

深入理解 Oracle RAC 12c

Expert Oracle RAC 12c

深入学习管理Oracle RAC的
专业知识

[美] Syed Jaffar Hussain
Tariq Farooq
Riyaj Shamsudeen
Kai Yu

著

ACOUG
All China Oracle User Group
中国 Oracle 用户组

赵燊
梁程
李真旭

译

张乐奕 组织审校



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Apress®

深入理解 Oracle RAC 12c

Expert Oracle RAC 12c

[美] **Syed Jaffar Hussain**
Tariq Farooq
Riyaj Shamsudeen
Kai Yu

著

ACOUG
Authorized Contributor
Oracle University

赵梁
梁程
程飞
飞真
真旭

译 张乐奕 组织审校

電子工業出版社

Publishing House of Electronics Industry

北京•BEIJING

内 容 简 介

本书介绍了Oracle RAC 12c技术的方方面面,涵盖了与RAC技术相关的集群件知识、数据库知识、存储知识、网络知识,并在基于RAC的应用软件设计、优化方面给出了大量的有价值的建议。

特别值得阅读的是,本书紧跟Oracle数据库新版本的发行,使用专门的章节描述了比如RAC One Node这样11g中的新特性,还有Flex集群这样12c中的新特性,是我们深刻了解RAC的基本知识,并紧跟技术发展潮流的优秀书籍。

本书适合有一定Oracle数据库经验的数据库管理员和开发者阅读。

Expert Oracle RAC 12c

By Syed Jaffar Hussain, Tariq Farooq, Riyaj Shamsudeen, Kai Yu, ISBN:978-14302-5044-9

Original English language edition published by Apress Media.

Copyright © 2013 by Apress Media

Simplified Chinese-language edition copyright © 2014 by Publishing House of Electronics Industry

All rights reserved.

本书中文简体版专有出版权由Apress Media, Inc.授予电子工业出版社,未经许可,不得以任何方式复制或抄袭本书的任何部分。

版权贸易合同登记号 图字: 01-2013-9177

图书在版编目(CIP)数据

深入理解 Oracle RAC 12c / (美) 赛义德 (Syed) 等著; 赵焱等译. —北京: 电子工业出版社, 2014.10

书名原文: Expert Oracle RAC 12c

ISBN 978-7-121-24066-9

I. ①深… II. ①赛… ②赵… III. ①关系数据库系统 IV. ①TP311.138

中国版本图书馆 CIP 数据核字 (2014) 第 187050 号

策划编辑: 张春雨

责任编辑: 刘 舫

印 刷: 北京丰源印刷厂

装 订: 三河市鹏程印业有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编: 100036

开 本: 787×980 1/16

印张: 30.5

字数: 683 千字

版 次: 2014 年 10 月第 1 版

印 次: 2014 年 10 月第 1 次印刷

定 价: 99.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 zlts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线: (010) 88258888。



译者序

很高兴《深入理解 Oracle RAC 12c》(英文版书名为 *Expert Oracle RAC 12c*) 这本书跟大家见面了, 这是我在英文技术书籍翻译领域做出贡献的第二本书, 之前的一本是 *Oracle Expert Exadata*, 那时是译者之一, 而这次则勉强算是组织者。

并无参与太多翻译工作, 于是说说翻译这本书的由来, 聊以作序。

本书的作者之一 Kai Yu, 相识已久, 华人一枚, 在美国 Dell 总部工作。其主要工作内容为研究 Dell 硬件环境中的 Oracle 数据库解决方案, 在 Oracle Virtual Machine 和 Oracle RAC 方面尤为擅长, 他还是 Oracle ACE Director。每年在旧金山的 Oracle Open World 上遇到, 总是颇感亲切, 毕竟能用中文交流的机会不多。Kai Yu 在 Facebook 上颇为活跃, 实际上大量的国外 Oracle 技术专家在 Facebook 上都彼此熟络, 交流甚多。

说回到书来, 在 2013 年的早些时候, 获悉 Kai Yu 正在跟另外几位 Oracle ACE Director 合作写一本关于 Oracle RAC 的书籍, 其中有 Riyaj Shamsudeen, 他的 orainternals 网站 (orainternals.wordpress.com) 我已经订阅多年, 还将他的 *RAC Object Remastering (Dynamic Remastering)* 这篇文章翻译成中文。这篇 DRM 的文章是我看过写得最全面、最细致的一篇文章。虽然我没有见到此书, 但这样豪华的作者团队也让我认定此书是一本高品质的技术书籍, 于是在美国见到 Kai Yu 的时候, 就直接敲定等他们出版了书籍我来负责组织中国的 Oracle 技术专家进行翻译。

岁月荏苒, 时光如梭, 很快就到了 2013 年 7 月, Kai Yu 来信说书很快要出版了, 正在最后审校。于是我在 ACOUG (中国 Oracle 用户组) 的核心成员邮件组中发了一封召集译者的邮件, 短短两天的时间集齐了 4 位译者, 他们是李真旭 (@oracledatabase12c)、赵焱 (@netbanker)、程飞 (@xifenfei_惜分飞)、梁涛 (@--梁涛--), Yong Huang (@yong321, 他是 ITPUB 的 Oracle 专题深入讨论版中最称职的版主) 也担任了翻译审校工作, 整个译者团队同样豪华。

2013 年 8 月该书在美国出版, 9 月 3 日电子工业出版社计算机图书出版分社的张春雨



目 录

第1章 Oracle RAC 概述.....	1
高可用性和可扩展性	2
什么是高可用性.....	2
数据库的可扩展性	3
Oracle RAC	5
数据库集群体系架构.....	5
RAC 架构.....	6
RAC 的硬件要求.....	8
RAC 的组件	10
Oracle RAC 的缓存融合	13
RAC 的后台进程.....	16
获得 Oracle RAC 的好处	19
高可用性和意外停机.....	19
高可用性和计划停机时间	23
使用 Oracle RAC One Node 实现高可用性.....	25
RAC 的可扩展性.....	25
使用 Oracle RAC 整合数据库服务.....	28
部署 RAC 时的注意事项.....	30
拥有成本.....	30
高可用性的注意事项.....	31
可扩展性的注意事项.....	32
是否选择 RAC	33
本章小结	34

第 2 章 Oracle 集群件堆栈的管理和故障诊断 35

 Oracle 12cR1 的集群件及其组件 36

 Oracle 集群件的存储组件 36

 集群件软件堆栈 38

 集群件启动顺序 40

 ASM 和集群件谁先启动 42

 集群件管理 43

 集群件的管理工具和实用程序 43

 启动和停止集群件 45

 管理 Oracle 集群件 45

 管理 OCR 和表决磁盘 48

 管理 CRS 资源 50

 添加和删除集群节点 50

 常见集群件启动问题的解决方法 53

 诊断、调试、跟踪集群件和 RAC 问题 58

 调试集群件的组件和资源 58

 网格架构中各组件的目录结构 61

 Oracle 集群件故障诊断工具 64

 CHM 69

 本章小结 77

第 3 章 Oracle RAC 运行实践 79

 工作负载管理 79

 服务 80

 服务指标 82

 负载均衡目标 83

 运行时的故障切换 86

 第二个网络中的服务 86

 服务的使用指导 86

 SCAN 和 SCAN 监听 87

 第二个网络中的 SCAN 监听 (12c) 91

 SCAN 监听使用指南 92

 全局数据库服务 (12c) 93

 RAC 中的故障切换 94

透明应用程序故障切换 (TAF)	95
快速连接故障切换 (FCF)	96
WebLogic Active GridLink	97
事务卫士 (12c)	97
应用程序的连续性 (12c)	98
策略管理的数据库	99
临时表空间	100
大量数据的修改	101
性能指标收集	102
参数文件管理	102
密码文件管理	103
管理数据库和实例	104
管理 VIP 和监听	106
其他主题	107
进程优先级	107
内存不足	108
SGA 的大小	109
文件系统缓存	110
本章小结	110
第 4 章 RAC 12c 的新特性	111
Oracle Flex 集群	112
Oracle Flex 集群的架构	112
Flex 集群的扩展性和可用性	114
配置 Flex 集群	115
Flex ASM 架构	120
Oracle Flex ASM 架构	120
Flex ASM 和 Flex 集群	122
配置 Flex ASM	122
ASM 客户端和重定位	124
新的 ASM 存储限制	125
在磁盘组中更换 ASM 磁盘	125
清理 ASM 磁盘组和文件	125
在 ASM 磁盘组中均匀地读取数据	126
衡量和优化 ASM 重新平衡操作	126

系统命令的假设分析和评估	126
Oracle RAC 中的可插拔数据库	128
可插拔数据库的体系结构概述	128
Oracle RAC 中的 PDB 数据库	132
12cR1: RAC 中的其他新功能	136
RAC 中的公共网络: 添加对 IPv6 的支持	136
全球数据服务	136
在线修改资源的属性	136
12cR1 RAC: 基于策略的数据库管理	136
ASM 磁盘组: 共享的 ASM 密码文件	137
节点的有效性检查: 限制服务的注册	137
12cR1: 共享的 GNS 服务	137
RAC 12cR1: 限制服务注册	137
Oracle ASM、ACFS 和 ADVM: 功能的改进以及新特性	137
NFS 的高可用性	138
12cR1: CHM 的增强	138
Windows: 支持 Oracle 安装用户	138
OUI 的增强和改进	138
12cR1: 安装和升级——自动运行脚本	139
12cR1: 应用的连续性	139
事务的幂等性和 Java 事务卫士	139
已废弃和不再支持的功能	139
本章小结	140
第 5 章 存储和自动存储管理	141
Oracle RAC 中的存储架构和配置	143
Oracle RAC 中的存储架构和 I/O	143
磁盘冗余阵列配置	146
存储协议	148
多路径设备配置	151
设置设备的所有权	153
自动存储管理	155
ASM 实例	156
ASM 存储结构	164
用 SQL 命令和 V\$ASM 视图管理 ASM	173

在 ASM 上存放 Oracle 集群注册表和表决磁盘	173
在安装网络架构时选择 ASM 存放 Oracle 集群注册表和表决磁盘	173
将 Oracle 集群注册表和表决磁盘迁移到新的 ASM 磁盘组	176
ASM 集群系统文件	179
建立 ACFS	181
用 ASMCA 为 Oracle RAC 创建 ACFS 的主目录	183
本章小结	185
第 6 章 应用设计上的问题	186
局部性插入操作	186
大量的 TRUNCATE 或 DROP 命令	189
序列缓存	191
空闲块链表和自动段表空间管理	193
过多的提交	194
长时间没有提交的事务	195
本地访问	196
小表的更新	197
索引设计	198
低效的执行计划	199
过多的平行扫描	199
全表扫描	199
应用之间的关联性	200
管道	201
应用改变的实施	201
本章小结	202
第 7 章 管理和调优一个复杂的 RAC 环境	203
比较共享和非共享的 Oracle 主目录的优点和缺点	204
服务器池	205
服务器池的类型	206
系统定义的服务器池	206
用户定义的服务器池	206
创建和管理服务器池	207
计划和设计 RAC 数据库	209
策略管理数据库	210

实例锁定	213
小规模和大规模的集群环境设定	214
裂脑案例和如何避免	215
理解、解决和防止节点驱逐	217
节点驱逐——梗概和综述	217
延伸距离（伸展）集群——摘要、概况和最佳实践	221
延伸距离（伸展）集群：创建和配置最佳实践	222
创建和配置	223
Oracle 图形界面	223
Oracle 企业管理器云控制 12c	225
RAC 的安装和设置——在不同操作系统：Linux、Solaris 和 Windows 中的考虑和 窍门	227
RAC 数据库性能调优：一个迅速简单的途径	228
性能调优的 3 个 A 工具	229
本章小结	234
第 8 章 RAC 的备份与恢复	235
RMAN 概要	235
介质管理层	237
联机备份和恢复的预备知识	238
非 RAC 数据库和 RAC 数据库的对比	239
重做日志和归档日志的共享存储位置	240
快照控制文件配置	241
为 RAC 配置多通道	242
RAC 中的并行机制	245
RAC 中的实例恢复和崩溃恢复	245
真实世界中的例子	250
使用 12c 的 OEM 云控制器来管理 RMAN	254
OCR 恢复	259
本章小结	261
第 9 章 网络实践	262
网络类型	262
网络层	263
协议	265

VIP.....	269
子网划分	270
集群内联	271
巨帧	274
负载均衡和故障转移	279
内核参数	282
网络测试工具	283
GC Lost Block 问题	288
配置 Oracle RAC 和集群件网络环境	290
建立 IP 和域名地址的解析	293
网格构架安装过程中的网络设置	297
集群件的网络配置	300
网络故障转移	306
第二网络配置	307
本章小结	308
第 10 章 优化 RAC 数据库	309
缓存融合介绍	309
缓存融合的处理	310
GRD	312
BL 资源和锁	313
性能分析	317
接收端的分析	318
RAC 等待事件	325
GC Current Block 2-Way/3-Way	325
GC CR Block 2-Way/3-Way	327
GC CR Grant 2-Way/GC Current Grant 2-Way	329
GC CR Block Busy/GC Current Block Busy	329
GC CR Block Congested/GC Current Block Congested	329
占位等待事件	329
发送端分析	330
曾用块的类型（被使用的块的类型）	333
GCS Log Flush Sync	334
保护 LMS 进程	335
GC Buffer Busy Acquire/Release	335

唯一索引	338
表块	339
DRM	341
DRM 进程概述	342
DRM 的阶段	344
GRD 冻结	345
参数	345
在 12c 中的改变	346
DRM 和 Undo	346
DRM 的故障诊断	347
AWR 报告和 ADDM	347
ASH 报告	348
本章小结	348
第 11 章 锁和死锁	350
资源和锁	350
SGA 的内存分配	352
资源类型	354
锁模式	356
锁相关的视图	357
可插拔数据库 (12c)	357
锁争用的故障排除方法	358
入队争用	360
TX 入队争用 (Enqueue Contention)	361
TM 入队争用	364
HW 入队争用	366
DFS Lock Handle	366
SV 资源	368
CI 资源	371
DFS lock handle 总结	373
Library Cache Locks/Pins	373
诊断 Library Cache Lock 争用	376
队列统计信息	377
v\$wait_chains	378
Hanganalyze	379

死锁	380
LMD 跟踪文件的分析	381
本章小结	385
第 12 章 RAC 中的并行查询	386
概述	386
RAC 中的并行执行	390
PX 服务进程的位置	391
测量 PX 通信	395
并行执行与缓存融合	397
PEMS	398
并行特性与 RAC	398
诊断并行执行问题	411
在 RAC 中创建索引	413
RAC 中的并行 DML	414
12c 中的并发联合处理	415
Partition-Wise Join	416
本章小结	417
第 13 章 集群件和数据库升级	419
配置	419
升级之前的检查	421
开始 Oracle 集群件升级	423
rootupgrade.sh 脚本的重要性	430
升级后的工作	433
集群件降级	434
数据库升级	437
手动升级数据库	438
数据库升级后的步骤	440
使用 DBUA 升级数据库	440
DBUA 的优势	443
数据库降级	443
本章小结	444

第 14 章 RAC One Node	445
RAC One Node 概述	445
升级到 11.2.0.2 或更高版本	446
配置 RAC One Node 环境	447
配置 RAC One Node 数据库	449
先决条件	449
开始 DBCA 创建过程	450
指定 RAC One Node 初始化参数	452
管理 RAC One Node 数据库	453
核实配置信息	453
验证在线迁移状态	454
停止和启动数据库	454
完成数据库在线迁移	455
处理计划外的节点和集群重启	457
RAC One Node 和标准 RAC 之间的转换	458
扩展为标准 RAC	458
降级到 RAC One Node	459
通过 12c 中的云控制管理 RAC One Node	460
通过 12c 中的云控制进行数据库迁移	460
第三方故障转移技术和 RAC One Node 的对比	463
本章小结	464

第 1 章



Oracle RAC 概述

作者：Kai Yu

在现今的商业世界中，随着互联网重要性的日益增加，越来越多的 IT 应用系统需要向客户提供不间断的服务。其中一个典型的例子就是网上商城。许多公司都希望自己的网上商城能够一年 365 天 7×24 小时运行，这样来自世界各地的客户，在不同的时区，都可以随时浏览产品的信息并下订单。

对于不直接面对客户的应用来说，系统的高可用性（HA）同样是至关重要的。IT 部门经常会有复杂的分布式应用程序，它们连接到多个数据源，比如在线商店应用中提取出销售数据，然后把这些数据汇总到报表系统。这些应用程序的一个共同特征是，任何意外的停机，都可能给客户带来巨大的损失，而损失的总数有时是难以量化的。Oracle 数据库作为这些应用系统的关键部分，它的可用性影响和决定了整个在线商城系统的可用性。

第二个方面是应用系统的可扩展性。伴随着业务的增长，交易量可能会比初始规划的容量有一倍甚至三倍的增长。此外，业务量可能会在很短的时间大幅动态变化，例如节假日的销量比平时会明显增多。为了适应业务的动态变化，Oracle 数据库需要变得足够灵活和容易扩展，即随着负载增加而扩展，当需求减少的时候数据库则需要减小和收缩。从历史上看，用作数据库服务器的传统大型 UNIX 服务器缺乏灵活性，难以适应这些变化。在过去的十年中，行业标准和技术方向已经逐渐转移到基于 x86-64 架构的 Linux，以满足应用系统日益增长的可扩展性和灵活性的需求。其中 Oracle 真正应用集群（RAC）运行在低成本的 Linux x86-64 服务器上，也逐渐成为广泛适应行业标准的数据库解决方案，能够实现数据库的高可用性和可扩展性。

本章主要介绍 Oracle RAC 技术，并讨论如何使用 Oracle RAC 来实现 Oracle 数据库的高可用性和可扩展性。在本章中将涵盖以下主题：

- 数据库的高可用性和可扩展性
- Oracle 真正应用集群（RAC）

- 获得 Oracle RAC 的好处
- 部署 Oracle RAC 时的注意事项

高可用性和可扩展性

本节讨论了数据库高可用性和可扩展性的要求和相关因素。

什么是高可用性

在前面的网上商城的例子中，为了满足业务的可用性要求，IT 部门需要提供相应的解决方案。数据库作为大多数业务应用系统的核心，它的可用性是保证所有应用系统可用的关键。

在大多数公司，业务部门和 IT 部门之间会定义系统的服务水平协议（SLA），包括整个应用系统的可用性。在协议中，它们的可用性可以用百分比来表示，也可以用每月或每年允许的最大停机时间来表示。例如，当可用性是 99.999% 的时候，意味着每年允许的停机时间要小于 5.26 分钟。有时候，一个 SLA 协议还会指定特殊时间作为停机窗口；例如为了硬件和软件升级，后台办公应用数据库可以在每季度的第一个星期六的凌晨零点至四点之间进行停机维护。

由于大多数的高可用性解决方案需要额外的硬件或软件，这些解决方案的成本可能会比较高。每个公司都应该根据应用系统的特点和预算决定其高可用性的要求。首先类似人力资源管理系统等一些后台办公应用系统可能不需要 7×24 小时运行。对于那些需要高度可用的关键业务系统，像停机 1 小时这样的停机成本是可以计算的。所以先得到停机时间的损失，然后再计算出不同级别中，可用性解决方案所需的采购、实施和运营的花费，并将两者进行比较。这种比较将帮助业务部门和 IT 部门提出更准确的 SLA，真正满足业务部门的需求和预算，IT 团队也能提供相应的服务。

许多应用程序包括了多层系统，并运行在多台计算机上，网络也是分布式的。应用程序的可用性不仅仅取决于支持这些应用系统的基础设施，包括服务器硬件、存储、网络、操作系统等，同时取决于应用的每一层系统，如 Web 服务器、应用服务器、数据库服务器等。在这一章中，我们将主要关注数据库服务器的可用性，这也是数据库管理员的职责。

对于应用程序的可用性来说，数据库的可用性起着至关重要的作用。使用停机时间来表示一个数据库不可用的时间，停机时间可以是意外停机时间或计划停机时间。当发生硬件或软件故障甚至是自然灾害（数据中心全部宕掉），这种系统管理员或数据库管理员没有预计到的意外，从而引起的宕机我们称之为非计划停机（或者意外停机）。大多数非计划停机是可以预期的；例如，当设计一个集群的时候，最好假定一切硬件都可能发生故障，因为大部分集群都是低成本的集群，硬件都存在着一定故障率。关键在于设计系统的可用性时，要确保每个组件都是由足够的冗余组成，甚至包括数据中心级别的冗余。计划停机时

间通常是计划性的维护活动，如系统升级或迁移。

Oracle 数据库服务的非计划停机可能由于数据丢失或服务器故障引起。而数据丢失的原因可能是存储介质失效、数据损坏、误删除数据，甚至是数据中心发生故障。数据丢失是非常严重的问题，有可能变成永久的数据丢失，也可能需要很长时间才能恢复。可以使用预防和恢复两种方法来解决数据丢失的问题。预防方法包括磁盘镜像的 RAID（独立磁盘冗余阵列），如磁盘阵列的 RAID 1（镜像）和 RAID 10（镜像和条带化），或者使用冗余的 Oracle ASM 磁盘组（自动存储管理）。在第 5 章我们将讨论为 Oracle 配置 RAID 和 ASM 的细节。恢复方法则是使用下面的三种办法进行：从之前的数据库备份恢复；闪回恢复（Flashback）；通过 Data Guard 切换到备用数据库。

服务器故障通常是由硬件或软件故障引起的。硬件故障有可能是物理设备故障、网络或存储连接故障；软件故障有可能是操作系统崩溃、Oracle 数据库实例或 ASM 实例失败。通常服务器出现故障时，数据库中的数据能够保持一致。在软件或硬件问题修复以后，故障服务器上的数据库服务可以恢复和启动。通过提供冗余的数据库服务器，当主服务器发生故障时，数据库服务可以切换到备用数据库服务器上，这样就能够提供不间断的数据库服务。通过提供冗余的网络和存储连接，可以避免网络和存储连接的故障。

系统升级或迁移会引起 Oracle 数据库的停机。数据库系统升级包括服务器硬件、网络、存储、操作系统、Oracle 数据库打补丁或升级。不同性质的升级需要的停机时间也不同。为了避免系统升级引起的数据库停机，一个方法是在系统升级期间使用冗余的系统对外提供服务。当数据库服务器需要使用新的服务器、新的存储，或者一个新的操作系统的时候我们就需要进行数据库迁移。虽然这种迁移并非频繁，但需要的停机时间可能更长，对业务有更大的影响。可以使用一些工具和方法来减少数据库迁移的停机时间，例如 Oracle 可传输表空间、Data Guard、Oracle GoldenGate 及 Quest SharePlex 等。

在这一章里，我们将关注 Oracle 数据库高可用性的特定区域：服务器的可用性，将讨论如何使用 Oracle RAC 减少因服务器故障和升级引起的数据库的停机时间。对于使用其他解决方案来最小化 Oracle 数据库的意外和计划停机时间，可以参考 Oracle 最高可用性体系结构（MAA）。想了解 MAA 的最新内容，请访问 Oracle MAA 架构的网页：www.oracle.com/technetwork/database/features/availability/maa-090890.html。

数据库的可扩展性

在数据库的调优理论里，有一种理论说更应该关注应用程序的数据库设计，SQL 的查询调优，然后才是数据库的实例调优，而不是仅仅增加新的硬件。确实如此，因为如果数据库应用程序设计得很差，或者有糟糕的 SQL 查询，添加额外的硬件不能从根本上解决性能问题。然而另一方面，即使是一些优化很好的数据库，当业务增加的时候也可能会超出系统规划的容量。