CISC 7332X T6

C13b: Routing Problem and Algorithms

Hui Chen

Department of Computer & Information Science

CUNY Brooklyn College

Acknowledgements

- Some pictures used in this presentation were obtained from the Internet
- The instructor used the following references
  - Larry L. Peterson and Bruce S. Davie, Computer Networks: A Systems Approach, 5th Edition, Elsevier, 2011
  - Andrew S. Tanenbaum, Computer Networks, 5th Edition, Prentice-Hall, 2010
  - James F. Kurose and Keith W. Ross, Computer Networking: A Top-Down Approach, 5th Ed., Addison Wesley, 2009
  - Larry L. Peterson's (http://www.cs.princeton.edu/~llp/) Computer Networks class web site

Outline

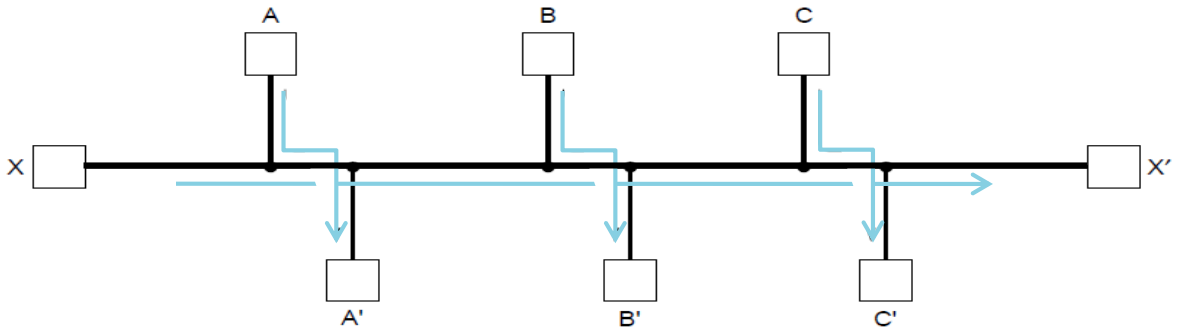
- Routing problem
- Optimality principle
- Routing algorithms

Routing: Motivation

- In a switched network, how does a route is determined?
- If there are multiple routes, which one is better?
- If network topology changes, does a route need to change, how?

Routing Problem

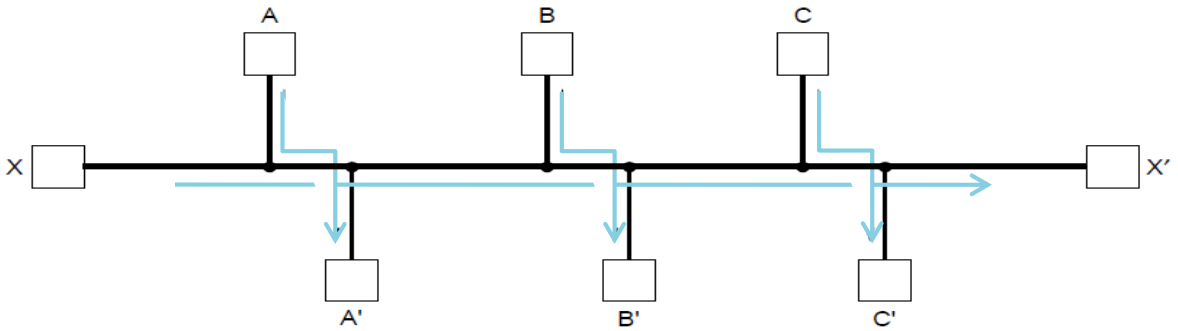
- Routing is the process of discovering network paths
  - Model the network as a graph of nodes and links
  - Decide what to optimize (e.g., fairness vs efficiency)
  - Update routes for changes in topology (e.g., failures)



- Forwarding is the sending of packets along a path

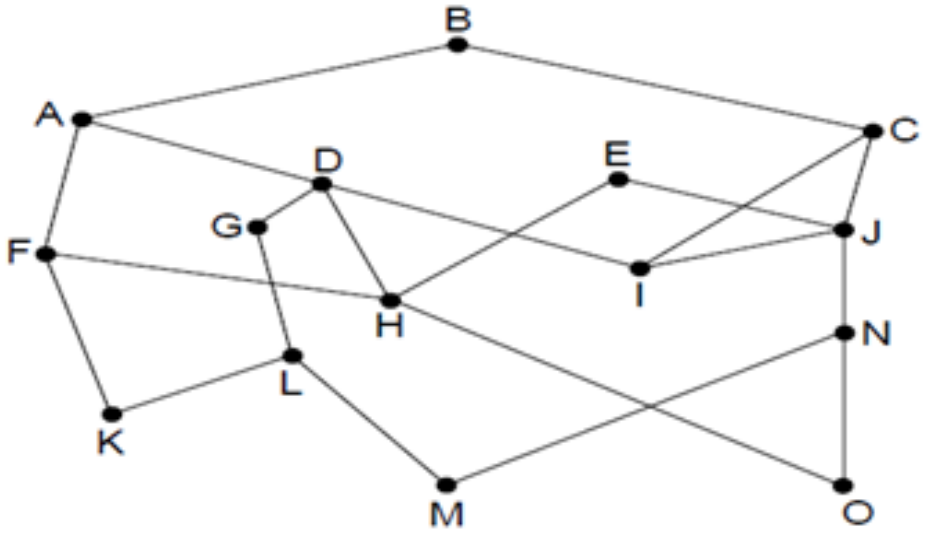
Routing: Objectives

- Correctness
- Simplicity
- Robustness
- Stability
- Fairness
- Efficiency



Optimality Principle

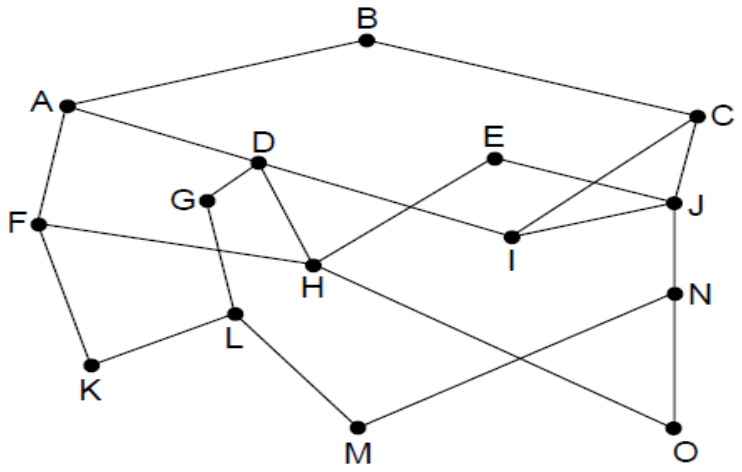
- Each portion of a best path is also a best path (Bellman, 1957)
  - Example: If J is on the optimal path from I to K, the optimal path from J to K also falls along the same route



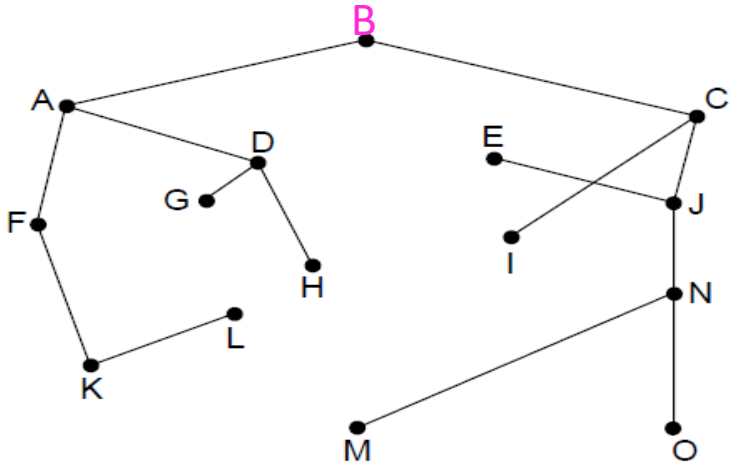
Network

Sink Tree

- The union of best paths to a router is a tree called the sink tree
  - In the example, define "best" as fewest hops



Network



Sink tree of best paths to router B

Sink Tree and Benchmark

- Not realistic, sometimes, useful to compute a sink tree
- However, a sink tree provides a benchmark against which other routing algorithms can be measured or evaluated

Routing Algorithms

- To discuss
  - Distance vector routing
    - Shortest path algorithm
    - Flooding
    - Distance vector routing
  - Link state routing
- On your own
  - Hierarchical routing
  - Broadcast routing
  - Multicast routing
  - Anycast routing
  - Routing for mobile hosts
  - Routing in ad hoc networks

Shortest Path

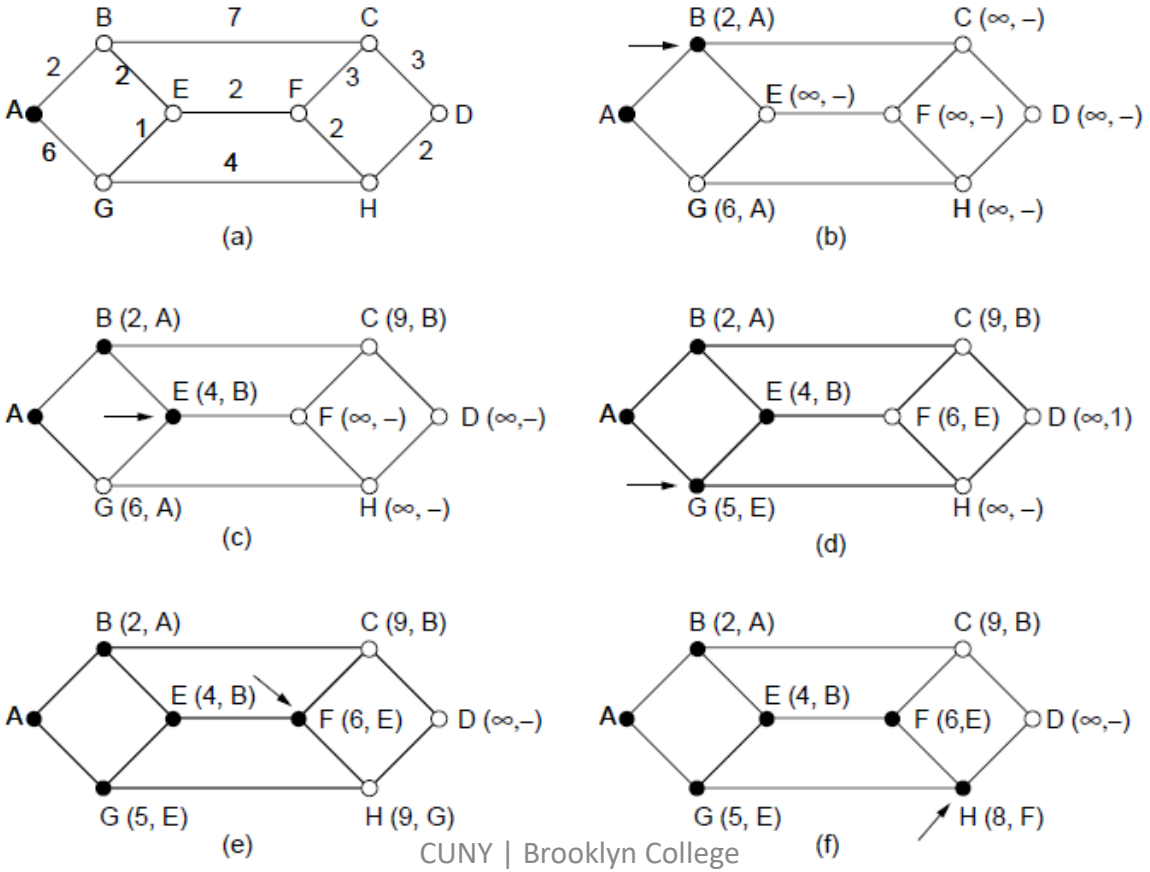
- Dijkstra's shortest path algorithm computes a sink tree on the graph:
  - Each link is assigned a non-negative weight/distance
  - Shortest path is the one with lowest total weight
  - Using weights of 1 gives paths with fewest hops

Shortest Path Algorithm

- Start with sink, set distance at other nodes to infinity
- Determine distances to other nodes
- Pick the lowest distance node, add it to sink tree
- Repeat until all nodes are in the sink tree

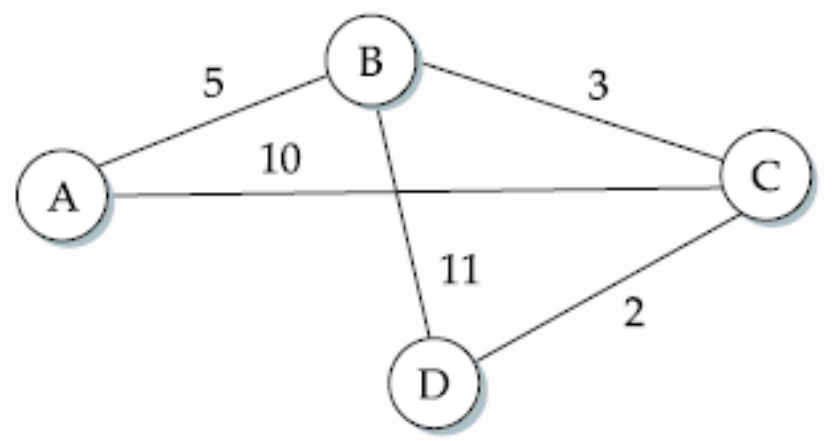
Shortest Path Algorithm: Example

- Notation: E(4, B): from sink at 4 via B



Exercise C13b-1

- Following the example illustrated and using the Dijkstra's shortest path algorithm, find the shortest path to all the other nodes from node D and show steps



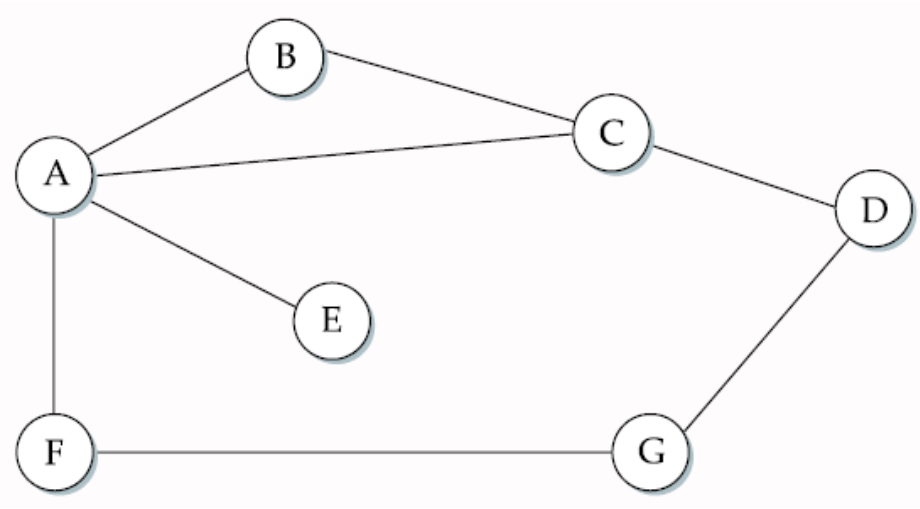
Flooding

- A simple method to send a packet to all network nodes
- Each node floods a new packet received on an incoming link by sending it out all of the other links
- Nodes need to keep track of flooded packets to stop the flood; even using a hop limit can blow up exponentially

Distance Vector

- Each node constructs a one dimensional array (a vector) containing the "distances" (costs) to all other nodes and distributes that vector to its immediate neighbors
- Starting assumption is that each node knows the cost of the link to each of its directly connected neighbors

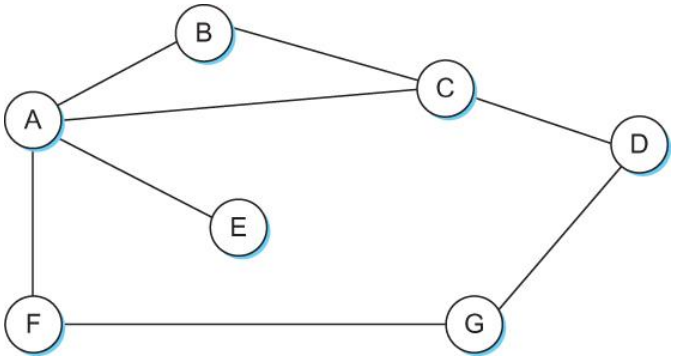
Distance From a Node to Other Nodes



- What is the (shortest) distance from A to B?
- What is the (shortest) distance from A to C?
- What is the (shortest) distance from A to D?

Distance Vector: Example

- Initial distances stored at each node (*global view*)

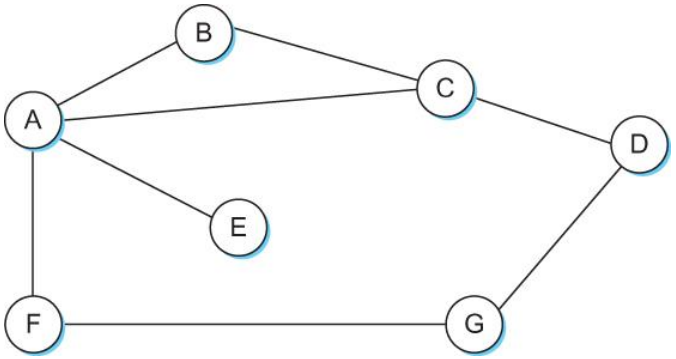


Information Stored at Node	Distance to Reach Node						
	A	B	C	D	E	F	G
A	0	1	1	∞	1	1	∞
B	1	0	1	∞	∞	∞	∞
C	1	1	0	1	∞	∞	∞
D	∞	∞	1	0	∞	∞	1
E	1	∞	∞	∞	0	∞	∞
F	1	∞	∞	∞	∞	0	1
G	∞	∞	∞	1	∞	1	0

- No node has this global view!

Distance Vector: Example of Initial Routing Table

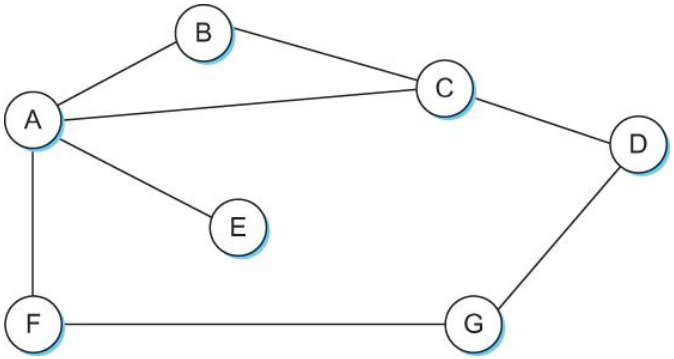
- Initial routing table at node *A*



Destination	Cost	NextHop
B	1	B
C	1	C
D	∞	—
E	1	E
F	1	F
G	∞	—

Distance Vector: Example of Final Routing Table

- Final routing table at node A

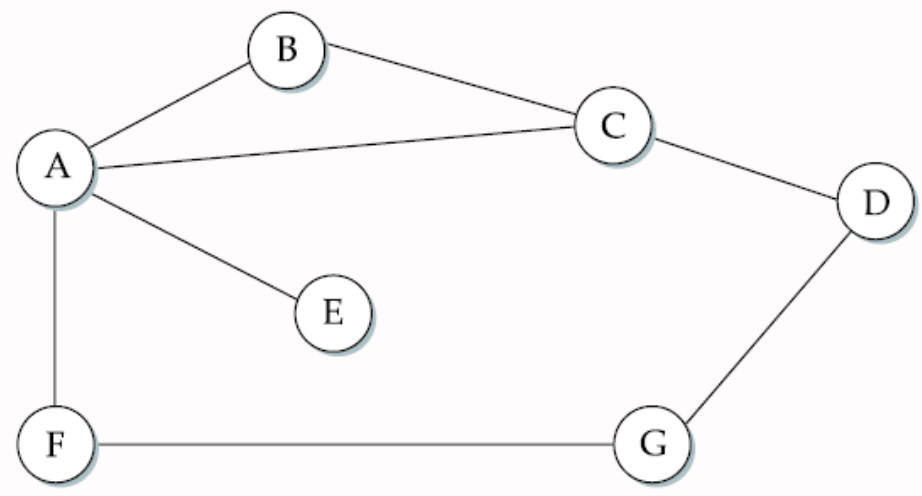


Distance vector: distances from A to the other nodes

Destination	Cost	NextHop
B	1	B
C	1	C
D	2	C
E	1	E
F	1	F
G	2	F

Exercise C13b-2

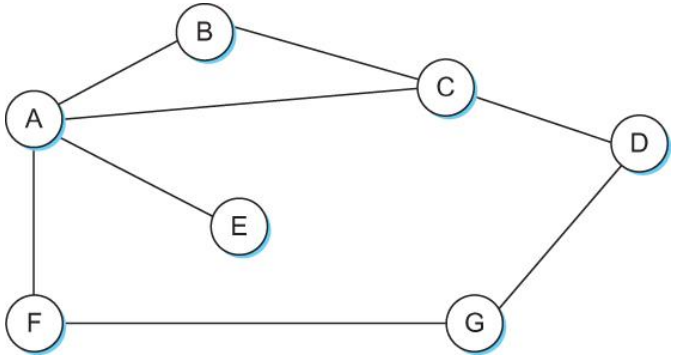
- Given an internetwork below, construct the *initial* routing table for the distance vector routing algorithm at *router C* (by filling the provided table below)



Destination	Cost	Next Hop
A		
B		
D		
E		
F		
G		

Distance Vector: Example

- Final distances stored at each node (*global view*)

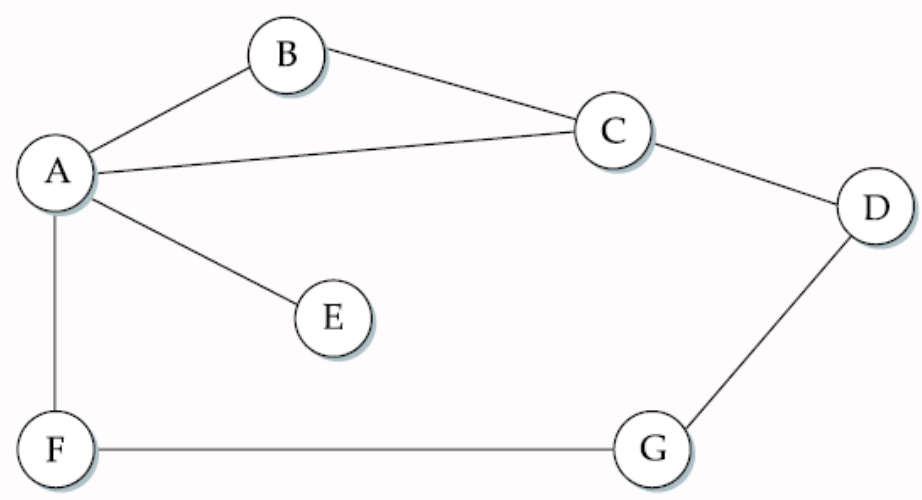


Information Stored at Node	Distance to Reach Node						
	A	B	C	D	E	F	G
A	0	1	1	2	1	1	2
B	1	0	1	2	2	2	3
C	1	1	0	1	2	2	2
D	2	2	1	0	3	2	1
E	1	2	2	3	0	2	3
F	1	2	2	2	2	0	1
G	2	3	2	1	3	1	0

- No node has this global view!

Exercise C13b-3

- Given an internetwork below, construct the *final* routing table for the distance vector routing algorithm at *router C* (by filling the provided table below)



Destination	Cost	Next Hop
A		
B		
D		
E		
F		
G		

Distance Vector Routing Algorithm

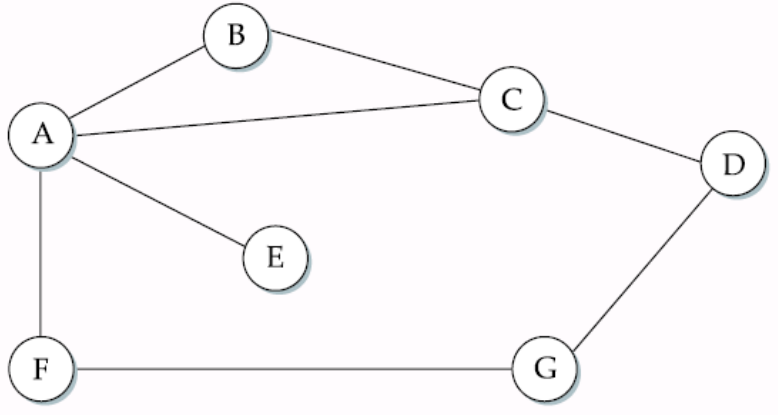
- Sometimes called as *Bellman-Ford* algorithm
- Main idea
  - Every T seconds each router sends its table to its neighbor each router then updates its table based on the new information
- Problems
  - Fast response to good news, but slow response to bad news
  - Also too many messages to update

Distance Vector Routing Algorithm: More Details

- Each node maintains a routing table consisting of a set of triples
  - (Destination, Cost, NextHop)
- Exchange updates directly connected neighbors
  - periodically (on the order of several seconds)
  - whenever table changes (called *triggered update*)
- Each update is a list of pairs:
  - (Destination, Cost): from sending router to destination
  - Update local table if receive a "better" route
    - smaller cost
    - came from next-hop
- Refresh existing routes; delete if they time out

Table Update

- Example: Exchange updates between A and C



- Then A sends an update to C

Destination	Cost
B	1
C	1
D	∞
E	1
F	1
G	∞

C's initial routing table

Destination	Cost	Next Hop
A	1	A
B	1	B
D	1	D
E	∞	-
F	∞	-
G	∞	-

C's updated routing table

Destination	Cost	Next Hop
A	1	A
B	1	B
D	1	D
E	2	A
F	2	A
G	∞	-

Table Update from A at C

Destination	Cost
B	1
C	1
D	∞
E	1
F	1
G	∞

+ 1 =

Destination	Cost	Next Hop
B	2	A
C	2	A
D	∞	A
E	2	A
F	2	A
G	∞	A

Destination	Cost	Next Hop
A	1	A
B	1	B
D	1	D
E	∞	-
F	∞	-
G	∞	-

Destination	Cost	Next Hop
A	1	A
B	1	B
D	1	D
E	2	A
F	2	A
G	∞	-

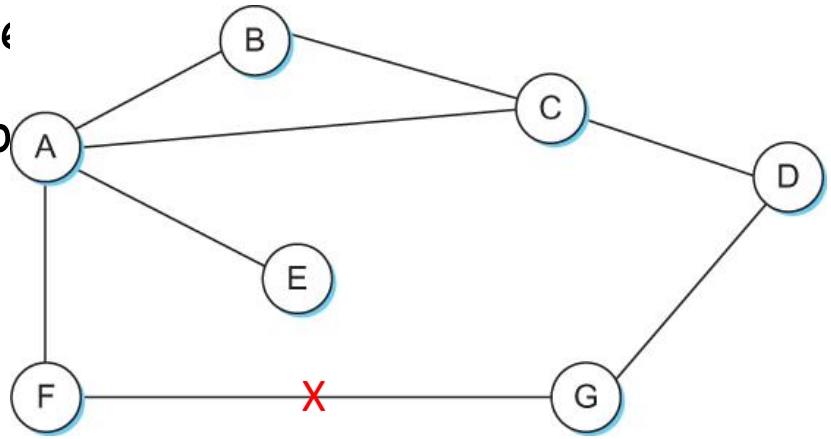
Convergence

- Process of getting consistent routing information to all the nodes
- Desired results: routing tables converges to a stable *global* table (no more changes upon receiving updates from neighbors)

Information Stored at Node	Distance to Reach Node						
	A	B	C	D	E	F	G
A	0	1	1	2	1	1	2
B	1	0	1	2	2	2	3
C	1	1	0	1	2	2	2
D	2	2	1	0	3	2	1
E	1	2	2	3	0	2	3
F	1	2	2	2	2	0	1
G	2	3	2	1	3	1	0

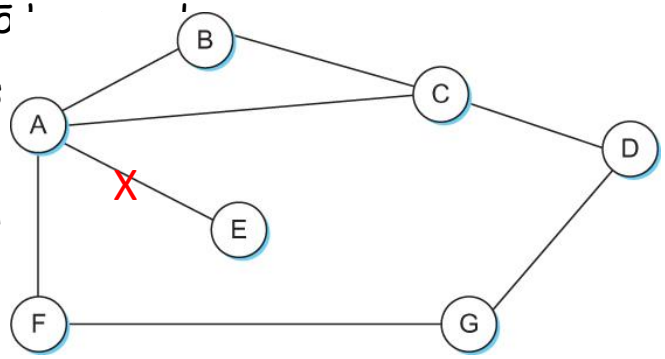
Link Failure: Example

- When a node detects a link failure
  - F detects that link to G has failed
  - F sets distance to G to infinity and sends update to A
  - A sets distance to G to infinity since it uses F to reach G
  - A receives periodic update from C with 2-hop path to G
  - A sets distance to G to 3 and s
  - F decides it can reach G in 4 ho



Count-to-infinity Problem

- Slightly different circumstances can prevent the network from *stabilizing*
  - Suppose the link from A to E goes down
  - In the next round of updates, A advertises a distance of infinity to E, but B and C advertise a distance of 2 to E
  - Depending on the exact timing of events, the following might happen
    - Node B, upon hearing that E can be reached in 2 hops from C, concludes that it can reach E in 3 hops and advertises this to A
    - Node A concludes that it can reach E in 4 hops and advertises this to C
    - Node C concludes that it can reach E in 5 '
    - This cycle stops only when the distances enough to be considered infinite
    - *called* **count-to-infinity problem**



Count-to-infinity Problem: Solutions

- Use some relatively small number as an approximation of infinity
- For example, the maximum number of hops to get across a certain network is never going to be more than 16
  - Set infinity to 16
  - Stabilize fast, but not working for larger networks
- One technique to improve the time to stabilize routing is called *split horizon*

Split Horizon

- When a node sends a routing update to its neighbors, it does *not* send those routes it learned from each neighbor *back* to that neighbor
- For example, if B has the route (E, 2, A) in its table, then it knows it must have learned this route from A, and so whenever B sends a routing update to A, it does not include the route (E, 2) in that update

Split Horizon with Poison Reverse

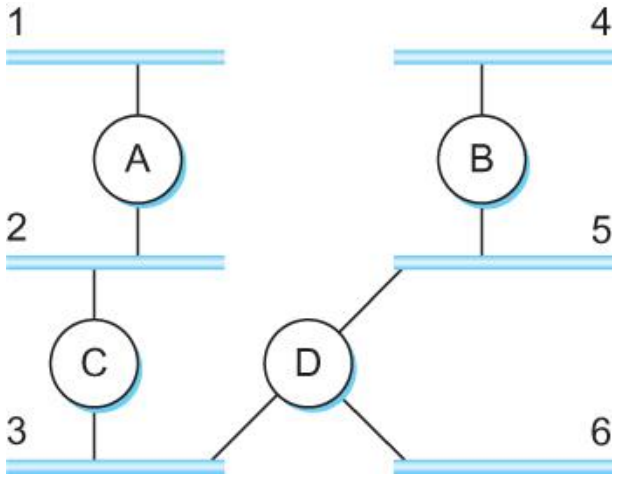
- In a stronger version of split horizon, called *split horizon with poison reverse*
  - B actually sends that back route to A, but it puts negative information in the route to ensure that A will not eventually use B to get to E
  - For example, B sends the route (E, ∞) to A

Routing Information Protocol

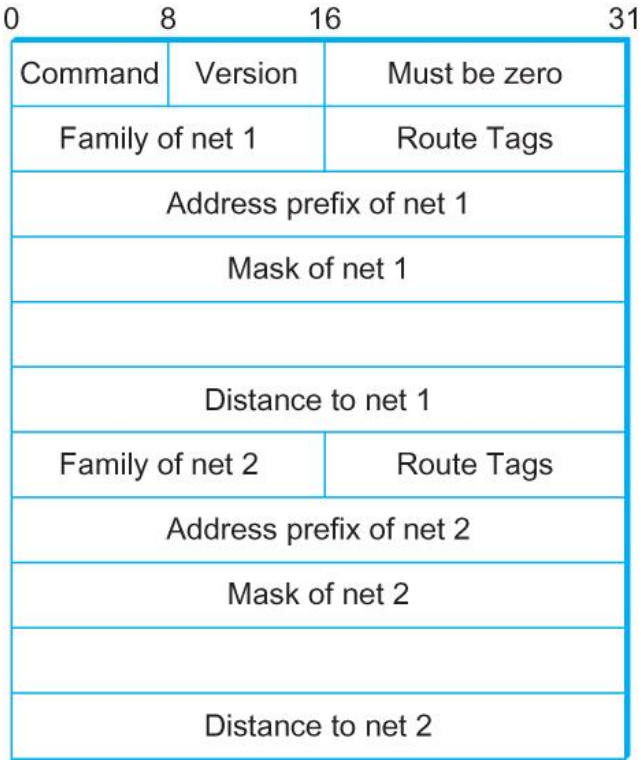
- Routing Information Protocol (RIP)
  - Initially distributed along with BSD Unix
  - Widely used
- Straightforward implementation of distance-vector routing

Routing Information Protocol (RIP)

- Distance: cost (# of routers) of reach a network
  - C → A
    - Network 2 at cost 0; 3 at cost 0
    - Network 5 at cost 1, 4 at 2



Example Network



RIPv2 Packet Format

Link State Routing

- Link state is an alternative to distance vector
  - More computation but simpler dynamics
  - Widely used in the Internet (OSPF, ISIS)
- Algorithm:
  - Each node floods information about its neighbors in LSPs (Link State Packets); all nodes learn the full network graph
  - Each node runs Dijkstra's algorithm to compute the path to take for each destination

Link State Routing

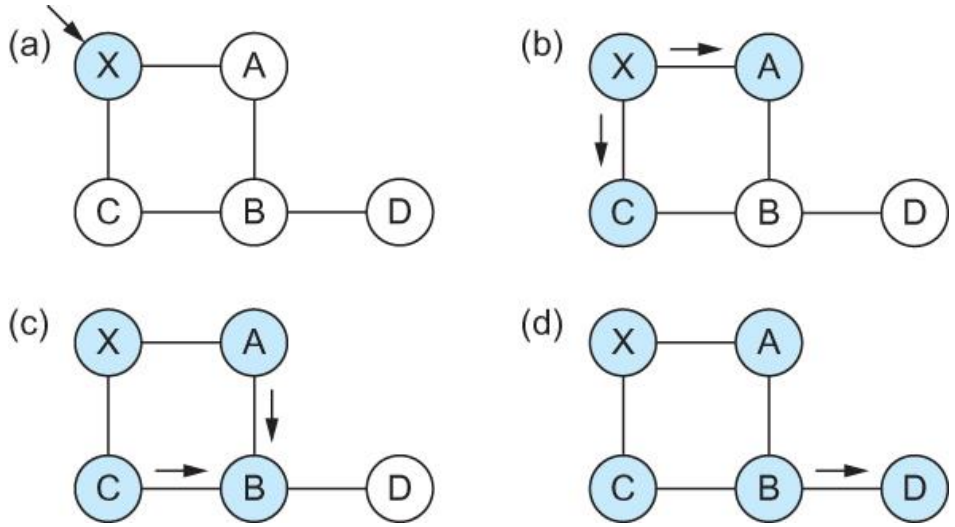
- Strategy: Send to all nodes (not just neighbors) information about directly connected links (not entire routing table).
- Link State Packet (LSP)
  - id of the node that created the LSP
  - cost of link to each directly connected neighbor
  - sequence number (SEQNO)
  - time-to-live (TTL) for this packet
- Reliable Flooding
  - store most recent LSP from each node
  - forward LSP to all nodes but one that sent it
  - generate new LSP periodically; increment SEQNO
  - start SEQNO at 0 when reboot
  - decrement TTL of each stored LSP; discard when TTL=0

Reliable Flooding

- Reliable flooding triggered by
  - Timer
  - Topology or link cost change
- increment SEQNO
  - start SEQNO at 0 when reboot
  - SEQNO does not wrap
    - e.g., 64 bits
  - decrement TTL of each stored LSP
- discard when TTL=0

Link State Routing: Example

- Reliable Flooding



- Flooding of link-state packets. (a) LSP arrives at node X; (b) X floods LSP to A and C; (c) A and C flood LSP to B (but not X); (d) flooding is complete

Shortest Path Routing Algorithm

- In practice, each switch computes its routing table directly from the LSPs it has collected using a realization of Dijkstra's algorithm called the *forward search algorithm*
- Specifically each switch maintains two lists, known as **Tentative** and **Confirmed**
- Each of these lists contains a set of entries of the form (Destination, Cost, NextHop)

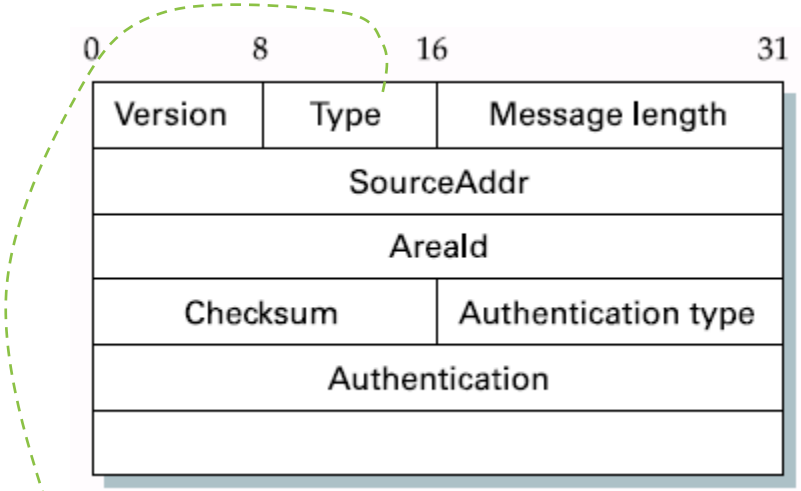
Link State in Practice

- Open Shortest Path First Protocol (OSPF)
  - "Open" → open, non-proprietary standard, created under the auspices of the IETF
  - "SPF" → Shortest Path First, alternative name of link-state routing
- Implementation of Link-State Routing with added features
  - Authenticating of routing messages
    - Due to the fact too often some misconfigured hosts decide they can reach every host in the universe at a cost of 0
  - Additional hierarchy
    - Partition domain into areas → increase scalability
  - Load balancing
    - Allows multiple routes to the same place to be assigned the same cost → cause traffic to be distributed evenly over those routes

Open Shortest Path First Protocol

OSPF Header Format

Advertisement



OSPF Link State

LS Age		Options	Type=1
Link-state ID			
Advertising router			
LS sequence number			
LS checksum		Length	
0	Flags	0	Number of links
Link ID			
Link data			
Link type	Num_TOS	Metric	
Optional TOS information			
More links			

Type	Packet name	Protocol function
1	Hello	Discover/maintain neighbors
2	Database Description	Summarize database contents
3	Link State Request	Database download
4	Link State Update	Database update
5	Link State Ack	Flooding acknowledgment

Metrics

- Original ARPANET metric
  - measures number of packets enqueued on each link
  - took neither latency or bandwidth into consideration
- New ARPANET metric
  - stamp each incoming packet with its arrival time (AT)
  - record departure time (DT)
  - when link-level ACK arrives, compute
- Delay = (DT - AT) + Transmit + Latency
  - if timeout, reset DT to departure time for retransmission
  - link cost = average delay over some time period
- Fine Tuning
  - compressed dynamic range
  - replaced Delay with link utilization

Questions?

- Distance Vector
  - Algorithm
  - Routing Information Protocol (RIP)
- Link State
  - Algorithm
  - Open Shortest Path First Protocol (OSPF)
- Metrics
  - How to measure link cost?