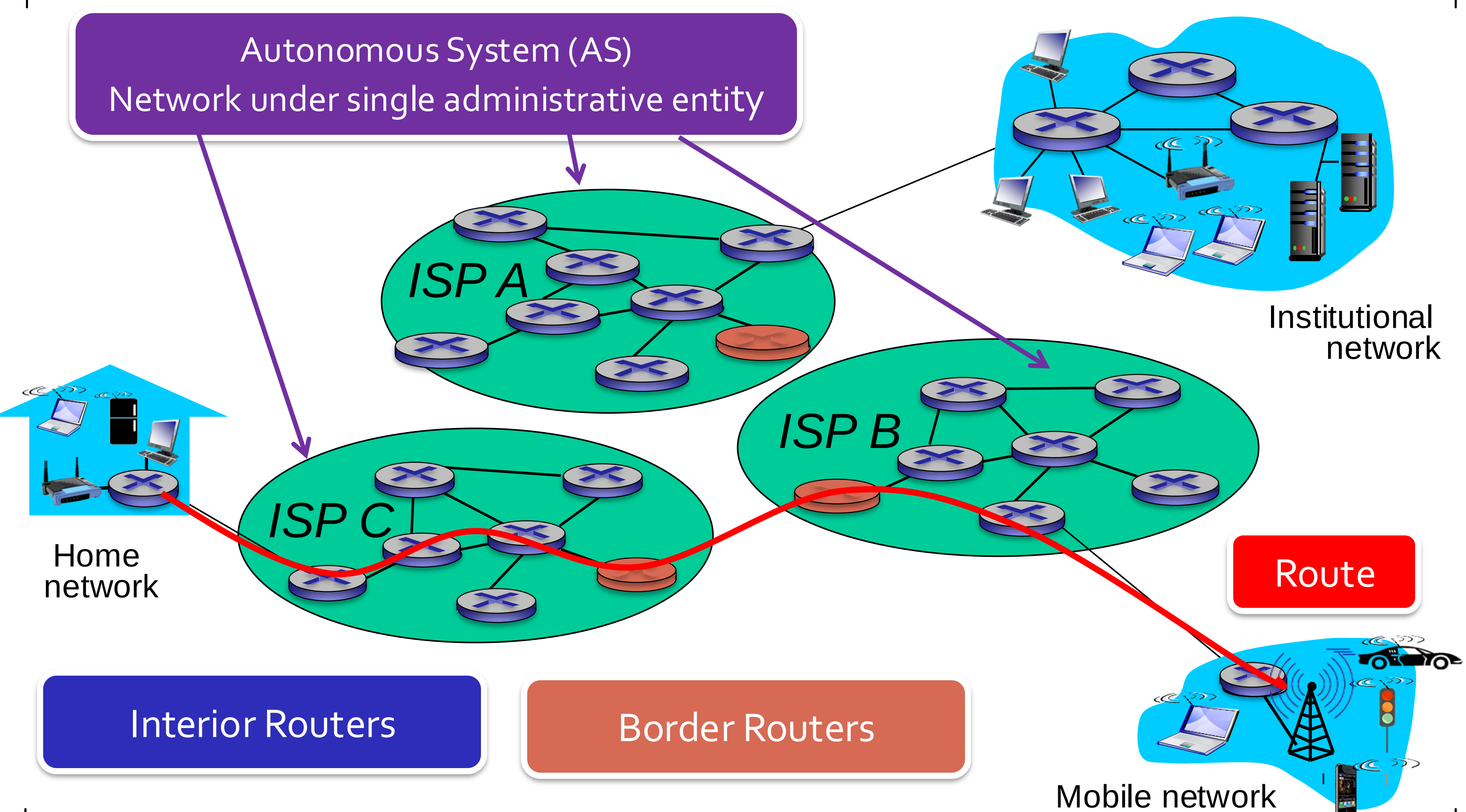
Intra - and Extra - AS (domain) routing

CE 352, Computer Networks

Salem Al-Agtash

Lecture 17

Recall (Lec12 - important context)



Network topology

Nodes

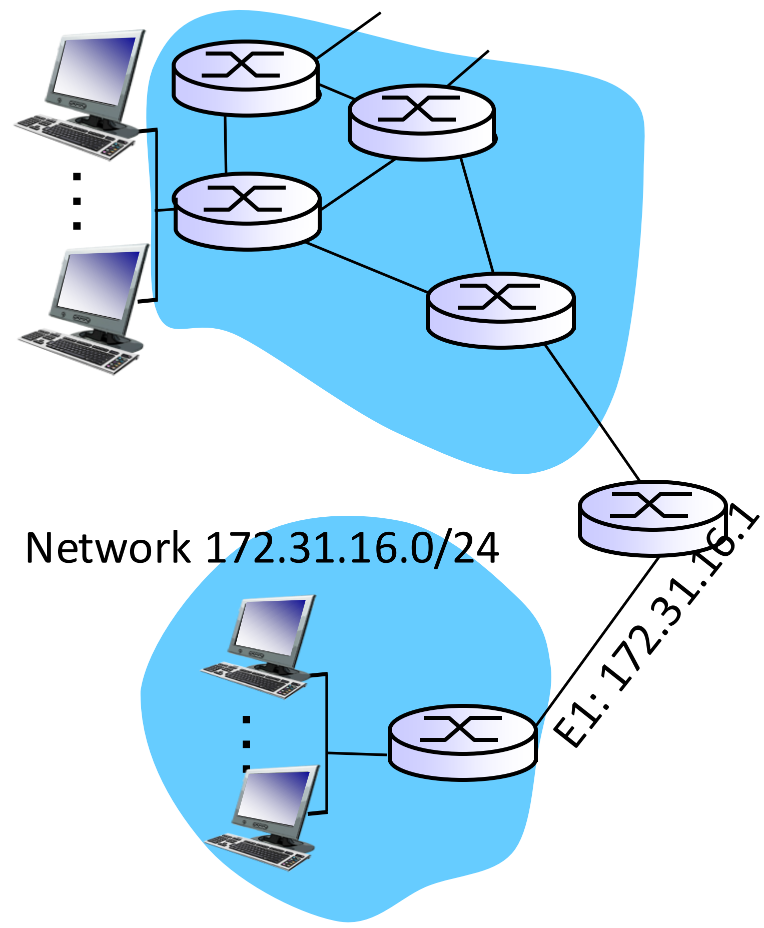
- Representing routers
- Mainly located in datacenters
- Destinations (IP prefixes)

Edges

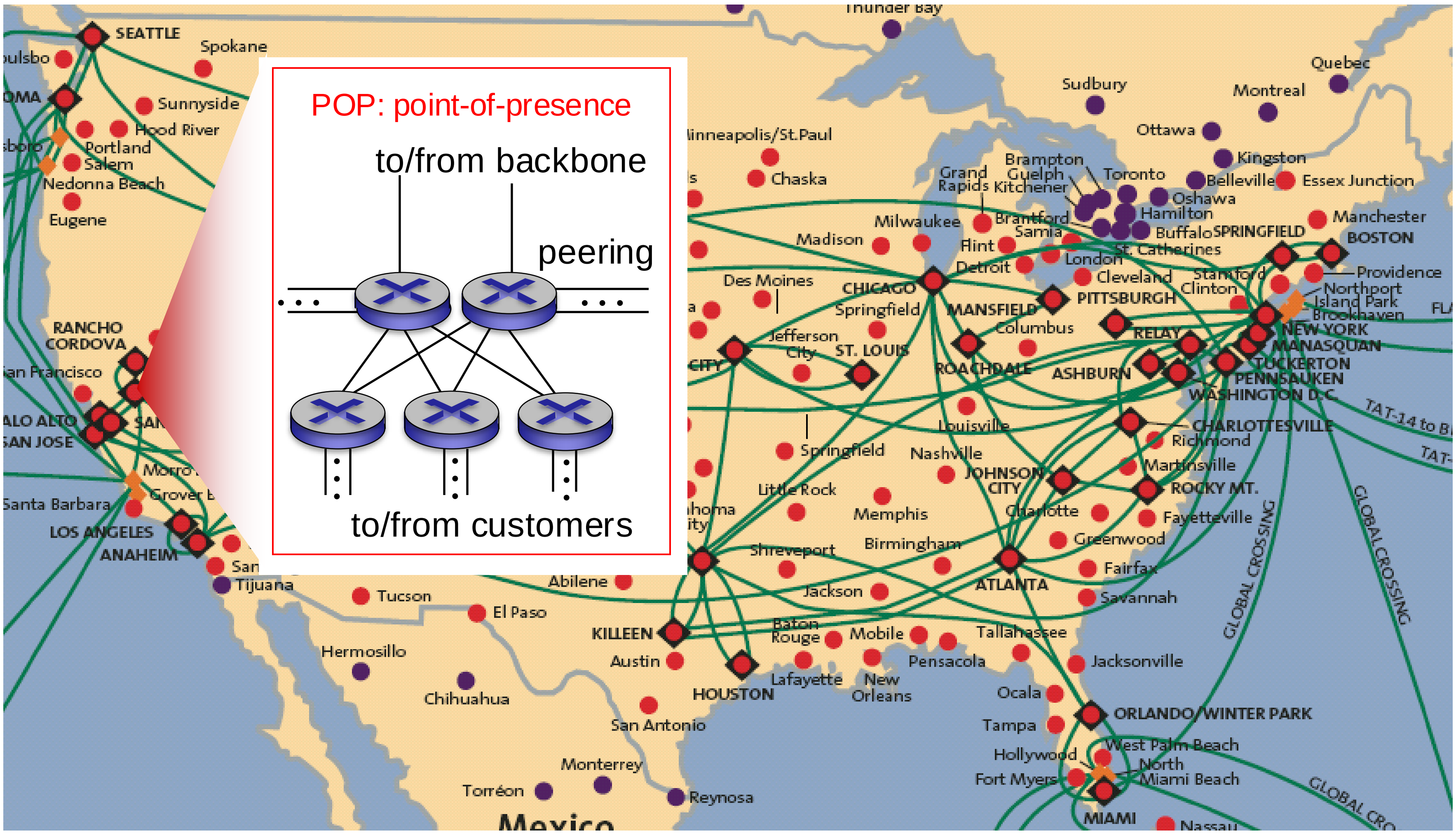
- Representing interconnecting links:
  - Metrics (distance, bandwidth, loss, congestion, etc.)
  - e.g. (depart time - arrival time) + transmission time + link propagation delay

Point-of-Presence (PoP)

- Cluster of routers in a datacenter (e.g. Equinix)
- Inter-PoP: High bandwidth links
- Intra-PoP: Cables between racks



Recall (Lec3. ISP: e.g. Sprint)



Routing in the Internet

So far "simple network assumption" with a set of homogenous interconnected routers running same algorithms to find best route

- Not realistic
  - Scale: Internet consists of millions of routers, so impossible for routers to store routing information and make cost updates
  - Administrative autonomy: ISPs desire to operate their own networks

The internet is managed and operated through Autonomous System (AS):

- 16-bit AS number → 65,536 (64,510 available for public use), recently assigned 32-bits, managed by Internet Assigned Numbers Authority (IANA), department in ICANN
  - e.g. MIT: 3, Harvard: 11, AT&T: 7018, 6341, 5074, ...
- AS Types: Stub AS (small corporation), Multihomed AS (large corporation – no transit), and Transit AS (ISP)
- Routers within the same AS all run the same routing algorithm and have information about each other

Regional Internet Registries

ICANN: Internet Corporation for Assigned Names and Numbers

http://www.icann.org/ operates via 5 Regional Internet Registries – ARIN (North America), APNIC (Asia Pacific), RIPE NCC (Europe and Middle East), LACNIC (Latin America), and AFRINIC (Africa) → (DNS, IP, Port, AS No.)

Intra and Inter AS Routing

Intra - AS routing (a.k.a. Interior Gateway Protocols (IGP)) – Within AS:

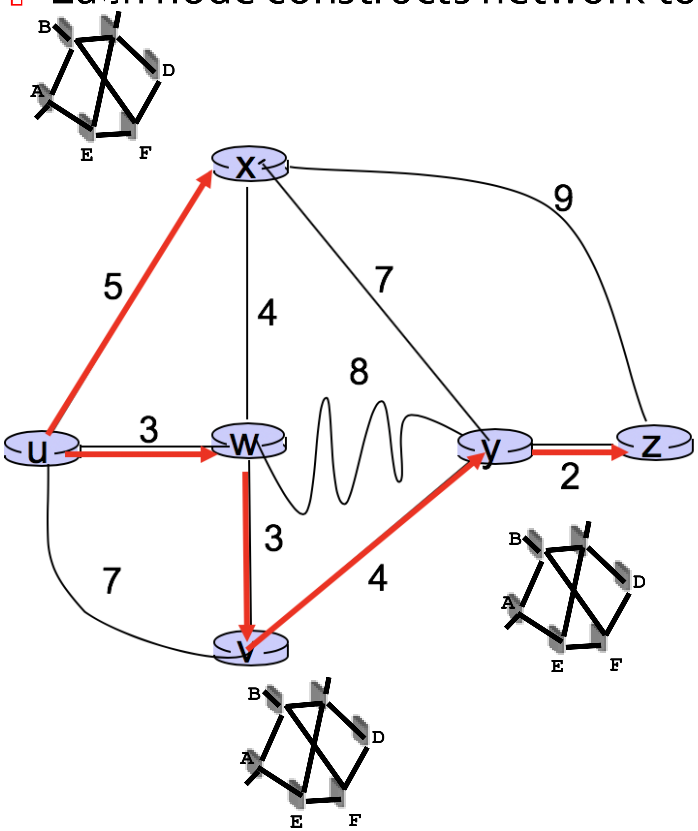
- Metric – based (link costs)
  - OSPF (Open Shortest Path First) uses LS
    - Widely used and each router constructs a complete topological map of AS
    - Individual link costs are configured by the network administrator
  - RIP (Routing Information Protocol) uses DV
  - EIGRP (Enhanced Interior Gateway Routing Protocol – Cisco) uses DV

Inter - AS routing – Between AS's:

- Policy-based (not link cost, but control over who and where to send traffic: transit, peering)
  - BGP (Border Gateway Protocol): *de facto* standards
  - Path Vector protocol: extension of DV
  - Provides path to destination as sequences of AS's
  - Scale: prefixes of 200,000, growing

Recap (LS)

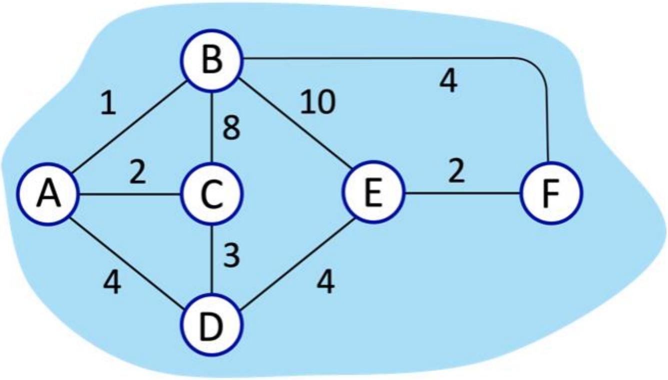
- Each node j, periodically creates LSP: ID, seq.#, TTL, List of neighbors and costs
- When node j receives LSP from node k, check seq#, if new, saves in its DB and forwards a copy to all links except node k, else discards LSP [Flooding]
- Each node constructs network topology, and uses Dijkstra to find shortest path



Step	N'	D(v) p(v)	D(w) p(w)	D(x) p(x)	D(y) p(y)	D(z) p(z)
0	u	7,u	3,u	5,u	∞	∞
1	uw	6,w		5,u	11,w	∞
2	uwx	6,w			11,w	14,x
3	uwxv				10,v	14,x
4	uwxvy					12,y
5	uwxvyz					

Example

- Consider the network shown below, and Dijkstra's link-state algorithm. Here, we are interested in computing the least cost path from node ***E*** to all other nodes. Using Dijkstra algorithm, complete the rows in steps 0, 1, and 2 in the table below showing the link state algorithm's execution and find the values at (a) .... (e)

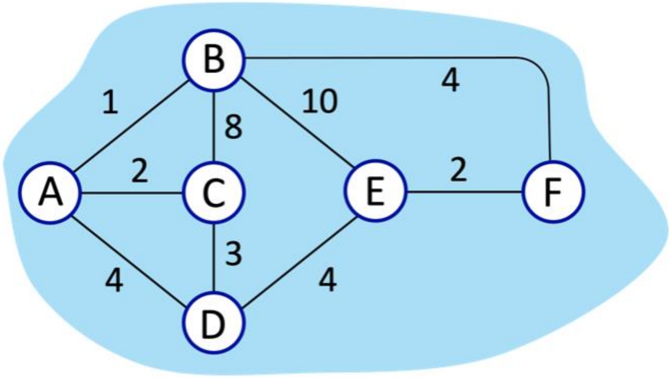


	A	B	C	D	F
N'	D(A),p(A)	D(B),p(B)	D(C),p(C)	D(D),p(D)	D(F),p(F)
E	*	*	∞	*	*
*	*	*	∞	*	*
(a)	(b)	(c)	(d)	(e)	2,E

Example

- Consider the network shown below, and Dijkstra's link-state algorithm. Here, we are interested in computing the least cost path from node ***E*** to all other nodes. Using Dijkstra algorithm, complete the rows in steps 0, 1, and 2 in the table below showing the link state algorithm's execution and find the values at (a) .... (e)

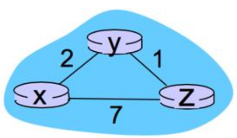
		A	B	C	D	F
Step	N'	D(A),p(A)	D(B),p(B)	D(C),p(C)	D(D),p(D)	D(F),p(F)
0	E	∞	10, E	∞	4, E	2, E
1	EF	∞	6, F	∞	4, E	
2	EFD	8, D	6, F	7, D		



Recap (DV – Good and bad news)

- Each node j, periodically creates distance table: destination and outgoing link to use and cost, by using Bellman-Ford dynamic programming
- Each node j notifies neighbors only when its least cost path to any destination changes
- Count to infinity problem (implicit path): poisoned reverse, split horizon, etc.

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	3	1	0



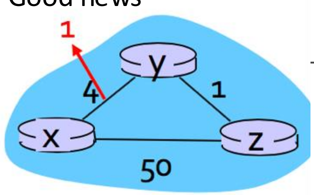
cost to x	y	z
0	1	2
1	0	1
2	1	0

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	3	1	0

cost to x	y	z
0	1	2
1	0	1
2	1	0

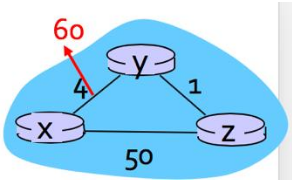
from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	3	1	0

Good news



cost to x	y	z
0	1	2
1	0	1
2	1	0

Bad news



from \ cost to	x	y	z
x	0	4	5
y	6	0	1
z	5	1	0

from \ cost to	x	y	z
x	0	4	5
y	6	0	1
z	5	1	0

from \ cost to	x	y	z
x	0	4	5
y	8	0	1
z	5	1	0

\ cost to	x	y	z
x	0	4	5
y	8	0	1
z	5	1	0

from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	5	1	0

from \ cost to	x	y	z
x	0	2	2
y	6	0	1
z	7	1	0

\ cost to	x	y	z
x	0	2	2
y	6	0	1
z	7	1	0

\ cost to	x	y	z
x	0	2	2
y	6	0	1
z	9	1	0

OSPF (Open Shortest Path First) routing

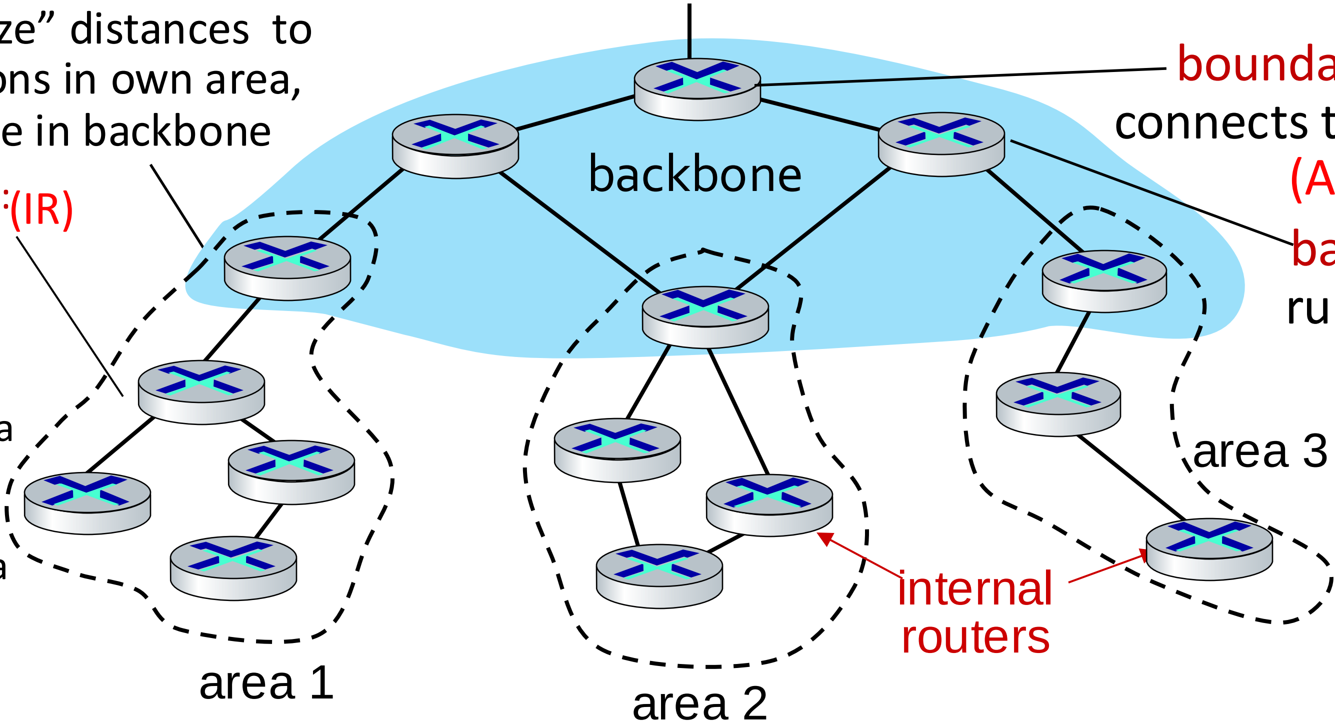
- LS and DV routing – ideal flat network, not true in practice, millions of destinations – cannot store in routing tables
- "open": publicly available
- classic link-state
  - each router floods OSPF link-state advertisements (directly over IP rather than using TCP/UDP) to all other routers in entire AS
  - multiple link costs metrics possible: bandwidth, delay
  - each router has full topology, uses Dijkstra's algorithm to compute forwarding table
- *security:* all OSPF messages authenticated (to prevent malicious intrusion)

Hierarchical routing - OSPF

- two-level hierarchy: local area (~300 routers), backbone.
  - link-state advertisements (LSA) flooded only in area, or backbone
  - each node has detailed area topology; only knows direction to reach other destinations

area border routers (ABR): "summarize" distances to destinations in own area, advertise in backbone

local routers:(IR)
- flood LS in area only
- compute routing within area
- forward packets to outside via area border router

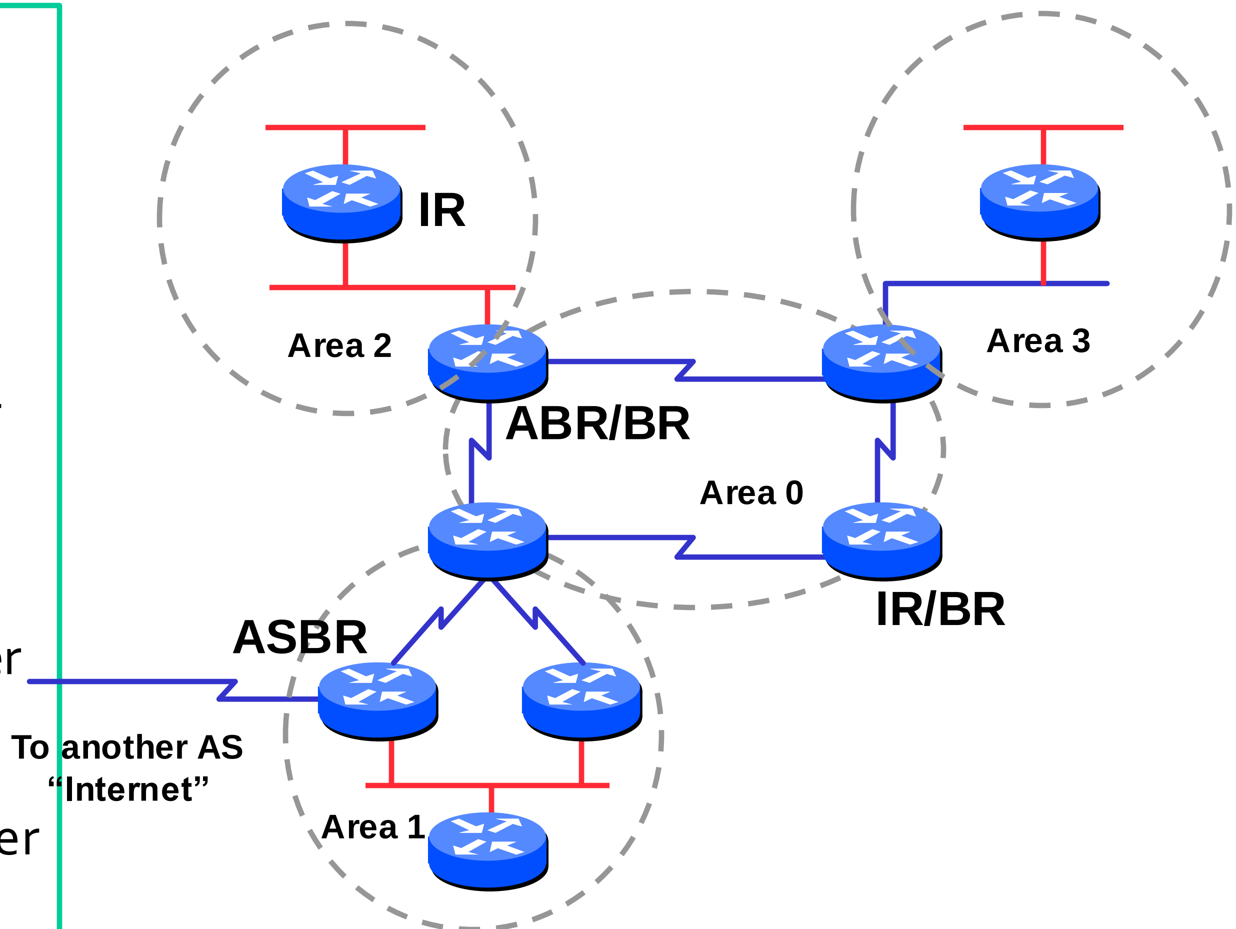


boundary router: connects to other Ases (ASBR)

backbone router: runs OSPF limited to backbone (BR)

Router Classification

- IR: Internal Router
  - Routes within area
- ABR: Area Border Router
  - Routes are announced from area to another
- BR: Backbone Router
  - Runs OSPF limited to backbone
- ASBR: Autonomous System Border Router
  - Connects to other AS's
- DR: Designated Router
  - Elected router in same area, multicasting info



Source: OSPF Presentation @ NSRC
https://learn.nsrc.org/bgp/ospf

Routing Protocol Packets

- Sent as IP packet with a TOS field = 0
- Five types of OSPF routing protocol packets
  - Link-state advertisement (LSA) – primary mean of communication between OSPF routers carrying information about the network topology (11 types)
    - Each router has information to communicate about the networks, building maps then running Dijkstra algorithm
  - Link-state DB Description (LSDB) – link-state information exchanged among the network. Routers in the same area have identical LSDB.
  - Link-state request (LSR) – neighbor router sends a request to claim a missing LSDB from a neighbor (adjacent router)
  - Link-state update – response to LSR on a specific piece of LSDB with neighbor
  - Link-state Acknowledgment – confirming receipt of LSU from neighbor

OSPF areas

Area is a group of contiguous hosts and networks

- Reduces routing traffic

Per area topology database

- Invisible outside the area

Backbone area contiguous and connects all areas

Types of areas:

- Regular (ISPs): Summary networks from other areas are injected, as external networks
- Stub: Summary networks from other areas are injected as default type 3 route – not connecting to other areas or to other AS (on the edge)
- Totally Stubby: Only a default route injected to closest area border router
- Not-So-Stubby: importing routes in a limited fashion

Types of LS advertisement (LSA)

- Type 1 : Router LSA to advertise directly connected networks
  - type (ABR, ASBR), links, costs in area, flooded in area
- Type 2 : Network LSA to represent each transit network
  - transit broadcast, all routers attached to network, flooded in area
- Type 3 : Summary LSA from one area to another to advertise a network in the source area (listing of networks)
  - inter-area routes advertised into backbone, ABR
- Type 4 : Summary ASBR LSA created by ABR to tell members of an area how to reach ASBR
  - destination outside area in AS
- Type 5&7: AS External LSA created by ASBR to advertise networks in a different AS
  - routes to destination external to AS, costs, LSA for one specific OSPF area type
- Type 6: Group membership LSA
- Type 9, 10 & 11: Link-Local, Area

OSPF details

"open": publicly available

two-level hierarchy: local area, backbone.

- link-state advertisements only in area
- each node has detailed area topology; only knows direction (shortest path) to nets in other areas.

uses link-state algorithm

- link state packet dissemination
- topology map at each node in an area
- route computation using Dijkstra's algorithm

router floods OSPF link-state advertisements to all other routers in *entire* AS

- carried in OSPF messages directly over IP (rather than TCP or UDP)
- link state: for each attached link

OSPF "advanced" features

- *security:* all OSPF messages authenticated (to prevent malicious intrusion), possible fictitious paths (hard to fix)
- multiple same-cost paths allowed
- integrated uni- and multi-cast support
- hierarchical OSPF in large domains

RIP (Routing Information Protocol)

- DV based RIP, earliest routing protocol
- UDP port 520
- As in DV, node maintains copies of neighbours' routing tables and uses iteratively to generate its own table (table is refreshed every 30 sec)
- When an entry is updated, router sends copy of entry to neighbours
- Link costs in RIP: 1 – 15, 16 represent infinity
- RIP is limited to fairly small networks
- Router timer: limit 180 seconds, if not updated, set to infinity
- When router or link fails, can take minutes to stabilize

- EIGPR widely used

Inter-AS routing: BGP

BGP (Border Gateway Protocol): the *de facto* inter-domain routing protocol

- "glue that holds the Internet together"
- TCP port 179

BGP provides each AS a means to:

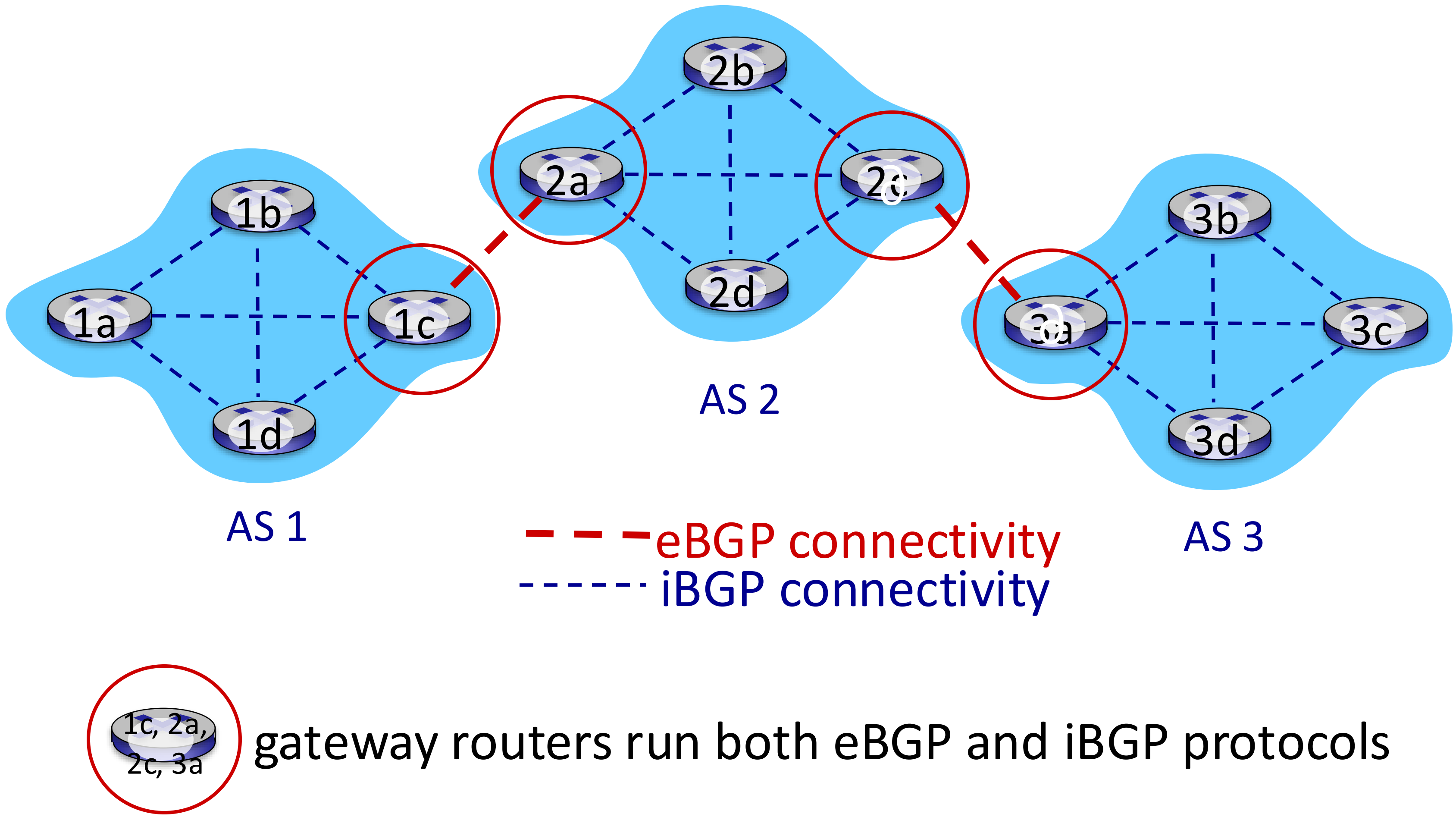
- eBGP: obtain subnet reachability information from neighboring ASes
- iBGP: propagate reachability information to all AS-internal routers.
- determine "good" routes to other networks based on reachability information and *policy*

Policy criteria

- Financial gains (Transit and Peering)
- Performance (smallest AS path length)
- Minimize use of own network bandwidth ("hot potato")

Path vector routing: extension of DV, gives entire path to destination (autonomous systems to go through)

eBGP, iBGP connections

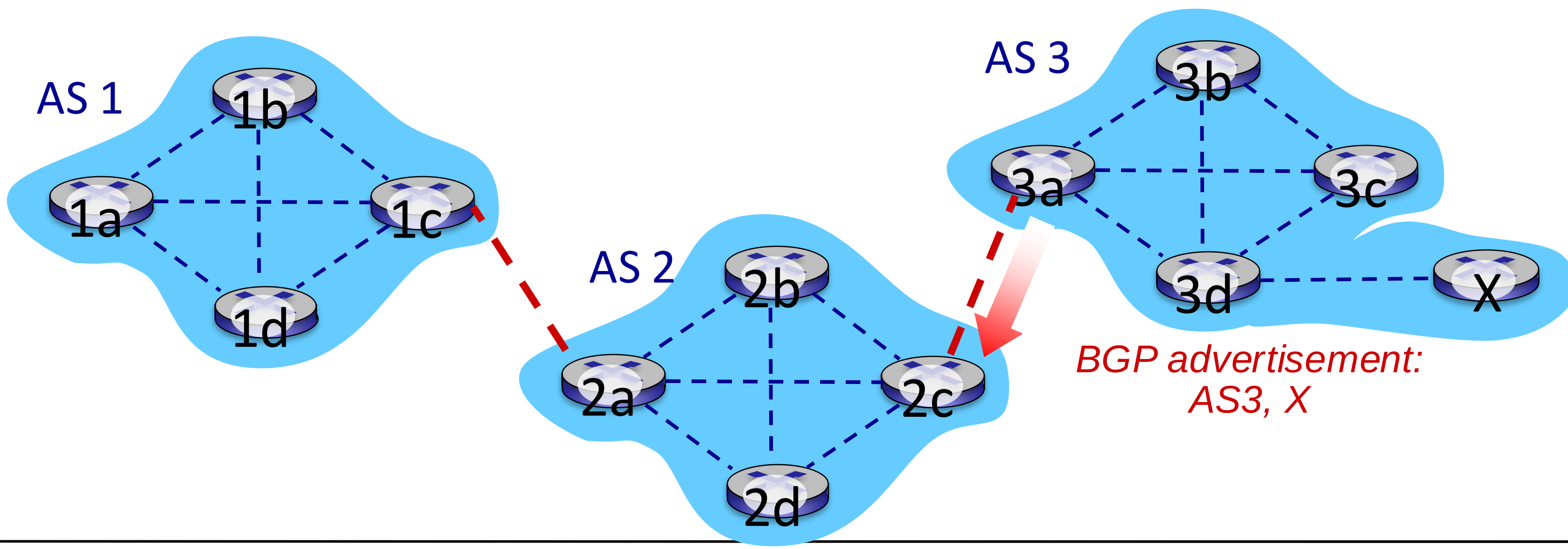


BGP basics

- BGP session: two BGP routers ("peers") exchange BGP messages over semi-permanent TCP connection:
  - advertising *paths* to different destination network prefixes (BGP is a "path vector" protocol)

when AS3 gateway router 3a advertises path AS3,X to AS2 gateway router 2c:

- AS3 *promises* to AS2 it will forward datagrams towards X



Path attributes and BGP routes

Advertised prefix includes BGP attributes

- prefix + attributes = "route" →

e.g.
128.112.0.0/16 (network X)
AS path = 7018 543 (AS2, AS3)
Next hop = 12.127.0.121 (1c)

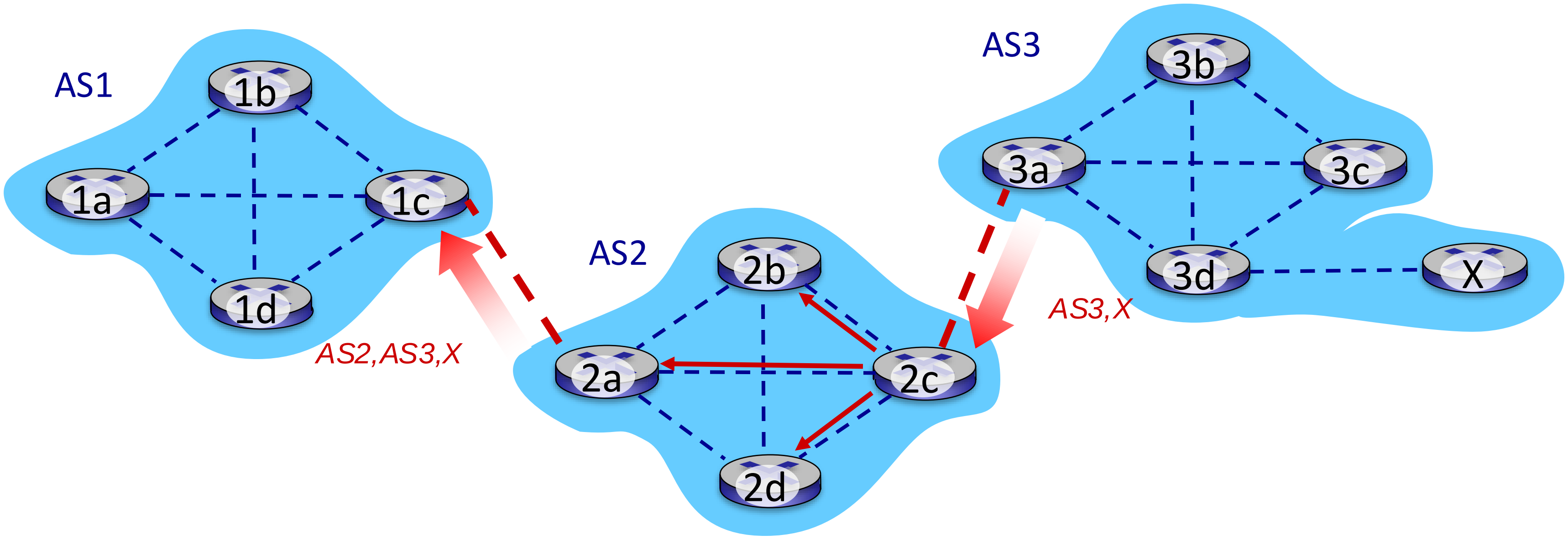
Two important attributes:

- AS-PATH: list of ASes through which prefix advertisement has passed
- NEXT-HOP: indicates specific internal-AS router to next-hop AS

Policy-based routing:

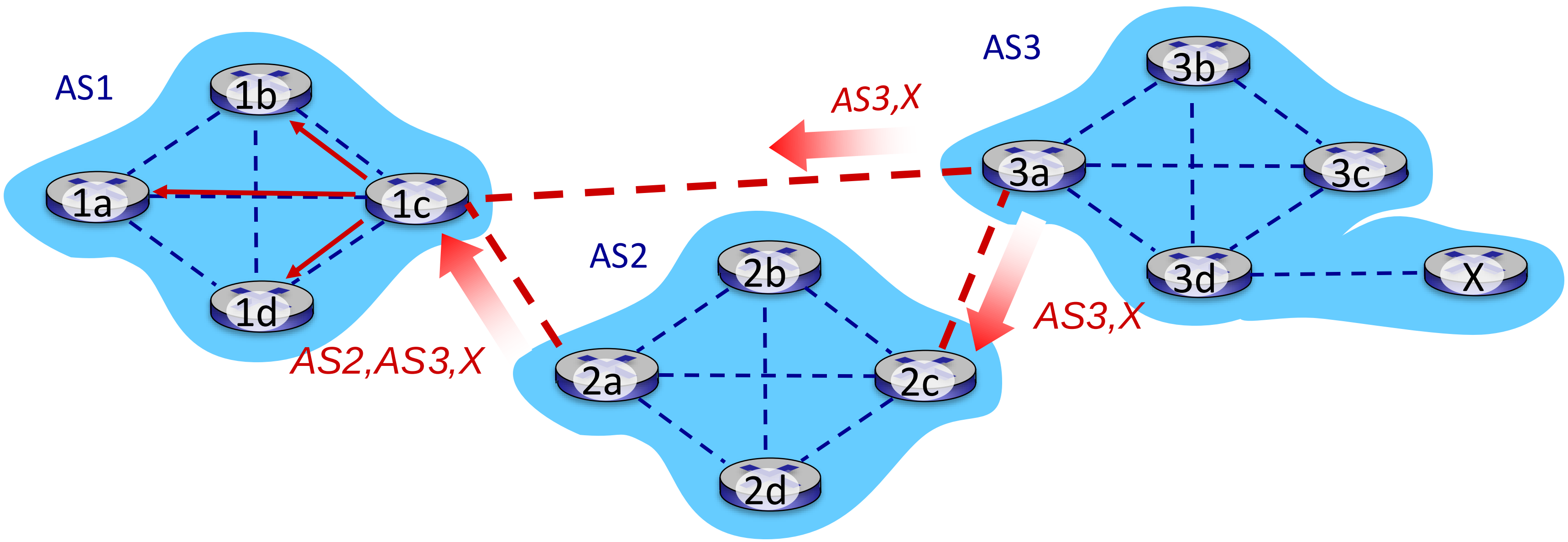
- gateway receiving route advertisement uses *import policy* to accept/decline path (e.g., never route through AS Y).
- AS policy also determines whether to *advertise* path to other other neighboring ASes

BGP path advertisement



- AS2 router 2c receives path advertisement AS3,X (via eBGP) from AS3 router 3a
- Based on AS2 policy, AS2 router 2c accepts path AS3,X, propagates (via iBGP) to all AS2 routers
- Based on AS2 policy, AS2 router 2a advertises (via eBGP) path AS2, AS3, X to AS1 router 1c

BGP path advertisement



gateway router may learn about multiple paths to destination:

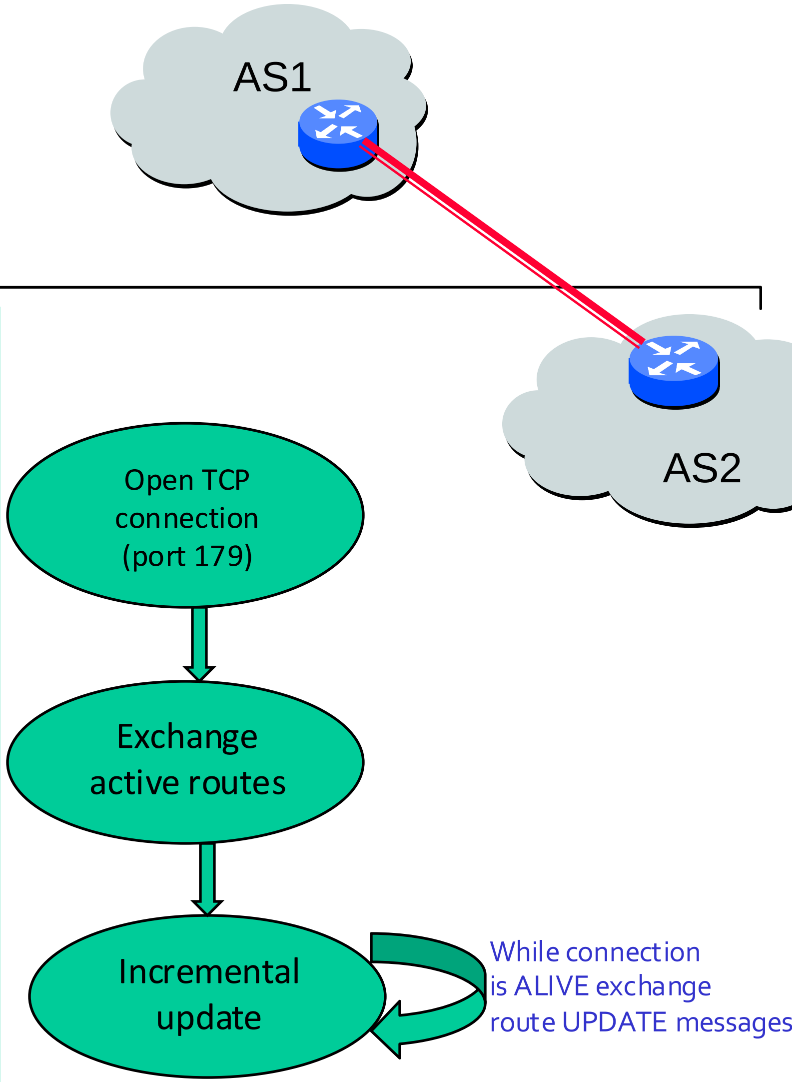
- AS1 gateway router 1c learns path *AS2,AS3,X* from 2a
- AS1 gateway router 1c learns path *AS3,X* from 3a
- Based on policy, AS1 gateway router 1c chooses path *AS3,X, and advertises path within AS1 via iBGP*

BGP operation

BGP messages exchanged between peers over TCP connection

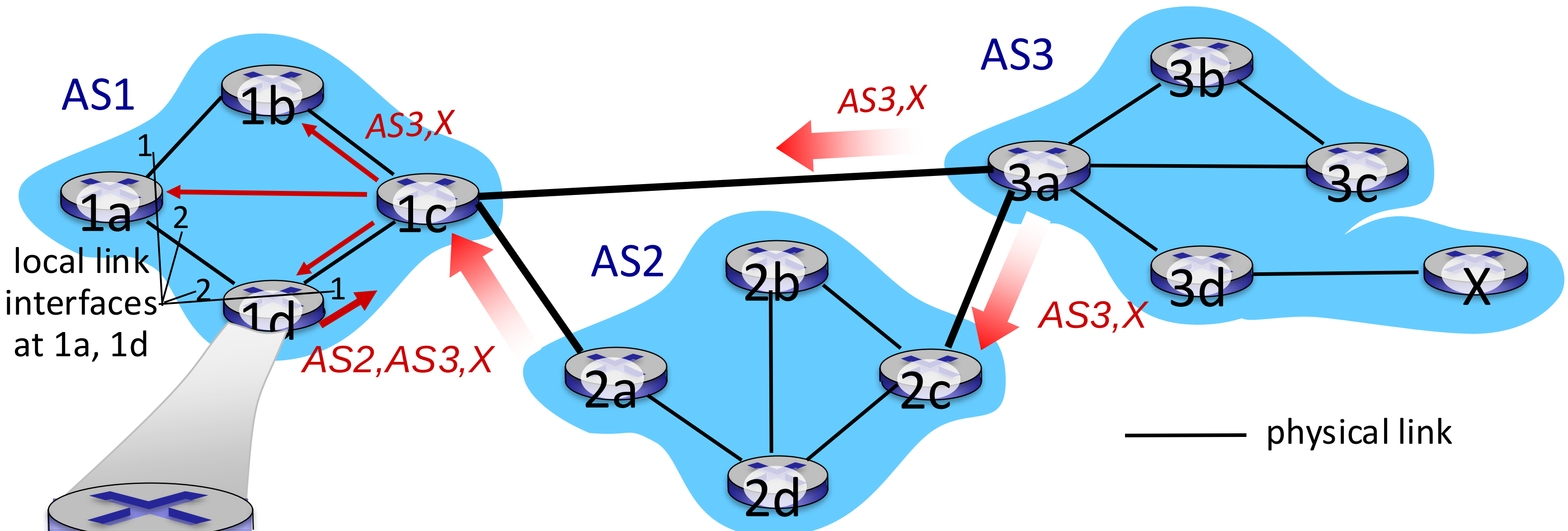
BGP messages:

- OPEN: opens TCP connection to remote BGP peer and authenticates sending BGP peer
- UPDATE: advertises new path (or withdraws old)
- KEEPALIVE: keeps connection alive in absence of UPDATES; also ACKs OPEN request
- NOTIFICATION: reports errors in previous msg; also used to close connection



BGP, OSPF, forwarding table entries

Q: how does router set forwarding table entry to distant prefix?



dest	interface
…	…
X	1
…	…

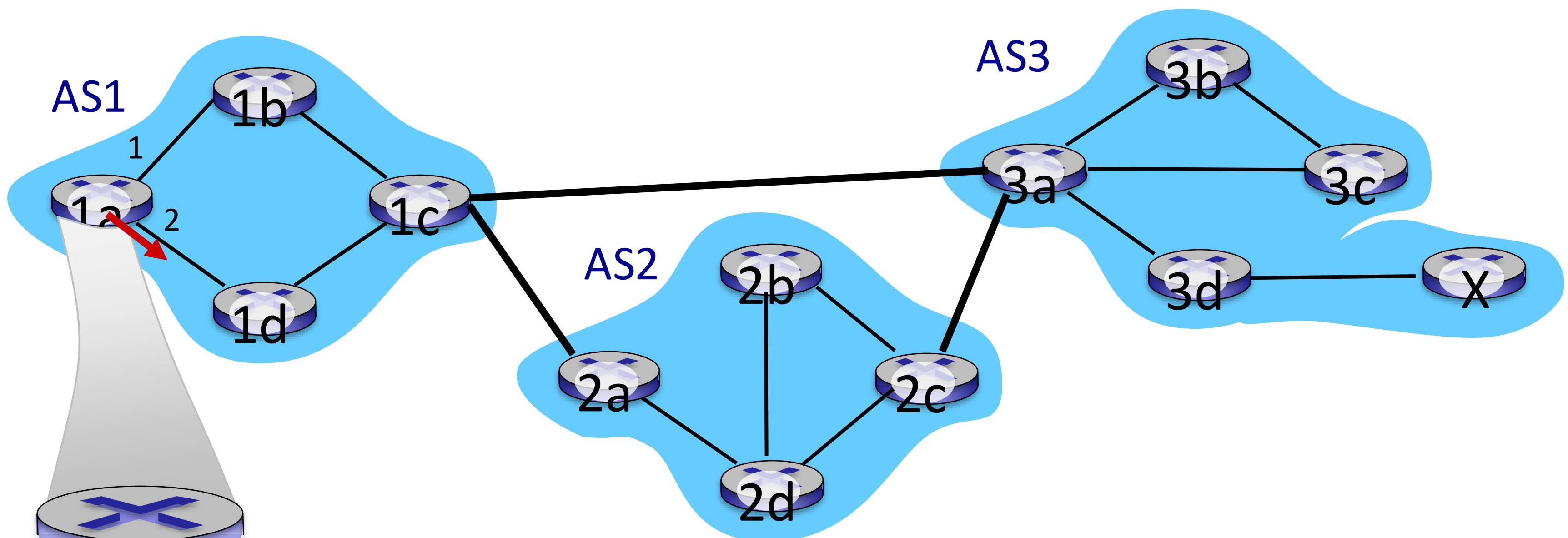
e.g.
128.112.0.0/16 (X)
AS path = 7018 543 (AS2, AS3)
Next hop = 12.127.0.121 (1c)

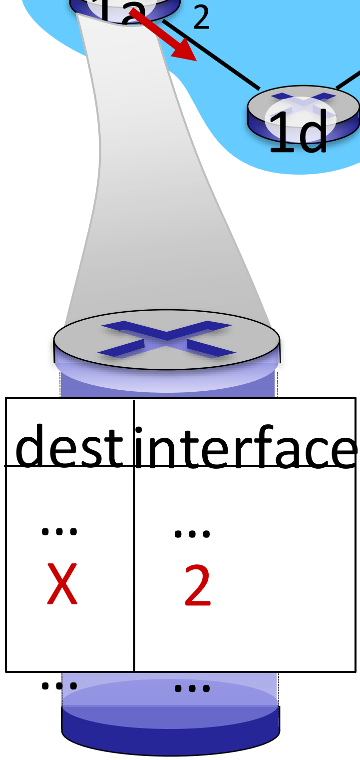
recall: 1a, 1b, 1c learn about dest X via iBGP from 1c: "path to X goes through 1c"

- 1d: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 1

BGP, OSPF, forwarding table entries

Q: how does router set forwarding table entry to distant prefix?





recall: 1a, 1b, 1c learn about dest X via iBGP from 1c: "path to X goes through 1c"

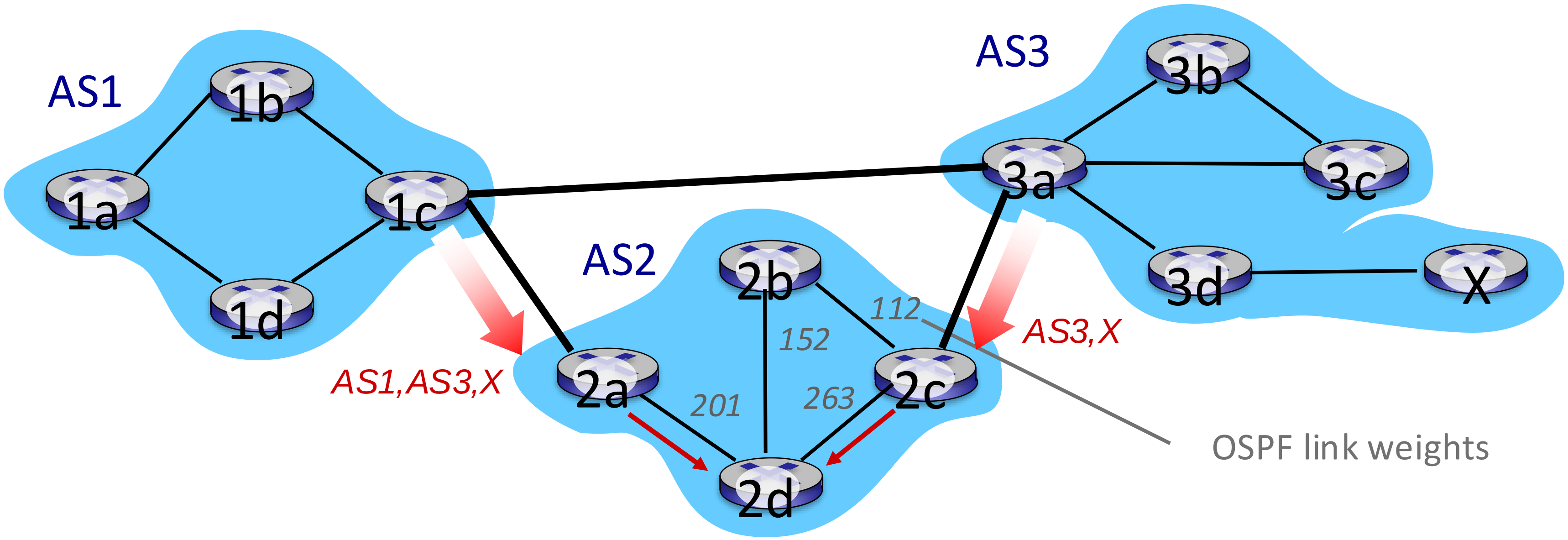
- 1d: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 1
- 1a: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 2

BGP route selection

router may learn about more than one route to destination AS, selects route based on:

1. local preference value attribute: policy decision
2. shortest AS-PATH
3. closest NEXT-HOP router: hot potato routing
4. additional criteria

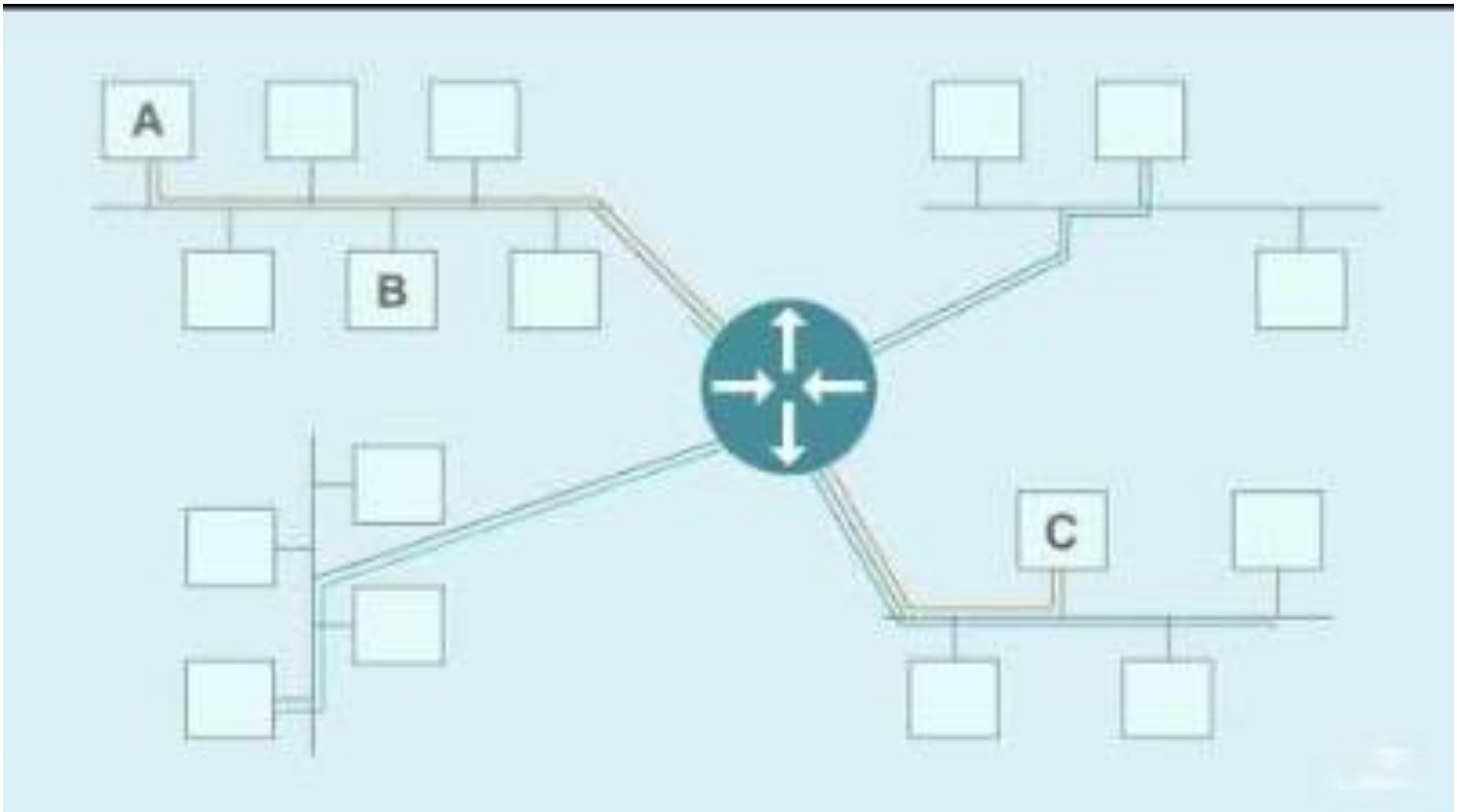
"Hot Potato" Routing



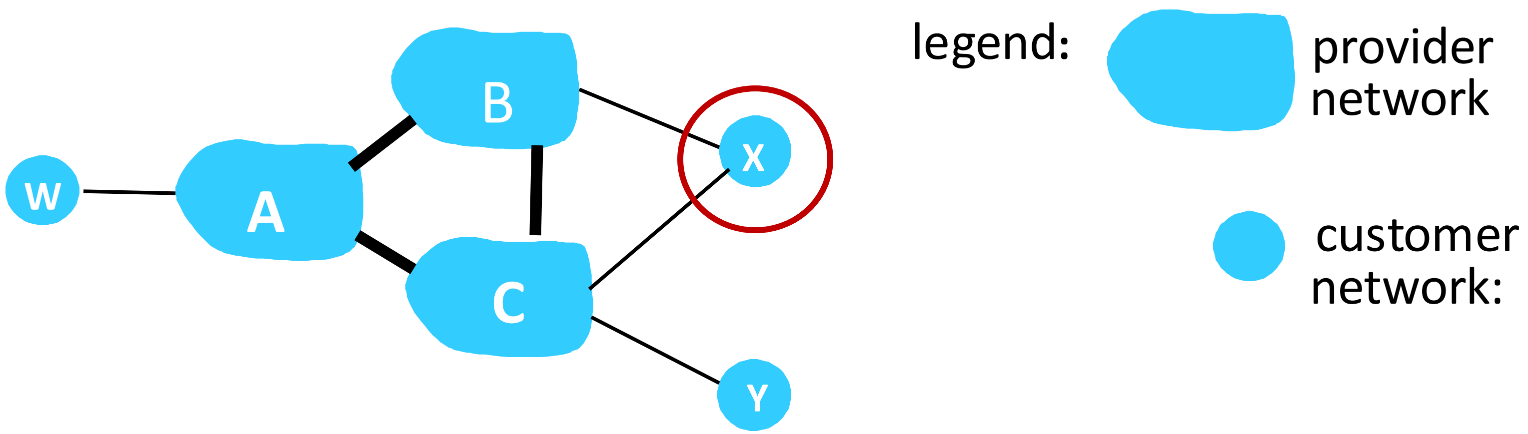
2d learns (via iBGP) it can route to X via 2a or 2c

hot potato routing: choose local gateway that has least intra-domain cost (e.g., 2d chooses 2a, even though more AS hops to *X*): don't worry about inter-domain cost! → routing instability

BGP https://youtu.be/A1KXPpqlNZ4



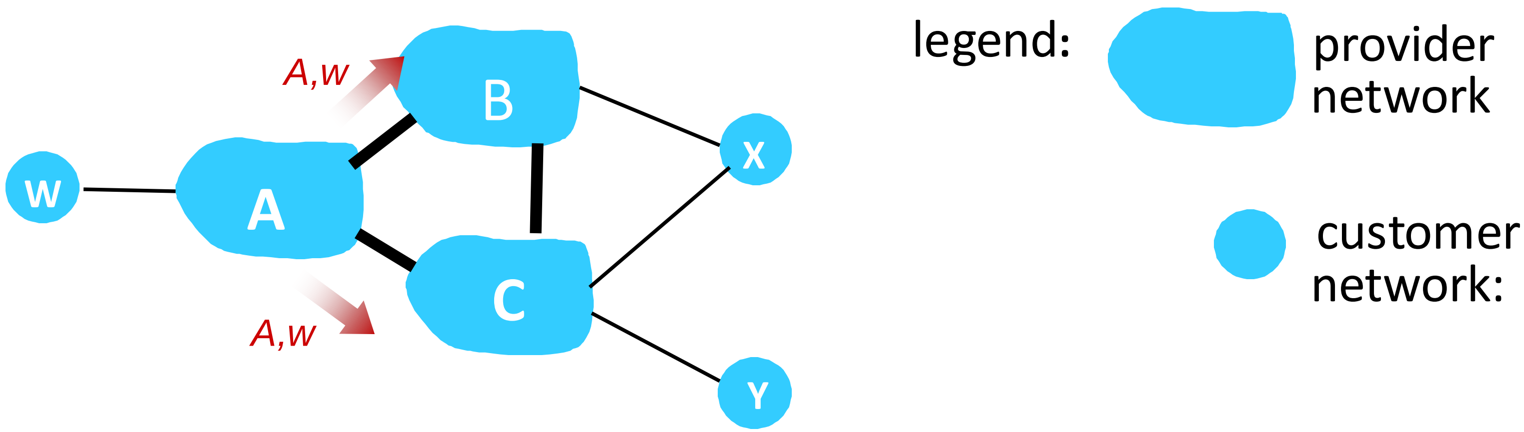
BGP: achieving policy via advertisements



Suppose an ISP only wants to route traffic to/from its customer networks (does not want to carry transit traffic between other ISPs)

- A,B,C are *provider networks*
- X,W,Y are customer (of provider networks)
- X is *dual-homed:* attached to two networks
- *policy to enforce:* X does not want to route from B to C via X
  - .. so X will not advertise to B a route to C

BGP: achieving policy via advertisements



Suppose an ISP only wants to route traffic to/from its customer networks (does not want to carry transit traffic between other ISPs)

- A advertises path Aw to B and to C
- B *chooses not to advertise* BAw to C:
  - B gets no "revenue" for routing CBAw, since none of C, A, w are B's customers
  - C does not learn about CBAw path
- C will route CAw (not using B) to get to w

Why different Intra-, Inter-AS routing ?

policy:

inter-AS: admin wants control over how its traffic routed, who routes through its net.

intra-AS: single admin, so no policy decisions needed

scale:

hierarchical routing saves table size, reduced update traffic

performance:

intra-AS: can focus on performance

inter-AS: policy may dominate over performance

Comparison:

Routing Protocol	Distance-Vector	Link-State	Path-Vector
RIP	✓		
OSPF		✓	
IS-IS		✓	
EIGRP	✓		
BGP			✓

Summary

Today:

- Intra – AS: OSPF
- Inter – AS: BGP

Canvas discussion:

- Reflection
- Exit ticket

Next time:

- read 5.5, 5.6 and 5.7 of K&R (SDN, ICMP, and SNMP)
- follow on Canvas! material and announcements

Any questions?