

引用格式:王昭,裴秉玺,何卫锋. 基于 AWGR 的光电混合交换高性能计算互连网络 [J]. 微电子学与计算机, 2019, 36(8): 19-24. [WANG Z, PEI B X, HE W F. An optical-electrical hybrid switching high performance computing interconnect network based on AWGR [J]. Microeletronics & Computer, 2019, 36(8): 19-24.]

基于 AWGR 的光电混合交换高性能计算互连网络

王 昭,裴秉玺,何卫锋

(上海交通大学 微纳电子学系,上海 200240)

摘 要: 为了满足高性能计算机中互连网络的高吞吐量、低延迟需求,本文基于阵列波导光栅路由器(AWGR),提出了一种低延迟的光电混合交换的互连网络结构. 该网络以“天河二号”超级计算机的互连网络拓扑结构为基础,使用无竞争阻塞的 AWGR 替代了部分电路路由器,通过改进设计电路路由器的光端口来实现与原有高阶路由器内部的电交换方式兼容. 在此基础上,使用 OMNeT++ 软件框架编写了新的光电混合交换网络的仿真模型. 仿真结果表明,与原始网络相比,光电混合交换网络的总延迟降低了 8%~10%.

关键词: 高性能互连网络;阵列波导光栅路由器;光分组交换;网络延迟

中图分类号: TP391.9

文献标识码: A

文章编号: 1000-7180(2019)08-0019-06

An optical-electrical hybrid switching high performance computing interconnect network based on AWGR

WANG Zhao, PEI Bing-xi, HE Wei-feng

(Department of Micro-Nano Electronics, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: To satisfy the throughput and latency requirement of high-performance-computing interconnect network, a low-latency optical-electrical hybrid switching network structure based on arrayed waveguide grating router (AWGR) is proposed. On the basis of “Tianhe-II” supercomputer’s network topology, part of electrical high-radix routers are replaced by non-blocking and contention-free AWGRs. The optical port of electrical router is re-designed to interface the optical packets with router’s internal switching logics. To evaluate the performance of proposed network, a simulation model is developed using OMNeT++ software framework. The simulation shows that the data latency of hybrid switching network is 8%~10% lower than the electrical switching network.

Key words: high-performance interconnect network; arrayed waveguide grating router; optical packet switching; network latency

1 引言

近年来,高性能计算机的性能遵循千倍定律,即每 10 年计算性能提升 1000 倍,根据预测在 2019 年达到 10^{18} Flops 每秒(Exa Flops,即 E 级). 高吞吐量、低延迟的互连系统是决定高性能计算机性能的关键硬件. 高性能计算互连系统普遍采用“电交换、光传输”的通信方式. 以我国自主研发的“天河二

号”超级计算机的互连系统为例^[1],它采用了三级交换机的胖树(fat-tree)拓扑,高阶路由芯片是每级交换机的组成单元,片内使用分布式缓存实现多级仲裁以缓解端口竞争,但也延长了硬件流水线,测试表明路由器的穿越延迟占据了总延迟的 30~50%^[1]. 此外“电交换、光传输”不可避免在交换机端口进行光电转换,引入了功耗、可靠性问题和带宽瓶颈. 具有直接光交换功能的开关器件在高性能计算的互连

系统中具有潜在的应用价值. 阵列波导光栅路由器 (Arrayed Waveguide Grating Router, AWGR) 是一种适用于光分组交换 (optical packet switching) 的新型无源光交换器件, 近年来被研究用于数据中心交换机的光互连^{[2][3]}. AWGR 器件具有循环波长路由特性, 控制光的波长可以将光路由到特定端口. AWGR 的器件工艺正在走向成熟, 目前业界的 Enablence 公司已有至多 32 通道的路由模块产品, 学术界报道了 512×512 开关^[4]和硅基工艺器件. 应用 AWGR 的波长路由特性, 需要解决产生和接收特定波长光信号的实际问题. 文献[2]在 AWGR 输入处安装可调波长转换器 (tunable wavelength converter), 由控制模块调节. 由于大规模互连网络需要上万转换器件, 总成本高. 文献[3]提出了适合于波长路由的定制化的光网络拓扑, 传输数据报文会经过若干个“接力”的计算结点, 每次“接力”都可以调整波长. 由于主流高性能计算互连网络普遍选取“胖树 (fat-tree)”拓扑, 全光互连网络的实现难度大, 定制化拓扑与已有互连系统的兼容性差.

本文提出了一种可适用于目前高性能计算互连系统的光电混合交换网络方案.

2 “天河二号”胖树网络拓扑

图 1 展示了“天河二号”互连系统的网络拓扑. 网络由三级交换机 (root、NRC 和 NRM) 组成, 可连接至多 18 432 个计算节点 (Node). 第一级交换机是 NRM (Network Routing Module), 它的 32 个电端口向下连接了 32 个计算节点. NRM 有 36 个光端口 (图中黑色小方块) 向上连接到第二级交换机. 第二级交换机是 NRC, 下侧和上侧各有 12 个光端口. 下侧通过 10 m 光纤连接到 NRM, 上侧通过 20 m 光纤连接到第三级交换机. 36 个 NRC 连接的 432 个结点组成一个计算集群 (cluster). 第三级是根交换机 (root switch), 一个根交换机有 48 个光端口, 分别通往不同的集群, 使整个网络连通. 在天河“胖树”网络中, 一条水平割线在任何位置穿过的链路数相等, 网络等分带宽正比于计算结点数量, 没有发生超额订阅 (oversubscription), 理论上可以提供最大 100% 的数据吞吐率.

24 端口的网络路由芯片 (Network Routing Chip, NRC) 是三级交换机的基本单元, NRM 和根交换机都由两层多个 NRC 拼接组成. 取决于通信结点之间的“距离”, 网络报文可能会穿越 1/3/5/7/9 个 NRC. 报文首先向上转发, 到达通信结点的某

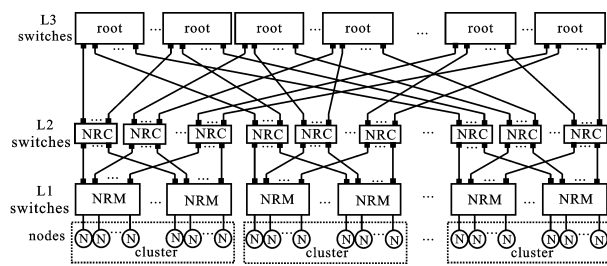


图 1 “天河二号”计算机互连网络的“胖树”拓扑

一“公共祖先”交换机, 然后向下转发, 到达目的结点. 任何一个“祖先”交换机都可以发现一条通向目的结点的路径, 因此报文向上转发到“祖先”的路径是任意的, 但向下通向目的结点的路径一般是唯一的. “胖树”丰富的链路资源提供了大量冗余链路, 流量自适应路由方法可以在大量数据传输时, 平衡链路负载. 另一方面, 当网络报文需要穿越第二级交换机时, 必须在两次光传输中间进行电交换, 发生光电转换.

3 电交换的路由器模型

3.1 NRC 的瓦片阵列高阶路由器结构

NRC 是网络三级交换机的基本单元, 由 24 个双向端口和一个片上网络 (Network on Chip, NoC) 组成. 路由器设计中, 增加端口数量可提升总吞吐量, 但因交叉开关、控制逻辑的复杂度与端口数成平方关系, 物理设计难度增大. 因此 NRC 采用了瓦片阵列结构^[5], 如图 2 所示. 缓冲区 (buffer) 和交叉开关被分布在 4×6 个完全相同的瓦片中, 每个瓦片包含一个 6 输入 4 输出的子开关 (sub-switch)、端口的输入缓冲区 (input buffer)、输出缓冲区 (output buffer), 子开关的交点处也设置了缓冲区 (cross-point buffer). 每个瓦片关联一个端口, 包含接收机 (receiver, RX) 和发射机 (transmitter, TX). 瓦片之间用行总线和列总线相连.

3.2 NRC 内数据分组的交换过程

网络中一个链路层分组 (packet) 由 5 个微片 (flit) 组成. 头微片包含源结点编号、目的结点编号的路由信息, 每个微片有表明微片位置的标志 (Head/Body/Tail) 和虚通道编号 (virtual channel, VC), 属于同一数据分组的微片的 VC 相同. 数据分组在网络缓冲区之间的传输采用了“虚切通 (virtual cut-through)”机制. 如图 2, 输入缓冲区、开关交点缓冲区、输出缓冲区都有 6 个 FIFO 队列, 用于 VC 为 1~6 的数据分组的缓存. 另一方面, 在 NRC 内部、多个 NRC 之间, 连接前后两级的缓冲区的只

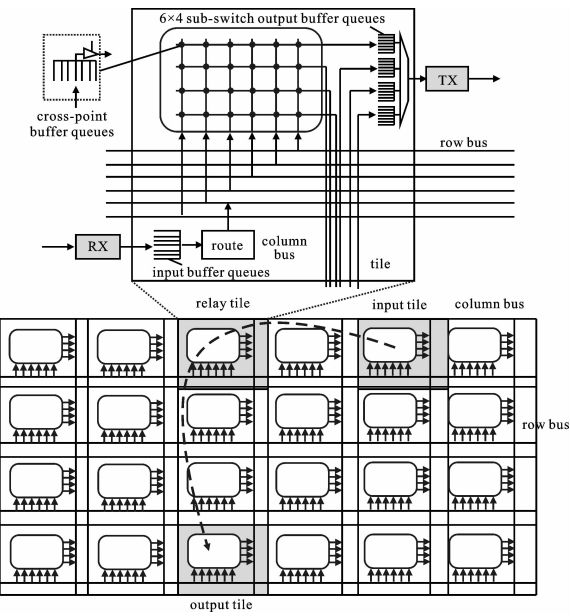


图 2 高阶路由器的瓦片阵列和瓦片内部结构

有一条物理信道。“虚切通”机制允许将单物理信道同时分配给几个不同虚通道上的数据分组. 获得分配的数据分组,其微片可以轮询传输. 这样,数据传输实际以微片为单位,多个分组分时复用单物理信道,在各自的虚拟信道上传输.

如图 2 所示,数据分组从“输入瓦片 (input tile)”的接收机进入 NRC,经行总线到达“接力瓦片 (relay tile)”、穿过子开关,再经过列总线到达“输出瓦片 (output tile)”,从发射机离开 NRC. 数据分组的传输是多级流水线过程:

(1)WB(Write Buffer):“输入瓦片”收到分组后,按照 VC 存放在相应输入缓冲区队列中;

(2)RC(Route Compute):“输入瓦片”的路由逻辑检查队首数据分组的路由信息、计算输出端口,行总线将数据分组分发到“接力瓦片”的开关缓冲区,存储在 VC 相应队列. “接力瓦片”和“输入瓦片”在同一行,和“输出瓦片”在同一列;

(3)VA-1(Virtual Channel Allocation 1):在 6 输入、4 输出的子开关中,数据分组和来自其它 5 个输入口的数据竞争同一个开关输出,将输出口的 6 个虚通道分配以轮询方式分配给 6 个数据分组;

(4)SA(Witch Arbitration):将开关输出权轮询地授予 6 个获胜的数据分组;

(5)ST(Switch Traversal):获胜的 6 个数据分组以微片为单位,穿过子开关,被列总线传送到“输出瓦片”的输出缓冲区;

(6)VA-2(Virtual Channel Allocation 2):由于

一列其它 3 个“接力瓦片”的数据分组也被发送到输出缓冲区,这些分组竞争输出 TX 的 6 个虚通道;

(7)PA(Port Arbitration):获得虚通道的 6 个数据分组,以微片为单位轮询地被发射机转发,离开 NRC.

采用分布式缓存、“虚切通”传输机制,提高了存储资源、链路资源的利用率. 当某一分组的数据由于自身原因无法传输时,其它分组的数据不会因此被阻塞(head-of-line blocking),因此也提升了数据吞吐量. 然而硬件流水线也显著延长. 测试表明“天河二号”单路由器的穿越延迟达到了 100 ns^[1].

4 基于 AWGR 的光电混合交换网络

4.1 AWGR 光交换器件

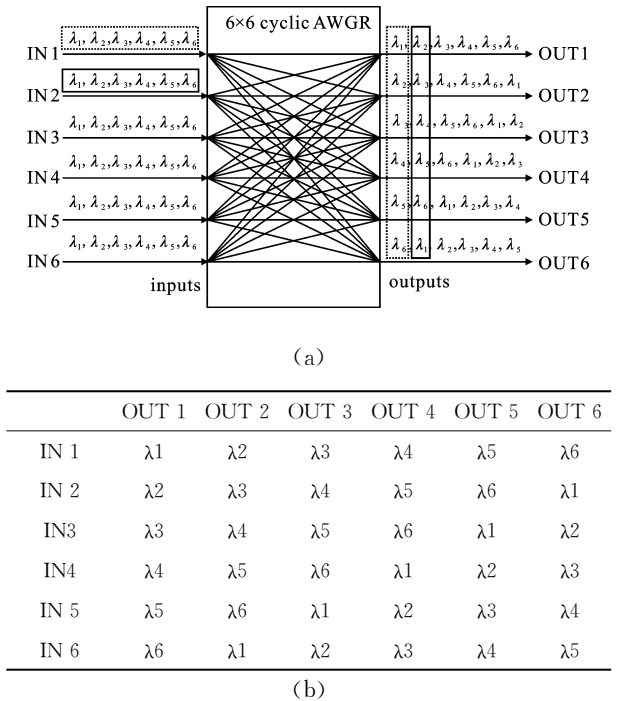


图 3 6×6 的 AWGR 波长路由示意

图 3 以一个 6 输入、6 输出的 AWGR 为例,展示了波长路由器的原理. AWGR 的路由波长数等于端口数. 如图 3(a),当来自输入 1 的光分别位于不同的波长上,会被路由到不同的端口;多个输入的路由方式是循环的. AWGR 内有弯曲的、长度不等阵列波导,引入了光相位差;不同波长的光在波导末端发生干涉时,增强和抵消的位置也不同,将输出端口设计在某一波长的光发生增强干涉的位置,即可实现波长的路由. 如输入 1 的 λ_5 光会被路由到输出 5,但输入 2 的 λ_5 光会被路由到输出 4. 波长与端口

路由的关系由图 3(b) 的路由表定义. 反过来看, 当来自输入 1 的光报文希望从输出 5 离开路由器, 则它必须被预先调制到 λ_5 波长上. 通过控制或调整光的波长, 可以实现 AWGR 的路由. 此外, 循环波长路由有一个重要性质: 来自不同输入、去往同一输出的光波长必不相同, 可以同时输出, 这样就消除了电路路由器的端口竞争, 光报文的穿越延迟仅取决于波导的长度. AWGR 独特的波长路由方式, 实现了光报文的无阻塞直接交换.

4.2 光电混合交换网络结构

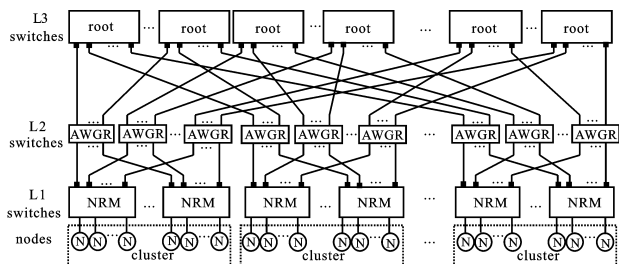


图 4 基于 AWGR 的互连网络拓扑

基于 AWGR 的光电混合交换网络结构如图 4 所示. 和图 1 相比, 光电混合交换网络的拓扑相同, 第一级、第三级交换机的基本结构不变, 但第二级交换机全部用 24 通道的 AWGR 替代, 同时对第一级、第三级交换机的光端口进行改进设计, 实现光分组的波分复用交换与电微片的片内交换的桥梁.

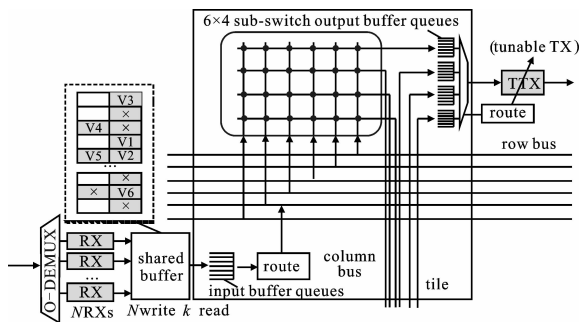


图 5 高阶路由器的光端口结构改进

4.3 适用于波长路由的电路路由器光端口设计

在光电混合交换网络中, 为了实现波长路由的光交换方式, 需要对第一级、第三级交换机 NRC 的光端口进行改进设计. 图 5 展示了改进光端口的结构. 在输入端口处, 1:N 的光解复用器 (O-DEMUX) 将 AWGR 输出的最多 N 路并行光报文进行分波, 每一路报文分别由 RX 接收. 然后, 报文以分组为单位, 存储在一个共享缓冲区 (shared buffer), 缓冲区具有 N 写 k 读的功能, k 不高于虚通道的数量. 从共享缓冲区读出的数据微片, 写入

到“输入瓦片”的输入缓冲区内. 在输出端口, 增加了一个路由计算模块, 能对输出缓冲区准备发送的数据分组, 预先执行在 AWGR 内的路由计算, 确定目标波长. 可调波长发射机 (tunable TX) 具有一个快速可调谐激光器 (Tunable Laser, TL) 结构, 可以根据路由计算的波长结果, 将数据分组转换成具有目标波长的光信号发射出去, 其波长切换时间在纳秒量级^[6]. 由于只对光端口进行了改进设计, 仍然保留了在 NRC 瓦片内的多级流水线逻辑, 端口仲裁和片上网络内“虚切通”的传输方式没有变化.

从光端口进入的数据分组, 在写入到输入缓冲区 (WB 阶段) 之前, 增加了两个阶段:

(1) WB-0 (Write Buffer 0): 一个 1:N 的光 DEMUX 分离出不同波长上最多 N 路数据分组, 同时写入到共享缓冲区中;

(2) VA-0 (Virtual Channel Allocation 0): 将 6 个虚通道分配给共享缓冲区中的至多 6 个数据分组 (如图 5 中标注 V1~V6 的数据分组), 其他分组在竞争中失败 (图 5 中标注 X).

同时, WB-0 阶段不再直接存储 RX 接收的数据微片. 在共享缓冲区中获得虚通道的数据分组, 其微片写入到输入缓冲区的对应队列中. 此后, 数据分组进入 NRC 片上网络的多级流水线中.

要从光端口输出的数据分组, 在进入到“输出瓦片”的输出缓冲区后, 将以数据分组为单位转发, 不再执行 VA-2 的过程. 因为快速可调谐激光器在波长切换时无法发送数据报文, 以分组为单位转发避免了频繁地波长切换导致的吞吐量损失.

(1) RC-2 (Route Compute 2): 在“输出瓦片”的输出缓冲区中, 路由模块检查缓存队首数据分组的路由信息, 预先计算出数据分组在 AWGR 内的输出端口. 由于“输出瓦片”映射的 NRC 端口固定连接到某一 AWGR 输入端口 p_i , 根据 p_i 和 p_o 查询如图 3(b) 的波长路由表, 即可得到实现正确路由所需的波长 λ . 路由计算模块寄存了所有可发送数据分组的请求波长值;

(2) PA (Port Allocation): 快速可调谐激光器的波长切换时间与两个波长距离成正比相关^[6]. 在多个可发送的数据分组中, 采用“懒惰”分配机制, 优先将端口分配给请求波长最接近当前调谐波长的数据分组, 可最小化波长距离.

在光端口的输入和输出侧增加的硬件结构, 具有“光适配器”功能. 在适配器的光信号侧, 可以直

接接收来自 AWGR 的多波长光数据分组,也可以直接向 AWGR 发送已调谐波长的光数据分组;在适配器的电信号侧,数据分组的传输方式与原 NRC 的瓦片结构相同。

5 网络流量控制

高性能计算互连网络应当能够处理突发的大量数据传输的流量场景,并且尽可能避免数据分组的丢失^[7]。与电信网络依靠传输层的 TCP 协议实现可靠数据传输和流量控制不同,高性能计算互连网络需要实现链路层的流量控制。目前学术界尚未提出可靠的光信号随机存储器(RAM)方案,因此光电混合交换网络的流量控制协议需要确保电交换路由器的缓冲区不发生溢出。光电混合交换网络采用了基于“信用(credit)”的流量控制和基于“通断(on/off)”的流量控制两种方法,可以确保当上游路由器向下路路由器的输入缓冲区发送数据之前,已经获知下游缓冲区的占用情况。

5.1 “信用”流量控制方法

在第一级交换机内部、第三级交换机内部的两层 NRC 之间,采用了细粒度的基于“信用”的流量控制方法。“信用”是指下游缓冲区的剩余容量,以数据微片为单位。在电交换机内部,由于某一个 NRC 电端口(称为上游电端口)唯一地向另一 NRC 电端口(称为下游电端口)发送数据微片,上游电端口维护了下游电端口缓冲区中每个缓存队列的剩余容量(即信用)。当上游电端口在虚通道 v 上发送了一个微片,上游电端口自觉地将维护的 v 队列的信用减少 1;当下游电端口中有一个微片离开了 v 缓冲队列,则下游电端口主动向上游电端口发送一个信用消息 credit-message(v)。上游电端口收到后,将维护的 v 队列的信用增加 1。当某一虚通道信用为 0 时,上游电端口不能在该虚通道上发送数据分组。信用消息只是局部地在直接相连的两个 NRC 电端口之间交换。

5.2 “通断”流量控制方法

在第一级和第三级交换机之间,采用了粗粒度的“通断”流量控制。此时上游光端口可以经过 AWGR 向任一下游光端口发送数据分组,上下游端口不再一一对应。当下游光端口的共享缓冲区占用率高于某一阈值时,下游光端口在每个波长上产生一个 traffic-message(OFF, k)并上溯发送,表示 AWGR 的 k 通道所连的光端口发生拥塞,临时关断。由于关断消息在所有波长上广播,所有可能向下游拥塞光端口发送数

据分组的上游光端口都会收到关断消息。当下游光端口的共享缓冲区占用率降低到阈值以下时,产生并广播 traffic-message(ON, k)消息,表示 AWGR 的 k 通道所连的光端口已恢复畅通。上游光端口根据收到的关断或开通消息决定是否在某一波长位置,向特定下游光端口发送数据分组。此外,“信用”和“通断”流量控制消息可以嵌入在数据分组的头微片中,利用已有的带内(in-band)数据链路进行传输,以最小化额外的网络负载。

6 网络性能仿真

本文使用 OMNeT++ 软件框架^[8],编写了“天河二号”互连网络链路层的仿真模型,以及提出的基于 AWGR 的光电混合交换的仿真模型,并进行了网络吞吐率和传输延迟的对比仿真。OMNeT++ 是基于 C++ 语言的离散事件仿真框架,广泛应用于通信网络中消息事件的建模和仿真。OMNeT++ 提供了带参数的链路、自定义消息、网络端口、拓扑连接、网络模块的软件基础设施。通过继承系统定义的网络模块,重载消息回调函数,可以建立网络中的主机、交换机模型;通过设定链路参数、组合网络模块、定义网络拓扑,可以搭建出完整的互连网络,实现链路层的流量仿真。

表 1 电交换网络和光电混合交换网络的仿真设定

仿真参数	仿真设定
流量模型	均匀随机(uniform random)、热点(hotspot)
端口和链路带宽	112 Gbps
链路延迟	一二级交换机间 50 ns,二三级交换机间 100 ns
NRC 零负载延迟	100 ns
网卡穿越延迟	320 ns
计算节点时钟	437.5 MHz
数据分组长度	5 个微片,每个微片 32 个字节
TL 波长切换时间	2~4 ns
NRC/AWGR 端口数	24
计算节点数量	18 432

网络仿真对比了光电混合交换网络、电交换网络的吞吐率和延迟性能。仿真的设定条件如表 1 所示,参考了天河超级计算机互连网络的设计与测试参数,根据仿真结果绘制出图 6 所示的性能-负载曲线。注入负载(offered load)和吞吐率(throughput rate)归一化到所有计算结点的总带宽,平均延迟(average latency)由计算结点一段时间内收到的所有微片的总延迟取算术平均得到,包含了路由器延

迟、链路延迟和网卡穿越延迟.

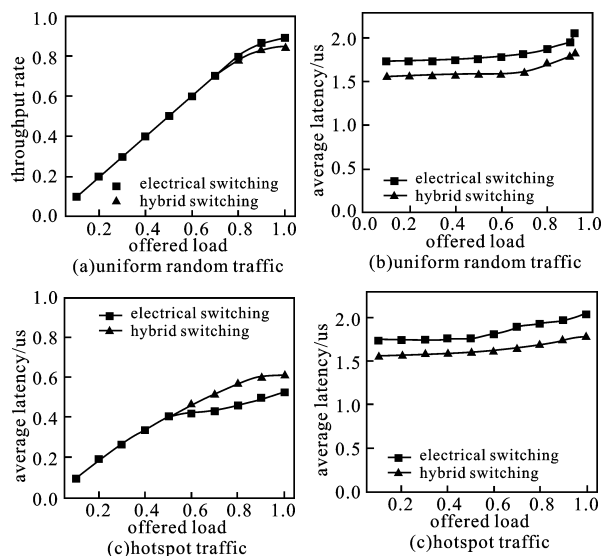


图6 随机均匀和热点流量下的吞吐率和延迟仿真

图6中,均匀随机(uniform random)流量模型下,光电混合交换网络和电交换网络的吞吐率-负载特性(a)和延迟-负载特性(b)表明:光电混合交换网络的延迟比电交换网络低160~200 ns(约8~10%),由于快速可调谐激光器波长切换的额外时间开销,最大吞吐率为84.56%,与电交换网络的88.7%相比略有降低。热点(hot-spot)流量模型下吞吐率-负载特性(c)和延迟-负载特性(d)表明:AWGR消除了端口竞争,混合交换网络的最大吞吐率为61.53%,高于电交换网络的51.26%,提升约10%;由于改善了端口的拥塞状况,平均延迟降低了180~250ns。光电混合交换网络保持了电交换网络的高吞吐量优势,降低了报文传输延迟。

7 结束语

本文在“天河二号”超级计算机互连网络的软硬件基础上,提出了一种基于阵列波导光栅路由器(AWGR)的光电混合交换的网络方案。仿真结果表明,光电混合交换网络的平均延迟降低了8%~10%的同时,热点流量下的吞吐量有所提升。

参考文献:

[1] LIAO X K, PANG Z B, WANG K F, et al. High per-

formance interconnect network for Tianhe System [J]. Journal of Computer Science and Technology, 2015, 30(2):259.

- [2] YE X, MEJIA P, YIN Y, et al. DOS - A scalable optical switch for datacenters [C]//ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS). La Jolla, CA, USA: IEEE, 2010: 1-12.
- [3] CAO Z, PROIETTI R, YOO S J B. Hi-LION: Hierarchical large-scale interconnection optical network with AWGRs [J]. IEEE/OSA Journal of Optical Communications and Networking, 2015, 7(1): 97-105.
- [4] CHEUNG S, SU T, OKAMOTO K, et al. Ultra-Compact Silicon Photonic 512×512 25 GHz Arrayed Waveguide Grating Router [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2014, 20(4): 310-316.
- [5] SCOTT S, ABTS D, KIM J, et al. The BlackWidow High-Radix Clos Network [C]//33rd International Symposium on Computer Architecture (ISCA'06). Boston, MA, USA: IEEE, 2006: 16-28.
- [6] SIMSARIAN J E, LARSON M C, GARRETT H E, et al. Less than 5-ns wavelength switching with an SG-DBR laser [J]. IEEE Photonics Technology Letters, 2006, 18(4): 565-567.
- [7] YE X, PROIETTI R, YIN Y, et al. Buffering and Flow Control in Optical Switches for High Performance Computing [J]. IEEE/OSA Journal of Optical Communications and Networking, 2011, 3(8): 59-72.
- [8] OMNeT++ Discrete Event Simulator [EB/OL]. Republic of Seychelles: Simulcraft Inc., 2018[2018-11-10]. <https://omnetpp.org>.

作者简介:

王昭男, (1994-), 硕士研究生。研究方向为高性能计算光互连网络分析与建模。

E-mail: wangzhaojack@163.com.

裴秉玺男, (1994-), 博士研究生。研究方向为片上光互连网络分析与建模。

何卫锋男, (1976-), 博士, 博士生导师, 副研究员。研究方向为光电混合集成电路技术, 硅光互连网络。