基于 FAST 架构的 TSN 介绍文档

主题	基于 FAST 架构的 TSN 介绍文档
文档号	
创建时间	2019-01-13
最后修改	2019-01-17
版本号	1.0
文件名	基于 FAST 架构的 TSN 介绍文档.pdf
文件格式	Portable Document Format

目录

一、FAST 2.0 流水线扩展模型的不足

由于 FAST 2.0 的流水线扩展模型难以满足确定性交换的要求，在保持 FAST 基本流水线架构不变的前提下，提出了 FAST 3.0 流水线扩展模型。

FAST 2.0 的流水线扩展模型如图 1 所示。其优点是在保持 FAST 标准五级流水线（GPP-GKE-GME-GAC-GOE）的基础上，支持用户定义解析（UDP）、用户定义关键字提取（UKE）、用户定义动作（UDA）和用户定义输出（UDO）等模块的插入，易于功能的扩展。但 FAST 2.0 在支持 TSN 方面主要存在两点不足。

第一个不足是在 GOE 和 UDO 之间存在用户逻辑难以控制的 FPGA OS 提供的分组缓冲区，在极端情况下，当一个输出接口发成拥塞后，可能会阻塞其他端口分组的发送，因此高优先级的 TSN 帧可能在 FPGA OS 中被阻塞，难以控制分组的延时；

第二个不足是 UDP 只能插入到 GPP 之后，只能在 GPP 支持的 IPv4、IPv6 和 ARP 三个解析树的基础上对分组的 L4-L7 协议进行进一步解析，而难以支持直接封装在以太网中的 PTP 协议（IEEE 1588）帧的解析。

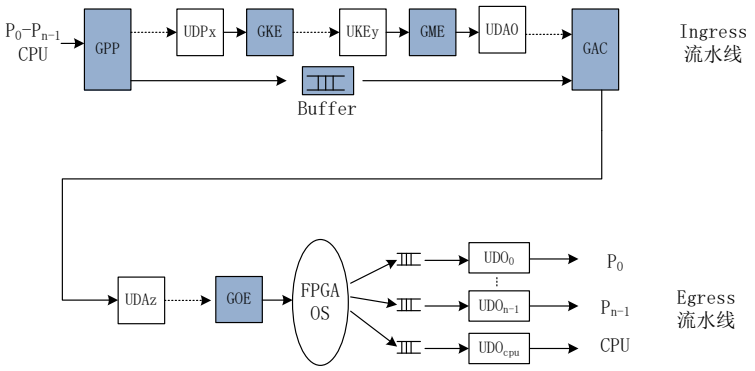

图 1　FAST 2.0 的流水线扩展模型

二、FAST 3.0 流水线扩展模型

FAST3.0 流水线扩展模型如图 2 所示。在两个方面对 FAST 2.0 扩展模型进行了改进。一是增加了 Pre-Ingress 流水线段，支持对 GPP 不支持的协议帧进行解析和处理，避免 GPP 将 1588 等未知协议帧定向到软件处理或丢弃；二是将 GOE 直接与 UDO 连接，避免了分组输出延时的不确定性，可

以有效支持在 UDO 中实现各种 QoS 保证功能，为基于 FAST 架构的 TSN 交换实现奠定了基础。

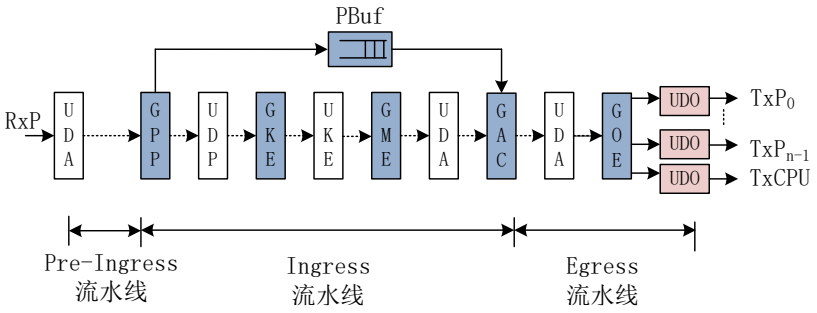

图 2　　FAST 3.0 的流水线扩展模型

FAST 3.0 流水线扩展模型不修改模块的接口规范，因此兼容基于 FAST2.0 的所有设计。

三、TSN 交换处理流程

在 802.1Q-2014 定义的以太网交换基本模型基础上，针对 TSN 的特定需求，802.1Qci 和 802.1Qbv 修订对交换模型中分组输出缓存的入队列操作和出队列调度机制进行了扩展，通过使用门控时间列表等机制对时间敏感分组入队和出队操作进行了限制。

FAST 3.0 的流水线可以在保持现有模块不变前提下，通过按需扩展插入新的模块支持用户定制的功能，因此可以方便地将 TSN 交换处理流程映射到 FAST 流水线上实现。

1. 标准以太网交换流程

802.1Q-2014 定义了标准以太网的交换流程，如下图所示。处理流程主要包含 10 个模块，每个模块的功能见下表。

序号	模块号	功能
1	拓扑管理	处理生成树协议帧，源 MAC 地址学习。
2	输入过滤	判断输入接口是否在到达帧 VID 对应的接口集合中，若不在，则丢弃分组。

3	转发交换	根据目的 MAC 和 VID 等确定输出接口集合。
4	输出过滤	判断输出接口集合是否在 VID 对应的接口集合中，去除不在 VID 端口集合中的输出接口。
5	流量测量	根据目的 MAC，VID，源 MAC 等信息对流量进行测量（metering）。
6	队列选择	根据优先级选择缓存分组的输出队列。
7	输出队列	定义了 8 个输出队列，分别存放不同优先级分组，例如 best effort 分组存放在 0 号队列，视频和音频分组分别存放在 3 号和 4 号队列，关键业务分组存放在最高优先级的 7 号队列。
8	队列管理	根据队列的状态和分组的属性决定是否丢弃分组。
9	输出调度	选择输出分组，支持优先级调度，加权调度等算法。
10	发送控制	将调度的分组发往以太网的 MAC 层输出

802.1Q 规范没有明确定义流量测量的粒度，无法对进入网络的流量进行细粒度的测量和管控。虽然支持多种输出调度算法，但更多是保证输出调度的优先级，或者按照预先确定的权值分配不同优先级队列占用的输出带宽，在调度中没有利用全局时间信息，无法实现确定性的延时控制。软件定义网络技术的应用可以简化交换流程，将生成树管理以及地址学习功能上载到控制器上实现，可以针对每条细粒度的流定义交换行为，但也难以实现确定性的延时控制。

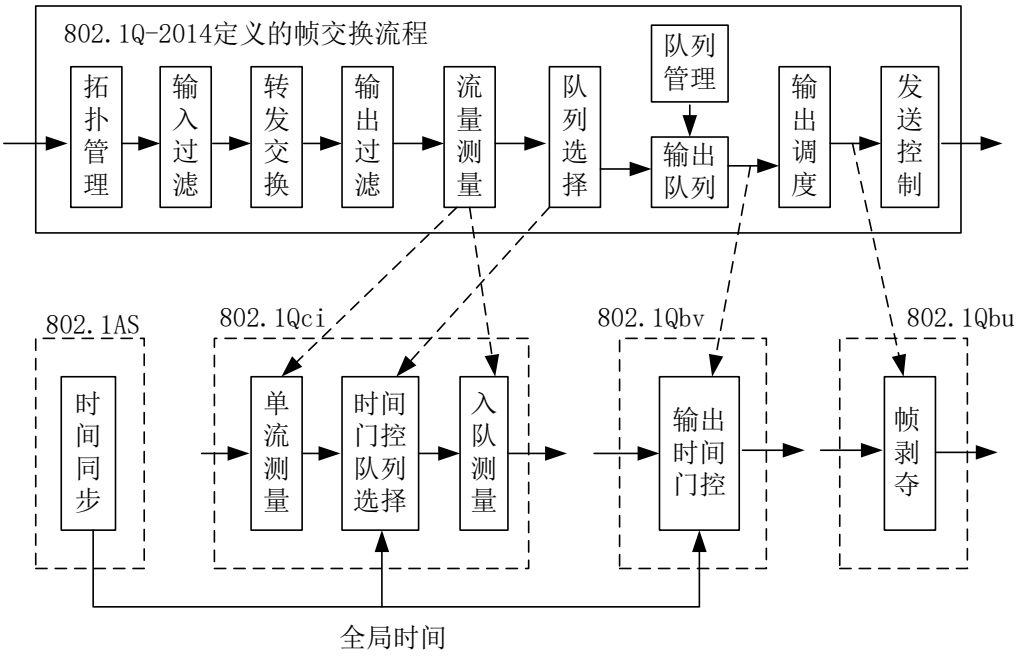

图 3　标准的帧交换流程及其 TSN 扩展

2. TSN 对以太网交换流程的扩充

针对确定性交换的目标，TSN 主要在时间同步（802.1AS）、单流的过滤和管控（802.1Qci，Per-Stream Filtering and Policing），时间敏感流量的调度（802.1Qbv Enhancement for scheduled traffic）以及帧剥夺（802.1Qbu）四个方面对标准以太网交换流程进行增强，除了时间同步标准外，其他三个标准都成为 802.1Q 的修订，并合并到最新的 802.1Q-2018 中。

时间同步机制采用 IEEE 1588 的 PTP 协议，为分组进入队列和输出调度的时间门控逻辑提供精确的全局同步时间。

TSN 在转发流程中扩充的单流过滤和管控（PSFP）机制主要实现三个功能，一是单流测量，使用令牌桶机制测量到达的每条流得流量和最大帧长度是否超过预定合约；二是时间门控队列选择机制，即将全局时间（分组到达的时刻）加入队列选择算法中考虑，重新计算分组内部优先级，并根据内部优先级而不是分组 VLAN 头或 IP 头中携带的外部优先级选择输出队列号；三是入队测量，基于令牌桶机制对进入特定队列的流量进行测量，保证进入相应队列缓存的分组流量满足一定的合约。

输出时间门控机制将全局时间用于输出调度，对于保存时间敏感帧的特定队列，是有在制定时刻才会打开。输出门控机制实际上是为每个输出队列设置了一个开关，只有开关打开时，队列调度请求才会发送到输出调度模块，该队列中的调度请求才能被响应。

帧剥夺机制主要是避免低优先级的长帧在发送时占用输出接口，影响高优先级帧的发送。例如在某个时刻，高优先级队列门的状态由关闭变成打开，因此输出调度逻辑可调度该队列中的高优先级帧发送。若在高优先级队列门打开前，一个低优先级的帧刚刚被调度，则该帧的发送可以立刻终止，在高优先级帧发送完成后，低优先级的帧可以继续发送。为了使以太网的 MAC 层支持帧剥夺机制（支持一个帧分多次发送，MAC 层负责这些分片的重新组合），802.3 工作组也推出了相应的规范（802.3br）。

四、FAST-TSN 实现模型

FAST 基本流水线包含协议解析（GPP），关键字提取（GKE），匹配查表（GME），通用动作（GAC）和通用输出控制（GOE）五个基本的模块，可为 TSN 交换提供基本的分组处理功能。而时间同步，以及流的测量整型、时间门控和输出调度逻辑分别由用户定义的 PTP UDA、CFQ UDO 和 PTPUDO 模块实现，如下图所示。

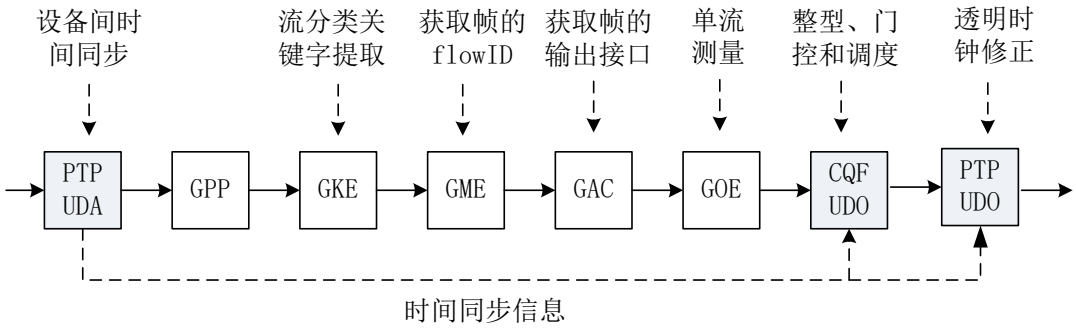

图 4　基于 FAST-TSN 交换实现模型

FAST-TSN 模型的特点是：

在硬件流水线中插入 PTP 协议处理模块，完全由硬件实现 PTP 同步帧（sync/delay-req/delay-resp 帧）处理，不需要软件参与，因此支持频率更高的时间同步操作，可获取优于 100ns 的同步精度。

将 TSN 的 PSFP 机制中的流分类和单流测量映射到 FAST 基本流水线中实现，通过 GME 实现基于五元组的流分类功能，为每个分组分配一个 flowID 并填写到分组的元数据中，后续的 GAC、GOE 和 UDO 模块可以利用 flowID 进行相关的操作。

采用独立的 UDO 模块实现核心的 TSN 门控和调度机制，通过 UDO 模块的重构可以支持多种 TSN 实现模型，满足不同 TSN 交换场景的需求。我们实现的 CQF-UDO 模型可以保证确定性的端到端交换延时。

五、TSN 循环队列转发（CQF）原理

1. CQF 工作原理

为了支持确定性的交换，TSN 对 802.1Q-2014 标准进行了扩充。其中单流过滤和管控机制（PSFP）中的时间门控逻辑控制了时间敏感分组进入缓存队列的时间，而时间敏感流增强调度（EST）机制中的输出门控机制控制

了分组离开输出队列的时间。基于对 PSFP 和 EST 机制的不同配置，TSN 交换机可以实现多样的确定性转发，满足不同场景的需求。

CQF 是 802.1Qch 定义的一种对 PSPF 和 EST 机制的配置，可以通过简单的计算实现确定性的转发延时。CQF 也是目前 TSN 规范中确定的唯一配置方式。

尽管对 PSFP 和 ETS 功能进行不同的配置可以实现不同的 TSN 控制，但 CQF 是目前 TSN 规范中给出的唯一一个实现模型，其最大特点是计算和配置简单，可以保证分组端到端交换的确定性延时。

（1）延时保证

CQF 模型将全网时间划分为长度为 d 的连续时间槽，用 i，i+1…，i+N 表示，若交换机 S0 在时间槽 i 中的 t1 时刻从链路上接收到数据帧 p，则必须在 i+1 时间槽中的某个时刻 t2 输出到链路上，如下图所示。

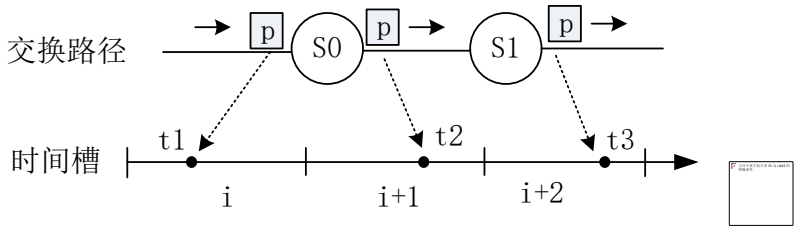

图 5　CQF 对交换机转发延时的要求

假设 t1 和 t2 可在时间槽 i 和 i+1 中任意分布，因此帧 p 经 S0 交换的延时 t2-t1 上限为 2d，下限为 0。同理，交换机 S1 必须在时间槽 i+2 中的某个点 t3 完成交换并输出到链路上，因此 p 经 S0 和 S1 交换机的延时 t3-t1 最大为 3d，最小为 d。更为一般的，基于 CQF 模型，帧 p 在网络中交换的最大延时为（h+1）*d，最小延时为（h-1）*d，其中 h 为传输路径跳数。

（2）时间敏感帧的处理

支持 CQF 模型的交换机只要在输出端口为时间敏感帧设置两个由时间门控制的队列 Q0 和 Q1。偶数时间槽，队列 Q0 保存输入端口接收的帧（接收模式，不发送帧），同时队列 Q1 发送在上一个奇数时间槽缓存的数据帧（发送模式，不接收帧）；奇数时间槽，两个队列的操作正好相反。因此，两个队列循环的进行分组缓存和调度输出操作，这也是 CQF 名称的来由。

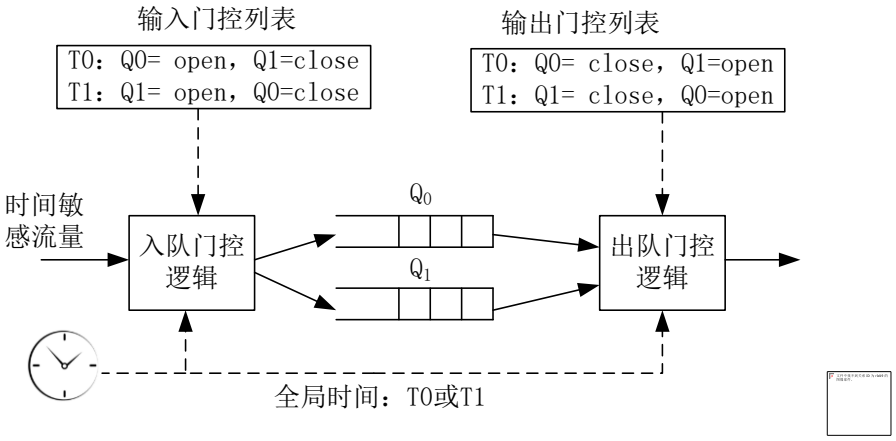

图 6　CQF 定义的循环队列结构及工作原理

针对 CQF 转发模型，PSFP 和 EST 机制定义的输入门控表和输出门控表示如图所示。在偶数时间槽，按照 T0 表项定义的动作执行，在奇数时间槽，按照 T1 表项定义的动作执行，具体操作如下表所示。

	偶数时间槽（T0）	奇数时间槽（T1）
入队门控逻辑	所有到达时间敏感分组进入 Q0 队列，关闭 Q1 队列入口。	所有到达时间敏感分组进入 Q1 队列，关闭 Q0 队列入口。
出队门控逻辑	将 Q1 输出请求发送给输出调度逻辑，屏蔽 Q0 的输出调度请求。	将 Q0 输出请求发送给输出调度逻辑，屏蔽 Q1 的输出调度请求。

显然，根据上面操作，每个时间敏感分组在交换中的延时不超过 2 个时间槽。当然设备间时间同步精度，非时间敏感帧传输占用输出链路对时间敏感帧的干扰，链路上分配的时间敏感业务量大小等因素都会对 CQF 模型中时间槽大小，Q0/Q1 队列长度等参数的选择有影响。

2. 支持 CQF 的交换机输出接口模型

（1）队列模型和入队出队控制

交换机每个输出接口除了时间敏感流量外，还有其他非时间敏感流量，如 best effort 流量，带宽预约流量等。为此，802.1Q-2014 的 Annex I（Priority and drop precedence）定义了 8 个优先级队列，分别缓存不同类型和优先级的流量，其中 Q7 的优先级最高，其次是 Q6，Q5…，优先级

最低的是 Q1。802.1Q-2014 规范中解释了 Q0 优先级高于 Q1 的原因。主要是网卡默认发出的 best effort 流量采用默认优先级 0，对应 Q0，而 Q1 用于存储优先级最低的背景流量，因此 Q1 的调度优先级低于 Q0。

为支持 CQF 模型，可将其中的两个最高优先级队列 Q7 和 Q6 设置缓存时间敏感流量。此时的交换机输出接口模型如下图所示。

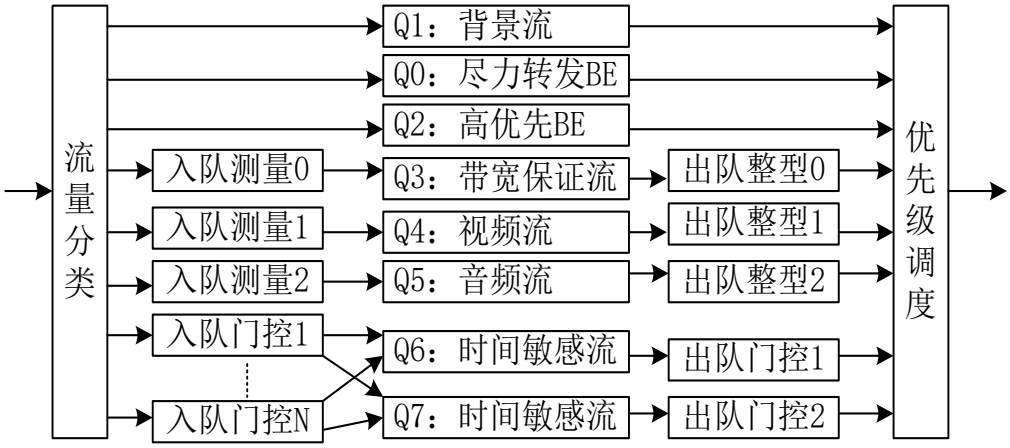

图 7　支持 CQF 的交换机输出接口模型

由上图可以看出以下几点。

一是优先级最高的 Q7 和第二高的 Q6 队列用于存储时间敏感流，而且只有这两个队列需要入队和出队的时间门控机制。由于不同的时间敏感流数据可能具有不同的发送周期（例如第一个流的周期时 125us，第二个流的周期是 250us），因此入队控制需要不同的门控逻辑。

二是 Q5，Q4 和 Q3 保存预约带宽的非时间敏感流量，其中 Q5 和 Q4 分别保存延时受限的音频和视频流，因此调度优先级比 Q3 要高。对于这些流量，在入队控制时需要增加流量测量逻辑，避免由于来自多个输入端口的多个单流汇聚后的流量超过输出接口预约的流量，同时在出队需要增加整型逻辑，减小流量的突发。

三是进出三个低优先级队列 Q2，Q0 和 Q1 的流量没有任何控制。当然，在队列将满时，队列管理逻辑会根据一定的算法选择分组丢弃。由于优先级低，这几个队列的流量也不会影响时间敏感流量和预约带宽的流量。

四是输出调度可采用绝对优先级调度。由于对高优先级队列采用了输出时间门控和输出整型机制，因此不会因为异常到达的高优先级流量“饿死”低优先级的流量。

（2）接口的配置管理

CQF 交换的输出接口是可管理的，即用户可以对优先级分类、入队门控，出队门控、入队测量和输出整型逻辑进行配置管理。

涉及的主要数据结构包括入队/出队门控列表，流量测量和整型的令牌桶参数，队列管理参数等。对 CQF 输出接口的配置管理抽象的介绍可以前往微信公众号进一步了解。

修改记录

版本号	修改人	日期	备注
1.0	肖智鹏	2019-01-13	初始版本
1.1	熊彩莉	2019-01-17	