第 32 卷 第 19 期 计 算 机 工 程 2006 年 10 月
Vol.32 № 19 Computer Engineering October 2006
·软件技术与数据库· 文章编号：1000—3428(2006)19—0111—03 文献标识码：A 中图分类号：TP316

分布式内存管理系统的研究与设计

王育坚[1]，刘 辰[2]，田 星[2]
(1. 北京联合大学信息学院，北京 100101；2. 北京邮电大学计算机科学与技术学院，北京 100876)

摘 要： 针对中小型企业数据处理的特点，提出了一种基于局域网的分布式内存管理系统，介绍了系统的数据组织方式、设计原理和各功能实体的具体实现。系统利用局域网内存资源存储数据，解决了磁盘输入/输出的性能瓶颈问题。把数据分成包，通过包管理实体对内存资源进行管理，利用同步协议机制保持主包和备包的同步。测试结果表明，利用系统对数据进行处理的效率是本机硬盘的 5 到 7 倍。
关键词： 分布式计算；内存管理；包；开放最短路径优先

Research and Design of Distributed Memory Management System

WANG Yujian[1], LIU Chen[2], TIAN Xing[2]
(1. Information College, Beijing Union University, Beijing 100101;
2. College of Computer Science & Technology, Beijing University of Posts and Telecommunications, Beijing 100876)

【Abstract】 According to the characteristics of data processing in the middle-small enterprises, a distributed memory management system based on LAN is proposed. The manner of data organizing, the design principle and the concrete implementation of every function entity in the system are introduced. The system uses the memory resource in the LAN to store data, solves the problem about I/O capability bottleneck. The data is divided into packages, and the package management entity is used for the management of memory resource. Main package and backup package are kept synchronization by synchronous protocol mechanism. The results of test show that the efficiency of data processing by using the system is 5 to 7 times as high as that by using the local fixed disk.
【Key words】 Distributed computing; Memory management; Package; Open shortest path first(OSPF)

分布式文件管理技术的应用愈来愈广泛，如文献[1]给出的分布式文件服务器和文献[2]给出的应用于因特网的分布式只读文件系统。但由于文件存储在磁盘上，其低效的 I/O 操作严重影响了文件管理系统作用的发挥。随着计算机内存技术和网络技术的发展，网络带宽和速度得到了大幅度的提高，一台计算机对分布于局域网中其它计算机内存的存取速度已经远远高于本机硬盘。研究和设计基于分布式技术的内存管理系统，特别是针对中小型企业局域网的分布式内存管理系统，将在很大程度上弥补文件管理系统的不足。中小型企业数据处理量不是很大，但用户访问频繁，且对数据的实时更新速度要求较高。在现有条件下，计算机和网络存在不稳定性，可能发生电源和网络的中断。因此，利用中小型企业的计算机和网络资源，提供一个可靠性好、维护方便和效率高的内存管理系统具有很大的实际意义。

本文提出的分布式内存管理系统(DMMS)在理论上接近于分布式数据库系统[3]，但DMMS把数据分布存储在分布式系统的内存介质中，在速度和性能方面比理论上的分布式数据库系统要好得多。同时，在数据的组织、结构和保护机制等方面，DMMS吸取了分布式数据库的优点，弥补了内存数据库的不足。设计内存管理系统，除了选择内存分配的策略，还必须考虑内存保护和内存回收的方法[4]。DMMS采用了开放最短路径优先(OSPF)的基本思想，通过包管理实体实现了内存数据的保护机制。

1 数据组织和工作原理

数据在 DMMS 中的组织结构分为应用数据、包数据、表数据或简单数据 3 个层次，如图 1 所示。这种数据分层结构也代表了系统中 3 个数据使用层次，即应用数据是原来存储在计算机中的用户数据，而把应用数据根据数据分包的协议打成一些包，包中的数据又被分为表数据和简单数据两种不同格式。表数据表示该数据是一个表，由一些记录组成。简单数据表示该数据是单个数据，是一些简单的值对。

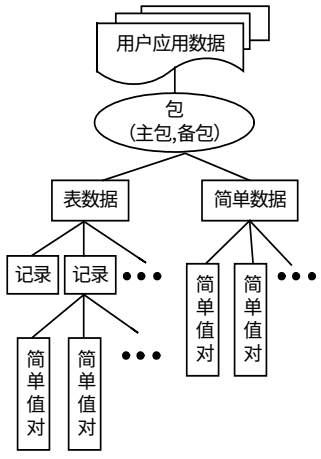

图 1 数据组织方式

包是内存数据存储、定位、保护的基本单位。一个包需

基金项目：北京市教育委员会科技发展计划基金资助项目(KM200611417001)
作者简介：王育坚(1963－)，男，副教授，主研方向：软件理论，分布式应用；刘 辰，副教授；田 星，硕士生
收稿日期：2005-10-20 **E-mail：**xxtyujian@buu.com.cn

要存储在两个不同 IP 地址的机器上，以保证一个包损坏后还有另一个内容相同的备份包能够提供使用。逻辑上的一个包对应一个主包和一个备包，主包和备包一直保持同步。当主包出现故障，启动备包成为主包，同时再生成一个备包；当备包出现故障，主包就再生成一个备包。DMMS 通过主包和备包这种相互保护机制实现包中数据的保护。

整个系统的设计采取灵活的框架结构，系统各个部分功能相对独立，以功能实体来实现。设计完成的协议和算法都被封装起来，可以随时替换，以实现灵活性和可扩展性。系统工作原理如图 2 所示，用户通过 API 接口向智能客户端发出操作指令，智能客户端根据用户指令进行操作，并将操作结果返回给用户。包管理实体运行在局域网内每台机器上，这些包管理实体通过协议机制协同对数据包进行管理。包定义了包的基本数据格式、属性和方法，包地址表提供了包的地址信息。

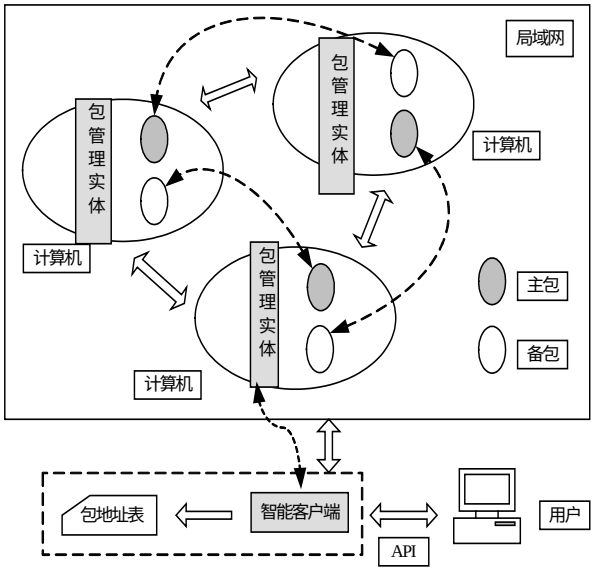

图 2 DMMS 工作原理

2 协议机制

内存管理需要一个有效的数据保护和备份机制，DMMS 数据保护机制设计分为如下两个层次：数据本身的保护机制和网络的保护机制[5]。对于数据本身的保护机制，系统将数据打包(主包)并备份(备包)，主包和备包分别存放在不同机器的内存中，利用同步协议机制一直保持同步，通过主包和备包的相互保护机制实现包内数据的保护。主包和备包的同步协议可以由主包发起，也可以由备包发起。例如，当备包作为同步协议发起方时，备包定时向主包发出同步申请。如果主包忙则返回忙消息，备包接到消息继续等待；如果主包不忙但主包数据没有改变，则主包返回数据没有变化的消息，备包接到消息继续等待；如果主包不忙且主包数据有改变，则主包发出开始同步消息，同时拒绝主包的其它操作，备包接到同步消息就开始同步操作，同时也拒绝其它操作。同步完成后主包向备包发出同步完成消息，同时可以接受其它操作，备包接到消息后结束对主包的访问，同时也可以接受其它操作。

对于网络的保护机制，系统采用开放最短路径优先 OSPF[6]的基本思想来设计局域网计算机之间的自适应协议，其基础是最短路径优先(SPF)路由算法。SPF路由器初始化路由协议数据结构，然后等待下层协议有关接口已可用的通知信息。当路由器确认接口已准备好，就通过OSPF Hello协议来获取邻接点信息。每个路由器周期性地发送链路状态公告LSA数据包，提供其邻接点的信息，且当其状态改变时通知其它路由器。通过对已建立的邻接关系和链路状态进行比较，失效的路由器可以很快地被检测出来，网络拓扑相应地进行变动。从LSA生成的拓扑数据库中，每个路由器求出最短路径树，由此生成路由表。

DMMS 将局域网内的计算机划分为 3 类：一个队长，一个副队长，若干队员。队长负责发布和维护数据表格，副队长、队员负责申请数据表格和刷新数据表格。这些计算机上都保存了最新的网络资源状态表，通过一个判决机制来确定网络中各个机器内存资源的可用性。当队长机出现故障时，副队长就升为队长，并从队员中选一个作副队长，且由新队长发布最新的网络资源状态表；当副队长机出现故障时，队长从队员中选出一个副队长，发布最新的网络资源状态表；当队员出现故障时，队长直接更新并发布最新的网络资源状态表。整个网络通过这种队长、副队长和队员的角色转换策略来保证内存资源的安全、可用。

3 系统功能实体设计

针对内存管理的特点，系统在功能实体的设计中主要考虑以下内存管理策略[7]：尽量扩大常规内存；利用虚拟存储技术进行数据的内存、外存转存；利用覆盖技术解决程序大小超过实际主存的问题；利用交换技术解决内存的不足；利用“空洞”内存建立高速磁盘缓存。除了图 2 中的智能客户端、包管理实体、包和包地址表，系统还包括包数据状态表和包数据访问接口等功能实体。

3.1 智能客户端

智能客户端以 API 的方式实现了包数据的定义、读取、修改和删除等方法，通过包地址表获取数据的存储地址，从而对存储在局域网中某台计算机上的包进行操作。智能客户端对包进行操作时还必须利用包管理实体提供的信息，如局域网内是否还有可用的内存资源，可用内存资源的数量和分布情况等。

智能客户端屏蔽了系统分布式数据管理和操作的复杂细节，用户只需要调用 API 就可以透明地使用局域网内存资源。DMMS 智能客户端定义的方法主要有判断访问主包或备包，增加表数据或简单数据，删除表数据或简单数据，读取表数据或简单数据，修改表数据或简单数据。

3.2 包管理实体

包管理实体是 DMMS 的核心功能实体，它通过动态调整的方法对网络内存资源进行维护和管理，如确定网络可用内存资源的大小和位置，获取每一台计算机的机器状态(机器正常或机器故障)和内存状态(容量未过门限、容量已满或容量超载需均衡)，发布最新的网络资源状态，保证网络中机器和内存资源的可用。

对于包的管理和分布于网络中各机器的管理，包管理实体采用了Socket通信机制，利用多线程技术实现多个实体的同时交互[8]。Socket运行在双向通信的每一端，它既可以接受请求，也可以发送请求。系统借助于Java类库JFC中的Socket类提供的方法来处理用户的请求和响应。Java提供了支持网络且与平台无关的软件包Java.net，Java有关网络的类及接口定义在Java.net包中。客户端程序通常使用Java.net包中的Socket类与服务器建立连接。服务器程序不同于客户端程序，

需要初始化一个端口进行监听，遇到连接呼叫，才与相应的客户机建立连接。服务器程序主要使用Java.net包提供的ServerSocket类。

包管理实体程序运行在局域网内的每台机器上，每个运行包管理实体的机器既是客户机也是服务器。对于网络其它机器而言它是客户机，如果其它机器和它通信，它就是服务器，因此每个包管理实体同时运行了客户端程序和服务器程序。客户端程序不断地发出心跳信息告诉其它机器自己的存在，服务器程序采用多线程机制，接收其它机器发出的心跳信息及故障信息。

DMMS 用包管理实体类 PkgMagEntity 来实现包管理协议的所有内容，PkgMagEntity 按照包管理实体之间的协议实现对包的管理。包管理实体协议主要协调队长、副队长和队员 3 种角色。系统整个设计采用了统一建模语言 UML，图 3 给出了用 UML 描述的包管理实体状态图。

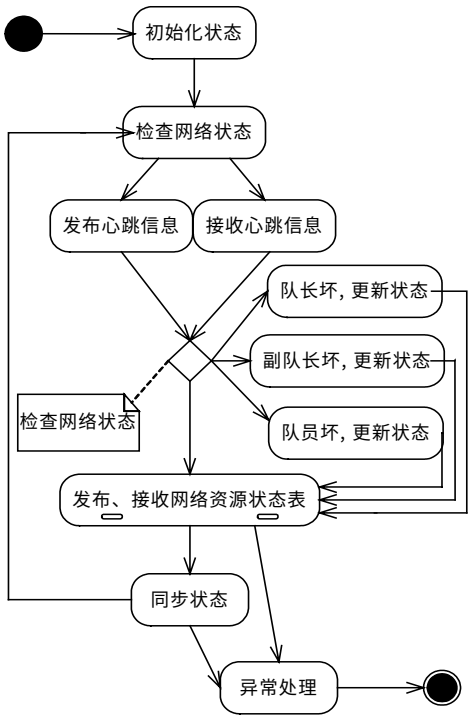

图 3 包管理实体状态

3.3 包和包其它数据实体

DMMS 中与包有关的数据实体包括包、包数据状态表、包地址表和包数据访问接口。包以类的定义来实现，包括包类、主包类和备包类。包类(Pkg)定义了包的共有属性和方法，是其它包类的父类。主包类(MainPkg)主要提供数据访问功能，备包类(BakPkg)主要提供数据备份功能。

包数据状态表供包管理实体使用，用以提供包数据当前的状态信息，其组成如图 4 所示。包的名称是数据包的 ID，包的 IP 地址是指包所在机器的 IP 地址，包端口是指包所在机器的实例占用的端口号，包的主备表示包是主包还是备包，包同步数指包同步时所用的标识同步进程的数字，包状态表示包是数据包还是协议包，包的使用状态标识包处于空闲、读写或者同步状态。包数据状态表通过包数据状态表类实现其方法，主要用于获得包的状态信息。

包地址表类(PkgAdd)提供包的地址信息、每台机器上包的数量和包的名称等，智能客户端可以通过查询包地址表找到数据所在的包。包数据访问接口类(VisitPkgData)提供对包中数据内容进行访问的接口，是对包中数据进行具体操作的接口，主要提供包的定义、读取、查找、修改和删除等数据服务，并且对上述操作进行时间戳记录。

包的名称	包的IP地址	包端口	包的主备	包同步数	包状态	包的使用状态

图 4 包数据状态表的组成

4 结束语

按照功能要求对 DMMS 进行了最终测试，测试用局域网为百兆以太网，11 台计算机组成局域网，1 台计算机作智能客户端。测试结果表明，用户能够利用局域网计算机的内存资源存储数据，并能够成功地访问数据，数据的保护机制也是有效的。性能测试结果表明，在较小数据量的情况下，DMMS 数据的存取效率是本机硬盘的 5 到 7 倍，系统数据的处理速度完全由网络的带宽和内存处理的速度所确定。

本文提出的 DMMS 能够充分利用网络中其它机器的资源，减少对磁盘等外部存储器的操作，实现了数据的高速读取，是分布式技术的一个有效应用。当然，DMMS 最适合于小型局域网，应用的数据量不是很大。针对大数据量的网络，主要需要解决的问题是系统负载平衡状态的管理，一个可行的方案是专门利用一个服务器建立监控中心，代替上述系统中的队长，这需要在下一步的工作中对协议机制进行改进。

参考文献

1 陈锡明, 卢显良, 宋　杰. 分布式内存文件服务器(DMFS)的研究和设计[J]. 小型微型计算机系统, 2000, 21(1): 60-63.

2 Fu K, Kaashoek M F, Mazieres D. Fast and Secure Distributed Read-only File System[J]. ACM Trans. on Computer Systems, 2002, 20(1): 1-24.

3 邵佩英. 分布式数据库系统及其应用(第 2 版)[M]. 北京: 科学出版社, 2005.

4 Colnet D, Coucaud P, Zendra O. Compiler Support to Customize the Mark and Sweep Algorithm[C]. Proc. of the 1st International Symposium on Memory Management, 1998, 34(3): 154-165.

5 Chang S J, Kapauan P T Z. Modeling and Analysis of Using Memory Management Unit to Improve Software Reliability[C]. Proc. of the 12th International Symposium on Software Reliability Engineering, 2001: 96-102.

6 Moy J. OSPF(Version 2)[S]. IETF RFC 2328, 1998.

7 Dan L C T, Srisa-an W, Chang J M. The Design and Analysis of a Quantitative Simulator for Dynamic Memory Management[J]. Journal of Systems and Software, 2004, 72(3): 443-453.

8 Ng M C, Wong W F. ORION: An Adaptive Home-based Software Distributed Shared Memory System[C]. Proc. of the Seventh International Conference on Parallel and Distributed Systems, 2000: 187-194.