Advanced Routing

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Xuetao Wei

weixt@sustech.edu.cn

Contents

- Routing Basics
- Intra-AS Routing
- Inter-AS Routing
- Multicast Routing
- MPLS

rouger：1.control plane 2. data plane

quiz考前三个

广东深圳电信(mcplay.cn) | ICMP | www.acm.org

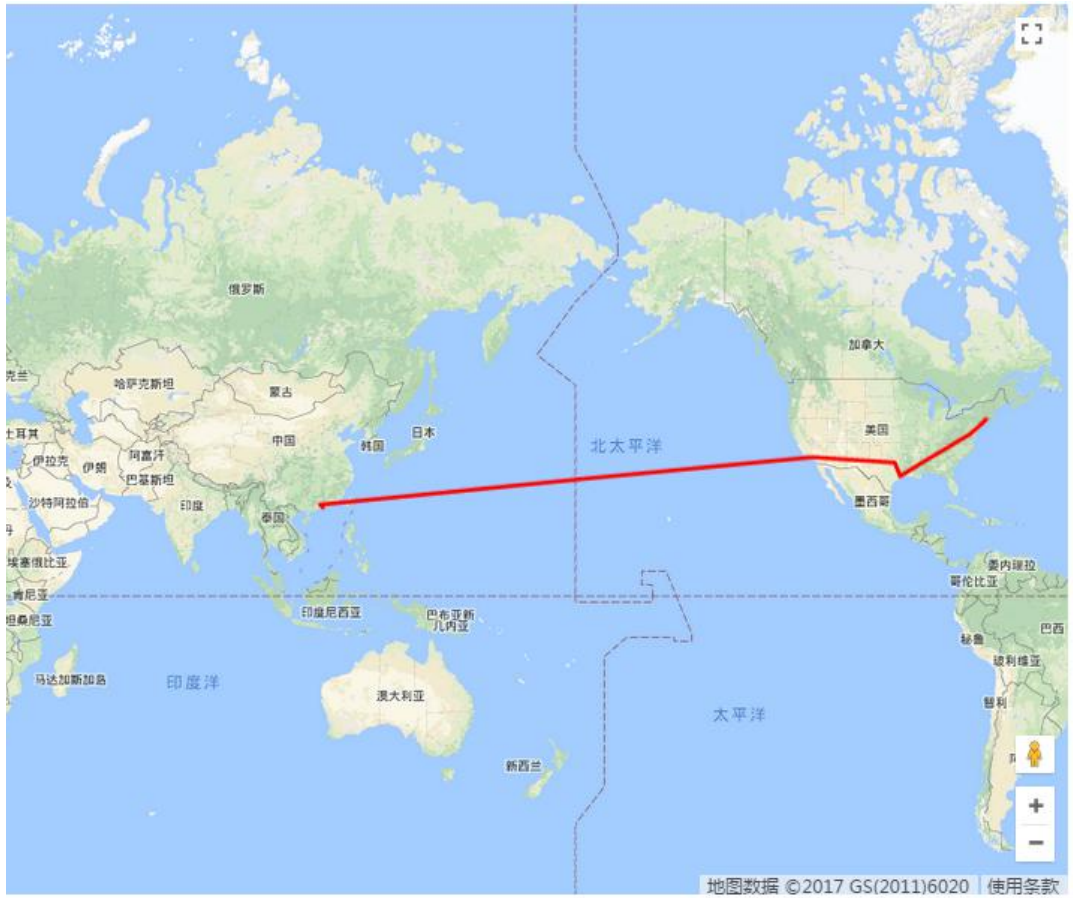
查看

traceroute

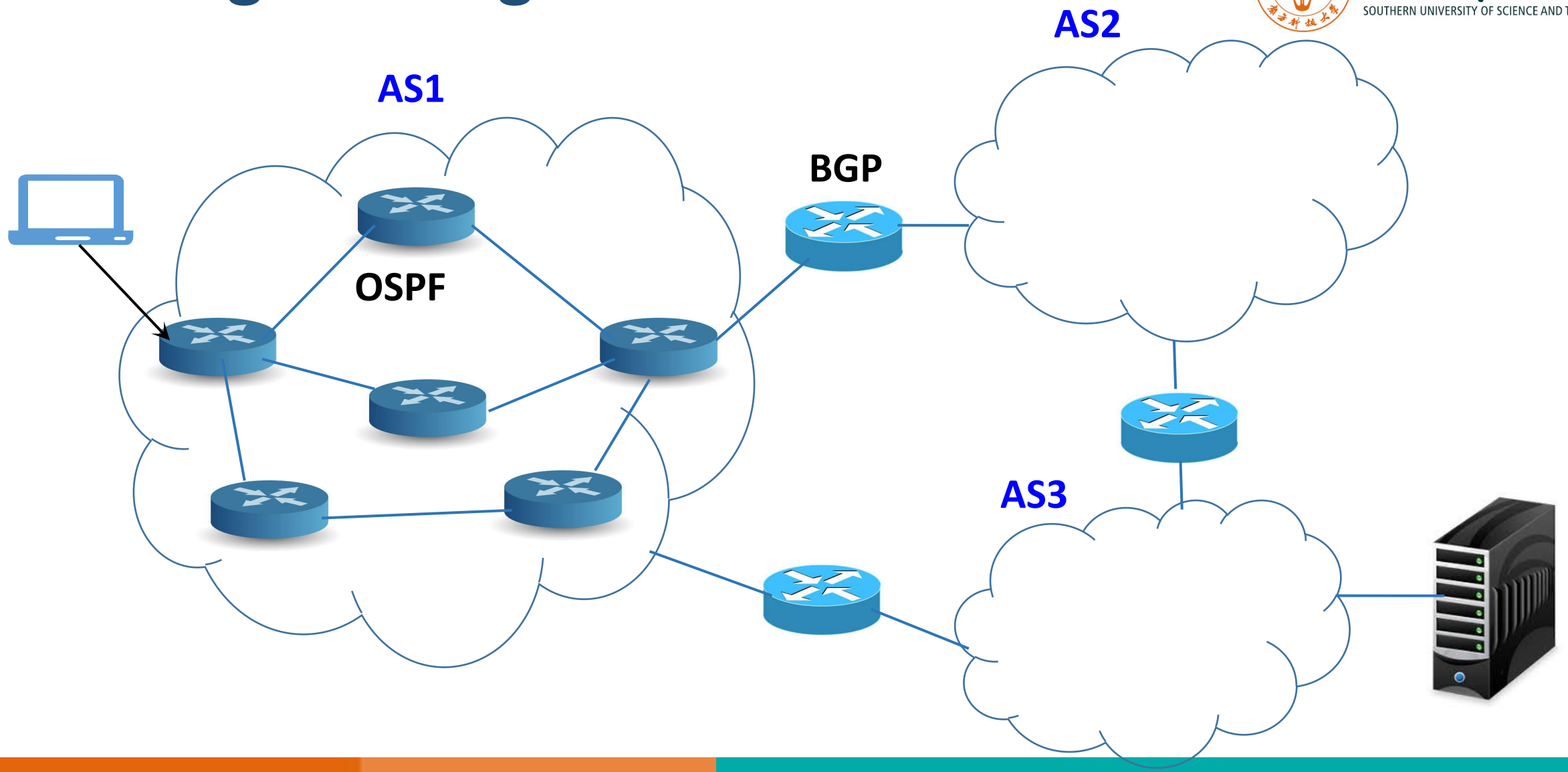
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

目标 IP：64.238.147.121　　监测点赞助商：速云

跳数	IP	主机名	地区（仅供参考）	AS号（仅供参考）	时间（毫秒）
1	172.17.11.254	172.17.11.254	局域网		1.2 / 1.1 / 1
2	172.17.2.1	172.17.2.1	局域网		0.4 / 0.8 / 1.2
3	61.144.170.169	61.144.170.169	广东深圳 电信	AS4134	1.3 / 1 / 1.8
4	219.133.30.82	219.133.30.82	广东深圳 电信	AS4134	2.1 / 2.1 / 1.8
5	58.60.24.41	58.60.24.41	广东深圳 电信	AS4134	2 / 2 / 2.1
6	183.56.65.70	183.56.65.70	广东深圳 电信	AS4134	7.6 / 5.3 / 7.1
7	*	*	N/A	*	*
8	202.97.94.114 * *	202.97.94.114 * *	广东广州 电信 N/A N/A	AS4134 * *	7.6 * *
9	202.97.51.106	202.97.51.106	美国加利福尼亚州洛杉矶 电信	AS4134	249.1 / 247.1 / 248.8
10	202.97.49.154	202.97.49.154	美国加利福尼亚州洛杉矶 电信	AS4134	282.1 / 280.7 / 277.7
11	218.30.54.174	218.30.54.174	美国加利福尼亚州洛杉矶 CHINATELECOM	AS4134	209 / 210.2 / 197.5
12	64.125.28.230	ae13.cs1.lax112.us.eth.zayo.com	美国加利福尼亚州洛杉矶 zayo.com	AS6461	284.9 / 277.2 / 284.6
13	64.125.29.52	ae3.cs1.dfw2.us.eth.zayo.com	美国德克萨斯州达拉斯 zayo.com	AS6461	271.1 / 281.3 / 248.1
14	64.125.28.98 64.125.28.98 *	ae5.cs1.iah1.us.eth.zayo.com ae5.cs1.iah1.us.eth.zayo.com *	美国德克萨斯州休斯顿 zayo.com 美国德克萨斯州休斯顿 zayo.com N/A	AS6461 AS6461 *	320.5 320.3 *
15	64.125.29.48	ae3.cs1.dca2.us.eth.zayo.com	美国华盛顿 zayo.com	AS6461	261.4 / 257.2 / 269.4
16	64.125.29.202 204.16.61.21 204.16.61.21	ae4.cs1.lga5.us.eth.zayo.com 204.16.61.21 204.16.61.21	美国纽约州纽约 zayo.com 美国康涅狄格州 cyrusone.com 美国康涅狄格州 cyrusone.com	AS6461 AS62 AS62	343.5 315.5 306.1
17	64.125.22.9	ae12.mpr1.bdl4.us.zip.zayo.com	美国康涅狄格州哈特福德 zayo.com	AS6461	320.4 / 320.4 / 320.4
18	64.238.144.18	64.238.144.18	美国 cyrusone.com	AS19479	375.4 / 374.8 / 372.9
19	204.16.61.21	204.16.61.21	美国康涅狄格州 cyrusone.com	AS62	262 / 270.5 / 266.9
20	64.238.147.121	64.238.147.121	美国 cyrusone.com	AS19479	375.4 / 380.4 / 377

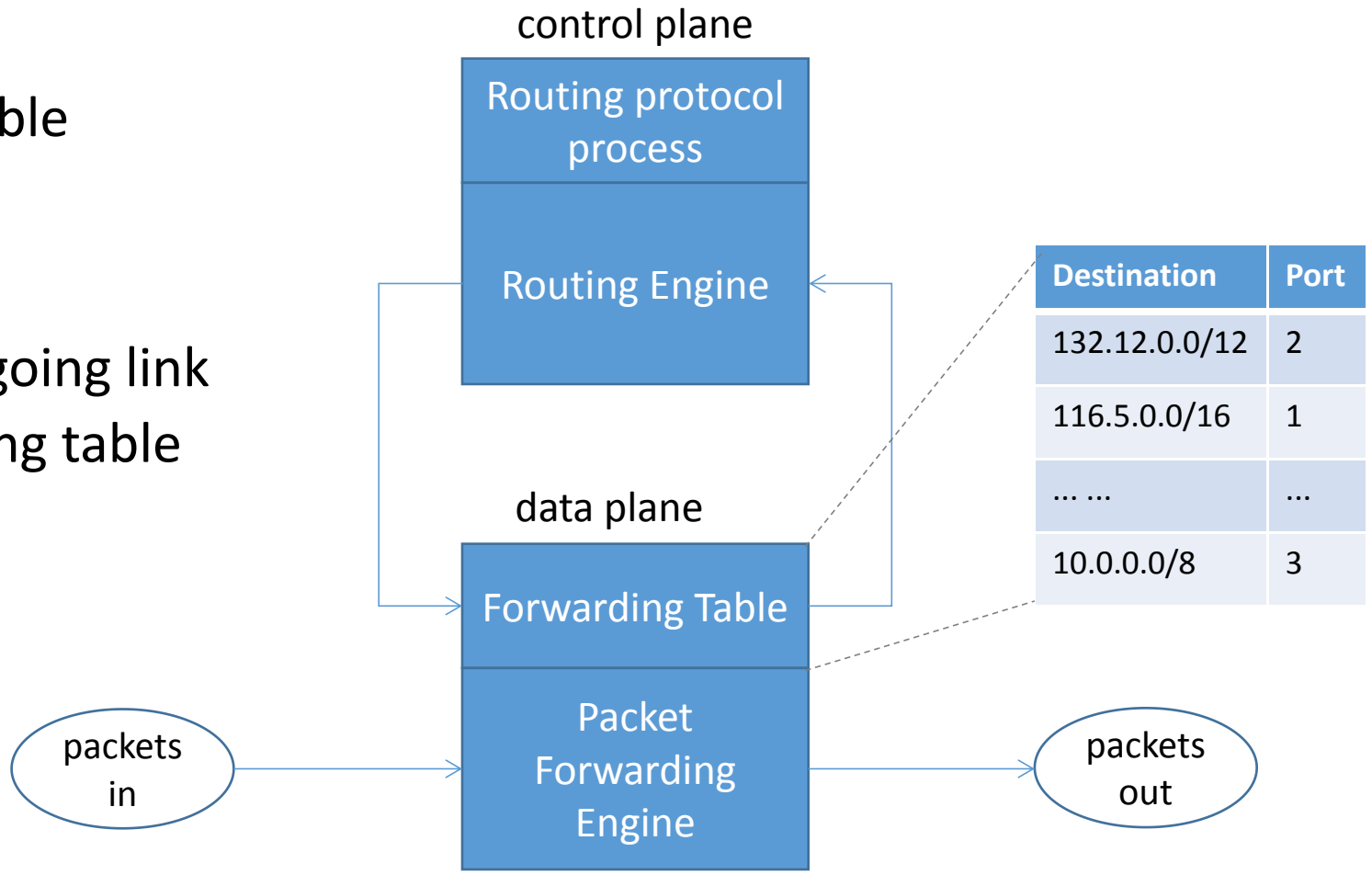


Routing: The Big Picture



Router

- Routing: control plane
  - Creating a forwarding table
  - Talking to other routers
- Forwarding: data plane
  - Send a packet to an outgoing link
  - Looking up the forwarding table

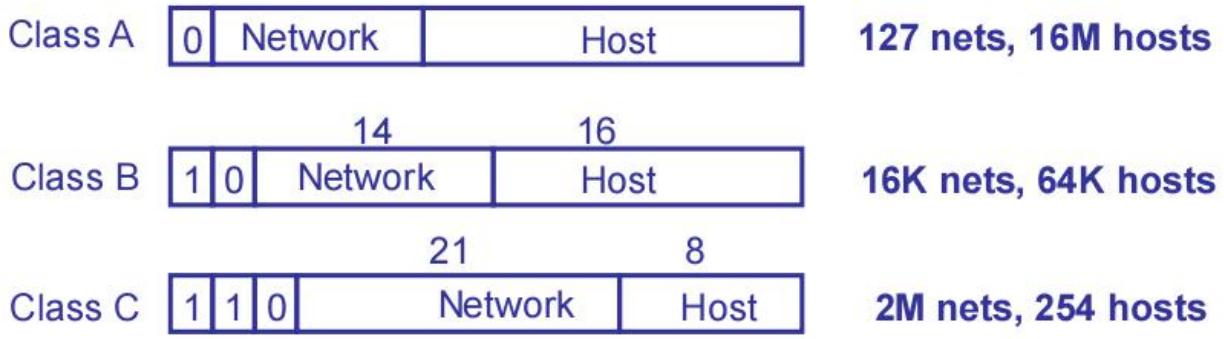


IP Addressing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- IP address: hierarchial or flat?
  - IPv4: dotted decimal format (a.b.c.d)
  - Internet is inherently hierarchical: Multiple ASes -> AS -> subnet
  - IP address: network + host

- Class-based addressing: class + network + host



Class D (1110) for multicast
Class E (1111) for experimental

CIDR Addressing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- CIDR: Classless Inter-Domain Routing
  - Allows arbitrary legnth of network prefix

  - e.g. 132.12.1.8/12 (**10000100 0000**1100 00000001 00001000)

Network Host

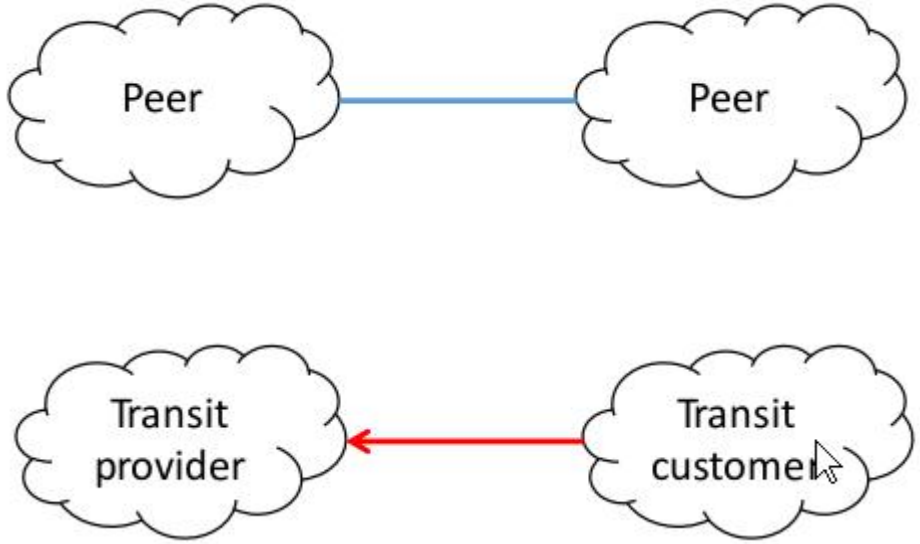
  - Pros: flexible, better utilization of IP resources
  - Cons: more complicated lookup: longest prefix match

Autonomous System

- Autonomous System (AS)
  - Unit of routing policy
  - ~58k ASes in use
  - E.g., AT&T has AS#144, Princeton has AS#88, CHINANET(backbone) has AS#4134

- AS relationships
  - Peering: jointly pay for costs

  - Transit: customer pay for upstream and downstream traffic



Routing Protocols in Scenarios

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Intra-AS routing
  - Distance Vector Algorithms (RIP)
  - Link State Algorithms (OSPF)

- Inter-AS routing
  - Border Gateway Protocol (BGP)

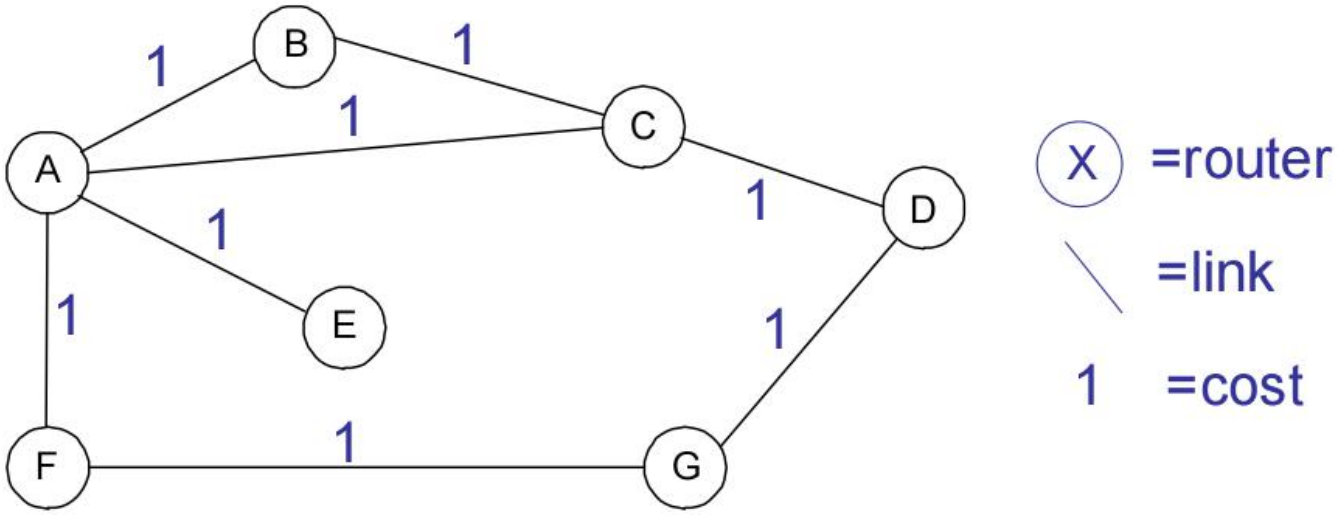
Contents

- Routing Basics
- Intra-AS Routing
- Inter-AS Routing
- Multicast Routing
- MPLS

Intra-AS Routing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Routing
  - Essentially a graph theory problem
  - Find “best” path between every pair of vertices (routers)



Global vs Decentralized Routing

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **Global routing algorithm**
  - Global view and complete knowledge of the network topology
  - Complete path for every pair of nodes
  - Also known as Link State routing (e.g., OSPF)

- **Decentralized algorithm**
  - Local view on the network
  - No complete path at each node, only the next hop to destination
  - Also known as Distance Vector routing (e.g., RIP)

1. Distance Vector Routing

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- A distance vector $\{d_x(y) \mid y \in N\}$ at each node x
  - $d_x(y)$ = cost of least-cost path from X to Y

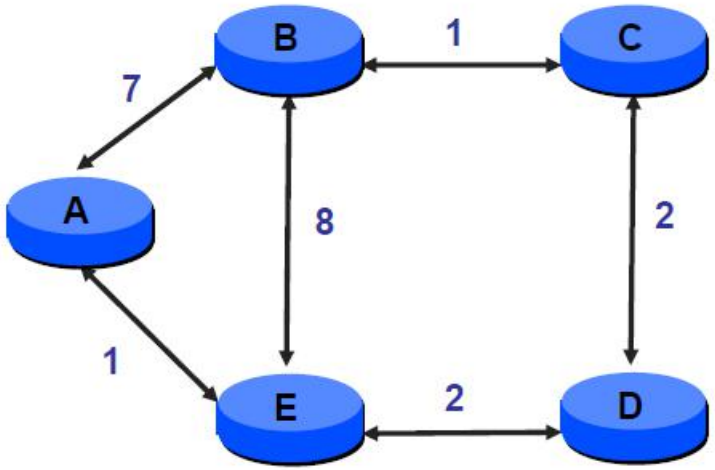
- Rippling vs flooding
  - Only notify neighbors on its updates, no explicit flooding to all nodes
  - But if a neighbor y gets updated as well, the update is rippled to y's neighbors

- Bellman-Ford Algorithm
  - $d_x(y) = \min \{c(x,v) + d_v(y)\}$ over all neighbors V

Example: Initial State

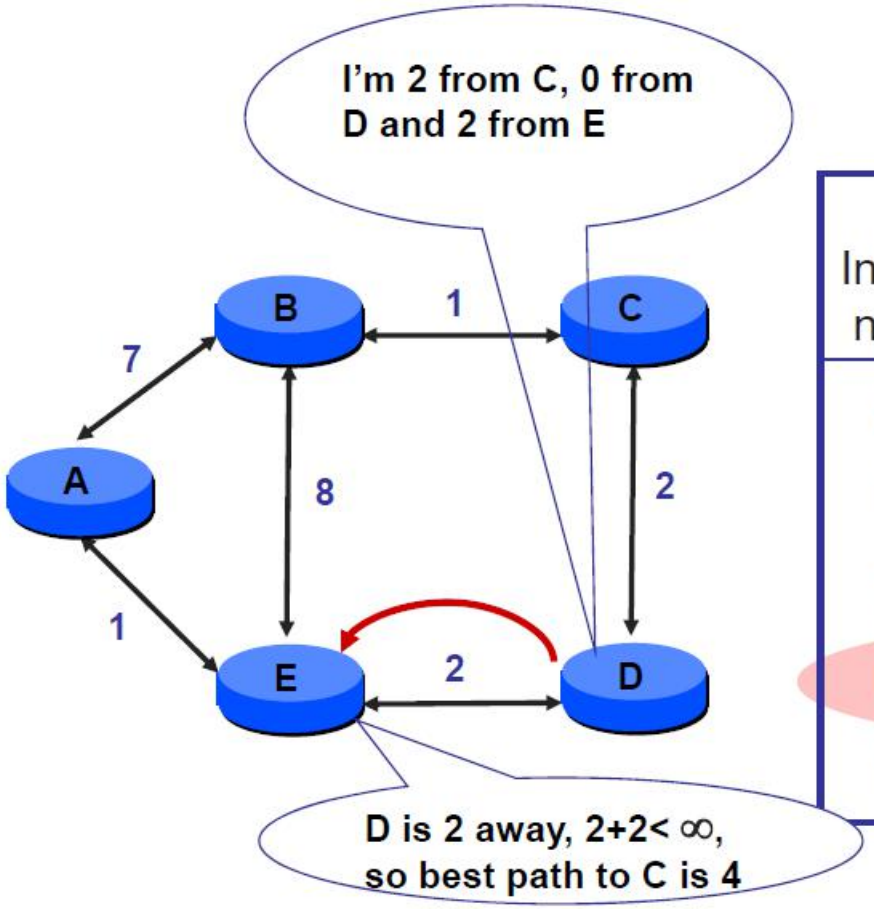
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

可能考



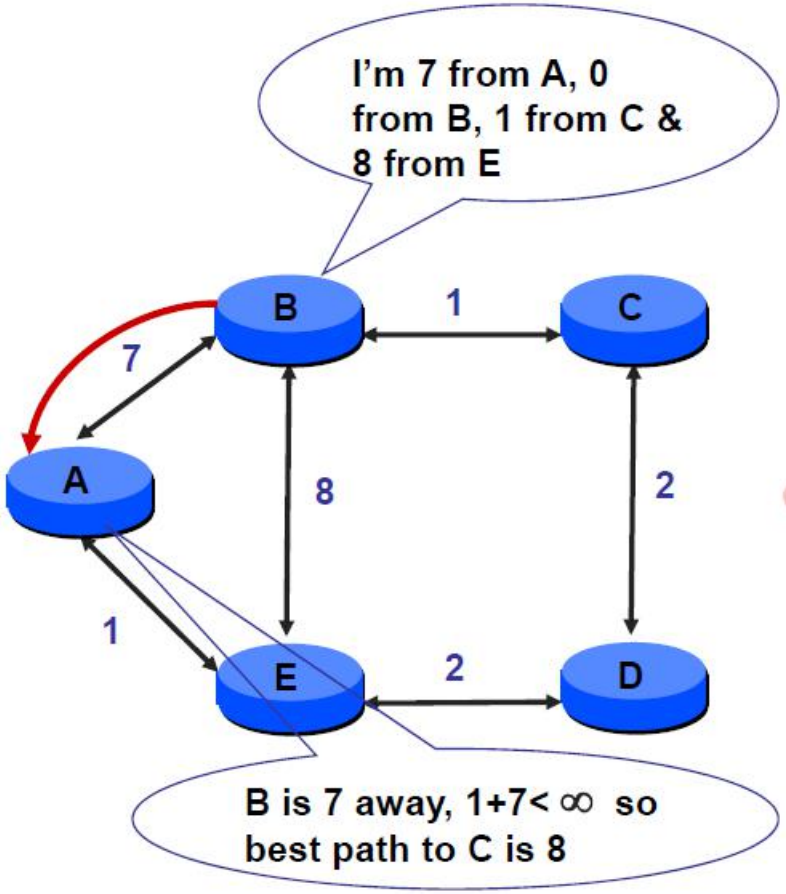
Info at node	Distance to Node				
	A	B	C	D	E
A	0	7	∞	∞	1
B	7	0	1	∞	8
C	∞	1	0	2	∞
D	∞	∞	2	0	2
E	1	8	∞	2	0

D Sends Vector to E



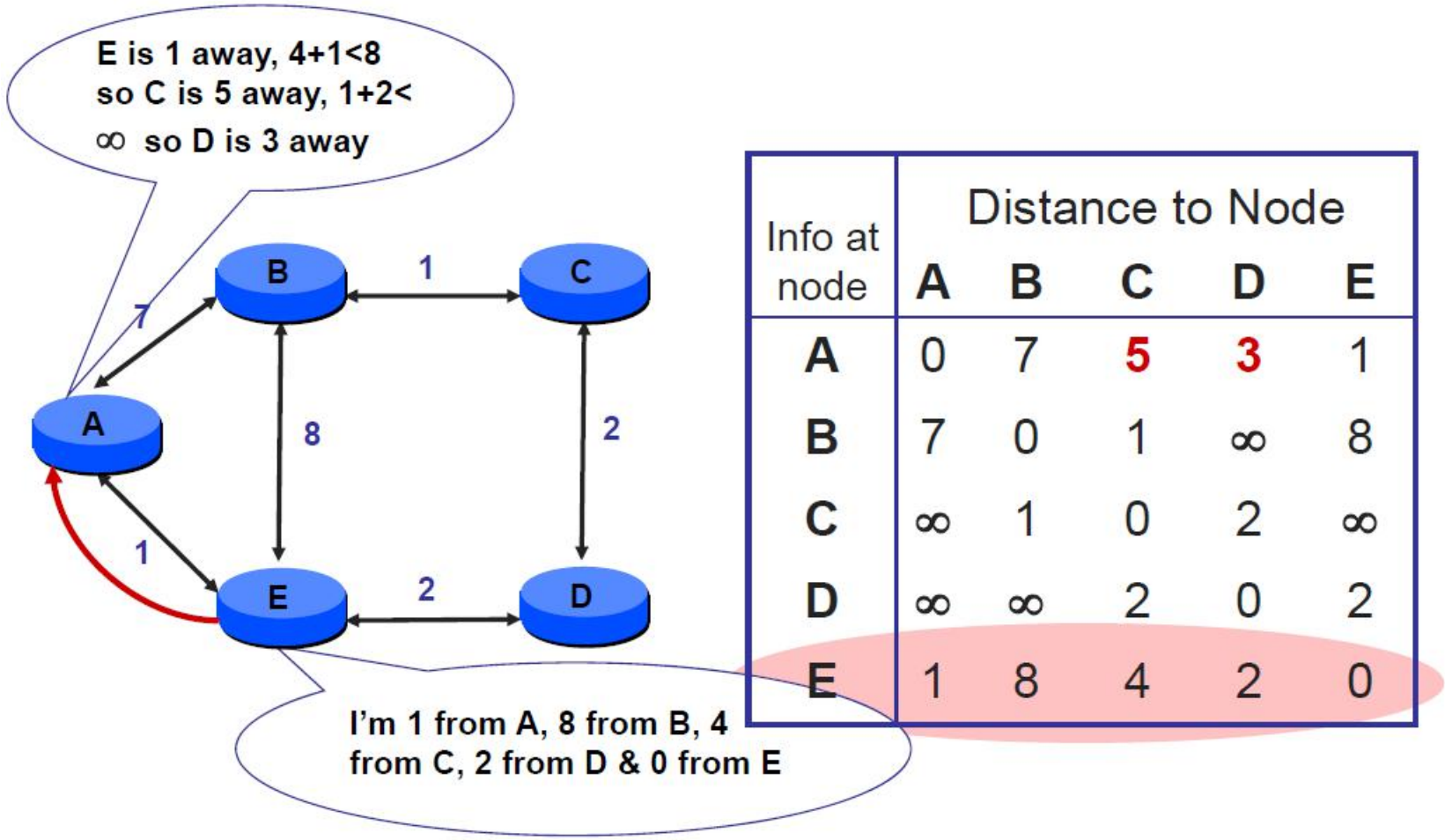
Info at node	Distance to Node				
	A	B	C	D	E
A	0	7	∞	∞	1
B	7	0	1	∞	8
C	∞	1	0	2	∞
D	∞	∞	2	0	2
E	1	8	**4**	2	0

B Sends Vector to A



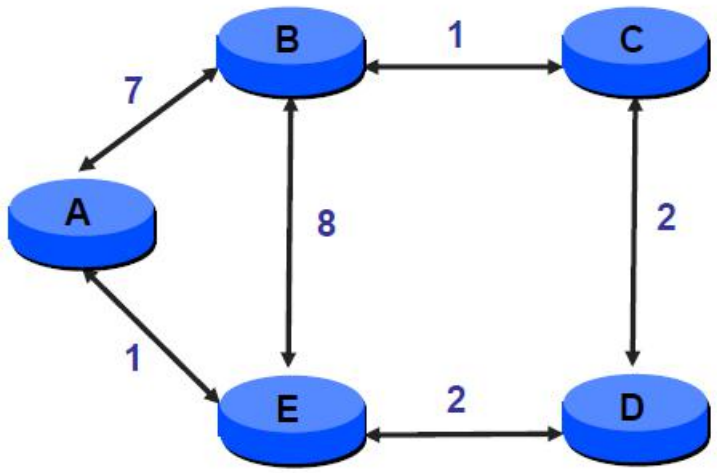
Info at node	Distance to Node				
	A	B	C	D	E
A	0	7	8	∞	1
B	7	0	1	∞	8
C	∞	1	0	2	∞
D	∞	∞	2	0	2
E	1	8	4	2	0

E Sends Vector to A



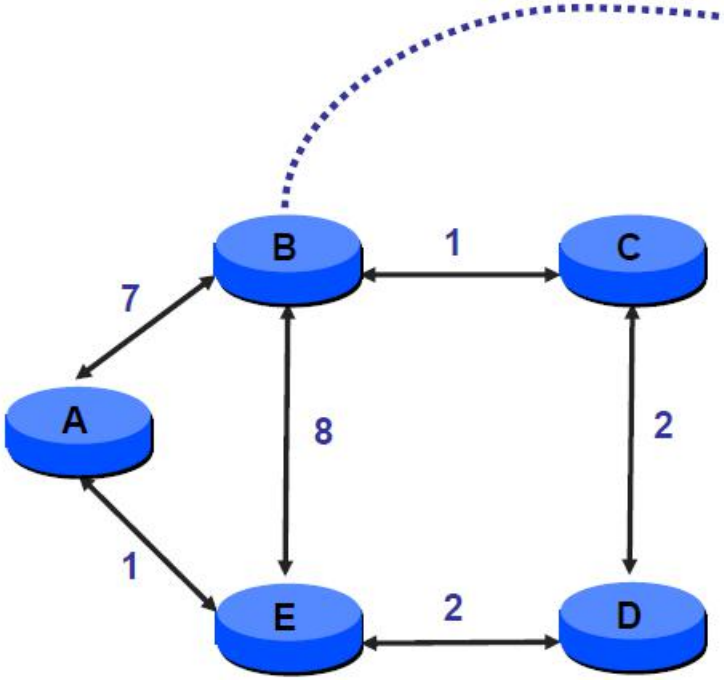
Info at node	Distance to Node				
	A	B	C	D	E
A	0	7	**5**	**3**	1
B	7	0	1	∞	8
C	∞	1	0	2	∞
D	∞	∞	2	0	2
E	1	8	4	2	0

… Until Convergence



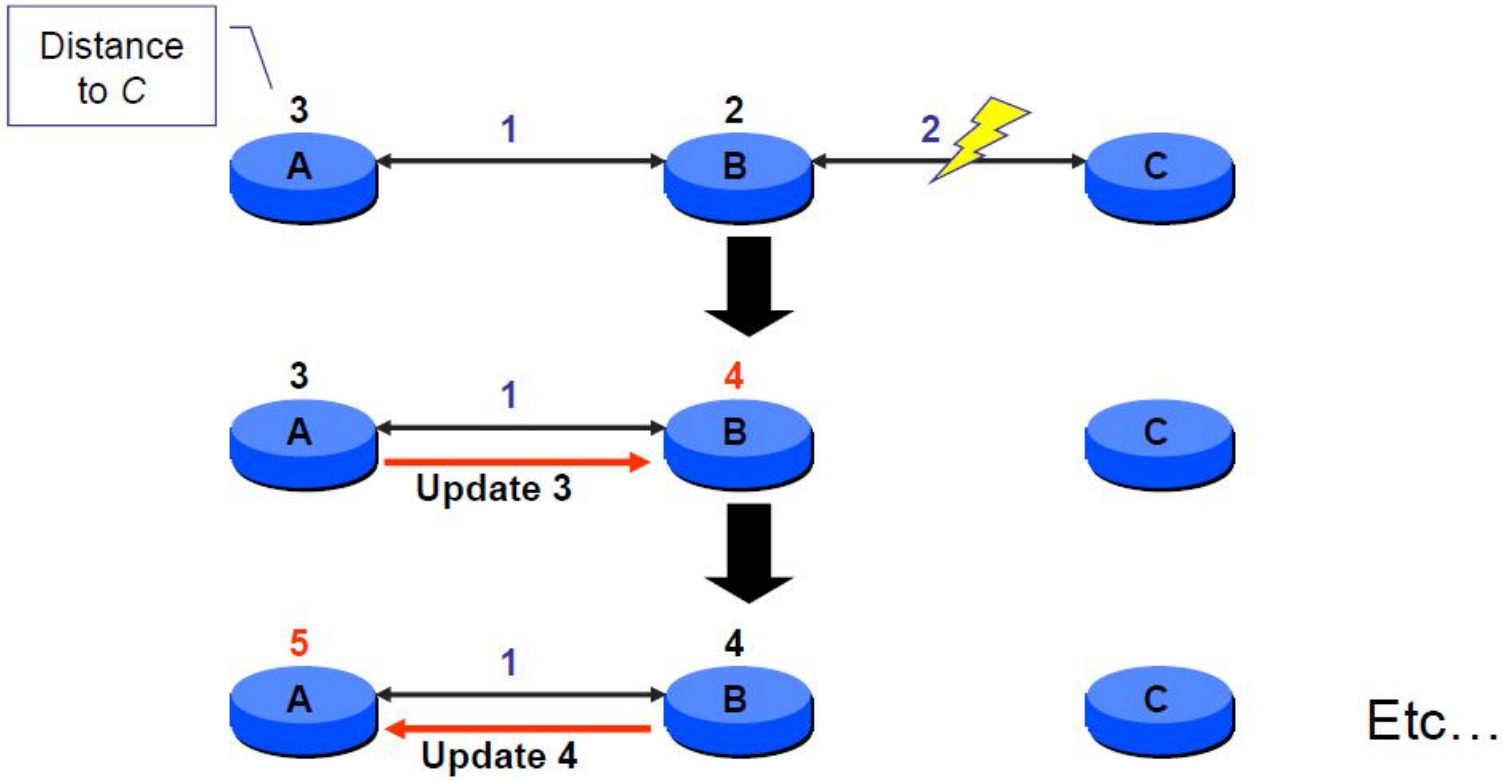
Info at node	Distance to Node				
	A	B	C	D	E
A	0	6	5	3	1
B	6	0	1	3	5
C	5	1	0	2	4
D	3	3	2	0	2
E	1	5	4	2	0

Node B's Distance Vectors



Dest	Next hop		
	A	E	C
A	7	9	**6**
C	12	12	**1**
D	10	10	**3**
E	8	8	**5**

Worse: Count to Infinity



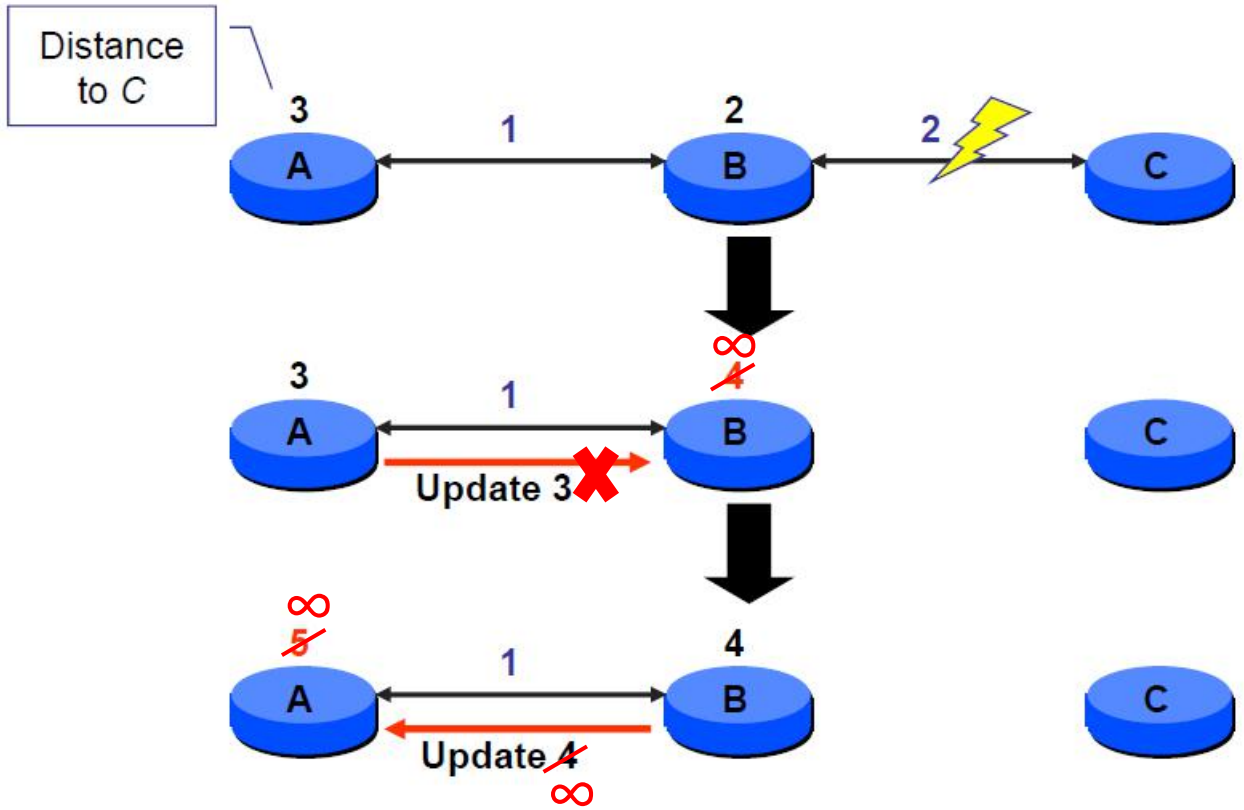
Why So and How to Solve?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Why So?
  - Updates don't contain enough information
  - B accepts A's path to C that is implicitly through B!

- How to Solve?
  - Split horizon: never advertise a destination through its next hop (Only works for node pairs)

Split Horizon



Why So and How to Solve?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

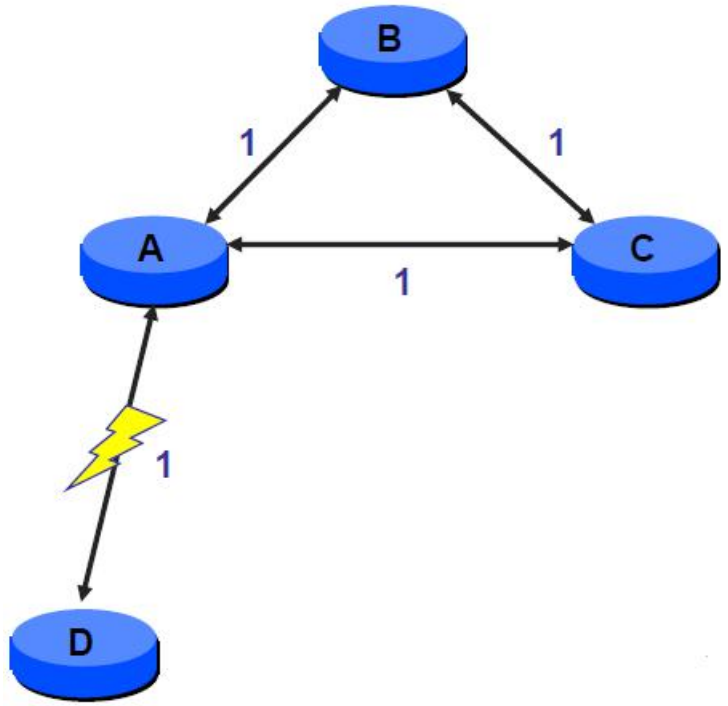
- Why So?
  - Updates don't contain enough information
  - B accepts A's path to C that is implicitly through B!

- How to Solve?
  - Split horizon: never advertise a destination through its next hop
    (Only works for node pairs)
  - Split horizon + poison reverse: send negative information (infinity cost) when advertising a destination through its next hop
    (Only works for a loop of size 2)

Only Works for Loop of Size 2

- A tells B & C that D is unreachable

- B computes new route through C
  - Tells C that D is unreachable (poison reverse)
  - Tells A it has path of cost 3 (split horizon doesn't apply)
  - A computes new route through B
  - A tells C that D is now reachable



Why So and How to Solve?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Why So?
  - Updates don't contain enough information
  - B accepts A's path to C that is implicitly through B!

- How to Solve?
  - Split horizon: never advertise a destination through its next hop
    (Only works for node pairs)
  - Split horizon + poison reverse: send negative information (infinity cost) when advertising a destination through its next hop
    (Only works for a loop of size 2)
  - Hold down: when a route is removed, no update of it accepted for some period of time (hold-down time) - to give everyone a chance to remove the route

Real DV Protocols

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- RIP: Routing Information Protocol
  - DV protocol with hop count as metric
  - Infinity value is 16 hops; limits network size
  - Routers send vectors every 30 seconds
  - With triggered updates for link failures

2. Link State Routing

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Tell all routers the topology and each computes best paths for source-destination pair of nodes

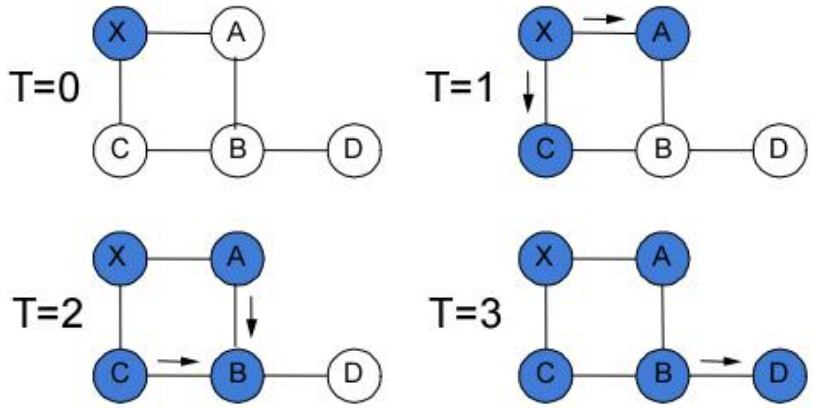
- Two phases
  1. Topology dissemination (by flooding)
  2. Shortest-path calculation (Dijkstra's algorithm)

1. Flooding

- **Link State Packet (LSP)**
  - ID of the node that created the LSP
  - Cost of link to each directly connnected neighbor
  - Sequence number (SEQNO)
  - Time-to-live (TTL) for this packet

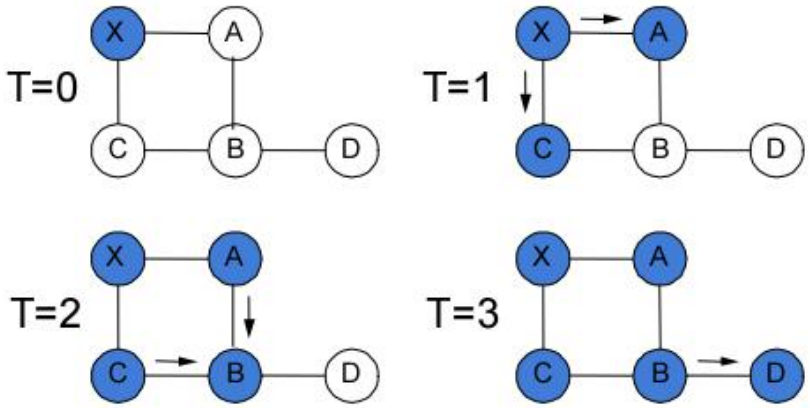
一定要知道区别

- **Relible flooding**
  - Store most recent LSP from each node
  - Forward LSP to all nodes but one that sent it
  - Generate new LSP periodically (increment SEQNO)
  - Decrement TTL of each stored LSP (discard when TTL = 0)



Why “Reliable”?

- **Complications**
  - Packet loss
  - Out-of-order arrival
- **Solutions**
  - Acknowledgments and retransmissions
  - Sequence numbers
  - Time-to-live for each packet



2. Shortest-path Calculation

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Dijkstra's Algorithm:

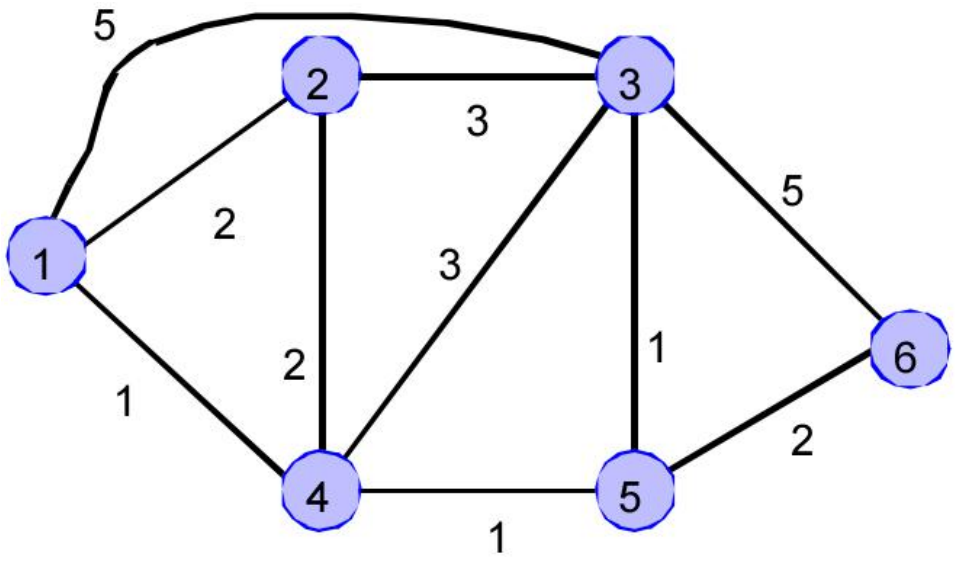
s source node.

D_n cost of the least-cost path from node s to node n

```
M = {s};
for each  n ∉ M
        D_n  = d_sn;
while (M ≠ all nodes) do
        Find w ∉ M for which D_w = min{D_j ; j ∉ M};
        Add w to M;
        for each  n ∉ M
                D_n = min_w [ D_n, D_w + d_wn ];
                Update route;
enddo
```

Example

Construct the routing table at node 1 using Dijkstra's algorithm

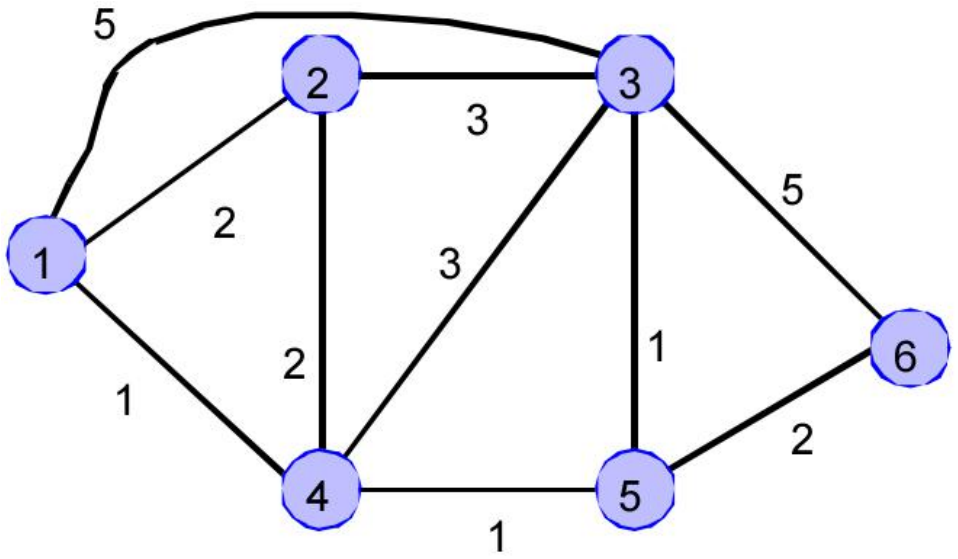


Iteration	M	D_1	D_2	D_3	D_4	D_5	D_6
Init							

Example

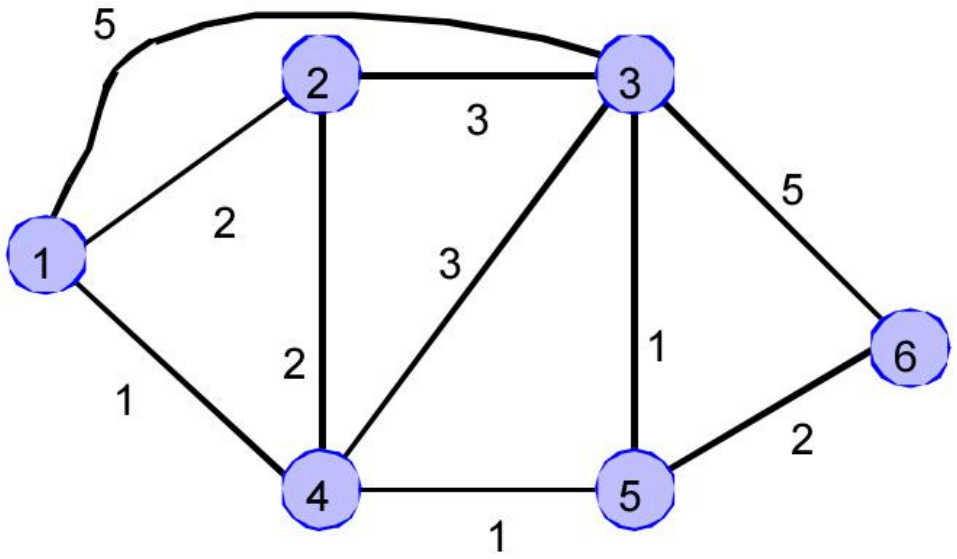
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

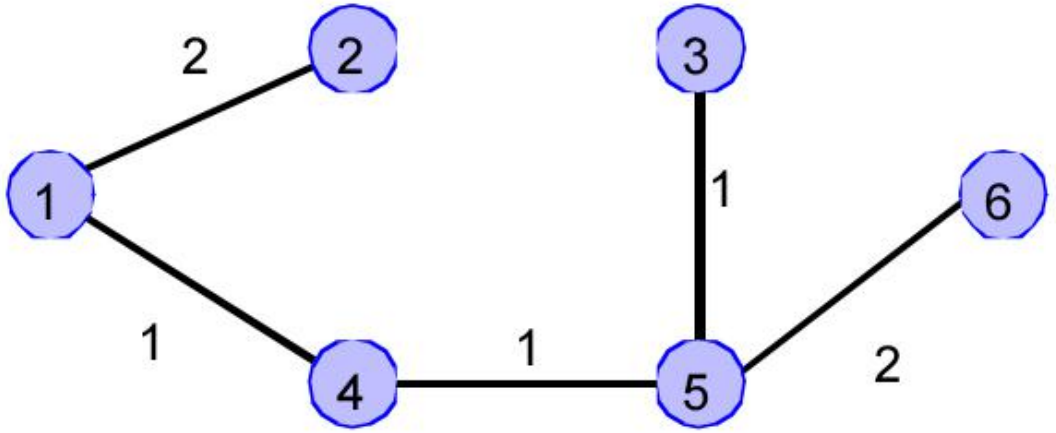
Construct the routing table at node 1 using Dijkstra's algorithm



	M	D1	D2	D3	D4	D5	D6
0	{1}	0	2	5	1	inf	inf
1	{1,4}	0	2	4	1	2	inf
2	{1,4,2,5}	0	2	3	1	2	4
3	{1,4,2,5,3}	0	2	3	1	2	4
4	{1,4,2,5,3,6}	0	2	3	1	2	4

Result Routing Table





The tree is translated into a routing table at node 1:

Destination	Next Hop
2	2
3	4
4	4
5	4
6	4

Real Link State Protocols

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **OSPF (Open Shortest Path First) and IS-IS**
  - Most widely used intra-domain routing protocols
  - Run by almost all ISPs and many large organizations

- **Implementational add-ons of OSPF**
  - Adds notion of areas for scalability
  - Area 0 is “backbone” area
  - Traffic between two areas must always go thrugh area 0
  - Only need to know how to route within area
  - Load balancing: multiple equanl cost routes

LS Routing vs DV Routing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **DV**: Tell your neighbors about the world
  - Global, consistent information
  - Easy to get confused
  - Slow convergence due to ripples and hold down
  - Simpler sometimes (only update neighbors on some DV changes)

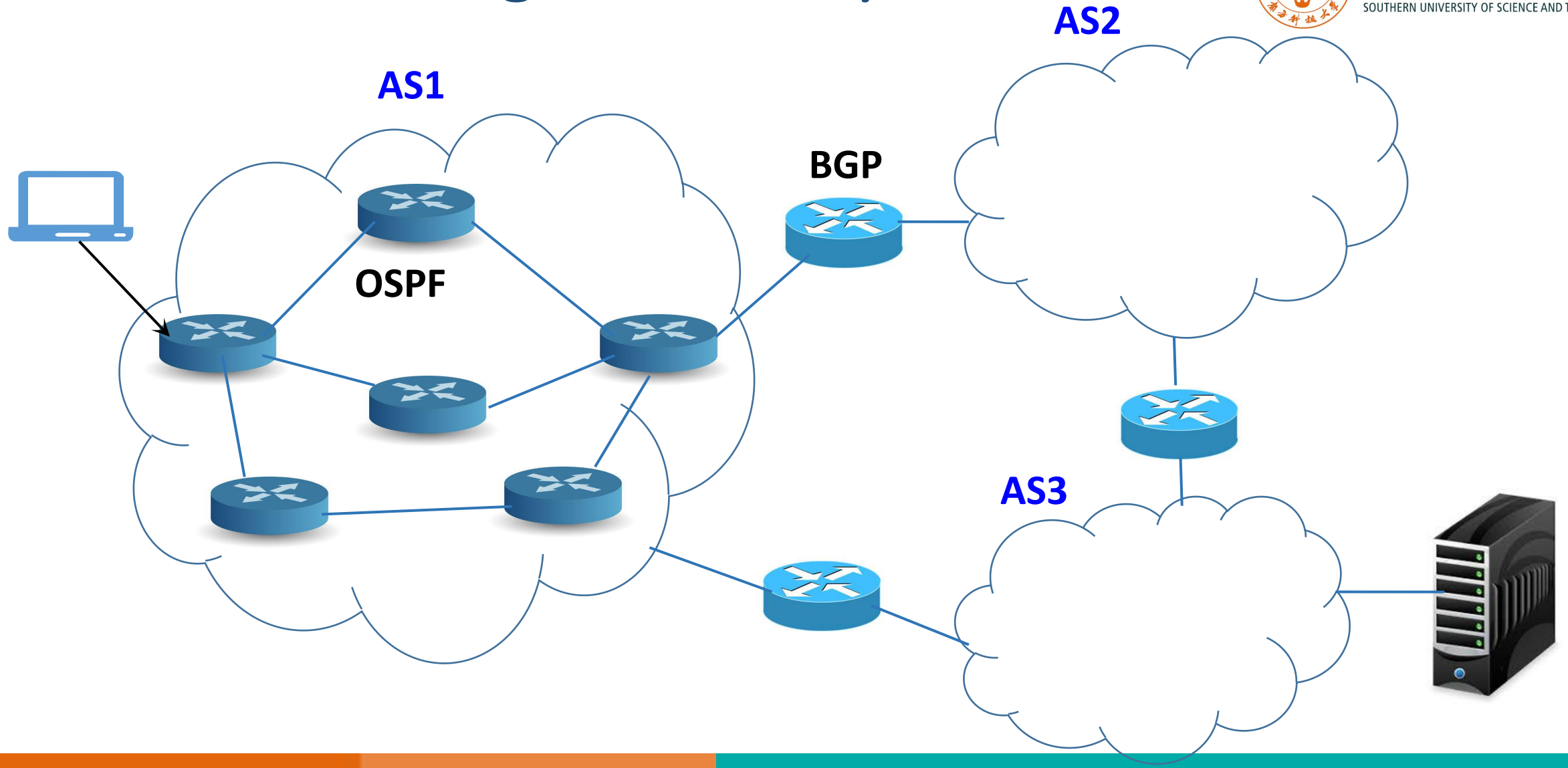
各自有什么优点和缺点

- **LS**: Tell the world about your neighbors
  - Iterative, asynchrnous and destributed information
  - Harder to get confused
  - Faster convergence (instantaneous update of link state changes)
  - More expensive sometimes (flooding to every one on link changes)

Contents

- Routing Basics
- Intra-AS Routing
- Inter-AS Routing
- Multicast Routing
- MPLS

Inter-AS Routing: Scalability Problem



Issues with Link State

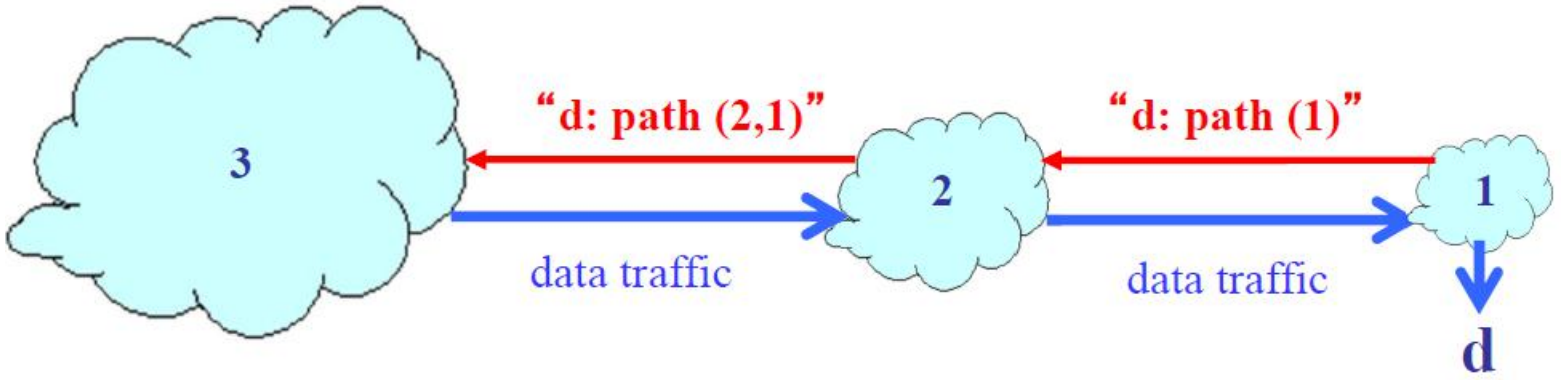
- Topology information is flooded
  - High bandwidth and storage overhead
- Entire path computed locally per node
  - High processing overhead in a large network
- Can't express policy

- What about Distance Vector?
  - Hides details of the network topology
  - Nodes determine only “next hop” toward the destination
  - But, slow convergence due to the counting-to-infinity problem
  - Idea: extend the notion of a distance vector

Path Vector Routing

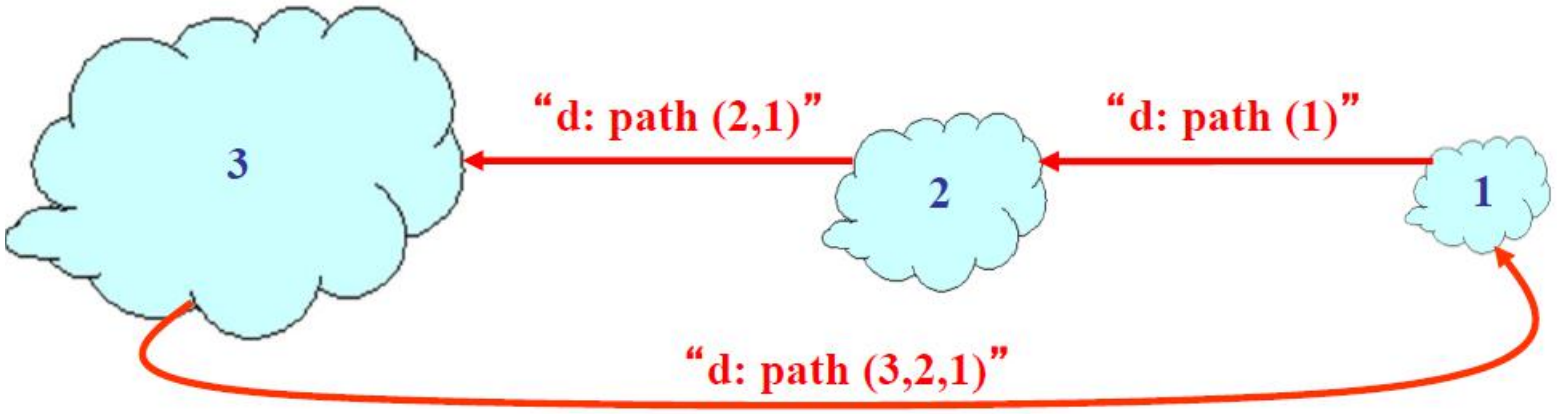
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Key idea: advertise the entire path (sequence of ASes)
  - Distance vector: send distance metric per destination
  - Path vector: send the entire path for each destination



Loop Detection

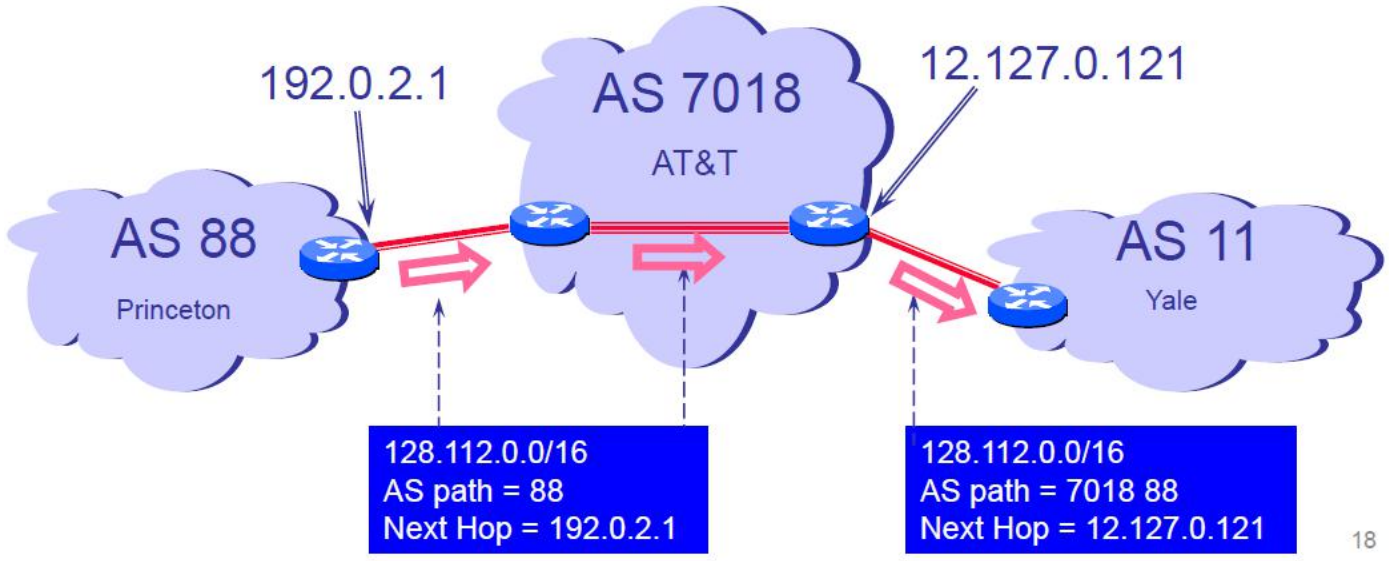
- Node can easily detect a loop
  - Look for its own node identifier in the path
  - E.g., node 1 sees itself in the path ‘‘3, 2, 1’’
- Node can simply discard paths with loops
  - E.g., node 1 simply discards the advertisement



Border Gateway Protocol (BGP)

- BGP uses Path Vector Routing

- A simple BGP route with destinatiion prefix: e.g., 128.112.0.0/16)
  - Route attributes, including AS path (e.g., “7018 88”) and next-hop IP address (e.g., 12.127.0.121)



Contents

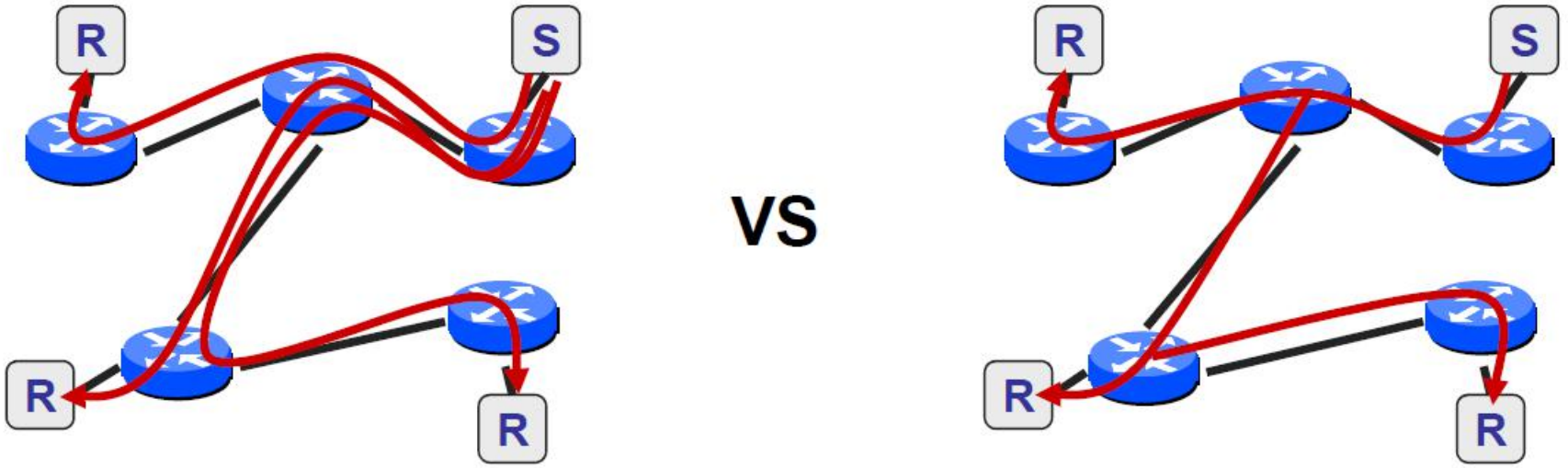
- Routing Basics
- Intra-AS Routing
- Inter-AS Routing
- Multicast Routing quiz不要求
- MPLS

Multicasting: Motivation

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Efficient delivery to multiple destinations (e.g. video broadcast)

- One-to-many addressing is needed for reducing traffic!



IP Multicast service model

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Communications based on groups
  - Special IP addresses (Class D in IPv4) represent “multicast groups”
  - Anyone can join group to receive packets
  - Anyone can send to group (senders need not be part of group)

- Unreliable datagram service
  - Extension to unicast IP
  - Group membership not visible to hosts
  - No synchronization

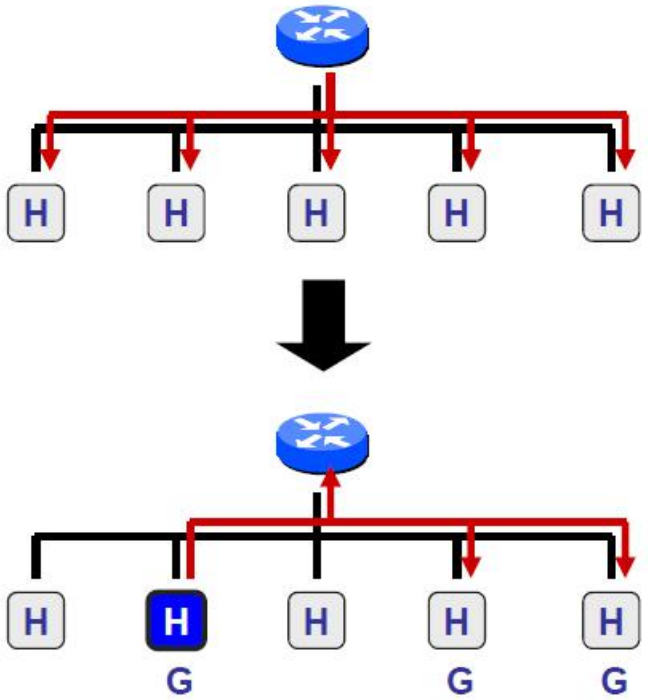
Elements of IP Multicast

- Host interface
  - Application visible multicast API
  - Multicast addressing
  - Link-layer mapping
- Host-Router interface
  - IGMP
- Router-Router interface
  - Multicast routing protocols

Internet Group Management Protocol (IGMP)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Goal: Communicate group membership between hosts and routers
  - Hosts explicitly inform their router about membership
  - Must periodically refresh membership report



Router broadcasts membership query to 224.0.01

Host sends membership report to group G when its timer expires

Router-Router Interface

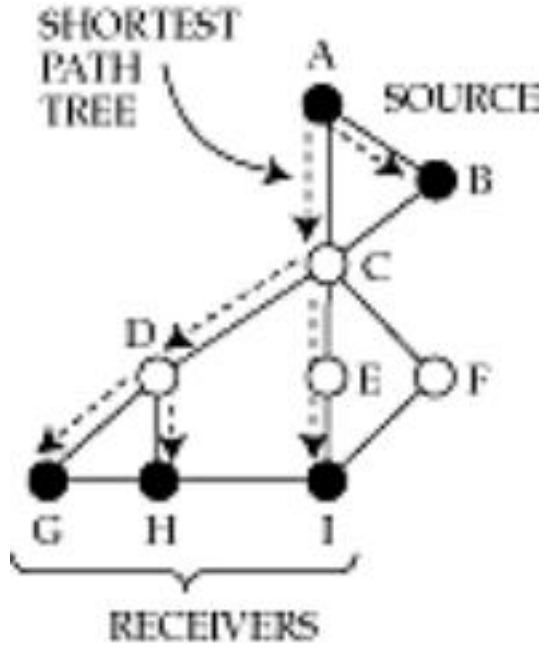
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Protocols for multicast packet forwarding

- A simple solution
  - Flood packets from a source to entire network
  - If a router has not seen a packet before, forward it to all interfaces except the incoming one

Problems

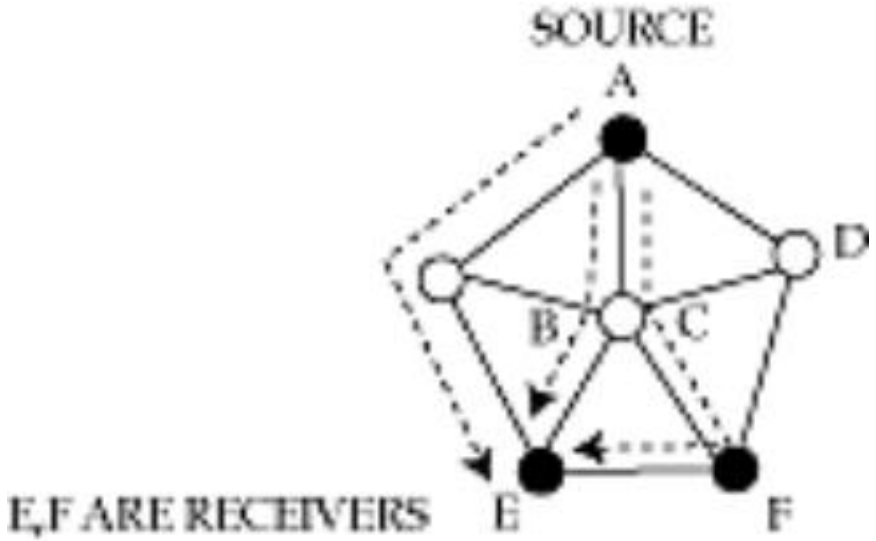
1. Routers receive duplicate packets
2. Detecting that a packet is a duplicate requires storage, which can be quite expensive



A Clever Solution

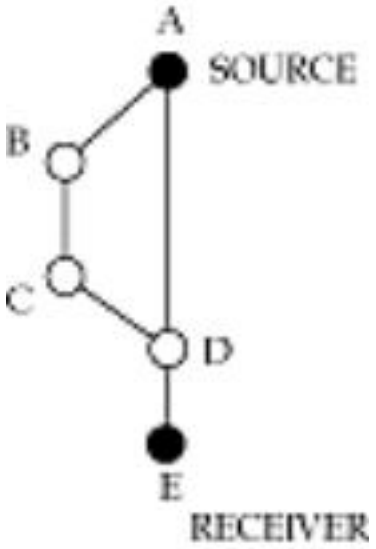
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Reverse path forwarding
  - Forward packet from S to all interfaces if and only if packet arrives on the interface that corresponds to the shortest path *to* S
  - No need to remember past packets (C needs not forward packet received from D)



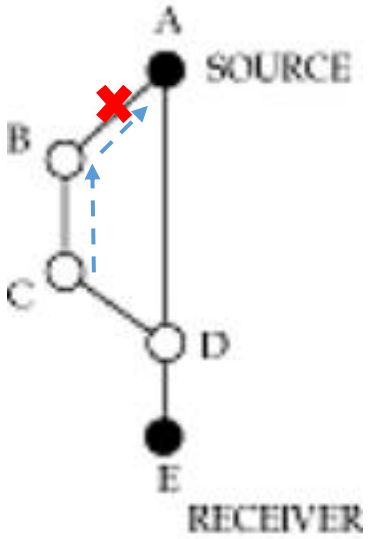
Cleverer

- Don’t send a packet downstream if you are not on the shortest path from the downstream router to the source
- C needs not forward packet from A to E



Even Cleverer: Pruning

- RPF does not completely eliminate unnecessary transmissions (B and C get packets even though they do not need it)
- Pruning => router tells parent in tree to stop forwarding



But How?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- A bunch of Mulitcast protocols...
  - MOSPF (Extenstion of OSPF)
  - DVMRP (Distance-vector Multicast routing protocol)
  - PIM (Protocol independent multicast)
  - … …

Contents

- Routing Basics
- Intra-AS Routing
- Inter-AS Routing
- Multicast Routing
- MPLS

MPLS Overview

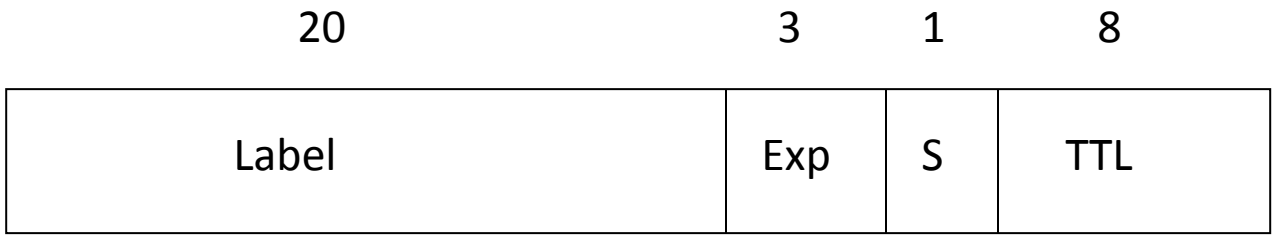
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- MPLS: Multi-Protocol Label Switching
- A forwarding scheme designed to speed up IP packet forwarding (RFC 3031)
- Idea: use a fixed length label in the packet header to decide packet forwarding
- Label is carried in an MPLS header between the link layer header and network layer header

MPLS Header Format

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Label: 20-bit label value
- Exp: experimental use
  - Can indicate class of service
- S: bottom of stack indicator
  - 1 for the bottom label, 0 otherwise
- TTL: time to live



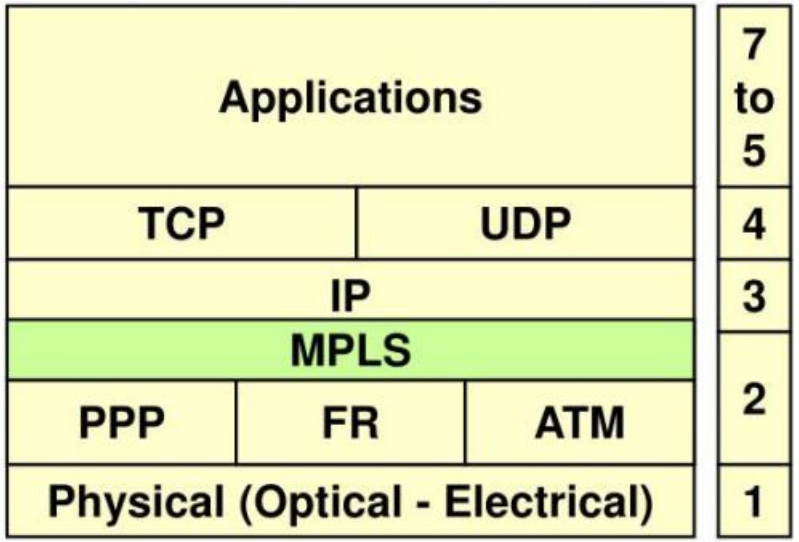
Forwarding Equivalence Class

- An MPLS capable router is called a **label switching router (LSR)**
- Forwarding Equivalence Class (FEC)
  - A subset of packets that are all treated the same way by an LSR
- A packet is assigned to an FEC at the ingress of an MPLS domain by either
  - Source and/or destination IP address
  - Source and/or destination port number
  - Protocol ID
  - Incoming interface

MPLS Operation

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- At ingress LSR of an MPLS domain, an MPLS header is inserted to a packet before the packet is forwarded
  - Label in the MPLS header encodes the packet's FEC



MPLS Operation

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- At subsequent LSRs
  - The label is used as an index into a forwarding table that specifies the next hop and a new label.
  - The old label is replaced with the new label, and the packet is forwarded to the next hop.
- Egress LSR strips the label and forwards the packet to final destination based on the IP packet header

Label Switched Path

- For each FEC, a specific path called *Label Switched Path (LSP)* is assigned
- For each FEC, a specific path called *Label Switched Path (LSP)* is assigned
- A forwarding table is constructed as the result of label distribution

LSP Route Selection

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Hop-by-hop routing
  - Use the route determined by the dynamic routing protocol
- Explicit routing (ER)
  - The sender LSR can specify an *explicit route* for the LSP

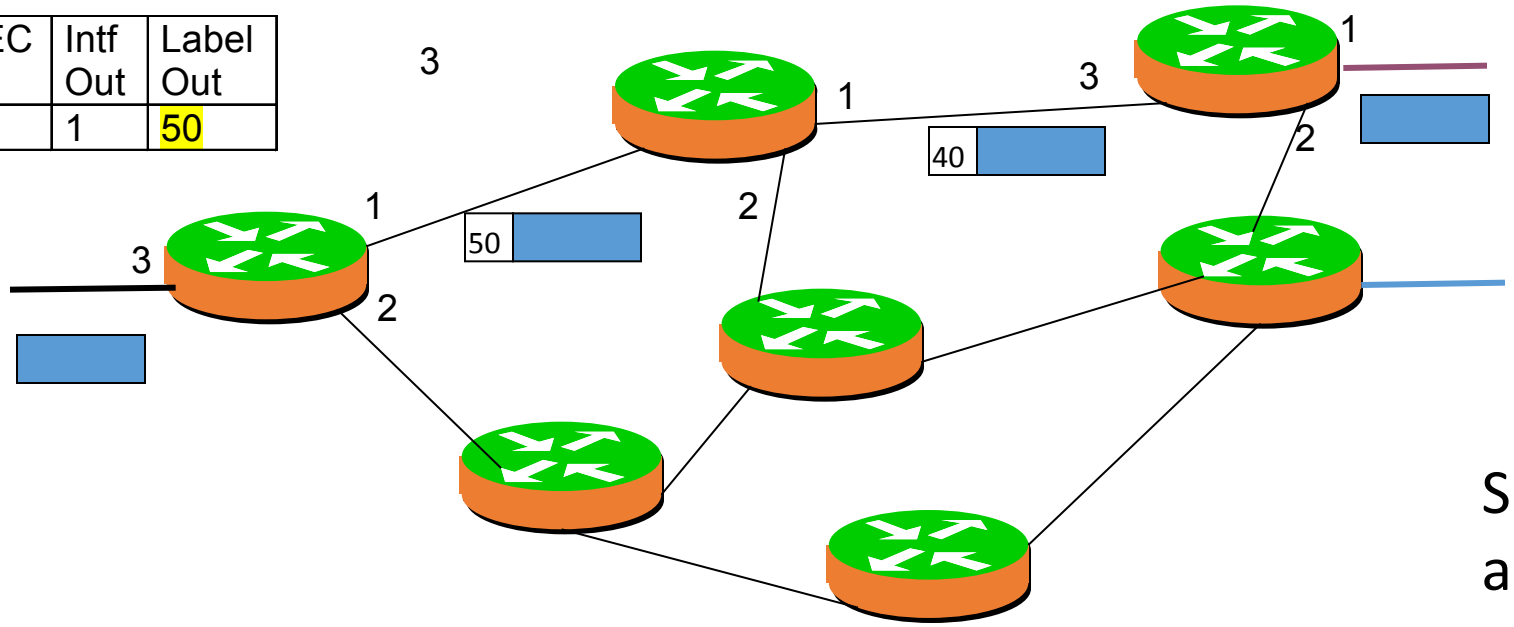
MPLS Operation

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intf In	Label In	Intf Out	Label Out
3	50	1	40

Intf In	Label In	Intf Out
3	40	1

FEC	Intf Out	Label Out
a	1	50



Simpler forwarding table and lookups

Summary

- Routing Basics
- Intra-AS Routing
- Inter-AS Routing
- Multicast Routing
- MPLS