文章编号：1005-8451(2009)10-0051-04

Oracle 10g RAC技术研究与分析

翟油华[1]，胡玉俊[2]
(1.南京医科大学第二附属医院 信息科，南京 210011；
2.上海铁路局 信息技术所南京电算站，南京 200037)

摘 要：以企业信息化面临的难题为背景，分析网格计算相对于传统计算的优势，提出利用网格技术解决信息化过程中面临的难题；研究Oracle 10g核心组件RAC的技术特点，对这些特点极其原理进行逐一分析。

关键词：网格计算；真正应用集群；共享磁盘；高速缓存融合；透明应用切换
中图分类号：TP39 文献标识码：A

Research and analysis of Oracle 10g RAC technology

ZHAI You-hua[1],HU Yu-jun[2]
(1. Second Affricated Hospital of Nanjing Medical University, Nanjing 210011, China;
2.Nanjing Computing Station, Information Technology Center of Shanghai Railway Administration, Nanjing 200037, China)

Abstract: Based on the requirements of enterprise IT, it was started with the comparison between enterprise grid computing and the traditional computing, elaborated the technological features of RAC, which was one of the great components of Oracle 10g.

Key words: grid computing ; real application cluster; shared disk; database cache fusion; transparent application failover

如何降低架设和使用信息技术基础架构所需的高昂成本，是IT用户最关心的问题。要降低IT成本，必须解决过剩的计算容量、昂贵的容量扩展以及高额的管理成本3大难题。受到传统企业计算的限制，用户只能针对高峰容量来构建计算容量，但又无法在平时有效地使用多余的容量，也无法在必要时以较低成本迅速地向模块单元增加容量，这些因素都是造成IT成本居高不下的原因。一种基于网格计算原理的企业网格计算正是企业所需要的，它很好地解决了企业IT面临的难题。

收稿日期：2009-03-11
作者简介：翟油华，工程师；胡玉俊，工程师。

1 网格计算

网格计算协调使用计算机集群来创建单个逻

了原有计算机联锁执行部分的继电器电路，以智能子执行单元，通过现场总线，计算机联锁系统对现场道岔、信号机等信号设备进行操纵，形成分布式控制。实现了控制、监督、监测一体化，为铁路信号的信息化发展提供有力的支持。

随着客运专线和高速铁路的大规模建设，随着我国铁路运行速度和运输效率的提高，对计算机联锁的安全性和控制性提出了更高的要求，这就要求计算机联锁系统及其它控制系统之间要高速度、高效率、高准确地传递信息，指导列车安全有序地运行。而以上这些通讯方式的运用为铁路快速发展提供了技术支持。

4 结束语

选择什么样的通讯方式是由计算机联锁通讯的需要决定的，其核心的目的都是为铁路运输服务。计算机联锁系统的通讯方式随着现代通讯手段的发展而发展，总的来说，它向着更高速、更安全、更可靠的方向发展，这样才能更好地提高运输效率，保证运输安全。

参考文献：

[1] 赵志熙. 车站信号控制系统[M]. 北京：中国铁道出版社，1997.
[2] 谢希仁. 计算机网络[M]. (5版) 北京：电子工业出版社，2008.

辑实体（如Database）。通过跨多台服务器分配工作，从而实现了可用性、可伸缩性的优点。由于单个逻辑实体跨多台服务器实施，因此可以在线增加或删除容量，实现更高的硬件利用率和更好的业务响应性。用户不需购买昂贵的高性能的主机，相反可以选择多台价廉且性能较低的服务器，并将它们群集起来；假如其中一台服务器发生故障，其它服务器仍可以让系统继续正常运行，提高了系统的可靠性。由于硬件技术发展迅速，未来用户可以用同样低的价格购买到性能比以前更好的服务器加入集群中，来提高系统的运算能力，而不必花巨资更换主机。这些优势解决了企业IT面临的难题。网格计算的创新之处主要来自硬件，但网格基础架构的功能必须在软件中得到体现，如果没有软件功能支持，则与目前的一些典型集群系统相似。企业网格的数据是真正共享的，实现了系统的可伸缩性，充分利用计算资源。

Oracle 10g是一个专门为企业网格计算开发的基础架构软件，真正应用集群（Real Application Cluster，RAC）是Oracle 10g的组件，也是其网格技术实现的核心。它具有高速缓存合并、共享磁盘、透明应用切换3大核心功能。

2 RAC体系结构

RAC的体系结构主要由：节点（Nodes），私有网格（Interconnect），共享磁盘（Shared Disk）3个主要部分组成，如图1。

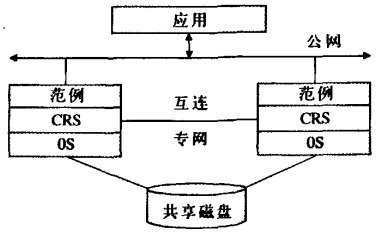

图1 RAC体系结构

节点之间通过私有网络连接来进行数据交换，并分别与共享磁盘存储进行连接。节点与应用层处在同一外部网络中，虽然每个节点有不同的物理IP地址，但应用客户仍可以在一个虚拟数据库服务名的水平上进行连接，而且客户端对于多服务器的多个地址可以不用关心，同时系统自动实现负载均衡（Load Balance）。

负载均衡能够自动适应快速变化的业务需求和随之而来的工作负荷的改变，通过动态地重新分配数据库资源，从而可以在节点之间用最小化的磁盘I/O和低的延迟通信来优化利用集群系统资源。

3 高速缓存融合

高速缓存融合（Database Cache Fusion）消除了多台服务器争用数据时产生的碰撞现象，极大地提高了RAC的可扩展性，使集群系统可以支持更多的节点，数据库应用不需要做任何复杂的修改或特殊设计就可以良好地运行在集群系统上，并且充分发挥多节点的处理性能。该技术把RAC数据库中的所有数据库缓存作为一个共享的数据库缓存，并被所有节点共享。RAC系统中每一个节点都运行一个数据库实例。每个数据库实例包含一组Oracle 进程和用于缓存的系统全局区（SGA）。用于多个Oracle实例的共享高速缓存技术不但提供了很高的性能，而且实现了群集系统的连续可用性。该技术改变了全局区域内部配置，把高速缓存的数据缓冲区从一个本地存储区转移到一个可被所有实例访问的共享高速缓存区域。高速缓存融合能够使集群中所有节点的磁盘共享对所有数据的访问，同步高速缓存，从而最大限度地降低磁盘I/O，优化数据读写。当然节点之间会产生不小的网络通信和CPU的开销。因此，双节点RAC的性能不会是单节点性能的2倍。

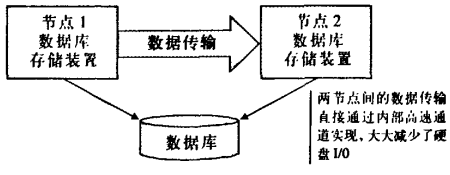

图2 缓存融合示意图

4 共享磁盘

RAC采用共享磁盘方式实现数据库群集，它的数据库文件、联机重做日志和数据库的控制文

件都能为集群中的每个节点所访问。同时，RAC允许多个实例同时访问同一数据库，因此一个实例的故障不会导致数据库无法访问。这种基于共享磁盘体系结构的特性，实现了按需增加和收缩集群服务器的特点，并实现了容错、负载均衡和性能效益等特性。群集环境中所有物理服务器共享且并发地对磁盘上的单个数据库进行更新，同时系统还额外地需要其它同步与串行机制，避免2个或多个服务器同时更新同一数据页上的记录。RAC处理数据同步模拟分析：

（1）假设Node1需要从Shard Disk读一个数据块B1，它向GCS发送锁请求，当Node1收到GCS的锁后，Node1便可以从Shared Disk读取B1。Node1读取并修改了B1里的数据行；

（2）此时Node2也需要访问B1，但该B1已经在Node1的缓存中，所以Node2不会再从Shared Disk读取B1。Node2向Node1的GCS发出锁请求；GCS要求Node1把B1给Node2，Node1直接通过Interconnect将B1新副本发送给Node2，Node2收到后通知GCS；此时Node2就可以读写B1并再次修改了B1里的数据行；

（3）当Node1需要再读取B1时，Node2直接通过Interconnect把该B1最新的副本传回给Node1。

5 透明应用切换

5.1 透明应用切换过程

透明应用切换(Transparent Application Failover，TAF)是RAC并行高可用性的体现，当一个节点发生故障时，连接在该节点上的终端用户会被自动重新连接到其它正常的数据库节点上，无需手工连接，应用端的应用及查询仍会继续执行，用户的注册信息得到保留，后续客户端的连接也会被指到正常节点。

应用端A1,A2在节点Node1中连接，A3,A4连向节点Node2；当节点Node1发生故障，A1,A2的事务将被回滚，但是它们可以继续工作而不必手工重新连接，因为A1,A2被TAF移植到节点Node2上。请看下面的模拟分析：

（1）RAC节点间都有心跳机制，用来监测其它节点是否正常运行；

（2）假定用户通过节点Node1准备执行一条命令并从数据库中返回1 000行记录；

（3）起初的500行记录被节点Node1执行并返回到该用户终端界面；

（4）当用户正在查看起初的这500行记录时，节点Node1发生故障；

（5）系统监测并确认Node1发生故障，并从集群中除去该故障点；

（6）此时用户并没有意识到故障的发生，并在自己的窗口中继续查看剩下的500行记录；

（7）RAC通过根据连接配置文件自动重新连接到节点Node2上；

（8）Oracle在节点Node2上重新执行该命令并返回剩下的500行记录给用户；如果记录在缓冲，将会被瞬间返回；否则，Oracle将重新执行一次I/O操作。

RAC群集中的一个节点发生了故障，故障节点上所有运行的事务会丢失，Oracle将故障节点所拥有数据块的控制权限重新转交给正常节点。此过程称为全局缓存服务重置。在全局缓存服务重置发生时，RAC中所有服务器都会被冻结，所有应用程序将被挂起，GCS将不会响应群集中任何节点发出的请求；重置后，Oracle读取日志记录，确定并锁定需要恢复的页面，并执行回滚，此时数据库恢复可用。

5.2 TAF的配置

TAF通过Oracle客户端TNSNAMES.ORA文件进行配置，几个主要参数是：TYPE、METHOD、BACKUP、RETRIES、DELAY。

配置方法如下示例：

```
RAC1 =
(DESCRIPTION =
(ADDRESS = (PROTOCOL = TCP)(HOST = NODE1_VIP)(PORT = 1521))
(ADDRESS = (PROTOCOL = TCP)(HOST = NODE2_VIP)(PORT = 1521))
(LOAD_BALANCE = no)
(CONNECT_DATA =
(SERVER = DEDICATED)
(SERVICE_NAME = RAC1)
(FAILOVER_MODE =
(TYPE = SELECT)
```

(METHOD = BASIC)
(RETRIES = 180)
(DELAY = 5))))

目前，Oracle RAC 支持对话期间的故障切换和Select操作故障切换，在进行中的Select请求在故障切换现场继续被处理，一个正在进行的业务处理在故障期间必须被返回，提供一个回叫信号，以便使应用程序能够继续执行。从Oracle 10g开始提供的VIP，可以实现IP的自动跳转来实现节点的失败切换，但不管采用哪种方式，对于失败的节点，Oracle必须先恢复失败节点中的事务，以便整个数据库处于一致状态，这个过程大致需要1 min～5 min，具体取决于应用的压力大小与复杂程度。

由于RAC采用高速缓存合并技术，节点间的数据传输量巨大，因此需要使用千兆网络并通过光纤互联来防止性能瓶颈。

5.3 TAF技术支持的主要应用接口

(1) OCI、OCCI；
(2) Java JDBC driver；
(3) ODBC Connection；
(4) SQL*Plus Connection。

6 负载均衡

Oracle10g RAC负载均衡有客户端负载均衡和服务器端负载均衡。其实现方法如下：

(1) Oracle10g RAC负载均衡通过Oracle Net Service实现；

(2) 通过各个实例的PMON进程向Oracle Net Service动态注册各个节点的负载；

(3) 客户端负载均衡

a.通过Oracle客户端TNSNAMES.ORA文件中的LOAD_BALANCE参数进行配置；

b.LOAD_BALANCE参数取值为ON/OFF；

c.客户端负载均衡将随机选择可用连接，类似于丢硬币；

(4) 服务器端负载均衡

a.需要在初始化参数中添加LOCAL_LISTENER、REMOTE_LISTENER；

b.服务器端负载均衡将根据RAC系统中各节点系统负载分配连接；

c.当采用共享服务器模式时，RAC也会考虑各个dispatcher的负载，然后分配连接。

7 RAC的优化

各个节点与instance需要很频繁的修改同一个数据块，因为还会有全局数据块状态修改，联机日志记录等过程，这些过程可能会导致大量的cache fusion与gc current/cr request。Gc buffer busy等待事件，会导致RAC的响应速度还不如单instance的响应速度快。

所以RAC优化最基本的原则是：避免大规模的不同节点频繁的修改同一个数据块，避免大量的cache Fusion与全局等待事件的出现。在RAC的优化中，有一个规则是可以适用的，就是让不同的业务单独运行在独立的节点上。虽然9i以后的RAC可以不用发生pin操作，但是，大量的全局等待事件还是可能对性能造成比较大的影响。

8 结束语

如果使用RAC，用户不必花巨资购买大型主机来满足高可靠性要求，也不必担心单节点系统故障对企业造成难以估计的损失。当系统需要进一步扩展时，可按需增加节点，无需对应用程序进行任何修改，也无需更换新的服务器。对于企业用户，可以选择多台刀片式服务器来组成集群环境，节省了服务器空间的占用，降低了硬件成本（PC服务器硬件价格只有普通小型机价格的1/10）；操作系统可以选择免费、开放、稳定的Linux系统，由于Oracle 10g是在Linux平台下开发测试的，因此它对Linux系统的支持是非常好的。企业网格计算的实现，解决了企业信息化过程中面临的难题，降低了企业信息化成本。这是企业网格计算带来的显著优点，也是未来信息技术发展的方向。

参考文献：

[1] Oracle Real Application Clusters 10g[EB/OL]. http://www.oracle.com/lang/cn/technologies/grid/index.html, 2006-01.

[2] 陈吉平. 构建高可用环境[M]. 北京：电子工业出版社，2008，1.

[3] Richard J.Niemiec. Oracle Database 10g性能调整与优化[M]. 薛 莹. 北京：清华大学出版社，2009，1.