

# 网络层

	设备	地址
直接交付	链路或网桥	链路层地址
间接交付	路由器	网络层地址

## 为什么不能用交换机做网络互联？

网桥不能解决的互联问题

- 不具备扩展性
  - 路由表大，每个主机一条记录
- 广播风暴
- 不能跨越多个链路层
  - Ethernet + 4G + 802.11
- 流量控制（带宽控制）
- 低开销路由
- QoS

## 网络层提供的两种服务

电信网：虚电路服务

- 面向连接的通信方式
  - 建立虚电路，保证双方通信所需的网络资源
  - 可靠传输的网络层协议，使发送的分组无差错按序到达终点
- 虚电路
  - 逻辑连接：分组沿逻辑连接按照存储转发方式传送，不是真正建立物理连接
- 电路交换的电话通建立物理连接

互联网提供数据报服务

互联网先驱提出

- 网络层向上提供简单灵活的、无连接的、尽最大努力交付的数据报服务
- 网络在发送分组时不建立连接
- 每个分组（即 IP 数据报）独立发送，与其前后分组无关
- 网络层不提供服务质量的承诺。
  - 传送的分组可能出错、丢失、重复和失序（不按序到达终点）
  - 不保证分组传送的时限
  - 可靠传输由位于网络边缘的通信双方的运输层保证

虚电路服务与数据报服务的对比

对比的方面	虚电路服务	数据报服务
思路	可靠通信应当由网络来保证	可靠通信应当由用户主机来保证
连接的建立	必须有	不需要
终点地址	仅在连接建立阶段使用，每个分组使用短的虚电路号	每个分组都有终点的完整地址
分组的转发	属于同一条虚电路的分组均按照同一路由进行转发	每个分组独立选择路由进行转发
当结点出故障时	所有通过出故障的结点的虚电路均不能工作	出故障的结点可能会丢失分组，一些路由可能会发生变化
分组的顺序	总是按发送顺序到达终点	到达终点时不一定按发送顺序
端到端的差错处理和流量控制	可以由网络负责，也可以由用户主机负责	由用户主机负责

网络层的两个层面

数据层面

- 路由器根据本路由器生成的转发表，把收到的分组从查找到的对应接口转发出去。
- 独立工作。
- 采用硬件进行转发，快。

控制层面

- 根据路由选择协议所用的路由算法计算路由，创建出本路由器的路由表。
- 许多路由器协同动作。
- 采用软件计算，慢。

网际协议 IP

与网际协议 IPv4 配套的 3 个协议

1. 地址解析协议 ARP
  - 从 IP 地址解析出 MAC 地址。
2. 网际控制报文协议 ICMP
3. 网际组管理协议 IGMP

虚拟互联网络(逻辑互连网络)

使用转发器或网桥不称为网络互连。转发器、网桥或交换机仅把一个网络扩大了，仍然是一个网络。网络互连使用路由器。

- 虚拟互连网络(逻辑互连网络)
  - 互连起来的各种物理网络的异构性客观存在
  - IP协议使性能各异网络从用户看是统一的网络

- IP 网
  - 使用 IP 协议的虚拟互连网络
  - 互联网主机进行通信时，像在一个网络上通信一样，看不见互连的具体网络异构细节
- 互联网 (Internet)
  - 覆盖全球的 IP 网及上层使用 TCP 协议

## IP 地址

- IP 地址及表示方法
  - 整个因特网看成为一个单一的、抽象的网络。
  - IP 地址是给每个连接在因特网上的主机（或路由器）分配一个在全世界范围唯一的 32 位标识符。
  - 没有IP地址，就无法和网上的其他设备进行通信。
- IP 编址演进
  - 分类的 IP 地址
    - 最基本的编址方法，1981年通过标准协议。
  - 子网划分
    - 改进的编址方法， [RFC950]在1985年通过。
  - 构成超网
    - 新的无分类编址方法，1993年提出得到推广。

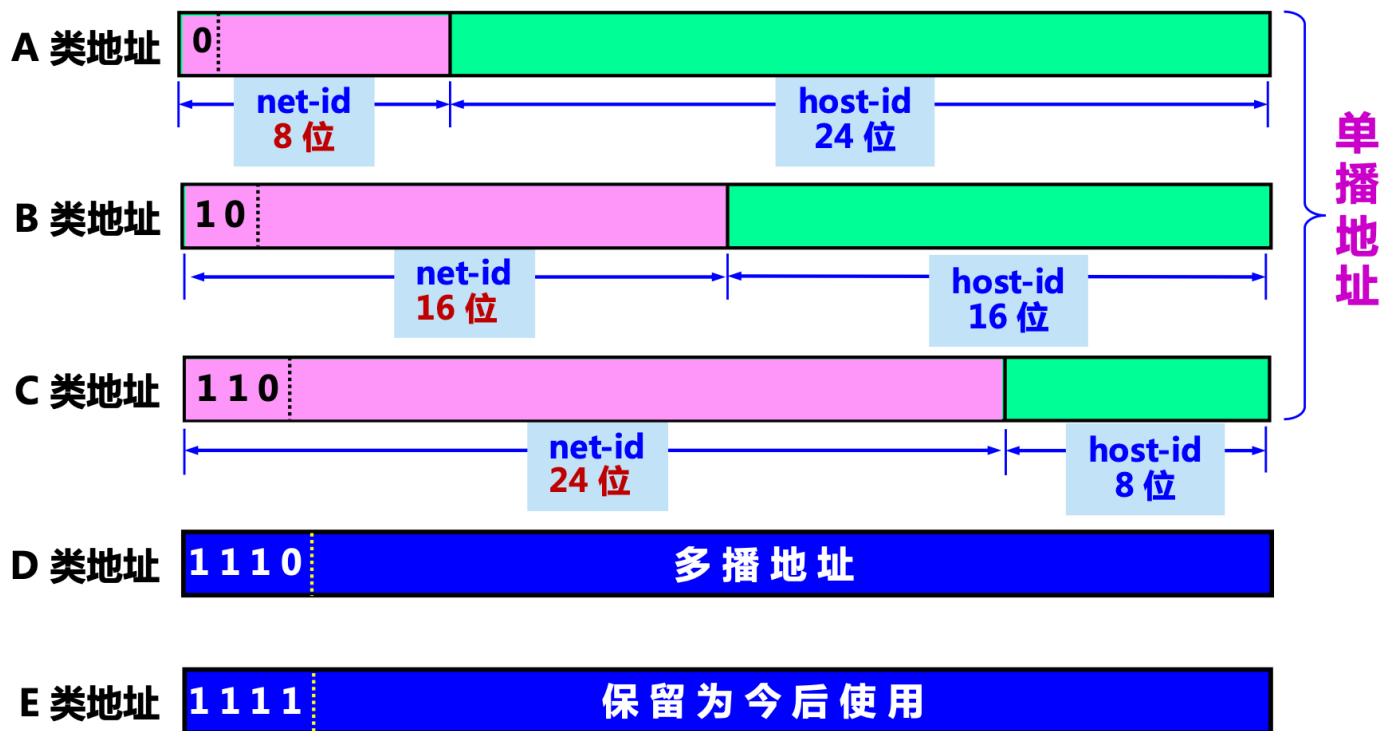
## 分类的 IP 地址

### IP 地址

32 位二进制代码，分为每 8 位为一组，将每 8 位的二进制数转换为十进制数，采用点分十进制记法

### IP 地址采用2级结构

IP 地址 ::= { <网络号>, <主机号> }



网络类别	最大可指派的网络数	第一个可指派的网络号	最后一个可指派的网络号	每个网络中最大主机数
A	126 ( $2^7 - 2$ )	1	126	16777214 ( $2^{24} - 2$ )
B	16383 ( $2^{14} - 1$ )	128.1	191.255	65534 ( $2^{16} - 2$ )
C	2097151 ( $2^{21} - 1$ )	192.0.1	223.255.255	254 ( $2^8 - 2$ )

#### 分类的 IP 地址的优点和缺点

- 优点
  - 管理简单
  - 使用方便
  - 转发分组迅速
  - 划分子网，灵活地使用
- 缺点
  - 设计上不合理
  - 大地址块，浪费地址资源
  - 即使采用划分子网的方法，也无法解决 IP 地址枯竭的问题

#### 划分子网

##### 基本思路

- 划分子网纯属单位内部事务
- 单位对外表现为没有划分子网的网络
- 从主机号借用若干个位作为子网号 subnet-id，而主机号 host-id 也就相应减少了若干个位
  - IP地址 ::= {<网络号>, <子网号>, <主机号>}

- 从其他网络发送给本单位某个主机的IP数据报，根据IP数据报的目的网络号 net-id，先找到连接在本单位网络上的路由器
- 该路由器收到 IP 数据报后，按目的网络号 net-id 和子网号 subnet-id 找到目的子网，将 IP 数据报直接交付目的主机。

**子网掩码：**IP地址的子网部分

- 路由器和相邻路由器交换路由信息时，把所在网络（或子网）的子网掩码告诉相邻路由器。
- 路由器转发表中的每个项目，包含目的网络地址和该网络的子网掩码。
- 路由器连接在两个子网上就有两个网络地址和两个子网掩码。
- 注意：子网掩码仅保存在路由器的转发表中，IP数据包中没有子网掩码表项。

## 无分类编址 CIDR

**CIRD：**无分类域间路由选择。

- 消除了传统的 A 类、B 类和 C 类地址以及划分子网的概念，可以更加有效地分配 IPv4 的地址空间，但无法解决 IP 地址枯竭的问题。

**要点：**

- 网络前缀
- 地址块
- 地址掩码

**网络前缀**

IP 地址 ::= { <网络前缀>, <主机号> }

**CIDR 记法：**斜线记法 (slash notation)

a.b.c.d / n: 二进制 IP 地址的前 n 位是网络前缀。

**地址块**

- CIDR 把网络前缀都相同的所有连续的 IP 地址组成一个 CIDR 地址块。
- 一个 CIDR 地址块包含的 IP 地址数目，取决于网络前缀的位数。

**地址掩码**

- 又称为子网掩码
- 32位
- 目的：让机器从 IP 地址迅速算出网络地址。
- 由一连串 1 和接着的一连串 0 组成，而 1 的个数就是网络前缀的长度。

**三个特殊的 CIDR 地址块**

网络前缀长度	点分十进制	说明
/32	255.255.255.255	就是一个 IP 地址。这个特殊地址用于主机路由
/31	255.255.255.254	只有两个 IP 地址，其主机号分别为 0 和 1。这个地址块用于点对点链路
/0	0.0.0.0	同时 IP 地址也是全 0，即 0.0.0.0/0。用于默认路由

IP 地址的特点

- 1. 每个 IP 地址都由网络前缀和主机号两部分组成。
- 2. IP 地址是标志一台主机（或路由器）和一条链路的接口。
- 3. 转发器或交换机连接起来的若干个局域网仍为一个网络
- 4. 在 IP 地址中，所有分配到网络前缀的网络都是平等的。

注意：

同一个局域网上的主机或路由器的 IP 地址中的网络号必须一样。

路由器的每一个接口都有一个不同网络号的 IP 地址。

IP 地址与 MAC 地址

IP 地址

- 虚拟地址、软件地址、逻辑地址。
- 网络层和以上各层使用。
- 放在 IP 数据报的首部。

MAC 地址

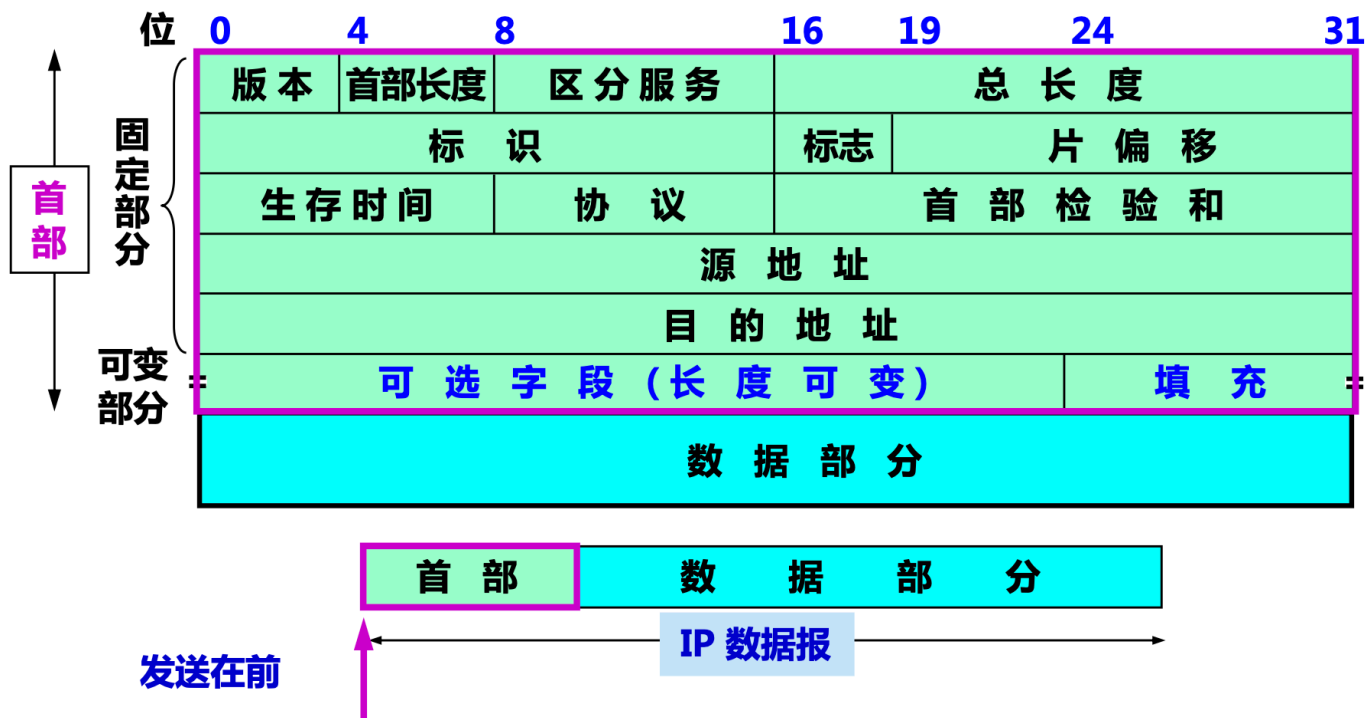
- 固化在网卡上的 ROM 中。
- 硬件地址、物理地址。
- 数据链路层使用。
- 放在 MAC 帧的首部。

IP 数据报的格式

IP 数据报由首部和数据两部分组成。首部的前一部分是固定长度，共 20 字节，是所有 IP 数据报必须具有的。可选字段，其长度是可变的。

IP 的首部长度的最大值是 60 字节。

片偏移以 8 个字节为偏移单位。



### 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。
- 把前缀最长的排在转发表的第 1 行。

### 转发表中的两种特殊路由

#### 主机路由

- 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

- 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。

### 二叉线索树

从二叉线索的根节点自顶向下的深度最多有 32 层，每一层对应于 IP 地址中的一位。

为检查网络前缀是否匹配，必须使二叉线索中的每一个叶节点包含所对应的网络前缀和子网掩码。

步骤：

1. 找到了一个叶节点
2. 将目的 IP 地址和该叶节点的子网掩码进行按位 AND 运算，看结果是否与叶节点的网络前缀相匹配。

3. 若匹配，就按下一跳的接口转发该分组。否则，就丢弃该分组。

## 路由选择协议

路由选择协议属于网络层控制层面的内容

自治系统之间的路由选择叫做域间路由选择

- 内部网关协议 EGP
- 使用最多：BGP-4

自治系统内部的路由选择叫做域内路由选择

- 内部网关协议 IGP
  - 常用：RIP, OSPF

## 内部网关协议 OSPF

- 开放最短路径优先 OSPF
  - 克服 RIP 的缺点在 1989 年开发出来的。
  - 现在使用 OSPFv2。
- 原理很简单，实现很复杂。
  - Dijkstra 提出的最短路径算法 SPF。
- 采用分布式的链路状态协议

### 三个特点

- 洪泛法
  - 向本自治系统中所有路由器发送信息。
- 发送的信息是与本路由器相邻的所有路由器的链路状态，但这只是路由器所知道的部分信息。
  - 链路状态：说明本路由器都和哪些路由器相邻，以及该链路的度量
- 当链路状态发生变化或每隔一段时间（如30分钟），路由器才用洪泛法向所有路由器发送此信息。

### 链路状态数据库

每个路由器最终都能建立。

- 全网的拓扑结构图。
- 在全网范围内是一致的（称为链路状态数据库的同步）。

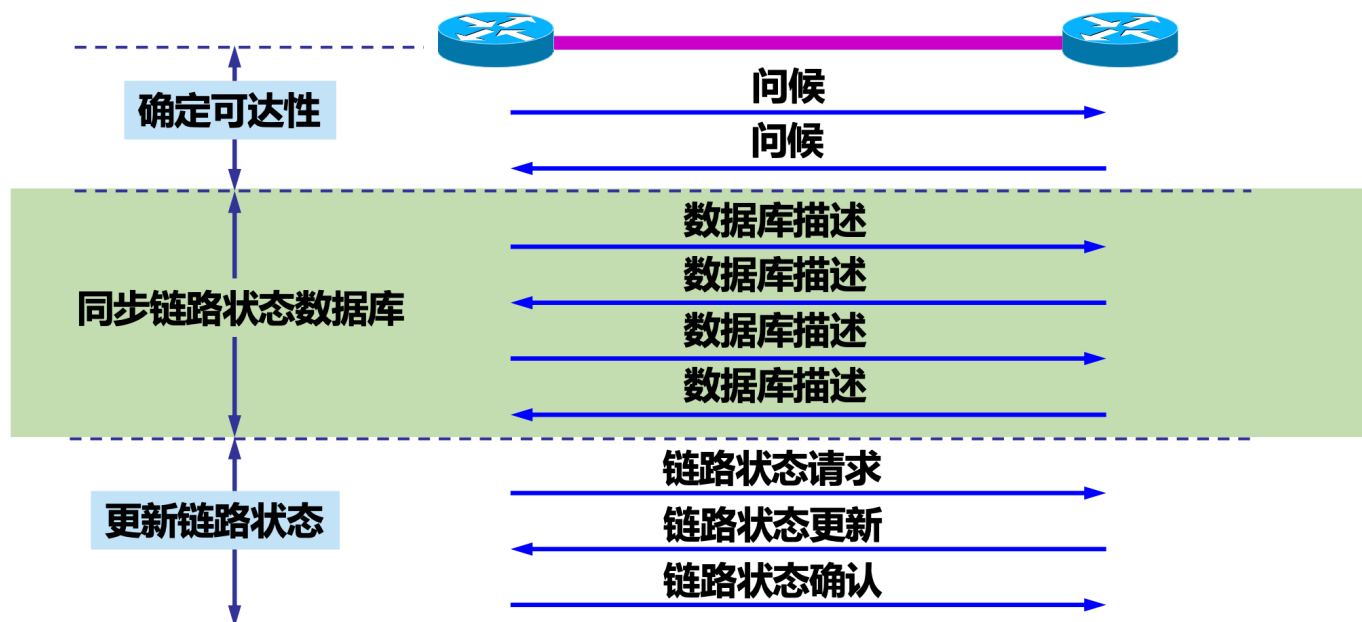
每个路由器使用链路状态数据库构造自己的转发（路由）表

- 例如，使用Dijkstra的最短路径路由算法。

重要优点：OSPF 更新过程收敛速度快。

### OSPF 工作过程





OSPF 使用可靠的洪泛法发送更新分组

划分区域优缺点

- 优点
  - 减少整个网络上的通信量。
  - 减少需要维护的状态数量。
- 缺点
  - 交换信息的种类增多。
  - 使 OSPF 协议更加复杂。

分层次划分区域的好处：

使每一个区域内部交换路由信息的通信量大大减小，因而使 OSPF 协议能够用于规模很大的自治系统中。

## 距离向量协议 RIP

路由信息协议 RIP

- 分布式的、基于距离向量的路由选择协议。
- 互联网的标准协议。

最大优点：简单

网络中的每个路由器都要维护从它自己到其他每一个目的网络的距离记录。

距离

- 路由器到直接连接的网络的距离= 1。
- 路由器到非直接连接的网络的距离= 所经过的路由器数 + 1。
- “距离”即“跳数”，每经过一个路由器，跳数就加 1。
- 一条路径最多只能包含 15 个路由器。
  - “距离”的最大值为 16 时即相当于不可达。
- 不能在两个网络之间同时使用多条路由，选择距离最短的路由。

## 距离向量算法

- Bellman-Ford 算法
  - 设  $V$  是结点  $X$  到  $Y$  的最短路径上的一个结点，若把路径  $X \rightarrow Y$  拆成两段路径  $X \rightarrow V$  和  $V \rightarrow Y$ ，则每一段路径  $X \rightarrow V$  和  $V \rightarrow Y$  也都分别是结点  $X$  到  $V$  和结点  $V$  到  $Y$  的最短路径。
- 每个结点定期向所有邻居发送自己的距离向量表
- 当结点从其邻居处获得距离向量表后，根据 B-F Equation 计算自己新的距离向量表

$$D_x(y) \leftarrow \min_v \{c(x,v) + D_v(y)\} \quad \text{for each node } y \in N$$

- 当结点规模有限时，所有结点的距离向量表均会收敛到最短路径的距离向量

## 距离向量协议特点

迭代、异步：

每次本地更新由以下情况引起

- 本地链路代价改变
- 收到来自邻居的距离向量表更新消息

分布式：

- 每个节点仅在自身距离向量表发生改变时通知邻居
  - 被通知的邻居根据自身距离向量表的更新结果决定是否继续通知他的邻居

## RIP 三个特点

- 仅和相邻路由器交换信息。
- 交换的信息是当前本路由器所知道的全部信息，即自己的路由表。
- 按固定时间间隔交换路由信息
  - 当网络拓扑发生变化时，路由器也及时向相邻路由器通告拓扑变化后的路由信息。

## 链路代价改变

好消息传得快而坏消息传播得慢。这是 RIP 的一个主要缺点。

## RIP 协议的优缺点

- 优点
  - 实现简单，开销小
- 问题
  - 坏消息传播得慢，收敛时间长
    - 网络出现故障，经过比较长时间才能故障信息传送到所有路由器。
  - 限制网络规模，最大距离为15（16表示不可达）。
  - 路由器之间交换的路由信息是路由器的完整路由表。
    - 网络规模扩大，交换开销增加。

# LS(OSPF)和DV(RIP)的比较

## 消息复杂度

- **LS:**  $n$ 个节点,  $e$ 条链路,  $O(e)$  条消息
- **DV:** 消息仅在邻居间传送

## 收敛速度

- **LS:** 接收拓扑更新后收敛
  - 可能存在摆动
- **DV:** 依具体情况而定
  - 可能存在路由回路
  - 坏消息传播慢

**鲁棒性:** 某个路由器故障（或错误估计代价），会发生什么情况？

## LS:

- 结点广播链路失效
- 每个结点重新计算自己的路由表

## DV:

- 每个结点的路由信息均被其他结点使用
- 错误在整个网络中传播

## 外部网关协议 BGP

- BGP 是不同自治系统的路由器之间交换路由信息的协议。
  - 用于自治系统 AS 之间的路由选择。
- 特点
  - 力求选择一条能够到达目的网络且比较好的路由（不能兜圈子），而并非要计算出一条最佳路由。
  - 互联网的规模太大，使得自治系统AS之间路由选择非常困难。
  - 自治系统AS之间的路由选择必须考虑有关策略。
  - 采用路径向量（path vector）路由选择协议。

## eBGP 连接和 iBGP 连接

- eBGP 连接
  - AS 之间， BGP 发言者在半永久性 TCP 连接（端口179）建立 BGP 会话(session)。
  - 运行 eBGP 协议，在不同 AS 之间交换路由信息。
- iBGP 连接
  - AS 内部，任何相互通信的两个路由器之间必须有一个逻辑连接（ TCP 连接）。
  - AS 内部所有的路由器之间的通信是全连通的
  - 运行 iBGP 协议，在 AS 内部的路由器之间交换 BGP 路由信息。

## IGP、iBGP 和 eBGP 的关系

- 在 AS 内部运行：
  - 内部网关协议 IGP（协议 OSPF 或 RIP）。
  - 协议 iBGP。
- 在 AS 之间运行：
  - 协议 eBGP。

# BGP 路由信息

- BGP 路由 = [ 前缀, BGP属性 ] = [ 前缀, AS-PATH, NEXT-HOP ]
- **前缀**：指明到哪一个子网（用 CIDR 记法表示）。
- **BGP 属性**：最重要的两个属性
  - ◆ 自治系统路径 AS-PATH
  - ◆ 下一跳 NEXT-HOP

在属性 AS-PATH 中，不允许出现相同的 AS 号。

## SDN

软件定义网络 SDN

- 是一个体系结构，是一种设计、构建和管理网络的新方法或新概念。
- 把控制层面和数据层面分离，而让控制层面利用软件控制数据层面中的许多设备。
- 优点：
  - 提高网络带宽利用率；
  - 网络运行更加稳定；
  - 管理更加高效简化；
  - 运行费用明显降低。

## 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。

报文种类

- 差错报告报文
- 询问报文

应用举例

- PING
  - 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。
- Traceroute：跟踪一个分组从源点到终点的路径。

## 虚拟专用网 VPN

- 虚拟专用网 VPN
  - 利用公用互联网作为本机构各专用网之间的通信载体的专用网。
  - 虚拟：表示实际上没有使用通信专线，只是在效果上和真正的专用网一样。
- 虚拟专用网 VPN 的构建
  - 如果专用网不同网点之间的通信必须经过公用的互联网，但又有保密的要求，那么所有通过互联网传送的数据都必须加密。

用隧道技术实现虚拟专用网

#### VPN 类型

- 内联网 (intranet)：同一个机构的内部网络所构成的 VPN。
- 外联网 (extranet)：一个机构和某些外部机构共同建立的。
- 远程接入 VPN (remote access VPN)：允许外部流动员工通过接入 VPN 建立 VPN 隧道访问公司内部网络，好像就是使用公司内部本地网络访问一样。

#### 网络地址转换 NAT

- 需要在专用网连接到互联网的路由器上安装 NAT 软件。
- 装有 NAT 软件的路由器叫做 NAT 路由器，它至少有一个有效的外部全球 IP 地址。
- 所有使用本地地址的主机和外界通信时，都要在 NAT 路由器上将其本地地址转换成全球 IP 地址，才能和互联网连接。

## 网络地址转换的过程



- 在内部主机与外部主机通信时，在 NAT 路由器上发生了两次地址转换：
  - 离开专用网时：替换源地址，将内部地址替换为全球地址。
  - 进入专用网时：替换目的地址，将全球地址替换为内部地址。
  - 当 NAT 路由器具有 n 个全球 IP 地址时，专用网内最多可以同时有 n 台主机接入到互联网。
  - 专用网内较多数量的主机轮流使用 NAT 路由器有限数量的全球 IP 地址。

**NAT 地址转换表举例**

方向	字段	旧的 IP 地址	新的 IP 地址
出 (发往互联网)	源 IP 地址	192.168.0.3	172.38.1.5
入 (进入专用网)	目的 IP 地址	172.38.1.5	192.168.0.3

NAT 并不能节省 IP 地址

网络地址与端口号转换 NAPT

网络地址转换特点

- 本地网络只要有一个公网IP就可以接入互联网：
  - 不需要向ISP申请连续的地址块，一个IP满足多个上网设备

- 局域网内可以自由修改网络设备的地址而不通知外部网络
- 自由更换ISP，而不需修改内部网络地址
- 局域网内部设备不需要被网络外主机寻址（即可见），增强了安全性
- 16bit端口号可以支持同时60000个的连接
- 工作在网络层速度快
- 破坏TCP/IP协议族端到端的可寻址的初衷——使用IPV6来解决地址不足的问题。

## IPv6

IPv6 仍支持无连接的传送；协议数据单元 PDU 称为分组

主要变化：

- 更大的地址空间，地址从 IPv4 的 32 位 增大到了 128 位。
- 扩展的地址层次结构，可以划分为更多的层次。
- 灵活的首部格式，定义了许多可选的扩展首部。
- 改进的选项，允许数据报包含有选项的控制信息，其选项放在有效载荷中。
- 允许协议继续扩充，更好地适应新的应用。
- 支持即插即用（即自动配置），不需要使用 DHCP。
- 支持资源的预分配，支持实时视像等要求保证一定的带宽和时延的应用。
- IPv6 首部改为 8 字节对齐，首部长度必须是 8 字节的整数倍。

首部长度的：固定的 40 字节，称为基本首部。首部字段数：只有 8 个。

### IPv6 的地址

- 单播：传统的点对点通信。
- 多播：一点对多点的通信。
- 任播：IPv6 增加的一种类型。任播的终点是一组计算机，但数据报在交付时只交付其中的一个。通常是按照路由算法得出的距离最近的一个。

## 冒号十六进制记法

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

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)：一串连续的零用一对冒号取代。

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

可压缩为：FF05::B3

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

**注意：在任一地址中，只能使用一次零压缩。**

## 点分十进制记法的后缀

- 结合使用点分十进制记法的后缀在 IPv4 向 IPv6 的转换阶段特别有用。
- CIDR 的斜线表示法仍然可用，但取消了子网掩码。
- 例如：60 位的前缀 12AB00000000CD3 可记为：

**12AB:0000:0000:CD30:0000:0000:0000:0000/60**  
或 **12AB::CD30:0:0:0:0/60**（零压缩）  
或 **12AB:0:0:CD30::/60**（零压缩）

**CIDR 记法地址：** 2001:0DB8:0:CD30:123:4567:89AB:CDEF/60，表示：  
**IPv6 的地址是：** 2001:0DB8:0:CD30:123:4567:89AB:CDEF  
**其子网号是：** 2001:0DB8:0:CD30::/60

### IP 多播

- 目的：更好地支持一对多通信。
- 一对多通信：一个源点发送到许多个终点。

### IP 多播

- 在互联网上进行多播就叫做 IP 多播
- 互联网范围的多播要靠路由器来实现。
- 能够运行多播协议的路由器称为多播路由器。
- 多播路由器也可以转发普通的单播 IP 数据报。

### 多播 IP 地址

- 在 IP 多播数据报的目的地址需要写入多播组的标识符。
- 多播组的标识符就是 IP 地址中的 D 类地址（多播地址）。
  - 地址范围：224.0.0.0 ~ 239.255.255.255
- 每一个 D 类地址标志一个多播组。

多播地址只能用于目的地址，不能用于源地址。

### IP 多播需要两种协议

- 网际组管理协议 IGMP
  - 使多播路由器知道多播组成员信息（有无成员）。
- 多播路由选择协议
  - 使多播路由器协同工作，把多播数据报用最小代价传送给多播组的所有成员。

### IGMP 使用 IP 数据报传递其报文

- 在 IGMP 报文加上 IP 首部构成 IP 数据报。



- 但 IGMP 也向 IP 提供服务。
- 因此，不把 IGMP 看成是一个单独的协议，而是整个网际协议 IP 的一个组成部分。

#### **IGMP 工作可分为两个阶段**

- 第一阶段：加入多播组。
- 第二阶段：探询组成员变化情况。

#### **IGMP 采用的一些具体措施，以避免增加大量开销**

- 分散响应。
- 采用抑制机制。

#### **多播路由选择**

转发多播数据报时使用三种方法：

- 洪泛与剪除
- 隧道技术
- 基于核心的发现技术