3. DSCP-BASED PFC

In this section we examine the issues faced by the original VLAN-based PFC and present our DSCP-based PFC solution. VLAN-based PFC carries packet priority in the VLAN tag, which also contains VLAN ID. The coupling of packet priority and VLAN ID created two serious problems in our deployment, leading us to develop a DSCP-based PFC solution. Figure 3(a) shows the packet formats of the PFC pause frame and data packets in the original VLANbased PFC. The pause frame is a layer-2 frame, and does not have a VLAN tag. The VLAN tag for the data packet has four parts: TPID which is fixed to 0x8100, DEI (Drop Eligible Indicator), PCP (Priority Code Point) which is used to carry packet priority, and VID (VLAN identifier) which carries the VLAN ID of the packet. For our purpose, although we need only PCP, VID and PCP cannot be separated. Thus, to support PFC, we have to configure VLAN at both the server and the switch side. In order for the switch ports to support VLAN, we need to put the server facing switch ports into trunk mode (which supports VLAN tagged packets) instead of access mode (which sends and receives untagged packets). The basic PFC functionality works with this configuration, but it leads to two problems.

在本节中，我们将研究原始基于 VLAN 的 PFC 所面临的问题，并介绍我们基于 DSCP 的 PFC 解决方案。基于 VLAN 的 PFC 在 VLAN 标记中携带数据包优先级，其中还包含 VLAN ID。数据包优先级和 VLAN ID 的耦合在我们的部署中造成了两个严重问题，导致我们开发了基于 DSCP 的 PFC 解决方案。

图3(a)显示了原始基于VLAN的PFC中PFC暂停帧和数据包的数据包格式。暂停帧是二层帧，没有VLAN标签。

数据包的VLAN标签由四部分组成：TPID固定为0x8100、DEI（Drop Eligible Indicator）、PCP（Priority Code Point）用于承载数据包优先级、VID（VLAN Identifier）承载VLAN ID数据包的。

就我们的目的而言，虽然我们只需要 PCP，但 VID 和 PCP 是不能分开的。因此，为了支持PFC，我们必须在服务器端和交换机端都配置VLAN。

为了使交换机端口支持 VLAN，我们需要将面向服务器的交换机端口设置为 trunk 模式（支持 VLAN 标记数据包），而不是 access 模式（发送和接收未标记数据包）。基本 PFC 功能适用于此配置，但会导致两个问题。

First, the switch trunk mode has an undesirable interaction with our OS provisioning service. OS provisioning is a fundamental service which needs to run when the server OS needs to be installed or upgraded, and when the servers need to be provisioned or repaired. For data centers at our scale, OS provisioning has to be done automatically. We use PXE (Preboot eXecution Environment) boot to install OS from the network. When a server goes through PXE boot, its NIC does not have VLAN configuration and as a result cannot send or receive packets with VLAN tags. But since the server facing switch ports are configured with trunk mode, these ports can only send packets with VLAN tag. Hence the PXE boot communication between the server and the OS provisioning service is broken. We tried several “hacks” to fix this problem, including letting the switches change the switch port configuration based on the guessed state of the servers, and letting the NICs accept all the packets with or without VLAN tag. However, all these proved to be complex and unreliable, and needless to say, non-standard.

首先，交换机中继模式与我们的操作系统配置服务存在不良交互。操作系统配置是一项基础服务，在需要安装或升级服务器操作系统、需要配置或修复服务器时需要运行该服务。
对于我们规模的数据中心，操作系统配置必须自动完成。
我们使用 PXE（预启动执行环境）引导从网络安装操作系统。
当服务器进行PXE启动时，其网卡没有VLAN配置，因此无法发送或接收带有VLAN标签的数据包。但由于面向服务器的交换机端口配置为 trunk 模式，因此这些端口只能发送带有 VLAN 标记的数据包。(冲突点)
因此，服务器和操作系统配置服务之间的 PXE 启动通信中断。我们尝试了几种“黑客”来解决这个问题，包括让交换机根据猜测的服务器状态更改交换机端口配置，以及让网卡接受所有带或不带 VLAN 标记的数据包。然而，所有这些都被证明是复杂且不可靠的，而且不用说，也是不标准的。

Second, we have moved away from a layer-2 VLAN, and all our switches including the ToR switches are running layer-3 IP forwarding instead of MAC-based layer-2 bridging. A layer-3 network has the benefits of scalability, better management and monitoring, better safety, all public and standard instead of proprietary protocols. However, in a layer-3 network, there is no standard way to preserve the VLAN PCP value when crossing subnet boundaries.

其次，我们已经放弃了第 2 层 VLAN，并且包括 ToR 交换机在内的所有交换机都运行第 3 层 IP 转发，而不是基于 MAC 的第 2 层桥接。第三层网络具有可扩展性、更好的管理和监控、更好的安全性、所有公共和标准而不是专有协议的优点。然而，在第 3 层网络中，没有标准方法可以在跨越子网边界时保留 VLAN PCP 值。

In both problems, the fundamental issue is that VLANbased PFC unnecessarily couples packet priority and the VLAN ID. We broke this coupling by introducing DSCP-based PFC. Our key observation is that the PFC pause frames do not have a VLAN tag at all. The VLAN tag in data packets is used only for carrying the data packet priority. In the IP world, there is a standard and better way to carry packet priority information, which is the DSCP field in the IP header.

在这两个问题中，根本问题是基于 VLAN 的 PFC 不必要地将数据包优先级和 VLAN ID 结合起来。我们通过引入基于 DSCP 的 PFC 打破了这种耦合。我们的主要观察结果是 PFC 暂停帧根本没有 VLAN 标记。数据包中的VLAN标签仅用于承载数据包的优先级。在IP世界中，有一种标准且更好的方式来承载数据包优先级信息，这就是IP报头中的DSCP字段。

The solution, as shown in Figure 3(b), is to move the packet priority from the VLAN tag into DSCP. As we can see, the change is small and only touches the data packet format. The PFC pause frame format stays the same. With DSCP-based PFC, data packets no longer need to carry the

> VLAN tag, which solves both of the problems mentioned earlier. The server facing ports no longer need to be in trunk mode, which means that PXE boot works without any issues. Also, the packet priority information, in form of DSCP value, is correctly propagated by IP routing across subnets.

> 解决方案如图 3(b) 所示，将数据包优先级从 VLAN 标记移至 DSCP。我们可以看到，变化很小，只涉及数据包格式。 PFC 暂停帧格式保持不变。基于DSCP的PFC，**数据报文不再需要携带VLAN Tag**，解决了前面提到的两个问题。**面向端口的服务器不再需要处于中继模式，这意味着 PXE 引导可以正常工作**。此外，**DSCP 值形式的数据包优先级信息可以通过 IP 路由在子网中正确传播**。

> Of course, DSCP-based PFC does not work for the designs that need to stay in layer-2, e.g., Fibre Channel over Ethernet (FCoE). This is not a problem for us since we do not have any layer-2 networks in our data centers.

> 当然，基于 DSCP 的 PFC 不适用于需要保留在第 2 层的设计，例如以太网光纤通道 (FCoE)。这对我们来说不是问题，因为我们的数据中心没有任何第 2 层网络。

> DSCP-based PFC requires both NICs and switches to classify and queue packets based on the DSCP value instead of the VLAN tag. In addition, the NIC needs to send out data packets with the right DSCP value. Fortunately, the switch and NIC ASICs are flexible enough to implement this. Internally, at each port, the switch or NIC maintains eight Priority Groups (PGs), with each PG can be configured as lossless or lossy. If a PG i (i ∈ [0, 7]) is configured as lossless, once its ingress buffer occupation exceeds the XOFF threshold, pause frame with priority i will be generated. The mapping between DSCP values and PFC priorities can be flexible and can even be many-to-one. In our implementation, we simply map DSCP value i to PFC priority i.

> 基于 DSCP 的 PFC 要求 NIC 和交换机都根据 DSCP 值（而不是 VLAN 标记）对数据包进行分类和排队。此外，NIC需要发送具有正确DSCP值的数据包。
> 幸运的是，交换机和 NIC ASIC 足够灵活来实现这一点。
> 在内部，交换机或 NIC 在每个端口维护八个优先级组 (PG)，每个 PG 都可以配置为无损或有损。
> 如果PG i (i ∈ [0, 7])配置为无损，一旦其入口缓冲区占用超过XOFF阈值，就会生成优先级为i的暂停帧。 [i -> i 哪个组超出阈值，哪个组就被对应暂停]
> DSCP值和PFC优先级之间的映射可以是灵活的，甚至可以是多对一的。在我们的实现中，我们简单地将 DSCP 值 i 映射到 PFC 优先级 i。

> Our DSCP-based PFC specification is publicly available, and is supported by all major vendors (Arista Networks, Broadcom, Cisco, Dell, Intel, Juniper, Mellanox, etc.). We believe DSCP-based PFC provides a simpler and more scalable solution than the original VLANbased design for IP networks.

> 我们基于 DSCP 的 PFC 规范已公开发布，并受到所有主要供应商（Arista Networks、Broadcom、Cisco、Dell、Intel、Juniper、Mellanox 等）的支持。我们相信基于 DSCP 的 PFC 为 IP 网络提供了比原始基于 VLAN 的设计更简单且更具可扩展性的解决方案。

TPID

DEI (Drop Eligible Indicator)

PCP (Priority Code Point) which is used to carry packet priority

VID (VLAN identifier) which carries the VLAN ID of the packet

PXE (Preboot eXecution Environment)（预启动执行环境）

ToR switches (ToR 交换机)

layer-3 IP forwarding (三层IP转发)

MAC-based layer-2 bridging. (基于 MAC 的第 2 层桥接)

“there is no standard way to preserve the VLAN PCP value when crossing subnet boundaries.” 在跨越子网边界时，没有标准方法来保留 VLAN PCP 值

“DSCP-based PFC does not work for the designs that need to stay in layer-2” 基于 DSCP 的 PFC 不适用于需要留在第 2 层的设计