一种基于网络分解的多播通讯路由方法*

谢 澎 朱怡安 康继昌

(航空工业总公司计算技术研究所 西安 710068) (西北工业大学计算机系 西安 710072)

王雅昆

(航空工业总公司计算技术研究所 西安 710068)

摘要 有效的消息通讯是提高分布存储器并行计算机性能的关键因素. 点对点通讯和广播通讯是 2 种常用的消息通讯方法,而多播通讯(Multicasting)是指从一个源节点同时给任意多个目标节点发送消息,这种通讯比点对点和广播 2 种方式更具一般性,适用于很多实际应用的需求. 本文针对 PAR95 并行计算机的二维网格结构,提出一种基于网络分解的多播消息通讯方法,并比较了该方法与用多个点对点方法实现多播通讯的性能.

关键词 并行计算,消息通讯,多播通讯.

基于消息传输的并行系统,其性能在很大程度上取决于下层的通讯机制. 多播通讯是指一条消息从一个源节点向任意多个目标节点的发送,这种通讯方式在许多实际应用中非常有效,如计算机网络仿真、数值并行算法等. 许多并行计算机系统,如美国 Caltech 的 Cosmic[1]都支持多播通讯功能.

我们在研制 PAR95 并行计算机系统[2]过程中,一方面由于航空气动数值计算应用需要有多播通讯功能的支持,另一方面为了有效地支持分布式共享存储器的并行程序设计模型[3],在 PAR95 并行操作系统中实现了多播通讯功能.

一般讲,多播通讯可以用多个点对点通讯或者广播通讯来实现[4],但是,在这 2 种情况下,都会在系统中产生过多不必要的通讯开销,影响多播通讯的效率,从而降低系统性能. 我们提出一种基于网络分解的双路径多播通讯方法,可以比上述 2 种实现方法大大改善通讯效率. 本文第 1 节介绍竞赛图和哈密顿路径的概念;第 2 节讨论如何根据一条哈密顿路径对网络节点进行标号排序,并将一个 T805 互连的二维网格拓扑结构分解为 2 个满足竞赛图性质的子网络. 第 3 节介绍双路径多播通讯算法. 本方法与采用多个点对点通讯来实现多播通讯的方法的性能比较在第 4 节中给出. 最后给出了结论与建议.

---

* 本文研究得到国防科委"八五"预研课题"并行处理技术研究"的资助. 作者**谢澎**,1957 年生,博士,高级工程师,主要研究领域为并行处理,分布式处理,并行程序设计环境. **朱怡安**,1960 年生,博士,副教授,主要研究领域为并行处理,流场计算,可视化计算. **康继昌**,1929 年生,博士导师,教授,主要研究领域为计算机系统结构,并行处理. **王雅昆**,1970 年生,助理工程师,主要研究领域为图形显示界面.

本文通讯联系人:谢澎,西安 710068,航空工业总公司计算技术研究所

本文 1995-08-29 收到修改稿

1　竞赛图和哈密顿路径

竞赛图(Tournment Graph)定义为一个有向图 $G(V,E)$,其中 V 为节点集合,E 为边集合.对 V 中每一对节点 i 和 j,或者 $\langle i,j\rangle \in E$ 或者 $\langle j,i\rangle \in E$,但 $\langle i,j\rangle$ 和 $\langle j,i\rangle$ 不能同时都在 E 中.

哈密顿路径(Hamilton Path)是指在一个有向图 $G(V,E)$ 中,从某节点 $i \in V$ 出发,可以访问 G 中所有节点一次而且仅仅一次的一条路径.

竞赛图具有许多重要性质,其中一条是有关哈密顿路径的,即对任意一个竞赛图 $G(V,E)$,都存在一条哈密顿路径(证明见文献[5]).

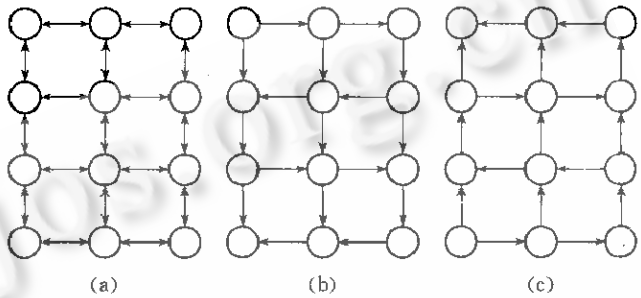

图1 (a)4×3二维网格;(b),(c)分解的2个竞赛图

PAR95 并行计算机的拓扑结构是采用英国 INMOS 公司生产的 T805[6]互连的二维网络,每 2 个相连节点之间存在 2 条反方向的通讯链路.它是一个有向图,但不是一个竞赛图.我们可以将这样一个拓扑结构分解为 2 个竞赛图,如图 1 所示,其中(a)表示一个 4×3 二维网格,(b)和(c)分别表示从(a)分解出来的 2 个竞赛图,每个竞赛图都包括原图中所有节点,但是只包括单方向的边.

2　基于哈密顿路径的节点排序与网络分解

一般情况下,二维网格结构的节点可用一个整数对 (x,y) 表示,为了下面讨论网络分解和多播算法的方便,我们这里介绍一种根据一条哈密顿路径对网络节点的表示方法.

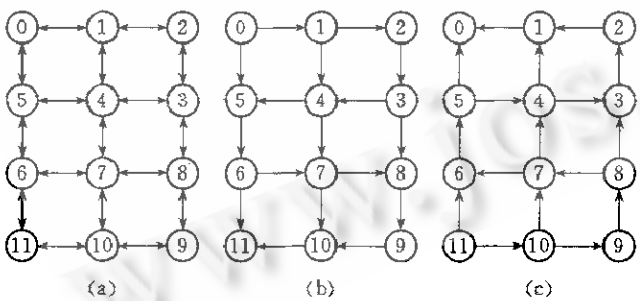

图2 (a)在4×3二维网格上选定一条哈密顿路径;(b),(c)对应的高节点网络和低节点网络

根据节点在一条哈密顿路径上出现的先后顺序,给其分配一个相应整数作为该节点的编号,即哈密顿路径的起始点的节点编号为 0,第 i 个节点的编号为 $i-1$,终节点的节点编号为 $N-1$(假设图中共有 N 个节点).例如,在图 1(a)的 4×3 二维网格中,选择哈密顿路径为{(0,0),(0,1),(0,2),(1,2),(1,1),(1,1),(1,0),(2,0),(2,1),(2,2),(3,3),(3,1),(3,0)},则相应的节点编号分配如图 2(a)所示.同时,图 1(b)和(c)2 个竞赛图就可定义为对应原来网络分解成的 2 个子网络(图 2(b)和(c)).我们把图 2(b)称为高节点网络而把图 2(c)称为低节点网络.从图 2 中可以看到,在高节点网络中,消息总是从低编号节点向高编号节点发送,而在低节点网络中,消息则总是从高编号节点向低编号节点发送.这样一种性质保证了在消息传送过程中决不会形成循环链路,从而不会发生死锁.

3 双路径多播通讯算法

为了描述双路径多播通讯算法，首先要讨论消息准备和路由策略.

3.1 消息准备

所谓双路径多播通讯，就是指在一次多播通讯中，允许消息同时沿着最多2条独立的路径传输.

一次多播通讯，有一个发出消息的源节点，记为 S_0，有 K 个接收消息的目标节点，记为 $D_K=\{d1,d2,\ldots,dk\}$. 我们根据上节所述的节点编号的大小，可以将目标节点集合 D_K 分解成2个子集合 D_H 和 D_L. 在 D_H 中的每个目标节点的节点编号都大于源节点 S_0 的节点编号. 而在 D_L 中的每个目标节点编号都小于源节点 S_0 的节点编号. 双路径算法的思想就是把 D_K 作为一个整体考虑，形成2条有序的消息传输路径，让源节点从高节点网络中向 D_H 中的目标节点发送消息，而从低节点网络中向 D_L 中的目标节点发送消息. 显然，这种方法不会产生多个单点通讯和广播通讯2种方法中的那种冗余消息传输.

为实现双路径多播通讯，源节点需要进行消息准备，主要完成以下工作：第1，把 D_K 分解为 D_H 和 D_L；第2，把 D_H 和 D_L 中节点分别按节点编号的升序和降序进行排列；第3，要把消息路由数据加入到欲多播发送的消息中.

在多计算机系统中，可以采取确定性路由和适应性路由2种方法. 确定性路由是由源节点决定本次通讯所要经过的所有中间节点；而适应性路由中在源节点仅仅决定向哪个邻节点发送消息而并不决定其余的中间节点. 确定性路由的优点是算法简单，路由开销小，其主要缺点是路由过程缺乏适应性，从而有可能造成网络负载不平衡，从而影响性能. 适应性路由方法可根据网络负载的实际情况，动态地选择合理的路径，但是这种方法一般比较复杂，从而路由本身的开销较大，一般在硬件支持下效果较好. 在PAR95并行机中，我们采用一种路由表方法[2]，它基本上属确定性路由方法，但是它结合了适应性路由的思想，用户可根据程序执行的动画显示和通讯负载统计图来调整通讯路由表[7]，在一定程度上克服确定性路由方法的缺点. 这种方法的唯一开销是路由表要包括在消息头中，从而增加消息长度，但是几个字节的传输量所产生的传输时间的增加，相对于整个通讯时延是微乎其微的.[6]

算法1. 双路径多播通讯消息准备算法

输入：$meg0$——欲多播发送的消息；D_K——目标节点集合；S_0——源节点编号；
输出：$megH$——通过高节点网络发送的消息；$megL$——通过低节点网络发送的消息
BEGIN；$D_H=\{di\mid di\in D_K \& di>S_0\}$；
$D_L=\{di\mid di\in D_K \& di<S_0\}$；
$D_{H'}=SORTH(D_H)$；对 D_H 按升序进行排序
$D_{L'}=SORTL(D_L)$；对 D_L 按降序进行排序
$dh=FIRST(D_{H'})$；
$dl=FIRST(D_{L'})$；
$megh=ROUTING(S_0,dh)$；取 S_0 到 dh 之间的路由
$megl=ROUTING(S_0,dl)$；取 S_0 到 dl 之间的路由
$megH=PutMulti(megh,D_{H'},meg0)$；将路径和目标节点集合放入高节点网络传输消息中
$megL=PutMulti(megl,D_{L'},meg0)$；将路径和目标节点集合放入低节点网络传输消息中
END；

3.2 路由策略与双路径多播通讯算法

在源节点准备消息时，除了要把经过排序的 D_H 和 D_L 分别包括到消息中，还要把路由

信息包括到消息中.因消息从源节点首先要发送给 D_H/D_L 中第 1 个目标节点,因此在源节点放入的路由信息是从源节点到第 1 个目标节点的路径,消息到达第 1 个目标节点后,该节点将根据下 1 个目标节点,填入相应的路径.这样一直到消息到达最后 1 个目标节点,本次多播通讯完成.

算法 2. 双路径多播通讯算法

```
输入:megH 或者 megL;d0——本节点编号;
输出:修改后的 megH 和 megL;
BEGIN:Di=GetMulti(megH/megL); 取消息中的目标节点集合
      dl=FIRST(Di);
      Switch (dl)
       Case dl=d0;本节点是目标节点
         Download(meg0);
         Remove(dl,Di);从目标节点集合中删除本节点
         dl=FIRST(Di);
         megi=Routing(d0,dl);
         PutMulti(megi,Di,megH/megL);
         Break;
       Case dl=φ;目标节点集合为空
         Stop Routing;停止消息传送
         Break;
       Default: 其余情况
         Break;
       end Switch
       TransMeg(megH/megL); 按路径转发消息
END;
```

4　性能比较

如前述,多播通讯也可用多个点对点通讯或者广播通讯来实现,但是,在这2种方法下,通讯效率很差,它们都产生过多的消息量,从而造成网络拥挤,延迟增大.我们用一个仿真程序测试和比较了用多个点对点算法和本文的双路径算法这2种不同的多播通讯实现方法在一个8×8二维网格结构上的性能,我们用的性能参数是完成一次多播通讯所使用的通讯链路总数目.

用一个随机数发生器产生[0,63]范围内的一个整数 K,它表示一次多播通讯中的目标节点数目,我们对每种算法运行10 000次,然后取平均值得到该算法所使用的通讯链路数.由图3可看到,我们的算法与多个点对点通讯这种实现方法比较,性能要优越很多.

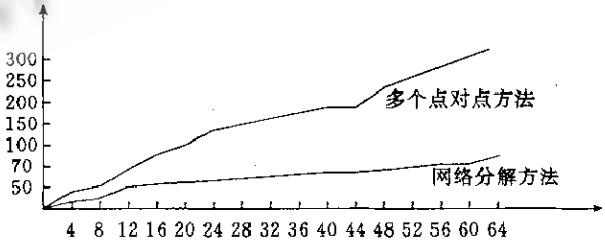

图3　8×8网格结构上2种多播通讯实现方法的性能比较

5 结论与建议

多播通讯是一种非常有用的通讯方式,虽然可用多个点对点和广播方法实现,但是都将引起过多通讯量从而性能较差.

本文提出一种针对 Transputer 互连二维 mesh 结构网络,将一个网络分解成具有竞赛图性质的2个子网络,使一条多播消息可在2个子网络中沿2条独立路径同时进行传送,仿真结果表明,这种方法的性能大大优于前2种实现方法.

如果采用虚通道概念和技术[4],则网络分解概念可进一步推广到将一个网络分解成若干个子网络,使多播通讯的目标节点集合分布到各个子网中,从而产生一种多路径的多播通讯算法,诸如分解为多少个子网及目标节点,如何在各子网中分布才能产生较好的通讯性能这些问题,都需进一步的研究.

参考文献

1 Seitz C L. The C programming's guide to multicomputer programming. Tech. Rept. Caltech_CS_TR_88_1, 1988.

2 谢澎. PAR95并行计算机软件设计报告. 631所技术报告,西安,1992.

3 谢澎. 面向气动数值模拟应用的并行程序设计环境研究与实现[博士论文],西安:西北工业大学,1995.

4 Dally W J. Deadlock—free message routing in multiprocessor interconnection networks. IEEE Trans. on Computer, May 1987, **C—36**:547~553.

5 唐善策. 并行图论算法. 安徽:中国科技大学出版社,1991.

6 INMOS Corp. T805 Engineering Data. England, 1993.

7 谢澎. 一个并行程序性能调试软件设计. 中国青年计算机研究新进展,西安:西北工业大学出版社,1994. 95~98.

A NETWORK—PARTITION BASED MULTICASTING ROUTING METHOD

Xie Peng

(*China Aeronautics Computing Technique Institute Xi'an* 710068)

Zhu Yi'an Kang Jichang

(*Department of Computer Science Northwestern Polytechnical University Xi'an* 710072)

Wang Yakun

(*China Aeronautics Computing Technique Institute Xi'an* 710068)

Abstract The efficient message communication is very important to improve the performance of distributed memory parallel computers. Point to point and broadcast communication are two message passing methods often used. Multicasting is to send messages from one source to any destinations, this kind of communication is more general than the two others, and required by many real world applications. A network partition based multicasting routing method for PAR95 parallel computer is provided in this paper, and the performance comparison is made.

Key words Parallel processing, message communication, multicasting.