计算机科学2001Vol.28№.9

高可用性系统软件HHA的服务级管理

Management Based on Servers of High Available System Software HHA

李 艺 李新明 刘绍南

(装备指挥技术学院 北京101416)

Abstract High Available System Software HHA is developed by Institution of Command and Technology of Equipment. The paper describes the management functions based on servers of HHA system, and analyses the main techniques of implementing those functions.

Keywords High Available System, Server, Migrate, Loaded server, Standby server

1. 背景

随着计算机应用日益广泛,银行、证券、油田、民航、海关、尤其是在军事领域,都要利用计算机系统来提供及时可靠的信息和服务,都希望计算机系统永远不停止地正常运行。但是,计算机硬件和软件都不可避免地会发生故障。这些故障有可能会造成整个服务的终止、网络的瘫痪,带来很大的损失。保证系统的可用性、保证系统服务器能为客户提供每周7天、每天24小时不间断的、可靠的服务,正是高可用性系统软件要解决的主要问题。

高可用性系统软件是高可用性系统的核心,是计算机界一个技术难度较大的课题。我国的军事信息系统迫切需要一个有自主版权、有一定通用性、可靠性高、可移植性好的高可用性系统软件,以满足军事及其它行业高可用性系统的需求。1998年,总装备部司令部对"高可用性系统软件"项目立项,装备指挥技术学院经过1年多的时间,成功地开发出"群英高可用性系统软件(Heroes High Availability System Software-HHA)"。该软件支持双服务器结构系统、具备服务级的管理能力、提供对系统软、硬件资源的实时监控和故障恢复等功能,为客户提供不间断的高可用性服务。它是军队自主研究开发的,有完全自主版权,填补了军队相关产品的空白,具有很大的实用价值,并可广泛投入军事应用。

高可用性系统的第一代产品是双机备份系统,是主机级的备份系统。系统通常由两台服务器(或称主机)通过串口或以太网互连,两台服务器配置相同,互为备份,并通过磁盘阵列保证数据的完整性。当一台服务器发生故障而不能提供服务时,由另一台服务器接管其所承担的所有任务。这种方式没有充分利用服务器资源,而且,服务器切换的时间也较长。因此,从九十年代后期开始研制第二代高可用性系统,其主要特点是以服务(或称任务)作为基本管理单位。系统以服务为单位进行监控、迁移,并使两台服务器上的负载相对平衡。这种方式,由于以服务作为管理粒度,一个服务的迁移时间比整个服务器上所有服务进行迁移的时间要少得多,而且服务可以分配在两个服务器上完成,更好地利用了服务器资源。

高可用性系统软件HHA属于第二代高可用性系统,服务级管理是本系统最显著的特点和最直观的功能,下面将着重对服务级管理的功能和实现的技术关键点进行说明。

2. 系统的硬件架构

群英高可用性系统软件运行的硬件平台是两台服务器、一个磁盘阵列和一些网络接口。典型的硬件平台架构如图1所示。

在上述体系结构下,包含以下一些关键内容:

1)服务器。服务器有两种类型:承载服务器和备份服务器。承载服务器是提供服务的主机,备份服务器对承载服务器上承载的服务进行监视,当承载服务器提供的服务不可用时,自动接替承载服务器的工作。本系统的两个服务器可同时作为某一服务的承载服务器和另一服务的备份服务器使用,即两台承载服务器互为备份。一台主机上面可以运行多个服务,也可作为多个

李 艺 硕士,副教授,主要从事网络和系统软件的研究。李新明 硕士,教授,主要从事操作系统、实时系统的研究。刘绍南 工程师,主要从事网络的研究。

服务的备份服务器。对客户机而言，两台服务器在逻辑上是一个整体。

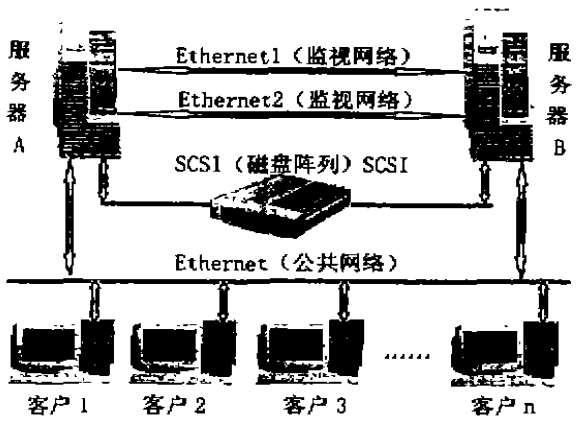

图1 系统硬件结构

2)网络连接。网络连接有两种类型：私用网(或监视网络)和公用网络。两台服务器通过私用网传送诊听信息，使两台服务器能够相互了解对方的运行情况。公用网用来向客户提供服务。

3)存储设备。存储设备有两种：自用存储设备和公用存储设备。自用存储设备是每台服务器自己的存储设备，用于存放操作系统软件和其他一些不需要被另一台服务器访问的软件和数据。公用存储设备是一个磁盘阵列RAID，两台服务器通过SCSI线与之相连，磁盘阵列上的信息可以被两台服务器分别访问，用于存放网络共享的重要数据。

4)软件系统。两台服务器上都包含以下几种类型的软件：操作系统--两台服务器都使用Solaris 2.5操作系统。应用软件--两台服务器上可正常运行各种数据库软件(如：Oracle、Sybase、Informix，SQL等)或其他应用软件(如：WWW服务、NFS服务等)，HHA软件--群英高可用性系统软件HHA同时在两个服务器上运行，用于监视系统的状态，协调两台服务器的工作，维护系统的可用性，保证为客户提供持续不断的服务。

3. 服务级管理功能

群英高可用性系统软件HHA的主要功能是要保证系统为客户提供持续不断的服务。为此，本系统提供了以下的主要功能：1)系统管理；2)服务管理；3)软件资源的实时监测与处理；4)硬件资源的实时监测与处理；5)系统报警。其中服务管理功能是本系统的一个最主要的功能，也是区别于第一代主机级双机备份系统的一个关键功能之一，本系统实现了服务级的粒度管理，以服务为单位，对服务进行增加、提交、迁移、停止和删除等操作，同时对每个服务的运行状态进行监测。服务级的管理功能包括以下一些内容：

1)查询服务的状态 在系统运行期间，系统管理员通过工具，可以随时查询各个服务的运行状态。

2)查询服务配置与重构服务 本系统可以承载各种服务、对此没有任何限制，既可以支持通用的WWW服务、数据库服务，也可以支持用户定制的其它服务。每个服务有一个相关的配置，包括服务的名字、对外服务的IP地址和网络端口、服务的优先级等内容。在系统运行中，可随时对系统承载的所有服务的配置情况进行查询和重构。

3)增加服务 系统管理员可以根据需要，随时在系统中增加服务。当增加服务时，必须提供相应的启动、停止、重新启动、监视等操作的脚本文件。这些脚本文件既可以是SHELL批命令，也可以是一个可执行的二进制文件。通常，一个服务器运行启动脚本，对客户提供指定的服务，另一个服务器则运行监视脚本，对服务的运行情况进行实时监测，当监视程序发现服务运行出现错误时，将通知服务承载方，停止服务，自动将服务迁移到对等服务器上。

4)提交服务 新增加的服务或被停止的服务，必须由系统管理员提交后，才能向外提供服务或开始监视服务在对等机上的运行。提交一个服务后可能结果在第4.5节详细说明。

5)迁移服务 迁移服务就是将服务从一个服务器迁移到另一个服务器上运行。如果一个服务在服务器A上处于运行态，在服务器B上处于监视态，则当服务迁移成功后，服务在服务器A上处于监视态，而在服务器B上处行运行态。服务迁移是进行服务级粒度管理中最常用的一种操作，当发现故障或当负载不均衡时，都要通过服务迁移来进行处理。服务迁移过程对使用服务的客户而言是透明的，从而保证用户接收到的服务是可靠的、不间断的。

6)停止服务 用户可以根据需要停止某个服务。如果是服务承载方停止服务，则对等服务器监测到服务被停止后，自动把服务迁移到本地服务器上。如果是服务监视方停止服务，则对等服务监测到此变化，服务仍由原承载服务器提供。

如果一个服务器上的HHA系统被关闭，则首先要停止所有的服务，因此，系统将自动把所有的服务迁移到对等服务器上，保证对客户的不间断服务。

7)删除服务 当系统管理员确认不再需要提供一个服务时，可以将服务删除。如果要重新提供该服务，则必须首先增加该服务，重新提交后才能再使用。当服务删除后，对应的所有配置文件和可运行脚本文件都被删除。

4. 服务级管理的实现要点

在实现服务级的管理功能时，有以下主要的技术要点：

4.1 服务状态的设计

本系统的服务可能处于以下一些状态中：

状态名	标识符	说明
活动态	ACTIVE	服务在本地服务器上被启动。
监视态	STANDBY	服务在对等服务器上被启动、本地服务器对服务进行监视。
下线态	OFFLINE	服务在本地服务器被停止。
转换态	TRANSITION	服务处于中间转换状态。
失败态	FAILED	服务在本地服务器上的启动或监视出现错误。可能是由于系统环境配置不当，也可能是用户提交脚本错误，或者是参数配置错误。

4.2 服务迁移的实现

服务迁移是进行服务级粒度管理中最常用的一种操作。服务迁移有两种时机：一种是自动迁移；当发现服务不能由原承载服务器正常提供时，系统将自动进行服务迁移。另一种是手动迁移；用户可以根据情况，随时手动地迁移服务。

需要进行自动迁移的情况很多，如服务监测方发现服务运行不正常，或者服务承载方发现服务所使用的文件系统、磁盘、网卡等资源工作不正常，或者服务被服务承载方停止等。需要手动迁移的情况由系统管理员确定，如当本地系统资源的使用达到最高上限而被报警时，或者用户要对系统进行一些日常维护或检修等。

服务迁移时解决了以下两个主要的技术难点：

1)对用户透明。针对每一个网络服务，都有唯一的一个 IP 地址与之对应。对客户来说，服务请求都发送给该 IP 地址对应的服务承载服务器。当服务迁移后，必须自动通知客户端：服务的 IP 地址对应的服务器已发生变化。本系统采用以下方法进行处理：当服务迁移完成后，发送一个新的 ARP 广播包到客户端系统，客户端收到 ARP 包后，在 ARP 协议层对 ARP 缓冲区做更新，从而确保了客户端的 IP 地址和物理地址对不过时。这个过程对应用层来说是透明的。在迁移的过程中，服务请求者可能会经历一个较慢的响应，但感觉不到服务被中断过，也感觉不到服务已经在一个新的服务器上提供。

2)临界状态的处理。服务迁移时，必须确保任何时候都不会出现两个服务器同时提供同一个服务的情况。因此，服务迁移时，首先要在原来服务承载服务器上停止服务，在停止操作完成前，服务在两个服务器上都处于一个转换态，此时，不对客户提供服务。当停止操作完成后，确信服务已经不再由原服务器提供时，从新的服务承载服务器上将启动该服务。启动服务时，必须首先构造与服务相关的环境，如网络构造、文件系统安装等工作。然后，调用相关的服务启动程序，启动服务。最后，记录有关服务的状态信息。

4.3 服务优先级和自动迁移标志的设计

如果一个服务只在一个服务器上被提交，则该服务器必须作为承载服务器启动该服务。如果一个服务同时在两个服务器上提交，则必须确定服务在哪一个服务器上启动。

为此，为每个服务设计一个优先级，由系统管理员在服务配置时确定。优先级可以为1或2，当优先级为1时，表示服务首先在该服务器上启动。如果一个服务在本地服务器上的优先级为1，则在对等服务器上的优先级应该为2。

当系统管理员提交一个服务时，如果该服务已经由另一个服务器提供，则系统管理员会有两种处理意见，一种是希望不管服务的优先级是多少，后提交服务的服务器作为备份服务器，执行对服务的监视工作。另一种是希望严格按照优先级的指定，当两个服务器上都提交服务时，服务总是在优先级为1的服务器上启动。

为此，系统设计了一个全局标志——“自动迁移”标志，由系统管理员在系统配置命令中设置。当提交一个服务时，如果该服务已经由另一个服务器提供，系统的处理策略由自动迁移标志来确定。当自动迁移标志为假时，不管服务的优先级是多少，后提交服务的服务器作为备份服务器，执行对服务的监视工作。当自动迁移标志为真时，如果提交的服务的优先级为1，则该服务总在本地服务器上启动，系统将自动把服务从对等机迁移到本地服务器上；如果提交的服务的优先级为2，则本地服务器作为服务监视方，运行服务的监视程序。

4.4 服务提交的实现

定义了服务的优先级和自动迁移标志后，提交一个服务的处理流程如下：

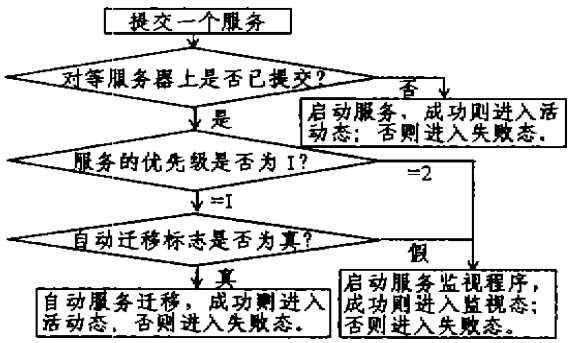

图2 服务提交的处理流程

4.5 以服务为单位的实时监测

本系统实现了以服务为单位的实时监测。当一个服务运行时，服务的承载方和备份方都有相应的监测手段，对服务状态进行监测。

服务的承载方要对服务使用的硬件资源，目前包括本服务使用的网络接口和 SCSI 接口的状态进行监测。由于本系统采用了多进程、多线程的软件结构，对上述硬件资源的监测可以由独立的线程来完成。当服务启动时，系统同时创建两个线程，分别对服务使用的网络接口和 SCSI 接口的状态进行监测。

同时服务的监视方也要对服务的运行状态进行监视，监视程序由服务提供，每个服务都应提供相应的启动、停止、重新启动和监视等操作的脚本文件，备份服务器上创建一个独立的进程，执行服务的监视脚本。

当上述任何一个线程或进程发现服务及其使用的资源出错时，将进行自动的服务迁移，在承载服务器上停止该服务，迁移到备份服务器上。

4.6 服务事件处理线程的设计

本系统采用了统一的事件处理机制对进程与进程之间、线程与线程之间、不同节点之间的通讯进行处理，事件是上述通讯传递的信息的总称。事件可能是由系统产生的，也可能是用户发起的，还可能是对等服务器向本地主机发送的。事件有两种类型，一种是全局事件，对所有服务有效，如监听到对等机失败、诊听线的所有本地 NIC 崩溃、诊听线的所有远程 NIC 崩溃、停止所有的服务、查询所有服务的状态等事件；另一种是局部事件，对某个或某几个服务有效，如提交一个服务、停止一个服务、删除一个服务、在本地启动一个服务、在本地关闭一个服务、查询服务的状态、查询服务的配置信息、将服务从一个主机迁移到另一个主机等事件。本系统设计了一个全局事件队列，事件全部发送到该队列中，并且按插入队列的次序，顺序处理。

由于各个服务之间是互相独立的，只与各个服务相关的事件完全可以并发处理。因此，为了提高并行处理能力，使事件得到及时处理，采用多线程技术。为每个服务设计了一个局部事件队列，并创建了一个线程来处理服务自己的局部事件。

这样，全局事件队列中的事件有以下几种处理方式：1）直接处理；2）复制成多个事件，发送到相关的服务的事件处理队列中；3）直接转发给对应服务的事件处理队列；4）生成新的事件，向对等服务器发送。

服务事件处理线程等待在服务自己的局部事件队列上，当有事件到来时，对事件进行处理，完成后，继续等待在局部事件队列上。

由于有多个线程要同时访问全局事件队列和每个服务的局部事件队列，因此必须分别对上述所有队列互斥访问，并确保不会产生死锁。

结束语 2000年11月25日，由总装司令部组织，北大杨芙清院士任鉴定委员会主任委员，对高可用性系统软件 HHA 进行了鉴定，鉴定委员会一致认为："该成果是具有完全自主版权的软件产品，对军事信息系统具有实际应用价值，对我国软件产业的发展有积极作用。该产品具有创新性，在双服务器结构、高可用性技术上达到了目前国内领先水平、国际先进水平。"服务级的管理是该系统中最主要的一个功能和特点。

参 考 文 献

1 Castagnera K, Cheng D, Fatochi R, et al. Clustered Workstations and Their Potential Role as High Speed Computer Processors: [NAS Computational Services Technical Report RNS-94-003]. April 1994

2 Sterling T, Becker D, Savarese D, et al. BEOWULF: A Parallel Workstation for Scientific Computation. In: Proc of the 1995 Int Conf. on Parallel Processing (ICPP), August 1999

（上接第118页）

在步骤(2)-(4)中，系统是按如下方法确定寄存器平面的：如果存在一个寄存器平面从未参与数据分布，则指定该寄存器平面，否则按照最远引用先分配的原则指定寄存器平面，如果该寄存器平面包含先前计算更新的数据，则应将数据存储到共享存储空间中。

结论 本文结合 LS SIMD 并行 C 编译器的具体设计，在分析了 LS SIMD 数据通信机制的基础上，对数据通信优化的关键技术进行了深入研究并提出了解决方法，主要包括以下内容：

1）给出了与数据通信优化相关的一些基本概念，并描述了这些概念间的相互联系，针对并行计算的一般特点，对问题进行了相应的简化操作。

2）针对自动数据分布给出了相应的数据通信优化算法，并对其中的寄存器空间的状态表示、各类数据通信的选择及生成方法进行了详细的描述。

参 考 文 献

1 High Performance Fortran Forum. High Performance Fortran Language Specification. Version 2.0. 1997

2 Kennedy K. Automatic Data Layout for High Performance Fortran. In: Proc. of Supercomputing, 1995

3 Kandemir M. A Linear Algebra Framework for Automatic Determination of Optimal Data Layouts. IEEE Trans. Parallel and Distributed Systems, 1999, 10(2)

4 Bal H E. Object Distribution in Orca using Compile and Run-Time Techniques. In: Proc. Conf. on Object-Oriented Programming Systems, Languages and Applications, 1993

5 沈绪榜 MPP 嵌入式计算机设计. 清华大学出版社, 1999

6 Wang Hui. Implementation of Data Parallel C Compiler of LS SIMD. In: Proc. of 7th Joint Int Computer Conf. 2000