

文章编号: 1005-0523(2015)05-0110-06

基于 OS 及存储的数据库容灾技术研究

陈小建

(华东交通大学轨道交通学院, 江西 南昌 330013)

摘要: 分析了数据库物理备份与逻辑备份、静态备份与动态备份的相关特性及影响数据库容灾性能的几个关键因素。重点对基于 OS(operation system)及存储的数据库容灾技术进行研究,并以结构图的方式显示了每种容灾技术的工作原理、特点及应用,该研究成果可为后续数据库容灾技术的理论与实践发展提供参考价值。

关键词: 数据库; 备份; 容灾

中图分类号: TP392

文献标志码: A

DOI:10.16749/j.cnki.jecjtu.2015.05.017

数据库的容灾技术是对数据库实现高可用性的重要保障^[1]。一些黑客或恐怖分子经常对数据库网络进行攻击,此外,地震、水灾、火灾、雷击等自然灾害也会对数据库造成损害而无法继续提供数据服务时,这对国家或企业将会造成不可估量的损失^[2]。如果拥有数据库容灾系统,可以及时接管受灾数据库系统,从而保证业务持续性。本质上,数据库系统的容灾技术就是当前对外服务的生产数据库环境的一个备份^[3]。基于此,本文重点对企业生产环境中基于操作系统及存储级别的数据库容灾技术进行阐述。

1 数据库备份分类分析

备份是保证数据库可用性的重要方式。根据备份对象的不同,可以将其分为物理备份和逻辑备份。物理备份的原理是基于数据块的,实现方法使用操作系统命令或者专门的数据库物理备份工具,一般情况下都支持增量备份,恢复对象主要是对数据块、数据文件等物理级别的恢复。逻辑备份的原理是基于具体的数据库对象,如表、索引、过程等;实现方法是使用专门的数据库逻辑备份工具进行备份,一般为数据库对象的备份,但难以提供增量备份,恢复也只能是数据库对象等逻辑层面的恢复。

根据备份的时间特点,又可分为静态备份和动态备份。静态备份基于特定的时间点(RPO),超过特定时间点之后的数据就不能恢复,至于恢复时间(RTO),视备份集的大小、系统 IO 能力等而定;备份的技术难度和维护成本都比较低。动态备份的原理是从某个基准备份开始动态持续备份,可以实现 0 或者接近于 0 的 RPO/RTO,备份的技术难度和维护成本都比较高。

2 影响数据库容灾性能的关键因素

影响数据库容灾的关键因素有: 网络带宽因素; 磁盘 IO 因素; RPO/RTO 指标; 网络安全。这些因素在文献[4]中已做了比较详细的分析,在此不再叙述。

收稿日期: 2014-12-25

基金项目: 江西省科技厅青年自然基金项目(20132BAB211013);江西省教育厅科技项目(GJJ13330);华东交通大学校立科研基金(14GD11;14GD02)

作者简介: 陈小建(1980—),男,工程师,研究方向为计算机应用技术。

3 操作系统级的容灾

操作系统级容灾可以基于以太网、SAN 网络来实现,如图 1 所示。



图 1 基于以太网/SAN 网络的操作系统级容灾原理

Fig.1 The operating system level disaster tolerance principle based on the Ethernet/SAN network

操作系统级容灾的工作方式是通过“基于以太网的容灾”方式,或“基于 SAN 网络的容灾”方式将源端(左边)数据库的变更传递到目标端数据库(右边)上,并在目标端数据库上重现源端的变更,从而使两边的数据达到一致,实现容灾目的^[5]。

应用服务器群代表一个或者多个应用程序运行在这些服务器上面,有的是出于业务种类的不一样,需要使用多台服务器,有的是因为某个业务访问压力太大,需要多个服务器来分担压力。而数据库服务器群可以是一个数据库的集群,也可以是多个数据库的简单组合。最下面的存储是用于存放数据库数据的硬件设备,带库是专门用做备份的磁带设备。

以下从两个方面阐述图 1 的工作原理。

3.1 基于以太网的容灾

“基于以太网的容灾”对应的是图中的“电信部门以太网专线或 Internet”那层。它可以采用诸如 VVR (Veritas Volume Replicator)的卷复制技术实现。VVR 是 Symantec 公司容灾套件 Storage Foundation 系列软件中的一个模块。VVR 复制基于逻辑卷(Logical Volume, LV),复制前先将 DBMS 所在的一个或多个卷定义为 RVG(Replicated Volume Group),数据库向 RVG 中的卷写入数据时,会先在 SRL (Storage Replicator Log)卷中写入日志,日志经过网络传输后在目标端的 RVG 上重现 I/O^[6-7]。SRL 保证数据复制严格按照写顺序进行,当 SRL 卷写满后,DCM(Data Change Map)开始记录变化过的数据块块号,以便在复制正常后仍保持主、备端的数据一致性。

VVR 复制的优点主要有:距离不限;支持异构的存储和操作系统;支持同步和异步复制;默认自适应;在网络延时情况较好、数据能够及时复制;工作在同步方式,当网络延时情况较差,数据不能及时复制;工作在异步方式下,保证主节点的 I/O 性能;VVR 复制严格按照 I/O 发生顺序进行,无论在同步还是异步工作方式下,都能保证数据的完整性。

3.2 基于 SAN 网络的容灾

“基于 SAN 网络的容灾”对应上图中的“电信部门高速光缆专线”那层。在操作系统级,也可以采用 LVM 镜像的方式实现基于 SAN 网络的异地容灾^[8]。业内较多地采用 VxVM(Veritas Volume Manager)卷镜像技术,像 AIX, HP-UX 等操作系统,本身也有 LVM 命令来实现镜像,但由于操作系统 LVM 的局限性,难以作为异

地容灾使用。VxVM 也是 Symantec 公司容灾套件 Storage Foundation 系列软件模块之一。利用 VxVM 卷镜像技术构建容灾系统比较简单,它只有一个条件,就是将生产中心和灾备中心之间的 SAN 存储区域网络通过光纤连接起来,建立城域 SAN 存储网络。然后,就可以通过 VxVM 提供的非常成熟的跨阵列卷镜像技术来实现异地容灾。从原理上讲,在城域 SAN 上的两套存储系统之间的卷镜像,和在同一个机房内的 SAN 上的两个存储系统之间的镜像并没有任何区别。

使用 VxVM 提供的逻辑卷镜像技术有如下优点:

- 1) 易操作,利用城域 SAN 网络和 VxVM 镜像功能,可以非常轻松的实现数据系统的异地容灾。
- 2) 维护方便,消除了复制技术(无论是同步还是异步)的切换的动作,从而保证零停机时间,零数据损失。
- 3) 由于磁盘或者链路等方面的故障导致主生产环境的逻辑卷不能访问时,应用或者数据库不会因为故障而停止,会继续在灾备中心的 LV 镜像上执行 I/O,整个过程对应用透明,不需要也不会中断业务系统的正常运行。
- 4) 当逻辑卷的镜像被破坏时,可以利用 VxVM 的 DCO (Data Change Object) 与 FMR (Fast Mirror Resync) 技术进行镜像的快速增量同步。

但是,逻辑卷镜像技术会受到距离限制,在较近范围内进行,生产中心与灾备中心通过裸光纤将两边的 SAN 环境联接起来。如果距离较远,无法直接部署光纤联接,可以租用运营商光纤将生产中心和灾备中心的 SAN 网络连接起来。常用的技术有 DWDM(Dense Wave Division Multiplexing)技术。

4 存储级别的容灾

在主机或者数据库层面实施数据库容灾方案,除了对存储 IO 性能会有一定影响外,还会消耗部分主机资源以及影响数据库性能。如果使用基于阵列的复制技术,则可以减轻主机与数据库在这方面的负担。

基于阵列的复制技术有本地复制和远程复制两种。

4.1 基于阵列的本地复制与容灾

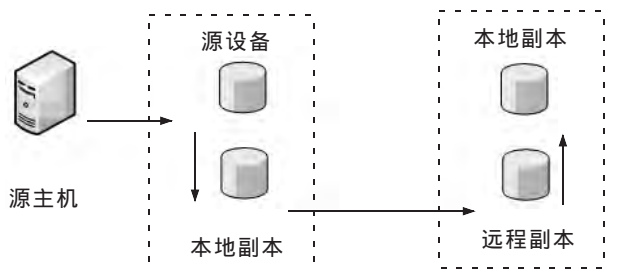
和正常的数据库备份一样,我们也可以在存储端定期对存储设备进行本地复制,创建存储设备的一致性 PIT(point in time)本地副本。再配合远程复制技术,就能实现满足容灾规划要求的 RPO 指标,该种实现方式远程站点的 RPO 通常在小时级。基于阵列的复制技术可以记录源设备和目标设备上的变化情况,因而,所有再同步操作都可以增量完成。容灾过程如图 2 所示。

在 PIT 上,数据或者备份的数据是满足一致性的。而增量备份,是在基准备份的基础上进行的,基准备份之后,如果源设备上面的数据发生了变更,则下次备份时,可以只备份基准备份以后发生变化的那部分数据。

可以用于阵列本地复制的技术主要有:EMC Symmetrix 阵列的 TimeFinder 系列产品以及 EMC CLARiiON 与 Celerra 阵列的 SnapView 与 SnapSure 产品;IBM 存储产品的 FlashCopy 以及 VolumeCopy 技术;HDS 的 ShadowImage 技术等。

4.2 基于阵列的远程容灾

基于阵列的远程存储复制主要有同步、异步模式,比较如表 1 所示。



源主机在源设备上写数据
创建一个源设备的一致性 PIT 副本
源阵列的本地副本远程复制为目标阵列上的远程副本
在目标阵列上创建一个远程设备的一致的 PIT 本地副本

图 2 基于阵列的本地复制与容灾原理图

Fig.2 Local replication and disaster tolerance principle based on the array

表 1 基于阵列的远程容灾复制比较

Tab.1 Comparison of remote disaster tolerance replication based on array

项 目	同步复制	异步复制
站点间距离限制	最好在 100 km 以内,否则距离越长响应延时越大	无限制
主机端 IO 响应	必须保证主站点与容灾站点的写操作都完成,相对较慢	主站点完成就算完成,相对较快
带宽要求	高,要求满足 IO 峰值要求	低
对主站点的影响	较大	小
RPO	0 或者接近 0	分钟或者小时级别

与同步复制相比,异步远程复制提供了一种非 0 的 RPO 灾难恢复解决方案,RPO 取决于站点缓冲区大小、网络带宽、源端写负载。

远程站点的数据副本可以在复制分离操作下被正常访问,可用于备份、测试、数据仓库报表生成、以及决策支持等。在数据副本使用的过程中,副本数据由于 IO 操作发生变化。当副本数据使用完毕,可以丢弃副本数据变化重新建立复制关系,或者采用还原操作保留远程数据副本而丢弃主站点数据变化继续进行复制。

典型的用于阵列远程复制的技术有:EMC Symmetrix 支持的 SRDF, CLARiiON 支持的 EMC MirrorView, EMC SAN Copy, HDS 厂商的 TrueCopy, HUR, 以及 IBM 的 PPRC, Remote Mirror, HP 的 BusinessCopy 等等。

为了进一步减少主站点上存储的 I/O 压力,一些存储厂商把容灾复制过程中的“推数据”方式改成“拉数据”方式,容灾环境会主动地将数据从主站点拉回,如 HDS HUR 复制技术。

通常在两站点存储复制中,当主站点出现故障,可以将应用切换到容灾站点继续服务,而当两站点间网络失效或者容灾站点出现故障时,主站点能继续工作,只是没有了远程灾备保障。在同步复制模式下,两站点间的距离通常较近,当发生区域性灾难时,两站点可能同时遭到破坏,从而导致额外的 RPO/RTO。区域性灾难一般不会影响到异步复制模式下的容灾站点,因此,在不考虑成本的情况下,三站点复制可以用来缓减两站点复制的风险同时又能满足低的 RPO/RTO 要求。

4.3 三站点容灾

三站点容灾又叫两地三中心容灾,有如图 3 所示三站点级联容灾和三站点星型容灾两种。

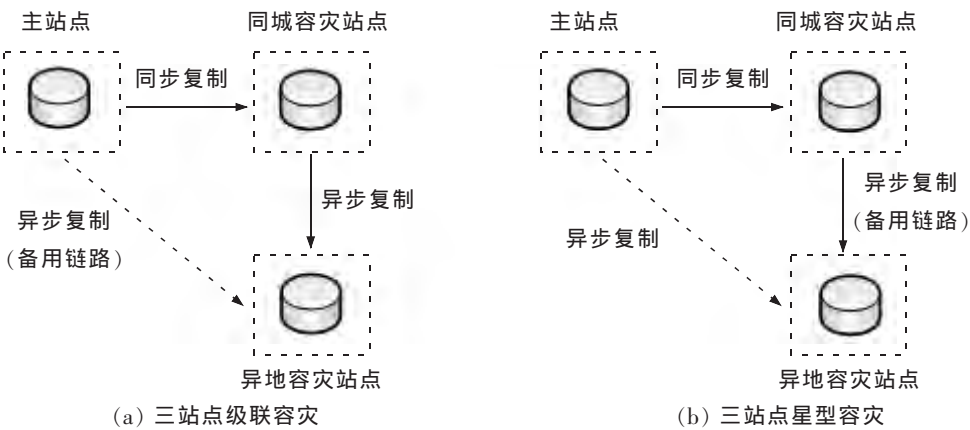


图 3 三站点容灾结构原理

Fig.3 Three-site disaster tolerance structure principle

三站点容灾中,主站点与同城容灾站点距离较近,采用同步复制模式,当主站点出现灾难性故障,可以快速切换到同城容灾站点,理论上可以实现 0 或者接近于 0 的 RPO 与 RTO。然而,在区域性的大灾难中,主

站点和同城站点都有可能被破坏,这时,异地容灾站点可以接管业务系统。由于主站点或同城容灾站点到异地容灾站点间采用的是异步复制模式,从而会存在有限的 RPO。在级联式容灾中,异地容灾站点的 RPO 取决于同城容灾站点失效与主站点失效之间的时间,以及异步复制过程的延时;在星型容灾环境中,异地容灾站点的 RPO 则主要取决于异步复制过程的延时。由于区域性灾难通常不会影响到异步复制模式下的目标站点,因此,三站点容灾可以有效预防区域性灾难。图 4 是一个级联式三站点容灾场景。

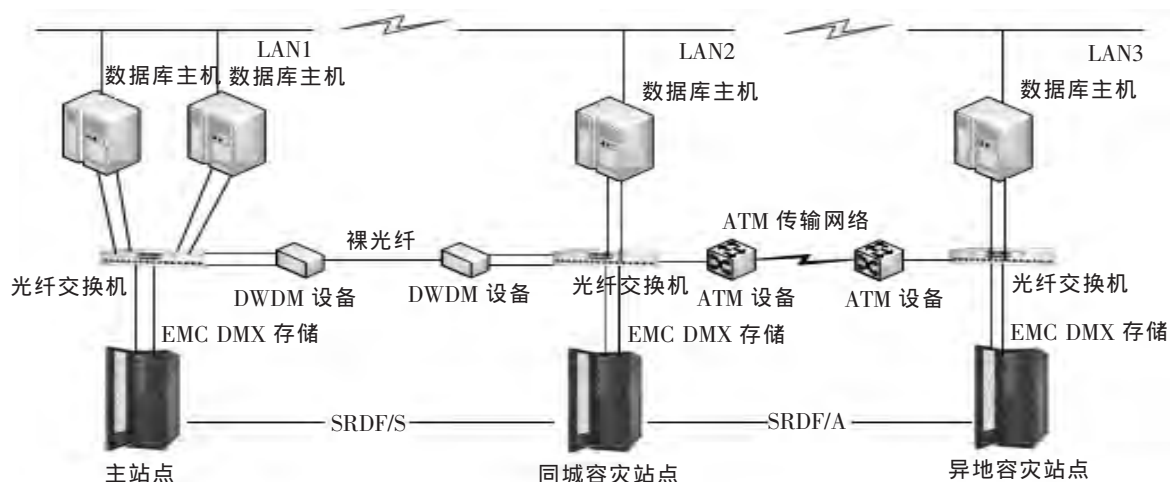


图 4 级联式三站点容灾结构

Fig.4 The cascading disaster tolerance structure principle of three sites

图 4 涉及到 3 个站点之间的数据复制,其中主站点与同城站点之间采用了 SRDF/S(同步复制)技术,该技术采用 DWDM 设备,能够在同一根光纤中,把不同的波长同时进行组合和传输。在给定的信息传输容量下,可以减少所需要的光纤的总数量。这是基于 EMC Symmetrix 阵列的复制技术。同城站点和异地容灾站点之间采用的是 SRDF/A(异步复制)技术,该技术采用了 ATM 传输网络,ATM 是以信元为基础的一种分组交换和复用技术。采用面向连接的传输方式,将数据分割成固定长度的信元,通过虚连接进行交换,集交换、复用、传输为一体,具有高速和支持许多种数据类型传输的特点。关于同步复制与异步复制的具体比较可参考上述“基于阵列的远程容灾”的分析。

三站点容灾形式比较灵活,不局限于某个存储厂商的某种技术,可视情况混合使用,甚至可以将主机 LVM 层的复制技术与存储复制技术混合起来构建三站点容灾环境,如图 5 所示。

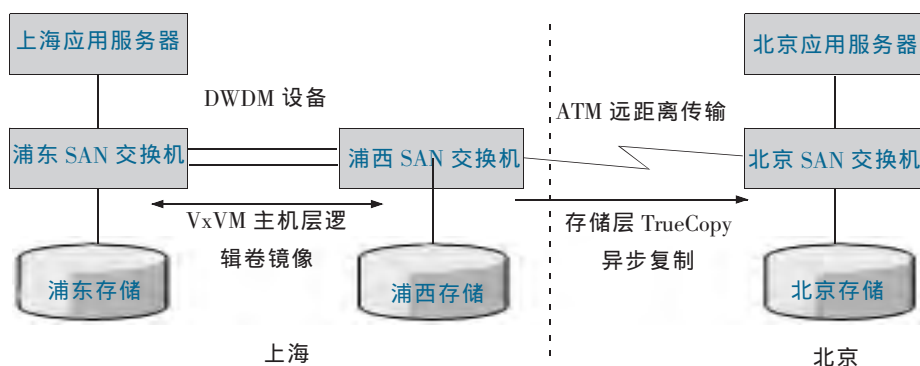


图 5 主机 OS 层复制技术与存储复制技术混合的三站点容灾实例

Fig.5 Three-site disaster tolerance example of the host OS layer replication technology combined with the storage replication technology

在实际实施中,基于对 IDC 相关的各项 IT 基础设施及配套设施建设的成本考虑。企业往往会适当降低对容灾站点的要求,甚至可能会由于异地容灾站点启用的可能性较小,使得异地容灾站点不(完全)具备业务接管的能力,只是作为一个数据备份站点投入使用。同时,为了缓减主站点的访问压力,将 ETL 或者数据查询等业务部署在容灾站点上,以充分利用容灾站点的资源。

5 结束语

实际上,数据库的容灾技术多种多样,不同种类的数据库又有其具体的解决方案,本文只是从 OS 以及存储级别综述了数据库中可以通用的部分容灾技术,尤其适合于中、大型企业级数据库。

随着近年来互联网行业的迅速发展,数据库技术也出现了前所未有的挑战。以 Oracle, DB2, SQL Server 等为代表的传统型集中式数据库越来越难以满足海量数据、高并发等要求,甚至在一些超大型企业(如阿里巴巴)中出现了去 IOE(IBM 主机, Oracle 数据库, EMC 存储)实践,分布式数据库技术的发展与应用逐渐兴起。在存储技术方面,也相应的也出现了分布式存储,大数据时代已经来临,从而数据库容灾技术也面临着更高的要求。

参考文献:

- [1] 刘丽洁.数据库容灾技术及容灾实施方案理论的研究[J].科技风,2014(13):25-26.
- [2] 马薇,姜雨.Oracle 数据库容灾备份中的流复制技术研究[J].科技通报,2012(2):182-184.
- [3] 王建波.数据容灾及其相关技术研究[J].数字化用户,2014(7):104-105.
- [4] 曹文琴,朱海燕,刘映球.基于 Oracle 数据库容灾技术的研究[J].制造业自动化,2012,11(21):61-64.
- [5] 刘淑鹤,王芳.数据容灾技术研究[J].网络安全技术与应用,2013(9):45-47.
- [6] 王津,蒋国良.容灾备份技术在人社信息化建设中的应用[J].城市建设理论研究,2014(1):109
- [7] 姜先贵,马瑞涛.IMS 容灾原理及部署策略探讨[J].邮电设计技术,2014(5):25-29.
- [8] 孙玲芳,王成文,徐会.基于语义 Web 的关系型虚拟社区服务发现模型分析[J].华东交通大学学报,2014,31(2):105-111.

Research on Database Disaster Tolerance Technology Based on OS and Storage

Chen Xiaojian

(School of Railway Tracks and Transportation, East China Jiaotong University, Nanchang 330013, China)

Abstract: The related characteristics of database backup, including the physical backup, the logical backup, the static backup and the dynamic backup, were firstly analyzed in this paper. Besides, key factors of database disaster tolerance were reviewed from the viewpoint of database backup. Then, this paper mainly discussed database disaster tolerance technology based on OS (operation system) and storage, and illustrated the working principal, structural features and specific applications through structural diagrams. Research results provide useful reference for the theory and practice of the database disaster tolerance technology.

Key words: database; backup; disaster tolerance

(责任编辑 姜红贵)