

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

H04L 12/56 (2006.01)

H04L 29/06 (2006.01)



# [12] 发明专利申请公开说明书

[21] 申请号 200510109556.5

[43] 公开日 2006年5月24日

[11] 公开号 CN 1777143A

[22] 申请日 2005.10.25

[21] 申请号 200510109556.5

[30] 优先权

[32] 2004.10.25 [33] US [31] 10/972,524

[71] 申请人 阿尔卡特公司

地址 法国巴黎市

[72] 发明人 奇安·耶 丹尼斯·韦弗

[74] 专利代理机构 北京市金杜律师事务所

代理人 朱海波

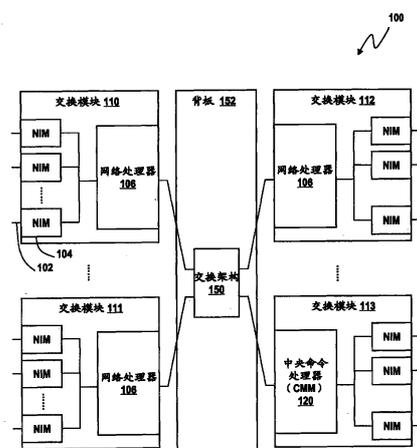
权利要求书 2 页 说明书 16 页 附图 6 页

## [54] 发明名称

使用分布式网络处理的数据交换机中的内部负载平衡

## [57] 摘要

本发明公开了一种用于在入口处理器与出口处理器之间动态地分布数据包处理操作以使负载平衡的数据通信交换机。在优选实施例中，本发明的特征表现为一种包括多个交换模块的交换设备，每个交换模块包括：数据包分类器，其用于识别将应用于入口数据包的一个或多个数据包处理操作；以及控制器，其适用于在第一组数据包处理操作与第二组数据包处理操作之间分配所识别的一个或多个数据包处理操作中的每个数据包处理操作，在接收数据包的入口处理器中执行第一组数据包处理操作，并发送指令到出口处理器以执行第二组数据包处理操作。然后，出口处理器执行第二组数据包处理操作，之后可以将数据包向其目的地节点发送。



1. 一种包括多个交换模块的交换设备, 每个交换模块包括:  
适用于接收协议数据单元 (PDU) 的至少一个外部端口;  
5 分类器, 适用于识别:  
应用于所述 PDU 的一个或多个 PDU 处理操作; 以及  
所述多个交换模块中接收所转发的数据包的第二交换模块; 以及  
控制器, 适用于:  
在第一组 PDU 处理操作与第二组 PDU 处理操作之间分配所述一  
10 个或多个 PDU 处理操作中的每个 PDU 处理操作;  
在从所述外部端口接收到所述 PDU 之后, 执行第一组 PDU 处理  
操作; 以及  
发送执行所述第二组 PDU 处理操作的命令到所述第二交换模块。
2. 根据权利要求 1 所述的交换设备, 其中每个交换模块还适用于  
15 响应于来自多个交换模块中的一个交换模块的命令而执行所述第二  
组 PDU 处理操作。
3. 根据权利要求 2 所述的交换设备, 其中每个交换模块还适用于  
发送所述 PDU 到接收所述命令的交换模块。
4. 根据权利要求 1 所述的交换设备, 其中所述 PDU 处理操作包  
20 括 PDU 转发操作。
5. 根据权利要求 1 所述的交换设备, 其中所述 PDU 处理操作选  
自如下操作: 报头转换、标记推送、标记弹出、服务质量、计费 and 记  
账、多协议标签交换 (MPLS) 管理、生成树操作、认证、访问控制、  
高层学习、警报生成、端口镜像、源学习、服务分类、色彩标记及其  
25 组合。
6. 一种对包括第一交换模块和第二交换模块的交换设备中的协  
议数据单元 (PDU) 交换操作进行分配的方法, 所述方法包括步骤:  
在所述第一交换模块的外部端口接收 PDU;  
识别应用于所述 PDU 的一个或多个 PDU 处理操作;

在第一组 PDU 处理操作与第二组 PDU 处理操作之间分配所述一个或多个 PDU 处理操作中的每个 PDU 处理操作；

在所述第一交换模块中执行所述第一组 PDU 处理操作；以及  
发送命令到所述第二交换模块以执行所述第二组 PDU 处理操作。

5 7. 根据权利要求 6 所述的方法，其中所述方法还包括步骤：响应于来自所述第一交换模块的命令，在所述第二交换模块中执行所述第二组 PDU 处理操作。

8. 根据权利要求 6 所述的方法，其中所述方法还包括步骤：发送所述 PDU 到所述第二交换模块。

10 9. 根据权利要求 6 所述的方法，其中所述方法还包括步骤：将包括标识符的报头附加到从所述第一交换模块发送到所述第二交换模块的所述 PDU，所述标识符表明所述一个或多个 PDU 处理操作在所述第一交换模块与所述第二交换模块之间的分配。

15 10. 根据权利要求 6 所述的方法，其中所述方法还包括步骤：从在所述第一交换模块与所述第二交换模块之间分配所述一个或多个 PDU 处理操作中的每个 PDU 处理操作的分配表中检索所述标识符。

## 使用分布式网络处理的数据交换机中的内部负载平衡

### 5 技术领域

本发明一般地涉及数据通信网络中用于在不同的数据包处理单元之间分布处理操作并由此分布处理负载的交换设备。特别地，本发明涉及一种用于在入口处理器和出口处理器之间分配数据包处理操作以使负载不平衡最小化的系统和方法。

10

### 背景技术

诸如路由器或交换机之类的联网设备内网络处理负载的分布通常是固定的，并且很大程度上由联网设备的体系结构所决定。例如，取决于厂商，一般预先确定转发决定（forwarding decision）和其他数据包处理是在入口执行还是在出口执行。由于固有的不对称处理负载和不对称业务模式，入口点处的网络处理负载很少会等于出口点处的网络处理负载。事实上，入口和出口之间的处理负载差异通常很大，以至于经常是交换机一端的处理器很忙碌，而另一端的处理器几乎保持空闲。

20

尽管一些联网设备可以与其他联网设备共享某些处理负载以平衡可用处理器上的处理工作量，但是这些方案不能动态地平衡数据包处理操作，因为：（a）网络处理典型地由硬连线的（hard-wired）专用集成电路（ASIC）设备实现，这种设备非常擅长重复性任务，但一般缺乏动态改变任务分布的智能；（b）采用分布式处理的网络集群都设计为处理一组大型且固定的任务，但这种网络集群相对较慢且不适合于数据包处理；以及（c）将大多数联网设备都简单地过度工程化（over-engineered）以适应最差的可能负载，而没有考虑到在系统级上浪费了过多的计算能力。

25

因此，需要一种能够监视数据包处理负载不平衡并在入口处理器

和出口处理器之间动态地分布负载以使不平衡最小化的网络交换设备。

### 发明内容

5 在优选实施例中，本发明的特征表现为一种包括多个交换模块的交换设备，每个交换模块包括：至少一个外部端口，其适用于接收数据包或其他协议数据单元（PDU）；数据包分类器，其适用于基于一个或多个数据包特性来检查数据包并识别应用于该数据包的一个或多个数据包处理操作，并适用于识别接收所转发的数据包的多个交换  
10 模块中的第二交换模块；控制器，其适用于在第一组数据包处理操作与第二组数据包处理操作之间分配所识别的一个或多个数据包处理操作中的每个数据包处理操作，在包括接收数据包的外部端口的交换模块中执行第一组数据包处理操作，并发送命令到第二交换模块指示该第二交换模块执行第二组数据包处理操作。如同优选实施例中多个  
15 交换模块中的每个交换模块一样，第二交换模块适用于响应于该命令而执行第二组数据包处理操作，之后可以将数据包从外部端口向其目的地节点发送。

在优选实施例中，交换设备是路由器、网桥或多层交换机，而数据包处理操作是为了准备用于发送到其目的地节点方向上的下一个  
20 节点的数据包而执行的数据包转发操作。根据优选实施例，可以在多个交换模块的入口交换模块或出口交换模块中串行地分布和执行数据包处理操作。数据包处理操作一般选自群组但不限于群组，该群组包括：报头转换、标记推送（push）、标记弹出（pop）、服务质量、计费 and 记账、多协议标签交换（MPLS）管理、生成树操作、认证、  
25 访问控制、高层学习、警报生成、端口镜像、源学习、服务分类、色彩标记及其组合。

在优选实施例中，在交换设备的第一交换模块与第二交换模块之间分配数据包交换操作的方法包括步骤：在第一交换模块的外部端口处接收数据包；识别应用于数据包的一个或多个数据包处理操作；在

第一组数据包处理操作与第二组数据包处理操作之间分配该一个或多个数据包处理操作中的每个数据包处理操作；在第一交换模块中执行第一组数据包处理操作；以及发送命令到第二交换模块以执行第二组数据包处理操作。本方法还包括响应在于该命令而第二交换模块中  
5 执行第二组数据包处理操作的步骤。

### 附图说明

通过示例和附图对本发明进行说明，并且本发明不限于附图的图形，并且其中：

10 图 1 是根据本发明的优选实施例的企业交换机的功能性框图；

图 2 是根据本发明的优选实施例用于执行串行分布的数据包处理的企业交换机的交换模块的功能性框图；

图 3 是根据本发明优选实施例的交换模块的本地存储器中所保存的数据库的功能性框图；

15 图 4 是根据本发明优选实施例的串行分布的数据包处理（SDPP）控制器的功能性框图；

图 5 是根据本发明的优选实施例用于在交换模块之间分配 SDPP 服务的报头的功能性框图；

20 图 6 是根据本发明优选实施例的入口交换模块处理入口数据流的方法的流程图；以及

图 7 是根据本发明优选实施例的出口交换模块处理出口数据流的方法的流程图。

### 具体实施方式

25 图 1 中示出的是企业交换机的功能性框图。企业交换机 100 包括多个节点中的一个节点和其他可寻址实体，这些节点和实体可操作地连接到数据通信网络，该数据通信网络具体为例如局域网（LAN）、广域网（WAN），或城域网（MAN）、网际协议（IP）网、因特网，或其组合。

企业交换机 100 优选地包括多个交换模块 110-113，有时称为刀片 (blade)，将这些交换模块固定在背板 152 上的插槽中。每个交换模块 110-113 包括一个或多个外部端口 102，每个外部端口都可以通过通信链路(未示出)可操作地连接到数据通信网络中的另一个节点，  
5 并且一个或多个内部端口通过共享的交换架构 150 将每个交换模块 110-113 连接到每个其他的交换模块。

交换模块 110-113 优选地包括至少一个网络处理器 (NP) 106，其能够进行如开放系统互联 (OSI) 参考模型中所定义的至少层 2 (数据链路层) 和层 3 (网络层) 的交换操作但不限于这些操作。用于将  
10 外部端口 102 可操作地连接到有线和/或无线通信链路的一种可能的层 2 协议是电气和电子工程师协会 (IEEE) 的 802.3 标准，而一组可能的层 3 协议包括因特网工程任务组 (IETF) 的请求注释 (RFC) 791 中定义的网际协议 (IP) 版本 4 和 IETF RFC 1883 中定义的 IP 版本 6。

对于本公开来说，从外部端口 102 到架构 150 的流入交换模块  
15 110-113 的数据在此称作入口数据，其包括入口 PDU。入口数据传播所通过的交换模块称作入口交换模块。相反，从架构 150 流向外端口 102 的数据称作出口数据，其包括出口 PDU。出口数据传播所通过的交换模块称作出口交换模块。对于不同的流，优选实施例的多个交换模块中的每个交换模块可以同时用作入口交换模块和出口交换模  
20 块。

企业交换机 100 还包括用于管理不同系统资源的中心命令处理器 (CMM) 120，管理不同系统资源包括下面将详细讨论的拥塞监控和操作分配。在优选实施例中，CMM 120 包含在多个交换模块 110-113 中的一个交换模块中，但是本领域的普通技术人员应当意识到，CMM  
25 执行的功能可以由合成在背板 152 上的一个或多个实体和/或一个单独的管理模块来执行。

图 2 中示出的是用于执行 PDU 流的串行分布式处理的交换模块的功能性框图。和交换模块 110-113 一样，该优选实施例的交换模块 200 包括一个或多个网络接口模块 (NIM) 104、一个或多个网络处理

器 106、一个管理模块 220 以及一个架构接口模块 208。为了接收入口数据业务和发送出口数据业务，将每个 NIM 104 可操作地连接到一个或多个外部端口 102。NIM 104 优选地包括适用于通过网络通信链路（未示出）交换例如以太网帧之类的 PDU 的一个或多个物理接口和媒体访问控制（MAC）接口。接收到入口 PDU 之后，通过一个或多个内部高速串行数据总线 206 将其从多个 NIM 104 传送到网络处理器 106。网络处理器 106 优选地对入口 PDU 进行分类，执行分配为在入口处执行的任意转发操作，并且在入口队列存储器 248 中将这  
5 PDU 排成队列，直到可获得带宽来通过交换架构 150 将这些 PDU 发送到适当的一个或多个出口刀片。

关于出口操作，交换模块 200 还适用于接收来自交换架构 150 的出口 PDU，并且在出口队列存储器 242 中将这 PDU 排成队列。在将 PDU 提交给缓冲器 250 中的一个或多个队列并将这 PDU 发送到适当的 NIM 104 和相应的出口端口 102 之前，出口处的交换模块 200  
15 的 NP 106 可以执行分配给它的一个或多个另外的转发操作。缓冲器 250 中的多个队列由统计管理器 252 进行主动监控，统计管理器 252 汇总拥塞信息 254 并通过管理模块 220 将其发送到 CMM 120。拥塞信息 254 包括例如队列深度，用于表现出口数据流的特征，评定模块的拥塞状态，并在入口网络处理器和出口网络处理器之间分配数据包  
20 处理操作。

管理模块 220 一般包括用于保持和实现业务策略的策略管理器 222，这些业务策略由网络管理员通过配置管理器 224 上传到交换模块 200。由策略管理器 222 所产生的策略还部分地基于由源学习操作得出的层 2 和层 3 寻址信息，这些寻址信息将 PDU 地址信息与接收  
25 这些信息的外部端口 102 相关联。如下面将更详细描述，管理模块 222 还适用于将更新 254 从 CMM 发送到网络处理器 106，使得在发送 PDU 到一个或多个下游交换模块之前，入口交换模块 200 可以在入口处执行一部分或全部数据包处理操作。更新 254 包括用于填充多种数据库的数据，这些数据库支持下面将更详细描述串行分布式数

据包处理操作。

该优选实施例的 NP 106 适用于利用 OSI 网络参考模型中定义的和层 2 到层 7 相关联的 PDU 特性来执行层 2 的交换操作和层 3 的路由操作。NP 106 优选地包括分类器 230、串行分布式数据包处理 (SDPP) 控制器 236 以及队列管理器 240。分类器 230 从数据总线 206 接收入口 PDU，检查 PDU 的一个或多个感兴趣的字段，将 PDU 归类为多个流中的一个流，并从保存在高速本地存储器 232 中的转发表中检索转发信息。转发信息优选地包括但不限于流标识符和出口端口标识符，即将发送 PDU 的外部端口标识符。

10 将分类器 230 检索到的转发信息发送到 SDPP 控制器 236，其中将该转发信息用于识别将在入口处执行的第一组一个或多个 SDPP 操作。在此使用的 SDPP 操作指的是数据包处理操作或其他由于 PDU、响应 PDU 或便于 PDU 从交换设备 100 发送而执行的转发操作。入口交换模块 200 或 CMM 120 也可以识别将在出口交换模块执行的第二组 SDPP 操作。本领域的普通技术人员应当意识到，SDPP 操作可以在入口交换模块或出口交换模块中执行。下面将结合图 5 更详细地讨论可能的 SDPP 服务的范围。

一般来说，入口交换模块 200 可以执行所有的、某一些或不执行入口交换模块中的 SDPP 操作。在从交换设备 100 中发送 PDU 之前，在出口交换模块中执行用于 PDU 而又不在入口交换模块 200 中执行的任意 SDPP 操作。

应用于入口的第一组一个或多个 SDPP 服务与将由出口交换模块执行的第二组 SDPP 服务之间的 SDPP 操作的分配可以在每个数据包和/或每个流的基础上动态地确定。

25 在优选实施例中，根据流 ID 以及入口交换模块和出口交换模块的拥塞状态确定将要在入口处执行的 SDPP 服务的分配。在有些实施例中，入口交换模块和出口交换模块的拥塞状态由 CMM 120 定期进行汇总并报告给每个交换模块以便每个交换模块可以动态地确定 SDPP 服务的最有利的分配方式。在其他实施例中，SDPP 操作的最优

分配由 CMM 120 确定，下载到多个交换模块 110-113 中的每一个交换模块，并结合下面将更详细讨论的各种共享数据库以 SDPP 分配表 238 的形式保存在图 3 所示的本地存储器 232 中。

例如，该 SDPP 分配表 238 可以例如在每个流或每个数据包的基础上为每对入口和出口交换模块明确定义 SDPP 服务的分配。可以以一定的间隔每秒一次或多次对 SDPP 分配表进行更新，以反映变化的业务模式和刀片间的负载不平衡。

在优选实施例中，根据入口交换模块和出口交换模块的相对拥塞状态在入口和出口之间分配 SDPP 服务。特别地，将 SDPP 分配设计为将处理负载从具有过度使用的 NP 的交换模块转移到具有未充分使用的 NP 的交换模块。因此，SDPP 服务分布的目标是在整个交换设备 100 中均匀地分布处理资源的消耗，并由此使任何交换模块由于超出限度的业务条件而丢失 PDU 的概率最小化。通过将处理负载的分布最优化，交换设备 100 不必付出将所有交换模块都过度工程化的代价，就可以适应可能发生在某些端口处的不成比例的高业务条件，这些端口例如用户所连接的端口或为因特网提供网关的端口。

图 4 中示出的是优选实施例的 SDPP 控制器的功能性框图。SDPP 控制器 236 包括包含在一个或多个硬件或软件计算单元中的入口 NP 处理器 410 和出口处理器 420、SDPP 分配表 238 以及适用于执行单个 SDPP 服务 430 的一个或多个模块。入口 NP 处理器 410 适用于从分类器 230 接收入口 PDU 460 和流标识符 462，并适用于向 SDPP 分配表 238 询问可应用于单个数据包或流的 SDPP 服务分配。在优选实施例中，指定给入口交换模块和出口交换模块的 SDPP 服务分配由下面详述的 SDPP 标识符来表示。随后，根据来自 SDPP 服务 430 的全部选择的第一组 SDPP 服务，在入口交换模块 200 中处理 PDU。

然后，将 PDU 464 和 SDPP 标识符 466 发送到出口交换模块，以指示出口模块执行分配给出口交换模块的第二组 SDPP 服务 430。在优选实施例中，在 PDU 或 PDU 描述符发送到交换架构 150 之前，由操作 SDPP 标记生成器 412 将 SDPP 标识符 466 附加到 PDU 或 PDU

描述符上。在其他的实施例中，可以通过例如带外通信信道来发送 SDPP ID。

在接收到 PDU 470 或其描述符之后，在出口交换模块中，出口交换模块的出口处理器 420 除去 SDPP 标识符 472，SDPP 标记读取器 5 422 通过查询 SDPP 分配表 238 来确定由 SDPP 标识符指定的 SDPP 服务，并且出口交换模块执行第二组 SDPP 服务 430，这组 SDPP 服务是完成向 PDU 476 的最终目的地方向发送 PDU 476 所必需的转发操作所需的 SDPP 服务。

例如，在优选实施例中，SDPP 服务 430 的列表包括但不限于下列转发操作：报头转换、标记推送、标记弹出、服务质量、计费 and 记账，多协议标签交换（MPLS）管理、生成树操作、认证、访问控制、高层学习、警报生成、端口镜像、源学习、服务分类，以及色彩标记，以及监控（policing）和整形（shaping）。

报头转换服务 431 一般包括步骤：（a）检索 PDU 的下一跳地址，其指定了到最终目的地的路径中的下一节点的物理地址；（b）用包括源地址和目的地地址的报头来封装 PDU，其中源地址等于交换设备 100 或出口交换模块的物理地址，目的地址等于到最终目的地的路径上的下一跳的物理地址；以及（c）递减例如 IP 包的生存时间计数器。例如，如果交换了在此包含的寻址信息或者由 CMM120 对表格进行了有规律的同步，则可以从入口交换模块或出口交换模块的本地存储器 20 232（见图 3）的转发表 302 中检索下一跳的地址。

VLAN 标记推送服务 432 通常包括步骤：（a）基于 PDU 特性来识别一个或多个 VLAN 标识符（VID），这些特性包括例如 PDU 源 MAC 地址、目的地 MAC 地址、协议和入口端口等；以及（b）用一个或多个 VLAN 标记来封装 PDU。例如，基于一个或多个 PDU 特性，可以从本地存储器 232（见图 3）中用于查找的 VID 的 VLAN 关联表 304 中检索到适当的标记。一旦例如由网络管理员定义的 VLAN 关联表 304 分布到交换机 100 的多个交换模块 110-113 中的每个交换模块或 VLAN 关联表 304 在每个交换模块之间取得同步，就可以在入口交

换模块或出口交换模块中执行 VLAN 标记推送服务。

VLAN 标记弹出服务 433 包括步骤：例如，如果交换机 100 是倒数第二跳，或者当 PDU 从公网转换到专用网的无标记域（untagged domain）时，将一个或多个 VLAN 标记从 PDU 上除去。如同上述 VLAN 5 标记推送服务 432，VLAN 关联表 304 中包含的 VLAN 弹出规则分布于多个交换模块 110-113 之间或由多个交换模块 110-113 共享。

服务质量（QoS）操作 434 包括步骤：采用例如资源保留协议（RSVP）之类的机制来保留包括例如存储器和带宽的网络资源。QoS 操作 434 还包括实现综合服务（Integrated Service）或区分服务（Differentiated Service）或同时实现这两种服务。关于区分服务，可以在入口交换模块或出口交换模块中执行各种操作，这些操作包括但不限于优先权标记、数据包监控、排队和调度。区分服务（DiffServ）的实现还可以包括识别与不同的流相关的服务分类和将这些流分配到专用于发送一个服务分类的业务的多条隧道中的一个隧道。当 PDU 15 进入隧道时，不同服务分类的隧道优选地是与用于封装 PDU 的不同 IP 报头相关联的 MPLS 隧道，其中 PDU 可以包括也可以不包括先前存在的 IP 报头。在优选实施例中，入口交换模块和出口交换模块都可以采用 DiffServ 模型中规定的过程来将 PDU 分配给多条 MPLS 隧道中的一个 MPLS 隧道，DiffServ 模型包括在因特网工程任务组（IETF）的请求注释（RFC）2474 和 RFC 2475 中提出的模型，在此 20 引用这两种模型作为参考。

计费 and 记账服务 435 优选地包括步骤：（a）基于 PDU 特性来识别客户业务数据流；（b）识别所提交的服务类型或特征；以及（c）在每客户的基础上和/或在每个流的基础上产生基于这些服务的累积 25 费用。例如，如果多个交换模块之间定时地交换信息，或者多个交换模块 110-113 的计费和记账数据库 306 由 CMM 120 进行同步，则可以由能够访问存储器 232（见图 3）中的本地计费和记账数据库 306 的入口交换模块或出口交换模块来执行费用的识别和跟踪。

MPLS 管理服务 436 优选地包括步骤：（a）采用诸如与 RSVP 相

关的协议、会话初始协议（SIP）或标签分布协议（LDP）之类的面向会话的协议来与相邻的标签交换路由器（LSR）交换 MPLS 绑定信息；（b）确定是否将非 MPLS 数据包转发到 MPLS 域中；（c）确定非 MPLS PDU 是否为转发等效类（FEC）的成员；（d）将 MPLS 标签加在是 FEC 成员的 PDU 上；（e）确定交换模块是否为来自 MPLS 域的 PDU 的倒数第二跳，并在必要时弹出 MPLS 标签；以及（f）将超过通信链路最大字节限制的 MPLS 数据包进行分段。管理 MPLS 的实现所需的数据包括用于确定 PDU 是哪个 FEC 的成员的标准、所要应用的可应用标签和下一跳的地址，由 CMM 集中维护，并分布到多个交换模块以实现入口和出口的执行。

生成树服务 437 一般包括用于产生中断会引起广播风暴的循环所需的生成树的方法。关于网络，企业交换机 100 适用于与网络中其他节点交换网桥协议数据单元（BPDU）。例如，企业交换机 100 使用 BPDU 来选出根桥（root bridge）并确定到该根桥的最短距离。对其他交换机的 BPDU 的响应可能要求交换机 100 的交换模块能够访问一个单独的共享数据库或维护共享生成树数据库 308（图 3）的本地副本，生成树数据库 308 包括相邻网桥和到达这些网桥的端口的列表。在优选实施例中，生成树由 CMM 120 产生，并定期地下载到每个交换模块 110-113。

如果在不同的交换模块或交换模块的端口之间发送 BPDU 和广播数据包会引起广播风暴，则优选实施例中的每个交换模块 110-113 还适用于防止这种发送。同样，每个交换模块 110-113 还实现生成树协议的简易版，用于识别哪个交换模块在其广播域内，并因此能够接收所发送的数据包而不存在广播风暴。在识别广播域内的交换模块后，交换模块一般会复制 BPDU 并将其发送到每一个所识别的交换模块。在优选实施例中，如果出口交换模块在同一个 BPDU 广播域中，则重新产生 BPDU 并将这些 BPDU 发送到其他交换模块的处理可以从入口交换模块委托给出口交换。指示 BPDU 的重新产生应当发生的地点的重要因素是入口交换模块和出口交换模块中缓冲器空间的可用

性。

认证服务 438 包括用于确定哪个 PDU 将获准进入并确定在获准进入的基础上所提供的访问级别的过程。在优选实施例中，每个交换模块都适用于查询访问控制列表 (ACL)，该列表用于确定是要将所接收的 PDU 发送到其目的地地址还是要将其过滤掉。优选实施例中的 ACL 基于一个或多个 PDU 特性来控制访问，这些特性优选地是包括源地址和目的地地址、广播比特、协议类型的层 2 和层 3 的特性，以便防止利用例如 RFC 2402、RFC 2463 和 RFC 1826 中所提出的包括认证报头的因特网控制消息协议 (ICMP) 消息和因特网群组管理协议 (IGMP) 数据包而进行的拒绝服务攻击。同上，每个交换模块 110-113 都可以维护包括用户 MAC 和/或 IP 地址、口令和相关联的访问权限的 ACL 310 (图 3)。

访问控制服务 439 包括认证服务的第二种形式，其基于更高层的特性来控制访问。在优选实施例中，访问控制服务 439 适于基于以下参数来授权或拒绝访问：(a) 协议 ID，从而使交换设备 100 可以阻止例如利用没有认证报头的 ICMP 和 IGMP 数据包而进行的拒绝服务攻击；(b) 协议所使用的端口号，这些协议例如包括文件传输协议 (FTP)、普通文件传输协议 (TFTP)、远程登录 (telnet) 和即时信息；以及 (c) 用于过滤掉诸如对等文件交换之类的不期望的应用的应用报头。在有些实施例中，访问控制服务 439 用于补充认证服务 438，并且可以在已经执行了初始认证服务 438 之后的任意时刻在入口交换模块或出口交换模块中执行。

高层学习服务 440 包括步骤：将在网络接口处学习到的信息报告给与高层操作相关的交换设备 100 中的应用。将通过例如地址解析协议 (ARP) 消息学习到的 MAC 地址报告给保存在本地存储器 232 (图 3) 中的 ARP 表 312，并通过在多个交换模块 110-113 之间或与 CMM 120 交换数据来定期地更新这些 MAC 地址。

警报服务 441 指的是用于将确保引起例如 CMM 120 或网络管理员的注意的条件通报给交换机 100 上的应用的系统范围的检查。例如，

这些条件可以包括从多个端口上接收到具有相同源地址的 PDU 的情形。

5 端口镜像服务 442 指的是用于例如复制在一个端口上接收到的 PDU 并由网络管理员将这些 PDU 发送到指定端口上的业务分析工具

源学习 443 一般是指下述处理：(a) 将 PDU 的源地址与接收 PDU 的入口端口相关联；(b) 识别在不同端口上接收到具有相同源地址的 PDU 的情形；以及 (c) 确定是拒绝在一个端口上接收但先前在另一个端口上学习到的 PDU，还是允许该 PDU 并简单地认为先前学习到的关联已经过期。可以在交换模块 110-113 之间定期地进行交换由每个交换模块 110-113 汇总的源学习表 314 (图 3)，以为每个模块提供对源学习端口关联的访问，而不管是在入口处还是在出口处执行源学习服务 443。

15 服务分类 (CoS) 444 操作适于进行步骤：(a) 在入口交换模块或出口交换模块中，基于一个或多个标准，例如包括到达 PDU 的 IEEE 802.1p 优先权值的标准，确定在另一个 PDU 更优选时是否将一个 PDU 过滤掉；以及 (b) 在入口交换模块或出口交换模块中将一个或多个流的 PDU 进行排队，并且随后根据流的服务分类需要采用调度器来释放应用级的流的 PDU。用于识别和规定 CoS 的分类标准统称为分类规则。在优选实施例中，CMM 维护全面的数据库，其包括分类规则

20 和定期地或在需要时散布到多个交换模块 110-113 的可应用分类数据。

25 色彩标记 445 服务用于：(a) 在入口交换模块或出口交换模块中确定是让 PDU 通过、过滤 PDU 还是用先前由网络中的上游节点所应用的三色标记为 PDU 重新着色；(b) 在入口交换模块或出口交换模块中，按照需要实现令牌桶算法 (token bucket algorithm) 用于将三色标记附加到 PDU 上以帮助下游节点选择性地过滤数据包。色彩标记 445 服务包括当前收录于由 Osama Aboul Magd 所著的 IETF 出版物中的双速率三色标记 (trTCM)，并且包括收录于 Juha Heinanen

草拟的 IETF 出版物中的单速率三色标记 (srTCM)，在此引用这两者作为参考。

图 5 中示出的是用于在交换模块间发送包数据的 SDPP 标记的示意图。在优选实施例中，入口交换模块 200 的 SDPP 标记发生器 412 将 SDPP 标记 510 附加到 PDU 500 或 PDU 描述符上，出口交换模块的 SDPP 标记读取器 422 读取 SDPP 标记 510。优选实施例中的 SDPP 标记 510 包括 SDPP 标识符 (ID) 512 和操作码 502。

如上所述，将 SDPP ID 512 作为命令用于指示接收 PDU 的出口交换模块为入口交换模块执行一个或多个转发操作。换句话说，在优选实施例中 SDPP 标记提供一种信令机制，通过这种机制，入口处理器将一个或多个 PDU 转发操作串行分布到出口处理器，从而减小了入口处理器所承受的处理负载。在优选实施例中，在入口处从 SDPP 分配表 238 中检索 SDPP ID 512，但其也可以根据一个或多个处理模块 110-113 中的处理负载动态地确定 SDPP ID 512。

在有些实施例中，SDPP ID 512 包括流 ID 504 和源处理器 ID 506。流 ID 504 唯一地定义一串具有相同 SDPP 服务处理需求的一个或多个 PDU，而源处理器 ID 506 表示接收 PDU 的入口交换模块 200 的 NP 106。交换机 100 可以利用流 ID 504 和源处理器 ID 506 一起唯一地定义将应用于相关联的 PDU 500 或为相关联的 PDU 500 而执行的特定 SDPP 服务。

在有些实施例中，SDPP 标记 510 还包括操作码，即操作码 502。在此使用的术语“操作码”指的是使得将出口交换模块配置为执行由与该操作码相关联的一个或多个 PDU 流所指定的转发操作。接收到操作码之后，出口 NP 将一个或多个计算机可读指令装载到 NP 的芯片内 (on-chip) 缓存 (未示出) 或其他本地存储器 232。由于大多数 NP 的片载缓存太小，不能保存处理交换机 100 可见的每个可能的流所需的计算机可读指令，所以操作码 502 只用来装载为发送到特定出口 NP 的有限数目的流而执行 SDPP 所需要的那些程序指令，这些流是交换机 100 所支持的流的子集。同样，可以将不同的指令集上传到

不同交换模块 110-113 的 NP 以使每个特定 NP 106 的转发操作最优化。仅通过发布新的操作码，还可以随着业务变化随意地更新单个 NP 106 的指令集。

5 作为示例，操作码可以由入口交换模块给出，指示 NP 106 对根据 MPLS 协议处理一个或多个流所必需的所有可执行的代码或算法进行缓存。一旦对与特定操作码 502 相关联的这些可执行代码进行了缓存，具有指定 MPLS 处理操作的流 ID 404 的每个随后的 PDU 就可以遵从相同的处理规则，直到接收到新的操作码为止。

10 在有些实施例中，交换机 100 的交换模块 110-113 适用于共同地处理成批的数据包，即与一个或多个 PDU 流相关联的多个数据包。为执行批处理，以多个相关数据包中的第一个数据包来发送包括操作码 502 和 SDPP ID 512 的 SDPP 标记。之后，应用于相关数据包的 SDPP 标记 510 只需要包括 SDPP ID 512。然后，出口交换模块就使多个 PDU 中的每一个都服从由其 SDPP ID 510 指定的处理规则和在多个 PDU  
15 中的第一个 PDU 中指定的计算机可读指令。以这种方式，免除了以多个 PDU 中的每一个 PDU 发送操作码 502 的需要，并且减少了附加、发送和读取操作码 502 所需的资源。

20 在备选的实施例中，使用带外通信信道（未示出）将 SDPP ID 512 和/或操作码 502 从入口交换模块发送到出口操作模块。带外通信信道对应于例如总线之类的信令信道，不同于用来在交换模块间发送 PDU 的数据信道。

图 6 中示出的是入口交换模块处理入口流的方法的流程图。接收到入口 PDU 之后（步骤 610），优选实施例中的入口交换模块对 PDU 进行分类以识别其将发送到的出口交换模块。获悉出口交换模块后，  
25 入口交换模块就可以确定（步骤 640）入口交换模块和出口交换模块上的相对需要以及负载平衡的需要。在有些实施例中，入口交换模块所承受的 PDU 处理负载（步骤 620）和出口交换模块所承受的 PDU 处理负载（步骤 630）可以由 CMM 120 从模块自身的拥塞状态中得出。如果入口交换模块的拥塞状态相对于出口交换模块是过度使用

的，则肯定地应答负载平衡变化查询（测试步骤 650），并且改变（步骤 660）在入口和出口之间执行的 SDPP 服务的分布以减少负载不平衡。例如，如果入口交换模块相对超载，则与不平衡的程度成正比地提高整个 SDPP 服务中分配给出口交换模块的百分比。在入口交换模块严重超载的情况下，可以将全部 SDPP 服务都分配到出口交换模块。如果入口交换模块和出口交换模块所承受的负载基本上相等，并且差异在预定的负载差异门限内，则否定地应答负载平衡变化查询，并且维护先前存在的 SDPP 服务分布。

在已经确定了 SDPP 服务的分配之后，入口交换模块执行（步骤 670）为入口指定的 SDPP 服务的子集。在 SDPP 标记的一个或多个字段中指定出口处待执行的其余 SDPP 服务，SDPP 标记是在将 PDU 发送到（步骤 690）出口交换模块之前附加到（步骤 680）PDU 上的。例如，如果出口交换模块由于其他业务条件而严重拥塞，则入口交换模块可以执行 PDU 要求的全部 SDPP 服务，并且附加到数据包上的 SDPP ID 为空值，表明出口交换模块中没有进行任何 SDPP 处理。

图 7 中示出的是出口交换模块处理出口流的方法的流程图。在从交换架构 150 处接收到（步骤 710）PDU 之后，该模块检查 PDU 以确定 SDPP 标记 510 是否存在，并且读取其中包含的字段。如果 SDPP 标记 510 存在，则肯定地应答 SDPP 标记查询（步骤 720），并且为了检查而除去（步骤 730）SDPP 标记。如果 SDPP 标记包括操作码 502，则肯定地应答操作码查询（步骤 740），并且将由操作码表示的计算机可读指令载入（步骤 750）出口网络处理器中的缓存。如果 SDPP 标记中没有操作码，则否定地应答操作码查询，并且 NP 106 根据由以前的同一批 PDU 的操作码所指定的指令集来执行由 SDPP ID 512 所标识的第二组 SDPP 服务。然后，从出口交换模块向 PDU 的目的地节点的方向发送（步骤 770）根据 SDPP 服务需求而处理的 PDU。本领域的普通技术人员应当意识到，在出口网络处理器中处理的 PDU 经历了相同的数据包处理，并且表现为与单独地在入口交换模块或出口交换模块中处理的 PDU 基本上相同。

尽管上述描述包含了许多特定内容，但这些特定内容不应当解释为对本发明范围的限制，而应解释为为本发明的一些优选实施例提供说明。

因此，以上已经以示例的方式并且非限制性地公开了本发明，并且应当参考以下权利要求以确定本发明的范围。

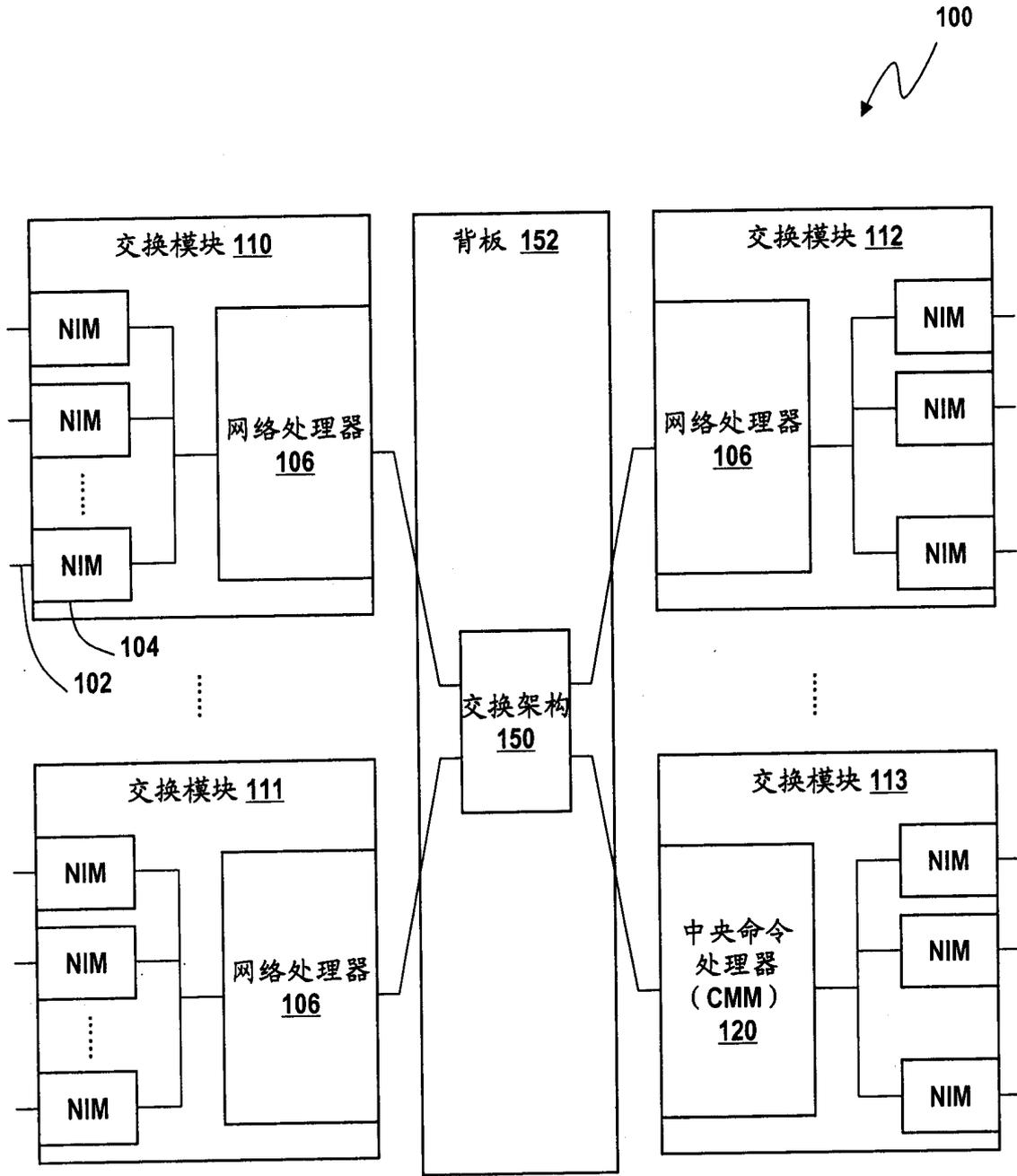


图1

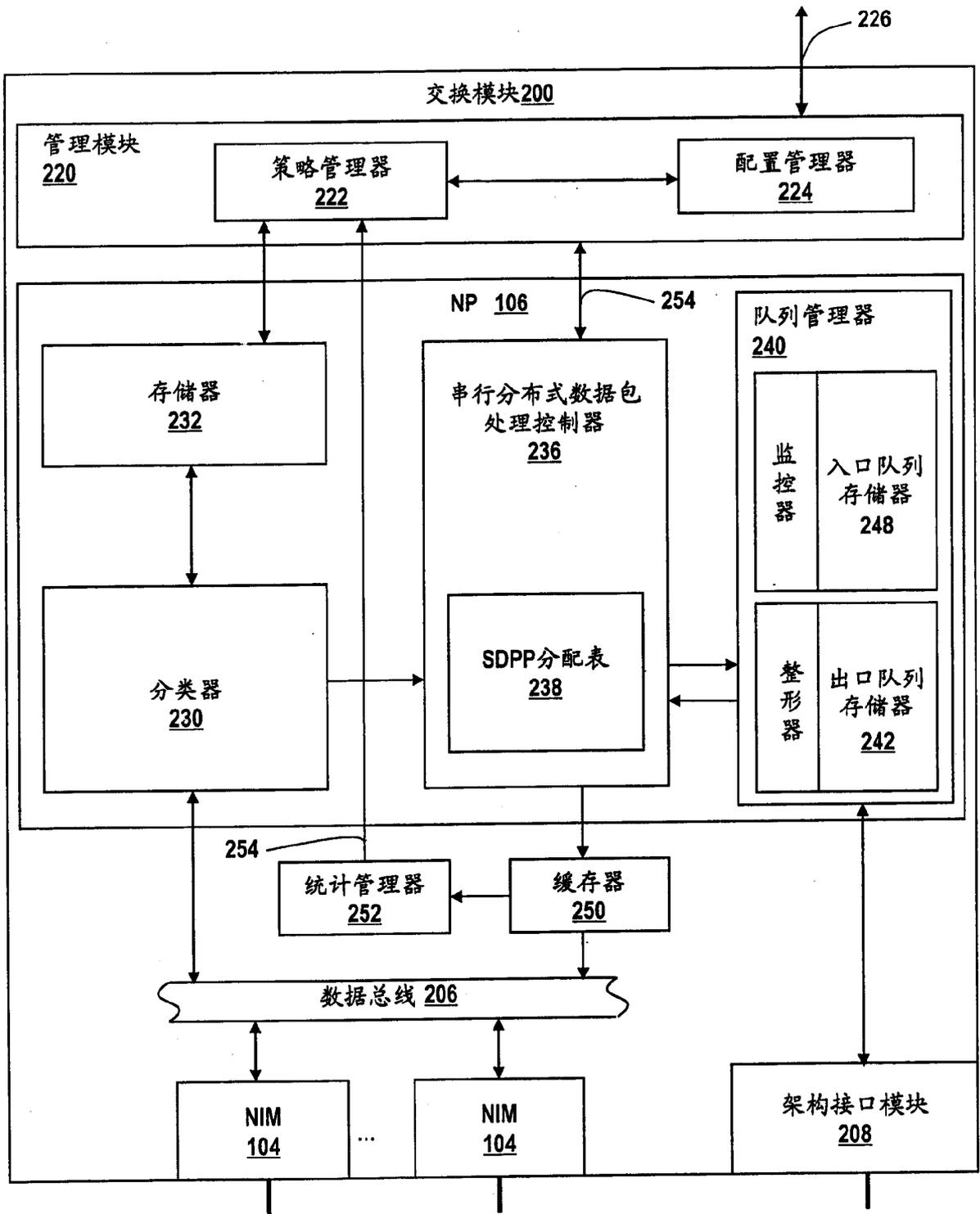


图2

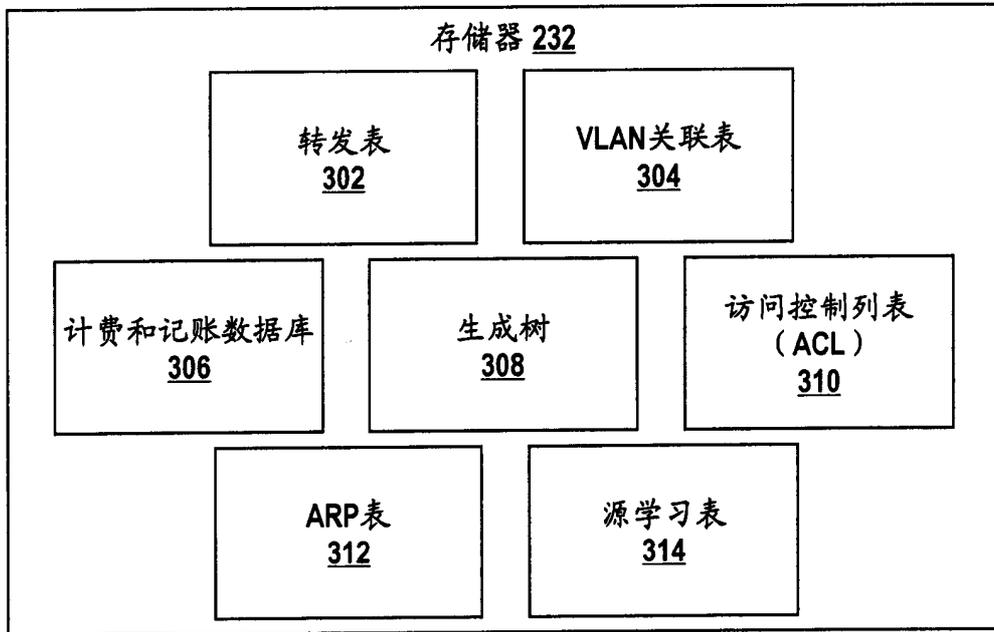


图 3

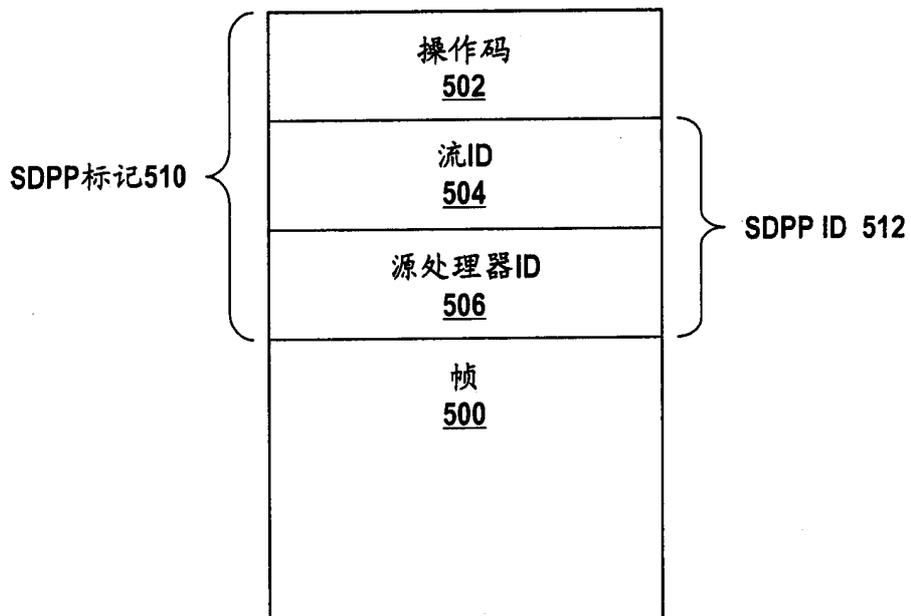


图 5

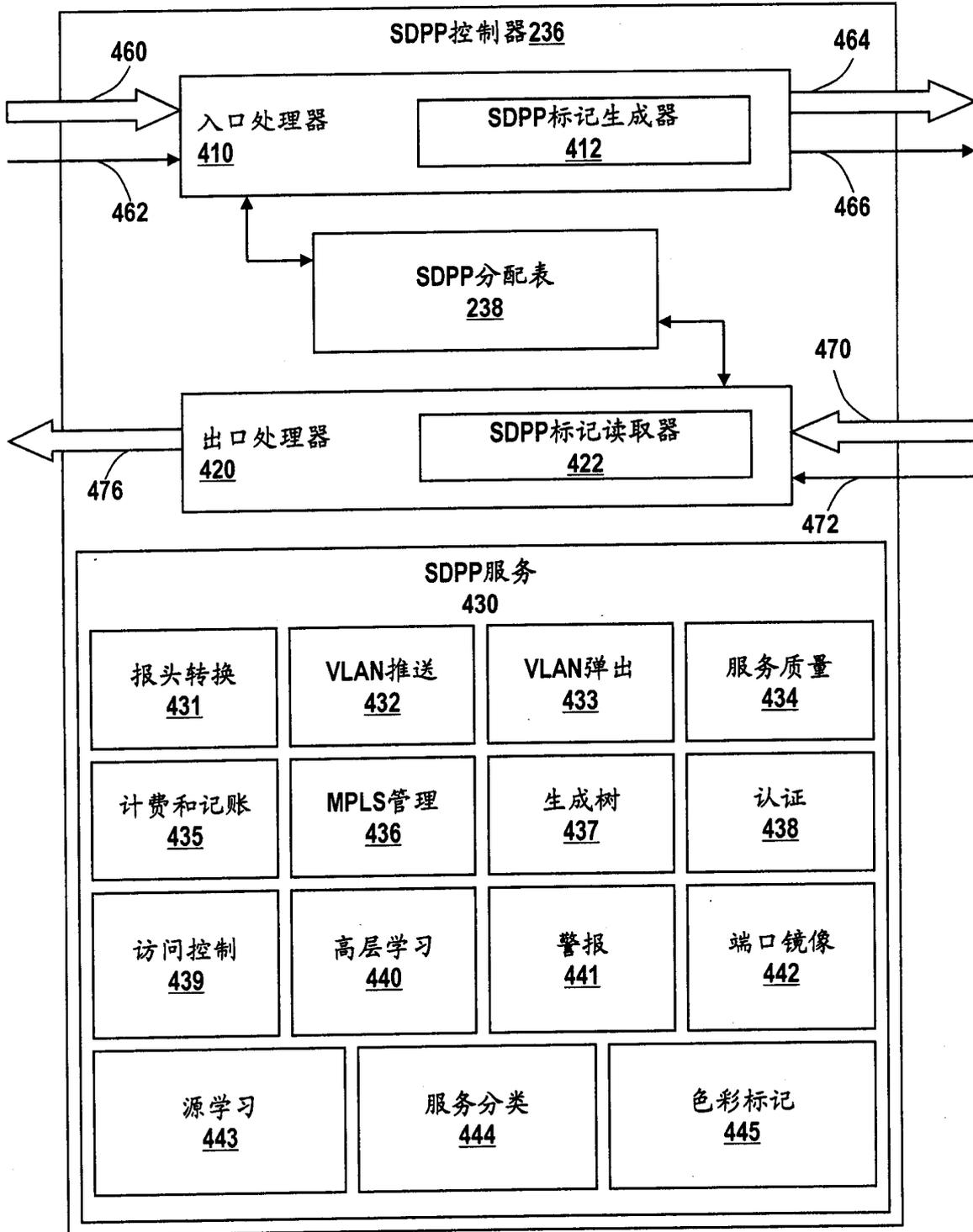


图 4

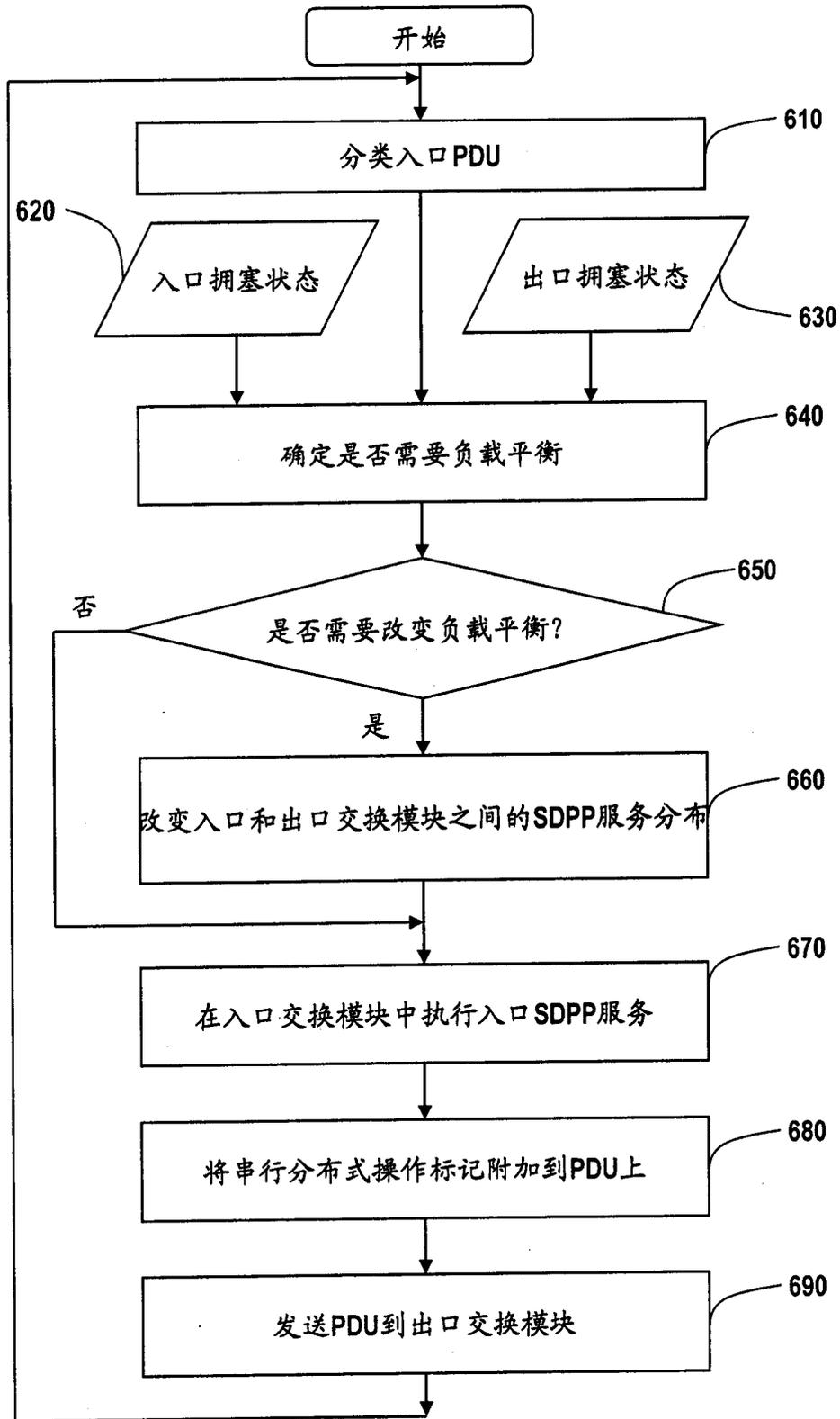


图6

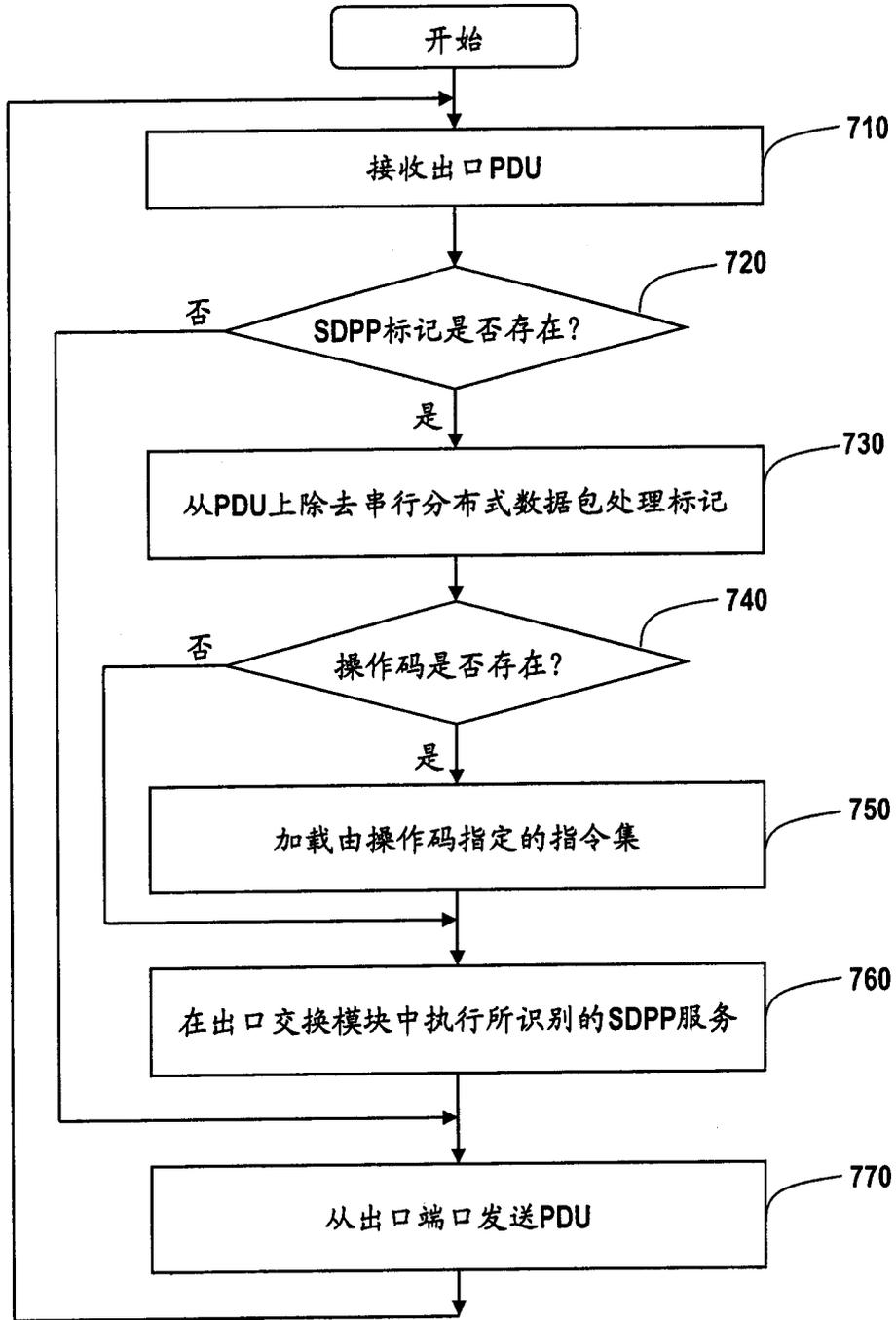


图7