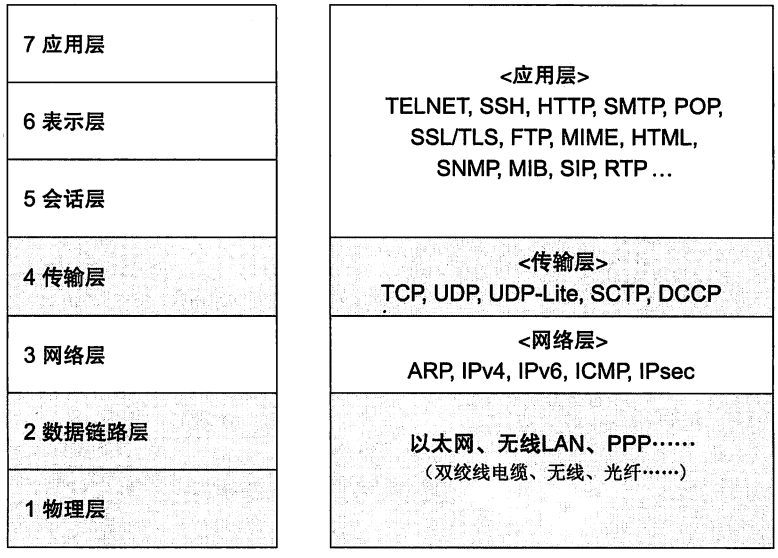
第7章

路由协议

在互联网世界中，夹杂着复杂的LAN和广域网。然而，再复杂的网络结构中，也需要通过合理的路由将数据发送到目标主机。而决定这个路由的，正是路由控制模块。本章旨在详细介绍路由控制以及实现路由控制功能的相关协议。



7.1 路由控制的定义

7.1.1 IP地址与路由控制

互联网是由路由器连接的网络组合而成的。为了能让数据包正确达地到达目标主机，路由器必须在途中进行正确地转发。这种向“正确的方向”转发数据所进行的处理就叫做路由控制或路由。

路由器根据路由控制表（Routing Table）转发数据包。它根据所收到的数据包中目标主机的IP地址与路由控制表的比较得出下一个应该接收的路由器。因此，这个过程中路由控制表的记录一定要正确无误。但凡出现错误，数据包就有可能无法到达目标主机。

7.1.2 静态路由与动态路由

▼ Static Routring
▼ Dynamic Routing

那么，是谁又是怎样制作和管理路由控制表的呢？路由控制分静态▼和动态▼两种类型。

静态路由是指事先设置好路由器和主机中并将路由信息固定的一种方法。而动态路由是指让路由协议在运行过程中自动地设置路由控制信息的一种方法。这些方法都有它们各自的利弊。

静态路由的设置通常是由使用者手工操作完成的。例如，有100个IP网的时候，就需要设置近100个路由信息。并且，每增加一个新的网络，就需要将这个新被追加的网络信息设置在所有的路由器上。因此，静态路由给管理者带来很大的负担，这是其一。还有一个不可忽视的问题是，一旦某个路由器发生故障，基本上无法自动绕过发生故障的节点，只有在管理员手工设置以后才能恢复正常。

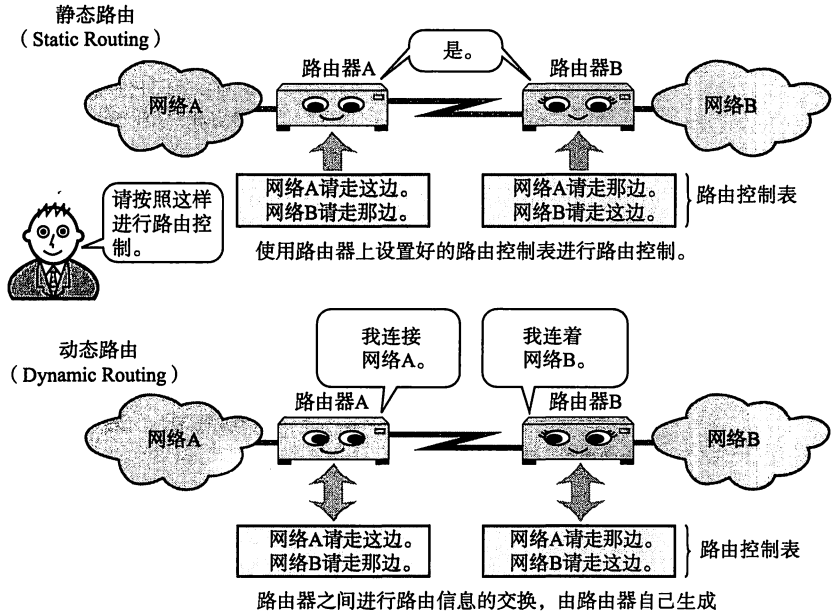


图7.1 静态路由与动态路由

使用动态路由的情况下，管理员必须设置好路由协议，其设定过程的复杂程度与具体要设置路由协议的类型有直接关系。例如在 RIP 的情况下，基本上无需过多的设置。而根据 OSPF 进行较详细路由控制时，设置工作将会非常繁琐。

如果有一个新的网络被追加到原有的网络中时，只要在新增加网络的路由器上进行一个动态路由的设置即可。而不需要像静态路由那样，不得不在其他所有路由器上进行修改。对于路由器个数较多的网络，采用动态路由显然是一个能够减轻管理员负担的方法。

况且，网络上一旦发生故障，只要有一个可绕的其他路径，那么数据包就会自动选择这个路径，路由器的设置也会自动重置。路由器为了能够像这样定期相互交换必要的路由控制信息，会与相邻的路由器之间互发消息。这些互换的消息会给网络带来一定程度的负荷。

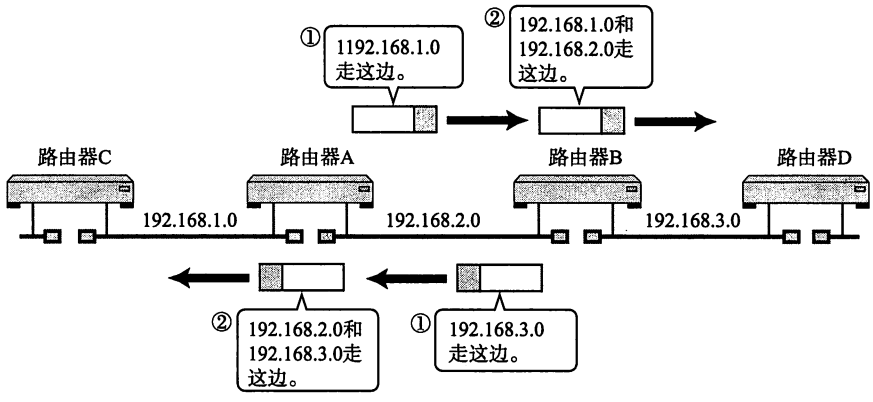
不论是静态路由还是动态路由，不要只使用其中一种，可以将它们组合起来使用。

7.1.3 动态路由的基础

▼图 7.2 中的传输，只有在没有循环的情况下才能很好地运行。例如路由器 C 和路由器 D 之间如果有连接，那么将无法正常工作。

动态路由如图 7.2 所示，会给相邻路由器发送自己已知的网络连接信息，而这些信息又像接力一样依次传递给其他路由器，直至整个网络都了解时，路由控制表也就制作完成了。而此时也就可以正确转发 IP 数据包了▼。

图 7.2 根据路由协议交换路由信息



*箭头（➡）表示路由信息流动的方向

7.2 路由控制范围

随着IP网络的发展，想要对所有网络统一管理是不可能的事。因此，人们根据路由控制的范围常使用IGP（Interior Gateway Protocol）和EGP（Exterior Gateway Protocol）▼两种类型的路由协议。

▼EGP是特定的路由协议名称，请不要与其他同名词汇混淆。

7.2.1 接入互联网的各种组织机构

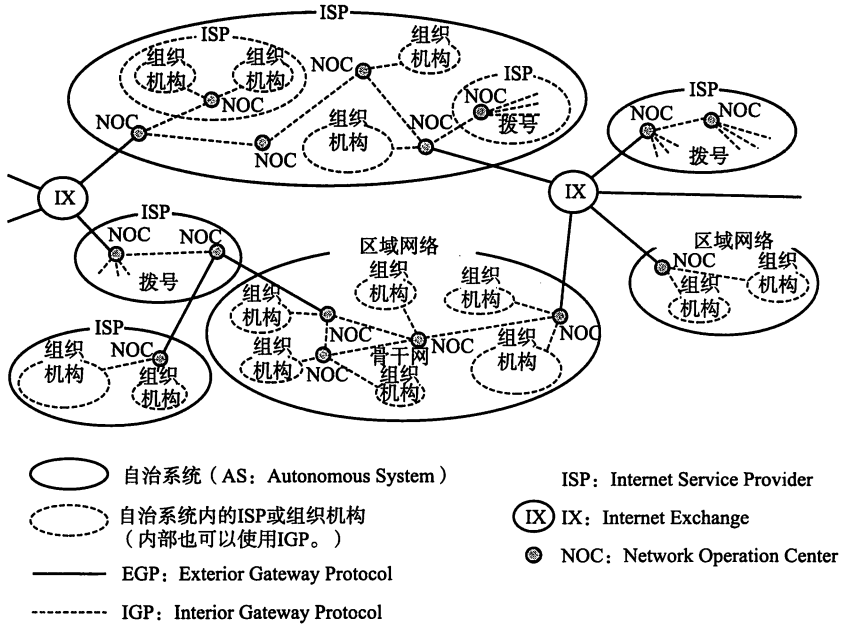
互联网连接着世界各地的组织机构，不仅包括语言不相通的，甚至包括宗教信仰全然不同的组织。没有管理者，也没有被管理者，每个组织之间保持着平等的关系。

7.2.2 自治系统与路由协议

企业内部网络的管理方针，往往由该企业组织内部自行决定。因此每个企业或组织机构对网络管理和运维的方法都不尽相同。为了提高自己的销售额和生产力，各家企业和组织机构都会相应购入必要的机械设备、构建合适的网络以及采用合理的运维体制。在这种环境下，可以对公司以外的人士屏蔽企业内部的网络细节，更不必对这些细节上的更新请求作出回应。这好比我们的日常生活，每个人对家庭内部的私事，都不希望过多暴露给外界，听从外界指挥。

制定自己的路由策略，并以此为准在一个或多个网络群体中采用的小型单位叫做自治系统（AS：Autonomous System）或路由选择域（Routing Domain）。

图7.3 EGP与IGP



说到自治系统，区域网络、ISP（互联网服务提供商）等都是典型的例子。在区域网络及ISP内部，由构造、管理和运维网络的管理员、运营者制定出路由控制相关方针，然后根据此方针进行具体路由控制的设定。

而接入到区域网络或 ISP 的组织机构，则必须根据管理员的指示进行路由控制设定。如果不遵循这个原则，会给其他使用者带来负面影响，甚至使自己也无法与任何组织机构进行通信。

自治系统（路由选择域）内部动态路由采用的协议是域内路由协议，即 IGP。而自治系统之间的路由控制采用的是域间路由协议，即 EGP。

7.2.3 IGP 与 EGP

如前面述，路由协议大致分为两大类。一类是外部网关协议 EGP，另一类是内部网关协议 IGP（Interior Gateway Protocol）。

IP 地址分为网络部分和主机部分，它们有各自的分功。EGP 与 IGP 的关系与 IP 地址网络部分和主机部分的关系有相似之处。就像根据 IP 地址中的网络部分在网络之间进行路由选择、根据主机部分在链路内部进行主机识别一样，可以根据 EGP 在区域网络之间（或 ISP 之间）进行路由选择，也可以根据 IGP 在区域网络内部（或 ISP 内部）进行主机识别。

由此，路由协议被分为 EGP 和 IGP 两个层次。没有 EGP 就不可能有世界上各个不同组织机构之间的通信。没有 IGP 机构内部也就不可能进行通信。

IGP 中还可以使用 RIP（Routing Information Protocol，路由信息协议）、RIP2、OSPF（Open Shortest Path First，开放式最短路径优先）等众多协议。与之相对，EGP 使用的是 BGP（Border Gateway Protocol，边界网关协议）协议。

7.3 路由算法

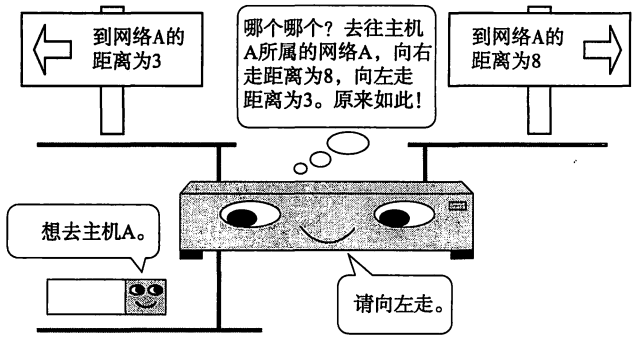
路由控制有各种各样的算法，其中最具代表性的有两种，是距离向量（Distance-Vector）算法和链路状态（Link-State）算法。

7.3.1 距离向量算法

▼Metric是指转发数据时衡量路由控制中距离和成本的一种指标。在距离向量算法中，代价相当于所要经过的路由器的个数。

距离向量算法（DV）是指根据距离（代价▼）和方向决定目标网络或目标主机位置的一种方法。

图7.4 距离向量



距离向量（Distance-Vector）中通过距离与方向确定通往目标网络的路径。

路由器之间可以互换目标网络的方向及其距离的相关信息，并以这些信息为基础制作路由控制表。这种方法在处理上比较简单，不过由于只有距离和方向的信息，所以当网络构造变得分外复杂时，在获得稳定的路由信息之前需要消耗一定时间▼，也极易发生路由循环等问题。

▼也叫做路由收敛。

7.3.2 链路状态算法

链路状态算法是路由器在了解网络整体连接状态的基础上生成路由控制表的一种方法。该方法中，每个路由器必须保持同样的信息才能进行正确的路由选择。

距离向量算法中每个路由器掌握的信息都不相同。通往每个网络所耗的距离（代价）也根据路由器的不同而不同。因此，该算法的一个缺点就是不太容易判断每个路由器上的信息是否正确。

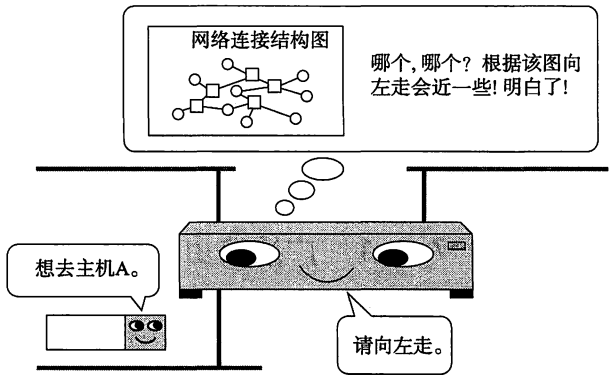
而链路状态算法中所有路由器持有相同的信息。对于任何一台路由器，网络拓扑都完全一样。因此，只要某一台路由器与其他路由器保持同样的路由控制信息，就意味着该路由器上的路由信息是正确的。只要每个路由器尽快地与其他路由器同步▼路由信息，就可以使路由信息达到一个稳定的状态。因此，即使网络结构变得复杂，每个路由器也能够保持正确的路由信息、进行稳定的路由选择。这也是该算法的一个优点。

▼同步一词常用于分布式系统，意指所有系统中保持同样的值。

为了实现上述机制，链路状态算法付出的代价就是如何从网络代理获取路由信息表。这一过程相当复杂，特别是在一个规模巨大而又复杂的网络结构中，管理和处理代理信息需要高速CPU处理能力和大量的内存▼。

▼为此，OSPF正致力于将网络分割为不同的区域，以减少路由控制信息。

图 7.5
链路状态



链路状态（Link-State）中路由器知道网络的连接状态，并根据该图的信息确定通往目标网络的路径。

7.3.3 主要路由协议

路由协议分很多种。表 7.1 列出了主要的几种路由协议。

▼此处的 EGP 不是区分 IGP 与 EGP 的那个 EGP，而是指一种叫做 EGP 的特定的协议。

其中，由于 EGP▼ 不支持 CIDR，现在已经不再用作互联网的对外连接协议了。在以后的章节中将详细介绍 RIP、RIP2、OSPF、BGP 等协议的基础知识。

表 7.1
几种路由协议的特点

路由协议名	下一层协议	方　式	适用范围	循环检测
RIP	UDP	距离向量	域内	不可以
RIP2	UDP	距离向量	域内	不可以
OSFP	IP	链路状态	域内	可以
EGP	IP	距离向量	对外连接	不可以
BGP	TCP	路径向量	对外连接	可以

7.4 RIP

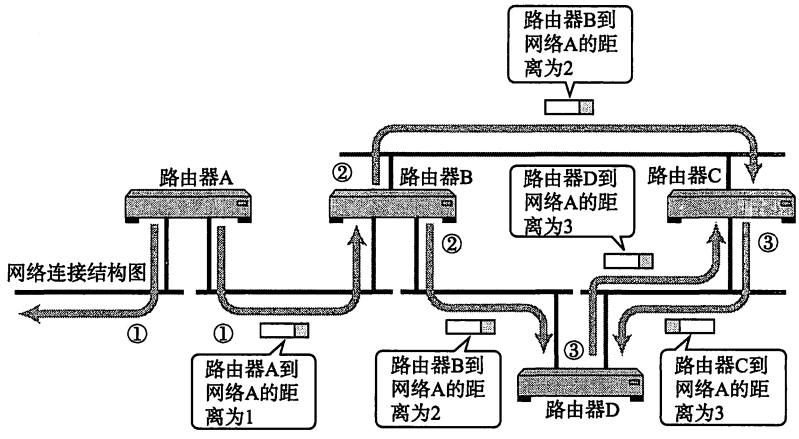
RIP（Routing Information Protocol）是距离向量型的一种路由协议，广泛用于LAN。被BSD UNIX作为标准而提供的routed▼采用了RIP，因此RIP得到了迅速的普及。

▼在UNIX系统上的一个守护进程。该进程实现了RIP协议。

7.4.1 广播路由控制信息

RIP将路由控制信息定期（30秒一次）向全网广播。如果没有收到路由控制信息，连接就会被断开。不过，这有可能是由于丢包导致的，因此RIP规定等待5次。如果等了6次（180秒）仍未收到路由信息，才会真正关闭连接。

图7.6
RIP概要

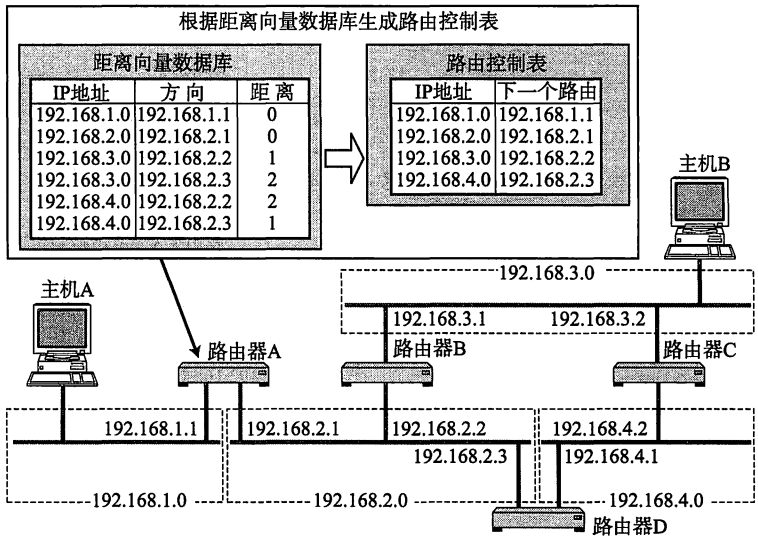


① 30秒一次，将自己所知道的路由信息广播出去。
② 将已知的路由信息经过一跳之后继续广播。
③ 以此类推，逐步传播路由信息。

7.4.2 根据距离向量确定路由

RIP基于距离向量算法决定路径。距离（Metrics）的单位为“跳数”。跳数是指所经过的路由器的个数。RIP希望尽可能少通过路由器将数据包转发到目标IP地址，如图7.7所示。根据距离向量生成距离向量表，再抽出较小的路由生成最终的路由控制表。

图 7.7
根据距离向量生成路由控制表



距离向量型的协议中根据网络的距离和方向生成路由控制表。

针对同一个网络如果有两条路径，那么选择距离较短的一个▼。

▼如果距离相等，那么根据路由器的类型选择的路由也会不同，通常是随机选择一个或是轮换使用。

7.4.3 使用子网掩码时的 RIP 处理

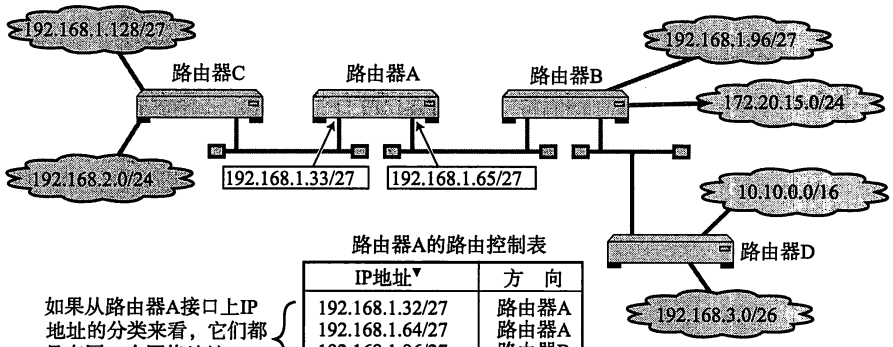
RIP 虽然不交换子网掩码信息，但可以用于使用子网掩码的网络环境。不过在这种情况下需要注意以下几点：

- 从接口的 IP 地址对应分类得出网络地址后，与根据路由控制信息流过此路由器的包中的 IP 地址对应的分类得出的网络地址进行比较。如果两者的网络地址相同，那么就以接口的网络地址长度为准。
- 如果两者的网络地址不同，那么以 IP 地址的分类所确定的网络地址长度为准。

例如，路由器的接口地址为 192.168.1.33/27。很显然，这是一个 C 类地址，因此按照 IP 地址分类它的网络地址为 192.168.1.33/24。与 192.168.1.33/24 相符合的 IP 地址，其网络地址长度都被视为 27 位。除此之外的地址，则采用每个地址的分类所确定的网络地址长度。

因此，采用 RIP 进行路由控制的范围内必须注意两点：一是，因 IP 地址的分类而产生不同的网络地址时；二是，构造网络地址长度不同的网络环境时。

图 7.8
RIP 与子网掩码



路由器A的路由控制表

	IP地址▼	方　向
如果从路由器A接口上IP地址的分类来看，它们都具有同一个网络地址。	192.168.1.32/27	路由器A
	192.168.1.64/27	路由器A
	192.168.1.96/27	路由器B
	192.168.1.128/27	路由器C
从分类的角度，它们与路由器A的网络地址不相同。	192.168.2.0/24	路由器C
	192.168.3.0/24	路由器B
	172.20.0.0/16	路由器B
	10.0.0.0/8	路由器B

▼当把 IP 地址分类表示的网络地址延长至子网掩码的长度时，所延长的部分如果为 0，称之为 0 子网；如果为 1，则称之为 1 子网。需要注意的是 0 子网与 1 子网在 RIP 中都无法使用。（但是它们可以被用于 RIP2 和 OSPF 以及静态路由中。）

7.4.4　RIP 中路由变更时的处理

RIP 的基本行为可归纳为如下两点：

- 将自己所知道的路由信息定期进行广播。
- 一旦认为网络被断开，数据将无法流过此路由器，其他路由器也就可以得知网络已经断开。

不过，这两点不论哪种方式都存在一些问题。

如图 7.9，路由器 A 将网络 A 的连接信息发送给路由器 B，路由器 B 又将自己掌握的路由信息在原来的基础上加 1 跳后发送给路由器 A 和路由器 C。假定这时与网络 A 发生了故障。

路由器 A 虽然觉察到自己与网络 A 的连接已经断开，无法将网络 A 的信息发送给路由器 B，但是它会收到路由器 B 曾经获知的消息。这就使得路由器 A 误认为自己的信息还可以通过路由器 B 到达网络 A。

像这样收到自己发出去的消息，这个问题被称为无限计数（Counting to Infinity）。为了解决这个问题可以采取以下两种方法：

- 一是最长距离不超过 16▼。由此即使发生无限计数的问题，也可以从时间上进行控制。
- 二是规定路由器不再把所收到的路由消息原路返还给发送端。这也被称作水平分割（Split Horizon）。

▼“距离为 16”这个信息只会被保留 120 秒。一旦超过这个时间，信息将会被删除，无法发送。这个时间由一个叫做垃圾收集计时器（Garbage-collection Timer）的工具进行管理。

图 7.9
无限计数问题

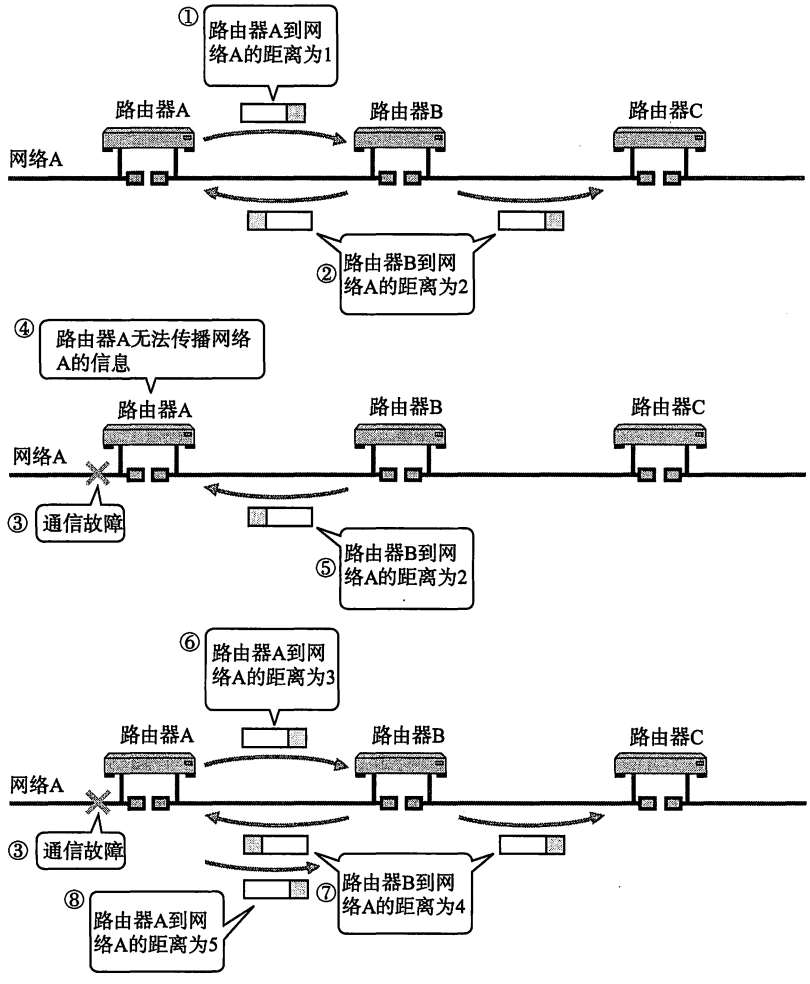
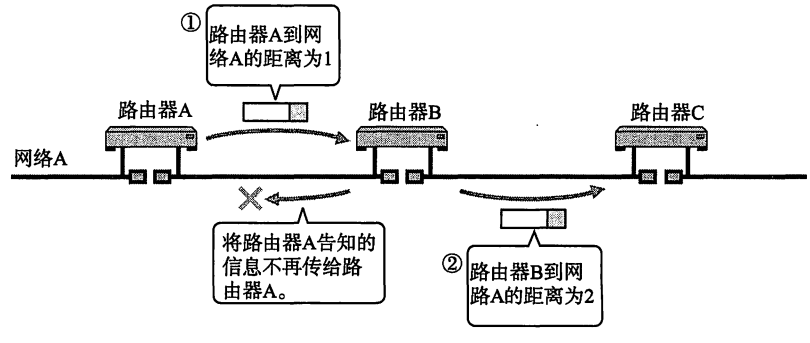


图 7.10
水平分割



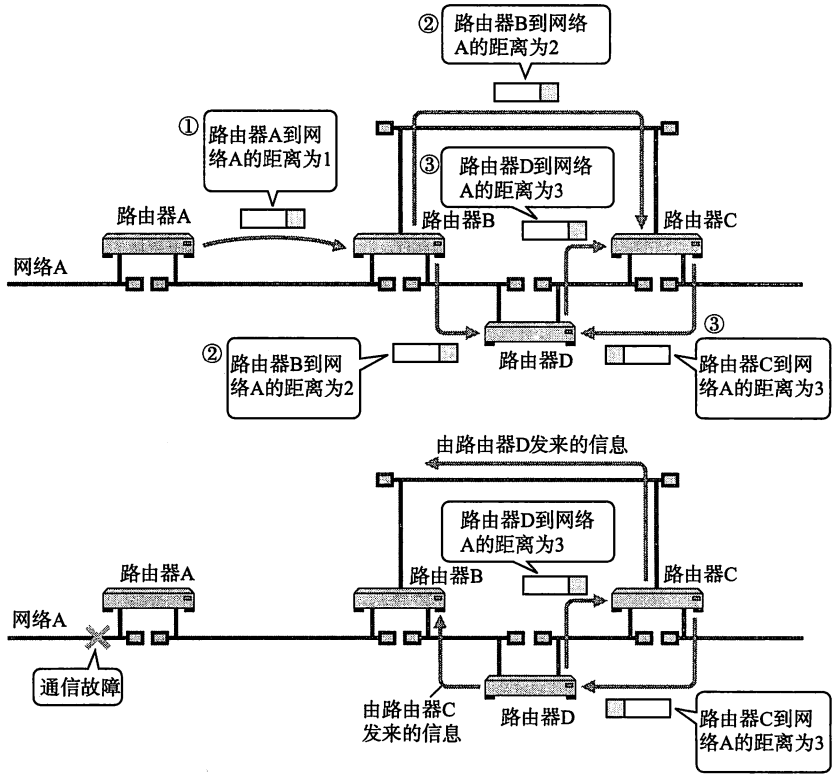
然而，这种方法对有些网络来说是无法解决问题的。如图 7.11 所示，在网络本身就有环路的情况下。

在有环路情况下，反向的回路会成为迂回的通道，路由信息会不断地被循环往复地转发。当环路内部某一处发生通信故障时，通常可以设置一个正确的迂回

通道。但是对于图7.11中的情况，当网络A的通信发生故障时，将无法传送正确的路由信息。尤其是在环路有多余的情况下，需要很长时间才能产生正确的路由信息。

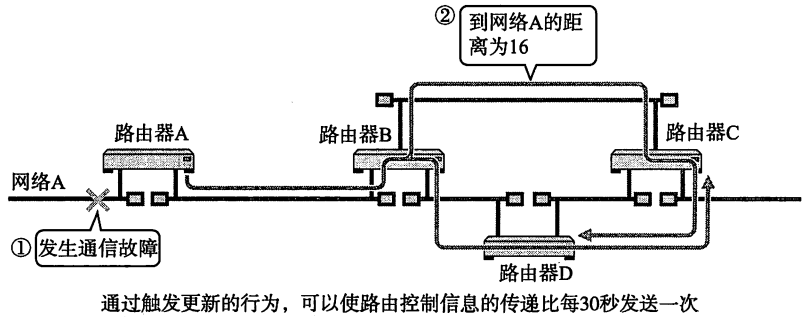
为了尽可能解决这个问题，人们提出了“毒性逆转”（Poisoned Reverse）和“触发更新”（Triggered Update）两种方法。

图7.11
带有环路的网络



毒性逆转是指当网络中发生链路被断开的时候，不是不再发送这个消息，而是将这个无法通信的消息传播出去。即发送一个距离为16的消息。触发更新是指当路由信息发生变化时，不等待30秒而是立刻发送出去的一种方法。有了这两种方法，在链路不通时，可以迅速传送消息以使路由信息尽快收敛。

图7.12
毒性逆转和触发更新



通过触发更新的行为，可以使路由控制信息的传递比每30秒发送一次的情况快很多，因此可以有效避免错误路由信息被不断发送。

然而，纵然使用了到现在为止所介绍的方法，在一个具有众多环路的复杂的网络环境中，路由信息想要达到一个稳定的状态是需要花一段时间的。为了解决

这个问题，必须明确地掌握网络结构，在了解究竟哪个链路断开后再进行路由控制非常重要。为此，可以采用 OSPF。

7.4.5 RIP2

RIP2 的意思是 RIP 第二版。它是在 RIP 使用过程中总结了经验的基础上进行改良后的一种协议。第二版与第一版的工作机制基本相同，不过仍有如下几个新的特点。

使用多播

RIP 中当路由器之间交换路由信息时采用广播的形式，然而在 RIP2 中改用了多播。这样不仅减少了网络的流量，还缩小了对无关主机的影响。

支持子网掩码

与 OSPF 类似的，RIP2 支持在其交换的路由信息中加入子网掩码信息。

路由选择域

与 OSPF 的区域类似，在同一个网络中可以使用逻辑上独立的多个 RIP。

外部路由标志

通常用于把从 BGP 等获得的路由控制信息通过 RIP 传递给 AS 内。

身份验证密钥

与 OSPF 一样，RIP 包中携带密码。只有在自己能够识别这个密码时才接收数据，否则忽略这个 RIP 包。

7.5 OSPF

▼ Intermediate System to Intermediate System Intra-Domain routing information exchange protocol，中间系统到中间系统的路由选择协议。

OSPF（Open Shortest Path First）是根据OSI的IS-IS▼协议而提出的一种链路状态型路由协议。由于采用链路状态类型，所以即使网络中有环路，也能够进行稳定的路由控制。

另外，OSPF支持子网掩码。由此，曾经在RIP中无法实现的可变长度子网构造的网络路由控制成为现实。

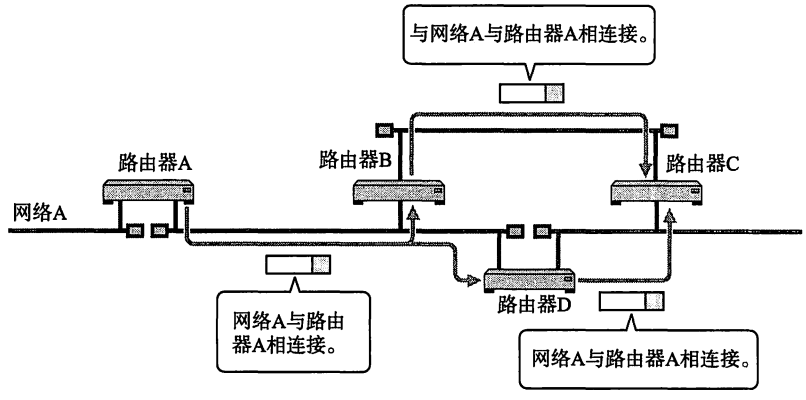
甚至为了减少网络流量，OSPF还引入了“区域”这一概念。区域是将一个自治网络划分为若干个更小的范围。由此，可以减少路由协议之间不必要的交换。

OSPF可以针对IP首部中的区分服务（TOS）字段，生成多个路由控制表。不过，也会出现已经实现了OSPF功能的路由器无法支持这个TOS的情况。

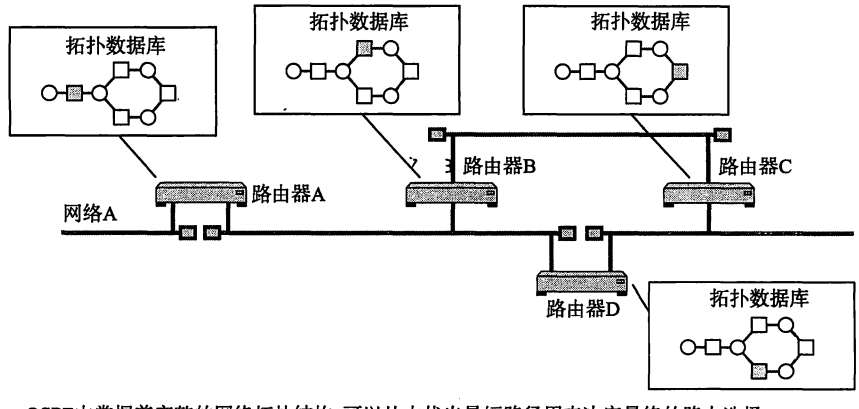
7.5.1 OSPF是链路状态型路由协议

OSPF为链路状态型路由器。路由器之间交换链路状态生成网络拓扑信息，然后再根据这个拓扑信息生成路由控制表。

图7.13 由链路状态确定路由



与哪个网络或与哪个路由器相连的信息要通过接力的方式进行发送。

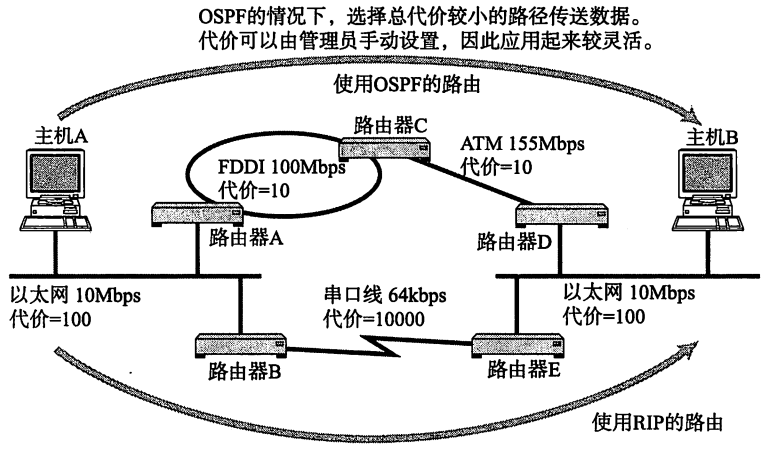


OSPF中掌握着完整的网络拓扑结构，可以从中找出最短路径用来决定最终的路由选择。

RIP 的路由选择，要求途中所经过的路由器个数越少越好。与之相比，OSPF 可以给每条链路▼赋予一个权重（也可以叫做代价），并始终选择一个权重最小的路径作为最终路由。也就是说 OSPF 以每个链路上的代价为度量标准，始终选择一个总的代价最小的一条路径。如图 7.14 对比所示，RIP 是选择路由器个数最少的路径，而 OSPF 是选择总的代价较小的路径。

▼实际上，可以为连到该数据链路（子网）的网卡设置一个代价。而这个代价只用于发送端，接收端不需要考虑。

图 7.14 网络权重与路由选择



7.5.2 OSPF 基础知识

在 OSPF 中，把连接到同一个链路的路由器称作相邻路由器（Neighboring Router）。在一个相对简单的网络结构中，例如每个路由器仅跟一个路由器相互连接时▼，相邻路由器之间可以交换路由信息。但是在一个比较复杂的网络中，例如在同一个链路中加入了以太网或 FDDI 等路由器时，就不需要在所有相邻的路由器之间都进行控制信息的交换，而是确定一个指定路由器（Designated Router），并以它为中心交换▼路由信息即可。

▼在专线网络中，路由器之间采用 PPP 相连。

▼邻接路由器中相互交换路由信息的关系叫做邻接（Adjancency）。

RIP 中包的类型只有一种。它利用路由控制信息，一边确认是否连接了网络，一边传送网络信息。但是这种方式，有一个严重的缺点。那就是，网络的个数越多，每次所要交换的路由控制信息就越大。而且当网络已经处于比较稳定的、没有什么变化的状态时，还是要定期交换相同的路由控制信息，这在一定程度上浪费了网络带宽。

而在 OSPF 中，根据作用的不同可以分为 5 种类型的包。

表 7.2 OSPF 包类型

类型	包　名	功　能
1	问候（HELLO）	确认相邻路由器、确定指定路由器
2	数据库描述（Database Description）	链路状态数据库的摘要信息
3	链路状态请求（Link State Request）	请求从数据库中获取链路状态信息
4	链路状态更新（Link State Update）	更新链路状态数据库中的链路状态信息
5	链路状态确认应答（Link State Acknowledgement）	链路状态信息更新的确认应答

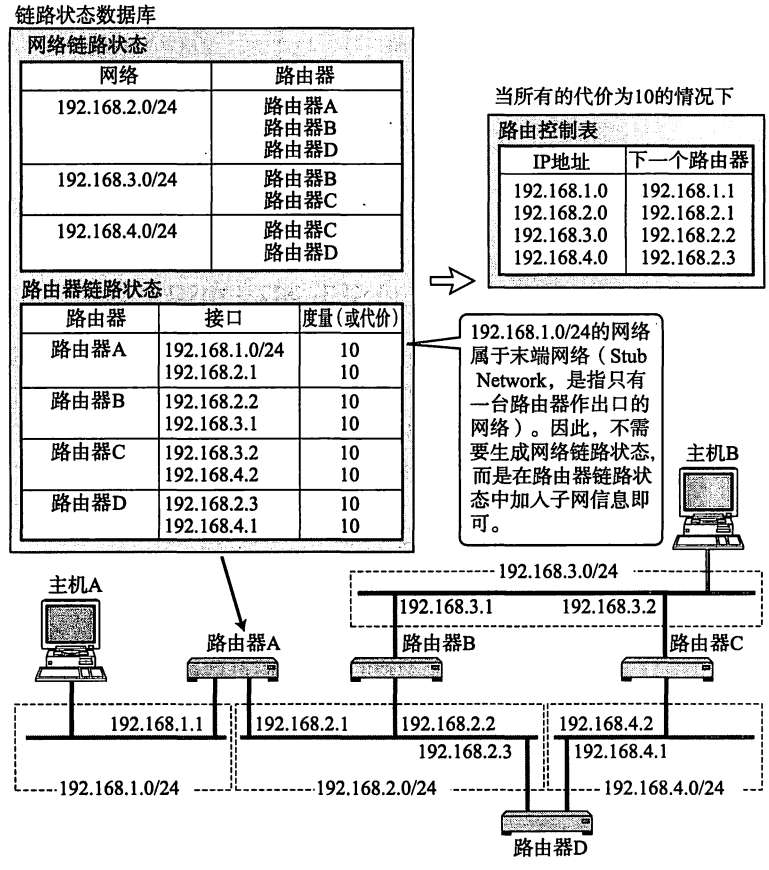
通过发送问候（HELLO）包确认是否连接。每个路由器为了同步路由控制信息，利用数据库描述（Database Description）包相互发送路由摘要信息和版本信息。如果版本比较老，则首先发出一个链路状态请求（Link State Request）包请求路由控制信息，然后由链路状态更新（Link State Update）包接收路由状态信息，最后再通过链路状态确认（Link State ACK Packet）包通知大家本地已经接收到路由控制信息。

有了这样一个机制以后，OSPF不仅可以大大地减少网络流量，还可以达到迅速更新路由信息的目的。

7.5.3　OSPF工作原理概述

OSPF中进行连接确认的协议叫做HELLO协议。

图7.15
OSPF中根据链路状态生成路由控制表



LAN中每10秒发送一个HELLO包。如果没有HELLO包到达，则进行连接是否断开的判断▼。具体为，允许空等3次，直到第4次（40秒后）仍无任何反馈就认为连接已经断开。之后在进行连接断开或恢复连接操作时，由于链路状态发生了变化，路由器会发送一个链路状态更新包（Link State Update Packet）通知其他路由器网络状态的变化。

▼管理员可以自定义HELLO包的发送间隔和判断连接断开的时间。只是在同一个链路中的设备必须配置相同的值。

链路状态更新包所要传达的消息大致分为两类：一是网络LSA▼，另一个是路由器LSA▼。

▼ Network Link State Advertisement，网络链路状态通告。

▼ Router Link State Advertisement，路由器链路状态通告。

网络 LSA 是以网络为中心生成的信息，表示这个网络都与哪些路由器相连接。而路由器 LSA 是以路由器为中心生成的信息，表示这个路由器与哪些网络相连接。

如果这两种信息▼主要采用 OSPF 发送，每个路由器就都可以生成一个可以表示网络结构的链路状态数据库。可以根据这个数据库、采用 Dijkstra 算法▼（最短路径优先算法）生成相应的路由控制表。

▼除这两种信息之外还有网络汇总 LSA（Summary LSA）和自治系统外部 LSA（AS External LSA）信息。

▼ Dijkstra 算法由提出结构化编程的 E. W. Dijkstra 发明。该算法用来获取最短路径。

相比距离向量，由上述过程所生成的路由控制表更加清晰不容易混淆，还可以有效地降低无线循环问题的发生。不过，当网络规模逐渐越大时，最短路径优先算法的处理时间就会变得越长，对 CPU 和内存的消耗也就越大。

7.5.4 将区域分层化进行细分管理

链路状态型路由协议的潜在问题在于，当网络规模越来越大时，表示链路状态的拓扑数据库就变得越来越大，路由控制信息的计算也就越困难。OSPF 为了减少计算负荷，引入了区域的概念。

区域是指将连接在一起的网络和主机划分成小组，使一个自治系统（AS）内可以拥有多个区域。不过具有多个区域的自治系统必须要有一个主干区域▼（Backbone Area），并且所有其他区域必须都与这个主干区域相连接▼。

▼主干区域的 ID 为 0。逻辑上只允许它有 1 个，可实际在物理上又可以划分为多个。

▼如果网络的实际物理构造与此说明不符时，需要采用 OSPF 的虚拟链路功能设置虚拟的主干或区域。

连接区域与主干区域的路由器称作区域边界路由器；而区域内部的路由器叫做内部路由器；只与主干区域内连接的路由器叫做主干路由器；与外部相连接的路由器就是 AS 边界路由器。

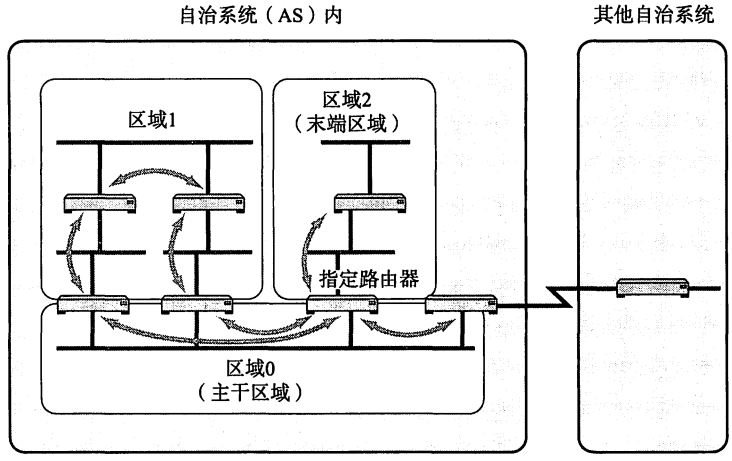
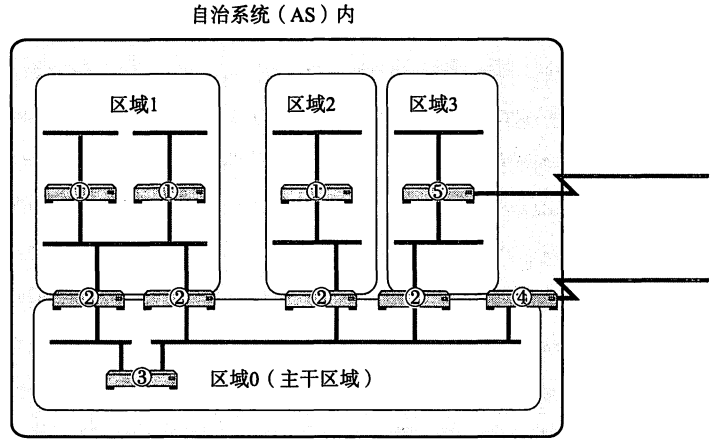


图 7.16 AS 与区域

每个区域内的路由器都持有本区域网络拓扑的数据库。然而，关于区域之外的路径信息，只能从区域边界路由器那里获知它们的距离。区域边界路由器也不会将区域内的链路状态信息全部原样发送给其他区域，只会发送自己到达这些路由器的距离信息，内部路由器所持有的网络拓扑数据库就会明显变小。

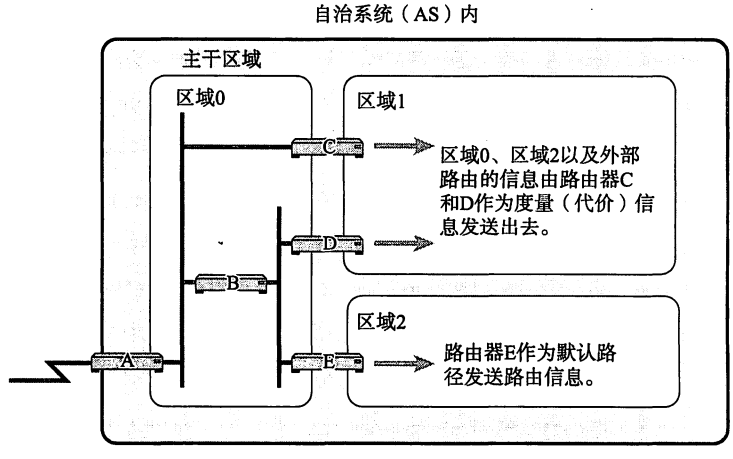
图7.17
OSPF的路由器种类



① 内部路由器
② 区域边界路由器
③ 主干路由器
④ AS边界路由器兼主干路由器
⑤ AS边界路由器兼内部路由器

换句话，就是指内部路由器只了解区域内部的链路状态信息，并在该信息的基础上计算出路由控制表。这种机制不仅可以有效地减少路由控制信息，还能减轻处理的负担。

图7.18
区域内的路由控制和区域之间的路由控制



此外，作为区域出口的区域边界路由器若只有一个的话叫做末端区域（如图7.18中的区域2）。末端区域内不需要发送区域外的路由信息。它的区域边界路由器（在本图中为路由器E）将成为默认路径传送路由信息即可。因此，由于不需要了解到达其他各个网络的距离，所以它可以减少一定地路由信息。

要想在OSPF中构造一个稳定的网络，物理设计和区域设计同样重要。如果区域设计不合理，就有可能无法充分发挥OSPF的优势。

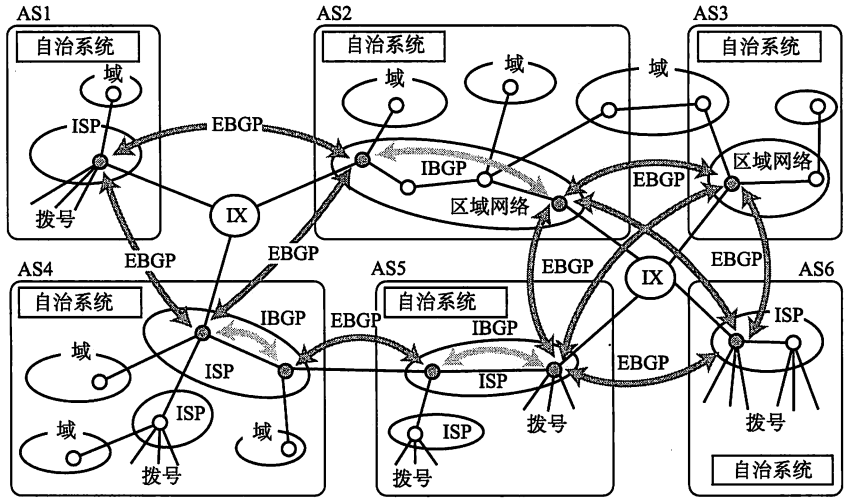
7.6 BGP

BGP（Border Gateway Protocol），边界网关协议是连接不同组织机构（或者说连接不同自治系统）的一种协议。因此，它属于外部网关协议（EGP）。具体划分，它主要用于ISP之间相连接的部分。只有BGP、RIP和OSPF共同进行路由控制，才能够进行整个互联网的路由控制。

7.6.1 BGP与AS号

在RIP和OSPF中利用IP的网络地址部分进行着路由控制，然而BGP则需要放眼整个互联网进行路由控制。BGP的最终路由控制表由网络地址和下一站的路由器组来表示，不过它会根据所要经过的AS个数进行路由控制。

图7.19
BGP使用AS号管理网络信息



ISP、区域网络等会将每个网络域编配成一个个自治系统（AS：Autonomous System）进行管理。它们为每个自治系统分配一个16比特的AS编号▼。BGP就是根据这个编号进行相应的路由控制。

▼在日本由JPNIC管理着这些AS编号。

由JPNIC管理的AS编号一览可以从如下网站获取：

```
http://www.nic.ad.jp/ja/ip/as-numbers.txt
```

有了AS编号的域，就相当于有了自己一个独立的“国家”。AS的代表可以决定AS内部的网络运营和相关决策。与其他AS相连的时候，可以像一位“外交官”一样签署合约再进行连接▼。正是有了这些不同地区的AS通过签约的相互连接，才有了今天全球范围内的互联网。

▼也叫对接（Peering）。

举一个例子，如图7.19所示，为了使AS1与AS3之间能够进行通信，需要

有 AS2 或者 AS4 与 AS5 组合起来的两者中的一者进行数据中转▼才能够实现。而这两者之间是否中转则由它们自己，即 AS2 或 AS4 与 AS5 决定▼。如果两者都不愿意中转，那么只能在 AS1 与 AS3 之间建立专线连接才能实现通信。

以下，我们将假定这两者都允许中转，详细介绍 BGP。

▼也叫转接（Transit）。

▼如果进行中转，就意味着网络负荷的加重以及成本的提升。因此，这种中转合约通常都会涉及中转费用。

7.6.2　BGP 是路径向量协议

根据 BGP 交换路由控制信息的路由器叫做 BGP 扬声器。BGP 扬声器为了在 AS 之间交换 BGP 信息，必须与所有 AS 建立对等的 BGP 连接。此外，如图 7.20 中的自治系统 AS2、AS4、AS5，它们在同一个 AS 内部有多个 BGP 扬声器。在这种情况下，为了使 AS 内部也可以交换 BGP 信息，就需要建立 BGP 连接。

BGP 中数据包送达目标网络时，会生成一个中途经过所有 AS 的编号列表。这个表格也叫做 AS 路径信息访问列表（AS Path List）。如果针对同一个目标地址出现多条路径时，BGP 会从 AS 路径信息访问列表中选择一个较短的路由。

在做路由选择时使用的度量，RIP 中表示为路由器个数，OSPF 中表示为每个子网的成本，而 BGP 则用 AS 进行度量标准。RIP 和 OSPF 本着提高转发效率为目的，考虑到了网络的跳数和网络的带宽。BGP 则基于 AS 之间的合约进行数据包的转发。BGP 一般选择 AS 数最少的路径，不过仍然要遵循各个 AS 之间签约的细节进行更细粒度的路由选择。

在 AS 路径信息访问列表中不仅包含转发方向和距离，还涵盖了途径所有 AS 的编号。因此它不是一个距离向量型协议。此外，对网络构造仅用一元化表示，因此也不属于链路状态型协议。像 BGP 这种根据所要经过的路径信息访问列表进行路由控制的协议属于路径向量（Path Vector）型协议。作为距离向量型的 RIP 协议，因为无法检测出环路，所以可能发生无限计数的问题▼。而路径向量型由于能够检测出环路，避免了无线计数的问题，所以令网络更容易进入一个稳定的状态。同时，它还有支持策略路由▼的优势。

▼路由进入稳定状态需要一定时间、网络跳数不可超过 15 等限制，导致无法应用于大型的网络等问题。

▼策略路由控制是指在发送数据包时，可以选择或指定所要通过的 AS 的意思。

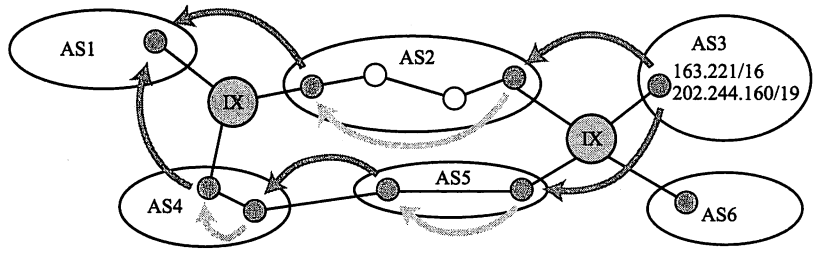


图 7.20　生成路由控制表时要用到 AS 路径信息访问列表

从邻接AS收到的AS路径信息访问列表中加入自己的AS编号，再发送给自己邻接的AS。

AS1到AS3的AS路径信息访问列表（AS Path List）

163.221/16	AS 2 — AS 3
	AS 4 — AS 5 — AS 3
202.244.160/19	AS 2 — AS 3
	AS 4 — AS 5 — AS 3

AS1 到AS3的AS路由控制表

163.221/16	→ AS2的路由器的IP地址
202.244.160/19	→ AS2的路由器的IP地址

通常选择较短的一方。

▩ 路由控制是跨越整个互联网的分布式系统

分布式系统是指多个系统协同完成一个特定任务的系统。

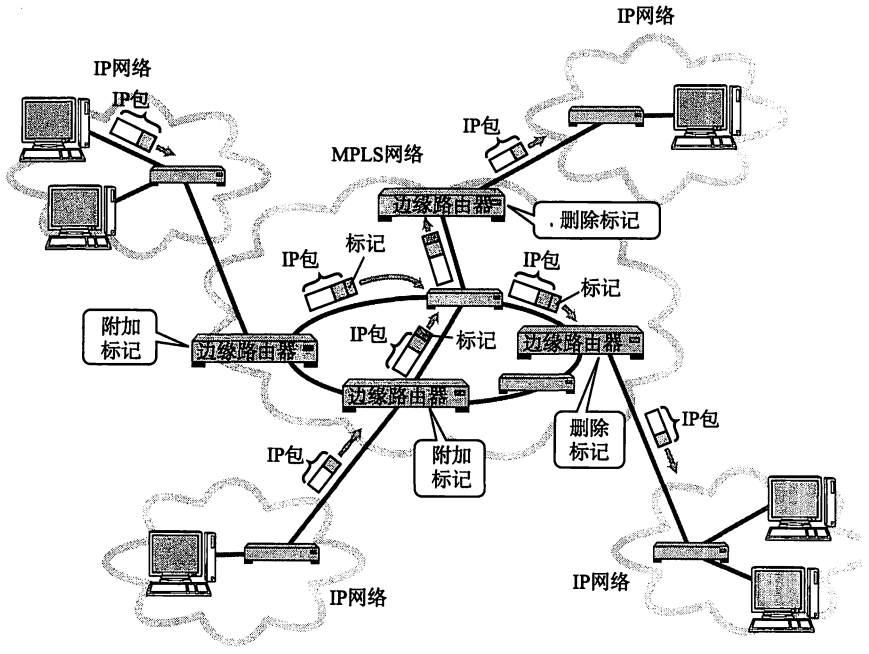
互联网中的路由控制，以网络内所有路由器都持有正确的路由信息为基础。使这些路由器的信息保持准确的协议就是路由协议。没有这些路由协议协同工作，就无法进行互联网上正确的路由控制。

总之，路由协议散布于互联网的各个角落，是支撑互联网正常运行的一个巨大的分布式系统。

7.7 MPLS

现如今，在转发IP数据包的过程中除了使用路由技术外，还在使用标记交换技术。路由技术基于IP地址中最长匹配原则进行转发，而标记交换则对每个IP包都设定一个叫做“标记”的值，然后根据这个“标记”再进行转发。标记交换技术中最具代表性的当属多协议标记交换技术，即MPLS（Multi Protocol Label Switching）。

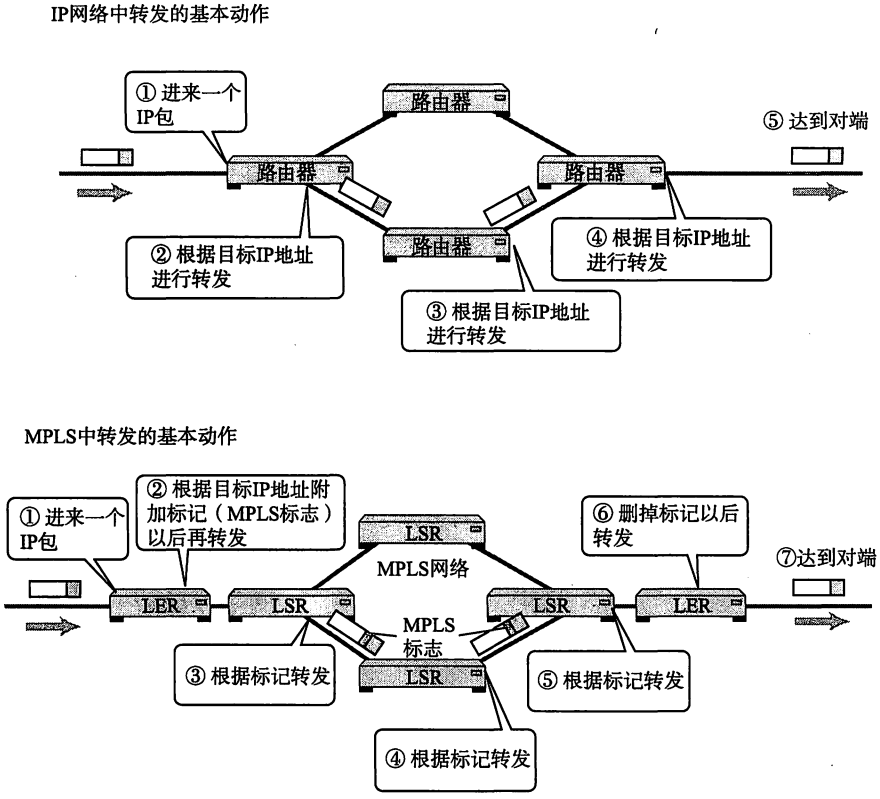
图7.21
MPLS网络



MPLS的标记不像MAC地址直接对应到硬件设备。因此，MPLS不需要具备以太网或ATM等数据链路层协议的作用，而只需要关注它与下面一层IP层之间的功能和协议即可。

由于基于标记的转发通常无法在路由器上进行，所以MPLS也就无法被整个互联网采用。如图7.22所示，它的转发处理方式甚至与IP网也有所不同。

图 7.22
IP 与 MPLS 转发的基本行为对比



7.7.1 MPLS 的网络基本动作

MPLS 网络中实现 MPLS 功能的路由器叫做标记交换路由器（LSR，Label Switching Router）。特别是与外部网路连接的那部分 LSR 叫做标记边缘路由器（LER，Label Edge Router）。MPLS 正是在 LER 上对数据包进行追加标记和删除标记的操作。

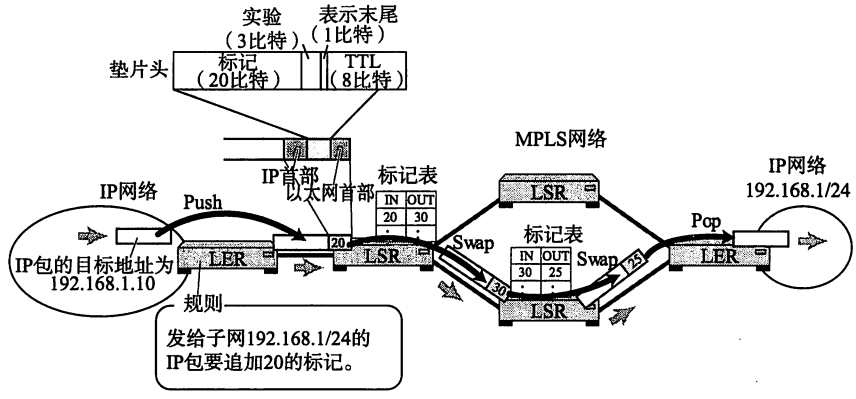
在一个数据包上附加标记是一个及其简单的动作。如果数据链路本来就有一个相当于标记的信息，那么可以直接进行映射。如果数据链路中没有携带任何相当于标记的信息（最典型的就是以太网），那么就需要追加一个全新的垫片头（Shim Header）。这个垫片头中就包含标记信息▼。

▼垫片头像个楔子一样介于 IP 首部与数据链路首部之间。

如图 7.23 展示了数据从以太网的 IP 网开始经过 MPLS 网再发送给其他 IP 网的整个转发过程。数据包在进入 MPLS 时，在其 IP 首部的前面被追加了 32 比特的垫片头（其中包含 20 比特的标记值）▼。MPLS 网络内，根据垫片头中的标记进一步进行转发。当数据离开 MPLS 时，垫片头就被去除。在此我们称附加标记转发的动作为 Push，替换标记转发的动作为 Swap，去掉标记转发的动作为 Pop。

▼有时也可能会被追加多个垫片头。

图 7.23
使用 Push、Swap 和 Pop 功能进行转发

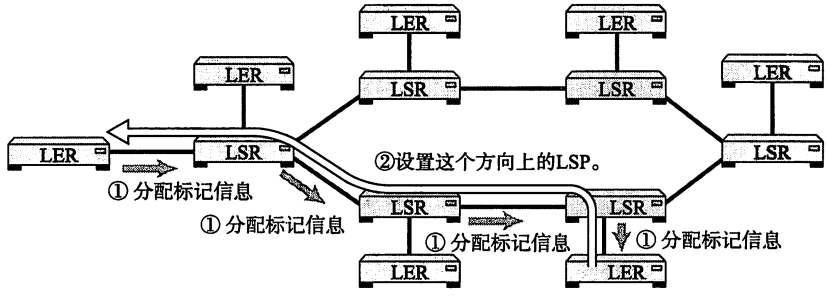


MPLS 中目标地址和数据包▼都要通过由标记决定的同一个路径，这个路径叫做标记交换路径（LSP，Label Switch Path）。LSP 又可以划分为一对一连接的点对点 LSP，和一对多绑定的合并 LSP 两类。

▼它们被称作 FEC（Forwarding Equivalence Class），是指具有相同特性的报文。

扩展 LSP 有两种方式。可以通过各个 LSR 向自己邻接的 LSR 分配 MPLS 标记，也可以由路由协议载着标记信息进行交互。LSP 属于单方向的通路，如果需要双向的通信则需要两个 LSP。

图 7.24
根据 MPLS 标记信息的分配情况设置的 LSP



·LSR之间进行标记信息交换时有两种方法：一是采用标记分配协议（LDP，Label Distribution Protocol）的方法；另一种是通过路由协议捎带信息的方法。本图展示了每个LSR独立生成标记表并将这个表传给上游LSR再进行分配的过程。

7.7.2　MPLS 的优点

MPLS 的优势可归纳为两点。第一个是转发速度快。通常，路由器转发 IP 数据包时，首先要对目标地址和路由控制表中可变长的网络地址进行比较，然后从中选出最长匹配的路径才能进行转发。MPLS 则不然。它使用固定长度的标记信息，使得处理更加简单，可以通过高速的硬件实现转发▼。此外，相比互联网中的主干路由器需要保存大量路由表才能进行处理的现状，MPLS 只需要设置必要的几处信息即可，所要处理的数据量也大幅度减少。而且除了 IPv4、IPv6 之外，针对其他协议，MPLS 仍然可以实现高速转发。

▼现在的路由器也更趋向于硬件化。

第二个优势在于利用标记生成虚拟的路径，并在它的上面实现 IP 等数据包的通信。基于这些特点，被称之为“尽力而为”（Best-Effort▼）的 IP 网也可以提供基于 MPLS 的通信质量控制、带宽保证和 VPN 等功能。

▼尽力而为服务是尽自己最大努力提供服务意思。具体请参考 4.2.4 节的最后部分。