基于虚拟化的应用容灾平台探索

许冠军

(台州科技职业学院信息中心 台州 318020)

摘 要 数据和信息系统是现代企业运营的基本要素,确保数据的完整性和信息系统的高可用性成为了信息化部门关注的焦点。针对中、小型企业应用容灾问题的要求和特点,提出了基于vSphere虚拟化的应用容灾平台建设方案。该方案的RTO、RPO指标和系统可靠性在局部故障和整体故障的条件下,都较好地满足了容灾需求。该平台在我院数据中心中的应用,初步证明了其技术的可行性。

关键词 应用容灾,虚拟化,数据中心,VMware,vSphere

中图法分类号 TP309 **文献标识码** A

Exploratory on Virtualization-based Application Disaster Recovery Platform

XU Guan-jun

(Information Center, Taizhou Vocational College of Science and Technology, Taizhou 318020, China)

Abstract Data and information systems are the basic elements of modern business operations. Ensuring the integrity of the data and the HA of the information systems becomes the focus attention of the information technology department. Aiming at the requirements and the characteristics of the small and medium enterprise application recovery problems, the thesis proposed vSphere virtualization based application disaster recovery platform. The solution's RTO, RPO and system reliability are better to meet the needs of disaster recovery. This platform is applied to our data center, and proved its technical feasibility.

Keywords Application disaster recovery, Virtualization, Data center, VMware, vSphere

1 引言

随着信息技术的发展,中、小型企事业单位纷纷建立独立的局域网和业务系统,使得信息管理系统成为了企业生产的关键一环。容灾系统将在不可预知的故障和灾难发生时,能最大程度地保障信息系统的正常服务,保障了生产系统的完整性和连续性。虽然市场上有许多成功的解决方案,但由于项目实施投入大、周期长、工程复杂度高,很多中、小企业望而却步。而不可控因素和日常维护带来的系统下线又不可避免,所以各中、小企业急需一套操作简单、性能可靠、投入有限、复杂度低的应用灾备解决方案。

2 容灾模型及容灾技术介绍

衡量容灾系统的安全性和可靠性有2个重要指标[1,2]:(1)恢复点目标(Recovery point object, RPO),是指从灾难发生到可以让业务恢复正常运行的时间段内,允许丢失的最大数据量;(2)恢复时间目标(Recovery time object, RTO),是指从信息系统下线开始,到系统恢复至正常运作,所能容忍的业务停止服务的最长时间,也就是从灾难发生到业务系统恢复服务所需的最短时间周期。因此,RPO与RTO越小,表示系统的可用性越高,也意味着容灾系统建设的投入越大。在实际容灾系统的建设中,企业应根据自身业务的性质和特点,确定合适的RPO和RTO目标。

根据容灾对象的不同,容灾系统包含3个层次[3],分别是数据容灾、系统容灾和应用容灾。数据容灾,就是构建异地的数据备份系统,保证工作数据能及时、完整地复制到备份系统中,保证数据的完整性、可靠性和安全性。数据容灾只保证关键工作数据的备份,并没有一整套冗余的可运营的业务系统,当灾难发生时,恢复业务需要较长时间,对于RTO要求高的企业就需要更高层次的容灾。系统容灾,就是通过对信息系统关键配置和关键进程的备份,保证运行信息系统本身的高可用性。系统容灾和数据容灾共同构成了基础容灾系统,要实现工作系统的快速灾难恢复,两者缺一不可。应用容灾,也称业务容灾,是指在基础容灾系统上,构建一整套与本地工作系统同构的异地备份应用系统,在正常工作的情况下,主、备系统间互为备份,当灾难发生时,备用系统能自动接管工作系统,提供连续、不间断的应用服务,从而保证了业务的连续性。一般对RPO和RTO目标较高的企业,如银行、电信、电力等,都需保证对应用的容灾。

应用容灾系统[4]往往需要融合以负载均衡、应用集中和隔离、系统运行参数监控等为基础的分布式资源调度(Distributed Resource Scheduler, DRS)技术,以发现、隔离故障,并及时迁移和重分配资源实现系统的连续运行,并在整个过程中尽可能实现自动化,而这些都是虚拟化技术的优势。

本文受浙江省教育厅科研项目(Y201431269),台州市教育科学规划研究课题(GG1414066)资助。

许冠军 (1981-),男,硕士,高级工程师,主要研究方向为图像处理、服务器虚拟化,E-mail:mathgary@gmail.com。

3 中、小企业容灾系统的基本要求

中、小型企业信息化的主要特点是以服务于生产和管理的各业务系统为主，历史数字资源和管理数据资源刚刚开始积累。因此，其对业务的连续性要求较高，而业务数据产生的速率和数据总备份量相对有限。根据此类信息系统的特征，对系统的恢复时间目标(RTO)有如下要求：(1)局部性故障，如部分服务器和部分机柜的设备故障 RTO≤10 分钟；(2)全局性故障，如自然灾害导致整个中心机房大部分机器故障，RTO≤7 天，并要求容灾系统在 RTO 内实现备用系统上线，平滑替换生产系统的功能，并及时发送告警信息。

针对上述容灾目标，对容灾系统提出如下要求：

(1)支持当前主流的操作系统如 Windows 系列和 Linux 常见发行版本，并在操作上实现统一的处理过程，降低平台系统配置和客户端配置的复杂度。

(2)应用容灾和数据备份相统一。最大程度增强重要应用系统的数据完整性和在线能力，解决除电力故障等整体性不可抗拒因素外的普通系统宕机引起的系统离线问题。

(3)系统可扩展性强。容灾系统基于 vSphere 虚拟化平台，该平台有很好的扩展性，保证了容灾系统具有较强的扩容功能。随着企业应用业务拓展，可以进行灵活的容灾系统调整和应用规模扩大。

(4)业务应用的高连续性。基于虚拟化平台，实现主、备机之间的短时间切换，并辅助数据备份功能，实现业务系统的高可用性(HA)。

(5)系统利用率高。该平台针对中小企业的业务规模和人员结构，充分利用虚拟化平台的高资源利用率，既在建设上节省投资，又在能耗上节能环保。

4 应用容灾系统构架和实施

在容灾系统的设计中，分为应用容灾和存储容灾两个子系统。应用容灾子系统采用了 vSphere 虚拟化平台[5-8]的 HA 和 DRS 解决方案。在 vSphere 的构架中，其由提供虚拟资源的 ESXi 服务器和集中管理界面 vCenter 组成。通过搭建同构的 ESXi 物理服务器和网络环境，实现物理资源池的统一管理和负载均衡，在某一台 ESXi 服务器出现物理故障或网络故障时，保证应用及时迁移到同构环境中，实现业务的连续性。

存储容灾子系统的实现，使用 HP 的 Data Protector 结合 VCB(VMware Consolidated Backup)以及虚拟机快照功能(snapshot)来实现虚拟机关键时间点的整体备份，并及时备份虚拟机文件到异地存储。

4.1 系统构架

综合企业的业务规模、性能要求以及物理硬件资源，可以确定生产系统中单台 ESXi 运行的虚拟机数目，可以参考如下经验公式：

$$N=2N_{cpu} \tag{1}$$

在环境部署中，由对等数量的服务器分别承担生产系统和备份系统。在初始设置中，每对(如 ESXi-01 和 ESXi-02)服务器构成一个独立应用容灾子系统，其逻辑架构如图 1 所示。容灾系统采用与生产系统在物理硬件和网络环境上完全同构，形成物理上相互独立、逻辑上相互统一的生产系统和备份系统。

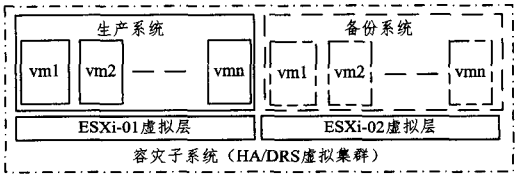

图 1 应用容灾子系统逻辑架构

在数据容灾上，部署一台异地 SAN 作为辅助存储，实现主要业务虚拟机的整机备份，并在关键时间点(如系统升级、数据更新等)和定期(备份周期)对所有业务系统的虚拟机做系统快照，并对相应的虚拟磁盘数据作异地备份，其系统物理架构如图 2 所示。

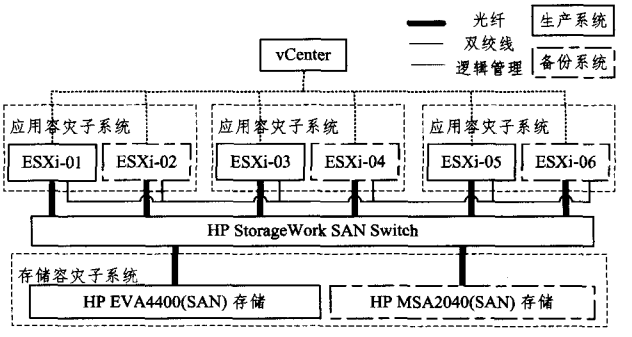

图 2 系统物理架构

4.2 系统设计的几点考虑

应用容灾子系统的设计：采用了一对 ESXi 服务器构成 HA、DRS 集群来实现应用容灾子系统，其目的在于隔离主备系统的同时，有效控制共享资源池的规模，使得 DRS 的虚拟机迁移在更小的范围内进行。在保证 HA 的前提下，减小容灾子系统的粒度，保障了虚拟机的运行性能，也大大降低了整体系统的耦合性。其代价是系统的可靠性有一定的损失，若子系统的两台物理服务器同时出现故障，会导致业务的离线。

存储容灾子系统中实时备份和关键点备份相结合：实时备份可以有效保证数据的完整性，可以让 RPO 降到很低。但由于其实现过程中需要备份代理(agent)做实时的文件操作，系统性能的开销较大。而关键点备份只是在系统运行到关键时间点，如软件升级、系统维护或者备份周期时，对运行系统的虚拟机做快照，然后对快照文件备份。快照操作对在线系统的影响可以忽略不计，但文件的备份会影响系统性能，针对这种特点，可以先做系统快照，再通过定期任务在系统运行的空闲期实现备份文件的复制。总体来说，关键点备份对系统影响较小，但没有实时性的保证，突然的系统故障会付出一定的 RPO 代价。

针对中、小企业环境的特点，建议将实时备份和关键点备份相结合。对 RPO 要求较高而性能要求相对较低的应用系统做实时备份，对 RPO 要求较低或性能要求较高的做关键点备份。

4.3 系统实施

根据系统架构(如图 2 所示)，对整个容灾系统做如下的物理部署和系统实施，如表 1 所列。

(1)6 台 HP 580G7 作为 ESXi 服务器，由其中 3 台 ESXi 承担生产系统，其他 3 台承担备份系统，并在运行 windows server2008 的一曙光服务器上安装 vCenter，统一管理 6 台

ESXi 服务器。

(2)在每台 ESXi 服务器上添加 HP EVA4400 存储,作为高速存储空间,主要用于虚拟机文件(vmdk 文件)的存放。

(3)在备用机房部署一台 HP MSA2040 存储,通过光纤交换机连接到每台 ESXi 服务器上,作为快照后虚拟机文件的远程备份存储。

(4)新建 ESXi 集群,开启 VMware HA 和 VMware DRS 功能,并设置成"全自动"级别和增强型 VMotion 兼容性(EVC)模式,以实现生产系统的高可用性和物理计算资源的负载均衡。

(5)确定远程备份周期,定期执行快照,并制定相应的存储备份策略,以保证定期复制相应的虚拟机文件到远端存储(HP MSA2040)。

表 1 容灾平台物理硬件表

设备	型号	数量/台
服务器	HP DL580 G5	6
	曙光 I620r-G	1
存储	HP EVA4400	1
	HP MSA2040	1
光纤交换机	HP StortageWork SAN switch	1

在系统部署中有如下几个注意点:

1)建议所有的 ESXi 服务器选择同一品牌的同一系列,以便于虚拟机在 ESXi 之间平滑高效迁移。

2)建议使用高计算性能、小硬盘容量的服务器,充分利用 SAN 等共享存储在访问性能、数据管理和容灾的便捷性上的优势,而 VMware HA 和 DRS 功能都以共享存储为前提。

3)对于 ESXi 集群中的所有服务器,配置同构的物理网络,并保持其虚拟交换机配置的一致性,这样可以保证在 vMotion 过程中网络的平滑切换。

5 容灾平台的可靠性分析和优势比较

5.1 常见故障和系统恢复

5.1.1 局部故障

应用容灾子系统的单机服务器或网络故障,不影响业务系统的运行,子系统会将该服务器上的虚拟机自动迁移到备份服务器上继续服务。此时,子系统失去了容灾性能,需要维护人员及时查看系统日志,并确定故障类型,及时恢复其上线。

子系统中两台服务器同时故障,会导致业务系统下线。此时若还有其他子系统正常运行,由于下线子系统的虚拟磁盘存放在公享存储上,只需把下线业务的虚拟机导入正常工作子系统中继续运行,并及时恢复故障子系统,手动迁移业务虚拟机到原子系统。若此时无子系统正常工作,则属于整体故障范畴。

5.1.2 整体故障

整体故障包括服务器或网络整体故障、存储系统故障、电力故障等全局性故障。对于服务器整体故障,其发生的可能性很小,若服务器故障比较严重,则需要启用额外服务器建立新的应用容灾子系统,因其虚拟机文件是在共享存储上,只需将其导入到新系统,即可继续工作。

存储系统故障中,若是备份存储故障,则不会影响应用系统的运行,系统会告警,管理员只需要及时恢复其上线。若是主存储系统故障,则会导致全部应用系统下线。此时,需要把主存储系统切换到备份存储系统。具体操作如下:对于实时备份的应用,重新导入应用子系统则可恢复;对于关键点备份系统,需要导入并恢复到最近关键点,中间会出现系统数据的部分丢失。

若整体网络故障和电力故障,则系统恢复时间完全依赖于网络或电力恢复时间。一般不建议对系统做任何操作。

5.2 可靠性分析

系统的可靠性[9]是指系统在规定的时间内无故障运行的概率,也就是系统维持其功能和性能水平的能力。电力系统中,市电供电系统的可靠性一般为 99.8%,加上机房的 UPS 设备作保障,可使电力的可靠性 $R_{power}\approx 99.9\%$;单机设备的故障与产品品牌和使用年限有很大的关系,新购 3 年内的主流设备其可靠性相对较高,服务器单机的可靠性记为 R_{server},存储设备的单机可靠性记为 R_{san};网络的可靠性依赖于网络的物理规划、网络规模和网络安全的防范措施,这里指的是全局网络的可靠性,记为 $R_{network}$。

根据本文容灾系统的设计,其系统的整体可用性也可以理解为业务系统在线的可靠性,为:

$$R_{online}=[1-(1-R_{server})^2]R_{power}R_{san}R_{network} \tag{2}$$

系统可恢复的可靠性为:

$$R_{recoverable}=[1-(1-R_{server})^{2n}][1-(1-R_{san})^2] \tag{3}$$

其中,n 为应用容灾子系统的数目。

5.3 平台的优势比较

当前市场上成熟的应用容灾产品有鼎联公司的 Lander Cluster、飞康公司的 CDP、浪潮公司的 HA Cluster 等。这些产品利用集群和内置虚拟化技术,适用于物理机和虚拟机环境的应用容灾。其应用范围更广,但部署相对复杂,投入成本较高。

本文中的应用容灾平台适用于中、小企业的纯虚拟化应用环境。在数据中心的虚拟化过程中,只需对物理部署做相应的容灾规划和基于 vSphere API 的简单开发,就可以实现系统的应用级容灾,节省了独立容灾系统的投入。

平台完成初始部署并设置备份周期后,无需人工干预。对于管理人员来说,其面对的是 VMware 虚拟化产品,而非第三方容灾系统。该平台在简化操作的同时,大大降低了容灾的复杂性。

结束语 本文基于 vSpere 虚拟化平台的 HA、DRS 和 VCB 技术,以及虚拟机快照技术,实现了应用子系统和存储子系统的冗余备份。该系统通过集成的 vSphere 虚拟化平台,来解决类似我院的中、小企业网络信息环境中,应用系统多而复杂、应用系统高可用性差、重要数据备份管理困难等问题,实现了主要应用系统的实时容灾和重要数据的实时备份,大大降低了普通容灾备份系统的操作复杂、投入大等问题,同时为提高数据安全级别和系统的稳定性提供技术平台上的支持。

该系统经过测试,在服务器单机故障的条件下,RTO≈0;在局部服务器故障、网络故障的条件下,RTO≈8min;在中心机房服务器群整体故障的条件下,由于本文环境中的快照备份周期为 1 周,因此 RPO≈1 周数据量,RTO≈2 天,系统可恢复的可靠性为 $R_{recoverable}\approx 99.999\%$,在线的可靠性为 $R_{online}\approx 99.7\%$。

参 考 文 献

[1] 杨晓红,李健,杨卫国. 信息系统容灾技术分析与研究[J]. 计算机工程与设计,2005,26(10):2727-2729

[2] 徐鹏,薛建锋. 数据中心容灾系统研究[J]. 计算机工程与设计,2007,28(22):5556-5558

[3] 刘其成. 郑纬民,陈康. 虚拟化技术在容灾系统中的应用[J]. 小型微型计算机系统,2010,31(10):1954-1957

[4] 杨义先,姚文斌,陈钊. 信息系统灾备技术综论[J]. 北京邮电大学学报,2010,31(10):1-5

[5] Sindoori R, Preetha Pallavi V, Abinaya P. An Overview of Disaster Recovery in Virtulization Technology[J]. Journal of Artificial Intelligence, 2010, 31(10): 1-5

[6] Maitra S, Shanker M, Mudholkar P K. Disaster recovery planning with virtualization technologies in banking industry[C]// Proceedings of the International Conference & Workshop on Emerging Trends in Technology (ICWET'11). ACM, New York, NY, USA, 2011: 298-299

[7] Guster D, Lee O F. Enhancing the Disaster Recovery Plan Through Virtualzation[J]. Journal of Information Technology Research, 2011, 4(4): 18-40

[8] 王晓东,康东明,顾晓鸣. 基于虚拟化技术的灾备系统模型研究[J]. 计算机与网络,2010,36 (1):53-56

[9] Rausand M. System Reliability Theory: Model, Statistical Methods, and Applications(2nd Edition)[M]. Wiley-Interscience, 2003

(上接第 410 页)

在内存中开辟一块区域,通过预调的方式,将正在运行的进程将要使用到的页面号存储到该区域中,在置换页面时,通过 hash 函数,先比较在内存中的页面号与预调存储区域中的页面号,若该页面号在预调存储区域中,则不置换,只置换出在预调页面中没有的页面。同时结合使用到的页面的访问位和修改位,在实现时给每个页面加一个寄存器用以存储该页面的访问位和修改位,若访问位为 1,则表示在最近一段时间内该页面被访问过,如果修改位为 1,则表明该页面被修改过;在淘汰时,将修改位为 1 的页面返回到外存储区。根据程序的局部性原理,若该页面被访问过,则该页面很有可能在下一个时间段还会被访问,所以不置换该页面。在实现时,用指针来指向某个页面。在淘汰时,选择访问位为 0 且不在预调区域中的页面,同时,此种页面为最佳淘汰页面。

淘汰算法的步骤如下:

(1)从指针所指向的当前位置开始扫描,来找最佳淘汰页面,最佳淘汰页面为:访问位为 0 且不在预调区域中的页面;找到第一个访问位为 0 且不在预调区域中的页面作为淘汰页面。

(2)若未找到最佳淘汰页面,则将指针重新返回开始位置,寻找访问位为 1 且不在预调区域中的页面;同时,将扫描过的页面的访问位置为 0。

(3)若还未找到,则再从原来的位置开始扫描。此时一定可以找到。

目前要解决的问题是,如何来比较预调中的页面号和在内存中的页面号。

定义:

$N=\{0,1,2,3,\cdots,n\}$

表示某进程在生存周期内所使用的所有页面的集合。

$M=\{0,1,2,3,\cdots,m\}$

表示系统为该进程分配的页框数中的页框号(从 0 开始编址)。由于内存中的页框数有限,而进程空间越来越大,所以通常情况下,有 $M<N$。

P 为预调页面号的集合,Q 为当前处在内存中的进程页面号的集合,则有 $P\in N, Q\in N$。

对于如何匹配预调页面中的页面号与已存储在内存中的页面号,如果该进程空间很大,可能需要预调的页面号就会有很多,对于不存在的页面号则需从预调表的开始一直查找到该表的表尾,这会给系统带来较大的开销。为了加快查找的速度,引入杂凑(hash 函数)技术,用 $hash(Q[i])$ 来找与 $Q[i]$ 集合中数据项相匹配的页面号。即找出在集合 Q 中但不在集合 P 中的页面号。设页面号集合为 T,则 T 可以用数学式表示为:

$T=\{t|t\in P \&\& t!\in Q, t\in N\}$

该算法的优点是不会淘汰将要使用到的页面,即会降低缺页率以及提高命中率。

结束语 每种算法都有自己的优缺点,同样,上面的这种算法也有缺点,即当预调中的页面没有用到时,同样会发生缺页。因此以上算法还需要进一步的完善及验证。

参 考 文 献

[1] 左万历,周长林,彭涛. 计算机操作系统教程(第 3 版)[M]. 北京:高等教育出版社,2011

[2] 彭青松,丁祥武. 一种改进的自适应页面置换算法[J]. 计算机应用与软件,2011,28(2):67-70

[3] 李芳,徐丽,陈亮亮. LRU 近似算法的研究[J]. 现代电子技术,2009,32(10)

[4] Bansal S, Modha D. CAR: Clock with adaptive replacement [OL]. http://theory.stanford.edu/~sbansal/pubs/fast04.pdf

[5] Wang Hong-bo. LRU-based Algorithm for Identifying and Measuring Large Flows[J]. Journal of Electronics and Information Technology, 2007, 39(10)

[6] Tanenbaum A S. Modern Operating Systems (Third Edition) [M]. 2009

[7] Chang Yuan-hao, Lin Jian-hong, Hsieh J W, et al. A Strategy to Emulate NOR Flash with NAND Flash[J]. ACM Transactions on Storage, 2010, 6(2): 1-23

[8] Juurlink B. Approximating the optimal replacement algorithm [C]// Pro 1th Conference on Computing Frontiers. April 2004: 313-319

[9] 李占胜,毕会娟,李艳平,等. 一种对 LRFU 置换策略的自适应改进[J]. 计算机工程与应用,2008,44(17):153-157

[10] Jiang Song, Chen Feng, Zhang Xiao-dong. CLOCK-Pro: an effective improvement of the CLOCK replacement[C]// Proceedings of 2005 USENIX Annual Technical Conference. 2005

[11] 蒋飞虎. 动态自适应页面置换算法[D]. 南京:东南大学计算机科学与工程学院,2006:19-21

[12] 张刚园. OS 中衡量页面转换算法的指标研究[J]. 西华师范大学学报:自然科学版,2012(12):403-407