第 33 卷 第 7 期 计 算 机 工 程 2007 年 4 月
Vol.33 No.7 Computer Engineering April 2007
·软件技术与数据库· 文章编号：1000—3428(2007)07—0053—03 文献标识码：A 中图分类号：TP311.132.3

# Oracle 10g RAC 核心技术研究与分析

周晓丹，冯少荣，薛永生
(厦门大学信息科学与技术学院，厦门 361005)

**摘 要：**以企业IT面临的难题为背景，分析了企业网格计算相对于传统计算的优势，提出了利用网格技术解决企业难题的观点；研究了Oracle 10g核心组件RAC的技术特点，对这些特点的原理进行逐一分析；架设RAC系统实例，并对其进行功能验证、性能测试以及结果分析。
**关键词：**网格计算技术；真正应用集群；共享磁盘；高速缓存合并；透明应用切换

# Research of Oracle 10g RAC Core Technology and Analysis

ZHOU Xiaodan, FENG Shaorong, XUE Yongsheng
(School of Information Science and Technology, Xiamen University, Xiamen 361005)

**【Abstract】** Based on the requirements of enterprise IT department, this paper starts with the comparison between enterprise grid computing and the traditional computing, elaborates the technological features of RAC, which is one of the great components of Oracle 10g. Further research is taken and a RAC system is set up, including functions verification, performance test and result analysis.
**【Key words】** Grid computing technology; 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，几乎是所有企业 IT 用户最关心的问题。然而，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三大难题。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。网格计算是利用网络技术，把分散在不同地理位置的计算机组成一台虚拟超级计算机。每一台参与计算的计算机就是其中的一个“节点”，所有的计算机组成了一张节点网就叫网格[1]。

## 1 传统计算与网格计算

传统计算与网格计算之间在结构上的区别(见图 1)。

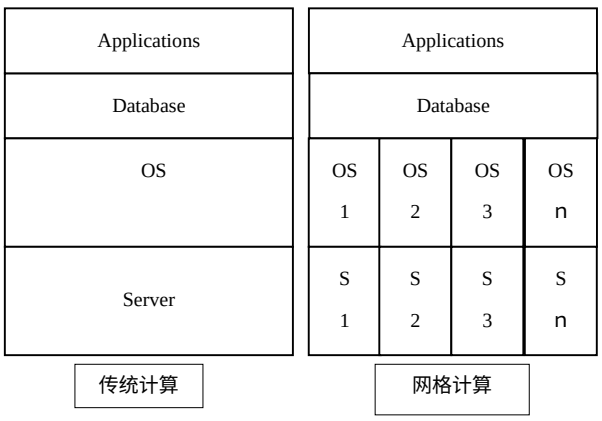


**图 1 结构比较**

网格计算协调使用计算机集群来创建单个逻辑实体(如Database)。通过跨多台服务器分配工作，从而实现了可用性、可伸缩性和性能方面的优点。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；而典型集群系统的数据不能被共享而只能被人为地分区，当增加服务器时，所有的数据都需要重新分区，并将数据分配给新的服务器；需要删除服务器时，又要重新对数据进行分区，不具备灵活性，操作复杂且管理成本高。

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[3]。它具有高速缓存合并、共享磁盘、透明应用切换三大核心功能[4,5]。

## 2 RAC 体系结构

RAC 的硬件体系结构，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 2 所示)。

**作者简介：**周晓丹(1978－)，女，硕士生，主研方向：数据库，数据仓库和数据挖掘；冯少荣，硕士、副教授；薛永生，教授
**收稿日期：**2006-06-17 **E-mail：**art3panda@163.com

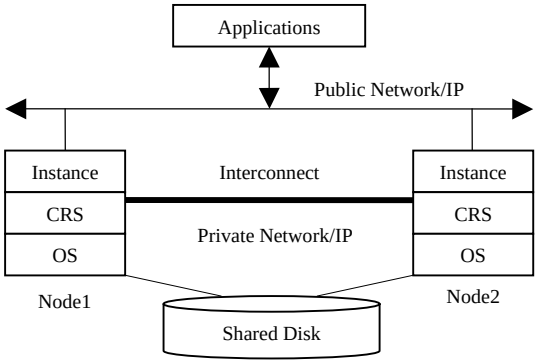


**图 2 RAC 体系结构**

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 高速缓存合并

高速缓存合并(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 集群系统的可扩展性。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。

高速缓存合并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两倍。

## 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，避免两个或多个服务器同时更新同一数据页上的记录。那么，RAC 是如何处理数据同步的？请看下面的模拟分析：

(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

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。

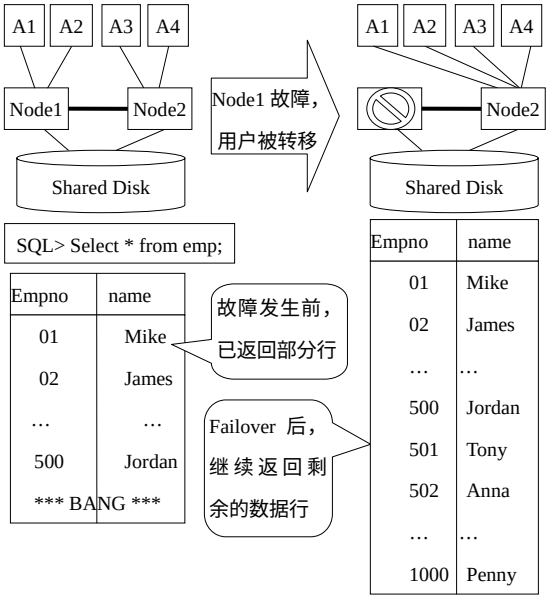


**图 3 TAF 示意图**

如图 3 所示，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)假定一个用户通过节点 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。

## 6 RAC 实例测试

为了更好地理解RAC架构，实例架设了RAC系统[6]，并进行功能与性能测试。架设结构类似于图 2。

### 6.1 环境

Database & Server:

Node1: Dell PE2850, 2CPU, 8GB memory

Node2: Dell PE6650, 4CPU, 4GB memory

Oracle database 10.1.0.2.0 with RAC

Storage: Dell | EMC CX300

Network:

Private interconnect: 100MB Ethernet

Public network: 100MB Ethernet

OS: Linux: Red Flag Server 4

### 6.2 RAC 功能及性能测试

6.2.1 透明应用切换(TAF)功能测试

测试方式：分别通过两种客户端连接方式对数据库进行连接：(1)通过 JDBC 连接 RAC 数据库；(2)通过 Oracle 自带的 SQL PLUS 连接数据库。

在两节点数据库 instance 正常运行情况下，分别使用两种方法使 Node1 节点发生故障(结果是类似的)：

(1)在 Node1 节点数据库上执行“ shutdown immediate ”命令，终止节点 Node1 的数据库 instance；

(2)在 Node1 节点操作系统上执行“ ifdown eth0 ”命令，直接将 Node1 节点 interconnect 端的网卡关闭。

测试结果及分析：

(1)对于当时连接在另一节点(如 Node2)上的客户端没有任何影响；

(2)对于正连接在 Node1 上的客户端分为两种情况：

1)正在执行 JDBC 应用端会报错，重新登录后可继续操作。这是由于 JDBC 采用的是 JDBC Thin 连接方式，这种连接方式不被 TAF 功能支持；

2)正在通过 SQL PLUS 执行数据库操作的用户不受影响，应用继续运行。说明 TAF 功能支持 SQL PLUS 类型的连接方式，将应用转移到另一个正常的节点上并继续运行。

6.2.2 负载均衡测试

测试方式 :通过 Loadrunner 模拟每隔 1s 登录一个用户并运行不同的 SQL 语句；动态跟踪两结点相关的 session 数量变化信息。

测试结果及分析：发现两节点上的 session 数量大致相等，新增加的 session 会自动连接到相对较为空闲的节点上。这说明负载均衡的功能起了作用，应用负载被自动均衡分布到所有的节点上并有效地利用资源。

6.2.3 RAC 性能 (单节点与双节点 RAC 的性能比较)

测试方式：采用 Loadrunner 定义多个用户并发，共使用 12 条 SQL 语句，分成 12 个组，每个组定义 1 个用户，共 12 个用户，每隔 1s 增加一组并发运行。分别测试以下两种环境的性能：(1)对 Dell PE2850 进行单节点性能测试(两台服务器的整体性能较为接近，所以

任选其一)；(2)对 RAC 系统进行双节点性能测试。

测试结果及分析(见图 4)：

运行所有 SQL 语句，单节点测试一共所需时间：1 分钟 21 秒。

运行所有 SQL 语句，双节点 RAC 测试一共所需时间：1 分钟 06 秒。

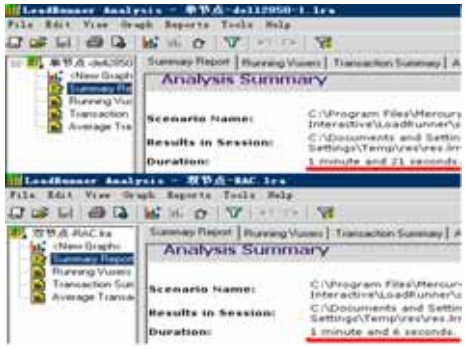


**图 4 单双节点性能测试结果截图**

双节点 RAC 的性能比单节点性能提高了 22.7%左右。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。由于实验条件限制，实例配置的私有网络为百兆，这无疑对 RAC 性能测试带来一定影响，如果提高网络带宽，相信 RAC 性能会有更大的提升。

## 7 总结

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 IT 面临的三大难题，降低了企业 IT 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 Ian F, Carl K. 网格计算[M]. 2 版. 金 海, 袁平鹏, 石 柯, 译. 北京: 电子工业出版社, 2004.

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

3 Murali V. Oracle Real Application Clusters [M]. USA: Digital Press, 2003.

4 Robert F. Oracle Database 10g New Features[M]. USA: Osborne Oracle Press Series, 2004.

5 Mike A, Madhu T. Oracle 10g Grid & Real Application Clusters[M]. USA: Rampant Tech. Press, 2004.

6 Kevin L. Oracle Database 10g: the Complete Reference[M]. USA: Osborne Oracle Press Series, 2004.