doi:10.11887/j.cn.201605007 http://journal.nudt.edu.cn

光互连网络中资源预留的多级光互连仲裁机制*

翦 杰,赖明澈,肖立权
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:基于资源预留策略提出一种多级光互连仲裁机制,通过将网络分级实现快速、高效的仲裁。多优先级数据缓存队列的传输节点设计,提供了不同类型流量的差异化传输;通过预约式两级仲裁机制,实现网络的完全公平与100%的高吞吐率。设计并对快速仲裁通道进行了合理布局,极大地缩短了仲裁延迟。仿真结果表明:采用基于资源预留的分级仲裁策略,在多种流量模式下所有节点均获得公平的服务。与FeatherWeight相比,分级仲裁策略吞吐率提高17%;与2-pass相比,仲裁延迟减少15%,同时,功耗减少5%。

关键词:多级光互连;快速仲裁通道;资源预留
中图分类号:TN95 **文献标志码**:A **文章编号**:1001-2486(2016)05-039-06

Multi-level arbitration in optical network-on-chip based on resource reservation

JIAN Jie, *LAI Mingche*, *XIAO Liquan*
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: A multi-level arbitration based on resource reservation in optical network-on-chip was proposed. The fast and high efficiency arbitration scheme divided the network into multi-stages. With the design of multi priority queues in every node, arbiters provided differential transmissions for different kinds of flows. The arbiter also presented the transmit bund resource reservation scheme to fairly reserve time slots for all nodes, thus achieving the throughput of 100%. The fast arbitration channels were proposed and designed to degrade the arbitration period, and the packet transmitting delay was reduced. The simulation results show that, with the multi-level arbitration based on resource reservation, all nodes are allocated with almost equal service under various patterns. This scheme improves throughput by 17% when compared with FeatherWeight under the self-similar traffic pattern, and eliminats arbitration delay by 15% to 2-pass arbitration, incurring total power overhead by 5%.

Key words: multi-level optical network-on-chip; fast arbitration channel; resource reservation

近几年来的硅基光子集成技术日臻成熟,低延迟、高吞吐率、低功耗的片上硅光网络成为研究热点。面向片上光互连结构的大量研究[1-3]表明,片上光互连网络作为未来“众核”处理器内部互连实体的趋势已初现端倪。在以Crossbar交换结构为主的当今主流光互连网络中,仲裁策略关系到网络的各种性能,包括通道利用率、功耗、公平性等,因此一直是需要重点解决的问题,现有的仲裁策略[4-5]采用的单一仲裁器进行分布式仲裁,随着网络规模和性能的不断发展,已经不能获得较好的仲裁效果,在解决网络公平性、通道利用、可扩展性等方面存在固有缺陷。

受复杂网络学科中分级网络的启发[6],本文给出的光互连网络中的服务质量(Quality of Service, QoS)支持策略,引入分级仲裁的概念,将网络进行两级管理,将仲裁过程分散到两级仲裁器分别实现,减少了仲裁延迟,同时利用很小的硬件开销和逻辑复杂度,解决了因节点位置导致的不公平性问题,解决了远端饥饿矛盾;在仲裁过程中,充分考虑不同应用请求的特点,为不同类型和优先级的请求维护不同的请求队列,保证了差别化服务的QoS要求,通过节点与一级仲裁、一级仲裁与总仲裁之间的信息交互,保证了网络资源能够完全被节点利用,达到100%的吞吐率。

1 支持QoS的分级仲裁结构

1.1 QoS设计规则

随着片上光网络的发展,网络上运行的应用

* 收稿日期:2015-11-25
基金项目:国家863计划资助项目(2015AA015302);国家重点基础研究发展计划资助项目(2012CB933504);国家自然科学基金资助项目(61572509)
作者简介:翦杰(1987—),男,湖南常德人,博士研究生,E-mail:jianj06@mails.tsinghua.edu.cn;
肖立权(通信作者),男,研究员,博士,博士生导师,E-mail:xiaoliquan@nudt.edu.cn

种类和数量越来越多,在不同优先级的网络报文同时竞争网络带宽时,为片上光网络提供 QoS[7] 的服务保证,成为新阶段光网络仲裁算法必须解决的一个问题。

但是,针对光互连网络的现有仲裁策略没有充分确保网络有较好的网络性能(最小带宽保证、最大延迟保证、公平性等),针对片上光网络的具体要求,该仲裁结构在保证对 QoS 支持的同时,令网络达到最佳性能。QoS 的设计规则为:①公平性。当高优先级队列还有数据等待传输时,低优先级队列数据被阻塞。②最小带宽和最大延迟保证。任何节点不能长时间无服务,以保证最小带宽;任何队列的数据从请求仲裁到获得仲裁结果的延迟不能太大,即仲裁延迟不能太大。③位置独立性。不同节点不因其在网络布局中的位置不同而影响其获得服务的概率,以保证整个网络节点之间的公平性。

假设传输节点有 3 种不同优先级的数据,它们分别存储在 3 个不同优先级的队列 P1,P2,P3 中(优先级 P1 > P2 > P3)。为保证公平性,高优先级队列的请求获得满足之后,低优先级队列的请求才会被满足,这一点在仲裁器中实现;为保证节点的最小带宽(用 B_m 表示),仲裁控制器会监测整个节点获得的服务情况,当服务很少,使得节点带宽达到 B_m 时,仲裁控制器向操作系统发送中断请求,减少节点的数据产生;为保证数据的最大传输延迟(用 D 表示),节点维护一个特殊请求队列 P0,其优先级大于所有普通队列,仲裁控制器会记录每个队列头报文的等待延迟,如果达到 D,则将普通队列的头报文迁移到特殊队列 P0 中以保证在下一个仲裁周期,数据能立即发送出去。

1.2　多优先级 QoS 两级仲裁器结构

基于上述 QoS 的设计规则,为满足 QoS 的公平性、最大延迟保证、最小带宽保证的支持,设计了两级仲裁机制。通过一级仲裁器将整个网络分成多个子网,每个子网通过一个一级仲裁器进行仲裁,所有的子网信息集中到总仲裁器;通过总仲裁器控制一级仲裁器,仲裁机制的总结构如图 1 所示。

在节点内部,节点控制逻辑维护 4 个不同等级的等待队列(P1,P2,P3,P0),不同类型的数据分别进入不同的等待队列,其中特殊请求进入队列 P0。每个队列的实际请求情况用向量 $\boldsymbol{q}$ 表示,同时,总仲裁的仲裁结果会分发到每个节点,节点收到的仲裁结果用向量 $\boldsymbol{g}$ 表示。为提供 QoS 支持,节点控制逻辑除了分别维护每个队列的请求信息,还要维护提供最小带宽和最大延迟保证的特殊请求信息。最大延迟保证的特殊请求信息由低等级队列在长时间未获得服务时发出,节点收到此信息后,会将发送此特殊信息的队列头报文迁移到特殊队列 P0;最小带宽保证的特殊请求信息由节点所有队列长时间未获得服务时发出,此时,节点控制器向软件发送中断请求,请求软件减少流量请求。

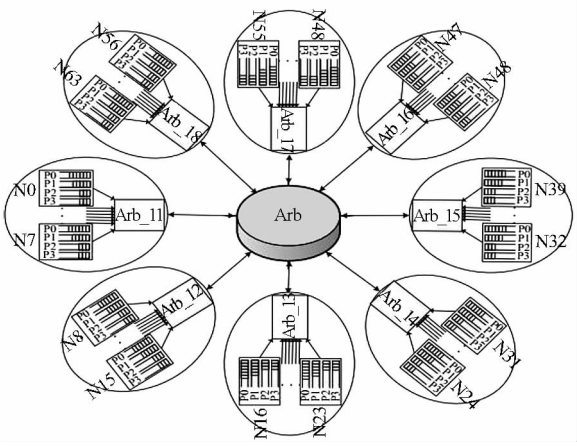


图 1　分级仲裁结构

Fig. 1　Architecture of hierarchical arbitration

节点在向仲裁器请求资源时,综合考虑每个队列的实时请求情况和收到的仲裁结果,无论是队列的实际请求 $\boldsymbol{q}$、收到的仲裁结果 $\boldsymbol{g}$,还是最终节点向一级仲裁器提供的请求信息 $\boldsymbol{r}$,均为一个四维向量。一级仲裁器管理由 8 个节点组成的子网,它收集 8 个节点发送过来的请求信息 $\boldsymbol{r}$,进行处理后,将请求信息发送到总仲裁器,并接收总仲裁器的仲裁结果,利用此结果对子网络进行仲裁,再将仲裁结果分发到每个节点。总仲裁器是通道资源的实际仲裁者,与 8 个一级仲裁器进行数据交互,根据 8 个一级仲裁器发送过来的请求信息,总仲裁器进行资源分配,并将分配结果分发给每个一级仲裁器。总仲裁器从 8 个一级仲裁器收到的是 8 个四维请求信息,总仲裁器进行仲裁时,特殊请求的优先级最高,接下来分别是 P1,P2,P3,资源分配的原则是先完全满足高优先级的请求,同一优先级的不同请求,按照 Max_Min 的原则进行分配。其仲裁的一般过程为:首先,利用 Max_Min 原则将资源分配给最高优先级的请求,如果资源还有多余,则以同样的原则分配给下一优先级的请求,直到所有资源被分配完。

2　资源预留的快速仲裁策略

2.1　仲裁策略设计

上述两级仲裁结构解决了 QoS 支持问题,但

基于上述思路实现的仲裁策略,存在吞吐率不高、仲裁延迟过长的问题。现提出一种基于资源预留的仲裁策略,根据网络的实时请求信息,针对性地仲裁预留资源,使得吞吐率能达到 100%;同时,简化仲裁信息量,提高仲裁信息的交换速度,减少仲裁延迟。高效的光网络仲裁策略一般采用预留[8]的方式进行通道资源的分配。仲裁策略首先利用仲裁通道向节点发送预约令牌,请求通道的节点获得令牌,即获得对令牌所代表资源的预约。节点从请求获得令牌到实际获得资源之间的时间间隔,称之为预留周期,它与网络的节点数目正相关。

随着网络规模的扩大,一级仲裁需要直接管理的节点数迅速增加,预留周期,即网络的仲裁延迟也迅速扩大。因此,提出针对光网络基于预留的两级仲裁策略,为减小单个仲裁器的复杂度以及整个网络的仲裁延迟,两级仲裁策略将网络的仲裁分散到多个不同的仲裁器上,从而极大缩短了仲裁延迟。仲裁策略所调度的通道使用权的最小单位,称之为传输槽(transmit slot),分级仲裁策略实际上就是对仲裁时刻后续多个传输槽的调度,称之为传输堆(transmit bund),其大小与预留周期一致。如图 2 所示,当前传输堆资源的仲裁结果,会用于下一个仲裁周期的资源调度。每个传输堆的仲裁过程分为 7 个阶段:①所有节点将自身的队列请求信息发往一级仲裁器;②一级仲裁器将收到的请求信息进行综合,得到子网请求信息;③一级仲裁器将子网请求信息发往总仲裁器;④总仲裁器进行仲裁;⑤总仲裁的仲裁结果分发到一级仲裁器;⑥一级仲裁器进行仲裁;⑦一级仲裁器的仲裁结果分发到每个节点,节点利用此结果调度下一个传输堆的数据传输。

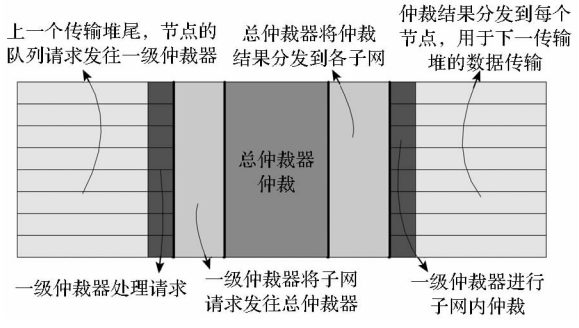


图 2　基于资源预留的两级仲裁

Fig. 2　Two level based on resource reservation

以实际网络状况为例,说明分级仲裁机制的工作原理。为简化描述,设定网络只含 2 个一级仲裁器,每个一级仲裁器管理的子网包含 3 个节点,每个传输堆由 8 个传输槽组成。时空图如图 3 所示。矩形内的数字分别代表 4 个队列的请求量或令牌量。在每个传输堆的第 1 传输槽,节点根据每个队列的等待信息 $\boldsymbol{q}$,结合从一级仲裁器得到的上一传输堆的仲裁结果 $\boldsymbol{g}$,计算出本传输堆向一级仲裁器发送的请求信息 $\boldsymbol{r}(\boldsymbol{r}=\boldsymbol{q}-\boldsymbol{g})$;子网内所有节点将各自的 $\boldsymbol{r}$ 值发送给一级仲裁耗时 1 个传输槽,因此,在第 2 传输槽,一级仲裁器进入 up 状态,它对收集到的所有请求信息进行处理,得到子网请求信息(1,1,3,3) 和 (2,3,3,4);在第 3 传输槽,子网请求信息由一级仲裁器传向总仲裁器,总仲裁器获得所有的请求信息;在第 4 传输槽,总仲裁器将传输堆内的资源,以传输槽为单位,按照优先级顺序,依次进行 Max_Min 仲裁;在第 5 传输槽,总仲裁器将仲裁结果(1,1,1,0) 和 (2,3,0,0) 分别传输至 2 个一级仲裁器;在第 6 传输槽,一级仲裁器利用总仲裁器的仲裁结果,用与总仲裁器一样的原则将获得的时间槽资源仲裁给每个节点;在第 7 传输槽,节点获得了仲裁结果,并按照仲裁结果,获得对下一个传输堆内传输槽的预约;在第 8 传输槽,节点统计每个请求队列的长度,为下一个传输堆资源的预约仲裁做准备。后续仲裁将重复上述过程,而第 9 ~ 16 传输槽的数据传输利用第 6 传输槽得到的仲裁结果。

图 3　仲裁时空图

Fig. 3　Time diagram of arbitration scheme

观察上述过程发现,在两级仲裁过程中,在每个传输堆的最开始时刻,节点处的队列等待信息向仲裁器发送之前,来自总仲裁器的当前传输堆的仲裁结果在上一传输堆末(第 6 传输槽)就已经分发到了每个节点处。因此,节点向仲裁器发送的请求是将实际请求信息减去仲裁结果后得到 $(\boldsymbol{r}=\boldsymbol{q}-\boldsymbol{g})$,这就保证了在下一传输堆开始后,实际所有队列的长度不会小于请求资源数量,因此,也保证了仲裁器仲裁给每个队列的资源最终肯定能够被完全利用,即保证了网络的吞吐率能够达到 100%。

2.2　快速仲裁通道布局

两级仲裁的主要优势在于能快速收集节点的

请求信息,经过仲裁之后又能快速将仲裁结果分发到每个节点。基于通用的 Corona[3] 光互连结构,设计了一种能够实现上述目标的快速仲裁通道(Fast Arbitration Channel,FAC),其物理布局如图 4 所示。

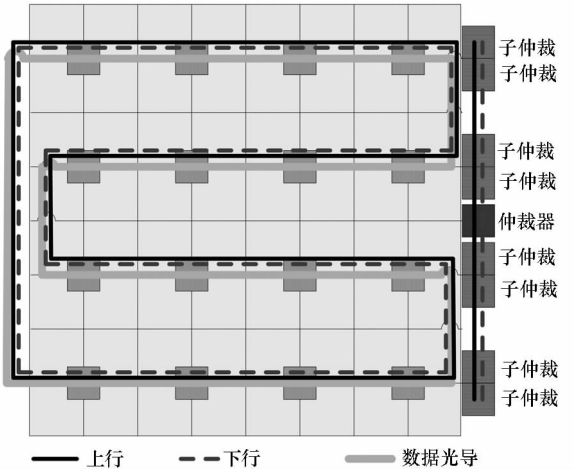


图 4 快速仲裁通道布局

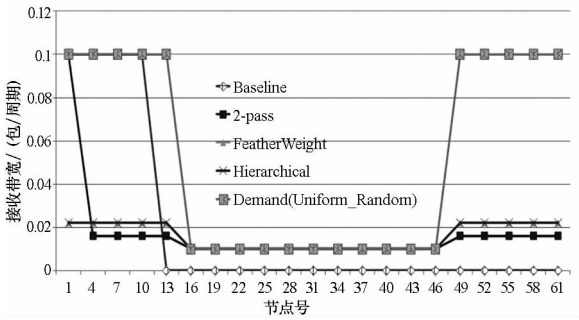
Fig. 4 Layout of FAC

快速仲裁通道充分考虑 Corona 结构的特点,利用两类光导快速收集和分发仲裁数据:S 型光导与数据光导平行,而垂直光导则用于连接所有仲裁器。两类光导都包含上行和下行 2 组,每个节点的请求信息由 40 bit 组成,其中包含 4 个优先级队列的目的地址(8 bit)和请求量(2 bit)。FAC 包含 8 条 40 波段的波分多路复用 S 型光导(4 条上行和 4 条下行),每个 S 型光导花费 4 传输周期发送请求信息至一级仲裁器,当一级仲裁器收到所有的请求信息并进行计算之后,它将子网请求信息通过垂直光导发送至总仲裁器,快速仲裁通道利用包含的 8 条 64 波段的垂直光导在 2 传输周期内完成子网请求信息的收集和总仲裁器仲裁结果的分发。整个仲裁过程需要 16 个传输周期。

3 仿真结果

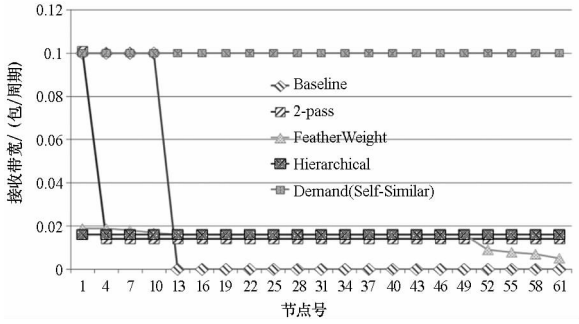
采用哥伦比亚大学的 PhoenixSim 模拟器模拟多写单读(Multi-Write-Single-Read, MWSR)光 Crossbar 分级仲裁网络,实现四种仲裁机制:Baseline (Corona)[3], FeatherWeight[9], 2-pass token stream[4] 以及 Hierarchical arbitration。两种流量模型分别是均匀流量和自相似流量,用以模拟真实网络流量。目的节点分为自然随机和热点两种。系统频率为 5 GHz,网络规模为 64 节点,数据包大小统一为 64 Bytes,队列深度大小为 8,采用随机分配多队列(Dynamic Allocate Multi Queue, DAMQ)技术,光链路传输率为 10 Gbps/link。光器件的物理参数来自文献[10]。

图 5 展示了在不同流量模型下,仲裁策略在公平性方面的仿真结果。图 5(a)和图 5(b)分别展示了节点 0 在通道负载饱和情形下,每个节点获得的带宽。对于源节点的注入率设置,两图有差异:图 5(a)在均匀流量模型下,一半节点(节点 16 ~48)具有低注入率 0.01,其余一半具有高注入率 0.1。图 5(b)在自相似流量模型下,所有节点均具有高注入率 0.1。仿真结果表明,在两种情形下,Baseline 只有开始的几个节点获得了服务,其余节点均饿死,2-pass 策略第一轮里给每个节点分配资源,在第二轮则收集第一轮未被使用的资源,按位置顺序分配给有需求的节点,没有完全解决公平问题,开始的节点获得了较多服务。FeatherWeight 通过历史信息和当前请求来节制高请求节点的流量。因此,其高请求节点获得流量大致相等,而低请求节点基本可以被满足。但是,当网络流量为自相似的突发流量时,此算法对流量变化反应不及时,导致不公平性依然存在。Hierarchical arbitration 利用集中式预留策略,在所有情形下,都获得了完全公平。



(a) 均匀流量

(a) Uniform random



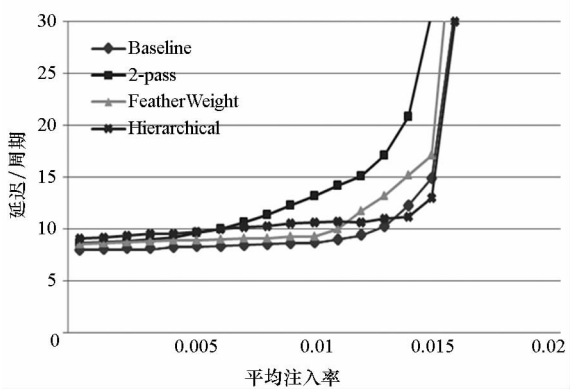
(b) 自相似

(b) Self-similar

图 5 不同流量模式下的带宽

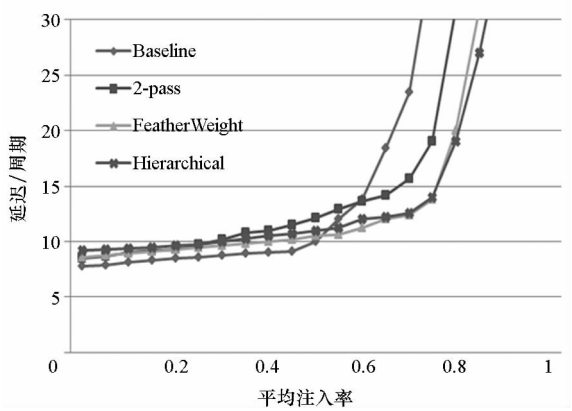
Fig. 5 Achieved bandwidth under

通过模拟 4 种仲裁策略在不同的流量模型和源节点与目的节点选择方式情形下网络获得的吞吐率和平均报文延迟来衡量仲裁策略的性能。在图 6 中,(a)、(b)两图给出了在均匀流量下的吞吐率,图 6(a)显示,在热点模式下,所有节点发送报文至目的节点 0,所有的仲裁策略获得1/N的吞吐率;图 6(b)显示,在自然随机选择策略下,所有的仲裁策略获得 100% 的吞吐率。分级仲裁策略在延迟上较 2-pass 有 15% 的提高。图 6(c)、图 6(d)显示在自相似流量模型下,2-pass 和 Baseline 在两种源目的节点选择机制下都能获得 100% 的吞吐率,说明基于时间片的仲裁机制能获得理想化的吞吐率。分级仲裁根据实时请求进行集中仲裁,因此也能获得 100% 的吞吐率。而 FeatherWeight 损失了约 20% 的吞吐率,这是因为 FeatherWeight 不能及时根据流量变化调整资源分配,存在带宽浪费。图 6(e)、图 6(f)展示了分级仲裁网络和数据网络的功耗。从图 6(e)中可以发现,相比于数据网络,仲裁网络的功耗几乎可忽略不计,这是由于与数据网络相比,仲裁网络的光导数目少,光导长度短。同时,如图 6(f)所示,与其他几种仲裁策略相比,分级仲裁的功耗很小,比 2-pass 仲裁功耗减少了 80% 。



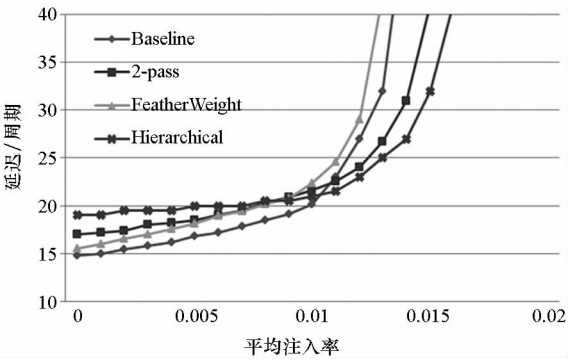
(a) 热点(均匀流量)

(a) Hotspot (uniform random)



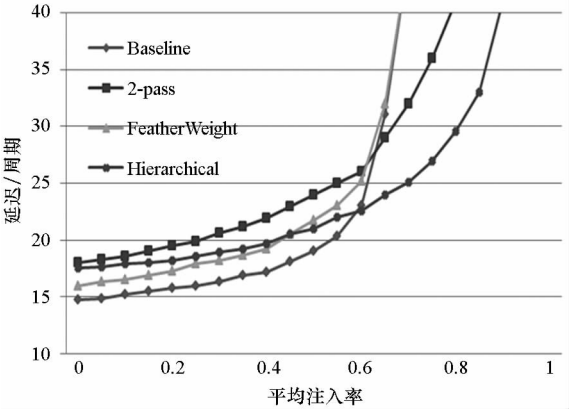
(b) 自然随机(均匀流量)

(b) Uniform(uniform random)



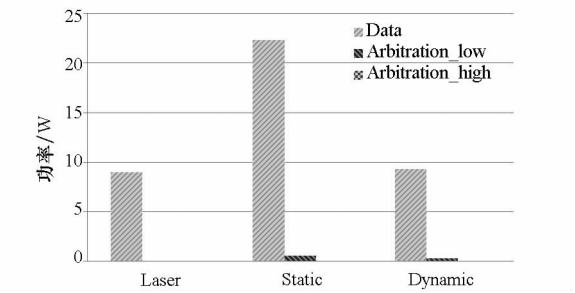
(c) 热点(自相似)

(c) Hotspot(self-similar)



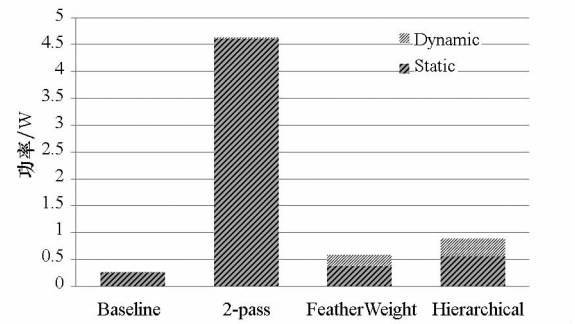
(d) 自然随机(自相似)

(d) Uniform(self-similar)



(e) Hierarchical 与数据网络对比

(e) Hierarchical vs. data network



(f) Hierarchical 与其他仲裁对比

(f) Hierarchical vs. other arbitrations

图 6　不同注入流量模型下的吞吐率和功耗

Fig. 6　Throughput and power consumption under different traffic patterns

4 结论

本文提出了一种分级快速光互连仲裁机制。通过具有多优先级数据缓存队列的传输节点设计,实现了数据传输差异化服务;利用请求驱动的预约式两级仲裁机制,实现网络的完全公平,和100%吞吐率;对设计的快速仲裁通道进行了合理布局,使得仲裁延迟缩短了15%,整个仲裁机制的功耗只占总功耗的5%。

参考文献(References)

[1] Calhoun D, Wen K, Zhu X, et al. Dynamic reconfiguration of silicon photonic circuit switched interconnection networks[C]// Proceedings of IEEE High Performance Extreme Computing Conference, 2014.

[2] Morris R W, Kodi A K, Louri A, et al. Three-dimensional stacked nanophotonic network-on-chip architecture with minimal reconfiguration [J]. IEEE Transactions on Computers, 2014, 63(1): 243 – 255.

[3] Vantrease D, Schreiber R, Monchiero M, et al. Corona: system implications of emerging nanophotonic technology [C]// Proceedings of the 35th Annual International Symposium on Computer Architecture, 2008: 153 – 164.

[4] Yan P, Kim J, Memik G. FlexiShare: channel sharing for an energy-efficient nanophotonic crossbar [C]//Proceedings of IEEE International Symposium on High Performance Computer Architecture, 2010: 1 – 12.

[5] Vantrease D, Binkert N, Schreiber R, et al. Light speed arbitration and flow control for nanophotonic interconnects[C]// Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture, 2009: 304 – 315.

[6] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98 – 101.

[7] Lee J W, Ng M C, Asanovi K. Globally synchronized frames for guaranteed quality-of-service in on-chip networks [J]. Journal of Parallel & Distributed Computing, 2012, 72(11): 1401 – 1411.

[8] Peh L S, Dally W J. Flit-reservation flow control [C]// Proceedings of International Symposium on High-Performance Computer Architecture, 2000: 73 – 84.

[9] Pan Y, Kim J, Memik G. FeatherWeight: low-cost optical arbitration with QoS support[C]// Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture, 2011: 105 – 116.

[10] Vantrease D M. Optical tokens in many-core processors[D]. USA: University of Wisconsin-Madison, 2010.