服务在线迁移系统研究与实现*)

张仲敏[1] 宋 凭[1] 许 鲁[2]

(西安通信学院一系 西安 710106)[1] (中国科学院计算技术研究所 北京 100080)[2]

摘 要 研究并实现了服务在线迁移(service online migration,SOM)系统。该系统能以指定的迁移策略在不影响生产中心正常提供服务的同时为其提供服务容灾能力,通过系统内核中的层次化请求过滤(Hierachical Request Filter,HRF)获取系统即时状态,保证迁移数据的一致性。测试与实际应用检验证明,该系统在迁移开销、迁移有效性及一致性等方面均达到预期目标。

关键词 服务容灾,服务迁移,数据一致性

Research and Implementation of System of Service Online Migration System

ZHANG Zhong-Min[1] SONG Ping[1] XU Lu[2]

(Dept. of Computer & Information Engineering, Xi'an Communications Institute, Xi'an 710106)[1]

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)[2]

Abstract System of service online migration (SOM) is researched and implemented. Capability of disaster recovery can be provided with specified strategy while the implementing service of the production center does not be interrupted, and consistency of migration can be ensured by HRF in system core which getting the system status. Test and practice in reality prove that SOM has achieved anticipated goals in expenses of migration, validity and consistency etc.

Keywords Disaster recovery for service, Service migration, Data consistency

时下,基于大型数据库的数据容灾研究较为普遍,但对于具有实时数据保护与灾后恢复需求的中小企业代价高昂。针对这一领域,提出一种新型的数据保护机制——服务在线迁移系统(service online migration,简称 SOM)。

从容灾的数据对象角度,容灾技术主要针对业务数据和系统本身进行容灾。前者是在灾难发生后尽快恢复此前的业务数据,使数据损失最小;后者是在灾难发生后在当前节点或其它节点上尽快恢复原系统环境的可用性,并最大程度与灾难发生前的状态保持一致。业务数据与系统数据共同构成系统服务,服务在线迁移机制可以同时对系统数据与业务数据实现容灾。在线迁移指实现服务迁移与生产活动的不相关性,二者可同时进行,互不干扰。

OneStat 在 2006 年的调查数据表明,Windows 操作系统在中小企业及个人用户中所占比例为 96.97%;同时,由于 Windows 内核数据组织与运行机制均不透明,其内核的封闭性导致实现数据迁移有较大困难。因此,基于 Windows 平台的容灾研究具有广泛的实际意义与一定的理论意义。

1 服务在线迁移系统架构

1.1 网络容灾系统

网络环境下容灾备份系统组成要素主要有:本地生产中心、本地备份中心、远程备份中心等,如图 1 所示。多个冗余的生产中心构成本地高可用系统;生产中心选择空闲时刻将保护的数据实时或定期迁移到本地备份中心,本地备份中心再定期集中向远程备份中心迁移备份的数据;远程备份中心通过灾难检测与数据同步机制对本地系统进行容灾。一旦本地生产系统的某些节点失效,备份中心立即进行灾难恢复。数据迁移在整个系统中扮演非常重要的角色。基于此架构,如何实现服务(系统数据与业务数据)的整体迁移,使服务迁移与生产活动互不影响,且灾难发生时能最小化服务的停顿时间,是 SOM 所关注的主要问题。

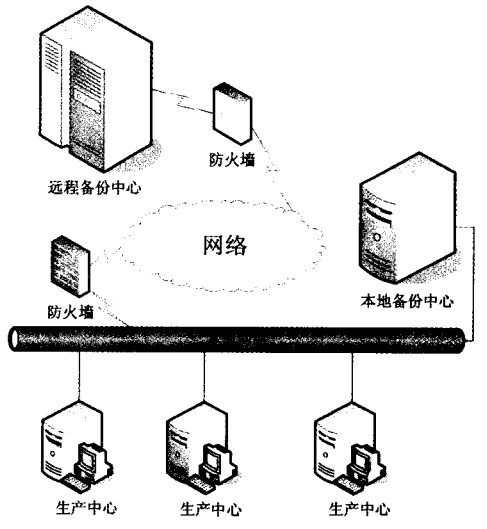

图 1 网络容灾系统

1.2 SOM 系统架构

SOM 系统架构如图 2 所示。

为了简化表示,图中只给出两个生产中心。每个生产节点的服务均运行于前端本地物理磁盘上。后端备份中心为采

*)国家自然科学基金资助项目(编号 60373045)。**张仲敏** 讲师,硕士,主研方向:备份容灾、网络存储;**宋 凭** 讲师,博士,主研方向:网络集群管理;**许 鲁** 研究员,博士生导师,博士,主要研究方向为操作系统、体系结构和网络存储。

用SAN架构的海量存储池，将特定的存储空间划分给生产中心，向上提供访问接口。在具体实现中，采用中科院计算所工程中心蓝鲸服务部署系统PC SAN服务器冗余地连接网络存储设备(FC盘阵)与前端自带物理磁盘的生产中心。PC SAN服务器使用IP SAN技术，以IP网络协议按需动态分配底层存储设备上的存储空间，并以虚拟磁盘(Virtual Disk，简称VD)的形式提供给生产中心，VD与生产中心自带的物理硬盘构成服务冗余，如图3所示。

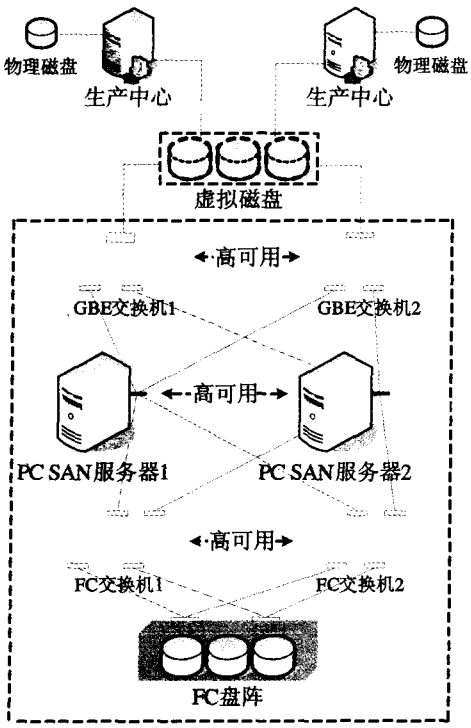

图2 SOM每张架构

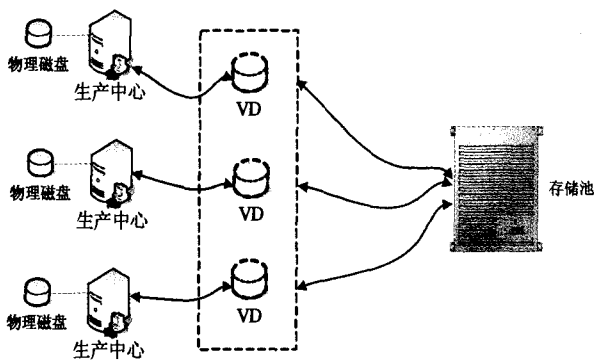

图3 SOM迁移流程图

正常生产时，SOM能够在生产系统正常提供生产服务的同时，根据指定的迁移策略透明地将当前系统即时状态下的系统服务(即系统数据与业务数据)迁移到VD，即备份中心；一旦不可抗的灾难发生导致前端生产中心宕机，将某生产节点的计算资源与备份的VD存储资源相绑定，由VD启动生产环境，可以构成宕机的生产中心的复本，继续向外提供服务。这个过程是非常快的，只须数十秒即可恢复生产系统，从而尽快恢复前端生产系统的正常生产，将前端应用停顿时间减到最小，保持业务的连续性。生产系统恢复生产后，可以按特定的策略向某生产中心(可以是修复后的源生产中心)在线回迁服务数据。服务迁移过程中，SOM能保证迁移数据的一致性。

该流程对于迁移源与迁移目标是对等的。当前端生产系统正常提供服务时，SOM解决服务备份的问题，运行服务的物理磁盘为迁移源，VD为迁移目标；一旦前端的生产系统由于灾难被破坏，后端VD绑定计算资源接管服务时，VD可以作为迁移源将服务备份迁移回前端的物理磁盘并保持服务的正常状态；若定期进行系统迁移，则可将生产系统由于灾难带来的损失降低到最小。

2 SOM设计与实现

数据一致性(可恢复性)是衡量备份容灾系统的关键因素。如果服务迁移后备份中心不能保持生产中心迁移前的一致性状态，即备份的服务处于不一致状态，则灾难发生后的生产中心的服务环境是不可恢复的。

2.1 服务数据一致性

定义 如果用时间戳$\tau^P(t)$和$\tau^B(t)$分别代表生产中心和备份中心在时刻t的服务映像，则在时刻t，存在$t'\leqslant t$，使得$\tau^P(t')=\tau^B(t)$，则认为生产中心与备份中心满足服务数据一致性(Data Consistency)。二者之间的相位差，称为主从相位差(记为V^i)，$V^i=t-t'_{max}$，其中$t'_{max}=max\{t'\mid t'<t\wedge\tau^P(t')=\tau^B(t)\}$。

服务数据一致性是服务运行与恢复所必须保证的。若生产中心由于某种故障不可用，需要利用备份中心的数据进行恢复或从备份中心启动服务，主从相位差可以用来以时间的形式评价丢失的数据量。

若系统或应用程序可以由备份数据重新启动，则称该数据为一致性数据。例如，数据包含一个文件系统，若fsck可以在其上正常运行，则为一致性数据。当且仅当数据包含有截止某一时刻的所有更新并且在该时刻之后没有新的更新到来，就认为数据是一致的。例如，一个文件系统重新启动之后，最近创建的文件就有可能消失。SOM备份的服务数据总是能够反映生产中心在过去某一时刻的状态，称备份中心的数据是一致的。

由于操作系统缓存的存在，部分写操作可能并没有写到磁盘上，而是被保存在缓存中。只有在上层应用暂停提交写请求并且系统缓存被清空的情况下，才能保证生产中心与备份中心应用级数据的一致性。

2.2 SOM实现及数据一致性保证

SOM基于文件粒度(file-based)设计，它较之数据块粒度(block-based)迁移方案，具有安全性受文件系统保护、生产系统额外开销小、迁移过程可由文件系统自动优化、保护对象符合用户使用逻辑、对后台网络性能影响低等优点。服务在线迁移过程中以及迁移完成前后维护了生产系统内部持续不间断的I/O操作，使迁移操作和生产活动相独立。

SOM数据在迁移时利用快照机制对生产中心的服务产生即时快照，将该即时快照状态下的服务在线迁移到备份中心。为了保证服务数据一致性，在Windows系统内核中实现一个层次化请求过滤(Hierachical request filter简称HRF)驱动模块。当上层模块准备产生即时快照时，由HRF负责预先完成所有打开的事务、更新事务日志、把系统Cache数据刷到硬盘、阻塞所有的文件操作等准备工作，使该时刻系统服务处于一致性状态，并产生一致性的即时快照。设计协调器负责与写入器、请求者、底层HRF间的通信机制。

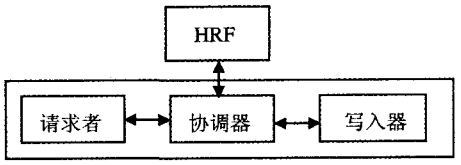

图 4　SOM 实现模块

请求者：创建系统即时快照请求的发起者，本文中指上层的服务迁移模块；

写入器：在系统即时快照创建过程中，如果有上层应用向磁盘迁移源区域写数据，将会影响迁移数据的一致性，写入器作为特定应用的软件模块能够确保该应用作用范围迁移数据的一致性；

协调器：一个中间组件与写入器等组件通信，以协调各组件的交互；

HRF：内核中实现的驱动程序，负责准备一致性数据与生成系统即时快照。

当服务迁移模块下达创建快照指令时，由 HRF 机制通知所有写入器准备各自的数据。写入器的动作包括预先完成所有打开的事务、更新事务日志、把系统 Cache 数据刷到硬盘等。准备好数据之后，各写入器通知协调器，再由协调器分别通知请求者和内核中的 HRF 模块，由 HRF 模块瞬间 freeze 所有磁盘的写请求并将其写入 freeze 队列，并在瞬间（几秒钟之内）产生即时快照。产生快照后，再经由协调器通知上层请求者和写入器，从 freeze 队列中 thaw 所有的 IO 请求，继续系统的正常 IO。整个过程在数十秒之内完成，对系统的正常 IO 不会造成大的影响，对系统的瞬间的写延迟可忽略不计。这样的机制保证了迁移数据的一致性状态；系统即时快照生成后，断开其与源数据的连接。源数据可以应用于服务继续访问。即时快照以只读的方式保存于存储介质上，供服务迁移模块调用。

2.3　HRF 层次结构

在 Windows 分层驱动模型的基础上，在设备层元上用 UpperFilters 来实现 HRF。

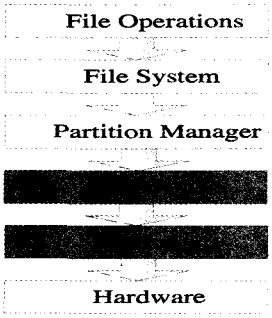

图 5　基于 Windows 分层驱动模型的 HRF

在 Windows 分层驱动模型中，文件系统接收到上层传来的文件操作（如创建、删除、读、写等）命令，针对每个操作产生相应的内核态处理单元 IRP（I/O Request Packet），IRP 是 Windows 系统中上层应用与内核态驱动及内核态驱动之间通信的基本单元。接着，将产生的 IRP 下传到分区管理器一层，由分区管理器将其发送到相应的设备上进行处理。处理完成后，再将处理结果逐层返回，直至顶层应用。这里的设备层由 Class/Port/Miniport 分层结构组成。

为了产生一致性的即时快照，在 Device 层之上用 UpperFilters 实现了 HRF，由 HRF 利用请求拦截或过滤的方法，对上层传来的 I/O 请求进行干预。当收到服务迁移模块创建快照指令时，HRF 拦截上层传来的所有 I/O 请求，不将其下传到设备层，而是将这些请求暂时阻塞到 Freeze 队列，此时生产中心处于短时的一致性状态。接着瞬间产生即时快照，产生快照后就可将 Freeze 队列上的请求 thaw，按正常流程将其下传到设备层处理，使之继续执行。整个过程在短短数十秒内完成，保证了迁移数据的一致性，同时保证了产生即时快照时生产中心的正常生产不受干扰。

产生即时快照后，可作为设备对象供迁移模块调用，将该只读设备对象上的一致性数据迁移到备份中心。由于实际迁移源是与生产中心系统相独立的快照对象，因此可以保证迁移时与生产中心的生产活动相独立。

2.4　SOM 迁移模块

以上设计的 HRF 在 Windows 内核产生了独立于生产中心生产环境的设备对象——即时快照，并保证了快照的一致性。SOM 迁移模块在应用层具体实现向备份中心的服务迁移。迁移源为只读的即时快照，迁移目标为备份中心提供给生产中心的 VD。迁移时设计了不同的迁移策略，如第一次向备份中心进行系统迁移时，可以采用全迁移策略，接下来定期迁移则可以采用差异迁移的策略，每次仅将自上次迁移以来发生改变的数据重新迁移。具体迁移过程采用了优化的迁移算法，配合不同的迁移策略最大限度地降低前端生产中心用于容灾的开销，将生产系统性能造成的影响最小化。

3　性能测试和功能评价

SOM 测试环境如表 1 所示。

表 1　服务在线迁移性能测试环境

	备份中心	生产中心
CPU	Intel Xeon 3G	Intel P4 1.5G
主板	TYAN Tiger-i7320-S5350	Gigabyte 8IDXI
内存	512M * 4 DDR	512M SDR
RAID	3Ware9500S8 RAID5	-
磁盘	8 * 120G SATA	SAMSUNG SP1604N IDE
网卡	Broadcom NetXtreme BCM5721 (1000M)	3com EtherLink XL 10/100 PCI TX NIC

使用全迁移策略下，服务迁移平均速度达到 8MB/s；在差异迁移策略下，基于差异快照技术，减少了实际迁移的数据量，整体迁移速度更快。服务迁移时对源生产中心性能造成的影响很小，同时 HRF 机制保证了数据迁移后的一致性。

服务迁移之后的目标生产中心具备源生产中心的所有服务能力，包括系统数据与业务数据。

表 2 是 SOM 与 PowerQuest V2i Protector 软件的对比结果。PowerQuest V2i Protector 是磁盘到磁盘的容灾解决方案，提供数据迁移的一致性。

表 2　服务在线迁移与 PowerQuest 的比较

	SOM	PowerQuest
创建快照时间	15s	180s
平均数据迁移速度	8MB/s	5MB/s
系统级克隆部署	支持	不支持
占用系统资源	少	多
需停止其它服务	否	是
支持的系统环境	Windows 2000/XP/2003	Windows 2000/XP/2003

（下转第 121 页）

序都能不做修改地在分布式环境下正常运行。这说明:第一,我们设计的DNMR能够为应用程序提供符合JBI规范的访问接口。第二:分布式DNMR确实屏蔽了组件物理位置的差异,实现了跨节点的消息路由和消息传输。

实验2 我们将第三方的BPEL引擎进行包装,使之成为JBI环境中的一个标准的组件,用BPEL文件描述个各服务之间的组合方式。当流程任务部署完成后,启动流流程运行,在某个JBI节点故障的情况下,只要总线内部有同名的备选服务,消息仍然能够被路由至提供该服务的运行组件。这说明系统的可用性得到了提高。

实验3 为了考察消息路由表的内存开销,我们按照实验2的方法,根据某数字园区的多个企业之间应用集成需求构建了100个流程任务。参与流程的组件个数从2个到100个不等。我们逐一将流程任务加载到分布式企业服务总线的环境中运行,并考察每个节点上路由表的内存开销。试验表明,可以认为消息传输路由表的大小随流程应用的任务数的增加成线形增长。当100个任务加载完毕后,路由表的内存开销约在800k左右。在实际应用中,这种内存开销是可以被企业用户接受的。

实验4 我们采用单位时间内消息路由表变化的项数来衡量由于消息路由表的变化而带来的网络资源的开销。仍采用试验3中的实验用例进行实验。实验表明:企业内部和跨企业之间的应用集成。路由表变更带来的网络开销主要是由JBI节点数目的变更、新流程任务的加载或者节点故障引起的。我们考察了8个企业的实际应用,发现在实际应用中以上三项变动都是很少的,因此由于消息路由表带来的网络资源开销可以忽略不计。

实验5 我们对随机等待-失效消息广播的机制进行了模拟实验。节点失效的响应时间定义为从某节点失效到第一个失效消息被发送到JMS广播队列的时间。节点失效的成本定义为一个节点失效后系统产生失效消息的个数。在泛洪机制中节点失效响应时间为JMS Topic的广播时间,在有n个节点的情况下,一个节点失效的广播消息个数为$n-1$。采用随机等待-失效消息广播的机制后,我们考察失效响应时间相对于泛洪机制的增量并考察失效消息的个数。模拟实验的结果如图5、6。从图5可以看出,随机等待-失效消息广播的机制的节点失效响应时间总是大于等于泛洪机制的响应时间,但响应时间的增量是很小的,并且随着节点数目的增多,响应时间有收敛的趋势,逐步逼近泛洪机制的响应时间。从图6可以看出,对于随机等待-失效消息广播的机制,随着节点数目的增多,节点失效成本在增加,但远远小于泛洪机制的开销。这说明随机等待-失效消息广播的机制在保持良好的响应时间的同时有效地降低了失效的开销。

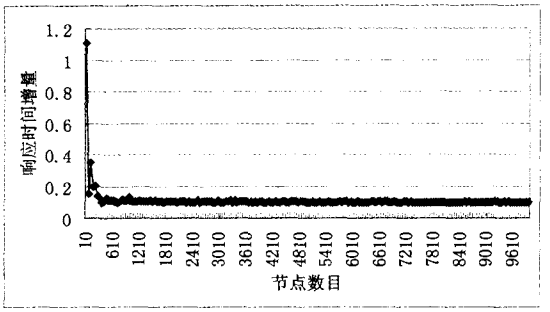

图5 失效消息响应时间图

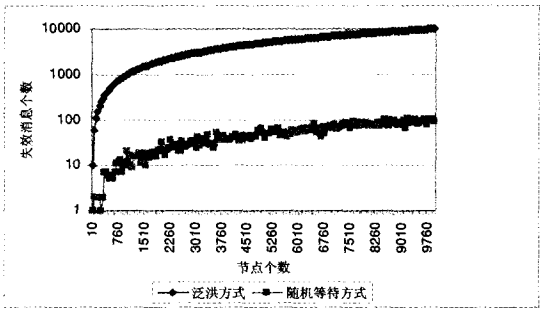

图6 两种失效机制产生失效消息个数试验结果

结论 基于JMS的分布式企业服务总线在满足JBI规范的基础上对企业服务总线的能力进行了扩充,突破了单JBI环境中的总线负载的性能瓶颈,实现了消息传输的持久性,克服了单点故障,提高了系统的可用性。充分利用了JMS发布订阅模式,在不依赖全局消息路由器的情况下实现了全局范围内的消息路由同步和消息转发。实验表明,我们提出的分布式企业服务总线的设计方案有良好的应用效果。

参考文献

1 Cherbakov L, Galambos G, Harishankar R, et al. Impact of Service Orientation at the Business Level [J]. IBM Systems Journal, 2005, 44(4)

2 Schmidt M-T, Hutchison B, Lambros P, et al. The Enterprise Service Bus: Making Service-oriented Architecture real [J]. IBM Systems Journal, 2005, 44(4): 781～797

3 Sun Microsystems Inc. Java Business Integration (JBI)1.0, 2005

(上接第113页)

由上表数据,SOM与PowerQuest相比,速度优势明显,占用系统资源少,且迁移时不影响正常的生产活动,因此在性能与功能方面,均占有优势。该系统目前已应用于中科院计算所工程中心蓝鲸服务部署系统,为生产前端应用服务器(AS)提供容灾服务,达到了预期效果。

结束语 设计并实现了服务在线迁移系统,为Windows生产中心提供容灾。SOM能够对包括系统数据与业务数据在内的系统服务进行完备的备份,灾难发生后能实时恢复生产环境与应用数据,HRF机制保证了迁移状态的一致性以及与生产中心生产活动的独立性。经检验在性能、兼容性等方面都达到了预期效果。

参考文献

1 Barkley P. Disaster-recovery Plans Focusing on Business Continuity [Z]. E-Commerce Times. http://www.ecommercetimes.com/story/35 993.html, 2 004-08-23

2 Ma Yili, Fu Xianglin, Han Xiaoming, et al. The Separation Between Storage and Computation. Journal of Computer Research and Development, 2005, 42(3)

3 Jiang X, Xu D. SODA: A service-on-demand architecture for application service hosting utility platforms. In: Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing, 2003. 174～183

4 Lu Chenyang, Alvarez G A. Aqueduct: Online Data Migration with Performance Guarantees. In: USENIX Conference on File and Storage Technologies, Monterey, CA, 2002-01. 219～230

5 Solomon D A, Russinovich M E. Inside Microsoft Windows 2000, 3rd edition

6 Microsoft®Windows® XP Driver Development Kit (DDK)