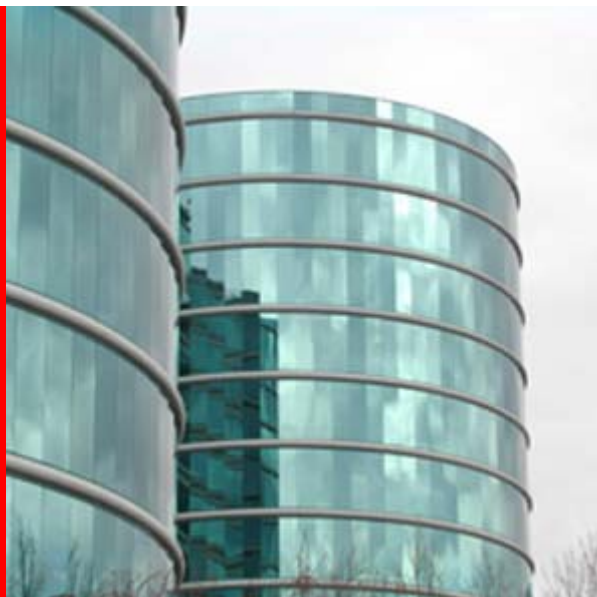


甲骨文



ORACLE®



RAC和ASM最佳实践

Kirk McGowan

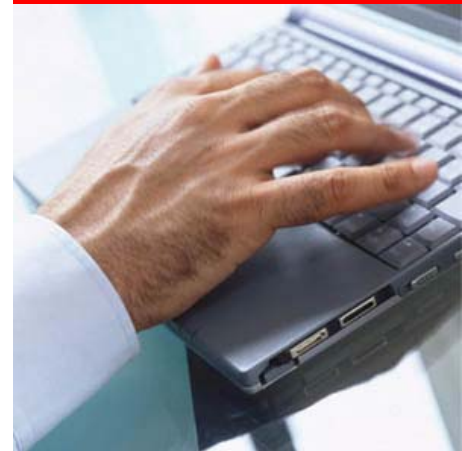
Technical Director – RAC Pack

Server Technologies Development

以下内容旨在概述我们产品的总体发展方向。这些信息仅供参考，不可纳入任何合同。它们不承诺提供任何资料、代码或功能，并且不应该作为制定购买决策的依据。所描述的有关 **Oracle** 产品的任何特性或功能的开发、发布和时间规划只能由 **Oracle** 决定。

议程

- 规划最佳实践
- 实施最佳实践
- 运行维护最佳实践



规划



1. 定义服务水平的目标

- 没有目标就无所谓成功
- 现实的和可度量的目标
 - 由业务目标驱动、与业务目标相关联
 - 现有的服务水平通常作为基线
 - 高可用性服务水平
 - 最大的计划内/计划外停机时间 **(RTO)**
 - 业务恢复点目标**RPO**
 - 性能/处理能力服务水平
 - 用户数、响应时间、某时间段完成的工作量/交易量等
 - **Speed-up (响应时间) vs scale-up (吞吐量)**
 - 整合, 降低成本

2. 设置现实的期望值

- 如果你的应用在SMP上能够线性扩展，则在RAC上也可以获得良好的扩展性，而不需要修改任何应用代码。
- RAC可以避免数据库实例、节点本身的单点故障，并能够保障在发生这些故障情况下数据库的完整性。

3. 无需听信传言

- **RAC**在多节点上扩展性不好
 - 超过**30**家可考察的客户运行**6**节点以上的**RAC**集群数据库
 - **Amazon, J2 Global – 16**节点
 - **Mercado Libre – 10**节点
 - **Overstock – 9**节点
 - **Gas Natural, Thomson – 8**节点
 - 大量的**4**节点以上的用户
- **RAC**不适合**DW**数据仓库应用
 - **TPC-H**基准测试 – **1, 3, 10TB**级别: **RAC**集群结果始终保持前**2**名
 - 数据量超过**1TB**的**DW**用户, 有超过**20**家可证明的用户使用**RAC**
 - 另外还有**20**家混合型业务系统使用**RAC**
 - 其中**8**家数据量 > **10TB**

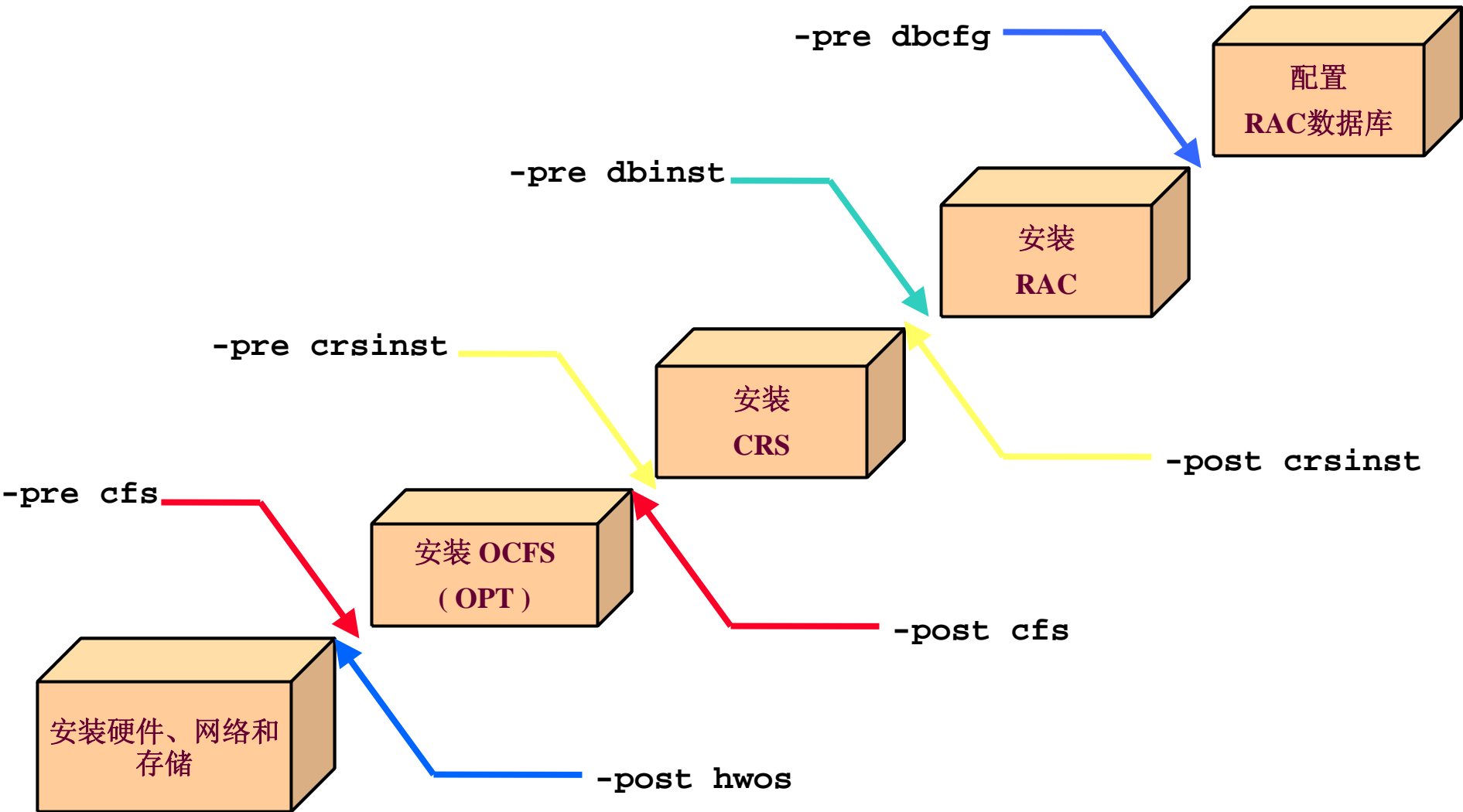
4. 尽量简单化

- 确认并使用被证实可行的配置方案
 - 不必太超前和太创造性
- 利用**Metalink**的**Certification matrix**进行认证
- 尽量减少供应商之间的集成点
 - 集群管理, 集群卷管理, **CFS**, 系统供应商, **OS**供应商, **HBA/NIC**, 存储, **MPIO**软件, 数据库, 应用
 - **B&R**工具, 配置管理工具, 系统监控工具, 灾难备份功能。
 - 可更换的部件越多, 集成的工作量越大越复杂
- 与供应商紧密合作
 - 共同的参与者, 不仅仅是一个响应团队

实施



5. 使用集群验证工具Cluster Verification Utility



CVU – 组件功能列表

- `$> ./cluvfy comp -list`
- 有效的组件:
 - **nodereach** : 检查节点间的可访问性
 - **nodecon** : 检查节点的连通性
 - **cfs** : 检查**CFS**的完整一致性
 - **ssa** : 检查共享存储的可访问性
 - **space** : 检查可用空间
 - **sys** : 检查最低的系统需求
 - **clu** : 检查集群的完整一致性
 - **clumgr** : 检查集群管理的完整一致性
 - **ocr** : 检查**OCR**的完整性
 - **crs** : 检查**CRS**的完整性
 - **nodeapp** : 检查存在的节点应用
 - **admprv** : 检查管理权限
 - ...

在哪里可以找到CVU ?

- 随10gR2一起提供
- Oracle DVD
 - clusterware/cluvfy/runcluvfy.sh
 - clusterware/rpm/cvuqdisk-1.0.1-1.rpm
- CRS Home目录下
 - \$ORA_CRS_HOME/bin/cluvfy
 - \$ORA_CRS_HOME/cv/rpm/cvuqdisk-1.0.1-1.rpm
- Oracle Home目录下
 - \$ORACLE_HOME/bin/cluvfy

部署 cluvfy

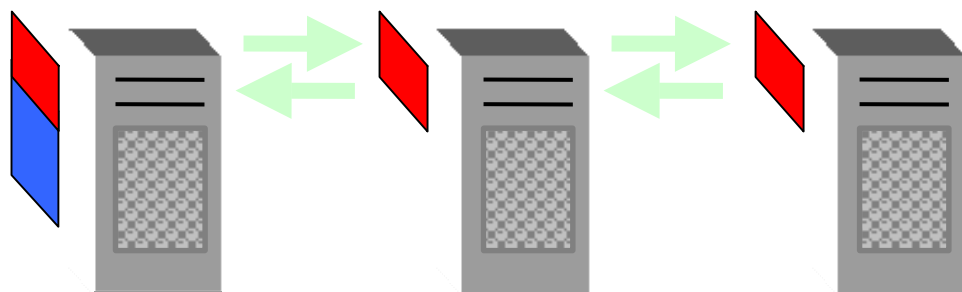
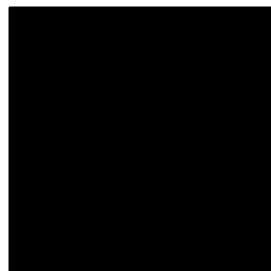
- 只需在本地节点安装。在运行工具时，根据需要、可以自动部署到远程节点上。

- 用户在本地图节点安装

- 执行多节点验证命令

- 工具复制必须的文件到远程节点

- 在所有的节点执行验证工作，并生成报告。



6. 正确地配置集群互联(Interconnect)

- 使用千兆以太网和**UDP**协议
 - 避免私有的低延迟协议
- 在**OS**级绑定多块网卡作为“虚拟”集群互联
 - 故障切换和均衡负载
- 设置最大支持的**UDP**协议的发送/接受缓冲区
 - 与平台相关 – 通常**256K**
- 使用交换机
 - 不支持直连电缆
- 消除传输的问题
 - 错包/丢包可能导致严重的损耗
- 使用集群互连来实现集群和**RAC**的通讯

7. 镜像OCR/Voting disk

- 保存关键集群配置信息的元数据库，用于解决**split brain**问题的机制
- 在**10gR1**中，需要通过硬件**RAID**技术或者**OS LVM**技术实现
- 在**Oracle 10gR2**提供镜像支持
 - >crsctl add css votedisk *path*
 - >ocrconfig -replace ocrmirror *destination_file or disk*
- 建议**3重镜像**
 - **split brain**问题需要通过磁盘少数服从多数的方式、允许来确定继续对外服务的子集群

8. 使用虚拟IP地址（VIP）

- 用来减少客户连接时**TCP/IP**超时所造成的延误。
- 只需要选择公共服务用的网卡(**VIPCA**)
- **VIP**地址必须在**DNS**域名服务器中注册并被识别
 - **VIP**用于 **tnsnames**连接
 - 监听程序监听**VIP**用于客户端连接
- 在多数平台上可以使用**ifconfig**命令来确认是否配置了**VIP**。
 - 你应该看到一个新的**VIP**网卡，例如: **eth0:1**
- 如果把集群迁移到一个新的数据中心（或子网络），需要变更**IP**地址。而**VIP**信息存储在**OCR**文件中，修改和变更**VIP**需要额外的管理工作。
 - 详细信息请参考 **Metalink Note:276434.1**

9. 使用自动存储管理(ASM)

- 专门针对**Oracle**数据文件优化的逻辑卷管理
- 可以替代传统意义上的集群文件系统(**CFS**)
- 通常会创建**2个diskgroups: database area**和**flash recovery area**，最好是物理分开的，不要共享相同的物理**spindles**
- 由多个相同大小的磁盘构成一个**DiskGroup**
- 如果在存储阵列上已经进行了镜像，设置**REDUNDANCY=EXTERNAL**
- 可能的话，**ASM**应使用**pseudo devices (multi-path IO)**
- 使用**OMF (Oracle Managed Files)**
 - **OMF**为所有的数据文件、日志文件、控制文件等生成唯一的文件名
- 使用用户或系统提供的模版，来简化和统一**ASM**文件的创建，例如：

```
Create tablespacetb1  
datafile '+group1/tb1(fine)' size 100M;
```

10. 实施工作负载管理

- 集群工作负载管理 – 不仅仅是均衡负载
 - 有效地利用集群资源
 - 最大限度地降低集群中同步的成本
 - 监控和优化响应时间
 - 在故障切换时减少网络延迟
- 充分利用AWR, Services, FAN and LBA

连接到服务

服务: **GL** 首选: 1,2,3,4

服务: **ERP**

首选: 1,2 可用: 3,4

服务: **CRM**,

首选: 3,4 可用: 1,2

实例:
FINPROD1

实例:
FINPROD2

实例:
FINPROD3

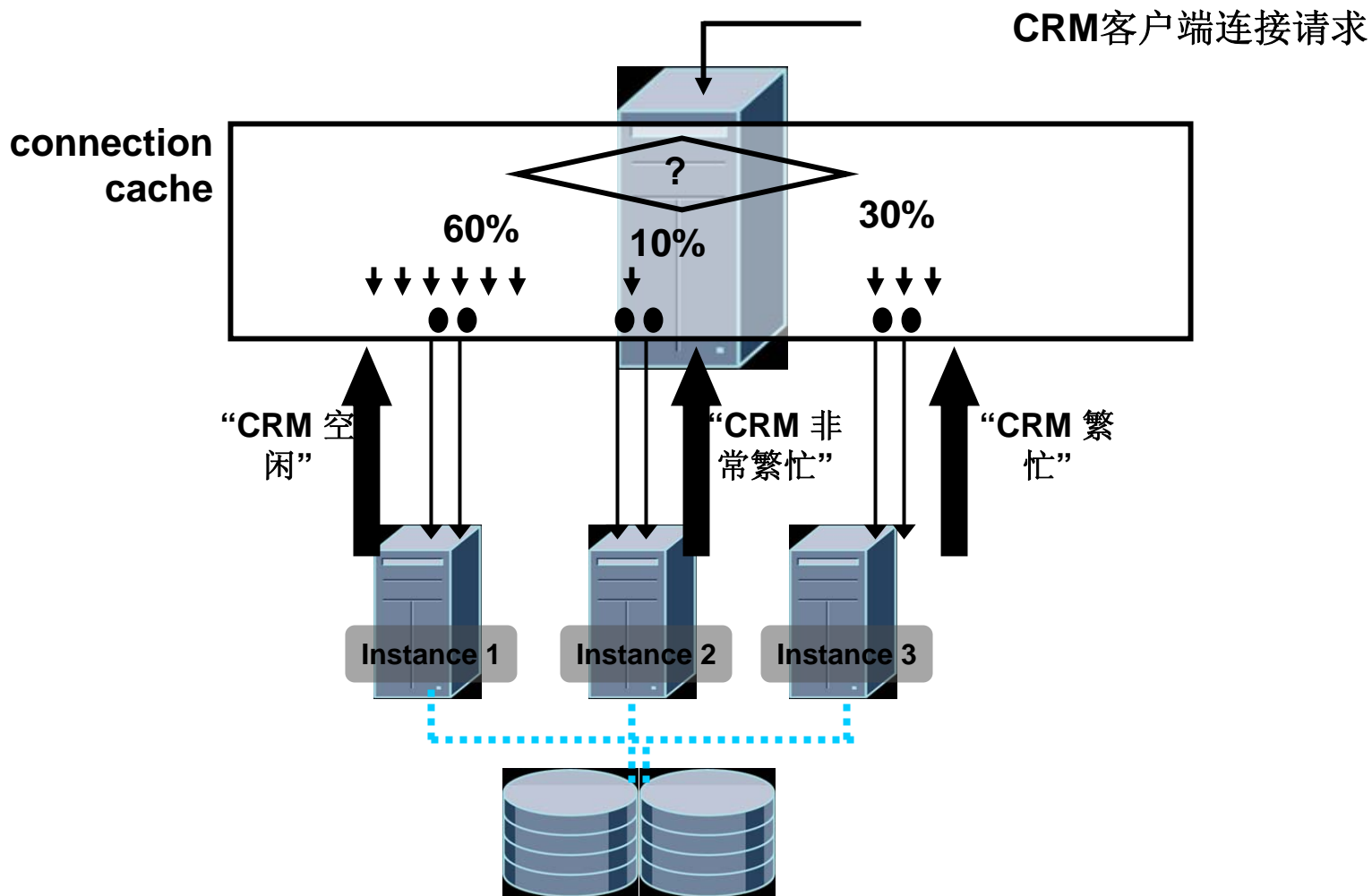
实例:
FINPROD4

数据库服务: FINPROD



FAN/FCF

运行时连接均衡负载



运行/维护



11. 测试、测试、测试...

- 为什么需要测试？

- 验证系统的基础架构满足**SLO**
- 验证安装/配置是否正确
- 期待发现“问题” – 每个应用程序对底层技术体系的考验都是不同的
- 积累经验和技能

- 如何测试？

- 测试计划
- 单独测试集群
- 产品在配置/技术体系上的匹配
- 现实的工作负荷
 - **最可靠，但是实现起来困难/昂贵**
- 功能、性能和破坏性测试
- 包括正常和非常规（例外）的操作流程

12. 实施变更控制

- 坚持强健的系统生命周期定律
 - 全面的测试计划 (功能、压力测试)
 - 演练生产迁移计划
- 变更控制
 - 独立的开发、测试、**QA/UAT**和生产环境
 - 系统和应用程序变更控制
 - 记录 **spfile**的变化
 - 建立性能/吞吐量的基线
 - 定义维护时间窗口(补丁、补丁集、升级)
 - 管理工作负荷的变更
 - 部署新的应用/功能模块
 - 业务周期变化
 - 业务高峰时期停止非核心的处理

13. 定义支持流程

- 和技术合作伙伴一起制定并签署技术支持的流程
- 侦测→捕获→重现→分析→解决→侦测 →...
- 重启系统之前捕获尽可能多的诊断信息。
 - 权衡恢复所需时间
 - 查看Metalink Note xxxxx诊断信息收集指南
 - 需要能够收集系统状态的Dump信息和做挂起分析

14. 性能监控

- 建立性能的基线
- AWR / Statspack
- 自动数据库诊断监控ADDM
- 活动会话历史ASH

诊断和确定问题

15. 补丁/软件维护

- 与当前CPU匹配
- 从售后服务部门获得当前推荐的补丁列表
 - 从售后服务部门获得特定平台的特定信息
 - >*opatch lsinventory -detail* 来确认不存在补丁冲突
- 认真阅读单个补丁的readme
 - 补丁安装的方式不完全一样
- 确人补丁成功安装到所有节点
- 应该首先在测试/QA环境中安装补丁

16. 避免错误的驱逐节点

- 跟单机系统一样，监视关键的系统资源
 - CPU, I/O, 内存, 交换区, 网络
- 关键进程如果不能及时响应，将导致‘心跳’失败
 - 系统不能够长时间运行在100% CPU利用率状态
 - 把控制文件和voting disk放置在高性能存储设备上，确保良好的I/O响应时间
 - 给LMS提供实时优先级

总结

1. 定义服务水平目标
2. 设置现实的期望值
3. 无需听信传言
4. 尽量简单化
5. 使用集群验证工具
6. 正确地配置集群互联
7. 镜像OCR/Voting disk
8. 使用虚拟IP地址
9. 使用自动存储管理(ASM)
10. 实施工作负载管理、不仅仅是负载均衡
11. 测试、测试、测试
12. 实施变更控制
13. 定义支持流程
14. 性能监控
15. 软件维护
16. 避免错误的驱逐节点



Q
QUESTIONS
A
ANSWERS
A



ORACLE®



ORACLE IS THE **INFORMATION** COMPANY

甲骨文

甲骨文

甲骨文