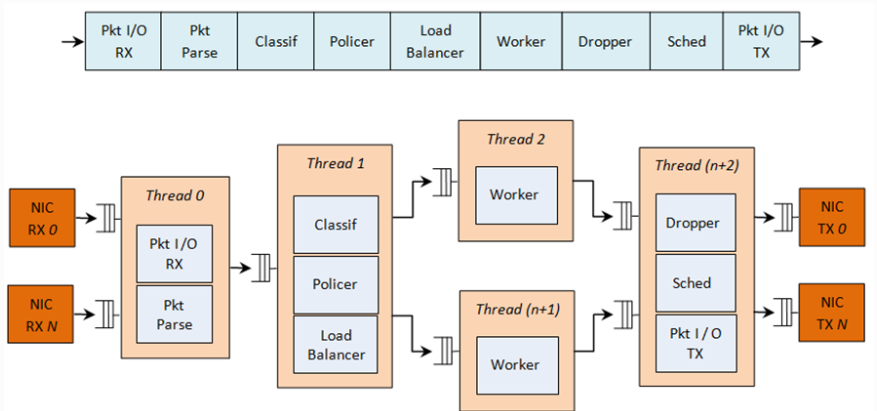
hqos设计说明文档

1. 实现原理及流程

1.1 hqos处理流程

hqos报文处理流水线的实例如下图所示。



Packet I/O RX & TX 多个NIC端口的报文接收/传输。用于Intel 1GbE/10GbE NIC的轮询模式驱动程序（PMD）。

Packet parser 识别输入数据包的协议栈。检查数据包头部的完整性。

Flow classification 将输入数据包映射到已知流量上。 使用可配置散列函数（jhash，CRC等）和桶逻辑来处理冲突的精确匹配表查找。

Policer 使用srTCM（RFC 2697）或trTCM（RFC2698）算法进行数据包测量。

Load Balancer 将输入数据包分发给应用程序worker。为每个worker提供统一的负载。 保持流量对worker的亲和力和每个流程中的数据包顺序。

Worker threads 客户指定的应用工作负载的占位符（例如IP堆栈等）。

Dropper 拥塞管理使用随机早期检测（RED）算法（Sally Floyd-Van Jacobson的论文） 或加权RED（WRED）。根据当前调度程序队列的负载级别和报文优先级丢弃报文。 当遇到拥塞时，首先丢弃优先级较低的数据包。

Hierarchical Scheduler 具有数千（通常为64K）叶节点（队列）的5级分层调度器 （级别为：输出端口，子端口，管道，流量类和队列）。 实现流量整形（用于子站和管道级），严格优先级（对于流量级别） 和加权循环（WRR）（用于每个管道流量类中的队列）。

1.2 hqos调度层次

hqos分为5层调度，依次为port、subport、pipe、traffic class（tc）、queue，下面一次介绍整体调度结构层次。

Port：以太网端口，多个端口轮询方式调度，所有端口优先级相同。

Subport：流量整型使用令牌桶算法（每个subport一个令牌桶），对每个tc强制执行带宽上限。

Pipe：流量整型使用令牌桶算法（每个pipe一个令牌桶），每个tc执行上限。

Tc：dpdk官方文档描述了4个tc，目前实际有13个tc，tc0-tc11前12个tc执行严格优先级（SP），tc12执行BE并且按照wrr算法进行预订权重分配。

Queue：与tc对应，目前tc0-tc11严格优先级队列每个tc对应一个queue，tc12为BE，有四个队列queue0-queue3，按照加权循环进行服务。

2、需求分析及方案设计

需求分析：确定性服务需求，业务划分为三大类：DE（确定性业务）、LS（时延敏感业务）、BE（尽力而为业务）。

DE服务优先级最高，分配到高优先级队列，LS优先级其次，分配到中优先级队列，BE优先级最低，刚好对应低优先级BE队列。

方案设计：

DE业务，占用tc0-tc7，8个tc，每个tc一个queue，支持8种不同DE业务（1个tc支持1个或多个dscp配置），优先级从tc0-tc7依次降低。

LS业务，占用tc7-tc11，4个tc，每个tc一个queue，支持4种不同LS业务（（1个tc支持1个或多个dscp配置），优先级从tc8-tc11依次降低。

BE业务，默认使用tc12，4个queue，支持wrr调度，默认走queue 0队列。

3、hqos配置下发：

参数：接口、user\_id、profile\_id、优先级、dscp值、带宽大小（MB）

接口：下发配置的接口，与frr中的接口对应，内部做了映射，映射到vpp中的实际物理网口。

user\_id：租户id，通过此id区分不同租户，租户id的取值需要控制器与实际数据包对应起来，采用内层dscp值，取值范围1- 62。

profile\_id：配置文件索引，从1开始递增，取值范围1- 64。

优先级：服务对应的优先级大小，优先级从1-12，值越小优先级越大，因此优先级1为最高优先级。

dscp值：区分不同类型的服务，报文中携带的内容。

带宽大小：配置不同服务的平均占用带宽大小，单位为，Mbps

配置下发注意事项：

（1）hqos配置目前分为两个步骤

步骤一：生成hqos配置文件profile，通过不同的profile\_id区分不同的配置文件。

步骤二：hqos配置文件profile应用，通过接口、profile\_id和user\_id来apply对应的配置。

提示:目前配置文件最大支持64（代码中自定义，可修改，与pipe数量一致）

（2）hqos配置文件为一组应用带宽分配，即每次下发配置需要将所有带宽规划好，如下，假设目前接口总带宽为1G

第一种配置：

配置下发：

网口：geb\_0\_0

profile\_id:1

hqos group配置：

dscp 10、priority 1、bandwidth 200M

dscp 20、priority 2、bandwidth 300M

dscp 30、priority 3、bandwidth 400M

其余流量 be队列

配置生效：

网口：geb\_0\_0

profile\_id:1

user\_id：1

第二种配置：

配置下发：

网口：geb\_0\_0

profile\_id:2

hqos group配置：

dscp 10、priority 3、bandwidth 200M

dscp 20、priority 2、bandwidth 300M

dscp 30、priority 1、bandwidth 400M

其余流量 be队列

配置生效：

user\_id：2

网口：geb\_0\_0

profile\_id:2

（3）hqos配置回退，可以直接执行第（2）部分的步骤二，通过profile\_id、接口和user\_id来直接生效。

（4）hqos配置修改，目前不支持直接修改hqos配置，因为目前dpdk不支持直接修改profile文件，也不对外提供任何修改profile文件的接口，所以可以通过新增、回退操作来设置想要的hqos配置。

（5）不同设备的网口初始hqos配置带宽需要配置，需要确保网口带宽大于目前初始带宽，这样资源才能全部利用，目前配置值为10G（可修改）。

（6）hqos配置很复杂，所以配置没有生效时先关注系统日志，是否出现了问题，根据对应的提示修改配置。

Srv6组网中hqos配置：

内层dscp映射为用户id、外层dscp映射为业务类型

1. 首节点为ipv4数据包，没有封装srv6层，因此内层配置和外层配置为：

Pipe：

set dpdk interface hqos pktfield GigabitEthernetb/0/0 id pipe offset 64 mask 00fc000000000000

业务：

set dpdk interface hqos pktfield GigabitEthernetb/0/0 id tc offset 8 mask 00000000000000fc

（2）中间节点，此部分数据为srv6隧道数据，因此为ipv6协议并且携带srv6封装层，因此配置为：

Pipe：

set dpdk interface hqos pktfield GigabitEthernetb/0/0 id pipe offset 192 mask 00fc000000000000

业务：

set dpdk interface hqos pktfield GigabitEthernetb/0/0 id tc offset 8 mask 00000000000000fc