文章编号: 1001- 2486(2008) 06- 0068- 05

一种无缓冲的光互连网络的吞吐率性能分析及优化*

齐星云, 窦 强, 陈永然, 温 俊, 窦文华
(国防科技大学 计算机学院, 湖南 长沙 410073)

摘 要: 针对当前高性能计算机光互连网络中存在的光缓冲不易实现的问题, 提出了一种无缓冲的光互连网络结构 BOIN, 并在对网络结构进行建模和分析的基础上, 研究了网络的吞吐率随不同的输入负载和网络规模而变化的规律, 给出了在一定的互连总规模和输入负载下, 网络实际吞吐率达到最大值时网络拓扑结构所必须满足的条件。最后用模拟实验证明了这一结果的正确性。

关键词: 高性能计算机系统; 处理器间互连; 无缓冲光互连; 网络吞吐率; 网络结构优化
中图分类号: TP303 文献标识码: A

Analysis and Optimization of the Throughput of a Kind of Bufferless Optical Interconnection Network

QI Xing-yun, DOU Qiang, CHEN Yong-ran, WEN Jun, DOU Wen-hua
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Aiming at the problem of implementing optical buffers in the present high performance computer optical interconnection, a bufferless optical interconnection, named BOIN, is put forward. Based on modeling and analyzing the network topology, the relationship between the network throughput and the offered loads as well as network sizes is studied. Then the condition of maximum throughput of the network in a fixed offered loads and total number of nodes is presented. Finally, the simulation results show that the correctness is confirmed.

Key words: high performance computer system; inter-processor interconnection; bufferless optical interconnection network; network throughput; optimization of network topology

由于传统的电互连技术存在着带宽小、速度低、抗干扰能力差等不足, 在高性能计算机系统中, 采用光互连技术实现各处理单元之间的高速互连已经成为一种趋势[1-3]。然而, 在当前技术条件下, 要实现系统的全光互连, 还存在两大技术障碍: 一是缺乏有效的光缓冲技术; 二是不能有效地直接对光数据信号进行逻辑处理[4]。在通常情况下, 需要在每个处理单元入口将到达的光数据信号转换为电信号, 然后对这些电信号进行逻辑处理和报文缓存。在处理单元的出口又需要将电数据信号重新转换为光信号并送至光链路上。一个数据报文在传输过程中需要重复进行光电/电光的转换, 这在对于互连网络性能要求极高的高性能计算机中, 造成了很大的性能损失。

针对这种问题, 本文提出了一种不需要对光数据信号进行光电转换并缓存的光互连结构 BOIN (Bufferless Optical Interconnection Network), 并在此基础上研究了如何进一步提高网络吞吐率。

1 BOIN 光互连网络结构

一个规模为 $m \times n$ 的 BOIN 网络[6] 由 $m \times n$ 个网络节点组成, 通过相互间的光链路组成一个 $m \times n$ 的单向 Torus 网络结构。每个网络节点 $N(x, y)$ ($0 \leqslant x \leqslant m-1$, $0 \leqslant y \leqslant n-1$) 由 3 个 2×2 的光开关及 1 个处理单元组成。3 个 2×2 的光开关根据在节点内部所处的位置和功能的不同, 分别记为 $S(x, y, X)$, $S(x, y, Y)$ 和 $S(x, y, F)$, 如图 1 所示。$S(x, y, F)$ 负责将从节点的 $X-$ 和 $Y-$ 端口到达的数据转发到

* 收稿日期: 2008- 05- 22
基金项目: 国家自然科学基金资助项目(60633050, 60603061)
作者简介: 齐星云(1979—), 男, 博士生。

节点的 $X+$ 或 $Y+$ 端口; $S(x, y, X)$ 负责将从 $S(x, y, F)$ 和处理单元 $P(x, y)$ 到来的数据报文送至节点的 $X+$ 端口; $S(x, y, Y)$ 负责将从 $S(x, y, F)$ 到来的数据送至节点的 $Y+$ 端口或处理单元 $P(x, y)$。每个光开关都具有两个状态: 直通(T)和交叉(C)。当处于直通状态时, 从 $X-$ 到达的信号送达 $X+$, 从 $Y-$ 到达的信号送达 $Y+$; 当处于交叉状态时, 从 $X-$ 到达的信号送达 $Y+$, 从 $Y-$ 到达的信号送达 $X+$。

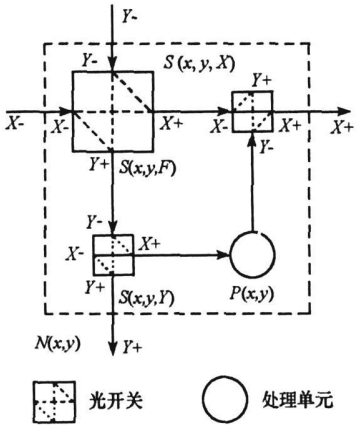

图 1 BOIN 网络节点

Fig. 1 The architecture of a network node

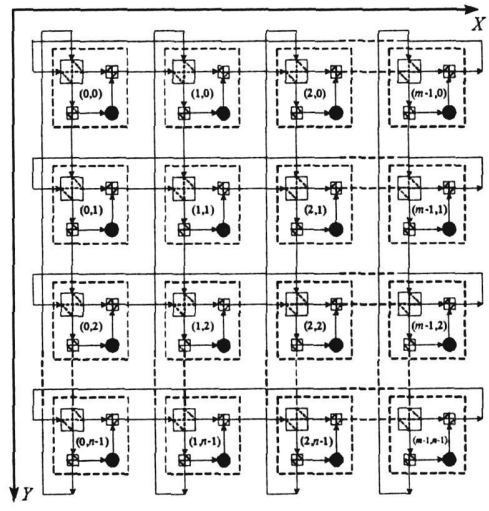

图 2 $m \times n$ 的 BOIN 网络结构

Fig. 2 An $m \times n$ BOIN network

mn 个网络节点在 X 方向和 Y 方向分别互相连接, 组成一个 $m \times n$ 的 BOIN 网络, 如图 2 所示。在 BOIN 网络中, 任何相邻两个节点之间的光链路都是等长的, 设其长度为 L。同时设光数据信号在链路上传输的带宽为 B, 前进速率为 c。同时令网络中传输的每个数据报文的长度为 $D = BL/c$, 这样, 每个数据报文从一个节点上发出的时间 D/B 恰好等于光信号在 L 长的链路上的传输时间 L/c。令 $\Delta T = D/B = L/c$, 称 ΔT 为 BOIN 网络中的一个单位时间片。网络中的所有数据传输操作都与单位时间片保持同步, 在一个 ΔT 的开始时刻开始发送一个数据报文, 在下一个 ΔT 开始时刻完成发送, 并开始发下一个报文。一个数据报文在一个 ΔT 的开始时刻由某个节点发出, 经过 ΔT 时间, 到下一个 ΔT 的开始时刻, 该报文的最后一位刚好发送完成, 同时第一位已经经过长度为 L 的链路, 到达下一个节点入口。这样, 整个报文相当于缓存在网络链路上, 不需要在每个节点处对数据进行缓存, 减少了光电/电光转换延时和缓冲延时。

BOIN 网络中, 在相邻两个节点之间, 除了有高速光链路进行数据传输外, 还有电链路, 用来在相邻节点之间进行控制报文的传送。一个数据报文在被从处理单元 $P(x, y)$ 发出的同时, 将通过电控制信号链路向下一个节点 $N((x+1) \bmod m, y)$ 发出一个控制信号, 告知其即将到来的光数据报文的目的地址。由于需要在光链路上缓冲一个数据报文, 因此光链路必须具有一定的长度($L = c \cdot \Delta T$), 而电控制链路无此要求, 可以采用很短的节点间链路, 使得电控制信号可以在光数据信号到达下一个节点之前提前到达。下一个节点可以根据控制信号来进行本节点内部各光开关的状态转换, 在光信号到达之前为其建立起下一跳的传输链路, 使报文不经过缓存就可以直接送往下一个节点。在 BOIN 网络中采用 XY 路由策略, 即任何一个报文先进行 X 方向的路由, 然后进行 Y 方向的路由。如果同时有报文分别从 $X-$ 和 $Y-$ 方向到达并要送往同一个输出端口, 则从 $X-$ 方向到达的报文具有较高的优先权。竞争失败的报文采用偏折路由[5] 技术在网络中绕道前进, 直至最终到达目的节点。如果处理单元 $P(x, y)$ 需要把一个报文发送到网络上, 此时正好有数据要通过 $S(x, y, X)$ 并到达其 $X+$ 端口, 则会在 $S(x, y, X)$ 上发生冲突, 此时, 应暂停 $P(x, y)$ 上数据的发送, 而优先允许 $S(x, y, X)$ 上的报文通过。

2 网络性能模型及吞吐率性能分析

在高性能计算机内部的高速互连网络中,吞吐率的高低对整个网络系统的性能具有重要影响。网络吞吐率已经成为衡量互连网络性能的一个重要指标。本节根据BOIN网络的性能模型对其吞吐率进行分析,并用来指导网络系统的优化设计。

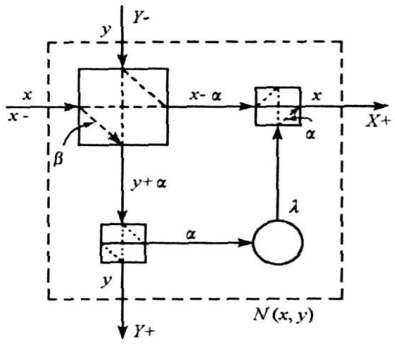

图3 网络节点中的参数

Fig. 3 Parameters in the network node

在 $m \times n$ 的网络中,对于任何一个节点 N,设到达平衡状态后,每个时间片内平均有 x 个报文从 $X-$ 到达节点 N,有 y 个报文从 $Y-$ 到达节点 N,其中 $0< x< 1, 0< y< 1$。设节点 N 上的处理器每个时间片内需要发出 λ 个报文,而实际上由于数据冲突的存在,只能发出 α 个报文($0< \alpha< \lambda< 1$)。x 和 y 分别为每个时间片内有数据报文从 $X-$ 和 $Y-$ 到达节点 N 的概率,λ 和 α 分别为每个时间片内节点 N 上的处理器需要发送一个数据报文和能够发送成功一个数据报文的概率。可知,α 即为网络的实际吞吐率。由于网络处于平衡状态,故从每个节点的 $X+$ 和 $Y+$ 送出的报文也分别为 x 和 y,每个处理单元接收的报文为 α。同时,设一个时间片内在某个节点处从 $X-$ 转向 $Y+$ 的报文有 β 个($0< \beta< 1$)。如图3所示。

根据文献[6]中的讨论结果可知,

$$\begin{cases} 1-\dfrac{\alpha}{\lambda}=x-\alpha \\ \beta-\alpha=y\beta \\ x-\dfrac{m(mn+n-2)}{2(mn-1)}\alpha=\dfrac{m+1}{2}y\beta \\ \dfrac{(m-1)\alpha}{mn-1}+\dfrac{2}{n}y=\alpha \end{cases} \tag{1}$$

可得

$$\alpha=\frac{(mn-1)[\theta-(mn^2+m^2n-2mn-2m+2)\lambda-2(mn-1)]}{mn(n-1)[(2mn+m-3)\lambda-2mn+2]} \tag{2}$$

其中,

$$\theta=\sqrt{[(mn^2-m^2n+2m-2)\lambda-2mn+2]^2+4mn(m+1)(n-1)(mn-1)\lambda^2} \tag{3}$$

根据式(2),图4(a)给出了当 $m=n=4, 8, 16, 32$ 时,网络的实际吞吐率 α 与网络的输入负载 λ 的关系。从图中可以看出,随着输入负载的增大,网络的实际吞吐率也随之增大。同时从图中可以发现,随着网络规模的增大,网络实际吞吐率逐渐降低。这是因为当网络规模增大时,每个报文在到达目的节点之前需要经过更多的中间节点,这就导致了网络中报文冲突的增加,从而使得网络的实际吞吐率降低。由前述可知,BOIN网络中一个报文能被处理单元成功发送到网络链路上的概率为 α/λ,$\alpha/\lambda\sim\lambda$ 反映了在不同的 λ 下,一个报文能被成功发出的概率。图4(b)给出了在不同的网络规模下 $\alpha/\lambda\sim\lambda$ 之间

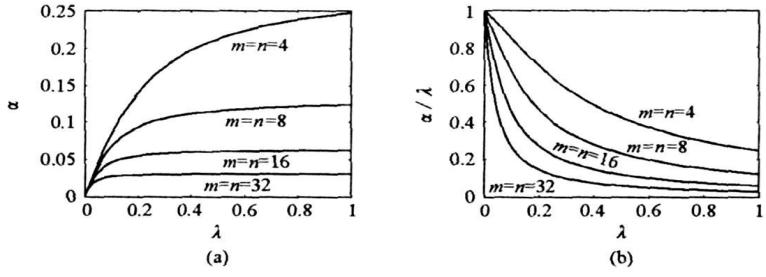

图4 不同规模下网络的吞吐率与输入负载的关系

Fig. 4 The relation between the throughput and the offered load in different network size

的关系。由图中可以看出, 随着网络输入负载或者网络规模的增大, 报文能被成功发送出去的概率逐渐减小。

图 5 给出了在 $m \times n = 64$ 的条件下, 当 m 和 λ 取不同值时, 网络的实际吞吐率 α 的变化趋势。从图中可以看出, 当给定网络的总规模 N 时(即 $m \times n = N$ 为定值时), 对于任何 λ, 都存在一个确定的 m 及与之对应的 $n(n = N/m)$, 使得网络的实际吞吐率 α 达到最大。下面研究对于一个确定的网络规模 N, 如何设计网络的结构, 选择合适的 m 和 n, 使得网络实际吞吐率最大。

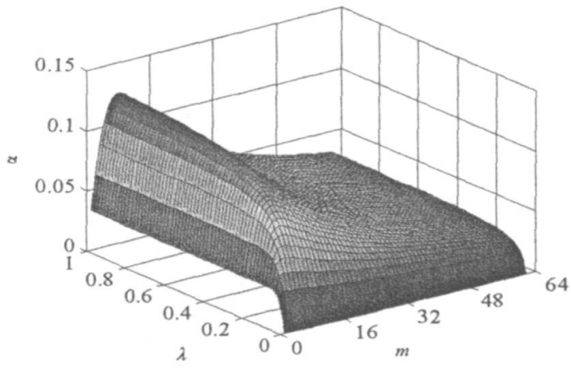

图 5 当 $m \times n = 64$ 时网络的吞吐率 α 与 m 及 λ 的关系

Fig. 5 The relation between α and m and λ when $m \times n = 64$

令 $\frac{d\left(\alpha\big|_{n=N/m}\right)}{dm} = 0$, 并考虑到在一般的高性能计算机互连网络中, $m = 2^k$ ($k = 1, 2, 3, \cdots$), 可得出当 α 取极大值时 m 对应的值。表 1 列出了当 $N = 16, 32, 64, 128, 256, 512, 1024$ 和 $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$ 时, 网络实际吞吐率 α 取最大值时 m 和对应的 n 的取值, 以及此时对应的网络最大实际吞吐率。

表 1 单位节点平均吞吐率的最大值及对应的 (m, n) 值

Tab. 1 The maximum value of the average throughput and the corresponding (m, n) of a node

λ \ N	16		32		64		128		256		512		1024	
	α	m, n	α	m, n	α	m, n	α	m, n	α	m, n	α	m, n	α	m, n
0.1	0.0892	2, 8	0.0802	4, 8	0.0698	4, 16	0.0586	8, 16	0.0456	16, 16	0.0357	16, 32	0.0268	32, 32
0.2	0.1404	2, 8	0.1197	4, 8	0.0936	8, 8	0.0721	8, 16	0.0544	16, 16	0.0391	16, 32	0.0292	32, 32
0.3	0.1752	4, 4	0.1377	4, 8	0.1055	8, 8	0.0768	8, 16	0.0577	16, 16	0.0402	16, 32	0.0301	32, 32
0.4	0.1977	4, 4	0.1472	4, 8	0.1121	8, 8	0.0790	8, 16	0.0594	16, 16	0.0407	16, 32	0.0305	32, 32
0.5	0.2131	4, 4	0.1529	4, 8	0.1161	8, 8	0.0804	8, 16	0.0604	16, 16	0.0410	16, 32	0.0307	32, 32
0.6	0.2240	4, 4	0.1567	4, 8	0.1189	8, 8	0.0813	8, 16	0.0611	16, 16	0.0412	16, 32	0.0309	32, 32
0.7	0.2322	4, 4	0.1594	4, 8	0.1209	8, 8	0.0819	8, 16	0.0616	16, 16	0.0414	16, 32	0.0310	32, 32
0.8	0.2385	4, 4	0.1626	8, 4	0.1225	8, 8	0.0824	8, 16	0.0619	16, 16	0.0415	16, 32	0.0311	32, 32
0.9	0.2435	4, 4	0.1654	8, 4	0.1237	8, 8	0.0829	16, 8	0.0622	16, 16	0.0416	16, 32	0.0312	32, 32
1.0	0.2475	4, 4	0.1677	8, 4	0.1246	8, 8	0.0834	16, 8	0.0624	16, 16	0.0417	32, 16	0.0312	32, 32

3 模拟验证

为了验证以上结果, 我们采用 OMNet++ [7] 网络性能模拟环境对 BOIN 网络的吞吐率特性进行了模拟。在实验中, 我们分别模拟了在不同的网络总规模 N 下, 当网络的横向规模 m 和纵向规模 n 取不同值时, 网络的实际吞吐率 α 随网络的输入负载 λ 的变化趋势。实验参数在表 2 中列出, 实验结果如图 6 所示。在实验中, 对于网络负载 $\lambda (0 < \lambda < 1)$, 相邻报文间隔时间 T (单位为时间片) 服从参数为 λ 的几何分布, 即 $\Pr(T = k) = \lambda(1 - \lambda)^{k-1} (k = 1, 2, 3, \cdots)$。

表 2 实验参数设置

Tab. 2 The parameters in the simulations

参数	参数值
模拟时间	100 000 时隙
网络规模	32, 64, 128, 256, 512, 1024
光链路延时	25.6ns
光链路带宽	10Gb/s
电链路延时	6.4ns
电链路带宽	200MHz× 32b
数据报文长度	256b

图 6(a) ~ (f) 分别给出了不同的网络总规模下, 在不同的配置结构时网络的实际吞吐率。

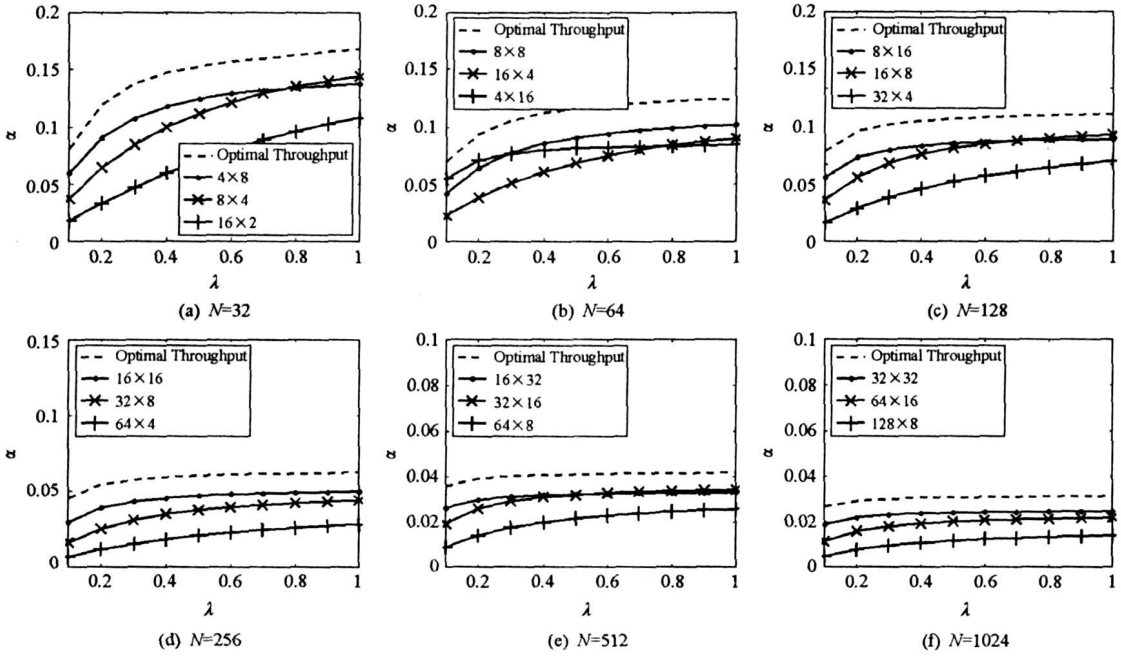

图 6 不同的配置结构下网络的实际吞吐率与输入负载的关系

Fig. 6 The results of the relation between throughput and offered load in different network size

图中虚线为理论上的最优吞吐率, 实线为在不同的网络配置结构下实验模拟得到的实际吞吐率曲线。从图中可以看出, 当网络总规模 N 一定时, 随着输入负载的增大, 网络的实际吞吐率也逐渐增大; 但网络的实际吞吐率随着网络总规模 N 的增大而逐渐减小。同时可以看出, 在 N 取不同的值时, 网络的最大实际吞吐率及对应的(m, n) 值与表 1 相吻合。这证明了模型分析的正确性, 同时也说明了实验是可信的。

4 结 论

本文基于一种无缓冲的高性能计算机光互连网络 BOIN, 在理论上分析了网络的性能模型, 研究了网络的吞吐率随网络负载和网络规模的变化情况, 给出了在一定的网络输入负载和网络总规模下达到网络最大吞吐率所需要满足的条件。最后通过模拟实验验证了这一结论。这一结论表明, 当需要将一定数目的处理单元通过 BOIN 网络连接起来组成高速互连网络时, 存在一个最优的网络结构, 在这个结构下, 网络具有最高的吞吐率。

参 考 文 献:

[1] Luijten R, Minkenberg C, Hemenway R, et al. Viable Opto-electronic HPC Interconnect Fabrics[C]//ACM/ IEEE Super Computing Conference, 2005.

[2] Kodi A K, Louri A. Design of a High-speed Optical Interconnect for Scalable Shared-memory Multiprocessors [J]. IEEE Micro, 2005, 25(1): 41 – 49.

[3] Hawkins C, Small B A, Wills D S, et al. The Data Vortex, an All Optical Path Multicomputer Interconnection Network [J]. IEEE Transactions on Parallel and Distributed Systems, 2007, 18(3): 409– 420.

[4] Papadimitriou G I, Papazoglou C, Pomportsis A S. Optical Switching: Switch Fabrics, Techniques, and Architectures [J]. Journal of Lightwave Technology, 2003, 21(2).

[5] Greenberg A G, Hajek B. Deflection Routing in Hypercube Networks[J]. IEEE Transactions on Communications, 1992, COM– 40(6).

[6] 齐星云. 新型高性能计算机光互连网络关键技术研究[D]. 长沙: 国防科技大学, 2008.

[7] OMNet + + [EB/OL]. http://www.omnetpp.org/.