基于半主动复制技术的组通信系统*)

袁 援 陈松乔 陈建二
(中南大学信息科学与工程学院计算机系 长沙410083)

Group Communication System Based on Semi-Active Replication

YUAN Yuan CHEN Song-Qiao CHEN Jian-Er
(Department of Computer, College of Information Science and Engineering, Zhongnan Univeristy, Changsha 410083)

Abstract Both active replication and passive replication in group communication system are limited to implementing the false-tolerance distributed system. In this paper, we have presented a duplicate technique called Semi-Active replication SAR, which takes in the advantage of two former techniques. As taking events log as synchronous object, employing reliable multicast as synchronous way between copies, and using the substitution method in passive replication, this new method implements the fault control to the processors in group communication system on the premise of assuring the consistence between duplicates and deceasing overhead of the system. It is best answer for large distributed system. Finally we apply the method in the distributed applications based on EJB specifications. It is shown that SAG provides a good solution for the false-tolerance distributed system where the communication model between clients and servers is point-to-point.

Keywords False-tolerance, Group communication system, Semi-active replication, Duplicate, EJB

1 引言

为提高分布式系统的可用性,研究者大都在系统中引入冗余,组通信技术是应用最广的一种空间冗余技术。其基本思想是在某一处理机组中备份客户请求,当某处理机出现错误时,副本组向客户提供错误屏蔽。通常将使用组通信的系统称为组通信系统(Group Communication System GCS),GCS 实现的关键是保证副本组副本间消息传递的同步,其实现有两种方式:主动复制和被动复制[1]。

主动复制技术采用组播实现客户向副本组成员的请求投递,副本组各成员执行请求后,将结果返回客户,客户首次得到结果后结束任务。主动副本技术采用虚拟同步通信[2]来保证副本组副本间事件投递的原子性和全序(Total Order)性,向副本组成员提供了全局一致的保障,因此是一种提供系统高可用性的复制技术。但主动副本技术要求各副本必须在每个事件上达到一致,所以采用主动副本技术的分布式系统中,副本组的全局一致性是以系统通信开销的加重为代价的。采用被动(主/从)复制技术,客户同处理机维持单点联系,只有被系统选出的主副本执行客户请求,从副本不执行客户请求而只保留请求,当主副本出错时,系统采用副本替换技术将选定从副本更新为主副本,执行客户请求。被动复制技术虽然可以在一定程度上克服主动复制技术系统通信开销过大的弊病,却不能保证副本间的同步。因此,两者都不大适用于大型分布式应用的容错设计。

而本文我们在给出 GCS 模型的前提下,结合两种基本复制技术,给出了一种适合于大型分布式系统容错应用的半主动复制技术。该技术的优点是:在保证副本同步的前提下,系统为容错所付出的通信开销较主动复制技术小。我们将其应用到 Enterprise JavaBeans 上的结果表明,本技术还为客户同处理机只能维持单点通信且用户有较高可用性要求的容错系统提供了解决方案。

2 GCS 模型

本文所考虑的 GCS 是一个消息传送的通讯模式。对系统模型的基本假设为:各处理机向客户提供完全相同的服务;处理机不共享内存,而是通过可靠的信道以 RPC 来进行相互通信,消息的传递方式采用组播(multicast);系统中的处理机遵从崩溃/失效模型,即处理机由于崩溃而失效;系统对处理机间的消息传递延时无限制,对处理机的执行速度无限制。

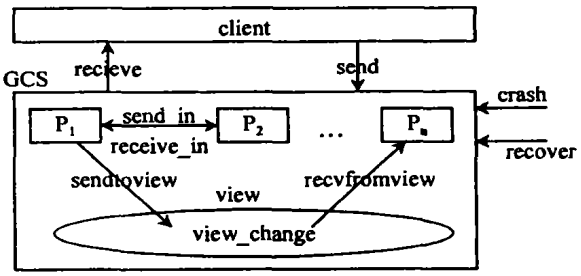

图1 GCS 的基本活动

系统交互活动是分布式系统运作的基础。组通信系统作为一种分布式系统,其基本活动包括:系统中同分布式应用以及环境之间发生的外部活动;系统内部各处理机之间发生的内部活动。GCS 的基本活动如图1所示。

定义1(组通信系统) 组通信系统被描述为三元组{P, M, VID}。其中,P—处理机集合;M—所传递的消息集合;VID—视图标示集合。下文中,我们采用集合论的方法来描述系统基本活动。

GCS 外部活动的一种形式是系统同分布式应用之间的

*)国家教育部重点科技课题(2000156)和海外杰出青年科学基金(69928201)资助项目。袁 援 博士研究生,主要研究领域为分布式计算及其通信机制。陈松乔 教授,博士生导师,主要研究领域为分布式软件系统。陈建二 教授,博士生导师,主要研究领域为容错计算和网络通信协议。

交互作用，包括应用向GCS发送消息和从GCS接收消息，发送消息函数 $send(p, m)$，$p\in P$，$m\in M$，接收消息函数 $receive(p, m)$，$p\in P$，$m\in M$；另一种形式是系统同外部环境之间的交互作用，包括 $crash(p)$，$p\in P$，$recover(p)$，$p\in P$。

GCS内部活动包括：

(1)采用面向视图的组通信服务与组成员关系变化密切相关，为了体现这种变化，引入视图变化函数 $view\_change(p, v)$，$p\in P$，$v\in VID\times 2^P$，$id\in VID$，$members\in 2^P$，$v=\langle id, numbers\rangle$。其中，$v$—活动中所传送的视图集，$id$—视图标示，$members$—视图的集合；

(2)处理机 p_1向 p_2发送消息 m：$Send\_in(p_1, p_2, m)$，$p_1, p_2\in P$，$m\in M$

(3)处理机 p_1接收到 p_2的消息 m：$Receive\_in(p_1, p_2, m)$，$p_1, p_2\in P$，$m\in M$

(4)处理机 p 将消息 m 送往视图 v：$Sendtoview(p, m, v)$，$p\in P$，$m\in M$，$v\in VID\times 2^P$

(5)处理机 p 从视图 v 接收到消息 m：$Recvfromview(p, m, v)$，$p\in P$，$m\in M$，$v\in VID\times 2^P$。

3 半主动复制

3.1 基本思想

为了克服主动复制和被动复制在分布式应用中的不足，我们给出一种对两者进行折中的副本复制方法——半主动复制(Semi-Active Replication SAR)。其基本思想是：将副本组织为一个视图同步组，GCS指定副本组中的某一成员为主副本(Leader Duplicate)，其它的副本则称为从副本(Follower Duplicate)，主副本执行客户请求，从副本保留但并不执行客户请求；当主副本正常工作时，主副本将载有事件的事件日志组播给从副本，实现副本间的同步采用全序Multicast方法，主副本在执行完客户请求后将处理结果返回客户；当主副本发生错误时，GCS采用被动复制中的副本替换技术，选取某个从副本为主副本，当前主副本在接收到失效主副本事件的基础上执行客户请求。

SAR首先在副本组中指定主副本，并且在主副本失效时，采用副本替换技术，以实现对错误的屏蔽，但在实现副本同步的时候采用全序组播的方式，因此系统在更替副本的时候，不会出现在被动复制中由于更新延迟产生的副本不一致现象，同时，它不像主动复制一样需要维持所有副本的全局一致性，且在执行过程中，只有主副本参与用户请求的处理，系统的通信负载不像采用主动复制那样重。因此，SAR是一种吸取两种复制技术长处的折中方法。

3.2 副本同步

主动副本技术的实现以副本同步传输对象的原子性和全序一致为前提[2]。它必须以事件为基本传输单位，即为了保证GCS的全局一致性，副本间必须实现所有事件的全序一致，这是主动复制技术通信开销巨大的主要原因。为了减小GCS的通信开销，在SAR中，我们将副本间的同步对象确定为事件日志(Events Log EL)[3]来进行副本间的全序一致性同步，因此，本技术中的副本同步指副本间事件日志的同步。

事件日志是副本对其上发生事件的记录，它确定了事件的执行顺序。主副本首先将事件记录入事件日志中，然后将EL组播给从副本，从副本在ML上取得同步。当EL到达从副本时，从副本执行消息日志中的事件，直至系统返回客户请求。由此所有副本具有相同的执行经历，最终状态将以相同的顺序出现在所有副本上。ML的同步是一种相对简单的顺序执行方法，因为它并不能保证副本间传输的原子性，大粒度消息在同步过程中的丢包概率更大，所以这种方法并不一定能够保证副本的完全顺序一致，但处理机处理消息日志中事件的执行顺序同记录顺序完全相同，因此消息日志方法只保证事件处理顺序的一致而不能实现事件传输顺序的一致。

3.3 系统容错

当处理机出错时，不失一般，假设出错的处理机是主副本所在的处理机，分两种情况进行讨论：

(1)在出错的时候，主副本所在的处理机未将EL组播到从副本，此时客户只能重新请求处理机。

(2)在出错的时候，主副本所在的处理机已将EL组播到各从副本，主副本首先将记录入EL中的事件挂起，从副本所在的处理机收到EL后，GCS选取某一从副本，并将其升级为主副本。当前主副本在执行完被挂起的事件后，将发生在其上的事件记录入其EL中，从副本并向其后继传送EL，即进入处理机正常工作状态，且主副本被排除在副本组之外。

鉴于以上两种情况，为方便讨论，假设错误发生在副本消息同步之后，图2给出了SAR的一个容错实例。

客户首先将请求reqA组播到副本组所有成员，GCS选取L为主副本，L在对reqA进行处理之前，将全部事件写入事件日志，并采用全序组播将其发送到从副本 F_1、F_2，然后向方法库发送远程请求ReqB并等待执行结果。但L在发送reqB后出错，L将全部事件挂起，L的错误引起视图变化，系统将L排除在副本组之外，F_1和 F_2在接收到视图变化消息后，系统在从副本中指定一副本为当前主副本(图中为 F_1，标记灰色者为当前主副本)，F_1重做事件日志中的事件，并从状态 S_1后(当前主副本在先前主副本同其之间的最后通信点处恢复执行)接续L执行客户请求reqA，如果 F_1在调用执行方法后没有发生错误，则直接向客户返回请求执行结果。图中 S_1、S_2、S_3分别对应主副本出错前状态、主副本出错后被排除到副本组外的状态、系统选取一个从副本并将其升级为主副本后的状态。

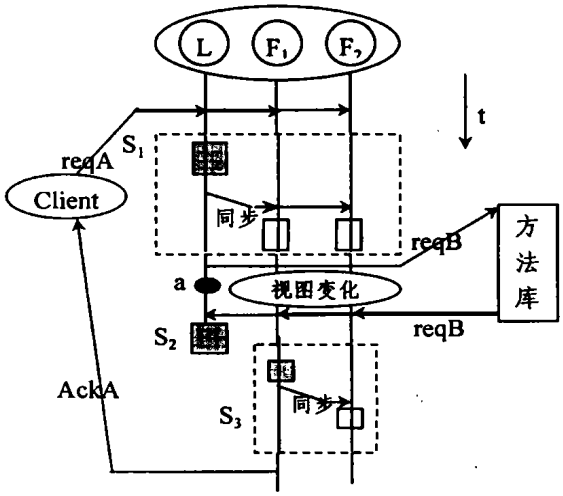

图2 SAR的一个容错实例

通过对上述容错实例的描述，可以看出使用SAR的GCS系统开销较主动复制技术小的原因在于：其一，系统指定一个主副本来执行客户请求，减小了处理机负荷；其二，副本间的同步内容是消息日志，而不是顺序事件，减小了处理机

(下转第183页)

有相同的带宽要求，但应用于短消息时ECC带宽要求却低得多。而公钥加密系统多用于短消息，例如用于数字签名和用于对称系统的会话密钥传递。带宽要求低使ECC在无线网络领域具有广泛的应用前景。

结论 目前，D－H密钥交换算法的专利已过期，RSA算法的专利期限也将面临结束，取而代之的是基于椭圆曲线的密码方案。ECC的这些特点使它必将取代RSA、D－H，成为通用的公钥加密算法。比如SET（Secure Electronic Transactions）协议的制定者已把它作为下一代SET协议中缺省的公钥密码算法。

参考文献

1 高品均，陈荣良. 加密算法与密钥管理[J]. 机界计算，2000，12：7～9

2 张克友，聂规划. NOVELL网的安全策略研究[J]. 电脑开发与应用，2001，14(10)：36～34

3 曾学蕾. 密码学的新方向[J]. 计算机应用，2002(12)：23～27

4 吴文玲，贺也平. 欧洲21世纪数据加密标准候选算法简评[J]. 软件学报，2001，12(1)：49～55

5 朱幼莲. 逻辑函数的计算机化简[J]. 计算机应用与软件，2003(2)：52～54

（上接第132页）

间的通信开销；第三，系统保证只有一个副本返回客户结果，减小了系统同外部应用间信道载荷。因此，同主动复制技术相比，SAR技术减小了整个系统的通信和处理开销，可以提高GCS性能，它特别适用于处理量大、消息交互频繁、对副本一致性要求不太严格的分布式应用。

4 SAR的一个分布式应用

现存的分布式应用大都采用了副本技术实现系统的容错。例如，CORBA定制了容错规范，CORBA2.2以上版本专门提出了容错CORBA[4]的概念，基于容错CORBA的GCS采用主动复制来提高分布式系统的可用性。而另外一种比较流行的中间件规范是Sun MicroSystem公司的EJB规范，在最新的EJB 2.0[5]版本中还没有正式的EJB容错规范，因此对EJB系统容错的探讨是一个新领域。而EJB是基于Server端构件的规范，因此提高基于EJB分布式系统的容错能力显得尤为重要，否则系统将出现由于Server崩溃而不可用的情况。一个直观的想法是采用CORBA规范中定制的组通信系统，但基于主动复制的组通信规范只适合于客户同处理机是多点通信的情形，而在EJB规范中，客户同处理机的交互是基于RMI-IIOP的点-点通信[5]，如果采用主从复制技术将不可能实现基于EJB规范的分布式应用容错，而我们认为本文提出的半主动副本技术是上述问题的一个解决方案。

客户同EJB服务器交互的基本过程为：客户stub通过JNDI找到Server（称为主服务器）后，客户向服务器发送RMI，服务器通过skeleton向EJB容器寻求方法调用，EJB容器将方法返回Server，Server在处理客户请求后，将结果返回客户。当主Server崩溃时，若系统是采用半主动复制技术进行错误屏蔽，有以下两种情形发生：

（1）当Server的崩溃发生在客户请求被备份之前，客户只有重新寻找Server。

（2）当客户请求已经被备份之后发生Server崩溃。如图3，类似于SAR应用于客户同处理机多点通信的情形（图2），但主副本向从副本发送的事件日志中，除了在主副本上所发生的事件外，还包含应用请求的信息，标记灰色者为当前主副本所在服务器。系统容错过程同3.3节所述相似，不再赘述。

由上可知，在这种客户同处理机点-点通信的分布式系统中，主动复制技术不再适用，而主/从复制技术的应用是一个可行方案，但对于有高可用性要求的系统而言，如上文提到的基于EJB规范的分布式应用，半主动复制技术是当然的首选。

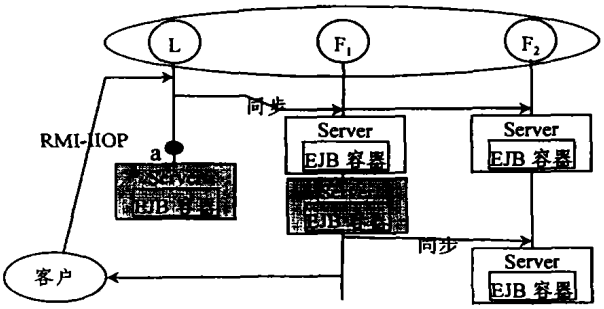

图3 SAR应用到基于EJB分布式系统容错

结束语 主动复制技术和被动复制技术在实现分布式系统容错时都有其不足之处，文中给出了一种兼具二者优点的半同步复制技术。此技术以消息为副本同步对象，在采用可靠的组播通信的前提下，应用被动复制的副本更替方案，很好地迎合了大型分布式应用对容错的需求。另外，我们将半主动复制技术应用于基于EJB规范中间件技术的容错问题，结果表明，此技术向此类分布式系统容错提供了一种很好的解决方案。因此，我们认为它是一种极具前景的分布式容错技术，对其的研究将成为分布式容错技术研究领域的热点。

文中并没有讨论对副本崩溃进行恢复的问题，这是我们下一阶段的研究工作。此外，在往后的研究工作中，我们还将致力于对SAR的副本同步做出更精确的描述、视图变化触发器等容错器件的设计等问题的研究，为此技术的实际应用奠定良好基础。

参考文献

1 Babaoglu O, et al. Group Communication in Partitionable Systems: Specification and Algorithms[J]. IEEE Transactions on Software Engineering, 2001, 27(4): 308～331

1 史殿习，等. 组通信中虚拟同步协议的研究与设计. 计算机研究与发展，2000，37(10)：1192～1196

3 Pedone F, et al. Exploiting Atomic Broadcast in Replicated Database. Processings of EuroPar. Southampton, England, Sep. 1998

4 CORBA 2.0 specification. http://www.omg.org

5 EJB 2.0 specification. http://java.sun.com/