基于 Fibre Chanel 组播的远程数据备份体系结构*)

秦磊华[1] 余胜生[1] 周敬利[1] 张宗平[2]
(华中科技大学计算机学院信息存储系统教育部重点实验室 武汉 430074)[1]
(广东出入境检验检疫局信息中心 广州 510623)[2]

摘 要 数据备份是容灾的基础。本文提出了一种基于 FC 组播服务的远程数据备份方法,对其工作原理和所涉及到的粗波分复用和流量控制等关键技术进行了研究,对基于 FC 组播的数据备份系统的性能和所属的容灾层次等内容进行了分析。

关键词 光纤通道,备份,组播,粗波分复用,流量控制

The Architecutre of Remote Data Backup Based on Multicast of Fibre Channel

QIN Lei-Hua[1] YU Sheng-Sheng[1] ZHOU Jing-Li[1] ZHANG Zong-Ping[2]
(Shool of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)[1]
(Computer Information Center, Guangdong Entry/Exit Inspection and Quarantine Bureau of the P. R, Guangzhou 510623)[2]

Abstract Backup is the basis of Disaster Recovery(DR). A new method based on multicast of Fibre Channel to back-up data to remote sites is proposed and key technology including CWDM(Coarse Wave Dense Multiplexing) and flow control is analyzed. Meanwhile, it's performance and hierarchie of DR are studied.

Keywords Fibre channel, Backup, Multicast, CWDM, Flow control

1 引言

信息时代,数据已成为企业重要的信息资源,其可用性已成为企业竞争能力的重要指标,越来越多的企业开始意识到信息可用性的重要意义。

2005 年,国务院信息化办招集了银行、电力、铁路、民航、证券、保险、海关、税务等八大重点行业的主管部门,制定了《重要信息系统灾难恢复规划指南》的框架及内容。在此推动下,国内业务连续与灾难备份应用快速增长,且应用领域正由传统的银行、电信、保险等数据密集型企业向制造、能源、烟草、物流等更为广泛的行业拓展。

建立数据的远程备份是容灾的基础。本文提出了一种基于光纤通道交换机组播机制的数据备份新方法,并对其工作原理、关键技术、数据备份的性能进行了分析。本文的研究结果为容灾提供了一种新型高性能的数据备份解决方案。

2 数据备份的基本方法

目前,容灾系统中常用的数据备份方法主要有基于主机、基于存储设备以及基于 SAN 的数据复制。

基于主机的数据复制工作在主机的卷管理层,通过运行于备份服务器中的卷管理软件数据复制功能,将业务系统的源数据复制到备份中心。这种方式的数据复制需要占用主机的 CPU 资源,对主机的性能有一定的影响。虽然与存储子系统的类型无关,但与业务系统及操作系统平台有关,且管理较复杂。

基于存储设备的数据复制工作在存储设备层,具有管理简单、对应用透明以及高性能等优点,但该方法对存储系统的兼容性差,不同厂家生产的存储设备之间往往不能支持基于存储设备的数据复制。

基于 SAN 的数据复制是基于 SAN 的虚拟卷管理工具并利用快照技术来实现数据的复制的一种方法,同基于存储设备的数据复制类似,复制过程中数据流不再经过网络而是通过 SAN 直接从一个存储设备复制到另一个存储设备,具有良好的存储设备兼容性和应用平台兼容性,是一种高性能、高可靠性的数据复制方法。

本文提出的基于 FC 组播的远程备份方法也是一种基于 SAN 层次的数据复制方法,但与非组播方式的基于 SAN 层次的数据复制相比,它的性能更高,灵活性更强。

3 基于 FC 组播的远程数据备份

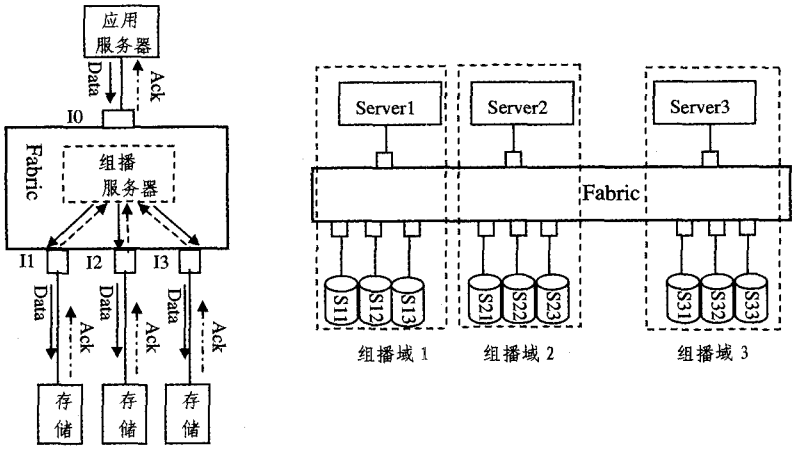
FC 支持的服务共有 6 类。其中第 6 类是可靠的组播服务[1],它是一种支持带确认组播的单向专线连接,对高可用系统中的有效数据复制具有重要的实用价值。本文提出的基于 FC 组播的远程数据备份就是利用 FC 交换机的第 6 类服务来设计的。

3.1 体系结构

基于 FC 组播的远程数据备份建立在光纤通道交换机基础上,其体系结构如图 1 所示。

图 1a)所示的基本结构中,应用服务器和 3 个存储设备通过 FC 交换机的端口 I0～ I3 相连并组成一个组播组,应用服务器作为组播的数据源节点,3 个存储设备作为组播的目标节点,源节点和目标节点间通过单向专线连接。

*)本文得到国家自然科学基金项目"基于冗于智能存储通道的简约容灾存储系统关键技术研究"(60373088)资助。秦磊华 博士生,副教授,研究方向为网络存储与网络安全。余胜生 教授,博导,研究方向为多媒体通信和网络存储系统。周敬利 教授,博导,研究方向为多媒体通信和网络存储系统。张宗平 工程师。



a) 基本结构　　b) 多组播域体系结构

图 1　基于 FC 组播的远程数据备份体系结构

组播过程中，应用服务器通过 I0 将数据发送到 FC 交换机，组播服务器复制该数据帧，并通过 I1～ I3 传送到与之相连的存储设备中。各目标存储节点正确收到组播数据帧后均会向组播服务器发确认 Ack 帧，组播服务器将来自于多个不同目标节点的确认帧汇集成一个确认帧并返回到源节点。

图 1a)所示的只是一种基本的结构，实际应用中可以根据容灾的需要建立多个组播域，实现灵活的数据分类备份。图 1b)为支持多组播域的基于 FC 组播的远程数据备份体系结构。图中共建立了三个组播域，彼此相互独立，互不影响。该体系的灵活性体现在 3 个方面：

1)可根据应用的需要构建多个独立的组播域，实现数据的分类备份。

2)在每个组播域中混合使用磁盘阵列和磁带机，同时支持数据备份和归档。

3)将每个组播域中的存储设备存放在不同地点，建立分处于不同存放地的多个数据备份，大大提高了数据的可用性。

3.2　组播配置

配置基于 FC 的组播包括下列几个步骤：

1)交换结构和 N 端口注册

只有通过注册才能获取相关的操作参数，注册的顺序是先执行与交换结构的注册再执行 N 端口注册。

交换结构注册完成三个功能：a)确定网络的拓扑结构，包括点到点、仲裁环还是交换结构。b)N 端口获得全网唯一的端口标识符。c)为链路协商缓冲区到缓冲区的可靠性值(BB_Credit)。

完成与交换结构的注册后，N 端口它将和每一个通信的 N 端口执行 N 端口注册，以完成下列功能：a)协商 N 端口之间通信的操作参数，主要包括信息源分配超时值 R_A_TOV、差错发现超时值 E_D_TOV、组播支持以及修改缓冲区到缓冲区的可靠性值等。b)初始化端到端可靠性值(EE_Credit)。

2)组播组建立

组播组的建立是通过别名服务来完成的[2]，其流程如图 2 所示。

(1)组播源 N 端口向别名服务器发出加入别名组的服务请求 JNA(Join Alias Group)。

(2)别名服务器向交换架构控制器申请别名组 GAID (Get Alias Group ID)的控制信号。交换架构控制器接受申请，并通过接受信号 ACC 向别名服务器传送别名组地址标识符。

(3)别名服务器向相应的 N 端口发出加入别名组的命令 NACT(Nx_Port Activate Alias Group)。

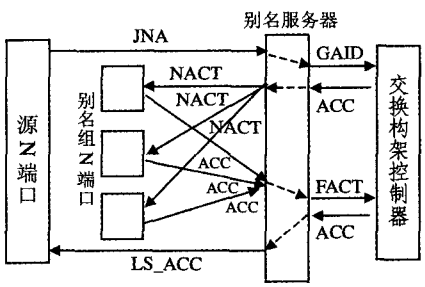


图 2　组播组的建立流程

(4)别名服务器接收到 N 端口接受加入别名组的请求信号 ACC 后，向交换架构控制器发出激活该别名组的信号 FACT(F_Port Activate Alias Group)。

(5)接收到交换架构控制器接受激活别名组的信号后，别名服务器向组播源 N 端反馈其别名组申请成功的信息 LS_ACC(Link_Service Accept)，组播组建立成功。

3)组播服务连接建立

源 N 端口向交换构架 Fabic 发出一个带 SOFc1 起始符的请求帧，交换构架通过识别该帧中的组播地址，在源 N 端口和处在该组播域中的其它目标 N 端口之间建立单向电路连接。如果各目标 N 端口接受该类服务连接，则向发连接确认信号 ACK_1，组播服务器汇集各目标 N 端口对该连接的响应，并向源 N 端口返回一个确认信号 ACK_1，至此，组播服务连接建立成功，接下来，源端 N 端口就可以通过组播实现数据备份操作了；反之，如存在部分目标 N 端口不支持该类服务连接，组播服务器会向源 N 端口发拒绝信号，此次组播服务连接失败。

3.3　组播过程中确认帧的处理

组播过程中，各目标节点反馈给组播服务器的确认帧受多种原因的影响可能出现异常，组播服务器对异常确认的处理方法如下：

1)如果多个目标节点的确认帧不能同时到达组播服务器，组播服务器需要缓存先期到达的确认信息，等所有的确认

都到达后再向源节点返回确认。

2)如果部分目标节点检测到数据帧出错,这时它向组播服务器发回的将是拒绝帧 P_RJT,组播服务器向源端 N 端口发拒绝信号。

3)部分目标节点在 R_A_TOV 规定的时间内还没有返回信号,组播服务器将向源端 N 端口发超时信号。

根据光纤通道协议的工作原理,事件 1)发生时会影响系统的性能,而事情 2)和 3)的发生将直接影响数据备份功能的实现。为了降低上述事件发生的概率,提高系统的性能和稳定性,在构建基于组播的数据备份应用时,还需要对所涉及的高速互连和流量控制等关键技术进行研究。

4 关键技术

4.1 基于粗波分复用 CWDM 的高速互连

通过前面的分析可知,基于组播的数据复制是一种同步的复制。为此,必须采用高质量和高速率的网络互连实现源、目标端与 Fabric 的连接。粗波分复用 CWDM 是目前最适合的互连技术。CWDM 是在密集波分复用的技术上发展而来的,它通过扩展波长间隔(从 DWDM 的 0.8nm 扩展为 CWDM 的 20nm)和优选波段范围(只采用 S+C+L 波段,避开光纤水峰 E 波段和损耗较大的 O 波段)等方法大大降低了 CWDM 的使用成本。使用 CWDM 作为存储设备与 FC 交换机之间的互连技术具有下列优势:

1)高带宽。目前 CWDM 系统单波长的带宽已达 2.5Gbps,可以适应 1G 和 2G FC 存储应用的需要。而采用多路复用,链路的带宽还能进一步提高。

2)高效数据传输。相对于采用高速以太网和同步光纤网,CWDM 具有更高效、简洁的协议处理流程。目前存储数据传输的协议处理主要有三条途径[3]:

- SAN→Ethernet →WDM(CWDM or DWDM)
- SAN→GFP→SONET/SDH→TON →WDM(CWDM of DWDM)
- SAN →WDM(CWDM or DWDM)

由此不难发现,采用 CWDM 实现存储设备与 FC 交换机的连接具有最高的效率和最简洁的协议处理流程。

3)便于实施,节约成本。

CWDM 比 SONET/SHD 更便于实施,它可以基于单位已有或重新敷设的黑光纤(指没有承载其它的通信业务)实施。目前,不带放大功能的 CWDM 模块单级可支持 80～120km 的距离,而带放大功能的 CWDM 模块可支持更远的距离(如 200km)[4]。因此,本文提出的备份方法非常适合在城域网范围内实施。

图 1b)中,虽然每个组播域中有多个存储设备,但采用波分复用后都只需要一根光纤即可实现与 FC 交换机的互连,大大降低了租用或敷设远程光纤的成本。

4.2 根据备份距离和链路带宽合理设置流量控制的信用值

组播服务中,当源端发送信息的速率高于目标端处理信息的速率时,将导致缓冲区溢出并丢掉部分数据,采用流控机制可以防止这类现象的发生。FC 的组播服务采用的是端到端基于信用度(EE_ Credit)的流控机制[5]。信用度的值即接收缓冲区的数量,代表接收设备接收帧的能力。基于 Credit 的流控原理如图 3 所示。

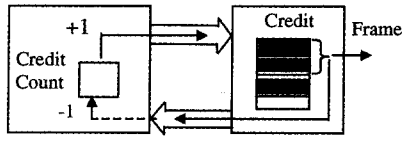


图 3 基于 Credit 的流控机制

1)两个通信节点登录时协商并交换 Credit,发送方将信用度计数变量 Credit count 初始为 0。

2)发送方每发送一个 FC 帧,其 Credit count 值加 1,每收到一个来自接收方的确认帧 Ack_N(N 的取值最小值为 1),Credit count 值减 N;当 Credit count 的值为 Credit 时,发送放将等待确认而不能发送新的 FC 帧。

3)接收方正确收到 N 个 FC 帧将 N 个帧输出或交由高层处理后,便向发送方发确认帧 Ack_N。

随着传输距离和网络带宽的增加,链路的传播延时对系统的传输性能会产生很大影响,为此需要设置较大的缓冲区来减少这种影响。文[5]对 Credit 与链路带宽和传输距离之间的关系进行了研究,得到了下列计算 Credit 的公式:

$$Credit = D(\mathrm{km}) \times LR(\mathrm{bps}) \times 10^{-5} / 1.25FS(\mathrm{bit})$$

式中 D 表示距离,LR 表示链路的带宽,FS 表示 FC 帧的大小,式中系数 1.25 是考虑了 FC 使用 8B/10B 编码对传输效率的影响而得到的。图 4 是根据上式计算得出的链路带宽度分别为 1G 和 2G,FC 帧大小为 1024B 时,传输距离与所需要 Credit 值之间的关系。

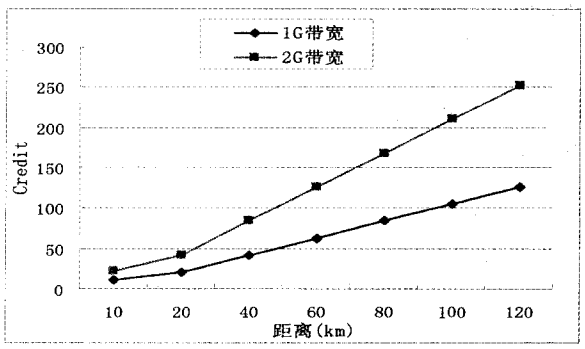


图 4 Credit 与连路带宽与距离之间的关系

由图 4 可知,为了保证链路带宽被充分利用,降低由于接收缓冲区溢出而导致的数据丢失,随着数据备分链路距离和带宽增加,所需要的 Credit 值也越大。图 4 显示,距离为 120km,1G 和 2G 带宽所需要的 Credit 值分别 126 和 252。实际应用中,应根据链路带宽和备份距离设置满足上式的 Credit 值,以提高备份系统的性能。

5 系统的性能分析

链路的最大数据传送速率与物理参数(如时钟频率和传输距离)、协议参数(如信令开销和控制开销)和通信模型等因素有关。由第 4.2 节的分析可知,在合理选择目标端的 Credit 值后,发送端可以处于饱和状态发送,因此可以不再考虑传播时延对传输性能的影响。

图 5 是基于 FC 协议的工作原理,并根据传输距离合理选择 Credit 值后,给出的一个用于计算在全速链路上可获得数据传送带宽的简单通信模型[6]。

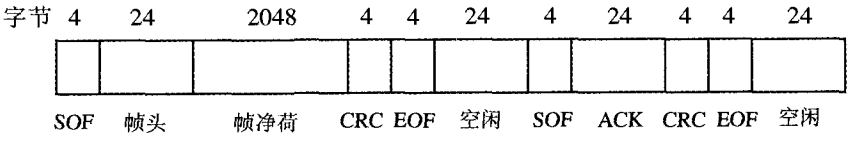


图5 带宽计算中的取样数据帧和ACK传送

SOF为帧头定界符;帧头表示源点、目标点、序列号和其它帧信息;CRC用于检验传送错误;EOF为帧尾界定符;空闲字节用于检错、同步和嵌入低层确认帧信息;ACK为确认信息。

光纤通道数据传输的基本时钟频率为1.0625GHz,2G的链路采用的是2倍频,考虑FC的0层使用8B/10B编码,则最大有效数据传输率为

$$S_{max}=2\times1.0625(\text{Gbit})\times2048(\text{B})\ /\ [(2048+120)(\text{B})\times10]=200.7\text{MByte/s}$$

尽管光纤通道的帧净荷长度可变,但传输帧需要的开销却是相同的。因此,随着净荷长度的缩短,有效数据传输率也会有所降低;另外,上述计算中没有考虑差错后的重传对性能的影响,由于光纤传输系统的误码率很低(如10^{-13}),因此,正常情况下差错重传的概率很低,重传对性能的影响很小。

结束语 本文利用FC的组播服务提出了一种基新的远程数据复制方法,该方法属于SAN层次的数据复制,能满足SHARE78[7]标准中第6级对数据进行实时备份的要求。该方法具有下列优点:

1)高度的灵活性。不仅与存储设备和应用服务平台无关,而且还可以根据应用的需要,建立多个组播域,实现不用类别信息的分类备份。

2)高性能。组播数据在交换结构层次而不是端口层次进行复制能大大提高带宽的利用率;另外,采用CWDM不仅可以提供存储设备与FC交换机之间的高带宽连接,还大大简化了数据复制的协议处理流程,提高了系统的效率。

3)高可用。可根据应用的需要,建立多个数据备份,提高系统数据的可用性。

随着4G和8G FC的普及,基于FC组播的数据复制方法的优点将更为明显,本文的研究结果将为城域网范围内的容灾提供一种高性能、高可靠的数据复制方法。

参考文献

1 2004 - FC-FS-2 Draft Standard. http://www.t11.org/ftp/t11/pub/fc/fs-2/04-045v0.pdf

2 Fibre Channel Generic Services - 5(FC-GS-5). http://www.t11.org/ftp/t11/pub/fc/ gs-5/06-192 v2.pdf

3 Qin Leihua, Zeng Dong, Liu Gang, et al. Network architecure of strage extension next generation SONET/SDH-based and GFP interface design of SONET/SDH with FPGA. In: Proceedings of SPIE, Network Architectures Management and Applications III, Shanghai, 2005. 965～973

4 Iannone P P, Reichmanna K C, Spiekman L H. Amplified CWDM systems. IEEE, 2003. 678～679

5 Qin Leihua, Yu Shengsheng, Zhou Jingli. Analysis and amendment of flow control credit-based in SAN extension. In: Proceedings of SPIE, Network Architectures Management and Applications III, Shanghai, 2005. 8595～973

6 Benner A F. 存储区域网光纤通路技术. 胡先志,胡佳妮,等译. 北京:人民邮电出版社,2003

7 IBM容灾白皮书. http://www-900.ibm.com/cn/ support/download/ Disaster-recovery.pdf

(上接第96页)

以从很大程度上减少报警数量和降低误报率,并且由于报警威胁度是基于可信度和危险度两方面的综合评价,这对管理员或其它工具进行进一步的分析和响应提供了相当坚实的依据。

表2 报警威胁度表

类别 / 威胁度	IDS报警数	正确报警数
0.1	348	3
0.2	152	5
0.3	55	31
0.4	33	37
0.5	38	38
0.6	30	17
0.7	16	14
0.8	7	7
0.9	11	9
1	3	2
合计	693	163

结论 本文针对当前入侵检测和响应领域的重要问题:过高的IDS误报率,提出了一种基于威胁度的动态报警管理TDAM模型。TDAM模型利用对所防护系统环境的感知,不仅对报警信息可信度而且对其危险度进行了评价,并将二者结合提出了报警威胁度。报警威胁度能更贴切地体现报警信息对系统的重要程度,对系统管理员来说有很高的参考价值,可以根据报警威胁度对报警信息进行统计、分类等管理。本文的实验进一步验证了TDAM模型在报警评价管理方面是有效而可行的。

参考文献

1 Anderson J P. Computer security threat monitoring and surveillance [R]. Fort Washington: James P. Anderson Company, 1980

2 Biermann E, Cloete E, Venter L M. A Comparison of Intrusion Detection Systems. Computers&Security, 2001, 20: 676～683

3 Phillip A P, Martin W F, Alfonso V. A mission-impact-based approach to INFOSEC alarm correlation [A]. In: Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID)2002[C]. Zurich Switzerland: Springer Verlag, 2002. 95～115

4 Qin X. A Probabilistic-Based Framework for INFOSEC Alert Correlation:[Ph. D. Dissertation]. College of Computing, Georgia Institute of Technology, USA, 2005

5 IETF Intrusion Detection Working Group. Intrusion detection message exchange format. http//www.ietf.org/internet-drafts/draft-ietf-idwg-idmef-xml-09.txt, 2002

6 Sandhu R S, Coyne E J, Feinstein H L, et al. Role-Based access control models. IEEE Computer, 1996,29(2):38～47