

CH4 网络层：数据平面

一、绪论

1. 网络层服务

- 在发送主机和接收主机之间传送段(segment)
- 在发送段，将段封装在数据报中
- 在接收端，将段上交给传输层实体
- 网络层协议存在于每一个主机和路由器
- 路由器检查每一个经过它的IP数据报的头部

关键功能：

- **转发**：将分组从路由器的输入接口转发到合适的输出接口（路口）
- **路由**：使用路由算法来决定分组从发送主机到目标主机的路径（规划路程）
 - 路由选择算法
 - 路由选择协议

2. 两个平面

数据平面：

- 路由器的功能
- 转发

控制平面：

- 网络逻辑
- 决定数据报如何路由

3. 网络服务模型

单个数据报：可靠+保证延迟

数据报流：可靠+最小带宽+保证分组之间的延迟差

连接服务区别：

网络层：2个主机之间，涉及到路径上的一些路由器

传输层：2个进程之间，体现在端系统上

二、路由器组成

1. 概况

路由：运行路由选择算法/协议（RIP, OSPF, BGP）生成路由表

转发：交换数据报，根据路由表进行分组转发

2. 输入端口

转发表：（目标地址范围，链接接口）

基于目标的转发： 仅仅依赖于IP数据报的目的IP地址

通用转发： 基于头部字段的任意集合进行转发

最长前缀匹配： 给定目标匹配转发表时，采用最长地址前缀匹配的目标地址表项（使用TCAMs硬件完成）

输入端口缓存： 存储待发送数据报

3. 交换结构

- 将分组从输入缓冲区传输到合适的输出端口
- 交换速率：输入到输出的速率
 - 运行速率经常是输入/输出链路速率的好几倍（倍数等于输入端口个数时不会成为瓶颈）
- 三种交换方式：memory、bus、crossbar

内存交换： 分组拷贝到内存，CPU找到对应输出端口，拷贝到输出端口（第一代路由器，一次仅一个分组）

总线交换： 数据报通过共享总线，从输入端口转发到输出端口（交换速度受限于总线带宽，一次处理一个分组）

互连网络交换： 同时并发转发多个分组，克服总线带宽限制。

4. 输出端口

使用输出端口缓存解决拥塞问题，由调度规则选择排队的数据报进行传输

调度机制： FIFO、优先级调度、Round Robin调度（循环扫描，发送同一类的分组）、Weighted Fair Queuing调度（为类别加权）

丢弃策略： 丢弃刚到达的分组、根据优先级丢失/移除分组、随机丢弃/移除

三、IPv4协议

1. IP数据报

传输TCP段时，20字节的TCP、20字节的IP

IP分片与重组： 将大的IP数据报分片为带有IP头部的小数据报，它们拥有相同的ID和不同的偏移量，将最后一个分片标记为0，在目标主机进行重组。

MTU = 网路链路最大传输单元

MTU = 1500bytes时，(20 + 3980)字节的IP数据包被分为以下三条小IP数据报：

	length =1500	ID =x	fragflag =1	offset =0	
	length =1500	ID =x	fragflag =1	offset =185	
	length =1040	ID =x	fragflag =0	offset =370	

$185 = ((1500 - 20) * 1) / 8$, $370 = ((1500 - 20) * 2) / 8$

2. IP编址

IP地址：32位标示，对主机或者路由器的接口编址，高位为子网部分，低位为主机部分

接口：主机/路由器和物理链路的连接处

一个IP和一个接口相关联（接口之间的连接手段属于数据链路层）

子网：一个子网内的节点IP地址高位部分相同，且无需路由器接入，子网内各个主机可以在物理上相互通达

子网个数判断：将每个接口从主机或者路由器上断开，剩下的每个都是一个子网

IP地址分类：

A类（126个网络， 2^{24} 个主机）【1.0.0.0 - 127.255.255.255】

B类（ 2^{14} — 2个网络， 2^{16} 个主机）【128.0.0 - 191.255.255.255】

C类（ 2^{11} — 2个网络，254个主机）【192.0.0.0 - 223.255.255.255】

D类（组播）【224.0.0.0 - 239.255.255.255】

E类（为未来使用）【240.0.0.0 - 247.255.255.255】

约定：0.0.0.0本主机，本网络；1.1.1.1广播地址，这个网络的所有主机

内网（专用）IP：永远不会被当成公网IP来分配，不会与公用地址重复，只在局部网络中有意义，区分不同的设备，路由器不对目标是专用地址的分组进行转发。

A类：10.0.0.0 - 10.255.255.255/8

B类：172.16.0.0 - 172.31.255.255/16

C类：192.168.0.0 - 192.168.0.0/24#分可以在任意的位置

a.b.c.d/x, x为子网号的长度。

子网掩码：1：表示该位置为子网部分，0：表示该位置为主机部分。/x表示IP地址中前x位为子网地址

3. 转发表和转发算法

schema：（目的子网号，掩码，下一条，接口）【有一条Default项】

目的：获取IP数据报的目标地址

对于转发表中的每一个表项，如果 $(\text{DestAddr} \& \text{Mask}) == \text{DestSubNet}$ ，按照该接口转发该数据报；都没有，命中default项

4. DHCP

DHCP（Dynamic Host Configuration Protocol）：从服务器中动态获取一个IP地址

目标：允许主机在加入网络的时候，动态地从服务器那里获得IP地址

可以更新对主机在用IP地址的租期

重新启动时，允许重新使用以前用过的IP地址

支持移动用户加入该网络

DHCP工作状态：

- 主机广播DHCP discover报文【optional】
- DHCP服务器用DHCP offer提供报文响应【optional】
- 主机请求IP地址：发送DHCP request报文
- DHCP服务器发送地址：DHCP ack报文

DHCP返回IP地址、第一跳路由器的IP地址（默认网关）、DNS服务器的域名和IP地址、子网掩码

请求过程：

1. 主机需要获取自己的IP地址，第一跳路由器地址和DNS服务器：DHCP
2. DHCP请求封装在UDP段中，封装在IP数据报中，封装在以太网的帧中
3. 以太网帧在局域网范围内广播(dest = fffffff), 被运行的DHCP服务的路由器收到
4. 以太网帧解封装为IP，IP解封装为UDP，解封装为DHCP
5. DHCP服务器生成DHCP ACK，包含主机需要的三个信息
6. DHCP服务器封装的报文所在的帧转发到客户端，在客户端解封装为DHCP报文
7. 客户端成功获取所需信息

获取子网部分：从ISP获得地址块中分配的一个小地址块

5. 层次编址

200.23.16.00 → 200.23.16.0, 200.23.18.0, 200.23.30.0

199.31.0.0 → 199.31.0.0

ISP获得地址块：ICANN（Internet Corporation for Assigned Names and Numbers）

- 分配地址
- 管理DNS
- 分配域名，解决冲突

6. NAT

NAT（Network Address Translation）

所有主机离开本地网络的数据包具有一个相同的源地址，但是有不同的端口号

好处：

- 不需要从**ISP**分配一块地址，可用一个**IP**地址用于所有的（局域网）设备--省钱
- 可以在局域网改变设备的地址情况下而无须通知外界
- 可以改变**ISP**（地址变化）而不需要改变内部的设备地址
- 局域网内部的设备没有明确的地址，对外是不可见的--安全

实现：

将发出去的数据包中的源地址和端口号替换为NAT IP地址和新端口号，目的地址和目的端口号不变

使用NAT转换表来保存（源IP，源端口）↔（NAT IP，新端口）的映射

将收到的数据包中的（NAT IP，新端口）替换为（源IP，源端口）

地址翻译：16位的端口字段

NAT协议没有遵守end-to-end原则（将复杂性放到网络边缘），导致外网机器可能无法连接到内网的机器上

解决NAT穿越问题：

静态配置NAT（总是转发到服务器的特定端口）

UPnP Internet Gateway协议：允许获知网络的公网IP、列举存在的端口映射、增/删端口映射

中继：NAT后面的服务器建立和中继的连接，外部的客户端链接到中继，中继在两个连接之间桥连

四、IPv6

1. 概述

32位的地址空间将会被很快用完，头部格式改变来加速处理和转发

IPv6数据报：固定的40字节头部，数据报传输过程中不允许分片

和IPv4的变化：移除了checksum，将options移至头部之外，使用ICMPv6

2. IPv4到IPv6

隧道：在IPv4路由器之间传输的IPv4数据报中携带IPv6数据报