网络路由技术白皮书

摘要

本文详细介绍了单播路由技术，包括各种路由协议、路由策略等，并探讨了路由技术的发展方向。

关键词

路由协议，路由策略，MPLS，VPN

1.概述

在基于TCP/IP协议栈的网络中，路由协议的目的就是保障其可达性和连通性。目前的路由协议其实都是基于一个假定：整个网络中的地址范围是按照层次合理分布的，所有的数据流流向都是由IP地址指定。IP地址的分配应该类似于现实世界中城市的地址分配，城市由很多街道组成，在每个街道上对具体的每幢建筑分配门牌号码。在网络的世界中网段（或者称为子网）等同与城市中的街道，连在网络上的各种主机和网络设备就好比是矗立街道上的建筑物，IP地址事实上就起了门牌号码的作用。在城市里要到某个大楼，我们往往是先到达这幢大楼所在的街道，然后根据门牌号码找到这幢大楼。在网络中也一样，要传送数据到某一主机，也是先把数据送到目的主机所在的网段，然后再在这个网段中找到目的主机。很多人根据经验就知道该如何走，但是不可能要求一个初到北京的人就知道怎么去“羊角胡同”。所以就需要很多路标，在每个路口指示方向。在网络世界里，数据的传送也需要在网络的每个分岔点给出路标（专业术语是路由）。路由协议就是用来生成路由，指导IP进行数据报文转发的。

于是动态路由协议应运而生，通过传播、分析、计算、挑选路由，来实现如下功能：

- 路由发现：通过交换信息得到整个网络的路由表
- 路由选择：从多条候选路由中得到最优
- 路由切换：自动检测网络故障，达到“自愈”目的
- 负载分担：把多条路由同时加到路由表，多条线路分担负载

具体实现的方法是给网络上每条路由赋一个数值称为权，以权值大小作为选取路由的标准。每个权值的赋给可以是人为指定，也可以是一个受带宽、延迟等因素影响的函数值。

图1是一个路由协议应用的简单例子。

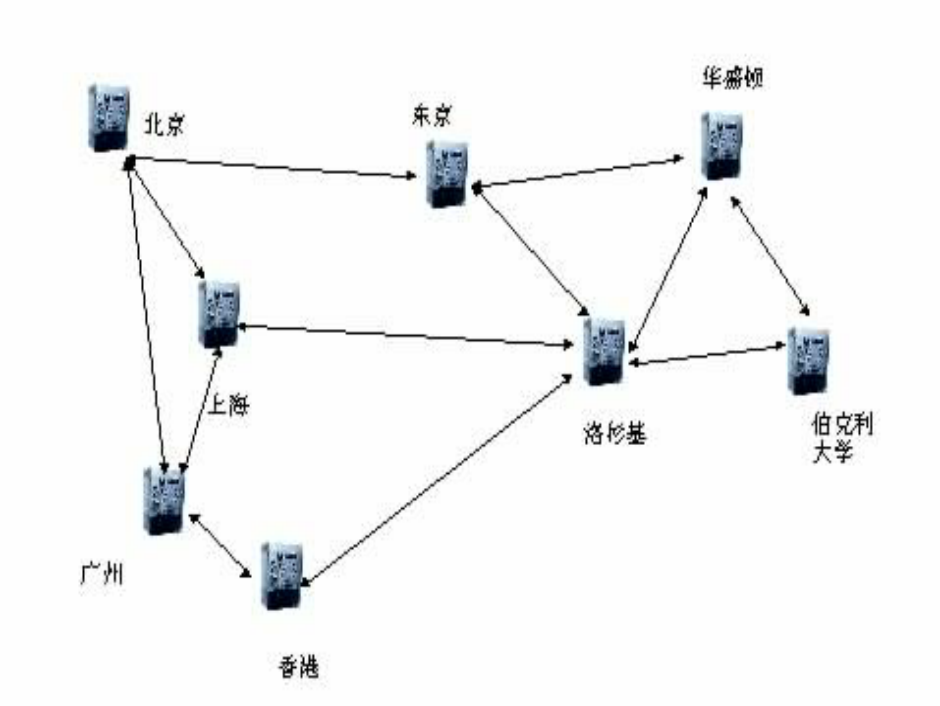


图1　网络实例（有向图）：每条线路均有权值

经过路由协议的处理，可以得到北京到各目的点的路径图，如图2所示。这里是OSPF协议处理的结果，如果采用RIP协议，细节略有不同。

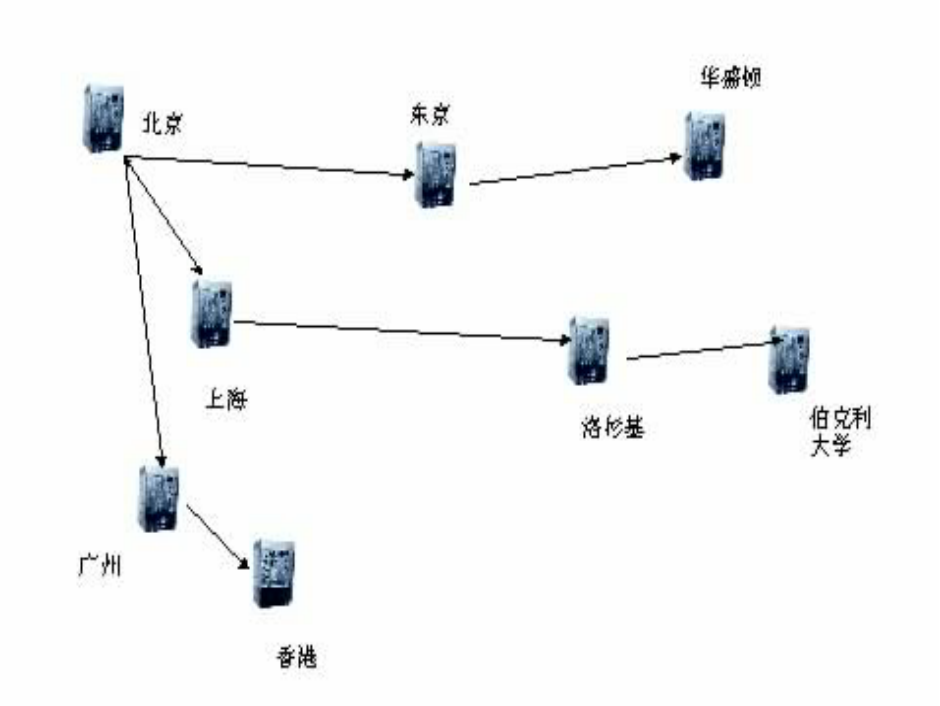


图2 在北京的路由器得到的最优路径图

针对不同的应用情况，现有很多种路由协议（RIP、OSPF、BGP、DVMRP、IS-IS等等），而且随着计算机网络的发展，肯定还会出现很多新的路由协议。每个路由协议的核心功能是根据其所知道网络拓扑结构信息和网络的各种参数指标信息（一般而言，运行该寻径协议的路由器所获得的只是一些局部的信息），寻找出最优的路径，生成路由，改变IP的路由表，从而指导数据包的转发。只不过每种协议的交换网络信息的方式和内容、判断最优的方法和标准不同罢了。

目前的路由协议很多，都是建立在对网络拓扑及网络规模的不同假定前提下发展起来的，不同的协议适合不同的应用场合，这就需要网络管理员对路由协议有一个全面的了解，根据实际情况选择合适的路由协议。

下面针对不同的分类依据，对各个路由协议分类。

1. 根据数据流的类型，可分成单播路由协议（Unicast Routing Protocol）和多播路由协议（Multicast Routing Protocol）。单播路由协议包括RIP、OSPF、BGP、IS-IS等。多播路由协议包括DVMRP、PIM-SM、PIM-DM等。

2. 根据网络规模大小，单播路由协议可分域内路由协议（IGP）和域外路由协议（EGP）。其中IGP有RIP、OSPF、IS-IS等。EGP目前只流行BGP。

3. 根据寻径算法，单播路由协议可分成距离矢量协议（Distance-Vctor）和链接状态协议（Link-State）。距离矢量协议包括RIP、BGP等。其中BGP是距离矢量协议变种，它是一种路径矢量协议。链接状态协议包括OSPF、IS-IS。

从IGP和EGP的定义，顾名思义EGP适用于大型网络，自治区域与自治区域之间。IGP适用于小型网络，自治区域内部。在这强调一个概念：自治系统（Autonomous System）。自治系统指在一个管理机构维护下的路由器集合，这些路由器之间使用一个IGP路由协议和统一的度量，用EGP路由协议来计算与其他自治系统的路由。随着协议的发展，自治系统内部并不局限与使用一个IGP，可以是几个不同的IGP，只要这个系统对外呈现出一种统一的路由机制。

同是IGP之间，也有区别。一般而言RIP适合于小型网络或网络拓扑结构简单网络，而OSPF由于引入了层次结构，适合于中型网络或网络拓扑复杂的网络，BGP适用骨干网络、ISP和大型企业网。

距离矢量协议和链接状态协议的主要区别在于它们传送的内容。距离矢量协议直接传送各自的路由表，各个路由器根据收到的路由表更新自己的路由表，每个路由器对整个网络拓扑不了解，它们只知道邻近的情况。而链接状态协议传送路由器之间的连接状态。这样每个路由器都知道整个网络拓扑结构，路由根据最短路径算法得出。

距离矢量协议无论是实现还是管理都比较简单，但是它的收敛速度慢，报文量大，占用较多网络开销，并且为避免路由环路得做各种特殊处理。链接状态协议则比较复杂，难管理，但是它收敛快，报文量少，占用较少网络开销，不会出现路由环路。

1 路由技术的设计原则

针对网络存在各种情况，路由技术必须具有如下的设计原则：

- 最优的路径选择
- 健壮稳定
- 低开销
- 快速收敛

1.1 最优的路径选择

路由协议的目的是在网络中寻找最优路径，保证网络的连通性。 各个路由协议都用自己的标准来衡量路由的好坏（有的采用下一跳次数、有的采用带宽、有的采用时延，一般在路由数据中用度量Metric来量化），所以要保证网络路径的最优，需要在不同的网络环境中选择恰当的度量。

1.2 健壮稳定

路由协议作为保障网络连通的信令协议，必须健壮稳定。在网络中会有各种异常情况出现，比如硬件错误、重负载等等。因为路由器位于网络的决策点，当路由器发生错误时，会导致网络行为的不可预测。路由协议必须能长时间承受各种异常情况的发生。

1.3 低开销

路由协议的运行在保证正确性的前提下，必须使开销尽可能的低。这包括对网络节点资源的开销和网络带宽的开销。路由协议在路由器上运行，必须少占用CPU处理时间和其他资源，路由协议之间需要互相传送协议报文以获知网络的拓扑变化，这会占用网络带宽，妨碍正常数据报文的传输，所以需要路由协议在设计时充分考虑对网络带宽的占用情况。比如在低速链路上，可以把周期性的更新报文改成按需发送。

1.4 快速收敛

路由设计时，整个网络必须做到全局路由同步，否则，会出现异常的路由循环或不可达。图3是一个路由不同步的例子。 北京路由器向伯克利大学发数据，上海路由器认为向洛杉基转发，洛杉基认为向东京转发，东京认为向上海转发，于是，形成一个路由循环。

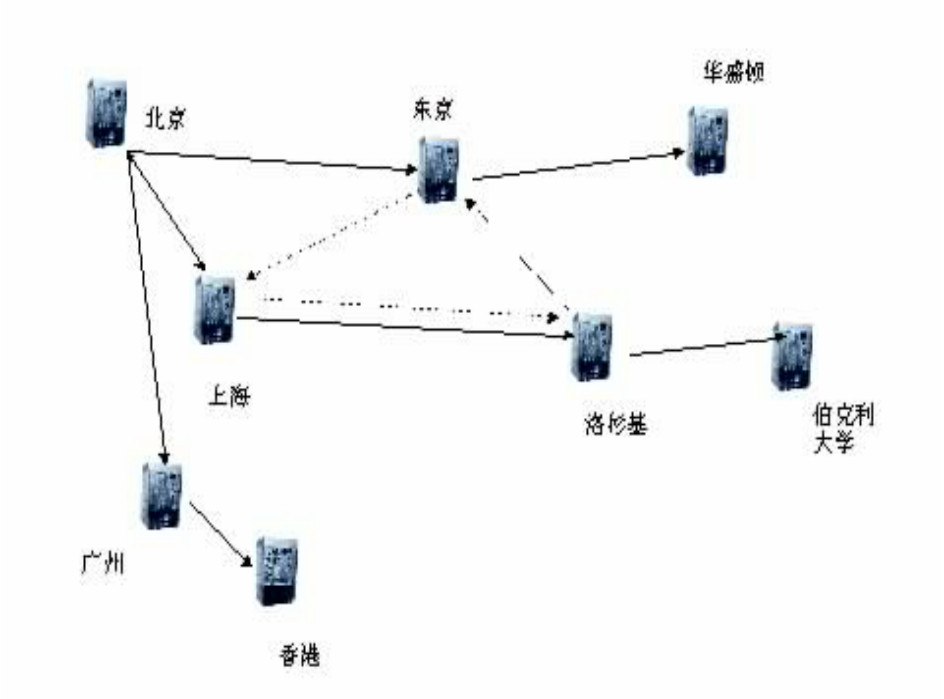


图3 路由不同步造成的结果

在表1所示三种情形下，路由循环/不可达情形可能出现：

可能性	解决办法
不同路由协议间不同步	用同一种路由协议或正确配置路由协议间的策略
指定错误静态路由	使用动态路由协议
网络拓扑变化，路由协议不能及时处理，RIP本身协议的缺陷使其可能出现循环	使用OSPF协议代替RIP，正确配置协议参数

一般来说，通过优化网络配置，能使循环/不可达路由出现的可能性降低到可以忽略的地步。路由的同步问题取决与路由协议的收敛快慢。所以路由协议的收敛特性也是判断该协议是否适用某中网络环境的重要指标。

3网络路由技术介绍

华为公司VRP平台提供一个全面的网络连接解决方案，包括常用的路由协议、灵活丰富的路由策略等等。华为公司VRP平台所采用的路由技术包括：

- RIP

- OSPF
- IS-IS
- BGP
- PIM-SM
- PIM-DM
- 路由策略
- 路由协议的安全性

下面将对各种路由技术作一个简单的介绍。

3.1RIP

RIP协议是内部网关协议（IGP）的一种，是动态路由协议中最早被实现并应用的一种协议。RIP协议的最著名的实现是伯克利分校的4BSD UNIX系统里包含的的一个软件包ROUTED（路由守护神），尽管当时还没有正式的协议，但随着4BSD UNIX系统的流行仍旧是RIP协议成为当时应用最广泛的IGP协议。不过RIP协议自身的弱点使它并不能适应大规模网络的应用，随着互联网的发展新兴的算法更优的OSPF协议逐渐取代RIP协议成为最流行的IGP协议。

RIP协议是基于V-D算法（矢量距离算法）的动态路由协议，该算法最早曾经于1969年在ARPANET上被用作路由计算算法。所谓V-D算法既是将路由信息用一组由目标网络和到达目标网络的开销（用路由权值——METRIC）所组成的矢量数据来表示，将这样的信息在相邻的路由器之间传递，根据自己所获得的信息利用矢量叠加的方式来计算本地的路由信息。每一台路由器都只从自己的邻居那里获得路由信息，在将这些信息连同本地的路由信息传递给其他邻居，这样一跳一跳的传递下去，最终会达到全网的收敛。

通过下面的例子我们可以更清楚的了解基于V-D算法的协议是如何获取和计算路由信息的。如图所示：路由器I与路由器J是相邻的，路由器I 从路由器J 获得到达远端目标网络N 的路由信息（N，M2），其中N 标示目标网络，M2 标示距离长短的Metric值。并且在这条矢量数据上叠加从I 到J 的矢量距离 （J，M1），形成I 上到目标网络N 的路由信息（N，M），其中M=M1+M2 。

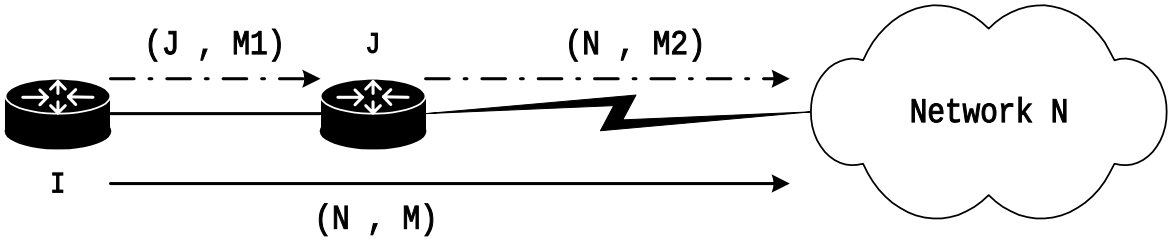


图4 矢量距离算法原理示意图

V-D算法本身存在着一定的缺陷，由于它依赖于从邻居路由器获得路由信息，如果网络中存在环路就很容易导致慢收敛、计算到无穷的问题，在这种情况下，几台路由器之间相互欺骗，路由信息的METRIC逐渐累加，一直到无穷大。

为解决这种问题RIP协议进行了若干改进。首先是设置METRIC的最大值，RIP的METRIC是简单的利用跳数来计算的，而没有考虑时间网络的带宽、负载、延时等诸多条件，为了避免无穷计算，协议规定路由的最大METRIC为15跳，大于15跳表示网络不可达。这种规定限制的RIP的应用范围，它只能适用于中小网络，网络规模太大路由信息就无法到达远端的路由器了。同时，RIP协议在实现中还使用了带毒性逆转的水平分割技术，所谓水平分割是指从某一个邻居获得的路由信息不再向这个邻居发送回去，而毒性逆转则是将这样的路由信息METRIC置为无穷大（大于或等于16）在发送回去，这两种措施都是为了让路由器不收到从自己发送出去的循环路由而产生错误路由。保持法，将不可达的路由信息在路由表中保持一段时间，以尽可能的扩展最坏情况。

RIP协议位于TCP/IP协议栈的最上层，与其它路由协议一样是应用层的协议，它在协议栈中的位置如下图所示。

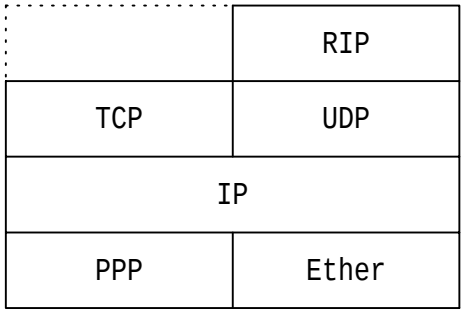


图5 RIP

RIP协议利用UDP协议承载自己的协议报文，由于UDP是无连接的，不保证报文的可靠传输，因此RIP采用周期性发送的方式来保证协议报文被邻居路由器接收。缺省的报文发送间隔是30秒，RIP协议收到周期性的更新报文，会将新的路由添加到路由表中，而将已经获得路由重新刷新一次。与周期性发送相适应的路由的超时处理，对于若干周期没有被刷新的路由就认为其已经不再可达，将它删除。

RIP协议分为RIP-1和RIP-2两个版本，这两个版本的最大区别就在于RIP-2协议支持无分类掩码路由，而RIP-1则不是，同时RIP-2还支持对报文的认证功能，因而具有更高的安全性。由于RIP-1 不支持无分类掩码路由，所以无法正确表示和解析子网路由。路由聚合功能对于RIP-2也是一项重要特性，对路由聚合的合理应用可以大量减少报文中路由的数目，降低对网络带宽的占用，同时可以减小路由表的大小，对于提高效率大有好处。对于属于同一网段的不同子网之间的路由交换则不能使用路由聚合，它们的聚合路由是一样的，聚合路由不能指导报文转发，这时必须交换子网路由。路由聚合对于RIP-1是必选属性，而对于RIP-2 则是一个可选属性。

RIP-2 的另一个特性是支持通过多播传递路由信息，而RIP-1 则只能采用广播或单播方式发送报文。RIP所使用的多播地址是224.0.0.9 。RIP-1 和RIP-2 都支持广播方式，单播方式则是对广播的补充。在实际的网络中，某些物理网络并不支持广播，比如NBMA网络，在这种网络中使用RIP协议就必须使用单播方式（定点发送）传递协议报文。

RIP-2支持两种报文的认证方式，一种的普通的明文认证，另一种是使用MD5加密算法的密文认证。其中MD5算法在RIP-2认证中的应用方式有两种，一种是在RIP-2协议中（RFC1723）中定义的报文格式，而另一种则是针对MD5在RIP-2中的应用所做的专门说明（RFC2082）。

在某些特定的环境下（如拨号接入网关），路由表中存在大量的主机路由，如果这些路由信息都放到报文中会占用大量的网络带宽，VRP软件支持一种对主机路由进行控制的特性，可以强制不将主机路由引入到协议报文中去，当然这一特性仅对RIP-2版本起作用（RIP-1不能正确传递主机路由）。

RIP协议通过报文的周期性发送和路由超时机制来保证路由信息的正确性，这种机制使它很难在按需拨号网络中应用，使用RIP协议会导致链路长时间不能被挂

断，即使再没有数据报文也是如此，增大了网络的开销，也违背了使用按需拨号的初衷。为了能够在按需拨号网络上很好的使用RIP协议，VRP软件增加了对快照功能（snapshot）的支持。快照功能是将RIP的报文交换过程分割成两个阶段，一个相对较短的激活阶段和一个很长时间的静止阶段（如下图所示）。在激活阶段，RIP进行正常的路由信息交换，对当前的路由进行快照保存，而在长时间的静止阶段，则不在这个网络上进行任何的路由信息交换，而先前被快照保存的路由也不进行超时处理，直到下一个激活阶段的到来才有开始正常的报文交换和路由超时。这样就保证了不会长时间占用线路而造成额外的网络使用开销。除了快照功能外，针对拨号网络的应用还产生了对RIP协议新的扩展，RIP的触发式扩展（Triggered RIP），VRP软件中也实现了这种新的扩展协议。

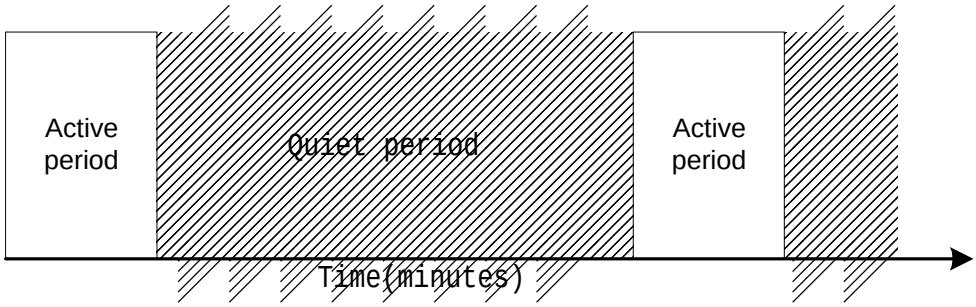


图6 快照功能路由交换过程

RIP协议虽然是最简单的动态路由协议，并且它的适用范围受到了一定程度的限制，但在中小网络的应用中仍然是一个实用的选择。

3．2OSPF

OSPF是Open Shortest Path First（即“开放最短路由优先协议”）的缩写。它是IETF组织开发的一个基于链路状态的自治系统内部路由协议。在IP网络上，它通过收集和传递自治系统的链路状态来动态地发现并传播路由。

每一台运行OSPF协议的路由器总是将本地网络的连接状态，（如可用接口信息、可达邻居信息等）用LSA（链路状态广播）描述，并广播到整个自治系统中去。这样，每台路由器都收到了自治系统中所有路由器生成的LSA，这些LSA的集

合组成了LSDB（链路状态数据库）。由于每一条LSA是对一台路由器周边网络拓扑的描述，则整个LSDB就是对该自治系统网络拓扑的真实反映。

根据LSDB，各路由器运行SPF(最短路径优先)算法。构建一棵以自己为根的最短路径树，这棵树给出了到自治系统中各节点的路由。在图论中，“树”是一种无环路的连接图。所以OSPF计算出的路由天生就是一种无环路的路由。

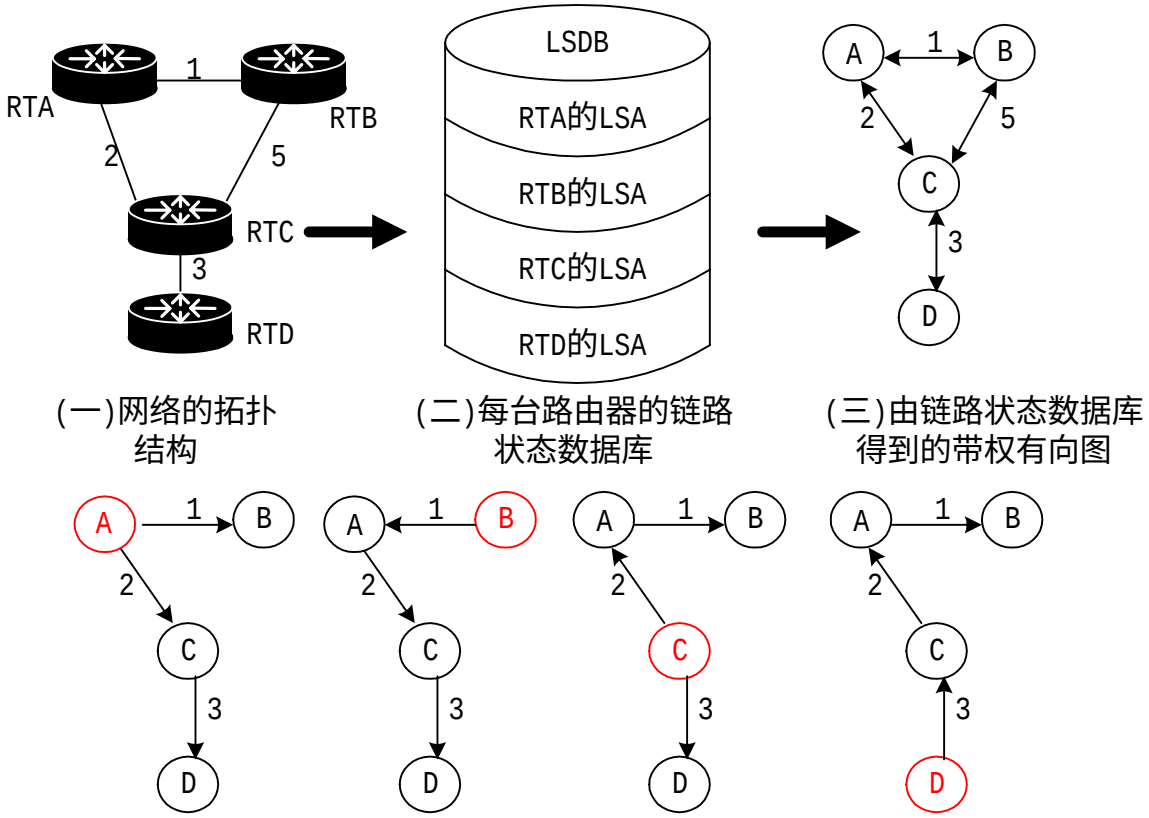


图7 OSPF协议的路由生成过程

上图中描述了通过OSPF协议计算路由的过程。

（一）由四台路由器组成的网络，连线旁边的数字表示从一台路由器到另一台路由器所需要的花费。为简化问题，我们假定两台路由器相互之间发送报文所需花费是相同的。

（二）每台路由器都根据自己周围的网络拓扑结构生成一条LSA（链路状态广播），并通过相互之间发送协议报文将这条LSA发送给网络中其它的所有路由器。这样每台路由器都收到了其它路由器的LSA，所有的LSA放在一起称作LSDB（链路状态数据库）。显然，4台路由器的LSDB都是相同的。

（三）由于一条LSA是对一台路由器周围网络拓扑结构的描述，那么LSDB则是对整个网络的拓扑结构的描述。路由器很容易将LSDB转换成一张带权的有向图，这张图便是对整个网络拓扑结构的真实反映。显然，4台路由器得到的是一张完全相同的图。

（四）接下来每台路由器在图中以自己为根节点，使用相应的算法计算出一棵最小生成树，由这棵树得到了到网络中各个节点的路由表。显然，4台路由器各自得到的路由表是不同的。

这样每台路由器都计算出了到其它路由器的路由。

OSPF协议为了减少自身的开销，提出了以下概念：

（1）DR：

在各类可以多址访问的网络中，如果存在两台或两台以上的路由器，该网络上要选举出一个“指定路由器”(DR)。“指定路由器”负责与本网段内所有路由器进行LSDB的同步。这样，两台非DR路由器之间就不再进行LSDB的同步。大大节省了同一网段内的带宽开销。

（2）AREA：

OSPF可以根据自治系统的拓扑结构划分成不同的区域（AREA），这样区域边界路由器（ABR）向其它区域发送路由信息时，以网段为单位生成摘要LSA。这样可以减少自治系统中的LSA的数量，以及路由计算的复杂度。

OSPF使用4类不同的路由，按优先顺序来说分别是：

- 区域内路由
- 区域间路由
- 第一类外部路由
- 第二类外部路由

区域内和区域间路由描述的是自治系统内部的网络结构，而外部路由则描述了应该如何选择到自治系统以外目的地的路由。一般来说，第一类外部路由对应于OSPF从其它内部路由协议所引入的信息，这些路由的花费和OSPF自身路由的花费具有可比性；第二类外部路由对应于OSPF从外部路由协议所引入的信息，它们的花费远大于OSPF自身的路由花费，因而在计算时，将只考虑外部的花费。

3．3IS-IS

IS-IS 是ISO组织最先提出的网络层选路协议，它是一个链路状态选路协议，采用SPF算法来计算最佳路由，属于IGP协议家族中的一员。它最初的提出是用于支持ISO的“无连接网络协议”（CLNP）分组的路由选择。由于当初ISO的出发点就是想能为今后发展变化的ISO网络制定出一个能支持在多种网络类型及大规模网络间选择路由路径的一个大而同的协议，而且在当初的1992年左右的时候，大家的预见仍是以为合理的未来应该属于OSI，使用IP只是作为“一个过渡策略”，正是在这样的环境下ISO风风火火地很快制定了一个预期支持OSI网络的协议IS-IS协议，协议文档先是在ISO/IEC 10589文档中阐述，后又在RFC 1142中进一步做了说明。IS-IS协议可翻译为“中介系统到中介系统协议”，对应于IP网络，可理解为“路由器到路由器之间的协议”，IS-IS协议将一个自制系统称为一个"路由域"（Routeing Domain）,一个域分为两层结构关系Level-1和Level-2，每个Level-1级组成一个路由子域成为“路由区域”（Routeing Area）,Level-2级的路由器负责区域间的路由，Level-1级的路由器负责区域内的路由，区域内的路由器与主机（在OSI术语中称为ES-终端系统）是通过运行ES-IS协议（End system - Intermeditate system ）互相发现对方的，ES-IS协议是在ISO 9542协议文档中描述的,在IS-IS协议中的选路地址是采用OSI体系中的 NSAP地址结构，与IP网络中的IP地址为4字节长度不一样，它是一个变长结构，长度为8到20字节之间，但在同一个"路由域"内的所有系统应采用等长的地址结构。这一特点也是此协议“闪光”的重要的特点之一。

IETF组织定义了RFC -1195，协议名称为"集成IS-IS"。它 实际上是在最近的网络发展变化中，为满足当前IP网络的飞速发展状况，而在原IS-IS协议文本的基础上作了一定的改动，使IS-IS协议能够既能在ISO的CLNP上跑，又能运行在IP上，利用此协议的健壮性及能满足支持大规模异种网络的优点使其也能支持IP路由功能而形成的一个新的协议，此协议仍利用当初的IS-IS协议的框架结构，只不过在原IS-IS相应的协议报文中添加了一些为支持路由IP报文功能所需的变长域。

为了能支持大规模的路由网络，IS-IS采用了两级结构：level1和level2。将一个大的路由选择域分解成一个或多个局域组成的区域（areas）。在区域内的路由采用level1级路由器来管理，level1级路由器 负责与在同区域内的其它level1路由器及在此区域内的ES通信； 在区域间的路由用level2 路由器来管理， 所有的level2路由器组成内部域的骨干网， 负责在不同区域间通信。每个区域至少有一个路由器同时属于level1和level2，并用来将区域连在主干网上。当一个区域内的NPDU（网络

协议数据单元）要发往另一个区域上的ES（终端系统），level1路由器将首先选择将数据包发到一个距本区域最近的一个level2路由器，而不管此数据包的目的区域是在何处，此数据包然后在level2骨干网上传送到达其目的区域level1路由器，再通过此level1路由器将数据包发到目的ES.

在OSI路由框架中没有任何子网描述，其地址包括的是区域和系统标识符，而不是子网或本地网络。因此路由器必须明确跟踪区域中主机的位置，主机通过“末端系统-中介系统路径选择协议”（ES-IS）向连接在子网上的路由器作自我申明，路由器在IS-IS的链路状态记录中描述这些附属情况，实际上level1路由器中保存的也正是这些信息。所有去往另一个区域的或另一个路由选择域的分组都简单传到最近的“2级路由器”。在主干看来，区域的所有连通点都是等价的，去往某个区域的分组总是传到属于该区域最近的"2级路由器"。

OSI 模型的IS-IS组织结构要求区域能够保持内部连接，如果level1级区域内部的某链路断裂而致使一个区域变得不相通时，可通过采用建立穿越主干的方法来修补。如果区域已经被分为两半，任何连接在这个区域上的2级路由器都可以接收去往该区域的另一半的分组。其采用的办法是将分组封装在发往连接 在另一半区域上 的路由器的分组里，经过主干传送。但值得注意的是，通常情况下，这种“修补”技术不能用来重新连接分裂的主干（level2）。主干(level2)分裂后只能等待，直到连接关系重新建立。特殊情形如：一个单个的主干（level2）路由器在可能与主干网失去连接时，在这种情形下，level2路由器可以在它的level1级链路状态数据记录（LSPs）中指明自己是“不可到达的”（注：levle1级LSP是由level1和level2中介系统共同产生的）。这样的话，有可能允许level1级路由器将报文发向能到达另一目的区域的level2路由器。因此也可总结为level1级路由器能且只能够与这样一些level2级路由器相互通讯，是这些level2路由器在其的level1级的LSP上标志为了“可达”。

3．4OSPF vs IS-IS

OSPF和IS-IS同为域内路由协议，都采用链路状态算法。两者非常相似。事实上是先有IS-IS协议，然后IETF才制定了OSPF协议。不过OSPF一开始就为为IP设计的，但是ISIS是为CLNP设计的，后来增加了对IP的支持。两个协议从邻居关系建立、层次划分、链路状态数据库同步、最短路径算法等都很相似，但是在协议实现细节上有一些不同。

OSPF和IS-IS都采用了层次结构，但是它们在区域Area之间的信息传播有差异。OSPF往一个区域中导入更多的其他区域的信息，所以一个区域内的路由器对于去其他区域的数据可以选择最优的区域边界路由器。而IS-IS协议采用了更严格的层次结构，一个Level 1路由器不知道任何其他Level 1区域的信息，所以要转发报文到其他区域只有先送到最近的Level 2路由器，这时的路由可能不是最优的。所以从选择路径的本领来看，OSPF要更好一些。但是OSPF需要维护更多的其他区域的信息，为了解决减少资源占用，OSPF有一些扩展用来控制信息的发布，如Stub Area和NSSA，在很大程度上解决了这个问题。

IS-IS不支持P-2-MP类型的网络，并且NBMA网络都只能设置为子接口模拟成P-2-P来运行；OSPF可以很好地支持实际中的各种网络类型：Broadcast，NBMA，P-2-P，P-2-MP。

标准的IS-IS 接口cost取值为0到63，对链路层区分不够细致，并且一个网络的metric达到1024就认为不可达；而OSPF接口cost取值范围为0到1024，一个网络的metric达到65535才认为不可达，使用范围要更广一些。后来IEFT在 draf-ietf-isis-traffic-02.txt中扩大了IS-IS的cost的取值范围和最大有效路径metric。

IS-IS虽然在《ISO 10589》中虽然提出了virtual-link来修复分开的骨干区域，但目前厂商基本没有实现，在RFC 1195中也没有提出；OSPF可以很好地支持virtual-link来修复分开的骨干区域或让远端的普通区域连接骨干区。

因为都是典型的链路状态协议，OSPF和IS－IS目前都支持TE（流量工程）的扩展。

在美国很多运营商如UUNET, Sprint, ICM, Digex, Verio, MIBH等目前都在运行IS-IS。这主要是为了能同时支持CLNP和IP两种网络。但是时至今日，运营商的骨干网络只需要支持IP，所以IS-IS比OSPF没有技术上的优势。 IS-IS历史上是为CLNS路由而制定的，发展比较缓慢，对于IP的支持很多地方需要改进，虽然已经提出了draft，但大部分还没有形成RFC，CLNP和IP双环境使用的优势并不明显，是一个不是很成熟的协议； OSPF是专门为IP设计的，更适合IP的路由，发展成熟，标准化程度高，支持厂商多，使用多缺点暴露多，改进也多。 我们推荐在毋需支持CLNP的情况下，应当采用OSPF协议，因为它能选择更好路径，提高网络的吞吐能力。

3．5BGP

BGP(Border Gateway Protocol——边界网关协议)是在自治系统之间传递路由信息的动态路由协议。其最初版本在1989年提出，用于取代以前的EGP(外部网关协议)，发展到1993年开始开发的BGP4，由RFC1771定义，并正在不断地完善着，已经成为事实上的外部路由协议的标准。华为VRP平台支持BGP2、3、4版本，并支持BGP的一系列扩展特性。

BGP没有对网络拓扑做任何限制，这意味着BGP可以方便和灵活地应用于各种复杂的环境之中，尤其是Internet。事实上，Internet并不是从上而下由某个组织建立起来的，而是一些网络自下而上互相连接而成的，每个这样的网络称为一个自治系统(简称AS)。

相对于使用于自治系统内部的路由协议，如OSPF、RIP、IS-IS而言，它是一种“外部”路由协议。BGP基于这样的事实，每个自治系统内部的路由已经由内部路由协议完成了，但是各个自治系统由独立的技术机构所管理，可能采用不同的内部路由协议和选路策略，并且，由于内部路由协议的种种限制，它们无法用于象Internet这样的巨型网络，因此，自治系统之间的路由问题交由BGP来处理。

BGP使用TCP作为其传输层协议，不仅提高了协议的可靠性，而且使得发送增量路由称为可能，这就大大减少了BGP传播路由所占用的带宽，并且适合于在自治系统之间交换大量的路由信息。

BGP支持CIDR（无类别域间选路），通过路由聚合可以有效的抑制因特网上路由的爆炸性增长，减少了Internet上的路由数目。BGP路由携带AS路径信息，也就是它所经过的的自治系统的序列，这样可以彻底解决路由循环问题。

作为一种外部路由协议，与内部路由协议不同，其着眼点不在于发现和计算路由，而在于控制路由的传播和选择最好的路由。由于经济的、政治的等原因，BGP以强制性的原则工作，也就是说，其策略由人工配置，来过滤路由和选择最佳的路由。因此，BGP路由携带了丰富的属性，提供给BGP的路由策略来使用，正是这一特性使得BGP是如此简明而又如此灵活和强大，它还使得BGP便于扩展，以支持因特网新的发展。

正如前面所述的，BGP用在自治系统之间。一般来说，在ISP之间才需要使用BGP，在这时，你要同多个ISP连接，需要在多个相同目的地的路由之间进行选择，并且为客户提供Internet路由。如果你只是ISP的客户，同一个ISP连接，最简单地，可以使用一个默认路由或静态路由来指向ISP，并不需要使用BGP。对于大多

数的局域网和Intranet来说，BGP是一种奢侈品，只有你准备同多个ISP连接或成为一个ISP时，才使用BGP。

两个自治系统边界路由器通过BGP传递路由信息，这两个路由器称为对等体。对等体之间交换多种报文来建立、维持连接和交换路由。在初始的时候，所有的有效路由被交换，之后，当网络信息改变时，只发送增量的更新。这有两种情况，一是路由的属性变化了，BGP重新发送这条路由，来更新它，一是路由不可达或者有了更好的路由，BGP撤销不可达路由，若需要则发送新的路由。如果没有路由发生改变，BGP周期性的交换KEEPALIVE报文来确认对等体之间的连接。

BGP路由携带了属性信息，这些属性有ORIGIN(起源)、AS路径、下一跳、MULTI-EXIT-DISC(多出口辨别符)、本地优先、原子聚合、聚合者、团体、ORIGINATOR_ID(起源者标识符)、群列表、扩展团体等，这些属性表达了路由的特征，它们可以分为过渡和非过渡的，过渡的属性被接收者转发给其它BGP发言者，而非过渡的属性仅被接收者使用，不再转发。属性还分为必遵的和可选的，必遵的属性每个路由都携带，而可选的属性不必要。属性的使用使得BGP非常容易扩展，以适应Internet和路由技术的发展。

一个BGP路由器上，到达某个目的地的路由可以有许多条，这些路由都被BGP保存，并从其中选择一个最优的路由，这种选择基于路由的属性和用户配置的策略，其它的路由作为候选的路由，一旦优选的路由失效，它就可能成为新的最优路由。BGP对等体之间只交换优选的路由。同时，BGP可以在发送和接收时过滤路由和设置路由的属性，以根据用户的需求有目的的选择路由。

BGP路由最初来自于自治系统内部，这需要BGP从内部路由协议引入路由，类似于BGP路由的接收，引入的路由也通过路由过滤和属性的设置。

BGP使用聚合来进一步减少路由的数目，聚合是指将相邻的IP前缀所表示的合并为掩码长度更短的路由，发送综合后的路由信息给对等体。BGP提供功能强大和灵活的聚合功能。

为了解决自治系统内所有的边界路由器需要全连接的问题，路由反射和自治系统联盟两种方案被提出来。为了保证BGP连接的安全性，BGP报文可以被加密。BGP使用路由衰减来来解决路由短时间内反复出现/消失的摆动现象。

BGP的一个典型应用环境如下图：

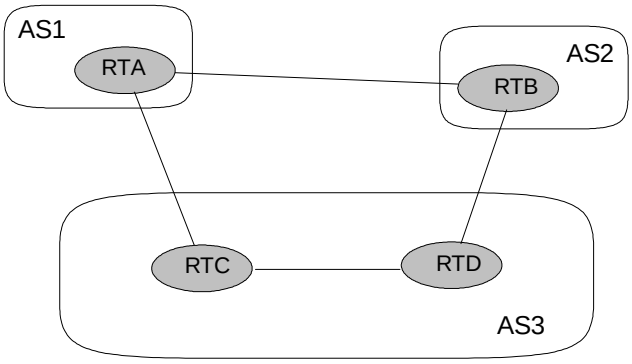


图8

自治系统AS1、AS2都连接到AS3中不同的边界路由器上，同时，它们之间直接相连，AS3内的两个边界路由器也通过BGP连接。路由环路的问题由AS路径属性来解决。到达这三个自治系统的每个目的地，都有两条BGP路由，其AS路径属性不相同，具有较短AS路径的路由被优先选择。

3．6路由策略

3.6.1路由策略概述

路由协议在与邻居路由器进行路由信息交换时，可能需要只接收或发布一部分满足给定条件的路由信息；路由协议在引入其它路由协议路由信息时，可能需要只引入一部分满足条件的路由信息，并对所引入路由信息的某些属性进行设置以使其满足本协议的要求。路由策略提供路由协议实现这些功能的手段。

路由协议的RFC并没有明确定义该协议所应实现的策略功能，因此，协议路由策略的实现并不是协议本身的必须部分，但是，它在管理上提供对协议进行控制的手段，是协议功能性的体现，一般都要求路由协议能够实现一定的策略功能。对于BGP等外部网关协议，提供完善的策略控制手段几乎是必须的。

路由策略由一系列的规则组成，这些规则定义了路由器之间的路由交换和协议间的相互动作，可大体上分成Import、Export、 Redistribute三类。

Import 规则定义了哪些路由是可接收的。其规则包括：接收关于某个自治系统的路由信息、接收关于某些目的网段的路由信息、接收指定路由器发布的路由信息...等等。

Exports 规则定义了哪些路由可被路由协议对外发布，其规则包括类似Import的规则。

Redistribute 规则对路由器内部各个路由协议之间的路由交换进行控制。在缺省状态下，一个路由协议只对外发布由该协议找回的路由信息。比如，RIP只发布由RIP找到的路由信息；BGP发布BGP协议的路由信息。Redistribute 规则控制各个协议之间的路由交换，使合乎条件的外部路由能被内部路由协议发布，反之亦然。

路由策略也常被称为路由过滤，因为定义一条策略规则通常等同于定义一组过滤列表，并在接收、发布一条路由信息或在不同协议间进行路由信息交换前应用这些过滤列表。

3.6.2BGP与路由策略

BGP被设计为可以方便的由用户来干预路由的选择，这主要是通过使用路由策略，利用BGP路由的属性来完成的。路由策略可以用于BGP路由的接收、发送、从其它路由协议的引入和路由聚合等。

除了独立于协议的路由策略，如基于目的地址、下一跳的策略之外，BGP路由属性可是用于策略中，使得BGP的路由策略非常丰富和灵活，为用户进行路由选择提供了最大的可能性。具体来说，BGP能够提供基于AS路径、团体、ORIGIN(起源)、本地优先、MULTI-EXIT-DISC(多出口辨别符)等属性的路由策略。

基于AS路径的策略允许用户过滤经过指定AS路径的路由，这种过滤运用了正则表达式匹配技术，可以灵活的选择AS路径的起始AS、邻居AS、中间AS等等。这种策略还允许用户为AS路径前置若干了自身AS号码，人为的加长这个AS路径，以改变路由选择的优先级。

基于团体属性的策略是BGP策略中功能最为强大的一种，用户可以过滤具有指定团体属性的路由，也可以为路由指定团体属性，这包括增加、删除、附加三种操作。团体属性可以用于区分用户感兴趣的任何路由，从而便于对这些路由施加其它策略。

ORIGIN、本地优先、MULTI-EXIT-DISC等属性被路由策略设置以影响路由选择的优先级。

下面是一个配置实例：

```
router bgp 100
neighbor 172.16.20.1 remote-as 200
```

```
        neighbor 172.16.20.1 route-map test out
        Exit
    route-map out perm 10
        match as-path  1
        set community 100:1 100:2
    Exit
    ip as-path access-list permit _300$
```

上述策略用户AS100中的一个边界路由器为AS200中的对等体172.16.20.1设置出口策略。它匹配从邻居AS300接收的路由，为它们设置两个团体属性。

3.6.3OSPF与路由策略

引入其它路由协议的路由：

路由器上各动态路由协议之间可以互相共享路由信息，由于OSPF的特性，其它的路由协议发现的路由总被当作自治系统外部的路由信息处理。

OSPF可以引入RIP，BGP等动态路由协议发现的路由以及静态路由和接口的直接路由作为自己的外部路由信息。 在引入命令中，可以指定路由的类型（一类外部路由或二类外部路由）、花费值、标记等参数。也可以通过路由映像（route-map）有选择的引入路由。

```
redistribute protocol [metric metric-value][tag tag-value] [route-map map-tag][ type 1 | 2 ]
```

对接收到的路由进行过滤：

- OSPF计算得到的路由在加入到本地路由表之前可以通过过滤列表进行选择。只有符合条件的路由才会被加入。

首先，配置一个访问列表（access-list）或地址前缀列表（prefix-list），在其中规定那些路由是可以被加入到路由表中的。

其次，使用distribute-list in 命令引用该列表即可。

```
distribute-list {access-list-number | prefix-list name} in
```

对发布的路由进行过滤

- OSPF在向外发布自治系统外部路由（ASE）时可以选择那些路由需要发布，那些不需要发布。

首先，可以配置一个访问列表（access-list）或地址前缀列表（prefix-list），也可以直接指定对哪些协议（routing-process）的路由进行过滤。

其次，使用distribute-list out 命令引用该列表即可。

distribute-list {access-list-number | prefix-list name} out [routing-process |autonomous-system-number]

3.7路由协议的安全性

随着信息技术的发展，网络在当代社会中发挥着日益重要的作用，网络的安全性也更加的被人们所重视，各种安全保障措施纷纷出台，如使用防火墙，构建VPN（虚拟私有网）等等。这里所要讲述的是路由协议的安全性保障措施。

由于路由协议的特殊用途，它所传递的不是用户数据，而是要将自己所知道的路由信息可靠的传递到目的地，因此路由协议的安全性主要侧重于可靠性的保障，而非对私有性和保密性的关注。也就是确保收到的路由报文是来自可靠的对等体，并且没有被人擅自改动，从而保证得到正确的路由信息，不被恶意的欺骗报文所误导。为了达到这样一个目的，大部分的动态路由协议都采用了报文认证的方式来验证报文的正确性。报文认证的基本思想是，交换路由信息的两个对等体上分别设置认证关键字，在发送的报文上携带这个关键字或由它生成的密文验证信息，当对等体受到携带认证信息的报文后，根据报文上的验证信息和本地设置的认证关键字来确定报文是否可靠，对于认证不通过的报文予以丢弃。

通用的报文认证方式有两种。一种是普通的明文认证方式，也就是将认证关键字直接放到报文的特定的区域，随报文一起发送。明文认证由于没有采取任何加密保护措施，既不能保证认证关键字不被人窃取，也不能确认报文是否在传输过程中被人篡改，不能起到很好的安全性作用。另一种则是使用MD5算法的密文认证方式。MD5算法是一种单向信息摘要算法，它能够根据任意长度的数据信息做出16字节的信息摘要来，对于MD5算法而言，不同内容的数据信息所得出的摘要是不可能相同的。当路由协议使用MD5密文认证方式时，它对连同认证关键字的整个报文做MD5摘要，并将得出的摘要随报文一起传递，报文并不携带认证关键字本身。这样如果报文在传输过程中被人修改，甚至是伪造的报文，那么在对等体那里算出的摘要信息就会与报文上所携带的摘要信息有所不同，从而保证报文的可靠性。但正如前面说过的，即使是采用MD5密文认证方式，报文的具体内容也都是不被加密的，因此报文的认证机制并不能提供保密性的支持。

路由信息本来就是要尽量广播出去的，因此，报文的保密性和私有性对路由协议来说意义不大。 目前流行的路由协议，如RIP、OSPF等都支持上述的报文认证机制，可以提供良好的安全性保证。

3.8路由协议的发展

3.8.1MPLS

MPLS（Muxtiprotocol Label Switch）最初是用来提高路由器的转发速度而提出一个协议，但是由于MPLS在流量工程（Traffic Engeering）和VPN这一在目前IP网络中非常关键的两项技术中表现。MPLS已日益成为扩大IP网络规模的重要标准。

MPLS协议的关键是引入了标签（Label）的概念。它是一种短的易于处理的、不包含拓扑信息、只具有局部意义的信息内容。Label短是为了易于处理，通常可以用索引直接引用。只具有局部意义是为了便于分配。熟识ATM的人可能很自然是想到ATM中的VPI/VCI。可以这么说ATM中的VPI/VCI就是一种标签。所以说ATM实际上就是一种标签交换。

在MPLS网络中，IP全在进入第一个MPLS设备时，MPLS边缘路由器就用这些标签封装起来。MPLS边缘路由器分析IP包的内容并且为这些IP包选择合适的标签，相对于传统的IP路由分析，MPLS不仅分析IP包头中的目的地址信息。它还分析IP包头中的其他信息。如TOS等。尔后所有MPLS网络中节点都是依据这个简短标签来作为转发判决依据。当该IP包最终离开MPLS网络时，标签被边缘路由器分离。

3.8.2路由协议与VPN

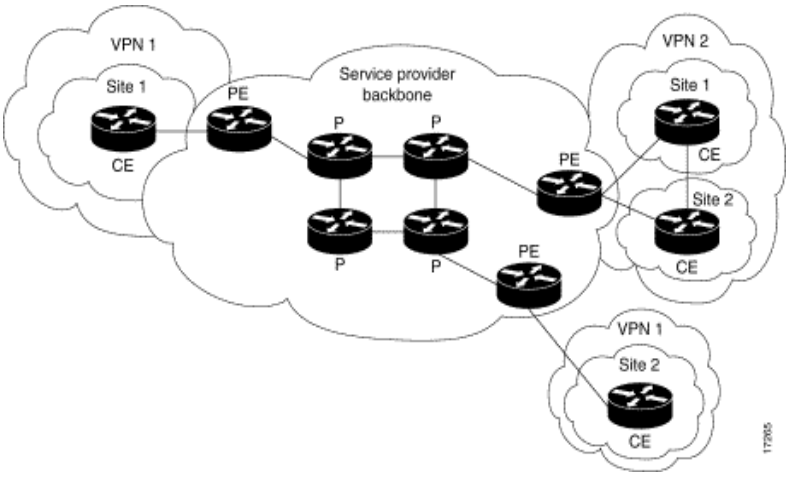
利用公共网络来构建的私人专用网络称为虚拟私有网络（VPN，Virtual Private Network），用于构建VPN的公共网络包括Internet、帧中继、ATM等。在公共网络上组建的VPN象企业现有的私有网络一样提供安全性、可靠性和可管理性等。

“虚拟”的概念是相对传统私有网络的构建方式而言的。对于广域网连接，传统的组网方式是通过远程拨号连接来实现的，而VPN是利用服务提供商所提供的公共网络来实现远程的广域连接。通过VPN，企业可以以明显更低的成本连接它们的远地办事机构、出差工作人员以及业务合作伙伴。

VPN技术也在不停的发展，最初的VPN是以IP Tunnel技术，在公网上为企业提供私有隧道的方式提供VPN。它的本质上类似与直接采用专线来构建企业网，由

企业自己来维护网络，操心网络安全等技术细节。随着虚拟路由器的概念出现，目前的发展趋势是VPN完全以一种增值业务的方式，由骨干网运营商向企业提供。类比主机托管，企业其实真正想得到的是VPN这种服务，并不关心里面实现的技术细节，更何况能节约的大量IT维护费用。

VPN可以由多种技术来实现，如基于IP Tunnel的VPN和基于MPLS/BGP的VPN。VPN的典型实例见下图：



运营商在骨干网边缘节点PE向用户提供VPN业务。 利用隧道来构造和实现VPN，实现VPN报文在SP网络中的透传，并保证VPN的私有性和安全性。隧道可以是IP Tunnel，如GRE、IPoverIP、IPSec，也可以是MPLS中的标签转发路径LSP（Label Switch Path）。在每个边缘节点PE为每个Site提供一套单独的路由表，不同Site来的数据报文按照不同的路由表转发。等于是向每个Site提供了一个虚拟的路由器。由于存在不同的路由表，允许不同VPN的地址可以冲突。在各个PE节点之间需要一种信令协议来传输VPN的信息。我们采用了BGP-4协议协议的扩展来实现。扩展后的BGP支持

- 能力协商
- 扩展团体属性
- 多地址组支持（MBGP）
- 携带标签

3.8.3路由协议与流量工程的融合

目前的路由协议专注于连通性，通常只支持“Best Effort”的数据包转发模式。在Internet中，路由器根据IP包的目的地址，选择一条到目的地最短的路径。这种方法非常简单，容易实现大型的网络互联。但这样选择出来的路径并不总是最优的。

在两个Internet节点之间可能有多条路径，它们的带宽、延时、当前的负载各不相同。Traffic Engineering的目的就是在多条路径中，选择一条较好的路径，以支持QOS，并避免拥塞。

选择出路径后，通过显式路由的配合或其它手段，实现Traffic Engineering的控制目标。

在实际的Internet中，经常发生拥塞，数据包可能大量丢失。拥塞的发生有两种可能：

1。网络资源不足，不能承载目前的业务。

对这种情况，可以通过增加资源，扩充带宽解决，也可以通过流量整形，滑动窗口，队列缓存，反压等拥塞控制算法，避免拥塞的发生。

2。线路负载分担不均匀，有的线路很繁忙，有的很空闲。在繁忙的线路上就会发生拥塞。

通过Traffic Engineering合理地分配负载，可以避免这种拥塞的发生。

实现Traffic Engineering后，可以提升网络的错误恢复能力，很快定位并自动解决网络故障，降低网络管理成本。

预计Traffic Engineering不久将具有下面的特征：

1。支持大型网络，优化算法，提高计算速度。

2。当发现网络结构变化时，重新快速计算Traffic Trunk的能力。

3。支持VPN，自动选择和改变路径，保证VPN的QOS。

4。对网络进行更好的预测，使计算出的Traffic Trunk具有较高的稳定性。

Traffic Engineering的研究内容有两个方面：

1。和其它协议的配合。Traffic Engineering需要和路由协议、RSVP等配合，它们之间的接口，对其它协议的扩展，都是需要研究的内容。 Traffic Engineering其实是网管的一部分，在一定的网管策略下和其它部件协同工作。

2。优化Traffic Engineering的计算方法，提高算法的速度、稳定性。