DB2 software

IBM

使用 IBM DB2® 通用数据库 V8.1，为需要处理大量事务的电信应用程序提供十秒的故障切换

客户案例分析

IBM 软件部多伦多实验室

目　录

概述

可用性是许多行业的关键成功因素。电信行业也不例外，实际上，它有可能是对系统可用性提出最苛刻要求的行业之一。在一家大型电信企业最近开展的客户项目中，在 @server xSeries™ 335服务器群上运行的IBM DB2 Universal Database for Linux V8.1 提供了超过客户需求的可用性。DB2 UDB V8.1 采用主动/被动式配置来在两台x335服务器上运行客户应用程序，只需9秒时间就能够切换到备用服务器并运行新事务处理。

近来在DBMS市场上，人们对不同厂商产品的可用性发表了不同的看法。本文将介绍不仅仅能够每秒处理3500多次事务，而且能够在少于10秒的时间内完成故障切换并继续处理相同级别的事务的配置，而不是大肆宣传和/或讨论理论上的体系结构优势。本文研究不是内部的可用性展示，而是客户设计用于仿真它们应用程序的数据库最高工作负载。这一故障切换时间足够快到能够适用于甚至最关键的客户环境，因为本系统每年可以处理30次以上的突发故障停机，同时继续提供99.999%的可用性（以每年5分钟的故障停机时间为例）。

系统配置

对于这一客户案例分析来说，我们使用以下硬件和软件配置：

- 两台 IBM @server xSeries 335 服务器，每台配置
  - 2 个 Intel® XEON™ 2.40GHz 处理器
  - 2GB RAM
- IBM TotalStorage FAStT700存储服务器
  - 456 GB 总存储容量
  - 通过 Qlogic 光纤信道主适配器连接两台 x335 服务器
- IBM DB2 UDB for Linux、UNIX、Windows V8.1
- SuSE® Linux™ 8.1 Professional

要　点

主动/被动式配置不能满足这一需求，因为在一个节点发生故障后系统性能将下降

我们配置了系统使用标准主动/被动式接管配置。与共享的数据体系结构不同，存储设备的并行接入可以通过DB2 UDB来实现。FAStT Storage 直接连接到两台x335服务器上的光纤信道适配器，而不是部署昂贵的光纤信道交换机，从而在故障切换时，备用服务器只需简单地安装文件系统并重新运行数据库即可。为了实现快速的故障切换时间，许多系统中都提供了并发磁盘接入并且 DB2 UDB 提供全面支持。但是，基于SMP服务器的一对标准故障切换是高可用性分布式系统最通用并通过实践证明的配置。

高性能应用程序通常要求在一台服务器发生故障的情形下，系统必须维持相同级别的性能。主动/被动式配置不能满足这一需求，因为在一个节点发生故障后系统性能将下降。但是，DB2 UDB V8.1 对备用节点上的所有处理器不需要许可证，从而成为一种经济高效的高性能解决方案。下面介绍的故障切换测试显示在主节点发生故障后系统将维持相同级别的性能。另外，许多客户使用备用服务器来运行其它工作(测试、分段等)。

为了实现故障切换自动化，我们使用高可用性 Linux 项目提供的 Heartbeat 软件(Heartbeat的详细信息请访问http://linux-ha.org/)。本次测试中使用的Heartbeat配置参数的详细信息请见附件。附件还包括可以调整用于实现高事务处理吞吐量比率和 9 秒的故障切换时间的数据库和网络配置参数。

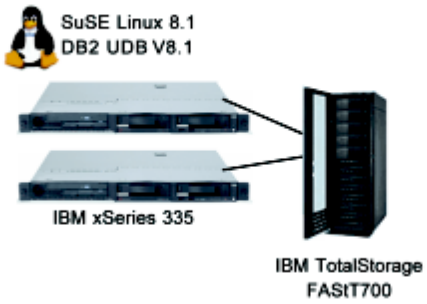

图 1.　高可用性 DB2 UDB 集群

要　点

通过尽可能快地向磁盘写入“脏”数据页面，可以减少故障切换期间需要重放的日志数量，从而缩短恢复时间。

该系统可以实现近乎不变的每秒3575次事务处理的吞吐量标准

吞吐量测试

在测试系统的可用性之前，我们将对吞吐量性能进行测试。这一测试旨在证明使用上面介绍的配置，可以实现本系统中需要的高事务处理比率。调整这一系统来均衡数据库恢复性能和最佳运行时间性能。为了实现最佳恢复性能，您必须尽可能维持磁盘上的数据库结构与存储器中的数据版本一致。通过尽可能快地向磁盘写入“脏”数据页面，可以减少故障切换期间需要重做的日志数量，从而缩短恢复时间。但是，频繁地向磁盘写页面将影响事务处理的性能，因为它们可能需要更多的I/O操作，进行多次处理以把数据写入到磁盘。

对于这一测试来说，我们对该数据库进行了配置，它能够提供高事务处理吞吐量和快速的崩溃恢复处理。该客户的工作负载在电信行业中是最普遍的，包括80%的读事务处理操作和20%的写事务处理操作。

图2显示了本次测试期间测量得到的事务处理吞吐量，本次测试期间系统可以实现近乎不变的每秒3575次事务处理的吞吐量标准。

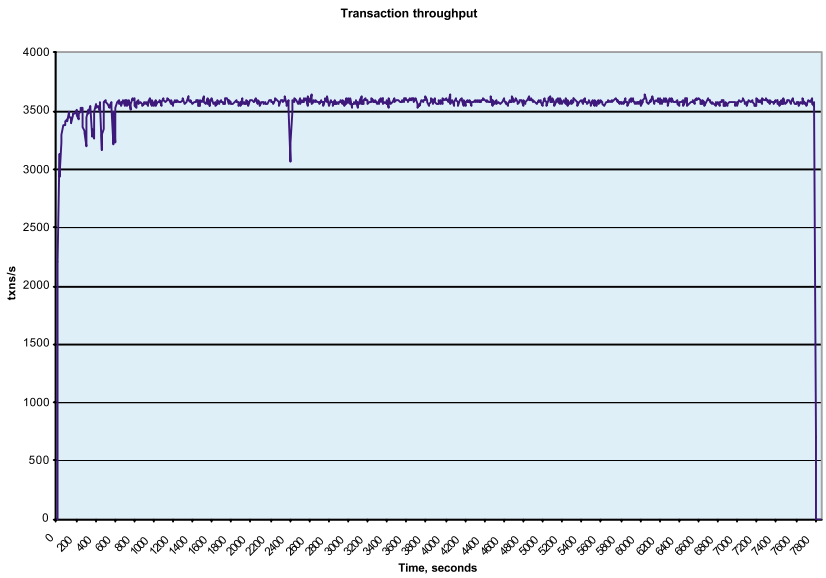

图2.　2小时的吞吐量测试

要　点

对于该客户的工作负载来说，其故障切换时间…… 测量仅为 9 秒

故障切换测试

在故障切换测试期间，我们对数据库、应用程序或配置参数不做任何更改。在测试中我们将进行同样的吞吐量测试并切断主服务器的电源。

故障切换步骤如下：

1. Heartbeat 检测到主服务器发生故障并启动备用服务器上的故障切换流程。
2. 在备用服务器上安装文件系统，从而数据库可用于该服务器。
3. 在故障切换之前启动数据库实例，因此，剩下的最后一步是数据库崩溃恢复流程。

对于这一客户的工作负载来说，其故障切换时间(主节点上的第一个事务处理发生故障和备用节点上第一个事务处理完成之前相差的时间)测量仅为 9 秒。对于这一切换时间来说，使用5个并行的重复操作流程，数据库的重复和撤销处理（称为崩溃恢复）只需 3 秒。

图 3 介绍了故障切换测试期间的事务处理比率。

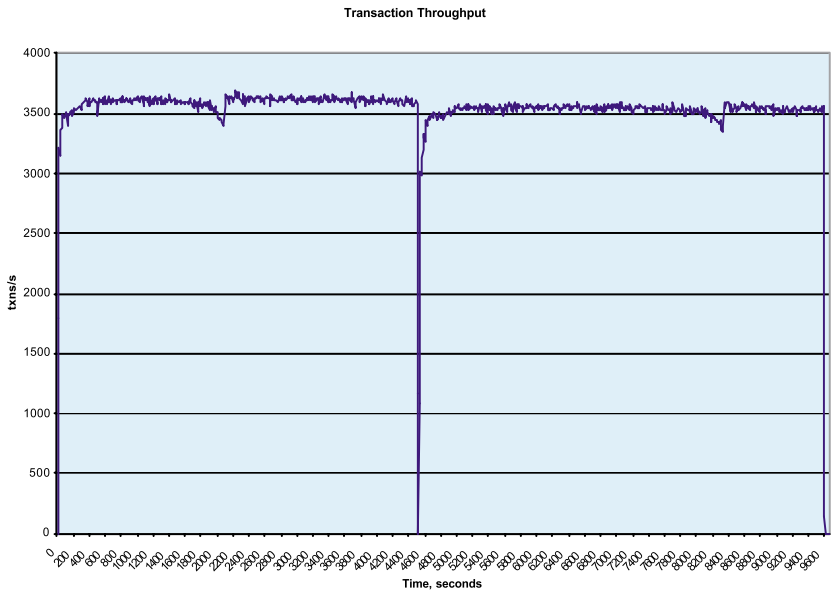

图 3.　2 小时的故障切换测试

正如图中所示，在故障切换前后系统每秒都进行3575次事务处理。这一配置能够使服务器故障远离应用程序性能需求。

要　点

…… 在 3 秒内完成整个数据库崩溃恢复

在如此短的时间内，DB2 UDB 如何能够故障切换这么高吞吐量的事务处理系统？DB2 UDB先进的页面刷新和日志记录功能支持把脏页面从缓冲池异步且频繁地写入到磁盘，从而减少重复处理。在恢复流程期间，DB2 UDB 使用并行重复处理来最大限度地减少崩溃恢复。在本次故障切换过程中，整个数据库崩溃恢复在 3 秒内完成。与大多数在线事务处理(OLTP)工作负载一样，也最大限度地减少了崩溃恢复的撤销处理(在本实例中，在 0.2 秒完成)。从而系统可以在打开数据库之前完成整个数据库恢复且不会对故障切换时间产生任何影响。

该电信客户部署了典型的电信数据库应用程序，它缩短了事务处理持续时间，增加了每秒运行的事务处理数量（这反过来意味着也增加了每秒数据库页面更新的数量）。DB2 UDB 能够与这些变化保持同步并把它们保存到磁盘，且不会显著影响性能，从而支持快速重启恢复，无需进行特殊的磁盘配置或增加软件组件。

结束语

DB2 UDB可以满足许多需要处理大量事务的客户的高可用性需求

正如这一真实的客户电信数据库实施所展示的，DB2 UDB可以满足需要处理大量事务的许多客户的高可用性需求，尤其是那些电信行业的客户。对于这一特殊的应用程序来说，DB2 UDB 提供近乎不变的每秒 3500 次以上的事务处理，能够在 9 秒内故障切换数据库并继续运行事务处理。

所有数据库配置都实现了性能、恢复时间和成本之间的某种均衡。在本次测试中，DB2 UDB 展示了极其快速的恢复，同时继续保留了高事务处理性能。对于要求不太严格的可用性的客户来说，使用相同的系统资源，DB2 UDB可以实现更高的吞吐量。同样，通过均衡恢复时间或吞吐量或两者可以降低系统成本。

在一个节点发生故障后，如果您的运行环境不需要100%的性能，您可以考虑部署了 DB2 UDB 的主动 / 被动式集群，它使用数据库分区特性或使用两个单独 DB2 UDB 实例的相互接管。

您的应用程序在性能和可用性特性方面可能不甚满意，因此我们邀请您使用 IBM 产品，使您能够满足性能和可用性需求，并帮助您的企业取得成功。

附件

SuSE 8.1 Linux **内核配置**

在可用性测试期间，在测试客户机上将 tcp_keepalive_time 和 tcp_keepalive_intvl 参数改为 10，以确保客户机 ip 层检测到服务器故障切换:

```
echo 10 > /proc/sys/net/ipv4/tcp_keepalive_time
echo 10 > /proc/sys/net/ipv4/tcp_keepalive_intvl
/etc/init.d/inetd restart
```

可用性软件

我们通过 ″Heartbeat" 产品来提供 HA 功能。Heartbeat 配置如下:

HeartBeat 版本:	heartbeat-0.4.9e-5
Heartbeat 之间的时间(秒)	2
宣布主机死亡时间(秒)	5
开始死亡时间(秒)	10 (当 Heartbeat 开始时宣布主机死亡时间)
NiceFailback:	on
Heartbeat 模式:	以太网、广播
文件系统:	ext3

文件系统

数据库文件和日志保存在 Third Extended File 系统(ext3fs)中。这是一种日志式文件系统，它与 ext2 文件系统兼容，可以被认为具有日志功能 ext2。由于在 ext3 文件系统中实现了日志功能，因此在崩溃恢复期间无需文件系统检查日志。

DB2 UDB **配置**

吞吐量和故障切换测试的数据库管理器配置是一样的。我们更改了以下参数。

数据库管理器配置			
缺省数据库监视交换机			
Timestamp	(DFT_MON_TIMESTAMP)	=	OFF
Log buffer size (4KB)	(LOGBUFSZ)	=	4
Changed pages threshold	(CHNGPGS_THRESH)	=	20
Number of asynchronous page cleaners	(NUM_IOCLEANERS)	=	20
Number of I/O servers	(NUM_IOSERVERS)	=	10
Log file size (4KB)	(LOGFILSIZ)	=	1000
Number of primary log files	(LOGPRIMARY)	=	100
Percent log file reclaimed before soft chckpt	(SOFTMAX)	=	1

数据页面到磁盘的清洗（flushing）由以下两种参数控制：

- 缓冲池中脏页面阈值(CHNGPGS_THRESH)，脏页面被异步写入到磁盘中。本次测试期间该值设为 20%。
- 崩溃恢复情况下必须从日志重放的数据的阈值，以在故障时恢复未记录在磁盘上的事务处理数据。本参数(SOFTMAX)设为 4MB (日志文件大小)的 1%，以便这些测试实现最佳的恢复时间。

缓冲池配置

所有测试都使用 25,000 4k 页面的缓冲池。

隔离级别

这些测试期间使用 Cursor Stability 的缺省隔离级别。

数据库物理设计

要点是:

- 在物理级，数据库文件和日志存储在所在磁盘中。
- 表格各行之间无明显的划分。

其它信息

使用的数据库接口为 ODBC。

在故障切换之前请先运行备用服务器上的 DB2 UDB 实例。通过运行这一实例，系统将装载并运行主要的 DB2 UDB 流程来进行数据库重启。这是一项可以在备用服务器接管之前执行的可选步骤。忽略这一步将使总重启恢复时间增加 2 秒。

国际商业机器中国有限公司。

■ **北京总公司**

地址：北京市朝阳区工体北路甲二号
盈科中心 IBM 大厦 25 层
邮编：100027
电话：010-65391188
传真：010-65391688

如需更多信息请拨打全国免费直拨电话：
800-810-1818 转 5017 ，或访问互联网：
www.ibm.com/software/cn
www.ibm.com/developerworks/cn

e business software