



# 基于 OpenFlow 协议的 MSE 设备实现方法研究

朱永庆,陈华南,唐 宏

(中国电信股份有限公司广东研究院 广州 510630)

**摘 要:**基于 OpenFlow 协议与 MSE 设备的技术原理研究,重点分析 OpenFlow 与 MSE 实现的关键技术,在对二者进行分析比较的基础上,提出了 MSE 设备基于 OpenFlow 协议向 SDN 架构演进的可行思路。

**关键词:**软件定义网络;多业务边缘;OpenFlow;流表

**doi:** 10.3969/j.issn.1000-0801.2013.12.022

## Research on Realization of MSE Equipment Based on OpenFlow Protocol

Zhu Yongqing, Chen Huanan, Tang Hong

(Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

**Abstract:** The theory of OpenFlow and MSE equipment were deeply and widely discussed. On the basis of the research, the MSE implementation with the extension of the OpenFlow protocol was proposed. It is a feasible solution for the IP network evolving to the SDN architecture.

**Key words:** software defined network, multi-service edge, OpenFlow, flow table

### 1 引言

软件定义网络(software defined network, SDN)是控制平面与实体数据转发平面分离的一种新型网络架构。SDN 架构包括应用层、控制层与基础设施层,其中,控制层具有网络全局视图,在对基础设施层的数据转发行为实施控制的同时,还通过 API 与应用层交互。OpenFlow 是 SDN 架构中目前最成熟的接口协议,负责控制层控制器与基础设施层交换机设备之间的交互。

现阶段,IP 路由器是软硬件紧耦合、相对封闭的网元,控制平面与转发平面集中于同一物理实体,功能强大,但开放性不足。MSE(multi-service edge,多业务边缘)设备作为 IP 城域网的关键网元,不仅担负 IP 流量的汇聚、转发职能,而且负责用户与业务的部分管理功能,实现方式与通用的 IP 路由器存在较大不同。在 SDN 架构下如何实现

MSE 设备,在业界还处于初步探索阶段。本文拟就 SDN 架构下 MSE 的实现思路进行探讨。

### 2 OpenFlow 与 MSE 设备关键技术

#### 2.1 OpenFlow 协议处理机制

OpenFlow 协议的核心是流表(flow table),其生成、维护、下发由控制器完成。如图 1 所示,流表中匹配域(match field)与指令(instruction)域是关键字段。匹配域包含报文匹配的关键字段,如端口号、VLAN、二层/三层/四层信息等;指令域决定对报文采取何种动作,包括报文转发、丢弃以及设置队列等。

|             |          |         |             |         |        |
|-------------|----------|---------|-------------|---------|--------|
| match field | priority | counter | instruction | timeout | cookie |
|-------------|----------|---------|-------------|---------|--------|

图 1 OpenFlow 流表格式

多级流表概念从 OpenFlow 1.1.0 开始引入,在此流表架构下,如图 2 所示,报文处理是一个“匹配—动作—再匹

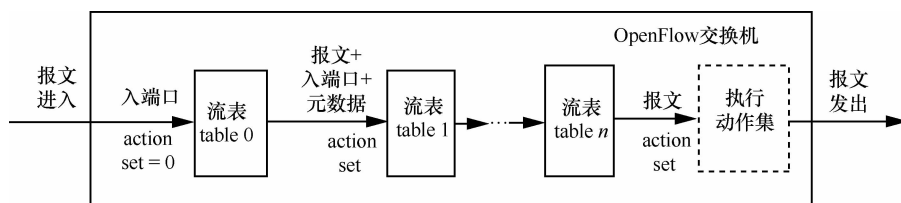


图2 OpenFlow 流表处理过程

配”的流水线过程:报文从入端口进入 SDN 交换机后,先查找第一级流表 table 0,如果报文的关键信息与 table 0 中匹配域匹配,则执行指令域中的动作,并通过动作中的跳转至表动作进入下一个流表;报文在最后一个流表执行完所有动作后被转发到出端口。为了保证这种多级流表/动作机制的有序性与有效性,OpenFlow 协议规定流表处理过程不可逆,即报文在某个流表处理完毕后不能再次返回该流表处理。

## 2.2 MSE 与 OpenFlow 关键特征比较

现阶段,主流 MSE 设备采用专用硬件架构实现,基于 NP(network processor)芯片进行报文处理,如图 3 所示。NP 内存存储的各种转发表(forwarding table)是流表策略与动作的集合。NP 接收到报文后,按照预定策略逐项对报文进行匹配转发。这种多级处理流程本质上也属流水线处理方式,与 OpenFlow 处理机制基本一致。

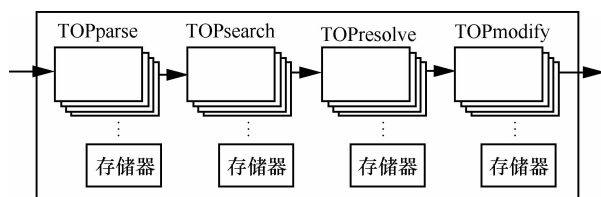


图3 NP 处理过程

需要指出的是,NP 的处理方式与芯片的具体实现强相关,每个厂商的 NP 芯片都基于自有芯片语言(类似于汇编语言)来实现报文处理流程,流表格式、大小与存储空间也各不相同。而 OpenFlow 协议将转发设备定义为交换机(不同于现有的交换机形态),采用软件编程方式实现高性能转发,需要统一的流表格式与处理机制。因此在报文处理方面,现有 MSE 设备与 OpenFlow 机制在流表查找、QoS、报文处理、信息交互以及转发机制等方面存在较大差异。

### (1) 流表查找

OpenFlow 协议建议使用 TCAM(三态内容寻址存储器)进行流表查找,以提升查找速度。MSE 设备承载业务复杂,表项众多,需存储众多表项的查找索引,若仅采用 TCAM,

将难以实现高效业务承载。通常,MSE 转发表项通过 NP 内部散列算法查找,只有当存储表项的关键字(key)超过特定长度时,才启用 TCAM 协助查找。例如,MSE 设备的路由转发表 FIB 通过 TCAM 方式查找,而宽带用户属性等相关表项则由 NP 自行查找。

### (2) QoS 机制

MSE 设备的 QoS 机制通过流量管理(traffic management, TM)芯片的队列与调度器实现,无需软件预定义,并支持动态层次化 QoS。OpenFlow 协议定义了软件 QoS 队列,流表通过设置队列动作将匹配的报文与队列相关联;OpenFlow 1.3.1 版本新增 meter table 以协助实现 QoS 限速或 DSCP-Mark 功能。尽管此种设计方式可在一定程度上实现 meter table 与软件 QoS 队列之间的联动,但二者弱关联,不具备 QoS 调度所特有的相互间动态控制机制,无法实现动态的层次化 QoS 调度。

### (3) 流水线处理机制

MSE 设备承载业务相对复杂,报文处理过程中会以关键字为基准多次查找同一张表,并将查找结果与报文关键信息进行多次比较(例如,在开启严格 uRPF 的情况下,报文转发需基于源 IP 地址查找转发表,将查找结果与报文接口信息相对比;若匹配,则依据目的 IP 地址再次查找转发表以确认转发路径)。OpenFlow 协议在报文处理过程中则严格按照多级流表顺序依次处理,若基于该协议实现 MSE 设备,则需改进其流表的不可逆机制,增加比较动作。

### (4) 信息交互

MSE 设备的控制平面与数据转发平面物理集中,主要与 AAA 服务器、网管等系统交互,实现认证、授权与信息统计等功能。OpenFlow 协议定义了 3 类消息实现控制器与交换机的交互:controller-to-switch 消息定义控制器从交换机获取诸如端口统计等信息的机制;asynchronous 消息定义交换机状态主动上报机制,实现对 SDN 转发设备的状态管理;symmetric 消息则定义交换机与控制器之间如何进行连通性交互。上述 3 类消息均缺乏对认证、授权与流量统计等信息的交互能力。



### (5) 转发机制

MSE 设备采用 CLOS 分布式交换架构,通过交换网实现接口卡间的数据交换,以提升设备转发性能。OpenFlow 协议定义的交换机是一个集中式处理网元,严格采用流水线式处理流程,若应用在现有的 MSE 设备转发架构上,必须进行相应的协议适配。

## 3 MSE 设备引入 OpenFlow 协议的实现

任何技术发展都存在“路径依赖”现象,MSE 设备向 SDN 架构的演进过程也不例外。OpenFlow 协议需要的是一种统一的与芯片语言无关的流表处理机制。现阶段,MSE 设备主要基于硬件(如 NP、ASIC 或 FPGA 等)实现报文处理与转发,短期内难以实现硬件统一,若以 OpenFlow 协议为基础向 SDN 架构演进,就必须考虑其硬件实现的基本特性。

### 3.1 MSE 设备作为 SDN 交换机的实现

在现有 MSE 设备基础上实现 OpenFlow 统一控制,如图 4 所示,需增加 OpenFlow 适配层以适配不同硬件,并对 OpenFlow 协议进行扩展以满足 MSE 设备复杂的业务需求。此种架构下,应用接口和 SDN 控制器由运营商统一定义,业务流程的抽象定义由 SDN 控制器完成,MSE 设备通过 OpenFlow 适配层解释 OpenFlow 协议的抽象流程定义,再依据不同的硬件实现对转发流表的控制;而增强型 OpenFlow 协议则是为满足 MSE 设备的业务承载需求在 OpenFlow 协议基础上进行的扩展。

MSE 设备的 OpenFlow 适配层基于硬件实现,该适配层的引入可屏蔽不同 MSE 设备的硬件差异,可在各设备商维持自身硬件实现的基础上,实现 MSE 设备控制平面

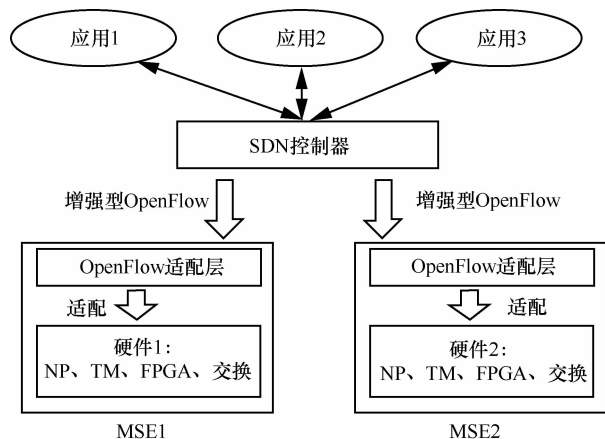


图 4 现有 MSE 设备基础上的 OpenFlow 统一控制架构

的统一。MSE 设备的 OpenFlow 适配层在对 OpenFlow 报文解析后,将 OpenFlow 的流表命令与 MSE 设备硬件进行指令适配,如图 5 所示。首先,将 OpenFlow 的抽象业务流表映射到 MSE 设备不同的硬件处理过程中,具体表现在业务 2 的流表 2、流表 3、流表 4 分别通过过程(3)、过程(4)、过程(5)映射至交换、TM 和 FPGA 单元,其中,过程(4)需 NP 和 TM 相互配合处理,映射至 NP 和 TM 中。其次,对 OpenFlow 的流表内容进行整合与分类下发。其主要原因在于 OpenFlow 协议的流表下发是针对单独业务的,而传统 MSE 设备的业务表项比较灵活,可以存放不同的流处理信息,不同业务也可以复用相同位置的表项(例如,MSE 设备的端口属性表可同时支持二层 VPN、三层 VPN 与 PPPoE 拨号业务的流量分流与流程分类,同一接口不同 VLAN 的二层 VPN 业务会下发一系列流表,该接口其他 VLAN 的三层 VPN 业务也会下发一系列流表),此种情况下,MSE 设备的 OpenFlow 适配层需要通过流表整合来适配 NP 硬件表项的存储,并根据 NP 的实现把不同业务的流表整合到 NP 的同一存储表保存。该过程具体表现在业务 2 和业务 1 的流表 1 被整合到了 NP 表 1。

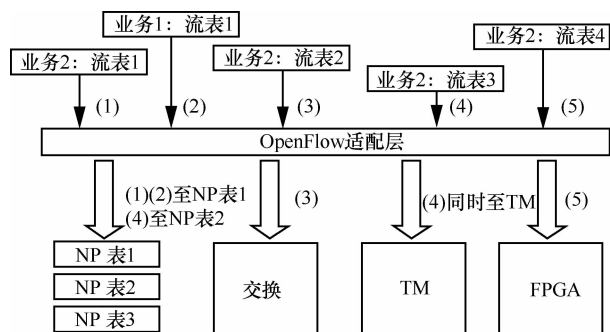


图 5 MSE 设备的 OpenFlow 适配过程

上述过程中,SDN 控制器使用 OpenFlow 协议下发控制信令,无需任何改动。但为满足 MSE 设备的业务承载需求,还需对 OpenFlow 协议进行相应扩展,如图 6 所示,具体表现如下。

- 匹配域增加多个自定义匹配字段,以支持对宽带接入的独特匹配(如用户 ID 与用户计费 ID 等)。
- 扩展流表不可逆的机制,支持对同一张表重复查找并允许查找结果与匹配域的内容相比较。
- 重新定义 SDN 转发设备的主动上报机制,允许 SDN 转发设备主动上报一些统计信息至应用层应用。

### 3.2 基于 OpenFlow 映射的 MSE 设备业务处理实现

MSE 设备的通用业务处理流程通常包括 uRPF 检验、

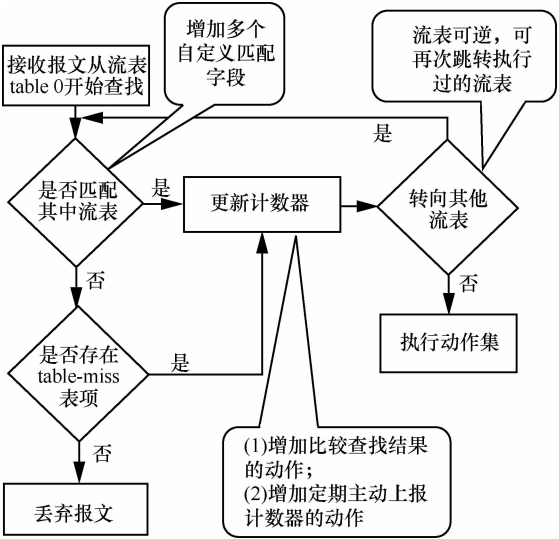


图 6 OpenFlow 协议扩展

ACL 匹配处理及用户业务 QoS 调度等环节,以下行流量为例的报文处理过程如下。

- (1)解析 IP 报文头,查找 MSE 设备的网络侧端口表,通过入端口(ingress port)与 VLAN 信息判断报文是否进入 IPv4/IPv6 报文转发流程。
- (2)端口属性表中若配置有 ACL 标志的信息,则构造五元组并通过 TCAM 查找进行 ACL 匹配处理。
- (3)根据源 IP 地址查找 FIB 表,进行 uRPF 安全检查。
- (4)根据目的 IP 地址查找 FIB 表,确定报文转发的出端口(egress port),报文通过交换网板送到用户侧板卡。

(5)用户侧板卡查找出端口端口表,校验 QoS 信息,通过五元组信息构造关键字,在 TACM 查找 QoS,并根据返回结果确定报文在 TM 的队列情况。

(6)查找 ARP 表,重新封装数据链路层信息后,报文送到 TM 的队列进行 QoS 处理。

(7)经 TM QoS 处理完毕,报文被转发至下一跳。

若基于 OpenFlow 协议实现此流程,MSE 设备 OpenFlow 适配层的转换处理对应关系见表 1。

实现 PPPoE 用户接入与承载是 MSE 设备最基本的功能。MSE 设备对 PPPoE 用户报文的处理与转发流程主要包括用户安全检查、ACL 匹配以及 QoS 处理等环节,为与 MSE 通用业务处理流程相对应,下面以 PPPoE 用户下行流量处理说明报文处理过程。

- (1)解析 IP 报文头,查找 MSE 设备的网络侧端口表,通过报文的入端口和 VLAN 确定 PPPoE 用户的报文转发流程。
- (2)根据用户报文的关键属性(如用户 IP 地址)查找用户表,核对用户信息,对用户进行安全检查。
- (3)用户表中如果配置有 ACL 标志的信息,则构造五元组并通过 TCAM 查找进行 ACL 匹配处理。
- (4)根据目的 IP 地址查找 FIB 表,确定转发出口为用户端口,报文通过交换网板送到用户侧板卡。
- (5)用户侧板卡根据目的 IP 地址查找 PPPoE 用户表,如果支持 QoS,则以五元组信息构造关键字,在 TACM 查找 QoS,并根据返回结果确定报文 TM 队列信息。

表 1 MSE 设备通用业务处理流程流表映射关系

| 流表匹配域      | 流表指令域                                      | MSE 设备的流表适配动作                                                                       |
|------------|--------------------------------------------|-------------------------------------------------------------------------------------|
| 匹配端口与 VLAN | 跳转至 FIB 表                                  | NP 查找网络侧端口表;<br>网络侧端口表标识支持 uRPF 和 ACL 处理                                            |
| 匹配源 IP 地址  | 比较匹配结果;跳转至 ACL 表                           | NP 根据源 IP 地址在 TCAM 中查找 FIB 表;<br>NP 实现匹配结果校验                                        |
| 匹配五元组      | 报文丢弃或跳转至 FIB 表;                            | NP 根据五元组在 TCAM 查找 ACL 并获取查找结果;<br>NP 实现匹配结果校验                                       |
| 匹配目的 IP 地址 | 确定出端口;<br>确定下一跳;<br>设置源 MAC 地址;<br>跳转至出端口表 | NP 根据目的 IP 地址在 TCAM 查找 FIB 表并获取查找结果;<br>NP 根据查找结果获取出端口和下一跳;<br>NP 通过交换网板将报文转发至用户侧板卡 |
| 匹配出端口      | 跳转至 QoS 表                                  | NP 查找报文出端口端口表;<br>端口表标识支持 QoS 处理;<br>NP 根据出端口端口表封装源 MAC 地址                          |
| 匹配五元组      | 设置队列;跳转至 ARP 表                             | NP 根据五元组在 TCAM 查找 QoS 表;<br>NP 根据查找结果设置队列                                           |
| 匹配下一跳      | 设置目的 MAC 地址;<br>执行所有动作;<br>转发至出端口队列        | NP 查找 ARP 表获取目的 MAC 地址;<br>报文被送至 TM 的队列,经处理后转发                                      |



表 2 PPPoE 用户业务处理流程流表映射关系

| 流表匹配域      | 流表指令域                                            | MSE 设备的流表适配动作                                                                          |
|------------|--------------------------------------------------|----------------------------------------------------------------------------------------|
| 匹配端口与 VLAN | 跳转至 PPPoE 业务处理                                   | NP 查找入端口端口属性表;<br>端口属性表标识该报文为 PPPoE 业务处理属性                                             |
| 匹配源 IP 地址  | 跳转至 ACL 表                                        | NP 查找用户表;<br>用户表标识需进行 ACL 处理                                                           |
| 匹配五元组      | 丢弃报文或跳转至 FIB 表                                   | NP 根据五元组在 TCAM 查找 ACL 并获取查找结果;<br>NP 实现匹配结果的校验                                         |
| 匹配目的 IP 地址 | 确定出端口;<br>确定下一跳;<br>设置源 MAC 地址;<br>跳转至 PPPoE 用户表 | NP 根据目的 IP 地址在 TCAM 查找 FIB 表;<br>NP 根据查找结果获取出端口和下一跳是 PPPoE 用户;<br>NP 通过交换网板将报文发送至用户侧板卡 |
| 匹配目的 IP 地址 | 设置队列;<br>跳转至 QoS 表                               | NP 查找 PPPoE 用户表;<br>用户表标识需进行 QoS 处理                                                    |
| 匹配五元组      | 设置队列;<br>跳转至下行端口表                                | NP 根据五元组在 TCAM 查找 QoS 表;<br>NP 根据查找结果设置队列                                              |
| 匹配出端口接口    | 设置源 MAC 地址;<br>跳转至 ARP 表                         | NP 查找用户侧端口表;<br>NP 根据用户侧端口表封装源 MAC 地址                                                  |
| 匹配下一跳      | 设置目的 MAC 地址;<br>执行所有动作;<br>转发到出端口队列              | NP 查找 ARP 表获取目的 MAC 地址;<br>报文被送至 TM 的队列,经处理后转发                                         |

(6)报文送到出端口端口表。查找 ARP 表,重新封装数据链路层信息后,报文送到 TM 的队列进行 QoS 处理。

(7)经 TM QoS 处理完毕,报文被转发至下一跳。

若 MSE 设备基于 OpenFlow 协议来完成 PPPoE 用户业务处理过程,OpenFlow 流表与 MSE 设备的流表映射关系见表 2。

需要特别指出的是,表 2 中的内容是一个抽象的过程对比,并非最终的业务流程,具体实现与 MSE 设备硬件实现密切相关。

## 4 结束语

MSE 设备是宽带用户接入管理与流量转发的重要节点,是 IP 城域网运营能力提升的关键。本文所探讨的基于 OpenFlow 协议的 MSE 设备实现方法,为现有设备向 SDN 架构演进提供了一种可行的方案,为下一代设备研发和网络演进提供了良好的思路。由于 OpenFlow 协议主要关注流量管理与转发,还难以满足 MSE 设备在用户与流量管理等方面的需求,后续将在信令响应实时性以及设备智能性提升方面,结合网络功能虚拟化(NFV)等技术的进展,进一步研究深化。

## 参考文献

1 McKeown N, Anderson T. OpenFlow: enabling innovation in

campus networks. <http://www.OpenFlow.org/documents/OpenFlow-wp-latest.pdf>

- OpenFlow switch specification(version 1.3.1).<https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/OpenFlow/OpenFlow-spec-v1.3.1.pdf>,2012
- OpenFlow switch specification (version 1.3.2).<https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/OpenFlow/OpenFlow-spec-v1.3.2.pdf>,2013
- Architecture and Transport Working Group. Migration to Ethernet-Based DSL Aggregation.
- 中国电信股份有限公司. 中国电信 IP 网络设备技术要求-MSE,2012

## [作者简介]



朱永庆,男,中国电信股份有限公司广东研究院高级工程师,主要研究方向为超宽带城域网技术研究。

陈华南,男,中国电信股份有限公司广东研究院工程师,主要研究方向为网络新技术与设备新技术研究。

唐宏,男,中国电信股份有限公司广东研究院高级工程师,主要研究方向为 IP 网络架构与新技术。

(收稿日期:2013-11-25)