4.1 网络层概述

- 网络层的主要任务是**实现网络互连**，进而**实现数据包在网络之间的传输**
- 要实现网络层任务，需要解决以下问题：
    - **网络层向运输层提供怎样的服务**（“可靠传输”还是“不可靠传输”）
    - **网络层寻址问题**
    - **路由选择**
- **因特网（Internet）**是目前全世界用户数量最多的互联网，它使用**TCP/IP协议栈**
- 由于TCP/IP协议栈的网络层使用**网际协议IP**，它是整个协议栈的核心协议，因此在TCP/IP协议栈中网络层常称为**网际层**
- 主要通过学习TCP/IP协议栈的网际层来学习网络层的理论知识

4.2 网络层提供的两种服务

对比	面向连接的虚电路服务	无连接的数据报服务
思路	可靠通信应当由网络来保证	可靠通信应由用户主机来保证
连接的建立	必须建立网络层连接	不需要建立网络层连接
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	每个分组可走不同路由
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
分组的排序	总是按发送顺序到达终点	到达终点时不一定按发送顺序
服务质量的保证	可以将通信资源提前分配给每一个虚电路，容易实现	很难实现

4.3.1 IPv4地址概述

- 在TCP/IP体系中，IP地址是一个最基本的概念，必须弄清楚
- **IPv4地址**就是给因特网（Internet）上的**每一台主机（或路由器）的每一个接口**分配一个在全世界范围内

唯一的32比特的标识符

- IPv4地址的编址方法经历了三个历史阶段：
    - 1981 分类编址
    - 1985 划分子网
    - 1993 无分类编址
- 32 比特的IPv4地址不方便阅读、记录等，因此IPv4地址采用**点分十进制表示方法**以便用户使用

4.3.2 分类编制的IPv4地址

网络类别	第一个可指派的网络号	最后一个可指派的网络号	最大可指派的网络数量	每个网络中的最大主机数量	不能指派的网络号	占总地址空间
A	1	126	126 ($2^{(8-1)}-2$)	16777214 ($2^{24}-2$)	0和127	1/2 ($2^{(32-1)}/2^{32}$)
B	128.0	191.255	16384 ($2^{(16-2)}$)	65534 ($2^{16}-2$)	无	1/4 ($2^{(32-2)}/2^{32}$)
C	192.0.0	223.255.255	2097152 ($2^{(24-3)}$)	254 ($2^{8}-2$)	无	1/8 ($2^{(32-3)}/2^{32}$)

网络类别	作 用	第一个地址	最后一个地址	地址数量	占总地址空间
D	多播地址	224.0.0.0	239.255.255.255	268435456 (2^{28})	1/16 ($2^{(32-4)}/2^{32}$)
E	保留为今后使用	240.0.0.0	255.255.255.255	268435456 (2^{28})	1/16 ($2^{(32-4)}/2^{32}$)

一般不使用的特殊IP地址				
网络号	主机号	作为源地址	作为目的地址	代表的意思
0	0	可以	不可	在本网络上的本主机（DHCP协议）
0	host-id	可以	不可	在本网络上的某台主机host-id
全1	全1	不可	可以	只在本网络上进行广播（各路由器均不转发）
net-id	全1	不可	可以	对net-id上的所有主机进行广播
127	非全0或全1	可以	可以	用于本地软件环回测试

4.3.3 划分子网的IPv4地址

- 为新增网络申请新的网络号会带来以下弊端：
    - 需要等待时间和花费更多的费用
    - 会增加其他路由器中路由表记录的数量
    - 浪费原油网络号中剩余的大量IP地址
- 可以从主机号部分借用一部分比特作为子网号
- **32比特的子网掩码可以表明分类IP地址的主机号部分被借用了几个比特作为子网号**
    - 子网掩码使用**连续的比特 1 来对应网络号和子网号
    - 子网掩码使用**连续的比特 0 来对应主机号
    - 将划分子网的**IPv4地址**与其相应的**子网掩码**进行**逻辑与运算**就可以得到IPv4地址**所在子网的网络地

址

- 给定一个分类的IP地址和其相应的子网掩码，就可知道子网划分的细节：**【重点】**
    - 划分出子网的数量
    - 每个子网可分配的IP地址数量
    - 每个子网的网络地址和广播地址
    - 每个子网可分配的最小和最大地址
- **默认的子网掩码是指在未划分子网的情况下使用的子网掩码** A类：255.0.0.0 B类：255.255.0.0 C类：255.255.255.0

4.3.4 无分类编址的IPv4地址

- 划分子网在一定程度上缓解了因特网在发展中遇到的困难，但是**数量巨大的C类网**，因为其**空间太小**并**没有得到充分使用**，而因特网的IP地址在加速消耗，整个**IPv4地址空间面临全部耗尽的威胁**
- 为此，提出**无分类编址**的方式来解决此问题
- 1993年，IETF发布**无分类域间路由选择CIDR**（Classless Inter- Domain Routing）的RFC文档：
    - **CIDR消除了传统的A类、B类和C类地址，以及划分子网的概念**
    - **CIDR可以更加有效的分配IPv4的地址空间**，并且可以在新的IPv6使用之前允许因特网的规模持续增长
- CIDR使用“**斜线记法**”，或称CIDR法，在IPv4地址后面加上斜线“/”，在**斜线后面写上网络前缀所占的比特数量**
- CIDR实际上是**将网络前缀都相同的连续的IP地址组成一个“CIDR”地址块**
- 只要知道CIDR地址块中的任何一个地址，就可以知道该地址块的全部细节：
    - 地址块最小地址
    - 地址块的最大地址
    - 地址块中的地址数量
    - 地址块聚合某类网络（A、B、C类）的数量
    - 地址掩码（子网掩码）
- **路由聚合**（构造超网）的方法是找**共同前缀**
- **网络前缀越长，地址块越小，路由越具体**
- 若路由器查表转发时发现有多条路由可选，则选择网络前缀最长的那条，这称为**最长匹配**，因为这样的路由更具体

4.3.5 IPv4地址的应用规划

- 定长的子网掩码FLSM（Fixed Length Subnet Mask）：
    1. 使用同一个子网掩码来划分子网

2. 子网划分方式不灵活，只能划分出2^n 个子网（n是从主机号部分借用的用来作为子网号的比特数量）
3. 每个子网所分配的IP地址数量相同，容易造成IP地址浪费

- 变长的子网掩码VLSM（Variable Length Subnet Mask）：

1. 使用不同的子网掩码来划分子网
2. 子网划分方式灵活，可以按需分配
3. 每个子网所分配的IP地址数量可以不同，尽可能减少对IP地址的浪费

4.4 IP数据报的发送和转发过程

- 主机发送IP数据报：
    - 判断目的主机是否与自己在同一个网络：
        - 若在，则属于**直接交付**
        - 若不在，则属于**间接交付**，传送给主机所在网络的**默认网关**（路由器），由默认网关帮忙转发
- 路由器转发IP数据报：
    - 检查IP数据报首部是否出错：
        - 若出错，则直接丢弃，并通报给源主机
        - 若没有，则进行转发
    - 根据IP数据报的目的地址在路由表中查找匹配的条目：
        - 若找到匹配的条目，则转发给条目中指示的下一跳
        - 若找不到，则丢失，并通报给源主机

4.5 静态路由配置及其可能产生的路由环路问题

- 静态路由配置是指用户或网络管理人员使用路由器的相关命令给路由器**人工配置路由表**
    - 这种人工配置方式简单、开销小。但**不能及时适用网络状态（流量、拓扑等）的变化**
    - 一般只在小规模网络中采用
- 使用静态路由配置可能出现以下**导致产生路由环路**的错误：
    - 配置错误
    - 聚合了不存在的网络
    - 网络故障
- 路由条目的类型：
    - 直连网络

- 静态路由（人工配置）
- 动态路由（路由选择协议）

- 特殊的静态路由条目

  - 默认路由（目的网络为0.0.0.0，地址掩码为0.0.0.0）
  - 特定主机路由（目的网络为特定主机的IP地址，地址掩码为255.255.255.255）
  - 黑洞路由（下一跳为null0）

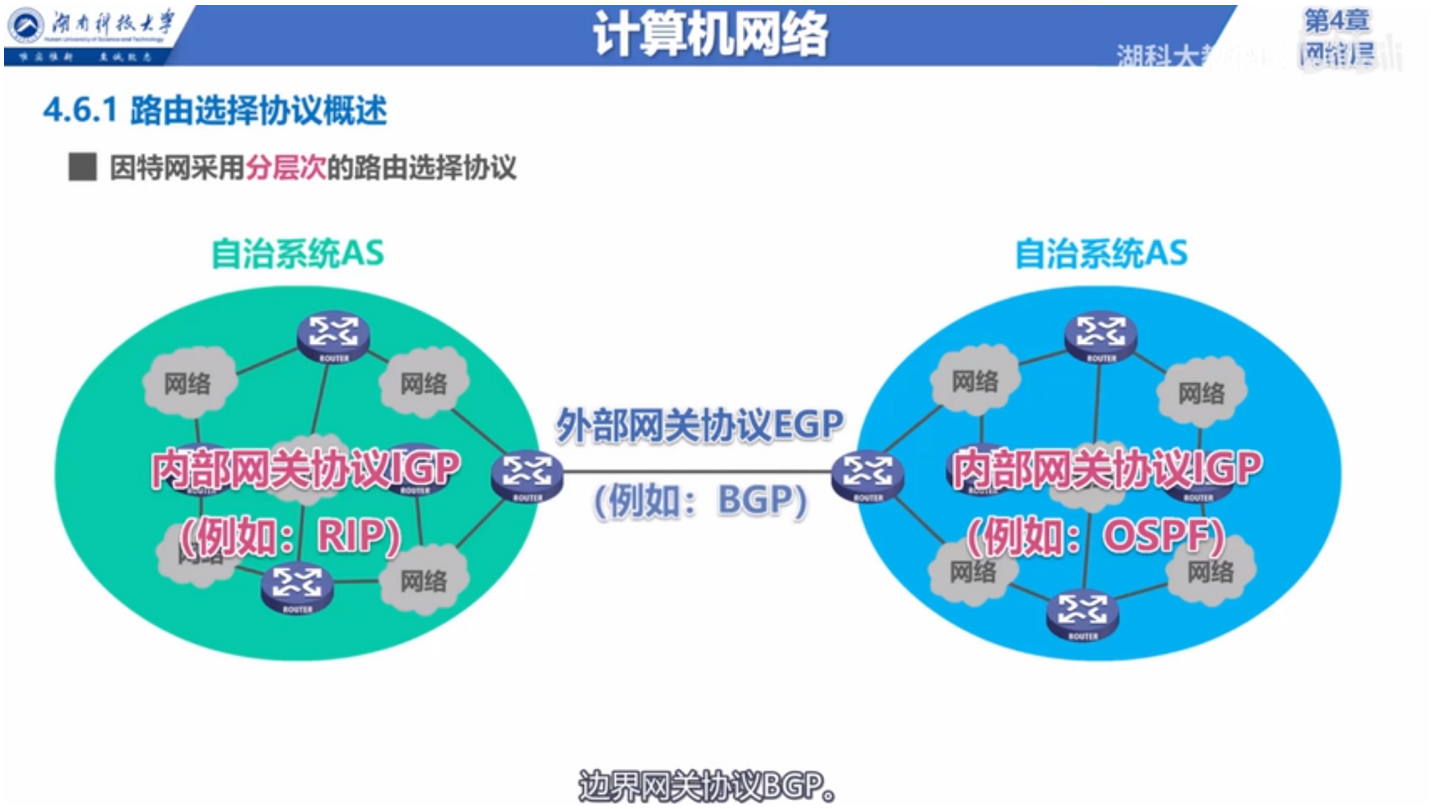
4.6.1 路由选择协议概述

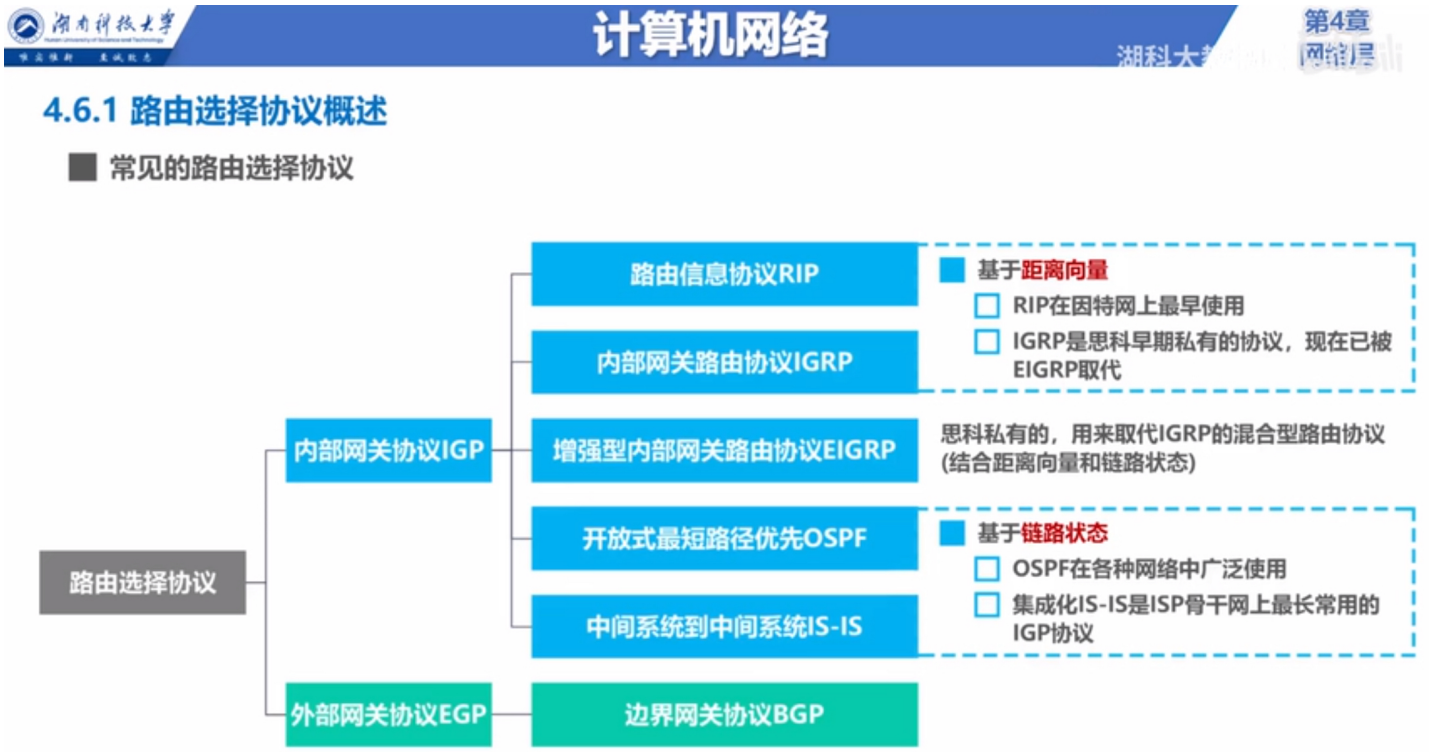
- 静态路由选择：

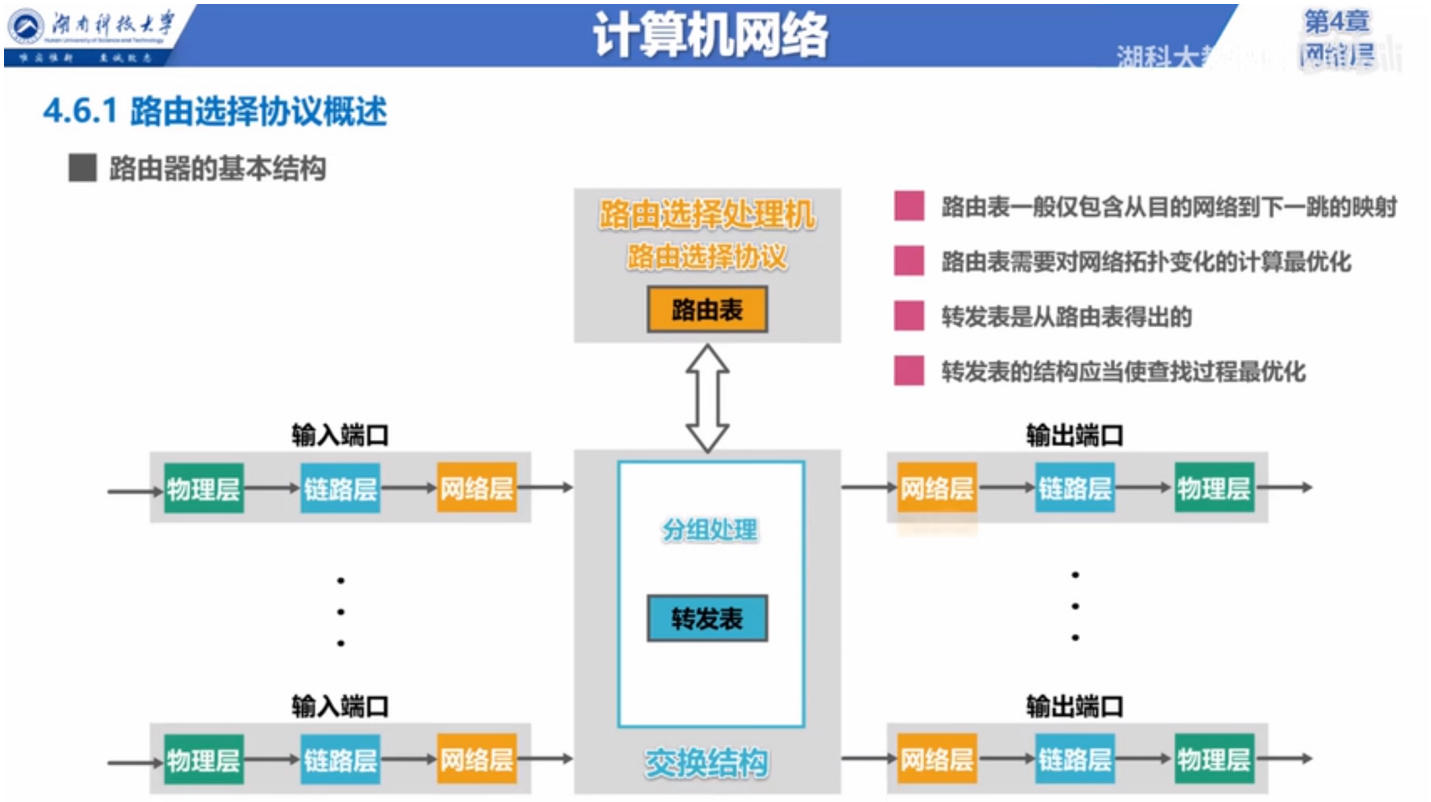
  - 由**人工配置**的网络路由、默认路由、特定主机路由、黑洞路由等都属于静态路由
  - 这种人工配置方式简单、开销小。但**不能及时适用网络状态（流量、拓扑等）的变化**
  - 一般只在小规模网络中采用

- 动态路由选择：

  - 路由器通过路由选择协议**自动获取路由信息**
  - 比较复杂、开销较大。**能较好地适应网络状态的变化**
  - 适用于**大规模网络**







4.6.2 路由信息协议RIP的基本工作原理

- **路由信息协议RIP**（Routing Information Protocol）是内部网关协议IGP中最先得到广泛使用的协议之一，其相关文档为 RFC 1058
- RIP要求自治系统AS内的每一个路由器都要维护从它自己到AS内其他每一个网络的距离记录。这是一组距离，称为“**距离向量**”

- RIP使用**跳数**（Hop Count）作为度量**来衡量到达目的网络的距离**
    - 路由器到直连网络的距离定义为1
    - 路由器到非直连网络的距离定义为所经过的路由器数加1
    - 允许一条路径最多只能包含15个路由器。**距离等于16时相当于不可达**，因此，**RIP只适用于小型网络**
- RIP认为**好的路由**就是“**距离短**”的路由，也就是**所通过路由器数量最少的路由**
- 当到达同一目的网络有多条“距离相等”时，可以进行“**等价负载均衡**”
- RIP包含以下三个要点：
    - **和谁交换信息** 仅和**相邻路由器**交换信息
    - **交换什么信息** 自己的**路由表**
    - **何时交换信息 周期性交换**（例如每30秒）
- **RIP的基本工作过程**：
    1. 路由器刚开始工作时，只知道自己到达直连网络的距离为1
    2. 每个路由器仅和相邻路由器周期性交换并更新路由信息
    3. 若干次交换和更新后，每个路由器都知道到达AS内网络的最短距离和下一跳地址，称为收敛
- **RIP的路由条目的更新规则**：
    - 发现了新的网络，添加
    - 到达目的网络，相同下一跳，最新消息，更新
    - 到达目的网络，不同下一跳，新路由优势，更新
    - 到达目的网络，不同下一跳，新路由劣势，不更新
    - 达到目的网络，不同下一跳，等价负载均衡
- RIP存在**坏消息传播得慢**的问题
- “坏消息传播得慢”，又称为**路由环路**或**距离无穷计数**问题，这是距离向量算法的一个固有问题。可以采取多种措施**减少**该问题的概率或减小该问题带来的危害
    - **限制最大路径距离为15**（16表示不可达）
    - 当路由表发生变化时，就立即发送更新报文（即“**触发更新**”），而不仅是周期性发送
    - 让路由器记录收到某特定路由信息的接口，而不是让同一路由信息再通过此接口向反向传送（即**水平分割**）

4.6.3 开放最短路径优先OSPF的基本工作原理

- 开放最短路径优先OSPF（Open Shorter Path First），是为克服RIP的缺点在1989年开发出来的
    - “开放”表明OSPF协议不是受一家厂商控制的，而是**公开发表**的
    - “最短路径优先”是因为使用了Dijkstra提出的**最短路径算法**SPF

- OSPF是基于**链路状态**的，而不像RIP那样是基于距离向量的
- OSPF采用SPF算法计算路由，从算法上保证了**不会产生路由环路**
- OSPF**不限制网络规模**，更新效率更高，**收敛速度更快**
- 链路状态是指本路由器都**和哪些路由器相邻**，以及相应**链路的“代价”**
    - “代价”用来表示费用、距离、时延、带宽等等，都由网络管理人员来决定
- 使用OSPF的每个路由器都会产生**链路状态通告LSA**（Link State Advertisement）L SA包含以下内容：
    - 直连网络的链路状态信息
    - 邻居路由器的链路状态信息
- LSA被封装在**链路状态更新分组LSU中**，采用**洪泛**法发送
- 通过各路由器洪泛发送封装有自己LSA的LSU分组，各路由器的LSDB最终将达到一致
- 使用OSPF的各路由器基于**LSDB进行最短路径优先SPF计算**，构建出各自到达其他路由器的最短路径，即构建各自的路由表
- OSPF有以下五种分组类型：
    - **问候**（hello）分组
    - **数据库描述**（Database Description）分组
    - **链路状态请求**（Link State Request）分组
    - **链路状态更新**（Link State Update）分组
- OSPF在多点接入网络中路由器邻居关系的建立
    - 选举**制定路由器DR**（designated router）和**备用的指定路由器BDR**（backup designated router）
    - **所有的非DR/BDR只与DR/BDR建立邻居关系**
    - 非DR/BDR之间通过DR/BDR交换信息
- 为了使OSPF能够用于规模很大的网络，OSPF把一个自治系统再划分为若干个更小的范围，叫做**区域**（Area）
    - 划分区域的好处就是**把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自治系统**，这样就减少了整个网络上的通信量

4.6.4边界网关协议BGP的基本工作原理

- 外部网关协议EGP（例如边界网关协议BGP）
    - 在不同自治系统内，度量路由的“代价”（距离、带宽、费用等）可能不同。因此，对于自治系统之间的路由选择，使用“代价”作为度量来寻找最佳路由是不行的
    - 自治系统之间的路由选择必须考虑相关策略（政治、经济、安全等）
    - BGP只能是力求寻找一条能够到达目的网络且比较好的路由（不能兜圈子），而并非要找一条最佳路由
- 在配置BGP时，每个自治系统的管理人员要选择至少一个路由器作为该自治系统的“**BGP发言人**”
- 不同自治系统的BGP发言人要交换路由信息，首先必须建立**TCP连接**，端口号179

    - 在此TCP连接上机哦啊还BGP报文以建立**BGP会话**
    - 利用BGP会话**交换路由信息**（例如，增加新的路由，或撤销过时的路由，以及报告出错的情况等）
    - 使用TCP连接交换路由信息的两个BGP发言人，彼此称为对方的**邻站**（neighbor）或**对等站**（peer）

- BGP发言人除了运行BGP外，还必须运行自己所在自治系统所使用的内部网关协议IGP，例如OSPF或RIP
- BGP发言人**交换网络可达性的信息**（要到达某个网络所要经过的一系列自治系统）
- 当BGP发言人互相交换了网络可达性的信息后，各BGP发言人就**根据**所采用的**策略**从收到的路由信息中**找出到达各自治系统的较好的路由**。也就是构造出树形结构，**不存在回路的自治系统连通圈**
- BGP适用于多级结构的因特网
- BGP-4有以下四种报文：
    - **OPEN（打开）报文**：用来与相邻的另一个BGP发言人建立关系，使通信初始化
    - **UPDATE（更新）报文**：用来通告某一路由的信息，以及列出要撤销的多条路由
    - **KEEPALIVE（保活）报文**：用来周期性地证实邻站的连通性
    - **NOTIFICATION（通知）报文**：用来发送检测到的差错

4.7 IPv4数据报的首部格式

- 版本 占4比特，表示IP协议的版本。通信双方使用的IP协议版本必须一致，目前广泛使用IPv4
- 首部长度 占4比特，表示IP数据报首部的长度，该字段的取值以4字节为单位
    - 最小十进制取值为5，表示IP数据报首部只有20字节固定部分
    - 最大十进制取值为15，表示IP数据报首部包含20字节固定部分和最大40字节可变部分

- 可选字段 长度从1个字节到40个字节不等，用来支持拍错，测量及安全等措施
    - 可选字段增加了IP数据报的功能，但这同时也使得IP数据报的首部长度成为可变的，这就增加了每一个路由器处理IP数据报的开销。实际上可选字段很少被使用

- 填充字段 确保首部长度为4字节的整数倍，使用全0进行填充
- 区分服务 占8比特，用来获得更好的服务。只有在使用区分服务时，该字段才起作用，一般情况下不使用
- 总长度 占16比特，表示IP数据报的总长度（首部+数据载荷）。最大取值为十进制的65535，以字节为单位
- 标识 占16比特，属于同一个数据报的各分片数据报应该具有相同的标识。IP软件维持一个计数器，每产生一个数据报，计数器值加1，并将此值赋给标识字段
- 标志 占3比特，指出分片数据报的数据载荷部分偏移在原书记报的位置有多少个单位。片偏移以8个字节为单位
- 生存时间TTL 占8比特，表示IP数据报的生存时间
    - 最初以秒为单位，最大生命周期为255秒；路由器转发IP数据报时，将IP数据报首部中的该字段的值减去IP数据报在本路由器上所耗费的时间，若不为0就转发，否则就丢弃。现在以“跳数”为单位，路由器转发IP数据报时，将IP数据报首部中的该字段值减1，若不为0就转发，否则就丢弃
    - IP数据报每经过一个路由器，路由器都要重新计算首部检验和因为某些字段（生存时间、标志、片

偏移等）的取值可能发生变化

- 协议 占8比特，指明IPv4数据报的数据部分是何种协议数据单元。常用的一些协议和相应的协议字段值如下

协议名称	ICMP	IGMP	TCP	UDP	IPv6	OSPF
协议字段值	1	2	6	17	41	89

4.8 网际控制报文协议ICMP

- 为了有效地转发IP数据报和提高交付成功的机会，在网际层使用了网际控制报文协议ICMP（Internet Control Message Protocol）
- 主机或路由器使用ICMP来发送**差错报告报文**和**询问报文**
- **ICMP报文被封装在IP数据报**中发送
- ICMP差错报告报文共有以下五种：
  - 终点不可达
  - 源点抑制
  - 时间超过
  - 参数问题
  - 改变路由（重定向）
- 以下情况**不应发送ICMP差错报告报文**：
  - 对ICMP差错报告报文不再发送ICMP差错报告报文
  - 对第一个分片的数据报片的所有后续数据报片都不发送ICMP差错报告报文
  - 对具有多播地址的数据报都不发送ICMP差错报告报文
  - 对具有特殊地址（127.0.0.0或0.0.0.0）的数据报不发送ICMP差错报告报文
- 常用的**ICMP询问报文**有以下两种：
  - 回送请求和回答
  - 时间戳请求和回答
- ICMP应用：
  - 分组网间探测PING
  - 跟踪路由traceroute

4.9虚拟专用网VPN与网络地址转换NAT

虚拟专用网VPN（Virtual Private Network）

- **利用公用的因特网作为本机构各专用网之间的通信载体**，这样的专用网又称为虚拟专用网

- 同一机构内不同部门的内部网络所构成的虚拟专用网VPN又称为**内联网VPN**
- VPN要保证传输数据的安全性，会将原始的**内部数据报进行加密**，然后再将其封装成为在因特网上发送到的外部数据报
- 有时一个机构的VPN需要有某些外部机构（通常就是合作伙伴）参加进来，这样的VPN就称为**外联网VPN**
- 在外地工作的员工需要访问公司内部的专用网络时，只要在任何地点接入因特网，进行驻留在员工PC中的VPN软件，在员工PC和公司的主机之间建立VPN隧道，即可访问专用网络中的资源。这种VPN称为**远程接入VPN**

网络地址转换NAT（Network Address Translation）

- 由于IP地址的紧缺，一个机构能够申请到的IP地址数量往往远小于本机构所拥有的主机数量。因此，**虚拟专用网中的各主机所分配的IP地址应该是本机可自由分配的专用地址**，而不是需要申请的，在因特网上使用的公有地址
- 虽然因特网采用了无分类编制方式来减慢IP地址空间耗尽的速度，但由于因特网用户量的激增，IPv4地址空间即将面临耗尽的危险依然没有解除
- 1994年提出了一种网络地址转换**NAT的方法再次缓解了IP地址空间耗尽的问题**
- NAT能**使大量使用内部专用地址的专用网络用户共享少量外部全球地址**来访问因特网上的主机和资源
- 由于绝大多数的网络应用都是使用运输层协议TCP或UDP来传送数据，因此可以**利用运输层的端口号和IP地址一起进行转换**。这样，使用**一个全球IP地址就可以使用多个拥有本地地址的主机同时和因特网上的主机进行通信**。这种将断口号和IP地址一起进行转换的技术叫做**网络地址与端口号转换NAPT**（Network Address and Port Translation）
- 对于一些P2P网络应用，需要**外网主机主动与内网主机进行通信，在通过NAT时会遇到问题**，需要网络应用字节使用一些特殊的NAT穿越技术来解决问题
- 由于**NAT对外网屏蔽了内网主机的网络地址，能为内网的主机提供一定的安全保护**