doi:10.3772/j.issn.1002-0470.2019.02.001

基于 WSS 的动态可重构光网络①

元国军②[*,**] 肖 鹏[**] 姜 涛[*] 王 展[*] 杨 帆[**] 曹 政[*] 张佩珩[*] 谭光明[*] 孙凝晖[*]

([*]中国科学院计算技术研究所 北京 100190)

([**]中国科学院大学 北京 100049)

摘 要 数据中心负载通信特征多种多样,可重构网络可以实现逻辑连接关系的动态变换以匹配不同的应用通信特征,是提高互连网络资源利用率、降低系统能耗、提高灵活性的有效方法之一。传统的电域互连技术很难实现物理拓扑的动态切换,本文基于新兴的光波长选择开关(WSS)技术提出了一种动态可重构光电混合网络结构,在不改变物理连接关系的条件下通过软件配置 WSS 实现拓扑连接的动态重构和互连带宽的最优调整;基于 Misra&Gries 算法和 Greedy 算法提出了面向可重构光网络的快速配置算法,可在满足光器件物理约束的条件下快速求解出典型拓扑对应的光波长配置参数。本文给出了多种物理拓扑的重构过程,1 024 个节点的仿真结果显示,合适的拓扑重构对网络带宽的提升超过 60%。

关键词 数据中心,光电混合网络,可重构网络,波长选择开关(WSS)

0 引 言

超级计算机和数据中心是现代信息时代的核心基础设施,承载着互联网、云计算、科学计算等多种类型的应用[1],应用之间通信模式差异很大[2],不同的网络拓扑适合用于不同的通信模式,因此实现互连网络拓扑与应用通信特征的动态匹配,即动态可重构网络,是提高数据中心资源利用率、降低系统能耗、提高网络灵活性的有效方法之一。

传统计算机系统域互连多采用电域交换技术,在完成布局布线后,拓扑结构就固定下来,灵活性较差,难以实现物理拓扑的动态重构。随着硅光子技术的突破,不少光交换技术已应用于计算机互连网络中,成为未来互连网络的新兴使能技术,例如基于波长选择开关(wavelength selective switch, WSS)[3]和阵列波导光栅(arrayed waveguide grating router, AWGR)[4]的光波长交换技术、基于大规模微机电系统(micro-electro-mechanical system, MEMS)的光空间交换技术等[5],都适合成为解决光网络动态重构问题的备选技术。这些光交换互连技术除了具备高带宽、低延迟和低损耗特征外,还支持端口之间连接关系的动态配置,适合用于实现动态可重构网络。

本文的主要贡献是:(1)基于新兴的 WSS 技术提出了一种全新的动态可重构光互连结构。将整个网络数据平面分为交换和互连两部分,融合电分组交换和光路交换技术,其中电交换器件负责交换功能,光交换器件负责连接功能,充分利用光器件的灵活性来动态改变节点间的连接关系,实现网络拓扑结构的动态重构,譬如同一个物理结构可实现 HyperX[6]和 Torus[7]之间的动态重构。(2)提出了一种快速的易扩展的光交换网络重构配置算法。对于可重构光网络结构,如何通过求解光交换器件的配置参数将网络配置为需要的拓扑结构,是一个重要

① 国家重点研发计划(2016YFB0200205),中科院战略性先导科技专项(XDB24050200)和国家自然科学基金(61572464, 61331008)资助项目。

② 男,1983 年生,博士生,高级工程师;研究方向:计算机系统结构,数据中心光互连网络等;联系人,E-mail: yuanguojun@ncic.ac.cn

(收稿日期:2018-04-12)

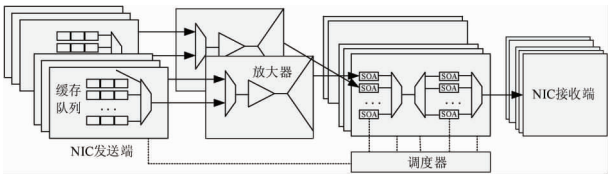
的研究内容,尤其当网络节点规模增大时,求解和配置都将变得更加困难。本文将光网络配置方法映射为图的边着色问题,实现了光网络拓扑结构动态配置和链路带宽的灵活调整,而且不需关注网络历史分配状态,算法时间复杂度低且灵活性高。

1　研究现状

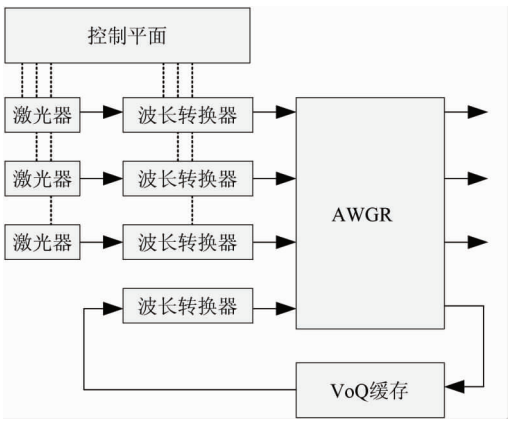
国内外已经开展了不少光交换网络的研究,主要分为两类:光电混合网络和全光交换网络。在全光交换网络中,所有的数据交换功能均在光域完成,典型结构比如 IBM 的 OSMOSIS 系统[8],该系统基于半导体光放大器(semiconductor optical amplifier, SOA)[9]阵列实现,如图 1(a)所示,采用广播-选择的交换策略,来自节点的光信号在广播阶段被广播至一个 SOA 阵列,然后由中央调度器根据网络接口卡(network interface card,NIC)的发送请求,将 SOA 阵列中的 SOA 配置到所需的开关状态来实现交换功能。加州大学戴维斯分校的 DOS (datacenter optical switch)[10]基于 AWGR 实现了单级交换结构,如图 1(b)所示,控制平面通过解析消息的目的地址,控制可调谐波长变换器(tunable wavelength converter, TWC)[11]进行相应的波长切换来完成网络交换功能,其他的全光交换结构还包括多级 AWGR 构成的 Petabit [12]和哥伦比亚大学的 Data vortex[13]等结构。

在光电混合网络(图 2)中,光网络和电网络同时参与数据交换,其中光网络使用毫秒级光线路交换(optical circuit switching,OCS)[14]技术,典型结构是莱斯大学 Wang 等人[15]提出的 c-Through 架构,将机顶交换机(Top of Rack, ToR)同时连接到基于包交换的电网络和基于 OCS 的光网络,电网络使用传统的带宽逐层等比缩减的胖树结构,光网络使用 MEMS 光交换机连接所有 ToR 的结构。加州大学圣地亚哥分校的 Farrington 等人[16]提出了 Helios 结构,光网络主要服务于 Pod(由若干个机柜组成)的长时间大数据量传输,如虚拟机迁移和数据备份。

已有的光网络研究工作致力于充分发挥光交换高带宽、低延迟特性,有些网络也具备一定灵活性,比如构建灵活的光快速链路来缓解热点通信,但对于如何构造面向所有节点全局拓扑可重构光网络(经典拓扑 Torus、HyperX、Dragonfly 等的动态配置)研究得不够深入和全面,特别是应用于数据中心或者高性能计算机中大规模节点之间的互连时,存在着扩展性、重构速度等方面的制约。



(a) IBM 的 OSMOSIS



(b) 基于 AWGR 的 DOS 结构

图 1　典型的全光网络结构

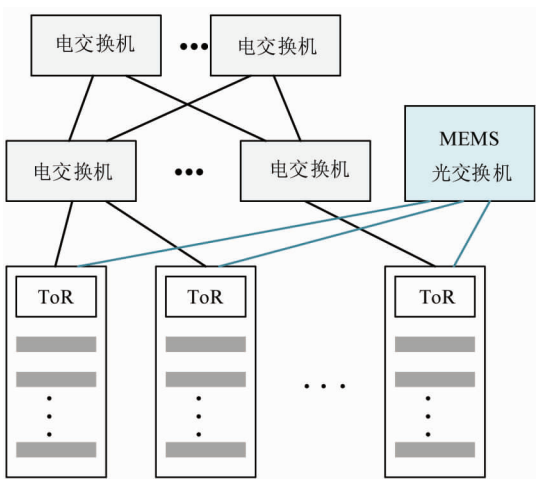


图 2　光电混合网络结构

2　可重构光网络物理结构

现有可重构光交换技术主要基于 2 种光交换原

理来实现:光波长交换和光空间交换。光空间交换器件切换延时较大,比如基于 MEMS 技术的光交换机,通常需要毫秒级或者更长时间,这也限制了它的应用范围;光波长交换器件切换速度可以达到微秒级,比如 WSS 技术。考虑到数据中心和超算中心很多场景下流量特征变化很频繁,数据流持续时间短(比如带宽小的"老鼠流"),因此本文采用光波长交换器件 WSS 来构建可重构光网络。

虽然 WSS 切换速度快,但用于构建大规模可重构网络时也面临不少问题,比如扩展性问题。现在商用 WSS 端口数量较少,主要规格为 1×4、1×9、1×20 或 1×40,而数据中心和超算中心节点数量众多,比如大型数据中心通常有几万台服务器,因此如何构建大规模易扩展的可重构网络是一个难题。

本文提出了一种基于 WSS 的动态可重构光互连结构,通过合理的层次化部署,大幅提升了扩展性,适合应用于节点数量众多的数据中心系统域互连网络。

2.1 基于 WSS 的基本交换单元

本文首先提出一种基于 WSS 的交换模块即 WSS-based Switch,在此基础上构建大规模的可重构光网络。N 端口 WSS-based Switch 由 $N+1$ 个 $1\times N$ 端口 WSS、Coupler 和波分复用(wavelength division multiplexing,WDM)光纤构成。所有 WSS 的下行端口之间使用 WDM 光纤进行全互连,假设第 $i(1\leqslant i\leqslant N+1)$ 个 WSS 的第 a 个端口和第 $j(1\leqslant j\leqslant N+1)$ 个 WSS 的第 b 个端口通过 WDM 光纤连接,连接规则如式(1)所示。

$$\begin{cases} a = (j - i - 1 + N) \bmod N \\ b = (i - j - 1 + N) \bmod N \end{cases} \tag{1}$$

WSS 的上行端口连接外部 WDM 光纤,这里要求 WDM 光纤输入的波长数目和类型符合 WSS 的波长选择范围。由 4 个 WSS 组成的 WSS-based Switch 如图 3 所示。

N 端口 WSS-based Switch 中 WSS 之间使用光纤全互连,因而每个端口与其他端口都存在可达的物理链路,两个单元是否存在真正的光路,取决于两个单元所连的 WSS 间的光纤是否分配了相同的波长,且这个波长对于一个 WSS 而言最多只能分配一次,否则会造成波长冲突;N 个端口之间可根据需要配置成全互连或者其他连接形式,以满足不同的带宽需求。

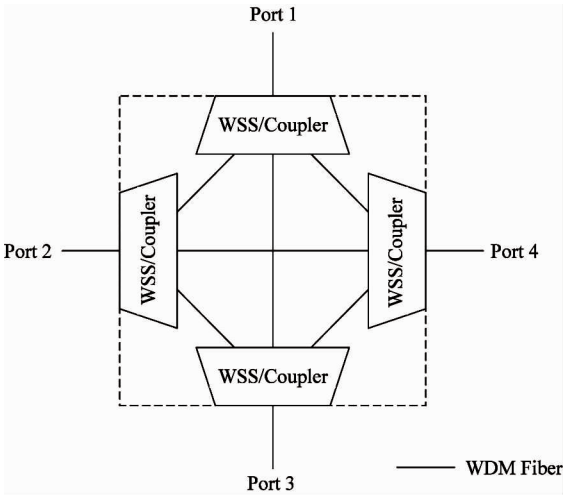


图 3 WSS-based Switch 基本结构

当给每个 WSS 的下行端口都分配了对应的 1 个波长,而且任意一根 WDM 光纤的两端分配的波长相同时,就可以实现交换单元 N 端口之间的无阻塞全互连。此时任意两个端口之间只有 1 个波长的带宽资源;如果某些端口之间需要更多的带宽资源时,可将多个波长分配到这些端口之间,但同时也需要避免波长冲突,具体分配方法将在下一节讨论。综上所述,WSS-based Switch 的内部结构支持全互连,利用每个端口上 WSS 波长选择分配使得端口间构成了无冲突的逻辑光路,将每个端口上的波长与连接到光交换机的波长进行对应。WSS-based Switch 具备极强的灵活性,包括动态的拓扑配置和带宽分配。通过在一条链路上分配多对波长,那么该链路的带宽就成倍增长,因而能够满足结构灵活性要求。

图 4 是 WSS-based Switch 物理结构和拓扑结构变换的例子。图 4(a)是物理连接示意图,由 4 个 WSS 构成;若交换机输入端包含 3 个波长 λ_0、λ_1、λ_2,按照图 4(b)所示的波长配置法则,该交换机内部能够实现 4 个端口之间的全互连结构;当波长较少的情况下,可以得到维度数较低的拓扑结构,譬如每个交换机输入端包含 2 个波长 λ_0、λ_1,按照图 4(c)所示的波长配置法则,4 个端口之间可以连接为

一维 Torus 互连结构；还可以根据负载需要灵活配置端口之间的连接带宽，如图 4（d）所示，不同的端口之间可以被分别配置为 2 个波长带宽或 1 个波长带宽。从该例可以看出，WSS-based Switch 提供了很强的拓扑重构能力，具体原理和配置方法在下一章有详细描述。

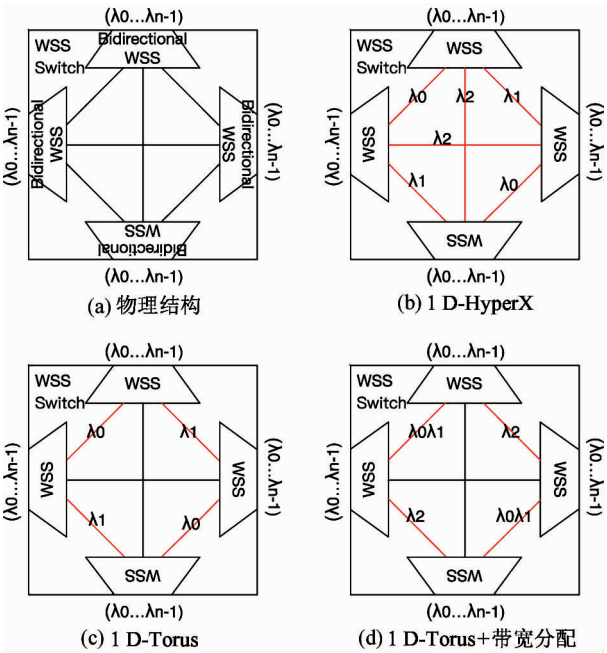


图 4　波长分配和拓扑结构变换关系

2.2　多层次可重构光网络

数据中心和超算中心中节点数量众多，而常见的 WSS 器件按照下行端口数量主要有 1×4、1×9、1×20 和 1×40 等，受到 WSS 端口数目和 WDM 波长数目的限制，单个WSS-based Switch构成的网络规模非常有限，需要开展面向大规模网络的层次化设计。

由于光缓存技术和光逻辑运算能力的制约，数据包的数据解析和缓存功能需要在电域完成，本文提出了一种利用了光路交换（optical circuit switching，OCS）技术和电分组交换（electrical packet switching，EPS）技术的层次化的动态可重构光电混合网络，网络分为多个层次，底层 L_0 层为电网络，其它 L_1，L_2，…，L_n 层为光域网络。其中电网络功能由 L_0 层的柜顶交换机 ToR 提供，参与报文的解析和转发；L_1、L_2 层光域网络由光复用/解复用器（Mux/Demux）和 WSS-based Switch 构成，参与光路的动态构建与光路路由。对于 ToR 交换机而言，光路的动态变换过程是透明的，ToR 交换机是通过网络协议解析下一跳步的交换机信息和端口信息；构建光路是通过结构的配置算法完成的，在所有 ToR 间构成了无冲突的光路，实际上是配置了 ToR 交换机间的连接关系，即拓扑连接关系。

图 5 详细描述了 3 层可重构网络结构，包括 p 个 Pod，每个 Pod 包含 m 个 Rack、$m+t$ 个 Mux/Demux 和 1 个 WSS-based Switch，每个 Rack 包含 n 个 Node、1 个 ToR 和 $n+t$ 个光收发器（TRX），不同 Pod 内同样位置的 TRX 的波长设置规则相同，令 $\lambda_{i,j}$ 表示第 i 个 Rack 的第 j 个 TRX 的波长，则按照式（2）进行分配。

$$\begin{cases} \lambda_{i,j} = \lambda_j & j \leqslant n \\ \lambda_{i,j} = \lambda_i & n < j \leqslant n \end{cases} \tag{2}$$

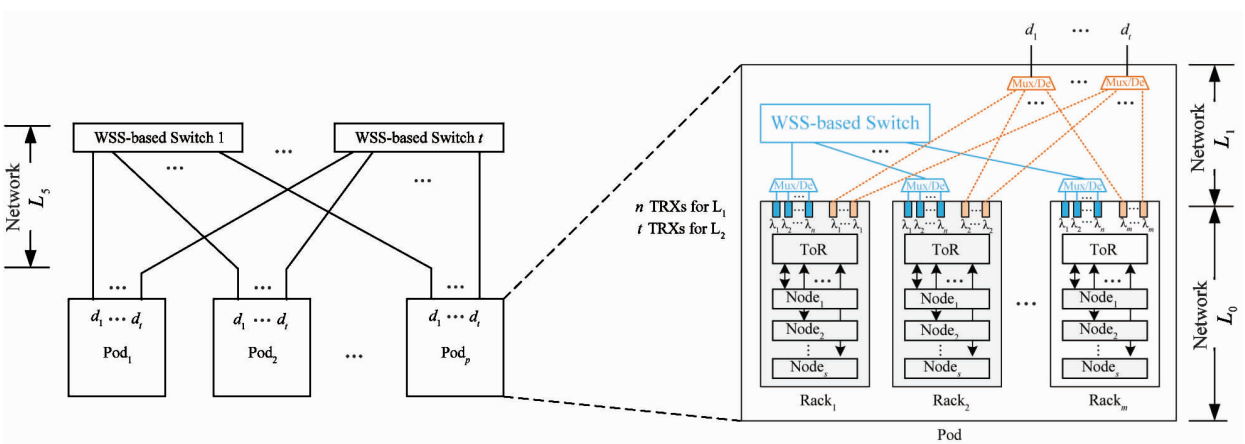


图 5　系统级可重构光电混合网络结构

L_0 层网络为基于 ToR 构成的电域网络,提供 Rack 内 Node 之间的互连;L_1 层网络基于 Mux/Demux 和 WSS-based Switch 构成,用于 Pod 内 Rack 之间的互连,每个 Rack 的前 n 个 TRXs 连接到同一个 Mux/Demux 进行波分复用解复用,上行信号连接到 WSS-based Switch;L_2 层网络基于 Mux/Demux 和 WSS-based Switch 构成,用于 Pod 之间的互连,每个 Rack 的后 t 个 TRXs 依次连接到 Pod 内的 t 个 Mux/Demux 进行波分复用解复用,上行输出信号依次记为 $d_1, d_2, \cdots$, 每个 Pod 的连接到对应的 WSS-based Switch_i 上。

上述 3 层结构支持 $p \times m \times s$ 个节点之间的可重构互连,共需要 $t+p$ 个 WSS-based Switch 和 $p \times (t+m)$ 个 Mux/Demux。若采用的 ToR 端口数目为 64,其中 48 端口连接到 Node 上,16 个上行端口用于链接到其它 Rack 和 Pod;采用 1×20 规格的 WSS 构成 21 端口的 WSS-based Switch,可支持 21 个 Pod,每个 Pod 含有 21 个 Rack,也即 $p=21$、$m=21$、$s=48$,整个网络共支持 20k 规模的节点。如果节点数量规模更大,可以增加网络层数,并根据局部和全局的不同通信需求分配用于 L_1 和 L_2 层网络的 TRX 数量。

为了增强灵活性,部分 TRX 可以替换为波长可调收发模块,如图 6 所示,通过分光器分别接到多个 Mux/Demux 上,这样就可以通过调谐波长来使得同一个 TRX 根据需要选择接入到不同层次的网络上,调整在不同层次的维度和带宽占比,增强网络灵活性。

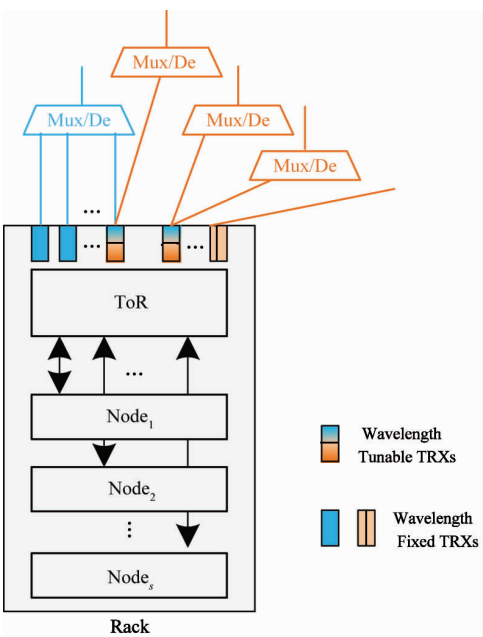


图 6　波长可调谐机制

2.3　光器件资源需求

以构建一个 10 k 计算节点规模的集群为例,设网络包含多个维度,1 个维度可连接 16 个 Pod,每个 Pod 内包含 16 个 Rack,每个 Rack 包含 64 个节点,Rack 内通过电域交换机连接,每个 Rack 有 16 个光端口用于 Pod 内 Rack 之间互连,1 个光端口用于连接到其他 Pod,那么所需要的资源如表 1 所示。

表 1　资源需求量

	规格	数量
电域交换机	64 端口	256
收发器	固定波长	4352
Mux/Demux	6 端口	256
Mux/Demux	16 端口	16
WSS	16 端口	272
Coupler	16 端口	272

3　拓扑重构配置算法

基于 WSS 的光电混合网络结构提供了可重构光互连的物理基础,但要将网络拓扑连接方式配置为匹配通信特征的目的拓扑,还需要完成对应于这个特定目的拓扑的所有 WSS 波长分配;由于波长资源宝贵且数量有限,因此最好的配置方式需要在保证连通度的同时尽可能利用每一个波长资源。本节将首先探讨如何根据通信特征需求选择合适的拓扑结构和参数,并将逻辑连接关系映射到不同的 WSS-based Switch 上;然后基于 Misra&Gries[17] 和 Greedy[18] 算法提出一种波长配置算法,可在多项式时间复杂度下自动生成所有波长交换模块的配置信息,完成对可重构网络的拓扑构建并且最优化剩余波长资源用于带宽调度。

3.1　拓扑重构策略

本文假定已知具体应用的通信特征,提取通信特征的具体方法不是本文讨论重点。首先来看下如何根据负载变换来进行网络全局拓扑重构,具体步

骤如下。

(1) 根据负载通信特征,统计每层或每维度网络对应的通信需求,通过拓扑和通信需求匹配来选择合适的网络拓扑结构,确定节点之间的逻辑连接关系;根据服务优先级或需求矩阵要求,构建节点之间带宽分配权重系数矩阵,空闲的光链路资源将被优先配置到权重系数高的节点之间。

(2) 以 WSS-based Switch 为基本单元,将整体拓扑结构和带宽参数进行分解和映射,获得每个 WSS-based Switch 对应的逻辑连接矩阵和带宽分配权重系数矩阵。

(3) 根据步骤(2)求解的 WSS-based Switch 的对应的逻辑连接矩阵求解出每个 WSS 的波长分配规则,形成无冲突的波长分配方案,本文采用基于 Misra&Gries 的算法,下文将进行详细描述。

(4) 根据步骤(2)求解的 WSS-based Switch 的对应的带宽分配权重系数矩阵来分配步骤(3)完成后剩余的波长资源,优先级越高的链路获得闲置波长资源的概率越大,本文采用基于 Greedy 的带宽分配算法,下文将进行详细描述。

(5) 根据步骤(3)和(4)中求解的波长分配结果,通过控制平面重置各个 WSS-based Switch 所有 WSS 的波长配置来完成拓扑重构。

3.2 波长配置算法

拓扑重构最为关键的步骤是如何将拓扑关系转换为具体的波长配置,本文创新地将 WSS-based Switch 的波长分配问题抽象为边着色问题,在保证 WSS-based Switch 构成拓扑具有连通性的基础上,进行剩余带宽的最优分配。

设 WSS-based Switch 端口数有 m 个,编号依次为 $0,\cdots,m-1$,输入波长数有 n 个,编号依次为 $0,\cdots,m-1$。构建多重图 $G(N, E)$,将所有端口表示为图中的节点 N,节点间的边表示为 E(可能存在多重边),所有的边代表了端口间的拓扑连通和带宽分配比例。将所有输入波长表示为颜色 C,现将这些颜色尽可能分配给图中每一条边,要求每个节点引出的边的颜色各不相同。根据实际情况,优先使节点间的单个边能够完成颜色分配,在此基础上再进行多重边的颜色分配。图 7 给出了基于 WSS-based Switch 构建的 3 × 3 的 2D-Torus 拓扑结构,其中顶点为光交换机的端口,边表示端口间的光路,边上的数字表示分配的波长编号,共使用了 5 种波长来完成拓扑构建,每对节点连接的边没有重复的编号,因而未出现波长冲突。

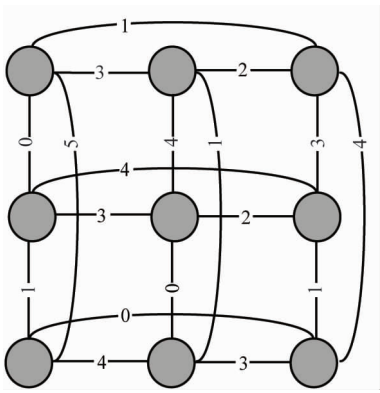


图 7 Torus 拓扑波长分配的例子

这个问题跟图论的多重图边着色问题很类似,但有几点特别之处需要引起注意:

- 边着色具有优先级,最好每次只对一条边进行着色,且不能使新的颜色分配出现冲突;
- 波长数目受到限制,即 K 着色问题,在进行多重边着色之前必须完成拓扑的着色,否则着色失败,即简单图必须满足 K 可着色;
- 参数 m 和 n 也即端口数和输入波长数涉及到物理结构的具体实现问题,对于波长的分配,合适的算法可以给出合适的参数,反作用于结构设计。

综上所述,波长配置算法重点解决 2 个问题,即拓扑构建和带宽分配。根据 Vizing 定理[19]可以证明仅需使用 Δ 或 $\Delta+1$ 种波长就肯定能够完成简单图的拓扑构建,如果放松了波长数约束条件,可确保在多项式时间内找到任意图的一种边着色方案;关于带宽分配,继续使用剩余波长进行多重图的边着色即可。

本文基于 Misra&Gries 算法和 Greedy 首次匹配算法提出一种波长配置算法,可以高效地解决波长配置问题。Misra&Gries[19]算法由 Misra 和 Gries 提出,证明了 Vizing 定理中可以使用 $\Delta+1$ 种对最大度数为 Δ 的任意简单图进行边着色。该证明显示了如何对合法图的任意无色边进行着色(可能需要

更改已着色边的颜色以保持合法性)。每次对一个边着色,通过重复此过程,直到所有边都着色完成。使用 $\Delta+1$ 种颜色进行边着色,这对于一些图是最优的,对于另外一些图效果较差,但最差也只需要多使用一种波长。对于边着色问题,该算法是已知最快的"几乎最优"算法。

Greedy 首次匹配算法直接判断尚未着色的两个顶点是否有相同的未使用颜色,并将该颜色分配给该边就完成了首次匹配。重复对每条需要着色的边进行着色,直到每条边都无法继续着色。Greedy 首次匹配算法使用最多 $2\Delta-1$ 种颜色进行着色,是最优的在线算法,该算法的解法仅与当前空余的颜色有关,而与之前的着色状态无关。

基于上述两种算法思想,本文提出表 2 所示的 WSS-based Switch 重构配置算法。其中拓扑配置算法使用的波长数最多比维度多 1,算法时间复杂度低(多项式时间)且波长利用率高;带宽调整算法不需关注网络历史分配状态,适应性好,不需要对图进行重着色,算法时间复杂度低。

表 2　WSS-based Switch 重构配置算法

WSS-based Switch 重构配置算法
条件:图 G 是有 N 个顶点和 $X+Y$ 条边的多重图,现有颜色集合 C,要求对 X 条边必须完成着色,之后 Y 条边根据优先级尽可能着色。
需求:拓扑构建是对 N 个顶点构成拓扑的 X 条边分配波长;带宽分配是对 N 个顶点的其他 Y 条边(可能有重边)按优先顺序分配波长。
算法步骤:每次只对一条边进行着色,先进行拓扑构建,再进行带宽分配,直到所有边都完成了着色过程。
(1) 以(u, v)表示简单图 G 的一条边,$F[0;k-1]$表示 u 的一个以 $F[0]=v$ 为首节点的最大扇,c 是 u 未使用的波长,d 是 $F[k-1]$上未使用的波长;
(2) 翻转 cd_u 路径并令 $F'=[F[0]\ldots w]$,其中 $w\in F$ 是一个扇,d 是 w 未使用的波长;
(3) 旋转扇 F'并令 $c(u,w)=d$;
(4) 以 (u,v) 表示任意图 G 的一条边,c 表示 u 和 v 都未使用的波长,求解函数,若存在 c 则 $c(u, v)=c$;否则放弃分配波长;
(5) 转到步骤(4)进行下一条边的波长分配直至所有波长分配完毕。

4　实验结果

本章首先对基于 WSS 的可重构网络的波长配置算法进行了展示,然后通过 1024 个节点的仿真平台评估了不同拓扑结构在多种典型流量特征下的网络延迟和平均跳步数。

4.1　拓扑重构和带宽分配

设一个 Pod 内有 9 个 Rack,由一个 WSS-based Switch 提供 Rack 间互连,如图 8 所示,每个 Rack 有 5 个不同波长光收发器($\lambda_0,\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5$)连接至 1 个 Mux/DeMux 下行端口,Mux/DeMux 上行端口连接至 WSS-based Switch。5 种波长可以分配给最大度数为 4 的图,能够配置为 2D-Mesh 和 2D-Torus 等维度为 4 的拓扑结构,如图 9(a)和(b)所示。

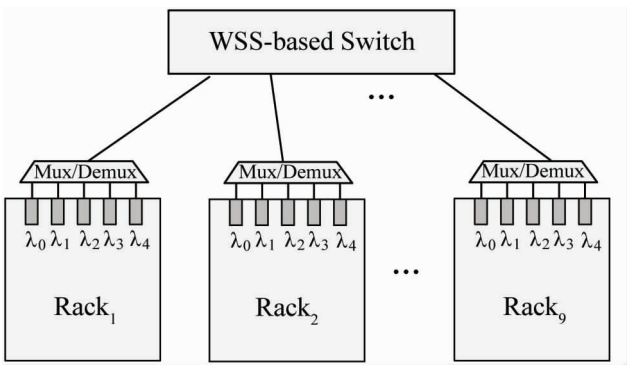


图 8　Torus 拓扑波长分配的例子

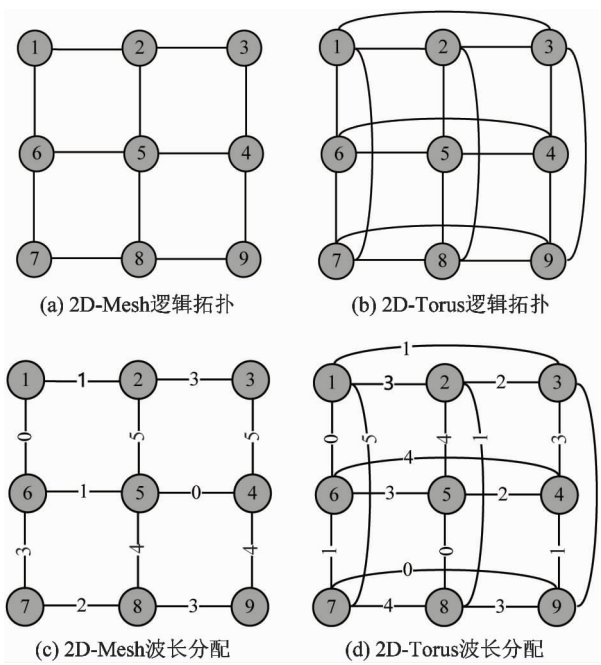


图 9　拓扑构建波长分配的例子

具体的波长配置细节可以通过运行第 2 节中的算法来进行求解，如图 9(c)和(d)所示，将 5 种波长分别表示为 0 ~ 4 数字分配到每条光路上，从而构建出 2D-Mesh 和 2D-Torus 拓扑，由图可见两种拓扑的构建过程都没有发生波长冲突，复合波长数目的要求。

网络中每个节点有 λ_0、λ_1、λ_2、λ_3、λ_4 共 5 个波长资源，分别将拓扑配置为 Mesh 和 Torus 后还有剩余波长资源。如果 9 个节点之间通信局部性很强，流量特征矩阵如图 10 所示，那么我们将把剩余的波长资源优先分配给相邻节点。

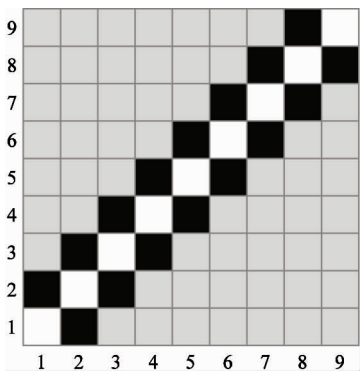


图 10　通信特征矩阵

剩余波长带宽分配时将按照节点 1 ~ 9 顺序依次进行分配，优先分配给相邻节点，比如 2 和 3 之间或者 5 和 6 之间的通信链路；由于同一个节点与其它节点进行连接的光波长不能相同，因此优先链路分配完毕后依然可能有剩余波长资源，此时将继续分配给其他节点。

如图 11 所示，以 2D-Mesh 的节点 1 为例，配置为 Mesh 时已经使用了 λ_0、λ_1，剩余 λ_2、λ_3、λ_4 可供分配，节点 1 和节点 2 通信需求大、优先级高，因此优先进行波长分配，考虑到节点 2 的 λ_1、λ_2、λ_3 均已

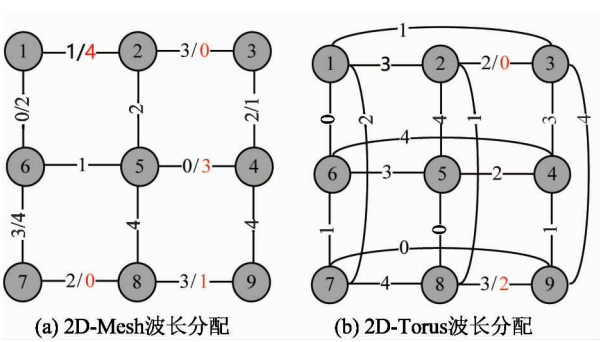


图 11　带宽资源分配

被使用，因此只有 λ_4 可以被分配到节点 1 和节点 2 之间；当节点 1 ~ 9 的优先链路(与相邻节点连接的链路)分配完毕后，节点 1 依然剩余 λ_2、λ_3 可供分配，考虑到节点 6 已经使用了 λ_0、λ_1、λ_3 因此在节点 1 和节点 6 之间分配 λ_2，这也就是前面章节提到的带宽分配算法，所有迭代完成后得到图 11 所示带宽分配图。

4.2　网络性能仿真

本文基于 cHPPNetSim 模拟器[20]对拓扑重构效果进行评估，该模拟器支持基本网络部件的细粒度模拟，包括网络接口控制器、交换机等，可以根据需求自定义输出仲裁算法和流控机制；支持对部件进行参数化配置，例如缓冲区大小、位宽、频率；支持数据包格式的灵活配置，可以添加各类管理信息，如路由信息、优先级信息等。本文的模拟平台配置参数如表 3所示。

表 3　仿真环境配置

类型	属性
CPU	2 个 Xeon E5-2658A
Memory	128 GB
OS	Linux 3.13.0-39
Compiler	GCC 4.8.4\MPI 3.0.4
Library	PTHREAD

模拟平台采用以下网络延迟定义：

- 消息传递的软件开销为 140 个时钟周期；
- 网络接口控制器的处理延迟设计为 12 个时钟周期，包括发送延迟、接收延迟和仲裁延迟；
- 交换延迟设计为 15 个周期，包括接收延迟、路由开销、仲裁开销和传输延迟；
- 链路延迟每跳定义为 100 ns；
- 链路带宽为 8 Gbps。

为了全面评估拓扑重构的收益，本文分别在 3 种通信特征下对不同拓扑结构的延迟性能和跳步数进行了仿真，通信特征定义如下：

- Random：所有节点之间通信流量随机分布；
- Local：将网络分为若干个区域，其中 $p\%$ 流量为区域内部的随机流量，$(1-p\%)$ 流量为区域之间的随机流量；

• Global:将网络分为若干个区域,其中 $p\%$ 流量为区域之间的随机流量,$(1-p\%)$ 流量为区域内部的随机流量。

本文分别对 1 024 个节点的网络和 4 096 个节点的网络进行了仿真。对于 1 024 个节点的网络规模,将整个模拟平台分为 16 个区域,每个区域包含 64 个节点;对于 4 096 个节点的网络规模,将整个模拟平台分为 64 个区域,每个区域包含 64 个节点,这两种情况下设定 p 都为 87.5。

(1) 1 024 节点仿真

拓扑维度数直接影响拓扑性能,为了精确评估拓扑重构效果,规定拓扑重构过程中均使用相同的端口数,保证不同拓扑的总维度数相同,此处定义每个节点有 12 个端口,3 种拓扑结构分别为:

• 拓扑 1:1 024 个节点之间的互连结构分成两层,内层和外层对应拓扑分别为 HyperX(4, 4,2,2)和 Torus(4,4);

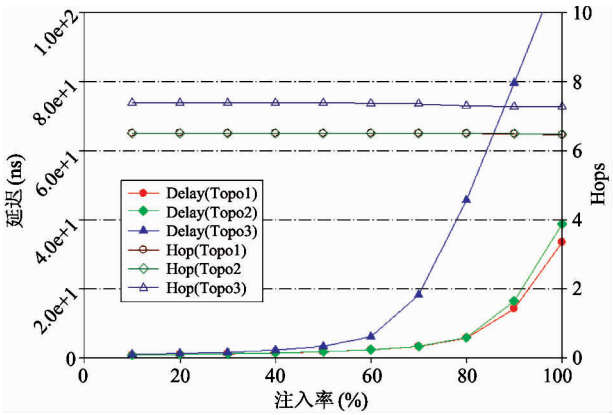
• 拓扑 2:1 024 个节点之间的互连结构分成两层,内层和外层对应拓扑分别为 HyperX (4, 4) 和 Torus (4, 4, 4);

• 拓扑 3:1 024 个节点之间的互连结构分成两层,内层和外层对应拓扑分别为 HyperX (8, 2) 和 Torus (8, 8)。

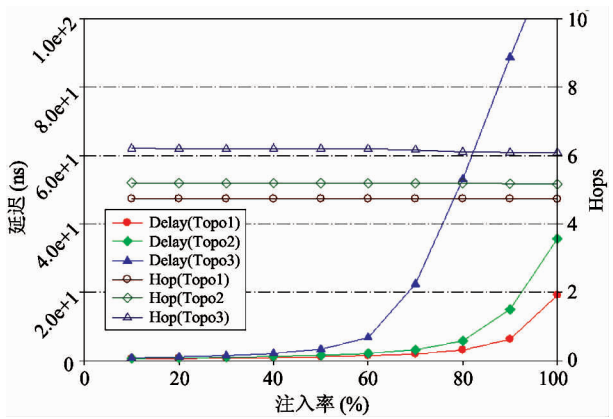
其中 HyperX$(a_1, a_2, \cdots, a_n)$ 表示拓扑种类为 HyperX,包含 n 个 Dimension,对应的节点数分别为 $a_1, a_2, \cdots, a_n$;Torus$(a_1, a_2, \cdots, a_n)$ 表示拓扑为 n 维 Torus 结构,对应的节点数分别为 $a_1, a_2, \cdots, a_n$。

图 12(a)、(b)和(c)分别为1 024个节点在 Random 通信特征、Local 通信特征和 Global 通信特征下的延迟性能(Delay)和跳步数(Hop)。

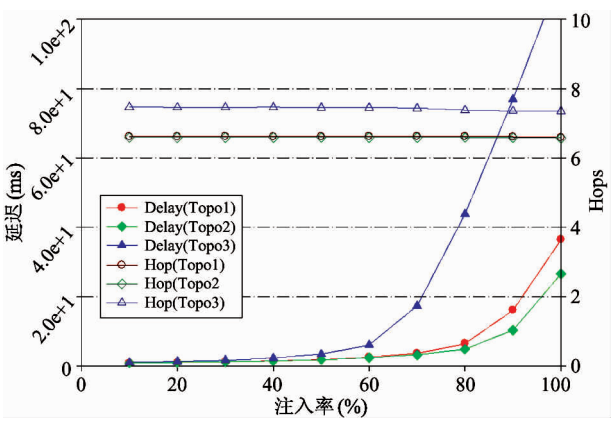
在 3 种流量模式下,拓扑 3 的延迟性能和平均跳数均为最差,这是由于外层拓扑 Torus(8, 8)维度低、性能差,因此整体性能受到较大影响;拓扑 1 的内层网络维度数比拓扑 2 内层网格维度更高,因此在局部流量占比更高的 Local 通信特征下性能更好;拓扑 2 的外层网络比拓扑 1 的外层网络维度更高,因此在全局流量占比更高的 Global 通信特征下性能更好。由上可见,不同拓扑结构适合不同的通信特征,通过拓扑重构能够提升网络性能,在高注入率情况下拓扑重构为最佳拓扑时,性能提高可达 60%。



(a) 1 024 节点 Random 通信特征



(b) 1 024 节点 Local 通信特征



(c) 1 024 节点 Global 通信特征

图 12　拓扑重构效果评测

(2) 4 096节点仿真

在 4 096 节点下,为保证不同拓扑的总维度数相同,此处定义每个节点有 18 个端口,3 种拓扑结构分别为:

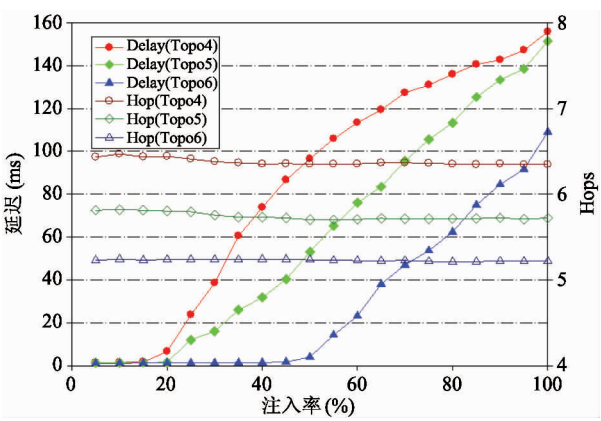
• 拓扑 4：4 096 个节点之间的互连结构分成两层。内层、外层对应拓扑分别为 HyperX(8,8)和 Torus (8,8)；

• 拓扑 5：4 096 个节点之间的互连结构分成两层，内层和外层对应拓扑分别为 HyperX(2, 4, 2, 4)和 Torus(2,2, 2,8)；

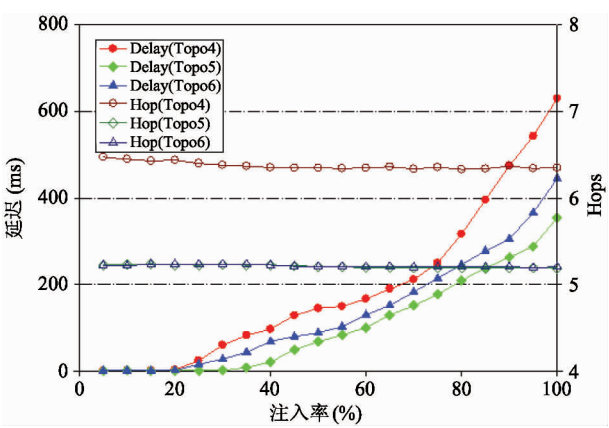
• 拓扑 6：4 096 个节点之间的互连结构分成两层，内层和外层对应拓扑分别为 HyperX(8, 2, 2, 2)和 Torus (2, 4, 2, 4)。

图 13(a)、(b)和(c)分别为 4 096 个节点在 Random 通信特征、Local 通信特征和 Global 通信特征下的延迟性能(Delay)和跳步数(Hop)。

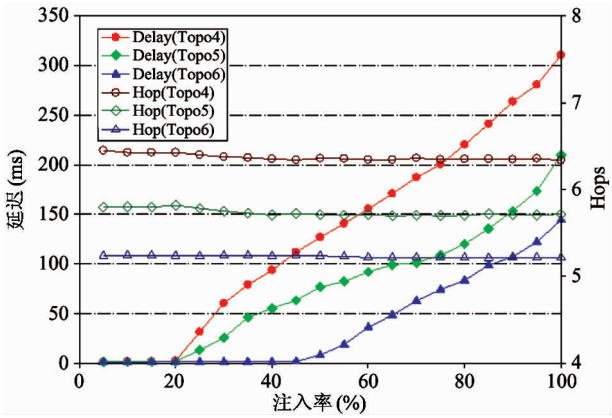
在 3 种流量模式中，拓扑 4 延迟性能和平均跳数均为最差，跟 1 024 个节点类似，这是由于该拓扑内外层维度数分布不均匀，全局拓扑 Torus(8, 8)性能差，影响了整体性能；拓扑 5 和拓扑 6 虽然内外层拓扑类型和维度数相同，但拓扑参数差异较大，因此适合于不同的通信特征，在Random流量和全局流量占比更高的 Global 通信特征下，拓扑 6 性能最优，延迟和跳步数都最小；在局部流量占比更高的 Local 通信特征下，拓扑 5 性能最优，延迟和跳步数都最小。



(a) 4 096 节点 Random 通信特征



(b) 4 096 节点 Local 通信特征



(c) 4 096 节点 Global 通信特征

图 13　拓扑重构效果评测

5　结　论

本文基于 WSS 器件提出了一种新型的可重构光网络结构，在不改变物理布局布线的情况下通过波长控制实现网络拓扑的动态构建和带宽的灵活分配；基于 Misra&Gries 和 Greedy 算法原理提出了一种重构配置算法，在多项式时间复杂度下可自动生成所有波长交换模块的配置信息。在 1 024 个节点和 4 096 个节点下的评测结果显示，不同拓扑在不同通信特征下延迟和跳步数差异明显，通过重构拓扑类型和参数能够显著改善网络性能，在注入率比较高时网络性能提升可达 60%。

参考文献

[1] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild[C]. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, Melbourne, Australia, 2010. 267-280

[2] Benson T, Anand A, Akella A, et al. Understanding data center traffic characteristics[C]. In: Proceedings of the ACM Workshop on Research on Enterprise Networking, Barcelona, Spain, 2009. 65-72

[3] Antoniades N, Ellinas G, Homa J, et al. ROADM architectures and WSS implementation technologies[J]. *Convergence of Mobile and Stationary Next-Generation Networks*, 2010: 643-674. doi: 10. 1002/9780470630976.

ch20
[4] Lea C T. A scalable AWGR-based optical switch[J]. *Journal of Lightwave Technology*, 2015, 33(22): 4612-4621
[5] Ma X, Kuo G S. Optical switching technology comparison: optical MEMS vs. other technologies[J]. *IEEE Communications Magazine*, 2003, 41(11): 16-23
[6] Azizi S, Safaei F, Hashemi N. On the topological properties of HyperX[J]. *The Journal of Supercomputing*, 2013, 66(1): 572-593
[7] Liu Y H, Zhu M F, Wang J, et al. Xtorus: an extended Torus topology for on-chip massive data communication[C]. In: Proceedings of the 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), Shanghai, China, 2012. 2061-2068
[8] Bergman K, Wang H. Optical Interconnects for Future Data Center Networks[M]. New York: Springer, 2013. 155-167
[9] Lin Y, Anthur A P, O'Duill S, et al. Fast reconfigurable SOA-based all-optical wavelength conversion of QPSK data employing switching tunable pump lasers[C]. In: Optical Fiber Communications Conference and Exhibition, Los Angeles, USA, 2017. 1-3
[10] Ye X, Yin Y, Yoo S J B, et al. DOS: a scalable optical switch for datacenters[C]. In: Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems, La Jolla, USA, 2010. 24
[11] Ju H L, Yusoff Z, Belardi W, et al. A tunable WDM wavelength converter based on cross-phase modulation effects in normal dispersion holey fiber[J]. *Photonics Technology Letters*, 2003, 15(3): 437-439
[12] Kang X, Kao Y H, Yang M, et al. Petabit optical switch for data center networks[J]. *Dca. fee. unicamp. br*, 2010. doi:10.1007/978-1-4614-4630-9-8
[13] Liboiron-Ladouceur O, Shacham A, Small B A, et al. The data vortex optical packet switched interconnection network[J]. *Journal of Lightwave Technology*, 2008, 26(13): 1777-1789
[14] Sato K I. Realization and application of large-scale fast optical circuit switch for data center networking[J]. *Journal of Lightwave Technology*, 2018, 36(7): 1411-1419
[15] Wang G, Andersen D G, Kaminsky M, et al. c-Through: part-time optics in data centers[J]. *ACM SIGCOMM Computer Communication Review*, 2010, 40(4): 327-338
[16] Farrington N, Porter G, Radhakrishnan S, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers[J]. *ACM SIGCOMM Computer Communication Review*, 2010, 40(4): 339-350
[17] Vizing V G. On an estimate of the chromatic class of a p-graph[J]. *Diskret Analiz*, 1964, 3: 25-30
[18] Bar-Noy A, Motwani R, Naor J. The greedy algorithm is optimal for on-line edge coloring[J]. *Information Processing Letters*, 1992, 44(5): 251-253
[19] Misra J, Gries D. A constructive proof of Vizing's theorem[J]. *Information Processing Letters*, 1992, 41(3): 131-133
[20] Cao Z, Xu J, Chen M, et al. HPPNetSim: a parallel simulation of large-scale interconnection networks. In: Proceedings of the 2009 Spring Simulation Multiconference, San Diego, USA, 2009. 32

WSS-based reconfigurable optical network

Yuan Guojun[*,**], Xiao Peng[**], Jiang Tao[*], Wang Zhan[*], Yang Fan[**], Cao Zheng[*], Zhang Peiheng[*], Tan Guangming[*], Sun Ninghui[*]
([*] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)
([**] University of Chinese Academy of Sciences, Beijing 100049)

Abstract

Data Centers run many kinds of applications that exhibit various communication patterns among the servers. For a better performance it would be admirable to choose the appropriate topology under different communication patterns. However data centers usually use a single architecture for various applications. In this paper a WSS-based reconfigurable optical network is proposed which dynamically matches the physical topology to various traffic patterns. It significantly improves the link utilization, power efficiency and flexibility of the interconnection network. Besides, a reconfiguration method combining Misra & Griesedge coloring algorithm and Greedy algorithm is recommended. It can calculate the proper wavelength parameters for each wavelength selective switch (WSS) during the reconfiguration process. Simulation results show that proper reconfiguration can be realized by tuning the control plane of WSS and improve the performance (1 024 nodes) beyond 60% under three typical communication patterns.

Key words: data center, hybrid optical-electrical network, reconfigurable network, wavelength selective switch (WSS)