一种保证服务连续性的容灾系统的设计和实现*

陈　倩，刘晓洁，李　涛，赵　奎，皮璐琳，顾启超
(四川大学　计算机学院，成都 610065)

摘　要：设计并实现了一种能够保证服务连续性的容灾系统。系统在设备驱动层使用实时监控写操作的方法，通过 Internet 将数据备份到多个远程容灾中心；采用多种数据恢复策略，实现数据的快速恢复；提供了服务自动切换机制，保证系统能够对外提供不间断服务。此外，该系统支持多种操作系统和数据库，并提供了基于 Web 的配置管理方式。

关键词：容灾；多点备份；失效检测；服务切换；数据恢复
中图分类号：TP393.08　　**文献标志码**：A　　**文章编号**：1001-3695(2008)08-2444-03

Design and implementation of disaster tolerant system guaranteeing service continuity

CHEN Qian, LIU Xiao-jie, LI Tao, ZHAO Kui, PI Lu-lin, GU Qi-chao
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: This paper presented a disaster tolerant system for supporting service continuity. It monitored writing requests on device driver level in real-time, and backuped them in several remote tolerant centers at the same time through Internet. Several recovery strategies had been used to support rapid data recovery. An automatic service switch mechanism was achieved to guarantee that the remote servers could provide continuous service instead of local servers during the disaster. Moreover, this system supports various operating systems and databases, and provides a Web-based system configuration.

Key words: disaster tolerant; multi-points backup; failure detection; service switch; data recovery

随着信息技术的高速发展，信息系统的可用性和灾难恢复能力逐渐成为企业生存的关键。一旦发生灾难，造成信息数据丢失，将带来无可估计的损失。人们在意识到灾难恢复重要性的同时，逐渐开始强调服务的连续性(service continuity)[1]。服务连续性不仅仅需要对数据进行灾难恢复[1]，还包括维持一个企业各项应用服务的正常运营，这对企业的生存发展至关重要。因此，建立容灾系统[1]，保证数据的完整性和服务的连续性，在现代信息社会中必不可少。

传统的备份技术[2,3]，如磁带备份[1]、RAID[4]等，只能在较短距离内实现备份，实施数据备份时一般需要停止服务，备份和恢复时间较长。NAS[4,5]等网络存储技术可实现数据的远距离备份，但需要光纤专线，成本十分昂贵。这些技术通常只是实现了数据的备份，对于服务容灾[1]没有很好地支持，不能保证其连续性。

目前，容灾系统，特别是服务容灾系统的研究和开发主要集中在国外，很多知名外企都有自己研制的灾难备份系统，虽然功能强大，但运行成本较高。国内在这方面的研究则还处于起步阶段，几乎没有自主研发的容灾产品。

基于上述因素，本文设计并实现了一种容灾系统，能够支持多种操作系统和数据库。它利用 Internet 实现数据的异地镜像和恢复，并能够在多点处同时进行数据的实时备份。此外，还提供了服务切换功能，当发生灾难时，可将本地服务转移到远程，由远程向外提供不间断服务，从而保证了服务的可连续性。

1　系统设计

1.1　体系结构

本文所提及的容灾系统包括本地生产中心和远程容灾中心两大部分，这两部分结构对称。其体系结构如图 1 所示。其中，远程容灾中心可以配置为多个。

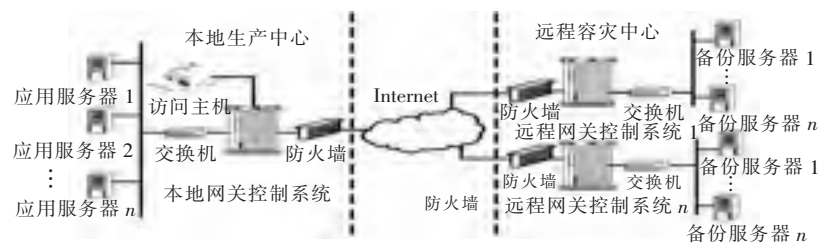

图 1　容灾体系结构

本地生产中心由本地网关控制系统和应用服务系统组成；应用服务系统由一个或多个应用服务器组成，它与本地网关控制系统之间，通过内部高速网络连接。远程容灾中心由一个或多个远程网关控制系统和远程备份系统组成。本地生产中心和远程容灾中心之间通过 Internet 连接。

此外，远程备份服务系统与本地应用服务系统结构相同，

收稿日期：2007-08-10；修回日期：2007-10-25　　基金项目：国家自然科学基金资助项目(60573130,60502011)；国家教育部新世纪优秀人才计划资助项目(NCET-04-0870)；国家“863”计划资助项目(2006AA01Z435)

作者简介：陈倩(1983-)，女，四川成都人，硕士，主要研究方向为网络安全技术及应用(cq1983120@163.com)；刘晓洁(1965-)，女，江苏南京人，副教授，主要研究方向为网络安全技术及应用；李涛(1965-)，男，教授，博导，主要研究方向为网络安全技术及应用；赵奎(1972-)，男，重庆人，博士，主要研究方向为网络安全技术及应用；皮璐琳(1983-)，女，重庆垫江人，硕士，主要研究方向为网络安全技术及应用；顾启超(1983-)，江苏扬州人，硕士，主要研究方向为网络安全技术及应用.

由多个备份服务器组成。灾难发生时,可由远程备份服务系统接管本地应用服务。系统中,远程网关控制系统可以配置为多个,保证了本地数据能够在多点同时进行备份。

1.2　模块结构

系统的模块关系如图 2 所示。

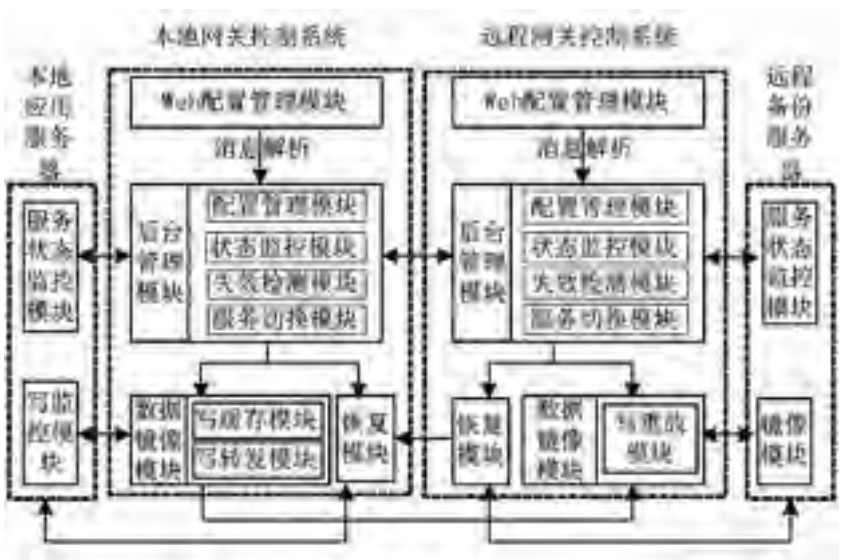

图 2　系统功能模块结构

1) 本地应用服务器和远程备份服务器

其结构对称,部分模块功能相同。

a) 服务状态监控模块。用于监控应用服务系统状态;

b) 写操作监控模块。用于监控并截获本地磁盘的写操作,并且能够将该写操作同步到本地网关相应的磁盘分区;

c) 镜像模块。为远程网关控制系统写操作重放模块提供操作对象。

2) 本地网关控制系统和远程网关控制系统

本地网关控制系统和远程网关控制系统可分为四大部分:Web 配置管理模块、后台管理模块、数据镜像模块、灾难恢复模块。各部分又包含了多个子模块,分别是:

a) Web 配置管理模块。提供系统管理的可视化界面,并根据用户的需求完成任务组的配置管理、系统管理等操作。

b) 配置管理模块。对应 Web 配置管理模块,与其一同完成容灾任务的配置。

c) 状态监控模块。监控本地镜像、远程复制以及恢复的状态等。

d) 失效检测模块。监测本地生产服务器的服务状态,当服务失效时激活服务切换功能。

e) 服务切换模块。灾难发生时将本地的应用服务切换到远程,灾难恢复之后再将应用服务切换回本地,以此来保证服务的连续性。

f) 写缓存模块。将本地应用服务器写监控模块截获的写操作缓存到缓冲区磁盘。

g) 写转发模块。将缓冲区中的写操作取出,并链接到多个发送队列,然后发送到远程。

h) 写操作重放模块。在远程容灾中心进行写操作重放,以实现数据的异地备份。

i) 灾难恢复模块。按照失效检测提供的有效数据源和失效点选择恢复策略。

2　系统实现

容灾是一项系统工程,其流程如图 3 所示。

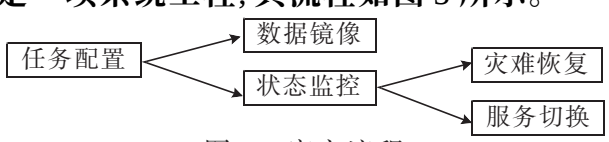

图 3　容灾流程

2.1　任务配置管理

任务配置主要是进行容灾任务的配置,由 Web 配置管理模块、后台管理模块以及数据镜像模块等共同负责,所有操作都是基于 Web 页面进行的。

容灾任务是指针对一种服务而进行的容灾功能配置,包括镜像配置、恢复策略配置,以及服务切换功能的配置。本系统是基于磁盘数据容灾之上的服务容灾。因此,容灾任务配置指一个操作单元,这个单元包含了特定逻辑关系,分布于多台主机的一组磁盘设备,还包括了对这些磁盘进行数据备份恢复以及服务切换等服务中的配置信息。

总的来说,容灾任务 $T = \langle G, M, R, S \rangle$。其中:$G$ 表示具体的一个任务组;M 表示镜像策略;R 表示为容灾任务配置的恢复策略;S 表示任务状态。

G 中包含了一组相关的镜像设备,设备之间有如下的联系:

$$G = \langle LDC, DRC_1, DRC_2, \cdots, DRC_n \rangle$$
$$LDC = \langle LSP, LGP, LGP' \rangle$$
$$DRC = \langle RGP, RSP \rangle$$

其中:LDC 表示本地数据生产中心;DRC 表示远程数据容灾中心;LSP 表示本地磁盘分区;LGP 表示本地网关磁盘分区;LGP′ 表示本地网关缓冲分区;RGP 表示远程网关数据备份分区;RSP 表示远程服务器备份分区,并有关系:LSP = LGP = RGP = RSP,即四者大小相等。

任务组配置好之后,就可以对每一个容灾任务进行相应的容灾操作,如状态监控、数据镜像等。

2.2　状态监控

状态监控负责对容灾任务状态、生产中心和容灾中心状态的监控。具体信息包括主机存活状态、网络状况、数据镜像情况、容灾任务组运行状态等。位于后台的状态监控模块从各个模块上收集到以上信息后,会以消息的形式将其传递给上层 Web 配置管理模块。通过对消息的解析、过滤等处理,配置管理模块会将状态信息显示在页面中,从而为用户提供了可视化界面。

2.3　远程数据镜像

数据镜像是保证服务连续性的基础,也是整个容灾系统的关键。这一功能由数据镜像模块来完成,整个镜像过程可以分为三个阶段,如图 4 所示。

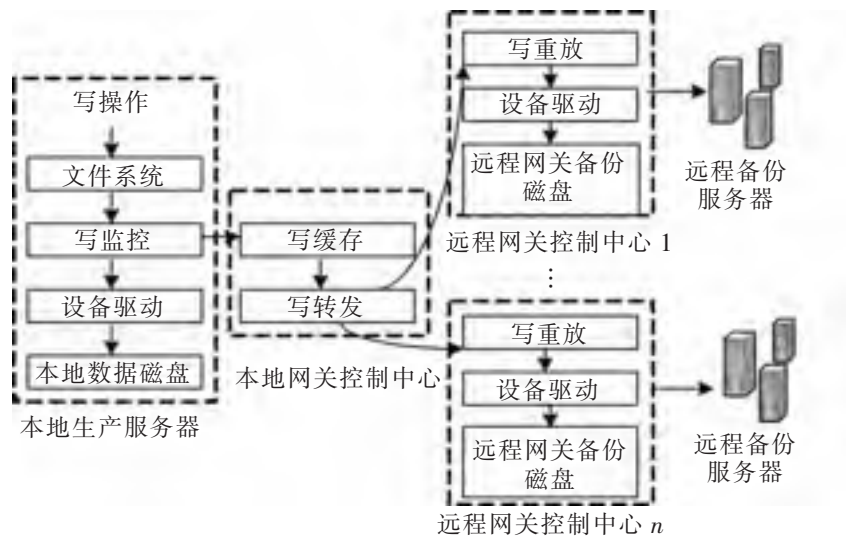

图 4　远程数据镜像过程

1) 本地应用服务器磁盘分区的写操作截获

当本地应用服务器发出写数据请求时,写操作监控模块截获写操作,并将截获的写操作进行封装,通过高速的本地网络发送到本地网关控制系统进行缓存,然后再将写操作实际向设备驱动提交,完成对物理磁盘的写操作。

2) 本地网关控制系统写操作的缓存

考虑到本地网关控制系统与远程网关控制系统之间通过

Internet 进行通信,网络速度慢、不稳定,因此在将截获到的写操作发送到远程网关之前,首先提交给本地网关的缓冲磁盘进行缓存。同时,本地网关上还有一个转发模块,负责从缓冲磁盘中取出写操作,并将其链接到多个发送队列,转发到多个远程网关控制系统并等待确认。待所有的远程控制系统确认后,销毁缓存中的数据。

3) 备份系统写操作的重建和提交

远程网关控制系统在收到写操作数据后,会完成写操作的重放,将远程网关上灾备分区的写操作同步到远程服务器上。最后向控制系统进行确认,从而完成本地数据到远程镜像的全过程。

2.4 服务切换

服务切换的目的在于提供一个有效的机制,使得故障或灾难发生时,远程容灾中心能够代替本地生产数据中心,提供连续性的服务,使外界觉察不到服务的中断,保证服务的可持续性。本系统中,服务切换主要由失效检测模块、服务切换模块共同完成。

2.4.1 失效检测模块

失效检测模块能够实现对网络通信和系统服务状态信息的检测,是服务切换功能实现的基础和条件。本系统中,失效检测[6,7]由检测客户端和检测服务端组成。失效检测客户端负责对本地服务监控模块获取的数据进行分析。将获取到的相关状态信息,以消息的形式发送到检测服务端,保证实时的失效检测。服务端根据所收到的容灾任务状态,将其分为三类,即正常运行状态(TRUST)、被怀疑状态(SUSPECT)和失效状态(FAILURE)。本系统中采用了基于 PUSH 模型的失效检测算法,并运用了 CHEN 等人[8]所提出的预测算法。其算法描述如下:

```
Failure_Detector( )
{
    s_i. state = SUSPECT;
    T_{i+1} = System Estimate time;
    for( s_i ∈ S)
    {
      receive messages m_j at time t;
    if ( t < = T_i &&j > = i || T_i < t < = T_{i+1}&&j > = i)
      s_i. state = TRUST;
    else
      s_i. state = FAILURE;
      }
}
```

其中:S 表示服务器状况列表;$s_i(s_i \in S)$ 保存着服务器的状况信息;t 表示实际收到心跳消息的时间。待检测端会按一定周期定时地向检测器发送心跳消息。检测器会估计出本次心跳时间 T_i 以及下一次心跳时间 T_{i+1}。如果在 T_i 时,已收到消息 $m_j(j \geq i)$,则将 s_i 为 TRUST,表示正常工作;反之,设为 SUSPECT 表示怀疑。如果在$[T_i, T_{i+1}]$时间内收到了消息,则也将 s_i 状态设为 TRUST。如果在上述两个阶段均未收到消息,则认为服务端失效,将其状态设为 FAILURE。

2.4.2 服务切换模块

当失效检测模块检测到生产中心应用服务系统中存在的失效状态,便将工作移交给服务切换模块,由其完成最后的服务切换工作。

基于系统的拓扑结构,本系统采用了 IPTables 中的 DNAT(destination NAT)将 IP 报文重定向,从而实现对数据内容提供访问切换的服务切换方式。这种切换方式下,只需要将服务请求进行重新定向,无须保持会话状态,而具体的事务由上层的应用来保证。该服务切换方式操作起来简单方便,可以大大降低系统的成本。

当服务器监测模块检测到本地服务器出现故障时,服务切换模块动态地在 IPTables 的防火墙 NAT 规则中添加一条 DNAT 规则,将访问本地生产中心的请求发送到远程容灾中心,由远程容灾中心的备份服务器对外提供服务;当本地服务器从故障中恢复并且数据恢复完成后,服务切换模块会删除该 DNAT 规则,访问本地容灾中心的请求仍然发送至本地应用服务器。

2.5 数据恢复

数据恢复模块负责灾难发生后,提供对生产中心数据的快速恢复。系统中提供了多种恢复策略和恢复方式。恢复策略有全恢复、快速恢复和定时恢复。全恢复是对所有的数据都进行一次重新拷贝。快速恢复是一种差异恢复,通过计算应用服务器和磁盘阵列上数据的校验值,找出更新过的数据,并针对这部分数据作拷贝。而定时恢复是由用户自行设定恢复的时间,当达到用户指定时间时,系统将自动进行恢复。恢复方式有手动恢复和策略恢复。手动恢复是由人工选择可靠的数据源进行恢复,多用于发现数据存在不一致的情况下;策略恢复只需人工选择灾难发生点,然后系统会智能选择恢复源,以最低的代价实现数据的恢复。

3 结束语

本文设计并实现的容灾系统,不仅仅提供了数据的异地备份和快速恢复功能,还使得数据能够同时备份在多个异地容灾中心,并提供了服务自动切换功能。灾难发生时,本地应用服务能够迅速切换到远程,由远程向外提供不间断服务,从而保证了系统服务的稳定性和连续性。同时,该系统能支持多种平台,具有广阔的发展前景。

参考文献:

[1] 李涛. 网络安全概论[M]. 北京: 电子工业出版社, 2004: 474-490.

[2] HUTCHNSON N C, MANLEY S, FEDRWISCH M, *et al*. Logical vs. physical file system backup[C]//Proc of the 3rd Symposium on Operating Systems design and implementation. Berkeley: USENIX Association, 1999: 239-249.

[3] QIAN Cun-hua, SYOUJI N, TOSHIO N. Optimal backup policies for a database system with incremental backup[J]. Electronics and Communications in Japan, Part III: Fundamental Electronic Science, 2002, 85(4): 1-9.

[4] 韩德志, 谢长生, 李怀阳. 存储备份技术探析[J]. 计算机应用研究, 2004, 21(6): 1-7.

[5] 陈凯, 白英彩. 网络存储技术及发展趋势[J]. 电子学报, 2002, 30(12A): 1928-1932.

[6] 董剑, 左德承, 刘宏伟, 等. 一种基于 QoS 的自适应网格失效检测器[J]. 软件学报, 2006, 17(11): 2362-2372.

[7] 王树鹏, 云晓春, 余翔湛, 等. 一种容灾中间件的设计与实现[J]. 通信学报, 2005, 26(7): 68-75.

[8] CHEN W, TOUEG S, AGUILERA M K. On the quality of service of failure detectors[J]. IEEE Trans on Computers, 2002, 51(5) 13-32.

[9] CHEN Yan, QU Zhi-wei, ZHANG Zhen-hua, *et al*. Data redundancy and compression methods for a disk-based network backup system[C]//Proc of International Conference on Information Technology. 2004: 778-785.

[10] KOTLA R, DAHLIN M. High throughput Byzantine fault tolerance[C]//Proc of International Conference on Dependable Systems and Networks. 2004: 575-584.