数据中心中面向光互联的流量识别与调度研究

郭秉礼[1]，赵宁[2]，朱志文[1]，宁帆[2]，黄善国[1]
（1. 北京邮电大学信息光子学与光通信研究院，北京 100876；2. 北京邮电大学信息与通信学院，北京 100876）

摘 要：为了解决数据中心链路拥塞问题，依据流量分布与类型的特点，提出了基于光互联架构的流量识别和调度方案，即 HCFD（host-controller flow detection），旨在识别出对网络性能影响较大的大象流。利用 SDN 控制器下发转发策略，对网络中的流量进行合理调度。 HCFD 首先在主机端利用 Linux 内核协议的 Netfilter 框架实现将超过阈值的数据流进行标记，然后在控制器端利用决策树分类模型再对标记流进行分类，最后利用光电混合网络的优势，实现深度融合的流量适配和切换机制。HCFD 方案整合了已有方法的优势进行大象流识别，同时保证了识别的实时性、准确性以及流信息的全面性。实验与仿真结果显示，在此方案场景下，能有效缓解网络拥塞情况，充分利用网络带宽，减少数据端到端时延，降低分组丢失率。

关键词：光互联架构；流量识别；软件定义网络；Linux 内核
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2018161

Research on traffic identification and scheduling based on optical interconnection architecture in data center

GUO Bingli[1], ZHAO Ning[2], ZHU Zhiwen[1], NING Fan[2], HUANG Shanguo[1]
1. Institute of Information Photonics and Optical Communication, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: In order to solve the data center link congestion problem, based on the characteristics of the flow distribution and flow types, a flow identification and scheduling scheme based on optical interconnect structure, named HCFD (host-controller flow detection), was proposed to identify the elephant flow which has a large impact on the network performance, and use the SDN controller to make forward strategy, and schedule the network traffic reasonably. The implementation of the scheme was to use the Netfilter framework in Linux kernel protocol on the host side to mark the flow that exceeds the threshold amount. Then, the classification model was used in the controller side to classify the marked flow. Finally, the appropriate forwarding strategy was developed based on the above results. With the advantage of the photoelectric network, mechanisms of flow depth fusion and switching could be realized. The scheme which integrates the advantage of the existing research results, was expected to identify elephant flow more accurately and comprehensively. It can effectively alleviate the network congestion, make full use of network bandwidth, reduce end-to-end delay and packet loss rate.

Key words: optical interconnection architecture, traffic identification, software define network, Linux kernel

1 引言

数据中心（DC, data center）作为云计算提供服务的主要基础设施，集中了大量的数据、存储资源及提高数据运行效率的计算资源，其通过网络设备（高速链路、路由器等）进行互连，为各种基础服务提供支持。数据中心在未来的互联网服务领域扮演着举足轻重的作用。近年来，在大量涌现的云应用的推动下，如实时视频、搜索引擎、Map-Reduce 计算和虚拟机迁移等，使数据中心流量呈飞速增长态势。思科 2018 年 2 月发布的年度云产业调研报告中预测[1]，到 2021 年全球云数据中心流量将达

收稿日期：2018−05−24；**修回日期**：2018−08−27

到每年 20.6 ZB，比 2016 年的每年 6.8 ZB 增长 3 倍。海量的数据以及复杂多变的业务对数据中心网络流量管理提出了巨大的挑战。现有数据中心网络（DCN, data center network）流量工程机制已经无法快速应对突发多变的流量形式，已有的研究成果[2]表明，86%的数据中心的链路因为突发的高带宽的数据流易而出现短暂的拥塞。网络的拥塞直接导致大量数据分组丢失、网络吞吐量下降、搜索时延变长、QoS 质量无法保证等问题。为了解决数据中心的流量问题，首先，对网络管控层来说，需要为数据中心的流量制定合理的流量调度策略。研究表明，数据中心虽然会产生巨大数目的流量，但是整体的网络状况是由少数的持续时间长的大流，也被称为大象流[2-3]（此处定义突发的高带宽的流量为大象流）所影响。从高效管理网络的角度考虑，控制器没有必要处理所有数据流的调度，只集中于对网络性能有重要影响的流操作。因此流量识别，尤其是对大象流的识别，就变得十分重要，通过转移少数突发的高带宽流量到空闲的通信链路上，能够有效缓解网络拥塞情况，充分利用网络带宽，减少时延。同时，近年来，许多基于光电路交换、光分组、光突发交换的数据中心内网络结构被提出。如混合光电交换网络架构通过集中式控制器对光链路的调整进行流量调度，一些大容量、持续时间长的大流（即大象流）会被引流到光网络上进行传输，而一些小数据量，对时延要求高的小流（老鼠流）仍采用电交换设备进行转发。基于光交换的混合光电网络一方面可以提供快速通路 shortcut 进行横向流量优化，另一方面光交换机的可重配能力也给整个网络架构带来了灵活性；不仅能有效利用光网络的大带宽、低时延、低功耗等优势，也保持了电域交换的灵活性。这能有效应对数据中心突发流量。总之，本文拟通过研究对数据中心流量进行识别和标记的方法，结合光电网络各自的优势，同时研究深度融合的流量适配与切换机制，可以实现数据流量到光电网络的高效适配，缓解网络拥塞情况，充分利用网络带宽，减少数据端到端时延。因此，本文实现了主机端的识别模块和控制端的分类模块，即在主机端能够标记超过阈值的大流功能以及控制器上实现了将标记流转移到光交换路径上，小流实现默认路由电转发的调度策略，最终达到网络流量的负载均衡以及优化。

2　相关的工作

目前的流量识别技术主要有如下 2 种，一种是基于交换机，在数据流传输过程中基于数据量或特征进行识别，如流采样技术，即在交换机上部署统计模块或使用第三方代理工具 sflow/Netflow，交换机周期性地采集网络流的数据分组，控制器则分析采样结果统计得到样本特征，基于样本特征推导出网络整体流量的特征，并制定流表规则，下发到交换机，交换机再依据所制定的流表规则进行数据分组转发，这些技术或多或少存在着交换机/控制器开销大，准确度低，实时性差等问题。而另一种是基于主机端进行流识别，如 Curtis 等[4]通过设计虚拟机的应用程序（mahout）来收集流信息，Yun 等[5]则是通过监控 TCP 发送队列的数据量来判断大象流。2 种方法都认定流缓存值超过指定阈值的网络流为大象流，第二种流量识别技术因为是在主机端识别，而且有阈值设定机制，所以相较于第一种提高了大象流识别的准确性，也加快了识别的速度，减少了交换机/控制器的系统开销。但是，值得注意的是，有些特定的大象流虽然数据量大，但是带宽占用并不高且传输速率低，因此此类流存在着被 Mahout 方法误检为大象流的风险，而且无法修正判断结果。而 Yun 的方法，虽然能通过控制器修正结果，但方法本身存在耗时长的问题，因为忽略了在主机端阻塞 TCP 队列带来的排队时间。另外，利用机器学习进行流量识别也是一个研究热点[6-7]。Chao 等[6]提出了 FlowSeer，一种能够进行快速、低开销的流量识别系统，仅通过分析流的前 5 个数据分组信息，使用预先训练好的分类模型，就能准确识别出该流的速率和持续时间。该方法的实现较复杂且 SDN 控制器/交换机的处理开销较大，因为需要在控制器和交换机上都部署模型，不符合 SDN 的数据面和控制面的分离思想。如表 1 所示 [8]，列举了目前研究下对大象流的各种识别方法以及性能评估。整合现有研究方法的优势，本文提出一种新的流量识别方案。

3　HCFD 模型

3.1　流量识别方案

为了解决现有研究的不足，在目前光电混合的数据中心架构下，提出一种结合主机终端和控制器

表 1　　**不同的大象流识别方法性能比较**

类别	相关技术	准确度	耗时	开销
基于交换机基于数据周期取样轮询	Hedera,Helios,HSP	高	一般	高
基于交换机基于数据周期取样抽样	Sflow	低	高	低
基于交换机基于数据全统计	FlowRader,DevFlow	高	高	一般
基于交换机基于特征标记识别	EiffiEye	低	一般	高
基于交换机基于特征流分类	文献[3-4]	高	一般	一般
基于终端基于数据无虚拟机应用	Mahout,MicroTE	高	一般	一般
基于终端基于数据有虚拟机应用	EMC2,VirtMonE	高	一般	一般

端的流量识别方案 HCFD（host-controller flow detection）。方案旨在吸取已有研究的优点，高效、实时、快速地识别出大象流。并有效地利用光电网络的优势，实现深度融合的流量适配和切换机制。光电混合架构的流量识别方案如图 1 所示。

方案由主机端模块和控制器端模块组成。通过在主机端配置监控模块起到流量预判断的作用，标记出数据量超过阈值的流，即可能成为大象流的数据流，这一步能够过滤掉大多数的小流，减小交换机的识别开销。因为交换机只需要专门处理被标记过的流。在控制器端的分类模块，能够分类得到标记流的带宽范围，然后制定转发策略，将大流转移到光网络传输，而小流默认在电网络中传输，这样就能满足大流的带宽需求，也能满足小流的时延需求，同时避免了链路拥塞[9]。

3.2　主机端模块

Linux 内核协议栈为网络开发者提供了大量的系统调用和函数库，如其中的 libnetfilte_queue 库。应用层可以通过 NFNETLINK 与内核进行交互，直接性地截获内核每个数据分组，并在应用层对数据分组进行逻辑处理，然后将裁决结果返回给内核。主机端模块如图 2 所示。

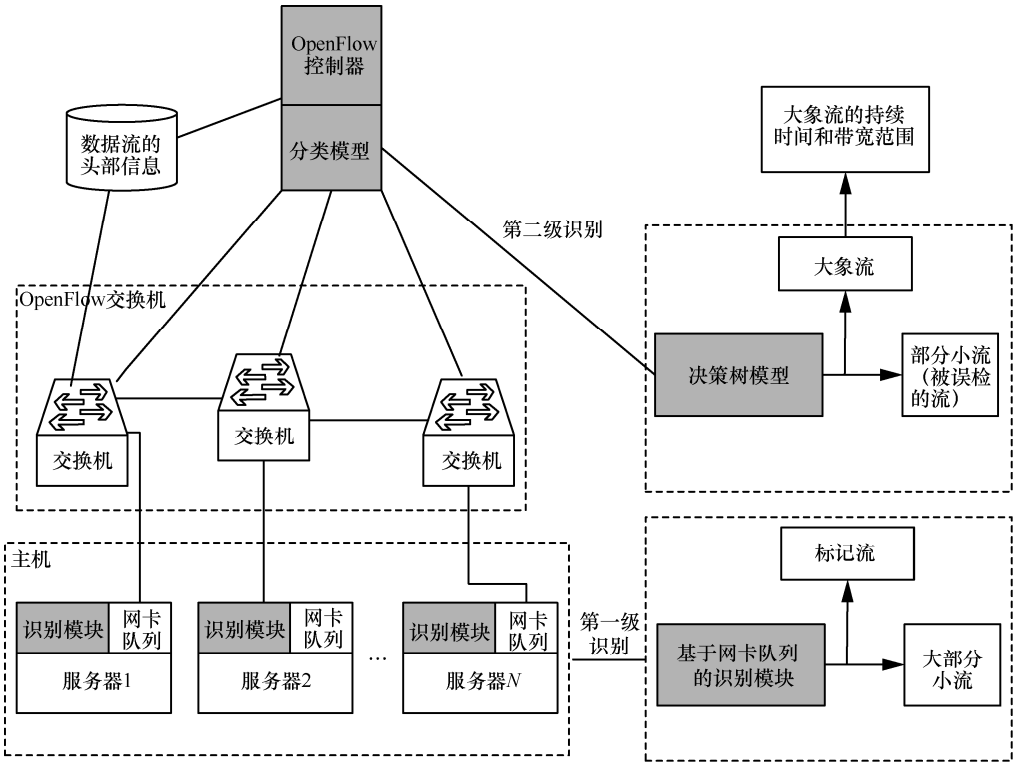

图 1　光电混合架构的流量识别方案

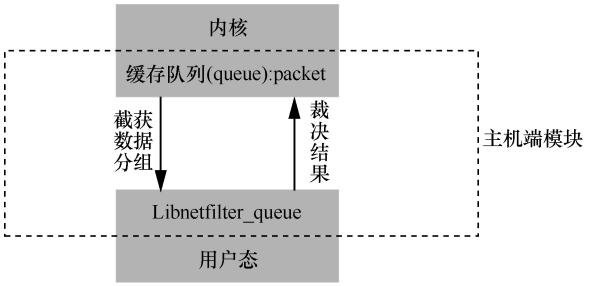

图 2　主机端模块

具体实现上，用户态进程 thread() 利用 libnetfilter_queue 库从内核缓存的队列中截获的数据分组，将数据分组的 IP 头部信息，通过散列函数映射成 *key* 值，*key* 将作为一条流的唯一标识，然后将数据分组的数据负载部分作为 *value*[10]。最后将 <*key*, *value*>保存在散列表内（链表数组）。如果新来的数据分组所在的流已经在散列表中，则将新来的数据分组的 *value* 值放入旧流中进行更新（原 *value* 值加上新来的数据分组的 *value* 值），反之，则新建一条流表项。在用户态进程持续处理每个数据分组的过程中，如果当某条流的数据量超过设定阈值，就对该流的 ToS 字段标记，视为标记流。如图 3 所示，在代码实现阶段利用 libnetfilter_queue 库提供的各种函数，第一步初始化，生成 handler，接着绑定 AF_INET 协议簇，指定数据分组缓存队列，并调用 nfd_handle_packet()取得每个数据分组，实现逻辑部分，再利用回调函数通知内核，内核依据裁决结果，选择对数据分组继续转发，然后退出程序。

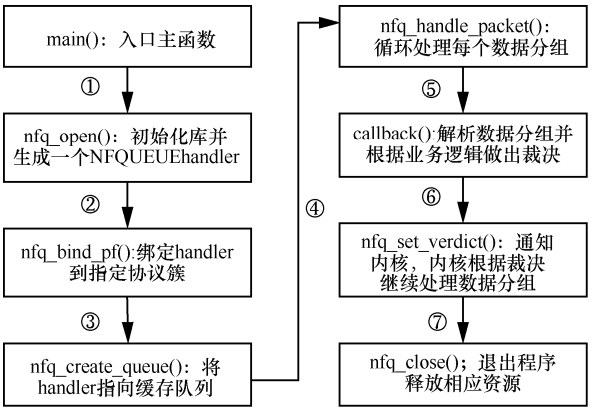

图 3　Libnetfilter_queue 处理数据分组流程

3.3　控制器端模块

利用 OpenFlow 控制器端做二级识别。在数据分组传输过程中，被标记过的标记流一旦经过交换机，交换机会直接将该流数据分组的 IP 头部信息上交给控制器进行处理，而未被标记过的数据流将按照 ECMP 等价多路径算法计算出转发规则而不必通过控制器制定流表项。控制器将拿到的 IP 头部信息输入事先训练好的分类模型（model），如图 4 所示，预测该流的带宽范围和持续时间范围，这里的分类模型采用的是决策树如图 4 所示，是用数据中心内该节点的历史流量数据进行训练所得。控制器根据链路带宽情况制定对该流调度策略，并通知给交换机，交换机将按照下发的流表规则转发后来的被标记的数据流。

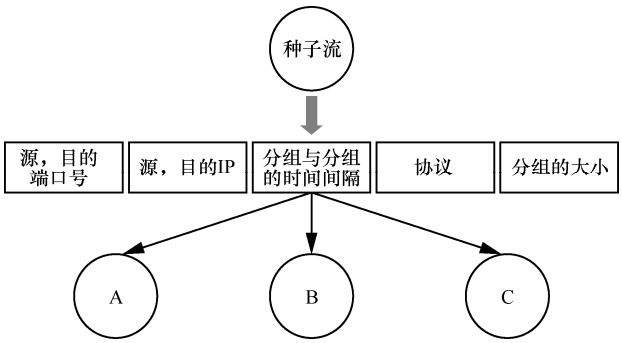

图 4　控制器端分类模型

基于决策树模型的控制器模块能够通过历史数据预测出标记流的带宽范围，比如 1~10 MB，10~100 MB 以及 100 MB 以上。对于控制器来说，针对不同带宽可以制定不同的路由策略，充分利用空闲链路，进行数据传输。目前研究表明，快速光交换模块已经被引入数据中心内，安装在栈顶交换机负责连接各个服务器，传输链路中的大象流。所以，控制器模块可以结合快速光交换模块实现光电路切换。

4　性能测试与分析

针对本文提出的 HCFD 方案，设计仿真实验测试其性能。

首先验证方案中主机端模块的标记功能。模拟实际网络环境中的流量分布，利用 Iperf 在配置好模块的主机生成 1~100 KB 大小的文件 100 个，100 KB~1 MB 大小的文件 10 个，1~100 MB 的文件 5 个。默认情况下，阈值设定为 1 MB，研究表明[7]，数据中心内超过 80%流的数据量都不超过 1MB。然后将数据源发送给接收端的主机（IP:10.108.48.14），发送端的主机（IP:10.108.49.201）会对超过阈值（1 MB）流的 IP 数据分组头 ToS 字段进行标记，默认标记值为 192。然后在接收端服务器利用 wireshark 进行

抓包分析，发现大流的数据分组的 ToS 字段都被标记成 0011xxxx，而小流的数据分组的 ToS 字段仍然是 0000xxxx，如图 5 和图 6 所示。

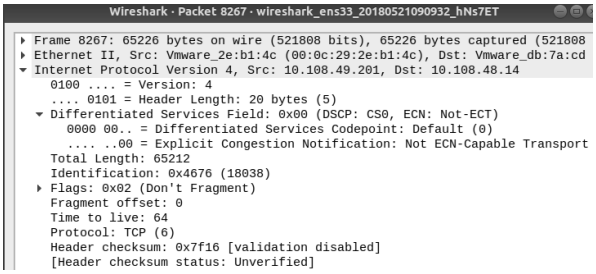

图 5　未被标记的数据分组

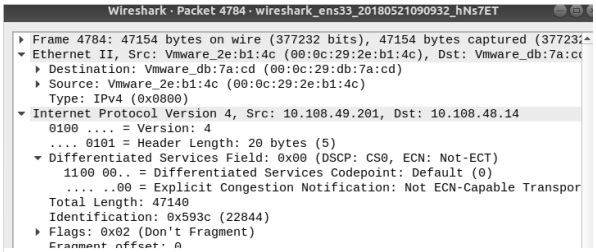

图 6　成功被标记的数据分组

由此可以验证主机端模块能够实现对超过阈值的数据流进行标记的功能。接下来，设计实验，验证 HCFD 方案，在网络分组丢失率以及端到端传输时间的性能表现。

Fat-tree 是实际数据中心场景中常见的拓扑结构，这种拓扑结构对于每个源、目的主机间都有多条链路可到达，如图 7 所示。

按照 Fat-tree 拓扑结构[11]，用 Mininet 配置如下模拟环境。

所有的主机均部署主机端识别模块，且与交换机相连。利用 Iperf 生成符合实际网络流量分布的数据源。实验在主机 A 生成了 100 条 10~100 KB 流以及 10 条 1~10 MB 的大流（大流会被主机端模块标记），分组间隔设为 1ms，分组的长度设为 1 KB，持续时间为 10 s[12]。然后将该数据源发送给主机 B。这里有 2 条路径可以选择（A—电交换机—B；A—电交换机—光交换机—B），路径的选择由控制器决定。

表 2　　模拟环境配置

参数	功能	数值
H_{total}/台	主机总数	128
H_{per}/台	每台交换机连接的主机数	8
S/台	交换机总数	16
Flow/条	每个终端每秒产生的新流量数目	10
T/s	持续时间	60

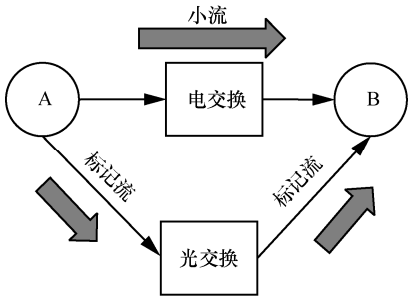

图 8　流 91CF 调度模型

实验分别验证使用 HCFD 模型以及未使用 HCFD 模型 2 种场景下的端到端时延和分组丢失率情况。在未使用 HCFD 模型的场景中，数据流将默认按照 A—电交换—B 路径传输。当在 HCFD 模型的场景下时，数据量大的流在发出网卡之前，会被主机端模块将 ToS 字段标记成 192，变成标记流，然后才会发出网卡。

默认情况下，当网络数据分组途径交换机时，如果是未标记的数据分组，则将直接被转发到 B 主

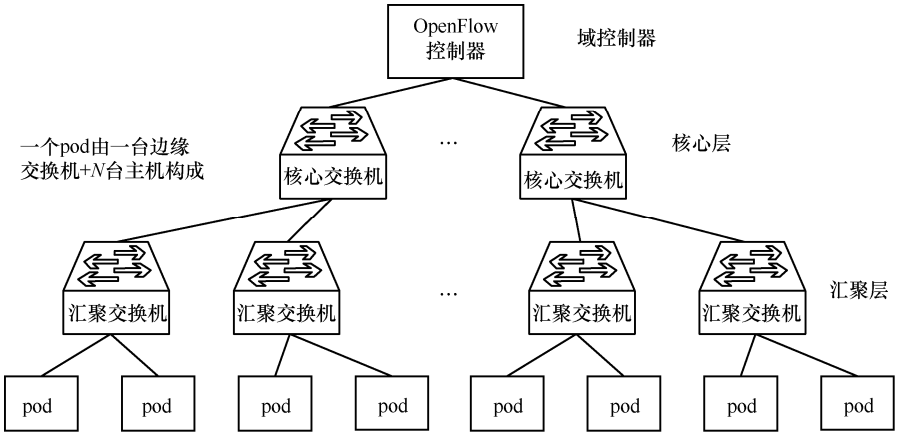

图 7　数据中心 Fat-tree 拓扑结构

机，如果数据分组 IP 头部 ToS 字段匹配到交换机设定的流表项，交换机便会将标记流的信息通过 OpenFlow 协议上传到控制器，控制器经过决策会下发转发规则到交换机，此处默认下发的规则是将标记流切换到光交换机上，然后再转发到主机 B。

通过实验数据，可以得到采用 HCFD 模块比未采用模块，数据流的平均端到端时延会减少大约 70%和 45%。如图 9 所示。

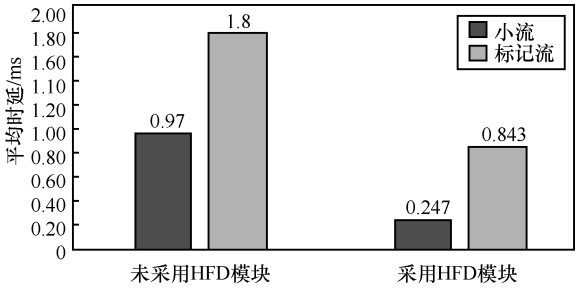

图 9　数据分组的端到端时延情况

而在分组丢失率上，小流和标记流分组发送数目分别是 9 093 和 8 845，在 2 种场景下进行验证，时延参数限制为 5 ms，发现在未采用 HCFD 模型的场景下，两者的分组丢失率分别达到 62%和 51.8%，而采用 HCFD 模型的场景下，分组丢失率降低到 5%和 17%。这说明 HCFD 模型能够有效降低网络分组丢失率。如图 10 所示。

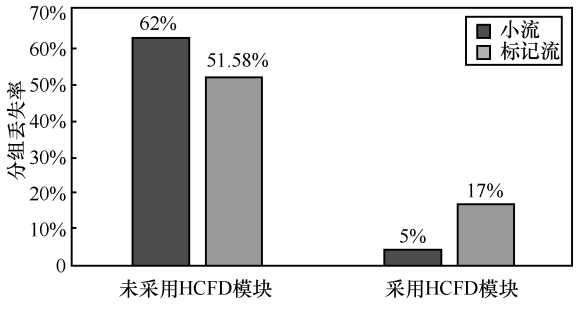

图 10　数据流的分组丢失率

通过控制大流的分组发送速率即分组发送时间间隔，单位为 ms，来验证大流和小流在接收端的时延变化。如图 11 和图 12 所示，当设置一组分组发送间隔，使大流的发送速率逐渐增加，可以看出有无模块下数据流的时延变化情况，加载模块的场景整个平均时延上都低于未加载模块的场景，且小流的平均时延也大幅度地降低并保持稳定（0.247 s）（此处设置了一定的网卡缓存队列长度，当数据分组过多就会导致大量分组丢失，时延也会有较大幅度上升。

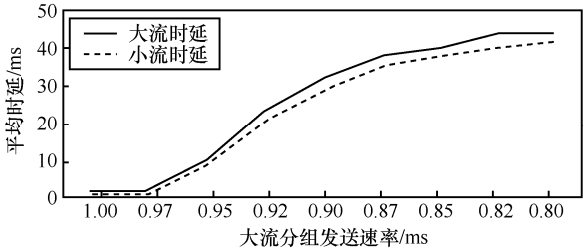

图 11　未采用模块的数据流时延情况

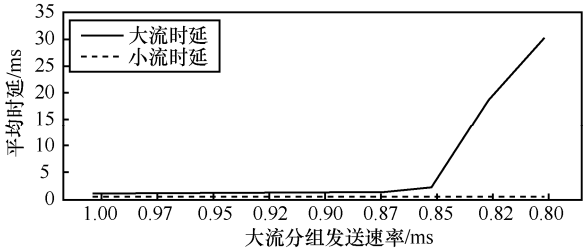

图 12　采用模块的数据流时延情况

这种方案相较于 Curtis 等[4]提出的 Mahout 主机端识别技术，不需要过高的大象流阈值（Mahout：10 MB）设定，这样便使处理的数据量更小，速度也更快，而且整体主机端开发难度会更简单，从代码实现上，不需要过多修改内核协议。同时，相比 Chao 等[6]的 FlowSee 和 Huang 等[7]的 Application-Round 的方案，他们需要在交换机上部署比如 C4.5 决策树模型进行预识别，然而本方案不需要交换机参与做分类识别，交换机只需负责基本的数据转发功能，更满足 SDN 逻辑面与数据面解耦合的思想。而且，HCFD 方案本身的准确度和速度预期也能达到目前高速网络流的要求

5　结束语

本文设计了一个快速、低开销大象流识别以及调度模块，能够在主机端标记大象流，并且让控制器在识别出链路中的标记流之后进行分类，基于全局的流量分布情况，合理地选择路径，采取光电切换的流量调度策略对大流进行处理，将大流重路由到其他的路径，以避免网络拥塞和提高网络链路利用率。

参考文献：

[1] CISCO. Cisco global cloud index: forecast and methodology, 2015-2020[R]. SanJose: Cisco Public, 2016.

[2] SRIKANTH K, SUDIPTA S, ALBERT G, et al. The nature of data center traffic: measurements & analysis[C]//The 9th ACM SIGCOMM Conference on Internet Measurement (IMC '09). 2009: 202-208.

[3] GREENBERG A, HAMILTON J R, JAIN N, et al. VL2: a scalable and flexible data center network[J]. Communications of the ACM, 2009, 54(4): 95-104.

[4] CURTIS A R, KIM W, YALAGANDULA P. Mahout: low-overhead datacenter traffic management using end-host-based elephant detection[C]// IEEE INFOCOM. 2011:1629-1637.

[5] 严军荣,叶景畅,潘鹏.一种大象流两级识别方法[J].电信科学,2017,33(03):36-43.
YAN J R,YE J C, PAN P. A two-level method for elephant flow identification[J]. Telecommunications Science,2017, 33(3):36-43.

[6] CHAO S C, LIN K C J, CHEN M S. Flow classification for software-defined data centers using stream mining[C]//IEEE Transactions on Services Computing.

[7] HUANG Y H, SHIH W Y, HUANG J L. A classification-based elephant flow detection method using application round on SDN environments[C]//2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS).2017:231-234.

[8] WANG B, SU J. A survey of elephant flow detection in SDN[C]// International Symposium on Digital Forensic and Security. 2018:1-6.

[9] 蔡岳平,樊欣唯,王昌平.光电混合数据中心网络负载均衡流量调度机制[J].计算机应用与软件,2017,34(08):145-150+166.
CAI Y P, FAN X W, WANG C P. Load balance traffic scheduling mechanism in an optical-electrical hybrid data center network[J]. Computer Applications and Software, 2017,34(8):145-150+166.

[10] RAN B B, EINZIGER G, FRIEDMAN R, et al. Optimal elephant flow detection[J]. IEEE INFOCOM .2017: 1-9.

[11] 罗军舟, 金嘉晖, 宋爱波, 等. 云计算: 体系架构与关键技术[J]. 通信学报, 2011, 32(7): 3-21.
LUO J Z, JIN J H, SONG A B, et al. Cloud computing: architecture and key technologies[J]. Journal on Communications, 2011, 32(7): 3-21.

[12] GANG D, ZHENG H G, HONG W. Characteristics research on modern data center network [[J]. Journal of Computer Research and Development, 2014, 51(2): 395-407.

作者简介：

郭秉礼（1982-），男，山西忻州人，北京邮电大学讲师，主要研究方向为网络存储、并行与分布式系统等。

赵宁（1995-），女，四川南充人，北京邮电大学硕士生，主要研究方向为 SDN、流量工程。

朱志文（1994-），男，安徽合肥人，北京邮电大学硕士生，主要研究方向为 SDN、数据中心网络。

宁帆（1962-），女，吉林长春人，北京邮电大学教授，主要研究方向为通信系统与网络的理论与应用、教学模式等方面。

黄善国（1978-），男，山东济南人，博士，北京邮电大学教授，主要研究方向为数据布局、并行与分布式系统。