The Network Layer: Routing

Slide Set were original prepared by Dr. Tatsuya Susa

Contents

7. Routing Algorithms

- An IMP executes a routing algorithm to decide which output line an incoming packet should be transmitted on
- In connection-oriented service, the routing algorithm is performed only during connection setup
- In connectionless service, the routing algorithm is performed as each packet arrives

Routing Algorithms *(cont'd)*

- Two types of routing algorithms:
  - Non-Adaptive Routing Algorithms
    - routes change slowly over time
  - Adaptive Routing Algorithms
    - routes change more quickly
      - periodic update
      - in response to link cost changes
- Hierarchical Routing is used to make these algorithms scale to large networks

7.1 Non-Adaptive Routing Algorithms

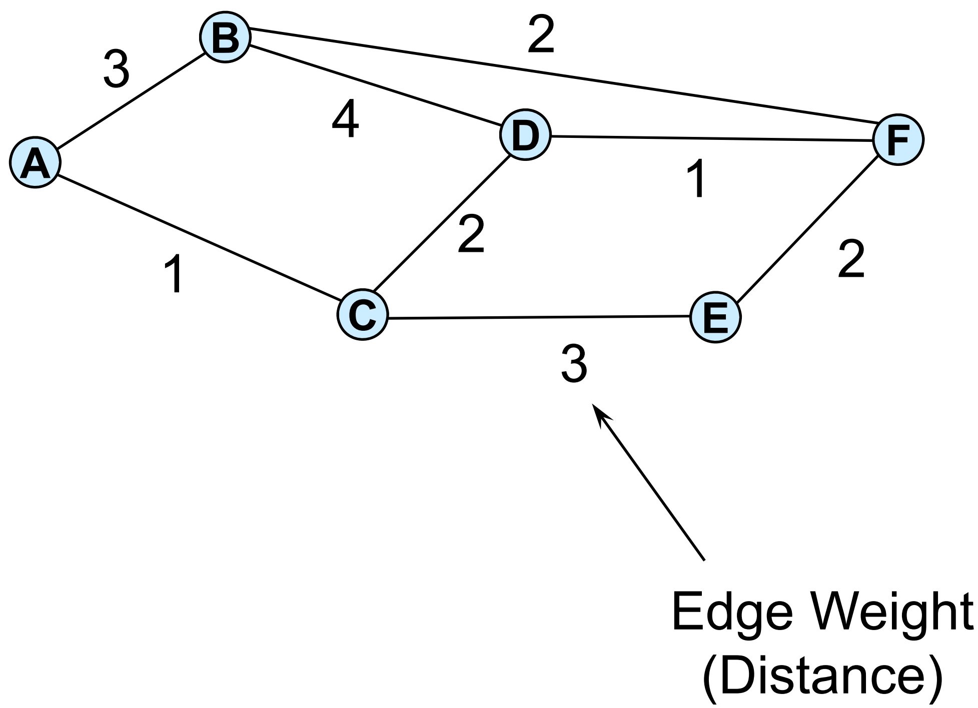
- Non-adaptive routing algorithms do not base their routing decisions on the current state of the network
- Examples:
  - Shortest Path Routing

7.1.1 Shortest Path Routing

- For a pair of communicating hosts, there is a shortest path between them
- Shortness may be defined by:
  - Number of IMP hops
  - Geographic distance
  - Link delay

Shortest Path

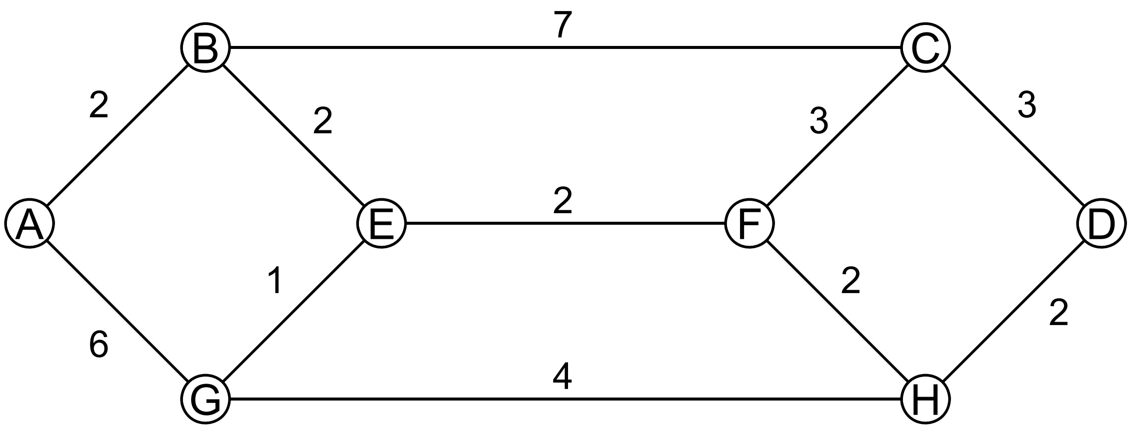
What is the shortest path between A and F?



Computing the Shortest Path

- Dijkstra's Shortest Path Algorithm:
  - Step 1: Draw nodes as circles. Fill in a circle to mark it as a "permanent node."
  - Step 2: Set the current node equal to the source node
  - Step 3: For the current node:
    - Mark the cumulative distance from the current node to each non-permanent adjacent node. Also mark the name of the current node. Erase this marking if the adjacent node already has a shorter cumulative distance marked
    - Mark the non-permanent node with the shortest listed cumulative distance as permanent and set the current node equal to it. Repeat step 3 until all nodes are marked permanent.

Dijskstra's Shortest Path Algorithm
Example



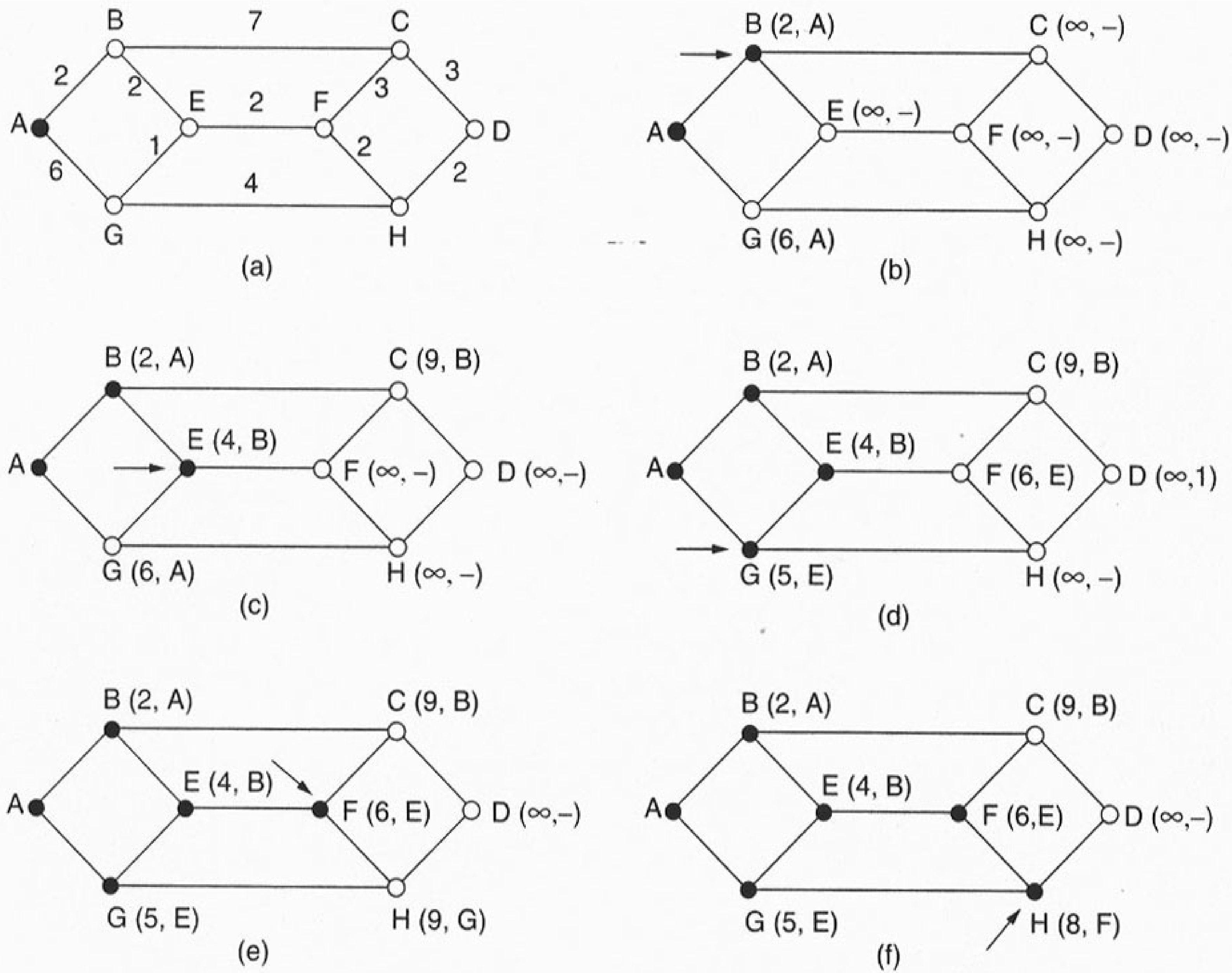


Fig. 5-6. The first five steps used in computing the shortest path from *A* to *D*. The arrows indicate the working node.

Shortest Path Routing *(cont'd)*

- Non-adaptive, if:
  - geographical distances are used as edge weights
  - maximum link throughputs are used as edge weights
  - Number of IMP hops are used as edge weights

7.2 Adaptive Routing Algorithms

- Problems with non-adaptive algorithms
  - If traffic levels in different parts of the subnet change dramatically and often, nonadaptive routing algorithms are unable to cope with these changes
  - Lots of computer traffic is bursty, but nonadaptive routing algorithms are usually based on average traffic conditions
- Adaptive routing algorithms can deal with these situations

Adaptive Routing Algorithms *(cont'd)*

- Two Types:
  - Centralized Adaptive Routing
    - one central routing controller
  - Distributed Adaptive Routing
    - routers periodically exchange information

- Distributed Adaptive Routing
  - routers periodically exchange information
  - Two types: Global or decentralized information?
    - Global
      - all routers have complete topology, link cost info
      - "link state" algorithms
    - Decentralized
      - router knows physically-connected neighbors, link costs to neighbors
      - iterative process of computation, exchange of info with neighbors
      - "distance vector" algorithms

7.2.1 Centralized Adaptive Routing

- Routing table adapts to network traffic
- A routing control center is somewhere in the network
- Periodically, each IMP forwards link status information to the control center
- The center can, with Dijkstra's shortest path algorithm, compute the best routes
- Best routes are dispatched to each IMP

Problem with Centralized Algorithms

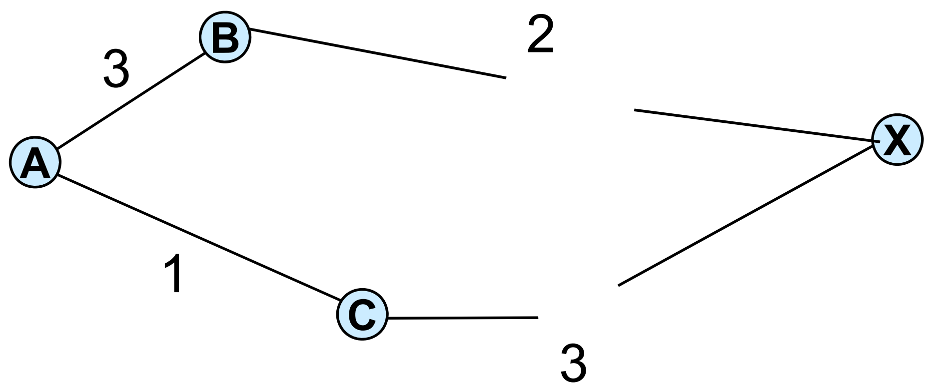
- Vulnerability
  - If the control center goes down, routing becomes nonadaptive
- Scalability
  - The control center must handle a great deal of routing information, especially for larger networks

7.2.2 Distributed Routing Algorithms

- Each IMP periodically exchanges routing information (e.g., estimated time delay, queue length, etc.) with its neighbors
- Examples:
  - Distance Vector Routing
    - original ARPA net routing scheme, often called RIP (route information protocol)
  - Link State Routing
    - base for the current Internet routing algorithm

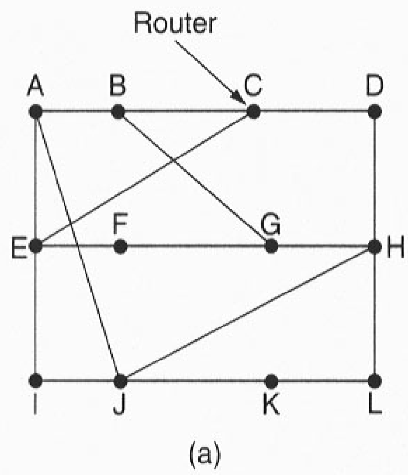
7.2.2.1 Distance Vectors

- Known as Bellman-Ford or Ford-Fullkerson algorithm
- Each IMP, or router, maintains lists of best-known distances to all other known routers. These lists are called “vectors.”
- Each router is assumed to know the exact distance (in delay,distance, etc.) to other routers directly connected to it.
- Periodically, vectors are exchanged between adjacent routers, and each router updates its vectors.



A -> X = min { (A -> B) + (B -> X), (A -> C) + (C -> X)}

Distance Vectors *(cont'd)*



To	A	I	H	K	New estimated delay from J	Line
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22	17	I
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	6	K
L	29	33	9	9	15	K
	JA delay is 8	JI delay is 10	JH delay is 12	JK delay is 6	New routing table for J	

Vectors received from J's four neighbors

(b)

Fig. 5-10. (a) A subnet. (b) Input from *A*, *I*, *H*, *K*, and the new routing table for *J*.

- Basic Idea
  - Each node periodically sends its own distance vector estimate to neighbors
  - When a node x receives new DV estimate from neighbor, it updates its own DV

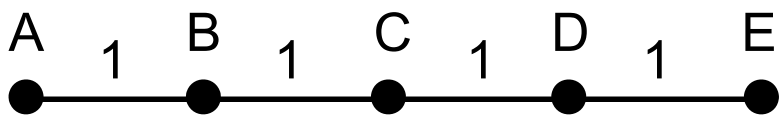
- Iterative, asynchronous
  - each local iteration caused by:
    - local link cost change
    - Distance vector update message from neighbor
- Distributed
  - each node notifies neighbors *only* when its distance vector changes
    - neighbors then notify their neighbors if necessary

Problem: Count-to-Infinity

- With distance vector routing, good news travels fast, but bad news travels slowly
- When a router goes down, it can take a really long time before all the other routers become aware of it

- In the following two examples, distance is measured in hops.

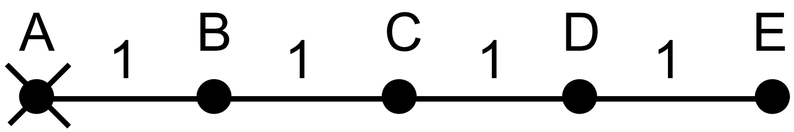
Count-to-Infinity



Infinity	Infinity	infinity	infinity	Initially (A is down)
				A comes up
1	infinity	infinity	infinity	After 1 exchange
1	2	infinity	infinity	After 2 exchanges
1	2	3	infinity	After 3 exchanges
1	2	3	4	After 4 exchanges

Good news travels fast.

Count-to-Infinity



1	2	3	4	Initially
				A goes down
3	2	3	4	After 1 exchange
3	4	3	4	After 2 exchanges
5	4	5	4	After 3 exchanges
5	6	5	5	After 4 exchanges
7	6	7	6	After 5 exchanges

etc… to infinity; bad news travels slow.

7.2.2.2 Link State Routing

- Each router measures the distance (in delay, hop count, etc.) between itself and its adjacent routers
- The router builds a packet containing all these distances. The packet also contains a sequence number and an age field.
- Each router distributes these packets using flooding

Link State Routing *(cont'd)*

- To control flooding, the sequence numbers are used by routers to discard flood packets they have already seen from a given router
- The age field in the packet is an expiration date. It specifies how long the information in the packet is good for.
- Once a router receives all the link state packets from the network, it can reconstruct the complete topology and compute a shortest path between itself and any other node using Dijsktra's algorithm.

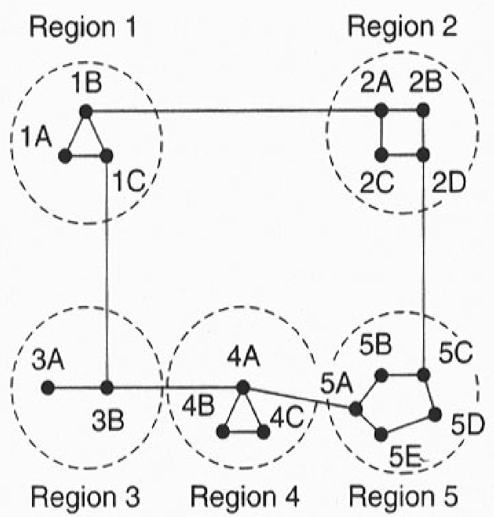
7.3 Hierarchical Routing

- All routing algorithms have difficulties as the network becomes large
- For large networks, the routing tables grow very quickly, and so does the number of flood packets
- How can this be reduced?
  - Hierarchical routing

Hierarchical Routing *(cont'd)*

- Segment the network into regions
- Routers in a single region know all the details about other routers in that region, but none of the details about routers in other regions
- Analogy: Telephone area codes

Hierarchical Routing *(cont'd)*



(a)

Full table for 1A

Dest.	Line	Hops
1A	–	–
1B	1B	1
1C	1C	1
2A	1B	2
2B	1B	3
2C	1B	3
2D	1B	4
3A	1C	3
3B	1C	2
4A	1C	3
4B	1C	4
4C	1C	4
5A	1C	4
5B	1C	5
5C	1B	5
5D	1C	6
5E	1C	5

(b)

Hierarchical table for 1A

Dest.	Line	Hops
1A	–	–
1B	1B	1
1C	1C	1
2	1B	2
3	1C	2
4	1C	3
5	1C	4

(c)

Fig. 5-17. Hierarchical routing.

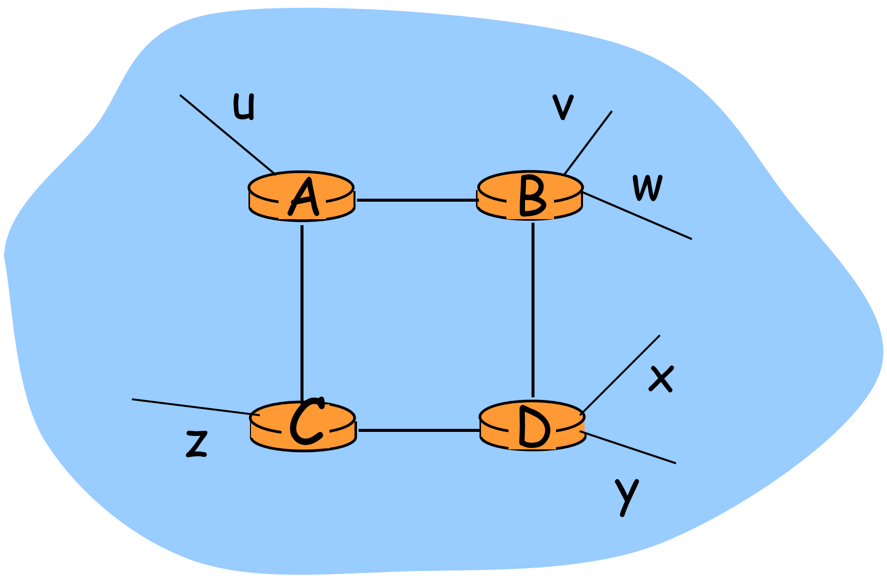
7.4 Routing in the Internet

- RIP (Route Information Protocol)
- OSPF (Open Shortest Path First)
- BGP (Border Gateway Protocol)

7.4.1 RIP (Routing Information Protocol)

- RIP
  - Route Information Protocol
  - One of the routing algorithms used by the Internet
  - Based on distance vector routing
  - Did not scale well, and it suffered the count-to-infinity problem
  - RIP is slowly being phased out

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)



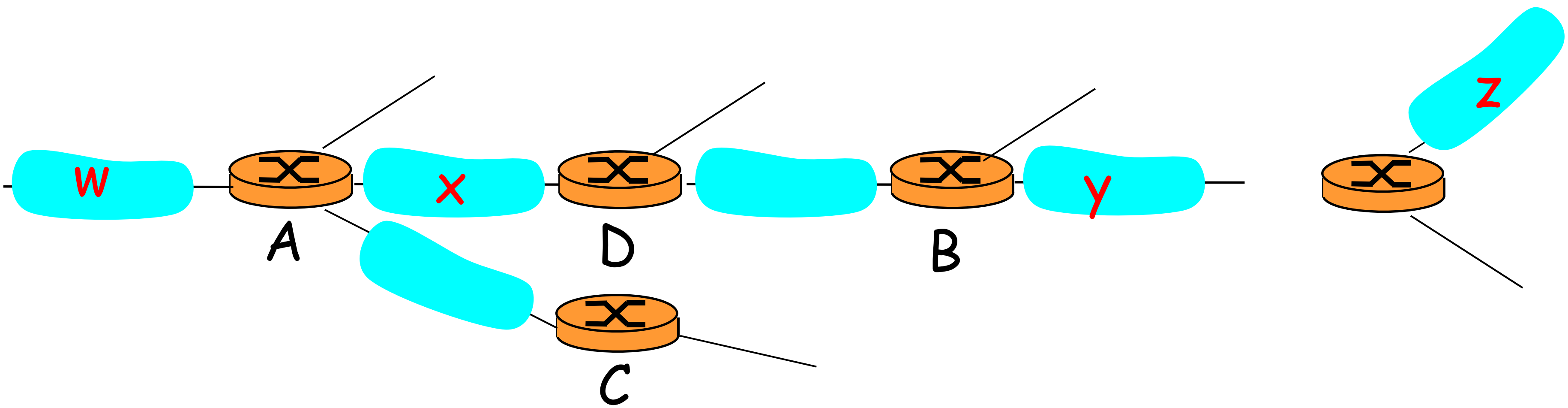
From router A to subsets:

destination	hops
u	1
v	2
w	2
x	3
y	3
z	2

RIP Advertisements*

- Distance vectors
  - exchanged among neighbors every 30 sec via Response Message (also called advertisement)
- Each advertisement
  - A list of up to 25 destination nets within AS

An Example

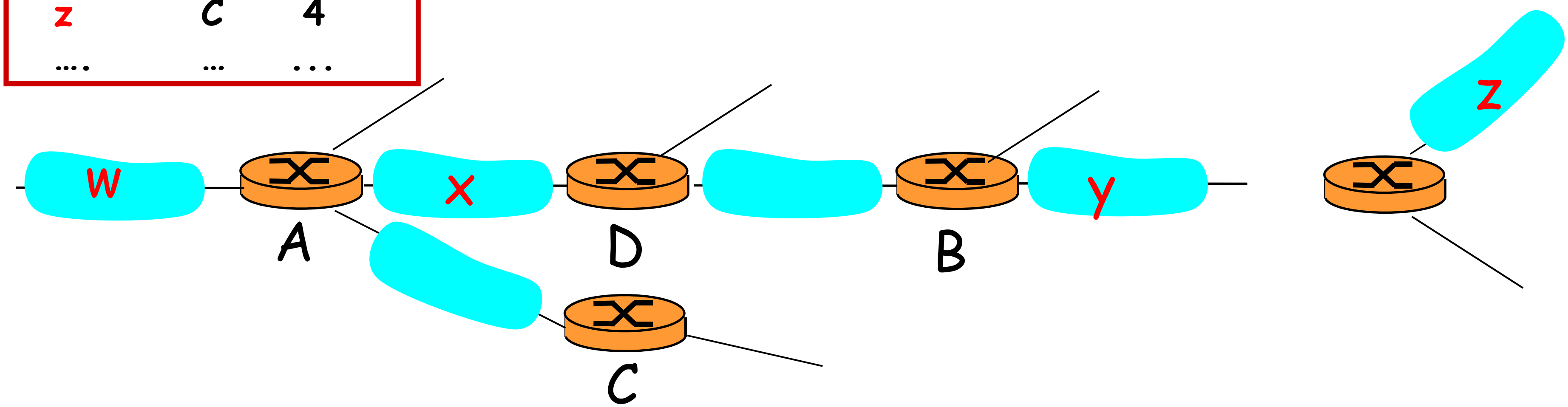


Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	B	7
x	--	1
….	….	….

Routing table in D

Advertisement from A to D

Dest	Next	hops
w	-	1
x	-	1
z	C	4
….	…	...



Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	~~B~~ A	~~7~~ 5
x	--	1
….	….	....

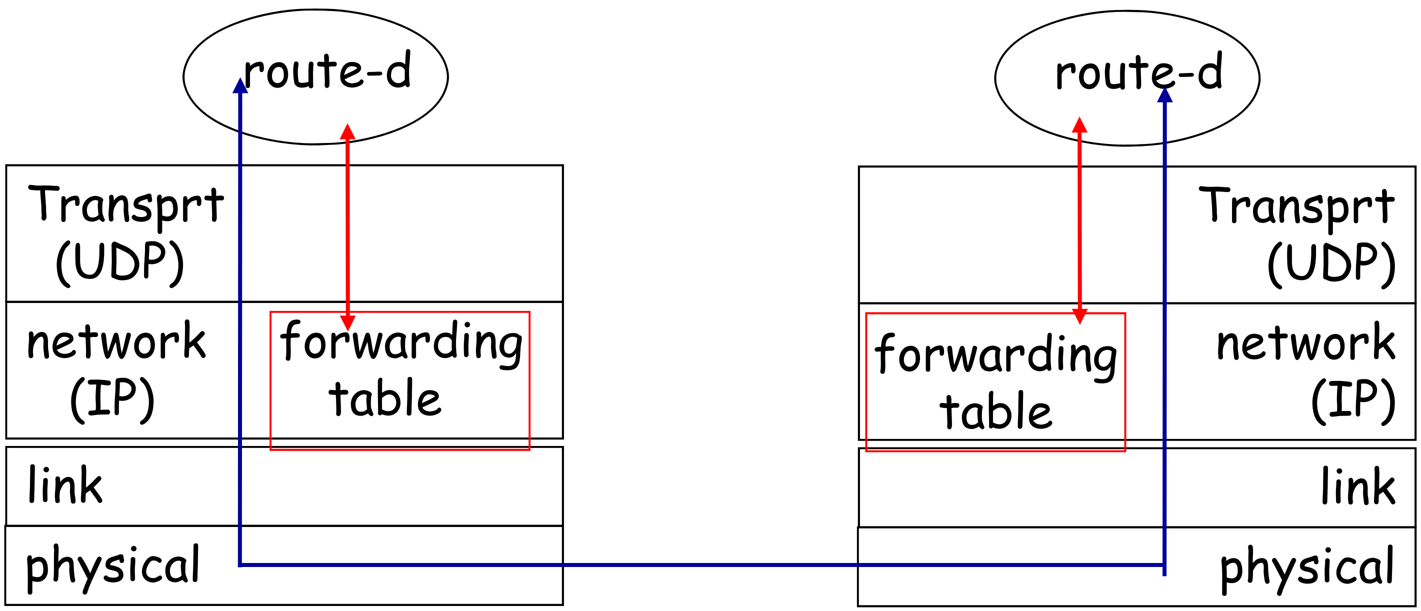
Routing table in D

RIP: Link Failure and Recovery

- If a node X does not hear advertisement from neighbor Y for 180 sec
  - Then, node X declares neighbor Y as dead, and
    - X sends new advertisements to its neighbors
    - neighbors in turn send out new advertisements (if tables changed)

RIP Table Processing*

- RIP routing tables managed by **application-level** process called route-d (daemon)
- Advertisements sent in UDP packets, periodically repeated



7.4.2 OSPF (Open Shortest Path First)

- Open Shortest Path First
- Routing algorithm now used in the Internet

OSPF "Advanced" Features (not in RIP)

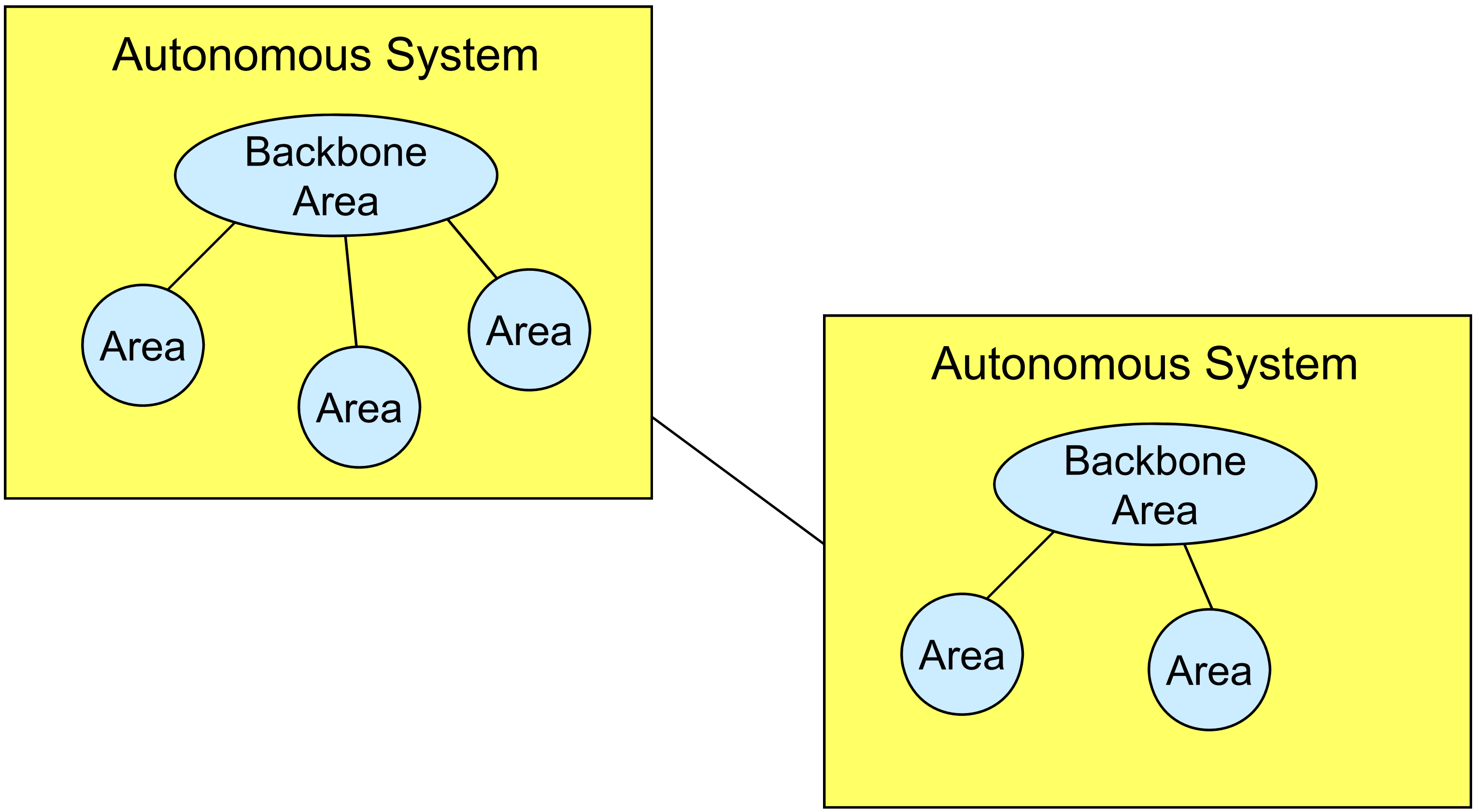
- OSPF uses the Link State Routing algorithm with modifications to support:
  - Multiple distance metrics (geographical distance, delay, throughput)
    - For each link, multiple cost metrics for different types of service (e.g., satellite link cost set "low" for best effort; high for real time)
  - Support for real-time traffic
  - Multiple same-cost paths allowed (only one path in RIP)

- Security
  - all OSPF messages authenticated (to prevent malicious intrusion)
- Hierarchical routing
  - Hierarchical OSPF in large domains.

OSPF: Hierarchical Routing

- OSPF divides the network into several hierarchies:
  - Autonomous Systems (AS’s)
    - groups of subnets
  - Areas
    - Groups of routers within an AS
  - Backbone Areas
    - Groups of routers that connect other areas together

OSPF *(cont'd)*

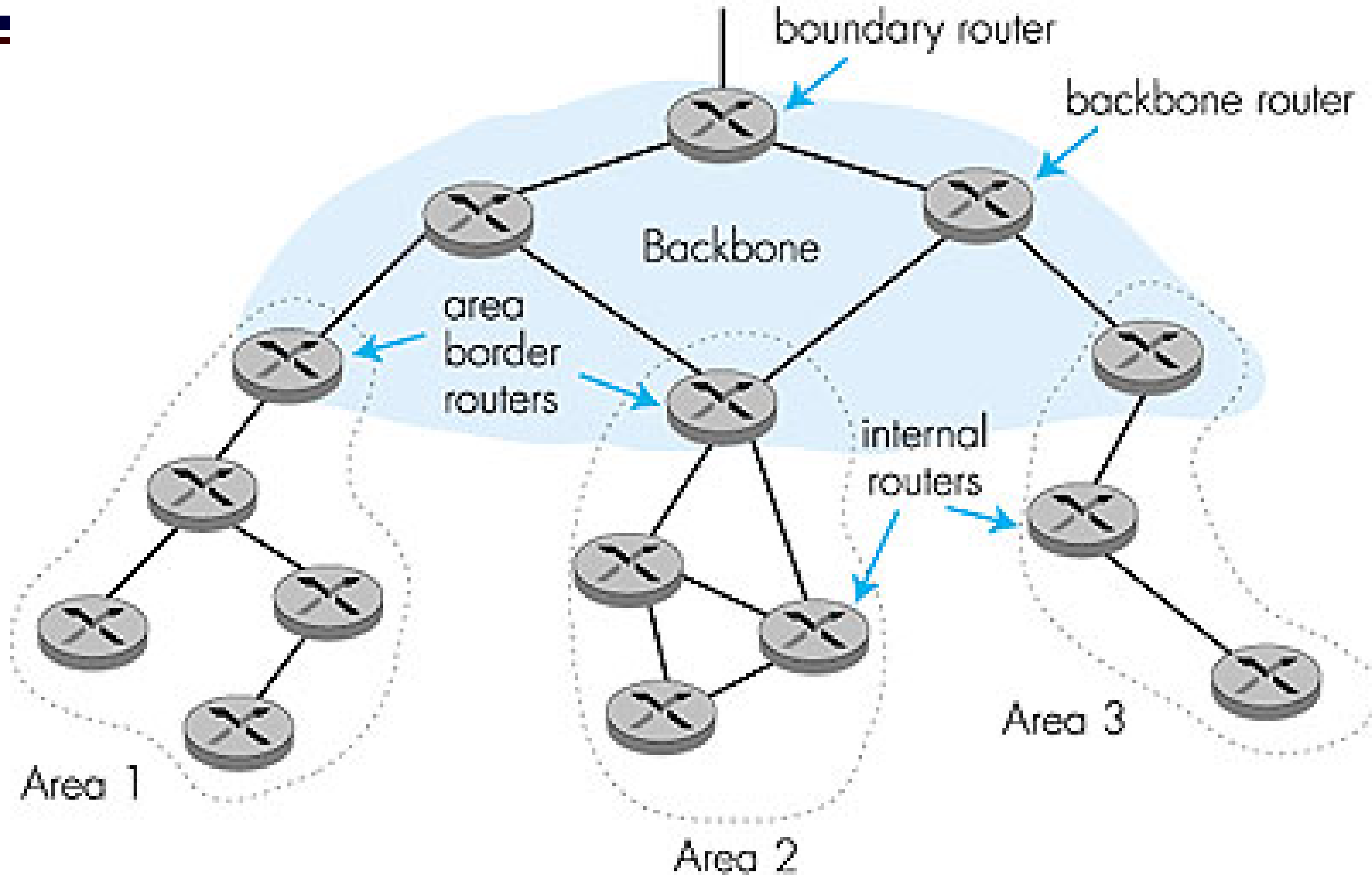


Autonomous System
Backbone Area
Area
Area
Area
Autonomous System
Backbone Area
Area
Area

OSPF *(cont'd)*

- Routers are distinguished by the functions they perform
  - Internal routers
    - Only route packets within one area
  - Area border routers *
    - Connect areas together
  - Backbone routers
    - Reside only in the backbone area
  - AS boundary routers
    - Routers that connect to a router outside the AS

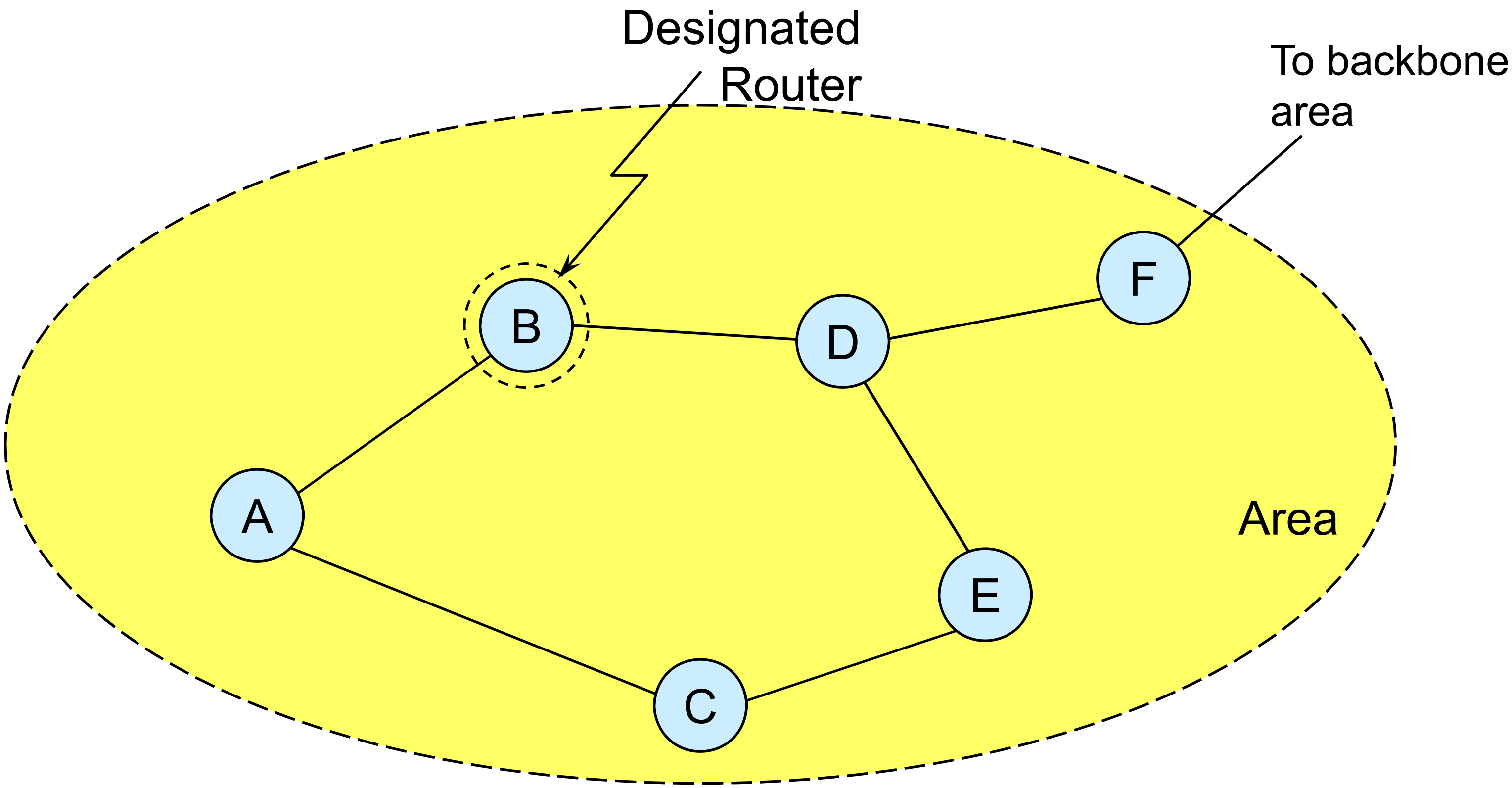
Hierarchical OSPF



OSPF: Modified Link State Routing

- Recall:
  - In link state routing, routers flood their routing information to all other routers in the network
- In OSPF, routers only send their information to “adjacent routers”, not to all routers.
- Adjacent does NOT mean nearest-neighbor in OSPF
- One router in each area is marked as the “designated router”
- Designated routers are considered adjacent to all other routers in the area
- OSPF combines link state routing with centralized adaptive routing

OSPF: Adjacency



Example:
C is “adjacent” to B but not to A or E
B is “adjacent” to all routers in the area

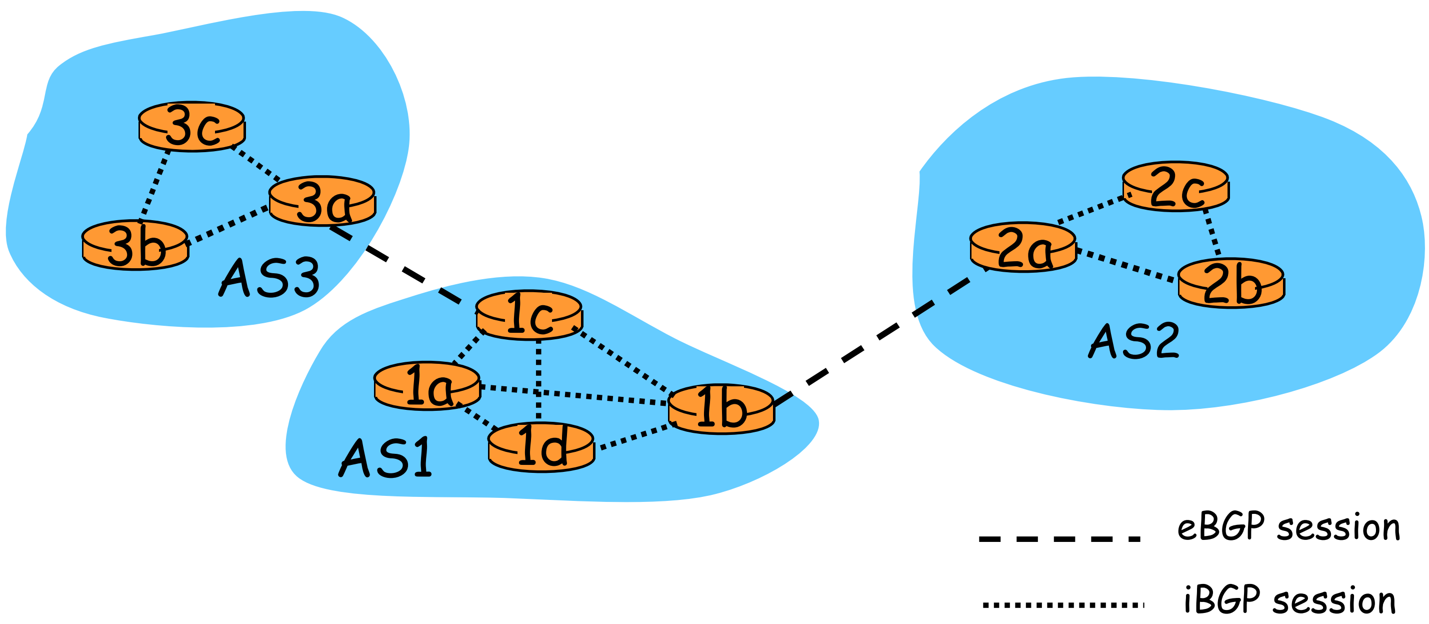
7.4.3 BGP (Border Gateway Protocol)

- BGP (Border Gateway Protocol)
  - the de facto standard
  - Internet inter-AS routing: BGP
- BGP provides each AS a means to:
  - Obtain subnet reachability information from neighboring ASs.
  - Propagate reachability information to all AS-internal routers.
  - Determine "good" routes to subnets based on reachability information and policy.
- BGP allows subnet to advertise its existence to rest of Internet: *"I am here"*

BGP Basics

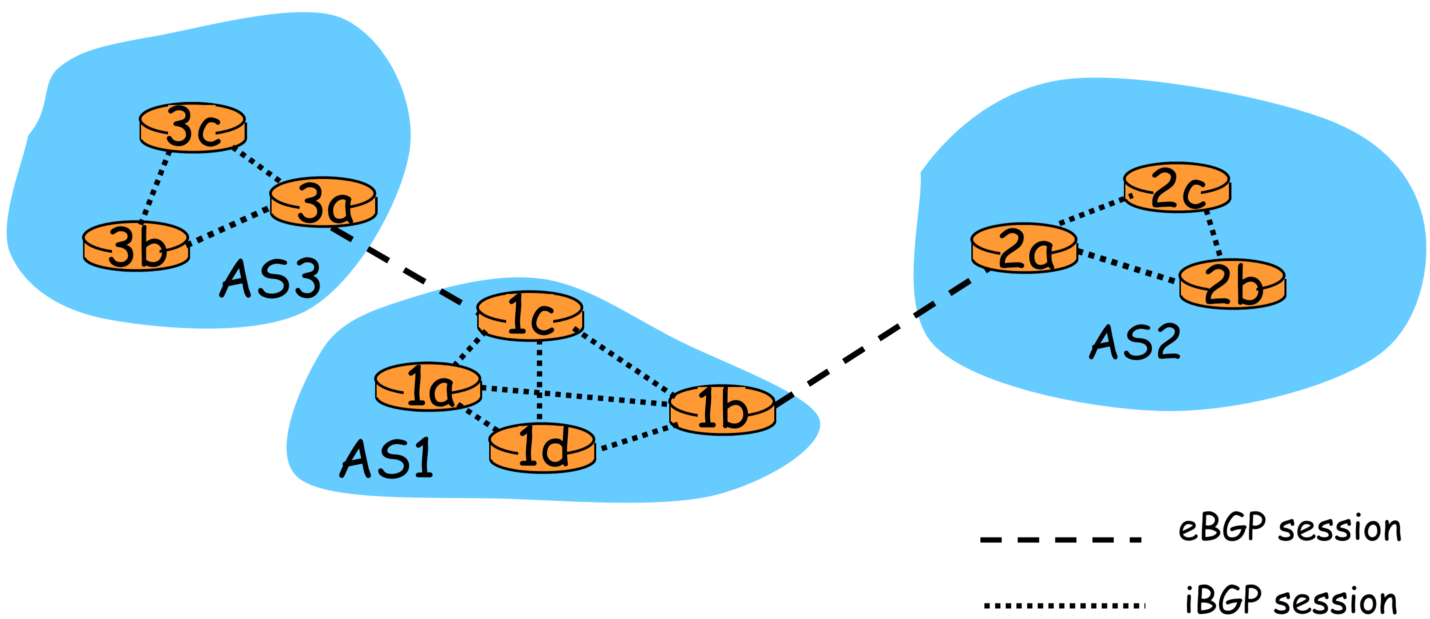
- Pairs of routers (BGP peers) exchange routing info over semi-permanent TCP connections: BGP sessions
  - BGP sessions need not correspond to physical links.

- When AS2 advertises a prefix (i.e., a subnet) to AS1, AS2 is *promising* it will forward any datagrams destined to that prefix towards the prefix.
  - AS2 can aggregate prefixes in its advertisement



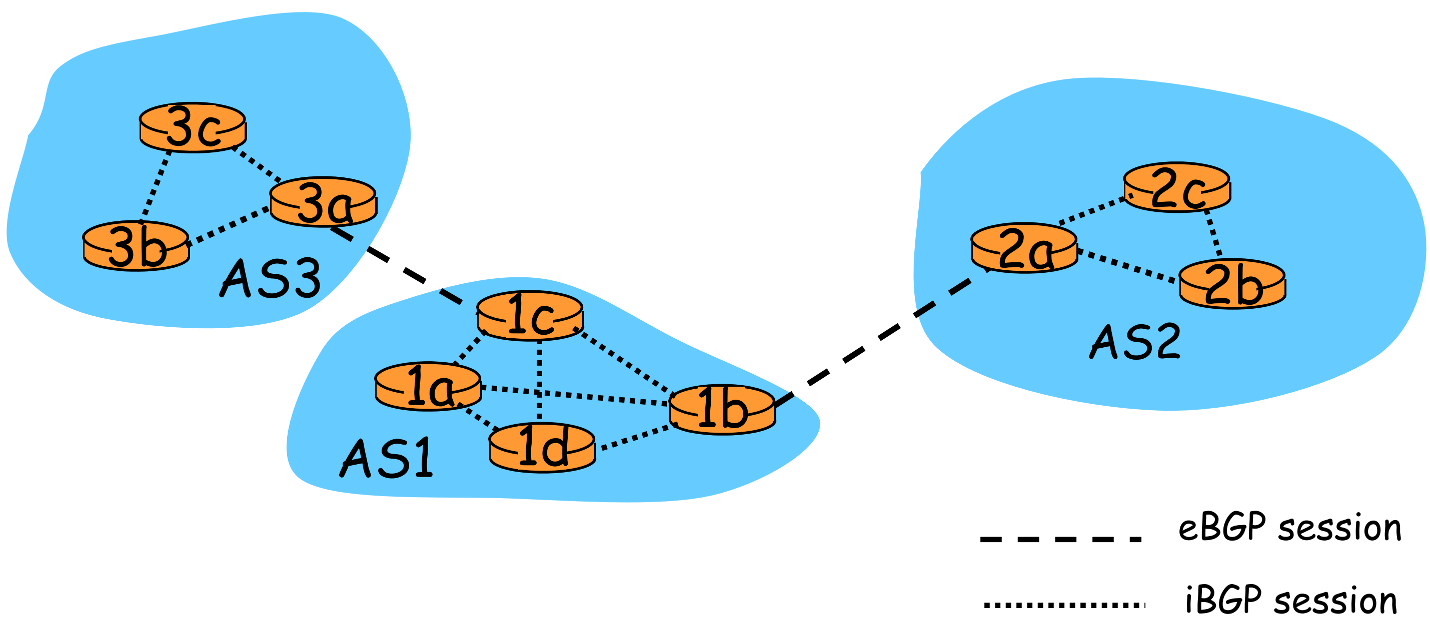
Distributing Reachability Information

- With eBGP (external BGP) session between 3a and 1c, AS3 sends prefix reachability info to AS1.



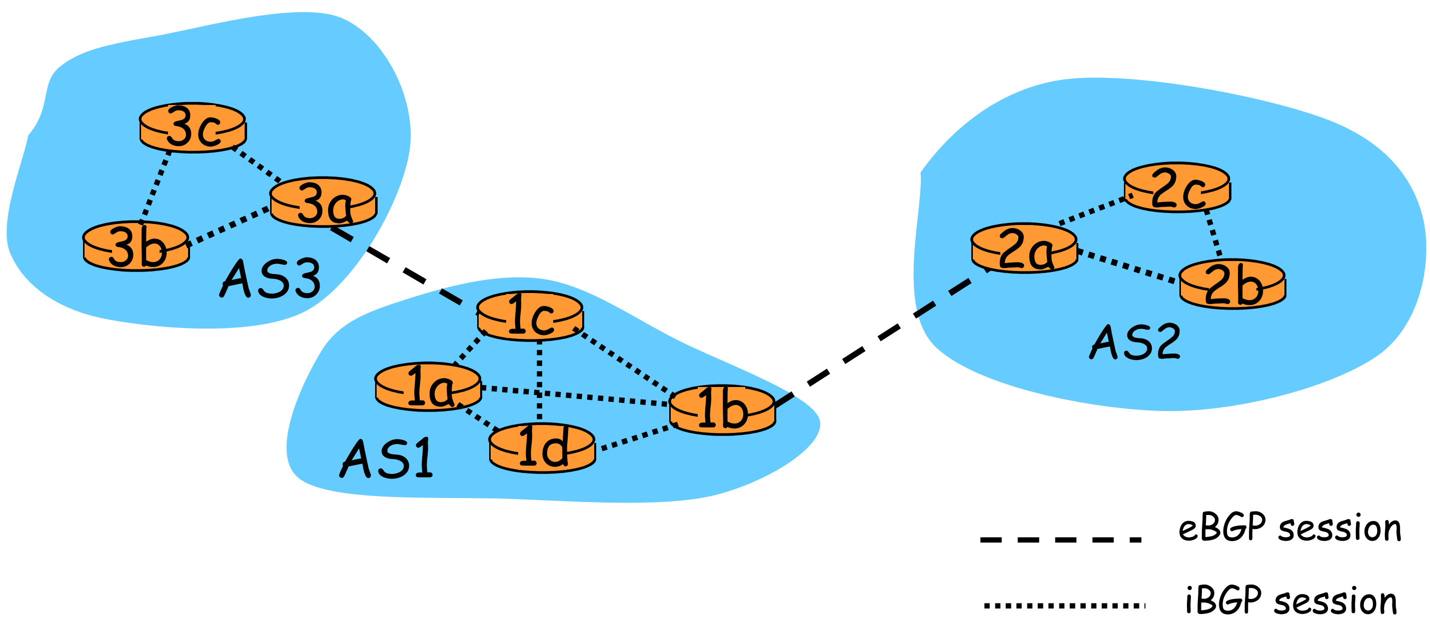
Distributing Reachability Information

- With eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.



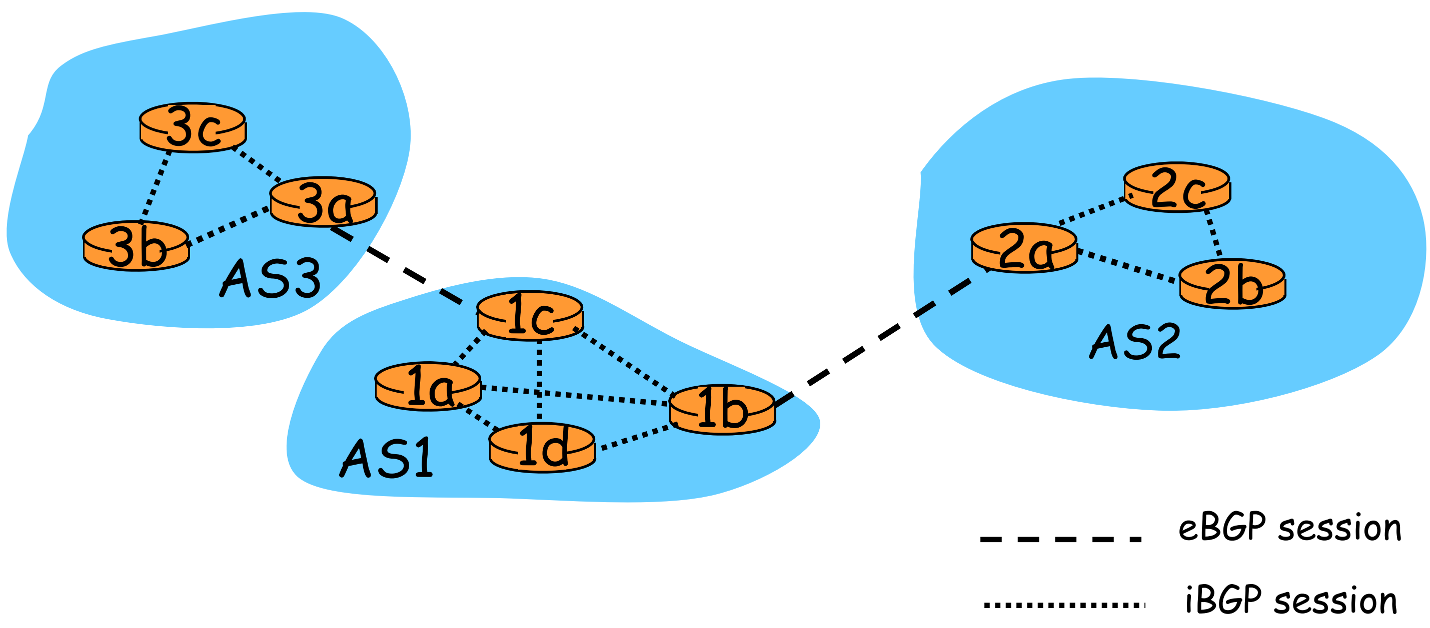
Distributing Reachability Information

- 1c can then use BGP to distribute this new prefix reach info to all routers in AS1



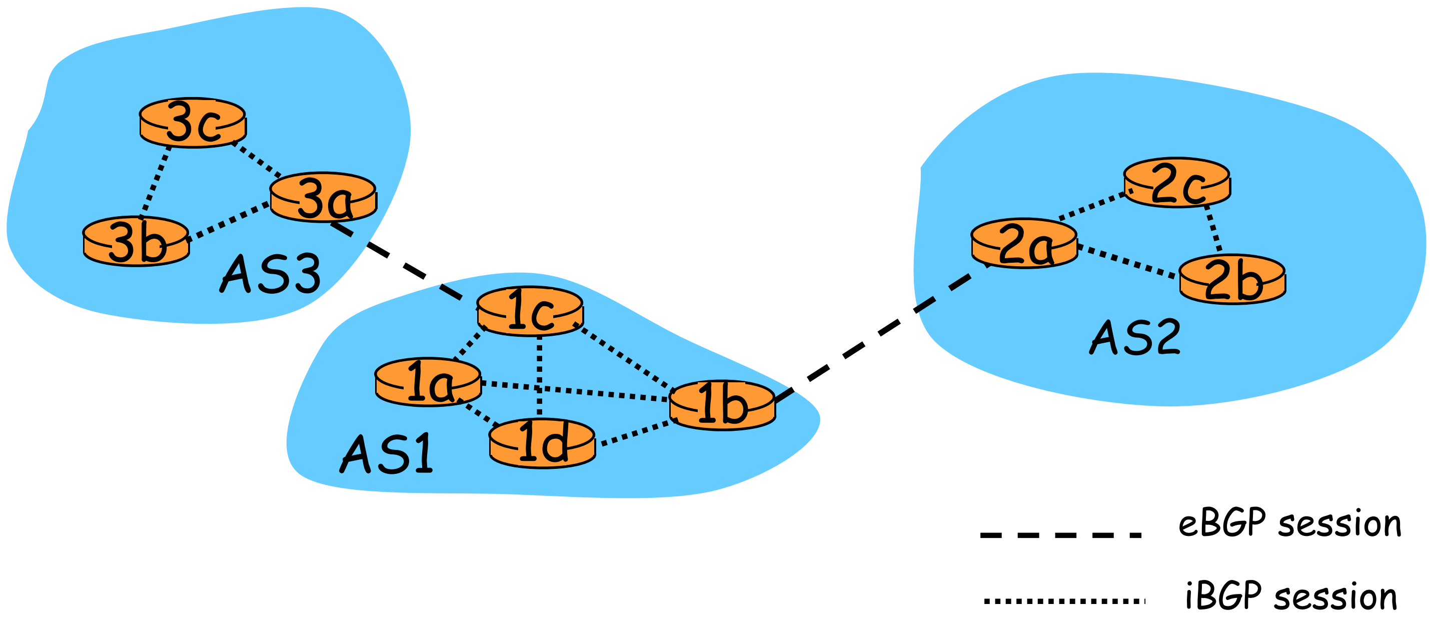
Distributing Reachability Information

- 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session



Distributing Reachability Information

- When a router learns of a new prefix, creates entry for the new prefix in its forwarding table.



Path Attributes and BGP Routes

- When advertising a prefix, advertisement includes BGP attributes.
  - prefix + attributes = "route"
- Two important attributes:
  - AS-PATH
    - contains ASs through which prefix advertisement has passed: AS 67 AS 17
  - NEXT-HOP
    - Indicates specific internal-AS router to next-hop AS.

- When a gateway router receives route advertisement, it uses import policy to accept/decline.

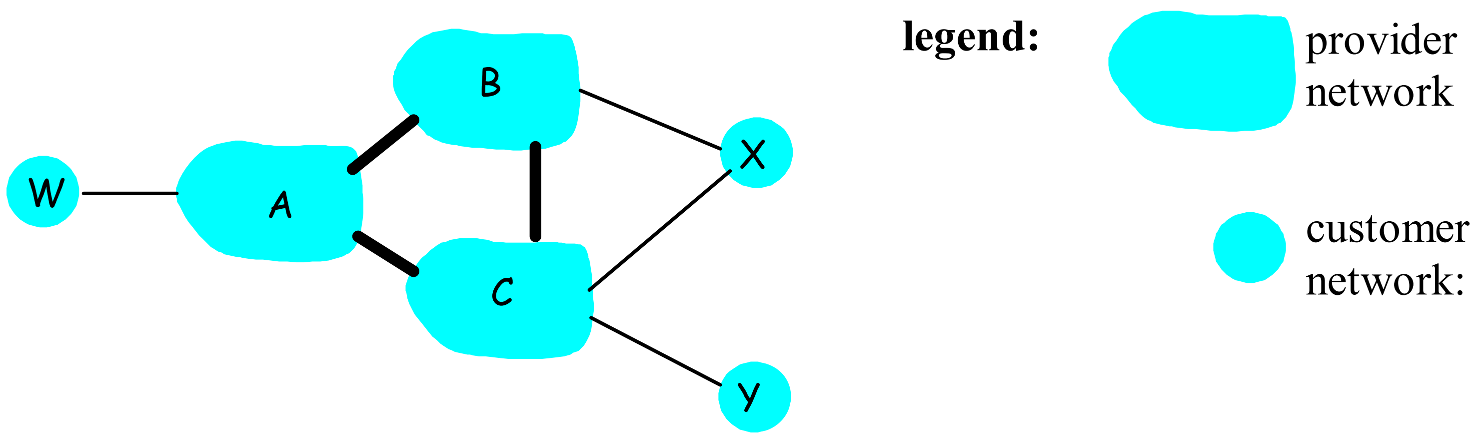
BGP Route Selection

- A router may learn about more than 1 route to some prefix.
  - A router must select route.
- Path selection (elimination) rules:
  - Shortest AS-PATH
  - Closest NEXT-HOP router: hot potato routing
  - Additional criteria

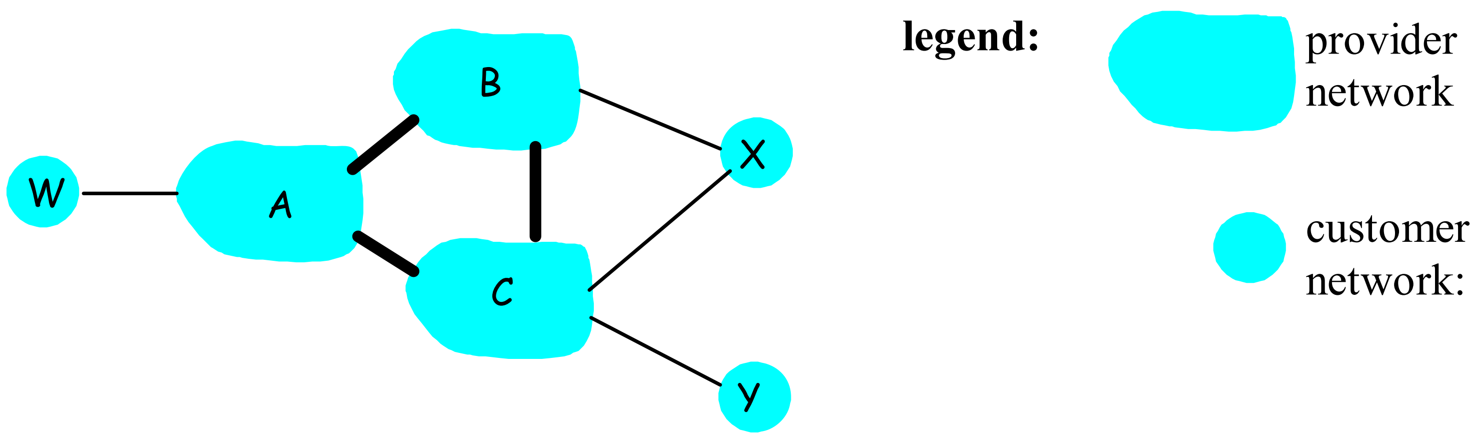
BGP Messages

- BGP messages exchanged using TCP.
- BGP messages
  - OPEN
    - opens TCP connection to a peer and authenticates the sender
  - UPDATE
    - advertises a new path (or withdraws an old path)
  - KEEPALIVE
    - keeps a connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION
    - reports errors in previous msg; also used to close connection

BGP Routing Policy



- A,B,C are provider networks
- X,W,Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B to C via X
  - so X will not advertise to B a route to C



- A advertises to B the path AW
- B advertises to X the path BAW
- Should B advertise to C the path BAW?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route *only* to/from its customers!

Why Different Intra- and Inter-AS Routing ?

- Policy
  - Intra-AS: single admin, so no policy decisions needed
  - Inter-AS: admin wants control over how its traffic routed, who routes through its net.
- Performance
  - Intra-AS: can focus on performance
  - Inter-AS: policy may dominate over performance