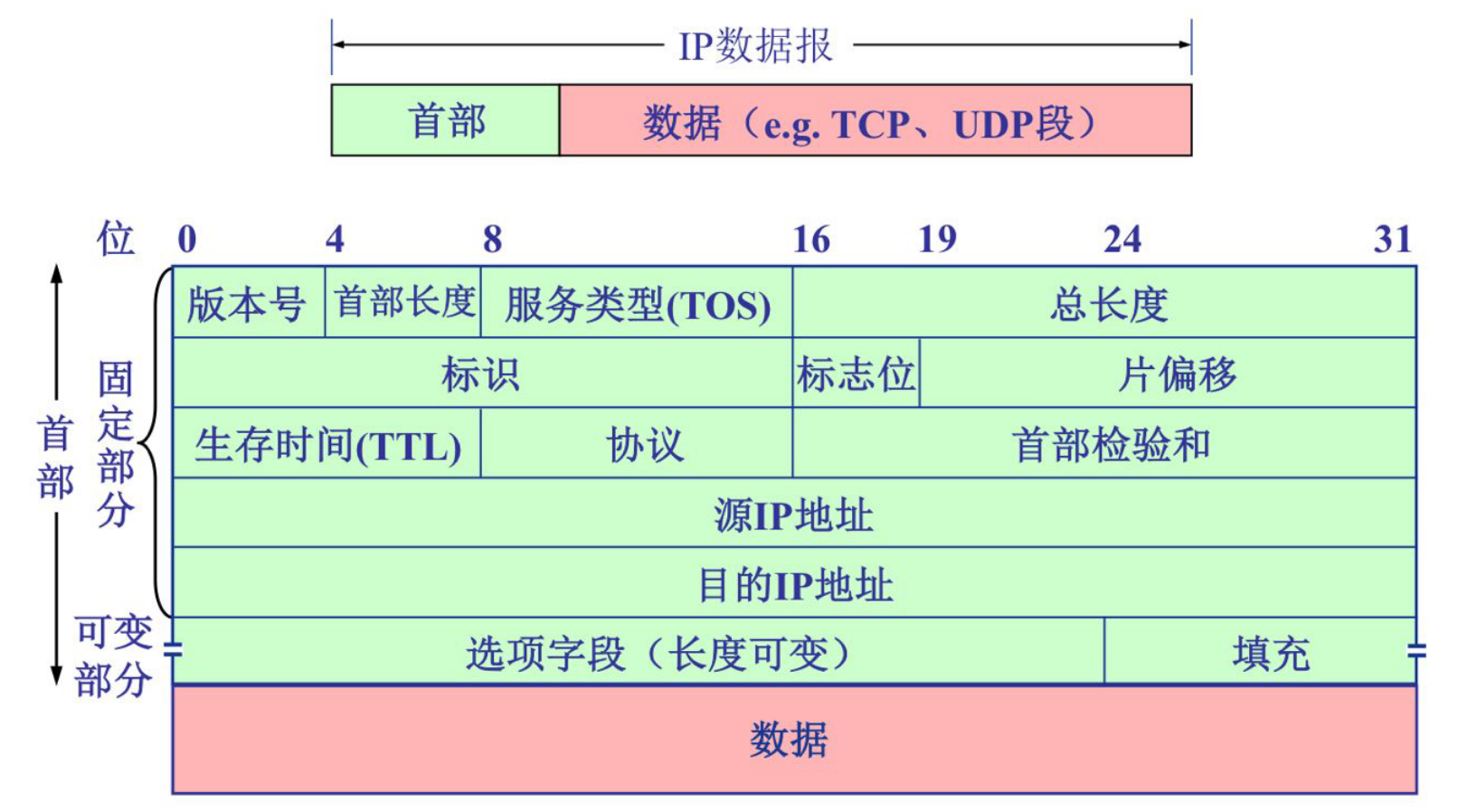
# 3.计算机网络-网络层

1. 网络层服务
   1. 网络层
      1. 从发送主机向接收主机传送数据
      2. 每个主机和路由器都运行网络层协议
      3. 路由器检验所有穿越它的IP数据报的头部域，决策如何处理这个数据报
   2. 网络层核心功能：转发与路由
      1. 路由算法确定通过网络的端到端路径，转发表确定在本路由器如何转发分组
   3. 某些网络的网络层重要功能：连接建立
      1. 数据传输前两端主机建立虚拟/逻辑连接，由网络设备参与连接的建立
      2. 传输层同样有连接建立功能，但传输层连接是两个应用进程间，对中间网络设备透明
   4. 网络层服务模型
      1. 无连接服务
         1. 不事先为系列分组确定传输路径，每个分组独立确定传输路径，不同分组传输路径可能不同
         2. 数据报网络
      2. 连接服务
         1. 首先为系列分组确定传输路径，沿该路径传输分组，各分组传输路径相同，传输后拆除连接
         2. 虚电路网络
   5. 拥塞控制：分组丢失/延迟过大
      1. 网络层（辅助）拥塞控制：路由器向发送方显式反馈拥塞信息，指示发送方应该采取何种速率
      2. 传输层端到端拥塞控制：端系统通过观察丢包等行为判断是否发生拥塞（TCP采取这种方法）
      3. 拥塞控制策略：流量感知路由、准入控制、流量调节、负载脱落等
2. 虚电路网络与数据报网络
   1. 连接服务与无连接服务
      1. 数据报网络：网络层无连接服务
      2. 虚电路网络：网络层连接服务
      3. 两者类似于传输层的UDP与TCP，但是网络层服务是主机到主机服务，由网络核心实现
   2. 虚电路（VC）：源主机到目的主机，类似电路的路径，是一种逻辑连接
      1. 过程：呼叫建立-数据传输-拆除呼叫
      2. 分组交换，每个分组利用链路全部带宽（独占）
      3. 源到目的经过的网络层设备共同完成虚电路功能
      4. 每个分组携带虚电路标识，虚电路经过的每个网络设备维护每条经过它的虚电路连接状态
      5. 链路、网络设备资源可以面向VC进行预分配
      6. 每条虚电路包括：
         1. 从源主机到目的主机的一条路径
         2. 虚电路号（VCID）：沿路每段链路一个编号
         3. 沿路每个网络层设备利用转发表记录经过的每条虚电路
      7. 同一条VC在每段链路上的VCID通常不同
      8. 虚电路信令协议
         1. 用于VC的建立、维护与拆除（路径选择），目前的Internet不采用
   3. 数据报网络
      1. 网络层无连接
      2. 每个分组携带目的地址，路由器根据分组的目的地址转发分组
         1. 基于路由协议/算法构建转发表，检索转发表，每个分组独立选路
      3. 数据报转发表：给出目的地址和链路接口之间的映射
      4. 最长前缀匹配优先：在检索转发表时，优先选择与分组目的地址匹配前缀最长的入口
   4. Internet没有严格时间需求，采用数据报网络，简化网络而复杂边缘
   5. ATM核心业务是实时对话，采用VC网络，简化边缘而复杂网络
3. 路由器体系结构
   1. 路由器的两个重要功能：运行路由算法/协议；从输入链路转发数据报到输出链路
   2. 输入端口排队：输入缓冲区溢出导致排队延迟和丢失
   3. 交换结构：将数据报从输入缓冲区传输到适当的输出缓冲区
      1. 经内存交换：第一代路由器，数据报被复制到系统内存中，速度受限
      2. 经总线交换：输入端口经一根总线将数据报传送到输出端口，可能出现总线争用
      3. 经网络交换（纵横式）：克服了单一总线带宽的设计
   4. 输出端口：同样存在输出端口排队
   5. 缓存如何设置？使用N个流，链路带宽为C，缓存大小等于：

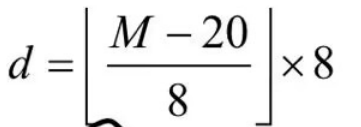


* 1. 调度机制
     1. FIFO：先进先出
     2. 优先级调度：设置优先级，选优先级最高的数据包发送
     3. 轮询调度：轮询扫描类队列，从每个类发送一个完整的数据报
     4. 加权公平队列：广义轮询，为每个类在周期中获得加权的服务量

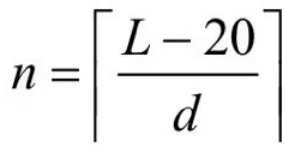
1. IP协议
   1. IP数据报格式



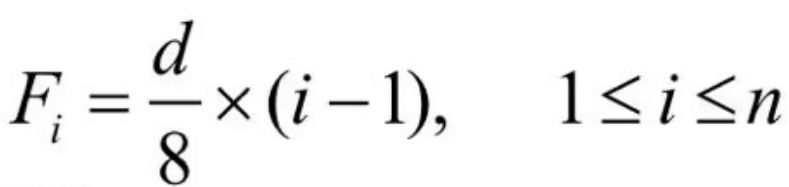
* + 1. 版本号（4位）：4-IPv4；6-IPv6
    2. 首部长度（4位）：以4字节为单位（5代表首部长度为20B）
    3. 服务类型（8位）：期望获得哪种类型服务，一般不使用
    4. 总长度（16位）：IP分组总**字节数**（首部+数据）
       1. 最大的IP分组总长度为65535B，最小IP分组首部为20B
    5. 标识（16位）：标识一个IP分组，用于分片
       1. IP协议利用计数器，每产生一个分组加一，作为该分组的标识
    6. 标志位（3位）：**MF | DF | 保留位（低|中|高）**
       1. DF（Don't Fragment）：取1则禁止分片，取0则允许分片
       2. MF（More Fragment）：取1则非最后一片，取0则是最后一片（或未分片）
    7. 片偏移（13位）：IP分组分片封装原IP分组数据的相对偏移量，以8字节为单位
       1. 没分片则该字段一定是0，分片则**一定是8字节的倍数（最后一片除外）**
    8. TTL（8位）：IP分组在网络中可以通过的路由器数（或跳步数），降为0时分组被丢弃
    9. 协议（8位）：IP分组封装的上层协议类型（TCP，UDP），实现复用/分解
    10. 首部校验和（16位）：对首部进行差错检测，计算时置全0，逐跳计算，逐跳校验
    11. 源/目的IP地址（各32位）
    12. 选项字段（长度可变，1-40B）：安全、源选路径、时间戳等，实际很少使用
    13. 填充字段（长度可变，0-3B）：补齐首部，使得32位对齐，保证首部长度是4B的倍数
  1. MTU（最大传输单元）：链路层数据帧可封装数据的上限，不同链路MTU不同
  2. IP分片与重组：大IP分组向小MTU链路转发时可以被分片（不是强制的）
     1. 分片后IP分组到达目的主机后被重组
     2. IP首部的总长度、标识、标志位、片偏移等字段用于标识分片及确定分片的相对顺序
  3. IP分片过程
     1. 假设IP分组总长度为L，待转发链路的MTU为M，若L>M，且DF=0，则可以/需要分片
     2. 分片时每个分片的标识复制原分组的标识
     3. 通常除最后一个分片外其它分片均分为MTU允许的最大分片
     4. 一个最大分片可封装的数据应该是8的倍数，而IP分组首部最小为20B，因此一个最大分片可封装的数据为：



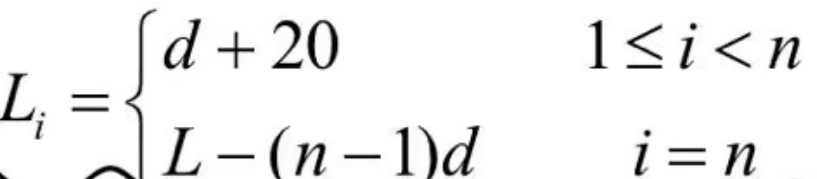
* + 1. 需要的总片数



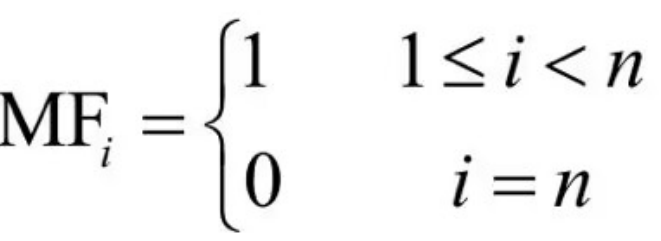
* + 1. 每片的片偏移字段取值为



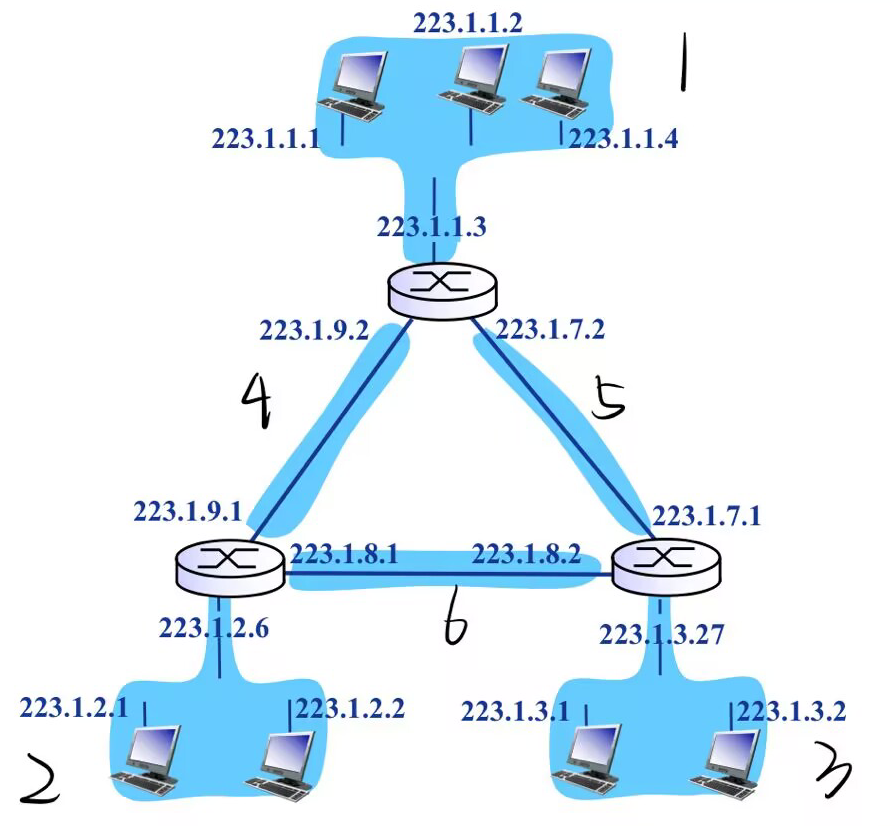
* + 1. 每片的总长度字段



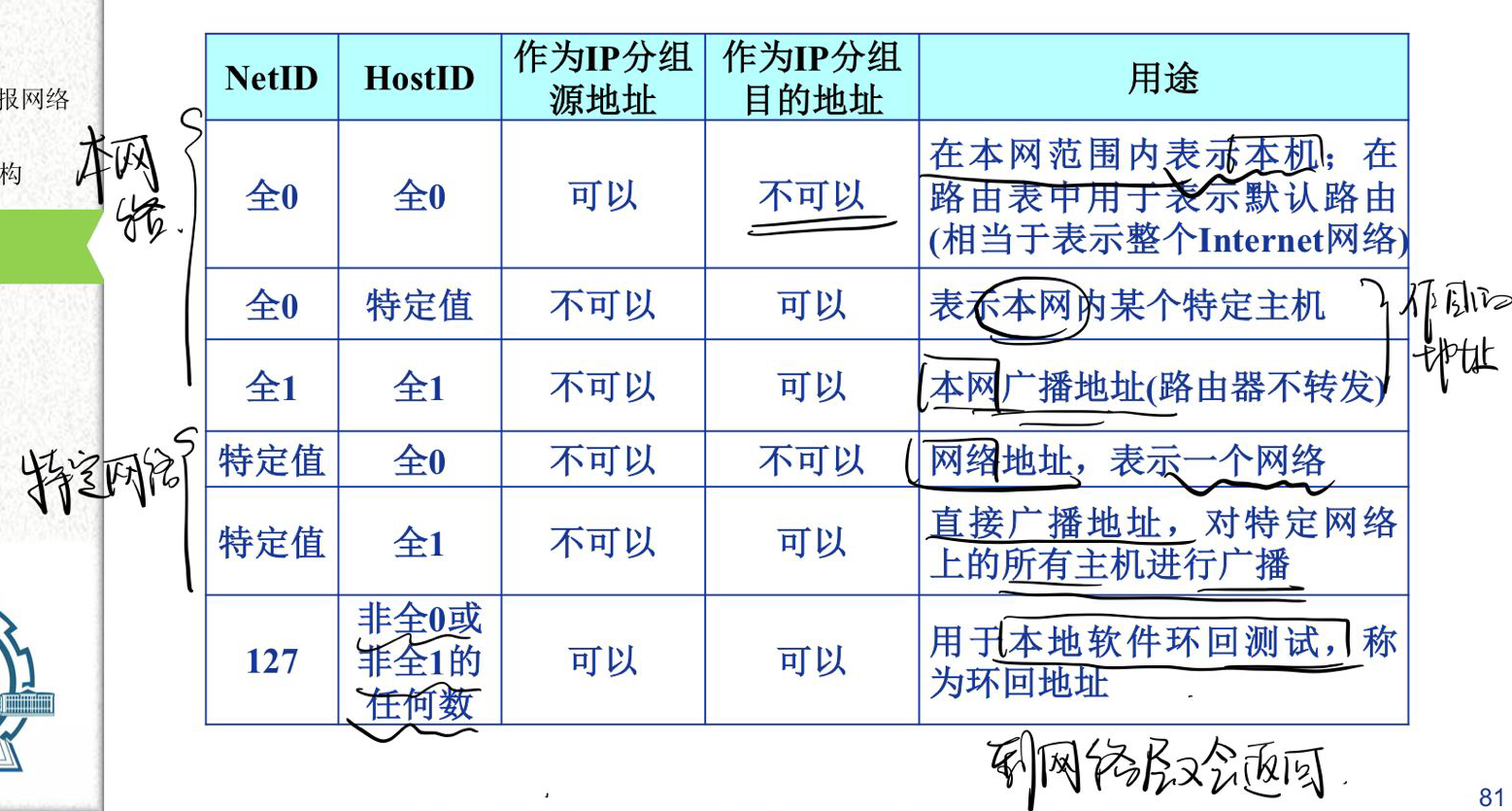
* + 1. 每片的MF标志位



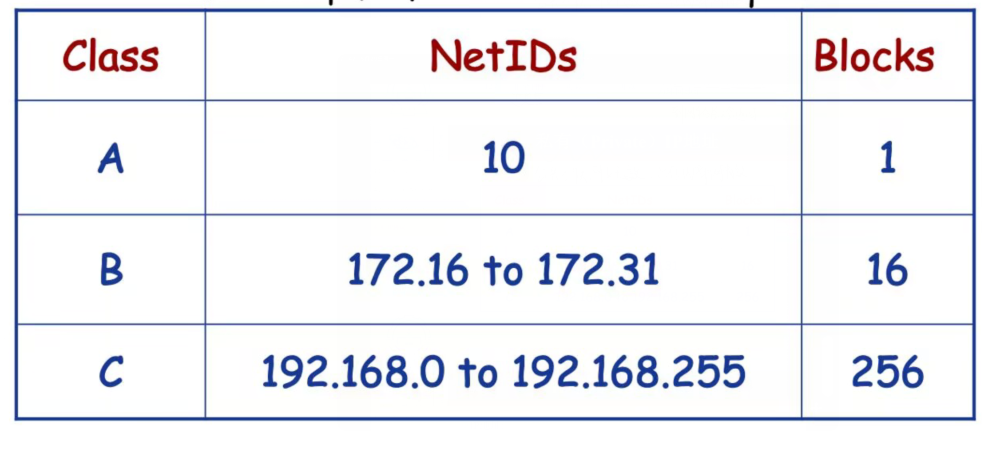
* 1. IP编址：编址实现网络层功能的接口，不关心不实现网络层的功能
     1. 32比特编号标识主机、路由器的接口，与每个接口关联
     2. 使用点分十进制
     3. NetID | HostID
  2. IP子网：
     1. IP地址具有相同网络号的设备接口
     2. 不跨越路由器（第三及以上层网络设备）可以彼此物理联通的接口
     3. e.g.图中网络有6个子网



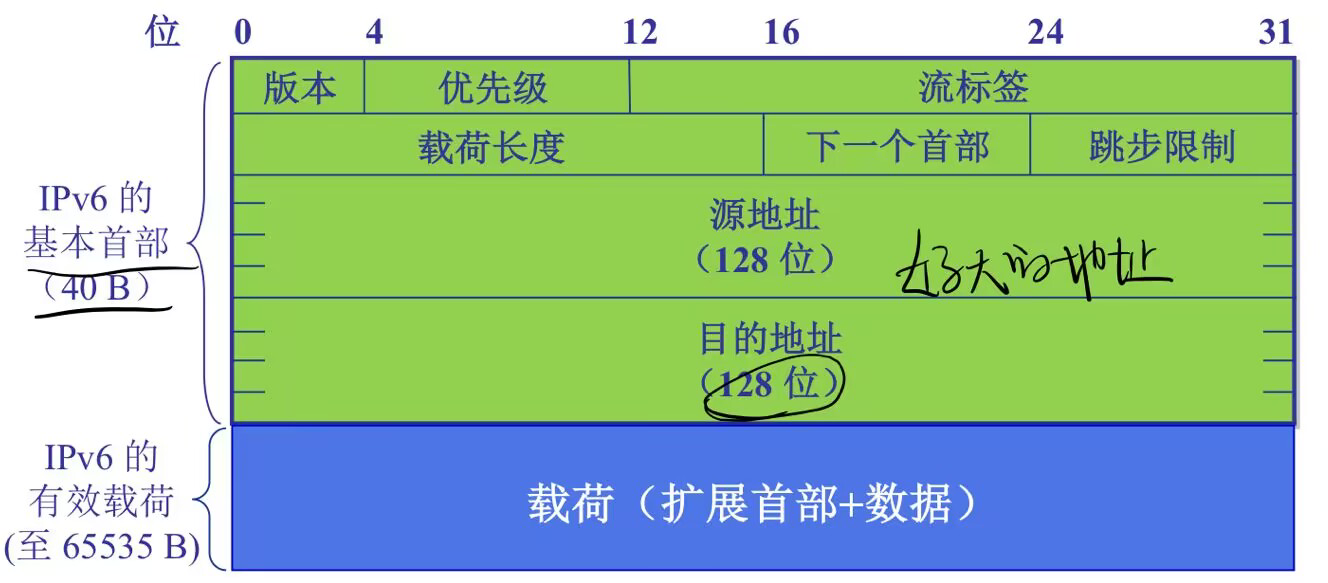
* 1. IP地址
     1. 有类编址：对NetID前4位进行划分
        1. A类地址：NetID（8位）| HostID（24位） 占50% 前4位为0xxx
        2. B类地址：NetID（16位）| HostID（16位） 占25% 前4位为10xx
        3. C类地址：NetID（24位）| HostID（8位） 占12.5% 前4位为110x
        4. D类地址：NetID（24位）| HostID（8位） 占12.5% 前4位为1110
        5. E类地址：NetID（24位）| HostID（8位） 占12.5% 前4位为1111
        6. A、B、C类可以用于标识网络接口，网络规模逐渐减小但网络数增加
        7. D类用于多点广播（组播），一次寻址一组计算机
        8. E类为将来使用保留
     2. **特殊IP地址**



* + 1. 私有IP地址：公共互联网上无效，只在内部网有效

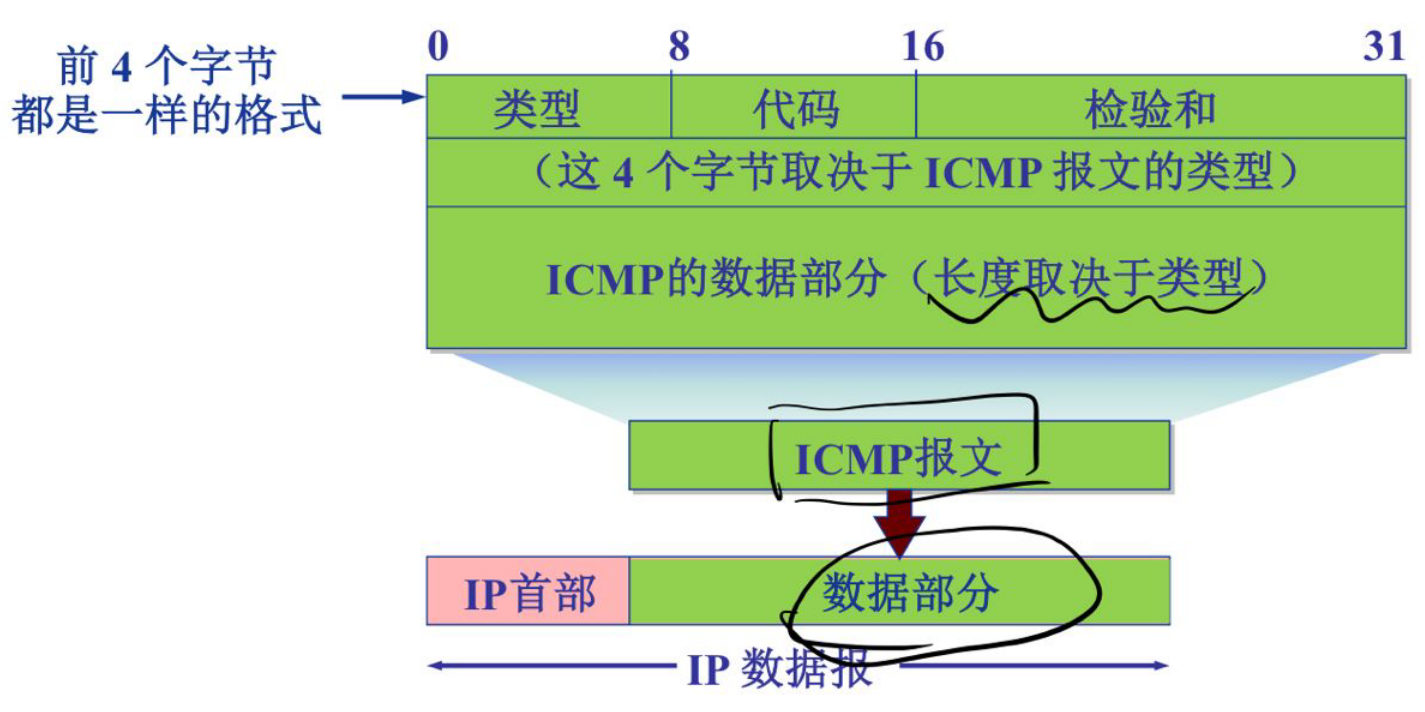


* 1. 子网划分：拆分原主机号为子网号和主机号，实现流量隔离
     1. 利用子网掩码确定是否划分子网，利用多少位划分子网。子网地址+子网掩码可准确确定子网大小
     2. 网络号和子网号取1，主机号取0
     3. 由有类划分可知
        1. A网的默认子网掩码255.0.0.0
        2. B网的默认子网掩码255.255.0.0
        3. C网的默认子网掩码255.255.255.0
  2. **子网掩码的应用**
     1. 将IP分组的目的IP地址与子网掩码按位与运算，提取子网地址（主机号和0与，变成0）
  3. 无类域间路由（CIDR）
     1. 消除传统的A、B和C类地址界限，NetID+SubID=prefix，前缀可以任意长度
     2. 融合子网地址和子网掩码，便于子网区分
     3. 地址格式：a.b.c.d/x，x为前缀长度
     4. 提高IPv4地址空间分配效率和路由效率
  4. 路由聚合：将多个子网地址聚合为一个IP地址，使路由信息通告更高效
     1. 同样选用最长前缀匹配优先
     2. 可能产生黑洞现象，某个地址不在聚合后的地址域对应的ISP中，此时路由器需单独指出该地址
  5. IPv6
     1. 动机：IPv4地址空间分配完了；改进IPv4首部格式
     2. IPv6数据报格式固定长度的40B基本首部，不允许分片



* + 1. 优先级：标识数据报的优先级
    2. 流标签：表示同一“流”中的数据报
    3. 下一个首部：标识写一个选项首部或上层协议首部（如TCP首部）
    4. 相较于IPv4
       1. 彻底移除了校验和，减少每一跳的处理时间
       2. 选项允许存在，但从基本首部中移出，定义多个选项首部，通过下一个首部字段指示
    5. ICMPv6：新版ICMP，附加报文类型“Packet Too Big”，不对分组分片，太大就扔掉，让源主机重做
    6. IPv6地址表示形式：
       1. 一般形式：8个“:”分割的十六进制 e.g. 1080:0:FF:0:8:800:200C:417A
       2. 压缩形式：FF01:0:0:0:0:0:0:43 -> FF01::43
       3. IPv4嵌入形式：0:0:0:0:0:FFFF:13.1.68.3或::FFFF:13.1.68.3 前96位固定，后32位用IPv4地址格式
       4. 地址前缀：2002:43c:476b::/48 Ipv6不再使用掩码，而是利用slider标示子网大小
    7. IPv6基本地址类型：IPv6中广播地址定义为一个特定的多播地址
       1. 单播（一对一通信）
       2. 多播（一对多通信）
       3. 任意播（一对一组之一）
    8. IPv4和IPv6共存的网络：“隧道”->IPv6数据报作为IPv4数据报的载荷进行封装，穿越IPv4网络

1. IP相关协议
   1. ARP（地址解析协议）：在同一个局域网内，已知目的接口的IP地址前提下确定其MAC地址
      1. ARP表：
         1. LAN中的每个IP结点（主机、路由器）维护一个表，存储某些LAN结点的IP/MAC映射关系
         2. 表中缓存的映射关系同样有TTL
      2. 工作过程：
         1. A想要给同一局域网内的B发送数据报，B的MAC地址不在A的ARP表中
         2. A广播ARP查询分组，其中包含B的IP地址
         3. B接收到ARP查询分组，IP地址匹配成功，向A应答B的MAC地址（利用单播帧发送）
      3. ARP是一个即插即用的协议
   2. 从一个LAN路由至另一个LAN：数据报中的源和目的MAC地址会随传输到的LAN变化为下一跳要到达的路由器端口MAC地址，而源和目的IP地址始终不变（到目的LAN后要根据IP地址来物理寻址）
   3. ICMP（互联网控制报文协议）：支持主机或路由器进行差错报告和网络探询
      1. ICMP报文会被封装到IP数据报中传输，格式如下：

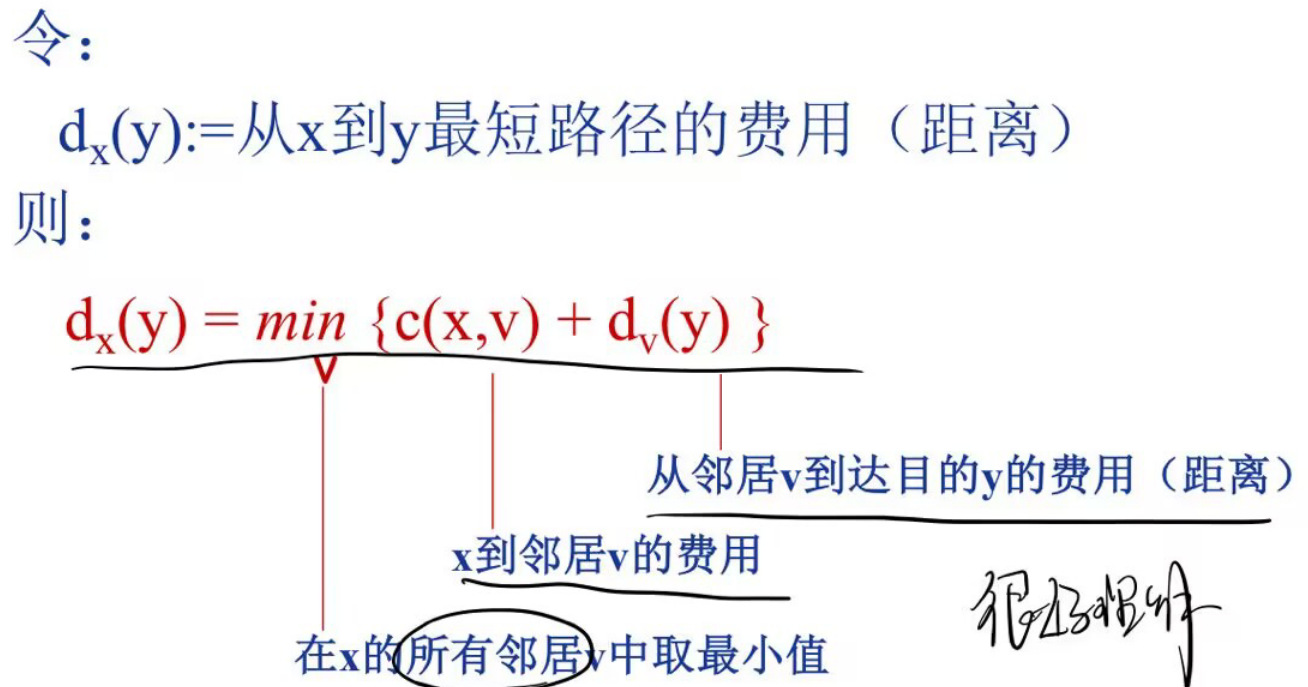


* + 1. 两类ICMP报文：
       1. 差错报告报文：
          1. 目的不可达
          2. 源抑制（缓存满了，请求主机慢点发）
          3. 超时/超期
          4. 参数问题
          5. 重定向
       2. 网络探询报文：探测网络可达性，ping命令的原理
          1. 回声（echo）与应答报文（reply）
          2. 时间戳与请求应答报文

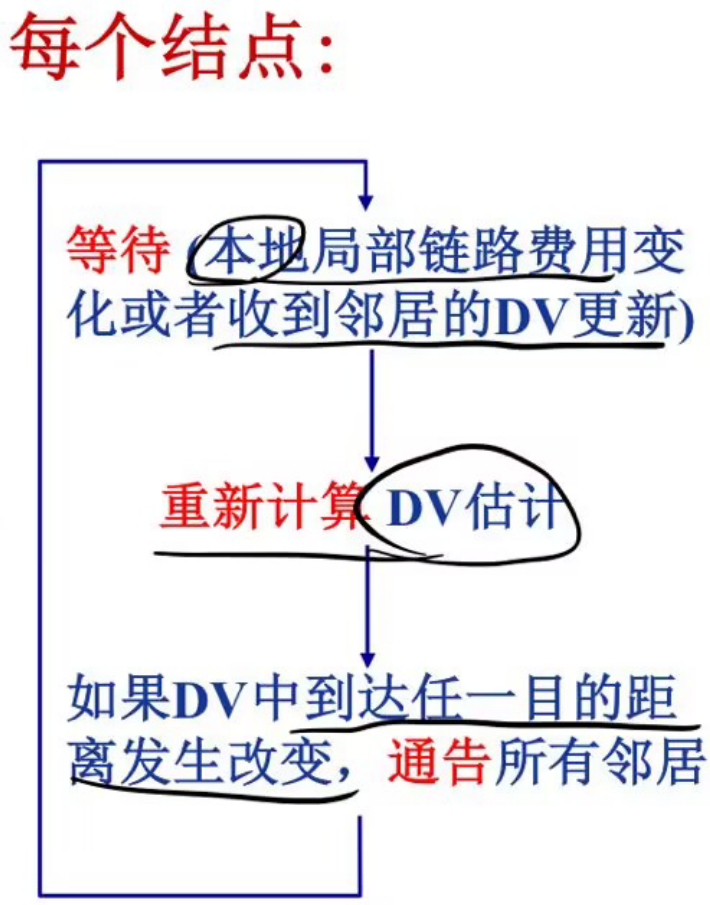


* + 1. 不发送ICMP差错报告报文的情况：
       1. 对ICMP差错报告报文不发送ICMP差错报告报文
       2. 只对第一个IP数据报分片发送ICMP差错报告报文，后续分片不发送
       3. 对所有多播IP数据均不发送ICMP差错报告报文
       4. 具有特殊IP地址（127.0.0.0,0.0.0.0等）的IP数据报不发送ICMP差错报告报文
  1. 主机如何获得IP地址？可以静态配置，也可以动态获得
  2. DHCP（动态主机配置协议）：主机通过与DHCP服务器报文交换获得IP地址
     1. 从服务器动态获取IP地址、子网掩码、默认网关地址、DNS服务器名称与IP地址等
     2. 即插即用，且允许地址重用，分配的IP地址有有效时长
     3. 工作过程：
        1. 主机广播DHCP discover（发现报文）
        2. DHCP服务器利用DHCP offer（提供报文）进行响应
        3. 主机请求IP地址：DHCP request（请求报文）（这个也是广播报文）
        4. DHCP服务器分配IP地址：DHCP ack（确认报文）（包括分配给客户的IP地址、子网掩码、默认网关、DNS服务器地址等）
     4. 协议在应用层实现，DHCP报文被封装到UDP报文中
  3. NAT（网络地址转换）：本地网络设备与互联网的IP地址转换
     1. 所有离开本地网络去往Internet的数据报的源IP地址被替换为相同的NAT地址以及不同的端口号
     2. 本地网络内通信的IP数据报的源与目的IP地址均在子网内部
     3. 地址转换实现：
        1. 替换：利用（NAT IP地址，新端口号）替换每个外出IP数据报的（源IP地址，源端口号）
        2. 记录：将每对（NAT IP地址，新端口号）（源IP地址，源端口号）的映射存储到NAT转换表中
        3. 替换：根据NAT转换表，替换每个进入内网的IP数据报的（目的IP地址，目的端口号）
        4. 注意这里要和ARP做区分，ARP协议的IP地址不变是在子网内部物理寻址，这个替换是在子网间实现IP地址转换映射，两者是不一样的东西
     4. 争议：违背端到端通信原则，给路由器增加了不属于第三层的功能
     5. NAT地址穿透：客户不能利用内网地址直接访问服务器，怎么办？
        1. 静态配置NAT，将特定端口的连接请求转发给服务器
        2. UPnP（互联网网关设备协议）自动配置：学习到NAT公共IP地址，在NAT转换表中增删映射
        3. 中继（如Skype）：NAT内部的客户和外部客户分别与中继服务器建立连接，由该服务器桥接两个连接的分组

1. 路由算法
   1. 我们可以将网络抽象成图结构，路由算法关键就是寻找最小费用路径
   2. 路由算法分类
      1. 全局信息vs分散信息
         1. 全局信息：所有路由器掌握完整的网络拓扑和链路费用信息 e.g. 链路状态路由算法（LS）
         2. 分散信息：路由器只掌握物理相连的邻居以及链路费用，需要有邻居间的信息交换与运算的迭代 e.g. 距离向量路由算法（DV）
      2. 静态路由vs动态路由
         1. 静态路由：手工配置，路由更新慢，优先级高
         2. 动态路由：路由定期更新，更新快，及时响应链路费用或网络拓扑变化
   3. 链路状态路由算法（LS）：**Dijkstra算法**在网络的应用
      1. 所有结点（路由器）掌握网络拓扑和链路费用（通过链路状态广播）
      2. 经过k次迭代后可以得到到达第k个目标结点的最短路径
      3. 路由器根据最后迭代得到的最短路径设置下一跳链路
      4. 算法存在震荡可能：状态改变可能导致链路传输方向改变（乒乓环路）最终导致丢包
   4. 距离向量路由算法（DV）：Bellman-Ford方程（动态规划）在网络上的应用



* + 1. 重点：结点获得最短路径的下一跳，并将该信息用于转发表
    2. 思路：
       1. 结点已知其到达每个邻居的费用，维护其所有邻居的距离向量
       2. 每个结点不定时将其自身的DV估计发给其邻居
       3. 当x收到邻居的新的DV估计时，即依据B-F方程更新自身的距离向量估计
    3. 分布式：每个结点只在DV变化时通告给邻居，不变就不必要通告



* + 1. 好消息传播快，坏消息传播慢，这就带来无穷计数问题，怎么办？
       1. 毒性逆转：一个结点Z到达X的最小费用路径经过Y，则Z通告给Y的到达X的距离为无穷大
       2. 定义最大度量：定义一个最大的有效费用值，将无穷转换为有效度量
  1. LS与DV比较：
     1. 报文交换：LS发送报文为O（mn）；DV只在邻居间交换DV
     2. 收敛速度：LS时间更复杂，可能产生震荡；DV收敛时间不定，可能出现路由环路和无穷计数
     3. 健壮性：LS结点可以通告，每个结点自己重新计算自己的转发表；DV每个结点基于邻居的转发表计算自己的，会导致故障失效在全网传播
  2. 层次路由：聚合路由器为一个自治系统AS，同一AS内的路由器运行相同的路由协议
     1. 网关路由器：AS边缘连接其他AS的路由器
     2. 转发表由AS内部路由算法和AS间路由算法共同配置
     3. 各自治系统需要学习不同的目的AS可以通过哪个AS到达：自治系统间AS路由
     4. 热土豆路由：若子网可通过多个AS到达，优先将分组发给最近的网关路由器（费用最小）

1. 路由协议
   1. AS内部路由：内部网络协议IGP
   2. 常见的AS内部路由协议：RIP、OSPF、IGRP（内部网关路由）
   3. RIP（路由信息协议）：基于距离向量路由算法（DV）
      1. 跳步数最大=15
      2. 每隔30s邻居间交换一次DV（通告），每次通告最多25个目的子网
      3. 180s没有收到通告：链路/邻居失效
      4. 邻居/链路失效后，经过该邻居的路由不可用，需重新计算路由，再向邻居发送新的通告
      5. 同样可能存在无穷计数问题，可以使用毒性逆转预防乒乓环路，无穷大距离设置为16
      6. RIP路由表由应用层进程实现，通告报文周期性通告UDP数据报发送，这再次印证了，层次模型只是一个功能上的逻辑划分，与实际实现无关
   4. OSPF（开放最短路径优先）：
      1. 开放：公众可用
      2. 采用链路状态算法（LS），LS分组扩散（通告），每个路由器构造完整拓扑并计算路由
      3. OSPF通告中每个入口对应一个邻居，在整个AS内泛洪，报文直接封装到IP数据报中
      4. 优点：
         1. 安全，所有OSPF可被认证
         2. 允许使用多条费用相同的路径（RIP只能选一条）
         3. 对于每条链路可以针对不同的TOS设置多个不同的费用度量
         4. 集成单播路由与多播路由，多播OSPF与OSPF利用的网络数据相同
         5. 支持大规模AS分层
      5. 主干路由器（运行OSPF算法）、AS边界路由器（连接其他AS）、区边界路由器（AS进一步分区，通告给其他区边界路由器）
   5. AS间路由协议：BGP
   6. BGP（边界网关协议）：事实上的标准域间路由协议
      1. 为每个AS提供一种手段
         1. eBGP：从邻居AS获取子网可达性信息
         2. iBGP：向所有AS内部路由器传播子网可达性信息
         3. 基于可达性信息与策略，确定到其他网络的较好路径
      2. 容许子网向Internet其他部分通告“我在这”
      3. BGP会话：两个BGP路由器交换BGP报文，通告去往不同目的前缀的路径
         1. BGP报文交换基于半永久的TCP连接
      4. BGP报文：
         1. OPEN：建立TCP连接、认证发送方
         2. UPDATE：通告新路径
         3. KEEPALIVE：保活连接，也用于确认OPEN请求
         4. NOTIFICATION：报告先前报文的差错，也用于关闭连接
      5. 当AS3通告一个前缀给AS1时，AS3承诺可以将数据报转发给该子网，在通告中会聚合网络前缀，AS1便可以获得可达性信息
      6. 当路由器获得新的前缀可达性时，即在其转发表中增加关于该前缀的入口（路由项）
      7. 通告的前缀属性包括BGP属性，前缀+属性=路由
         1. AS-PATH（AS路径）：包含前缀通告经过的AS序列
         2. NEXT-HOP（下一跳）：开始一个AS-PATH的路由器接口，指向下一跳AS
      8. 网关路由器在收到通告后，会利用其输入策略决策接受还是拒绝
      9. 路由器也可能获得到达某AS的多条路由，基于以下准则选择：
         1. 本地偏好，策略决策
         2. 最短AS-PATH
         3. 最近NEXT-HOP（热土豆）
         4. 附加准则
      10. 桩网络：只与一个其他AS相连
      11. 双宿网络：与两个其他AS相连
      12. 路由选择策略要基于收益，若W和C均不是B的客户，则B没必要向C通告到达W的路径
      13. 采用不同的AS内与AS间路由协议：
          1. 策略：AS间期望管理路由流量，路由经过者，而AS内无所谓
          2. 规模：层次路由节省路由表大小，适应大规模互联网
          3. 性能：AS间策略主导，而AS内侧重性能