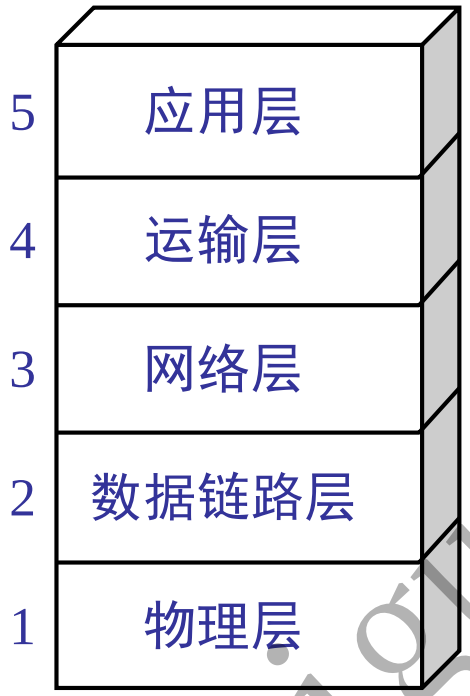
计算机网络

顾 军

计算机学院

jgu@cumt.edu.cn

专题4：数据包怎么在互联网中寻路和转发



- 应用层(application layer)
- 运输层(transport layer)
- 网络层(network layer)
- 数据链路层(data link layer)
- 物理层(physical layer)

Q18：如何进一步提高IP地址利用率？

- 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。
- 然而在 1992 年因特网仍然面临三个必须尽早解决的问题，这就是：
  - B 类地址在 1992 年已分配了近一半，眼看就要在 1994 年 3 月全部分配完毕！
  - 整个 IPv4 的地址空间最终将全部耗尽。
  - 因特网主干网上的路由表中的项目数急剧增长（从几千个增长到几万个）。

无分类的两级编址

- 在 VLSM 的基础上又进一步研究出无分类编址方法，它的正式名字是无分类域间路由选择 CIDR (Classless Inter-Domain Routing)，也称为无分类编址。
- CIDR 使用各种长度的“网络前缀”(network-prefix)来代替分类地址中的网络号和子网号，IP 地址从三级编址（使用子网掩码）又回到了两级编址。

IP地址 ::= {<网络前缀>, <主机号>}

CIDR 记法

- CIDR 使用“斜线记法”(slash notation)，又称为CIDR记法，即在 IP 地址面加上一个斜线“/”，然后写上网络前缀所占的位数（这个数值对应于三级编址中子网掩码中 1 的个数）。
  - CIDR 虽然不使用子网了，但仍然使用“掩码”这一名词（但不叫子网掩码）。

10.0.0.0/10

它的掩码是 10 个连续的 1，斜线记法中的数字就是掩码中1的个数

CIDR 记法的形式

- 10.0.0.0/10 隐含地指出 IP 地址 10.0.0.0 的掩码是 255.192.0.0。此掩码可表示为

  11111111 11000000 00000000 00000000

  255 192 0 0

  掩码中有 10 个连续的 1

- 10.0.0.0/10 可简写为 10/10，也就是把点分十进制中低位连续的 0 省略。

CIDR 记法的其他形式

- CIDR还可以采用在网络前缀的后面加一个星号 * 的表示方法
- 如 00001010 00*，在星号 * 之前是网络前缀，而星号 * 表示 IP 地址中的主机号，可以是任意值。

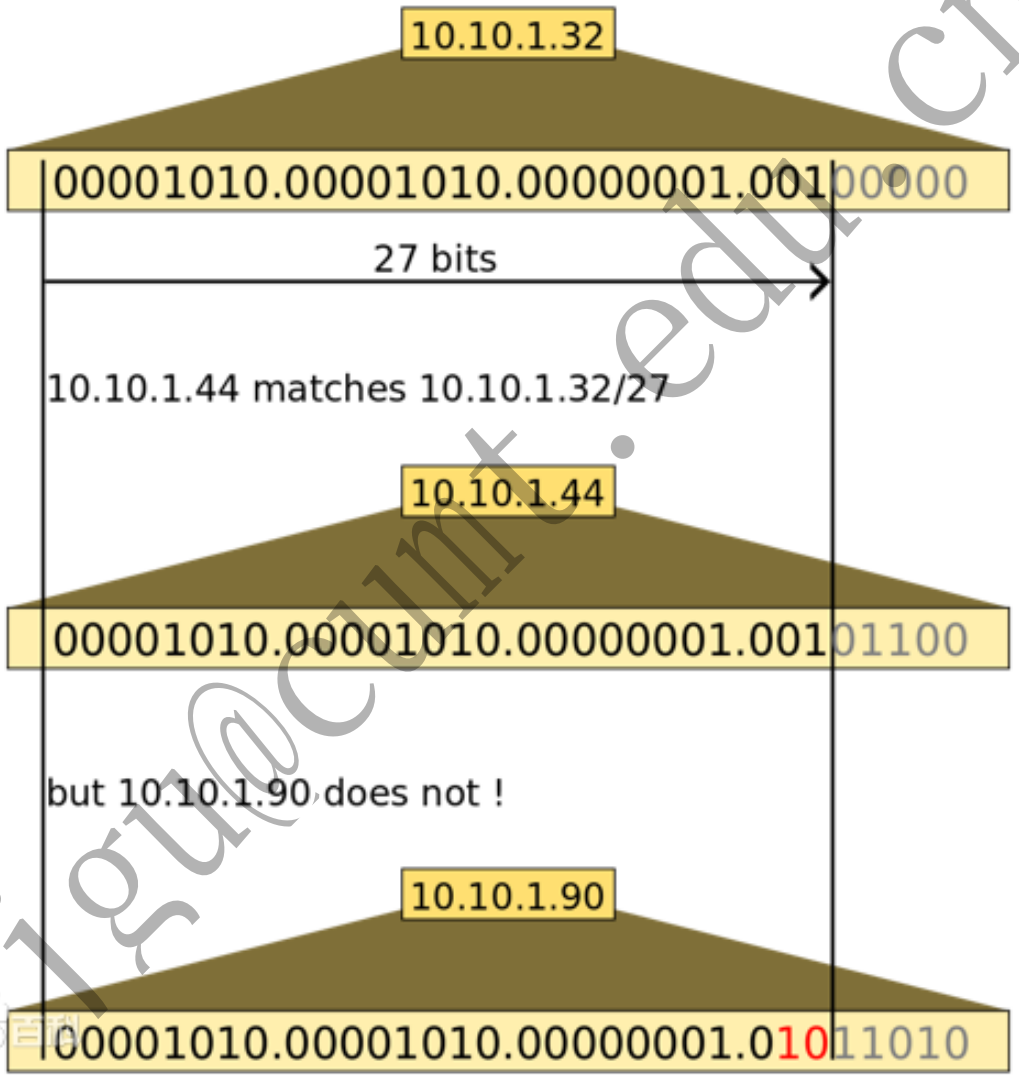
CIDR 地址块

- CIDR 把网络前缀都相同的连续IP地址组成“CIDR地址块”，地址数一定是2的整数次幂。
- 128.14.32.0/20 表示的地址块共有 2^{12} 个地址（斜线后面的 20 是网络前缀的位数）。
  - 这个地址块的起始地址是 128.14.32.0。
  - 128.14.32.0/20 地址块的最小地址：128.14.32.0
  - 128.14.32.0/20 地址块的最大地址：128.14.47.255
  - 全 0 和全 1 的主机号地址一般不使用。
- 在不需要指出地址块的起始地址时，也可将这样的地址块简称为“/20 地址块”。

128.14.32.0/20 表示的地址块（2^{12} 个地址）

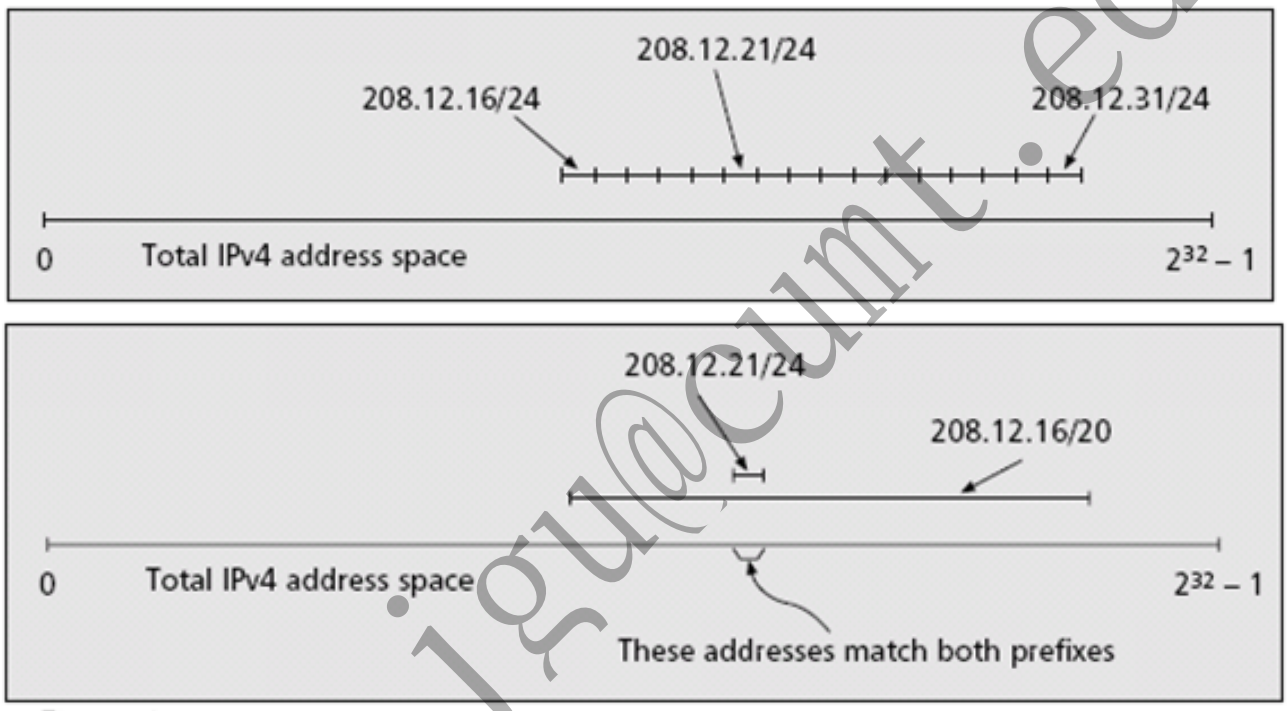
所有地址的 20 位前缀都是一样的

CIDR 地址匹配



Q19：如何理解CIDR技术？

- CIDR格式可以理解为一个从0到(2^{32}-1)长的“刻度尺”。如：掩码24表示2^{24}个小线段，每个线段包含2^{32-24}个IP点。



208.12.16/24 } 相离
208.12.21/24 }

208.12.16/24 } 相邻
208.12.17/24 }

208.12.21/24 } 包含
208.12.16/20 }

任意2个CIDR地址块都不能相交

CIDR技术要点1

- CIDR 消除了传统的 A 类、B 类和 C 类地址及划分子网的概念，对原来的有类别路由选择进程进行了重新构建，因而可以更加有效地分配 IPv4 的地址空间，并且在新的IPv6使用前容许因特网的规模继续增长。
- 如果 IP 地址的分配从一开始就采用 CIDR，那么我们可以按网络所在的地理位置来分配地址块，这样就可以大大减少路由表中的项目数。

基于地理位置划分地址块

例如，可以将世界划分为四大地区，每一个地区分配一个CIDR 地址块（每个地址块包含约 3200 万个地址 (2^{32-7}）：

- 194/7(194.0.0.0至195.255.255.255)分配给欧洲
- 198/7(198.0.0.0至199.255.255.255)分配给北美洲
- 200/7(200.0.0.0至201.255.255.255)分配给中美洲和南美洲
- 202/7(202.0.0.0至203.255.255.255)分配给亚洲和太平洋地区

IP地址与地理位置关联好处是可以大大压缩路由表中的项目数。例如从中国发往北美的数据报（不管它是地址块 198/7 中的哪一个地址）都先送到美国的一个路由器，因此在路由表中使用一个项目就行了。

显然，已分配的IP地址全部回收再重新分配是不可能的，但是部分调整还是可以的

江苏 徐州 电信IP地址段（部分）

原A类IP地址

IP地址段：58.218.0.0 - 58.218.13.255 共有【3584】个IP记录

详细地址：江苏省徐州市 电信

58.218.0.17 58.218.0.16 58.218.0.15 58.218.0.14 58.218.0.13 58.218.0.12 58.218.0.11 58.218.0.10

58.218.0.9 58.218.0.8 58.218.0.7 58.218.0.6 58.218.0.5 58.218.0.4 58.218.0.3 58.218.0.2 58.218.0.1

58.218.0.0 更多

IP地址段：58.218.82.56 - 58.218.144.57 共有【15874】个IP记录

详细地址：江苏省徐州市 电信

58.218.82.73 58.218.82.72 58.218.82.71 58.218.82.70 58.218.82.69 58.218.82.68 58.218.82.67

58.218.82.66 58.218.82.65 58.218.82.64 58.218.82.63 58.218.82.62 58.218.82.61 58.218.82.60

58.218.82.59 58.218.82.58 58.218.82.57 58.218.82.56 更多

IP地址段：58.218.150.15 - 58.218.152.13 共有【511】个IP记录

详细地址：江苏省徐州市 电信

58.218.150.32 58.218.150.31 58.218.150.30 58.218.150.29 58.218.150.28 58.218.150.27 58.218.150.26

58.218.150.25 58.218.150.24 58.218.150.23 58.218.150.22 58.218.150.21 58.218.150.20 58.218.150.19

58.218.150.18 58.218.150.17 58.218.150.16 58.218.150.15 更多

IP地址段：58.218.152.15 - 58.218.152.33 共有【19】个IP记录

详细地址：江苏省徐州市 电信

58.218.152.32 58.218.152.31 58.218.152.30 58.218.152.29 58.218.152.28 58.218.152.27 58.218.152.26

58.218.152.25 58.218.152.24 58.218.152.23 58.218.152.22 58.218.152.21 58.218.152.20 58.218.152.19

58.218.152.18 58.218.152.17 58.218.152.16 58.218.152.15

CIDR技术要点2

- CIDR用13～27位长的前缀取代了原来地址结构对地址网络部分的限制（3类地址的网络部分分别被限制为8位、16位和24位）。
  - 在管理员能分配的地址块中，主机数量范围是32～500,000，从而能更好地满足机构对地址的特殊需求。
- 前缀变长，则可以使地址块变小，把一个大的地址块划分为多个小的地址块，达到子网划分的目的
- 前缀变短，则可以使地址块变大，把多个连续地址块聚合为一个大的地址块，达到子网聚合的目的

Q20：如何进一步划分CIDR地址块?

- “CIDR不使用子网”，是指CIDR中并没有在32位地址中指明若干位作为子网字段。
- 但分配到一个CIDR地址块的单位，仍然可以在本单位内根据需要划分出一些子网。这些子网也都只有一个网络前缀和一个主机地址号，但子网的网络前缀比整个单位的网络前缀要长一些。
- CIDR通过网络前缀在Internet上创建附加地址，这些地址提供给服务提供商（ISP），再由ISP分配给客户。
  - 可以按照实际需要进行网络地址分配，提高地址空间的利用率。

CIDR地址块可以进行多次划分

CIDR 地址块划分方法

- CIDR地址块的划分也是从主机号中借走一定的位数即可，与基本子网划分不同的是：借走n位可以划分成2^n个子网，不用减2。
  - 主机号为全0的地址是地址块的最小地址，加上“/前缀”后用来表示该地址块，不能分配给任何网络接口。
  - 主机号为全1的地址是地址块的最大地址，是该地址块的广播地址，不能分配给任何网络接口。

CIDR 地址块划分举例1

求网络地址块212.110.96.0/20包含的最大主机数，以及8等分子网后，各子网的掩码及主机数。

地址块： 212.110.01100000.0 /20

最小地址 212.110.01100000.00000000 表示该地址块

最大地址 212.110.01101111.11111111 地址块广播地址

最大主机数 $2^{32-20}-2 = 2^{12}-2$

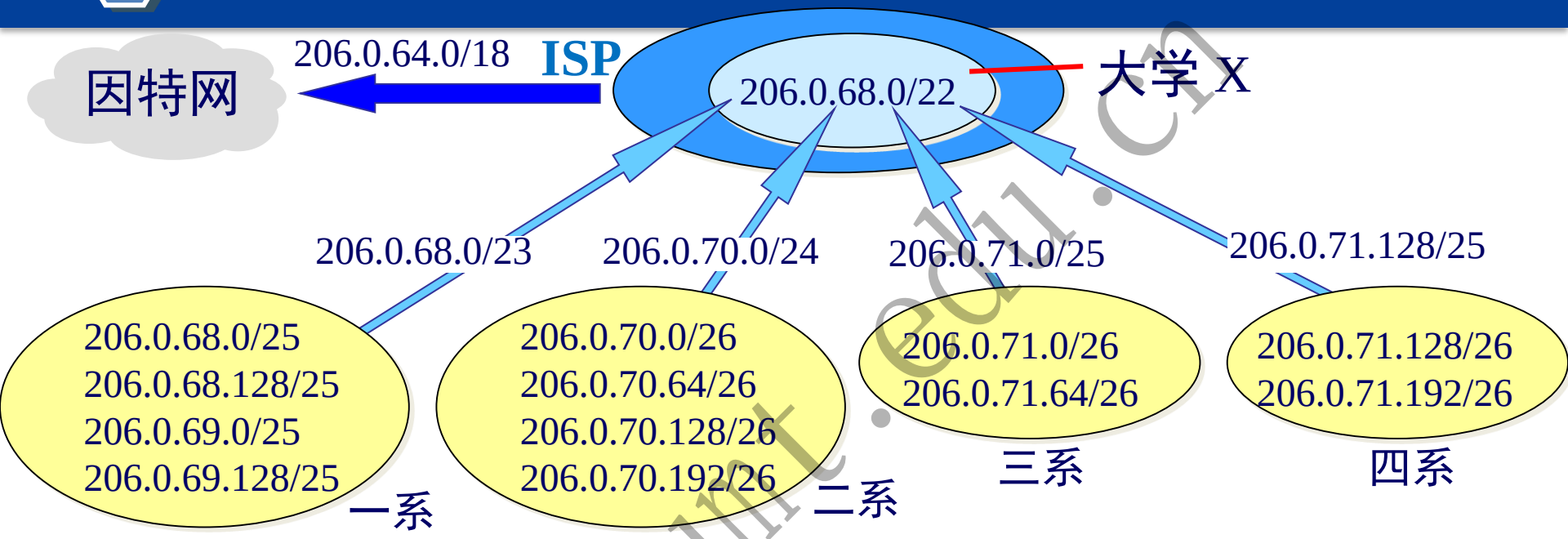
8等分子网

212.110.01100000.00000000 /23
212.110.01100010.00000000 /23
212.110.01100100.00000000 /23
212.110.01100110.00000000 /23
212.110.01101000.00000000 /23
212.110.01101010.00000000 /23
212.110.01101100.00000000 /23
212.110.01101110.00000000 /23

子网掩码 /23 或 255.255.254.0 主机数 $2^9-2 = 512-2 = 510$

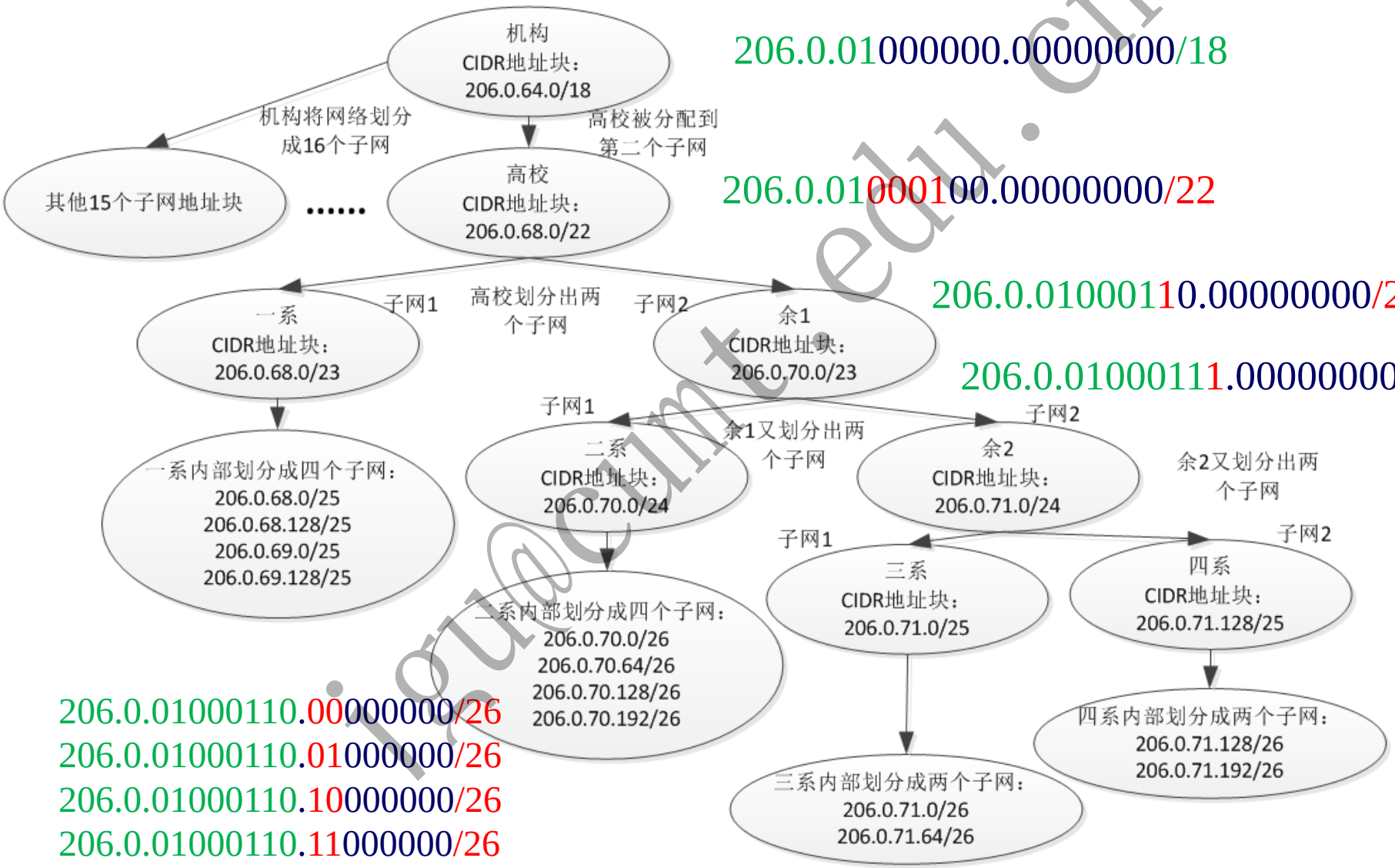
CIDR 地址块划分举例2

某个ISP拥有一个大的CIDR地址块，即206.0.64.0/18，现在某个高校需要申请一个较大的CIDR地址块以供学校使用，学校内部又分为4个系，由于每个系的人数不一样，所以要给人数较多的系分配较多的IP地址，人数较少的系分配较少的IP地址。



单位	地址块	二进制表示	地址数
ISP	206.0.64.0/18	11001110.00000000.01*	16384
大学	206.0.68.0/22	11001110.00000000.010001*	1024
一系	206.0.68.0/23	11001110.00000000.0100010*	512
二系	206.0.70.0/24	11001110.00000000.01000110.*	256
三系	206.0.71.0/25	11001110.00000000.01000111.0*	128
四系	206.0.71.128/25	11001110.00000000.01000111.1*	128

某高校CIDR划分过程



206.0.01000000.00000000/18

206.0.01000100.00000000/22

206.0.01000110.00000000/23

206.0.01000111.00000000/24

206.0.01000110.00000000/26
206.0.01000110.01000000/26
206.0.01000110.10000000/26
206.0.01000110.11000000/26

Q21：如何利用CIDR构成超网？

- 将多个C类的网络聚合起来，构成一个单一的、具有共同地址前缀的网络，从而减少这些地址在路由表中的表项数，也称构成超网(Super Netting)。
  - 比如，规模在254个结点以上（但远小于 64K）的网络，可分配一个由若干 C 类地址聚合成的超网地址空间块，而不必占用一个完整的B类地址空间。
- 构成超网是“子网划分”的派生词，可看作子网划分的逆过程。

CIDR构成超网举例1：

一个机构有30000台主机，且只能申请C类地址，该如何配置？

- 30000台主机需要多少主机位？
  - 15位
- 可以利用C类地址进行地址块聚合
  - 申请一组连续的C类地址
  - C类地址的前缀长度为24
  - 32-15=17，前17位相同，则子网掩码为17个1，15个0

网络前缀越短，其地址块所包含的地址数就越多

CIDR构成超网举例2：

- 假设有一组C类地址为192.168.8.0－192.168.15.0，如果用CIDR将这组地址聚合为一个网络，其网络地址和子网掩码应该为：
  - A. 192.168.8.0/21
  - B. 192.168.8.0/20
  - C. 192.168.8.0/24
  - D. 192.168.8.15/24

- **KEY：A 要求将192.168.8.0－192.168.15.0这组C类地址聚合为一个网络，需要先将C类地址的第三个八位组转换成二进制：**

192.168.8.0	192.168.00001	000.0
192.168.9.0	192.168.00001	001.0
192.168.10.0	192.168.00001	010.0
192.168.11.0	192.168.00001	011.0
192.168.12.0	192.168.00001	100.0
192.168.13.0	192.168.00001	101.0
192.168.14.0	192.168.00001	110.0
192.168.15.0	192.168.00001	111.0

- 相同的网络前缀为21位，最小的起始地址为192.168.8.0，因此聚合后的网络地址为192.168.8.0/21。
- 剩下的11位作为主机位，主机位地址代表一个主机，只有网络地址才有聚合的意义。

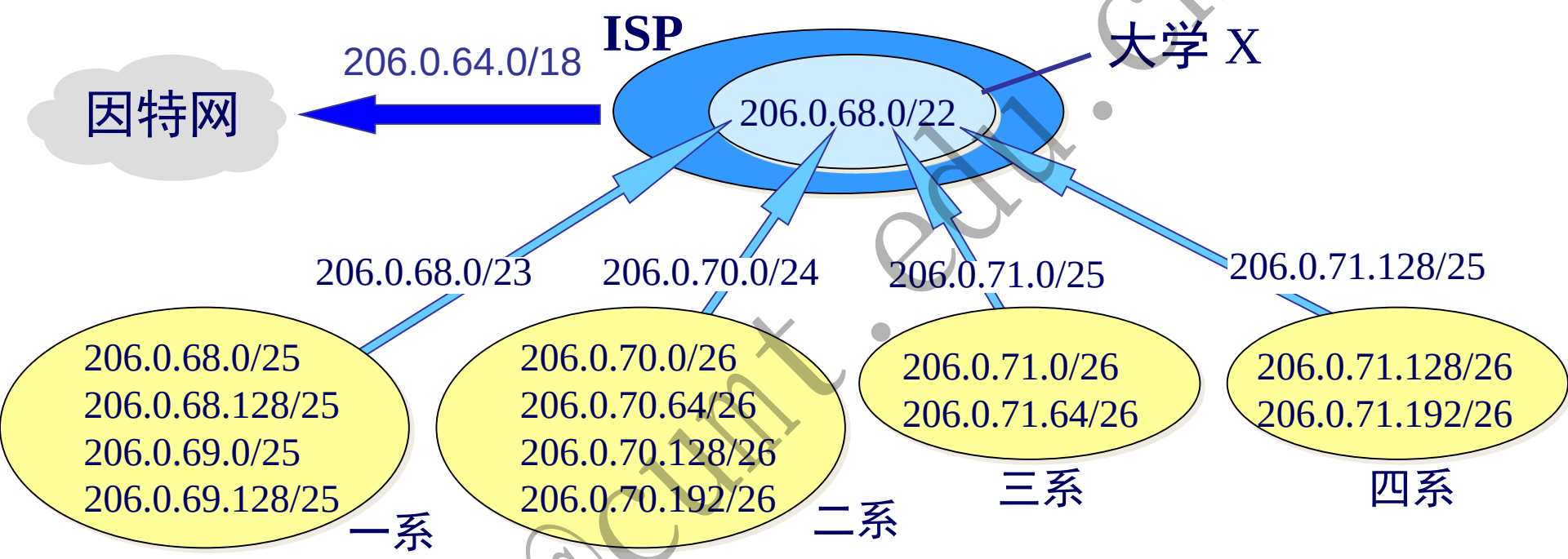
Q22：如何利用CIDR聚合路由？

- 超网使得一个CIDR地址块中有很多地址，所以路由表中的一个项目可以表示很多个（例如上千个）原来传统分类地址的路由，减少路由表表项的数量，节省路由器中的资源。
- 这种地址的聚合称为路由聚合(route aggregation)。
  - 比如，1990年，Internet上约有2000个路由。五年后，Internet上有3万多个路由。
  - CIDR可以将路由集中起来，使一个IP地址可以代表主要骨干提供商服务的几千个IP地址，从而减轻Internet路由器的负担。
  - 如果没有CIDR，路由器就不能支持Internet网站的增多。

CIDR路由聚合举例1：

- 直接给某单位分配16个C类地址块（/24），那么就相当于给该单位分配了16个C类网络。这个单位对外界来说，是16个C类网络。而每一个C类网络都要在本单位外面的路由表中占有一个表项，使得路由表更大了。
- 当本单位内的许多主机相互通信时，由于跨越了不同的网络，都必须使用路由器来转发IP数据报，由此造成的开销是很大的。因此，单位一般不愿意接受16个C类地址，但也很难申请到一个B类地址块，而B类地址块也太大。
- 所以，可以采用CIDR，分配一个地址块/20。

CIDR路由聚合举例2：



这个 ISP 共有 64 个 C 类网络。如果不采用 CIDR 技术，则在与该 ISP 的路由器交换路由信息的每一个路由器的路由表中，就需要有 64 个项目。但采用地址聚合后，只需用路由聚合后的 1 个项目 206.0.64.0/18 就能找到该 ISP。

Q23：CIDR中怎么匹配路由表项？

- 使用 CIDR 时，路由表中的每个项目由“网络前缀”和“下一跳地址”组成。在查找路由表时可能会得到不止一个匹配结果。
- 网络前缀越短，其地址块所包含的地址数就越多。而在三级结构的IP地址中，划分子网是使网络前缀变长。网络前缀越长，其地址块就越小，因而路由就越具体(more specific) 。
- 应当从匹配结果中选择具有最长网络前缀的路由：最长前缀匹配(longest-prefix matching)，又称为最长匹配或最佳匹配。

最长前缀匹配举例

收到的分组的目的地址 D = 206.0.71.128

路由表中的项目：206.0.68.0/22　　（ISP）

206.0.71.128/25　（四系）

查找路由表中的第 1 个项目

第 1 个项目 206.0.68.0/22 的掩码 M 有 22 个连续的 1。

M = 11111111. 11111111. 11111100. 00000000

因此只需把 D 的第 3 个字节转换成二进制。

	M =	11111111 .	11111111 .	11111100 .	00000000
AND	D =	206.	0.	01000111.	0
		206.	0.	01000100.	0

与 206.0.68.0/22 匹配

最长前缀匹配举例

收到的分组的目的地址 D = 206.0.71.128

路由表中的项目：206.0.68.0/22　　（ISP）

206.0.71.128/25　（四系）

再查找路由表中的第 2 个项目

第 2 个项目 206.0.71.128/25 的掩码 M 有 25 个连续的 1。

M = 11111111. 11111111. 11111111. 10000000

因此只需把 D 的第 4 个字节转换成二进制。

	M =	11111111.	11111111.	11111111 .10000000
AND	D =	206.	0.	71. 10000000
		206.	0.	71. 10000000

与 206.0.71.128/25 匹配

最长前缀匹配

D AND (11111111 11111111 11111100 00000000)

= 206.0.68.0/22　　匹配

D AND (11111111 11111111 11111111 10000000)

= 206.0.71.128/25　　匹配

- 选择两个匹配的地址中更具体的一个，即选择最长前缀的地址。

Q24：如何加速CIDR路由表的查找？

- 当路由表的项目数很大时，怎样设法减小路由表的查找时间就成为一个非常重要的问题。
- 例如：
  - 连接路由器的线路的速率： 10Gbit/s
  - 分组的平均长度： 2000 bit
  - 匹配的路由器处理能力：
    - 5Mpps (500万个分组/秒)
    - 或　处理一个分组的平均时间只有　200ns

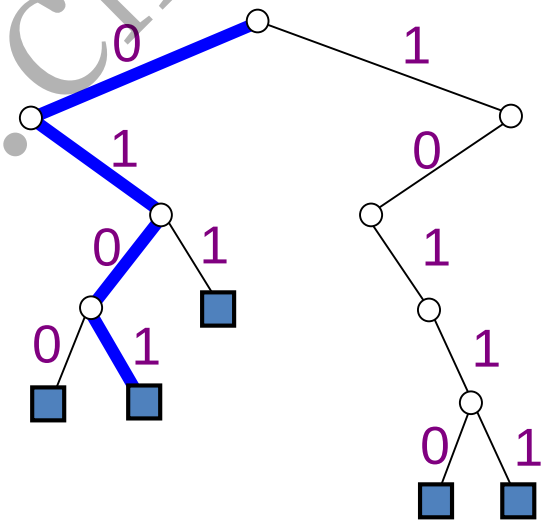
- 对CIDR的路由表的最简单的查找算法就是对所有可能的前缀进行循环查找。
  - 给定一个目的地址D；
  - 对每一个可能的网络前缀长度M，路由器从D中提取前M个比特位，形成一个网络前缀；
  - 查找路由表中的网络前缀；
  - 所找到的最长匹配就对应于要查找的路由。

- 明显缺点就是查找的次数太多了。
- 最坏的情况是路由表中没有这个路由。
  - 此时，算法仍要进行32次（具有32位的网络前缀是一个特定主机路由）
- 对于经常使用的默认路由，要经历31次不必要的查找。

- 为了进行更加有效的查找，通常是将无分类编址的路由表存放在一种层次的数据结构中，然后自上而下地按层次进行查找。这里最常用的就是二叉线索(binary trie)。
- IP 地址中从左到右的比特值决定了从根结点逐层向下层延伸的路径，而二叉线索中的各个路径就代表路由表中存放的各个地址。

用 5 个前缀构成的二叉线索

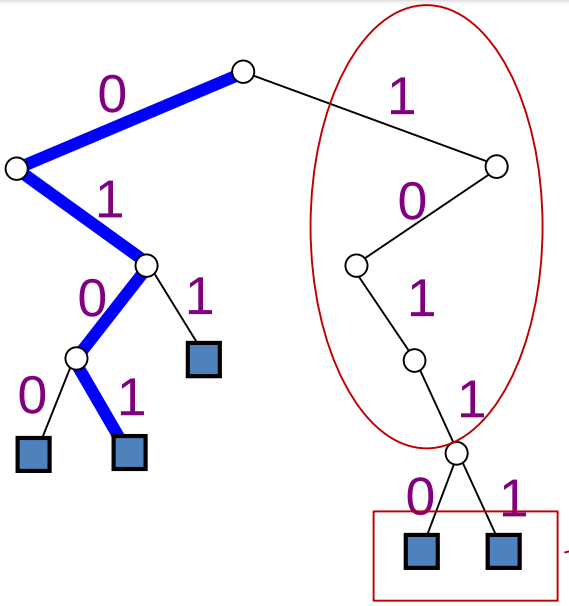
32 位的 IP 地址	唯一前缀
01000110 00000000 00000000 00000000	0100
01010110 00000000 00000000 00000000	0101
01100001 00000000 00000000 00000000	011
10110000 00000010 00000000 00000000	10110
10111011 00001010 00000000 00000000	10111



- 唯一前缀(unique prefix)就是在路由表中所有的IP地址中，该前缀是唯一的。
- 这样就可以用这些唯一前缀来构造二叉线索，在进行查找时，只要能够何唯一前缀相匹配就行了。
- 每个叶结点代表一个唯一前缀。
- 10011011 01111010 00000000 00000000

不在这个二叉线索中

提高二叉线索的查找速度



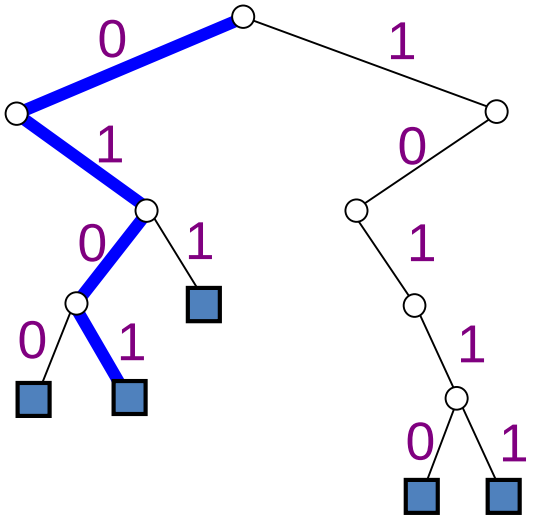
为了提高二叉线索的查找速度，广泛使用了各种压缩技术。

这两个地址具有相同的前4位 1011

- 只要一个地址的前4位是1011，就可以跳过前面的4位（即压缩了4个层次）而直接从第5位开始比较，这样可以减少查找的时间。

完成压缩的计算开销与查找路由表的速度提升的权衡

将二叉线索用于路由



要将二叉线索用于路由表中，还必须使二叉线索中的每一个叶结点包含所对应的网络前缀和子网掩码。

- 当搜索到一个叶结点时，就必须将寻找匹配的目的地址和该叶结点的子网掩码进行逐位“与”运算，看结果是否与对应的网络前缀相匹配。
- 若匹配，就按下一跳的接口转发该分组。
- 否则，就丢弃该分组。

Q25：如何控制IP数据报的交付？

- 为了提高 IP 数据报交付成功的机会，在网际层使用了网际控制报文协议 ICMP (Internet Control Message Protocol)。
- ICMP 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 不是高层协议，而是 IP 层的协议。
- ICMP 报文作为 IP 层数据报的数据，加上数据报的首部，组成 IP 数据报发送出去。

ICMP 差错报告报文

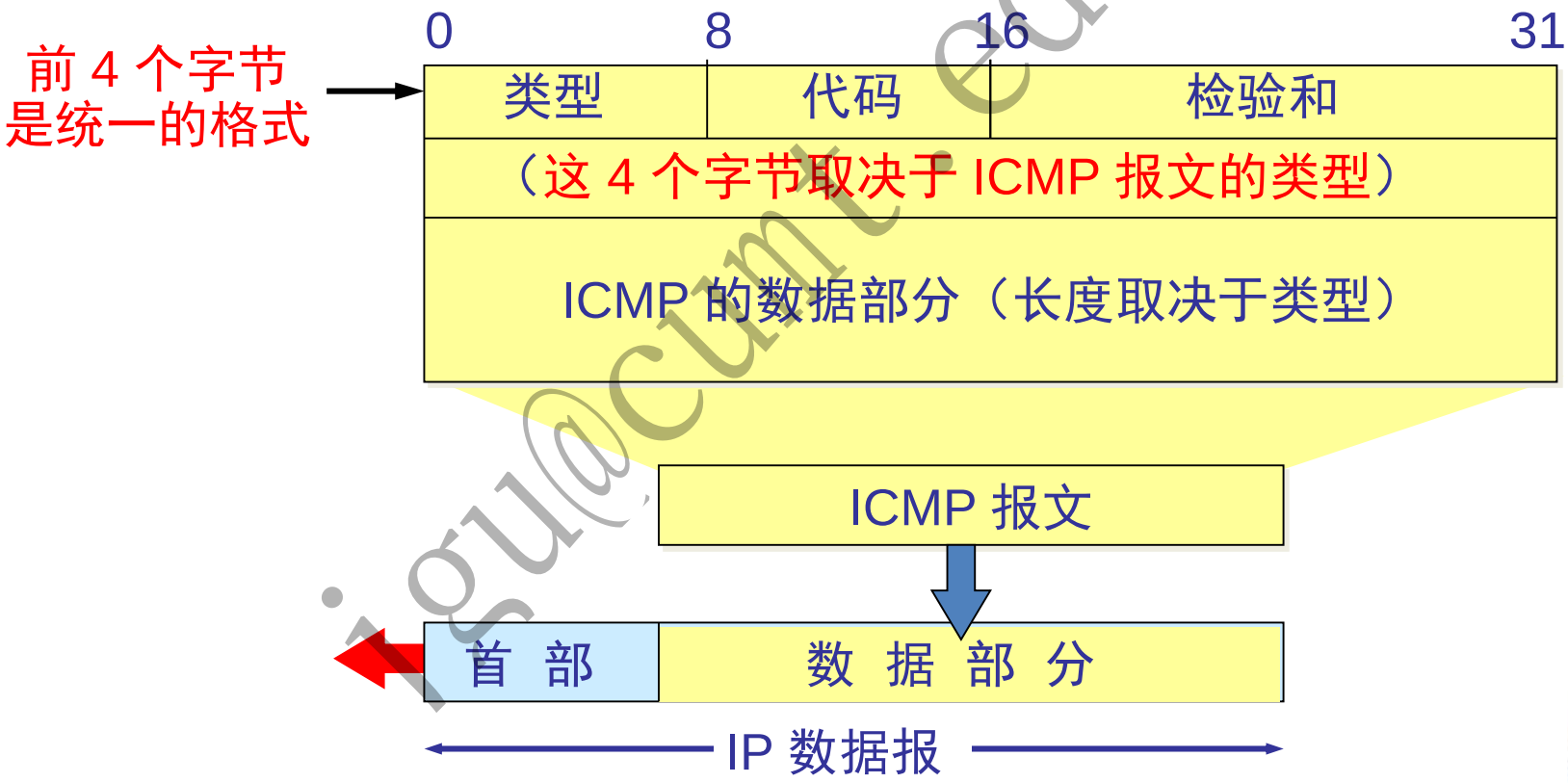
- 终点不可达
  - 网络不可达，主机不可达，协议不可达，端口不可达，需要分片但DF比特已置为1，以及源路由失败等
- 源站抑制(Source quench) ——不再使用
  - 当路由器或主机由于拥塞而丢弃数据报时，让源站知道应当将数据报的发送速率放慢。
- 时间超过
  - 当路由器收到生存时间为零的数据报时
- 参数问题
  - 当路由器或目的主机收到的数据报的首部中的字段的值不正确时
- 改变路由（重定向）(Redirect)
  - 让主机知道下次应将数据报发送给另外的路由器

ICMP询问报文

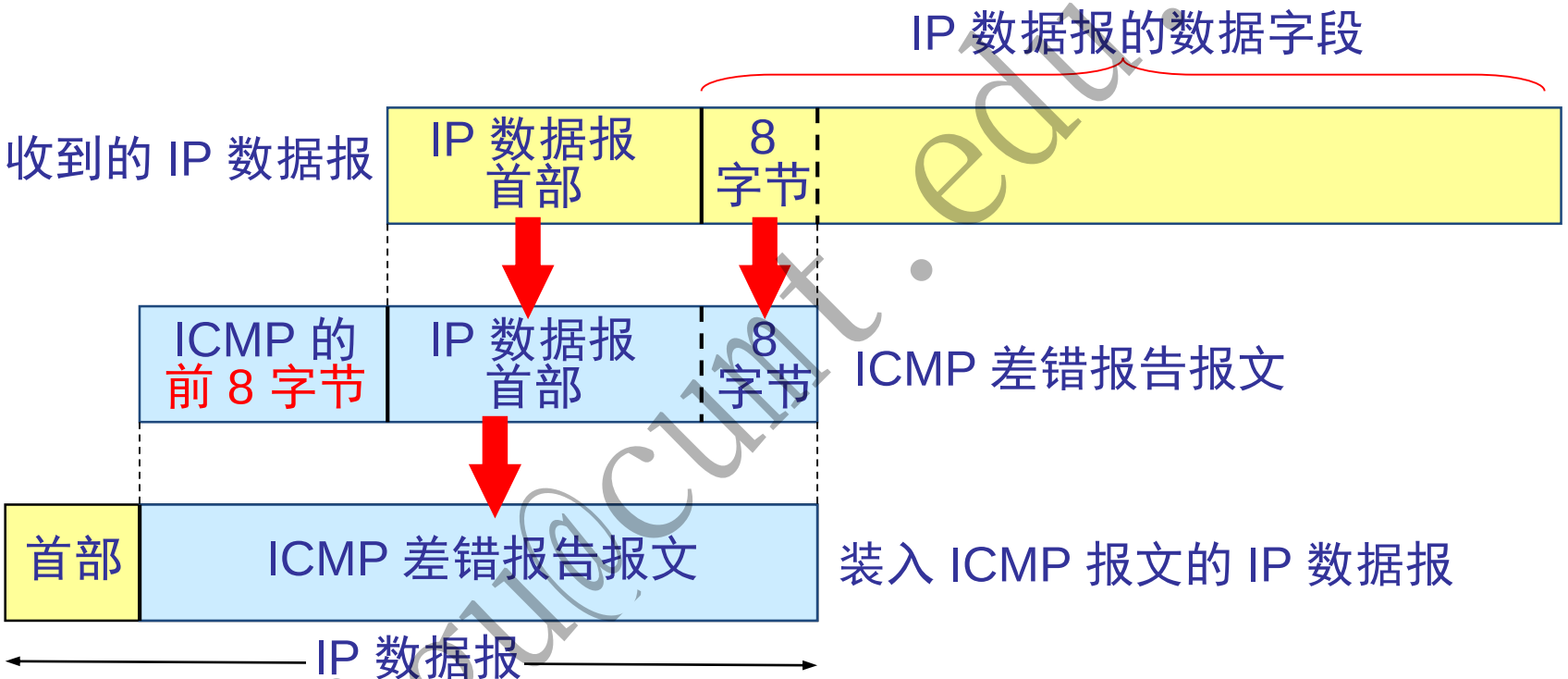
- 回送请求和回答报文
  - 由主机或路由器向一个特定的目的主机发出的询问，用来测试目的站是否可达以及了解其有关状态
- 时间戳请求和回答报文
  - 允许系统向另一个系统查询当前的时间，可以提供毫秒级的分辨率
- 掩码地址请求和回答报文——不再使用
  - 系统广播掩码地址请求报文，子网掩码服务器回答某个接口的地址掩码。
- 路由器询问和通告报文 ——不再使用
  - 了解连接在本网络上的路由器是否正常工作。主机广播（或多播）路由器询问报文，收到询问报文的一个或几个路由器使用通告报文广播其路由选择信息

ICMP 报文的格式

- ICMP 报文的前 4 个字节是统一的格式，共有三个字段：即类型、代码和检验和。接着的 4 个字节的内容与 ICMP 的类型有关。



ICMP 差错报告报文的数据字段的内容



traceroute(Linux)/tracert(Windows)

- 用来跟踪一个分组从源点到终点的路径，获得目的主机的路由信息。
- 利用 IP 数据报中的 TTL 字段和 ICMP超时差错报告报文实现对从源点到终点的路径的跟踪。
  - 首先，送出一个TTL是1的IP 数据包到目的地，当第一个路由器收到这个数据包时，将TTL减1。此时，TTL变为0，所以会将此数据包丢掉，并送回一个「ICMP time exceeded」消息（包括发IP包的源地址，IP包的所有内容及路由器的IP地址），tracert 收到这个消息后，便知道这个路由器存在于这个路径上。
  - 接着tracert 再送出另一个TTL是2 的数据包，发现第2 个路由器..... tracert 每次将送出的数据包的TTL 加1来发现另一个路由器，这个重复的动作一直持续到某个数据包抵达目的地。

“生存时间”（每途经一个路由器结点自增1）

```
C:\Documents and Settings\XXR>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.
```

途经路由器的IP地址（如果有主机名，还会包含主机名）

三次发送的ICMP包返回时间，相差大说明网络情况变化大，如果有带有星号（*）的信息表示该次ICMP包返回时间超时。

不应发送 ICMP 差错报告报文的几种情况

- 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（如127.0.0.0 或 0.0.0.0）的数据报不发送 ICMP 差错报告报文。

PING (Packet InterNet Groper)

- PING 用来测试两个主机之间的连通性。
- PING使用 ICMP 回送请求与回送回答报文。

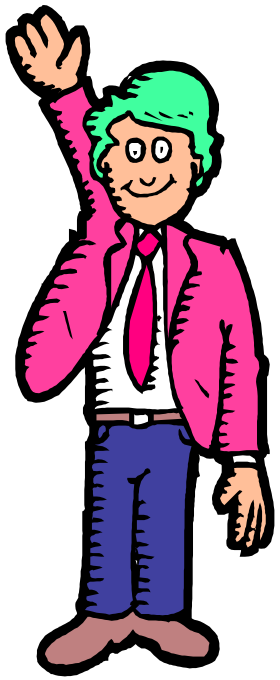
```
C:\Documents and Settings\XXR>ping mail.sina.com.cn

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
```

- PING 是应用层直接使用网络层 ICMP ，没有通过运输层的 TCP 或UDP。



THANK YOU!