RDMA over Commodity Ethernet at Scale(Conclusion)

Conclusion

> In this paper, we have presented our practices and experiences in deploying RoCEv2 safely at large-scale in Microsoft data centers. Our practices include the introducing of DSCP-based PFC which scales RoCEv2 from layer-2 VLAN to layer-3 IP and the step-by-step onboarding and deployment procedure. Our experiences include the discoveries and resolutions of the RDMA transport livelock, the RDMA deadlock, the NIC PFC storm and the slow-receiver symptom. With the RDMA management and monitoring in place, some of our highlyreliable, latency-sensitive services have been running RDMA for over one and half years

> 在本文中，我们介绍了在 Microsoft 数据中心大规模安全部署 RoCEv2 的实践和经验。我们的实践包括引入基于 DSCP 的 PFC，将 RoCEv2 从第 2 层 VLAN 扩展到第 3 层 IP，以及逐步的加入和部署过程。我们的经验包括发现并解决 RDMA 传输活锁、RDMA 死锁、NIC PFC 风暴和缓慢接收器症状。随着 RDMA 管理和监控到位，我们的一些高度可靠、延迟敏感的服务已经运行 RDMA 超过一年半了

Layer-2 VLAN & Layer-3 IP

背景引入

RDMA 允许用户态的应用程序直接读取和写入远程内存，避免了数据拷贝和上下文切换；并将网络协议栈从软件实现 offload 到网卡硬件，实现了高吞吐量、超低时延和低 CPU 开销的效果。

当前 RDMA 在以太网上的传输协议是 RoCEv2，RoCEv2 是基于无连接协议的 UDP 协议，相比面向连接的 TCP 协议，UDP 协议更加快速、占用 CPU 资源更少，但其传输是不可靠的，*一旦出现丢包会导致 RDMA 的传输效率降低*，这是由 RDMA 的 Go-back-N 重传机制决定的。RDMA 接收方网卡发现丢包时，会丢弃后续接收到的数据包，发送方需要重发之后的所有数据包，这导致性能大幅下降。所以**要想 RDMA 发挥出其性能，需要为其搭建一套不丢包的无损网络环境**

差异化流量分类

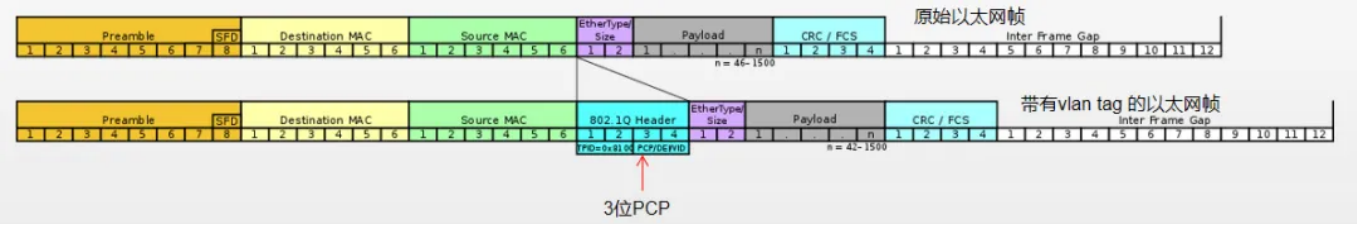
构建无损网络，首先需要对网络流量进行分类，然后针对不同类别流量采用具体流控策略，实现精确控制，避免相互影响。

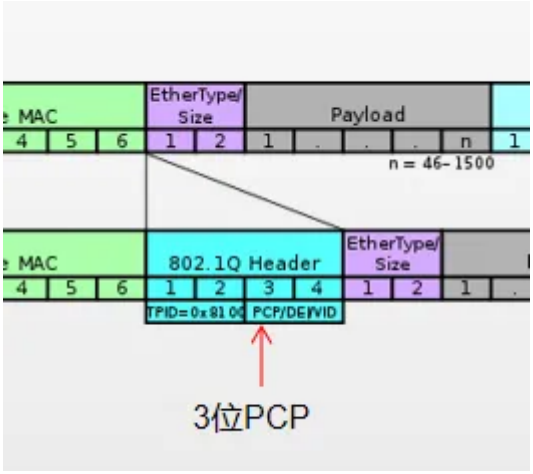
流量分类有两种不同的分类方法：传输层（Layer 2）和网络层（Layer 3）。

【1】Layer2 通过 vlan header（802.1q）里的 PCP（802.1p）位进行分类，对应 CoS（Class of Service）
【2】Layer3 通过 IP header 里的 DSCP 进行分类，对应 DSCP

Layer 2 流量分类

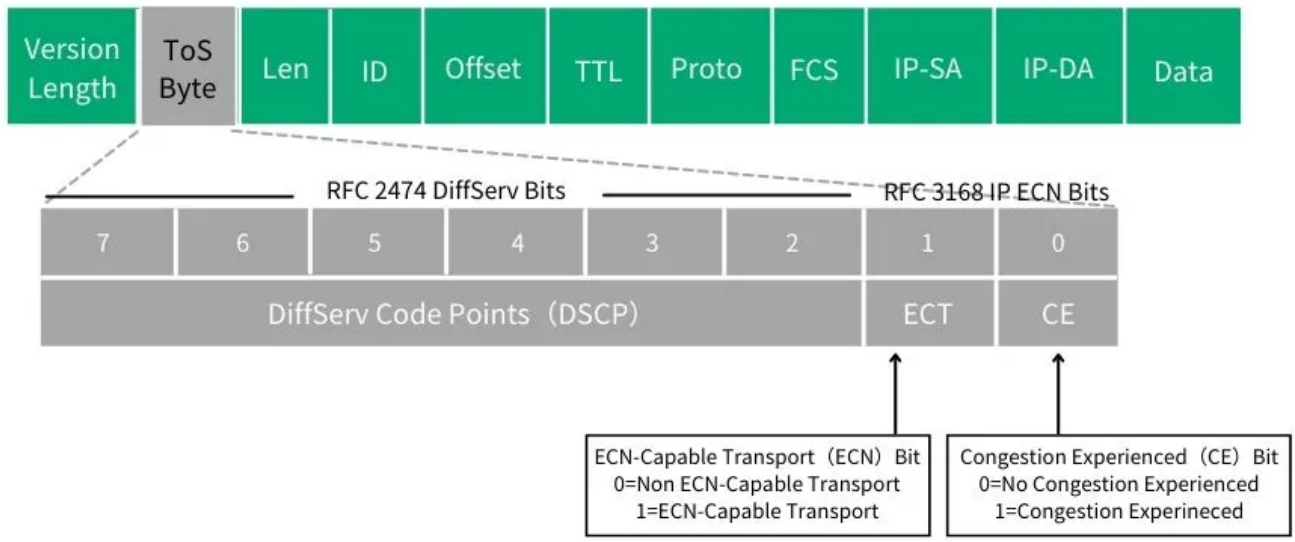
Layer2 层流量分类依据的是 vlan tag 中的 3 位 PCP bit，总共有 8 个类别。3 个 bit 是 Header 中第 3 个 byte 的前三位，如下图。在使用 Layer 2 流量分类时，主机端发出的包需要带有 vlan tag。因此要对网卡配置 vlan，并且设置优先级。因为 L2 层 PFC 需要依靠 vlan，因此包经过三层交换机时可能存在 tag 失效等问题。





Layer 3 流量分类

Layer3 使用 IP 包头中的 TOS 前 6 位（DSCP），支持 64 种不同的流量分类方式，TOS 的后两位用作 Explicit Congestion Notification（ECN）Field，ECN 是一种端到端的流控方式，后面会有介绍。



选择 Layer2 还是 Layer3 层流控

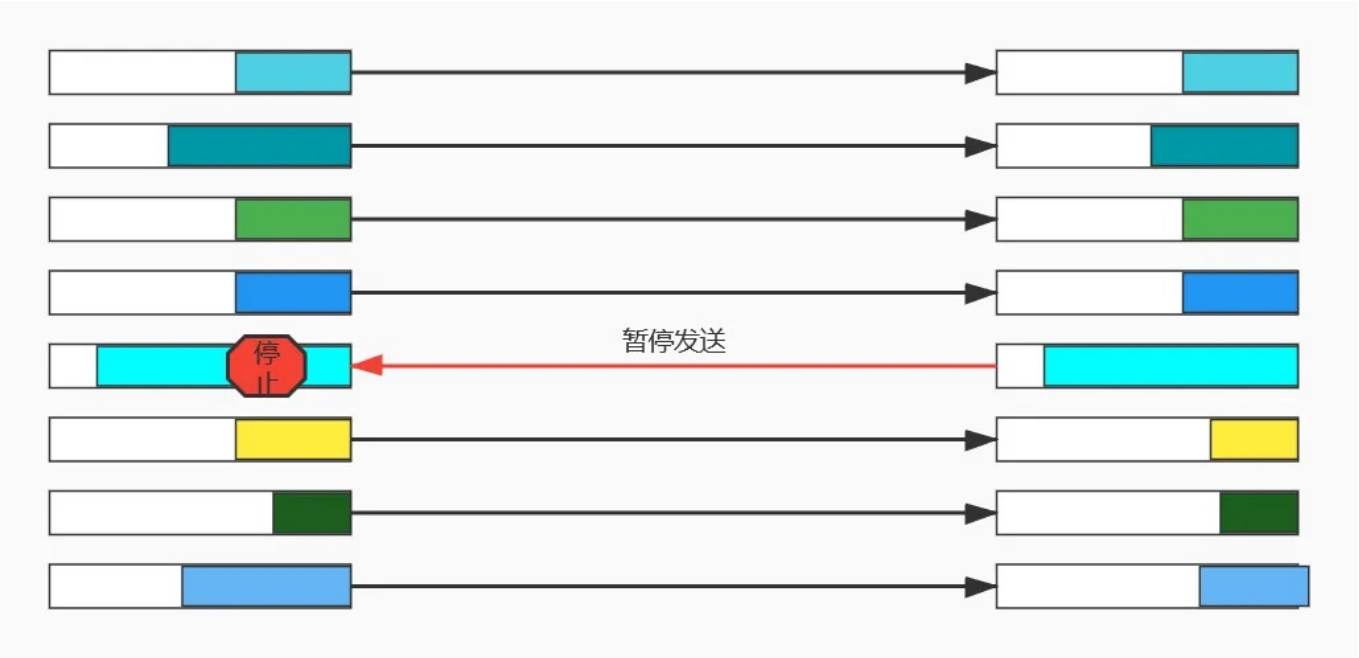
在交换机支持 DSCP 的条件下，建议使用 L3 层的流控方式。从前面的介绍可以看到，**L3 层的控制方式可跨多层交换机，DSCP 值在端到端的传输过程中不会发生变化**。RoCE 使用 UDP 报文进行数据传输，建议 RoCE 的流控使用基于 DSCP 的方式。

构建无损网络—基于 DSCP 或 PCP 的 PFC 流控机制

IEEE 802.1Qbb（Priority-based Flow Control，基于优先级的流量控制）简称 PFC，是 IEEE 数据中心桥接（Data Center Bridge）协议族中的一个技术，是流量控制的增强版。

我们先看一下 IEEE 802.3X（Flow Control）流控的机制：当接收者没有能力处理接收到的报文时，为了防止报文被丢弃，接收者需要通知报文的发送者暂时停止发送。IEEE 802.3X 协议存在一个**缺点：一旦链路被暂停，发送方就不能再发送任何数据包，如果是因为某些优先级较低的数据流引发的暂停，结果却让该链路上其他更高优先级的数据流也一起被暂停了，这是得不偿失的。**

PFC 在基础流控 IEEE 802.3X 基础上进行扩展，【1】允许在一条以太网链路上创建 8 个虚拟通道，并为每条虚拟通道指定相应优先级，【2】允许单独暂停和重启其中任意一条虚拟通道，同时允许其它虚拟通道的流量无中断通过。【3】PFC 将流控的粒度从物理端口细化到8个虚拟通道，分别对应 Smart NIC 硬件上的 8 个硬件发送队列，如下图。



对比二层与三层流控

在二层网络中，PFC 使用 vlan 中的 PCP 位来对数据流进行区分；在三层网络中，PFC 既可以使用 PCP，也可以使用 DSCP，使得不同数据流可以享受到独立的流控制。

当下数据中心因多采用三层网络，且 DSCP 值在端到端的传输过程中不会发生变化，故推荐使用 DSCP。

RDMA 无损网络中利用 PFC 流控机制，实现了交换机端口缓存溢出前暂停对端流量，阻止了丢包现象发生，但因为需要一级一级反压，效率较低，而且存在不公平问题和 Head-of-Line 堵塞问题。此外，PFC 是通过下游网络设备对上游设备的控制方式达到不丢包的目的，但最有效的流控应该是控制产生数据的源端主机的发送速度，使得主机往网络中注入数据速度放缓，这是解决问题的根本方法。