

第六章 应用层 万维网概述

万维网是一个大规模、联机式的信息储藏所/资料空间/资源（文字、视频、音频。。。），是无数个网络站点和网页的集合。

万维网的内核部分是由三个标准构成的：

(1) 统一资源定位符（URL）。负责标识万维网上的各种文档，并使每个文档在整个万维网的范围内具有唯一的标识符URL。

(2) 超文本传输协议（HTTP）：一个应用层协议，它使用TCP链接进行可靠的传输，HTTP是万维网客户程序和服务器之间交互所必需严格遵守的协议。

(3) 超文本标记语言（HTML）：一种文档结构的标记语言，它使用一些约定标记对页面上的各种信息、格式进行描述。

统一资源定位符**URL**：可以**唯一标识一个资源**。

URL一般形式：**<协议>://<主机>:<端口>/<路径>**

协议：如http ftp

主机：域名 或 IP地址

端口和路径有时可以省略：

URL不区分大小写。

如：<http://www.pku.edu.cn>

用户通过**点击超链接**（<http://www.baidu.com>）获取资源，这些资源通过**超文本传输协议（HTTP）**传送给使用者。

万维网使用超文本标记语言**HTML**，使得**万维网页面设计者可以很方便地从一个界面的链接转到另一个界面**，并能够在自己的屏幕上显示出来。

万维网**以客服/服务器方式**工作，用户使用的**浏览器就是万维网客户**

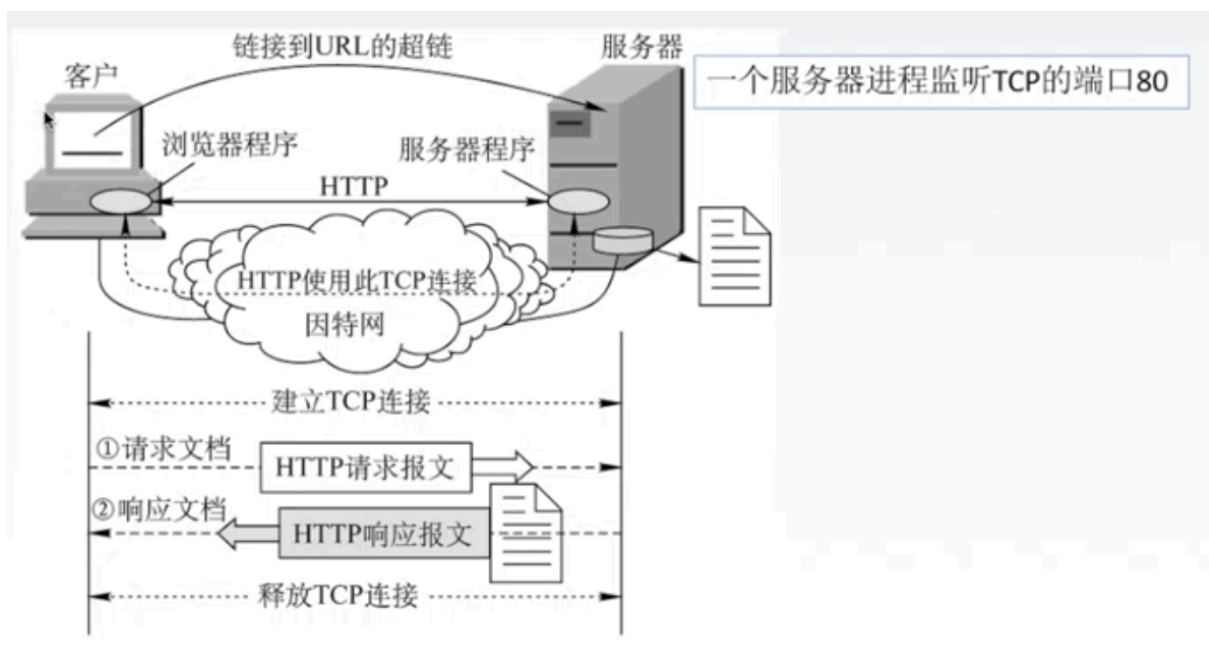
程序，万维网文档所驻留的主机运行服务器程序。

工作流程如下：

- Web用户使用浏览器（指定URL）与Web服务器建立连接，并发送浏览请求
- Web服务器把URL转换为文件路径，并返回信息给Web浏览器。
- 通信完成，关闭连接

超文本传输协议HTTP

HTTP定义了浏览器（万维网客户进程）怎样向万维网服务器请求万维网文档，以及服务器怎样把文档传送给浏览器。



用户浏览页面方法：

- 1、输入URL
- 2、点击超链接

具体过程：

- 1.浏览器分析URL
- 2.浏览器向DNS请求解析IP地址
- 3.DNS解析出IP地址
- 4.浏览器与服务器建立TCP连接
- 5.浏览器发出取文件命令
- 6.服务器响应
- 7.释放TCP连接
- 8.浏览器显示

HTTP的特点：

1、HTTP协议是**无状态**的。即**同一个客户第二次访问同一个服务器上的页面时，服务器的响应与第一次被访问时的相同**。通常使用Cookie加数据库的方式来跟踪用户的活动。

Cookie是存储在用户主机中的文本文件，**纪录一段时间内某用户的访问记录（使用识别码识别，如“123456”）**。目的是提供个性化服务。

2、**HTTP采用TCP作为运输层协议，但HTTP协议本身是无连接的（通信双方在交换HTTP报文之前不需要先建立HTTP连接）**。

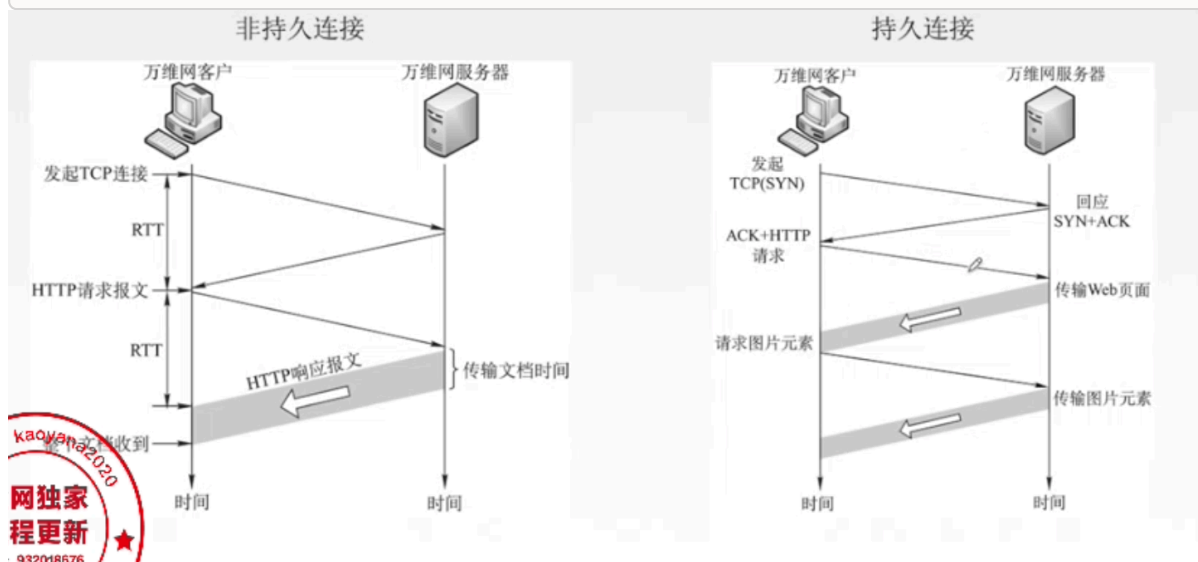
HTTP连接方式

- 持久连接（Keep-alive）
- 非持久连接(Close)

非持久连续，如下图左所示

(1) 每个网页元素对象的传输都需要单独建立一个TCP连接。(第三次握手的报文段中稍带了客户对万维网文档的请求)

(2) 也就是说, 请求一个万维网文档所需的时间是该文档的**传输时间**(与文档大小成正比) 加上**两倍往返时间RTT**(一个RTT用于TCP连接, 另一个RTT用于请求和接受文档)。



持久连接:

分为流水线和非流水线两种方式。

对于**非流水线方式**, 客户在收到前一个响应后才能发出下一个请求。

对于**流水线方式**, 客户每遇到一个对象引用就立即发出一个请求, 因而客户可以逐个地连续发出对各个引用对象的请求。

如果**所有的请求和响应都是连续发送的**, 那么**所有引用的对象共计经历1个RTT延迟**。而不是像非流水线方式那样, 每个引用都必须有1个RTT延迟。

HTTP报文格式

HTTP报文是**面向本文的**, 因此在**报文中的每一个字段都是ASCII码串**。并且**每一个字段的长度都是不确定的**, 有两类HTTP报文。

HTTP的报文结构:

HTTP报文分为:

- (1) 请求报文：从客户向服务器发送的请求报文。
- (2) 响应报文：从服务器到客户的回答。

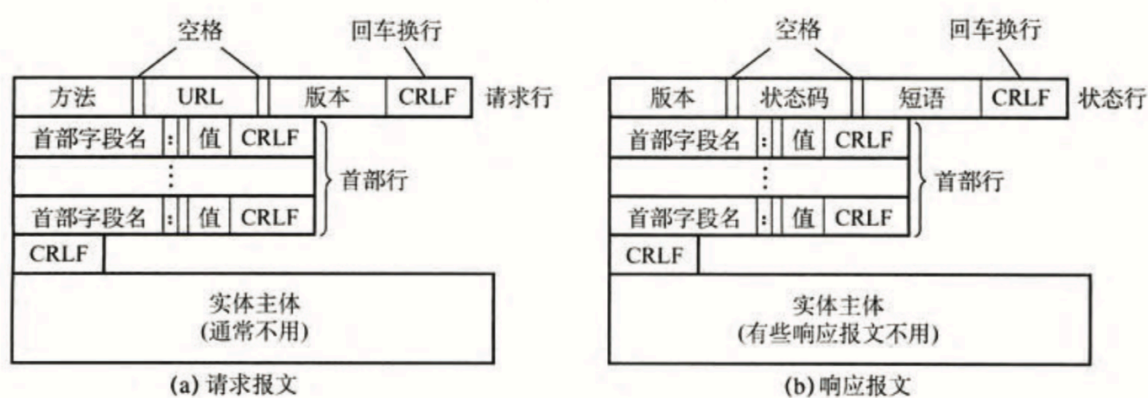


图 6.14 HTTP 的报文结构

HTTP请求报文和响应报文都由三个部分组成。
这两种报文格式的区别是开始行不同。

开始行定义：用于区别是请求报文还是响应报文。在请求报文中的开始行称为请求行，而在响应报文中开始行称为状态行。

开始行的三个字段之间都以空格分隔，最后的“CR”代表“回车”，“LF”代表“换行”。

1、请求报文

(1) 请求行：有三个内容：方法、请求资源的URL及HTTP的版本。

其中，“方法”是对所请求对象进行的操作，这些方法实际上也就是一些命令。如下图为HTTP请求报文中常用的几个方法；

表 6.1 HTTP 请求报文中常用的几个方法

方法（操作）	意 义
GET	请求读取由 URL 标识的信息
HEAD	请求读取由 URL 标识的信息的首部
POST	给服务器添加信息（如注释）
CONNECT	用于代理服务器

(2) 首部行：用来说明浏览器、服务器或报文主体的一些信息。首部可有几行，但可不使用。在每个首部行都有首部字段名和它的值，每行在结束的地方都要有“回车”和“换行”。整个首部结束时，还有一空行将首部行和后面的实体主体分开。

(3) 实体主体：在请求报文中的一般不用这个字段，而在响应报文中也可能没有这个字段。

2、响应报文

(1) 响应报文中的版本：为HTTP的版本。一般为HTTP/1.1

(2) 状态码：

状态码：

1xx表示通知信息的，如请求收到了或正在处理。

2xx表示成功，如接受或知道了。

3xx表示重定向，如要完成请求还必须采取进一步的行动。

4xx表示客户的差错，如请求中有错误的语法或不能完成。

5xx表示服务器的差错，如服务器失效无法完成请求。

202-> Accepted 接收

301->Moved Permanently 转移到新的地址

404->Not Found

常见应用层协议小结如表 6.2 所示。

表 6.2 常见应用层协议小结

应用程序	FTP 数据链接	FTP 控制链接	TELNET	SMTP	DNS	TFTP	HTTP	POP3	SNMP
使用协议	TCP	TCP	TCP	TCP	UDP	UDP	TCP	TCP	UDP
熟知端口号	20	21	23	25	53	69	80	110	161