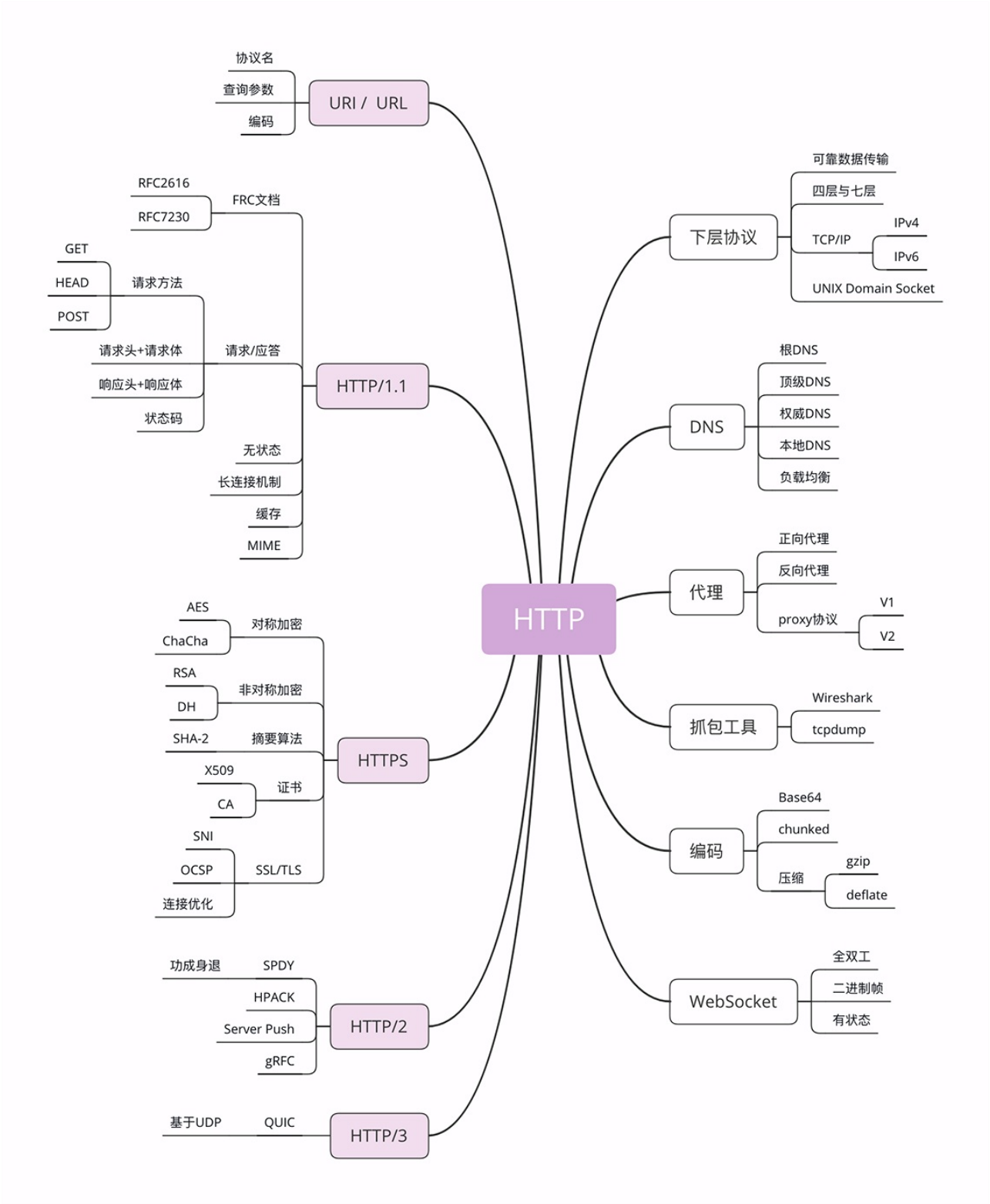
04-HTTP世界全览（下）：与HTTP相关的各种协议

在上一讲中，我介绍了与HTTP相关的浏览器、服务器、CDN、网络爬虫等应用技术。

今天要讲的则是比较偏向于理论的各种HTTP相关协议，重点是TCP/IP、DNS、URI、HTTPS等，希望能够帮你理清楚它们与HTTP的关系。

同样的，我还是画了一张详细的思维导图，你可以点击后仔细查看。



TCP/IP

TCP/IP协议是目前网络世界“事实上”的标准通信协议，即使你没有用过也一定听说过，因为它太著名了。

TCP/IP协议实际上是一系列网络通信协议的统称，其中最核心的两个协议是**TCP**和**IP**，其他的还有UDP、ICMP、ARP等等，共同构成了一个复杂但有层次的协议栈。

这个协议栈有四层，最上层是“应用层”，最下层是“链接层”，TCP和IP则在中间：**TCP属于“传输层”，IP属于“网际层”**。协议的层级关系模型非常重要，我会在下一讲中再专门讲解，这里先暂时放一放。

IP协议是“Internet Protocol”的缩写，主要目的是解决寻址和路由问题，以及如何在两点间传送数据包。IP协议使用“**IP地址**”的概念来定位互联网上的每一台计算机。可以对比一下现实中的电话系统，你拿着的手机相当于互联网上的计算机，而要打电话就必须接入电话网，由通信公司给你分配一个号码，这个号码就相当于IP地址。

现在我们使用的IP协议大多数是v4版，地址是四个用“.”分隔的数字，例如“192.168.0.1”，总共有2^32，大约42亿个可以分配的地址。看上去好像很多，但互联网的快速发展让地址的分配管理很快就“捉襟见肘”。所以，就又出现了v6版，使用8组“:”分隔的数字作为地址，容量扩大了很多，有2^128个，在未来的几十年里应该是足够用了。

TCP协议是“Transmission Control Protocol”的缩写，意思是“传输控制协议”，它位于IP协议之上，基于IP协议提供可靠的、字节流形式的通信，是HTTP协议得以实现的基础。

“可靠”是指保证数据不丢失，“字节流”是指保证数据完整，所以在TCP协议的两端可以如同操作文件一样访问传输的数据，就像是读写在一个密闭的管道里“流动”的字节。

在第2讲时我曾经说过，HTTP是一个"传输协议"，但它不关心寻址、路由、数据完整性等传输细节，而要求这些工作都由下层来处理。因为互联网上最流行的是TCP/IP协议，而它刚好满足HTTP的要求，所以互联网上的HTTP协议就运行在了TCP/IP上，HTTP也就可以更准确地称为“**HTTP over TCP/IP**”。

DNS

在TCP/IP协议中使用IP地址来标识计算机，数字形式的地址对于计算机来说是方便了，但对于人类来说却既难以记忆又难以输入。

于是“**域名系统**”（Domain Name System）出现了，用有意义的名字来作为IP地址的等价替代。设想一下，你是愿意记“95.211.80.227”这样枯燥的数字，还是“nginx.org”这样的词组呢？

在DNS中，“域名”（Domain Name）又称为“主机名”（Host），为了更好地标记不同国家或组织的主机，让名字更好记，所以被设计成了一个有层次的结构。

域名用“.”分隔成多个单词，级别从左到右逐级升高，最右边的被称为“顶级域名”。对于顶级域名，可能你随口就能说出几个，例如表示商业公司的“com”、表示教育机构的“edu”，表示国家的“cn”“uk”等，买火车票时的域名还记得吗？是“www.12306.cn”。

但想要使用TCP/IP协议来通信仍然要使用IP地址，所以需要把域名做一个转换，“映射”到它的真实IP，这就是所谓的“**域名解析**”。

继续用刚才的打电话做个比喻，你想要打电话给小明，但不知道电话号码，就得在手机里的号码簿里一项一项地找，直到找到小明那一条记录，然后才能查到号码。这里的“小明”就相当于域名，而“电话号码”就相当于IP地址，这个查找的过程就是域名解析。

域名解析的实际操作要比刚才的例子复杂很多，因为互联网上的电脑实在是太多了。目前全世界有13组根DNS服务器，下面再有许多的顶级DNS、权威DNS和更小的本地DNS，逐层递归地实现域名查询。

HTTP协议中并没有明确要求必须使用DNS，但实际上为了方便访问互联网上的Web服务器，通常都会使用DNS来定位或标记主机名，间接地把DNS与HTTP绑在了一起。

URI/URL

有了TCP/IP和DNS，是不是我们就可以任意访问网络上的资源了呢？

还不行，DNS和IP地址只是标记了互联网上的主机，但主机上有那么多文本、图片、页面，到底要找哪一个呢？就像小明管理了一大堆文档，你怎么告诉他是哪个呢？

所以就出现了URI（Uniform Resource Identifier），中文名称是 **统一资源标识符**，使用它就能够唯一地标记互联网上资源。

URI另一个更常用的表现形式是URL（Uniform Resource Locator）， **统一资源定位符**，也就是我们俗称的“网址”，它实际上是URI的一个子集，不过因为这两者几乎是相同的，差异不大，所以通常不会做严格的区分。

我就拿Nginx网站来举例，看一下URI是什么样子的。

```
http://nginx.org/en/download.html
```

你可以看到，URI主要有三个基本的部分构成：

1. 协议名：即访问该资源应当使用的协议，在这里是“http”；
2. 主机名：即互联网上主机的标记，可以是域名或IP地址，在这里是“nginx.org”；
3. 路径：即资源在主机上的位置，使用“/”分隔多级目录，在这里是“/en/download.html”。

还是用打电话来做比喻，你通过电话簿找到了小明，让他把昨天做好的宣传文案快递过来。那么这个过程中你就完成了一次URI资源访问，“小明”就是“主机名”，“昨天做好的宣传文案”就是“路径”，而“快递”，就是你要访问这个资源的“协议名”。

HTTPS

在TCP/IP、DNS和URI的“加持”之下，HTTP协议终于可以自由地穿梭在互联网世界里，顺利地访问任意的网页了，真的是“好生快活”。

但且慢，互联网上不仅有“美女”，还有很多的“野兽”。

假设你打电话找小明要一份广告创意，很不幸，电话被商业间谍给窃听了，他立刻动用种种手段偷窃了你的快递，就在你还在等包裹的时候，他抢先发布了这份广告，给你的公司造成了无形或有形的损失。

有没有什么办法能够防止这种情况的发生呢？确实有。你可以使用“加密”的方法，比如这样打电话：

> 你：“喂，小明啊，接下来我们改用火星文通话吧。”
> 小明：“好啊好啊，就用火星文吧。”
> 你：“巴拉巴拉巴拉巴拉……”
> 小明：“巴拉巴拉巴拉巴拉……”

如果你和小明说的火星文只有你们两个才懂，那么即使窃听到了这段谈话，他也不会知道你们到底在说什么，也就无从破坏你们的通话过程。

HTTPS就相当于这个比喻中的“火星文”，它的全称是“**HTTP over SSL/TLS**”，也就是运行在SSL/TLS协议上的HTTP。

注意它的名字，这里是SSL/TLS，而不是TCP/IP，它是一个负责加密通信的安全协议，建立在TCP/IP之上，所以也是个可靠的传输协议，可以被用作HTTP的下层。

因为HTTPS相当于“HTTP+SSL/TLS+TCP/IP”，其中的“HTTP”和“TCP/IP”我们都已经明白了，只要再了解一下SSL/TLS，HTTPS也就能够轻松掌握。

SSL的全称是“**S**ecure **S**ocket **L**ayer”，由网景公司发明，当发展到3.0时被标准化，改名为TLS，即“**T**ransport **L**ayer **S**ecurity”，但由于历史的原因还是有很多人称之为SSL/TLS，或者直接简称为SSL。

SSL使用了许多密码学最先进的研究成果，综合了对称加密、非对称加密、摘要算法、数字签名、数字证书等技术，能够在不安全的环境中为通信的双方创建出一个秘密的、安全的传输通道，为HTTP套上一副坚固的盔甲。

你可以在今后上网时留心看一下浏览器地址栏，如果有一个小锁头标志，那就表明网站启用了安全的HTTPS协议，而URI里的协议名，也从“http”变成了“https”。

代理

代理（Proxy）是HTTP协议中请求方和应答方中间的一个环节，作为“中转站”，既可以转发客户端的请求，也可以转发服务器的应答。

代理有很多的种类，常见的有：

1. 匿名代理：完全“隐匿”了被代理的机器，外界看到的只是代理服务器；
2. 透明代理：顾名思义，它在传输过程中是“透明开放”的，外界既知道代理，也知道客户端；
3. 正向代理：靠近客户端，代表客户端向服务器发送请求；
4. 反向代理：靠近服务器端，代表服务器响应客户端的请求；

上一讲提到的CDN，实际上就是一种代理，它代替源站服务器响应客户端的请求，通常扮演着透明代理和反

向代理的角色。

由于代理在传输过程中插入了一个“中间层”，所以可以在这个环节做很多有意思的事情，比如：

1. 负载均衡：把访问请求均匀分散到多台机器，实现访问集群化；
2. 内容缓存：暂存上下行的数据，减轻后端的压力；
3. 安全防护：隐匿IP,使用WAF等工具抵御网络攻击，保护被代理的机器；
4. 数据处理：提供压缩、加密等额外的功能。

关于HTTP的代理还有一个特殊的“代理协议”（proxy protocol），它由知名的代理软件HAProxy制订，但并不是RFC标准，我也会在之后的课程里专门讲解。

小结

这次我介绍了与HTTP相关的各种协议，在这里简单小结一下今天的内容。

1. TCP/IP是网络世界最常用的协议，HTTP通常运行在TCP/IP提供的可靠传输基础上；
2. DNS域名是IP地址的等价替代，需要用域名解析实现到IP地址的映射；
3. URI是用来标记互联网上资源的一个名字，由“协议名+主机名+路径”构成，俗称URL；
4. HTTPS相当于“HTTP+SSL/TLS+TCP/IP”，为HTTP套了一个安全的外壳；
5. 代理是HTTP传输过程中的“中转站”，可以实现缓存加速、负载均衡等功能。

经过这两讲的学习，相信你应该对HTTP有了一个比较全面的了解，虽然还不是很深入，但已经为后续的学习扫清了障碍。

课下作业

1. DNS与URI有什么关系？
2. 在讲**代理**时我特意没有举例说明，你能够用引入一个“小强”的角色，通过打电话来比喻一下吗？

欢迎你通过留言分享答案，与我和其他同学一起讨论。如果你觉得有所收获，欢迎你把文章分享给你的朋友。

课外小贴士

01 IP 协议曾有 v1、v2、v3 等早期版本，但因为不够完善而没有对外发布，而 v5 则是仅用于实验室内部研究，也从未公开，所以我们看到的只有 v4 和 v6 两个版本。

02 2011 年 2 月，互联网管理组织 ICANN 正式宣布 IPv4 的地址被“ 用尽 ”。

03 如果使用 UNIX/Linux 操作系统，HTTP 可以运行在本机的 UNIX Domain Socket 上，它是一种进程间通信机制，但也满足 HTTP 对下层的“ 可靠传输 ”要求，所以就成了“HTTP over UNIX Domain Socket”。

精选留言：

- 壹笙☞漂泊 2019-06-05 10:00:31

  课后题：

  1、URI DNS

  DNS 是将域名解析出真实IP地址的系统

  URI 是统一资源标识符，标定了客户端需要访问的资源所处的位置，如果URI中的主机名使用域名，则需要使用DNS来讲域名解析为IP。

  2、打电话给小明，请小明找小王拿一下客户资料。小明处于代理角色。

  内容笔记

  1、四层模型：应用层、传输层、网际层、链接层

  2、IP协议主要解决寻址和路由问题

  3、ipv4，地址是四个用“.”分隔的数字，总数有2^32个，大约42亿个可以分配的地址

  4、ipv6，地址是八个用“:”分隔的数字，总数有2^128个。

  5、TCP协议位于IP协议之上，基于IP协议提供可靠的(数据不丢失)、字节流(数据完整)形式的通信，是HTTP协议得以实现的基础

  6、域名系统：为了更好的标记不同国家或组织的主机，域名被设计成了一个有层次的结构

  7、域名用“.”分隔成多个单词，级别从左到右逐级升高。

  8、域名解析：将域名做一个转换，映射到它的真实IP

  9、URI：统一资源标识符；URL：统一资源定位符

  10、URI主要有三个基本部分构成：协议名、主机名、路径

  11、HTTPS：运行在SSL/TLS协议上的HTTP

  12 、SSL/TLS：建立在TCP/IP之上的负责加密通信的安全协议，是可靠的传输协议，可以被用作HTTP的下层

  13、代理(Proxy)：是HTTP协议中请求方和应答方中间的一个环节。既可以转发客户端的请求，也可以转发服务器的应答。

  14、代理常见种类：匿名台历、透明代理、正向代理、反向代理

  15、代理可以做的事：负载均衡、内容缓存、安全防护、数据处理。 [7赞]

  作者回复2019-06-05 14:16:41

  总结的非常详细，也很准确，鼓掌！

- 一步 2019-06-05 09:41:15
  Http协议不是依赖tcp/ip的拆包和封包吗？Unix domain socket可以做到吗？ [1赞]

  作者回复2019-06-05 11:03:52
  当然可以，如果在Linux上跑Nginx，就可以指定用Unix domain socket。

  关键要理解协议栈，http不强制要求下层必须是tcp。

- 不靠谱 ~ 2019-06-05 07:23:10
  1.URI是相当于网络资源的位置，由协议类型，域名或ip和具体位置构成。
  DNS相当于电话簿，会将解析URI中域名部分解析为具体的ip地址。
  2.代理可以理解为一个中转，a要向b发送消息，实际是先发到代理，由代理发给b。反向由b返回给代理，代理返回给a。 [1赞]

  作者回复2019-06-05 09:37:19
  √

- 一粟 2019-06-05 06:57:58
  小强家钥匙丢了，需要找一家开锁公司开门。于是小强打电话给114，114给小强提供一家有资质的开锁公司，并将电话转接过去。这里的114就是代理。 [1赞]

  作者回复2019-06-05 09:37:01
  √

- zjajxzg 2019-06-05 00:15:55
  1、dns是用来解析uri中的域名部分，将人能够记住的域名解析为计算机能够认识的ip地址，才能让 [1赞]

  作者回复2019-06-05 09:35:25
  说的挺好，写完就更好了。

- 业余草 2019-06-06 11:09:49
  写的很好，期待疯狂更新！

- 小美 2019-06-06 09:05:30
  1. URL 包含了协议+主机名+路径，DNS 会将其中的主机名解析为 IP，进而方便根据 IP 协议进行寻址、路由；
  2. 我们为了更安全的和小明交流，选择通过和小强交流，让其再告诉小明，这是匿名代理，也是正向代理，而如果让小明知道我们的存在则不是匿名代理，是透明代理；小明由于某些原因不能直接响应我们，找了小强来代为响应我们，这是反向代理；
  3. 另外回答一下楼下同学关于 URI 和 URL 区别的疑惑，URI 是 Identifier，即标识符，URL 是 Location，即定位，所以定位只是标识符的一种，打个比方，我们找到小明可以通过其家庭地址（Location）也可以通过名字（比如上课点名）来找到他，所以后者也可以成为 URN。因此 URL 和 URN 都是 URI 的子集。

  作者回复2019-06-06 09:26:34
  说的很好，不过现在urn用的很少，现在的uri基本上就是url，除非写论文，否则不用特意区分。

- Atomic 2019-06-06 07:47:51
  打个比方：我让老婆帮我去楼下超市买瓶水，DNS可以帮她找到楼下超市，URI可以帮她找到水放在超市的具体位置

  作者回复2019-06-06 09:18:26

比喻的好生动，笑。

- 右耳朵猫咪 2019-06-06 07:30:28
  老师，uri和url什么关系？

  作者回复2019-06-06 09:19:05
  现在可以认为uri就是url，以前的区分比较严格，现在没有这个必要了。

- -W.LI- 2019-06-05 23:43:21
  URI为了方便拥有记忆可以采用域名代替IP。
  当用户使用域名访问时，就需要DNS技术找到对应的IP地址。然后找到对应的服务器或者代理。DNS域名解析发生在客户端。服务端接受到的还是用户输入的域名，或者IP。服务器(代理)可开启限制，只采用域名访问。
  小刚替小明找小张，小刚就是正向代理。
  小刚说我就是小张(私下问小张)。反向代理

  作者回复2019-06-06 09:19:34
  说的很好。

- bywuu 2019-06-05 23:22:39
  我的理解，DNS是一个地址，是一个总章一样的地址，就像一本书的目录总则，他是最大的那个目录，但是没有更具体的章节目录。而URI是具体的章节目录，所要找的东西，需要URI才能找到。
  代理：A想打电话给B，但是需要通过小强这个代理，这个小强，就很像是传呼机时代的接线员，A要说的话，必须先说给小强听，之后再由小强把A的话传递给B，之后再由B传给小强，再由小强传回给A。

  作者回复2019-06-06 09:19:52
  √

- Carson 2019-06-05 19:47:55
  Dns负责解析uri中主机名为ip地址，这样才能使用ip协议来完成通信

  在早起电话时代，小强给朋友打电话，要先拨通总机，让总机转接，总机就是代理。

  作者回复2019-06-06 09:24:04
  后一个不太准确，总机是中转的作用，和代理还是不太一样的，代理要能够代替另一方。

- 死后的天空 2019-06-05 17:41:11
  我印象中https的加密，只是对报文头进行加密，在三成和四层之间，如果消息被篡改，将会改变头信息，这样报文到对端的时候会被丢弃，这样的效果就是实现了，文件传输不被篡改，但是还是会被截获。如果记忆没错的话，“火星文”这个比喻，会不会有一点歧义，让不知道同学理解为，https将消息本身也加密了

  作者回复2019-06-06 09:25:37
  https对所有数据都会加密，包括header和body，你的理解不太正确。

- Mavericker 2019-06-05 15:07:08
  老师讲的真好, 真好~

  作者回复2019-06-05 18:03:47
  不敢当，多提意见。

- benying 2019-06-05 14:17:28
  打卡打卡，期待搞定http，谢谢老师

  作者回复2019-06-05 16:55:42
  不客气，还要自己努力。

- 瑞 2019-06-05 11:02:34
  第一个问题: dns是域名解析，uri包含协议，域名，路径，因此dns只是帮忙映射域名这块为ip地址

  第二个问题: 比如我要联系小强，但是我没有小强的电话，而小明有小强的电话，因为我找到了小明，让小明帮忙传达信息，小明联系了小强，在又小明把与小强沟通的信息告诉我

  作者回复2019-06-05 14:12:42
  对，补充一下，uri里不仅可以有域名，也可以直接使用ip地址。

- 无野 2019-06-05 10:19:53
  当我们说“域名解析”的时候，域名已经暗指不是IP地址了吧！？

  作者回复2019-06-05 11:05:40
  域名就是字符串名字，ip地址是数字，你说的对。

- 肥low 2019-06-05 09:49:21
  1 dns是用来解析host uri是在dns基础上进一步对资源进行区分 他们分工不同
  2 小强饿了叫外卖 外卖小哥就是个透明代理

  作者回复2019-06-05 11:05:06
  第一个，只有当uri里使用了域名时才会用到dns，两者不是强关系。

- sakila 2019-06-05 09:45:25
  老师我想问一下，同一个域名下的不同文件可以指定不同的ip吗？比如test.com/a.html指向1.1.1.1，test.com/b.html指向2.2.2.2

  作者回复2019-06-05 11:02:12
  这个是由域名解析来决定的，如果解析出多个ip地址，那么文件就在不同的地址，但无法强制指定。

- Geek_d4dee7 2019-06-05 09:44:10
  老师 cdn 与今天讲的代理是什么关系 都有负载均衡的作用 能再多说一些么

  作者回复2019-06-05 11:02:54
  后面会专门讲，cdn综合了很多技术，有dns、代理、负载均衡。