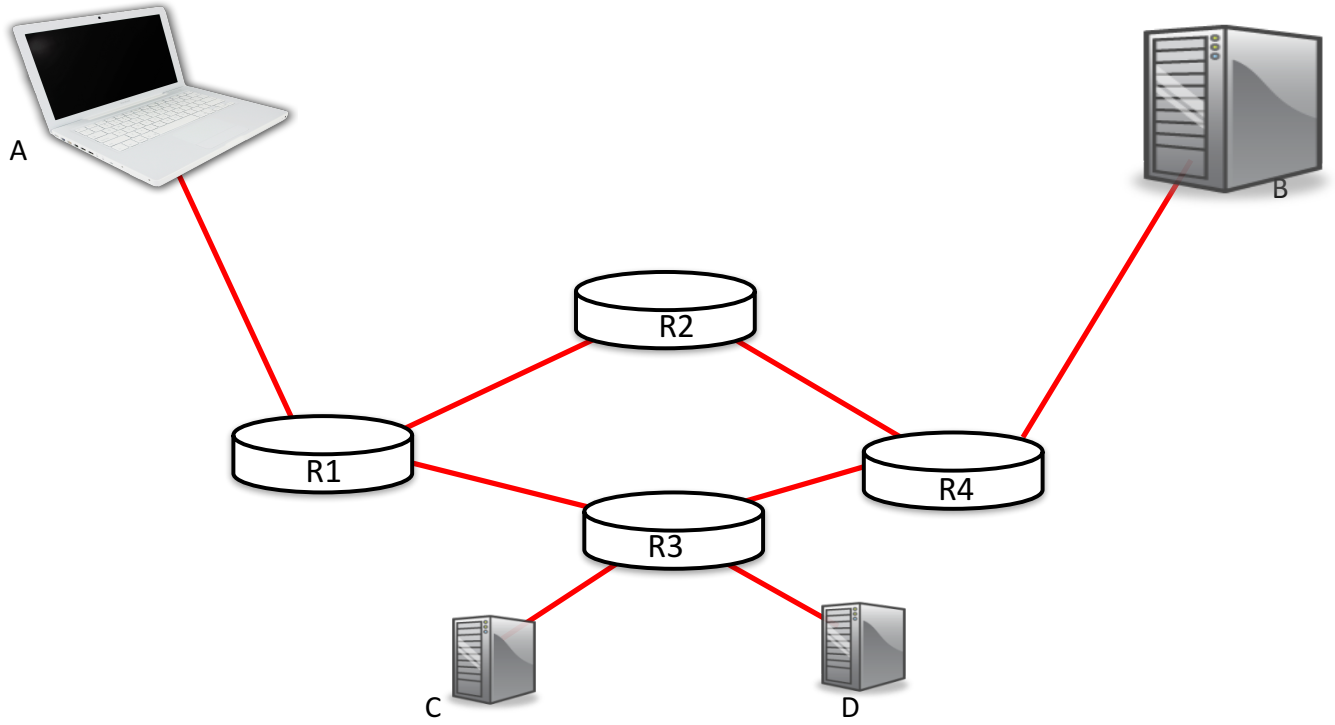
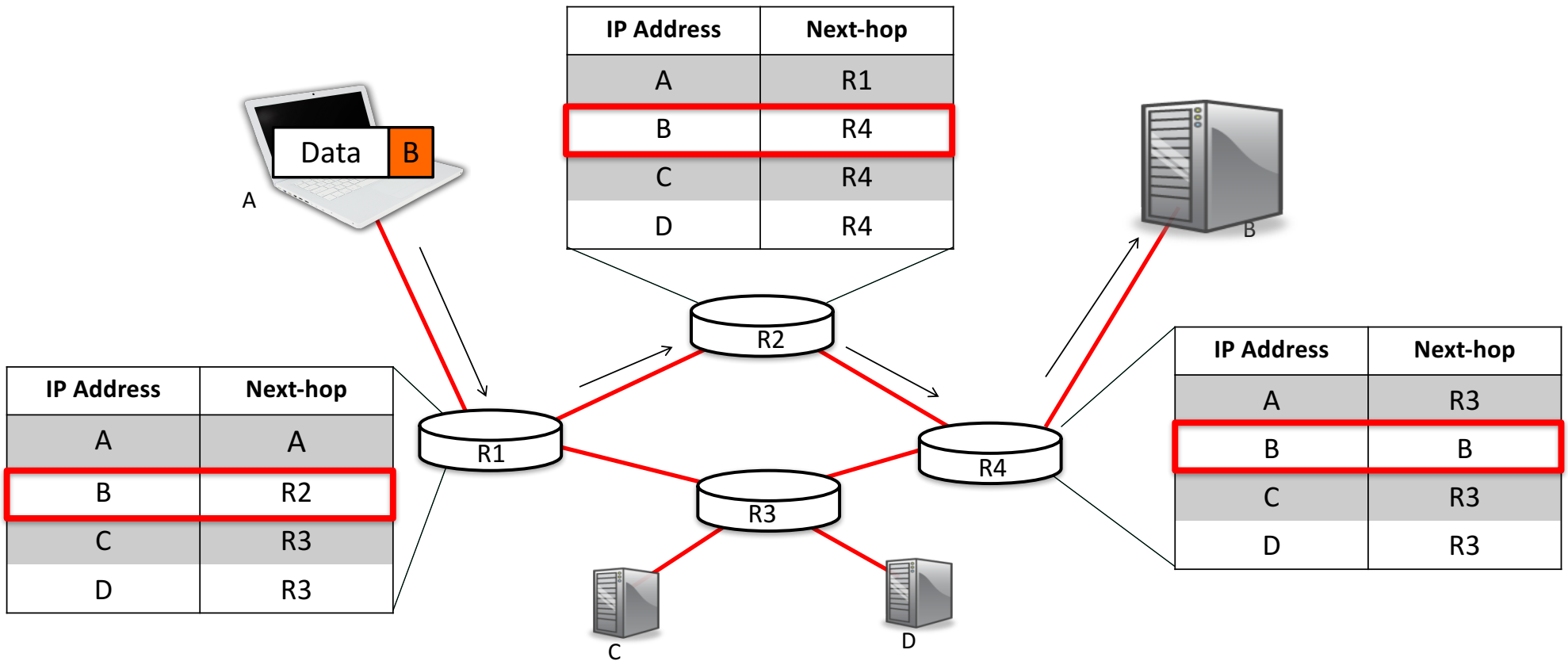
How does a router know where to send a packet next?

The Problem

Which path should packets take from A to B?

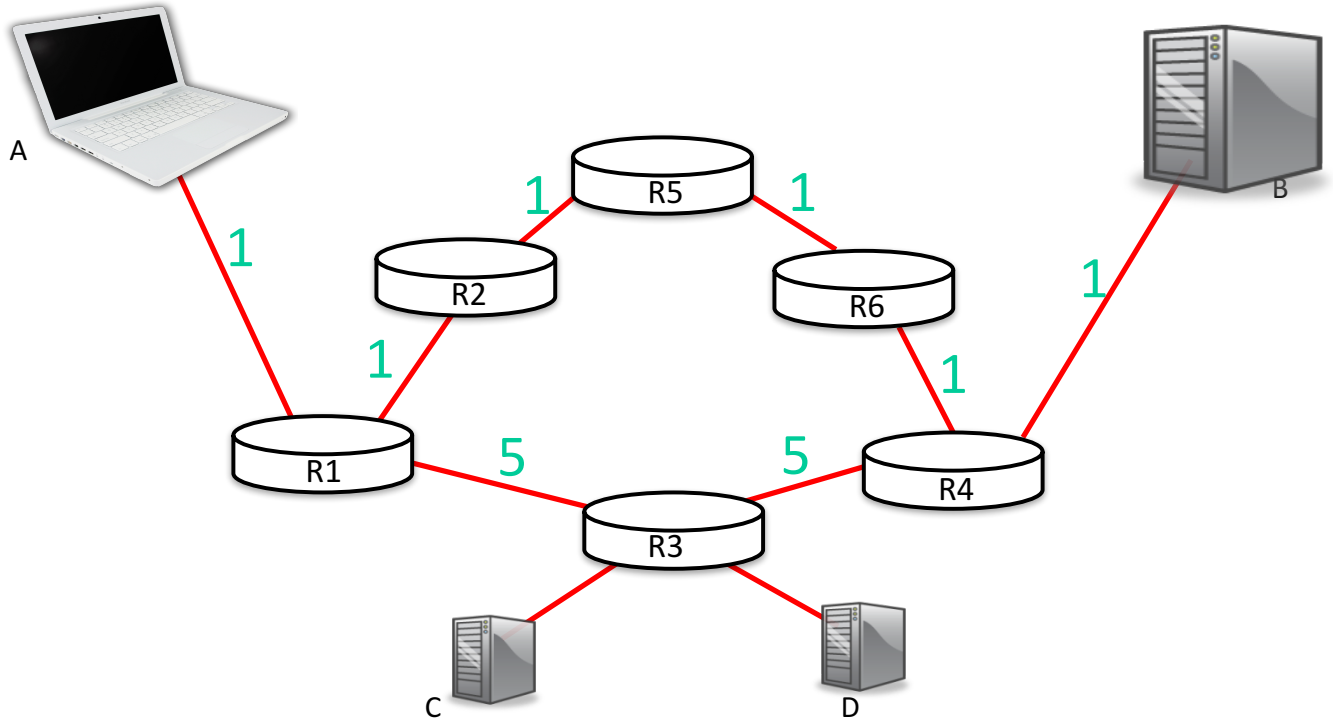


The Internet forwards packets **hop-by-hop**



It's not always obvious

Which path should packets take from A to B?

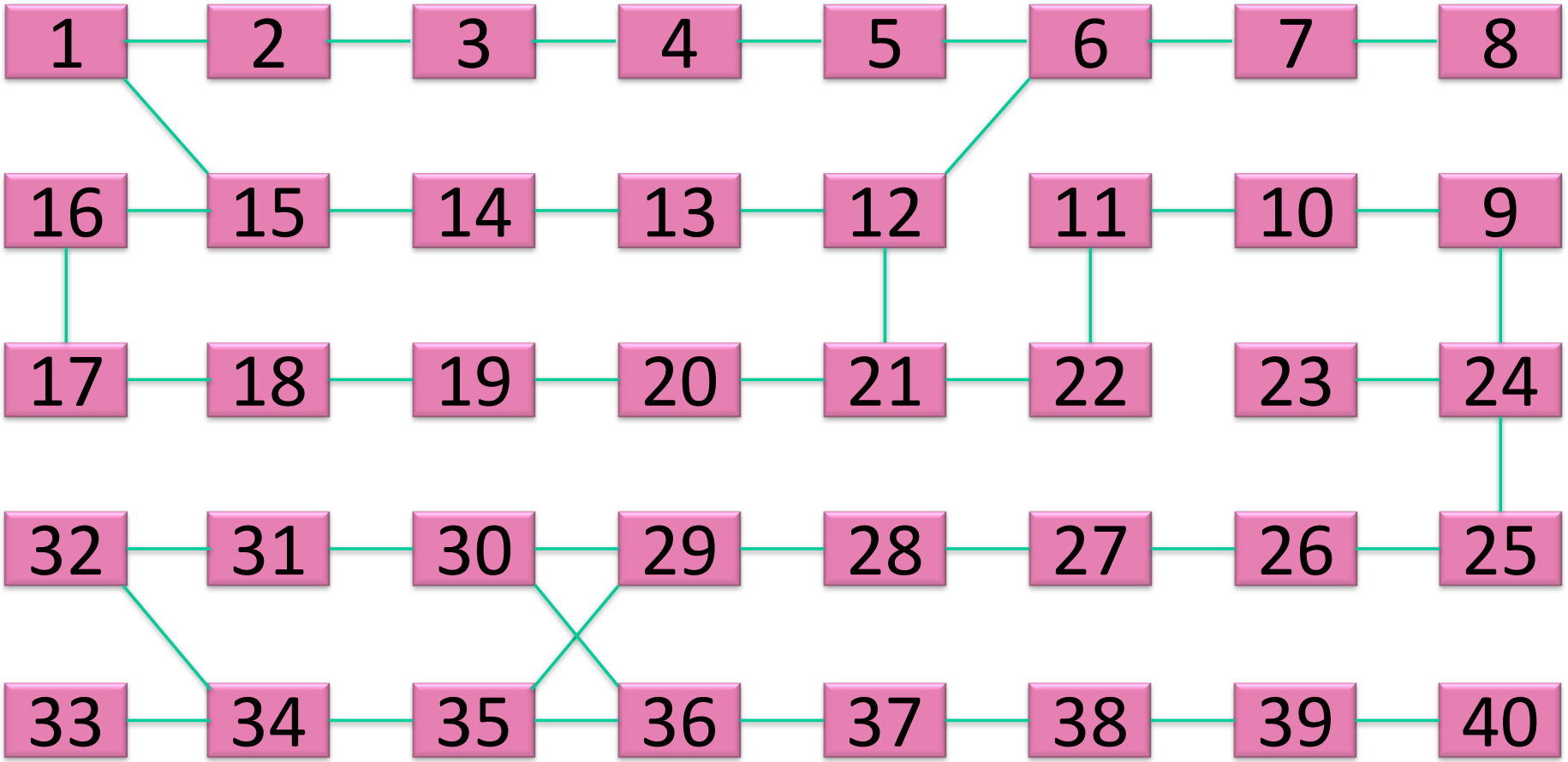


How do the routers **know**
how to populate the forwarding table?

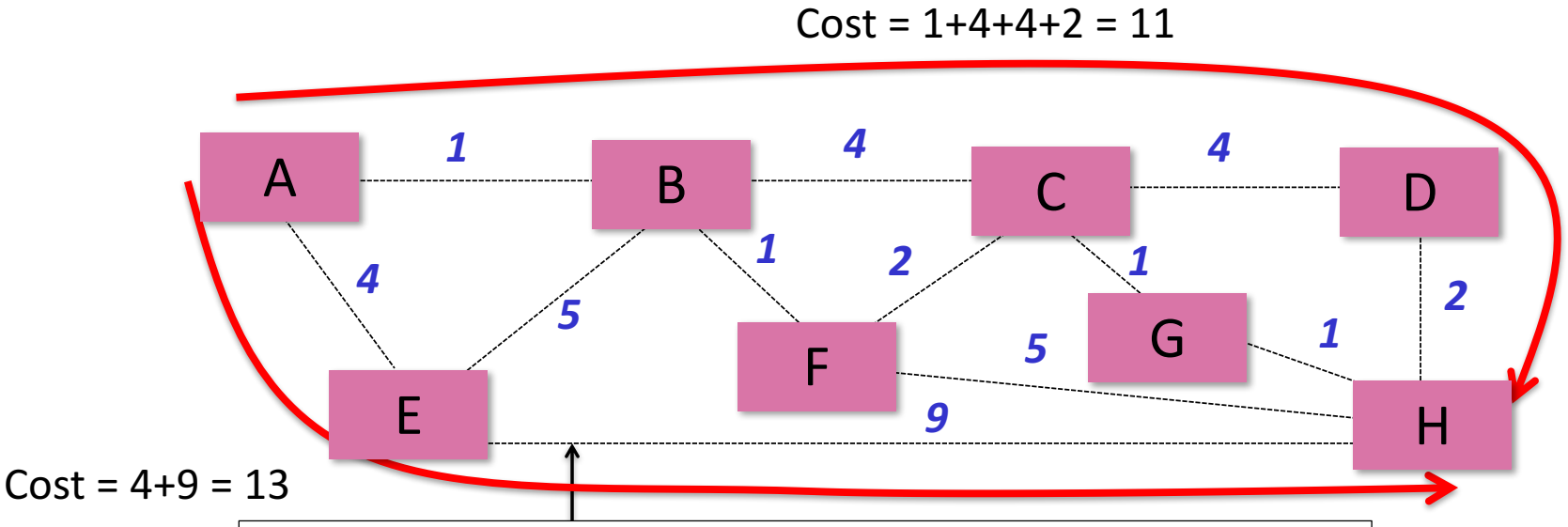
Here are four common ways

1. **Flooding**: Every router sends every arriving packet to every neighbor
2. **Source Routing**: End host adds a list of routers to visit along the way
3. **Static Entries**: If the network owner “knows”, then simply write the entries into the table.
4. Routers talk to each other and construct forwarding tables using a clever algorithm

Find the shortest path from 1 to 40



What if each link has a “cost”?

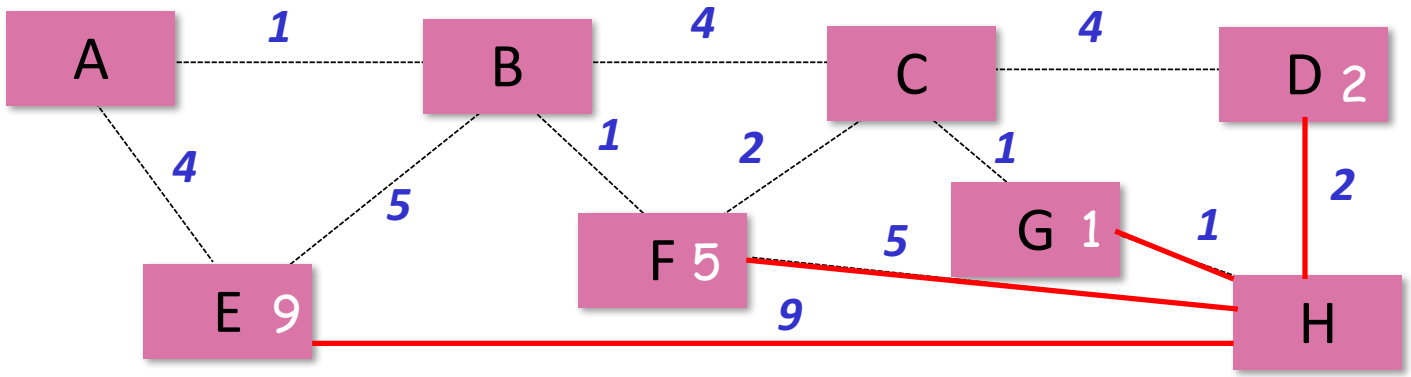


“Expensive link”:

It might be very long. e.g. a link from Europe to USA.
Or it might be very busy. e.g. it connects to Google or CNN.
Or it may be very slow. e.g. 1Mb/s instead of 1Gb/s.

A distributed algorithm to find the lowest cost spanning tree

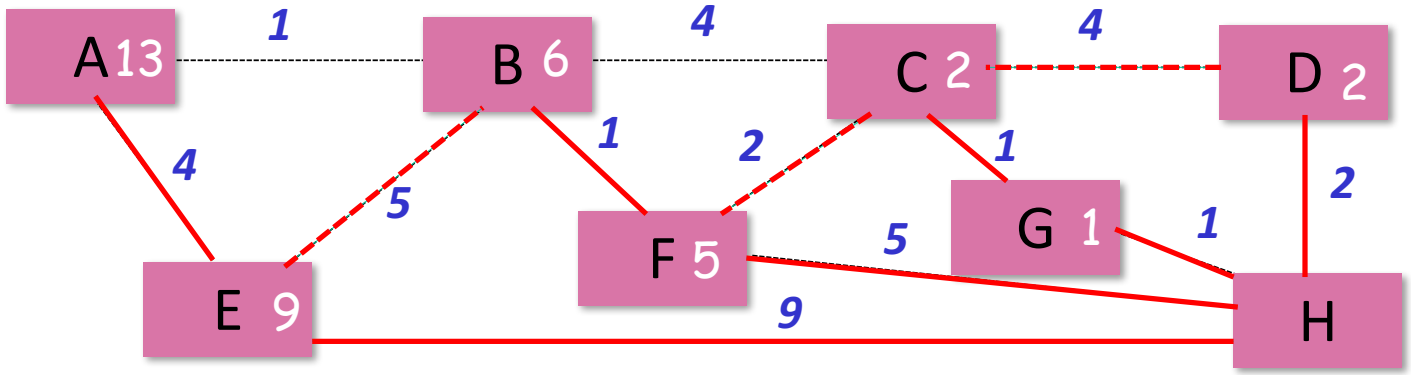
Find lowest cost path from A to H



"Remember the lowest cost we have heard, and the next hop to reach H.
Tell our neighbors the lowest cost we know to reach H."

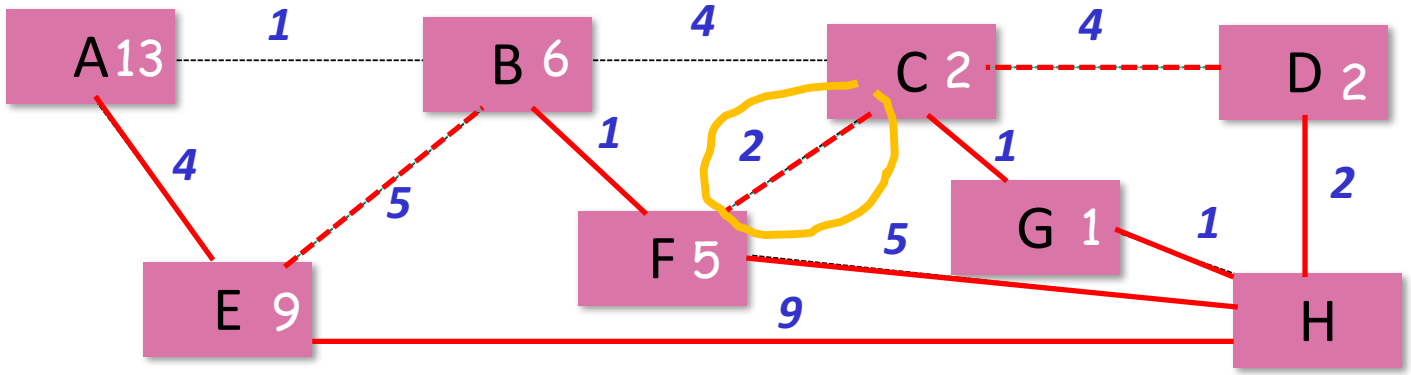
Round #1

Find lowest cost path from A to H



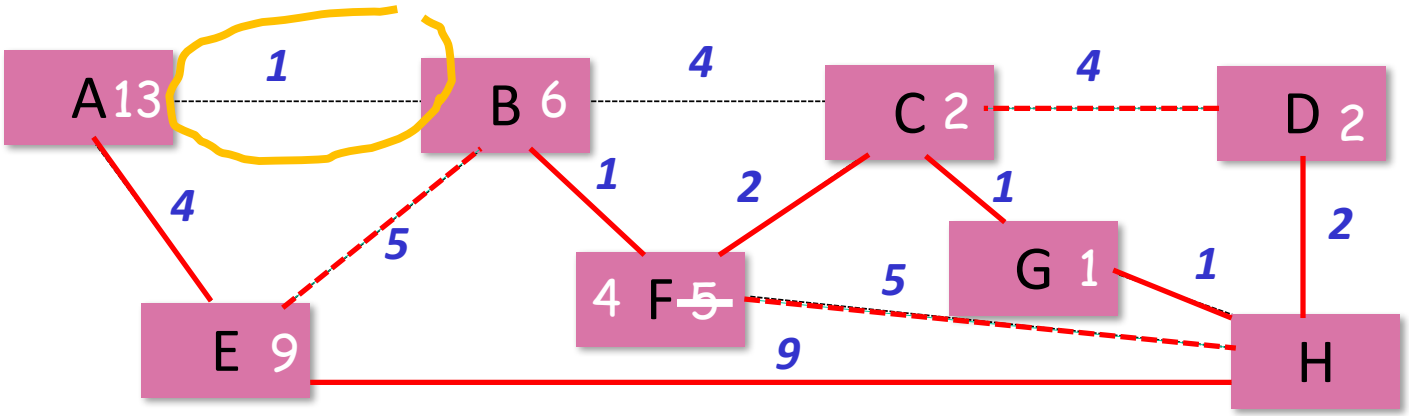
Round #2

Find lowest cost path from A to H



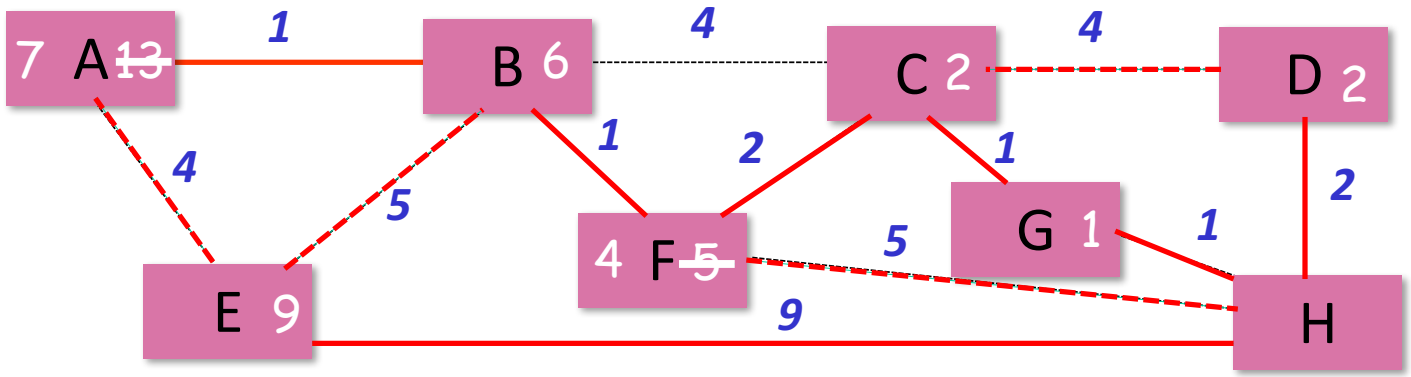
Round #3

Find lowest cost path from A to H



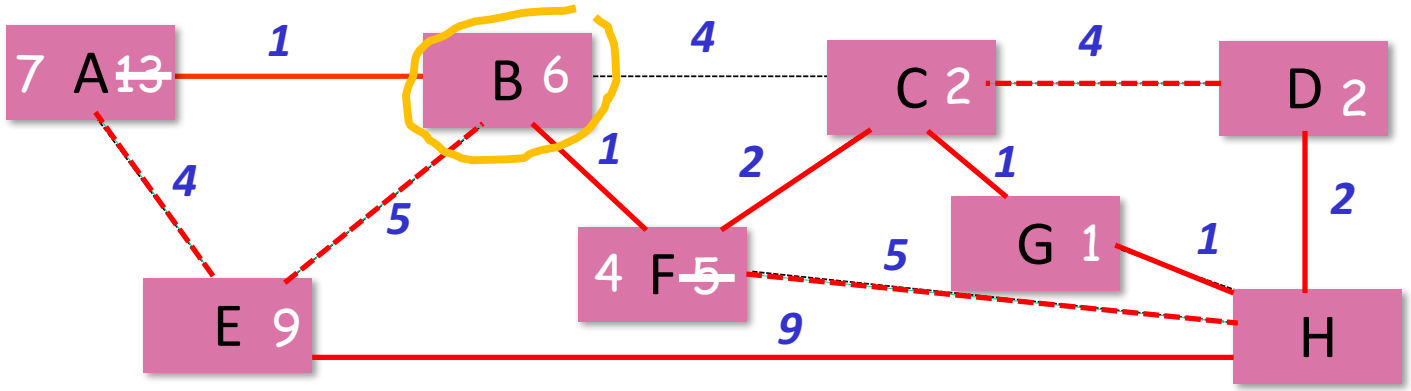
Round #3

Find lowest cost path from A to H



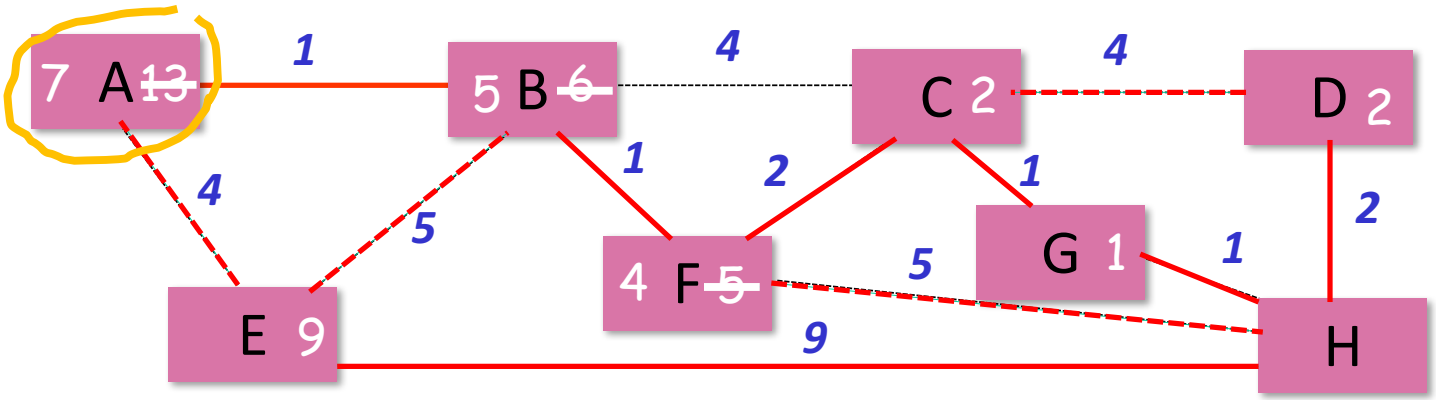
Round #3

Find lowest cost path from A to H



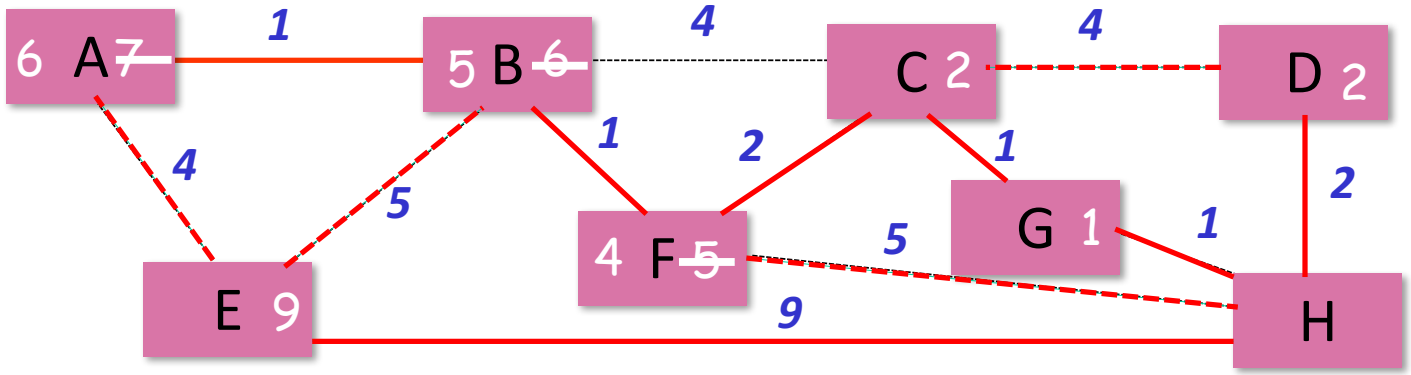
Round #4

Find lowest cost path from A to H



Round #5

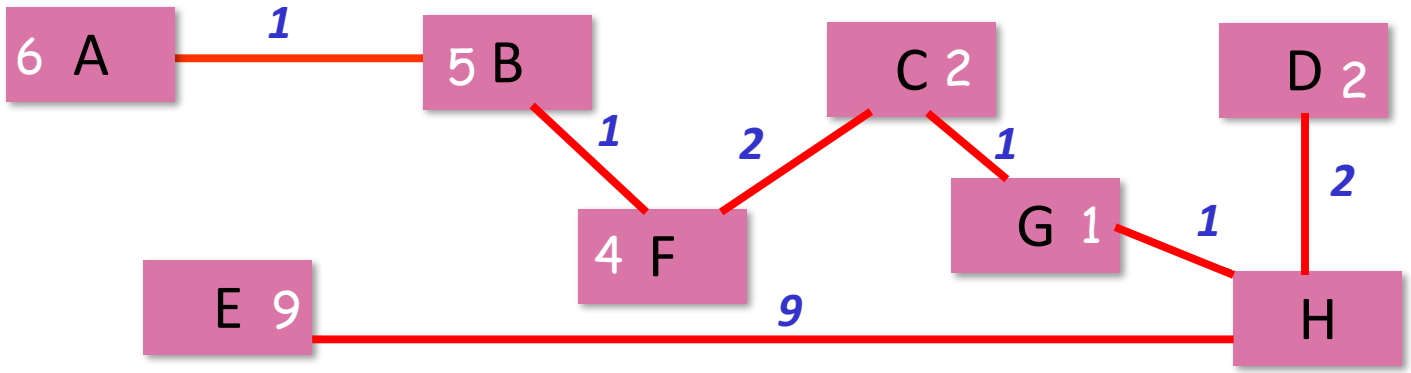
Find lowest cost path from A to H



Round #5

Find lowest cost path from A to H

Known as the Bellman-Ford “Distance Vector” Algorithm

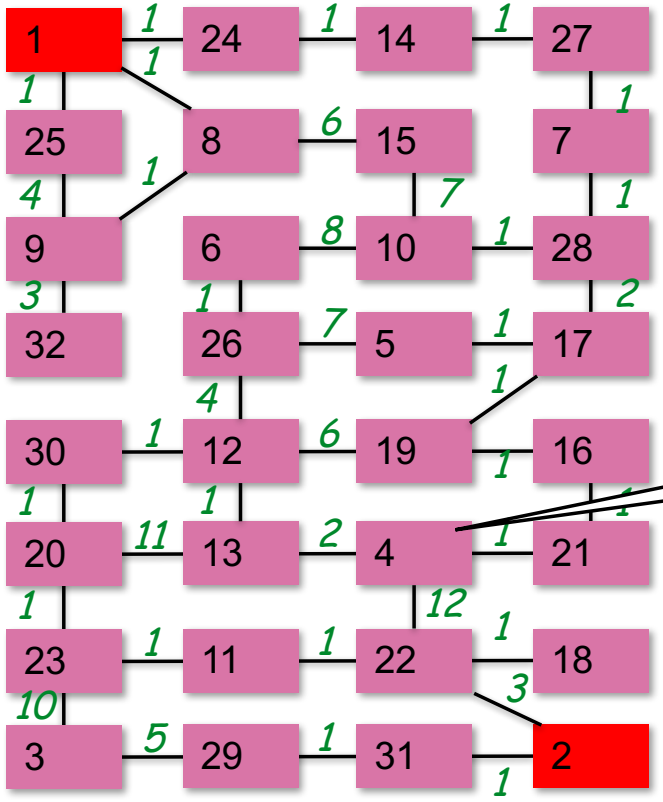


Questions

1. How do we know we finished?
2. In general, what is the maximum number of steps the algorithm needs to complete?
3. How do we turn this into a distributed algorithm between the routers?
4. What happens when the topology changes?

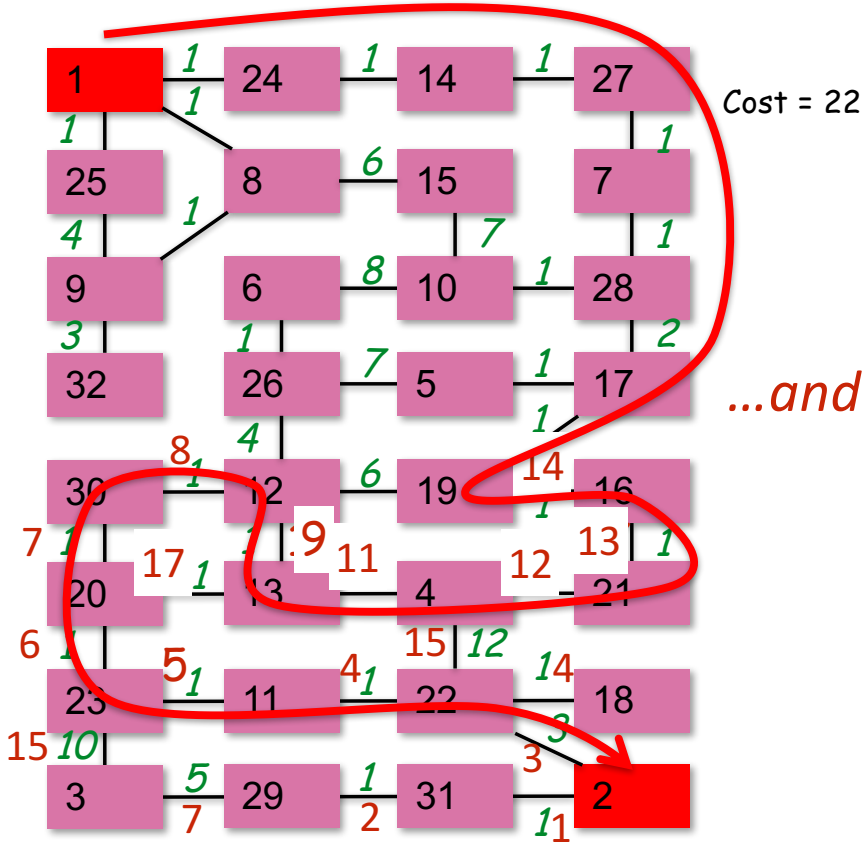
A more complicated topology

Find the lowest cost path from 1 to 2



Router 4 tells its neighbors:
"I can reach 2 with a cost of 15"

Solution


1
24
14
27
25
8
15
7
9
6
10
28
32
26
5
17
30
12
19
16
20
13
4
21
23
11
22
18
3
29
31
2
Cost = 22
...and so on!

A second algorithm

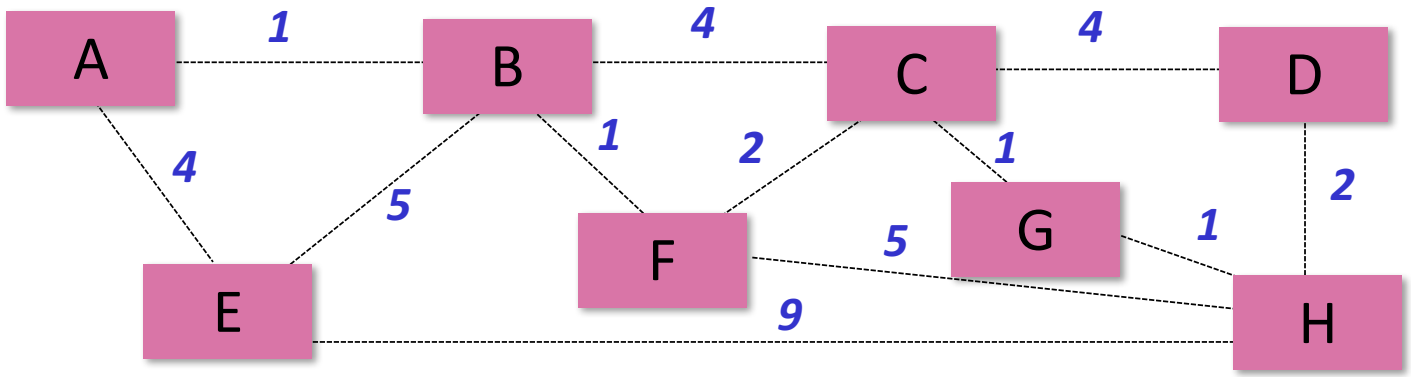
Dijkstra's shortest path first algorithm

(An example of a "Link State Algorithm")

1. **Exchange link state**: A router floods to every other router the state of links connected to it.
   - Periodically
   - When link state changes
2. **Run Dijkstra's algorithm**: Each router independently runs Dijkstra's shortest path first algorithm.

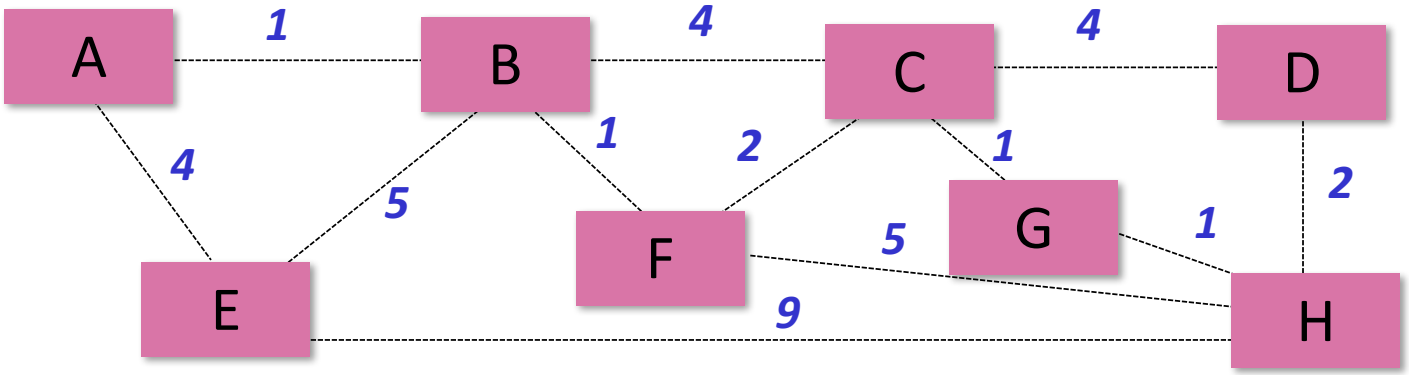
Each router runs an algorithm to find the lowest cost spanning tree to reach every other router.

Find lowest cost path from H to every router

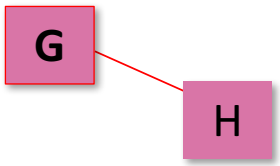


"Every router learns the topology from the flooded link-state"

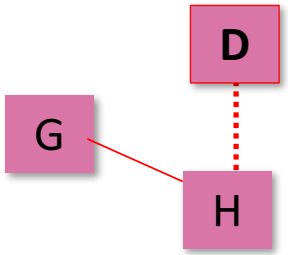
Find lowest cost path from H to every router



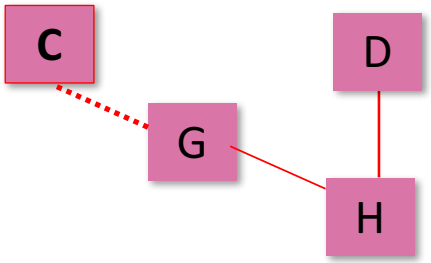
Add path of cost 1



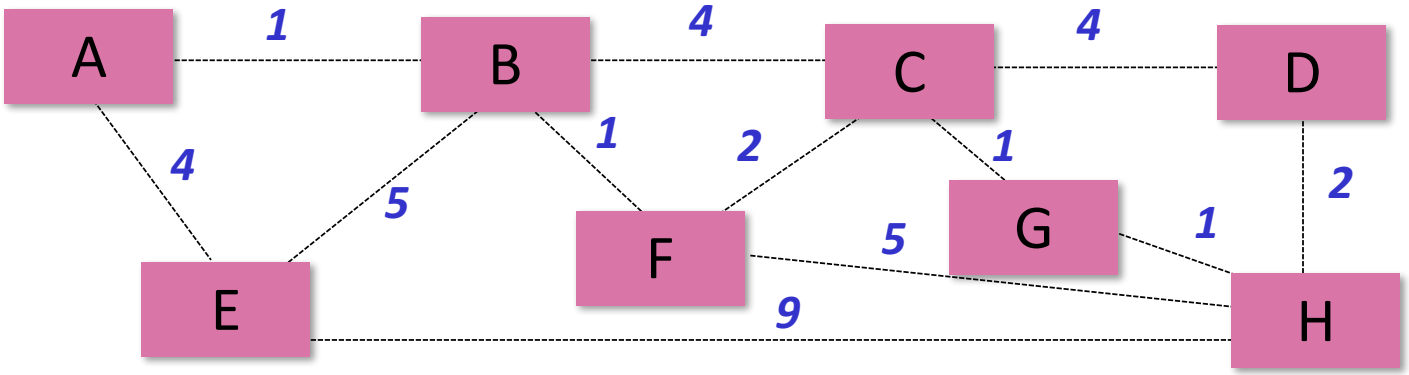
Add path of cost 2



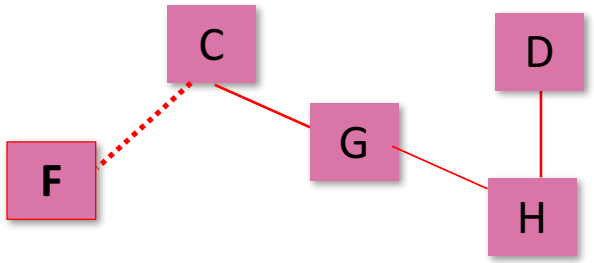
Add path of cost 2



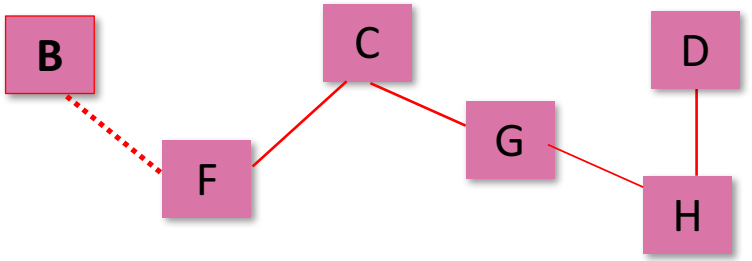
Find lowest cost path from H to every router



Add path of cost 4

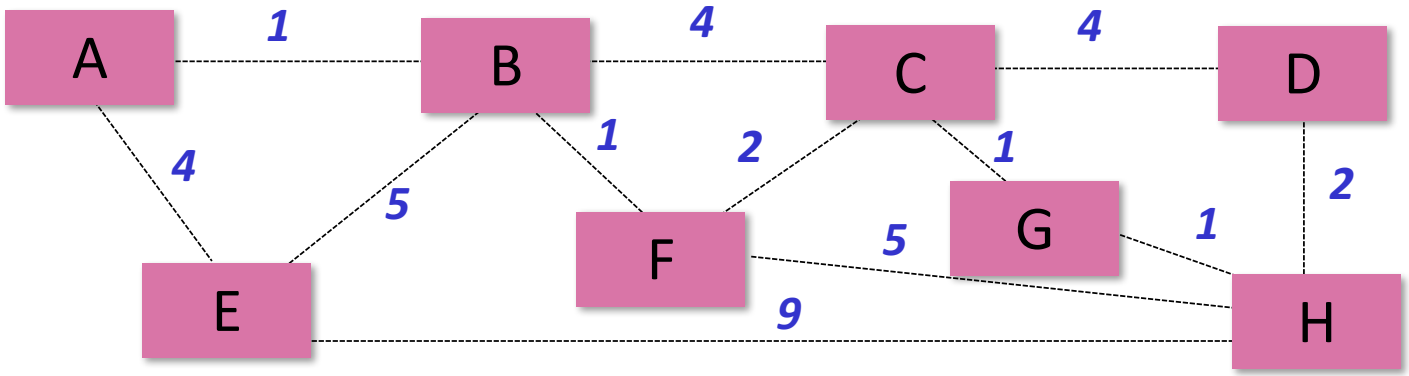


Add path of cost 5



.....

How to find the lowest cost tree manually



	0	1	2	3	4	5	6	7
Shortest Path Set	H							
Candidate Set	DEFG							
Add	G							

Dijkstra's Algorithm

Questions:

1. How long does the algorithm take to run?
2. What happens when link costs change, or when routers/links fail?

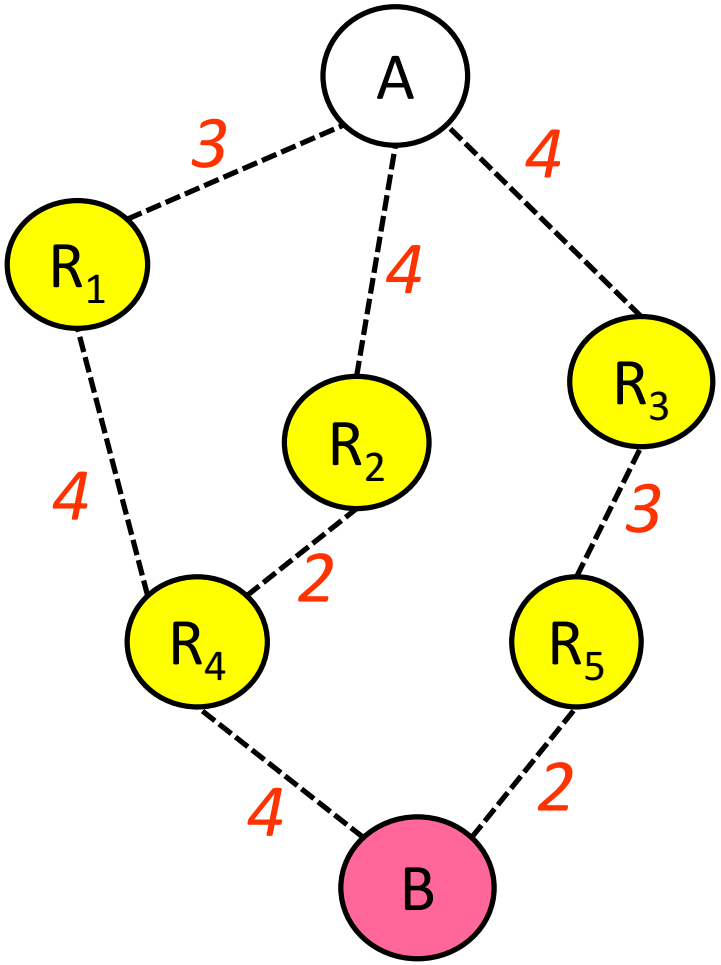
Dijkstra's algorithm in practice

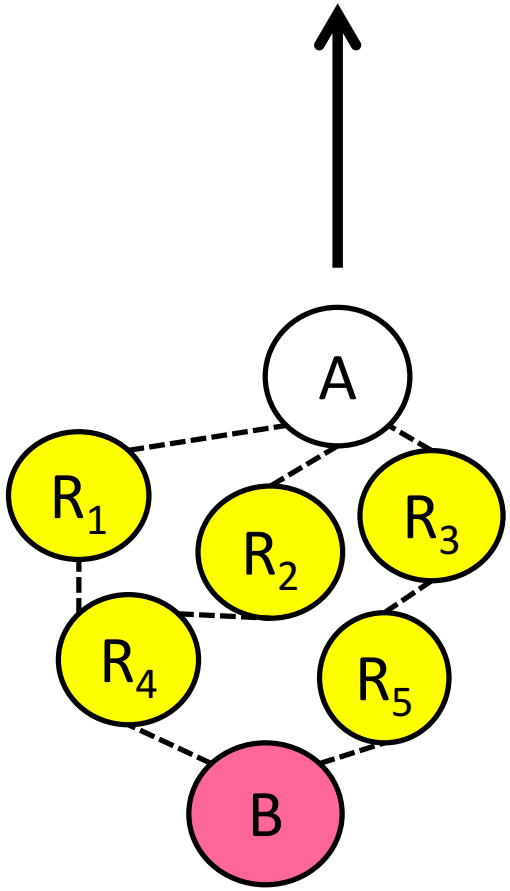
Dijkstra's algorithm is an example of a Link State algorithm.

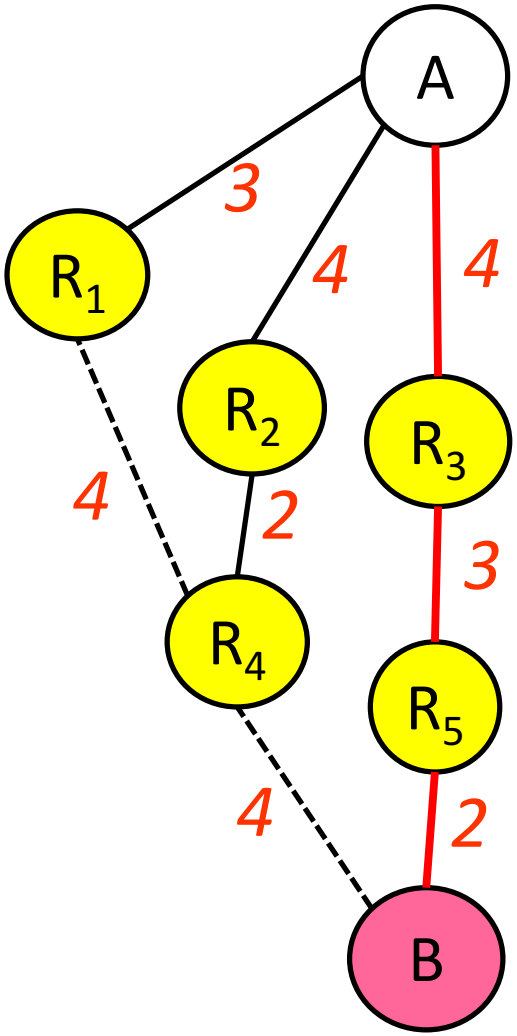
- Link state is known by every router.
- Each router finds the shortest path spanning tree to every other router.

It is the basis of OSPF (Open Shortest Path First), a very widely used routing protocol.

Another view of Dijkstra…


A
3
4
R1
4
R2
R3
4
3
2
R4
R5
4
2
B

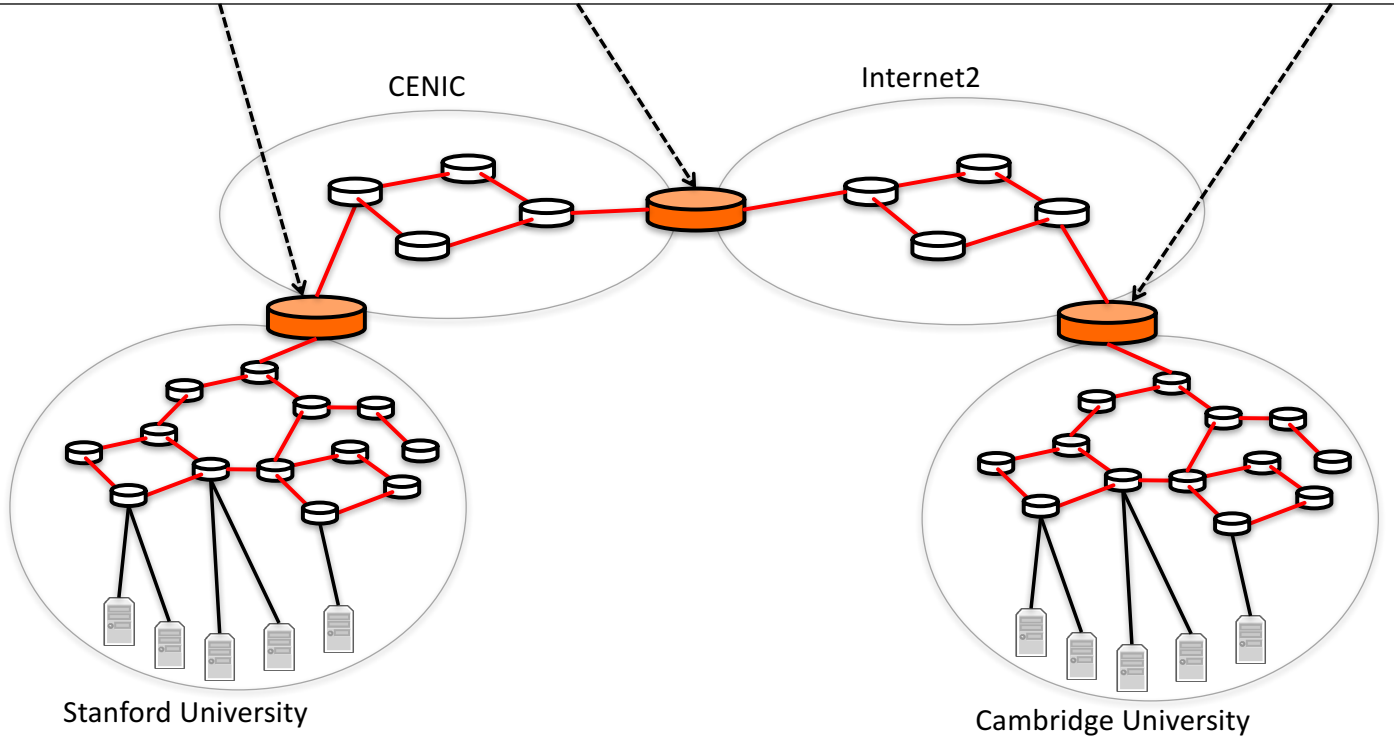

A
R_1
R_2
R_3
R_4
R_5
B


A
R1
R2
R3
R4
R5
B
3
4
4
4
2
3
4
2

Routing between Autonomous Systems

The Border Gateway Protocol (BGP)

All organizations (Autonomous Systems) use the same algorithm to talk to each other



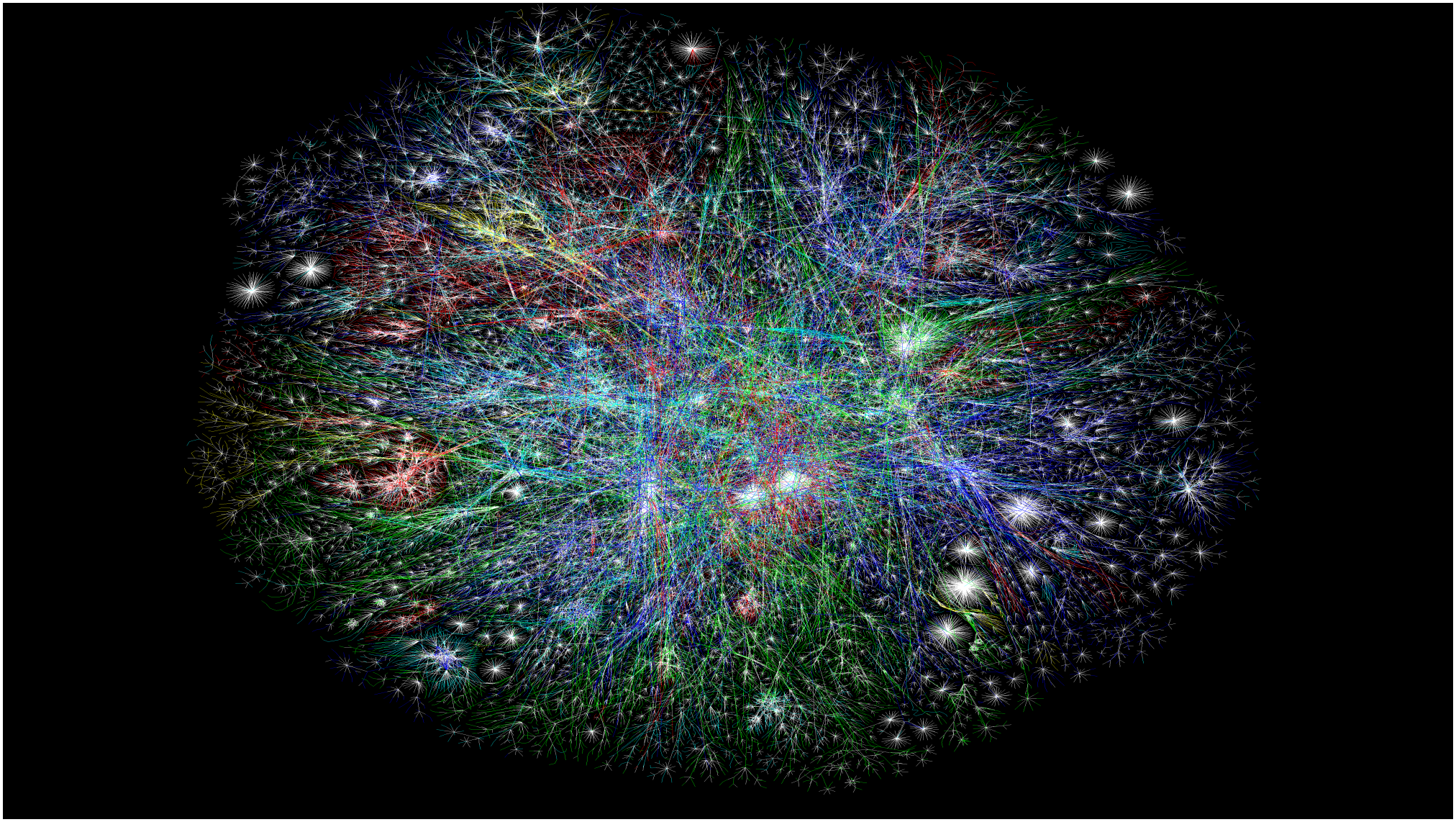
AS (Autonomous System) numbers

nickm> traceroute -q1 www.cam.ac.uk

```
traceroute to www.cam.ac.uk (131.111.150.25),
    30 hops max, 40 byte packets
 1  csmx-west-rtr.SUNet (171.64.74.2)  8.567 ms
 2  dc-svl-rtr-vl8.SUNet (171.64.255.204)  0.334 ms
 3  dc-svl-agg4--stanford-100ge.cenic.net
    (137.164.23.144)  1.041 ms
 …
 7  et-4-0-0.4079.sdn-sw.lasv.net.internet2.edu
    (162.252.70.28)  14.320 ms
…
14  internet2.mx1.lon.uk.geant.net
    (62.40.124.44)  144.085 ms
15  janet-gw.mx1.lon.uk.geant.net
    (62.40.124.198)  144.552 ms
…
24  primary.admin.cam.ac.uk (131.111.150.25)  150.353
    ms
```

nickm> whois -h whois.cymru.com 62.40.124.198

```
[Querying whois.cymru.com]
[whois.cymru.com]
AS      | IP               | AS Name
20965   | 62.40.124.198    | GEANT The GEANT IP Service, GB
21320   | 62.40.124.198    | GEANT_IAS_VRF, EU
```



Border Gateway Protocol (BGP)

- BGP neighbors ("peers") establish a TCP connection.
- BGP is not a link-state or a distance-vector routing protocol.
- Instead, BGP uses what is called a "Path vector".

- For each prefix, a BGP router advertises a path of AS's to reach it.
  - This is the "path vector"
  - Example of path vector advertisement:

  "The network 171.64/16 can be reached via the path {AS1, AS5, AS13}"

- When a link/router fails, the path vector is "withdrawn"

Border Gateway Protocol (BGP)

"The network 171.64/16 can be reached via the path {AS1, AS5, AS13}"

Paths with loops are detected locally and ignored.

Local policies pick the preferred path among all advertised paths.

Current Version

BGP Version 4 or “BGP-4”

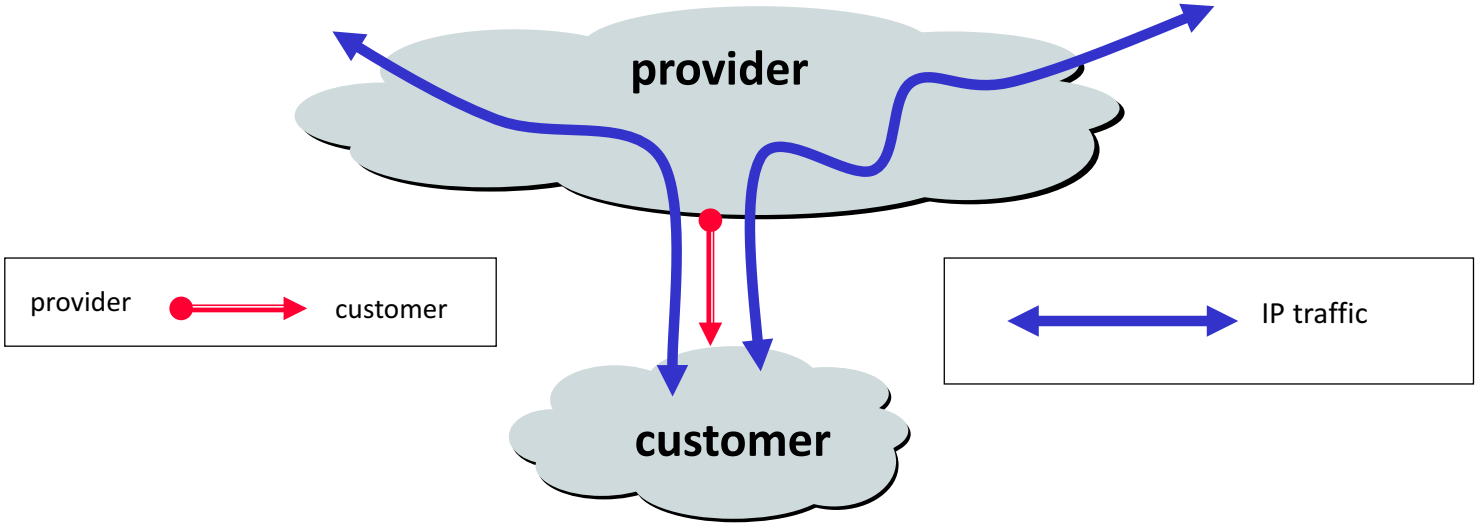
Described in RFC 4271 (2006)

Used by the Internet since 1994

Used to connect all Border Gateways in the Internet

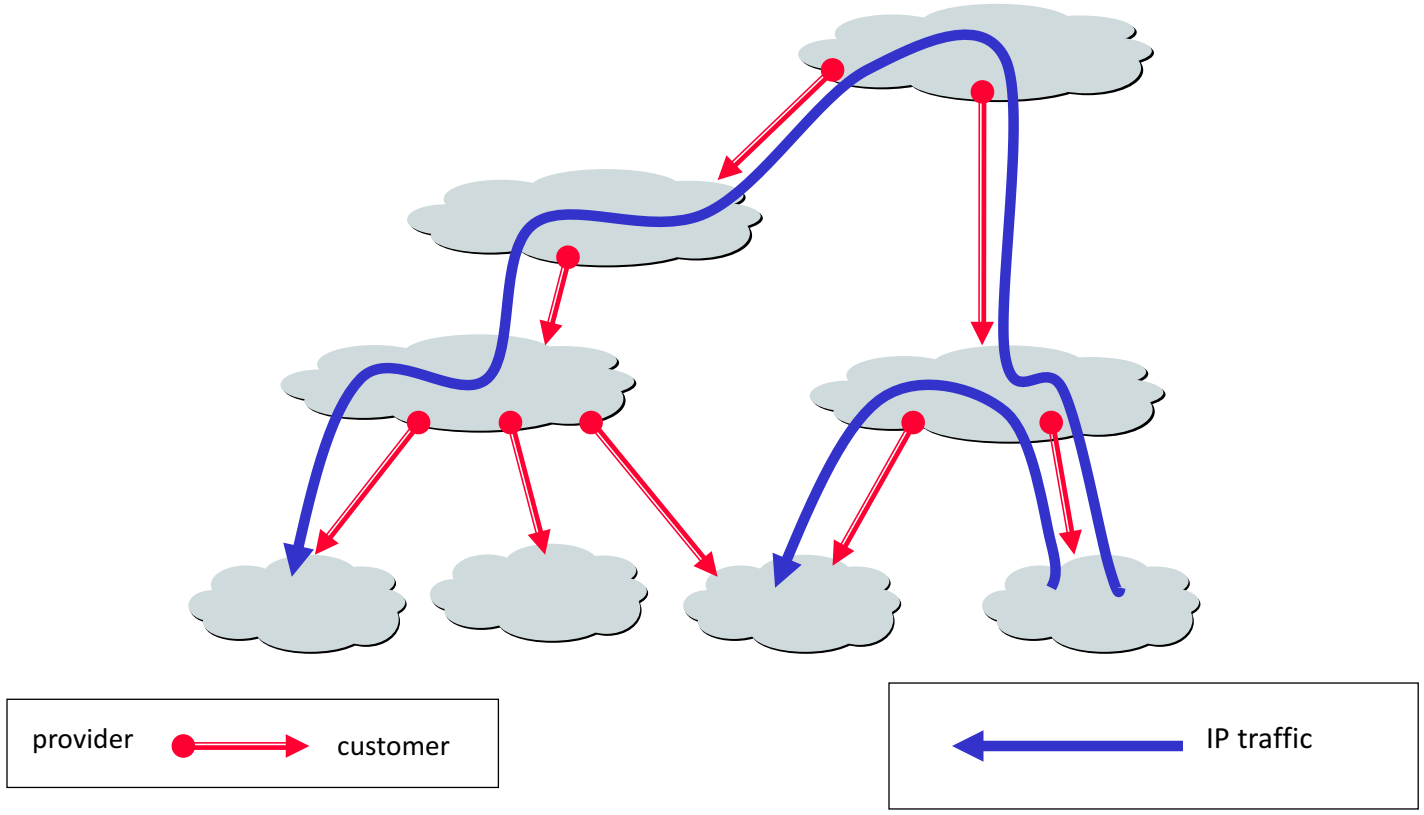
Used internally in some very large private networks, such as data-centers

Customers and Providers

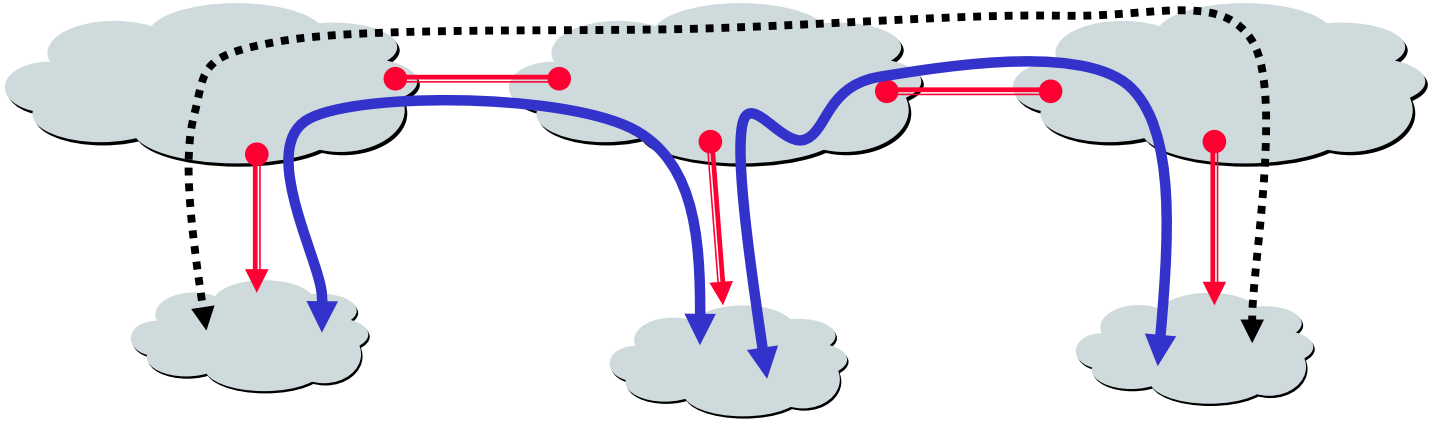


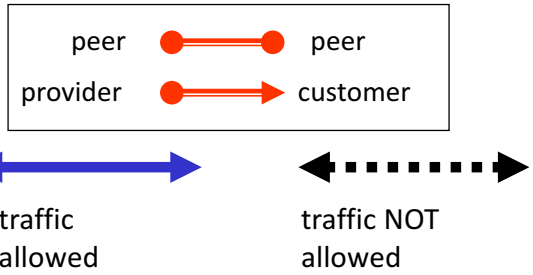
Customer pays provider to carry its packets.

Customer-Provider Hierarchy



The Peering Relationship





Peers provide transit between their respective customers

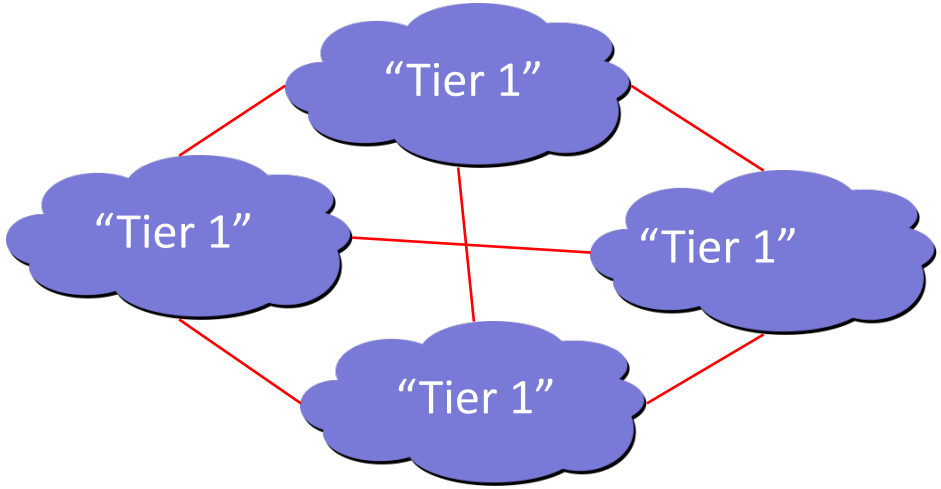
Peers do not provide transit between peers

Peers (typically) do not exchange $$$

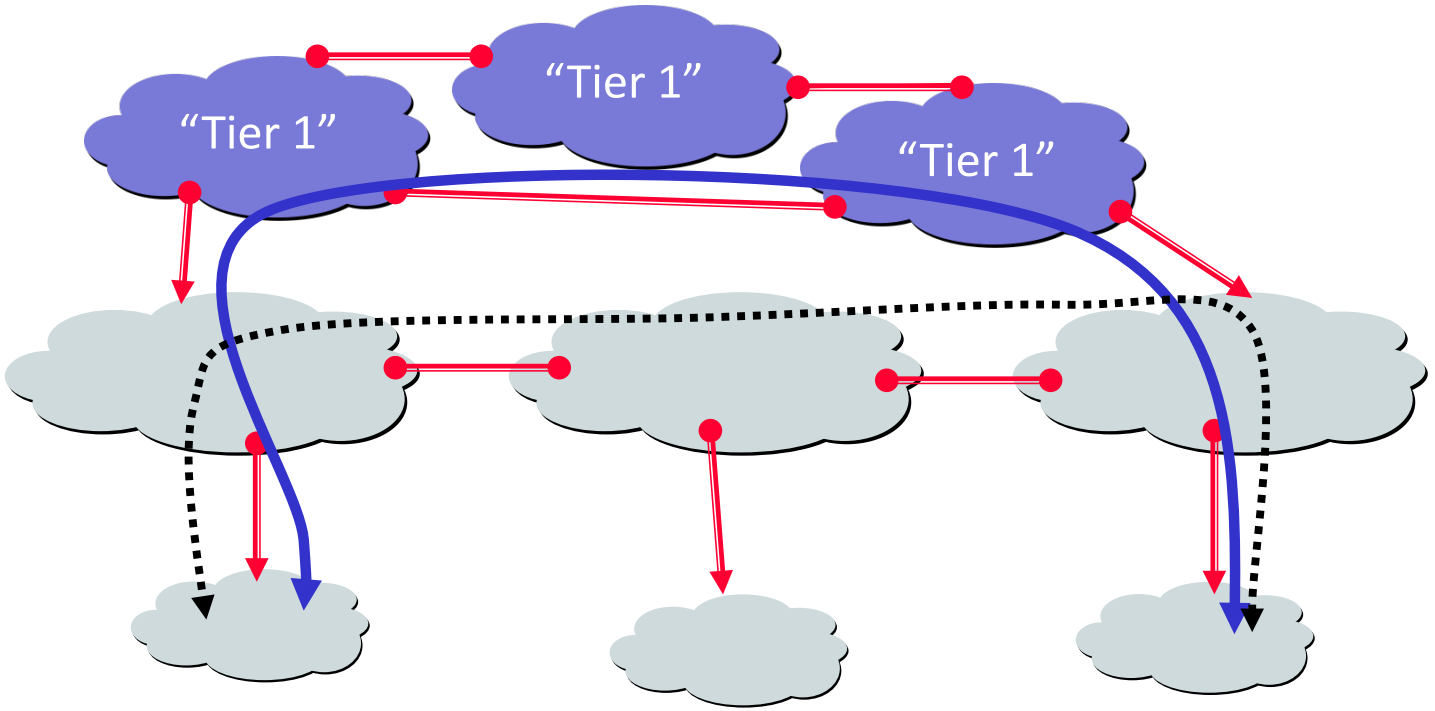
So how does traffic from the left side reach the right side?

“Tier 1” Providers

A tier 1 network is a transit-free network that peers with every other tier 1 network



Examples: AT&T, CenturyLink, Level 3, Deutsche Telekom, Tata, Verizon, …



peer — peer

provider — customer

traffic allowed

traffic NOT allowed

Peers provide transit between their respective customers

Peers do not provide transit between peers

Peers (typically) do not exchange $$$

BGP Messages

Open : Establish a BGP session.

Keep Alive : Handshake at regular intervals.

Notification : Shuts down a peering session.

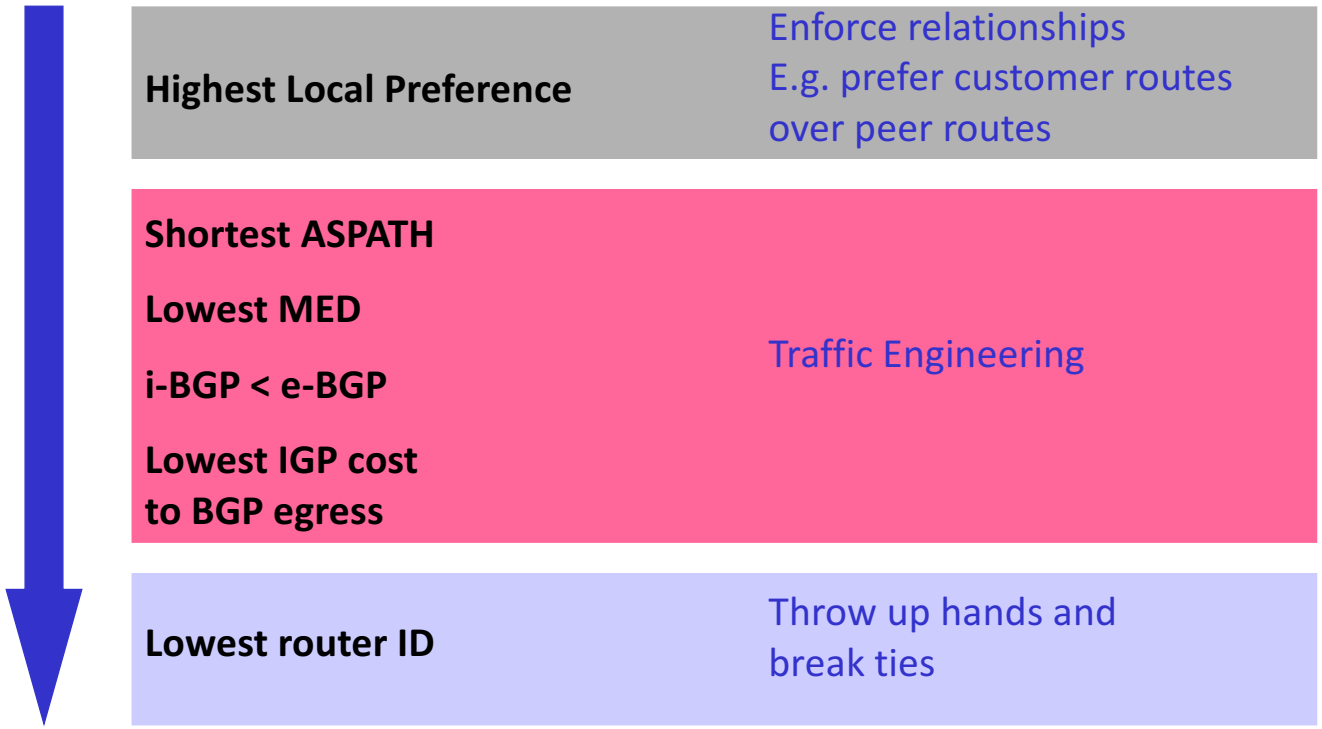
Update : Announcing new routes or withdrawing previously announced routes.

BGP announcement = prefix + path attributes

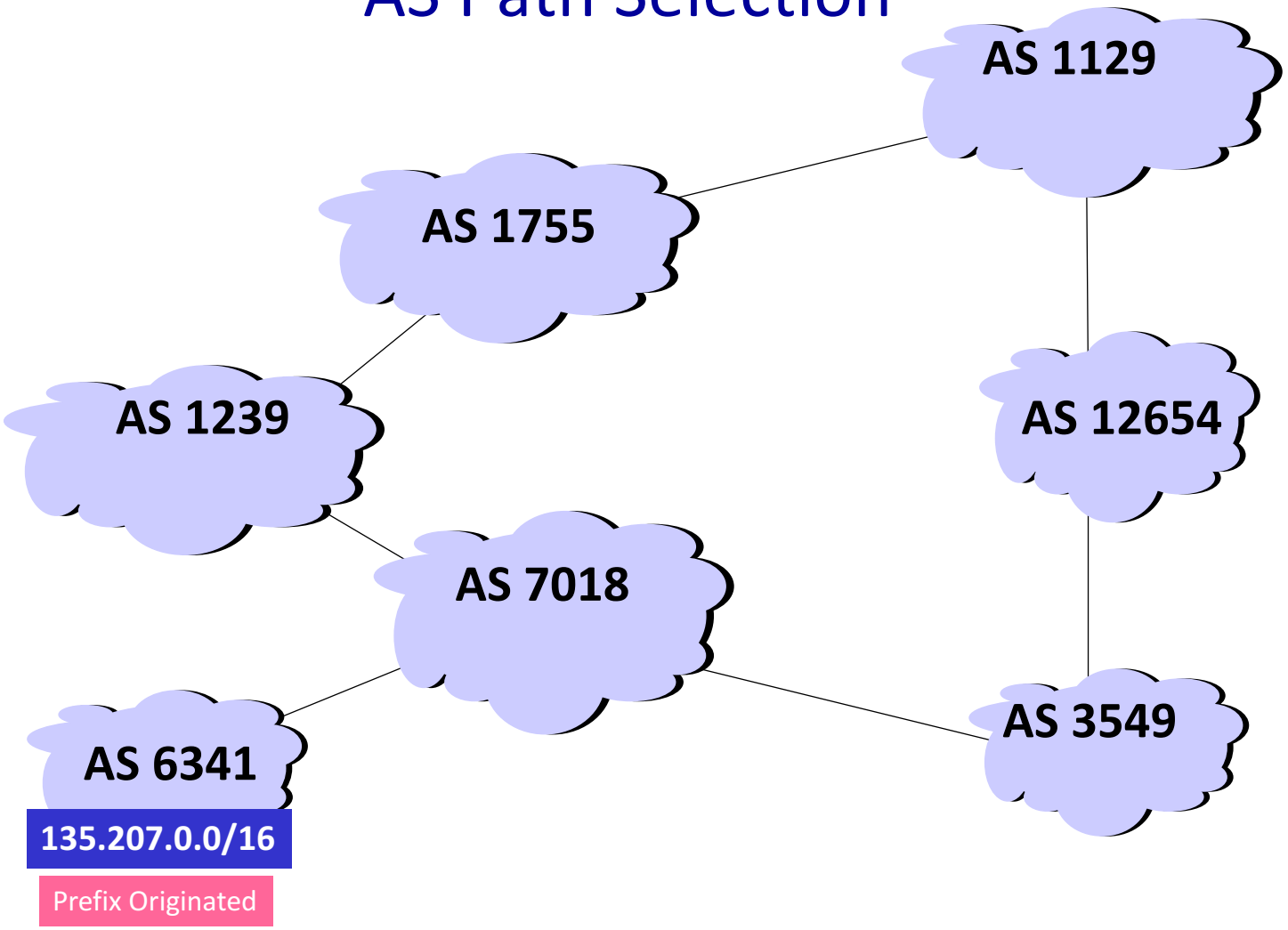
Path attributes

- Include: next hop, AS Path, local preference, Multi-exit discriminator, …
- Used to select among multiple options for paths.

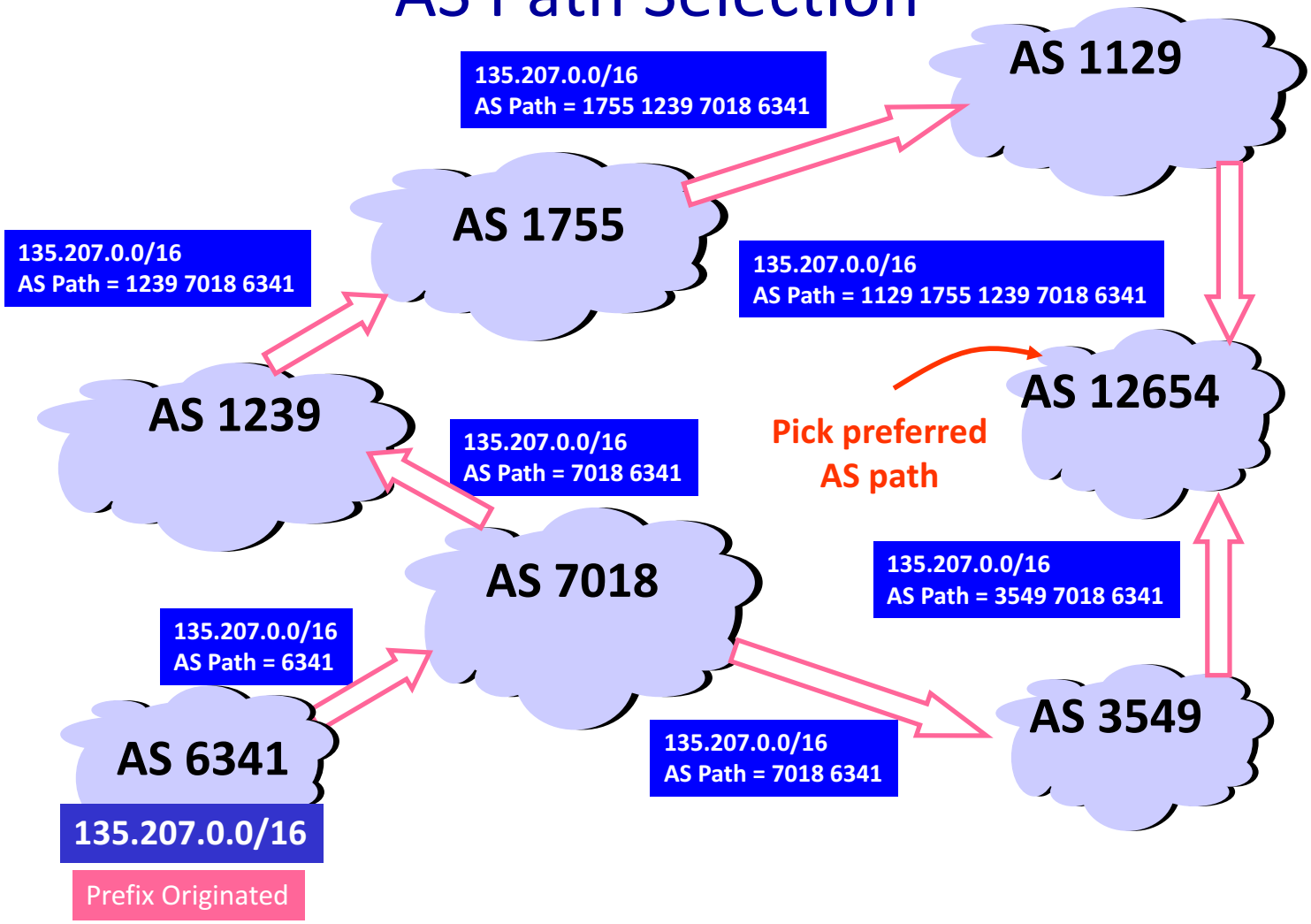
BGP Route Selection Summary



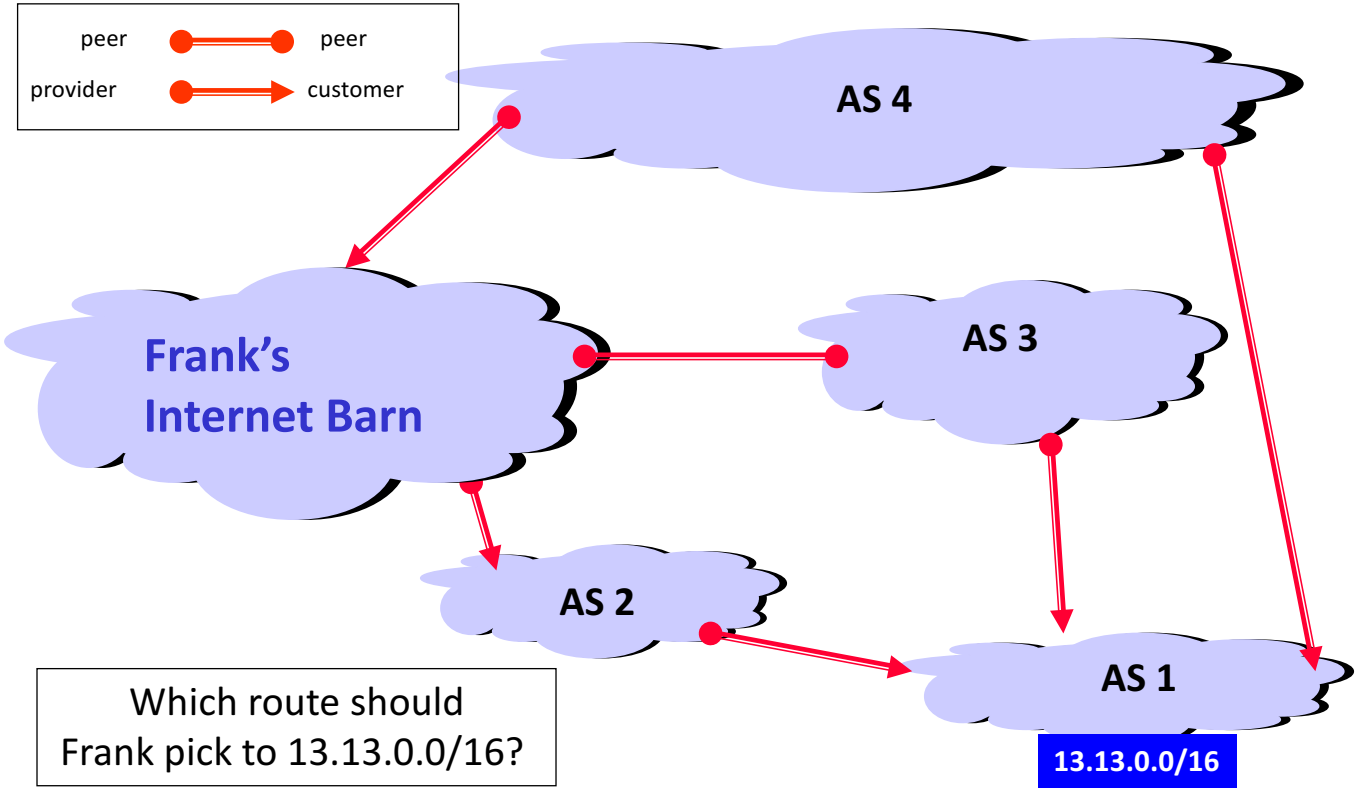
AS Path Selection



AS Path Selection



So Many Choices…



Summary of BGP

All AS's in the Internet must connect using BGP-4.

BGP-4 is a path vector algorithm, allowing loops to be detected easily.

BGP-4 has a rich and complex interface to let AS's choose a local, private policy.

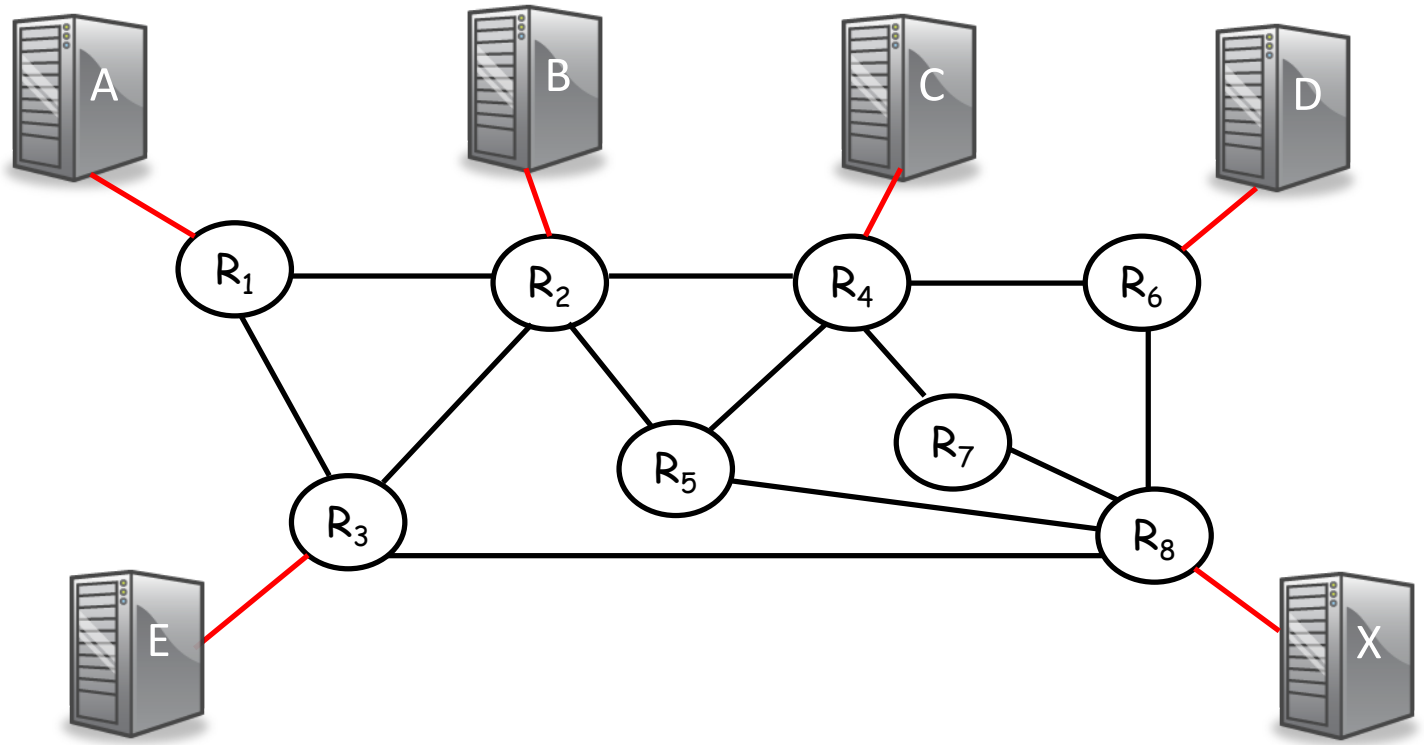
Each AS decides a local policy for traffic engineering, security and any private preferences.

Thank you!

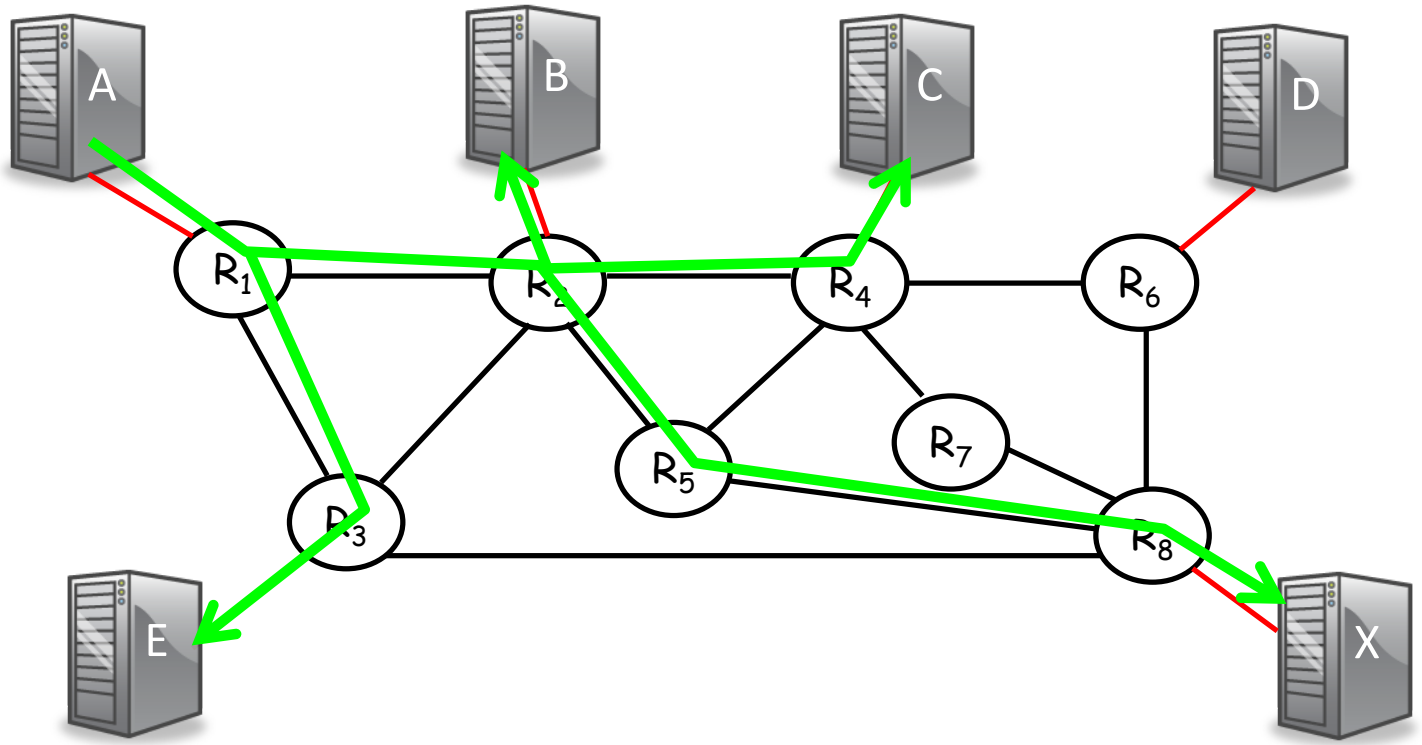
If there is time….

Multicast routing in the Internet

Multicast



Multicast



Multicast

Techniques and Principles

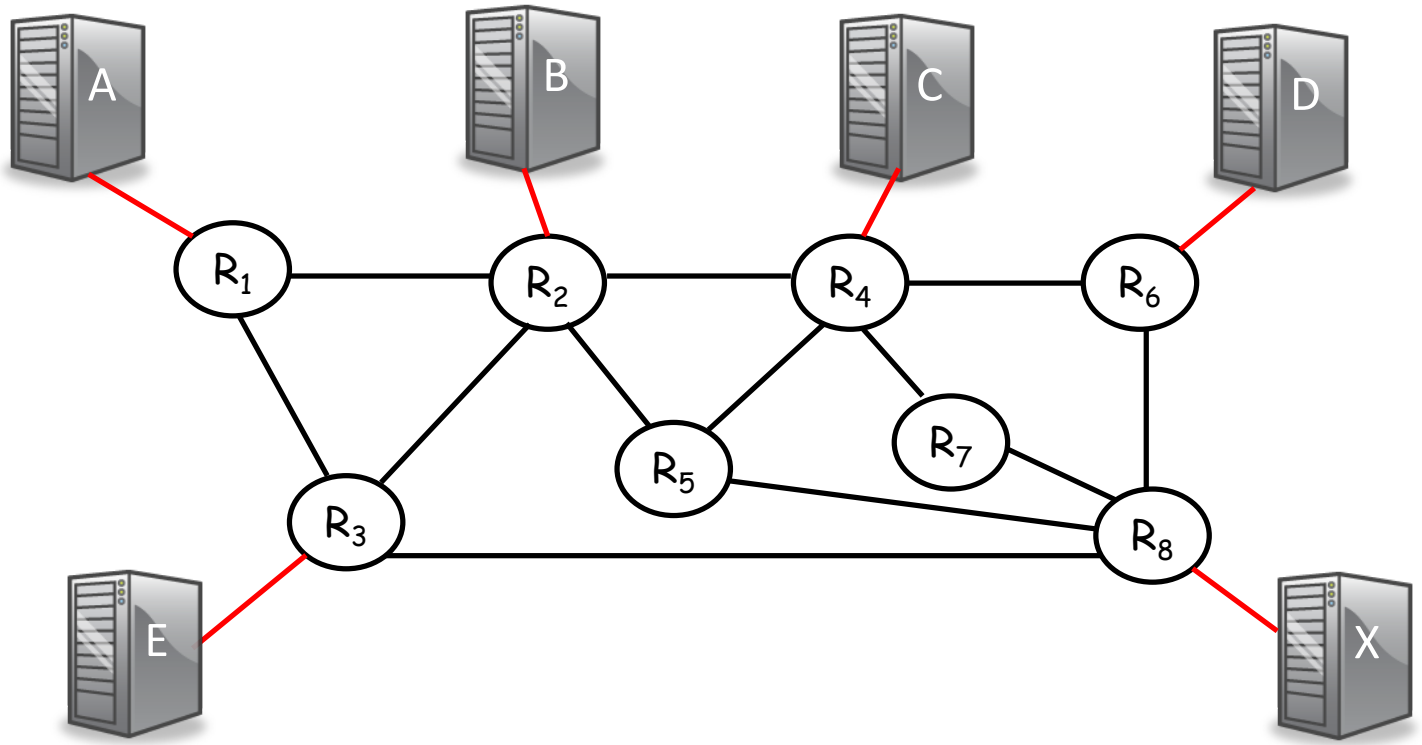
- Reverse Path Broadcast (RPB) and Pruning
- One versus multiple trees

Practice

- IGMP – group management
- DVMRP – the first multicast routing protocol
- PIM – protocol independent multicast

Reverse Path Broadcast (RPB)

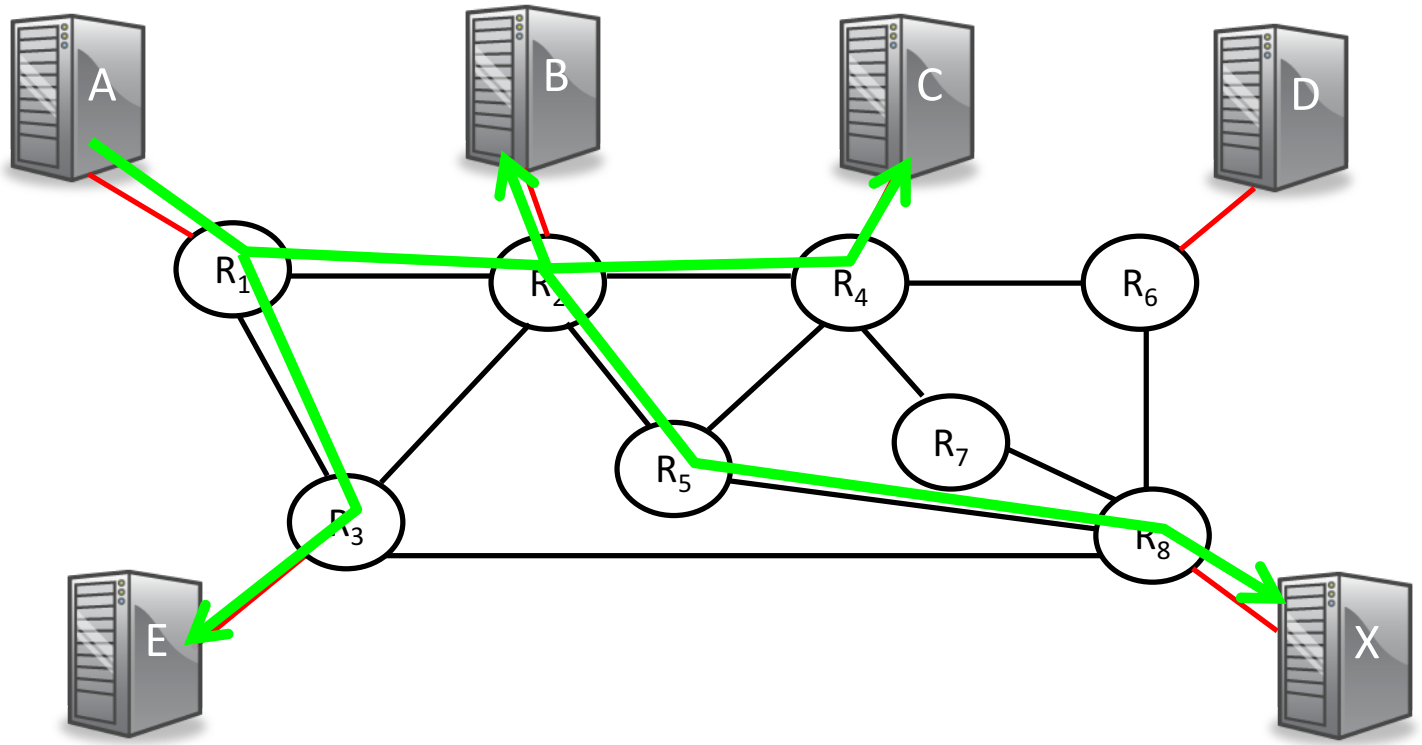
aka Reverse Path Forwarding (RPF)



RPB + Pruning

1. Packets delivered loop-free to every end host.
2. Routers with no interested hosts send prune messages towards source.
3. Resulting tree is the minimum cost spanning tree from source to the set of interested hosts.

One tree versus several trees?



Addresses and joining a group

IPv4: Class D addresses are set aside for multicast.

IGMP (Internet group management protocol)

- Between host and directly attached router.
- Hosts ask to receive packets belonging to a particular multicast group.
- Routers periodically poll hosts to ask which groups they want.
- If no reply, membership times out (soft-state).

Multicast in practice

Multicast used less than originally expected

- Most communication is individualized (e.g. time shifting)
- Early implementations were inefficient
- Today, used for some IP TV and fast dissemination
- Some application-layer multicast routing used

Some interesting questions

- How to make multicast reliable?
- How to implement flow-control?
- How to support different rates for different end users?
- How to secure a multicast conversation?