第27卷 第1期 电子科技大学学报 Vol.27 No.1
1998年2月 Journal of UEST of China Feb. 1998

基于程控纵横交换开关直通的计算机系统*

刘心松**
（电子科技大学计算机系 成都 610054）

【摘要】 在用纵横交换开关进行节点间互连的大规模并行处理系统中，引入可程控的节点间直通通信概念，相对于存储转发、虚拟直通、蛀孔寻径和电路交换，系统性能均有进一步提高。文中对基于程控纵横交换开关直通信息交换技术的计算机系统之加速比、效率、延迟时间和使用频带等系统性能参数进行了研究。结果表明，在大规模并行处理系统中，程控纵横交换开关直通是更具吸引力的节点间通信技术。

关 键 词 大规模并行处理； 纵横交换开关； 程控； 直通
中图分类号 TP302.7； TP311

在并行处理和分布式计算机系统中，较为流行的节点间信息交换技术有存储转发、虚拟直通、蛀孔寻径和电路交换等[1]。1）存储转发是一种交换技术，它把要在源节点和目的节点间传递的信息分为若干个包，每一包到达其间的每一中间节点时，都先把整个包存储在相应的缓存中，待通往下一节点的相应通道可用并且有足够的缓存空间时，则将该包传送到下一节点，如此直至该信息的所有包都抵达目的节点并且正确为止。这种交换技术的缺点是：当源节点和目的节点间的距离增加时，通信延迟时间也增大；同时每个节点的路由器需要较大的缓存空间。2）相对于存储转发，虚拟直通的特点是，当信息包进行时，如果相应于下一节点的通道可用，则信息包无需在本节点缓存而可直接传送到下一节点。虚拟直通虽然仍需较大的节点路由器缓存空间，但通信延迟时间已有显著减少。3）相对于虚拟直通，蛀孔寻径是把信息包分成许多小段，这种小段被称为 flit，它可以小到一个 flit 只为一个字节。一个信息包的第一个 flit 用于寻径控制，称之为头 flit，当头 flit 前进时，后面的 flit 一同前进。如果头 flit 不能前进时（如去下一节点的通道被占用），后面的 flit 亦停止前进，都就地暂存于相应的节点路由器缓存中，和虚拟直通相比，通信延迟时间也小，节点路由器缓存的需求量小。4）电路交换的过程是：在开始发送信息前建立源节点和目的节点间的物理路径，该路径一旦建立，信息即从源节点直接发向目的节点，直至该信息发完撤消该物理路径。电路交换技术对节点路由器缓冲的需求量小。

存储转发对节点路由器缓存需求量大，通信延迟时间也长。虚拟直通虽然仍需大的节点路由器缓存空间，但通信延迟时间已有减小。蛀孔寻径只需很小的节点路由器缓存空间，且通信延迟时间也小，但允许的节点间物理距离太小。电路交换无需节点路由缓存空间，但在大规模并行处理系统中，中间节点很多，一个节点一个节点地接通开关建立源节点和目的节点的路径，所需时间太长。本文提出的基于程控纵横交换开关计算机系统具有以下特点：1）通信延迟时间小；2）无需节点路由器缓存空间；3）节点间的物理距离相对于蛀孔寻径大有增加。

我们可借助于多个性能参数来评估一个并行处理或分布式计算机系统，但针对特定应用，可集

* 1997年5月19日收稿，1997年7月29日修改定稿
* 国家科委863高科技项目
** 男 56岁 大学 教授

中研究对其应用效果有重要影响的少数几个参数。本文根据应用需求研究加速比、效率、通信延迟和频带。

1 基本结构

我们的目标是研制大规模并行处理或分布式并行处理系统。若将系统中的所有节点都直接连入一级纵横交换开关，则会发生工程实施方面的困难。通常是将少量的节点例如64个节点连入一个纵横交换开关同时通过总线连入一管理机，如图1所示。我们称这一级为纵横交换开关计算机族，用 C_j 表示，族内节点数目用 N_j 表示。管理机的功能是：1）通信路由管理；2）由上层模块接收并向模块内节点传递任务；3）从模块内各节点汇集结果并传给上级模块。

在传递消息本体时，管理节点由一程控开关短接（旁通）。

将若干 C_j 通过 O 点连到新的纵横交换开关即构成更大规模的并行计算机，如图2所示。我们称这一级为纵横交换开关计算机群，用 C_g 表示，群内 C_j 的数目用 N_g 表示，M_g 为群内管理机。

如果将若干 C_g 通过 O 点连到第三级纵横交换开关，我们称为纵横交换开关计算机系统，用 C_s 表示，如图3所示。系统内 C_g 的数目 N_s 表示，系统管理机为 M_s。依此类推，可以构成更多级数纵横交换开关计算机系统。

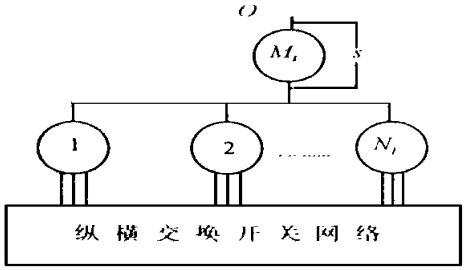

图1 交叉开关计算机族 C_j 结构图

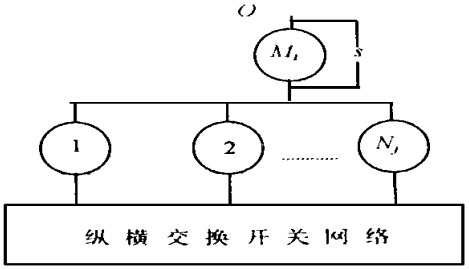

图2 交叉开关计算机群 C_g 结构图

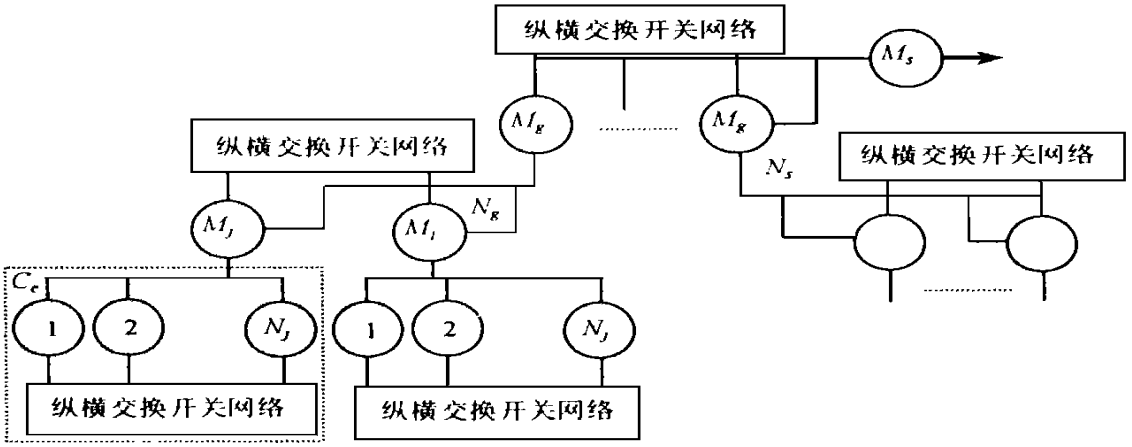

图3 三级纵横交换开关计算机系统 C_s 结构图

若 $N_j = N_g = N_s = 64$，则节点总数达131 072。就工程实现可行性和可预见的需求而言，这已经是一个足够大的数字。

族内节点的通信路由由该族管理机管理，群内不同节点间的路由由群族两级管理机共同管理，

不同群节点间的路由由系统群族三级管理机协同管理。

节点间需要通信时，经由总线将请求提交给相应最高级管理机，该管理机选定路由后即下达给相应下层管理机接通相应各级纵横交换开关之相应程控开关通知请求节点启动通信。管理机有其所属范围内的空闲路由信息。一次通信完成后即释放相应路由，更新相应管理机中的路由信息。算法应尽量考虑局部化以降低通信开销。

工程实现时，以图3 中虚拟框所示 C_c 部分为一基本单元，这样整个系统便可由一种部件连成。

2　加速比和效率

加速比和效率是并行处理中的重要性能指标。

定义 1　正则纵横交换开关计算机单元 C_0 每个节点有三条双向 I/O 路径，通过可程控的纵横行列数相等的纵横交换开关互连。

我们称这种基于程控直通的纵横交换开关计算机单元为正则纵横交换开关计算机单元，用 C_0 表示，如图 1 所示用一粗线表示三条双向 I/O 路径。

定义 2　总线纵横开关计算机单元 C_b 若 C_0 中所有节点又通过第四条双向 I/O 通道连到一条总线上，则称总线纵横交换开关计算机单元，用 C_b 表示。

定义 3　纵横交换开关计算机族 C_j 若总线上再挂一管理机，则为纵横交换开关计算机族，用 C_j 表示。

定义 4　纵横交换开关计算机群 C_g，以 C_j 为节点构成的新纵横交换开关计算机系统称为纵横交换开关计算机群 C_g。

定义 5　纵横交换开关计算机系统，以 C_g 为节点构成的新纵横交换开关计算机系统称为纵横交换开关计算机系统 C_s。依此类推，任意级数的系统用 C 表示。

定义 6　在电气上直接连通的两节点间无论经过了多少个开关，均称该两节点为相邻节点。

定义 7　程控纵横交换开关直通计算机系统，相应各管理节点在程序控制下，根据路由信息，接通源节点和目的节点间的选定纵横开关使之成为相邻节点达直通目的之系统。

定义 8　加速比 S_a 对一给定任务，单一节点完成的时间与程控纵横交换开关直通计算机系统完成时间之比。

定义 9　效率 $E-S_a$ 与节点数之比。

为论述方便起见，我们作如下假设：

假设 1　上述各种纵横交换开关系统中的节点总数为 $(2n)^i$。其中，$2n$ 为 C_0 中的节点数，最大值为 N_j，i 为系统中的级，设最大级数为 N。

假设 2　某一给定任务能等分为 $(2n)^i$ 个可并行执行的段，每一段的处理时间为 t_p。忽略任务处理期间(除原始任务数据广播和结果汇集外)的消息通信开销。

2.1　基于 C_j 和 C_s 的 S_a、E

若整个任务由一个节点处理，则处理时间为 $(2n)^i t_p$；若分为 $(2n)^i$ 段并行处理，则处理时间为 t_p，但要加上任务信息的传送接收时间和处理结果的汇集传送时间。在节点数为 $(2n)^i$ 的系统中，最大独立通信路径的数目为 $\sum_{i=1}^{N} 3n^i 2^{i-1}$。系统的加速比为

$$S_a = \frac{(2n)^i) t_p}{t_p \sum_{i=1}^{N} \frac{t_b + t_r}{3n^i 2^{i-1}}}$$

式中 t_p 为任务信息数据在两邻节点间的传送接收时间; t_r 为一段的处理结果在两邻节点间的传送接收时间。

效率为

$$E=\frac{t_p}{t_p+\sum_{i=1}^{N}\frac{t_b+t_r}{3n^i2^{i-1}}} \tag{2}$$

2.2 粒度对加速比和效率的影响

定义 10 粒度 假设式(2)为特例, 某一给定任务的分解情况一般可表示为 XYZ。整个任务分为 X 步, 每一步有 Y 段可并行执行, 一段的长度为 Z, Z 的大小即为任务的粒度, 相应的处理时间为 t。

假设 3 假设任务处理期间的通信开销为 $aX(A^{Y}_{(2n)^i})t_e$。其中, A 为排列符号, a 为系数, t_e 为一次信息交换的时间开销。

假设 4 在前 $X-1$ 步中, $Y=(2n)^i$。

假设 5 采用广播原始任务数据和汇集结果仅分别进行一次的算法。根据定义 7 和假设 3~5, 式(1)、(2)应改写为

$$S_a=\frac{XYt}{Xt+\sum_{i=1}^{N}\frac{t_b+t_r}{3n^i2^{i-1}}+aX(A^{Y}_{(2n)^i})te} \tag{3}$$

根据定义 10 和式(3)

$$E=\frac{XYt}{X_t+\sum_{i=1}^{N}\frac{t_b+t_r}{3n^i2^{i-1}}+aX(AY_{(2n)^i})te}(2n)^i \tag{4}$$

定理 **1** 在基于程控纵横交换开关直通计算机系统中, 加速比 S_a 和效率 E 均随粒度即 t 的增长而增加, 其极限值为 $S_a\to(2n)^i$, $E\to1$, 但永远不可能达到极限值。

3 通信延迟

节点间通信延迟时间是否足够短是并行处理系统能否成功的关键因素之一。

定义 11 通信延迟一节点(源节点)自提出发送消息请求开始至消息自源节点完全发送至目的节点为止所经历的时间即为此次通信的通信延迟, 用 T_d 表示。

定义 12 两节点间的直接的一条物理路径称为一个通道。

为对通信延迟进行数学分析, 特作如下假设。

假设 6 若干请求的到达服从泊松到达过程且相互独立, 以便用排队论进行分析。

假设 7 系统中消息的长度服从指数分布规律, 以适应固定长度的缓冲区。

假设 8 节点(含管理节点)的存储容量无限大, 因为有限的存储容量可能引起阻塞、重发和消息的丢失等。

假设 9 系统中所有节点相同, 通道相同, 以简化分析。

假设 10 通信优先级最高, 一旦有请求节点立即响应。

假设 11 预处理时间和后处理时间均不包括在通信延迟时间内。

假设 12 不考虑在物理线路上的纯传输时间, 因为相对说来它很小。

在程控纵横交换开关直通的计算机系统中，节点间通信延迟时间[3]分析如下。

3.1　源节点经由总线向管理机发请求的时间

该时间包括逐级的请求等待时间和发送请求时间两部分之和，即

$$\left(\sum_{i=1}^{N} \left(\frac{p_{b_i}}{\mu_m c_{ci}(1-p_{bi})} + \frac{1}{\mu_h c_{ci}} \right) \right. \tag{5}$$

式中　p_{b_i}为一 i 级总线忙的概率；$1/\mu_m$ 为经总线传输的消息平均长度；Cc_i 为一 i 级总线的传输容量；ρ_{b_i}为一 i 级总线的利用系数；$1/\mu_h$ 为消息头长度。

若系统由一级纵横交换开关和一级总线组成，则式(5)成为

$$\frac{\rho_b}{\mu_m C_c(1-\rho_b)} + \frac{1}{\mu_h c_c} \tag{6}$$

式中　第二项中出现 $1/\mu_h$ 是因为此时只需向管理机发送消息头即可。

3.2　路径建立时间

若源节点和目的节点属于一 i 级系统，则其间的路径将包含$(2i-1)$个通道，根据排队论，等待路径$(2i-1)$个通道都闲而建立可用路径的时间为

$$\frac{\rho_{c_i}^{2i-1}}{\mu_m c_p(1-\rho_{c_i}^{2i-1})} \tag{7}$$

式中　C_p 为一路径的传输容量；C_i 是通道的平均利用系数或忙的概率。

3.3　管理机向源节点和目的节点发“命令”的时间

为简化分析，设向源节点和目的节点发“命令”的时间相等，即同时向两节点广播。鉴于路径已经建立，故无等待时间。命令发送时间则为

$$\frac{1}{\mu_0 c_p} \tag{8}$$

式中　$1/\mu_0$ 为命令的平均长度。

3.4　源节点向目的节点直接发送消息的时间

在基于程控纵横交换开关直通的计算机系统中，一旦源节点和目的节点间的直通路径建立，该路径实质上变成了一物理通道，故其消息发送时间为

$$\frac{1}{\mu_m c_p} \tag{9}$$

由此可得，基于程控纵横交换开关直通的并行处理系统中节点间平均通信延迟时间为

$$T_d = t_{sm} + \sum_{i=1}^{N} \left(\frac{\rho_{b_i}}{\mu_m c_{c_i}(1-\rho_{b_i})} + \frac{1}{\mu_h c_{c_i}} \right) + \frac{\rho_{c_i}^{2i-1}}{\mu_h c_p(1-\rho_{c_i}^{2i-1})} + \frac{1}{\mu_0 c_p} + \frac{1}{\mu_m c_p} \tag{11}$$

式中　t_{sm}为在源节点和管理节点的处理时间。

当 $i=1$，并且 $\rho_{b_i}=\rho_{c_i}=\rho_b$，$c^0=c_{c_i}=c^0$ 时

$$T_d = t_{sm} + \frac{2\rho_b}{\mu_m c_p(1-\rho_b)} + \frac{1}{c_p}\left(\frac{1}{\mu_h} + \frac{1}{\mu_0} + \frac{1}{\mu_m} \right) \tag{12}$$

通常，命令和消息头的长度都远小于消息本身的长度，故

$$T_d \approx t_{sm} + \frac{1}{\mu_m c_p}\left(\frac{1+\rho_b}{1-\rho_b} \right) \tag{13}$$

定理 2　在基于程控纵横交换开关直通的并行处理计算机系统中，路径(含总线和纵横交换开

关通道)的利用率越高, 节点间的通信延迟时间越长, 极限为∞。路径利用率越低, 节点间的通信延迟越短, 极限为 $t_{sm}+1/(\mu_m c_p)$。

4 系统使用频带

系统使用频带是并行处理的又一重要性能指标。

定义 13 请求系统中两节点间的发送或接收要求。

定义 14 周期从一个请求的产生开始直至完成该请求之任务所经历的时间。

定义 15 系统使用频带每个周期系统接收请求的平均数即为系统的使用频带。

定义 16 利用率每个周期一个节点接收请求的平均数。

假设 13 每个节点随机产生请求, 请求间彼此独立, 请求在节点间均匀分布。

假设 14 在一个周期中, 一节点产生仅涉及族内节点的一个请求的概率为 P_j, 出族但仅涉及族内节点的概率为 P_g, 出族涉及系统的概率为 P_s。依此类推, 对任意级, $(p_j+p_g+p_s+\cdots)$即为每个节点在每个周期产生的平均请求数。

假设 15 周期间彼此独立, 即在一个周期内, 产生请求的概率不受对前一周期中所产生请求的接收情况的影响。

假设 16 系统内所有节点之结构相同。

假设 17 一节点发送一消息或接收一消息的服务率为 μ, 并采用先来先服务的非剥夺调度算法。

根据上述定义和假设, 在一个周期内有 k 个请求的概率为

$$q(k)=\left\{\begin{matrix}(2n)^i\\ k\text{建立}\end{matrix}\right\}(p_j+p_g+p_s+\cdots)^k[1-(p_j+p_g+p_s+\cdots)]^{(2n)^{i-k}} \tag{14}$$

在一个周期中有 k 个请求的期望值为

$$E(k)=\frac{A^k_{(2n)^i}-A^k_{(2n)^{i-1}}}{A^k_{(2n)^i}} \tag{15}$$

期望的频带[3] 为

$$B[(2n)^i]=\sum_{0\leqslant k\leqslant(2n)^i}E(k)q(k) \tag{16}$$

对系统而言, 这就是请求到达率。

节点利用率为

$$u=\sum_{0\leqslant k\leqslant(2n)^i}q(k)\left\{\frac{A^k_{(2n)^i}-A^k_{(2n)^{i-1}}}{A^k(2n)^i}\right. \tag{17}$$

到此可以顺便指出, 请求的平均响应时间为

$$R=\frac{c(\rho,u)}{\mu(2n)^i-B}+\frac{1}{\mu} \tag{18}$$

式中

$$c(\rho,u)=\frac{(\frac{B}{\mu})(2n)^i}{(\frac{B}{\mu})^{(2n)^i}+[(2n)^i\ !\left\{1-\frac{\lambda}{\mu(2n)^i}\ \mathrm{e}^{B/\mu}\right.}$$

一个请求被接收的概率为

$$p_r = \frac{B}{(p_j + p_g + p_s + \cdots)(2n)^i} \tag{19}$$

定理 3　在基于程控纵横交换开关直通的计算机系统中，系统的使用频带与节点数成正比。

5 结束语

在大规模并行处理系统中，加速比、效率、通信延迟和系统使用频带是其主要的系统性能参数，通信延迟和频带又更为重要。通信延迟的大小决定系统能否成功，频带宽窄表示实际并行度和吞吐力的大小。引入程控纵横交换开关直通，使任何两节点间的通信距离均可变为 1（一个物理通道），由此通信延迟小，加之频带又宽，促成了大的加速比和高效率。因此，这是大规模并行处理系统的一种优秀系统结构方案。

在 863 和其他预研项目的资助下，我们已经研制成功了两种实验系统。在 8 个节点和使用纵横交换开关通信网络的情况下，相对于存储转发，直通的通信速度提高一个数量级以上，系统使用频带、加速比和效率均有大幅度提高。可以肯定地指出，在大节点数情况下，效果会更好。

参 考 文 献

1　程代杰. Wormhole routing — 大规模并行处理系统中的一项关键技术. 计算机应用，1994，14(2)：5～8

2　刘心松等. 超立方体计算机直通通信技术. 电子科技大学学报，1993，22(4)：400～406

3　Hwang Kai, Briggs Faye A. Computer architecture and parallel processing. N Y：McGrawHill Book company

Computer System Based on Programmable Crossbar Switches Cut-through Communication

Liu Xin Song

(Computer Dept, UEST of China　Chegndu　610054)

Abstract　Massively parallel processing system using crossbar switches to interconnect nodes & introducing programmable cut-through communication between nodes has better performance as compared with store- &forward, virtual cut-through, wormhole routing and circuit switching. Presented in this paper are speedup ratio, efficiency, delay & bandwidth of the computer system based on the programmable crossbar switches cut-through communication. The result shows that the communication is an attractive communication technology for nodes of the massively parallel processing system.

Key words　massively parallel processing;　crossbar switches;　programmable;　cutthrough

编辑　黄　辛