

第22讲 | VPN：朝中有人好做官

2018-07-06 刘超

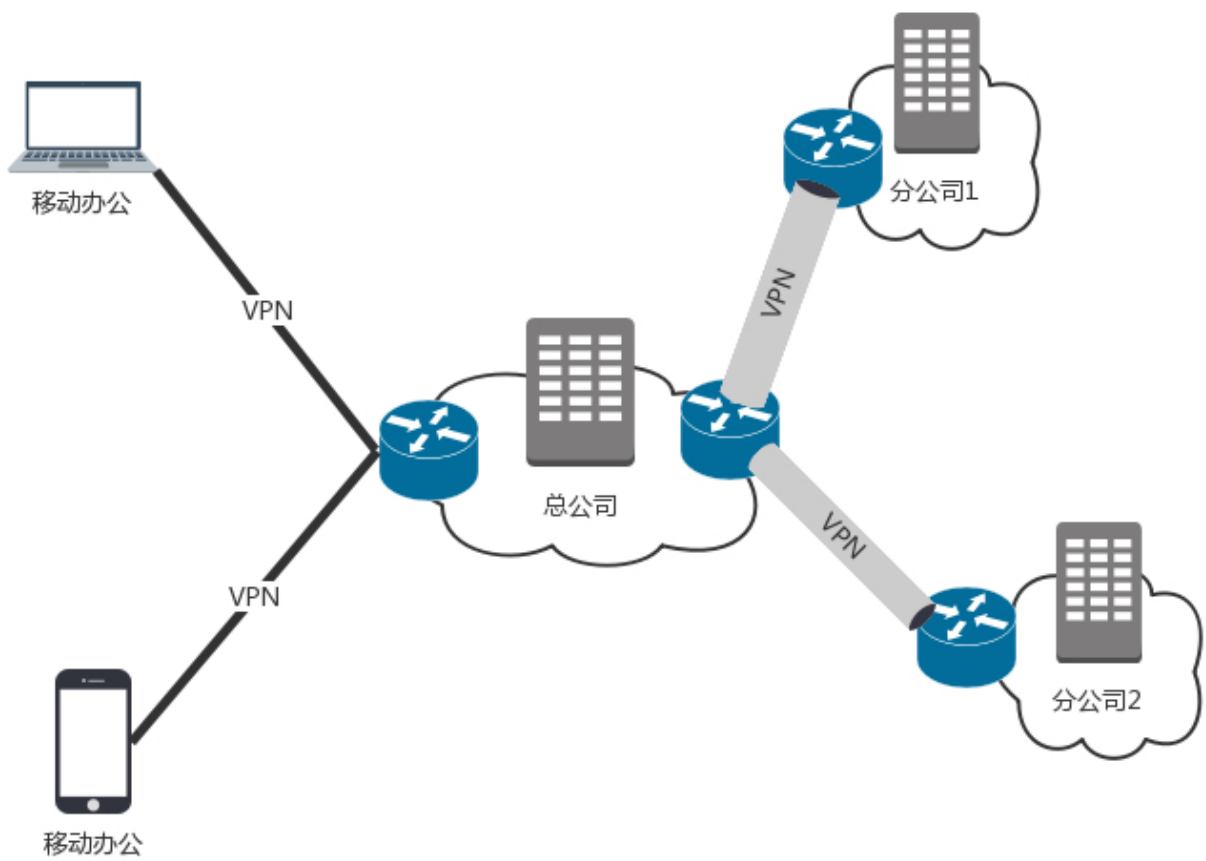


第22讲 | VPN：朝中有人好做官

朗读人：刘超 21'05" | 8.46M

前面我们讲到了数据中心，里面很复杂，但是有的公司有多个数据中心，需要将多个数据中心连接起来，或者需要办公室和数据中心连接起来。这该怎么办呢？

- 第一种方式是走公网，但是公网太不安全，你的隐私可能会被别人偷窥。
- 第二种方式是租用专线的方式把它们连起来，这是土豪的做法，需要花很多钱。
- 第三种方式是用 VPN 来连接，这种方法比较折中，安全又不贵。



VPN，全名Virtual Private Network，虚拟专用网，就是利用开放的公众网络，建立专用数据传输通道，将远程的分支机构、移动办公人员等连接起来。

VPN 是如何工作的？

VPN 通过隧道技术在公众网络上仿真一条点到点的专线，是通过利用一种协议来传输另外一种协议的技术，这里面涉及三种协议：乘客协议、隧道协议和承载协议。

我们以 IPsec 协议为例来说明。



你知道如何通过自驾进行海南游吗？这其中，你的车怎么通过琼州海峡呢？这里用到轮渡，其实这就用到隧道协议。

在广州这边开车是有“协议”的，例如靠右行驶、红灯停、绿灯行，这个就相当于“被封装”的乘客协议。当然在海南那面，开车也是同样的协议。这就相当于需要连接在一起的一个公司的两个分部。

但是在海上坐船航行，也有它的协议，例如要看灯塔、要按航道航行等。这就是外层的承载协议。

那我的车如何从广州到海南呢？这就需要你遵循开车的协议，将车开上轮渡，所有通过轮渡的车都关在船舱里面，按照既定的规则排列好，这就是隧道协议。

在大海上，你的车是关在船舱里面的，就像在隧道里面一样，这个时候内部的乘客协议，也即驾驶协议没啥用处，只需要船遵从外层的承载协议，到达海南就可以了。

到达之后，外部承载协议的任务就结束了，打开船舱，将车开出来，就相当于取下承载协议和隧道协议的头。接下来，在海南该怎么开车，就怎么开车，还是内部的乘客协议起作用。

在最前面的时候说了，直接使用公网太不安全，所以接下来我们来看一种十分安全的 VPN，IPsec VPN。这是基于 IP 协议的安全隧道协议，为了保证在公网上面信息的安全，因而采取了一定的机制保证安全性。

- 机制一：私密性，防止信息泄漏给未经授权的个人，通过加密把数据从明文变成无法读懂的密文，从而确保数据的私密性。

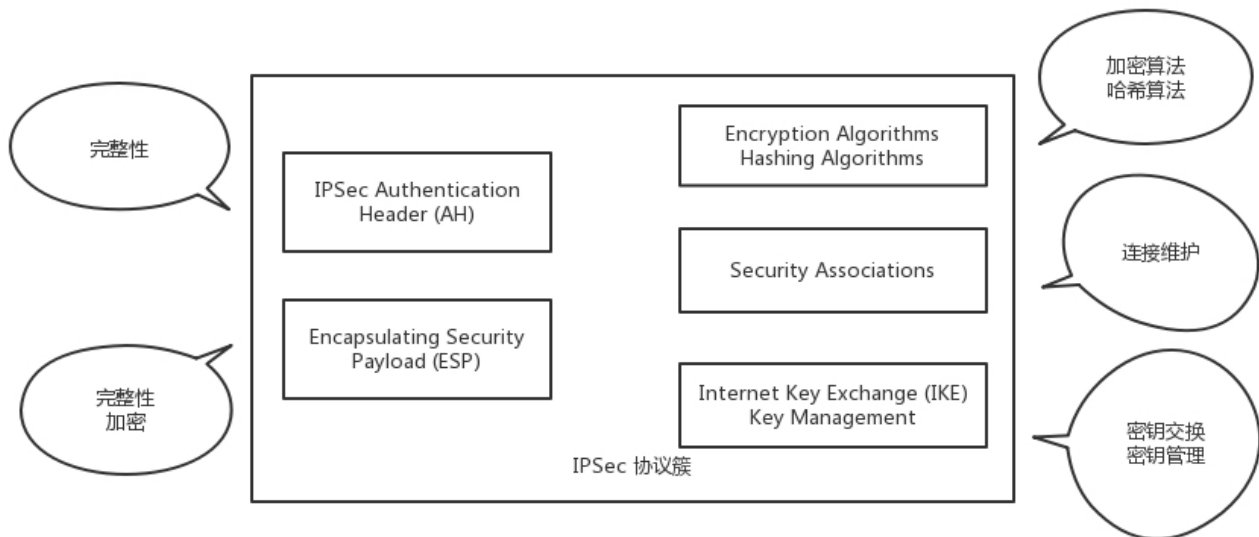
前面讲 HTTPS 的时候，说过加密可以分为对称加密和非对称加密。对称加密速度快一些。而 VPN 一旦建立，需要传输大量数据，因而我们采取对称加密。但是同样，对称加密还是存在加密密钥如何传输的问题，这里需要用到因特网密钥交换（IKE，Internet Key Exchange）协议。

- 机制二：完整性，数据没有被非法篡改，通过对数据进行 hash 运算，产生类似于指纹的数据摘要，以保证数据的完整性。
- 机制三：真实性，数据确实是由特定的对端发出，通过身份认证可以保证数据的真实性。

那如何保证对方就是真正的那个人呢？

- 第一种方法就是预共享密钥，也就是双方事先商量好一个暗号，比如“天王盖地虎，宝塔镇河妖”，对上了，就说明是对的。
- 另外一种方法就是用数字签名来验证。咋签名呢？当然是使用私钥进行签名，私钥只有我自己有，所以如果对方能用我的数字证书里面的公钥解开，就说明我是我。

基于以上三个特性，组成了 IPsec VPN 的协议簇。这个协议簇内容比较丰富。



在这个协议簇里面，有两种协议，这两种协议的区别在于封装网络包的格式不一样。

- 一种协议称为AH（Authentication Header），只能进行数据摘要，不能实现数据加密。
- 还有一种ESP（Encapsulating Security Payload），能够进行数据加密和数据摘要。

在这个协议簇里面，还有两类算法，分别是加密算法和摘要算法。

这个协议簇还包含两大组件，一个用于VPN的双方要进行对称密钥的交换的IKE组件，另一个是VPN的双方要对连接进行维护的SA（Security Association）组件。

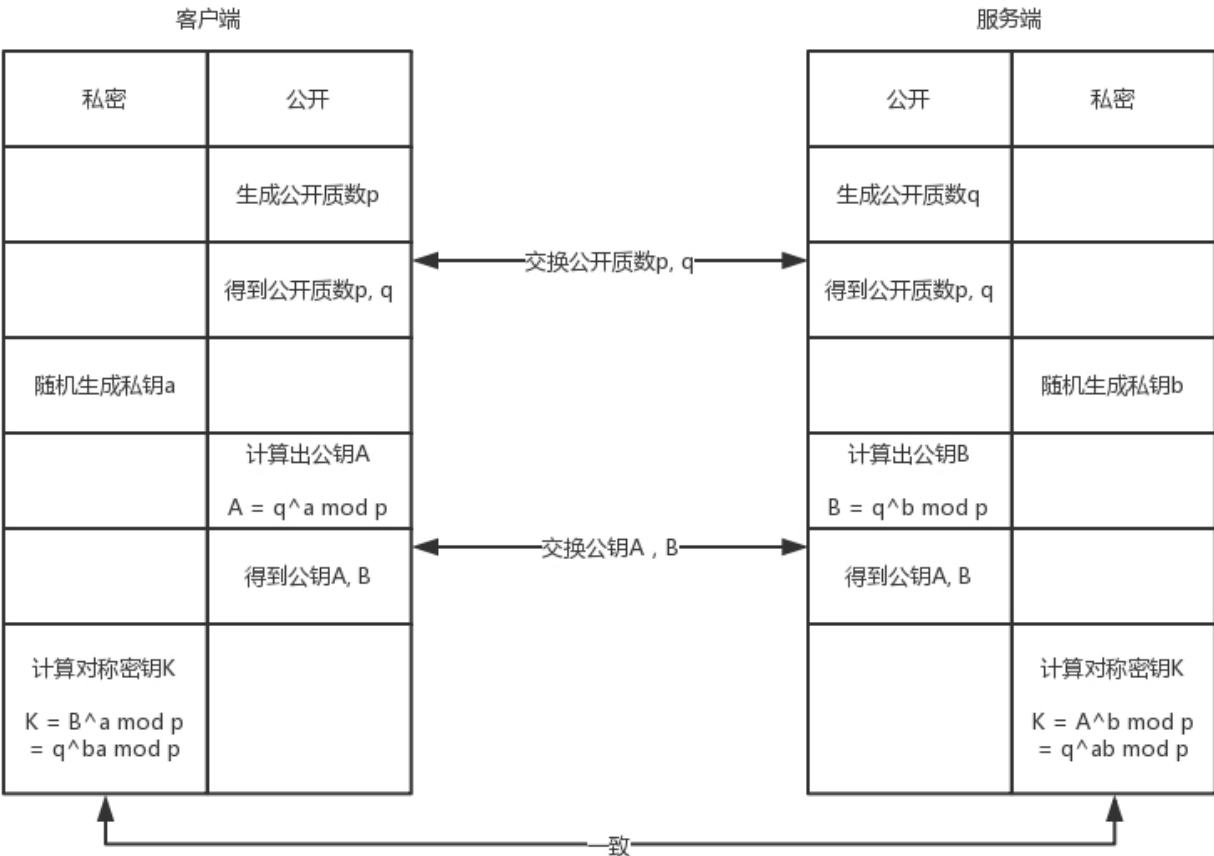
IPsec VPN 的建立过程

下面来看IPsec VPN的建立过程，这个过程分两个阶段。

第一个阶段，建立IKE自己的SA。这个SA用来维护一个通过身份认证和安全保护的通道，为第二个阶段提供服务。在这个阶段，通过DH（Diffie-Hellman）算法计算出一个对称密钥K。

DH算法是一个比较巧妙的算法。客户端和服务端约定两个公开的质数p和q，然后客户端随机产生一个数a作为自己的私钥，服务端随机产生一个b作为自己的私钥，客户端可以根据p、q和a计算出公钥A，服务端根据p、q和b计算出公钥B，然后双方交换公钥A和B。

到此客户端和服务端可以根据已有的信息，各自独立算出相同的结果K，就是对称密钥。但是这个过程，对称密钥从来没有在通道上传输过，只传输了生成密钥的材料，通过这些材料，截获的人是无法算出的。



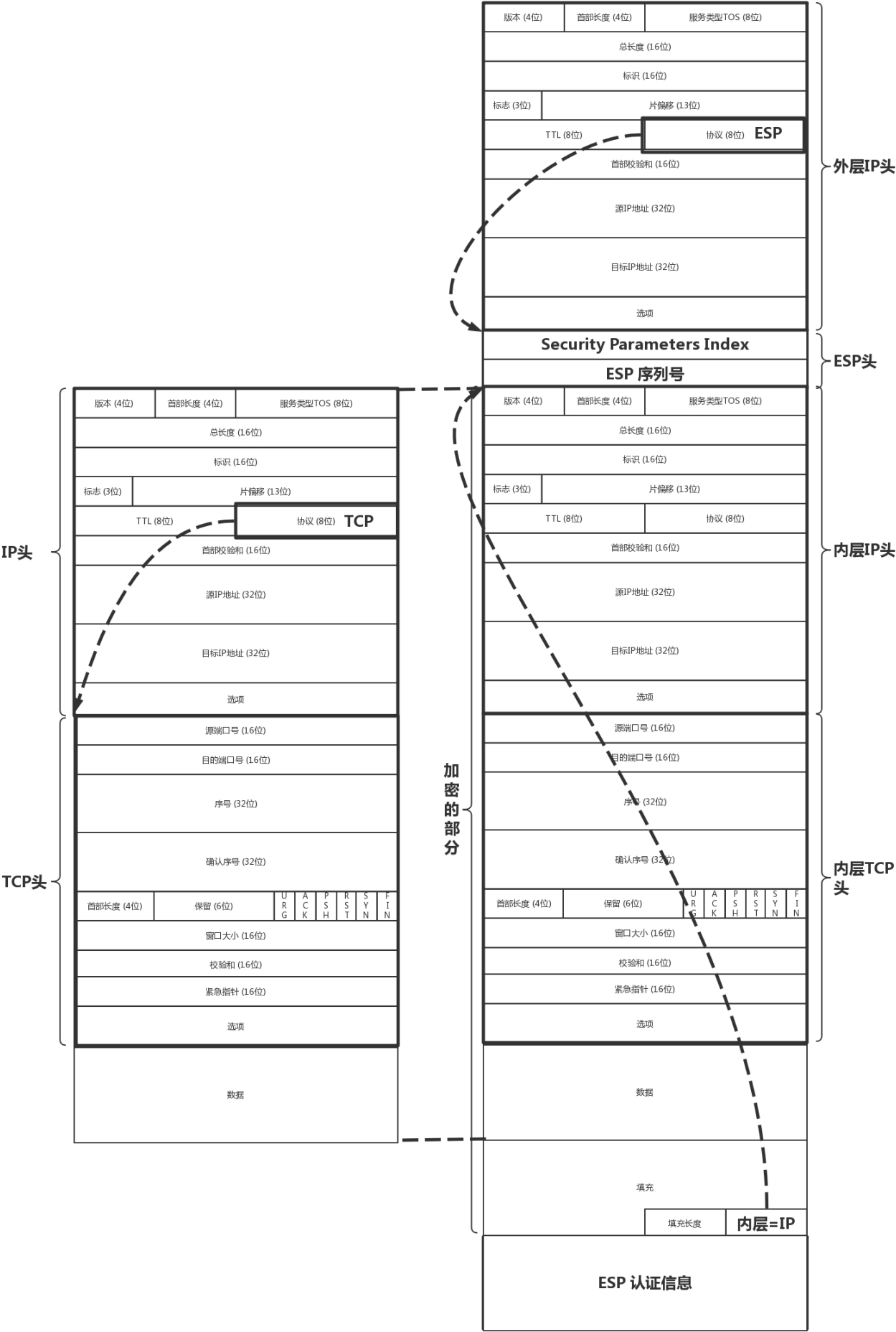
有了这个对称密钥 K，接下来是第二个阶段，建立 IPsec SA。在这个 SA 里面，双方会生成一个随机的对称密钥 M，由 K 加密传给对方，然后使用 M 进行双方接下来通信的数据。对称密钥 M 是有过期时间的，会过一段时间，重新生成一次，从而防止被破解。

IPsec SA 里面有以下内容：

- SPI (Security Parameter Index)，用于标识不同的连接；
- 双方商量好的加密算法、哈希算法和封装模式；
- 生存周期，超过这个周期，就需要重新生成一个 IPsec SA，重新生成对称密钥。



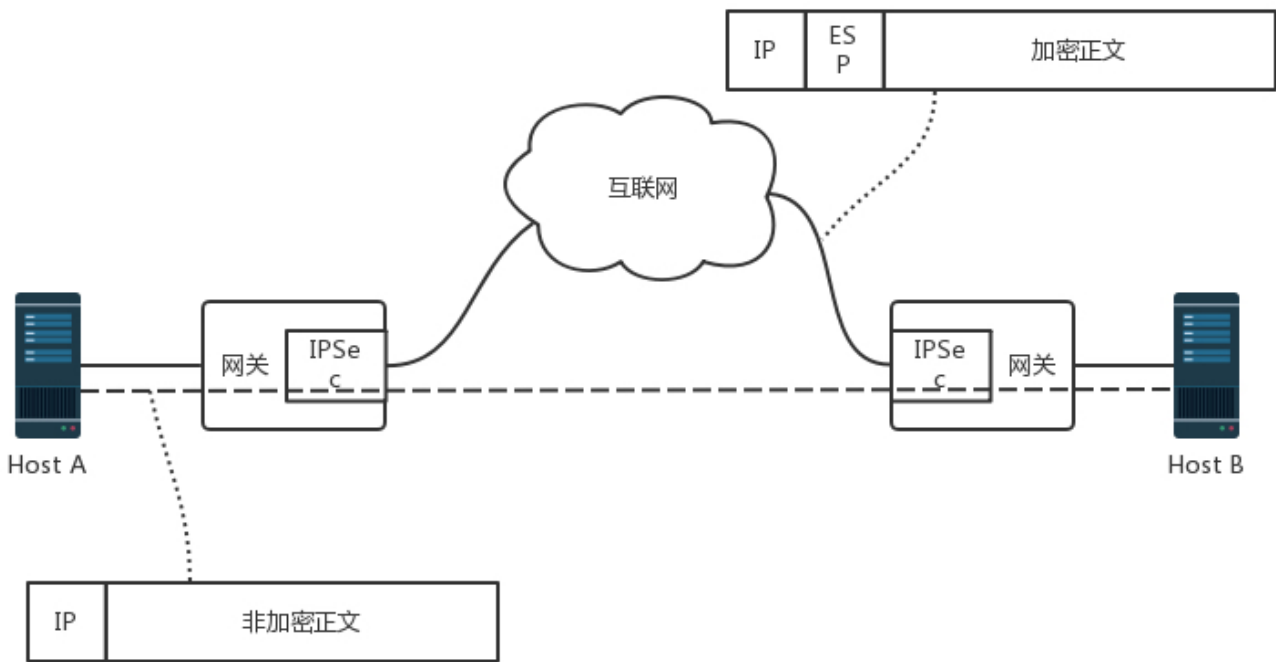
当 IPsec 建立好，接下来就可以开始打包封装传输了。



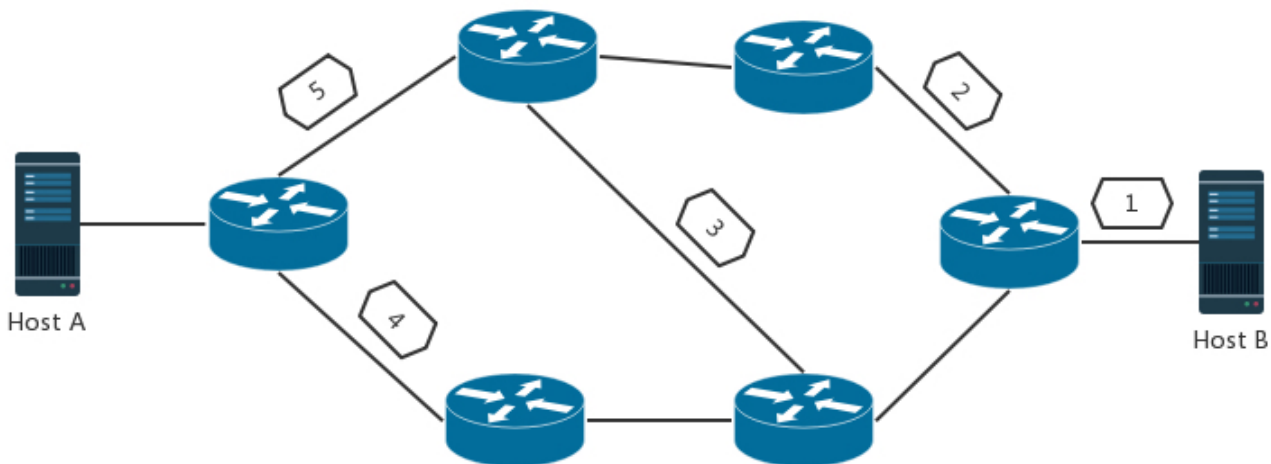
左面是原始的 IP 包，在 IP 头里面，会指定上一层的协议为 TCP。ESP 要对 IP 包进行封装，因而 IP 头里面的上一层协议为 ESP。在 ESP 的正文里面，ESP 的头部有双方商讨好的 SPI，以及这次传输的序列号。

接下来全部是加密的内容。可以通过对称密钥进行解密，解密后在正文的最后，指明了里面的协议是什么。如果是 IP，则需要先解析 IP 头，然后解析 TCP 头，这是从隧道出来后解封装的过程。

有了 IPsec VPN 之后，客户端发送的明文的 IP 包，都会被加上 ESP 头和 IP 头，在公网上传输，由于加密，可以保证不被窃取，到了对端后，去掉 ESP 的头，进行解密。



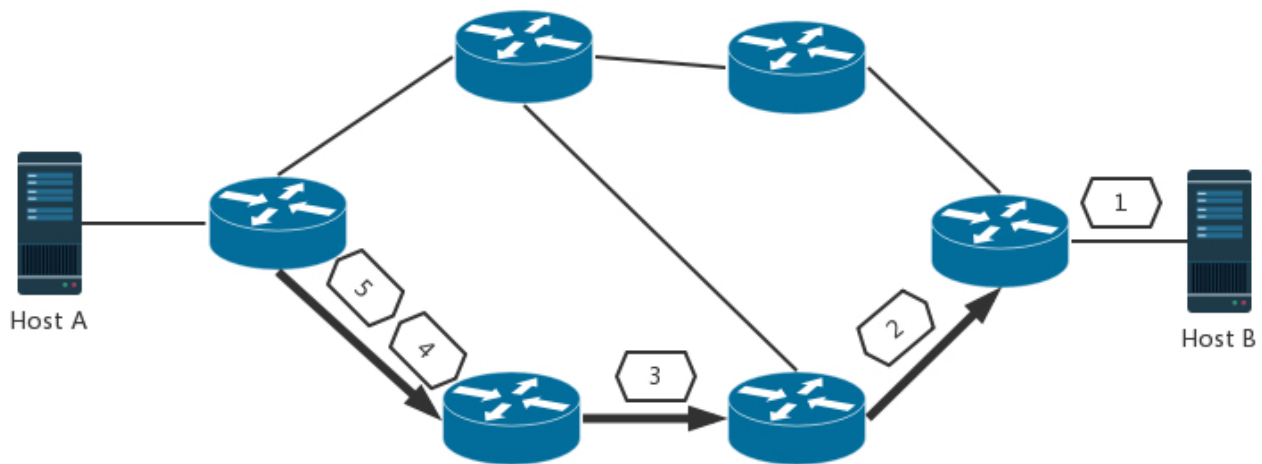
这种点对点的基于 IP 的 VPN，能满足互通的要求，但是速度往往比较慢，这是由底层 IP 协议的特性决定的。IP 不是面向连接的，是尽力而为的协议，每个 IP 包自由选择路径，到每一个路由器，都自己去找下一跳，丢了就丢了，是靠上一层 TCP 的重发来保证可靠性。



因为 IP 网络从设计的时候，就认为是不可靠的，所以即使同一个连接，也可能选择不同的道路，这样的好处是，一条道路崩溃的时候，总有其他的路可以走。当然，带来的代价就是，不断的路由查找，效率比较差。

和 IP 对应的另一种技术称为 ATM。这种协议和 IP 协议的不同在于，它是面向连接的。你可以说 TCP 也是面向连接的啊。这两个不同，ATM 和 IP 是一个层次的，和 TCP 不是一个层次的。

另外，TCP 所谓的面向连接，是不停地重试来保证成功，其实下层的 IP 还是不面向连接的，丢了就丢了。ATM 是传输之前先建立一个连接，形成一个虚拟的通路，一旦连接建立了，所有的包都按照相同的路径走，不会分头行事。

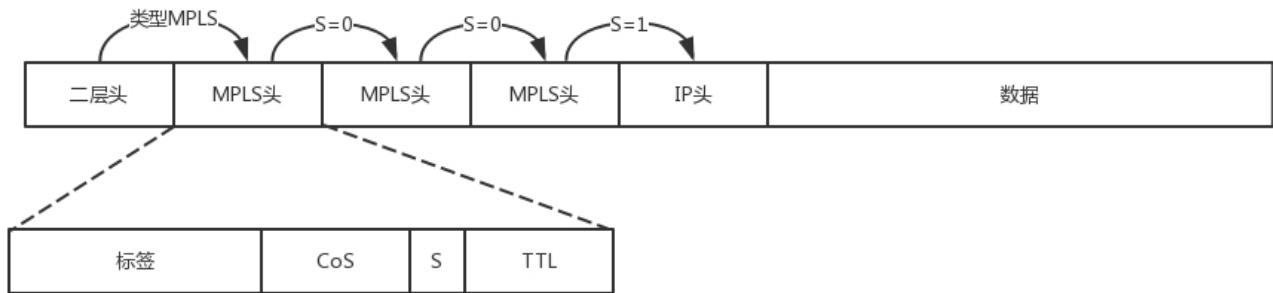


好处是不需要每次都查路由表的，虚拟路径已经建立，打上了标签，后续的包傻傻的跟着走就是了，不用像 IP 包一样，每个包都思考下一步怎么走，都按相同的路径走，这样效率会高很多。

但是一旦虚拟路径上的某个路由器坏了，则这个连接就断了，什么也发不过去了，因为其他的包还会按照原来的路径走，都掉坑里了，它们不会选择其他的路径走。

ATM 技术虽然没有成功，但其摒弃了繁琐的路由查找，改为简单快速的标签交换，将具有全局意义的路由表改为只有本地意义的标签表，这些都可以大大提高一台路由器的转发功力。

有没有一种方式将两者的优点结合起来呢？这就是多协议标签交换（MPLS，Multi-Protocol Label Switching）。MPLS 的格式如图所示，在原始的 IP 头之外，多了 MPLS 的头，里面可以打标签。

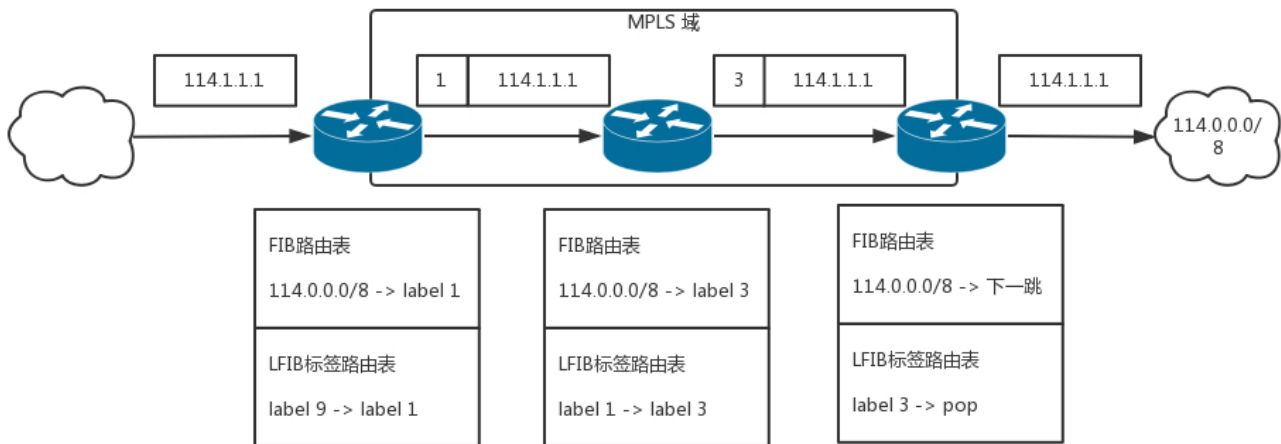


在二层头里面，有类型字段，0x0800 表示 IP，0x8847 表示 MPLS Label。

在 MPLS 头里面，首先是标签值占 20 位，接着是 3 位实验位，再接下来是 1 位栈底标志位，表示当前标签是否位于栈底了。这样就允许多个标签被编码到同一个数据包中，形成标签栈。最后是 8 位 TTL 存活时间字段，如果标签数据包的出发 TTL 值为 0，那么该数据包在网络中的生命期被认为已经过期了。

有了标签，还需要设备认这个标签，并且能够根据这个标签转发，这种能够转发标签的路由器称为标签交换路由器（LSR，Label Switching Router）。

这种路由器会有两个表格，一个就是传统的 FIB，也即路由表，另一个就是 LFIB，标签转发表。有了这两个表，既可以进行普通的路由转发，也可以进行基于标签的转发。



有了标签转发表，转发的过程如图所示，就不用每次都进行普通路由的查找了。

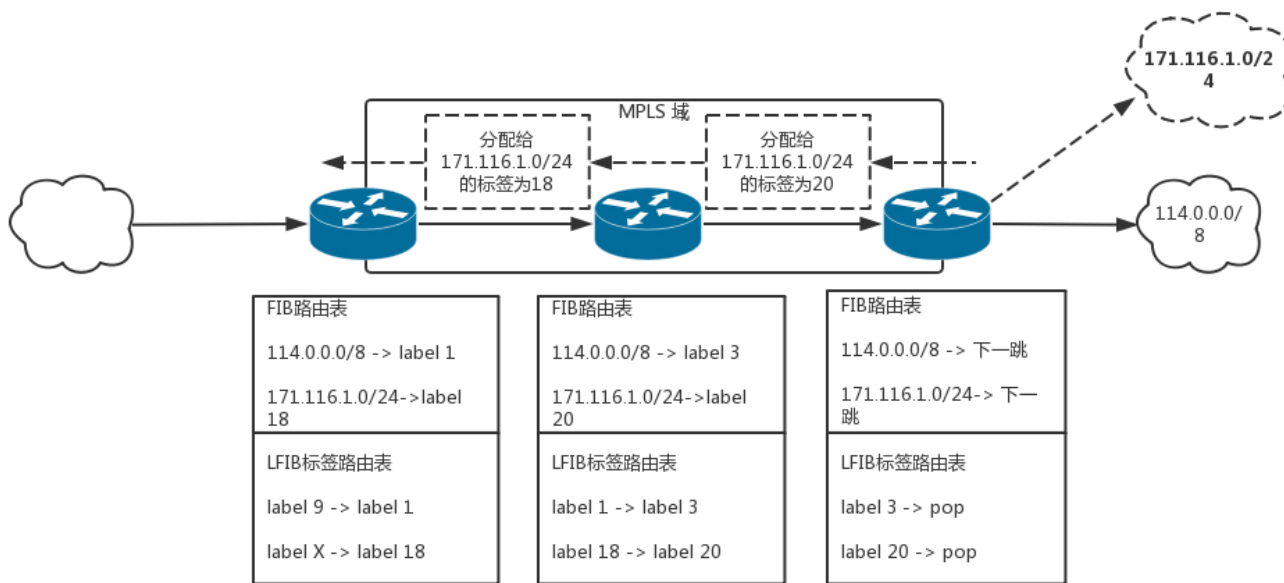
这里我们区分 MPLS 区域和非 MPLS 区域。在 MPLS 区域中间，使用标签进行转发，非 MPLS 区域，使用普通路由转发，在边缘节点上，需要有能力将对于普通路由的转发，变成对于标签的转发。

例如图中要访问 114.1.1.1，在边界上查找普通路由，发现马上要进入 MPLS 区域了，进去了对应标签 1，于是在 IP 头外面加一个标签 1，在区域里面，标签 1 要变成标签 3，标签 3 到达出口边缘，将标签去掉，按照路由发出。

这样一个通过标签转换而建立的路径称为 LSP，标签交换路径。在一条 LSP 上，沿数据包传送的方向，相邻的 LSR 分别叫上游 LSR（upstream LSR）和下游 LSR（downstream LSR）。

有了标签，转发是很简单的事，但是如何生成标签，却是 MPLS 中最难修炼的部分。在 MPLS 秘笈中，这部分被称为 LDP（Label Distribution Protocol），是一个动态的生成标签的协议。

其实 LDP 与 IP 帮派中的路由协议十分相像，通过 LSR 的交互，互相告知去哪里应该打哪个标签，称为标签分发，往往是从下游开始的。



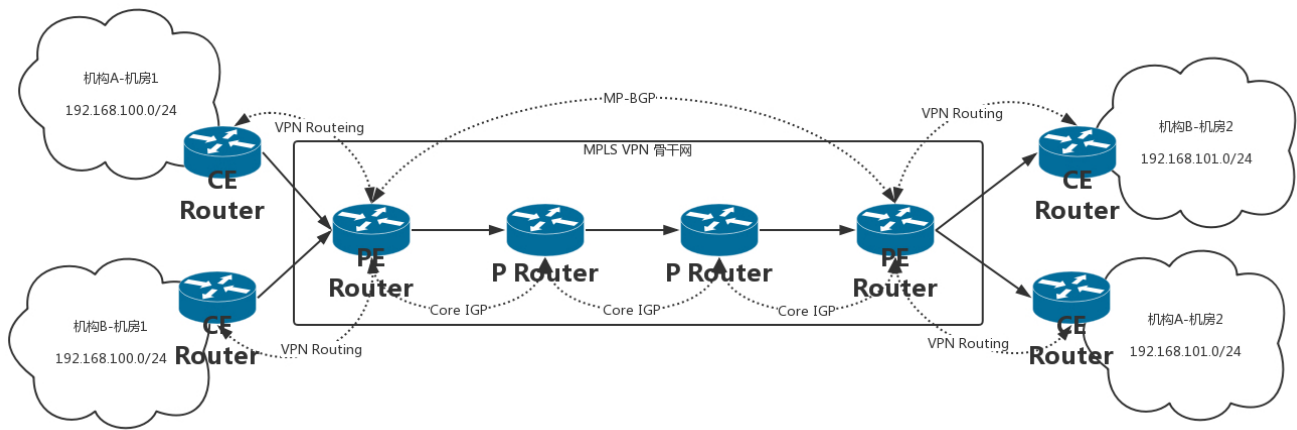
如果有一个边缘节点发现自己的路由表中出现了新的目的地址，它就要给别人说，我能到达一条新的路径了。

如果此边缘节点存在上游 LSR，并且尚有可供分配的标签，则该节点为新的路径分配标签，并向上游发出标签映射消息，其中包含分配的标签等信息。

收到标签映射消息的 LSR 记录相应的标签映射信息，在其标签转发表中增加相应的条目。此 LSR 为它的上游 LSR 分配标签，并继续向上游 LSR 发送标签映射消息。

当入口 LSR 收到标签映射消息时，在标签转发表中增加相应的条目。这时，就完成了 LSP 的建立。有了标签，转发轻松多了，但是这个和 VPN 什么关系呢？

可以想象，如果我们 VPN 通道里面包的转发，都是通过标签的方式进行，效率就会高很多。所以要想个办法把 MPLS 应用于 VPN。



在 MPLS VPN 中，网络中的路由器分成以下几类：

- PE (Provider Edge) ：运营商网络与客户网络相连的边缘网络设备；
- CE (Customer Edge) ：客户网络与 PE 相连接的边缘设备；
- P (Provider) ：这里特指运营商网络中除 PE 之外的其他运营商网络设备。

为什么要这样分呢？因为我们发现，在运营商网络里面，也即 P Router 之间，使用标签是没有问题的，因为都在运营商的管控之下，对于网段，路由都可以自己控制。但是一旦客户要接入这个网络，就复杂得多。

首先是客户地址重复的问题。客户所使用的大多数都是私网的地址 (192.168.X.X;10.X.X.X;172.X.X.X)，而且很多情况下都会与其它的客户重复。

比如，机构 A 和机构 B 都使用了 192.168.101.0/24 网段的地址，这就发生了地址空间重叠 (Overlapping Address Spaces)。

首先困惑的是 BGP 协议，既然 VPN 将两个数据中心连起来，应该看起来像一个数据中心一样，那么如何到达另一端需要通过 BGP 将路由广播过去，传统 BGP 无法正确处理地址空间重叠的 VPN 的路由。

假设机构 A 和机构 B 都使用了 192.168.101.0/24 网段的地址，并各自发布了一条去往此网段的路由，BGP 将只会选择其中一条路由，从而导致去往另一个 VPN 的路由丢失。

所以 PE 路由器之间使用特殊的 MP-BGP 来发布 VPN 路由，在相互沟通的消息中，在一般 32 位 IPv4 的地址之前加上一个客户标示的区分符用于客户地址的区分，这种称为 VPN-IPv4 地址族，这样 PE 路由器会收到如下的消息，机构 A 的 192.168.101.0/24 应该往这面走，机构 B 的 192.168.101.0/24 则应该去另外一个方向。

另外困惑的是路由表，当两个客户的 IP 包到达 PE 的时候，PE 就困惑了，因为网段是重复的。

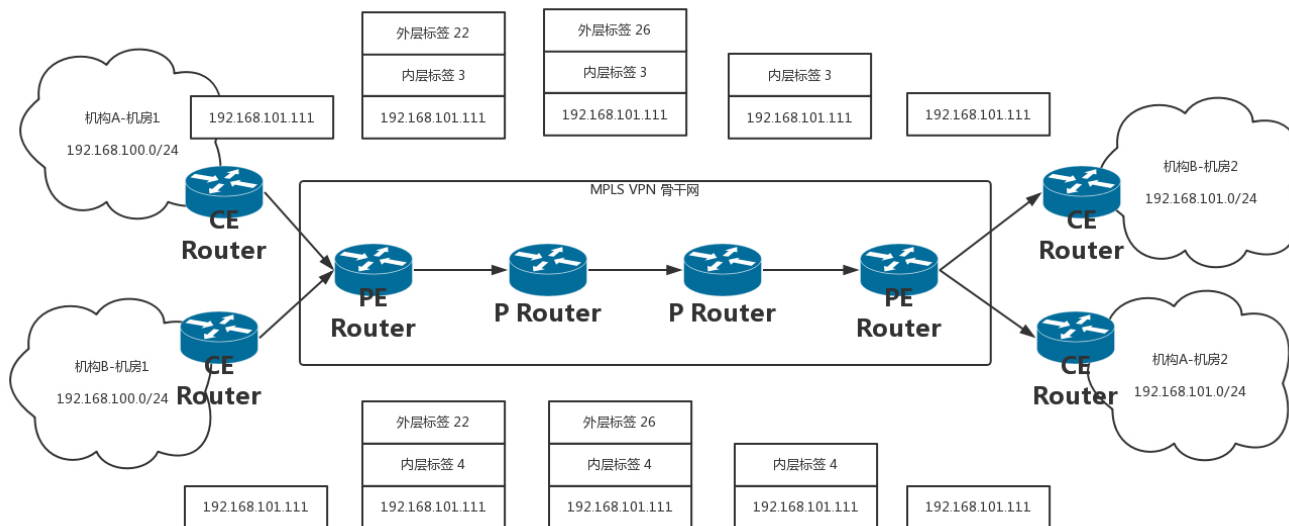
如何区分哪些路由是属于哪些客户 VPN 内的？如何保证 VPN 业务路由与普通路由不相互干扰？

在 PE 上，可以通过 VRF（VPN Routing&Forwarding Instance）建立每个客户一个路由表，与其它 VPN 客户路由和普通路由相互区分。可以理解为专属于客户的小路由器。

远端 PE 通过 MP-BGP 协议把业务路由放到近端 PE，近端 PE 根据不同的客户选择出相关客户的业务路由放到相应的 VRF 路由表中。

VPN 报文转发采用两层标签方式：

- 第一层（外层）标签在骨干网内部进行交换，指示从 PE 到对端 PE 的一条 LSP。VPN 报文利用这层标签，可以沿 LSP 到达对端 PE；
- 第二层（内层）标签在从对端 PE 到达 CE 时使用，在 PE 上，通过查找 VRF 表项，指示报文应被送到哪个 VPN 用户，或者更具体一些，到达哪一个 CE。这样，对端 PE 根据内层标签可以找到转发报文的接口。



我们来举一个例子，看 MPLS VPN 的包发送过程。

1. 机构 A 和机构 B 都发出一个目的地址为 192.168.101.0/24 的 IP 报文，分别由各自的 CE 将报文发送至 PE。
2. PE 会根据报文到达的接口及目的地址查找 VPN 实例表项 VRF，匹配后将报文转发出去，同时打上内层和外层两个标签。假设通过 MP-BGP 配置的路由，两个报文在骨干网走相同的路径。
3. MPLS 网络利用报文的外层标签，将报文传送到出口 PE，报文在到达出口 PE 2 前一跳时已经被剥离外层标签，仅含内层标签。

4. 出口 PE 根据内层标签和目的地址查找 VPN 实例表项 VRF，确定报文的出接口，将报文转发至各自的 CE。
5. CE 根据正常的 IP 转发过程将报文传送到目的地。

小结

好了，这一节就到这里了，我们来总结一下：

- VPN 可以将一个机构的多个数据中心通过隧道的方式连接起来，让机构感觉在一个数据中心里面，就像自驾游通过琼州海峡一样；
- 完全基于软件的 IPsec VPN 可以保证私密性、完整性、真实性、简单便宜，但是性能稍微差一些；
- MPLS-VPN 综合和 IP 转发模式和 ATM 的标签转发模式的优势，性能较好，但是需要从运营商购买。

接下来，给你留两个思考题：

1. 当前业务的高可用性和弹性伸缩很重要，所以很多机构都会在自建私有云之外，采购公有云，你知道私有云和公有云应该如何打通吗？
2. 前面所有的上网行为，都是基于电脑的，但是移动互联网越来越成为核心，你知道手机上网都需要哪些协议吗？

我们的专栏更新到第 22 讲，不知你掌握得如何？每节课后我留的思考题，你都有没有认真思考，并在留言区写下答案呢？我会从已发布的文章中选出一批认真留言的同学，赠送**学习奖励礼券**和我整理的**独家网络协议知识图谱**。

欢迎你留言和我讨论。趣谈网络协议，我们下期见！



版权归极客邦科技所有，未经许可不得转载

通过留言可与作者互动