Fleschier 渐寒๑

计算机网络-自底向上

2018-12-24 | 2018-12-27 | Computer | 1

> 一些概念性的名词用 *斜体* 标出了

互联网概述

- 计算机网络由若干的结点(node)和连接这些结点的链路(link)组成.结点可以是计算机,集线器,交换机或路由器.
- 网络之间可以通过路由器互连起来,因此互联网是 *网络的网络*.
- 网络把许多计算机连接在一起,而互联网把很多网络通过路由器连接在一起.与网络相连的计算机通常称为 *主机*.
- 以小写字母 `i` 开始的 `internet`(互连网) 是一个通用名词,泛指由多个计算机网络互连而成的计算机网络.
- 以大写字母 `I` 开始到 `Internet`(互联网) 是一个专有名词,特指采用TCP/IP协议族作为通信规则的互联网.
- RFC(Request For Comments)就是 “请求评论”
- 主机A和主机B进行通信,更为全面的说法是 **主机A的某个进程和主机B的某个进程进行通信**

互联网的组成

- 互联网的拓扑结构非常复杂,从其工作方式上来看分为以下两大块:

1. 边缘部分: 这部分是 用户直接使用的 ,由所有连接在互联网上的主机构成
2. 核心部分: 这部分是 为边缘部分提供服务的 ,由大量网络和链接这些网络的路由器组成

- 在客户机和服务器之间通信的过程中,服务请求方和服务提供方 **都要使用网络核心部分所提供的服务**

路由器

- 路由器是一种专用计算机,实现 *分组交换* 的关键部件.其任务就是转发收到的分组.

电路交换

- 必须经过 建立连接(占用通信资源) -> 通话(一直占用通信资源) -> 释放连接(归还通信资源)
  这三个步骤的交换方式就叫 *电路交换*.

分组交换

- 分组交换采用 *存储转发* 技术.
- 通常把要发送的一整块数据叫做 *报文*.
- 把较长的报文划分成一个个更小等长的数据段,在每一个数据段的前面,加上一些必要的控制信息组成 *首部* 后,就构成了一个 *分组*.(分组又称为 包)
- 路由器收到一个分组,先暂时存储一下,检查其首部,查找转发表,按照首部中的目的地址找到合适的端口发送出去,把分组交给下一个路由器.

三种交换方式的比较

1. 电路交换: 整个报文的比特流连续地从源点直达终点,好像在一个管道中传送.
2. 报文交换: 整个报文先传送到相邻结点,**全部存储下来后查找转发表**,转发到下一个结点.
3. 分组交换: 单个分组(**整个报文的一部分**)传送到相邻的结点,存储下来后查找转发表,转发到下一个结点.

计算机网络的类别

- 计算机网络的定义: 计算机网络是有一些通用的,可编程的硬件互联而成的,而这些硬件并非专门用来实现某一特定的目的.
- `WAN` (Wide Area Network):广域网 `MAN` (Metropolitan Area Network):城域网 `LAN` (Local Area Network):局域网 `PAN` (Personal Area Network):个人局域网

计算机网络性能

带宽

- 带宽本来是指某个信号具有的 频带宽度 .信号的带宽是指该信号所包含的各种不同的频率成分所占据的频率 **范围**.例如,传统通信线路上的电话信号的标准带宽是3.1kHz(从300Hz到3.4kHz,这个值代表的是一个 **范围的大小**),这时带宽的单位是 `Hz`
- 带宽的另一个表示是网络中某通道传送数据的能力,即传输数据的速度,此时单位是 `bit/s`

时延

- 总时延 = 发送时延 + 传播时延 + 处理时延 + 排队时延

利用率

- 信道利用率是指某信道有百分之几的时间是被利用的(有数据通过)
- 令 `D0` 表示网络空闲时的时延(也就是网络所能达到的最低时延), `D` 表示网络当前的时延, `U` 为网络的利用率,则有: `D = D0/(1-U)`
- 可以看出如果 `U` (利用率)太高则时延会成指数级增长,一般尽量保证网络的利用率不超过50%.

计算机网络体系结构

OSI的七层协议

- 物理层 - 数据链路层 - 网络层- 运输层 - 会话层 - 表示层 - 应用层

TCP/IP的四层协议

- 网络接口层 - 网际层IP - 运输层(TCP或UDP) - 应用层(各种应用层协议如TELNET,FTP,SMTP等)

五层协议

- 物理层 - 数据链路层 - 网络层 - 运输层 - 应用层

物理层

波特率和比特率

- 比特率(bit rate)又称传信率、信息传输速率(简称信息速率，information rate)。 其定义是：通信线路(或系统)单位时间(每秒)内传输的信息量，即每秒能传输的二进制位数，通常用Rb表示，其单位是比特/秒(bit/s或b/s，英文缩略语为bps)。

- 在二进制系统中，信息速率(比特率)与信号速率(波特率)相等，例如，当系统以每秒50个二进制符号传输时，信息速率为50bit/s，信号速率也为50Bd(波特)。在无调制的情况下，比特率等于波特率；采用调相技术时，比特率不等于波特率。通信系统的发送设备和接收设备必须在相同的波特率下工作，否则会出现帧同步错误。

- 波特率(Baud rate)又称传码率、码元传输速率(简称码元速率)、信号传输速率(简称信号速率，signaling rate)或调制速率。 其定义是：通信线路(或系统)单位时间(每秒)内传输的码元(脉冲)个数；或者表示信号调制过程中，单位时间内调制信号波形的变换次数，通常用RB表示，单位是波特(Bd或Baud，前者规范)。 如果每秒传输1个码元就称为1Bd；如果1码元的时间长短为200ms，则每秒可传输5个码元，那么码元速率(波特率)就是5Bd。

- 波特率(码元速率)并没有限定是何种进制的码元，所以给出波特率时必须说明这个码元的进制。对于M进制码元，比特率(信息速率)Rb与波特率(码元速率)RB的关系式为: $Rb=RB\cdot lbM$.式中：lbM=log2M，表示M的以2为底的对数。显然，对于二进制码元，由于lb2=1，所以Rb=RB，即波特率与比特率在数值上相等，但单位不同，也即二者代表的意义不同。

- 例如，波特率为600Bd，则在二进制时，比特率也为600bit/s；**在四进制时，由于lb4=2，所以比特率为1200bit/s。可见，在一个码元中可以传送多个比特。**

- **可以理解为,比特率在数值上等于二进制编码的波特率**.

信道极限容量&香农公式

- 码元的传输速率越高,或者信号传输距离越远,或噪声干扰越大,或传输媒体质量越差,在接收端的波形失真越严重.

- 限制码元在信道上传输速率的因素有两个:

1. 信道能通过的频率范围:在任何信道中,码元传输的速率是有上限的,传输速率超过此上限,就会产生严重的码间串扰(码元之间的界限变得不再明确,前后都拖了”尾巴”).
2. 信噪比: (突然的噪声会导致误判,0变成1,1变成0).信噪比就是信号的平均功率和噪声的平均功率之比.

- 码元:时间轴上的一个信号的编码单位

香农公式

- 信道的极限传输速率C是: `C = W log2(1+S/N) (bit/s)`

数据编码

常用编码方式

- 不归零制,归零制,曼彻斯特编码,差分曼彻斯特编码

基本的带通调制方法(将数字信号转化为模拟信号)

- 调幅(AM):载波振幅随基带数字信号而变化.例如0或1对应于无载波输出和有载波输出
- 调频(FM):载波频率随基带数字信号而变化.0或10或1对应频率f1和f2
- 调相(PM):载波的初始相位随基带数字信号而变化.例0或1分别对应相位0度或180度

常用的通信介质

导引型传输媒体

- 双绞线:绞合可减少对相邻导线的电磁干扰.绞合度越高最大传输速率就越高,但是价格也越高.
- 同轴电缆:主要用在有线电视网的居民小区.
- 光缆:多模光纤(只适合近距离传输),单模光纤.

非导引型传输媒体

- 短波通信
- 微波接力
- 微信通信

多路复用技术

- 频分复用(FDM):同一时间不同用户使用不同的频带宽度(就是信号的频率)进行通信.
- 时分复用(TDM):不同用户在不同的时间点占用过同样的频带宽度(表现为周期性出现用户的时隙)
- 波分复用(WDM) => 就是光的频分复用

- 码分复用(CDM, Code Division Multiplexing):更常用的是码分多址CDMA.码分复用即将0和1对应到一串特殊的二进制,每个用户不重复(更准确地说,要求是正交的).比如将1表示为001100这一串二进制,则对应的0就是将其每一位都取反,即110011.

数据链路层

- 数据链路层的基本传输单元叫做 *帧*

帧同步问题

> 帧的开始和结束的确认

- 数据链路层的三个功能: 透明传输, 流量控制, 差错检测(只检错不纠错)

封装成帧

- 所有在互联网上传送的数据都以分组(IP数据报)为传送单位,网络层的IP数据报送到数据链路层就成为帧的 数据部分 .对帧的数据部分加上首部和尾部就构成了一个完整的帧.
- 首部和尾部的一个重要作用就是帧定界(确认帧的开始和结束)
- 最大传输单元 `MTU(Maximum Transfer Unit)` : MTU是的是帧的 数据部分 的最大长度.

帧定界

- `SOH(Start Of Header)` :帧开始符(由一串不与数据部分重合的二进制位组成)
- `EOT(End Of Transmission)` :帧结束符
- 帧定界符可以检测数据是否有差错,例如一个帧是否完整

透明传输

- 即用作帧定界的控制字符的比特编码不能出现在数据部分中
- 透明: 某一个实际存在的事物看起来却好像不存在一样
- 在数据链路层透明传输数据表示:**无论什么样的比特组合的数据,都能够按照原样没有差错地通过这个数据链路层**

- 为了解决透明传输的问题,对于数据中出现的控制字符’ EOT’ 和’ SOH’ 的前面插入一个转义字符’ ESC’ ,接收端的数据链路层会在将数据交付网络层之前删除这个转义字符.这种方法称为 字节填充 或者 字符填充 .(如出现转义字符则再加转义字符)

差错检测

- 目前在数据链路层广泛地使用了CRC(Cyclic Redundant Check)的检错技术.

点对点协议ppp

> point to point protocol

- 互联网用户必须连接到某个ISP才能接入互联网,ppp协议就是用户计算机和ISP进行通信时使用的 **数据链路层协议**
- ppp协议支持多种网络层协议(如IP和IPX)和多种类型链路(例如串行(一次发送一个比特)和并行(一次并行地发送多个比特))
- 差错检测: ppp协议对接收端收到的帧进行检测,并 **立即丢弃有差错的帧**
- 最大传送单元: ppp协议对每一种类型的点对点链路设置最大传送单元MTU的标准默认值,如果高层协议发送的分组过长并超过MTU的数组,则 **丢弃这样的帧**

ppp协议的帧格式

- 首部为四个字段, 尾部为两个字段.
- 首部的第一个字段和尾部的第二个字段(最后一个字段)都是 **标志字段**,规定为 `0x7E` .标志字段表示一个帧的开始或结束.因此标志字段就是ppp帧的定界符. **连续两个帧之间只需要一个标志字段,连续两个标志字段表示这是一个空帧**
- 首部中的第二个字段(地址字段A)和第三个字段(控制字段C)没有实际的含义
- 首部的第四个字段是协议字段,占2字节.
- 数据部分(信息字段)长度是可变的,最长不超过1500字节
- 尾部的第一个字段(2字节)是使用CRC的检验序列FCS
- 字节填充: ppp使用异步传输时,转义符定义为 `0x7D` ,填充方法:将信息字段中每一个0x7E字符转化为(0x7D, 0x5E),将出现的0x7D转化为(0x7D, 0x5D),如果出现ASCII码的控制符(即数值小于

0x20),则在该字符前增加一个0x7D,例如0x03转化为(0x7D, 0x03)

局域网的数据链路层

- 局域网的主要特点:网络为一个单位所拥有,且地理范围和站点数目均有限.
- 局域网使用的是数据链路层的协议
- 局域网的优点:

1. 具有广播功能,从一个站点可以很方便地访问全网.
2. 便于系统的扩展和逐渐演变,各设备的位置可以灵活调整和改变.
3. 提高了系统的可靠性,可用性和生存性.

- 局域网的拓扑结构: 星型网,环形网和总线网.
- 局域网工作的层次跨越了数据链路层和物理层.
- 数据链路层在局域网中分为两个子层:

1. 逻辑链路控制层(LLC) (现在使用的局域网协议–*以太2型* 取消了LLC层,只留下MAC层)
2. 介质访问控制层(MAC) (MAC层与物理层使用的介质密切相关)

共享信道

- 由于局域网的广播信道是一人发送,所有人都可以接收,所以要制定发送规则.

静态信道划分

- 使用诸如频分复用,时分复用等技术,用户只要分配到了信道就不会和其他用户发生冲突,但代价太高,不适合局域网使用.

动态媒体接入控制

- 又称为 *多点接入(multiple access)*
- 又分为两类:

1. 随机接入: 所有用户都可以随机地发送信息,但可能发生碰撞,需要解决碰撞的网络协议.
2. 受控接入: 用户发送信息要服从一定的控制,例如集中控制的多点线路 *探寻(polling)*

> 受控接入在目前的局域网中使用的较少

适配器

- 计算机与外界局域网的连接是通过适配器(adapter)进行的.
- 适配器实现的功能包括了数据链路层和物理层.
- 适配器在接收和发送各种帧的时候,不使用计算机的CPU.当适配器收到有差错的帧的时候,就直接丢弃而不通知计算机.**当接收到正确的帧的时候,就使用中断来通知计算机,并交付协议栈中的网络层.当计算机要发送IP数据报时,就由协议栈把IP数据报向下交给适配器,组装成帧后发送到局域网**
- **注意: 计算机的硬件地址(MAC地址)就在适配器的ROM中**

CSMA

- 为了通信简便,以太网使用以下两个措施:

1. 采用 **无连接** 的工作方式,即尽最大努力交付(不可靠的交付).接收站对收到的差错帧直接丢弃(采用CRC检错),至于是否需要重传由高层协议决定(例如TCP协议)
2. 以太网发送的数据都是使用曼彻斯特编码的信号

CSMA/CD

- CSMA/CD: Carrier Sense Multiple Access Protocl with Collision Detection.意思是载波监听多点接入/碰撞检测
- 要点:

1. “多点接入”:许多计算机是以多点接入的方式连接在一根总线上.
2. “载波监听”:就是用电子技术检测总线上有没有其他计算机也在发送.**不管在发送前,发送中,每个站都必须不停地检测信道**.在发送中检测信道,是为了及时发现有没有其它站的发送和本站发送的碰撞,这称为 *碰撞检测*.
3. “碰撞检测”:也就是边发送边监听,也称为 “冲突检测” .一旦发生冲突,总线上传输的信号产生了严重的失真,无法恢复出有用的信息来,因此任何一个正在发送的站一旦检测到冲突其适配器立即停止发送.

- 检测时信道空闲不代表没有站在通信,因为存在传播时延.
- **电磁波在1km电缆的传播时延约为5us**
- 若总线上单程端到端的传播时延记为 `t` ,则最多经过 `2t` 就能确定有没有发生冲突.(发送数据后监听 `2t` 时间,没有冲突即发送成功)

- 使用CSMA/CD协议时,一个站不能边发送边接收,因此使用CSMA/CD的以太网只能进行 *半双工通信*

- 以太网规定最短帧长度为64字节(512bit).因为 **对于10Mbit/s的以太网,争用期为51.2us,而发送512bit需要恰好51.2us,如果数据短于64字节,则可能无法检测出发生冲突的帧**. 因此规定 凡是长度小于64字节的帧都是由于冲突而异常终止的无效帧 .

- 名词概念: `10BASE-T` 这里 `10` 表示10Mbit/s的数据率, `BASE` 表示连接线上的信号是基带信号, `T` 代表双绞线.

- 星型网使用 *集线器(hub)* 作为中心.使用集线器的以太网在逻辑上仍然是一个总线网,各站共享逻辑上的总线,使用的还是CSMA/CD协议.

- 集线器工作在物理层,它的每个接口仅仅简单地转发比特,**不进行碰撞检测**(碰撞检测由各站的适配器完成)

以太网的MAC层

以太网V2的MAC帧

- 以太网V2的MAC帧由五个字段组成.

1. 第一个字段为6个字节的目的地址
2. 第二个字段为6个字节的源地址
3. 第三个字段为2个字节的类型字段(用来标志上一层使用了什么协议,以便把收到的MAC帧交给上一层的这个协议)
4. 第四个字段数据字段,**长度在46-1500字节之间**(由于之前提到帧的最小长度是64字节,所以46字节就是用64减去首部尾部得到的).数据字段就是网络层传下来的IP数据报.
5. 最后一个字段是4字节的帧检验序列FCS(使用CRC检验)

- 从MAC子层向下传到物理层还要添加8个字节(包括7字节的前同步码(1和0的交替码)和1字节的帧开始定界符(定义为10101011)),它的作用是使接收端的适配器在接收MAC帧时能够迅速调整其时钟频率,使它和发送端同步.

- **设置MAC帧长度的限制的原因**:最短64字节的原因在之前以太网帧长度处规定了.最长不能超过1518字节是因为如果数据帧太长,则单一站点占用信道的时间太长影响其他站点分享带宽,同时太长的帧在传输时容易出错.

网桥(bridge)

- 中继器只能识别出位(bit),网桥能够识别数据帧.
- 网桥的升级版为交换机,两者都是工作在数据链路层.

交换机(switch)

- 1990年问世的 *交换式集线器(switching hub)* 很快就淘汰了网桥.交换式集线器称为 以太网交换机 或 第二层交换机
- 以太网交换机实际上就是一个 **多接口的网桥**, 和工作在物理层的转发器,集线器有很大的区别,以太网交换机的每个接口都直接与一个单台主机或另一个以太网交换机相连,并且一般都工作在 **全双工方式**.其内部的交换表(又称为地址表)是通过 *自学习* 算法自动建立起来的.

虚拟局域网

- 使用了特殊的交换机,第三层交换机,即搭载了路由表的交换机,可以提供网络间通讯的功能.

网络层

- 网络层向上只提供简单灵活的,无连接的,尽最大努力交付的数据报服务.
- 传输单元为分组(即IP数据报)

虚电路和数据报的比较

对比方面	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	尽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
当结点出现故障时	所有通过出故障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

对比方面	虚电路服务	数据报服务
分组的顺序	总是按发送顺序到达终点	到达终点的时间不一定按发送顺序
端到端的差错处理和流量控制	可由网络负责,也可以由用户主机负责	由用户主机负责

网际协议IP

网络互联

- 将互联网连接起来需要使用一些中间设备,不同的层次使用的中间设备有不同的名称:

1. 物理层: 转发器(repeater)
2. 数据链路层: 网桥或者桥接器(bridge)
3. 网络层: 路由器(router)
4. 网络层以上各层: 网关(gateway)

IP地址的分类

- A类,B类,C类地址都是由两个固定长度的字段组成: 第一个字段是网络号(net-id), 一个网络号在整个互联网范围内必须是唯一的. 第二个字段是主机号(host-id), 一台主机号在它前面的网络号所指明的网络范围内必须是唯一的.

- IPV4地址长度是32位,A类,B类和C类的网络号分别是1个字节(8位),2个字节(16位)和3个字节(24位)长,在网络号最前面有1-3位类别位,A,B,C类网络分别为0,10和110.

A类地址

- A类地址网络号占一个字节,只有7位可以使用(第一位固定为0),但可指派的网络号是126个(2^7 - 2),**减2的原因是:** (1)网络号全为0的IP地址是个保留地址,意思是”本网络”.(2) 网络号为127的网络地址保留作为本地软件 *环回测试(loopback test)* 本机的进程之间的通信之用.

- A类地址主机号占3字节,可以指派的主机数为2^24-2, 这里 **减2的原因是:** (1)全0的主机号字段表示该IP地址是”本主机”所连接到的 *单个网络地址* (例如: 一台主机的IP地址是 `5.6.7.8` ,则 `5.0.0.0` 表示这个主机所在的 *网络地址*) (2)全1的主机号表示 **所有的(all)**, 因此全1的主机号表示该网络上的所有主机.

> 例如: B类地址128.7.255.255表示”在网络128.7.0.0上的所有主机”,A类地址0.0.0.35表示”在这个网络上主机号为35的主机”

- **A类地址第一个可指派的网络号为1.0.0.0,最后一个可指派的网络号为126.0.0.0**

B类地址

- B类地址网络号有2个字节,最前面两位已经固定(为10),值剩下14位.**由于B类地址128.0.0.0是不指派的**,所以可以指派的 **B类最小网络地址是128.1.0.0**.因此B类可指派的网络数为2^14-1.
- B类地址主机号占2字节, 由于全0和全1的主机号不能指派(理由同A类地址),所以可以指派的主机数为2^16-2
- **B类地址第一个可指派的网络号为128.1.0.0,最后一个可指派的网络号为191.255.0.0**

C类地址

- C类地址有3个字节的网络号,前3个字节固定(为110),C类地址192.0.0.0也是不指派的,指派的 **最小的C类网络地址是192.0.1.0**,因此C类地址可指派的网络总数是2^21-1.
- 每个C类地址的最大主机数为2^8-2
- **C类地址第一个可指派的网络号为192.0.1.0, 最后一个可指派的网络号为223.255.255.0**
- 一般不使用的特殊IP地址
网路号	主机号	源地址使用	目的地址使用	代表意思
—	—	—	—	—
0	0	可以	不可以	在本网络上的本主机
0	host-id	可以	不可以	在本网络上的某台主机(host-id)
全1	全1	不可以	可以	在本网络上进行广播
net-id	全1	不可以	可以	对net-id上的所有主机进行广播
127	非全0或全1的 任何数	可以	可以	用于本地软件环回测试

IP地址特点

- 实际上,IP地址标志一台主机(或路由器)和一条链路的 *接口*.当一台主机同时连接到两个网络上时,该主机就必须同时具有两个相应的IP地址,其网络号必须是不同的.
- 按照互联网的观点, 一个网络就是具有相同网路号net-id的主机的集合.

地址解析协议ARP

- ARP协议根据IP地址获取物理地址.

- 由于网络层使用的是IP地址,但是在实际网络的链路上传送数据帧时,最终还是必须使用该网络的硬件地址.但IP地址和下面的网络的硬件地址之间由于格式的不同而不存在简单的映射关系(IP地址32位而局域网的硬件地址是48位).此外,在一个网络上可能经常会有新的主机加进来或者撤走旧的主机.更换网络适配器也会使主机的硬件地址改变.

- 地址解析协议ARP解决这个问题的方法是,在主机ARP高速缓存中存放一个从IP地址到硬件地址的映射表,并且这个映射表还经常动态更新.

即IP数据报

格式

- 一个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.在首部固定部分后面是一些可选字段,其长度是可变的.

首部固定部分中的各字段

> 一共12个字段

- 版本: 占4位,指IP协议的版本,通信双方使用的IP协议的版本必须一致.

- 首部长度: 占4位,可表示的最大十进制数值是15.**注意,这个长度的单位是32位字(1个32位字是4字节),即当值为1111(15)时,表示首部长度为60字节**.当IP分组的首部长度不是4字节的整数倍时,必须利用最后的填充字段加以填充.

- 区分服务: 占8位, 用来获得更好的服务.(一般情况下都不使用这个字段)

- 总长度: 占16为,因此最大值为2^16-1=65535.总长度指首部和数据之和的长度,单位为字节.一般下面的数据链路层都规定了MTU(Maximum Transfer Unit),即最大传送单元.一个IP数据包封装成帧之后长度一定不能超过这个值,否则就要进行分片.

- 标识(identification): 占16位.IP软件在存储器中维持一个计数器,每产生一个数据报,计数器就加1,并将此值赋给标识字段.但这个不是序号,因为数据报不存在顺序接收.

- 标志(flag): 占3位,目前只有前两位有意义.

1. 标志字段中的最低位记为 `MF(More Fragment)` . MF=1表示后面“还有分片”的数据报.MF=0表示这已是若干数据报片的最后一个.
2. 标志字段中间一位记为 `DF(Don't Fragment)` ,意思是”不能分片”.只有当DF=0时才允许分片.

- 片偏移: 占13位.片偏移指出:较长的分组在分片后,某片在原分组中的相对位置.也就是说,相对于用户数据字段的起点,该片从何处开始.**片偏移的单位是8字节**,*所以说每个分片的长一定是8字节(64位)的整数倍.*

- 生存时间: 占8位.英文缩写是TTL(Time To Live),表明这是数据包在网络中的寿命.其目的是防止无法交付的数据包无限在网络中兜圈子.路由器每转发一次就将数据报的TTL的值减1,如果某个数据报的TTL=0,则路由器接收之后直接丢弃而不转发.

- 协议: 占8位.协议字段指出此数据包携带的数据使用的是何种协议, 以便使目的主机的IP层知道将数据部分上交给哪个协议进行处理.

- 首部校验和: 占16位.**这个字段只检验数据报的首部,但不包括数据部分**.每经过一个路由器,都要重新计算一遍校验和(因为生存时间,标志,片偏移等可能会变化).

- 源地址:占32位

- 目的地址: 占32位

IP转发分组的流程

- 在路由表中,对每一条路由最主要的是以下两个信息(还会包括其他的一些信息):(目的网络地址, 下一跳地址)

- 只有到达最后一个路由器时,才试图向目的主机进行交付.

划分子网和构造超网

子网

- 在IP地址中增加了一个”子网号字段”,使得两级IP地址变为三级IP地址.这种做法叫做 *划分子网(subnetting)*.

- `IP地址 ::= {<网络号>,<子网号>,<主机号>}`

- 子网号字段占用了主机号字段的长度,因此划分子网会使主机数目减少,**但是不改变原来的网络号**,因此对外这个网络还是原来的网络.

子网掩码

- 从IP数据报的首部无法看出源主机或者目的主机所连接的网络是否进行了子网划分,就必须使用*子网掩码*

- 例如将IP地址145.13.3.10与子网掩码255.255.255.0做与操作,得到的网络地址是145.13.3.0.
- 使用子网划分后,路由表必须包含一下三项内容:目的网络地址,子网掩码和下一跳地址.

超网

无分类编址CIDR

- 使用 *变长子网掩码VLSM(Variable Length Subnet Mask)* 可以进一步提高IP地址资源的利用率.在VLSM的基础上又进一步研究出无分类编址方法,正式名称是 *无分类域间路由选择CIDR(Classless Inter-Domain Routing)*
- CIDR的主要特点有两个:

1. CIDR消除了传统的A类, B类和C类的地址以及划分子网的概念.CIDR把32位的IP地址分成前后两个部分,前面部分是 网络前缀(`network-prefix`),用来指明网络,后面部分用来指明主机.因而CIDR是无分类的两级编址,记法为:
   `IP地址:={<网络前缀>,<主机号>}`.CIDR还使用”斜线记法”,即在IP地址后面加上斜线 `/`,然后写上网络前缀所占的位数.
2. CIDR把网络前缀都相同的连续的IP地址组成一个”CIDR地址块”,我们只要知道CIDR地址块中的任何一个地址,就可以知道这个地址块的起始地址和最大地址,以及地址块中的地址数.例如:
   128.14.35.78/20 = **10000000 00001110 0010** 0011 00000111 (20位网络前缀已加粗)
   所以最小地址可以很方便地写出: 128.14.32.0(**10000000 00001110 0010**0000 00000000),
   最大地址也可以很方便写出:128.14.47.255(**10000000 00001110 0010** 1111 11111111)

- 为了更方便进行路由选择,CIDR使用32位 *地址掩码*(address mask), 地址掩码由一串1和一串0组成,1的个数就是网络前缀的长度.(原理与子网掩码一致,都是通过与操作获取网络地址)
- CIDR仍然可以划分子网,但子网的网络前缀比整个单位的网络前缀长一些.
- CIDR记法有很多种,例如地址块 `10.0.0.0/10` 可以简写为 `10/10`,也就是把点分十进制中低位连续的0省略.
- 另一种简化表示法是在网络前缀后加一个 `*`,如 `00001010 00*`,表示 `*` 之前是网络前缀,而 `*` 表示IP地址中的主机号,可以是任意值.
- 使用CIDR时, 由于采用了网络前缀这种记法,IP地址由网络前缀和主机号这两个部分组成,因此路由表中的项目也要相应地改变.这时每个项目由”网络前缀”和”下一跳地址”组成.

- 但是查找路由表 **可能会得到不止一个匹配结果**, 因此 **应当从匹配结果中选择具有最长网络前缀的路由**,这叫做 *最长前缀匹配(longest-prefix matching)*. 这因为网络前缀越长,其地址块就越小,因而路由也就越具体.因此在转发的时候是选择匹配到的地址中更具体的.

- 最长前缀匹配又称为 *最长匹配* 或 *最佳匹配*

路由选择算法

- 从路由算法能否随网络的通信量或拓扑自适应地进行调整变化来划分, 则只有两大类: *静态路由选择策略* 和 *动态路由选择策略*

- 静态路由选择也叫 *非自适应路由选择*, 适用于简单的小网络.

- 动态路由选择也叫 *自适应路由选择*, 适用于复杂的大网络.(因为实现较复杂,开销也比较大)

分层次的路由选择协议

- 内部网关协议IGP(Interior Gateway Protocol): 在一个自治系统内部使用的路由选择协议.(如RIP和OSPF协议)

- 外部网关协议EGP(External Gateway Protocol): 若源主机和目的主机处在不同的自治系统中(这两个自治系统可能使用不同的内部网关协议),当数据包传到自治系统的边界时,就要使用外部网关协议.

IP多播

- 多播和单播的区别: 例如要向100台主机发送同一份数据, 单播则要发送90份相同的副本给每个主机,而多播只要发送一个副本就可以了(不需要复制分组).

虚拟专用网VPN和网络地址转换NAT

- RFC1918指明了一些 *专用地址*(private address),这些地址只能用于一个机构内部通信,即只能用作本地地址不能用作全球地址.

- 例如:

1. `10.0.0.0` 到 `10.255.255.255`
2. `172.16.0.0` 到 `172.31.255.255`
3. `192.168.0.0` 到 `192.168.255.255`
   这三个地址相当于一个A类网络,16个连续的B类网络和256个连续的C类网络.

VPN(Virtual Private Network)

- 使用IP隧道技术实现虚拟专用网.
- 例子: 某个机构两个场所建立了专用网A和B,其网络地址分别为专用地址10.1.0.0和10.2.0.0.现在这两个场所需要 **通过公用的互联网构成一个VPN**.显然 **每个场所至少有一个路由器具有合法的全球IP地址**,这个具有全球IP地址的路由器在专用网内部的接口地址则是专用网的本地地址.即两个场所内主机的相互通信借由连到公网的路由器来转发和接收实现.

NAT(Network Address Translation)

- 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全球IP地址.**所有使用本地地址的主机和外界通信的时候,都要在NAT路由器上将其本地地址转换成全球IP地址**
- 当NAT路由器具有N个全球IP地址的时候,专用网内部最多 **同时** 有N台主机可以接入互联网.
- 现在常用的NAT转换表把运输层的 *端口号* 也用上, 这样就可以使多个拥有本地地址的主机公用一个NAT路由器上的全球IP地址.

运输层

- 网络层只是将分组从一台主机传输到另一台主机,而运输层则再将这一通信过程精确到主机中的进程.

> IP协议虽然能把分组送到目的主机, 但是这个分组还停留在主机的网络层而没有交付主机在的应用程序.从运输层角度来看,通信真正的端点并不是主机而是主机中的进程.也就是说, 端到端的通信是应用进程之间的通信.

- TCP/IP运输层的两个主要协议: **用户数据报协议UDP(User Datagram Protocol)** 和 **传输控制协议TCP(Transmission Control Protocol)**
- 传送的数据单元根据协议的不同,分别称为 **TCP报文段** 和 **UDP用户数据报**
- 使用UDP和TCP协议的各种应用和应用层协议
应用	应用层协议	运输层协议
—	—	—
名字转换	DNS(域名系统)	UDP
文件传送	TFTP(简单文件传送协议)	UDP
路由选择协议	RIP(路由信息协议)	UDP

IP地址配置	DHCP(动态主机配置)	UDP
网络管理	SNMP(简单网络管理协议)	UDP
远程文件服务器	NFS(简单网络管理协议)	UDP
IP电话	专用协议	UDP
流式多媒体通信	专用协议	UDP
多播	IGMP(网际组管理协议)	UDP
电子邮件	SMTP(简单邮件传送协议)	TCP
远程终端接入	TELNET(远程终端协议)	TCP
万维网	HTTP(超文本传送协议)	TCP
文件传送	FTP(文件传送协议)	TCP

运输层的端口

- 由于进程的创建和撤销是动态的,所以通信的一方几乎无法识别对方机器上的进程.解决这个问题的方法就是在运输层使用 *端口协议号(protocol port number)*,通常简称为 *端口*.
- 不同的端口用来标志不同的应用进程.虽然通信的终点是应用进程,但是只要把所传送的报文交到目的主机的某个合适的目的端口,剩下的工作(交付目的进程)就由TCP和UDP来完成.
- 这种在协议栈层间的抽象的协议端口是 **软件端口**,和路由器或交换机的硬件端口是完全不同的概念.硬件端口是不同硬件设备进行交互的接口, 而软件端口是应用层的各种协议进程与运输实体进行层间交互的一种地址.
- TCP/IP的运输层用一个16位的端口号来标志一个端口. **注意,端口号只有本地意义,只是为了标记本计算机应用层中各个进程在和运输层交互时的层间接口**
- 互联网上的计算机通信采用客户-服务器方式(C/S方式).客户在发起通信请求的时候,必须知道对方服务器的IP地址和端口号.
- 运输层端口号分为两大类:

一. 服务器端使用的端口号

- 这里又可以分为两类: **熟知端口号** 和 **登记端口号**.
- 熟知端口号是最重要的一类, 又叫做 *系统端口号* ,数值为0-1023.
- **需要记住的常用端口号**:

应用程序	熟知端口号	协议类型

应用程序	熟知端口号	协议类型
FTP	21	TCP
TELNET	23	TCP
SMTP	25	TCP
DNS	53	UDP
TFTP	69	UDP
HTTP	80	TCP
HTTPS	443	TCP
SNMP	161	UDP
SNMP(trap)	162	UDP

二. 客户端使用的端口号

- 数值为49152-65535
- 由于这类端口号仅在客户进程运行时才动态选择,因此又叫做 *短暂端口号*
- 当服务器进程收到客户进程的报文时,就知道了客户进程现在所使用的端口号,因而可以把数据发送给客户进程.当通信结束后,刚才使用的客户端口号就不复存在了,这个端口号就可供其他客户进程使用.

连续ARQ协议

- 将流量控制和差错控制结合了起来.
- ARQ协议规定,发送方每收到一个确认,就把发送窗口向前滑动一个分组的位置.于是就可以接着发送下一个分组进入窗口的分组.
- 接收方一般采用累积却认的方式,也就是说接收方不必对收到的分组逐个发送确认,**只要对按序到达的最后一个分组发送确认**,这表示:到这个分组为止的所有分组都已经正确收到了.
- 累积确认的优点是:容易实现,即使确认丢失也不必重传.缺点是:不能向发送方反映出接收方已经正确收到的所有分组的信息.(因为可能存在可能没有按序到达的分组,比如收到了1,2,3,5四个分

组,只能对前三个进行确认,而5号分组到达的消息发送方没有办法知道,而且很有可能会重传5号分组)

TCP可靠传输的实现

以字节为单位的滑动窗口

- 发送窗口表示: 在没有收到B的确认的情况下,A可以连续地把窗口内的数据都发送出去.**凡是已经发送过的数据,在未收到确认之前都必须暂时保留以便在超时重传时使用.**
- 接收方会把自己的接收窗口的数值放在窗口字段中发送给对方,这相当于告诉发送方,自己的剩余缓冲区域的大小.**因此A的发送窗口一定不能超过B的接收窗口的值.**
- 发送窗口的位置由窗口的前后沿共同决定.发送窗口的后沿不动有两种可能:**(1)没有收到新的确认. (2)收到了新的确认但是对方通知窗口缩小了,使得前沿刚好不动.**
- 发送缓存用来暂时存放:

1. 发送应用程序发送给发送方TCP准备发送的数据
2. TCP已发送出但尚未收到确认的数据

- 接收缓存用来暂时存放:

1. 按序到达的,但尚未被接收应用程序读取的数据
2. 未按序到达的数据

拥塞控制

- 拥塞控制与流量控制的区别: 流量控制管理的是两个点,而拥塞控制管理的是整个网络.
- 计算机网络中的链路容量(即带宽),交换结点中的缓存和处理机等,都是网络的资源.在某段时间,对网络中某一资源的需求超过了该资源所能提供的可用部分,网络性能就要变坏,这种情况就叫 *拥塞(congestion)*
- 问题的实质是,整个系统的各个部分不匹配,只有所有部分平衡了,问题才会得到解决.
- *拥塞控制* : 防止过多的数据注入到网络中,这样可以使网络中的路由器或链路不致过载.拥塞控制是一个全局性的过程,涉及到所有的主机,所有的路由器,以及降低网络传输性能的所有因素.
- 拥塞控制可以分为 *开环控制* 和 *闭环控制*.
- 开环控制就是事先考虑所有的因素,一旦开始运作就不再更改.

- 闭环控制是基于反馈回路的概念.

TCP的拥塞控制方法

慢开始和拥塞避免

- 发送方维护一个叫做 *拥塞窗口cwnd(congestion window)* 的状态变量. **发送方让自己的发送窗口等于拥塞窗口**.
- 判断网络拥塞的依据是 **出现了超时**.
- 慢开始: 由小到大逐渐增加拥塞窗口(发送窗口).
- 从慢开始开始指数增长,直到一个阈值(慢开始门限ssthresh),开始加法增长,最后出现超时,再从慢开始从头再来.
- 当cwnd小于ssthresh时,使用慢开始算法(指数增长).当cwnd>ssthresh时,停止使用慢开始算法**改用拥塞避免算法**.(cwnd = ssthresh时两个算法任意选取)
- *拥塞避免算法* 的思路是让拥塞窗口cwnd缓慢地增大,即每过一个RTT就把发送方的拥塞窗口加1.注意拥塞避免并不是完全避免了拥塞,而是是吧拥塞窗口控制为按线性增长,使网络比较不容易出现拥塞.

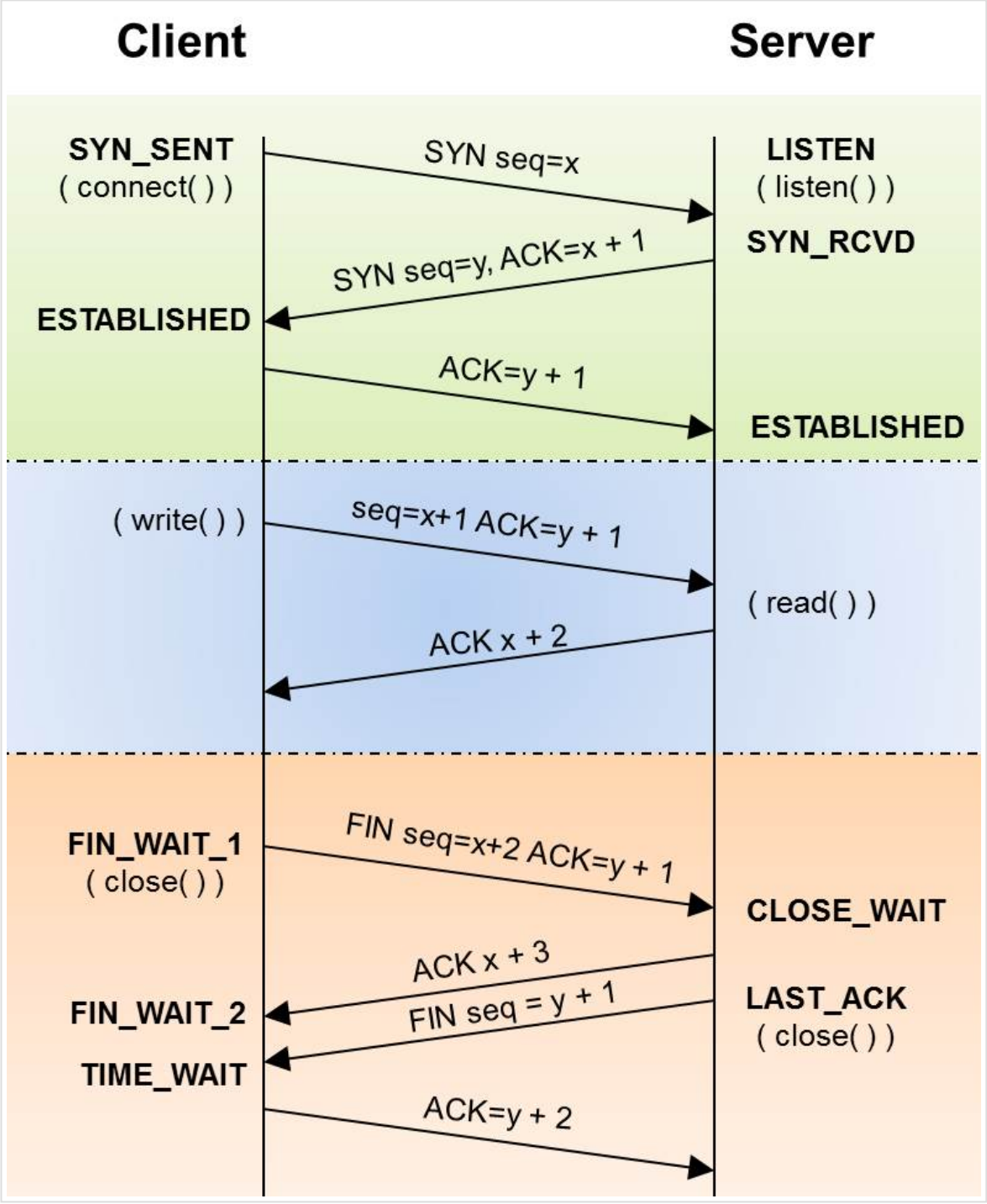
主动队列管理AQM

- *随机早期检测RED(Random Early Detection)*

TCP连接的建立和断开(重点)

> 三次握手, 四次挥手

-



- ACK=1 表示这是一个确认报文, 对之前收到的报文进行确认.
- SYN=1 表示这是一个同步报文段,用于请求同步.
- SYN报文段 **不能携带数据, 并且要消耗一个序号**
- ACK报文段可以携带数据,如果不携带数据 **则不消耗序列号**
- FIN报文段可以携带数据, **但是不管携带与否,都要消耗一个序列号**

三报文握手(TCP连接建立)

- 一开始,B的TCP服务器进程先创建一个 *传输控制块TCB(Transmission Control Block)*, 准备接收客户进程的连接请求. *然后服务器就处于LISTEN(收听)状态*

- A的TCP客户进程也是首先创建传输控制模块TCB,然后在打算建立TCP连接时,向B发出连接请求报文段,**这时首部中的同步位SYN=1**,同时 **选择一个初始序号seq=x**.(SYN报文不能携带数据,但要消耗掉一个序号). *TCP客户进程进入SYN-SENT(同步已发送)状态*

- B收到连接请求报文段后,如同意连接,则向A发送确认.这个确认报文段中SYN位和ACK位都置为1.**确认号ack=x+1(由A发送的确认号为x的请求报文而得到的)**,同时也为自己选择一个初始序号seq=y(**注意B的确认报文的seq字段与A的同步报文的seq无关**),同时,**这个报文也不能携带数据,而且同样要消耗掉一个序列号.** *这时TCP服务器进程进入SYN-RCVD(同步收到)状态*

> 确认报文的确认号ack总是等于被确认的报文的序号seq的值加1

- TCP客户进程收到B的确认后,还要向B给出确认.(**防止A发出的连接请求报文在网络中长时间逗留而没有丢失,在A重新发送请求并且完成连接,结束通信后,B又收到了A之前误以为被丢失而实际只是滞留了的连接请求报文,这时B会以为这是A的一个新的连接请求,从而发送确认给A.而A没有请求,因此会忽略B的这个确认.设想如果没有A的第三次确认,则B以为连接已经建立而无限地等待A发送数据,从而造成资源的浪费.**) 确认报文的ACK置为1,确认号为ack=y+1(确认号为y+1是因为这是对B发送过来的报文的确认报文),自己的序号seq=x+1.(*如果这个报文没有携带数据,则下一个报文的序号仍然是x+1*)这时,TCP连接已经建立,A进入ESTABLISHED(已建立连接)状态.

- 当B收到A的确认后,也进入ESTABLISHED状态.

四报文挥手(TCP连接释放)

- 数据传输结束后,通信双方都可以释放连接(释放连接是单向的,可以一方停止向另一方发送数据,但另一方还可以向这一方发送数据).(**同时也说明,两个方向都要释放一遍连接**)

- A的应用进程先向其TCP发出连接释放报文段,并停止再发送数据,主动关闭TCP连接.A把连接释放报文段首部的终止控制位FIN置1,其序号seq=u(**u等于前面A已经传送过的数据的最后一个字节的序号加1**).*这时A进入FIN-WAIT-1(终止等待1)状态,等待B的确认.*

- B收到连接释放报文段后立即发出确认,确认号ack=u+1.这个报文段自己的序号为v(v等于B前面已经传送过的数据的最后一个字节的序号加1).*然后B就进入CLOSE-WAIT(关闭等待)状态.* **这时TCP服务器进程会通知高层应用进程,因而从A到B这个方向上的连接就释放了,这时TCP处于半关闭(half-close)状态,即A已经没有数据要向B发送了,但是如果B要发送数据,A仍然要接收.也就是说从B到A这个方向的连接还未关闭,这个状态可能会持续一段时间.**

- A收到来自B的确认后,*就进入FIN-WAIT2(终止等待2)状态*,等待B发出的连接释放报文段.

- 若B也没有要向A发送的数据,其应用进程就通知TCP释放连接.这时B发出的连接释放报文段FIN=1.假定B的序号为w(因为在半关闭状态B可能又发送了一些数据),此时B还必须重复上次发送过的确认号ack=u+1.*这时B就进入了LAST-ACK(最后确认)状态*,等待A的确认.

- A在收到B的连接释放报文段后,在确认报文段中把ACK置1,确认号ack=w+1,序号seq=u+1.*进入TIME-WAIT(时间等待)状态*. **注意!此时TCP连接没有释放掉,必须经过时间等待计时器(TIME-WAIT timer)设置的时间2MSL后,A才进入CKISED状态**.时间 *MSL* 叫做 *最长报文段寿命(Maximum Segment Lifetime)*.

- 这里A需要等待2MSL(**B在收到确认后就立即进入CLOSED状态无需等待**)有两个理由:

1. 为了保证A发送的最后一个ACK报文段能够到达B(A发送的ACK确认报文段可能丢失,B会超时重传这个FIN+ACK报文段,而A就能在2MSL时间段内收到重传的FIN+ACK报文段,并重新发送一次确认,重启2MSL计时器).
2. 防止”已失效的连接请求报文段”出现在本连接中.再经过2MSL,就可以使本连接持续时间内所产生的所有报文段从网络中消失,确保下一个连接不会收到旧的无效报文段而引发错误.

- 最后,TCP还设置 *保活计时器(keepalive timer)*,在服务器长时间(一般为2小时)没有收到客服的数据(服务器每收到一次客户数据就重置这个等待时间),服务器就发送一个探测请求报文,之后每隔75秒发送一次,若连续10次没有收到客户响应,则关闭这个连接.(为了防止客户机出现故障)

应用层

域名系统DNS

- 域名系统DNS(Domain Name System)能够把互联网上的主机名字转换为IP地址.

- 每一个域名都由 *标号(label)* 序列组成,各标号之间用点隔开.例如 `mail.cctv.com` ,其中 `mail` 为三级域名, `cctv` 为二级域名, `com` 为一级域名.(级别最低的域名写在最左边)

- 每一个标号不超过63字符,也不区分大小写.

域名解析

- 过程如下:当某一个应用进程需要把主机名解析为IP地址时,该应用进程就调用 *解析程序(resolver)*, 并称为DNS的一个客户,把待解析的域名放在DNS请求报文中,以UDP用户数据报方式发给本地域名服务器.(使用UDP是为了减小开销) 如果本地域名服务器不能回答该请求,则此域名服务器就暂时成为DNS中的另一个客户,向其他域名服务器发出查询请求.

文件传送协议FTP

- FTP的工作流程:

1. 打开熟知端口(21),使客户进程能够连接上.
2. 等待客户进程发出连接请求
3. 启动从属进程处理客户进程发来的请求.
4. 回到等待状态

万维网WWW

- *万维网WWW(Worl Wide Web)* 是一个大规模的,联机式的信息储藏所.

- **超文本**: 所谓超文本是指 包含指向其他文档的链接的文本(text).也就是说,一个超文本由多个信息源链组成, 而这些信息源可以分布在世界各地,并且数目不受限制.

- **超媒体**: 超媒体与超文本的区别是文档内容不同.超文本文档仅包含文本信息, 而超媒体文档还包含其他表示方式的信息,如图形,图像,声音,动画以及视频图像等.

- 万维网以客户服务器方式工作,浏览器就是在用户主机上的万维网客户程序. **客户程序向服务器程序发出请求,服务器程序就向客户程序送回客户所要的万维网文档**.在一个客户程序主窗口上显示出的万维网文档称为 *页面(page)*.

三要素URL,HTML和HTTP

- 万维网使用 *统一资源定位符URL(Uniform Resource Locator)* 来标志万维网上的各种文档,并使每一个文档在整个互联网的范围内具有唯一的标识符URL.(解决怎样标志分布在整个互联网上的万维网文档)

- *超文本传送协议HTTP(HyperText Transfer Protocol)*: HTTP是一个应用层协议,使用TCP连接进行可靠的传送.HTTP协议使得万维网客户程序与万维网服务器程序之间的交互严格遵守这一标准.(解决用什么样的协议来实现万维网上的各种链接)

- 万维网使用 *超文本标记语言HTML(HyperText Markup Language)*,使得万维网页面的设计者可以很方便地用链接从本页面的某处链接到互联网上的任何一个万维网页面.(解决各种不同风格的文档都能在各种不同的主机上显示出来的问题)

电子邮件

- TCP/IP协议规定的电子邮件地址的格式如下: 用户名@邮件服务器域名

- 电子邮件最重要的两个标准就是: *简单邮件传送协议SMTP(Simple Mail Transfer Protocol)* [RFC 5321] 和 *互联网文本报文格式* [RFC 5322]

- 邮件读取协议POP3和IMAP(Internet Message Access Protocol).

- 不要把POP3或IMAP与SMTP弄混,发件人的用户代理向发送方邮件服务器发送邮件,以及发送方邮件服务器向接收方邮件服务器发送邮件,都使用SMTP协议. 而POP3或IMAP是用户代理从接收方邮件服务器上读取邮件时所使用的协议.

动态主机配置协议DHCP

- DHCP提供了一种机制,即 *即插即用网(plug-and-play networking).*这种机制允许一台计算机接入新的网络和获取IP地址而不用手工参与.

简单网络管理协议SNMP

- SNMP(Simple Network Management Protocol)的网络管理由三个部分组成: 即SNMP本身, *管理信息结构SMI(Structure of Management Information)* 和 *管理信息库MIB(Management Information Base)*

Blog

C++学习笔记——指针

Total Words Of All Blogs: 89.6k words