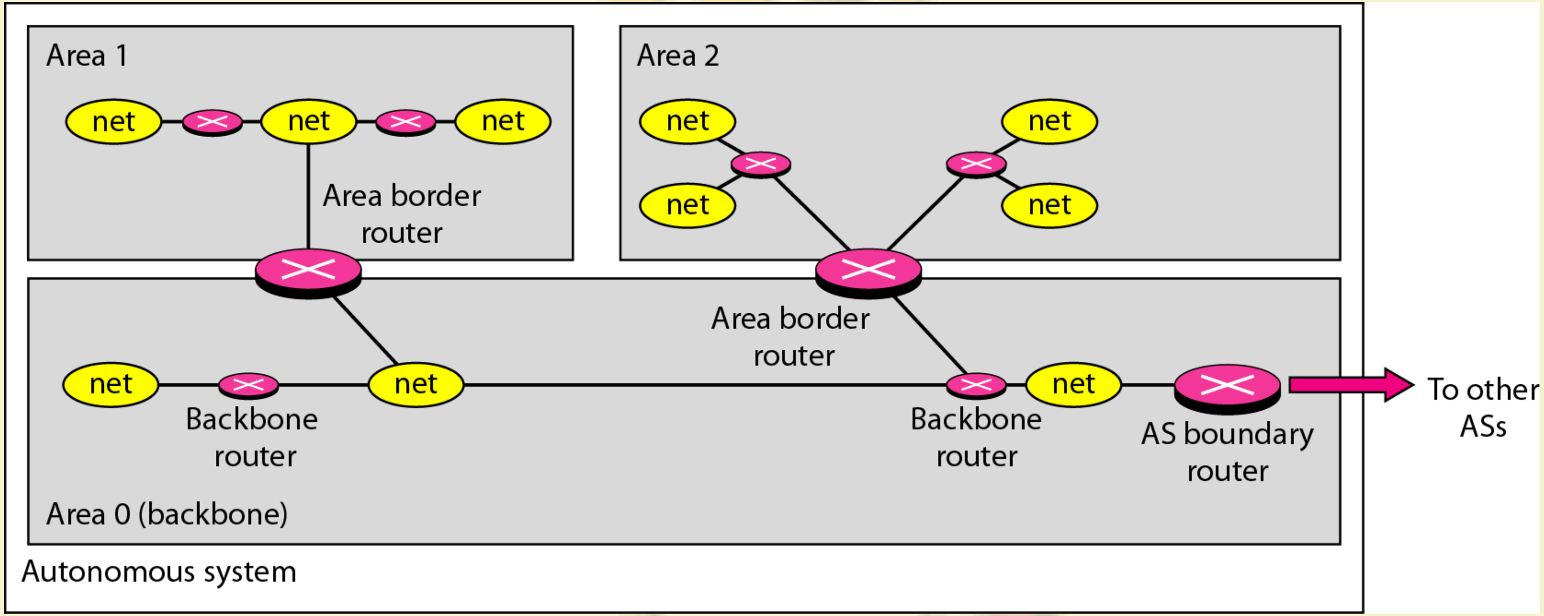
Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

Areas in AS

- **Area** is a collection of routers, networks and hosts within an AS. Each area is an area identification.
- An AS can be divided into different Areas. All network inside an area must be connected
- Routers inside an area flood the area with routing information. Special routers called **Border Area Routers** summarize the information about the area and send it other areas
- **Backbone** area – Special area inside AS. All areas in AS must be connected to the backbone. Backbone area acts as primary area and other areas as secondary.
- Routers within the backbone area are called **backbone routers**. A backbone router can also be a area border router.

Ref: Data communications and networking by Behrouz A. Forouzan

Areas in AS



Ref: Data communications and networking by Behrouz A. Forouzan

Path Vector Routing

- In Path Vector routing, the routing table contains <destination address>, <next router> and the <path to reach the destination>
- Path is defined as an ordered list of Autonomous Systems (ASs) that the packet need to travel through.

Path Vector Messages

- AS routers that participate in path vector routing advertise the reachability of the networks in their Ass to neighbors autonomous boundary routers. Two autonomous boundary routers connected to the same network are neighbors.
- Autonomous boundary routers receive information from interior routing protocols, like RIP, OSPF
- Each router that receives a path vector verifies that the advertised path is in agreement with the defined policy

Loop Prevention

- When a router receives a message, if there is loop, the message is ignored.

Policy Routing

- Policy routing can be implemented in path vector routing.

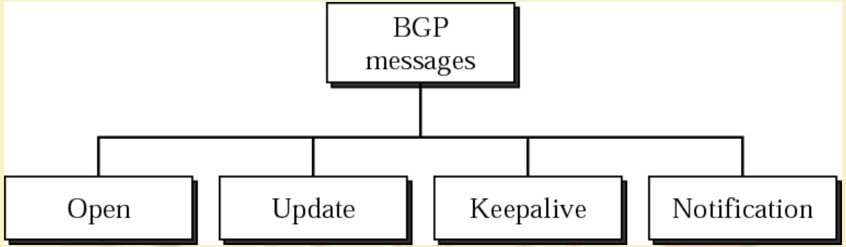
Path Attributes

- Two broad categories: (i) Well-known attribute; (ii) Optional atrribute
- Well-known attribute: Should be recognized by every BGP router. Well-known attribute: Mandatory and Discretionary

Border Gateway Protocol (RFC 1771)

- Based on the path vector routing.
- Distance-vector protocol not preferred for inter-AS routing (exterior routing protocol)
  - Assumes all routers have a common distance metrics to judge route preferences.
    - If routers have different meanings of a metric, it may not be possible to create stable, loop free routes.
  - A given AS may have different priorities from another AS.
  - Gives no information about the ASs that will be visited.
- Link-state routing protocol
  - Different metrics.
  - Flooding is not realistic.
- Path vector routing
  - No metrics,
  - Information about which networks can be reached by a given router and ASs to be crossed.
- Differs from DVA
  - Path vector approach does not include a distance or cost estimate
  - Lists all of the ASs visited to reach destination network.

BGP (continued)



- Messages are sent over TCP connections on port 179.
- Functional procedures
  - Neighbor acquisition (open message, acceptance through Keepalive message)
  - Neighbor reachability (periodic Keepalive messages)
  - Network reachability (broadcast an update message)
    - Each routers maintains a database of networks that can be reached
    - + preferred route to this network.
- RFC does not address
    - How a router knows the address of another router.
    - Up to network admin.

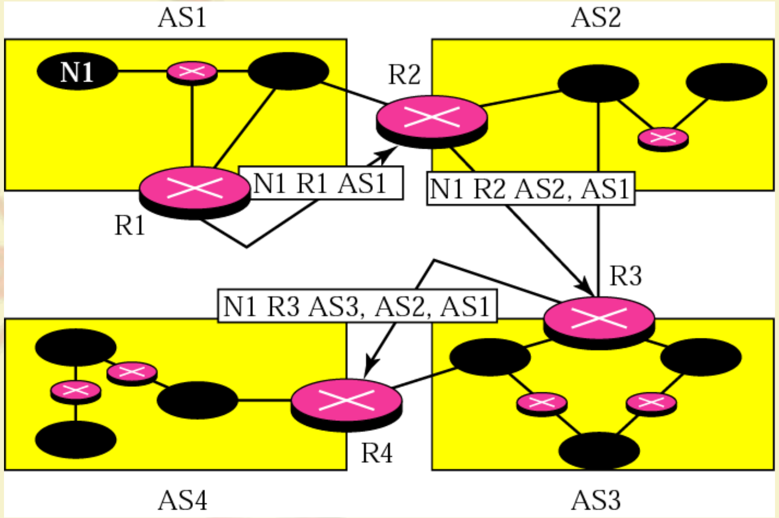
BGP (cont.)

Example of Network Reachability

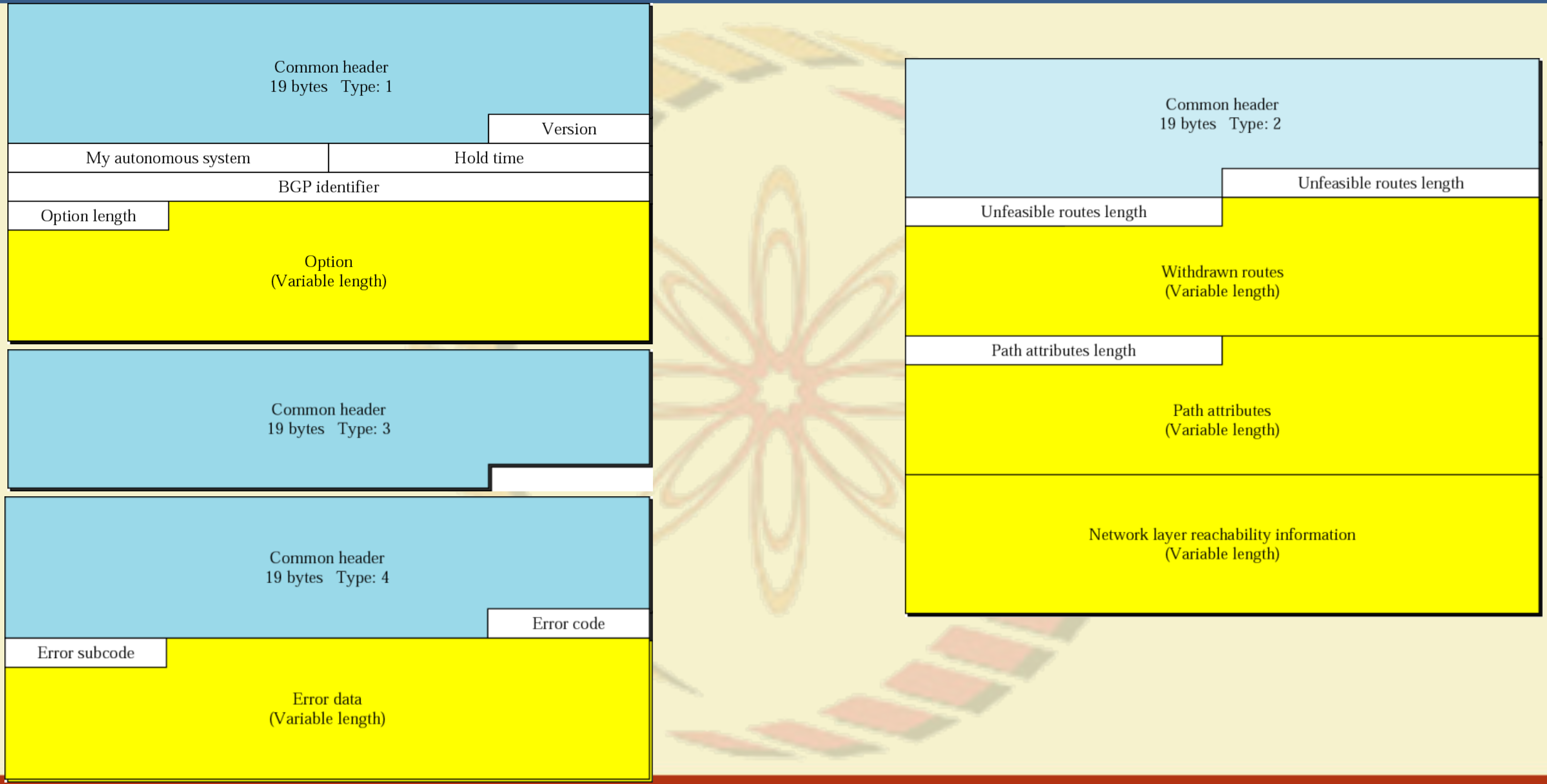
Network	Next router	Path
N1	R1	AS14,AS23,AS67
N2	R5	AS22,AS67,AS5,AS89
N3	R6	AS67,AS89,AS9,AS34
N4	R12	AS62,AS2,AS9

- Loop Prevention in BGP:
  - Checks the Path before updating its database. (If its AS is in the path ignore the message)
- Policy Routing:
  - If a path consist of an AS against the policy of the current AS, message discarded.

Example of Message adverstisements

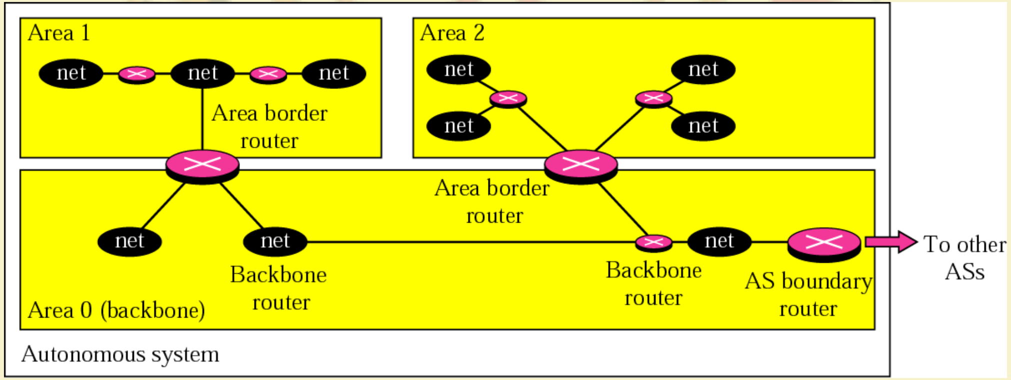


BGP message format (Open, Keepalive, Update, Notification

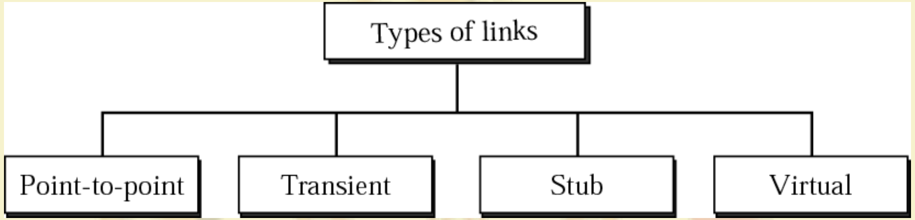


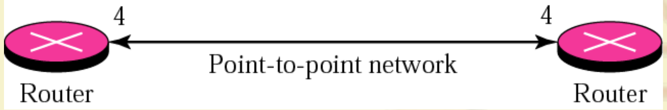
Open Shortest Path First (RFC 1247)

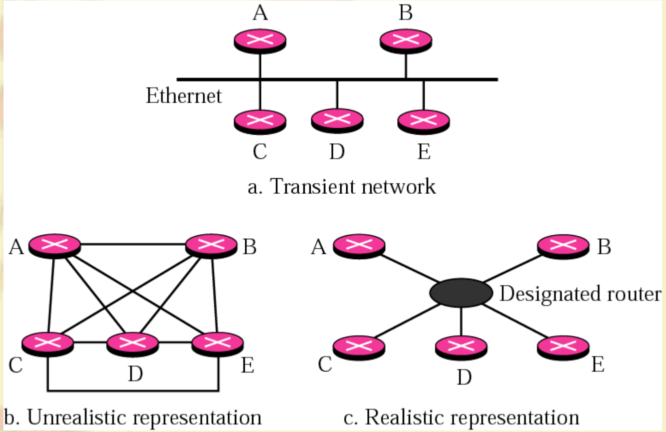
- Uses IP, has a value in the IP Header (8 bit protocol field)
- Interior routing protocol, its domain is also an autonomous system (AS)
- Special routers (autonomous system boundary routers) or backbone routers responsible to dissipate information about other AS into the current system.
- Divides an AS into areas
- Metric based on type of service
  - Minimum delay (RTT), maximum throughput, reliability, etc..

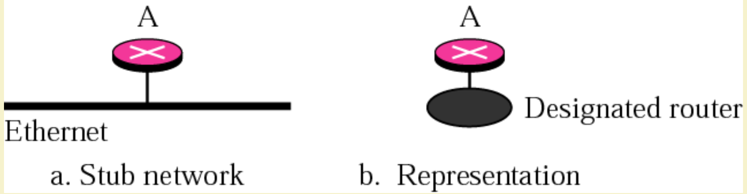


OSPF (Type of Links)

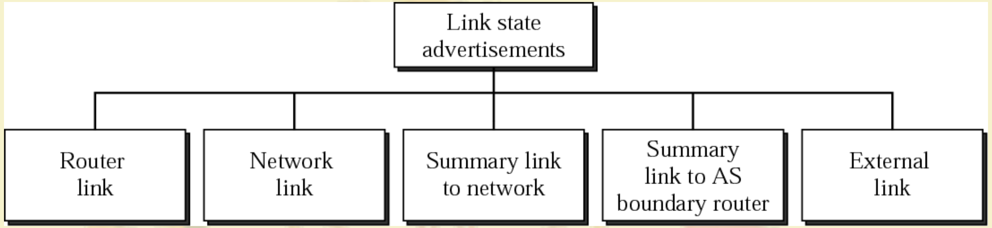


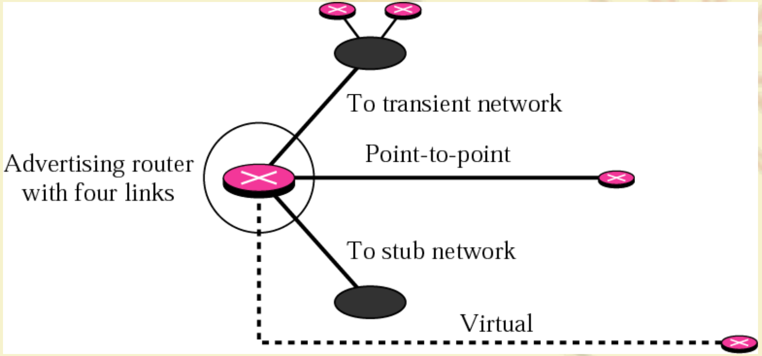




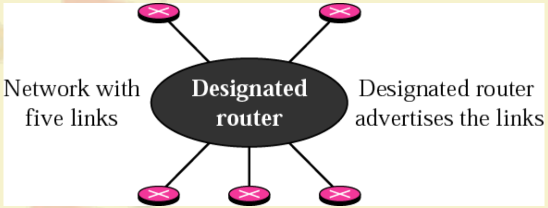


OSPF (Link State Advertisement)



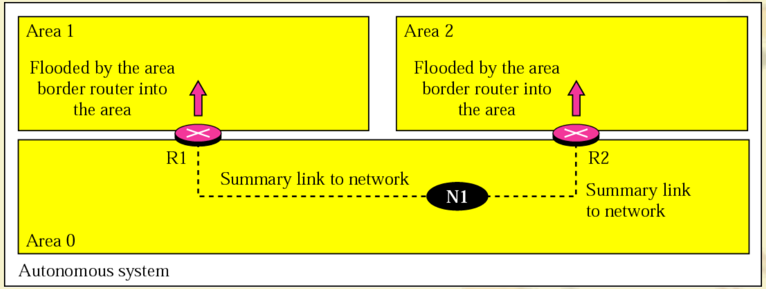


Router Link

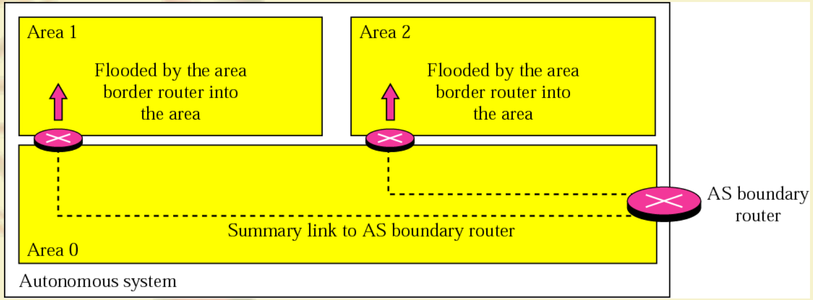


Network Link

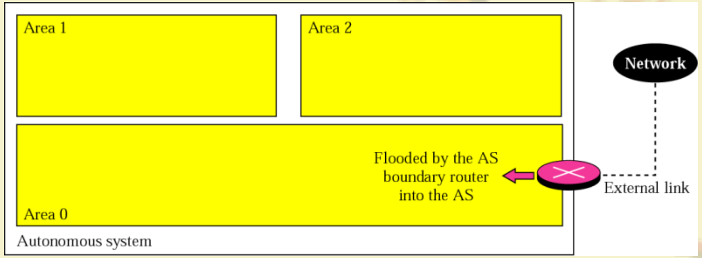
OSPF (LSA cont.)



Summary link to Network

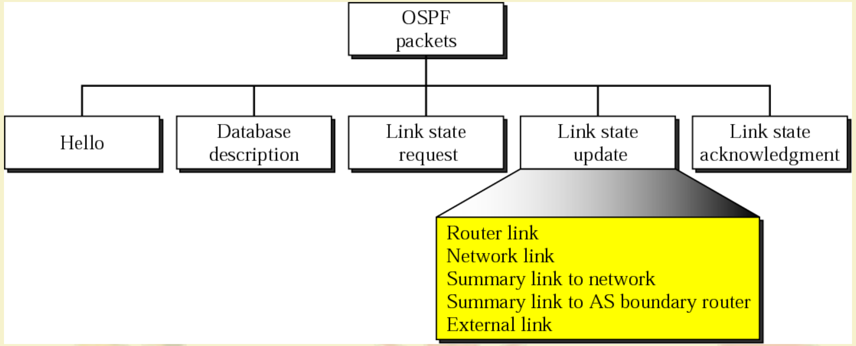


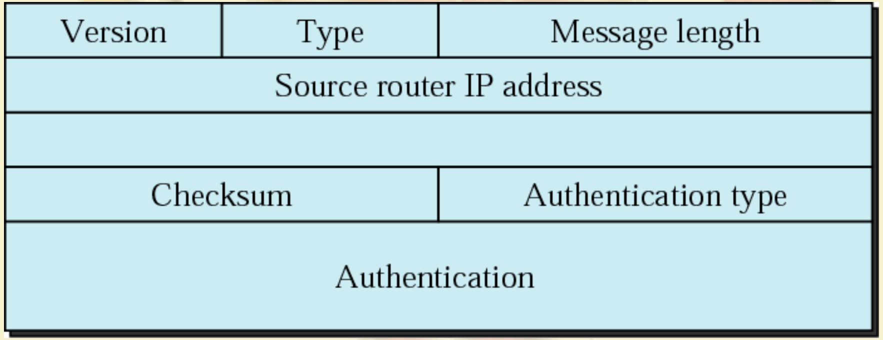
Summary link to AS boundary router



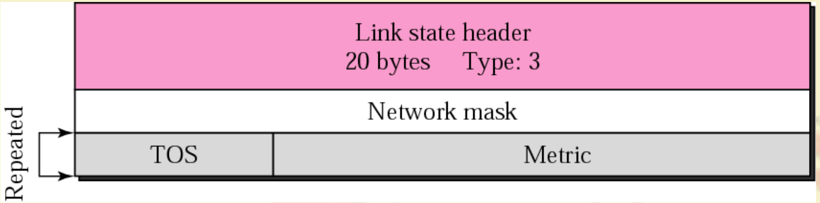
External Link

Types of OSPF Packets

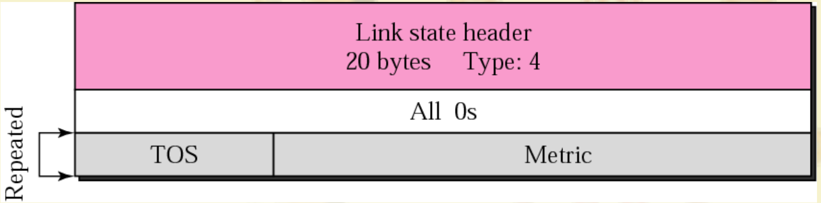




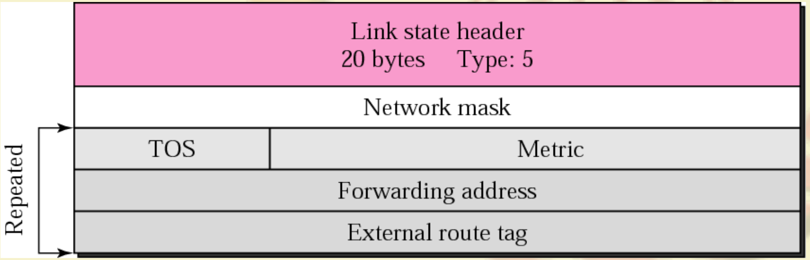
Links state Advertisements - Summary



Summary link to network



Summary link to AS boundary



External Link

thank you!

COMPUTER NETWORKS AND INTERNET PROTOCOLS

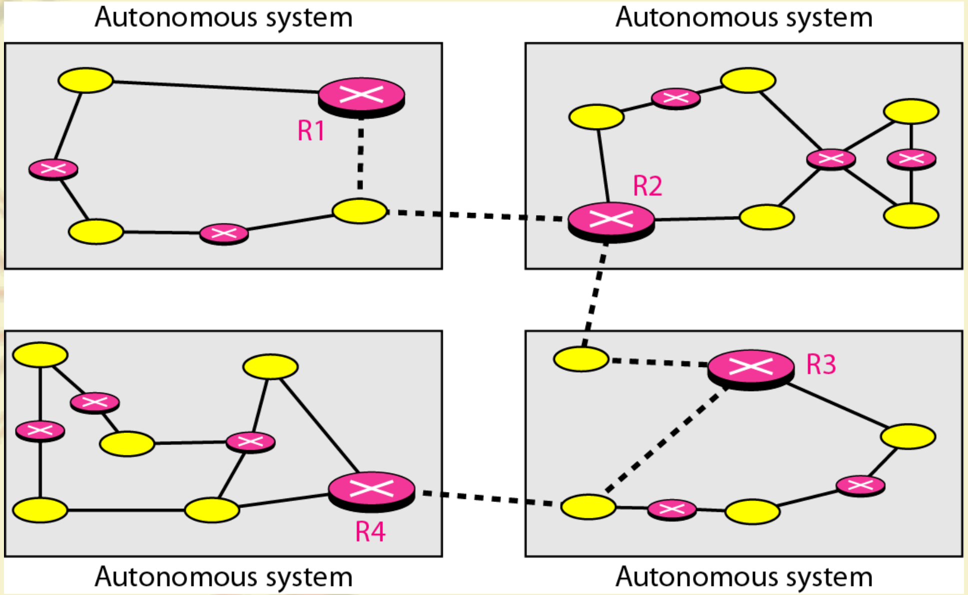
IP Routing – IV [Border Gateway Protocol - BGP]

SOUMYA K GHOSH
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

SANDIP CHAKRABORTY
COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Autonomous System (AS)

AS is a logical portion of a larger IP network. An AS normally consists of an internetwork within an organization. It is administered by a single management authority. An AS can connect to other autonomous systems managed by the same organization or other public or private networks.



Ref: *Data communications and networking by Behrouz A. Forouzan; TCP/IP Tutorials and Technical Overview, IBM Redbooks*

Routing Protocols in AS

Two sets routing protocols are used – (i) to determine routing paths within an AS; (ii) others are used to interconnect a set of autonomous systems:

Interior Gateway Protocols (IGPs): Interior Gateway Protocols allow routers to exchange information within an AS. Examples of these protocols are Open Short Path First (OSPF) and Routing Information Protocol (RIP).

Exterior Gateway Protocols (EGPs): Exterior Gateway Protocols allow the exchange of summary information between autonomous systems. An example of this type of routing protocol is <u>**Border Gateway Protocol (BGP).**</u>

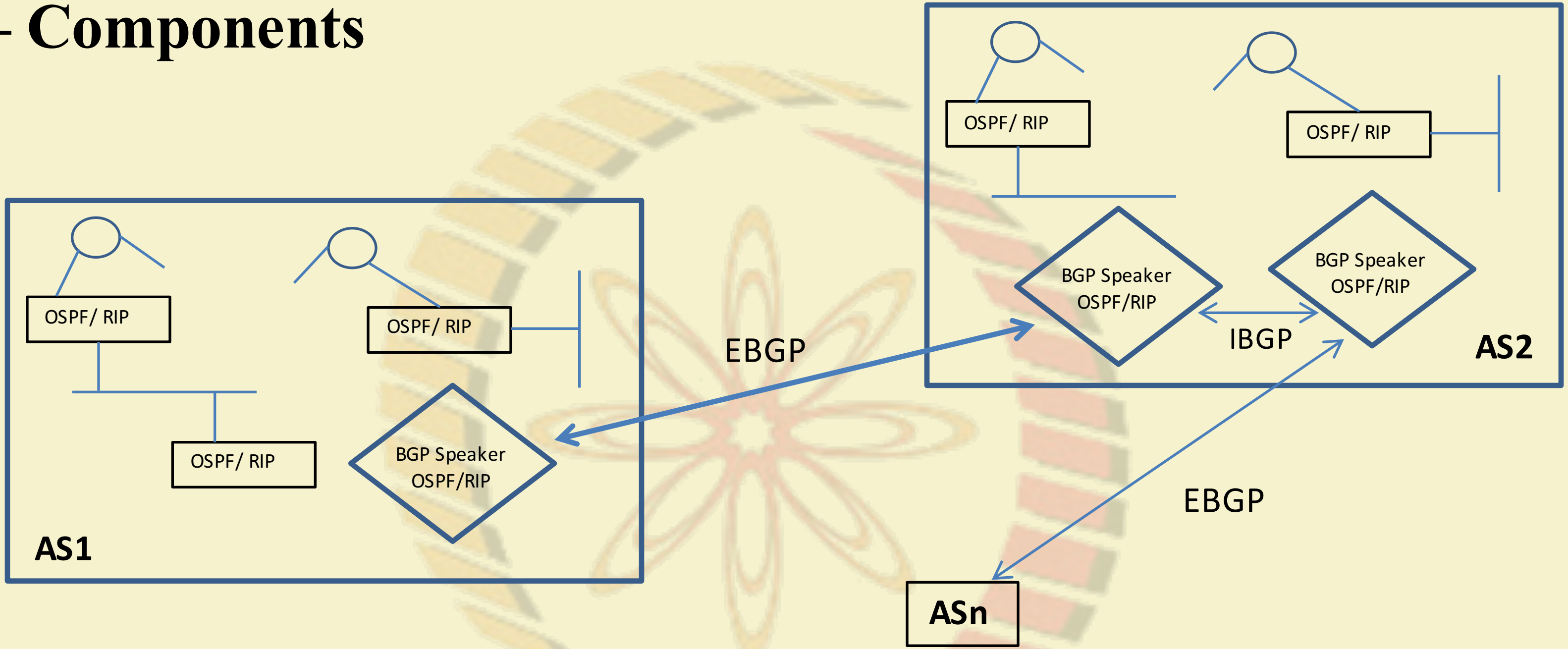
Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

Border Gateway Protocol (BGP)

- Border Gateway Protocol (BGP) is an exterior gateway protocol.
- It was originally developed to provide a loop-free method of exchanging routing information between autonomous systems.
- BGP has since evolved to support aggregation and summarization of routing information.
- BGP is an IETF draft standard protocol described in RFC 4271 (BGP Version 4).

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP – Components



Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP – Basic Concepts

- **BGP speaker**: A router configured to support BGP.
- **BGP neighbors** (peers): A pair of BGP speakers that exchange routing information.
- Two types of BGP neighbors:
  - Internal (**IBGP**) neighbor: A pair of BGP speakers within the same AS.
  - External (**EBGP**) neighbor: A pair of BGP neighbors, each in a different AS. These neighbors typically share a directly connected network.
- **BGP session**: A TCP session connecting two BGP neighbors. The session is used to exchange routing information. The neighbors monitor the state of the session by sending *Keepalive* messages.
- **AS number**: A 16-bit number uniquely identifying an AS.
- **AS path**: A list of AS numbers describing a route through the network. A BGPneighbor communicates paths to its peers.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP – Basic Concepts (contd…)

- **Traffic type**: BGP defines two types of traffic:
  - **Local**: Traffic local to an AS either originates or terminates within the AS. Either the source or the destination IP address resides in the AS.
  - **Transit**: Any traffic that is not local traffic is transit traffic. One of the goals of BGP is to minimize the amount of transit traffic.
- **AS type**: BGP defines three types of autonomous systems:
  - ***Stub***: A stub AS has a single connection to one other AS. A stub AS carries only local traffic.
  - ***Multihomed***: A multihomed AS has connections to two or more autonomous systems. However, a multihomed AS has been configured so that it does not forward transit traffic.
  - ***Transit***: A transit AS has connections to two or more autonomous systems and carries both local and transit traffic. The AS can impose policy restrictions on the types of transit traffic that will be forwarded.
  - The autonomous system can be either a multihomed AS or a transit AS.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP – Basic Concepts (contd…)

- **Routing policy**: A set of rules constraining the flow of data packets through the network. Routing policies are not defined in the BGP protocol. Rather, they are used to configure a BGP device. For example, a BGP device can be configured so that:
  - A multihomed AS can refuse to act as a transit AS. This is accomplished by advertising only those networks contained within the AS.
  - A multihomed AS can perform transit AS routing for a restricted set of adjacent autonomous systems. It does this by tailoring the routing advertisements sent to EBGP peers.
  - An AS can optimize traffic to use a specific AS path for certain categories of traffic.
- **Network layer reachability information (NLRI)**: NLRI is used by BGP to advertise routes. It consists of a set of networks represented by the tuple <length,prefix>. For example, the tuple <14,202.124.116.0> represents the CIDR route 202.124.116.0/14.
- **Routes and paths**: A route associates a destination with a collection of attributes describing the path to the destination. The destination is specified in NRLI format. The path is reported as a collection of path attributes. This information is advertised in UPDATE messages.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

EBGP – IBGP Communications

- BGP does not replace the IGP operating within an AS. Instead, it cooperates with the IGP to establish communication between autonomous systems.
- BGP within an AS is used to advertise the local IGP routes. These routes are advertised to BGP peers in other ASs.

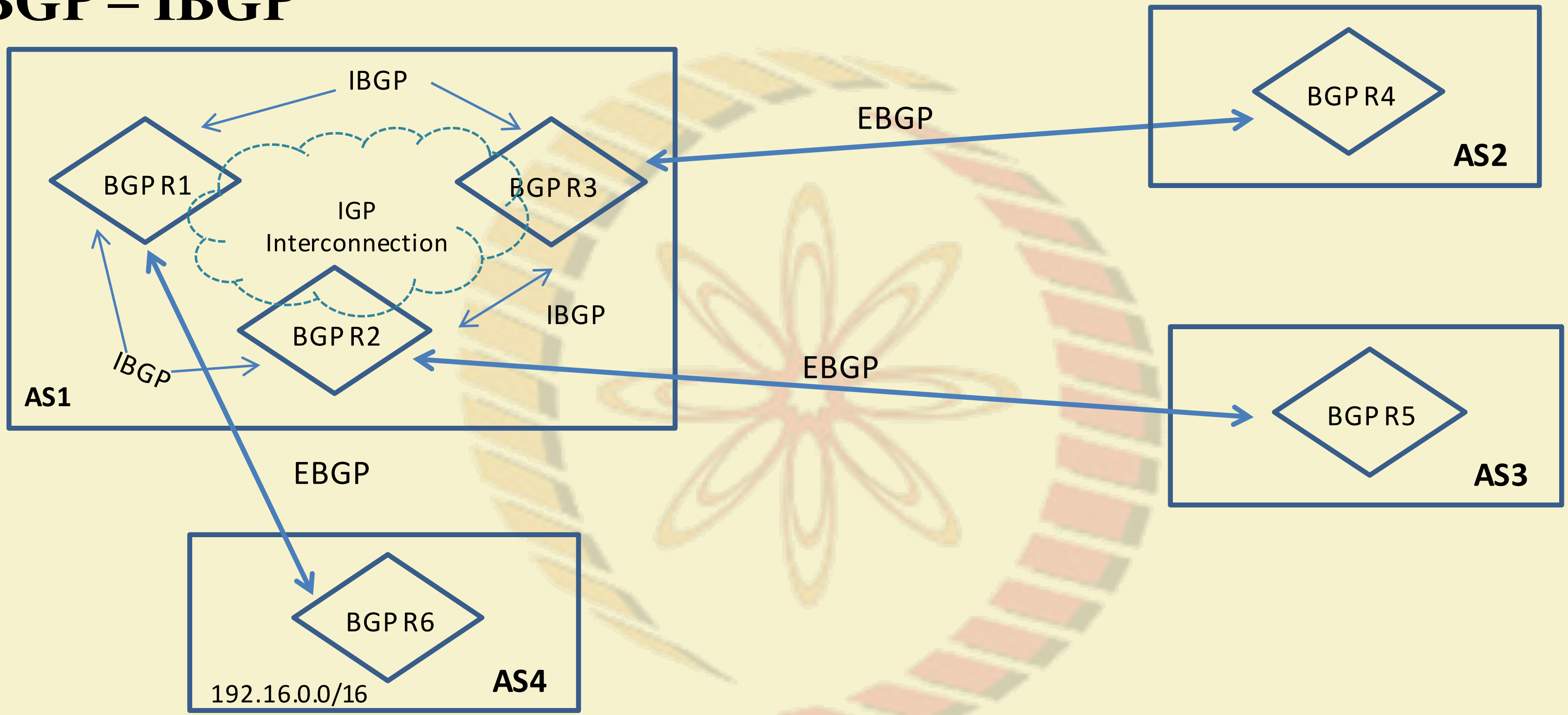
Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

EBGP – IBGP Communications

- Role of BGP and the IGP: Both BGP and the IGP are used to carry information through an AS.
- Establishing the TCP session between peers: Before establishing a BGP session, a device verifies that routing information is available to reach the peer:
    - **EBGP peers**: EBGP peers typically share a directly connected network. The routing information needed to exchange BGP packets between these peers is trivial.
    - **IBGP peers**: IBGP peers can be located anywhere within the AS. They do not need to be directly connected. BGP relies on the IGP to locate a peer. Packet forwarding between IBGP peers uses IGP-learned routes.
- Full mesh of BGP sessions within an AS: IBGP speakers assume a full mesh of BGP sessions have been established between peers in the same AS.
- When a BGP speaker receives a route update from an IBGP peer, the receiving speaker uses EBGP to propagate the update to external peers. Because the receiving speaker assumes a full mesh of IBGP sessions have been established, it does not propagate the update to other IBGP peers

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

EBGP – IBGP



Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP Protocol

- BGP establishes a reliable TCP connection between peers. Sessions are established using TCP port 179.
- BGP assumes the transport connection will manage fragmentation, retransmission, acknowledgement, and sequencing.
- When two speakers initially form a BGP session, they exchange their entire routing table. This routing information contains the complete AS path used to reach each destination. This information avoids the routing loops and counting-to-infinity behavior observed in RIP networks.
- After the entire table has been exchanged, changes to the table are communicated as incremental updates.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP Packet types

- OPEN: This message type establishes a BGP session between two peer nodes.
- UPDATE: This message type transfers routing information between GP peers.
- NOTIFICATION: This message is sent when an error condition is detected.
- KEEPALIVE: This message determines if peers are reachable.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP Function

- **Opening and confirming a BGP connection**: After a TCP session has been established between two peer nodes, each router sends an OPEN message to the neighbor.
- **Maintaining the BGP connection**: BGP does not use any transport-based keepalive to determine if peers are reachable. Instead, BGP messages are periodically exchanged between peers. If no messages are received from the peer for the duration specified by the hold timer, the originating router assumes that an error has occurred. When this happens, an error notification is sent to the peer and the connection is closed.
- **Sending reachability information**: Reachability information is exchanged between peers in UPDATE messages. An UPDATE message is used to advertise feasible routes or withdraw infeasible routes.
- **Notification of error conditions**: A BGP device can observe error conditions impacting the connection to a peer. NOTIFICATION messages are sent to the neighbor when these conditions are detected. After the message is sent, the BGP transport connection is closed. This means that all resources for the BGP connection are de-allocated. The routing table entries associated with the remote peer are marked as invalid. Finally, other peers are notified that these routes are invalid.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP Selection

- BGP is a path vector protocol. In path vector routing, the path is expressed in terms of the domains (or confederations) traversed so far.
- The best path is obtained by comparing the number of domains of each feasible route.
- There are no universally agreed-upon metrics that can be used to evaluate external paths.
- Each AS has its own set of criteria for path evaluation.

Path attributes

- Path attributes are used to describe and evaluate a route.
- Peers exchange path attributes along with other routing information.
- When a BGP router advertises a route, it can add or modify the path attributes before advertising the route to a peer.
- The combination of attributes are used to select the best path.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP Path Attribute

Four path attribute categories:

- Well-known mandatory: The attribute must be recognized by all BGP implementations. It must be sent in every UPDATE message.
- Well-known discretionary: The attribute must be recognized by all BGP implementations. However, it is not required to be sent in every UPDATE message.
- Optional transitive: It is not required that every BGP implementation recognize this type of attribute. A path with an unrecognized optional transitive attribute is accepted and simply forwarded to other BGP peers.
- Optional non-transitive: It is not required that every BGP implementation recognize this type of attribute. These attributes can be ignored and not passed along to other BGP peers.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

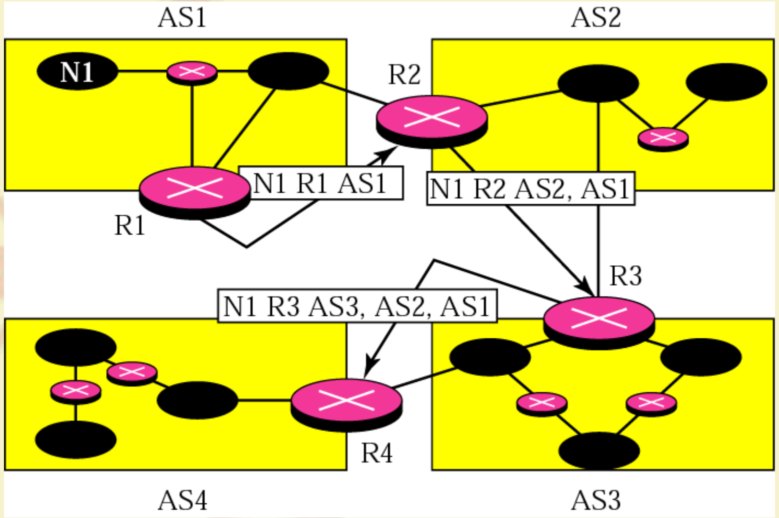
BGP

Example of Network Reachability

Network	Next router	Path
N1	R1	AS14,AS23,AS67
N2	R5	AS22,AS67,AS5,AS89
N3	R6	AS67,AS89,AS9,AS34
N4	R12	AS62,AS2,AS9

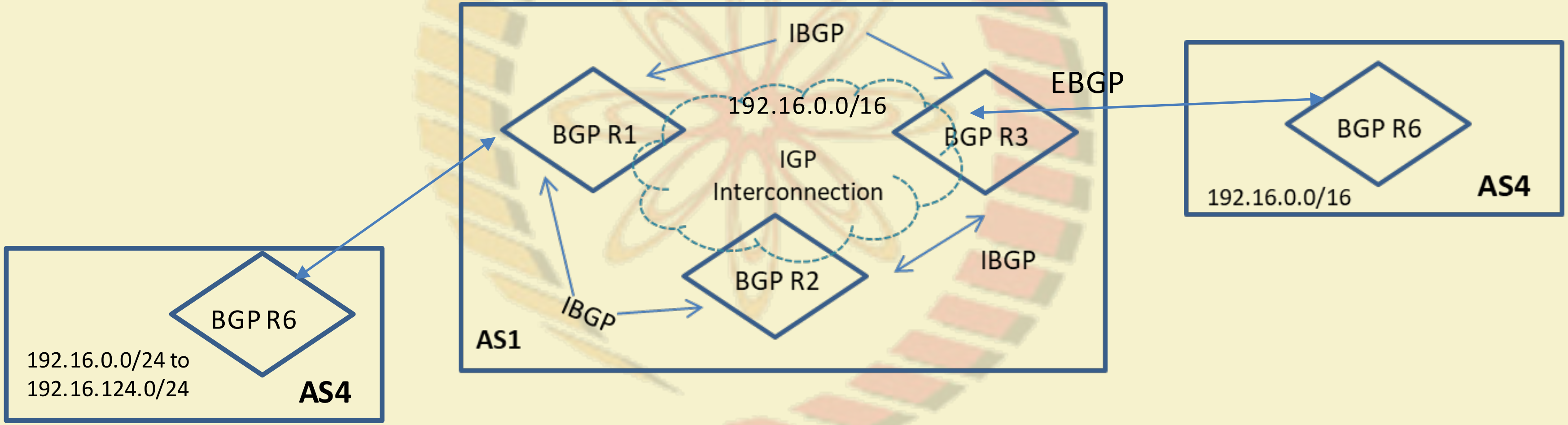
- Loop Prevention in BGP:
  - Checks the Path before updating its database. (If its AS is in the path ignore the message)
- Policy Routing:
  - If a path consist of an AS against the policy of the current AS, message discarded.

Message Advertisements



BGP Aggregation

- The major improvement introduced in BGP Version 4 was support for CIDR and route aggregation.
- This feature allows BGP peers to consolidate multiple contiguous routing entries into a single advertisement.
- It significantly enhances the scalability of BGP into large internetworking environments.



Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

BGP Confederations

- BGP requires that all speakers within a single AS have a fully meshed set of IBGP connections.
- This can be a scaling problem in networks containing a large number of IBGP peers.
- A BGP confederation creates a set of autonomous systems that represent a single AS to peers external to the confederation.
- This removes the full mesh requirement and reduces management complexity.

Ref: TCP/IP Tutorials and Technical Overview, IBM Redbooks

thank you!