

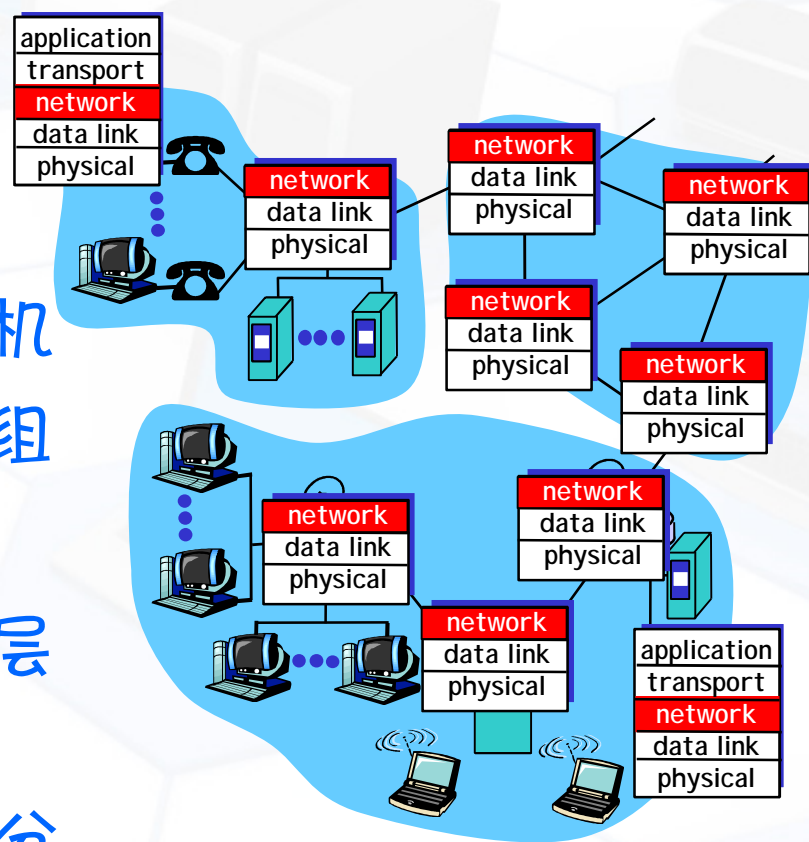
○、网络层问题

1. 网络层为什么提供的是不可靠服务？
2. 网络层只有IP协议吗？
3. 使用ARP协议查找对方的MAC地址时，如何让对方收到所发出的询问报文？
4. 如果被查询的站点B和查询站点A在不同的网段，B将如何回应A的ARP请求报文？

一、网络层简介

1. 概况

- § 将上层数据传送到对方主机
- § 在发送方将数据包装为分组
- § 在接收方将数据交传输层
- § 每个主机、路由都有网络层协议
- § 路由检查所有经过它的IP分组的头部字段



2. 两个网络层的关键功能

§ 转发: 将分组自路由器入口送到合适的路由器出口

§ 路由: 根据路由算法确定分组自源到目的路线

p 网络层主要解决的问题

§ 路由选择

§ 网络互连

§ 拥塞控制

§ 向上层提供服务

p 网络层的服务模式

§ 服务应与通信子网的技术无关

§ 通信子网的数量、类型和拓扑结构对于传输层是隐蔽的

§ 传输层能获得的网络地址应采用统一的编号方式, 即使跨越了多个LAN和WAN



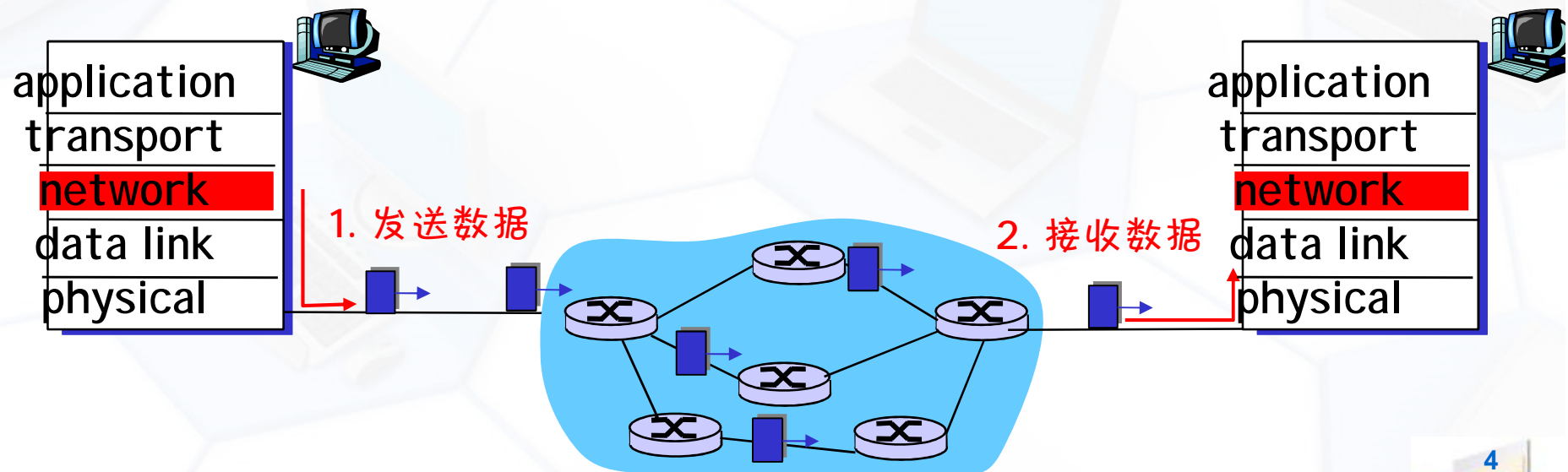
3. 基于数据报的网络

§ 网络层不事先建立连接

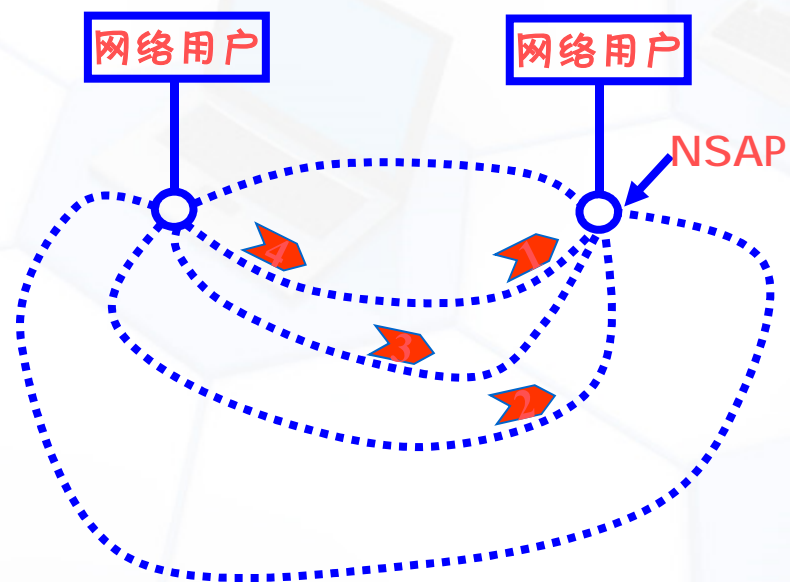
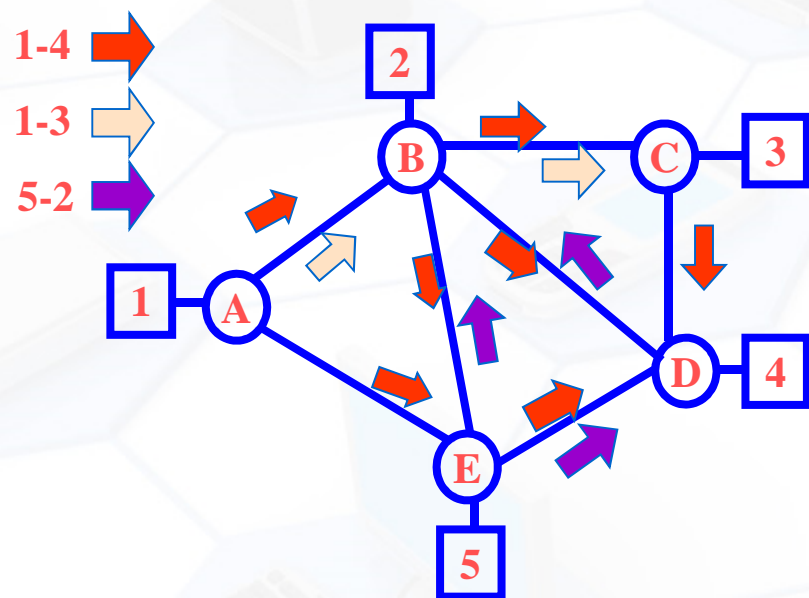
§ 路由器没有端到端连接的概念

§ 使用目标主机的地址来发送分组

ü 同一目的与源之间的分组传递可以经过不同的路线



p 每个数据报包含全部的地址，自行寻找路径

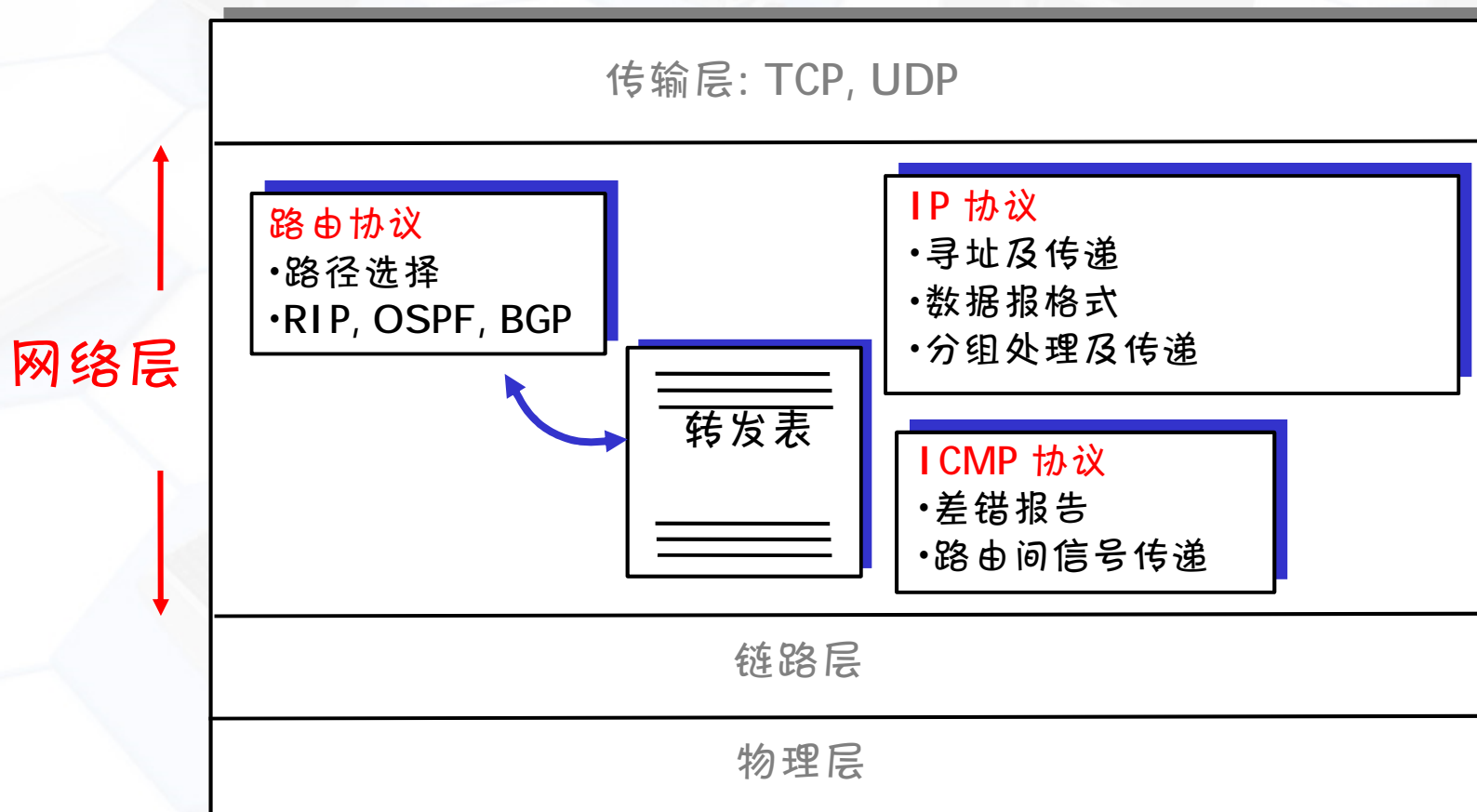


数据报子网



网络层

4. 主机、路由器的网络层功能结构

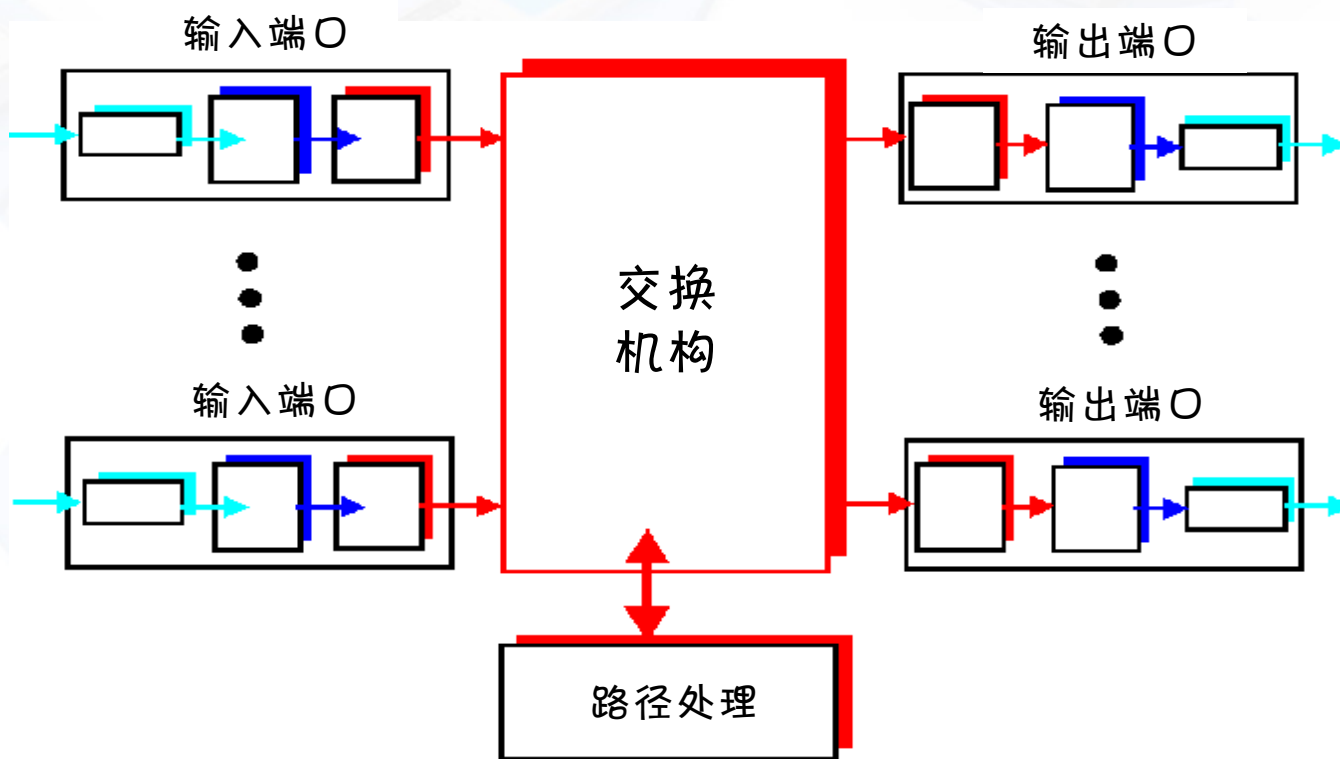


5. 路由器结构概述

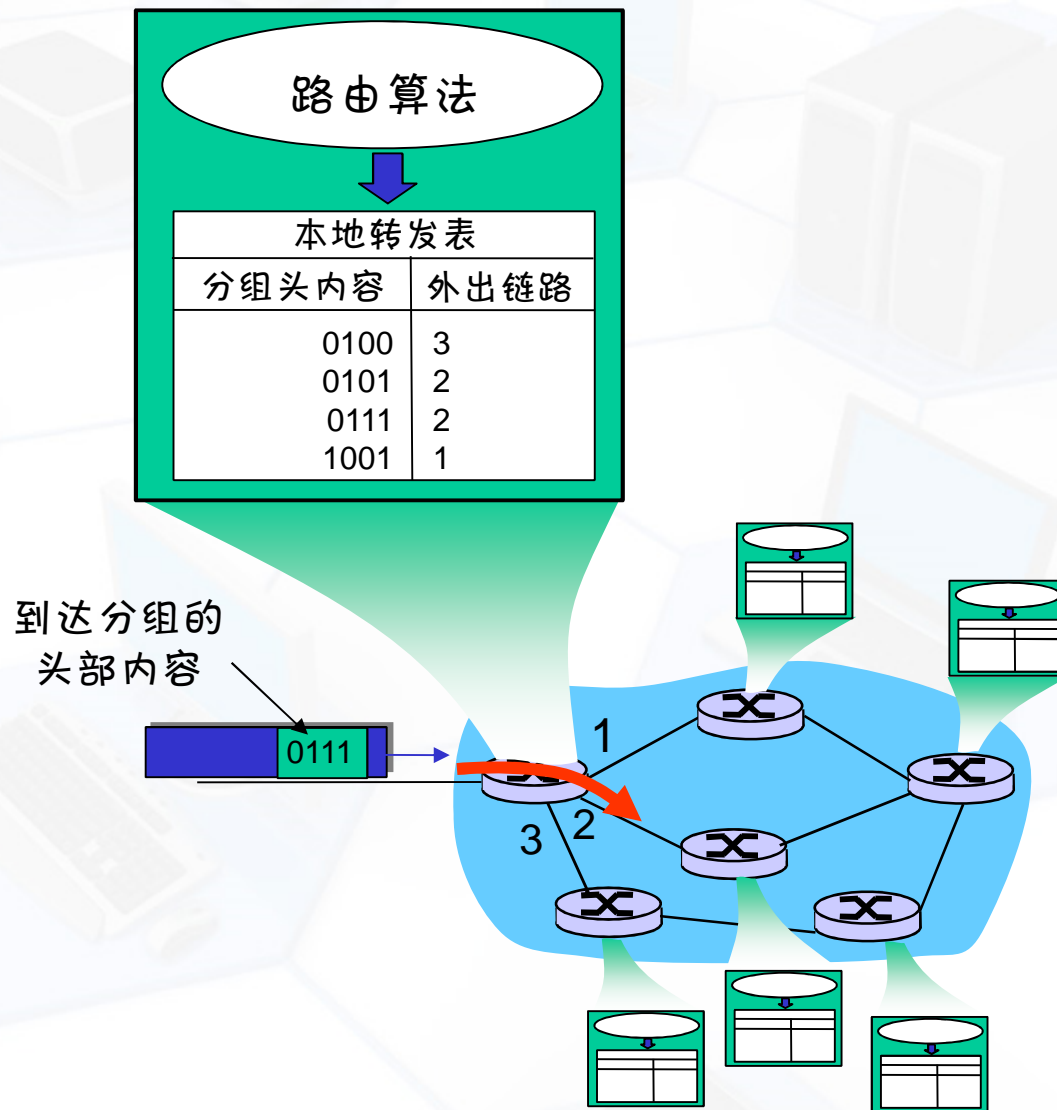
p 路由的两个关键功能

ü 运行路由算法/协议 (RIP, OSPF, BGP)

ü 向相关的外出链路转发分组



p 路由与转发的内部操作



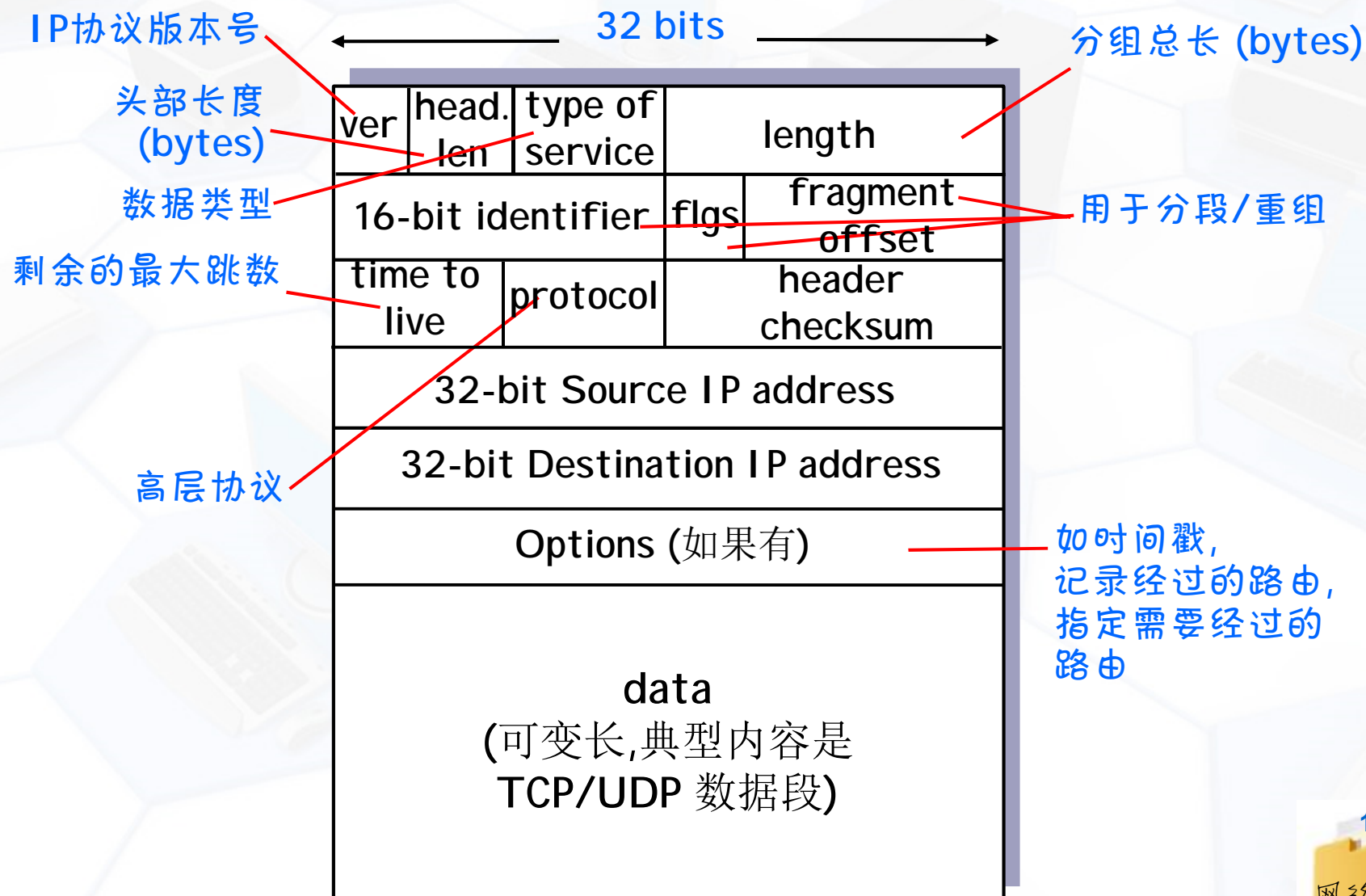


通过路由连接时的数据流向变化

二、IP协议

1. IP数据报格式
2. IP地址
3. 划分子网
4. CIDR无类域间路由
5. IPv6

1. IP数据报格式



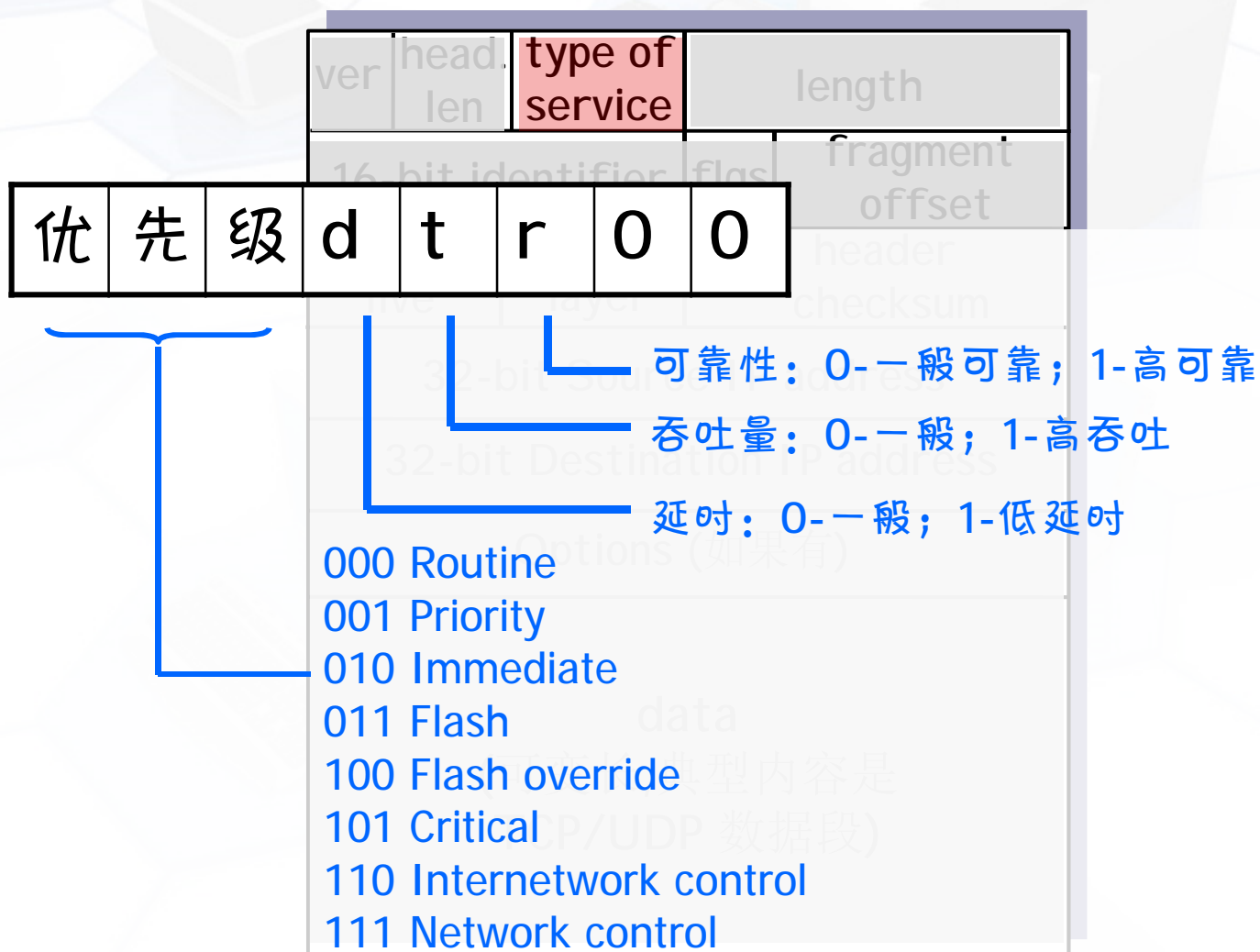
1) 版本号与头标长度

ver	head. len	type of service	length	
16-bit identifier		flgs	fragment offset	
time to live	upper layer	header checksum		
32-bit Destination IP address				
data				
(可变长,典型内容是 TCP/UDP 数据段)				

§ 版本号 (VERS) : 4bits, IPv4协议填4, IPv6协议填6

§ IP分组头长度 (LEN) : 4bits, 单位为4字节, 取值范围5-15 (确省值为5, 即标准头标长20字节), 指示IP分组头的长度

2) 服务类型 (TOS)



3) 总长度、分段功能

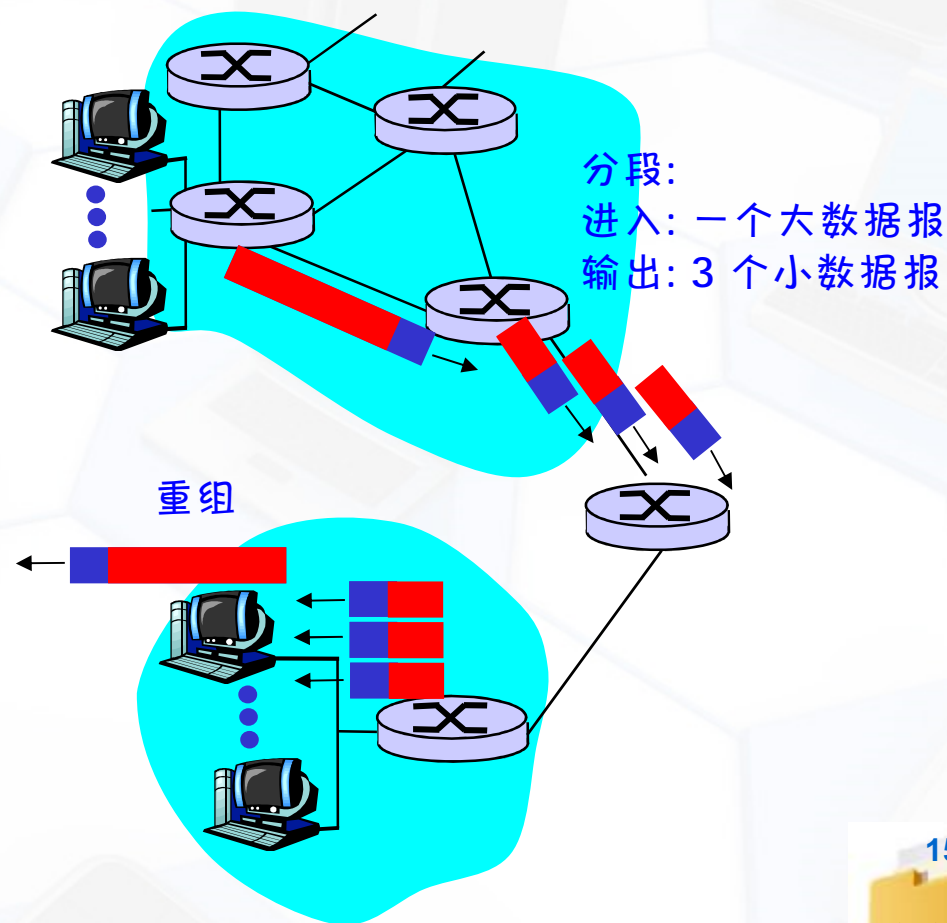
ver	head. len	type of service	length	
16-bit identifier			flgs	fragment offset

- n 总长度：16bits**，单位字节，描述IP分组的总长（包括头和数据），最大分组长度为65535字节
- n 标识符：16bits**，用于唯一标识该分组
- n 标志：3bits**，第1位未定义，第2位为0表示该分组可分段，否则表示不可分段；第3位为0表示这是最后分段，否则则表示还有后续分段
- n 段偏移：13bits**，单位8字节。取值0-8191，标明当前分段在原分组中的位置

(a) 为什么要分段?

- § 网络链路存在MTU (最大可传输尺寸)限制
 - ∅ 不同链路类型的MTU各不相同
- § 过大的IP分组在子网中被分割 (分段)
 - ∅ 一个数据报会变成几个小的数据报
 - ∅ 重组仅在目的地进行
 - ∅ 在IP分组头部标识各分段的信息

MTU: Maximum Transmission Unit, 最大传输单元, 即一帧所能携带的最大IP分组, 包括IP头标



(b) 怎样分段?

p 按MTU及数据包的实际负载长度计算所需段数，并划分，分段应满足两个条件：

§ 各段在不大于MTU的前提下，尽可能的大

§ 段的长度为8的整倍数

p 原数据包的报头作为每段的数据包报头，并修改其中的某些字段，指明：

§ 属原来的哪个段

§ 属原来段中的第几个分段

§ 哪一个为段尾

(c) 分段过程举例



MTU包括IP头标，此处假设IP头标取固定长度20字节

假设

ü 4000 字节数据报

ü MTU = 1500 字节

1480 bytes in
data field

offset =
 $1480/8$

	length	ID	fragflag	offset	
	=4000	=x	=0	=0	

	length	ID	fragflag	offset	
	=1500	=x	=1	=0	

	length	ID	fragflag	offset	
	=1500	=x	=1	=185	

	length	ID	fragflag	offset	
	=1040	=x	=0	=370	

IP分段示例

(d) 数据报分段的重组

p 重组是在各分段都到达目的地后才进行

§ 每个分段可以走不同的路径

§ 减少路由器中保存的信息量及路由器的工作量

p 途中的任意一个路由器都无法重组

§ 重组必须在所有的分段全部收到后，才可进行

p 互联网层是遵循尽力而为来传送IP包的，
也存在力不从心的时候，此时只能丢弃

§ 重组主机将遵循：要么重组成功，要么全部
丢弃的原则

4) 生存时间TTL、协议类型

§ 生存时间(Time to Live): 8bits, 单位秒, 表示分组的生存时间。实际操作时, 分组每经过一个路由器, TTL值减1, 当TTL值为0时, 该分组被丢弃

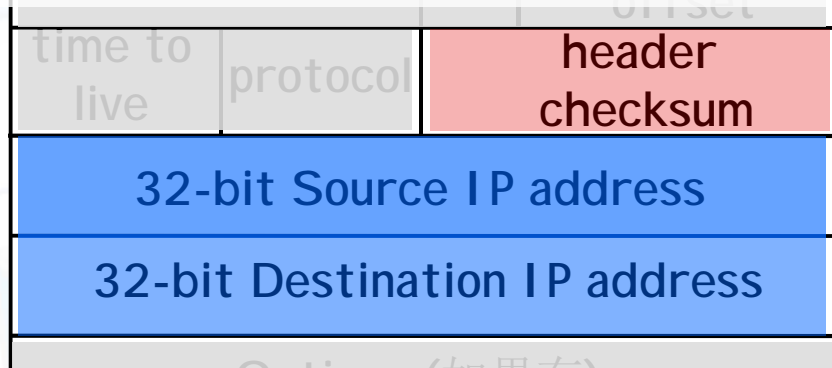
time to live	protocol	header checksum
--------------	----------	-----------------

§ 协议 (Protocol) : 表示高层协议类型

- Ø 0 Reserved
- Ø 1 Internet Control Message Protocol (ICMP)
- Ø 2 Internet Group Management Protocol (IGMP)
- Ø 3 Gateway-to-Gateway Protocol (GGP)
- Ø 4 IP (IP encapsulation)
- Ø 5 Stream
- Ø 6 Transmission Control (TCP)
- Ø 8 Exterior Gateway Protocol (EGP)
- Ø 9 Private Interior Routing Protocol
- Ø 17 User Datagram (UDP)
- Ø 89 Open Shortest Path First(OSPF)

5) 分组头校验、源/目的地址

分组头校验：16bits，用来检验IP头标在传输过程中是否被破坏；按16位相加，结果求反



源地址：32bits，分组发送者的IP地址

目的地址：32bits，分组接收者的IP地址

(可变长, 典型内容是
TCP/UDP 数据段)

分组头校验举例

<pre> IP: ...0 = normal delay IP: 0... = normal throughput IP: 0.. = normal reliability IP: 0. = ECT bit - transport protocol will not be marked ECN-capable IP: 0 = CE bit - no congestion IP: Total length = 52 bytes IP: Identification = 8143 IP: Flags = 4X IP: .1... = don't fragment IP: ..0. = last fragment IP: Fragment offset = 0 bytes IP: Time to live = 64 seconds/hops IP: Protocol = 6 (TCP) IP: Header checksum = 3CA9 (correct) </pre>	<pre> 45 00 00 34 1F CF 40 00 40 06 00 00 DA F6 BC FD DA 1E 6C 39 ----- 3 C3 53 </pre>
---	---

00000000: 00 01 30 ff b7 c0 00 fd 00 00 1b 45 08 00 45 00 ..

00000010: 00 34 1f cf 40 00 40 06 3c a9 da f6 bc fd da 1e .4

00000020: 6c 39 06 2f 00 50 c5 2a fa 5a 00 00 00 00 80 02 19

00000030: ff ff ca c5 00 00 02 04 05 b4 01 03 03 02 01 01

00000040: 04 02 ..

把进位3加到结果后: 3 + C3 53 = C3 56

C 3 5 6
 1100 0011 0101 0110 求反码: 0011 1100 1010 1001 得到 3CA9

路由器每次都会重新计算

6) 选项

- § 安全性 (Security) : 指明分组的机密性
- § 严格的源路由选择 (Strict Source routing) : 给出分组经过的完整路由
- § 松散的源路由选择 (Loose Source routing) : 给出分组经过的某些路由器列表
- § 路由记录 (Route recording) : 使每个路由器都附上它的IP地址
- § 时间标记 (Time stamping) : 使每个路由器都附上它的IP地址和时间标记

Options (如果有)

- § 填充 (padding) : 分组头长度必须为4字节的整数倍, 如果选项的长度不是4字节的整数倍, 那么就要进行填充

TCP/UDP 数据段)

No.	Time	Source	Destination	Protocol	Info
1	0.000000	14:da:e9:f1:3e:48	Spanning-tree-(for-br	STP	Conf. Root = 32768/0/14:da:e9
2	0.356421	192.168.1.2	61.172.201.195	ICMP	Echo (ping) request
3	0.359541	61.172.201.195	192.168.1.2	ICMP	Echo (ping) reply
4	1.358871	192.168.1.2	61.172.201.195	ICMP	Echo (ping) request
5	1.361808	61.172.201.195	192.168.1.2	ICMP	Echo (ping) reply
6	1.452201	Giga-Byt_7hidd:b9	14:da:e9:f1:3e:48	ARP	Who has 192.168.1.12? Tell 192

Frame 3 (74 bytes on wire (588 bits) captured (472 bits) on 0

Ethernet II, Src: 14:da:e9:f1:3e:48 (14:da:e9:f1:3e:48), Dst: Giga-Byt_7 (6c:f0:49:71:b9)

Destination: Giga-Byt_7 (6c:f0:49:71:b9)

Source: 14:da:e9:f1:3e:48 (14:da:e9:f1:3e:48)

Type: IP (0x0800)

Internet Protocol, Src: 61.172.201.195 (61.172.201.195), Dst: 192.168.1.2 (192.168.1.2)

Version: 4

Header length: 20 bytes

Differentiated Services Field: 0x00 (DSCP 0x00: Default; ECN: 0x00)

0000 00.. = Differentiated Services Codepoint: Default (0x00)

.... ..0. = ECN-Capable Transport (ECT): 0

.... ...0 = ECN-CE: 0

Total Length: 60

Identification: 0x1185 (4485)

Flags: 0x00

0.. = Reserved bit: Not Set

.0. = Don't fragment: Not Set

..0 = More fragments: Not Set

Fragment offset: 0

Time to live: 247

Protocol: ICMP (0x01)

Header checksum: 0xe921 [correct]

[Good: True]

[Bad : False]

Source: 61.172.201.195 (61.172.201.195)

Destination: 192.168.1.2 (192.168.1.2)

Internet Control Message Protocol

0000 6c f0 49 71 b9 14 da e9 f1 3e 48 08 00 45 00

0010 00 3c 11 85 00 00 f7 01 e9 21 11 85 00 00 00

0020 01 02 00 00 55 2c 00 01 00 2f 61 62 63 64 65 66

0030 67 68 69 6a 6b 6c 6d 6e 6f 70 71 72 73 74 75 76

0040 77 61 62 63 64 65 66 67 68 69 6a 6b 6c 6d 6e 6f

IP 报文示例

24

网络层

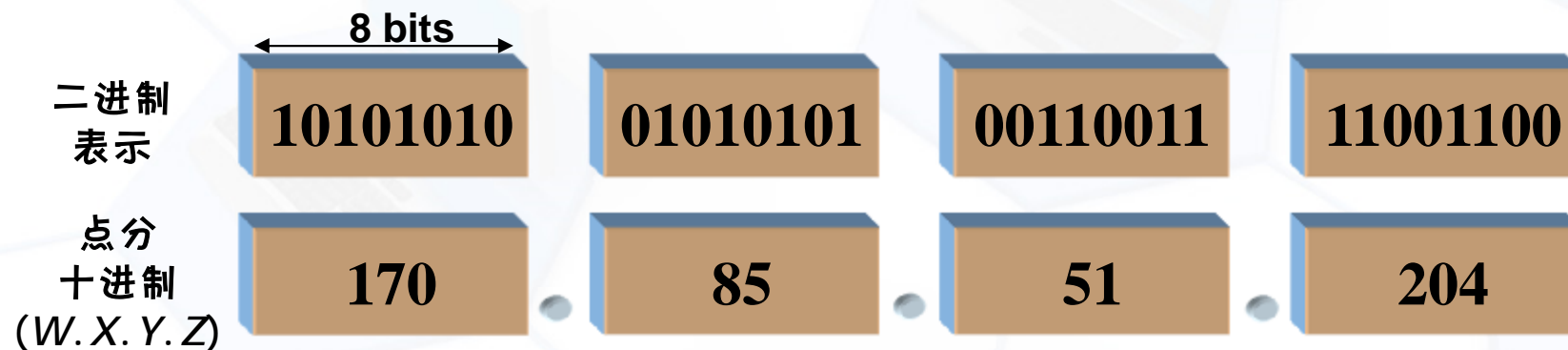
2. IP地址

IP地址构成(逻辑地址)

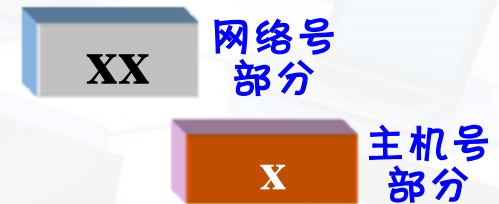
10011010 10011110 00001111 10101010



32位



1) IP地址分类



Class A	0XXXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXXX	0.0.0.0 – 127.255.255.255
Class B	10XXXXXX	XXXXXXXXXX	XXXXXXXXX	XXXXXXXXX	128.0.0.0 – 191.255.255.255
Class C	110XXXXX	XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXX	192.0.0.0 – 223.255.255.255
Class D	1110XXXX	组 XXXXXXXXXX	播 XXXXXXXXXX	地址 XXXXXXXXXX	224.0.0.0 – 239.255.255.255
Class E	11110XXX	保 XXXXXXXXXX	留 XXXXXXXXXX	地址 XXXXXXXXXX	240.0.0.0 – 247.255.255.255

- § A类：最高位为0，随后7位为网络号，最后24位表示主机号。可以标识126个A类网络，每个网络可以有 $2^{24}-2$ （约1600万）个主机
- § B类：最高两位10，随后14位为网络号，最后16位表示主机号。可以标识 $2^{14}-2$ （约16000）个B类网络，每个网络可以有 $2^{16}-2$ （约65000）个主机
- § C类：最高三位为110，随后21位为网络号，剩下8位为主机号。可以标识200万个C类网络，每个网络最多只能有254个主机
- § D类：最高四位为1110，是组播地址，标识一个组的地址
- § E类：最高五位为11110，是保留地址

地址类别	网络数	主机数
A	0~127 (128)	16777216
B	128~191 (16384)	65536
C	192~223 (2097152)	256

2) 特殊IP地址

§ IP地址中网络号或主机号为全0或全1的一般用做特殊处理，不用来标识网络或主机

有限广播

全1

定向广播

网络号

全1

回环地址
(loopback)

127

任意

用做测试

全0

主机号

全0

3) IP地址分配

p Internet上的IP地址由IANA负责分配和管理

ü IANA将super-block（网络号较短）的地址块分配给地区Internet注册机构RIR，例如APNIC

ü RIR将地址块细分，将网络号更长的地址块分配给Internet服务提供商ISP或者企业

p IP地址配置

ü 手动配置

ü 拨号用户通过PPP协议分配

ü 基于DHCP（Dynamic Host Configuration Protocol）的配置

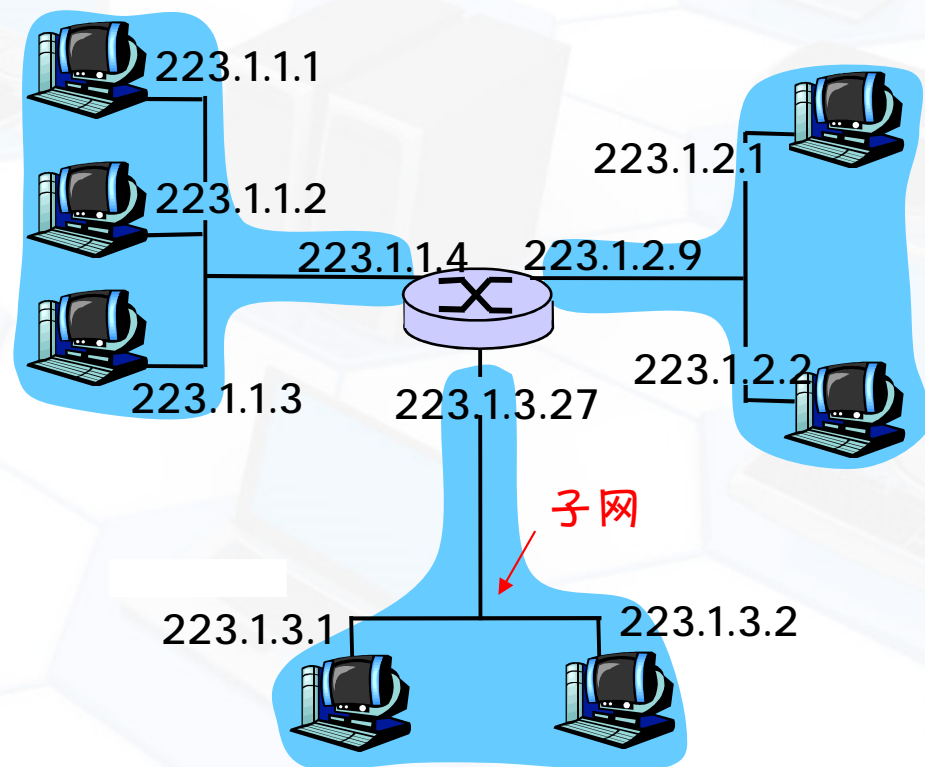
注：IANA (Internet Assigned Numbers Authority)
RIR (Regional Internet Registry)
ISP (Internet Service Provider)

Table 1: The Percentage of IP Address Ownership by Country in 2007

ISO Country Code	Country Name	Percentage	Ranking
US	UNITED STATES	37.7270%	1
UK	UNITED KINGDOM	12.8334%	2
JP	JAPAN	7.6421%	3
CN	CHINA	5.7362%	4
DE	GERMANY	3.8051%	5
FR	FRANCE	3.6514%	6
CA	CANADA	2.8189%	7
KR	KOREA REPUBLIC OF	2.7441%	8
NL	NETHERLANDS	1.9983%	9
IT	ITALY	1.6730%	10
AU	AUSTRALIA	1.6198%	11
ES	SPAIN	1.0997%	12
SE	SWEDEN	1.0846%	13
BR	BRAZIL	1.0495%	14
CH	SWITZERLAND	0.9685%	15
TW	TAIWAN	0.9313%	16

4) 子网

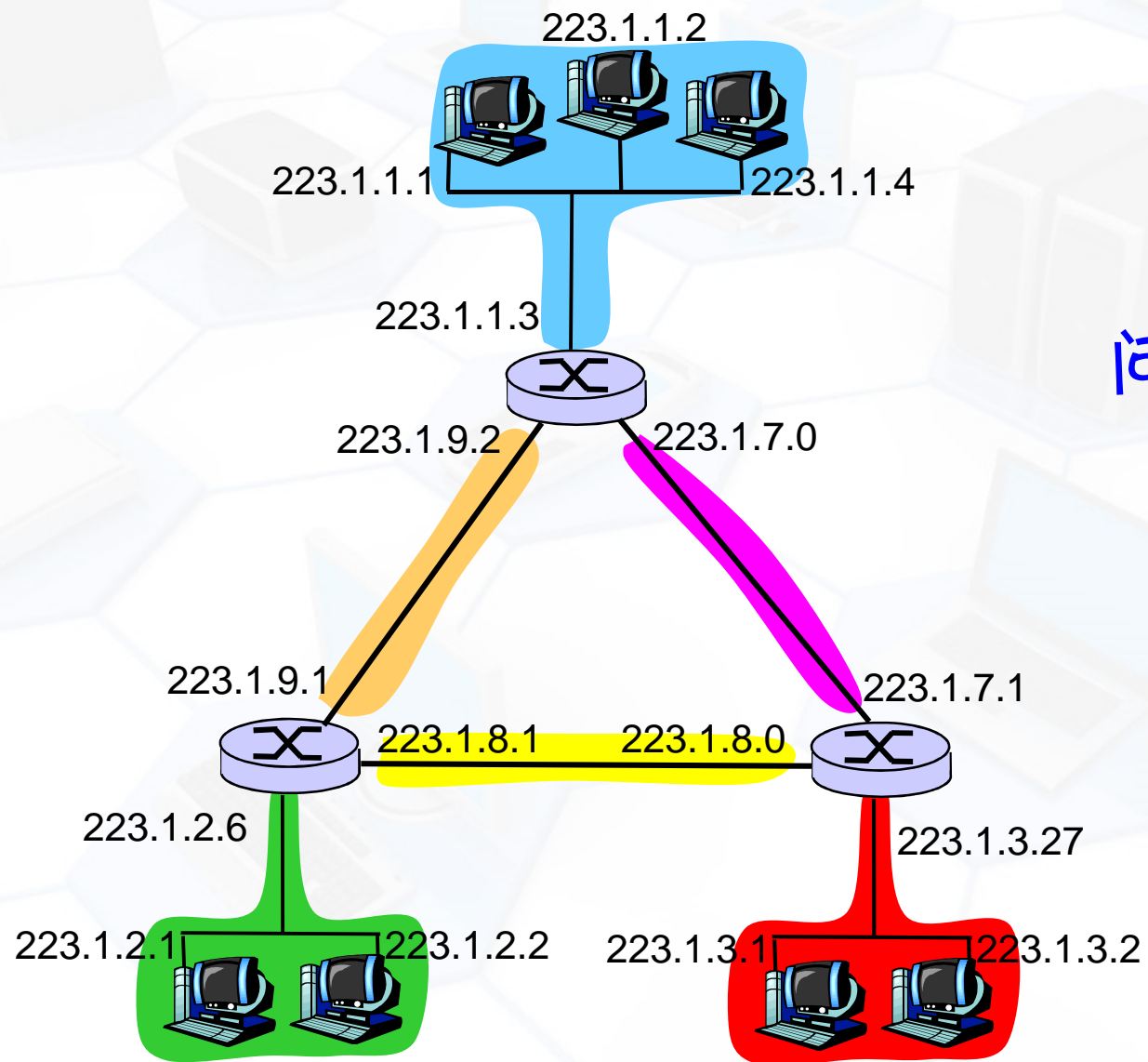
- § IP地址反映在每个物理接口上
- § 路由器通常有多个接口
- § 主机通常只有一个接口
- § IP地址分为网络号和主机号两部分
- § IP地址具有相同网络号的设备接口属于同一个子网
- § 这些接口不需要路由参与就可以相互访问



由3个子网构成的网络

5) 子网掩码

- p** 子网掩码的作用：因为子网部分位数不是固定的，需要告知设备IP地址的哪一部分是属于子网地址段，哪一部分是主机地址段
- p** 子网掩码使用与IP编址相同格式：子网掩码的网络地址部分全为1，主机部分全为0
- p** 默认的各类IP地址对应的掩码为：
 - ü A类：255.0.0.0
 - ü B类：255.255.0.0
 - ü C类：255.255.255.0
- p** 简便表示方法：IP地址/网络号位数，如：
192.168.1.1/24



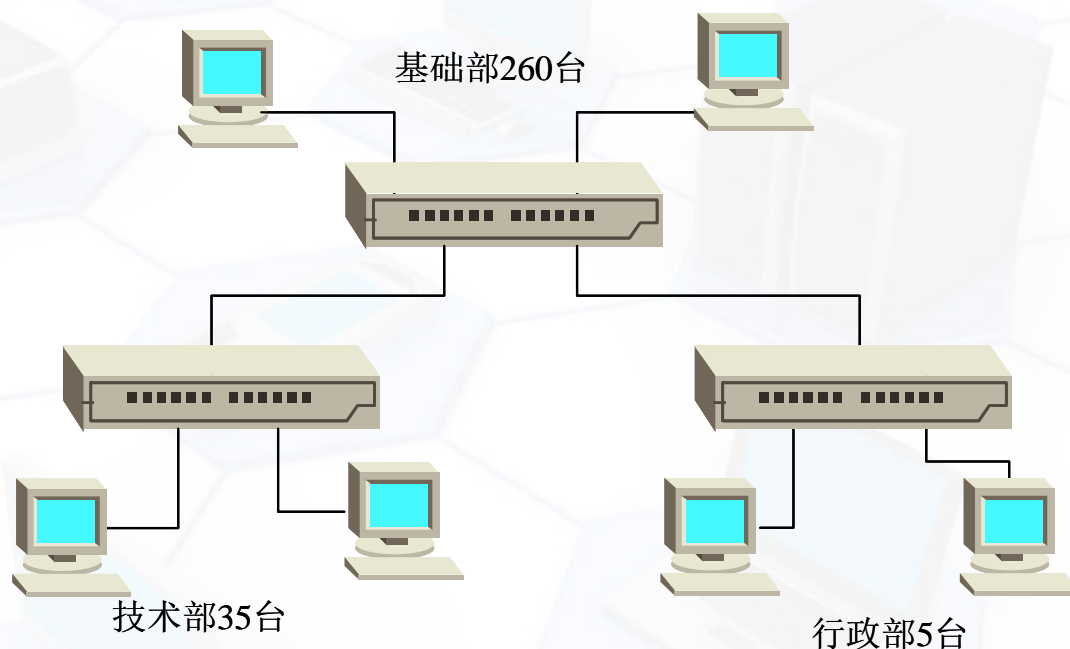
问题：有多少
个子网？

3. 划分子网

1) IP地址的规划

p 在没有特殊要求情况下，划分IP地址时可以有很大的自由度，只要满足子网间相互独立的要求，并且子网中IP地址的数量能够覆盖（或超过）实际主机的数量即可，当然还应该把那些保留地址排除在外

p IP地址划分举例



某单位现有三个部门，其中，基础部有260台电脑，技术部有35台电脑，行政部有5台电脑，所有的电脑均连接在同一个物理网络，需要根据合理的IP地址设置，使部门间相互独立，即各部门属于独立的IP子网

pIP地址划分结果

部 门	IP地址范围	子网掩码
基础部（260台）	172.1.1.1 ~ 172.1.2.5	255.255.0.0
技术部（35台）	192.1.1.1 ~ 192.1.1.35	255.255.255.0
行政部（5台）	10.1.1.1 ~ 10.1.1.5	255.0.0.0

2) IPv4地址问题

p 地址空间出现不足

ü 32位的IPv4地址总共有 2^{32} (4,294,967,296) IP地址，但根据地址的分类方式，实际可用的地址要少得多

p 骨干网路由表急剧膨胀

ü 路由器对于每一个网络都需要在路由表中增加一条表项

解决思路

p 采用层次地址结构

ü 提高IP地址的利用率

ü 对IP地址进行汇聚，减少路由协议中携带的和路由器的路由表中储存的网络号的数量

措施1：划分子网、无类域间路由

p 多台主机共享一个全局的IP

措施2：网络地址转换NAT

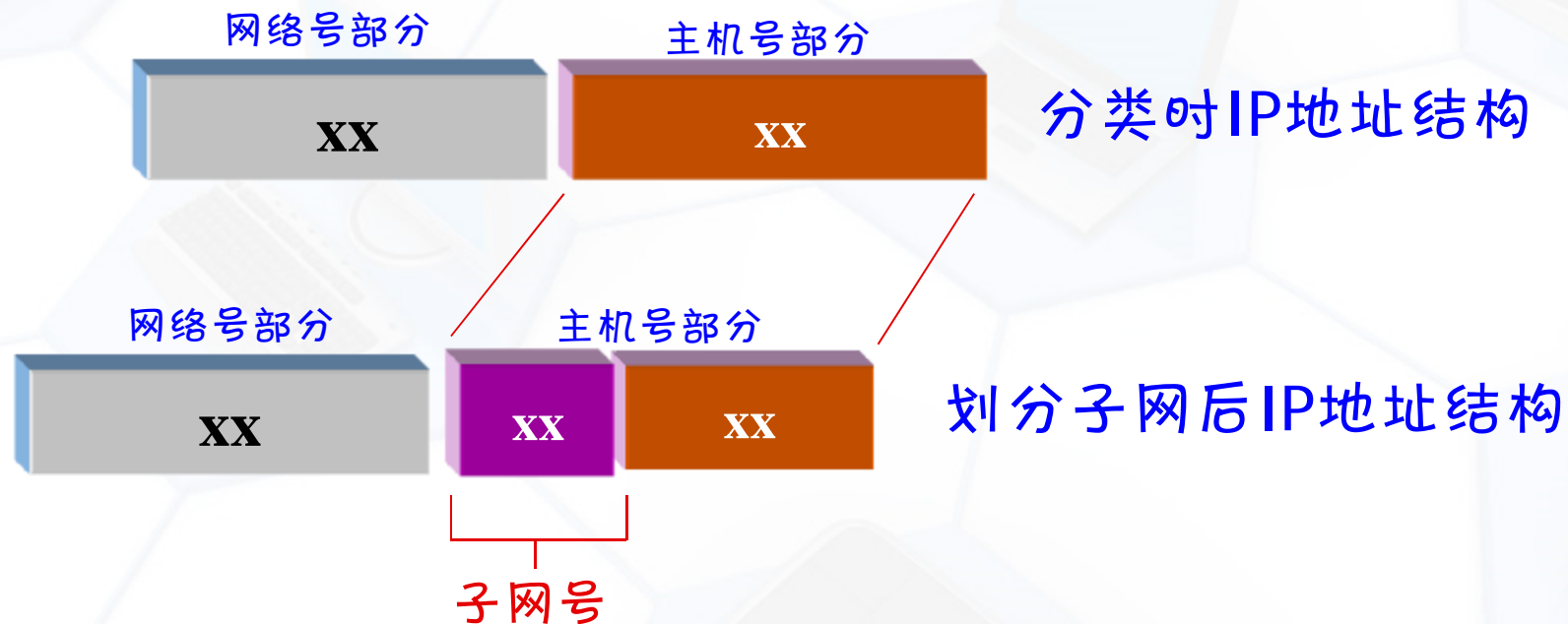
p 最终的解决方案是采用IPv6

3) 划分子网

p 大的A/B/C类网络划分为多个小子网（subnet），
对外仍然表现为一个单独的网络，只有一个网络号

p IP地址=网络号+子网ID+主机号

p 子网ID部分向原主机号部分借用



子网地址位数的确定

p 确定子网号的长度或者主机号的长度

ü 子网数量决定子网ID的长度，从而决定了子网号的长度

ü 主机数量决定了主机号的长度

p 根据子网号的长度确定子网掩码，然后根据子网掩码确定子网号

p 根据子网掩码，子网号等信息配置每个路由器上的转发表

p 子网地址位数的分析举例—— 使用1位

C类地址

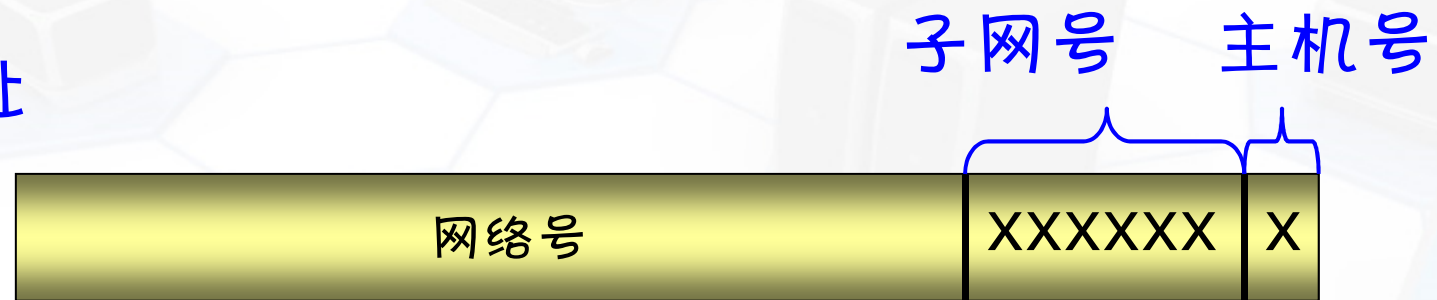


- ü 子网地址=0：表示本子网主机，不可作为有效目的地址使用
- ü 子网地址=1：子网地址全1，不可用
- ü 故：至少要借2位

注：全“0”的地址保留为识别子网自身，全“1”的子网地址用于在子网内广播

p 子网地址位数的分析举例—— 使用7位

C类地址



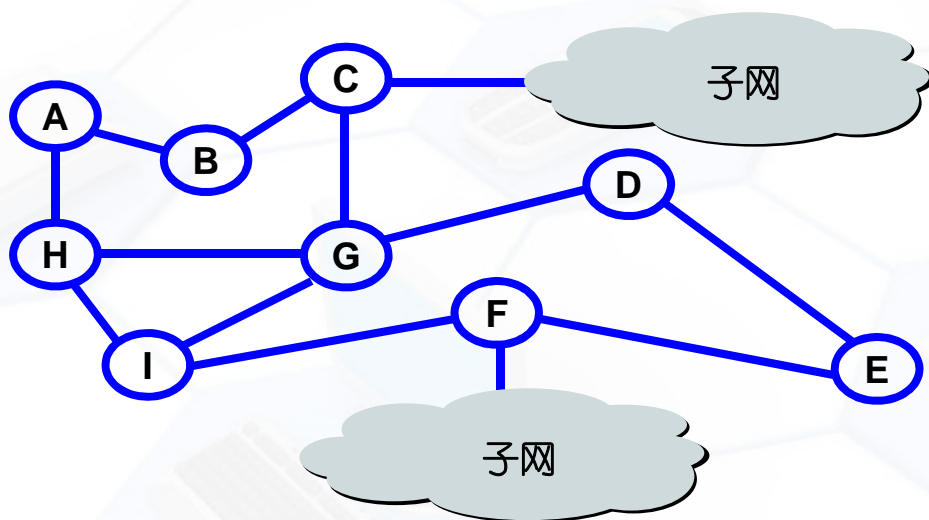
- ü 主机地址=0：子网地址，不可作为地址分配
- ü 主机地址=1：广播地址，不可分配
- ü 因此最多借6位

p 举例1：一个主机地址为202.120.3.99，子网地址 = 011的子网掩码是：

	网络号			子网号	主机号
C类IP地址	11001010	01111000	00000011	011	00011
	202	120	3	99	
掩码	11111111	11111111	11111111	111	00000
	255	255	255	224	

简洁的方式表示：202.120.3.99/27，其中27表示掩码中1的个数

p 例2：某主干网有一C类地址201.122.8.0，主干网的结构如下所示，有A、B、C、...H、I等9个路由器，每个路由器都有一个局域网与之连接，应如何设计子网掩码，以分配路由器的端口地址？



一个原则：两个路由器的相互连接的两个端口必须
在同个子网

所以，掩码应为255.255.255.248
即 201.122.8.0/29

4. CIDR无类域间路由(Classless InterDomain Routing)

- p 采用可变长度的网络号来取代地址分类中网络号长度固定的做法
- p VLSM: Variable-Length Subnet Masking, 等同于CIDR概念, 置1的位数等于网络号长度
- p 具有相同网络号的IP地址组成CIDR段, 表示为A.B.C.D/X, 其中X为网络号部分1的位数

例如202.38.208.0/20



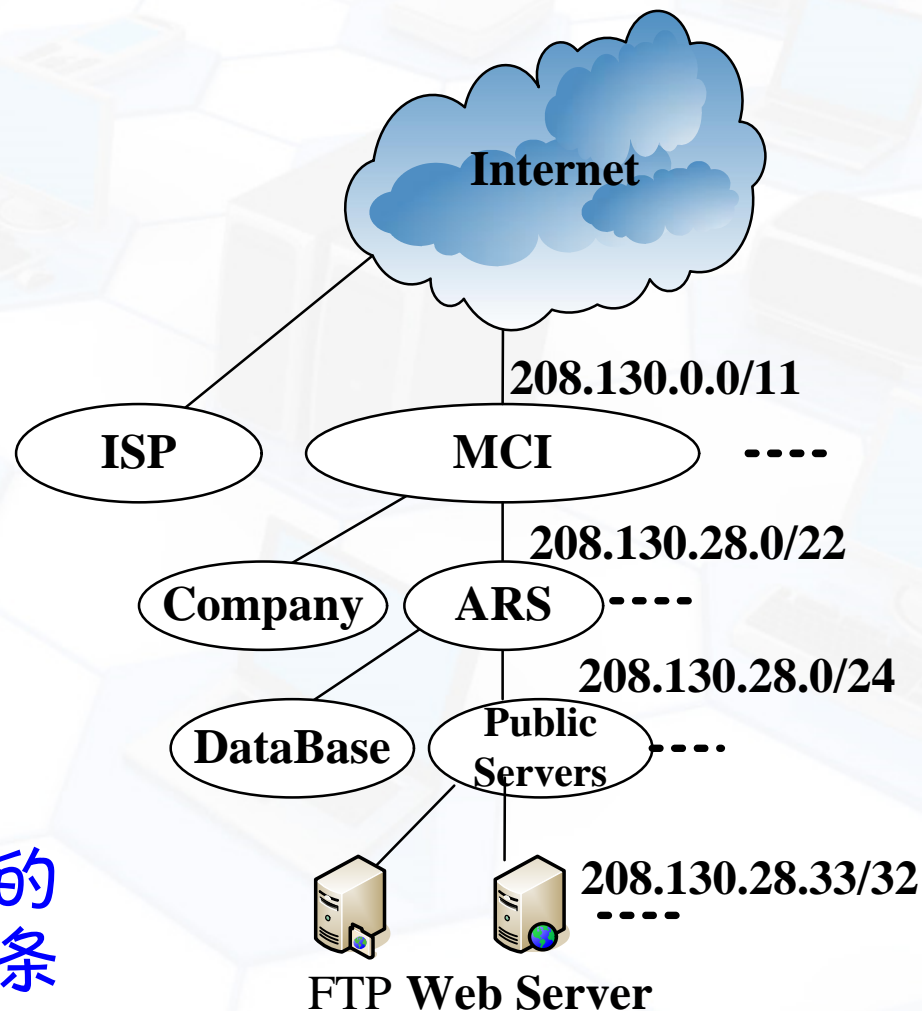
200.23.16.0/23

1) 地址汇聚

p 采用网络号汇聚

p 具有相同网络号的多个连续的CIDR段可以汇聚成一个更短路由表项

p 例如：16个连续的/24的CIDR段可以汇聚成一条/20的路由



208.130.16.0/24~208.130.31.0/24 → 208.130.16.0/20

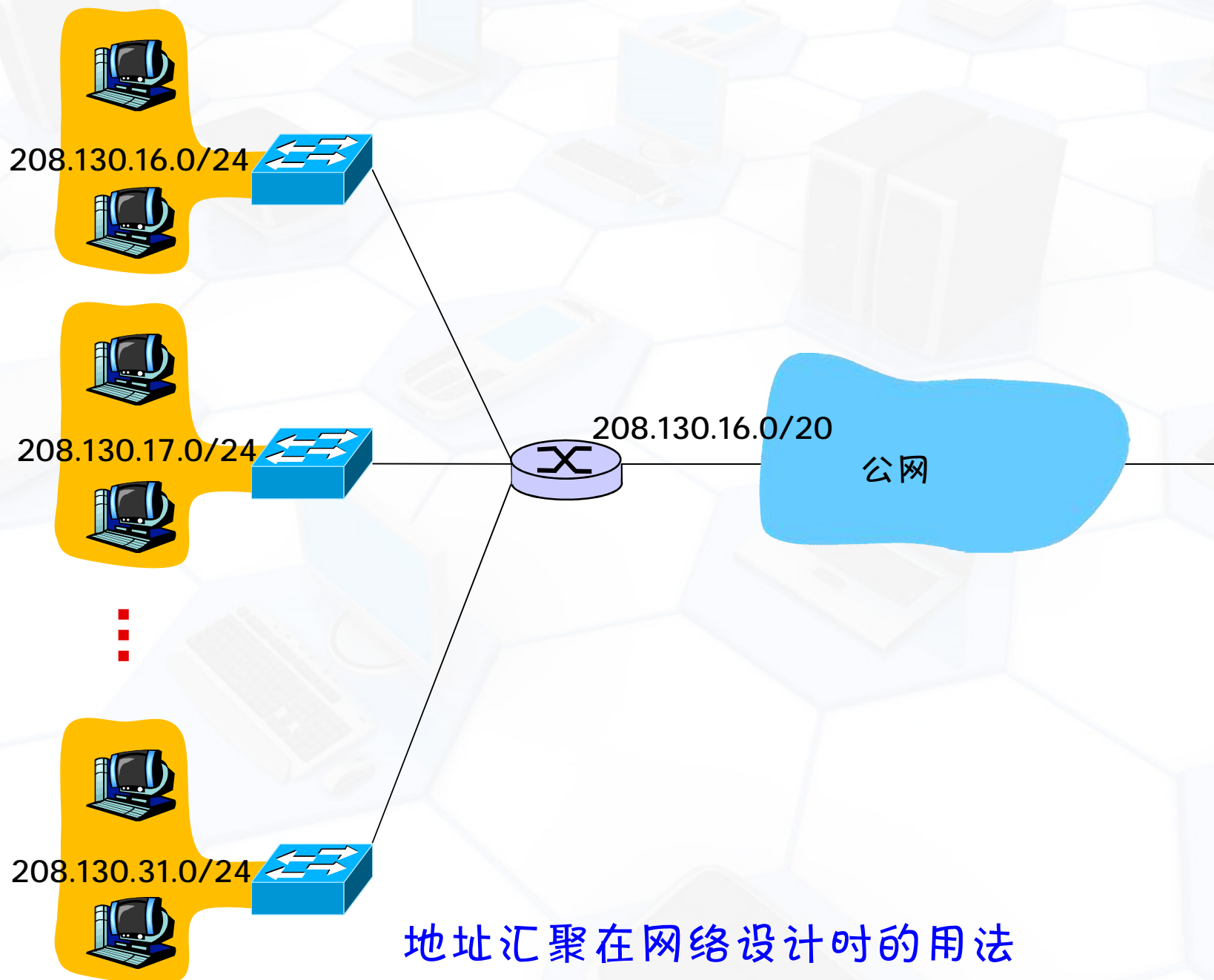
208	130	16	0
11010000	10000010	00010000	00000000
208	130	17	0
11010000	10000010	00010001	00000000
208	130	18	0
11010000	10000010	00010010	00000000
⋮	⋮	⋮	⋮
208	130	31	0
11010000	10000010	0001 1111	00000000

原网络号 (24位)



208	130	16	0
11010000	10000010	0001 0000	00000000

新网络号 (20位)



地址汇聚在网络设计时的用法



Organization 0

208.130.16.0/24

Organization 1

208.130.17.0/24

Organization 2

208.130.20.0/24

⋮

Organization 7

208.130.30.0/24

A-ISP

以208.130.16.0/20
开始的数据发往这里

Internet

B-ISP

以199.31.0.0/16开
始的数据发往这里

地址汇聚的示意

2) 匹配原则

p 路由表结构

<网络号/网络号长度、下一跳>

p 在CIDR中，如果路由器上的路由表中有多条表项满足要求，则采用最长前缀匹配规则

如：对于目的地址为128.96.195.70的分组，匹配的路由表表项包括两项，但是使用最长匹配规则选择的出口为R1

目的网络	出口
128.96.0.0/16	R0
128.96.192.0/18	R1
128.96.128.0/18	R2

11000011

11000000

10000000

单位	地址数	地址范围	掩码
计算机	2048	194.24.0.0 ~ 194.24.7.255	255.255.248.0(21位)
电子	4096	194.24.16.0 ~ 194.24.31.255	255.255.240.0(20位)
基础	1024	194.24.8.0 ~ 194.24.11.255	255.255.252.0(22位)

当目的地址为194.24.17.4的分组到达时的处理

	194.24.17.4	11000010	00011000	00010001	00000100
基础系掩码	255.255.252.0	11111111	11111111	11111100	00000000
	以上掩码对应网络号	11000010	00011000	00010000	00000000
基础系网络号194.24.8.0		194.24.16.0 不是基础系的起始网络号			
计算机系掩码	255.255.248.0	11111111	11111111	11111000	00000000
	以上掩码对应网络号	11000010	00011000	00010000	00000000
计算机系网络号194.24.0.0		194.24.16.0 不是计算机系的起始网络号			
电子系掩码	255.255.240.0	11111111	11111111	11110000	00000000
	以上掩码对应网络号	11000010	00011000	00010000	00000000
电子系网络号194.24.16.0		194.24.16.0 是电子系的起始网络号			

Organization 0

200.23.16.0/23

Organization 2

200.23.20.0/23

Organization 7

200.23.30.0/23

A-ISP

以200.23.16.0/20
开始的数据发往这里

Internet

B-ISP

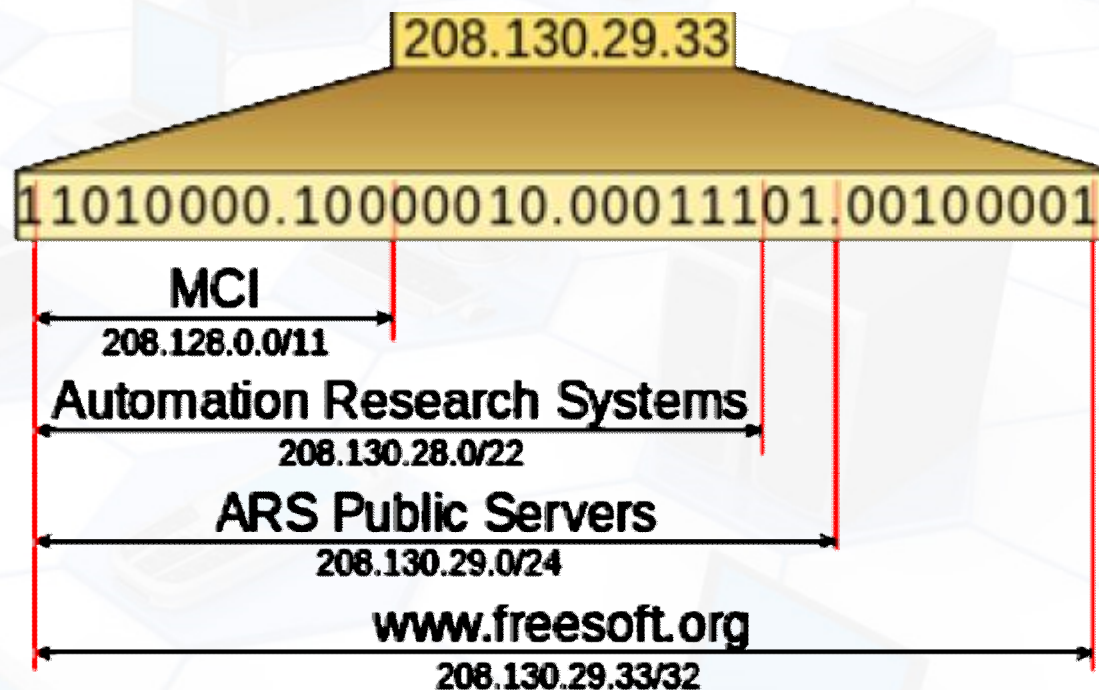
以199.31.0.0/16或
200.23.18.0/23开始
的数据发往这里

Organization 1

200.23.18.0/23

地址汇聚及最长前缀匹配的示意

一个例子：



在90年代末，IP地址208.130.29.33是被www.freesoft.org使用。后来，包含两百万地址的208.128.0.0/11块被ARIN（北美的RIR）分配给了MCI。MCI又将208.130.28.0/22 分配给了从MCI 租用互联网连接的Automation Research Systems。ARS则用了208.130.29.0/24这个地址块，其中就包含208.130.29.33这个地址。在MCI 的网络之外，208.128.0.0/11这个前缀会用于路由MCI 的数据流。这些数据流不仅会去到208.130.29.33，也会去到其他那些前11位相同的近两百万的地址里。在MCI 的网络里，208.130.28.0/22则会被路由到属于ARS租用的连接。最后，只有在ARS自己的网络内，208.130.29.0/24这个前缀才会被使用。

5. IPv6

1) IPv4的不足

- p 地址基本耗尽，这是当前最棘手的问题
- p 功能不足，缺少对多媒体信息传输的支持
- p 缺少对高速传输的支持
- p 缺少对安全的支持
- p 寻找路径的功能不强

2) IPv6的主要改进

- p 更大的地址空间：128位
- p 灵活的首部格式：用一系列固定格式的扩展首部取代了IPv4中可变长度的选项字段
- p 简化了协议：如取消了首部的校验和字段，分段只能在源端进行，节省了处理时间
- p 允许对网络资源的预分配，支持实时图像等要求保证一定的带宽和时延的应用
- p 允许协议继续演变，增加新的功能

IPv4到IPv6的数据结构变化

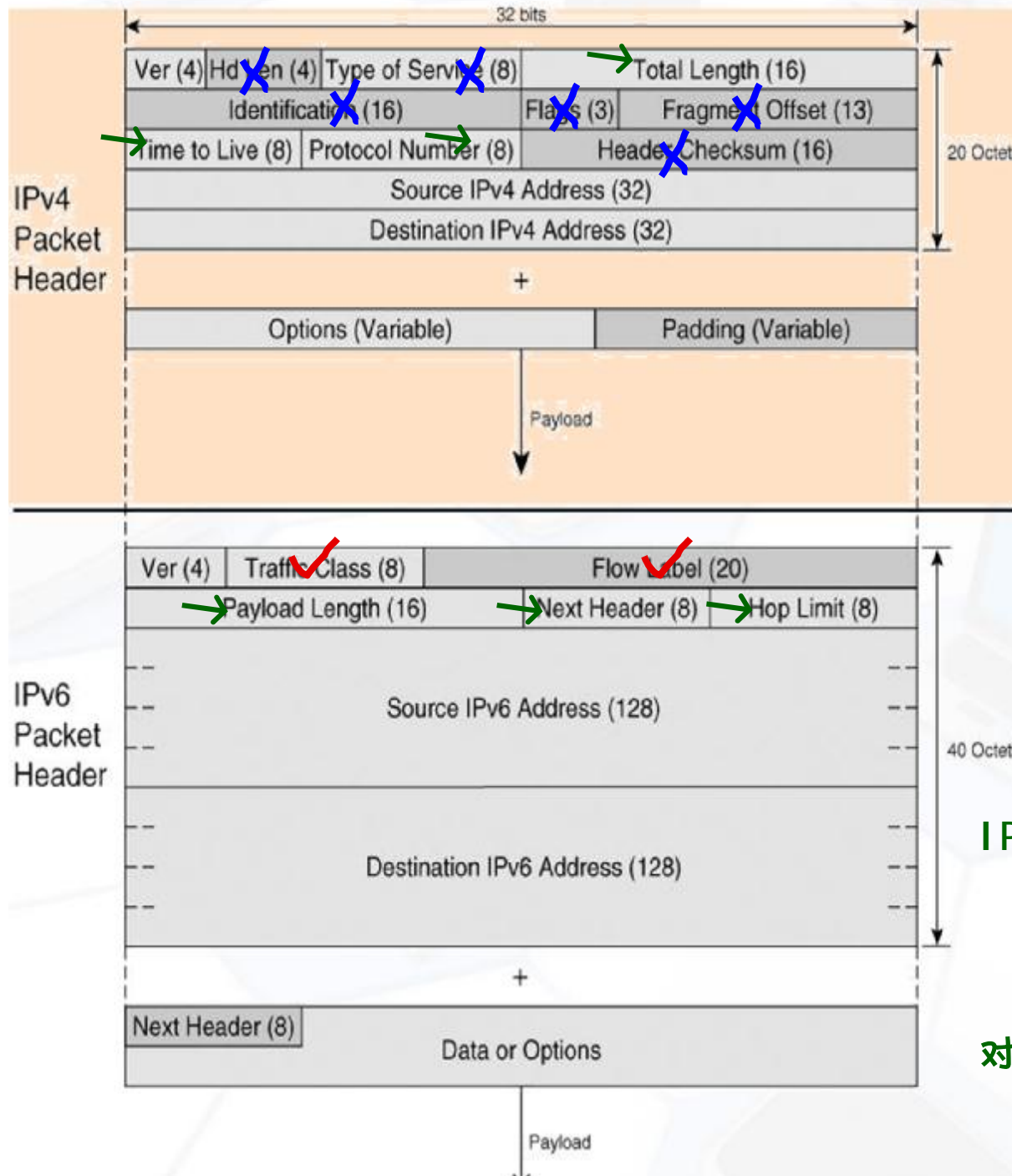
IPv6: 6 字段 + 2 地址
 IPv4: 10 字段 + 2 地址 + 选项

IPv6取消了以下内容:
 Header length
 type of service
 identification, flags,
 fragment offset
 Header Checksum

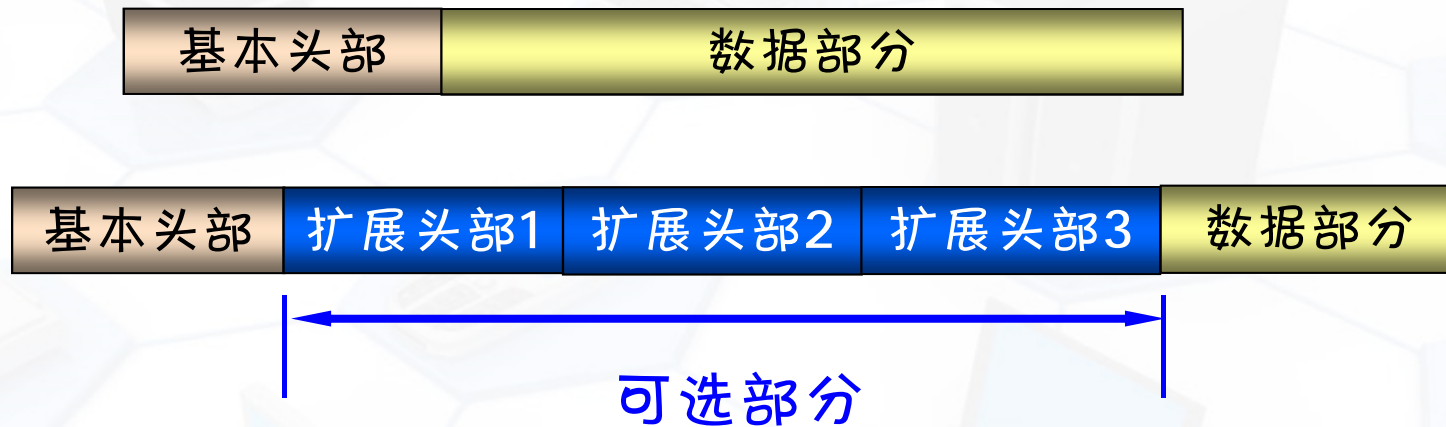
IPv6增加了:
 Traffic class
 Flow label

IPv6重命名了以下内容:
 length -> Payload length
 Protocol -> Next header
 time to live -> Hop Limit

对Option机制进行了再定义

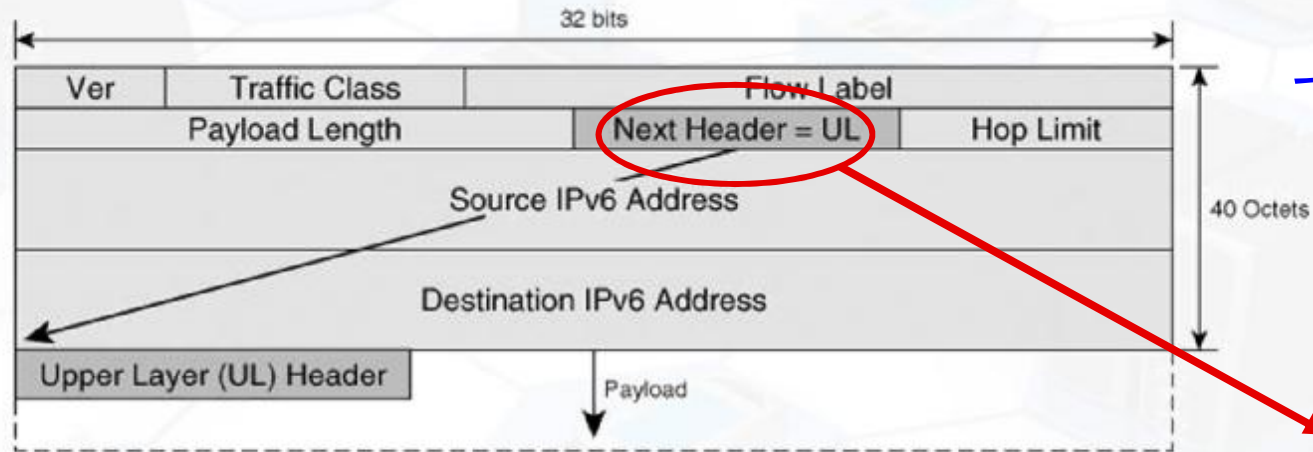


3) IPv6数据的基本与扩展头部



一个IPv6报文可以带有零个、一个或多个扩展头部，由前一个头部中的Next Header字段进行说明；如有扩展头部，则说明扩展头部的类型，如没有其它扩展头部，则说明数据报中携带的数据类型

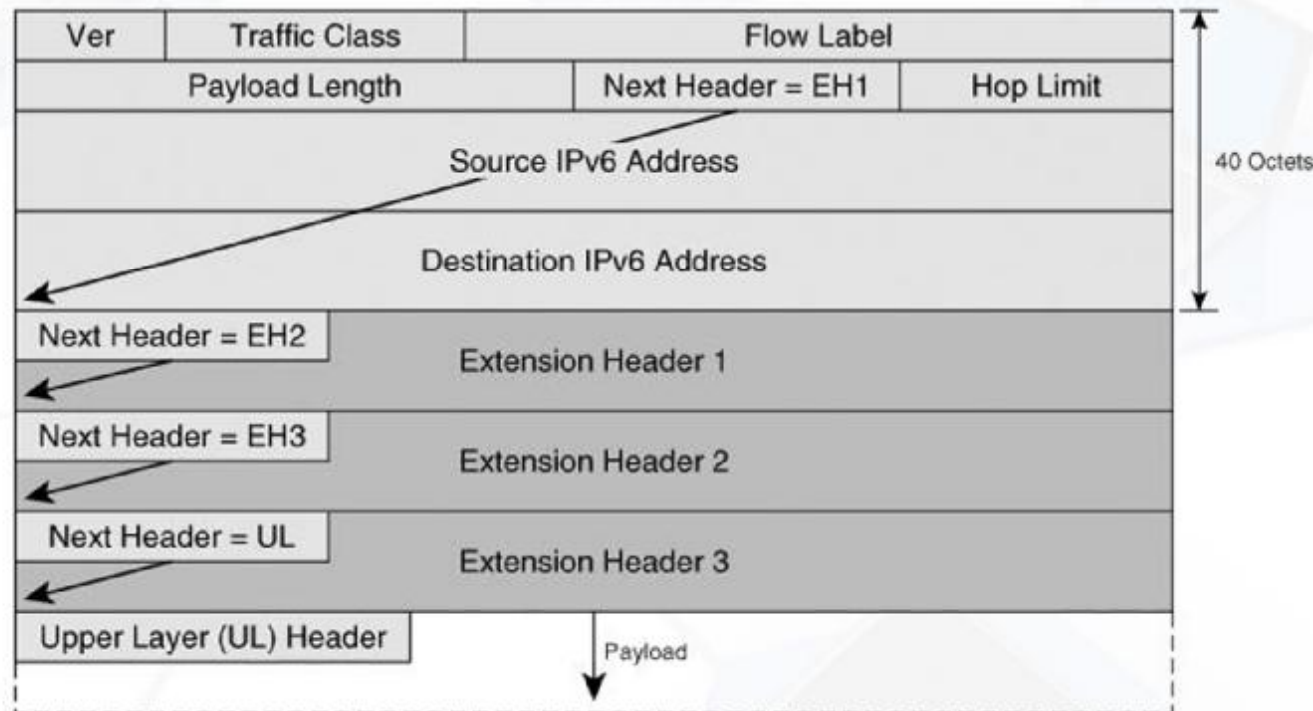
Packet without Extension Header



下一头部取值定义

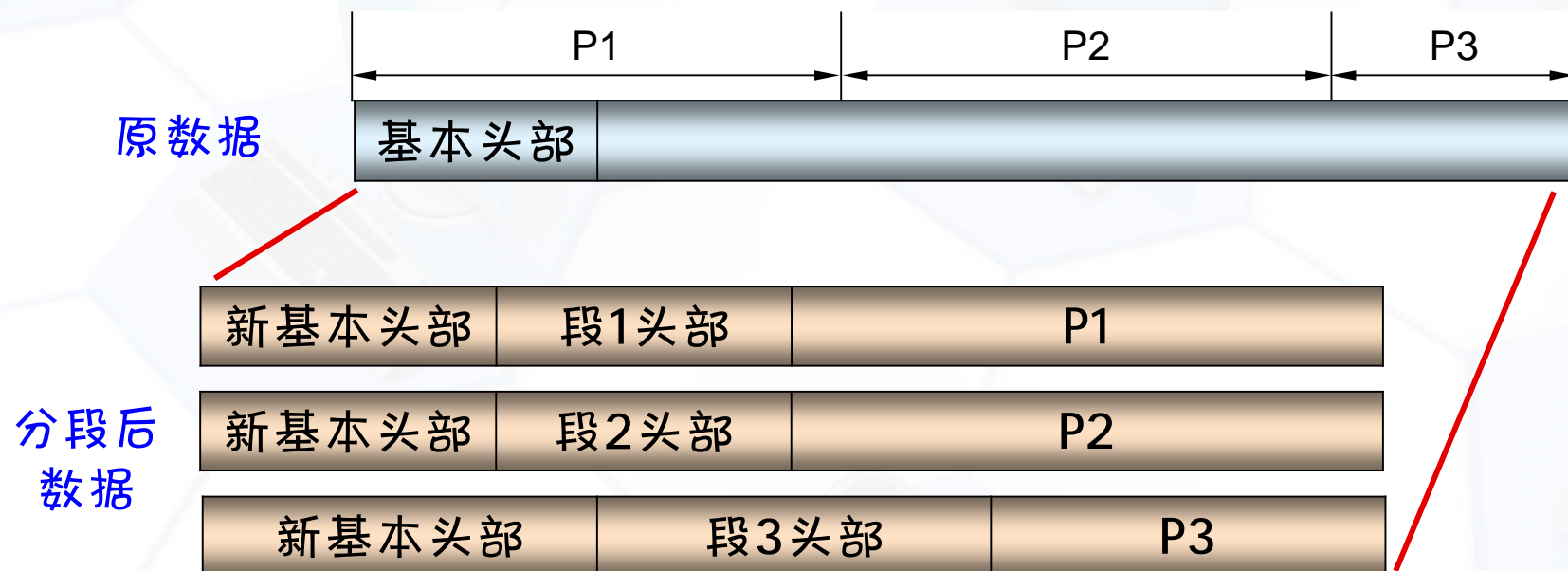
取值	含义
0	中继点选项头标
4	IP
6	TCP
17	UDP
43	寻路头标
44	分片头标
45	IDRP
46	RSVP
50	封装化安全净荷
51	认证头标
58	ICMPv6
59	无下一个头标
60	信宿选项头标

Packet with Extension Header



4) IPv6的分段与重组

- p** 源站点负责分段，中途的路由不再允许分段，节省了时间，减少了差错风险
- p** 路由器发现一个大于MTU的数据包，则丢弃，并发一ICMP消息给源站，由源站重新划分数据包并重发
- p** 每个分段将增加一个分段头部，并紧随基本头部之后
- p** 每个分段的基本头部即原基本头部，但其中的负载长度需作修改



5) IPv6地址

p 128位地址可产生 2^{128} 个地址

p 理论上说，地球上每平方米有
665,570,793,348,866,943,898,599个IPv6地址

p 表示方法采用冒分十六进制法

X:X:X:X:X:X:X:X 其中X表示地址中16位二进制数的
十六进制值

例：FEDC:BA98:7654:3210:FEDC:BA98:7654:3210

p 若其中有多连续的零，则可用零压缩法

如：1080:0:0:0:8:800:200C:417A 可写成

1080::8:800:200C:417A

p IPv4与IPv6共存的表示

X:X:X:X:X:X:d.d.d.d 最后32位地址为IPv4的点分十进制

0:0:0:0:0:0:202.120.5.100(称为IPv4兼容IPv6地址)

0:0:0:0:0:FFFF:202.120.5.100(称为IPv4映象IPv6地址)

p IPv6地址空间分配:

ü 单播地址(Unicast)

标识单个接口,单播地址的分组被发送到该接口

ü 组播地址(Multicast)

标识通常属于不同节点的一组接口,组播地址的分组被发送到所有的接口

ü 任播地址(Anycast)

标识通常属于节点上的一组接口,任播地址的分组被发送到最近的那个接口(依据路由协议度量的最近距离)

IPv6中没有广播地址,其功能被组播地址所代替

6) 由IPv4过渡到IPv6

- 无法做到所有的路由同时升级到IPv6
- 需要解决 IPv4与IPv6共存的问题
- 隧道技术: 将IPv6分组作为IPv4的数据, 然后经仅支持IPv4的路由传递

逻辑功能示意



物理网络示意

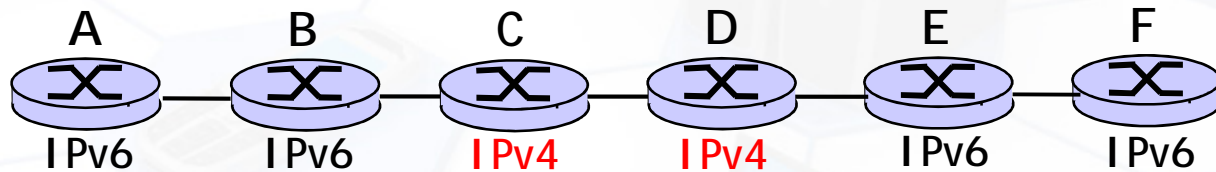


隧道功能详解

逻辑隧道



实际通讯



Flow: X
Src: A
Dest: F
data

A-to-B:
IPv6

Src:B
Dest: E

Flow: X
Src: A
Dest: F
data

B-to-C:
IPv6 inside
IPv4

Src:B
Dest: E

Flow: X
Src: A
Dest: F
data

B-to-C:
IPv6 inside
IPv4

Flow: X
Src: A
Dest: F
data

E-to-F:
IPv6

三、 IP控制协议

p IP协议只负责传送IP数据包，无法监视和控制网络中出现的一些问题，这些工作由Internet的控制协议来完成

1. ICMP协议（Internet Control Message Protocol）

p IP在网络层提供尽力服务（best effort service），当分组由于各种原因无法投递而遭丢弃时，就用ICMP发送差错报告，尽管ICMP也是网络层协议，但它也需要经过IP协议封装；同样ICMP也不能保证可靠传输

1) ICMP协议特点

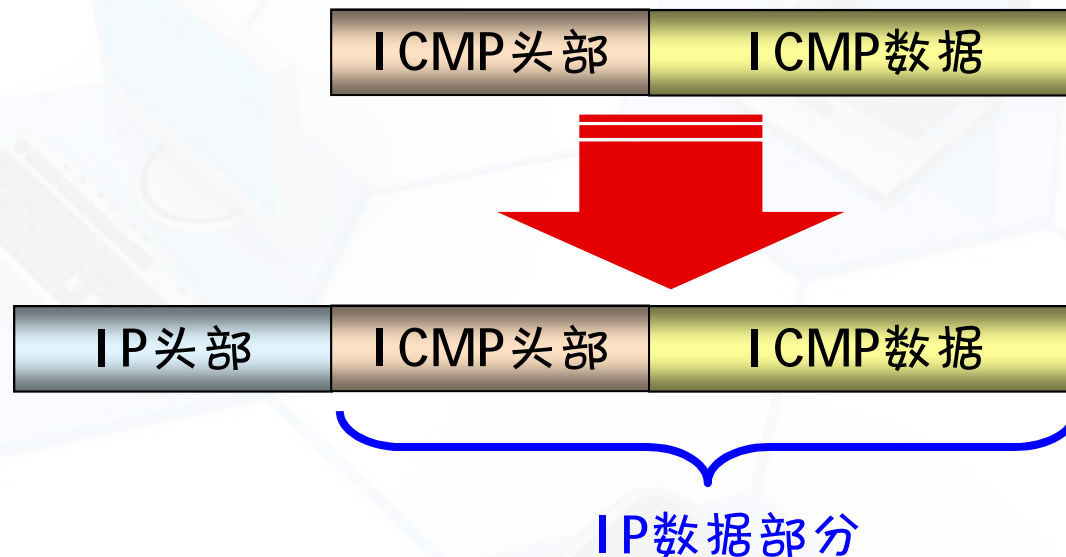
p 主机及路由间用于进行网络层级别的信息交流

 ü 差错报告:无法达到的主机、网络、端口、协议

 ü 回应请求及应答(ping命令)

p 在网络层中位置高于IP协议:

 ü ICMP 消息需要IP数据报来封装



2) ICMP报文格式

p ICMP消息构成：类型，代码 + 出错IP包的前8个字节



3) ICMP报文主要类型

类型	类型值	ICMP报文类型
差错报文	3	目的站点不可达
	11	数据报超时
	12	数据报参数错
控制报文	4	源抑制
	5	重定向
请求/应答 报文	8	回应请求
	0	回应应答
	13	时间戳请求
	14	时间戳应答
	17	地址掩码请求
	18	地址掩码应答

a) 目的站点不可达(Destination Unreachable)

p 目的站点已关机

p 目的地址不存在

p 不知道目的站点的路径

0	8	16	31
类型（3）	代码（0~12）	校验和	
未用（全0）			
出错数据报报头 + 前64 bit数据			
... ..			

目的站点不可达报文的格式



p 目的站点不可达报文类型

代码值	意义
0	网络不可达
1	主机不可达
2	端口不可达
3	协议不可达
4	需分段，但DF=1（不允许分段）
5	源寻径失败
6	目的站点网络未知
7	目的主机网络未知
8	原主机被隔离
9	与目的站点网络的通信被禁止
10	与目的站点主机的通信被禁止
11	对请求的服务类型，网络不可达
12	对请求的服务类型，主机不可达

b) 数据报超时 (Time Exceeded)

p 代码为0: TTL超时

p 代码为1: 分段重组超时

0	8	16	31
类型（11）	代码（0/1）	校验和	
未用（全0）			
出错数据报报头 + 前64 bit数据			
... ..			

数据报超时报文的格式

c) 数据报参数错

p代码为0: 指出数据报中现有的某参数出错, 其指针域指向数据报中引起故障的字节

p代码为1: 指出数据报中缺少某一必须的选项 (代码为1时无指针域)

0	8	16	31
类型（12）	代码（0/1）	校验和	
未用（全0）			
出错数据报报头 + 前64 bit数据			
... ..			

数据报参数错报文的格式

d) 重定向(Redirect)

p 重定向机制将保证主机拥有一个动态的既小且优的路由表，但对路由器之间的路由表优化无能为力

0	8	16	31
类型（5）	代码（0 ~ 3）	校验和	
网关地址			
出错数据报报头 + 前64 bit数据			
... ..			

重定向报文的格式

代码值	意 义
0	对网络的重定向报文 (已不用)
1	对主机的重定向报文
2	对服务类型和网络的重定向报文
3	对服务类型和主机的重定向报文

其中的“网关IP地址”是告诉源站点“去目的站点最优路径中第一个网关的地址”

e) 回应请求/应答报文的格式

p 回应请求/应答用于测试目的站点的可达性

0	8	16	31
类型（8或0）	代码（0）	校验和	
标识符		序号	
任选数据			
... ..			

p 通常：某源站点主机向某目的站点主机发出一个回应请求，包含一段任选数据，如与收到的应答报文中的任选数据相同，则证明目的站点可达

4) ICMP协议的运用

p 测试报文的可达性
ping命令

p 路由跟踪命令
tracert (Unix下为traceroute) 命令

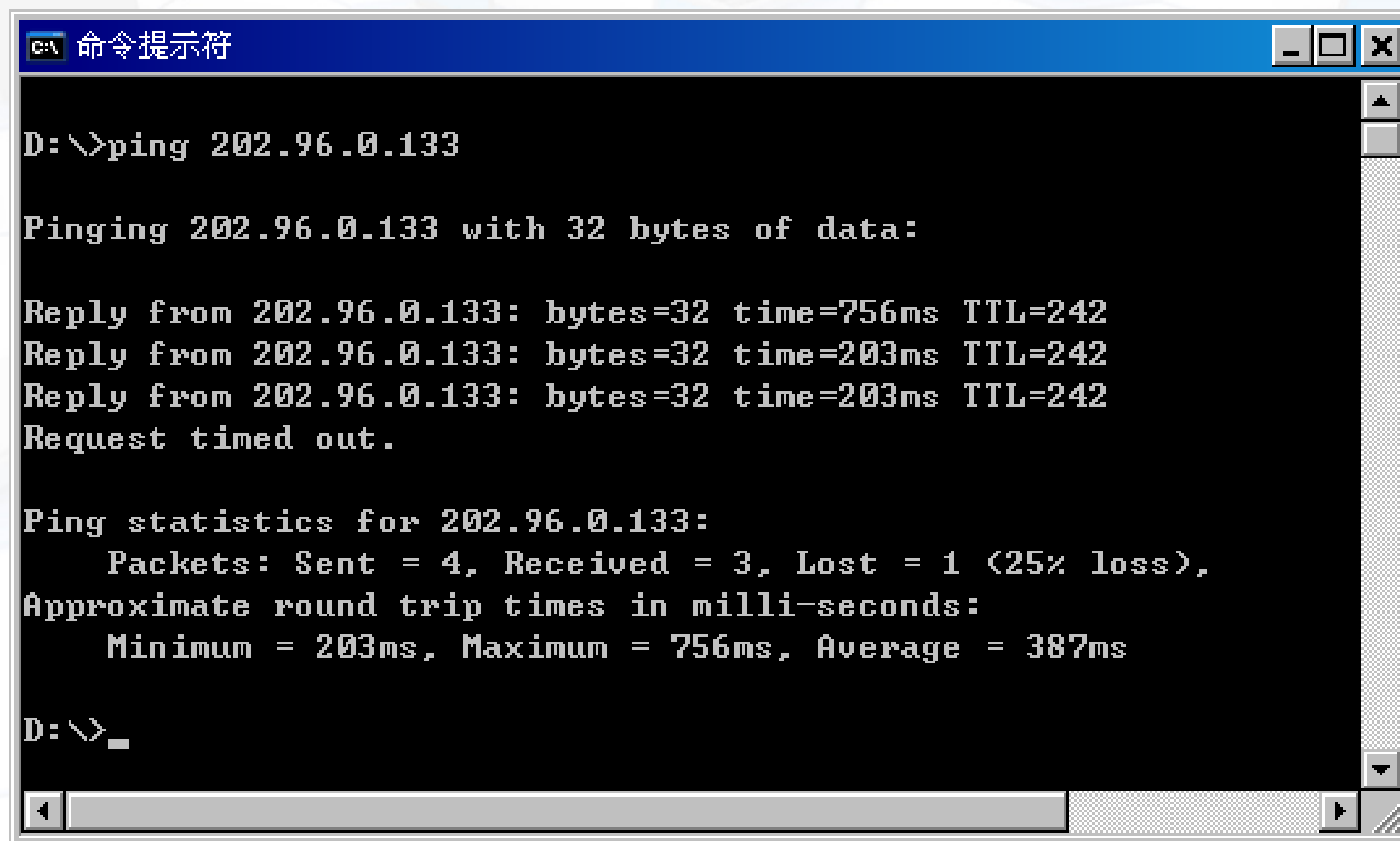
p 得到路径中最小的MTU

a) Ping程序

p 功能：测试主机的可达性和往返延迟等网络信息

p ping程序采用ICMP的回应请求/应答(Echo request/reply)报文，通过向目的主机发送回应请求，对方返回响应报文，来测试目的主机的可达性、往返延迟以及丢包率

p 使用ping命令时，将向目的站点发送一个ICMP回应请求报文（包括一些任选的数据），若目的站点接收到该报文，必须向源站点发回一个ICMP回应应答报文，源站点收到应答报文（且其中的任选数据与所发送的相同），则认为目的站点是可达的，否则为不可达



```
C:\>命令提示符

D:\>ping 202.96.0.133

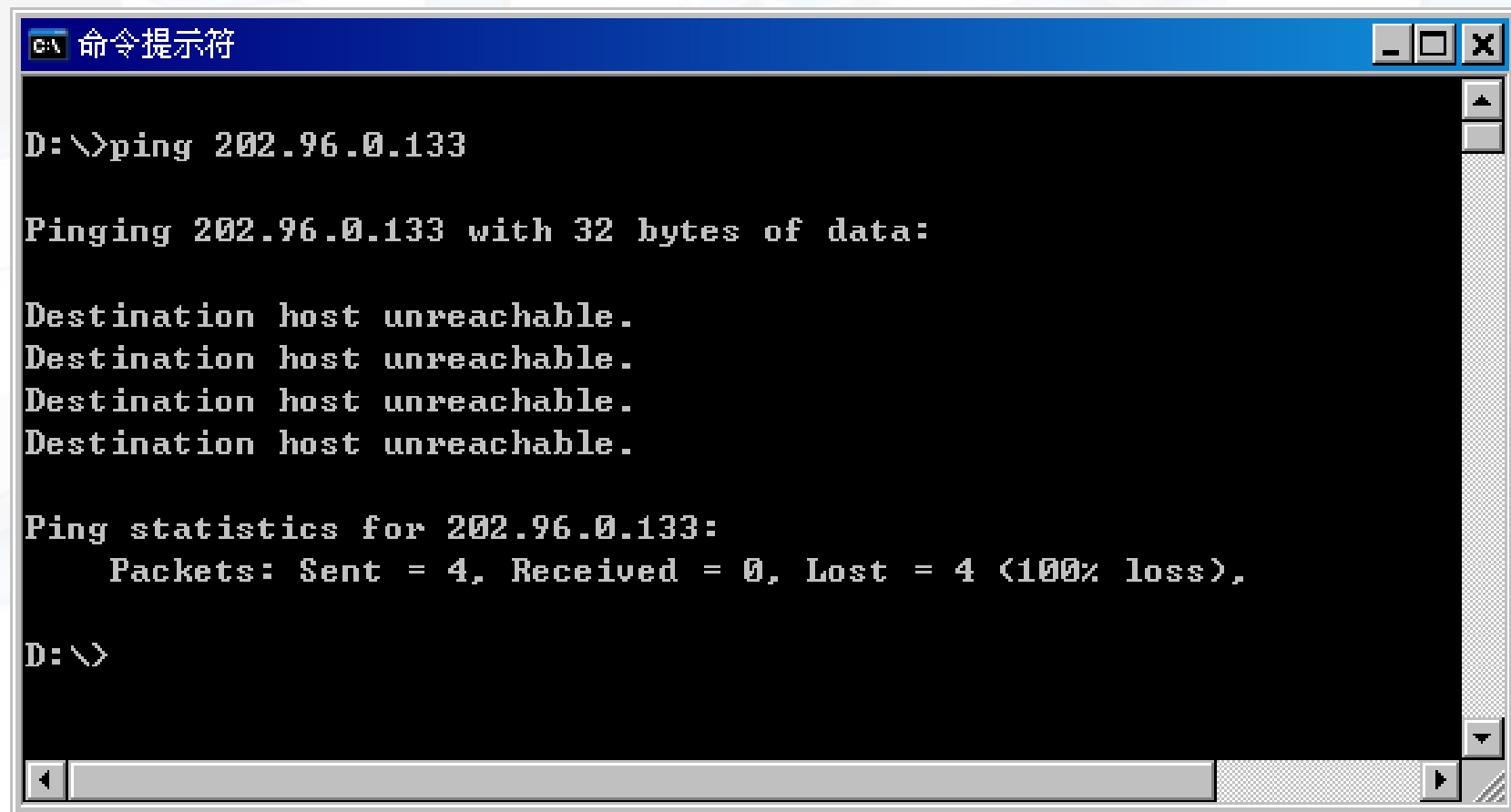
Pinging 202.96.0.133 with 32 bytes of data:

Reply from 202.96.0.133: bytes=32 time=756ms TTL=242
Reply from 202.96.0.133: bytes=32 time=203ms TTL=242
Reply from 202.96.0.133: bytes=32 time=203ms TTL=242
Request timed out.

Ping statistics for 202.96.0.133:
    Packets: Sent = 4, Received = 3, Lost = 1 (25% loss),
    Approximate round trip times in milli-seconds:
        Minimum = 203ms, Maximum = 756ms, Average = 387ms

D:\>_
```

目的站点可达的情况



```
C:\>命令提示符

D:\>ping 202.96.0.133

Pinging 202.96.0.133 with 32 bytes of data:

Destination host unreachable.
Destination host unreachable.
Destination host unreachable.
Destination host unreachable.

Ping statistics for 202.96.0.133:
    Packets: Sent = 4, Received = 0, Lost = 4 (100% loss),

D:\>
```

目的站点不可达的情况

b) Tracert程序

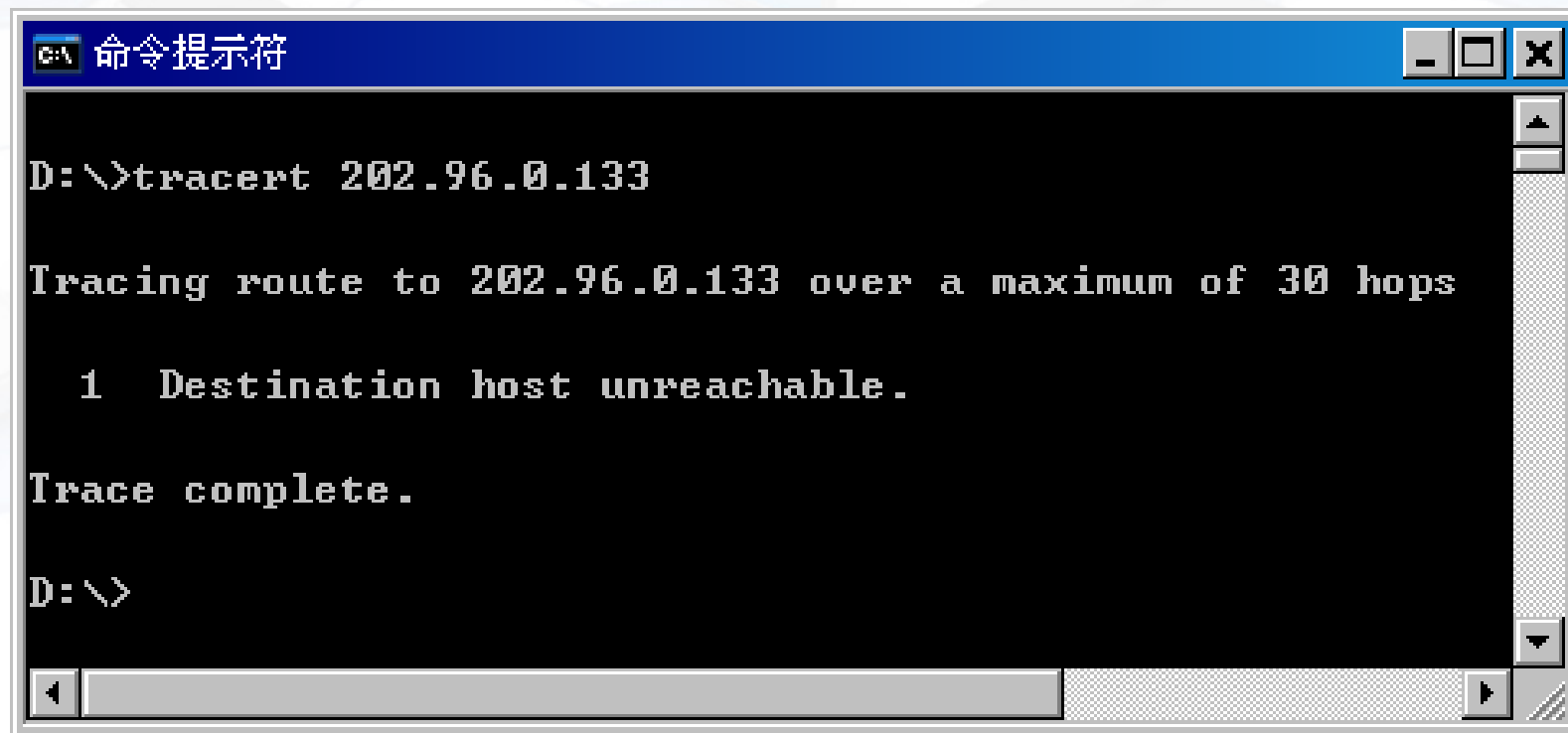
p 功能：发现从源主机到目的主机路径上的网络节点

p tracert过程是通过ICMP数据报超时报文来得到一张途经路由器列表的

p 源主机向目的主机发一个IP报文，并置hop为1，到达第一个路由器时，hop减1，为0，则该路由器返回一个ICMP数据报超时报文，源主机取出路由器的IP地址即为途经的第一个路由端口地址

p 接着源主机再向目的主机发第二个IP报文，并置hop为2，然后再发第三个、第四个IP数据报，...直至到达目的主机

但互联网的运行环境状态是动态的，每次路径的选择有可能不一致，所以，只有在相对较稳定网络中，tracert才是有意义的



```
命令提示符

D:\>tracert 202.96.0.133

Tracing route to 202.96.0.133 over a maximum of 30 hops

  1  Destination host unreachable.

Trace complete.

D:\>
```

tracert目的站点不可达的情况

```
c:\ 命令提示符

D:\>tracert 202.96.0.133

Tracing route to ns.bta.net.cn [202.96.0.133]
over a maximum of 30 hops:

  1    25 ms    25 ms    25 ms    61.152.64.119
  2     *       *       *       Request timed out.
  3    28 ms    27 ms    26 ms    ppp91-93-109-202.online.sh.cn [202.109.93.91]
  4    26 ms    26 ms    27 ms    branch-2-h130.sta.net.cn [61.152.46.130]
  5    27 ms    28 ms    27 ms    202.109.8.93
  6    27 ms    27 ms    26 ms    iso0-0-jnpr-px.online.sh.cn [202.109.0.157]
  7    27 ms    27 ms    27 ms    202.109.0.142
  8    27 ms    27 ms    28 ms    202.101.63.130
  9    27 ms    27 ms    27 ms    202.101.63.214
 10    28 ms    27 ms    27 ms    202.97.36.38
 11   211 ms   213 ms   211 ms   219.158.28.117
 12   209 ms   208 ms   208 ms   202.96.12.42
 13     *      210 ms   213 ms   202.106.192.62
 14   211 ms   213 ms   211 ms   ns.bta.net.cn [202.96.0.133]

Trace complete.

D:\>_
```

tracert目的站点可达的情况

c) 得到路径中最小的MTU

p源主机发送一系列的探测IP数据报，并置DF=1，即不允许分段，如途径某个网络的MTU较小，则路由器将丢弃该数据报并发回一个ICMP数据报参数错，要求分段。源主机则逐步减小数据报长度，并仍置DF = 1，直至某个探测报文成功到达目的主机，即得到路径中的最小MTU

2. 地址解析协议

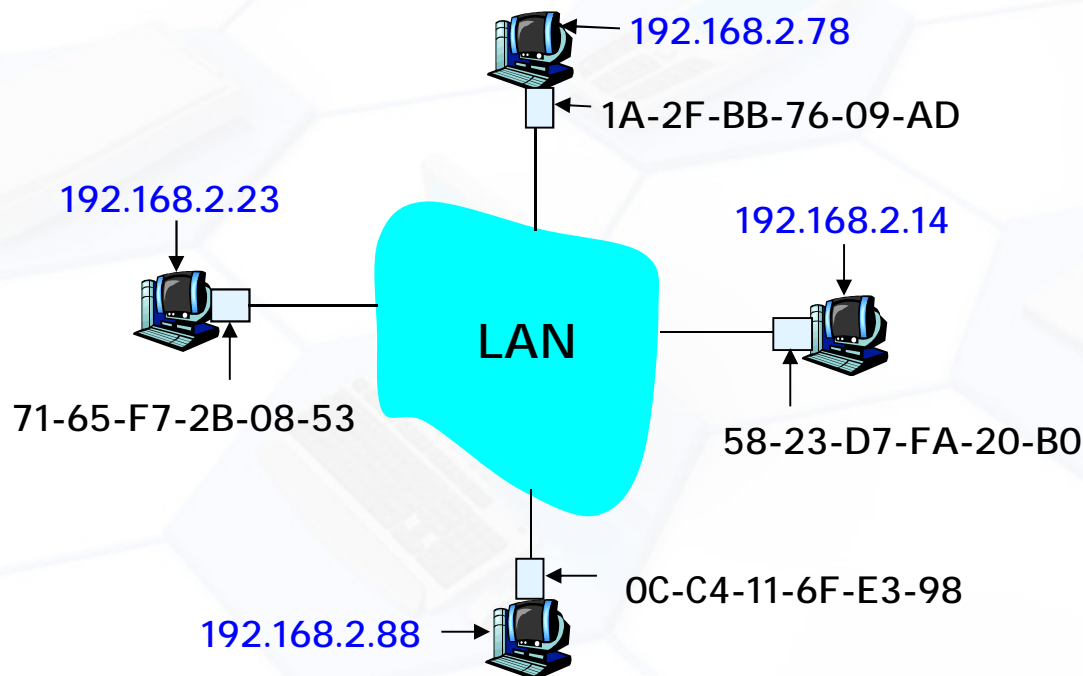
1) 地址解析的作用

p 协议地址：软件提供的抽象地址，如IP地址，它使整个互联网看成一个网络，但真正的物理网络并不能通过IP地址来定位机器

p 协议地址和物理地址之间的转换，如IP地址和MAC地址之间的转换

问题1: 若已知MAC地址, 如何知道对应的IP地址?

问题2: 问题1反之又如何处理?



IP分组封装: 分组中需要填写对方的IP地址

帧的封装: 帧结构中需要填写对方的MAC地址

? 如何知道对方IP与对方MAC的对应关系

2) 地址解析技术

p 查表

- ü 一个物理网络，即IP的一个子网，对应一张地址解析表

IP地址	MAC地址
202.120.1.102	0A:07:4B:12:82:36

p 相近形式计算

- ü 适用于动态硬件地址，它在指定IP地址和硬件地址时，使它们保持一定的关系，在解析时可通过IP地址计算出硬件地址
- ü 如：硬件地址与IP地址的最后一个字节相同，
则：硬件地址 = IP地址 & 0xff

p 消息交换法

- ü 服务器方式：由服务器提供解析结果(ATM网络)
- ü 分布式方法：每台计算机负责对本机地址的解析
- ü 相比较而言，前者对地址的配置和管理较容易，但需要额外的服务器，一旦网络繁忙程度加大，服务器会成为瓶颈

3) 地址解析协议ARP

p 每个IP节点 (主机/路由) 都有一个ARP表

p ARP表保存IP/MAC地址的映射关系

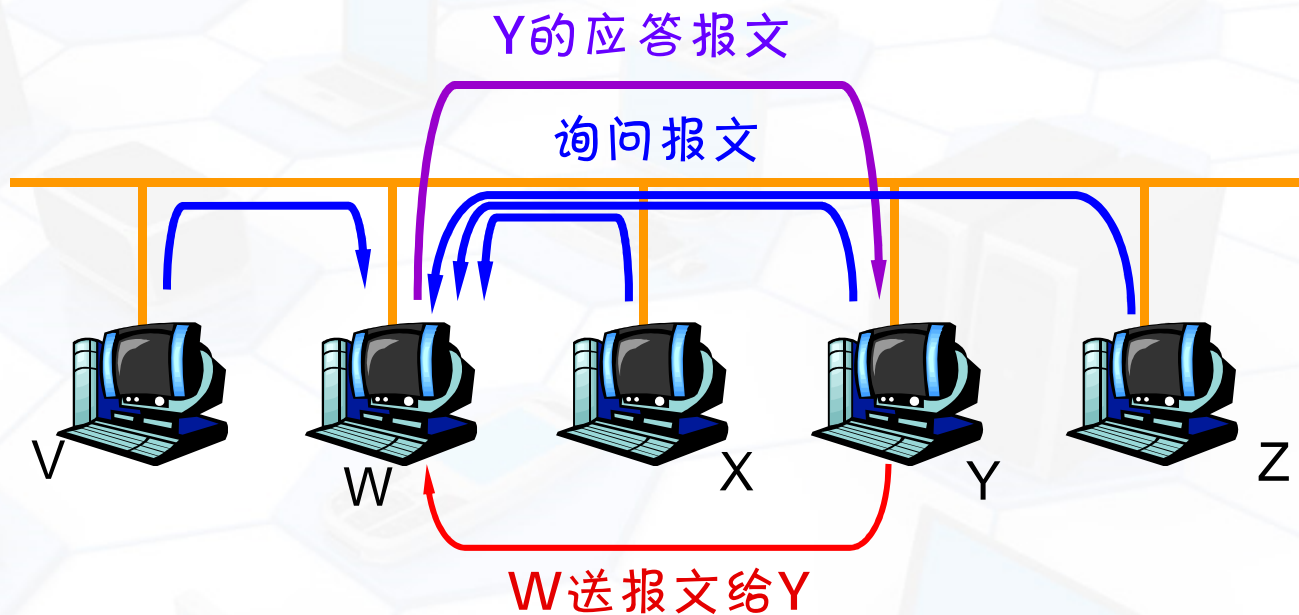
< IP, MAC, TTL >

p TTL (Time To Live): 该映射项失效的时间 (典型为 20 分钟)

p 主机根据分组头上的目的IP地址查阅自己的ARP缓存, 如果没查到, 就用广播地址发送ARP请求

p 被请求的IP地址所对应的主机返回一个ARP响应

p 主机收到响应后, 就可发送数据帧, 并将该IP地址与MAC地址的映射存放在ARP缓存中



p 一个ARP请求消息是一个数据帧，其中包含发送站的MAC地址和IP地址，以及目的地址的IP地址，并把此数据帧在本物理网络内广播

p 一个ARP应答消息是一个数据帧，其中包含应答站的MAC地址和IP地址，以及原发送站点的MAC和IP地址，并把此数据帧发送给原发送站

4) ARP的报文格式

0	8	16	31
硬件地址类型		协议地址类型	
硬件地址长度	协议地址长度	操 作	
发送站硬件地址（字节0~3）			
发送站硬件地址（字节4~5）		发送站协议地址（字节0~1）	
发送站协议地址（字节2~3）		目的站硬件地址（字节0~1）	
目的站硬件地址（字节2~5）			
目的站协议地址全部（字节0~3）			

ARP报文举例

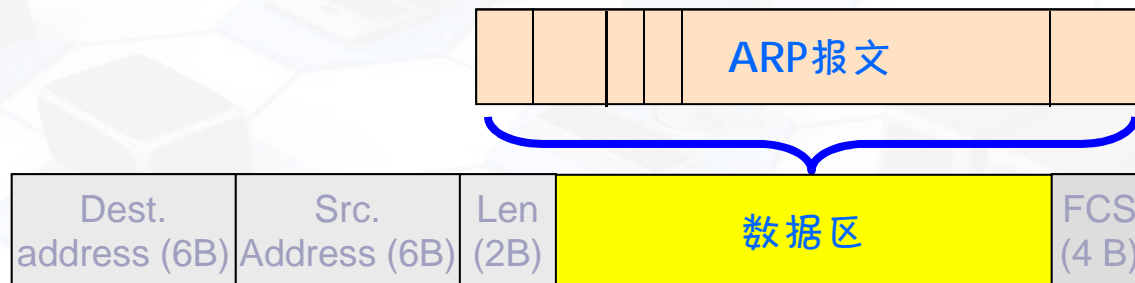
发送站ARP请求报文中内容

硬件地址类型	1
协议地址类型	0X0800
硬件地址长度	6
协议地址长度	4
操作	1 (请求)
发送站硬件地址	MAC地址
发送站协议地址	IP地址
目的站硬件地址	全0
目的站协议地址	IP地址

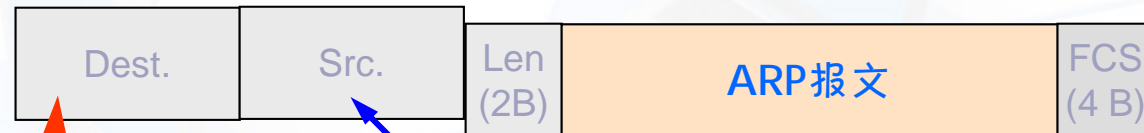
目的站ARP应答报文中内容

硬件地址类型	1
协议地址类型	0X0800
硬件地址长度	6
协议地址长度	4
操作	2 (应答)
发送站硬件地址	MAC地址
发送站协议地址	IP地址
目的站硬件地址	MAC地址
目的站协议地址	IP地址

5) ARP消息需要在以太网帧中封装



以太网帧的完成



发送方填写自己的MAC地址

问题：发送方在发出ARP询问报文时，ARP报文被封装到以太网帧中，帧的目的地址填什么（发送方还不知道该MAC地址）？——但没有具体物理地址不能发送数据帧

6) ARP命令

```
C:\WINDOWS\system32\cmd.exe

C:\>arp -a

Interface: 192.168.1.101 --- 0x60007
    Internet Address      Physical Address      Type
    192.168.1.1           00-21-27-1b-78-0a     dynamic
    192.168.1.102         00-19-d2-41-25-a1     dynamic

C:\>
```

p 暂存ARP应答于Cache或内存中，以后即可查表，不必再发询问报文，以减少网络的通信量

p 从消息中取出发送方的协议地址和硬件地址，更新cache中已有的信息

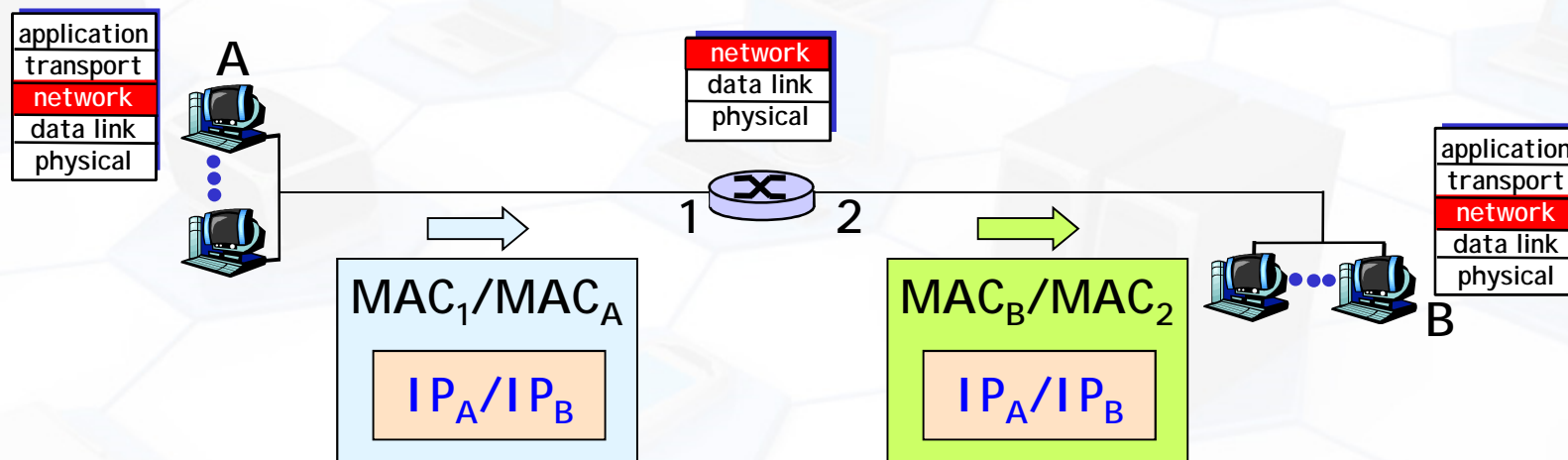
7) 不同网络间的ARP

p 使用缺省路由：主机通过识别目的IP地址的网络号，知道它是子网外的主机，直接发给缺省路由器

p 代理ARP：路由器有ARP代理功能，它代理网络外的主机响应ARP请求，可以实现路由器拦截目的IP地址为其代理的主机的分组

这种情况下，目的IP地址对应的MAC地址均为路由器的MAC地址

ARP只工作在同一子网中

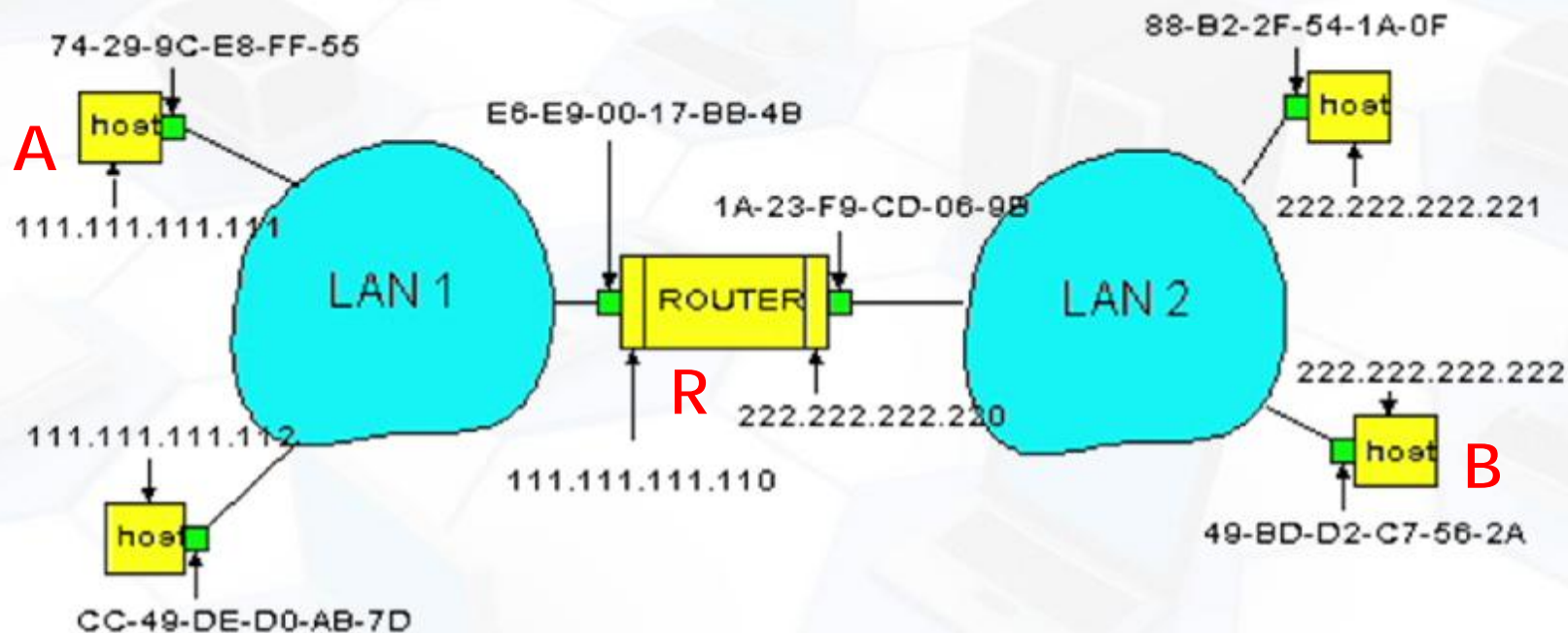


p 数据帧在链路上传输使用MAC地址，源MAC地址为发送帧的接口的MAC地址，而目的MAC地址为同一链路上接收帧的接口的MAC地址

p IP分组在Internet中传输使用IP地址，源IP地址为发送IP分组的主机的IP地址，而目的IP地址为接收IP分组的主机的IP地址

注意：数据帧在不同链路上传输时，源和目的MAC地址变化，而源和目的IP地址不变

举例：假设A已知B的IP地址，要发送数据给B



- p 路由R有两个ARP表, 分别对应不同的IP网络 (LAN1/LAN2)
- p 主机A中有去向外网的路由地址信息 111.111.111.110
- p 主机A的ARP表中可能保存有MAC地址E6-E9-00-17-BB-4B与IP地址111.111.111.110的映射, 也可能没有

pA 产生一个分组，源IP是A, 目的地IP是B

pA利用ARP 来得到IP地址是111.111.111.110的路由端口所对应的MAC地址

pA将刚才获得的路由R的MAC地址填入其帧的目的地MAC位置, 该帧中包含着A-to-B的IP分组

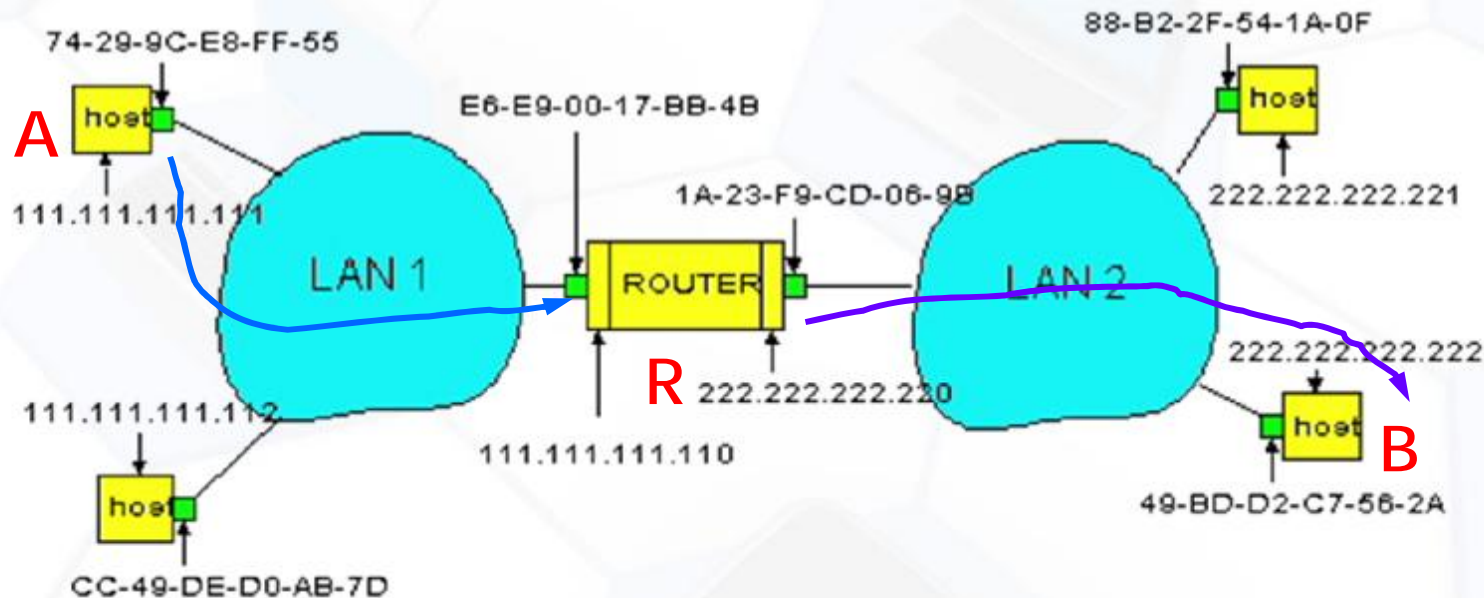
pA的网卡发送该帧

p路由R的局域网接口接收该帧

p路由R从帧中取出 IP分组, 检查出它的目的地是B

p路由R利用 ARP来取得B的MAC地址 (B的IP地址已经知道)

p路由R用刚才得到的B的MAC地址形成一个包含A-to-B IP分组的帧发往 B



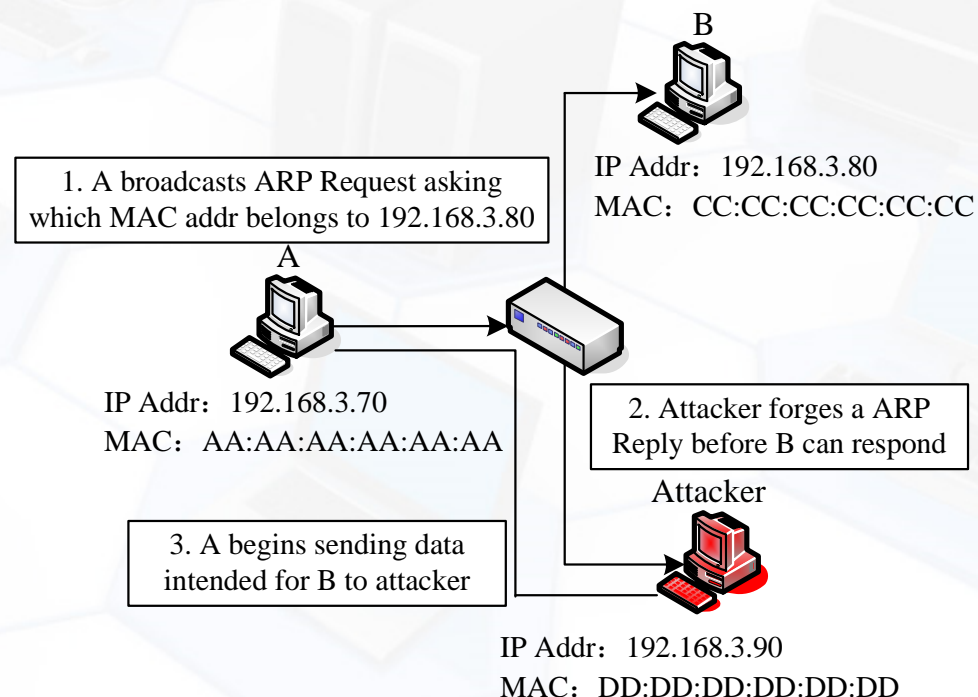
8) ARP欺骗

防范措施

p 监测可疑的ARP流量，特别是网关和路由器等关键设备的MAC地址的变化

p 通过划分子网、VLAN等措施限制ARP的广播域

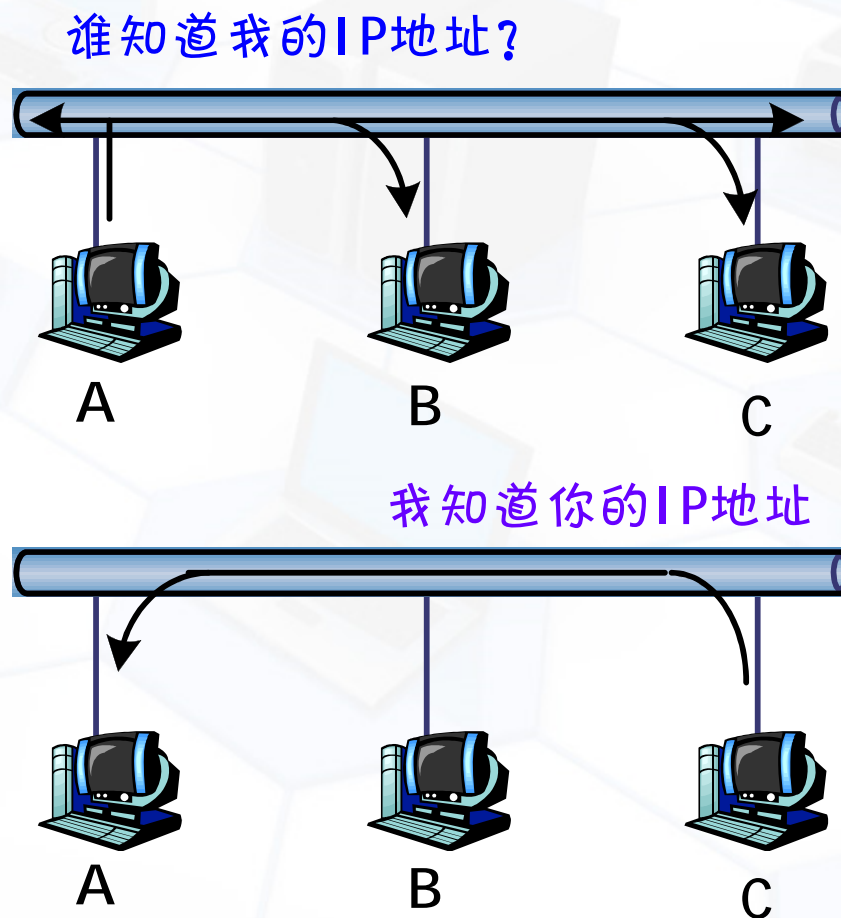
p 在ARP Cache中静态配置IP和MAC地址的映射表项



9) 反向ARP

p Reverse ARP

p 用于查找物理地址所对应的IP地址，例如对于无盘机，启动时需要知道自己的IP地址



3. BOOTP、DHCP和IGMP

a) 引导协议BOOTP

p 引导协议（BOOTP）是一种基于UDP/IP的协议。这种协议允许正在启动的主机动态配置而无需用户监督

p BOOTP主要用于客户机从服务器获得自己的IP地址、服务器的IP地址以及启动映像文件名

p 其他的一些配置信息，如本地子网掩码、本地时间偏移量、默认路由器地址和各种Internet服务器地址，都能通过BOOTP协议与客户机交流

b) 动态主机配置协议DHCP

p 从一个地址池中把IP地址分配给请求主机

p 该协议既允许手工分配IP地址，也允许自动分配IP地址

p DHCP也能提供其他信息，如网关IP、DNS服务器、缺省域和网络范围内HOSTS文件的位置

c) Internet组管理协议IGMP

p IGMP用来帮助组播路由器识别加入到一个组播群组的成员主机，用于IP主机向任一个直接相邻的路由器报告他们的组成员情况

p IGMP使用IP数据报传递其报文（即IGMP报文加上IP首部构成IP数据报）

四、 拥塞控制

p 当通信子网中有太多的分组，导致其性能降低，这种情况叫拥塞

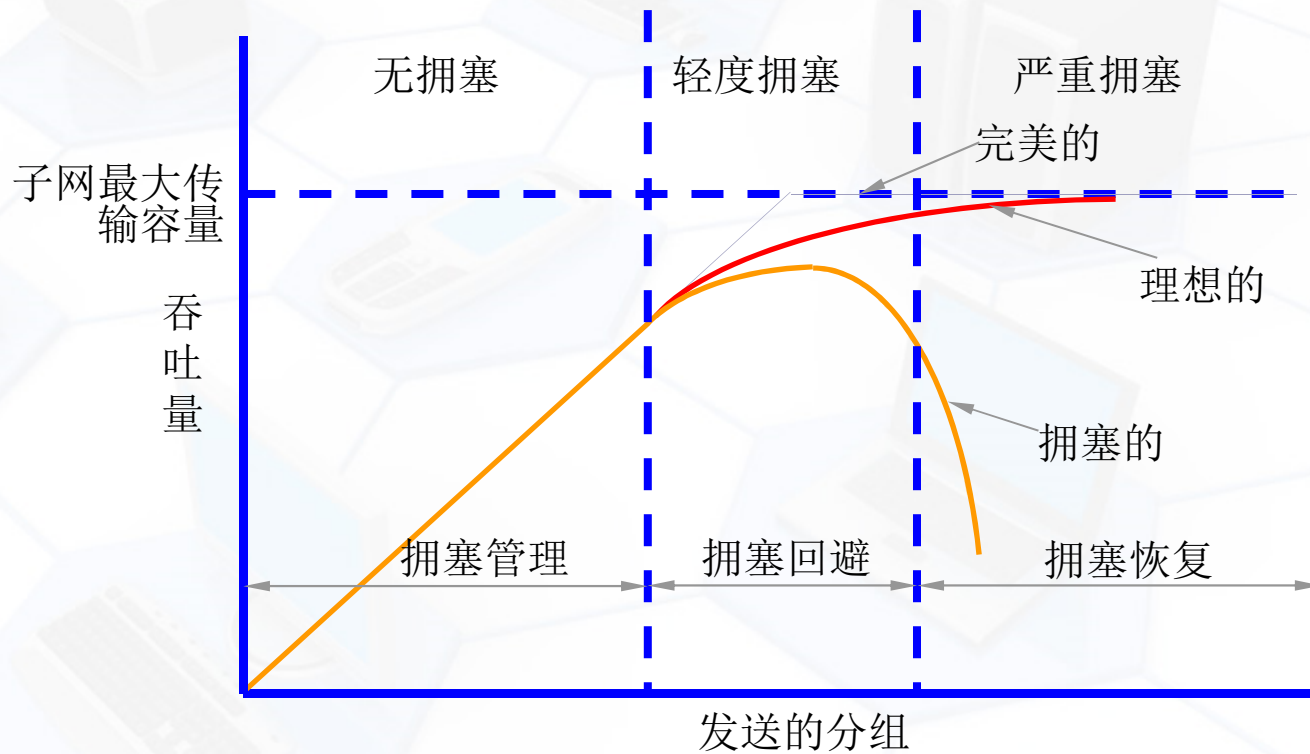
p 造成拥塞的原因

- ü 节点存储容量（缓冲区）不够

- ü 处理机速度太低

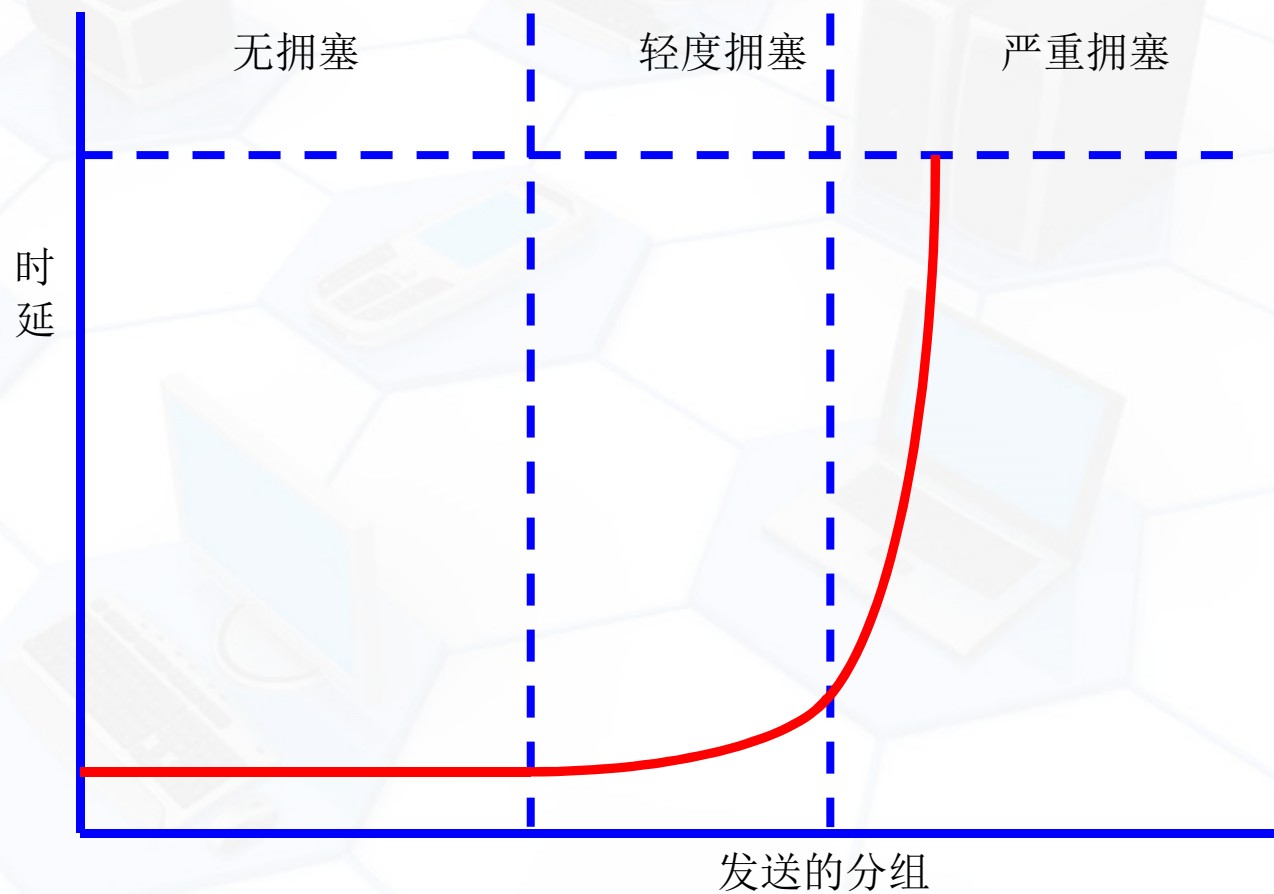
- ü 线路容量（带宽）不够

p 当通信量太大时，发生拥塞，性能显著降低



拥塞示例

p 拥塞对延迟的影响



1. 拥塞控制的基本原理——开环控制

p 通过良好的设计来避免问题的出现，确保问题在一开始就不会出现

p 基于源节点的控制：由发送端决定进入网络的速率，确保按照这个速率网络不会拥塞（常用）

p 基于目的节点的控制：控制进入目的主机的流量

2. 拥塞控制的基本原理——闭环控制

p 建立在反馈的基础上，由三部分组成：

ü 监视系统，检测何时何地发生了拥塞

ü 将此信息传送到可能采取行动的地方

ü 调整系统操作以更正系统

ü 闭环控制方式

显式反馈：当某一节点发现拥塞时，它发一个回答帧给相关节点，通知它们网络拥塞了

隐式反馈：通过定时器方法，发送端每发一个帧，就启动一个定时器，如在规定的时间内没有收到相应的ACK，则认为该帧丢失，如丢失率相当高，则认为网络发生了拥塞

3. 流量整型

p 拥塞发生的原因往往是通信量的突发性，如果主机能以恒定的速率发送数据，则拥塞将会少得多

p 流量整型是调整数据传输的平均速率，客户与传输载体之间需进行协商

比如：流量控制的漏桶算法(Leakey Bucket Algorithm) ——平滑输入流量，控制进入网络的流量

4. 流说明

- p** 为避免拥塞，通信前主机与网络之间要协商一下，以保证网络的服务质量
- p** 流说明用来说明流量的模式，它描述流量输入模式，也描述了应用程序的服务质量要求
- p** 在建立连接或传输数据以前，源端向子网和接受端提交流说明以求批准，子网可以接受，也可以拒绝，或者提出一个建议与源端协商

5. 抑制报文

p 当路由器发现出现拥塞时，向源端主机发一个抑制报文，通知它们降低发送速率，如ICMP中的源抑制报文

p 当源主机不减慢其发送速率时，则应采取强迫手段

ü 公平排队

强迫各发送源公平地占用信道

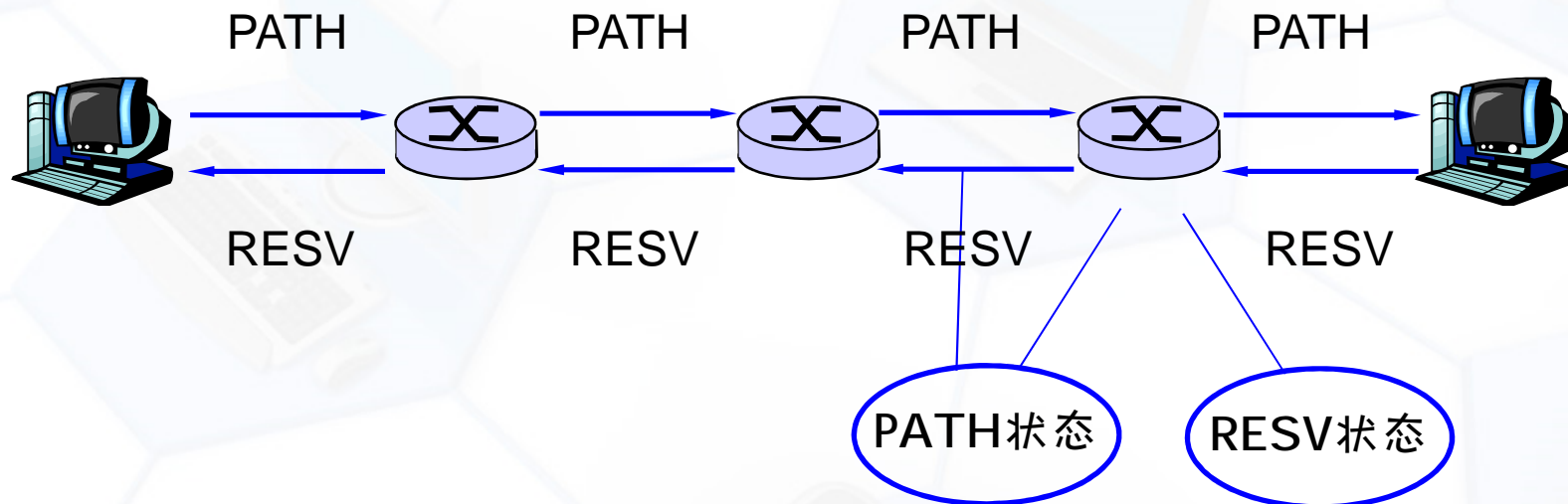
ü 加权公平排队

每个发送源有不同的优先级，即不同的流有不同的带宽

6. 资源预留协议

p最常用的资源预留协议是RSVP（Resource reSerVation protocol），协议将在沿途的路由器上预留一定的资源，包括带宽、缓冲区、表空间等

pRSVP是一种基于接收端，并由接收端发起的资源预留协议



网络层

IP (IPv4 IPv6) ▪ ICMP ▪ ICMPv6 ▪
IGMP ▪ IS-IS ▪ IPsec ▪ BGP ▪ RIP
▪ OSPF ▪ ARP ▪ RARP ▪ 更多

数据链路层

Wi-Fi (IEEE 802.11) ▪
WiMAX (IEEE 802.16) ▪ ATM ▪ DTM
▪ 令牌环 ▪ 以太网 ▪ FDDI ▪
帧中继 ▪ GPRS ▪ EV-DO ▪ HSPA ▪
HDLC ▪ PPP ▪ L2TP ▪ ISDN ▪ SPB
▪ STP ▪ 更多

物理层

以太网 ▪ 调制解调器 ▪ 电力线通信
▪ 同步光网络 ▪ G.709 ▪ 光导纤维
▪ 同轴电缆 ▪ 双绞线 ▪ 更多

目前层次各层协议汇总