Chapter 5
Network Layer: Control Plane

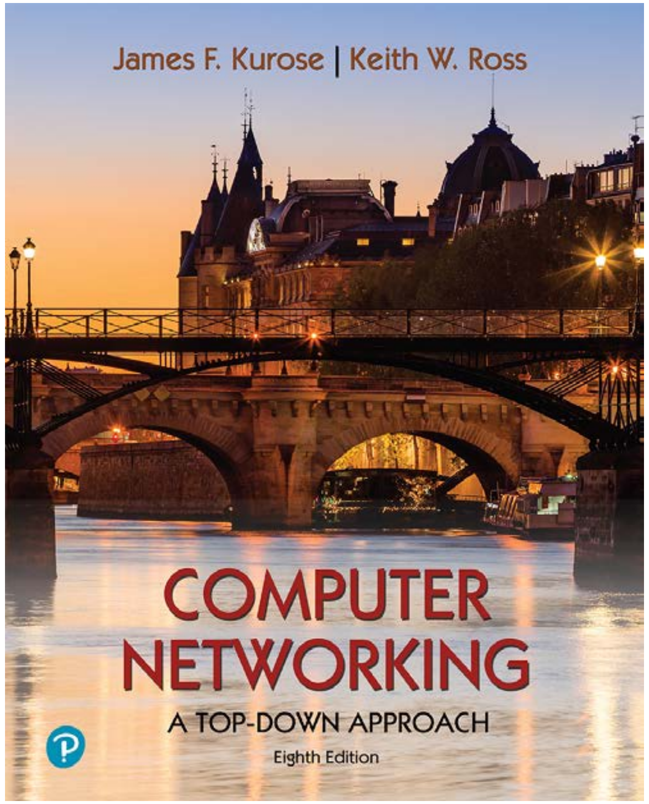
A note on the use of these PowerPoint slides:

We're making these slides freely available to all (faculty, students, readers). They're in PowerPoint form so you see the animations; and can add, modify, and delete slides (including this one) and slide content to suit your needs. They obviously represent a *lot* of work on our part. In return for use, we only ask the following:

- If you use these slides (e.g., in a class) that you mention their source (after all, we'd like people to use our book!)
- If you post any slides on a www site, that you note that they are adapted from (or perhaps identical to) our slides, and note our copyright of this material.

For a revision history, see the slide note for this page.

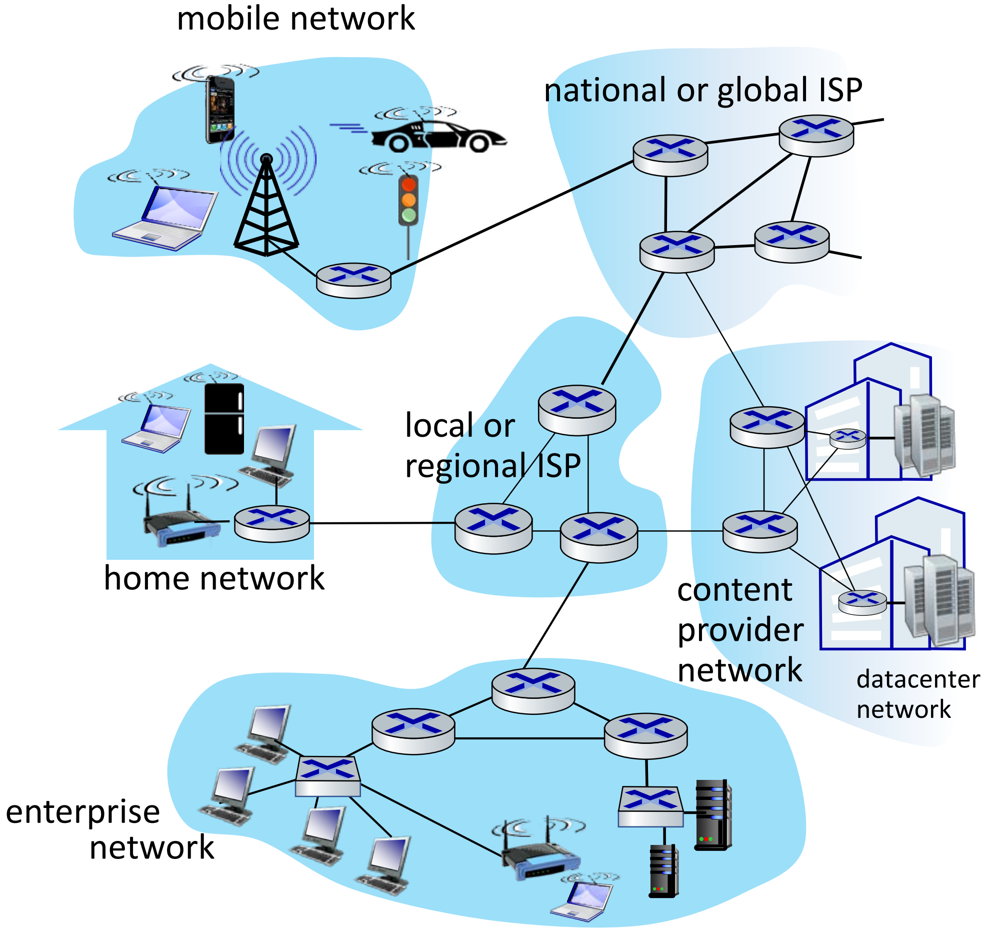
Thanks and enjoy! JFK/KWR



Computer Networking: A Top-Down Approach
8th edition
Jim Kurose, Keith Ross
Pearson, 2020

Internet structure: a "network of networks"

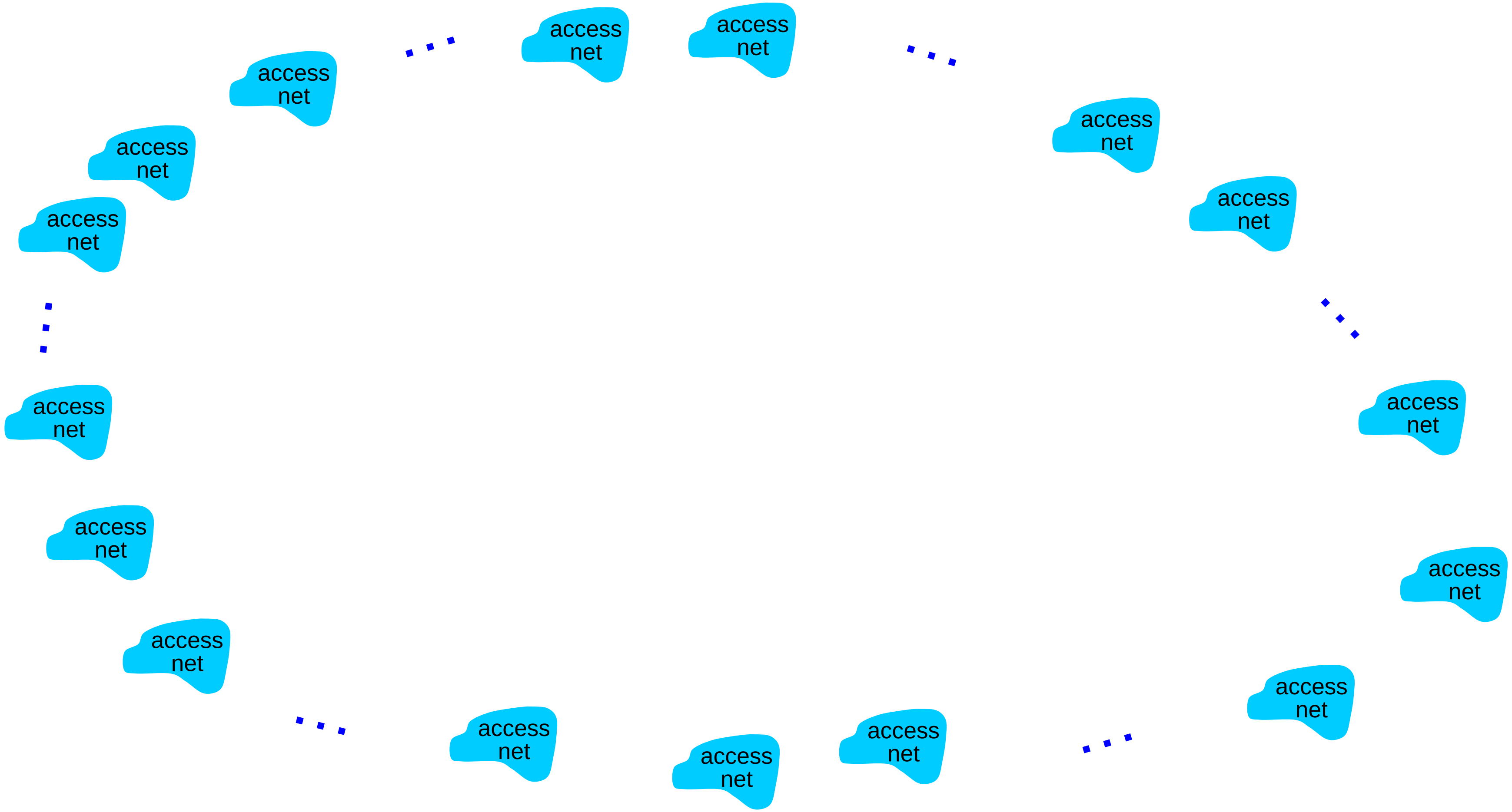
- hosts connect to Internet via access Internet Service Providers (ISPs)
- access ISPs in turn must be interconnected
  - so that *any* two hosts *(anywhere!)* can send packets to each other
- resulting network of networks is very complex
  - evolution driven by economics, national policies



Let's take a stepwise approach to describe current Internet structure

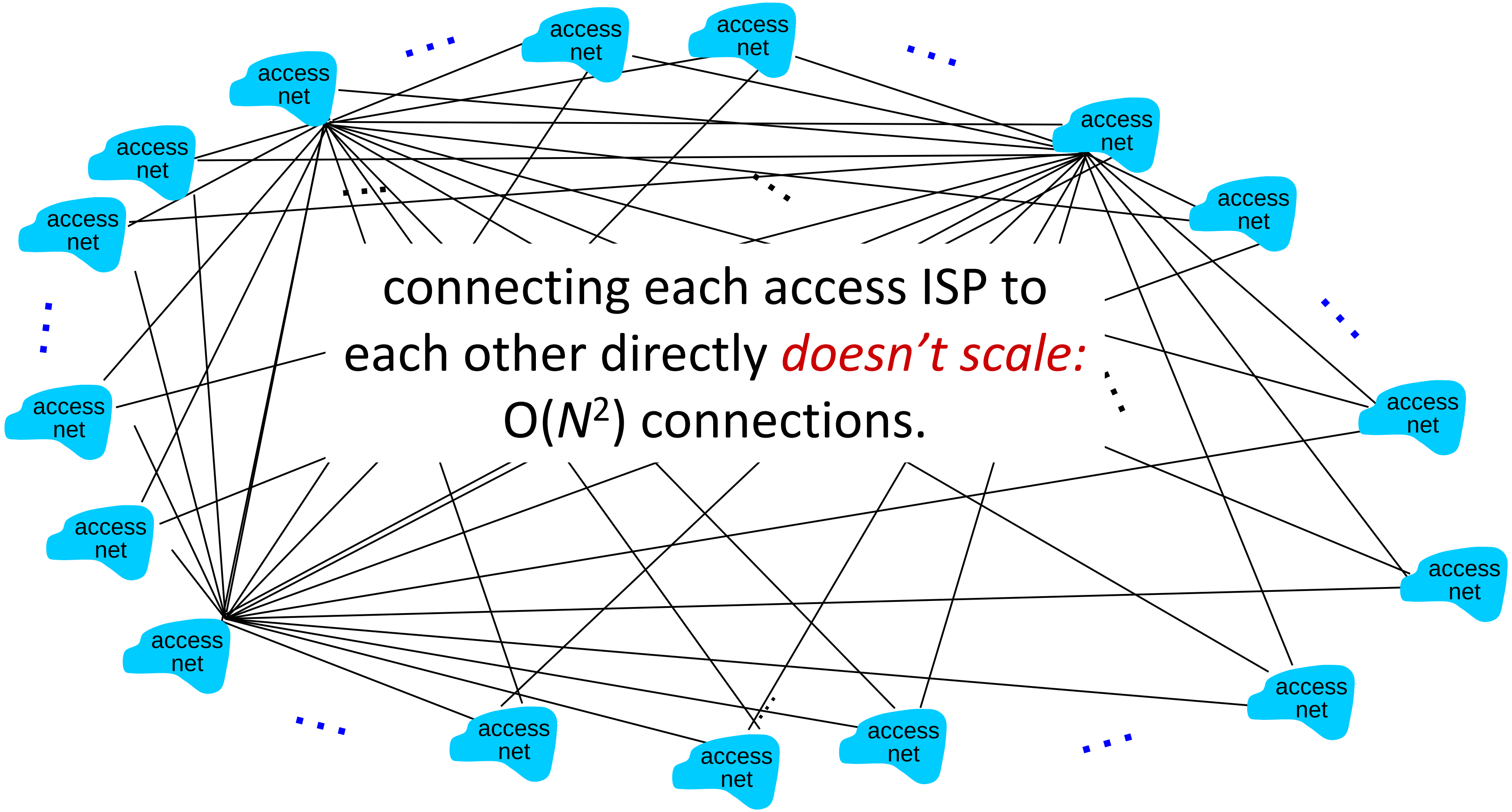
Internet structure: a "network of networks"

Question: given *millions* of access ISPs, how to connect them together?



Internet structure: a "network of networks"

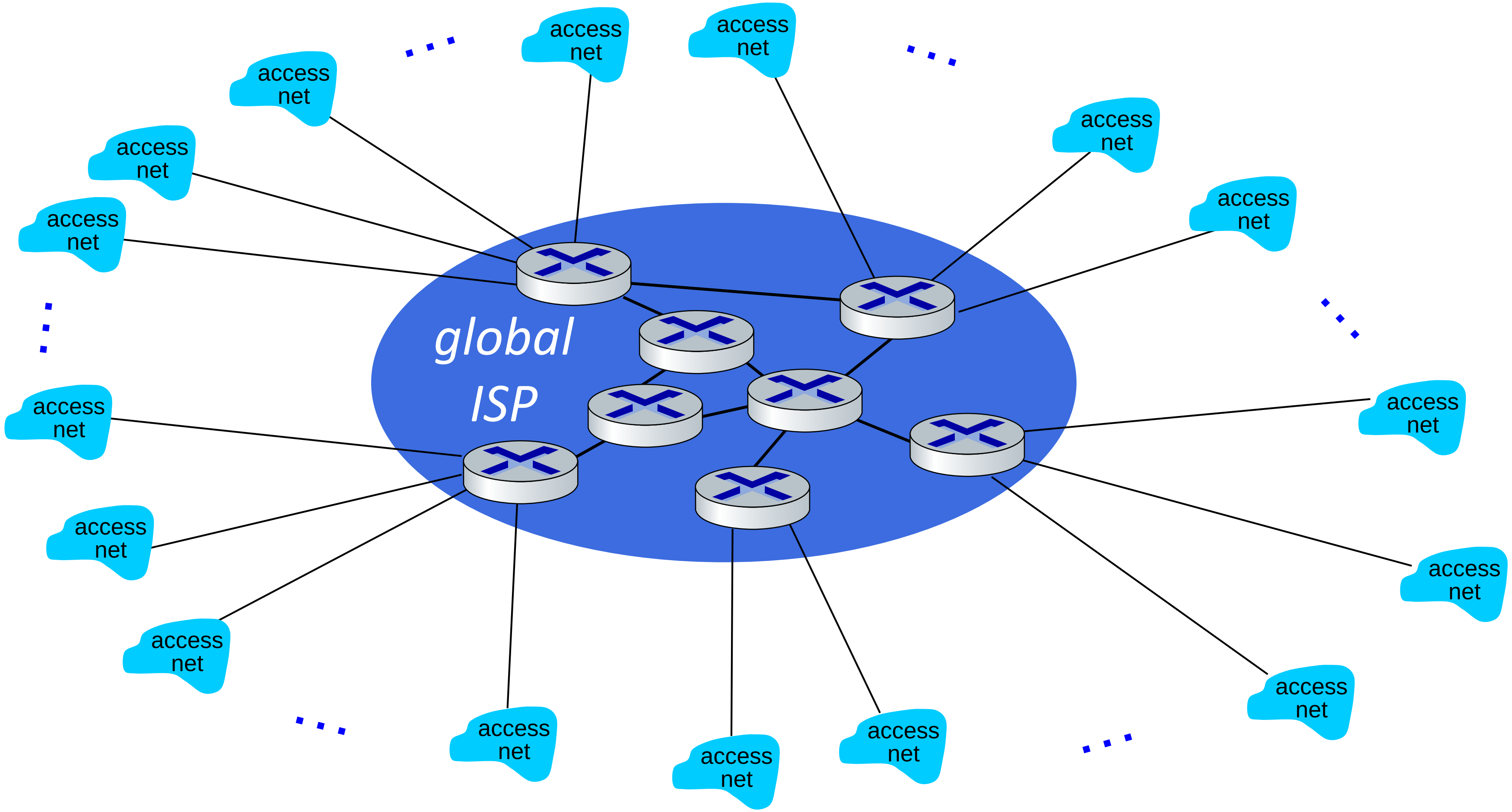
Question: given *millions* of access ISPs, how to connect them together?



Internet structure: a “network of networks”

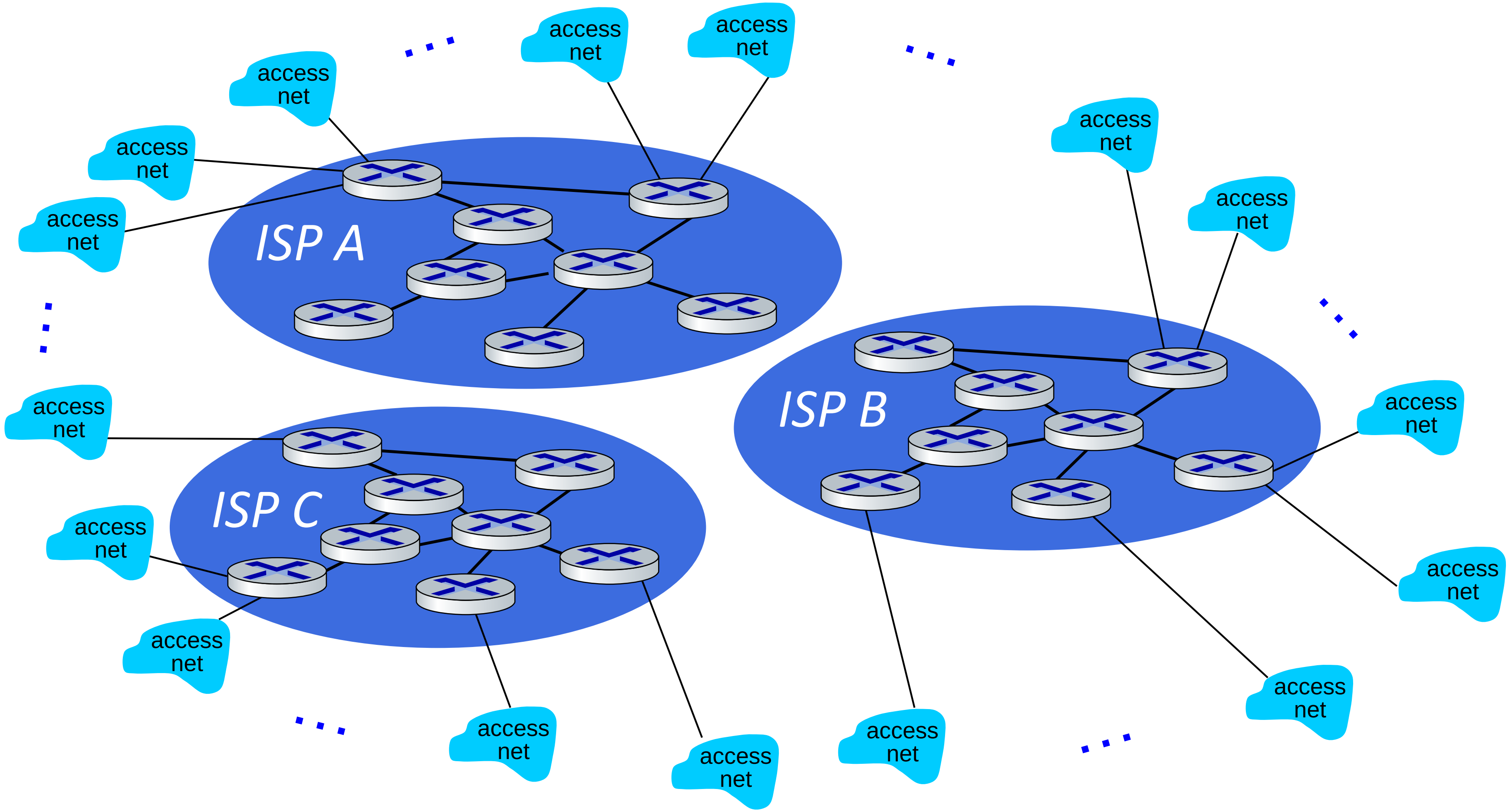
Option: connect each access ISP to one global transit ISP?

Customer and provider ISPs have economic agreement.



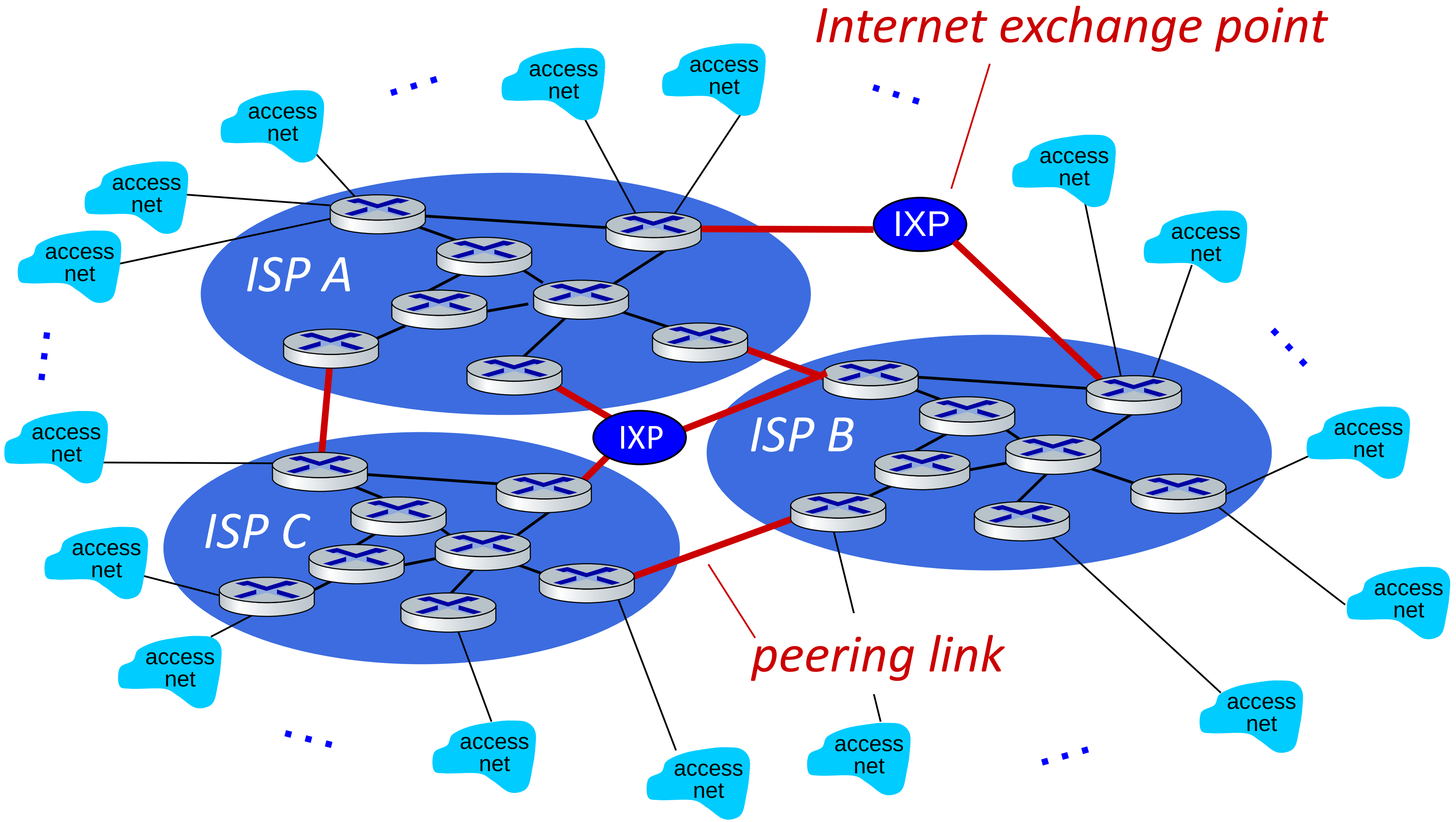
Internet structure: a “network of networks”

But if one global ISP is viable business, there will be competitors ….



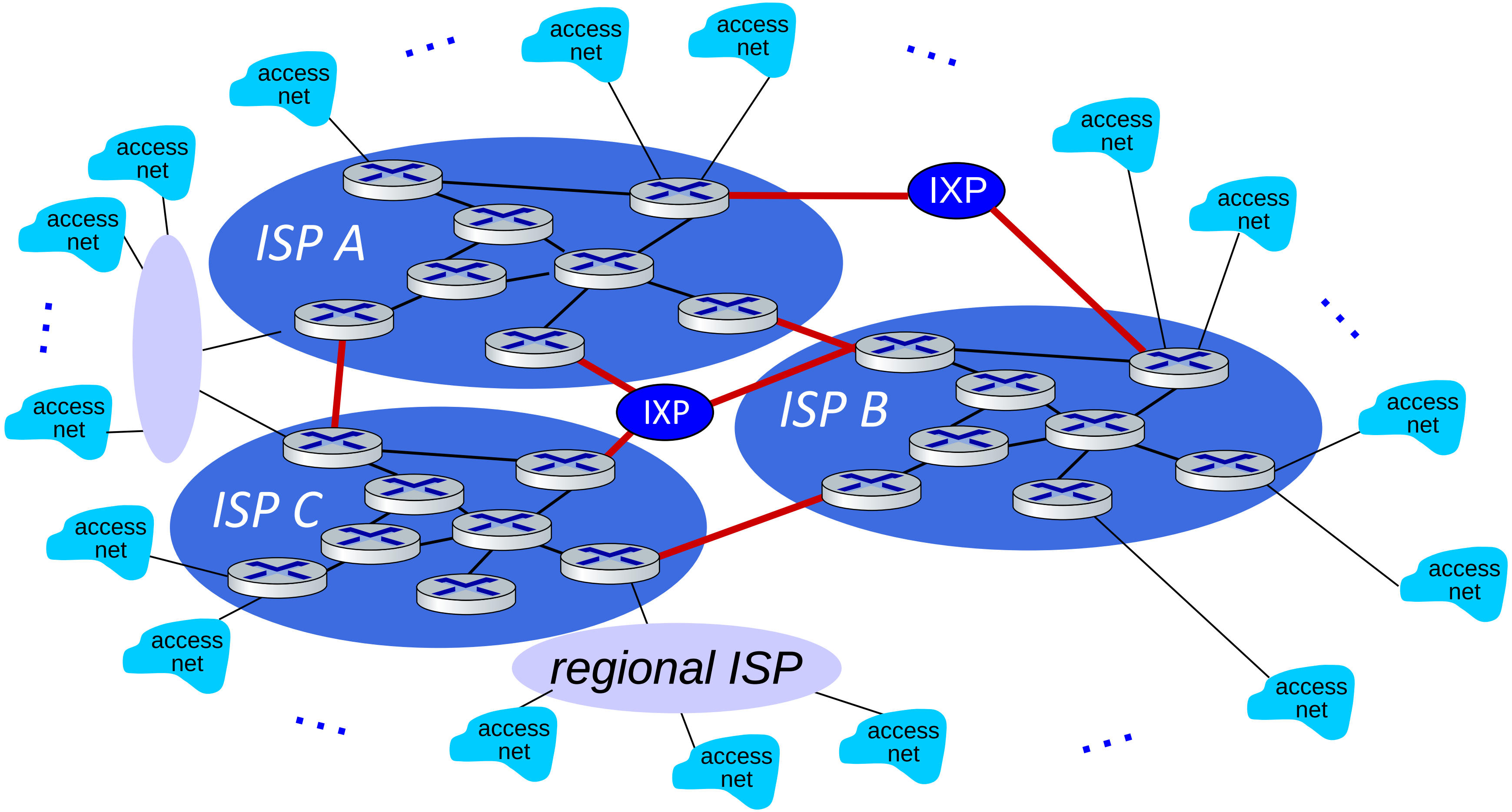
Internet structure: a "network of networks"

But if one global ISP is viable business, there will be competitors …. who will want to be connected



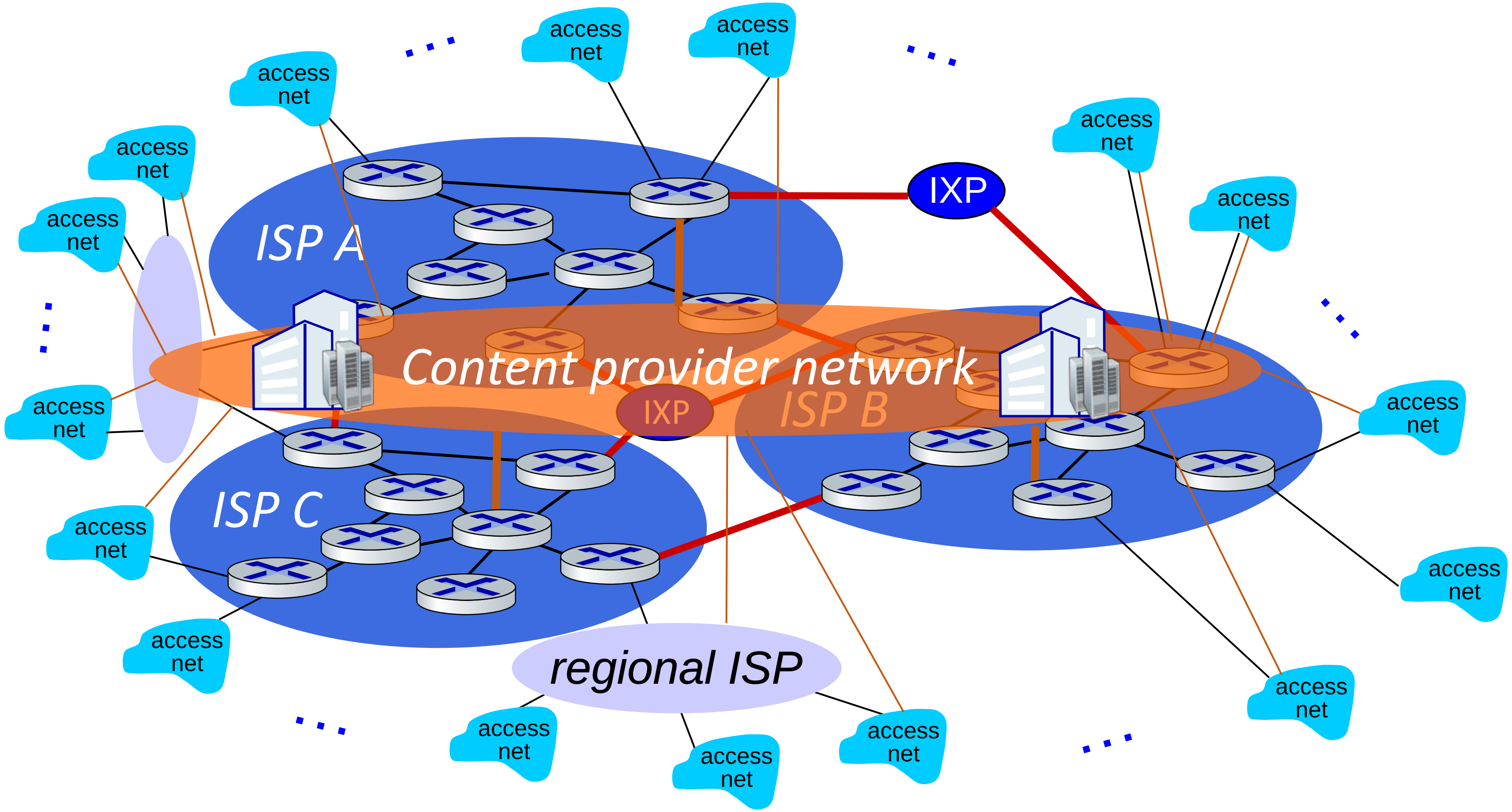
Internet structure: a "network of networks"

… and regional networks may arise to connect access nets to ISPs

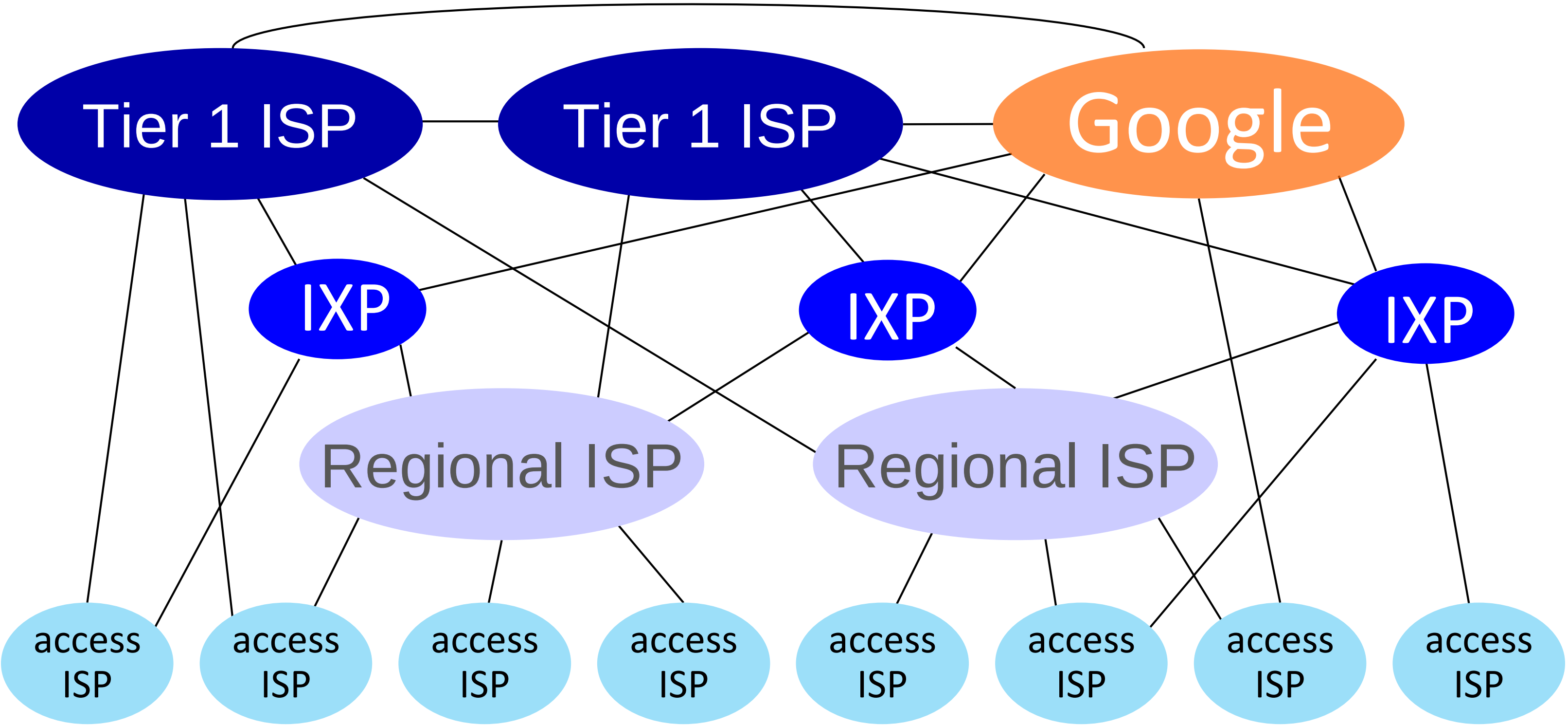


Internet structure: a "network of networks"

… and content provider networks (e.g., Google, Microsoft, Akamai) may run their own network, to bring services, content close to end users



Internet structure: a "network of networks"



At "center": small # of well-connected large networks

- "tier-1" commercial ISPs (e.g., Level 3, Sprint, AT&T, NTT), national & international coverage
- content provider networks (e.g., Google, Facebook): private network that connects its data centers to Internet, often bypassing tier-1, regional ISPs

Making routing scalable

our routing study thus far - idealized
- all routers identical
- network "flat"

… not true in practice

scale: billions of destinations:
- can't store all destinations in routing tables!
- routing table exchange would swamp links!

administrative autonomy:
- Internet: a network of networks
- each network admin may want to control routing in its own network


A
B

Internet approach to scalable routing

aggregate routers into regions known as "autonomous systems" (AS) (a.k.a. "domains")

intra-AS (aka "intra-domain"):
routing among *within same AS ("network")*

- all routers in AS must run same intra-domain protocol
- routers in different AS can run different intra-domain routing protocols
- gateway router: at "edge" of its own AS, has link(s) to router(s) in other AS'es

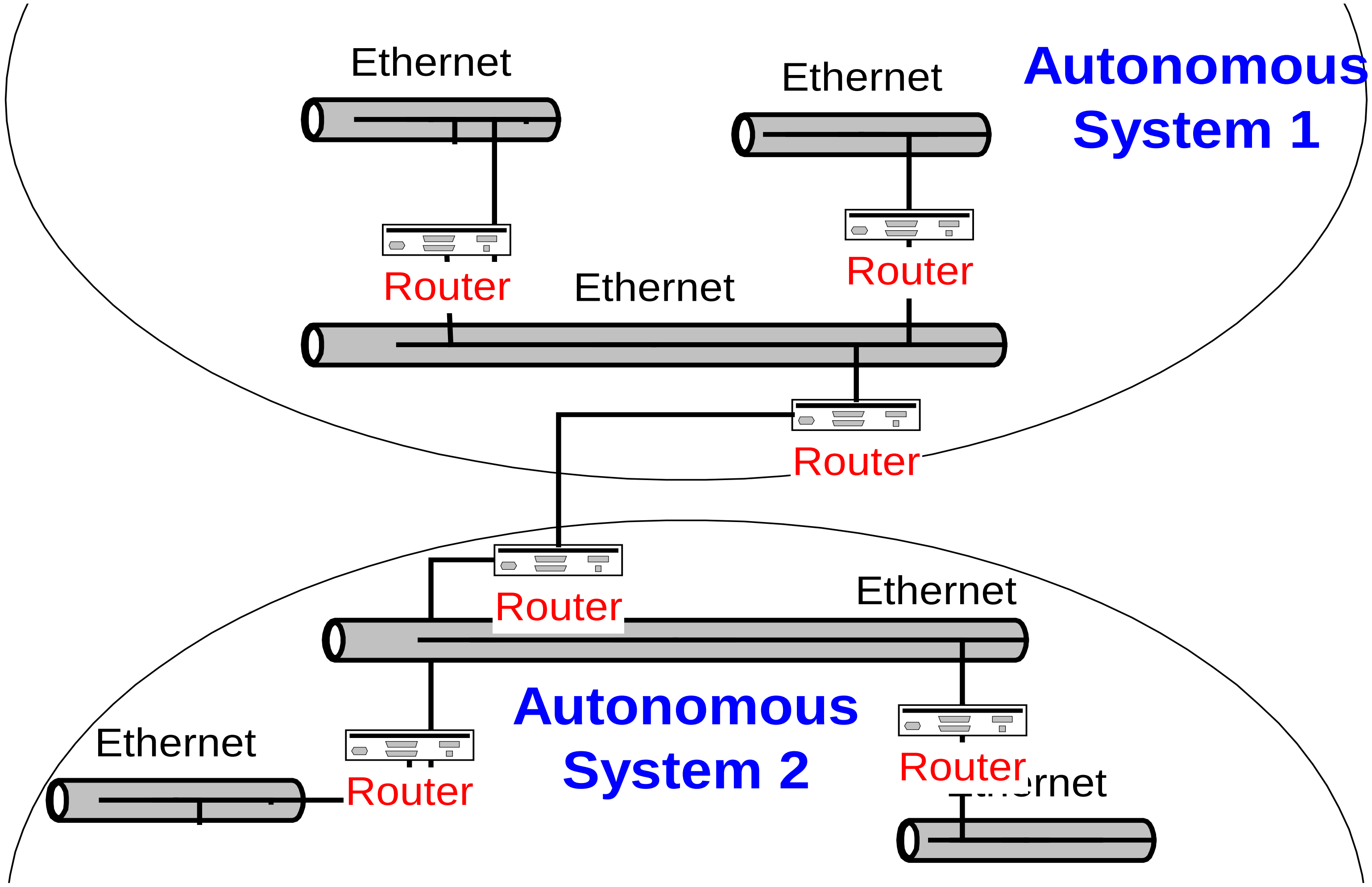
inter-AS (aka "inter-domain"):
routing *among* AS'es

- gateways perform inter-domain routing (as well as intra-domain routing)

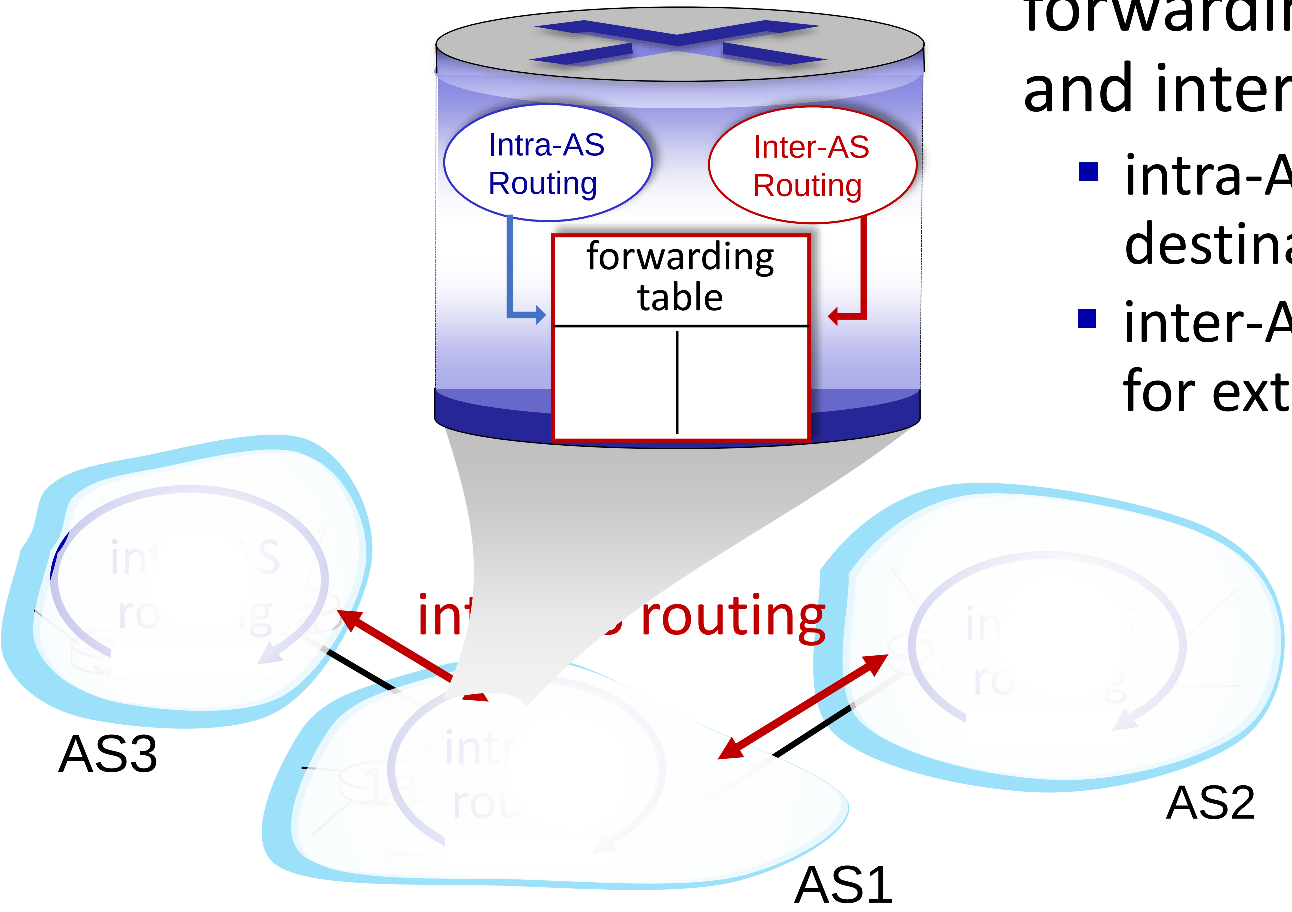
Autonomous Systems

- An **autonomous system** is a region of the Internet that is administered by a single authority.
- Examples of autonomous regions are:
  - UVA’s campus network
  - MCI’s backbone network
  - Regional Internet Service Provider
- Types of autonomous system (AS):
  - Stub AS: has connection to only one AS, only carry local traffic
  - Multihomed AS: has connection to >1 AS, but does not carry transit traffic
  - Transit AS: has connection to >1 AS and carries transit traffic
- Routing is done differently within an autonomous system (**intradomain routing**) and between autonomous system **(interdomain routing**).

Autonomous Systems (AS)



Interconnected ASes



forwarding table configured by intra- and inter-AS routing algorithms

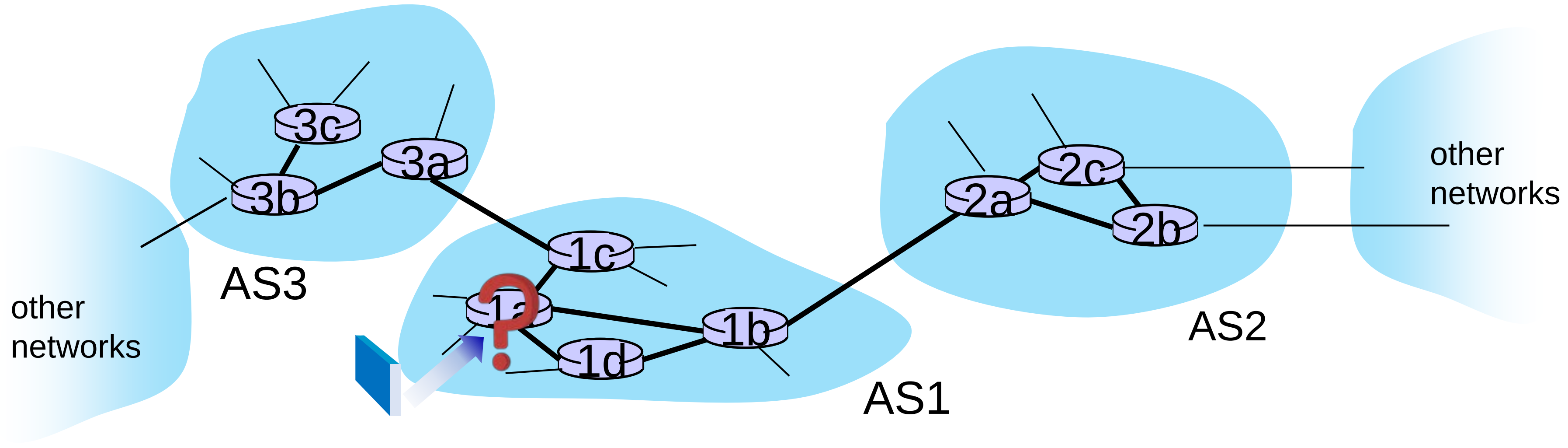
- intra-AS routing determine entries for destinations within AS
- inter-AS & intra-AS determine entries for external destinations

Inter-AS routing: a role in intradomain forwarding

- suppose router in AS1 receives datagram destined outside of AS1:
  - router should forward packet to gateway router in AS1, but which one?

AS1 inter-domain routing must:

1. learn which destinations reachable through AS2, which through AS3
2. propagate this reachability info to all routers in AS1



Inter-AS routing: routing within an AS

most common intra-AS routing protocols:

- RIP: Routing Information Protocol [RFC 1723]
  - classic DV: DVs exchanged every 30 secs
  - no longer widely used
- EIGRP: Enhanced Interior Gateway Routing Protocol
  - DV based
  - formerly Cisco-proprietary for decades (became open in 2013 [RFC 7868])
- OSPF: Open Shortest Path First [RFC 2328]
  - link-state routing
  - IS-IS protocol (ISO standard, not RFC standard) essentially same as OSPF

OSPF (Open Shortest Path First) routing

- "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)

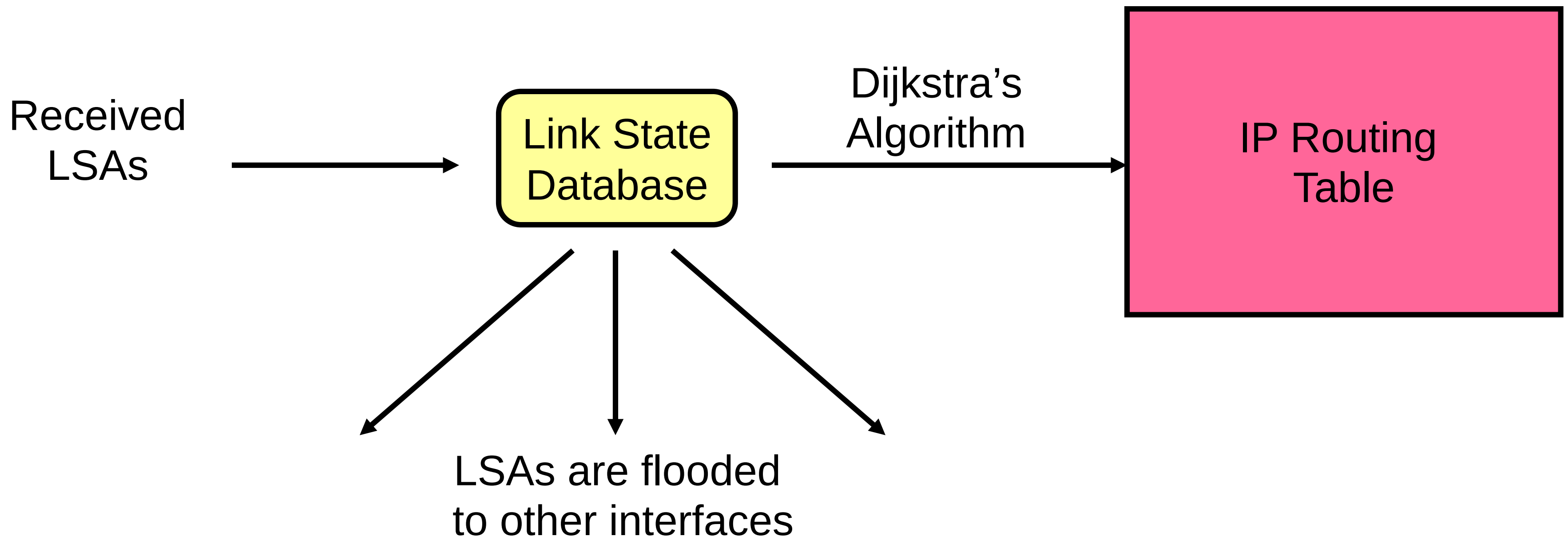
Link State Routing: Basic princples

1. Each router establishes a relationship *("adjacency")* with its neighbors
2. Each router generates *link state advertisements (LSAs)* which are distributed to all routers

   LSA = (link id, state of the link, cost, neighbors of the link)

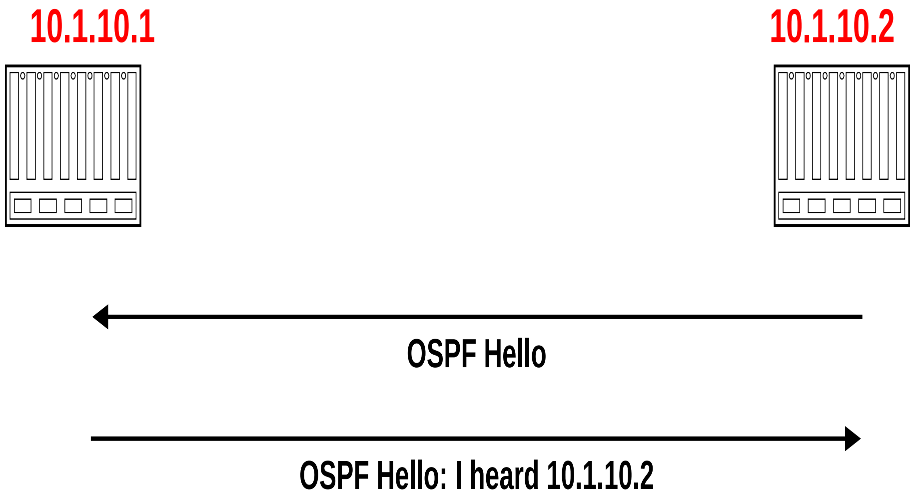
3. Each router maintains a database of all received LSAs (*topological database* or *link state database*), which describes the network as a graph with weighted edges
4. Each router uses its link state database to run a shortest path algorithm (Dijikstra's algorithm) to produce the shortest path to each network

Operation of a Link State Routing protocol



Discovery of Neighbors

- Routers multicast OSPF Hello packets on all OSPF-enabled interfaces.
- If two routers share a link, they can become neighbors, and establish an adjacency



Scenario:
Router 10.1.10.2 restarts

- After becoming a neighbor, routers exchange their link state databases

Synchronizing OSPF Databases

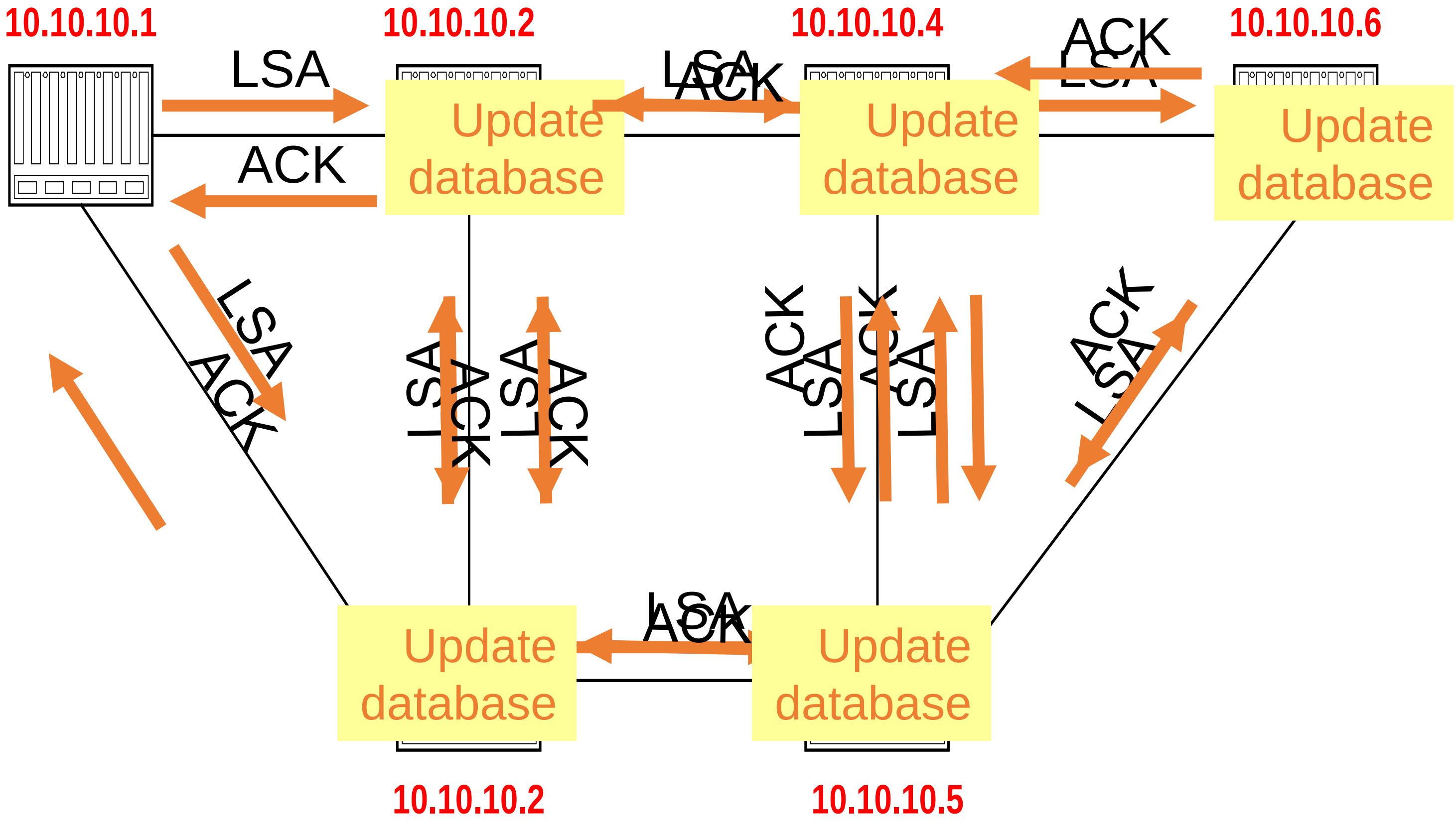
- While the Hello packet was used to establish neighbor adjacencies, the other four types of OSPF packets are used during the process of exchanging and synchronizing link state databases.

OSPF Packet Descriptions

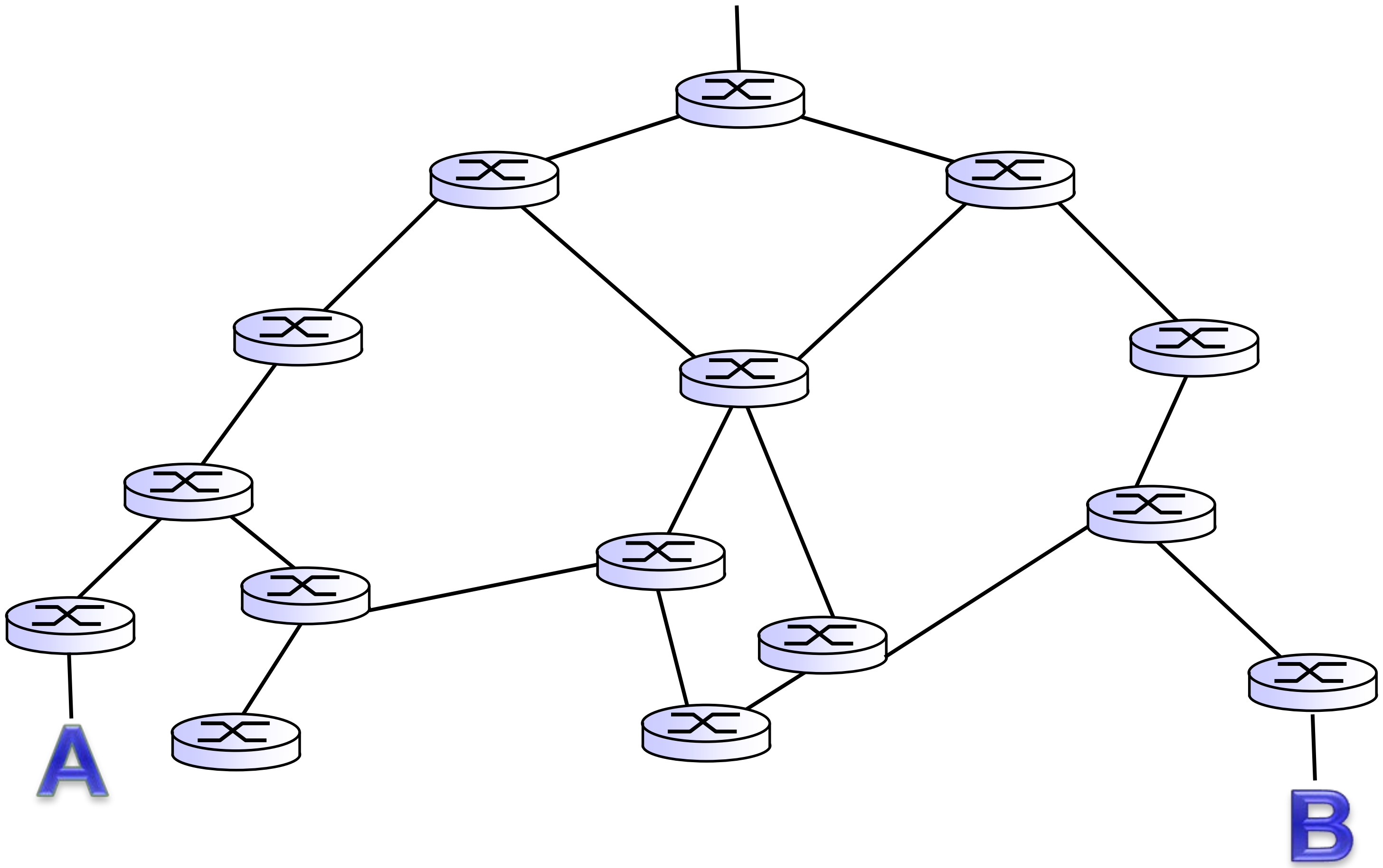
Type	Packet Name	Description
1	Hello	Discovers neighbors and builds adjacencies between them
2	Database Description (DBD)	Checks for database synchronization between routers
3	Link-State Request (LSR)	Requests specific link-state records from router to router
4	Link-State Update (LSU)	Sends specifically requested link-state records
5	Link-State Acknowledgment (LSAck)	Acknowledges the other packet types

Routing Data Distribution

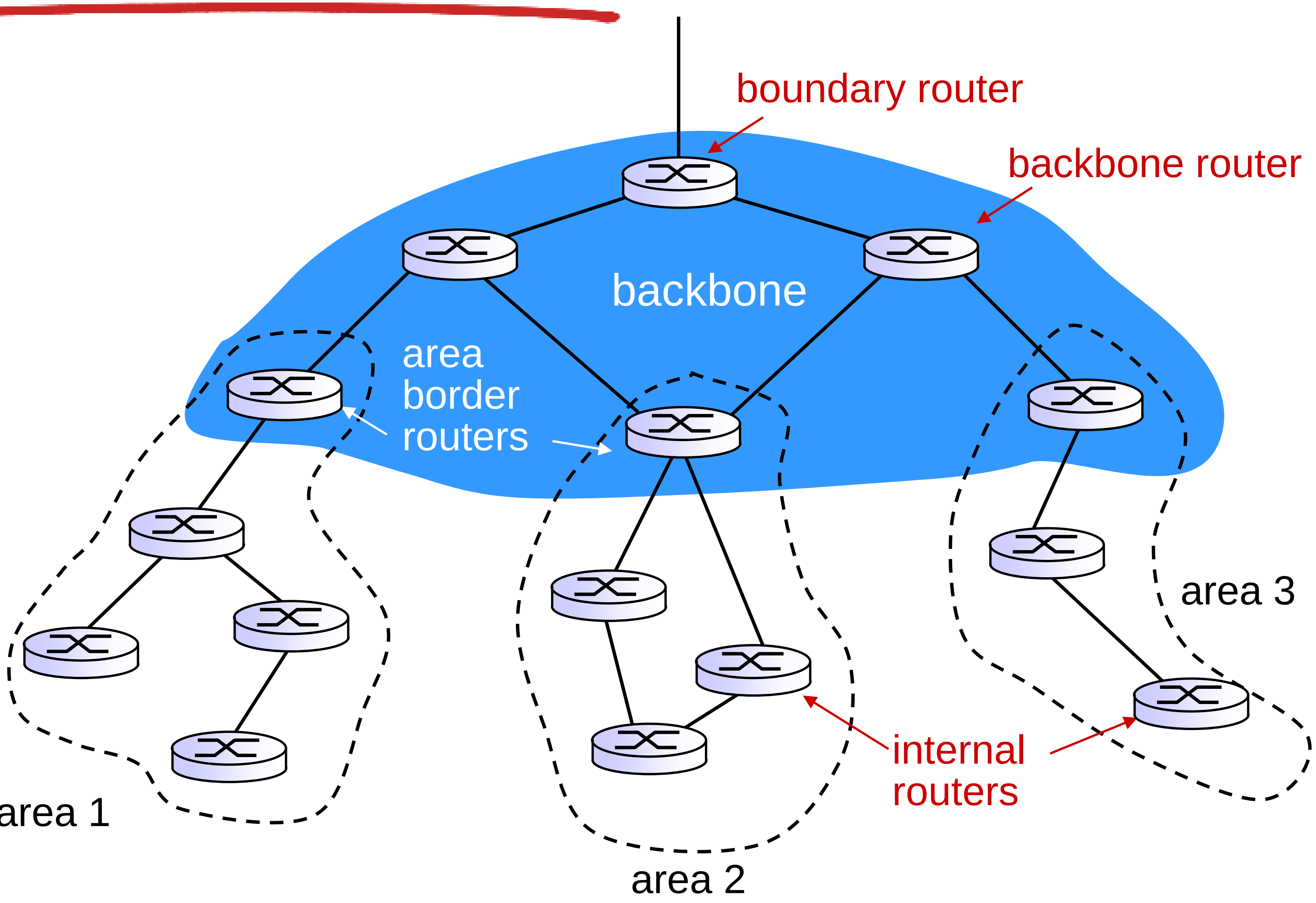
- LSA-Updates are distributed to all other routers via **Reliable Flooding**
- **Example:** Flooding of LSA from 10.10.10.1



Enterprise Network

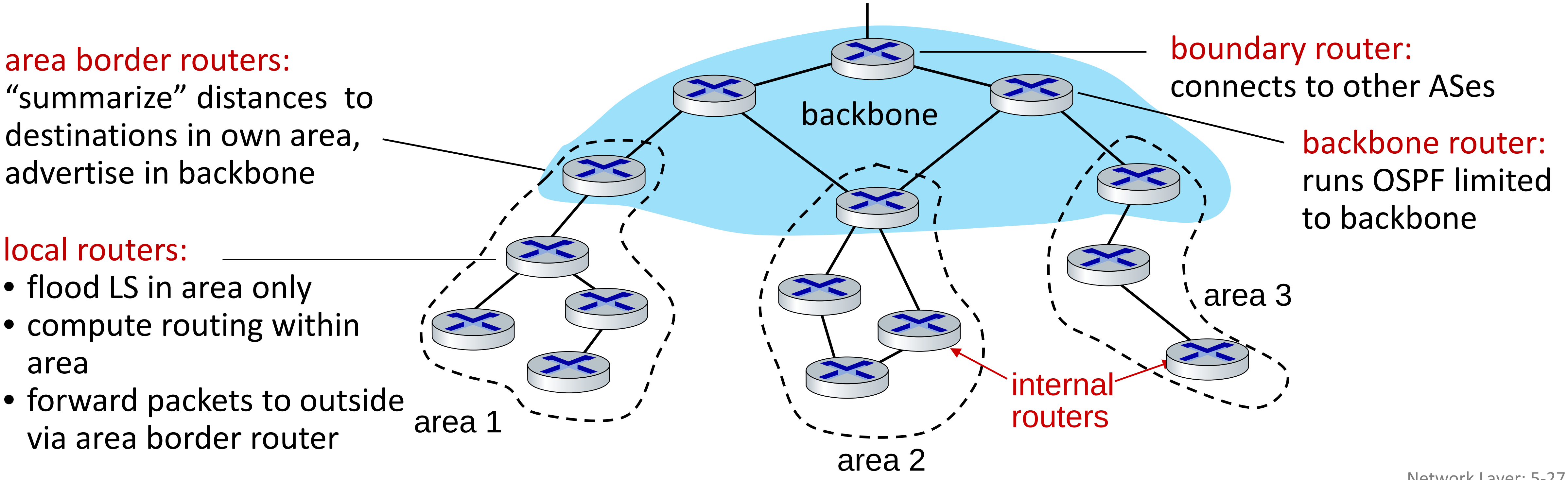


Hierarchical OSPF

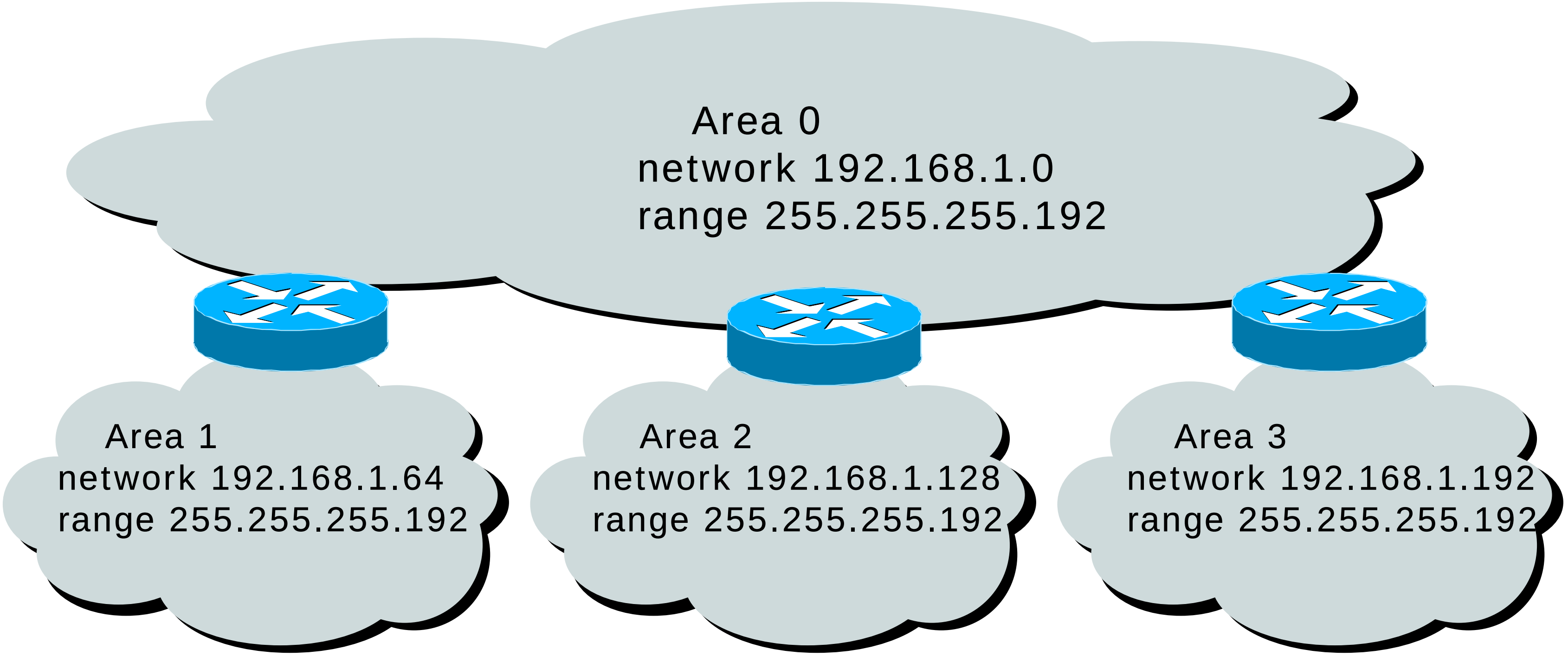


Hierarchical OSPF

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



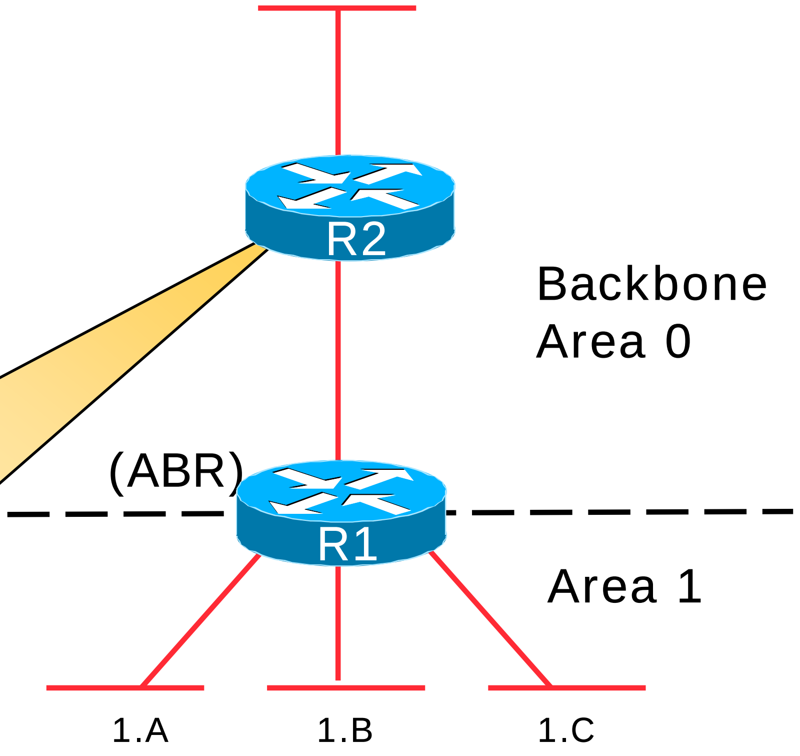
Addressing for Areas



Assign contiguous ranges of subnets per area to facilitate summarisation

Inter-Area Route Summarisation

- Prefix or all subnets
- Prefix or all networks
- ‘Area range’ command



With summarisation

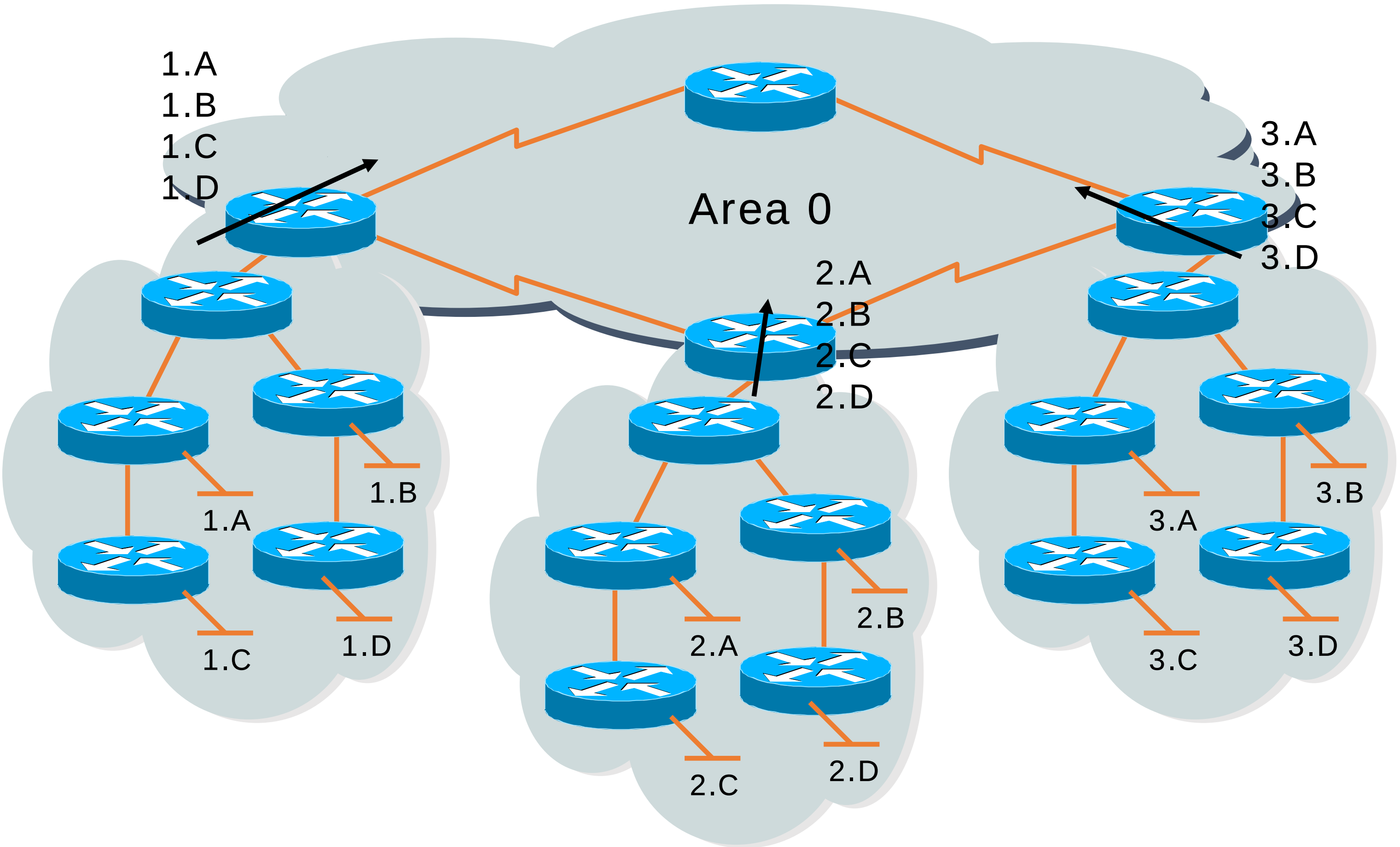
Network	Next Hop
1	R1

Without summarisation

Network	Next Hop
1.A	R1
1.B	R1
1.C	R1

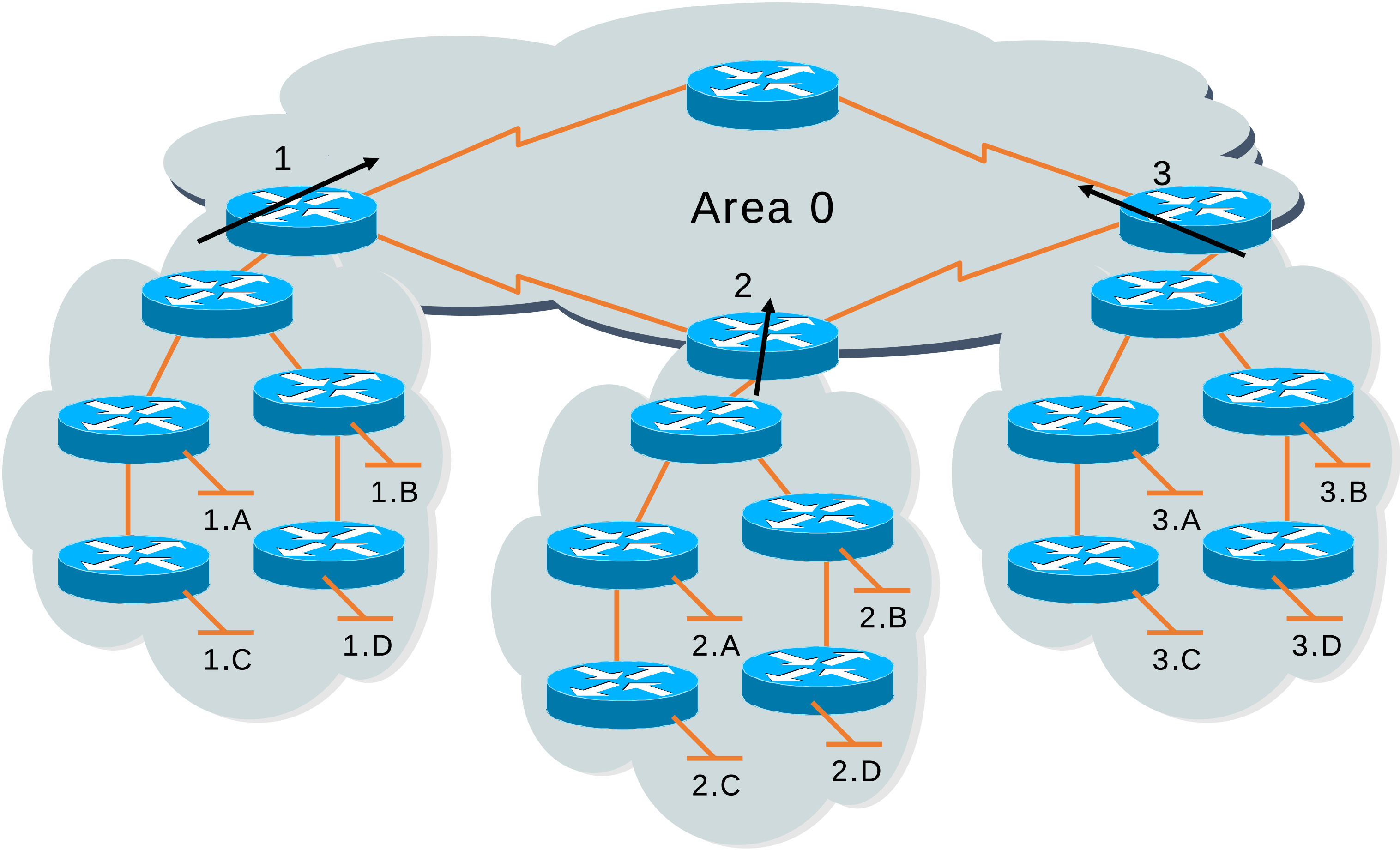
No Summarisation

- Specific Link LSA advertised out of each area
- Link state changes propagated out of each area



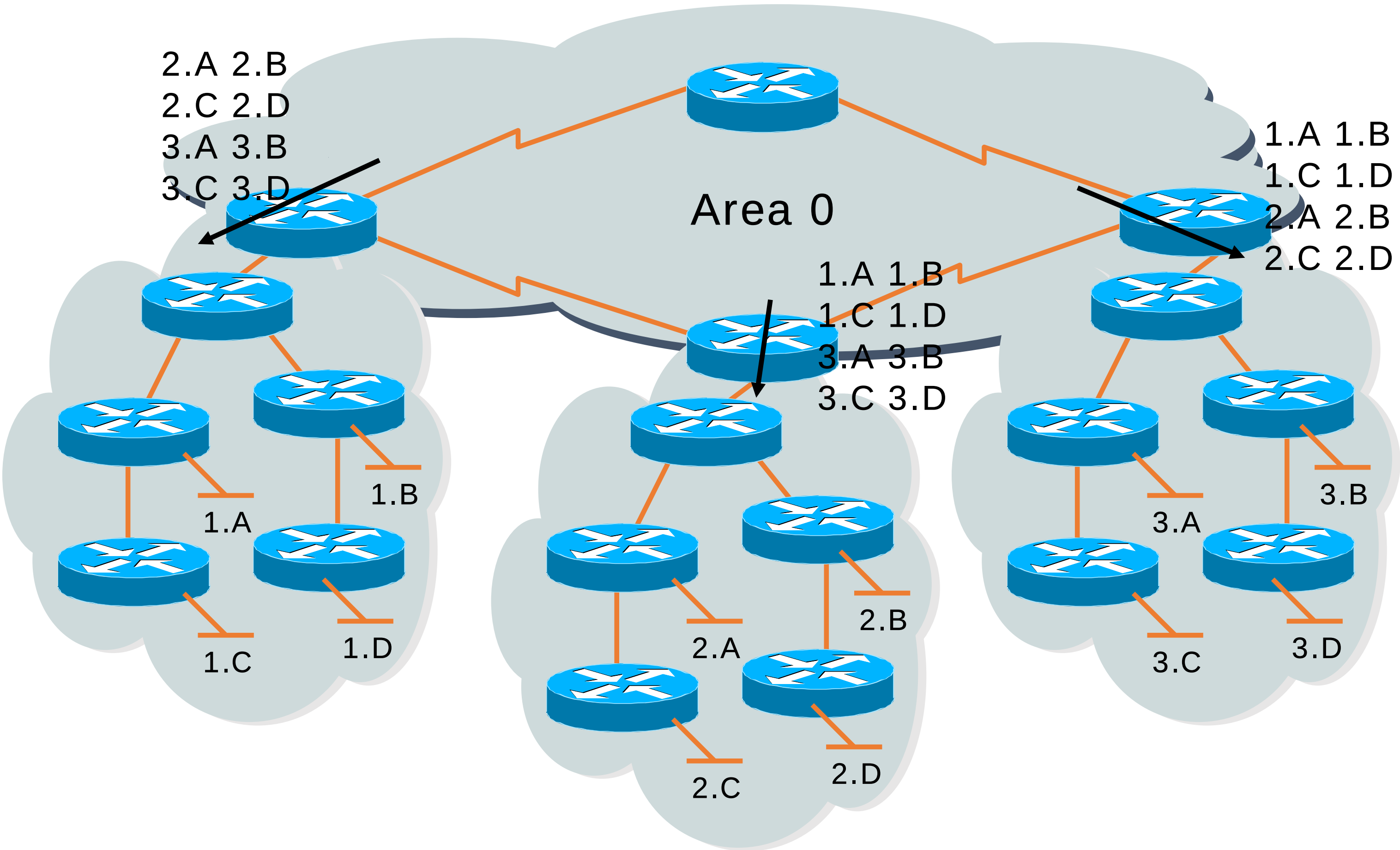
With Summarisation

- Only summary LSA advertised out of each area
- Link state changes do not propagate out of the area



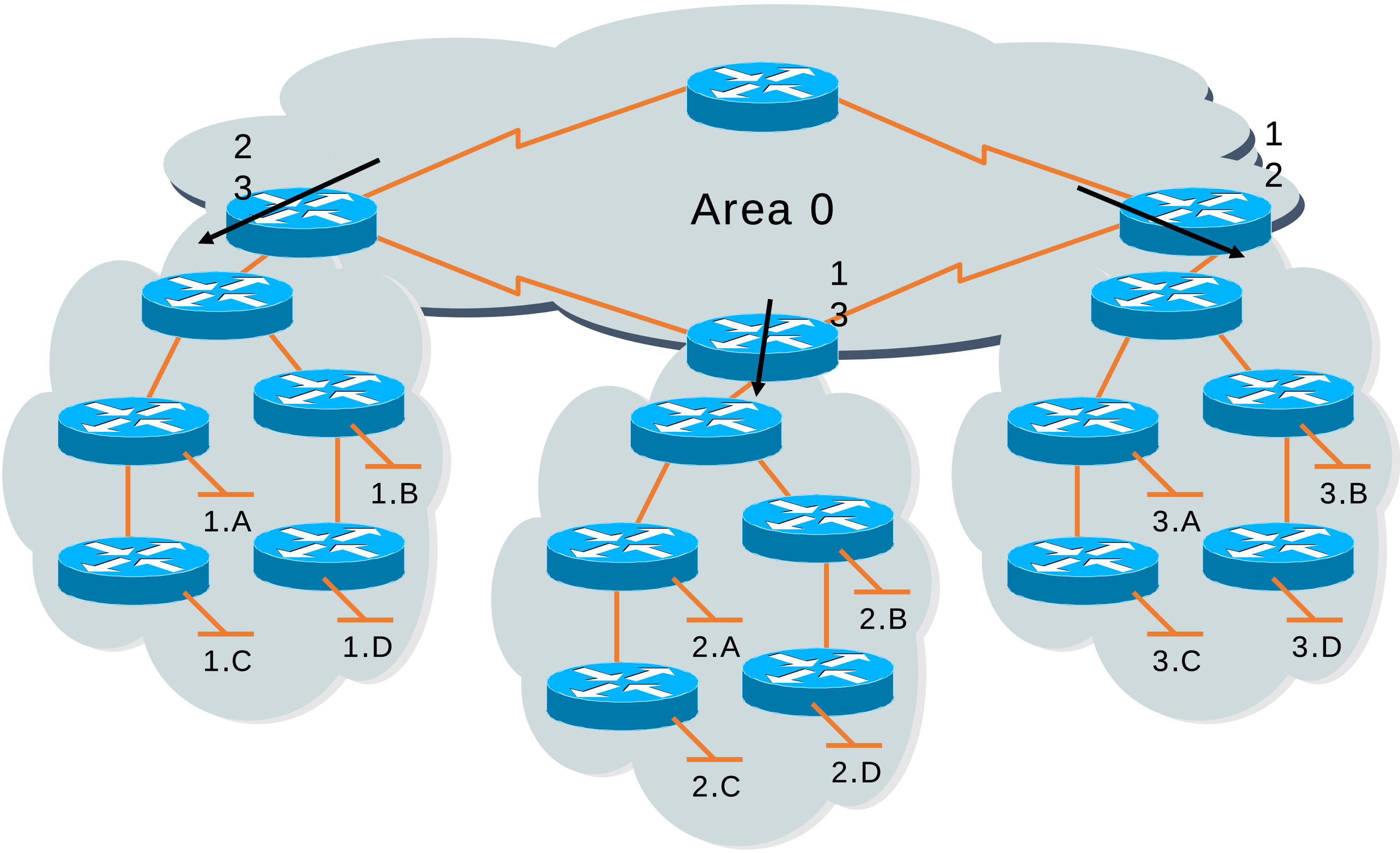
No Summarisation

- Specific Link LSA advertised in to each area
- Link state changes propagated in to each area

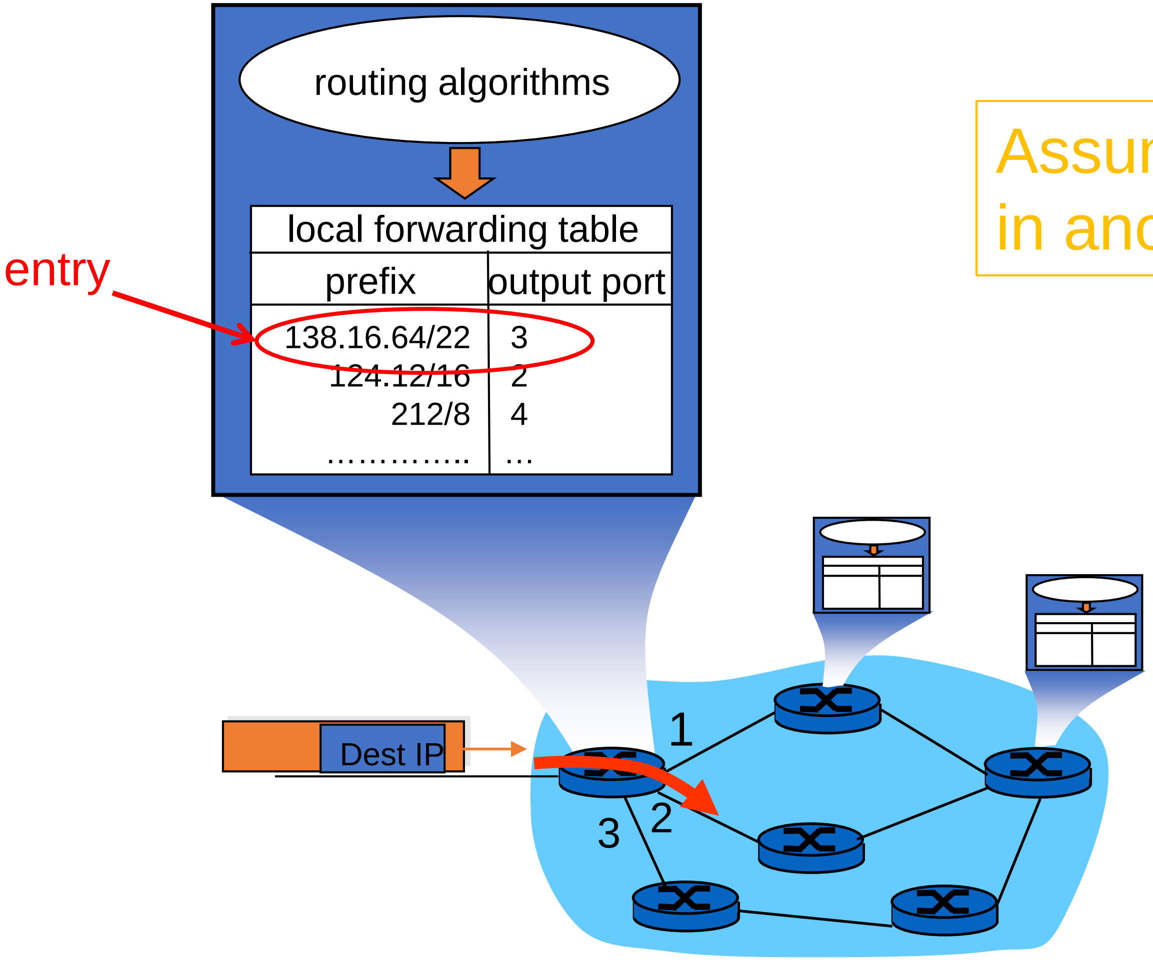


With Summarisation

- Only summary link LSA advertised in to each area
- Link state changes do not propagate in to each area



How does entry get in forwarding table?

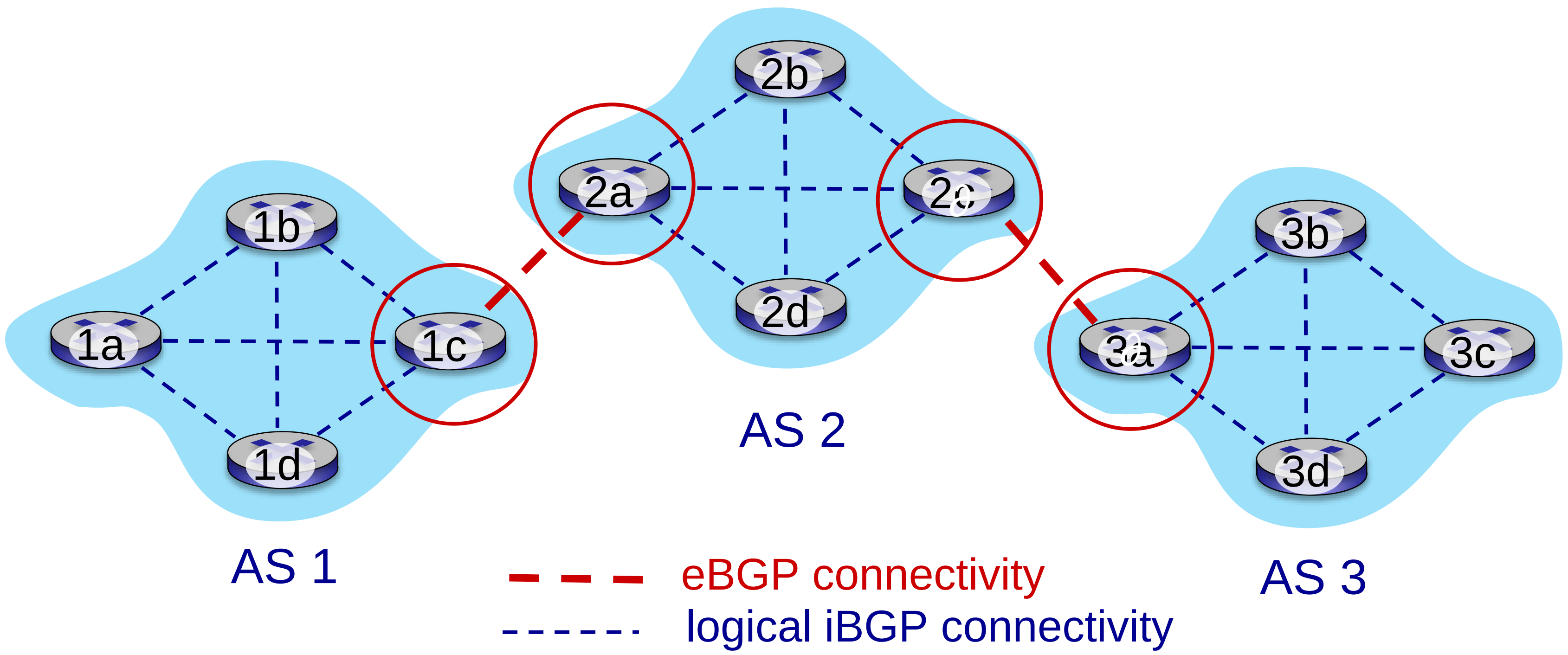


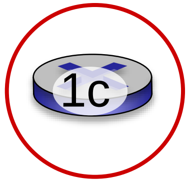
Assume prefix is in another AS.

Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the* de facto inter-domain routing protocol
  - "glue that holds the Internet together"
- allows subnet to advertise its existence, and the destinations it can reach, to rest of Internet: *"I am here, here is who I can reach, and how"*
- 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*

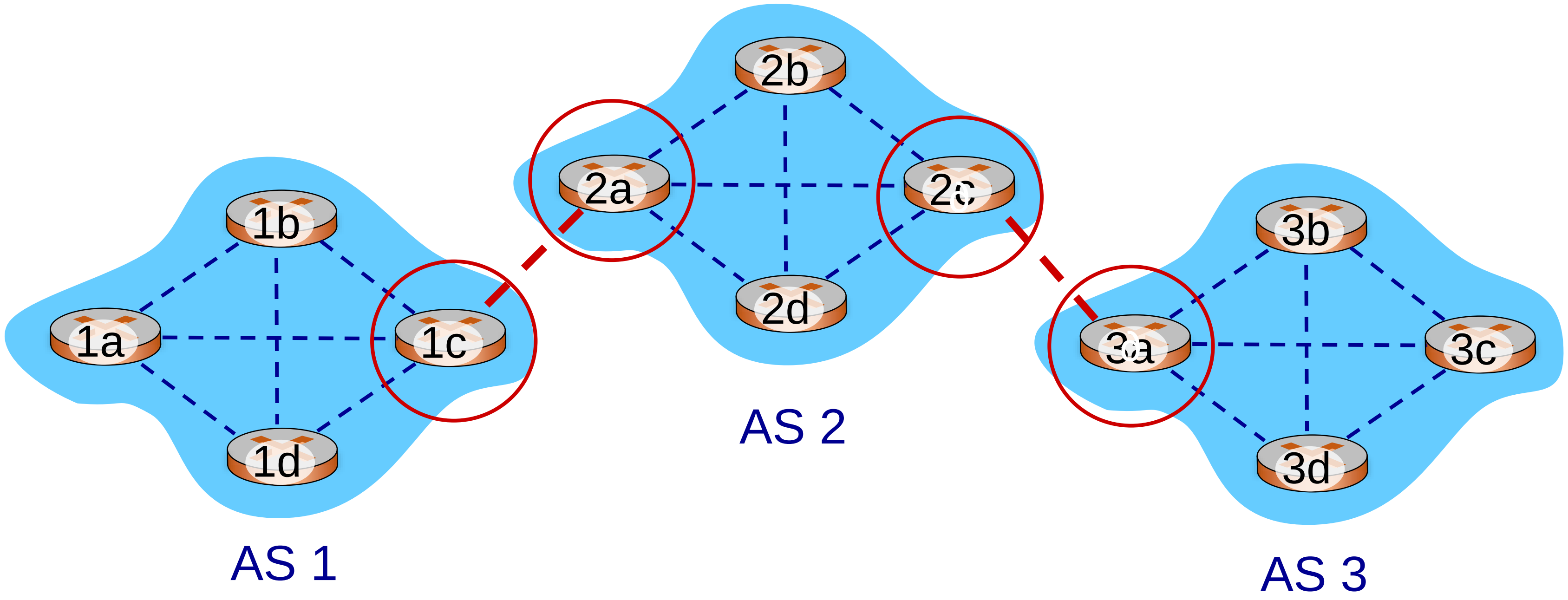
eBGP, iBGP connections





gateway routers run both eBGP and iBGP protocols

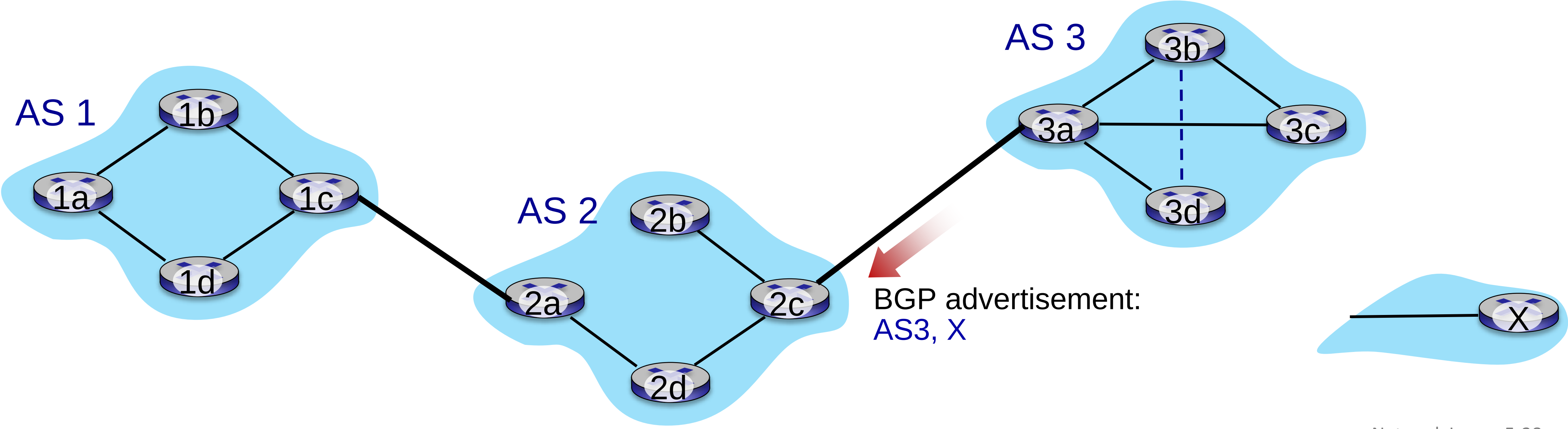
Route establishment in BGP



- For networks in AS1 and AS2 to communicate:
  - AS1 must announce a route to AS2
  - AS2 must accept the route from AS1
  - AS2 must announce a route to AS1
  - AS1 must accept the route from AS2

BGP basics

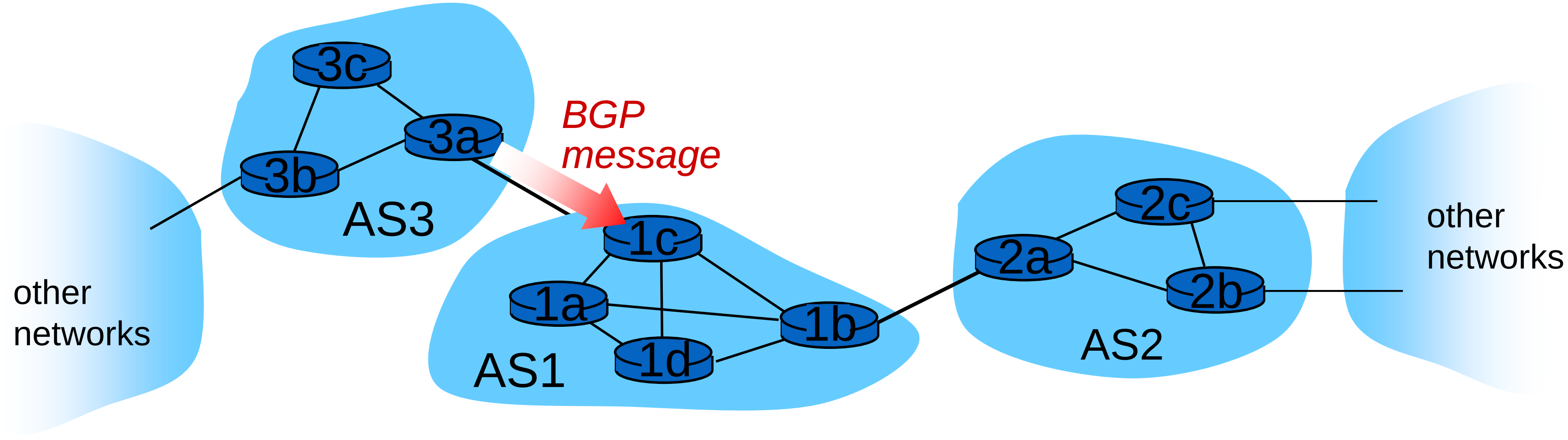
- BGP session: two BGP routers ("peers") exchange BGP messages over semi-permanent TCP connection:
  - advertising *paths* to different destination network prefixes (BGP is a "path vector" protocol)
- when AS3 gateway 3a advertises path AS3,X to AS2 gateway 2c:
  - AS3 *promises* to AS2 it will forward datagrams towards X



Path attributes and BGP routes

- BGP advertised route: prefix + attributes
  - prefix: destination being advertised
  - 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

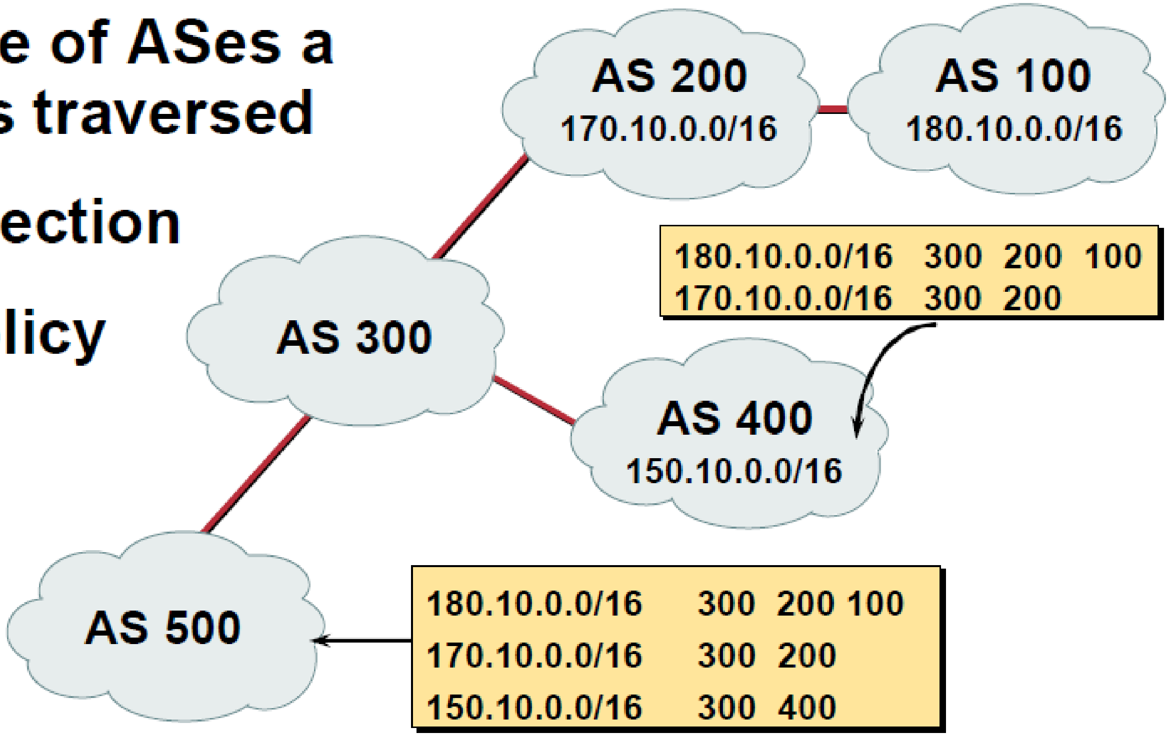
BGP route advertisement



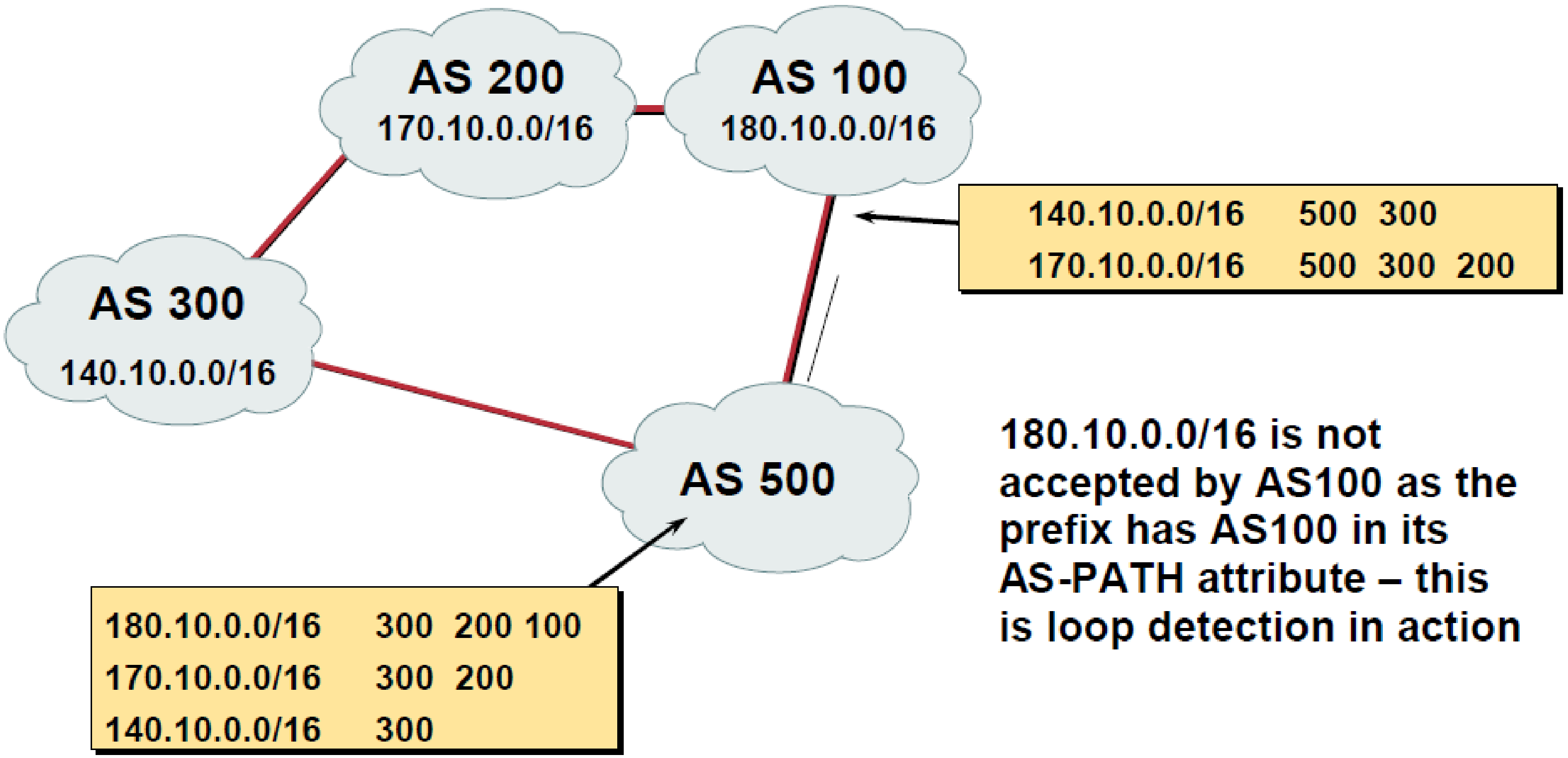
- BGP message contains "routes"
- "route" is a prefix and attributes: AS-PATH, NEXT-HOP,…
- Example: route:
  - Prefix:138.16.64/22 ; AS-PATH: AS3 AS131 ; NEXT-HOP: 201.44.13.125

AS Path

- Sequence of ASes a route has traversed
- Loop detection
- Apply policy



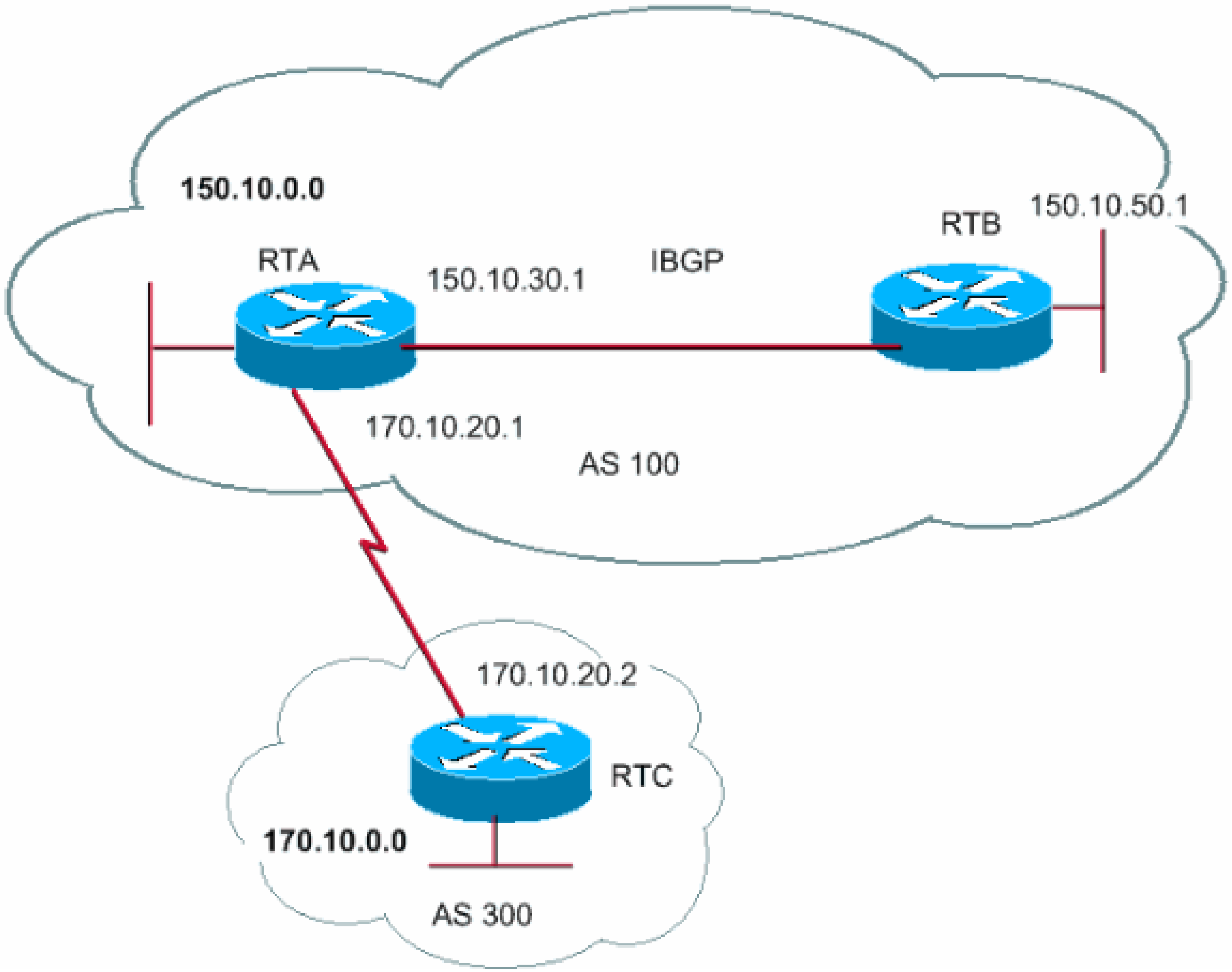
AS-Path loop detection



Next Hop

- IGP should carry route to next hops
- Recursive route look-up
- Unlinks BGP from actual physical topology
- Allows IGP to make intelligent forwarding decision

Next Hop

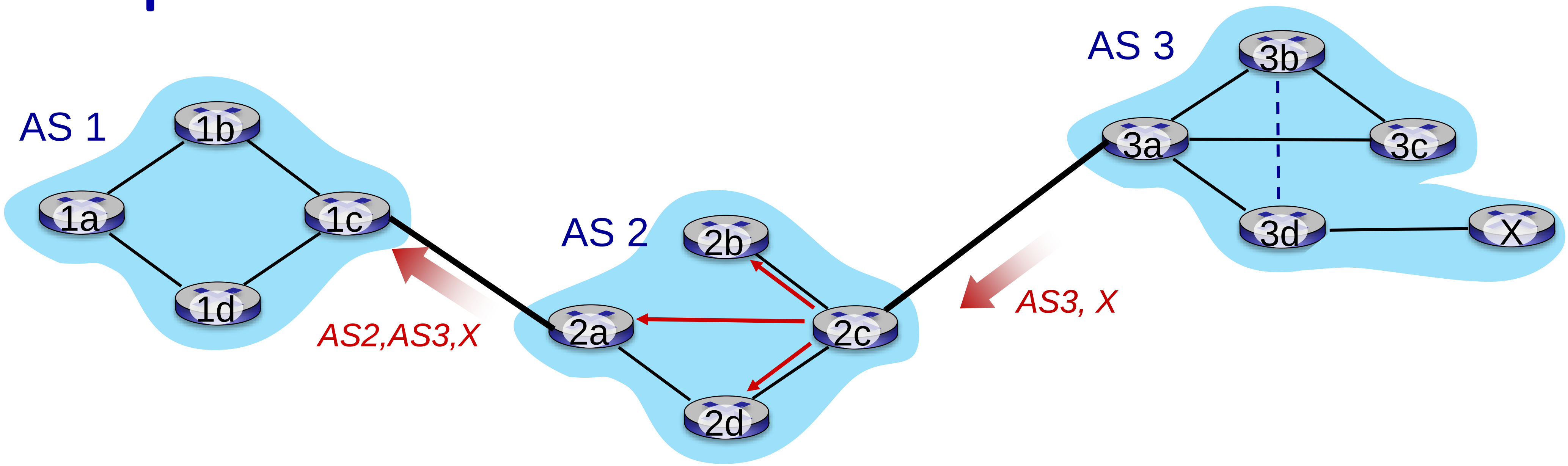


150.10.0.0
RTA
150.10.30.1
IBGP
RTB
150.10.50.1
170.10.20.1
AS 100
170.10.20.2
RTC
170.10.0.0
AS 300

Policy-based Routing

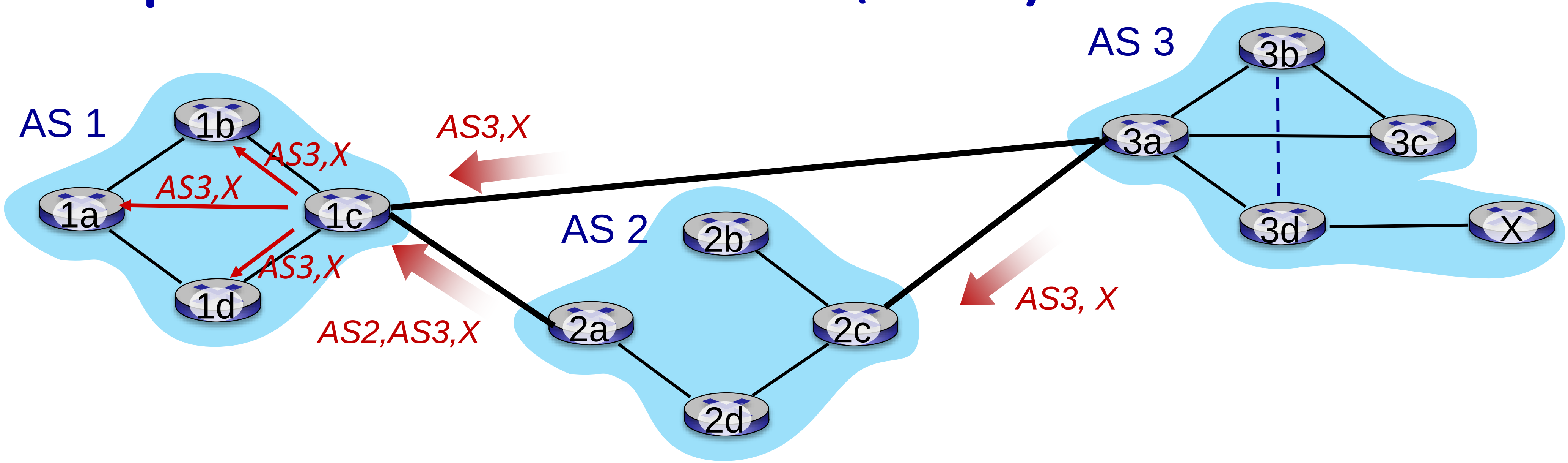
- gateway receiving route advertisement uses *import policy* to accept/decline path (e.g., never route through AS Y).
- AS policy also determines whether to *advertise* path to other neighboring ASes

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

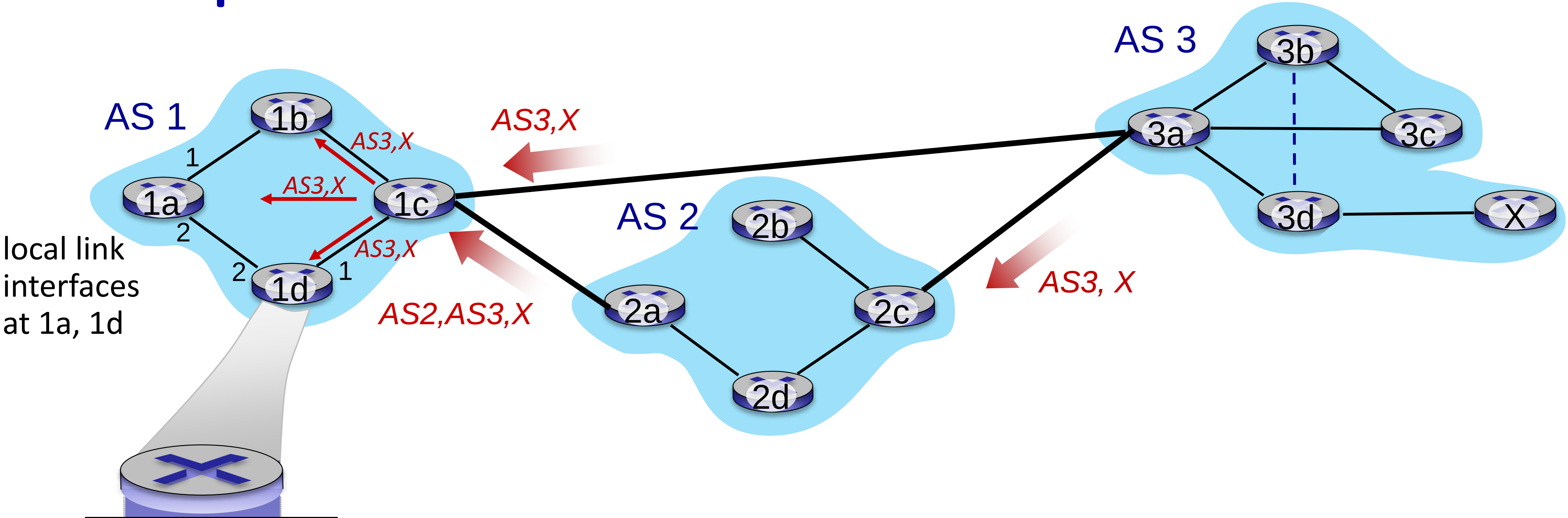
BGP path advertisement (more)



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

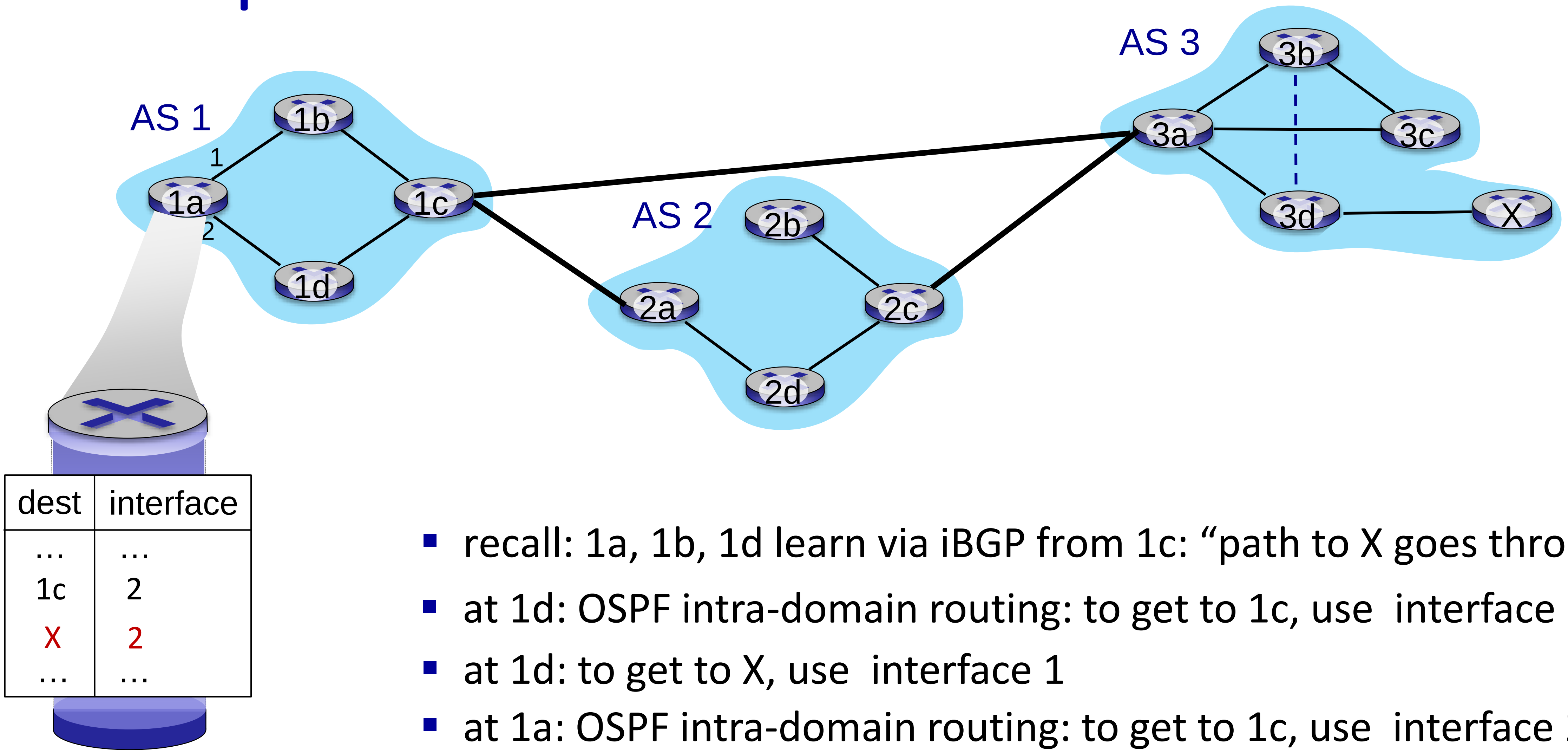
BGP path advertisement



dest	interface
…	…
1c	1
X	1
…	…

- recall: 1a, 1b, 1d learn via iBGP from 1c: “path to X goes through 1c”
- at 1d: OSPF intra-domain routing: to get to 1c, use interface 1
- at 1d: to get to X, use interface 1

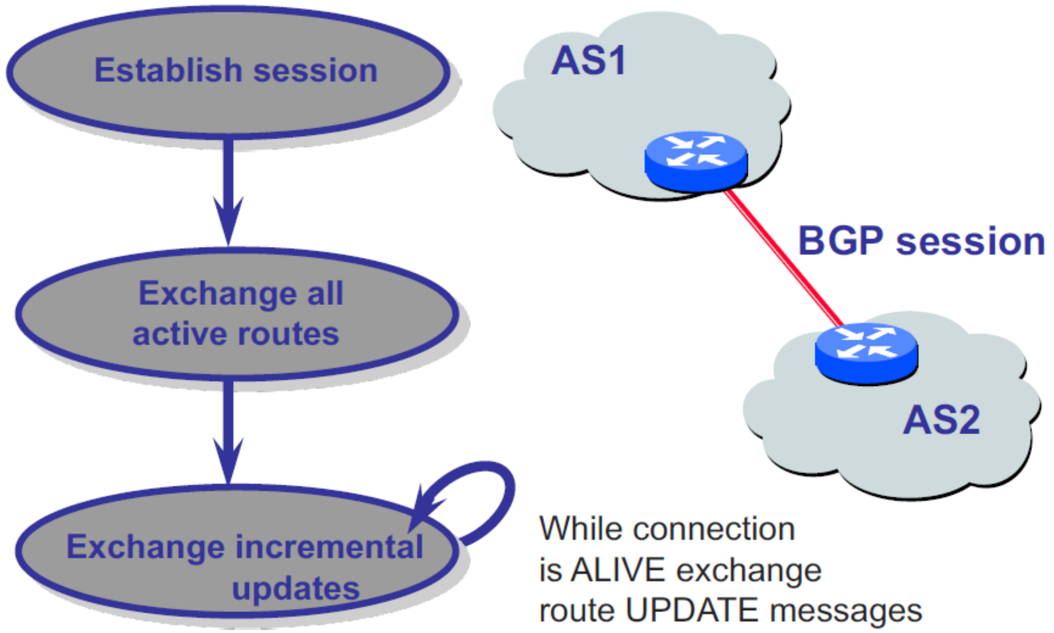
BGP path advertisement



dest	interface
…	…
1c	2
X	2
…	…

- recall: 1a, 1b, 1d learn via iBGP from 1c: “path to X goes through 1c”
- at 1d: OSPF intra-domain routing: to get to 1c, use interface 1
- at 1d: to get to X, use interface 1
- at 1a: OSPF intra-domain routing: to get to 1c, use interface 2
- at 1a: to get to X, use interface 2

Route establishment and maintenance



BGP messages

- 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

Why different Intra-, Inter-AS routing ?

policy:

- inter-AS: admin wants control over how its traffic routed, who routes through its network
- intra-AS: single admin, so policy less of an issue

scale:

- hierarchical routing saves table size, reduced update traffic

performance:

- intra-AS: can focus on performance
- inter-AS: policy dominates over performance

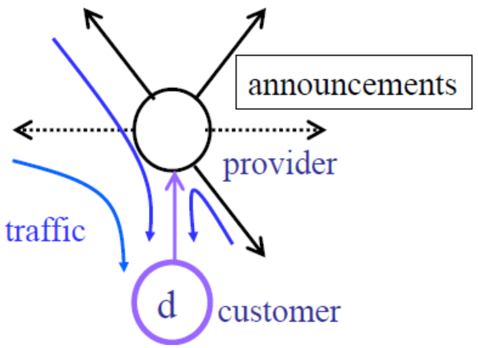
Business Relationships

- Neighboring ASes have business contracts
  - How much traffic to carry
  - Which destinations to reach
  - How much money to pay
- Common business relationships
  - Customer-provider
    - E.g., Princeton is a customer of USLEC
    - E.g., MIT is a customer of Level3
  - Peer-peer
    - E.g., UUNET is a peer of Sprint
    - E.g., Harvard is a peer of Harvard Business School

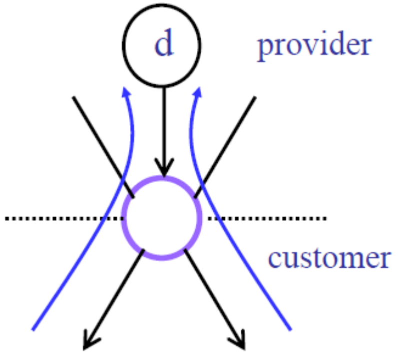
Customer/Provider

- Customer needs to be reachable from everyone
  - Provider tells all neighbors how to reach the customer
- Customer needs to be able to reach to everyone
  - Provider tells the customer how to reach to others

Traffic **to** the customer

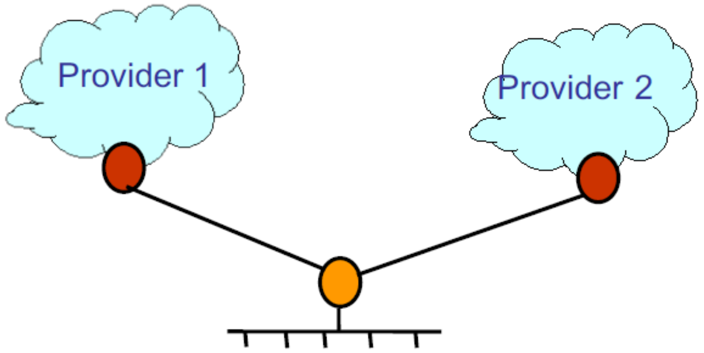


Traffic **from** the customer



Multi-Homing

- Customers may have more than one provider
  - Extra reliability, survive single ISP failure
  - Financial leverage through competition
  - Better performance by selecting better path
  - Gaming the 95_{th}-percentile billing model
- Customer does not want to provide transit service
  - Customer does not let its providers route through it



Export Policies

- Provider to Customer
  - All routes so as to provide transit service
- Customer to Provider
  - Only customer routes
  - Why?
  - Only transit for those that pay
- Peer to Peer
  - Only customer routes

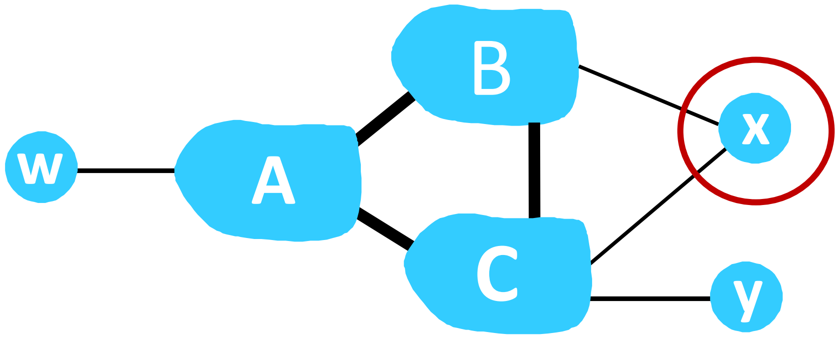
Import Policies

- Same routes heard from providers, customers, and peers, whom to choose?
  - customer > peer > provider
  - Why?
    - Choose the most economic routes!
    - Customer route: charge $$ J
    - Peer route: free
    - Provider route: pay $$ L

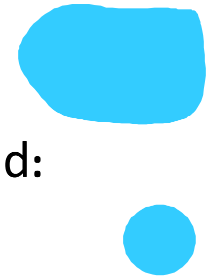
BGP route selection

- 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

BGP: achieving policy via advertisements (more)



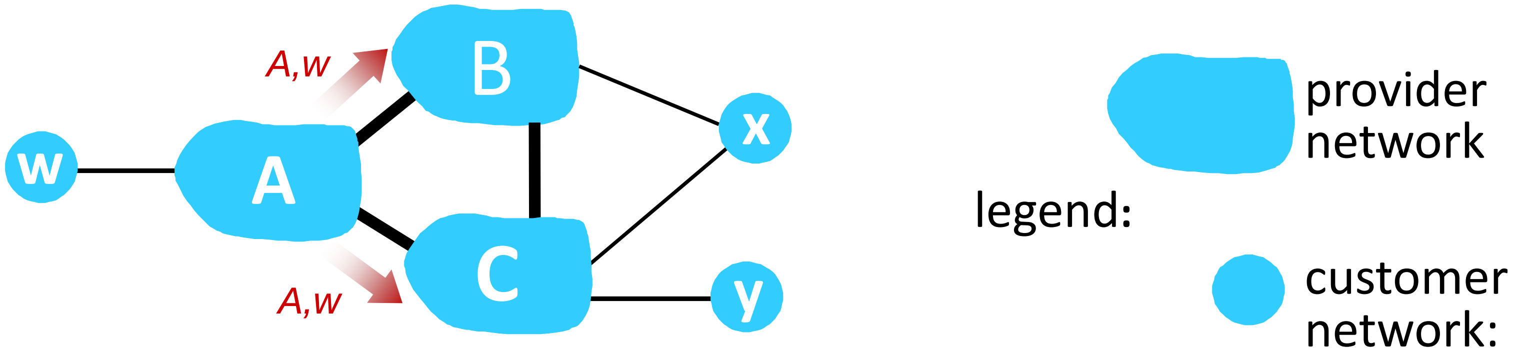
legend:



ISP only wants to route traffic to/from its customer networks (does not want to carry transit traffic between other ISPs – a typical "real world" policy)

- 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



ISP only wants to route traffic to/from its customer networks (does not want to carry transit traffic between other ISPs – a typical "real world" policy)

- 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

Local Preferences

- Local to an AS – non-transitive

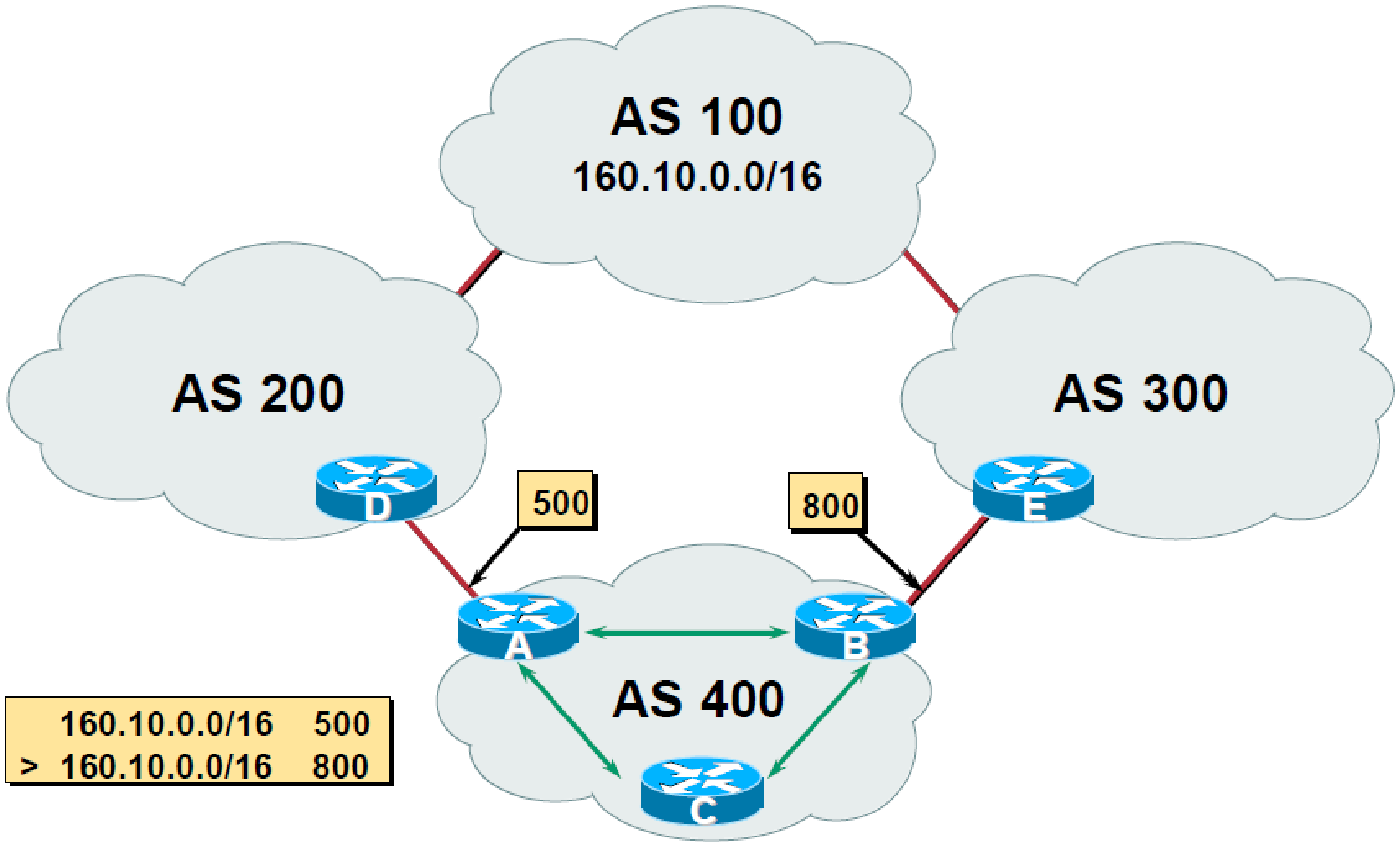
  Default local preference is 100 (IOS)

- Used to influence BGP path selection

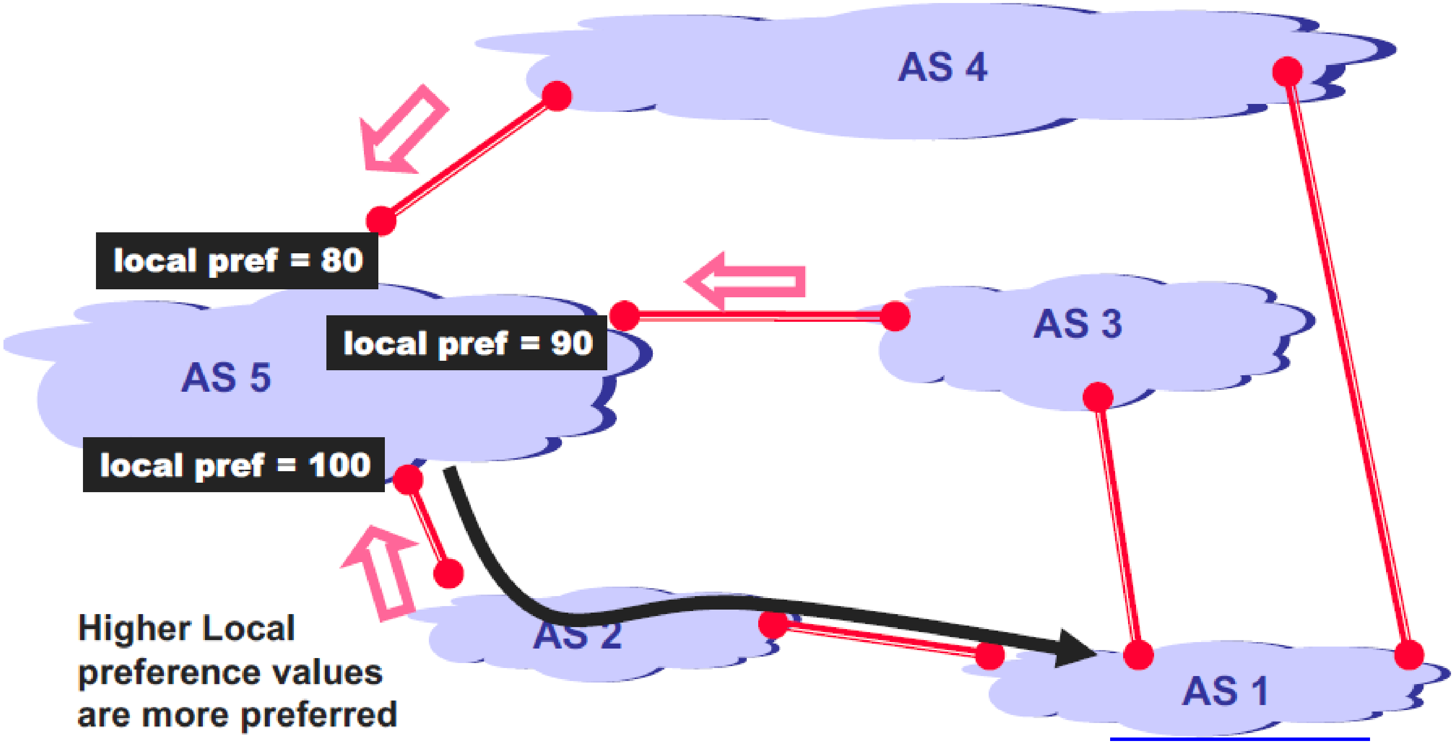
  determines best path for *outbound* traffic

- Path with highest local preference wins

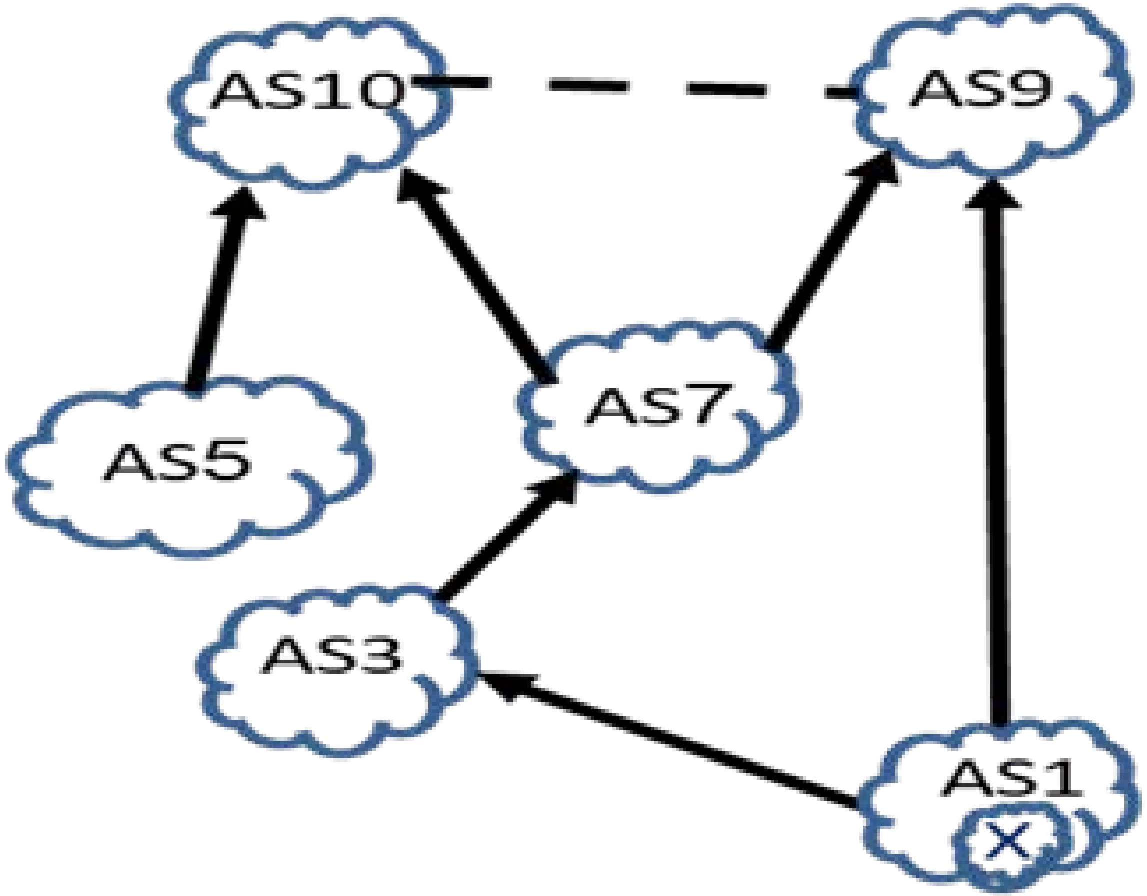
Local Preferences



Local preferences



Example

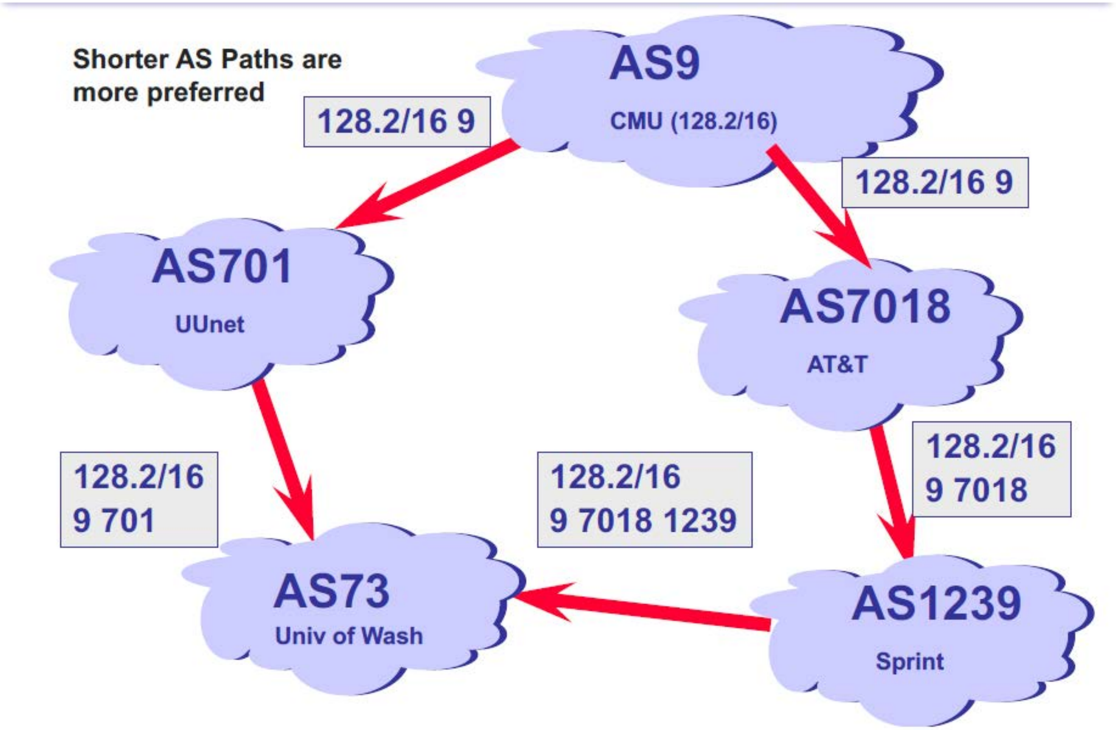


در شبکه شکل روبرو خطوط پر نشان دهنده ارتباط Customer-Provider و خط چین نشان دهنده ارتباط Peer-Peer است.

AS5 چه مسیر (یا مسیرهایی) را برای سابنت x دریافت خواهد کرد؟

کدام مسیر را انتخاب خواهد کرد؟

Shorter AS path selection



Select best BGP route to prefix

- Router selects route based on shortest AS-PATH

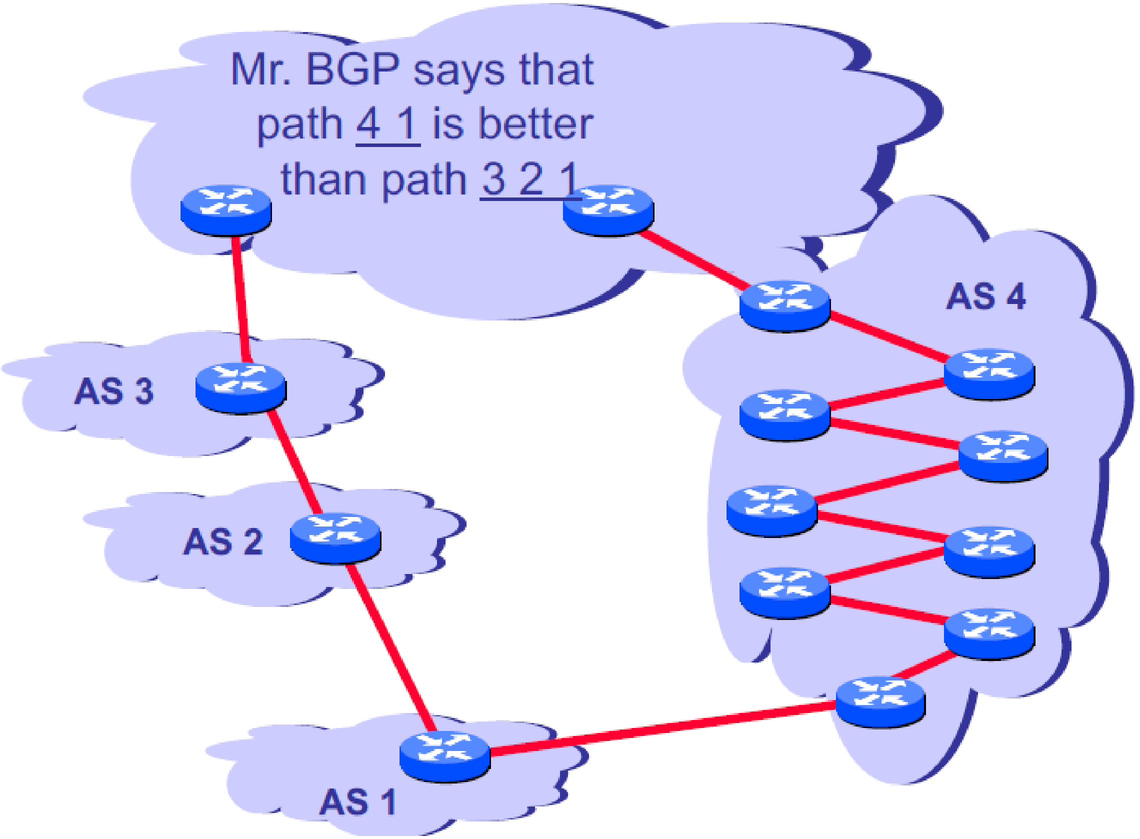
- Example:

  select

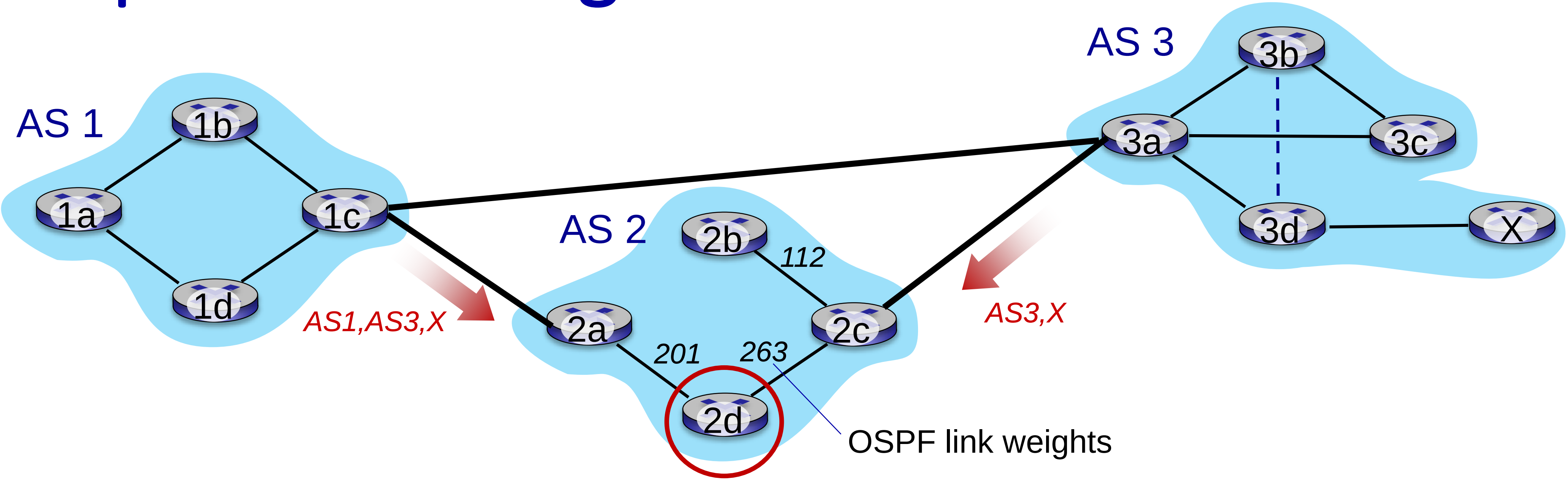
  - AS2 AS17 to 138.16.64/22
  - AS3 AS131 AS201 to 138.16.64/22

- What if there is a tie?

Shorter AS path vs shorter route

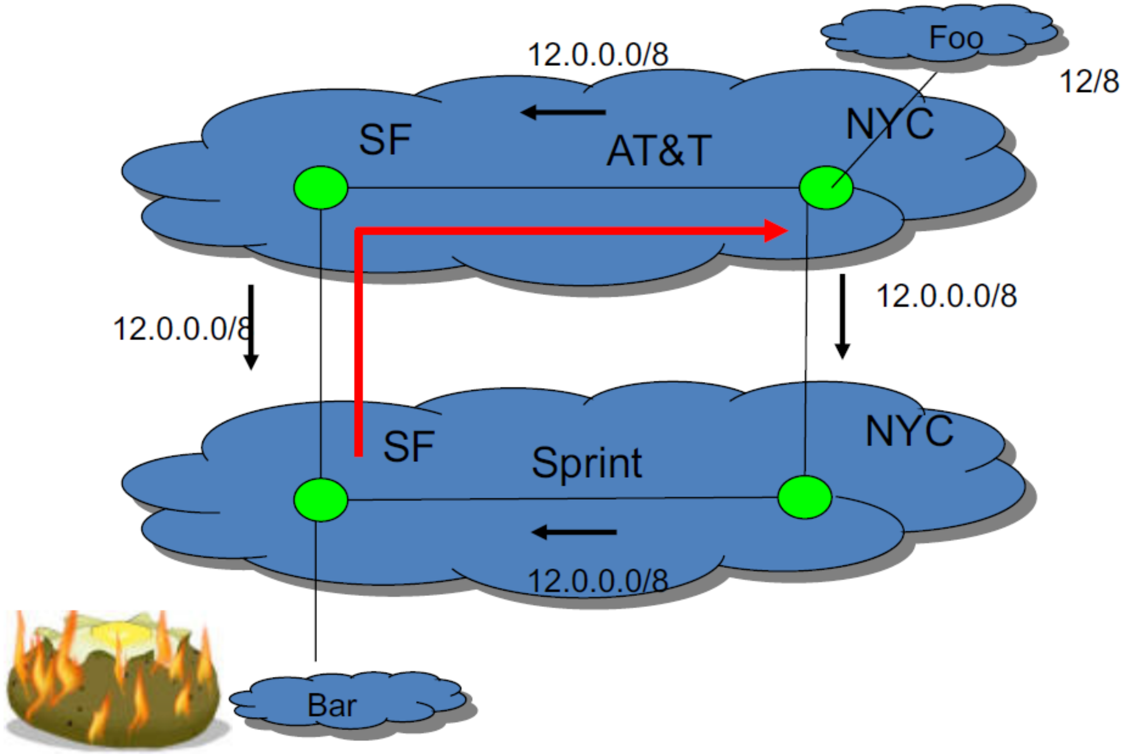


Hot potato routing

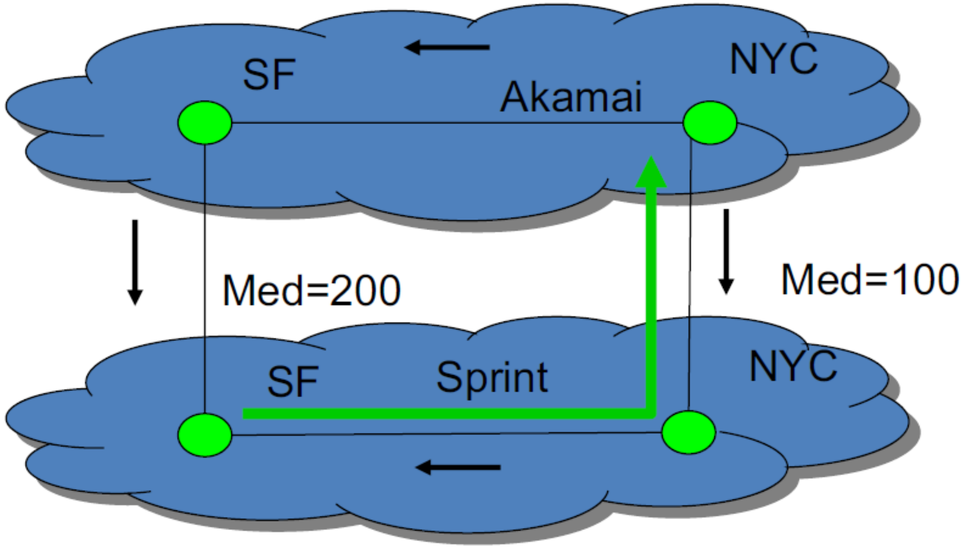


- 2d learns (via iBGP) it can route to X via 2a or 2c
- hot potato routing: choose local gateway that has least *intra-domain* cost (e.g., 2d chooses 2a, even though more AS hops to *X*): don't worry about inter-domain cost!

Hot potato routing

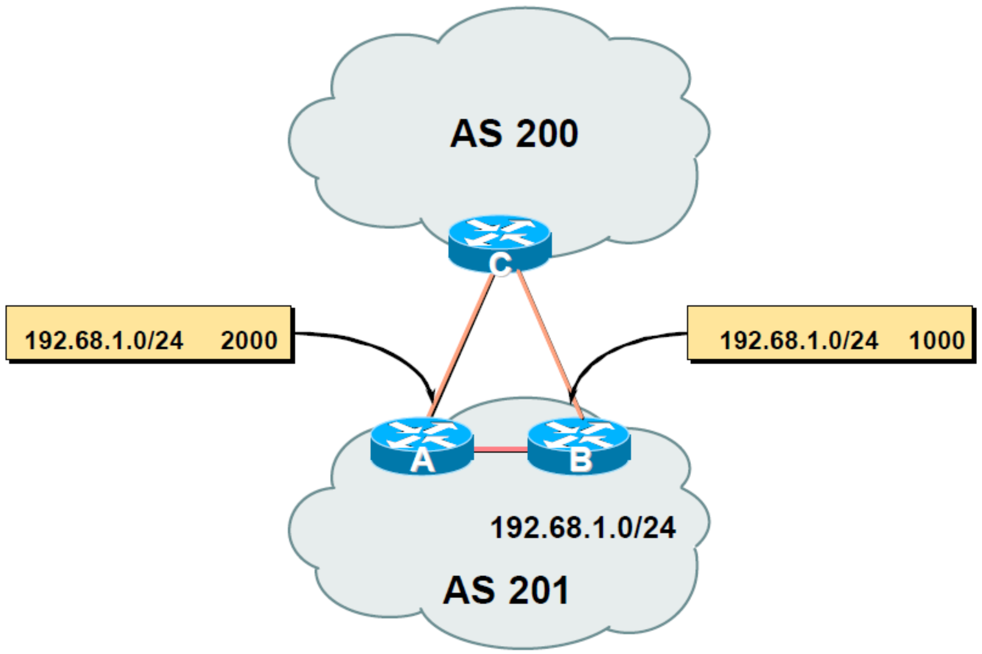


Cold potato routing



MED: Multi Exit Discriminator

MED: Multi-Exit discriminator



MED

- Inter-AS – non-transitive
- Used to convey the relative preference of entry points

  determines best path for *inbound* traffic
- Comparable if paths are from same AS
- IGP metric can be conveyed as MED

  set metric-type internal in route-map

Routing Protocols

- IGP:
  - Intra-AS routing protocols
    - OSPF
      - Dijkstra
- EGP:
  - Inter-AS routing protocol
    - BGP-4
      - eBGP
      - iBGP