计算机科学 1997 Vol. 24 No. 6

分布式系统中组与组通信机制的研究*)

Research on Group and Group Communication Mechanisms in Distributed Systems

王兴伟 张应辉 刘积仁 李华天

(东北大学软件中心研究部 沈阳 110006)

摘 要 Group and group communication is an important research field in distributed systems. In this paper, the basic concepts on group and group communication are presented at first, and then how to classify group and its communication mechanisms is described. The problems which should be solved when designing group and group communication systems are discussed. The new challenges posed by distributed multimedia group applications are introduced. Finally, some conclusions are given.

关键词 Distributed systems, Group, Group communication, Distributed multimedia

近年来,随着分布式系统的发展,组与组通信机制不仅在传统的分布式组应用领域(如复制文件系统、分布式名字服务系统)中得到进一步发展,而且在新型的分布式多媒体组应用领域(如计算机会议系统、远程学习系统、远程会诊系统等)也日益受到人们的重视。本文研究组与组通信的有关概念和机制以及多媒体对组机制提出的新的需求,以期推动组机制的研究与发展。

1 组与组通信机制基本概念

开放分布处理参考模型 RM-ODP 将组定义为一些对象的集合,这些对象由于结构上的原因或者是由于其行为具有共性而集结成组[1]。⟨x⟩组就是由具有特征关系⟨x⟩的多个对象组成的集合。⟨x⟩描述

通过对内和对外两方面使用情况的统计,为我们解决"访问瓶颈"问题和"通信费用"问题提供了基础素材。

5. 展望

分布式信息系统信息组织的优化问题包含诸多因素,本文分析并成功地解决了其中两个主要问题。但事实并非如此简单。随着企业内部网的发展和完善,有可能碰到以下问题:内部网的各个 Web 服务器容量太小,不能容纳与日俱增的信息内容。同时,各个 Web 服务器上每块信息的容量和被访问次数都不相同,问如何将这些信息分布到不同的 Web 服务器上,在不超过各个服务器容量的情况下,使得每个服务器被访问次数尽可能相等?

企业的不断发展,必然会提出新的问题,新的问题需要新的方法来解决。本文提出的数学模型为企业信息规划奠定了解决各种问题的基础,并为解决问题提供了有关思路。

参考文献

[1] Rick Stout 著,阎下兵、卢茂译,World Wide Web 参考大全,海洋出版社,1996

[2] 郁松年、邱伟,组合数学,国防工业出版社,1995

[3] Michael R. Gatey, David S. Johnson, Computers and Intractability: A Guide to the NP-Completeness, W. H. Freeman and Company, San Francisco, 1979

[4] R. L. Graham, Bounds on Multiprocessing Timing Anomalies, SIAM, Appl. Math, 17(2)

[5] B. L. Deuemayer et al., Scheduling to Maximize the Minimum Processor Finish Time in a Multiprocessor System, SIAM, J. Alg. Disc. Math, Vol. 3, 190-196, 1982

*)"九五"国家科技攻关项目 96-B08 资助。

对象之间的结构关系或者预期的对象之间的共同行为。例如，编址组就是由按相同方式编址的多个对象构成的集合；通信组是由多个对象组成的集合，这些对象与其所处环境的交互作用序列完全相同；容错复制组也是一种通信组，其目的是提供对某些故障的一定程度的容错能力。

在使用组机制时，客户希望组机制能够实现透明性，即组服务应该提供与单一服务尽可能完全一致的语法和语义，使客户使用组服务就如同使用单一服务而毋需知道组内部的协调细节；因此，组需要提供必要的机制来协调参与多方联编(multiparty binding)的多个对象之间的交互作用。交互作用组是参与由组管理的多方联编的对象的一个子集。对于联编到同一交互作用组的每个对象集合来说，组需要提供如下几方面的管理功能：

* 交互作用：根据交互作用策略，决定组中的哪些成员参与哪些交互作用。

* 整理：根据整理策略，导出交互作用的一致视图。

* 定序：根据定序策略，确保组成员之间的交互作用排序正确。

* 成员属性：根据成员属性策略，确定成员的故障与恢复、成员的增加与删除。组成员关系在某一特定时刻的"快照"称为组视图。

由于组可以抽象出组成员及其所提供服务的共同特征，可以向客户封装组的内部状态并隐蔽组成员间的相互作用从而向外界提供一致的界面，可以作为构造更大的系统对象的组件块，因此我们可以将相关的对象集结成组。组通信是一种高级的通信抽象机制，可以向应用隐蔽组内部协调工作的细节(如组成员的变化情况)，因而可以简化用户程序和组的交互作用，改善通信效率，提高使用通信机制的方便性。

实现组通信的方法有三种。一是使用网络硬件提供的多点播送能力实现组通信，通过一次多点播送将一个组报文同时提交给多个接收方，可以降低发送方和网络的开销。这是一种比较理想的实现方法，但对分布应用支撑环境的要求较高。二是利用一对一进程间通信机制实现。这需要发送方追踪接收方组成员的变化情况。此外，该方法的发送方和网络的开销大，因为有多少个组成员，发送方就要发送多少次报文。三是使用网络广播机制实现。该方法不仅降低了组通信的安全性(组外用户也能收到组报文)，而且会增加主机开销(所有上网的机器都要增加相应的"过滤器"以检查到来的报文是否是发往自己所属的组)。

组通信机制通常与多点播送机制密切相关，因此在很多文献中两者经常出现混用。我们认为多点播送和组通信是一对需要加以区分的概念。多点播送是一种能将报文从源地发送给多个目的地的网络通信机制，应该由下层基础网络硬件提供相应的支持。同多点播送这种低层网络通信抽象相比，组通信则是一种操作系统级的高层抽象机制，最好有多点播送机制作为其下层实现基础。区分这对概念有助于我们从高层理解组和组通信机制。实际上，组就是由共享相同应用语义、具有相同组标识和(或)同一多点播送地址的对象组成的集合，其中的每一对象都是组的一个成员。

2 组与组通信机制分类

为了更好地理解和应用组与组通信机制，有必要对其进行适当地分类。我们基于客户/服务器模式介绍几种分类标准。

服务器组中各成员之间的通信称为组内通信，客户和服务器组之间的通信称为组间通信，如图1(a)；由于客户也可以是由多个成员组成的组，因此也可以出现客户组对服务器组的通信，如图1(b)所示。

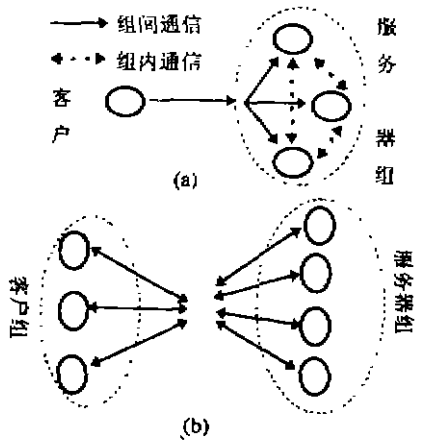

图1 组间通信、组内通信、组对组通信

封闭型组只允许组内通信，不允许外界客户对组进行访问，在国防、公安、银行等安全保密性强的部门有较高的应用价值。开放型组不仅允许组内通

信，而且允许外界客户以适当的方式访问该组，因此应用场合更广泛一些。

文[2]提出了两种比较好的分类标准。第一种是根据每个组成员所管理的数据对象以及所支持的对数据对象的操作而将组分成四类：

·数据与操作同构 DOH(Data and Operation Homogeneous)型组。组中各成员都维护完全相同的数据对象，支持完全相同的操作。该型组的作用主要是提高服务的可靠性和可用性，如采用对等成员模式的完全复制文件系统。

·仅操作同构 OHO(Operation Homogeneous Only)型组，组中各成员所维护的数据对象可能相同、可能不同、可能部分重叠，但各成员支持的对数据对象的操作完全相同。该型组的主要作用是在组成员之间分布工作负载，如部分复制文件系统。

·仅数据同构 DHO(Data Homogeneous Only)型组。组中各成员维护的数据对象完全相同，但各成员所支持的操作可能相同、可能部分相同、也可能完全不同。该型组一般是通过相互协作的工作者进程集合提供组服务，如采用主/从模式的完全复制文件系统。

·异构 Het(Heterogeneous)型组，组中各成员维护的数据对象和支持的操作可能都是不同的。组成员之间可能存在协作，也可能不存在协作。组成员内部状态可能彼此完全独立。该型组的目的主要是为了方便系统控制并且简化客户与服务器组之间的交互作用，而不是为了提供协作式组服务。Usenet 中的新闻组等属于该型组应用。

第二种是根据为客户所见的组的外部行为将组分成确定型组和非确定型组两类。如果组中各成员都必须接收并处理每一个组请求，那么该组就是确定型组。在多数确定型组中，组中各成员是等价的，即各成员在相同的状态下接收相同的请求，调用相同的过程，转换到相同的新状态，并且产生相同的响应和外部影响。对等模式下的完全复制文件系统和复制过程调用等都属于确定型组应用。

非确定型组的各成员不一定是等价的，因而不要求组中各成员都接收并处理每一个组请求。根据应用的具体要求，每个组成员可以依照各自的功能和当前所处的状态对同一个组请求作出不同的响应。非确定型组可以根据特定的应用环境和组本身的性质适当地放松一致性需求，因此维护开销比确定型组低。V 核中的分布式时间服务就属于非确定型组应用。

组是确定型组还是非确定型组取决于使用组机制的应用的特定需求，与其结构特征无关。确定型组主要用于复制数据与服务以增强系统的可靠性；非确定型组主要用于在多个服务器之间分布数据与负载，以实现资源共享并改善服务性能。这两种分类标准也可以分别构成分类矩阵的行和列。例如，对等模式下的完全复制文件系统就属于确定型 DOH 组应用；采用主/从模式的完全复制文件系统属于确定型 DHO 组应用；Usenet 中的新闻发布应用则属于非确定型 Het 组应用。

3 设计组通信机制应该解决的问题

由于组机制的客户通常并不关心组的内部结构而只是关心组机制对客户所呈现的外部属性，因此下面我们就针对确定型和非确定型组来讨论设计组与组通信机制时应该解决的问题。

3.1 确定型组通信机制

确定型组需要完整的组视图以便在组成员之间进行协调与同步，维护强一致性，因此要求组通信具有高可靠性。在设计确定型组时，组通信机制应该满足如下要求：

·组报文提交的原子性和有序性　原子性是指一个组报文要么为服务器组中的所有成员正好接收并处理一次，要么就根本未被服务器组的任何成员接收与处理。原子性通过将部分失败转换成完全失败，向客户隐蔽部分失败的发生。由于组通信系统向组成员提交报文的顺序可能会影响组成员所维护的数据对象在报文处理之后所处的状态，因此组通信系统需要对报文在各个组成员处的提交顺序进行同步。组中各成员需要看到相同的对相关数据的请求序列，并且据此调整其内部状态。因此，报文的提交需要具有有序性。

·应答处理透明性　由于客户和服务器之间的交互作用一般都遵循请求-响应模式，因此只有从客户到服务器的可靠组通信是不够的。要想实现组通信的透明性，还必须正确地收集和处理来自服务器组各成员的应答。这些应答可能完全相同，可能完全不同，可能部分相同。应答处理透明性确保客户毋需知道对其请求可能存在多个应答的事实。客户只看到一个应答，而不必关心该应答是如何导出的(例

如，通过加权投票）。

·组命名透明性　这涉及到既动态又透明地将组成员联编到同一个组名字上。组命名机制应该提供如下两种功能：组名对服务器组中各成员的映射；组成员关系动态变化的管理机制。随着成员的故障与恢复、插入与删除等情况的出现，组视图在不断地发生变化，同其他组报文活动并行进行。因此，有关各方都要参与组视图的维护。

由于组报文提交的原子性需要依靠一致的组视图来验证是否所有活跃成员都已确认对每个原子组报文的接收，因此必须以一致的方式检测组视图的变化。在 Isis 系统中，当有某个组成员出现故障时，系统就代表该故障成员发出一个通告，该通告紧随该成员在发生故障之前发出的所有报文之后。相对于其他组报文，该通告以相同的顺序到达每一个成员，这样，所有成员就在同一"虚拟"时间看到相同的组视图变迁序列。因此，Isis 系统可以确保组成员在收到有关某个成员发生故障的通告之后，就再不会收到来自该故障成员的报文。如果故障成员从故障中恢复过来，那么其状态必须及时更新，以便同所属组的最新视图保持一致。

·故障透明性　在发生故障时，系统根据组的意图进行适当的故障处理。当组是用于增强数据可靠性时，如复制文件系统，由于需要在组成员之间维持文件复制品的强一致性，因此通常需要将部分失败放大成完全失败。但是，当组是用来增强服务的可靠性时，例如复制过程调用，维护对客户的服务就是至关重要的；在这种情况下，在故障成员恢复期间，其他活跃组成员应该继续各司其职，以便将损害降低到最低点。

·实时需求　在确定型组应用中，一致性通常比及时性更为重要。当需要在两者之间进行权衡时，系统通常赋予一致性更高的优先级。但是，在实时系统中，报文通常必须在客户指定的最后期限内提交；否则，报文就会过时，系统将触发定时故障。在一对一客户/服务器应用中，服务器可以简单地忽略定时故障报文或者向客户作出操作失败响应。在客户/服务器组应用中，由于网络和服务器负载等因素的影响，有些服务器可能及时收到了报文，而其他服务器看到的却是定时故障；即使是确保原子性和有序性，也可能如此。在发生这种情况时，服务器组中的各成员必须相互协调以便对每个定时故障报文作出一致响应。组报文的传送时间应该有界，以便系统可以对组操作进行适当的调度，保证组操作可在同一虚拟时间在所有组成员处原子地进行。

3.2　非确定型组通信机制

非确定型组可以根据应用的实际需要以各种特定于应用的方式放松对一致性的要求，因此通常只需要基本的多点播送数据报传送支持。

·组报文提交的原子性和有序性　在非确定型组应用中，请求/应答依然是主要的通信模式，但是通常不需要绝对可靠的报文提交或报文定序。在这类应用中，要求客户在服务器组的所有成员达到同步并且准备好接收请求之前一直处于等待状态通常既不必要也不实际。客户通常也不知道服务器组的成员关系情况。有的服务器可能根本就没接收请求。因此，非确定型组通信从本质上讲是异步的。应用本身可能具备方便地检测不一致性并从中予以恢复的能力。数据的一致性也不再是一个非常严重的问题。在设计非确定型组应用时，灵活性较强，但要增加处理部分失败的复杂性。

·应答处理透明性　来自非确定型组的多个应答可能是不一致的，通常由客户以特定于应用的方式处理，因而会在一定程度上牺牲应答处理的透明性。

·组命名透明性　非确定型组应用也希望使用逻辑名来标识组服务，因此同样希望能以独立于组成员数目的方式来处理请求和应答。然而，非确定型组的成员关系可能动态地发生变化，而且这种变化通常并不为当前活跃的组成员所知，这就相应地增加了应答处理透明性的实现难度。

·故障透明性　非确定型组成员故障具有同确定型组成员故障不同的语义。在某个成员出现故障时，通常是由其他活跃成员透明地接管该故障成员未完成的任务以增强服务的可用性。

·负载均衡　在非确定型组中，由于缺乏适当的内部协调，因此组成员之间的工作负载可能非常不平衡，需要在应用级提供适当的机制予以解决。

4　多媒体对组机制提出的新要求

随着高速网络技术和多媒体工作站技术的不断进步，出现了很多新型的分布式多媒体组应用，对组机制提出了新的挑战。

多媒体数据可以分成静态媒体和连续媒体两

类。静态媒体(如文本)是没有时间维的媒体,播放速度不会影响所含信息的再现。静态媒体也称离散媒体。连续媒体(如视频和音频)是由媒体"量子"(如视频帧和音频采样)组成的,具有隐含的时间维,需要在一段特定的时间里按特定的速度播放;如果播放速度得不到满足,媒体信息的完整性就会受到影响。实际上,多媒体对组机制提出的新的挑战主要来源于连续媒体。

·高带宽　同传统数据相比,多媒体数据、特别是连续媒体数据对网络的带宽需求高。例如,一条CD质量的音频连接需要1.4M比特/秒的带宽(2个声道,每个声道的采样速率为44.1KHz,每次采样16比特),而一条未经压缩的NTSC制式的视频连接需要高达27M字节/秒的带宽(每秒30帧,每帧640*480象素,每个象素24比特)。因此,对于分布式多媒体组应用来说,不仅需要传统的组管理功能和组通信机制,而且需要有高速网络作为其下层基础支撑环境,才能满足其高带宽需求。

·实时性与等时性　连续媒体数据的录制、访问、传送与播放过程具有实时性和等时性。例如,NTSC质量的视频应用不仅要求每秒产生30个视频帧而且应该是每隔1/30秒产生1个视频帧。这就要求通信机制不仅要支持连续媒体的实时传送,而且要支持连续媒体的等时传送,保证媒体内时间连续性的实现。同传统的滑动窗口机制相比,基于速率的传送机制是一种比较好的选择,因为基于速率的机制不仅可以建立连续媒体数据流与速率受控传送之间的自然对应关系,使发送方的通信量平稳地进入网络以便与接收方的处理能力相匹配,而且可以将流控和差错控制机制解耦。

·差错控制　由于视频和音频之类的连续媒体数据具有内在的信息冗余性,因此对于多数多媒体组应用(如远程培训)来说,偶尔丢失几个视频帧或音频采样,或者是视频帧和音频采样在传送过程中出现少量的比特错,都不会严重影响信息的可用性,通常依然能够为用户所接受。当然,也有一些多媒体组应用需要信息(如远程会诊中的患者X射线影像)的实时与等时无差错传送。因此,需要采用面向应用的差错控制策略,允许用户对差错检测、差错指示和差错纠正进行适当的组合,满足应用的特定需求。还应指出的是,超时重传机制根本不适应连续媒体通信的实时性和等时性要求,因此在多媒体应用中广泛采用的是前向纠错模式。

·压缩和解压缩　多媒体、特别是连续媒体信息源产生的实时数据量非常大,直接进行传送或存储对网络带宽和存储空间的压力太大。因此,在网络上传送或在多媒体计算机上存储多媒体数据之前,一般都要在信息源对其进行压缩,而在目的地解压缩后播出。压缩和解压缩的标准很多,如JPEG、MPEG系列等。应该指出的是,压缩技术虽然大大降低了多媒体数据对网络带宽和存储空间的要求,但同时也大大降低了多媒体数据的内在冗余性,因此相应地增加了差错控制的实现难度。

·媒体间同步　在多媒体应用中,还必须解决不同媒体间的同步问题。例如,在远程播放一部影片时,不仅需要维持视频和音频信号本身的时间连续性,而且需要在视频与音频信号之间进行严格的同步,维持严格的"对口型"关系。

·服务质量异构　服务质量QoS(Quality of Service)是指服务性能的聚集效应,它决定用户对特定服务的满意程度。在分布式多媒体组应用中,由于各成员使用的端系统的能力(CPU处理能力、显示器的分辨率与灰度级、存储器的速度与容量等)、所上网络(ATM、FDDI、FDDI-II、IBM令牌环、Ethernet、快速Ethernet等)的能力以及所愿支付的费用不同(通常服务质量越高,收费也越高),因此在不同成员处实际达到的QoS水平应该有所不同。如何在同一分布式多媒体组应用中满足不同组成员的异构QoS需求依然是一个值得我们深入研究的问题。

结束语　组和组通信机制是分布式系统中的一个非常活跃的研究领域,人们已经取得了很多研究成果,为今后的研究工作打下了良好的基础。多媒体信息处理技术与组机制的结合为分布式系统的研究与发展注入了新的活力,同时也提出了新的挑战。这就需要我们进一步研究与探索,以推动组与组通信机制的发展。

参考文献

[1] Draft Recommendation ITU-T X.901/ISO 107461: Basic Reference Model of Open Distributed Processing-Part-1:Overview

[2] Luping Liang et al., Process Groups and Group Communications: Classifications and Requirements, IEEE Computer, Feb. 1990