ELEC / COMP 177 – Fall 2012

Computer Networking → Routing Protocols (2)

Some slides from Kurose and Ross, *Computer Networking*, 5^{th} Edition

Schedule

- **Project #2** – Due Tuesday, Nov 6th
- **Homework #5** – Due Tuesday, Nov 13th
- ***Later this semester:***
  - ***Homework #6 - Presentation on security/privacy***
    - ***Topic selection*** *– Due Tuesday, Nov 20th*
    - **Slides** – Due Monday, Nov 26th
    - **Present!** – Tuesday, Nov 27th (and Thursday)
  - ***Project #3*** *– Due Tuesday, Dec 4th*

Recap – Forwarding versus Routing

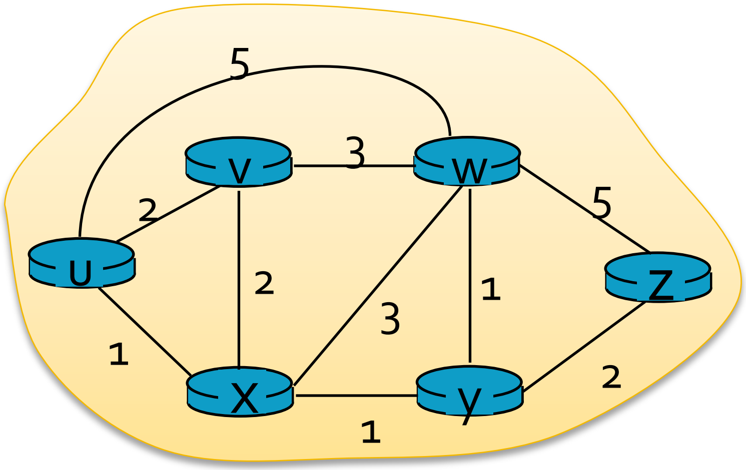
- Forwarding
  - Move packets from router's input to appropriate router output
  - Router does a *longest prefix match* (LPM) on entries in the forwarding table to determine output port
- Routing
  - Determine path (route) taken by packets from source to destination
  - Routing algorithms

Recap – Routing Algorithm Classification

- **Global Information**
  - All routers have complete topology, link cost info
  - **“link state” algorithms**
- **Decentralized**
  - Router knows physically-connected neighbors and link costs to neighbors
  - Iterative process of computation, exchange of info with neighbors
  - **“distance vector” algorithms**

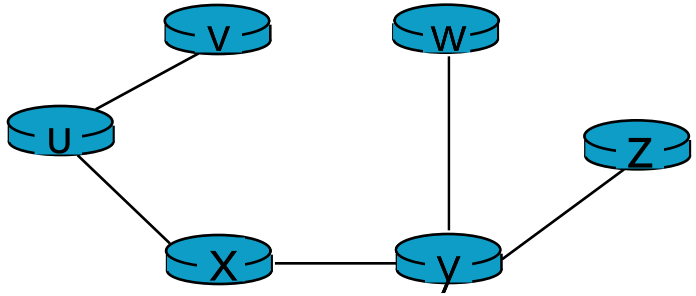
Recap – Link State – Dijkstra's Algorithm

- Network topology and link costs are known to all nodes
  - Accomplished via "link state broadcast"
  - All nodes have same info
- Computes least cost paths from one node (source) to all other nodes
  - Produces **forwarding table** for that node
- Iterative: after k iterations, know least cost path to k destinations



Recap – Link State – Dijkstra's Algorithm

Resulting shortest-path tree from u:



Resulting forwarding table in u:

destination	link
v	(u,v)
x	(u,x)
y	(u,x)
w	(u,x)
z	(u,x)

Recap – Distance Vector Algorithm

Iterative, asynchronous:

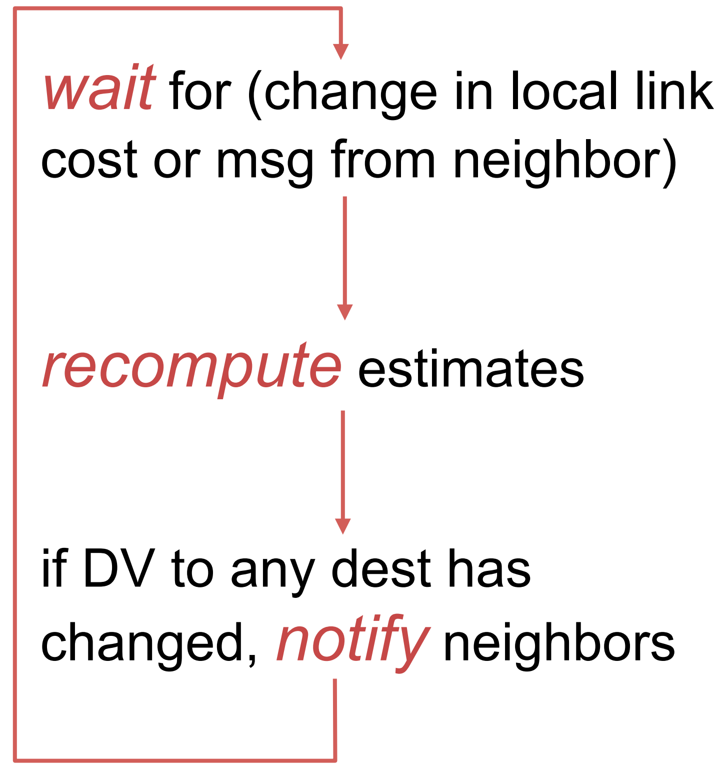
each local iteration caused by:

- local link cost change
- DV update message from neighbor

Distributed:

- each node notifies neighbors *only* when its DV changes
  - neighbors then notify their neighbors if necessary

Each node:



Recap – Distance Vector – Bellman-Ford Equation

Define:

$d_x(y)$:= cost of least-cost path from *x* to *y*

Then:

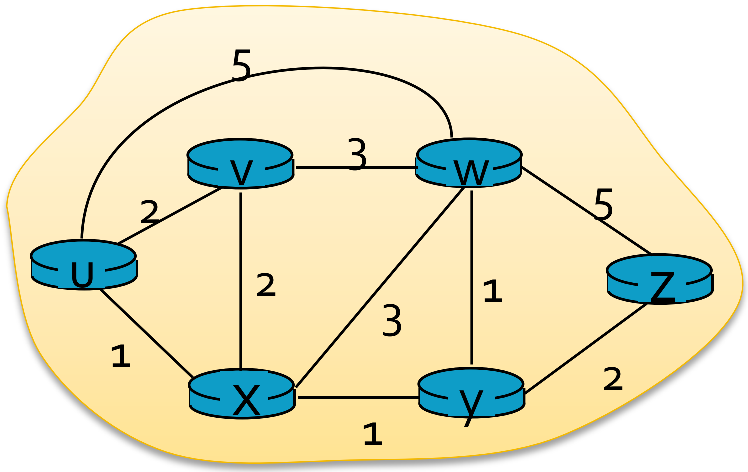
Something I know...

Something my neighbor told me...

$$d_x(y) = \min_v \{c(x,v) + d_v(y)\}$$

where min is taken over all neighbors *v* of *x*

Recap – Distance Vector – Bellman-Ford



Clearly, $d_v(z) = 5$, $d_x(z) = 3$, $d_w(z) = 3$

B-F equation says:

$$d_u(z) = \min \{ c(u,v) + d_v(z), \; c(u,x) + d_x(z), \; c(u,w) + d_w(z) \}$$

$$= \min \{2 + 5, \; 1 + 3, \; 5 + 3\} = \mathbf{4}$$ *(by way of x!)*

The node that provides the minimum cost is entered in the router forwarding table as the next hop

Today

- Continue discussing network layer
- **Routing algorithms used in the Internet**
  - **Routing Information Protocol (RIP)**
  - **Open Shortest Path First (OSPF)**
  - **Border Gateway Protocol (BGP)**

Recap – Hierarchical Routing

- Our routing discussion thus far has been idealized
  - All routers are identical
  - The network is “flat”
- This is not true in practice!

- Problem 1 – **Scale**
  - Hundreds of millions of destinations:
  - Can’t store all destinations in routing tables!
  - Routing table exchange would swamp links!
  - Distance-vector would never converge

- Problem 2 - **Administrative autonomy**
  - Internet = network of networks
  - Each network admin wants to control routing in his/her own network

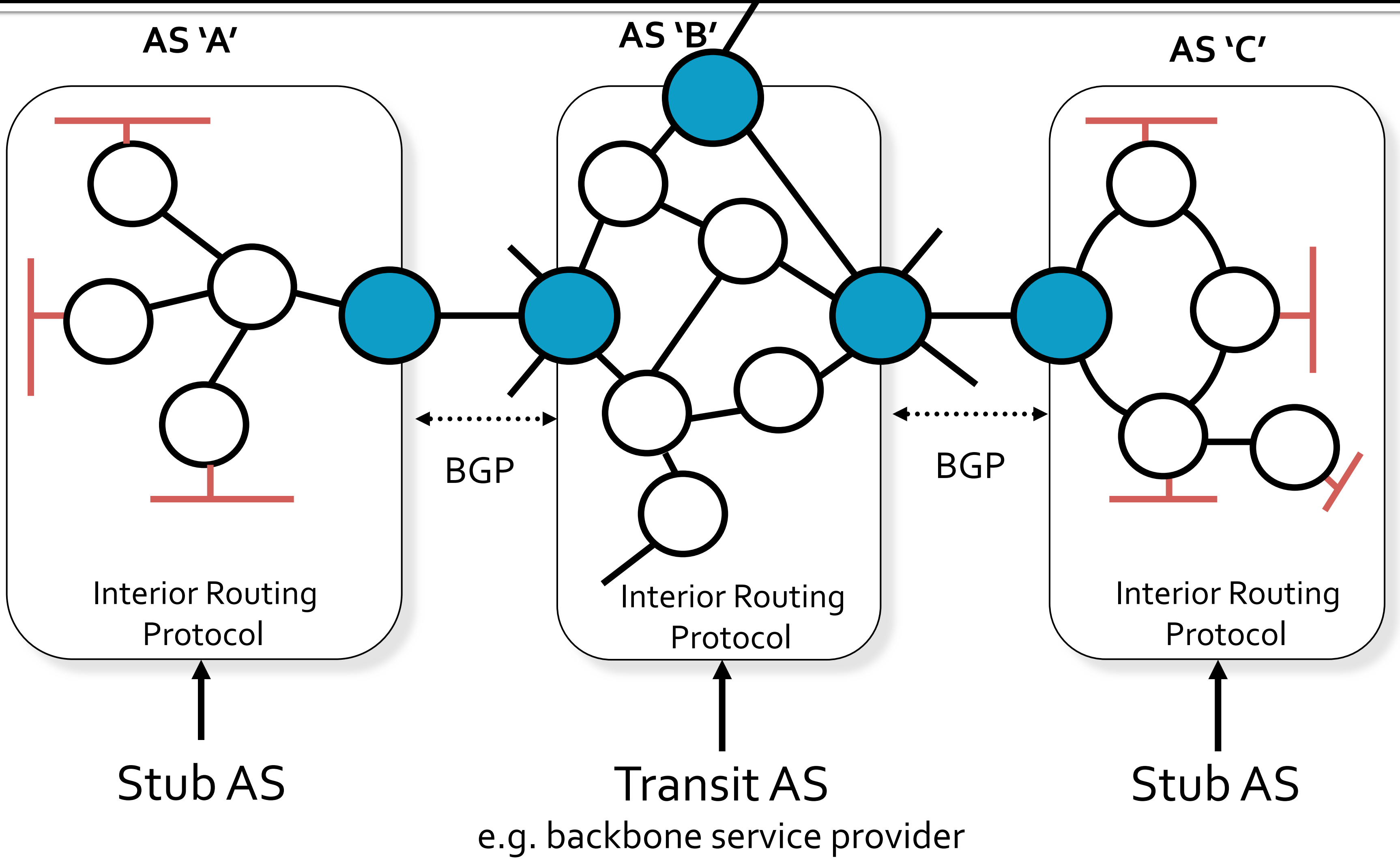
Recap – Hierarchical Routing

- Aggregate routers into regions (aka "**autonomous systems**" - AS)
- Routers inside autonomous system run same routing protocol
  - "Intra-AS" routing protocol
  - Routers in different AS can run different intra-AS routing protocol
- Border Router
  - Direct link to router in another AS

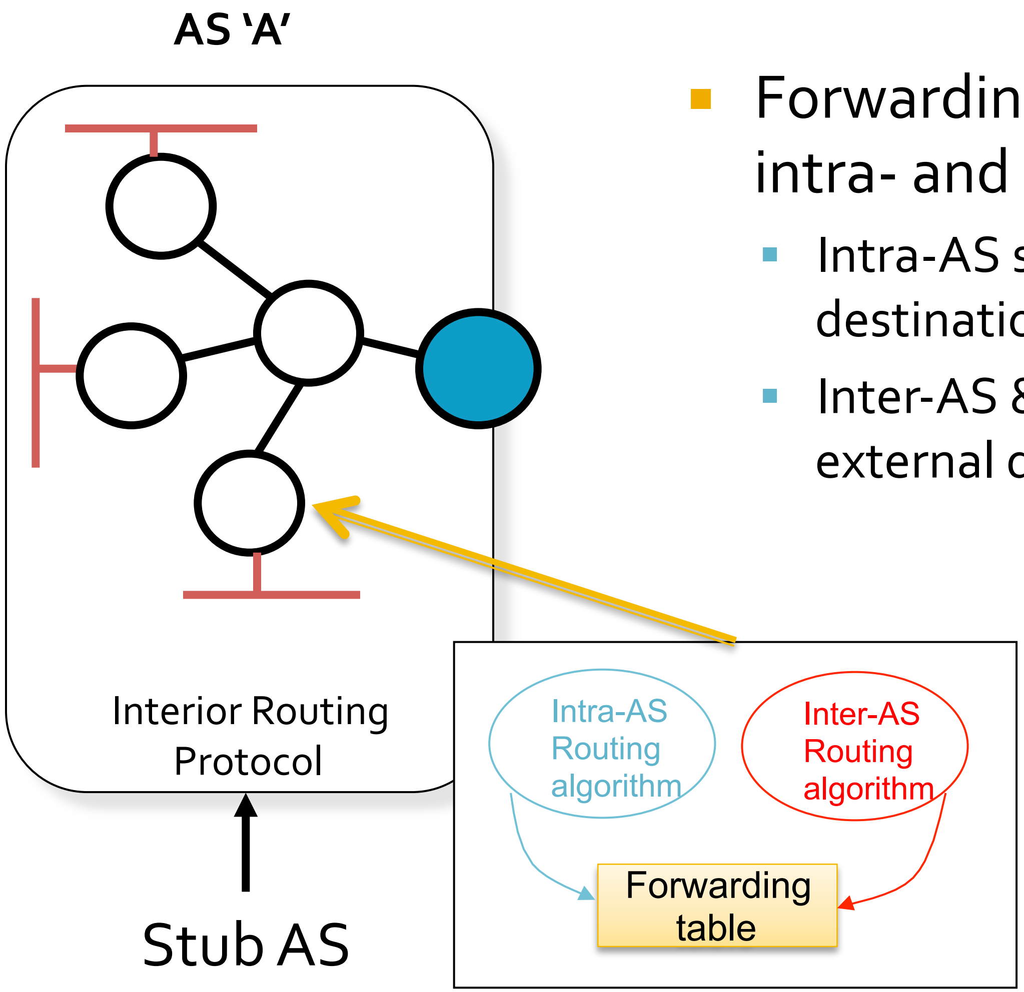
Routing in the Internet

- The Internet uses hierarchical routing
- The Internet is split into Autonomous Systems
  - "Independent" networks on the Internet
  - Typically owned/controlled by a single entity
  - Share a common routing policy
- Example autonomous systems
  - Pacific (18663), Exxon (1766), IBM (16807), Level3 (3356)
- Different routing protocols within and between autonomous systems
  - Interior gateway/routing protocol (e.g. OSPF)
  - Border gateway protocol (e.g. BGP)

Autonomous Systems



Forwarding Table



- Forwarding table configured by both intra- and inter-AS routing algorithm
  - Intra-AS sets entries for internal destinations
  - Inter-AS & intra-As sets entries for external destinations

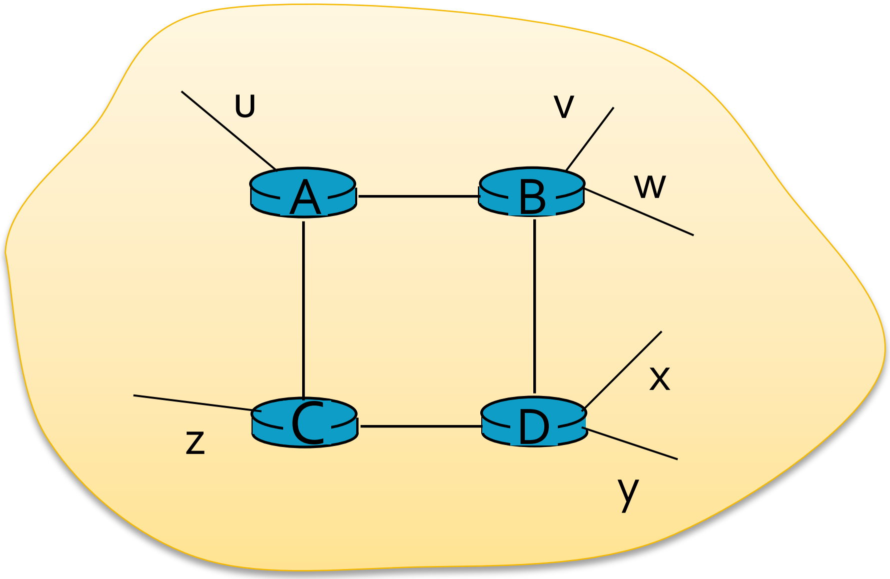
Intra-AS Routing

- Routing *inside* the autonomous system
- Also known as **Interior Gateway Protocols (IGP)**
- Most common Intra-AS routing protocols:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco proprietary)

Routing Information Protocol (RIP)

Routing Information Protocol (RIP)

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



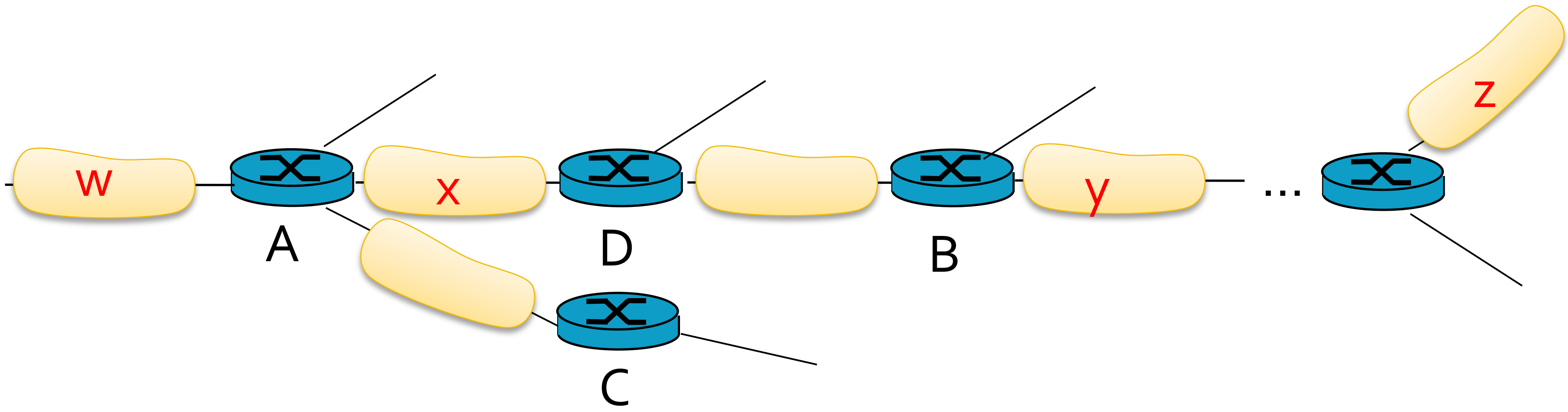
From router A to subnets:

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

RIP advertisements

- Distance vectors
  - Exchanged among neighbors every 30 seconds via Response Message (also called **advertisement**)
- Each advertisement lists up to 25 destination subnets within AS

RIP: Example



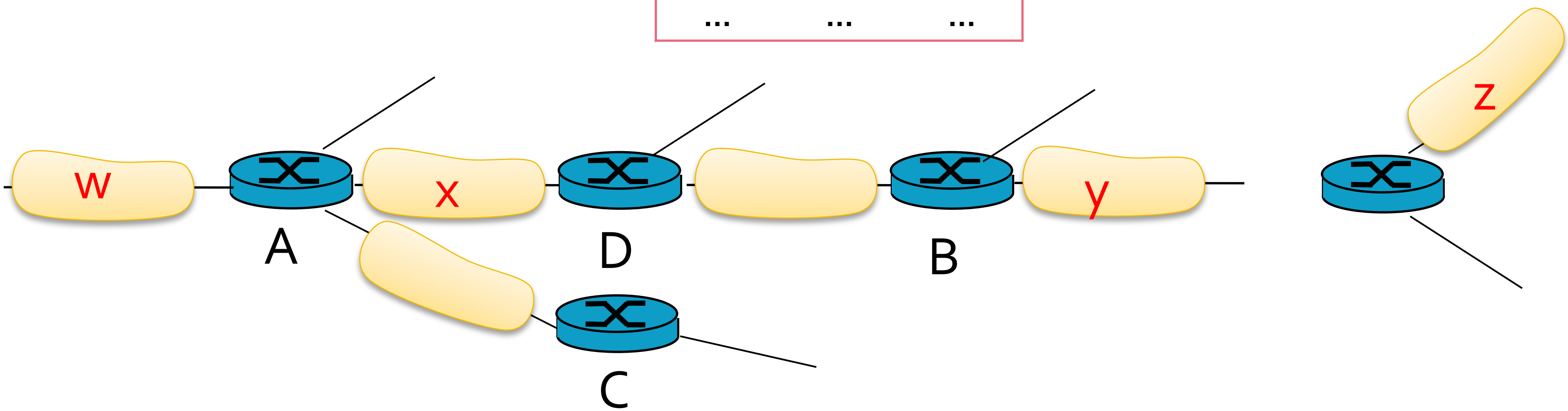
Routing/Forwarding table in D:

Destination Network	Next Router	# of Hops to Destination
w	A	2
y	B	2
z	B	7
x	--	1
...	...	...

RIP: Example

Dest	Next	Hops
w	--	1
x	--	1
z	C	4
…	…	…

Advertisement from A to D



Routing/Forwarding table in D:

Destination Network	Next Router	# of Hops to Destination
w	A	2
y	B	2
z	~~B~~ A	~~7~~ 5
x	--	1
…	…	…

RIP: Link Failure and Recovery

- If no advertisement heard after 180 sec, the neighbor/link declared dead
- Failure recovery
  - Routes via neighbor invalidated
  - New advertisements sent to neighbors
  - Neighbors in turn send out new advertisements (if tables changed)
  - Link failure info "quickly" propagates to entire net

Open Shortest Path First (OSPF)

Open Shortest Path First Routing

- Networks are partitioned into “areas”
  - OSPF only runs within a specific area
  - Other protocols (i.e., BGP) used to route outside an area
- Link-state algorithm
  - Each node has full topology map
  - Route computation using **Dijkstra’s algorithm**

Open Shortest Path First Routing

- Routers periodically send “hello” and “link state” packets to their neighbors
  - Learn who your neighbors are dynamically
  - Decide link/router down if no more hellos
  - Announce changes to the topology
  - Broadcast throughout the area
  - Carried in OSPF messages directly over IP (rather than TCP or UDP)

Reliable Flooding of LSPs

- Link state packets (LSP) delivered throughout the area
  - Flooded throughout the area
  - Sequence numbers and TTLs
- Reliable Flooding
  - If newer sequence number, then forward packet over all links other than the ingress link, otherwise drop packet
  - Resend unacknowledged packets
- Link State Detection
  - If no hello packets during dead interval, assume link is down

OSPF Features (not in RIP)

- **Security**: all OSPF messages authenticated
  - To prevent malicious intrusion
- **Multiple** same-cost **paths** allowed
  - Only one path in RIP
- For each link, multiple cost metrics for different **TOS** (e.g., satellite link cost set "low" for best effort; high for real time)
- Scalable to larger networks (can divide 1 large AS into multiple OSPF "areas")

Routing Across Borders

- **Can we use OSPF Internet-wide?**
- No! OSPF still has scalability limits
  - Broadcasts all link states to all routers
    - Consumes bandwidth
  - Calculates shortest path to all routers
    - Consumes router CPU time?
- Autonomous systems are independent
  - Run by different organizations
  - May use different link cost metrics

Routing Across Borders

- Need a "border gateway protocol"
  - Global routing protocol across autonomous systems
- Global connectivity is at stake!
  - Must settle on one protocol
- What are the requirements?
  - Scalability
  - Flexibility in choosing routes

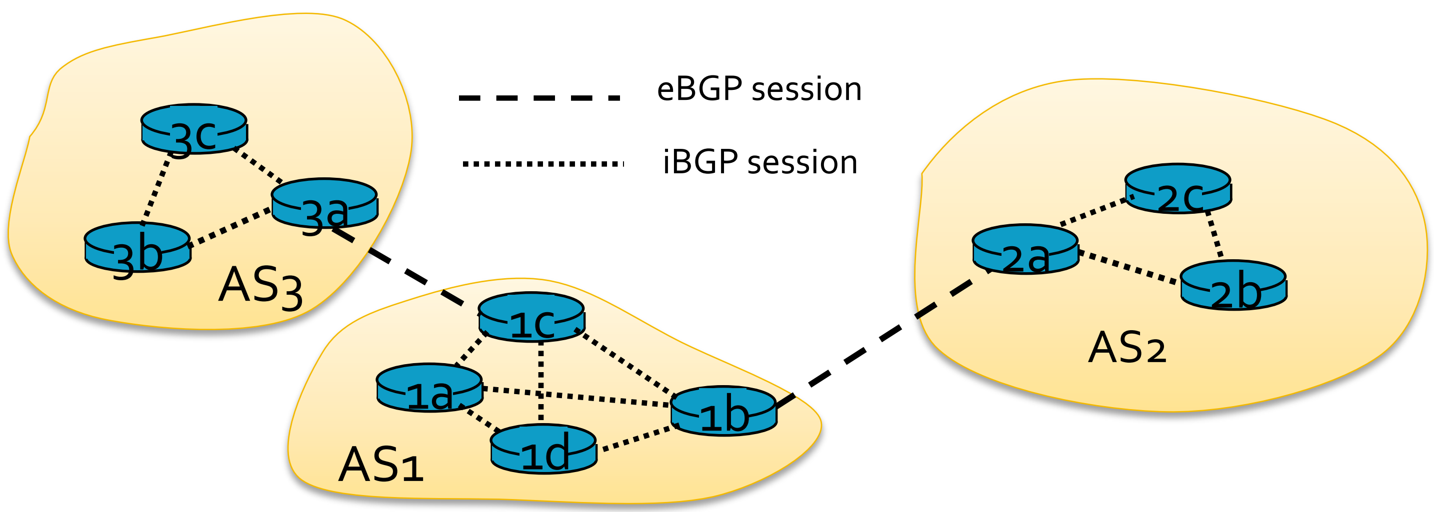
Border Gateway Protocol (BGP)

Internet Inter-AS routing: BGP

- BGP is the **de facto standard**
- BGP provides each AS a means to:
  - Obtain subnet reachability information from neighboring ASs
  - Propagate reachability information to all routers inside an AS
  - Determine “good” routes to subnets based on reachability information and policy
- **Allows subnet to advertise its existence to rest of Internet: “I am here”**

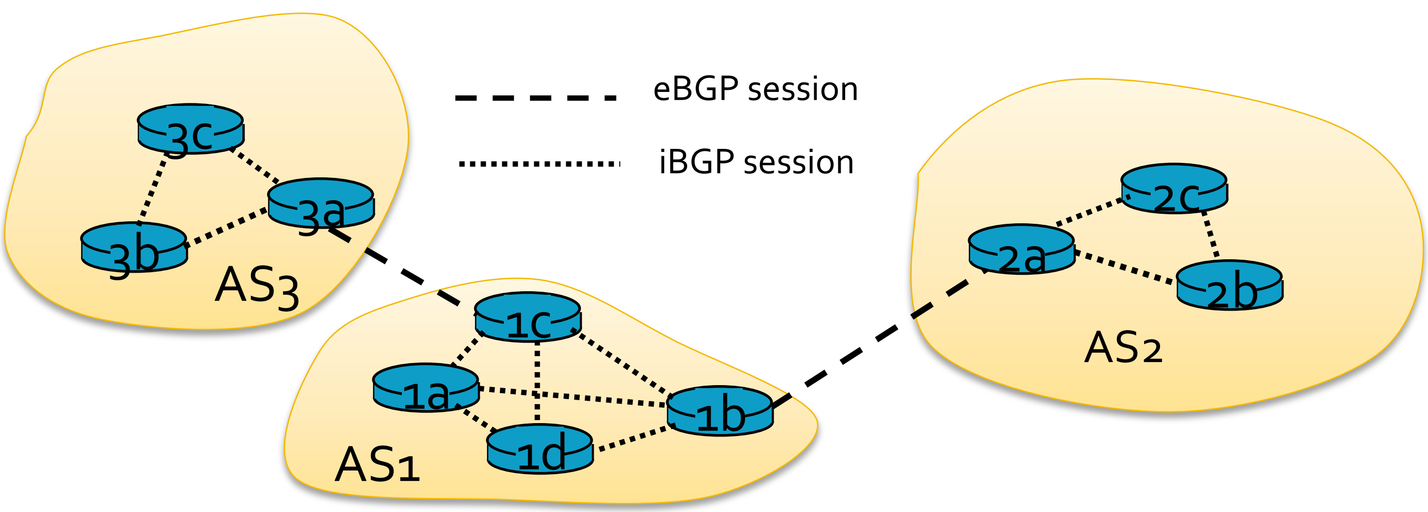
BGP Basics

- 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 to AS1:
  - AS2 *promises* it will forward datagrams towards that prefix.



Distributing Reachability Info

- Using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
  - 1c can then use iBGP do distribute new prefix info to all routers in AS1
  - 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session
- When router learns of new prefix, it creates entry for prefix in its forwarding table.



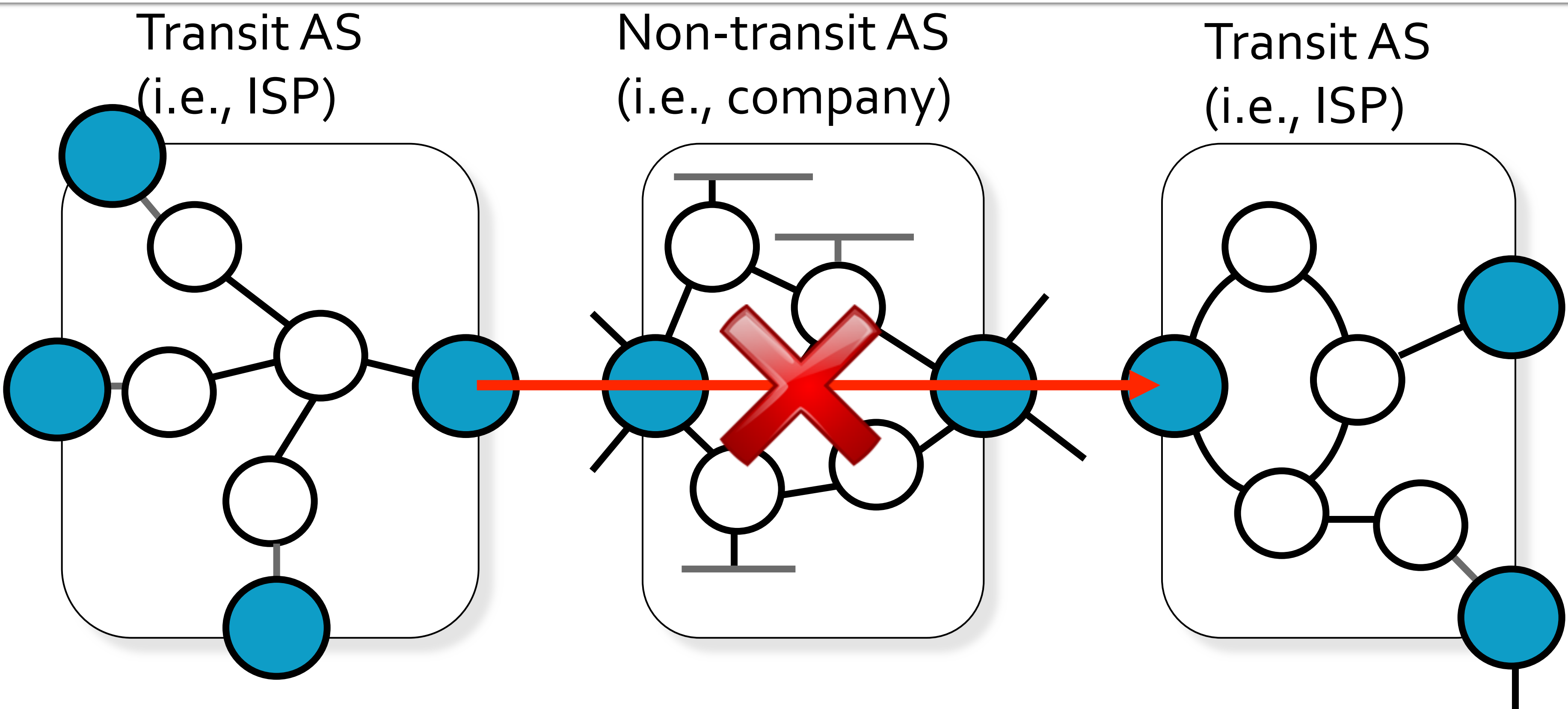
Border Gateway Protocol (BGP-4)

- BGP uses "path vectors" (AS_PATH)
  - Advertises complete "paths" – a list of autonomous systems
  - **"The network 171.64/16 can be reached via the path {AS1, AS5, AS13}"**
  - Makes no use of distance vectors or link states
- Path selection
  - Supports CIDR (classless inter-domain routing)
    - Most specific entry wins
  - Paths with loops are detected locally and ignored
  - Local policies pick the preferred path among options
  - When a link/router fails, the path is "withdrawn"

BGP route selection

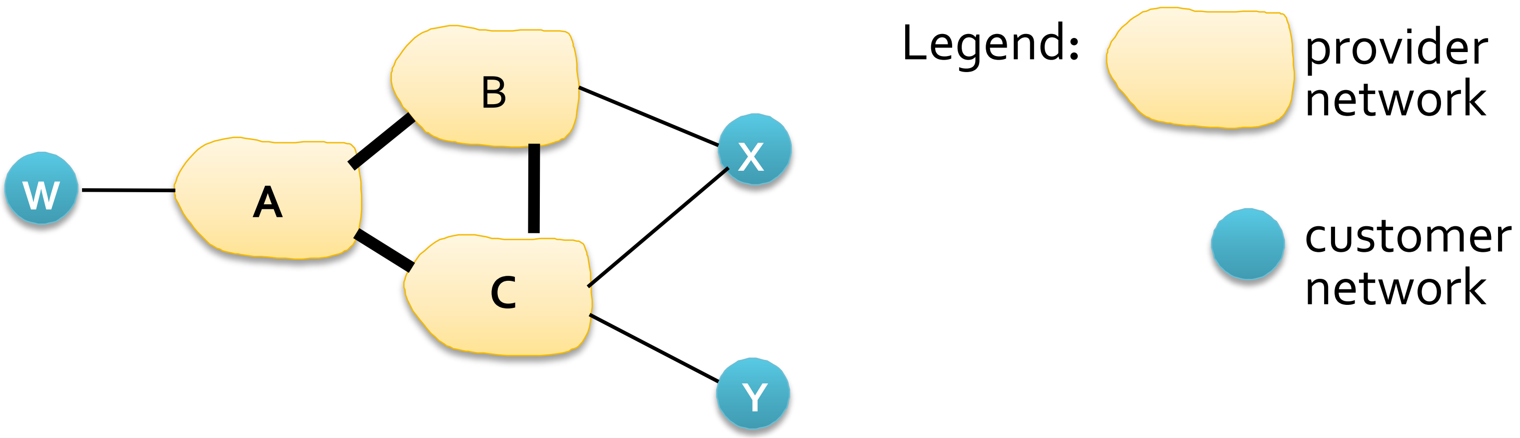
- Router may learn about more than 1 route to some prefix
  - Must select best route
- Elimination rules:
  1. Local preference value attribute: policy decision
  2. Shortest AS-PATH
     1. <u>Not</u> counting routers, but counting AS!
  3. Closest NEXT-HOP router: hot potato routing
  4. Additional criteria (varies by administrator)

BGP Routing Policy (1)



- Traffic shouldn't flow *through* the non-transit AS
  - Paying ISPs for connectivity, not to route traffic for them!
  - Don't advertise any BGP routes between transit AS's
  - Pacific is dual-homed to TCTC (Time Warner) and SWIS (AT&T)

BGP Routing Policy (2)



- A advertises path AW to B
- B advertises path BAW to X
- Should B advertise path BAW to C?
  - No way! B gets no $$$ for routing CBAW since neither W nor C are customers of B
  - 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**
  - Inter-AS: admin wants control over how its traffic is routed and who routes through its net
  - Intra-AS: single admin, so no policy decisions needed
- **Scale**
  - Hierarchical routing saves table size and reduces update traffic
- **Performance**
  - Intra-AS: can focus on performance
  - Inter-AS: policy may dominate over performance

Traceroute with AS Numbers

```
dhcp-10-6-162-134:~ shafer$ traceroute -a -q 1 www.msu.ru
traceroute to www.msu.ru (93.180.0.18), 64 hops max, 52 byte packets
 1  [AS65534] 10.6.163.254 (10.6.163.254)  1.677 ms
 2  [AS1] 10.0.0.141 (10.0.0.141)  1.116 ms
 3  [AS1] 10.0.0.90 (10.0.0.90)  1.053 ms
 4  [AS0] 138.9.253.252 (138.9.253.252)  5.200 ms
 5  [AS0] 74.202.6.5 (74.202.6.5)  8.137 ms
 6  [AS4323] pao1-pr1-xe-1-2-0-0.us.twtelecom.net (66.192.242.70)  13.241 ms
 7  [AS3356] te-9-4.car1.sanjose2.level3.net (4.59.0.229)  92.772 ms
 8  [AS3356] vlan70.csw2.sanjose1.level3.net (4.69.152.126)  8.440 ms
 9  [AS3356] ae-71-71.ebr1.sanjose1.level3.net (4.69.153.5)  11.130 ms
10  [AS3356] ae-2-2.ebr2.newyork1.level3.net (4.69.135.186)  80.992 ms
11  [AS3356] ae-82-82.csw3.newyork1.level3.net (4.69.148.42)  77.316 ms
12  [AS3356] ae-61-61.ebr1.newyork1.level3.net (4.69.134.65)  74.584 ms
13  [AS3356] ae-41-41.ebr2.london1.level3.net (4.69.137.65)  147.127 ms
14  [AS3356] ae-48-48.ebr2.amsterdam1.level3.net (4.69.143.81)  151.779 ms
15  [AS3356] ae-1-100.ebr1.amsterdam1.level3.net (4.69.141.169)  152.848 ms
16  [AS3356] ae-48-48.ebr2.dusseldorf1.level3.net (4.69.143.210)  156.349 ms
17  [AS3356] 4.69.200.174 (4.69.200.174)  168.386 ms
18  [AS3356] ae-1-100.ebr1.berlin1.level3.net (4.69.148.205)  167.652 ms
19  [AS3356] ae-4-9.bar1.stockholm1.level3.net (4.69.200.253)  192.668 ms
20  [AS3356] 213.242.110.198 (213.242.110.198)  176.501 ms
21  [AS3267] b57-1-gw.spb.runnet.ru (194.85.40.129)  198.827 ms
22  [AS3267] m9-1-gw.msk.runnet.ru (194.85.40.133)  204.276 ms
23  [AS3267] msu.msk.runnet.ru (194.190.254.118)  202.454 ms
24  [AS2848] 93.180.0.158 (93.180.0.158)  201.358 ms
25  [AS2848] 93.180.0.170 (93.180.0.170)  200.257 ms
26  [AS2848] www.msu.ru (93.180.0.18)  204.045 ms !Z
```

AS Numbers in Traceroute

AS	Name
0	Reserved (local use) – Pacific is here…
4323	Time Warner Telecom
3356	Level 3 Communications
3267	Runnet - State Institute of Information Technologies & Telecommunications (SIIT&T "Informika")
2848	Moscow State University

Problems

- BGP designed for policy, not performance
- Susceptible to misconfiguration
  - Intentionally / accidentally announce routes to networks you cannot reach
- Incompatible policies might render networks unreachable

BGP, Censorship, and You (February 2008)

1. Pakistan government orders Pakistan Telecom (AS 17557) to block access to YouTube
2. Pakistan Telecom advertises a route for 208.65.153/24 (YouTube) to its customers leading to a black hole
3. That route is accidentally advertised to its provider (PCCW)
   - This is more specific than YouTube's (AS 36561) real advertisements (208.65.152/22)
   - Multiple routes → More specific route preferred
4. PCCW failed to verify that Pakistan Telecom actually owned YouTube's netblock (very common)
   - BGP uses transitive trust – PCCW trusted P.T., and upstream providers trusted PCCW
5. Within ~3 minutes, large fraction of the Internet had bad route
   - YouTube traffic was routed to AS 17557 instead of AS 36561
   - AS 17557 can then just drop the received traffic

We Want Our Videos Back!

6. ~1 hour later, YouTube advertises that its addresses have been hijacked to its providers
   - YouTube verifiably owns that address space and its AS number
7. Autonomous systems stop using the bad route
   - YouTube also advertises its own /25 routes
8. ~1 hour later, Pakistan Telecom's provider (Hong Kong-based PCCW) withdraws bogus routes to AS 17557