MPLS TE 技术白皮书

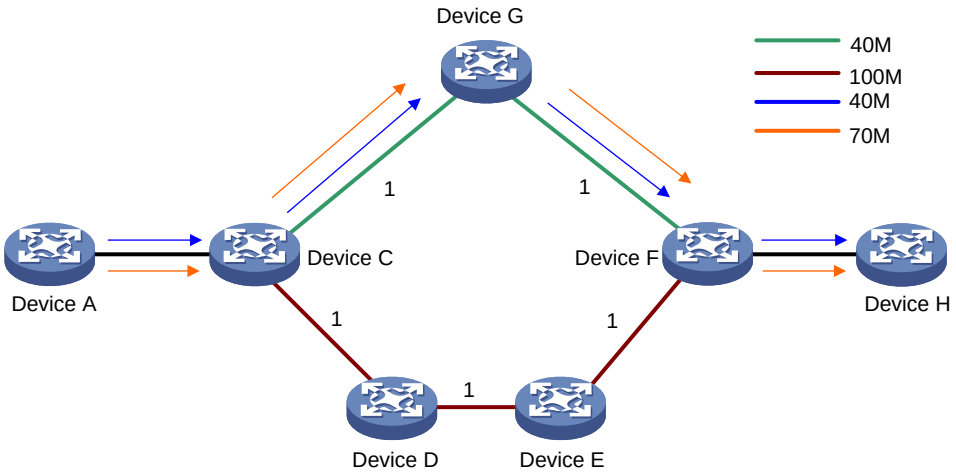
目 录

1 概述

1.1 产生背景

路由器根据传统路由协议计算出的最短路径转发流量，即使某条路径发生拥塞，也不会将流量切换到其他的路径上。在网络流量比较小的情况下，这种问题不是很严重，但是随着 Internet 的应用越来越广泛，传统的最短路径优先的路由问题暴露无遗。

图1 传统路由问题



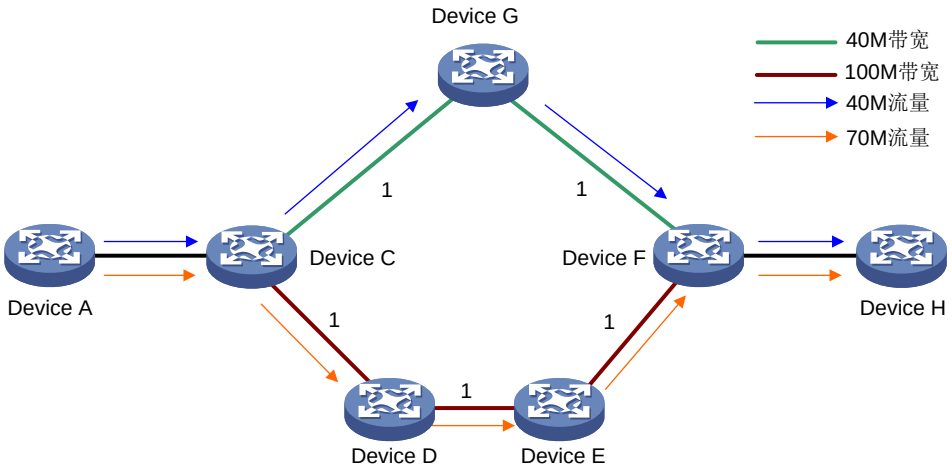
如图 1 所示，假设每条链路的 Metric 值相同，则从 Device A 到 Device H 的最短路径为 Device A－Device C－Device G－Device F－Device H。尽管存在 Device A－Device C－Device D－Device E－Device F－Device H 这条路径，但是流量转发只会从最短路径 Device A－Device C－Device G－Device F－Device H 上经过。这样就有可能形成一条路径 Device A－Device C－Device G－Device F－Device H 过载，一条链路 Device A－Device C－Device D－Device E－Device F－Device H 闲置。

可以通过负载分担，即修改 Metric 值使得到达同一目的地存在多条开销相同路由的方法，解决上述问题。但是这种方法可能会引起其他链路拥塞。在拓扑复杂的网络中，Metric 值的调整更加困难，一条链路的改动可能会引起多条路由变动。

1.2 流量工程的解决办法

通过流量工程，服务提供商可以精确地控制流量流经的路径，从而可以避开拥塞的节点，解决一部分路径过载，另一部路径空闲的问题，使现有的带宽资源得到充分利用。如图 2 所示，从 Device A 到 Device H 存在两条路径：Device A－Device C－Device G－Device F－Device H 和 Device A－Device C－Device D－Device E－Device F－Device H，前者的带宽为 40M，后者的带宽为 100M。流量工程可以根据带宽等因素合理地分配流量，从而有效地避免链路拥塞。例如，Device A 到 Device H 存在两种业务，流量分别为 40M 和 70M，流量工程可以把前者分配到带宽为 40M 的路径上，将后者分配到带宽为 100M 的路径上。

图2 建立流量工程的网络



流量工程关注网络整体性能的优化，其主要目标是方便地提供高效、可靠的网络服务，优化网络资源的使用，优化网络流量。流量工程分两个层面：一是面向流量的，关注如何提高网络的服务质量；二是面向资源的，关注如何优化网络资源的使用，最主要是带宽资源的有效利用。通过实施流量工程，可以减少网络的管理成本，使网络资源充分有效的使用，可以在网络拥塞或者抖动的情况下动态调节资源，同时还可以实现增值服务和附加业务。

1.3 MPLS TE技术优点

流量工程可以通过在 IGP 上使用重叠模型来实现，如 IP over ATM、IP over FR 等。重叠模型在网络的物理拓扑结构之上提供了一个虚拟拓扑结构，从而扩展了网络设计的空间，为支持流量与资源控制提供了许多重要功能，可以实现多种流量工程策略。然而，由于协议之间往往存在很大差异，重叠模型在可扩展性方面存在不足。

为了在大型骨干网络中部署流量工程，必须采用一种可扩展性好、简单的解决方案。MPLS TE 就是针对这一需求而提出的。

MPLS 本身具有一些不同于 IGP 的特性，其中就有实现流量工程所需要的，例如：

- MPLS 支持显式 LSP 路由；
- LSP 较传统单个 IP 分组转发更便于管理和维护；
- 基于 MPLS 的流量工程的资源消耗较其它实现方式更低。

MPLS TE 结合了 MPLS 技术与流量工程，具备以下优势：

- 在建立 LSP 隧道的过程中，可以预留资源，保证服务质量；
- LSP 隧道有优先级、抢占等多种属性，可以方便地控制 LSP 隧道的行为；
- 通过备份路径和快速重路由技术，在链路或节点失败的情况下，提供保护；
- 建立 LSP 隧道的负荷小，不会影响网络的正常业务。

正是这些优势，使得 MPLS TE 成为非常吸引人的流量工程方案。通过 MPLS TE 技术，服务提供商能够充分利用现有的网络资源，提供多样化的服务。同时可以优化网络资源，进行科学的网络管理。

2 MPLS TE 技术实现

2.1 概念介绍

1. CRLSP

CRLSP（Constraint-based Routed Label Switched Paths，基于约束路由的 LSP）是基于一定约束条件建立的 LSP。与普通 LSP 不同，CRLSP 的建立不仅依赖路由信息，还需要满足其他一些条件，比如带宽需求、显式路径等。

可以通过静态方式、动态方式或 PCE 方式建立 CRLSP。

2. MPLS TE 隧道

MPLS TE 隧道是从头节点到目的节点的一条虚拟点到点连接。通常情况下，MPLS TE 隧道由一条 CRLSP 构成。在部署转发路径备份或需要将流量通过多条路径传输等情况下，需要为同一种流量建立多条 CRLSP，在这种情况下，MPLS TE 隧道由一组 CRLSP 构成。

头节点上 MPLS TE 隧道由 MPLS TE 模式的 Tunnel 接口标识。当流量的出接口为 Tunnel 接口时，该流量将通过构成 MPLS TE 隧道的 CRLSP 来转发。

2.2 静态建立CRLSP

静态建立 CRLSP 是指在流量经过的每一跳设备上（包括 Ingress、Transit 和 Egress）分别手工指定入标签、出标签、流量所需的带宽等信息，从而建立满足约束条件的 CRLSP。该方式的优点是配置简单，缺点是不能根据网络的变化动态调整建立的 CRLSP。

2.3 动态建立CRLSP

动态建立 CRLSP 是指根据链路状态信息计算出路径后，通过标签分发协议（如 RSVP-TE）通告标签，并在经过的节点上为流量预留所需的带宽资源，从而建立满足约束条件的 CRLSP。该方式的优点是能根据网络的变化动态调整建立的 CRLSP，且支持 CRLSP 备份、快速重路由等功能，缺点是配置复杂。

采用动态方式建立 CRLSP 时，MPLS TE 需要实现如下功能：

- 发布包含链路 TE 属性的信息，以便根据这些信息选择满足约束条件的路径。
- 计算出到达某个节点的满足 TE 属性要求的最短路径。
- 通过标签分发协议沿着计算出的路径建立 CRLSP，并预留资源。

2.3.1 发布 TE 属性

除了网络的拓扑信息外，流量工程还需要知道网络上各链路的 TE 相关属性以及负载信息。为此，需要对现有的 IGP 进行扩展，来发布链路状态信息，包括最大链路带宽、最大可预留带宽、当前预留带宽、亲和属性等。这些 TE 相关信息通过 IGP 在网络上泛洪，在需要进行 CSPF 计算的设备上形成流量工程（TE）使用的链路状态数据库 TEDB。TEDB 数据库中不仅收集了网络拓扑信息，还增加了 TE 所关心的链路属性，使用 TEDB 可以监控整个网络中使能了 TE 功能的链路状态，并通过 CSPF 算法计算出以自己为根节点的、基于限制的到目的网络的路径。

对于 OSPF 协议而言，它使用 Opaque Type 10 LSA 携带链路的 TE 属性信息。接口上的 TE 相关属性变化会及时通过这类 LSA 更新并泛洪到 OSPF 的其它邻居，最终在每个运行 TE 的 LSR 上形成 TEDB。

同样的，对于标准的 IS-IS 协议，也对其进行扩展来承载 TE 所需要的各种信息。IS（Router）通过 IS-IS Link State Protocol Data Units（LSPs）发布路由信息，在 LSP 中增加新的 TLV 用于承载 TE 相关的链路属性。如，TLV Type 22 携带 wide 类型的开销值；TLV Type 134 携带 TE Router ID；TLV Type 135 携带可达的 IP 信息。

2.3.2 路径计算

MPLS TE 利用 CSPF 算法，根据通过 IS-IS TE 或 OSPF TE 扩展产生的 TEDB，计算符合带宽、亲和属性、抢占/保持优先级、显式路径等约束条件的路径。

1. 约束条件

(1) 带宽

带宽要求是指经过 MPLS TE 隧道的流量所属的服务类型及其所需的带宽。只有链路上针对流量所属服务类型的可预留带宽大于等于流量所需带宽时，该链路才满足带宽约束条件。

(2) 亲和属性

MPLS TE 隧道的亲和属性和链路的属性配合，决定了该隧道可以使用哪些链路。

链路属性、亲和属性、亲和属性的掩码都是 32 位的二进制数。如果希望某条链路能够被隧道所用，则需要满足如下要求：

- 对于掩码为 1 的位，亲和属性为 1 的位中链路属性至少有 1 位也为 1，亲和属性为 0 的位对应的链路属性位不能为 1。
- 对于掩码为 0 的位，不对链路属性的相应位进行检查。

例如，亲和属性为 0xFFFFFFF0，掩码为 0x0000FFFF，则可用链路的链路属性高 16 位可以任意取 0 或 1，17～28 位中至少有 1 位为 1，且低 4 位不能为 1。

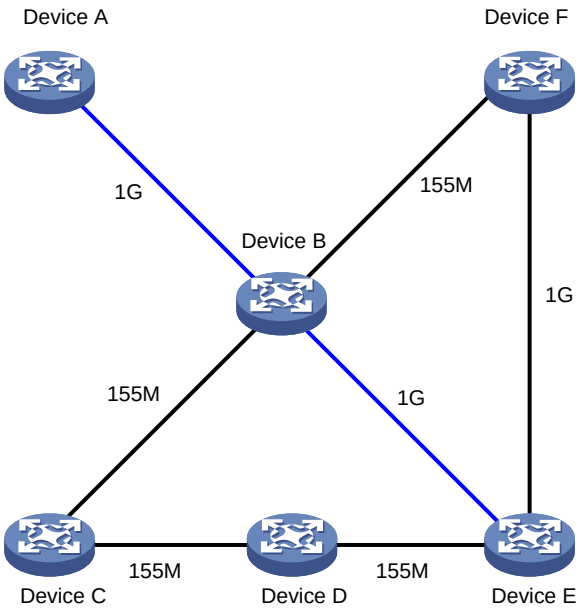
(3) 建立和保持优先级

- 优先级

  LSP 隧道有一个建立优先级和一个保持优先级。优先级的范围从 0 到 7，7 为最低优先级。需要建立多条 LSP 隧道的情况下，建立优先级高的 LSP 隧道优先占有资源、优先建立。当带宽等资源不够时，保持优先级低的、已建立的 LSP 隧道的带宽资源可能被一个建立优先级高的 LSP 隧道抢占。

- 抢占

图3 LSP 抢占



如图 3 所示，标明了链路的带宽（假设两个方向的带宽相同），每条链路的 Metric 值都相同。存在两条 TE 隧道，Tunnel1：Device A－Device B－Device E，带宽需求为 155M，优先级为 0；Tunnel2：Device C－Device B－Device F，带宽需求为 155M，优先级为 7。

假设 Device B－Device E 的链路 down 了，Device B 通过信令通知 Device A 链路故障，Device A 会计算出新的路径 Device A－Device B－Device F－Device E，链路 Device B－Device F 的带宽不够 Tunnel1、Tunnel2 共同使用，Tunnel2 会被抢占，新的 Tunnel1 会优先建立。

(4) 显式路径

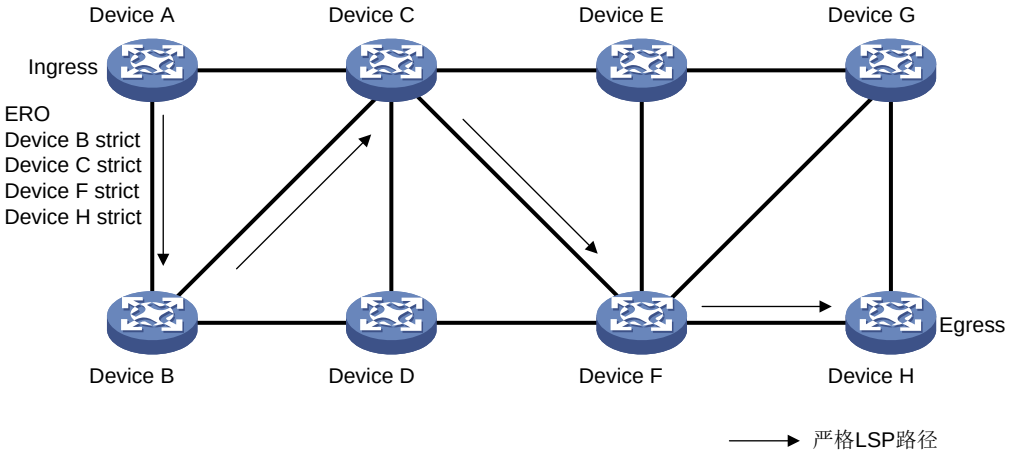
通过显式路径技术，可以指定到达某个目的地所必须经过的路径、不能经过的路径等。将显式路径作为约束条件，可以动态计算出所规划的 LSP 路径。

- 严格显式路径

  严格显式路径是指定必须经过哪些节点，并且指定的下一跳与前一跳必须直接相连。通过严格显式路径，可以最精确地控制 MPLS TE 隧道所经过的路径。

  如图 4 所示，“Device B strict”表示该 CRLSP 必须经过 Device B，并且 Device B 的前一跳是 Ingress LSR（Device A），“Device C strict”表示该 LSP 必须经过 Device C，并且 Device C 的前一跳是 Device B。

图4 严格显式路径

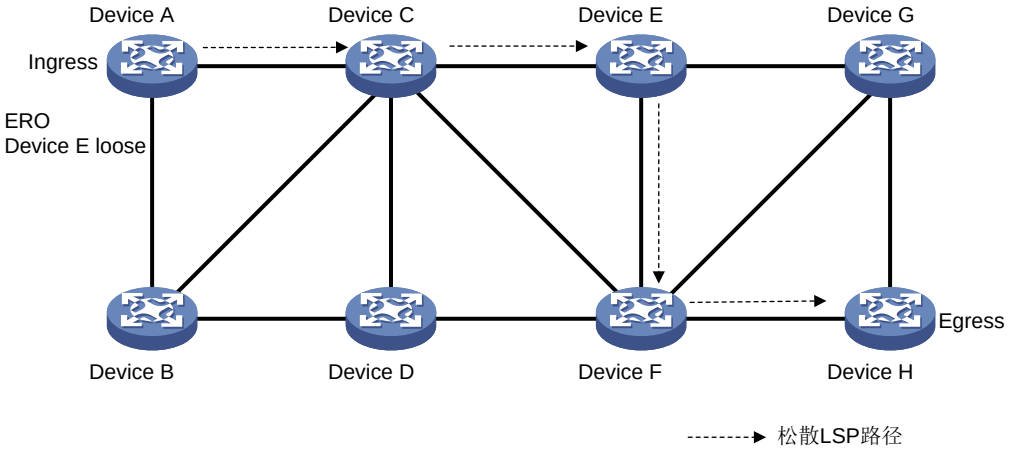


- 松散显式路径

  松散显式路径是指定必须经过哪些节点，并且指定的下一跳和前一跳之间可以存在其他节点。通过松散显式路径，可以模糊地限制 MPLS TE 隧道所经过的路径。

  如图5所示，“Device E loose”表示该LSP必须经过Device E，但是Device E与Ingress LSR（Device A）之间可以经过多个路由器，不必直接相连。

图5 松散显式路径

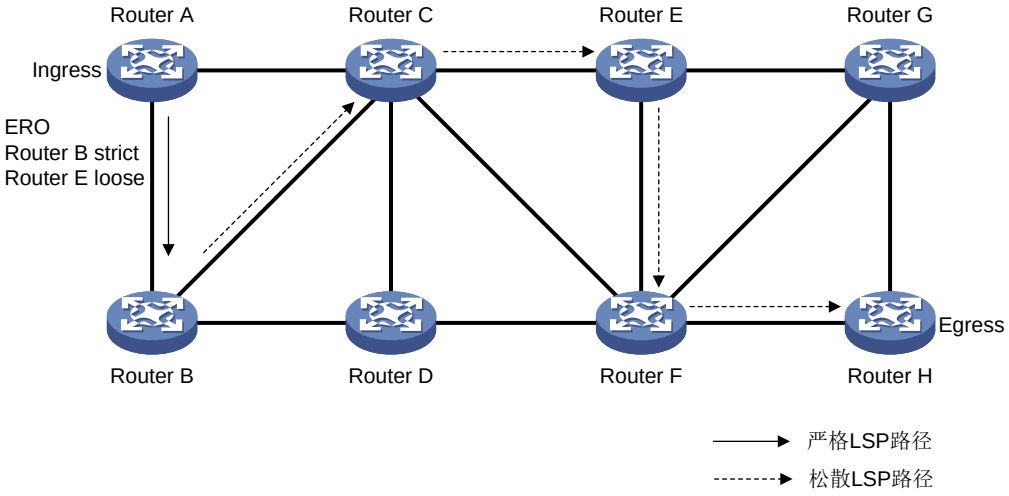


- 严格与松散混合

  严格显式路径和松散显式路径还可以配合使用，即在显式路径中部分节点之间必须直接相连，部分节点之间可以存在其他节点。

  如图6所示，“Device B strict”表示该LSP必须经过Device B，并且Device B与Ingress LSR（Device A）必须直接相连；“Device E loose”表示该LSP必须经过Device E，但是Device E与Device B之间可以经过多个路由器，不必直接相连。

图6 混合显示路径

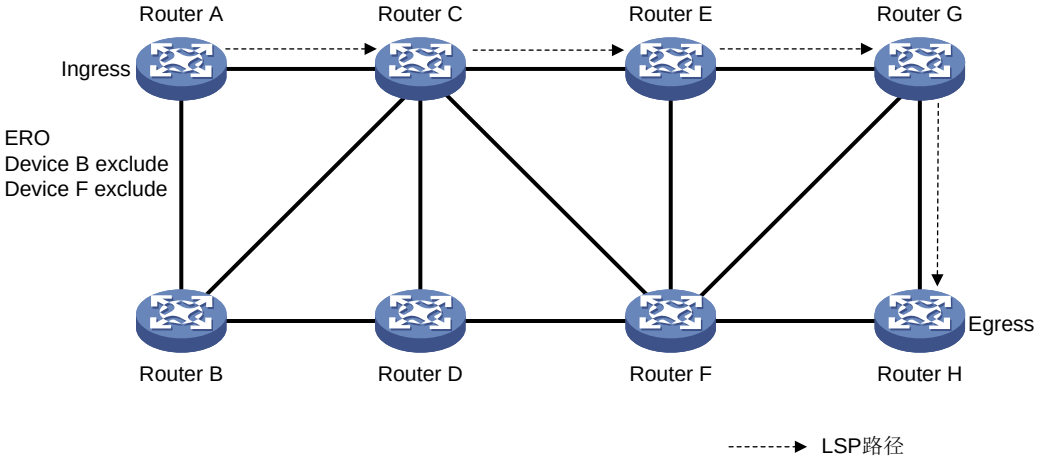


- 路径中排除节点

  显式路径除了可以指定必须经过的节点外，还可以指定显式路径不能经过的节点，以便更加灵活地控制 MPLS TE 隧道所经过的路径。

  如图 7 所示，“Device B exclude”表示该 LSP 不能经过 Device B；“Device F exclude”表示该 LSP 不能经过 Device F。Device A 到达 Device H 的路径为 Device A－Device C－Device E－Device G－Device H。

  图7 显式路径中不能包括节点



2. 选路过程

CSPF 是一种改进的最短路径优先算法，在计算通过网络的最短路径时，将特定的约束也考虑进去。CSPF 基于资源的可用性和所选部分是否违反用户策略约束，在当前拓扑结构中删除不满足条件的节点和链路，然后再通过 SPF 算法计算出一条满足约束条件的最短路径，包括一组 LSR 地址。

路径计算的具体过程为：

(1) 对比 TEDB 中的每一条链路，裁减不满足带宽和亲和属性等要求的链路；

(2) 在剪切以后的拓扑中采用最短路径算法（SPF 算法），得到一条满足 LSP 约束条件的最短路径；

(3) 如果仍有多条路径，选择跳数最少的路径；

(4) 如果仍有多条路径，根据配置的负载分担策略进行选择。

IS-IS 或 OSPF 的 SPF 计算出来的下一跳就是直接的下一跳，每一个路由器都需要运行 SPF 算法。而 CSPF 计算的结果是一条满足约束条件的完全明确的路由，它通常只在需要建立 LSP 的入口节点（TE 的头节点）进行计算。

2.3.3 通过 RSVP-TE 建立 CRLSP

Comware 支持 RSVP-TE 动态信令协议。RSVP-TE 使用 DoD 模式（下游设备按需分配）进行标签发布。通过 RSVP-TE 信令协议实现资源预留和 LSP 建立等功能。

1. RSVP-TE 概述

RSVP 采用 Integrated Service 模型，用于在一条路径的各节点上进行资源预留。RSVP 工作在传输层，但不参与应用数据的传送，是一种 Internet 上的控制协议，类似于 ICMP。

RSVP 经扩展后可以支持 MPLS 标签的分发，并在传送标签绑定消息的同时携带资源预留信息，这种扩展后的 RSVP 称为 RSVP-TE。RSVP-TE 作为一种信令协议，用于在 MPLS TE 中建立 LSP 隧道，可以实现：

- CRLSP 的建立和维护
- CRLSP 路径的拆除
- 错误通告

2. RSVP-TE 消息类型

RSVP-TE 使用 RSVP 的消息类型，并进行了扩展。RSVP 使用以下消息类型：

- Path 消息：由发送者沿数据报文传输的方向向下游发送，在沿途所有节点上保存路径状态。
- Resv 消息：由接收者沿数据报文传输的方向逆向发送，在沿途所有节点上进行资源预留的请求，并创建和维护预留状态。
- PathTear 消息：此消息产生后马上向下游发送，并立即删除沿途节点的路径状态和相关的预留状态。
- ResvTear 消息：此消息产生后马上向上游发送，并立即删除沿途节点的预留状态。
- PathErr 消息：如果在处理 Path 消息的过程中发生了错误，就会向上游发送 PathErr 消息，PathErr 消息不影响沿途节点的状态，只是把错误报告给发送者。
- ResvErr 消息：如果在处理 Resv 消息的过程中发生了错误，或者由于抢占导致预留被破坏，就会向下游节点发送 ResvErr 消息。
- ResvConf 消息：该消息发往接收者，用于对预留消息进行确认。

3. RSVP-TE 对 RSVP 消息的扩展

RSVP-TE 对 RSVP 消息的扩展主要是在 Path 消息和 Resv 消息中增加了新的对象。新增对象除了可以携带标签信息外，还可以携带在沿途寻找路径时的限制信息，从而实现对约束条件和快速重路由的支持。

Path 消息新增的对象包括：

- LABEL_REQUEST：用来请求下游节点分配标签。

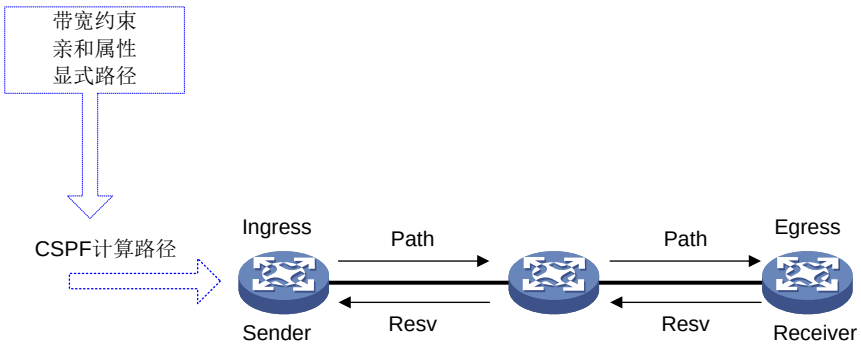
- EXPLICIT_ROUTE：用来携带 Ingress 节点计算出的路径信息，确保沿着该路径建立 CRLSP。
- RECORD_ROUTE：用来记录 CRLSP 实际经过的路径及各个节点分配的标签。
- SESSION_ATTRIBUTE：用来携带 MPLS TE 隧道的属性信息，如建立优先级、保持优先级、亲和属性等。

Resv 消息新增的对象包括：

- LABEL：用来将下游节点分配的标签通告给上游节点。
- RECORD_ROUTE：用来记录 CRLSP 实际经过的路径及各个节点分配的标签。

4. 建立 CRLSP

图8 使用 RSVP-TE 建立 CRLSP 的示意图



如图 8 所示，使用 RSVP-TE 建立 CRLSP 隧道的过程可以简单描述为：

(1) 在 Ingress 上依据 TE 隧道所配置的约束，如带宽约束、亲和属性、显式路径等条件，CSPF 计算出 CRLSP 隧道所要经过的路径。

(2) Ingress LSR 产生携带相应带宽预留信息和路径信息的 Path 消息，依据计算的路径向 Egress LSR 方向发送。Path 消息经过的 LSR，都依据 Path 消息生成路径状态。

(3) Egress LSR 收到 Path 消息后，产生携带预留信息和标签的 Resv 消息，沿 Path 消息发送的相反路径逐跳返回到 Ingress LSR。同时，Resv 消息在沿途的 LSR 上预留资源，并生成预留状态，生成标签交换路径。

(4) 当 Ingress LSR 收到 Resv 消息时，CRLSP 建立成功。

采用 RSVP-TE 建立的 CRLSP 具有资源预留功能，沿途的 LSR 可以为该 CRLSP 分配一定的资源，使在此 CRLSP 上传送的业务得到保证。

5. RSVP 刷新机制

(1) Refresh 消息

由于 RSVP 是软状态协议，因此需要定时发送消息来维护节点上的资源预留状态。

资源预留状态包括路径状态和预留状态，分别保存在如下状态块中：

- PSB（Path State Block，路径状态块）：由 Path 消息创建，用来保存 LABEL_REQUEST 对象。
- RSB（Reservation State Block，预留状态块）：由 Resv 消息创建，用来保存 LABEL 对象。

路径状态和预留状态分别由周期性发送的 Path 消息和 Resv 消息来刷新。对于某个状态，如果在一定时间内没有收到刷新消息，则 PSB 或 RSB 中相应的状态将被删除，根据该状态建立的 CRLSP 也将被删除。

用来刷新资源预留状态的 Path 和 Resv 消息，统称为 Refresh 消息。Refresh 消息除了刷新资源预留状态外，还可以用于恢复丢失的 RSVP 消息。

由于 Refresh 消息是周期性发送的，当网络中的 RSVP 会话比较多时，Refresh 消息会加重网络负担，此时 Path 和 Resv 消息的刷新时间间隔不易过小；而对于时延敏感的应用，当 RSVP 消息丢失时，希望能够尽快通过 Refresh 消息恢复丢失的消息，此时 Path 和 Resv 消息的刷新时间间隔不易过大。简单地调整刷新间隔并不能同时解决这两类问题。

Srefresh（Summary Refresh，摘要刷新）和 RSVP 消息的可靠传递功能可以很好地解决上述问题。

(2) 摘要刷新功能

摘要刷新功能的工作机制为：发送 Path 和 Resv 消息时，在消息中携带 Message ID，用来唯一标识一个消息；RSVP 通过发送携带待刷新消息 Message ID 的 Srefresh 消息，来刷新对应的 Path 和 Resv 消息。

采用摘要刷新功能后，不必传送标准的 Path 和 Resv 消息，只需传递携带 Path 和 Resv 消息摘要的 Srefresh 消息，即可实现对 RSVP 路径和预留状态进行刷新，减少了网络上的 Refresh 消息流量，并加快了节点对刷新消息的处理速度。

(3) RSVP 消息的可靠传递功能

RSVP 消息没有重传机制，消息丢失后发送端无法获悉，无法重传丢失的消息。通过 RSVP 消息的可靠性传递功能可以提高消息传递的可靠性。

RSVP 消息的可靠传递功能是指对端设备需要应答本端发送的 RSVP 消息，否则将会重传此消息。其工作机制为：节点发送了携带 Message_ID 对象的消息，且 Message_ID 对象的 ACK_Desired 标识（是否需要应答标识）置位后，如果在重传时间 Rf 内没有收到携带对应 Message_ID_ACK 对象的消息，则重传时间 Rf 超时后重传此消息，并将重传时间置为（1＋Delta）×Rf。节点持续按照上述方法重传此消息，直到节点在重传时间超时前接收到对应的应答消息，或消息传送次数达到 3 次。

2.4 采用PCE计算的路径建立CRLSP

在 MPLS TE 网络中，作为 PCC（Path Computation Client，路径计算客户端）的 LSR 需要获取到达目的地的 CRLSP 路径时，向 PCE（Path Computation Element，路径计算单元）发起路径计算请求，PCE 执行路径计算后对该请求进行应答，并提供计算后的路径。PCC 根据 PCE 计算后的路径使用 RSVP-TE 建立 CRLSP。

2.4.1 基本概念

- PCE：网络中的一个实体，用于为网络上的设备提供路径计算服务，可进行区域内的路径计算，也可在复杂的网络环境中计算完整的 CRLSP 路径，比如，在区域间的 ABR 上部署 PCE，用来计算跨区域的 CRLSP。PCE 分为以下两种类型：
  - Stateless PCE（Stateless Path Computation Element，无状态 PCE）：该类型 PCE 仅提供路径计算服务。

- Stateful PCE（Stateful Path Computation Element，有状态 PCE）：该类型 PCE 掌握了网络内所有 PCC 维护的 CRLSP 信息，可以重新计算和优化域内的 CRLSP，以达到最大程度分配和使用网络资源的目的。Stateful PCE 包括 Active-Stateful PCE（Active-Stateful Path Computation Element，主动有状态 PCE）和 Passive-Stateful PCE（Passive-Stateful Path Computation Element，被动有状态 PCE）两种类型。被动有状态 PCE 仅维护 PCC 的 CRLSP 信息，不能接受 PCC 的 CRLSP 托管并对 CRLSP 进行优化；主动有状态 PCE 可以接受 PCC 的 CRLSP 托管并对 CRLSP 进行优化。

- PCC：请求 PCE 执行路径计算，并根据 PCE 返回的路径信息建立 CRLSP。PCC 缺省为 Stateless PCC（Stateless Path Computation Client，无状态 PCC）。如果 PCE 为 Stateful PCE，PCC 也需要为对应的 Stateful 类型，即 Active-Stateful PCC（Active-Stateful Path Computation Client，主动有状态 PCC）和 Passive-Stateful PCC（Passive-Stateful Path Computation Client，被动有状态 PCC）。
- PCEP（Path Computation Element Protocol，路径计算单元通信协议）：运行于 PCC 与 PCE 之间、或者 PCE 与 PCE 之间的通信协议，用于建立 PCEP 会话，交互 PCEP 消息。该协议基于 TCP。

2.4.2 PCE 发现机制

PCE 的发现有两种方式：

- 静态指定：在 PCC 上静态指定 PCE。
- 动态发现：通过 OSPF TE 通告 PCE 信息，使得网络上的其它 LSR 可自动发现 PCE。

2.4.3 PCE 路径计算方式

PCE 路径计算有两种方式：

- EPC（External Path Computation，外部路径计算）：此方式由单台 PCE 独立完成 CRLSP 的计算，通常用于区域内的路径计算。
- BRPC（Backward-Recursive PCE-Based Computation，反向递归路径计算）：此方式通过多台 PCE 配合完成 CRLSP 的计算，通常用于跨区域的路径计算。

下面以 BRPC 为例介绍 PCE 的路径计算过程。如图 9 所示，两台 ABR 分别被配置为 PCE 1 和 PCE 2。PCE 1 可计算 Area 0 和 Area 1 内的路径，PCE 2 可计算 Area 0 和 Area 2 内的路径。当 PCC 需要获取到达 Area 2 的 CRLSP 路径时，路径计算步骤为：

(1) PCC 向 PCE 1 发起路径计算请求。

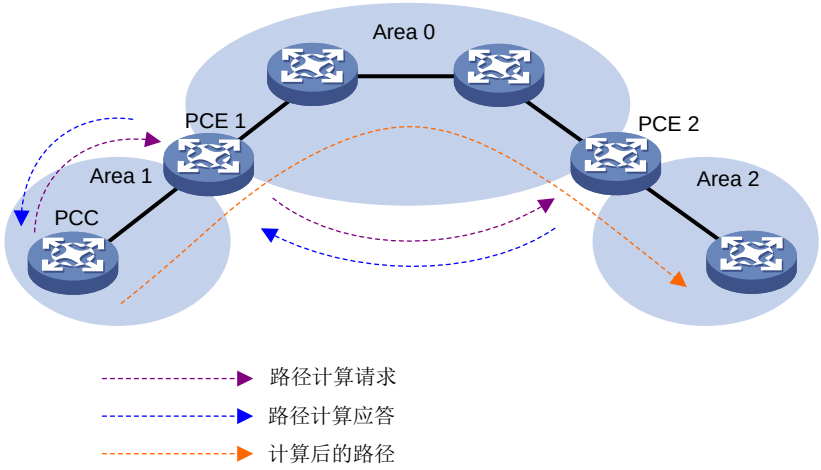
(2) PCE 1 收到该请求后，发现无法计算 Area 2 内路径，则继续向 PCE 2 发起到达 Area 2 的路径计算请求。

(3) PCE 2 应答该请求，并提供到达 Area 2 的路径。

(4) PCE 1 收到 PCE 2 的应答后，汇总路径信息，并对 PCC 的路径请求进行应答，提供到达 Area 2 的路径。

(5) PCC 根据 PCE 计算后的路径使用 RSVP-TE 建立 CRLSP。

图9 路径计算过程示意图



2.5 数据转发

当 MPLS TE 隧道建立之后，流量不会自动通过 MPLS TE 隧道转发，需要通过如下方法配置流量沿 MPLS TE 隧道转发。

2.5.1 静态路由指定

使用静态路由转发流量，是指在 MPLS TE 隧道的头节点定义一条到达目的网络地址、通过 TE Tunnel 接口的静态路由，以便将流量引入到 TE Tunnel 上进行转发。

静态路由是将流量引入 MPLS TE 隧道的最简便、直观的方法。该方法的缺点是：如果多个目的网络的流量都需要引入到 MPLS TE 隧道上，则需要配置多条静态路由，配置和维护难度比较大。

2.5.2 策略路由指定

策略路由指定是指在 MPLS TE 隧道的头节点定义策略路由，在策略路由中将匹配 ACL 规则的流量的出接口指定为 Tunnel 接口，并在流量的入接口上应用该策略路由，从而实现将流量引入到 MPLS TE 隧道上进行转发。

策略路由方式不仅可以根据目的 IP 地址来匹配需要通过 Tunnel 接口转发的流量，还可以根据源 IP 地址、协议类型等来匹配流量。与静态路由方式相比，策略路由方式更加灵活，但是配置比较复杂。

2.5.3 自动路由发布

自动路由发布是指将 MPLS TE 隧道发布到 IGP（OSPF 或 IS-IS）路由中，让 MPLS TE 隧道参与 IGP 路由的计算，使得流量可以通过 MPLS TE 隧道转发。自动路由发布方式的配置和维护都比较简单。

自动路由发布包括以下两种方式：

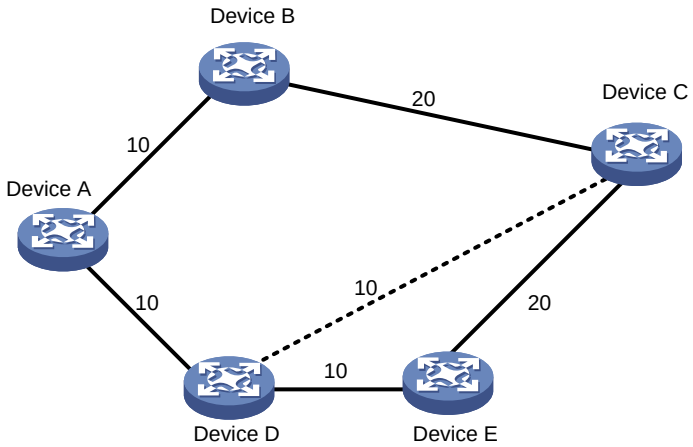
- IGP Shortcut：也称为自动路由宣告（AutoRoute Announce），该功能将 MPLS TE 隧道当作一条直接连接隧道 Ingress 节点（头节点）和 Egress 节点（尾节点）的链路，在隧道的 Ingress 节点上进行 IGP 路由计算时考虑该 MPLS TE 隧道。

- 转发邻接：该功能将 MPLS TE 隧道当作一条直接连接隧道 Ingress 节点和 Egress 节点的链路，通过 IGP 路由协议将该链路发布到网络中，以便网络中的节点在路由计算时使用 MPLS TE 隧道。

IGP Shortcut 和转发邻接功能的区别在于：

- 在隧道的 Ingress 节点上开启 IGP Shortcut 功能后，只有 Ingress 节点计算 IGP 路由时会考虑 MPLS TE 隧道。IGP Shortcut 功能不会通过 IGP 路由协议将 MPLS TE 隧道作为一条链路发布出去。因此，其他设备在路由计算时不会考虑 MPLS TE 隧道。
- 在隧道的 Ingress 节点上开启转发邻接功能后，Ingress 节点会通过 IGP 路由协议将 MPLS TE 隧道作为一条链路发布出去。因此，IGP 网络中的所有设备在路由计算时都会考虑 MPLS TE 隧道。

图10 IGP Shortcut 与转发邻接示意图



如图 10 所示，Device D 到 Device C 之间存在一条 MPLS TE 隧道，IGP Shortcut 只能使 Ingress 节点 Device D 在计算 IGP 路由时利用这条隧道，Device A 并不能利用这条隧道到达 Device C。如果配置了转发邻接功能，则 Device A 也能够知道这条 MPLS TE 隧道的存在，从而可以利用该隧道将到 Device C 的流量转发到 Device D 上。

3 Comware 实现的技术特色

3.1 make-before-break

make-before-break 是一种在尽可能不丢失数据，也不占用额外带宽的前提下改变 MPLS TE 隧道的机制。

在隧道重优化、自动带宽调整等情况下，如果在新的 CRLSP 建立之前拆除旧的 CRLSP，则会导致流量转发中断。通过 make-before-break 机制可以确保新 CRLSP 建立、并将流量切换到新的 CRLSP 后，再拆除旧 CRLSP，从而有效地避免流量转发中断。此时，存在的问题是：如果新的 CRLSP 和旧 CRLSP 部分路径相同，则在这些路径上需要重复为新旧 CRLSP 预留带宽，造成带宽资源的浪费。make-before-break 机制采用 SE 资源预留风格解决这个问题。

资源预留风格是 RSVP-TE 协议在建立 CRLSP 时预留带宽资源的方式。MPLS TE 隧道使用的资源预留风格由隧道的 Ingress 节点决定，并通过 RSVP 协议通知给各个节点。

目前，设备支持以下两种资源预留风格：

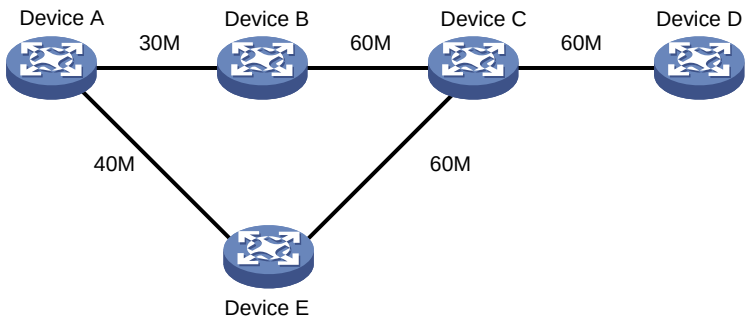
- FF（Fixed-Filter，固定过滤器）：为每个发送者单独预留资源，同一会话中的不同发送者不能共享资源。
- SE（Shared-Explicit，共享显式）：为同一个会话中的不同发送者预留同一个资源，不同发送者之间可以共享资源。该方式主要用于 make-before-break。

在图 11 中，假设需要建立一条 Device A 到 Device D 的 CRLSP，保留 30M 带宽，起初建立的路径是 Device A-Device B-Device C-Device D。

现在希望将带宽增大为 40M，Device A-Device B-Device C-Device D 路径不能满足要求。而如果选择 Device A-Device E-Device C-Device D，则 Device C-Device D 需要同时预留 30M 和 40M 带宽，也存在带宽不够的问题。

采用 make-before-break 机制，新建立的 CRLSP 在 Device C-Device D 可以共享原 CRLSP 的带宽，不需要为新 CRLSP 和旧 CRLSP 重复预留带宽。新 CRLSP 建立成功后，流量切换到新 CRLSP 上，之后拆除原 CRLSP，从而有效地避免了流量中断。

图11 make-before-break 示意图



3.2 路由固定

路由固定是指 CRLSP 创建成功后，即使路由发生变化，也不重新选择最优路径，而是沿用已创建的 CRLSP。

在路由变化频繁的网络中，如果不希望 CRLSP 随着路由频繁变化，则可以通过本功能确保只要已建立的 CRLSP 可用就不重新创建 CRLSP。

3.3 隧道重优化

流量工程一个主要的目标就是优化网络上流量的分布。隧道建立之后，可以根据网络上的带宽变化、流量变化、管理策略变化等对已经建立的 CRLSP 隧道进行优化。

在优化时，用户的业务流不中断是非常重要的，即新的 CRLSP 隧道必须先建立，业务在旧的 CRLSP 隧道被拆除前切换到新的 CRLSP 隧道上。在新旧 CRLSP 隧道共享的链路上，由于旧的 CRLSP 隧道使用的资源不能在新的 CRLSP 隧道建立前释放，共享链路上资源不能被重复计算，否则可能会由于资源缺乏而导致新的 CRLSP 隧道无法建立。

RSVP-TE 信令的 SE 预留风格能够非常好地解决这个问题。SE 预留风格允许新旧的 CRLSP 隧道共享资源，使新的 CRLSP 隧道不会因为链路资源缺乏而必须等到旧的 CRLSP 隧道拆除才能建立。

3.4 自动带宽调整

通常情况下，用户最初不能确定有多少业务需要通过服务提供商的网络传输。因此，服务提供商需要具备这样一种功能：能在最初时为用户请求的带宽建立 MPLS TE 隧道；当用户业务增多时，能够根据用户的业务量自动调整分配给 MPLS TE 隧道的带宽，这种调整不会影响当前通过隧道的流量。

自动带宽调整功能的工作机制为：

(1) 设备定时地对隧道的出口速率进行采样，计算采样时间间隔内的平均出口速率；

(2) 自动带宽调整时间间隔到达后，将隧道的带宽设置为该时间间隔内多次采样中的最大平均出口速率；

(3) 根据调整后的隧道带宽建立新的 CRLSP；

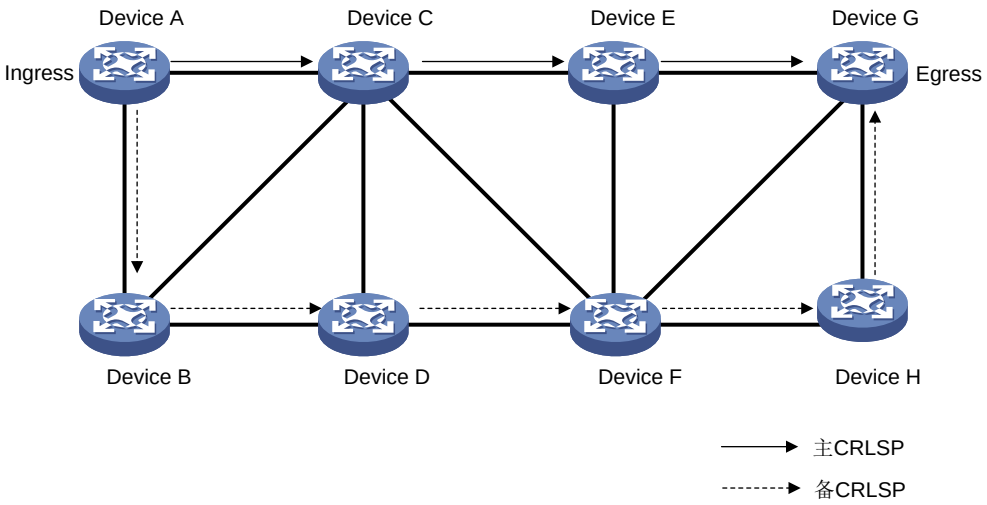
(4) CRLSP 建立成功后，将流量从旧的 CRLSP 切换到新的 CRLSP，并拆除旧的 CRLSP。

为了减少不必要的调整，用户可以指定允许调整到的最大带宽值和最小带宽值。如果自动带宽调整计算出的隧道带宽大于最大值，则采用最大带宽值建立新的 CRLSP；如果小于最小带宽值，则采用最小带宽值建立新的 CRLSP。

3.5 CRLSP备份

CRLSP 备份是指通过备份 CRLSP 对主 CRLSP 进行保护。如图 12 所示，当 Ingress 感知到主 CRLSP 故障时，将流量切换到备份 CRLSP 上，当主 CRLSP 路径恢复后再将流量切换回来，以实现对主 CRLSP 的备份保护。

图12 CRLSP 及其备份



1. 主 CRLSP

主 CRLSP 是期望的最优的路径，它是 CRLSP 备份的对象。主 CRLSP 故障后，直接将流量切换至备份 CRLSP，并重新建立主 CRLSP。

2. 热备份 CRLSP

热备份 CRLSP 在主 CRLLSP 建成之后，发起建立。当主 CRLSP 故障后，流量会切换到热备份 CRLSP。当主 CRLLSP 恢复后，将流量切换回去。

热备份通过建立额外的 CRLSP，消耗成倍的资源，来保证主 CRLSP 发生故障时快速进行切换。热备份适用于对时延敏感的业务。

3. 普通备份 CRLSP

普通备份 CRLSP 在主 CRLSP 故障后，发起建立。建立成功后，将流量切换到该 CRLSP 上。当主 CRLLSP 恢复后，将流量切换回去。

普通备份中备份 CRLSP 和主 CRLSP 通过的路径不同，从而绕过可能的失效链路。但是由于没有提前建立备份 CRLSP，普通备份的切换速度较慢，适用于对时延不敏感的业务。

3.6 快速重路由

3.6.1 功能简介

FRR（Fast Reroute，快速重路由）是 MPLS TE 中实现网络局部保护的技术。FRR 的切换速度可以达到 50ms，能够最大程度减少网络故障时数据的丢失，以满足时延敏感业务（如 VoIP）的需求。。

开启 FRR 功能后，当 CRLSP 链路或者节点失效时，通过保护链路或者节点的 Bypass 隧道继续转发流量，以保证数据传输不中断。头节点在数据传输不受影响的同时继续发起主 CRLSP 的重建。FRR 的最终目的就是利用 Bypass 隧道绕过失效的链路或者节点，从而达到保护主路径的目的。

3.6.2 基本概念

- 主 CRLSP：被保护的 CRLSP。
- Bypass 隧道：旁路隧道，保护主 CRLSP 中某条链路或某个节点的 MPLS TE 隧道。
- PLR（Point of Local Repair，本地修复节点）：Bypass 隧道的 Ingress 节点，必须在主 CRLSP 的路径上，并且不能是主 CRLSP 的 Egress 节点。
- MP（Merge Point，汇聚点）：Bypass 隧道的 Egress 节点，必须在主 CRLSP 的路径上，并且不能是主 CRLSP 的 Ingress 节点。

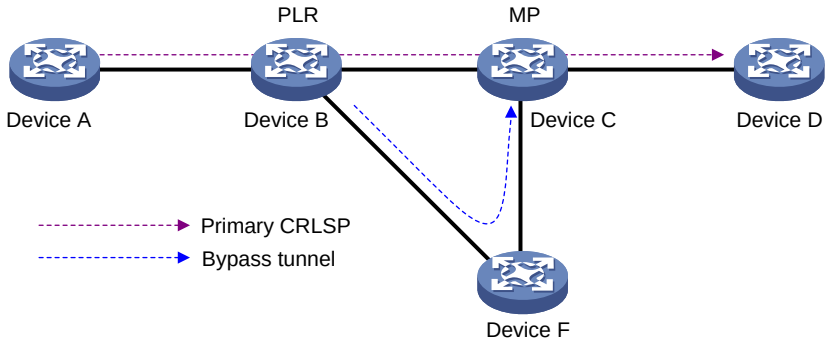
3.6.3 保护方式

根据保护的对象不同，FRR 分为链路保护和节点保护两类。

1. 链路保护

又称为 Next-hop（NHOP）保护。PLR 和 MP 之间有直接链路连接，主 CRLSP 经过这条链路。当这条链路失效时，流量可以切换到 Bypass 隧道上。如图 13 所示，主 CRLSP 是 Device A-Device B-Device C-Device D，Bypass 隧道是 Device B-Device F-Device C。

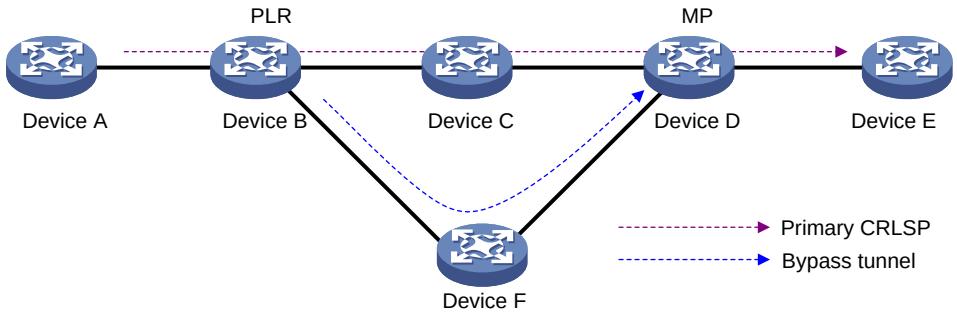
图13 FRR 链路保护示意图



2. 节点保护

又称为 Next-next-hop（NNHOP）保护。PLR 和 MP 之间通过一台设备连接，主 CRLSP 经过这台设备。当这台设备失效时，流量可以切换到 Bypass 隧道上。如图 14 所示，主 CRLSP 是 Device A-Device B-Device C-Device D-Device E，Bypass 隧道是 Device B-Device F-Device D，Device C 是被保护的设备。

图14 FRR 节点保护示意图



3.6.4 FRR 的切换时间

FRR 的切换时间由两部分组成：

- 一部分是链路/节点失效的检测时间，可以通过硬件检测、BFD 或 RSVP hello 来进行检测。失效检测速度从高到低依次为硬件检测、BFD 检测和 RSVP hello 检测。
- 另一部分是切换流量的时间，该时间主要由 CPU 以及系统的负载程度来决定。

3.6.5 FRR 的局限性

FRR 的保护是一种临时性措施，因为它可能无法提供足够的带宽等资源，或者会给其他链路带来拥塞。在被保护 LSP 恢复正常后，FRR 将不起作用。

FRR 的 Bypass 隧道不能够提供首尾节点的保护，这种保护可以通过前面介绍的 CRLSP 备份来实现。

另外，同时有多个链路/节点失败的情况下，FRR 本身也可能失效。

3.7 DiffServ-Aware TE

3.7.1 功能简介

DiffServ 作为一种 QoS 解决方案，其主要实现机制是对流量按照服务类型（class of service）进行划分，基于服务类型提供不同的 QoS 保证。而 MPLS TE 作为流量工程解决方案，主要用于对网络资源的使用进行优化。

DiffServ-Aware TE，简称 DS-TE，结合上述两者的优势，能够基于按服务类型划分的流量进行网络资源优化，即对不同的服务类型进行不同的带宽约束。概括来说，DS-TE 将不同服务类型的流量与 CRLSP 进行映射，使流量经过的路径符合对其服务类型的流量工程约束条件。

目前，Comware 支持两种 DS-TE 模式：

- 自定义的 Prestandard 模式
- 根据 RFC 4124、RFC 4125、RFC 4127 实现的 IETF 模式

3.7.2 DS-TE 基本概念

- CT（Class Type，服务类型）：流量所属的业务类别，用来实现对不同的流量进行分类。DS-TE 根据业务流所属的 CT 为其分配链路带宽、实施约束路由及进行准入控制。对于一个给定的业务流，在其经过的所有链路上，该业务流都属于相同的 CT。
- BC（Bandwidth Constraint，带宽约束）：用来对各种服务类型流量所能使用的带宽进行限制。
- 带宽约束模型（Bandwidth Constraint Model）：用来实现对不同 CT 的业务流进行带宽约束的算法。带宽约束模型由两部分内容决定：最大 BC 数目、BC 与 CT 的对应关系。DS-TE 支持三种带宽约束模型 RDM（Russian Dolls Model，俄罗斯套娃模型）、MAM（Maximum Allocation Model，最大分配模型）和 Extended-MAM（Extended Maximum Allocation Model，扩展的最大分配模型）。
- TE class：CT 及优先级的组合。如果流量属于某个 CT，则传输该流量的 MPLS TE 隧道的建立优先级或保持优先级必须是该 CT 对应的优先级。

说明

Prestandard 模式和 IETF 模式具有如下区别，请根据服务类型的数量、所需带宽约束模型等选择合适的 DS-TE 模式。

- Prestandard 模式支持 2 个 CT（CT 0 和 CT 1），8 种优先级，最大支持 16 个 TE class；IETF 模式支持 8 个 CT（CT 0 ~ CT 7），8 种优先级，最大支持 16 个 TE class。
- Prestandard 模式下不可以通过配置改变 TE class；IETF 模式下可以通过配置改变 TE class。
- Prestandard 模式只支持 RDM 模型；IETF 模式支持 RDM 模型、MAM 模型和 Extended-MAM 模型。
- Prestandard 模式为自定义模式，无法与所有厂商设备互通；IETF 模式为根据 RFC 标准实现的模式，可以与其他厂商设备互通。

3.7.3 DS-TE 工作原理

根据流量的服务类型建立 MPLS TE 隧道的过程如下：

(1) 判断流量所属的 CT

设备上根据配置实现不同业务流量的分类：

- 对于动态建立的 MPLS TE 隧道，在隧道接口下执行 **mpls te bandwidth** 命令，可以配置通过该隧道接口的流量所属的 CT。
- 对于静态建立的 MPLS TE 隧道，配置静态隧道时，可以通过 **bandwidth** 参数指定通过该静态隧道转发的流量所属的 CT。

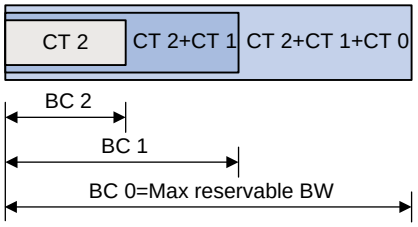
(2) 检查 CT 对应的 BC 中是否存在足够的带宽

用户可以在接口下通过 **mpls te max-reservable-bandwidth** 命令，配置该接口的带宽限制。设备根据流量所属的 CT 及接口的带宽限制，判断是否存在足够的带宽为该流量建立 MPLS TE 隧道。

不同带宽约束模型下，BC 与 CT 的关系不同：

- RDM：限制多种服务类型流量（CT）的共用带宽，允许多种 CT 间共享使用带宽，而不是限制某一种 CT 的带宽。如图 15 所示，以三个 CT（CT 0、CT 1 和 CT 2）为例，BC 2 为 CT 2 的带宽限制，即属于 CT 2 流量的带宽总和不能超过 BC 2；BC 1 为 CT 2 和 CT 1 两种业务的带宽限制，即属于 CT 2 和 CT 1 流量的带宽总和不能超过 BC 1；BC 0 为 CT 2、CT 1 和 CT 0 三种业务的带宽限制，即属于 CT 2、CT 1 和 CT 0 流量的带宽总和不能超过 BC 0。在 RDM 中，BC 0 即为链路的最大可预留带宽。RDM 与建立优先级/保持优先级配合，可以实现 CT 间的带宽隔离。RDM 比较适用于属于 CT 的流量不平稳、可能存在突发流量的情况。

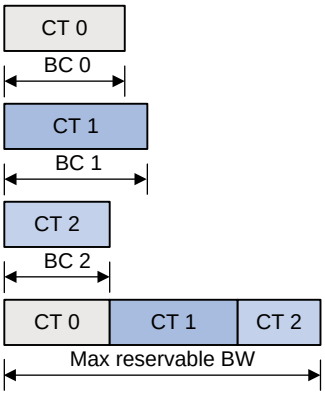
  图15 RDM 带宽约束模型示意图



- MAM：限制某一 CT 在接口上占用的带宽总和，即隔离 CT 之间的带宽使用。如图 16 所示，以三个 CT（CT 0、CT 1 和 CT 2）为例，BC 0 为 CT 0 的带宽限制，即属于 CT 0 流量的带宽总和不能超过 BC 0；BC 1 为 CT 1 的带宽限制，即属于 CT 1 流量的带宽总和不能超过 BC 1；以此类推。并且，属于 CT 0、CT 1 和 CT 2 流量的带宽总和不能超过最大可预留带宽。MAM 不需要与建立优先级/保持优先级配合，就可以实现 CT 间的带宽隔离。MAM 的特点是比较直观，配置较为容易。MAM 比较适用于属于 CT 的流量较为平稳、不存在突发流量的情况。
- Extended-MAM：和 MAM 模式类似，Extended-MAM 限制某一 CT 在接口上占用带宽总和，CT 间不共享带宽。与 MAM 模型不同，Extended-MAM 支持 8 个 CT，即 CT 0～CT 7，且支持在一条 LSP 上为多个 CT 预留带宽。Extended-MAM 支持 16 个 TE Class，前 8 个 TE

Class 的默认映射与 MAM 模型相同，后 8 个 TE Class 为 CT0 与优先级 0～7 的映射，对应 TE Class [8]～TE class[15]，且不支持用户配置。

图16 MAM/Extended-MAM 带宽约束模型示意图



(3) 检查流量是否与已经存在的 TE class 匹配

根据服务类型建立 MPLS TE 隧道时，还需要检查流量所属的 CT 及 LSP 的建立优先级/保持优先级是否与已经存在的 TE class 匹配。要想为该流量建立隧道，必须同时满足下面两个条件：

- 隧道经过的节点上都存在与流量所属 CT、LSP 建立优先级匹配的 TE class；
- 隧道经过的节点上都存在与流量所属 CT、LSP 保持优先级匹配的 TE class。

3.8 CBTS

3.8.1 CBTS 简介

CBTS（Class-based Tunnel Selection，基于服务类型的隧道选择）有别于传统的隧道选择方式，它基于流量的隧道转发类选择相对应的隧道进行转发，以便根据业务的不同提供不同的转发服务。

3.8.2 CBTS 工作原理

CBTS 工作原理为：

(1) 在设备入方向上通过流行为指定流量所属的隧道转发类。

(2) 配置隧道的隧道转发类（Service-class 属性），与隧道转发类匹配的流量可以通过该隧道转发。

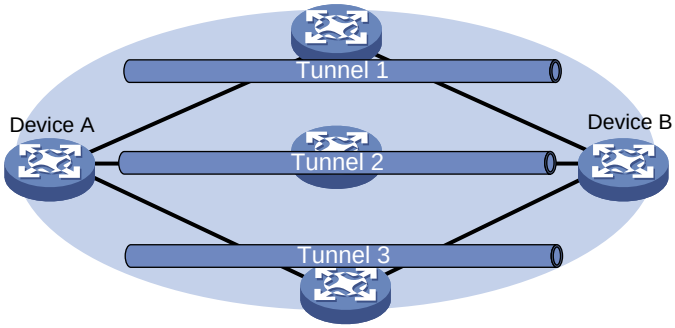
3.8.3 CBTS 优选规则

CBTS 的优选规则为：

- 设备会优先选择与流量的隧道转发类值相同的隧道转发该流量。
- 如果存在多条与流量的隧道转发类值相同的隧道，只有一条流且为逐流转发则随机选择一条隧道转发；有多条流或者一条流但是为逐包转发，则相同转发类的隧道进行负载分担。
- 如果没有与流量的隧道转发类值相同的隧道，则选择隧道转发类值最小的隧道转发流量，未配置隧道转发类的隧道转发类值最小。

3.8.4 CBTS 示例

图17 CBTS 示意图



Tunnel 1隧道转发类：未配置

Tunnel 2隧道转发类：3

Tunnel 3隧道转发类：6

如图 17 所示，隧道的选择原则为：

- 从 Device A 到 Device B 隧道转发类值为 3 的流量通过 Tunnel2 转发。
- 从 Device A 到 Device B 隧道转发类值为 6 的流量通过 Tunnel3 转发。
- 从 Device A 到 Device B 未配置隧道转发类的流量通过 Tunnel1 转发。

3.9 非均衡负载分担

当设备上存在到达同一目的地的多条等价隧道时，设备在转发去往该目的地的报文时，依次通过各条隧道发送报文，从而实现流量的负载分担。在 Tunnel 接口下配置隧道非均衡负载分担带宽值，可以使得到达同一目的地址的多条等价 MPLS TE 隧道能够按照指定的负载分担比例转发流量，更好地满足业务的需求和资源的整合。例如，到达某一目的地址存在三条等价隧道：Tunnel1、Tunnel2 和 Tunnel3，这三条隧道的非均衡负载分担带宽值分别为 10000kbps、10000kbps 和 20000kbps，则三条隧道承担的到达此目的地址的流量比重分别为 1/4、1/4 和 1/2。

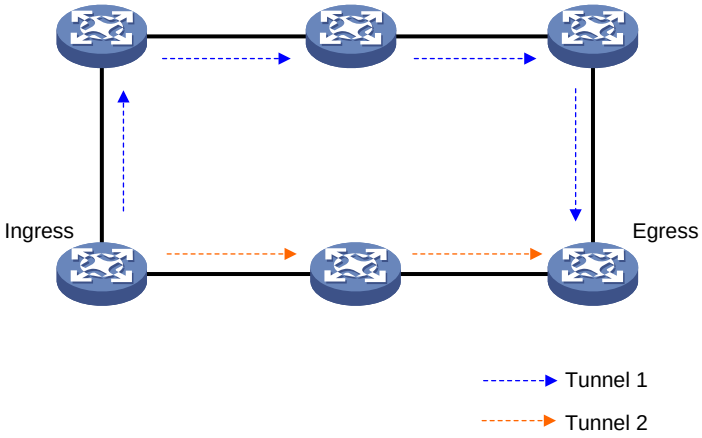
4 典型组网应用

4.1 带宽保证

如果网络中两点间的不同业务要求不同的带宽保证，可以通过部署到同一目的地的多条 TE 隧道，实现为每种业务提供独立的带宽保证。

如图 18 所示，在 Ingress 和 Egress 之间存在数据和语音两种业务，通过建立 TE 隧道分别为两种业务提供服务。Tunnel 2 所在的链路为低延迟链路，可以使用 Tunnel 2 为语音业务提供带宽保证；其他数据业务通过 Tunnel 1 转发，避免数据与语音经过相同的路径造成拥塞。

图18 带宽保证

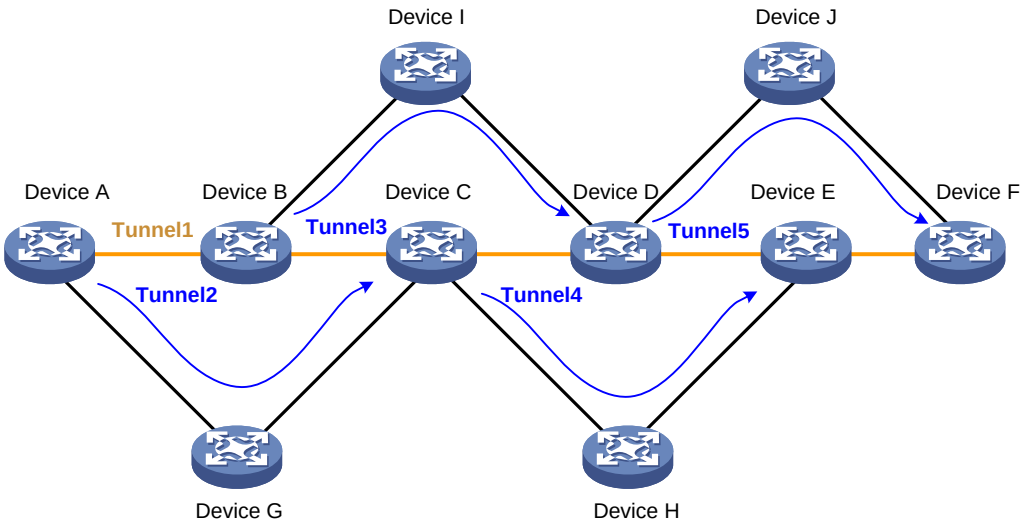


4.2 MPLS TE FRR组网

通过 FRR 对网络中的关键节点或链路进行保护，从而实现对通过关键节点的流量的保护。

如 4.2 图 19 所示，运营商使用 MPLS TE 隧道接入用户，使异地的用户网络通过运营商网络连接。因为承载业务为重要的业务，因此需要对主 CRLSP Tunnel1 经过的路径进行保护。利用 MPLS TE FRR，可以实现通过保护路径 Tunnel2、Tunnel3、Tunnel4 和 Tunnel5 分别保护主 CRLSP 路径上的 Device B、Device C、Device D 和 Device E。

图19 MPLS TE FRR 组网

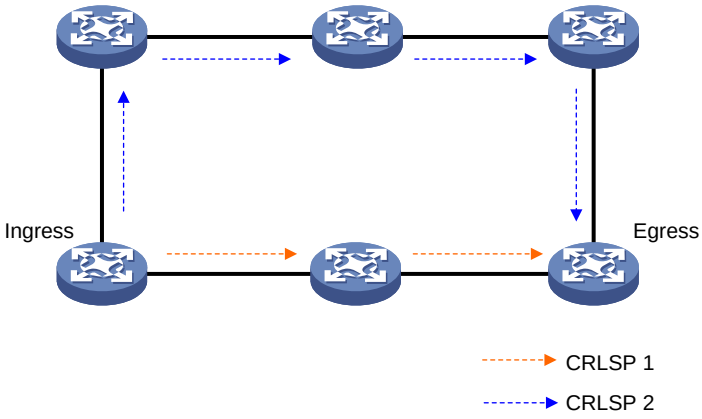


4.3 CRLSP备份组网

CRLSP 备份用于端到端的路径保护，对整条 CRLSP 提供保护。

如 4.3 图 20 所示，为 CRLSP 1 建立备份路径 CRLSP 2，当检测到 CRLSP 1 发生故障时，可以将通过 CRLSP 1 的流量切换到备份路径 CRLSP 2 上，实现对整条 CRLSP 1 路径的保护。

图20 CRLSP 备份组网



4.4 TE隧道与MPLS VPN结合

建立的域内或跨域的 TE 隧道可以应用于 MPLS VPN。

如图 21 和图 22，对于 MPLS L2VPN/L3VPN，TE 隧道作为其公网隧道可以为其提供带宽保证。TE 隧道也可以将不同 VPN 业务隔离到不同的隧道中，提供不同的带宽保证和 QoS 服务。

图21 VPN over TE 组网

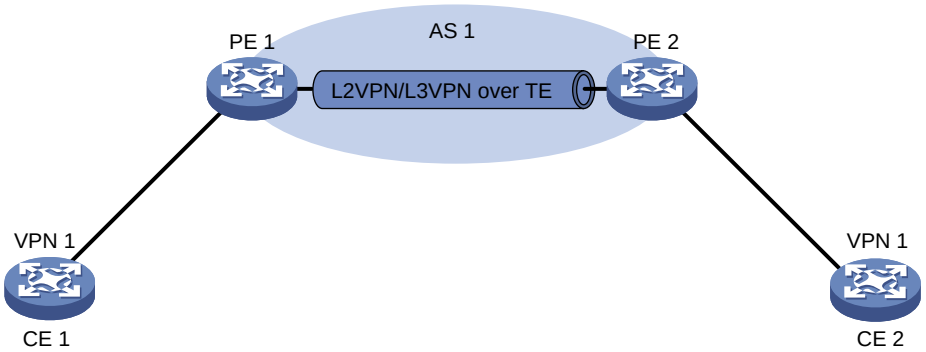
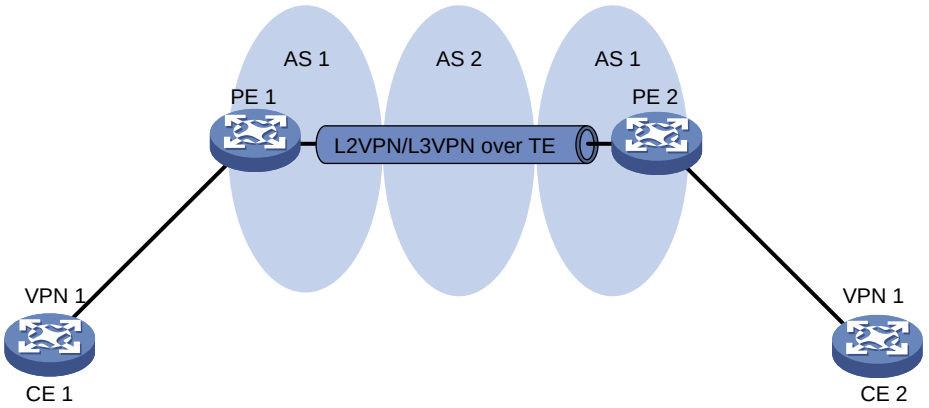


图22 VPN over TE 组网（跨域）



5 参考文献

- RFC 2702：Requirements for Traffic Engineering Over MPLS
- RFC 3564：Requirements for Support of Differentiated Service-aware MPLS Traffic Engineering
- RFC 3812：Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Management Information Base (MIB)
- RFC 4124：Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering
- RFC 4125：Maximum Allocation Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering
- RFC 4127：Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering
- ITU-T Recommendation Y.1720：Protection switching for MPLS networks
- RFC 4655：A Path Computation Element (PCE)-Based Architecture
- RFC 5088：OSPF Protocol Extensions for Path Computation Element Discovery
- RFC 5440：Path Computation Element (PCE) Communication Protocol (PCEP)
- RFC 5441：A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering LSP
- RFC 5455：Diffserv-Aware Class-Type Object for the Path Computation Element Communication Protocol
- RFC 5521：Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions
- RFC 5886：A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture
- draft-ietf-pce-stateful-pce-07