HTTP协议有关常识

原创 GeorgeKai 2018-02-02 18:06:00 评论(0) 417人阅读

HTTP协议基础原理

1.1 面试题：简述用户访问网站的过程？

1. 利用DNS协议进程域名解析
2. 建立TCP协议三次握手过程
3. 客户端发出访问网站相应页面请求
   系统架构部署情况（可以说一说）
4. 服务端响应访问页面的请求信息
5. 断开TCP协议四次挥手过程

1.2 HTTP协议报文结构

1.2.1 请求报文结构：

查看请求和响应报文的结构：两种方法

```
curl www.baidu.com  -v
wget --debug www.baidu.com
```

1. 请求行：

1）请求的方法—get：客户端请求指定资源信息，服务器返回指定信息（没有请求主体）

post：将客户端的数据提交到服务器（如：查询,注册）（有请求主体）

表 6-1 常用的 HTTP 请求方法

HTTP 方法	作用描述
GET	客户端请求指定资源信息，服务器返回指定资源
HEAD	只请求响应报文中的 HTTP 首部
POST	将客户端的数据提交到服务器，例：注册表单
PUT	用从客户端向服务器传送的数据取代指定的文档内容。
DELETE	请求服务器删除 Request-URI 所表示的资源。
MOVE	请求服务器将指定的页面移至另一个网络地址。

2）请求的数据信息（默认请求index.html）

3）请求http协议版本

TCP协议分为长连接（http1.1）和短连接（http1.0）

可以看看http2.0版本特性

2. 请求头：客户信息 user-agent：用户使用的浏览器/代理

Host：www.baidu.com 要访问的网站信息

3. 空行：表示请求头结束

4. 请求主体：get没有请求主体，post有请求主体

1.2.2 响应报文结构：

1. 起始行：

1）HTTP协议版本信息

2）相应的状态码信息：

表 6-2 不同范围的状态码及其对应的作用

状态码范围	作用描述
100 - 199	用于指定客户端相应的某些动作
200 - 299	用于表示请求成功
300 - 399	用于已经移动的文件，并且常被包含在定位头信息中指定新的地址信息
400 - 499	用于指出客户端的错误
500 - 599	用于指出服务端的错误

状态代码	详细描述说明
200 - OK	服务器成功返回网页，这是成功的 HTTP 请求返回的标准状态码
301 - Moved Permanently	永久跳转，所请求的网页将永久跳转到被设定的新位置， 例如：从 etiantian.org 跳转到 www.etiantian.org
302 - Moved Temporarily	临时跳转，所请求的网页将临时跳转到被设定的新位置 例如：从 www.jd.com 跳转到 https://www.jd.com
403 - Forbidden	禁止访问，虽然这个请求是合法的，但是服务器端因为匹配了预先设置的规则而拒绝响应客户端的请求，此类问题一般为服务器或服务权限配置不当所致。 Nginx 403 forbidden 多种原因及故障模拟重现 http://oldboy.blog.51cto.com/2561410/1633952 apache 服务 Forbidden 403 问题精彩总结 http://oldboy.blog.51cto.com/2561410/581383
404 - Not Found	服务器找不到客户端请求的指定页面，可能是客户端请求了服务器上不存在的资源所致。
500 - Internal Server Error	内部服务器错误，服务器遇到了意料不到的情况，不能完成客户的请求。 这是一个较为笼统的报错，一般为服务器的设置或内部程序问题导致。 例如：SELinux 开启，而又没有为 HTTP 设置规则许可，客户端访问就是 500
502 - Bad Gateway（重点）	坏的网关，一般是代理服务器请求后端服务时，后端服务不可用或没有完成响应网关服务器。这通常为反向代理服务器下面的节点出问题所致。 反向代理服务器无法与后面的 web 服务节点服务器建立联系
503 - Service Unavailable	服务当前不可用，可能是服务器超载或停机维护导致的，或者是反向代理服务器后面没有可以提供服务的节点
504 - Gateway Timeout	网关超时，一般是网关代理服务器请求后端服务时，后端服务没有在特定的时间内完成处理请求。多数是服务器过载导致没有在指定的时间内返回数据给前端代理服务器。
生产环境常见 HTTP 状态码的博客文章见 http://oldboy.blog.51cto.com/2561410/716294 http 权威指南	

PS：www.jd.com以前的域名www.360buy.com

2. 响应头部：服务器有关信息

3. 空行：表示响应头结束

4. 响应主题内容：一般为html、css、js等代码信息

1.3 HTTP协议资源类型和名词概念介绍

1. 媒体资源类型
   对于web服务可以处理的用户请求资源信息（html、 xml）
2. URL
   全称为Uniform Resource Location（统一资源定位符）
3. URI
   全称为Uniform Resource Identifier（统一资源标识符）
   channel.jd.com / women.html ：nginx的识别方法
   url部分 uri部分

4. 静态网页资源
   静态资源特点：
   1. 纯文本类程序或文件，如.html、.htm、.xml、.shtml、.js、.css等；
      图片类文件或数据文档，如.jpg、.gif、.png、.bmp、.txt、.doc、.ppt等；
      视频类流媒体文件，如.mp4、.swf、.avi、.wmv、.flv等。
   2. 每个网页的内容都是保存在网站服务器文件系统上的，
      也就是说，静态网页是实实在在保存在服务器上的文件实体
   3. 网页内容是固定不变的，因此，容易被搜索引擎收录 （优点）
   4. 因为网页没有数据库的支持，所以在网站制作和维护方面的工作量较大，（缺点）
      当网站信息量很大时，完全依靠静态网页比较困难
   5. 网页的交互性较差，在程序的功能实现方面有较大的限制 （缺点）
   6. 网页程序在用户浏览器端解析
      当客户端向服务器请求数据时，服务器会直接从磁盘文件系统上返回数据（不做任何解析） （优点）

应用场合：

1. 门户新闻业务

新闻网站的特点是一旦发布完成，几乎不会再改动网页内容，因此，新闻业务内容静态化相对比较简单。

第一步：程序要支持发布的内容由动态转成静态的功能。

第二步：运营编辑人员发布新闻网页后，后台程序立刻将动态网页生成静态文件。

第三步：运维人员通过发布或事件触发把运营编辑生成的静态网页发布到事先搭建好的公司缓存集群服务器上，或者把静态内容同步到购买的全国所有 CDN 服务器节点上，然后，再提供给用户进行访问浏览。

5. 动态网页资源（所谓的动态网页是与静态网页相对而言的）

   动态资源特点：
   1. 网页扩展名后缀常见为：.asp、.aspx、.php、.js、.do、.cgi等
      动态资源网页中会出现? &等特殊符号信息
   2. 网页一般以数据库技术为基础，大大降低了网站维护的工作量。
   3. 采用动态网页技术的网站可以实现更多的功能，
      如用户注册、用户登录、在线调查、投票、用户管理、订单处理、发博文等
   4. 动态网页并不是独立存在于服务器上的网页文件
   5. 动态网页资源不便于被搜索引擎收录
   6. 网页程序在服务架构端进行解析

应用场合：

2. **视频网站业务**

视频网站和新闻网站类似，特点都是一旦发布完成，几乎不会再改动网页内容。因此，实现视频业务网站高效访问也很简单。

以优酷视频网为例，用户在上传视频时，需要经历转码---审核的过程（大概 1 个小时）。此外，一些热点视频也可能会被提前推送同步到 CDN 的核心节点或全国所有 CDN 服务器节点，这样用户访问时才会更快。

6. 伪静态网页
取长补短（伪静态资源实质是动态资源）
集合了动态和静态的优点：便于搜索引擎查询， 可以处理动态代码，有数据库的支持

应用场合：

3. **Blog/BBS/SNS/微博社区业务/电商（淘宝，京东）**

这几类业务由动态转静态是比较困难的，因为，用户发布内容后，可能会随时更新并查看，对于这种情况，一般会通过异步的方式来处理，例如通过消息中间件技术加上 NoSQL 集群技术来实现转换。较为详细的说明见博客文章《浅谈千万级 PV/IP 规模高性能高并发网站架构》（http://oldboy.blog.51cto.com/2561410/736710）和（http://edu.51cto.com/course/course_id-3093.htr

1.4 网站访问统计

1.4.1 网站流量度量语

1）IP
IP（独立IP），即Internet Protocol，这里指独立IP数，
独立IP数是指不同IP地址的计算机访问网站时被计的总次数
一般一天内（00:00-24:00）相同IP地址（公网IP）的客户端访问网站页面只会被计一次
2）PV
PV（访问量）即Page View，中文翻译为页面浏览
即页面浏览量或点击量，不管客户端是不是相同，也不管IP是不是相同
用户只要访问网站页面就会被计算PV，一次计一个PV。
3）UV
UV（独立访客）即Unique Visitor，同一个客户端（PC或移动端）访问网站被计为一个访客。
一天（00：00-24:00）内相同的客户端访问同一个网站只计一次UV。
UV一般是以客户端Cookie等技术作为统计依据

1.4.2 cookie和session的区别：

• cookie：根据用户信息，为用户设定一个身份标识信息，便于统计和识别用户信息（保存在客户端本地）， 可以理解为钥匙
• session：记录用户信息，用户情况（保存在服务端本地），可以理解为锁头

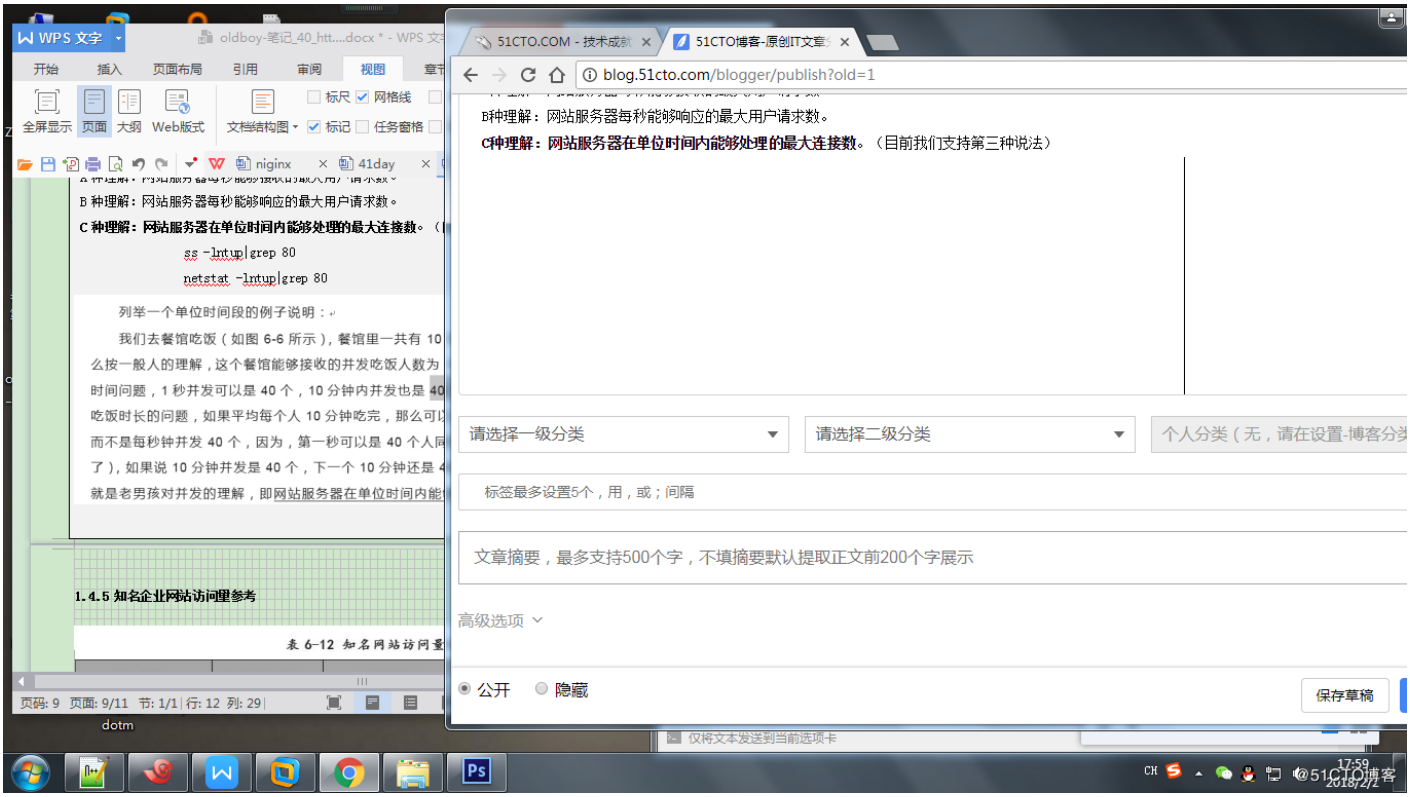
1.4.3 网站访问统计方法

常用的统计工具：

- 网页信息统计软件-piwik：
  pwiki统计工具：（https://piwik.org/）
  pwiki演示页面：https://piwik.org/demo
- ELK软件介绍说明：
  http://blog.oldboyedu.com/elk/

1.4.4 并发连接概念

A种理解：网站服务器每秒能够接收的最大用户请求数。
B种理解：网站服务器每秒能够响应的最大用户请求数。
C种理解：网站服务器在单位时间内能够处理的最大连接数。（目前我们支持第三种说法）



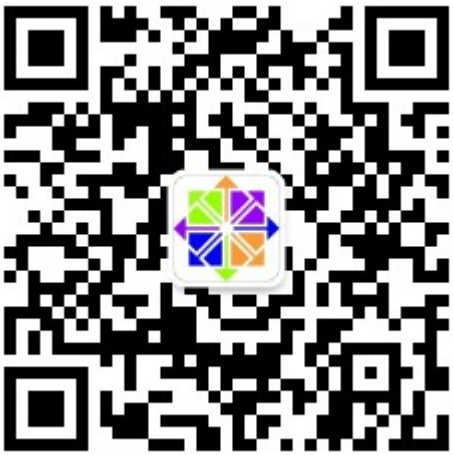
1.4.5 知名企业网站访问量参考

表 6-12 知名网站访问量信息参考

网　站	独立 IP 万/日	PV 数万/日	网站并发级别	机器数量
www.51cto.com	582'000	1'338'600	10000	数十台
www.ganji.com	1'734'000	13'872'000	10000~30000	几百台
www.58.com	1'398'000	22'927'200	10000~30000	几百台
www.weibo.com	30'180'000	166'593'600	几十万	千台
www.taobao.com	46'620'000	489'510'000	几十万~百万	万台
www.jd.com	6'108'000	98'949'600	数万	千台
www.163.com	10'320'000	79'154'000	十万	千台
www.suning.com	623,250	3,365,550	10000~30000	@51CTO博客

网站访问量统计地址：http://alexa.chinaz.com/alexa_more.aspx

小伙伴们可以关注我的微信公众号：linux运维菜鸟之旅



关注“中国电信天津网厅”公众号，首次绑定可免费领2G流量，为你的学习提供流量！