

## 第37讲 | 知识串讲：用双十一的故事串起碎片的网络协议（上）

2018-08-10 刘超

趣谈网络协议

[进入课程 >](#)



讲述：刘超

时长 09:51 大小 4.52M

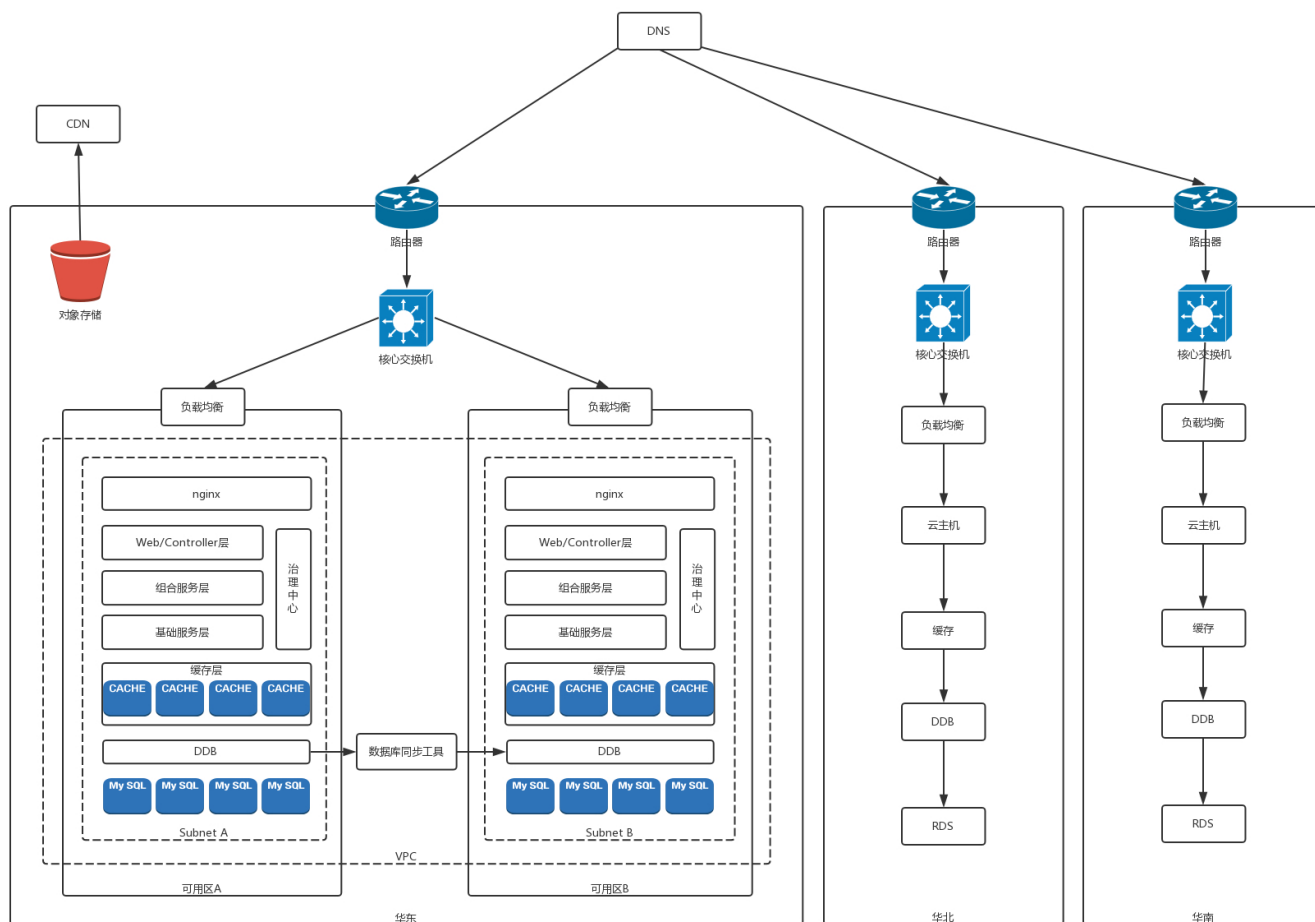


基本的网络知识我们都讲完了，还记得最初举的那个“双十一”下单的例子吗？这一节开始，我们详细地讲解这个过程，用这个过程串起我们讲过的网络协议。

我把这个过程分为十个阶段，从云平台中搭建一个电商开始，到 BGP 路由广播，再到 DNS 域名解析，从客户看商品图片，到最终下单的整个过程，每一步我都会详细讲解。这节我们先来看前三个阶段。

### 1. 部署一个高可用高并发的电商平台

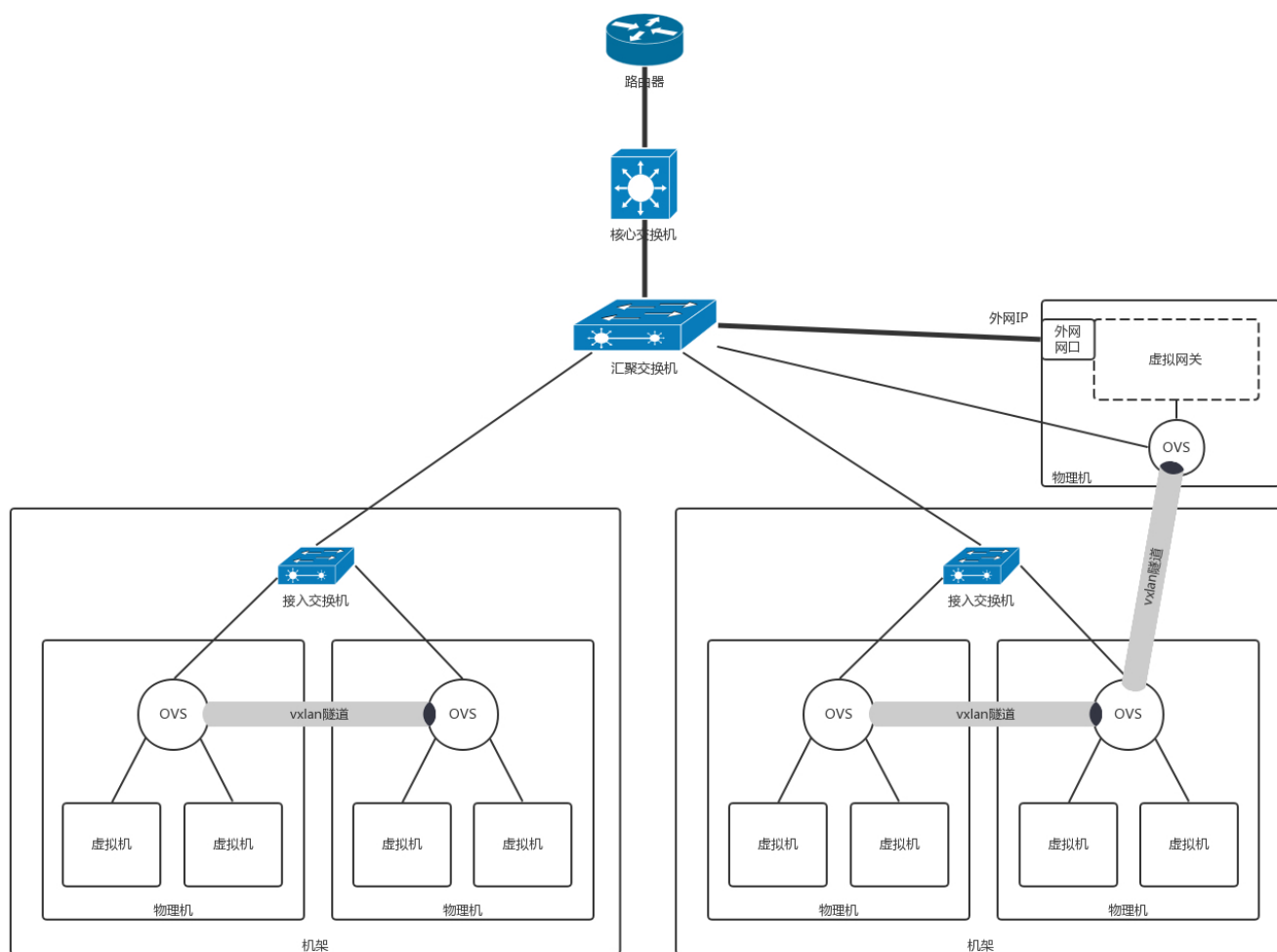
首先，咱们要有个电商平台。假设我们已经有了一个特别大的电商平台，这个平台应该部署在哪里呢？假设我们用公有云，一般公有云会有多个位置，比如在华东、华北、华南都有。毕竟咱们的电商是要服务全国的，当然到处都要部署了。我们把主站点放在华东。



为了每个点都能“雨露均沾”，也为了高可用性，往往需要有多多个机房，形成多个可用区（Available Zone）。由于咱们的应用是分布在两个可用区的，所以假如任何一个可用区挂了，都不会受影响。

我们来回想[数据中心](#)那一节，每个可用区里有一片一片的机柜，每个机柜上有一排一排的服务器，每个机柜都有一个接入交换机，有一个汇聚交换机将多个机柜连在一起。

这些服务器里面部署的都是计算节点，每台上面都有 Open vSwitch 创建的虚拟交换机，将来在这台机器上创建的虚拟机，都会连到 Open vSwitch 上。



接下来，你在云计算的界面上创建一个 **VPC**（Virtual Private Cloud，虚拟私有网络），指定一个 IP 段，这样以后你部署的所有应用都会在这个虚拟网络里，使用你分配的这个 IP 段。为了不同的 VPC 相互隔离，每个 VPC 都会被分配一个 VXLAN 的 ID。尽管不同用户的虚拟机有可能在同一个物理机上，但是不同的 VPC 二层压根儿是不通的。

由于有两个可用区，在这个 VPC 里面，要为每一个可用区分配一个 Subnet，也就是在大的网段里分配两个小的网段。当两个可用区里面网段不同的时候，就可以配置路由策略，访问另外一个可用区，走某一条路由了。

接下来，应该**创建数据库持久化层**。大部分云平台都会提供 PaaS 服务，也就是说，不需要你自己搭建数据库，而是采用直接提供数据库的服务，并且单机房的主备切换都是默认做好的，数据库也是部署在虚拟机里面的，只不过从界面上，你看不到数据库所在的虚拟机而已。

云平台会给每个 Subnet 的数据库实例分配一个域名。创建数据库实例的时候，需要你指定可用区和 Subnet，这样创建出来的数据库实例可以通过这个 Subnet 的私网 IP 进行访问。

为了分库分表实现高并发的读写，在创建的多个数据库实例之上，会**创建一个分布式数据库的实例**，也需要指定可用区和 Subnet，还会为分布式数据库分配一个私网 IP 和域名。

对于数据库这种高可用性比较高的，需要进行跨机房高可用，因而两个可用区都要部署一套，但是只有一个主，另外一个备，云平台往往会提供数据库同步工具，将应用写入主的数据同步给备数据库集群。

接下来是**创建缓存集群**。云平台也会提供 PaaS 服务，也需要每个可用区和 Subnet 创建一套，缓存的数据在内存中，由于读写性能要求高，一般不要求跨可用区读写。

再往上层就是**部署咱们自己写的程序**了。基础服务层、组合服务层、Controller 层，以及 Nginx 层、API 网关等等，这些都是部署在虚拟机里面的。它们之间通过 RPC 相互调用，需要到注册中心进行注册。

它们之间的网络通信是虚拟机和虚拟机之间的。如果是同一台物理机，则那台物理机上的 OVS 就能转发过去；如果是不同的物理机，这台物理机的 OVS 和另一台物理机的 OVS 中间有一个 VXLAN 的隧道，将请求转发过去。

再往外就是**负载均衡**了，负载均衡也是云平台提供的 PaaS 服务，也是属于某个 VPC 的，部署在虚拟机里面的，但是负载均衡有个外网的 IP，这个外网的 IP 地址就是在网关节点的外网网口上的。在网关节点上，会有 NAT 规则，将外网 IP 地址转换为 VPC 里面的私网 IP 地址，通过这些私网 IP 地址访问到虚拟机上的负载均衡节点，然后通过负载均衡节点转发到 API 网关的节点。

网关节点的外网网口是带公网 IP 地址的，里面有一个虚拟网关转发模块，还会有一个 OVS，将私网 IP 地址放到 VXLAN 隧道里面，转发到虚拟机上，从而实现外网和虚拟机网络之间的互通。

不同的可用区之间，通过核心交换机连在一起，核心交换机之外是边界路由器。

在华北、华东、华南同样也部署了一整套，每个地区都创建了 VPC，这就需要有一种机制将 VPC 连接到一起。云平台一般会提供硬件的 VPC 互连的方式，当然也可以使用软件互连的方式，也就是使用 VPN 网关，通过 IPsec VPN 将不同地区的不同 VPC 通过 VPN 连接起来。

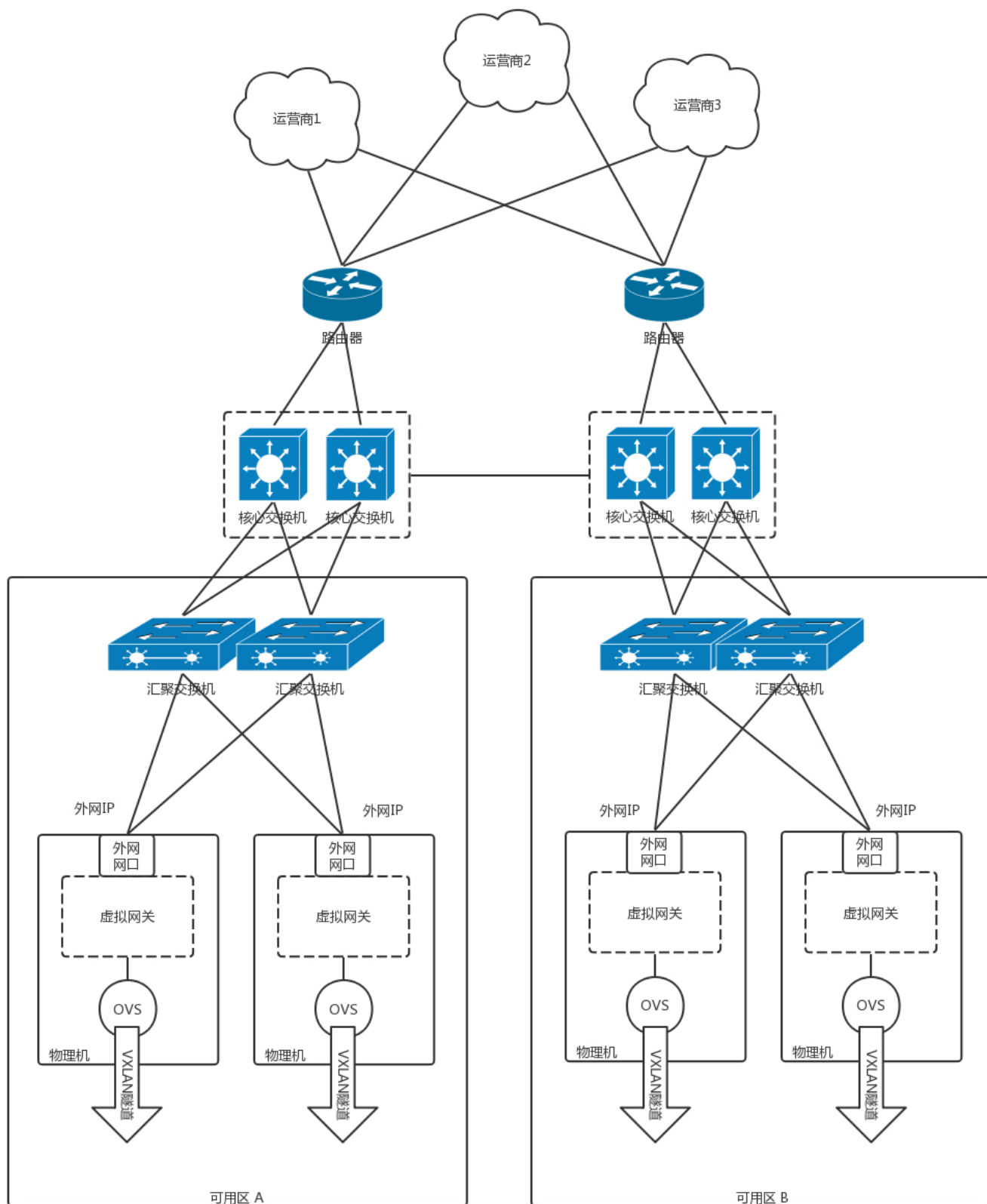
对于不同地区和不同运营商的用户，我们希望他能够就近访问到网站，而且当一个点出了故障之后，我们希望能够在不同的地区之间切换，这就需要有智能 DNS，这个也是云平台提供的。

对于一些静态资源，可以保持在对象存储里面，通过 CDN 下发到边缘节点，这样客户端就能尽快加载出来。

## **2. 大声告诉全世界，可以到我这里买东西**

当电商应用搭建完毕之后，接下来需要将如何访问到这个电商网站广播给全网。

刚才那张图画的是一个可用区的情况，对于多个可用区的情况，我们可以隐去计算节点的情况，将外网访问区域放大。



外网 IP 是放在虚拟网关的外网网口上的，这个 IP 如何让全世界知道呢？当然是通过 BGP 路由协议了。

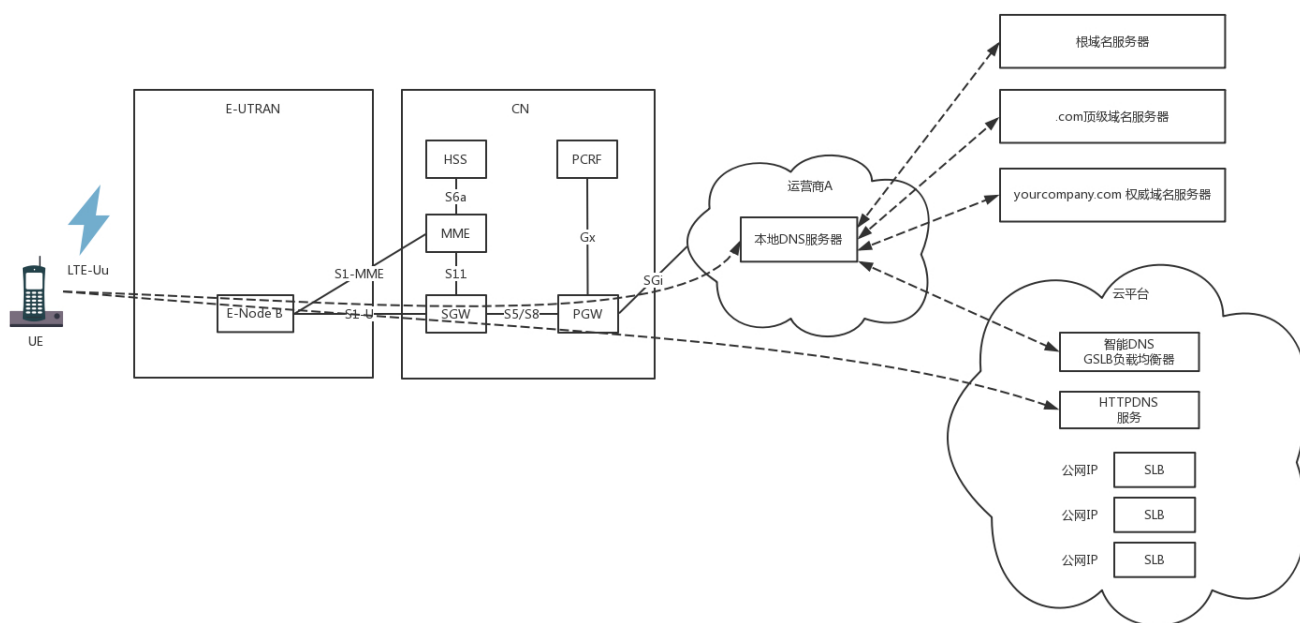
每个可用区都有自己的汇聚交换机，如果机器数目比较多，可以直接用核心交换机，每个 Region 也有自己的核心交换区域。

在核心交换外面是安全设备，然后就是边界路由器。边界路由器会和多个运营商连接，从而每个运营商都能够访问到这个网站。边界路由器可以通过 BGP 协议，将自己数据中心里面的外网 IP 向外广播，也就是告诉全世界，如果要访问这些外网 IP，都来我这里。

每个运营商也有很多的路由器、很多的点，于是就可以将如何到达这些 IP 地址的路由信息，广播到全国乃至全世界。

### 3. 打开手机来上网，域名解析得地址

这个时候，不但你的这个网站的 IP 地址全世界都知道了，你打的广告可能大家也都看到了，于是有客户下载 App 来买东西了。



客户的手机开机以后，在附近寻找基站 eNodeB，发送请求，申请上网。基站将请求发给 MME，MME 对手机进行认证和鉴权，还会请求 HSS 看有没有钱，看看是在哪里上网。

当 MME 通过了手机的认证之后，开始建立隧道，建设的数据通路分两段路，其实是两个隧道。一段是从 eNodeB 到 SGW，第二段是从 SGW 到 PGW，在 PGW 之外，就是互联网。

PGW 会为手机分配一个 IP 地址，手机上网都是带着这个 IP 地址的。

当在手机上面打开一个 App 的时候，首先要做的事情就是解析这个网站的域名。



在手机运营商所在的互联网区域里，有一个本地的 DNS，手机会向这个 DNS 请求解析 DNS。当这个 DNS 本地有缓存，则直接返回；如果没有缓存，本地 DNS 才需要递归地从根 DNS 服务器，查到.com 的顶级域名服务器，最终查到权威 DNS 服务器。

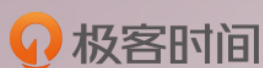
如果你使用云平台的时候，配置了智能 DNS 和全局负载均衡，在权威 DNS 服务中，一般是通过配置 CNAME 的方式，我们可以起一个别名，例如 [vip.yourcompany.com](http://vip.yourcompany.com)，然后告诉本地 DNS 服务器，让它请求 GSLB 解析这个域名，GSLB 就可以在解析这个域名的过程中，通过自己的策略实现负载均衡。

GSLB 通过查看请求它的本地 DNS 服务器所在的运营商和地址，就知道用户所在的运营商和地址，然后将距离用户位置比较近的 Region 里面，三个负载均衡 SLB 的公网 IP 地址，返回给本地 DNS 服务器。本地 DNS 解析器将结果缓存后，返回给客户端。

对于手机 App 来说，可以绕过刚才的传统 DNS 解析机制，直接只要 HTTPDNS 服务，通过直接调用 HTTPDNS 服务器，得到这三个 SLB 的公网 IP 地址。

看，经过了如此复杂的过程，咱们的万里长征还没迈出第一步，刚刚得到 IP 地址，包还没发呢？话说手机 App 拿到了公网 IP 地址，接下来应该做什么呢？

欢迎你留言和我讨论。趣谈网络协议，我们下期见！



# 趣谈网络协议

像小说一样的网络协议入门课

刘超 网易研究院  
云计算技术部首席架构师



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。



上一篇 第36讲 | 跨语言类RPC协议：交流之前，双方先来个专业术语表

下一篇 第38讲 | 知识串讲：用双十一的故事串起碎片的网络协议（中）

## 精选留言 (20)

写留言



固态U盘

2018-08-10

18

意犹未尽！老师再开个其他专栏吧。

展开



zcpromisi...

2018-08-11

12

讲的太精彩了，老师还能不能来个kubernetes相关的专栏



trlllllll

2018-08-23

8

极客时间学过的5个专栏里最好的一个了，赞

展开

作者回复: 谢谢



程启

2018-08-10

3

请教，全球或者国内加海外双十一，调度和数据库层会否有大的变动。

这个系列这么快结束了，真是依依不舍。再开个续集吧 😊

展开





夜空中最亮...  
2018-12-18



精彩

展开 ▾



Enterpriz...  
2018-11-22



云中网络看的我像在云里，终于到能看懂的东西了

展开 ▾

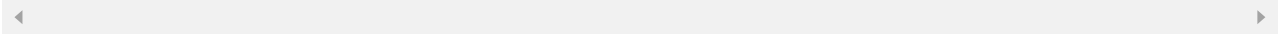


叶  
2018-11-17



老师，我买你前面的课了，那个趣谈协议，我买了，没看到你总结的协议图，不过你讲的生动易懂

作者回复: 有啊，再往后看



凌天  
2018-09-23



老师太博学了

展开 ▾



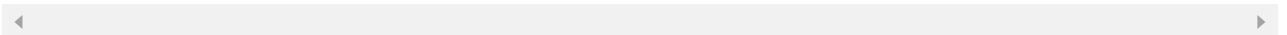
西斯  
2018-08-30



请教个问题 不同机房里的数据库存储的是自己机房的数据 还是包括其他所有机房的数据 如果是全量 跨机房间的数据同步是怎么做的？

展开 ▾

作者回复: 可以全量，就是比较贵，买个存储设备都是有同步机制的，距离近，带宽能保证，就实时同步，不能保证就异步同步。



horsechest...



2018-08-10

这一章开始看不懂了~不过还没工作，等工作了应该在实践中可以得出真知

---



**谢晋**

2019-05-12



喜欢老师文章的通俗、易懂、有趣，表达非常有激情

展开 ▾

---



**大坏狐狸**

2019-03-26



撑到现在 给自己鼓鼓掌

展开 ▾

---



**周平**

2019-01-15



信息量巨大，老师赞

展开 ▾

---



**Hurt**

2018-08-15



老师太赞了 学学学100遍

展开 ▾

---



**hhq**

2018-08-12



把vxlan，vpn和bgp的使用场景都串联起来了，期待下一期的串讲

---



**空档滑行**

2018-08-10



有个问题请教下，我是可以每个虚拟机都绑定一个外网ip的。但是图上一个物理机只分配了一个，如果要多个是怎么实现的

展开 ▾

---



空档滑行

2018-08-10



拿到公网ip后，应该要建立连接了吧。走TCP协议发SYNC

---



blackpikle...

2018-08-10



补充些内容，应该是继续通过 enodeb 接入网络，然后向 CDN 获取静态内容，浏览到动态部分时，向负载均衡发起请求。

展开 ▾

---



赤脚小子

2018-08-10



有ip了要建立连接了，http协议，tcp协议udp协议，长连接，短连接。

---



blackpikle...

2018-08-10



App 应该是先发http 请求进行登陆和身份验证吧，如果是匿名的话，就直接可以请求首页浏览商品了。