第 5 章　SDN 管控技术

知识导读

传统承载网络的管理和控制功能有限，主要是通过网络管理系统（Network Management System，NMS）实现网络设备的硬件管理、性能监控、参数配置等。随着客户及运营商需求的变化以及网络功能的增加，5G 承载网要求能快速实现网络调整，即实现网络的可编程、自动化。

承载网的 SDN（软件定义网络）管控技术是指可编程实现对承载网设备的管理和控制，实现网络自动化运维，提升管控效率。其中，管理功能主要指硬件、软件、告警、性能、日志、配置等各类资源和数据的管理，控制功能主要指通过控制器实现无须人为参与的闭环决策控制、隧道托管给 NMS 的重路由功能等。

本章将讲解承载网的 SDN 背景、架构和关键技术，并介绍 SPN 的 SDN 方案。

学习目标

- 了解 SDN 的网络管理架构。
- 了解 SDN 管控的关键技术。
- 掌握 NETCONF、RESTCONF 接口的基本概念及网络位置。
- 掌握网络 IS-IS 分域规则。
- 掌握网络 BGP-LS、PCEP 部署方案。
- 了解 SPN 的 SDN 管控架构。

能力目标

- 掌握 BGP-LS 现网规划的能力。
- 掌握 IS-IS 现网分域的能力。

5.1　SDN 管控技术背景

SDN 管控技术是一种将设备控制与传送分离并直接可编程的新兴网络架构。5G 承载网的 SDN 管控技术是将 SDN 的集中化智能控制与 5G 承载网面向数据优化的高效多业务传送能力、电信级的高可靠性、端到端的 QoS 保障结合起来的全新承载网络架构。通过开放性的应用和服务，增强网络资源的智能化调度能力，使客户与网络资源之间的关系扁平化，从而提升运维管理和业务运营效率。

SDN 控制功能的核心为应用软件参与对网络行为的定义，通过自动化业务部署简化网络运维，通过开放的 APP 进行快速业务创新，即使承载网具备控制与传送分离、逻辑集中控制、开放的编程接口等技术特征，具体如下：

（1）控制与传送分离。控制平面与分组传送设备在逻辑上独立部署，减少分组传送设备上的分布式网络协议，从而降低分组传送设备的复杂性。通过控制与传送分离，可支持控制平面与传送平面的独立发展。

（2）逻辑集中控制。为达到全网资源的高效利用，SDN 实现了控制功能逻辑集中化。相比传统分布式控制平面，集中控制可以掌握全局网络资源，进行最优的控制决策，实现网络资源全局最优利用。

（3）开放的编程接口。通过标准的可编程网络控制接口，SDN 使承载网可向外部应用开放网络资源信息，允许第三方业务应用灵活利用网络资源，实现网络和业务的持续演进和不断创新。开放的编程接口有利于承载网走向开放和合作的业务开发和经营模式。

5.2 SDN 管控架构

SDN 管控架构如图 5-1 所示。

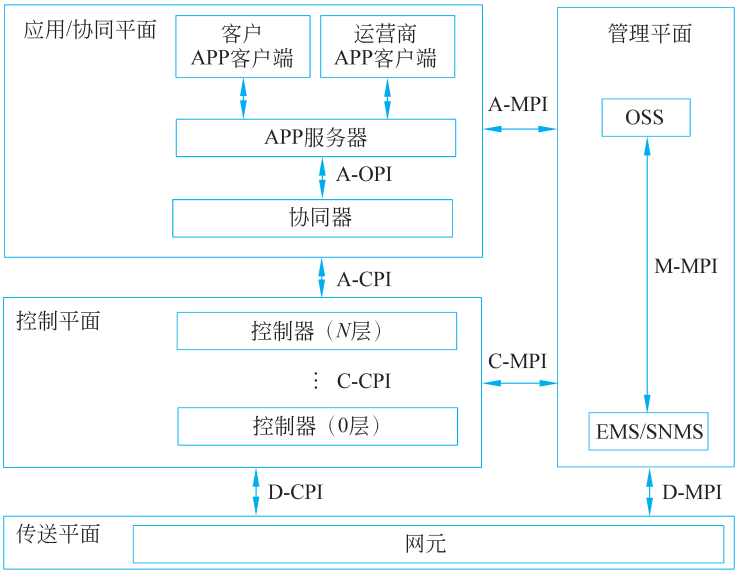


图 5-1 SDN 管控架构

采用 SDN 管控技术，在现网部署中可对传送平面、控制平面、管理平面、应用/协同平面进行软件服务化设计、共云平台部署，以实现 SDN 管理、控制、应用集成部署。

1. 传送平面

简单理解，传送平面就是由承载网设备连接成的具有业务传送能力的网络。

传送平面提供两点或多点之间的双向或单向的用户业务传送能力，也可以提供控制和网络管理信息的传送。此外，传送平面提供信息传输过程中的 OAM、保护恢复、业务 SLA 保证、时钟同步等功能。传送平面承载和传送客户层的各种业务，并保证客户业务信息的透明性。

2. 控制平面

控制平面由支持分层部署的控制器组成，对传送平面的转发行为进行无须人为参与的

闭环决策控制，并向上层应用/协同平面提供控制策略北向接口。

控制器根据业务需求生成转发行为控制数据，并逐层分解控制粒度，最终下发到传送平面各节点，控制网络的业务转发、保护、恢复等行为。控制器应支持分层分域部署，以满足大规模网络的组网要求，分层部署的控制器之间通过带外控制通道互连。上层控制器可连接多个下层控制器，完成跨下层控制域的业务统一控制。

控制平面应具备高可靠性、高安全性，应配置防火墙以防止外部网络或客户对网络的攻击，并且控制平面失效不应影响传送平面业务转发。

3. 管理平面

管理平面包括 EMS/SNMS、OSS 系统，完成对承载网的网络管理和维护。在网络演进过程中，为保持和已有系统的兼容性，应支持与传统管理平面进行协同工作，并保持数据的一致性。

4. 应用/协同平面

应用/协同平面由协同器（orchestrator）、应用服务器（APP server）、应用客户端（APP client）组成。应用/协同平面通过调用应用/协同平面与控制平面接口（A-CPI）对网络进行操作。协同器是应用/协同平面中负责业务协同的组件，包括业务编排、业务策略管理等功能，屏蔽网络技术差异，实现网络资源的协同应用和全网资源的动态可视化构建。协同器向 APP 提供面向业务模型的应用与协同器接口（A-OPI），方便第三方 APP 灵活定制运营商的网络资源。

5. 系统平面间接口

在 SDN 管控系统架构中，各平面通过软件接口进行交互并协同工作。各平面间的接口有以下 5 个：

（1）传送平面与控制平面接口（D-CPI）。用于控制平面对传送平面设备资源的调度、配置以及状态获取。D-CPI 定义了网元级资源信息模型，实现了控制平面对传送资源的统一调度。D-CPI 应能支持多样化接口形式并能兼容现有网元接口。

（2）应用/协同平面与控制平面接口（A-CPI）。控制平面通过 A-CPI 对应用/协同平面提供网络级的抽象能力和服务。A-CPI 定义了网络级资源信息模型，实现了网络能力抽象。

（3）管理平面与传送平面接口（D-MPI）。管理平面通过该接口可对网元实施管理，实现光通道的建立、确认和监视，并在需要时对其进行保护和恢复。

（4）控制平面与管理平面接口（C-MPI）。用于控制平面与管理平面的信息交互，管理平面通过 C-MPI 对控制器进行管理和维护。

（5）应用/协同平面与管理平面接口（A-MPI）。用于应用/协同平面与管理平面的信息交互，通过 A-MPI，应用/协同平面可从管理平面获取存量资源信息，实现与管理平面的协作。

各平面内不同部件通过软件接口进行交互。平面内接口有以下 4 个：

（1）控制平面层间接口（C-CPI）。用于分层部署的控制器之间进行资源的协同调度和控制，每层控制器和其他层控制器之间均可通过 C-CPI 交互。C-CPI 定义了基于控制域的网络级资源信息模型，实现了全网控制器分层分域协同控制。上层控制器通过 C-CPI 对多个下层控制器的网络资源进行调度。

（2）应用与协同器接口（A-OPI）。用于协同器向应用提供基于业务级的抽象能力。A-

OPI 定义了业务级接口和模型，与具体使用的网络技术无关。业务级接口便于应用聚焦子业务开发，而不必关心具体的网络技术，降低了应用开发难度。

（3）应用客户端与应用服务器间接口。用于应用/协同平面内应用客户端与应用服务器之间的信息交互。

（4）管理平面层间接口（M-MPI）。现存的 EMS/SNMS 的北向接口用于 OSS 从 EMS/SNMS 获取网络存量资源信息和进行自动化业务配置，相关内容不在本书讨论范围之内。

5.3 SDN 管控关键技术

SDN 管控关键技术包括控制器、BGP-LS、PCEP、NetConf、RestConf 和数据建模语言 YANG。前 5 项技术在网络中的位置如图 5-2 所示。

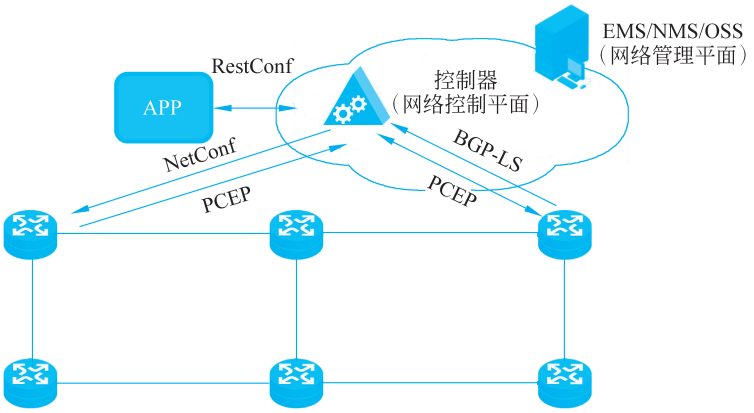


图 5-2 SDN 管控关键技术在网络中的位置

5.3.1 控制器

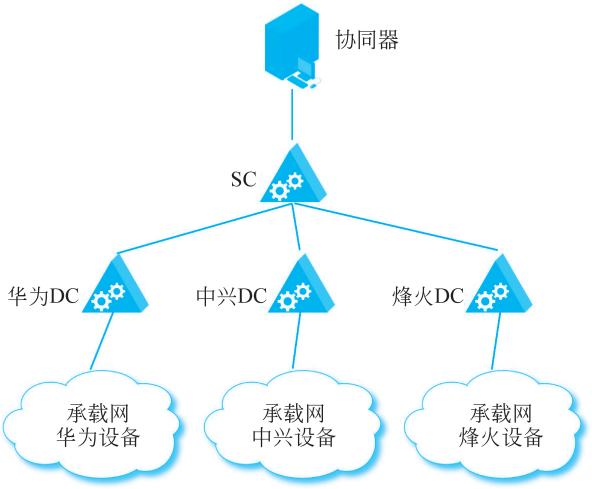
控制器可分层部署，实现网络的集中控制功能。控制器实际上是运行在服务器上的软件系统，同时连接着应用、管理平面和承载网设备，具有承上启下的桥梁作用。运营商可以通过该桥梁实现灵活的网络部署，例如实现跨 IS-IS 域、跨设备厂商甚至跨承载网设备类型的灵活网络部署。

1. 简介

控制器对网络进行集中控制。当承载网由多家厂商的设备组网时，厂商控制器与 NMS 合并设置，构成区域控制器（Domain Controller，DC）实现区域控制，由超级控制器（Super Controller，SC）管理 DC 实现跨厂商管控，数个控制器构成网络的控制平面，如图 5-3 所示。控制器实现网络拓扑和资源统一管理、网络抽象、路径计算、策略管理等功能，同时提供协议适配层和应用接口适配层。

2. 逻辑架构

控制器逻辑架构应包括北向接口协议适配层、业务层、策略层、网络层、资源抽象层以及南向接口协议适配层 6 个层次，同时还包括告警、性能、日志、数据库以及 S-SCN 接口和 S-MCN 接口等通用模块，如图 5-4 所示。



SC：超级控制器（Super Controller），管理多个DC
DC：区域控制器（Domain Controller），与NMS合设

图 5-3　多控制器部署示意图

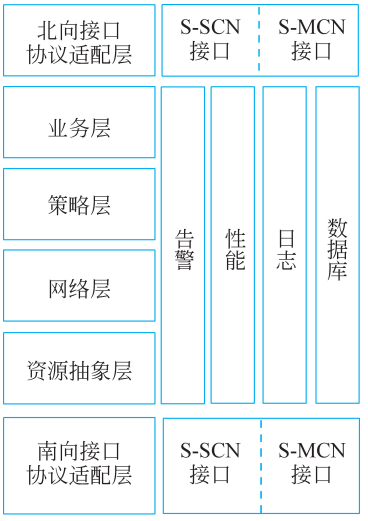


图 5-4　控制器逻辑架构

控制器逻辑架构中各层和各模块支持的功能如下：

（1）北向接口协议适配层支持对业务适配的网络资源编程和控制接口，北向接口建议采用 RestConf 协议。

（2）业务层支持 L2VPN、L3VPN、TDM 仿真等业务的建立、拆除、修改功能，业务层使用策略层提供的各种策略完成业务选路和保护恢复、QoS、OAM 属性设置。

（3）策略层支持网络和业务所需的策略管理功能，包括路由策略、保护恢复策略、QoS 策略、OAM 策略、安全策略等。

（4）网络层支持对网络进行抽象，屏蔽物理网络细节，实现网络拓扑管理、管道连接管理。资源抽象层可提供物理网络抽象模型和虚拟网络抽象模型，用于路径计算。

（5）资源抽象层支持收集网元资源（包括网络中节点、端口等资源）的信息。

（6）南向接口协议适配层支持对下层控制器、传送平面网元的接口适配和下发。在网络演进过程中，也可采用控制器与网管的私有接口进行传送平面南向接口适配。

（7）通用功能模块提供网络必要的告警、性能和日志查询功能。

（8）S-SCN 接口是控制器与控制器之间以及控制器与网元之间安全可靠的传输通道。控制器还应支持与管理平面互通的接口。

（9）S-MCN 接口是管理平面之间以及管理平面与网元之间安全可靠的传输通道。

（10）数据库负责业务、策略、资源、告警、性能等数据的存储和同步。

3. 控制平面

数个控制器构成控制平面。控制平面总体架构如图 5-5 所示，包括分层部署的控制器以及相关接口。

控制平面使用的接口如下：

（1）控制平面通过 D-CPI 与传送平面进行交互，对传送平面的网元进行控制操作并获取网元相关资源的状态信息。

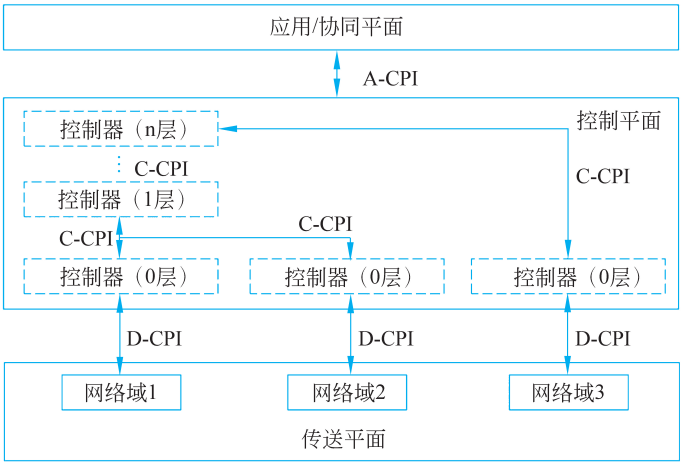


图 5-5 控制平面总体架构

（2）控制平面通过 A-CPI 与应用/协同平面进行交互，通过开放可编程接口向应用/协同平面提供网络服务。

（3）在控制平面分层部署的情况下，不同层次的控制器之间通过 C-CPI 进行交互，完成全网 SPN 资源的协同控制。

5.3.2 BGP-LS

BGP-LS（Border Gateway Protocol Link State，携带链路状态的边界网关控制协议），是一种集中控制协议，是 BGP 的扩展应用，用于控制器搜集网络实时拓扑状态。该协议汇总 IGP 收集的拓扑信息并传送给上层控制器，实现网络拓扑监控、流量调优、重路由等功能。

1. 简介

每个路由器都维护一个或多个数据库，用于存储任何给定区域内与节点、链路相关的链路状态信息。存储在这些数据库中的链路属性包括本地/远端 IP 地址、本地/远端接口标识符、链路度量和 TE（Traffic Engineering，流量工程）度量、链路带宽、保留带宽、CoS（Class-of-Service，服务等级）保留状态、优先级、共享风险链路组（Shared Risk Link Group，SRLG）。路由器的 BGP 进程可以从这些数据库中恢复拓扑并将其发布给使用者。

链路状态和 TE 信息的收集和发布如图 5-6 所示。

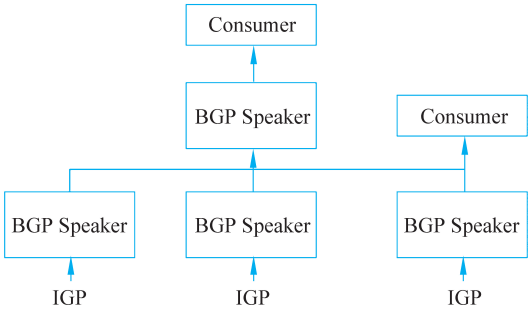


图 5-6 链路状态和 TE 信息的收集和发布

发送 BGP 报文的设备称为 BGP Speaker，相互交换报文的 Speaker 之间互称对等体

(peer)。BGP Speaker 支持信息发布策略的配置。BGP Speaker 可以从 LSDB(Link-State Database)或 TED(Traffic Engineering Database,流量工程数据库)中分发真实的物理拓扑,也可以创建一个抽象拓扑[由虚拟路径连接的虚拟聚合节点,聚合节点可以是同一个 POP(Point Of Presence,因特网接入点)的多台路由器]。抽象拓扑也可以是物理节点(或链路)和虚拟节点(或链路)的组合。通过配置 BGP Speaker 的拓扑信息更新频率,可以减小网络中拓扑更新信息的流量。

BGP-LS(RFC7552)产生前,网元使用 IGP(OSPF 协议或 IS-IS 协议)收集网络的拓扑信息,IGP 将各个域的拓扑信息单独上送给上层控制器。在这种拓扑收集方式下,存在以下几个问题:

(1) 对上层控制器的计算能力要求较高,且要求控制器也支持 IGP 及其算法。

(2) 当涉及跨 IGP 域拓扑信息收集时,上层控制器无法看到完整的拓扑信息,无法计算端到端的最优路径。

(3) 不同的路由协议分别上送拓扑信息给上层控制器,控制器对拓扑信息的分析处理过程比较复杂。

BGP-LS 产生后,IGP 发现的拓扑信息由 BGP 汇总后上送给上层控制器,利用 BGP 强大的选路和算路能力带来以下几点优势:

(1) 降低对上层控制器计算能力的要求,且不再对控制器的 IGP 能力有要求。

(2) BGP 将各个进程或各个自治系统(Autonomous System,AS)的拓扑信息进行汇总,直接将完整的拓扑信息上送给控制器,有利于路径选择和计算。

(3) 网络中所有拓扑信息均通过 BGP 上送控制器,使拓扑上报协议归一化。

2. BGP-LS 路由

BGP-LS 在原有 BGP 的基础上引入了一系列新的 NLRI(Network Layer Reachability Information,网络层可达性信息)携带链路、节点和 IPv4/IPv6 前缀相关信息,这种新的 NLRI 称为链路状态 NLRI(Link-State NLRI)。BGP-LS 采用 MP_REACH_NLRI(多协议可达 NLRI)和 MP_UNREACH_NLRI(多协议不可达 NLRI)属性作为链路状态 NLRI 的容器,即链路状态 NLRI 是作为 MP_REACH_NLRI 或者 MP_UNREACH_NLRI 属性携带在 BGP 的 Update 消息中的。

一共有 6 种 BGP-LS 路由,分别用来携带节点、链路、路由前缀信息、IPv6 路由前缀信息、SRv6 SID 路由信息和 TE 策略路由信息。这几种路由相互配合,共同完成拓扑信息的传输。

BGP-LS 主要定义了如下几种链路状态 NLRI:Node NLRI(节点 NLRI)、Link NLRI(链路 NLRI)、IPv4 Topology Prefix NLRI(IPv4 拓扑前缀 NLRI)、IPv6 Topology Prefix NLRI(IPv6 拓扑前缀 NLRI)。下面给出各种 NLRI 的报文格式。

(1) Node NLRI 报文格式如图 5-7 所示。

Node NLRI 报文各字段解释如下:

- Protocol-ID:协议标识,如 IS-IS、OSPF 和 BGP 等,8 位。
- Identifier:标识符,在运行 IS-IS、OSPF 多实例时,用于标识不同的协议实例,64 位。
- Local Node Descriptors:本地节点描述符,由一系列节点描述符 sub-TLV 组成,长度可变。

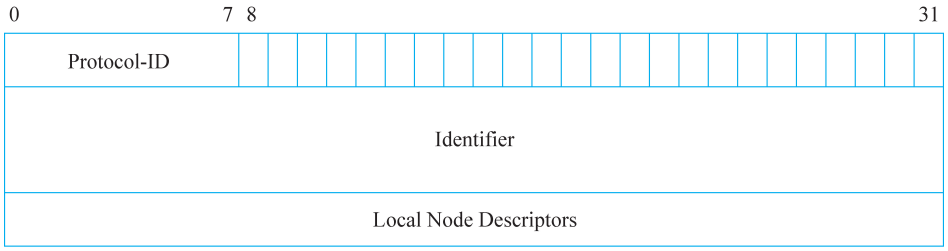


图 5-7 Node NLRI 报文格式

(2) Link NLRI 报文格式如图 5-8 所示。

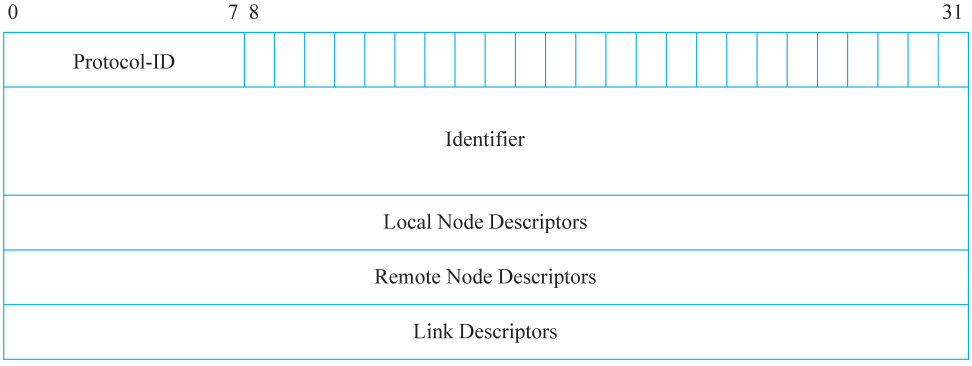


图 5-8 Link NLRI 报文格式

Link NLRI 报文各字段解释如下：

- Protocol-ID：协议标识，如 IS-IS、OSPF 和 BGP 等，8 位。
- Identifier：标识符，在运行 IS-IS、OSPF 多实例时，用于标识不同的协议实例，64 位。
- Local Node Descriptors：本地节点描述符，由一系列节点描述符 sub-TLV 组成，长度可变。
- Remote Node Descriptors：远端节点描述符，长度可变。
- Link Descriptors：链路描述符，长度可变。

(3) IPv4/IPv6 Topology Prefix NLRI 报文格式如图 5-9 所示。

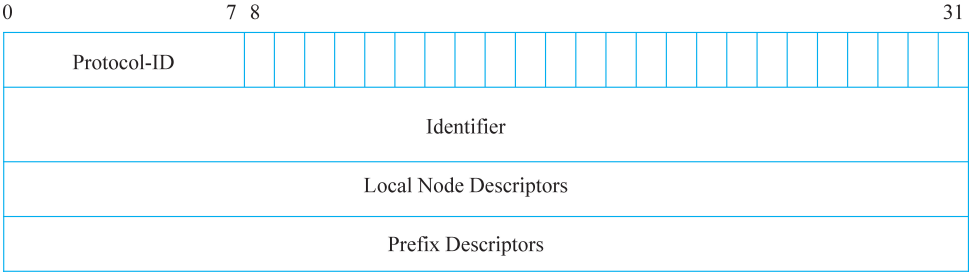


图 5-9 IPv4/IPv6 Topology Prefix NLRI 报文格式

IPv4/IPv6 Topology Prefix NLRI 报文各字段解释如下：

- Protocol-ID：协议标识，如 IS-IS、OSPF 和 BGP 等，8 位。
- Identifier：标识符，在运行 IS-IS、OSPF 多实例时，用于标识不同的协议实例，64 位。

- Local Node Descriptors：本地节点描述符，由一系列节点描述符 sub-TLV 组成，长度可变。
- Prefix Descriptors：前缀描述符，长度可变。

与此同时，针对上述 NLRI，BGP-LS 还定义了相应的属性，用于携带节点、链路和 IPv4/IPv6 前缀相关的参数和属性。BGP-LS 属性是以 TLV[Tag(标签)、Length(数据的长度)、Value(数据)]的形式和对应的 NLRI 携带在 BGP-LS 消息中。这些属性都属于 BGP 可选非传递属性，主要包括 Node Attribute(节点属性)、Link Attribute(链路属性)和 Prefix Attribute(前缀属性)。

3. 典型组网

1) IGP 域内拓扑信息收集

如图 5-10 所示，A、B、C 和 D 之间通过 IS-IS 协议达到 IP 网络互联的目的。A、B、C 和 D 同属于域 10，都是 Level-2 设备。在这种情况下，只需要 A、B、C 和 D 中的任何一台设备部署 BGP-LS 特性并与控制器建立 BGP-LS 邻居关系便可以达到整个网络拓扑收集和上送的目的。但是，为了拓扑上送的可靠性，往往选择两台或两台以上设备都部署 BGP-LS 特性并与控制器建立 BGP-LS 邻居关系。由于网络中的设备收集的拓扑信息相同，所以它们之间可以互相作为备份，当有设备出现故障时依然保证拓扑信息的及时上送。

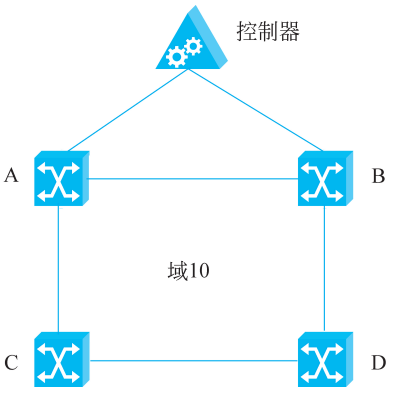


图 5-10　IGP 域内拓扑信息收集

2) BGP 自治域间拓扑信息收集

如图 5-11 所示，A 和 B 属于同一自治系统，两者之间建立 IS-IS 邻居关系。A 为自治系统内部的一台非 BGP 设备。B 和 C 之间建立 EBGP(External Board Gateway Protocol，外部边界网关协议)连接。在这种情况下，由于 BGP(未使能 BGP-LS)不能传递拓扑信息，所以 AS100 内的设备和 AS200 内的设备上收集的拓扑信息不同(都只能收集本自治系统的拓扑信息)，所以此时要求 AS100 和 AS200 两个自制系统中都至少有一台设备使能 BGP-LS 特性并与控制器建立 BGP-LS 邻居关系。每个自治系统中有两台或两台以上设备与控制器相连，则可以保证拓扑收集与上送的可靠性。

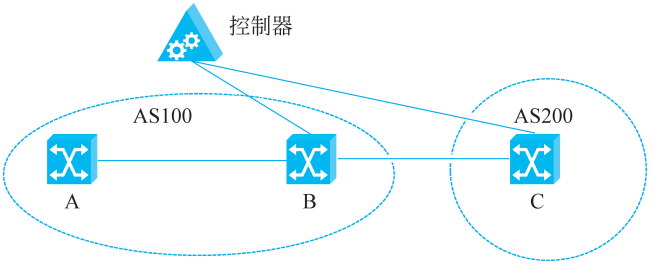


图 5-11　BGP 自治域间拓扑信息收集

4. 部署实例

图 5-12 给出了 BGP-LS 部署实例。

每个 IS-IS 域内至少选取两台设备与控制器建立 BGP-LS 连接(逻辑连接)，设备将通

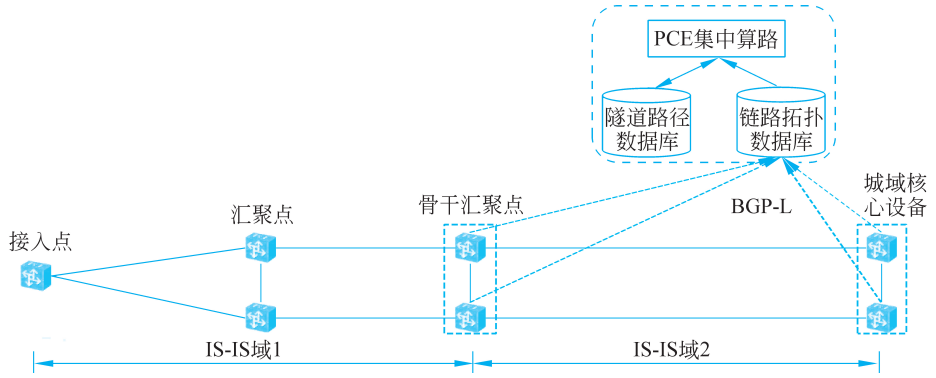


图 5-12 BGP-LS 部署实例

过 IS-IS 域搜集的网络拓扑信息 LSDB 等通过 BGP-LS 上报给控制器。核心 IS-IS 域选取一对核心设备启用 BGP-LS，接入汇聚 IS-IS 域选取骨干汇聚对（IS-IS 分层点）设备启用 BGP-LS。

同一对骨干汇聚点带多个 IS-IS 进程时，启用一个 BGP-LS 将多个进程绑定到一个 BGP-LS 会话上。

以华为 SPN 设备的 BGP-LS 全局配置为例，其脚本示例如下，其中 3.3.3.3 为 BGP-LS Server 的 IP 地址。

```
#
bgp 10
  peer3.3.3.3 as-number 10
#
link-state-family unicast
  peer3.3.3.3 enable
#
```

同时，需要在对应的 IS-IS 配置中使能 BGP-LS：

```
#
isis 1
  ****(此部分配置省略)
  bgp-ls enable level-2
#
```

5.3.3 PCEP

PCEP（Path Computation Element communication Protocol，路径计算单元通信协议）是一种集中动态控制协议，用于设备向控制器请求隧道算路以及控制器向设备下发隧道标签栈信息，即网络路径的请求和计算。PCEP 面向连接，效率高。

1. 简介

为实现更好的网络控制，承载网网络设备和 SDN 控制器应支持 PCEP，并通过 PCEP 将集中算路结果实时下发到网络设备，或由网络设备向控制器提交算路申请。在 PCEP 模