基于可编程数据平面的PFC算法实现

耿俊杰,颜金尧
(中国传媒大学 信息与通信工程学院,北京 100024)

摘 要:当前高吞吐量、超低延迟的高性能无损数据中心网络成为研究的热点。传统TCP/IP协议是为广域网设计的,在高速网络条件下(特别是随着10 Gb/s的网络接口的普及)会存在I/O瓶颈问题;远程直接数据存取技术RDMA(remote direct memory access)是为了解决网络传输中终端主机的数据处理延迟、降低CPU负载而产生的,最早应用在高性能计算领域。RDMA技术基于优先级流量控制PFC算法实现了无损传输网络。首先介绍了可编程数据平面技术和高性能数据中心网络的研究现状,并基于可编程数据平面以软件定义的方式实现了PFC算法,进而实现了可编程的无损数据中心网络,并在仿真网络环境下对实现的PFC算法进行了性能测试。实验结果显示在可编程数据平面下实现的PFC算法达到了无损传输的目标。同时证明了可编程数据平面技术在高性能数据中心网络的实现中可以发挥巨大作用,相对于传统的网络架构,可编程数据平面技术由于采用了软件定义的方式,因此更加灵活、高效。

关键词:可编程数据平面;数据中心;无损传输;高性能网络;流量控制
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2021)01-0116-06
doi:10.3969/j.issn.1673-629X.2021.01.021

Implementation of Priority Flow Control Algorithm Based on Programmable Data Plane

GENG Jun-jie, YAN Jin-yao
(School of Information and Communication Engineering, Communication University of China, Beijing 100024, China)

Abstract: At present, the lossless data center network with high throughput, ultra-low delay and high performance has become a research hotspot. Traditional TCP/IP protocol is designed for wide area networks. TCP/IP stacks will suffer I/O bottlenecks under high-speed network conditions (especially with the popularity of 10 Gb/s network interfaces). RDMA (remote direct memory access) is developed to solve the data processing delay of the terminal host in network transmission and reduce the CPU load, which was first applied in the field of high performance computing. RDMA realizes lossless transmission network based on PFC (priority flow control). We first introduce the research status of programmable data plane technology and high-performance data center network and realize the PFC algorithm based on programmable data plane in the way of software definition, and then realize the programmable lossless data center network and test the performance of the realized PFC algorithm under the simulation network environment. The experiment shows that the PFC framework realized under the programmable data plane achieves the goal of lossless transmission. At the same time, it is proved that the programmable data plane technology can play a huge role in the realization of high-performance data center network. Compared with the traditional network architecture, the programmable data plane technology is more flexible and efficient since adoption of the way of software definition.

Key words: programmable data plane; data center; lossless transmission; high performance network; flow control

0 引 言

当前互联网中存在大量的在线业务需要网络对高频率的用户请求做出快速应答,对数据中心提出了超低时延的要求;随着近年来机器学习和人工智能技术的高速发展,对于数据中心计算能力的需求大幅上升,因此数据中心部署了大量的分布式计算集群[1]以满足日益复杂的神经网络和深度学习模型,大量的分布式计算集群采用的并行程序会导致通讯延迟,进而严重影响数据中心的计算效率;同时,随着近年来数据中心流量的激增[2],数据中心往往会利用以太网融合组网的分布式存储技术来解决数据存储和读取效率问题。而大象流占据分布式存储网络流量的主要比例,

收稿日期:2020-01-01 **修回日期**:2020-05-07
基金项目:国家重点研发计划(2019YFB1804300);国家自然科学基金面上项目(61971382)
作者简介:耿俊杰(1987-),男,博士研究生,CCF会员(A7065G),研究方向为计算机网络架构;颜金尧,教授,博导,研究方向为宽带信息网络。

当分布式存储网络中一旦因拥塞发生数据包丢失导致大象流重传,会严重影响数据中心效率并且会加重拥塞程度,进一步影响网络性能,因此高吞吐量、超低延迟的高性能无损数据中心网络成为现在研究的热点[3-5]。RDMA(remote direct memory access)技术[6]是目前实现超低时延、高吞吐量的高性能数据中心网络最常用的技术,RoCE(RDMA over converged Ethernet)协议[7]因为具备明显的性能和成本优势,目前在融合以太网数据中心中占据主流市场地位。而 RoCEv2 基于 PFC(priority flow control)[8]算法实现了无损传输。

另一方面,传统网络架构封闭、固化,网络协议开发部署周期长,很难适应当前对于网络创新的要求。在此背景下,为了应对当前网络面临的挑战,开放网络可编程能力,扩大网络创新的空间,2008 年,斯坦福大学 Nick McKeown 教授为首的研究团队提出了 OpenFlow 以及软件定义网络 SDN(software defined networking)技术[9-10]。软件定义网络技术获得了学术界和工业界的高度关注,2014 年,研究者在 SDN 基础上又提出了可编程数据平面技术[11-12],将网络编程能力扩展到数据平面。在网络领域的各种创新和发展,对应对当前数据中心面临的各种问题和挑战提供了新的思路和方法。该文基于可编程数据平面技术实现了 PFC 算法,进而实现了无损数据中心网络,为实现高性能数据中心网络提供了新的思路和实现方法。

1　可编程数据平面技术

在传统网络架构中,网管系统被作为网络管理平面,而控制平面和数据平面则被分别部署在每个网络设备上。在这种部署方式下,网络管理十分复杂繁琐。另外,除了标准协议外,各个厂家都会有一些私有协议,这样就进一步加大了网络管理的复杂性。同时,传统网络架构下的控制平面和数据平面分布式地部署在网络中的各个设备上,并且控制平面和数据平面是固化、封闭的,实现网络创新需要的新功能部署周期非常长(往往是几年),显然传统网络架构已经完全不能满足当前对于网络创新的需求。在此背景下,软件定义网络技术以及可编程数据平面技术应运而生。

1.1　软件定义网络技术

软件定义网络技术(SDN)是斯坦福大学 Nick McKeown 教授团队在 Clean Slate 项目提出的一个概念,特别是 2009 年 SDN 南向接口协议 OpenFlow1.0 的发布,标志着软件定义网络技术进入高速发展的阶段。

软件定义网络技术将网络设备控制平面和数据平面分离,通过逻辑集中的控制器实现对网络转发设备的集中管理。在软件定义网络架构中,控制平面具有全局化视野,通过南向接口协议实现与数据平面通信,通过开放控制平面的可编程特性,使得网络功能更加灵活和易于扩展,因此对于网络功能的部署更加灵活,同时也简化了网络的管理。软件定义网络技术是当前网络领域最为活跃的技术创新,被 MIT 评为“改变世界的十大创新技术之一”[13]。

如图 1 所示,软件定义网络架构实现了控制平面和数据平面的分离,同时通过北向接口协议开放 API,允许用户编程实现网络功能自定义。同时使用南向接口协议 OpenFlow 实现控制平面与数据平面的通信。

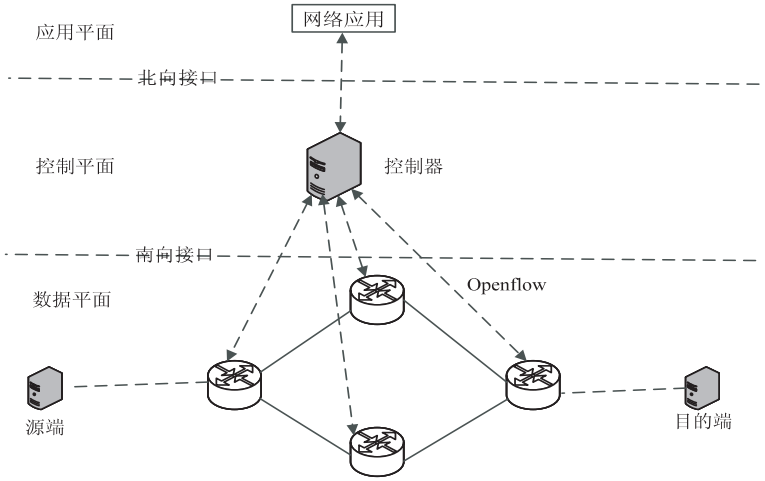

图 1　SDN 网络架构

1.2　可编程数据平面

软件定义网络技术实现了数据平面和控制平面的分离,开放了控制平面编程能力,实现了控制平面的逻辑集中,这些特点为网络的管理和开放带来了一定的灵活性。但是在软件定义网络架构下,数据平面并没有被开放,其行为仍然是固定的。OpenFlow 协议已经从 2008 年 OpenFlow1.0 版本演进到 1.5 版本,其中匹配域中支持的元组数量也从支持 12 元组增加到目前

支持 45 个元组,其支持的元组数量随着 OpenFlow 协议的更新也不断增加。但 OpenFlow 协议支持的匹配域都是协议设定的,并不能支持灵活的弹性增加,对于新匹配元组的支持都需要重新编写数据平面与控制平面两端的协议栈以及数据平面的数据包处理逻辑,这种局限性导致了 OpenFlow 交换机的设计难度大大增加。OpenFlow 协议的版本稳定性也存在很大问题,对软件定义网络技术所追求的网络创新是一种阻碍。

Nick McKeown 教授等人提出了协议无关(programming protocol-independent packet processors)的高级编程语言 P4[11]。P4 是一种声明性的高级编程语言,通过编写 P4 代码可以自定义网络数据平面数据包的处理流程,也就是开放了数据平面的可编程能力。

图 2 描述了可编程数据平面的抽象转发模型,主要由输入、输出端口、解析器、Ingress/Egress 控制流水线、队列缓存组成。解析器负责报文解析,当数据包进入交换机时,解析器按照解析表对进入的数据包进行报文解析,解析表是由 P4 代码定义,并由编译器编译生成,通过 P4 代码自定义报文头和报文头解析顺序,实现数据包的报文解析逻辑,解析器从进入交换机的数据包的报文首部解析出自定义的报文头,并赋值给 P4 定义的实例化首部。

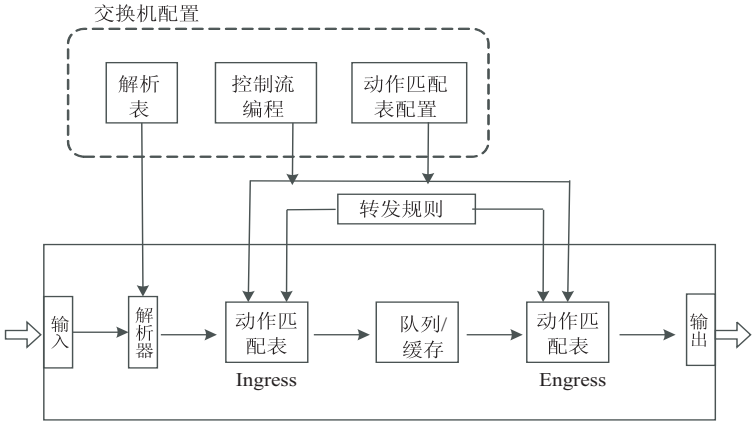

图 2 可编程数据平面抽象转发模型

匹配动作表是数据平面执行转发逻辑的基本单元,对进入交换机并匹配成功的数据包执行相应的动作,实现对数据包的处理。可编程数据平面的抽象转发模型实现了全流水线可编程,控制程序负责定义匹配动作表的执行顺序,进而实现转发逻辑的实现,控制流水线分为 Ingress control 和 Egress control 两部分。

可编程数据平面真正实现了协议无关的数据转发,并且作为一种描述性的高级编程语言,无需关心底层设备的具体信息,实现了设备无关性和代码可移植性,使得网络更加灵活和开放。

2 高性能数据中心网络及 PFC 算法

随着各种网络应用以及移动互联网的快速发展,近些年网络流量也出现了爆发式的增长。网络流量的激增对于作为基础设施的网络提出了更高的性能要求。网络中的不同应用对于网络性能的要求也不同,例如交互式应用需要网络提供更小的时延,分布式存储应用需要网络更小的丢包率,但概括来说,网络性能指标主要包括时延、吞吐量以及丢包率。因此实现超低时延、高吞吐量以及无丢包的高性能数据中心网络成为当前研究的热点。

2.1 远程直接数据存取技术(RDMA)

远程直接数据存取技术(RDMA)最早应用在高性能计算领域,是为了解决网络传输中服务器侧数据处理延迟、降低服务器 CPU 负载而设计的一种技术协议。如图 3 所示,RDMA 技术通过允许用户态的应用程序直接读取和写入远程服务器端内存,避免了 CPU 多次介入拷贝内存,在没有双方服务器端操作系统参与的情况下,可以绕过服务器内核,直接向网卡写数据,将数据直接从网络中一台服务器的内存传输到网络中另一台服务器。这样就释放了部分服务器的计算能力。并且,通过避免双方操作系统的介入可以消除外部存储器复制和上下文切换的开销,从而可以解放服务器内存带宽,有助于应用系统性能提升。RDMA 技术可以实现低时延、高吞吐量的高性能数据中心网络。

目前 RDMA 有三种实现方式,分别是 InfiniBand、iWARP(internet wide area RDMA protocol)和 RoCE(RDMA over converged Ethernet)。

(1)InfiniBand 技术[14]是由 IBTA(InfiniBand trade association)行业协会在 1999 年提出的,其标准规范在 1999 年开始起草并于 2000 年正式发表,经过不断发展,InfiniBand 架构在 2005 年之后开始在集群式超级计算机上广泛应用。

InfiniBand 技术目前主要应用于高性能计算数据中心网络。InfiniBand 技术在 RDMA 三种实现方式中

具有最好的性能,但需要有定制的硬件设备来实现,因此也是成本最高的一种实现方案。

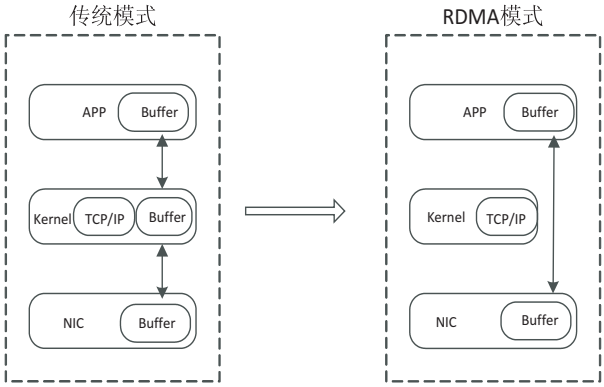

图 3　RDMA 模式与传统模式对比

(2)iWARP 协议[15]是 RDMA 技术的另外一种实现方案,是由 RDMA 联盟在 2002 年向 IETF 提出的,iWARP 协议通过在标准 TCP/IP 协议栈上定义一个多层处理栈实现将 RDMA 的特性集成到以太网上。iWARP 协议利用 RDMA 技术的内核旁路、零内存拷贝以及避免 CPU 介入等特点,有效地降低了网络延迟和服务器 CPU 的负载,释放了服务器部分 CPU 的计算能力。

iWARP 协议是第一个在标准以太网基础设施上实现了 RDMA 技术的方案,另外,iWARP 协议没有具体指定底层物理层设备信息,因此所有工作在使用 TCP/IP 协议栈的上层协议都可以被支持,在 iWARP 协议中,TCP/IP 协议栈在网卡中设计并实现。因此需要专有网卡来实现。当数据中心网络规模较大时,使用 iWARP 协议会产生大量的 TCP 连接,大量的 TCP 连接会占用服务器大量的内存资源,进而影响系统性能,所以会带来性能和成本问题。

(3)RoCE 技术是 IBTA 提出的在融合以太网上实现 RDMA 技术的一个实现方案。目前主流版本是 RoCEv2 版本。

RoCEv2 是基于 UDP/IP 协议实现的。RoCEv2 由于在性能和成本上占据明显优势,因此占据了目前融合以太网数据中心市场的主流地位。RoCEv2 是基于不可靠传输的 UDP 协议实现,与 TCP 协议相比,UDP 协议更加快速、占用较少的计算资源,但其是不可靠传输,没有 TCP 协议的滑动窗口、确认应答等机制,当出现丢包时会大大降低RDMA技术的工作效率。图4显示了当发生丢包时 RDMA 技术的吞吐率情况,可以看出丢包会严重影响 RDMA 的性能。所以为了实现 RDMA 技术的真正性能,必须实现无损传输。RoCEv2 协议依靠基于优先级的流量控制 PFC 算法实现无损传输。

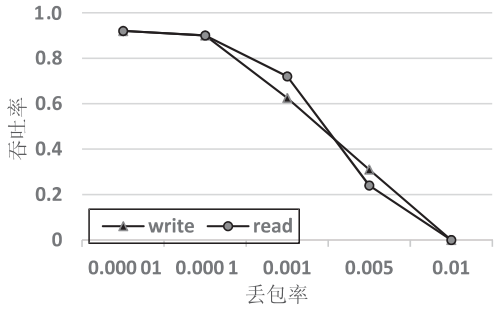

图 4　RDMA 的丢包率与吞吐率关系

2.2　基于优先级的流量控制算法 PFC

基于优先级的流量控制算法 PFC 是由电气和电子工程师协会在 2008 年提出的,PFC 算法通过划分虚拟实现是对传统流控协议的优化。当出现拥塞时,传统流控协议会将链路上的所有流量禁止发送。而 PFC 协议通过在以太网链路上创建 8 个优先级通道,并实现对每个优先级通道进行暂停发送,同时不会影响其他优先级通道数据的发送。

优先级流控 PFC 的基本运行机制如下:当网络链路中发生拥塞时,即交换机的入端口队列长度超过设定的阈值时,交换机生成一个暂停帧,并将暂停帧发送给上一跳交换机,上一跳交换机接收到暂停帧后,会根据暂停帧中的信息,暂停交换机中对应优先级通道中数据的发送,暂停的时间根据暂停帧中携带的字段信息设定;同时,如果上一跳交换机中也发生了拥塞则会向其上一跳交换机发送暂停帧,如果拥塞继续则逐级暂停。同样的,如果交换机中拥塞消除,则向上一跳交换机发送一个暂停时间为 0 的暂停帧,以恢复上一跳交换机中对应优先级队列中数据的发送,以此类推,逐级发送。PFC 算法[16]被广泛应用于高性能数据中心网络,以确保数据中心网络的无损传输。

图 5 显示了优先级流量控制 PFC 的报文格式,其中 DA 字段是目的 MAC 地址,SA 是源 MAC 地址,type 字段是报文类型,其中有关暂停优先级和暂停时间的信息包含在 Parameters 字段中。

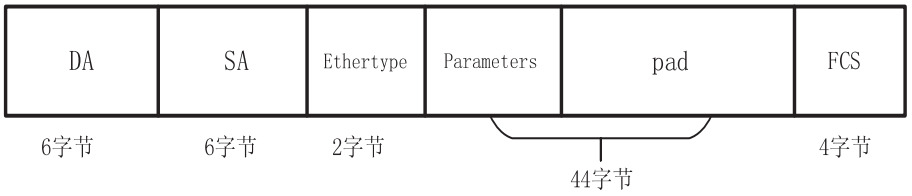

图 5　PFC 报文格式

3 基于可编程数据平面的 PFC 算法实现

该文在可编程数据平面的协议无关架构下实现了基于优先级的流量控制 PFC 算法,进而实现了无损传输网络,并在 Mininet 仿真环境下进行了仿真实验。

3.1 PFC 算法实现模型

该文基于可编程数据平面协议无关架构的特点,设计实现了基于优先级的流量控制 PFC 算法。实现模型如图 6 所示。

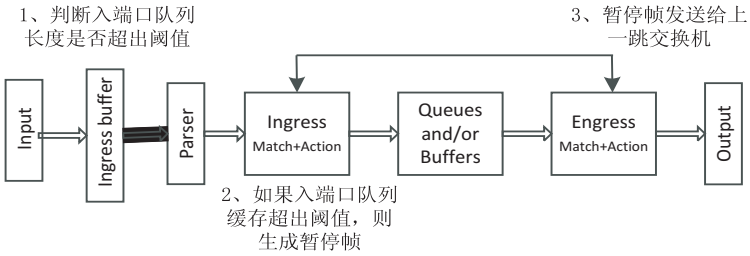

图 6 PFC 算法实现模型

首先在协议无关架构交换机的入端口队列对队列长度进行判断,根据设定的阈值判断交换机是否发生拥塞,如果入端口队列长度超出阈值,则判定交换机发生了拥塞,交换机生成 PFC 暂停帧,并发送给上一跳交换机。其中,通过自定义报文头实现对暂停帧的报文解析。

```
header pause_t {
    bit<48>   dstAddr;
    bit<48>   srcAddr;
    bit<16>   etherType;
    bit<16>   conCode;
    bit<16>   priEnable;
    bit<16>   time0;
    bit<16>   time1;
    bit<16>   time2;
    bit<16>   time3;
    bit<16>   time4;
    bit<16>   time5;
    bit<16>   time6;
    bit<16>   time7;
    bit<208>  reserved;
    bit<32>   fcs;
}
```

上述代码为使用 P4 自定义的暂停帧报文头 Header 字段,经过 Parser 解析器将进入交换机的数据包按照 Header 字段定义进行解析。获得相应的报文头实例,报文头实例将会在匹配动作表中使用。其中暂停帧的生成与发送,以及接收到暂停帧后暂停相应优先级通道数据发送等动作是在匹配动作表和控制流水线中完成。

3.2 实验验证

该文基于 Mininet 工具搭建了一个仿真网络环境,采用了数据中心网络常用的叶脊结构拓扑,包括 2 个 spine 交换机,4 个 leaf 交换机,4 个 tor 交换机和 8 个终端主机。实验拓扑如图 7 所示。

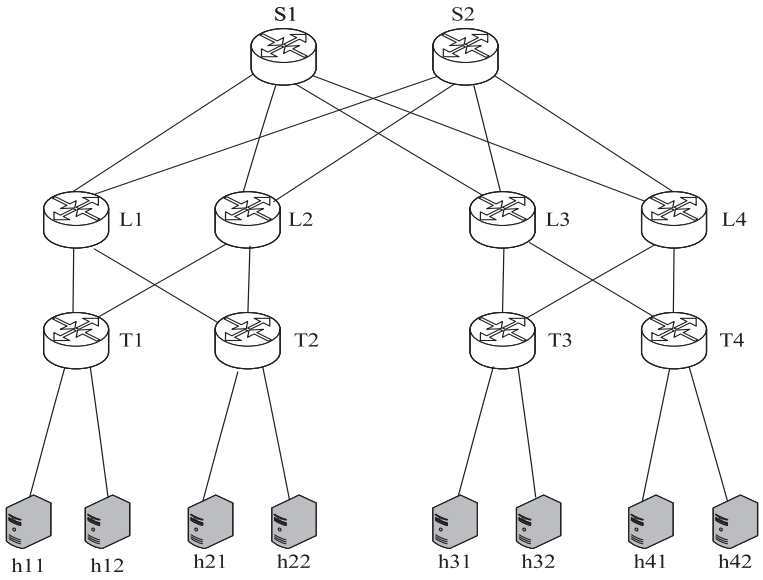

图 7 仿真实验拓扑

其中交换机使用 bmv2 软件交换机,实验拓扑中链路带宽设置为 5 Mbps,分别使用 Iperf 从主机 h11 和 h31 以 3 Mbps 的速度向主机 h41 发送 UDP 流,则在交换机 L3、L4、T4 中必有一个会发生拥塞。此时测试网

络丢包率,实验结果如图 8 所示。

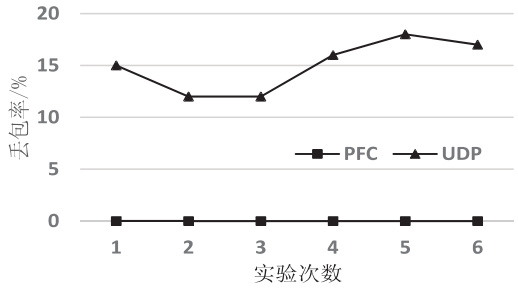

图 8　PFC 算法和 UDP 丢包率对比

通过网络丢包率的实验结果可以看出,在没有部署 PFC 时,使用 UDP 协议在发生拥塞时产生了丢包,而在部署了 PFC 后,可以实现无丢包网络。

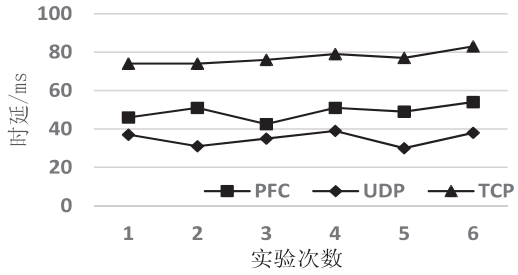

图 9　PFC 算法和 UDP、TCP 协议时延对比

同时,在仿真网络环境下对比了基于可编程数据平面实现的 PFC 算法与 UDP、TCP 协议的平均端到端时延,实验结果如图 9 所示。UDP 协议由于没有使用丢包重传以及流控机制,因此具有最小的端到端时延,但是不能保证无丢包传输,而基于可编程数据平面实现的 PFC 算法相对于 TCP 协议具有更低的端到端时延。

该文验证了在可编程数据平面下采用软件定义的方式可以实现基于优先级的流量控制 PFC 算法,并能确保网络无丢包传输。对网络功能的开发升级周期短,相对于传统的 PFC 算法实现方式更加灵活和高效。

4　结束语

当前各种快速发展的应用对数据中心网络提出了性能上的挑战,因此高性能数据中心网络是当前研究的热点。该文介绍了当前高性能数据中心网络研究的现状以及实现无损传输的意义,同时介绍了当前在网络领域最具活力的软件定义网络和可编程数据平面技术,最后在可编程数据平面下实现了基于优先级的流量控制 PFC 算法,并在 Mininet 仿真网络环境下对采用软件定义方式实现的 PFC 算法进行了实验验证。实验结果显示使用可编程数据平面可以很灵活、便捷地实现基于优先级的流量控制 PFC 算法,并确保了网络的无损传输,说明可编程数据平面技术在实现高性能数据中心网络中可以发挥巨大的作用。传统网络架构下,网络功能的开发升级周期长,往往需要花费数年的时间,该文使用软件定义的方式实现网络功能,更加高效、便捷,也证明了可编程数据平面技术为网络创新带来的巨大空间。

参考文献:

[1] 王　健,陈　威,汤卫东,等. 分布式并行网络拓扑计算关键技术研究[J]. 电力系统保护与控制,2017,45(2):117-122.

[2] 马铭冀,张晓蕾,杨继家. 云时代下数据中心网络技术研究[J]. 科技创新与应用,2017(15):99.

[3] 李　丹,陈贵海,任丰原,等. 数据中心网络的研究进展与趋势[J]. 计算机学报,2014,37(2):259-274.

[4] HANDLEY M,RAICIU C,AGACHE A,et al. Re-architecting datacenter networks and stacks for low latency and high performance[C]//Proceedings of the conference of the ACM special interest group on data communication. Los Angeles,CA,USA:ACM,2017:29-42.

[5] 曾　珊,陈　刚,齐法制. 高性能云数据中心弹性网络研究[J]. 计算机工程与应用,2018,54(7):89-95.

[6] RECIO R,METZLER B,CULLEY P,et al. A remote direct memory access protocol specification[S]. [s. l.]:[s. n.],2007.

[7] 陈淑平,吴志兵,张运德. RDMA over Ethernet 技术研究[J]. 高性能计算技术,2015(4):26-31.

[8] IEEE. 802. 11Qbb. Priority based flow control[S]. [s. l.]:IEEE,2011.

[9] 左青云,陈　鸣,赵广松,等. 基于 OpenFlow 的 SDN 技术研究[J]. 软件学报,2013,24(5):1078-1097.

[10] 俞慧春. SDN 技术的发展和应用浅析[J]. 中国新通信,2014(16):85.

[11] BOSSHART P,DALY D,IZZARD M,et al. Programming protocol-independent packet processors[J]. ACM SIGCOMM Computer Communication Review,2013,44(3):87-95.

[12] 张朝昆,崔　勇,唐骞祎,等. 软件定义网络(SDN)研究进展[J]. 软件学报,2015,26(1):62-81.

[13] 屠海令,孙棕檀,姚　源,等. 近六年《麻省理工科技评论》"全球十大突破性技术"解析与启示[J]. 中国工程科学,2017,19(5):85-91.

[14] 倪永军. InfiniBand 技术分析与应用研究[J]. 计算机应用研究,2003,20(12):4-6.

[15] 陈彦灵,吴　安,张　斌,等. 面向云服务器系统的分布式网络架构与技术研究[J]. 电信网技术,2017(8):8-11.

[16] 武　磊. 一种基于无损以太网的流量控制机制[J]. 中国新通信,2015(11):110-111.