文章编号：0258-7025(2001)10-0932-05

并行计算机系统的光 WDM 互连链路*

党明瑞 周明拓 毛幼菊
(重庆邮电学院光通信研究所 重庆 400065)

提要 光波分复用(WDM)互连接是并行计算机系统克服‘电子瓶颈’的可行方案，光纤传输的特性可使大容量、低时延、低误码率传输的并行计算机互连成为现实。探讨了并行计算机光 WDM 互连的一种结构，分析研究了其中的关键技术，并设计了一个实验系统。分析和实验表明采用光 WDM 互连可以实现超大容量的并行计算机系统。

关键词 波分复用，光互连链路，并行计算机系统
中图分类号 TP 393 文献标识码 A

WDM-optical Interconnection Link for Parallel Computer System

DANG Ming-rui ZHOU Ming-tuo MAO You-ju
(*Optical Communication Institute*, *Chongqing University of Posts & Telecommunications*, *Chongqing* 400065)

Abstract WDM-optical interconnection is a feasible method overcoming the problem of electronic bottlenecks for parallel computer system, the transmission properties of fiber make high capacity, low-latency, low error rate parallel computer interconnection realistic. This paper discussed one kind of structures for parallel computer WDM-optical interconnection, researched key technologies of it, and designed a test system. Analysis and test show that parallel computer system with ultrahigh capacity can be achieved by WDM-optical interconnection.

Key words WDM, optical interconnection link, parallel computer system

1 引 言

分布并行计算系统是实现超高速和超大容量计算机的关键部件之一，研究并行计算系统对发展超高速超大容量计算机有着重大的意义。美国 IBM 公司最近报道研制成功了峰值运算速度达 12.3×10^{12}次/s 的计算机；与此同时，我国也报道了国家并行计算机研究中心研制成功了超高速高性能的计算机。

在并行计算机中，有多个中央处理单元(CPU)和多个存储器(MEM)。中央处理单元和存储器之间要交换数据，因此需要一个通信网络在它们之间构成数据交换与传输通道。超大规模的并行计算机系统要求此通信网络具有特别高的通信容量。众所周知，由于电子系统传输速率的限制，在电交叉连接传输网中存在‘电子瓶颈’问题，它限制了通信容量的进一步提高，并且由于铜缆的损耗和时延在高速数据传输时更为严重，这使并行计算机系统的发展受到限制。光纤具有大带宽、低时延、低损耗、重量轻、不受电磁干扰等优点，可作为并行计算机系统中的节点连接介质。在并行计算机系统中采用同步光纤传输，大大简化了电子技术的复杂性，并且能轻易地突破‘电子瓶颈’，实现超大容量、低时延和极低误码率的并行计算机光纤连接[1~3]。

近年来光波分复用(WDM)技术应用于光纤通信，得到了很大的发展[4]，解决了干线传输容量和超高速交换的问题。在并行计算机系统中如果采用不同波长传输计算机的各个数据位，并采用光开关矩阵实现节点之间的连接，则使并行计算机内的 CPU-MEM 数据交换大大简化[5]。本文探讨了在并行计算机中采用光 WDM 连接的一种通信网方案，分析、讨论了其中的关键技术并且报道了一个实验系统及实验结果。

* 信息产业部重点科技计划发展项目资助(项目号：97021)。
收稿日期：2000-05-30；收到修改稿日期：2000-09-21

2　理论基础

对有 n 个 CPU 和 n 个 MEM 的并行计算机系统，如果数据字节长为 m 比特，则系统总连接线数目为 $m \cdot n^2$。在大型并行计算系统中，n 值可能非常大（如 1024 等），而比特位数 m 通常也是 32 或 64，采用电连接其总连线数可能高达数十兆条，此时控制电路变得异常复杂，而且时钟同步也成为一大难题。

采用光 WDM 技术，则可使这一交叉连接系统大大简化，巧妙实现大容量、低时延传输的并行计算机传输，结构如图 1 所示。

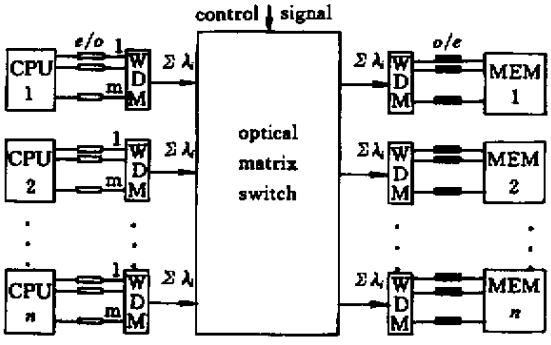

图 1　并行计算系统光 WDM 单向（CPU 到 MEM）交叉互连网络

Fig. 1　Optical WDM uni-directional C. I. net for parallel computer system

系统中，使用一个波长传输一位比特，则每个 CPU 需要 m 个不同波长的光。数据从 CPU（i）输出后，调制光成为多路光信号，再由 WDM 将它们汇入一根光纤，在控制信号的作用下，通过光矩阵开关选择到达某一指定 MEM（j）的路由，通过前置于 MEM 的光 WDM 将各比特位光信号分开，经过光电转换进入 MEM，实现了 CPU（i）到 MEM（j）的单向传输。MEM 向 CPU 的传输结构与此完全相同。如果采用两个不同波长（1310 nm 和 1550 nm）窗口的密集波分复用，则可同时实现 CPU→MEM 和 MEM→CPU 的双向光 WDM 互连，使并行计算机的 CPU-MEM 互连网络更加简化。

3　关键技术研究

3.1　光 WDM 技术

3.1.1　采用 DWDM 技术

并行计算机的比特位数已从 8 位发展到 32，64 位甚至更高，因此系统中 WDM 器件对应的波长数目是 8，32 甚至 64，对此需要采用密集波分复用（DWDM）器件。由于 1550 nm 窗口的 DWDM 技术目前已相当成熟，因而在实用上一般都选择 1550 nm 波长窗口。

3.1.2　特性优化措施

由于光纤具有色散特性，在各不同比特位（波长）之间产生传输时延，其值为

$$\Delta t = D_c \cdot L \cdot \delta\lambda \quad (\mathrm{ps}) \tag{1}$$

式中 D_c 为光纤的色散系数，L 为传输距离，$\delta\lambda$ 为两个不同光信号的波长间距。

对于 G.652 单模光纤，传输 1550 nm 窗口波长光的色散系数约为 20 ps/(nm·km)，对信道间隔为 0.8 nm 的 64 波长系统，其最大波长 $\lambda_{\max}$ 与最小波长 $\lambda_{\min}$ 之间的波长间隔的大小为 $(64-1)\times 0.8$ nm $= 50.4$ nm，由公式(1)可知，经过 100 m 传输距离后产生的最大传输时延为

$$\Delta t = D_c \cdot L \cdot \delta\lambda = 20\ \mathrm{ps/(nm \cdot km)} \times 0.1\ \mathrm{km} \times 50.5\ \mathrm{nm} \approx 100\ \mathrm{ps}$$

对于 2.5 Gbit/s 的信道速度，信号周期 $T = 1/B = 1/2.5\ \mathrm{Gbit/s} = 400\ \mathrm{ps/bit}$，可见 Δt 为 100 ps 的时延已经相当于 2.5 Gbit/s 速率信道一个比特位信号周期的 1/4，其影响不可忽略。

在实际工程中，可以通过使用色散系数较小的光纤，如色散位移光纤（DSF）和非零色散位移光纤（NZDSF）减小多波长的传输时延，当然也可减小信道间隔，但这意味着需要使用密集度更大的波分系统。

3.1.3　啁啾影响的仿真研究

激光器在稳态输出时光中心频率比较稳定，但当处于调制时其输出中心频率将发生偏移。对采用正弦小信号 $I_m = I_p \sin(\bar{\omega}_m t)$ 的直接电流调制，啁啾可表示为

$$\delta\gamma(t) = \delta\gamma_0 \sin(\bar{\omega}_m t + \theta_c) \tag{2}$$

式中 $\delta\gamma_0$ 为最大频率滑移量，可以看出，当调制信号的频率 $\bar{\omega}_m$ 很高时，输出光表现出强烈的啁啾效应。

在高速并行计算机中由于信道速率很高，当采用直接调制时，激光器的调制电流强度会随着比特取值（"0"或"1"）变化而快速变化，则由于啁啾的原因，中心波长会产生动态偏移，而使输出光谱展宽。光谱展宽不仅会引起信道间串扰的增加，也会在本信道内由于光纤色散作用使光脉冲波形展宽，其结果是导致在接收端眼图发生劣化，眼图闭合度的大小近似为[6]

$$\Delta = \left(\frac{4}{3}\pi^2 - 8\right) \cdot t_c \cdot K \cdot B \cdot \left[1 + \frac{2}{3}(K - t_c B)\right] \tag{3}$$

式中 t_c 表示激光弛豫振荡周期的一半，B 表示速率，K 由下式确定

$$K = D_c \cdot L \cdot \Delta\lambda \cdot B \tag{4}$$

式中 $\Delta\lambda$ 表示啁啾的波长偏移量。

如果眼图发生劣化，则在接收端会引起功率代价，由于信号幅度降低而引起系统信噪比(SNR)下降，当光探测器使用雪崩管 APD 时，系统的功率代价为[7]

$$P_c = -10\left(\frac{X+2}{X+1}\right) \cdot \lg(1-\Delta) \tag{5}$$

式中 X 表示雪崩光电管的过度噪声系数，对于 InGaAs-APD，$X = 0.8$。

图 2 示出了 2.5 Gbit/s 和 10 Gbit/s 两种信道速率时啁啾色散导致功率代价的仿真曲线。仿真表明，在 10 Gbit/s 速率下，啁啾色散超过 0.005 ns 时即有 1 dB 的功率代价，而对于 2.5 Gbit/s，即使啁啾色散大于 0.03 ns 时也无明显的功率代价。由图 2 可见，曲线随信道速率的增大而陡然上升。实际上由公式(3)和(4)知道眼图闭合度和信道速率的平方近似地成正比关系，因此当信道速率增加时，功率代价增加很大，这说明信道在高速率下，啁啾色散会造成较大的功率代价。

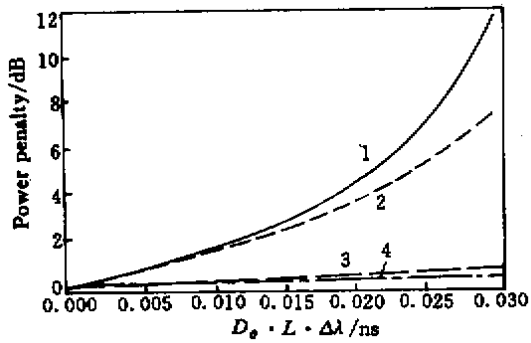

图 2　2.5 Gbit/s 与 10 Gbit/s 速率下的功率代价和啁啾的仿真曲线

Fig.2　Emulation curve of power penalty vs chirp dispersion at 2.5 Gbit/s and 10 Gbit/s

1：10 Gbit/s (0.1 ns)；2：10 Gbit/s (0.05 ns)；
3：2.5 Gbit/s (0.1 ns)；4：2.5 Gbit/s (0.05 ns)

在高速系统设计中，采用电吸收(EA)技术或马赫-曾德尔(M-Z)外调制，可以有效地抑制啁啾。

3.2　光矩阵开关的设计分析

在并行计算机中，光矩阵开关是按照系统控制信息指令提供不同 CPU-MEM 路由连接的器件。结构可采用由 2×2 光开关构成的空间矩阵光开关，或由 n^2 个 1×2 光开关和 n 个星型耦合器构成的 $n\times n$ 矩阵开关网络(即分送耦合光开关)。由于分送耦合光开关的控制简单，并能同时提供 n 条不同路由链路的连接，因此并行计算机采用分送耦合光矩阵开关较为理想。图 3 示出了这一结构。

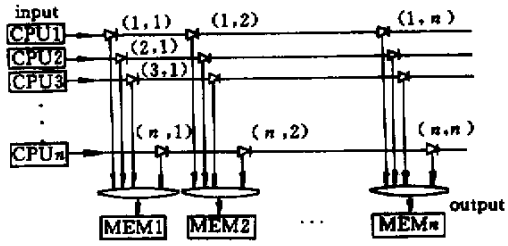

图 3　分送耦合光矩阵开关

Fig.3　Delivering-and-coupling optical matrix switches

由图 3 可知光纤矩阵网由 $n\times n$ 根纵横交叉而不互连的光纤组成，一个 CPU 对 n 个 MEM 构成一个星形网。每一个光开关处于矩阵网的一个相关位置，两个输出分别与纵向和横向光纤相连，并根据控制信息指令，将 CPU 的数据横向“连通”(纵向“断开”)或纵向“连通”(横向“断开”)。在光矩阵开关中，需要光开关对复用在一根光纤中的多个波长进行“连通”或“断开”，因此使用的光开关应是宽带的。

在图 3 所示的结构中，任何一个 1×2 光开关 (ij) 在“连通”状态时，即构成一个由 CPU (i) 到 MEM (j) 的链路，显然与同一个 MEM 相连的所有光开关在同一时刻只能有一个处于“连通”状态，否则将会造成多个 CPU 同时与它连通而发生链路冲突。

若 CPU (i) 希望向 MEM (j) 传递数据，则需要光开关 (ij) 处于“连通”状态，光开关 (ij) 的控制信号 Q_{ij} =“1”，若 CPU (i) 不向 MEM (j) 传递数据，则光开关 (ij) 应处于“断开”状态，需要 Q_{ij} =“0”。由图 3 可见，具有 n 个 CPU 和 n 个 MEM 的光矩阵开关中，总共有 n^2 个光开关，将全部光开关的控制信号 Q 按照它们的空间位置排列，可表示为如下矩阵

$$\begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1n} \\ Q_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ Q_{n1} & \cdots & \cdots & Q_{nn} \end{bmatrix} \tag{6}$$

式(6)称为光矩阵开关的控制矩阵。在某一时刻，按照所有 CPU 的通信请求，控制电路产生控制矩阵，矩阵元素值取“0”或“1”，它们对应 CPU 和 MEM 的位置排列和控制命令，构成所需链路，实现并行计算机的 CPU 与 MEM 传输。

4　实验与结果

本文对上述结构进行了实验探讨，设计了一个数据字节为 8 bit 和具有 2 个节点的并行计算机光 WDM 连接试验系统，并进行了实验研究，相关设计及实验结果如下。

4.1　系统设计

系统设计如图 4 所示。实验中，数据发送模块模拟 CPU，它包含随机数据信号发生器、激光管阵列和接口控制电路等。发送数据的每一位调制一个波长，然后将一个字节的光信号同时送入光 WDM，合波后由一根光纤传输。数据接收模块模拟 MEM，它包含雪崩管 APD 光探测器、接口控制电路等。控制电路产生时钟信号以及控制数据发送、数据接收模块和光矩阵开关的控制信号，它的功能是按控制信息指令分配路由，并负责网络的传输协调工作。实验用的主要元器件性能参数如表 1 列出。

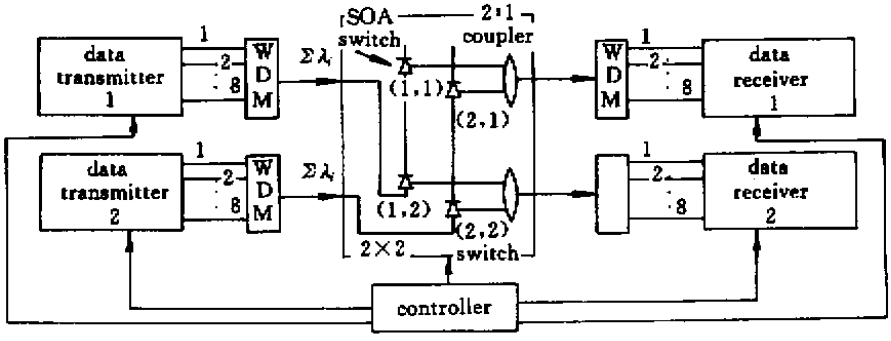

图 4　二节点 8 位并行计算机的光 WDM 链路系统(CPU 到 MEM)

Fig.4　Optical-WDM link system for 8 bits parallel computer with 2 notes

表 1　主要元器件参数

Table 1　Main devices′ parameter

Device	Parameter	Value	Test condition
LD set	Threshold current/mA	< 25	CW
	Wavelength/nm	1549.32,1550.92,1552.52,1554.13, 1555.75,1557.36,1558.98,1560.61	
	Spectral width/MHz	< 20	P_f = 20 mW
EA	Chirp coefficient	0.5	
WDM	Band profile	Gaussian	
	Insertion loss	< 4.5 dB	
	Insulation	> 25 dB	
	3 dB band width	> 0.48 nm	
	Temperature sensitivity	< +/− 1 pm/deg.C	
APD	Detection range/nm	1000 ~ 1600	
	Sensitivity/dBm	− 33	
	Responsivity/A·W^{-1}	0.95	λ = 1.55 μm

每个波长被调制在 1.2 Gbit/s，数据总容量为 2 × 8 × 1.2 Gbit/s = 19.2 Gbit/s。

4.2　实验结果及分析

实验结果表明，除了光纤色散带来的时延抖动外，还有激光二极管(LD)开通时的延迟抖动和电子电路信号处理产生的延迟抖动，前者是由于实际器件的差异，各激光器的阈值电流并不完全相同，即使在相同电流同时驱动下，各激光器发出光脉冲的时间仍然不完全一致，产生的时延抖动约为数百皮秒 后者是由于光发送模块、光接收模块、电平转换及印刷电路板的差异而产生的，其量级也在数百皮秒，实验时可以通过重定时电路减少其影响。

传输线采用长 100 m 的 G.652 单模光纤，色散造成的时延抖动≤10 ps，各种因素造成的总时延抖动被控制在 ± 1 ns 以下，小于时钟的 ± 10%，信道误码率测得为 10^{-12}量级，根据以上几个方面实验结果与数据分析，采用光 WDM 这一结构应用于并行计算机数据传输是完全可行的。

参 考 文 献

1 A. Takai, T. Kato, S. Yamashita *et al*.. 200-Mb/s/ch 100-m optical subsystem interconnections using 8-channel 1.3 μm, laser diode arrays and single-mode fiber arrays. *J. Lightwave Technol.*, 1994, **12**(2): 260 ~ 270

2 Shinji Nishimura, Hiroaki Inoue, Shoichi Hanatani *et al*.. Optical interconnections for the massively parallel computer. *IEEE Photon. Technol. Lett.*, 1997, **9**(7): 1029 ~ 1031

3 Shinji Nishimura, Hiroaki Inoue, Shoichi Hanatani *et al*.. Error-free optical inter-node connection for the massively parallel computer. *IEEE Photon. Technol. Lett.*, 1998, **10**(1): 147 ~ 149

4 Youju Mao, Mingrui Dang. Technology of Wave Division Multiplexing, Beijing: People's P&T Press, 1996. 7 ~ 10 (in Chinese)

5 J. Y. Fan, X. Zhao, J. P. Zhang *et al*.. Wavelength-division-multiplexed (WDM) data-block switching for parallel computing and interconnects. *SPIE*, 1998, **3491**: 634 ~ 636

6 Shu Yamamoto, Masakuni Kuwazuru, Hiroharu Wakabayashi *et al*.. Analysis of chirp power penalty in 1.55-μm DFB-LD high-speed optical fiber transmission systems. *J. Lightwave Technol.*, 1987, **5**(10): 1518 ~ 1524

7 S. Yamamoto, H. Sakaguchi, M. Nunokawa *et al*.. Studies of design for 1.55-μm optical fiber submarine cable systems. *J. Opt. Comm.*, 1986, **7**(1): 2 ~ 10