第1章　网络 IGP 路由设计

IGP 是指在一个自治域（Autonomous System）系统内部运行的内部路由协议。对于一个运营商来讲，BGP 协议用来控制 AS 之间的流量，整体调控流量如何到达其他 AS 已经如何通过多个入口进入本 AS；而 IGP 则控制的是 AS 内部的流量，控制这些流量如何穿越内部的链路和设备到达另外网关最终离开本 AS，已经进入本 AS 的流量如何到达网内的其他目的地。

IGP 的设计不是独立存在的，其很大程度上依赖于物理拓扑，所以一个好的 IGP 设计一定是建立在好的物理层设计之上的，与物理层的设计一起完成流量在网络内部穿越的规划。这就像城市道路建设和交通管理指挥，网络物理层规划等同于城市道路建设，而 IGP 规划设计就是交通管理指挥了。这两种是需要统一规划的，否则在不合理的道路建设下再怎么进行交通管理也没有用。网络的 IGP 规划将通过 IGP 本身的特性如分层、分域和设置链路属性等方式引导流量按照规划的路径来穿越网络。

IGP 动态路由协议可以分为两类：距离矢量和链路状态路由协议。RIP、RIPv2 和 IGRP 均是距离矢量路由协议的代表，OSPF、IS-IS 和 EIGRP 则都是链路状态路由协议。由于本书讨论的是大型网络设计，所以距离矢量路由协议不在本书的讨论范围。EIGRP 虽然是链路状态路由协议，但它属于私有协议范畴，所以也不适合作为一个开发标准。基于公共协议的标准是进行大型网络设计时的第一标准，否则以后的扩容发展只会是痛苦的抉择过程。

那么，我们现在的 IGP 选择无可厚非地只剩下两种：OSPF 和 IS-IS。

1.1　OSPF 和 IS-IS 的选择

关于选择 OSPF 还是 IS-IS，这样的争论已经发生过无数次，我相信以后还会发生很多。这是两个势均力敌的选项，所以大家才会争论，才会有一些人力挺 OSPF，另外一些人一定要选择 IS-IS。其实这些选择无所谓正确还是错误，只是哪一种协议更适合自己的网络，更适合运营维护。

对大多数人来说，OSPF 更常见一些，因为大多数的企业网络往往采用这种协议。相比 IS-IS，大多数人更加熟悉 OSPF。同时 OSPF 拥有更长的部署历史以及广泛的使用范围，足以变成路由器厂商必须实现的标准。

但是对于运营商级别的网络来讲，使用 IS-IS 的用户会更多一些。很多大的运营商或者企业网络都经历从 OSPF 迁移到 IS-IS 的一个过程。原因有很多种，但最为主要的两个原因是 IS-IS 有更强的可扩展性，具备更多的大型部署实践案例。

是该用 OSPF 还是 IS-IS，没有标准的答案，需要综合考虑网络规模、业务规模种类、以后的规划、运维知识结构等多方面的因素。下面我们来详细了解一下这两种协议。

1.1.1 OSPF 和 IS-IS 的发展历史

最早的计算机路由网络诞生于 1976 年，应用在英国陆军发起的一个名为“韦弗尔”的项目中。晚些时候，在 BBN 加入了区域的概念以适用于分层的网络系统。再后来，该技术被 IS-IS 和 OSPF 采用，成为了目前主力的 IGP 路由协议。

两种协议在 20 世纪 80 年代发展起来，经过了 20 多年的发展日益完善。表 1-1 概括了 OSPF 和 IS-IS 的发展历史。

表 1-1 OSPF 和 IS-IS 发展历史

时间	IS-IS	OSPF
1987	DEC 公司为 CLNP 选用的域内协议	
1988	IP 扩展属性加入到 IS-IS	OSPF 标准工作开始动工
1989	IS-IS 成为 ISO 路由协议	OSPF v1 RFC1131 发布
1990	支持 ISO 和 IP 的 IS-IS RFC1195 发布	
1991	IGP 内全局影响	OSPF v2 RFC1247 发布 Cisco 支持 OSPF
1992	少量 IS-IS 开始部署 Cisco 发布支持 IP 的 IS-IS	OSPF 开始大量部署
1993	Novell 发布支持 IPX 的 IS-IS：NLSP	复杂，需要通过新特性实现
1994	Cisco 开始发布 NLSP	更新版本 OSPF v2 RFC1583 发布

1995	大型运营商开始部署 IS-IS	
1996~1998	IS-IS 用户开始壮大，很多运营商从 OSPF 迁移到 IS-IS。 Juniper 支持 IS-IS	
1999 以后	不断完善、新增各种特性： 流量工程支持、IPv6 等	不断完善、新增各种特性： 流量工程支持支持、IPv6 等

1.1.2 OSPF 和 IS-IS 各自的特点

OSPF 和 IS-IS 同为链路状态路由协议，这两种协议有着非常多的相似之处，同时又由于出身不同而带来了很多不同的特点。掌握了这些异同点，我们才能更好地完成 IGP 的设计工作。

下面先介绍一下 OSPF 和 IS-IS 的相同点，具体如下所示。

- IS-IS 和 OSPF 都是链路状态路由协议的典型代表，每一个区域都维护一个链路状态数据库（LSBD），以 Dijkstra 算法为核心计算出 SPF 树来进行路由加载。
- 由于具有快速收敛、无环路、支持 VLSM、触发更新等特点，IS-IS 和 OSPF 都能很好地支持大型网络。但从全球部署来看，企业采用 OSPF 的多些，IS-IS 主要应用在大型 ISP 网络中。
- IS-IS 和 OSPF 都使用 Hello 报文来维护邻居关系。
- IS-IS 和 OSPF 都采用两级层次化的拓扑结构，设有骨干区域，并且在区域之间提供地址汇总的能力，为网络规划提供了灵活且实际的设计方案。
- IS-IS 和 OSPF 都具备两种链路类型：点对点链路和广播型链路。
- 为了控制链路状态数据库的规模和复杂度，IS-IS 和 OSPF 在广播型链路上都选举特定设备——DIS/DR 来担任数据同步的工作。
- 对于每个链路状态，数据库中的链路状态记录单元都有一个存活时间与之关联，发送该记录的源设备会在其存活时间倒计时到以前发送新的链路状态报文来刷新存活时间计时器。如果存活时间倒计时到 0 时都没有收到新的更新，将从数据库中清除该链路状态报文，不再用做路由计算。

- 处于边缘区域中的路由器，一是通过区域划分，二是通过设置区域类型来减少对路由器资源的需求。IS-IS 可以将区域中的路由器设置为单纯 level-1 类型，OSPF 可以将整个区域设置为 stub、完全 stub 或 NSSA 来减少数据库的大小，同时保证路由的可达性。

OSPF 和 IS-IS 的不同点如下所示。

- IS-IS 协议直接在链路层上运行，报文直接封装在链路层报文中，支持 CLNS 和 IP 协议。OSPF 报文封装在 IP 报文中，协议运作以 IP 地址为基础。
- IS-IS 采用问候报文（Hello）来维护邻居关系。在 IS-IS 中，邻居之间的问候报文间隔不必一样，邻居存活时间（Dead Timer）也是如此。一端可以设置邻居存活时间为 3 秒，而另外一端可以使用 30 秒，只需各自维护各自的状态就好；在做 IS-IS 平滑重启（GR：Graceful Restart）的过程中会用到这个特性。OSPF 相对复杂，必须要求两端的所有计时器一致才能形成邻居关系 。
- 在点对点链路上，IS-IS 默认使用两次握手（Two-way）建立邻居关系，即收到对端的问候报文就认为邻居关系已经建立了，而不去理会对端是否收到本端的问候报文，这样可能存在单通的问题。OSPF 使用的是 3 次握手的机制，确保双方都可见。IETF 发布 RFC3373 来解决 IS-IS 的 3 次握手的邻居建立问题。
- OSPF 通过特殊的区域 Area 0 来定义骨干区，而 IS-IS 通过连续的 L2 路由器来组成骨干区，但 IS-IS 的 Level 2 路由器可以不在一个区域内。在 IS-IS 协议中，一台路由器只能属于一个区域，区域边界在链路上，路由器的链路状态数据库按 Level-1 和 Level-2 分别维护。而 OSPF 按接口来分区域，这样一个路由器可以属于多个区域，该路由器为每个区域维护一个链路状态数据库，这样的路由器称之为区域边界路由器（ABR）。
- IS-IS 协议的伪节点（DIS）选举比较简单，且抢占结果可预见，接口优先级最高的即为伪节点。而为了保证较小的变动，OSPF 协议的指定路由器（DR）选举机制复杂且不可预见。由于 OSPF 协议没有抢占机制，优先级最高的不一定是指定路由器。OSPF 设有备份指定路由器，指定路由器失效后备份指定路由器立即变为正选的。而 IS-IS 没有备份伪节点，如果伪节点失败，则重新选举新的伪节点。

- IS-IS 路由器在一个链路状态报文（LSP）协议报文中插入所有它发布的 IP 前缀信息。如果大于发布链路的最大传输单元（MTU），将进行分片。当前一个 IS-IS 链路状态报文最多分为 255 片，限制到大约只有 30000 个 IP 地址前缀。
- 在 IS-IS 的 Level-2 的区域中， 没有 Area ID 的限制。即使两个位于 Level-2 区域的路由器不在同一区域内也能形成邻居，这一点与 OSPF 要求邻居区域唯一性的严格限制不同。

上面这些不同点是一些无关痛痒的差别，不应该成为选择 IGP 的判断点。

下面两点会作为大型网络中考量 IGP 稳定程度的参考点。由于存在网络规模、网络链路抖动情况等差别，所以两种协议在不同的组合下会各有优劣。

- 在 OSPF 中，每条链路状态广播（LSA）的存活时间最多为 1 小时，刷新时间为 30 分钟；而 IS-IS 的链路状态报文存活时间是从 20 分钟倒计时，刷新时间由各厂家自行设定，改进后的 IS-IS 的 LSP 存活时间最长是 65535 秒，即超过 18 小时。
- OSPF 使用多个 LSA 来描述路由器自身和周边连接的状态，IS-IS 则使用一个 LSP 来描述，所以出现局部信息需要更新的时候，两种协议发送的信息量是不同的：IS-IS 需要重新发送整个 LSP，OSPF 则发送改变的某个 LSA 即可。

下面的这三点是 OSPF 和 IS-IS 在设计时需要考虑的区别。通常，我们会在设计上避免对一种协议的依赖。对于一些极端情况，没有办法通过设计来消除这种区别，那么就没有选择的余地了。例如，由于存在帧中继（Frame Relay）网络，网络必须支持非广播多点（NBMA）链路类型，此时就只能使用 OSPF 了。

- IS-IS 的 Level-1 区域只能对应 OSPF 中的 Total Stub 区域，依赖最近的 Level-1-2 路由器作为该区域的出口，有可能造成次优路由。而 OSPF 非常灵活，普通区域既可保持原状，选择最优路由，也可以设置为 Stub 区域、完全 Stub 区域、NSSA 区域和完全 NSSA 区域。不过 IETF 对 IS-IS 已经提供了一个改进方案，就是将 Level-2 的路由注入到 Level-1 区域中去。

- OSPF 可以很好地支持虚拟链路来修复分开的骨干区域或让隔离开的普通区域连接到骨干区，但 IS-IS 没有这样的实现。
- IS-IS 只支持两种链路类型，而 OSPF 可以支持更多的网络类型，包括非广播多点（NBMA）、点到多点（P2MP）和虚拟链路（Virutal Link） 等。

1.1.3 IS-IS 和 OSPF 在对比上的认知误区

在选择使用 IS-IS 协议还是 OSPF 协议时，我们一定会对比 1.1.2 节提到的种种不同点。对于每一个用户来说，其现有状况、需求和未来趋势等组合形成了独特的使用环境，所以不是每一个不同点都能适用你的使用场景，不要让不相关的标准影响你的最终判断。

- **因为 IS-IS 直接封装在链路层，所以比 OSPF 更安全。**IS-IS 报文直接封装在链路层，OSPF 报文则是标准的 IP 报文，所以从理论上来讲，OSPF 的安全性更差一点。但实际上，在没有物理安全的环境下，两者都不具备安全性。在实际的部署上面，合理的安全策略也可以杜绝目前已知的攻击手段。

相反，一个使用 IS-IS 的实际案例在缺省配置的情况下，由于人为操作失误导致的危害比 OSPF 大得多。

- **OSPF 支持的区域类型更多。**IS-IS 只有一种 Level 1 区域类型，相当于 OSPF 的 Total Stub 区域。但经过不断的演进，配合上 IS-IS 的路由泄露技术，IS-IS 的这么一种 Level 1 区域能实现 OSPF 多种区域的功能。

表 1-2 OSPF 和 IS-IS 区域类型对比

OSPF		IS-IS	
Area 类型	Area 类型解释	IS-IS Level	Level 解释
Area 0	骨干区域	Level 2	骨干区域
普通区域	该区域内存在的路由： • 区域内路由	Level 1 + 路由策略	泄露所有的 IS-IS 路由进入 Level 1： • 区域内自由路由

	• 区域间路由 • 区域间外部路由 • 本区域的外部路由		• 分发进来的区域间路由 • 分发进来的外部路由 • 本区域的外部路由
Stub 区域	该区域内存在的路由： • 区域内路由 • 区域间路由	Level 1 + 路由策略	只泄露 IS-IS 区域间路由进入 Level 1： • 区域内自由路由 • 分发进来的区域间路由
完全 Stub 区域	该区域内存在的路由： • 区域内路由	Level 1 + 人为控制	人为不发布外部路由进入 Level 1
NSSA	该区域内存在的路由： • 区域内路由 • 区域间路由 • 本区域的外部路由	Level 1 + 路由策略	只泄露 IS-IS 区域间路由进入 Level 1: • 分发进来的区域间路由
完全 NSSA	该区域内存在的路由： • 区域内路由 • 本区域的外部路由	Level 1	缺省 Level 1

- OSPF 支持的链路类型更多。回头看看 15 年前的网络，那个时候 ATM 和帧中继（Frame Relay）正在盛行，为了考虑灵活地支持 ATM 和帧中继，OSPF 协议使用了两种特殊的链路类型——P2MP 和 NBMA。在当今的网络中，骨干链路是 Ethernet 和 POS 的天下，如果你的网络里面没有 P2MP 和 NBMA 的特殊要求的话，OSPF 的这个优点对你就没有任何好处。

- IS-IS 比 OSPF 扩展性更好。每台 OSPF 路由器的每条链路都有一条 LSA，该链路变化了，只需要重新泛洪该条 LSA。而对于 IS-IS 来说，IS-IS 路由器的所有链路都在一个 LSP 里，如果一条链路变化了，即使其他链路无变化，就需要泛洪所有链路的信息，从而引起的计算量也是针对该路由器全部的链路。

就网络拓扑而言，如果是前缀的改变，则对 IS-IS 有利，如果是链路的改变，则对 OSPF 有利，而实际中，链路不变，只改地址前缀的情况并不多。比如，由 100 台路由器组成的网络，OSPF 和 IS-IS 都可以用。IS-IS 在理论上只是支持前缀数大，而不是支持链路数大，而一个网络的规模其实很大程度上是由链路数来决定的。

1.1.4 OSPF/IS-IS 实际应用

由于历史原因，IS-IS 在大型网络中的使用较为广泛。目前，国内几乎所有大型骨干网络均运行 IS-IS 协议。在国际上，很多一级运营商也运行 IS-IS。下面按北美地区、大洋洲地区、中国地区和亚洲其他国家来概要介绍一下。

- 北美地区：美国的一级运营商大都使用 IS-IS 作为 IGP 路由协议，只有个别运营商使用 OSPF。
- 大洋洲地区：由于人口密度稀少，网络规模相对来说比较小，大洋洲的运营商都使用 OSPF，没有运营商使用 IS-IS。
- 中国地区：中国的大型运营商都使用 IS-IS 作为 IGP 协议，很多运营商并不是一开始就使用 IS-IS 的，而是先使用 OSPF 然后再迁移到 IS-IS 的。企业用户和金融用户大都使用 OSPF 协议。
- 亚洲其他国家：亚洲的其他国家大多使用 IS-IS 作为 IGP 协议。

1.2 分区设计

无论使用 OSPF 还是 IS-IS，骨干区域一定是需要的，所以首先一定要有 Area 0 或者 IS-IS Level 21。是不是需要再划分更多的区域？这可以有很多原因和考虑，如果你认为某一点是异常重要的因素，那么它可能会成为决定分区的主因。下面我们就详细讲解一下哪些是考虑分区的评估点。

1.2.1 单区设计和分区设计的对比

在我们对 IGP 进行进一步设计的时候，第一个问题就是是否分区。

可能你还在犹豫到底是选择 IS-IS 还是 OSPF，我们可以暂时把那个问题放一边，讨论一下分区的问题。因为即使 IS-IS 和 OSPF 有很多不同点，但两种协议大体上还是比较一致的。两者都具备分区的概念，虽然在分区的边界划分上略微有所不同，但不妨碍我们进行网络概要设计。实际上，很多网络实施了 OSPF 到 IS-IS 的 IGP 迁移，它们在逻辑上并未有大的改动，这从另外一个角度佐证了两种协议的类似性。

分区与否各有利弊。对于不同的用户，关系的重点是不一样的，所以最终得出的结论也是不一样的。当然，在网络运维过程中如果发现新的重点导致结论不一样了，可能就会推翻以前的设计。

回答这个问题之前，我们看看需要考虑哪些方面的问题，通过这些问题的权重和结果分析就能获得你需要的答案了。表 1-3 从各个方面将单一区域结构和多区域分区结构进行了对比，提供了一个概要的对比结构。

表 1-3 单一区域和多区域分区的对比

评估点	单一区域结构	多区域分区结构
组织结构	顺应集中管理的组织架构	分成不同的团队管理
可扩展性	受到设备路由引擎的硬件限制	有非常好的扩展性
资源限制	网络内部具备一致的门槛要求	可以将低端设备隔离
路由控制	路由发起端控制	路由发起端和区域边界控制
路由安全性	IGP 内全局影响	可以控制在区域内部

[1] 纯理论角度来看，OSPF 可以没有 Area 0，IS-IS 可以只运行 Level 1，但这都不是我们设计一个好网络应该考虑的实现。

收敛时间	具备一致性的路由更新	区分不同的事件
流量工程 TE	容易实现端到端的 TE	复杂，需要通过新特性实现
运维管理	简单	较为复杂
路由优化性	IGP 内最优化路由	可能存在次优化路由问题

针对表 1-3 概要的对比描述，下面将详细从这 9 个方面来阐述两种方式的区别、优劣以及实现的难易程度等，这样设计者可以挑选适合自己网络的衡量权重，最终挑选出合理的区域架构。

1. 组织结构

人的因素在任何地方都是第一要素，网络设计也是。网络出现融合、拆分、组织架构调整等，可能都会影响网络的运维，从而改变网络设计。如果同一个网络由多个团队的人去运维和管理，那么分区就变得非常合理了。在国内的大型网络中，常常存在一个 IGP 里面的设备会由总部团队和省或者大区团队各自负责一部分的情况，此时往往是总部负责核心设备，省内团队负责该省所属的设备。对于这种管理架构，为了与核心区域区别开，每一个省或者大区就需要各自分配一个路由区域。

2. 可扩展性

在未来 5 年，你的网络会变成多大规模？在你的设计中，参与 IGP 运行的设备数量会有多少？在实际的部署案例里面，IS-IS 中单一分层目前最多运行了 2000 多台设备，OSPF 中有超过 500 台设备在单一区域里面的案例。这只是一个参考数字而已，如果你的网络设备使用的都是高级路由引擎，那么单一区域里面的设备还能多一些。当然，如果有很多比较低端的设备运行在你的网络里面，单一域里面能够容纳的设备就小很多。

需要注意的是，这里统计的是参与 IGP 路由的设备数量。如果网络在边缘使用很多静态路由设备，那么那些设备不在你的衡量范围内。

这一点比较简单，小于这个数字就不需要考虑扩展性问题了，大于这些数字就需要考虑分区了。

3. 资源限制

路由引擎是负责路由计算的，任何网络内的拓扑变化都会触发相应区域内的设备重新进行 SPF 计算。区域内的设备数量将直接影响到路由引擎完成计算的时间，从而影响网络收敛的时间。对于低端设备，如果将其放在一个只有很少设备的单独区域内，它们就不会影响网络整体的收敛。

对比 10 年前的网络，那个时候多是分区域的设计，其中一个最重要的原因是设备的硬件限制。下面我们看看 10 年前高端路由器的顶级路由引擎的硬件指标：

- 200MHZ RISC5000 CPU
- 256MB CPU 内存
- 512KB 二级 Cache

现在，路由器的路由引擎和主流 CPU 相差无几，那么区别显而易见。

总结这一点，如果网络内都是高端路由引擎，那么单一区域不会有影响，但如果存在很多低端设备，就需要结合网络规模来综合考虑分区设计了。

4. 路由控制

单一区域内所有设备的拓扑信息是一致的，所以对于单一区域来说，路由的控制点就是发起路由的设备。对于分区网络设计，如果使用汇总路由来屏蔽详细路由的话，路由的控制点会多一个——边界路由器，这个多出来的控制点就可以隔离相互直接的影响。对于不同团队管理的同一个 IGP 网络的不同区域来讲，这个控制点往往是他们需要的安全隔离带。

分区的另外一个好处就是：在适当的路由汇总下，区域内的网络波动不会泛及全网，例如某些链路不稳定长期抖动，每一次链路的波动变化都会被限制在本区域内部。

5. 路由安全性

路由的安全考量和上面的路由控制是相关的。在单一区域网络内，如果除了源设备没有任何控制点，那么源设备的问题会影响整个网络；如果整个网络都是在一个团队管理范围内，这时不会存在问题；但

如果是多个团队各自管理一部分的话，这就存在安全隐患了。网络需要额外的控制点来确保错误不会影响全网，所以合理的分区设计可以给网络带来更好的安全性。

6. 收敛时间

到底哪个会有更好的收敛时间呢？这点还不能一概而论。

对于单一区域，所有拓扑变化都是洪泛到所有设备。对于分区的网络，有两种情况：有边界设备执行路由汇总，没有路由汇总。分区网络的边界设备需要再次生产、转发 IS-IS 或者 OSPF 的链路状态报文，此时会带来额外的延迟。对于分区的两种情况，它们和单一区域的对比如下所示。

- 对于分区设计有汇总路由的情况：
  - 影响的路由在汇总内，对于其他区域没有变化，分区收敛更快；
  - 影响的路由不在汇总内，单一区域收敛更快。
- 对于分区设计没有汇总路由的情况，区域边界带来的额外延迟使得收敛变慢。

7. 流量工程 TE

为了支持流量工程（RSVP-TE），OSPF 通过加入了新的 Opaque LSA 来实现，IS-IS 则通过加入新的 TLV 22 和 TLV135 来实现。但这都是在一个区域内部传播的信息，并不能跨域区域边界。单一区域能够获得更好的 TE 部署环境。

TE 只能部署在一个区域内部这个限制已经存在很多年，现在 IETF 已经发布新的标准来解决跨区域的 TE 的实现（RFC 4105）。不过这是相对比较新的标准，目前实际部署异常稀少。

8. 运维管理

对于单一区域的设计，IGP 的配置比较简单，因此运维的复杂度也比较简单。但如果改成多区域的设计，尤其是存在很多路由汇总和路由泄露等配置时，运维的复杂度就加倍增加了，那么出现配置错误的几率就增大了许多。这对于运维管理来说是一个不小的挑战。单一区域新增、更改设备不需要额外配

置，但对于分区域网络来说，除了新增和更改设备以外，还需要对一系列边界路由器进行更改。涉及的设备范围取决于改动的内容和本身的策略配置。

9. 路由优化性

当 IGP 划分为多个区域的时候，区域间使用的就是距离矢量算法，这带来的后果就是可能存在次优化路由问题。通过合理的网络设计，可以尽量避免次优化路由问题，但当网络由于故障而发生了拓扑变化后，就很难确保不会再次出现这个问题。

对于单一区域 IGP 网络，任何时候拓扑信息在所有设备是一致的，不会存在这个问题。

1.2.2 如何进行有效的分区设计

通过上面的分析，你可能已经获得了想要的答案。如果只需要单一区域设计，就可以略过这一小节。现在假设你的答案是需要分区设计，我们来看看如何分区。

分区的大小如何判断，这个很难有统一的标准。按照区内设备引擎能力的高低，网络链路稳定度可以有所不同。如果区内设备有很多低端设备或者链路经常波动，建议区内设备不要超过 100 台；如果区内设备都配备了高端路由引擎，那么区内设备多到 300~500 台也是可行的。

按照什么方式来分区呢？下面会介绍三种方式，实际设计中可能会用其中几种方式的组合。

- **按照地理位置来分区。当网络规模足够大时，单域的过多数量的路由器会影响网络收敛、降低整体性能，通常使用分区来解决这种扩展性问题。对于这种情况，按照物理位置来分区是一种简单易行的分区方法。对于国内网络来讲，可以按照大区为单位来进行网络分区，这对于中型网络就足够了。大区可以这样简单的：华北、西北、西南、华东、华南、华中等。对于大型网络，按照大区划分还是不够，同一个区内的设备还是太多，那么按照省来划分是更好的划分方式。**

图 1-1 是按照地理位置进行分区的示例，其中骨干核心设备在骨干域内（IS-IS Level 2 或者 OSPF Area 0），其他设备按照各自的地理所属的大区/省来分区。对于按照省来进行的分区，我们可以使用省会城市的电话号码区号来作为 Area ID，这样便于日后的运维和设计的交接、理解。

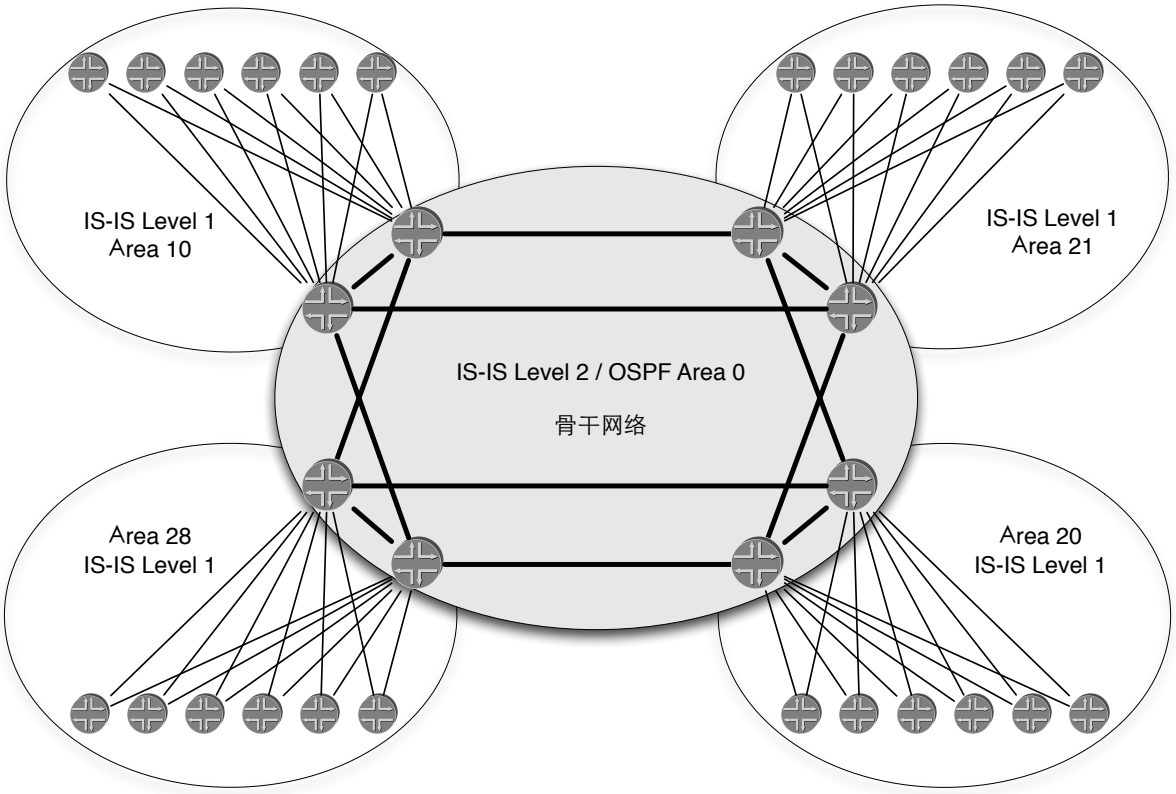

图 1-1 按照地理位置来分区

- **按照业务类型来分区。对于不同的业务来讲，网络都是其基础架构，我们可以将多种业务放到同一张网络中，用 VPN 加以区分，也可以混合在一起使用服务质量（QoS）控制手段来加以区分。但是对于一些相对特别的业务来说，希望相对独立，这个时候就可以通过 IGP 分区的方式来加以隔离。**

例如整个网络主要承载互联网访问业务，但其中一块子网络是为国际业务服务的。该国际业务相对较为独立，这个时候把所有与国际业务相关的路由器划分到一个分区内更能适应业务的发展和管理。图 1-2 展示了这种分区的划分：左边为其国内业务的网络，右边部分是为国际业务服务的设备，这些设备有的位于国内的城市，有的位于海外的节点。

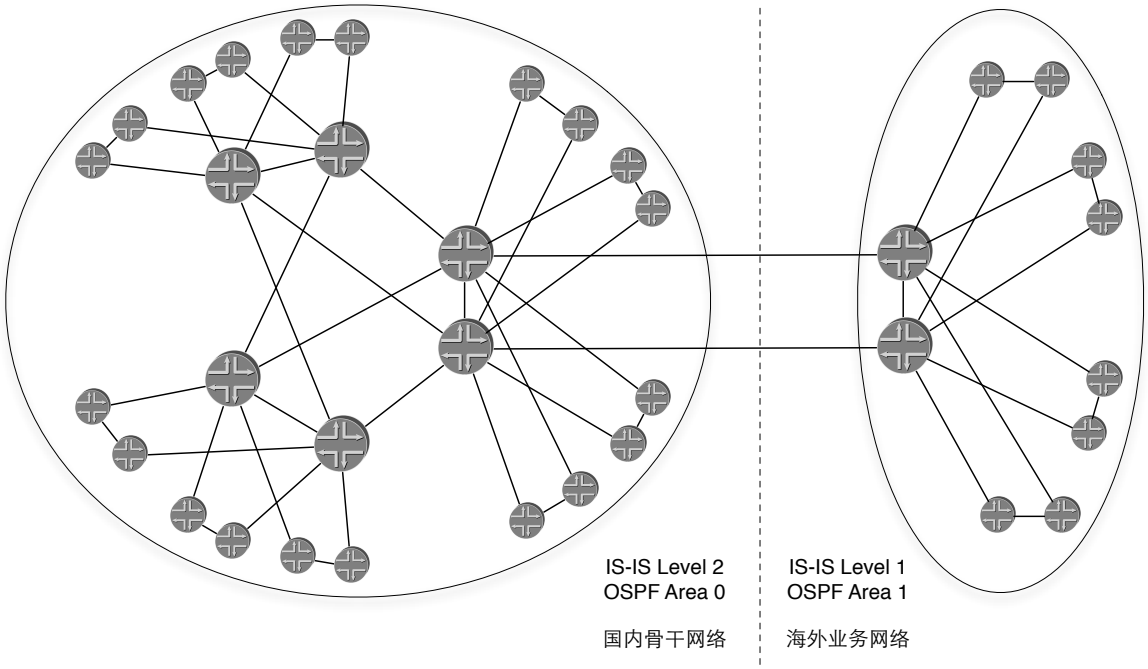

图 1-2 按照业务类型来分区

❑ 按照设备功能类型来分区。当网络承载了多种业务，同时这些业务又分别位于不同的独立设备时，我们就可以用第三种方式来划分区域：按照设备功能类型。网络中某些相同功能的设备划分到一个区域中。图 1-3 为这种分区的描述：该网络具有多个节点，图中下半部左边为其中一个节点，该节点具有核心设备和众多的业务路由器。业务路由器大体分为两类： VPN PE 路由器和专线接入类设备。对于 PE 路由器，由于其特殊要求，将网络内所有 PE 设备划入同一个区域（IS-IS Level 2 或者 OSPF Area 0）会便于操作。对于节点内的其他专线设备——业务路由器（SR），我们可以将其划分在一个独立的分区内。

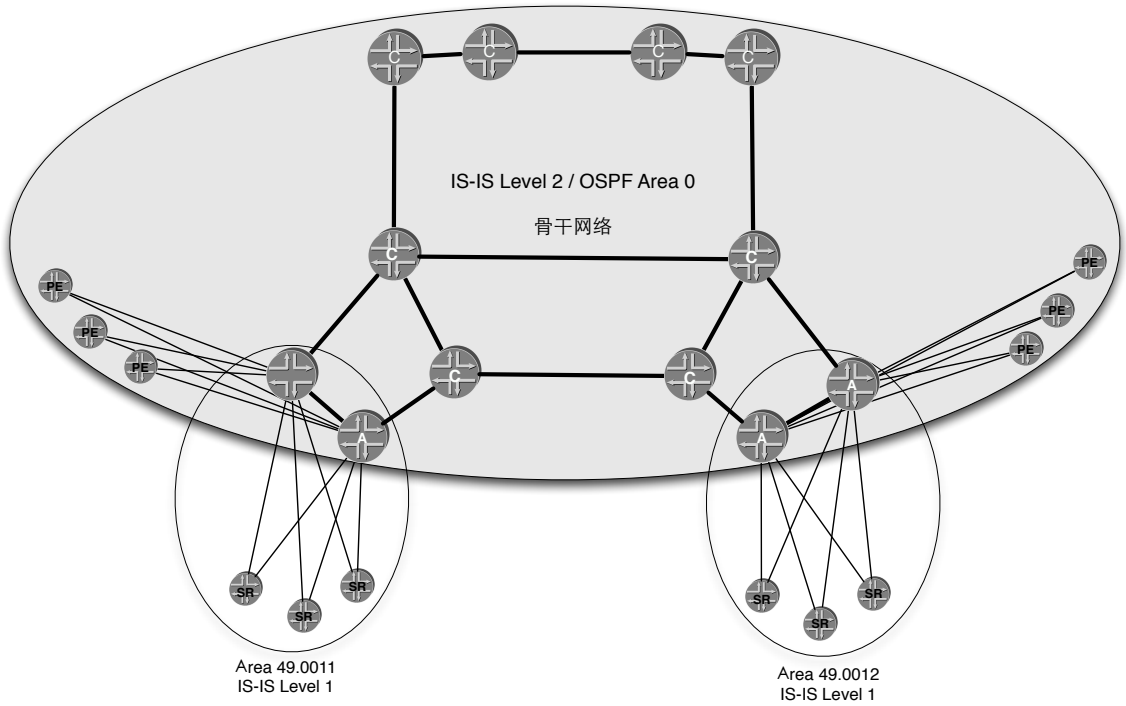

图 1-3 按照设备功能类型来分区

1.2.3 OSPF 和 IS-IS 的分区的选择

由于 OSPF 和 IS-IS 各自的特性不一样，采用分区设计的时候，它们的边界是不一样的。分区意味着统一的数据库分割成了几块，跨区的路由就需要特别考虑，否则就会出现次优化路由问题。

1. 次优化路由问题

当 IGP 划分为多个区域的时候，区域间使用的就是距离矢量算法，这带来的后果就是可能存在次优化路由问题。通过合理的网络设计，可以尽量避免次优化路由问题，但当网络由于故障发生了拓扑变化后，就很难确保不会再次出现。

图 1-4 展示了跨区的次优化路由问题。从源“S”到目的地“D”的最佳路径是通过 C1->C2 之间的链路。但由于 C1 和 C2 之间的链路属于 Area 0，该链路不在 Area 10 的链路数据库中，所以当 C1 路由器收到源“S”的报文以后，实际上会直接转发到 Area 10 中去，而不是使用 Area 0 里面的 GE 链路。通过增加一条属于 Area 10 的 GE 链路互联 C1 和 C2 就可以解决这个问题，如图 1-5 示。

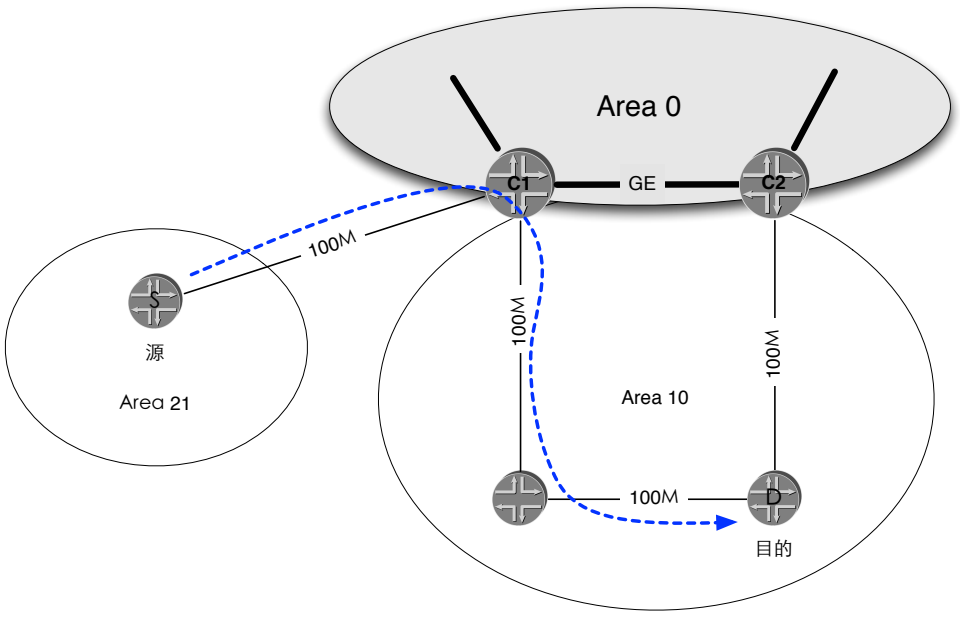

图 1-4 次优化路由问题

对于 IS-IS 来说，解决这个问题比较简单：将图 1-4 中 C1 和 C2 直接的链路同时运行 Level 1 和 Level 2 即可，不需要另外增加一条链路。因为在 OSPF 中，每一个链路只能属于一个 Area，但 IS-IS 中一条链路可以同时运行 Level 1 和 Level 2。

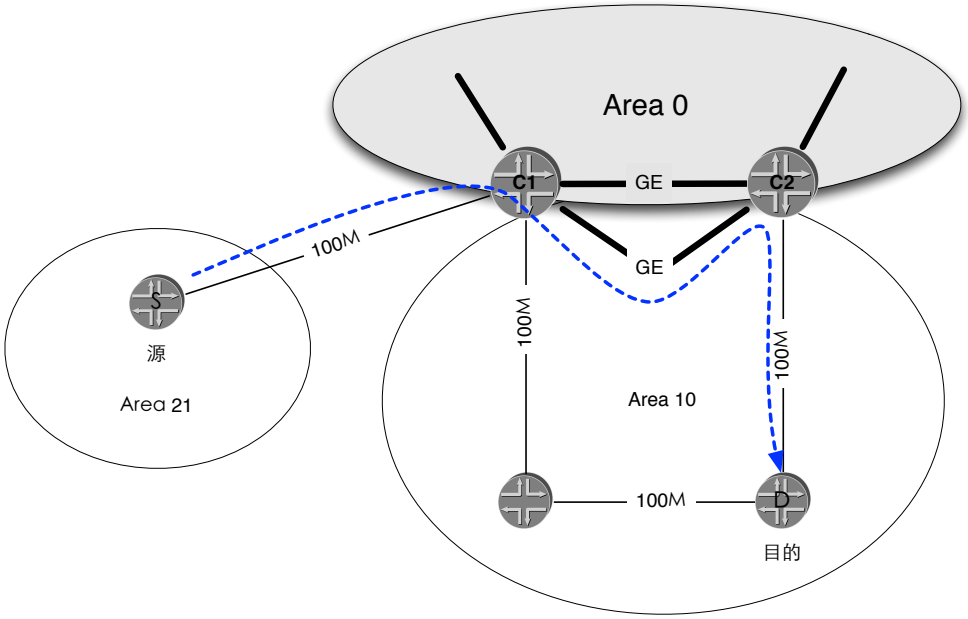

图 1-5 解决次优化路由问题

2. 分区边界的划分

OSPF 和 IS-IS 的区域划分方式不太一样，如图 1-6 所示。OSPF 的单个路由器能同时属于多个区域，但通常情况下 IS-IS 路由器只属于一个区域（IS-IS 多区域的支持并非所有厂家都支持），所以 OSPF 的区域边界是在路由器上，IS-IS 的边界则在链路上。

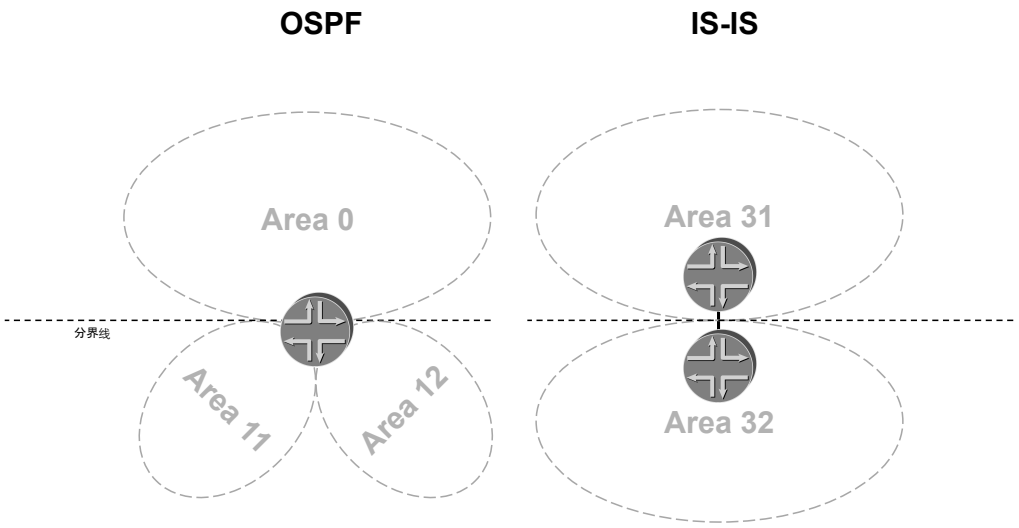

图 1-6 OSPF 和 IS-IS 的边界划分

与 OSPF 不一样，IS-IS 还有分层的概念，参见图 1-7。IS-IS 的 Level 1 和 Level 2 有重复覆盖的区域，但 OSPF 的区域则不会有重复的覆盖。

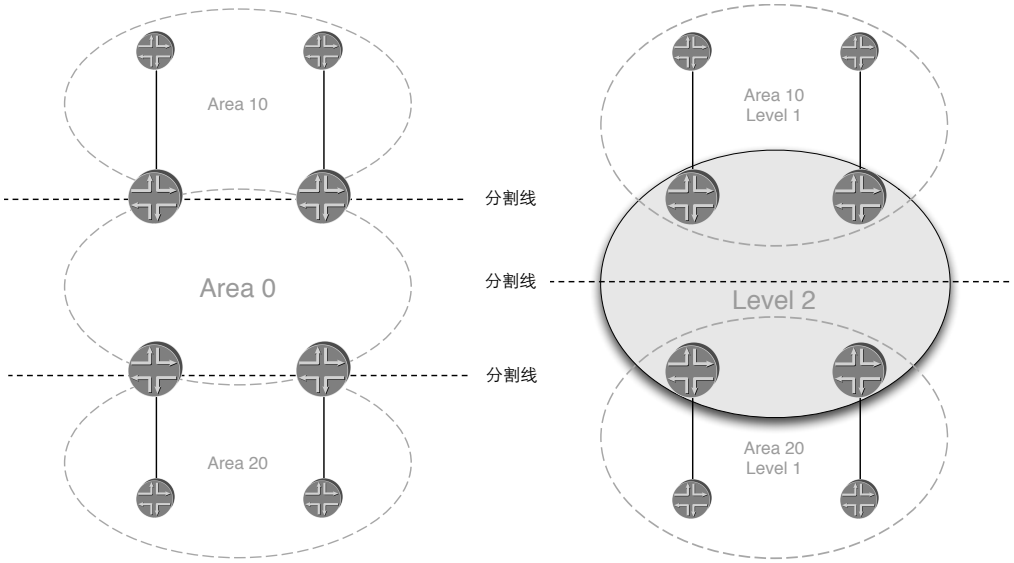

图 1-7 OSPF 和 IS-IS 区域的边界

3. 分区划分的原则

分区划分的原则很简单：1. 分区内拓扑相对完整；2. 跨区链路只应该出现在骨干区域。

分区内应该将拓扑尽量完整，路由器的链路状态数据库是按照每个区域进行计算的，所以设计的时候也需要分区来考虑。如果区内的拓扑不够完整，那么就需要另外添加链路来完善，就像图 1-5 所示的解决方案。

跨区的流量应该被骨干核心设备来承载，所以一定应该将这些设备和链路划分到 OSPF Area 0 和 IS-IS 的 Level 2。对于已经存在的非骨干区域的跨区链路，应该断开或者割接到骨干设备上去。

1.2.4 单区设计和分区设计实际应用

现在运行中的实际网络有很多单一区域的设计，也有很多分区域的设计。有趣的事情是：一些运营商经过若干年的运维、演进和发展，有从多区域迁移到单区域的，当然也有从单区域迁移到多区域的。

下面有两个实际案例分别是单一区域和多区域的案例。

1. 单一区域设计案例

图 1-8 展现的是一个中型全国网络，共有大约 350 台设备，遍布全国 31 个省，共 120 个节点。该网络为典型的 3 层构架：中间为核心层，分布在北京、上海、广州等 8 个中心城市；汇集节点为每个省的两个城市——省会城市以及省内另一重点城市；每个省有 2~8 个接入层节点直接连接省内的汇聚节点。

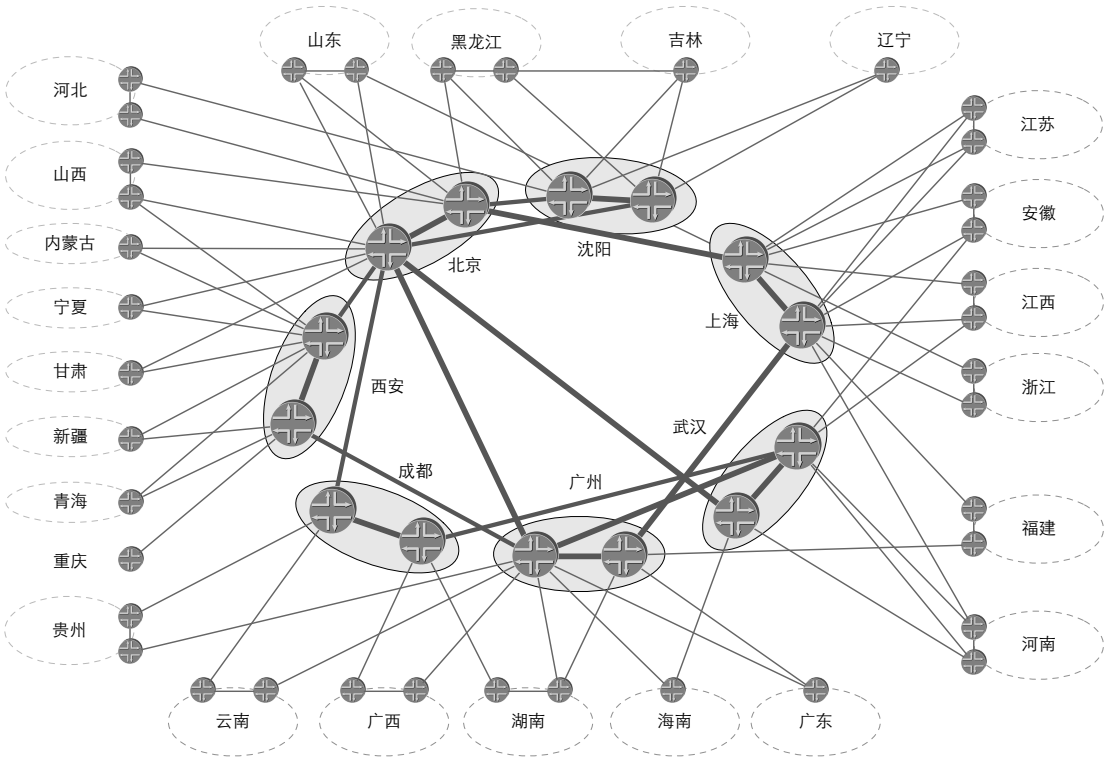

图 1-8 单一 IGP 区域网络案例

网络中的业务路由器需要承载综合业务类型，承载专线业务的同时也要承载 MPLS-VPN 业务，所以将所有路由器放置在一个单一域内可以便于开展全程全网的 MPLS-VPN 业务。由于总路由器数量相对不多，并且都相对是高端设备，所以不会出现扩展性的问题。

全网采用 IS-IS 作为 IGP 协议，所有设备都在 Level 2 中。为了便于以后存在分区的可能性，同处于 Level 2 的设备按照各自地址位置配置了不同的 Area ID。

2. 分区设计案例

图 1-9 是一个位于亚太的运营商的全国网络拓扑图。该网络由于存在较多的低端设备、以及很多低速的不稳定链路，所以最终设计采用分区的 IGP 网络，用以减小区域内的链路数据库尺寸、同时隔离稳定链路的频繁波动。使用 IS-IS 作为 IGP 协议。网络分成了 6 个区域，图中的 1 至 5 个区域是按照地理位置来划分的，所有区域提供同样的业务类型。

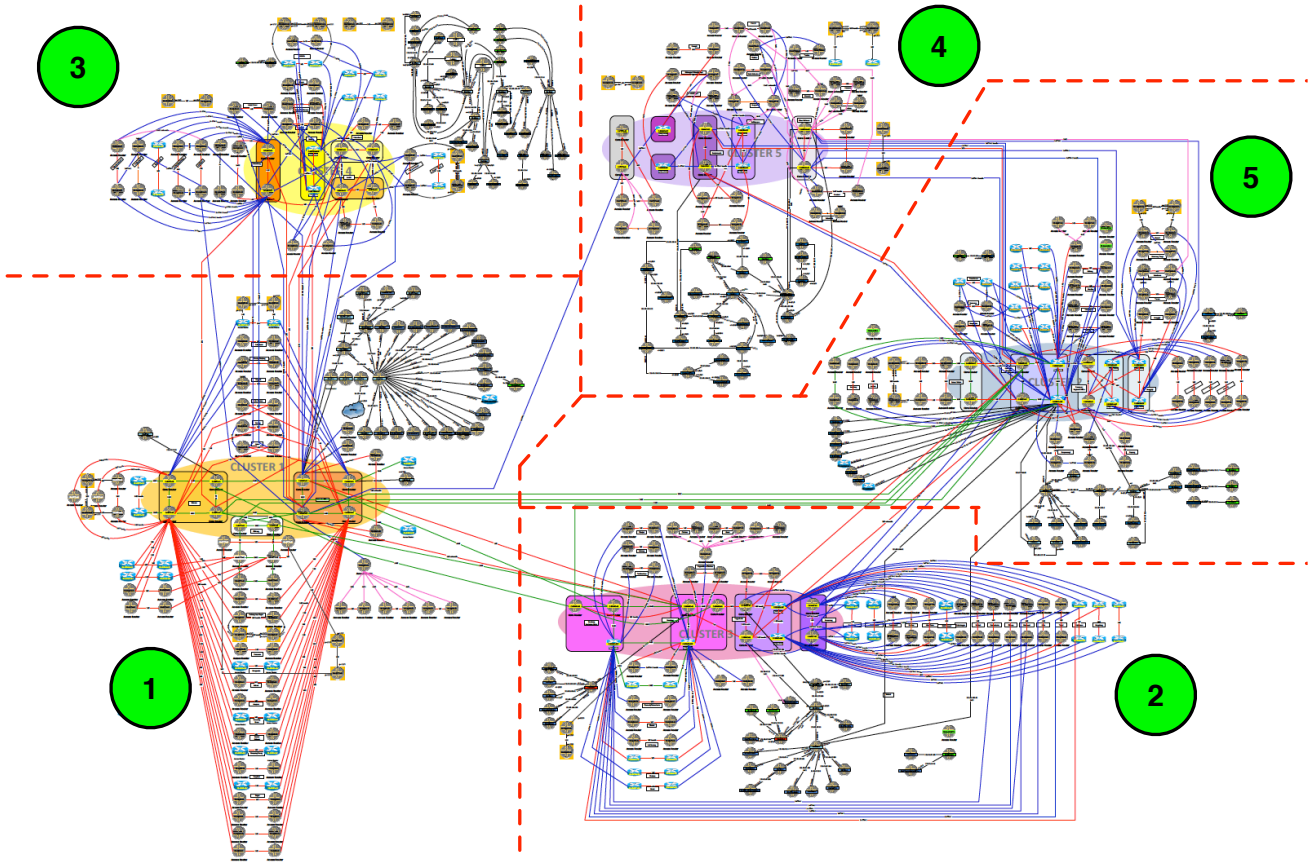

图 1-9 多分区 IGP 网络案例

每一个区域内的核心设备属于 Level 2，MPLS-VPN PE 设备属于 Level 2，其他设备属于 Level 1。区域内所有设备的 Area ID 相同，因此区域内所有设备共享同样的链路拓扑数据库。这里使用了前面章节提及的两种方法来进行了分区：按照设备类型分区的方式和按照地理位置分区的方式。

第 6 个区域没有展现是按照分区类型 1 来实现的，将国际带宽批发这么一钟截然不同的业务单独划分到一个独立的 Level 1 里面去了。所以这个运营商融合前面提到的三种分区方法来进行分区设计。

1.3 链路类型、Metric 值设计

在一个分区内，所有的设备共享同一个链路状态数据库。该数据库描述设备直接是不是有直接链路，已经该链路的 Metric 值。链路的 Metric 值越小，被使用的可能性越大。路由器进行最短路径计算的度量值就是从起点到终点所有链路的 Metric 的总和，所以 Metric 越小的链路将可能承载更多的流量。但链路的 Metric 和链路带宽是可以没有关系的，由此 Metric 的设计需要全网统一考量，融合物理层的链路带宽和整网的流量流向策略来规划。考虑不周全的 Metric 设计会导致人为的网络拥塞。我们不仅要考虑这个 Metric 设计能否在正常状况下工作，还需要考虑这个设计能否应付各种极端的网络异常状况，比如：链路中断、设备断电和光缆中断。

1.3.1 路由协议链路类型的规划

IS-IS 路由协议支持 2 钟基本链路类型：广播型和点对点型。OSPF 除了支持那 2 个以外还支持：点对多点（P2MP）类型、非广播多点类型（NBMA）。

在有选择的情况下，建议尽量不选用点对多点和非广播多点类型。否则等到某一天你需要从 OSPF 迁移到 IS-IS 的时候就会异常的痛苦了；第二，尽量减少广播型链路的使用。

点对点（P2P）链路

这是 IS-IS 和 OSPF 最基本的链路类型，也是在进行最短路径计算的时候最简单的链路类型。因此是 IGP 协议设计中最推荐使用的链路类型。

广播型（Broadcast）链路

当通过以太网端口建立 IGP 邻居关系的时候，IGP 端口类型缺省就变成了广播型链路。这里存在两种情况：

1. 通过交换机互联，多台路由器形成广播网络的关系；
2. 两台设备逻辑上直接互联，没有真正的形成广播网络的关系；

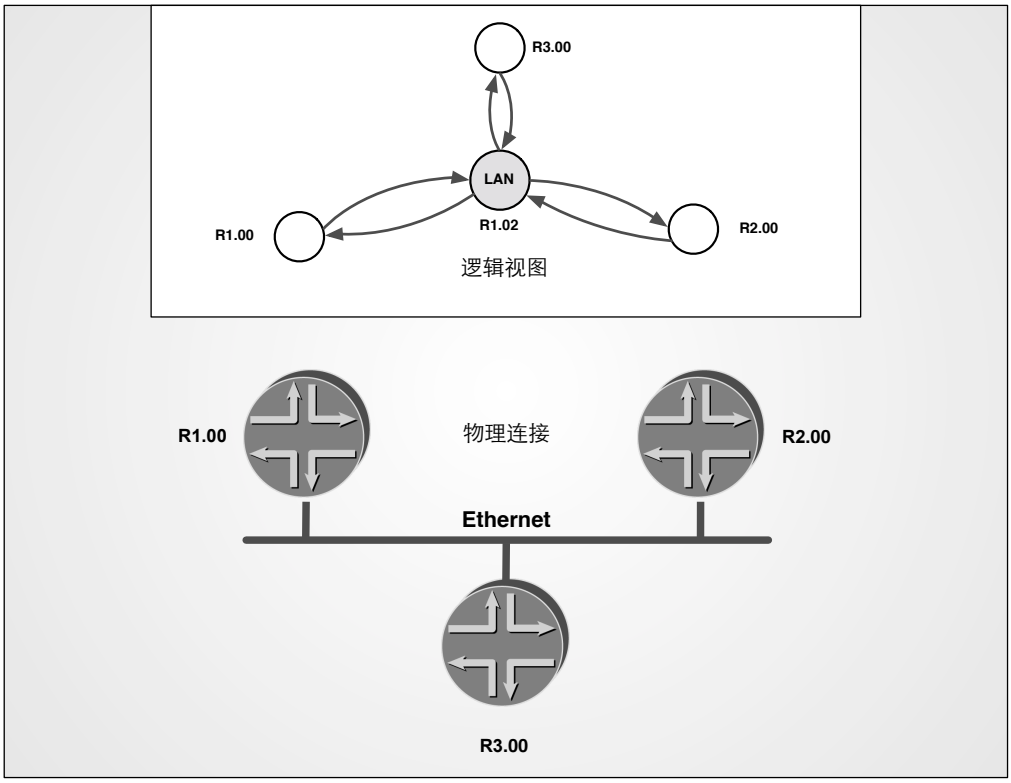

图 1-10 广播型（Broadcast）链路

路由器若使用广播型链路建立邻居关系，如图 1-10 所示了，广播型链路上会选举一个伪节点或者指定路由器作为虚拟节点。所有链路上的其他设备会和这个虚拟节点建立邻居关系，将广播型链路最终模拟成多个点对点链路。

对于第二种情况，其本质还是点对点互联，只是由于使用了以太网端口 IGP 协议缺省认为是广播型链路。这种情况下可以强制设定 IGP 下端口类型为点对点类型来避免伪节点或者指定路由器的选择，从而简化链路数据库的维护和最短路径的计算。图 1-11 中可以看到在逻辑视图上，这个物理上的广播链路实际是点对点链路。

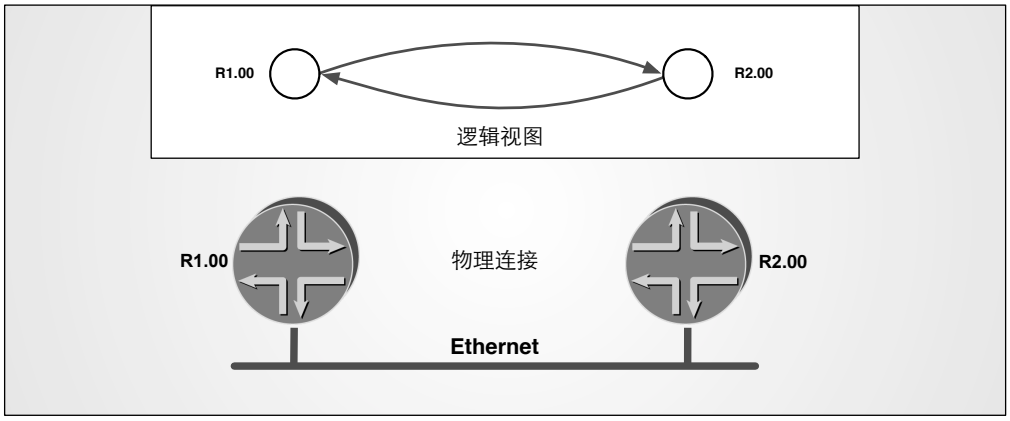

图 1-11 配置广播型链路成为点对点类型

对于第一种情况，最好是从物理设计上避免这种拓扑结构。多台 IGP 路由器通过交换机互联虽然简单，但实际上带来很多问题：

- 流量不可控：N 台设备之间的流量流向是 N(2) 的关系，没有办法有效利用设备之间的端口；对于异常情况下的流量分布是难以预测的，这个带来的直接问题就是异常情况下的网络拥塞，出现了本应通过设计而避免的网络拥堵。
- 服务质量（oS）无法保障：中间经过交换机，容易形成多个端口输出到一个端口的情况，CoS 没有办法保障。
- 增加中间环节、收敛缓慢：通过交换机互联的另外一个最大问题，就是即使对端路由器端口已经失效了，但本端端口由于连接交换机，所以还以为是能用的，整个过程要等 IGP 协议的邻居存活计数器失效才能收敛；这个问题要通过一些额外附加的办法才能解决这个问题，例如：双向转发检测（BFD）。

1.3.2 路由协议 Metric 值的规划种类

IS-IS 和 OSPF 这种链路状态路由协议使用链路状态数据库来计算每个节点直接的距离，每一条链路的 Metric 度量值就成了如何进行选路的关键。而网络的 IGP 设计的一个重点就是要规划如何设定网络内部所有链路的 Metric 值，通过这些设定达到什么样的流量策略，如何实现规划好的流量流向。

IGP 的 Metric 需要和物理链路设计相辅相成、相互作用，其基本划分方式有下面 3 种，实际的运用上可能是某 2 种的组合运用。

1.3.3 缺省的 Metric 值规划

最简单的 Metric 设定莫过于使用参考带宽，让路由器自己算计每一条链路的 Metric 值。如 图 1-12 所示， 假设参考带宽是 100G，那么图上的每种链路的 Metric 为：

- 10G： Metric 10
- 2 * GE： Metric 50
- GE： Metric 100
- 2 * STM1： Metric 323
- STM1： Metric 645

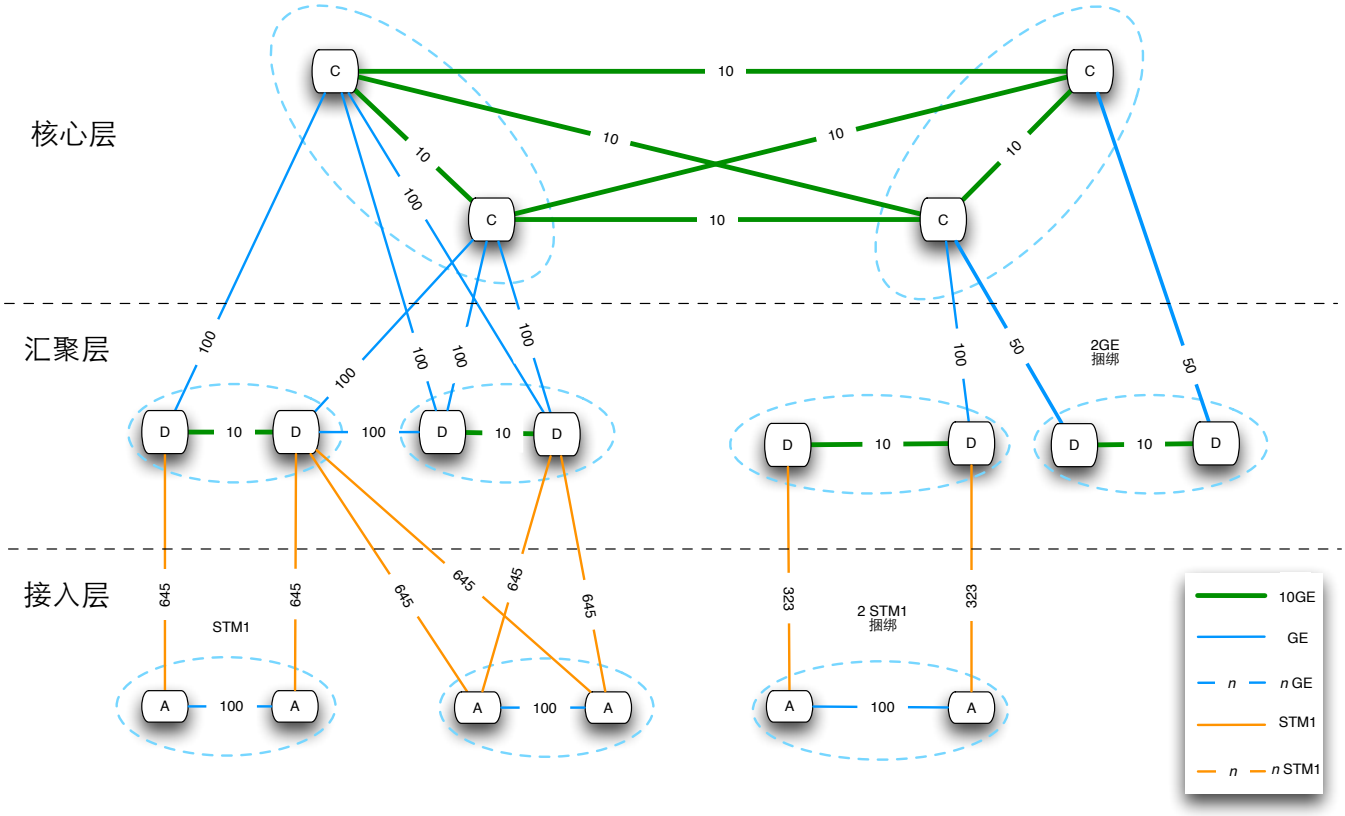

图 1-12 使用参考带宽来设定 IGP Metric

参考带宽值是计算 Metric 的公式的分母，所以必须要求每一台设备的参考带宽值是一致的。

使用参考带宽的好处是简单，但问题却是多多。最大的问题是：由于是自动计算，所以网络的局部链路带宽的变化极有可能影响到全网；对于双上联的节点来说需要对称扩容，如果单边扩容会造成另外一边上联链路不会有流量。

1.3.4 按照链路角色进行的 Metric 值规划

对于大部分的网络拓扑来说，IGP 流量很难通过参考带宽值去实现。设计者需要流量清晰的按照自己设计的路径行进。按照链路角色进行的 Metric 值规划很好的满足了这一需求。这是目前被广泛使用的 IGP Metric 值设计方法。

参加示例网络拓扑图 1-13：典型的 3 层网络结构具备：核心层、汇聚层、接入层；同时按照物理位置有节点内链路，节点间链路。所以整体上我们可以吧所有的链路按照其角色不同分成下面的种类：

表 1-4 按照链路角色来静态设定 Metric

类型	节点内（Intra-Hub）	节点间（Inter-Hub）
核心—核心	50	1000
核心—汇聚	—	1500
汇聚—汇聚	20	1500
汇聚—汇聚（跨区）	—	2500
汇聚—接入	—	3000
接入—接入	10	—

如何来解读这个 Metric 设定？首先是 Metric 设定和链路带宽没有关系了，所以链路的选定需要满足节点流量的要求，这是基本要求；在这个基础上我们才能通过 Metric 来控制流向。

第二、节点内的互联链路的 Metric：级别越高（例如核心节点内），则 Metric 约大。

第三、节点间链路的 Metric：级别越高（例如核心直接），则 Metric 约小。

最后一点、节点间的 Metric 远大于节点内的 Metric 值。

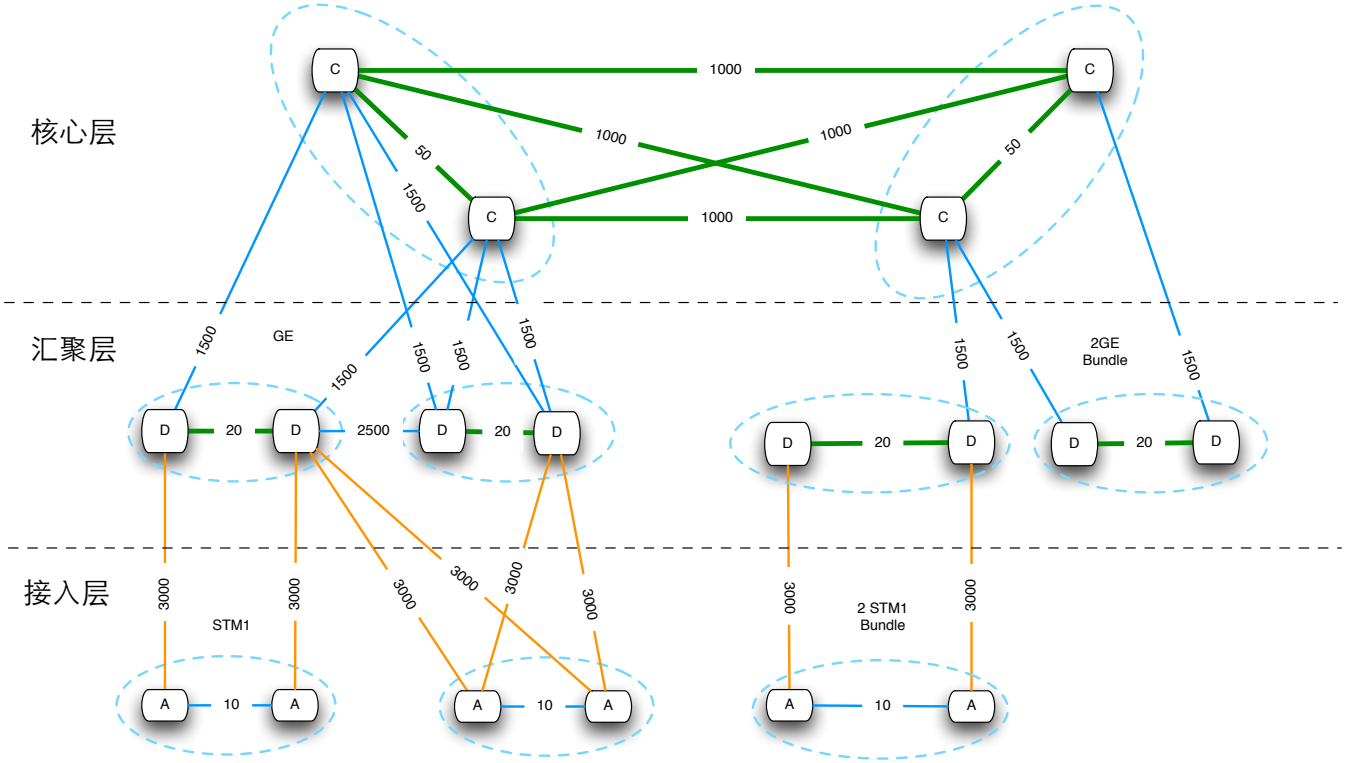

图 1-13 按照链路角色静态配置链路 Metric

在这种 Metric 的设计下的流量流向是什么样子的呢?

1. 骨干直接的跨节点链路讲承载所有跨区的流量;
2. 核心层之间的流量不会通过汇聚层绕行;
3. 汇聚层之间的流量不会通过接入层绕行;
4. 本节点内互联链路优先，如图 1-14 所示，不会加重上面一层的互联压力;

这样设计的好处在于:

1. 各个层次功能分明，不会混淆各自的职责；核心层担负跨区的流量中转，汇聚层担负区内跨节点的流量分发，接入层只负责本节点内的最终到达。

2. 流量流向按照设定路线行进；每条链路承载的流量很清晰；设计者、运维人员可以清楚的知道每条链路会承载的流量、以及不会存在的流量；对于流量监控、扩容规划等可以提供宝贵的数据。
3. 扩容简单易行，任意单链路扩容不会影响整体流向策略；扩容可以按需而来，由于 Metric 都是固定设置，所以增加链路带宽不会改变 IGP 的拓扑数据库，都是相对独立的操作。实现了整体设计的强壮型。

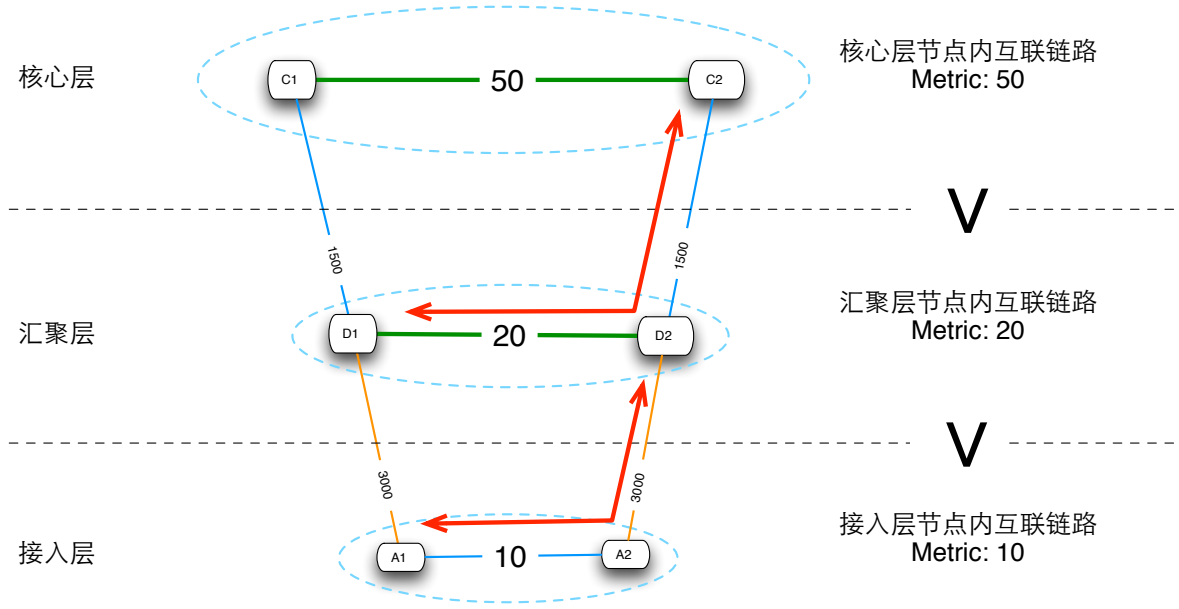

图 1-14 按照链路角色静态节点内互联链路配置 Metric 的示例

图 1-14 所示的目前广泛采用的节点内 Metric 设计方式：对于口字型连接的对角流量，由于：接入层的节点内互联 Metric < 汇聚层的节点内互联 Metric < 核心层的节点内互联 Metric，所以会优选低一级的节点内互联链路。这样做的原因是接入层节点数量巨大、汇聚层相对较多、核心层设备最少，如果把大量的穿透流量置于上一层，极可能导致拥塞。所以才形成这样的 Metric 设计。

1.3.5 按照链路延迟进行的 Metric 值规划

图 1-13 展示的按照链路角色来指定 Metric 的规划方法存在一个相对较大的问题。由于链路角色本身没有距离因素的引入，所以在延迟上不会是最优化的结构。参见图 1-15，按照 1.3.4 描述的方式来设定核心节点直接的 Metric 值。从北京到杭州的流量将会有两条负载均担的路径：1.北京->上海->杭州；2.北京->广州->杭州。但显而易见通过上海的路径延迟会比去穿过广州的路径小的多，我们希望应该优选传输延迟小的路径。

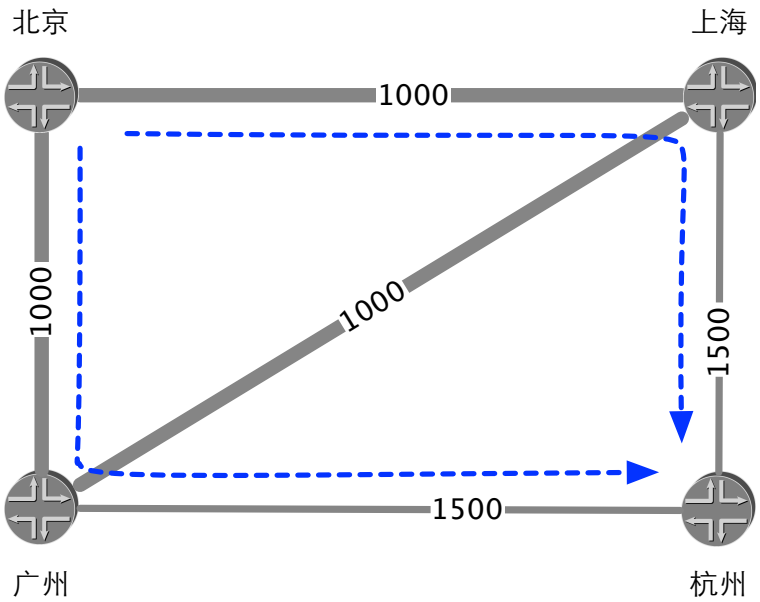

图 1-15 按照链路角色设置 Metric 的问题

基于传输延迟的 Metric 规划方法很好的缓解了这个问题。链路的传输延迟作为生成最终 Metric 的一个重要因子。对于相同角色的链路按照延迟等比例的设定 Metric，参见示例图 1-16：网络核心层的 Metric 设定，虽然都是核心节点之间的链路，但也因为传输延迟的不同而有不同设定。

需要注意的一点是，节点之间的地理距离不是参考的因素，传输延迟才是；因为传输路径可能存在绕道等情况，不会和地理距离等同。

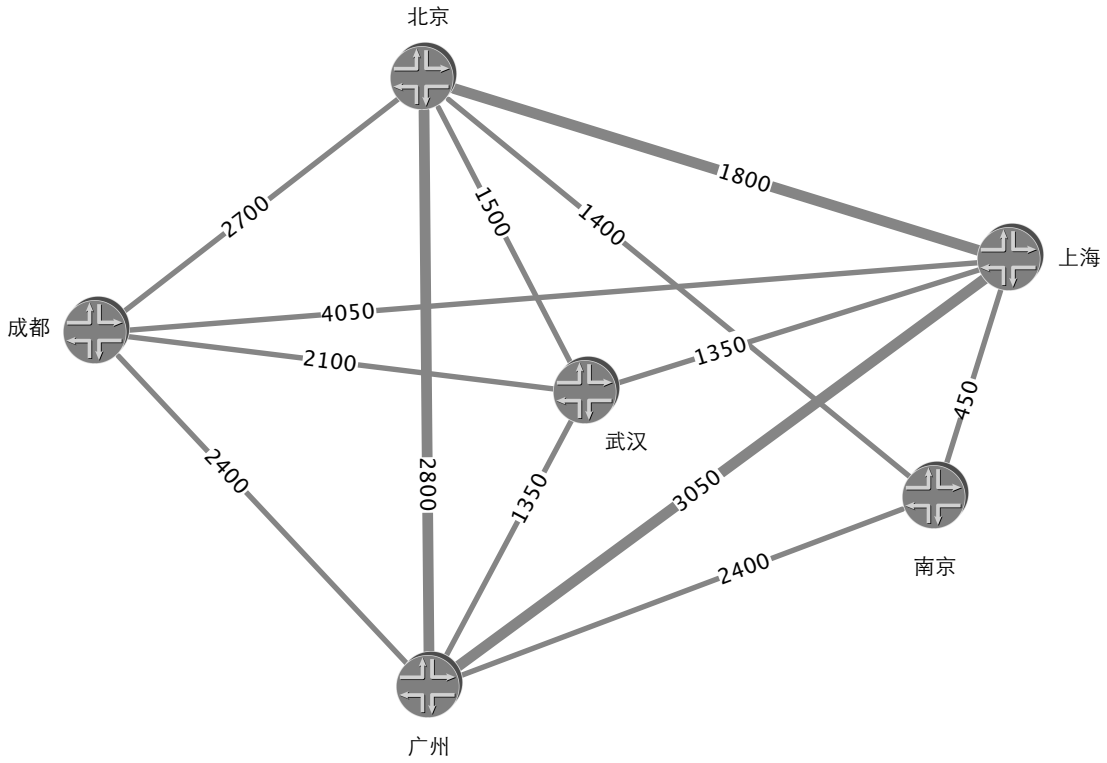

图 1-16 按照传输链路延迟来设计 Metric

采用传输延迟来设计 Metric 的实际应用中，通常是和角色设定组合使用，来兼备两者的优点，表 1-5 是这种组合方式的示例。这种方式只有一个缺点：就是设计复杂，需要仔细设计、小心验证。

表 1-5 传输距离和角色组合定义链路的 Metric

类型	节点内（Intra-Hub）	节点间（Inter-Hub）
核心—核心	50	传输延迟 * 200
核心—汇聚	—	传输延迟 * 200 + 10000
汇聚—汇聚	20	传输延迟 * 200 + 20000
汇聚—汇聚（跨区）	—	传输延迟 * 200 + 30000
汇聚—接入	—	传输延迟 * 200 + 50000
接入—接入	10	—

1.4 IGP 设计汇总

<要点汇总>

==========