基于 AWGR 的 OCS/EPS 数据中心光电混合网络

臧大伟[1),2)] 曹 政[1)] 王 展[1),2)] 刘小丽[1),2)]
付斌章[1)] 孙凝晖[1)]

[1)](中国科学院计算技术研究所 北京 100190)
[2)](中国科学院大学 北京 100190)

摘 要 随着云计算和大数据应用技术的发展,数据中心的数量和规模迅速发展,为了满足服务器之间大规模数据流动的需求,数据中心网络的通信能力面临巨大的挑战.传统数据中心中,网络的路由交换设备一般仅采用电域交换技术,电域交换技术虽然可以快速地、灵活地切换数据包的传输路径,但其本身存在通信带宽低、交换容量有限、高能耗等缺点.为了提高数据中心网络的性能、降低网络的能耗,最近的研究提出了若干基于慢速路径切换光器件的光电混合网络结构.它们通常只能将小部分数据量非常大的网络流放在高带宽的光网络上传输,其他的网络流仍然需要电域网络传输.随着快速可调波长激光器以及光波长路由器件的成熟,使光电混合网络结构灵活应对动态、多样的流量模式成为可能.该文基于快速可调波长激光器 TWC(Tunable Wavelength Converters)和光波长路由器 AWGR(Arrayed-Waveguide Grating Router),首次提出了一种 OCS(Optical Circuit Switching)/EPS(Electrical Packet Switching)光电混合网络结构 Ace-net.在文中详细描述了光电混合网络的结构以及带宽测量、仲裁控制、流量分配等机制,这些机制利用 TWC 器件快速波长变换的特性,能快速地应对网络流量的变化,使更多的网络流量在光域网络上传输;同时使用模拟器对此结构进行了评测,模拟结果表明此网络结构具有很好的网络性能.

关键词 数据中心网络;光电混合网络;阵列波导光栅路由器;光线路交换
中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2016.01868

AWGR-Based OCS/EPS Hybrid Datacenter Network

ZANG Da-Wei[1),2)] CAO Zheng[1)] WANG Zhan[1),2)] LIU Xiao-Li[1),2)]
FU Bin-Zhang[1)] SUN Ning-Hui[1)]

[1)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2)](*University of Chinese Academy of Sciences, Beijing* 100190)

Abstract With the increasing requirements of cloud services, the size and quantity of datacenters are growing rapidly. In order to meet the needs of large-scale data exchange among servers, datacenter network faces enormous challenges in the communication capabilities. In traditional data center, the routing and switching equipment typically uses electronic switching technology which is fast and flexible to switch packets between ports. But it inherently has low communication bandwidth and limited switching capacity. In order to improve the performance of datacenter network and reduce its energy consumption, many MEMS-based hybrid network architectures have been proposed recently. In these architectures, a fraction of flows which is very big is placed on the optical network while the rest flows are placed on the electronic network. With the maturity of

收稿日期:2015-03-18;最终修改稿收到日期:2015-07-09.本课题得到国家自然科学基金(61572464,61331008)、国家"八六三"高技术研究发展计划项目基金(2015AA01A301)、华为项目(YB2015070066)资助.**臧大伟**,男,1988 年生,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为计算机系统结构、数据中心网络.E-mail:zangdawei@ncic.ac.cn.**曹 政**,男,1982 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为计算机系统结构、高性能计算机网络.**王 展**,男,1986 年生,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为计算机系统结构.**刘小丽**,女,1986 年生,硕士,助理工程师,中国计算机学会(CCF)会员,主要研究方向为计算机系统结构.**付斌章**,男,1983 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为计算机系统结构、计算机网络.**孙凝晖**,男,1968 年生,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为计算机系统结构、文件系统.

Tunable Wavelength Converters (TWC) and Arrayed-Waveguide Gating Router (AWGR), the hybrid network has the opportunity to serve dynamic and various flows. In this paper, we propose Ace-net which is a hybrid OCS/EPS network architecture basing AWGR and TWC optical device. We first describe current research status of the hybrid network architecture in detail. Then we present the design and the key characteristics of Ace-net, which include traffic demand estimation, arbitration control and traffic distribution. Also, we evaluate this structure using our simulator and the simulation results show that the network structure has good performance.

Keywords datacenter network; hybrid optical-electrical network; arrayed waveguide grating router; optical circuit switching

1 引　言

数据中心作为支撑现代社会的基础设施，对人类的生活产生越来越大的影响. 近年来，随着新的服务需求不断涌现、服务质量要求的提高以及云计算等新兴计算模式的发展，数据中心的数量和规模得到了迅速的发展，大型数据中心运行着数万台服务器[1]. 为了满足服务器之间的大规模通信需求，数据中心网络面临成本、结构和性能等多个方面的挑战.

首先，现在数据中心网络的交换设备一般采用电域交换技术，电域交换技术虽然可以快速、灵活的切换数据包的传输路径，但其本身存在通信带宽低、交换容量有限等缺点. 当前数据中心网络一般采用两层或者 3 层的树形结构，为了降低成本以及实际部署的复杂性，导致网络的各层之间存在严重的网络带宽缩减①. 因此汇聚层和核心层交换机成为网络通信的瓶颈，并带来了网络整体利用率低以及数据传输时延较长等问题[2].

其次，现在的数据中心网络系统使用大量的电域交换机来搭建拓扑结构，交换机之间一般使用光纤来连接，数据包每通过一级交换机都要经过光-电-光的转换，存在着巨大的能量消耗. 在 2010 年，全球的数据中心消耗了全球用电量的 1.1%～1.5%，并且这个比例在持续增加[3]. 随着服务器节能技术的发展，处理器等主要部件的能耗已得到有效的控制，而网络设备的能耗则日益突出. 例如，在 Google 的数据中心中，当服务器处理器的利用率为 100%时，网络系统的能耗占总能耗的 20%；而当处理器的利用率为 15%时，网络系统的能耗达到数据中心总能耗的 50%. 在 Google 数据中心中，服务器处理器的利用率只有 10%～50%[4-5]，网络能耗已经成为现在数据中心能耗的重要组成部分.

随着集成电路工艺和硅光技术的发展，光交换技术逐渐进入实用阶段，像 MEMS(Micro-Electro-Mechanical Systems)光交换机和 AWGR(Arrayed-Waveguide Grating Router)光路由器已在工业环境中使用. 与电域交换技术相比，光交换技术具有极高的传输带宽以及几乎不受限制的交换能力，非常适合于缓解当前数据中心网络通信带宽低、交换容量有限和高能耗的问题[6]. 基于以上特征，最近的研究提出了若干面向数据中心的光电混合网络结构，如 c-Through[7]、Helios[8]、OSA[9]等；它们结合高带宽的光域网络和传统的电域网络，并构建统一的逻辑控制平面，将网络流量按照需求分配到电域网络和光域网络上传输. 当前光电混合网络结构通常基于慢速光线路切换器件，在机柜之间建立一条持续时间很长的光链路，直接将小部分数据量非常大的网络流由一个机柜传输到另一个机柜中，有效地缓解了现在数据中心网络的问题.

然而当前光电混合网络所使用的慢速路径切换光交换机的重配置延迟很高，严重限制了混合网络的适用性. 在光线路的重新配置过程中，必须缓存分配到光域网络上传输的网络包，等待光线路配置的完成. 因此这些结构对运行的应用程序的通信模式有较大的限制，需要稳定、持续的网络流来分摊线路配置所引起的延迟. 如 c-Through 使用的 MEMS 光交换机采用机械方式控制光线路的建立，光线路的配置过程需要数十毫秒的时间，主要面向于虚拟机迁移等持续时间较长的网络流，并不能适应于不同大小和快速变化的网络流[10-11].

随着快速可调波长激光器 TWC 技术[12]以及

① Cisco Data Center Infrastructure 2.5 Design Guide. http://www.cisco.com/univercd/cc/td/doc/solution/dcidg21.pdf

光波长路由器 AWGR 技术[13]逐渐的成熟,使光电混合网络结构应对动态、多样的流量模式成为可能.快速可调波长激光器具有纳秒级延迟的波长变换特性,而光波长路由器是能够基于波长进行路由的被动光器件.使用这两种器件构建的光域网络结构,配合适当的控制平面,可以快速、低延迟地重配置光线路,使通信模式复杂的应用程序也可以利用光网络,提高整个数据中心网络的普适性.

为了提高现在数据中心光电混合网络应对复杂网络流量模式的能力,使更多的网络流能在光域网络上传输,本文首次提出了一种使用 TWC 和 AWGR 两种光器件构建的 OCS/EPS 异构光电混合网络结构 Ace-net;为了充分利用 TWC 光器件纳秒级的波长变换特性和 AWGR 基于波长路由的特性,设计了快速的控制平面,降低线路重配置的延迟,在机柜之间快速的建立起一条持续的光链路.相比于现有的光电混合网络结构,Ace-net 使用实时的流量需求测量和带宽信息收集方法,并使用快速的算法进行线路仲裁.在 Ace-net 中,电域网络和光域网络并没有主次之分,其中光域网络主要用于传输数据量稍大的网络流;电域网络主要用来传输未建立光线路的机柜之间的流量和光域网络重建期间的网络流量.

与现有的光电混合网络系统相比,Ace-net 的设计主要有 3 个创新点.首先,它首次使用快速可调波长激光器 TWC 和光波长路由器 AWGR 替代慢速路径切换光交换机,构建 OCS/EPS 光电混合网络结构.其次,它使用实时流量需求测量、收集方法,能灵活应对服务器的流量变化.第三,它采用以目的地址为索引的虚拟队列管理方式,将数据包缓存在主机内存中,但所占用的内存量远远小于现有的光电混合网络结构.

本文的主要贡献有:(1)提出了一种使用 TWC 和 AWGR 光路由器的 OCS/EPS 异构光电混合网络结构 Ace-net,其使用可快速变换波长的 TWC 光器件和光波长路由器 AWGR 在两个机柜间建立起一条持续的光线路,使更多样的网络流可以在高带宽的光域网络中传输,进一步减轻电域网络的压力;(2)设计了统一控制平面,并使用了实时的流量需求测量机制和快速的光链路调度配置算法,充分发挥 TWC 可快速变换波长的特性;(3)采用了终端缓存的方法,使用较少的服务器内存缓存待发送的数据包,同时使用以目的地址为索引的虚拟队列管理方式,对每一台服务器中的网络流进行统一的管理,减少服务器内存的额外消耗;(4)对此结构进行了详细的论证分析,并使用模拟器进行了评测,结果表明此结构可以很好地降低网络中端到端的延迟,极大地增加了网络的吞吐率.

本文第 2 节介绍光器件的相关背景知识;第 3 节介绍数据中心混合网络的研究现状及相关工作;第 4 节详细阐述光电混合网络的组成和结构;第 5 节对光电混合网络的性能进行测评;第 6 节对本文进行总结.

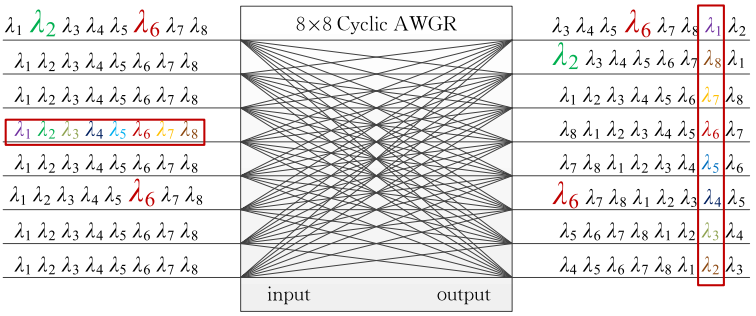
2 光器件简介

随着集成电路工艺和硅光技术的发展,光器件的设计和制造技术逐渐的成熟,许多光传输技术和光交换技术已在通信领域广泛使用.本部分将对现在光电混合网络结构中所使用的光器件和涉及的光通信技术进行介绍.

MEMS-switches(Micro-Electro-Mechanical System Switches).MEMS 光交换机是一种由微型机械装置控制的光器件;在光交换机内部有一个 $N\times N$ 的反光镜阵列,这些反光镜依附在一些微小的电机上,由一个嵌入式处理器来控制.根据外部的控制命令,嵌入式处理器通过控制反光镜阵列的旋转角度,将一个端口输入的激光束重定向到一个输出端口,在两个端口之间建立一条持续的光链路;由于它使用的是微型的机械控制系统,配置一次需要数毫秒至数十毫秒的时间.

AWGR(Arrayed-Waveguide Grating Router),一种根据光的波长信息进行路由的被动光学器件,其循环波长路由特性允许不同输入端口的光束路由至同一输出端口.一个 N 端口的 AWGR,由端口 i 输入的波长为 w 的光束将被路由至 $[(i+w-2) \bmod N]+1$ 端口,如图 1 所示一个 8 端口的 AWGR 及其波长路由表[14].不同波长的光线通过同一输出端口时,这些光线将被复用在一条光纤上传输,而不发生网络竞争.AWGR 是被动器件,具有极低的功耗,延迟仅皮秒至纳秒级.更为关键的是,AWGR 可以实现较高的维度,在 250 nm SOI(Silicon-On-Insulator)技术下,加州大学戴维斯分校的 Ben Yoo 教授团队流片成功的 AWGR 具有 512 端口,面积仅 16 mm×11 mm[13].

OSM(Optical Switching Matrix).一般情况下,OSM 模块采用二分矩阵形式,有 N 个输入端口和 N 个输出端口,通过合适的配置可以将任何一个



(a) 8端口AWGR结构图

	Output 1	Output 2	Output 3	Output 4	Output 5	Output 6	Output 7	Output 8
Input 1	λ_3	λ_2	λ_1	λ_8	λ_7	λ_6	λ_5	λ_4
Input 2	λ_2	λ_1	λ_8	λ_7	λ_6	λ_5	λ_4	λ_3
Input 3	λ_1	λ_8	λ_7	λ_6	λ_5	λ_4	λ_3	λ_2
Input 4	λ_8	λ_7	λ_6	λ_5	λ_4	λ_3	λ_2	λ_1
Input 5	λ_7	λ_6	λ_5	λ_4	λ_3	λ_2	λ_1	λ_8
Input 6	λ_6	λ_5	λ_4	λ_3	λ_2	λ_1	λ_8	λ_7
Input 7	λ_5	λ_4	λ_3	λ_2	λ_1	λ_8	λ_7	λ_6
Input 8	λ_4	λ_3	λ_2	λ_1	λ_8	λ_7	λ_6	λ_5

(b) 8端口AWGR路由表

图 1　8 端口 AWGR 及其波长路由表

输入端口的光线路传输到任何一个输出端口. 现在常用的 OSM 器件一般采用 MEMS 技术来实现，使用机械控制来重新配置端口的连接关系，光线路的重配置时间与 MEMS 光交换机类似[15].

WSS(Wavelength Selective Switch). 通常是一个 $1\times N$ 的光交换机，包含一个光输入端口和 N 个输出端口，配置一次的时间需要数毫秒. 它可以将输入的一系列波长的光束划分成不同的组，通过不同的输出端口输出. 例如，如果输入端口输入的光线中包括 100 个波长，那么它可以将波长 1～30 路由到 1 号输出端口，50～80 路由到 2 号输出端口.

TWC(Tunable Wavelength Converters). 快速可调波长激光器，TWC 可以将一个输入的光信号变换成给定的波长输出. 每个 TWC 包含一个可调激光器、一个 SOA(Semiconductor Optical Amplifier)和一个 MZI(Mach-Zehnder Interferometer). 在波长调制过程中，首先由 MZI 生成给定波长的激光束，然后由 SOA 来完成激光的调制工作；它接受某个输入波长的激光束，然后按照给定的控制，将其输出为给定波长的激光束. 现在的 TWC 可以实现 160 Gbps 的波长转换带宽[12]，其重构时间与输入输出的波长相关，延迟波动范围在 1 ns 到 30 ns 之间.

OCA(Optical Channel Adapter). 光发送和接收器件，在发送端它会将电信号转换成光信号，在接收端它会将光信号转换成电信号. 每一个 OCA 器件有一个 $1:N$ 的分光器，可以将一束混合光分离成多束单波长的光，同时最多可以处理 N 种波长；每一个器件都有一个接收器阵列，可以将光信号转换为电信号. 使用该器件可以将一束由多种波长混合而成的光信号转变为电信号，使其继续在电域网络中传输.

3　相关工作

为了解决网络通信能力不足的问题，最近的研究提出了若干新的数据中心网络结构[16-19]，从不同方面弥补了传统树形网络结构的不足. 但是这些网络结构仍然使用纯电域交换技术，无法解决现有电域网络通信带宽低、交换容量有限、能耗高等缺点. 随着集成电路工艺和硅光技术的发展，光交换技术逐渐进入实用阶段，像 MEMS (Micro-Electro-Mechanical Systems)光交换机和 AWGR (Arrayed-Waveguide Grating Router)光路由器已在工业环境中使用. 与电域交换技术相比，光交换技术具有极高的传输带宽以及几乎不受限制的交换能力，非常适合于缓解当前数据中心网络通信带宽低、交换容量

有限和高能耗的问题[6]. 基于以上特征,最近的研究提出了若干面向数据中心的光电混合网络结构,如c-Through[7]、Helios[8]、OSA[9]、REACToR[20]等;它们结合高带宽的光域网络和传统的电域网络,并构建统一的逻辑控制平面,将网络流量按照需求分配到电域网络和光域网络上传输,以此来缓解纯电域网络中的问题.

c-Through是一种基于MEMS光交换机构造的光电混合网络结构,它在树形电域网络的基础上增加一层光域网络,数据中心中的ToR(Top of Rack)交换机同时连接到电域网络和光域网络;通过控制MEMS光交换机在两个机柜之间建立一条持续的光线路,使数据量很大的网络流在光域网络上传输,其网络拓扑结构如图2所示. 通过流量监控系统监测任意两个机柜之间的流量需求,形成一个流量需求矩阵;为了尽可能地利用高带宽的光链路,将机柜之间光线路的配置关系转变成了一个最大权值的完美匹配问题,使用Edmonds算法求出这个问题的最优解;根据计算结果配置MEMS光交换机,在机柜之间建立光线路连接,然后使用基于VLAN的路由方法将ToR交换机中的流量分配给电域网络和光域网络. c-Through使用了慢速的MEMS交换机,在很长的间隔后进行一次新的配置,且配置一次的时间需要数百毫秒的时间,系统缺乏灵活性且只适合于处理数据量很大的网络流,适应性有限.

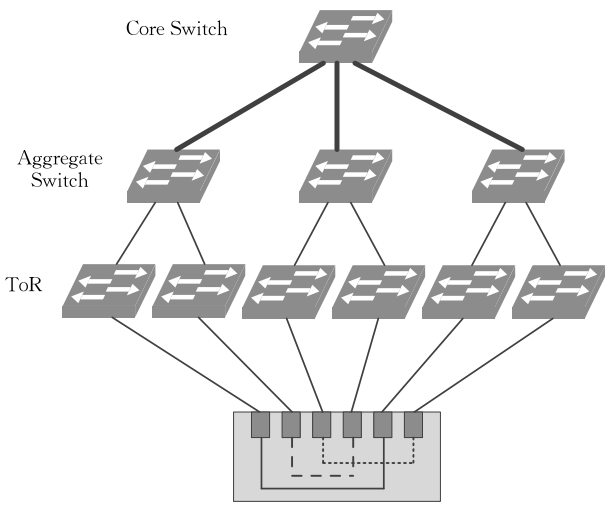


图2 c-Through拓扑结构

Helios是为模块化的数据中心设计的一种光电混合网络结构,采用了与c-Through类似的MEMS光交换机,在PoD之间动态建立光线路,同时它使用了WDM(Wavelength Division Multiplex)技术增加光线路的带宽,网络拓扑结构如图3所示. 其网络采用两层的多根树结构,核心层同时采用电域交换机和光域交换机:电域交换机用于处理all-to-all的通信,而光域交换机用于处理机柜间高流量需求的大流量通信. 在交换机中使用TM(Topology Manager)模块统计两个业务单元之间的流量需求和连接数,形成一个流量需求矩阵;使用Edmonds算法计算出最优的光链路连接关系后,光交换机管理模块CSM(Circuit Switch Manager)在PoD之间建立光链路;最后通过接入层交换机控制软件PSM(PoD Switch Manager)在交换机中修改路由关系. Helios与c-Through存在相似的问题,整个配置流程需要数百毫秒的时间,网络缺乏灵活性.

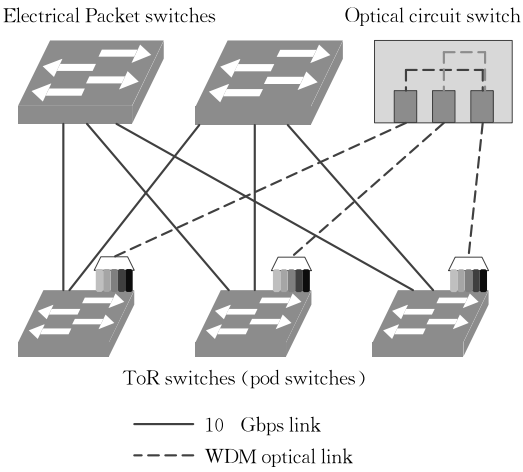


图3 Helios拓扑结构

OSA是一种使用WSS和OSM光交换机在一个PoD内的机柜之间构建的多跳光网络结构,其结构如图4所示. 它利用光线路交换的方式,将网络根据流量的特征重构成直连网络拓扑结构,数据包从源端点到目的端点需要经过多跳步. 通过使用WSS来配置不同波长的光线,其连接的每条链路的带宽可以动态的调节. 它周期性地收集机柜间网络的需求信息,并根据这些信息将拓扑的计算转换成一个求解带权值的b-matching问题,然后通过控制OSM光交换机和WSS光器件重新配置拓扑并为链路分配带宽. OSA结构同样缺乏灵活性和普适性,配置一次的时间需要数百毫秒的时间,仍然需要稳定、持续的网络流的支持.

DOS[21]是一种基于AWGR光路由器件的全光域包交换OPS(Optical Packet Switching)网络结构,它只在ToR交换机中采用了电域交换,而上层网络采用全光域互连,如图5所示. 每一个ToR交换机通过TWC连接到AWGR光路由器的一个端口,在每一个TWC的前端有一个标签提取器LE

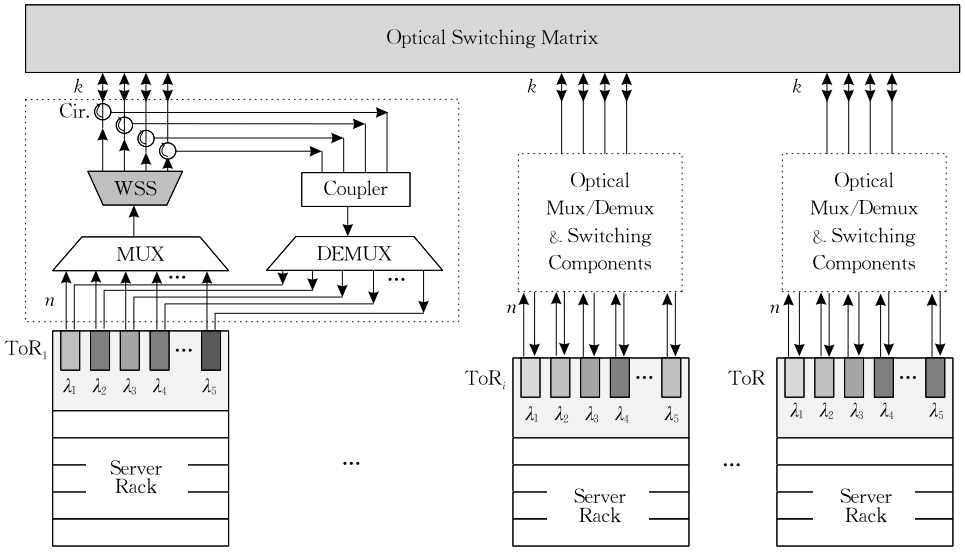


图 4　OSA 结构

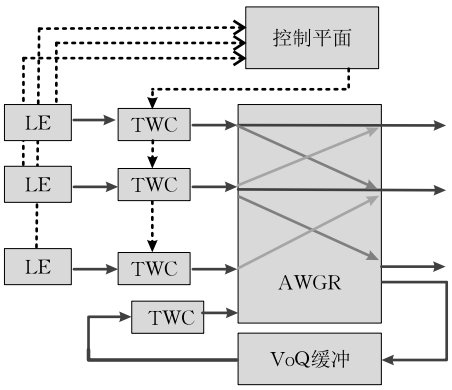


图 5　DOS 拓扑结构

(Label Extractor)，它可以提取每一个数据包的目的地址. 当一个数据包要传输时，LE 会提取其目的地址信息，同时将信息传给中央控制平面；然后由中央控制平面调整 TWC 的输出波长，经由 AWGR 光路由器将数据包传输到相应的端口；为了防止波长竞争，在系统中存在一个单独的电域缓存，将不能发送的数据包暂时缓存. DOS 是一种基于数据包交换的光域网络，它所转发的每个数据包都要经过 LE 和控制平面的处理，增加了数据包的延迟；同时它所采用的电域缓存需要进行光电转换，增加了能耗.

c-Through、Helios、OSA 等网络系统在原有电域网络的基础上使用了高带宽的光链路来缓解当前电域网络通信带宽低、交换容量有限、高能耗的问题. 他们虽然在结构上各有差异，但是他们都是依靠慢速 MEMS 光交换机来重构网络连接关系，计算连接关系、拓扑重配置需要数百毫秒的时间. 因此这些结构对运行的应用程序的通信模式有较大的限制，需要稳定、持续的网络流来分摊线路配置所引起的延迟. Ace-net 虽然也是基于线路交换的光电混合网络结构，但其使用的纳秒级波长变换器件 TWC 和基于波长路由的光路由器 AWGR，使用快速的控制平面在机柜之间快速的建立起一条持续的光链路，可以灵活的应对多样、复杂的网络流量模式.

DOS 虽然也是使用快速波长变换器件 TWC 和 AWGR 光路由器构建的网络，但是其实质是一种纯光域网络结构；而 Ace-net 是结合电域网络灵活性的特点和光域网络高带宽、低能耗的特点所构建的光电混合网络结构，二者具有本质的区别. 同时，DOS 网络系统使用单一的控制器对每个数据包进行进行地址解析，系统的扩展性有限；Ace-net 是以一台服务器中到达同一目的地址数据包的总和作为仲裁对象，仲裁复杂度显著降低.

Ace-net 与 c-Through、DOS 有相似之处，但是有着本质的区别. Ace-net 借鉴 c-Through 的光线路交换(OCS)，在机柜之间建立起一条较持续的光线路；但其借助 TWC 纳秒级的波长变换特性，可以快速的切换光线路，适应网络流的动态变化. 同时，Ace-net 借鉴了 DOS 所使用的网络结构，但不采用光包交换(OPS)的策略，不需要对每个数据包进行解析；同时在处理光波长竞争时，Ace-net 消除了 DOS 中的光电转换、电域缓存的问题.

4　系统设计

Ace-net 是一种光电混合网络结构，它在现有数据中心电域网络的基础上增加光域网络，有区别地将特征不同的网络流分配到电域网络和光域网络

上传输,电域网络主要用来传输数据量非常小的网络流,而光域网络用来传输数据量稍大的网络流;其利用快速的控制平面和 TWC 快速的波长变换特性将更多的网络流量分配到光域网络上传输,以适应更复杂的流量模式,进一步缓解现有纯电域网络带宽低、交换容量有限和高耗能的问题.本节主要介绍 Ace-net 的结构以及所使用的主要策略和机制.

4.1 系统结构概览

在使用 Ace-net 混合网络结构的数据中心中,N 台服务器都被放置到 R 个机柜中,每个机柜中放置 N/R 台服务器.在机柜内部的每台服务器使用电域链路连接到本机柜内的 ToR 交换机上,机柜之间分别使用电域网络和光域网络互连;电域网络可以使用任意的拓扑结构,而光域网络中所有的 ToR 交换机都连接到同一个 AWGR 光路由器.

系统的可扩展性依赖于 AWGR 光路由器的端口数.NTT 的研究人员展示了使用 400 个通道 400×400 的 AWGR 路由器,覆盖的波长范围为 1530 nm 至 1610 nm[22],而 512×512 端口的 AWGR 光器件已经流片成功[13];如果一个机柜内放置 40～60 台服务器,那么使用一个 AWGR 光路由器的混合网络系统最多可以实现 3 万多台服务器的互连,完全满足一个超大规模数据中心的需求.因此我们假设 512×512 端口的 AWGR 已经完全满足我们的扩展性需求.

图 6 展示了 Ace-net 网络结构的一个简单例子,它使用传统的三层树形电域网络作为基础网络,使用单层的光域网络作为拓展网络.每一个 ToR 交换机有一个连接到上层电域交换机的电域端口;同时它有一个光域端口,端口的发送链路通过 TWC 光器件连接到 OCA 器件的发送端(Tx),端口的接收链路连接到 OCA 器件的接收端(Rx).

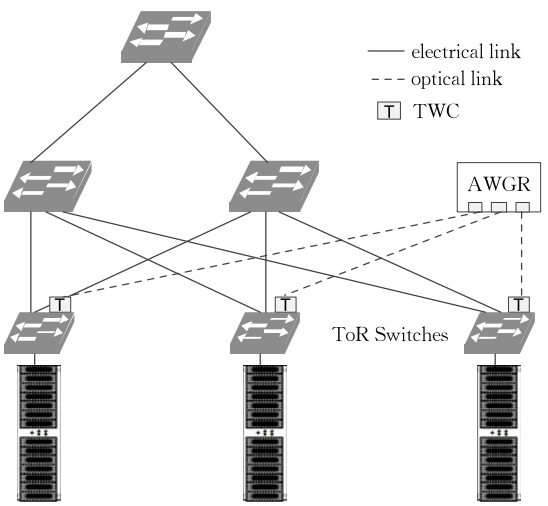


图 6 Ace-net 网络拓扑结构示意图

Ace-net 中使用一个快速仲裁控制系统对混合网络进行控制,它主要完成如下几个工作:(1)收集服务器的通信需求信息,用来作为光线路配置的依据;(2)根据收集到的通信需求信息,决定光线路的调度配置;(3)控制 TWC 光器件和 ToR 交换机,在机柜之间建立光线路;(4)控制每台服务器开始或者停止网络数据的发送.仲裁控制系统在逻辑上包含 3 个部分:网络流量测量模块、仲裁控制模块和流量分配控制模块.

4.2 网络流量测量

Ace-net 根据机柜之间对网络资源的需求,在两个机柜之间建立一条光链路来传输数据量较大的网络流,需要对服务器的流量进行测量,并控制网络包的传输时机,以等待光链路的建立.

一种理想的网络流量测量机制是采用应用程序通知的方法:一个进程在进行网络操作时,首先要将其发送的数据量和目的地址等信息通知网络控制系统.这种方法可以准确和及时地测量出应用程序对带宽的需求,但是其对应用程序不透明,需要重新定义编程模型,不能很好的兼容现有的各种应用程序和适应现在数据中心中多样化的应用程序运行环境.

为了适应现在数据中心的特点和确保应用程序的兼容性,Ace-net 采用一种对上层应用程序透明的监测方法:通过在操作系统内核中监控每一个 socket 缓冲区的占用情况,并将缓冲区的占用值按照目的地址划分、聚合到虚拟队列中,通过虚拟队列对相同目的地址的流集中管理,完成网络流量测量和网络流传送时机的控制.虚拟队列是存在于每台服务器操作系统中的一个表结构,它包含若干的项,其中的任意一项 E_i 对应数据中心中编号为 i 的机柜.每一项是一个三元组 $\{subnetID, socketList, dataVolume\}$,记录着本机与对应机柜之间的网络流信息;其中 $subnetID$ 是对应机柜的子网号,$socketList$ 是所有与对应机柜建立连接的 $socket$ 列表,$dataVolume$ 表示 $socketList$ 在此段时间内所累积的网络数据量.流量监测模块根据虚拟队列中的信息,完成对网络流的集中控制.

在操作系统中,每一个 TCP 连接建立时,都会有一个独占的 socket 缓冲区来缓存数据,从而保证几个并发的连接不会因为某个阻塞而全部无法发送数据.在 Ace-net 中,建立一个 TCP 连接的同时会根据目的 IP 地址将这个连接注册到相应的虚拟队列的对应项中,并添加到 $socketList$ 列表当中,使虚

拟队列可以管理和控制此连接. 同时，为了便于对大流的检测，Ace-net 将服务器操作系统中每一个连接的 socket 缓冲区增大至 10 MB，通过在运行时监测缓冲区的占用情况，对网络流的流量需求进行监测.

当一个进程调用 write 函数将数据写入 socket 缓冲区时，流量监测模块会提取写入的数据量（即 socket 缓冲区的占用量），如果数据量大于 64 KB，则认为这是一个大流，将此流量值与对应虚拟队列三元组中的 *dataVolume* 相加，作为流量的积累值；如果数据量小于 64 KB，则认为这个流是一个实时性较强的小流，虚拟队列不对其进行管理，由操作系统调度发送. 流量监测模块每隔 200 μs 会对虚拟队列中的项进行扫描，如果数据量的累积值不超过 2 MB，则控制这些网络流直接发送，不需要光链路的申请；如果数据量的累计值大于 2 MB，则有必要将数据放在光域网络上传输，此时需要向中央控制器申请光链路. 它会将本机累积的数据量、目的地址等信息通知中央控制器，以便根据这些信息进行光线路的仲裁和分配.

在将这些信息通知中央控制器时，如果实时的建立 TCP 连接，需要数十乃至数百微秒的时间开销，增加了光线路建立的延迟. 为了降低通知过程的延迟，Ace-net 采用了两种优化方法：第一，采用重复 UDP 数据包来传递信息. 由于整个控制系统运行在一个网络条件良好的专用网络环境中，具有较低的丢包率，同时我们选择采用重复 UDP 数据包来传递信息，即每个 UDP 数据包同时发送 3 次，在一段时间内接收端在接收到任何一个数据包后，会将后续接收到的相同内容的数据包丢弃，从而降低数据通信的不可靠性；由于在控制系统中，在短时间内不会有完全相同的控制命令发出，因此这种方法不会造成混淆. 假如网络的丢包率为 10^{-2}，则采用重复 UDP 数据包机制的失效率将降为 10^{-6}. 即使偶尔产生丢包，只会使流量统计信息产生短时间的错误，使本次流量的报告失败，可能短暂地影响系统的性能，而不会对系统产生致命的损害. 为了防止由于控制数据包丢失或者中央控制器故障等原因产生的中央控制器长时间不响应虚拟队列的链路请求的情况，在通知中央控制器的同时，虚拟队列会为自身中的这一项开启定时器，当 300 μs 内仍未得到中央控制器的响应，虚拟队列直接将这一项管理的网络流发送.

其次，为了减少把通知信息封装成数据包所带来的延迟，采用预生成-缓存-发送的模式，即将通知信息预先封装成数据包，在需要的时候直接发送. 系统中设置 15 个数据量门限值，分别为{2,4,6,8,10,14,18,22,26,30,38,46,54,62,100}；如果累积数据量在 2 MB 与 4 MB 之间，则选择数值 2 作为通知值，中央控制器解析时会使用中值 3 作为累积值. 将数据中心每一个机柜中所管理主机的子网地址与 15 个门限值进行组合，最多可以形成 7500 个组合数据，然后把每一个组合数据封装成一个网络包，使用子网地址和门限值的组合作为 hash 值将其保存在操作系统中，如图 7 所示. 当检测到某累积数据量

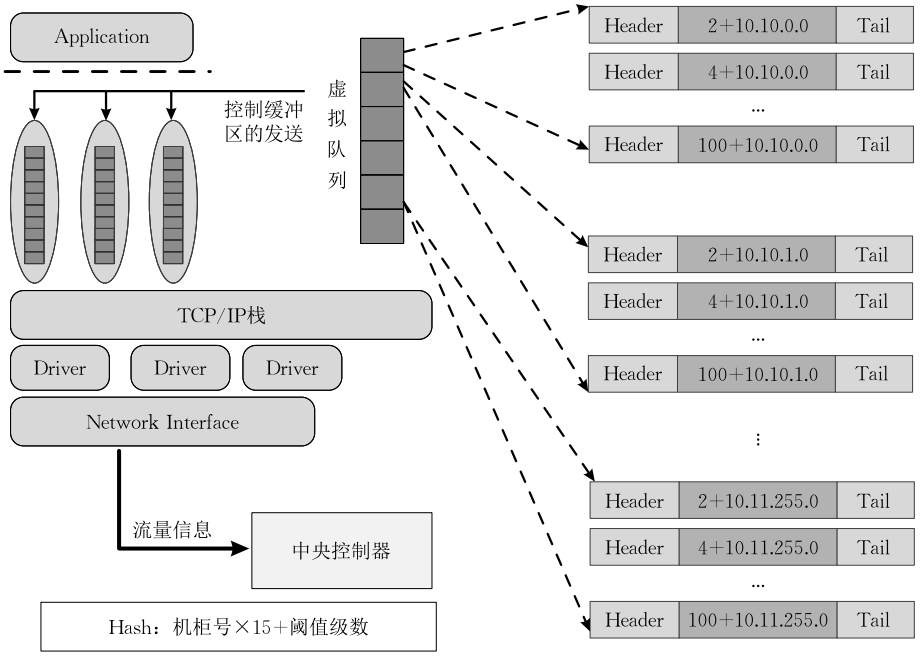


图 7　流量需求测量原理图

达到门限值时,使用此虚拟队列对应项的目的机柜号和数据量的门限值作为 hash 值,使用 hash 函数查找对应的网络包,并发送给中央控制器.

虽然 c-Through 等系统也将网络数据暂时缓存在主机的内存中,但是它将缓冲区扩大到上百兆字节来监测极大的网络流[7],以应对 MEMS 光线路建立速度较慢的问题,极大地增加了系统内存的消耗,而 Ace-net 的光线路建立速度极快,具有较强的实时性,从而缓冲区可以大大缩小.同时,c-Through 等系统以服务器中的流为单位进行管理,而 Ace-net 以服务器为单位,采用虚拟队列的方式进行管理,从而突出流的聚合效应,降低中央控制器控制的复杂度.

4.3 仲裁控制

在中央控制器获得每台服务器累积数据量的信息后,需要对这些信息进行解析处理,以便根据累积值在机柜之间建立一条持续的光链路,并与服务器操作系统中的虚拟队列管理模块相配合,使用光域网络来传输大块数据.

在现有的光电混合网络系统中,为了适应 MEMS 光交换机速度较慢的特性,将一段时间收集到的数据量信息形成流量需求矩阵,然后使用图的完美匹配算法计算出系统中应该如何建立光链路,在计算过程中流量需求矩阵不会变化,并且计算过程需要数百毫秒的时间[7-8]. Ace-net 所使用的光器件与 MEMS 光交换机具有本质的不同,为了使光网络能灵活的应对持续时间较短的网络流,采用"光链路无空闲"的策略,即所做出的任何光线路的仲裁和调度都要尽可能地使光链路有足够的数据传输,使光网络有较高的利用率.因此中央控制器必须向光域网络提供足够多的数据并减少光网络调度之间的时间间隙.

Ace-net 采用了实时的流量需求解析方法,如图 8 所示.每台服务器中的流量监测模块每隔 200 μs 会对虚拟队列中的项进行扫描,如果数据量的累积值不超过 2 MB,则控制这些网络流直接发送,不需要光链路的申请;如果数据量的累计值大于 2 MB,则有必要将数据放在光域网络上传输,此时需要向中央控制器申请光链路.它会将本机柜累积的数据量的门限值、目的机柜子网地址等信息通知中央控制器,以便根据这些信息进行光线路的仲裁和分配.在中央控制器中为每一个机柜维护了一个实时的 N 维流量累积向量(其中 N 表示系统中有多少个机柜)$(1,2,\cdots,k,\cdots,N)$,其中的第 k 个元素记录了本机柜与第 k 个机柜之间的累积流量需求.当中央控制器接收到新的流量通知数据包时,首先会提取其携带的门限值,并将处理过的门限值与相应流量需求向量中对应的元素值相加,作为新的累积流量需求.

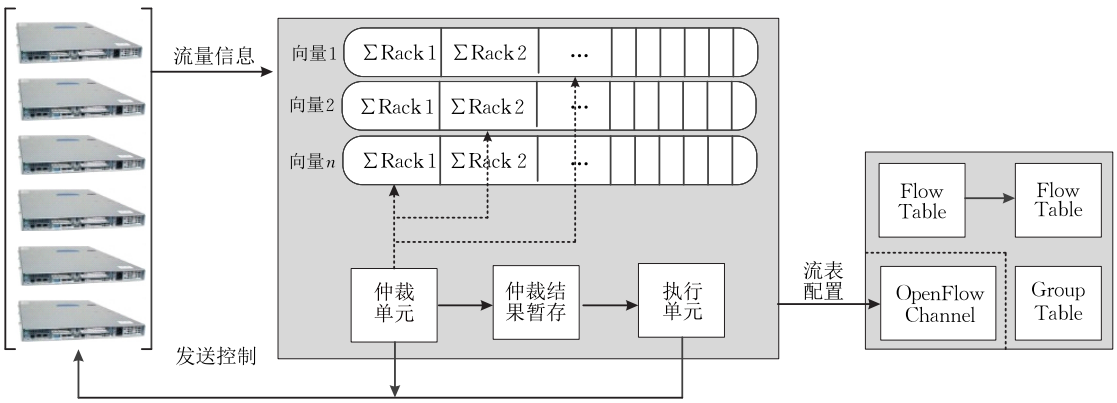


图 8 流量需求测量原理图

中央控制器中的仲裁控制模块采用非阻塞的方式扫描一个向量中的每个元素,一个向量扫描完毕后将其清零;由于向量之间的无关性,可以并行的对 N 个向量同时进行扫描,以提高扫描的速度.仲裁控制模块在扫描的过程中仍然会有新的流量产生,因此其会接受新的流量通知数据包,并持续更新向量中的元素值,以保证灵活性和实时性.

在对每个向量扫描时,会挑选出此向量中数值最大的元素,即找到与本机柜累积流量最多的机柜;然后控制对应的 TWC 在两个机柜之间建立一条持续的光线路,使这些流量在光域网络上传输,仲裁控制模块会将建立光线路的请求发送给执行单元,由执行单元控制 TWC 进行波长变换,在机柜之间建立起一条持续的光线路.

执行单元建立链路时,会通知对应机柜内的所有服务器的虚拟队列管理模块,让其开始发送分配到光域网络上的网络流.交换机持续监视每条光线路上传输的网络数据量,如果单位时间内传输的数

据量持续下降，使光线路的利用率低于一个阈值 T 时，则认为本次数据流的传输已经结束，需要重新的计算、分配光线路.

在扫描一个向量的过程中，只选取最大累积值对应的元素作为线路建立的依据，而对于同一个向量中的其他元素，其所对应的服务器中的流需要在电域网络中传输，而不再需要中央控制器的调度和控制. 仲裁控制模块在找到最大累积值后，会将最大累积值对应的元素编号发送到本向量对应的机柜中的所有服务器，以此告知虚拟队列管理系统，会在这两个机柜之间建立一条持续的光链路. 而对于目的地址是其他机柜的网络流，虚拟队列管理系统会直接让操作系统调度发送.

由于中央控制器控制所有大流的发送，仲裁控制模块对向量的扫描间隔直接导致网络的延迟，有两种事件可以触发仲裁控制模块的扫描操作. 第一种是定时器，仲裁控制模块维护了一个扫描定时器，每隔 200 μs 会触发一次扫描操作；此时的光线路可能正处在忙碌状态，扫描产生的结果会被暂存，如果到下次定时器触发时此扫描结果仍未被使用，则会被新的扫描结果所取代. 第二种是光线路的利用率接近阈值 T 且现在并没有可用的扫描结果时，会触发一次新的扫描过程以建立新的光链路；这种情况主要发生在累计数据量较少时，光线路在 200 μs 内将数据传输完毕，其线路建立周期小于向量扫描周期.

在系统中可能产生仲裁的“轮空现象”，即本应在光域网络传输的数据却在电域网络中传输，产生此现象主要有 3 种原因. 第一，执行模块在通知虚拟队列管理模块开始发送数据时，由于系统采用 UDP 作为通知控制协议，虽然采用了重复数据包的方法，仍可能会发生控制包丢失的问题. 因此虚拟队列中的定时器超时之后才会发送对应的网络流，光线路在规定时间内的利用率并未达到阈值 T，从而重新建立光链路. 第二，由于控制网络可能产生拥塞，控制数据包到达时间延迟较高，同样也会引起虚拟队列管理模块的超时；第三，由于交换机持续监视每条光线路上传输的网络数据量，如果光线路的利用率低于一个阈值 T，则会重新仲裁建立光链路，本应在光域网络上传输的网络流仍然有部分数据未完成传输，这些数据会在电域网络中传输.

4.4　流量分配

执行单元会根据仲裁控制单元的处理结果，为两个机柜之间建立一条持续的光链路，并通知服务器端开始发送数据. 因此在 ToR 交换机中必须采用相应的方法将数据包有区别的分别放到电域网络和光域网络上传输. 在 Ace-net 中，使用 OpenFlow[23] 对 ToR 交换机的数据包转发路径进行控制.

流量分配工作由运行在中央控制器中的执行单元和支持 OpenFlow 的 ToR 交换机共同完成. 执行单元为两个机柜建立光链路后，会向对应的 ToR 交换机发出数据流表更新的命令，将添加的路由规则通知 ToR 交换机，使其更新数据流表；如果需要切断两个机柜之间的光链路，同样会通过执行单元向对应的 ToR 交换机发出数据流表更新的命令，对其中的数据流表进行更新. 为了避免光链路建立、切断过程中的丢包问题，对交换机中数据流表采用保守的更新操作：光链路建立时，待建立的光链路稳定后，再更新数据流表；光链路切断时，首先要更新数据流表，然后才能调整光链路.

ToR 交换机中的数据流表更新完成后，会根据流表表项描述的转发规则将经过交换机中的流量分配到对应的光域网络和电域网络上，完成流量分配工作.

5　系统评测

在 Ace-net 的性能评测中，采用了为评测混合网络性能所设计的离散事件网络模拟器 Ace-sim. 为了模拟一个接近于真实的网络环境，在其中构建了一个网络数据生成模块来近似模拟应用程序和操作系统的行为. 同时，在模拟器中集成了网络协议栈和支持 OpenFlow 的交换机. 模拟器中网络的拓扑结构、延迟、带宽等参数可以动态配置.

本文着重于对光域网络结构和工作原理的评测，对于电域网络的结构并没有限制，在评测中搭建了一个如图 6 所示的树形三层电域网络架构，具体配置参数如表 1 所列.

表 1　网络拓扑结构参数

参数	参数值	单位
节点数	20480	个
机柜数	512	个
节点/机柜	40	个
电域链路	1 和 10	Gbps
光域链路	20	Gbps
AWGR 端口	512	个

在模拟网络中共有了 512 个机柜，其中每一个

机柜中放置40台服务器,系统中一共有20480个节点.每一台ToR交换机有40个1Gbps的下行端口用于与服务器相连,一个10Gbps的上行电域端口与聚合层交换机相连,一个20Gbps的上行端口用于连接TWC光器件,假设每一个TWC输出的单波长光束的带宽为20Gbps.每一台聚合层交换机有33个10Gbps的端口,其中32个下行端口用于与ToR交换机相连,一个上行端口用于与核心层交换机相连;核心层交换机共有16个10Gbps的端口,都用于与聚合层交换机相连.光电混合系统中使用512×512端口的AWGR光交换机,分别与512个机柜中的光器件相连.

为了与纯电域网络的性能以及基于MEMS光交换机的混合网络的性能作比较,我们参照c-Through的网络结构,在Ace-net中实现了基于MEMS交换机的光电混合网络结构.在实验中使用注入率来描述服务器节点上数据量产生的强度,其定义为每台服务器每秒钟产生的平均数据量.

5.1 平均包延迟

此小节对网络系统的平均包延迟进行评估,数据包的平均延迟反映了Ace-net所使用的策略对网络性能的影响程度.

根据现在数据中心网络流量的特点,按照一定的近似比例生成网络流.其中每一台服务器以2%的概率随机生成数据量达到20MB网络流,以5%的概率生成数据量达到10MB的网络流,以10%的概率生成数据量1MB至10MB的网络流,以15%的概率生成64KB到1MB的网络流,以68%的概率生成小于64KB的网络流.在模拟中,假设每一个网络流都随机地发送到系统中的某一台服务器.在模拟中,将控制延迟定义为:从服务器向中央控制器发出链路申请到获得中央控制器反馈这段时间;将光域网络的控制系统延迟设置为30μs,光线路的利用率接近阈值T设置为5%.

从图9平均包延迟测试结果可以看出,在随机注入模式下,Ace-net混合网络的平均包延迟要明显小于纯电域网络和基于MEMS交换机的光电混合网络的平均包延迟.在注入率达到0.2Gbps时,Ace-net的平均包延迟只有纯电域网络的0.6倍;而随着注入率的增加,其平均的包延迟缓慢的增长,不仅远低于纯电域网络结构,而且低于基于MEMS交换机的光电混合网络.

这主要有三方面的原因:首先,Ace-net利用了快速波长变换器件的纳秒级波长变换特性,在机柜之间迅速地建立路径,路径切换速度优于基于MEMS光交换机的光电混合网络.其次,Ace-net仍然是对数据量较大的网络流进行加速,在两机柜之间建立的是一条持续时间较长的光线路,每一次链路的建立都会传输数十兆的数据;虽然有控制延时的存在,但由于传输的数据量较大,平均包延迟并没有显著地升高;最后,Ace-net在两个机柜的ToR交换机之间建立桥接的光线路,数据包的传输不需要经过聚合层和核心层的电域交换机,缩短了数据包的传输路径.

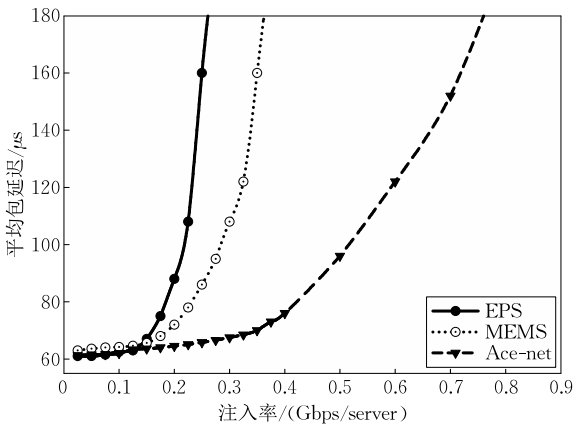


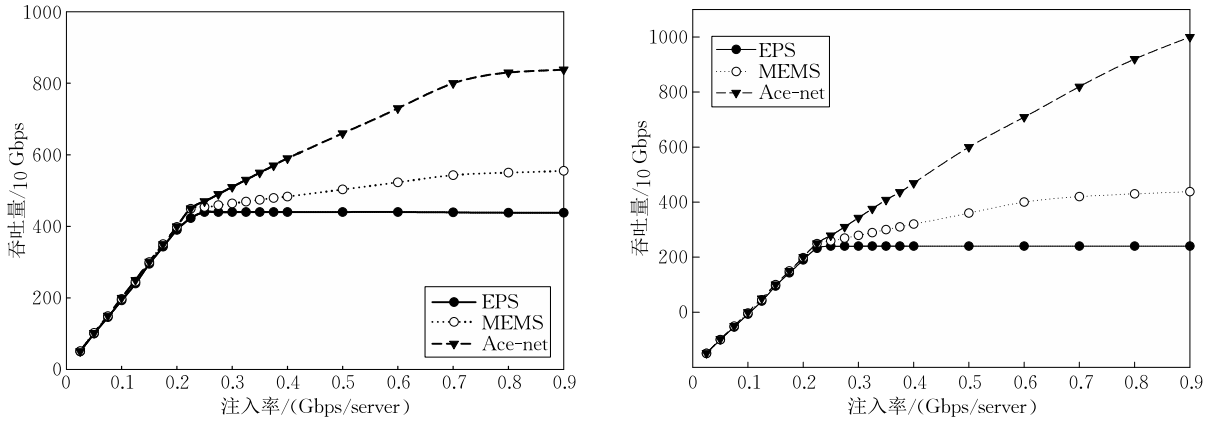
图9 平均包延迟测试结果

5.2 吞吐量

在吞吐量测试中,以网络流的大小作为变量,分别测试了3种网络的吞吐量指标.在测试中选择两种不同数据量的网络流作为对比,使用数据量分别为200KB和10MB的网络流作为测试集,其中测试中随机生成网络流的目的地址.

图10所示为采用不同数据量网络流的吞吐量的测试结果,从两图中可以看出注入不同数据量的网络流对网络的吞吐量有非常大的影响:当网络流的大小为200KB时,纯电域网络和基于MEMS光交换机的混合网络的性能只有18%左右的差距,而Ace-net的性能要显著高于其他两种网络,甚至可以达到纯光域网络性能的两倍.当注入数据量达到10MB的网络流时,Ace-net的吞吐率的优势更加的明显.

对比图10(a)、(b)两图可以发现,增大网络流的大小,混合网络系统的性能会显著增加,网络流的大小与网络的性能在一定范围内呈正相关.大的网络流使光网络中有更持续的网络包传输,使建立的线路更加稳定持续,缩短了由于仲裁、控制延迟引起的光线路空闲的时间.



(a) 200 KB的网络流通信吞吐量结果　　(b) 10 MB的网络流通信吞吐量结果

图 10　采用不同数据量网络流的吞吐量结果

5.3　控制延迟影响

在 Ace-net 系统中，控制延迟对混合网络系统的性能有较大的影响. 为了防止出现由于控制数据包丢失或者中央控制器故障等原因引起的中央控制器长时间不响应虚拟队列的链路请求的情况，在通知中央控制器的同时，虚拟队列会为虚拟队列中的这一项开启定时器，当 300 μs 内仍未得到中央控制器的响应，虚拟队列会将这一项管理的到对应机柜的网络流直接发送. 在执行模块通知虚拟队列管理模块开始发送数据时，由于数据包丢失或者控制延迟较高，虚拟队列中的定时器已经超时，而数据流已经开始发送. 因此这些数据流量可能会在电域网络上传输而未在光域网络中传输，使本次建立的光链路并未完整加速网络数据的传输，从而降低光域网络的利用率. 由于系统采用 UDP 作为通知控制协议，虽然采用了重复数据包的方法，仍可能会发生控制包丢失或者延迟过高等问题，由此可能会发生“轮空现象”.

在评估中，定义“轮空比率”为：在仲裁模块为两机柜分配光线路后，两机柜之间本应在光线路上传输的网络流量而实际未在光线路上传输的网络流量占这段时间内两机柜之间所有流量的百分比. 即由于控制延迟和控制机制的问题，有部分分配到光域网络流会在电域网络上传输.

设定每台服务器的注入率为 0.2 Gbps，控制延迟变化区间为 10 μs 至 320 μs，定时器的值设置为 300 μs，UDP 数据包的丢包率设置为 1%. 根据现在数据中心网络流量的特点，按照一定的近似比例生成网络流. 其中每一台服务器以 2%的概率随机生成数据量达到 20 MB 网络流，以 5%的概率生成数据量达到 10 MB 的网络流，以 10%的概率生成数据量 1 MB 至 10 MB 的网络流，以 15%的概率生成 64 KB 到 1 MB 的网络流，以 68%的概率生成小于 64 KB 的网络流；在模拟中假设每一个网络流都随机地发送到系统中的某一台服务器.

首先需要评估在给定注入率时，控制延迟对轮空比率的影响，在模拟中不断的增加混合网络的控制延迟，绘制的轮空比率曲线如图 11 所示.

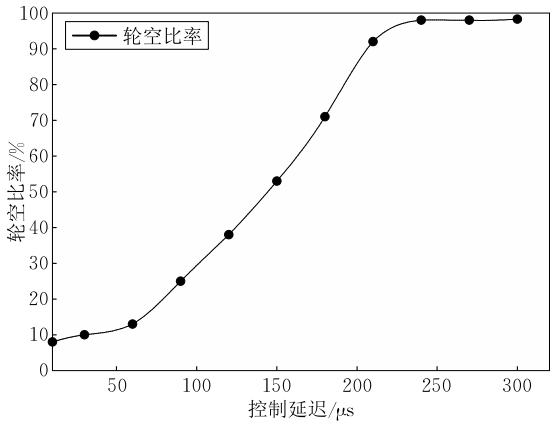


图 11　控制延迟与轮空比率关系

从图 11 不同控制延迟的轮空比率结果中可以看出，当控制延迟大于 60 μs 时，轮空比率开始大幅上升；在 230 μs 时轮空比率达到 98%，光网络上只有极少的数据传输，控制系统基本失效. 因此在实际系统中，必须要降低控制延迟的开销，以提高系统的性能. 微秒级延迟的点对点通信系统已经商用，如 Infiniband 网络可以实现低于 1 μs 的点对点通信延迟[①]；同时，由于控制系统的算法简单，可以使用全硬件的中央控制系统，从而将系统的控制延迟控制在 20 μs 以下.

① Mellanox IB. http://www.mellanox.com/page/infiniband_cards_overview

其次需要评估在不同的丢包率的情况下,对轮空比率的影响.在控制系统中为了降低延迟,通知机制使用 UDP 协议来实现,虽然采用了重复数据包的技术,但是在丢包率较高的情况下仍然会对混合网络的性能产生影响.在测试中设置控制延迟为 30 μs,定时器的设置值为 300 μs,轮空比率与丢包率的关系如图 12 所示.

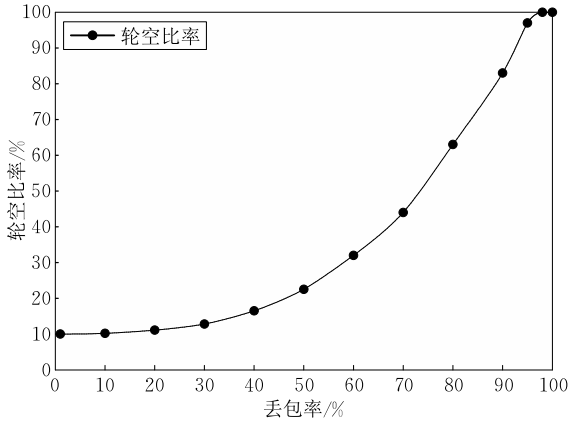


图 12 UDP 丢包率与轮空比率关系

从图 12 所示的 UDP 丢包率与轮空比率的关系中可以发现,只有当丢包率非常高的情况下才会引发较高的轮空比率;只要将 UDP 数据包的丢包率控制在 30%以下,对系统的性能影响有限.而在实际的网络系统中比较容易实现将丢包率控制在 30%以下,因而采用 UDP 协议设计的通知机制是可行和高效的.

6 结束语

本文提出了一种使用 TWC 光器件和 AWGR 光路由器的 OCS/EPS 光电混合网络结构 Ace-net.相比于现有的混合网络结构,此结构使用可快速变换波长的 TWC 光器件和光波长路由芯片 AWGR 在两个机柜间建立起一条光链路,实现纳秒级的链路切换,使更多的网络流可以在高带宽的光域网络中传输,进一步缓解了现在数据中心网络系统通信带宽低、交换容量有限、高能耗等缺点.在文中详细描述了 Ace-net 的整体设计以及流量需求测量、仲裁控制、流量分配等各个子系统的运作方式;同时使用模拟器对此结构进行了评测.实验结果表明,相对于传统的电域网络结构,此结构能够很好地降低网络中平均包延迟,提高网络的吞吐率.但由于光信号直接存储、灵活控制的问题在物理上仍然没有解决,光域网络对数据量较小的网络流(如小于 64 KB)的应对能力较差,如何构建更高效的光电混合网络系统仍然是一个需要进一步探索的问题.

Ace-net 中所使用的光器件在现阶段都是成熟可用的,在本文中我们主要探索混合网络结构和控制机制,因此只使用模拟器来评估 Ace-net 的性能特征.在未来我们将采用 FPAG 以及真实的 AWGR 光路由器构建原型系统,进一步验证系统的工作机制和实际性能.

一方面,现在大型数据中心内部,10 GigE 已经基本普及,并且正在向 40 GigE 和 100 GigE 的方向迈进.另一方面,随着大数据等新型应用的发展,数据中心的规模不断扩大.上述两方面的因素正在加速推进数据中心网络的变革.高带宽的光域网络成为现在数据中心网络的一个重要选择,混合网络结构也将是数据中心网络发展的一个重要方向.

致 谢 感谢各位审稿人为本文的完善所提出的有益建议!

参 考 文 献

[1] Chen Kai, Hu Cheng-Chen, Zhang Xin, et al. Survey on routing in data centers: Insights and future directions. IEEE Network, 2010, 25(4): 6-10

[2] Deng Gang, Gong Zheng-Hu, Wang Hong. Characteristics research on modern data center network. Journal of Computer Research and Development, 2014, 51(2): 395-407(in Chinese)
(邓罡, 龚正虎, 王宏. 现代数据中心网络特征研究. 计算机研究与发展, 2014, 51(2): 395-407)

[3] Jonathan K. Worldwide electricity used in data centers. Environmental Research Letters, 2008, 3(034008): 8

[4] Dennis A, Marty M R, Wells P M, et al. Energy proportional datacenter networks. ACM SIGARCH Computer Architecture News, 2010, 38(3): 338-347

[5] Barroso L A, Hölzle U. The case for energy-proportional computing. Computer, 2007, 40(12): 33-37

[6] Kachris C, Tomkos I. A survey on optical interconnects for data centers. IEEE Communications Surveys and Tutorials, 2012, 14(4): 1021-1036

[7] Wang Guo-Hui, Andersen D G, Kaminsky M, et al. c-Through: Part-time optics in data centers. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 327-338

[8] Nathan F, Porter G, Radhakrishnan S, et al. Helios: A hybrid electrical/optical switch architecture for modular data centers. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 339-350

[9] Chen Kai, Singla A, Singh A, et al. OSA: An optical switching architecture for data center networks with unprecedented flexibility. IEEE/ACM Transactions on Networking, 2014, 22(2): 498-511

[10] Theophilus B, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild//Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. Melbourne, Australia, 2010: 267-280

[11] Srikanth K, Sengupta S, Greenberg A, et al. The nature of data center traffic: Measurements and analysis//Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. Chicago, USA, 2009: 202-208

[12] Pina J, Silva H, Monteiro P, et al. Performance evaluation of wavelength conversion at 160 Gbit/s using XGM in quantum-dot semiconductor optical amplifiers in MZI configuration// Proceedings of the 2007 Photonics in Switching. San Francisco, USA, 2007: 77-78

[13] Cheung S, Su Tie-Hui, Okamoto K, Yoo S J B. Ultra-compact silicon photonic 512×512 25 GHz arrayed waveguide grating router. IEEE Journal of Selected Topics in Quantum Electronics, 2014, 20(4): 310-316

[14] Cao Z, Proietti R, Yoo S. Hi-LION: Hierarchical large-scale interconnection optical network with AWGRs. Journal of Optical Communications and Networking, 2015, 7(1): A97-A105

[15] Truex T, Bent A A, Hagood N W. Beam steering optical switch fabric utilizing piezoelectric actuation technology// Proceedings of the National Fiber Optic Engineers Conference. Orlando, USA, 2003: 203-208

[16] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 63-74

[17] Guo Chuan-Xiong, Wu Hai-Tao, Tan Kun, et al. Dcell: A scalable and fault-tolerant network structure for data centers. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 75-86

[18] Guo Chuan-Xiong, Lu Guo-Han, Li Dan, et al. BCube: A high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 63-74

[19] Zang Da-Wei, Cao Zheng, Wang Zhan, et al. Decentralized NIC-switching architecture using SR-IOV PCIe network device. IEEE Micro, 2014, 34(5): 52-56

[20] Liu He, Lu Feng, Forencich A, et al. Circuit switching under the radar with reactor//Proceedings of the ACM/USENIX Symposium on Networked Systems Design and Implementation. Seattle, USA, 2014: 1-15

[21] Ye Xiao-Hui, Yin Ya-Wei, Yoo S J B, et al. DOS: A scalable optical switch for datacenters//Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems. San Diego, USA, 2010: 24

[22] Hida Y, Hibino Y, Kitoh T, et al. 400-channel arrayed-waveguide grating with 25 GHz spacing using 1.5%-Δ waveguides on 6-inch Si wafer. Electronics Letters, 2011, 37(9): 576-577

[23] McKeown N, Anderson T, Balakrishnan H, et al. OpenFlow: Enabling innovation in campus networks. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 69-74

ZANG Da-Wei, born in 1988, Ph. D. candidate. His research interests include computer architecture and datacenter networks.

CAO Zheng, born in 1982. Ph. D., associate professor. His research interests include high performance computer architecture, high performance interconnection, and optical interconnection.

WANG Zhan, born in 1986, Ph. D. candidate. His main research interests include virtualization technology and high performance interconnection networks.

LIU Xiao-Li, born in 1986, M. S., assistant engineer. Her major interests focus on I/O virtualization and high performance interconnection networks.

FU Bin-Zhang, born in 1983, Ph. D., associate professor. His research interests include high performance computer architecture, high performance computing and computer network.

SUN Ning-Hui, born in 1968, Ph. D., professor, Ph. D. supervisor. His main research interests include computer architecture, high performance computing and distributed OS.

Background

Nowadays, the requirements of cloud services are increasing rapidly. In order to meet the needs of large-scale data exchange among servers, data center network faces enormous challenges. In traditional data center, the switching equipment typically used electronic switching technology which is fast and flexible to switch packets between ports. But it inherently has low communication bandwidth and limited switching capacity. In order to reduce the cost and complexity of the industrial deployment, there is a serious oversubscription among layers. The aggregation layer and core layer switches become the bottleneck of network communication, which results in the long delay of data transmission. At the same

time, the energy consumption of the network is an important part of the datacenters' power consumption because of the multi-stages O-E-O (optical-electronic-optical) conversion. Our aim is to improve the performance of the datacenter network and increase its flexibility.

Optical interconnects have gained attention recently as a promising solution offering high throughput, low latency and reduced energy consumption compared to current networks. At present, the research on optical network focuses on two topics. One is using the MEMS switch to construct an OCS (Optical Circuit Switching) network. For example, a hybrid electrical-optical network has been presented by Wang G. et al. called c-Through. But, the circuit switch network is inflexible, which can only provide a matching on the graph of racks and the reconfiguration time is rather high. The other is using the AWGR (Arrayed-Waveguide Grating Router) to construct an OPS (Optical Packets Switching) network. For example, Ye X et al. present the DOS architecture which allows contention resolution in the wavelength domain. But the DOS architecture resolves the header of each packet and do not eliminate the O-E-O (optical-electronic-optical) conversion.

In the paper, we propose a hybrid network architecture using AWGR optical device, called Ace-net. We present the design and the key characteristics of Ace-net, which include traffic demand estimation, arbitration control and traffic distribution. Also, we evaluate this structure using a simulator and the simulation results show that the network structure has good performance.

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 61572464, 61331008. And it is supported Partly by the National High Technology Research and Development Program (863 Program) of China under Grant No. 2015AA01A301. As well, the authors gratefully acknowledge support from Huawei Technologies Co. Ltd. under Grant No. YB2015070066.