云计算数据中心光互连网络:研究现状与趋势

余晓杉[1)] 王 琨[2)] 顾华玺[1)] 王 曦[1)]

[1)](西安电子科技大学 ISN 国家重点实验室 西安 710071)

[2)](西安电子科技大学计算机学院 西安 710071)

摘 要 数据中心是云计算的核心支持平台,云计算的发展对数据中心网络架构提出了严峻的挑战,传统电互连网络架构难以在带宽、设备开销、能耗、管理复杂度等方面同时满足云应用的要求,因此以低能耗、低开销、高带宽为特点的光互连网络架构出现并受到研究人员的广泛关注.该文在对比电互连技术和光互连技术基本特点的基础上,对数据中心光互连网络的研究现状展开综述,具体介绍了现有的光电混合网络、集中式全光网络和分布式全光网络,并从交换机制、扩展性、技术可行性、设备开销等方面对上述网络架构进行了对比分析.最后该文对数据中心光互连网络的未来研究趋势进行了总结和展望.

关键词 云计算;数据中心网络;光互连技术

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2015.01924

The Optical Interconnection Network for Cloud Computing Data Centers: State of the Art and Future Research

YU Xiao-Shan[1)] WANG Kun[2)] GU Hua-Xi[1)] WANG Xi[1)]

[1)](*State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an* 710071)

[2)](*Department of Computer Science, Xidian University, Xi'an* 710071)

Abstract Data center is the key platform to support cloud computing. The rapid development of cloud-based services and applications has brought great challenges to data center network. Traditional electronic interconnection network can hardly satisfy all requirements for bandwidth, cost of devices, power consumption, and management. The optical interconnection network, which is characterized by low power consumption, low cost and high bandwidth, is proposed to solve this problem. Recently it has become a hot topic in the field of research. Based on the comparison of the basic characteristics of the optical and electronic interconnection techniques, this paper presents a survey on the research progress of the optical interconnects for cloud computing data center. The proposed optical interconnect schemes include the optical/electronic hybrid networks, the centralized all-optical networks, and the distributed all-optical networks. We analyze and compare these optical interconnects in terms of the switching strategy, scalability, technical feasibility, equipment cost, and so on. Finally, this paper proposes some future research topics in the optical data center network for cloud computing.

Keywords cloud computing; data center network; optical interconnects

收稿日期:2014-12-11;最终修改稿收到日期:2015-05-17.本课题得到国家自然科学基金(61472300)、中央高校基本业务经费(JB150318,JB142001-5)和高等学校学科创新引智计划(B08038)资助.**余晓杉**,男,1986 年生,博士研究生,主要研究方向为云计算数据中心光互连网络技术.E-mail:yuxiaoshan.21@163.com.**王 琨**,女,1982 年生,硕士,讲师,主要研究方向为云计算、虚拟网络技术.**顾华玺**,男,1977 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为云计算网络互连技术、高性能光互连技术.**王 曦**,男,1991 年生,硕士研究生,主要研究方向为光互连网络、软件定义网络.

1 引　言

通过广泛的接入模式、共享的资源架构、按需的服务部署及灵活的容量扩展，云计算在近年来获得了广泛的部署和应用[1]. 数据中心是云计算的核心支撑平台，随着云应用的广泛部署，数据中心的通信模式和业务需求出现了根本性变化. 这些变化具体包括：(1) 数据中心的网络规模和负载出现了指数级增长①；(2) 主要的流量模式由传统“南北向流量”转变为“东西向流量”；(3) 更多时延敏感和数据密集型业务在数据中心内运行；(4) 一些虚拟化技术，如虚拟机实时迁移，需要网络提供更好的支持. 这些变化对数据中心网络架构提出了更高的要求. 传统数据中心网络在对分带宽、传输时延、网络可扩展性、容错性、资源利用率等方面均无法满足云业务的需求[2-3]. 对此，研究人员提出了新的电互连网络架构，如 Fat Tree[4]、VL2[5]、DCell[6]、BCube[7]、CamCube[8] 和 Snowflake[9] 等. 尽管上述架构能够有效满足新的云业务要求并改善数据中心的网络性能，但这些网络架构同时也带来了拓扑结构复杂、线缆开销过大、设备数量过多、网络能耗难以优化等问题. 究其根本原因在于，随着网络容量的指数级增长，基于 COMS 的电子元件几乎达到了其带宽的上限[10]. 因此，光互连技术得到研究人员的极大关注. 与电互连技术相比，光互连技术能够更好地满足云计算数据中心对能耗和带宽的需求. 尤其随着绿色计算[11]、Green Cloud[12] 等概念的提出，数据中心光互连技术成为网络节能的重要方式[13]. 近年来，结合云计算数据中心的流量模式和新型光交换器件，研究人员提出了多种新的光互连网络架构. 实验和仿真表明，这些架构在吞吐、时延、灵活性、能耗等方面优于传统的电互连网络架构. 但相对于电互连网络，工业界和学术领域对于数据中心光互连网络的研究尚处于起步阶段，其中很多技术挑战尚未得到很好的解决. 随着云计算的发展，服务、计算、存储、网络将进一步融合为一个整体方案，相对于发展迅速的计算技术和存储技术，网络技术的革新相对缓慢. 因此，深入研究数据中心网络，尤其是具有革新性的光互连网络，对于未来网络技术和云计算技术的创新发展都具有重要的意义.

本文将详细介绍数据中心光互连网络的最新研究进展，对比和分析现有网络架构的特点，总结并讨论数据中心光互连网络的未来研究趋势.

2 电互连网络的技术挑战

云计算的发展使数据中心内的流量出现了爆炸式增长，因此在接入层采用 10 GigE 交换架构，在核心层采用 40 G/100 GigE 交换架构已基本成为数据中心网络未来的发展趋势. 在这种情况下，数据中心电互连架构将面临以下技术挑战：

(1) 高带宽要求

对于铜缆而言，在额定功率下其通信带宽和传输距离之间存在着一定的权衡，即随着通信带宽的提升，铜缆的传输距离会下降. 对于 10 Gbps 的通信带宽，铜缆的传输距离会小于 10 m. 若要获得更长的传输距离，则发射器的功率需要大于 6 W/端口[14]. 因此，随着链路带宽需求的提升，铜缆不再是一种理想的传输介质.

(2) 高交换容量要求

由于在单一线路上，电信号的传输速率受到信号损耗及码间串扰的限制，因此芯片设计中常通过增加线路位宽的方式提升信号的传输速率. 但这种方式最终会受到芯片封装面积的限制. 因此，随着信号速率的提升，交换芯片所能支持的端口数目会逐步降低. 例如，文献[15]提出，在现有工艺水平及封装限制下，若端口速率为 80 Gbps，则基于电集成技术的交换芯片最大支持的端口数目仅为 64.

(3) 低开销要求

随着链路带宽的提升，电信号在铜缆或背板线路的传输损耗增大，因此信号需要进行复杂的预加重处理和差错控制. 另外，由于端口密度和交换容量的限制，电交换网络需要使用更多的设备和更加复杂的互连方式以满足大容量交换的需求，这增加了网络的设备开销和布线开销.

(4) 低能耗要求

能耗问题已经成为制约数据中心发展的瓶颈因素之一. 根据预测，从 2012 年到 2020 年，高性能计算系统的峰值计算能力和带宽需要分别以 10 倍/4 年和 20 倍/4 年的速度增长，但其能耗仅允许以 2 倍/4 年的速度增长. 这对数据中心的设计提出了严峻的挑战[16]. 数据中心的网络能耗占总能耗的 23% 左右，随着链路速率和交换机容量的提升，网络能耗所占的比例将继续增大[17]. 由于目前电交换机的能耗

① Cisco Global Cloud Index: Forecast and Methodology, 2012～2017. www.intercomms.net/issue-21/pdfs/articles/cisco.pdf

不会随负载的降低而成比例缩减,因此电互连网络常用的节能策略主要包括[13]:①将网络流量整合至更少的链路和交换设备,关闭空闲的网络设备;②降低链路速率以节省发射机能耗.上述两种策略都会在一定程度上带来性能的损失.

光互连技术具备巨大的潜力解决上述问题.例如文献[14]指出:在最优化的情况下,光电混合网络Helios能够比相同规模的电网络节省2倍的成本、5倍的设备开销和8倍的网络能耗.光互连架构能够实现上述优化的主要原因包括以下几个方面:(1)在传输带宽方面,目前基于100 Gbps PM-QPSK调制的相干接收器已经商用[18].结合DWDM技术,一对单模光纤的传输带宽可达12 Tbps[19].进一步就传输距离而言,目前多模光纤和单模光纤的无中继传输距离可轻易满足数据中心内的互连需求;(2)在交换容量方面,光交换架构可以实现更高的交换容量.这是因为:①高速光信号的损耗和串扰远小于电信号;②通过波分复用技术,单一光波导内承载的信道数目可以实现数十倍的增长;(3)在网络开销方面,光纤具有更高的带宽、更小的横截面积和更轻的重量,因此能够带来更好的散热并降低网络的布线开销.同时,由于多数光交换单元对于信号的速率、调制模式、协议等具有透明传输的特性,因此网络能够在不更换光交换设备的情况下进行链路带宽的升级.另外由于光交换机能够实现更大的交换容量和更高的端口密度,因此很多全光互连网络或光电混合网络能够实现扁平化的架构设计.这在很大程度上降低了网络的设备和管理开销;(4)在能耗方面,光互连技术能够很好地解决能耗和性能的权衡问题.这主要因为:①光信号具有更低的损耗和更长的传输距离,因此光链路可以使用更低的发送功率;②对于全光交换架构,信号在中间节点不用经历O/E(光-电)和E/O(电-光)转换的过程;③光交换架构可以采用无源或低能耗光器件构建,因此可进一步缩减网络能耗.

光互连技术和电互连技术的综合对比如表1所示.可以看到,由于受到通信带宽、交换容量、设备开销和能耗的限制,电互连网络已较难满足云计算数据中心的通信需求,因此具备高带宽和低能耗的光互连技术已受到研究人员的重点关注.尤其随着硅光技术和光集成技术的发展,光交换设备的成本日益降低,这更促进了光互连技术在数据中心的应用和部署[20].

表1 电互连技术和光互连技术的对比

对比条目	电互连技术	光互连技术
带宽	链路带宽受到传输距离和发射机功率的限制.	单一信道的带宽可达100 Gbps,进一步结合DWDM技术,链路带宽可成倍增加.
交换容量	在高信号速率下,交换容量受到芯片封装面积限制.	(1)光交换芯片不需要以增加线路位宽的方式提升线路速率;(2)光信号可以通过DWDM技术成倍增加芯片的交换容量.
网络开销	(1)随着信号速率的提升,对信号的处理过程更加复杂;(2)网络需要使用更多的设备和线缆以满足应用对带宽和吞吐的需求.	(1)光纤具备更高的带宽、更小的横截面积和更轻的重量,因此可有效降低线缆方面的开销;(2)光交换单元对于信号的透明传输特性可降低网络升级方面的开销;(3)大容量的光交换架构可降低网络设备方面的开销.
能耗	目前电交换设备的能耗不会随负载降低而成比例缩减,而常用的节能策略都存在能耗与性能方面的权衡.	(1)光互连架构可有效降低信号在传输过程中的能耗(包括信号发射功率,中间节点的光电转换能耗等);(2)通过使用无源器件,光互连网络可进一步缩减能耗.

3 研究现状分析

3.1 光电混合互连网络架构

为有效兼容现有数据中心的电互连技术,同时通过光互连技术解决核心层流量负载过大的问题,研究人员提出了光电混合互连架构.该架构中,电交换设备和光交换设备构成两个并行网络,不同流量将通过不同的网络完成传输.由于混合架构中,光互连网络一般采用商用全光交换设备构建,因此具有较高的技术可行性.典型的光电混合互连网络包括c-Through[21]、Helios[14]、基于混洗-交换[22]和基于光突发环路[23]的光电混合网络.

为有效地将光电路交换网络与传统电分组交换网络结合,以创建适合同时承载突发性业务和数据密集型业务的网络环境,Wang等人[21,24]提出了如图1所示的c-Through网络架构.在该架构中,每个ToR(Top of Rack)交换机同时连接至一个电分组交换网络和一个光电路交换网络.其中,电分组交换网络由多个传统以太网交换机互连成树形拓扑,光电路交换网络则使用单一微机电系统(Micro-Electro-Mechanical System,MEMS)交换机构建.由于光电路交换机所提供的并行连接极为有限,同时在配置周期内光链路无法承载任何数据,因此光连接必须提供给网络中具有最大通信量的源-目的机架.c-Through使用一套集中式控制系统完成光连接的

优化配置. 在该控制系统中，光配置管理器连接至所有服务器的统计信息接口以获取流量信息，同时该管理器连接至光交换机的控制接口以实现配置指令的下发. 具体到控制流程，首先各服务器所部署的后台管理模块会实时监测每条 socket 队列所缓存的分组数目，并将该信息周期性汇报给光配置管理器. 对于属于同一应用的 TCP 连接，其分组会存储在指定 socket 队列中，因此通过收集所有服务器 socket 队列的缓存状态，光配置管理器可准确了解全网待发送的流量. 完成状态信息收集后，光配置管理器将进一步生成流量矩阵. 该流量矩阵以源机架号为行坐标、以目的机架号为列坐标，矩阵内的元素是从源机架到目的机架待发送的分组数目总和. 根据该流量矩阵，光配置管理器将决定光交换机的最优配置. 该优化问题实际是一个二分图的最大权重匹配问题. 二分图中 A 子集内的节点代表源机架，B 子集内的节点代表目的机架，从 A 子集到 B 子集的一条边代表一条连接，边的权重为待发送的分组数目. 对于该问题的求解可以使用经典的 Edmonds 算法. 在运行该算法获得配置结果后，光配置管理器将下发配置命令并通知服务器传输相应分组. 未获得光连接的服务器将通过电网络传输分组. c-Through 在 ToR 交换机位置采用基于 VLAN 的路由算法从逻辑上分离电网络和光网络，VLAN-s 负责处理通过电分组交换网络的分组，VLAN-c 负责处理通过光电路交换网络的分组.

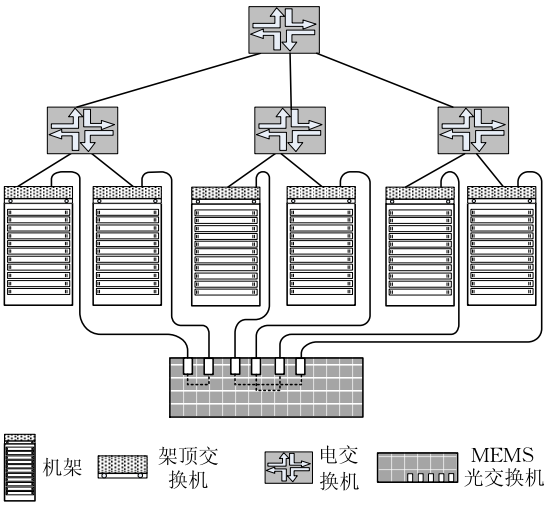


图 1　c-Through 网络架构

尽管 c-Through 能够在提供高通信带宽的同时保持网络的低复杂度，但在实际部署中却遇到了较多的挑战. 主要原因在于数据中心的实际流量远比初始设计架构时的假设流量复杂. 为解决这一问题，在文献[25]中，研究人员提出了 Observe-Analyze-Act 控制架构，该控制架构将光交换机的配置过程划分为 3 个阶段，即观察阶段(Observe)、分析阶段(Analyze)、配置阶段(Act). 在观察阶段，光配置管理器与不同设备进行通信并获取相应的信息. 这些信息具体包括：从交换机位置获取的链路利用率信息，从任务调度器位置获取的应用状态信息，从服务器位置获取的应用优先级及 QoS 要求信息. 随后，光配置管理器将上述信息进行汇总和分析. 在该阶段，控制器将准确剖析应用的流量需求，探测病态(ill-behaved)负载，发现相关性流量，生成有效支持应用要求的优化配置. 最后在配置部署阶段，光配置管理器会将配置命令下发至网络中光交换设备. 该控制架构充分考虑了数据中心实际流量的复杂性特点，提出从应用层、任务调度层、网络层多个维度解析应用的流量需求，但该文章仅提出了控制架构的总体思想，并未就具体实现细节进行讨论.

Farrington 等人[14]提出了如图 2 所示的光电混合互连架构 Helios. 该架构用于实现 pod 之间的混合互连. 这里的 pod 是一个包含 250～1000 台服务器的集群，pod 内服务器通过铜缆连接至 pod 交换机，不同 pod 之间通过核心层交换机进行通信. pod 交换机和核心层交换机相互连接构成一个 2 层多根树拓扑. 其中，核心层交换机可以是电分组交换机或是基于 MEMS 的光电路交换机. 电分组交换机用于处理 all-to-all 的突发性流量，光电路交换机用于处理 pod 间具有高带宽需求和长持续周期的流量.

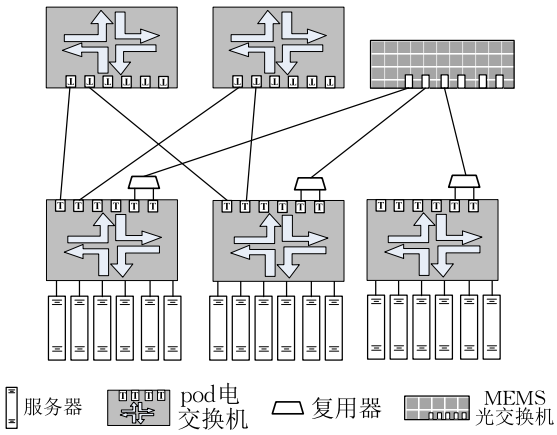


图 2　Helios 网络架构

Helios 中用于配置光连接的控制系统包括 3 个软件模块：拓扑管理(Topology Manager，TM)模块、电路管理(Circuit Switch Manager，CSM)模块、pod 交换机管理(Pod Switch Manager，PSM)模块. 其中，PSM 模块部署在 pod 交换机上，每个模块使

用流计数器记录从本 pod 发往不同 pod 的流量(按字节计数).TM 模块周期性使用远程过程调用协议从所有 PSM 模块获取流量的统计信息.随后将这些信息填入以源 pod 号为行坐标、以目的 pod 号为列坐标的字节计数矩阵中.通过当前周期的字节计数矩阵和前一个周期的字节计数矩阵,可计算出流速率矩阵.随后,TM 移除矩阵中流速率小于 15 Mbps 的元素,即获得只记录大流的流速率矩阵.该流速率矩阵不能直接应用于光电路的配置决策,TM 需要进一步使用 Max-Min 公平带宽分配算法处理该矩阵以获得能够更好反映应用需求的流量需求矩阵.最后,TM 根据流量需求矩阵计算光电路的最优配置,该建模过程和使用的求解算法与 c-Through 架构相同.在获得配置结果后,TM 将通知 CSM 配置光电路交换机的输入输出端口,同时通知 PSM 将未获得光连接的流量转发至核心层的电分组交换机.

Helios 架构通过使用商用全光交换机和 WDM 收发器实现网络成本和能耗的缩减.但由于传统 MEMS 全光交换机主要应用于电信骨干网[26],在数据中心环境中,MEMS 交换机主要存在以下缺陷:(1)端口规模有限,较难支持大规模的网络互连;(2)交换机配置时延过长(10 ms~100 ms),光链路的带宽利用率相对较低;(3)交换机的插入损耗较大,可能会限制交换机的级联.文献[27]对 MEMS 光交换机的配置时延进行了详细的分析.研究发现,通过使用先进的交换器件可以进一步缩减交换机的配置时延,这种优化将带来网络吞吐的显著提升.

由于 MEMS 交换机较长的配置时延会带来光电路利用率的下降,同时一旦交换机响应某个连接请求,则在后续一段时间内,交换机的特定输入输出端口被该连接独占.这种有限的连通性在部署至大规模网络时会成为挑战.为解决上述问题,Lugones 等人[22]提出了基于混洗-交换拓扑的光电混合网络.在该网络架构中,具有点到点通信、持续时间长和高负载的数据流通过光电路交换网络,低速率的数据流和控制消息通过电分组交换网络.为保证光网络的可扩展性,光互连架构部分采用了如图 3 所示的单级混洗-交换(Single-stage Shuffle eXchange,SSX)拓扑.该拓扑由边缘层和核心层构成,其中边缘层包括 N 个 k 端口电分组交换机,核心层包括一个或多个 MEMS 光交换机.电分组交换机的每个输入输出端口通过单向光纤连接至光电路交换机,而光电路交换机内部的交叉矩阵按照特定的混洗规则进行配置.与传统混洗-交换网络不同的是:在该架构中,分组首先在电交换机内执行交换操作,即通过路由计算选择合适的输出端口;随后分组在光交换机内执行混洗操作.由于光交叉开关总是按照某种混洗模式配置输入输出端口,某些分组无法通过一次交换-混洗操作到达目的节点.因此文章提出了多跳路由的方式,若分组在经过一次光交换机后并未到达目的节点,则该分组需要在中间节点(电分组交换机)再次路由并重新注入光交换网络.该过程会进行多次直到分组到达目的节点.这种多跳路由的方式一方面增加了光电路交换机的逻辑连通性,另一方面减少了交换机的配置频率,能够在一定程度上提高光链路利用效率.另外,该架构中光交叉开关矩阵的混洗连接模式并非固定不变的.文章提出了一种启发式算法,该算法将根据网络的流量状态动态改变光交叉矩阵的混洗模式,以保证当前网络中具有最大通信负载的源、目的节点对通过一跳光连接到达,其余节点对则通过多跳的光连接到达.该启发式算法以流量矩阵 λ 和光交换机当前的混洗连接状态为输入,具体操作包括以下几步:第 1 步,找出该流量矩阵中 N 对具有最大通信量的源、目的节点对,并将对应元素 $\lambda[i][j]$ 写入子集 η 中.其中 $\lambda[i][j]$ 表示源节点 i 到目的节点 j 的流量;第 2 步,针对子集 η 中的一个元素 $\lambda[i][j]$,检查源节点 i 到目的节点 j 的跳数 $d(i,j)$ 是否大于 1.若是,则寻找替代节点 s,s 需要满足条件 $d(s,j)=1$,且 $\lambda[i][j]>\lambda[s][j]$;第 3 步,若在第 2 步中找到替代节点,则进行混洗模式的重新配置,将混洗拓扑中 s 到 j 的连接修改为 i 到 j 的连接;第 4 步,将第 2 步和第 3 步重复执行 N 次.若混洗拓扑已经改变了 m 对连接的配置($m\leq N$),则原有的 m 对连接被拆除.此时该混洗拓扑中存在 m 个空闲的输入端口和 m 个空闲的输出端口,将这些空闲输入输出端口按照混洗模式重新连接.

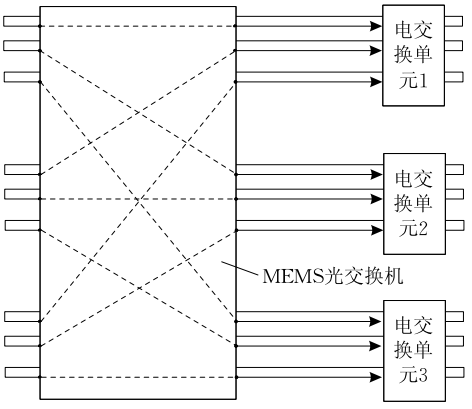


图 3 混洗-交换网络架构

尽管混洗-交换网络能够在一定程度上缓解光电路交换所带来的性能问题，但该网络仍然对通过光连接的流量有两点要求：(1) 流量的持续时间远大于 MEMS 交换机的配置时延；(2) 流量的带宽需求不能小于一个波长信道的带宽. 这些限制条件意味着在实际中只有部分业务能够通过光电路交换获得收益. 采用新的交换机制是解决上述问题的根本方法. 对此 Li 等人[23,28]提出了结合光突发交换和电分组交换的混合网络架构. 如图 4 所示，该架构的电互连网络部分采用两层的多根树拓扑. 底层由 ToR 交换机或 End of Row (EoR)交换机组成，每台底层交换机连接数十台服务器构成 pod. 上层由高性能核心交换机组成，每台核心交换机连接所有底层交换机. 底层交换机除连接上层的核心交换机外，还通过光纤链路连接成环形拓扑，该环形拓扑即构成混合架构的光互连部分. 在该混合架构中，突发性流量将通过核心交换机转发，而相对稳定的流量则在光突发环路上传输. 每台底层交换机配置有线卡(Line Card，LC)、交换卡(Switch Card，SC)、光突发线卡(Optical Burst Line Card，OBLC)和光突发交换卡(Optical Burst Switch Card，OBSC). 其中，LC 用于连接服务器和核心交换机，SC 用于转发 pod 内分组并进行流量汇聚，OBSC 用于实现光突发块上下环路，OBLC 用于生成或接收光突发块. 架构采用面向连接的方式传输光突发块，分组在传输之前需要建立连接并预约该连接的持续时间. 在光突发环路上，该持续时间可以用连接所支持传输的光突发块数目衡量. 由于业务的流量会持续地产生，因此 OBLC 所请求的传输容量(以光突发块的数目计算)会大于当前所存储的分组数目. 在连接建立后，源节点将记录已发送光突发块的数目. 当 OBLC 接收到新产生的流量后，首先检查本 pod 是否已经建立到达目的 pod 的连接. 若连接存在且剩余的传输容量支持新到达的流量，则 OBLC 将流量封装成光突发块并通过 OBSC 发送. 否则，OBLC 请求建立一条新的连接. 光突发环路使用多个波长信道增加链路带宽，控制信息在特定的波长信道上传输. 环路上某个 pod 中的一台服务器将作为超级节点从控制信道收集所有节点的连接请求，随后为每对连接分配波长信道和光突发块的发送时隙. 配置信息会封装在信道每个突发帧的首个光突发块中，所有节点需要在指定时隙接收在该信道内广播的配置信息，随后按照该配置信息进行光突发块的发送或接收. 由于该架构并未使用具有较长配置时延的 MEMS、波长选择交换机(Wavelength Selective Switch，WSS)等光器件，因此具有更好的灵活性和更高的带宽利用率. 但面向连接的光突发交换仍然存在链路建立的时延开销和控制开销. 同时由于架构采用了集中式仲裁的方式，这些开销会随着节点规模的上升而增大.

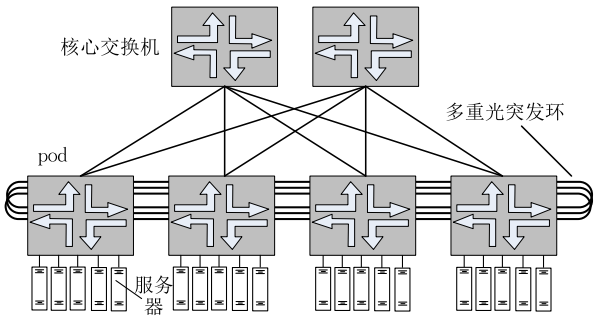


图 4　基于多重光突发环的混合架构

3.2　全光互连网络架构

光电混合互连网络的优势在于：(1) 为传统电互连网络提供了渐进式升级的方案；(2) 该类光网络一般使用成熟的光交换技术和设备构建，因此能够较快部署到实际网络中. 但由于光电混合网络中仍存在大量的电交换设备，因此网络并未在能耗和设备开销方面取得显著的优化. 目前大规模集群系统的能耗水平已经超过预期，为进一步降低能耗，需要在服务器级互连、板级互连、甚至在芯片级互连领域广泛使用光技术[29]. 因此，研究人员提出了多种面向下一代云计算数据中心的全光网络架构. 值得注意的是，由于目前全光互连网络的扩展性极为有限，这些架构最多提供数千个交换端口，无法实现服务器级的全光互连. 在这些网络中，仍然需要使用少量电交换机完成机架内互连或簇内互连. 因此，目前我们主要根据网络的簇间(或机架间)通信方式来区分光电混合网络和全光网络. 若位于不同簇(或机架)的服务器既可通过电网络通信，又可通过光网络通信，则该网络属于光电混合互连网络. 若位于不同簇(或机架)的服务器只能通过光网络进行通信，则该网络属于全光互连网络.

从互连方式上看，可以将全光互连网络划分为集中式光互连网络和分布式光互连网络. 根据所使用的核心光器件，集中式光互连网络可进一步分为基于 MEMS、基于阵列波导光栅路由器(Array Waveguide Grating Router，AWGR)、基于半导体光放大器(Semiconductor Optical Amplifier，SOA)和基于微环谐振器的光互连网络. 下面将分别介绍各类网络架构的典型设计方案.

3.2.1　集中式全光互连网络架构

考虑到数据中心的实际流量往往呈现局部性特

点,采用全对分带宽的网络拓扑不仅会带来成本、能耗、布线复杂度等问题,而且可能造成带宽资源的浪费.设计一种灵活可重配置的网络架构能够在性能和成本上实现更好的权衡.基于此,Singla 等人[30-31]提出了如图 5 所示的 Proteus(或 OSA)架构.该架构使用 MEMS 光交换机直接连接所有 ToR 交换机构成星形拓扑.每个 ToR 交换机配置有数个工作在不同波长的光收发器,从这些收发器发送的多波长信号经过复用后输入 $1\times k$ 波长选择交换机(WSS).波长选择交换机根据配置将波长重分为 k 组,并通过 k 个端口将信号送入 MEMS 光交换机.通过配置 MEMS 交换机的交叉开关矩阵,每个 ToR 交换机可以与其他 k 个 ToR 交换机直接连接.对于非直接互连 ToR 交换机,Proteus 提出了 hop-by-hop 的通信方式.源 ToR 交换机选择当前 k 个连接中的一个作为中间节点,该中间 ToR 交换机接收到信号后进行光电转换,读取分组头部并重新向目的 ToR 交换机转发.为避免转发流量过大造成中间节点缓存溢出的问题,高通信流量的节点对之间需要采用直接互连的方式.除通过动态配置光交换机改变网络拓扑的连通性外,Proteus 还通过波分复用技术和 WSS 实现链路带宽的灵活配置.在该架构中拓扑管理器(Topology Manager,TM)负责完成 MEMS、WSS 和 ToR 交换机的优化配置工作.为最大化网络吞吐,TM 需要根据流量矩阵完成以下配置:(1)MEMS 光交换机的配置,保证具有最高通信量的 ToR 交换机通过一跳完成通信;(2)WSS 的配置,确保每个端口的带宽满足 ToR 交换机的发送需求;(3)ToR 节点对之间的路由配置,保证流量低时延、无阻塞到达目的节点.文章将这个优化配置问题转换为一个混合整数线性规划,并提出使用启发式算法求解该问题.Proteus 的优势在于网络具备较高的带宽灵活度,能够根据流量需求提供较高的对分带宽.但由于 Proteus 使用 MEMS 交换机,因此对于业务的流量特点仍然有较高的要求[32].

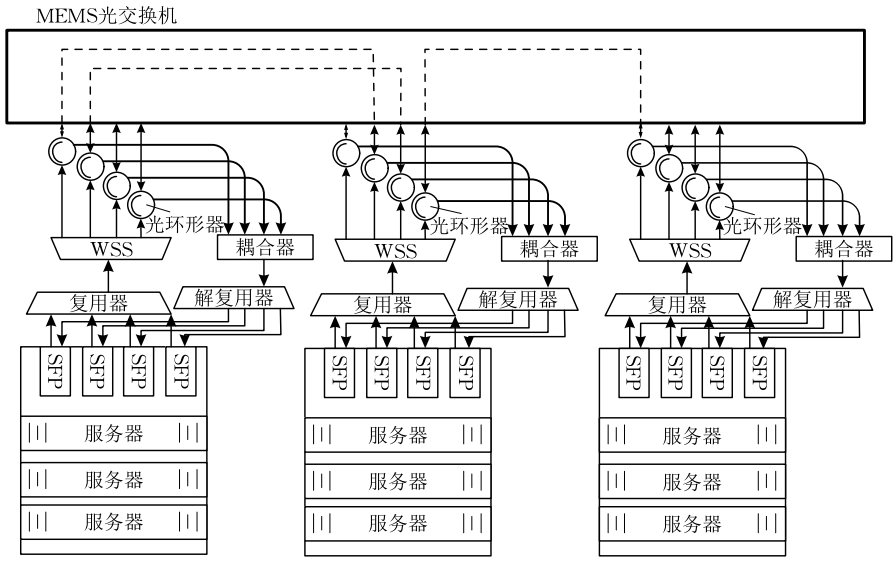


图 5 Proteus 全光交换架构

尽管 MEMS 全光交换机是比较成熟的商用设备,该交换机的使用能够带来网络成本的降低.但较长的配置时延会严重影响交换粒度和网络的性能,因此研究人员基于其他的快速光交换单元提出了新的交换架构.

Ye 等人提出的 DOS(或 LION[33])架构基于 AWGR 构建[34].AWGR 是一种基于波长交换的元件,通过干涉相消和干涉加强效应实现光信号从任意输入端口到任意输出端口的路由.对于某端口的输入信号,其输出端口取决于信号的波长.通过在输入端口部署可调波长转换器(Tunable Wavelength Converter,TWC)即可实现对信号路由的控制.如图 6 所示,DOS 使用核心光交换架构直接互连所有服务器节点.该核心光交换架构具体包括 TWC、AWGR 和共享式环回缓存.源节点发送的分组通过光通道适配器进行电光转换,随后部署在输入端口的光标签提取器将提取分组头域信息,并将该信息送入控制层进行路由计算和仲裁.控制层将根据仲裁结果配置输入端口的 TWC,光分组经过 TWC 时被转换到特定波长并路由到对应的输出端口.根据 AWGR 的波长路由特性,同一输入端口的信号通过不同的波长到达不同输出端口.若输出端口接收到不同输入端口的信号,则这些信号的波长必然不同.因此,对于 N 端口 AWGR,若在输出端口配置 1:N

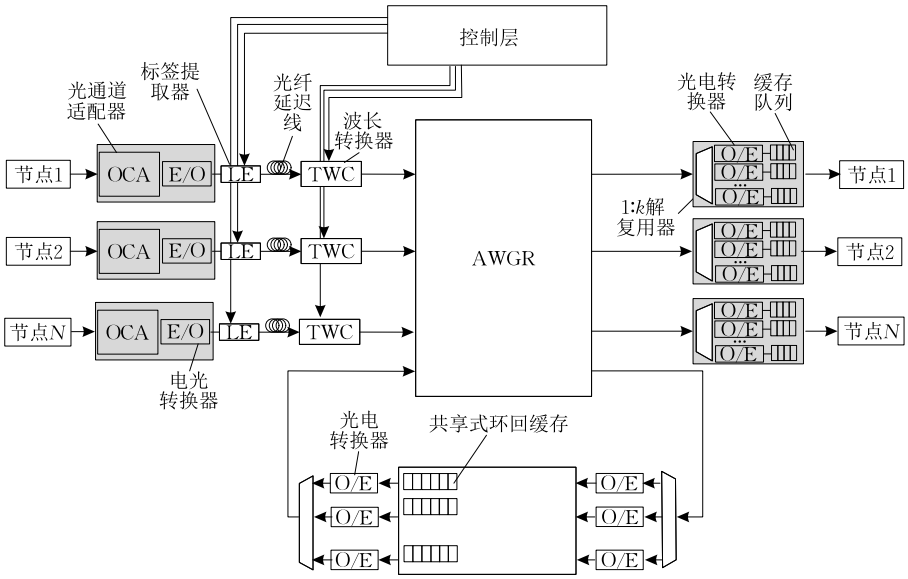


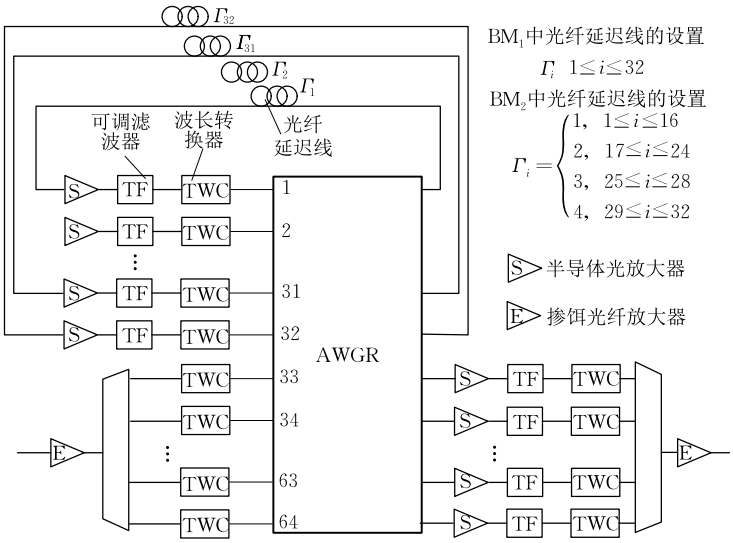
图 6　DOS 全光互连架构

解复用器，则 AWGR 可以实现 all-to-all 无阻塞通信. 但这会带来较高的器件成本，因此在实际应用中 AWGR 的输出端口会配置 $1:k(k<N)$ 解复用器. 这将导致使用某些波长的分组无法同时到达输出端口. 为解决波长竞争问题，DOS 使用了共享式环回缓存(Shared Loopback Buffer，SLB)结构. 竞争分组将被路由至环回缓存的输入端口，通过光电转换临时存储在环回缓存中. 缓存控制器负责为竞争失败的分组发送重新传输的请求，收到响应后，相关分组将再次进行电光转换并通过 AWGR 传输到目的节点. 尽管上述共享式环回缓存结构具有控制简单、资源利用率较高的优点，但这种缓存结构必须在内部设置 N 倍的加速带宽以处理多个分组同时到达一条队列的情况. 缓存的 I/O 带宽会随着交换机端口数目的增加而增大，这会限制交换机的扩展规模. 为解决这一问题，文献[34]提出了分布式环回缓存(Distributed Loopback Buffer，DLB)结构和混合式环回缓存(Mixed Loopback Buffer，MLB)结构. 其中，DLB 是一种基于输入队列的分布式缓存结构，它使用 N 个分离的缓存单元实现 N 条分离的缓存队列，每个队列对应存储一个特定输入端口的延迟分组. 在最简单的情况下，每条队列无须使用虚拟输出队列，仅配置一个发射器. 但这会导致缓存控制器设计难度的增加，同时队头阻塞的问题不可避免. 作者利用波长的并行性来缓解队头阻塞问题，通过在每条队列部署多个发射器，可以实现使用不同波长的多个分组并行向不同输出端口发送. 另外前文所述的 SLB 仅占用 AWGR 的一个端口，而 DLB 中每条缓存队列都需要占用 AWGR 的一个端口. 因此对于连接 N 个终端节点的分布式环回缓存架构，需要使用 $2N\times 2N$ 的 AWGR 分别连接终端节点和缓存队列. 与 SLB 相比，DLB 能够实现更低的端到端传输时延，同时缓存的 I/O 带宽仅与输入端口的数据速率相关. 但 DLB 需要占用更多的端口和发射机单元. 为结合两种缓存结构的优势，作者进一步提出了 MLB 结构. MLB 仍需要占用多个 AWGR 端口以支持多条分离的队列，但与 DLB 不同的是，在 DLB 中每条回收缓存队列仅服务一个输入端口的延迟分组，在 MLB 中每条回收缓存队列服务 r 个输入端口的延迟分组. 因此，若互连 N 个服务器节点，混合式回收缓存队列需要占用 N/r 个端口. MLB 中每条队列同样使用多个发射器以缓解队头阻塞的问题. 综合对比上述 3 种缓存结构，SLB 需要最高的缓存 I/O 带宽，但仅占用 AWGR 一个端口；DLB 需要最小的缓存 I/O 带宽，但需要更多的可调波长转换器. 同时 AWGR 需要使用 $2N$ 个端口以分别连接 N 个终端节点和 N 条回收队列；MLB 实现了 SLB 和 DLB 之间的权衡. 文献[35]建立了 DOS 的硬件实验平台，实验发现环回缓存的复杂结构会成为交换机扩展的主要限制因素. 为解决这一问题，作者进一步提出了基于全光 NACK(All Optical Negative Acknowledgement，AO-NACK)的交换架构[35-36]. 该交换架构与基于共享环回缓存的 AWGR 架构基本相似，对于一个 $(N+1)\times(N+1)$ 的 AWGR，其中 N 个端口通过 TWC 连接至服务器节点，剩余一个端口用于处理竞争延迟的分组. 与共享环回缓存架构的不同之处在于：(1) 每个输入端口配置有两个光环形器用于分离发往 AWGR 的分组

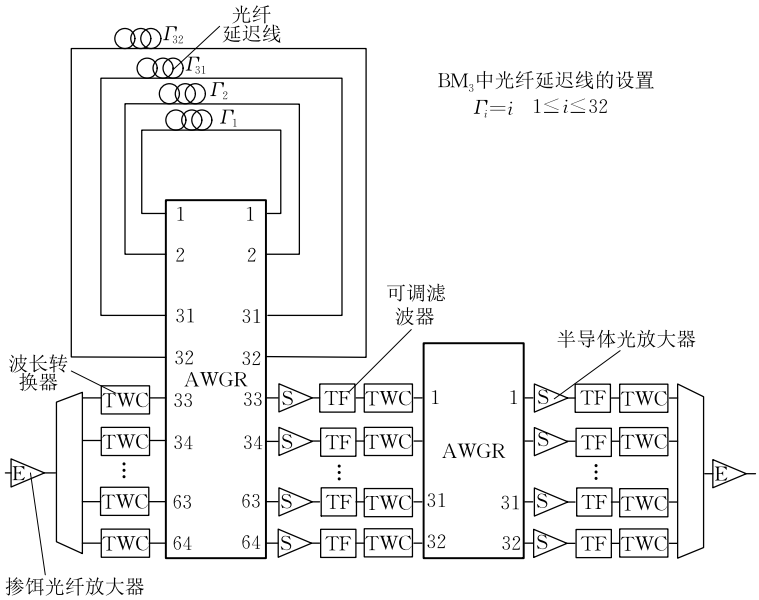
和反向传输的 AO-NACK 分组；(2) 处理竞争分组的端口被称为反射端口，该端口通过使用一个环形器实现所有输入信号的反向传输. AO-NACK 架构中不存在电缓存. 当发往同一输出端口的两个分组发生波长竞争时，控制层将调节 TWC 使其中一个分组正确发往目的端口，使另一个分组发往反射端口. 反射端口所连接的光环形器会将分组反向发送至输入端口，连接在输入端口的光环形器进一步将分组反向发送至源服务器. 当源服务器探测到该分组时，知道分组发送失败，因此会重新发送该分组. AO-NACK 架构能够并行处理所有输入端口的竞争分组，因此能获得与 DLB 相似的网络性能，同时避免了环回式缓存带来的成本和控制开销.

由于基于 AWGR 和 TWC 的波长交换架构能够实现细粒度下的光分组交换，因此得到研究人员的广泛关注. 在解决分组竞争的问题上，除上述 DOS 中提出的电环回缓存机制、AO-NACK 机制，还可以使用基于光纤延迟线(Fiber Delay Lines，FDL)的光环回缓存机制.

Rastegarfar 等人在文献[37]中提出了 3 种基于 FDL 的光环回缓存结构. 如图 7 所示，这 3 种光环回缓存结构分别被命名为 BM1、BM2、BM3. BM1 使用一个 $2N\times 2N$ 的 AWGR 作为核心交换单元，其中 N 个端口作为交换端口实现输入光分组的直接转发，其余 N 个端口作为环回端口实现竞争分组的缓存和二次转发. 每个交换端口的输入位置连接有 TWC 以实现分组的波长路由. 在输出位置依次连接有 SOA、可调滤波器(Tunable Filter，TF)和



(a) BM1及BM2架构



(b) BM3架构

图 7 基于光纤延迟线的 AWGR 环回架构

TWC. 其中 SOA 用于补偿光信号的传输损耗，TF 用于限制自发放大辐射噪声，TWC 用于调整输出分组的波长. 每个环回端口的输出位置连接有长度为一个时隙的 FDL，在输入位置依次连接有 SOA、TF 和 TWC. 在每个时隙，BM1 通过合理配置 TWC 可最多实现 N 个高优先级的分组到达交换端口的输出端，其余分组则需要丢弃或进入环回缓存. BM1 的最大特点在于环回缓存的共享性，即任意输入分组可以进入任意的 FDL.

在 BM1 中，所有 FDL 仅提供一个时隙的延迟，同时也仅能存储一个分组，若延长 FDL 的长度，则可以存储更多的分组. 因此 BM2 采用了与 BM1 类似的结构，唯一不同在于 BM2 中 $N/2$ 个环回端口连接长度为 1 个时隙的 FDL，$N/4$ 个环回端口连接长度为 2 个时隙的 FDL，$N/8$ 个环回端口连接长度为 3 个时隙的 FDL，剩余 $N/8$ 个环回端口连接长度为 4 个时隙的 FDL. 在每个时隙，控制层将检查目的地址为同一输出端口的分组的优先级. 若某个分组的优先级为 k，则交换架构中存在 $k-1$ 个分组优先级高于该分组，因此该分组将被延迟 k 个时隙后发送. 在 BM1 中该分组需要循环存储 k 次，而在 BM2 中，该分组可以存储在长时隙 FDL 中以降低分组循环存储的次数. 如图 7(b)所示的 BM3 使用两级 AWGR 构建. 第 1 级为 $2N\times 2N$ 的 AWGR，其中 N 个端口为交换端口，N 个端口为环回端口，第 2 级为 $N\times N$ 的 AWGR，所有端口全部为交换端口. 第 1 级 AWGR 的每个交换端口在输入位置部署有 TWC，在输出位置依次部署有 SOA、TF、TWC 模块. 第 2 级 AWGR 的输出端口同样依次部署 SOA、TF、TWC 模块. 与 BM1、BM2 不同的是，BM3 中环回端口位置仅部署不同长度的 FDL，不部署任何波长转换器件，因此在 BM3 中竞争分组只能在环回缓存中存储一次. 同时从环回端口输出的分组无法实现基于波长的路由，只能通过第 2 级 AWGR 路由到指定的目的端口. 尽管基于 FDL 的光环回缓存能够有效避免光电转换器和电缓存的使用，从而降低能耗并实现信息的透明传输. 但在实际部署中由于受到技术和成本的限制，交换机能够连接的 FDL 数目有限，这使光路由器不得不降低吞吐以避免分组的大量丢失. 对此，Rastegarfar 等人[38]进一步提出了基于 FDL 和 WDM 的共享环回缓存结构. 该结构同样采用 $2N\times 2N$ 的 AWGR 作为核心交换单元，其中 N 个端口为交换端口，另外 N 个端口为环回端口. 每个环回端口的输出位置连接长度为一个时隙的 FDL，随后依次连接掺铒光纤放大器(Erbium-doped Fiber Amplifier，EDFA)和缓存模块(Buffer Module，BM)，最后连接到该端口的输入位置构成一条环路. 每个交换端口的输入位置配置有输入模块(Input Module，IM)和 EDFA. 每个 IM 和 BM 依次由 $1\times 2N$ 解复用器、TWC、$2N\times 1$ 耦合器构成. 在一个周期内，该架构可以在输入端口实现 $2N$ 个分组的动态路由. 由于在环回端口的输入位置使用了 BM 模块，在每个时隙，每条 FDL 内可最多存储 $2N$ 个不同波长的光分组. 另外每个时隙，竞争分组可以选择任意可用的环回缓存，这种方式有效提高了缓存的利用效率. 为进一步提高交换机的吞吐，作者提出了负载均衡的调度策略. 该策略可将具有相同目的端口的分组均匀缓存在不同 FDL 中，这样在下个时隙，这些竞争分组就可以并行地输出. 仿真结果表明基于 WDM 的共享 FDL 光路由器能够有效降低分组丢失率并能较好地适应多种流量模式.

另外通过细化交换粒度，可以有效提高 AWGR 的并行交换能力，进而实现 all-to-all 流量模式下的无阻塞交换. 基于这种思想，Ji 等人[39]提出了基于正交频分复用(Orthogonal Frequency Division Multiplexing，OFDM)的 AWGR 交换架构. 该架构的交换核心使用 $N\times N$ 的 AWGR，每个 AWGR 的端口连接一台机架，机架内部署数十台服务器. 机架内的通信通过 ToR 交换机完成，机架间的通信通过 AWGR 完成. 服务器发往其他机架的分组首先在 ToR 交换机位置汇聚并发送至一个发射器，该发射器所包含的 OFDM 模块会使用合适的子载波将汇聚信号调制成 k 路 OFDM 数据流，其中 k 是汇聚流量所发往的目的机架的数目. 随后，这 k 路 OFDM 数据流被转换为 k 路多波长信号，通过 WDM 耦合器，所有信号耦合为一路 WDM 信号并被发送至 AWGR. 根据 AWGR 的循环波长路由特性，不同波长的信号被发送到不同的目的机架. 集中式 OFDM 子载波分配机制可以确保同一波长信道内的子载波不会相互冲突. 在接收端，OFDM 信号解调为原始电分组并通过 ToR 交换机将分组发送到相应的目的服务器. 由于提供了比波长更加精细的交换粒度，AWGR 位置不存在分组的竞争. 同一波长信道内的分组对于子载波的竞争可以在源节点位置进行仲裁，因此整个 AWGR 架构未使用环回缓存，这在一定程度上降低了设备和控制的开销.

尽管基于 TWC 和 AWGR 的光交换架构可以极大地简化拓扑结构设计，实现优于严格无阻塞交换的通信能力. 但在实际部署中，AWGR 的端口数目却会受到下述条件的限制：(1) 随着端口数目的增加，

AWGR 波长信道的通带中心频率偏移会更加严重，这会给 TWC 的设计和控制带来极大的挑战；(2) 光信号之间的串扰会随着 AWGR 规模的增大而增加. 为解决上述问题，研究人员提出了基于低基数 AWGR 模块的级联结构.

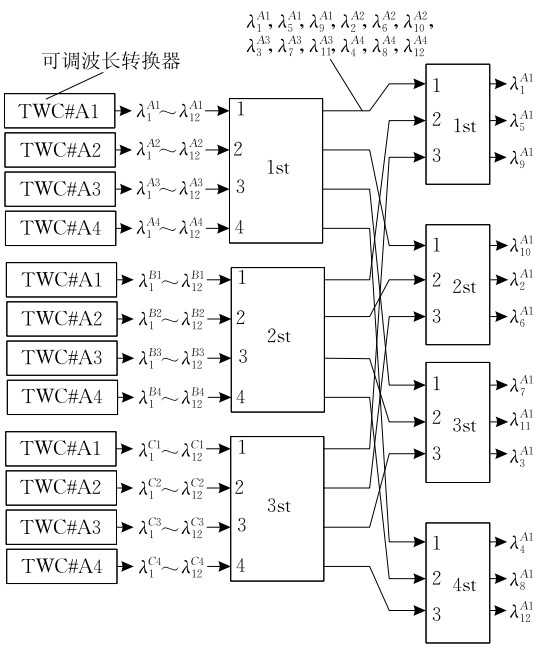


图 8 两级 AWGR 架构

Sato 等人[40]提出的两级 AWGR 交换网络如图 8 所示. 在该网络中，N 个 $M\times M$ 的 AWGR 模块构成前置交换级，M 个 $N\times N$ 的 AWGR 模块构成后置交换级. 前置交换级中每个 AWGR 模块(称为前置 AWGR)的 M 个端口分别连接至后置交换级的每个 AWGR 模块(称为后置 AWGR). 整个架构共提供 $MN\times MN$ 个交换端口. 每个前置 AWGR 模块的输入端口配置有 TWC 单元. 根据路由计算，TWC 可以将分组的波长调制到 $\lambda_1\sim\lambda_{MN}$ 范围内的特定波长输出. 其中波长的序号将决定分组的输出端口. 值得注意的是，前置 AWGR 仅有 M 个端口，根据循环波长路由的规则，MN 个波长中有 N 个波长将被路由到每个前置 AWGR 的同一个输出端口，这些波长的序号依次间隔 M. 经过后置 AWGR 模块后，上述间隔为 M 的波长组将分别被路由至不同的输出端口. 因此从整体上看，前置 AWGR 模块每个输入端口的不同波长 $\lambda_1\sim\lambda_{MN}$ 将被路由至后置 AWGR 的不同输出端口. 另外，为保证前置 AWGR 同一输出端口的波长组 $\{\lambda_i,\lambda_{i+M},\lambda_{i+2M},\cdots\}$ 在经过后置 AWGR 后被彻底分离，M 和 N 的取值为互质数. 上述两级 AWGR 交换网络完全实现了单一 AWGR 的波长路由功能，但与单一 AWGR 相比具有以下两点优势：(1) 由于架构使用了小端口 AWGR 模块，因此减小了波长偏移效应；(2) 这种两级式交换架构可以采用渐进的扩展方式构建，因此降低了网络成本开销.

Chao 等人[41-42]则提出了基于 AWGR 的三级交换网络 Petabit. 如图 9 所示，该网络由 IM 模块、CM(Central Module)模块和 OM(Output Module)模块互连成 Clos 拓扑. 每个模块使用 AWGR 作为核心交换单元. CM 模块和 OM 模块的输入端口位置配置有 TWC 以进行路由的控制，由于线卡的发射器已经包含可调激光器，因此连接线卡的 IM 模块不需要在输入端口配置 TWC 单元. 相对基于单一 AWGR 的交换架构，Petabit 需采用更加复杂的配置过程来建立输入端口到输出端口的光路径. 具体包括：(1) IM 模块需要根据分组选择的 CM 模块确定输出端口，并进一步确定线卡输出端口的 TWC 配置；(2) CM 模块需要根据分组目的端口所在的 OM 模块确定输出端口，并进一步确定本模块输入端口的 TWC 配置；(3) OM 模块需根据分组的目的端口确定本模块输入端口的 TWC 配置.

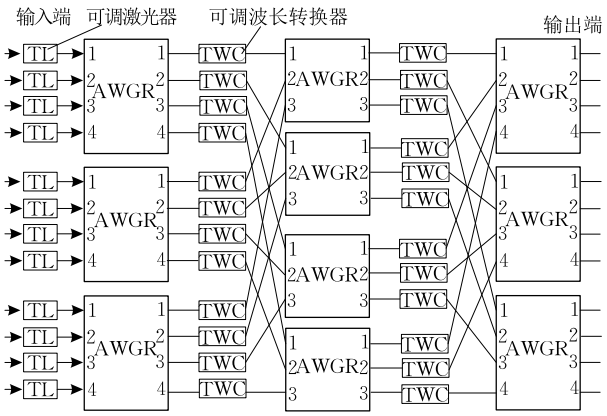


图 9 Petabit 全光交换架构

Petabit 在响应多个分组请求配置光路径的过程中，在 CM 模块和输出端口位置都会产生竞争，对此 Petabit 设计了一套电域的仲裁机制解决该竞争问题. 每个 IM 模块、CM 模块和 OM 模块分别对应于一个电域的调度模块，这些调度模块可相应地被称为 SIM(Scheduler at IM)模块、SCM(Scheduler at CM)模块和 SOM(Scheduler at OM)模块. 这些调度模块同样互连成 Clos 拓扑结构. 分组到达线卡后首先被存储在虚拟输出队列中，随后每个 IM 模块的输入端口选择 H 个非空的虚拟输出队列，并将这些队列的传输请求发送给对应的 SIM 模块. SIM 模块使用 H 个周期处理这 H 个请求，在第 i 个周期，SIM 模块将第 i 个请求发送到随机选择的 SCM 模块. 若 SCM 收到多个到达相同 OM 模块的请求，

则采用轮询的方式选择一个请求并发送到相应的SOM模块.当SOM模块接收到请求后,若请求的目的输出端口在前$(i-1)$个周期内没有做出响应,则SOM采用轮询的方式响应首个请求,否则SOM模块不做出响应.最后,SIM模块在收到某个输入端口的响应后,在后面的$(H-i)$个周期内不会处理该端口的请求.使用上述的调度机制,分组在光域的传输过程中不存在任何竞争,因此架构避免了对电环回缓存和FDL的使用,这在一定程度上降低了单一光交换模块的能耗和设计复杂度.

由于采用了低基数AWGR模块,上述两级和三级AWGR交换网络能够有效克服信号串扰和带宽偏移等物理层限制,因此能够扩展至数千节点规模.但若继续增大网络规模至数万节点以上,上述架构所需的波长资源和收发器的数量将成为严重的限制.对此,Yoo等人[43]提出了基于AWGR的分层网络.该架构如图10所示,其最底层的基本单元被称为picoBlade(pB).pB模块包括一个嵌入式电交叉开关和多个收发器(Transceiver,TRX).电交叉开关负责执行分组的路由功能并根据分组的目的地址选择合适的TRX发送该分组.p个pB模块构成一个microBlade(uB),m个uB模块进一步构成一个miniBlade(mB).在同一mB内,每个pB模块使用p个TRX完成同一uB内不同pB模块之间的通信,使用u个TRX完成不同uB中pB模块之间的通信.每个uB模块部署一个无源$(p+u)\times(p+u)$的AWGR,其中p个端口连接该uB内所有pB模块,其余u个端口连接其他uB模块的AWGR形成uB模块之间的通信连接.另外,每个uB模块存在一个中继pB模块.该模块包括一个TRX用于mB之间的通信,由于每个mB包括m个uB模块,因此每个mB共包括m个TRX可用于连接其他mB模块.为实现大规模mB模块之间的互连,作者提出了一种扁平分布式全光互连架构Thin-CLOS拓扑,$m\times W$个miniBlade模块被分成m个分离的小组.每个小组内部署有m个有源$W\times W$ AWGR交换机,本组内mB模块的输出端口和交换机输入端口之间形成完全二分图,而本组交换机的$m\times W$个输出端口依次连接所有mB模块的一个输入端口.与其他基于AWGR的交换网络相比较,该分级架构不仅具有更高的扩展性,可支持100 000节点以上的网络规模,同时具有较低的网络直径和较高的对分带宽.因此在均匀流量下该架构能够保证低时延传输,同时其饱和点可达到节点最大输出带宽的70%.

除利用AWGR外,还可以使用其他的光交换元件如半导体光放大器(SOA)或微环谐振器构建全光交换网络.其中,SOA是一种基于硅介质的PN结,光信号注入SOA后,在不同的外加电压下信号将获得增益或者产生损耗,由此实现光开关的"开"状态和"关"状态.Wang等人[44]提出了基于SOA的光交换网络.利用SOA本身所具备的双向透明传输特性,作者仅使用6个SOA单元设计了双向2×2光交叉开关(即传统的4×4光交叉开关).如图11所示,该交叉开关采用广播-选择式空分交换结构.在输入端口位置上,光信号经过1∶3分离器分割为3路信号并分别发往其他3个端口.部署在3条传输路径上的SOA根据配置对信号进行放大或衰减.在传统4×4光交叉开关中,每个输入端口都需要通过SOA连接到其他3个输出端口.而双向2×2光交叉开关允许两个输入端口共享同一个SOA单元,即若端口0的输入信号通过SOA传输到端口1,端口1的输入信号同样可通过该SOA传输到端口0.这种方式极大减小了SOA的使用数量,进而在交叉开关的成本、能耗和面积等方面获得了改进.利用该交叉开关单元,可以进一步构建Fat Tree等树形网络以实现大规模节点的互连.

微环谐振器是另外一种可实现快速空分交换的硅光器件.该器件是一个由波导制作的环形谐振腔,

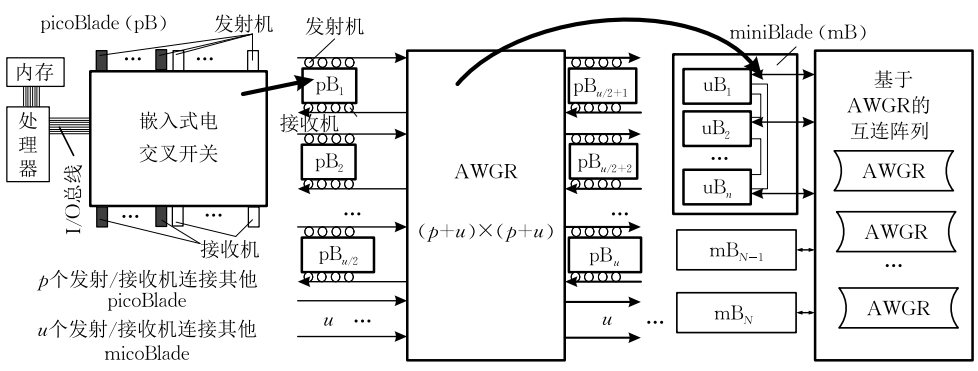


图10　基于AWGR的分层架构

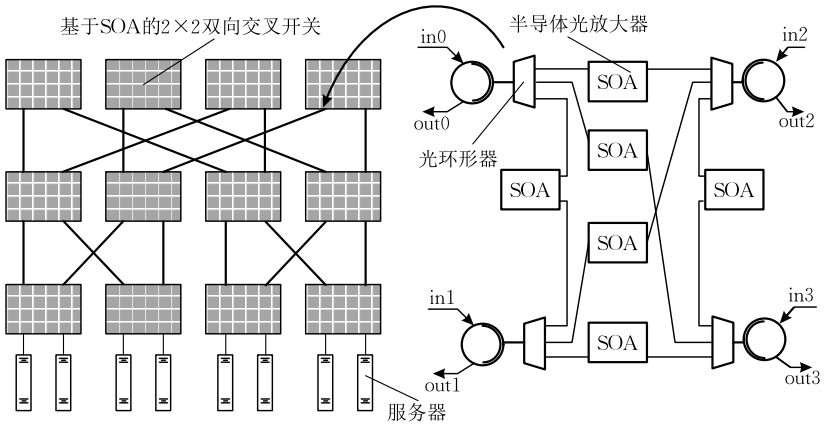


图 11 基于 SOA 的 Fat Tree 网络

其材料和直径将决定微环本身的谐振波长，通过注入载流子也可改变该微环的谐振波长. 微环部署在两根波导的交叉点位置. 若信号波长与微环的谐振波长相同，则信号首先耦合进微环，随后耦合到另外一根波导. 若信号的波长与微环的谐振波长不同，则信号沿着原来的波导传输. 由于微环谐振器能够良好地兼容 CMOS 平台，因此能够在制作、成本、面积等方面取得良好的特性. 基于微环谐振器的 $N\times N$ 光交叉开关由 $2N$ 根波导和 N^2 个微环谐振器组成. N 根输入波导和 N 根输出波导交叉布局形成 $N\times N$ 网格，微环部署在输入波导和输出波导的交叉点位置实现信号的耦合转向. 基于单晶硅制作的微环谐振器光交叉开关具备良好的光电特性. 但由于单晶硅无法实现多层沉积，所有的波导和微环必须部署在同一平面内，波导之间的交叉不可避免，这会带来较大的插入损耗. 对此，Biberman 等人[45]提出了基于微环的多层沉积光交叉开关结构. 在该结构中，输入波导和输出波导使用氮化硅材料构建并部署在最顶层和最底层，微环谐振器使用多晶硅材料构建并部署在中间层. 由于上下两层波导之间没有交叉，该多层沉积光交叉开关有效降低了插入损耗. 利用该光交叉开关模块，作者进一步构建了集中式数据中心光互连网络. 该网络使用多个机架构成一个簇结构，每个簇使用 WDM 技术汇聚簇间通信的流量，随后这些流量通过部署在顶层的 $N\times N$ 光交叉开关模块完成传输和交换.

随着云计算的发展，多种具有不同流量特性的新应用开始在数据中心内部署，灵活性成为网络设计的一个重要因素. 未来的数据中心网络应该能够满足每种应用所要求的服务质量，同时能够有效地利用资源并降低能耗. 混合全光互连网络针对不同的业务流量采用不同光交换机制，因此能够同时满足数据密集型业务对网络带宽的需求和突发性业务对传输时延的需求. Wang 等人[46]提出了基于 SOA 的 4×4 光分组/光电路混合交换节点. 如图 12 所示，该交换节点使用 16 个 SOA 单元构成广播-选择式门阵列. 这 16 个 SOA 被分成 4 个子集，每个子集对应于一个输入端口，每个子集内的 SOA 分别部署在该输入端口到其他输出端口的线路上. 该 4×4 交换节点中，每对输入到输出的连接都可以灵活配置为光电路交换模式或光分组交换模式. 若输入端口配置为光分组交换模式，部署在该输入端口的滤波器和光电探测器将提取每个分组的头信息进行路由计算和交叉开关配置. 但若输入端口配置为电路交换模式，该端口将忽略后续分组的头部信息，并在一定时长内维持 SOA 的配置状态. 另外，该交换节点采用丢弃分组的方式处理输出端口的竞争问题.

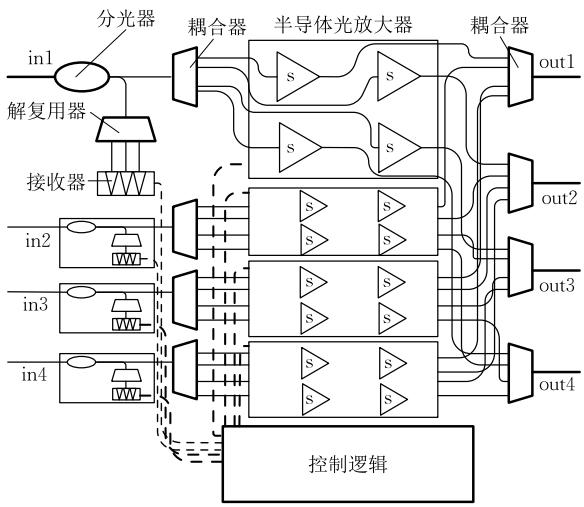


图 12 基于 SOA 的混合交换架构

Fiorani 等人[47]提出了一种结合光电路交换、光突发交换和光分组交换的混合交换网络 HOS. 该网络使用如图 13 所示的 3 层树形拓扑. 底层的每个机架内部署 N_s 台刀片式服务器，每台服务器通过 1Gbps 链路连接至 ToR 交换机. 多台 ToR 交换机

通过 40 Gbps 链路连接至汇聚层的 HOS 边缘交换机，最后所有 HOS 边缘交换机通过 WDM 光纤连接至顶层的一个核心交换机. 该核心交换机可以在逻辑上划分为 3 个模块：控制模块、交换模块和其他有源模块. 控制模块负责处理控制信息并控制数据的传输. 该模块具体包括 3 个单元：通用多协议标记交换(Generalized Multiprotocol Label Switching，GMPLS)控制单元、HOS 控制单元和交叉开关控制单元. 其中，GMPLS 控制单元用于实现 WAN 内与其他 HOS 核心节点的交互. HOS 控制单元用于管理光电路、光突发及光分组的传输. 交叉开关控制单元负责在光交叉开关中建立端到端的光路径. 交换模块包括两种全光交叉开关：慢速交叉开关用于实现光电路交换并负责传输长光突发块，快速交叉开关用于传输短突发块和光分组. 快速光交叉开关使用 SOA 构建，慢速光交叉开关使用 3D-MEMS 构建. 其他有源模块包括光放大器、可调波长转换器、控制信息提取器等. HOS 的边缘交换机主要用于实现流量的分类和汇聚等. HOS 能够根据应用的流量特点选择最合适的光传输机制，从而保证了网络的灵活性和较高的带宽利用效率. 但在核心层采用单一的交换节点可能会带来网络可靠性的问题，在文章中作者尚未提出该问题的解决方案.

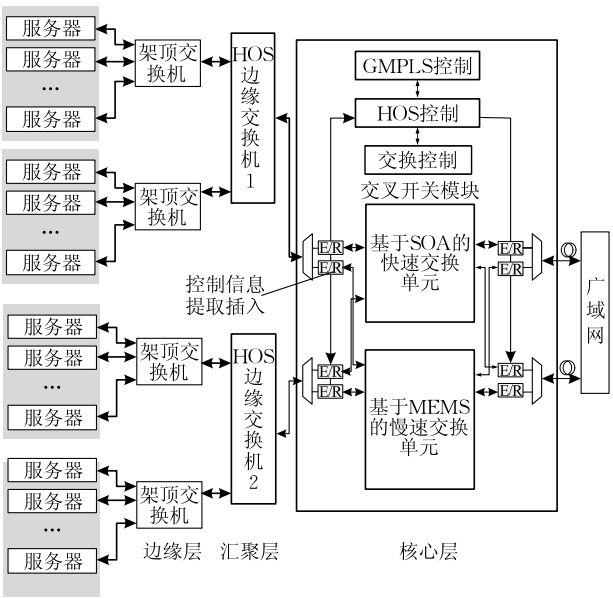


图 13　HOS 混合全光交换架构

3.2.2　分布式全光互连网络架构

尽管上述全光互连架构基于不同的光器件和不同光交换机制设计，但所有方案都使用单一的核心光交换机直接互连所有机架或簇结构. 这种互连方式的优势在于能够极大的简化网络管理和路由控制，符合目前工业界所提倡的扁平化网络设计思想. 但其劣势在于网络扩展性有限、核心光交换机内部结构设计复杂、网络可靠性较差等. 对此，一些研究人员提出了分布式全光互连网络架构.

Gumaste 等人提出的 FISSION 是一种基于多波长光总线的可扩展架构[48-49]. 如图 14 所示，该架构在核心层使用光环形链路互连多个载波以太网交换域，每个交换域被称为一个扇区. 每个扇区包括可重构光分插复用器(Reconfigurable Optical Add-Drop Multiplexer，ROADM)、汇聚交换机和边缘交换机等设备. 边缘交换机到 ROADM 的单向链路用于发送本扇区到其他扇区的流量，而 ROADM 到汇聚交换机的单向链路用于接收其他扇区到本扇区的流量. 扇区内通信由边缘交换机和汇聚交换机完成，服务器和存储设备连接到 $N\times 2N$ 的边缘交换机，边缘交换机进一步通过汇聚交换机相互连接. 对于扇区间的通信，在发送侧每个扇区通过 ROADM 和耦合器连接到一个特定的光纤环路上，在接收侧每个扇区可通过耦合器接收所有光纤环路的信号. 每个扇区被分配固定数目的波长用于发送扇区间流量，这些流量通过边缘交换机到达电-光交换机，在电-光交换机内转换成指定波长的光信号，随后通过复用器复用成一路 WDM 信号并通过耦合器发送到光纤环路. 作者进一步使用改进的载波以太网协议实现 FISSION 架构内的通信，该协议将整个网络转换为一系列二元树并为网络内的每个节点分配唯一的二元标签. 这样 FISSION 内的服务器可以利用这些标签实现源路由算法.

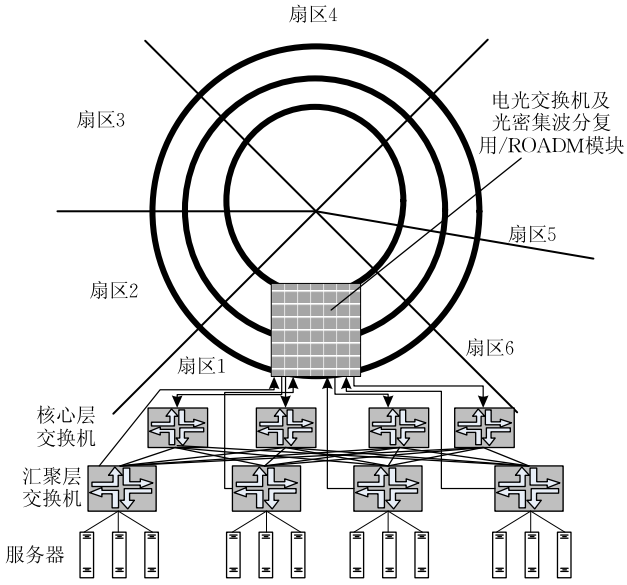


图 14　FISSION 全光互连架构

Hamedazimi 等人[50-51]提出了一种基于自由空间光技术的可重配置网络 FireFly. 通过利用自由空间光(Free Space Optical,FSO)技术,FireFly 可以避免使用集中式光交换架构处理机架间流量. 所有机架间的通信链路可以根据流量需求实现灵活的配置. 如图 15 所示,该网络使用传统电分组交换实现机架内通信,使用 FSO 链路实现机架间的灵活通信. 每台机架的顶部配置有一系列可调 FSO 设备. 为确保 FSO 发射器所发出的光束不会相互阻挡,FireFly 在机架顶部的天花板部署了反射镜面(Ceiling Mirror). 机架间通信的光信号首先从源机架发射至天花板镜面,经过反射后到达目的机架的接收端. 为节省能耗和成本,在发送侧光信号直接从光纤输出到自由空间,在接收侧光信号同样采用直接耦合的方式进入输入光纤. 输出端和输入端不部署任何光电转换器. FireFly 在光纤的焦距位置上部署了透镜以解决光束从光纤进入自由空间时的散射效应,在接收光纤的焦距上同样部署有透镜以实现信号的耦合. 为实现 FSO 链路的可重配置,架构使用了两种设备:可调镜面(Switchable Mirror,SM)和 Galvo 镜面. 其中,SM 是一种特殊的液晶材料,能够在电信号的控制下在反射和透射两种状态转换. 该元件采用如下方式动态配置 FSO 链路:每个 FSO 输出光纤配置有多个 SM 模块,每个 SM 在预配置阶段提前对准天花板镜面的一个目标反射点,该反射点进一步对准一台机架的 FSO 接收器. 在工作阶段,一个 SM 被设置为反射状态而其他 SM 被设置为透射状态,这样即可建立从源机架到特定目的机架的 FSO 链路. 改变 SM 模块组的状态即可建立到不同目的机架的 FSO 链路. Galvo 镜面可以在

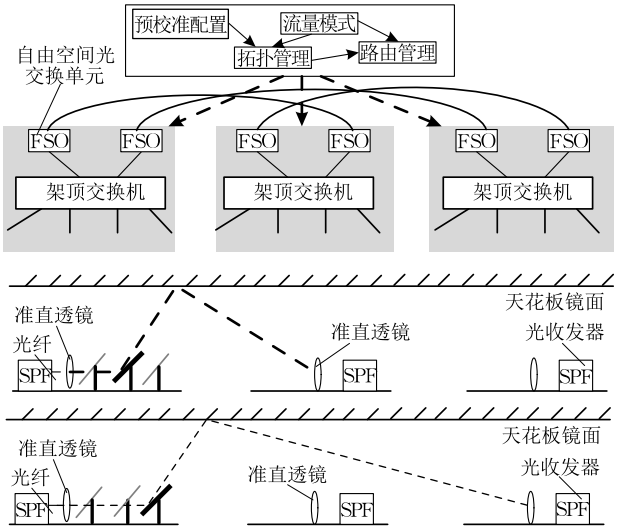


图 15 自由空间光网络 FireFly 架构

电信号的控制下绕固定轴转动,因此经过 Galvo 镜面和天花板镜面反射的光信号可以到达一个锥形区域. 通过配置 Galvo 镜面的角度,光束可以到达预定接收区域的任意一台机架. 除构建上述可重配置 FSO 链路外,还需要在控制层设计算法实现网络拓扑的优化配置. 在 FireFly 中,网络的配置包括两个阶段:在预配置阶段网络需要设置每个 FSO 可调镜面的反射点或 Galvo 镜面的覆盖区域,以形成多条备选网络链路;在实际运行阶段网络需要实时调整工作的 FSO 链路以满足当前的流量需求.

除上述两种典型的分布式光互连网络外,研究人员还提出了其他分布式光互连网络. Karthi 等人[52]提出了 POST 网络架构. 该网络使用传统的多层树形拓扑. 但与传统电互连网络不同的是,在 POST 中每层的多个节点(服务器或交换机节点)构成一个簇结构,簇内的所有节点直接通过一个无源光路由器(Passive Optical Router,POR)进行通信. 簇间的通信通过上层的交换设备完成. POST 的主要网络设备包括两种:POR 和互连交换机(Inter Connection Switch,ICS). 每个连接至 POR 的节点配置有一定数量的线卡,所有线卡通过电交换机进一步连接到一张光集成电路板卡(Photonic Integrated Circuit,PIC)上. 该电交换机负责将有通信需求的线卡接入 PIC 上的激光器阵列进行调制,不同波长的光信号进一步通过 N 端口 AWGR 实现交换. 在输出端,光信号通过 PIC 板卡上的光电探测器转换为电信号并进一步通过电交换机转发到目的节点的线卡. 每层簇内的节点除连接到 POR 外,还连接到上层的 ICS. 该 ICS 负责处理簇间输入流量和输出流量,同时负责控制簇内节点之间的通信. 因此,在 POST 网络的最底层(L0 层),N 台服务器构成一个簇. 簇内所有服务器连接至一台 POR,同时所有服务器连接到上层(L1 层)的一个 ICS. 在 L1 层,N 个 ICS 通过一台 POR 互连,同时这 N 个 ICS 进一步连接到 L2 层的一个 ICS. 这种互连方式逐级向上延续,直到根节点位置. 这种互连方式一方面保证了簇内节点的低时延、高带宽通信,另一方面有效增强了网络的可扩展性. Zhang 等人[53]提出了 OpenScale 网络架构. 该架构将“小世界”理论引入方案设计以充分实现应用对于数据中心网络高带宽、低时延、低平均距离的要求. 小世界网络能够通过在规则的网格拓扑中增加随机链路实现. 因此,作者首先使用光突发交换(Optical Burst Switching,OBS)环路构建基本的多边形网格拓扑,进一步使用波长交换技术随机地为远距离节点提供直连路径. 为同时实现光突发

交换和波长交换的功能，每个 OToR(Optical Top of Rack)交换机同时配置波长交换模块和光突发交换模块. 输入信号首先通过 WSS 分离为两组，一组波长信号被传输至波长交换模块，另外一组波长信号被传输至光突发交换模块. 其中，光突发交换模块采用简单的广播-选择式结构. 进入该模块的光信号通过 1∶2 耦合器复制为两路信号：一路信号发送至本地 OToR 收发器阵列；另外一路信号发送至下一跳节点，该路信号在到达输出端口前会通过可调光衰减器(Variable Optical Attenuator，VOA)以滤除信号中的无用信息. 波长交换模块将根据集中式控制器的命令配置交叉开关的输入输出端口，多个节点的波长交换模块通过协调配置即可形成端到端通信路径. 由于相邻节点之间可通过 OBS 环路实现快速灵活的光交换，远距离节点之间可通过波长交换形成直连路径. 因此 OpenScale 网络有效地实现了网络扩展性和网络直径之间的权衡. Saha 等人提出了基于递归定义的光互连网络 HyScale[54] 和 HyScaleII[55]. 这两种网络均使用 OToR 交换机作为基本交换设备. 每台 OToR 交换机配置有光交叉连接开关以连接机架内服务器和其他机架，该光交叉连接开关同时支持光电路交换和光突发交换. 其中，光电路交换用于传输象流，光突发交换用于传输鼠流. HyScale 使用了基于递归定义的拓扑结构，该拓扑可以表示为 $\phi(k,\Omega,\Gamma)$. 其中，参数 k 为拓扑的级数，Γ 为拓扑 ϕ 内子拓扑的数目，Ω 定义了 ϕ 内所有节点的地址空间. 在拓扑 ϕ 中每个节点使用一个 $(k+1)$元数组表示，同时 $\phi(k,\Omega,\Gamma)$通过互连 Γ 个 $\phi(k-1,\Omega,\Gamma)$构成. $\phi(0,\Omega,\Gamma)$为构建拓扑 ϕ 的基本单元，该基本单元使用 Γ 个机架互连成完全二分图. 随后 $\phi(k,\Omega,\Gamma)$使用 Γ 个下一级子拓扑 $\phi(k-1,\Omega,\Gamma)$构建. 若拓扑 ϕ 内节点 $I=\langle i_k,i_{k-1},\cdots,i_0\rangle$和 $J=\langle j_k,j_{k-1},\cdots,j_0\rangle$满足条件：$F_1(i_k)=j_k$，$F_2(i_k)=i_0$或 $F_3(i_k)=i_0$和 $i_0=j_0$，则节点 I 与节点 J 通过第 k 级链路互连. 其中，函数 F_1，F_2，F_3分别定义如下：

$$F_1(i)=(2z+1+i)\bmod\Gamma,\ z\in[0,1,\cdots,\Gamma/2-1],$$

$$F_2(i)=(z+i\bmod\Gamma/2)\bmod\Gamma,\ z\ \text{满足}\ F_1(i_k)=j_k,$$

$$F_3(i)=(F_2(i)+\Gamma/2)\bmod\Gamma.$$

HyScaleII 同样使用递归的方式构建，该拓扑可以使用 $\psi(k,\Omega,\Gamma)$表示. 其中 $\psi(0,\Omega,\Gamma)$与 $\phi(0,\Omega,\Gamma)$的结构相同. Γ 个子拓扑 $\psi(k-1,\Omega,\Gamma)$使用第 k 级链路互连构成 $\psi(k,\Omega,\Gamma)$. $\psi(k,\Omega,\Gamma)$中的每个节点的地址空间定义为一个 $k+1$ 元数组：$\Omega=\{\langle i_k,\cdots,i_0\rangle\,|\,i_l\in[0,1,\cdots,(\Gamma-1)]\}$，$\forall l=\{0,\cdots,k\}$. 若节点 $I=\langle i_k,\cdots,i_0\rangle$ 与 $J=\langle j_k,\cdots,j_0\rangle$ 满足条件 $F(i_k,i_0)=j_k$和 $i_0=j_0$，则节点 I 与节点 J 通过第 k 级链路互连. 其中，函数 F 的定义如下：

$$F(x,y)=\begin{cases}x\oplus y, & \text{若 } y\neq 0\\ \bar{x}, & \text{其他}\end{cases}.$$

HyScale 以及 HyScaleII 具有类似于 BCube、DCell、FiConn[56] 等网络拓扑的高扩展性、低网络直径、高对分带宽以及高容错性. HyScaleII 与 HyScale 相比，HyScaleII 具有更低的分组丢失率和更高的网络吞吐.

3.3　光互连架构的综合比较

表 2 综合对比了近年来研究人员所提出的数据中心光互连网络架构.

就网络架构类型而言，目前的设计方案主要包括光电混合互连网络和全光互连网络. 其中，光电混合互连网络在传统电互连网络架构的基础上额外增加了光互连网络，因此这些架构对传统数据中心网络具有较强的兼容性. 但光电混合互连网络无法在设备开销、能耗、网络管理等方面做出显著改进，因此研究全光互连网络架构成为主要的趋势. 表 3 使用实例对比了光电混合网络 Helios 和全光网络 Proteus 的开销、能耗和成本. 从该表格可以看出，与光电混合互连网络相比，全光网络能够提供更高的带宽并有效降低网络的设备开销和能耗. 但由于全光互连网络的技术革新性更强，需要替换所有传统电交换设备以实现网络的部署. 因此全光互连网络在设计中需要权衡更多的因素，如扩展性、可升级性、设备开销、技术可行性等.

就网络的扩展性而言，由于目前商用全光交换机的最大规模在数百端口[58]，AWGR 模块的端口数目也会受到信号串扰、波长漂移等因素的限制[59]，因此在实际部署中，集中式交换架构的扩展规模十分有限. 为解决该问题，研究人员需要设计定制的交换机结构，即使用 Clos[41]、Benes[60]、Spank[61] 等多级拓扑进一步将单一光交换模块互连成大规模多级光交换矩阵，但这会在一定程度上增加网络的成本. 分布式光交换网络可提供更加灵活的扩展方式，同时可支持的网络规模也远大于集中式架构，但分布式光交换网络需要复杂的通信协议实现交换机之间的协调.

可升级性也是数据中心网络架构设计考虑的重要因素之一，它主要是指随着链路速率和服务器接口速率的提升，单一的交换节点是否能够快速升级以支持更高的交换容量. 在现有的设计方案中，基于 MEMS 光电路交换和基于自由空间光交换的网络，如 Helios、c-Through、Proteus、FireFly 等，其核心光

表 2 数据中心光互连架构的综合比较

架构名称	类型	交换机制	扩展性	容量限制	技术进展	设备开销	突出特点
Helios[14]	分布式光电混合架构	电分组/光电路交换	中	速率透明	商用设备	低	集中式流量调度
c-Through[21]	集中式光电混合架构	电分组/光电路交换	低	速率透明	商用设备	低	ToR 位置上基于 VLAN 的流量调度
混洗-交换网络[22]	集中式光电混合架构	电分组/光电路交换	高	速率透明	商用设备	中	基于多跳的光电路交换
基于光突发环的光电混合网络[23]	分布式光电混合架构	电分组/光突发交换	高	速率透明	商用设备	高	基于广播-选择式光突发交换
Proteus(OSA)[30-31]	集中式全光互连架构	光电路交换	低	速率透明	商用设备	低	可实现波长信道和网络连接的灵活配置
DOS[33-34]	集中式全光互连架构	光分组交换	低	TWC 的调制速率	实验原型[56]	低	利用 AWGR 波长路由特征实现高吞吐交换
基于 FDL 的 AWGR 架构[37-38]	集中式全光互连架构	光分组交换	低	TWC 的调制速率	软件仿真	低	使用不同 FDL 结构实现光分组缓存
基于 OFDM 的 AWGR 架构[39]	集中式全光互连架构	光分组交换	低	子载波的调制速率	实验原型	低	使用 OFDM 技术解决分组的竞争问题
两级 AWGR 架构[40]	集中式全光互连架构	分组交换	中	TWC 的调制速率	实验原型	中	使用小规模 AWGR 模块解决串扰及波长漂移问题
Petabit[41-42]	集中式全光互连架构	分组交换	高	TWC 的调制速率	软件仿真	高	利用电域的仲裁解决分组的竞争问题
分层 AWGR 架构[43]	集中式全光互连架构	分组交换	高	TWC 的调制速率	软件仿真	高	多层光交换架构,扩展规模较高
基于 SOA 的 Fat Tree 架构[44]	分布式全光互连架构	分组/突发交换	中	SOA 支持的传输速率	实验原型	中	利用 SOA 的双向传输特性减小器件开销
基于微环谐振器的光交换架构[45]	集中式全光互连架构	可实现分组粒度交换	中	微环谐振器支持的传输速率	建模分析	低	新硅光器件-微环谐振器引入数据中心光互连
基于 SOA 的光分组/光电路混合交换架构[46]	集中式全光互连架构	光电路/光分组交换	低	SOA 支持的传输速率	实验原型	低	在控制层实现光电路和光分组的可重配置选择
HOS[47]	集中式全光互连架构	光电路/光突发/光分组交换	中	光电路部分对速率透明,光突发/光分组部分速率受 SOA 单元限制	建模分析	中	在数据传输层实现光电路/光突发/光分组的混合交换
FISSION[48-49]	分布式全光互连架构	光电路交换	高	速率透明	软件仿真	中	利用多波长技术及光总线技术提高网络吞吐,同时引入改进的载波以太网协议
FireFly[50-51]	分布式全光互连架构	光电路交换	中	速率透明	实验原型	低	将基于自由空间的光交换技术引入数据中心网络
POST[52]	分布式全光互连架构	类似于光电路交换	高	激光器阵列的调制速率	建模分析	高	新的基于 AWGR 分布式互连结构
OpenScale[53]	分布式全光互连架构	光突发/光电路交换	高	速率透明	软件仿真	高	利用光突发交换和光电路交换实现小世界网络
HyScale[54]、HyScaleII[55]	分布式全光互连架构	光突发/光电路交换	高	——	软件仿真	高	使用递归定义的方式构建光网络拓扑

表 3 Helios 和 Proteus 的能耗、成本、开销对比实例

架构名称	服务器数目	收发器数目	交换机端口数目	光纤数目	成本/$	能耗/W
Helios	65536	98304	36864	36864	39936000	508887.04
Proteus	65536	65536	8192	8192	17612800	67502.08

注:本表格采用文献[14]所提供的模型及基本数据进行计算.

交换单元对于信号的速率完全透明.这类网络中单一交换节点的交换容量仅取决于交换机的端口数目,因此最易实现网络交换容量的升级.基于广播-选择式光突发网络,如基于光突发环的混合网络、FISSION 等,由于网络中并不存在实质的光交换单元,因此这类网络也对信号速率透明.基于 AWGR 的网络架构中,波长转换器所支持的数据速率将限制波长信道的传输带宽.最后,在基于 SOA 和微环谐振器的网络架构中,SOA 单元和微环谐振器所支持的数据传输速率将决定网络支持的最大数据传输速率.

各网络架构的技术研究进展在一定程度上反映了该设计方案能否快速部署到实际应用之中.相对而言,基于 MEMS 的网络架构直接使用商用交换机

构建，因此易于实现和部署. 基于 AWGR 和 SOA 的光交换网络大多处于实验原型阶段，尽管这些架构已经具备技术可行性，但由于成本过高，因此该类光器件仍然会限制网络的实际部署. 基于微环谐振器的网络架构和复杂的分布式光互连架构尚处于软件仿真阶段，这些网络架构需要经历更长的时间才能实现实际的应用和部署.

网络的设备开销将关系到网络的构建成本和管理复杂度. 集中式光互连架构能够有效降低网络所需的设备数量并简化控制管理. 但如前所述，该类架构的扩展规模受到较大限制. 分布式光互连架构所需的网络设备较多，同时需要设计更复杂的通信协议，但分布式架构支持更灵活的扩展模式，同时能够避免对于定制设备的使用.

4　研究趋势

与电互连技术相比，研究人员对数据中心光互连网络的研究尚处于起步阶段. 这主要体现在：(1) 多数光互连方案仅处于原型验证或软件仿真阶段，这些方案缺乏低成本、成熟的商用光器件支持部署；(2) 光互连低成本、低能耗、高带宽的优势尚未充分体现；(3) 工业界主要通信设备提供商，如思科、中兴、华为、Juniper 等，尚未就数据中心光互连网络推出相关产品设备. 国际标准化组织如 IEEE、IETF、ITU-T 等也尚未提出与数据中心光互连相关的技术标准或建议. 因此，数据中心光互连技术还需要进行更加广泛深入的研究，未来的研究方向将主要集中于以下几个方面：

(1) 新型光交换机结构的研究

光交换节点的性能会直接影响网络拓扑、交换机制和通信策略的设计. 目前商用 MEMS 全光交换机的配置时延过长(数十毫秒)，因此基于 MEMS 的光互连架构难以实现细粒度的信息交换，网络灵活性和链路利用率受到严重限制. AWGR 需要在每个输入端口配置 TWC 以实现分组的正确路由，这些有源器件会增大光交换网络的能耗. 另外，基于光纤延迟线的缓存结构无法实现光分组的灵活存取，缓存的容量也十分有限. 现存光交换机的上述缺点会导致光互连架构设计中存在各种限制，也导致目前光网络无法获得与电网络相似的灵活性和信息交换粒度. 因此，未来对于新型光交换机的研究会成为重点，低能耗、低成本、细粒度、快速配置的光交换机结构会极大改善光互连网络的性能.

(2) 分布式光互连网络的研究

目前数据中心光互连网络多采用集中式架构，这些方案在网络的扩展性、可靠性、技术可行性等方面存在较大的挑战. 分布式光互连网络具有更灵活的扩展模式，同时能够避免对定制光交换设备的使用. 但目前研究领域对于分布式光互连网络的研究较少，因此未来在数据中心分布式光互连网络设计方面仍存在较大的研究空间. 研究人员需要解决的问题包括网络拓扑的设计、通信协议的设计、控制系统的设计等.

(3) 与云计算特定应用相结合或与新技术相结合的光互连网络架构设计

目前的光互连架构设计主要关注网络性能、扩展性、成本等基本因素，结合云环境或特定应用要求的光互连网络架构研究较少. 例如，虚拟化是云计算的主要特征[62]，随着服务器虚拟化、存储虚拟化技术的发展，网络虚拟化技术备受关注[63]. 另外由于云计算数据中心面向全球用户提供不间断服务，可靠性成为网络设计的重要要求. 最后，云计算数据中心是一个典型的多租户环境，不同应用的通信模式、流量特征和 QoS 要求互不相同，网络的流量调度策略和交换传输策略应该充分考虑不同应用的要求. 因此，基于光互连网络架构的虚拟化技术、可靠性保障技术和改进型传输协议都值得进行深入研究. SDN 技术近年来受到工业界和学术界的极大关注[64]，它为网络运营者和研究人员提供了一个开放性平台，可使创新性技术和协议实现快速地应用部署. 将 SDN 技术应用于数据中心光互连网络成为一个新的研究课题，对此，北京邮电大学就这一课题展开了研究[65]，可以预期这一方向将产生更多的研究成果.

5　总　结

云计算将带来数据中心内应用、网络、计算、存储的全面融合. 随着 Hadoop、MapReduce 等分布式计算模式的部署，服务器之间的任务协作增加，网络成为决定云计算性能的关键因素之一. 与电互连技术相比，光互连技术在能耗、带宽、传输透明性等方面占有较大的优势. 尤其随着链路带宽的提升，光纤成为数据中心的主要传输媒质，全光互连架构将进一步降低网络的传输时延和能耗. 因此，数据中心光互连架构受到广泛关注，已获得的研究成果表明，数据中心光互连技术具有较大的潜力解决云计算网络

所面临的一系列问题.尽管该领域仍存在一系列待解决的问题,但可以预期的是,随着光器件的成熟和研究的深入,低成本、低能耗、高带宽、低时延的数据中心光互连架构将有力推动云计算技术的发展.

参 考 文 献

[1] Zhang Q, Cheng L, Boutaba R. Cloud computing: State-of-the-art and research challenges. Journal of internet services and applications, 2010, 1(1): 7-18

[2] Zhang Y, Ansari N. On architecture design, congestion notification, TCP incast and power consumption in data centers. IEEE Communications Surveys & Tutorials, 2013, 15(1): 39-64

[3] Li Dan, Chen Gui-Hai, Ren Feng-Yuan, et al. Data center network research progress and trends. Chinese Journal of Computers, 2014, 37(2): 259-274(in Chinese)
(李丹,陈贵海,任丰原等. 数据中心网络的研究进展与趋势. 计算机学报, 2014, 37(2): 259-274)

[4] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture//Proceedings of the Special Interest Group on Data Communication(SIGCOMM). Seattle, USA, 2008: 63-74

[5] Greenberg A, Hamilton J R, Jain N, et al. VL2: A scalable and flexible data center network//Proceedings of the Special Interest Group on Data Communication(SIGCOMM). Barcelona, Spain, 2009: 95-104

[6] Guo C, Wu H, Tan K, et al. Dcell: A scalable and fault-tolerant network structure for data centers//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). Seattle, USA, 2008: 75-86

[7] Guo C, Lu G, Li D, et al. BCube: A high performance, server-centric network architecture for modular data centers//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). Barcelona, Spain, 2009: 63-74

[8] Abu-Libdeh H, Costa P, Rowstron A, et al. Symbiotic routing in future data centers//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 51-62

[9] Liu Xiao-Qian, Yang Shou-Bao, Guo Liang-Min, et al. Snowflake: A new-type network structure of data center. Chinese Journal of Computers, 2011, 34(1): 76-86 (in Chinese)
(刘晓茜,杨寿保,郭良敏等. 雪花结构:一种新型数据中心网络结构. 计算机学报, 2011, 34(1): 76-86)

[10] Glesk I, Buis A, Davidson A. How photonic networking can help data centers//Proceedings of the 16th International Conference on Transparent Optical Networks (ICTON). Graz, Austria, 2014: 1-4

[11] Guo Bing, Shen Yan, Shao Zi-Li. The redefinition and some discussion of green computing. Chinese Journal of Computers, 2009, 32(12): 2311-2319(in Chinese)
(郭兵,沈艳,邵子立. 绿色计算的重定义与若干讨论. 计算机学报, 2009, 32(12): 2311-2319)

[12] Baliga J, Ayre R W A, Hinton K, et al. Green cloud computing: Balancing energy in processing, storage, and transport. Proceedings of the IEEE, 2011, 99(1): 149-167

[13] Bilal K, Malik S U R, Khalid O, et al. A taxonomy and survey on green data center networks. Future Generation Computer Systems, 2014, 36(294): 189-208

[14] Farrington N, Porter G, Radhakrishnan S, et al. Helios: A hybrid electrical/optical switch architecture for modular data centers//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 339-350

[15] Binkert N, Davis A, Jouppi N P, et al. The role of optics in future high radix switch design//Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA). San Jose, USA, 2011: 437-447

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

[17] Kliazovich D, Bouvry P, Khan S. DENS: Data center energy-efficient network-aware scheduling. Cluster Computing, 2013, 16(1): 65-75

[18] Lam C F. Optical network technologies for datacenter networks (invited paper)//Proceedings of the Optical Fiber Communication Conference collocated National Fiber Optic Engineers Conference (OFC/NFOEC). San Diego, USA, 2010: 1-3

[19] Zhao X, Vusirikala V, Koley B, et al. The prospect of inter-data-center optical networks. IEEE Communications Magazine, 2013, 51(9): 32-38

[20] Hong L, Lam C F, Johnson C. Scaling optical interconnects in datacenter networks opportunities and challenges for WDM//Proceedings of the IEEE 18th Annual Symposium on High Performance Interconnects (HOTI). Mountain View, USA, 2010: 113-116

[21] Wang G, Andersen D G, Kaminsky M, et al. c-Through: Part-time optics in data centers//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 327-338

[22] Lugones D, Katrinis K, Collier M. A reconfigurable optical/electrical interconnect architecture for large-scale clusters and datacenters//Proceedings of the 9th Conference on Computing Frontiers. Cagliari, Italy, 2012: 13-22

[23] Li C Y, Deng N, Li M, et al. Performance analysis and experimental demonstration of a novel network architecture using optical burst rings for interpod communications in data centers. IEEE Journal of Selected Topics in Quantum Electronics, 2013, 19(2): 3700508-3700508

[24] Wang G, Andersen D G, Kaminsky M, et al. Your data center is a router: The case for reconfigurable optical circuit switched paths//Proceedings of the ACM Workshop on Hot Topics in Networks (Hotnets). New York, USA, 2009: 1-6

[25] Bazzaz H H, Tewari M, Wang G, et al. Switching the optical divide: Fundamental challenges for hybrid electrical/optical datacenter networks//Proceedings of the 2nd ACM Symposium on Cloud Computing. Cascais, Portugal, 2011: 1-8

[26] Porter G, Strong R, Farrington N, et al. Integrating microsecond circuit switching into the data center//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). New York, USA, 2013: 447-458

[27] Farrington N, Fainman Y, Liu H, et al. Hardware requirements for optical circuit switched data center networks//Proceedings of the Optical fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC'11). Los Angeles, USA, 2011: 1-3

[28] Deng N, Xue Q, Li M, et al. An optical multi-ring burst network for a data center//Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Anaheim, USA, 2013: 1-3

[29] Tekin T, Pleros N, Apostolopoulos D. Photonic interconnects for data centers//Proceedings of the Optical Fiber Communications Conference and Exhibition (OFC). San Francisco, USA, 2014: 1-3

[30] Singla A, Singh A, Ramachandran K, et al. Proteus: A topology malleable data center network//Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks. Monterey, USA, 2010: 1-6

[31] Chen K, 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

[32] Xu L, Singh A, Zhang X. Optically interconnected data center networks//Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Los Angeles, USA, 2012: 1-3

[33] Yoo S J B, Yin Y, Wen K. Intra and inter datacenter networking: The role of optical packet switching and flexible bandwidth optical networking//Proceedings of the 16th International Conference on Optical Network Design and Modeling (ONDM). Colchester, UK, 2012: 1-6

[34] Ye X, Proietti R, Yin Y, et al. Buffering and flow control in optical switches for high performance computing. IEEE/OSA Journal of Optical Communications and Networking, 2011, 3(8): A59-A72

[35] Yin Y, Proietti R, Ye X, et al. LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers. IEEE Journal of Selected Topics in Quantum Electronics, 2013, 19(2): 3600409-3600409

[36] Proietti R, Nitta C J, Yin Y, et al. Scalable and distributed contention resolution in AWGR-based data center switches using RSOA-based optical mutual exclusion. IEEE Journal of Selected Topics in Quantum Electronics, 2013, 19(2): 3600111-3600111

[37] Rastegarfar H, Leon-Garcia A, LaRochelle S, et al. Cross-Layer performance analysis of recirculation buffers for optical data centers. Journal of Lightwave Technology, 2013, 31(3): 432-445

[38] Rastegarfar H, Ann Rusch L, Leon-Garcia A. WDM recirculation buffer-based optical fabric for scalable cloud computing. Journal of Lightwave Technology, 2014, 32(21): 3451-3465

[39] Ji P N, Dayou Q, Kanonakis K, et al. Design and evaluation of a flexible-bandwidth OFDM-based intra-data center interconnect. IEEE Journal of Selected Topics in Quantum Electronics, 2013, 19(2): 3700310-3700310

[40] Sato K, Hasegawa H, Niwa T, et al. A large-scale wavelength routing optical switch for data center networks. IEEE Communications Magazine, 2013, 51(9): 46-52

[41] Chao H J, Kang X. Bufferless optical Clos switches for data centers//Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Los Angeles, USA, 2011: 1-3

[42] Kachris C, Bergman K, Tomkos I, et al. Optical Interconnects for Futhure Data Center Networks. New York: Springer Science & Business Media, 2013

[43] Proietti R, Cao Z, Li Y, et al. Scalable and distributed optical interconnect architecture based on AWGR for HPC and data centers//Proceedings of the Optical Fiber Communications Conference and Exhibition (OFC). San Francisco, USA, 2014: 1-3

[44] Wang H, Bergman K. A bidirectional 2×2 photonic network building-block for high-performance data centers//Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Los Angeles, USA, 2011: 1-3

[45] Biberman A, Hendry G, Chan J, et al. CMOS-compatible scalable photonic switch architecture using 3D-integrated deposited silicon materials for high-performance data center networks//Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Los Angeles, USA, 2011: 1-3

[46] Wang H, Garg A S, Bergman K, et al. Design and demonstration of an all-optical hybrid packet and circuit switched network platform for next generation data centers//Proceedings of the Optical Fiber Communication and National Fiber Optic Engineers Conference (OFC/NFOEC). San Diego, USA, 2010:1-3

[47] Fiorani M, Casoni M, Aleksic S. Large data center interconnects employing hybrid optical switching//Proceedings of the 8th Conference on Network and Optical Communications (NOC) and 18th European Conference on and Optical Cabling and Infrastructure (OC&I). Graz, Austria, 2013: 61-68

[48] Gumaste A, Bheri B M K. On the architectural considerations of the FISSION (Flexible Interconnection of Scalable Systems Integrated using Optical Networks) framework for data-centers//Proceedings of the 17th International Conference on Optical Network Design and Modeling (ONDM). Brest, France, 2013: 23-28

[49] Gumaste A, Bheri B M K, Kshirasagar A. FISSION: Flexible interconnection of scalable systems integrated using optical networks for data centers//Proceedings of the IEEE International Conference on Communications (ICC). Budapest, Hungary, 2013: 3963-3968

[50] Hamedazimi N, Gupta H, Sekar V, et al. Patch panels in the sky: A case for free-space optics in data centers//Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks. College Park, USA, 2013: 1-7

[51] Hamedazimi N, Qazi Z, Gupta H, et al. FireFly: A reconfigurable wireless data center fabric using free-space optics//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). Chicago, USA, 2014: 319-330

[52] Karthi D, Das G. POST: A scalable optical data center network//Proceedings of the 2013 IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS). Kattankulathur, India, 2013: 1-3

[53] Zhang D, Wu J, Guo H, et al. An optical switching architecture for intra data center interconnections with ultra-high scalability//Proceedings of the IEEE Optical Interconnects Conference. San Diego, USA, 2014: 45-46

[54] Saha S, Deogun J S, Xu L. HyScale: A hybrid optical network based scalable, switch-centric architecture for data centers//Proceedings of the IEEE International Conference on Communications (ICC). Ottawa, Canada, 2012: 2934-2938

[55] Saha S, Deogun J S, Xu L. HyScaleII: A high performance hybrid optical network architecture for data centers//Proceedings of the 35th IEEE Sarnoff Symposium (SARNOFF). Newark, USA, 2012: 1-5

[56] Li D, Guo C, Wu H, et al. FiConn: Using backup port for server interconnection in data centers//Proceeding of the IEEE INFOCOM 2009. Rio de Janeiro, Brazil, 2009: 2276-2285

[57] Proietti R, Ye X, Yin Y, et al. 40 Gb/s 8×8 low-latency optical switch for data centers//Proceedings of the Optical Fiber Communication Conference and National Fiber Optic Engineers Conference (OFC/NFOEC). Los Angeles, USA, 2011: 1-3

[58] Vahdat A, Liu H, Zhao X, et al. The emerging optical data center//Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Los Angeles, USA, 2011: 1-3

[59] Niwa T, Hasegawa H, Sato K, et al. Large port count wavelength routing optical switch consisting of cascaded small-size cyclic arrayed waveguide gratings. IEEE Photonics Technology Letters, 2012, 24(22): 2027-2030

[60] Garao D G, Maier G, Pattavina A. Modular architectures of optical multi-stage switching networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Turin, Italy, 2013: 555-559

[61] Luo J, Di Lucente S, Ramirez J, et al. Low latency and large port count optical packet switch with highly distributed control//Proceedings of the Optical Fiber Communication Conference and Exposition (OFC) and the National Fiber Optic Engineers Conference (NFOEC). Los Angeles, USA, 2012: 1-3

[62] Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. Communications of the ACM, 2010, 53(4): 50-58

[63] Jain R, Paul S. Network virtualization and software defined networking for cloud computing: A survey. IEEE Communications Magazine, 2013, 51(11): 24-31

[64] Hyojoon K, Feamster N. Improving network management with software defined networking. IEEE Communications Magazine, 2013, 51(2): 114-119

[65] Zhang J, Zhao Y, Yang H, et al. First demonstration of enhanced software defined networking (eSDN) over elastic grid (eGrid) optical networks for data center service migration//Proceedings of the Optical Fiber Communication Conference and Exposition (OFC) and the National Fiber Optic Engineers Conference (NFOEC). Anaheim, USA, 2013: 1-3

YU Xiao-Shan, born in 1986, Ph. D. candidate. His current research interest is optical interconnects for cloud computing data center.

WANG Kun, born in 1982, M. S., lecturer. Her current research interests include cloud computing, network virtualization technologies.

GU Hua-Xi, born in 1977, Ph. D., professor, Ph. D. supervisor. His current research interests include data center network architecture, high-performance optical interconnects.

WANG Xi, born in 1991, M. S. candidate. His current research interests include optical interconnections and software defined network.

Background

Cloud computing enables a novel model for accessing a shared pool of computing resources and enjoying the on-demand self-services. Data center is the main platform to host various computing resources and provides cloud-based applications. In cloud computing data centers, the servers need the high interaction with each other. Thus the increasing data-intensive traffic flows pose a great challenge to the network. Limited by the bandwidth and power consumption, the electronic networks cannot provide a long-term solution for cloud computing. The optical interconnections, which potentially sustain the increased bandwidth with low latency and reduced power consumption, have attracted great attention. A survey on optical interconnection for cloud computing data center is presented in this paper. Moreover, a comparison of the proposed optical schemes is conducted and the future trends are discussed.

This work is supported by the National Science Foundation of China under Grant No. 61472300, the Fundamental Research Funds for the Central Universities under Grant Nos. JB150318 and JB142001-5, and the 111 Project under Grant No. B08038.