Chapter 4. Internetworking

- The Internet Protocol
- IP Address
- ARP and DHCP
- ICMP
- IPv6
- Mobile IP
- Internet Routing
- IP Multicasting
- Multiprotocol Label Switching (MPLS)

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

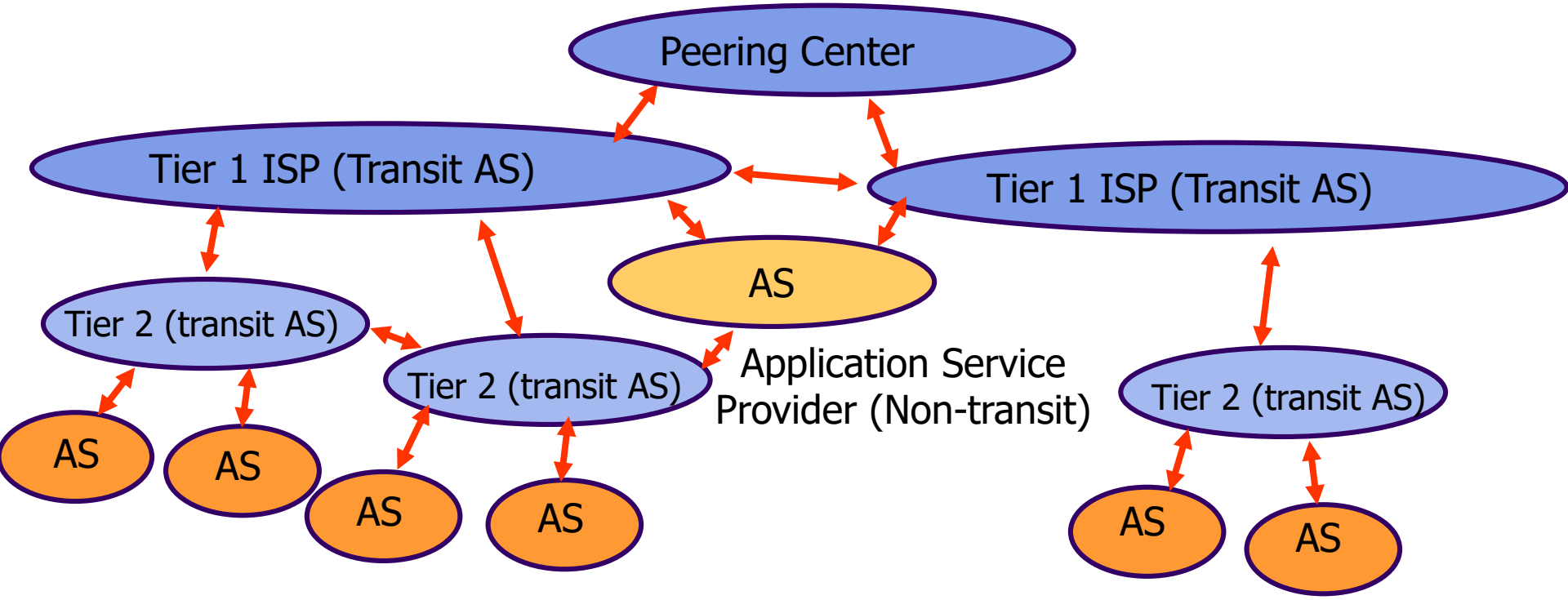
Hierarchical Routing

- 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)

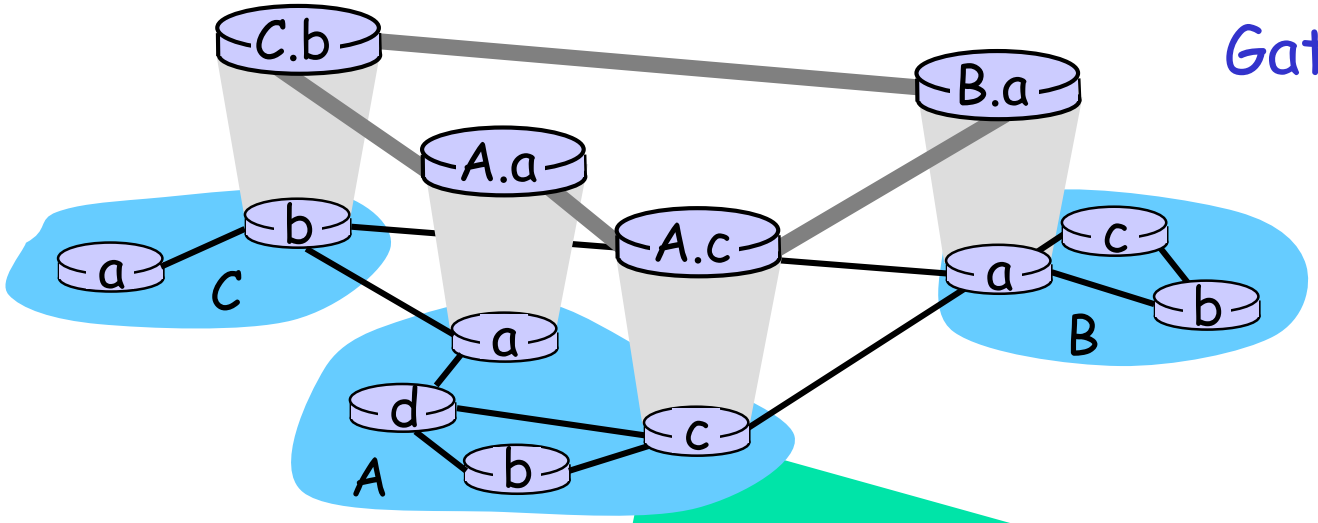
- Set of routers and networks managed by single ISP or large organization
- A connected internets uniquely assigned a 16-bit or 32-bit AS Number
  - There is at least one route between any pair of nodes
- Use common routing protocol

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

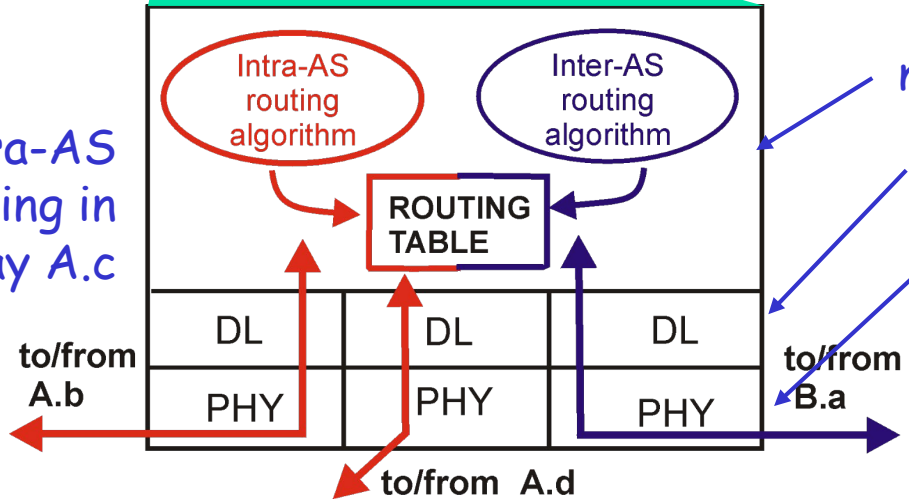
Intra-AS and Inter-AS routing



Gateways:

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

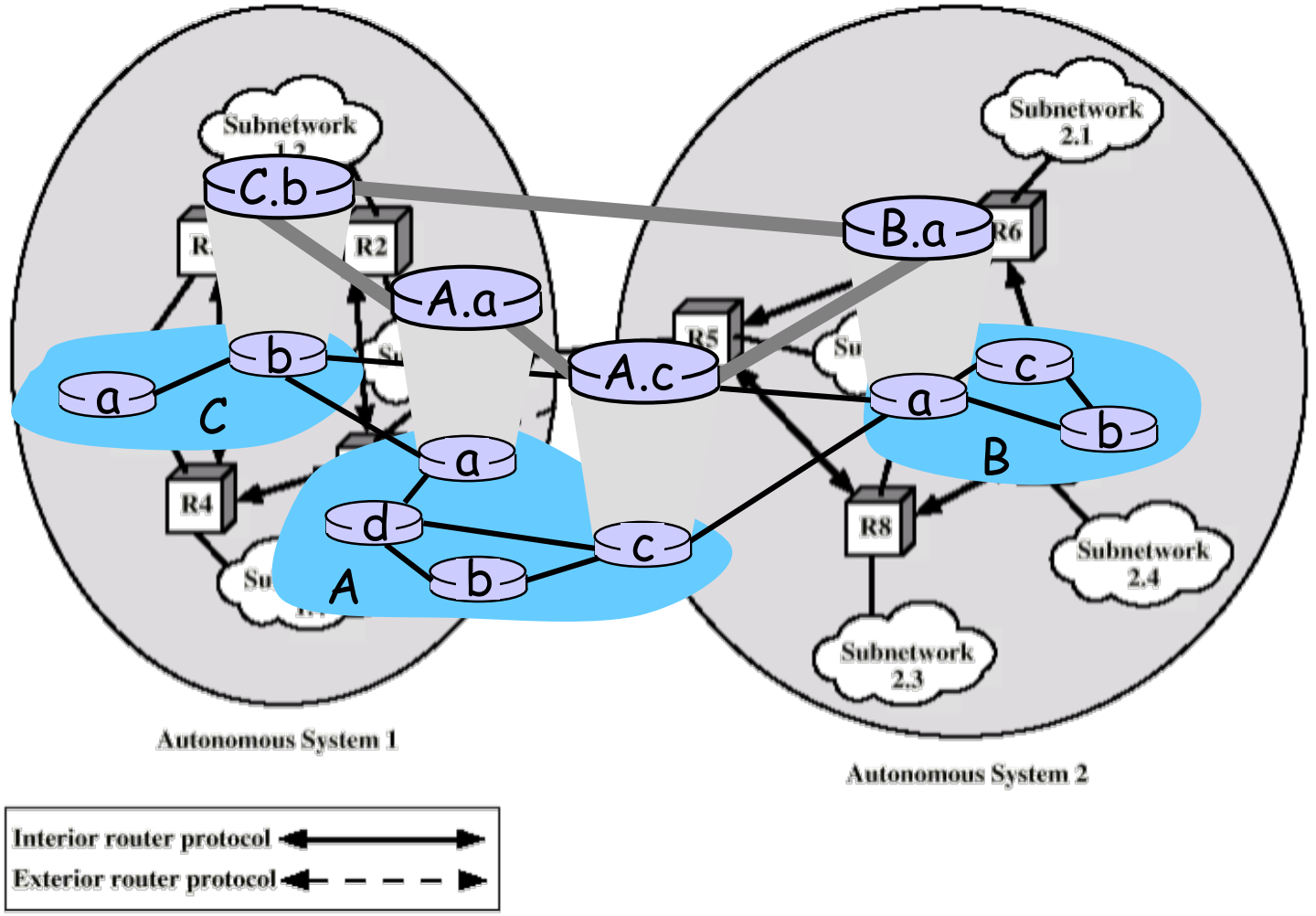
inter-AS, intra-AS routing in gateway A.c



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)

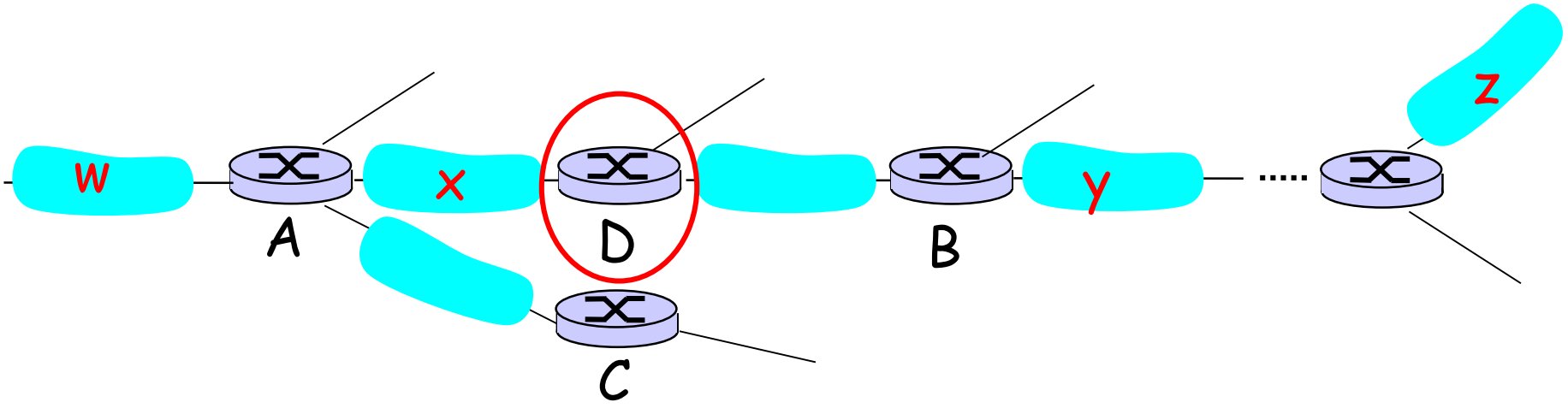
- Use Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982

- 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)



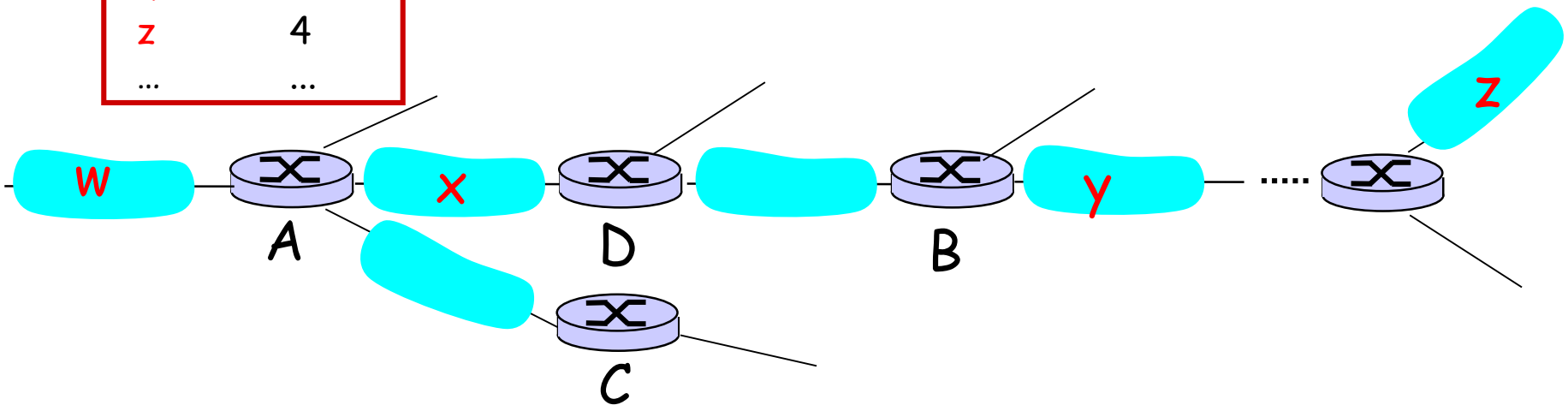
Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	B	7
x	--	1
....	....	....

Routing table in D

RIP: Example (2)

Dest	hops
w	1
x	1
z	4
…	…

Advertisement from A to D

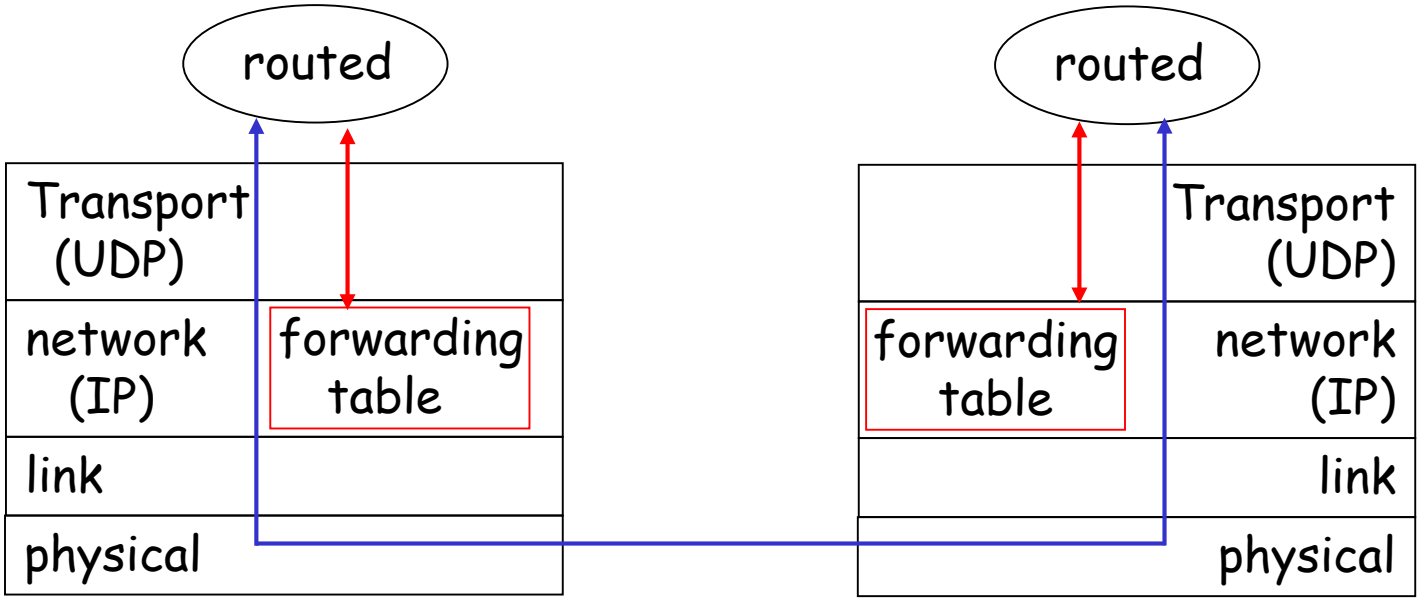


Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	~~B~~ A	~~7~~ 5
x	--	1
….	….	….

Routing table in D

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



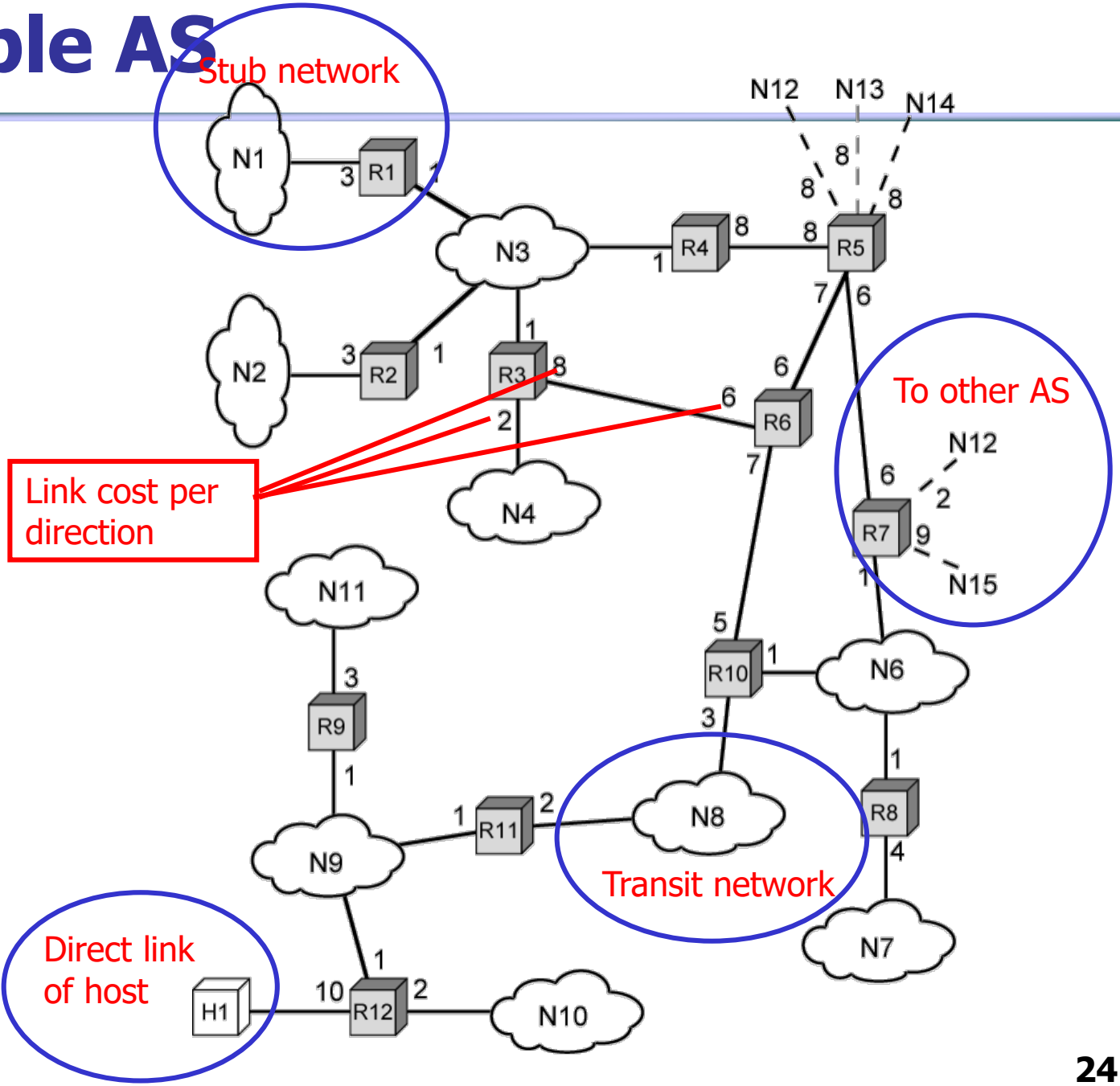
Open Shortest Path First (1)

- 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 10s
  - Carried in OSPF messages directly over IP, Not UDP
- Uses cost metric assigned on each link

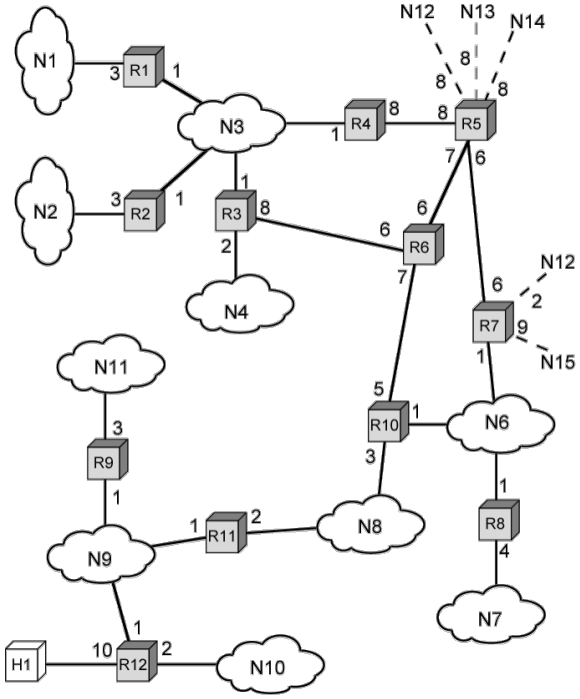
Open Shortest Path First (2)

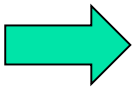
- Topology map stored as directed graph on each node
- Router nodes
- Network nodes: (Transit vs. Stub)
- Edges: router—router, router—network
- Dijkstra's algorithm used to compute the shortest path to each destination

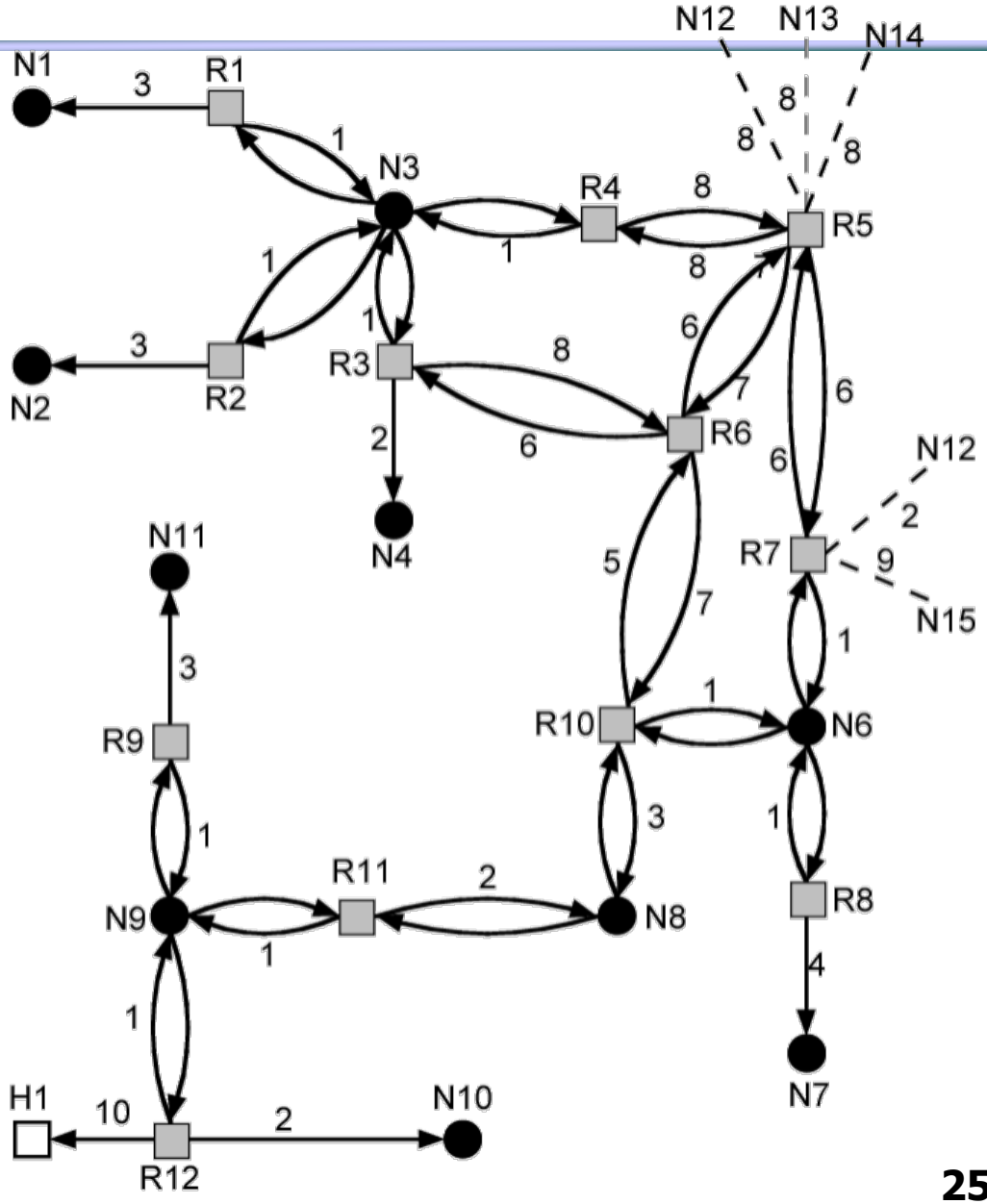
Sample AS


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



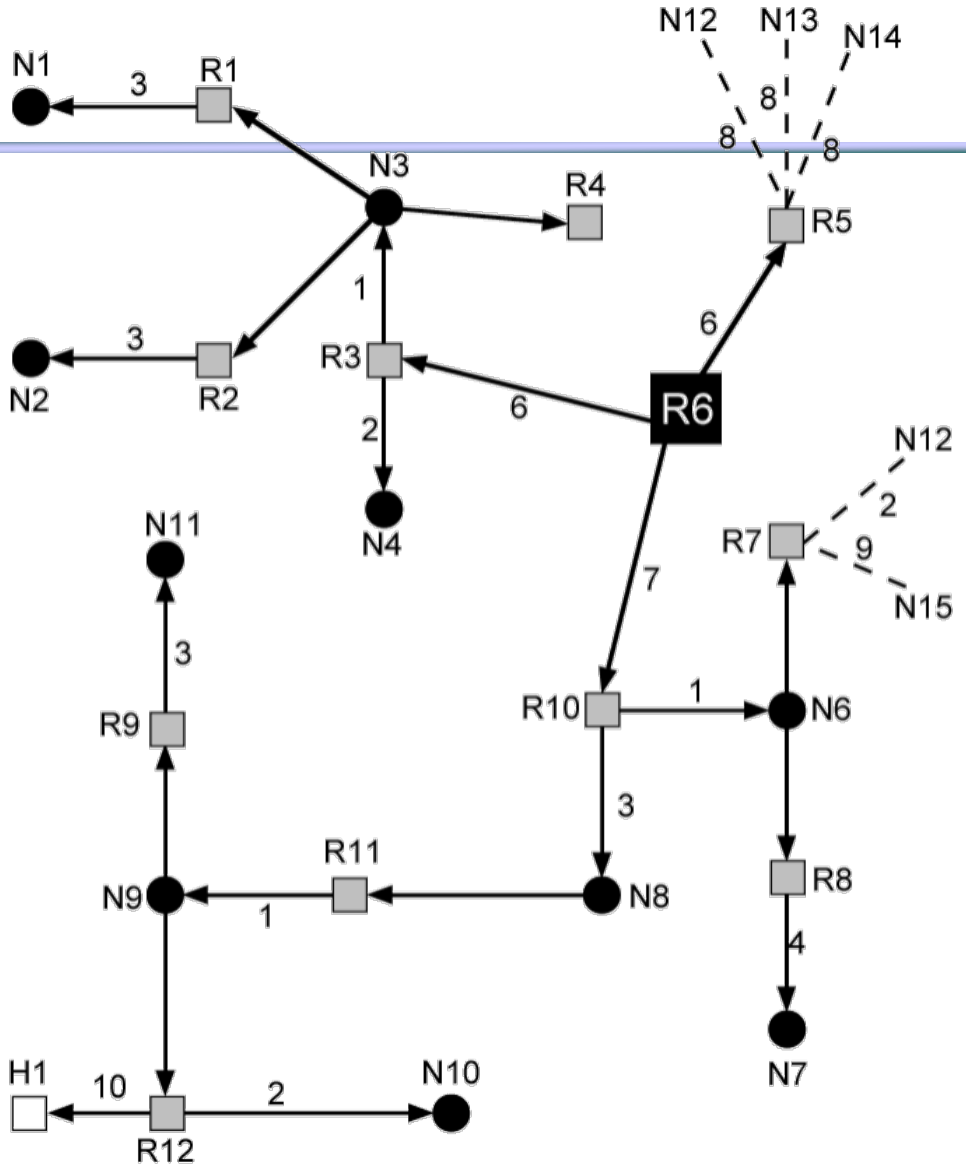




SPF Operation

- 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*

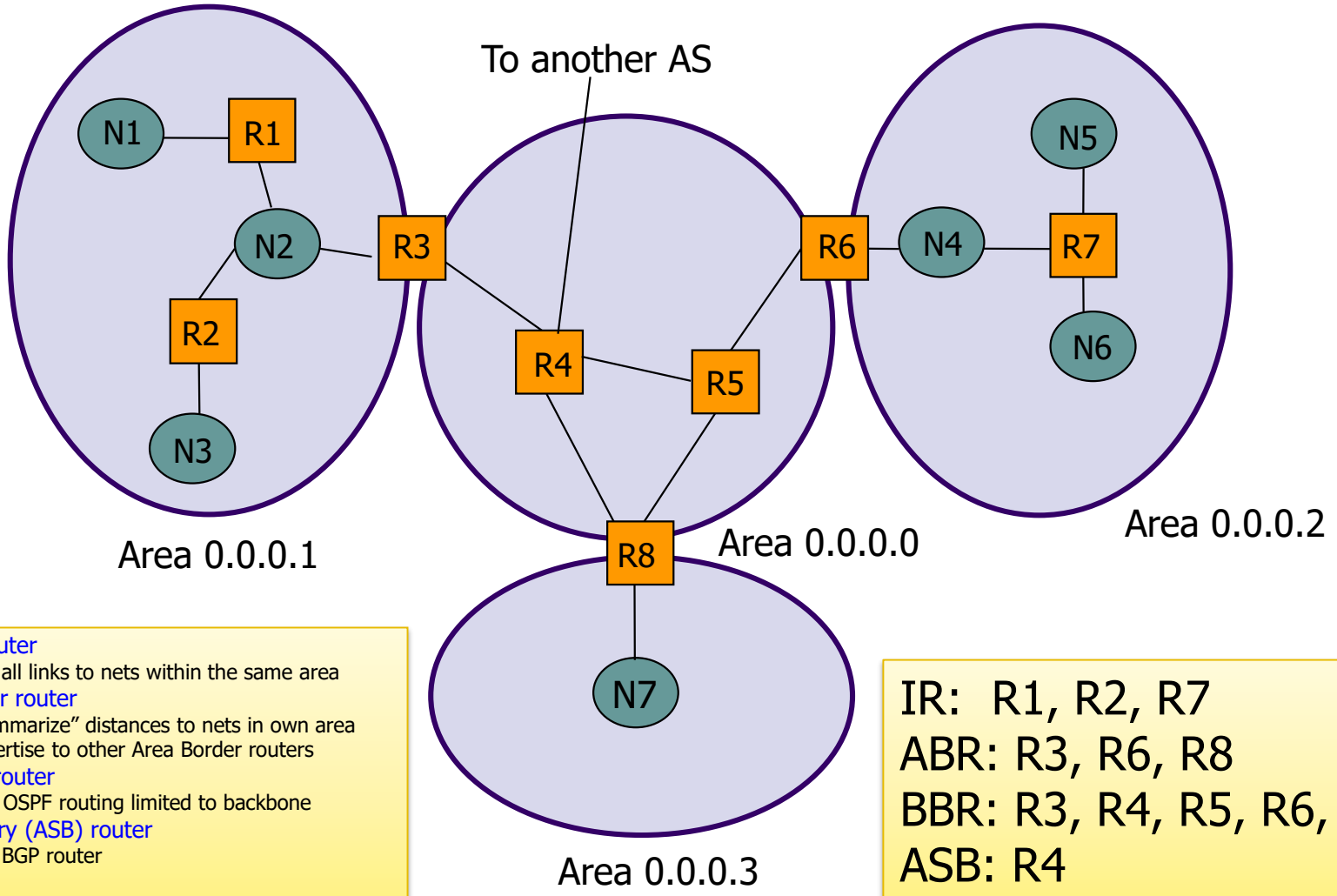
- Hierarchical OSPF in large domains

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

Border Gateway Protocol

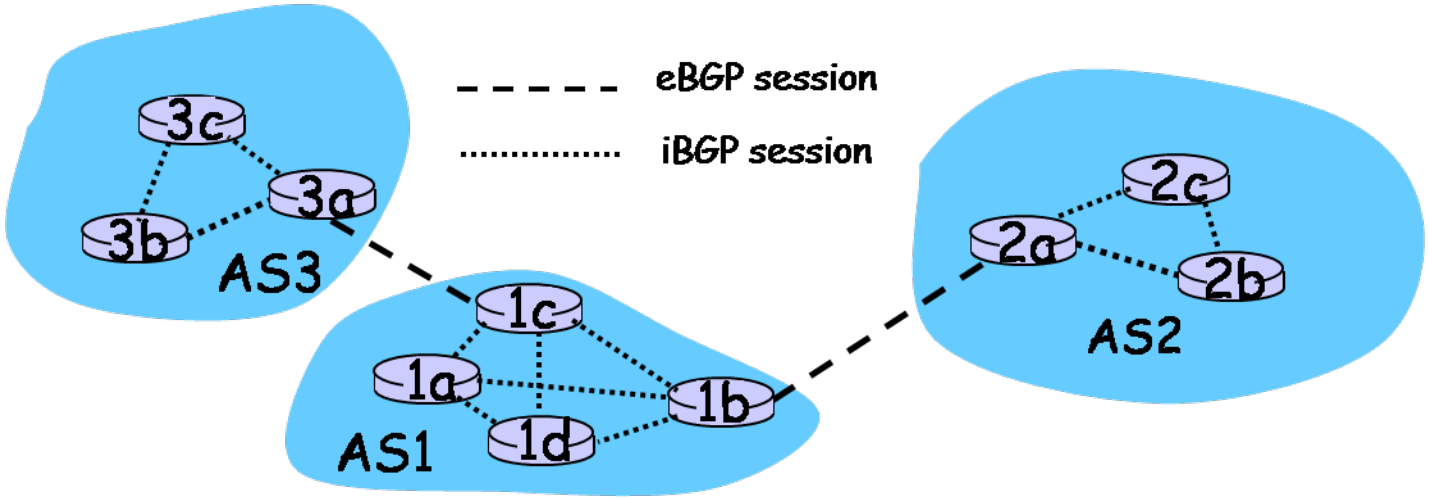
- Subnets use BGP to advertise its existence to rest of Internet

- BGP provides each AS a means to
  - Obtain subnet reachability information from neighboring ASs using eBGP
  - Propagate reachability information to all AS-internal BGP-routers using iBGP
  - Determine "good" routes to subnets based on reachability information and policy

BGP Basics

- Pairs of BGP routers exchange routing info over TCP connections: BGP sessions

- When AS2 (2a) advertises a net prefix to AS1 (1b)
  - AS2 promises it will forward any datagrams addressed towards that prefix
  - Net prefixes can be aggregated during advertisement



BGP – the Protocol

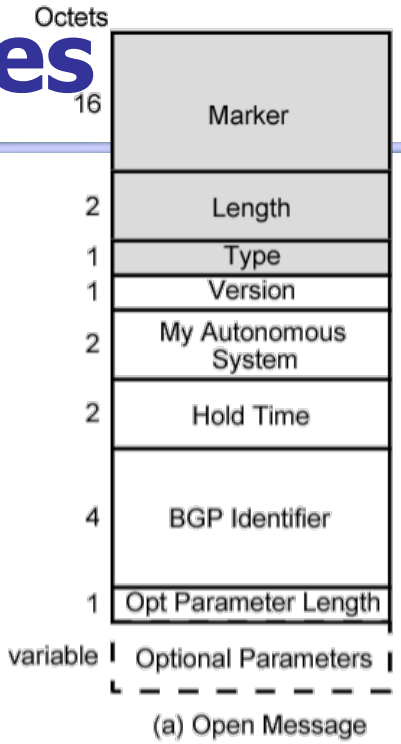
4 types of BGP messages

- Open: opens TCP connection to peer and authenticates sender
- Update: (1) advertises new path; (2) withdraws old
- Keep-alive: (1) ACKs OPEN request; (2) keeps connection alive in absence of UPDATES
- Notification: (1) closes connection; (2) reports errors in previous message

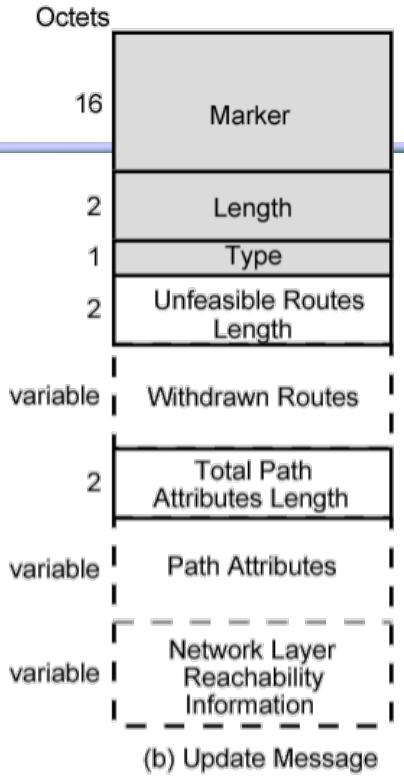
Procedures of BGP

- Neighbor acquisition
  - One router sends an *Open* message to another
  - If the target router accepts the request, it returns a *Keep-alive* message
- Neighbor reachability
  - The two routers periodically issue *Keep-alive* or *Update* messages to each other
- Network reachability
  - Each router maintains a database of networks that it can reach and the list of *ASs* passed
  - The router issues an *Update* message whenever a change is made to this database

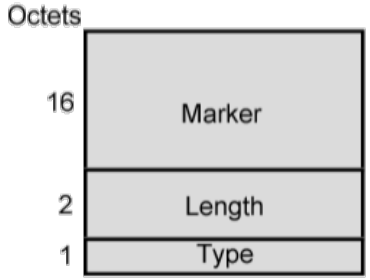
BGP Messages



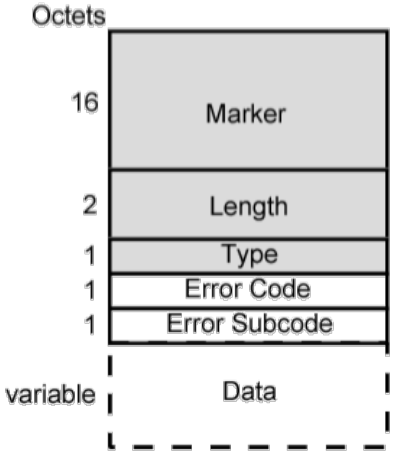
(a) Open Message



(b) Update Message



(c) Keepalive Message

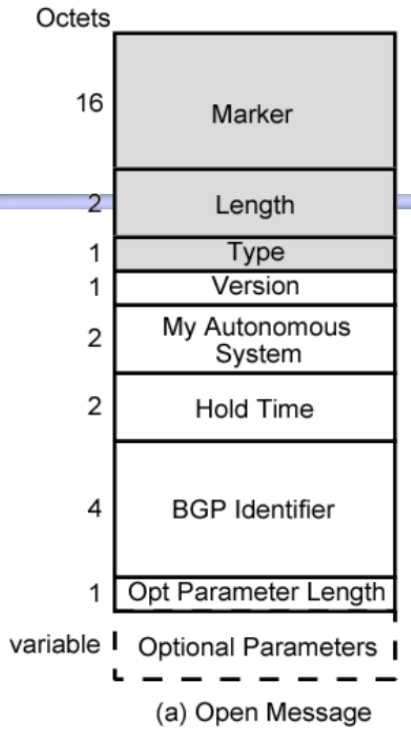


(d) Notification Message

BGP Messages

- 3 common fixed-size fields in each header
- Marker (16 octets)
  - Detect loss of synchronization between a pair of BGP router
  - Authenticate incoming BGP messages
- Length (2 octets)
  - Length of message in octets, including the header
- Type (1 octets)
  - 1.Open, 2.Update, 3.Notification, 4.Keep-alive

Open Message



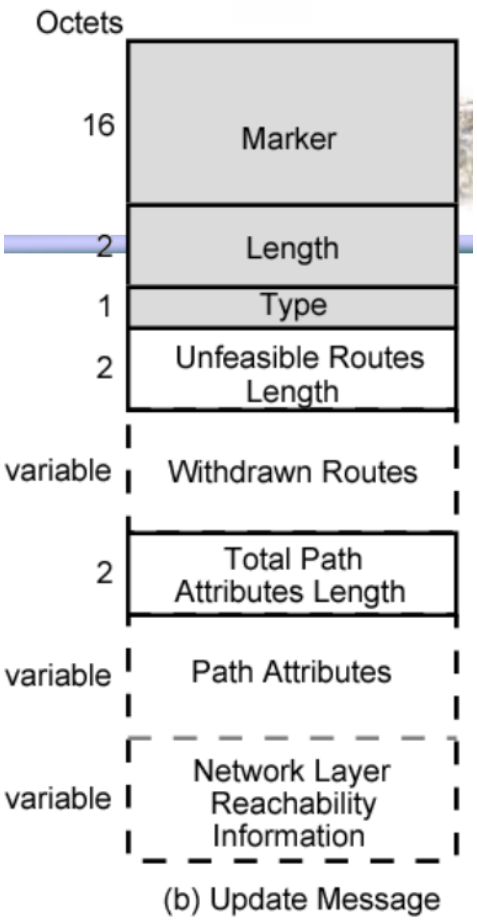
- Version (1 octet)
  - Current BGP version (v4)
- My Autonomous System (2 octets)
  - AS number the sender belongs to
- Hold time (2 octets)
  - Max time between Keep-alive and/or update messages
- BGP Identifier (4 octets)
  - Identifier of the sender, one of its IP addresses
- Opt parameter length (1 octet)
  - Total length of the Optional parameter field in octet

8 16

Parm. Type	Parm. Length	Parameter Value (Variable)

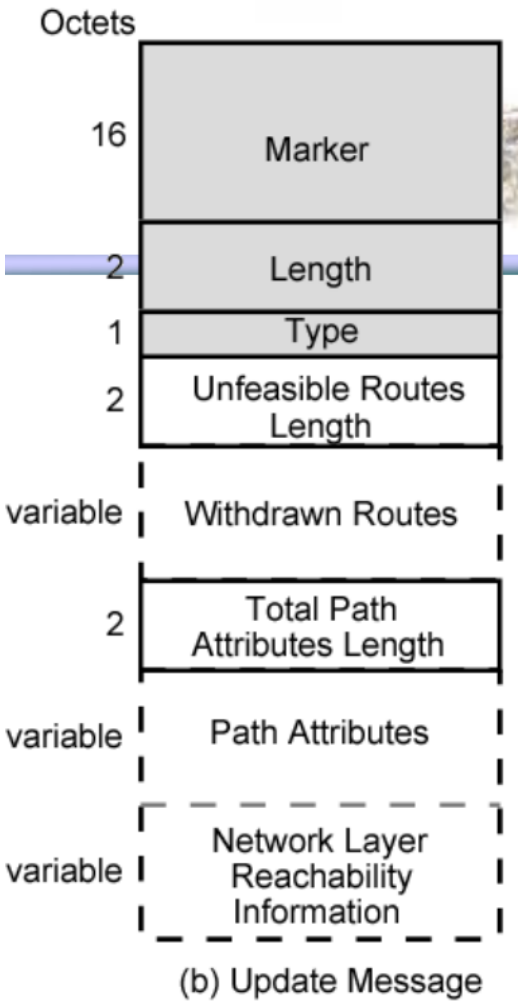
Update Message (1)

- Unfeasible Routes Length (2 octets)
  - Total length of withdraw routes in octets
- Withdrawn route (variable length)
  - A list of IP address prefixes, 2-tuple of the form <length, prefix>
  - Each prefix identifies a group of subnets
  - e.g. <10, D8CA> means 16 bits length, 216.202.0.0 network
- Total Path Attribute Length (2 octets)
  - Total length of path attribute field in octets for new path



Update Message (2)

- Path Attribute (variable length)
  - A list of path attributes, each path attribute is a triple <attribute type, attribute length, attribute value>
  - Attributes that apply to the particular router or route
- Network Layer Reachability Information (variable length)
  - A list of IP address prefixes, each one is 2-tuple of the form <length, prefix>
  - A single route through the internet

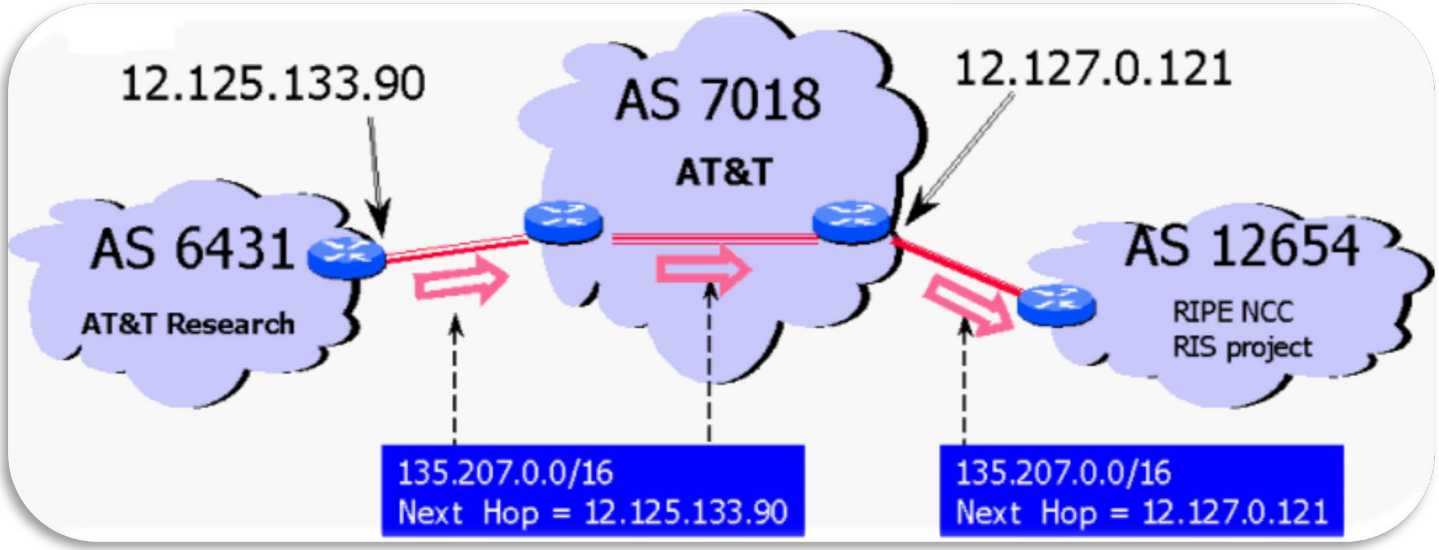


Defined Path Attributes (1)

- Origin
  - Learned from IGP or EGP
- AS_Path
  - A list of AS traversed, in ordered or unordered way
- Next_hop
  - IP address of the border router that are used as the next hop
  - Responsible for informing outside routers of the route to other networks

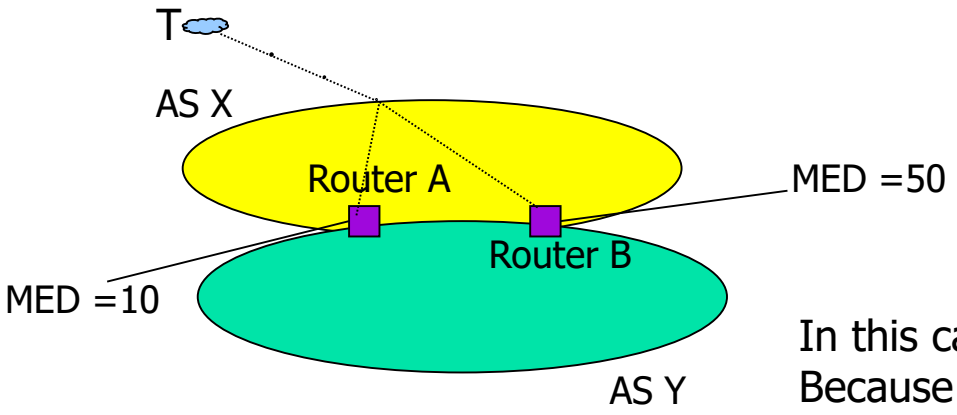
Defined Path Attributes (2)

- Next_hop



Defined Path Attributes (3)

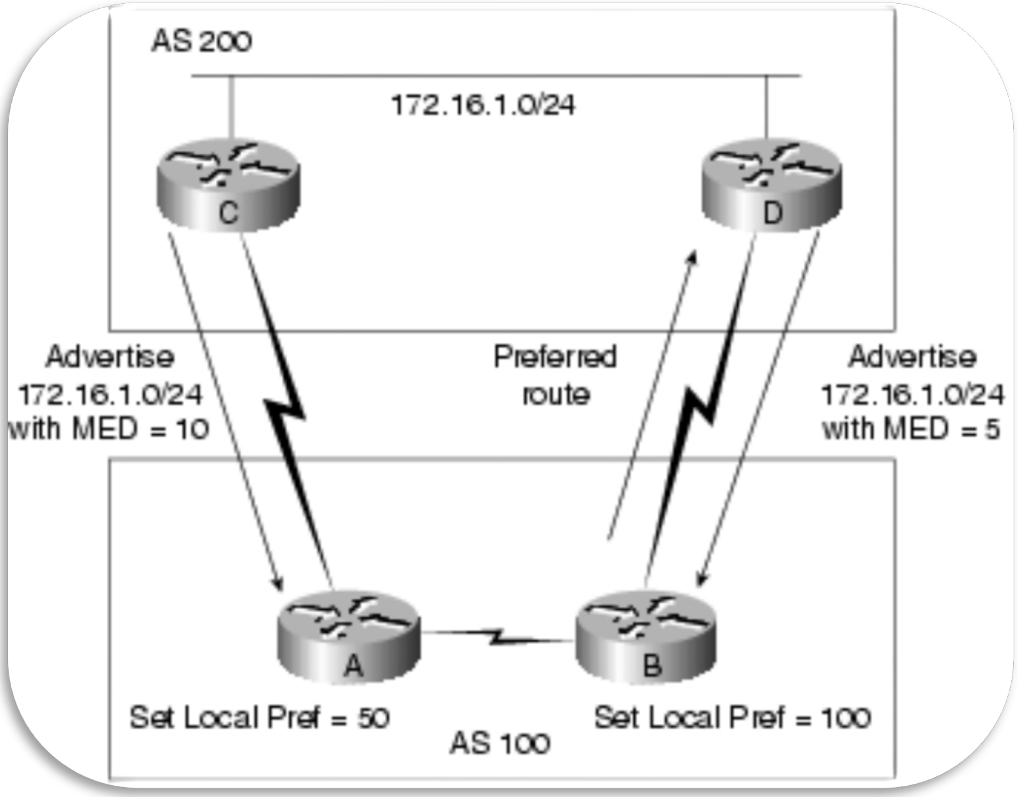
- Multi_Exit_Disc (MED)
  - There may be multiple border points in one *AS* available to another *AS*
  - MED is a metric value computed by certain routing policy within the *AS*
  - It can be used (by eBGP) to discriminate among multiple exit points



In this case, it selects route used router A.
Because router A's MED < router B's MED

Defined Path Attributes (4)

- Local_pref
  - Should be included when the 2 BGP speakers located within the same AS, i.e. for iBGP

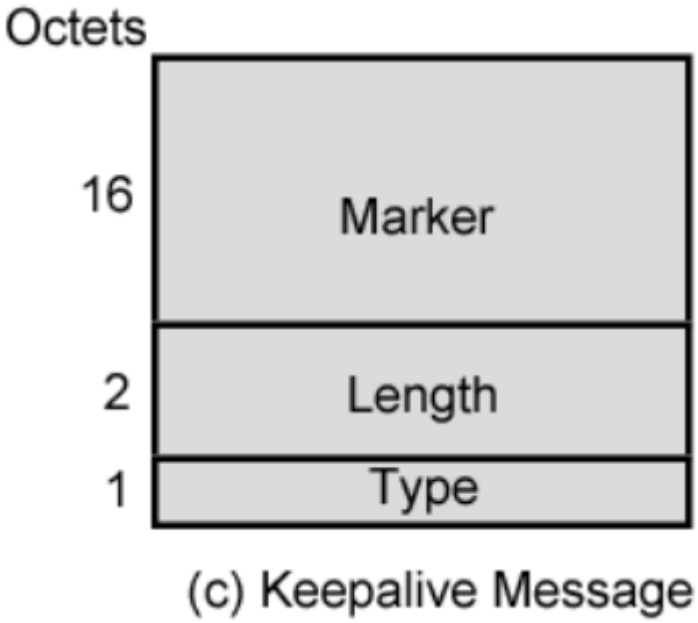


Defined Path Attributes (5)

- Atomic_Aggregate
  - Informs others that the local AS has aggregate the addresses of subnets
  - i.e. some interim specific route is hided
- Aggregator
  - Contains the last AS number and IP address of the BGP router that formed the aggregates

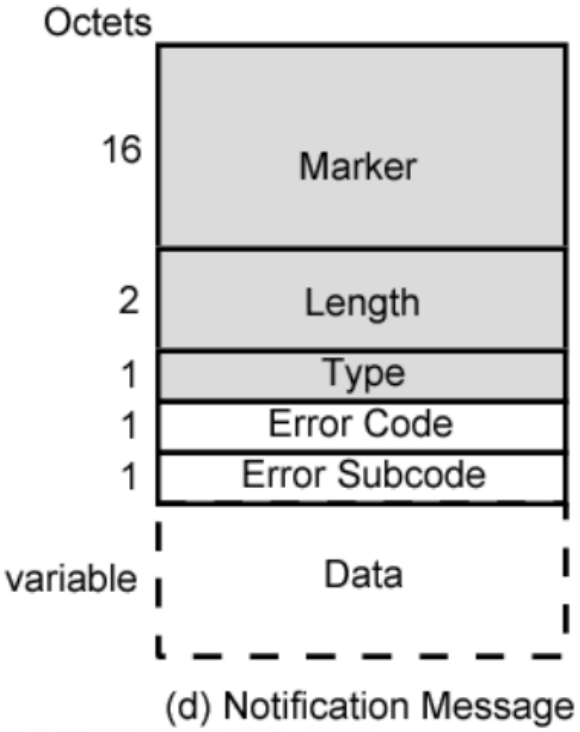
Keep Alive Message

- To tell other routers that this router is still here
- BGP speaker send Keep-Alive message periodically to keep conne



(c) Keepalive Message

Notification Message (



- Message header error
  - Authentication and syntax, subtypes:
  - Connection Not Synchronized
  - Bad Message Length
  - Bad Message Type
- Open message error
  - Syntax and option not recognized, Unacceptable hold time, subtypes:
  - Unsupported Version Number
  - Bad peer AS
  - Bad BGP identifier
  - Unsupported Optional Parameter, …

Notification Message (2)

- Update message error
  - Syntax and validity errors
- Hold time expired
  - Connection is closed
- Finite state machine error
  - Any procedural errors: wrong message at wrong states
  - e.g. got *Open* message at *Connect* state
- Cease
  - Used to close a connection when there is no error

BGP Routing Information Exchange

- Within *AS*, router builds topology picture using IGP
- BGP router issues Update message for rising subnet within to other BGP routers outside *AS*
- These routers exchange info with BGP routers in other *ASs*
- Routers then decide best routes use policy routing

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

Homework

- 第四章：R23, R29, P35, P37