流水——减少MPP系统广播延迟的有效方法*

刘宏伟 李晓明 崔光佐

(哈尔滨工业大学计算机系 哈尔滨 150001)

摘要 大规模并行处理系统中的通讯开销是影响系统性能的一个重要因素.广播操作是MPP系统中常用的通讯方式,快速地实现广播将有助于提高系统的性能.本文基于TORUS互连网络,提出了实现广播的流水方式,并分析了它的性能,指出采用流水方式进行广播,可以减少广播延迟,改进系统的性能.

关键词 广播,流水,多计算机系统,TORUS网络.

在可扩展的并行体系结构中,当处理机的数目增加时,系统的性能也相应增加.当处理机的数目很多时(>100),称为大规模并行处理系统MPP(massively parallel processing),它由大量的具有分布存储器的结点处理机构成,当系统中处理机的个数增加时,系统的通讯带宽、存储器带宽及处理能力也相应提高.

在MPP系统中,处理机间有效的通讯对系统的性能至关重要.通讯开销将严重地限制系统的加速比.本文研究一种特殊的通讯方式——广播,广播是指一个处理机将数据发送给所有的处理机,它是科学计算和商业领域等应用问题中最为重要的通讯方式之一.目前对广播的研究越来越多,虽然有的机器提供了实现广播的硬件,如广播总线等,更多的研究是基于没有硬件支持的广播算法[1~11],如文献[1]提出的Dual—Path和Multi—Path算法及文献[3]提出的U—mesh,U—cube算法等.这些广播算法都有其不同的出发点,Dual—Path,Multi—Path算法主要是避免wormhole路由开关技术可能造成的死锁,U—mesh和U—cube算法则是在减少通道拥挤的前提下,尽快地实现广播.本文提出的流水广播方式由于按互连结构组织广播,可以避免路径冲突,同时,采用流水方式减少了广播的延迟时间.

1 分析模型

评价MPP系统的一个重要标准是通讯延迟,它是3个值的和:启动延迟、网络延迟和阻塞时间.启动延迟(Start—up Latency)是源和目的处理机用来处理信息的时间,又可以分为发送延迟和接受延迟.网络延迟(Network Latency)等于信息的第1个数据进入网络到最

* 本文研究得到国家自然科学基金资助.作者刘宏伟,1968年生,博士生,主要研究领域为大规模并行处理的体系结构.李晓明,1957年生,教授,博士导师,主要研究领域为大规模并行处理,群机系统,并行编译.崔光佐,1968年生,博士生,主要研究领域为大规模并行处理的体系结构.

本文通讯联系人:刘宏伟,北京 100083,北京科技大学计算机系

本文1995—07—07收到修改稿

后 1 个数据流出网络的时间. 阻塞时间(Blocking Time)是信息在网络中传输时,因网络的通道拥挤等原因可能造成的延迟. 广播延迟是指从处理机发送第 1 个广播数据到最后 1 个处理机接受到全部广播数据并完成计算的时间. 如何减少广播延迟依赖于不同的体系结构,本文将依赖于下述体系结构:

(1)互连网络的结构为二维花环(TORUS)结构. 一个二维花环结构有 $M\times N$ 个处理机,每个处理机由坐标(X,Y),$0\leqslant X\leqslant M-1$,$0\leqslant Y\leqslant N-1$ 标识. 如果 2 个处理机 p,q,它们的坐标满足 $Xp=Xq$,$Yp=(Yq\pm 1)\ mod\ N$ 或 $Yp=Yq$,$Xp=(Xq\pm 1)\ mod\ M$,则称 2 个处理机是相邻的处理机.

(2)采用 Store-and-Forward 路由开关技术. 在这种方法中,当信息包到达一个中间处理机时,全部信息存放在处理机的信息包缓冲区(Packet Buffer)中. 如果信息传送所需要的通道是空闲的,并且相邻处理机的缓冲区有足够的空间,信息包送往下一个处理机. 假设信息的长度为 M,通道的带宽为 w,那么所有信息发往通道所需的时间为 M/w. 如果信息通过 2 个处理机互连线路的时间为 r,则信息在 2 个相邻处理机间的网络延迟为 $M/w+r$. 用 $Tsnd$,$Trcv$ 分别标识信息的发送延迟和接受延迟,得到信息在 2 个相邻处理机间的通讯延迟 D 为

$$D=Tsnd+M/w+r+Trcv \tag{1}$$

(3)每个结点处理机用一个独立的路由器(Router)进行处理机间的通讯. 如图 1 所示.

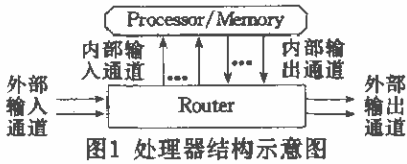

图1 处理器结构示意图

几对外部通道用于和相邻处理机的路由器相连. 路由器通过一对或多对内部通道和处理机/存储器相连,每对通道一个用于输入,一个用于输出. 本文假设每个结点处理机只有一对内部通道,因此处理机必须串行地接受或发送数据,连续的 2 次发送数据和接受数据之间的时间间隔,文中以 $g(gap)$ 表示,它反映处理机和路由器之间的带宽. 本文用 C 表示处理机对广播数据的计算量. 通常,C 与广播数据的长度 M 有关. 例如,在矩阵向量乘中,C 为 $O(M)$ 量级,而在 N 体问题中,C 为 $O(M^2)$ 量级. 本文假设 C 为 $O(M)$ 量级,因此计算其中 m 个数据的计算量可估算为 $C\times(m/M)$.

2 流水广播方式的性能分析

一个 MPP 系统可以用图 $G(V,E)$ 表示. V 是结点集,每个结点 v 代表一个处理机,E 是边集,代表处理机间的互连线路. 如果$(v_i,v_j)\in E$,则 v_i,v_j 所代表的处理机间有直接的通讯线路. 一个 4×4 的 TORUS 网络如图 2 所示.

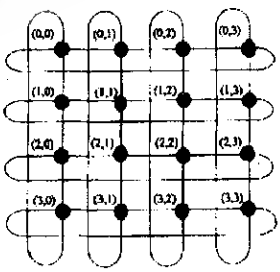

图 2 TORUS 网络示意图

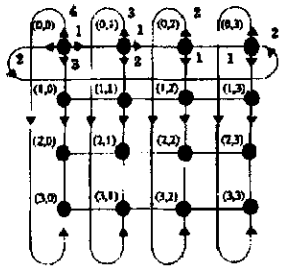
图 3 TORUS 网络上的广播示意图

由于 TORUS 网络是对称的,不失一般性,可以取(0,0)结点为广播结点. 广播的路由函数 R 为:R: $N\to P(N)$,N 为结点集,$P(N)$为结点集的有序子集合. 在 $n\times n$ 的 TORUS 网络中,以(0,0)结点为广播结点的路由函数 R 为:

$$R((i,j))=\begin{cases}\{(0,1),(0,n-1),(1,0),(n-1,0)\} & i=j=0\\ \{\ \} & i=n/2 \text{ or } i=n/2+1\\ \{(0,j+1),(1,j),(n-1,j)\} & i=0,0<j<n/2\\ \{(1,j),(n-1,j)\} & i=0,j=n/2 \text{ or } i=0\ j=n/2+1\\ \{(0,j-1),(1,j),(n-1,j)\} & i=0,n/2+1<j\leqslant n-1\\ \{(i+1,j)\} & 0<i<n/2\\ \{(i-1,j)\} & n/2+1<i<n-1\end{cases}$$

广播时,每个结点(i,j)按 $R((i,j))$定义的有序集合依次发送数据,一个结点上连续的 2 次发送之间要经过时间间隔 g. 4×4 TORUS 网络上的广播如图 3 所示,图中箭头旁的标号是结点接到一个新的数据块时,向相邻结点发送的顺序. 可以直接证明,采用这种广播算法,广播结点和任意结点之间的路径,并且广播不会遇到路径冲突.

当不采用流水方式进行广播时,广播数据首先由(0,0)结点依次发送给它相邻的 4 个处理结点,每个结点接到全部数据后,先按 R 所定义的有序集合将数据依次发送到相邻节点,然后再对广播数据进行处理. 当每个结点都接到数据并完成了计算时广播结束. 由图 3 可以看出,从$(0,0)\to(n/2,n/2)$ 的路径最长,为 n. 假设广播数据的长度为 M,经过处理机间直接互连的延迟为 D,则这条路径的通讯延迟为 nD,代入(1)式,得:

$$t=n\times(Tsnd+M/w+r+Trcv) \tag{2}$$

虽然路径$(0,0)\to(n/2+1,n/2)$的长度为 $n-1$,但是由于在$(0,n/2)$结点,数据先发往$(1,n/2)$,再发往$(n-1,n/2)$就必须间隔 g,因而这条路径的通讯延迟为:

$$t_1=(n-1)\times(Tsnd+M/w+r+Trcv)+g \tag{3}$$

比较(2)、(3)式,可知 t,t_1 的大小取决于 g 和 $Tsnd+M/w+r+Trcv$ 的大小.

表 1 给出了几个典型系统的通讯参数,从表中可以看出,g 通常小于 $Tsnd+M/w+r+Trcv$,即便对 DASH 和 J—machine,如果 M 稍大,g 小于 $Tsnd+M/w+r+Trcv$ 也成立,因而我们假设 $t_1<t$. 同样分析其它路径的通讯延迟,可知路径$(0,0)\to(n/2,n/2)$的通讯延迟决定了广播延迟.

表 1　几个典型系统的通讯参数

系统	Cycle(ns)	w(bits)	Tsnd+Trcv	g	r
NCUBE/2	25	1	6400	6400	40
CM—5	25	4	3600	3600	8
DASH	30	16	30	40	2
J—machine	31	8	16	40	2
Monsoon	20	16	10	10	2

如果处理机对广播数据的计算量为 C,广播延迟为

$$t=n\times(Tsnd+M/w+r+Trcv)+C \tag{4}$$

当采用流水方式进行广播时,(0,0)结点把 M 个广播数据分 P 次广播,每次 m 个数据(称为一块数据),$P=[M/m]$. 每次(0,0)结点都将 m 个数据依次发往它相邻的 4 个结点:(0,1),(0,$n-1$),(1,0),($n-1$,0),每个结点接受到数据后再按 R 定义的有序集合传送给

其它的结点.对于其中的任何一条路径,多块数据象在工厂中的流水线上一般流过路径上的处理机,因此我们称这种广播方式为流水方式.由于(0,0)处理机在连续的 2 次发送之间必须有一定的时间间隔,所以每条路径上前后 2 块数据的间隔时间至少为 $4g$,如图 4 所示.

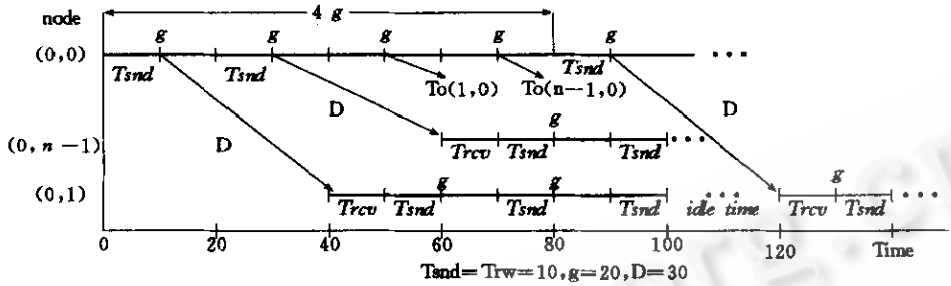

图4 流水广播的示意图

如果处理机 $Tsnd$,$Trcv$,g 的值不随发送数据量的多少变化时,流水方式的广播延迟为

$$t'=n\times(Tsnd+m/w+r+Trcv)+([M/m]-1)\times 4g+C \tag{5}$$

式中第 1 项是 m 个数据流过路径(0,0)→($n/2$,$n/2$)的时间,即流水线的注入时间.以后每隔 $4g$,都会有一块数据到达结点($n/2$,$n/2$),这样经过($[M/m]-1$)×$4g$ 时间后,全部数据到达结点($n/2$,$n/2$),再加上结点对数据的处理时间就是广播延迟.

从图 4 也可以看出,中间处理机在接受连续的 2 个数据块之间有很长一段空闲时间,这段空闲时间降低了处理机的利用率,也给流水方式带来了副作用,但是如果利用这段时间完成对刚刚接受到的一块数据的处理,使通讯和计算并行,会进一步减少广播延迟.通过对流水广播过程的分析可知,中间处理机每接到一块数据后,用于通讯的时间最多为 $Trcv+3Tsnd$(如图 3 中的(0,1)结点),因此当利用空闲时间进行计算时,处理机接受下一块数据的间隔时间应是 $4g$ 和 $Trcv+3Tsnd+C\times m/M$ 中的较大值,此时通讯延迟为:

$$t''=n\times(Tsnd+m/w+r+Trcv)+([M/m]-1)\times\mathrm{MAX}(4g,Trcv+3Tsnd+C\times m/M)$$

t''加上结点($n/2$,$n/2$)对最后一块数据的处理时间 $C\times(m/M)$,就是新的广播延迟:

$$t_1=n\times(Tsnd+m/w+r+Trcv)+([M/m]-1)\times\mathrm{MAX}(4g,Trcv+3Tsnd+C\times m/M)+C\times m/M \tag{6}$$

(4)~(6)得:

$$\triangle=n\times(M-m)/w+C\times(1-m/M)-([M/m]-1)\times\mathrm{MAX}(4g,Trcv+3Tsnd+C\times m/M) \tag{7}$$

式中前 2 项是采用流水方式减少的广播延迟,而最后一项是流水方式所增加的时间.如果流水方式能减少广播延迟,则 $m\leqslant M$,并且$\triangle\geqslant 0$,由(7)可得:

$$n\times M/w+C\geqslant 4g \text{ 或 } n\times M/w\geqslant Trcv+3Tsnd \tag{8}$$

当系统的通讯参数确定时,由上式可以求出 M 的下限.当 M 满足(8)式时,对$\triangle$求导,并令$\triangle'=0$,可得 m 的最佳值:

$$m=\begin{cases}\sqrt{4g/(n/w+C/M)} & 4g\geqslant Trcv+3Tsnd+C\times m/M\\ \sqrt{(Trcv+3Tsnd)\times w\times M/n} & 4g<Trcv+3Tsnd+C\times m/M\end{cases} \tag{9}$$

当 $M=1024(bits)$,$C=1024(cycles)$,$n=16$,通讯参数参考 CM−5,DASH,Monsoon 3 种机器时,采用流水方式进行广播对广播延迟的影响(假设 $Tsnd=Trcv$):

(1)对 CM−5,$g=3600$,可以假设 $4g\geqslant Trcv+3Tsnd+C\times m/M$,将 M,C 代入(8)式,

得 $n\times M/w+C<4g$,所以不宜采用流水方式进行广播,从中也可知流水方式的采用受限于处理机和路由器之间的通讯带宽.

(2)对 DASH,$g=40$,我们先假设 $4g<Trcv+3Tsnd+C\times m/M$,将 M 代入(8)式,得 $n\times M/w\geqslant Trcv+3Tsnd$,可以采用流水方式进行广播. m 的最佳值为(m 取整数):

$$m=\sqrt{60\times 16\times 1024/16}=248$$

m 的取值使假设 $4g<Trcv+3Tsnd+C\times m/M$ 成立,此时

$$t=16\times(30+1024/16+2)+1024=2560$$

$$tp=16(30+248/16+2)+([1024/28]-1)(60+1024\times 248/1024)+1024\times 248/1024=2240$$

$\triangle=320$,流水方式可以减少 12.5%的广播延迟. 当 $M=2048$,$C=2048$ 时,同样计算可得流水方式减少了 29%的广播延迟.

(3)对 Monsoon,$g=10$,我们还是先假设 $4g<Trcv+3Tsnd+C\times m/M$,将 M,C 代入(8)式,得 $n\times M/w\geqslant Trcv+3Tsnd$,故可以采用流水方式进行广播. m 的最佳值为(m 取整数):

$$m=\sqrt{20\times 16\times 1024/16}=144$$

m 的取值使假设 $4g<Trcv+3Tsnd+C\times m/M$ 成立,此时

$$t=16\times(10+1024/16+2)+1024=2240$$

$$tp=16(10+144/16+2)+([1024/144]-1)(20+1024\times 144/1024)+1024\times 144/1024=1628$$

$\triangle=612$,流水方式可以减少 27.3%的广播延迟. 当 $M=2048$,$C=2048$ 时,同样计算可得流水方式减少了 34%的广播延迟,这是非常乐观的数字. 由此也可以看出,g 值的减小,即处理机和路由器之间的通讯带宽增加及广播数据量 M 的增加,采用流水方式进行广播的收益也相应增大.

当通讯参数参考 DASH,Monsoon 和 J-machine 3 种机型时,如果处理机个数等于 256,$M=1000(bits)$,$C=1000(cycles)$,用(7)式可以推出流水方式和非流水方式的广播延迟的差随 m 变化的曲线. 如图 5 所示. 从图中可以看出,3 条曲线都呈抛物线形,m 存在一个值使得广播延迟的差最大.

3　结束语

在一个 MPP 系统中,只要广播的数据量 M 满足(8)式,采用流水方式进行广播就可以减少广播延迟. 当 M=2048(bits)时,取 DASH,Monsoon 和 J-machine 3 种机型的通讯参数,广播延迟可以减少 30%左右. 虽然本文对流水方式的分析主要基于 TORUS 网络,Store-and-Forward 路由技术,实际上对其它的互连网络和其它的路由技术以及其它的通讯方式,采用流水方式都会对系统的性能有所改进. 同时,随着硬件技术的不断提高,系统的通讯参数 $Tsnd$,$Trcv$,g,w 也在不断改进,另外硬件代价的降低,使得 MPP 系统中处理机数目不断增多,与其相应的是处理的问题的规模也越来越大,即 M 的值也在不断增大,这些都为采用流水方式提供了更充分的理由.

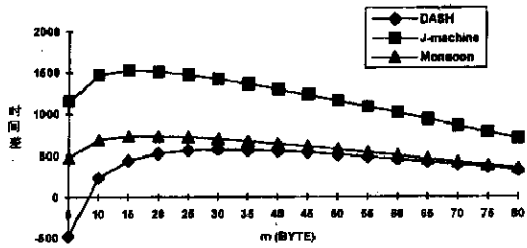

图 5 广播延迟的差随 m 变化的曲线(非流水方式－流水方式)

参考文献

1 Lin Xiaola, Ni L M. Deadlock－free multicast wormhole routing in multicomputer networks. ISCA'91, 1991. 116～125.

2 Ni L M, McKinley P K. A survey of wormhole routing techniques in direct networks. IEEE Trans. Computers, 1993, **C－42(2)**: 62～76.

3 McKinley P K, Xu Hong, Esfahanian A H *et al*. Unicast－based multicast communication in wormhole－routed networks. ICPP'92, II－10－II－19, 1992.

4 Culler D, Patterson D. LogP: towards a realistic model of parallel computation. In: Patterson David, Hot Topics on Advanced Computer Architecture, 国家智能计算机研究开发中心技术资料, 1993.

5 Lin Xiaola, Ni L M. Multicast communication in multicomputer networks. ICPP'90, III－114－III－118, 1990.

6 Bruck Jehoshua, Cypher Robert, Ho Ching－Tien. Multiple message broadcasting with generalized fibonacci trees. Frontiers of Massively Parallel Processing, 1992. 424～431.

7 L Ju－Yong, Lee Park Sang－Kyu, Choi Hyeong－Ah. Circuit－switched broadcasting in d－dimensional tori and meshes. International Conference on Parallel and Distributed Computing, 1994. 554～560.

8 Rajeev Thakur, Alok Choudhary. All－to－all communication on meshes with wormhole routing. International Conference on Parallel and Distributed Computing, 1994. 561～565.

9 Izidor Jerebic. Optimal broadcasting in toroidal networks. Frontiers of Massively Parallel Processing, 1992. 671～676.

10 Byrd G T, Saraiya Nakul P, Delagi B A. Multicast communication in multiprocessor systems. ICPP'89, I－196－I－200, 1989.

11 刘宏伟, 李晓明. 多机系统中"分发"和"流水"两种广播方式的性能分析. 哈工大并行计算技术实验室技术报告, PACT－TR－94－015, 1994.

PIPELINING——AN EFFECTIVE METHOD TO REDUCE BROADCASTING DELAY OF MPP SYSTEMS

Liu Hongwei Li Xiaoming Cui Guangzuo

(*Department of Computer Science and Technology University of Harbin Industry Harbin* 150001)

Abstract Communication overhead in massively parallel processing systems is an important factor which affects the performance of MPP. Broadcasting is a communication method used frequently in MPP systems. If broadcasting can be implemented as quickly as possible, then the performance of system will be increased. Based on the TORUS interconnection network, this paper proposes the pipelining broadcasting, analyses its performance and shows that pipeline broadcasting can derease the broadcasting delay and improve the performance of MPP systems.

Key words Broadcast, pipeline, multicomputer, TORUS network.