The Network Layer

Chapter 5

- Provide facilities for getting data from a source to a destination
- May require making many hops at intermediate routers along the way

5.1 Network Layer Design Issues

- Store-and-forward packet switching
- Services provided to transport layer
- Implementation of connectionless service
- Implementation of connection-oriented service
- Comparison of virtual-circuit and datagram networks

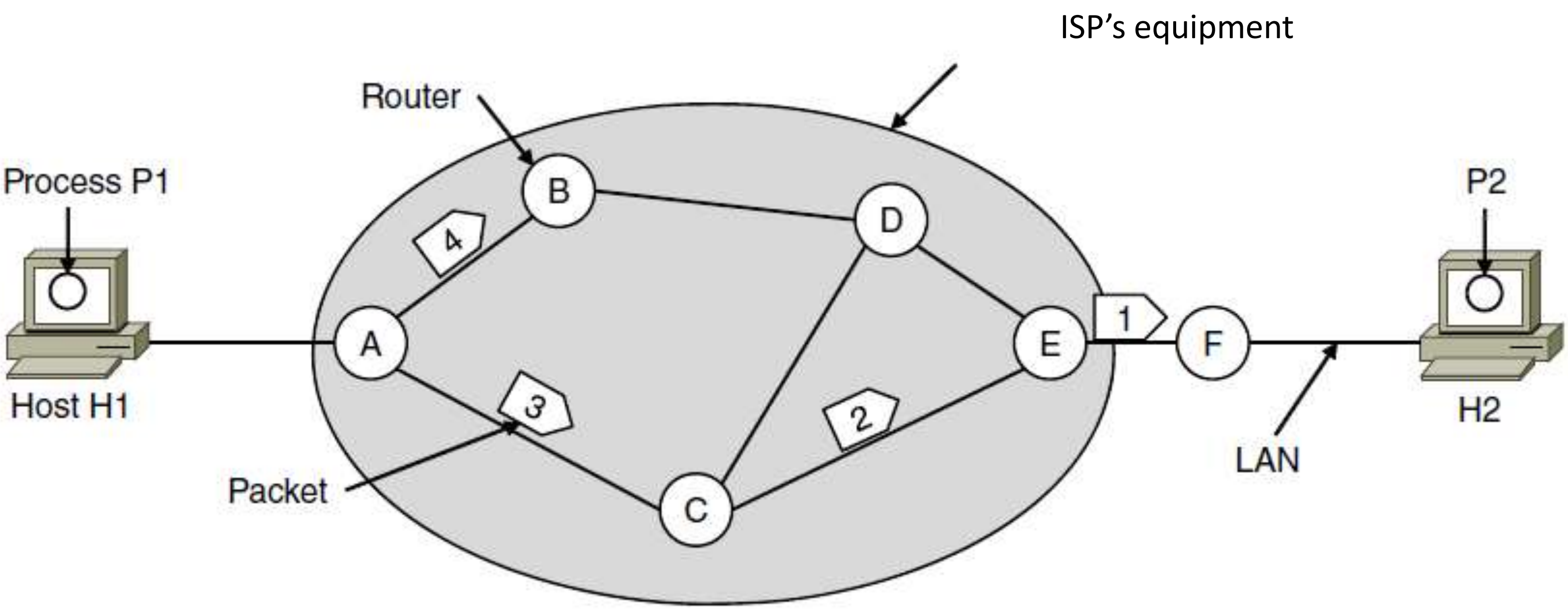
Flashback….

- 电路交换:直接利用可切换的物理通信线路。
  - 三个阶段：建立电路、传输数据、拆除电路
- 存储交换（存储-转发）
  - 报文交换：信息以报文（逻辑上完整的信息段）为单位进行存储转发；
  - 分组交换：比报文还小的信息段，通常有最大长度的限制；
    - ✓ 数据报:分组独立路由
    - ✓ 虚电路：所有分组只作一次路由（建立虚电路）
  - 信元交换：大小固定的信息段

Services Provided to the Transport Layer

1. Services independent of router technology.
2. Transport layer shielded from number, type, topology of routers.
3. Network addresses available to transport layer use uniform numbering plan
   - even across LANs and WANs

Implementation of Connectionless Service



A's table (initially)

Dest.	Line
A	
B	B
C	C
D	B
E	C
F	C

A's table (later)

A	
B	B
C	C
D	B
E	B
F	B

C's Table

A	A
B	A
C	
D	E
E	E
F	E

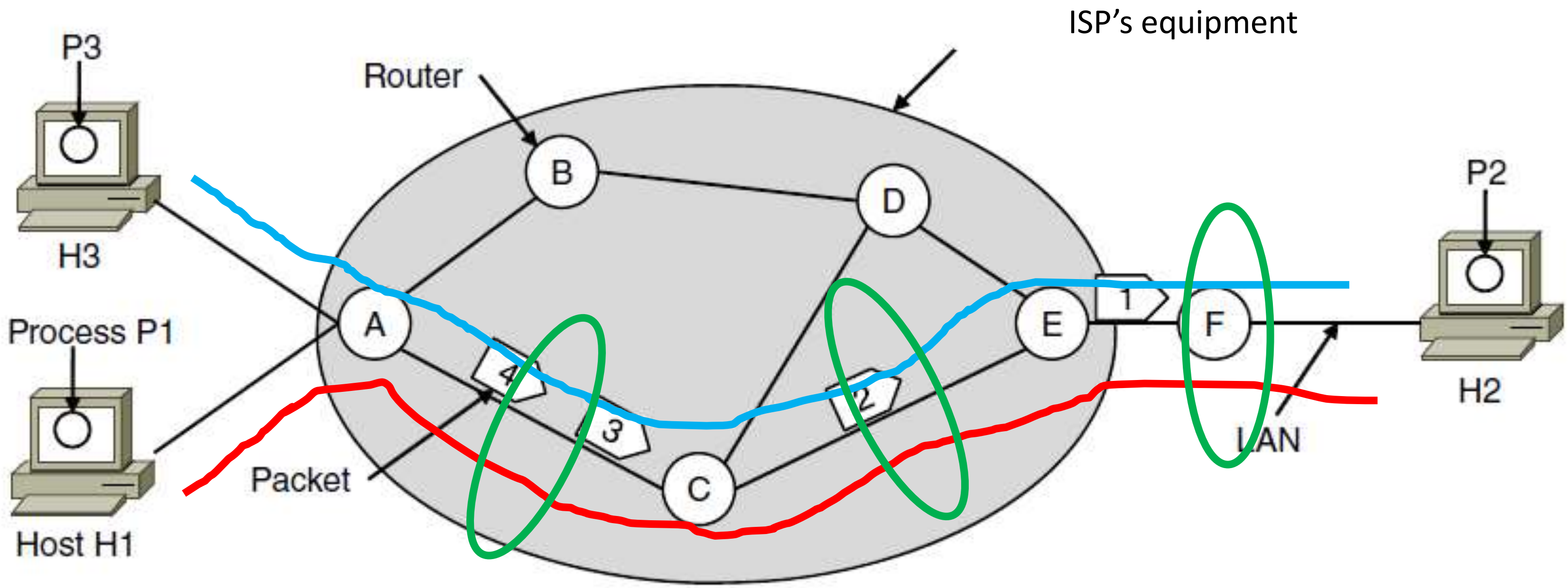
E's Table

A	C
B	D
C	C
D	D
E	
F	F

管理路由表并做出路由选择的就是路由算法

Routing within a datagram network

Implementation of Connection-Oriented Service



A's table

In		Out	
H1	1	C	1
H3	1	C	2

C's Table

A	1	E	1
A	2	E	2

E's Table

C	1	F	1
C	2	F	2

局部分配，标签交换

Routing within a virtual-circuit network

Comparison of Virtual-Circuit and Datagram Networks

权衡
- 建立时间
- 地址解析时间
- 表空间的数量
- 脆弱性

Issue	Datagram network	Virtual-circuit network
Circuit setup	Not needed	Required
Addressing	Each packet contains the full source and destination address	Each packet contains a short VC number
State information	Routers do not hold state information about connections	Each VC requires router table space per connection
Routing	Each packet is routed independently	Route chosen when VC is set up; all packets follow it
Effect of router failures	None, except for packets lost during the crash	All VCs that passed through the failed router are terminated
Quality of service	Difficult	Easy if enough resources can be allocated in advance for each VC
Congestion control	Difficult	Easy if enough resources can be allocated in advance for each VC

5.2 Routing Algorithms in a Single Network (1)

- Optimality principle
- Shortest path algorithm
- Flooding
- Distance vector routing
- Link state routing
- Routing in ad hoc networks

Routing Algorithms in a Single Network（2）

- Broadcast routing
- Multicast routing
- Anycast routing
- Routing for mobile hosts
- Routing in ad hoc networks

Routing Algorithms in a Single Network（3）

- **Main issue:** Routers that constitute the network layer of a network, should cooperate to find the **best routes** between all pairs of stations

跳数，是延时最小和吞吐量最大的一个折衷

- Properties of Routing Algorithm

- **正确性（correctness）**
- **简单性（simplicity）**
- **健壮性（robustness）**
- **稳定性（stability）**
- **公平性（fairness）**
- **最优性 有效性（optimality）**

Routing Algorithms in a Single Network（4）

- Two major classes of routing algorithms
  - Non-adaptive algorithms(static routing): route is computed in advanced, off-line, and downloaded to routers, not depends on current traffic or topology.

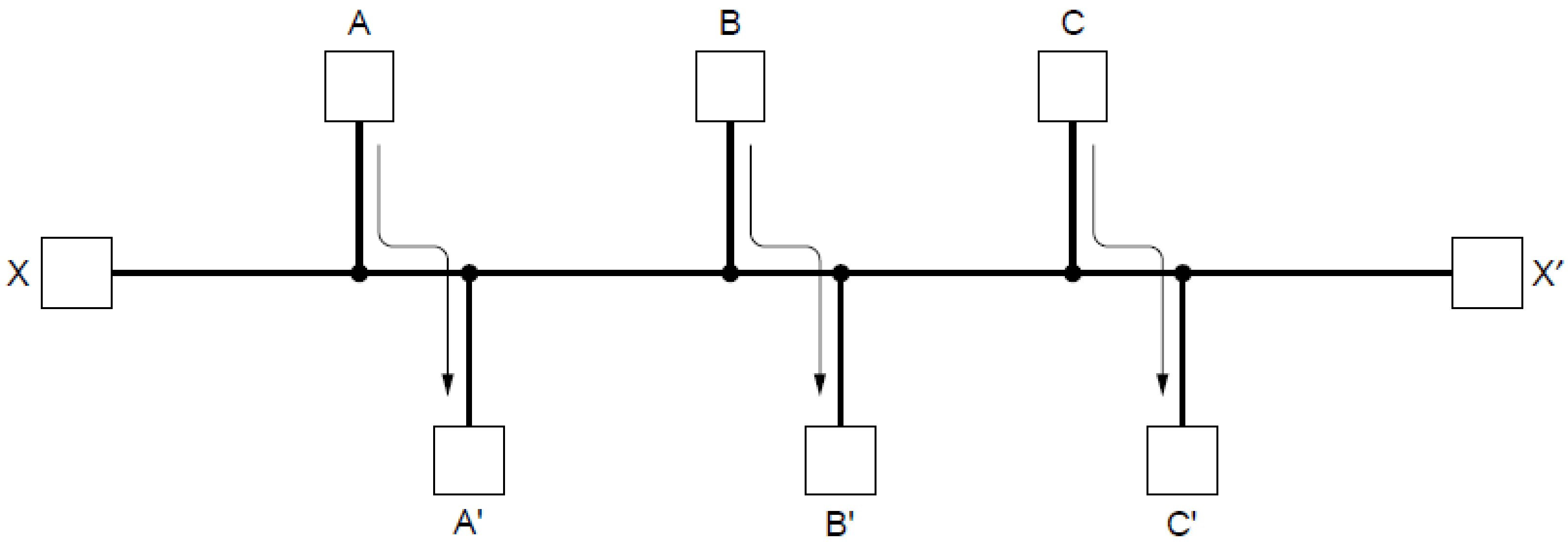
    简单，开销小；灵活性差

  - Adaptive algorithms(dynamic routing): route is computed dynamically, on-line, according to current traffic and network topology

    开销大；健壮性、灵活性好

Routing Algorithms in a Single Network（5）

Fairness vs. Efficiency



Network with a conflict between fairness and efficiency.

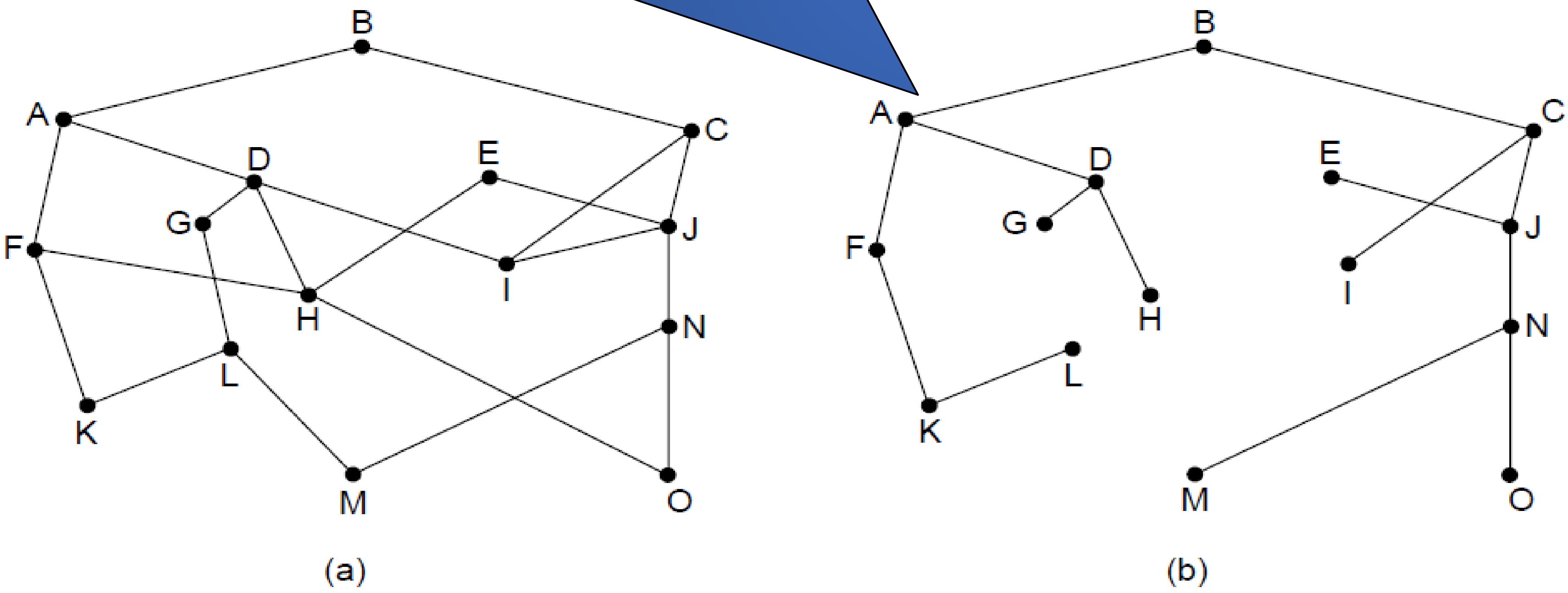
The Optimality Principle

- One can make a general statement about optimal routes without regard to network topology or traffic: if router *J* is on the optimal path from router *I* to router *K*, then the optimal path from *J* to *K* also falls along on the same route.

- **Sink tree**
  - The set of optimal routes from all source station to a given destination forms a tree: **sink tree**

The Optimality Principle

This means: Routers have to collaborate to build the sink tree (or something that comes near to that) for each source station.



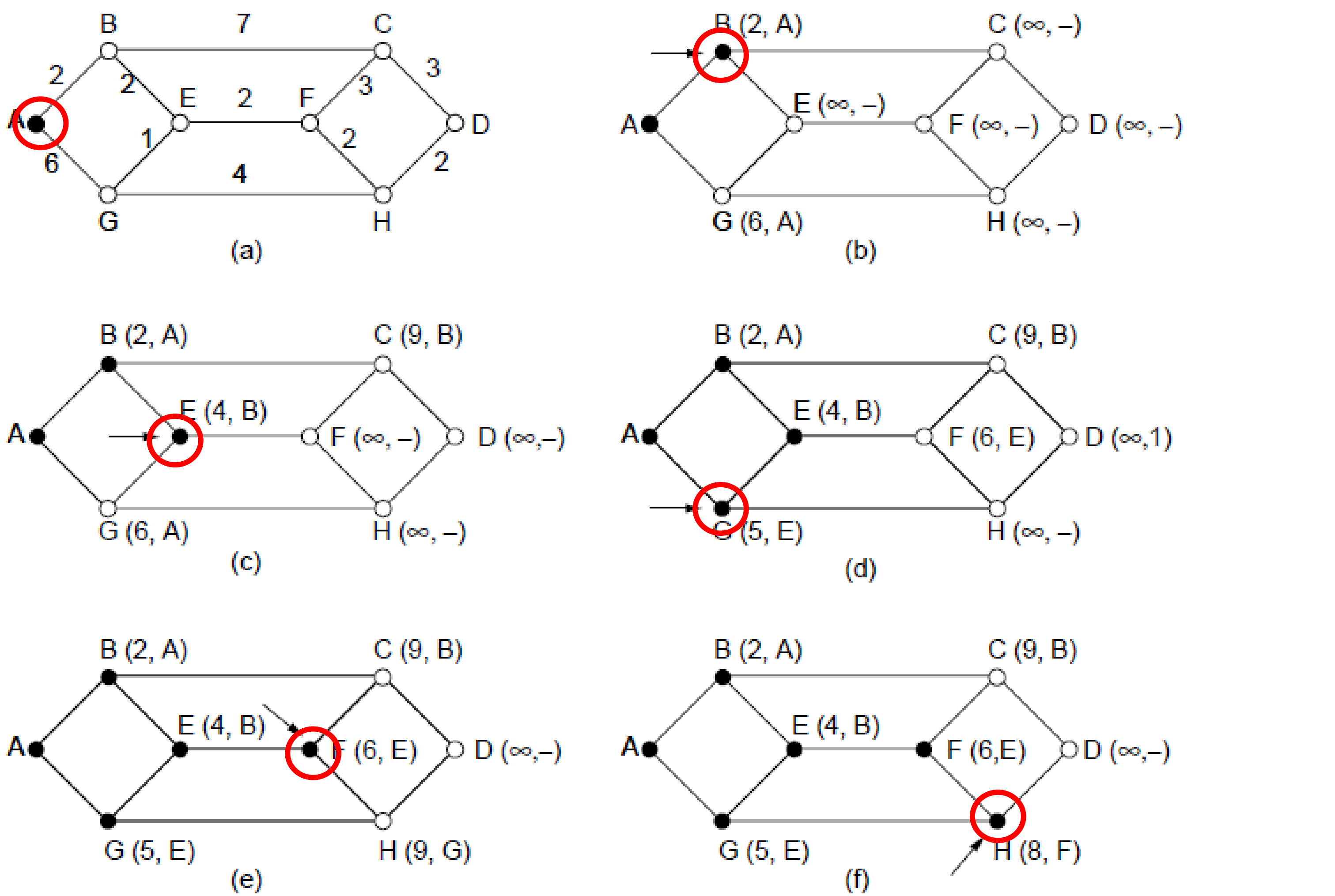
(a) A network. (b) A sink tree for router *B.*

Shortest Path Algorithm (1)

- **Basic idea:** During each step, select a newly reachable node at the lowest cost, and add the edge to that node, to the tree built so far.

给定一个完整的网络视图，可以用来计算最优路径

Shortest Path Algorithm (1)



The first five steps used in computing the shortest path from *A to D.*
The arrows indicate the working node

Shortest Path Algorithm (2)

Edsger Wybe Dijkstra

1 提出“goto有害论”；

2 提出信号量和PV原语;

3 解决了有趣的“哲学家聚餐”问题;

4 最短路径算法(SPF)和银行家算法的创造者;

5 第一个Algol 60编译器的设计者和实现者;

6 THE操作系统的设计者和开发者;

与D. E. Knuth并称为我们这个时代最伟大的计算机科学家的人。

在与癌症进行了多年的斗争之后，伟大的荷兰计算机科学家Edsger Wybe Dijkstra已经于2002年8月6日在荷兰Nuenen自己的家中与世长辞！享年72岁。

Flooding

- **Basic idea:** Forward an incoming packet across every outgoing line, except the one it came through.

好处？

鲁棒性好，延时少，适合广播（无线网络）

大量重复包

- **Basic problem:** how to avoid “drowning by packets”?
  - **Use a hop counter**: after a packet has been forwarded across *N* routers, it is discarded. Got to find the right hop count, though.

    计数器，每经过一站计数器减1，为0时则丢弃该包

  - Be sure to forward a packet only once (i.e. avoid directed cycles). Requires sequence numbers per source router. Each router keeps track of the **last sequence number per source router.**

    记录包经过的路径

Distance Vector Routing

Basic idea: Take a look at the costs that your **direct neighbors** are advertising to get a packet to the destination. Select the neighbor whose advertised cost, added with the cost to get to that neighbor, is the lowest. Advertise that new cost to the other neighbors.

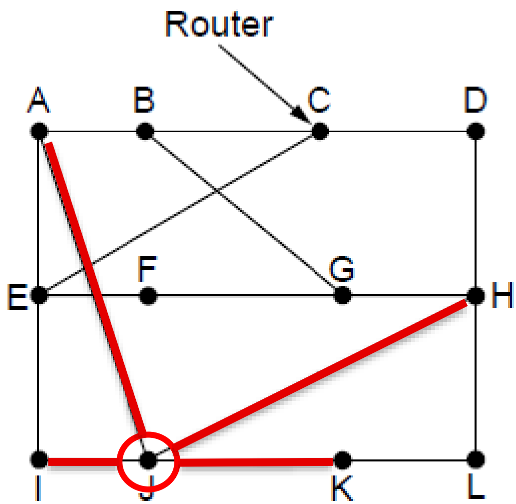
Ex:

Neighbor:	R_1	R_2	R_3
Link cost:	12	8	5
Advertised:	28	25	39
Total:	40	33	44

我只跟邻居沟通，但是想知道的是全局的，**道听途说**

A-（R1，R2，R3）-B

Distance Vector Routing



(a)

New estimated delay from J

To	A	I	H	K		Line
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22		
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	6	K
L	29	33	9	9	15	K
	JA delay is 8	JI delay is 10	JH delay is 12	JK delay is 6	New routing table for J	

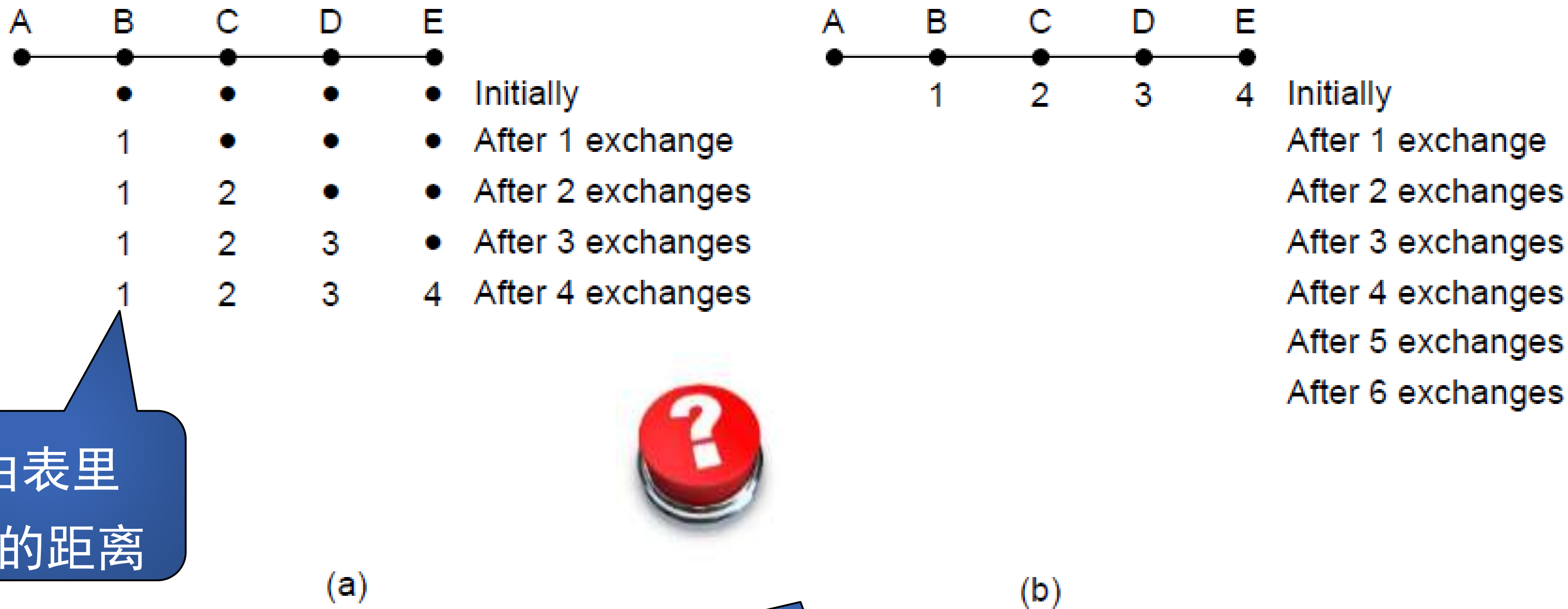
Vectors received from J's four neighbors

(b)

(a) A network.

(b) Input from *A, I, H, K, and the new routing table* for *J.*

The Count-to-Infinity Problem



路由表里
到A的距离

跳数限制；
防止路由器向邻居返回一个从该邻居获得的最佳路径。

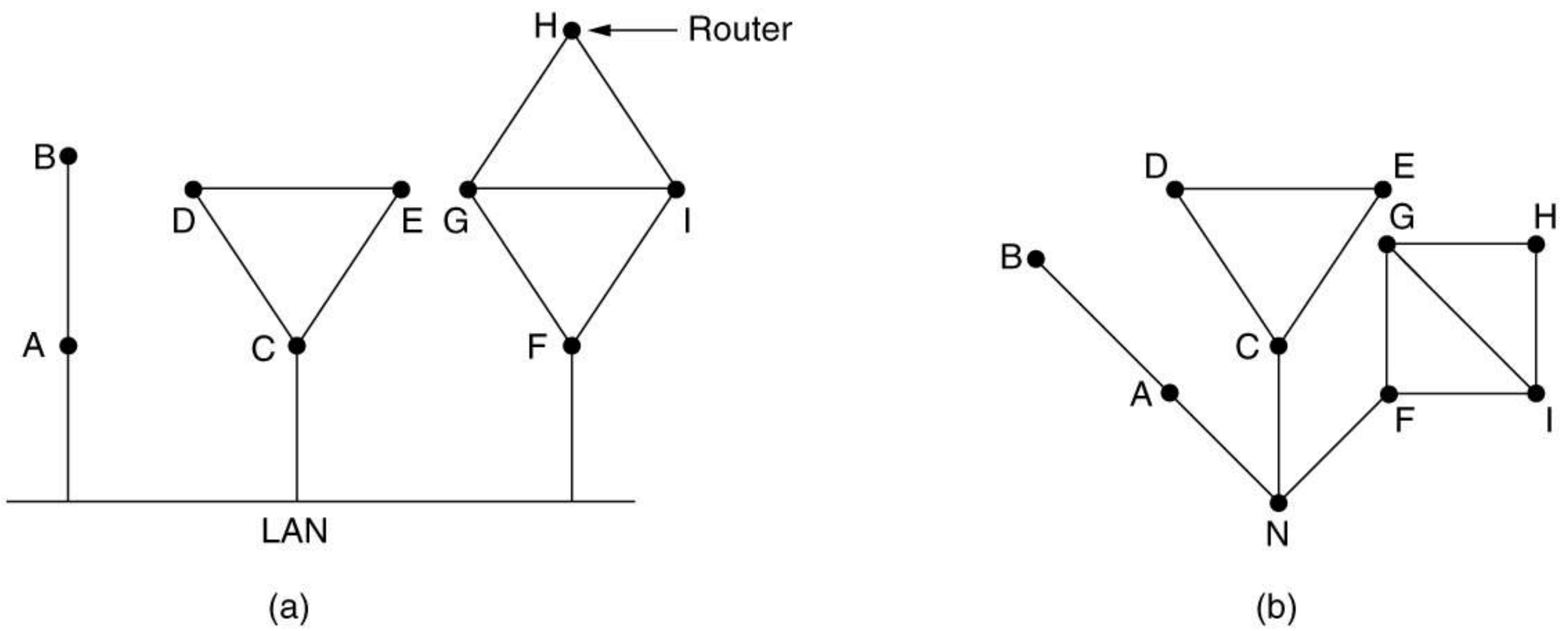
The count-to-infinity problem

Link State Routing

1. Discover neighbors, learn network addresses.
2. Set distance/cost metric to each neighbor.
3. Construct packet telling all learned.
4. Send packet to, receive packets from other routers.
5. Compute shortest path to every other router.

第一手资料发给所有人

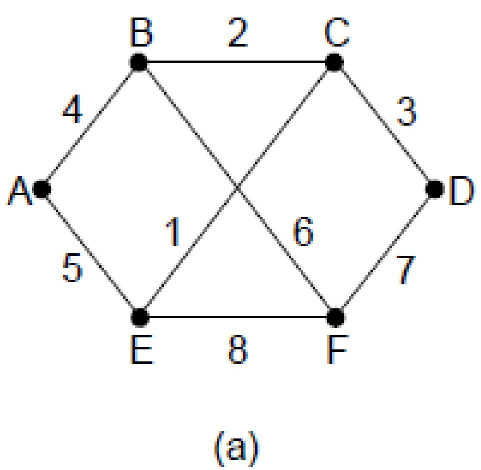
Learning about the Neighbors



(a) Nine routers and a broadcast LAN. (b) A graph model of (a).

Building Link State Packets

- just put in a sequence number and aging information. The hard part is when to build them. Practice shows that once an hour is often enough.



Link State Packets

A	B	C	D	E	F
Seq.	Seq.	Seq.	Seq.	Seq.	Seq.
Age	Age	Age	Age	Age	Age
B 4	A 4	B 2	C 3	A 5	B 6
E 5	C 2	D 3	F 7	C 1	D 7
	F 6	E 1		F 8	E 8

(b)

(a) A network. (b) The link state packets for this network.

Distributing the Link State Packets

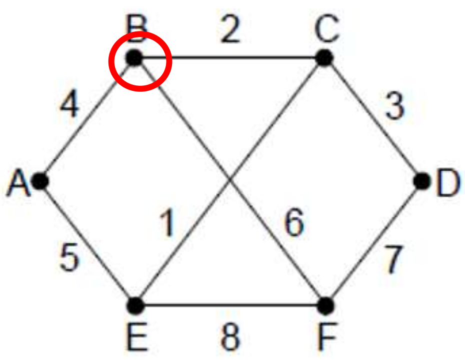
- **Basic idea:** use a flooding algorithm, and dam the flood through sequence numbers: all routers maintain a list of (source, seq. number)-pairs.

路由器崩溃，
序号破坏

- 00000000，00000000，000000000，00000100
- 10000000，00000000，000000000，00000100

  – To safeguard against old data, down links, etc., **an age** is added to an LSP. The age is decremented once a second, and every time it is forwarded by a router. When the age hits zero, the LSP is discarded.

Distributing the Link State Packets



- 从E发来的链路状态包有两个，一个经过EAB，另一个经过EFB；
- 从D发来的链路状态包有两个，一个经过DCB，另一个经过DFB；

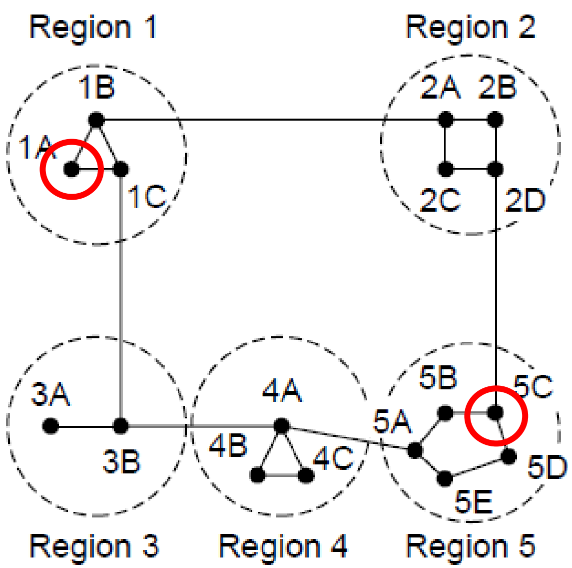
Source	Seq.	Age	Send flags A	Send flags C	Send flags F	ACK flags A	ACK flags C	ACK flags F	Data
A	21	60	0	1	1	1	0	0	
F	21	60	1	1	0	0	0	1	
E	21	59	0	1	0	1	0	1	
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

The packet buffer for router *B* in previous slide

Computing the New Routes

- Use Dijkstra's algorithm to construct the shortest path to all possible destinations.

Hierarchical Routing



1A-5C

(a)

Full table for 1A

Dest.	Line	Hops
1A	–	–
1B	1B	1
1C	1C	1
2A	1B	2
2B	1B	3
2C	1B	3
2D	1B	4
3A	1C	3
3B	1C	2
4A	1C	3
4B	1C	4
4C	1C	4
5A	1C	4
5B	1C	5
5C	1B	5
5D	1C	6
5E	1C	5

(b)

Hierarchical table for 1A

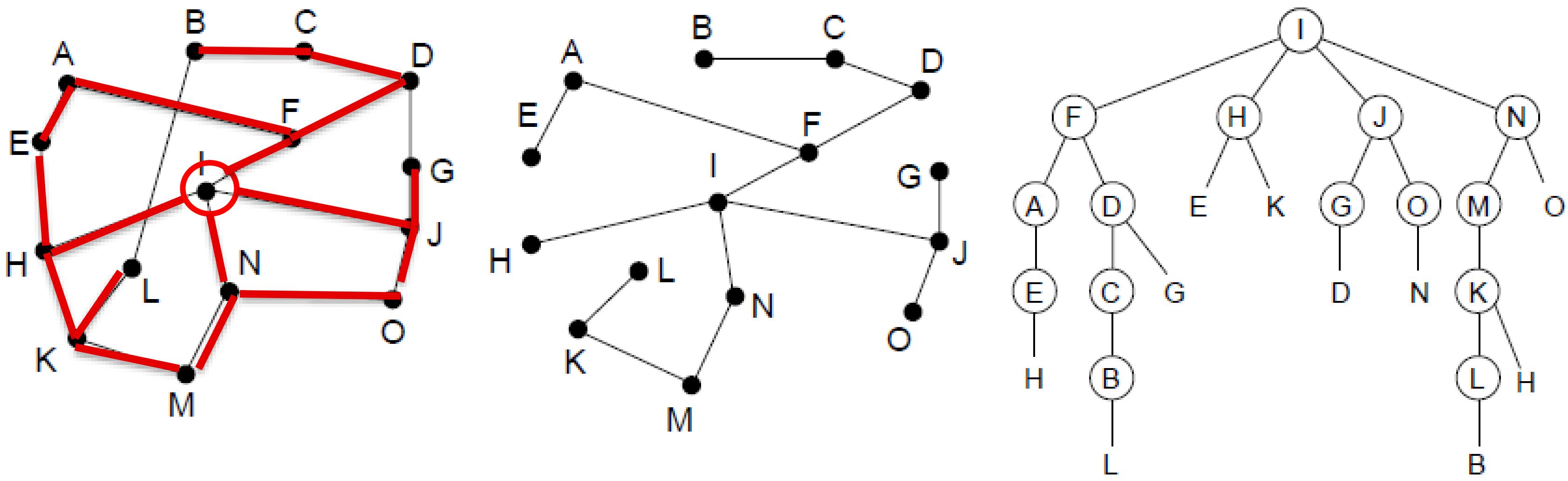
Dest.	Line	Hops
1A	–	–
1B	1B	1
1C	1C	1
2	1B	2
3	1C	2
4	1C	3
5	1C	4

(c)

- 对于一个包含N个路由器的网络，最优的层数是ln N

Hierarchical routing.

Broadcast Routing



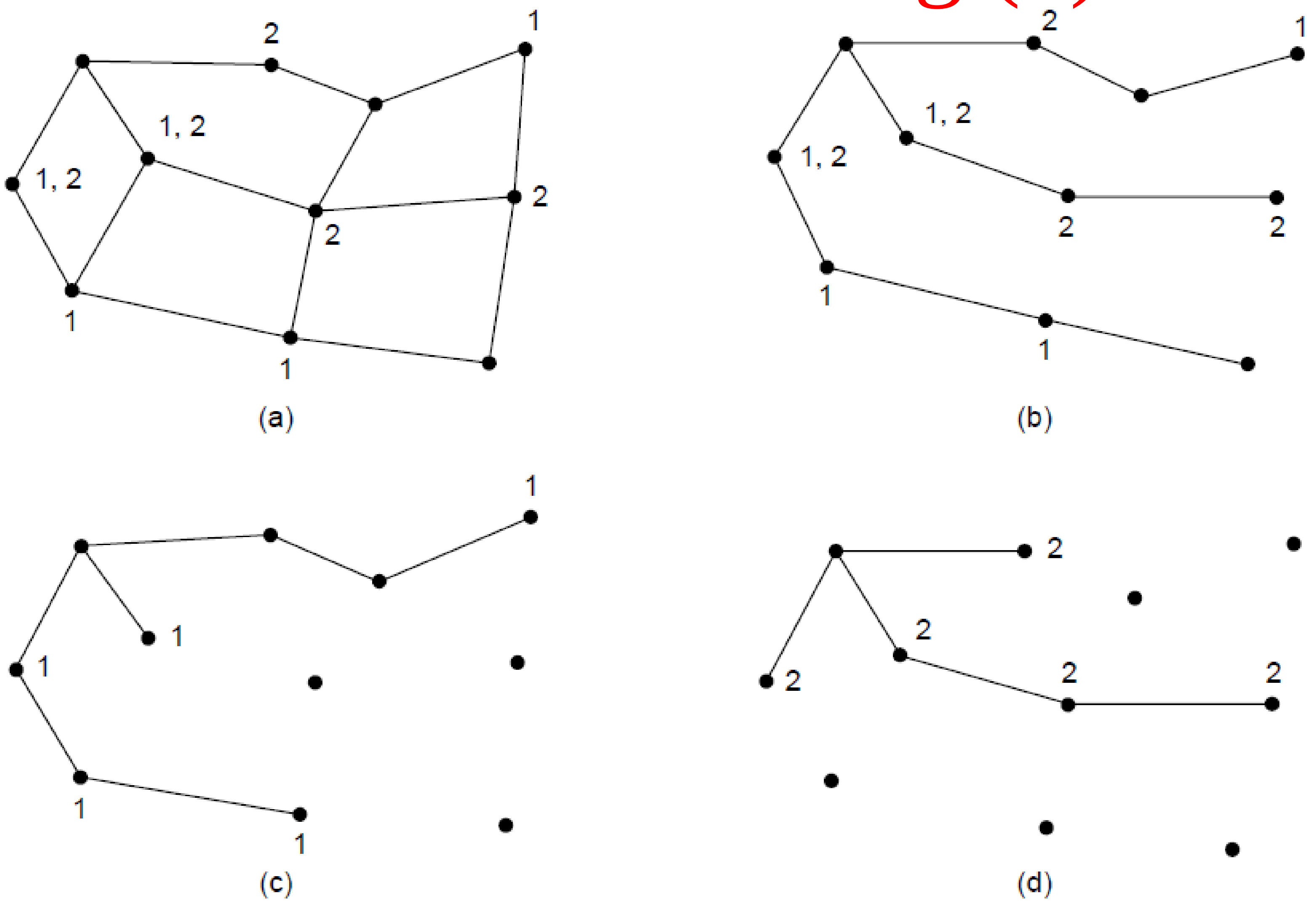
(a) (b) (c)

路由器检查它到来的那条线路是否是通常用来给广播源发数据包的那条线路。 是，继续广播转发。不是，丢弃。

Reverse path forwarding. (a) A network. (b) A sink tree.
(c) The tree built by reverse path forwarding.

逆向路径转发

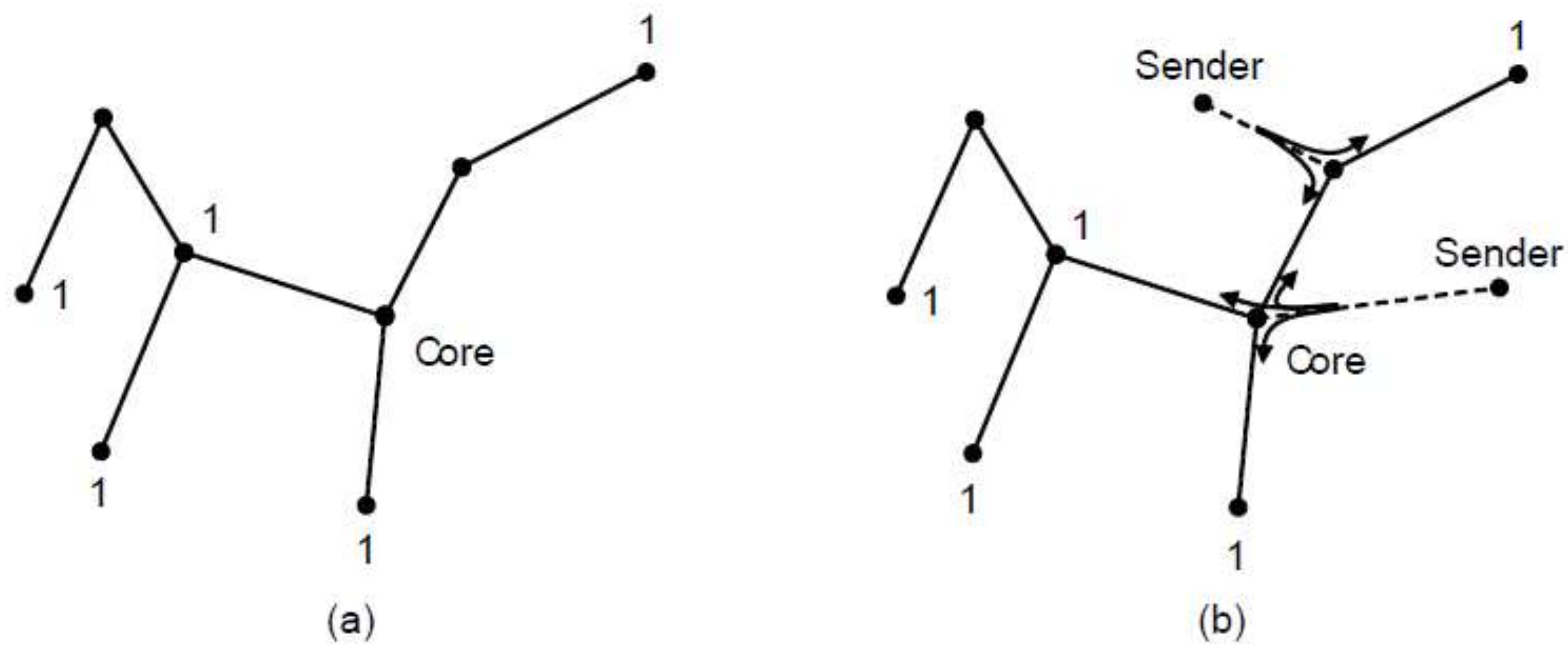
Multicast Routing (1)



(a) A network. (b) A spanning tree for the leftmost router. (c) A multicast tree for group 1. (d) A multicast tree for group 2.

Multicast Routing (2)

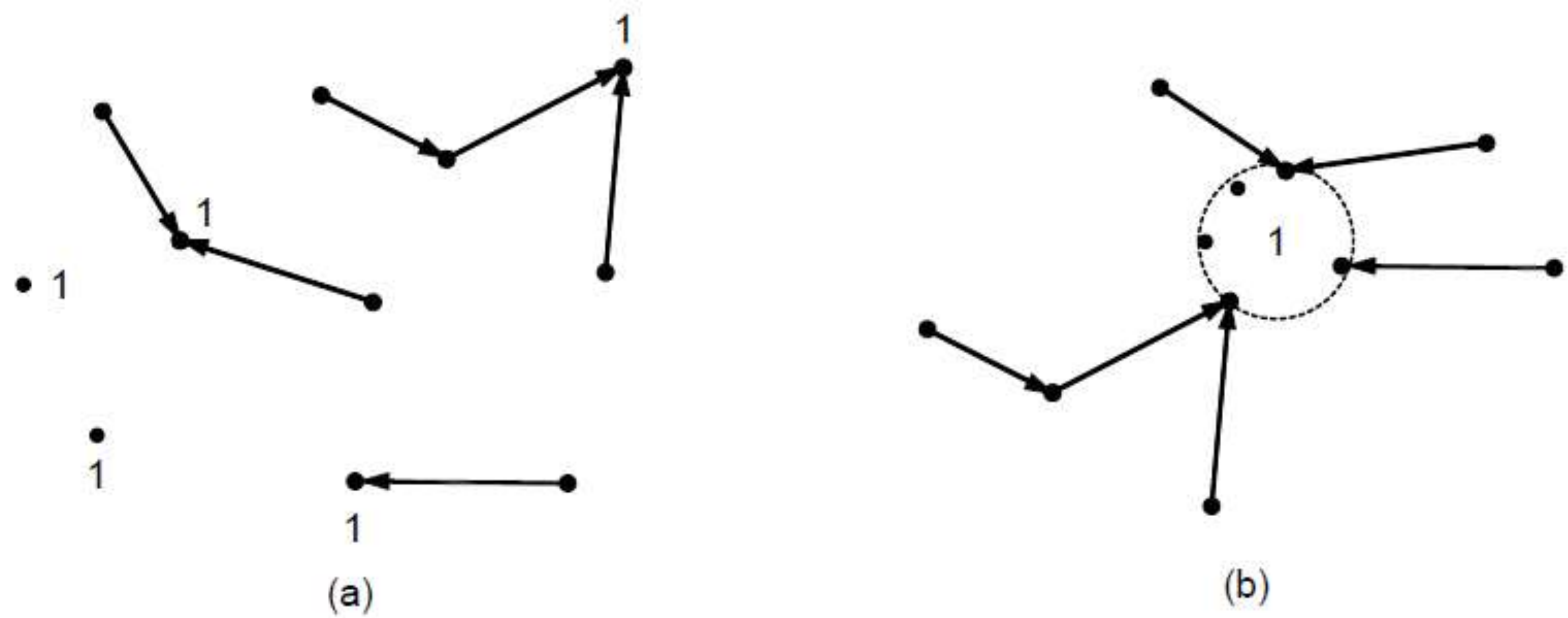
无法达到最优
只存一棵树，节省存储开销、消息发送和计算



(a) Core-based tree for group 1.

(b) Sending to group 1.

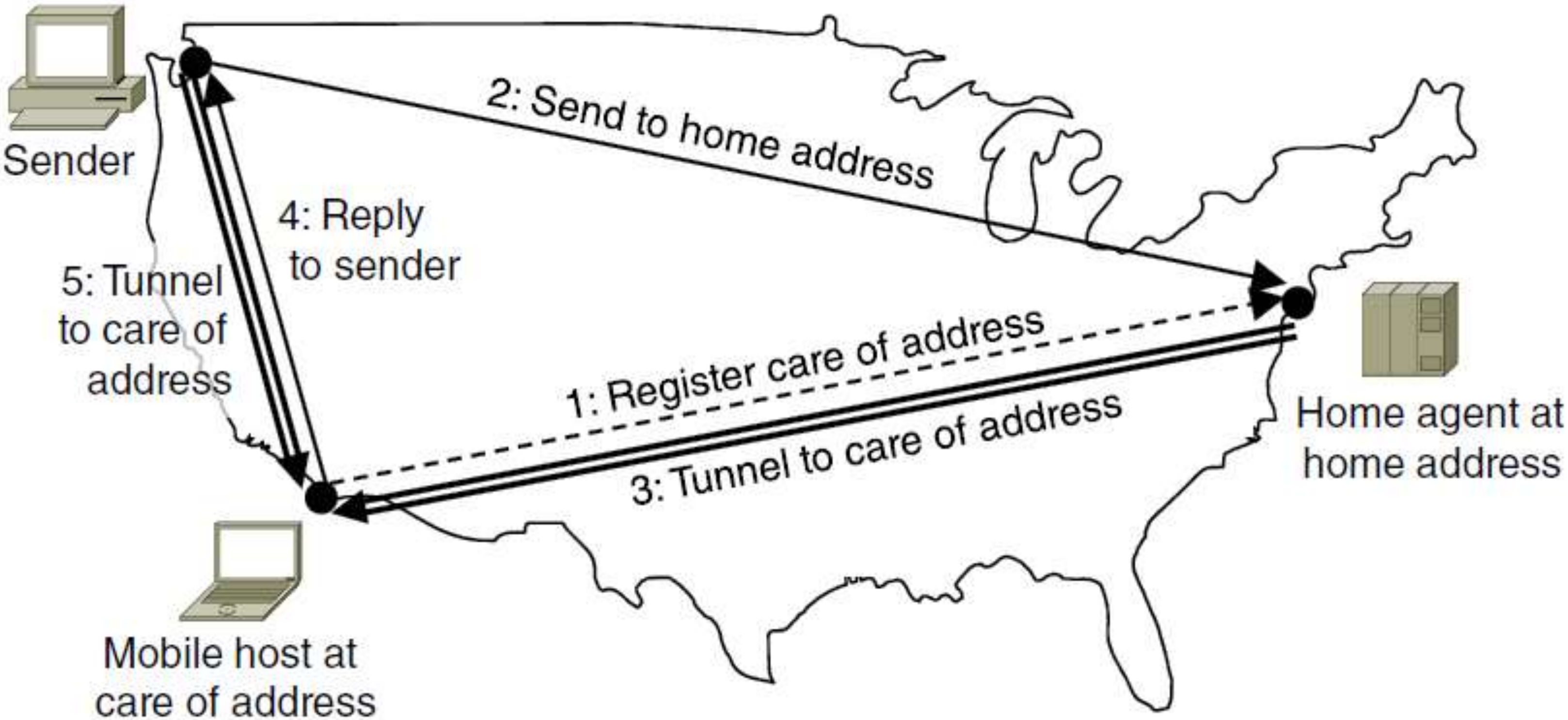
Anycast Routing



(a) Anycast routes to group 1.

(b) Topology seen by the routing protocol.

Routing for Mobile Hosts（补）

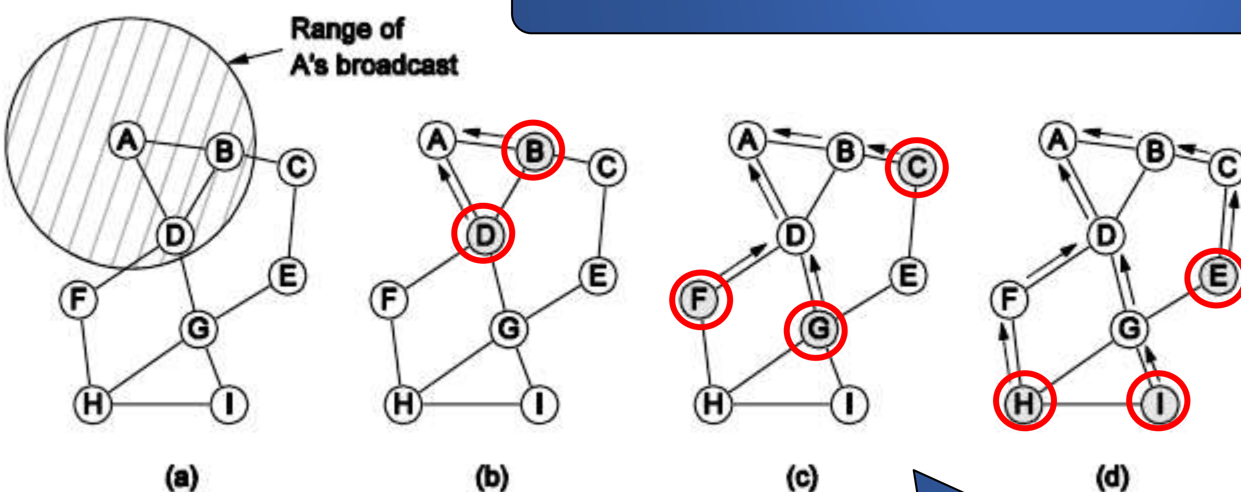


转交地址

Packet routing for mobile hosts

Routing in Ad Hoc Networks （补）

AODV ad hoc 按需距离矢量路由算法



(a) Range of A’s broadcast.

(b) After B and D receive it.

(c) After C, F, and G receive it.

(d) After E, H, and I receive it.

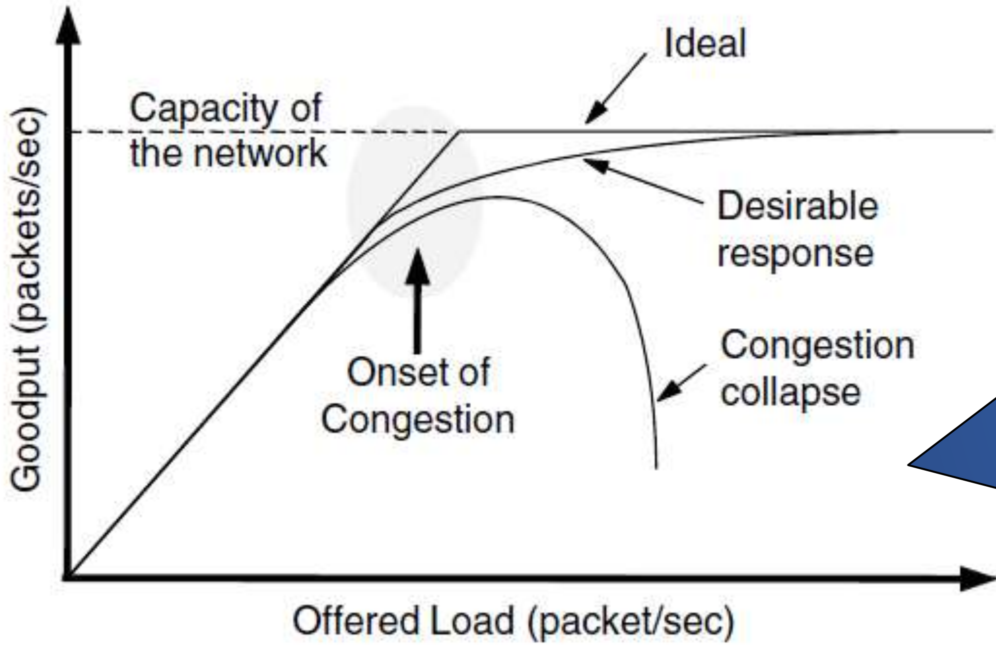
路由器都是动态的，网络拓扑没有意义

- 按需发现：路由请求包，泛洪，带序号；节点I发路由应答包
- 路由维护：周期Hello

5.3 Traffic Management at the Network Layer

- The need for traffic management: congestion
- Approaches to traffic management
  - Traffic-aware routing
  - Admission control
  - Load shedding
  - Traffic shaping
  - Active queue management
  - Random early detection
  - Choke packets
  - Explicit congestion notification
  - Hop-by-hop backpressure

Congestion (1)



全局拥塞

局部拥塞

- 多个输入对应一个输出；
- 慢速处理器；
- 低带宽线路。

When too much traffic is offered, congestion sets in and performance degrades sharply.

Congestion (2)

- **拥塞控制（congestion control）**
  - **需要确保通信子网能够承载用户提交的通信量，是一个全局性问题，涉及主机、路由器等很多因素；**
- **流量控制（flow control）**
  - **与点到点的通信量有关，主要解决快速发送方与慢速接收方的问题，是局部问题，一般都是基于反馈进行控制的。**

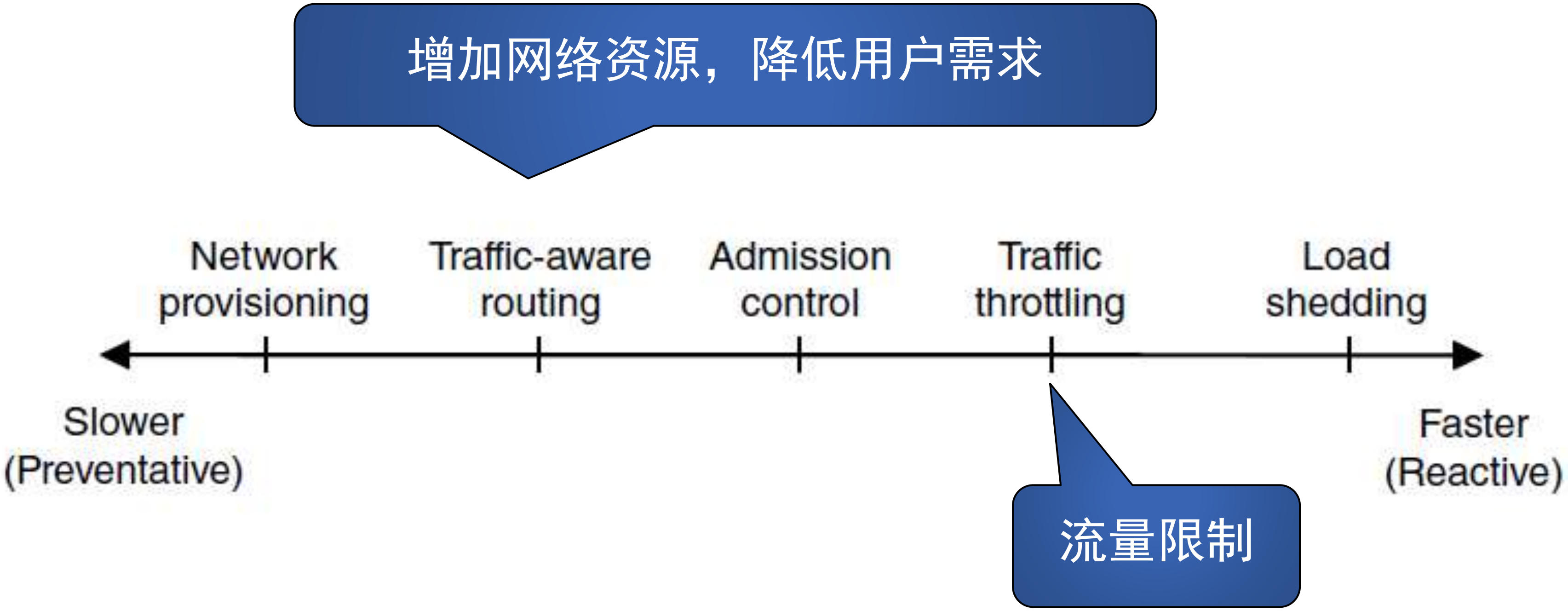
Congestion (3)

Layer	Policies
Transport	• Retransmission policy • Out-of-order caching policy • Acknowledgement policy • Flow control policy • Timeout determination
Network	• Virtual circuits versus datagram inside the subnet • Packet queueing and service policy • Packet discard policy • Routing algorithm • Packet lifetime management
Data link	• Retransmission policy • Out-of-order caching policy • Acknowledgement policy • Flow control policy

Congestion (4)

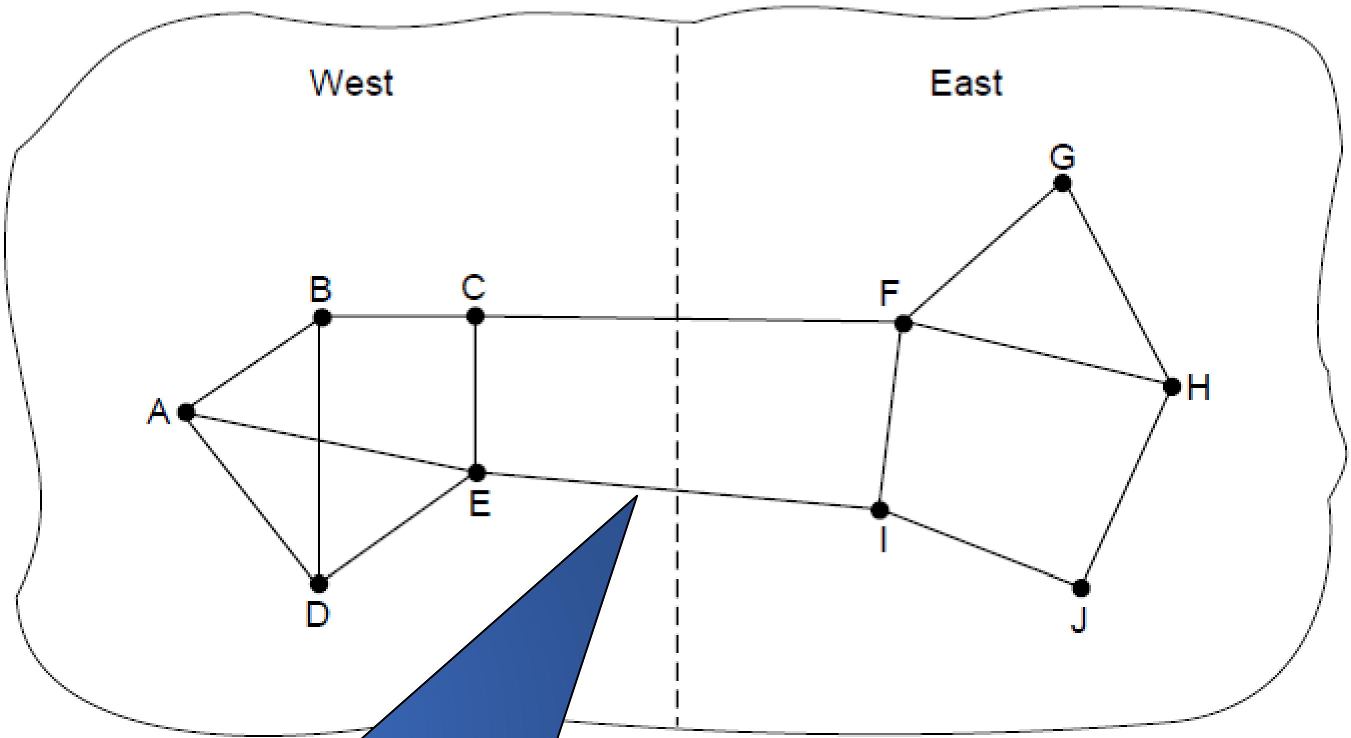
- Monitor the system to detect when and where congestion occurs
  - Percentage of all packets discarded for lack of buffer space
  - The average queue lengths
  - The number of packets that time out and are retransmitted
  - The average packet delay
- Pass information to places where action can be taken
  - Transfer the information about the congestion from the point where it is detected to the point where something can be done
- Adjust system operation to correct the problem
  - Increase the resources or decrease the load

Approaches to Traffic Management



Timescales of approaches to traffic and congestion management

Traffic-Aware Routing

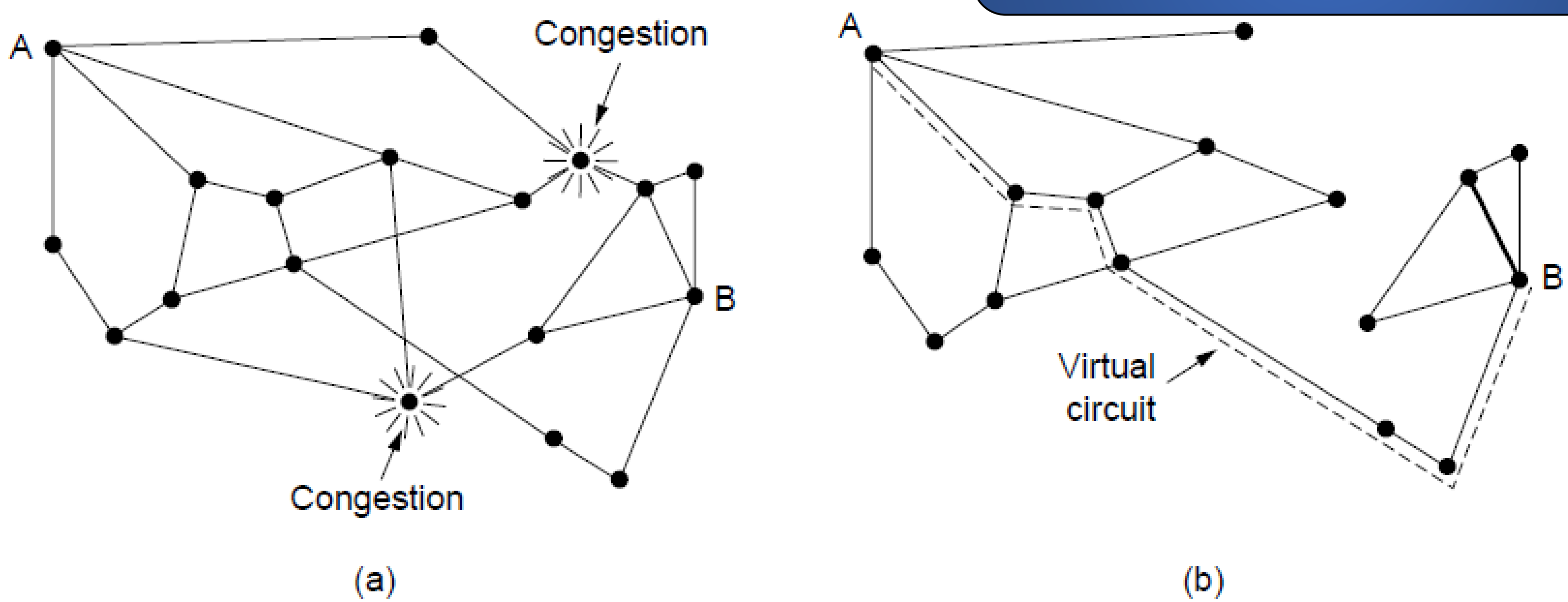


- 避免震荡
- 流量慢慢迁移，多径路由

A network in which the East and West parts are connected by two links.

Traffic-Aware Routing
Admission Control

应用于虚电路，可以携带额外的流量而不会变的拥塞



(a) A congested network. (b) The portion of the network that is not congested. A virtual circuit from A to B is also shown.

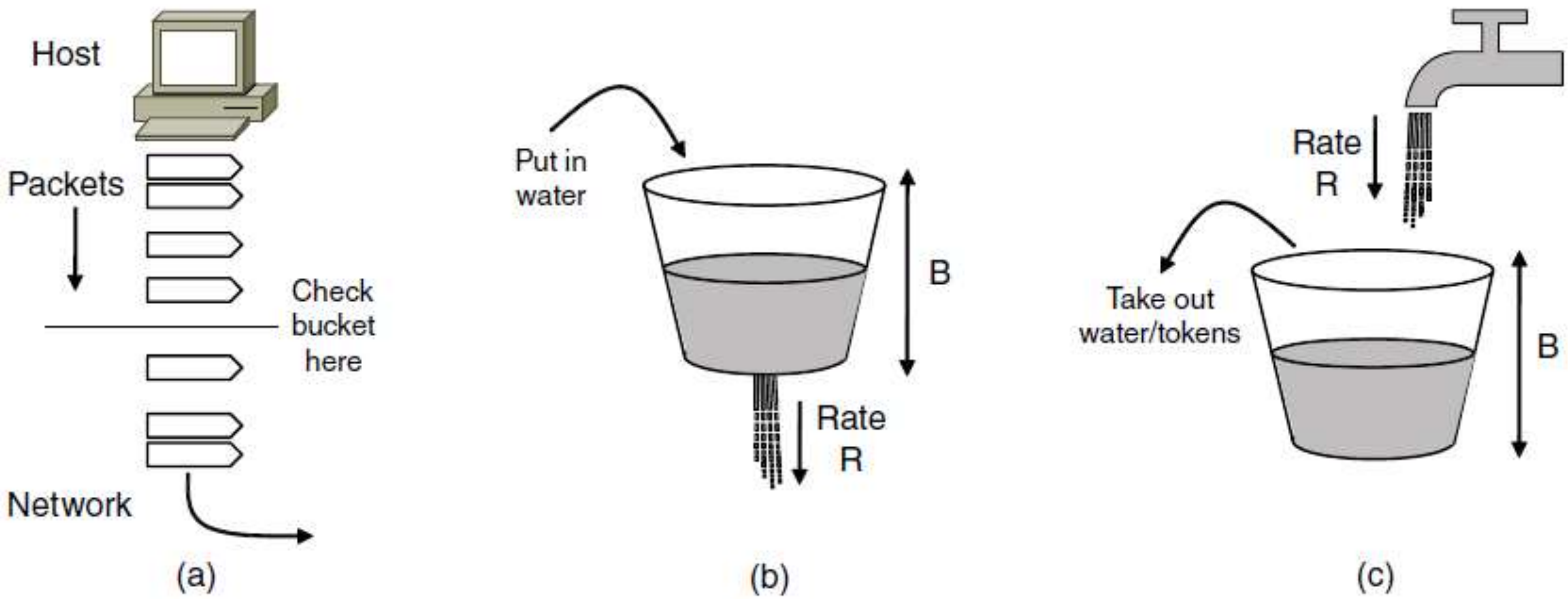
Load Shedding (1)

- **Wine-drop tail**
- **Milk**

1, 被动的丢包

Traffic Shaping (1)

- 造成拥塞的主要原因是网络流量通常是突发性的；
- 强迫包以一种可预测的速率发送；



(a) Shaping packets. (b) A leaky bucket. (c) A token bucket

Traffic Shaping (2)

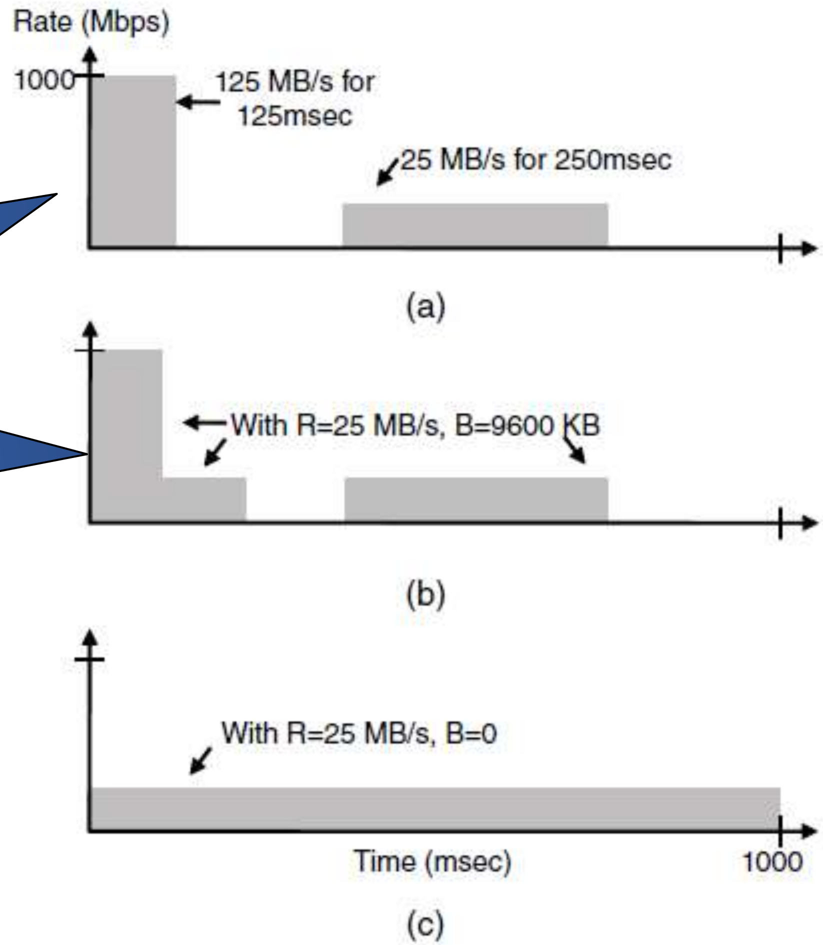
- **流量整形策略不同：漏桶算法不允许空闲主机积累发送权，以便以后发送大的突发数据；令牌桶算法允许，最大为桶的大小。**
- **漏桶中存放的是数据包，桶满了丢弃数据包；令牌桶中存放的是令牌，桶满了丢弃令牌，不丢弃数据包。**

漏桶规范到网络的流量； 但有时候，情况没那么糟，如路由器可以接收突发流量。

Traffic Shaping (3)

1000Mbps*0.125s = 125Mb=15.625MB= 16000KB

假设令牌桶快速注满，可以突发取9600KB，再按R来取。

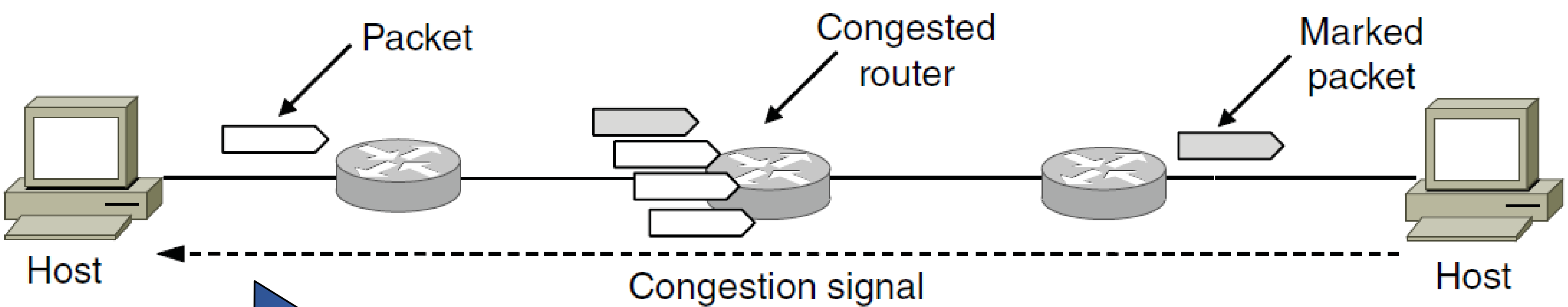


(a) Traffic from a host. Output shaped by a token bucket of rate 200 Mbps and capacity (b) 9600 KB, (c) 0 KB.

Explicit Congestion Notification

- 不一定要拥塞了才丢
- 2，Red通过丢包来通知主机拥塞了。

除了被动丢包，
如何通知源端？



3,在数据包和确认中打标记

ECN首选， ECN不可用的时候RED

Explicit congestion notification

Explicit Congestion Notification

Biography [edit]

Dr. Floyd received a BA in Sociology from the University of California - Berkeley in 1971. She received an MS in Computer Science in 1984 and a PhD in 1987, both from UC - Berkeley.[2]

Floyd is best known in the field of congestion control as the inventor of Random Early Detection ("RED") active queue management scheme, thus founding the field of Active Queue Management (AQM) with Van Jacobson.[1] Almost all Internet routers use RED or something developed from it to develop data paths between different networks.[1] Floyd devised the now-common method of adding jitter to message timers to avoid synchronization.[3]

Floyd, with Vern Paxon, in 1997 identified the lack of knowledge of network topology as the major obstacle in understanding how the Internet works.[4] This paper, "Why We Don't Know How to Simulate the Internet", was re-published as "Difficulties in Simulating the Internet" in 2001 and won the IEEE Communication Society's William R. Bennett Prize Paper Award.

Floyd is also a co-author on the standard for TCP Selective acknowledgement (SACK), Explicit Congestion Notification (ECN), the Datagram Congestion Control Protocol (DCCP) and TCP Friendly Rate Control (TFRC).

She received the IEEE Internet Award in 2005 and the ACM SIGCOMM Award in 2007 for her contributions to congestion control.[1] She has been involved in the Internet Advisory Board, and is one of the top-ten most cited researchers in computers science.[1]

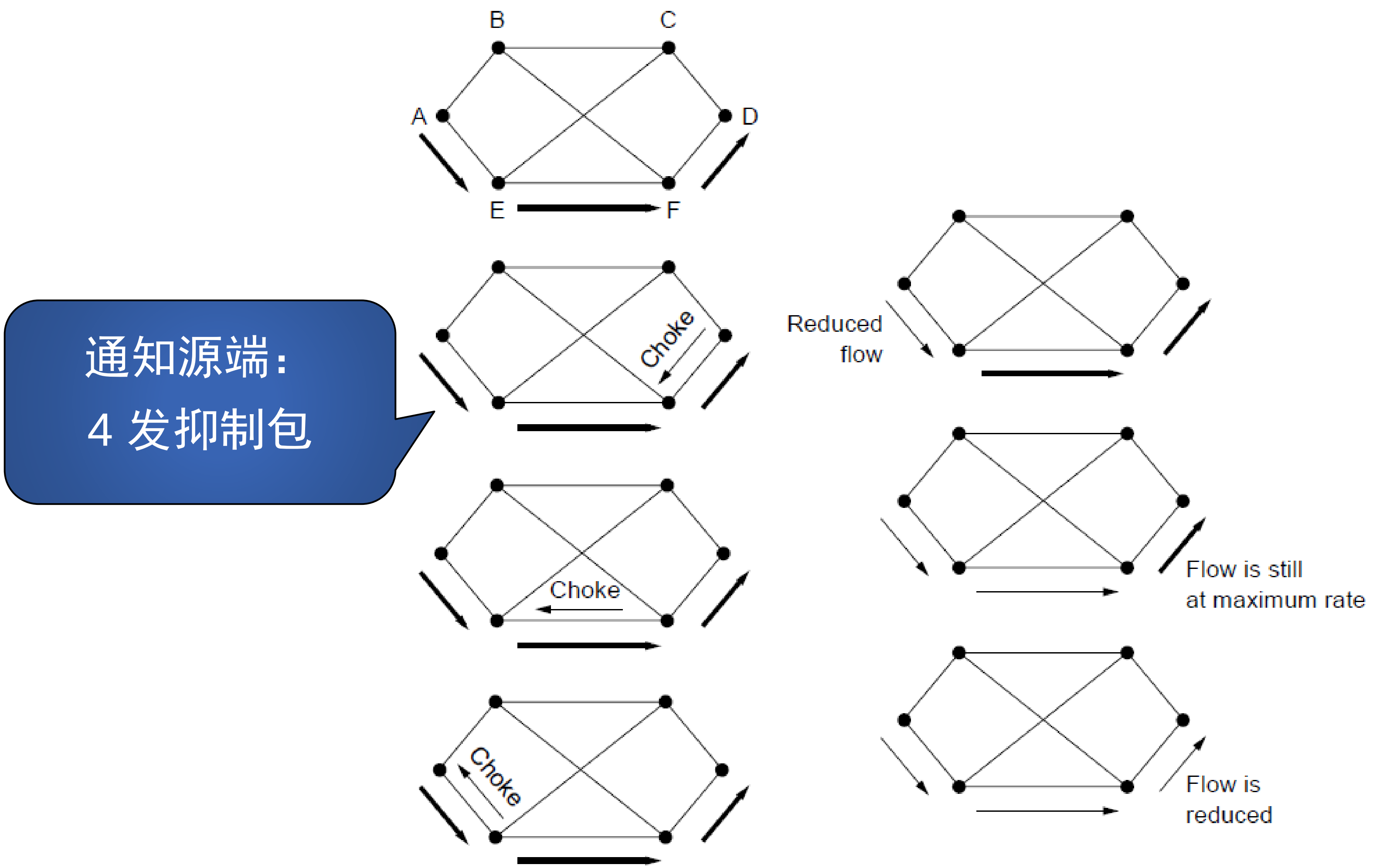
Sally Floyd - Information

- Email: floyd at acm.org, contact information, and my PGP key.
- ICSI, and the ICSI Networking Group.

Work:

- Papers (RFCs only), talks, informal notes, and travel.
- Sally's papers at: ACM Digital Library, Microsoft Academic Search, ResearchGate.
- Resume and short biography.
- Research projects (e.g., DCCP, ECN, HighSpeed TCP, Models, NS-2 [NS-3], Quick-Start, RED, RED-PD, TBIT (TCP behavior), TFRC, TMRG).
- Past professional activities (e.g., in the IAB, IETF, IRTF, SIGCOMM).
- Pointers to the literature (e.g., TCP, measurement studies of end-to-end congestion control).

Hop-by-Hop Backpressure

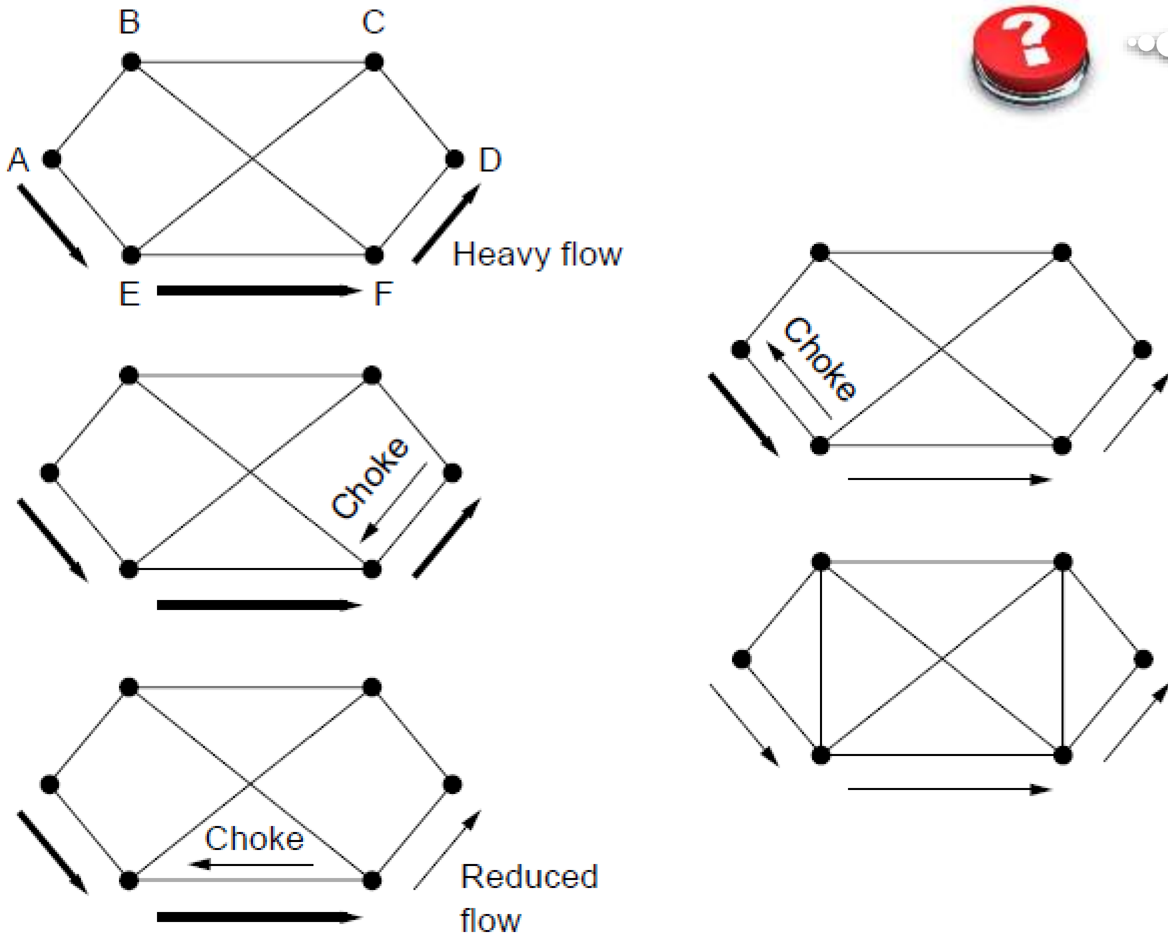


A choke packet that affects only the source..

Traffic Throttling

代价

上游路径需要消耗更多的缓冲区



A choke packet that affects each hop it passes through.

5.4 Quality of Service and Application QoE

- Application QoS requirements
- Overprovisioning
- Packet scheduling
- Integrated services
  - RSVP—The Resource reSerVation Protocol
- Differentiated services
  - Expedited forwarding
  - Assured forwarding

- 需要啥质量
- 如何规范进入网络的流量
- 如何在路由器预留资源
- 网络能否接受更多的流

e.g.交通、餐馆、食堂

确保服务质量
要解决的问题

Application Requirements

Application	Bandwidth	Delay	Jitter	Loss
Email	Low	Low	Low	Medium
File sharing	High	Low	Low	Medium
Web access	Medium	Medium	Low	Medium
Remote login	Low	Medium	Medium	Medium
Audio on demand	Low	Low	High	Low
Video on demand	High	Low	High	Low
Telephony	Low	High	High	Low
Videoconferencing	High	High	High	Low

How stringent the quality-of-service requirements are.

Categories of QoS and Examples

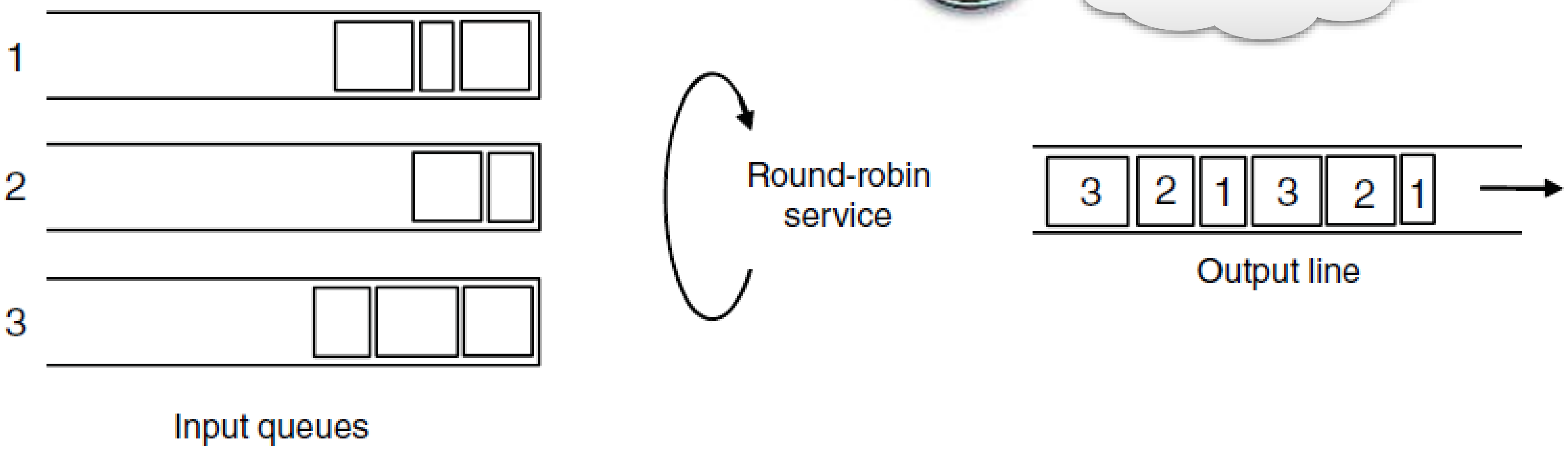
- Constant bit rate
  - Telephony
- Real-time variable bit rate
  - Compressed videoconferencing
- Non-real-time variable bit rate
  - Watching a movie on demand
- Available bit rate
  - File transfer

Packet Scheduling (1)

- 同一个流的数据包之间或者竞争流之间分配路由器资源的算法。

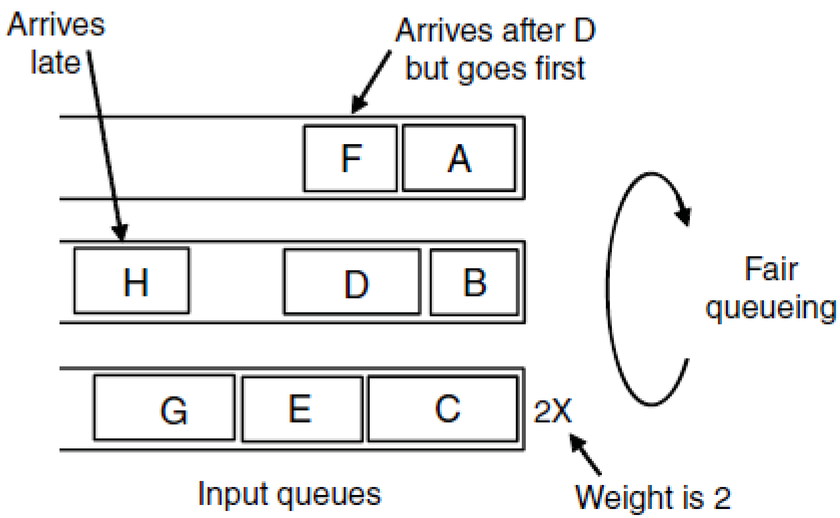
- Router resources reserved for different flows
  - Bandwidth
  - Buffer space
  - CPU cycles
- Algorithms
  - First-In First-Out (FIFO) scheduling
  - Fair queueing
  - Weighted fair queueing
  - Putting it together

Packet Scheduling (2)



Round-robin Fair Queuing

Packet Scheduling (3)



(a)

Packet	Arrival time	Length	Finish time	Output order
A	0	8	8	1
B	5	6	11	3
C	5	10	10	2
D	8	9	20	7
E	8	8	14	4
F	10	6	16	5
G	11	10	19	6
H	20	8	28	8

(b)

(a) Weighted Fair Queueing.

(b) Finishing times for the packets.

Putting it Together (1 of 2)

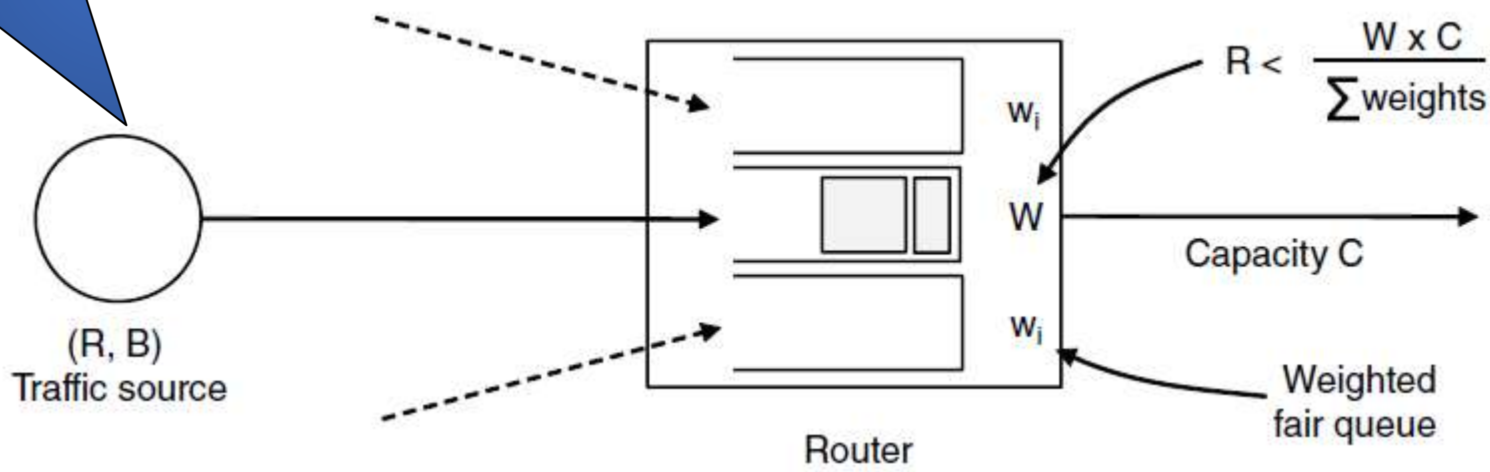
Parameter	Unit
Token bucket rate	Bytes/sec
Token bucket size	Bytes
Peak data rate	Bytes/sec
Minimum packet size	Bytes
Maximum packet size	Bytes

An example flow specification

Putting it Together (1 of 2)

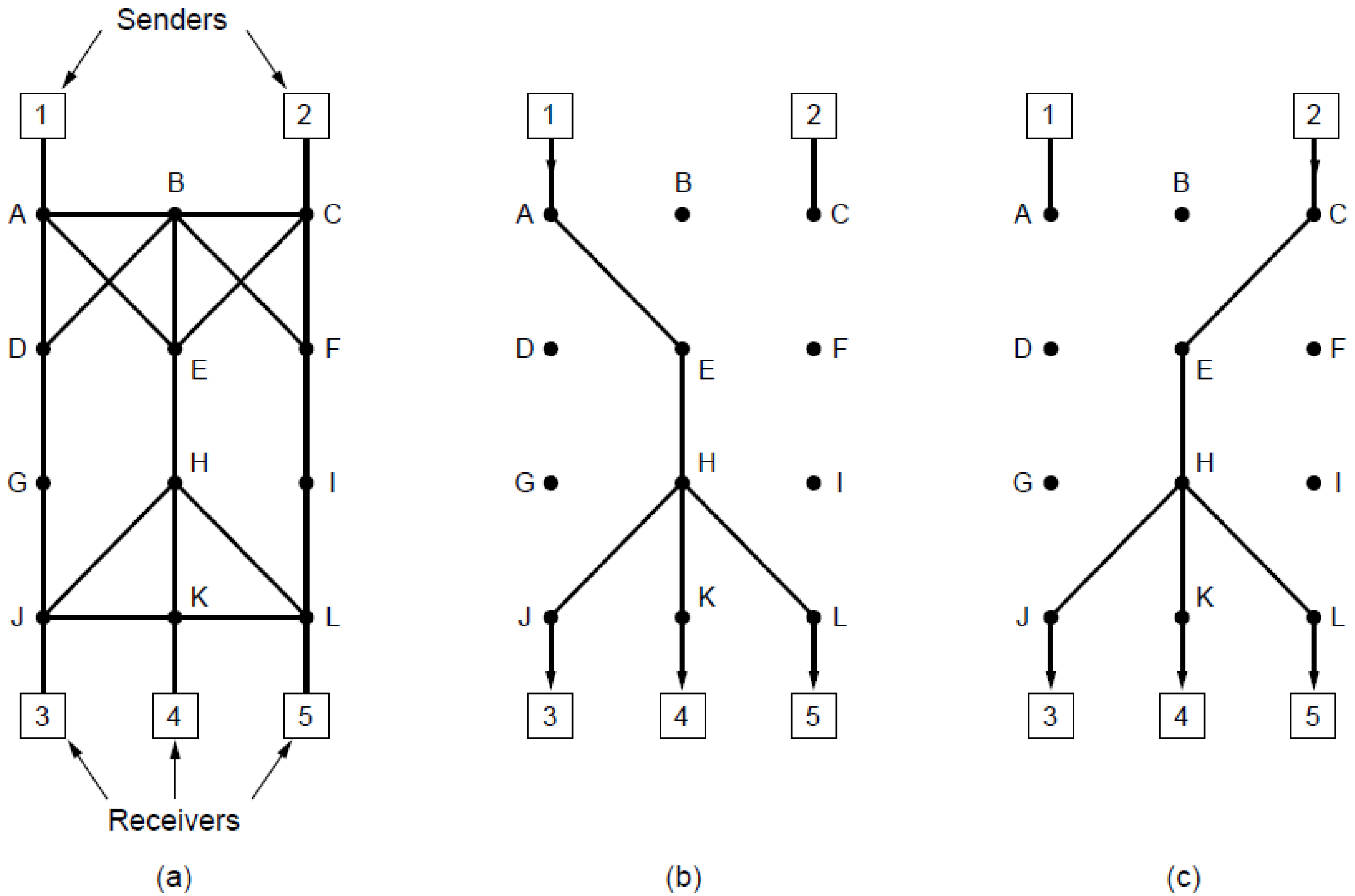
- 分的带宽比R大就可以满足带宽要求

- 最大的延迟 B/R， R只保障带宽。



Bandwidth and delay guarantees with token buckets and WFQ.

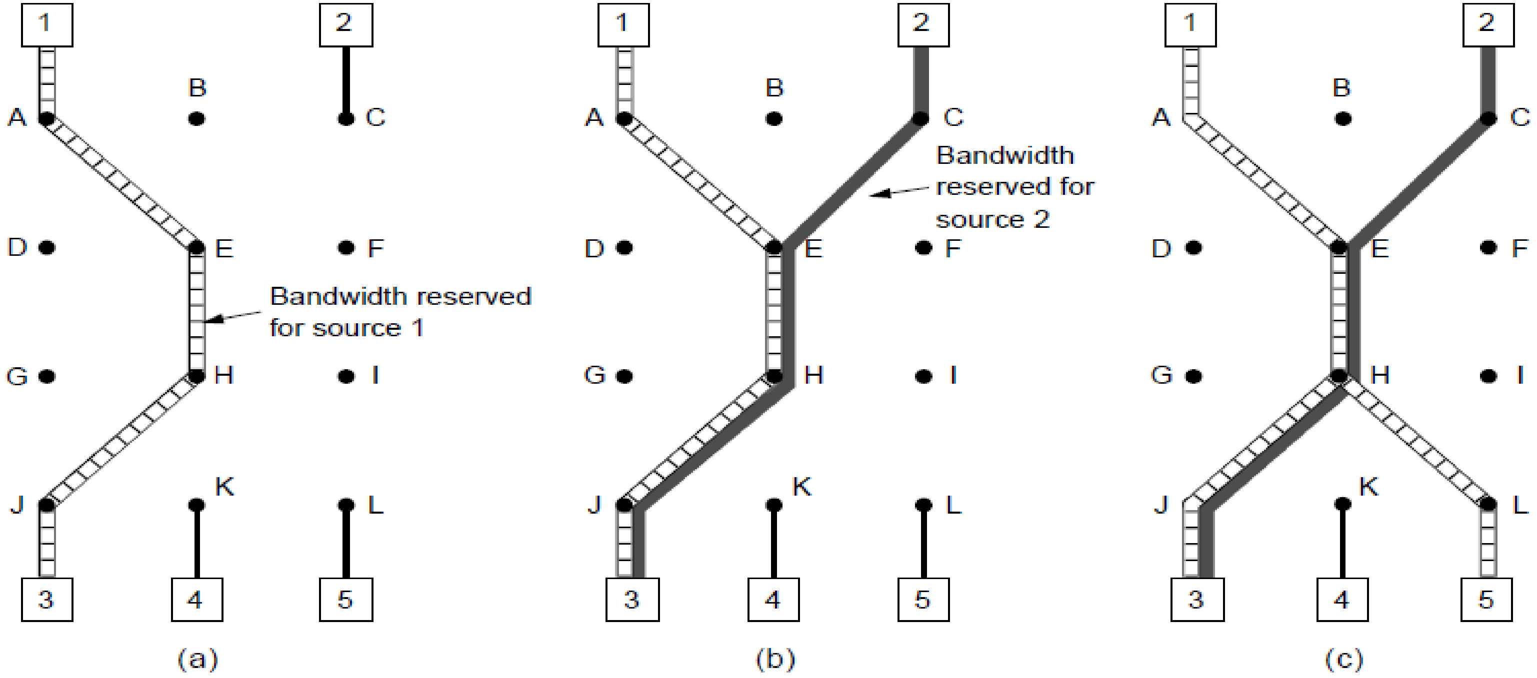
Integrated Services (1)



(a) A network. (b) The multicast spanning tree for host 1. (c) The multicast spanning tree for host 2.

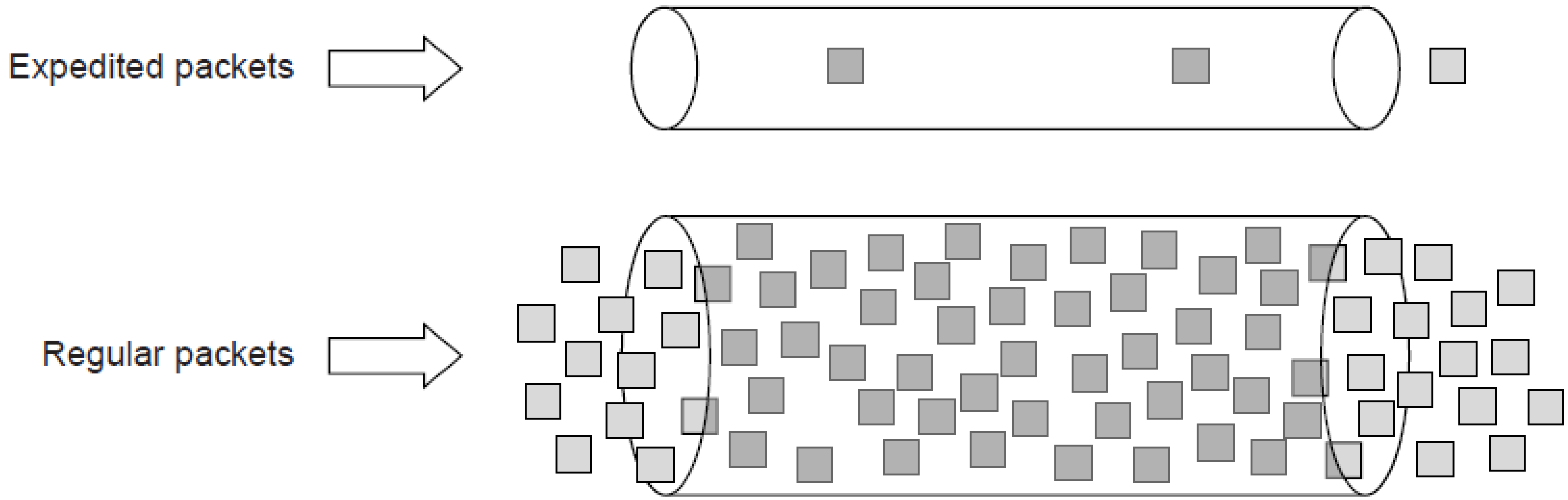
Integrated Services (2)

- 1，预留，不好扩展
- 2，路由器维护内部状态



(a) Host 3 requests a channel to host 1. (b) Host 3 then requests a second channel, to host 2.
(c) Host 5 requests a channel to host 1.

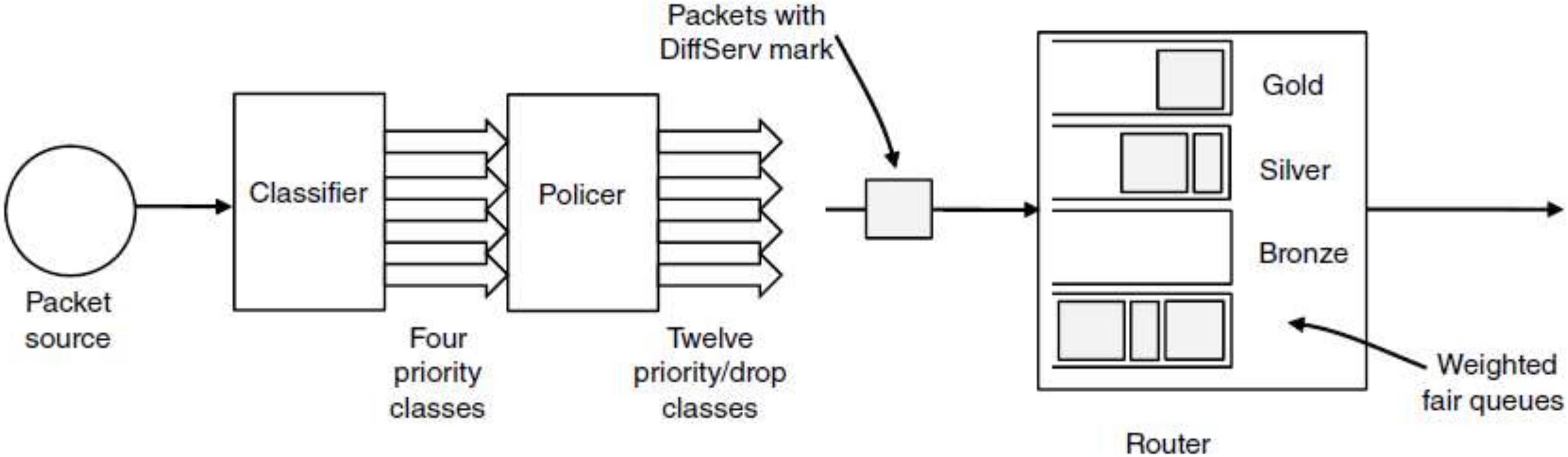
Differentiated Services (1)



Expedited packets experience a traffic-free network

Differentiated Services (2)

- 4个优先级，3个丢包的优先级，4*3=12个



A possible implementation of assured forwarding

5.5 Internetworking

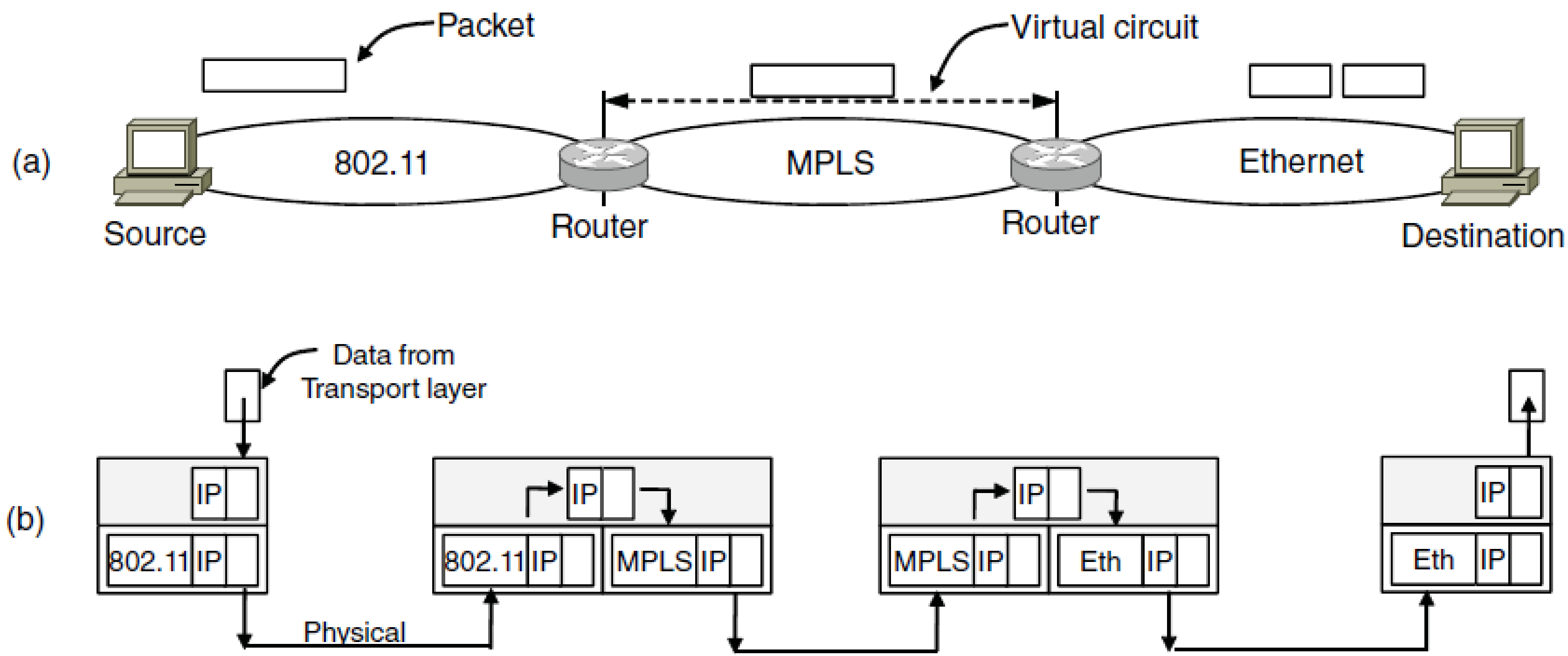
- Internetworks: an overview
- How networks differ
- Connecting heterogeneous networks
- Connecting endpoints across heterogeneous networks
- Internetwork routing: routing across multiple networks
- Supporting different packet sizes: packet fragmentation

How Networks Differ

Item	Some Possibilities
Service offered	Connectionless versus connection oriented
Addressing	Different sizes, flat or hierarchical
Broadcasting	Present or absent (also multicast)
Packet size	Every network has its own maximum
Ordering	Ordered and unordered delivery
Quality of service	Present or absent; many different kinds
Reliability	Different levels of loss
Security	Privacy rules, encryption, etc.
Parameters	Different timeouts, flow specifications, etc.
Accounting	By connect time, packet, byte, or not at all

Some of the many ways networks can differ

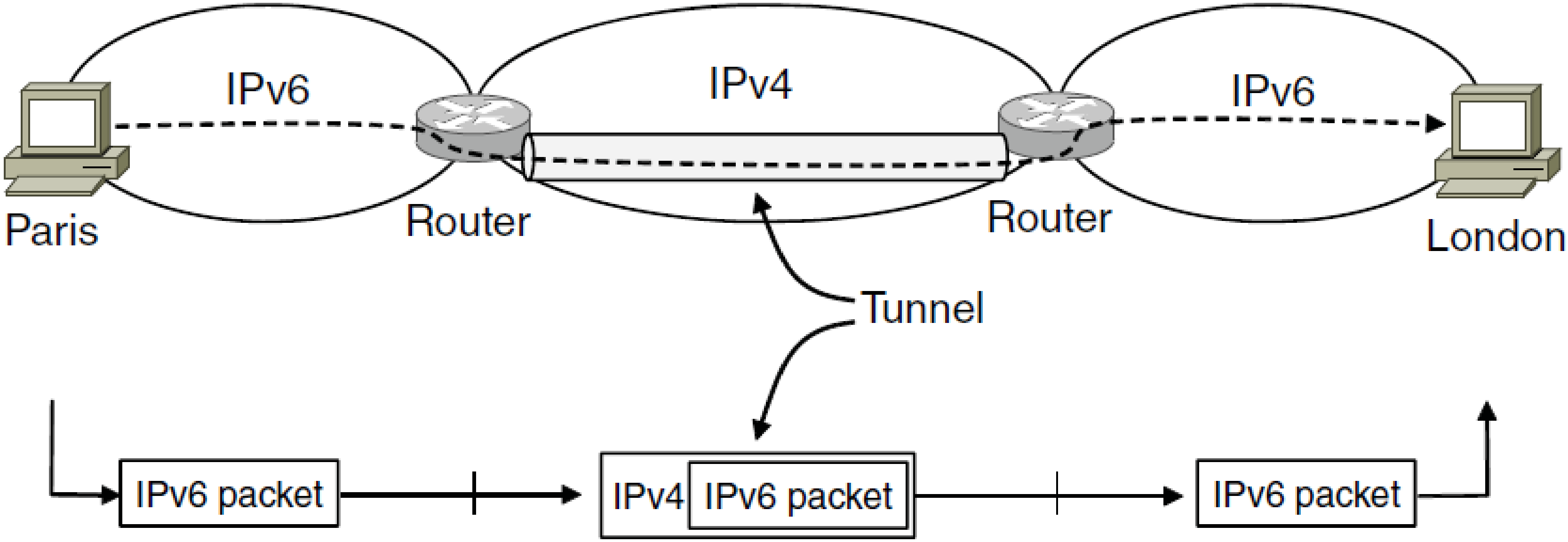
Connecting Heterogeneous Networks



(a) A packet crossing different networks.

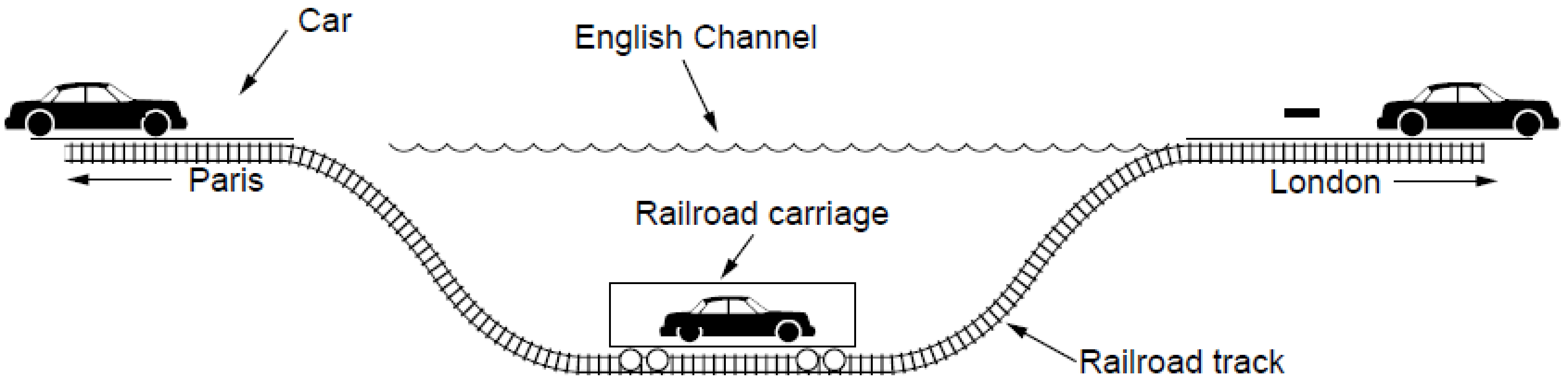
(b) Network and link layer protocol processing.

Connecting Endpoints Across Heterogeneous Networks (1 of 2)



Tunneling a packet from Paris to London.

Connecting Endpoints Across Heterogeneous Networks (1 of 2)



Tunneling a car from France to England

Supporting Different Packet Sizes: Packet Fragmentation (1 of 3)

Packet size issues:

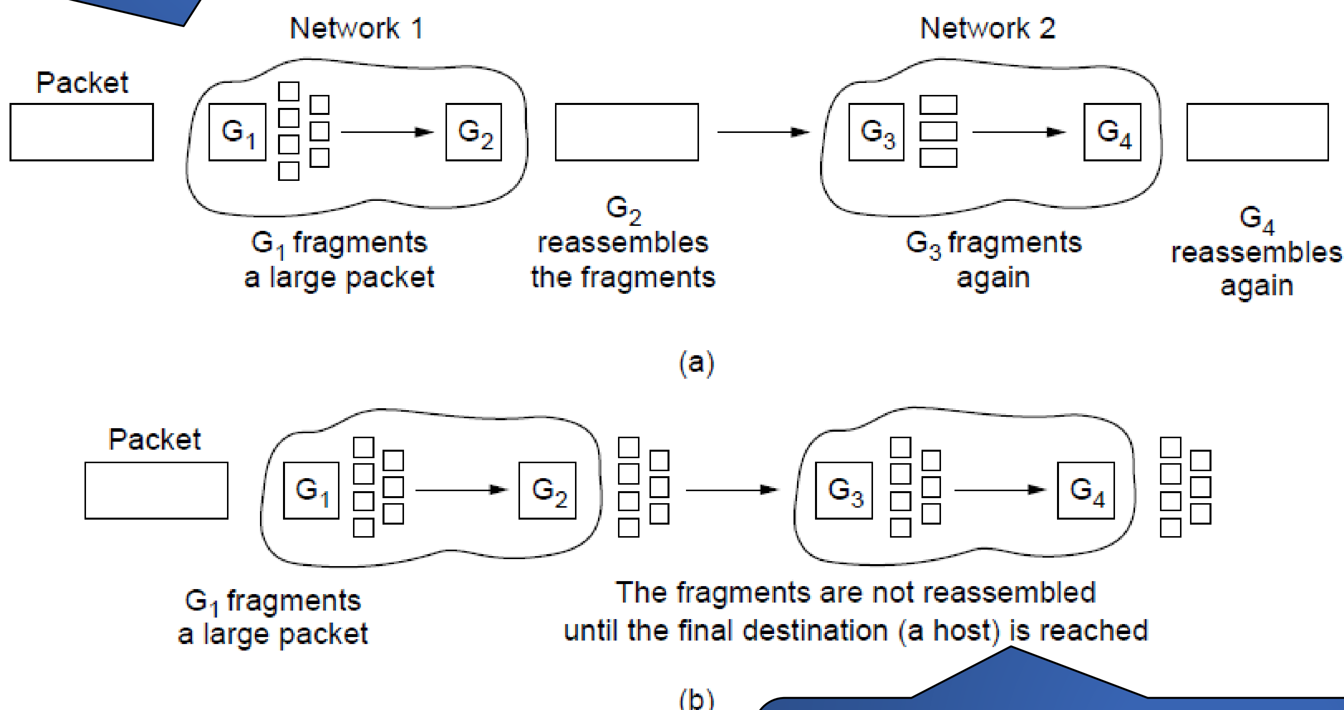
MTU: Path Maximum Transmission Unit
路径传输最大单元

- Hardware
- Operating system
- Protocols
- Compliance with (inter)national standard.
- Reduce error-induced retransmissions
- Prevent packet occupying channel too long.

Supporting Different Packet Sizes: Packet Fragmentation (2 of 3)

1，计数或识别数据包结束
2，路由受到限制
3，缓冲

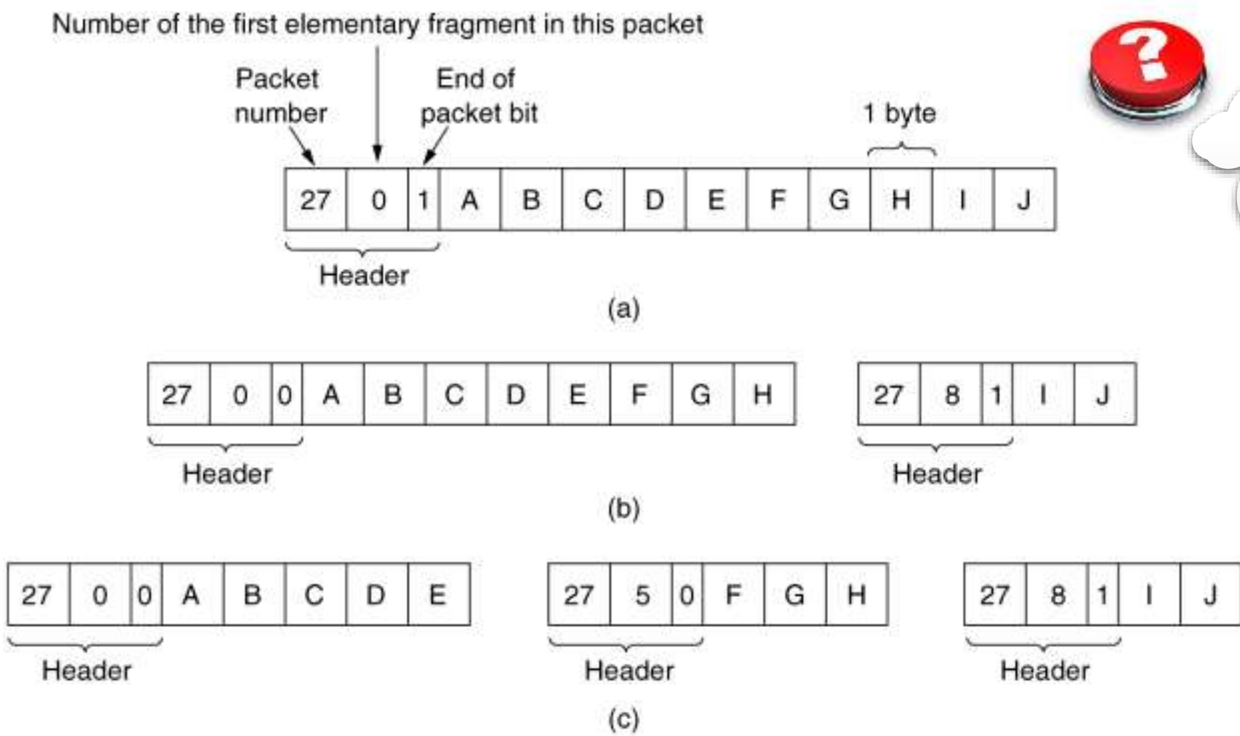
优缺点？



1, 分段开销，有的线路不需要
2,丢失的可能 e.g. IP

(a) Transparent fragmentation. （b）Nontransparent fragmentation

Supporting Different Packet Sizes: Packet Fragmentation (3 of 3)

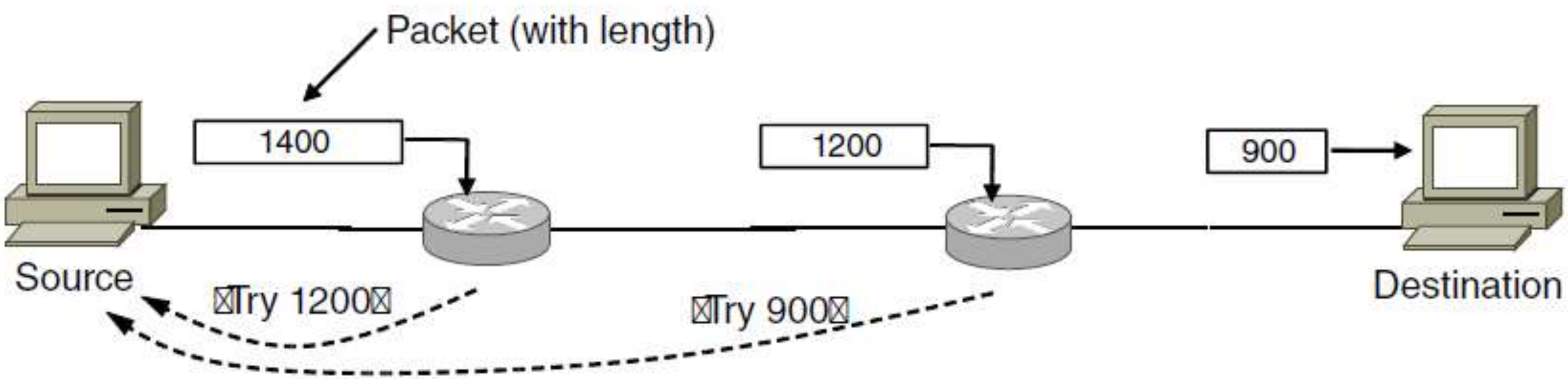


非透明分段方式下，可不可能不分段？

按链路最小的发

Fragmentation when the elementary data size is 1 byte. (a) Original packet, containing 10 data bytes. (b) Fragments after passing through a network with maximum packet size of 8 payload bytes plus header. (c) Fragments after passing through a size 5 gateway.

Supporting Different Packet Sizes: Packet Fragmentation (4 of 3)



Path MTU Discovery

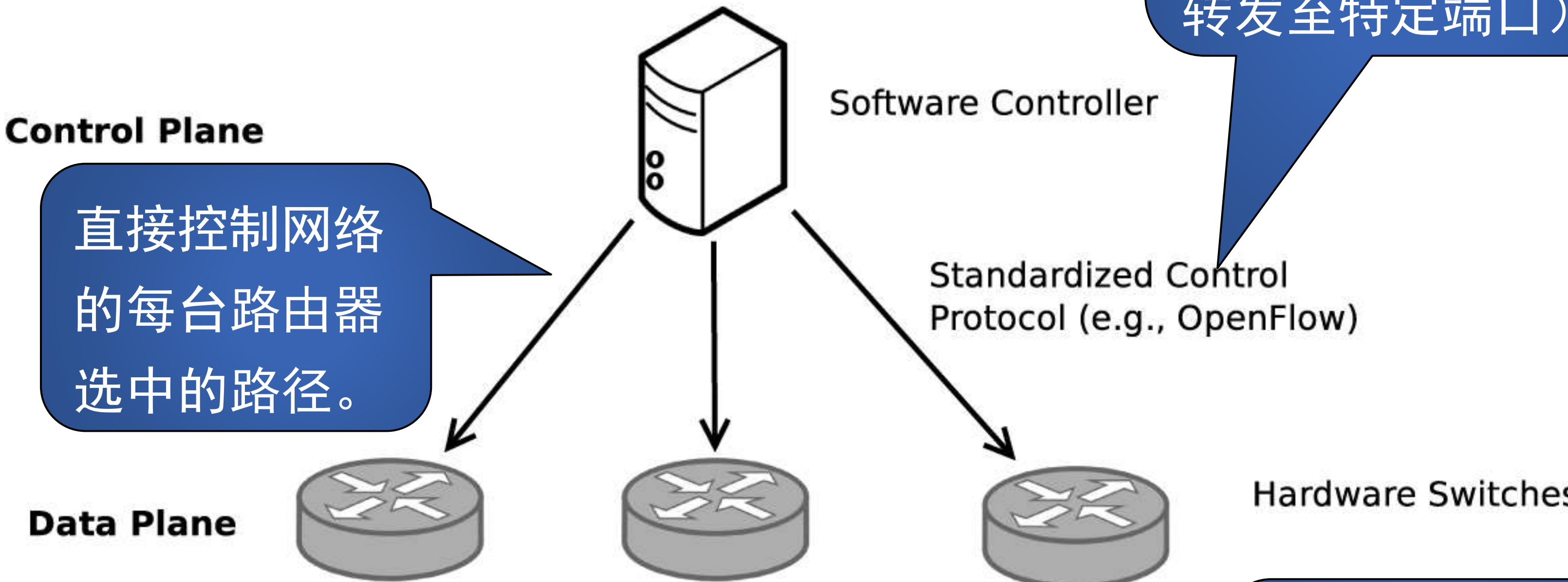
5.6 Software-Defined Networking

- Overview
- The SDN control plane: logically centralized software control
- The SDN data plane: programmable hardware
- Programmable network telemetry

Overview

匹配-动作表：匹配头部字段（MAC，IP），执行动作（如转发至特定端口）

直接控制网络的每台路由器选中的路径。



Control and data plane separation in SDN

适用于数据中心网络，广域网等因操作有限受限

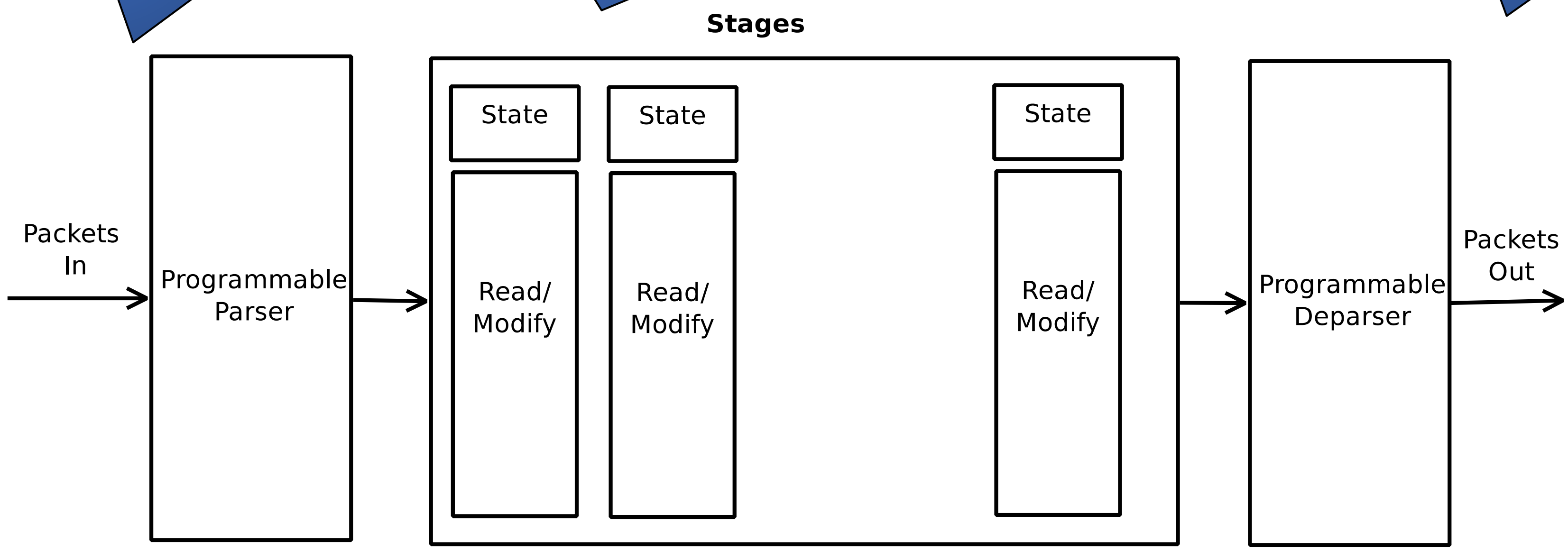
The SDN Data Plane: Programmable Hardware (1 of 2)

1 可编程解析器：数据包头部读取

2 一组匹配步骤：修改、转发、丢弃

硬件本身更加可编程

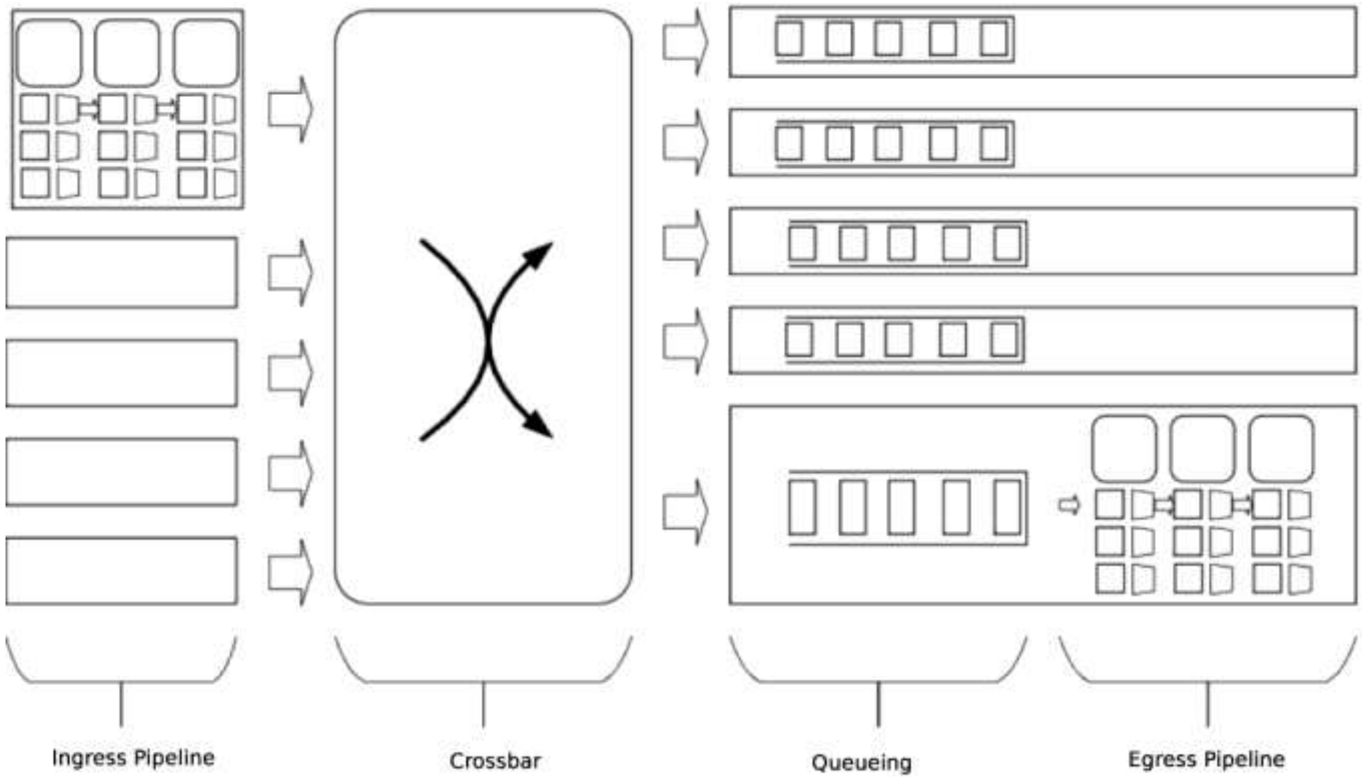
3可编程反解析器：写回数据包



Reconfigurable match-action pipeline for a programmable data plane

可编程数据平面的可重配置匹配-动作流水线。

The SDN Data Plane: Programmable Hardware (2 of 2)



Reconfigurable match-action pipelines on both ingress and egress

进入和离开时都能执行定制化的处理

5.7 The Network Layer in the Internet (1 of 3)

- The IP Version 4 Protocol
- IP Addresses
- IP Version 6
- Internet Control Protocols
- Label Switching and MPLS
- OSPF—An Interior Gateway Routing Protocol
- BGP—The Exterior Gateway Routing Protocol
- Internet Multicasting
- Mobile IP

The Network Layer in the Internet (2 of 3)

Top 10 principles

1. Make sure it works
2. Keep it simple
3. Make clear choices
4. Exploit modularity
5. Expect heterogeneity

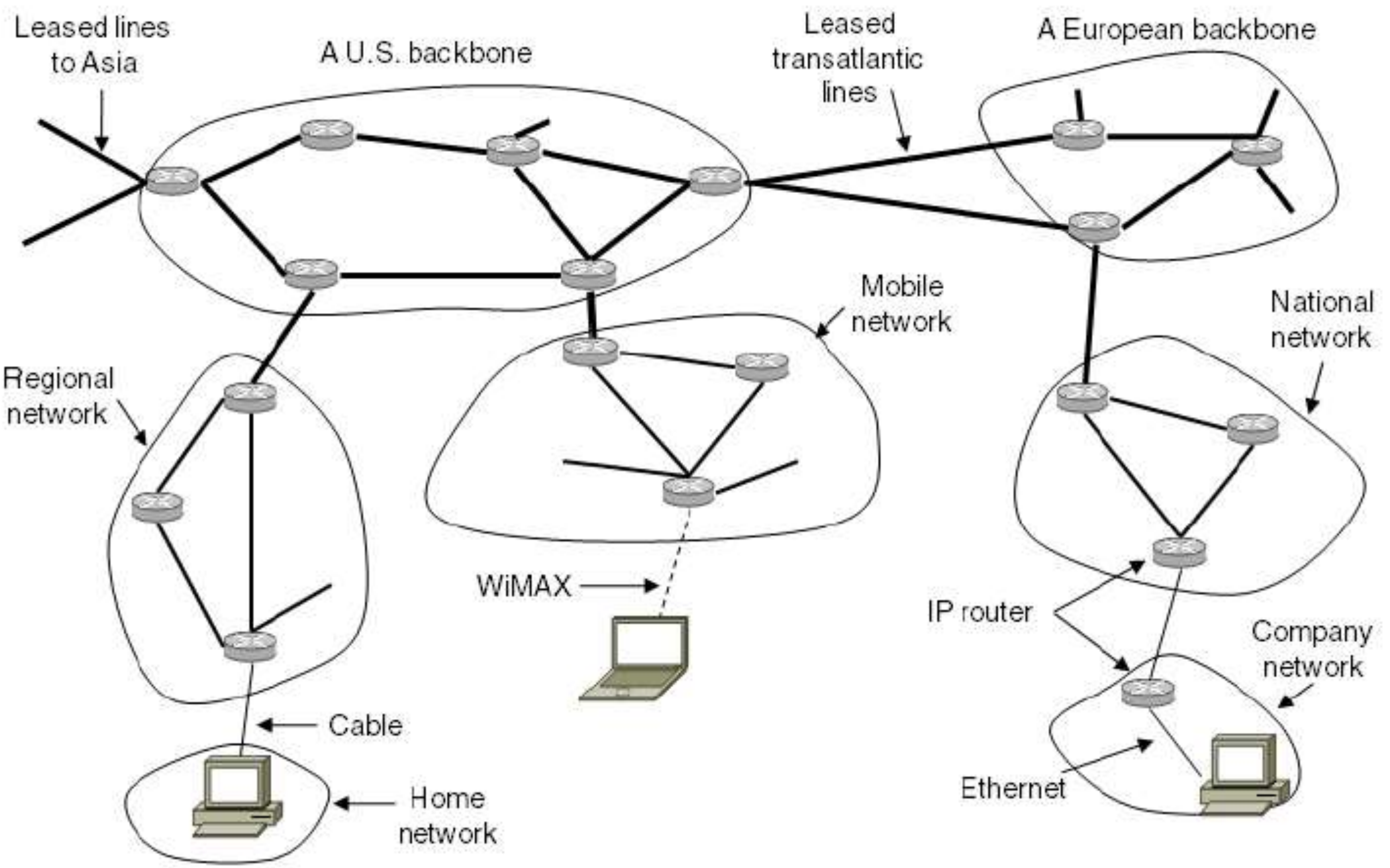
   . . .

The Network Layer in the Internet (2 of 3)

. . .

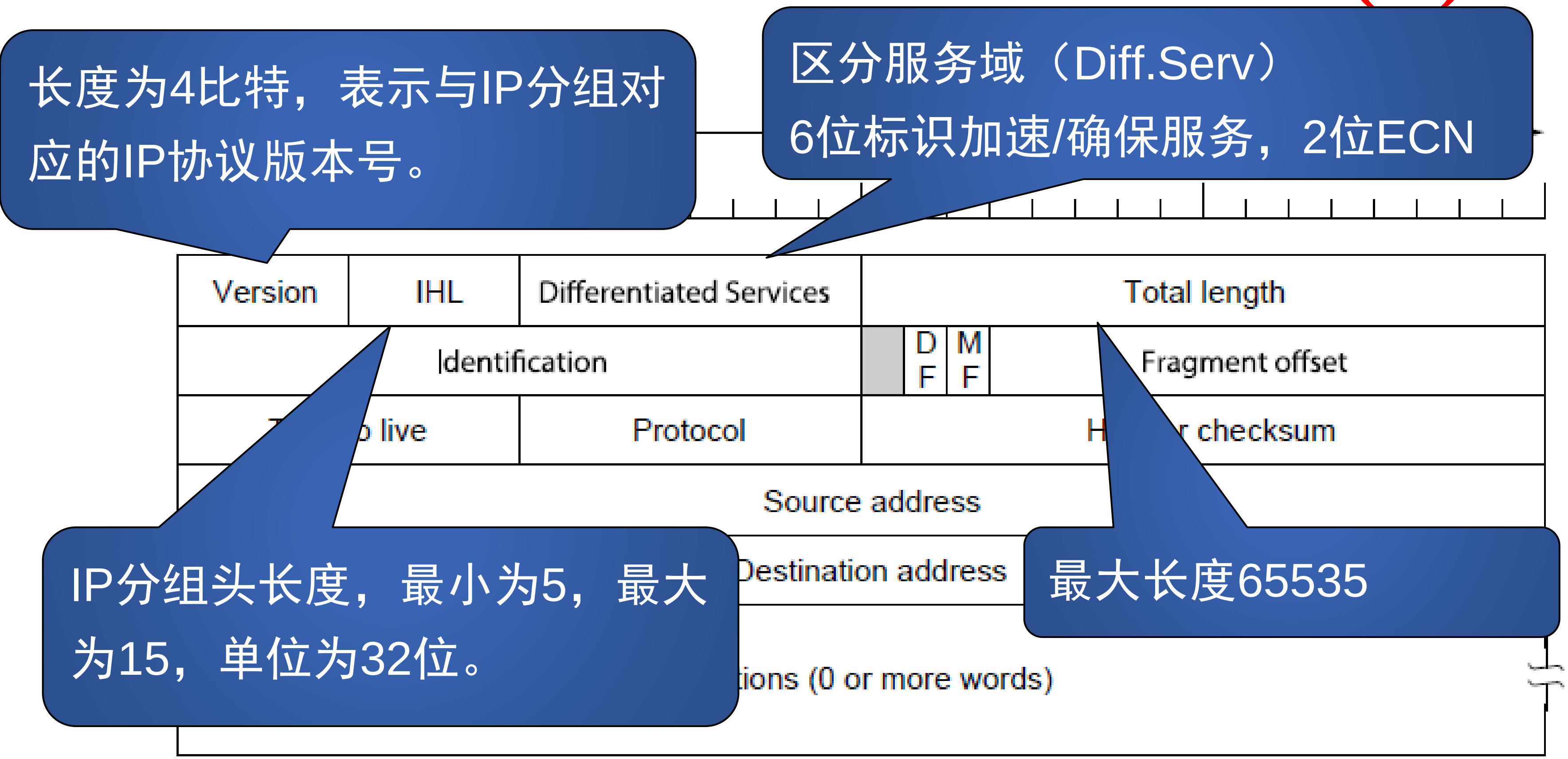
6. Avoid static options and parameters
7. Look for good design (not perfect)
8. Strict sending, tolerant receiving
9. Think about scalability
10. Consider performance and cost

The Network Layer in the Internet (3 of 3)



The Internet is an interconnected collection of many networks.

The IP Version 4 Protocol (1)



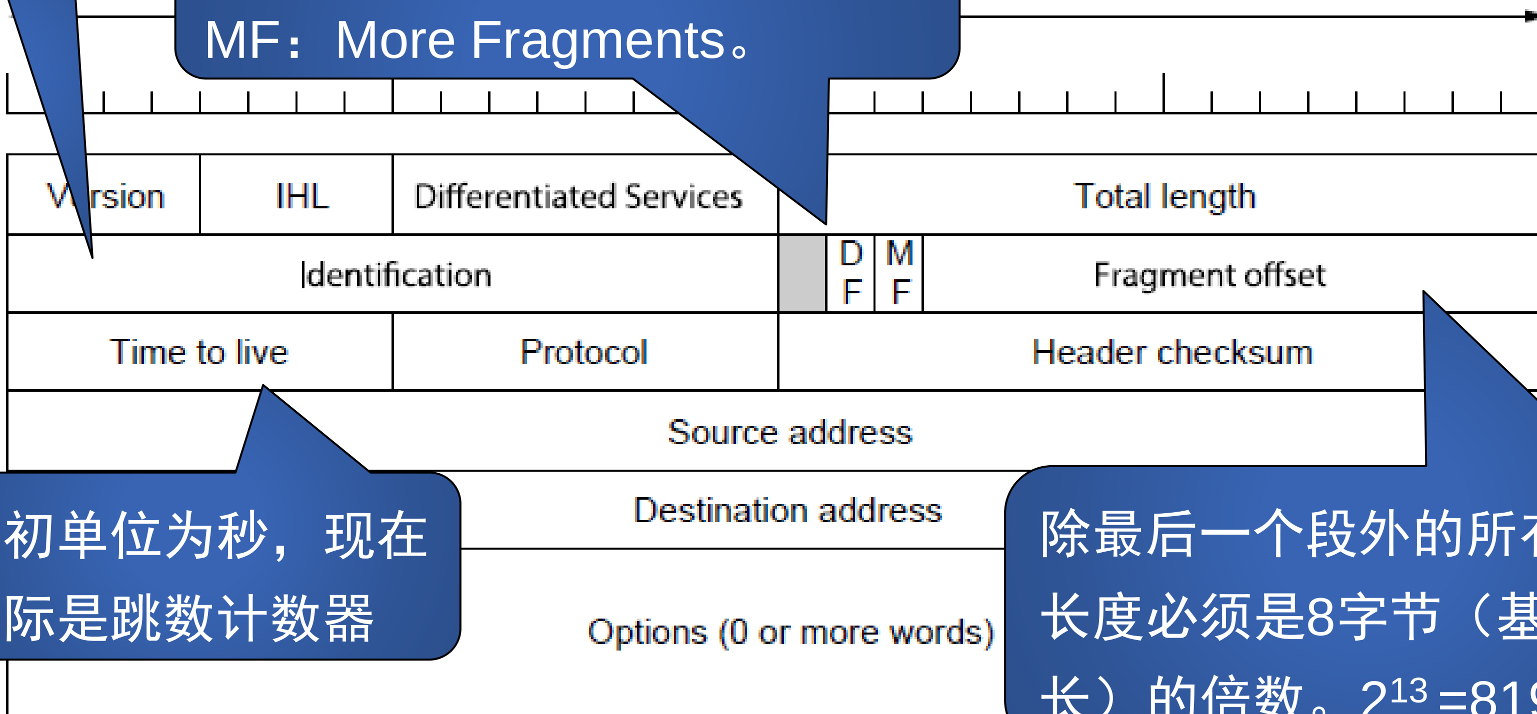
The IPv4 (Internet Protocol) header.

The IP Version 4 Protocol (1)

标识域

DF：Don’t Fragment；

MF：More Fragments。



最初单位为秒，现在实际是跳数计数器

除最后一个段外的所有段的长度必须是8字节（基本段长）的倍数。2^{13}=8192

The IPv4 (Internet Protocol) header.

例子

一个长度为1500字节的UDP段，通过IP分组进行传输，不使用头部扩展选项。现串行经过两个物理网络发往目的主机，这两个网络的MTU分别为1000字节和600字节。请写出到达目的结点时，IP分组和各IP分片的首部下列字段或标志的具体内容。

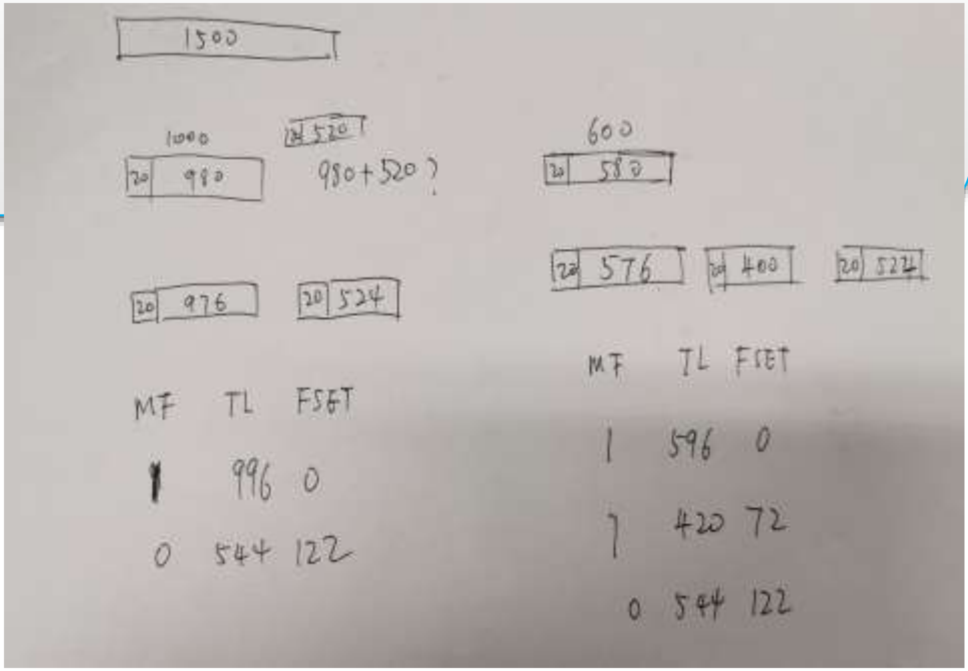
MF标志；
分组总长度TL；
分段偏移量Offset。

第一步:

MF	TL	FSET
1	996	0
0	544	122

第二步:

MF	TL	FSET
1	596	0
1	420	72
0	544	122



The IP Version 4 Protocol (2)

Option	Description
Security	Specifies how secret the datagram is
Strict source routing	Gives the complete path to be followed
Loose source routing	Gives a list of routers not to be missed
Record route	Makes each router append its IP address
Timestamp	Makes each router append its address and timestamp

Some of the IP options.

IP Addresses

- Prefixes
  - A contiguous block of IP address space
- Subnets
- CIDR—Classless InterDomain Routing
- Classful and special addressing
- NAT—Network Address Translation

Prefixes

4字节32位点分10进制：128.208.2.151

层次路由的优缺点？

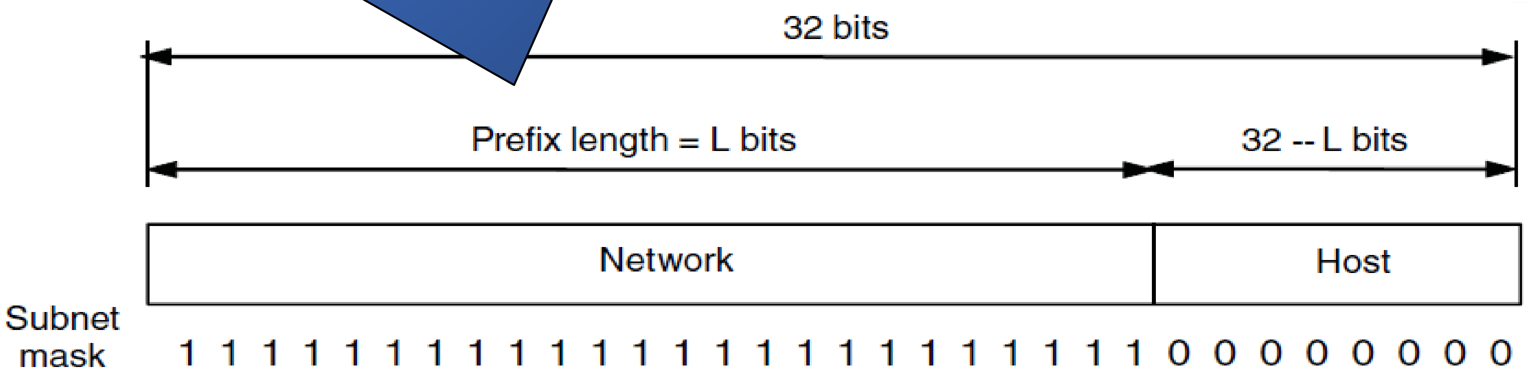
- Ip地址并不指向主机，而是网络接口。1个主机多个接口，路由器都有IP地址。
- 一个网络对应一块连续的地址空间，这块地址空间就称为地址的前缀。

层次路由显著降低路由表项。

问题：

1） 地址和位置绑定了

2） 浪费地址，不灵活。



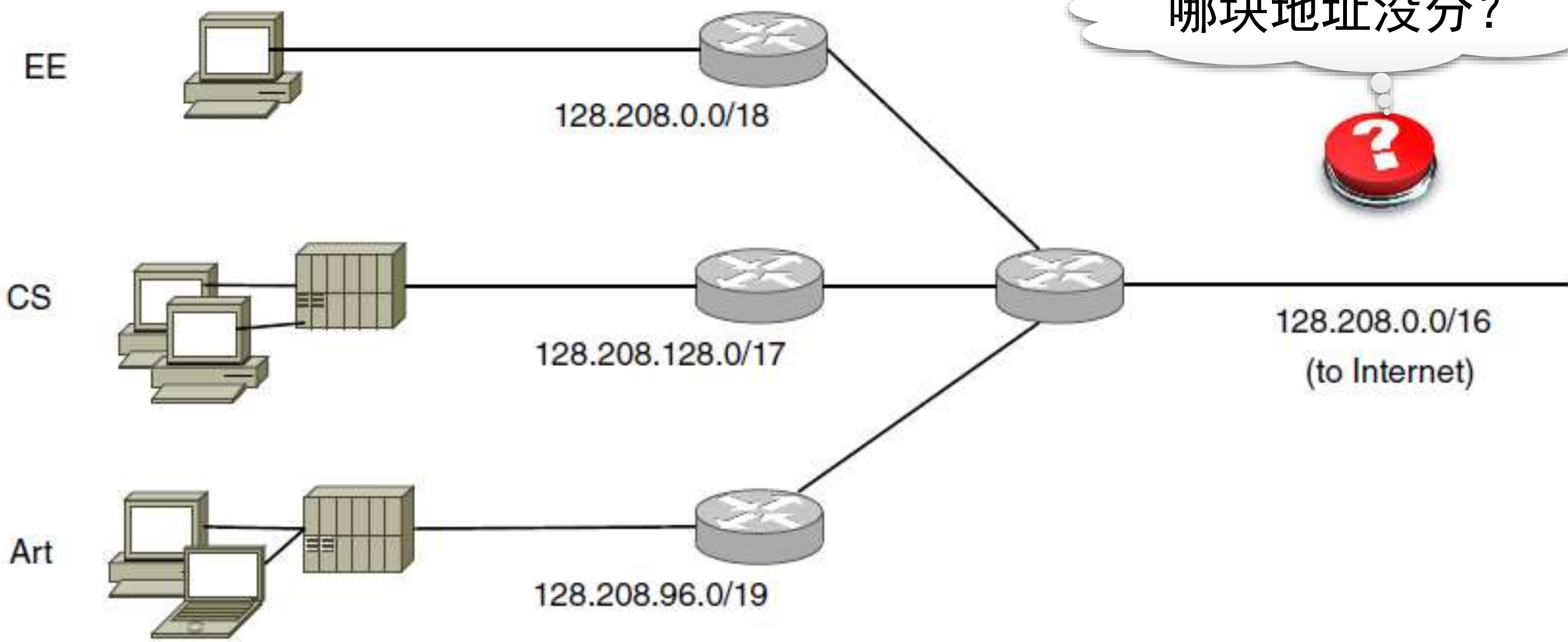
An IP prefix. **128.208.0.0/24**

Subnets

EE   10000000 11010000   00xxxxxx xxxxxxxx
CS   10000000 11010000   1xxxxxxx xxxxxxxx
ART 10000000 11010000   011xxxxx xxxxxxxx

010|xxxxx xxxxxxxx
128.208.64.0/19

哪块地址没分？



Splitting an IP prefix into separate networks with subnetting.

CIDR—Classless InterDomain Routing (1 of 2)

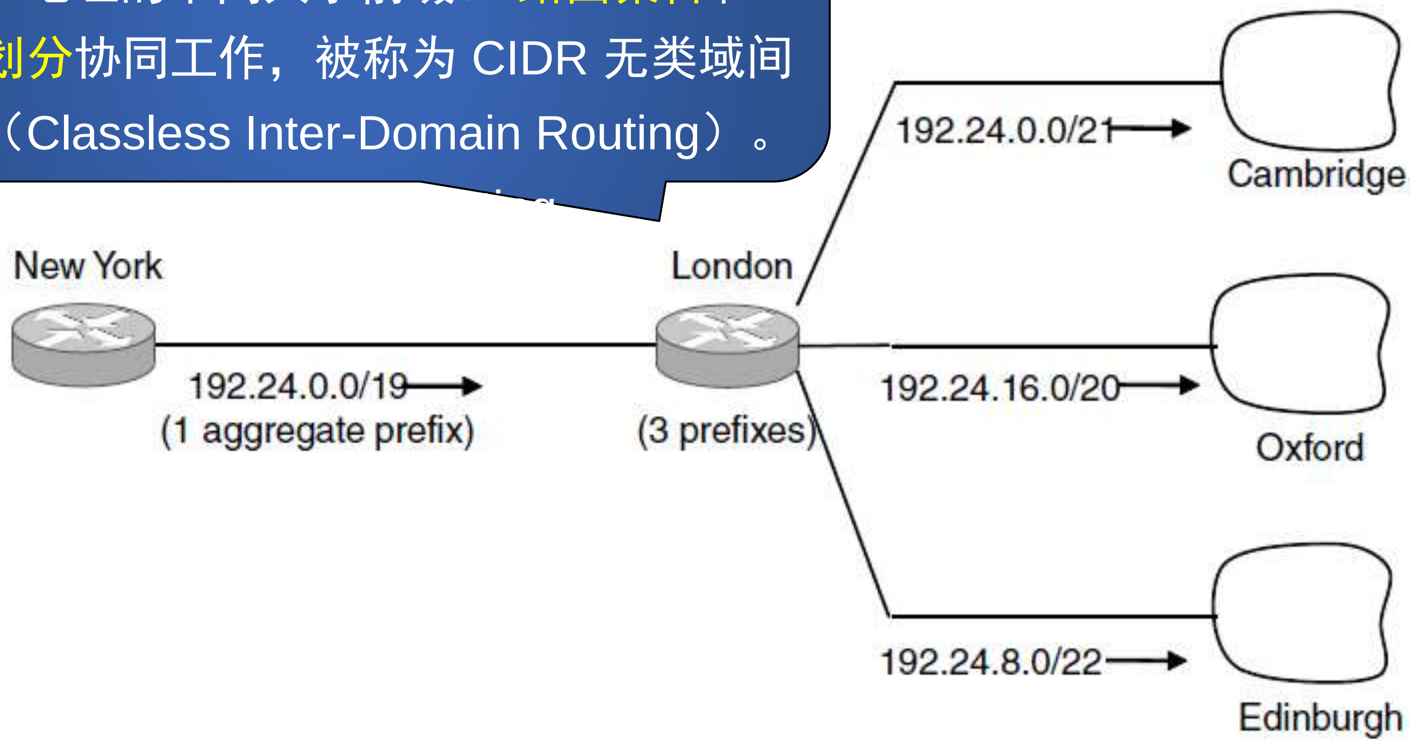
Camb 00000xxx xxxxxxxx
Edin 000010xx xxxxxxxx
Avai 000011xx xxxxxxxx
Oxfo 0001xxxx xxxxxxxx

University	First address	Last address	How many	Prefix
Cambridge	194.24.0.0	194.24.7.255	2048	194.24.0.0/21
Edinburgh	194.24.8.0			194.24.8.0/22
(Available)				
Oxford	194.24.16.0			194.24.16.0/20

A set of IP address assignments

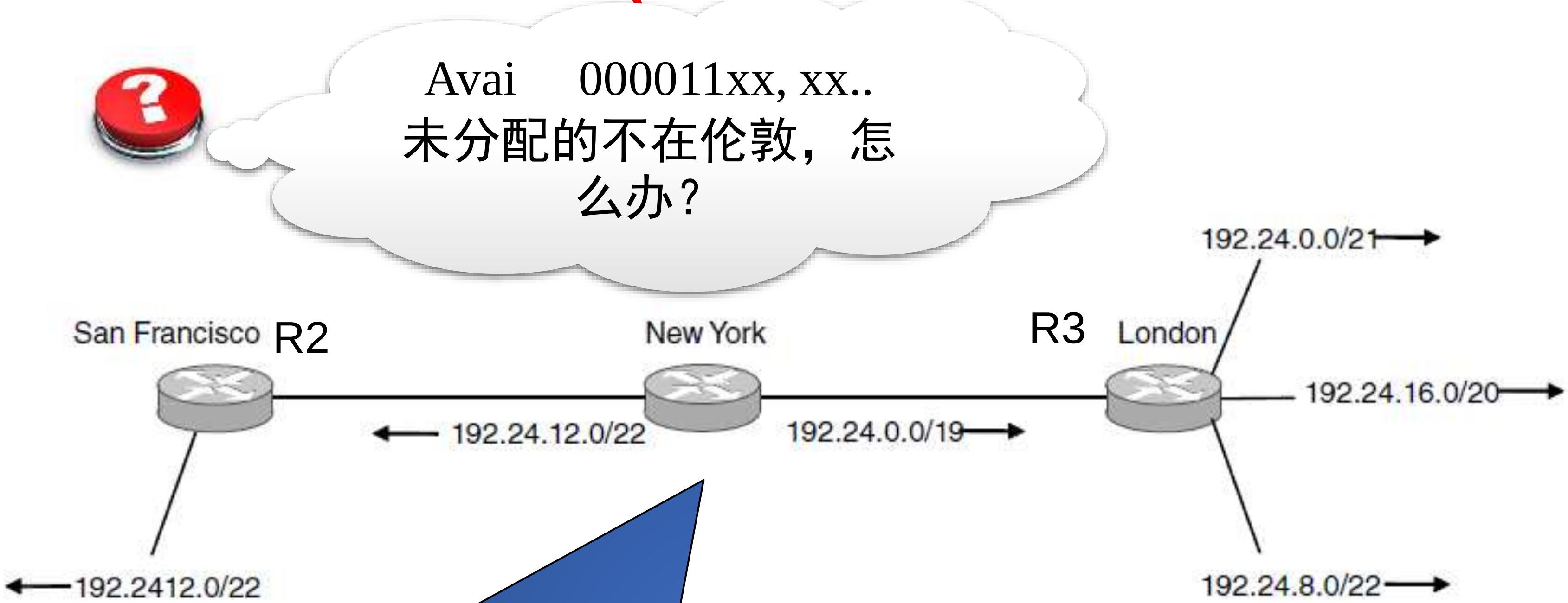
CIDR—Classless InterDomain Routing (2 of 3)

不同的路由器看待的不一样，可以知道一个给定IP地址的不同大小前缀。 路由聚合和子网划分协同工作，被称为 CIDR 无类域间路由（Classless Inter-Domain Routing）。



- Aggregation of IP prefixes

CIDR—Classless InterDomain Routing (3 of 3)



地址/掩码	下一跳
192.24.0.0/19	R3
192.24.12.0/22	R2

Longest matching prefix routing at the New York router.

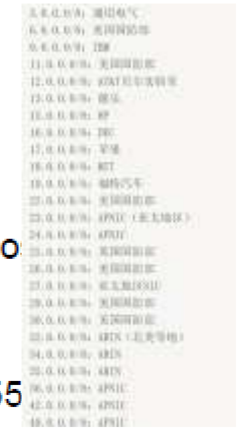
Classful and Special Addressing (1 of 2)

A: 128网络，1600万主机

B：16384网络，65535主机

C：200万个网络，256主机

Class				Range of ho addresses
A	0	Network	Host	1.0.0.0 to 127.255.255
B	10	Network	Host	128.0.0.0 to 191.255.255.255
C	110	Network	Host	192.0.0.0 to 223.255.255.255
D	1110	Multicast address		224.0.0.0 to 239.255.255.255
E	1111	Reserved for future use		240.0.0.0 to 255.255.255.255



IP address formats

Classful and Special Addressing (2 of 2)

Host全0表示网络本身，全1是广播地址

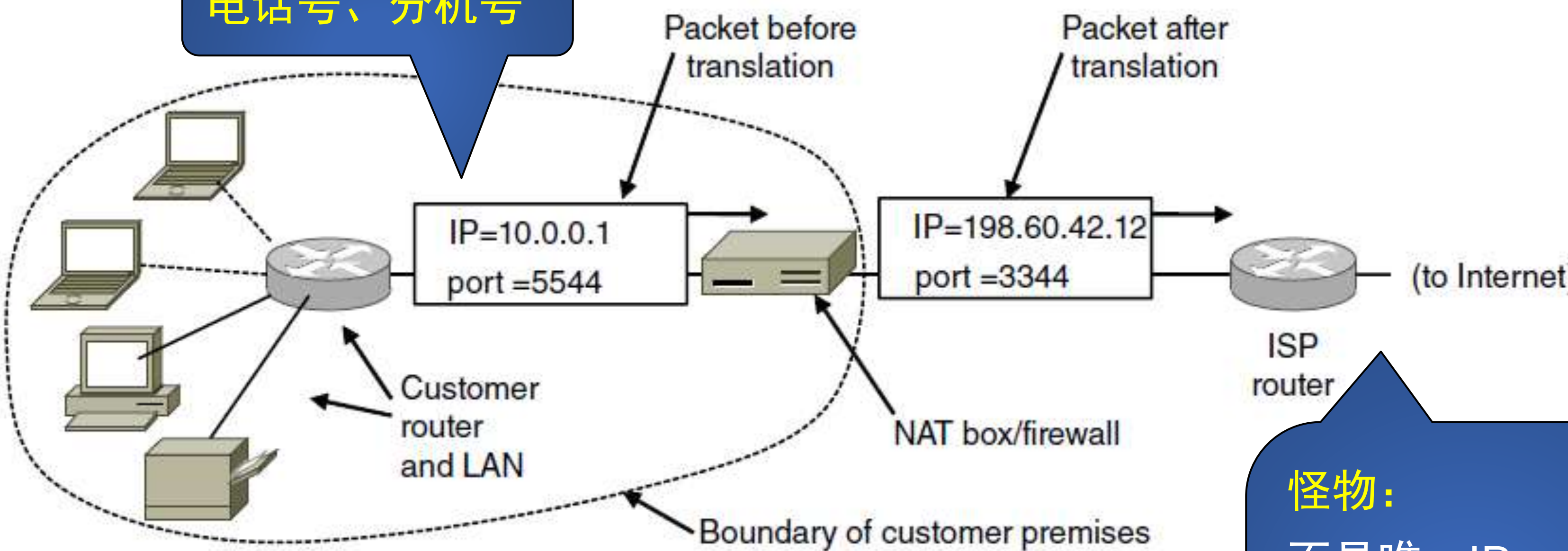
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		This host
0 0 . . . 0 0	Host	A host on this network
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		Broadcast on the local network
Network	1 1 1 1 . . . 1 1 1 1	Broadcast on a distant network
127	(Anything)	Loopback

Special IP addresses

NAT—Network Address Translation

- NAT is a quick fix for the problem of running out of IP address

电话号、分机号



Placement and operation of a NAT box.

怪物：

不是唯一IP，

有连接

违背分层原则，

必须TCP / UDP

连接有限，65535

NAT—Network Address Translation

- NAT is to assign each company a single IP address for Internet traffic. Within the company, every host gets a unique IP address, which is used for routing for intramural traffic. When a packet exits the company and goes to the ISP, an address translation takes place. To make this scheme possible, three ranges of IP addresses have been declared as private.
- The reserved ranges are:
  - 10.0.0.0 -10.255.255.255/8 (16,777,216 hosts)
  - 172.16.0.0 -172.31.255.255/12 (1,048,576 hosts)
  - 192.168.0.0 -192.168.255.255/16 (65,536 hosts)

P 349

NAT—Network Address Translation (补)

- Full Cone NAT
- Restricted Cone NAT
- Port Restricted Cone NAT
- Symmetric NAT

NAT—Network Address Translation (补)

1、Full Cone NAT：内网主机建立一个 socket(LocalIP:LocalPort) 第一次使用这个 socket 给外部主机发送数据时 NAT 会给其分配一个公网(PublicIP:PublicPort),以后用这个 socket 向外面任何主机发送数据都将使用这对(PublicIP:PublicPort)。此外，任何外部主机只要知道这个(PublicIP:PublicPort)就可以发送数据给(PublicIP:PublicPort)，内网的主机就能收到这个数据包。

2、Restricted Cone NAT：内网主机建立一个 socket(LocalIP:LocalPort) 第一次使用这个 socket 给外部主机发送数据时 NAT 会给其分配一个公网(PublicIP:PublicPort),以后用这个 socket 向外面任何主机发送数据都将使用这对(PublicIP:PublicPort)。此外，如果任何外部主机想要发送数据给这个内网主机，只要知道这个(PublicIP:PublicPort)并且内网主机之前用这个 socket 曾向这个外部主机 IP 发送过数据。只要满足这两个条件，这个外部主机就可以用自己的(IP,任何端口)发送数据给(PublicIP:PublicPort)，内网的主机就能收到这个数据包。

NAT—Network Address Translation（补）

3、Port Restricted Cone NAT：内网主机建立一个 socket(LocalIP:LocalPort) 第一次使用这个 socket 给外部主机发送数据时 NAT 会给其分配一个公网(PublicIP:PublicPort),以后用这个 socket 向外面任何主机发送数据都将使用这对(PublicIP:PublicPort)。此外，如果任何外部主机想要发送数据给这个内网主机，只要知道这个(PublicIP:PublicPort)并且内网主机之前用这个 socket 曾向这个外部主机(IP,Port)发送过数据。只要满足这两个条件，这个外部主机就可以用自己的(IP,Port)发送数据给(PublicIP:PublicPort)，内网的主机就能收到这个数据包。

4、Symmetric NAT：内网主机建立一个 socket(LocalIP,LocalPort),当用这个 socket 第一次发数据给外部主机 1 时,NAT 为其映射一个(PublicIP-1,Port-1),以后内网主机发送给外部主机 1 的所有数据都是用这个(PublicIP-1,Port-1)，如果内网主机同时用这个 socket 给外部主机 2 发送数据，NAT 会为其分配一个(PublicIP-2,Port-2)，以后内网主机发送给外部主机 2 的所有数据都是用这个(PublicIP-2,Port-2).如果 NAT 有多于一个公网 IP，则 PublicIP-1 和 PublicIP-2 可能不同，如果 NAT 只有一个公网 IP,则 Port-1 和 Port-2 肯定

QQ 在同一个内网之间发消息，和qq在不同内网之间发消息，咋传输的？

IP Version 6 (1 of 3)

- The main IPv6 header
- Extension headers
- Controversies

IP Version 6 (2 of 3)

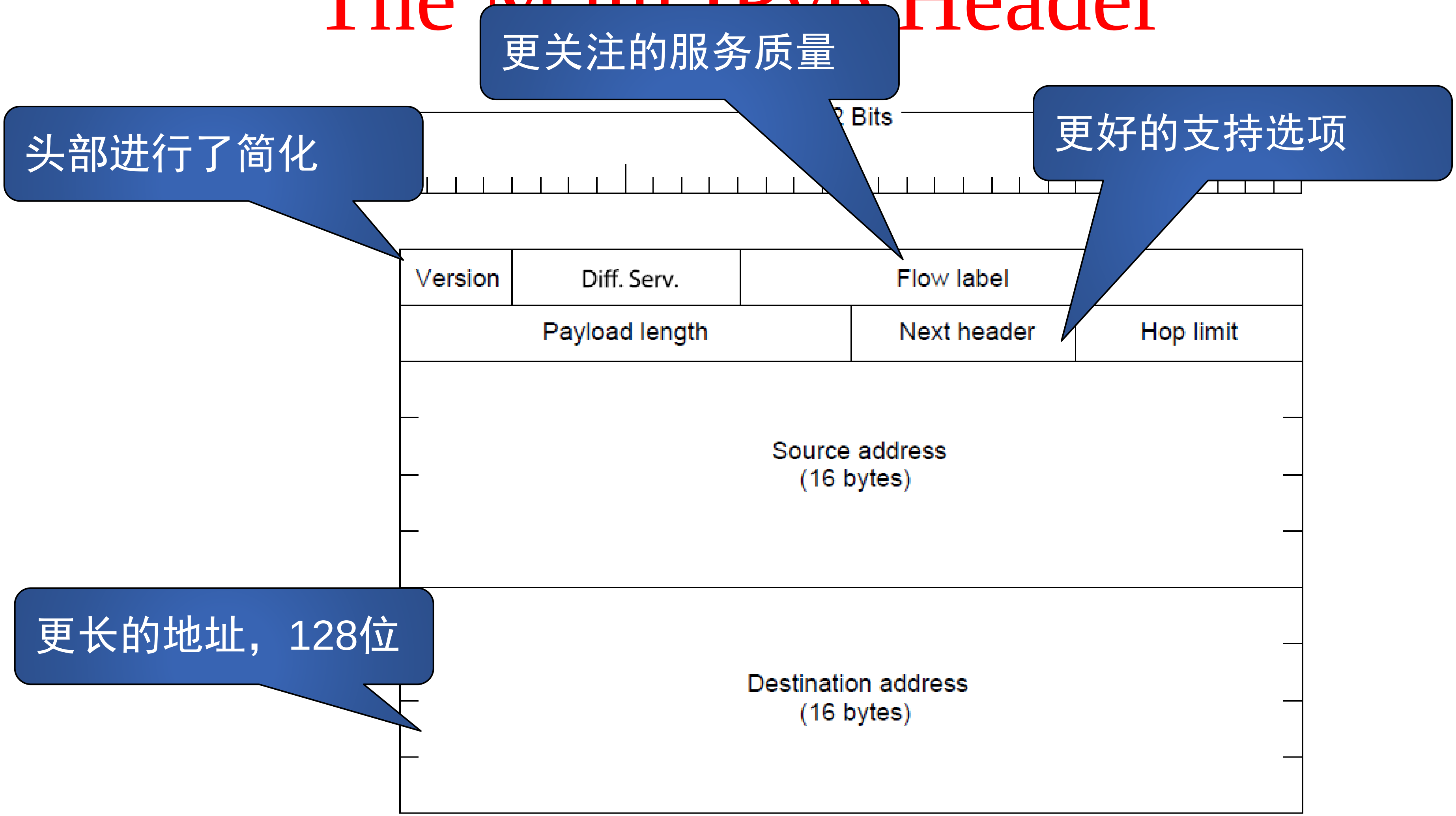
２的１２８次方
每平方米，７＊１０的２５次方（1000）

- IPv6 major goals
  - Support billions of hosts
  - Reduce routing table size
  - Simplify the protocol
  - Provide better security
  - Attention to type of service
  - Aid multicasting
  - Roaming host without changing address
  - Allow future protocol evolution
  - Permit coexistence of old and new protocols for years

IP Version 6 (3 of 3)

- IP version 6 improvements
  - Longer addresses than IPv4
  - Simplification of the header
  - Better support for options
  - Big advance is in security
  - Quality of service

The Main IPv6 Header



The IPv6 fixed header (required).

The Mai

区分服务域（Diff.Serv）

6位标识加速/确保服务，2位ECN

长度为4比特，表示与IP分组对应的IP协议版本号。

额外可选的扩展，如果是最后一个，传输层协议

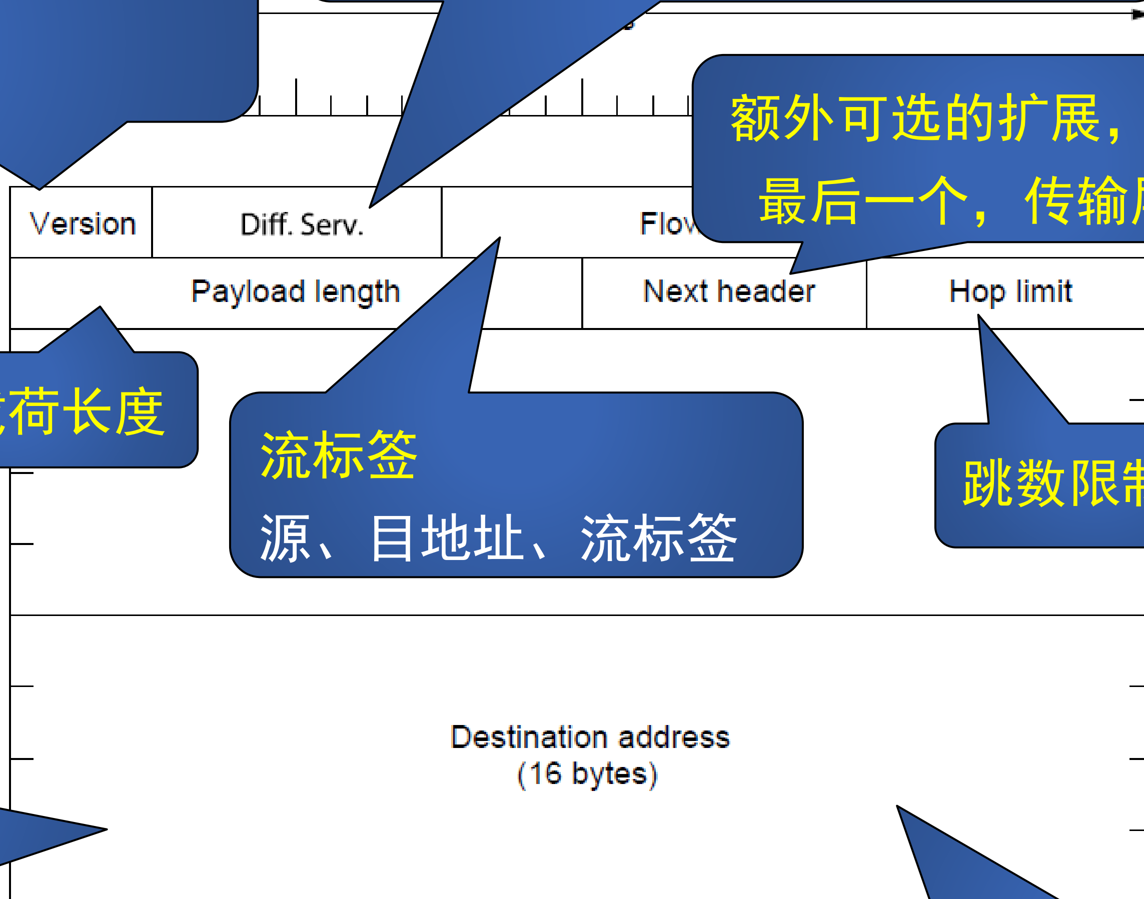
有效载荷长度

流标签

源、目地址、流标签

跳数限制

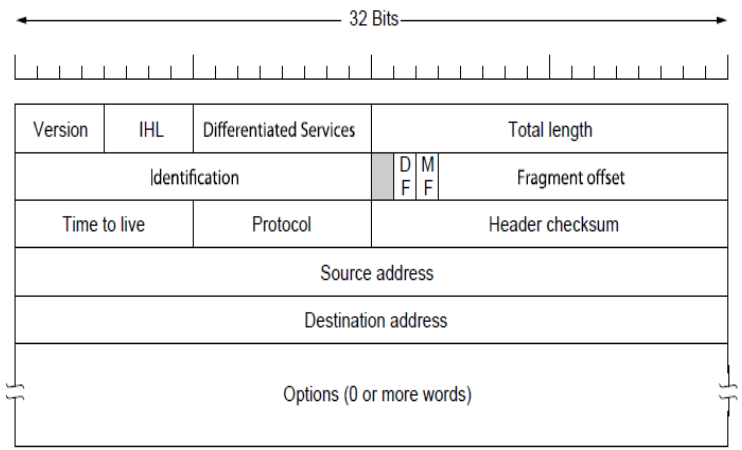
16个字节被冒号分成8组，每一组用4个16进制数



The IPv6 fixed header (required

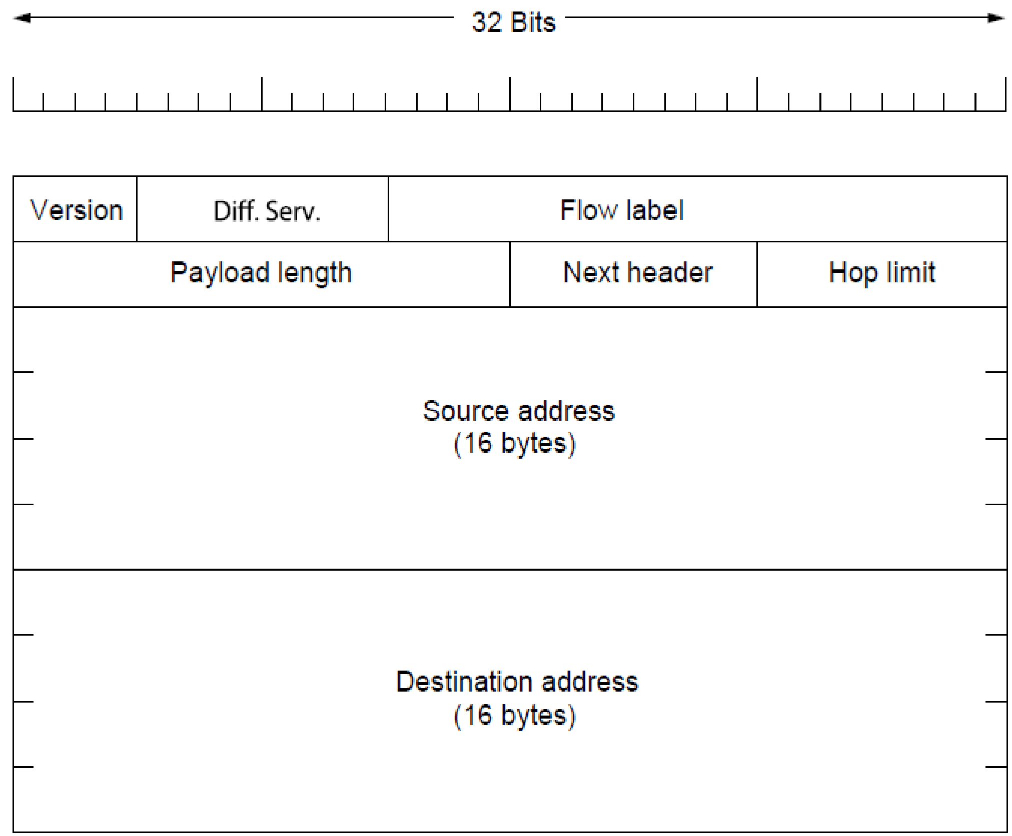
和IPv4相比少了什么：？

The Main IPv6 Header



和IPv4相比：
分段、校验和没了

快速、灵活，具有强大
地址空间的协议



The IPv6 fixed header (required).

Extension Headers

混杂信息：存放路由器必须要检查的信息

Extension header	Description
Hop-by-hop options	Miscellaneous information for routers
Destination options	Additional information for the destination
Routing	Loose list of routers to visit
Fragmentation	Management of datagram fragments
Authentication	Verification of the sender's identity
Encrypted security payload	Information about the encrypted contents

IPv6 extension headers

Internet Control Protocols

- ICMP—The Internet Control Message Protocol
- ARP—The Address Resolution Protocol
- DHCP—The Dynamic Host Configuration Protocol

ICMP—The Internet Control Message Protocol

不能定位，或者 DF标志写了不能分段， 而经过小数据包网络

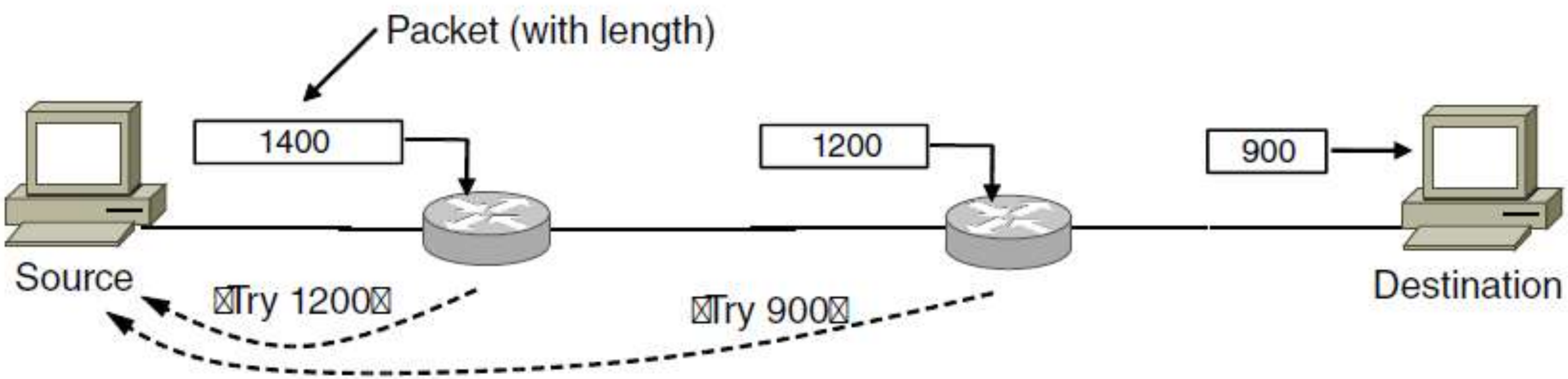
Traceroute/
Tracert

Message type	Description
Destination unreachable	Packet could not be delivered
Time exceeded	Time to live field hit 0
Parameter problem	Invalid header field
Source quench	Choke packet
Redirect	Teach a router about geography
Echo and Echo reply	Check if a machine is alive
Timestamp request/reply	Same as Echo, but with timestamp
Router advertisement/solicitation	Find a nearby router

Ping

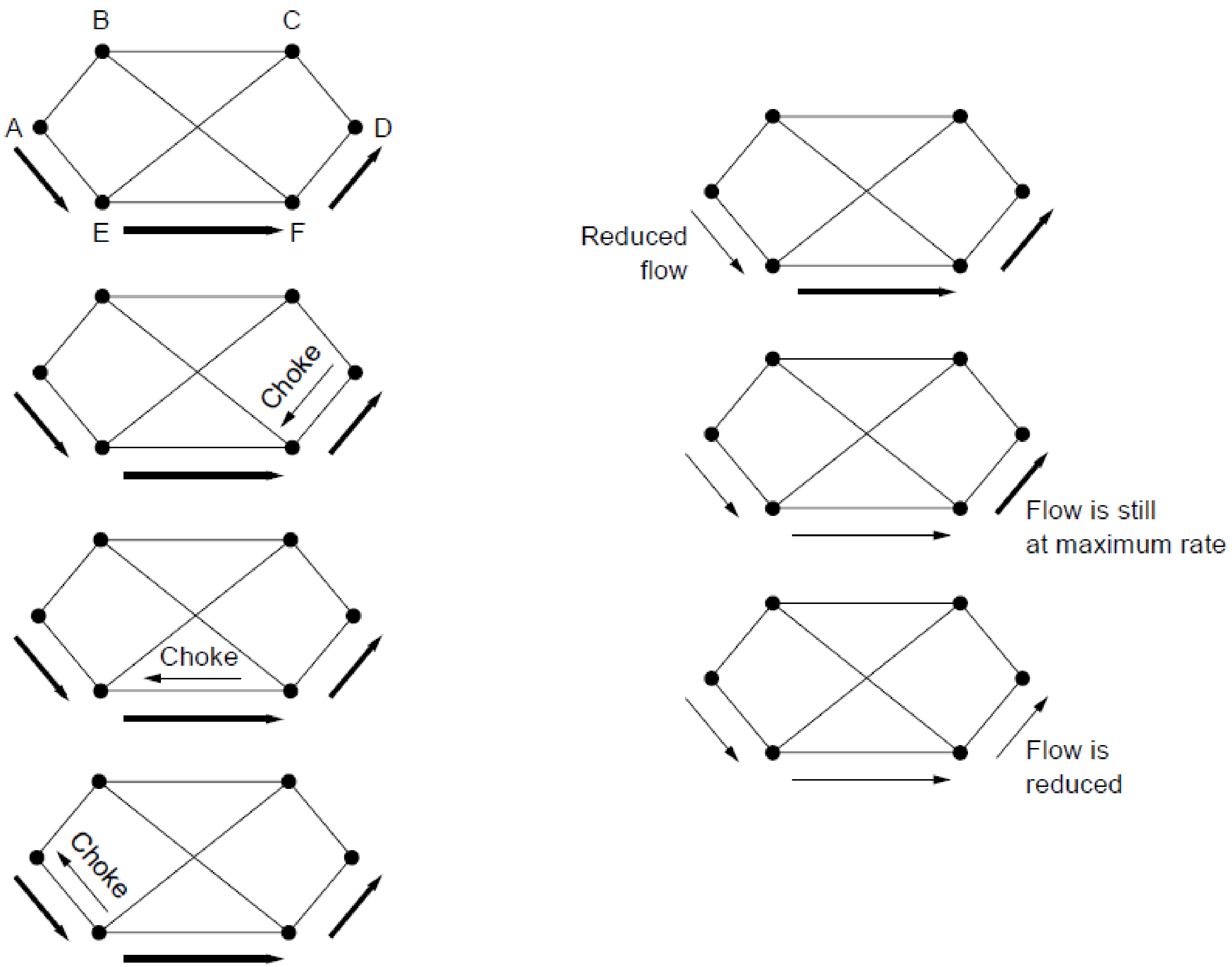
The principal ICMP message types.

Flashback…



Path MTU Discovery

Flashback…



A choke packet that affects only the source..

ARP—The Address Resolution Protocol

Filter:

No.	Time	Source			
6	4.363453	44:ad:			0x8004
7	4.508215	fe80::d			
8	5.335129	fe80::d5c4:f4e4:81(	ff02::1:2	DHCPv6	Solicit
9	5.506608	fe80::d5c4:f4e4:81(	ff02::1:fff7:9623	ICMPv6	Neighbor solicitation
10	5.932119	RealtekS_70:70:59	Broadcast	ARP	Who has 192.168.1.21? Tell 192.168.1.22
11	5.932535	RealtekS_68:69:00	RealtekS_70:70:59	ARP	192.168.1.21 is at 00:e0:4c:68:69:00
12	5.932562	192.168.1.22	192.168.1.21	ICMP	Echo (ping) request
13	5.933303	192.168.1.21	192.168.1.22	ICMP	Echo (ping) reply
14	6.364664	44:ad:d9:66:69:84	Spanning-tree-(for-	STP	Conf. Root = 32769/44:ad:d9:66:69:80 Cost = 0 Port = 0x8004
15	6.941842	192.168.1.22	192.168.1.21	ICMP	Echo (ping) request
16	6.942382	192.168.1.21	192.168.1.22	ICMP	Echo (ping) reply
17	7.819915	fe80::d897:cf06:12(	ff02::1:fff7:9623	ICMPv6	Neighbor solicitation
18	7.955822	192.168.1.22	192.168.1.21	ICMP	Echo (ping) request
19	7.956358	192.168.1.21	192.168.1.22	ICMP	Echo (ping) reply
20	8.042485	44:ad:d9:66:69:84	44:ad:d9:66:69:84	LOOP	Reply
21	8.323119	fe80::d897:cf06:12(	ff02::1:fff7:9623	ICMPv6	Neighbor solicitation
22	8.369533	44:ad:d9:66:69:84	Spanning-tree-(for-	STP	Conf. Root = 32769/44:ad:d9:66:69:80 Cost = 0 Port = 0x8004
23	8.969799	192.168.1.22	192.168.1.21	ICMP	Echo (ping) request
24	8.970257	192.168.1.21	192.168.1.22	ICMP	Echo (ping) reply
25	9.228038	fe80::d897:cf06:12(	ff02::1:2	DHCPv6	Solicit
26	9.321492	fe80::d897:cf06:12(	ff02::1:fff7:9623	ICMPv6	Neighbor solicitation
27	9.578203	fe80::d5c4:f4e4:81(	ff02::1:fff7:9623	ICMPv6	Neighbor solicitation

ARP协议：Address resolution protocol

主机1广播请求，主机2以自己的MAC应答

第二次？
反向？
动态？

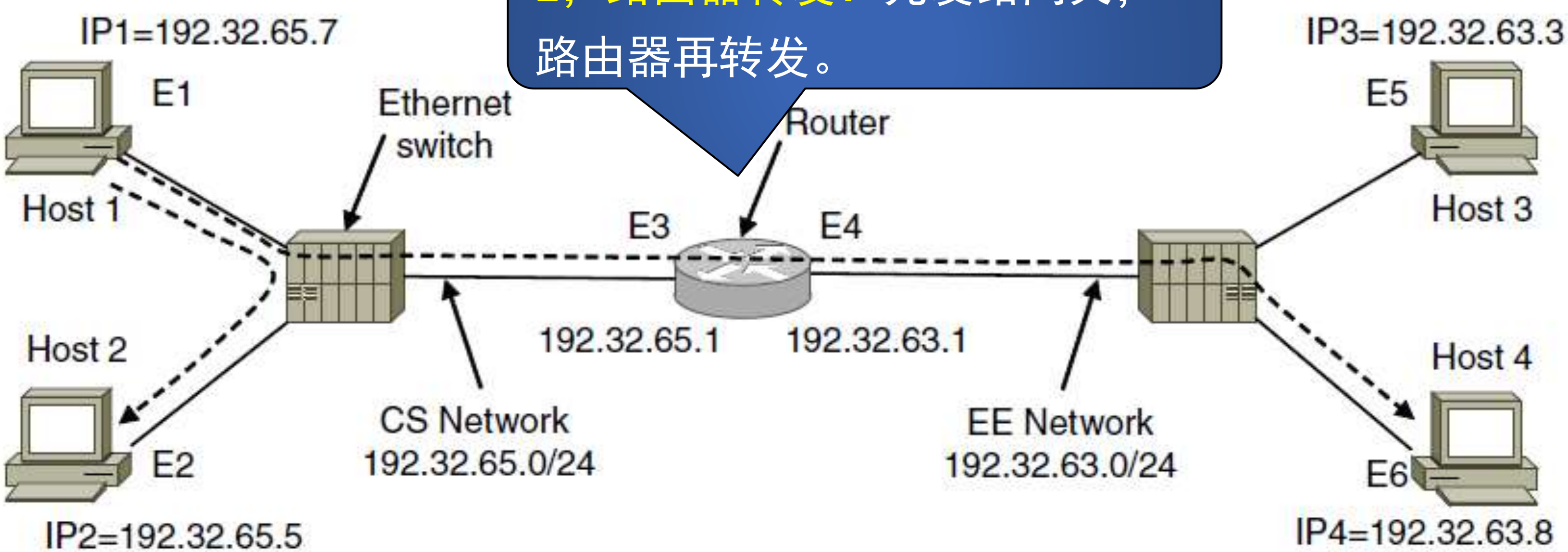
1） 运行过ARP，可以缓存起来

2） ARP-IP映射包含在它的ARP包里，反向就不用广播了

3）缓存超时，以允许动态。

ARP—The Address Resolution Protocol

1，路由器转发：先发给网关，路由器再转发。

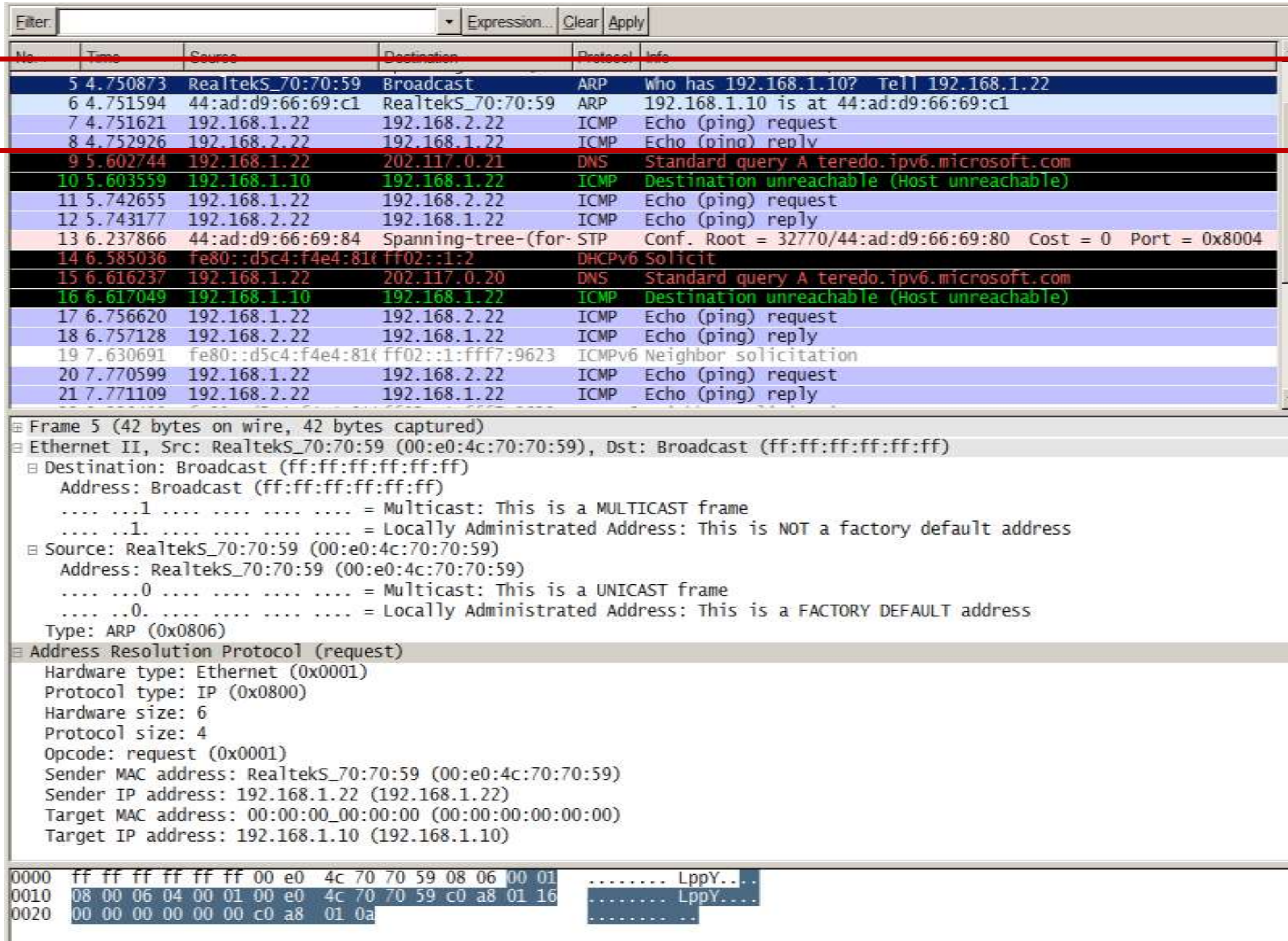


Frame	Source IP	Source Eth.	Destination IP	Destination Eth.
Host 1 to 2, on CS net	IP1	E1	IP2	E2
Host 1 to 4, on CS net	IP1	E1	IP4	E3
Host 1 to 4, on EE net	IP1	E4	IP4	E6

Two switched Ethernet LANs joined by a router

2，ARP代理：路由器回答自己的MAC，再转发。IP地址没变，而 MAC地址变了

ARP—The Address Resolution Protocol



DHCP

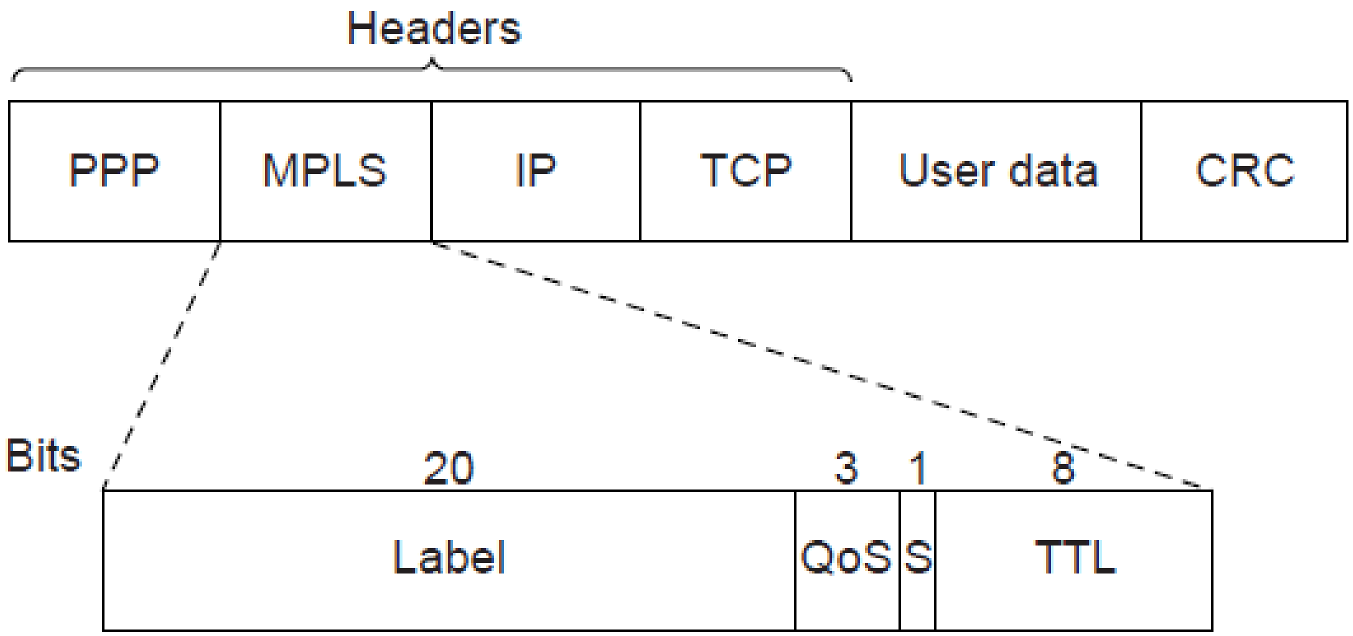
Dynamic Host Configuration Protocol

DHCP DISCOVER   DHCP OFFER

Label Switching and MPLS (1 of 3)

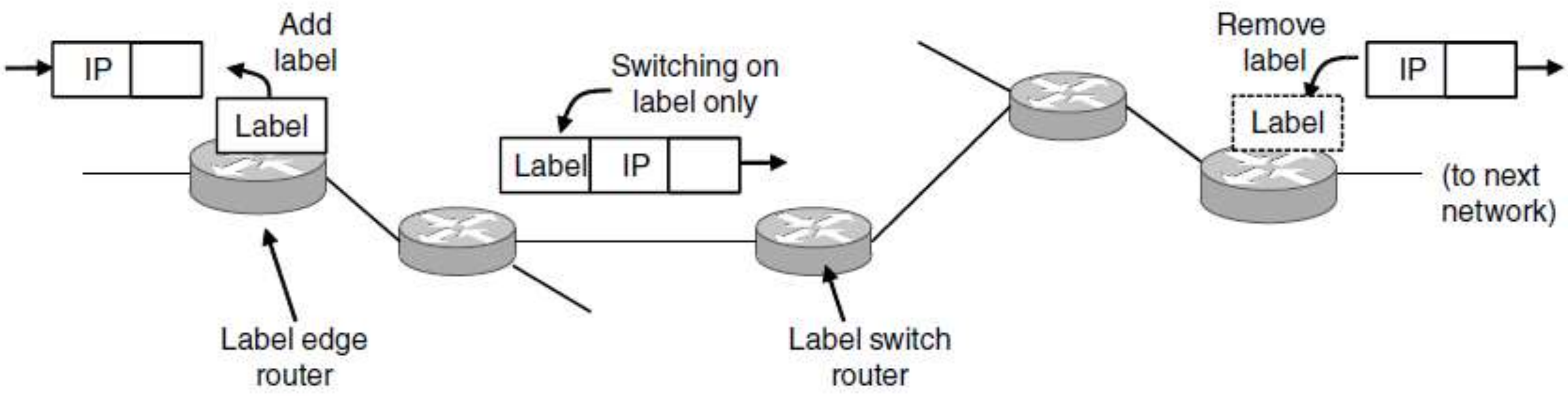
- MPLS (MultiProtocol Label Switching)
  - Perilously close to circuit switching
  - Adds a label in front of each packet
  - Forwards based on the label (not the destination address)
  - Forwarding can be done very quickly
- New MPLS header is added in front of the IP header

Label Switching and MPLS (2 of 3)



Transmitting a TCP segment using IP, MPLS, and PPP.

Label Switching and MPLS (3 of 3)



Forwarding an IP packet through an MPLS network

OSPF—An Interior Gateway Routing Protocol (1)

- Intradomain routing
  - IGP (Interior Gateway Protocol)
- RIP (Routing Information Protocol)
  - Works well in small systems
- OSPF (Open Shortest Path First)
  - Widely used in company networks
- IS-IS (Intermediate-System to Intermediate-System)
  - Widely used in ISP networks

OSPF—An Interior Gateway Routing Protocol (2)

- OSPF
  - Published in the open literature
  - Supports a variety of distance metrics
  - Dynamic
  - Supports routing based on type of service
  - Performs load balancing, splitting the load over multiple lines
  - Supports hierarchical systems
  - Provides security
  - Provision for dealing with routers that were connected to the Internet via a tunnel
- OSPF supports multiaccess networks

- 内部网关协议IGP（interior gateway prot

  自治系统AS内使用的路由算法，RIP、OSPF

- 外部网关协议EGP（exterior gateway pro col）

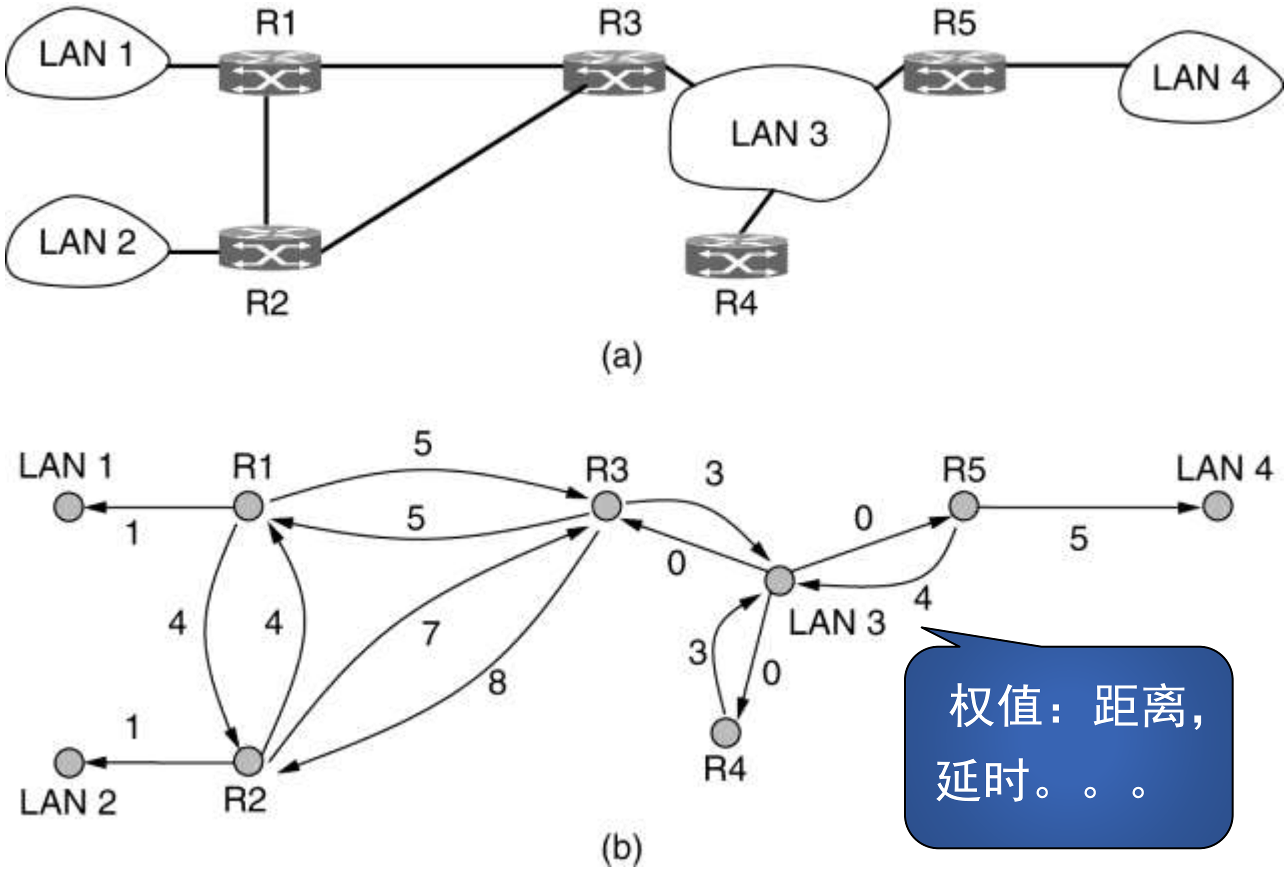
  自治系统AS之间使用的路由算法

RIP，UC Berkeley 的实验平台 Python

开放最短路径优先OSPF（Open Shortest Path First）

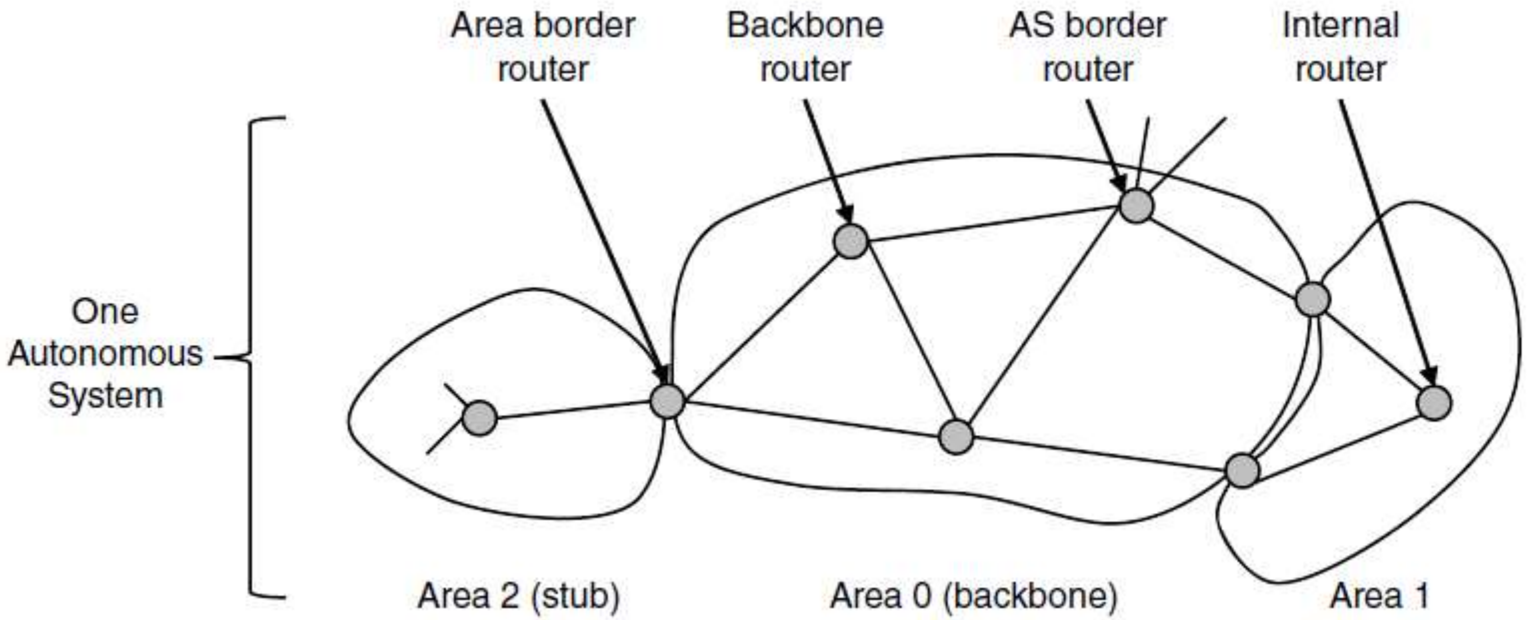
- 开放，公开发表；
- 支持多种距离衡量尺度，例如，物理距离、延迟等；
- 动态算法；
- 支持基于服务类型的路由；
- 负载平衡；
- 支持分层系统；

OSPF—An Interior Gateway Routing Protocol (3)



(a) An autonomous system. (b) A graph representation of (a).

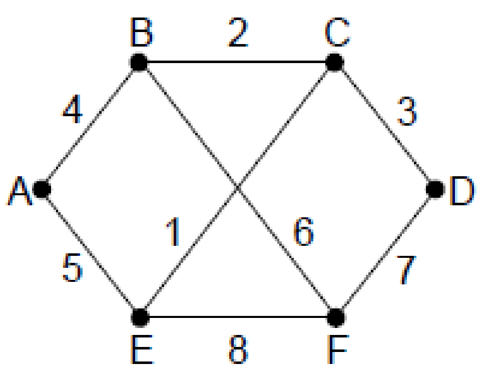
OSPF—An Interior Gateway Routing Protocol (4)



The relation between ASes, backbones, and areas in OSPF.

Flashback…

- just put in a sequence number and aging information. The hard part is when to build them. Practice shows that once an hour is often enough.



(a)

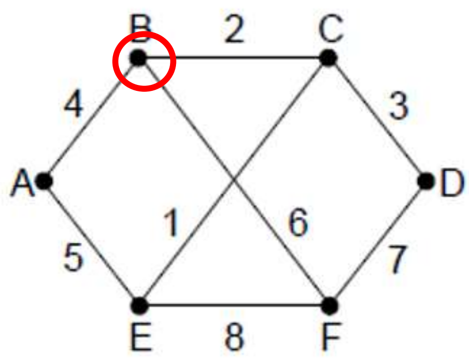
Link State Packets

A		B		C		D		E		F	
Seq.		Seq.		Seq.		Seq.		Seq.		Seq.	
Age		Age		Age		Age		Age		Age	
B	4	A	4	B	2	C	3	A	5	B	6
E	5	C	2	D	3	F	7	C	1	D	7
		F	6	E	1			F	8	E	8

(b)

(a) A network. (b) The link state packets for this network.

Flashback…



- 从E发来的链路状态包有两个，一个经过EAB，另一个经过EFB；
- 从D发来的链路状态包有两个，一个经过DCB，另一个经过DFB；

Source	Seq.	Age	Send flags A	Send flags C	Send flags F	ACK flags A	ACK flags C	ACK flags F	Data
A	21	60	0	1	1	1	0	0	
F	21	60	1	1	0	0	0	1	
E	21	59	0	1	0	1	0	1	
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

The packet buffer for router *B* in previous slide

OSPF—An Interior Gateway Routing Protocol (5)

Message type	Description
Hello	Used to discover who the neighbors are
Link state update	Provides the sender's costs to its neighbors
Link state ack	Acknowledges link state update
Database description	Announces which updates the sender has
Link state request	Requests information from the pa r

声明发送者的链路状态更新情况（链路状态表项的序号），通过与自己相应值的比较，决定谁拥有最新的值。

The five types of OSPF messages

OSPF—An Interior Gateway Routing Protocol (6)

No.	Time	Source	Destination	Protocol	Length	Info
236	208.638111	168.1.1.2	168.1.1.1	OSPF	66	DB Description
240	208.643046	168.1.1.2	168.1.1.1	OSPF	86	DB Description
241	208.643412	168.1.1.2	168.1.1.1	OSPF	70	LS Request
245	208.644946	168.1.1.2	168.1.1.1	OSPF	98	LS Update
246	208.645314	168.1.1.2	168.1.1.1	OSPF	78	LS Acknowledge
237	208.638524	168.1.1.1	168.1.1.2	OSPF	66	DB Description
238	208.640599	168.1.1.1	168.1.1.2	OSPF	86	DB Description
242	208.643799	168.1.1.1	168.1.1.2	OSPF	66	DB Description
243	208.644189	168.1.1.1	168.1.1.2	OSPF	70	LS Request
244	208.644578	168.1.1.1	168.1.1.2	OSPF	110	LS Update
223	167.226287	168.1.1.1	224.0.0.5	OSPF	78	Hello Packet
224	168.232424	168.1.1.1	224.0.0.5	OSPF	78	Hello Packet
227	178.333461	168.1.1.1	224.0.0.5	OSPF	78	Hello Packet
228	188.434497	168.1.1.1	224.0.0.5	OSPF	78	Hello Packet
229	198.535530	168.1.1.1	224.0.0.5	OSPF	78	Hello Packet
232	207.993789	168.1.1.2	224.0.0.5	OSPF	78	Hello Packet
233	208.636565	168.1.1.1	224.0.0.5	OSPF	82	Hello Packet

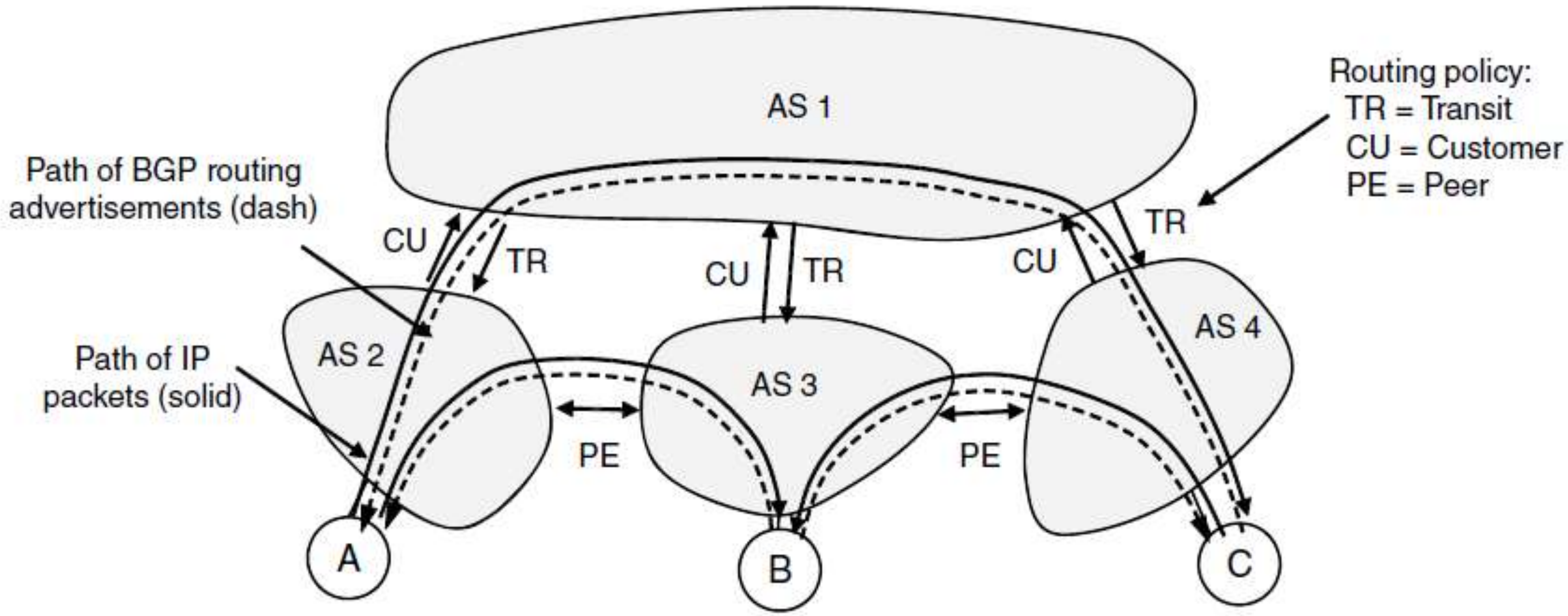
组播地址

BGP—The Exterior Gateway Routing Protocol (1 of 3)

BGP通告路由（Border Gateway Protocol）

- Possible routing constraints
  - Do not carry commercial traffic on the educational network
  - Never send traffic from the Pentagon on a route through Iraq
  - Use TeliaSonera instead of Verizon because it is cheaper
  - Don’t use AT&T in Australia because performance is poor
  - Traffic starting or ending at Apple should not transit Google

BGP—The Exterior Gateway Routing Protocol (2 of 3)



Routing policies between four Autonomous Systems

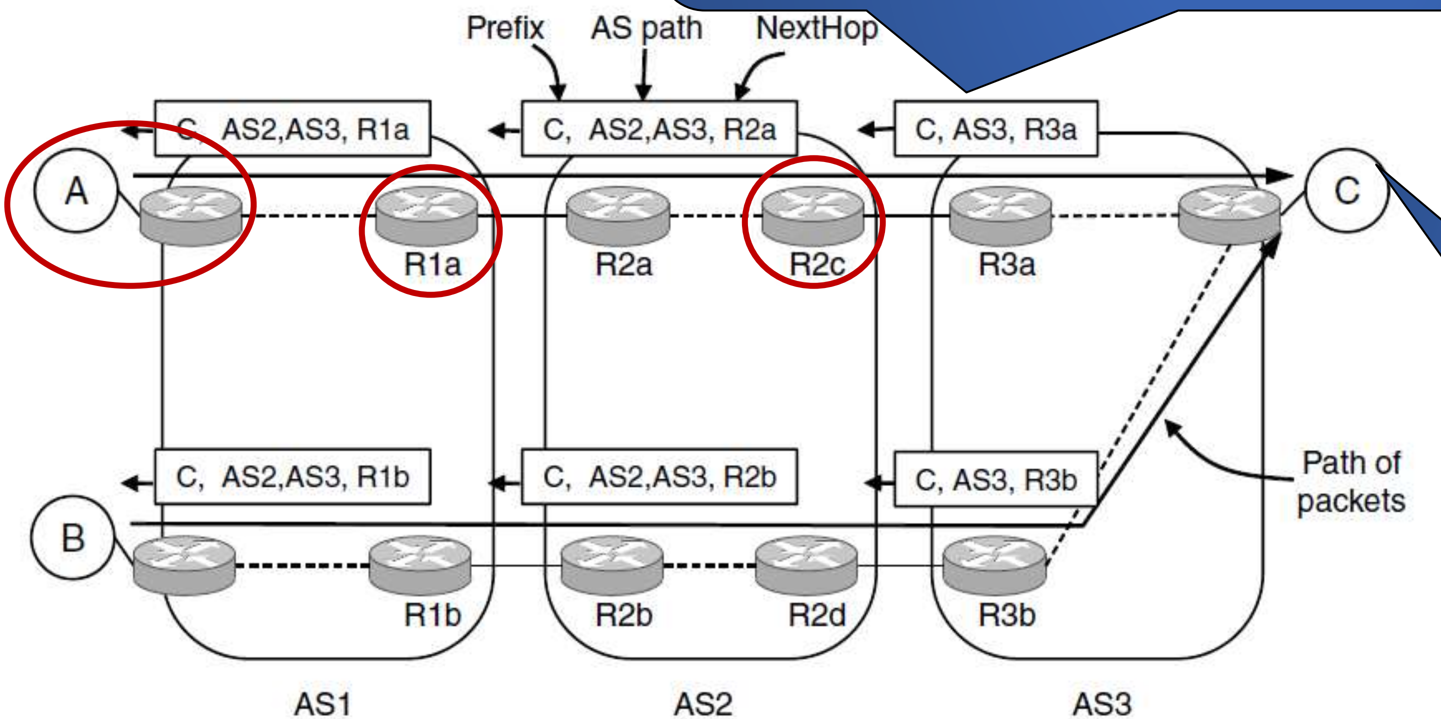
BGP—The Exterior Gateway Routing Protocol (3 of 3)

1）根据政策来

2）维护到每个目的地的成本，还跟踪路径，距离矢量

3）路由器通过TCP通信

发起到前缀C的路由通告



Propagation of BGP route advertisements

Interdomain Traffic Engineering

- Tune parameters and configuration network protocols to manage utilization and congestion
- Inbound traffic engineering
  - Selects routes to control how traffic enters the network
  - Set the local preference attribute for individual routes
  - Use AS path prepending
  - Leverage longest prefix match
    - Split a prefix into multiple smaller (longer) prefixes, so that upstream routers prefer the routes with longer prefixes
- Outbound traffic engineering
  - How traffic leaves the network

Internet Multicasting

- Internet multicasting
  - One-to-many communication using class D IP addresses
- Each class D address identifies a group of hosts
- Twenty-eight bits available for identifying groups
  - Over 250 million groups can exist at the same time
- Process sends a packet to a class D address
  - Best-effort attempt is made to deliver it to all the members of the group addressed, but no guarantees are given

Policy at the Network Layer

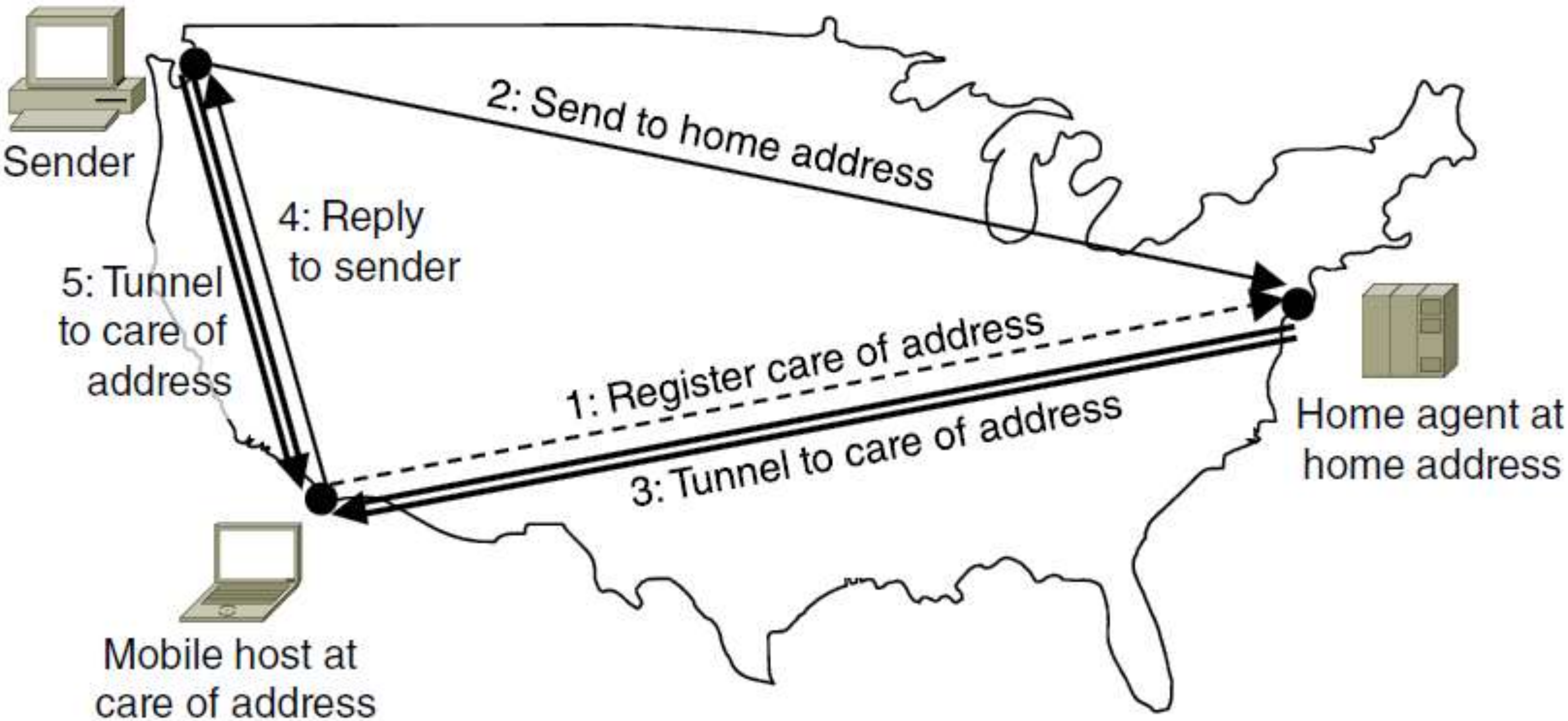
- Peering disputes
  - A breakdown in negotiations over paying for transit
- Traffic prioritization
- Generally agreed upon bright-line rules
  - No blocking
  - No throttling
  - No paid prioritization
  - Disclosure of any prioritization practices

Mobile IP

Goals

- Mobile host use home IP address anywhere.
- No software changes to fixed hosts
- No changes to router software, tables
- Packets for mobile hosts – restrict detours
- No overhead for mobile host at home.

Mobile IP



Sender
2: Send to home address
4: Reply
to sender
5: Tunnel
to care of
address
1: Register care of address
3: Tunnel to care of address
Home agent at
home address
Mobile host at
care of address

End

Chapter 5

- 服务
- 路由
- 拥塞控制
- QoS
- 不同网络的连接
- Internet