

A large, semi-transparent abstract network graph is positioned on the left side of the slide. It consists of numerous small, light-blue circular nodes connected by thin white lines, forming a complex web-like structure that suggests a global or local network.

第4章

网络层

计算机网络体系结构

OSI 的七层协议体系结构



(a)

TCP/IP 的四层协议体系结构



(b)

五层协议的体系结构



(c)

4.1	网络层的几个重要概念
4.2	网际协议 IP
4.3	IP 层转发分组的过程
4.4	网际控制报文协议 ICMP
4.5	IPv6
4.6	互联网的路由选择协议
4.7	IP 多播
4.8	虚拟专用网 VPN 和网络地址转换 NAT
4.9	多协议标记交换 MPLS
4.10	软件定义网络 SDN 简介

4.3

IP 层转发分组的过程

4.3.1

基于终点的转发

4.3.2

最长前缀匹配

4.3.3

使用二叉线索查找转发

主机 H₁

在局域网的链路层，只能看见 MAC 帧。

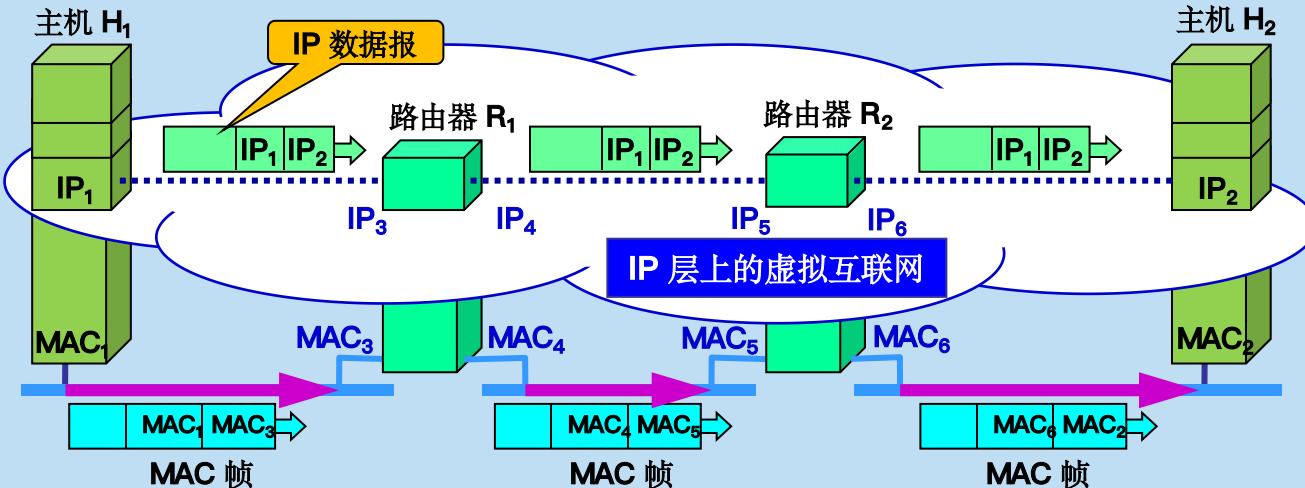
主机 H₂



主机 H₁

IP 数据报

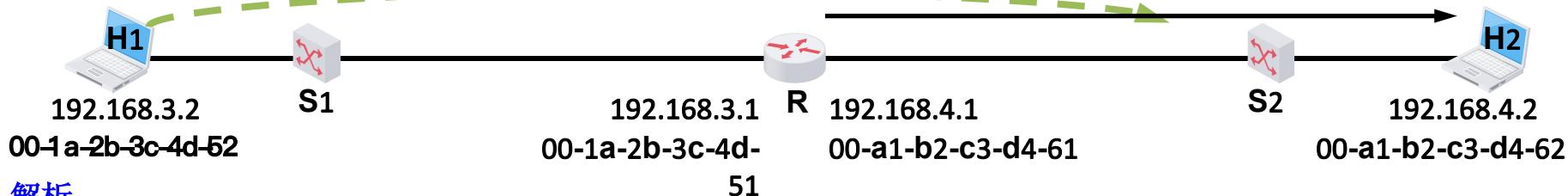
主机 H₂



注意：帧中 MAC 地址是否有变化？对 IP 层有何影响？

【2018年 题37】 路由器R通过以太网交换机S1和S2连接两个网络，R的接口、主机H1和H2的IP地址与MAC地址如下图所示。若H1向H2发送一个IP分组P，则H1发出的封装P的以太网帧的目的MAC地址、H2收到的封装P的以太网帧的源MAC地址分别是（**D**）。

- | | | | |
|----------------------|-------------------|----------------------|-------------------|
| A. 00-a1-b2-c3-d4-62 | 00-1a-2b-3c-4d-52 | B. 00-a1-b2-c3-d4-62 | 00-1a-2b-3c-4d-61 |
| C. 00-1a-2b-3c-4d-51 | 00-1a-2b-3c-4d-52 | D. 00-1a-2b-3c-4d-51 | 00-a1-b2-c3-d4-61 |

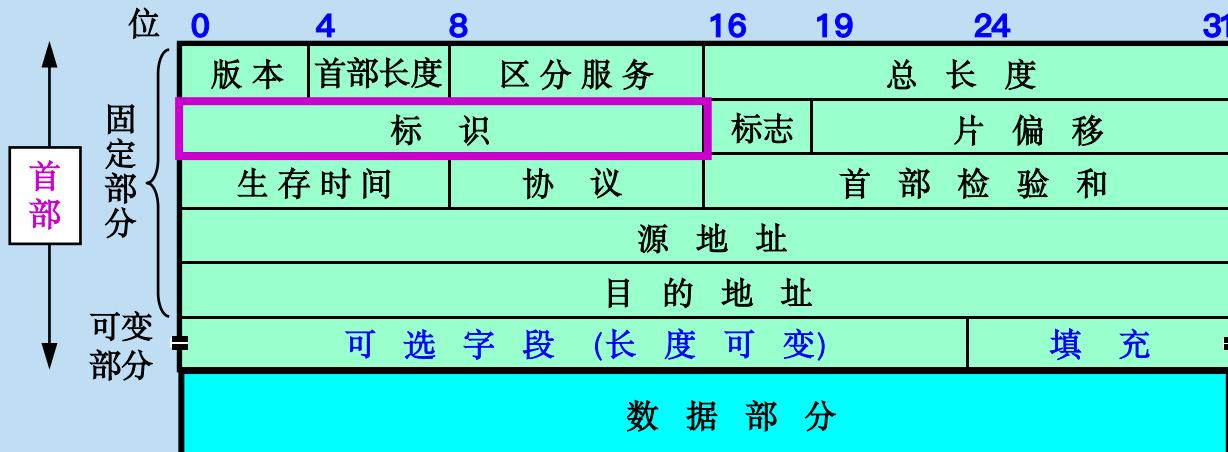


解析

在数据包的传送过程中，**源IP地址和目的IP地址保持不变**，而**源MAC地址和目的MAC地址逐链路（或逐网络）改变**。

数据包传输区间	在网络层写入IP数据报首部的IP地址		在数据链路层写入帧首部的MAC地址	
	源IP地址	目的IP地址	源MAC地址	目的MAC地址
H1→R	192.168.3.2	192.168.4.2	00-1a-2b-3c-4d-52	00-1a-2b-3c-4d-61
R→H2	192.168.3.2	192.168.4.2	00-a1-b2-c3-d4-62	00-a1-b2-c3-d4-61

1. IP 数据报首部的固定部分中的各字段



标识 (identification) —— 占 16 位,
它是一个计数器，用来产生 IP 数据报的标识。

1. IP 数据报首部的固定部分中的各字段



标志(flag)——占 3 位, 目前只有前两位有意义。

- 标志字段的最低位是 **MF** (**M**ore **F**ragment)。
MF=1 表示后面还有分片, **MF=0** 表示最后一个分片。
- 标志字段中间的一位是 **DF** (**D**on't **F**ragment)。
只有当 **DF=0** 时才允许分片。

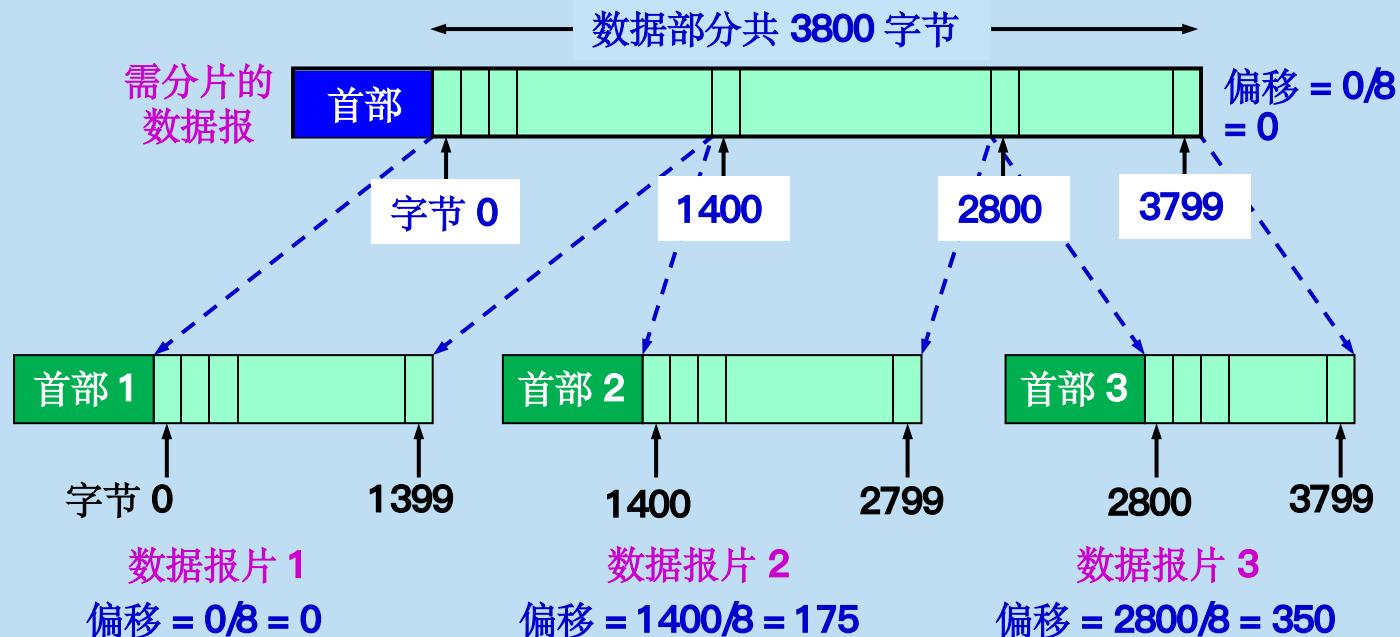
1. IP 数据报首部的固定部分中的各字段



片偏移——占 **13** 位，指出：较长的分组在分片后
某片在原分组中的相对位置。

片偏移以 **8个字节** 为偏移单位。

【例4-1】IP 数据报分片



【例4-1】IP 数据报分片

IP 数据报首部中与分片有关的字段中的数值

	总长度	标识	MF	DF	片偏移
原始数据报	3820	12345	0	0	0
数据报片1	1420	12345	1	0	0
数据报片2	1420	12345	1	0	175
数据报片3	1020	12345	0	0	350

01

IPv4数据报的头部格式

【2021年题36】若路由器向MTU=800B的链路转发一个总长度为1580B的IP数据报（首部长度为20B）时，进行了分片，且每个分片尽可能大，则第2个分片的总长度字段和MF标志位的值分别是（ B ）。

- A. 796, 0
- B. 796, 1
- C. 800, 0
- D. 800, 1

解析

总长度为1580B的IP数据报			
首部20B	0	数据载荷1560B	1559

01

IPv4数据报的头部格式

【2021年题36】若路由器向MTU=800B的链路转发一个总长度为1580B的IP数据报（首部长度为20B）时，进行了分片，且每个分片尽可能大，则第2个分片的总长度字段和MF标志位的值分别是（**B**）。

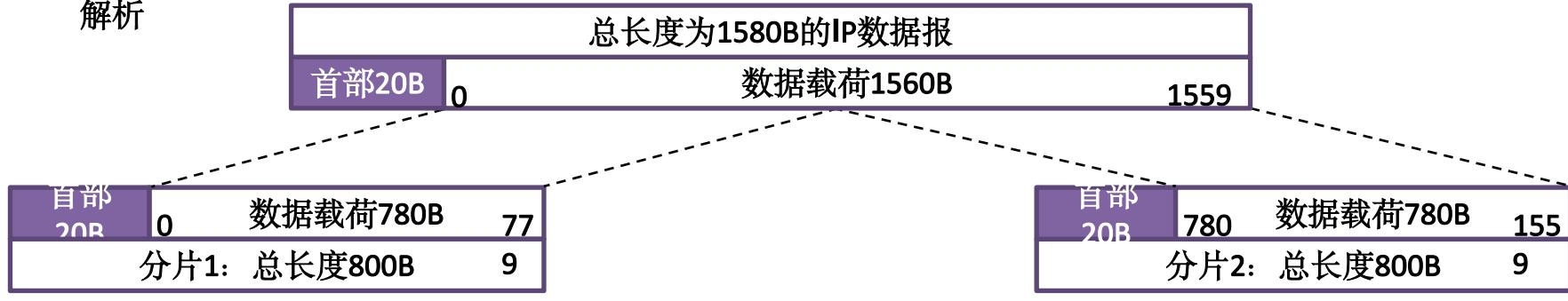
A. 796, 0

B. 796, 1

C. 800, 0

D. 800, 1

解析



$$\text{片偏移} = 0 / 8 = 0$$

片偏移必须为整数，因此这种分配方案不行！

$$\text{片偏移} = 780 / 8 = 97.5$$

分片的数据载荷的最大长度取小于780且能整除8的最大整数

$$\lfloor 780 \div 8 \rfloor \times 8 = 776$$

01

IPv4数据报的头部格式

【2021年题36】若路由器向MTU=800B的链路转发一个总长度为1580B的IP数据报（首部长度为20B）时，进行了分片，且每个分片尽可能大，则第2个分片的总长度字段和MF标志位的值分别是（B）。

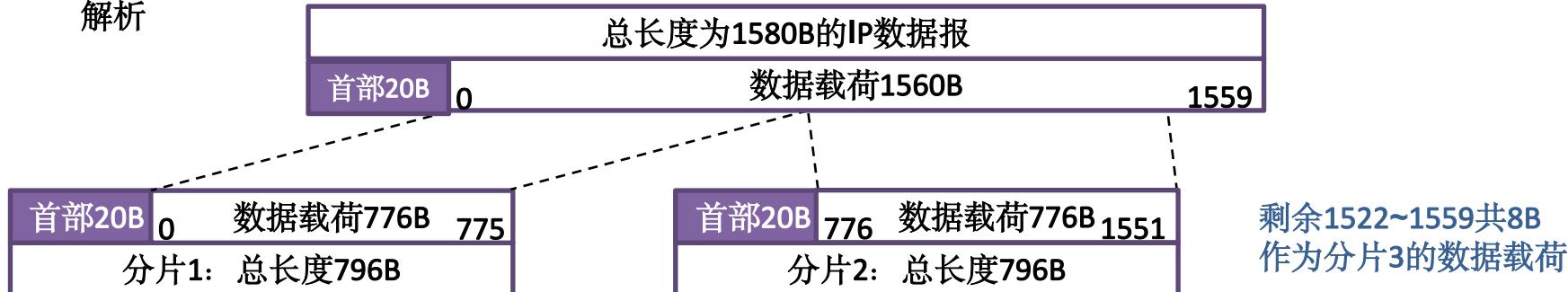
A. 796, 0

B. 796, 1

C. 800, 0

D. 800, 1

解析



$$\text{片偏移} = 0 / 8 = 0$$

$$\text{片偏移} = 776 / 8 = 97$$

$$\text{MF} = 1$$

1. IP 数据报首部的固定部分中的各字段



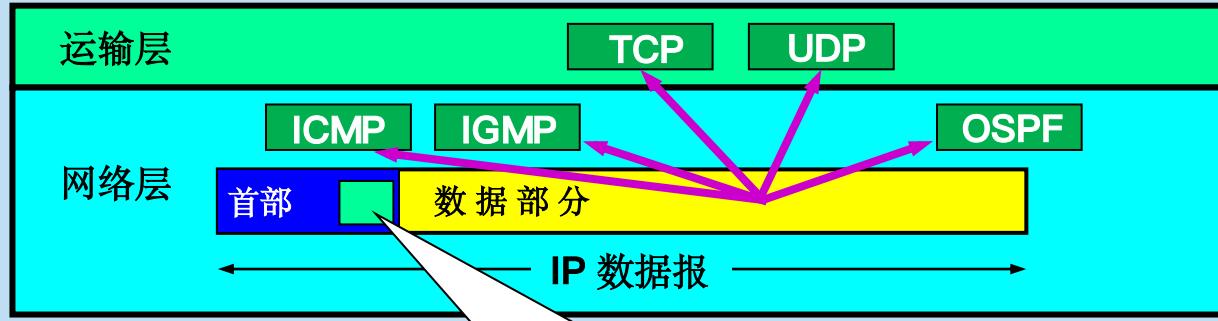
生存时间——占 8 位, 记为 TTL (Time To Live),
指示数据报在网络中可通过的路由器数的最大值。

1. IP 数据报首部的固定部分中的各字段



协议——占 8 位，指出此数据报携带的数据使用何种协议，以便目的主机的 IP 层将数据部分上交给那个处理过程

IP 协议支持多种协议，IP 数据报可以封装多种协议 PDU.



协议字段指出应将数据部分交给哪一个进程

常用的一些协议和相应的协议字段值

协议名	ICMP	IGMP	IP	TCP	EGP	IGP	UDP	IPv6	ESP	AH	ICMP-IPv6	OSPF
协议字段值	1	2	4	6	8	9	17	41	50	51	58	89

1. IP 数据报首部的固定部分中的各字段



首部检验和——占 16 位，只检验数据报的首部，
不检验数据部分。这里不采用 CRC 检验码而采用简单的计算方法。

1. IP 数据报首部的固定部分中的各字段

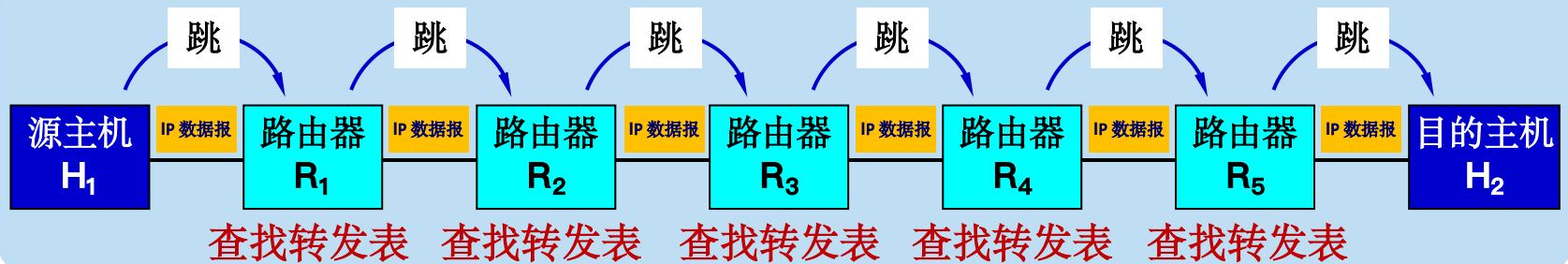


源地址和目的地址都各占 **32** 位。

4.1	网络层的几个重要概念
4.2	网际协议 IP
4.3	IP 层转发分组的过程
4.4	网际控制报文协议 ICMP
4.5	IPv6
4.6	互联网的路由选择协议
4.7	IP 多播
4.8	虚拟专用网 VPN 和网络地址转换 NAT
4.9	多协议标记交换 MPLS
4.10	软件定义网络 SDN 简介

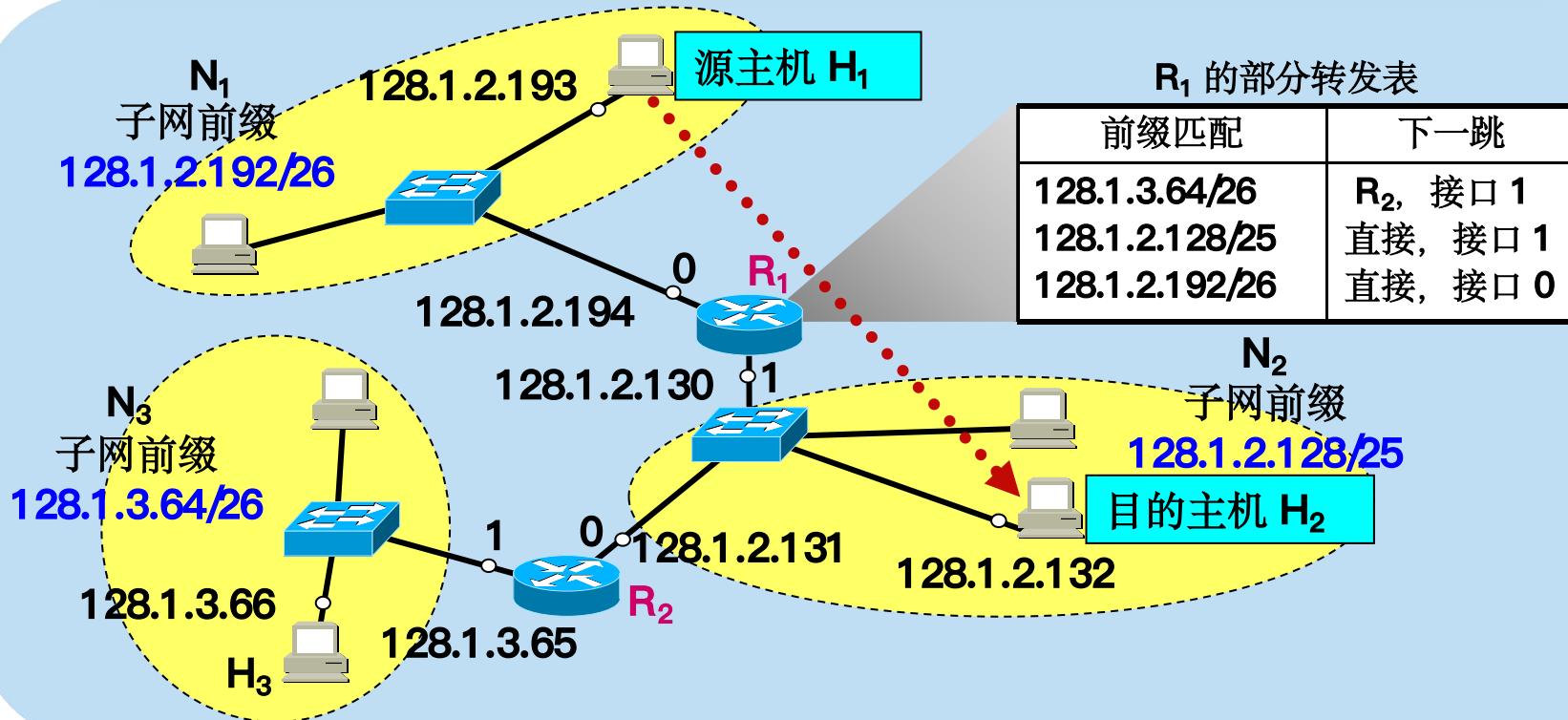
4.3.1 基于终点的转发

- 分组在互联网中是逐跳转发的。
- 基于终点的转发：基于分组首部中的目的地址传送和转发。



为了压缩转发表的大小，
转发表中最主要的路由是（目的网络地址，下一跳地址），
而不是（目的地址，下一跳地址）。
查找转发表的过程就是逐行寻找前缀匹配。





主机 H_1 发送出的、目的地址是 $128.1.2.132$ 的分组是如何转发的?

H_1 首先检查 **128.1.2.132** 是否连接在本网络上。
如果是，则直接交付；否则，就送交路由器 R_1 。

N_1 的网络地址为 **128.1.2.192**

N_1 的网络掩码为 **/26 = 255.255.255.192**

目的地址与网络掩码 **128. 1 . 2 .132**

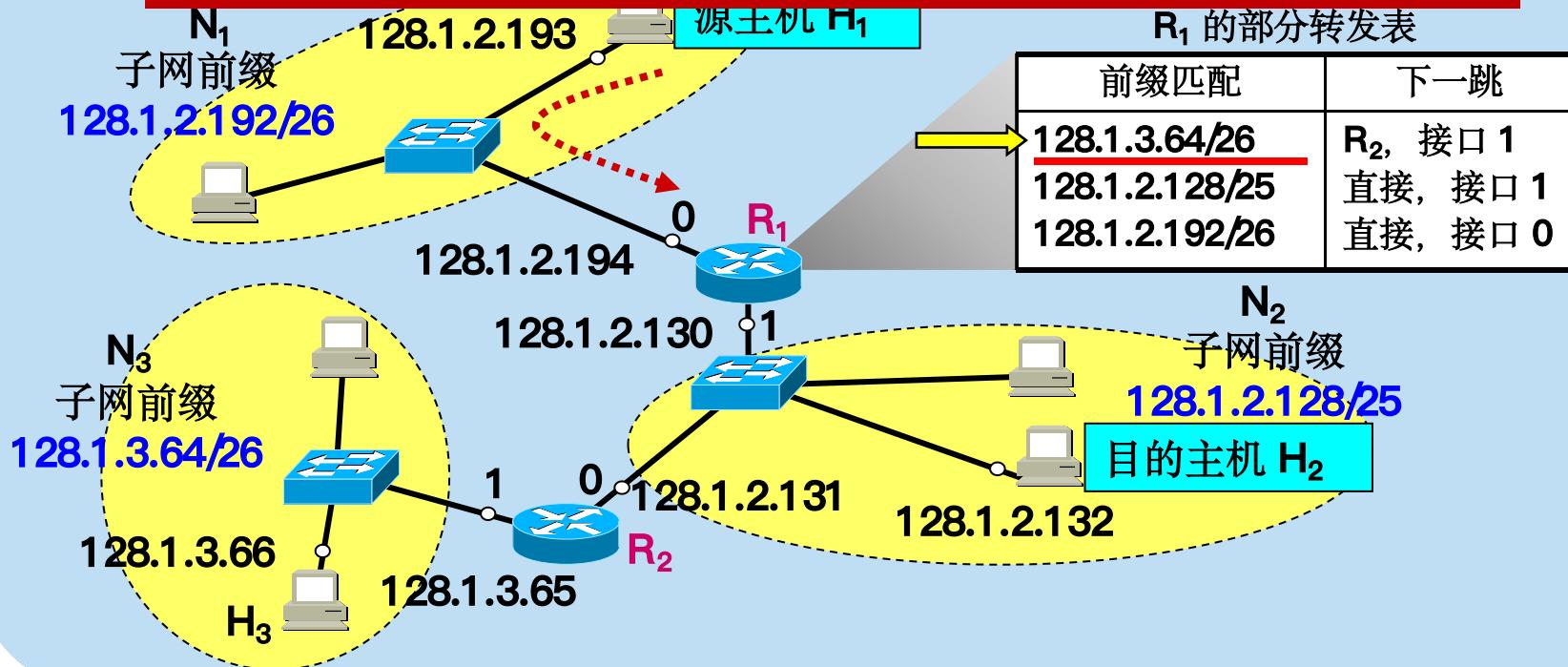
逐比特 AND **255.255.255.192**

128. 1 . 2 .128 $\neq H_1$ 的网络地址

128.1.2.132 不在本地网络上。

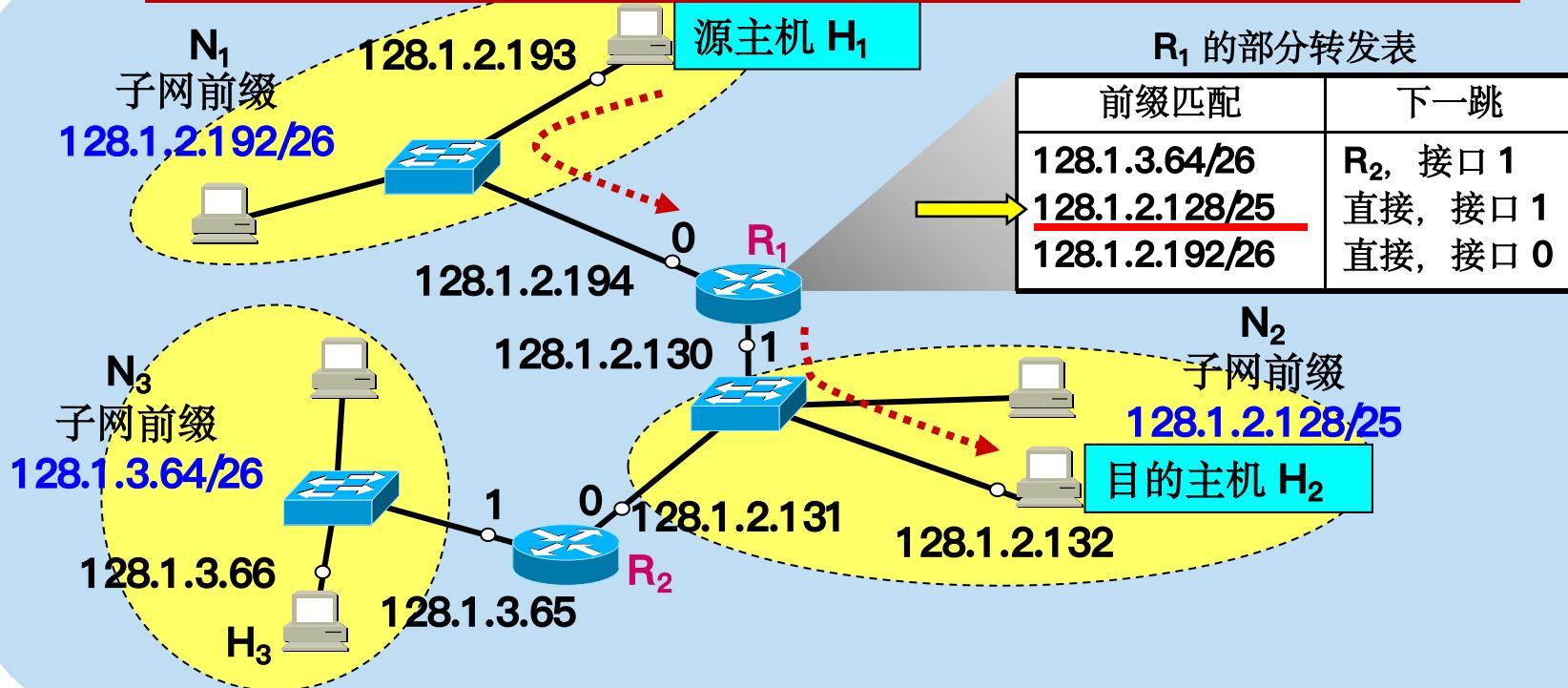
源主机 H_1 必须把分组发送给路由器 R_1 。

路由器 R_1 收到分组后查找转发表。先检查第 1 行。



$128.1.2.132 \text{ AND } 255.255.255.192 = 128.1.2.128$ 不匹配!

路由器 R_1 收到分组后查找转发表。接着检查第 2 行。

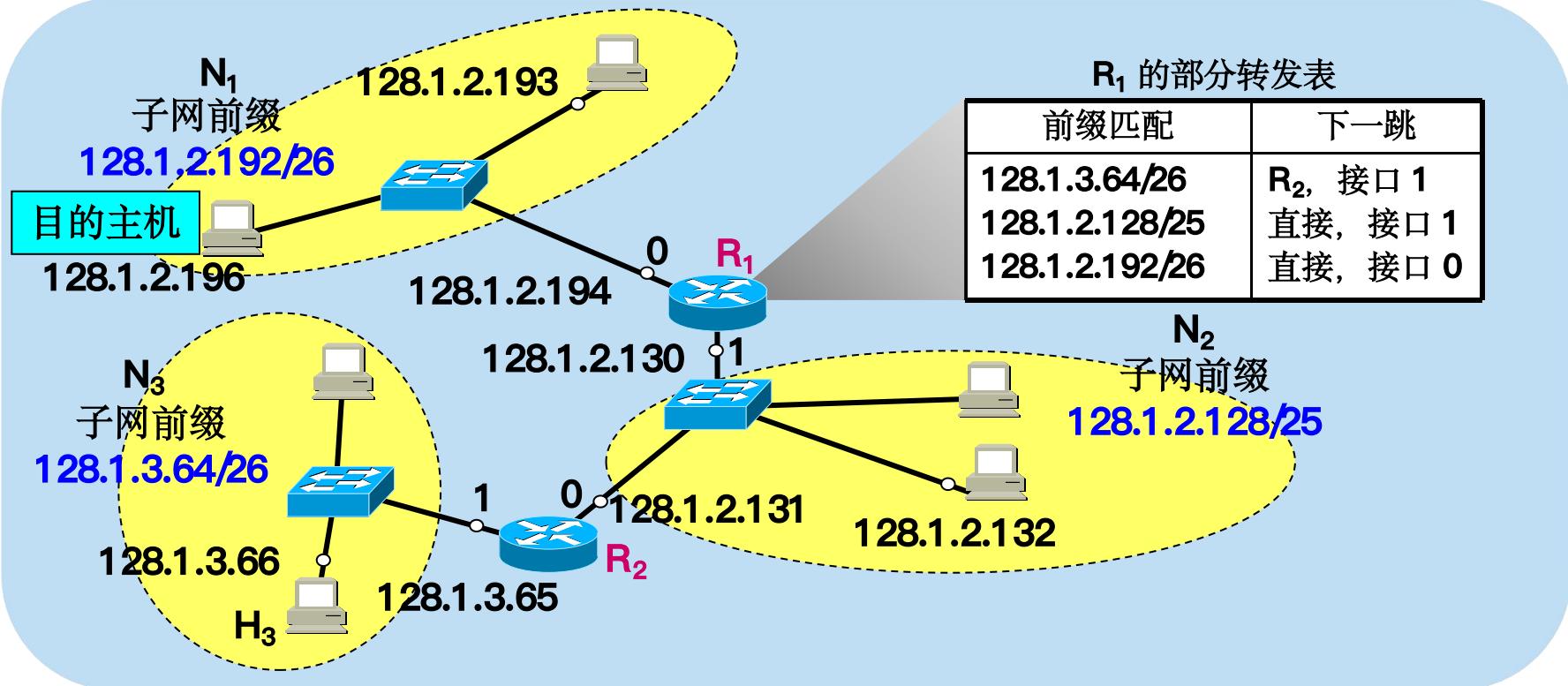


$128.1.2.132 \text{ AND } 255.255.255.128 = 128.1.2.128$ 匹配!

进行分组的直接交付（通过路由器 R_1 的接口 1）。

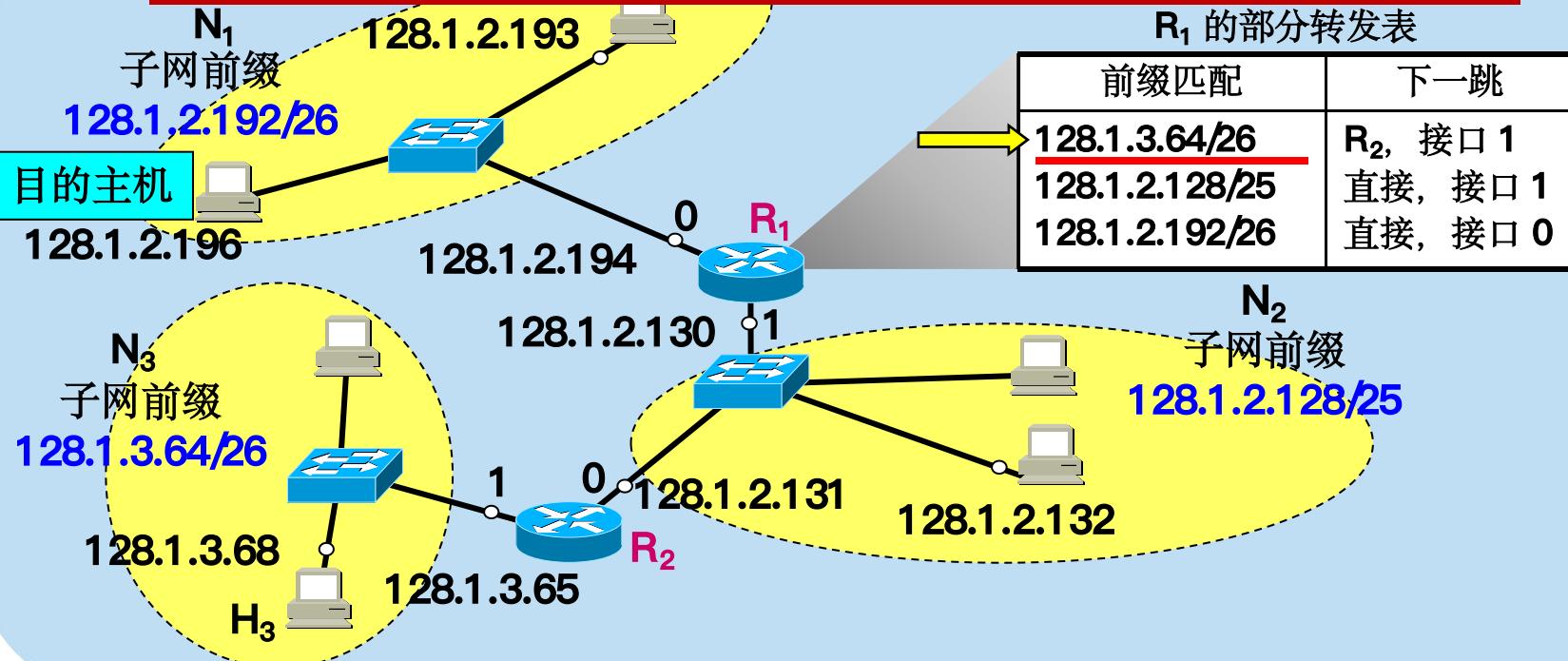
4.3.2 最长前缀匹配

- 使用 CIDR 时，在查找转发表时可能会得到不止一个匹配结果。
- **最长前缀匹配 (longest-prefix matching) 原则：**选择前缀最长的一个作为匹配的前缀。
- 网络前缀越长，其地址块就越小，因而路由就越具体。
- 可以把前缀最长的排在转发表的第一行。



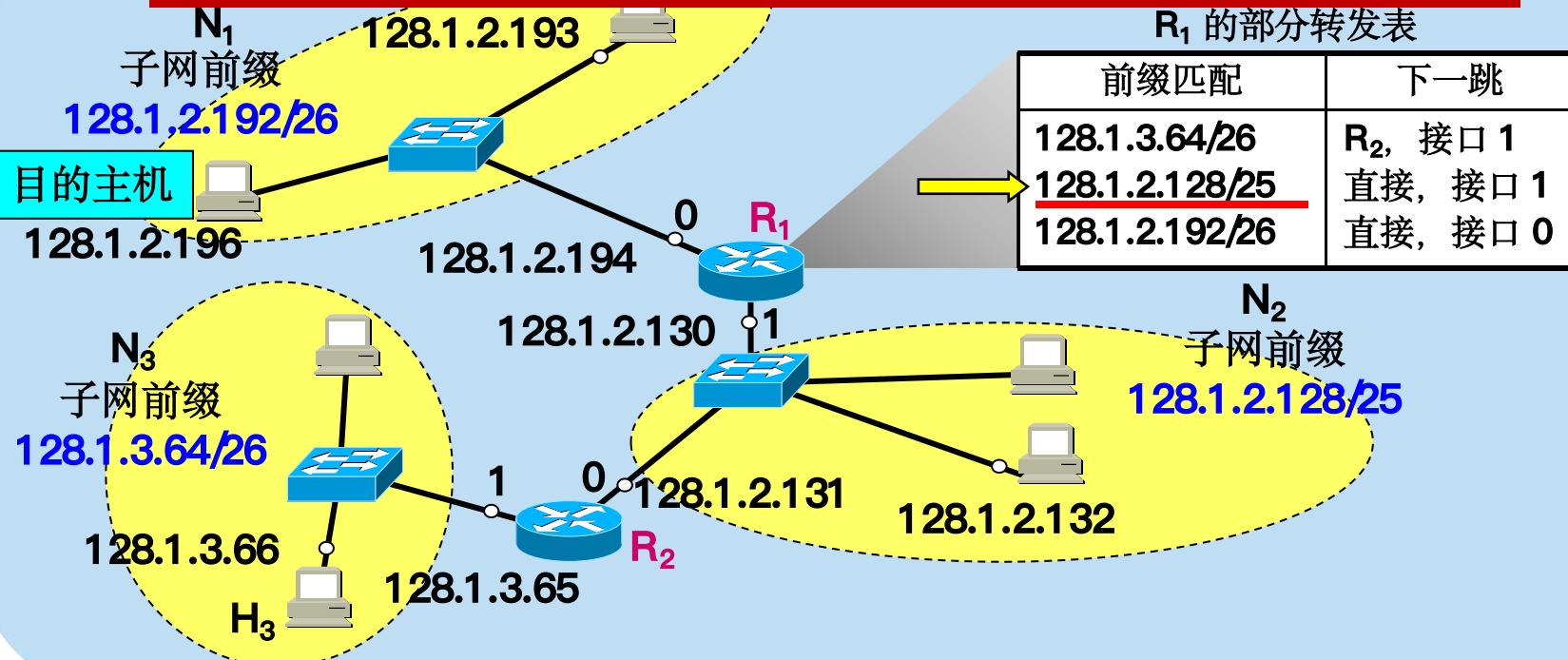
路由器 R₁ 如何转发目的地址是 128.1.2.196 的分组？

路由器 R_1 收到分组后查找转发表。先检查第 1 行。

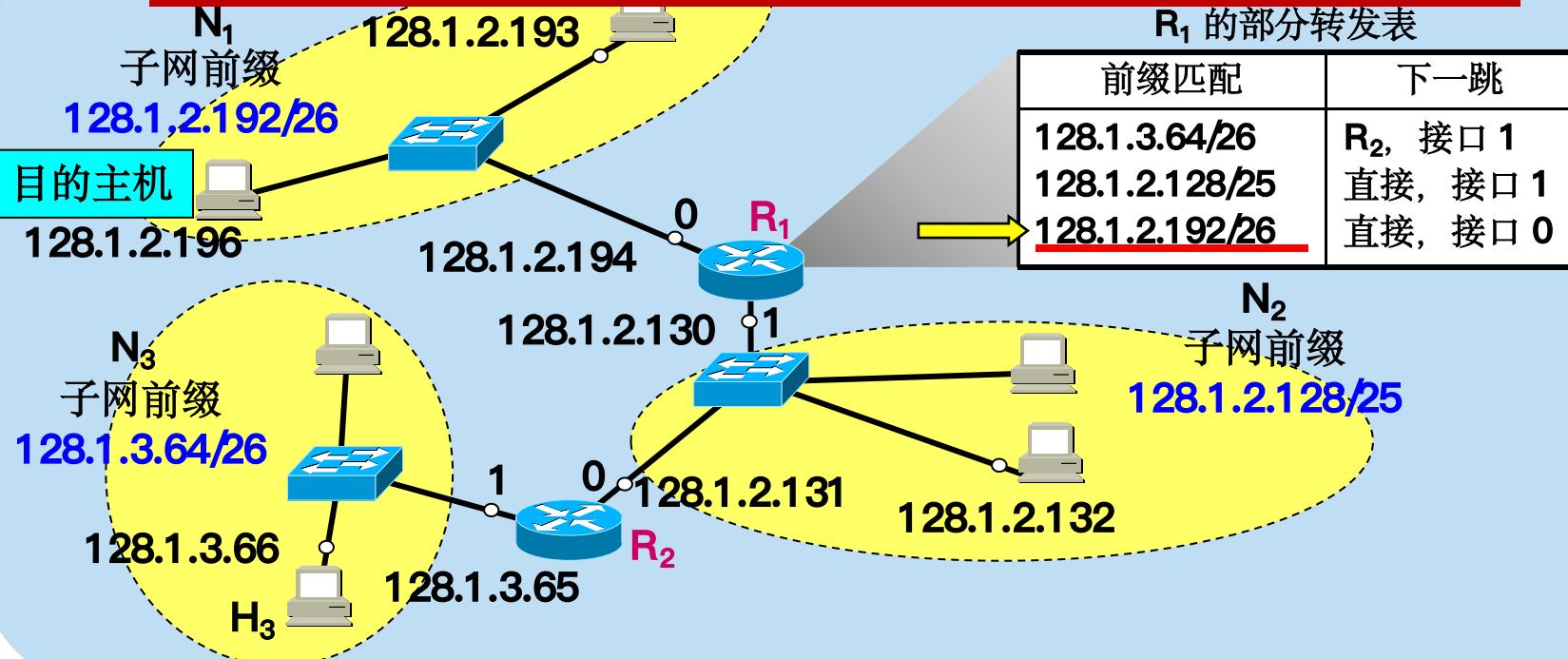


$128.1.2.196 \text{ AND } 255.255.255.192 = 128.1.2.192$ 不匹配!

路由器 R_1 收到分组后查找转发表。接着检查第 2 行。



路由器 R_1 收到分组后查找转发表。接着检查第 3 行。



128.1.2.196 AND 255.255.255.192 = 128.1.2.192 匹配!

4.3.2 最长前缀匹配

- 问题: R_1 从哪个接口向外转发分组?



A. 接口 0 ? 匹配的前缀最长



B. 接口 1 ?

最长前缀匹配:

选择前缀最长的一个作为匹配的前缀

4.3.2 最长前缀匹配

128 . 1 . 2 . 196

目的主机 IP 地址

10000000 00000001 00000010 11000100

128.1.2.192/26 的最小地址

10000000 00000001 00000010 11000000

128.1.2.192/26 的最大地址

10000000 00000001 00000010 11111111

128.1.2.128/25 的最小地址

10000000 00000001 00000010 10000000

128.1.2.128/25 的最大地址

10000000 00000001 00000010 10111111

网络前缀越长，其地址块就越小，路由就越具体(**more specific**)

可以把前缀最长的排在转发表的第 1 行，以加快查表

转发表中的 2 种特殊的路由

- 主机路由 (**host route**)

- ◆ 又叫做**特定主机路由**。
- ◆ 是对特定目的主机的 IP 地址专门指明的一个路由。
- ◆ 网络前缀就是 **a.b.c.d/32**
- ◆ 放在转发表的**最前面**。

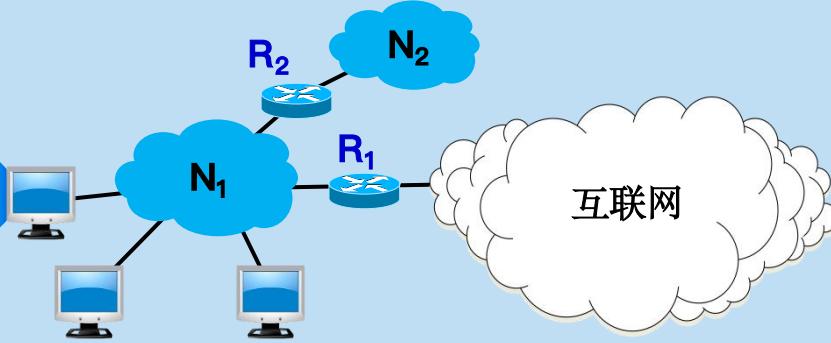
- 默认路由 (**default route**)

- ◆ 不管分组的最终目的网络在哪里，都由指定的路由器 R 来处理
- ◆ 用特殊前缀 **0.0.0.0/0** 表示。

默认路由举例

主机路由表

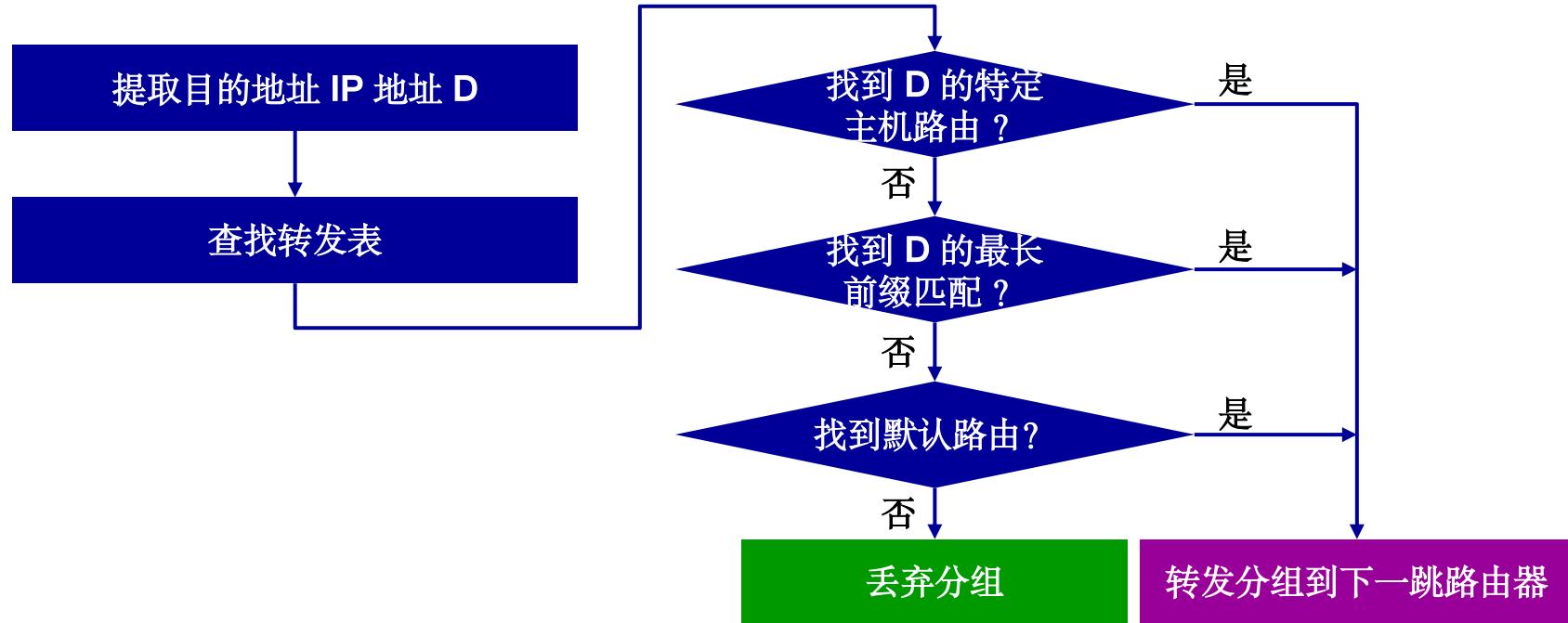
网络前缀	下一跳
N ₁	直接
N ₂	R ₂
0.0.0.0/0	R ₁



路由器 R₁ 充当到达互联网的默认路由器

只要目的网络不是 N₁ 和 N₂, 就一律选择默认路由, 把 IP 数据报先间接交付默认路由器 R₁, 让 R₁ 再转发给下一个路由器。

路由器分组转发算法

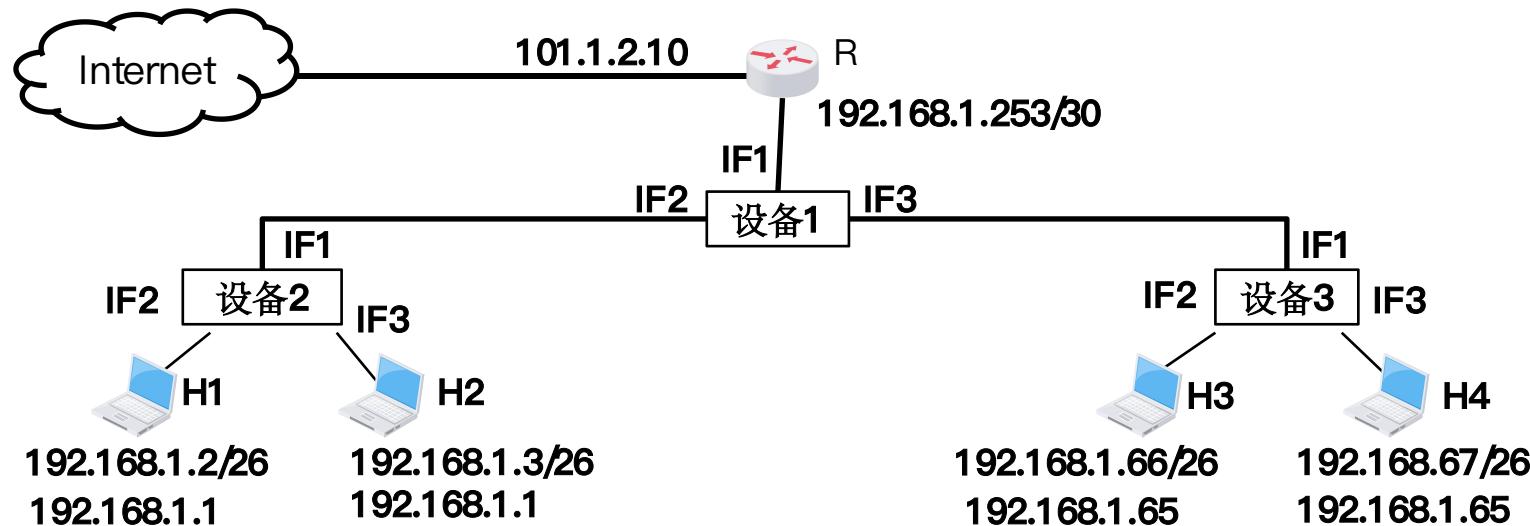


为什么要使用两种地址：IP 地址和 MAC 地址？

- 不同网络使用不同的 **MAC 地址**。**MAC 地址**之间的转换非常复杂。
- 对以太网 **MAC 地址**进行寻址也是极其困难的。
- **IP 编址**把这个复杂问题解决了。
 - ◆ 连接到互联网的主机只需各自拥有一个唯一的 **IP 地址**，它们之间的通信就像连接在同一个网络上那样简单方便，即使必须多次调用 **ARP** 来找到 **MAC 地址**，但这个过程都是由计算机软件自动进行的，对用户来说是看不见的。
- 因此，在虚拟的 **IP 网络**上用 **IP 地址**进行通信非常方便。

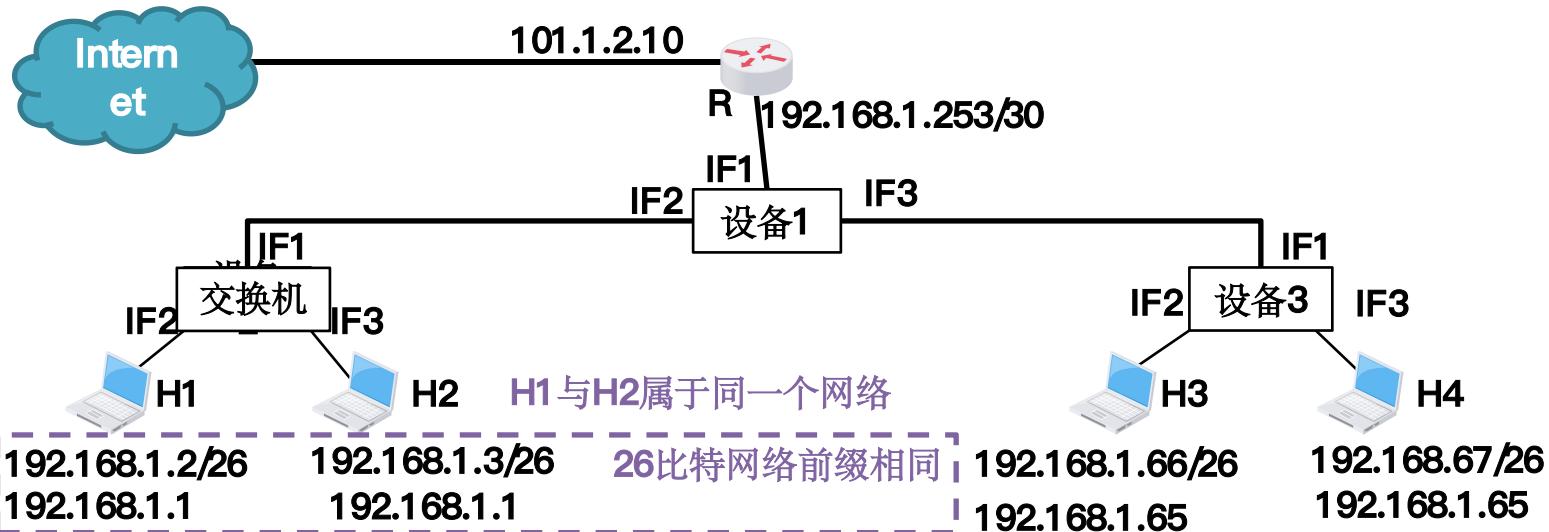
IP数据报的发送和转发过程

【2019年 题47】某网络拓扑如下图所示，其中R为路由器，主机H1~H4的IP地址配置以及R的各接口IP地址配置如图中所示。现有若干台以太网交换机（无VLAN功能）和路由器两类网络互连设备可供选择。



- (1) 设备1、设备2和设备3分别应选择什么类型网络设备？
- (2) 设备1、设备2和设备3中，哪几个设备的接口需要配置IP地址？并为对应的接口配置正确的IP地址。
- (4) 若主机H3发送一个目的地址为192.168.1.127的IP数据报，网络中哪几个主机会收到该数据报？

【2019年 题47】某网络拓扑如下图所示，其中R为路由器，主机H1~H4的IP地址配置以及R的各接口IP地址配置如图中所示。现有若干台以太网交换机（无VLAN功能）和路由器两类网络互连设备可供选择。

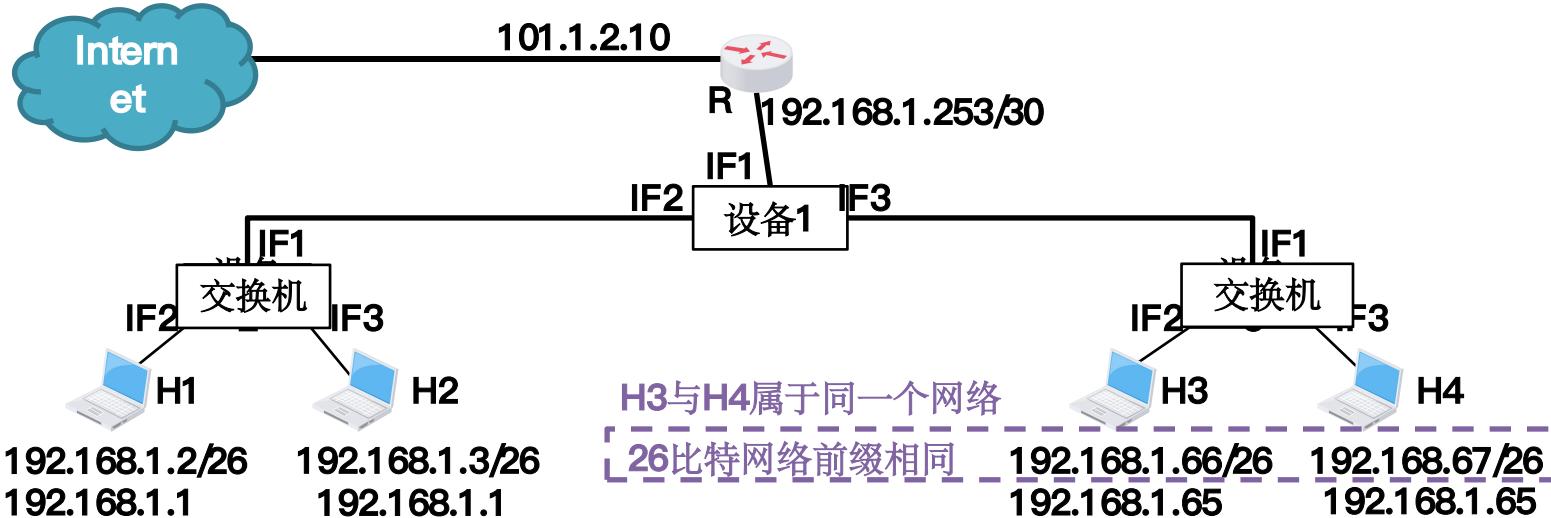


请回答以下问题：

- (1) 设备1、设备2和设备3分别应选择什么类型网络设备？
- (2) 设备1、设备2和设备3中，哪几个设备的接口需要配置IP地址？并为对应的接口配置正确的IP地址。
- (4) 若主机H3发送一个目的地址为192.168.1.127的IP数据报，网络中哪几个主机会收到该数据报？

IP数据报的发送和转发过程

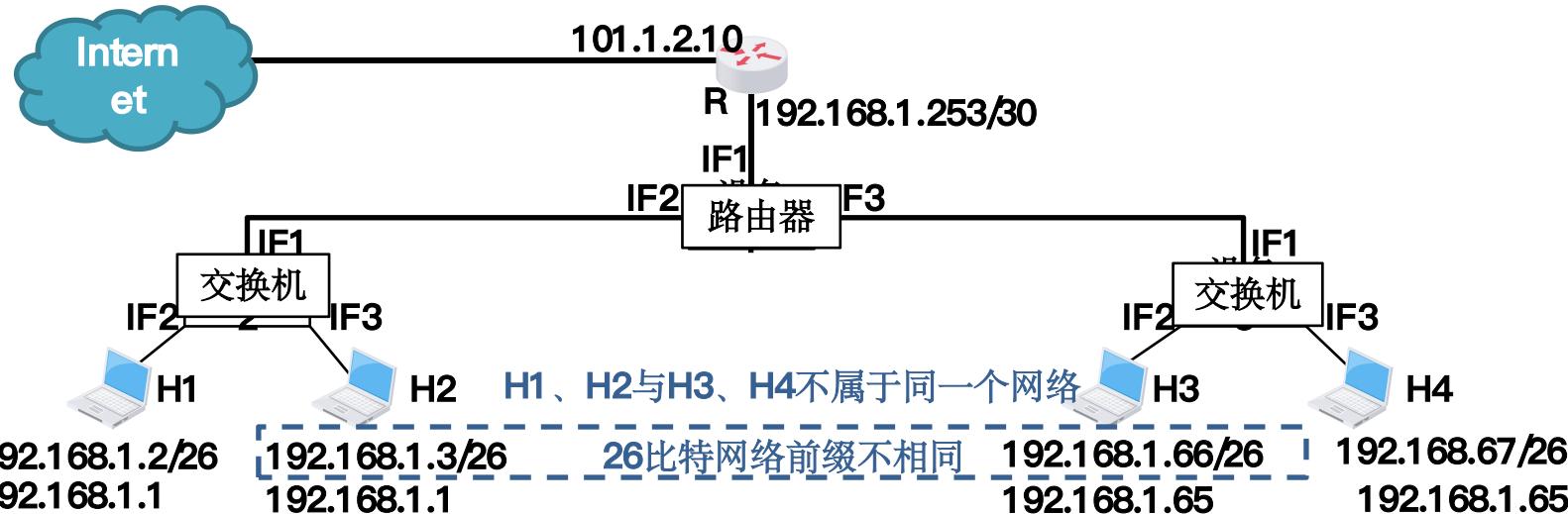
【2019年题47】某网络拓扑如下图所示，其中R为路由器，主机H1~H4的IP地址配置以及R的各接口IP地址配置如图中所示。现有若干台以太网交换机（无VLAN功能）和路由器两类网络互连设备可供选择。



- (1) 设备1、设备2和设备3分别应选择什么类型网络设备？
- (2) 设备1、设备2和设备3中，哪几个设备的接口需要配置IP地址？并为对应的接口配置正确的IP地址。
- (4) 若主机H3发送一个目的地址为192.168.1.127的IP数据报，网络中哪几个主机会收到该数据报？

U1 IP数据报的发送和转发过程

【2019年题47】某网络拓扑如下图所示，其中R为路由器，主机H1~H4的IP地址配置以及R的各接口IP地址配置如图中所示。现有若干台以太网交换机（无VLAN功能）和路由器两类网络互连设备可供选择。



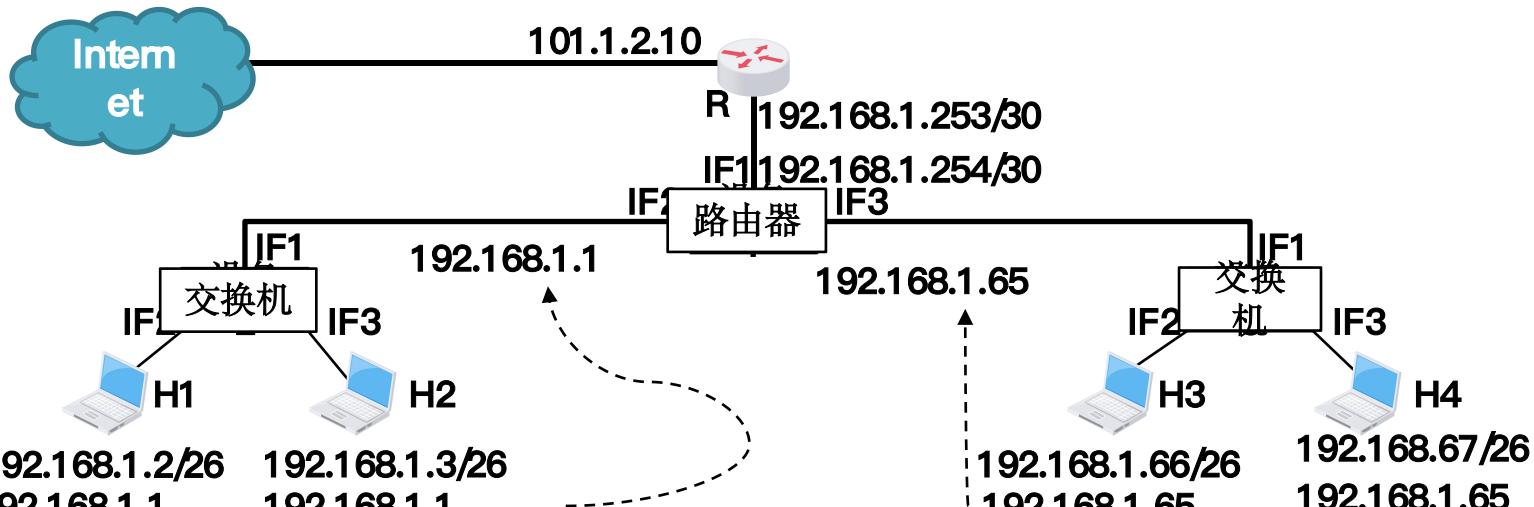
请回答以下问题：

- (1) 设备1、设备2和设备3分别应选择什么类型网络设备？
- (2) 设备1、设备2和设备3中，哪几个设备的接口需要配置IP地址？并为对应的接口配置正确的IP地址。
- (4) 若主机H3发送一个目的地址为192.168.1.127的IP数据报，网络中哪几个主机会收到该数据报？

01

IP数据报的发送和转发过程

【2019年 题47】某网络拓扑如下图所示，其中R为路由器，主机H1~H4的IP地址配置以及R的各接口IP地址配置如图中所示。现有若干台以太网交换机（无VLAN功能）和路由器两类网络互连设备可供选择。



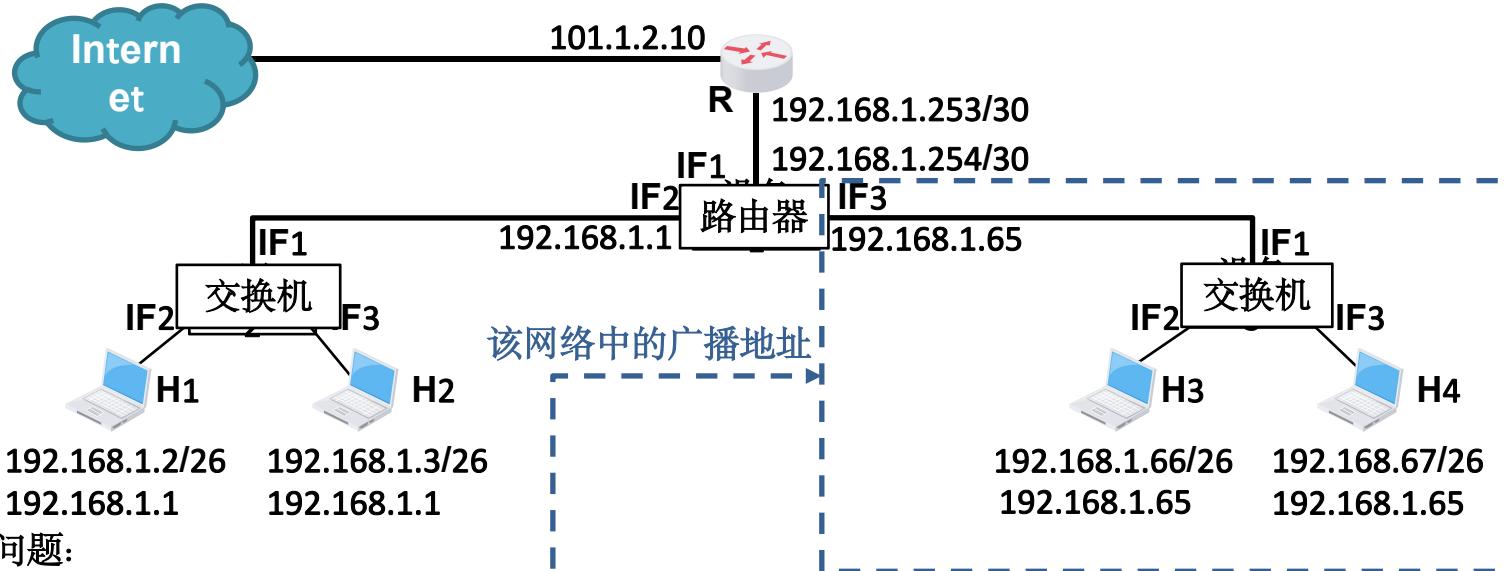
请回答以下问题:

- (1) 设备1、设备2和设备3分别应选择什么类型网络设备?
- (2) 设备1、设备2和设备3中，哪几个设备的接口需要配置IP地址？并为对应的接口配置正确的IP地址。
- (4) 若主机H3发送一个目的地址为192.168.1.127的IP数据报，网络中哪几个主机会收到该数据报？

01

IP数据报的发送和转发过程

【2019年 题47】某网络拓扑如下图所示，其中R为路由器，主机H1~H4的IP地址配置以及R的各接口IP地址配置如图中所示。现有若干台以太网交换机（无VLAN功能）和路由器两类网络互连设备可供选择。



请回答以下问题：

- (1) 设备1、设备2和设备3分别应选择什么类型网络设备？
- (2) 设备1、设备2和设备3中，哪几个设备的接口需要配置IP地址？并为对应的接口配置正确的IP地址。
- (4) 若主机H3发送一个目的地址为192.168.1.127的IP数据报，网络中哪几个主机会收到该数据报？

地址分配、聚合、路由表转发相关内容是本书、本章的重点内容。请同学们课后自己练习，老师抽空课堂上讲解。

路由表生成及路由转发相关课后习题：

课本 P205 4-37 P207 4-48、4-49

4.4

网际控制 报文协议 **ICMP**

4.4.1

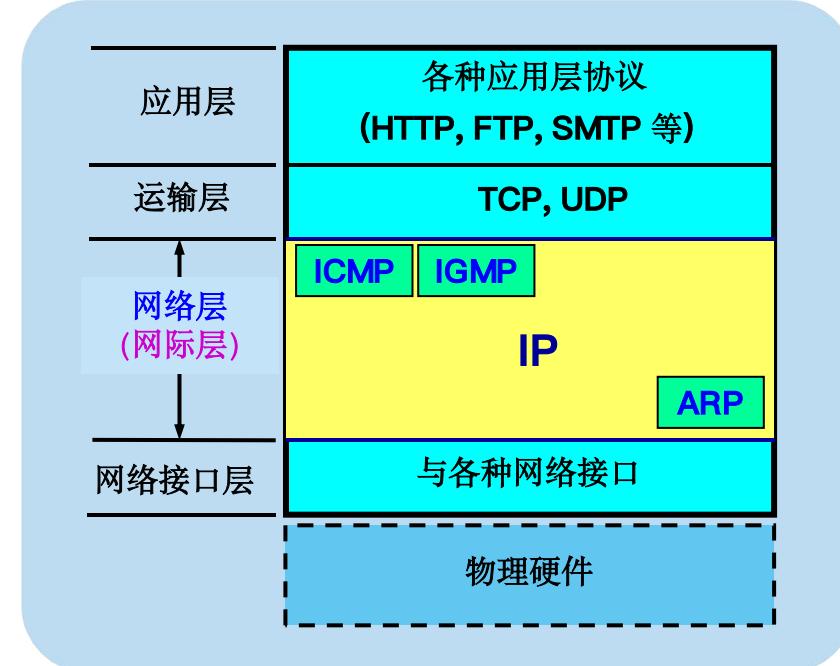
ICMP 报文的种类

4.4.2

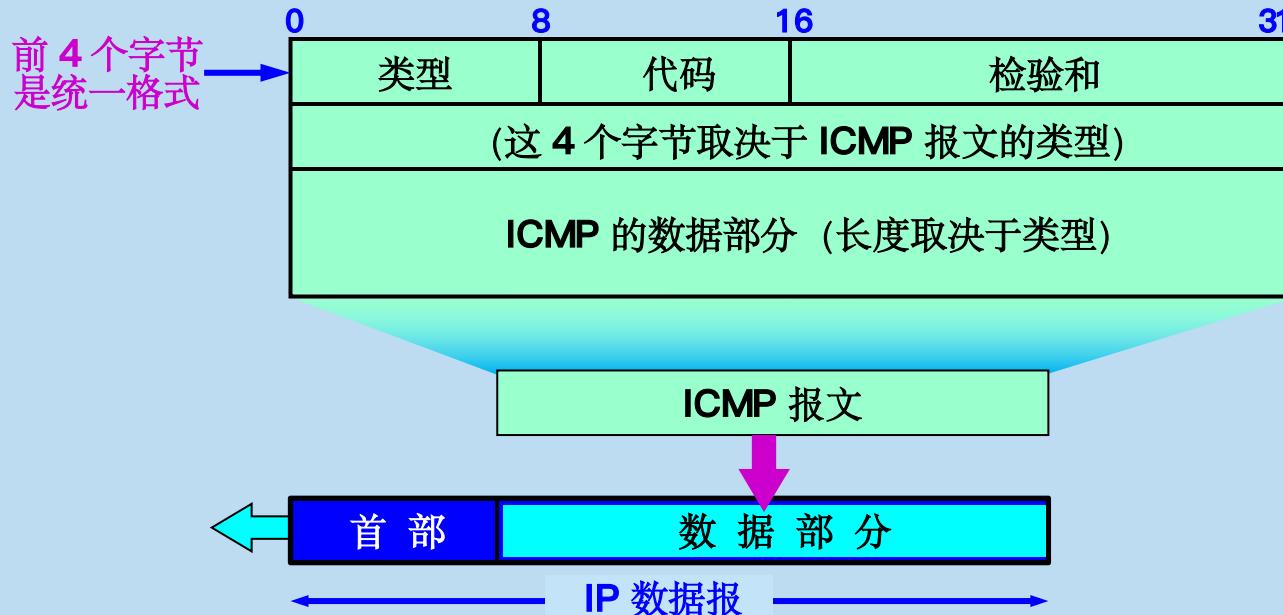
ICMP 的应用举例

4.4 网际控制报文协议 ICMP

- ICMP (Internet Control Message Protocol) 允许主机或路由器报告差错情况和提供有关异常情况的报告。
- ICMP 是互联网的标准协议。
- 但 ICMP 不是高层协议，而是 IP 层的协议。



ICMP 报文的格式



4.4.1 ICMP 报文的种类

- 2 种：差错报告报文，询问报文。

几种常用的 ICMP 报文类型

ICMP 报文种类	类型的值	ICMP报文的类型
差错报告报文	3	终点不可达
	11	时间超过
	12	参数问题
	5	改变路由 (Redirect)
询问报文	8 或 0	回送 (Echo) 请求或回答
	13 或 14	时间戳 (Timestamp) 请求或回答

4.4.2 ICMP 的应用举例

PING (Packet InterNet Groper)

- 用来测试两个主机之间的连通性。
- 使用了 ICMP 回送请求与回送回答报文。
- 是应用层直接使用网络层 ICMP 的例子，没有通过运输层的 TCP 或 UDP。

PING 的应用举例

```
C:\Documents and Settings\XXXR>ping mail.sina.com.cn

Pinging mail.sina.com.cn [202.108.43.230] with 32 bytes of data:

Reply from 202.108.43.230: bytes=32 time=368ms TTL=242
Reply from 202.108.43.230: bytes=32 time=374ms TTL=242
Request timed out.
Reply from 202.108.43.230: bytes=32 time=374ms TTL=242

Ping statistics for 202.108.43.230:
    Packets: Sent = 4, Received = 3, Lost = 1 (25% loss),
Approximate round trip times in milli-seconds:
    Minimum = 368ms, Maximum = 374ms, Average = 372ms
```

用 PING 测试邮件服务器 mail.sina.com.cn 的连通性

4.4.2 ICMP 的应用举例

Traceroute

- 这是**UNIX**操作系统中名字。在 **Windows** 操作系统中这个命令是 **tracert**。
- 用来跟踪一个分组从源点到终点的路径。
- 它利用 **IP** 数据报中的 **TTL** 字段、**ICMP** 时间超过差错报告报文和 **ICMP** 终点不可达差错报告报文实现对从源点到终点的路径的跟踪。

4.4.2 ICMP 的应用举例

```
C:\Documents and Settings\XXXR>tracert mail.sina.com.cn

Tracing route to mail.sina.com.cn [202.108.43.230]
over a maximum of 30 hops:

 1  24 ms    24 ms    23 ms  222.95.172.1
 2  23 ms    24 ms    22 ms  221.231.204.129
 3  23 ms    22 ms    23 ms  221.231.206.9
 4  24 ms    23 ms    24 ms  202.97.27.37
 5  22 ms    23 ms    24 ms  202.97.41.226
 6  28 ms    28 ms    28 ms  202.97.35.25
 7  50 ms    50 ms    51 ms  202.97.36.86
 8  308 ms   311 ms   310 ms  219.158.32.1
 9  307 ms   305 ms   305 ms  219.158.13.17
10  164 ms   164 ms   165 ms  202.96.12.154
11  322 ms   320 ms   2988 ms  61.135.148.50
12  321 ms   322 ms   320 ms  freemail43-230.sina.com [202.108.43.230]

Trace complete.
```

用 **tracert** 命令获得到新浪网的邮件服务器 **mail.sina.com.cn** 的路由信息

4.5 IPv6

4.5.1

IPv6 的基本首部

4.5.2

IPv6 的地址

4.5.3

从 IPv4 向 IPv6 过渡

4.5.4

ICMPv6

冒号十六进制记法

- 在 IPv6 中，每个地址占 128 位，地址空间大于 3.4×10^{38} 。
- 使用**冒号十六进制记法**(colon hexadecimal notation, 简写为 colon hex): 16 位的值用十六进制值表示，各值之间用冒号分隔。

点分十进制数记法: 104.230.140.100.255.255.255.255.0.0.17.128.150.10.255.255

冒号十六进制记法: 68E6:8C64:FFFF:FFFF:0000:1180:960A:FFFF

冒号十六进制记法: 68E6:8C64:FFFF:FFFF:0:1180:960A:FFFF

两个技术: 零压缩, 点分十进制记法的后缀。

零压缩

- 零压缩 (zero compression): 一串连续的零可以用一对冒号取代。

FF05:0:0:0:0:0:B3

可压缩为: FF05::B3

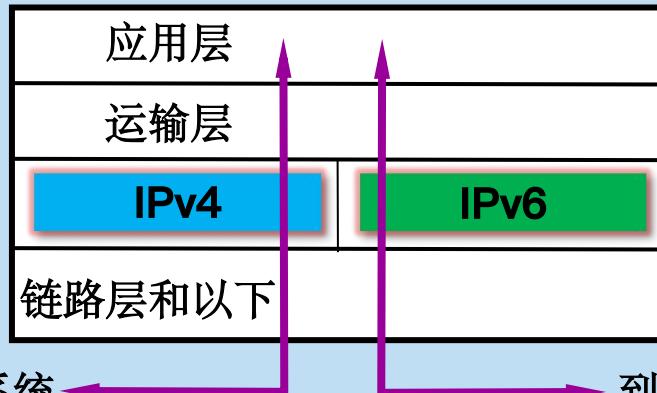
0:0:0:0:0:128.10.2.1	→ ::128.10.2.1
1080:0:0:0:8:800:200C:417A	→ 1080::8:800:200C:417A
FF01:0:0:0:0:0:101 (多播地址)	→ FF01::101
0:0:0:0:0:0:1 (环回地址)	→ ::1
0:0:0:0:0:0:0 (未指明地址)	→ ::

注意: 在任一地址中, 只能使用一次零压缩。

4.5.3 从 IPv4 向 IPv6 过渡

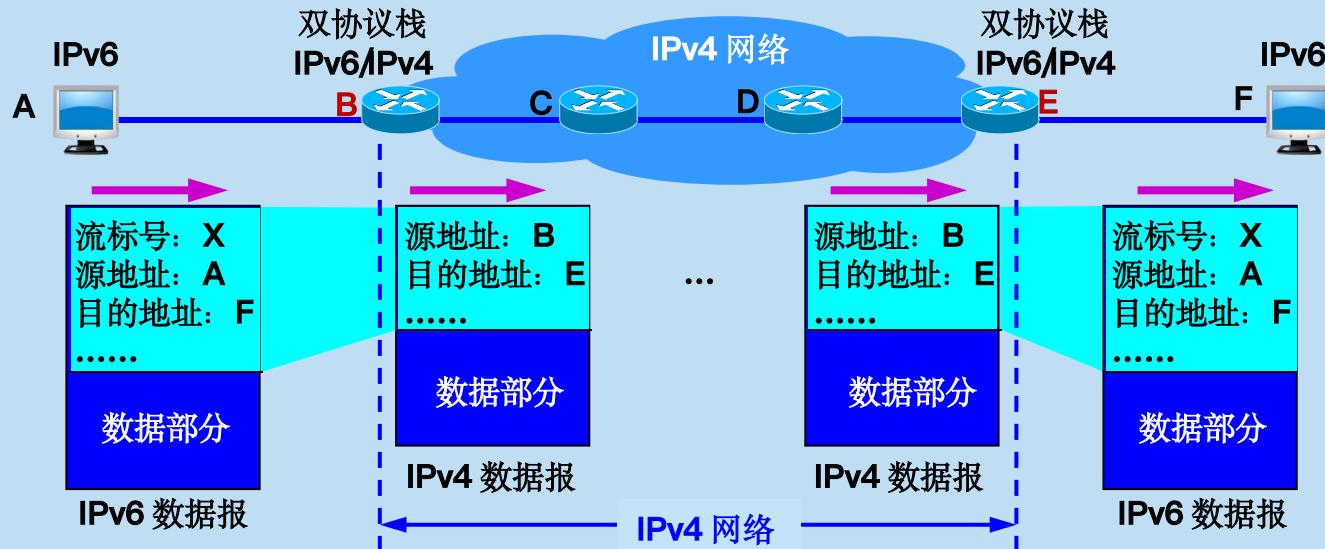
- 方法: 逐步演进, 向后兼容。
- 向后兼容: IPv6 系统必须能够接收和转发 IPv4 分组, 并且能够为 IPv4 分组选择路由。
- 两种过渡策略:
 1. 使用双协议栈
 2. 使用隧道技术

双协议栈



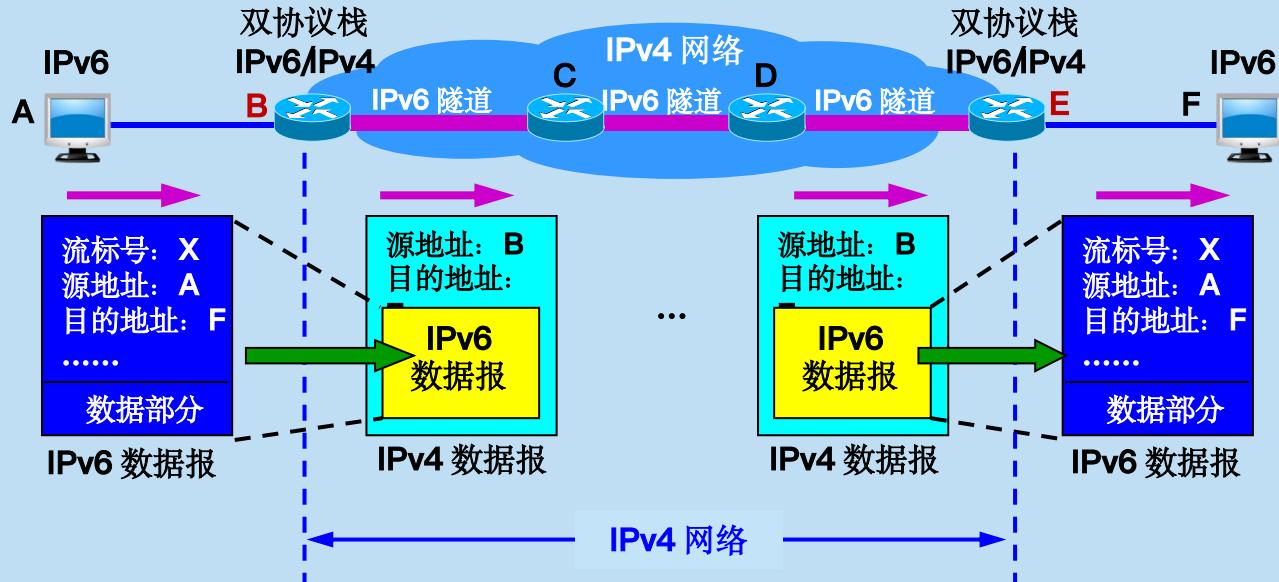
IPv6/IPv4 双协议栈主机 (或路由器)

双协议栈



使用双协议栈进行从 IPv4 到 IPv6 的过渡

隧道技术



使用隧道技术进行从 IPv4 到 IPv6 的过渡

本讲小结

掌握:

- IP数据报格式
- 路由转发的过程

了解:

- ICMP的工作过程
- IPv6地址