使用IBM DB2 pureScale 实现透明的应用扩展

2009年11月

简介

在经济复苏的过程中，对核心业务数据的即时访问始终是企业赖以生存，乃至获得成功的关键因素。随着越来越多的美元流入国内市场，企业需要借助具备高可用性和灵活的架构来提高业务的敏捷性，以便能够抓住新的发展机会。

大多数分布式软件公司在营销时都将可用性水平与“类大型机”或“5-9”可用性这样的术语联系在一起。这些短语都在试图传达业界公认的高可用性“黄金”标准(即DB2® for z/OS®)所设定的持续可用性目标。

可用性	每年宕机时间
99.999%	5分钟
99.99%	50分钟
99.9%	8小时20分钟
99%	3天11小时18分钟
95%	18天16小时
90%	34天17小时17分钟
85%	54天18小时

如今，可用性并不仅仅意味着能够从容应对组件故障并恢复正常的事务处理。如果您的服务水平协议(SLA)指定预期的查询响应时间应在数秒之内，而服务器却花费了1分钟才返回查询，那么这就是可用性方面的问题。要确保可用性，您的系统不仅需要提供事务服务，还需要在SLA指定的期限内提供服务。

举例来说，如果业务周期中的季节性波动造成了扩展方面的可用性问题，则真正具备可用性的架构必须能透明地增加资源，同时避免更改应用，以满足不断变化的性能需求。透明性是一个关键因素：在提高产能时，不应该让应用具备集群感知性(应用知道哪些数据在哪个节点上，以避免节点之间的争用)。企业无法投入足够的资金来开发这些复杂的应用，因此无法实现合理的扩展。这是为什么呢？首先，显而易见的是，集群感知的应用需要适应数据量和分布状态的不断变化。集群感知的应用并不仅仅要求代码随着集群的发展而改变：这些代码还需要经历测试、质量保证(Q/A)、部署和认证等过程。这可能造成企业花费数周时间来进行协调，并且不可避免地会耗尽基础设施中本应该有更好用途的资源。

用于在分布式平台(非大型机)上扩展数据库事务的其他产品通常采用过时的架构，因此会为扩展带来不必要的困难(比如说增加开销)，从而无法确保符合SLA协议。

IBM DB2 pureScale技术(以下称为DB2 pureScale)可以将**高可用性与真正的透明应用扩展**结合在一个系统中，以便满足您当前和未来对持续可用性的需求。IBM® Power™ Systems服务器和IBM存储解决方案的整合是DB2 pureScale架构交付这种高价值解决方案的内在基础。

到目前为止，“类大型机”仍然是一个引人注目的市场营销词汇。DB2 pureScale标志着真正透明的扩展架构首次应用于分布式平台。本文将介绍DB2 pureScale技术的基本概念、背景信息，以及它在高可用性和透明应用扩展方面具备独特优势的奥秘。

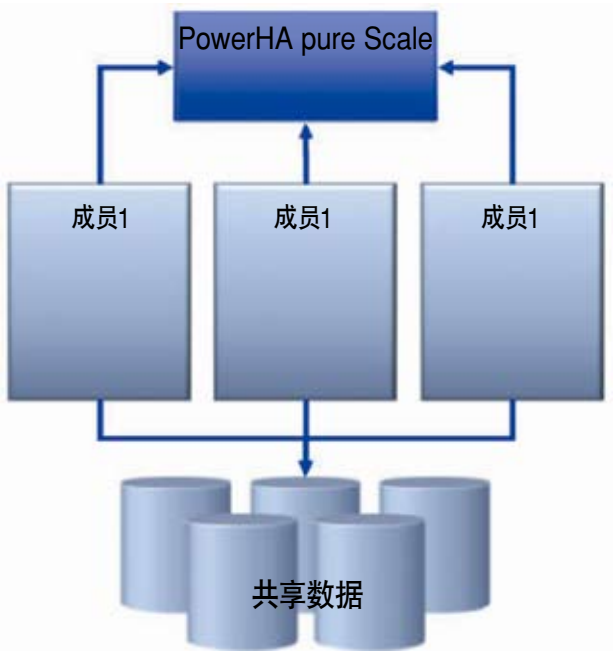
DB2 pureScale基本信息

DB2 pureScale是一种新的DB2 V9.8可选特性，它允许您通过“双机(active-active)”配置将数据库扩展到一组服务器上，以便交付高水平的可用性和可伸缩性。在这种配置中，运行于各主机(或服务器)上的DB2副本可以同时读取和写入相同的数据。

共享DB2数据的一台或多台DB2服务器被称作*数据共享组*。数据共享组中的DB2服务器是该组的成员。目前，数据共享组支持的最大成员数量是128。

除了DB2成员外，PowerHA pureScale™组件还提供了整合的锁管理以及针对数据分页的全局缓存(称作分组缓冲池)。

数据共享组中的各成员可以通过一个非常有效的InfiniBand™网络直接与PowerHA pureScale组件交互，如下图所示。这意味着各成员与集中化的锁和缓存设备之间建立了点到点(P2P)连接。



DB2 pureScale的起源

您所听到或看到的任何关于大型机可用性的描述均指的是DB2 for z/OS设定的高可用性黄金标准。事实上，世界上还没有任何一款数据库解决方案能在可用性方面与运行DB2 for z/OS的System z®服务器相提并论。

DB2 for z/OS数据共享所采用的底层技术确保服务器持续满足SLA [I1]需求，因为Coupling Facility提供了集中化的锁管理和全局缓存，这为快速从故障中恢复提供了保障。事实上，DB2 for z/OS从严格意义上讲已经实现了“5-9s”级的可用性，同时在无缝线性扩展工作负载方面享有很高的声望。

说起DB2 for z/OS，很多人都会想到广泛的可伸缩性和极高的可用性。这种市场声誉并非空穴来风，而是源于这些系统在数据库工作负载可用性方面的市场领先地位始终无人憾动。或许，最能佐证DB2 for z/OS强大功能的莫过于Oracle创始人兼CEO Larry Ellison的评论[1]:

我取笑过其他许多数据库，但唯独对大型机版本的DB2抱有尊重之心。它是当之无愧的一流技术。

DB2 for z/OS究竟有何独特之处，让Ellison对它如此赞赏有加？DB2 for z/OS在数据共享领域中的“独门秘笈”对其用户来说再熟悉不过了，那就是众所周知的Coupling Facility。Coupling Facility不仅为DB2 for z/OS赋予了线性扩展的能力，还提供了一个集中化设备来管理锁。除此之外，它还充当脏页(dirty page)的全局共享缓冲池(有助于可伸缩性和可恢复性操作)。

Coupling Facility技术为DB2 Z/OS贴上了可用性和可伸缩性方面的“黄金”标准的标签，DB2 PureScale技术秉承了DB2 for z/OS Coupling Facility的血脉。这是如何做到的呢？DB2 pureScale提供了一个IBM powerHA pureScale组件，该组件提供了同样集中化的锁管理和真正的全局共享缓冲池架构。

其他供应商实现了采用共享磁盘架构的数据库，其中最有影响力的是Oracle Real Application Clusters (Oracle RAC)。但是，当时在开发和设计Oracle RAC时，分布式平台技术还不允许有效地访问*集中*共享缓存。结果，Oracle RAC的设计最终成为了一次模拟DB2 for z/OS的一次尝试；这也是Oracle RAC的*分布式*锁管理技术和分布式缓存架构的起源。Oracle RAC在引入横向扩展的共享磁盘架构之后失去了DB2 for z/OS解决方案的简洁性优势。另一方面，DB2 for z/OS和DB2 pureScale提供了相同的*集中化*资源管理，因此解决了这些复杂的可伸缩性和可用性问题。本文将在稍后讨论这方面的内容。

起初市场上只有一种架构交付了真正透明的应用可伸缩性和高可用性。随着现代硬件在分布式平台上实现了互连，以及基于InfiniBand的无中断Remote Direct Memory Access (RDMA)的深入发展，DB2 for z/OS所采用的集中锁和缓冲缓存算法已经不再是它所独享的专利。DB2 pureScale将这项久经行业考验的技术引入到了分布式平台中，而这也代表了整个IBM家族的进步。

DB2 pureScale实现透明的应用可伸缩性

在横向扩展的数据库环境中节省成本的关键是实现真正透明的应用扩展机制。透明的扩展意味着数据库引擎可以为OLTP应用提供更大的吞吐量和更快的响应速度，而对数据*本地性*没有要求。

数据的本地性表示应用所需的数据位于它所连接的服务器上，并且节点之间很少会争用相同的数据分页。在横向扩展架构中，如果采用基于网络的消息架构共享集群中的数据，数据的本地性就显得格外重要。

依靠数据本地性实现有效扩展的横向扩展架构要求开发人员创建复杂的事务应用来实现*集群感知性*。集群感知的应用在开发和部署方面不仅更加复杂，而且成本更加高昂，同时当集群发生更改时还要求重新设计应用。一些供应商可能声称它们的架构能运行任何应用，而不需要修改；但是，如果在设计时没有实现某种形式的集群感知性，它们将不能*扩展*任何应用。

透明的应用扩展意味着应用不需要具备集群感知性便可利用横向扩展架构。DB2 pureScale是分布式平台上所特有的，其高效性源于对现代网络和硬件架构，以及pureScale的集中化锁和缓存机制的利用。

为了减少集群中各节点之间的通信，以便实现锁管理和全局缓存服务，DB2 pureScale使用powerHA pureScale集群加速设备(以下简称为CF)和RDMA技术来提供透明的应用可伸缩性。

RDMA允许集群中的各个成员直接访问CF中的内存，而CF也可以直接访问各成员的内存。举例来说，假定集群中的某成员(成员1)希望读取未存储在本地缓冲池中的数据分页。DB2会分配一个代理(或线程)来执行此事务；然后，代理使用RDMA直接向CF的内存写入数据，声称自己需要读取某个特定分页。如果成员1希望读取的分页位于CF的全局集中缓冲池中，则CF会将该分页直接推送到成员1的内存中，而不是让该成员的代理执行I/O操作从磁盘读取它。通过使用RDMA，成员1的代理只需向远程服务器发起一个memcopy

[1] http://www.eweek.com/c/a/Database/In-Larrys-Own-Words/2/

(内存复制)调用，从而避免了成本较高的进程间通信、处理器中断、IP栈调用等。简单来说，pureScale允许成员的代理通过执行看似本地的内存复制操作来执行远程内存复制操作。

这些轻量级的远程内存调用，连同集中缓冲池和锁管理设备，意味着应用不需要连接到已经包含数据的成员。集群中的任何成员都可以有效地从全局缓冲池接收数据分页，无论集群有多大。大多数RDMA调用都非常迅速，这使得发起调用的DB2进程在等待CF的响应时不需要让出已分配的CPU时间，并且不需要重新调度便可完成任务。举例来说，为了向CF通知某行即将更新(因此需要一个X锁)，某个成员的代理需要执行Set Lock State (SLS)请求，也就是将锁信息直接写入到CF上的内存中。CF会确认集群中的其他成员没有锁定这个行，并直接修改请求成员的内存以批准锁请求。

这个SLS只需15微秒就可以完成整个过程，因此代理不需要让出已分配的CPU时间。代理可以持续高效运行，而不需要像其他横向扩展架构那样等待IP中断(避免不必要的上下文切换)。对于长时间运行的批量事务等特定操作来说，DB2代理有必要让出CPU时间，而DB2会自动决定是否动态让出CPU时间。

DB2 pureScale内置的针对集群成员的负载均衡机制是另一个重要的DB2可伸缩性特性。应用不需要具备集群感知性便可利用负载均衡机制。DB2 for z/OS数据共享客户如今所使用的客户端驱动程序可以为DB2 pureScale提供集群负载均衡特性。

DB2 pureScale实现可用性

横向扩展架构的作用并不仅为了处理能力的增加。采用这种架构设计的系统在遇到组件故障时可以继续处理事务，从而能够交付更高的可用性。

与分布式平台上的其他产品相比，DB2 pureScale将可用性提升到了一个新的高度。DB2 pureScale允许访问所有不需要恢复的数据分页，并且随时可以洞察哪些分页需要恢复，而不需要执行任何I/O操作。这是通过集中化CF的独特功能实现的另一项重要创新。

每当成员将一个页读取到它的缓冲池中时，CF都会感知到这一事件并持续对其进行跟踪。任何时候当成员希望更新一页中的行，CF同样能够知晓相关事件。当一个应用执行事务时，成员会将脏页直接写入到CF的内存中。此流程允许集群中希望读取这些经过更改的页的任何其他成员直接从CF获取更新。更加重要的是，从恢复的角度来说，如果任何成员出现故障，CF中会保留该失败成员正在处理更新的页列表，同时还有一些页已经完成更新和提交，但尚未写入磁盘。

任何关系数据库管理系统(RDBMS)的恢复流程首先都需要重新执行任何已提交的事务，以确保磁盘上这些事务的页是最新的(此流程称作redo恢复)。此外，任何数据库服务器都需要撤销任何未完成的事务，即在故障之前对磁盘数据执行了更改但尚未提交(此流程称作undo恢复)。

在共享磁盘集群中，非常关键的一点是要确保集群中的其他节点没有读取或更新尚未恢复的磁盘中的任何分页(恢复这些分页之后才可以对这些行执行新的事务)。这正是CF的闪光之处：由于CF知道哪些页正处于故障节点的更新过程之中，并且CF已经将该节点提交的脏页保存在它的集中缓冲池中，因此DB2 pureScale在确定哪些分页需要恢复时不必阻塞其他成员持续处理事务。其他架构则需要了解哪些节点占用的处理时间较多，以便根据锁信息的分布来确定哪些节点必须恢复。

从较高的层面来看，可以很容易地解释DB2 pureScale环境中的这种恢复进程。每个成员都有空闲的进程，但它们都随时准备着处理故障事件。如果某个成员出现故障时，其中一个恢复进程便会激活；既然这些进程已经存在，因此操作系统不必浪费宝贵的系统时间来创建进程，为它分配内存等。此恢复进程会立即将CF中的脏页预取到它自己的本地缓冲池中。大部分恢复过程都不需要I/O操作，因为需要恢复的页已经在CF的集中缓冲池中了。此外，页预取机制使用轻量级的RDMA在CF与恢复成员之间实现迅速有效的传输。在这段时间内，所有其他成员上的所有其他应用将继续处理请求。如果它们需要从不需要恢复的任何页获取任何数据，那么它们可以继续执行自己事务。因此，它们可以继续从磁盘读取页，因为CF已经知道磁盘上的哪些页是干净的，以及哪些页需要恢复。然后，恢复进程读取故障成员的日志文件，以便于重放必要的事务来重做或撤销故障成员所做的更新。

对于典型的事务工作负载来说，从成员出现故障到故障节点未更新完的分页可供其他事务使用的时间间隔通常在20秒以内。注意，这同时还包括故障检测时间，而某些供应商在提到恢复时间时都排除了这一时间。数据库中的所有其他分页无时不刻(甚至在成员出现故障之后)都是完全可用的。

此外，系统中像PowerHA pureScale集群加速器这样的组件是冗余的。DB2 pureScale支持双重CF功能，这样锁和共享缓存信息就可以存储在两个相互独立的位置，以应对主CF出现故障的情况。

结束语

通过利用现代化的硬件架构，DB2 pureScale可以将之前仅在DB2 for z/OS上可用的集中锁和缓存功能引入到分布式平台中。对硬件和网络的利用提高了并发性水平并显著降低了开销，从而提供了更高水平的可伸缩性。此外，集中锁和分页缓存允许DB2 pureScale持续感知在成员遇到故障时需要恢复哪些分页。因此，在遇到故障时，所有不需要恢复的数据仍然能供其他应用使用，而故障节点正在更新的分页将更加迅速地被恢复。

对于需要高可用性以及通过横向发展实现成本收益的应用来说，DB2 pureScale提供了一个可以满足这些需求的解决方案，并且已经过了市场的考验。

使用DB2 pureScale实现透明的应用扩展–技术比较，2009年10月

8200 Warden Avenue
Markham, ON
L6G 1C7
加拿大

在中国印刷

2009年11月