



(12) 发明专利申请

(10) 申请公布号 CN 102843299 A

(43) 申请公布日 2012. 12. 26

(21) 申请号 201210335080. 7

(22) 申请日 2012. 09. 12

(71) 申请人 盛科网络(苏州)有限公司
地址 215021 江苏省苏州市苏州工业园区星
汉街 5 号 B 幢 4 楼 13/16 单元

(72) 发明人 赵怡 闫兴安

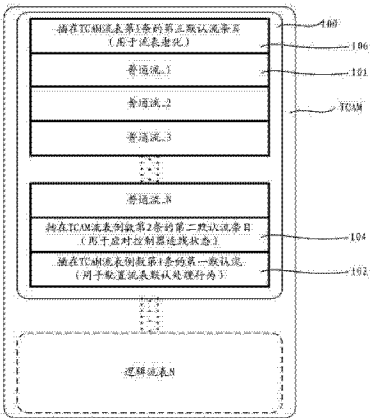
(74) 专利代理机构 苏州威世朋知识产权代理事
务所(普通合伙) 32235
代理人 杨金

(51) Int. Cl.
H04L 12/56(2006. 01)

权利要求书 1 页 说明书 6 页 附图 3 页

(54) 发明名称
基于 TCAM 实现 Openflow 多级流表的方法及
系统

(57) 摘要
本发明提供一种基于 TCAM 实现 Openflow 多
级流表的方法及系统,其方法包括:S1、将一张在
Openflow 交换机上维护的 TCAM 表划分为多个逻辑
流表;S2、为每一逻辑流表分别设定其所匹配的
域类型。本发明通过将 TCAM 表划分为多个逻辑
流表,且每一流表的匹配字段长度包括多种形式,
这样有效增加了 TCAM 表的存储条目数,从而满足
不同应用需求,进而可有效减小 Openflow 交换机的
制造成本。



1. 一种基于 TCAM 实现 Openflow 多级流表的方法,其特征在于,该方法包括:

S1、将一张在 Openflow 交换机上维护的 TCAM 表划分为多个逻辑流表;

S2、为每一逻辑流表分别设定其所匹配的域类型。

2. 根据权利要求 1 所述的方法,其特征在于,所述步骤 S1 还包括:在所述逻辑流表中的最后一条配置一个用于配置流表默认处理行为的第一默认流条目。

3. 根据权利要求 2 所述的方法,其特征在于,所述默认处理行为包括丢弃、或者重定向到下一张逻辑流表、或者重定向到另一张逻辑流表、或者发送到控制器。

4. 根据权利要求 3 所述的方法,其特征在于,所述步骤 S1 还包括:在所述流表中的倒数第二条配置一个用于应对控制器断线状态的第二默认流条目,其中,“应对控制器断线状态”的步骤具体为:

配置第二默认流条目的默认状态为“禁用”;

判断当前控制器是否为断线状态,若是,则使能所述第二默认流条目,并且在当然默认处理行为为“发送到控制器”时,将默认处理行为更改为“丢弃”;若否,不作使能动作。

5. 根据权利要求 3 所述的方法,其特征在于,所述步骤 S1 还包括:在所述流表中的最前一条配置一个用于监测流表老化时间的第三默认流条目,其中,“监测流表老化时间”的步骤具体包括:

指定一个定时器对所述第三默认流表的老化时间进行计时,并当老化时间来临时,将所述第三默认流条目的行为设置为所述默认处理行为中相应行为。

6. 一种基于 TCAM 实现 Openflow 多级流表的系统,其特征在于,该系统包括:

流表划分单元、用于将一张在 Openflow 交换机上维护的 TCAM 表划分为多个逻辑流表;

流表功能配置单元、用于为每一逻辑流表分别设定其所匹配的域类型。

7. 根据权利要求 6 所述的系统,其特征在于,所述流表中还包括一个用于配置流表默认处理行为的第一默认流条目,该第一默认流条目位于所述流表的最后一条。

8. 根据权利要求 7 所述的系统,其特征在于,所述默认处理行为包括丢弃、或者重定向到下一张逻辑流表、或者重定向到另一张逻辑流表、或者发送到控制器。

9. 根据权利要求 8 所述的系统,其特征在于,所述流表中还包括一个用于应对控制器断线状态的第二默认流条目,该第二默认流条目位于所述流表的倒数第二条,其中,该第二默认流条目具体用于:

配置第二默认流条目的默认状态为“禁用”;

判断当前控制器是否为断线状态,若是,则使能所述第二默认流条目,并且在当然默认处理行为为“发送到控制器”时,将默认处理行为更改为“丢弃”;若否,不作使能动作。

10. 根据权利要求 8 所述的系统,其特征在于,所述流表中还包括一个用于监测流表老化时间的第三默认流条目,该第三默认流条目位于所述逻辑流表的最前一条,其中,该第三默认流条目具体用于:

指定一个定时器对所述第三默认流条目的老化时间进行计时,并当老化时间来临时,将所述第三默认流条目的行为设置为所述默认处理行为中相应行为。

基于 TCAM 实现 Openflow 多级流表的方法及系统

技术领域

[0001] 本发明涉及以太网领域技术,尤其涉及一种基于 TCAM 实现 Openflow 多级流表的方法及系统。

背景技术

[0002] 目前互联网发展迅速,其应用规模的庞大及其地位的重要性,远远超出了设计者们的想象,而现在的互联网主要有以下弊端:特定的范围内没有统一的管理、网络转发机制与策略单一、运营商难以最大限度优化自身网络、以及传统交换机用于转发的功能和协议多且杂,很难配置,系统容易出错等问题。为应对上述弊端,人们提出了 Openflow 交换机,其将传统交换机上的报文转发和转发策略分离开来,采用专门的一台控制器(controller)通过网线和交换机连接。这样原来同在一台交换机设备上的报文转发功能(硬件芯片实现)和报文转发策略(各种软件协议)就被分开到了不同的硬件设备上。而一台控制器还可以控制多台 Openflow 交换机,从而实现了统一的转发控制端,更有效地控制了网络。

[0003] 通常的 Openflow 交换机上的流表是通过把流表配置到 TCAM 硬件上来实现的,TCAM(ternary content addressable memory)是一种三态内容寻址存储器,主要用于快速查找 ACL、路由等表项。其匹配字段长度可以是 80bit、160bit、320bit 等,TCAM 存储器容量一般为 2.5M,5M,10M,18M,20M,36M,72M 等,TCAM 可支持的流条目数 = 存储容量 ÷ 匹配字段长度。在 Openflow 交换机中,假设要匹配所有 12 个包头域,那么至少需要用到 TCAM 的匹配字段长度为 320bit。

[0004] 然而,现有技术中,一般的 Openflow 交换机只有一张普通流表,其所匹配的匹配字段长度为固定值(320bit),那么此技术的缺陷便在于:TCAM 流表存储条目数不够多,造成 TCAM 利用率低,难以满足不同应用需求。

发明内容

[0005] 本发明的目的在于提供一种基于 TCAM 实现 Openflow 多级流表的方法。

[0006] 相应于上述方法,本发明的另一目的在于提供一种基于 TCAM 实现 Openflow 多级流表的系统。

[0007] 为实现上述发明目的,本发明的技术方案如下:一种基于 TCAM 实现 Openflow 多级流表的方法,其包括:

S1、将一张在 Openflow 交换机上维护的 TCAM 表划分为多个逻辑流表;

S2、为每一逻辑流表分别设定其所匹配的域类型。

[0008] 作为本发明的进一步改进,所述步骤 S1 还包括:在所述流表中的最后一条配置一个用于配置流表默认处理行为的第一默认流条目。

[0009] 作为本发明的进一步改进,所述默认处理行为包括丢弃、或者重定向到下一张逻辑流表、或者重定向到另一张逻辑流表、或者发送到控制器。

[0010] 作为本发明的进一步改进,所述步骤 S1 还包括:在所述流表中的倒数第二条配置

一个用于应对控制器断线状态的第二默认流条目,其中,“应对控制器断线状态”的步骤具体为:

配置第二默认流条目的默认状态为“禁用”;

判断当前控制器是否为断线状态,若是,则使能所述第二默认流条目,并且在当然默认处理行为为“发送到控制器”时,将默认处理行为更改为“丢弃”;若否,不作使能动作。

[0011] 作为本发明的进一步改进,所述步骤 S1 还包括:在所述流表中的最前一条配置一个用于监测流表老化时间的第三默认流条目,其中,“监测流表老化时间”的步骤具体包括:指定一个定时器对所述第三默认流条目的老化时间进行计时,并当老化时间来临时,将所述第三默认流条目的行为设置为所述默认处理行为中相应行为。

[0012] 相应地,本发明提供一种基于 TCAM 实现 Openflow 多级流表的系统,其包括:

流表划分单元、用于将一张在 Openflow 交换机上维护的 TCAM 表划分为多个逻辑流表;

流表功能配置单元、用于为每一逻辑流表分别设定其所匹配的域类型。

[0013] 作为本发明的进一步改进,所述流表中还包括一个用于配置流表默认处理行为的第一默认流条目,该第一默认流条目位于所述流表的最后一条。

[0014] 作为本发明的进一步改进,所述默认处理行为包括丢弃、或者重定向到下一张逻辑流表、或者重定向到另一张逻辑流表、或者发送到控制器。

[0015] 作为本发明的进一步改进,所述流表中还包括一个用于应对控制器断线状态的第二默认流条目,该第二默认流条目位于所述流表的倒数第二条,其中,该第二默认流条目具体用于:

配置第二默认流条目的默认状态为“禁用”;

判断当前控制器是否为断线状态,若是,则使能所述第二默认流条目,并且在当然默认处理行为为“发送到控制器”时,将默认处理行为更改为“丢弃”;若否,不作使能动作。

[0016] 作为本发明的进一步改进,所述逻辑流表中还包括一个用于监测流表老化时间的第三默认流条目,该第三默认流条目位于所述逻辑流表的最前一条,其中,该第三默认流条目具体用于:指定一个定时器对所述第三默认流条目的老化时间进行计时,并当老化时间来临时,将所述第三默认流条目的行为设置为所述默认处理行为中相应行为。

[0017] 本发明的有益效果是:本发明通过将 TCAM 表划分为多个逻辑流表,且每一流表的匹配字段长度包括多种形式,这样有效增加了 TCAM 表的存储条目数,从而满足不同应用需求,进而可有效减小 Openflow 交换机的制造成本。

附图说明

[0018] 图 1 是本发明 Openflow 交换机系统的网络架构示意图;

图 2 是本发明具体实施方式中基于 TCAM 实现 Openflow 多级流表的方法的流程图;

图 3 是本发明具体实施方式中 TCAM 多级流表的结构示意图;

图 4 是本发明具体实施方式中基于 TCAM 实现 Openflow 多级流表的系统的单元示意图。

具体实施方式

[0019] 以下将结合附图所示的具体实施方式对本发明进行详细描述。但这些实施方式并不限制本发明,本领域的普通技术人员根据这些实施方式所做出的结构、方法、或功能上的变换均包含在本发明的保护范围内。

[0020] 参图 1 所示,其为本发明 Openflow 交换机系统的网络架构,其包括一个或者多个控制器(controller1、controller2、controller3...controllerN),控制器与 Openflow 交换机(一个或多个)之间通过控制网络相连,Openflow 交换机另一端通过多业务端口连接多个终端设备(终端 1、终端 2、终端 3...终端 N),上述终端设备可包括手机、电脑、PDA(个人数字处理)等。所述终端可通过数据通道与 Openflow 交换机相互通信,实现报文发送。

[0021] 参图 2 及图 3 所示,基于上述的交换机系统,在本发明具体实施方式中,一种基于 TCAM 实现 Openflow 多级流表的方法,其包括:

S1、将一张在 Openflow 交换机上维护的 TCAM 表划分为多个逻辑流表 100,所述逻辑流表 100 可以包括多个普通流条目 101(普通流_1、普通流_2...普通流_N);

S2、为每一逻辑流表 100 分别设定其所匹配的域类型。其中,域类型包括其支持的域可为 2 层域、4 层域等。

[0022] 本发明的 Openflow 交换机上会维护一张 TCAM 表 100,该 Openflow 交换机上所提出来的 12 个包头域包括:入端口(INGRESS_PORT),源以太网地址(ETH_SRC),目的以太网地址(ETH_DST),以太网类型(ETH_TYPE),VLAN 标示 ID(VLAN_ID),VLAN 优先级(VLAN_PCP),源 IP 地址(IP_SRC),目的 IP 地址(IP_DST),IP 协议号(IP_PROTO),IP 服务类型(IP_TOS),4 层源端口(TP_SRC_PORT),4 层目的地址(TP_DST_PORT)。那么,根据实际应用环境和经验,一般应用上用户并不需要全部用到上述 12 个域来做匹配,如用户只需要匹配 2 层域(如 VLAN_ID,ETH_SRC,ETH_DST),或者只需要匹配 3 层域或 4 层域(如 IP_SRC,IP_DST,TP_SRC_PORT,TP_DST_PORT)等。

[0023] 本方法根据不同应用需求,可以把 TCAM 表划分为多张逻辑流表 100 做不同域匹配,每张逻辑流表使用的匹配字段长度可以不同,从而可以增加可使用的流条目数,进而大幅提高 TCAM 的使用效率。其中,在本发明具体实施例中:以一块 5M 的 TCAM 为例,如果其同时支持匹配 12 个域,相应需要 320bit 的字段匹配长度,那么其可支持的流条目数为: $5 \times 1024 \times 1024 \div 320 = 16384$ 条(可支持的流条目数 = 存储容量 $\div$ 匹配字段长度),若采用本发明的技术方案,即把 5M 的 TCAM 划分成相等容量的 2 张逻辑表,每张表的容量为 2.5M,一张表专门匹配 2 层域(如 VLAN_ID,ETH_SRC,ETH_DST),另一张表专门匹配 4 层域(如 TP_SRC_PORT,TP_DST_PORT),由于单独的 2 层域只需要 80bit 就能够实现匹配,而单独的 4 层域匹配也只需要 160bit,那么多级流表可以支持的条目数为: $2.5 \times 1024 \times 1024 \div 80 + 2.5 \times 1024 \times 1024 \div 160 = 32768 + 16384 = 49150$ 条。可明显看出,多级流表可支持的条目数远远高于一般的单流条目数。此外,在本发明其他具体实施例中,用户完全可以根据不同运行场景,划分出更多的逻辑流表来支持不同的应用,从而最大程度的利用 TCAM 资源。

[0024] 继续参图 2 及图 3,在本发明另一实施方式中,步骤 S1 还包括:在所述逻辑流表 100 的最后一配置一个用于配置流表默认处理行为的第一默认流条目 102(步骤 S12)。本发明中,对于每张流表,需要指明其默认处理行为,这样可以更加灵活的对每张表的整体行为进行控制。具体地,上述默认处理行为可包括丢弃、或者重定向到下一张逻辑流表、或者重定向到另一张逻辑流表、或者发送到控制器等。

[0025] 继续参图 2 及图 3, 在本发明另一实施方式中, 对于每张逻辑流表 100, 需要对控制器的连线状态做出判断, 防止在控制器断线状态下, 大量未匹配报文发送和冲击到 CPU, 而造成系统不稳定。故, 上述步骤 S1 还包括: 在所述逻辑流表 100 的倒数第二条配置一个用于应对控制器断线状态的第二默认流条目 104 (步骤 S14), 其中, “应对控制器断线状态”的步骤具体为:

配置第二默认流表 104 的默认状态为“禁用”;

判断当前控制器是否为断线状态, 若是, 则使能所述第二默认流条目 104, 并且在当然默认处理行为为“发送到控制器”时, 将默认处理行为更改为“丢弃”, 这样在控制器断线的时候, 需要发送到控制器的报文会自动丢弃, 而不会发送到 CPU, 来影响系统性能; 若否, 不作使能动作。

[0026] 继续参图 2 及图 3, 在本发明另一实施方式中, 上述步骤 S1 还包括: 在所述逻辑流表 100 的最前一条配置一个用于监测流表老化时间的第三默认流条目 106 (步骤 S16), 其中, “监测流表老化时间”的步骤具体包括: 指定一个定时器对所述第三默认流条目 106 的老化时间进行计时, 并当老化时间来临时, 将所述第三默认流条目 106 的行为设置为所述默认处理行为中相应行为(丢弃、重定向到下一张逻辑表、重定向到另一张逻辑表、发送到控制器)。如此, 本发明也可有效地对每张表的时效性进行控制。

[0027] 接下来, 请参图 3 及图 4 所示, 在本发明具体实施方式中, 一种基于 TCAM 实现 Openflow 多级流表的系统, 其包括:

流表划分单元 10、用于将一张在 Openflow 交换机上维护的 TCAM 表划分为多个逻辑流表 100, 所述逻辑流表 100 可以包括多个普通流条目 101 (普通流_1、普通流_2...普通流_N);

流表功能配置单元 20、用于为每一逻辑流表 100 分别设定其所匹配的域类型。其中, 域类型包括其支持的域可为 2 层域、4 层域等。

[0028] 本发明的 Openflow 交换机会维护一张 TCAM 表 100, 该 Openflow 交换机上所提出来的 12 个包头域包括: 入端口 (INGRESS_PORT), 源以太网地址 (ETH_SRC), 目的以太网地址 (ETH_DST), 以太网类型 (ETH_TYPE), VLAN 标示 ID (VLAN_ID), VLAN 优先级 (VLAN_PCP), 源 IP 地址 (IP_SRC), 目的 IP 地址 (IP_DST), IP 协议号 (IP_PROTO), IP 服务类型 (IP_TOS), 4 层源端口 (TP_SRC_PORT), 4 层目的地址 (TP_DST_PORT)。那么, 根据实际应用环境和经验, 一般应用上用户并不需要全部用到上述 12 个域来做匹配, 如用户只需要匹配 2 层域 (如 VLAN_ID, ETH_SRC, ETH_DST), 或者只需要匹配 3 层域或 4 层域 (如 IP_SRC, IP_DST, TP_SRC_PORT, TP_DST_PORT) 等。

[0029] 本系统根据不同应用需求, 可以把 TCAM 表划分为多张逻辑流表 100 做不同域匹配, 每张逻辑流表使用的匹配字段长度可以不同, 从而可以增加可使用的流条目数, 进而大幅提高 TCAM 的使用效率。其中, 在本发明具体实施例中: 以一块 5M 的 TCAM 为例, 如果其同时支持匹配 12 个域, 相应需要 320bit 的字段匹配长度, 那么其可支持的流条目数为: $5 \times 1024 \times 1024 \div 320 = 16384$ 条 (可支持的流条目数 = 存储容量 $\div$ 匹配字段长度), 若采用本发明的技术方案, 即把 5M 的 TCAM 划分成相等容量的 2 张逻辑表, 每张表的容量为 2.5M, 一张表专门匹配 2 层域 (如 VLAN_ID, ETH_SRC, ETH_DST), 另一张表专门匹配 4 层域 (如 TP_SRC_PORT, TP_DST_PORT), 由于单独的 2 层域只需要 80bit 就能够实现匹配, 而单独的 4

层域匹配也只需要 160bit,那么多级流表可以支持的条目数为: $2.5 \times 1024 \times 1024 \div 80 + 2.5 \times 1024 \times 1024 \div 160 = 32768 + 16384 = 49150$ 条。可明显看出,多级流表可支持的条目数远远高于一般的单流条目数。此外,在本发明其他具体实施例中,用户完全可以根据不同运行场景,划分出更多的逻辑流表来支持不同的应用,从而最大程度的利用 TCAM 资源。

[0030] 如图 3 所示,在本发明另一实施方式中,所述逻辑流表 100 还包括一个用于配置流表默认处理行为的第一默认流条目 102,该第一默认流条目 102 位于所述流表的最后一条。本发明中,对于每张逻辑流表,需要指明其默认处理行为,这样可以更加灵活的对每张表的整体行为进行控制。具体地,上述默认处理行为可包括丢弃、或者重定向到下一张逻辑流表、或者重定向到另一张逻辑流表、或者发送到控制器等。

[0031] 继续参图 3,在本发明另一实施方式中,对于每张逻辑流表,需要对控制器的连线状态做出判断,防止在控制器断线状态下,大量未匹配报文发送和冲击到 CPU,而造成系统不稳定。故,所述流表中还包括一个用于应对控制器断线状态的第二默认流条目 104,该第二默认流条目 104 位于所述流表的倒数第二条,其中,该第二默认流条目 104 具体用于:

配置第二默认流条目 104 的默认状态为“禁用”;

判断当前控制器是否为断线状态,若是,则使能所述第二默认流条目 104,并且在当然默认处理行为为“发送到控制器”时,将默认处理行为更改为“丢弃”,这样在控制器断线的时候,需要发送到控制器的报文会自动丢弃,而不会发送到 CPU,来影响系统性能;若否,不作使能动作。

[0032] 继续参图 3,在本发明另一实施方式中,所述逻辑流表 100 还包括一个用于监测流表老化时间的第三默认流条目 106,该第三默认流条目 106 位于所述流表的最前一条,其中,该第三默认流条目 106 具体用于:指定一个定时器对所述第三默认流条目的老化时间进行计时,并当老化时间来临时,将所述第三默认流条目的行为设置为所述默认处理行为中相应行为(丢弃、重定向到下一张逻辑表、重定向到另一张逻辑表、发送到控制器)。如此,本发明也可有效地对每张表的时效性进行控制。

[0033] 综上所述,本发明的系统及相应方法均通过将 TCAM 表划分为多个逻辑流表,且每一流表的匹配字段长度包括多种形式,这样有效增加了 TCAM 表的存储条目数,从而满足不同应用需求,进而可有效减小 Openflow 交换机的制造成本。

[0034] 以上所描述的装置实施方式仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施方式方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0035] 值得一提的是,本发明所介绍的方法中,所提及的步骤序号“S1”、“S2”...“Sn”之间并不一定存在先后顺序,特此声明。

[0036] 应当理解,虽然本说明书按照实施方式加以描述,但并非每个实施方式仅包含一个独立的技术方案,说明书的这种叙述方式仅仅是为清楚起见,本领域技术人员应当将说明书作为一个整体,各实施方式中的技术方案也可以经适当组合,形成本领域技术人员可以理解的其他实施方式。

[0037] 上文所列出的一系列详细说明仅仅是针对本发明的可行性实施方式的具体说

明,它们并非用以限制本发明的保护范围,凡未脱离本发明技艺精神所作的等效实施方式或变更均应包含在本发明的保护范围之内。

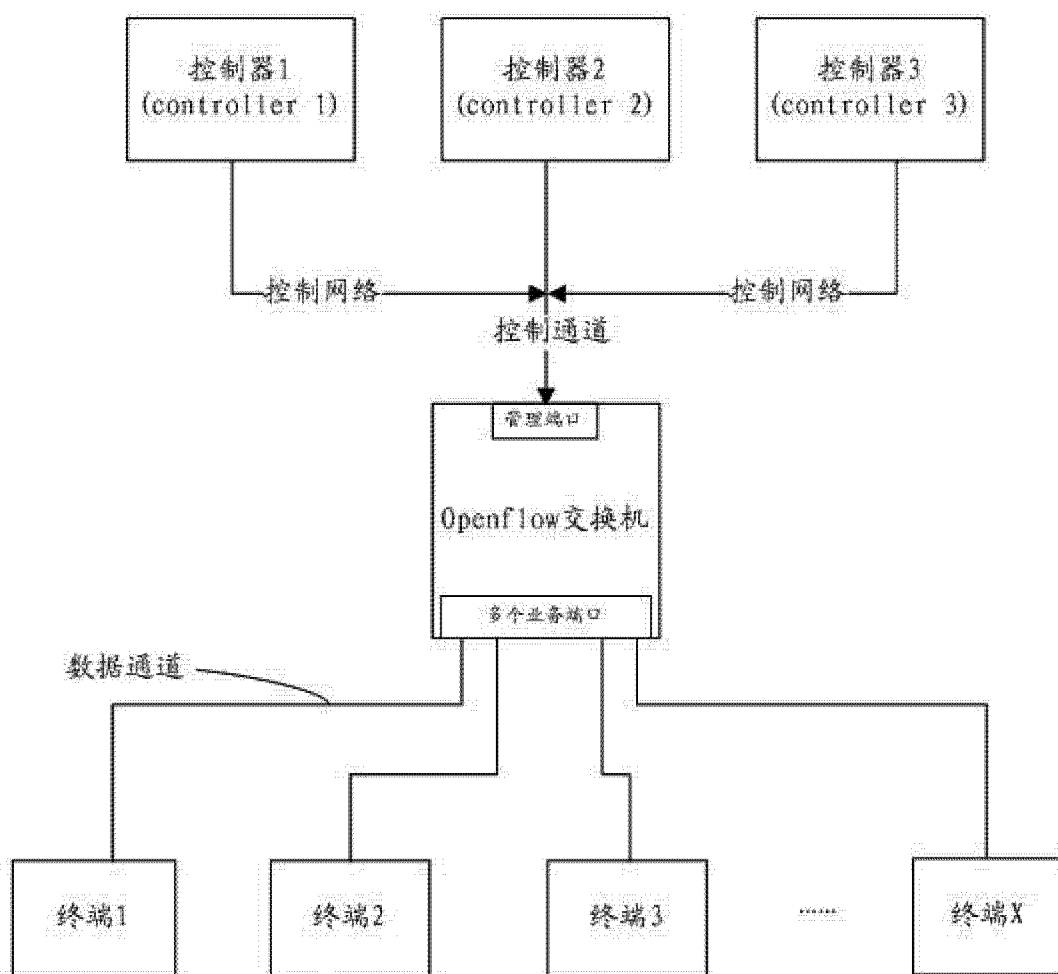


图 1

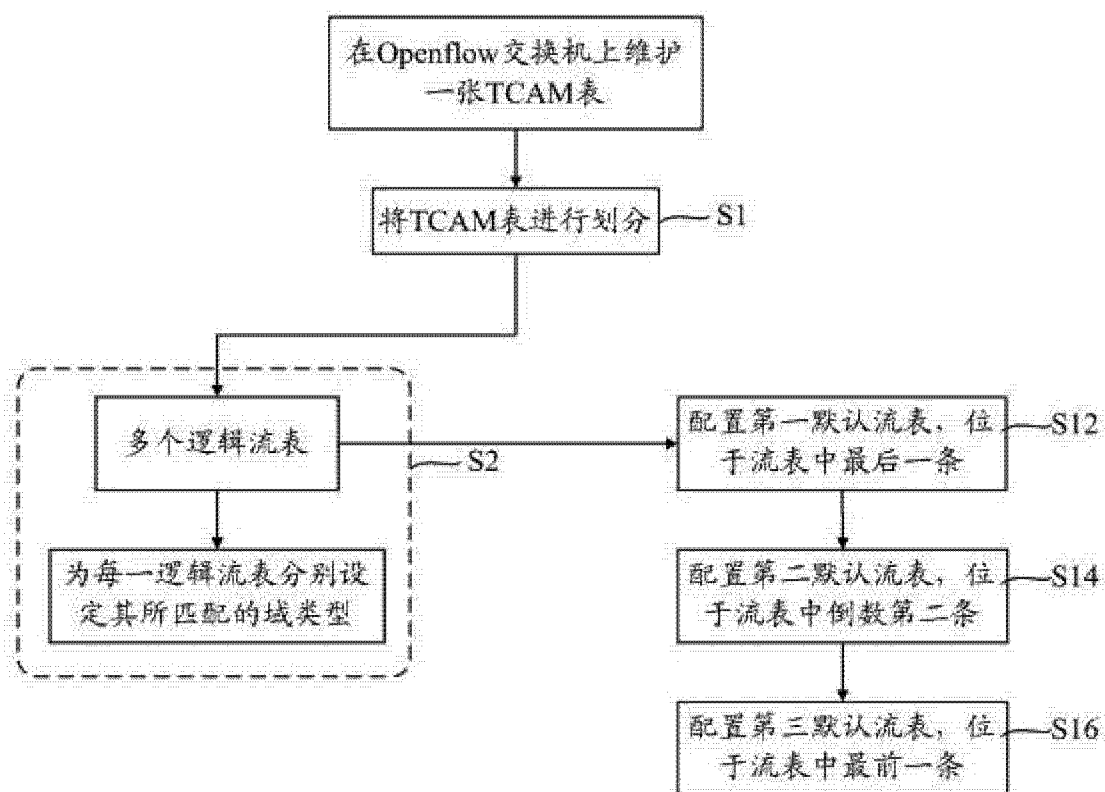


图 2

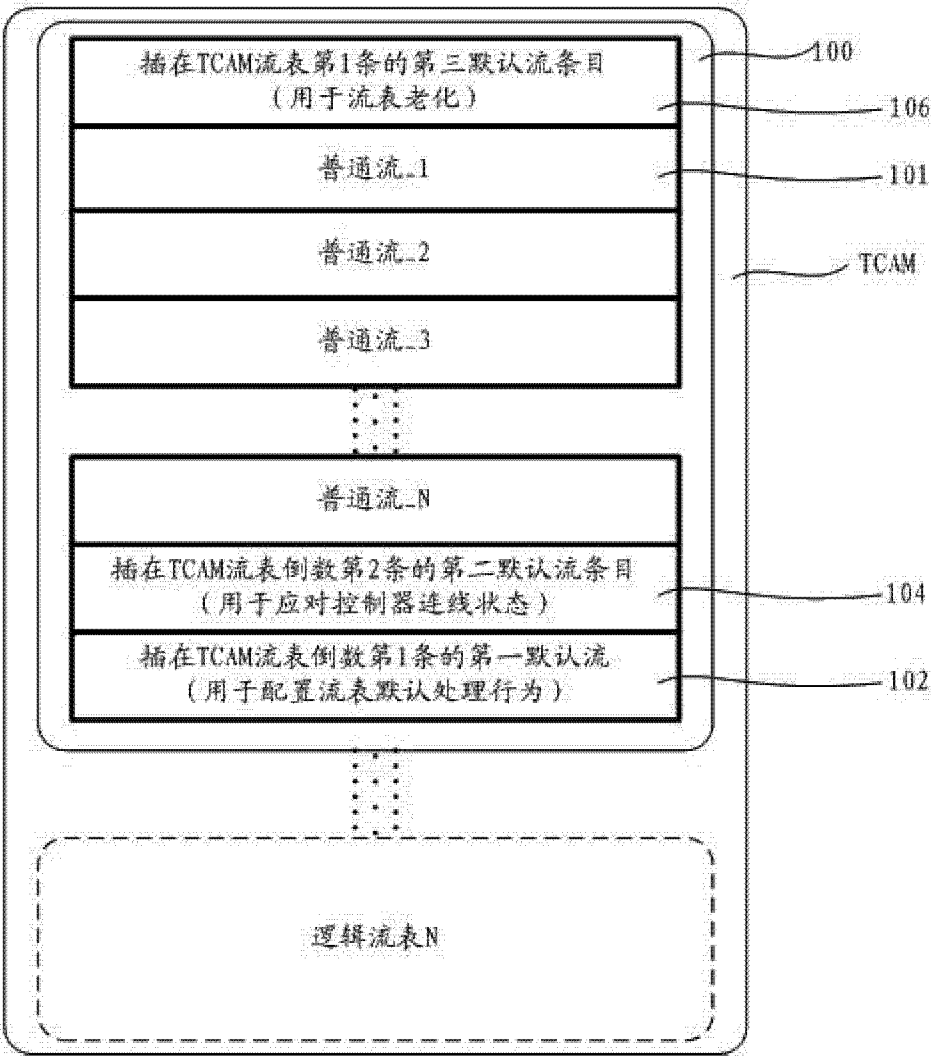


图 3

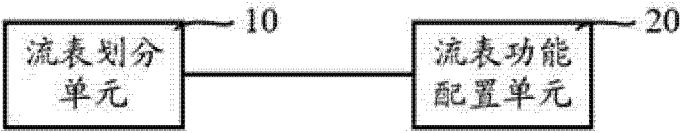


图 4