·所庆论文选·

分布式计算机系统·

吴荣泉

（华东计算技术研究所 上海201800）

摘要 提出了一种具有安全性、高可靠性、高可用性的异构分布式计算机系统（HDCS）结构，讨论了HDCS各子系统的功能和组成。

关键词 分布式计算机系统 体系结构 安全 容错

A Distributed Computer System（HDCS)

Wu Rongquan

(East-China Institute of Computer Technology Shanghai 201800)

【Abstract】 First the paper presents a heterogeneous distributed computer system (HDCS)architecture with security, high reliability and availability, then discusses the functions and composition of every subsystem.

【Key words】 Distributed computer system; System architecture; Security; Fault tolerance

越来越多的应用要求计算机系统能够实现功能分布、地域分布，且分布部件之间能够密切配合、协同工作。尤其是一些特殊应用领域，还要求计算机系统具有高可靠性、高可用性和较好的安全性。为了实现异构分布式应用支撑环境，为特殊应用领域的计算机系统建设提供技术支撑，开展相关技术研究非常必要。

1 系统结构

1.1 系统的层次关系

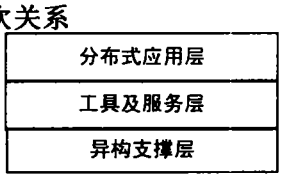

图 1 系统层次关系

各层及其相互关系具有如下特征：

- 低层为服务提供者，高层为服务使用者；
- 较高层也可以直接调用不相邻的低层服务；
- 每层中可以包含多种结构的软件实体，且以统一的服务接口支持高层；
- 每层提供的服务都是其以下各层服务的集成；
- 每层中都包含有安全功能，各安全功能套接形成自应用层至支撑层的安全通道；
- 各层中都包含有管理功能，各层管理功能的集合构成了系统的完整管理。

1.2 工具及服务层

- 结构

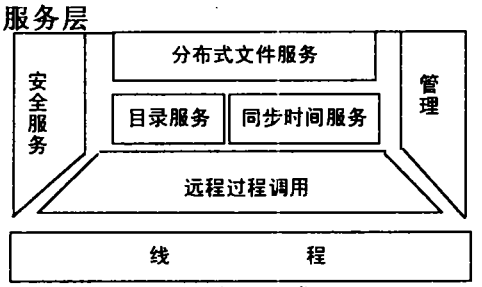

图 2 工具及服务层组成

- 工具及服务子系统之间的关系

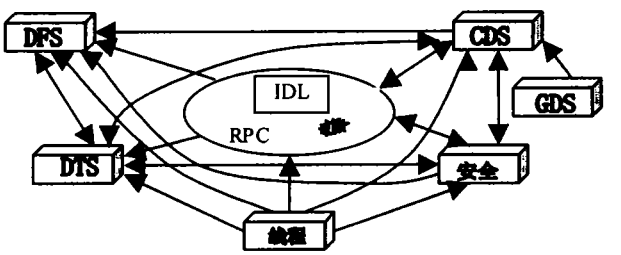

图 3 工具及服务关系

1.3 异构支撑层

（1）结构（局域网）

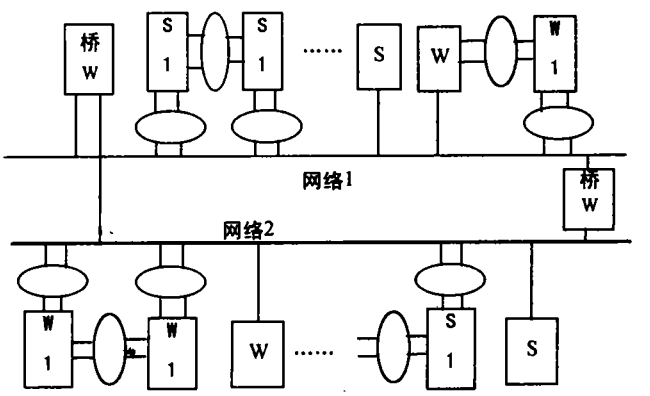

图 4 支持层结构

图4中：表示双端口规划

表示监测通路

* 项目得到了"九五"预研经费资助

吴荣泉 男，46岁，高级工程师，主要从事分布式、网络、容错等方面的研究工作。

收稿日期：1998-09-14

桥　表示双网线连接设备，实现网1与网2连接

W1　表示重要工作站

W　表示一般工作站

S1　表示重要服务器（数据或文件）

S　表示一般服务器

(2) 支撑软件

- Solaris操作系统
- HP-UX操作系统
- Windows NT操作系统
- Windows 95操作系统
- Linux操作系统
- TCP/IP网络协议
- 上述系统支持的数据库和程序设计语言及其它工具

(3) 支撑硬件

- SUN工作站
- HP工作站
- PC 服务器
- PC机
- 交换器
- 集线器

2 工具、服务子系统的功能

2.1 工具子系统功能

(1) 线程

线程模块是为提高系统效率而设计的。我们采用了多种调度策略和多种线程通信方式，由直接调用操作系统（因为当今的主流操作系统都有线程机制）中的线程实现。

线程管理涵盖线程的创建、初始化和终结，线程的优先权调度，线程的中止以及线程与相关性数据的关联。

(2) 远程过程调用

远程过程调用是分布式计算机系统的主要功能部件之一，系统中实现的各个服务都是建立在远程过程调用之上的。主要特点在于与系统中其它组件之间的集成性。例如，远程过程调用使用线程来提高效率；使用安全服务来提供验证的远程过程调用；使用目录服务来查找服务器。

对一个分布式应用，有下面几个基本需求：客户机找到合适的服务器、在异构环境中的数据转换、网络通信。远程过程调用机制满足了这些要求。

2.2 服务子系统功能

(1) 目录服务

目录服务是分布式计算机系统的又一主要功能部件。在HDCS中，以一个局域网为一个单元（也可以以多个局域网为一个单元、或远程网为一个单元），而目录服务是以单元为单位进行组织的。每个单元中拥有一个单元目录服务，它负责管理单元内资源的名字与属性。目录服务是一种分布的、复制的数据库服务，每个单元内可有一个或多个目录服务服务器，每个服务器中存放不同目录信息的副本。每个资源拥有其唯一标识的名字，包括了单元名字和该资源在单元内的名字。HDCS的目录服务支持两种机制，即全局目录服务和域名服务系统，全局目录服务基于X.500标准，用来将独立的单元连接起来，组成一个全局的目录系统。

(2) 同步时间服务

时间服务为加入分布式计算机系统的机器之间实现时间同步。同步时间服务把主机的时间同UTC（世界标准时，一种国际时间标准）同步。用来协调事件发生的顺序，以免因为时钟的漂移，造成误操作，甚至引起系统崩溃。

HDCS的每个单元中都有若干个同步时间服务服务器，装在几台主机上，同步时间服务服务器之间互相校对时间，其它的主机向服务器校对时间。每个单元中还设有一个专门负责提供标准时间的时间提供者（Time provider），它可以是一台机器，也可以是管理员。

由于系统中存在不可预计的时间，例如，主机向同步时间服务服务器校对时间时，系统响应并返回结果的时间。所以，HDCS中的时间是用间隔（Interval）来表示的。

(3) 安全服务

HDCS的安全服务是防止非法用户入侵，同时也防止合法用户进行授权之外的操作和访问，分别由操作系统的安全机制、网络系统的安全机制、HDCS工具及服务层等提供的安全机制形成的安全体系来保护整个系统的安全。由于受篇幅所限，我们将专文讨论HDCS 的安全体系。

(4) 分布文件服务

分布文件服务提供透明的本地和远程文件存取，它允许系统内的进程访问系统内任何地方有存取权限的文件。主要功能包括：

1) 统一文件存取

HDCS 具有全局统一的名字空间，用户不需要了解网络的地址和文件所在计算机的名字就可以在整个分布式系统中透明地访问合法的文件。

2) 单元内文件位置的透明性

数据可以在不被应用人员所感知的情况下在单元内自由移动，具有位置透明性。

3) 对分布式应用程序开发的支持

分布文件服务本身是一个分布式应用，且支持分布式应用的开发。程序员利用分布文件服务可实现数据共享。

4) 高性能

通过HDCS 客户机上对文件和目录数据进行缓存不仅可以减少用户对文件的访问时间，同时能减小网络流量以及服务器的负载，使用户获得了快速系统响应。

5) 高可用性

通过复制技术，一个文件可以存储在多个文件服务器上。当某个文件服务器发生故障时，可以使用存储在另一服务器上的文件拷贝。客户机缓存技术则将文件的一个拷贝缓存在客户机中，即使服务器发生故障，客户机也能够使用缓存中的文件拷贝而继续工作。

(5) 容错子系统

为了适应特殊应用领域的要求，HDCS 提供的容错功能大大地提高了系统的可靠性。由于本文篇幅所限，我们

将专文讨论HDCS的容错结构、设计和实现。

3 各功能子系统的组成

3.1 远程过程调用子系统

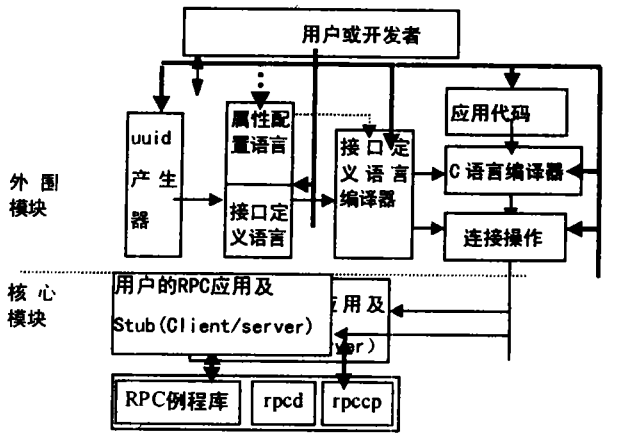

图 5 远程过程调用子系统组成

• 核心模块组成

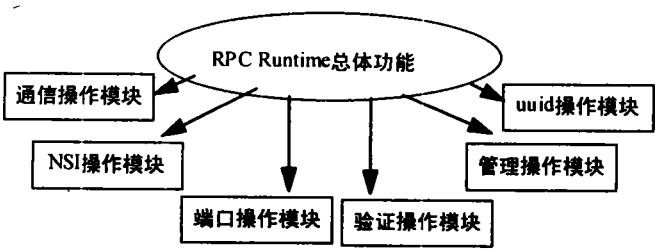

图 6 核心模块组成

3.2 目录服务功能子系统

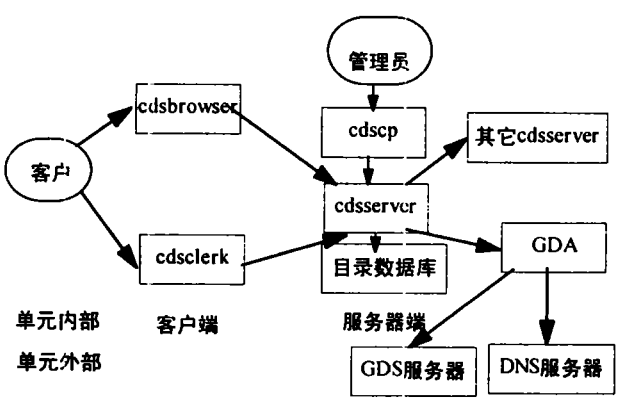

图 7 目录服务子系统组成

3.3 同步时间功能子系统

• 时间职员：同步时间服务客户端，询问时间服务器来调准当地时间；

• 时间服务器：回答时间询问。

• 典型配置为一个LAN三台时间服务器

(1)局部时间服务器

(2)全局时间服务器

(3)通信员时间服务器

(4)后备通信员时间服务器

3.4 安全服务功能子系统

HDCS的安全服务体系由操作系统、网络系统和工具及服务层提供的安全机制组成。除用户注册机制外，还提供了论证、授权机制、加/解密机制。

论证服务由3个重要部分组成：中心数据库、验证服务器和准行批准服务器，其中的中心数据库是安全服务的关键部分。

授权服务是通过存取控制表进行，通过使用该管理器，可设定和更改用户对相关系统资源的访问和存取权限。

3.5 分布文件服务功能子系统

分布文件服务分为两部分：本地部分和远程部分。其本地部分是一个高效的、基于log的本地文件系统，它与UNIX文件系统相似。但提供了比传统UNIX文件系统更多的功能，例如支持文件卷和支持访问控制列表。本地文件系统把影响文件和数据的行为都记录在日志中，以加强文件系统的鲁棒性，一旦系统崩溃了，则利用日志可以正确地恢复系统。

分布文件服务是建立在系统其它服务基础上的分布式客户机/服务器应用，由以下部分组成：

• 缓存管理器

缓存管理器作为分布文件服务的客户端运行于分布文件服务客户机之上，接受用户的文件访问请求。它首先在缓存区中查询是否存在文件的一个拷贝，如果不存在，它就向文件服务器发送访问请求并且缓存该文件。以后就可以直接访问该文件的拷贝了。

• 文件输出器

文件输出器作为分布文件服务的服务器端运行于分布文件服务的文件服务器之上，处理远程客户机的文件访问请求。它接受RPC调用，存取局部文件系统（可以是本地文件系统或者是兼容的文件系统如UNIX File System），并且通过令牌管理器控制多客户机对文件的并发存取。

• 局部文件系统

局部文件系统是系统的物理文件系统，在分布文件服务中占有重要的作用，通过访问控制列表进行权限控制、文件复制，透明地在系统中移动文件以及快速故障恢复。

• 令牌管理器

令牌管理器运行于文件服务器之上，通过向客户发放不同权限的令牌（读、写）来控制多客户的并发文件存取。主要有以下4种令牌：用于访问文件和目录访问的数据令牌（data tokens）；用于查询文件和目录状态的状态令牌（status tokens）；用于文件加锁的锁令牌（lock tokens）和用于打开文件的开令牌（open tokens）。

• 文件夹服务器

文件夹服务器用于对文件夹的管理操作，如创建、删除文件夹。

• 复制服务器

复制服务器是用于对文件夹复制的管理，使管理员可以创建文件夹的拷贝以及第二个文件服务器。它定期地对复制文件夹进行更新操作，即使主文件服务器发生故障，系统也能够依靠第二文件服务器继续工作。

• 更新服务器

更新服务器提供了向各个分布文件服务节点分配二进制文件和管理信息的能力。它是由upclient和upserver两部分组成的，upclient软件运行在那些需要接收新的二进制文件和管理信息的计算机上；而upserver运行于主文件服务器上，自动向upclient机发送更新文件内容的信息。

• 监测器

监测器收集并显示运行在文件服务器上有关文件的信息，便于系统管理员监控整个分布文件服务系统。

• 备份服务器

备份服务器用于文件系统的备份，维护备份数据库的备份记录。

• 文件夹定位服务器

文件夹定位服务器是一个复制目录服务，记录了在每一台文件服务器上有哪些文件夹。可以提供文件夹查询服务，通过文件夹定位服务器，只需要知道文件夹的名字，而不必准确地了解它的位置就可以访问文件，由此实现了文件夹位置的透明性。文件夹在系统中移动或者更新位置操作对用户是透明的。

3.6 容错功能子系统

(1) 硬件

为避免单点故障对系统关键部件的影响，HDCS系统中的每一关键部件（如重要数据服务器、重要文件等服务器或重要工作站）可设计为两模冗余，即为这种关键部件提供两台物理机器、每台机器安装两块网卡、其相应的集线端口规划为一组。用于每一关键部件的两台机器之间另加两条通路，专用于其相互监测。

(2) 软件

通过软件与硬件相配合共同实现系统的容错功能。用于容错功能的软件包括：

• 双端口规划软件
• 双机监测软件
• 双桥切换软件
• 重要信息及时复制软件（确保信息的一致性）
• 容错管理软件

4 结束语

分布式计算机系统是实现分布式应用的基础。我们对分布式应用环境涉及的技术进行研究的目的，试图解决互操作性、高可靠性和高可用性。目前基本实现了异构分布式计算机系统的互操作性，并已完成了高可靠性和高可用性要求的详细设计。

参考文献

1 Open Software Foundation. Introduction to DCE. Prentice Hall Inc 1992

2 伍 宁. DCE分布式文件服务高可用性的分析及其改进[硕士论文].上海：华东计算技术研究所，1998

3 叶茂常. 利用LDAP协议对DCE目录服务的改进[硕士论文].上海：华东计算技术研究所，1998

4 王克伟. 分布式计算环境DCE的安全策略[硕士论]. 上海：华东计算技术研究所，1998

5 吴荣泉. 一种分布式容错系统的设计. 上海市计算机学会第四届学术年会论文集，1994—09

≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈

（上接第37页）

地重现随时间的推移，进程的状态和进程之间的相互通信的变化关系。借助于这些直观的方式，程序员能从更高的角度分析和观察并发程序的行为，从而能有效地判断出进程间的同步和通信的错误。

4 结论

由于应用程序、并发程序设计语言、支持并发的软硬件环境也各异，以上提到的各种单一的方法没有一种能完全解决并发程序调试中的所有的问题。因此，调试并发程序最好能有一个将各种调试技术结合在一起的集成的调试环境，而且在底层的硬件、操作系统设计和编译系统的开发时都应考虑为并发程序的调试问题提供接口。

当然，避免使用会引起程序错误的并发的语言结构则更好，但这需要相应的并行化技术能有突破的进展，如数据流计算机体系结构的发展，数据流和函数式语言的执行的效率能进一步提高或采用自动化的并行编译工具。

参考文献

1 刘炳文. 程序设计语言Ada. 北京：国防工业出版社，1993

2 Andrews G R, Schneider F B. Concepts and Notations for Concurrent Programming. IEEE Computing Surveys, 1983, 15(1): 1

3 LeBlanc T J, Mellor-Crummey J M. Debugging Parallel Programs with Instant Replay. IEEE Trans. Comput., 1987, 36(4): 471—482

4 Miller B P, Choi J D. A Mechanism for Efficient Debugging of Parallel Programs. Proceedings of the SIGPLAN'88 Conference on Programming Language Design and Implementation. Atlanta Georgia, 1988—06: 22—24

5 McDowell C E, Helmbold D P. Debugging Concurrent Programs. ACM Comput. Survey, 1989, 21(4): 593—622

6 Brindle A F, Taylor R N, Martin D F. A Debugger for Ada Tasking. IEEE Trans. Software Eng., 1989, 15(3): 293—304

7 Gait J. A Probe Effect in Concurrent Programs. Software-practice and Experience, 1986, 16(3): 225—233

8 German S M. Monitoring for Deadlock and Blocking in Ada Tasking, IEEE Trans. Software Eng., 1984, 10(3): 764—777

9 Helmbold D, Luckhan D. Debugging Ada Tasking Programs. IEEE Software, 1985, 2(2): 47—57

10 Tai K C. On Testing Concurrent Programs. In Proc. COMPSAC 85, 1985—10,310—317

11 Tai K C, Carver R H, Obaid E E. Debugging Concurrent Ada Programs by Deterministic Execution. IEEE Trans. Software Eng., 1991, 17(1): 1

12 Conti R A. Debugging Ada Tasking Programs. Annual National Conference on Ada Technolgy, 1985,72—81

13 Carver R H, Tai K C. Replay and Testing for Concurrent Programs.IEEE Software, 1991—05,66—74