§5. 网络层

5.1. 网络层的设计问题

★ 与数据链路层的区别

数据链路层：将帧从线路的一端传输到另一端

网络层：将完整的数据包从发送方传送到接收方

- 完整的数据包可能拆分为多帧
- 完整的数据包可能要经过多个不同的数据链路层

★ 在网络协议中的位置

...

传输层

网络层　　=> 处理端到端数据的最底层

数据链路层

...

★ 路由器的基本概念

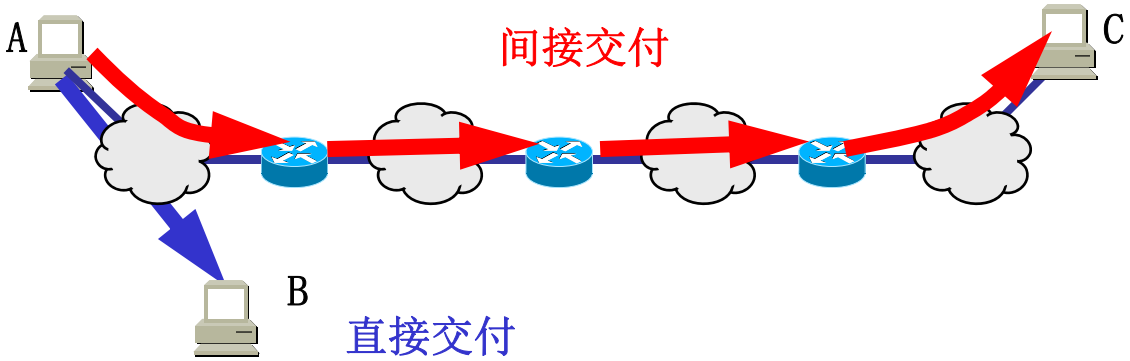
- 网络层进行完整数据包的分组/转发的设备
- 当使用中继器/集线器/网桥/以太网交换机(二层)时，一般不称为网络互连，因为这仅仅是把一个网络扩大了，逻辑上仍当做一个网络看待
- 传输网关/应用网关由于较复杂，使用得较少
- 一般情况互联网都是指用路由器进行互连的网络
- 由于历史的原因，许多有关 TCP/IP 的文献将网络层使用的路由器称为网关
- 以太网交换机(三层)在讨论时当做路由器看待

§5. 网络层

5.1. 网络层的设计问题

★ 直接交付与间接交付

- 直接交付：源和目的主机间不经过路由器而直接传递数据包(会经过网桥/HUB等)
- 间接交付：源和目的主机间通过1-n个路由器的转发来传递数据包(也有网桥/HUB等)



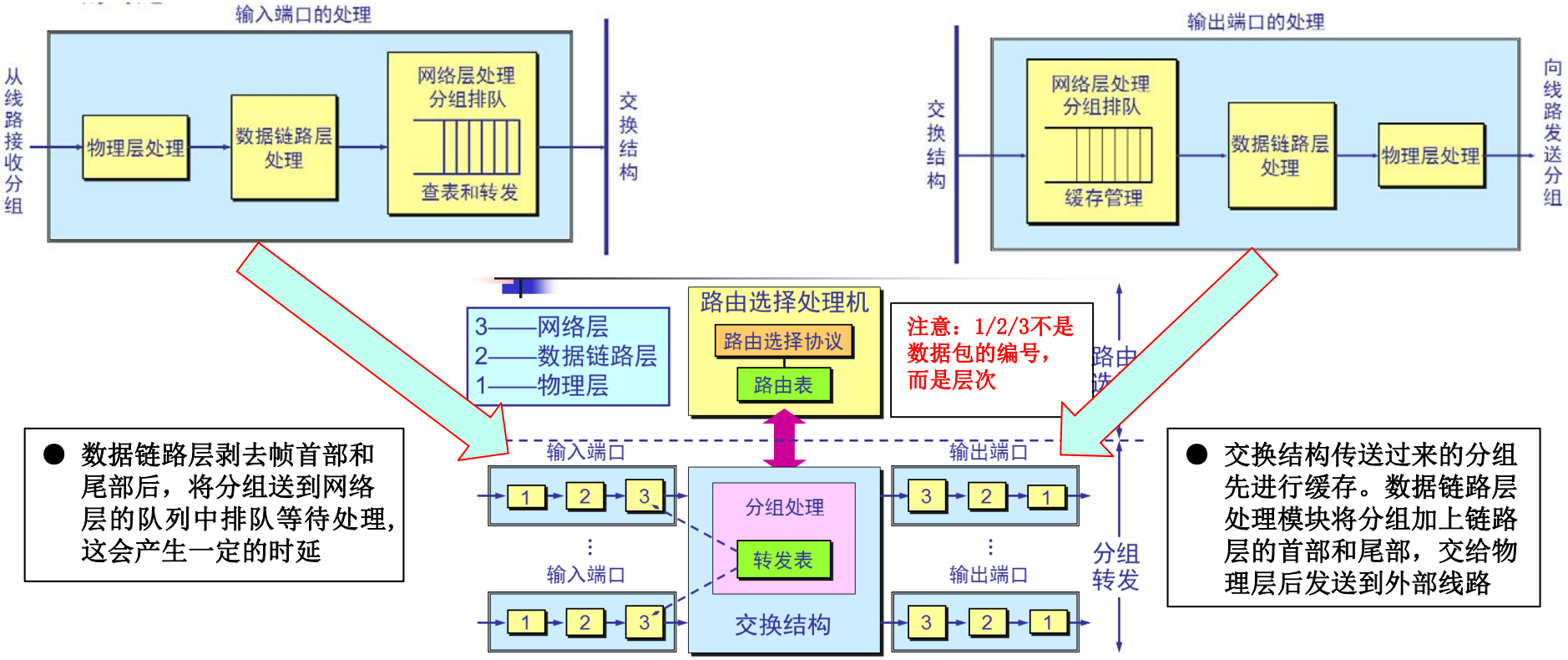
§5. 网络层

5.1. 网络层的设计问题

★ 路由器的工作原理

- 当主机 A 要向另一个主机 B 发送数据报时，先要检查目的主机 B 是否与源主机 A 连接在同一个网络上(对IP地址，就是是否同一网段)
- 如果是，就将数据报直接交付给目的主机 B 而不需要通过路由器(可能会经过网桥/HUB等)
- 如果不是，则应将数据报发送给本网络上的某个路由器，由该路由器按照转发表指出的路由将数据报转发给下一个路由器，多次转发直到到达目的主机B为止(也可能会丢弃)

=> 路由器至少两个接口，连接两个网段



§5. 网络层

5.1. 网络层的设计问题

★ 路由器的工作原理

- "转发"(forwarding)就是路由器根据转发表将用户的 IP 数据报从合适的端口转发出去
- "路由选择"(routing)则是按照分布式算法，根据从各相邻路由器得到的关于网络拓扑的变化情况，动态地改变所选择的路由
- 路由表是根据路由选择算法得出的。而转发表是从路由表得出的
- 在讨论路由选择的原理时，往往不区分转发表和路由表的区别

★ 分组丢弃

- 若路由器处理分组的速率赶不上分组进入队列的速率，则队列的存储空间最终必定减少到零，这就使后面再进入队列的分组由于没有存储空间而只能被丢弃，称为分组丢弃
- 路由器中的输入或输出队列产生溢出是造成分组丢弃的重要原因

★ 虚拟互联网络

- 所谓虚拟互连网络也就是逻辑互连网络，含义是互连起来的各种物理网络的异构性本身是客观存在的，但可以利用跨不同物理链路层的网络层协议使这些性能各异的网络从角度用户看起来好像是一个统一的网络
- 使用虚拟互连网络的好处是，当互联网上的主机进行通信时，就好像在一个网络上通信一样，而看不见互连的各具体的网络异构细节
- 使用 IP 协议的虚拟互连网络可简称为 IP 网

§5. 网络层

5.1. 网络层的设计问题

5.1.1. 存储转发数据包交换

★ 网路层的数据包传输采用存储-转发机制

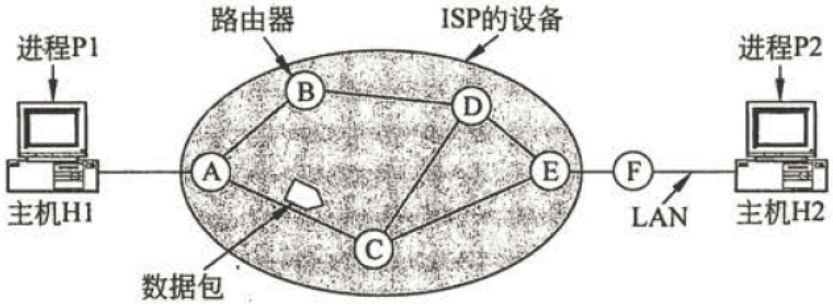


图 5-1　网络层协议的环境

- 所提供的服务与数据链路层类型无关
- 网络层的数量、类型和拓扑结构对传输层透明
- 网络层的编址方式应当跨越多个LAN及WAN

5.1.2. 提供给传输层的服务

★ 网络层为连接在网络上的主机所提供的服务有两类：

- 无连接的网络服务（数据报服务）
- 面向连接的网络服务（虚电路服务）

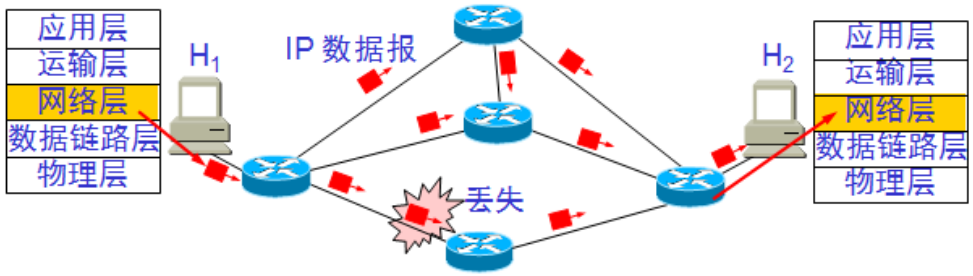
§5. 网络层

5.1. 网络层的设计问题

5.1.3. 无连接服务的实现

★ 基本思路

- 传输层传输的数据可以被分为若干数据报(datagram)
- 每个数据报在网络层传输时是独立的
- 对应的网络层称为数据报网络(datagram network)



H_1 发送给 H_2 的分组可能沿着不同路径传送

★ 无连接服务的特点

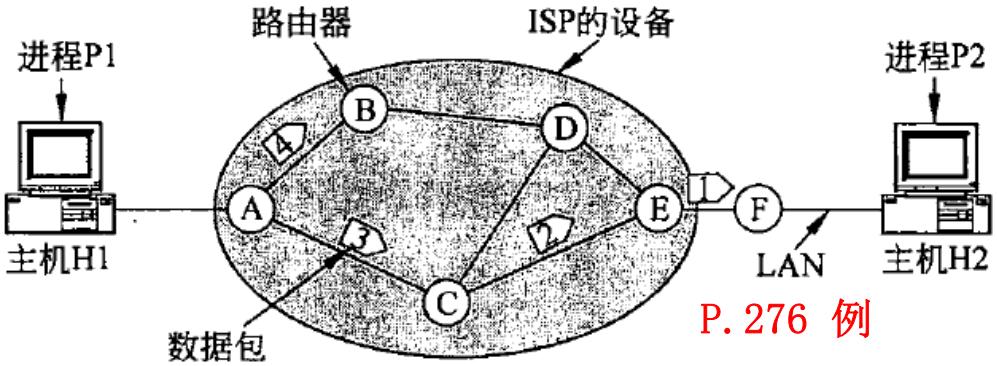
- 网络层向上提供简单灵活的、无连接的、不可靠的数据报服务（尽最大努力交付）
- 网络层在发送分组时不需要先建立连接。每一个分组独立发送，与其前后的分组无关（不编号）
- 网络层不提供服务质量的承诺。即所传送的分组可能出错、丢失、重复和失序(不按序到达终点)，当然也不保证分组传送的时限
- Internet即使此方式，数据报为IP数据报

§5. 网络层

5.1. 网络层的设计问题

5.1.3. 无连接服务的实现

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



P.276 例

隐含的信息

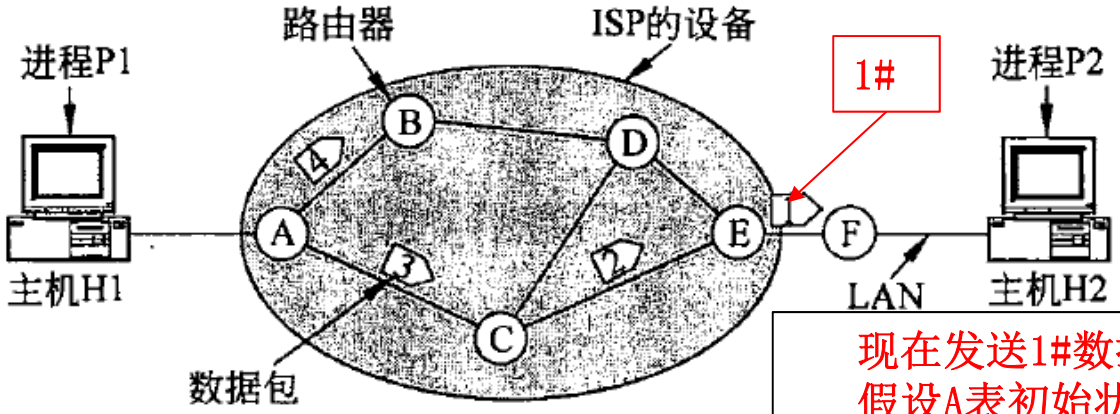
1、ABCDEF至少有3/2/3/3/3/2个接口
（每条线代表一对同网段地址，不同线不同）

2、主机H1和H2不在同一网段（需要间接交付）

3、主机H1与A的一个接口在同一网段，主机H1的缺省网关设置为A的同网段地址

4、主机H2与F的一个接口在同一网段，主机H2的缺省网关设置为F的同网段地址

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



现在发送1#数据包：
假设A表初始状态如左图

A的表(初始化)

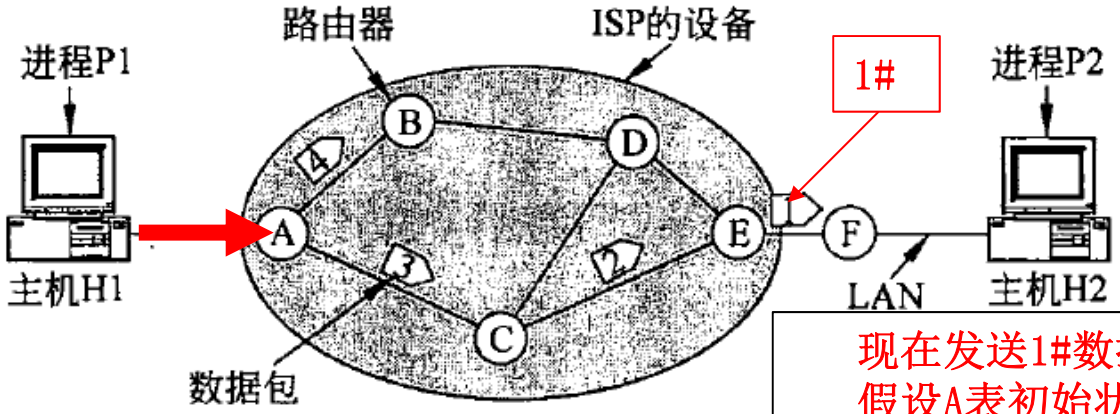
A	—
B	B
C	C
D	B
E	C
F	C

直接转发

间接转发

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



现在发送1#数据包：
假设A表初始状态如左图
1、A收到H1的1#包，存储

A的表(初始化)

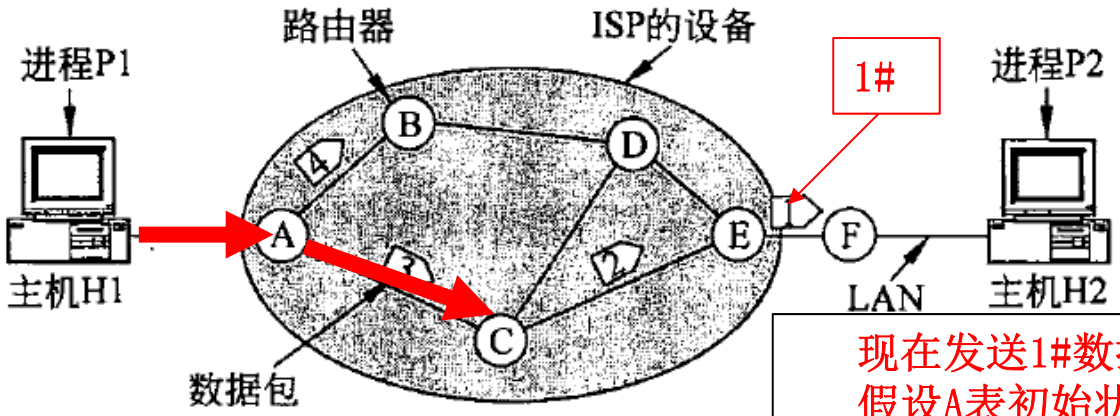
A	—
B	B
C	C
D	B
E	C
F	C

直接转发

间接转发

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



现在发送1#数据包：
假设A表初始状态如左图
1、A收到H1的1#包，存储
2、A查询自己的路由表，得知F表有一个接口与1#包的目的地址(H2)在同一网段，因此向C转发

A的表(初始化)

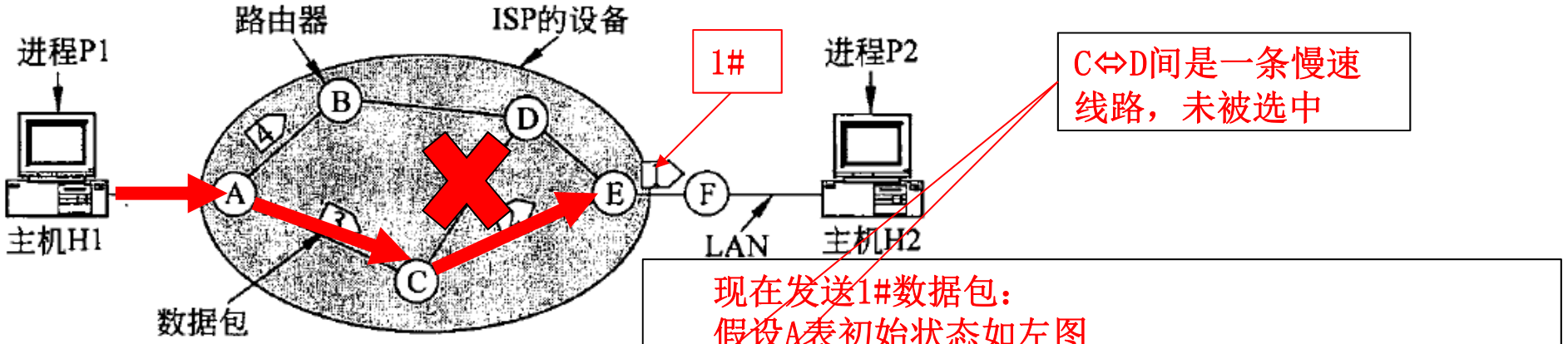
A	—
B	B
C	C
D	B
E	C
F	C

直接转发

间接转发

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器

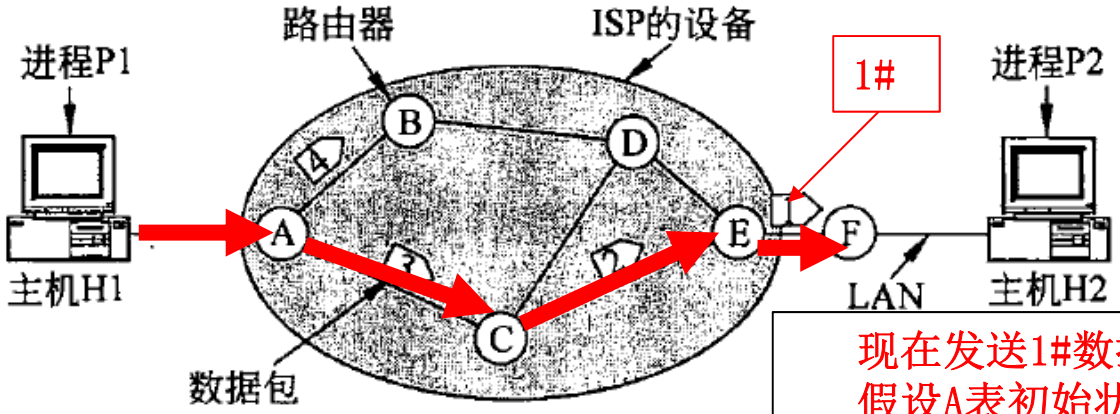


C的表

A	A
B	A
C	—
D	E
E	E
F	E

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



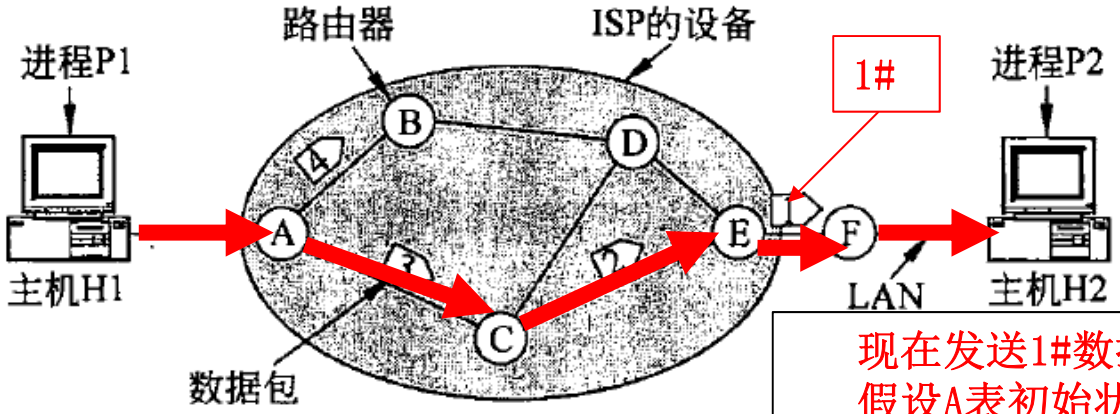
现在发送1#数据包：
假设A表初始状态如左图
1、A收到H1的1#包，存储
2、A查询自己的路由表，得知F表有一个接口与1#包的目的地址(H2)在同一网段，因此向C转发
3、C从A收到、存储、查表、向E转发
4、E从C收到、存储、查表、向F转发
(E的路由表中向F应直接转发)

E的表

A	C
B	D
C	C
D	D
E	—
F	F

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



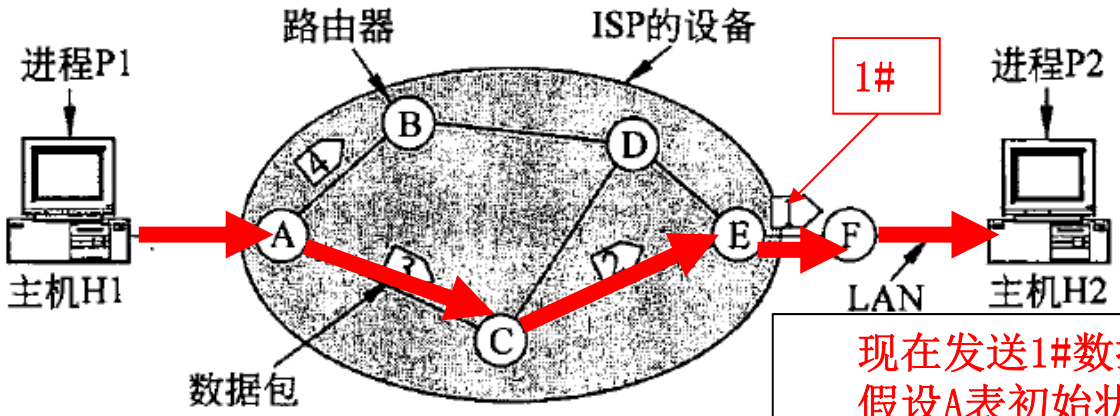
F的表

A	E
B	E
C	E
D	E
E	E
F	–

现在发送1#数据包：
假设A表初始状态如左图
1、A收到H1的1#包，存储
2、A查询自己的路由表，得知F表有一个接口与1#包的目的地址(H2)在同一网段，因此向C转发
3、C从A收到、存储、查表、向E转发
4、E从C收到、存储、查表、向F转发
5、F从E收到、存储、查表、向H2转发
(F的路由表中向H2应直接转发)

图 5-2　数据报网络中的路由过程

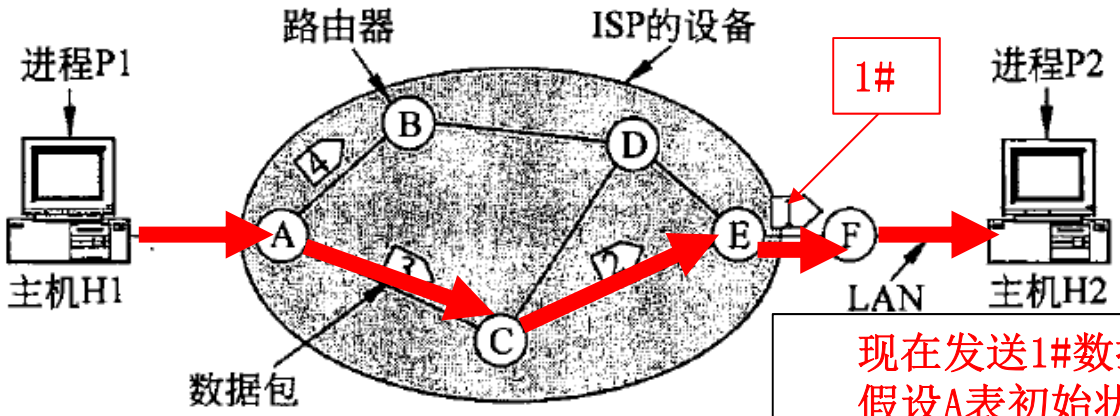
★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



现在发送1#数据包：
假设A表初始状态如左图

1#包：
H1 -> A -> C -> E -> F -> H2

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器

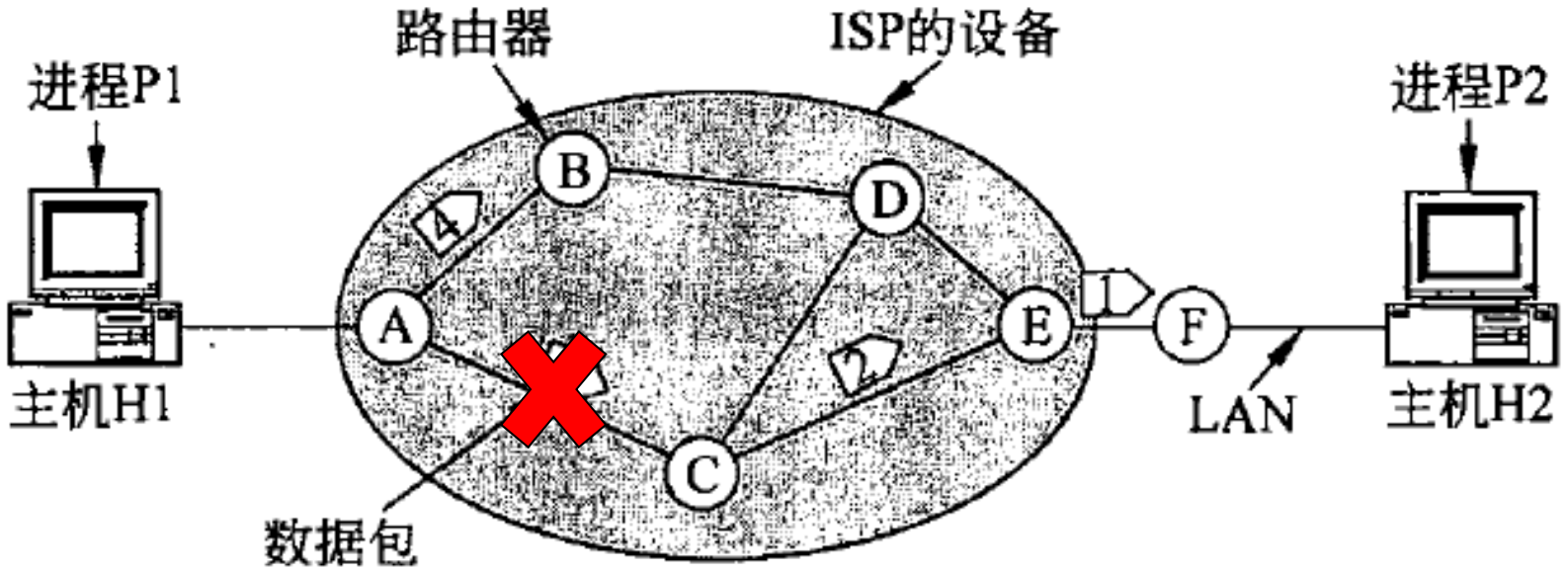


现在发送1#数据包：
假设A表初始状态如左图

1#包：
H1 -> A -> C -> E -> F -> H2

假设2#/3#相同路径转发

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



A的表(初始化)

A	—
B	B
C	C
D	B
E	C
F	C

直接转发

间接转发

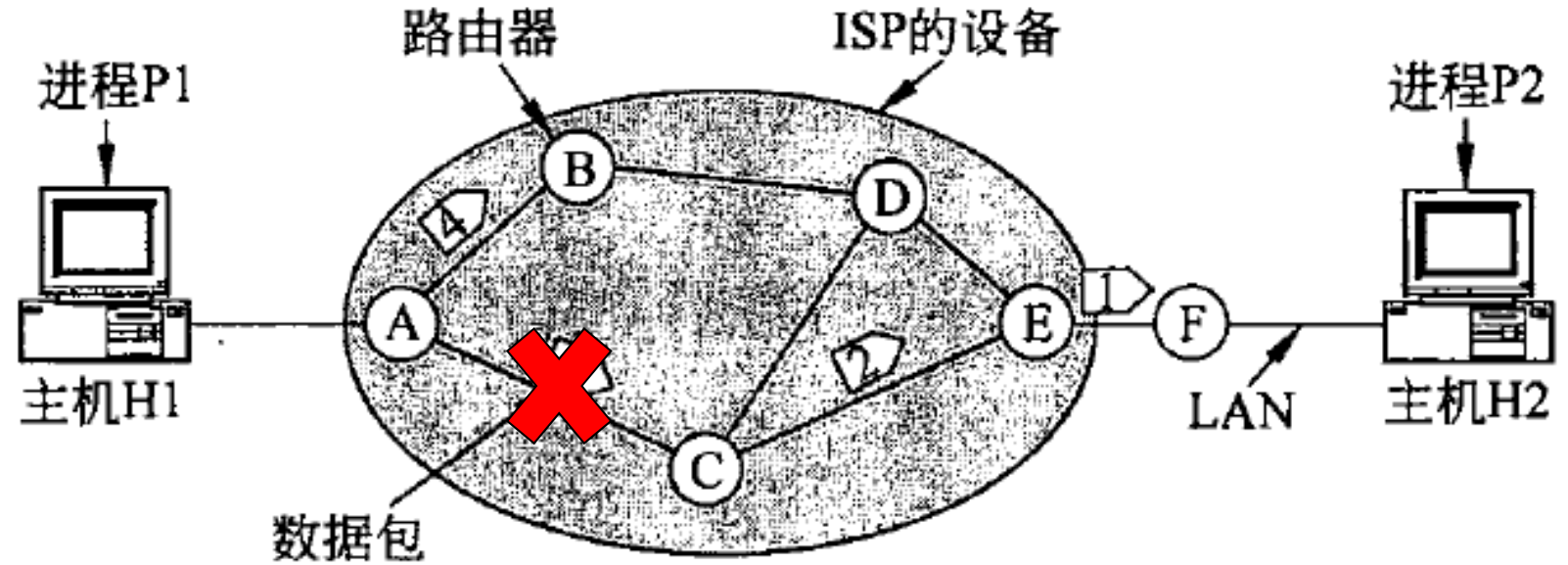
A的表(稍后)

A	—
B	B
C	C
D	B
E	B
F	B

假设A⇔C间线路断开
或 A⇔C⇔E间拥塞

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



A的表(初始化)

A	—
B	B
C	C
D	B
E	C
F	C

直接转发

间接转发

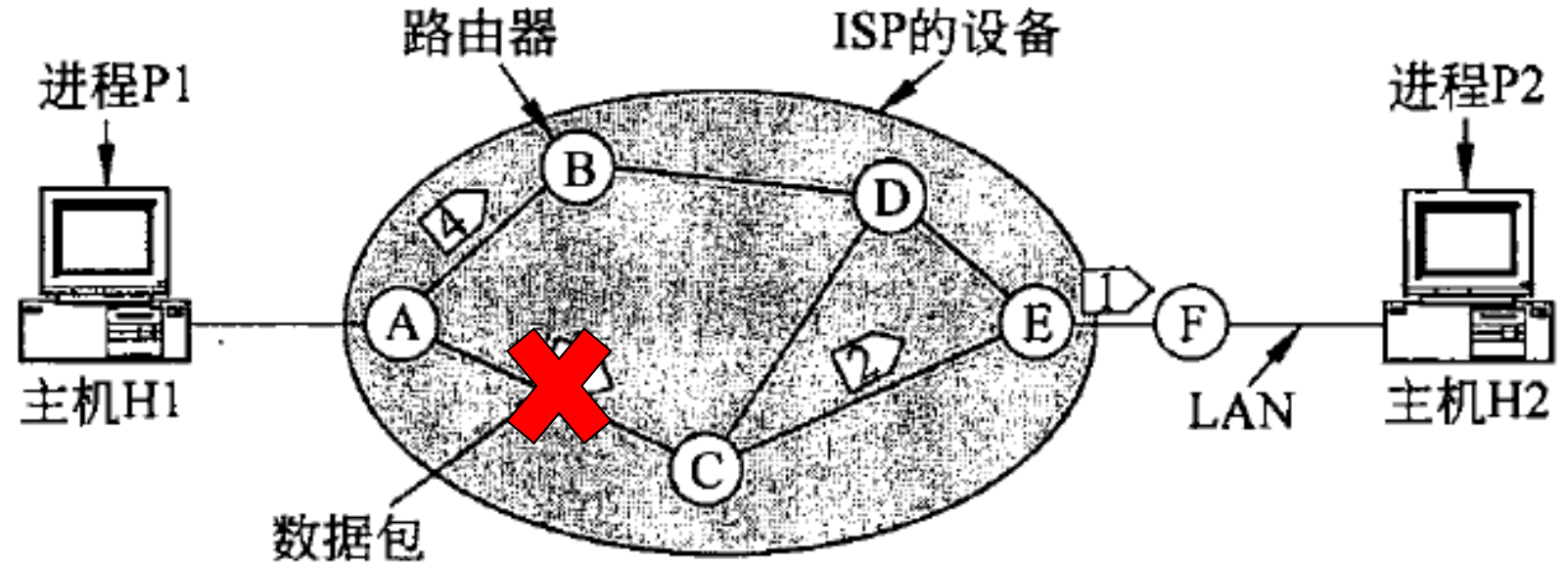
A的表(稍后)

A	—
B	B
C	C
D	B
E	B
F	B

假设A⇔C间线路断开
或 A⇔C⇔E间拥塞
A的路由表被更新

图 5-2　数据报网络中的路由过程

★ 例：设P1/P2为传输层进程，P1准备向P2发送的数据被分为4个网络层数据报发送，ABCDEF为路由器



A的表(初始化)

A	—
B	B
C	C
D	B
E	C
F	C

直接转发

间接转发

A的表(稍后)

A	—
B	B
C	C
D	B
E	B
F	B

假设A⇔C间线路断开
或 A⇔C⇔E间拥塞
A的路由表被更新
此时发送4#包，则路径为：
H1->A->B->D->E->F->H2

图 5-2　数据报网络中的路由过程

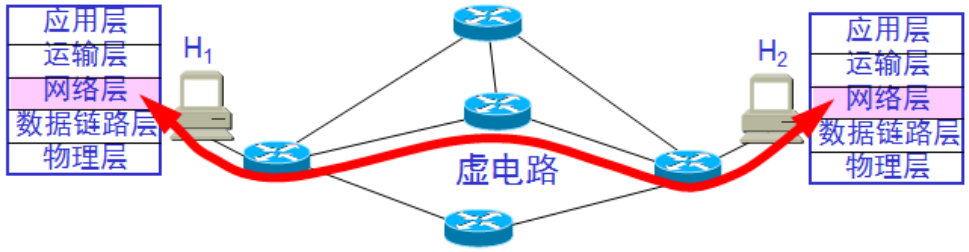
§5. 网络层

5.1. 网络层的设计问题

5.1.4. 面向连接服务的实现

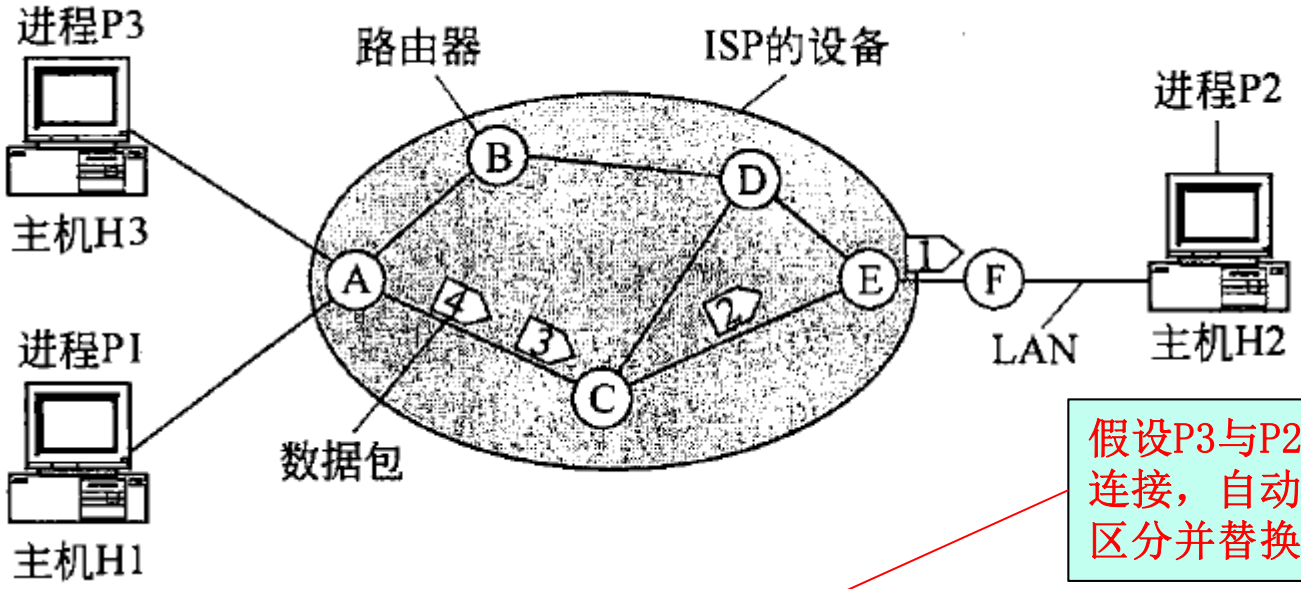
★ 基本思路

- 建立虚电路(Virtual Circuit)，以保证双方通信所需的一切网络资源
- 面向连接的通信方式，当连接建立后，从源机器到目标机器的一条路径就被当做这个连接的一部分被确定下来并保存在中间路由器的表中
- 所有在这个连接上通过的流量都使用这条路径
- 如果再使用可靠传输的网络协议，就可使所发送的分组无差错按序到达终点



H_1 发送给 H_2 的所有分组都沿着同一条虚电路传送

★ 例：设P1/P2/P3为传输层进程，ABCDEF为路由器，P1/P2间已建立连接1，
P3/P2间已建立连接2 （均为ACEF）



假设P3与P2的连接是P3的第1条连接，自动被编号为1，但A要能区分并替换为2

A的表

入境		出境	
H1	1	C	1
H3	1	C	2

C的表

A	1	E	1
A	2	E	2

E的表

C	1	F	1
C	2	F	2

图 5-3 虚电路网络的路由过程

§5.网络层

5.1.网络层的设计问题

5.1.5.虚电路与数据包网络的比较

	数据报网络	虚电路网络
电路建立	不需要	需要
寻址	每个包包含全部的源和目的地址	每个包包含简短的VC号
状态信息	路由器不保留连接状态	针对每个连接，每条VC都需要路由器保存其状态
路由方式	每个数据包被单独路由	建立VC时选择路由，所有包都遵循该路由
路由器失效的影响	没影响，除了那些路由器崩溃期间丢失的包	穿过故障路由器的所有VC都将中断
服务质量	困难	容易，如果在预先建立每条VC时有足够的资源可分配
拥塞控制	困难	容易，如果在预先建立每条VC时有足够的资源可分配

§5.网络层

5.2.路由算法

★ 路由选择算法的特征

- 正确性：算法必须是正确的和完整的
- 简单性：算法在计算上应简单
- 适应性（健壮性）：算法应能适应通信量和网络拓扑的变化，这就是说，要有自适应性
- 稳定性（收敛性）：算法应具有稳定性
- 公平性：算法应是公平的
- 最优性：算法应是最佳的

★ 路由代价

- "代价"是由一个或几个因素综合决定的一种度量值(metric)，如链路长度、数据率、链路容量、是否要保密、传播时延等，甚至还可以是一天中某一个小时内的通信量、结点的缓存被占用的程度、链路差错率等
- 在研究路由选择时，需要给每一条链路指明一定的代价

★ 最佳路由

- 不存在一种绝对的最佳路由算法。
- 所谓"最佳"只能是相对于某一种特定要求下得出的较为合理的选择而已。
- 实际的路由选择算法，应尽可能接近于理想算法
- 路由选择是个非常复杂的问题，是网络中的所有结点共同协调工作的结果
- 路由选择的环境往往是不断变化的，而这种变化有时无法事先知道

§5. 网络层

5.2. 路由算法

★ 静态路由与动态路由

- 静态路由：在离线环境下人为事先计算好并写入路由器中，是非自适应路由选择，其特点是简单和开销较小，但不能及时适应网络状态的变化
- 动态路由：在路由器加入网络后，与相邻路由器交换路由信息数据包并按一定规则计算并动态得到，是自适应路由选择，其特点是能较好地适应网络状态的变化，但实现起来较为复杂，开销也比较大

★ 分层次的路由选择协议

- 在网络规模很大（例如：Internet）时，如果让所有路由器相互交换信息，则代价很大（CPU/链路）（以至于不可行）
- 有时单位希望隐藏自己的网络布局细节，但同时希望能与外部相连

★ 自治系统AS（Autonomous System）

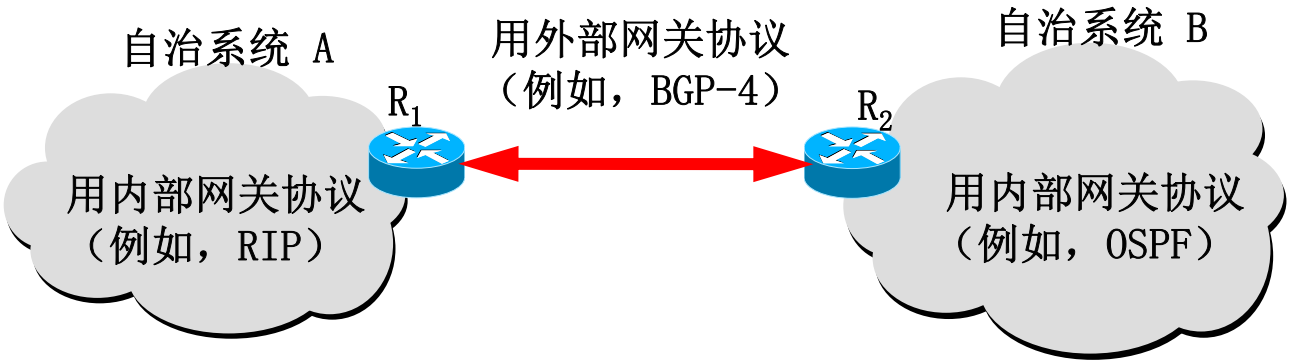
- AS是指在单一的技术管理下的一组路由器，这些路由器使用一种 AS 内部的路由选择协议和共同的度量以确定数据报在该 AS 内的路由，同时还使用一种 AS 之间的路由选择协议来确定数据报在 AS之间的路由
- AS内部可以使用多种内部路由选择协议和度量，但AS之间要有一个单一的和一致的路由选择策略

§5. 网络层

5.2. 路由算法

★ IGP与EGP

- IGP(Interior Gateway Protocol)：内部网关协议，即在一个自治系统内部使用的路由选择协议，目前这类路由选择协议使用的最多（例如：RIP 和 OSPF 协议）
- EGP (External Gateway Protocol)：外部网关协议，即源站和目的站处在不同的自治系统中，当数据报传到一个自治系统的边界时，就需要使用一种协议将路由选择信息传递到另一个自治系统中。这样的协议就是外部网关协议 EGP（例如：BGP）



- 自治系统间的路由选择也称为域间路由选择(interdomain routing)
- 自治系统内部的路由选择也称为域内路由选择(intradomain routing)

§5. 网络层

5.2. 路由算法

5.2.1. 优化原则

★ 最优化原则：如果路由器J在从路由器I到路由器K的最优路径上，则从J到K的最优路径也必定遵循同样的路由

- 证：(略)，可参考数据结构中最小代价生成树中必然包含最短路径的反证法

★ 汇集树(sink tree)：所有源端到目的端最佳路由集合，形成以目的地为根的树

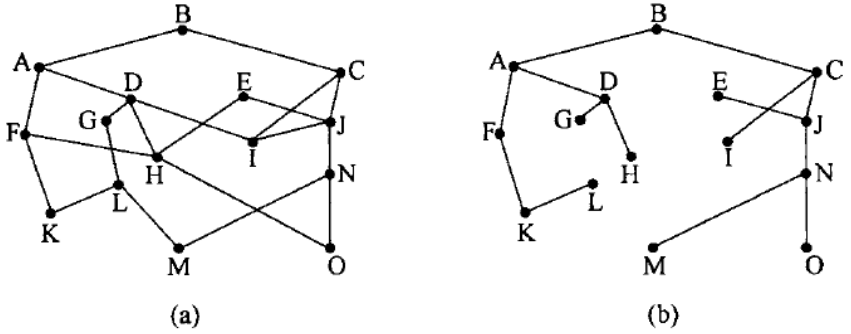


图 5-6

(a) 一个网络；(b) 路由器 B 的汇集树

- 汇集树不唯一

§5.网络层

5.2.路由算法

5.2.2.最短路径算法

★ 路径度量：度量值一般称为cost，可以是距离、站点数（hop）、信道带宽、平均通信量、通信开销、队列平均长度、传输延时等，或是它们的函数

★ 算法：Dijkstra算法（某一顶点到其它顶点的最短路径，某一顶点 = 路由器自身）

§5. 网络层

5.2. 路由算法

5.2.3. 泛洪算法

★ 算法：路由器收到主机发来的数据包，复制后，再向除到达线路以外的所有输出线路发送

- 如果不加以控制，就会产生大量的重复数据包，称为过度扩散（二次扩散）

★ 控制过度扩散（二次扩散）的方法

- 在数据包中设置跳计数器，初值为最大网络直径，每经过一台路由器减1，若为0丢弃
  - 无法预估网络直径
  - 重复数据包会指数级增长，即使有丢弃机制，仍会占用大量的带宽和处理时间
- 在数据包中设置序号，通过序号避免二次转发
  - 源路由器对来自主机的每个数据包设置序号
  - 每台路由器对每台源路由器设一张序号表
  - 每台路由器对经过的分组检查序号，发现重复则丢弃，不再转发

★ 泛洪算法的应用

- 军事领域：鲁棒性好（robust=健壮性）
- 一定能选出最短路径

§5.网络层

5.2.路由算法

5.2.4.距离矢量算法

★ 距离矢量路由（distance vector routing）

- 每台路由器维护一张表（一个矢量），表中列出本路由器到当前已知的每个目标的最佳距离以及所使用的链路，相邻路由器不断交换路由信息并根据收到的新信息更新自己，最终每个路由器都能了解到到达每个目的地的最佳链路
- 常见的是分布式 Bellman-Ford 路由算法
- 相应的路由协议称为 RIP 协议 (RIP = Routing Information Protocol)

§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

- 简介
  - 内部网关协议中最先得到广泛使用的协议
  - 是基于距离矢量算法的路由选择协议
  - 协议要求网络中的每一个路由器都要维护从自己到其他每一个目的网络的距离记录
    （一个路由器有多个接口，对应多个IP网段，因此一个路由器有多个目的网络）
- 距离
  - 从路由器到直接连接的网络的距离定义为1
  - 从路由器到非直接连接的网络的"距离"定义为所经过的路由器数加1（跳数=hop count）
  - RIP认为一个好的路由就是通过的路由器数目少，即"距离短"（距离 = 最短距离）
  - RIP允许一条路径最多只能包含 15 个路由器，即"距离"为16 时即相当于不可达
    （适用于小型网络）
  - RIP不能在两个网络间同时使用多条路由，RIP的选择标准是"距离短"而不是"速度快"
- 邻接路由器
  - 在点对点网络中互连的两台路由器称为邻接路由器，P2P网络不强制要求双方必须在同一网段
  - 在广播网络中存在物理连接，且某个接口与本路由器的某个接口在同一网段的路由器也称为邻接路由器

§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

- 收敛

  整个网络中最佳路径的寻找过程称为收敛

- 路由器间信息的交换
  - RIP仅和邻接路由器交换信息
  - 交换的信息是本路由器所知道的全部信息，即自己的路由表（包括自身端口的直连网络信息和已经通过RIP学习到的信息）
  - 每条路由信息都包含四项，分别是IP地址、子网掩码、下一跳路由器地址、距离
  - 按固定的时间间隔来进行路由信息的交换(一般是30秒)
  - 通过不断与相邻路由器交换信息，使路由器到每个目的网络的距离(跳数)都是最短的
  - 网络中每个路由器的路由表是不同的

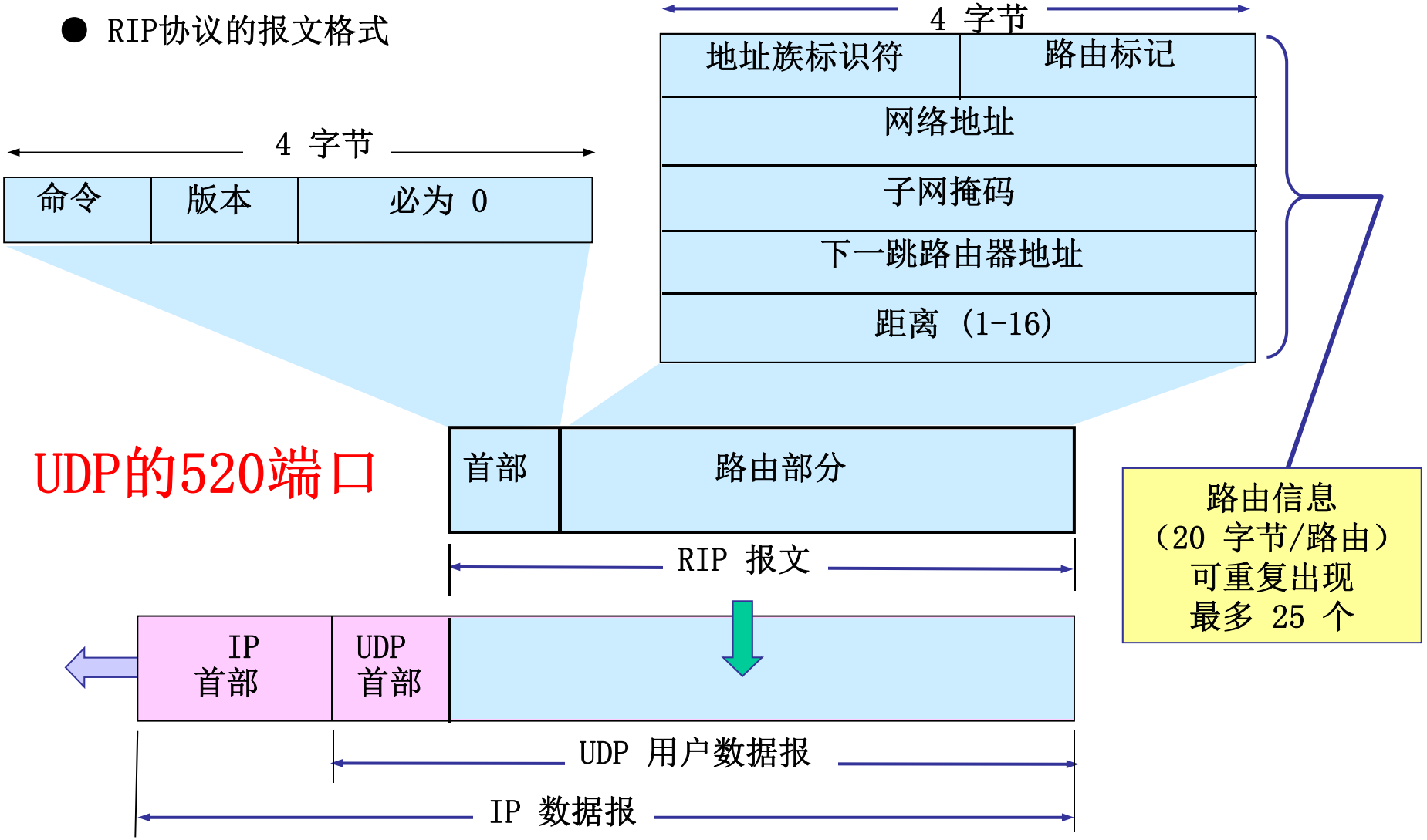
§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

● RIP协议的报文格式



§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

● RIP协议的算法

1）初始填入直连网络的距离(距离=1)

2）收到邻接路由器(假设地址为 X)的RIP报文后，修改此RIP报文中的所有项目，把"下一跳"字段中的地址都改为 X，并把所有的"距离"值加1

3）对修改后的 RIP 报文中的每一个项目，重复以下步骤：

a）若项目中的目的网络不在路由表中，则把该项目加到路由表中

b）若项目中的目的网络已在路由表中，则检查下一跳的地址，若下一跳的地址相同则替换路由表中对应项(不检查报文中的距离和现有路由表的距离，直接替换)

c）若项目中的目的网络已在路由表中，且与现有路由表的下一跳地址不同，则检查报文中该项的距离是否小于现有路由表中的距离，若是则替换路由表，否则什么也不做

4）若180秒还未收到邻接路由器的更新，则把此邻接路由器的距离标记为16(不可达)

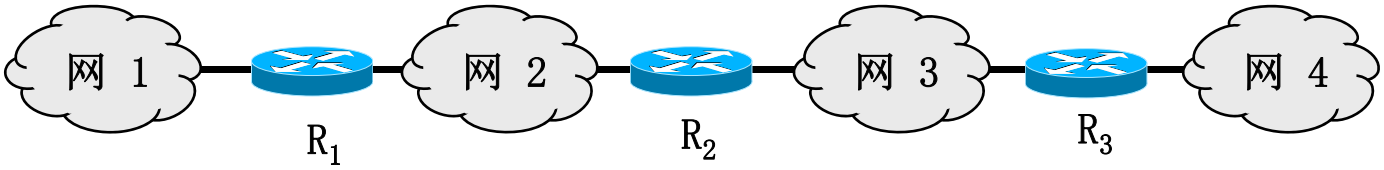
§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

● 快扩散与慢收敛



◆ 网1/2/3/4是不同网段

◆ R1/R2、R2/R3互为邻接路由器，R1/R3不是

◆ 初始
R1中有网1/2的信息，直连，距离为1
R2中有网2/3的信息，直连，距离为1
R3中有网3/4的信息，直连，距离为1

◆ 邻接路由器第一次交换信息后
R1有网1/2的信息，直连，距离为1
有网3的信息，下一跳R2，距离为2
R2有网2/3的信息，直连，距离为1
有网1的信息，下一跳R1，距离为2
有网4的信息，下一跳R3，距离为2
R3有网3/4的信息，直连，距离为1
有网2的信息，下一跳R2，距离为2

◆ 邻接路由器第二次交换信息后
R1有网1/2的信息，直接，距离为1
有网3的信息，下一跳R2，距离为2
有网4的信息，下一跳R2，距离为3
R2有网2/3的信息，直接，距离为1
有网1的信息，下一跳R1，距离为2
有网4的信息，下一跳R3，距离为2
R3有网3/4的信息，直接，距离为1
有网2的信息，下一跳R2，距离为2
有网1的信息，下一跳R2，距离为3

◆ 邻接路由器第三次（及以后）交换信息后
在网络不出故障，拓扑结构不改变的情况下，
R1/R2/R3的路由表保持稳定，不再发生变化

◆ 每交换一次扩散一跳，称为好消息的快扩散

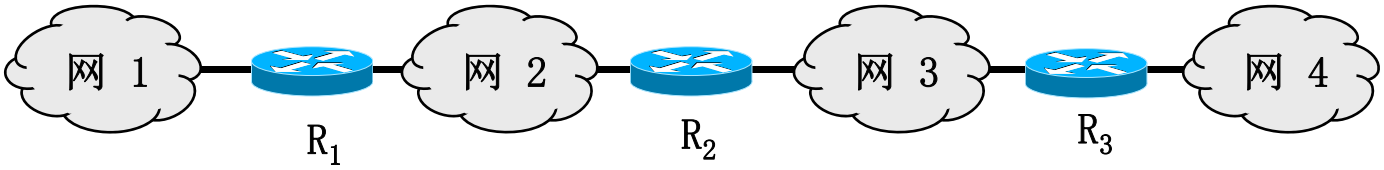
§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

● 快扩散与慢收敛



◆ 经过15次交换信息后，路由表稳定

R2 中有网1的信息，下一跳R1， 距离2

R15中有网1的信息，下一跳R14，距离15

=> R1出故障 或 R1与R2间线路故障

◆ 下一次交换信息时(讨论R2的情况) R2中现有网1的信息为：下一跳R1，距离2 ┌无法从R1收到网1的信息 └从R3收到网1的信息，下一跳自己，距离3 => 改为下一跳R3，距离4，再和现有路由表比较 => R2路由表中网1项不改变，仍为下一跳R1，距离2 (R3-R15中关于网1的路由也不变)	◆ 再次交换时，R2从R3收到网1的信息，下一跳自己，距离3 => 改为下一跳R3，距离4，再和现有路由表比较 => R2路由表中网1项更新为下一跳R3，距离4 => R3收到R2的报文后，到网1改为下一跳R2，距离5
◆ 再次交换 ……	◆ 再次交换 => R2路由表中网1项更新为下一跳R3，距离6 => R3收到R2的报文后，到网1改为下一跳R2，距离7 …
◆ 180秒后，R2仍未收到R1，置路由表中到网1的距离为16	◆ 再次交换 … 直到都是16，才知道网1不可达

◆ 称为坏消息的慢收敛

§5. 网络层

5.2. 路由算法

5.2.4. 距离矢量算法

★ RIP协议（实用中已淘汰，但可用于理解算法思想）

- 快扩散与慢收敛

  思考：如果网1出现问题，会怎样?

  1. 网1是通过P2P方式连接R1
  2. 网1通过广播方式连接R1

  分别讨论

- RIP协议的缺陷
  - 网络的规模小
  - 慢收敛问题

§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ 基本步骤

- 发现邻居节点，得到邻居的网络地址
- 设置到每个邻居节点的距离或者成本度量值
- 构造一个包含刚才获知信息的链路信息包
- 将这个包发送给所有的其它路由器，并接收来自其它路由器的数据包
- 计算到每个其它路由器的最短路径

§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

- 简介
  - 目前使用最广泛的分布式链路状态路由
  - 发送的信息是与本路由器相邻的所有路由器的链路状态（该链路的度量称为metric）
  - 只有当链路状态发生变化时，路由器才用洪泛法向所有路由器发送此信息（不变化则不使用泛洪法，简化处理）
  - 如果到同一目的网络有多条相同代价的路径，则可以将通信量分配给这几条路径。这叫作多路径间的负载均衡
  - 所有在 OSPF 路由器之间交换的分组都具有鉴别的功能（安全性高）

- 链路状态数据库
  - 由于各路由器之间频繁地交换链路状态信息，因此所有的路由器最终都能建立一个链路状态数据库(每个路由器必须具有唯一标识)
  - 这个数据库实际上就是全网的拓扑结构图，它在全网范围内是一致的（称为链路状态数据库的同步）
  - OSPF 的链路状态数据库能较快地进行更新，使各个路由器能及时更新其路由表。OSPF 的更新过程收敛快是其重要优点

§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

- ● OSPF的区域
  - ◆ 为了使 OSPF 能够用于规模很大的网络，OSPF 将一个自治系统再划分为若干个更小的范围，叫作区域
  - ◆ 划分区域的好处就是将利用洪泛法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，减少了网络上的通信量
  - ◆ 在一个区域内部的路由器只知道本区域的网络拓扑，而不知道其他区域的网络拓扑的情况
  - ◆ 每一个区域都有一个 32bit的区域标识符（点分十进制形式，同IPv4）
  - ◆ OSPF 使用层次结构的区域划分。在上层的区域叫作主干区域(backbone area)。主干区的标识符规定为0.0.0.0，作用是用来连通其他在下层的区域
  - ◆ 区域不能太大，建议一个区域内的路由器不超过200个

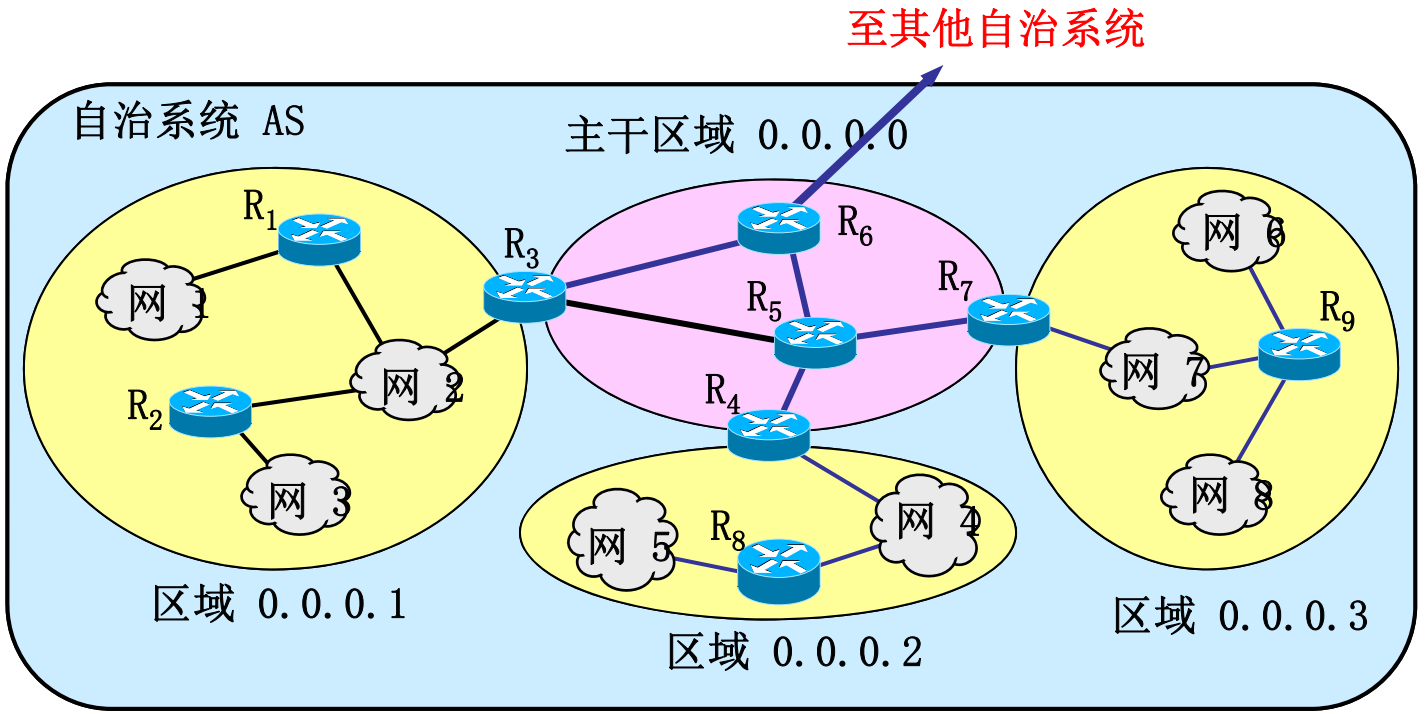
§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

- OSPF的区域
- 主干路由器与区域边界路由器
  - 主干路由器：含0区的路由器 (R3/4/5/6/7)
  - 区域边界路由器：含0区和其它区的路由器 (R3/4/7)



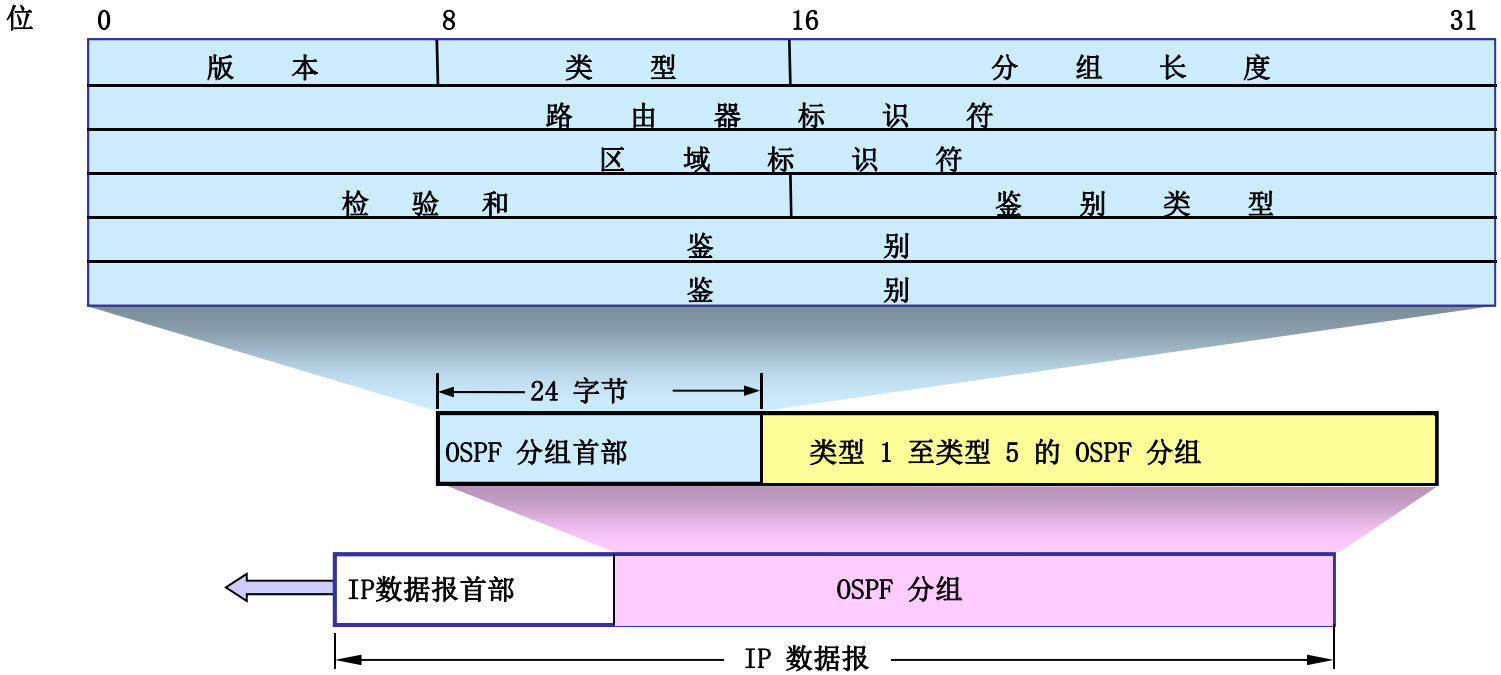
§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

- ● OSPF的报文类型
  - ◆ 类型1：问候(Hello)分组
  - ◆ 类型2：数据库描述(Database Description)分组
  - ◆ 类型3：链路状态请求(Link State Request)分组
  - ◆ 类型4：链路状态更新(Link State Update)分组(泛洪法对全区域更新链路状态)
  - ◆ 类型5：链路状态确认(Link State Acknowledgment)分组



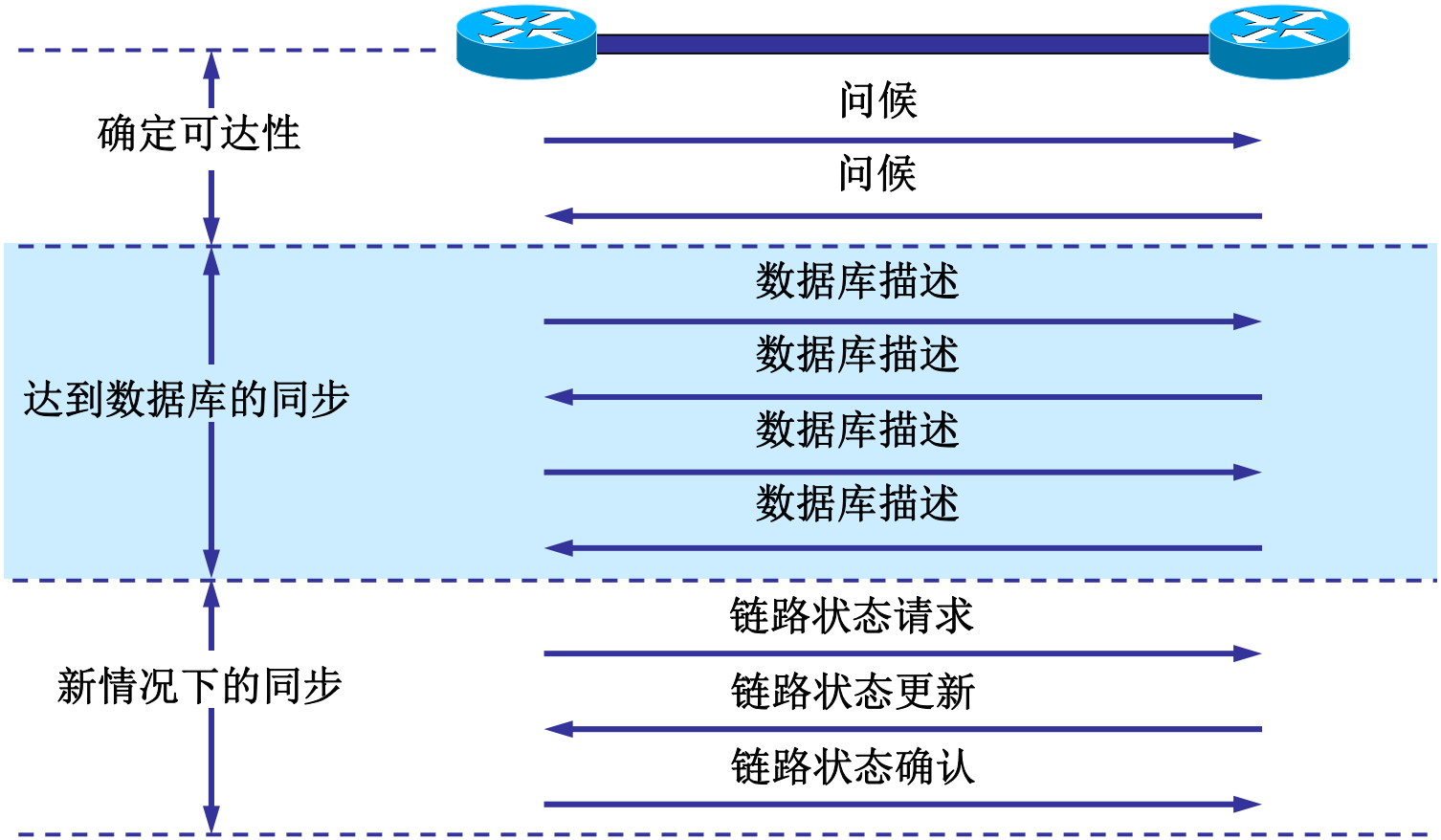
§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

● OSPF的基本流程



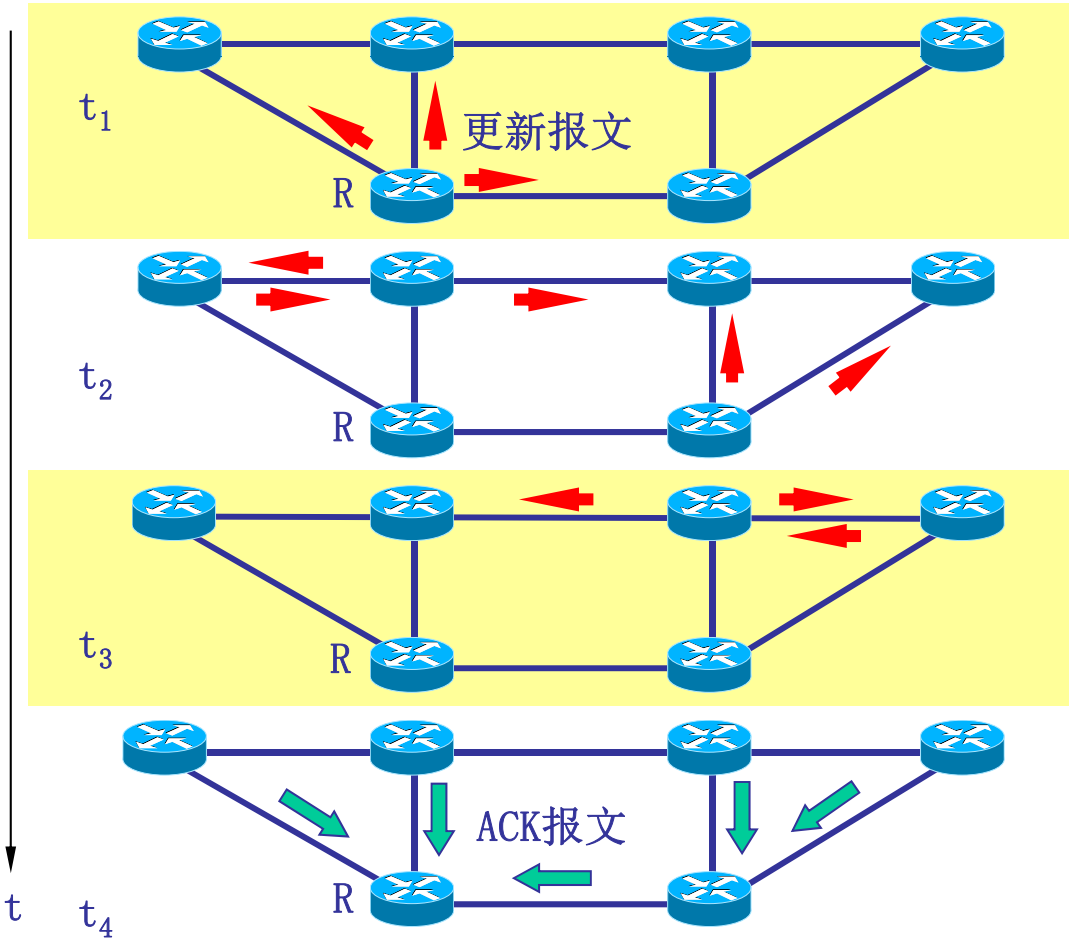
§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

● OSPF的可靠泛洪法（收到的路由器发ack报文）



§5. 网络层

5.2. 路由算法

5.2.5. 链路状态路由

★ OSPF（Open Shortest Path First）

- ● 指定路由器（designated router）
  - ◆ OSPF在多点接入的局域网采用了指定路由器的方法，使广播的信息量大大减少
  - ◆ 指定的路由器代表该局域网上所有的链路向连接到该网络上的各路由器发送状态信息

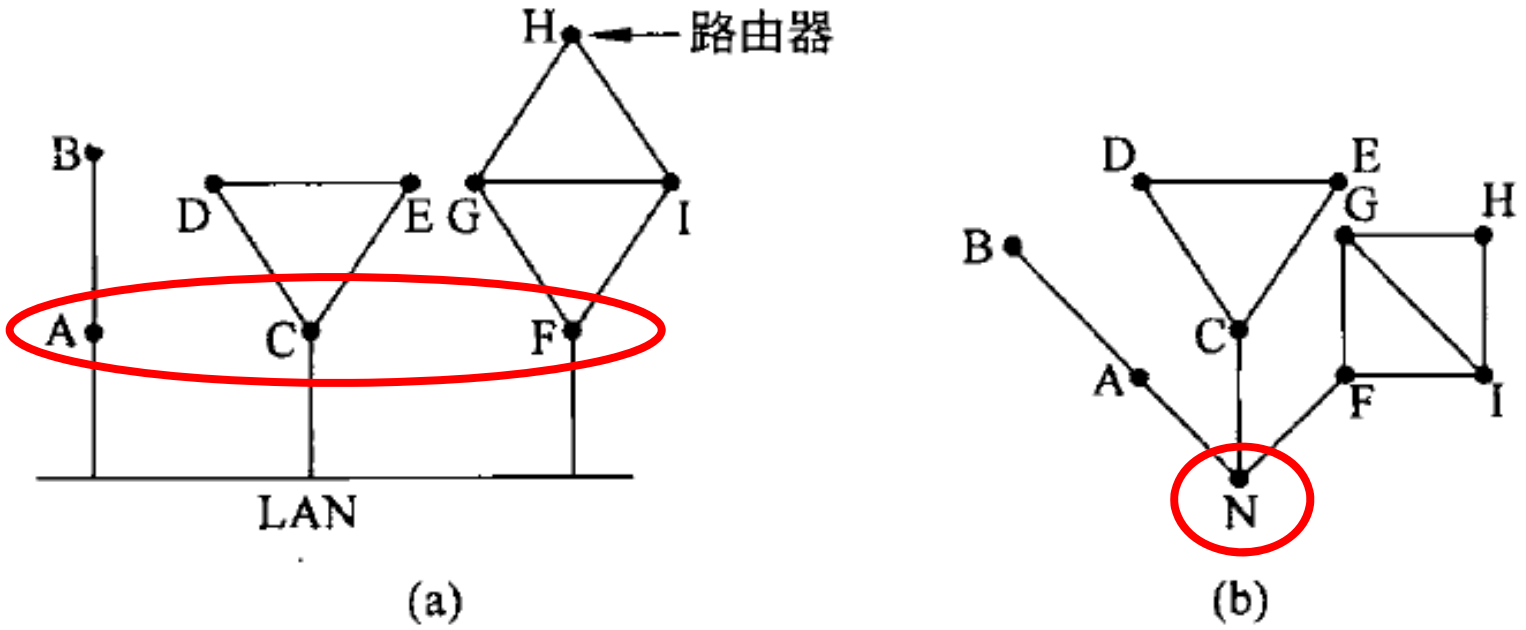


图 5-11

（a）9 个路由器和一个广播 LAN；（b）左侧网络的模型图

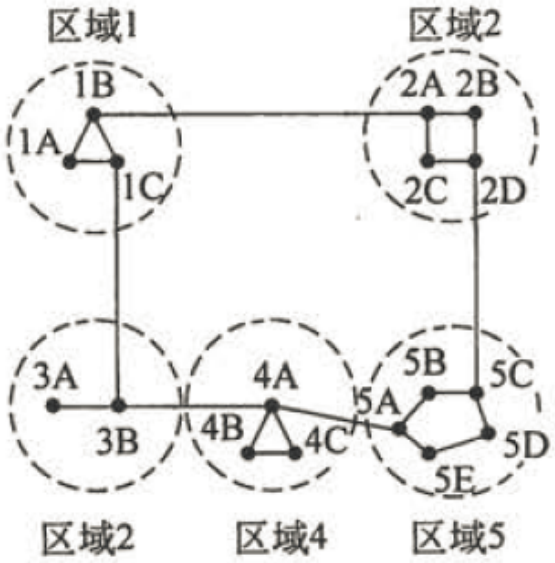
§5. 网络层

5.2. 路由算法

5.2.6. 层次路由

★ 引入

在网络规模较大时，采用分级路由的方式来简化路由表的查找与转发



17台路由器非层次路由
1A路由器中17个表项

1A的完整表

目标地址	线路	跳数
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

17台路由器层次路由
1A路由器中7个表项

1A的层次表

目标地址	线路	跳数
1A	—	—
1B	1B	1
1C	1C	1
2	1B	2
3	1C	2
4	1C	3
5	1C	4

★ 分层定理：

N台路由器的子网最优级数为lnN，每台路由器的总表项为elnN

§5.网络层

5.2.路由算法

5.2.7.广播路由

★ 引入：希望给所有的机器发送相同的信息

★ 实现方法：

- N次单播：给每台机器单独发送数据包，重复N次
- 多目标路由：每个数据包含有一组目标地址，路由器根据目的地址清单确定输出线路集合，为每条输出线路生成一个数据包的副本
- 泛洪法：路由器向除输入线路外的其它端口复制并发送数据包（通过序号控制二次扩散）
- 沿生成树发送：在网络中构造一颗(最小代价)生成树，路由器只向属于生成树的线路发送

★ IP广播地址：255.255.255.255

★ 以太网广播地址：FF:FF:FF:FF:FF:FF

§5. 网络层

5.2. 路由算法

5.2.8. 组播路由

★ 引入：希望给部分的机器发送相同的信息

★ 实现：

- 通过某种方式创建/删除组信息
- 每个组有特定的组播地址标识
- 用类似广播的方法发送组播信息

★ IP组播地址：

- D类地址为组播地址(224.0.0.0-239.255.255.255)
- 常见组播地址

  224.0.0.5 ：OSPF

  224.0.0.9 ：RIPv2

  224.0.0.12：DHCP

★ 以太网组播地址：许多MAC组播地址由IP组播地址转换而来

例：OSPF => 224.0.0.5, 则对应的MAC组播地址：

(1) 224.0.0.5 = 11100000: 00000000: 00000000: 00000101

(2) IP地址的后23bit为：　　0000000: 00000000: 00000101

(3) IEEE定义的MAC组播前24bit为 01:00:5e

(4) 组播MAC = MAC前24bit + 0 + IP后23bit　=> 01:00:5e:00:00:05

以太网上发送的OSPF广播报文的目标MAC地址

§5.网络层

5.2.路由算法

5.2.9.选播路由

★ 基本概念：点到多点的通信，数据包被输送到离源主机"最近"的目的主机处

★ 实现：现有的距离矢量/链路状态路由算法可生成

§5. 网络层

5.2. 路由算法

5.2.10. 移动主机路由

★ 基本概念

- 区域：对整个网络的划分，每个区域可以是一个LAN或一个无线单元
- 移动主机：在不同区域间进行移动的主机
- 移动主机归属：每台移动主机归属的区域
- 外地代理：对进入本区域的、来自于其它主区域的移动主机进行管理的主机
- 主代理：对归属于本区域，但目前不在本区域的移动主机进行管理的主机

★ 登录外地代理的过程

- 外地代理定期广播自己的存在，或移动主机广播寻找外地代理
- 移动主机登录到外地代理
- 外地代理通知移动主机的主代理
- 主代理审核安全信息
- 外地代理获知确认

★ 通信过程

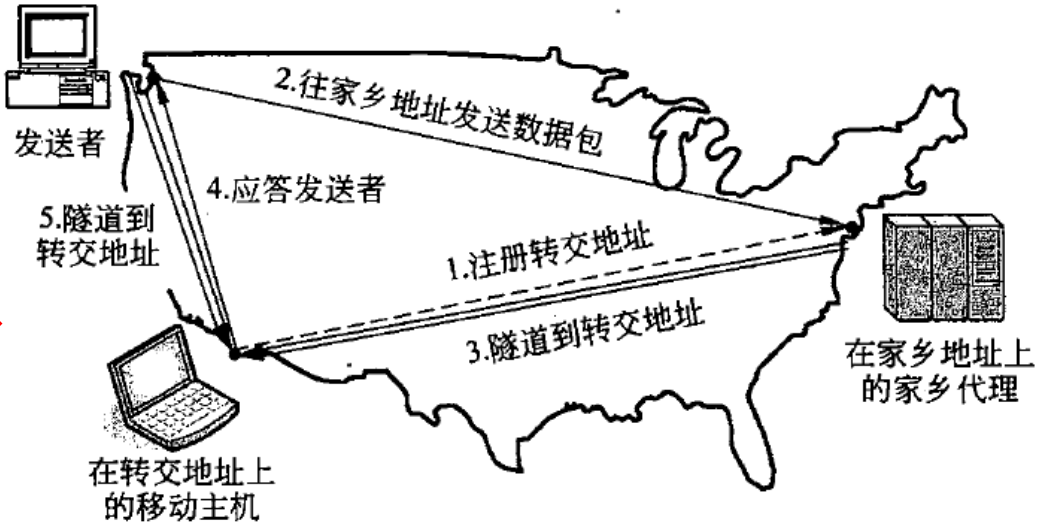


图 5-19　移动主机的数据包路由过程

§5. 网络层

5.2. 路由算法

5.2.11. 自组织网络路由(略)

§5. 网络层

5.3. 拥塞控制算法(略)

§5. 网络层

5.4. 服务质量(略)

§5. 网络层

5.5. 网络互联

5.5.1. 网络如何不同

项目	某些可能性
提供的服务	无连接与面向连接
寻址	不同大小，扁平或层次
广播	提供或者缺乏(组播同样)
数据包尺寸	每个网络有自己的最大尺寸
有序性	有序和无序传递
服务质量	提供或缺乏；许多不同种类
可靠性	丢包的不同级别
安全性	隐私规则，加密等
参数	不同超时值，流规范等
记账	按连接时间、包数、字节数或不收费

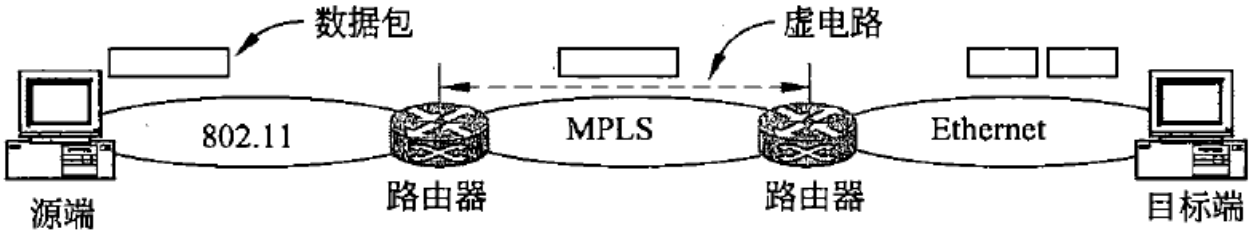
图 5-38　网络的某些不同之处

§5. 网络层

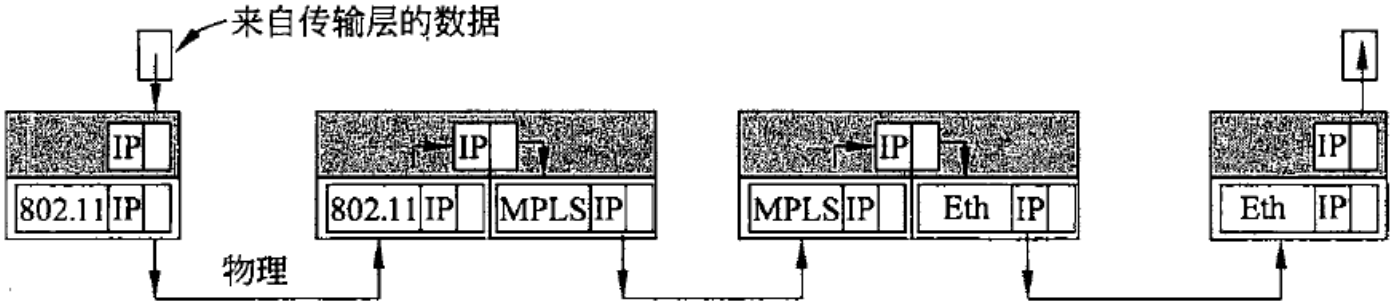
5.5. 网络互联

5.5.2. 何以连接网络

★ 构造公共的网络层在不同的物理链路层上传输



(a)



(b)

图 5-39

（a）跨越不同网络的数据包；（b）网络层和链路层的协议处理

§5. 网络层

5.5. 网络互联

5.5.3. 隧道

★ 含义：在同一层次上封装异种数据包的协议

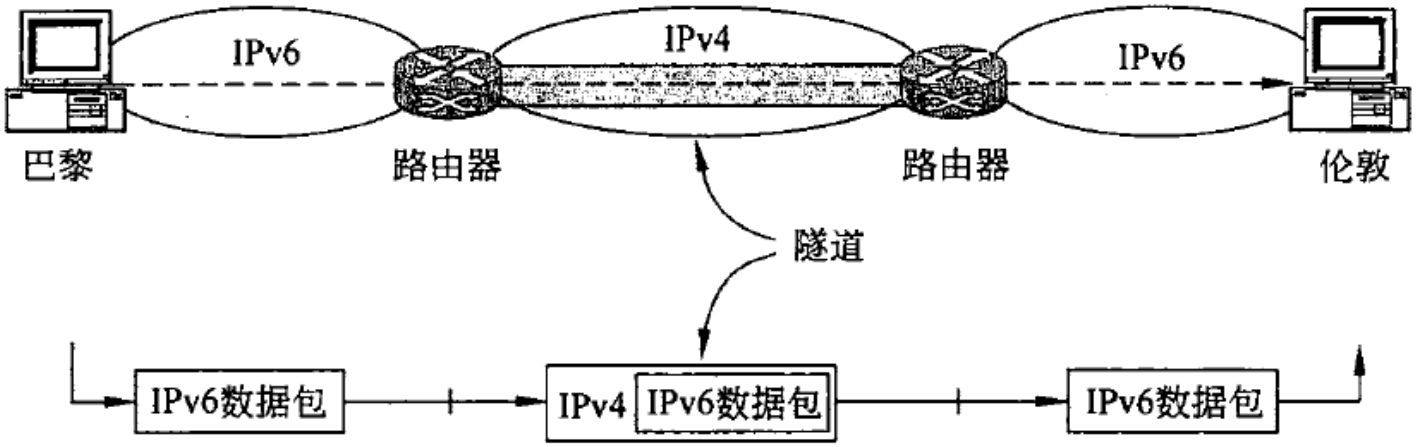


图 5-40　从巴黎隧道一个数据包到伦敦

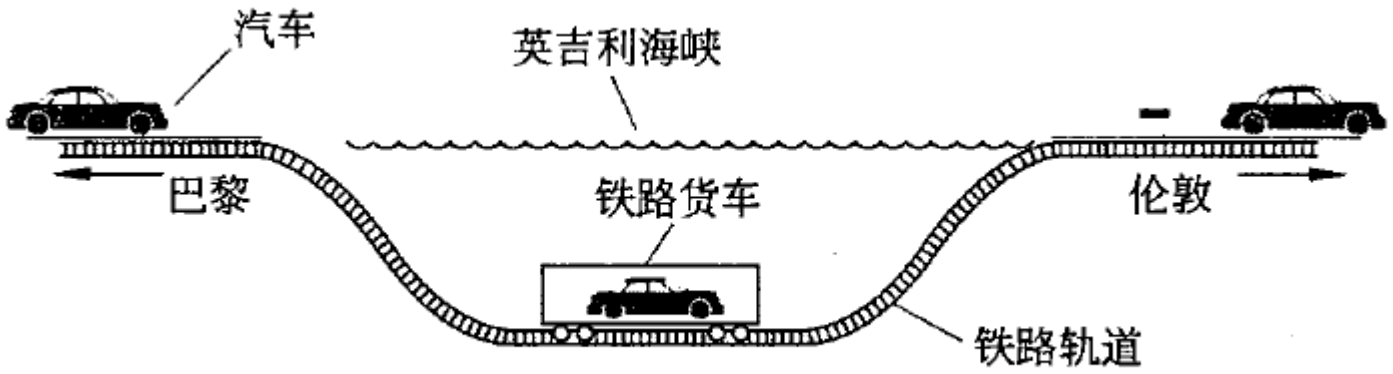


图 5-41　通过隧道把一辆车从法国运到英国

§5.网络层

5.5.网络互联

5.5.3.隧道

★ 含义：在同一层次上封装异种数据包的协议

★ Windows下的IPv6隧道设置

- 管理员方式运行cmd
- 禁止IPv6隧道

```
netsh interface teredo set state disabled
netsh interface 6to4 set state disabled
netsh interface isatap set state disabled
```

- 允许IPv6隧道

```
netsh interface teredo set state client teredo.ipv6.microsoft.com 60 34567
netsh interface 6to4 set state enable
netsh interface isatap set state enable
```

★ VPN（Virtual Private Networks）

采用隧道技术可以构造虚拟专用网络VPN，即建立一个独立于网络物理拓扑结构的逻辑网络，它允许地理位置上分布的一组主机互相交互并且可以作为一个单独的网络进行管理，不用关心主机在网络中所处的位置

- 必须具备的功能：加密、认证和保证数据完整性

§5. 网络层

5.5. 网络互联

5.5.4. 互联网路由（前面介绍过）

- 自治系统AS（Autonomous System）
  - AS是指在单一的技术管理下的一组路由器，使用同一AS的内部应使用相同的路由协议
  - 在一个AS内部使用的路由协议称为内部网关协议(IGP = Interior Gateway Protocol)
  - 在不同AS之间使用的路由协议称为外部网关协议(EGP = External Gateway Protocol)

§5. 网络层

5.5. 网络互联

5.5.5. 数据包分段

★ 引入：每种不同的网络，因为物理特性或协议约定，其数据包最大长度(MTU = Maximum Transmission Unit) 是不同的，当在不同网络间传输数据时，就可能会出现无法传输的问题

★ 解决：将大数据包拆分为若干段，每段封装为完整的数据包后分别传送

★ 透明分段与非透明分段

- 透明分段：路由器在入口处重组收到的分段，再根据出口的情况完整转发或重新分段后转发(要求分段序列必须经过同一个路由器)
- 不透明分段：数据包被分段后，再转发过程中不重组，仅在目的主机上重装分组

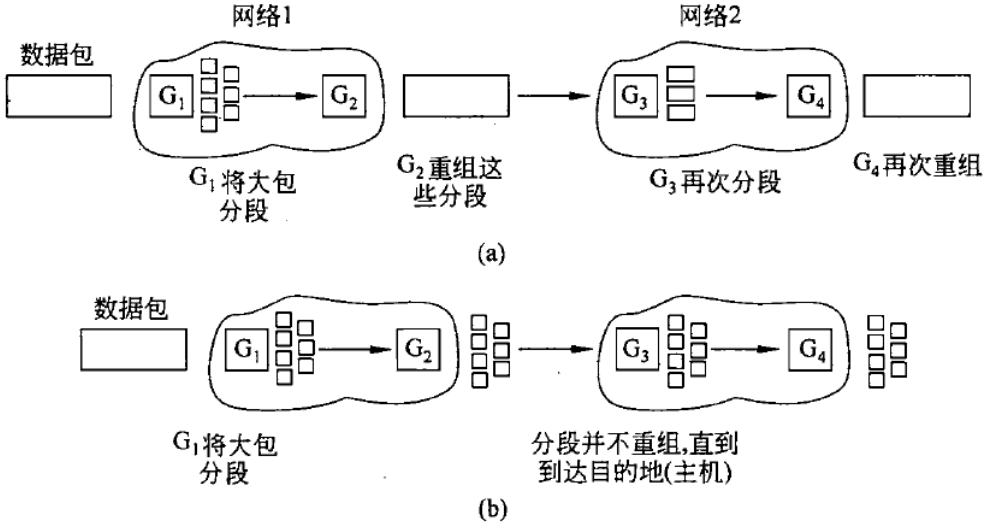


图 5-42

(a) 透明分段；(b) 非透明分段

§5. 网络层

5.6. Internet的网络层

5.6.0. 设计的基本准则

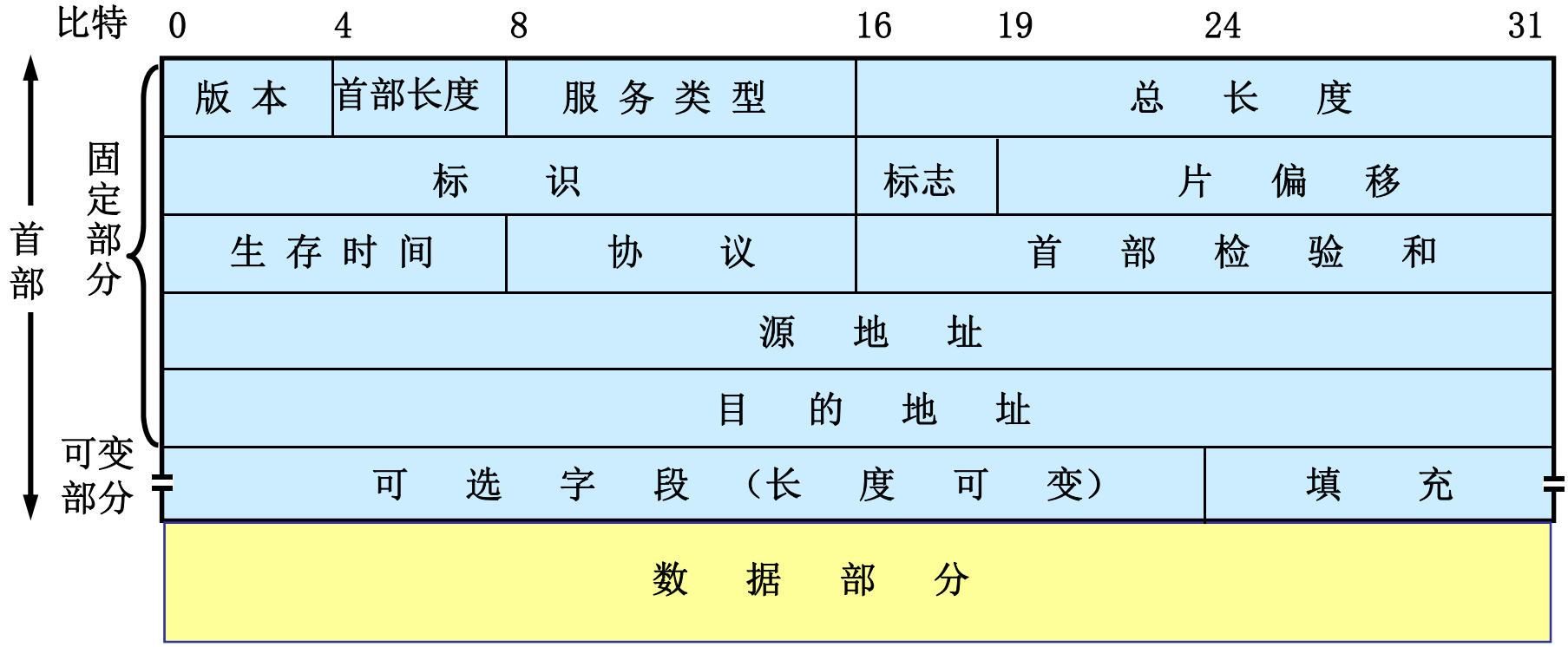
★ 保证工作：验证原型的可行性后再定义规范
★ 保持简单：非必须即取消，可组合即取消
★ 明确选择：不要给出多种方案
★ 模块开发：高内聚、低耦合
★ 期望异构性：考虑异种形式的接入接口
★ 避免静态选项和参数：尽量协商，不要固定
★ 寻找好的而不是完美的设计：单独进行特殊处理
★ 严格发送，宽容接收：保证自己的正确，宽容别人
★ 考虑可扩展性：未来的发展(IPv4网段划分是失败的例子)
★ 考虑性能和成本：快速发展的关键

§5. 网络层

5.6. Internet的网络层

5.6.1. IPv4协议

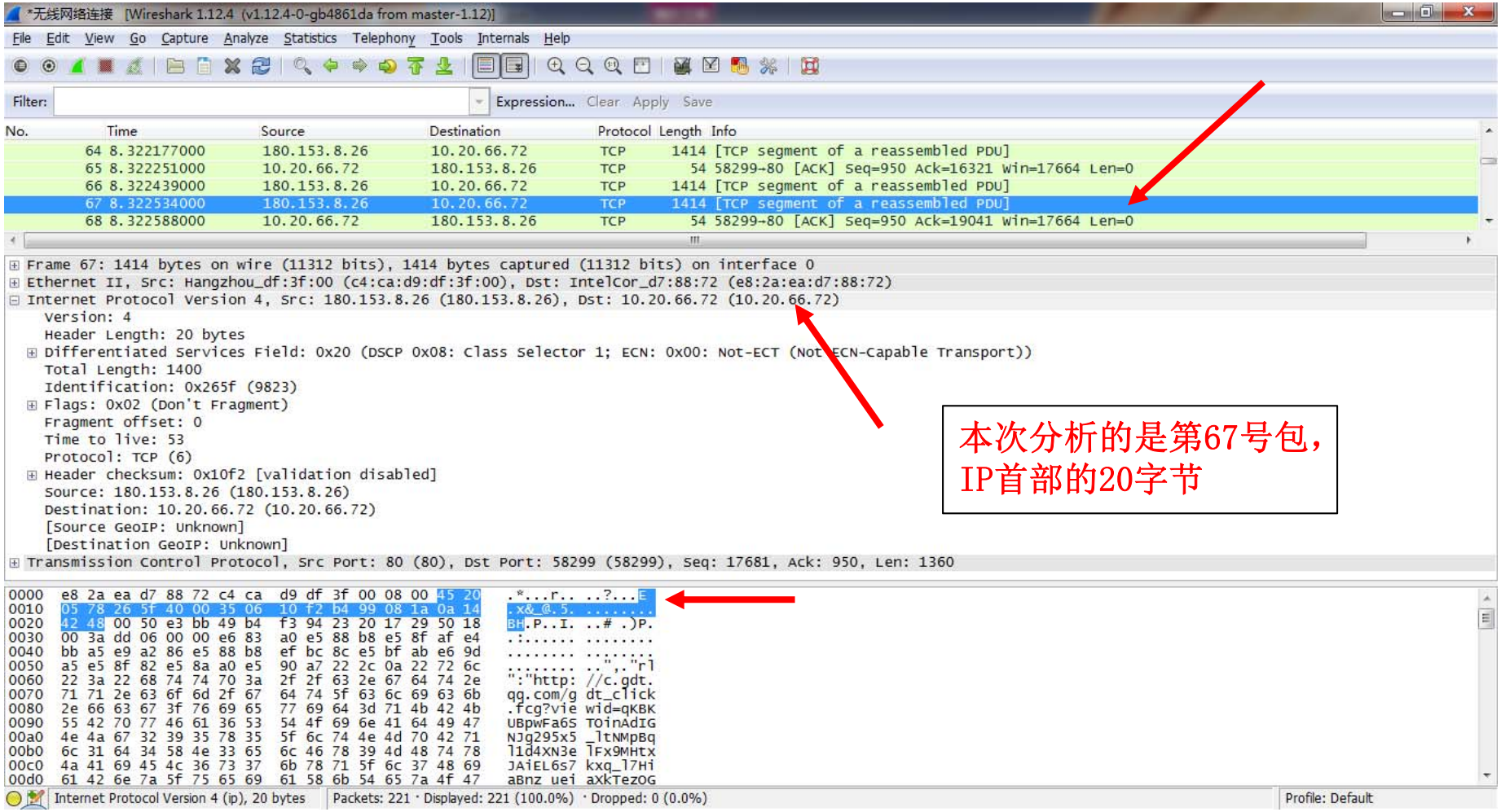
★ IPv4数据包的格式

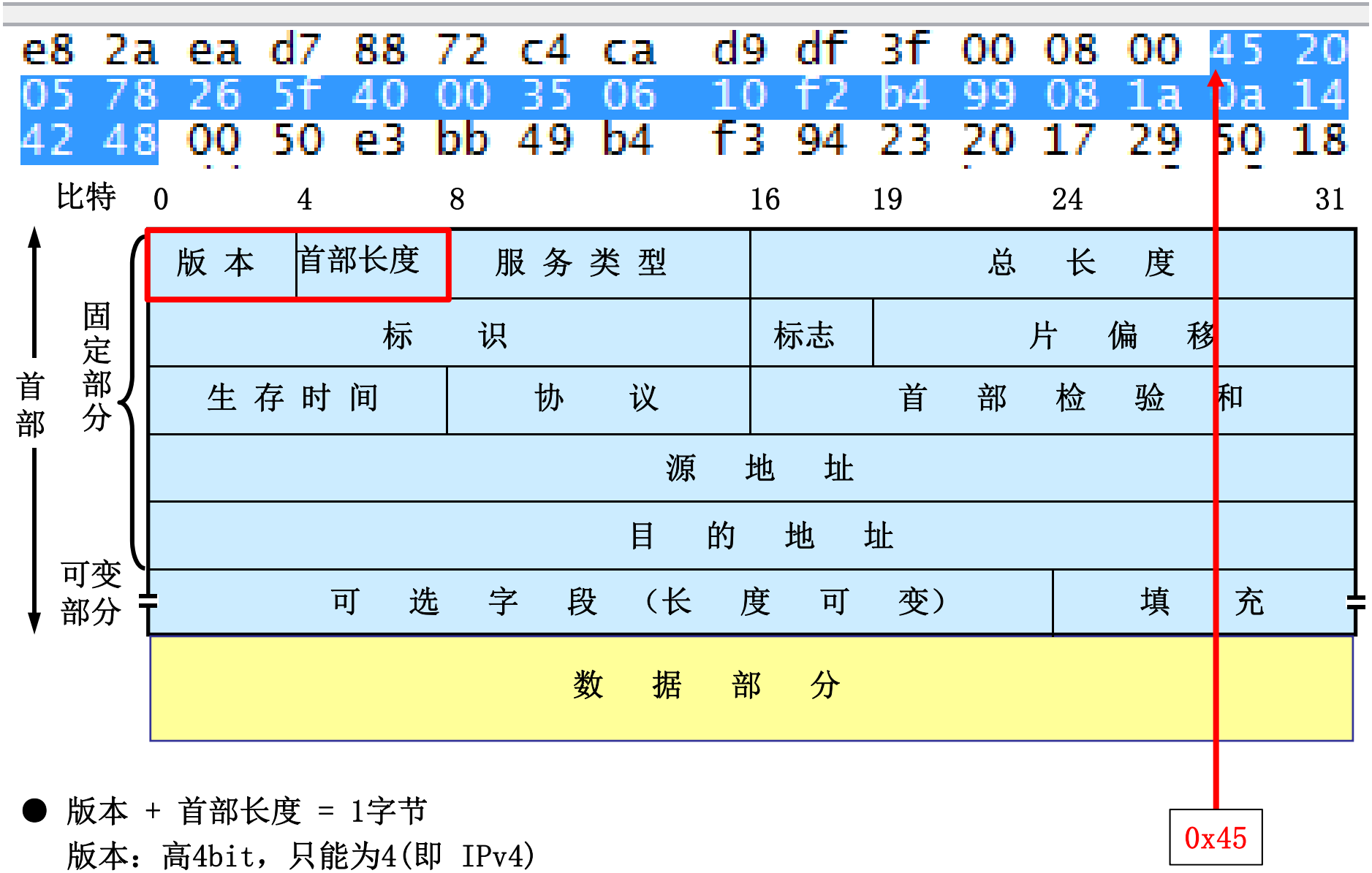


★ 以IPv4数据包的解析为例来说明作业要求

★ 以IPv4数据包的解析为例来说明作业要求

- 首先指出本次分析的是哪个pcap文件、第几个包、该数据包的哪部分





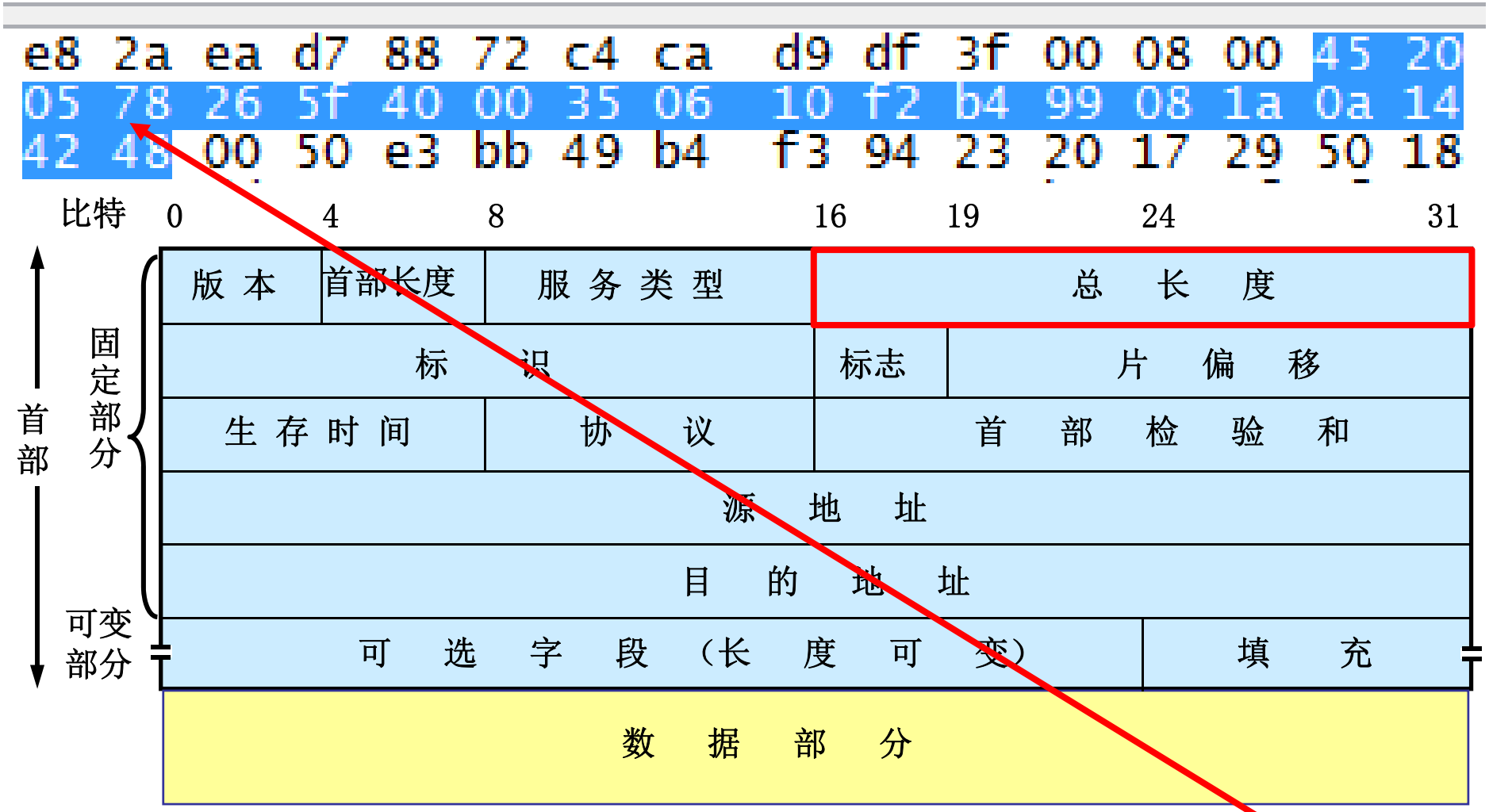
- 版本 + 首部长度 = 1字节

  版本：高4bit，只能为4(即 IPv4)

  首部长度：低4bit，取值5-15，单位4字节

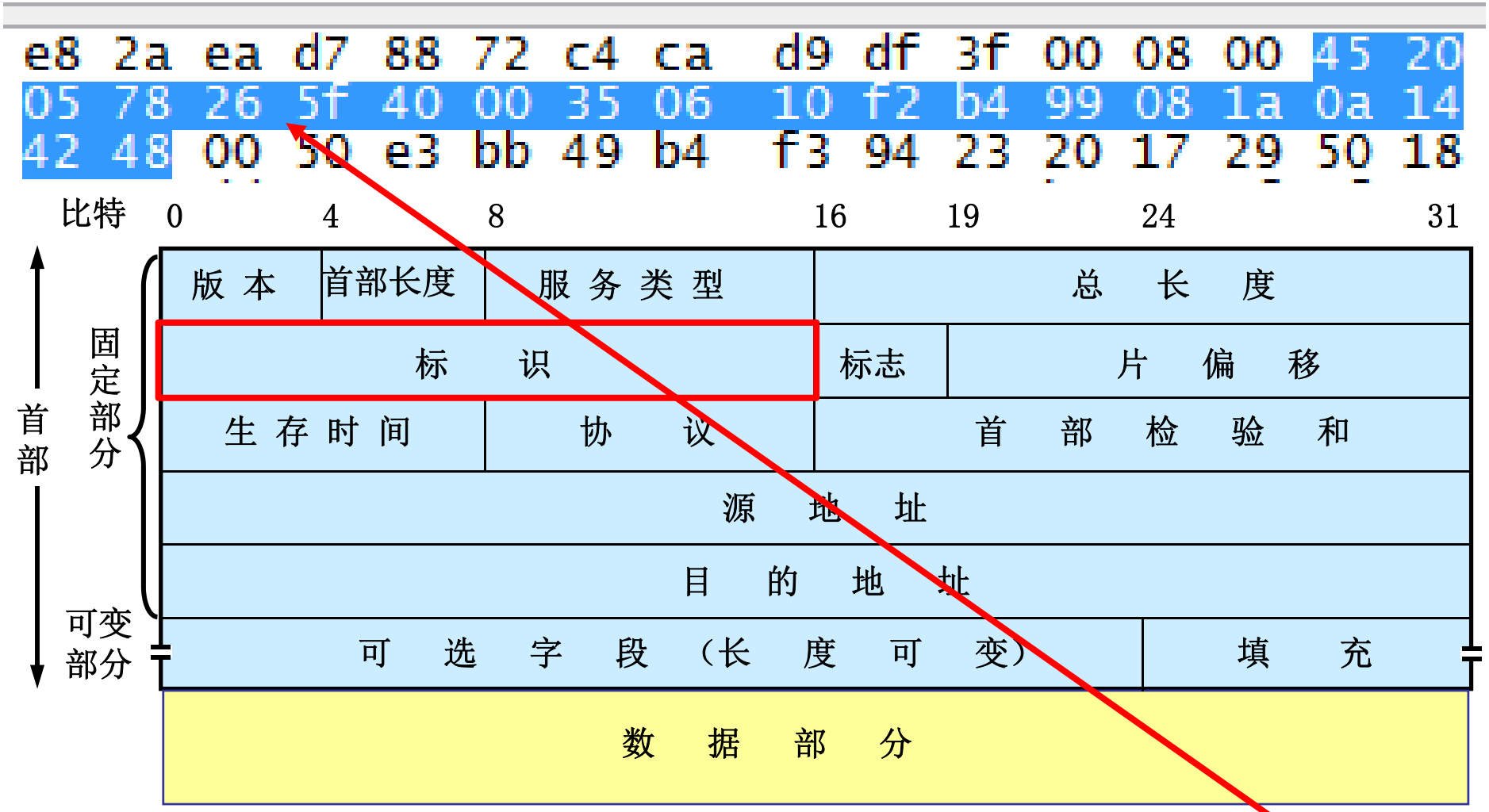
  （固定部分20字节，可变部分0-40字节）

- 服务类型 = 1字节，为了用来获得更好的服务。
  本例中：优先级为1，D/T/R/C均未置位



0x05 78

- 总长度 = 2字节，指首部和数据的总字节长度
  => 总长不可能超过65535
  => 要求不能超过MTU大小
  本例中：0x0578 = 1400 = 20(头) + 1380(数据)
- 关联知识验证：在Wireshark中查看本数据帧对应的以太帧长、TCP长

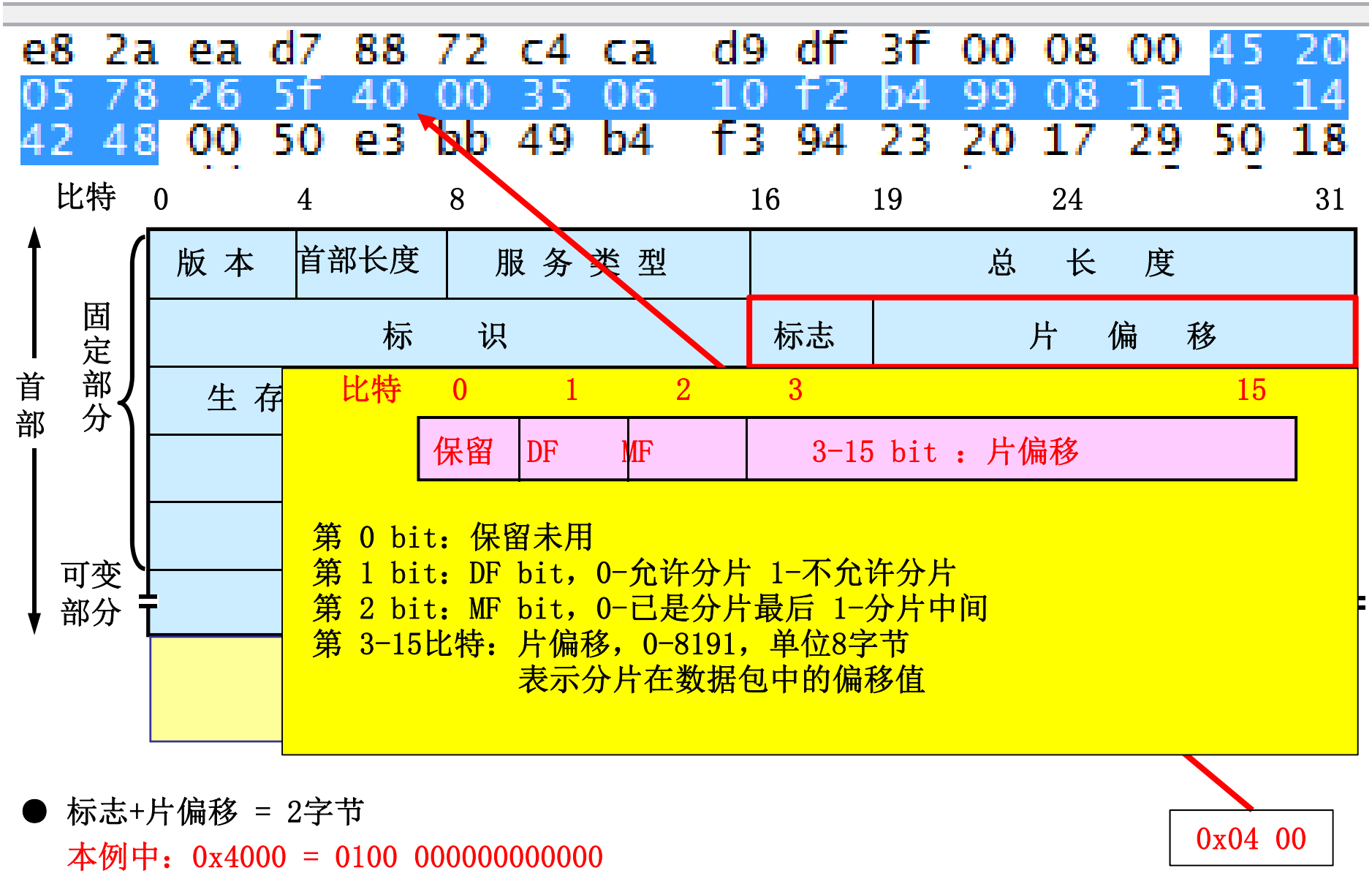


0x26 5f

- 标识 = 2字节 ：计数器，用来产生数据报的标识

  本例中：0x265f = 9823

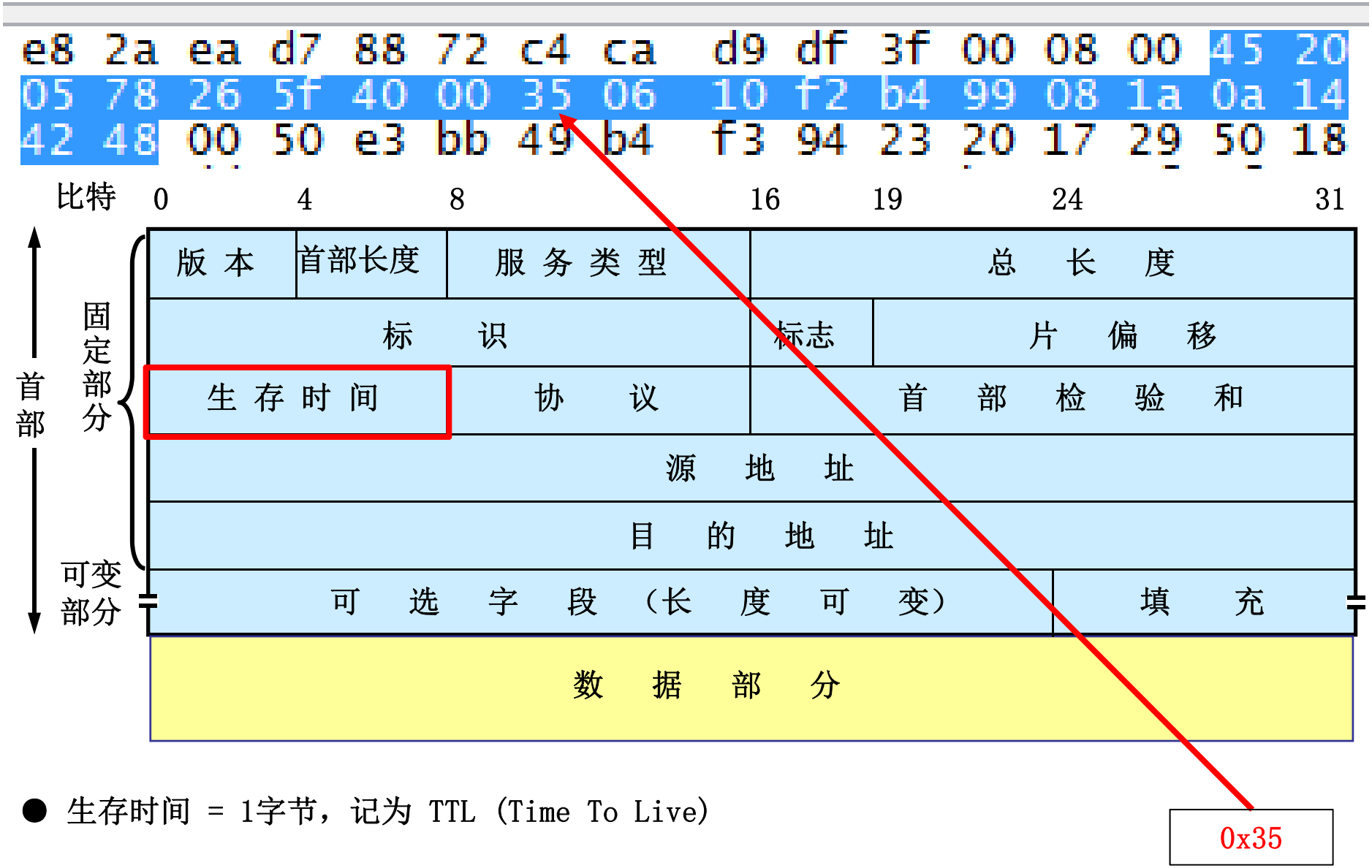
- 关联知识验证：在Wireshark中查看相同源/目的IP地址的前后帧的序号



- 标志+片偏移 = 2字节

  本例中：0x4000 = 0100 000000000000

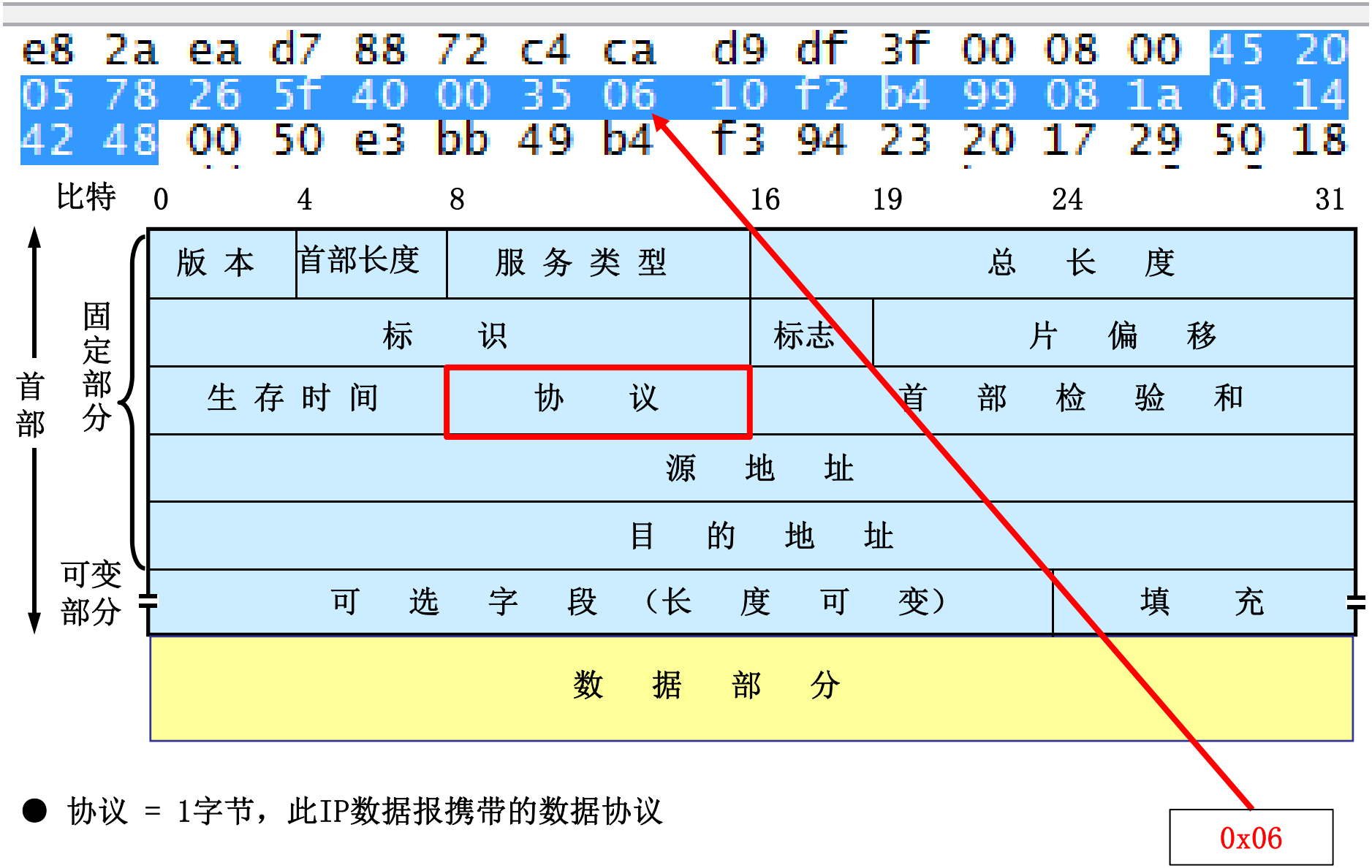
  不允许分片，已是分片最后，偏移0 => 无分片



● 生存时间 = 1字节，记为 TTL (Time To Live)

数据报在网络中的寿命，转发一次减一，到0则丢弃

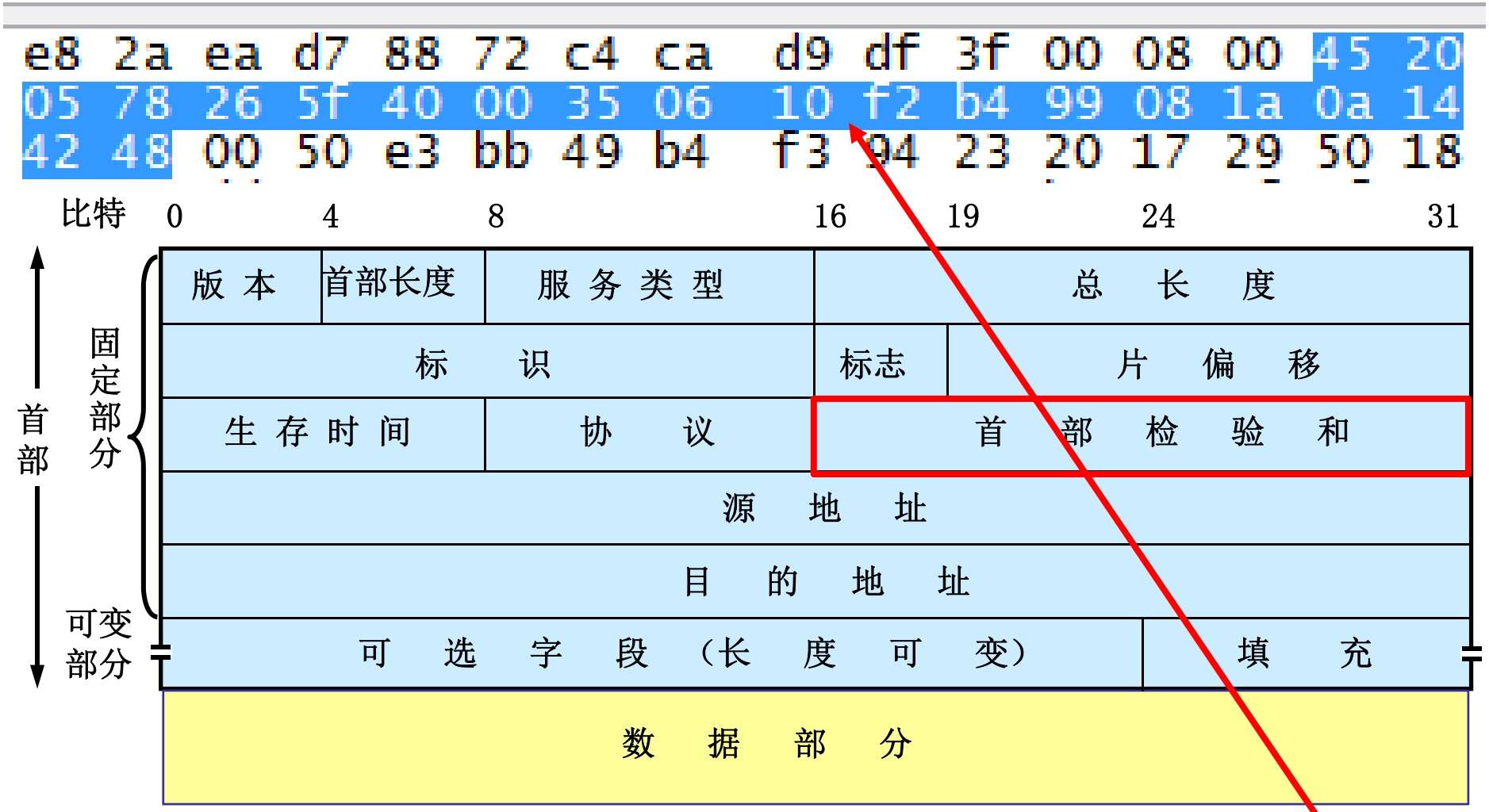
本例中：0x35 = 53



● 协议 = 1字节，此IP数据报携带的数据协议

TCP(6)/UDP(17)/ICMP(1)/OSPF(89)/...

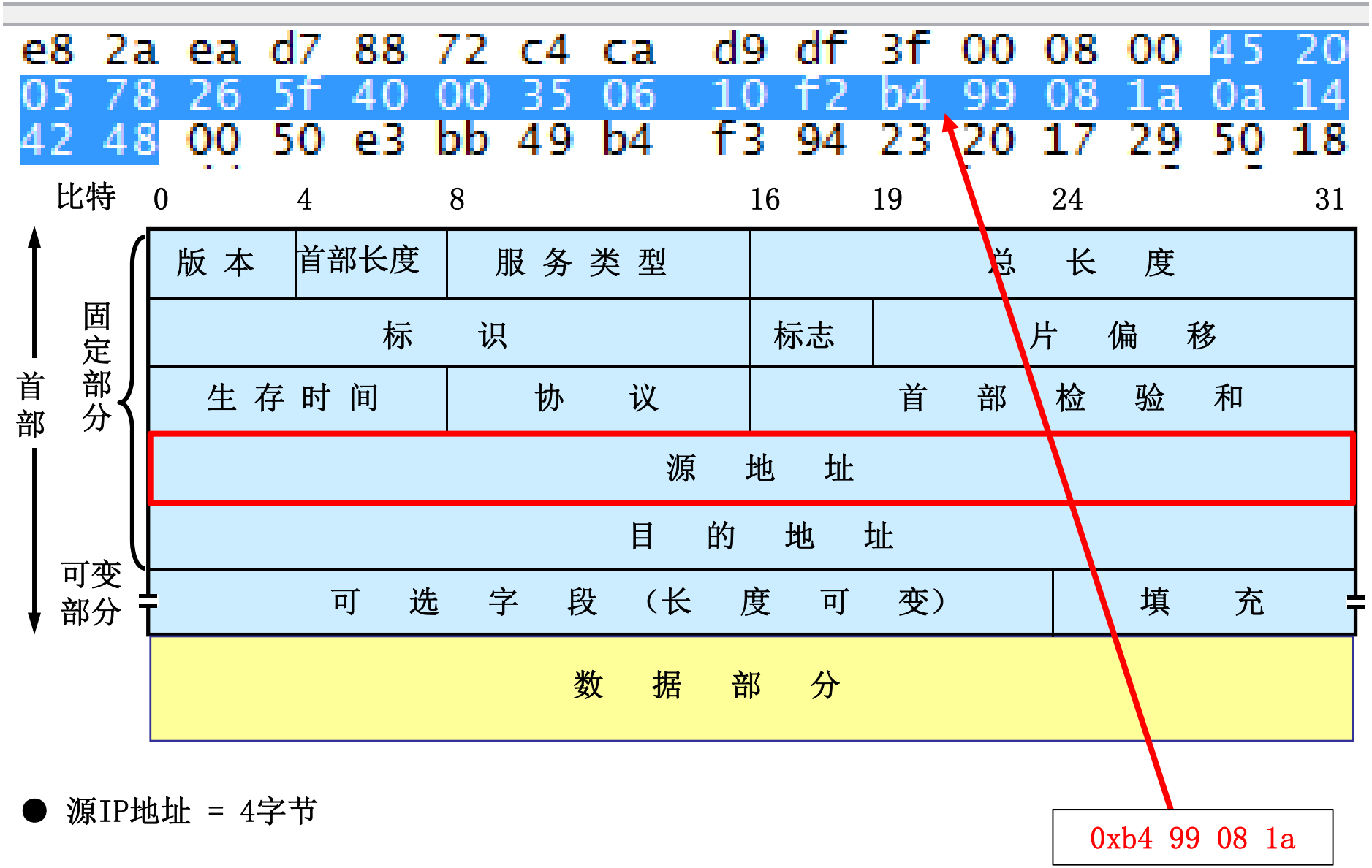
本例中：0x06 = TCP



- 首部检验和 = 2字节，数据报首部检验码(不含数据)

  具体校验方式：...

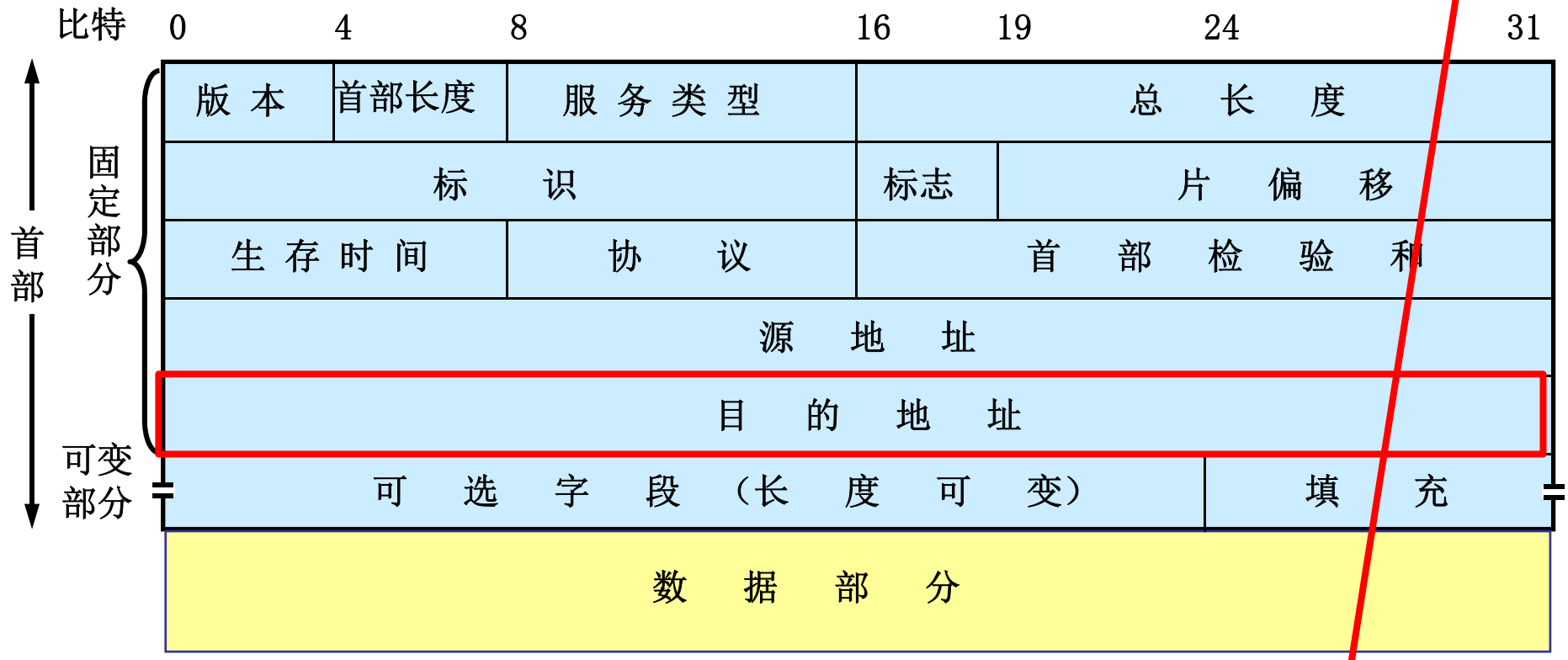
  本例中：0x10f2



● 源IP地址 = 4字节

本例中：0xb499081a = 180.153.8.26

```
e8 2a ea d7 88 72 c4 ca   d9 df 3f 00 08 00 45 20
05 78 26 5f 40 00 35 06   10 f2 b4 99 08 1a 0a 14
42 48 00 50 e3 bb 49 b4   f3 94 23 20 17 29 50 18
```

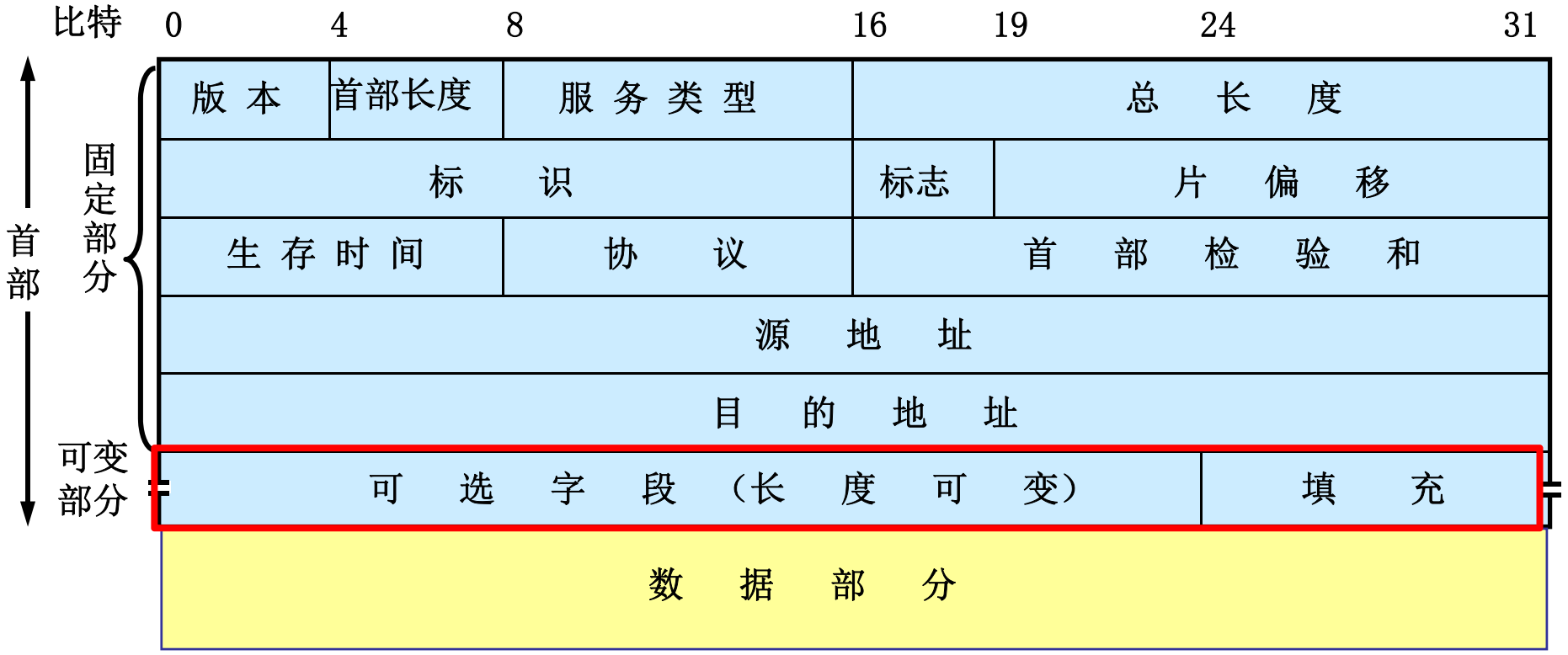


- 源IP地址 = 4字节

0x0a 14 42 48

本例中：0x0a144248 = 10.20.66.72

★ IPv4数据包的格式



● 可变部分：基本不用，某些低端路由器不支持(略)

§5. 网络层

5.6. Internet的网络层

5.6.2. IP地址

★ IP地址的划分、子网、变长子网掩码等（已讲）

★ NAT（讲课作业）

5.6.3. IPv6协议（讲课作业）

5.6.4. Internet控制协议

★ ICMP （ping包的格式）

5.6.5. 标签交换和MPLS（略）

5.6.6. OSPF - 内部网关路由协议（略，概念已介绍）

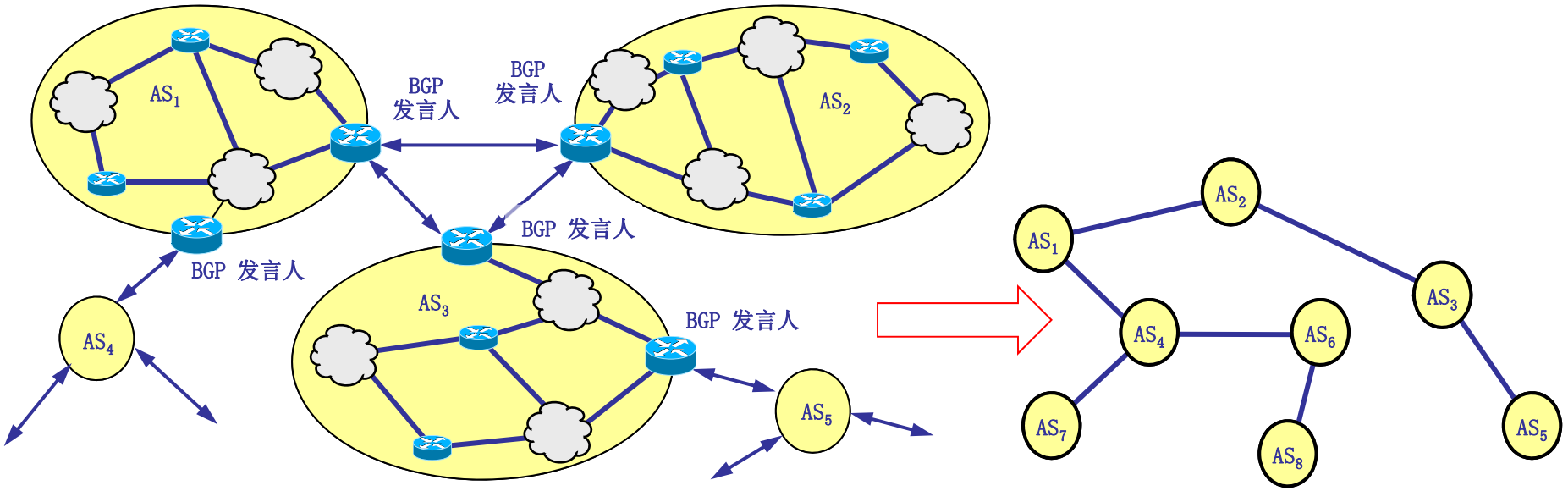
§5. 网络层

5.6. Internet的网络层

5.6.7. BGP – 外部网关路由协议

★ 基本概念

- BGP 是不同自治系统的路由器之间交换路由信息的协议
- BGP 的较新版本是 1995 年发表的 BGP-4（BGP 的第 4 个版本，可直接简写为 BGP）
- Internet 规模太大，使得自治系统之间路由选择非常困难，因此自治系统之间，不追求寻找最佳路由，而是力求寻找一条能够到达目的网络且比较好的路由
- 每一个自治系统的管理员要选择至少一个路由器作为该自治系统的"BGP 发言人"，两个自治系统的 BGP 发言人都是通过共享网络连接在一起，BGP 发言人往往就是 BGP 边界路由器(也可以不是)
- BGP发言人交换网络可达性的信息后，各BGP发言人就可找出到达各自治系统的较好的路由



§5. 网络层

5.6. Internet的网络层

5.6.8. Internet组播（略，概念已介绍）

5.6.9. 移动IP（略，概念已介绍）