第二讲 Web基础知识

陈伟
Email: chenwei@njupt.edu.cn
Tel: 18951896489

课程内容

1. Web基本概念
2. 静态/动态网页
3. Web系统架构
4. HTTP/HTTPS协议概述
5. Web编码和加密

Web基本概念

- Web是万维网（World Wide Web）的简称
- 利用HTTP（Hyper Text Transfer Protocol，超文本传输协议）来建立用户和服务器之间的标准交互方式
- HTML（Hyper Text Markup Language，超文本标记语言）规定了Web应用的页面格式
  - HTML5是HTML的最新版本，HTML5不仅大幅提升Web 应用在交互、系统能力调用、多媒体、语义化等方面的能力，用户无需安装纷繁的插件就可以获得更为丰富的Web 应用。

Web使用了很多新技术

- 真实大型网站会利用很多技术，否则一台服务器无法支持海量用户的访问
  - 负载均衡
  - CDN
  - 云技术
  - ……

静态网页

① 浏览者在浏览器地址栏中输入HTTP请求或链接到该网页地址，该请求通过网络从浏览器传送到Web服务器中。

② Web服务器在服务器中定位该.html或.htm文件，将其转化为HTML流。

③ Web服务器将HTML流通过网络传送到浏览者的浏览器中。

④ 浏览器解析HTML，并显示网页。

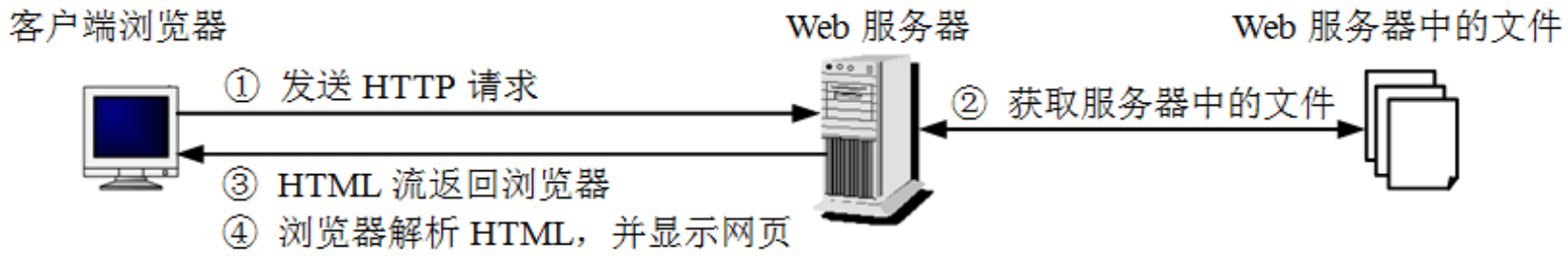


图 1-3　静态网页技术的工作过程

动态网页-1

- 动态网页技术主要分为两种
  - 客户端动态网页技术
  - 服务器端动态网页技术

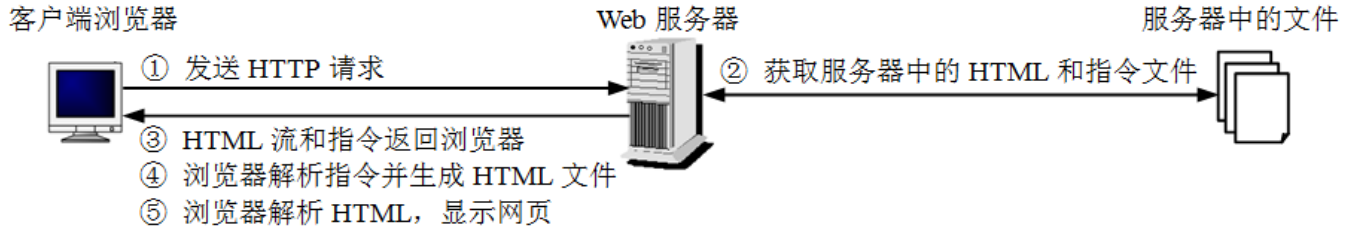


图 1-4　客户端动态网页技术的工作过程

- 客户端动态网页技术
  - 客户端动态网页技术是指Web服务器把原始的HTML页面及一组包含了页面逻辑的脚本、组件等一起发送到客户端，这些脚本和组件包含了如何与浏览者交互并产生动态内容的指令，由客户端的浏览器及其插件解析HTML页面并执行这些指令。典型的客户端动态网页技术包括JavaScript、VBScript、ActiveX控件、Java Applet、Ajax等。

动态网页-2

❖服务器端动态网页技术

- 服务器端动态网页技术是指在Web服务器端根据客户端浏览器的不同请求，动态地生成相应的内容，然后发送给客户端浏览器。

① 用户在客户端浏览器中输入一个HTTP网页请求，通过网络传送到Web服务器中。

② Web服务器在服务器中定位指令文件。

③ Web服务器根据指令生成HTML流。

④ Web服务器将生成的HTML流通过网络传送到浏览者的浏览器中。

⑤ 浏览器解析HTML，显示网页。

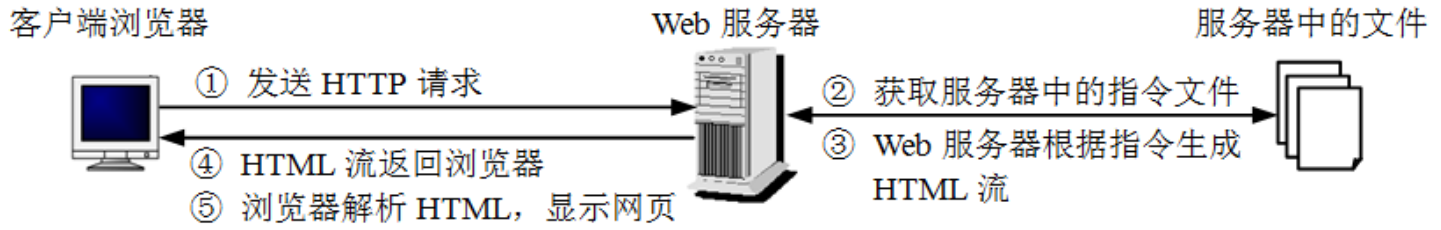
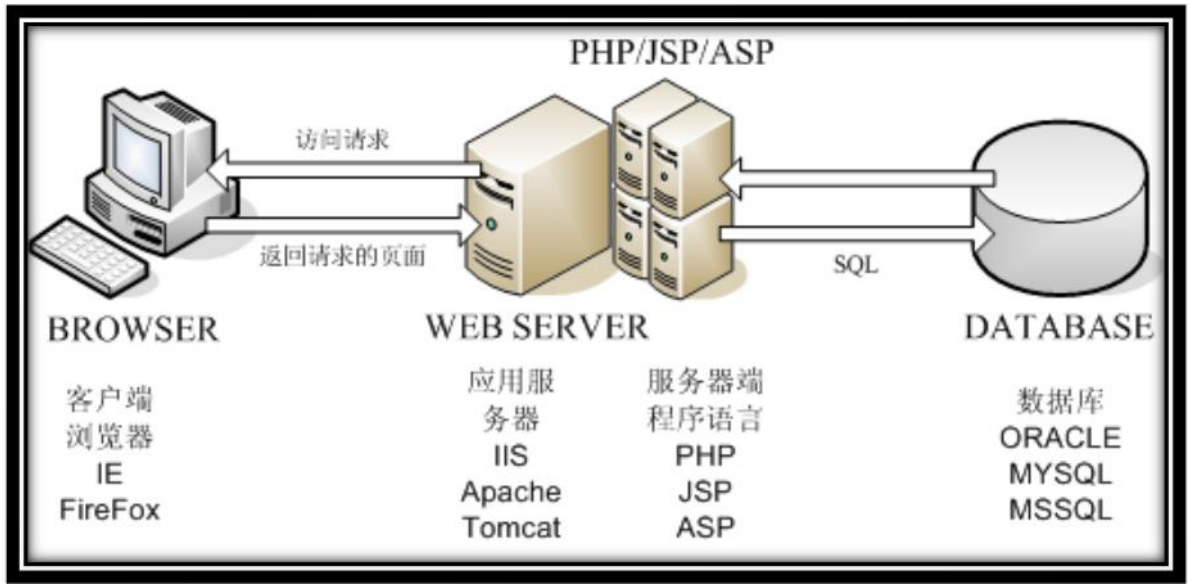


图 1-5　服务器端动态网页技术的工作过程

Web应用系统架构

- 实际上一个Web应用并不是我们所理解的那么简单，它是一个由多个要素构成的系统。Web应用程序的设计者、Web应用服务器、动态脚本引擎、数据库是构成Web应用必不可少的要素



Web服务器选择

1. IIS： 微软公司的Web服务器，内含Gopher服务器和FTP服务器，与Windows Server完全集成
   - 经典组合： Windows Server+IIS+SQLServer+ASP/ASP.NET
2. Apache：一种免费服务器，目前市场占有率排名第一。Apache由一个完全通过互联网运作的非盈利机构Apache Group公布发行
   - 经典组合： Linux+Apache+Mysql+PHP
3. Tomcat：一个JSP和Servlet的运行平台，不仅是一个Servlet容器，同时也具有传统Web服务器的功能，即处理HTML网页。但是与Apache相比，它处理静态网页的能力稍逊。但可以将Tomcat和Apache集成到一块，让Apache处理静态网页，而Tomcat处理JSP和Servlet。

操作系统

项目	Linux	Windows
性能	好	差
安全性	好	差
维护难度	较难	容易
开发难度	较难	较容易

数据库

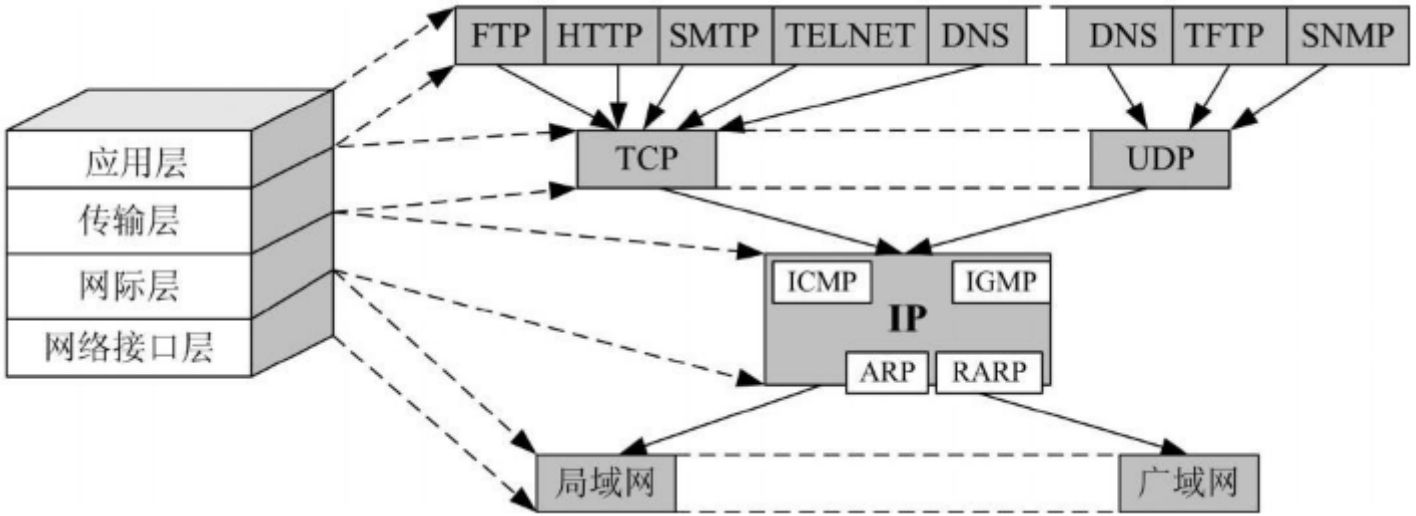
- 相比于操作系统和Web服务器，数据库的选择面较宽，微软的SQL Server、Oracle、Mysql、Sybsae、DB2等都可以使用
- 通常大型平台选择Oracle、Sybase较多；
- 采用Windows操作系统的中小型平台采用微软SQL Server较多；
- 而采用Linux作为操作系统的平台则大多数选用Mysql。Linux/Apache Tomcat/ Mysql构成一个很好低成本、高性能组合，不过相比于Windows/IISSQL Server组合，需要的技能更高一些

Web常用开发语言介绍

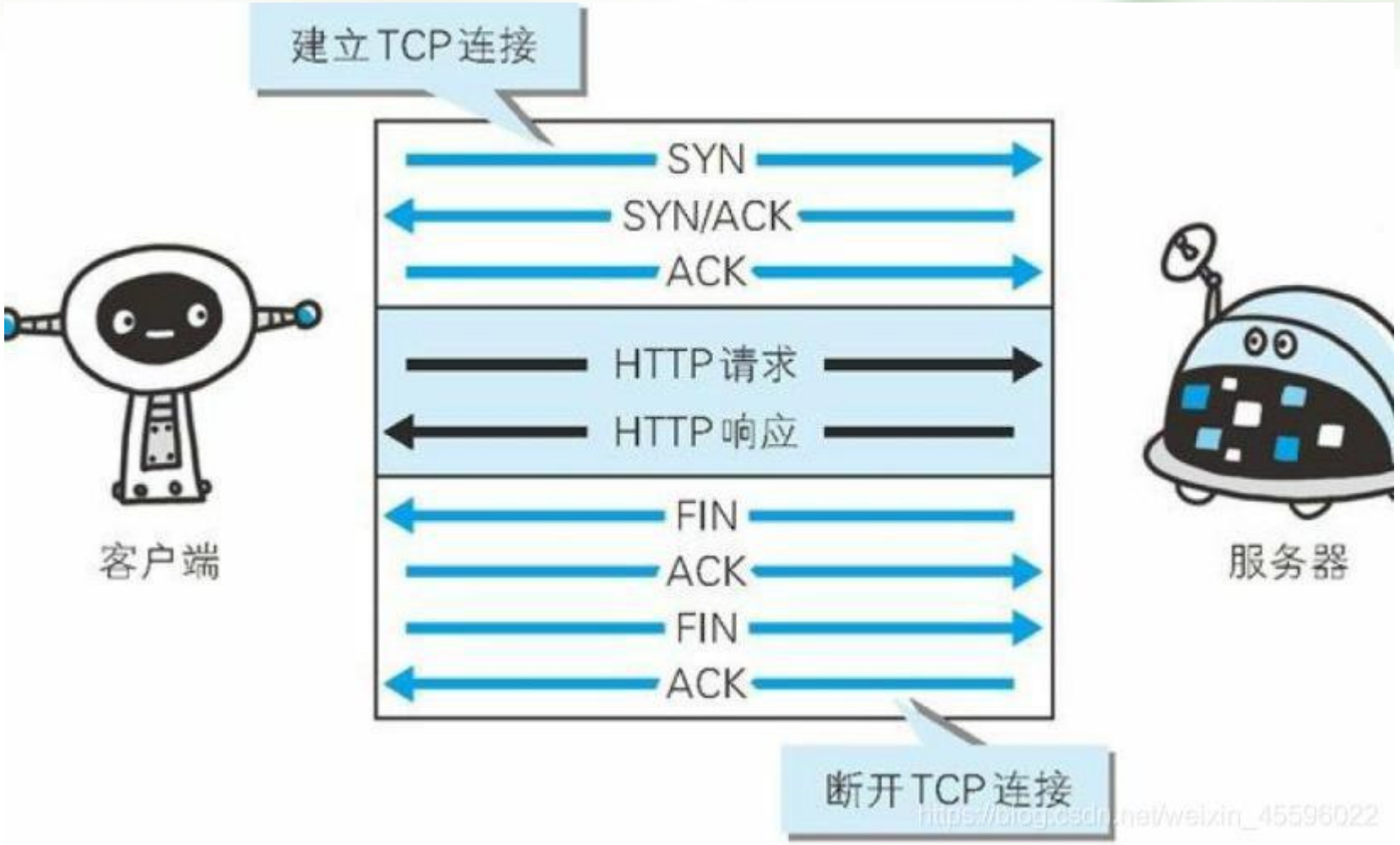
- **ASP（Active Server Pages）**
  - 使用VBScript脚本语言，可以将脚本语言直接嵌入HTML文档中，不需要编译就可以直接运行。
- **ASP.NET**
  - 一种用于创建动态Web页的强大的服务器端新技术，它可为WWW站点或企业内部互联网创建动态的、可进行交互的HTML页面。
- **JSP**
  - JSP页面由HTML代码和嵌入其中的Java代码组成，具有良好的跨平台性。
- **PHP**
  - PHP程序最初是用Perl语言编写的简单程序，后来经其他程序员不断完善，于1997年发布了功能基本完善的PHP3。

HTTP协议

- HTTP是一个应用层的面向对象的协议
- 于1990年提出，是互联网中应用最为广泛的应用层协议

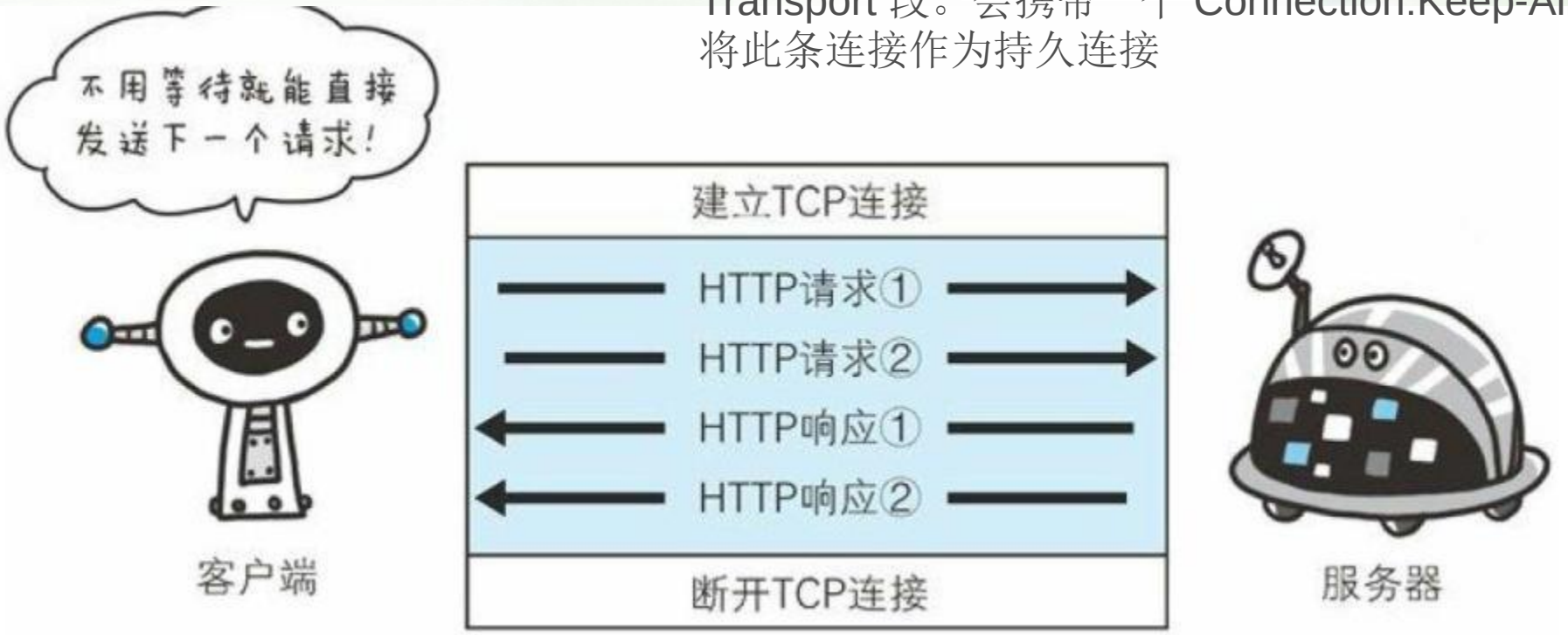


每次请求都要建立连接吗？



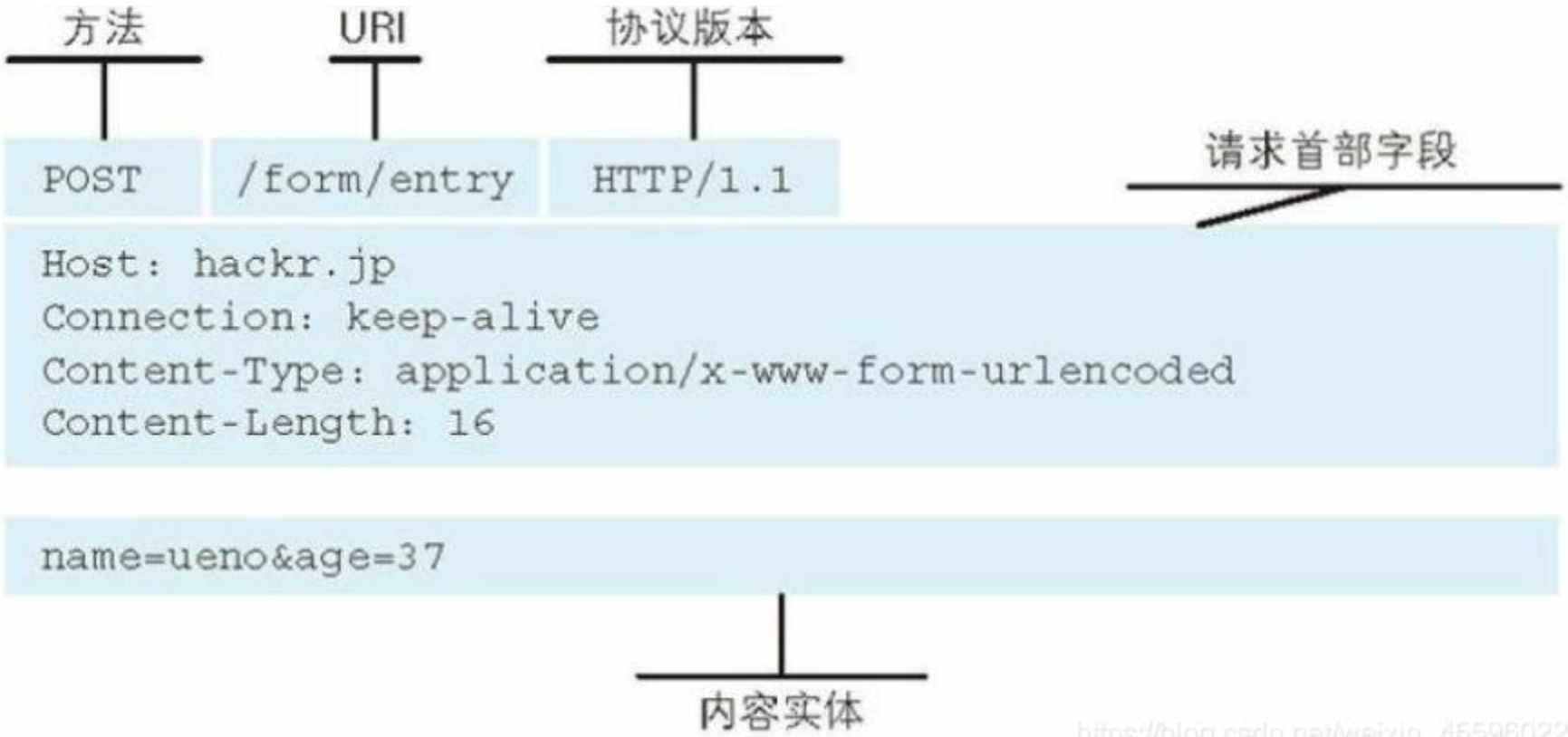
持久连接

所以在 HTTP/1.1 中改用了持久连接，就是在一次连接建立之后，只要客户端或者服务端没有明确提出断开连接，那么这个 tcp 连接会一直保持连接状态持久连接的一个最大的好处是：大大减少了连接的建立以及关闭时延。 HTTP1.1 中有一个 Transport 段。会携带一个 Connection:Keep-Alive，表示希望将此条连接作为持久连接



HTTP协议包含两个报文

- 一个是请求报文，一个是响应报文



响应报文

状态码

200：一切正常
301：永久重定向
404：请求资源不存在
500：服务端内部错误

	类别	原因短语
1xx	信息性状态码	接收的请求正在处理
2xx	成功状态码	请求的正常处理完毕
3xx	重定向状态码	需要进行附加操作以完成请求
4xx	客户端错误状态码	服务器无法处理请求
5xx	服务器错误状态码	服务器处理请求出错

HTTP协议的主要特点

- HTTP协议足够简单，每次请求均为独立行为，无状态的特点
- HTTP协议支持B/S模式，只要有浏览器，就可以工作。用户使用简单，易于操作
- HTTP协议灵活性好，可以用于数据传输、视频播放、交互等
- 但HTTP本身并没有太好的防范措施，大量的安全问题都是HTTP应用带来的

URL的基本格式

- schema://host[:port#]/path/…/?[url-params]#[query-string]
- http://www.baidu.com:80/java/index.html?name=aaa
- **MIME Type：描述消息内容类型的因特网标准**
  - 文本文件：text/html,text/plain,text/css,application/xhtml+xml,application/xml
  - 图片文件：image/jpeg,image/gif,image/png.
  - 视频文件：video/mpeg,video/quicktime

媒体类型（ MIME ）是一种标准，用来表示文档、文件或字节流的性质和格式。

GET与POST请求的区别

- HTTP中的GET与POST之间有很多的区别
- GET
  - 例如：在开发一个新闻模块时，我们需要动态的获取数据，一般就会利用GET请求中的参数来传递需要获得的新闻ID
  - http://www.domain.com/news.php?id=3

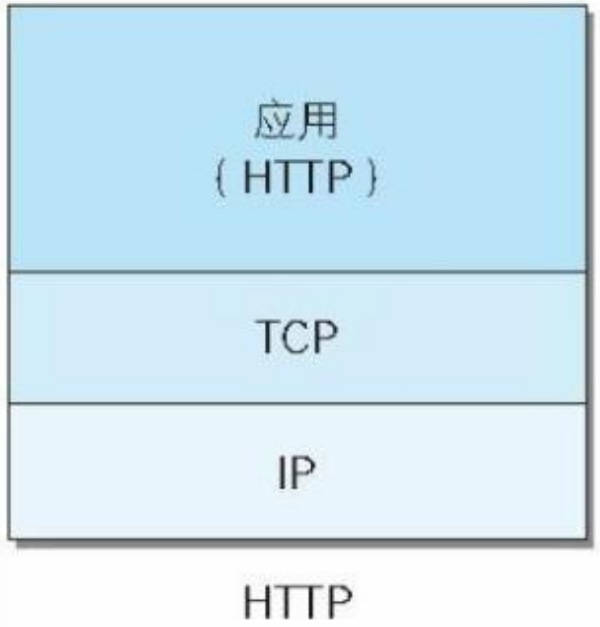
GET与POST请求的区别

- **POST**
  - 从某个角度来看，GET请求能传递的数据是有限的，因为它的数据是在URL中，而URL的长度是有限的。而POST请求的数据是放置在请求体中，也就是说POST请求能承载更多的数据，理论上来说POST请求是没有大小限制的。
  - 例如，发送邮件、文件上传时，我们就必须使用POST请求来完成。
- **简单来说：GET是向服务器发索取数据的一种请求，而POST是向服务器提交数据的一种请求。**

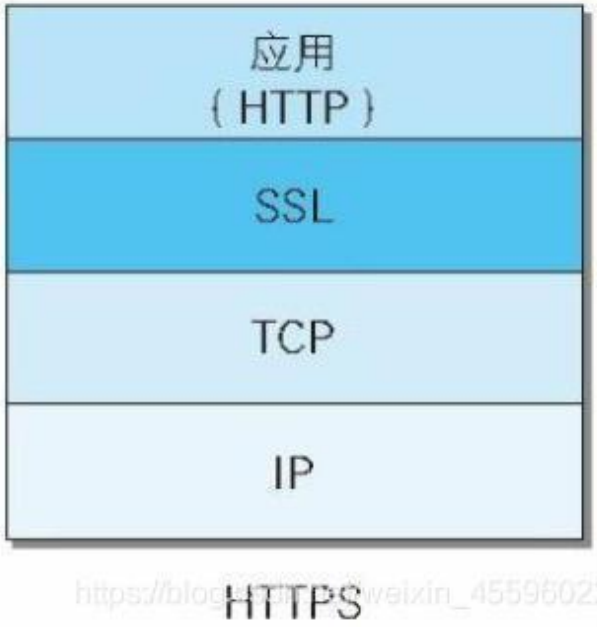
HTTPS协议的安全性分析

- 两个核心问题
  - 如何建立安全的传输通达
  - 如何确认双方的身份

https 是一种加密的超文本传输协议，它与 HTTP 在协议差异在于对数据传输的过程中，https对数据做了完全加密。由于 http 协议或者 https 协议都是处于 TCP 传输层之上，同时网络协议又是一个分层的结构，所以在 tcp 协议层之上增加了一层 SSL（Secure Socket Layer，安全层）或者 TLS（Transport Layer Security） 安全层传输协议组合使用用于构造加密通道。



HTTP



HTTPS

简化的SSL协议

Alice

我们能交谈吗?,密码算法列表, R_A

证书, 选择的密码算法, R_B

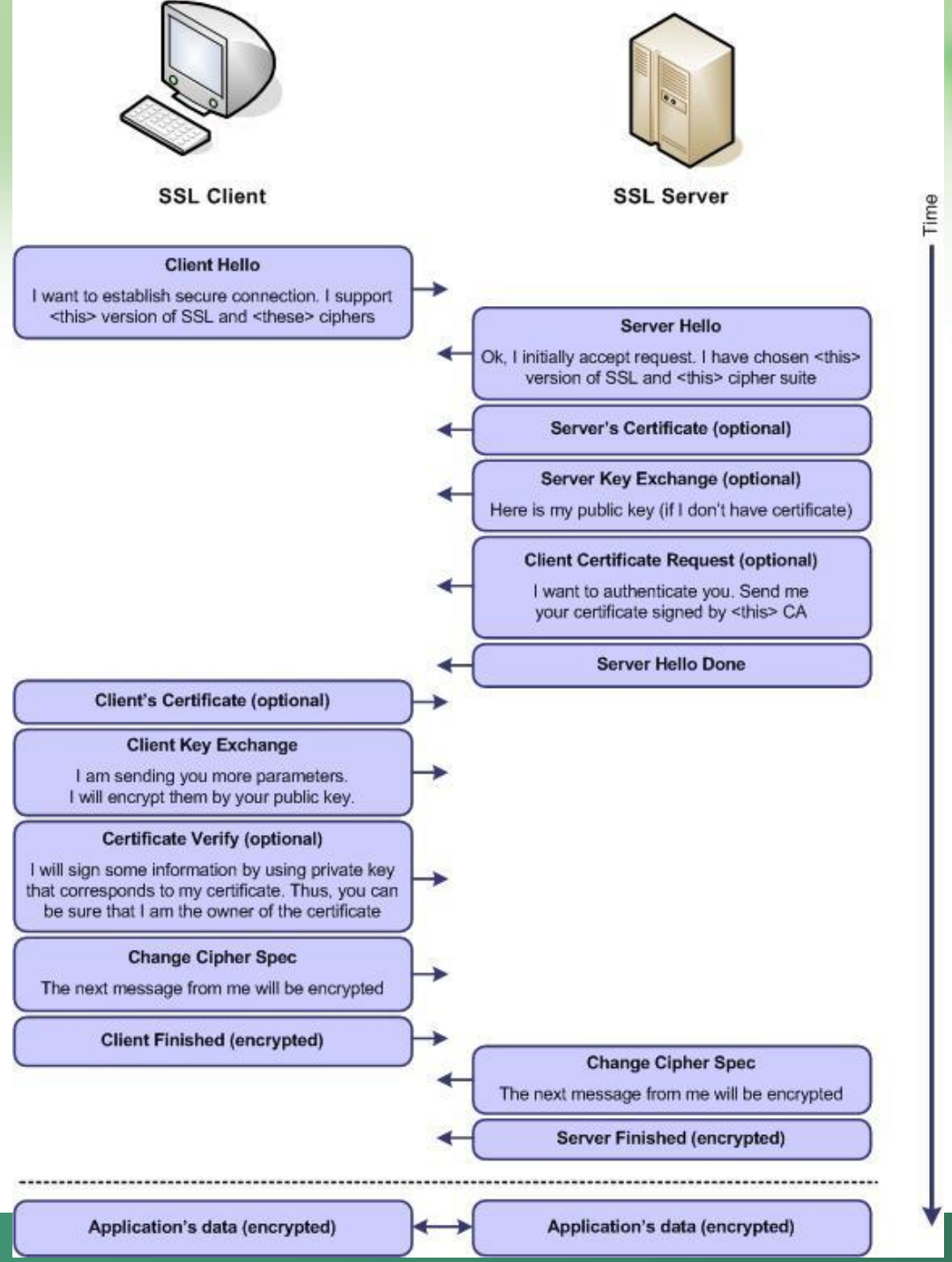
$\{S\}_{Bob}$, E(h(msgs,CLNT,K),K)

h(msgs,SRVR,K)

使用密钥 K保护的数据

Bob

- S 是pre-master secret
- $K = h(S,R_A,R_B)$
- msgs = 所有以前的消息
- CLNT 和 SRVR 是常量

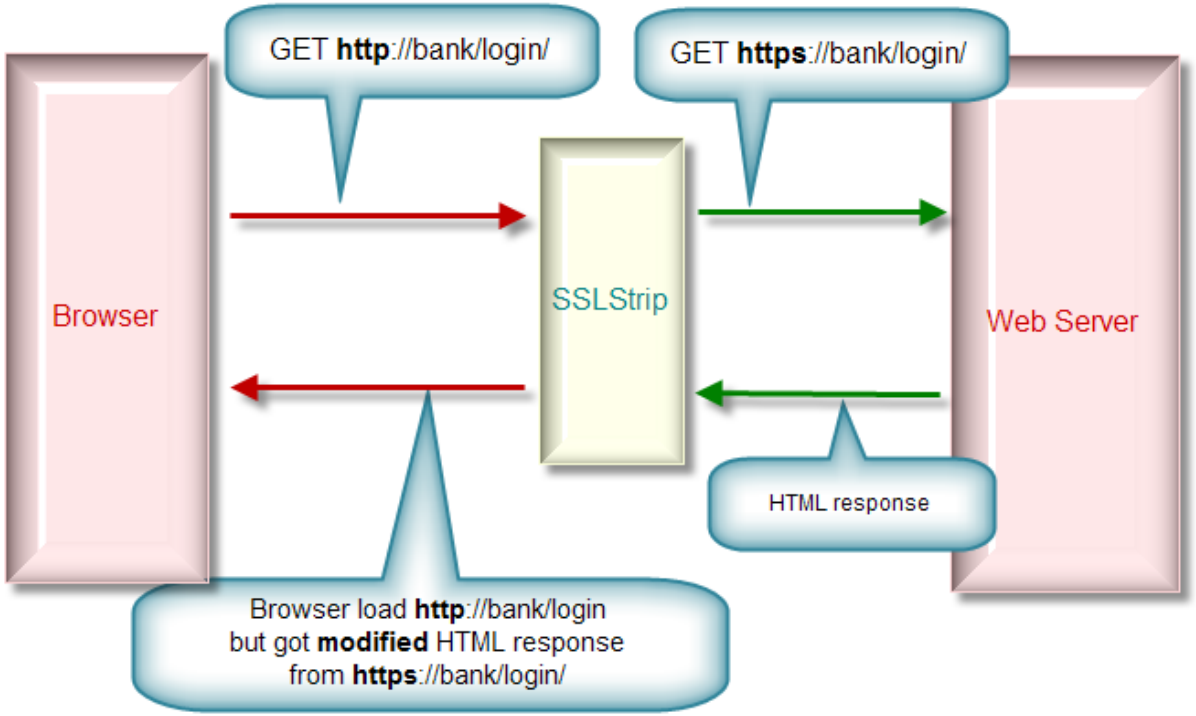

SSL Client
SSL Server
Time
Client Hello
I want to establish secure connection. I support <this> version of SSL and <these> ciphers
Server Hello
Ok, I initially accept request. I have chosen <this> version of SSL and <this> cipher suite
Server's Certificate (optional)
Server Key Exchange (optional)
Here is my public key (if I don't have certificate)
Client Certificate Request (optional)
I want to authenticate you. Send me your certificate signed by <this> CA
Server Hello Done
Client's Certificate (optional)
Client Key Exchange
I am sending you more parameters. I will encrypt them by your public key.
Certificate Verify (optional)
I will sign some information by using private key that corresponds to my certificate. Thus, you can be sure that I am the owner of the certificate
Change Cipher Spec
The next message from me will be encrypted
Client Finished (encrypted)
Change Cipher Spec
The next message from me will be encrypted
Server Finished (encrypted)
Application's data (encrypted)
Application's data (encrypted)

HTTPS认证

- 单项认证
  - 对服务器进行认证
- 双向认证
  - 对服务器认证
  - 对客户端认证

SSL剥离攻击(SSL Strip)

- SSL剥离攻击也叫 https 降级攻击，攻击者拦截用户流量后，欺骗用户与攻击者进行 http 通信，攻击者与服务器保持正常通信 （http 或 https），从而获取用户信息。



HTTPS特点总结

- HTTPS并没有改变HTTP协议本身的特性，只是利用了SSL
- HTTPS可以有效保障用户信息不被泄露
- HTTPS主要防护传输过程中的安全，并不会有效提升服务器的安全性
  - 例如使用Burpsuite
  - 如果采用双向认证呢？
- 最常见的实现套件：OpenSSL

课程内容

1. Web基本概念
2. 静态/动态网页
3. Web系统架构
4. HTTP/HTTPS协议概述
5. Web编码和加密

针对字符的编码

- 单字节编码
  - ASCII：8 bit
- 双字节字符集
  - GBK字符集：16 bit，支持65536个汉字
- 不同国家和地区采用的编码不一致，导致乱码
- Unicode编码主要解决多语言环境下的统一集合
  - UTF-8
    - 主流编码方式，3字节
  - UTF-16

传输过程的编码

- URL编码
- RFC3986文档规定，Url中只允许包含英文字母（a-zA-Z）、数字（0-9）、-_.~4个特殊字符以及所有保留字符
- Http协议中参数的传输是“key=value”这种键值对形式的，如果要传多个参数就需要用“&”符号对键值对进行分割
- 如果参数值中就包含=或&这种特殊字符的时候该怎么办
  - name1= va&lu=
- 在特殊字符的各个字节前加上%，例如，我们对上述会产生奇异的字符进行URL编码后结果：
  - name1= va%26lu%3D

Base64编码

- 网络上常见的用于传输8bit字节代码的编码方式
- 将3个8bit转化为4个6bit的字节
- 编码后是4的倍数，不足4bit用等号（=）填充
- 好辨识，含有大小写字母和+、-、=等符号

Base64，顾名思义，就是包括小写字母a-z、大写字母A-Z、数字0-9、符号"+"、"/"一共64个字符的字符集，（另加一个“=”，实际是65个字符，至于为什么还会有一个“=", 这个后面再说）。任何符号都可以转换成这个字符集中的字符，这个转换过程就叫做base64编码。

HTML字符实体

- HTML字符实体（Character Entity）是用来表示HTML中危险字符的方案，也是解决跨站脚本（XSS）攻击的有效手段
  - 不能使用小于号(<)和大于号(>)
- HTML字符实体的特点是以&开头，并以分号结尾
  - < <
  - > >

在 HTML 中，某些字符是预留的。
在 HTML 中不能使用小于号（<）和大于号（>）， 这是因为浏览器会误认为它们是标签。
如果希望正确地显示预留字符，我们必须在 HTML 源代码中使用字符实体（character entities）。

Web系统中的加密措施

- 标准的加密方法是对用户提交的参数（如密码）进行加密后再传输，到Web服务器后，再将参数解密后处理。
- 不需要服务器知道明文的内容
  - 常见用户的隐私信息，如用户密码
  - 散列函数 MD5/SHA-1
  - 简单的加盐也会存在一定安全隐患
- 需要服务器知道明文的内容
  - 如订单信息、留言
  - 前沿研究方向：云服务器上的密文检索