Computer Networks

Wenzhong Li, Chen Tian

Nanjing University

Material with thanks to James F. Kurose, Mosharaf Chowdhury, and other colleagues.

Chapter 3. Network Layer

- Network Layer Functions
- IP Protocol Basic
- IP Protocol Suit
- Routing Fundamentals
- Internet Routing Protocols
- IP Multicasting

Chapter 3. Network Layer

- Internet Routing
- Intra-AS protocol: RIP and OSPF
- Inter-AS protocol: BGP

Internet Routing

Internet Routing

- Our routing study thus far – idealization
  - All routers identical, network "flat"
  - Not true in practice
- Scale: with 200 million destinations
  - Cannot store all destinations in routing tables
  - Routing table exchange would swamp links
- Administrative autonomy
  - Internet = network of networks
  - Each network admin may want to control routing in its own networks

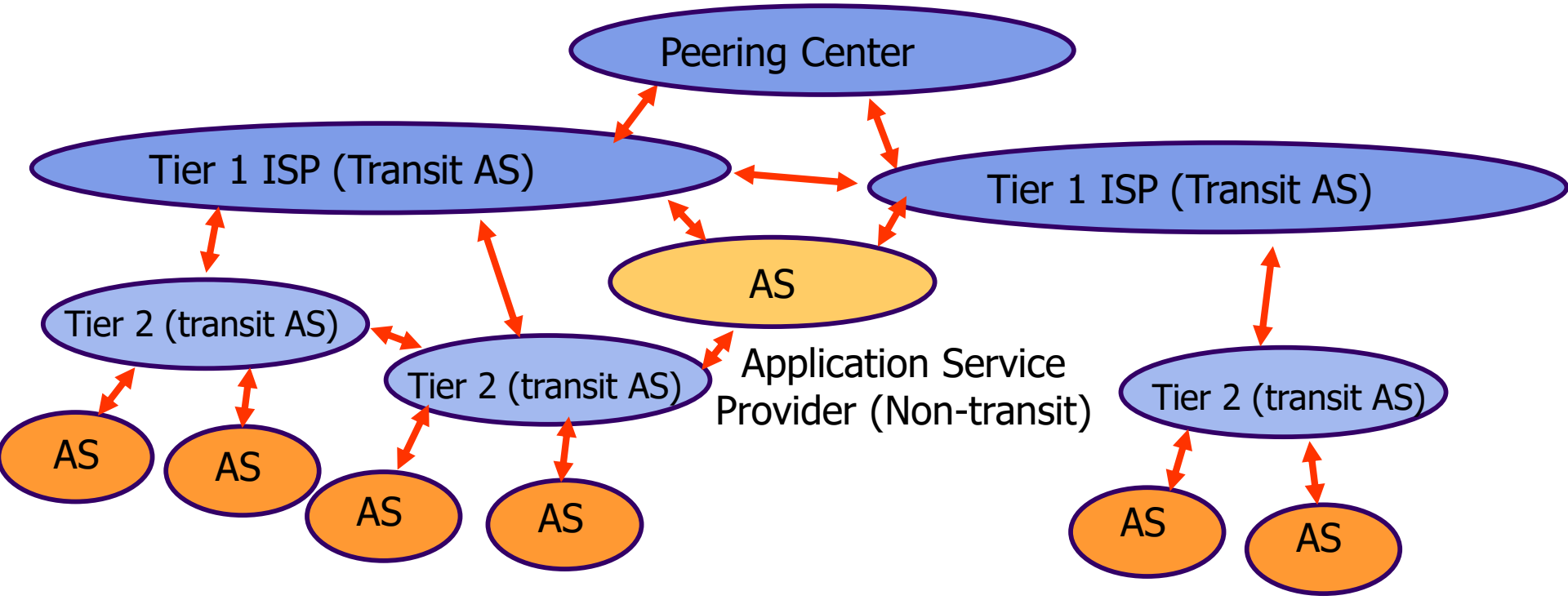
Hierarchical Routing

- Aggregate routers into regions, i.e. autonomous systems (AS)
- Routers in same *AS* run same routing protocol
  - Intra-AS routing protocol
  - Routers in different *AS* can run different intra-AS routing protocol
- Gateway routers
  - Routers in *AS* responsible for routing to destinations outside *AS*
  - Run inter-AS routing protocol with other gateway routers
  - Run intra-AS routing protocol with routers in *AS*

Autonomous Systems (AS)

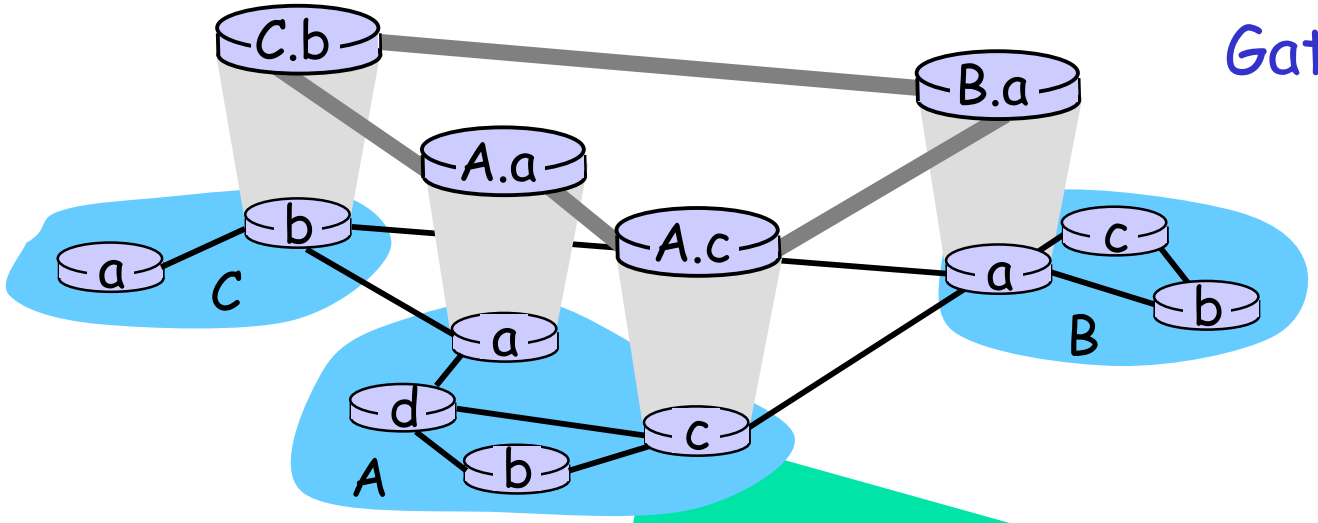
- Set of routers and networks managed by single ISP or large organization
- A connected internets uniquely assigned a 16-bit or 32-bit AS Number
  - There is at least one route between any pair of nodes
- Use common routing protocol

Internet AS-Structure



- Tier 1 ISPs peer with each other, privately & peering centers
- Tier 2 ISPs peer with each other & obtain transit services from Tier 1s
- Non-transit AS's (stub & multi-homed) do not carry transit traffic

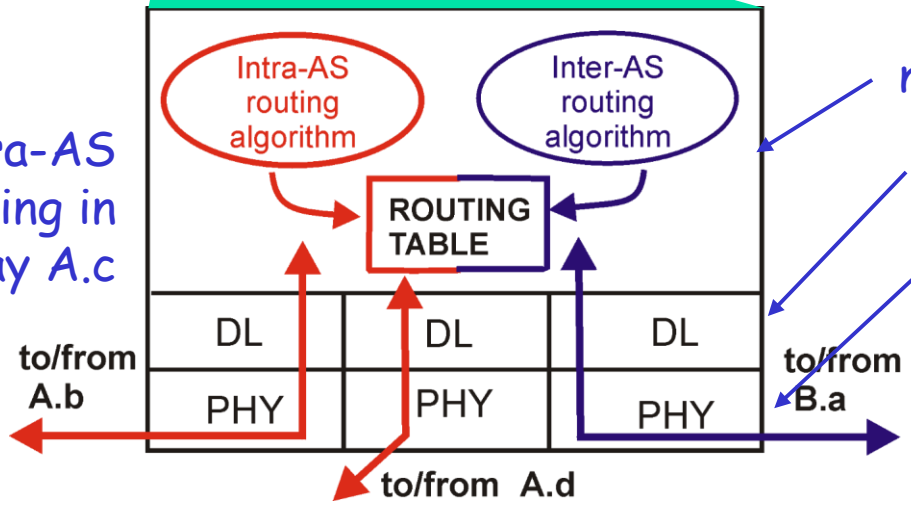
Intra-AS and Inter-AS routing



Gateways:

- Perform inter-AS routing amongst themselves
- Perform intra-AS routing with other routers in their AS

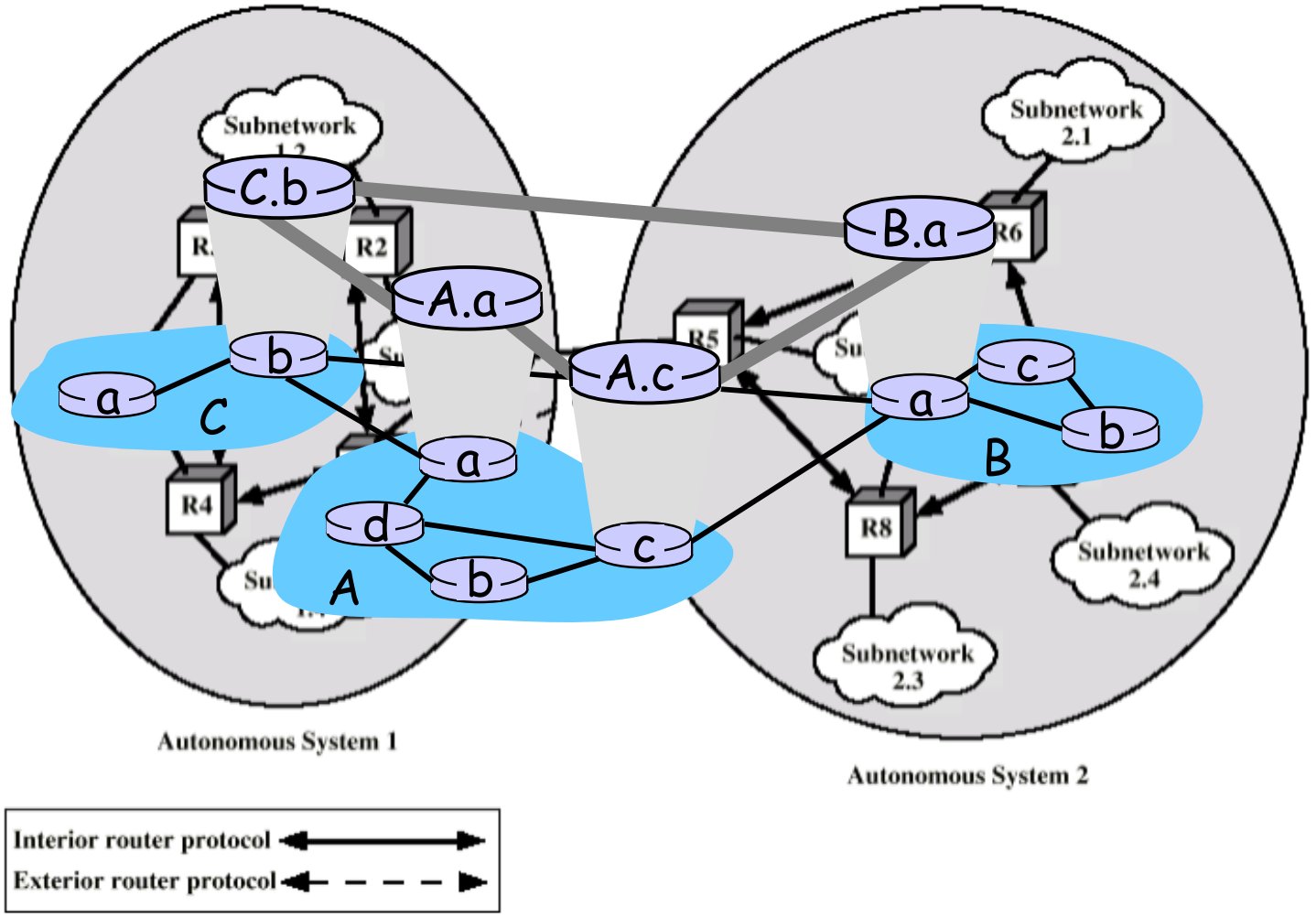
inter-AS, intra-AS routing in gateway A.c



IGP and EGP

- IGP (Interior Gateway Protocol): for Intra-AS routing
  - Passes routing information between routers within *AS*
  - Can focus on performance
  - Routing algorithms and tables may differ between different *AS*
- EGP (Exterior Gateway Protocol): for Inter-AS routing
  - Routers need some info about networks outside their *AS*
  - Supports summary information on reachability
  - Policy may dominate over performance

Application of IGP and EGP



Common Protocols

- IGP – Intra-AS protocols
  - RIP: Routing Information Protocol, use distance vector
  - OSPF: Open Shortest Path First, use link state
  - IGRP: Interior Gateway Routing Protocol (Cisco proprietary)
- EGP – Inter-AS protocols
  - BGP: Border Gateway Protocol

Distance-Vector

- First generation routing algorithm for ARPANET
- Each node (router or host) exchange information with neighboring nodes
  - Neighbors are both directly connected to same network
- Node maintains vector of
  - Link costs for each directly attached network
  - Estimated distance and next-hop vectors for each destination
- DV update messages exchanged between neighbors to build/update routing tables
  - Changes take long time to propagate

Link-State

- Second generation routing algorithm for ARPANET
- When router initialized, it determines link cost on each interface
- Advertises set of link costs to all other routers in topology
  - Not just neighboring routers
- From then on, monitor link costs
  - If significant change, router advertises new set of link costs

Link-State

- Each router can construct topology of entire configuration
  - Can calculate shortest path to each destination network
- Router constructs routing table, listing first hop to each destination
- Router does not use distributed routing algorithm
  - Use any routing algorithm to determine shortest paths
  - In practice, Dijkstra's algorithm

EGP Requirements

- Link-state and distance-vector not effective for exterior gateway protocol
  - Different *ASs* may use different metrics and have different restrictions
  - Not all subnets want or need to be known to all
- Distance-vector
  - Gives no information about *ASs* visited on route
- Link-state
  - Flooding of link state information to all routers unmanageable

EGP – Path-Vector

- The most concern is the *ASs* passed through
  - Dispense with routing metrics
- Each gateway router broadcasts to neighbors entire path to destination
  - Each block of information lists all *ASs* visited on the route
  - Needs not include distance or cost estimate
- Enables gateway router to perform policy routing
  - Avoid path to avoid transiting particular *AS*
  - Minimizing number of transit *ASs*
  - Others, e.g. link speed, net capacity, tendency to become congested, overall quality of operation, and security

BGP and OSPF

- BGP: Border Gateway Protocol
  - The de facto Internet standard used for inter-AS routing
- OSPF: Open Shortest Path First
  - The most-used intra-AS protocol in Internet

RIP and OSPF

Inter and Intra AS Routing

- Exterior Gateway Protocol (EGP): routing between AS's
  - BGPv4
- Interior Gateway Protocol (IGP): routing within AS
  - RIP, OSPF

RIP (Routing Information Protocol)

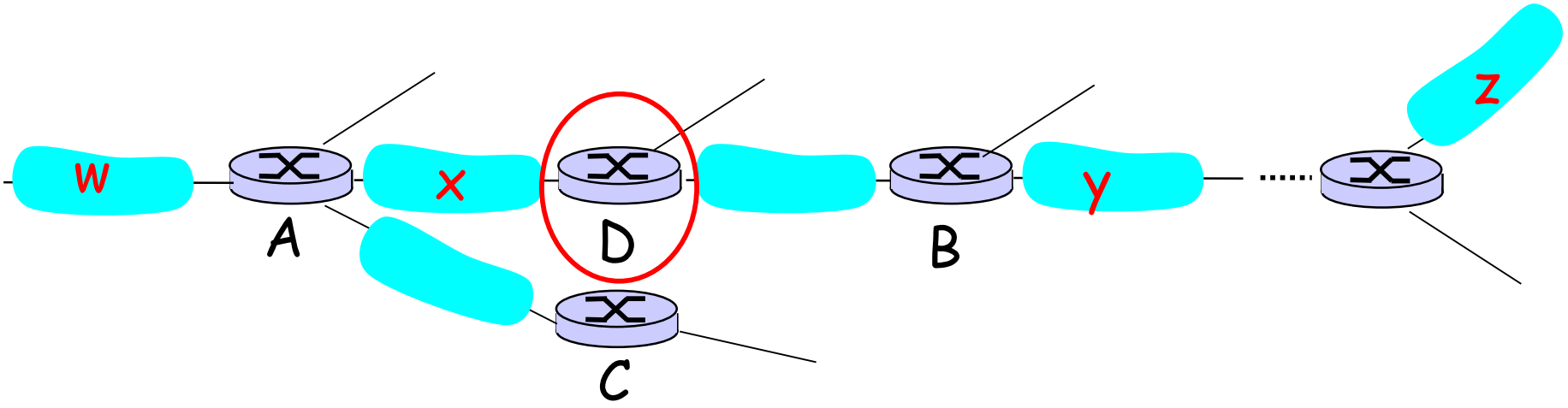
- Use Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982

- Distance metric: # of hops (max=15 hops)
- Distance vectors: exchanged among neighbors every 30 *sec* via RIP update message

- Fail to receive the update message within 180 *sec* means the link to the neighbor is lost

- Each advertisement: list of up to 25 destination nets
- Advertisements sent in UDP packets

RIP: Example (1)



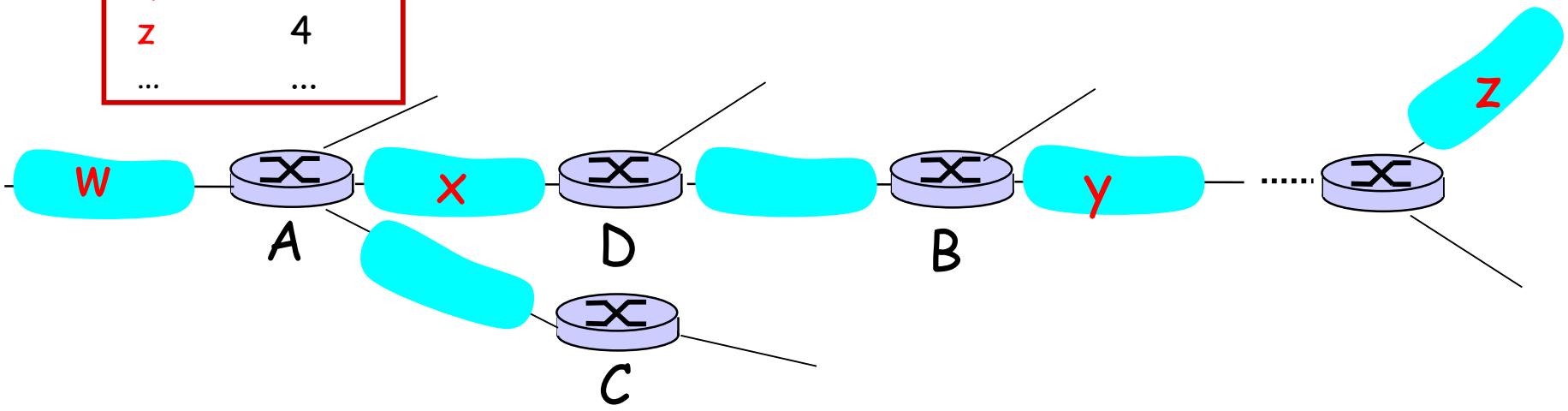
Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	B	7
x	--	1
....	....	....

Routing table in D

RIP: Example (2)

Dest	hops
w	1
x	1
z	4
…	…

Advertisement from A to D

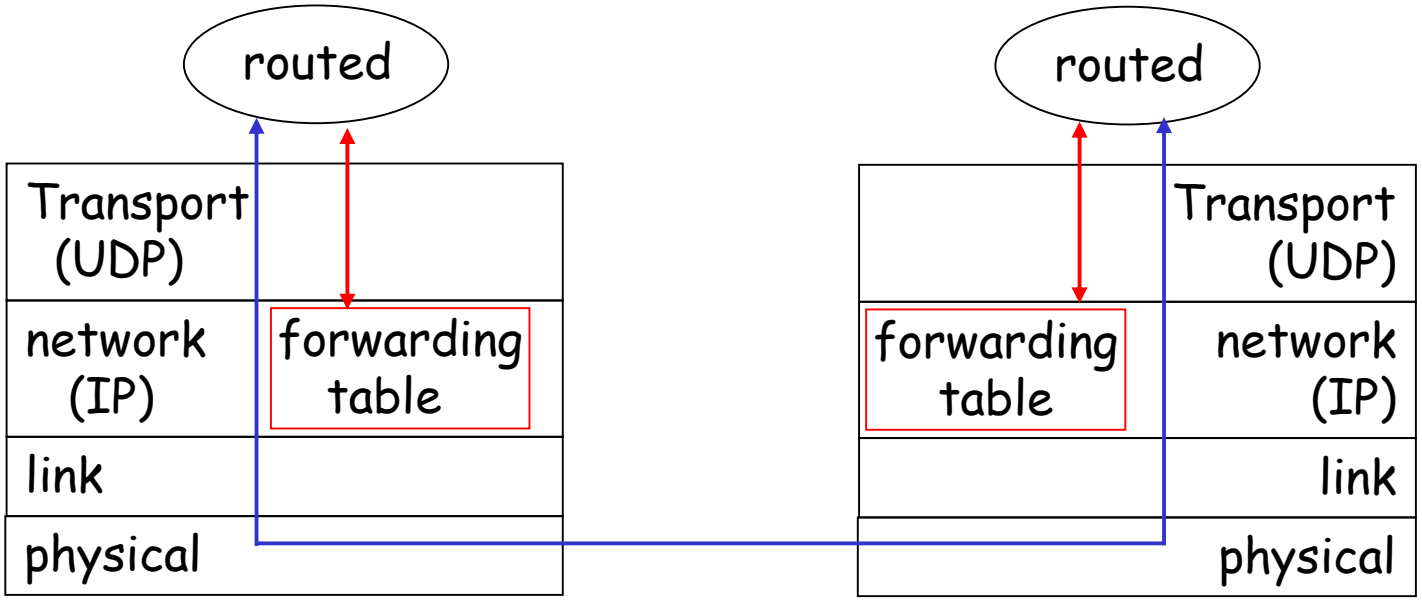


Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	~~B~~ A	~~7~~ 5
x	--	1
….	….	….

Routing table in D (update based on A's advertisement)

RIP Table Processing

- Later queue length is used for link cost, instead of just hops
- RIP routing tables managed by application-level process called routed (daemon)
- Advertisements sent in UDP packets, periodically repeated



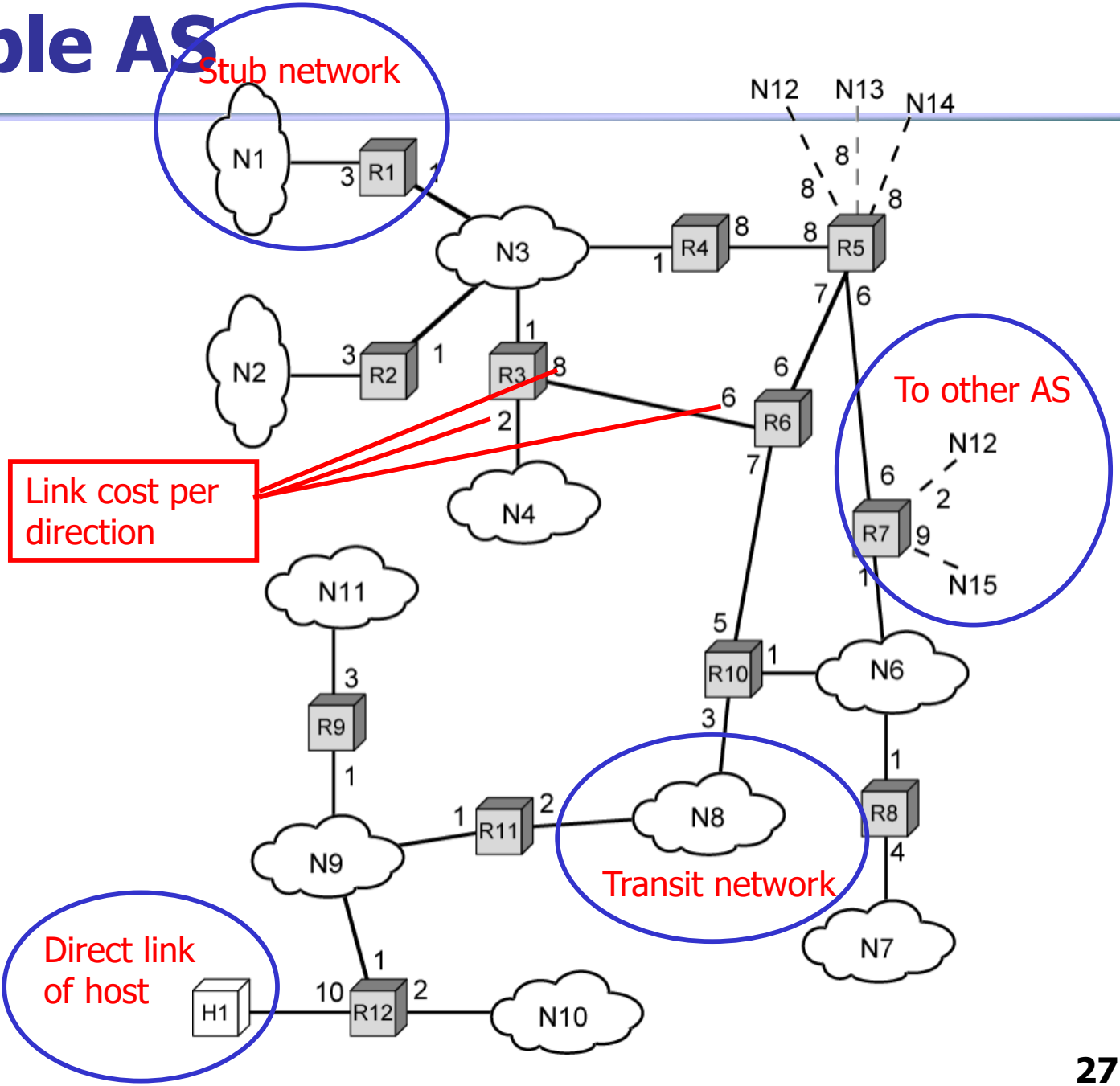
Open Shortest Path First (1)

- OSPF (RFC 2328), replaced Routing Information Protocol (RIP)
- Uses Link-State routing algorithm
  - Each router keeps list of state of local links to neighbor routers
  - Transmits update state info (advertisement) to entire *AS* via flooding per 10*s*
  - Carried in *OSPF* messages directly over IP, Not UDP
- Uses cost metric assigned on each link

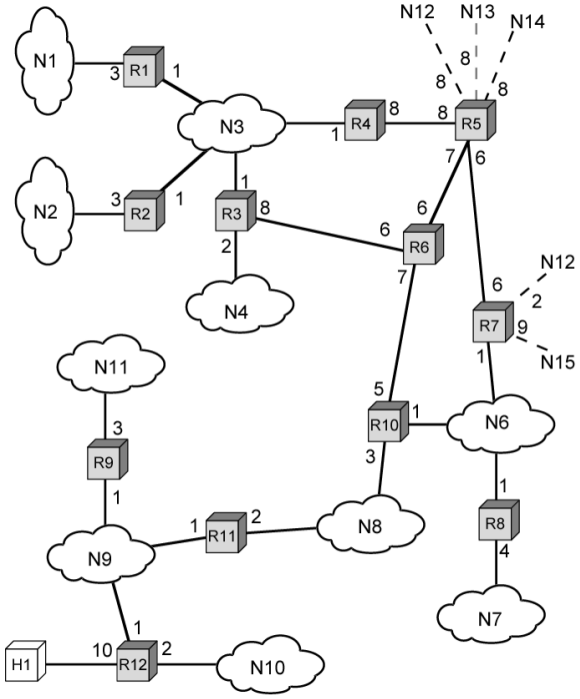
Open Shortest Path First (2)

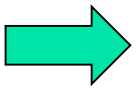
- Topology map stored as directed graph on each node
- Router nodes
- Network nodes: (Transit vs. Stub)
- Edges: router—router, router—network
- Dijkstra's algorithm used to compute the shortest path to each destination

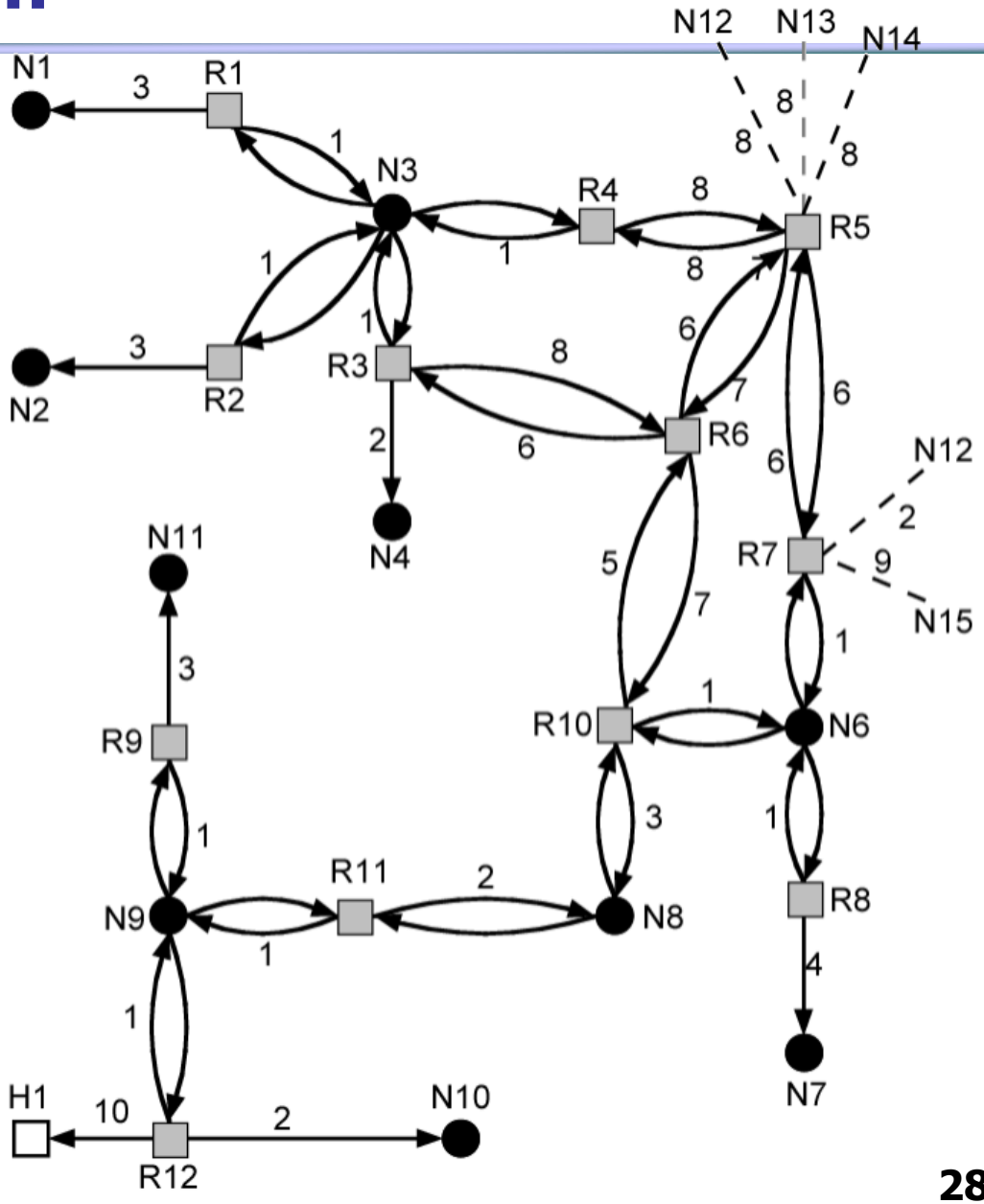
Sample AS


Stub network
N1
R1
N3
R4
R5
N12
N13
N14
N2
R2
R3
R6
To other AS
N12
R7
N15
Link cost per direction
N4
N11
R10
N6
R9
N8
R8
N9
R11
Transit network
N7
Direct link of host
H1
R12
N10

The Directed Graph







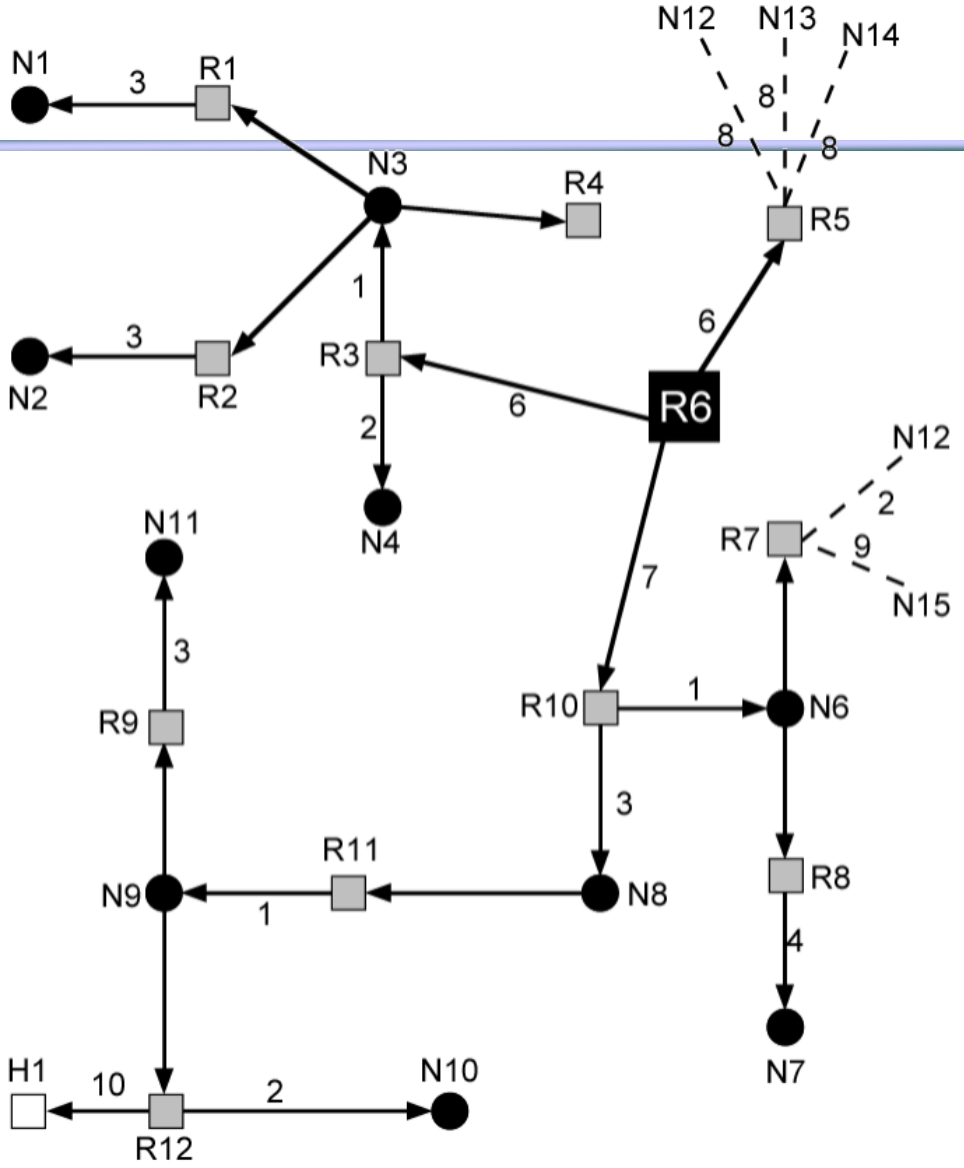
SPF Operation

- Networks, hosts and BGP routers as destinations
- Each router compute its SPF tree showing the least cost path to all other destination
- Only next hop used in routing packets

SPF Tree for Router 6

Destination:

Network
Direct Host
Border Router



Routing Tables of Router 6

Destination	Next Hop	Distance	Destination	Next Hop	Distance
N1	R3	10	N11	R10	14
N2	R3		H1		21
N3	R3		R5		6
N4	R3		R7		8
N6	R10		N12		10
N7	R10		N13		14
N8	R10		N14		14
N9	R10	11	N15	R10	17
N10	R10	13			

Routing Tables of Router 6

Destination	Next Hop	Distance	Destination	Next Hop	Distance
N1	R3	10	N11	R10	14
N2	R3	10	H1	R10	21
N3	R3	7	R5	R5	6
N4	R3	8	R7	R10	8
N6	R10	8	N12	R10	10
N7	R10	12	N13	R5	14
N8	R10	10	N14	R5	14
N9	R10	11	N15	R10	17
N10	R10	13			

OSPF Advanced Features

- Security: all *OSPF* messages authenticated to prevent malicious intrusion

- Multiple same-cost paths allowed
- For each link, multiple cost metrics for different TOS (type of service)
  - e.g. satellite link cost set "low" for best effort; "high" for real time

- Integrated uni- and multicast support
  - Multicast *OSPF* (*MOSPF*) uses same topology database as *OSPF*

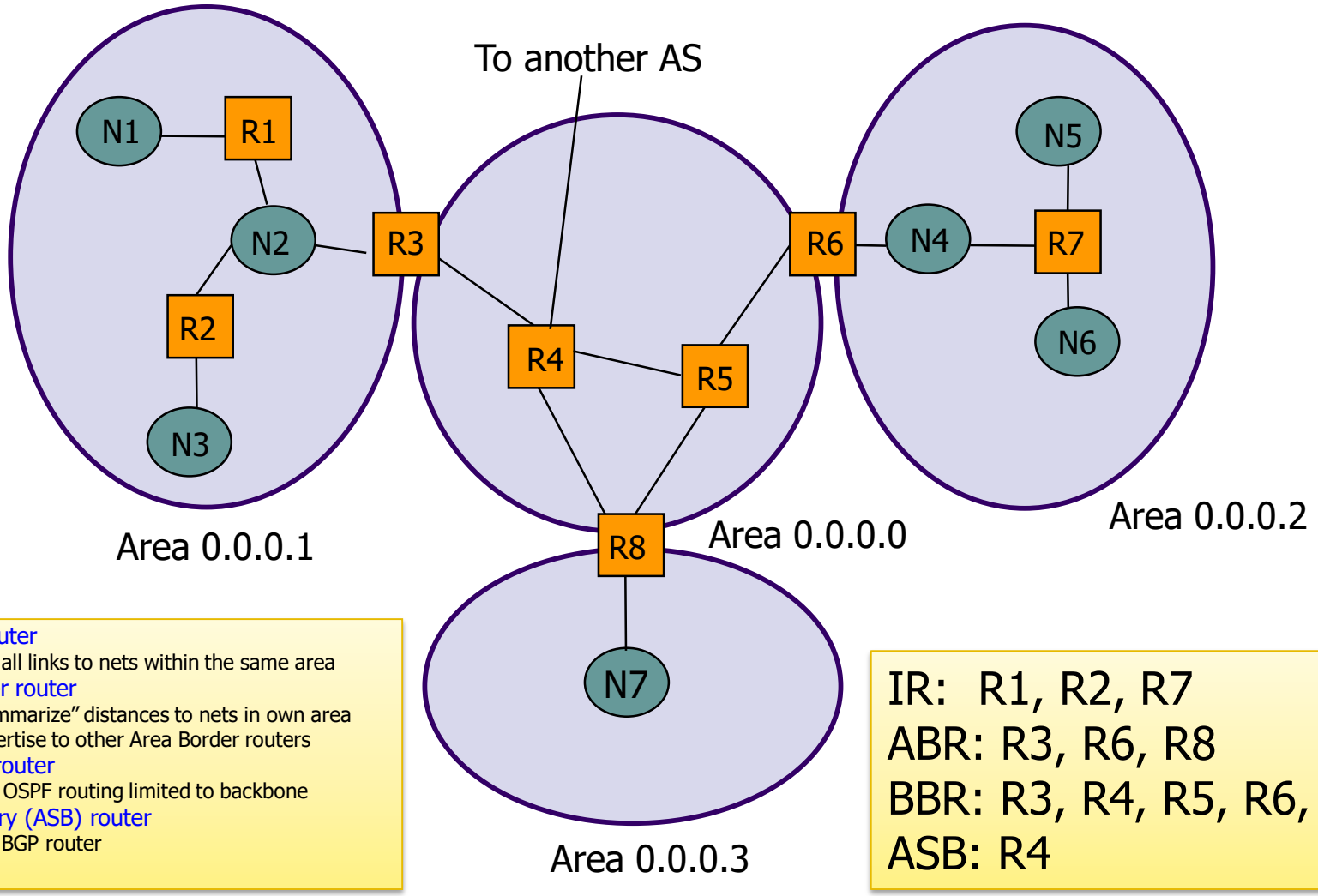
- Hierarchical OSPF in large domains

Hierarchical OSPF

- To improve scalability, AS may be partitioned into areas
  - Area is identified by 32-bit Area ID
  - Router in area only knows complete topology inside area
  - Limits the flooding of link-state information to other area
  - Area border routers summarize info from other areas
- Each area must be connected to backbone area (0.0.0.0)
  - Distributes routing info between areas

- 划分区域的好处是将洪泛交换链路状态信息的范围局限于每一个区域而不是整个的自治系统。
- 在一个区域内部的路由器只知道本区域的完整网络拓扑，而不知道其他区域的网络拓扑的情况。
- 主干区域用于连通其他下层区域

OSPF Areas



- Internal router
  - Has all links to nets within the same area
- Area border router
  - "Summarize" distances to nets in own area
  - Advertise to other Area Border routers
- Backbone router
  - Run OSPF routing limited to backbone
- AS boundary (ASB) router
  - The BGP router

IR: R1, R2, R7
ABR: R3, R6, R8
BBR: R3, R4, R5, R6, R8
ASB: R4

Link State Advertisements

- Router link ad: generated by all OSPF routers
  - State of router links within area, flooded within area
- Net link ad: generated by the designated router
  - Lists routers connected to net, flooded within area
- Summary link ad: generated by area border routers
  - Routes to destinations in other areas
  - Routes to ASB routers
- AS external link ad: generated by ASB routers
  - Describes routes to destinations outside the OSPF net
  - Flooded in all areas in the OSPF net

RIP vs OSPF

- RIP
  - 配置简单，适用于小型网络（小于15跳）
  - 可分布式实现
  - 收敛速度较慢
  - 网络是一个平面，不适用于大规模网络
- OSPF
  - 收敛速度快，无跳数限制
  - 支持不同服务类型选路
  - 支持身份认证
  - 支持层次式网络，适用于大规模复杂网络
  - 集中式算法
  - 每个节点需要维护全局拓扑
  - 配置复杂

BGP: Border Gateway Protocol

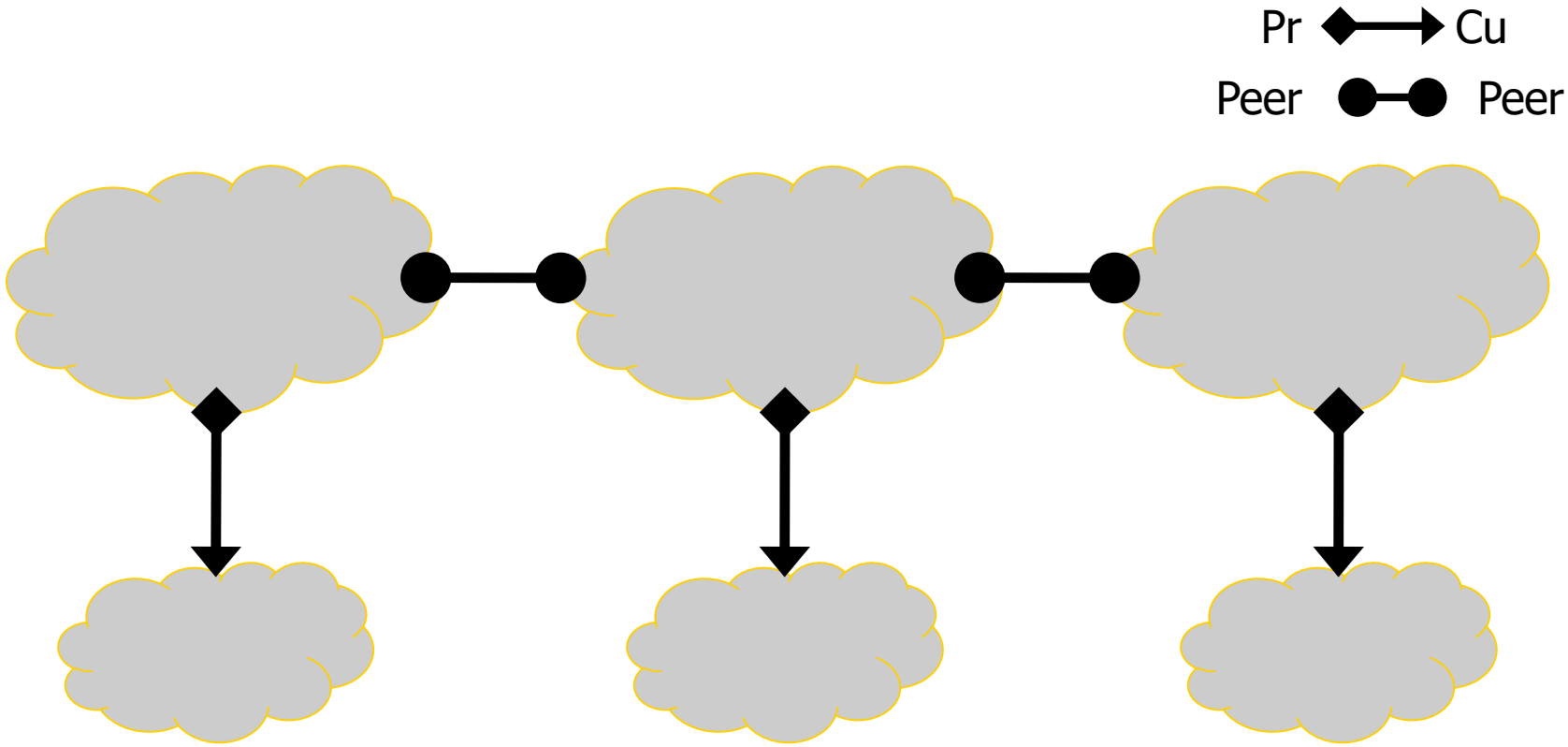
Administrative structure shapes Inter-domain routing

- ASes want freedom to pick routes based on policy
- ASes want autonomy
- ASes want privacy

Topology & policy shaped by inter-AS business relationship

- Three basic kinds of relationships between ASes
  - AS A can be AS B's customer
  - AS A can be AS B's provider
  - AS A can be AS B's peer
- Business implications
  - Customer pays provider
  - Peers don't pay each other
    - Exchange roughly equal traffic

Business relationships



Relations between ASes

provider ⟷ customer

peer ⟷ peer

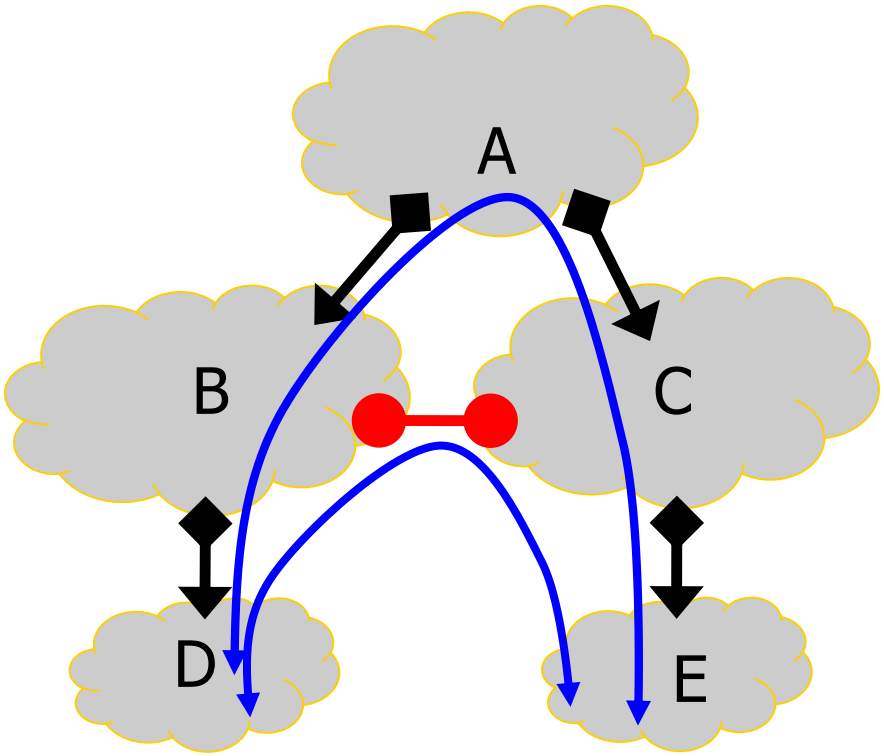
Business implications

- Customers pay provider
- Peers don't pay each other

Why peer?

Pr ◆—▶ Cu

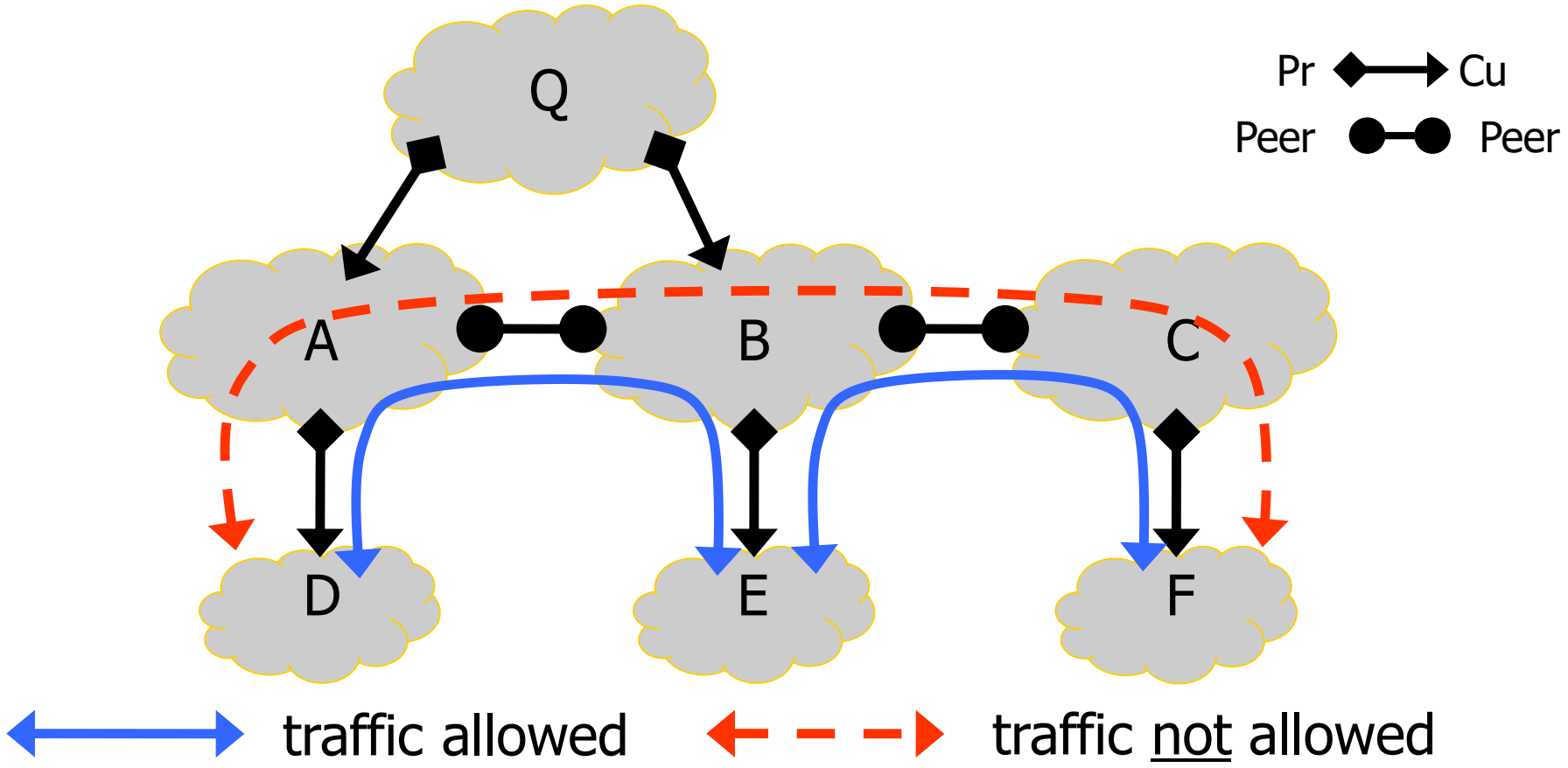
Peer ●—● Peer



D and E communicate a lot

Peering saves B *and* C money

Routing follows the money!



- ASes provide “transit” between their customers
- Peers do not provide transit between other peers

In short

- AS topology reflects business relationships between ASes
- Business relationships between ASes impact which routes are acceptable

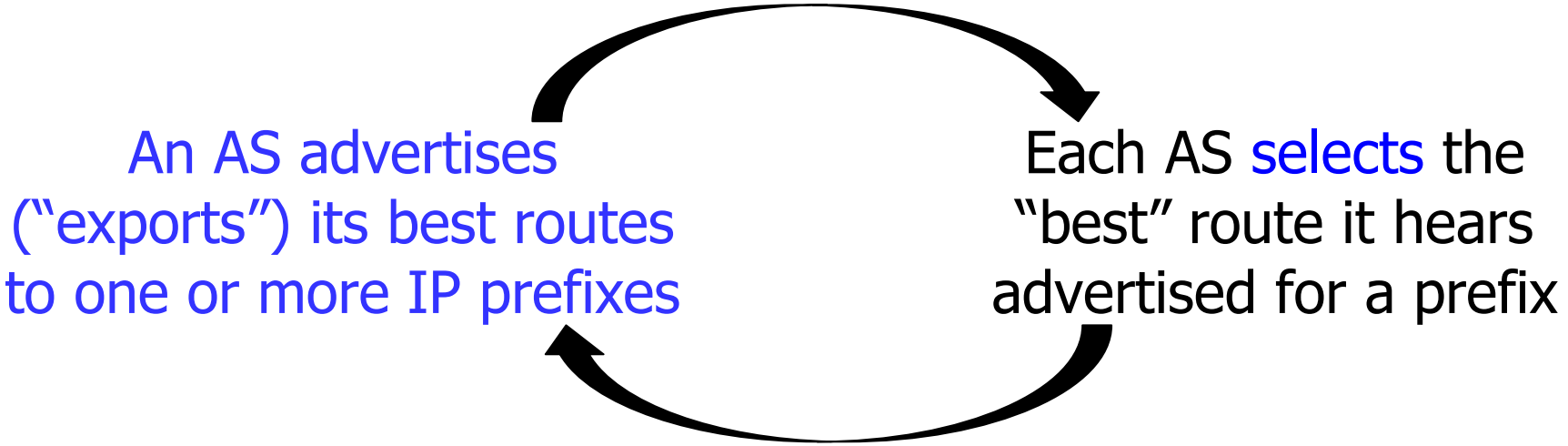
BGP (Today)

- The role of policy
  - What we mean by it
  - Why we need it
- Overall approach
  - Four non-trivial changes to DV

Inter-domain routing: Setup

- Destinations are IP prefixes (12.0.0.0/8)
- Nodes are Autonomous Systems (ASes)
  - Internals of each AS are hidden
- Links represent both physical links and business relationships
- BGP (Border Gateway Protocol) is the Inter-domain routing protocol
  - Implemented by AS border routers

BGP: Basic idea



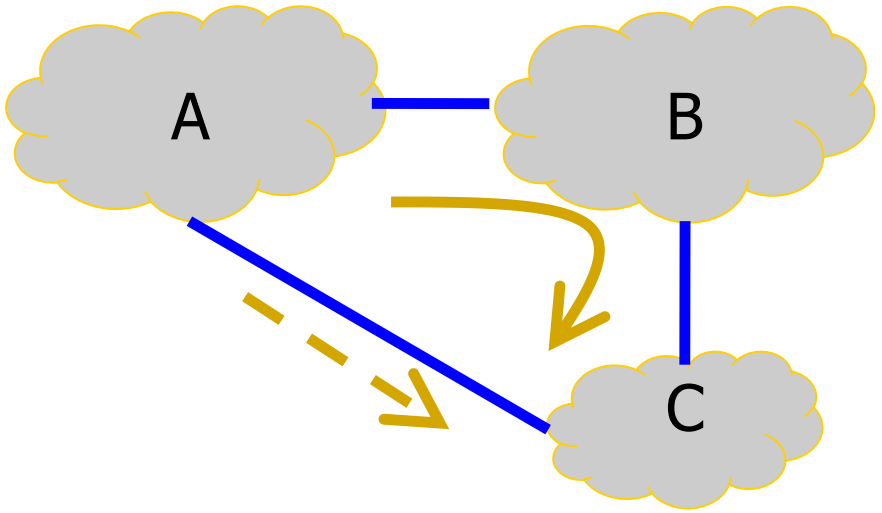
You've heard this story before!

BGP inspired by Distance-Vector

- Per-destination route advertisements
- No global sharing of network topology information
- Iterative and distributed convergence on paths
- With four crucial differences!

BGP & DV differences: (1) Not picking shortest-path routes

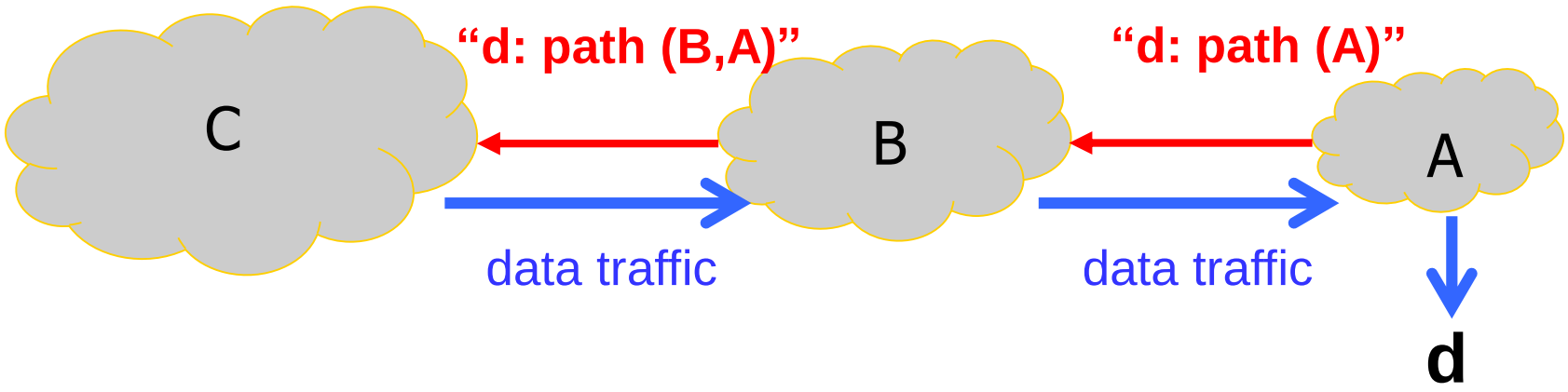
- BGP selects the best route based on policy, not shortest distance (i.e., least-cost)
- AS A may prefer "A,B,C" over "A,C"



- How do we avoid loops?

BGP & DV differences: (2) Path-Vector routing

- Key idea: advertise the entire path
  - Distance vector: send distance metric per dest d
  - Path vector: send the entire path for each dest d

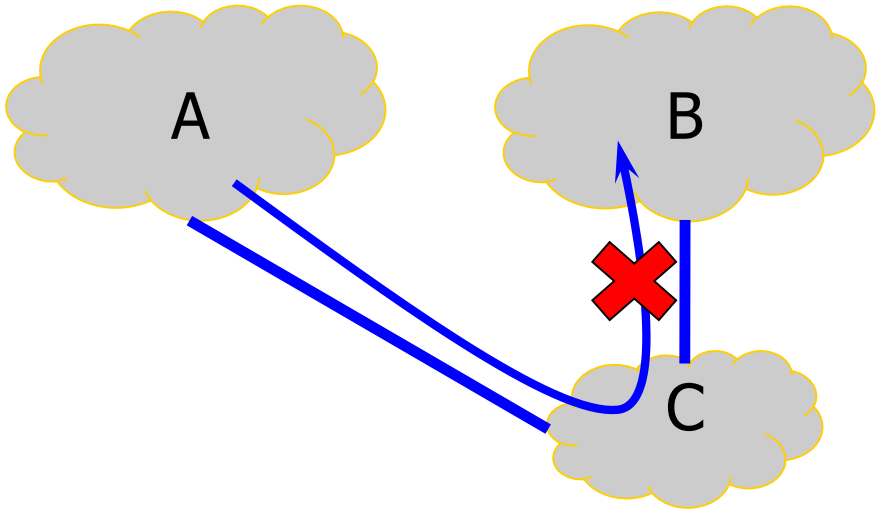


BGP & DV differences: (2) Path-Vector routing

- Key idea: advertise the entire path
  - Distance vector: send distance metric per destination
  - Path vector: send the entire path for each destination
- Benefits
  - Loop avoidance is straightforward (simply discard paths with loops)
  - Flexible and expressive policies based on entire path

BGP & DV differences: (3) Selective route advertisement

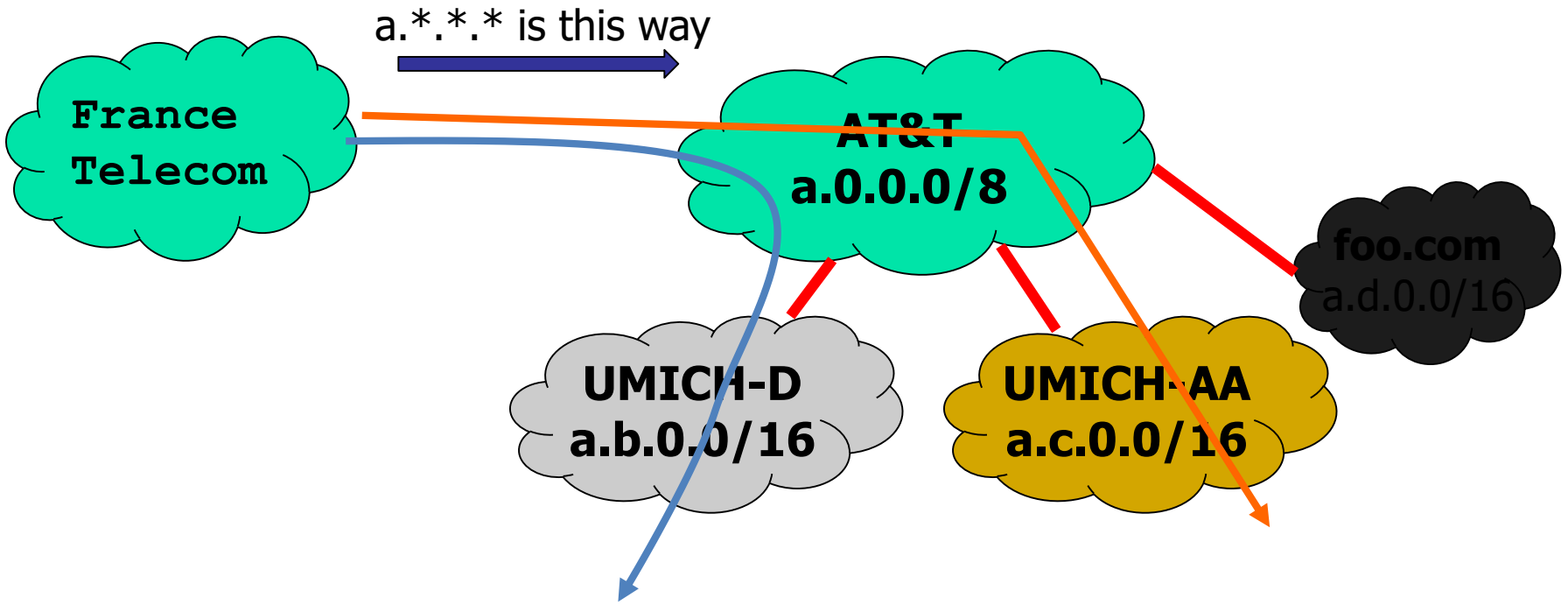
- For policy reasons, an AS may choose not to advertise a route to a destination
- Hence, reachability is not guaranteed even if graph is physically connected



AS-C does not want to carry traffic to AS-B

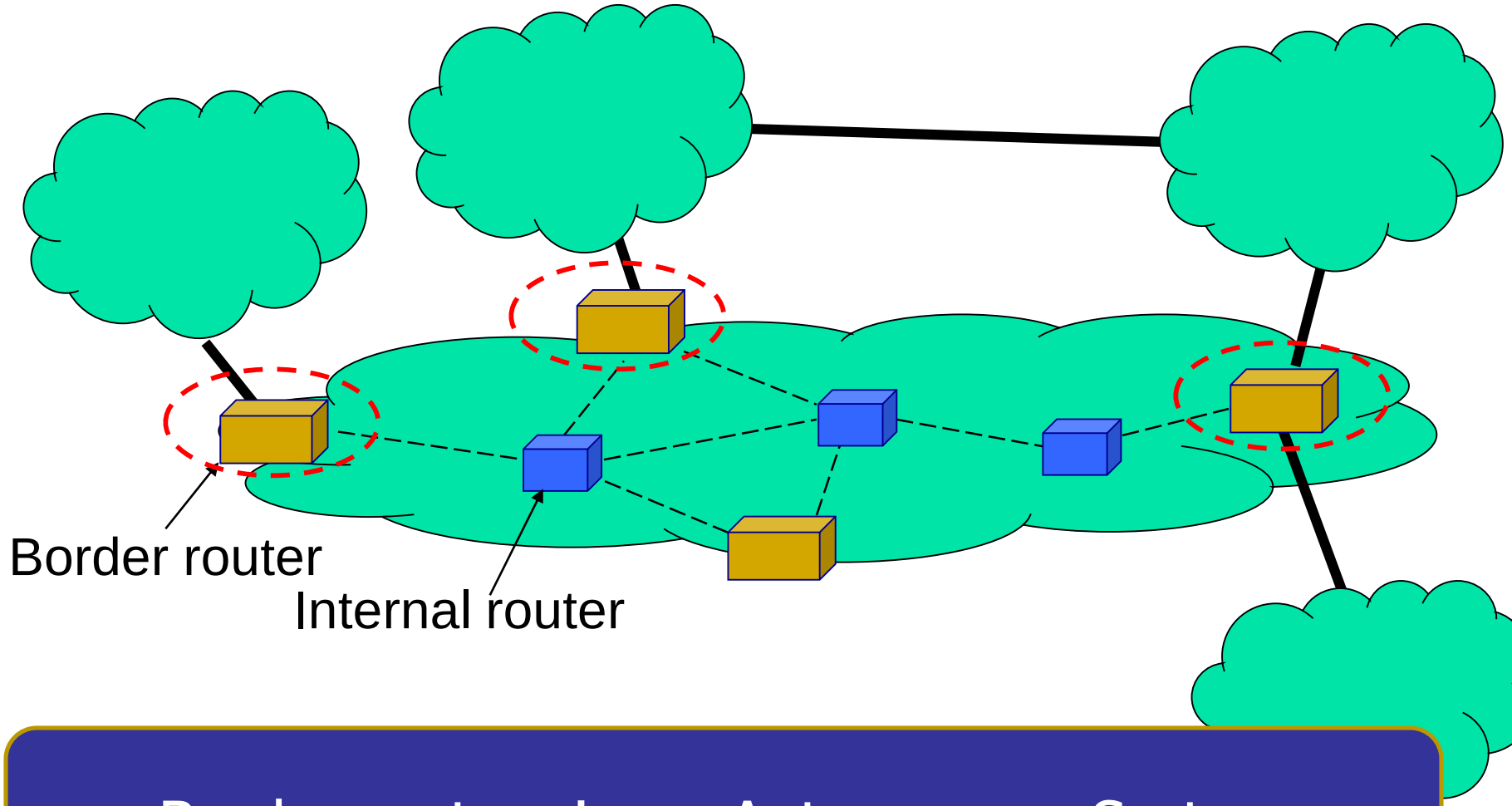
BGP & DV differences: (4) BGP may aggregate routes

- For scalability, BGP may aggregate routes for different prefixes



BGP Protocol details

Who speaks BGP?

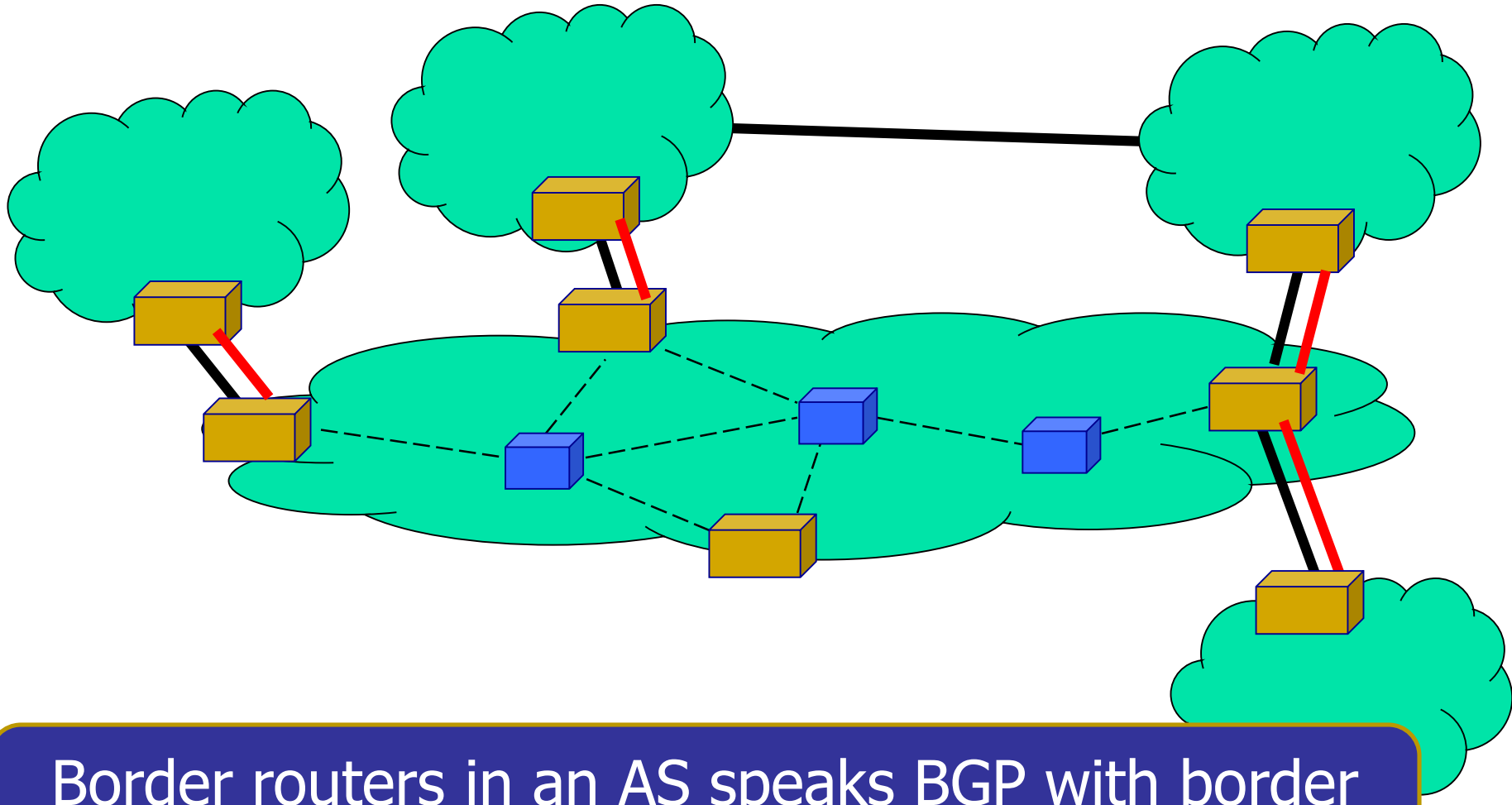


Border routers in an Autonomous System

What does "speak BGP" mean?

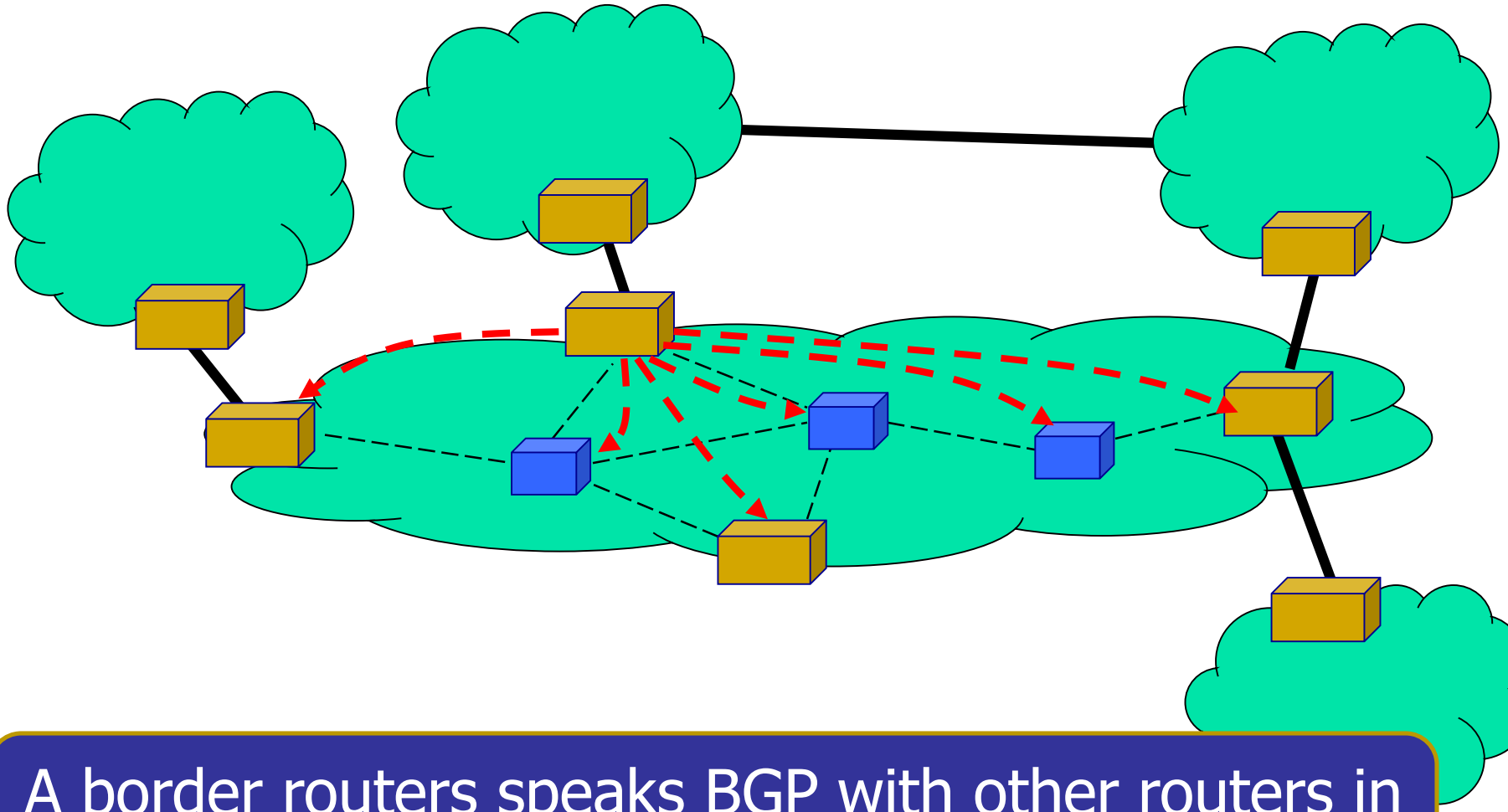
- Implement the BGP protocol standard
  - Read more here: http://tools.ietf.org/html/rfc4271
- Specifies what messages to exchange with other BGP "speakers"
  - Message types (e.g., route advertisements, updates)
  - Message syntax
- How to process these messages
  - E.g., "when you receive a BGP update, do…. "
  - Follows BGP state machine in the protocol spec + policy decisions, etc.

BGP sessions: External



Border routers in an AS speaks BGP with border routers in other ASes using eBGP sessions

BGP sessions: Internal



A border routers speaks BGP with other routers in the same AS using iBGP sessions

eBGP, iBGP, and IGP

- eBGP: BGP sessions between border routers in different ASes
  - Learn routes to external destinations
- iBGP: BGP sessions between border routers and other routers within the same AS
  - Distribute externally learned routes internally
- IGP: "Interior Gateway Protocol" = Intra-domain routing protocol
  - Provide internal reachability
  - E.g., OSPF, RIP

eBGP, iBGP, and IGP together

- Learn routes to external destination using eBGP
- Distribute externally learned routes internally using iBGP
- Travel shortest path to egress using IGP

Basic messages in BGP

- Open
  - Establishes BGP session (BGP uses TCP)
- Notification
  - Report unusual conditions
- Update
  - Inform neighbor of new routes
  - Inform neighbor of old routes that become inactive
- Keep-alive
  - Inform neighbor that connection is still viable

Route updates

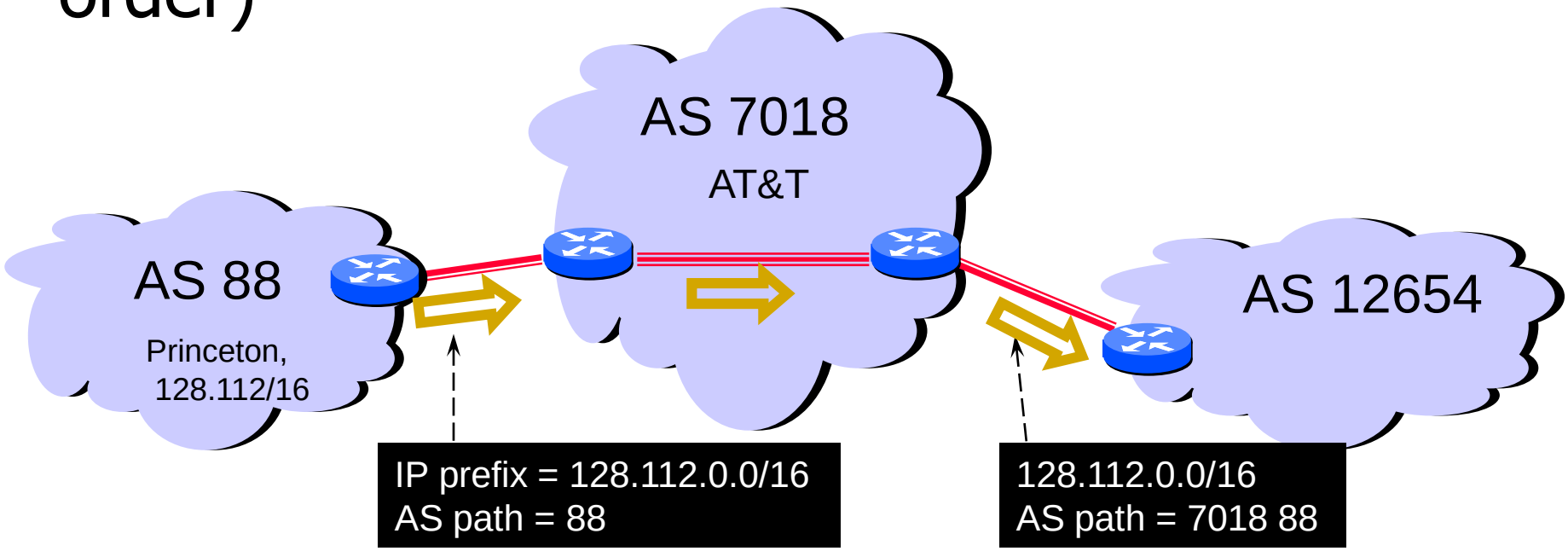
- Format <IP prefix: route attributes>
  - Attributes describe properties of the route
- Two kinds of updates
  - Announcements: new routes or changes to existing routes
  - Withdrawal: remove routes that no longer exist

Route attributes

- Routes are described using attributes
  - Used in route selection/export decisions
- Some attributes are local
  - I.e., private within an AS, not included in announcements
- Some attributes are propagated with eBGP route announcements
- There are many standardized attributes in BGP
  - We will discuss a few

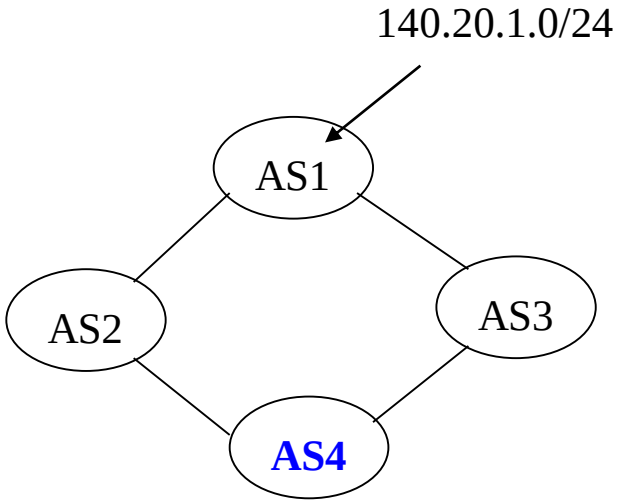
Attributes: (1) ASPATH

- Carried in route announcements
- Vector that lists all the ASes a route advertisement has traversed (in reverse order)



Attributes: (2) LOCAL PREF

- Local preference in choosing between different AS paths
  - Local to an AS; carried only in iBGP messages
- The higher the value the more preferred

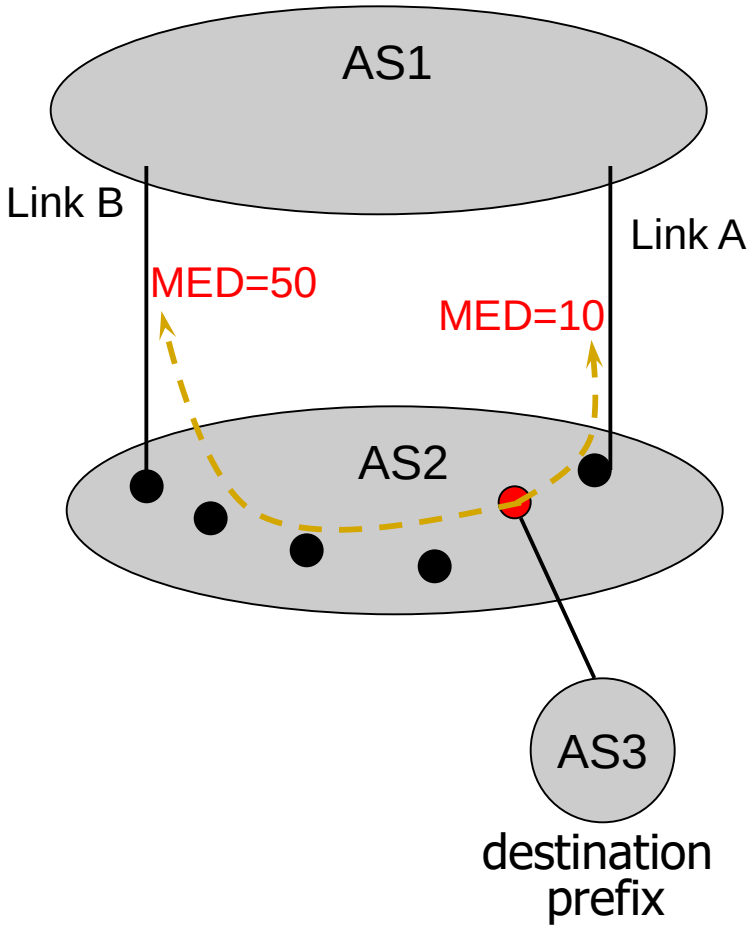


BGP table at AS4:

Destination	AS Path	Local Pref
140.20.1.0/24	**AS3 AS1**	**300**
140.20.1.0/24	**AS2 AS1**	**100**

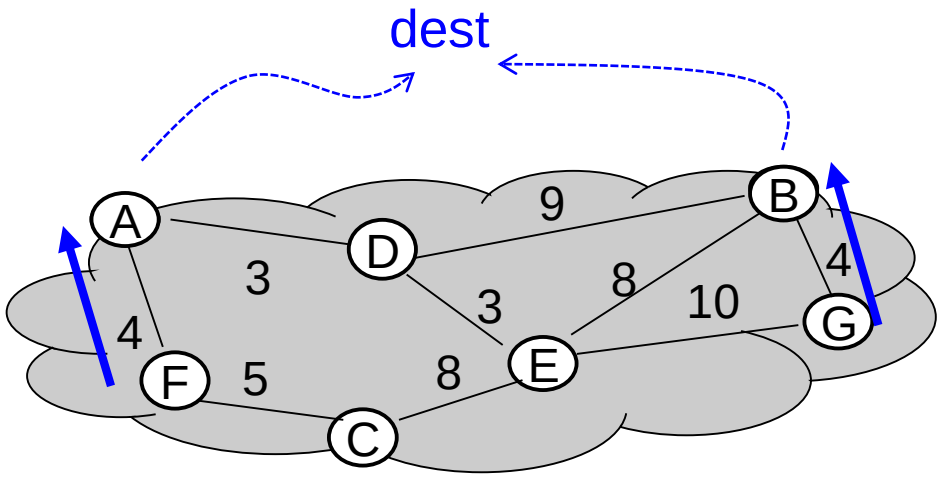
Attributes: (3) MED

- Multi-exit discriminator is used when ASes are interconnected via 2 or more links; it specifies how close a prefix is to the link it is announced on
- Lower is better
- AS that announces a prefix sets MED
- AS receiving the prefix (optionally!) uses MED to select link

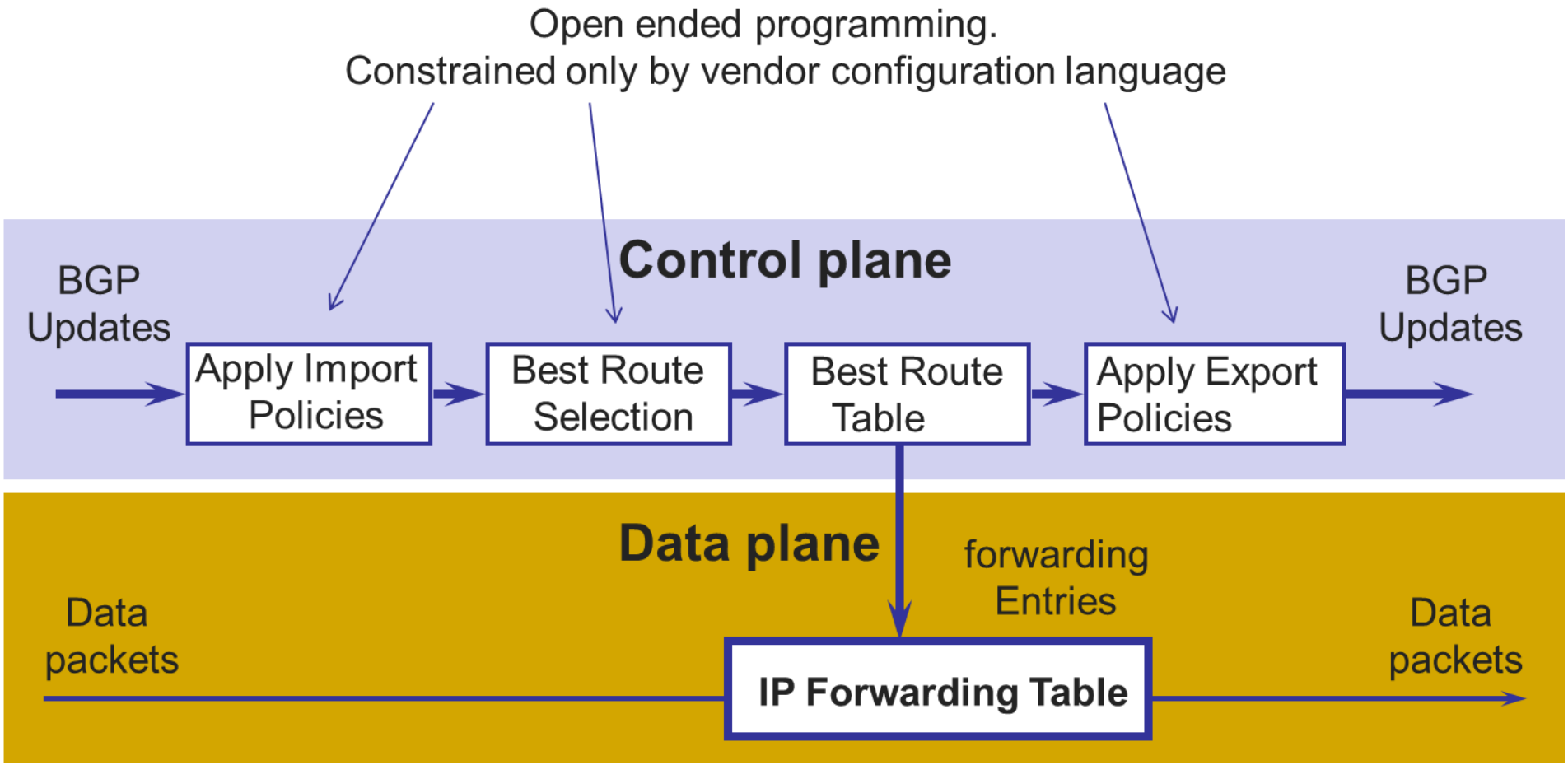


Attributes: (4) IGP cost

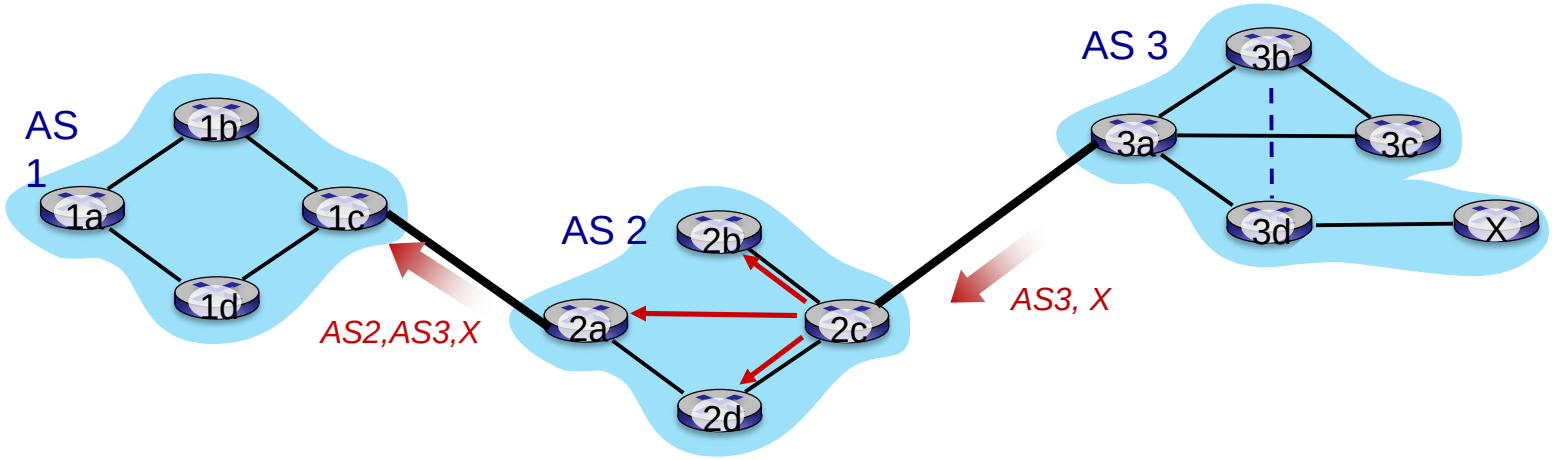
- Used for hot-potato routing
  - Each router selects the closest egress point based on the path cost in intra-domain protocol



BGP UPDATE processing

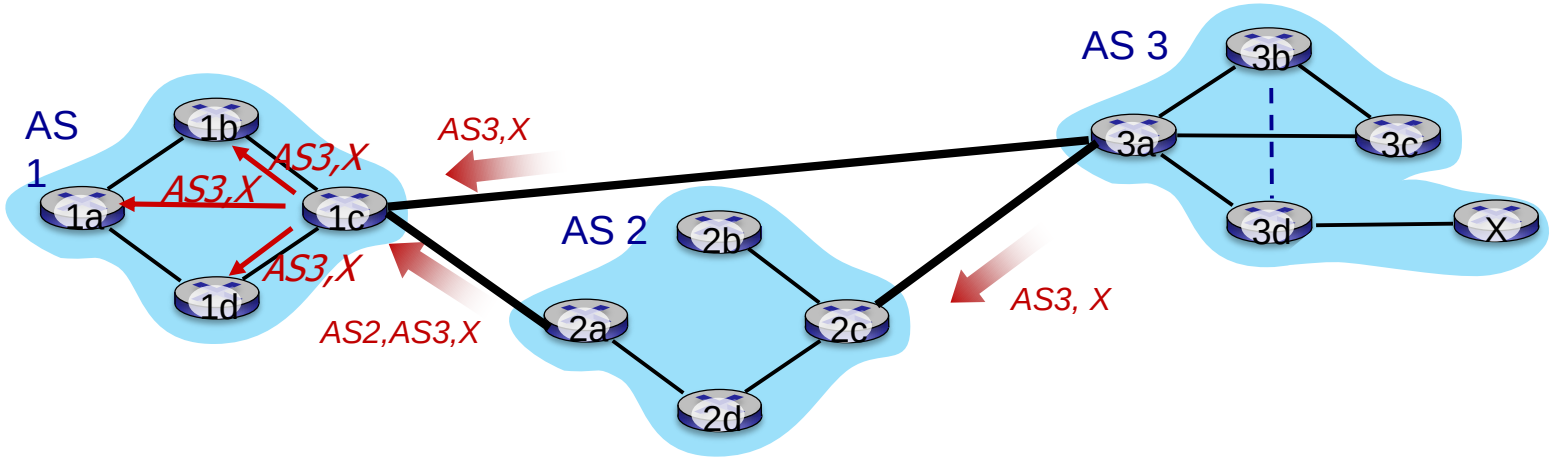


BGP example: path advertisement



- 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

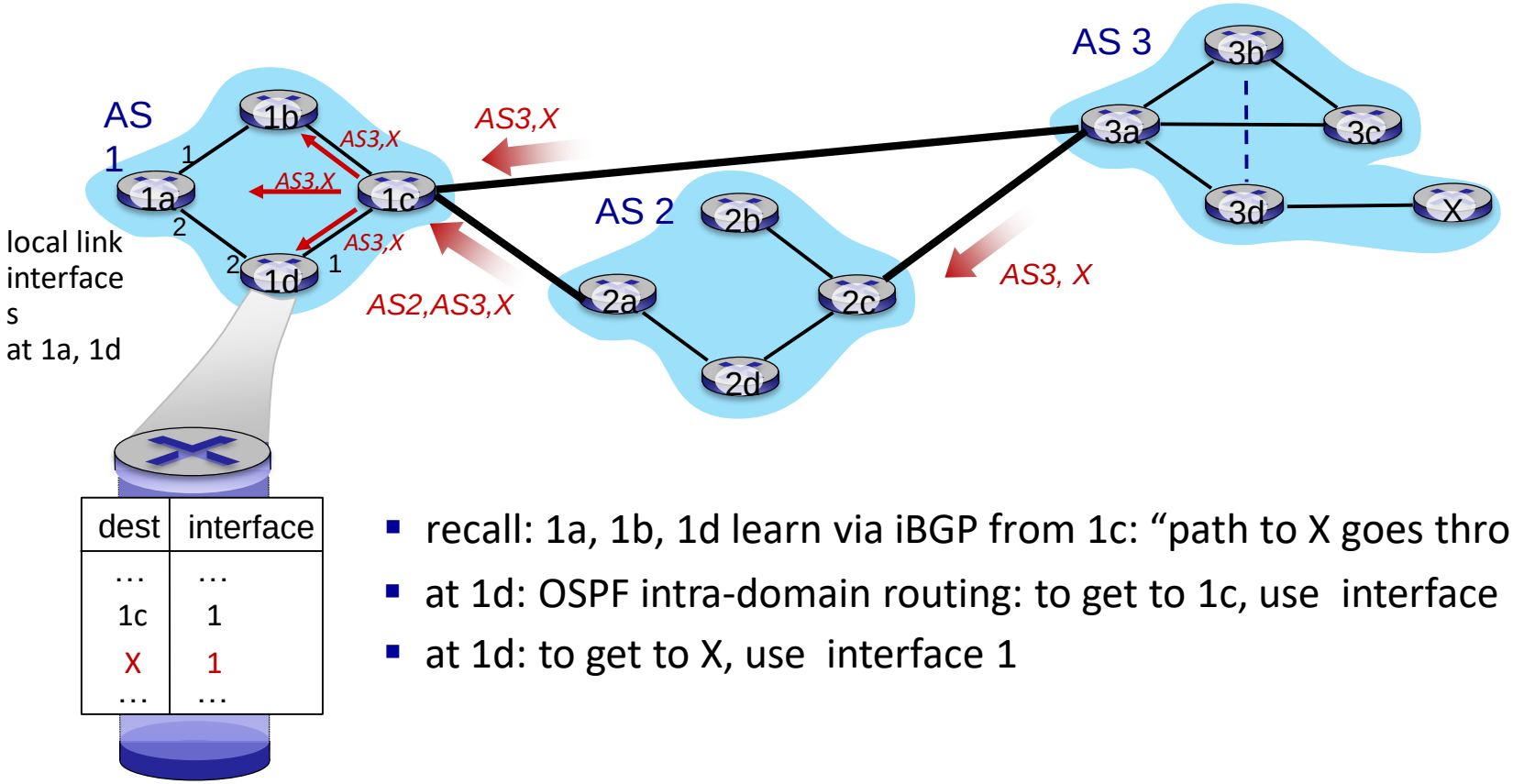
BGP example: multiple paths



gateway router may learn about multiple paths to destination:

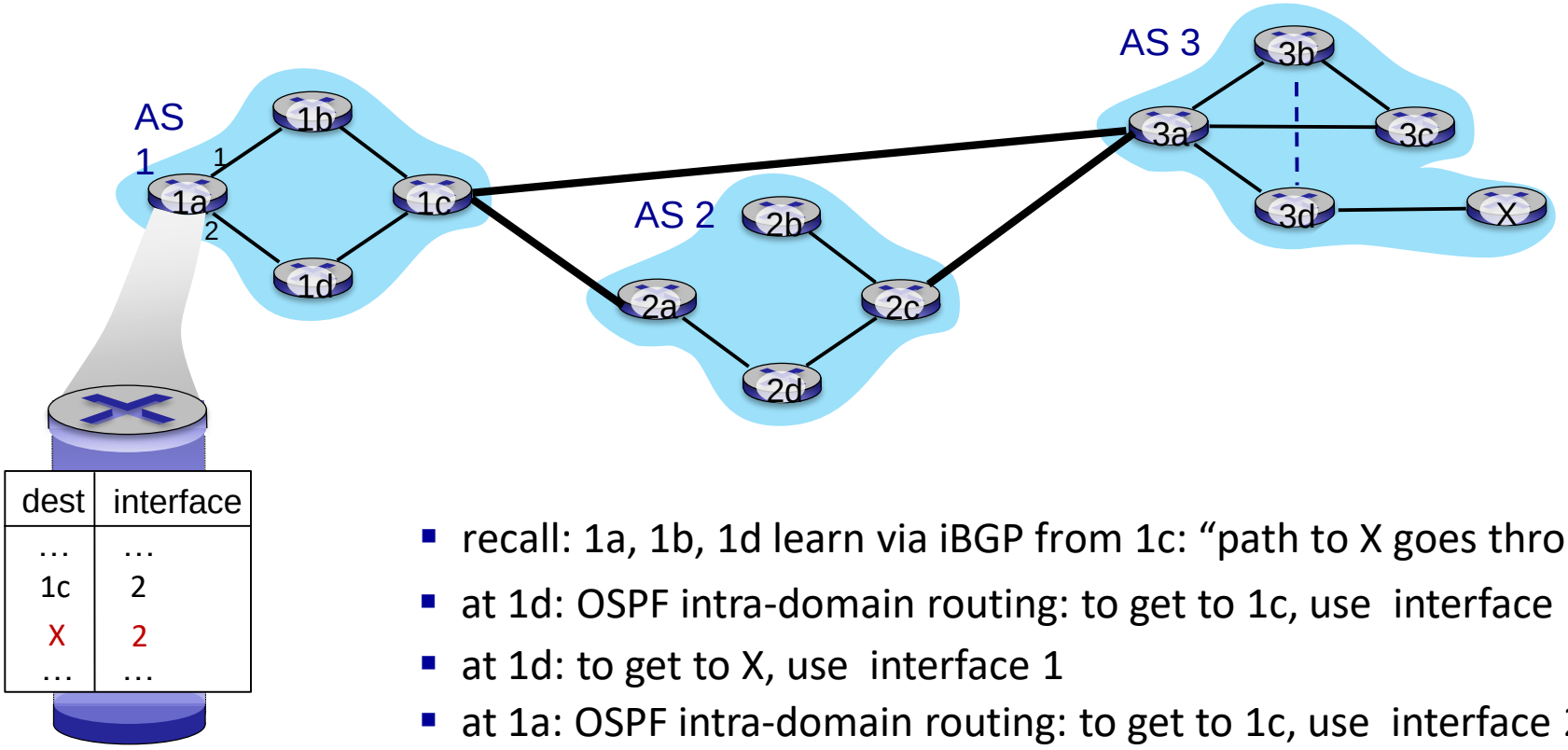
- AS1 gateway router 1c learns path *AS2,AS3,X* from 2a
- AS1 gateway router 1c learns path *AS3,X* from 3a
- based on *policy,* AS1 gateway router 1c chooses path *AS3,X* and advertises path within AS1 via iBGP

BGP example: populating forwarding tables



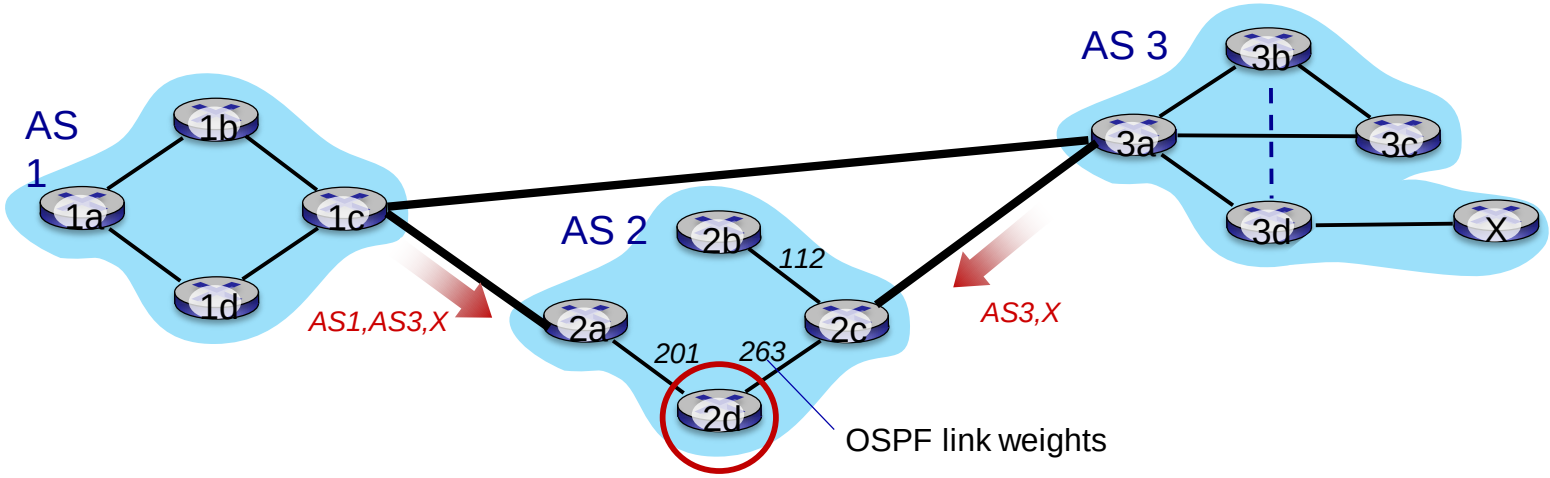
- recall: 1a, 1b, 1d learn via iBGP from 1c: “path to X goes through 1c”
- at 1d: OSPF intra-domain routing: to get to 1c, use interface 1
- at 1d: to get to X, use interface 1

BGP example: populating forwarding tables



- 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

BGP example: 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!

BGP协议特点

- 主要用于表示路由的可达性，不一定走最短路径路由
- 通过携带AS路径信息，可以解决路由循环问题
- 使用TCP协议，端口号179，可实现协议的可靠性
- BGP 协议交换路由信息的结点数量级是自治系统数的量级。
- 由于自治系统中 BGP Speaker（或边界路由器）的数目是很少的，使得自治系统之间的路由选择不致过分复杂。
- 支持 CIDR，可以进行路由聚合。
- 在BGP 刚刚运行时，BGP 的邻站交换整个的 BGP 路由表。但以后只需要在发生变化时增量更新有变化的部分，减少开销。

Summary

- IGP – Intra-AS protocols
  - RIP: Routing Information Protocol, use distance vector
  - OSPF: Open Shortest Path First, use link state
  - IGRP: Interior Gateway Routing Protocol (Cisco proprietary)
- EGP – Inter-AS protocols
  - BGP: Border Gateway Protocol