淘宝数据仓库环境介绍

陈吉平
淘宝网首席DBA，资深技术专家
Oracle ACE
《搭建Oracle高可用环境》一书作者

淘宝网 Taobao.com
阿里巴巴旗下网站

淘宝介绍

- 淘宝网（www.taobao.com）是国内领先的个人交易网上平台，由全球最佳B2B公司阿里巴巴公司投资4.5亿创办，致力于成就全球最大的个人交易网站。自2003年5月10日成立以来，淘宝网基于诚信为本的准则，从零做起，在短短的半年时间，迅速占领了国内个人交易市场的领先位置...

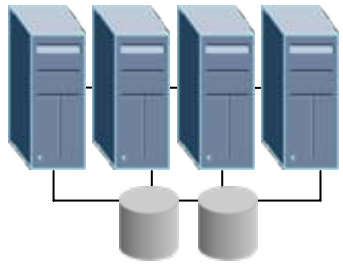
议程

- 淘宝数据仓库的发展历史
- 淘宝数据仓库的现状
- 淘宝数据仓库的维护
- 淘宝数据仓库的未来发展

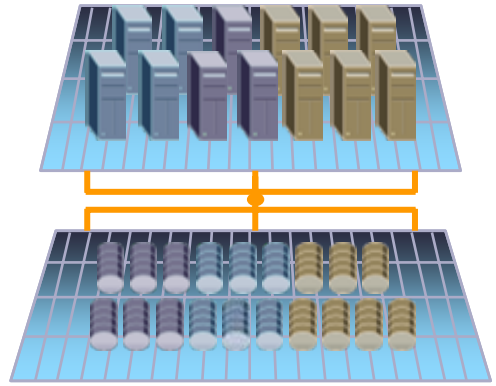
淘宝数据仓库的发展历史

单机-
DELL6650

IBM-P550



4节点RAC环境



12节点RAC环境
计划扩展到20节点

2004 **2006** **2007** **2008**

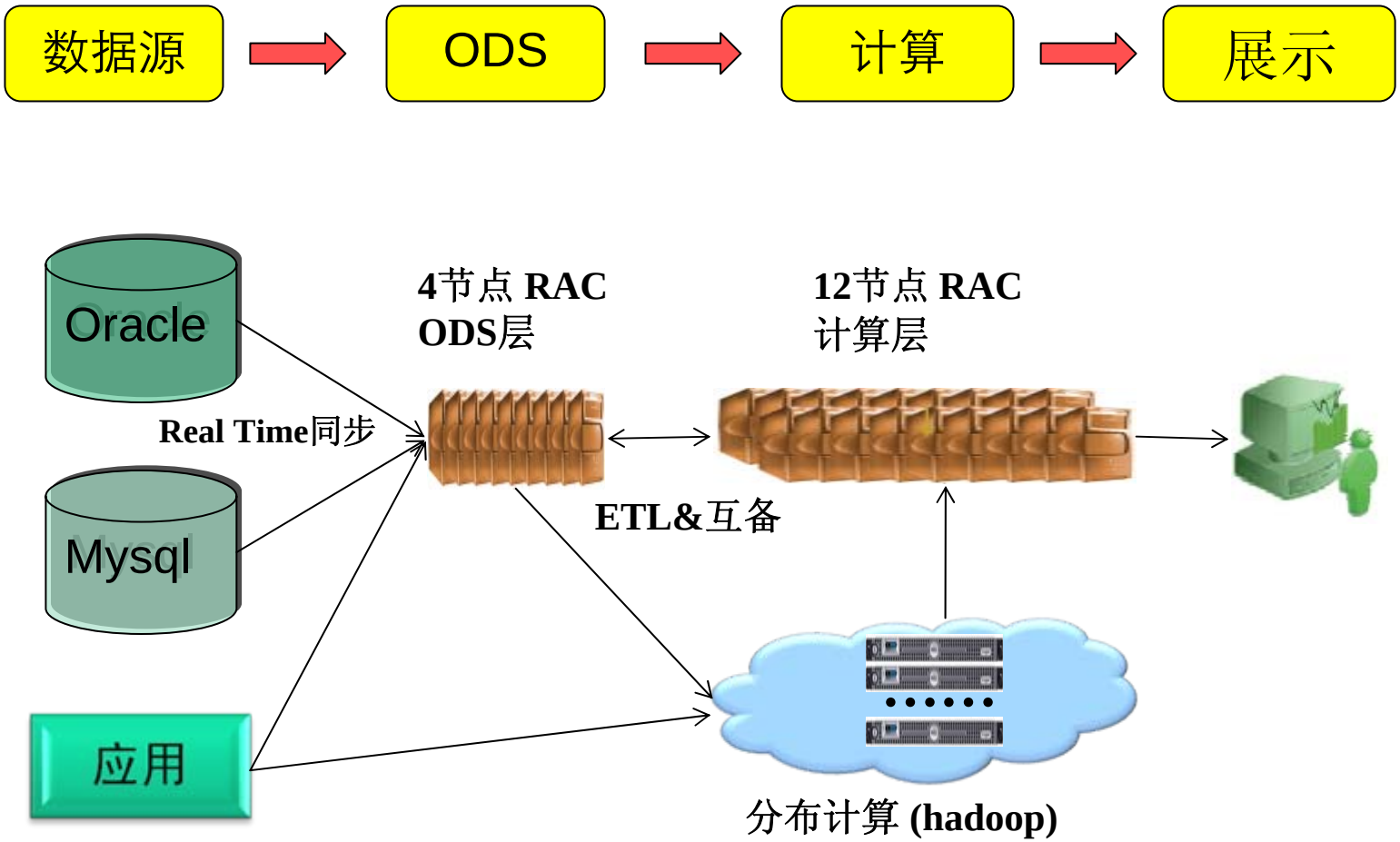
为什么选择Oracle RAC

- 当时的可选方案
  - Teradata，在ebay有成功案例
  - Oracle Rac，对淘宝来说，技术比较成熟
- 主要考虑的问题：
  - 1. 效率，是否能快速的满足日常需求
  - 2. 开放性，开放系统还是选择封闭系统
  - 3. 可扩展性，包括线性扩展能力，对外接口等
  - 4. 成本因素

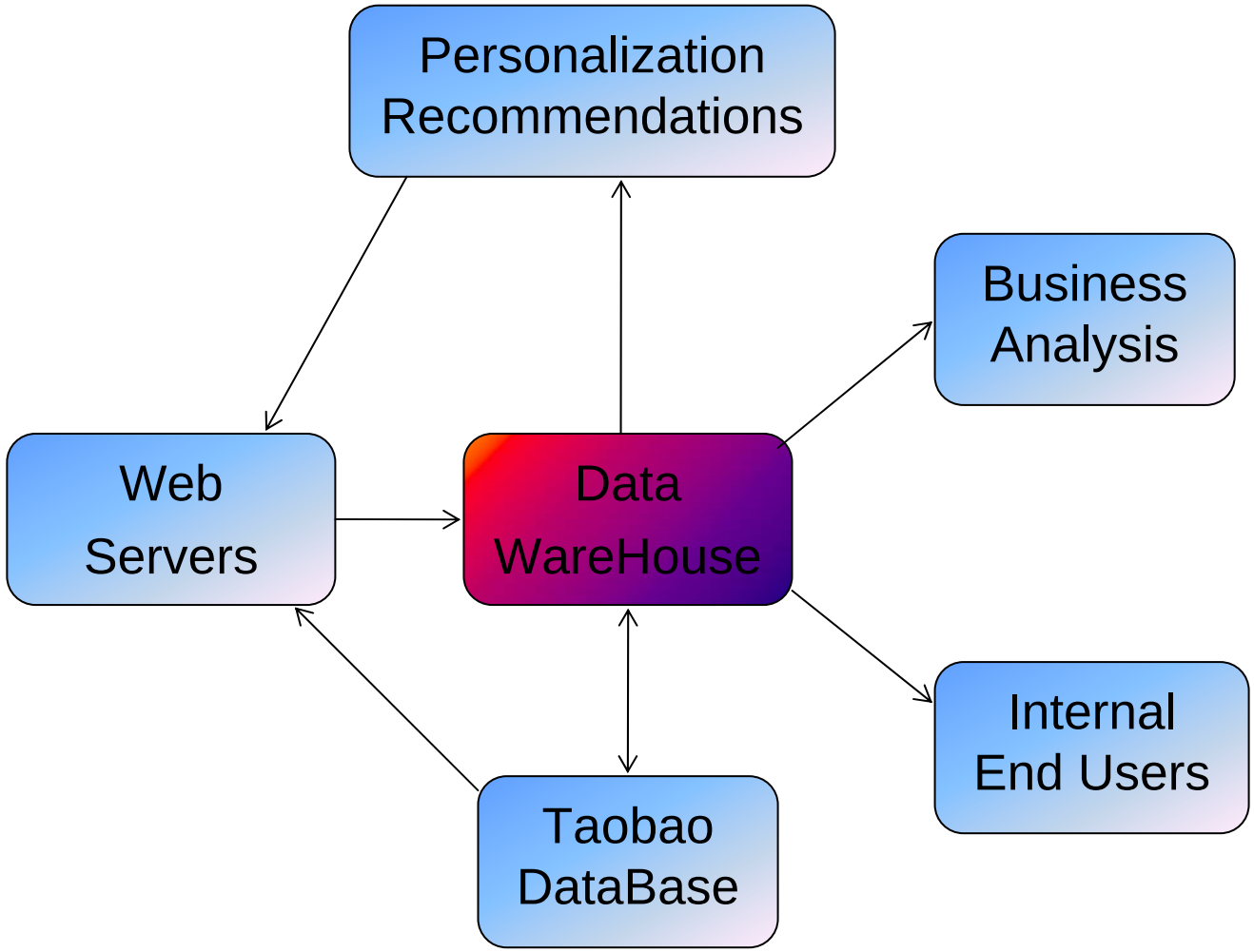
为什么选择Oracle RAC

- Oracle RAC优点
  - Oracle是一个相对开放的系统，文档很多
  - Taobao本身对Oracle的技术比较了解
  - 在实际的测试中，Oracle表现比较优越
  - 可以做到线性的扩展能力
  - 成本不会高于Teradata
- Oracle RAC缺点
  - 对技术的要求相对比较高，需要DBA深入的优化

淘宝数据仓库现状


数据源
ODS
计算
展示
Oracle
Mysql
应用
Real Time同步
4节点 RAC
ODS层
12节点 RAC
计算层
ETL&互备
分布计算 (hadoop)

淘宝数据仓库的应用场景



淘宝数据仓库应用案例（一）—淘我喜欢

淘宝数据仓库应用案例（二）—猜你喜欢

收藏该店铺的用户最喜欢的其它宝贝：

￥6.99

h61116【特】 Fruit of the Earth 100%纯芦荟胶56g

收藏人气: 5552

加入我的推荐

￥1.90

J00905脚丫双面吸盘~颜色随机发送

收藏人气: 738

加入我的推荐

￥9.49

h60321【特】最爱瘦身教馆冲凉瘦★消脂收紧去橘皮

收藏人气: 1092

加入我的推荐

￥1.29

h61119【折】女人我最大韩伊橄榄特润美白保湿护手霜

收藏人气: 2472

加入我的推荐

淘宝猜你还喜欢什么店铺:

双生儿四周年双金冠 来

掌柜: 双生儿香港平价店

信用: 1013848

心蓝t透双金冠冲击！全

掌柜: 心蓝t透0

信用: 980570

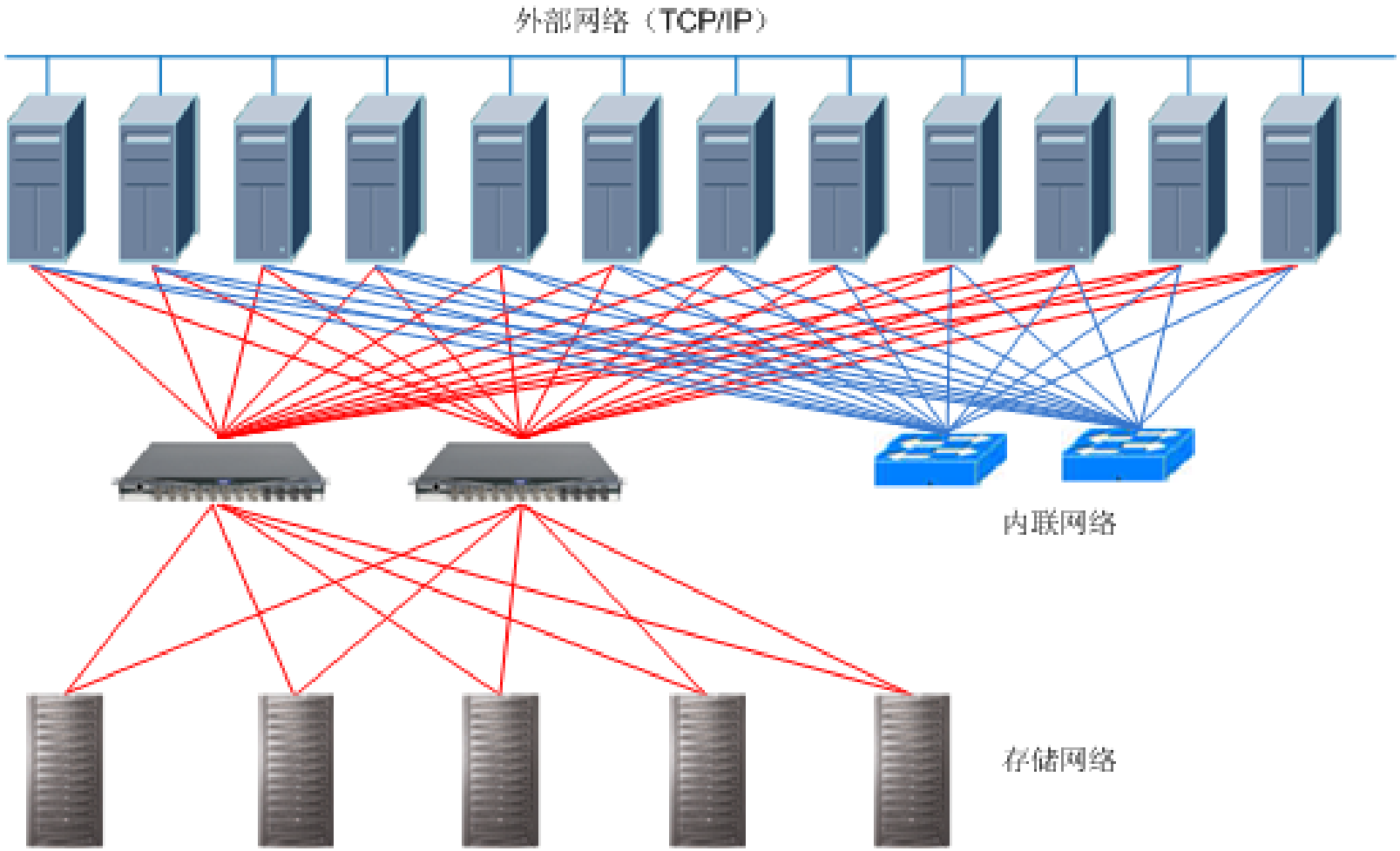
『女人天生是妖精』店内

掌柜: 妖精de国度

信用: 567326

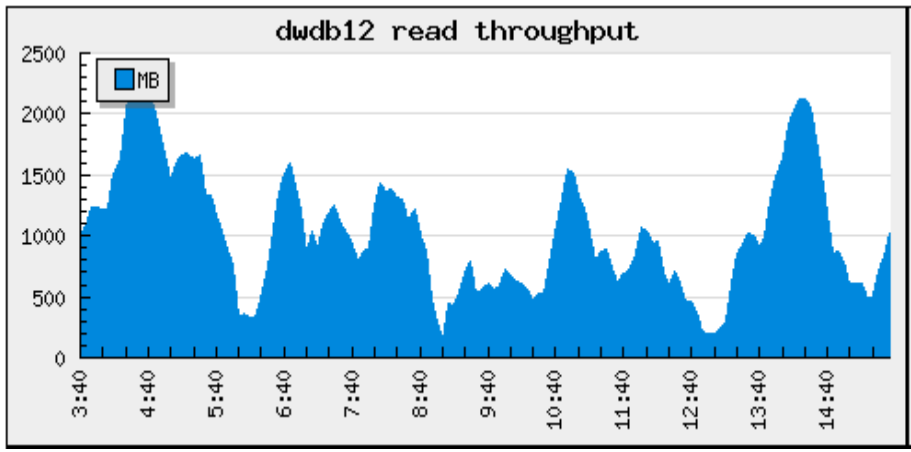
淘宝数据仓库应用案例（三）—内部系统

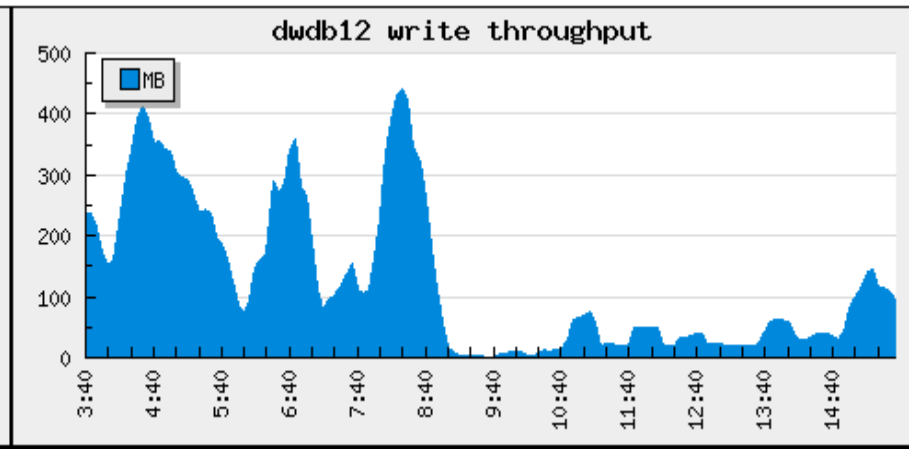
淘宝数据仓库12节点物理拓扑图



淘宝数据仓库12节点的承载量

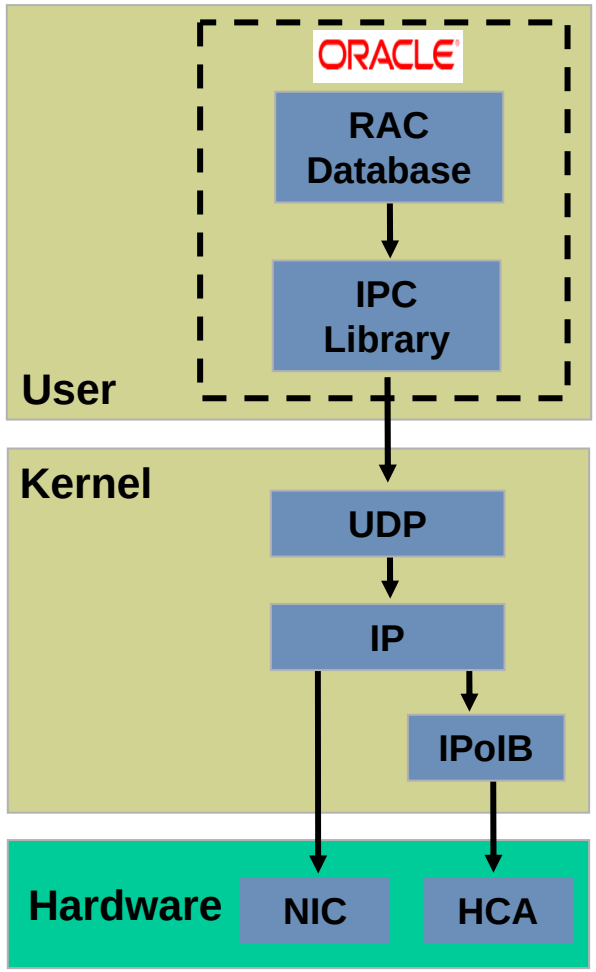
- 上线前测试情况
  - 5台CX3-80，每台120块146G 15K光纤硬盘，12个节点同时跑Orion测试：
  - orion -run advanced -testname mbps_all_point_20 -num_disks 500 -size_small 1024 -size_large 1024 -matrix point -num_small 0 -num_large 20 -type seq -verbose
  - 读总吞吐量达到7721.26MB/s
- 目前使用量：读2.5GB/s，写500MB/s





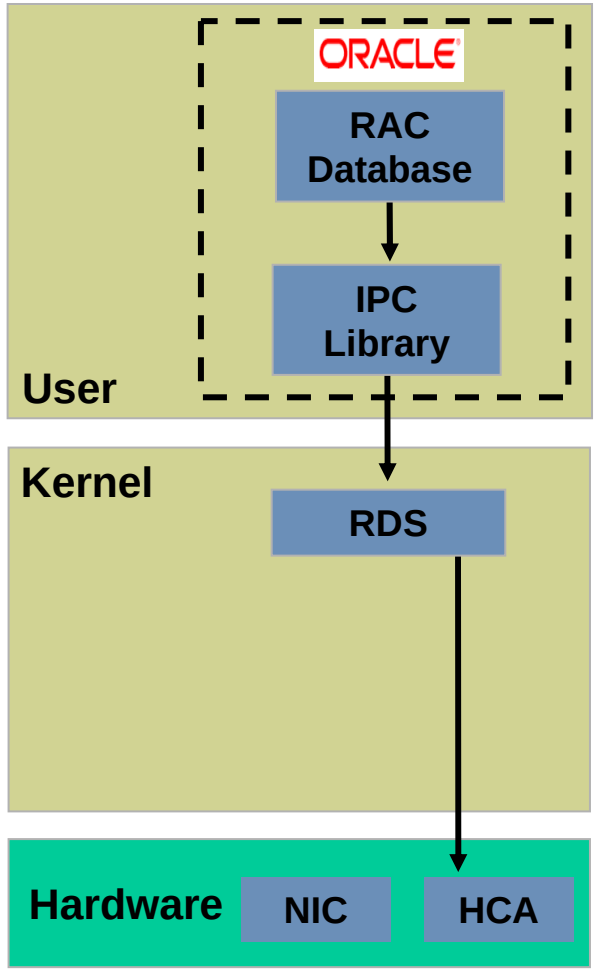
实施细节（一）——内联网络

- 普通千兆网卡
- Infiniband高速交换



IPOIB

300-500MB/s，延迟几十us



RDS(certified by oracle 10.2.0.2)

速度1GB/s，延迟10us

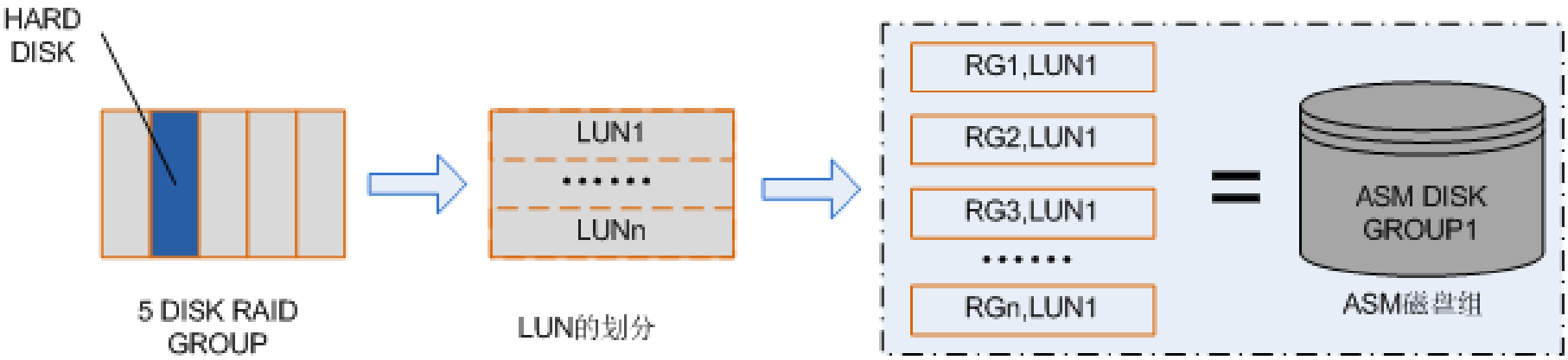
关于infiniband

- 对于IPOIB
  - oracle端不需要做任何调整
  - 检测
    - oradebug setmypid
    - oradebug ipc
    - oradebug tracefile_name
    - netperf -H 10.0.100.104 -l 60(Linux tool)
- 对于RDS
  - 安装
    - cd $ORACLE_HOME/rdbms/lib
    - make -f ins_rdbms.mk ipc_rds ioracle
    - 详细见“Oracle Real Application Clusters 10g Release 2: Installation and Configuration of Linux Clusters Using RDS over InfiniBand Interconnect” http://www.oracle.com/technology/products/database/clustering/pdf/rds_installation_on_oracle_rac_10g_release_2_linux_clusters.pdf
  - 对于RDS速度的检测
    - crload/crtest(Oracle Developer team’s tool)
    - Oradebug中无法看到RDS信息，oracle BUG 5610067
    - paul.tsien@oracle.com is the PM of RAC & RDS team

实施细节（二）——网格存储ASM

- 关于ASM
  - ASM的性能十分好和裸设备基本没有差别，在测试中可以把存储性能压榨干
  - ASM的strip做的也很好，通过存储监控可以看到I/O分布很均匀
- 对于rebalance
  - 在RAC中可以利用ASM来rebalance
  - alter diskgroup taobao rebalance power 11;
  - Rebalance速度很快，测试过程中，新增加3TB磁盘的同步用时1.5个小时

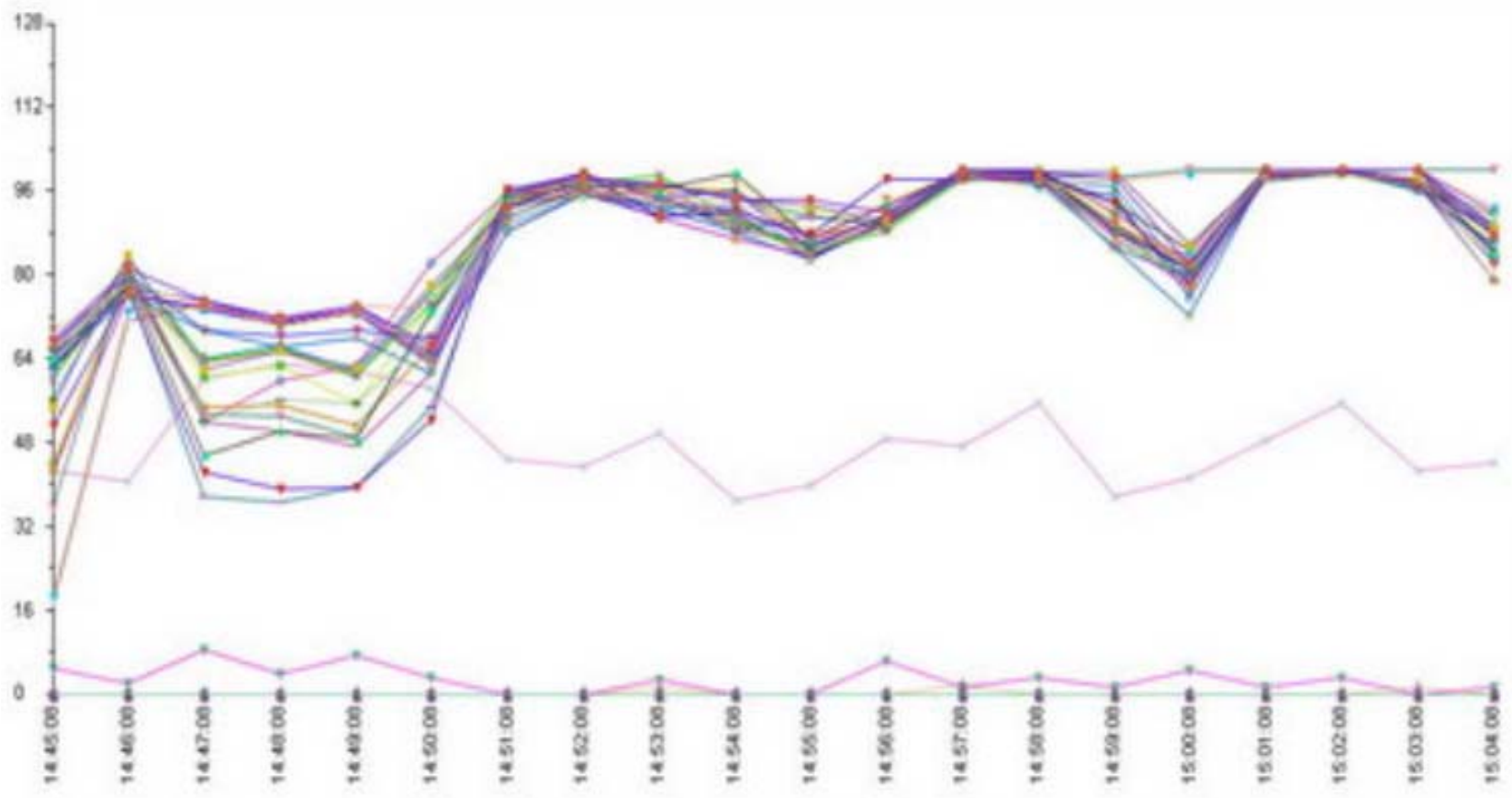
ASM磁盘组的划分



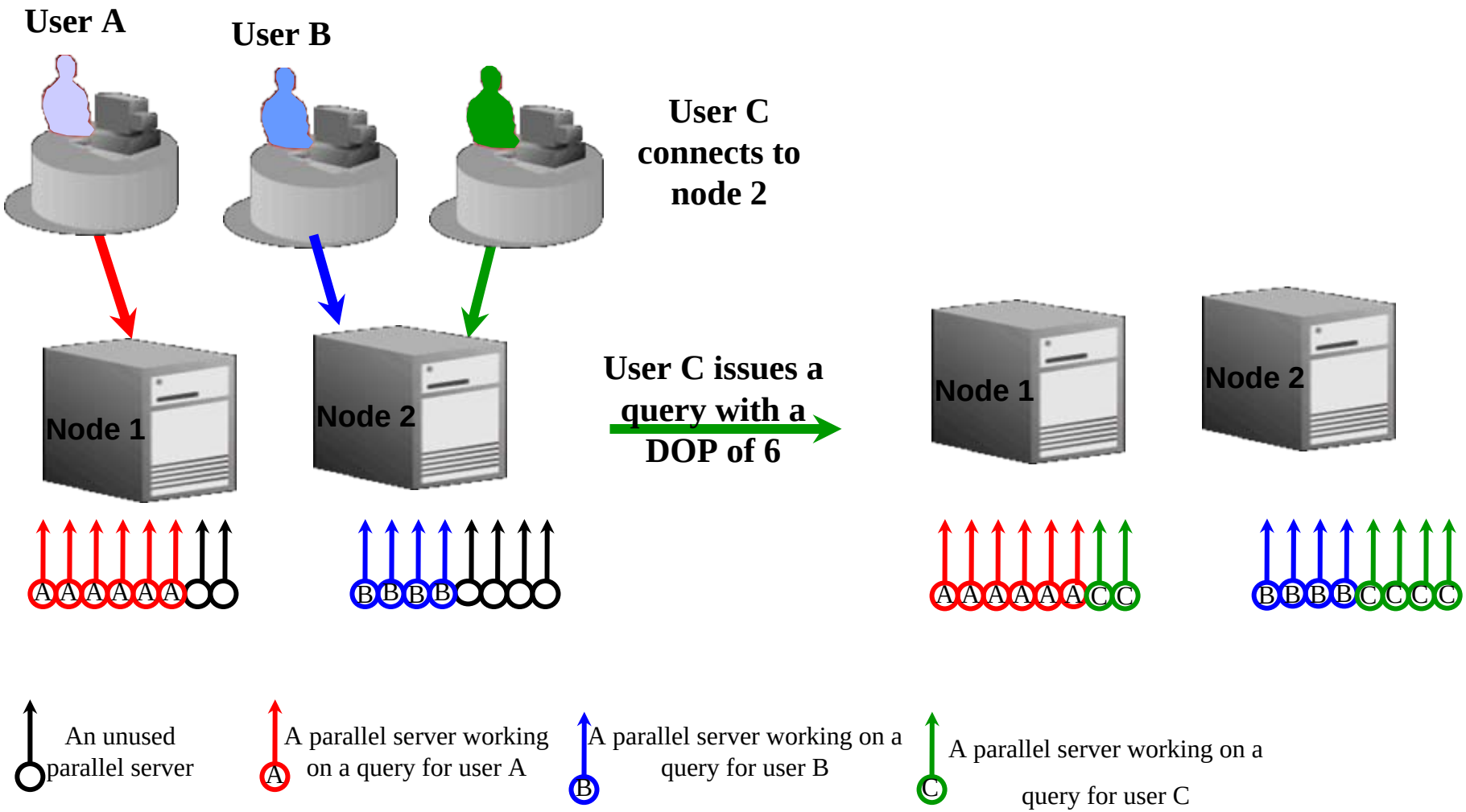
- 4+1 的RAID5
- 每个RAID组划分2-4个LUN
- 每个RAID组的第一个LUN组成ASM磁盘组1，第二个LUN组成ASM磁盘组2，依次类推
- 数据库建立在ASM磁盘组上
- 可以确保数据分布在所有的物理磁盘上

ASM的测试与使用效果

- 数据基本上分布到了所有的LUN上面



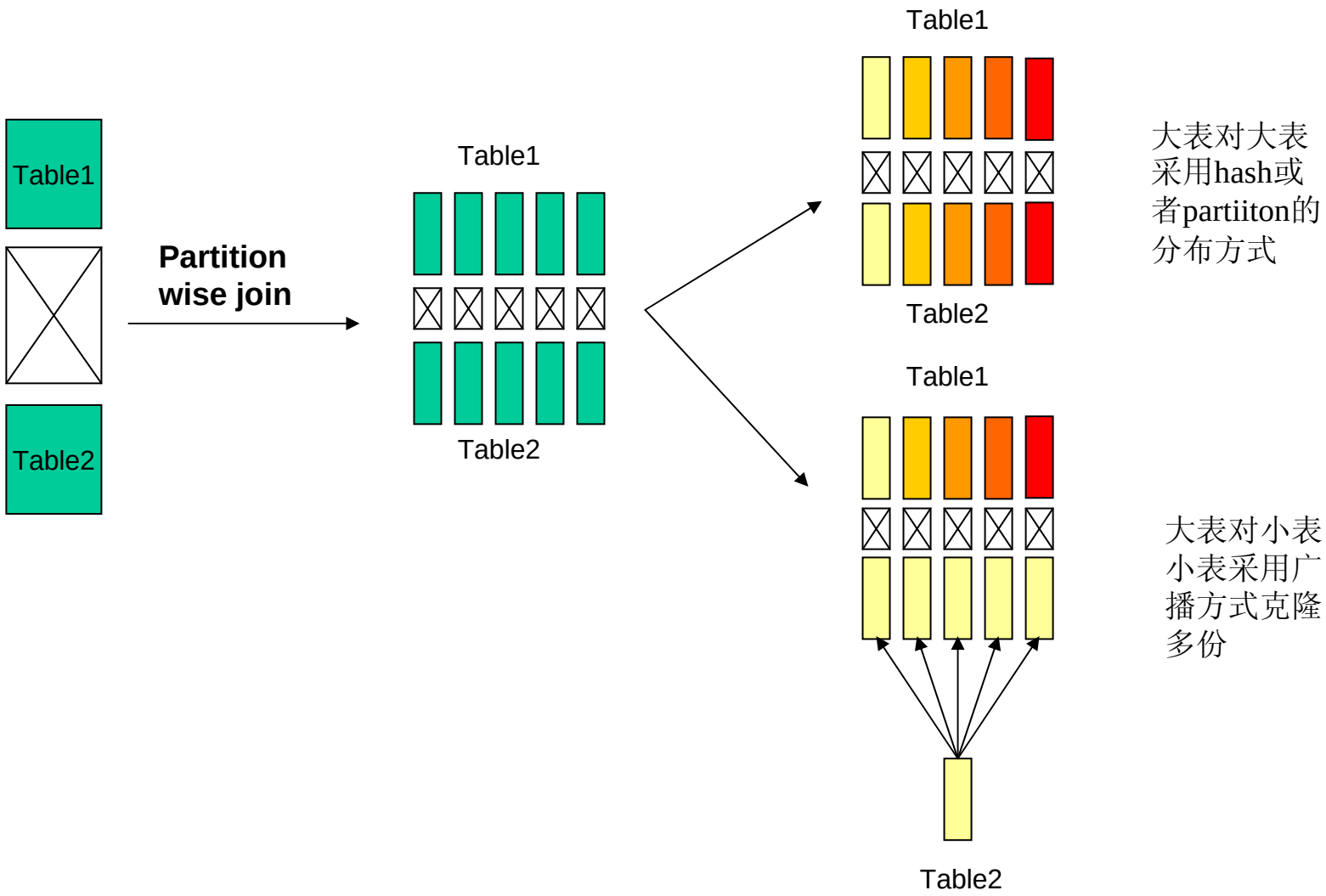
实施细节（三）——并行执行


User A
User B
User C connects to node 2
Node 1
Node 2
User C issues a query with a DOP of 6
Node 1
Node 2
A A A A A A
B B B B
A A A A A A C C
B B B B C C C C
An unused parallel server
A
A parallel server working on a query for user A
B
A parallel server working on a query for user B
C
A parallel server working on a query for user C

并行执行相关参数

- 并行度参数，如
  - parallel_max_servers=256
  - parallel_min_servers=16
  - parallel_threads_per_cpu=2
  - parallel_execution_message_size=16384
- 跨节点并行的参数
  - instance_groups=crm,erp,oltp
  - parallel_instance_group=crm
  - 关于跨节点并行的更多资料，可以参考我们团队的Blog：http://rdc.taobao.com/blog/dba/html/217_rac-span-paralle.html

并行执行的Join方式



案例分析（一）——升级中的问题

- 08年数据仓库从10.2.03升级到10.2.04
  - 在升级之后碰到了一个新的BUG，数据库会报如下600错误：
  - ORA-00600: internal error code, arguments: [kddummy_blkchk], [47], [935468], [18038], [], [], [], []
  - 数据库部分查询报块损坏，无法执行

升级中的问题—解决办法

- 临时解决办法是将db_block_checksum这个参数改为FALSE
- 后面ORACLE新发布了一个补丁5386204，打上补丁后问题得以解决。

案例分析（二）——节点异常退出

- 08年12节点刚开始使用cisco的infiniband交换机，总是出现如下的情况：
  - [CSSD]2008-09-02 17:48:15.706 [1241577824] >WARNING: clssnmPollingThread: node dw6 (6) at 50 3.118329e-317artbeat fatal, eviction in 29.490 seconds
    .....
  - [CSSD]2008-09-02 17:49:15.677 [1241577824] >WARNING: clssnmPollingThread: node dw5 (5) at 50 3.118519e-317artbeat fatal, eviction in 29.780 seconds
    .....
  - [CSSD]2008-09-02 17:49:46.460 [1262557536] >ERROR: clssnmCheckDskInfo: Aborting local node to avoid splitbrain.
    [CSSD]2008-09-02 17:49:46.460 [1262557536] >ERROR: : my node(11), Leader(1), Size(10) VS Node(6), Leader(1), Size(12)
  - 节点6与其它节点内部通信超时，甚至不通，而自动重启。
  - 节点6重启之前，节点5也出现重启
  - 在节点6重启之前，有更多的节点被踢出（重启），直到开始正式重启

节点异常退出—解决办法

- 原因分析
  - 怀疑cisco的infiniband在高压下不稳定，网络出现延迟。
  - 当一个节点负载压力太高，被迫重启的时候，与其它节点内部网络出现超时甚至不通
  - 在该节点重启之前，其它节点认为该节点还正常（Voting Disk是正常的）
  - 其它节点会出现脑裂，也会不同程度的把自己从RAC环境中踢出
- 解决办法
  - 将infiniband网卡替换成普通网卡，内部互联就恢复正常，再也没有节点自动宕机的情况了

未来的计划

- 扩展到20节点
- Oracle 11g的新特性
- 继续尝试使用Infiniband交互技术
- Oracle新硬件Exadata的测试

谢谢

- Q&A