基于 Oracle RAC 的数据库架构分析与企业应用①

闫 黎

(中国空空导弹研究院, 洛阳 471009)

摘 要: 阐述了 Oracle RAC 的基本概念、分类和结构体系和基本原理. 结合业务需求提出了基于 Oracle RAC 的解决方案. 实施了基于 Oracle RAC 数据库的共享存储、文件存放和网络设置等规划与部署. 实现了无用户干预的自动故障切换, 极大提升了数据库的可靠性. 对企业数据库的应用构建具有一定指导意义.

关键词: Oracle RAC; 数据库; 集群; 高可用; 负载均衡

Analysis and Construction of the Database Architecture based on Oracle RAC

YAN Li

(Luoyang Optoelectro Technology Development Center, Luoyang 471009, China)

Abstract: Describes the Oracle RAC's basic concepts, classifications and structure system and the basic principles. Combined with the demand for business solutions based on Oracle RAC. Implementation of the basic Oracle RAC database shared storage, file storage, and network settings, planning and deployment. Achieved automatic failover without user intervention, which greatly enhance the availability of the database. Have a certain significance to Enterprise database application and Construction.

Key words: Oracle RAC; cluster; database; high availability; load balance

随着信息技术的不断发展, 信息系统服务的可持续性、系统的高可用性成为衡量一个企业服务能力的重要标准, 作为后台支撑的数据库系统其作用更不容忽视. 如何提供高可用性、高性能以及灵活的数据库应用系统, 成为各数据库厂商、集成厂商致力解决的问题, Oracle RAC 就是其中一种优秀的解决方案.

1 Oracle RAC介绍

Oracle集群, 也称Oracle RAC, 全称Real Application Cluster, 译为“真正应用集群”, 其主要思想是在需要的时候插入 Oracle 节点以支持更多的工作载荷, 该体系架构实现了多个节点(实例)同时访问同一数据库, 如果其中某个节点发生故障, RAC 能够通过后台的监控进程将连接自动切换到另外一个或多个节点上, 从而实现应用的无缝切换.

1.1 集群的分类

目前主流的集群技术主要分为三类:

① 高性能计算集群: 将计算任务分解后分配到不同计算节点来提高整体计算能力, 因而主要应用在科学计算领域.

② 负载均衡集群: 核心是把负载流量尽可能平均合理地分摊到集群各个节点, 每个节点都可以处理一部分负载, 并且可以根据节点负载进行动态平衡.

③ 高可用性集群: 侧重于提高系统的可用性, 如果集群中的某个一节点发生了故障, 那么将另外的节点代替它. 只要一个节点正常运行, 就能够对外提供服务.

Oracle RAC 兼具了 2 和 3 的特点.

1.2 Oracle RAC 体系结构

随着 Oracle10g 及 Oracle11g 的推出, Oracle 开发了自己的集群组件(Cluster Ware), 称为集群就绪服务组件(Cluser Ready Services, CRS). CRS 由 Oracle 提供, 能够将集群中的各节点通过共享存储的硬件支持集合在一起. 支持的平台包括 AIX、Linux、Windows 等主流平台. 如图 1 为体系结构示意图.

① 收稿时间:2013-04-25;收到修改稿时间:2013-05-27

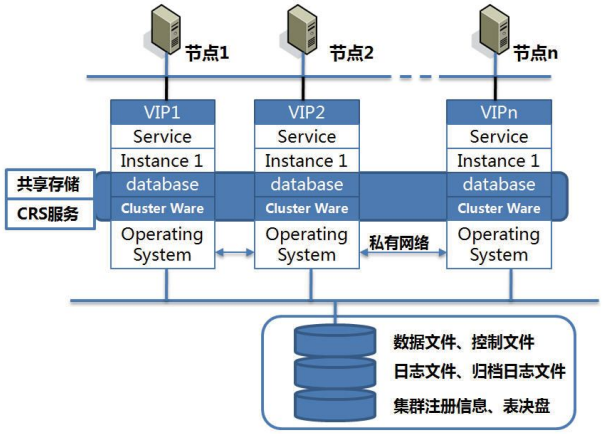

图 1 Oracle RAC 体系结构

在通过网络连接的各个节点上，运行着不同的数据库实例，每个实例都可以操作放置在共享存储上的集群数据库，这通过 Oracle 的集群就绪服务实现. 集群之间通过内部网络连接进行数据交换，称为缓存融合(Cache Fustion)，在保证效率的前提下，同步了各个节点的数据.

缓存融合本质上通过私有互联网络在集群内各节点的 SGA 间进行数据块的传递，而不是把数据先写到共享存储，再在另一节点重新读入缓存中的策略. 当一个数据块被读入 RAC 集群中的某个节点时，该数据块会获得一个锁资源，以避免集群中的其他实例在使用该块时发生冲突. 另一节点如果要请求该块，而该块已经存储在某一实例的缓存中，则该块直接通过私有网络传递过去，如果这个块的内容已经改变但未提交，那么将传递一个一致性副本.

因此，数据块只需要通过私有网络在各实例的 SGA 缓存间互动，从而有效降低所有节点的磁盘的 IO，这也是为什么集群数据库在硬件上要求高速私有网络的原因，至少不能让网络 IO 低于磁盘 IO 速度.

2 针对Oracle RAC的操作系统配置

2.1 共享存储配置

在 RAC 模式下，Oracle 要求所有的控件文件、联机重做日志、数据文件都存储在共享存储上，被集群中所有的节点访问. 数据库的其它相关文件(如归档重做日志、参数文件、数据库闪回日志)则根据其存储类型被存储在共享存储或者本地磁盘上，无特定的要求.

根据特定的平台，一般存在以下 4 种基本的集群存储实现方式:

① 裸设备方式: 通过共享卷组方式，在磁盘阵列上创建并发卷组及裸设备，供 CRS、Oracle 使用. 这种方式在 Oracle11gR2 版本中已经不被支持;

② 自动存储管理方式: 通过 Oracle 的 ASM 技术，将磁盘阵列设置为 ASM 磁盘组;

③ 集群文件系统方式: 例如通过 AIX 并行文件系统 GPFS 实现创建并发的文件系统;

④ 网络文件系统方式 例如 NFS 方式，在磁盘阵列创建 NFS 文件系统实现.

在 Linux 和 Windows 平台中，Oracle 提供了一种专用于 RAC 环境的集群文件系统 OCFS/OCFS2，这需要从 Oracle 网站下载安装.

2.2 网络配置

为了配置 RAC 环境，每个节点必须至少有两块网上，一块网卡连接公共网络用于对外服务，另一块网卡用于专用内部通信.

一般 RAC 节点间的内部通信要使用两个以上的专用交换机，即使一个 RAC 节点失败或者关机，另一个节点上的网卡仍然可以和交换机保持活动状态. 双交换机保证了互通私有网络的冗余. 另外这种方法也提供了超出两节点集群结构的可能性.

3 需求分析

笔者所在单位某信息系统为 C/S 架构，由一台 IBM X3650M3 提供数据库服务，日常在线人数维持在 200 人以上. 此种构架下存在如下问题:

① 系统可用性的问题: 只要服务器出现问题，在停机期间，所有客户端无法登陆，严重影响了正常业务的开展.

② 数据安全性的问题: 硬件故障导致数据丢失的风险极大，虽然采用虚拟磁带库的方式进行备份，但是一旦出现硬件问题，系统的恢复和重建工作也将耗费宝贵的正常业务时间.

③ 系统性能上的问题: 由于在线人数较大，有时做一条业务查询需要 10 多分钟，他们希望改进系统的响应时间，提高工作效率.

4 解决方案

4.1 架构设计

根据该系统存在的问题，结合实际条件，设计如

图 2 所示的系统架构.

硬件组成:

主机环境: 两台 IBM X3650 M3,作为运行 RAC 的主机.

存储环境: EMC CX480 存储, 两台光纤交换机, 搭建 SAN 存储网络, 主机配置两块光纤卡, 分别连接到两个光纤交换机, 保证冗余及故障切换.

私有网络环境: 每个主机分配 2 个网卡连接私有网络, 采用网络交换机或者专用内联交换机(成对出现), 保障私有网络的可靠性与失败切换.

服务网络环境: 提供对外服务的通道. 通常情况下, 主机都连接在一个或者是两个网络交换机上, 负责与外界通信. 网卡的失败冗余保护可以通过网卡绑定实现自动接管, 证外部网络的可靠性与故障切换.

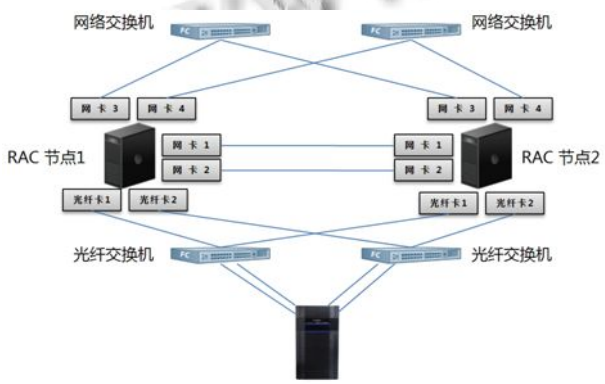

图 2 改造后的系统架构

4.2 操作系统和数据库选型

出于稳定性、安全性的考虑, 选择开源操作系统:RedHat EnterPrise Linux Server 5.5.

Oracle 版本: 10.2.0.1

Clusterware 版本: 10.2.0.1

共享磁盘使用 ocfs2+ASM 的方式来管理, orc 和 votingdisk 使用 ocfs2, 数据文件使用 ASM.

4.3 主机及 IP 规划

RAC 环境由两个节点组成, 每个节点有两块网卡, 分别需要 3 个 IP 地址 public、Private、VIP, Private 和 VIP 分配到 Public NIC 上;每个 IP 对应一个网络名. 具体规划如表 1 所示.

4.4 Oracle RAC 布署过程

由于篇幅限制, 在此不做详细描述, 具体可见参考文献或者网上相关资料.

表 1 主机及 IP 具体规划

	节点 1	节点 2
主机名	Rac1	Rac2
网卡：public	Eth0	Eth0
网卡：private	Eth1	Eth1
IP/网络名：public	192.168.10.1/rac1	192.168.10.2/rac2
IP/网络名：vip	192.168.10.11/rac1-vip	192.168.10.12/rac2- vip
IP/网络名：private	10.0.0.1/rac1-priv	10.0.0.2/rac2-priv

5 高可用性测试

Oracle RAC 故障切换技术基础是 Failover, 是指集群中任何一个节点的故障都不会影响对外提供服务, 连接到故障节点的用户会被自动转移到正常节点, 对用户来说感觉不到这种切换, 这个功能在 Oracle 中被称为 Failover(故障转移). 本系统采用的是 Oracle 推荐的 Server side TAF 的方式. 下面是测试过程:

① 从一个客户端连接, 并且窗口一直保持打开:

Sqlplus system/admin@racl-vip/oacl;

② 执行查询:

SQI>select instance_name from v$instance;

INSTANCE_NAME

Rac1

可以看到当前连接到实例 1.

③ 在节点 1 上, 强制结束掉这个会话:

SQL>select pid, spid from v$process Where addr in(select paddr from v$session

where username=‘system’):

PID SPID

36 24247

该连接对应的 PID 是 24247, 在操作系统中 kill 掉这个进程.

④ 在会话中执行语句:

SQI>select instance_name from v$instance;

ERROR at 1ine l:

ORA-03113:end of file on communication channel;

可见此时会话已经断开.

⑤ 稍等一会再执行查询:

SQI>select instance_name from v$instance;

INSTANCE_NAME

Rac2

可以看到, 用户连接自动切换到了节点 2, 这样解决了在某个节点故障后, 保持服务连续的需求.

除此之外, 还可以利用Oracle RAC环境下的Load Balance 特性, 解决负载问题. oracle10g 提供了两种手段来实现分散负载:其一是通过 Connection Balancing, 在客户端按照某种算法把用户分配到不同的节点;其二是通过 service, 在应用层进行分散.

6 小结

数据库的高可用性、负载能力与许多方面是相关的, 如存储、网络和主机等等, Oracle RAC 是在本地环境进行的数据库冗余措施, 对于地震、火灾等不可抗力导致的机房的损毁, 将无法得到可用性保障. 因此需要其他的容灾方案进行保护, 如 Oracle DataGuard、存储级别的 Mirror 镜像等等, 可以实现快速切换与灾难恢复. 这是每个企业数据中心所必须考虑和面对的.

参考文献

1 文平.Oracle 大型数据库在 AIX/UNIX 上的实战详解.北京:电子工业出版社.2012:446–449.

2 陈吉平.Oracle 高可用环境.北京:电子工业出版社.2009:35–36.

3 张晓明.大话 Oracle RAC 集群高可用性备份与恢复.北京:人民邮电出版社.2009:71–73.

4 何明.Oracle DBA 培训教程—从实践中学习 Oracle 数据库管理与维护.北京:清华大学出版社.2009:98–99.

5 红帽软件(北京)有限公司.RedHat Enterprise Linux 系统管理.北京:电子工业出版社.2009:112–117.

(上接第 189 页)

解异常产生的原因, 以确定是否存在逻辑错误, 而对于隐蔽性错误则需根据现象进行分析. 将异常时的函数执行时间与该函数正常执行时间相比较, 当出现执行时间明显增加的情况时, 则死锁与该函数有关, 并进一步将与该函数有关的函数、锁信息进行上报到分析软件中进行处理, 通过分析软件进行可视化调试有助于快速定位问题. 对于资源竞争造成的死锁, 还可以通过自旋其他核实现核数量上的变化辅助定位.

4 结束语

本文结合 OCTEON 的多核硬件平台, 分析该多核处理器的软硬件工作方式, 阐述异常产生的原因, 提出了一种调试方法, 在不使用 JTAG 调试器以降低系统硬件复杂度的基础上, 具有占用内存资源小等优点, 缩短系统开发时间的同时, 方便了在工程现场的调试, 具有很高的实用价值.

参考文献

1 曾令将,王继红,舒红霞.并行嵌入式系统可视化性能分析工具的设计与实现.计算机与数字工程,2012,40(3):130–132.

2 刘磊,黄河,唐志敏.支持多核并行程序确定性重放的高效访存冲突记录方法.计算机研究与发展,2012,49(1):64–75.

3 杨旭,刘江,钱诚,苏孟豪,吴瑞阳,陈云霁,胡伟武.一种面向多核处理器的通用可调试性架构.计算机辅助设计与图形学学报,2011,23(10):1656–1644.

4 杨启军,鲁士文.基于多核的入侵防御系统的设计与实现.计算机工程与设计,2010,31(21):4595–4598.

5 林贻珀.可视化并行性能调试环境的设计与实现[硕士学位论文].北京:清华大学,2009.