TP-LINK®

QoS 技术白皮书

TP-LINK 管理型交换机

REV1.0.0
1910040556

目录

1 概述

1.1 QoS 产生背景

互联网发展日新月异，IP 网络上的新应用层出不穷。除了传统的 WWW、FTP 、E-Mail 应用外，IP 电话、电子商务、多媒体游戏、远程教学、远程医疗、可视电话、电视会议、视频点播、在线电影等各种新业务也已经被广泛应用在我们的生活中。这些新业务有一个共同的特点，即对带宽、延迟、抖动和丢包率等传输性能有着特殊的需求， 比如电视会议、视频点播等业务需要高带宽、低延迟和低延迟抖动。这就要求网络能够将关键业务与普通业务区分开来，为不同的通信业务提供不同的服务。

而在传统的 IP 网络中，所有的报文都被无区别地等同对待，每个转发设备对所有的报文均采用先入先出（FIFO）的策略进行处理，它尽最大的努力（Best-Effort）将报文送到目的地，但对报文传送的丢包率、传送延迟等性能不提供任何保证。

QoS 技术就是在这种背景下发展起来的。QoS 是 Quality of Service（服务质量）的简称，其目的是针对各种业务的不同需求，为其提供有区别的服务质量。

1.2 QoS 度量指标

每种业务都对网络有特定的要求，业务 A 可能对某些指标的要求高一些，业务 B 则可能对另一些指标的要求高一些。QoS 旨在为特定的业务提供其所需要的服务。

QoS 采用如下指标来度量网络为各项业务提供的服务质量：

- 带宽
- 时延
- 抖动
- 丢包率

带宽

指网络的两个节点之间特定应用业务流的平均速率，单位为比特 / 秒（bps）。我们可以将带宽类比为高速公路：一般用高速公路上的车道数来衡量其可承载的车流量，车道越多，则同一时间内可承载的车流量越大。这里车道数就好比是带宽，车流量则好比是网络传输的数据。

带宽的大小直接影响用户的上网体验，特别是实时业务，对带宽有更高的要求。例如，当视频业务的可用带宽低于视频源的编码速率时，图像质量就无法保证。

时延

指报文从网络的一端传送到另一端所需要的时间。

实时性业务如 VoIP、在线视频、在线对抗游戏等对时延有较高的要求。以 VoIP 为例，时延就是从说话者开始说话到对方听到所说内容之间的间隔。一般人们能忍受小于 250ms 的时延，时延太大会导致通话声音不清晰、不连贯，通话过程变成类似对讲机式的对话。

抖动

也叫时延变化，指同一业务流中不同报文时延不同。抖动主要是由业务流中前后报文的排队等候时间不同引起的。

与时延的区别：抖动是时延的变化。以 VoIP 为例，若时延为 300ms 而无抖动，表现为需要等待一会才能听到对方说话，但听到的每一句话都是连贯的；若时延为 300ms 且带抖动，则除了开始的停顿外，还表现为听到的语句是断断续续的。

实时性的业务，如 VoIP 和在线视频对抖动非常敏感，稍许的抖动就会导致通话质量或视频质量迅速下降。

丢包率

指在网络传输过程中丢失报文占总传输报文的百分比。

FTP、电子商务网上支付等业务对丢包率很敏感，一般使用有重传机制的 TCP 协议以保证数据完整性。而视频电话会议、流媒体等业务次之，一般要求满足可预计的丢包率。

1.3 QoS 服务模型

网络中的通信本质上都属于端到端的通信。两个主机要进行通信，中间可能要跨越多个网络设备，所以，要实现端到端的 QoS，必须要从全局考虑。QoS 服务模型就是用于表示采用什么样的模式来实现全局 QoS 的一个概念。

根据网络对应用程序的控制能力的不同，QoS 有如下三种模型：

- Best-Effort 模型
- IntServ 模型
- DiffServ 模型

Best-Effort 模型

Best-Effort 模型（Best-Effort Service，尽力而为服务模型）是最简单的服务模型。应用程序可以在任何时候，发出任意数量的报文，而且不需要事先获得批准，也不需要通知网络。

对 Best-Effort 服务，网络尽最大的可能性来发送报文，但对带宽、时延、丢包率等性能不提供任何保证。

Best-Effort 模型是现在 Internet 的缺省服务模型，适用于绝大多数网络应用，如 FTP、E-Mail 等。它通过 FIFO 队列来实现。

IntServ 模型

IntServ 模型（Integrated Service，综合服务模型）可以满足多种 QoS 需求。这种服务模型在发送报文前，需要向网络申请资源预留，确保网络能够满足数据流的特定服务要求。

IntServ 模型中网络资源的申请是通过 RSVP（Resource ReservationProtocol，资源预留协议）信令来完成的，过程如下：

1) 应用程序通过 RSVP 信令向网络描述它自己的流量参数，申请特定的 QoS 服务（包括带宽、时延等性能要求）。

2) 网络收到应用程序的资源请求后，执行资源分配检查，即基于应用程序的资源申请和网络现有的资源情况，判断是否为应用程序分配资源，然后决定是否回复确认信息。

3) 应用程序在收到网络的确认信息（即确认网络已经为这个应用程序的报文预留了资源）后才开始发送报文。

4) 网络为每个应用流维护状态信息，只要该应用程序的报文控制在流量参数描述的范围内，网络将承诺满足应用程序的 QoS 需求。

IntServ 模型为应用程序提供了一套端到端的保障制度，保证满足应用程序的带宽、时延等性能需求，优点显而易见。但是，其缺点也一样很明显：

- IntServ 模型中端到端经过的所有的网络节点都必须支持 RSVP 信令协议，但核心层、汇聚层和接入层的设备功能参差不齐，很难统一；
- IntServ 模型要求为每个提出申请的应用流预留连接路径上的带宽资源，而当前 Internet 上有千千万万条应用流，如果所有的应用流都申请 QoS 服务，网络就肯定应付不过来了。

正是由于上述的局限性，IntServ 模型不适用于大规模网络。

DiffServ 模型

DiffServ 模型（Differentiated Service，区分服务模型）是一个多服务模型，它可以满足不同的 QoS 需求。在 DiffServ 模型中，网络会根据服务要求对不同业务的数据进行分类，对报文按类进行优先级标记，然后有差别地提供服务。

与 IntServ 模型不同，DiffServ 模型不需要通知网络为其预留资源，也不需要为每个流维护状态，它根据每个报文的自身携带的服务等级参数信息来提供服务，有差别地控制和转发流量。

DiffServ 模型一般用来为一些重要的应用提供端到端的 QoS，它通过下列技术来实现：

- 流量标记与控制技术：它可以根据数据帧的 802.1Q Tag 中的 CoS（Class of Service，服务等级）域，或者 IP 报文的 ToS（Type of Service，服务类型）域以及五元组（协议、

源地址、目的地址、源端口号、目的端口号）等信息进行报文分类，完成报文的标记和流量监管；

- 拥塞管理技术：采用 SP（Strict Priority，严格优先级）、WRR（Weighted Round Robin，加权轮询优先级）、SP+WRR 等队列技术对拥塞的报文进行缓存和调度实现拥塞管理；
- 拥塞避免技术：采用 TD（Tail Drop，尾丢弃）、RED（Random Early Detection，早期随机检测）、WRED（Weighted Random Early Detection，加权早期随机检测）等算法，在交换机缓存区被填满之前对各端口各队列的报文进行随机丢弃，避免拥塞的产生。

IntServ 模型中端到端建立连接后，沿途的各网络节点需要一同预留出网络资源并维护该连接的状态信息。而 DiffServ 模型利用报文自身携带的信息将流量分为有限的几种类别，把全局参与的 QoS 服务过程转化为单跳行为，实现简单，扩展性较好，是当前网络中的主流服务模型。TP–LINK 交换机的 QoS 功能所采用的，也正是 DiffServ 模型。

2 基于 DiffServ 模型的 QoS 技术实现

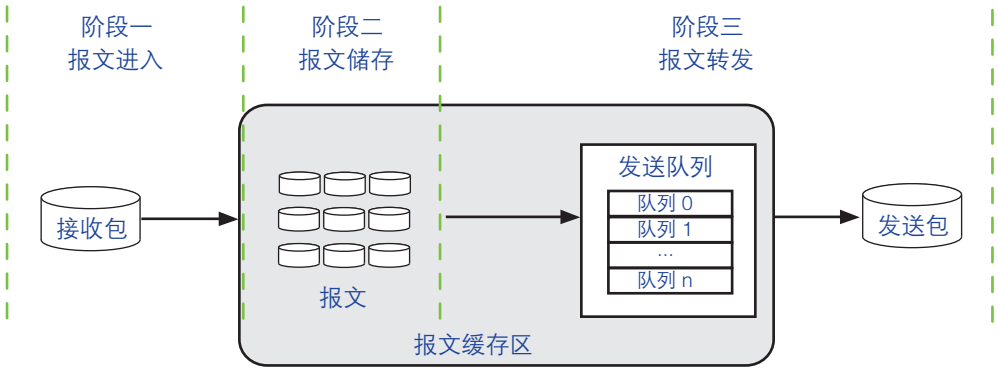
2.1 存储转发原理

交换机一般采用存储转发方式来实现网络接口间的数据交换。存储转发交换是交换机 QoS 功能实现的关键，下面我们将简单介绍交换机的存储转发原理以更好地理解 QoS 技术实现过程。

在存储转发交换中，所有的输入和输出端口都共享一个缓存模块，所有需要经过交换机的数据都在缓存模块中存储转发。交换机收到数据帧时会先将数据存储在缓存区中，直到收下完整的数据帧为止。存储过程期间，交换机会读取该数据帧的目的地址的信息，同时使用以太网帧的循环冗余校验（Cyclic Redundancy Check，简称 CRC）帧尾部分来进行错误检查。在确认了该数据帧的完整性之后，交换机才将该数据帧从对应的端口转发出去，发往目的地址。如果在该数据帧中检测到了错误，那么交换机将丢弃该数据帧以减少不必要的带宽损耗。

综上，交换机从接收报文到转发出去，经历了三个阶段：报文进入、报文储存和报文转发，如图 2-1 所示：

图 2-1　存储转发



QoS 主要作用于报文转发阶段。默认情况下，报文转发过程中只使用一个发送队列，报文按照“先进先出”的规则被转发出去。当启用 QoS 配置之后，剩余的发送队列也将发挥作用：系统根据报文首部的优先级字段将报文分配到不同的发送队列，每个队列按照一定的比例分享带宽，而这个比例由调度算法决定。

2.2 拥塞的产生与控制

就 Internet 的体系结构而言，发生拥塞是其固有属性。拥塞控制是 QoS 技术研究的重点内容，主要通过拥塞避免、流量控制和拥塞管理等技术来实现。

拥塞管理是指采用 WRR、SP、SP+WRR 等队列技术对拥塞的报文进行缓存和调度实现拥塞管理，详细内容请参考 QoS 处理流程之调度出列。下面将详细介绍拥塞的产生原因、拥塞避免技术和流量控制技术。

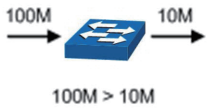
2.2.1 拥塞的产生

拥塞是分组交换网络中传送报文的数目太多时，由于存储转发节点的资源有限而造成网络传输性能下降的情况。

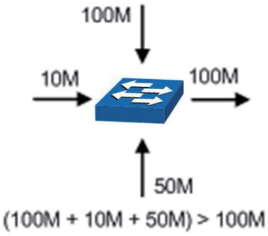
对于交换机而言，当其输入的报文流量大于输出的报文流量，缓存区中存储的报文数量逐步增加直至缓存区被填满时，交换机发生了拥塞。

交换机发生拥塞主要有如下两种情形：

- 高速端口向低速端口发送报文



- 多个端口向一个端口发送报文



当拥塞产生时，由于交换机的缓存区已经被填满，只有当缓存区中有一个报文被转发出去时才能有下一个报文进入缓存区，这样，很多发往非阻塞端口的报文因为不能进入缓存区也被丢弃。这种现象被称为队头阻塞（Head of Line Blocking）现象。

出现队头阻塞现象时，交换机缓存区报文的清除速度由阻塞的出端口的转发速率决定。通常此时交换机其他端口的带宽利用率并没有饱和，但由于缓存区空间不足，发往这些端口的报文无法进入交换机，从而产生大量丢包，严重影响转发性能，降低带宽利用率。

2.2.2 拥塞避免

拥塞避免即避免交换机缓存区发生队头阻塞。我们需要在缓存区被填满之前抢先一步主动丢包，防止缓存区被填满，从而保证最大的吞吐量和带宽利用率。

拥塞避免算法

常见的拥塞避免算法有尾丢弃（Tail Drop，简称 TD）、早期随机检测（Random Early Detection，简称 RED）以及加权早期随机检测（Weighted Random Early Detection，简称 WRED）三种。

» 尾丢弃（TD）

尾丢弃的做法是：为共享缓存区以及各端口的各发送队列分别设置丢包门限，当报文堆积达到门限时，丢弃新到达的报文，直至拥塞解除。

尾丢弃算法的优点是实现简单：它不需要选择丢弃的包，只是在拥塞时丢弃新到达的包。但是这种丢包策略会引发 TCP 全局同步现象：当出现拥塞时，所有新到达的 TCP 报文都被丢弃，这样所有经过这个设备的 TCP 连接同时进入慢启动状态，设备网络接口的吞吐量陡降。随着 TCP 拥塞窗口慢慢增加，网络吞吐量也慢慢增加，增加到一定程度时设备再次出现拥塞，所有 TCP 连接再次进入慢启动状态，如此反复。所有 TCP 连接以相同的步调执行“慢启动、丢包、再次慢启动”的过程，吞吐量随之剧烈波动，网络带宽不能得到充分的利用，这就是“TCP 全局同步”现象。

» 早期随机检测（RED)

RED 为共享缓存区以及各端口的各发送队列分别设置上限和下限，对到来的报文进行如下处理：

- 当报文存量小于下限时，不丢弃报文；
- 当报文存量大于上限时，丢弃所有新到达的报文；
- 当报文存量在上限和下限之间时，随机丢弃新到来的报文，缓存区（包括发送队列中）的报文存量越多，丢弃的概率越大，但有一个最大的丢弃概率。

RED 通过随机丢弃报文，避免了拥塞时所有的 TCP 连接同时进入慢启动状态，从而可以有效防止 TCP 全局同步的发生。

» 加权早期随机检测（WRED)

在 RED 算法中，所有的报文不分优先级，被丢弃的概率是相等的。WRED 在 RED 的基础上对数据流中的 IP 报文进行优先级划分：它根据 IP 报文首部的 ToS 字段把报文分成优先级不同的几个类别，不同类别的报文被丢弃的概率不同，优先级越高，被丢弃的概率越低，从而保证优先级高的报文获得更高的带宽。

拥塞避免算法比较

从上文可知，尾丢弃算法实现简单，但若应用于数据转发量大的设备中容易引发 TCP 全局同步问题；RED 算法采用随机丢弃报文的方式，将丢包行为在时间轴上进行散列，从而避免了尾丢弃带来的 TCP 全局同步问题；WRED 算法则在 RED 算法基础上，引入优先级概念，按

照优先级高低来设置报文丢弃的概率，是 RED 算法的改进版本。目前使用较广的算法为尾丢弃算法和 WRED 算法。

尾丢弃算法一般应用于接入层交换机和中小型网络的汇聚层交换机中。此类交换机由于数据转发量相对较小，同一时间内经过设备的 TCP 报文也比较少，所以很少发生 TCP 全局同步现象。尾丢弃算法因实现简单，被广泛应用于此类交换机中。

WRED 算法一般应用于核心交换机和路由器中。此类设备位于网络的中心位置，数据转发量大，采用 WRED 算法可避免 TCP 全局同步现象，同时保证优先级高的报文获得更高的带宽。

TP-LINK 交换机的拥塞避免功能采用的是尾丢弃算法，由内部芯片自动处理，无需配置。

2.2.3 流量控制

流量控制是一种用于防止网络拥塞时丢包的机制。当交换机缓存区开始溢出时，系统会向源地址发送阻塞信号，让其暂停发送数据直至拥塞解除。

半双工模式下，交换机采用基于 CSMA/CD（Carrier Sense Multiple Access with Collision Detection，带冲突检测的载波监听多路访问）机制的背压技术来实现流控：当缓存区被填满时，系统会向源地址发送一个假冲突信号，使之延迟一个随机时间再继续发包，从而为自己赢取更多的时间处理缓存区中的报文。

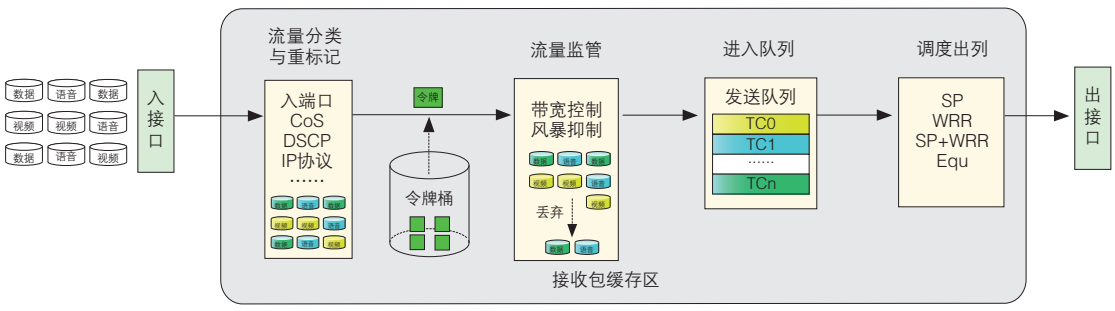
全双工模式下，因为没有 CSMA/CD 机制，所以不能采用背压技术来解决拥塞。为此，IEEE 802.3x 提供了一套解决方案：当交换机缓存区被填满时，系统会向源地址发送 Pause 帧。Pause 帧中有一个时间参数，表示收到 Pause 帧的一方要停多长时间；如果时间参数不为 0，则停止发送数据报文；如果时间参数等于 0，表示收到 Pause 帧的一方可以马上发送。

流量控制可以有效控制传送和接收节点间的数据流量，防止数据包丢失。TP-LINK 交换机可以在**端口管理**模块启用各端口的流控功能。

2.3 QoS 处理流程

如图 2-2 所示，QoS 的处理流程可分为流量分类与重标记、流量监管、进入队列和调度出列四个过程。

图 2-2 QoS 处理流程



1) 流量分类与重标记：报文进来后，交换机根据报文的入端口或其中的特定字段将报文分类，赋予不同的优先级。流量分类是进行差分服务的基础。对于满足 ACL 匹配规则的 IP 报文，交换机还可以对其进行 QoS 重标记，改变报文的优先级取值并传递给下游网络。

2) 流量监管：通过令牌桶机制约束各类数据流占用的带宽，确保各类业务没有滥用带宽资源。典型应用有带宽控制、风暴抑制等。

3) 进入队列：报文经过中间处理环节后被转发到相应的出端口，若出端口堵塞，则根据优先级取值进入不同的发送队列等待。

4) 调度出列：根据调度算法（SP、WRR、SP+WRR 等）将各发送队列中的报文转发出去，一般优先级高的报文会得到优先处理。

2.3.1 流量分类与重标记

流量分类

数据进入交换机缓存区后，交换机根据一定的规则将报文分类，赋予不同的优先级。流量分类就是流量的优先级设置，它是有区别地实施服务的前提。优先级与发送队列之间存在一种映射关系，分类完毕的数据流经过中间处理环节后按照映射表被送往相应的发送队列。

可以将流量分类分为简单流分类和精细流分类两种。

- 简单流分类是指采用简单的规则，例如根据数据流的入端口、数据帧的 802.1Q Tag 优先级取值或者 IP 报文的 DSCP（Differentiated Services Codepoint，差分服务编码点）值对数据流进行分类。
- 精细流分类是指采用复杂的匹配规则，例如根据报文的源 MAC 地址、目的 MAC 地址、VLAN ID、源 IP 地址、目的 IP 地址、采用的 IP 协议、源端口号、目的端口号等信息的组合对报文进行精细的分类。

流量重标记

流量重标记是指改变报文中的优先级字段。交换机可以选择接受上游网络对流量的优先级标记，也可以重新标记流量的优先级，并将标记传递给下游网络。

流量分类与重标记的关系

交换机可以对满足精细流分类匹配规则的 IP 报文进行 QoS 重标记，改变报文中的 DSCP 值。交换机按照新的 DSCP 值决定该报文的优先级，送往相应的 TC 队列。

TP-LINK 交换机中，精细流分类和流量重标记在**访问控制**（ACL）模块中实现，而简单流分类以及后续的流量监管、进入队列、调度出列在**服务质量**（QoS）模块中实现。**数据流进入交换机后，先经过 ACL 模块的处理，然后再进入 QoS 模块进行处理**。

说明：

TP-LINK 交换机的 ACL 模块除了可以对满足匹配规则的数据流进行重标记和带宽控制外，还有访问控制、对数据流进行端口镜像、重定向等丰富的功能。这里只简单介绍属于 QoS 范畴的精细流分类、QoS 重标记和带宽限制三部分内容，其他内容不做介绍。

简单流分类的优先级模式

流量分类的方法也叫优先级模式。对于简单流分类，TP-LINK 交换机支持三种优先级模式，分别是基于端口的优先级、基于 802.1P 优先级和基于 DSCP 优先级。

» 基于端口的优先级

指根据数据流的入端口来设置优先级。此模式下，需要先设置交换机各端口的优先级，交换机会根据入端口的优先级将数据流映射到发送队列。发送队列用 TC0~TCn 标记，编号越高，优先级也越高。

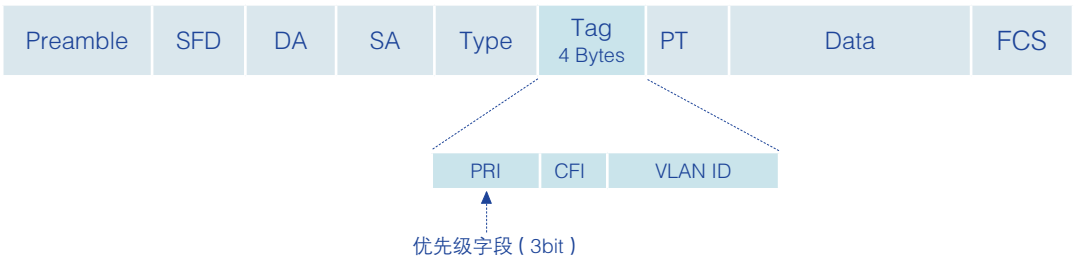
» 基于 802.1P 优先级

指根据数据帧的 802.1Q Tag 中的 PRI 域来设置优先级，适用于带 VLAN Tag 的数据帧。

802.1P 为 802.1Q VLAN 标准的扩充协议。图 2-3 所示为 802.1Q 的帧格式 ，可以看到它的 Tag 字段由 PRI、CFI 和 VLAN ID 组成。其实 802.1Q VLAN 标准中没有定义和使用 PRI 域，它是由 802.1P 补充定义的。

802.1P 将 802.1Q Tag 的最高的 3 个 bit 定义为用户优先级（User Priority），取值范围为 0~7，用于决定数据帧的优先级。

图 2-3　802.1Q 的帧格式



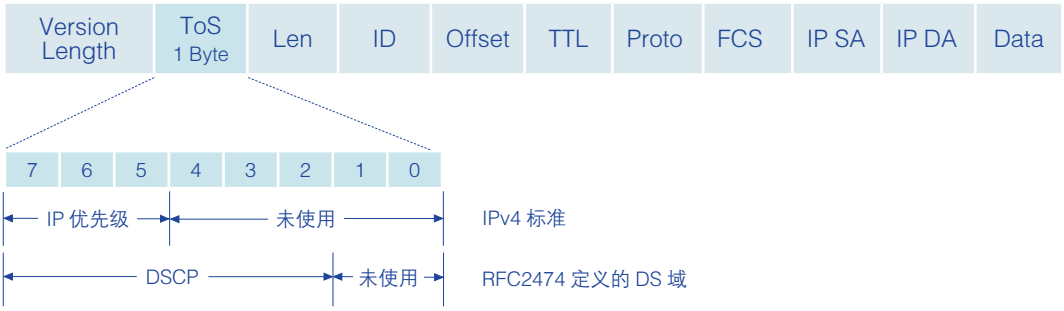
» 基于 DSCP 优先级

指根据 IP 报文头部 ToS（Type of Service，服务类型）字段的 DS 域来设置优先级，适用于 IP 报文。

如图 2-4 所示，IP 报文头部的 ToS 字段共有 8 bit，可以表征不同优先级特征的报文，最高的 3 个 bit 表示的是 IP 的优先级，取值范围是 0~7。RFC2474 重新定义了 IP 报文头部的

ToS 域，称之为 DS 域。DSCP 优先级用的就是该域的前 6 个 bit，取值范围为 0 ~ 63，后 2 个 bit 是保留位。

图 2-4　IP 报文

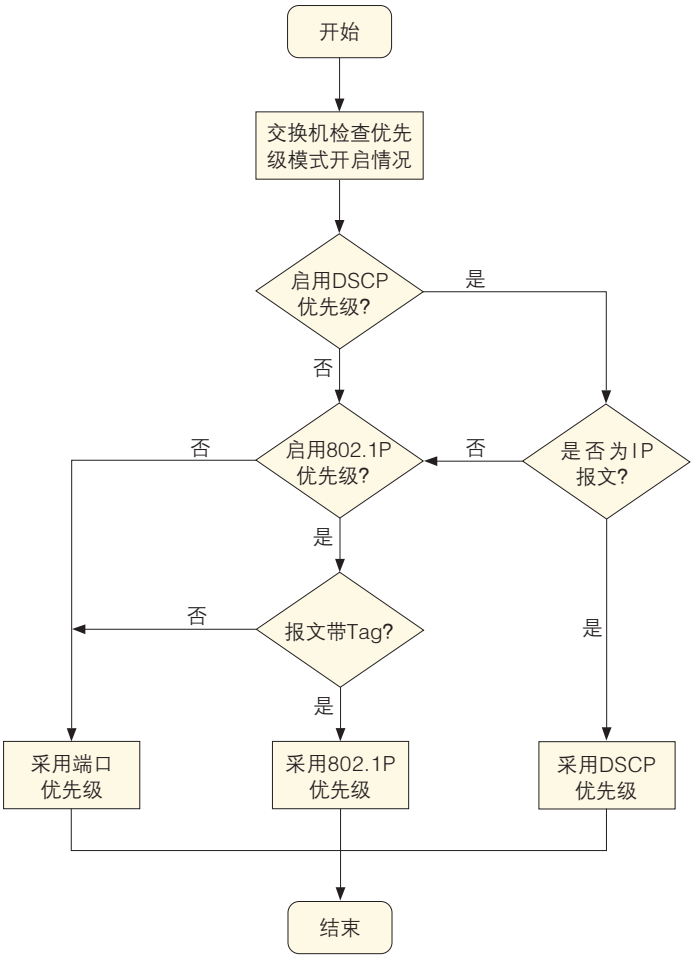


优先级模式匹配顺序

TP-LINK 交换机基于端口的优先级是始终开启的，其他优先级模式不同机型默认开启情况不同，一般需要手动开启。若同时开启了多种优先级模式，则按如下匹配顺序优先匹配前面的优先级模式：

DSCP 优先级 > 802.1P 优先级 > 端口优先级

图 2-5　优先级匹配顺序



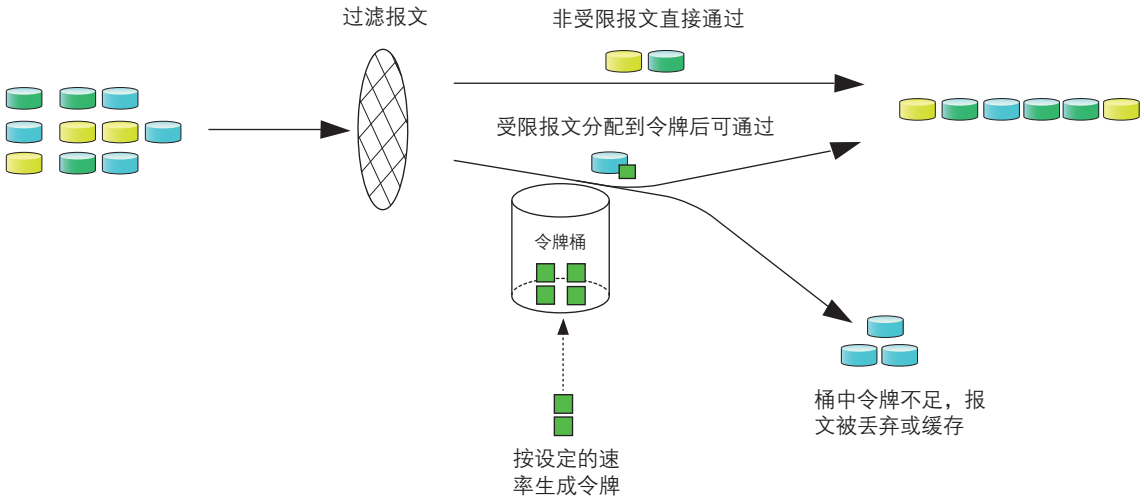
2.3.2 流量监管

流量监管是指约束各类数据流所占的带宽，即监测数据流中的报文，若该数据流所占带宽超过了允许的范围，则丢弃该数据流中的报文。交换机采用令牌桶机制来实现对流量的控制。

令牌桶机制

令牌桶机制的处理过程如图 2-6 所示。

图 2-6　令牌桶机制



令牌桶机制下的处理过程可总结为如下几点：

- **令牌桶只作用于匹配其过滤规则的流量。**

  每设置一条流量限制规则，系统将对应启动一个令牌桶作用于该规则。令牌桶对进来的流量进行过滤，非受限报文可直接通过，受限报文则需要等分配到令牌后才能通过。

- **令牌生成速度和令牌桶容量可设。**

  令牌生成速率即为用户对流量设置的限制速率。一个令牌可发送一个字节的数据，假设用户限制该流量传输速率不得大于 R，则每隔 1/R 秒有一个令牌被放入桶中。

  令牌桶容量即令牌桶中最多可存放的令牌数量，当桶中令牌的数量达到令牌桶容量时，令牌不再增加。令牌桶容量用于限制数据流的最大突发流量。假设令牌桶容量为 B，则最多允许 B 个字节的突发数据。令牌桶容量应大于报文的最大长度（极端情况是，最大的一个报文取走了桶中的全部令牌）。

- **报文过来之后，通过比较报文长度和桶中令牌数量来决定是否发送报文。**

  若报文长度 L 不超过桶中令牌数 T，则发送报文，同时桶中令牌减少为 T-L。

  若报文长度 L 超过桶中令牌数 T，则报文被丢弃或缓存，桶中令牌数不变。一般情况下，如果是入口方向的限速令牌桶，若令牌不够，报文会被直接丢弃；如果是出口方向

的限速令牌桶，则由交换机的数据包缓存管理单元（Memory Management Unit，简称MMU）决定是直接丢包还是将报文缓存。

流量监管的应用

流量监管的典型作用是限制网络中某一连接的流量占用的带宽，以及防止流量突发。其应用也围绕这两点展开。TP-LINK 交换机中，流量监管的应用包括**带宽控制**和**风暴抑制**两部分。

» 带宽控制

带宽控制是指对特定的数据流进行限速，约束其占用的带宽资源。

对于 TP-LINK 交换机，可以在**访问控制**（ACL）模块中配置 Policy，对满足 ACL 规则的数据流进行限速；也可以在**服务质量**（QoS）模块中分别设置各端口的入口带宽和出口带宽，进行基于端口的带宽分配。

» 风暴抑制

广播风暴现象：由于感染蠕虫病毒、遭受 ARP 攻击或交换机端口故障等原因，网络上的广播报文被不断转发，占用大量的网络带宽，导致正常的网络通讯受影响甚至瘫痪的现象。

风暴抑制功能是指通过限制端口中的广播报文的可用带宽来防止广播风暴。用户可以限制各端口允许接收的广播流量的大小，当该类流量超过用户设置的阈值后，系统将丢弃超出流量限制的广播帧，防止广播风暴的发生，从而保证网络的正常运行。

局域网内可能引起广播风暴的报文有三种类型：广播包、组播包和 UL 包。

- 广播包：目的 MAC 地址为 FF-FF-FF-FF-FF-FF 的数据帧。当交换机收到广播包时，会将其转发到除报文进来的端口之外的所有端口（此处不考虑在交换机上进行了 VLAN 划分的情况）。
- 组播包：目的 MAC 地址为组播 MAC 地址的数据帧，常见的组播 MAC 地址以 01-80-C2 和 01-00-5E 开头。当交换机收到组播包时，会将其转发给组播组内的所有成员。
- UL 包：目的 MAC 地址不在交换机 MAC 地址表中的数据帧。当交换机收到此类数据帧时，会将其转发到除报文进来的端口之外的所有端口。

TP-LINK 交换机可以设置进入端口的广播包、组播包和 UL 包的速率上限，有效防止广播风暴的发生。

2.3.3 进入队列

交换机读取报文的目的地址后，将报文转发到相应的出端口。当流量的速率超过出端口的带宽或超过为该流量设置的带宽时，报文就以队列的形式暂存在缓存中，这个队列就是发送队列。

经过不同优先级模式分类出来的数据，最终都集合在端口的发送队列中。优先级取值与发送队列（TC 队列）之间的映射存在如下两种模式：

- 直接映射：直接建立优先级取值与 TC 队列之间的映射关系。
- 间接映射：先将优先级取值映射为 CoS 取值，再建立 CoS 取值与 TC 队列之间的映射关系。其中 CoS（Class of Service，服务等级）是交换机用于建立优先级取值与 TC 队列之间映射关系的中间参数，取值范围为 0~7，数值上与 802.1P 优先级取值一一对应。

2.3.4 调度出列

调度出列是 QoS 流程的最后一个环节。当报文被送往端口的不同 TC 队列后，交换机将根据调度算法来发送 TC 队列中的报文，一般优先级高的报文会得到优先处理。

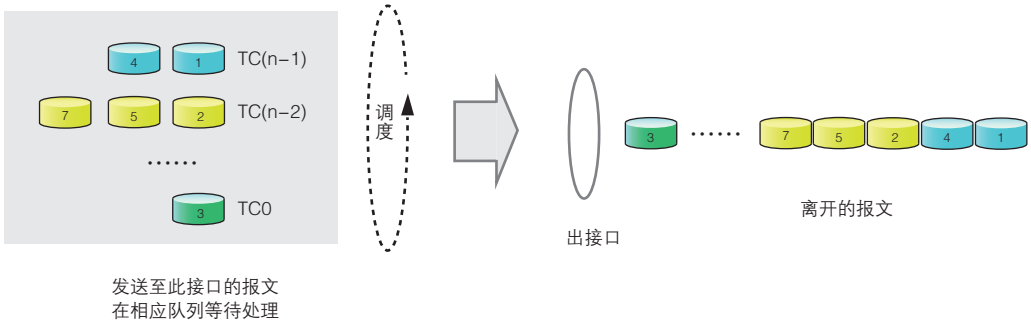
TP-LINK 交换机采用的调度算法有：严格优先级模式（SP）、加权轮询优先级模式（WRR）、SP+WRR 模式和等优先级模式（Equ）四种，下面将详细介绍。

严格优先级模式（SP）

严格优先级模式（Strict Priority Mode），简称 SP 模式。此模式下，交换机优先转发当前优先级最高的数据帧，等最高优先级数据帧全部转发完后，再转发次高优先级的数据帧。

如图 2-7 所示，交换机支持 n 个发送队列，TC0~TC(n-1)优先级依次升高。在 SP 模式下，交换机优先转发 TC(n-1)中的报文，待 TC(n-1)队列为空之后再转发 TC(n-2)队列中的报文，以此类推，TC0 中的报文最后发送。

图 2-7　SP 模式



SP 模式保证关键业务的报文拥有绝对的优先级，其缺点是，拥塞发生时，若较高优先级队列中长时间有报文需要发送，那么较低优先级中的报文就会由于得不到服务而“饿死”。

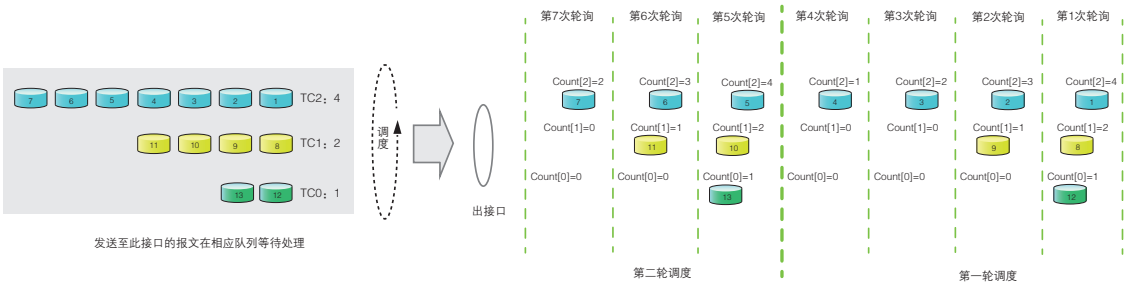
加权轮询优先级模式（WRR）

加权轮询优先级模式（Weighted Round Robin Mode），简称 WRR 模式。WRR 模式采用轮询的方式，在 TC 队列之间按照权重比值进行调度，以保证每个队列都得到一定的服务时间。默认情况下，TC0~TC(n-1)的权重比值为 1:2:4:...:$2^{(n-1)}$。

WRR 的实现方法：每个队列都有一个计数器 Count，根据权重进行初始化。系统每次轮询将检查所有队列，从计数器不为 0 的队列中取走一个报文，同时该队列计数器减一。对于空队列则直接跳过，以节省轮询时间。当所有队列的计数器都减为 0 时，所有队列的计数器重新根据权重初始化，系统开始新一轮的调度。

假设有 3 个 TC 队列，分别为 TC0、TC1 和 TC2，它们的权重比为默认的 1:2:4，其调度过程将如图 2–8 所示。

图 2–8　WRR 模式



从统计上看，WRR 模式中各队列被调度的次数与该队列的权值成正比，既实现了差分服务，也保证了权重低的业务可以得到服务时间，避免了 SP 模式中被“饿死”的情况。缺点是，它不提供绝对优先级队列，无法满足那些对延时比较敏感、需要有绝对优先级的业务（如语音业务）的需求。

SP+WRR 模式（SP+WRR）

SP+WRR 模式为前两种模式的混合模式。此模式下，交换机提供了两个调度组：SP 组和 WRR 组。SP 组内部队列遵循 SP 调度规则，WRR 组内部队列遵循 WRR 调度规则，而 SP 组与 WRR 组之间遵循 SP 调度规则。

此模式下，可将语音等关键业务放在 SP 组，其余业务放在 WRR 组。这样，一方面可以保证关键业务的带宽；另一方面，只要 SP 组所需带宽小于接口带宽，WRR 组内的队列也可按照权重分享到剩余带宽而不会被“饿死”。

等优先级模式（Equ）

等优先级模式（Equal Priority Mode），简称 Equ 模式。Equ 模式可以认为是 WRR 模式的一种特殊情况，所有队列的权重比是 1:1:...:1。

此模式下，所有队列所占带宽相同，无法体现优先级，所以也叫无优先级模式。

3 参考文献

本文在编写过程中，主要参考如下 RFC 文件：

- RFC791：Internet Protocol
- RFC1349：Type of Service in the Internet Protocol Suite
- RFC1633：Integrated Services in the Internet Architecture: an Overview
- RFC2474：Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers
- RFC2697：A Single Rate Three Color Marker
- RFC2698：A Two Rate Three Color Marker
- RFC3644：Policy Quality of Service (QoS) Information Model

声明