CS348: Computer Networks

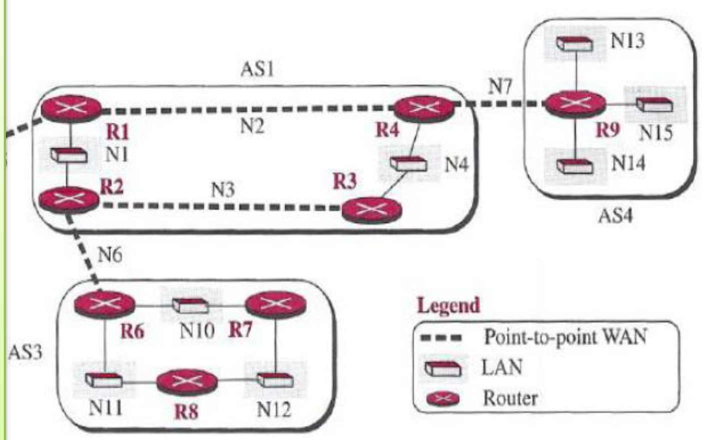
Routing in the Internet

Dr. Manas Khatua
Assistant Professor
Dept. of CSE, IIT Guwahati
E-mail: manaskhatua@iitg.ac.in

Introduction

- So far we've viewed the network simply as a collection of interconnected routers.
- All routers executed the same routing algorithm to compute routing paths through the entire network.
- In practice, this model is too simplistic and not efficient!
  - *Scalability:*
    - Today's public Internet consists of hundreds of millions of hosts.
    - As the number of routers becomes large, the overhead involved in computing, storing, and communicating routing information becomes prohibitive.
    - Sometimes, routing algorithms would surely never converge!
  - *Administrative autonomy:*
    - company's desire to run whatever routing algorithm it chooses in its routers

- Solution::
  - organizing routers into **autonomous systems** (**ASs**)
  - each AS consisting of a group of routers that are typically under the same administrative control (e.g. ISP)
  - typically all routers in an AS run the same routing algorithm
  - to connect ASs to each other, one or more routers in an AS will have the added task; these routers are called gateway routers.



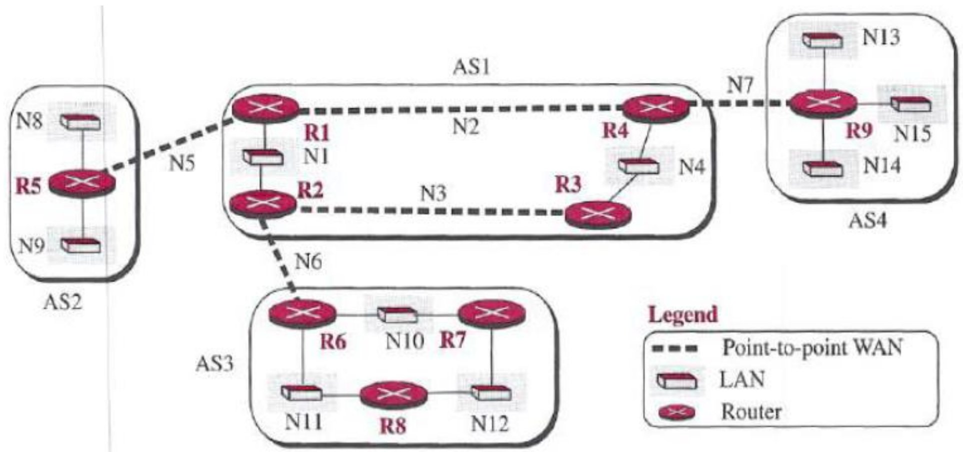
Cont…

- The routing algorithm running within an autonomous system (AS) is called an **intra-AS routing protocol** (e.g. RIP, OSPF).

- How does a router, within some AS, know how to route a packet to a destination that is outside the AS?

- Solution (two cases):
  - Each AS has only one gateway router:
    - Intra-AS routing algorithm has determined the least-cost path from each internal router to the gateway router
    - Upon receiving the packet, gateway router forwards the packet on the one link that leads outside the AS.
    - The AS on the other side of the link then takes over the responsibility

- Each AS has multiple gateway routers:

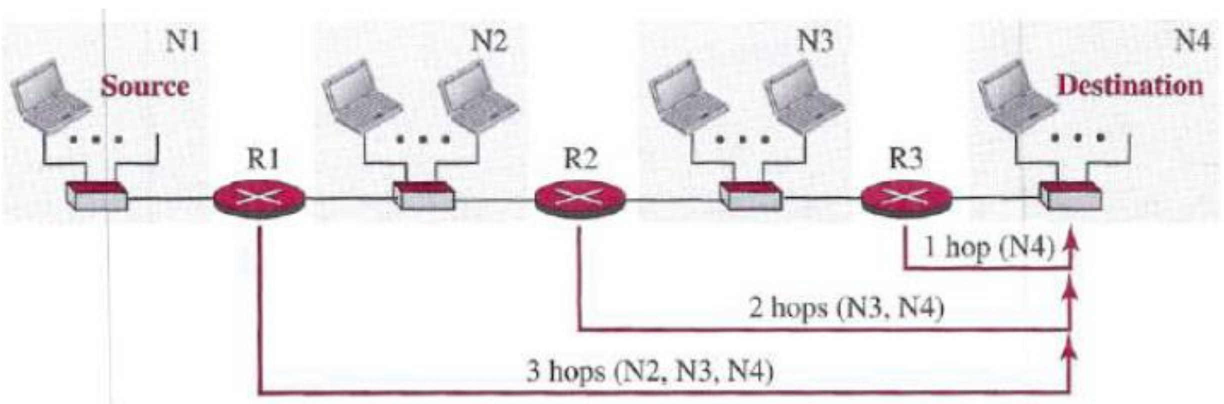


- To solve this problem, AS1 needs
  - (1) to learn which destinations are reachable via AS2/AS3/AS4, and
  - (2) to propagate this reachability information to all the routers within AS1, so that each router can configure its forwarding table to handle external-AS destinations.

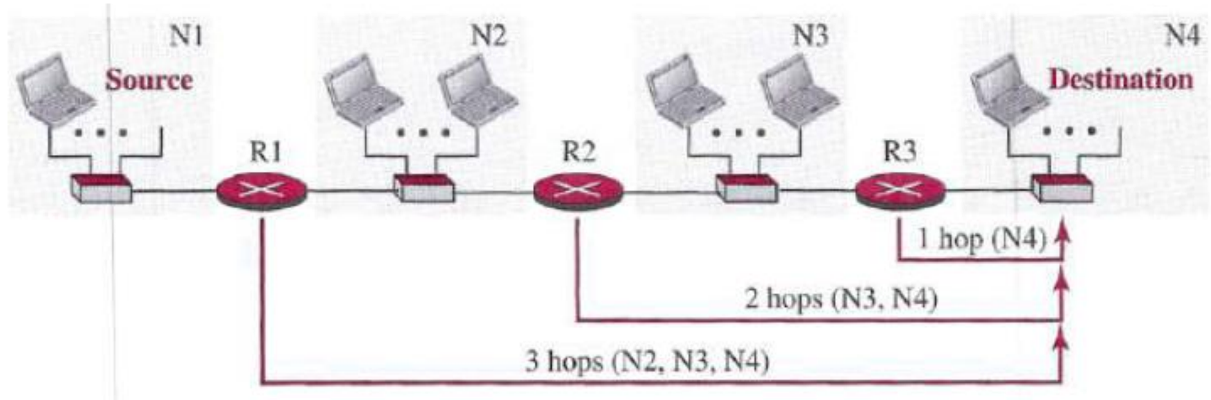
- These two tasks are handled by **inter-AS routing protocols** (e.g. BGP)

Routing Information Protocol (RIP)

- RIP is an intra-AS routing protocol
- Intra-AS routing protocols are also known as interior gateway protocols
- Follow distance-vector routing with the following modifications:
  - Router advertise the cost to reach different networks instead of individual node
  - Cost is defined as number of **hops** which is the number of subnets traversed along the shortest path from source router to destination subnet, including the destination subnet
  - The maximum cost of a path is limited to 15
  - Routing/Forwarding table instead of distance-vector is exchanged between neighbors approximately every 30 sec.



Cont…



- Final Routing Tables for the above network

Forwarding table for R1

Destination network	Next router	Cost in hops
N1	—	1
N2	—	1
N3	R2	2
N4	R2	3

Forwarding table for R2

Destination network	Next router	Cost in hops
N1	R1	2
N2	—	1
N3	—	1
N4	R3	2

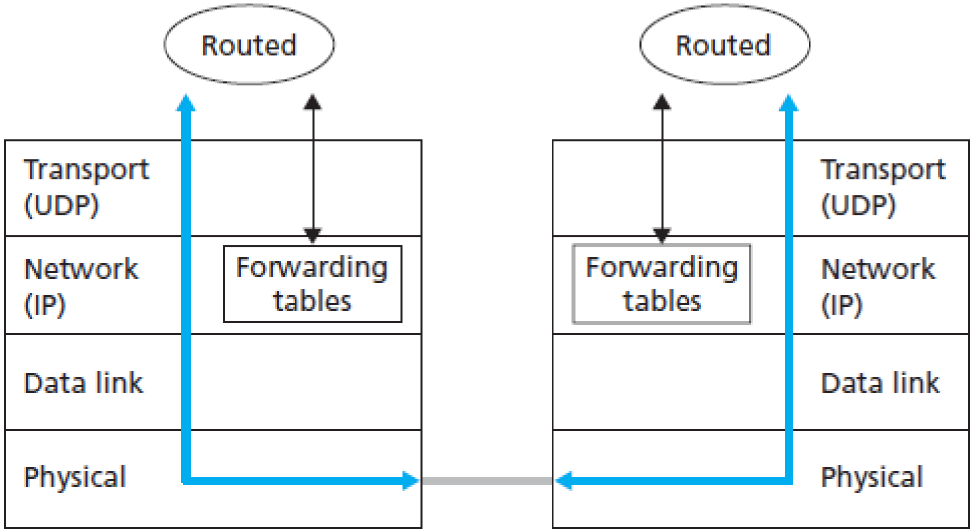
Forwarding table for R3

Destination network	Next router	Cost in hops
N1	R2	3
N2	R2	2
N3	—	1
N4	—	1

RIP Implementation

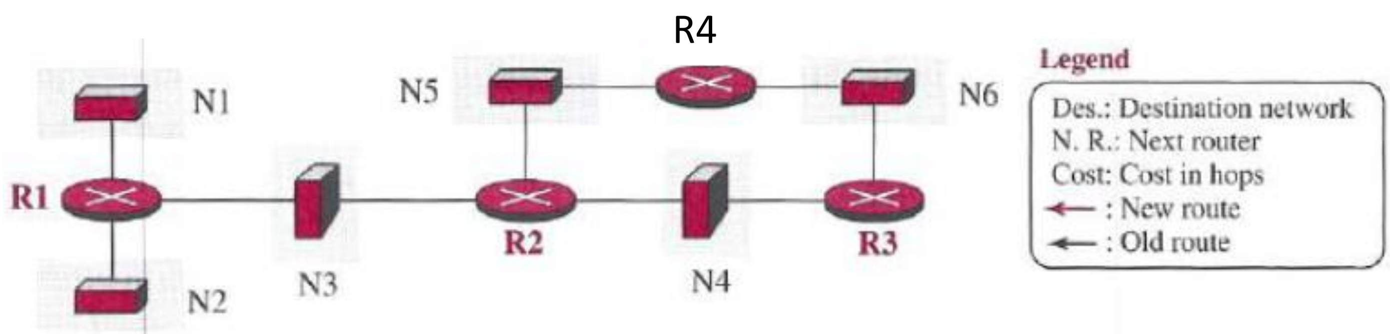
- RIP is implemented as an application layer process
- RIP uses transport layer protocol (UDP) on top of a network layer protocol (IP) to implement network layer functionality (a routing algorithm)
- Uses the service of UDP on port 520
- RIP uses two types of messages: request and response

Implementation of RIP as the *routed* daemon in a UNIX system



RIP Algorithm

- RIP uses timers:
  - Update timer: for advertising update message regularly. Default period 30 sec.
  - Expiration timer / Invalid timer : specifies how long a routing entry can be in the routing table without being updated. Default value 180 sec.
  - Hold-down Timer: This allows the route to get stabilized. During this time no update can be done to that routing entry. Default value 180 sec.
  - Flush Timer: The flush timer controls the time between the route is invalidated or marked as unreachable and removal of entry from the routing table. Default time 240 sec.



R1

Des.	N. R.	Cost
N1	—	1
N2	—	1
N3	—	1

R2

Des.	N. R.	Cost
N3	—	1
N4	—	1
N5	—	1

R3

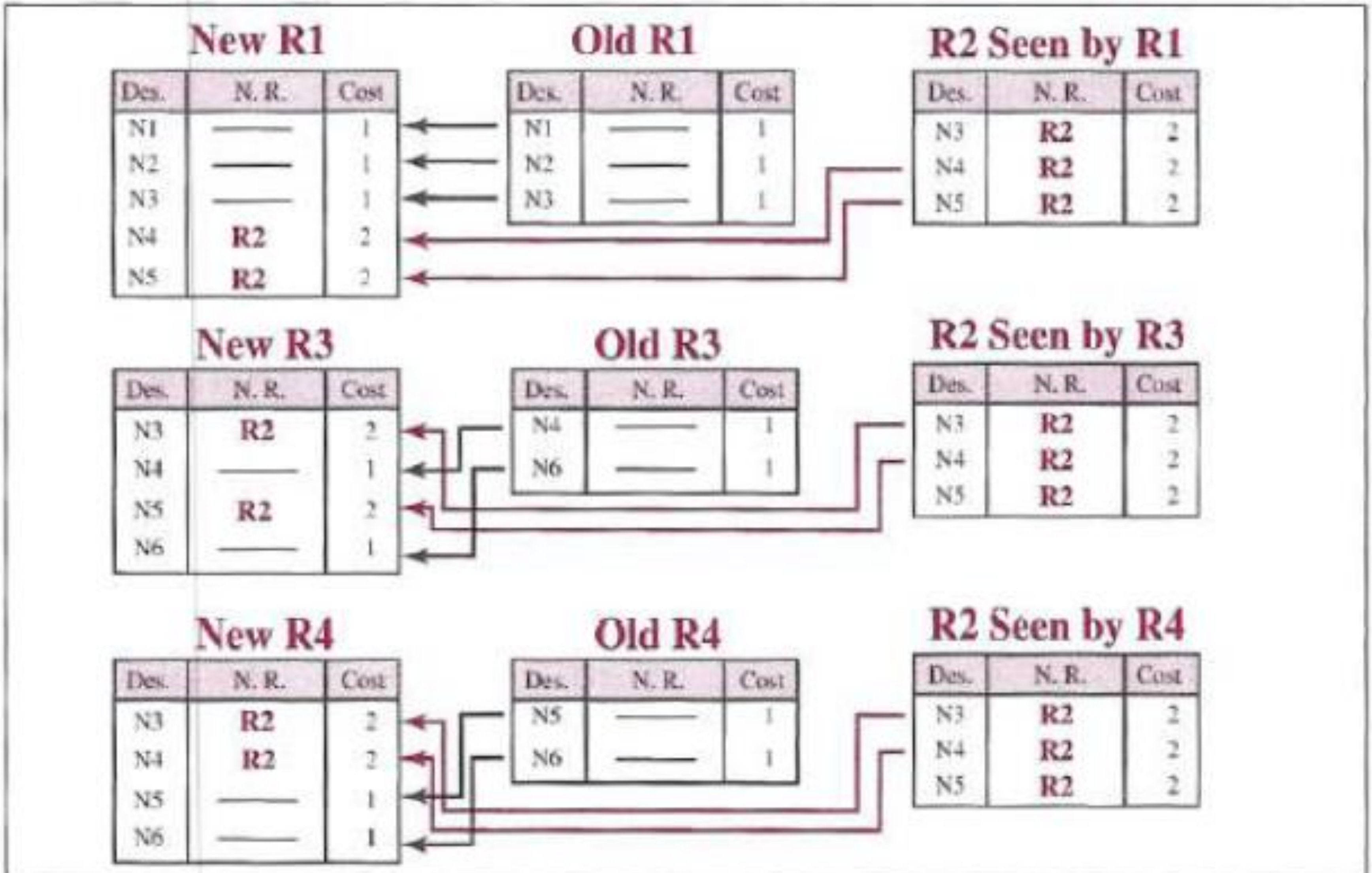
Des.	N. R.	Cost
N4	—	1
N6	—	1

R4

Des.	N. R.	Cost
N5	—	1
N6	—	1

Forwarding tables after all routers booted

Updates after R2 Advertisement



New R1

Des.	N. R.	Cost
N1	——	1
N2	——	1
N3	——	1
N4	R2	2
N5	R2	2

Old R1

Des.	N. R.	Cost
N1	——	1
N2	——	1
N3	——	1

R2 Seen by R1

Des.	N. R.	Cost
N3	R2	2
N4	R2	2
N5	R2	2

New R3

Des.	N. R.	Cost
N3	R2	2
N4	——	1
N5	R2	2
N6	——	1

Old R3

Des.	N. R.	Cost
N4	——	1
N6	——	1

R2 Seen by R3

Des.	N. R.	Cost
N3	R2	2
N4	R2	2
N5	R2	2

New R4

Des.	N. R.	Cost
N3	R2	2
N4	R2	2
N5	——	1
N6	——	1

Old R4

Des.	N. R.	Cost
N5	——	1
N6	——	1

R2 Seen by R4

Des.	N. R.	Cost
N3	R2	2
N4	R2	2
N5	R2	2

Open Shortest Path First (OSPF)

- OSPF is a link-state protocol that uses
  - flooding of link-state information
  - Dijkstra algorithm to find least-cost path
- Every router constructs a complete topological map of the entire AS
- Link cost are defined by the administrator
- OSPF is typically deployed in upper-tier ISPs whereas RIP is deployed in lower-tier ISPs and enterprise networks
- A router broadcasts link-state information whenever there is a change in a link's state
- It also broadcasts a link's state periodically at least once in every 30 minutes
- OSPF does not use a transport protocol (TCP/UDP) but encapsulates its data directly in IP packets, using protocol number 89.
- OSPF implements its own transport layer error detection and correction functions

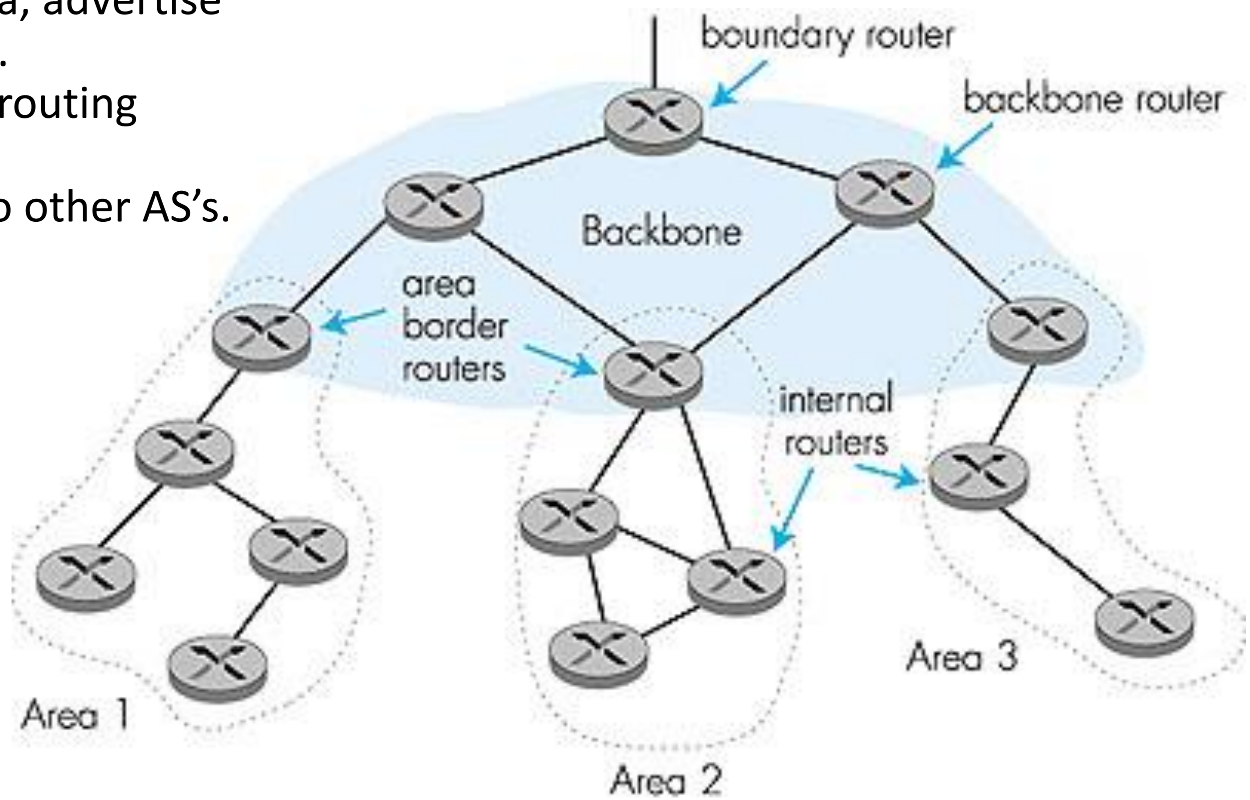
OSPF "advanced" features (not in RIP)

- Some of the advances embodied in OSPF:
  - *Security:* Exchanges between OSPF routers (for example, link-state updates) can be authenticated using MD5.
    - MD5 authentication is based on shared secret keys that are configured in all the routers.
  - *Multiple same-cost paths:* When multiple paths to a destination have the same cost, OSPF allows multiple paths to be used
  - *Integrated support for unicast and multicast routing:* Multicast OSPF (MOSPF) provides simple extensions to OSPF to provide for multicast routing
  - *Support for hierarchy within a single routing domain:* OSPF has the ability to structure an AS hierarchically into areas for a large domain.

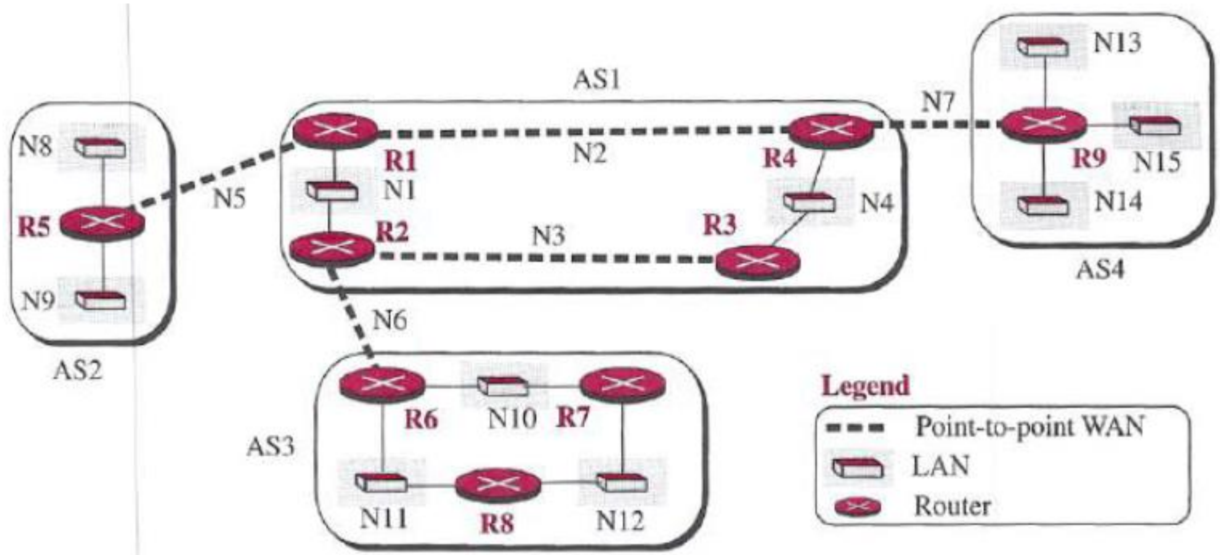
Hierarchical OSPF

- Two-level hierarchy: local area, backbone.
  - Link-state advertisements only in area
  - each node has detailed area topology;

- Area border routers: "summarize" distances to nets in own area, advertise to other Area Border routers.
- Backbone routers: run OSPF routing limited to backbone.
- Boundary routers: connect to other AS's.



Border Gateway Protocol (BGP)



Stub AS: connects to one other AS only

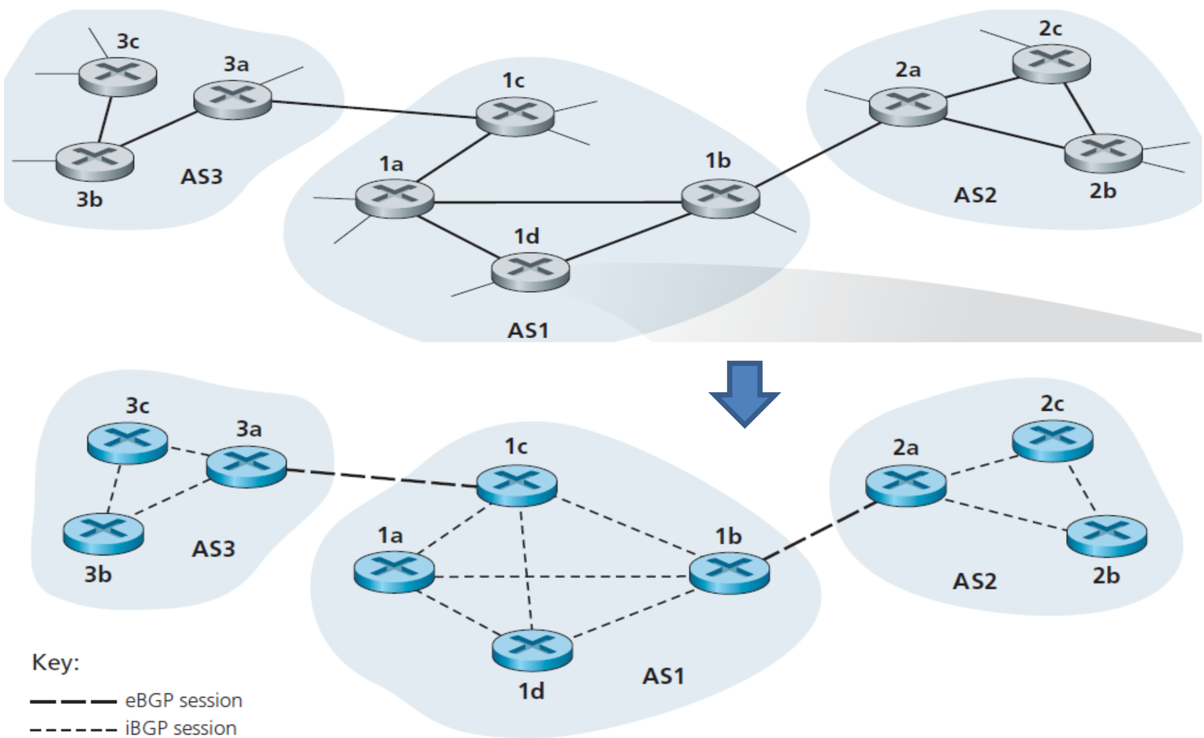
Multihomed AS: connects more than one AS, but refuses to carry transit traffic

Transit AS: connects more than one other AS, and carries local and transit traffic

- N1 - N14 are subnets in ASs
- AS2, AS3, AS4 : stub AS;   AS1 : transient AS
- Each AS uses RIP / OSPF for intra-domain routing
- All AS use BGP for inter-domain routing
  - eBGP: on each **border router**
  - iBGP: on **all routers**

Cont...

- BGP allows each subnet to advertise its existence to the rest of the Internet.
- BGP makes sure that all the ASs in the Internet know about a subnet and how to get there.
- BGP allows each AS to learn which destinations are reachable via its neighboring ASs.
- In BGP, pairs of routers exchange routing information over semipermanent TCP connections using port 179.

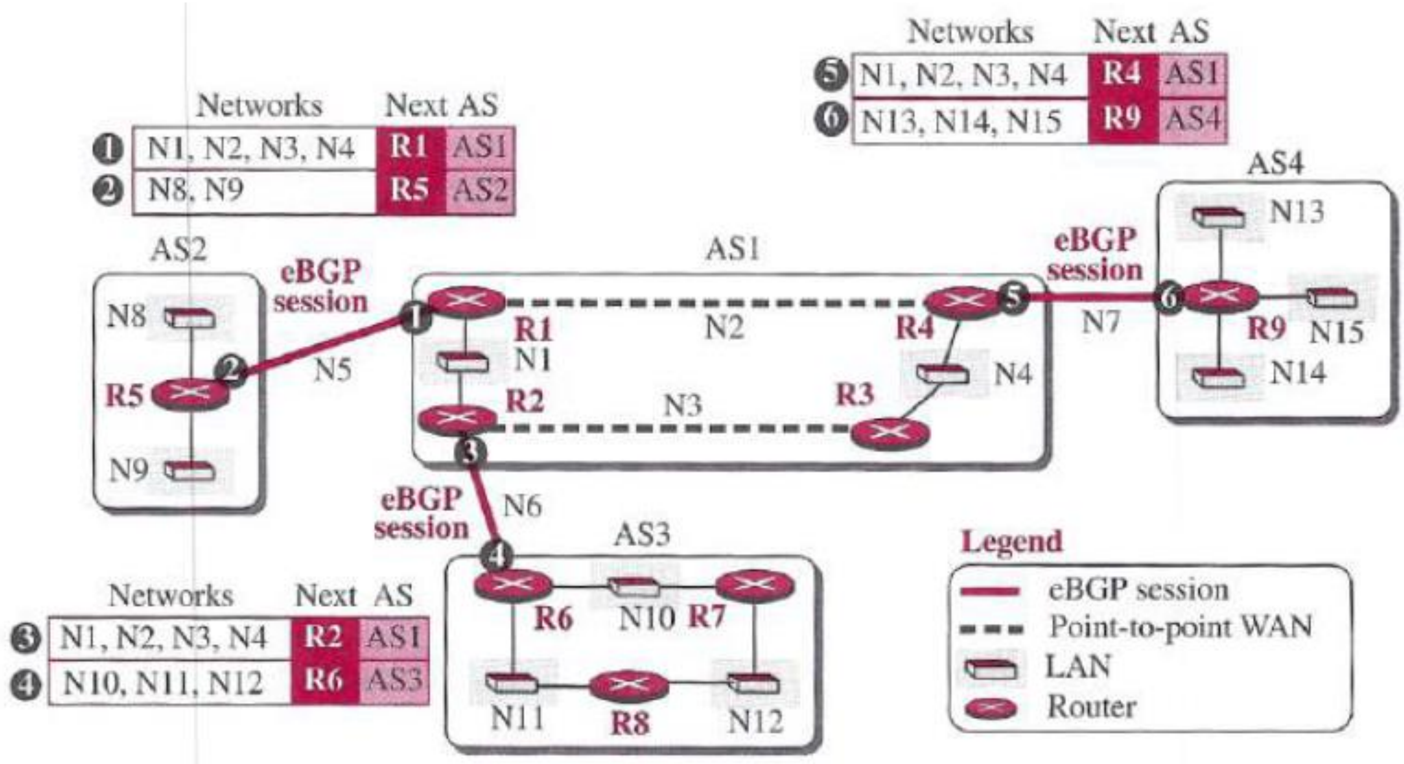


Two types of pairing:

- two routers in two different ASs
- two routers within an AS

The TCP connection along with all the BGP messages sent over the connection is called a **BGP session**

External BGP (eBGP)



- 3 pairs: R1-R5, R4-R9, R2-R6 => **3 eBGP sessions** (using TCP)
- Message 1 is sent by router R1 and tells router R5 that N1, N2, N3, and N4 can be reached through router R1. → Then, R5 updates its table.
- Message 2 ….. Message 3 ….. So on.

Cont…

- In BGP, destinations are not hosts but instead are CIDRized **prefixes (i.e. subnets)**

- **Example-1**:
  - Suppose there are four subnets attached to AS2
    - 138.16.64.0/24, 138.16.65.0/24, 138.16.66.0/24, and 138.16.67.0/24.
  - Then AS2 could aggregate the prefixes for these four subnets, and
  - use BGP to advertise the single prefix to 138.16.64.0/22 to AS1

- **Example-2**:
  - suppose that only the first three of those four subnets are in AS2 and the fourth subnet, 138.16.67.0/24, is in AS3.
  - AS3 could advertise to AS1 the more specific prefix 138.16.67.0/24, and
  - AS2 could still advertise to AS1 the aggregated prefix 138.16.64.0/22.

  - because routers use **longest-prefix matching** for forwarding datagrams, there will not be any ambiguity !

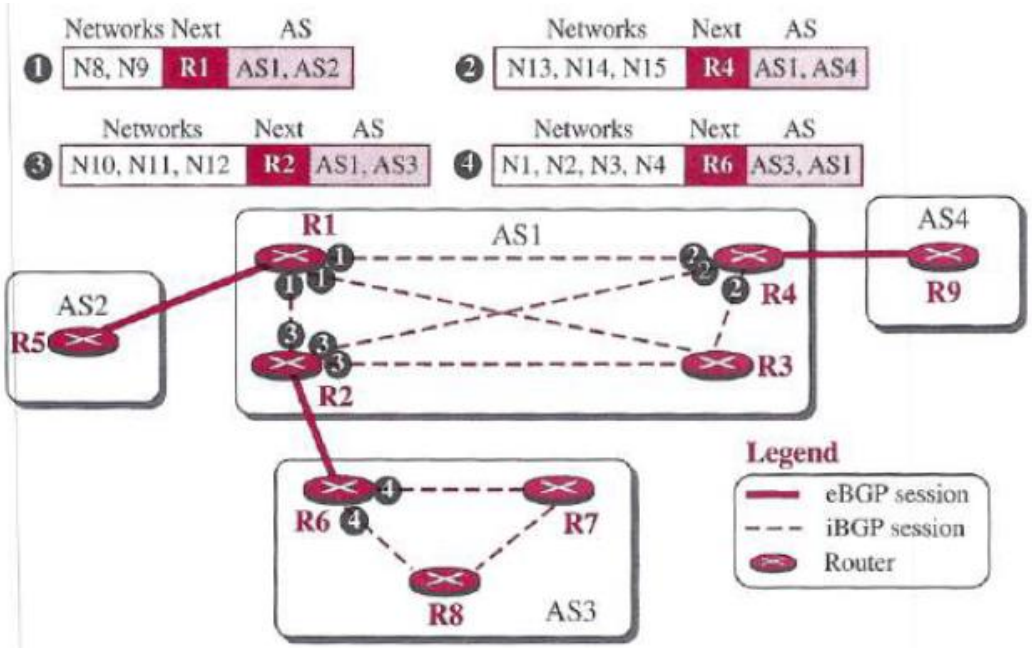
Limitation in eBGP

1. Border router does not know how to route a packet destined for non-neighbour AS.
   - E.g., R5 does not know about networks in AS3, AS4

2. None of the non-border routers know how to route a packet destined for any network in other ASs.
   - E.g., R3 does not know about networks in AS2, AS3, AS4

Solution: Internal BGP (iBGP)

Internal BGP (iBGP)



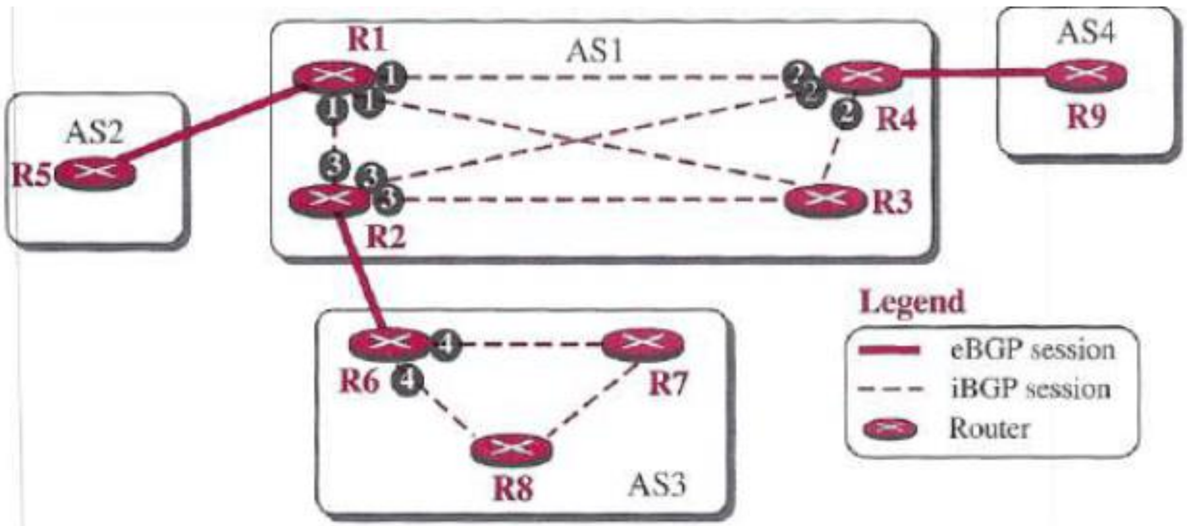
No message for R3, R7, R8

Note that BGP session lines do not always correspond to the physical links.

- For n router in an AS: n(n-1)/2 sessions
- No iBGP session for single router in an AS.

- Message 1 sent by R1 tells that N8 and N9 are reachable through the path AS1-AS2, but the next router is R1.
- Message 2 ….. Message 3 …… so on

Finalizing iBGP Path Table



- R1 receives:

Networks	Next	AS
N8, N9	R5	AS2

Networks	Next	AS
N10, N11, N12	R2	AS1, AS3

Networks	Next	AS
N13, N14, N15	R4	AS1, AS4

Networks	Next	Path
N8, N9	R5	AS1, AS2
N10, N11, N12	R2	AS1, AS3
N13, N14, N15	R4	AS1, AS4

Path table for R1

Cont…

Networks	Next	Path
N8, N9	R5	AS1, AS2
N10, N11, N12	R2	AS1, AS3
N13, N14, N15	R4	AS1, AS4

Path table for R1

Networks	Next	Path
N8, N9	R1	AS1, AS2
N10, N11, N12	R6	AS1, AS3
N13, N14, N15	R1	AS1, AS4

Path table for R2

Networks	Next	Path
N8, N9	R2	AS1, AS2
N10, N11, N12	R2	AS1, AS3
N13, N14, N15	R4	AS1, AS4

Path table for R3

Networks	Next	Path
N8, N9	R1	AS1, AS2
N10, N11, N12	R1	AS1, AS3
N13, N14, N15	R9	AS1, AS4

Path table for R4

Networks	Next	Path
N1, N2, N3, N4	R1	AS2, AS1
N10, N11, N12	R1	AS2, AS1, AS3
N13, N14, N15	R1	AS2, AS1, AS4

Path table for R5

Networks	Next	Path
N1, N2, N3, N4	R2	AS3, AS1
N8, N9	R2	AS3, AS1, AS2
N13, N14, N15	R2	AS3, AS1, AS4

Path table for R6

Networks	Next	Path
N1, N2, N3, N4	R6	AS3, AS1
N8, N9	R6	AS3, AS1, AS2
N13, N14, N15	R6	AS3, AS1, AS4

Path table for R7

Networks	Next	Path
N1, N2, N3, N4	R6	AS3, AS1
N8, N9	R6	AS3, AS1, AS2
N13, N14, N15	R6	AS3, AS1, AS4

Path table for R8

Networks	Next	Path
N1, N2, N3, N4	R4	AS4, AS1
N8, N9	R4	AS4, AS1, AS2
N10, N11, N12	R4	AS4, AS1, AS3

Path table for R9

BGP Attributes

- In RIP/OSPF, a destination has two associated information:
  - next hop and
  - cost

- But, inter-domain routing needs more information. In BGP, those information are called attributes

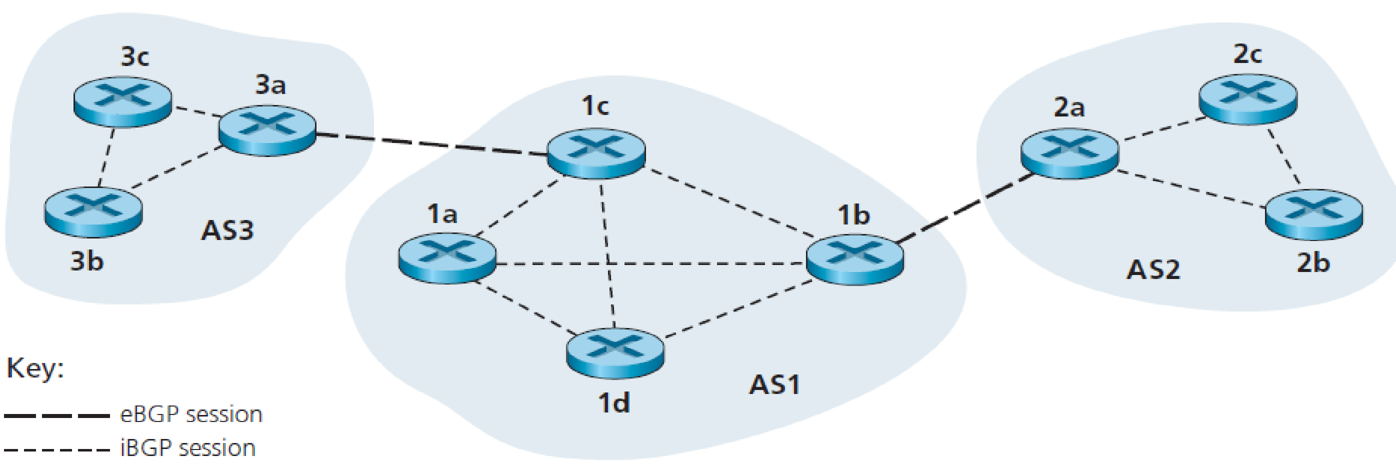
- BGP peers advertise routes to each other when they advertise prefix across BGP sessions

- Two important **attributes**:
  - AS-PATH
  - NEXT-HOP

- When a prefix is passed into an AS, the AS adds its ASN (unique autonomous number) to the **AS-PATH** attribute.
  - E.g., suppose that prefix is first advertised from AS2 to AS1; if AS1 then advertises the prefix to AS3, AS-PATH would be AS2 AS1.
  - Routers use the AS-PATH attribute to detect and prevent looping advertisements
  - Routers also use the AS-PATH attribute in choosing among multiple paths to the same prefix

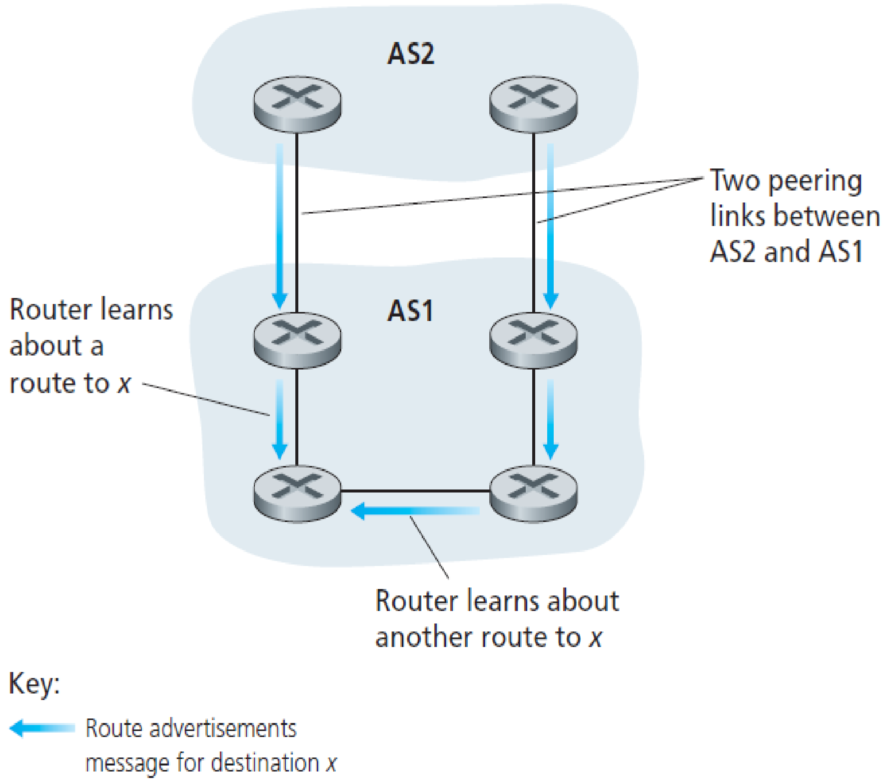
Cont…

- *The **NEXT-HOP** is the router interface that begins the AS-PATH.*
  - Providing the critical link between the inter-AS and intra-AS routing protocols, the NEXT-HOP attribute has a subtle but important use.
  - Example:
    - When the gateway router 3a in AS3 advertises a route to gateway router 1c in AS1 using eBGP, the route includes the NEXT-HOP, which is the IP address of the router 3a interface that leads to 1c.
    - Let router 1d learns about this route from iBGP.
    - Let router 1d may want to include the entry (*x*, *l*) in its forwarding table, where *l* is its interface that begins the least-cost path from 1d towards the gateway router 1c. To determine l, 1d provides the IP address in the NEXT-HOP attribute to its intra-AS routing module.



Cont…

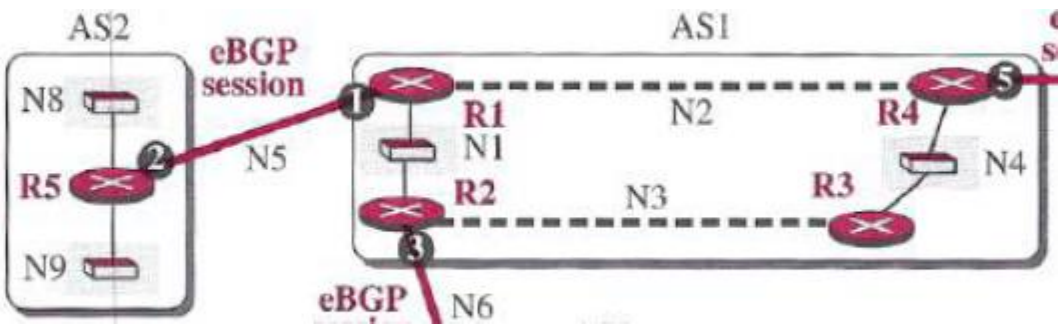
- In this figure, AS1 and AS2 are connected by two peering links.
- A router in AS1 could learn about two different routes to the same prefix *x*.
- These two routes could have the same AS-PATH to *x*, but could have different NEXT-HOP
- The router can determine the cost of the path to each peering link using the NEXT-HOP values and the intra-AS routing
- Apply hot-potato routing to determine the appropriate interface
  - The AS gets rid of the packet (the hot potato) as quickly as possible
  - This is done by having a router send the packet to the gateway router that has the smallest router-to-gateway cost among all gateways with a path to the destination.



Injection of Information into Forwarding Table

- The role of BGP is to help the routers inside the AS to augment their routing table
- Cost update problem for mixing of RIP and OSPF

- For stub AS:



Des.	Next	Cost
N8	—	1
N9	—	1

Table for R5

Networks	Next	Path
N1, N2, N3, N4	R1	AS2, AS1
N10, N11, N12	R1	AS2, AS1, AS3
N13, N14, N15	R1	AS2, AS1, AS4

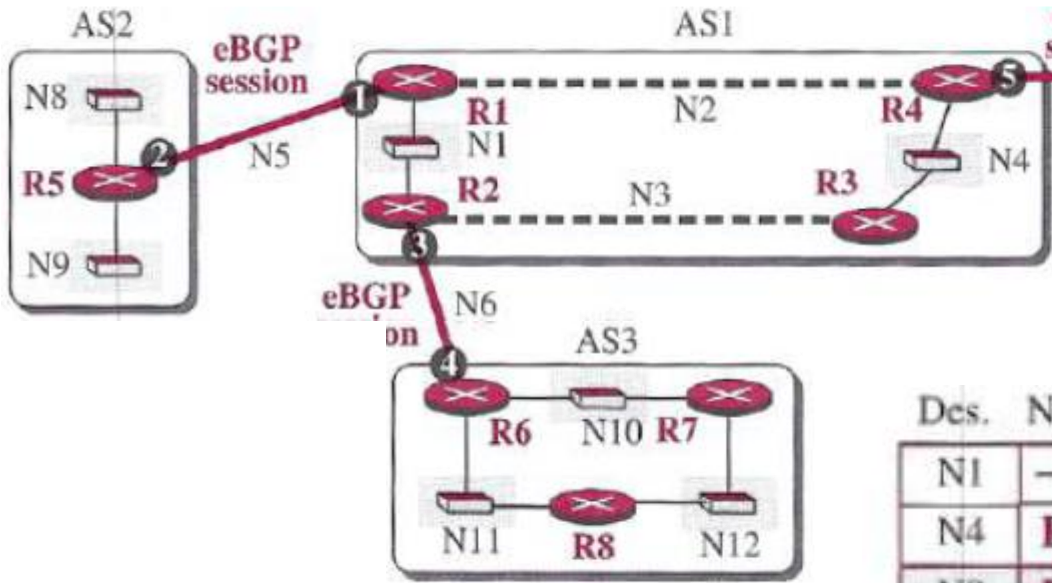
Path table for R5

Des.	Next	Cost
N8	—	1
N9	—	1
0	R1	1

Forwarding Table for R5

Cont…

For transient AS:



Des.	Next	Cost
N1	—	1
N4	R4	2

Table for R1

Networks	Next	Path
N8, N9	R5	AS1, AS2
N10, N11, N12	R2	AS1, AS3
N13, N14, N15	R4	AS1, AS4

Path table for R1

Des.	Next	Cost
N1	—	1
N4	R4	2
N8	R5	1
N9	R5	1
N10	R2	2
N11	R2	2
N12	R2	2
N13	R4	2
N14	R4	2
N15	R4	2

Forwarding Table for R1

BGP Route Selection

- BGP uses eBGP and iBGP to distribute routes to all the routers within ASs.
- A router may learn about more than one route to any one prefix, in which case the router must select one of the possible routes
- BGP sequentially invokes the following elimination rules until one route remains
  - local preference value: The local preference of a route could have been set by the router or could have been learned by another router
  - shortest AS-PATH: the route with the shortest AS-PATH is selected from the remaining routes (all with the same local preference value)
  - closest NEXT-HOP router: From the remaining routes the route with the closest NEXT-HOP router is selected.
    - closest means the router for which the cost of the least-cost path is the smallest
  - BGP identifiers: If more than one route still remains, the router uses BGP identifiers to select the route

Thanks!