第 8 章　应用层(2)

课程名称：计算机网络

主讲教师：姚烨

课程代码：U10M11016.02　　第49-50讲

E-MAIL　：yaoye@nwpu.edu.cn

2021 – 2022 学年第一学期

本节内容提要

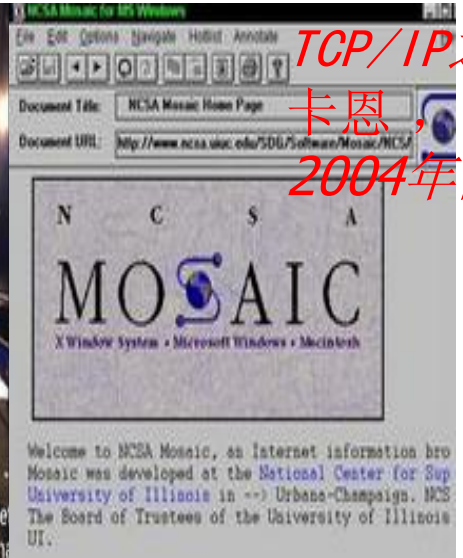
万维网 WWW

- 一、 万维网概述
- 二、 统一资源定位符 URL
- 三、 超文本传输协议 HTTP
- 四、 万维网的文档（HTML）

引言

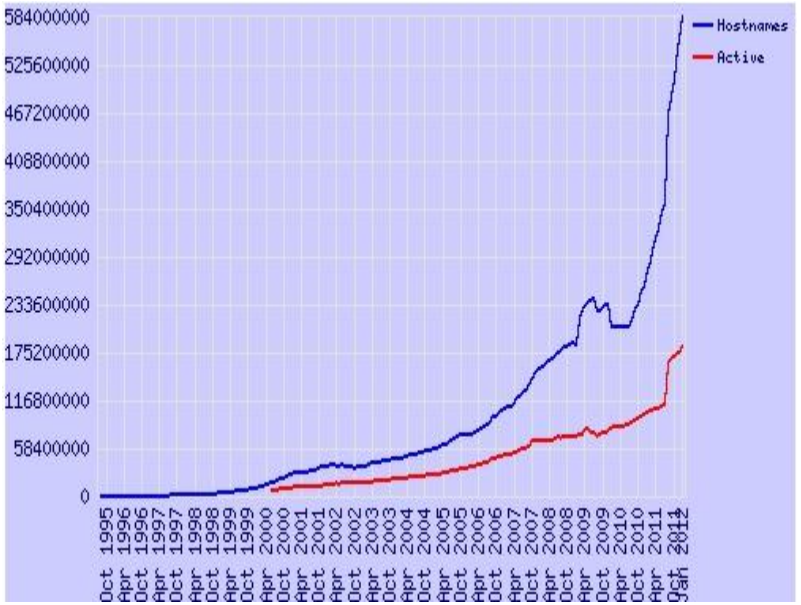
- 在因特网发展早期，FTP传输文件大约占了网络通信流量1/3，到了1995年，万维网（WWW）通信量首次超过了FTP。
- 万维网是欧洲粒子物理实验室的Tim Berners-Lee 于1989.3提出；1993.2月，第一个图形界面浏览器（ Mosaic ）开发成功；1995年Netscape Navigator 浏览器上市。

TCP/IP之父：罗伯特·卡恩，温顿·瑟夫
2004年图灵奖



引言

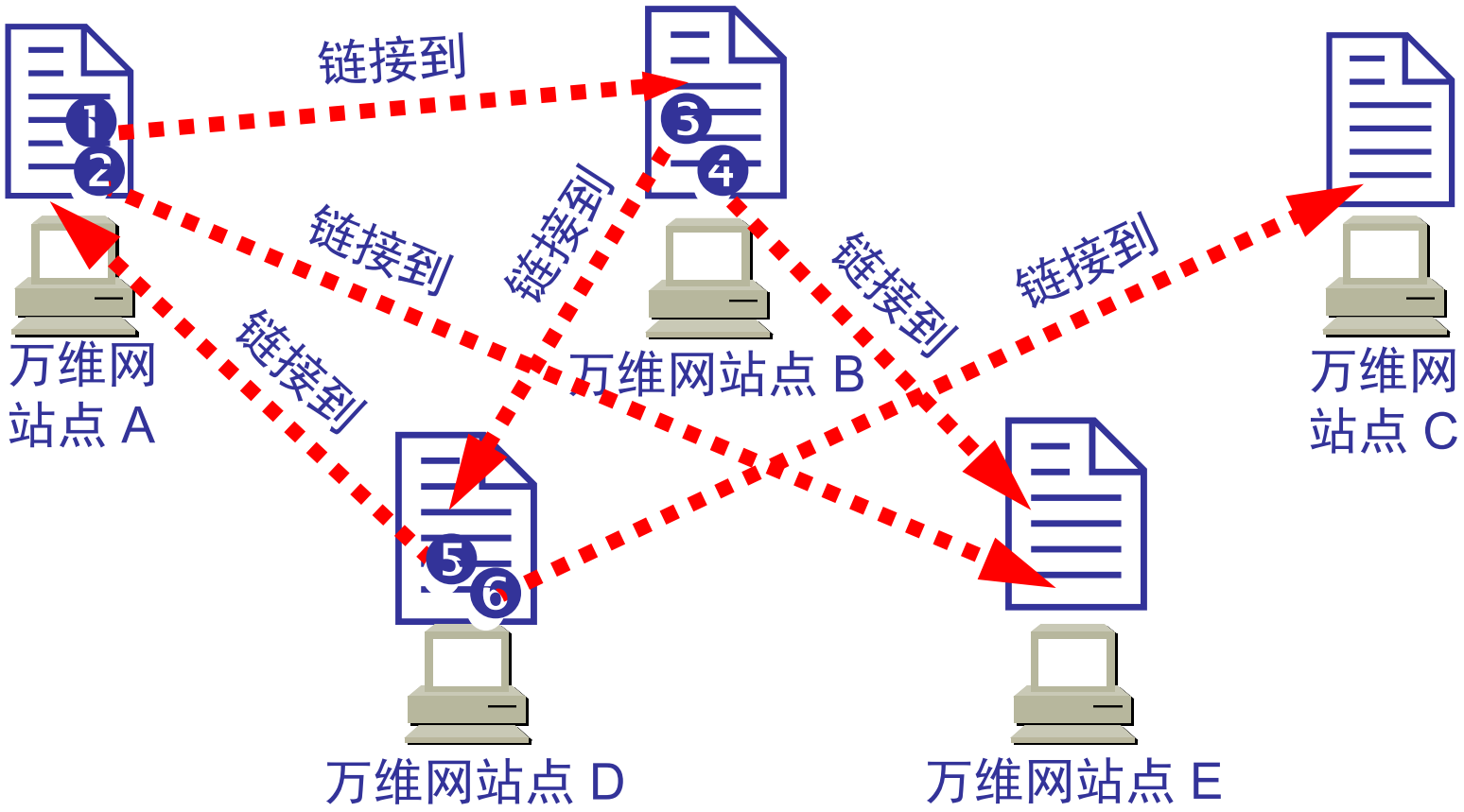
- 万维网出现是因特网发展中一个非常重要里程碑
  - 因特网由少数计算机专家使用变成普通百姓也可参与，实现资源共享。
  - 网站数量呈指数规律增长。



一、 万维网概述

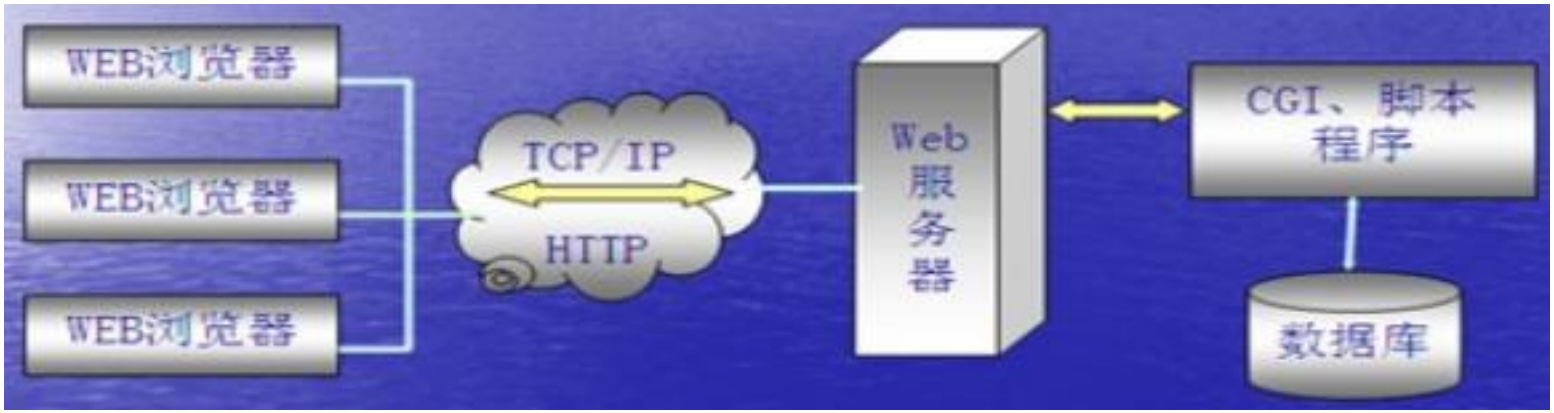
- 万维网 WWW (World Wide Web)并非特殊计算机网络；
- 万维网是一个大规模的、联机式的信息存储与管理方式。
- 采用“链接”方法用户从因特网上一个站点访问另一个站点的网页（超媒体信息:文本、视频、音频、图象），用户可以主动地按需获取丰富的信息。
- 这种访问方式称为“链接”（超链）。

万维网提供分布式服务


链接到
链接到
链接到
链接到
链接到
链接到
1
2
3
4
5
6
万维网
站点 A
万维网站点 B
万维网
站点 C
万维网站点 D
万维网站点 E

万维网的工作方式

- 万维网以客户/服务器方式工作。
- 浏览器就是在用户计算机上的万维网客户程序。
- 万维网服务器由两部分构成。
  - Web服务器程序：万维网文档文档进行管理和传输；
  - Web服务器应用程序：网站（网页+数据库），万维网文档。
- 客户程序向服务器程序发出请求，服务器程序向客户程序返回客户所要的万维网文档或处理结果。



万维网必须解决四个问题

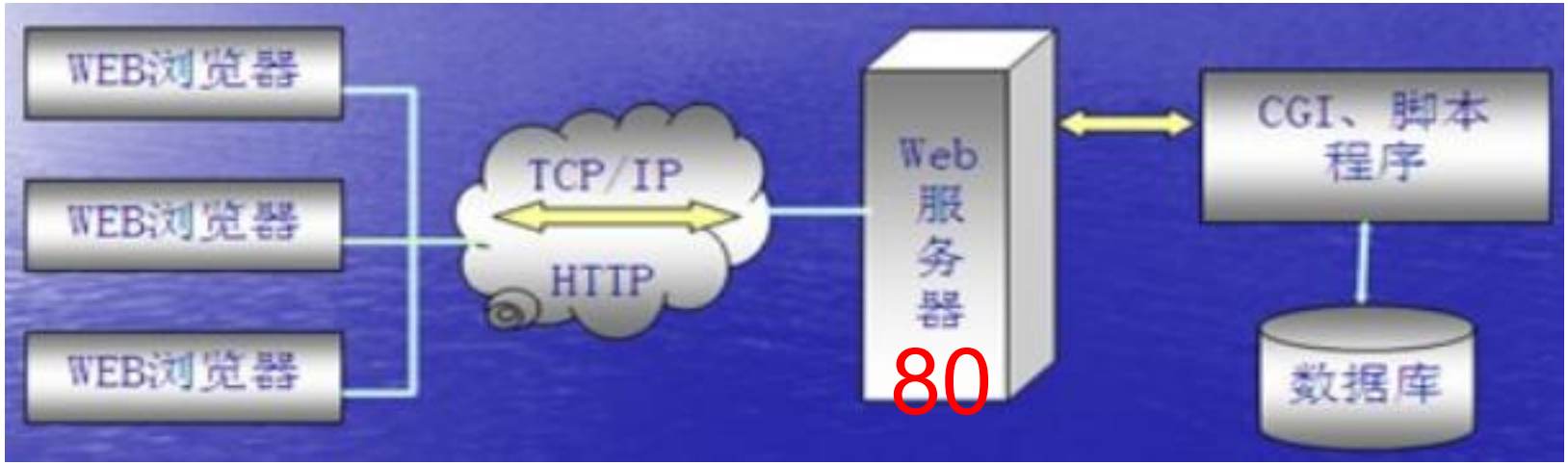
(1) 怎样标识分布因特网上的万维网文档？

- 使用统一资源定位符 URL (Uniform Resource Locator) 来标识万维网上的各种文档。
- 一个文档(文件)因特网围内具有唯一标识符 URL。

万维网必须解决四个问题

(2) 用何协议实现万维网上各种超链的交互通信？

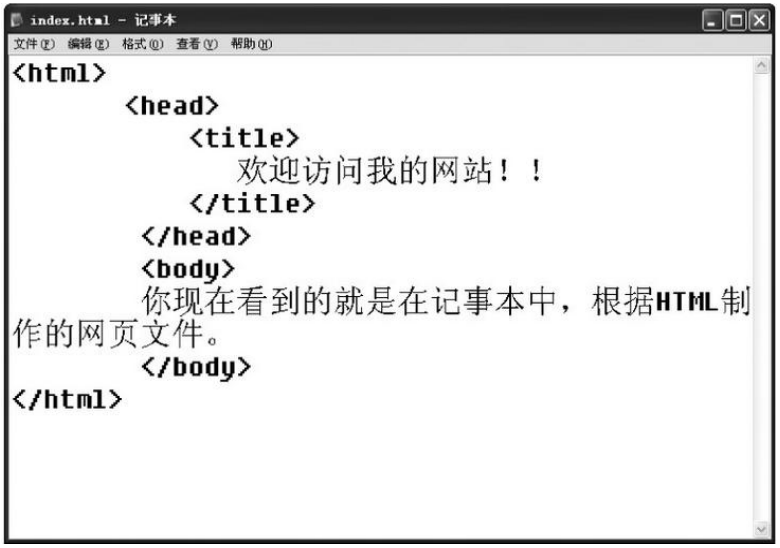
- 万维网客户程序与万维网服务器程序之间进行交互所使用的协议：超文本传送协议 HTTP (HyperText Transfer Protocol)。
- HTTP 是一个应用层协议，在传输层使用 TCP 连接进行可靠的通信。



万维网必须解决的问题

(3) 万维网文档如何在各种计算机上显示，用户如何知道在什么地方存在着超链？

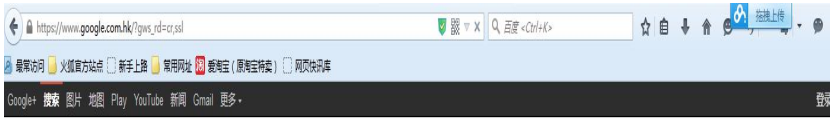
- 万维网文档采用超文本标记语言 HTML (HyperText Markup Language)编写，浏览器利用自身解析器对文档进行解析显示；超链有特殊标识（鼠标手形、特殊颜色，下划线）

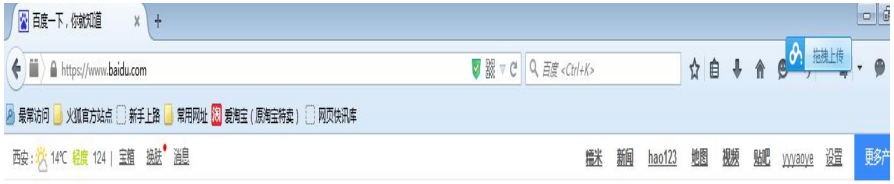


万维网必须解决的问题

(4) 怎样使用户能够很方便地找到所需的信息？

- 为了在万维网上方便地查找信息，用户可使用各种的搜索工具（即搜索引擎）。





本节内容提要

万维网 WWW

- 一、 万维网概述
- 二、 统一资源定位符 URL
- 三、 超文本传送协议 HTTP
- 四、 万维网的文档（HTML）
- 五、 万维网的信息检索系统

二、统一资源定位符 URL

- 统一资源定位符 URL ：可以唯一的标识因特网上任一网络资源（位置+名称）和访问方法。
  - URL 给资源（位置+名称）提供一种抽象的标记和识别方法，用该方法可对资源定位。
  - 只要能够对资源定位，系统就可以对资源进行各种操作：存取、更新、替换和查找其属性。
  - URL 相当于一个文件名在网络范围的扩展。

1. URL 一般形式

- 由以冒号隔开的两大部分组成，并且在 URL 中的字符对大写或小写没有要求。
- URL 的一般形式是：

<协议>://<主机>:<端口>/<路径+名称>

ftp —— 文件传送协议 FTP

http —— 超文本传送协议 HTTP

News —— USENET 新闻

2. 使用 HTTP 的 URL

- 使用 HTTP 的 URL 的一般形式

http://<主机>:<端口>/<路径+名称>

若再省略文件的<路径>项，则 URL 就指到因特网上的某个主页(home page)。

举例：http://www.nwpu.edu.cn:80/index.html(.htm)
http://www.nwpu.edu.cn
www.nwpu.edu.cn

不论协议是HTTP、FTP、USENET（新闻组），客户端都可以是浏览器。

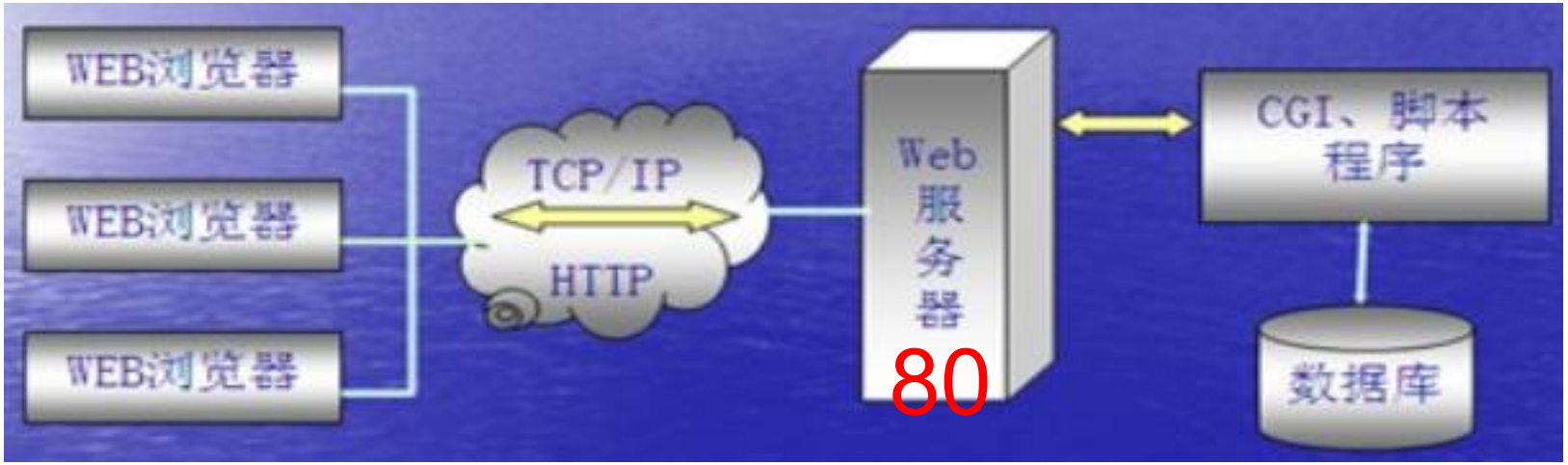
本节内容提要

万维网 WWW

- 一、 万维网概述
- 二、 统一资源定位符 URL
- 三、 超文本传输协议 HTTP
- 四、 万维网的文档（HTML）
- 五、 万维网的信息检索系统

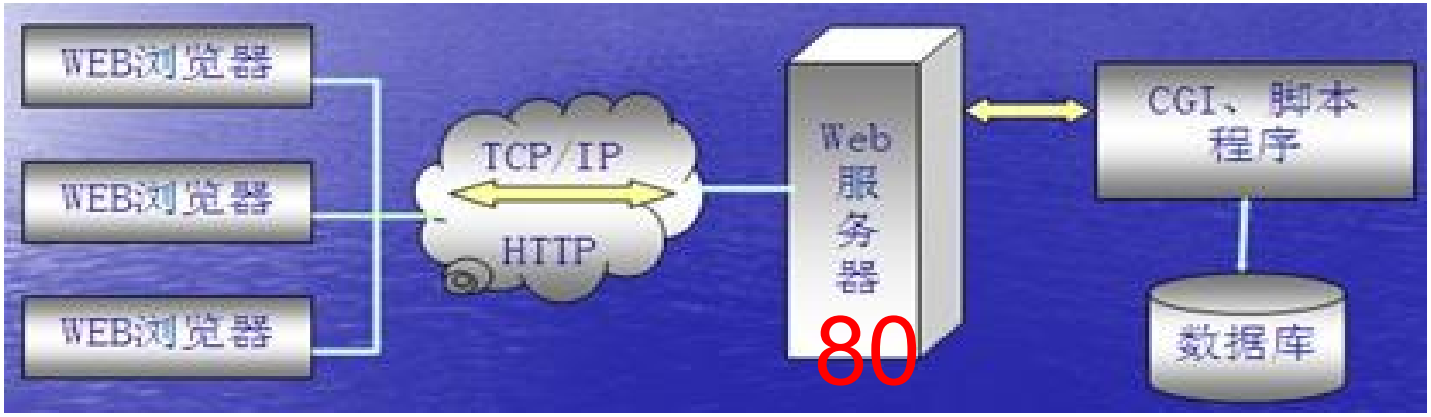
三、超文本传输协议 HTTP

- 万维网中客户与服务器之间采用 HTTP 协议。
  - 客户端将用户的请求、数据发送给服务器；
  - 服务器将文档、或处理结果发送给客户端。

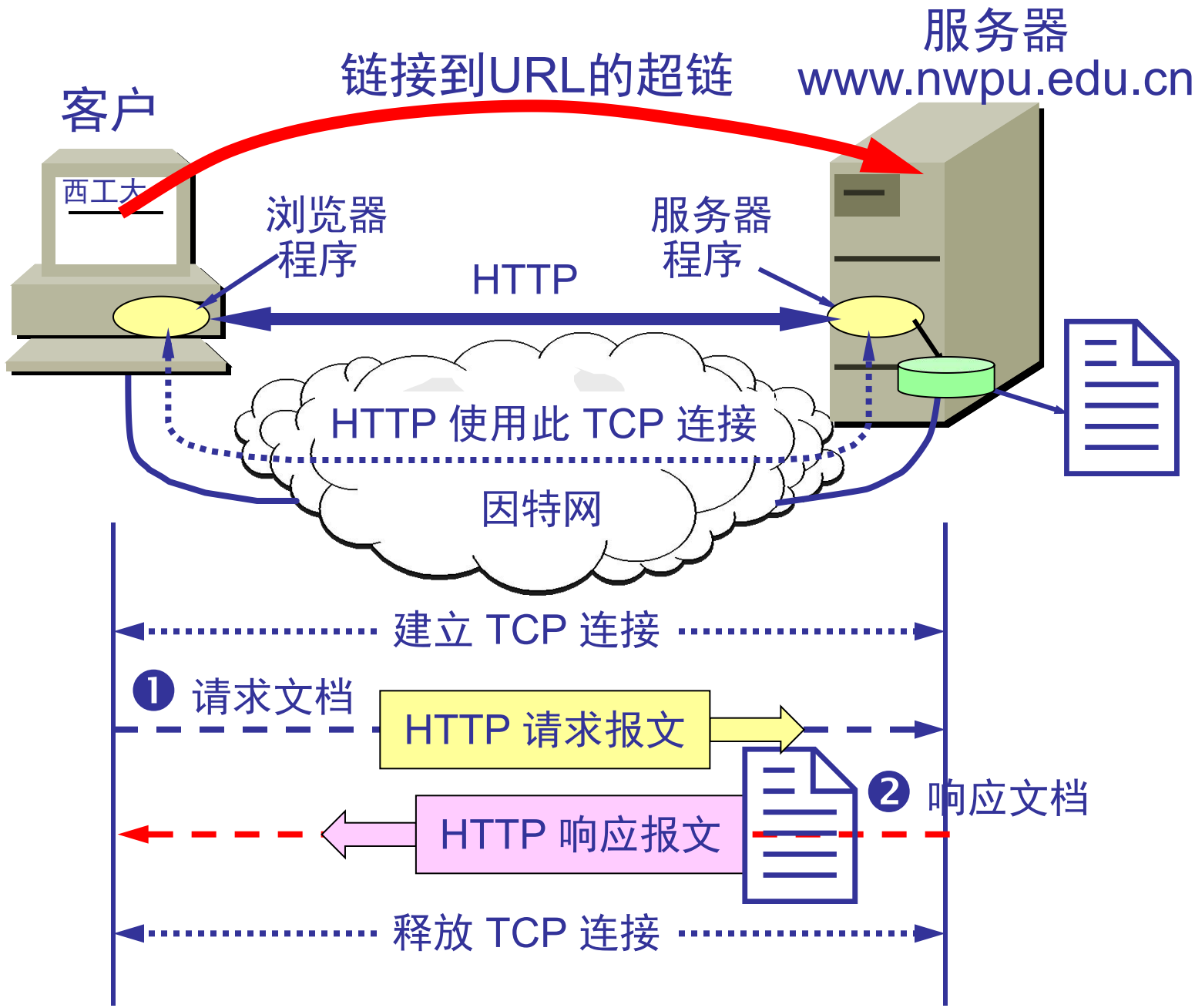


三、超文本传输协议 HTTP

- HTTP 是面向事务的(transaction-oriented)应用层协议，它是万维网上能够可靠地传输文档文件（包括文本、声音、图像等各种媒体内容）重要基础。
  - 事务：一系列不可分割的信息交换；要么所有的信息交换均完成，要么一次交换也不进行。
- 思考：浏览地址栏输入WWW.NPWU.EDU.CN后发生了什么事情？



1. 万维网的工作过程


服务器
www.nwpu.edu.cn
链接到URL的超链
客户
西工大
浏览器
程序
服务器
程序
HTTP
HTTP 使用此 TCP 连接
因特网
建立 TCP 连接
❶ 请求文档
HTTP 请求报文
❷ 响应文档
HTTP 响应报文
释放 TCP 连接

用户鼠标超链发生的事件

(1) 浏览器分析超链指向页面的 URL。

(2) 浏览器向 DNS 请求解析 www.nwpu.edu.cn 的 IP 地址。

(3) 域名系统 DNS 解析出西工大Web服务器的 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：

GET /index.htm。

(6) 服务器发送HTTP响应，把文件 index.htm 网页文件发给浏览器。

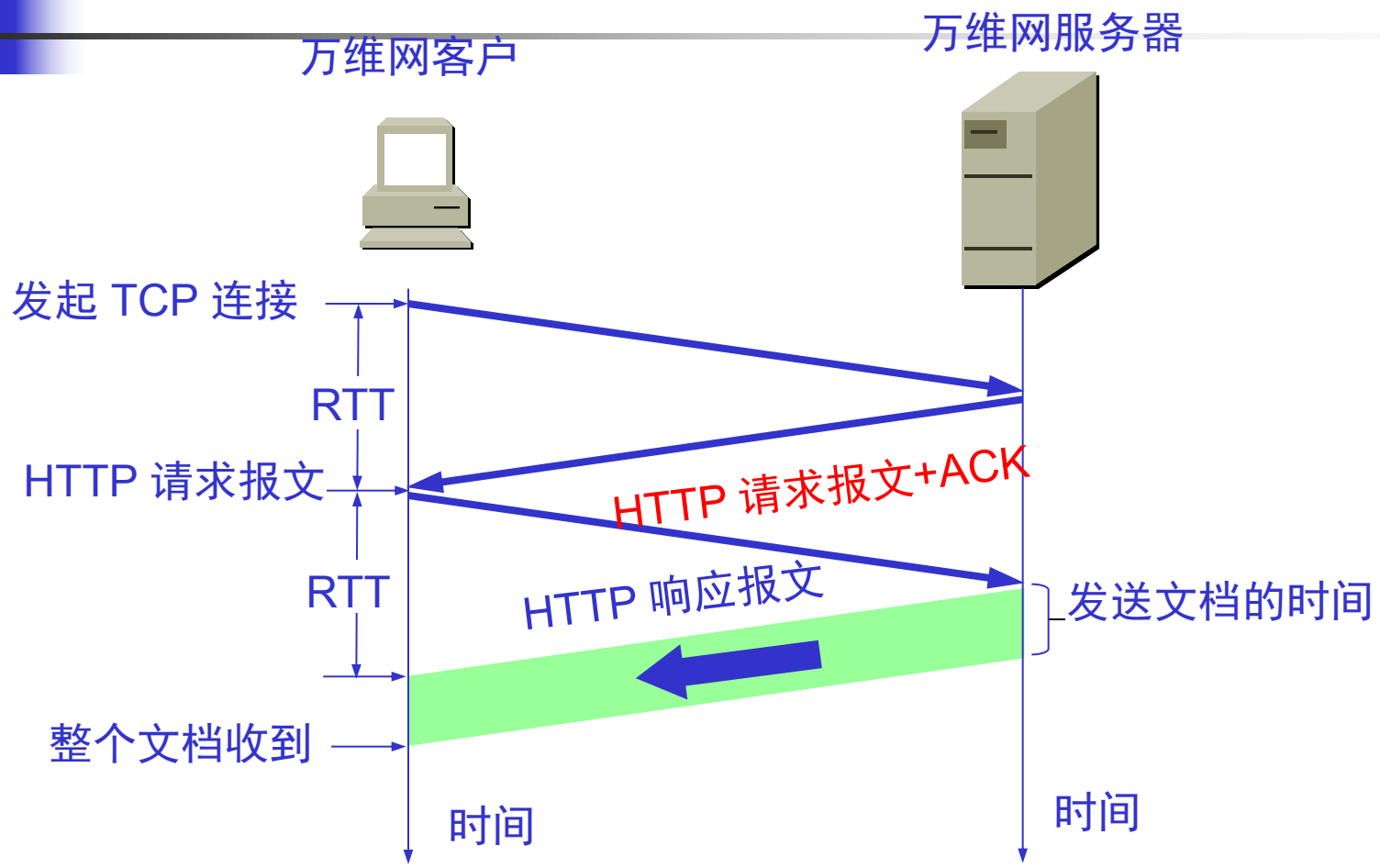
(7) TCP 连接释放。

(8) 浏览器显示“西工大”首页文件 index.htm 中的所有信息。

HTTP/1.0 的主要特点

- HTTP 是面向事务客户/服务器协议。
- HTTP 1.0 协议是无状态的(stateless)。
  - 同一用户第二次访问同一页面时,服务器第二次应答和第一次应答一样;
  - 服务器不记录谁曾访问过,访问过几次，访问哪些;
  - 简化了服务器设计,服务器容易支持大量并发HTTP请求.
- HTTP 协议本身无连接的，在传输层选用了面向连接的 TCP ，保证数据传输的可靠性。

计算访问一个万维网文档所需时间



用户获得一个文挡的时间：2RTT+文档发送时间(第三次握手应答信号同时捎带客户对文档请求)

HTTP/1.0 的主要特点

- 采用一次一连接
  - 当一个万维网文挡中有9个图片链接对象时，共需10个TCP连接
    - 下载该文挡的文本信息需要1个TCP连接；
    - 下载9个图片需要9个TCP连接;
  - 好在浏览器一般允许同时建立5~10个TCP连接,实现并行传输,缩短了响应时间；
  - 客户端和服务器消耗了大量网络资源(TCB)；
- HTTP/1.1很好的解决了资源消耗问题:采用持续连接方式.
  - 1997年以前一般采用HTTP/1.0([RFC 1945]);1998年升级为HTTP/1.1([RFC 2616]).

持续连接（HTTP/1.1）(persistent connection)

- 客户端与服务器建立好一个TCP连接后,客户端可在该条连接上发送多个HTTP请求报文,服务器也可以发送多个HTTP响应报文,在此会话期间,TCP连接一直保持.
  - 实现在一个TCP连接上传输多个资源对象（文本、图像、视频等），每个资源对象必须在同一个服务器上。
- 目前一些流行的浏览器（例如，IE 6.0）的默认设置就是使用 HTTP/1.1。

浏览器-工具-Internet选项-高级

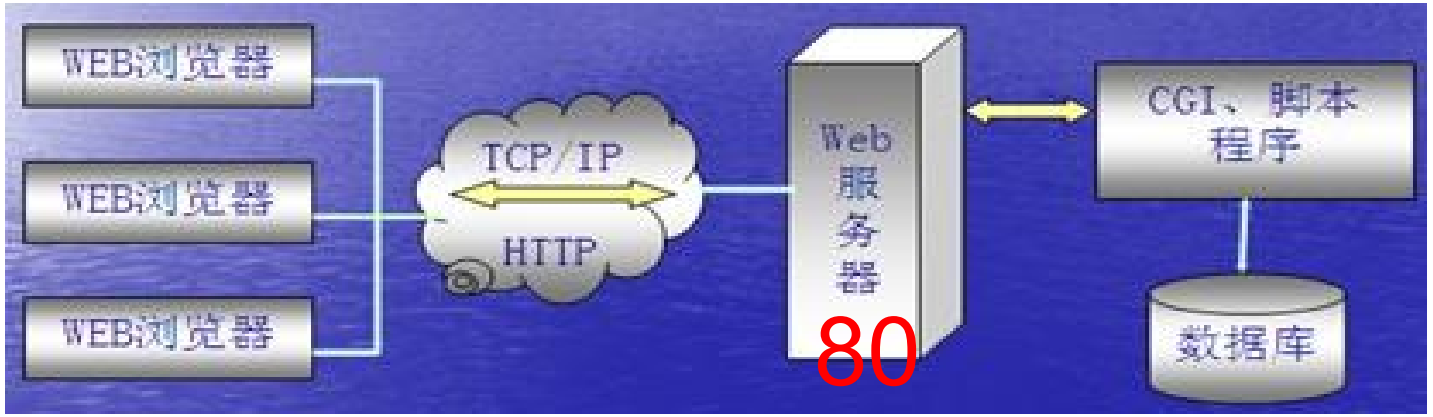
如果不采用该选项,则使用HTTP 1.0

持续连接两种工作方式

- 非流水线方式（类似停止-等待）
  - 在一个TCP连接上，客户在收到前一个响应后才能发出下一个请求。这比非持续连接的两倍 RTT 的开销节省了建立 TCP 连接所需时间。但服务器在发送完一个对象后，其 TCP 连接就处于空闲状态，浪费了服务器资源。
- 流水线方式（类似连续ARQ）
  - 在一个TCP连接上，客户在收到 HTTP 的响应报文之前就能够接着发送新的请求报文；一个接一个的请求报文到达服务器后，服务器就可连续发回响应报文；使用流水线方式时，客户访问所有的对象只需花费一个 RTT时间，使 TCP 连接中的空闲时间减少，提高了下载文档效率。

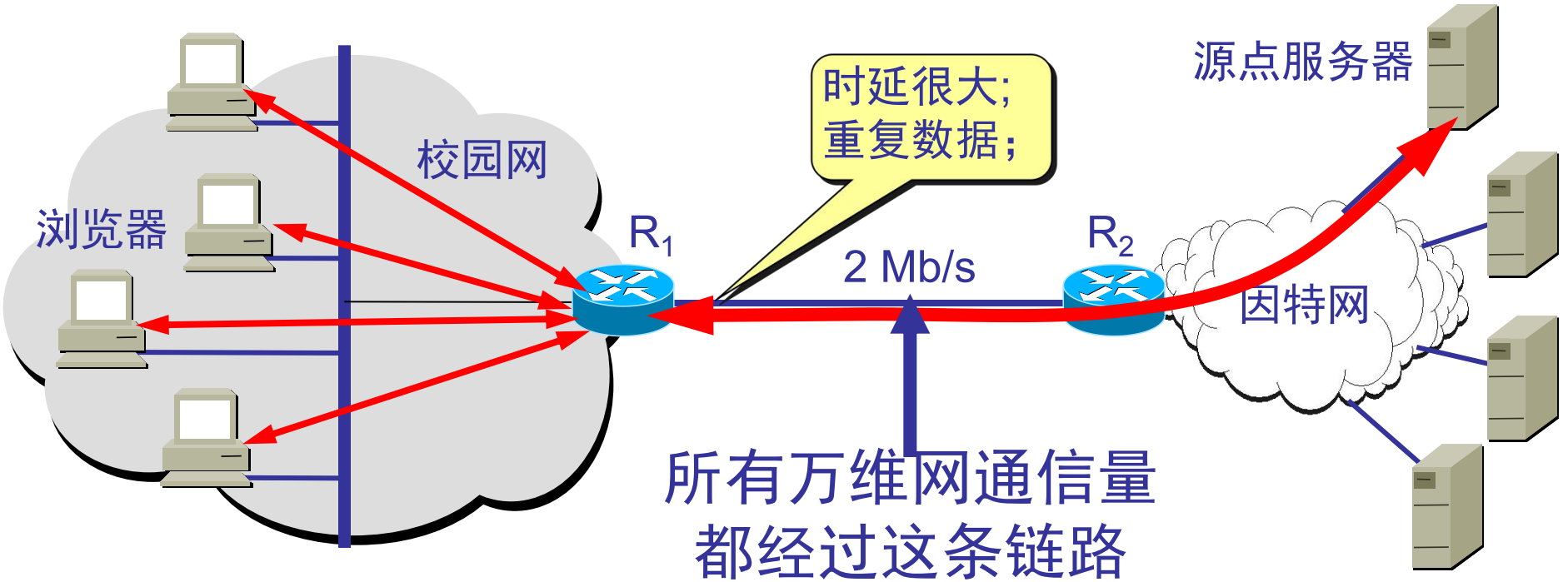
2. 代理服务器 (proxy server)

- 代理服务器(proxy server)又称为万维网高速缓存(Web cache) 。
- 代理服务器把最近的一些HTTP请求和响应报文暂存在本地(代理服务器)磁盘缓存中。
- 当新请求到达时，代理服务器如果发现该请求与暂存的某一请求相同,则返回暂存的相应响应, 不需要按 URL 的地址再去因特网访问该资源。
- 代理服务器可部署在客户端、本地网络系统中。



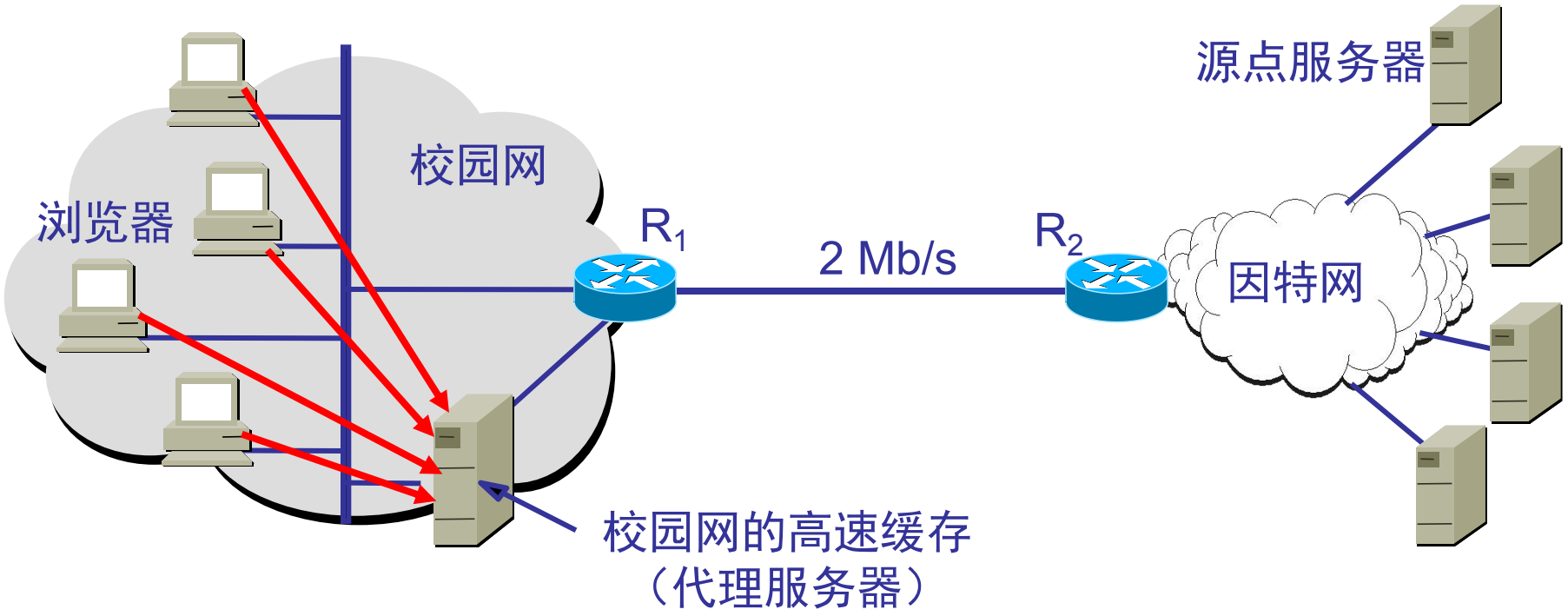
使用高速缓存可减少访问因特网服务器的时延

没有使用高速缓存的情况



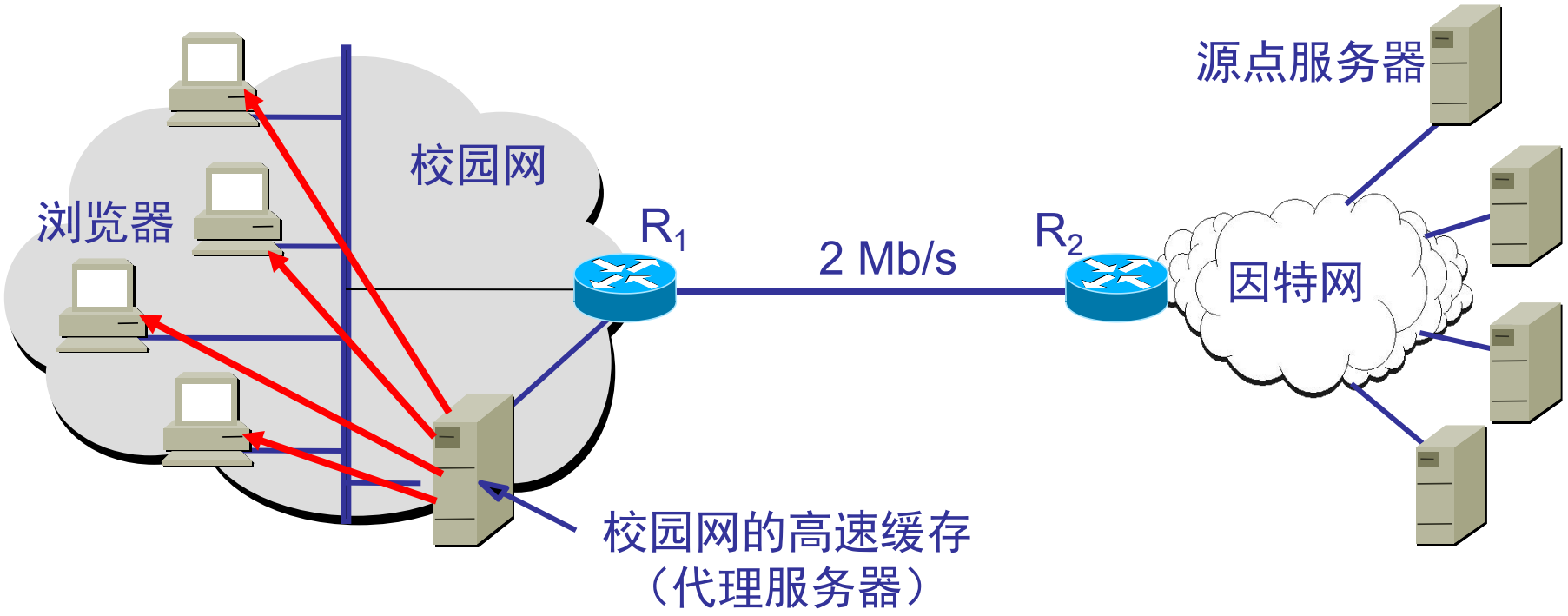
使用高速缓存的情况

(1) 浏览器访问因特网的服务器时，要先与校园网的高速缓存建立 TCP 连接，并向高速缓存发出 HTTP 请求报文



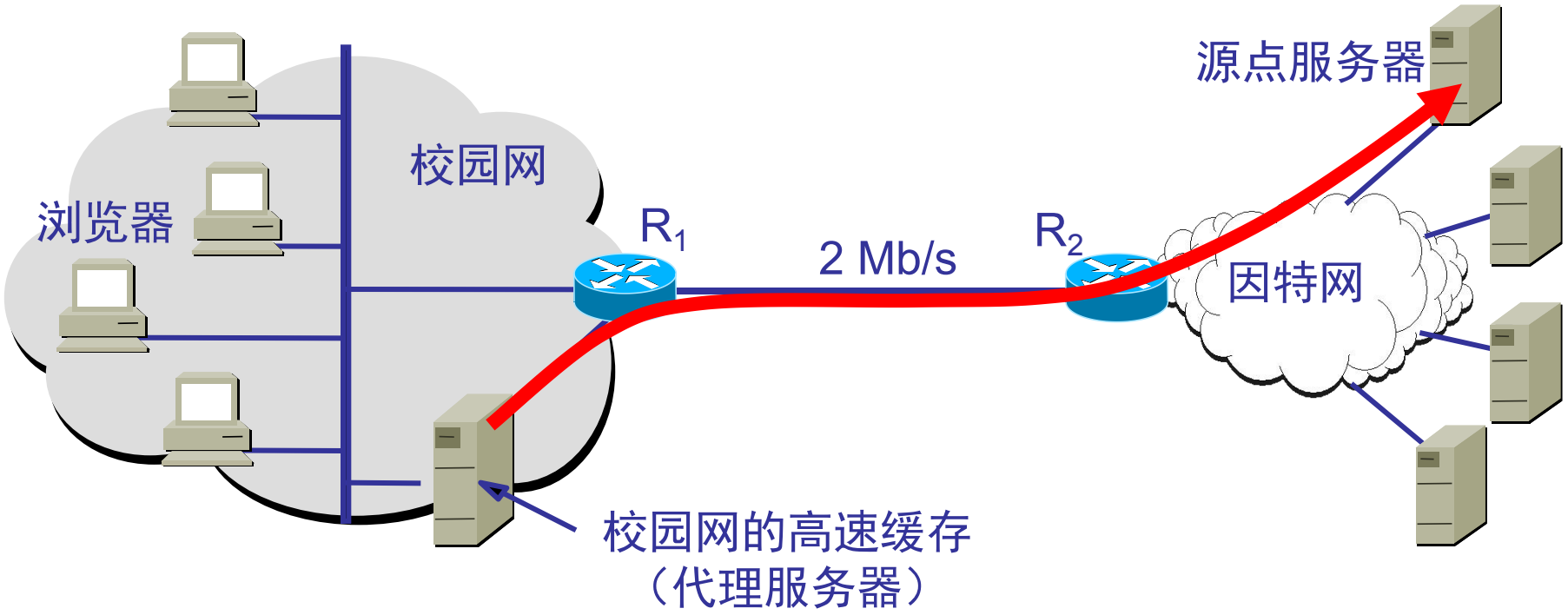
使用高速缓存的情况

(2) 若高速缓存已经存放了所请求的对象，则将此对象放入 HTTP 响应报文中返回给浏览器。



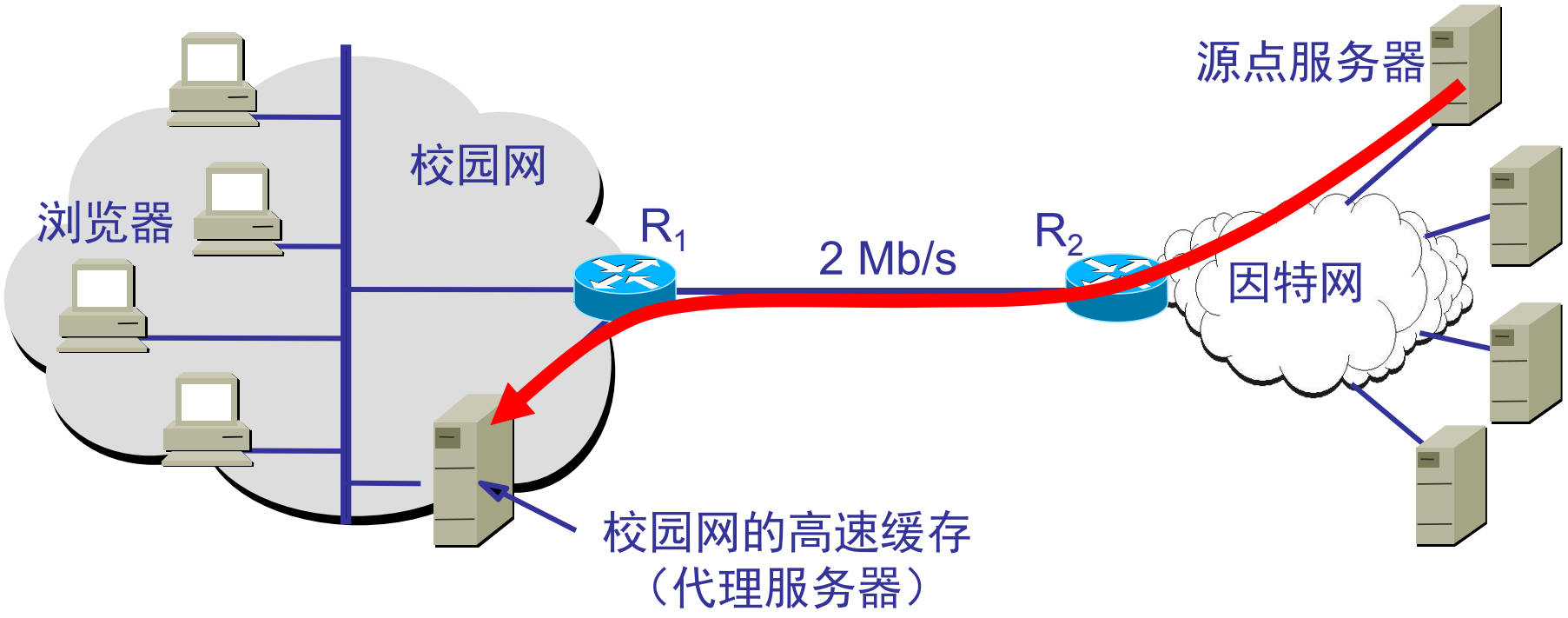
使用高速缓存的情况

(3) 否则，高速缓存就代表发出请求的用户浏览器，与因特网上的源点服务器建立 TCP 连接，并发送 HTTP 请求报文。



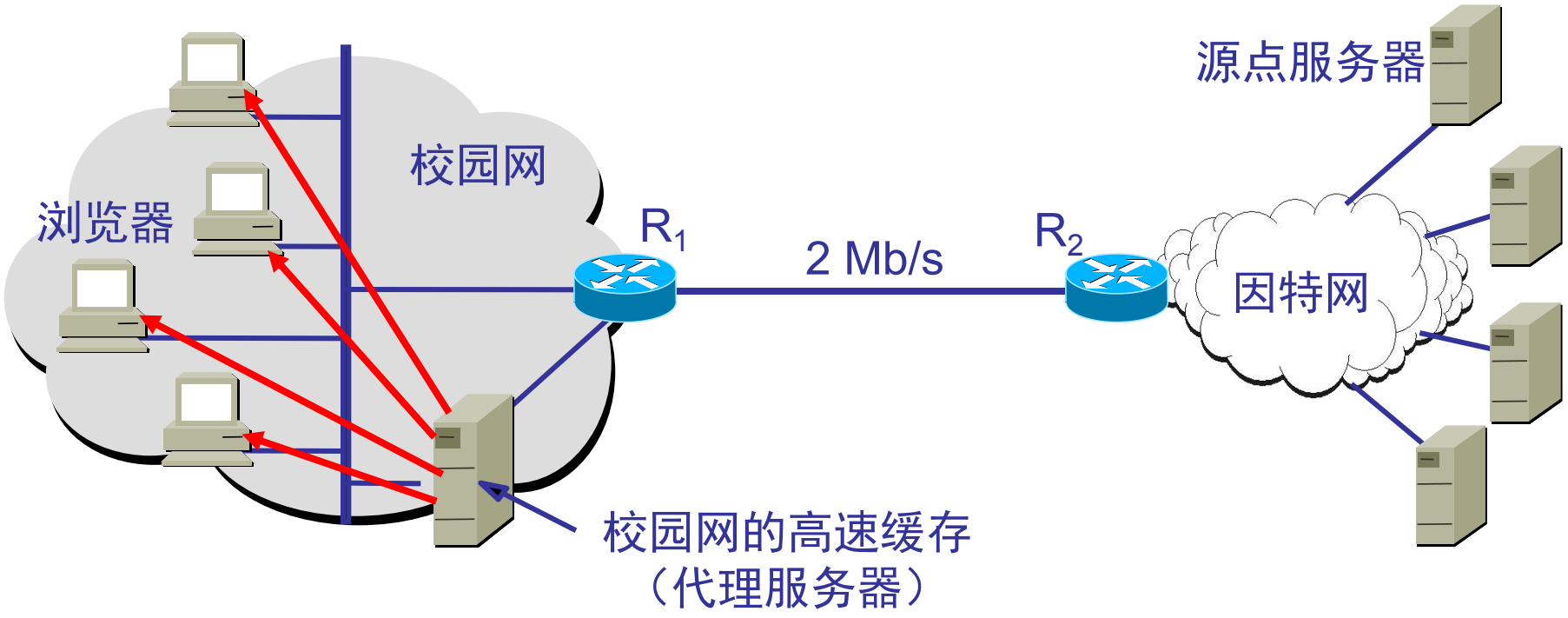
使用高速缓存的情况

(4) 源点服务器将所请求的对象放在 HTTP 响应报文中返回给校园网的高速缓存。

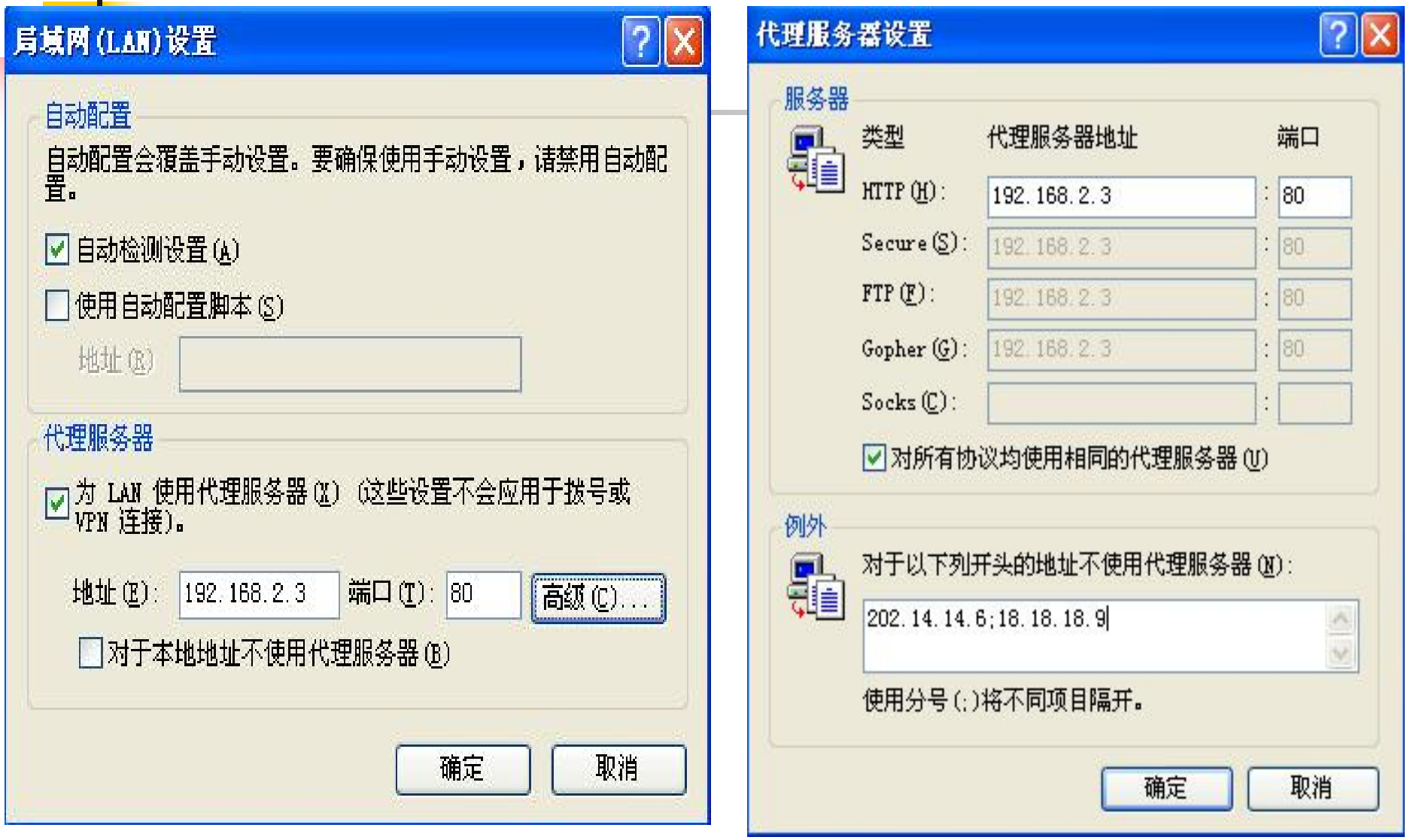


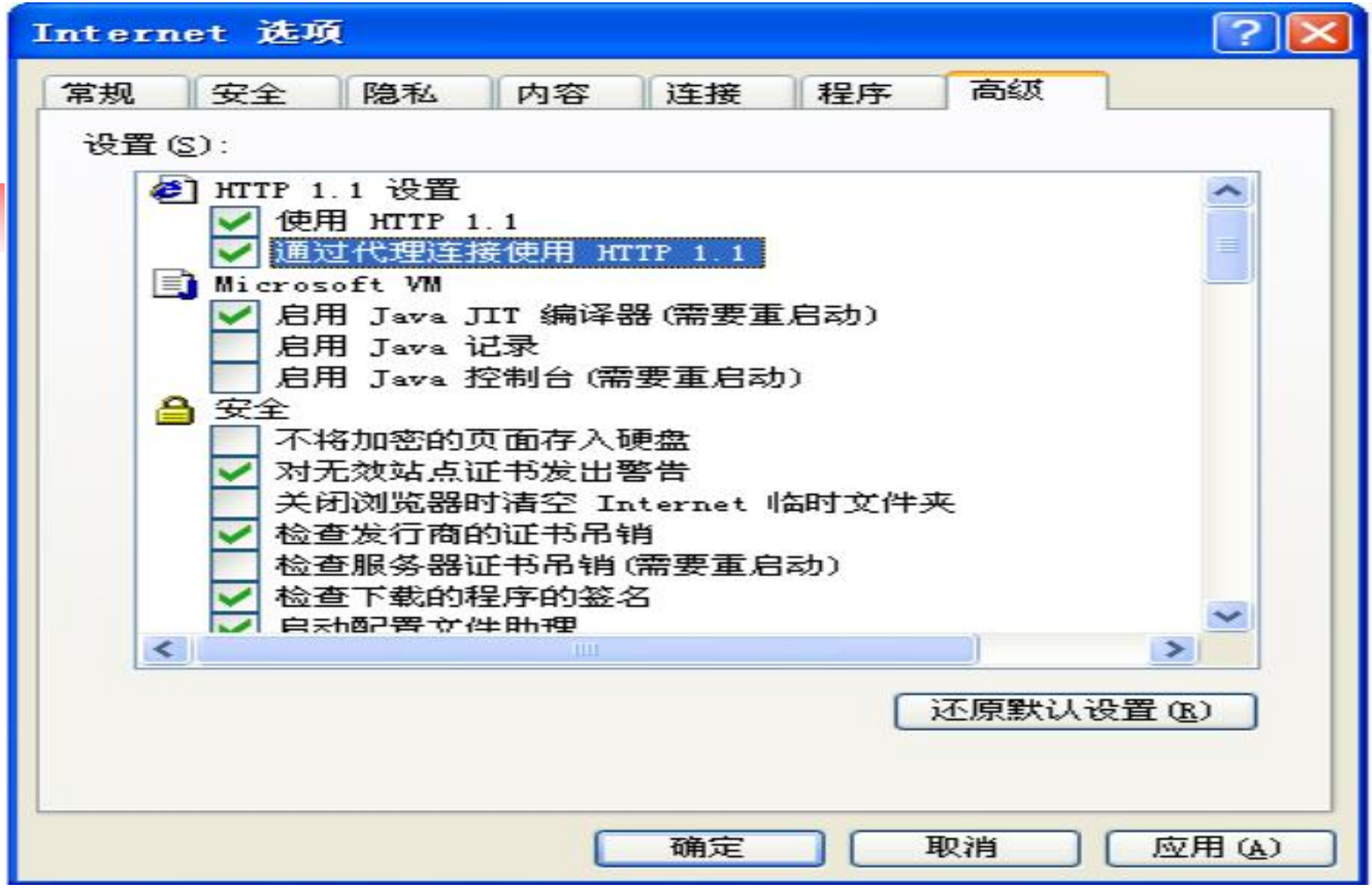
使用高速缓存的情况

(5) 高速缓存收到此对象后，先复制在其本地存储器中（为今后使用），然后再将该对象放在 HTTP 响应报文中，通过已建立的 TCP 连接，返回给请求该对象的浏览器。



浏览器-工具-internet选项-连接-LAN设置(1)-高级





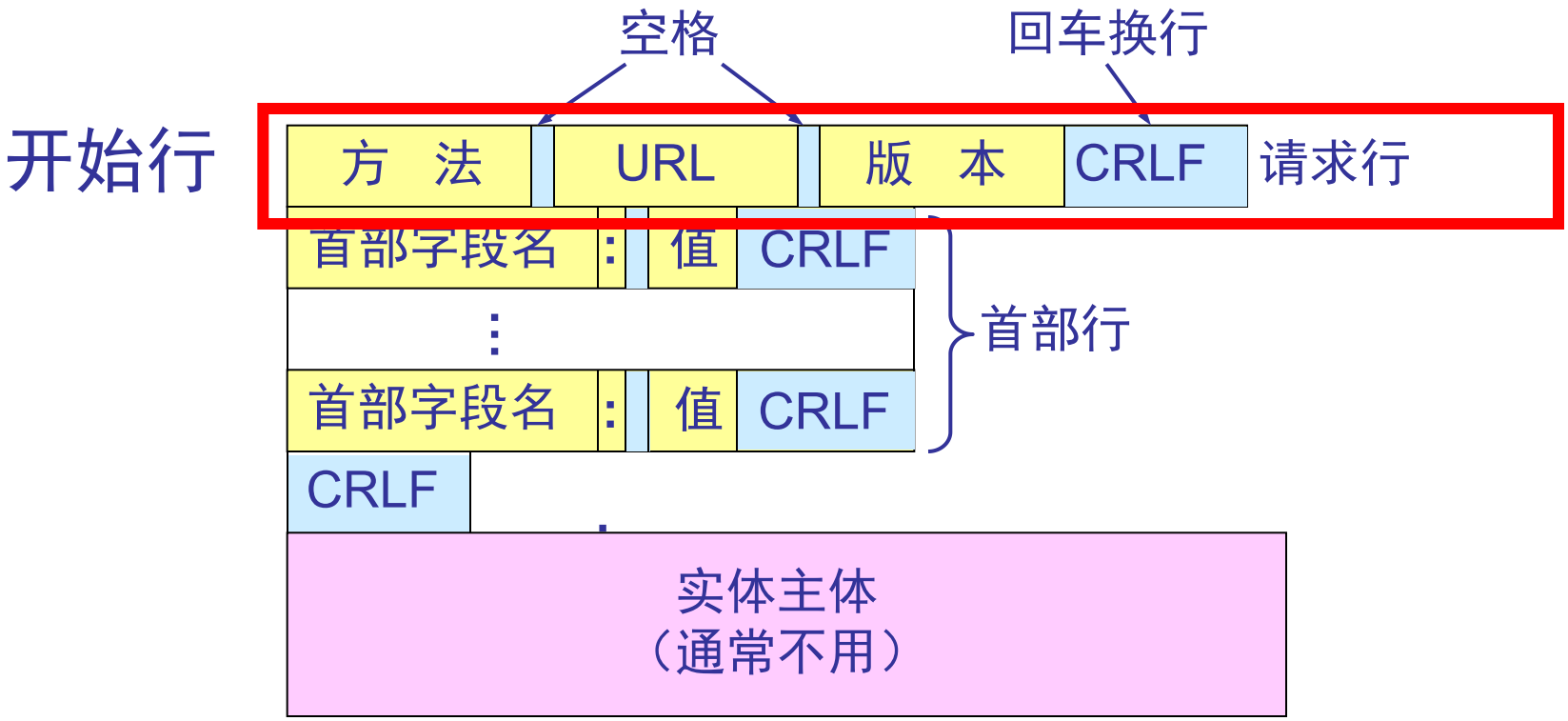
浏览器-工具-internet选项-高级

3. HTTP 的报文语法

HTTP 有两类报文（语义）

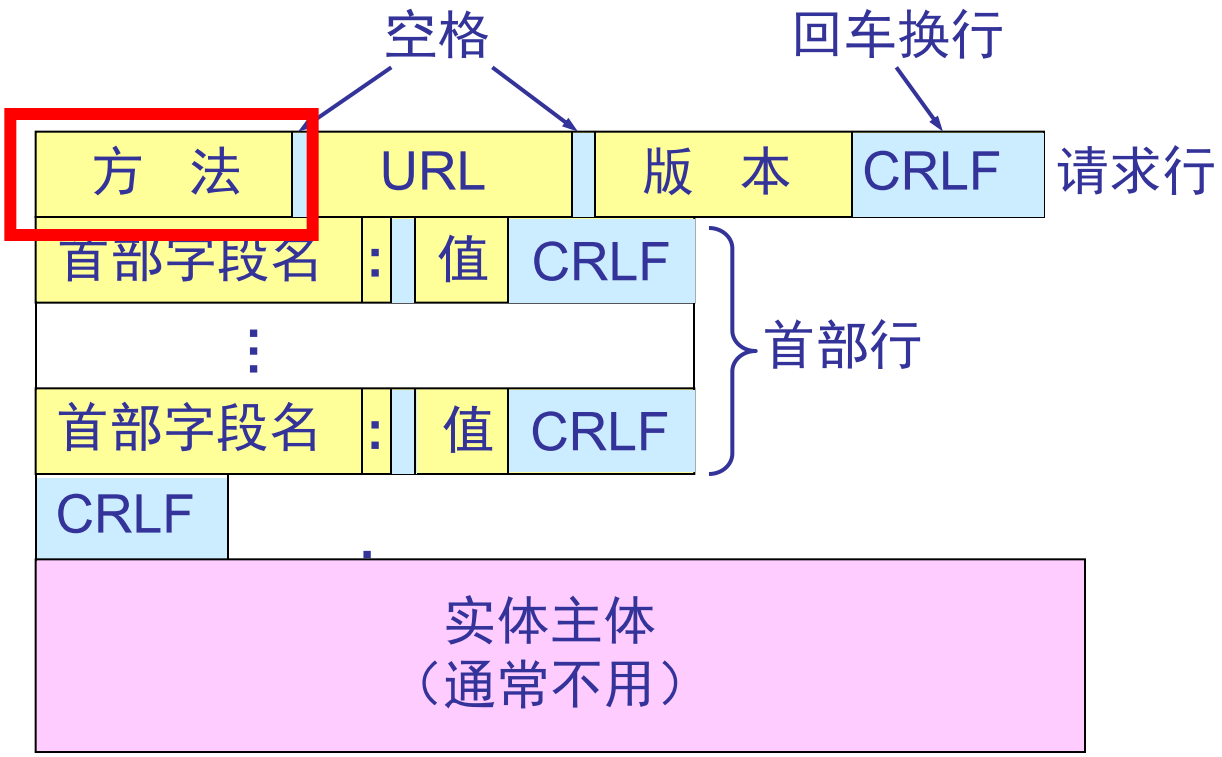
- 请求报文——从客户向服务器发送请求报文。
- 响应报文——从服务器到客户的应答。
- HTTP报文由三个部分组成：
  - 请求报文：开始行、首部行和实体主体
  - 相应报文：状态行、首部行和实体主体

HTTP 的报文语法（请求报文）



在请求报文中，开始行就是请求行，且实体主体不用。

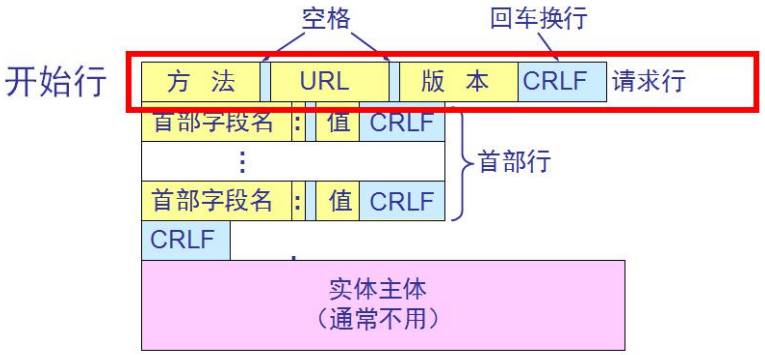
HTTP 的报文语法（请求报文）



“方法”是面向对象技术中使用的专门名词。所谓“方法”就是对所请求的对象进行的操作，因此这些方法实际上也就是一些命令。因此，请求报文的类型是由它所采用的方法决定的。

HTTP 请求报文的一些方法

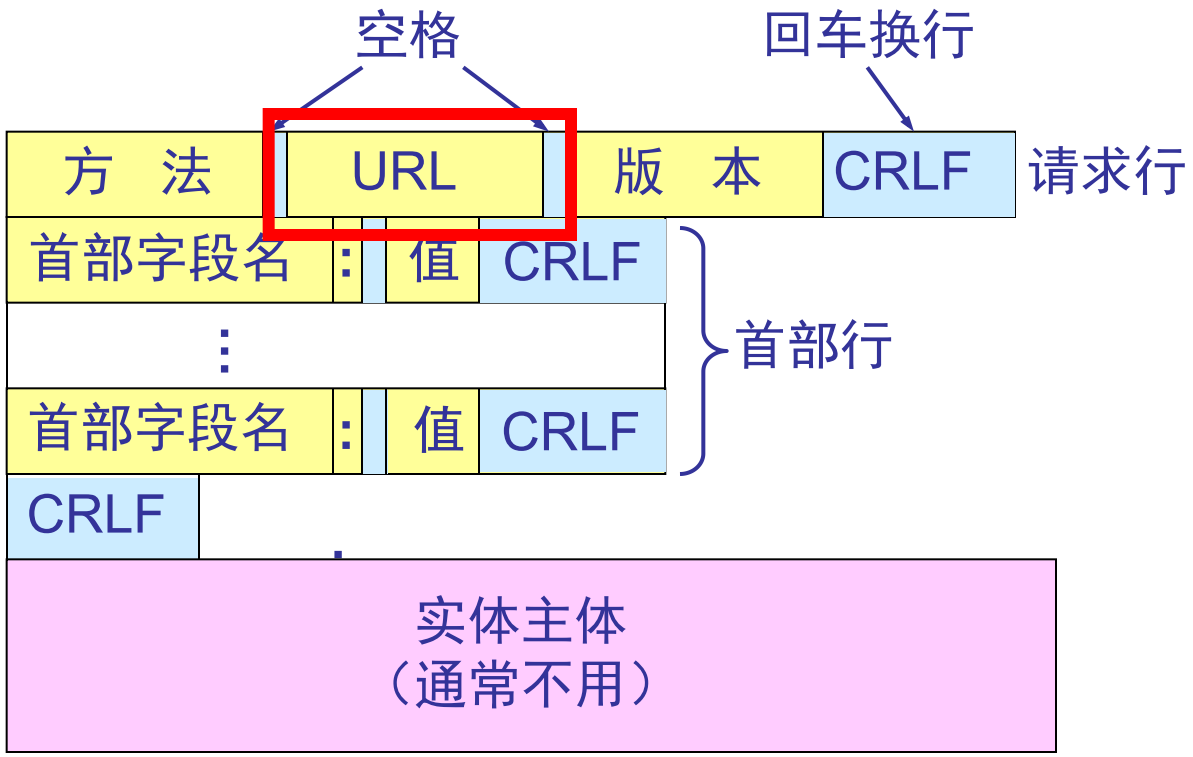
方法（操作）	意义
OPTION	请求一些选项的信息
GET	请求读取由 URL所标志的信息
HEAD	请求读取由 URL所标志的信息的首部
POST	给服务器添加信息（例如，注释）
PUT	在指明的 URL下存储一个文档
DELETE	删除指明的 URL所标志的资源
TRACE	用来进行环回测试的请求报文
CONNECT	用于代理服务器



GET/POST方法

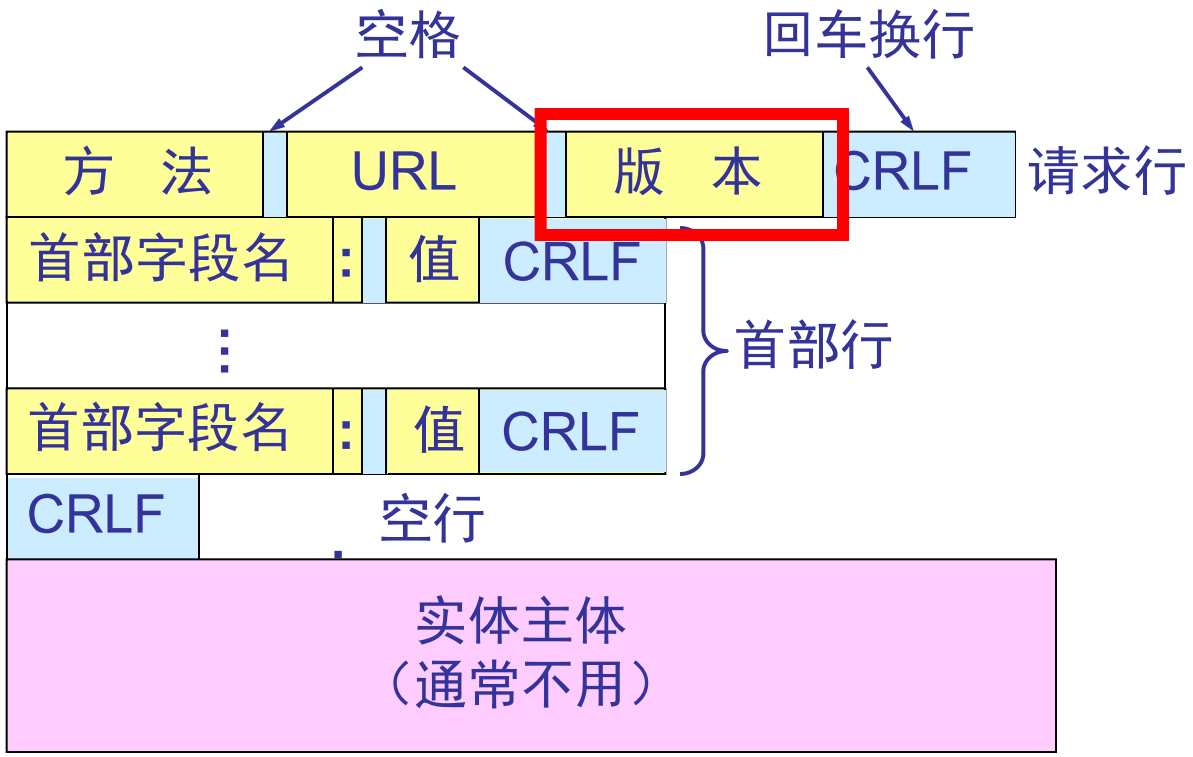
- GET方法：取回由URL指定的资源对象，主要用于取回由一个超文本链所定义的对象。
  - 如果对象是文件，则GET取回的是文件内容；
  - 如果对象是程序，则GET取回的是该程序执行的结果；
  - 如是对象是数据库查询，则GET取回的是本次查询的结果。
- POST方法：当客户向服务器传送大量的数据，并要求服务器和公共网关接口CGI(Common GatewayInterface)程序作进一步处理时要使用POST方法。

HTTP 的报文结构（请求报文）



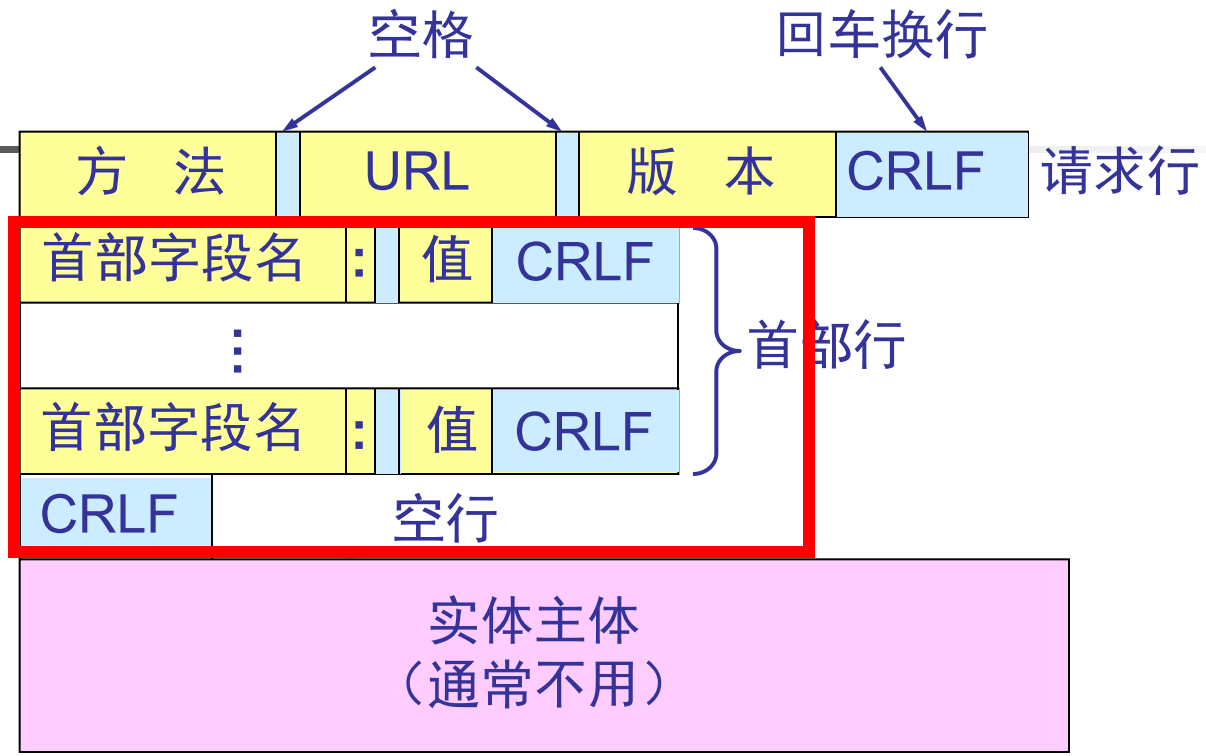
“URL”是所请求的资源的 URL。

HTTP 的报文结构（请求报文）

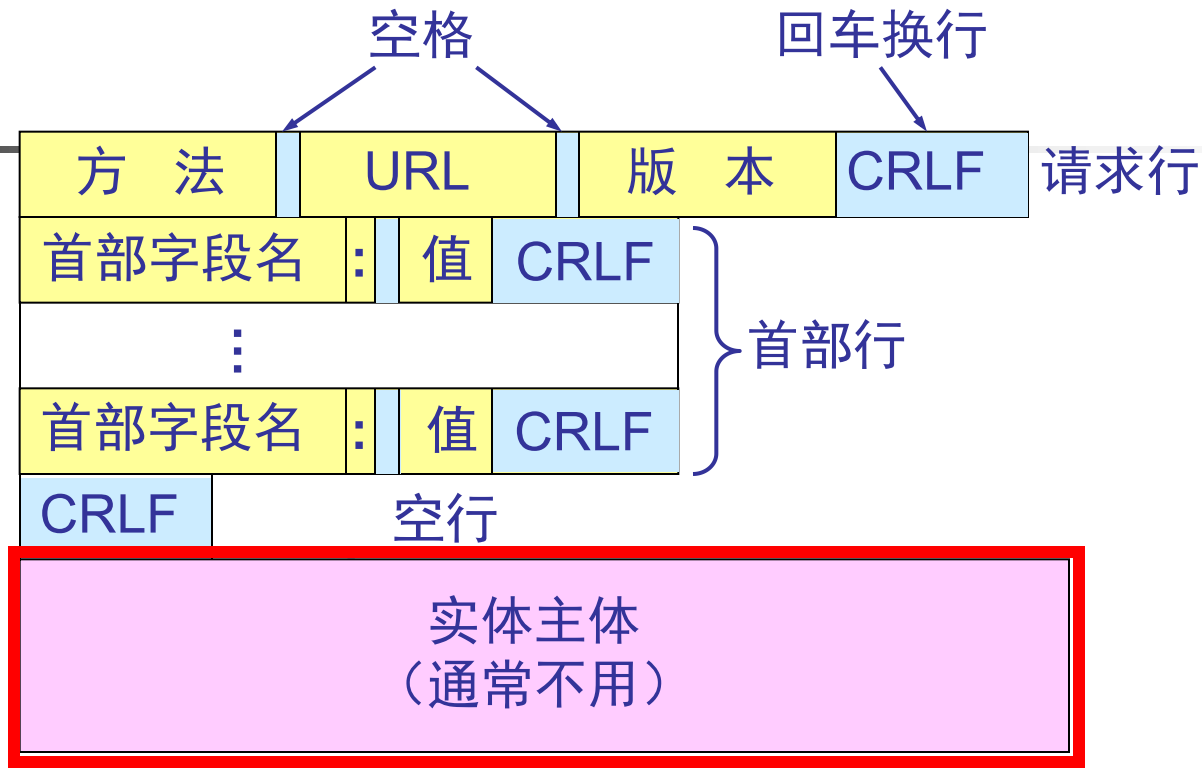


“版本”是 HTTP 的版本。

HTTP 的报文结构（请求报文）



HTTP 的报文结构（请求报文）



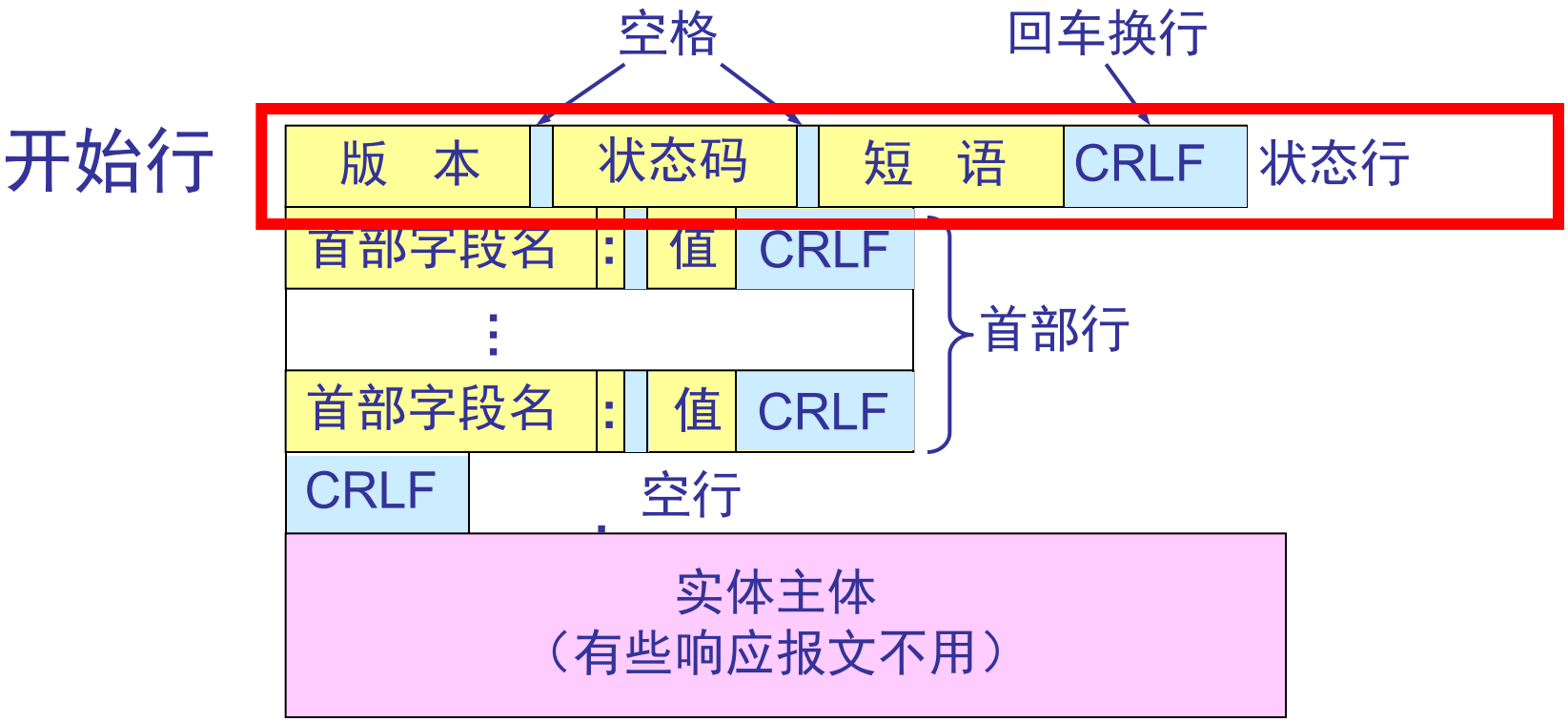
www.nwpu.edu.cn

```
GET /index.htm HTTP/1.0
Host:www.nwpu.edu.cn
Connection:close                          //采用一次一连接
User-agent: Mozilla/4.0                   //用户浏览器采用Netscape
Accept: text/html, image/gif,image/jpeg   //可接受数据类型
Accept-language:cn                        //用户希望优先得到中文版
                                          //空行
```

(extra carriage return, line feed)

```
HTTP/1.0 200 OK
Date: Thu, 06 Aug 1998 12:00:15 GMT     //发送时间
Server: Apache/1.3.0 (Unix)             //服务器类型
Last-Modified: Mon, 22 Jun 1998 ...     //文档最后修改时间
Content-Length: 6821                    //文档长度
Content-Type: text/html                 //文挡类型
                                        //空行
data data data data data ...            //文档内容（数据）
```

HTTP 的报文语法（应答报文）



应答报文：开始行、首部行和实体主体；
应答报文的开始行是状态行。
状态行包括三项内容，即 HTTP 的版本，状态码，以及解释状态码的简单短语。

状态码都是三位数字

- 1xx：表示通知信息，如请求收到了或正在进行处理。
- 2xx：表示成功，如接受或执行成功。
- 3xx：表示重定向，表示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。

www.nwpu.edu.cn

```
GET /index.htm HTTP/1.0
Host:www.nwpu.edu.cn
Connection:close                          //采用一次一连接
User-agent: Mozilla/4.0                   //用户浏览器采用Netscape
Accept: text/html, image/gif,image/jpeg   //可接受数据类型
Accept-language:cn                        //用户希望优先得到中文版
                                          //空行
```

(extra carriage return, line feed)

```
HTTP/1.0 200 OK
Date: Thu, 06 Aug 2020 12:00:15 GMT     //发送时间
Server: Apache/1.3.0 (Unix)             //服务器类型
Last-Modified: Mon, 22 Jun 2020 ...     //文档最后修改时间
Content-Length: 6821                    //文档长度
Content-Type: text/html, image/gif,     //文挡类型
                                        //空行
data data data data data ...            //文档内容（数据）
```

4. 服务器上存放用户信息

- HTTP协议无状态，服务器无法记录用户访问服务器的行为：谁在什么时间、访问了哪些网页，作了哪些操作。
- 目前一些应用需要对用户访问服务器的行为进行跟踪，如在“亚麻逊”网站购买书，服务器需要记录用户身份，将该用户选择书放入一个“购物车”（服务器需要对用户选书行为进行跟踪），然后集中结帐。
- 解决方法：万维网站点使用 Cookie[RFC2109] 来跟踪用户行为，并在万维网服务器保存，在服务器和客户之间传递用户行为状态信息。
- 现在许多网站都支持Cookie。

Cookie工作原理

- （1）当用户张三浏览某个支持Cookie 的网站时，该服务器为张三产生一个唯一识别码Cookie id:123 - 保护隐私；
- （2）服务器利用此识别码，在其后台数据库中为该用户创建一个表，记录拥有此识别码用户访问网站的行为，达到跟踪该用户在该网站的活动。
- （3）服务器给张三的响应报文中增加一个首部行：

```
HTTP/1.0 200 OK
Date: Thu, 06 Aug 2013 12:00:15 GMT      //发送时间
Server: Apache/1.3.0 (Unix)              //服务器类型
Last-Modified: Mon, 22 Jun 2012 ...      //文档最后修改时间
Content-Length: 6821                     //文档长度
Content-Type: text/html                  //文挡类型
Set-cookie:123                           //识别码

data data data data data ...             //文档内容（数据）
```

Cookie工作原理

- 张三客户端收到该响应，浏览器在特定Cookie文件（c:\document and settings\zhangsan\Cookie\文件名称）增加一行：服务器域名，识别码（ www.amazon.cn, 123）.
- 当张三在同一台机器继续网络该网站时，每发送一个请求报文，浏览器从该特定Cookie文件读取识别码，增加在Cookie首部

```
GET /index.htm HTTP/1.0
Host:http://www.amazon.cn
Connection:close                          //采用一次一连接
User-agent: Mozilla/4.0                   //用户浏览器采用Netscape
Cookie: 123                               //识别码
Accept: text/html, image/gif,image/jpeg //可接受数据类型
Accept-language:cn                        //用户希望优先得到中文版
                                          //空行
(extra carriage return, line feed)
```

Cookie工作原理

- 服务器收到该请求报文后，利用识别码，跟踪张三访问行为，并记录到数据库表中：访问哪些网页，访问顺序等。
- 注意：服务器管理程序并不知道该用户是张三，也不知道其他信息，只知道识别码：123。
- 服务器应用程序可利用识别码在后台数据库中为张三（张三与识别码之间建立映射关系）维护一个购物清单表，实现购物车的功能，方便张三集中付帐。
- 如果张三几天后使用同一计算机再次访问同一购物网站，HTTP请求有**Cookie：123**，服务器可利用识别码以及历史访问记录，为张三推荐产品；

Cookie工作原理

- 如果张三以前在该网站使用过信用卡付费，张三的注册信息（姓名+密码，电话，E-MAIL地址、信用卡号码等）已经被服务器记录并和张三的识别码绑定；
- 这时，张三如果还利用同一计算机访问该购物网站，由于请求中识别码（**Cookie：123**），服务器可验证该用户是张三（**Cookie id +IP地址**），用户登录系统或者付费时不需要用键盘输入用户名，信用卡号，有时甚至密码也不输入，服务器从后台数据库中利用识别码和注册信息已经替用户写上了
- **Cookie争议**
  - 用户帐户不安全-利用**Cookie** 实现了身份验证；
  - 个人隐私泄露：对用户访问网站行为进行跟踪；

Cookie设置

- 浏览器-工具- Internet选项-隐私
- 滑动标尺最高：阻止所有Cookie；最低：接受所有Cookie；中间：有条件接受Cookie。

本节内容提要

万维网 WWW

- 一、 万维网概述
- 二、 统一资源定位符 URL
- 三、 超文本传送协议 HTTP
- 四、 万维网的文档（HTML）

四、万维网的文档（超文本标记语言 HTML）

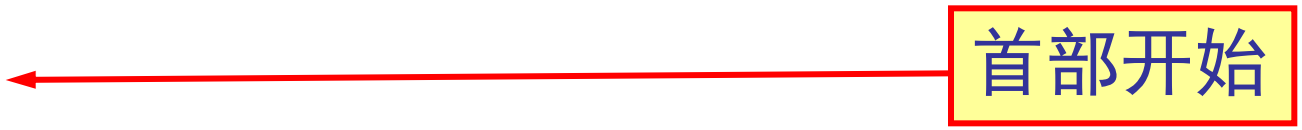
- 超文本标记语言 HTML（HyperText Markup Language） 中的 Markup 的意思就是“设置标记” - 标签。
- HTML 定义了许多用于排版的命令（即标签）。
- HTML 把各种标签写入到万维网页面，构成了HTML 文档。
- HTML 文档可以用任何文本编辑器创建为ASCII 码文件保存。
- 仅当 HTML 文档是以.html 或 .htm 为后缀时，浏览器才对此文档的各种标签进行解释，并在屏幕上显示网页信息。
- 若 HTML 文档改换以 .txt 为其后缀，则浏览器解释程序就不对标签进行解释，在浏览器只能看到原来的文本文件。

HTML 文档中标签的用法


```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```



HTML 文档中标签的用法


```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

首部结束

HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

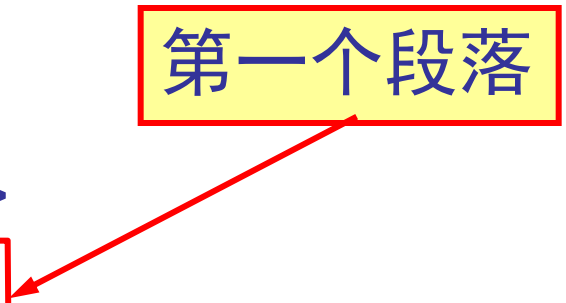
主体开始

HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

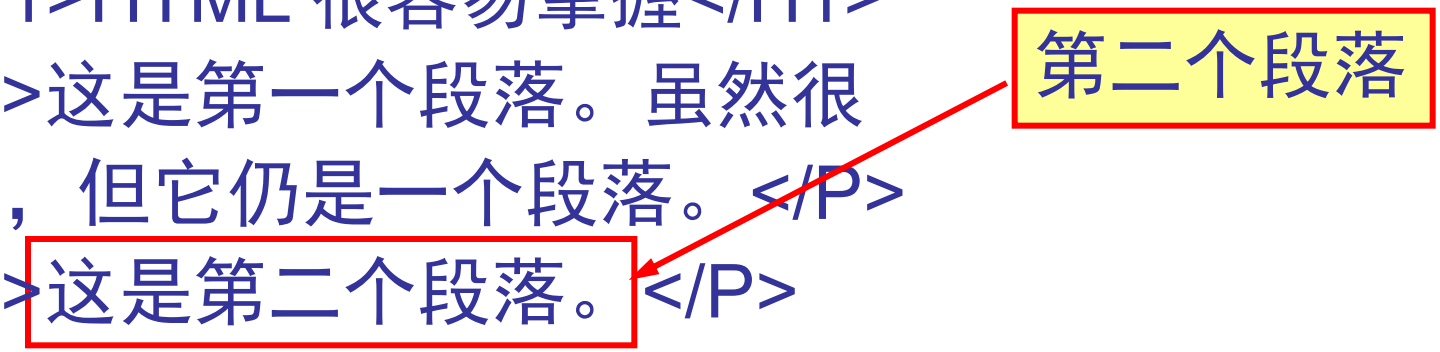
HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```



HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```



HTML 文档中标签的用法

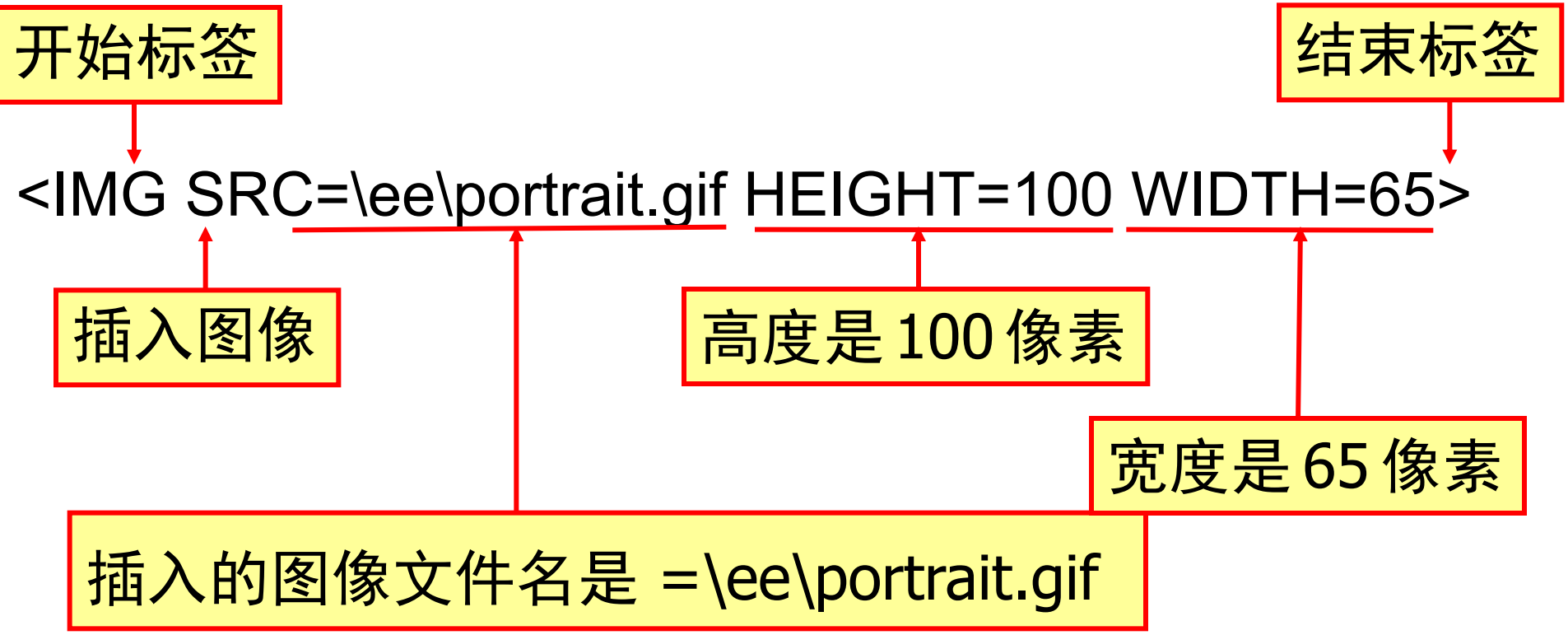
```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

主体结束

HTML 文档中标签的用法

```
<HTML>
<HEAD>
    <TITLE>一个 HTML 的例子</TITLE>
</HEAD>
<BODY>
    <H1>HTML 很容易掌握</H1>
    <P>这是第一个段落。虽然很
     短，但它仍是一个段落。</P>
    <P>这是第二个段落。</P>
</BODY>
</HTML>
```

插入图片


开始标签
结束标签
<IMG SRC=\ee\portrait.gif HEIGHT=100 WIDTH=65>
插入图像
高度是 100 像素
宽度是 65 像素
插入的图像文件名是 =\ee\portrait.gif

万维网页面中的超链
（链接到其他网点上的页面）

- 定义一个超链的标签是<A>。字符A表示锚(Anchor)。
- 在HTML文档中定义一个超链的语法是：

```
<A HREF="...  "> X </A>
```

超链的起点

这个地方填写超链终点的 URL

超链(链接)举例

<A HREF="http://www.nwpu.edu.cn">西工大</A>

超链的终点是西工大的主页

超链的起点

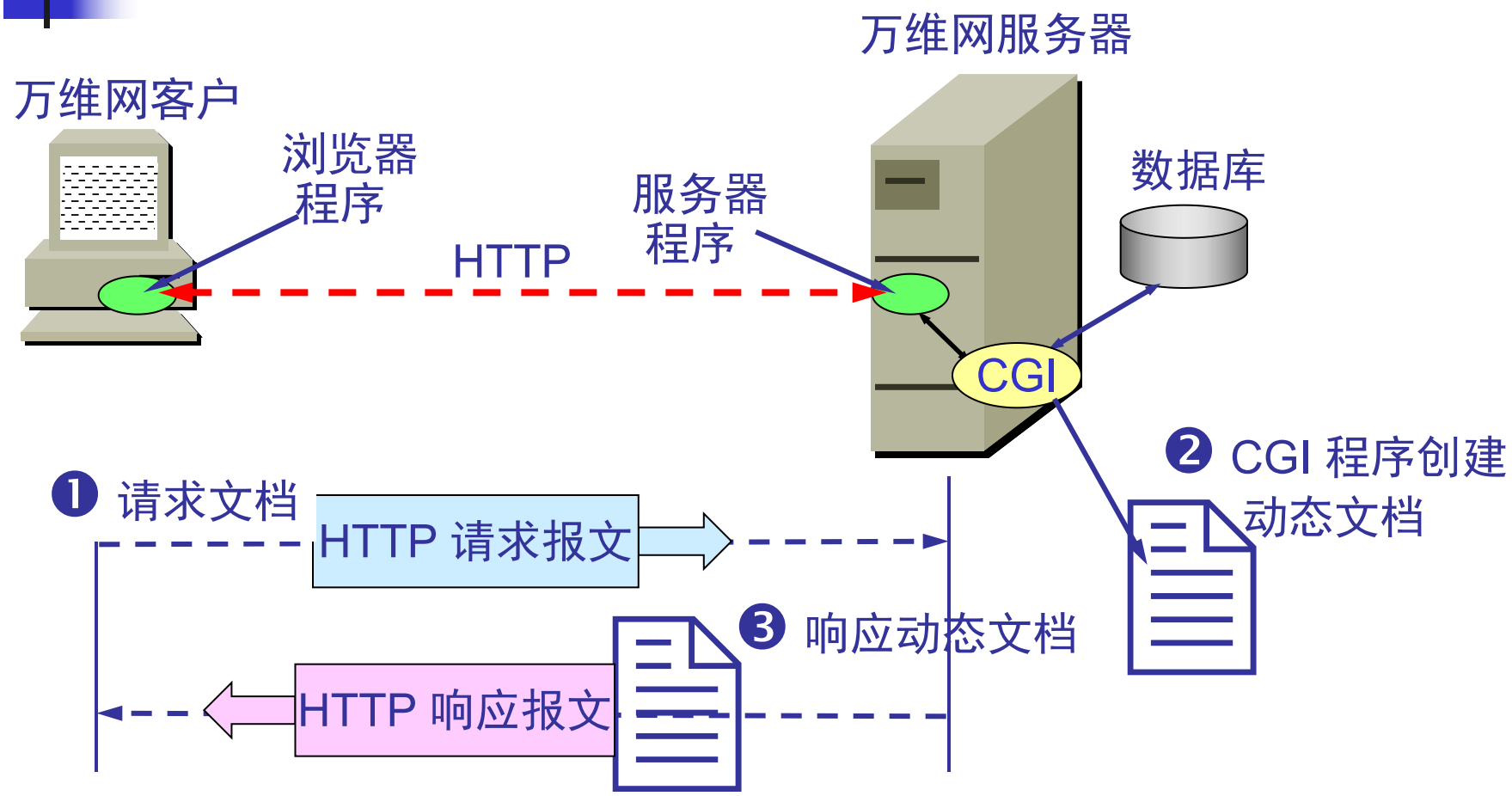
```
<A HREF = "http://www.nwpu.edu.cn">
    <IMG  ALT = "THIS IS A PICTURE"
SRC = /picture.gif  ALIGN = MIDDLE>  </A>
```

- 本地链接：超链指向本服务器中的某个文件。
- 远程链接：超链的终点是其他网点上的页面。

2. 动态万维网文档

- 静态文档：是指该文档创建完毕后就存放在万维网服务器中，在被用户浏览的过程中，内容不会改变。
- 动态文档是指文档的内容是在浏览器访问万维网服务器时才由特殊应用程序(CGI：通用网关接口程序)动态创建。
  - 当浏览器请求到达服务器时，服务器要运行一个特殊应用程序CGI ，并把控制全交给此应用程序；
  - CGI程序对浏览器来的请求进行处理，并输出HTML格式文档，服务器把CGI程序的输出作为对浏览器的响应发送给浏览器。
  - 由于对浏览器每次请求得到的响应都是临时产生，所以用户通过动态文档看到的内容都是不断变化的。

扩充了功能的万维网服务器



CGI 程序

- CGI 程序可以是脚本程序(script)。
  - 利用脚本语言专门编写CGI程序：Perl、JavaScrip 、ASP 、 Tcl\Tk等；
  - “脚本”指一个程序，它可被另一个程序（解释程序）-而不是计算机的处理机-来解释或执行。
- CGI程序也可用一些常用的编程语言编写：C、C++等
- 脚本运行起来要比一般的编译程序要慢，因为它的每一条中间指令先要被解释程序来处理（需要一些附加的指令），而不是直接被指令处理器来处理。

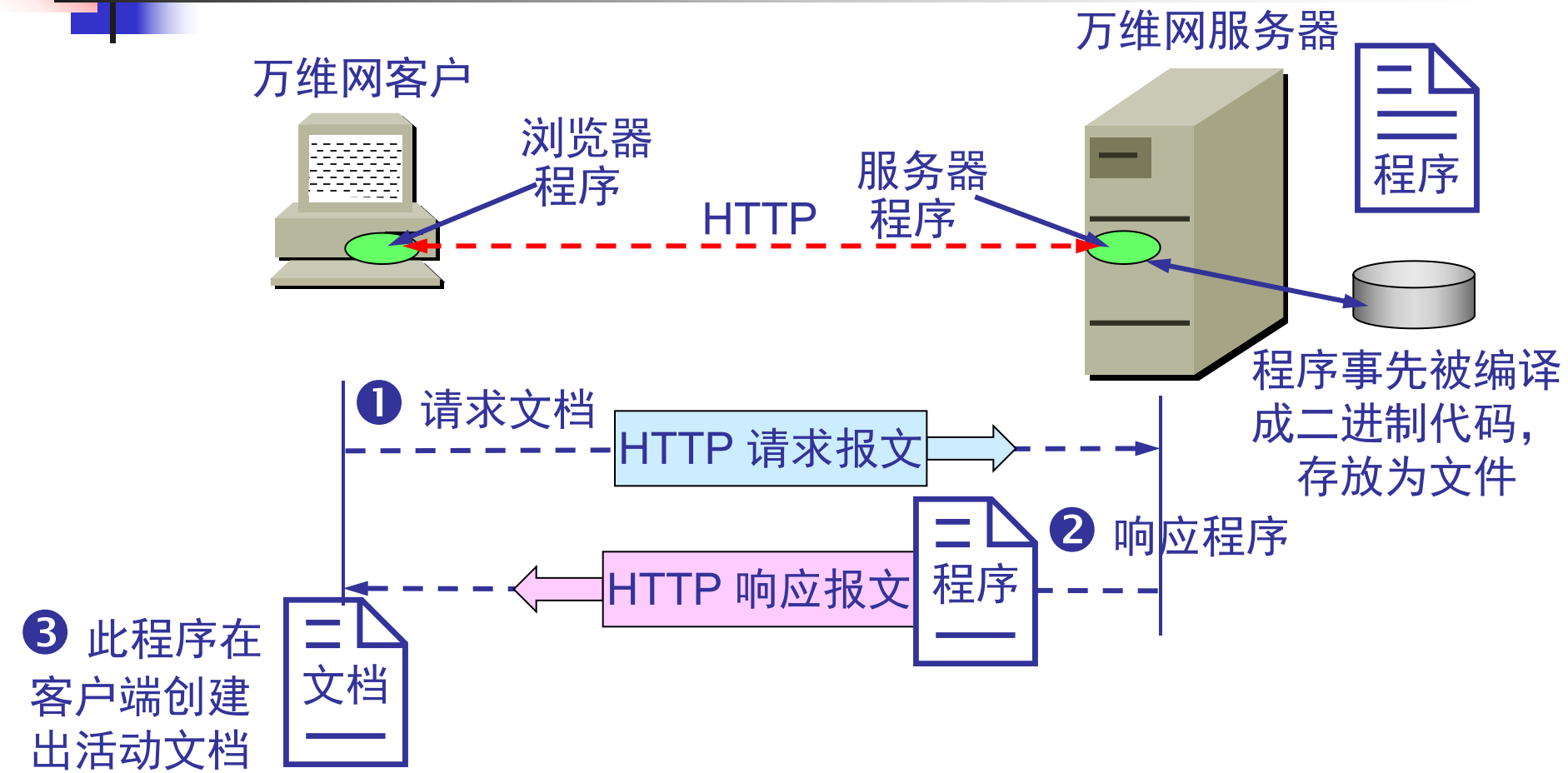
3. 活动万维网文档

- 提出原因
  - 随着HTTP和万维网浏览器、服务器发展，动态文档已不能满足应用需要，因为动态文档一旦建立，所包含的内容就固定下来无法自动刷屏变化，如无法显示动画效果。
- 有两种方法可解决浏览器屏幕显示连续更新
  - 方法一：服务器推送（Server push）-所有工作交给服务器负责。
  - 服务器不断的运行与动态文档关联的CGI程序，由CGI程序定期将更新信息产生更新文档并发送给浏览器。
  - 存在问题：（1）服务器存在大量CGI程序同时运行，造成服务器开销大；（2）更新文档推送时采用TCP连接，该TCP连接在一段时间一直保持，不能释放，而且数量大。
  - 方法二：活动文档技术（Active Document）

3. 活动万维网文档

- 活动文档(active document)技术把网页内容发生变化的工作交给浏览器负责。
- 每当浏览器请求一个活动文档时，服务器就返回一段程序副本（活动文档程序），并在浏览器端运行。
- 活动文档程序可与用户直接交互，并可连续地改变屏幕显示。
- 由于活动文档技术不需要服务器连续发送更新文档，对网络带宽的要求也不会太高，不会占用大量服务器资源。
- 注意：活动文档程序本身并不包括其运行所需要的全部软件，大部分的支持软件（运行环境）都由客户端浏览器提供

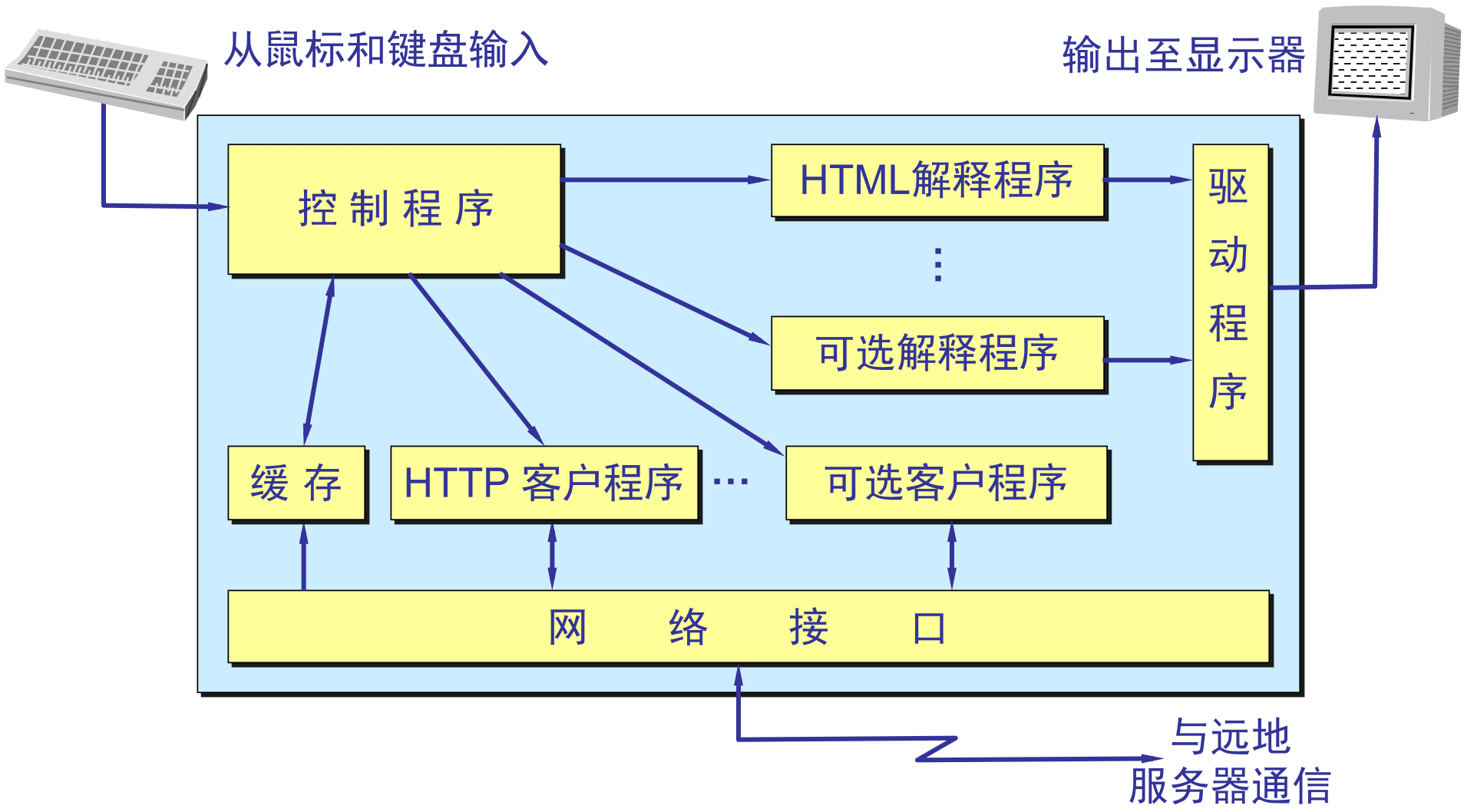
活动文档在客户端创建


万维网客户
浏览器
程序
HTTP
服务器
程序
万维网服务器
程序
程序事先被编译
成二进制代码，
存放为文件
❶ 请求文档
HTTP 请求报文
❷ 响应程序
程序
HTTP 响应报文
❸ 此程序在
客户端创建
出活动文档
文档

用 Java 技术创建活动文档

- 由美国 Sun 公司开发的 Java 语言可以用于创建活动文档。
- 在 Java 技术中使用 “小应用程序”(applet)来编写活动文档程序。
- 用户从万维网服务器下载嵌入了 Java 小应用程序的 HTML 文档后，可在浏览器的屏幕上点击某个图像，就可看到动画效果；或在下拉式菜单中点击某个项目，就可看到计算结果。

4. 浏览器的结构


从鼠标和键盘输入
输出至显示器
控 制 程 序
HTML解释程序
可选解释程序
驱动程序
缓 存
HTTP 客户程序
可选客户程序
网 络 接 口
与远地服务器通信

五、万维网的信息检索系统

- 万维网是一个大规模、联机式信息存储所，用户采用什么方法才能找到自己需要的信息？
- 在万维网中用来进行信息搜索的程序叫做搜索引擎。
- 搜索引擎大体分两类：全文检索搜索引擎和分类目录搜索引擎
- 全文检索搜索引擎是一种纯技术型的检索工具，工作原理：
  - 通过搜索软件（Spider）到因特网上的各网站收集信息，找到一个网站后可以从这个网站再链接到另一个网站。然后按照一定的规则建立一个很大的在线数据库供用户查询。
  - 用户在查询时只要输入关键词，就从已经建立的索引数据库上进行查询（并不是实时地在因特网上检索到的信息），有可能查询到的信息是过时的，需要定期更新。
- 优点：根据关键词可找到具体信息；缺点：量大，不准确。

分类目录搜索

- 分类目录搜索引擎:
  - 不采集网站的任何信息，而是利用各网站向搜索引擎提交的网站信息时填写的关键词和网站描述等信息，经过人工审核编辑后，如果认为符合网站登录的条件，则输入到分类目录的数据库中，供网上用户查询。
- 分类目录搜索也叫做分类网站搜索。
- 分类目录搜索引擎好处：用户可根据搜索网站设计好的目录有针对性的逐级查询所需要的信息，不需要关键词，只需按类查找（先找大类，再找小类），因此查询准确；
- 缺点：分类目录查询结果并不是具体页面，而是被收录网站主页的URL，所得到的内容有限。

一些著名的搜索引擎

- 最著名的全文检索搜索引擎：
  - Google（谷歌）(www.google.com)
  - 百度 (www.baidu.com)
- 最著名的分类目录搜索引擎：
  - 雅虎 (www.yahoo.com)
  - 雅虎中国 (cn.yahoo.com)
  - 新浪 (www.sina.com)
  - 搜狐 (www.sohu.com)
  - 网易 (www.163.com)

两类新的搜索引擎

- 垂直搜索引擎(Vertical Search Engine)
  - 针对某一特定领域、特定人群或某一特定需求提供搜索服务。
  - 垂直搜索也是提供关键字来进行搜索的，返回结果更倾向于基于特殊需求信息等。
  - 举例：对买房子的人讲，希望查找到房子具体供求关系信息（面积、地点、价格）；而不是有关房子一般新闻、事件、论文等。
  - 目前热门垂直搜索行业：购物、旅游、汽车、求职、房产、交友等。

两类新的搜索引擎

- 元搜索引擎(Meta Search Engine)
  - 它把用户提交的搜索请求发送给多个独立的搜索引擎，
  - 把不同的搜索结果集中统一处理，以统一格式提供给用户；
  - 是一种搜索引擎之上的搜索引擎；
  - 目的是：提高搜索速度、实现智能化处理搜索结果、个性化搜索功能的设置和用户搜索界面友好。
  - 元搜索引擎的查全率和查准率均比较高。

本节内容提要

万维网 WWW

- 一、 概述
- 二、 统一资源定位符 URL
- 三、 超文本传送协议 HTTP
- 四、 万维网的文档(HTML)
- 五、 万维网的信息检索系统