基于ATM的群通信问题的研究*

吴礼发 周笑波 谢 立 孙钟秀

(南京大学计算机科学与技术系 南京 210093)
(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘要 群通信在并行计算中起着重要的作用.ATM(asynchronous transfer mode)网络有许多特点使之适合群通信.如何有效地利用ATM的这些特点来实现群通信操作是一个重要的研究课题.该文提出了一种基于ATM的群通信结构——混合树(Hybrid-Tree),该结构适合动态组中的群通信,并且能有效地利用ATM的特点.文中提出的组管理协议和树结构维护方法很好地解决了树结构的维护问题.

关键词 ATM,群通信,多目发送.
中图法分类号 TP393

在基于报文传递的并行和分布式计算环境中,通信可以分为两类:一类是点到点通信(Point-Point),参加通信的对象只有源和目的进程;另一类是群通信(Collective),主要用于一组进程同时参与的通信.群通信可分为3类:①进程控制,如路障同步(Barrier Synchronization);②数据移动,如多目发送(Multicast)、分发(Scatter)、收集(Gather)、多对多广播以及多对多分发-收集等;③全局计算操作,如归约(Reduction)和扫描(Scan).其中最基本的群通信操作是多目发送.

群通信在科学计算应用中显得特别重要,因为在这些应用中常常需要划分,分布大的数组到执行程序的结点上.结点通过群通信操作来实现分布、收集和交换数据;在全局数据上执行全局计算;在程序流中某一执行点同步.由于它的重要性,报文传递接口标准MPI对群通信作了详细的规定.

近几年来,人们对大规模并行计算机MPCs(massively parallel computers)中的群通信问题作了大量的研究,产生了很多行之有效的算法.因为工作站网络NOW(network of workstations)还是近几年随着工作站性能的不断提高以及高速网络的出现(如ATM)而迅速发展起来的一种并行计算结构,所以,研究基于ATM(asynchronous transfer mode)的NOW环境中如何实现有效的群通信操作是当前的一个热门研究课题.

本文讨论基于ATM的群通信问题,提出一种基于ATM的群通信结构模型——混合树(Hybrid-Tree).第1节介绍一些背景;第2节是本文的重点,讨论群通信结构模型;第3节对维护群通信结构的组管理协议进行描述.最后讨论一些相关的工作.

1 背 景

1.1 ATM简介[1]

异步转移模式ATM是一种采用固定大小的数据单元(信元,53字节)快速分组交换.ATM网络有许多特点使之适于进行群通信.首先,因为ATM网络是基于交换的网络,所以,多个报文可以同时通过交换机,这一特点使得网络竞争减少,从而使得群通信操作的执行时间缩短.其次,基于交换的光纤局域网有一个显著的特点是输入输出通路在物理上是独立的,这一特点使得我们能够有效地实现交换操作(Swap),在很多群通信操作中要用到这一操作.第3,ATM虚通路VC(virtual channel)的预先建立也能改善通信性能.如果一个应用的通信模式是可知的,则我们可以在应用开始执行之前,利用虚通路建立起应用的虚拟拓扑结构,这将大大提高群通信的性能.最后,一些ATM交换机对多目发送的支持也是提高群通信性能的一个重要方面.

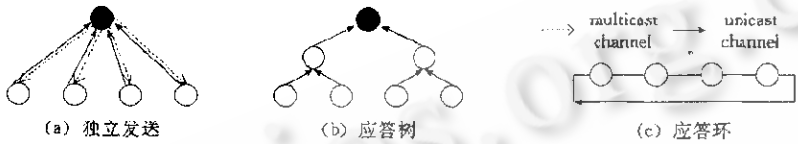
* 本文研究得到国家攀登计划基金资助.
作者吴礼发,1968年生,博士生,主要研究领域为计算机网络,高速通信.周笑波,1973年生,博士生,主要研究领域为并行与分布式系统.谢立,1942年生,教授,博士导师,主要研究领域为分布式系统.孙钟秀,1936年生,教授,博士导师,中国科学院院士,主要研究领域为并行与分布式系统.

本文通讯联系人:吴礼发,南京 210093,南京大学计算机科学与技术系
本文1997-06-16收到原稿,1997-07-21收到修改稿

1.2 可靠的点到多点 multicast 连接和多点到多点连接

随着有关 ATM 信令协议的国际标准的逐步完善，几乎所有的 ATM 交换机都从硬件上支持 Multicast. 这种 Multicast 连接是一种一对多(One-to-Many)的树型连接，树根是发送报文的源结点，所有的叶结点是报文的接收方. 按照 ATMUNI3.0/3.1 的规定，只有根结点才能增加叶结点，并且连接是单工的，即只有根结点才能发送报文. 尽管新的 UNI 标准 4.0 中允许叶结点可以自己加入到一个 Multicast 连接中去，但是，连接仍然是单工的. 因为 ATM 硬件并不提供任何形式的应答和流控措施，所以连接是不可靠的.

为了实现可靠的点到多点 multicast 连接，必须建立其他的连接来实现应答的传递. 传输应答的方式有很多种，如图 1 所示.



(a) 独立发送 (b) 应答树 (c) 应答环

图1 multicast应答传输

当叶结点数目很大时，图 1(a)所示的独立发送法显然不合适，因为根必须建立 n 条连接(假定有 n 个叶结点)，将消耗大量的资源，且对一个大的组而言，容易引起报文源的应答风暴，并且这种方法对应答的处理是串行的. 但这种结构易于维护和处理. 图 1(b)所示的树结构有比较好的可伸缩性和性能，但当有 multicast 成员发生改变时(加入或离开)，需要一个好的维护算法. 对于图 1(c)的环形结构，如果 multicast 结点太多，将会带来较大的时延.

除了一对多的 multicast 连接，还有一种多对多(Many-to-Many)的 multicast 连接. 多对多连接指的是，一个组的任何一个成员都可以利用这个连接来发送报文，而其他的组员都能收到. 目前，ATM 硬件并不提供这种连接. 要实现它，可以用 multicast 服务器和 ATM 交换机提供的一对多的 multicast 功能.

在服务器方式下，每一个组成员建立一条 unicast 连接到服务器，服务器建立一条一对多的 multicast 连接到所有的组成员. 这种方式存在着如图 1(a)所示的方式所面临的同样问题. 还有一种途径是，每一个报文源建立一条一对多的 multicast 连接到所有其他的组成员. 但这种途径很快会用完 VCI(virtual channel identity)，并且因为一旦一个成员加入或退出组，则要求这个多对多的连接中的所有的一对多的连接必须也随之改变，从而很难管理.

为了实现多点到多点连接(点到多点连接是它的一种特殊情况)以及解决应答的传输问题，本文提出一种混合树(Hybrid-Tree)结构.

2 群通信结构模型

2.1 基本的群通信结构模型

我们利用混合树(Hybrid-Tree)结构来实现群通信操作. 首先，我们来看看如图 2 所示的树结构，图中所有结点均为端系统(主机)上的进程，而不是交换机.

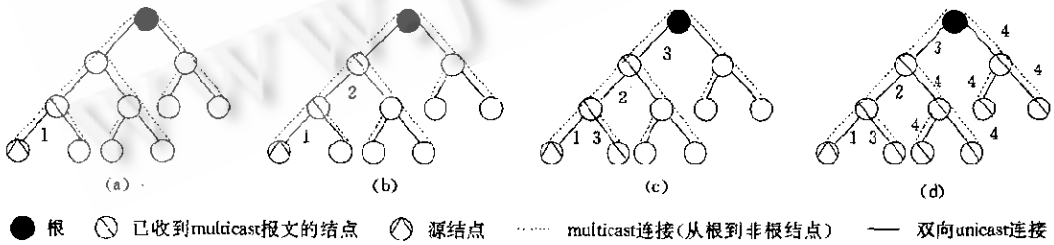


图2 基于混合树的multicast

在图 2 中，所有结点形成一棵树(一般为二叉树)，非根结点同它的父结点之间都有一条双向 unicast 连接，非叶结点同它的子结点之间也有一条双向 unicast 连接. 从根结点到所有非根结点之间还存在一条硬件提供的 multicast 连接(从逻辑上看，这些非根结点是硬件提供的 multicast 连接的叶结点，这条一对多连接同时形成了另一棵树)，如虚线所示. 我们称这种树为混合树(Hybrid-Tree). 这样，这棵混合树就形成了一条多点到多点的连接. 任何一个结点均可给其他所有结点发送报文. 当有结点加入或退出时，树结构发生改变，一个结点可能有两个以上的儿子结点，点到多

点连接的叶结点的数目也发生改变.

如果非根结点在它的 unicast 链路上收到报文，则它在所有除收到该报文的链路以外的链路上转发报文. 如果它是在 multicast 链路上收到报文，则不转发. 如果根结点收到一个报文，则用 multicast 链路转发报文. 在非 mulitcast 链路上转发报文的目的有 3 个：(1) 将要发送的报文传至根结点；(2) 提高报文传输的可靠性，如果一个结点没有从 multicast 链路上收到报文，则它还有机会从 unicast 上收到报文；(3) 一些结点可以在更短的时间内收到报文，如与报文源在根结点的同一侧的子树上的一些结点. 因为结点既可从 unicast 连接上接收报文，也可从 multicast 连接上接收报文，所以，一些结点可能会收到一些重复的报文，根据报文序号可以保证结点不会重复接收. 尽管这样会消耗一些额外的网络带宽，但它增加了传输的可靠性，降低了时延. 我们以最左边的叶结点作为源结点来说明在这样的结构中是如何实现 multicast 通信的. 这里，我们假定一个结点一次只能发送一个报文. 如图 2(a)～(c)所示，报文经过 3 步($\log N$)到达根结点，根结点收到报文后，利用 multicast 连接在一个时间步内将报文送至所有非根结点. 图中链路边上的数字代表着报文步数，在本例中，完成 multicast 需要 4 个报文步.

应答的传送也是通过树来完成的，即执行一个 multireceive 操作，结果存放在报文源结点. 一个结点只有在收到它的输出链路返回的应答后，才向它的输入链路发送应答. 注意，报文源不同，结点的输入输出链路是不同的. 为了解决应答风暴问题，将输出链路返回的应答合并后，再向它的输入链路发送应答.

利用上述结构，我们还可以实现其他群通信操作，如路障同步. 实现方法如下：执行一个 $N/1$ 归约操作，结果存放在根结点. 一个结点只有在收到它的所有子结点(包括它自己)的到达同步点的指示后，才向它的父结点发送到达同步点的指示报文，这个过程如图 3(a)所示. 根结点收到所有子结点到达同步点的报文后，利用它的 multicast 链路发送所有结点均已到达同步点的指示报文，如图 3(b)所示. 收到这个报文的结点后就可继续向前推进了. 所有收到根结点到达同步点的指示报文的结点均发送一个应答给根结点，这个过程同图 3(a)是一样的.

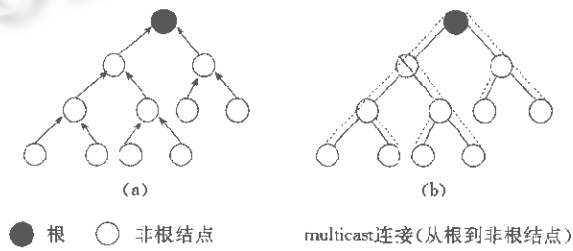


图3　基于混合树的路障同步

2.2　树结构维护

当有结点失败、新结点加入或结点退出时，树结构必须发生改变. 改变方法如下.

当有非根非叶结点失败或退出时，该结点的所有儿子结点的父结点改为失败或退出结点的父结点. 这一改变要求受影响的结点拆除原有的 unicast 连接，同新的父结点建立新的 unicast 连接. 而从根到非根结点间的 multicast 连接只需从其叶结点中去掉失败或退出结点即可. 最坏的结果是除根结点外，其余结点都是叶结点. 这一过程如图 4(a)(b)所示.

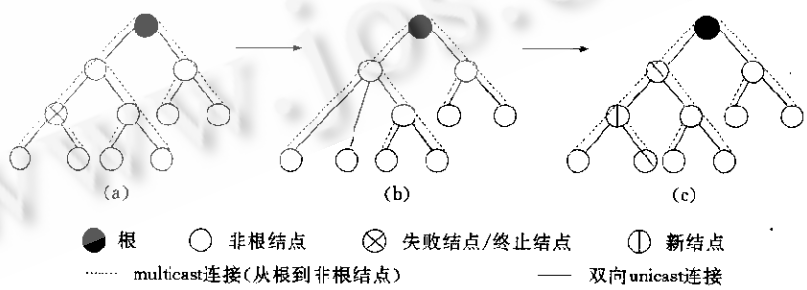


图4　树结构的维护

当叶结点失败或退出时，失败或退出的叶结点的父结点只需拆除同它的 unicast 连接即可. multicast 连接的改变同非根非叶结点时一样.

当有新结点要加入时，如果先前有退出或失败的结点，则该新结点取代退出或失败的结点在原树中的位置，如图 4(c)所示. 如果没有这样的结点，则将新结点作为叶结点. 对于 multicast 连接，根将新的结点加入到这个 multicast 连接的叶结点中去.

不同组的混合树是可以复用的. 复用方法是这样的：如果两个进程同时属于两个不同的组，则两个组的混合树共

用这两个进程之间的 unicast 连接.通过报文头中的组号即可实现复用.在树结构维护时,只有当两个进程不同在任何一个组时,才拆除它们之间的 unicast 连接.

在图 4(b)所示的树中作各种群通信操作的算法仍然是一样的.

3 组管理

群通信是一组进程参与的操作,因而需要对组进行管理.我们首先简要介绍几个概念.

拓扑信息中心 TIC(topology information center):在全网结点中,有一个结点存有全网的拓扑信息,并且这种拓扑信息是最新的(即被定期更新),我们称这种结点为拓扑信息中心.一般来说,拓扑信息中心是运行在网管结点上的一个进程,它能够从网管进程处得到最新的全网拓扑信息.进程组管理主要在拓扑信息中心中得以实现.

操作代理 OA(operating agent):每一个参与群通信操作的计算结点(有计算任务的网络结点)上都有一个进程负责该结点上的群通信报文的发送和接收,我们称这个进程为操作代理.操作代理与用户进程之间的通信可以采用共享内存、管道等来实现.操作代理可以支持一个结点上运行多个用户进程.

组代理 GA(group agent):每一个组都有一个组管理者,负责组员的加入和退出.这个组管理者称为组代理.在组代理中,存放着所有组员的信息.所有组代理上的成员信息的总和就等于拓扑信息中心中存放的成员信息.组代理一般作为该组的群通信树的树根.

进程用组名来申请加入组,由 TIC 分配组号,并维护组号与组名的一致性.通信时,用组号(GN)来标志组.这种集中式的分配策略保证了唯一性.每一个进程有一个标志号 PN(进程在组中的 rank 值),二元组(GN,PN)唯一地标志组中的一个进程.PN 由组代理 GA 分配.每一个组代理中存有它所在组的所有组员的 PN 号.

3.1 加入一个组

我们用一个例子来说明进程申请加入一个组的过程,假定进程 $U1$ 申请加入组 g,其加入过程如图 5 所示.

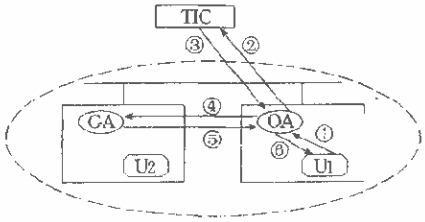


图5 进程 $U1$ 申请加入/退出查询组 g

(1) $U1$ 向本机上的操作代理发申请加入组 g 的申请报文;

(2) OA 收到申请报文后,即向 TIC 发申请报文;

(3) TIC 收到该报文后,查组定义表:如果组 g 已存在,则作如下处理:将 $U1$ 加入组 g 的组成员表,并告知组 g 的组代理的地址.否则,产生组 g 的定义项,并将该 OA 作为组 g 的代理;

(4) OA 收到 TIC 发回的报文后,如果自己不是该组的根结点(组代理),则向该组的 GA 发注册报文;

(5) GA 收到注册报文后,记录该组成员信息,产生相应的表项,发应答报文给 OA,告知该成员在群通信树中的位置;

(6) OA 收到 GA 发回的应答报文后,修改树结构,并通知 $U1$ 加入组 g 成功.

3.2 退出一个组

退出一个组的过程同加入一个组的过程类似.我们仍以 $U1$ 为例来说明.

(1) $U1$ 向本机上的操作代理发退出组 g 的申请报文;

(2) OA 分别向 TIC 和本组的 GA 发 $U1$ 退出组 g 的申请报文;

(3) TIC 收到申请后,将 $U1$ 从组 g 成员表中删除,并发应答给 OA;

(4) OA 分别向 TIC 和本组的 GA 发 $U1$ 退出组 g 的申请报文;

(5) GA 收到申请后,同样将 $U1$ 从组 g 的组员表中删除,告知 $U1$ 的儿子结点的新的父结点.并发应答给 OA;

(6) OA 收到应答后,通知 $U1$ 离开组 g 成功.

3.3 查询组信息

查询组信息的过程则比较简单.具体过程如下.

(1) $U1$ 向本机上的操作代理发查询组 g 的信息申请;

(4) OA 收到申请后向 GA 发查询申请;

(5) GA 收到查询申请后将信息返回给 OA;

(6) OA 收到应答后将信息返回给 $U1$.

3.4 失败检测

对组员的失败检测主要通过以下 3 种机制来实现.

(1) TIC处理从网管送来的结点失败事件.如果是非组代理所在的结点失败,则其上的所有组员进程即告失败,应将这种情况通知进程所在组的GA;如果失败的是GA所在的结点,则应重新指定一个GA,并修改树结构或删除该组.

(2) GA定期向TIC发状态报告.GA必须定期向TIC报告其状态,如果TIC在规定时间内没有收到报告,则按一定策略决定是否发一个查询报文给该GA,如果同样没有收到响应,则认为该代理已失败.它必须寻找一个新的GA或删除该组.

(3) OA定期向本组的GA发状态报告.为了使GA了解在本组内的所有OA的状态,规定所有OA必须定期向所在组的GA发状态报告.如果GA在规定时间内没有收到报告,则按一定策略决定是否发一个查询报文给该OA,如果同样没有收到响应,则认为该OA已失败.

4 相关工作

基于ATM的NOW环境中如何实现有效的群通信操作的研究是当前的一个热门研究课题.Y. Huang, Chengchang Huang等人对在ATM网络上作并行计算所涉及的通信问题作了大量的研究.[2~4]他们将研究重点放在如何利用ATM交换机的基于交换的特点来更有效地实现并行计算中的群通信这一问题上.他们定义了一种multicast虚拟拓扑结构M-array以及从M-array演变而来的、资源利用率更高的REM-array结构.在这种结构上,可以有效地实现多对多multicast连接以及各种群通信操作.但是,这种结构要求比较多的一对多的硬件multicast连接,当组比较多,或组成员数目较大时,所需要的资源可能得不到满足.并且,这种结构不容易改变,因而只适合于静态组.

本文提出的混合树结构模型,不仅适合静态组也适合动态组.就效率而言,在最坏情况下(深度最大的叶结点发送multicast),能够在$(\log N+1)$个报文步完成multicast操作.对通信资源(ATM的unicast连接和点到多点multicast连接)的需求量不大,因而实现起来比较可行.并且,在这种结构模型上也较易实现其他的群通信操作,如路障同步、归约等.文中提出的树结构维护方法和组管理协议较好地解决了树结构的维护问题.

参考文献

1 Othmar Kyas. ATM Networks. London: International Thomson Computer Press, 1995

2 Huang Cheng-chang, Mckinley P K. Communication issues in parallel computing across ATM networks. IEEE Parallel & Distributed Technology, 1994,2(4):73~86

3 Huang Cheng-chang, Huang Yih, Mckinley P K. A thread-based interface for collective communication on ATM Network. In: Proceedings of the International Conference'95 on Distributed Computing System. Vancouver, British, 1995

4 Huang Yih, Huang Cheng-chang, Mckinley P K. Multicast virtual topologies for collective communication in MPCs and ATM clusters. In: Proceedings of Supercomputing'95. San Diego, CA, Dec. 1995. http://www.supercomp.org/sc95/proceedings/

Research of Collective Communications on ATM Networks

WU Li-fa ZHOU Xiao-bo XIE Li SUN Zhong-xiu

(*Department of Computer Science and Technology Nanjing University Nanjing* 210093)

(*State Key Laboratory for Novel Software Technology Nanjing University Nanjing* 210093)

Abstract Collective communications play an important role in parallel computing. Many features of an ATM (asynchronous transfer mode) switch environment can be exploited in the design of collective operations. In this paper, the authors present a collective communications framework based on ATM——hybrid-tree, which fully takes advantage of the features supporting multicast of ATM and is suitable to implement collective communications in dynamic groups. The problems of the maintance of the hybrid-tree are completely solved by the group management protocol and methods to maintain the hybrid-tree presented in this paper.

Key words ATM(asynchronous transfer mode), collective communication, multicast.