

基于 SDN 的光电混合数据中心流量调度方法

张天亿, 刘逢清

(南京邮电大学 光电工程学院, 南京 210023)

摘要: 面对大量的东西向流量, 过去采用的电域交换无法有效适应流量拥塞的问题, 当前光电混合数据中心所采用的基于路径切换的光交换结构无法适应突发的实时流量。为此, 在数据中心内采用快速可调波长激光器(TWC)和光波长路由器(AWGR)来适应流量模型, 并基于软件定义网络(SDN)技术, 采用统一的控制平面, 感知全网状态信息, 下发扩展的流表项动态实现流量的分类识别调度。结果表明, 该设计方法能够有效改善数据中心网路拥塞的问题。

关键词: 软件定义网络; 光波长路由器; 数据中心; OpenFlow

中图分类号: TN915 **文献标识码:** A **文章编号:** 1002-5561(2018)04-0025-04

DOI: 10.13921/j.cnki.issn1002-5561.2018.04.007

Traffic scheduling method in hybrid optical-electronical data centers based on SDN

ZHANG Tianyi, LIU Fengqing

(School of Optoelectronic Engineering,

Nanjing University of Posts & Telecommunications, Nanjing 210023, China)

Abstract: The electric switching adopted in the past cannot meet the demands of transporting large numbers of east-west flows and face the problems of traffic congestion. The present MEMS-based optical switches in data center (DC) also cannot adapt to highly bursty real-time traffic. This paper uses the tunable wavelength converters(TWC) and arrayed waveguide grating router(AWGR) to quickly adapt to changes of network traffic. Based on the software defined network (SDN) technology, a unified control plane is used to find out the whole network state information, and the extended flow table items are used to realize the classification, identification and scheduling of traffic. The results show that the design method can effectively improve the congestion performance of data center networks.

Key words: SDN; AWGR; DC; OpenFlow

0 引言

近年来云计算技术广泛铺开, 数据中心作为承载云计算服务的基础设施也随之成为企业和科研社区的研究热点。现有数据中心通常采用电域交换来实现机架内部的网络连接, 在机架上采用光纤连接到汇聚层交换机, 但面对大量东西向流量时会出现局部瓶颈。常见的等价多路径转发方法虽然在一定程度上缓解了流量拥塞, 但是面对突发流量仍有局限性。

根据数据中心流量特征, 本文提出一种设计方

案, 在原有胖树结构中增加基于波导阵列光栅(AWGR)的大容量光波长快速路由, 同时采用电分组交换实现大量突发分组、小数据流对链路带宽的共享和利用。软件定义网络(SDN)^[1]已由学术圈研究转向成熟的商业应用。基于SDN技术控制和转发分离的特点, 本文在控制器端建立全网拓扑和流量监控, 实时采集网络的流量, 动态划分大小流量, 将判决门限和大流量的路由策略流表项周期下发到交换机上, 实现流量的动态调度。

1 研究现状

1.1 数据中心的流量特征

早期 80%的数据中心流量为承载用户和服务

收稿日期: 2017-12-05。

基金项目: 江苏省科技计划项目(BY2016010-01)资助。

作者简介: 张天亿(1993-), 男, 硕士研究生, 主要研究方向为光通信技术、SDN虚拟化。

交互的南北向流量,满足对大容量存储数据的访问需求。而如今数据中心作为云计算基础设施,大量的流量属于内部的东西横向流量^[1],只有 20% 的流量通往外部,这种流量变化是因为云平台的分布式计算存储业务是由服务器集群协同完成的^[2]。进一步分析可以发现超过 90% 的流量是小于 1MB 的具有高突发性的小流量,98% 的流量不大于 100MB^[3]。此外,研究表明现有数据中心链路利用率偏低,核心层链路的利用率少于 25%。考虑普遍的情况,我们用 2 个 90 来描述数据中心流量模型,即 90% 的流量是小于 1MB 的小流量,90% 的数据流量集中在大于 100MB 的数据流中。

这种流量分布会造成单一路径上大流量对小流量的阻塞而其它可用路径处于闲置的问题,目前多采用胖树 Fattree 网络拓扑配合等价多路径转发策略^[4]来应对。这种方案在节点对间提供多条路径,然而不同路径的带宽时延和可靠性并不一致,而且该方案一般通过静态配置,不能适应动态的流量变化。

1.2 光交换技术

传统光交换中光电转换会增加大量时延,无法满足数据中心的时延要求。光无法缓存,而纯光网络都是点到点结构或环路拓扑,因此我们采用光电混合网络,数据可以选择光路由传输,也可以通过电交换传输。现有的慢速路径切换光交换的重配置时延过高,基于微电子机械系统 (Micro-Electro-Mechanical System, MEMS) 的光交换机采用机械方式控制光链路的建立会增加数十毫秒的延迟,无法适应数据中心中越来越复杂的突发流量变化。

AWGR 路由器^[5]是一种无需配置、根据光波长进行快速路由的无源光器件,它根据循环波长路由特性,将不同波长的光束路由至同一输出端口,这些光束复用在同一光纤上不发生竞争。6×6 端口 AWGR 如图 1 所示,6 个输入端口 1 分别输入 6 个波长信号: $\lambda_1^x, \lambda_2^x, \lambda_3^x, \lambda_4^x, \lambda_5^x, \lambda_6^x$,经过 AWGR 分别输出到不同端口,来自不同端口的输入光波也能够输出到同一端口,复用在同一光纤中。因此,我们只需要利用快速可调波长激光器,将输入光信号变成给定的波长输出,其延迟在皮秒至纳秒级,可以满足数据中心低时延高带宽的光互连要求。

现有的可调谐波长转换器 (Tunable Wavelength Converters, TWC) 已经实现 160Gb/s 的波长转换带宽,延迟增长控制在 30ns 内。此外,AWGR 可以实现的交换端口数量较高,加州大学的团队已将 512 端口的

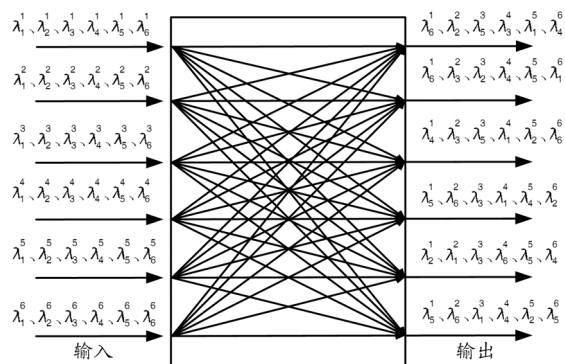


图 1 6×6 端口 AWGR 结构图

AWGR 成功流片^[6]。我们将每个架顶交换机的上行端口均与 AWGR 相连,1 个机顶交换机 (Top of the Rack, ToR) 下平均放置 40 台服务器,那么采用 512×512 端口的 AWGR 可以满足数据中心内约两万台服务器的连接需求。

1.3 SDN 技术

基于 OpenFlow 的 SDN^[7]概念最初是斯坦福大学 Nick 教授团队提出的网络虚拟化技术。通过将资源虚拟化的思想引入到网络中,将网络分为应用层、控制层和数据转发层。在底层的数据平面,交换机只负责高速转发,不考虑智能功能。控制器利用南向接口 OpenFlow 流表协议^[8]实现交换机和控制器之间的信息交互,下发流表到交换机实现各种路由策略。上层应用通过北向接口与控制器交互,可以更快地下发各种增值业务,感知网络状态部署策略,进行维护和管理。我们通过改写控制器模块增加了感知流量并分类,以实现分流的功能来适应数据中心流量特征。

2 方案设计

2.1 流量调度策略

本文在胖树结构的基础上增加了能够高速转发的光链路,设计了集中流量调度管理机制,交换机不计算路由,只根据流表中的判决门限将大流量由光纤输送到 AWGR 路由,而小流量由电交换处理。数据流门限和调度都是由 SDN 控制器决定的,具体工作流程如图 2 所示。

初始化时,交换机首先通过链路层发现协议 (Link Layer Discovery Protocol, LLDP) 上传拓扑信息和链路能力给控制器,控制器下发 100MB 为大流量的初始转发策略,100MB 的流量会走 AWGR 光路由,而 10MB 的流量会走汇聚层交换机。当虚拟机迁移的大数据流造成链路利用率接近 90% 时认为拥塞,控制器重新计算门限,以链路利用率较高链路上待调度流量的加权

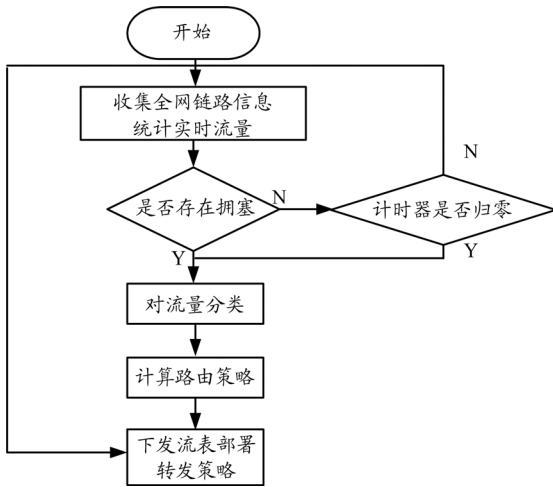


图2 流量调度策略流程图

平均值作新门限,得到符合当前网络状态的新门限,从而将造成拥塞的流量切换到光链路上去。同时在控制器中设定了计时器,即使未发生拥塞,也会定时选取链路利用率较高链路上的流量进行重新计算。

2.2 控制器设计

SDN 三层平面如图3所示,本方案选取适合仿真设计的模块化控制器 RYU。方案实现主要涉及拓扑发现模块、流量感知模块、路径计算模块和流表模块。

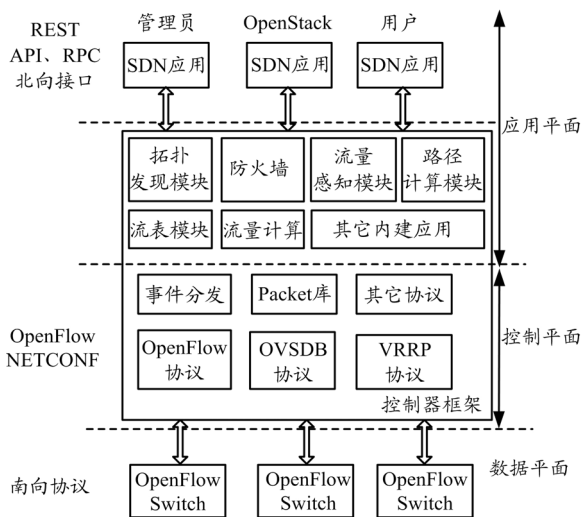


图3 SDN 三层网络模型

在拓扑发现模块中,控制器通过流表的Hello报文与Open vSwitch虚拟交换机建立连接协商版本号;通过Feature报文得到交换机的交换能力;通过LLDP获得网络拓扑,构建全网的拓扑图表。

在路径计算模块采用生成树协议以跳数和链路状态做参数计算最短路径。本文中大流量和小流量的路由不同,因此在计算模块中,只需计算小流量的路径,大流量通过调制到不同波长即可实现光链路转发。

我们的流量调度策略是基于流量感知分类的,在流量感知模块内,我们设计让控制器向交换机通过Openflow报文定时轮询流量,计算链路利用率 η_i 。对链路 i 上 m 个流量 B 求和后除以链路最大带宽 $Maxload$ 得到利用率。

$$\eta_i = \frac{\sum_{j=1}^m B_j}{Maxload} \quad (1)$$

选取链路利用率最高的 n 条(本文中 $n=10$)链路计算权值 a_i , B_{imax} 代表链路 i 上最大的流量,流量越大所占比重最大。

$$a_i = \frac{B_{imax}}{\sum_{i=1}^n B_{imax}} \quad (2)$$

由式(3)计算出新门限交由流表模块下发。调节因子 k 取值为0.9以取得低于大流量的门限值 s 。

$$s = k \sum_{i=1}^n a_i B_{imax} \quad (3)$$

一般数据中心内链路的流量并不平均,同一租户的虚拟机间会交换较多流量,而不同租户之间的流量较少。因此,我们在仿真中假设计算拥塞链路时所取样的链路利用率较高的链路中,所占比重较大的流量均为同一租户虚拟机之间产生的交互流量以及虚拟机迁移所造成的长时大流量。通过选取这些高利用率的链路计算门限值仍较为粗略,在实验条件允许的情况下可以进一步在控制器中设计根据不同租户的虚拟机分布情况去统计端到端流量。

3 实验仿真结果分析

本文在传统数据中心胖树结构的基础上增加了快速交换的光链路,并采用SDN控制器对流量感知分类采取不同传输策略的方案进行了仿真实验分析。与等价多路径转发(ECMP)策略进行对比,本文设计的改进的流量调度方案在吞吐量和端到端时延方面具有显著改善。

3.1 实验环境搭建

本文使用Mininet对数据中心进行仿真,搭建的Fat-tree网络如图4所示,链路带宽设置为10Gb/s。底层接入层每台TOR柜顶交换机接入两台主机,TOR同时与汇聚层交换机和AWGR建有链路连接,两台顶层核心交换机与每一台汇聚层交换机均建立链路。交换机之间均使用光纤连接,AWGR与TOR之间亦使用光线互联。小数据流由TOR转发到汇聚层交换机再经由

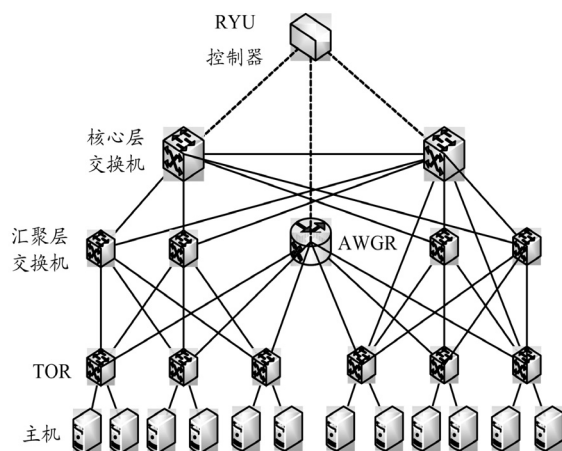


图4 仿真拓扑结构

核心层交换机路由,采用传统的胖树路由转发模式。数据流在TOR处根据流表被调制到光载波上,通过AWGR路由到相应的输出端口沿光纤传输到目的TOR。

仿真实验选取RYU作为控制器,改写了部分模块,并下发了默认流表项。根据数据中心流量模型,改写iperf工具,设置节点发送数据包大小分别为1kB、10kB、1MB、10MB和100MB时的概率依次为0.5、0.3、0.1、0.05和0.05。添加命令iperfmulti(Num),该函数接收发送数据包数目作为参数,数目越大则模拟的网络负载越大。通过random函数选择节点列表里的其它主机作为iperf的服务器端,并以客户端身份按照指定参数发送TCP和UDP流,我们采用Python脚本利用正则表达式根据抓包结果计算出网络吞吐量,统计不同负载情况下的网络状态。

3.2 性能分析

ECMP和改进方案的时延对比图如图5所示,单个主机的发送数据量代表网络负载(横坐标),纵坐标为主机的端到端时延,通过增加发包数量来增加负载。可以看出,随着网络负载增大,端到端时延相应增加,负载较小时两方案的时延接近;当负载增大时,ECMP的时延增加明显更高。这是由于ECMP静态的流量均衡,无法检测当前链路负载,只能固定地配置流量,而本方案将大流量分流,在负载增加时,放宽门限,分流更多流量,时延没有明显增大。

不同负载下吞吐量对比图如图6所示。随着网络负载即主机发送数据量的增加,网络吞吐量相应增加。峰

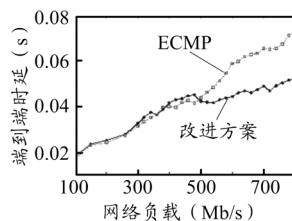


图5 分组端到端时延对比图

值时,采用改进方案的网络吞吐量比ECMP高10%。通过实时感知网络的负载情况,SDN控制器能够及时将更多的流量由光交换转发从而增加了网络的吞吐量,而传统多路径转发很快就达到了自己的吞吐量极限。

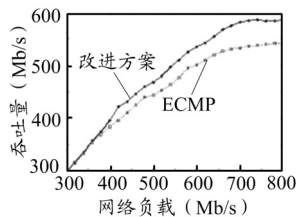


图6 不同负载下吞吐量对比图

4 结束语

本文针对数据中心流量拥塞问题,提出了一种基于SDN技术的光电混合数据中心流量调度方案,设计了光电混合数据中心拓扑结构,在胖树形拓扑的基础上利用AWGR路由器部署额外的光链路,采用扩展的OpenFlow流表进行控制,将跨机架的大流量由AWGR路由,避免链路中长时间占用带宽的大流量对时间敏感的小流量造成阻塞。在SDN控制器感知流量的基础上将流量分类,确定了不同流量的调度策略,并下发流表到交换机上。通过实验仿真验证了光电混合的拓扑结构下实施分流调度的策略相比传统的ECMP策略有更好的对分带宽和时延。

参考文献:

- [1] KREUTZ D, RAMOS F M V, ESTEVES V P, et al. Software-Defined Networking: A Comprehensive Survey [J]. Proceedings of the IEEE, 2014, 103(1): 10-13.
- [2] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. Open-Flow:enabling innovation in campus networks[J]. Acm Sigcomm Computer Communication Review, 2008, 38(2): 69-74.
- [3] GREENBERG A, HAMILTON J R, JAIN N, et al. VL2: a scalable and flexible data center network [J]. Acm Sigcomm Computer Communication Review, 2009, 39(4): 51-62.
- [4] BENSON T, ANAND A, AKELLA A, et al. Understanding data center traffic characteristics [J]. ACM Sigcomm Computer Communication Review, 2010, 40(1): 92-99.
- [5] YIN Y, PROIETTI R, YE X, et al. LIONS: An AWGR-Based Low-Latency Optical Switch for High-Performance Computing and Data Centers [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2013, 19(2): 3600409-3600409.
- [6] CHEUNG S, SU T, OKAMOTO K, et al. Ultra-Compact Silicon Photonic 512×512 25GHz Arrayed Waveguide Grating Router[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2014, 20(4): 310-316.
- [7] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. Open-Flow:enabling innovation in campus networks[J]. Acm Sigcomm Computer Communication Review, 2008, 38(2): 69-74.
- [8] 孙三山,汪帅,樊自甫. 软件定义网络架构下基于流调度代价的数据中心网络拥塞控制路由算法[J]. 计算机应用, 2016, 36(7): 1784-1788.