CS4450

Computer Networks: Architecture and Protocols

Lecture 12
Intra-domain Routing: Deep Dive

Rachit Agarwal

Announcements

- Problem Set 3 will be released today

- **Prelim: 28th March, In-class**
    - **Nobody should be in conflict**
        - **Let me know immediately if there is a conflict**

- **Project 1 will be out next week**
    - Same idea: do when you get time (no deadline)
    - We will provide all the help

- We are at a stage where **you have a strong foundation in networks**
    - Sharing networks, architectural principles, design goals
    - And, you don't realize it, but **you understand many new tradeoffs**

Goals for Today's Lecture

- **Learning about Routing Protocols**
- **Link State (Global view, Local computation)**
- **Distance Vector (Local view, Local computation)**
- Maintain sanity: its one of the "harder" lectures
  - I'll try to make it -less- hard, but …
    - Pay attention
    - Review again tomorrow
    - Work out a few examples

Recap from last few lectures

Recap: Spanning Tree Protocol …

- Used in switched Ethernet to avoid broadcast storm
- Can be used for routing on the Internet (via “flooding” on spanning tree)
- **Three fundamental issues**:
    - Unnecessary processing at end hosts (that are not the destination)
    - Higher latency
    - Lower available bandwidth

Recap: Routing Tables

- **Routing table:**
    - Each switch: the next hop for each destination in the network
- **Routing state**: collection of routing tables across all nodes
- Two questions:
    - How can we **verify** given routing state is valid?
    - How can we **produce** valid routing state?
- Global routing state valid **if and only if**:
    - There are no **dead ends** (other than destination)
    - There are no **“persistent” loops**

Recap: The right way to think about Routing Tables

- Routing tables are nothing but ….
    - A collection of (directed) spanning tree
    - One for each destination

- **Routing Protocols**
    - Mechanisms to producing valid routing tables
    - What we will see:
        - “n” spanning tree protocols running in parallel

Recap: Three flavors of protocols for producing valid routing state

- **Create Tree, route on tree**
    - E.g., Spanning tree protocol (switched Ethernet)
    - **Good:** easy, no (persistent) loops, no dead ends
    - **Not-so-good:** unnecessary processing, high latency, low bandwidth

- **Obtain a global view:**
    - E.g., Link state (last lecture)

- **Distributed route computation:**
    - E.g., Distance vector
    - E.g., Border Gateway Protocol

Recap: Producing Valid Routing State

- Easy to avoid dead ends
- Avoiding loops is hard
- **The key difference between routing protocols is how they avoid loops!**

Recap: Global View

- **Questions**:
    - Where to create global view
    - How to create global view
    - When to run route computation

Recap: Where to create global view?

- One option: Central server
    - Collects a global view
    - Computes the routing table for each node
    - "Installs" routing tables at each node
    - **Software-defined Networks: later in course**

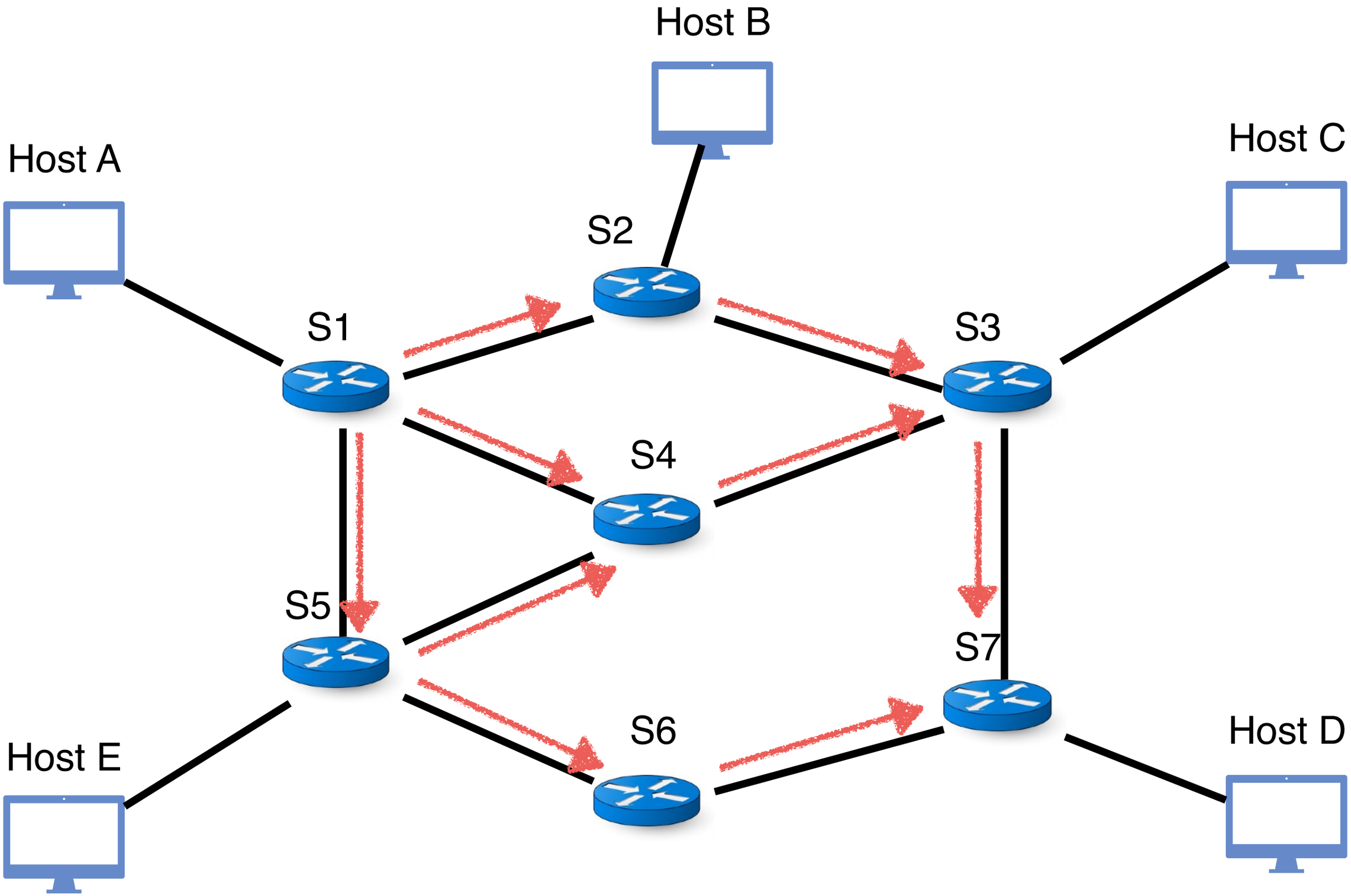
- Second option: At each router
    - Each router collects a global view
    - Computes its own routing table using Link-state protocol

- **Link-state routing protocol**
    - OSPF is a specific implementation of link-state protocol
        - IETF RFC 2328 (IPv4) or 5340 (IPv6)

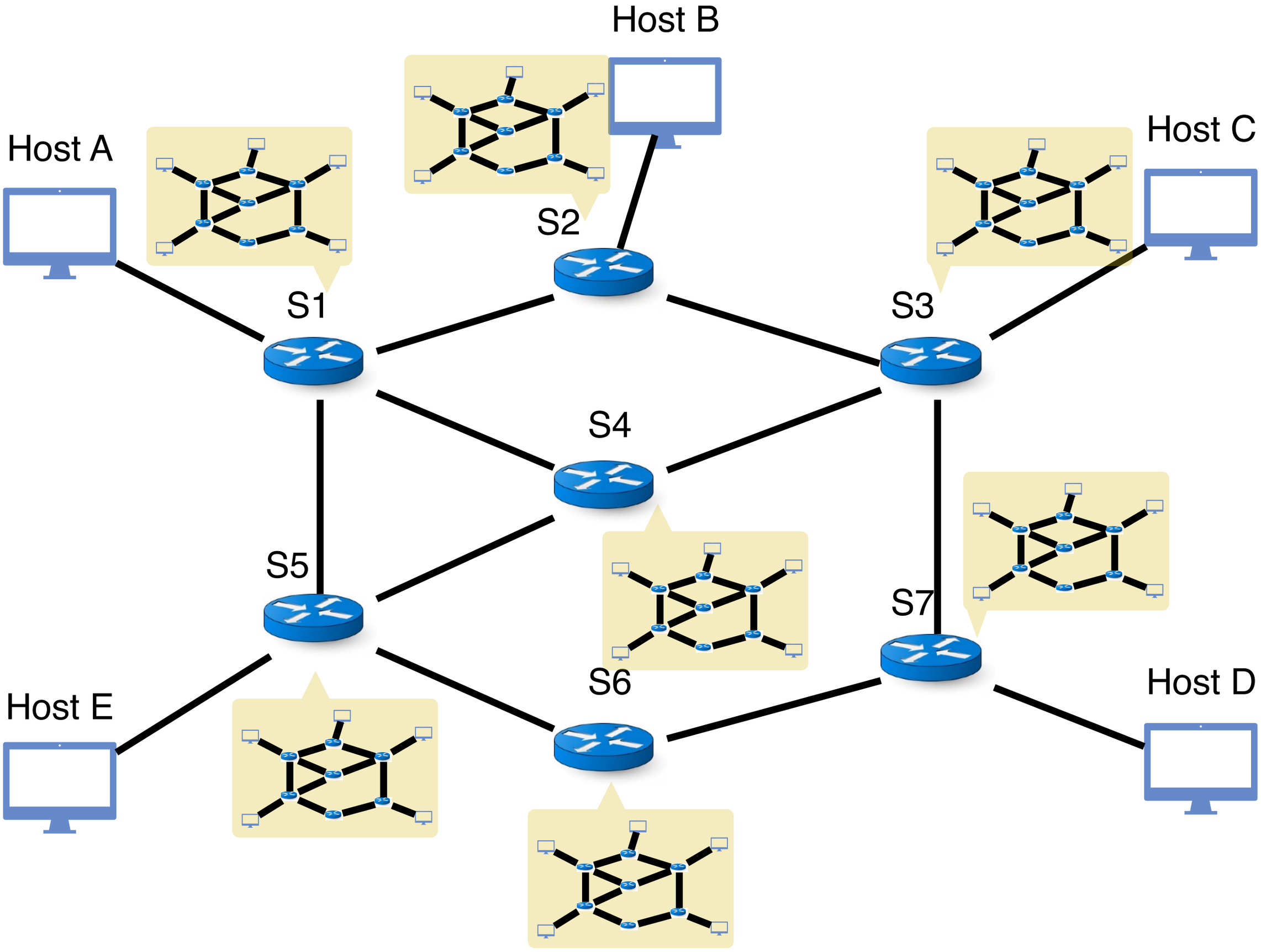
Recap: Overview of Link-State Routing

- **Every router knows its local "link state"**
    - Knows state of links to neighbors
    - Up/down, and associated cost
- **A router floods its link state to all other routers**
    - Uses a special packet — Link State Announcements (LSA)
    - Announcement is delivered to all nodes (next slide)
    - Hence, every router learns the entire network graph
- **Runs route computation locally**
    - Computing least cost paths from them to all other nodes
    - E.g., using Dijkstra's algorithm

Recap: Link-State Routing



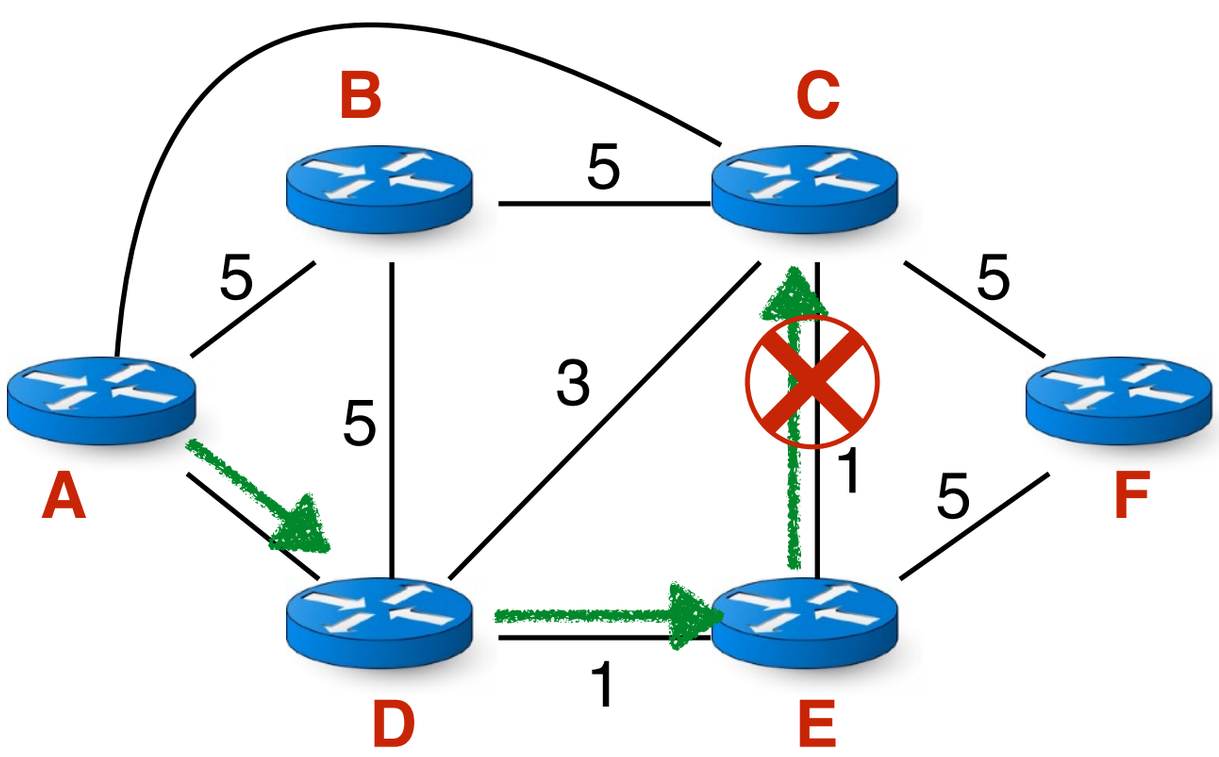
Recap: Each Node Then has a Global View



Recap: When to Initiate Flooding of announcements?

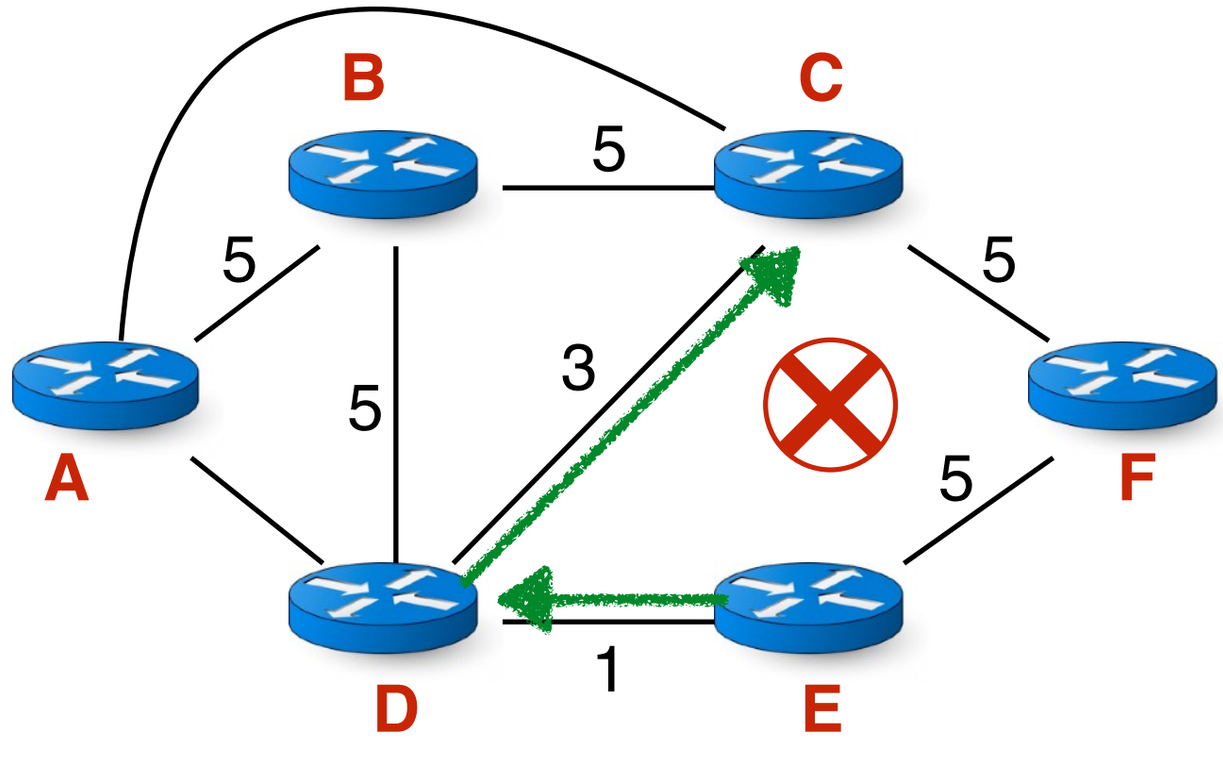
- **Topology change**
    - Link failures
    - Link recovery
- **Configuration change**
    - Link cost change (why would one change link cost?)
- **Periodically**
    - Refresh the link-state information
    - Typically (say) 30 minutes
    - Corrects for possible corruption of data

Recap: Are Loops Still Possible?



A and D think this is the path to C

E-C link fails, but D doesn't know yet

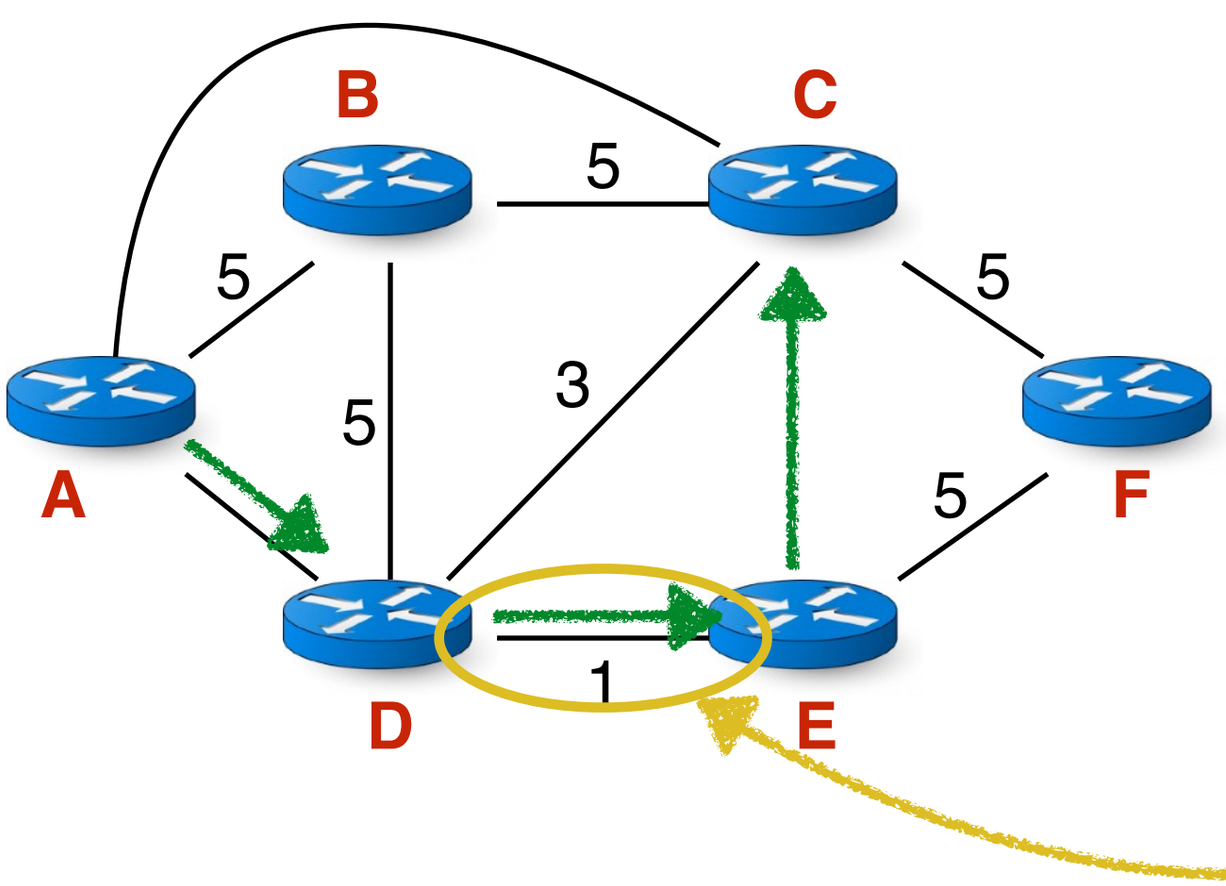


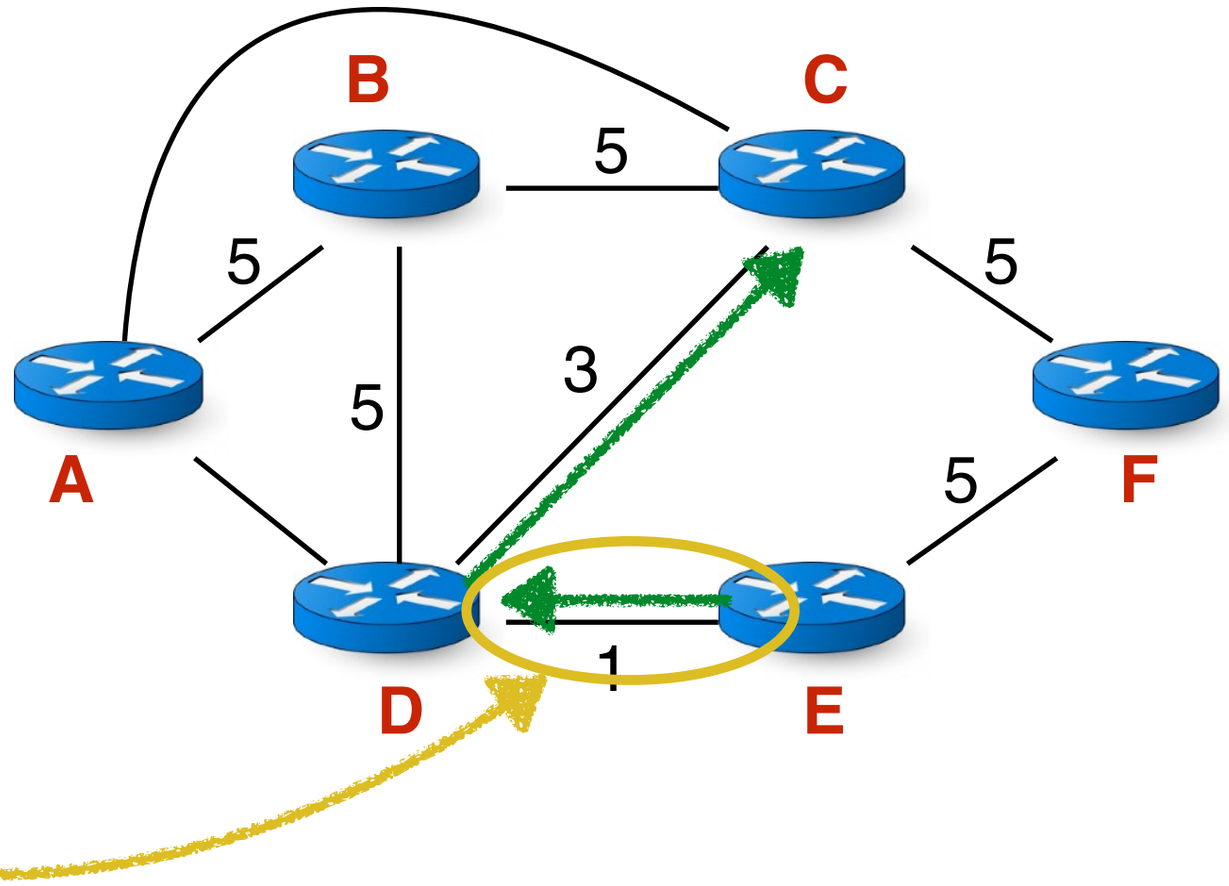
E thinks that this the path to C

E reaches C via D, D reaches C via E

Loop!

Recap: Transient Disruptions





- Inconsistent link-state views
  - Some routers know about failure before others
  - The shortest paths are no longer consistent
  - Can cause **transient forwarding loops**
    - **Transient loops are still a problem!**

Questions?

Convergence in Link-State Routing

- **Eventually**, all routers have consistent routing information
    - E.g., all nodes having the same link-state database
- Here, eventually means “if nothing changes after a while”
- Forwarding is consistent after convergence
    - All nodes have the same link-state database
    - All nodes forward packets on same paths
- **But while still converging, bad things can happen**

Time to Reach Convergence

- Sources of convergence delay?
    - Time to detect failure
    - Time to flood link-state information (~longest RTT)
    - Time to recompute forwarding tables
- Performance problems during convergence period?
    - Dead ends
    - Looping packets
    - And some more we'll see later ….

Final words: Link State is Conceptually Simple

- Everyone floods links information
- Everyone then knows graph of the network
- Everyone independently computes paths on the graph
- **All the complexity is in the details**

Distributed Route Computation

Distributed Computation of Routes

- Each node computes the outgoing links (for each destination) based on:
    - Local link costs
    - Information advertised by neighbors
- Algorithms differ in what these exchanges contain
    - **Distance-vector**: just the distance (and next hop) to each destination
    - **Path vector**: the entire path to each destination
- We will focus on distance-vector for now

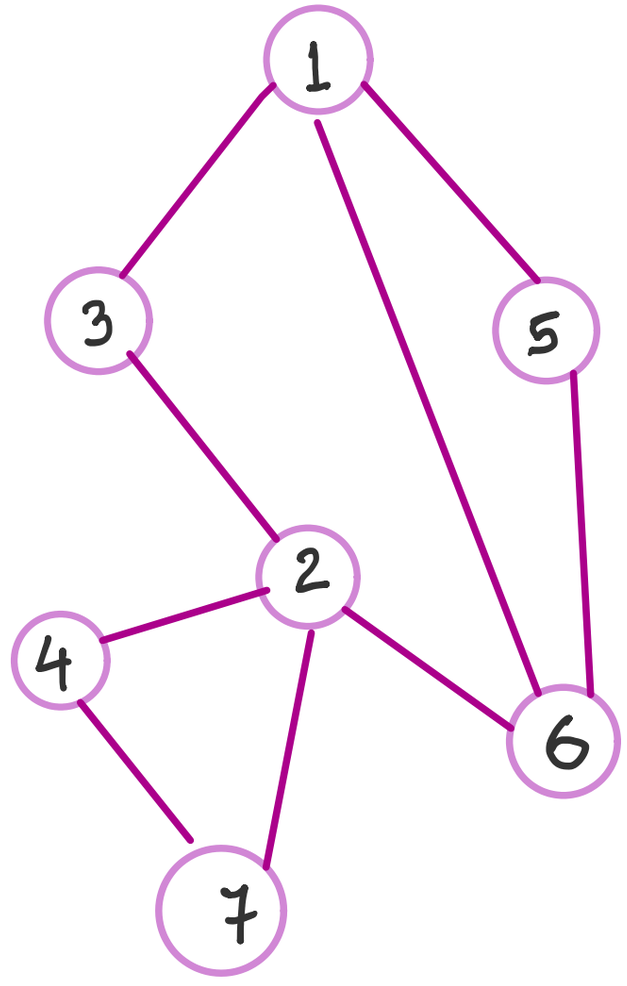
Recall: Routing Tables = Collection of Spanning Trees

- **Can we use the spanning tree protocol (with modifications)?**
- **Messages (Y,d,X): For root Y; From node X; advertising a distance d to Y**
- Initially each switch X announces (X,0,X) to its neighbors

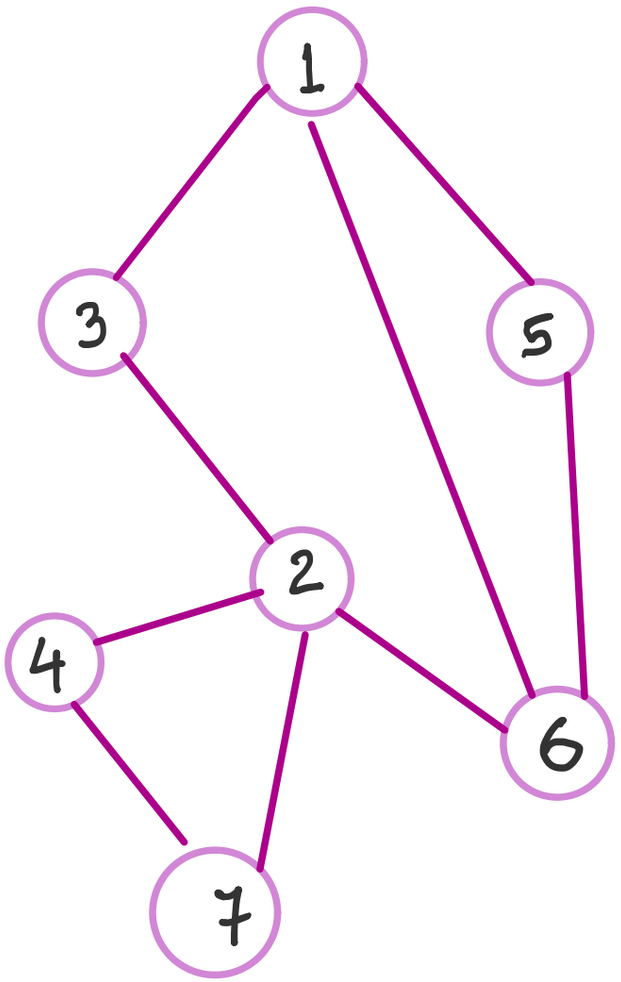
Distance vector: a collection of “n” STP in parallel

Lets run the Protocol on this example

(destination = 1)

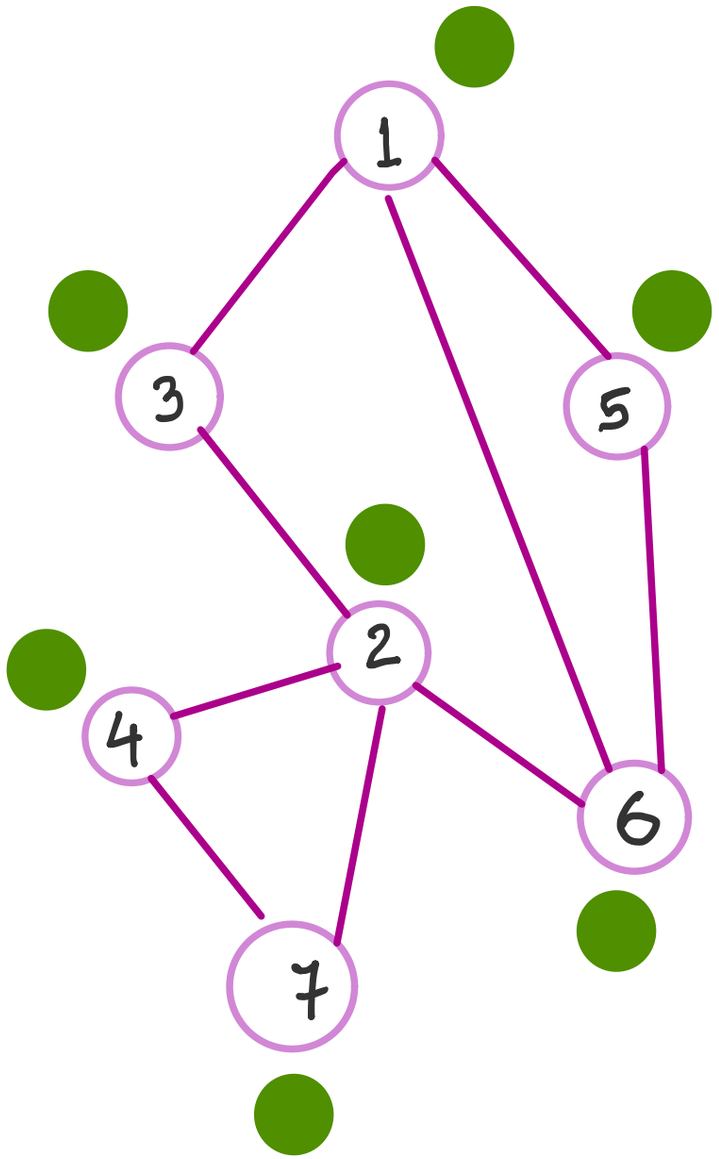


Round 1



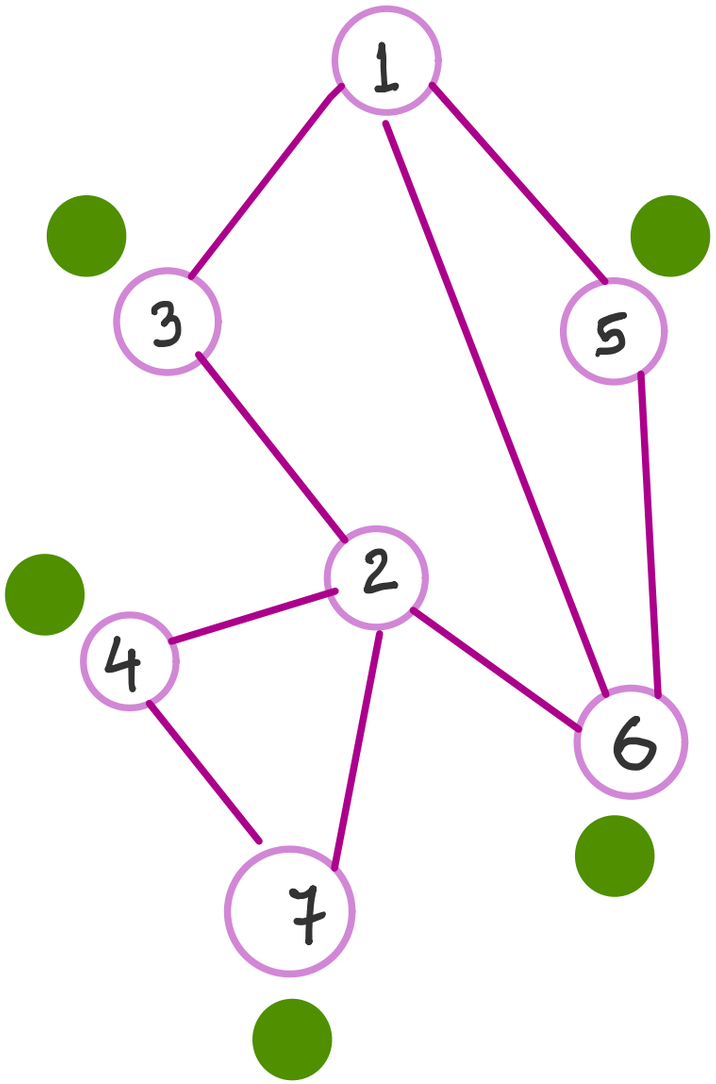
	Receive	Send
1		(1, 0, 1)
2		
3		
4		
5		
6		
7		

Round 2



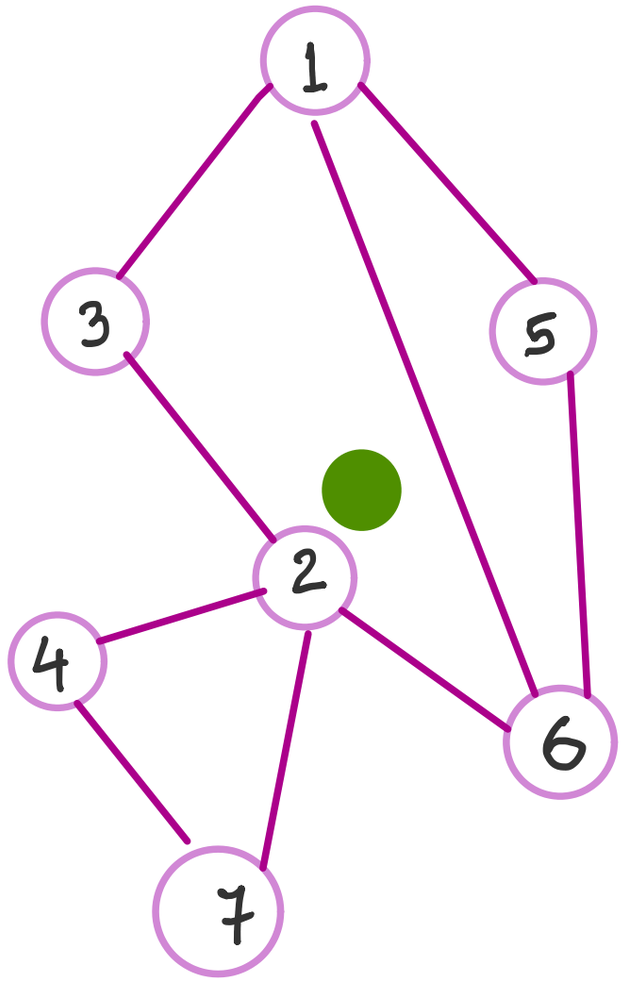
	Receive	Send
1 (1, 0, 1)		
2		
3	(1, 0, 1)	**(1, 1, 3)**
4		
5	(1, 0, 1)	**(1, 1, 5)**
6	(1, 0, 1)	**(1, 1, 6)**
7		

Round 3



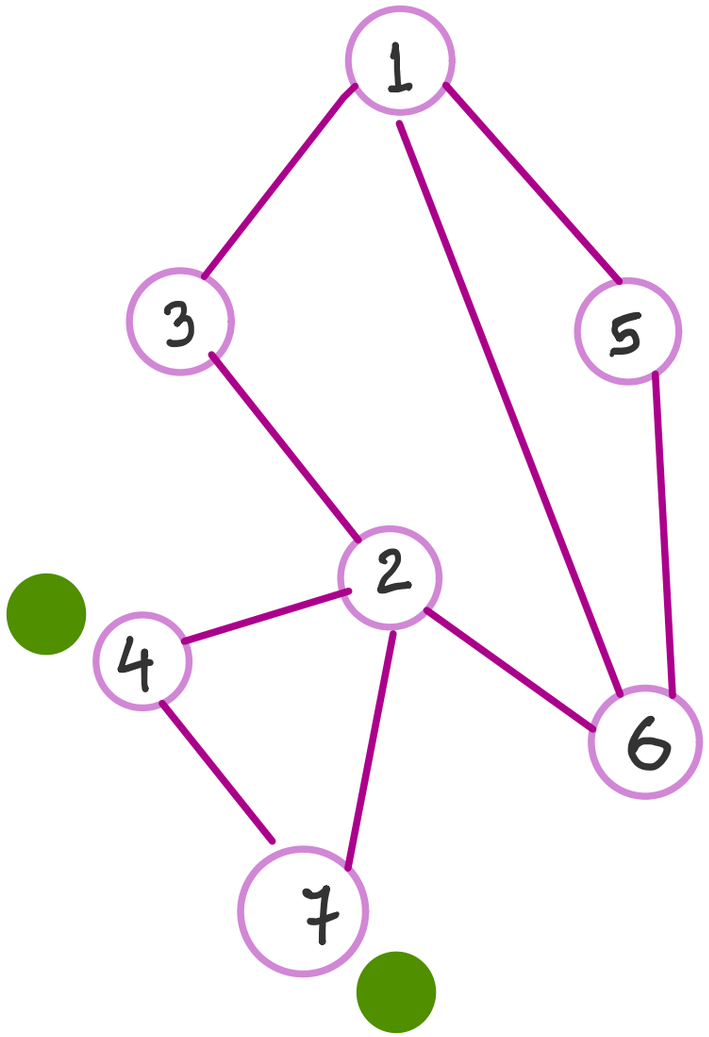
	Receive	Send
1 (1, 0, 1)	(1, 1, 3), (1, 1, 5), (1, 1, 6)	
2	(1, 1, 3), (1, 1, 6)	**(1, 2, 2)**
3 (1, 1, 3)		
4		
5 (1, 1, 5)	(1, 1, 6)	
6 (1, 1, 6)	(1, 1, 5)	
7		

Round 4



	Receive	Send
1 (1, 0, 1)		
2 (1, 2, 2)		
3 (1, 1, 3)	(1, 2, 2)	
4	(1, 2, 2)	**(1, 3, 4)**
5 (1, 1, 5)		
6 (1, 1, 6)	(1, 2, 2)	
7	(1, 2, 2)	**(1, 3, 7)**

Round 5



	Receive	Send
1 (1, 0, 1)		
2 (1, 2, 2)	(1, 3, 4), (1, 3, 7)	
3 (1, 1, 3)		
4 (1, 3, 4)	(1, 3, 7)	
5 (1, 1, 5)		
6 (1, 1, 6)		
7 (1, 3, 7)	(1, 3, 4)	

Why not Spanning Tree Protocol? Why Distance "Vector"?

- The same protocol/algorithm applies to all destinations
- Each node announces distance to **each** dest
    - I am 4 hops away from node A
    - I am 6 hops away from node B
    - I am 3 hops away from node C
    - …
- Nodes are exchanging a **vector** of distances

Towards Distance Vector Protocol (with no failures)

- **Messages (Y,d,X): For root Y; From node X; advertising a distance d to Y**

- Initially each switch X announces (X,0,X) to its neighbors

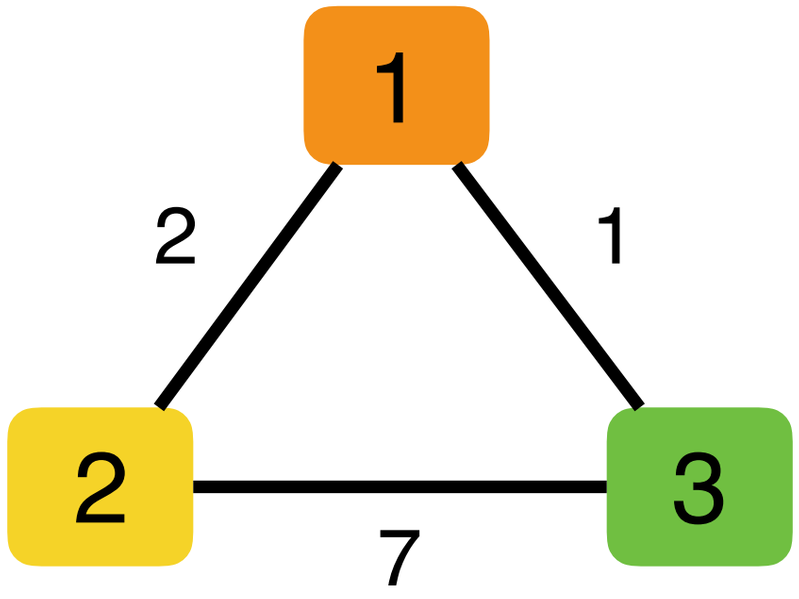
- Switches update their view
    - Upon receiving message (Y,d,Z) from Z~~, check Y's id~~
    - ~~If Y's id < current root: set root destination = Y~~

- Switches compute their shortest distance from the ~~root~~ destination
        - If current_distance_to_Y > d + cost of link to Z:
            - update current_distance_to_Y = d + cost of link to Z

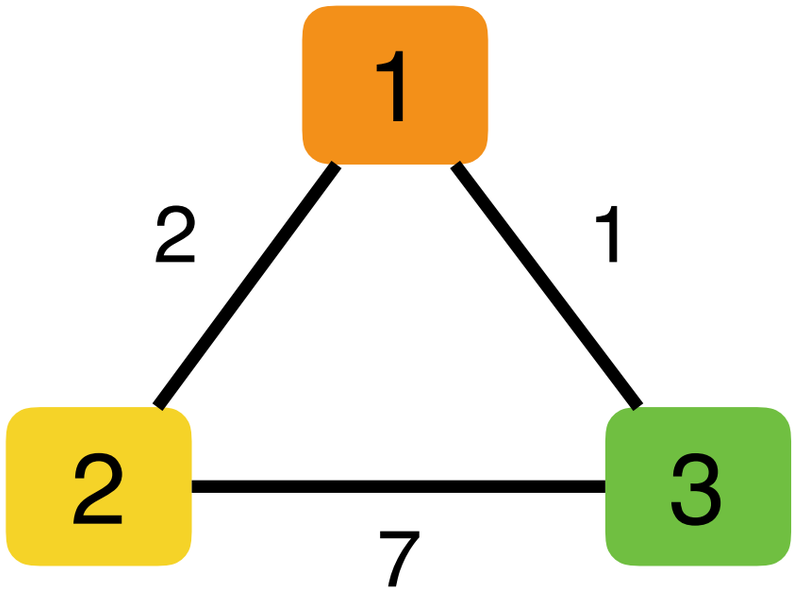
- If ~~root changed OR~~ shortest distance to the ~~root~~ destination changed, send all neighbors updated message (Y,d+c,X)

Group Exercise:

Lets run the Protocol on this example

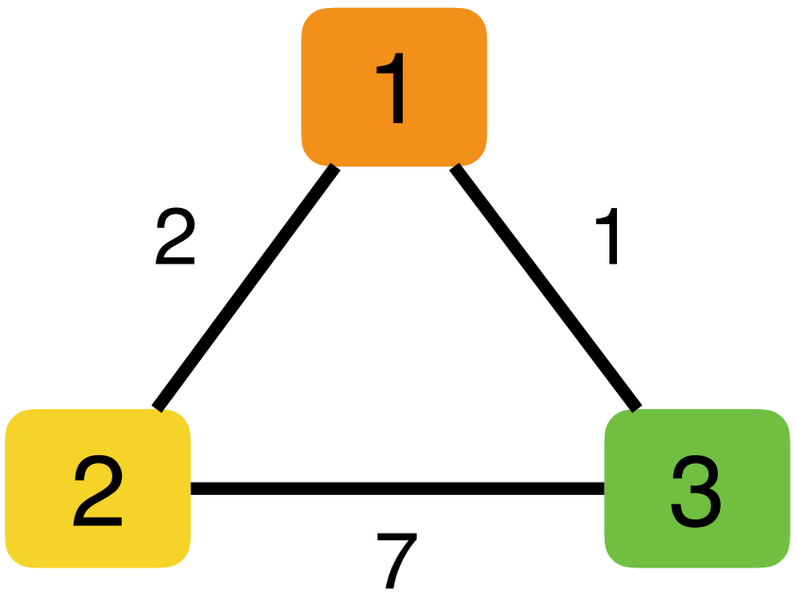


Round 1



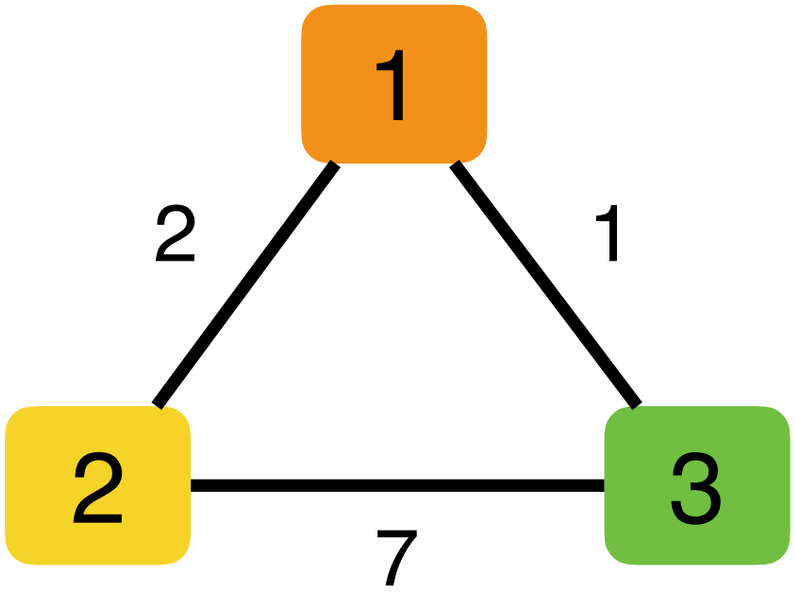
	Receive	Send
1		(1, 0, 1)
2		(2, 0, 2)
3		(3, 0, 3)

Round 2



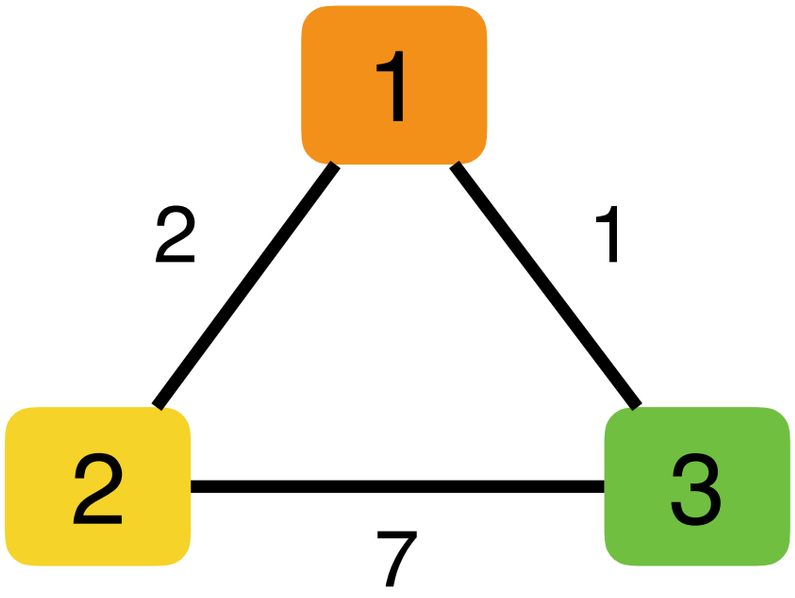
	Receive	Send
1 **(1, 0, 1)**	(2, 0, 2), (3, 0, 3)	(2, 2, 1), (3, 1, 1)
2 **(2, 0, 2)**	(1, 0, 1), (3, 0, 3)	(1, 2, 2), (3, 7, 2)
3 **(3, 0, 3)**	(1, 0, 1), (2, 0, 2)	(1, 1, 3), (2, 7, 3)

Round 3



	Receive	Send
1 **(1, 0, 1)** **(2, 2, 1),** **(3, 1, 1)**	(1, 2, 2), (3, 7, 2), (1, 1, 3), (2, 7, 3)	
2 **(1, 2, 2),** **(2, 0, 2),** **(3, 7, 2)**	(2, 2, 1), (3, 1, 1), (1, 1, 3), (2, 7, 3)	(3, 3, 2)
3 **(1, 1, 3),** **(2, 7, 3),** **(3, 0, 3)**	(2, 2, 1), (3, 1, 1), (1, 2, 2), (3, 7, 2)	(2, 3, 3)

Round 4



	Receive	Send
1 (1, 0, 1) (2, 2, 1), (3, 1, 1)	(3, 3, 2), (2, 3, 3)	
2 (1, 2, 2), (2, 0, 2), (3, 3, 2)	(2, 3, 3)	
3 (1, 1, 3), (2, 3, 3), (3, 0, 3)	(3, 3, 2)	

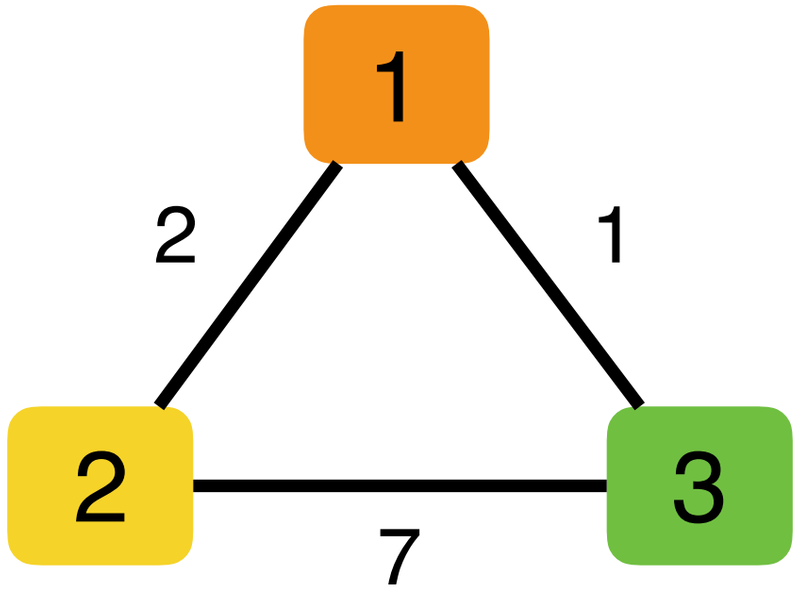
Towards Distance-vector protocol with next-hops (no failures)

- **Messages (Y,d,X): For root Y; From node X; advertising a distance d to Y**
- Initially each switch X announces (X,0,X) to its neighbors
- Switches update their view
  - Upon receiving message (Y,d,Z) from Z~~, check Y's id~~
  - ~~If Y's id < current root: set root destination = Y~~
- Switches compute their shortest distance from the ~~root~~ destination
  - If current_distance_to_Y > d + cost of link to X:
    - update current_distance_to_Y = d
    - **update next_hop_to_destination = X**
- If ~~root changed OR~~ shortest distance to the ~~root~~ destination changed, send all neighbors updated message (Y,d+c,X)

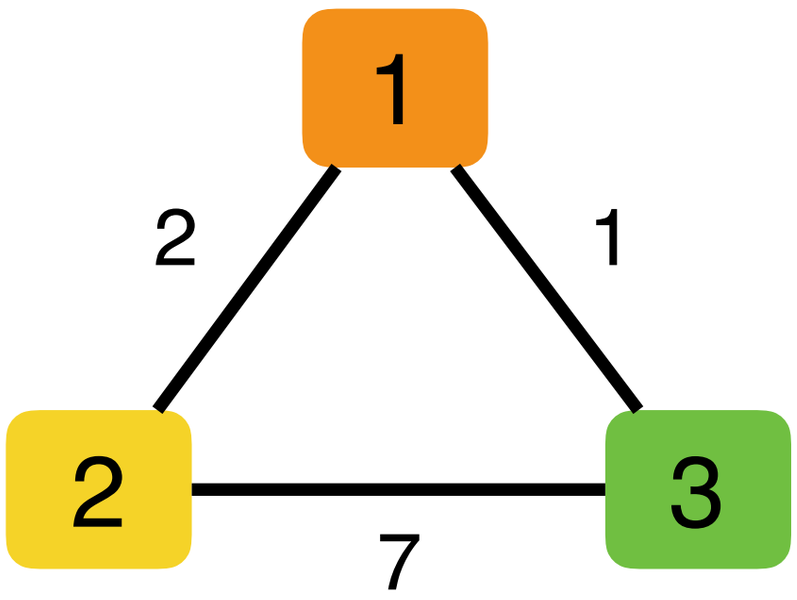
Group Exercise:

Lets run the Protocol on this example

(this time with next-hops)

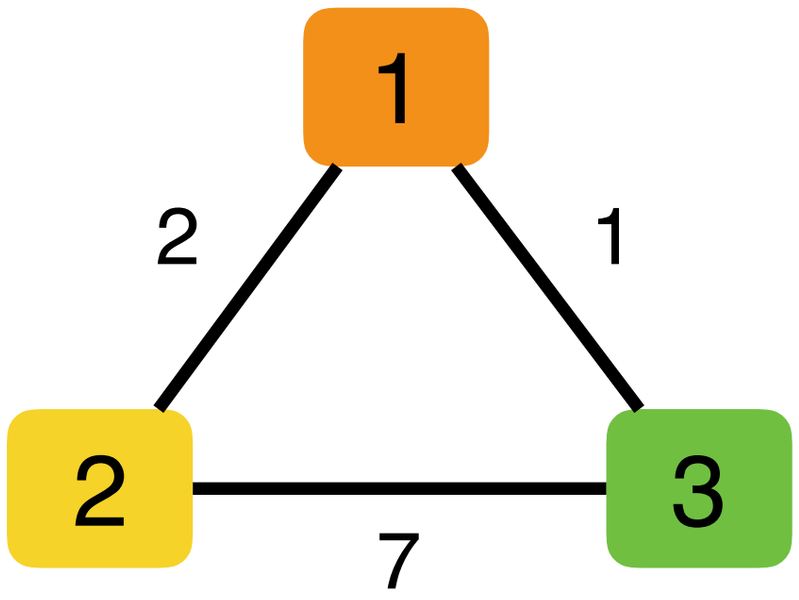


Round 1



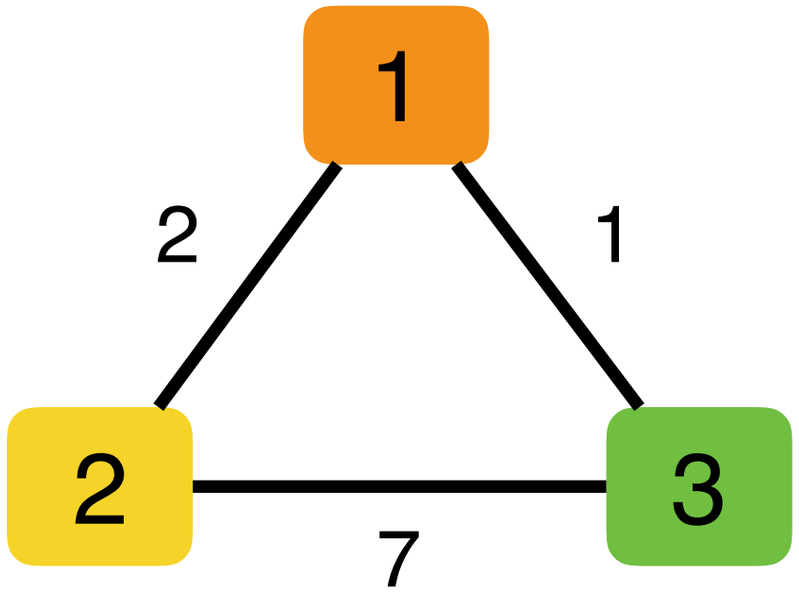
	Receive	Send	Next-hops
1		(1, 0, 1)	[-]
2		(2, 0, 2)	[-]
3		(3, 0, 3)	[-]

Round 2



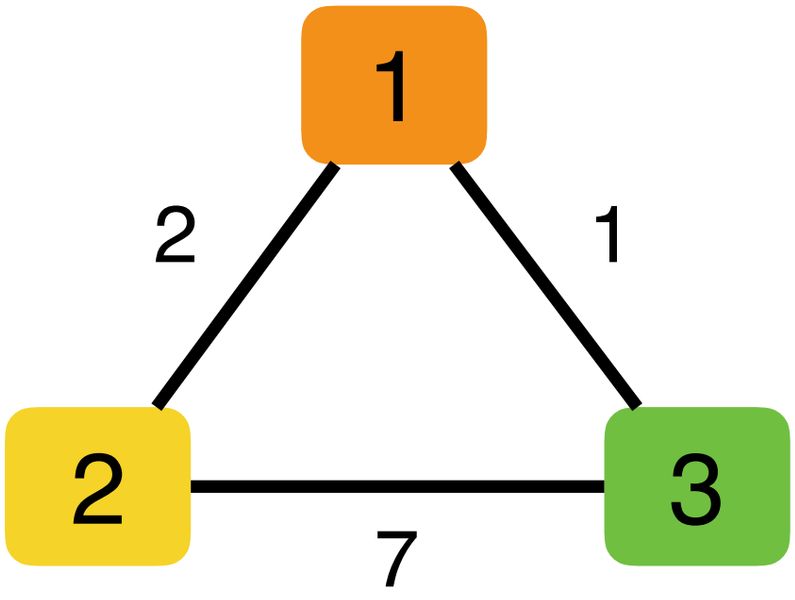
	Receive	Send	Next-hops
1 (1, 0, 1)	(2, 0, 2), (3, 0, 3)	(2, 2, 1), (3, 1, 1)	[-, 2, 3]
2 (2, 0, 2)	(1, 0, 1), (3, 0, 3)	(1, 2, 2), (3, 7, 2)	[1, -, 3]
3 (3, 0, 3)	(1, 0, 1), (2, 0, 2)	(1, 1, 3), (2, 7, 3)	[1, 2, -]

Round 3



	Receive	Send	Next-hops
1 (1, 0, 1) (2, 2, 1), (3, 1, 1)	(1, 2, 2), (3, 7, 2), (1, 1, 3), (2, 7, 3)		[-, 2, 3]
2 (1, 2, 2), (2, 0, 2), (3, 7, 2)	(2, 2, 1), (3, 1, 1), (1, 1, 3), (2, 7, 3)	(3, 3, 2)	[1, -, **1**]
3 (1, 1, 3), (2, 7, 3), (3, 0, 3)	(2, 2, 1), (3, 1, 1), (1, 2, 2), (3, 7, 2)	(2, 3, 3)	[1, **1**, -]

Round 4



	Receive	Send	Next-hops
1 (1, 0, 1) (2, 2, 1), (3, 1, 1)	(3, 3, 2), (2, 3, 3)		[-, 2, 3]
2 (1, 2, 2), (2, 0, 2), (3, 3, 2)	(2, 3, 3)		[1, -, 1]
3 (1, 1, 3), (2, 3, 3), (3, 0, 3)	(3, 3, 2)		[1, 1, -]

Why not Spanning Tree Protocol? Why Distance “Vector”?

- The same algorithm applies to all destinations
- Each node announces distance to **each** dest
    - I am distance d_A away from node A
    - I am distance d_B away from node B
    - I am distance d_C away from node C
    - …
- Nodes are exchanging a **vector** of distances

Distance Vector Protocol

- **Messages (Y,d,X): For root Y; From node X; advertising a distance d to Y**
- Initially each switch X initializes its routing table to (X,0,-) and distance infinity to all other destinations
- Switches announce their entire distance vectors (routing table w/0 next hops)
- Upon receiving a routing table from a node (say X), each node does:
  - For each destination Y in the announcement (distance(X, Y) = d):
    - If current_distance_to_Y > d + cost of link to X:
      - update current_distance_to_Y = d
      - update next_hop_to_destination = X
- If shortest distance to any destination changed, send all neighbors your distance vectors

Two Aspects to This Approach

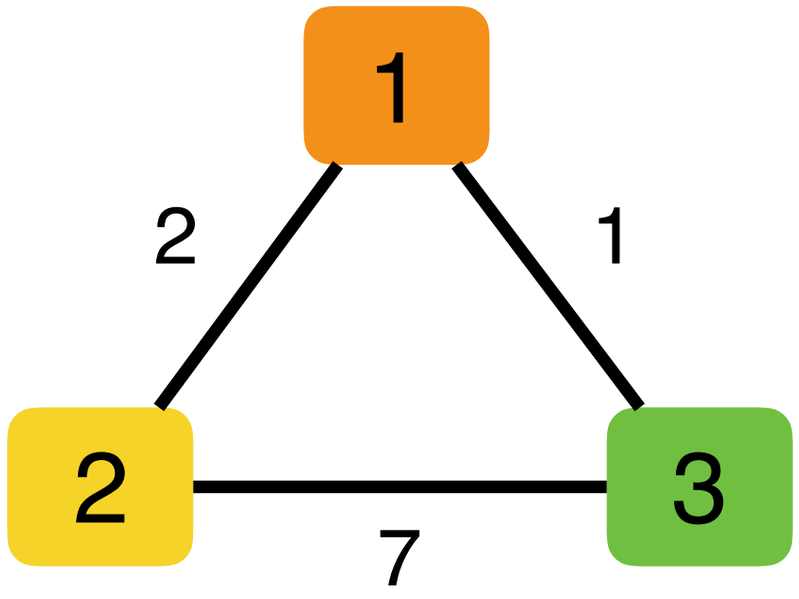
- **Protocol**:
    - Exchanging that routing information with neighbors
    - What and when for exchanges
    - RIP is a protocol that implements DV (IETF RFC 2080)

- **Algorithm**:
    - How to use the information from your neighbors to update your own routing tables?

Group Exercise:

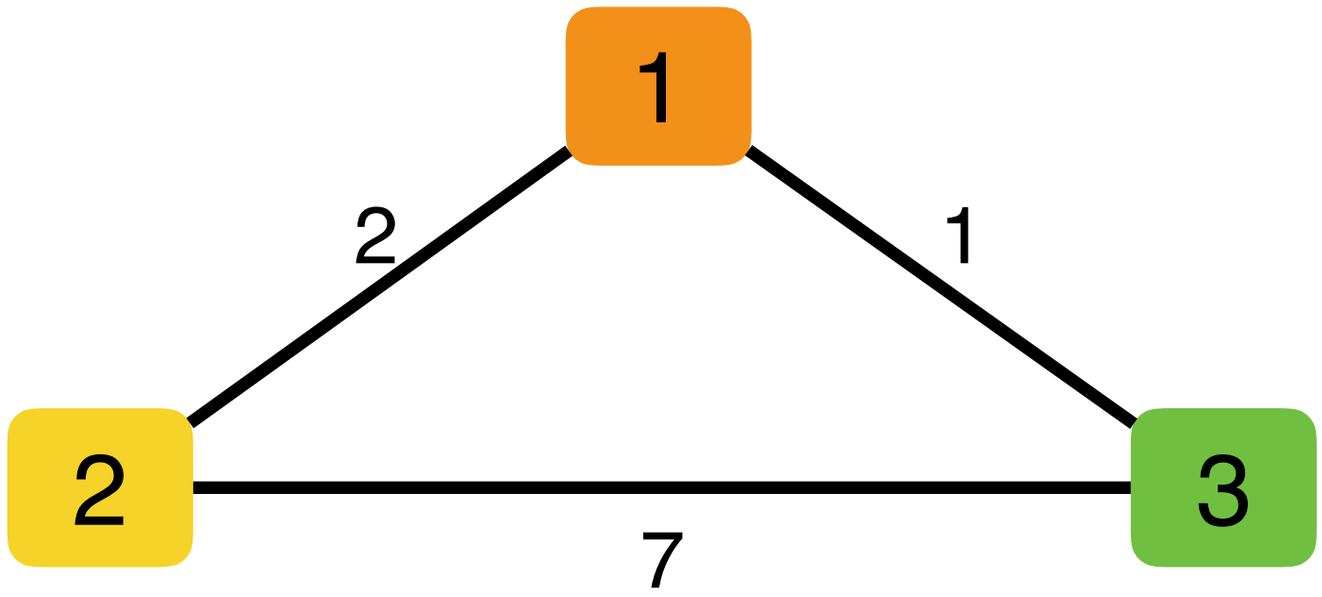
Lets run the Protocol again on this example

(this time with distance vectors)



Round 1

	distance	next-hop
1	0	-
2	infinity	
3	infinity	

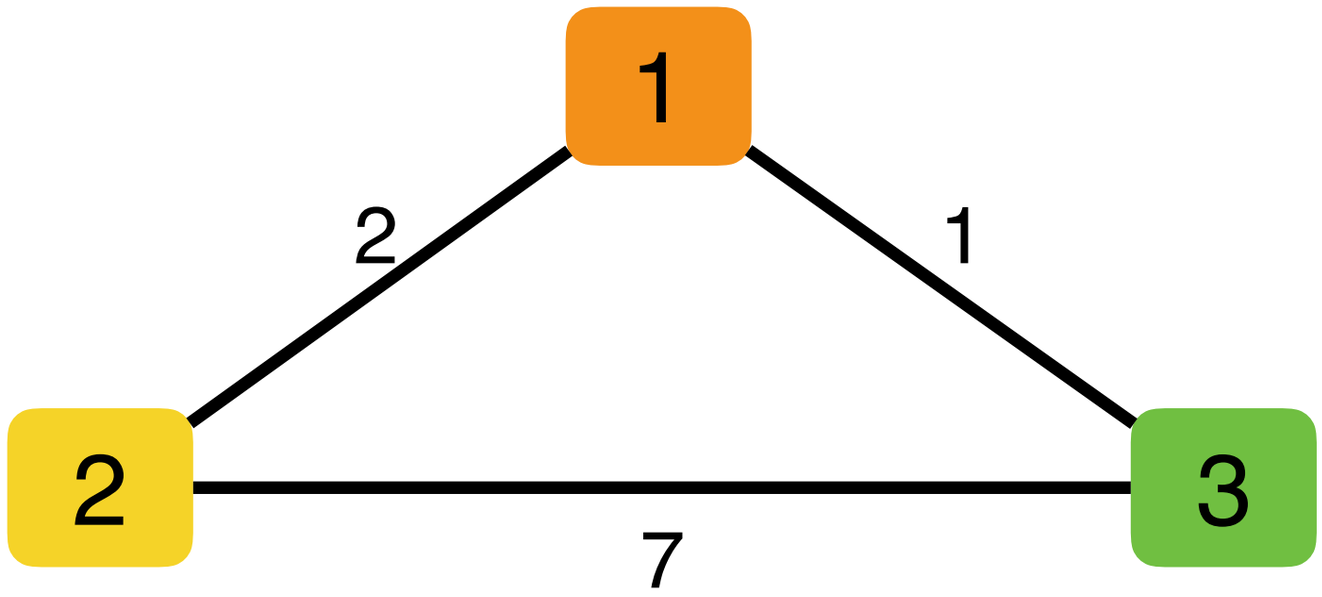


	distance	next-hop
1	infinity	
2	0	-
3	infinity	

	distance	next-hop
1	infinity	
2	infinity	
3	0	-

Round 2

	distance	next-hop
1	0	-
2	**2**	**2**
3	**1**	**3**

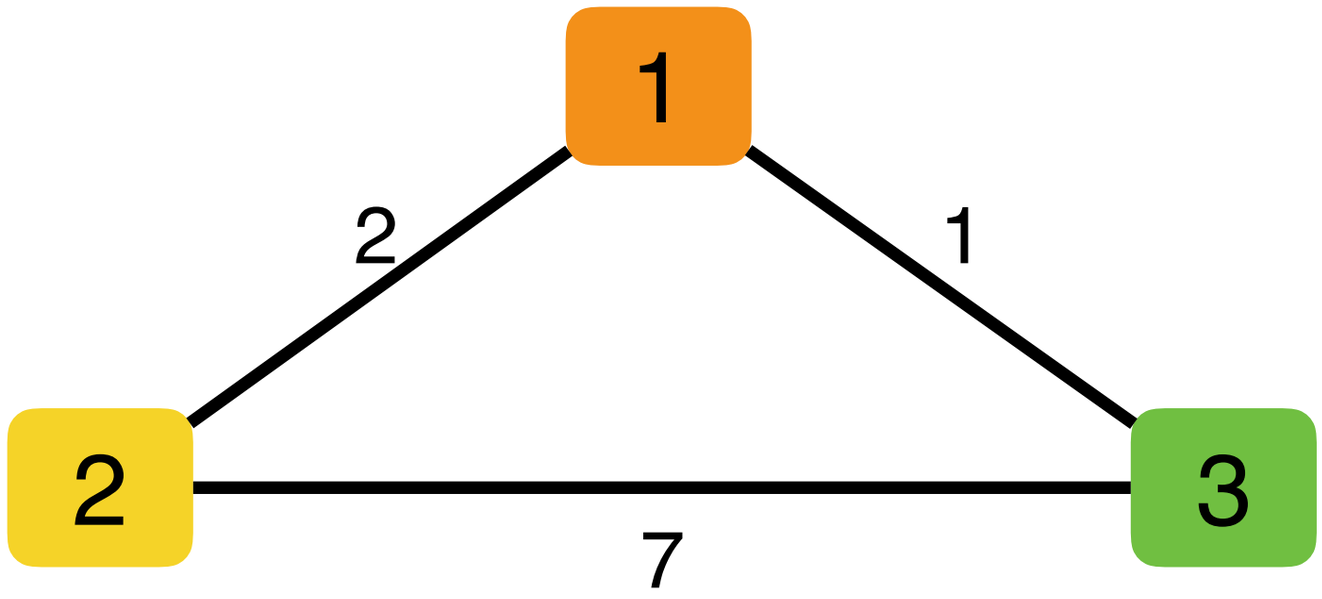


	distance	next-hop
1	**2**	**1**
2	0	-
3	**7**	**3**

	distance	next-hop
1	**1**	**1**
2	**7**	**2**
3	0	-

Round 3

	distance	next-hop
1	0	-
2	2	2
3	1	3

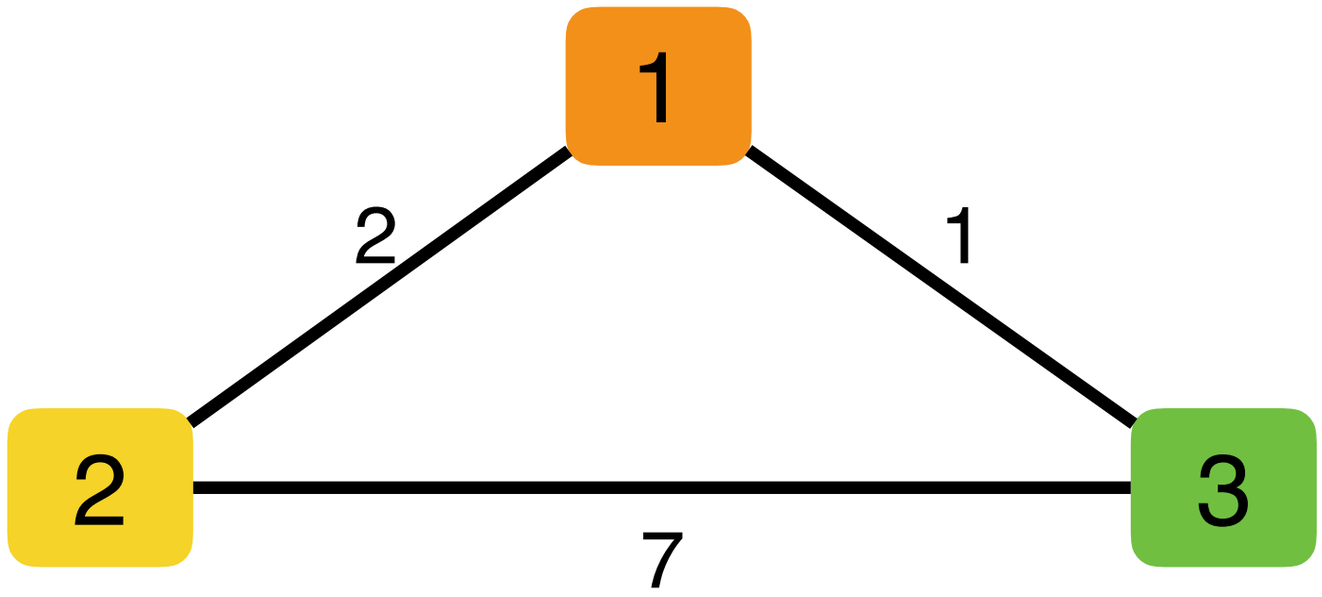


	distance	next-hop
1	2	1
2	0	-
3	**3**	**1**

	distance	next-hop
1	1	1
2	**3**	**1**
3	0	-

Round 4

	distance	next-hop
1	0	-
2	2	2
3	1	3



	distance	next-hop
1	2	1
2	0	-
3	3	1

	distance	next-hop
1	1	1
2	3	1
3	0	-

From Algorithm to Protocol

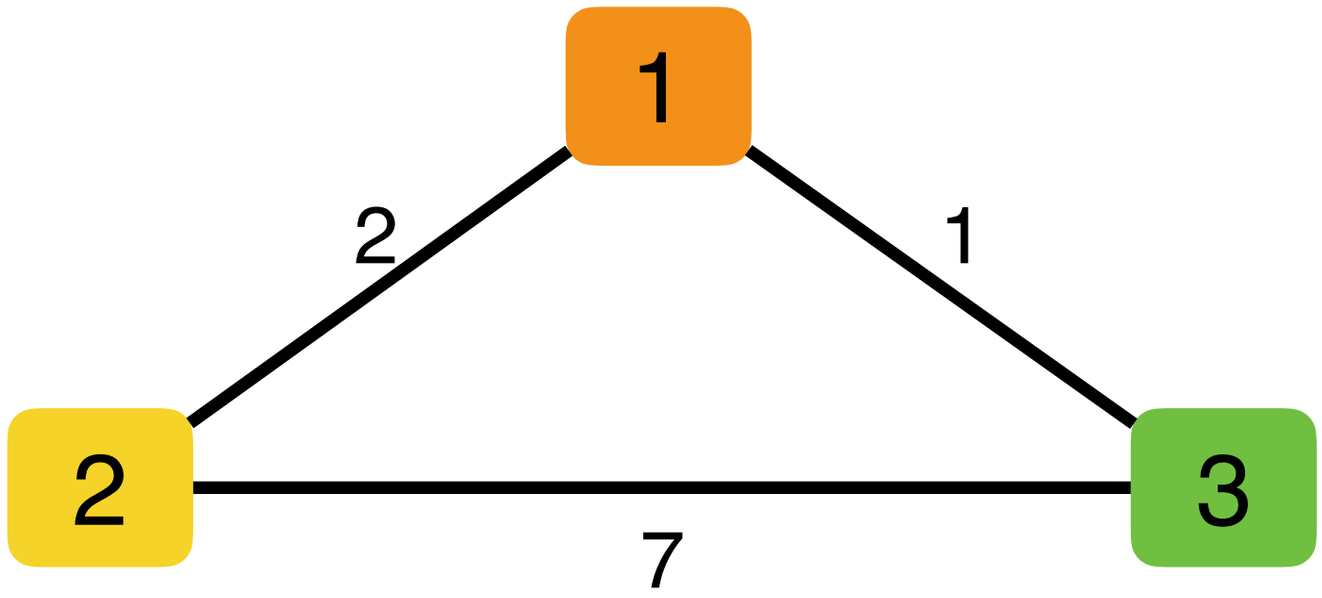
- Algorithm:
    - Nodes use Bellman-Ford to compute distances
- Protocol
    - Nodes exchange distance vectors
    - Update their own routing tables
    - And exchange again…
    - Details: when to exchange, what to exchange, etc….

Other Aspects of Protocol

- When do you send messages?
    - When any of your distances d(u,v) change
        - What about when c(u,v) changes?
    - Periodically, to ensure consistency between neighbors
- What information do you send?
    - Could send entire vector
    - Or just updated entries
- Do you send everyone the same information
    - Consider the following slides

Three node network

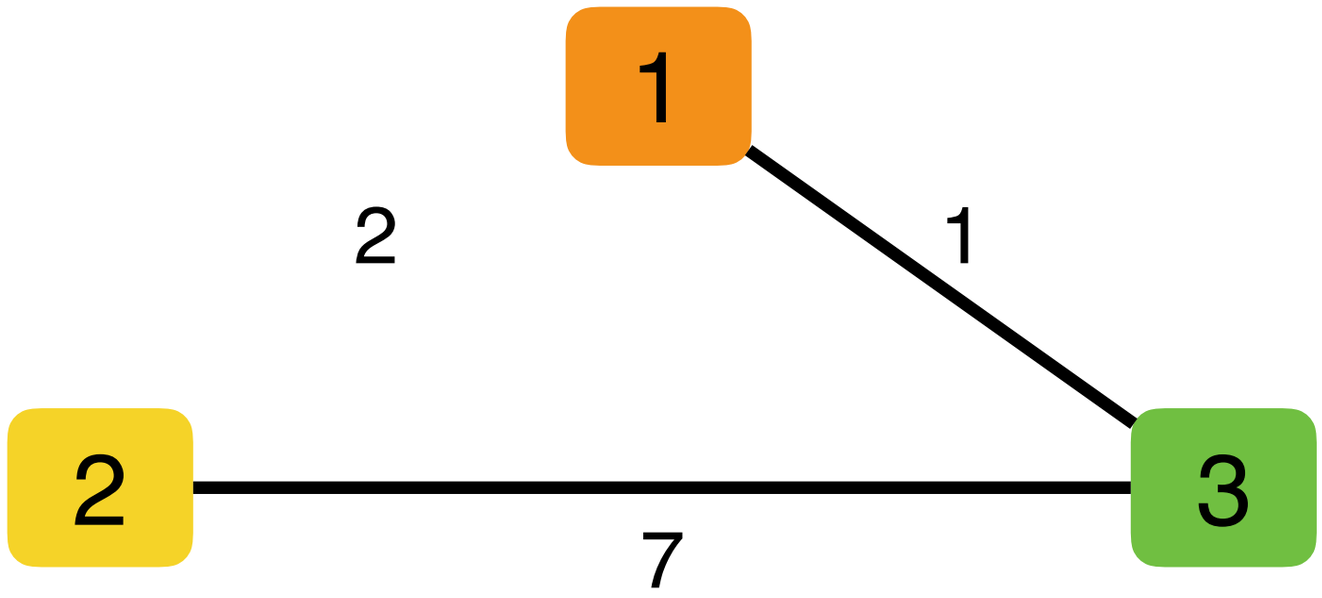
	distance	next-hop
1	0	-
2	2	2
3	1	3



	distance	next-hop
1	1	1
2	3	1
3	0	-

Three node network

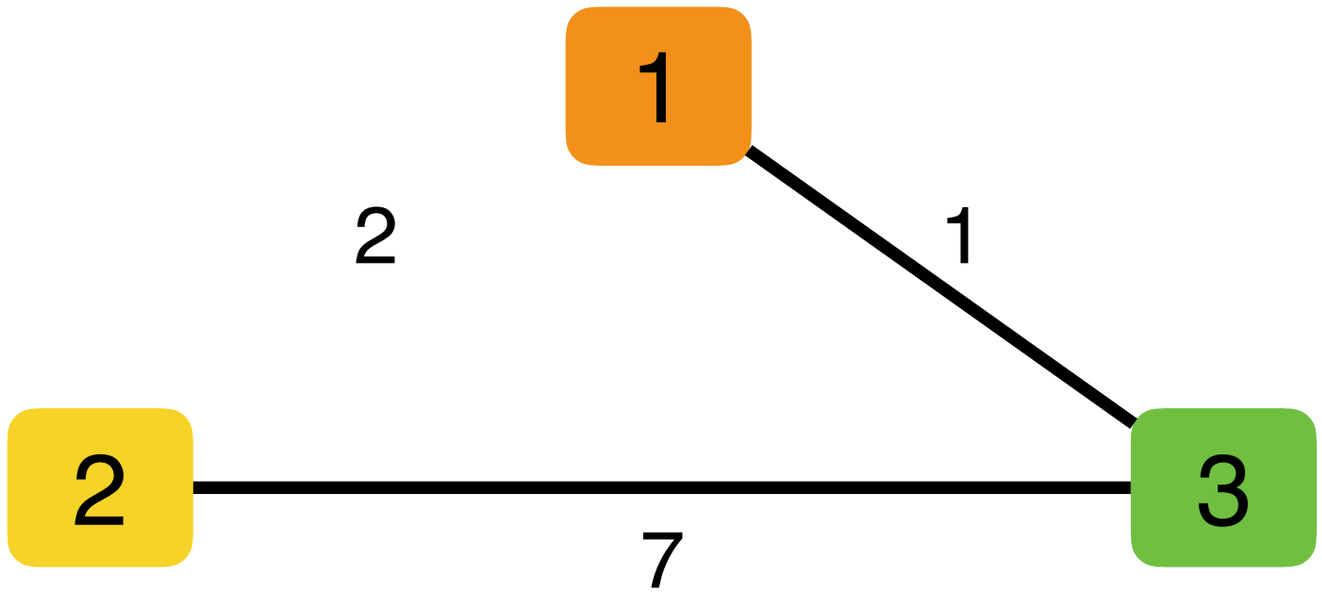
	distance	next-hop
1	0	-
2	infinity	
3	1	3



	distance	next-hop
1	1	1
2	3	1
3	0	-

Round 1

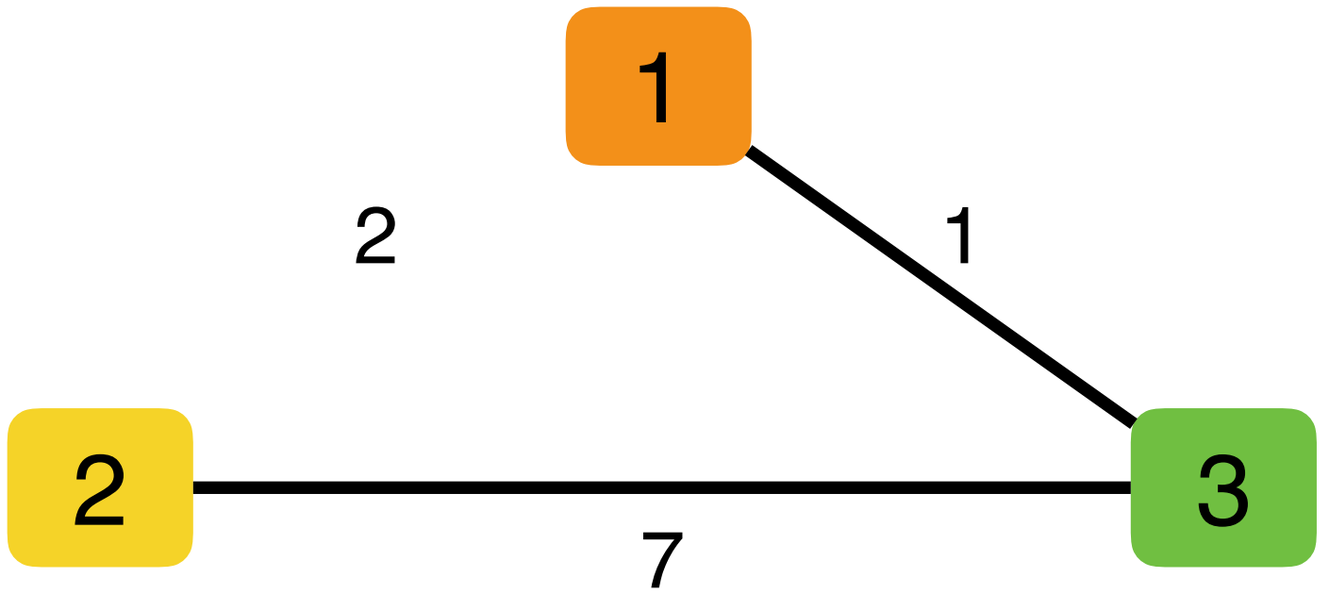
	distance	next-hop
1	0	-
2	**4**	**3**
3	1	3



	distance	next-hop
1	1	1
2	3	1
3	0	-

Round 2

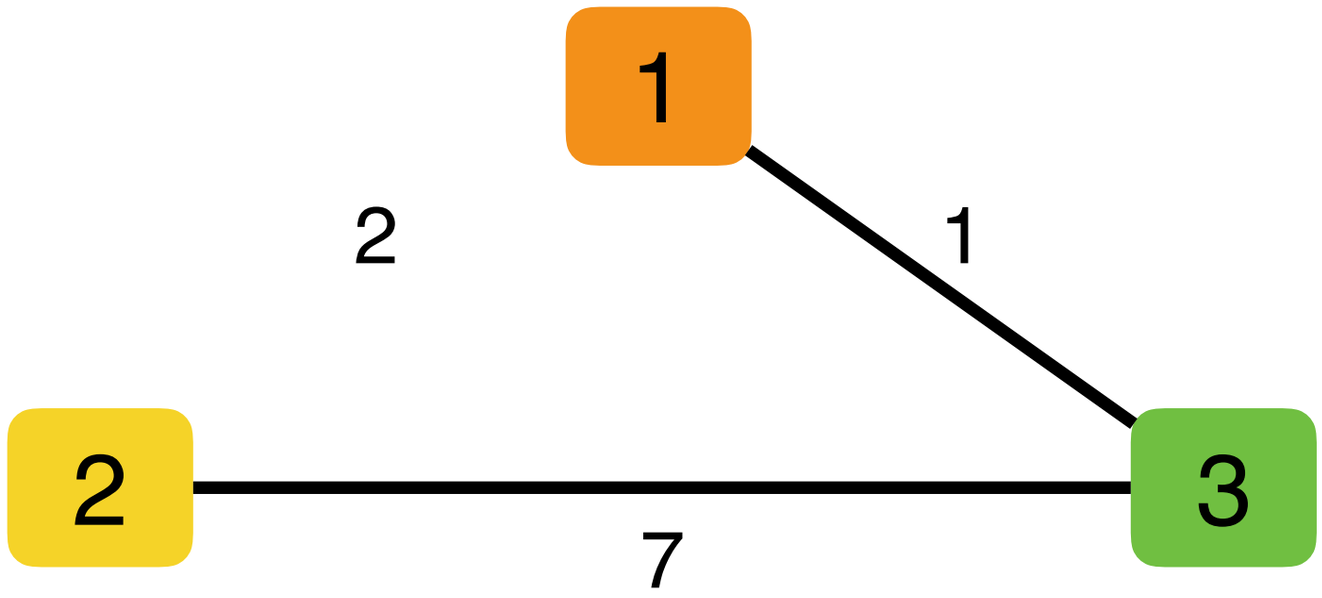
	distance	next-hop
1	0	-
2	4	3
3	1	3



	distance	next-hop
1	1	1
2	**5**	**1**
3	0	-

Round 3

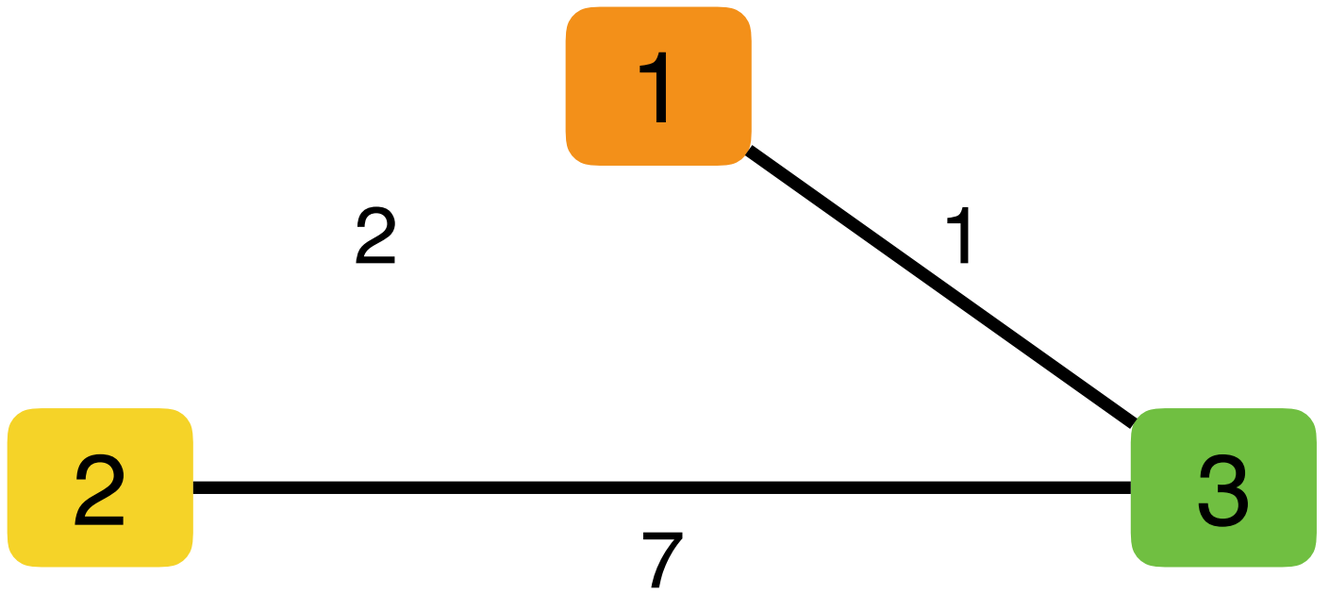
	distance	next-hop
1	0	-
2	**6**	**3**
3	1	3



	distance	next-hop
1	1	1
2	5	1
3	0	-

Round 4

	distance	next-hop
1	0	-
2	6	3
3	1	3

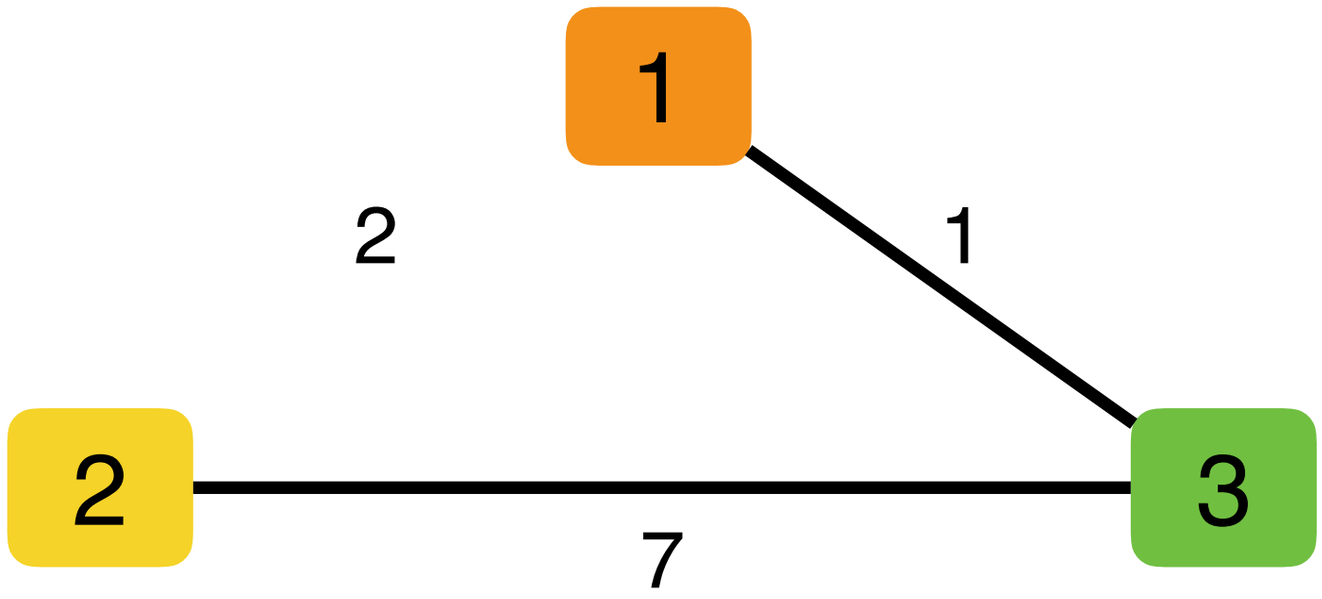


COUNT-TO-INFINITY problem!!!!

	distance	next-hop
1	1	1
2	7	1
3	0	-

Count-to-infinity problem

	distance	next-hop
1	0	-
2	6	3
3	1	3



Not just due to failures:
Can happen with changes in cost!

	distance	next-hop
1	1	1
2	7	1
3	0	-

How Can You Fix This?

- **Do not advertise a path back to the node that is the next hop on the path**
    - Called **“split horizon”**
    - Telling them about your entry going through them
        - Doesn’t tell them anything new
        - Perhaps misleads them that you have an independent path
- **Another solution: if you are using a next-hop’s path, then**:
    - Tell them not to use your path (by telling them cost of infinity)
    - Called “**poisoned reverse**”

Convergence

- Distance vector protocols can converge slowly
    - While these corner cases are rare
    - The resulting convergence delays can be significant

Comparison of Scalability

- Link-State:
    - Global flood: each router's link-state (#ports)
    - Send it once per link event, or periodically
- Distance Vector:
    - Send longer vector (#dest) just to neighbors
        - But might end up triggering their updates
    - Send it every time DV changes (which can be often)
- Tradeoff:
    - LS: Send it everywhere and be done in predictable time
    - DV: Send locally, and perhaps iterate until convergence

End of Distance-vector Routing

Now you know just as much as my PhD students :-)