第 41 卷 第 14 期/2021 年 7 月/光学学报 研究论文

光学学报

一种新型的高容量光互连架构

杨晓雪[1]，胡冰[1,2*]，魏晓强[3]，尚迎春[3]

[1]浙江大学信息与电子工程学院，浙江 杭州 310027；

[2]之江实验室智能网络研究中心，浙江 杭州 310027；

[3]中兴通讯股份有限公司移动网络和移动多媒体技术国家重点实验室，浙江 杭州 518055

摘要 光互连具有低功耗、大带宽等优越性能，可以实现数据中心节点数与交换容量的大幅增大。提出了一种基于阵列波导光栅(AWGR)的新型大容量光互连架构，通过可调波长转换器与 AWGR 提供波长路由，并利用分布式控制实现快速配置与低延迟。无缓冲的光交换机可能产生数据包争用，基于光纤延迟线的光缓冲可用于争用解决。详细描述了分别适用于严格无阻塞网络与大规模互连的两种实现方案，对所提架构在不同网络规模、流量模式、缓冲容量下的性能进行分析对比，仿真结果表明，该架构可以互连 32768 个节点，且具有低延迟与大吞吐量。

关键词 光通信；光互连架构；无阻塞网络；低延迟；高吞吐量；丢包率

中图分类号 TN915.02 **文献标志码** A **doi**：10.3788/AOS202141.1406002

Novel Architecture for High Capacity Optical Interconnects

Yang Xiaoxue[1], Hu Bing[1,2*], Wei Xiaoqiang[3], Shang Yingchun[3]

[1] *College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*;

[2] *Intelligent Network Research Center, Zhejiang Lab, Hangzhou, Zhejiang* 310027, *China*;

[3] *State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation, Hangzhou, Zhejiang* 518055, *China*

Abstract Optical interconnects have advantages of low power consumption and large bandwidth and can achieve a substantial increase in the number of nodes and switching capability for data centers. An arrayed waveguide grating router (AWGR) based architecture for high capacity optical interconnects is proposed in this paper. The tunable wavelength converter and AWGR are utilized to provide wavelength routing, and a distributed control is used to achieve fast configuration and low latency. Packet contention may occur at the buffer-less optical switches, and the optical buffering approach based on fiber delay lines (FDLs) is introduced for contention resolution. Two implementations are employed to support strictly non-blocking network and large-scale interconnection. In this paper, we analyze and compare the performance of the proposed architecture in terms of the network size, traffic mode, and buffer capacity. Simulation results indicate that the proposed architecture can interconnect 32768 nodes while providing low latency and high throughput.

Key words optical communications; optical interconnect architecture; non-blocking network; low latency; high throughput; packet loss probability

OCIS codes 060.1155; 060.4258

收稿日期：2020-12-29；**修回日期**：2021-02-02；**录用日期**：2021-02-22

基金项目：国家重点研发计划(2019YFB1802905)、国家自然科学基金面上项目(61971377)

通信作者：*binghu@zju.edu.cn

1 引　　言

随着云计算、大数据、物联网的发展，用户的带宽需求显著增长，大容量、高可靠性的光通信系统获得了研究人员的广泛关注[1-2]。通信网络的常见交换架构有 Clos[3]、Benes[4]、胖树[5]等，但随着互连节点数的增加，这些含有集中控制器的多层体系结构具有极大的延迟和极高的复杂性，与其相比，具有高度分布式控制的扁平化架构在提供大容量和可扩展性的同时，缩短了控制器的处理时间，进而从缓冲、吞吐量、延迟等方面优化了系统的整体性能[6]。

AWGR 是一种无争用的波长路由器件，其将空间域与波长域相结合，每个端口可以同时处理具有不同波长的多个信号，这种方式有效增大了交换结构的带宽，单个 AWGR 的吞吐量达 Tbit 级[7]。为互连更多节点，文献[8]提出一种基于小基数 AWGR 的 DC 交换机和子系统，用于组建大端口光交换机。Hipolaos 将基于 AWGR 的波长路由与基于 Spanke 的广播与选择路由相结合，该架构具有较多端口数、大吞吐量与低延迟[1]。基于 AWGR 的体系结构具有很多优势，但仍有一定局限性。当网络阻塞时，光数据包难以缓冲，因此研究人员设计了多种争用解决模块，DOS[9]利用环回共享缓冲区降低争用概率，Flex-LIONS 利用空间交换机直连争用节点对，以形成带宽可重新配置的网络[10]。AWGR 多与可调波长转换器(TWC)结合进行波长路由，但 TWC 具有极高的能耗，为提高能源效率，文献[11]提出了一种基于 AWGR 和光空间交换机的交换架构。

目前大量研究的关注点为大吞吐量，关于无阻塞架构的研究不够深入。无阻塞架构在提供高性能的同时，大幅度降低了网络的阻塞概率与排队延迟。考虑到 AWGR 的无争用特性，本文设计了一种基于 AWGR 与 TWC 的交换架构，其通过配置 TWC 进行信号的波长转换，进而利用 AWGR 的波长路由特性形成交换矩阵，并引入了光纤延迟线(FDL)进行争用解决，同时，本文给出了此架构在不同网络规模下的实现方案。

2 架　　构

2.1 阵列波导光栅

AWGR 是一种无源光学设备，可以为输入与输出节点对提供无争用的连接，其具有一些出色的特性，如低通道串扰、低插入损耗、长期稳定性及易于与光纤耦合等。AWGR 易于集成，文献[12]提出了一种具有 512 个端口的 AWGR，其面积仅为 16 mm×11 mm。SiPh 层可用于垂直集成 AWGR，以减少空间占用。AWGR 具有循环波长路由特性，可描述为

$$j=(i+w)\bmod M, \tag{1}$$

式中：j 为输出端口序号；i 为输入端口序号；w 为波长编号；M 为 AWGR 的端口数；mod 为求余运算函数。

通过将信号转换为合适的波长，可以利用 AWGR 的循环波长路由特性将其路由至相应输出端口处。图 1 所示为 4×4 端口的 AWGR，其中，$\lambda_w^{(i)}$ 中的 λ 表示此信号为具有特定波长的光信号，其上角标 i 为信号的输入端口序号，下角标 w 为信号的波长编号。

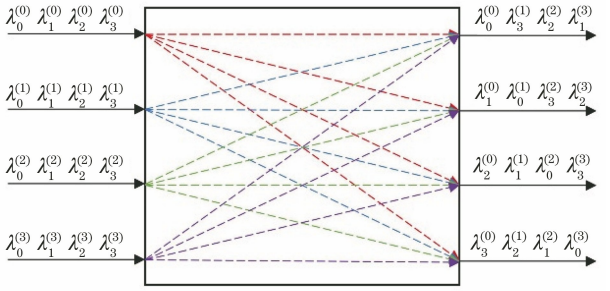

图 1　AWGR 的循环波长路由特性

Fig. 1　Cyclic wavelength routing characteristic of AWGR

2.2 架构设计

$M\times M$ 端口的 AWGR 可以同时路由 M^2 个数据包，从而用更少的光学设备传输更多的数据。为扩展端口数量以互连更多的服务器，本文提出了一种基于 AWGR 和 TWC 的两级交换架构，名为 AA，如图 2 所示。此架构的主要组成为：与外部网络相连的线卡及路由信号的交换结构，其中，线卡的作用是缓冲与波长转换，交换结构的作用是将光信号路由至对应的输出端口。此交换结构分为两级：输入模块(IM)与输出模块(OM)，每个模块由 M 个 $M\times M$ 端口的 AWGR 构成，由 TWC 模块相连。图中，AWGR 的上角标表示其所属模块(IM 或 OM)，下角标表示其在 IM(OM)中的编号。该交换结构允许光信号从任意输入端口传输至任意输出端口，其规模为 $M^2\times M^2$，且具有良好的可扩展性。IM(OM)的每个输入(输出)端口都有一条光纤连接至线卡，波分复用(WDM)技术可以实现多波长光信号的同时传输，这不仅增大了端口带宽，且提供了一对多、多对一的通信。与传统架构相比，AA 架构允许器件直连，从而无需波导交叉，且其采用独立的分布式控制机制，这种机制简单且延迟较低。

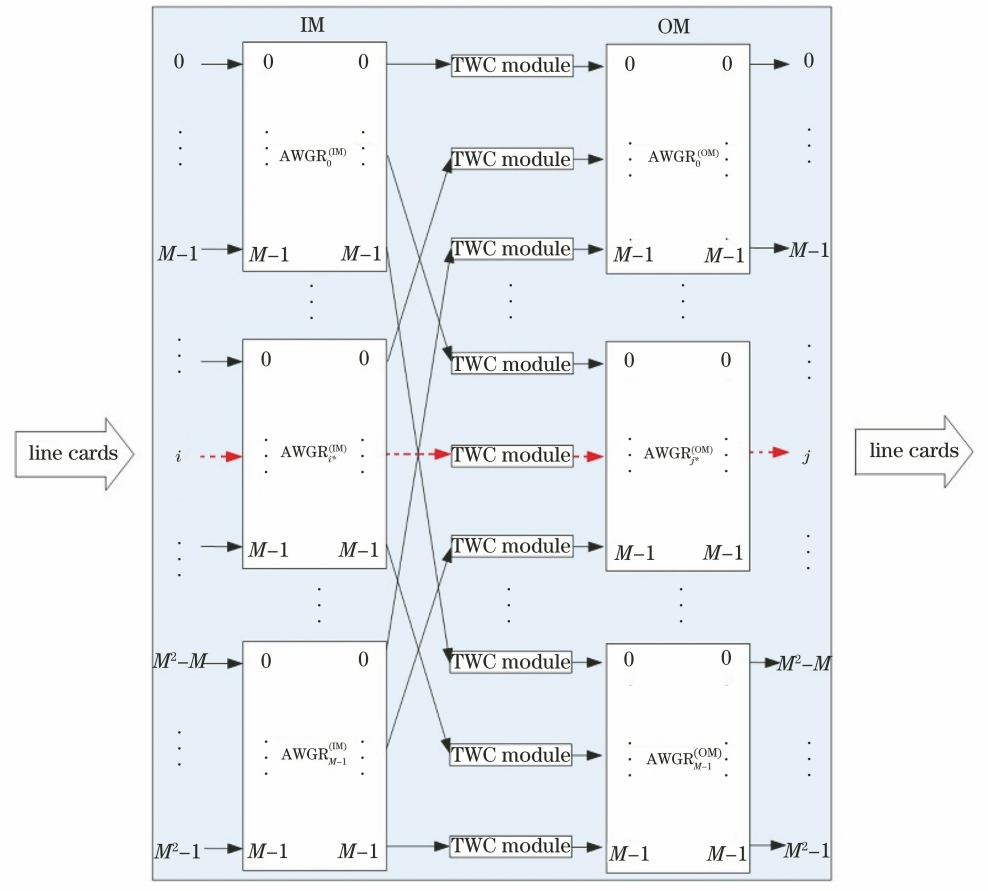

图 2　AA 架构

Fig. 2　AA architecture

在每个时隙内，如果每个输入端口仅处理一个数据包，则 AA 架构仅能同时传输 M^2 个数据包。基于 WDM 技术，每个 AWGR 可以同时处理 M^2 个数据包，则此架构最多支持 M^3 个数据包的传输。本文将在第三节对这两种实现方法进行详细描述。

2.3　路由算法

一个在交换结构上的输入端口为 i、输出端口为 j 的信号，需要经由 IM 级 $A_{i^*}^{(\mathrm{IM,AWGR})}$、OM 级 $A_{j^*}^{(\mathrm{OM,AWGR})}$ 进行路由，到达目的地址。其中，$A_{i^*}^{(\mathrm{IM,AWGR})}$、$A_{j^*}^{(\mathrm{OM,AWGR})}$ 的上角标分别表示其为 IM、OM 级 AWGR，其下角标分别为

$$i^* = i/M, \tag{2}$$

$$j^* = j/M。 \tag{3}$$

此信号在与其相连的 AWGR 上对应的本地端口号为

$$i' = i \bmod M, \tag{4}$$

$$j' = j \bmod M, \tag{5}$$

图 2 中的虚线表示从输入端口 i 到输出端口 j 的路径。AWGR 是一种不可配置的光器件，可以通过转换输入信号的波长来确保其到达目的端口，因此，在两级 AWGR 前均需要配置 TWC，即分别进行两次波长转换。

第一次波长转换在 IM 级 AWGR 之前的线卡上。对于上述信号，当其到达 $A_{i^*}^{(\mathrm{IM,AWGR})}$ 上的本地输入端口 i' 后，应将其发送至本地输出端口 j^*，此端口与 $A_{j^*}^{(\mathrm{OM,AWGR})}$ 相连。此时，TWC 应配置为

$$\lambda_{\mathrm{IM}}(i,j) = \lambda_{(j^*-i'+M) \bmod M}, \tag{6}$$

式中：$\lambda_{\mathrm{IM}}(i,j)$ 为输入端口为 i、输出端口为 j 的信号在经第一次波长转换后的波长。

第二次波长转换在 OM 级 AWGR 之前的 TWC 模块上。因此，$A_{j^*}^{(\mathrm{OM,AWGR})}$ 应将信号从本地输入端口 i^* 路由至本地输出端口 j'，此时，TWC 应配置为

$$\lambda_{\mathrm{OM}}(i,j) = \lambda_{(j'-i^*+M) \bmod M}, \tag{7}$$

式中：$\lambda_{\mathrm{OM}}(i,j)$ 为输入端口为 i、输出端口为 j 的信号在经第二次波长转换后的波长。

2.4　线卡设计

线卡是交换架构与外部网络的接口，多个数

据包发往同一地址时会发生阻塞，为避免产生丢包，线卡需要为信号提供缓冲。光信号的缓冲有两种方法：1）将光信号转换为电信号，并将其存储于队列中，等待转发，这种方式需要进行光-电-光转换，这大幅度增加了延迟；2）利用 FDL，此时延迟时间取决于光纤的长度，这种方式具有可编程、范围大、精度高的优势[13]。以 FDL 为主要器件的缓冲模块分为前馈式与反馈式，两者各有优劣：前馈式结构不支持优先级路由，而反馈式结构具有更大的衰减与串扰。文献[14]提出了一种前馈与反馈的混合式结构，使缓冲模块同时具备两者的优势。本文参考这种混合式结构，设计了一种基于 AWGR 与 TWC 的争用解决模块。其中，线卡采取分布式控制的方式调度信号，各线卡间的信号不会发生阻塞。

考虑一种简单的情况：每个端口仅处理一个数据包，即单波长输入，如图 3 所示。线卡不具备路由功能，即由输入端口 i 进入线卡的信号会经由输出端 i 发往交换结构。部分输入信号被分离并发送至控制器，其余数据被存储于 FDL 中，直到控制器完成 TWC 配置。在这种情况下，需要避免两种争用：线卡输出端口 TWC 争用与目的地址争用。线卡输出端口 TWC 争用是指 TWC 仅能处理单个光信号；目的地址争用是指单个线卡与服务器间不可有两组数据同时传输。争用的信号将被缓冲至 FDL 中，使其具有不同的延迟。输入信号发往 $\mathrm{AWGR_M}$ 后，没有争用或在争用中胜出的信号将会被直接发往对应的输出端口；被阻塞的信号将会被发往前馈缓冲，其中，前馈缓冲含 K 条 FDL，每条 FDL 的延迟数不等，第 1，2，…，K 条 FDL 的延迟分别用 τ_1，τ_2，…，τ_K 表示。经前馈缓冲后，这部分信号经由 $\mathrm{AWGR_A}$ 发往对应的输出端口；如果没有空闲的前馈缓冲可用于存储信号，则其将会被发往反馈缓冲，其中，反馈缓冲含 N 条 FDL，第 1，2，…，N 条 FDL 的延迟分别用 τ_1'，τ_2'，…，τ_N' 表示。经过反馈缓冲的信号将会被再次发往 $\mathrm{AWGR_M}$ 的输入处进行调度。前两级 TWC 用于辅助 AWGR 进行波长路由，第三级 TWC 用于控制信号的输出波长，即进行 2.3 节所述的第一次波长转换，以便在交换结构中进行路由。AWGR 的端口数决定了线卡的缓冲容量，但因其受限于 TWC 的调谐范围与 AWGR 的串扰，端口数最大可达 64 个，故设置 $M+K+N\leqslant 64$，其中，K 与 N 分别为前馈缓冲、反馈缓冲中 FDL 的数量。对于 $M<32$ 的网络，将 K 与 N 均设置为 M；对于 $M=32$ 的网络，因受限于 AWGR 的端口数，故将 K 设置为 24，将 N 设置为 8。

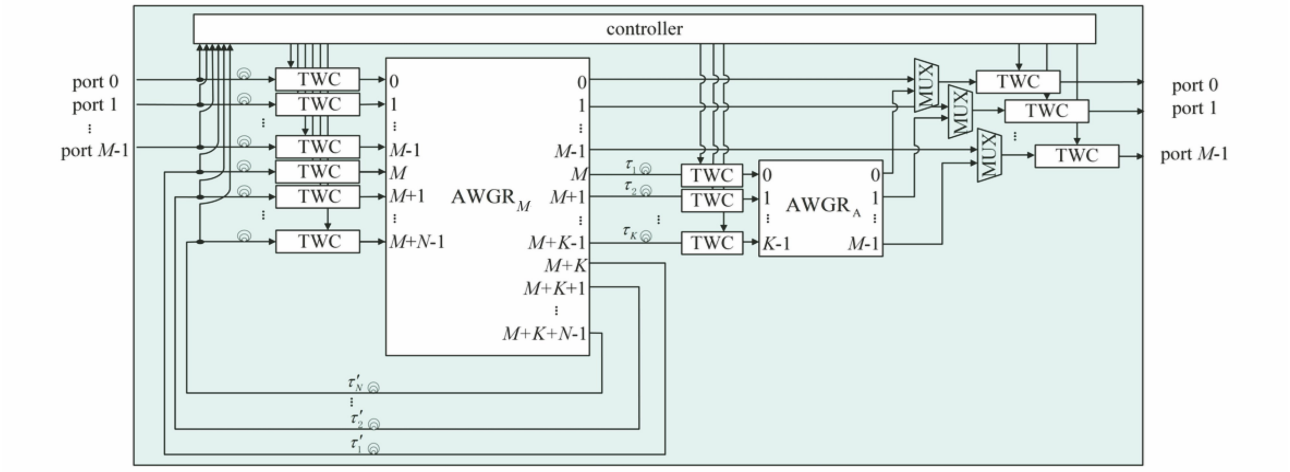

图 3 单波长输入的线卡设计

Fig. 3 Line card design for single-wavelength input

利用 WDM 技术，使线卡的每个端口可以同时处理包含 M 个波长的信号，对 WDM 输入的线卡结构在单波长输入的基础上进行了扩展，如图 4 所示。线卡的输入、输出处分别增加了解复用与复用模块，争用解决模块与单波长输入相似。

不同之处在于：

1）控制器需要避免的争用情况增加，除线卡输出端口 TWC 争用与目的地址争用外，增加 OM 争用。其中，OM 争用是指单个线卡输入端口的数据必须发往不同的 OM，这是由 AWGR 的循环波长路由特性决定的。

2）缓冲容量增大。因 WDM 输入是有阻塞网络，其阻塞概率远高于单波长输入，故而需要设置更多 FDL，以降低丢包率。对于 $M<32$ 的网络，将 K

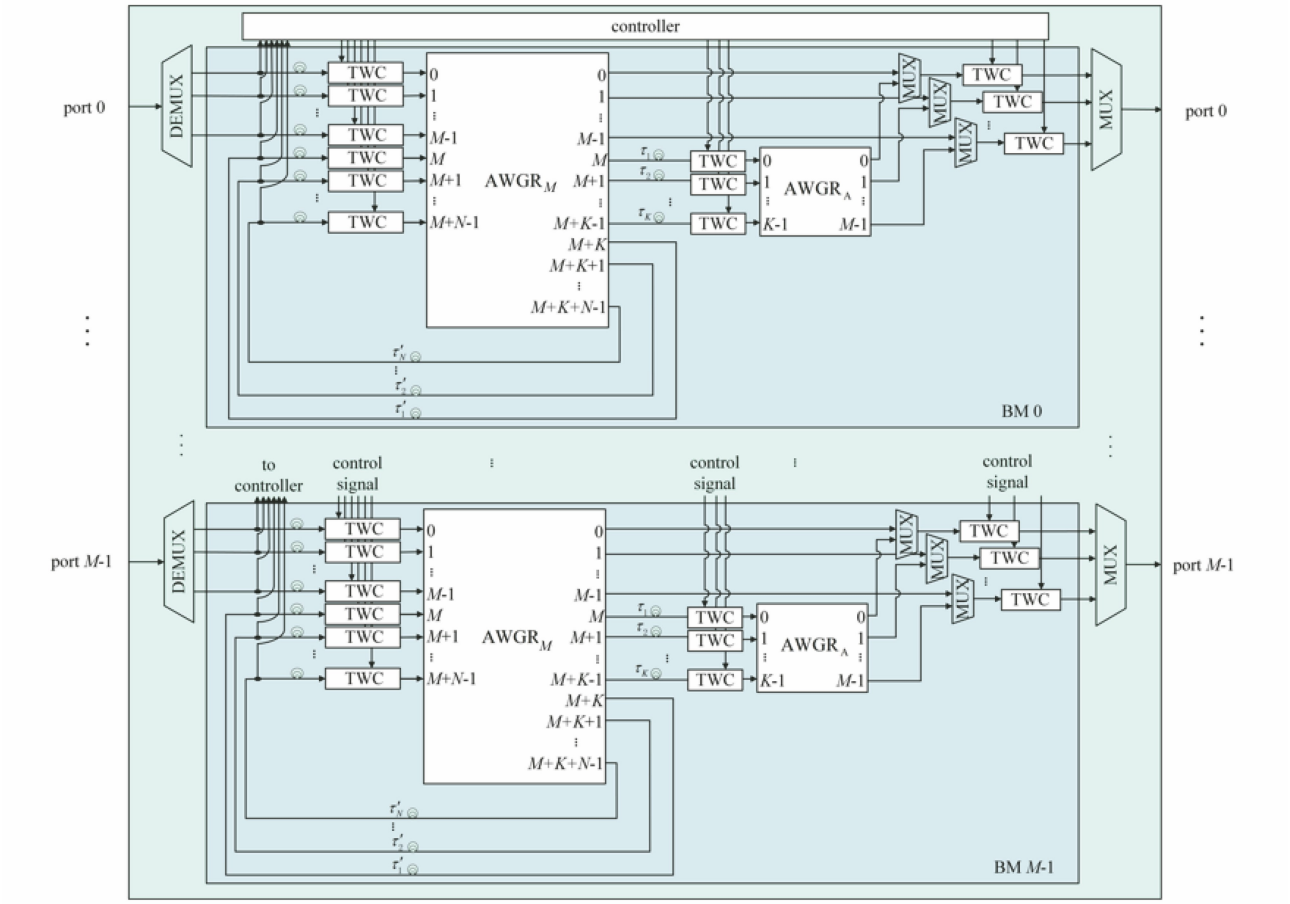

图 4 WDM 输入的线卡设计

Fig. 4 Line card design for WDM input

设置为 32，N 设置为 16；对于 $M=32$ 的网络，将 K 设置为 24，N 设置为 8。

2.5 TWC 模块设计

TWC 是一种波长调制器件，其重构速度可达纳秒级，它常与 AWGR 结合，进行波长路由，同时，TWC 是为所有节点建立全连接的关键器件。对两级 AWGR 而言，由特定输入端口进入交换结构的 M 个信号仅能被路由至 M 个特定输出端口，无法实现全连接，TWC 模块允许数据被路由至任意输出端口。每个 TWC 模块都是独立的，且利用分布式控制配置器件与调度信号。

图 5 为 TWC 模块的内部结构。每根波导可以同时传输 M 个不同波长的光信号，这些光信号进入

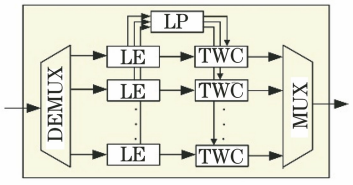

图 5 TWC 模块

Fig. 5 TWC module

TWC 模块后，经过解复用（DEMUX），根据其波长发往不同的标签提取器（LE）。标签随后被发往标签处理器（LP），各 LE 所传输的标签数据经独立处理后，LP 生成控制信号并对其进行 TWC 配置，信号经波长转换后，被复用并发往相应的 OM 级 AWGR。

3 实 现

3.1 单波长输入

如图 6 所示，线卡的每个输入（输出）端口都有两条光纤连接至架顶式交换机（TOR），TOR 生成的信号被传输至背板上的相应端口。信号经线卡调度后，由 $M^2\times M^2$ 交换矩阵进行路由，并再次经线卡发往目的服务器。

与单个 AWGR 相比，AA 架构可以互连更多的服务器，达到更大的吞吐量。受限于串扰等因素，设置 AWGR 的大小不高于 32×32，且设置单波长带宽为 25 Gbit·s^{-1}，则 1024 个节点可以通过两级架构相互通信。

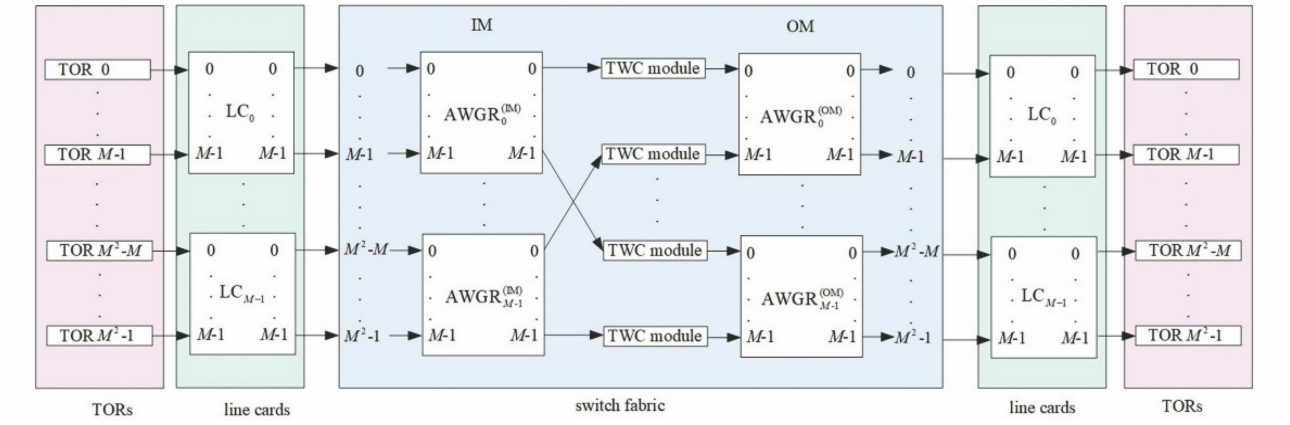

图 6 基于 AWGR 与 TWC 光网络的单波长输入实现

Fig. 6 Single-wavelength input implementation of optical network based on AWGR and TWC

如果背板的每个输入端口仅处理单波长信号，而不限制输出端口的连接，则 AA 是严格无阻塞架构，且背板可以同时处理 M^2 个数据包，证明如下。

根据信号的输入输出端口对，可以分别为 IM 和 OM 选择合适的波长。相同波长的信号必定具有不同的输入端口，故而具有不同的输出端口。也就是说，单个 IM 级 AWGR 上的输出端口只可能含有不同波长的信号，而不会发生波长阻塞。这些信号经 TWC 进行波长转换后，传输至对应的 OM 级 AWGR。同样地，因 AWGR 的无争用波长路由特性，这些信号不会在 OM 级发生阻塞。这表明，不论网络处于何种状态，可以为任意空闲输入输出端口建立连接，而不会影响网络中的其他连接，即此网络为严格无阻塞网络。

3.2 WDM 输入

为增大交换机规模，在架构的输入、输出端口分别设置复用模块(MUX)与波长选择开关(WSS)，使背板的每个输入(输出)端口可以同时发送(接收)具有不同波长的 M 个信号，以实现更大的吞吐量。

图 7 为 WDM 输入的实现。在输入端，最多 M 个并发光数据包通过 1∶M MUX 复用至单根光纤，并传输至线卡的相应输入端口，这些光信号须在不同的波长上传播，否则会产生争用。同样地，线卡的每个输出端口均部署了 WSS，信号经 WSS 路由被发送至目的服务器。

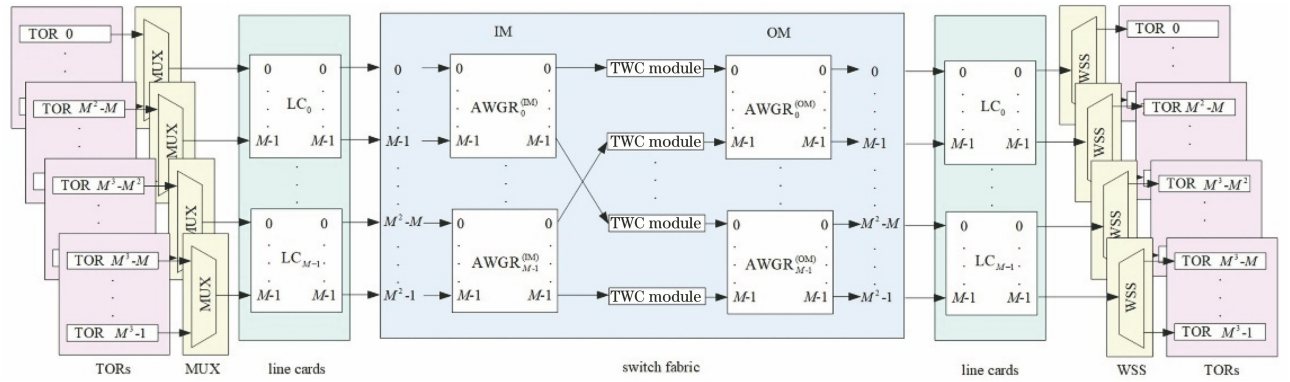

图 7 基于 AWGR 与 TWC 光网络的 WDM 输入实现

Fig. 7 WDM input implementation of optical network based on AWGR and TWC

此实现不再是无阻塞架构：对于某个输入端口而言，AWGR 的波长路由特性使得 M 个信号会被路由至不同 OM，即如果 $m(1\leqslant m\leqslant M)$ 个数据包的目的地址位于同一 OM，网络会发生阻塞。根据这种特性，当同一输入端口的数据包具有相同目的 OM 时，需要在线卡处进行缓冲，以避免争用，同时，来自单个 IM 级 AWGR 的 M^2 个数据包应被发往不同的输出端口。因此，此架构最多可以同时路由 M^3 个信号，即可以进行 M^3 个服务器间的通信。网络的可扩展性依赖于 AWGR 的端口数，如果使用 32×32 端口的 AWGR，则此架构可以互连 32768 个节点。

4 仿　　真

本文利用 OMNeT++对上述架构进行仿真，研究其在不同网络规模、流量模式、缓冲容量下的吞吐量与端到端延迟等网络性能指标。为模拟真实的网络环境，设定各输入的发包模式为具有固定概率的伯努利模式，以 32 ns 的间隔将数据包发送至网络中，每个连接可以同时承载 M 个具有不同波长的数据包，且单波长带宽为 25 Gbit·s^{-1}。控制器的调度时间为 450 ns，TWC 的配置时间为 4 ns。

4.1 网络模型与性能估算

为分析 AA 架构的网络性能，将解决争用的 $\mathrm{AWGR_M}$ 结构简化为仅具有反馈 FDL 的缓冲模块，如图 8 所示。参考文献[6]，对网络的阻塞率及吞吐量进行分析。在单波长输入情况下，网络负载为固定值，各输入以概率 p 发送数据包，且在随机流量模式下，数据包被发送至各目的输出的概率为 $\frac{p}{M^2}$。因阻塞仅发生在单张线卡的 M 个数据包中，故对于某输入 i，其阻塞概率 P_{cont} 为

$$P_{\mathrm{cont}}=p\cdot\sum_{k=1}^{M-1}C_{M-1}^{k}\left(\frac{p}{M^2}\right)^{k}\left(1-\frac{p}{M^2}\right)^{M-1-k},\tag{8}$$

式中：k 为与此输入具有相同目的输出的数据包个数。

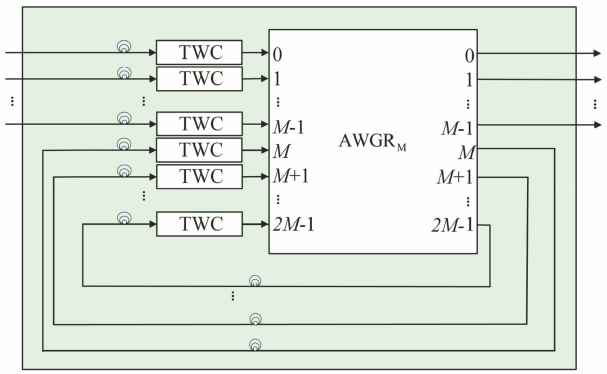

图 8　简化的缓冲模块

Fig. 8　Simplified buffer module

对于 $k+1$ 个数据包，仅有 1 个数据包被直接发送，其余 k 个数据包可发往反馈 FDL 的任意端口，但此数据包在线卡上的输出端口仍为 i，即反馈 FDL 的作用与重传相同，则重传概率 P_{fb} 为

$$P_{\mathrm{fb}}=p\cdot\sum_{k=1}^{M-1}C_{M-1}^{k}\left(\frac{p}{M^2}\right)^{k}\left(1-\frac{p}{M^2}\right)^{M-1-k}\frac{k}{k+1}。\tag{9}$$

在 WDM 输入情况下，各端口会产生 OM 争用，则对于每个端口的 M 个不同波长数据包，OM 争用导致重传的概率 P_{fb1} 为

$$P_{\mathrm{fb1}}=p\cdot\sum_{k=1}^{M-1}C_{M-1}^{k}\left(\frac{p}{M}\right)^{k}\left(1-\frac{p}{M}\right)^{M-1-k}\frac{k}{k+1}。\tag{10}$$

除去因 OM 争用被重传的数据包外，每个端口平均有 $32(p-P_{\mathrm{fb1}})$ 个数据包，每个数据包由目的地址争用导致重传的概率 P_{fb2} 为

$$P_{\mathrm{fb2}}=(p-P_{\mathrm{fb1}})\cdot\sum_{k=1}^{(M^2-M)(p-P_{\mathrm{fb1}})}C_{(M^2-M)(p-P_{\mathrm{fb1}})}^{k}\left(\frac{p}{M^2}\right)^{k}\cdot\left(1-\frac{p}{M^2}\right)^{(M^2-M)(p-P_{\mathrm{fb1}})-k}\frac{k}{k+1}。\tag{11}$$

结合(10)式和(11)式，可得 WDM 输入的重传概率为

$$P_{\mathrm{fb}}=P_{\mathrm{fb1}}+P_{\mathrm{fb2}}。\tag{12}$$

数据包重传导致 $\mathrm{AWGR_M}$ 各输入端口接收到数据包的概率 $p(i)$ 增大为

$$p(i)=p+P_{\mathrm{fb}}(i-1)。\tag{13}$$

对 $p(i)$ 与 $P_{\mathrm{fb}}(i)$ 进行迭代计算，直到其达到稳定值，分别记为 p_{R} 与 P_{fbMax}，二者分别表示各端口的实际流量与在此流量下的最大重传概率。吞吐量为网络在单位时间内可以成功传输的数据量，因各节点的网络模型一致，本文采取归一化方式，将吞吐量 T 表示为每个节点成功发送数据包的比例，即

$$T=p_{\mathrm{R}}-P_{\mathrm{fbMax}}。\tag{14}$$

对上述简化模型进行仿真，图 9 所示为吞吐量理论结果与仿真结果的对比。因单波长输入的阻塞率极低，其重传概率 P_{fbMax} 最高仅为 1.5%，故其理论与仿真结果均达近 100%；而因 WDM 输入的阻塞率较高，数据包重传进一步增大了输入端口的实际负载，在网络负载 $l_{\mathrm{load}}=0.5$ 时，网络已接近饱和状态，重传

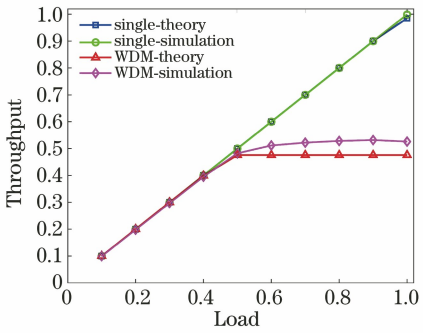

图 9　理论结果与仿真结果对比

Fig. 9　Comparison of theoretical results and simulation results

概率 P_{fbMax} 高达 52.4%，此网络的理论吞吐量为 47.6%，仿真吞吐量为 52.6%，偏差在 5%以内。

4.2 单波长输入的相关仿真

在单波长输入情况下，以随机流量模式生成数据包，即将各服务器所生成的数据包以相等的概率发送至 M^2 个目的端口。缓冲容量无疑会影响仿真结果，本文采取分组线性分布：将具有 Q 条 FDL 的前馈/反馈式缓冲分为具有不同延迟的 P 组，每组 FDL 的数量尽量相等且延迟线性增加：$P=\mathrm{lb}\,Q+1$[15]。如对于具有 32 条 FDL 的缓冲，令 6，6，5，5，5，5 条 FDL 分别具有 1，2，3，4，5，6 个时隙的延迟。

对于单波长输入在不同网络规模，即 $M=2^t$（$1\leqslant t\leqslant 5$）时的吞吐量、端到端延迟进行仿真。其中，t 为时间，M 为 AWGR 的端口数，即网络节点数为 $M^2=2^{2t}$（$1\leqslant t\leqslant 5$）。每个节点仅可使用单个波长，单波长带宽为 25 Gbit·s^{-1}，则理论吞吐量最高可达 51.2 Tbit·s^{-1}。当数据包不产生争用时，其端到端延迟为 6.7×10^{-7} s。仿真结果如图 10 所示，AA 架构的吞吐量与流量负载在不同网络规模下均呈线性关系，且接近 100%；文献[6]所提出的架构（以 CRB 代指）在 $l_{\mathrm{load}}\geqslant 0.7$ 时的吞吐量明显低于 AA 架构，且性能差距最高可达 40%。从延迟来看，低负载与 $l_{\mathrm{load}}=1$ 时，CRB 架构的延迟相比 AA 架构较低，高负载时，AA 架构的延迟相比 CRB 架构较低；当负载低于 0.9 时，AA 架构的丢包率极低，网络规模的增大反而会降低端到端延迟，这是由于随着网络增大，具有低延迟的 FDL 数量更多，且阻塞率降低，因此数据包在 FDL 中经历的平均延迟降低；但当负载为 1 时，节点数为 4、16 的网络因阻塞率较高，其丢包率急剧增加至 10^{-3} 级，且被丢弃的数据包具有较高的延迟，这使得当 $1\leqslant t\leqslant 3$ 时，网络规模较小的架构反而延迟更低，从曲线图可知，$M=8$ 的网络具有最高的延迟。

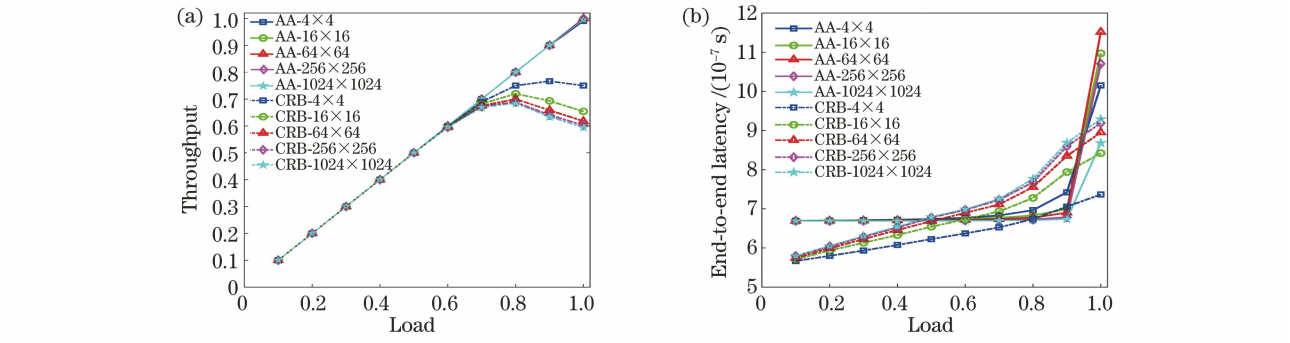

图 10 单波长输入在不同网络大小下的仿真结果。(a)吞吐量；(b)延迟

Fig. 10 Simulation results of single-wavelength input under different network sizes. (a) Throughput; (b) end-to-end latency

缓冲容量也极大影响了网络性能，其分布一般为线性分布与均等分布。其中，线性分布表示对具有 Q 条 FDL 的缓冲结构，第 1，2，…，Q 条 FDL 的延迟分别为 1，2，…，Q 个时隙；均等分布表示每条 FDL 的延迟均为 1 个时隙。为探究不同分布方式对网络性能的影响，对 $M=32$ 的网络分别配置分组线性、线性与均等分布的缓冲，如图 11 所示。由仿真结果可知，三种缓冲分布的性能相似，主要区别为

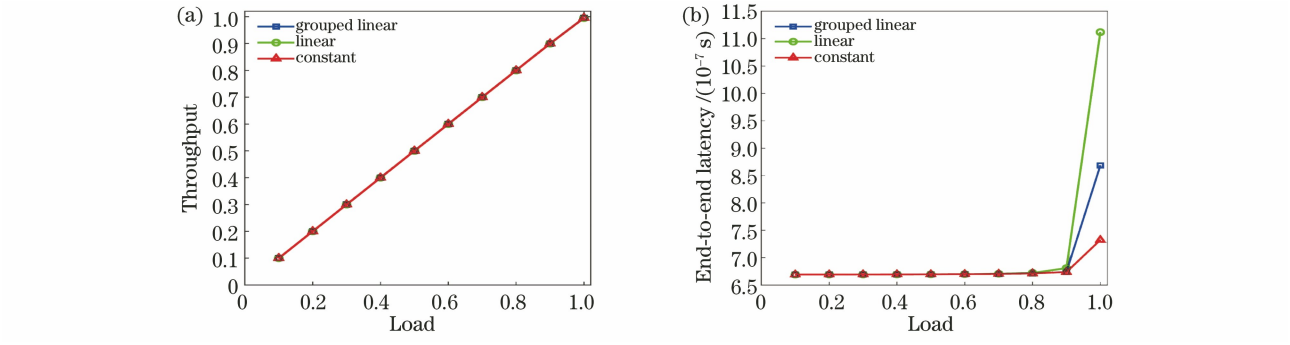

图 11 单波长输入在不同 FDL 分布下的仿真结果。(a)吞吐量；(b)延迟

Fig. 11 Simulation results of single-wavelength input under different FDL distributions. (a) Throughput; (b) end-to-end latency

负载为 1 时的延迟。线性分布的缓冲容量最高，可以容纳更多数据包，但 FDL 平均延迟也最高，这使得其丢包率极低，但延迟相比较高；均等分布的缓冲容量最低，因此其丢包率最高，延迟最低。这表示，当 FDL 数量有限时，网络需要平衡延迟与丢包率，由曲线图可知，分组线性分布是一种较优的平衡方式。

OM 级 AWGR 的每个输入端口最多收到 M 个数据包，故每个 TWC 模块包含 M 个 TWC，以构成无阻塞网络。但在实际网络中，平均每个 TWC 模块仅会收到单波长数据，即 TWC 利用率极低，因此为减少器件数量、成本与功耗，研究当 $M=32$ 时，TWC 数量对网络性能的影响，结果如图 12 所示。可以发现，当 TWC 的数量从 1 增加到 5 时，吞吐量逐渐增大，延迟基本不变，而丢包率逐渐降低。由仿真结果可知，TWC 数量为 5，6，32 的性能曲线产生重叠，这表示采用 5 个 TWC，即可在满足网络性能要求的同时，实现较低的成本与功耗。

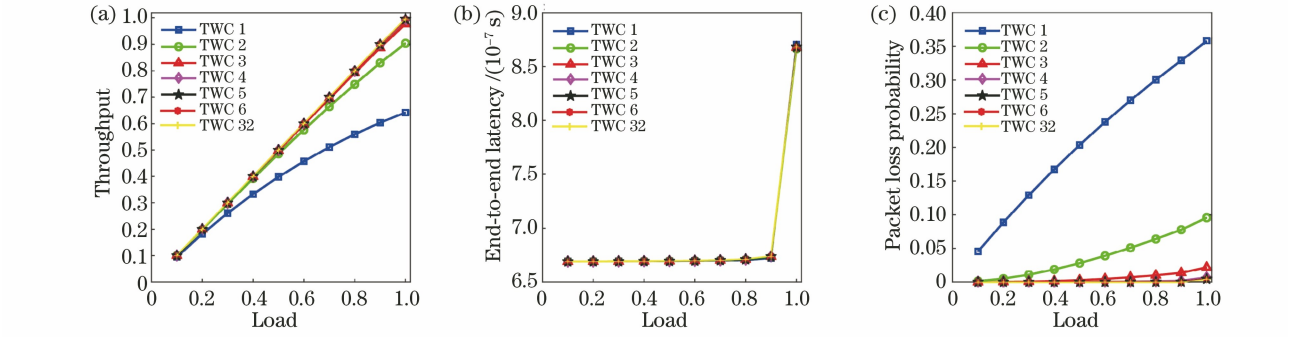

图 12　单波长输入在不同 TWC 个数下的仿真结果。(a)吞吐量；(b)延迟；(c)丢包率

Fig. 12　Simulation results of single-wavelength input under different numbers of TWC. (a) Throughput; (b) end-to-end latency; (c) packet loss probability

图 13 所示为 $M=32$ 的情况下，随机流量模式与位反转流量模式的比较。在位反转模式下，某输入的所有数据包被发往某个特定的目的端口，端口地址被预先确定为

$$d_n = s_{b-n-1},\ 0 \leqslant n \leqslant b-1, \tag{15}$$

式中：b 为地址的位数；d_n 为目标地址的第 n 位；s_{b-n-1} 为源地址的第 $b-n-1$ 位。目标地址的每一位 d_n 是源地址的每一位 s_n 的函数。

两种流量模式的吞吐量没有明显差异，但位反转模式的端到端延迟保持稳定，这是因为数据包之间没有争用，因此在线卡中，所有数据包将被直接发送而无需缓冲，故延迟不会受到流量负载变化的影响。然而，随机模式的阻塞概率随负载的增加而增加，故其负载为 1 时的延迟远高于位反转模式。

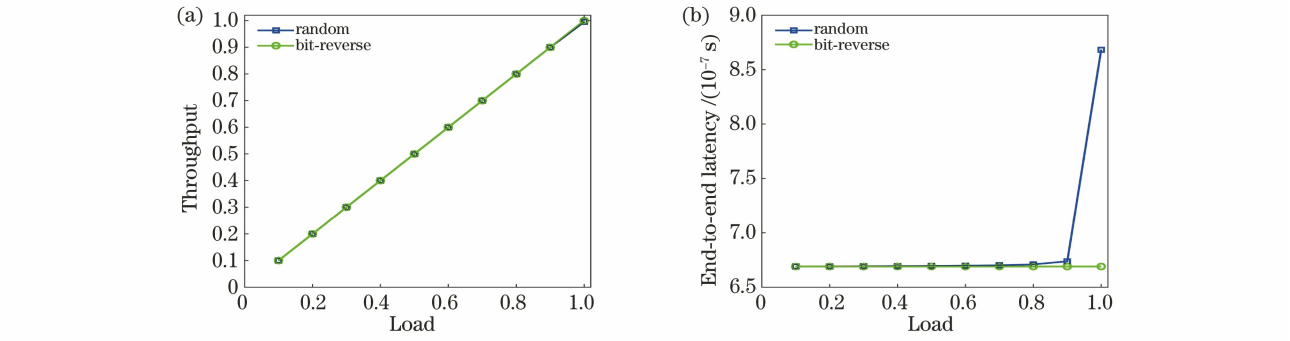

图 13　单波长输入在不同流量模式下的仿真结果。(a)吞吐量；(b)延迟

Fig. 13　Simulation results of single-wavelength input under different traffic patterns. (a) Throughput; (b) end-to-end latency

4.3　WDM 输入的相关仿真

为扩大网络规模并充分利用背板资源，本节仿真了架构在 WDM 输入时的性能，其中 $M=2^t$ ($1 \leqslant t \leqslant 5$)，即其所互连的节点数 $M^3=2^{3t}$ ($1 \leqslant t \leqslant 5$)。受限于 AWGR 串扰等，节点的可用波长数为 M，单波长带宽为 25 Gbit·s^{-1}。光互连的相关研究[1,6,10]所提出架构的互连节点数均为 1024，本文实现了 3 万节点的互连，且在吞吐量与延迟方面均具有较优的性能，图 14 所示为不同网络规模下的仿真结果。由曲线图可以看出，随着网络规模的增加，吞吐量急剧降低，当节点数为 8 时，吞吐量可达 99.3%，而当节点数达 32768 时，吞吐量仅为

66.4%，实际上，其吞吐量与延迟在 $l_{load}=0.7$ 附近已基本稳定。节点数为 64，512，4096 的网络均在负载为 0.9 时出现峰值，这是由于在满载情况下，丢包率极高，导致其吞吐量反而下滑。此外，在负载低于 0.7 时，各网络的吞吐量与负载基本等同，这说明此时丢包率极低，在同等节点数条件下，其性能仍优于 CRB 架构。由延迟结果可以看出，在负载低于 0.7 时，网络规模的增大会导致延迟增加，这是因为在 WDM 输入时，数据包的争用情况增多，故与单波长输入不同，网络规模的增大会带来更高的阻塞率，但当负载高于 0.7 时，因丢包率的急剧增加，大规模的网络反而具有较低的延迟。

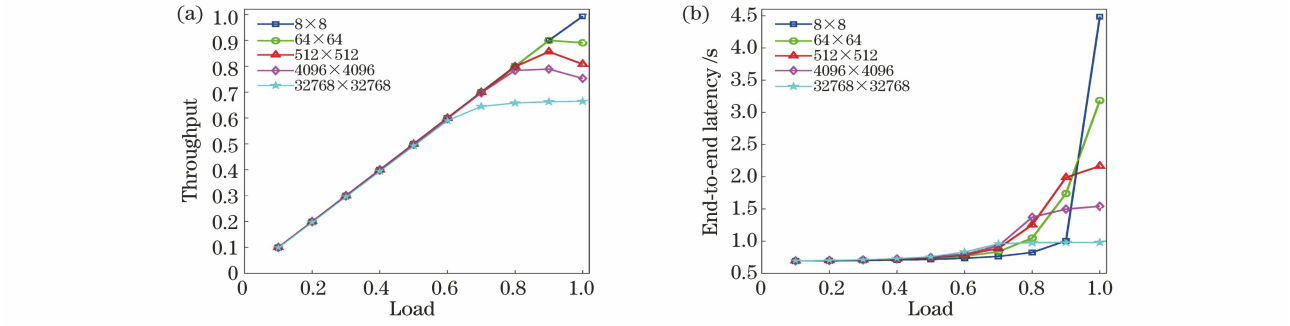

图 14　WDM 输入在不同网络大小下的仿真结果。(a)吞吐量；(b)延迟

Fig. 14　Simulation results of WDM input under different network sizes. (a) Throughput; (b) end-to-end latency

与单波长输入类似，本文探讨了在 WDM 输入下，FDL 分布对网络性能的影响，结果如图 15 所示。可以发现，FDL 的容量越小，吞吐量越低，即丢包的主要原因是阻塞率较高而缓冲容量不足。从延迟曲线可以看出，在负载低于 0.7 时，分组线性分布的性能最优，而当负载高于或等于 0.7 时，均等分布的丢包率过高，导致其延迟不断降低，线性分布的丢包率最低、延迟最高，而分组线性分布在两方面均具有较好的性能。

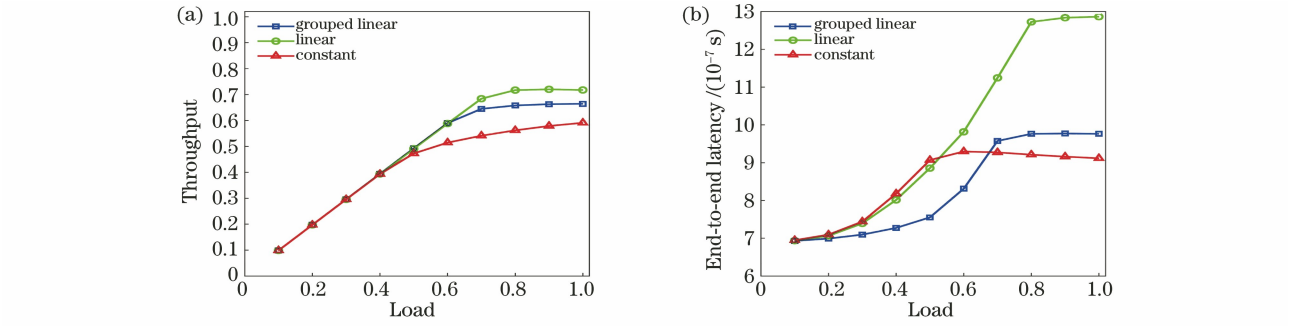

图 15　WDM 输入在不同 FDL 分布下的仿真结果。(a)吞吐量；(b)延迟

Fig. 15　Simulation results of WDM input under different FDL distributions. (a) Throughput; (b) end-to-end latency

图 16 所示为不同流量模式下的网络性能。对于位反转模式而言，预先设计的输入输出端口对避免了数据包间的争用，故所有数据包都被直接发送，没有产生延迟或丢包，其吞吐量与负载成线性，且延

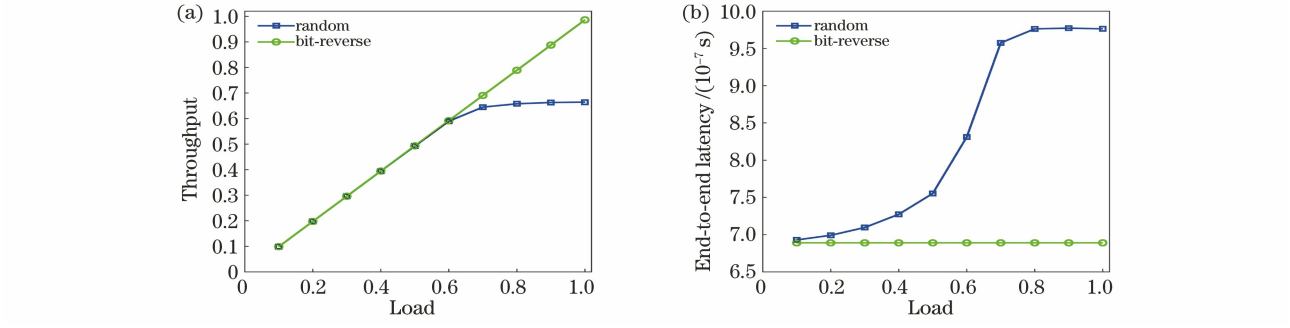

图 16　WDM 输入在不同 FDL 分布下的仿真结果。(a)吞吐量；(b)延迟

Fig. 16　Simulation results of WDM input under different FDL patterns. (a) Throughput; (b) end-to-end latency

迟与负载无关。随着负载的增大，随机模式的吞吐量越来越明显地低于位反转模式，这表明丢包率在逐渐增加。但是，受技术限制，很难增加 FDL 的个数，尽管从理论上讲，可以使用更多的 FDL 来实现接近 0 的丢包率，且随着阻塞率的增加，数据包需要在 FDL 上等待更长的时间以避免争用，但这会导致更大的延迟。

5 结 论

对大容量、高可靠性、低端到端延迟的全光网络进行了研究，提出了一种基于小基数 AWGR 的架构。考虑到网络规模，提出了两种实现方法。该架构在每个端口单波长输入的情况下，可以实现严格无阻塞网络。而每个端口在 WDM 输入情况下，可以互连更多的服务器，其中，每个 AWGR 都可以路由 M^2 个数据包，因此交换架构可以同时处理 M^3 个数据包。仿真结果表明，在高流量负载下，该架构在吞吐量和延迟方面仍具有良好的性能。如果进一步增加 AWGR 端口数量，网络规模将会更大。该架构为数据中心的扁平化做出了创新性的贡献，提高了现有交换架构的吞吐量，且具有高可扩展性。

参 考 文 献

[1] Tsakyridis A, Terzenidis N, Giamougiannis G, et al. 25.6 Tbps capacity and sub-μsec latency switching for DataCenters using ＞ 1000-port optical packet switch architectures[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2021, 27(2): 1-11.

[2] Zhao X Y, Lu L, Wu C X, et al. Ring fiber network based multipoint time-frequency dissemination method with high precision[J]. Acta Optica Sinica, 2019, 39(6): 0606002.
赵晓宇, 卢麟, 吴传信, 等. 基于光纤环形网的多点高精度时频传递方法[J]. 光学学报, 2019, 39(6): 0606002.

[3] Krishnamoorthy A V, Thacker H D, Torudbakken O, et al. From chip to cloud: optical interconnects in engineered systems [J]. Journal of Lightwave Technology, 2017, 35(15): 3103-3115.

[4] Zhang J H, Wu B J, Qiu K. Constrained link routing algorithm for dilated Benes optical switching chips under non-full configuration [J]. Laser & Optoelectronics Progress, 2019, 56(21): 211301.
张金花, 武保剑, 邱昆. 扩张型 Benes 光交换芯片未满配置情形下的约束链路路由算法[J]. 激光与光电子学进展, 2019, 56(21): 211301.

[5] Misra S, Mondal A, Khajjayam S. Dynamic big-data broadcast in fat-tree data center networks with mobile IoT devices[J]. IEEE Systems Journal, 2019, 13(3): 2898-2905.

[6] di Lucente S, Calabretta N, Resing J A C, et al. Scaling low-latency optical packet switches to a thousand ports[J]. IEEE/OSA Journal of Optical Communications and Networking, 2012, 4(9): A17-A28.

[7] Pitris S, Mitsolidou C, Moralis-Pegios M, et al. 400 Gb/s silicon photonic transmitter and routing WDM technologies for glueless 8-socket chip-to-chip interconnects[J]. Journal of Lightwave Technology, 2020, 38(13): 3366-3375.

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

[9] Ye X H, Yin Y W, Yoo S J B, et al. DOS: a scalable optical switch for datacenters [C] // Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems-ANCS'10, October 25-26, 2010, La Jolla, California. New York: ACM Press, 2010: 1-12.

[10] Xiao X, Proietti R, Liu G C, et al. Silicon photonic flex-LIONS for bandwidth-reconfigurable optical interconnects[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(2): 1-10.

[11] Lea C T. A scalable AWGR-based optical switch[J]. Journal of Lightwave Technology, 2015, 33(22): 4612-4621.

[12] Cheung S, Su T H, 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.

[13] Li Y Y, Gao Y Z, Li Z, et al. Characteristics of programmable optical fiber delay system[J]. Acta Optica Sinica, 2019, 39(8): 0806002.
李炎炎, 高彦泽, 李卓, 等. 可编程光纤延时系统特性[J]. 光学学报, 2019, 39(8): 0806002.

[14] Reza A G, Lim H. Throughput and delay performance analysis of feed-forward and feedback shared fiber delay line based hybrid buffering optical packet switch[C]//The International Conference on Information Networking 2011 (ICOIN2011), January 26-28, 2011, Kuala Lumpur, Malaysia. New York: IEEE Press, 2011: 414-418.

[15] Haas Z. The 'taggering switch': an electronically controlled optical packet switch [J]. Journal of Lightwave Technology, 1993, 11(5/6): 925-936.