Information Management

IBM

Oracle与DB2那些事（二）
- DB2 Purescale群集

Data Management

IBM 中国软件开发实验室
王飞鹏

DTCC2012

议程

- **数据库群集与DB2 pureScale的起源**
- **DB2 pureScale 技术概览**
- **DB2 pureScale 群集技术的特点**

Information Management

IBM

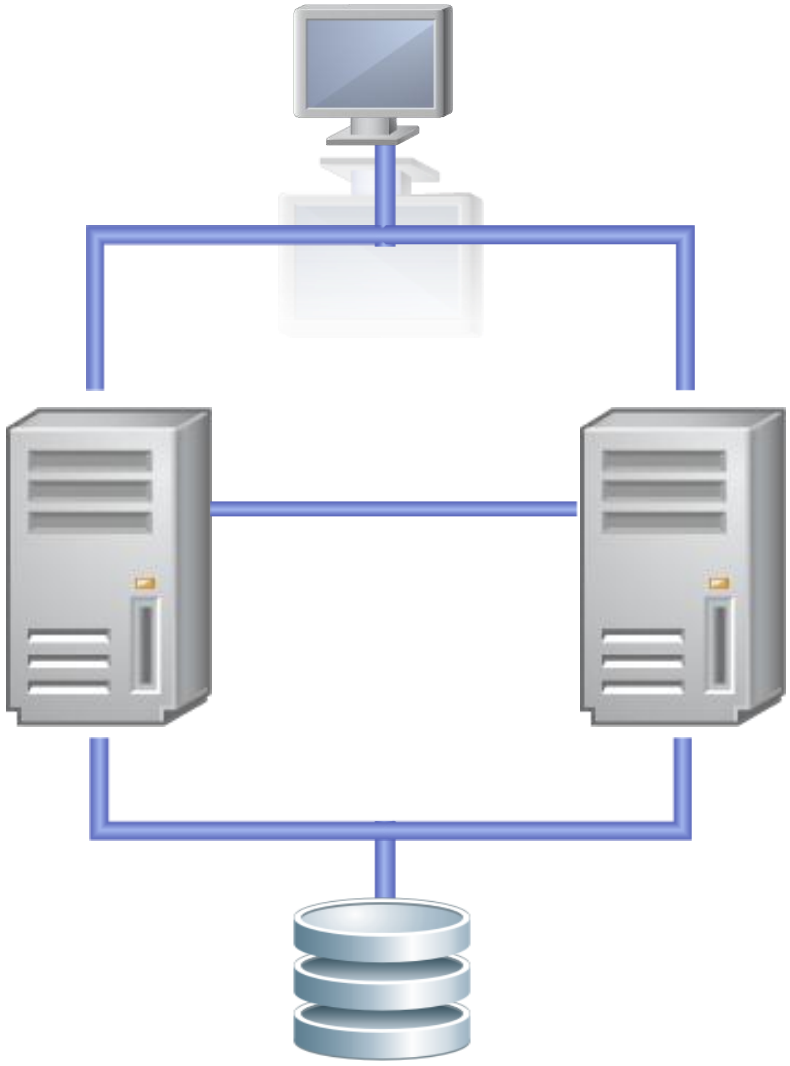
Data Management

群集数据库与 DB2 pureScale

DTCC2012

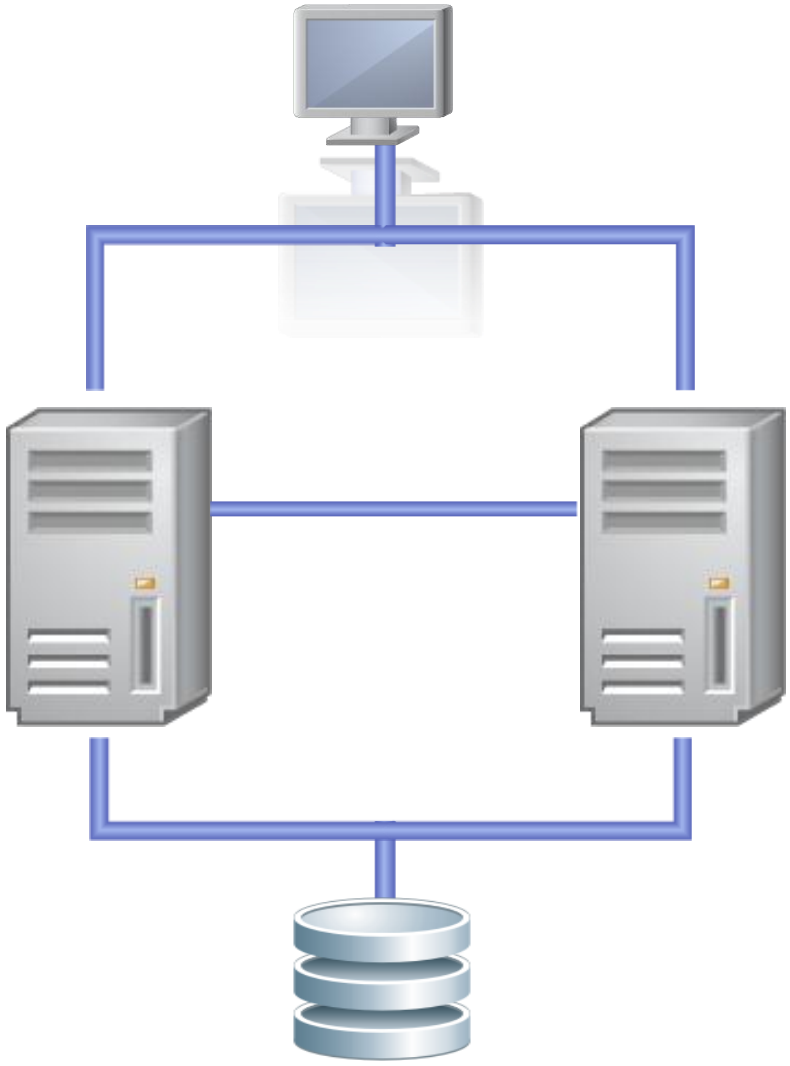
群集环境下的数据库系统

- **群集环境的数据库系统起源**
  - 在单处理机系统的时代，CPU常常成为数据库系统的瓶颈
  - 多端口存储设备及相应控制软件出现后
  - 多服务器群集的数据库系统应运而生
- **群集数据库系统的用途**
  - 提高吞吐量
    - 分担用户的访问请求
  - 高可靠性
    - 故障接替
      - 高可靠性处于从属的地位
- **群集数据库系统的技术挑战**
  - 通信交互导致扩展性的损失
    - 访问封锁机制
      - 分布式锁管理器
    - 页面缓存
      - 服务器间的通信、交换



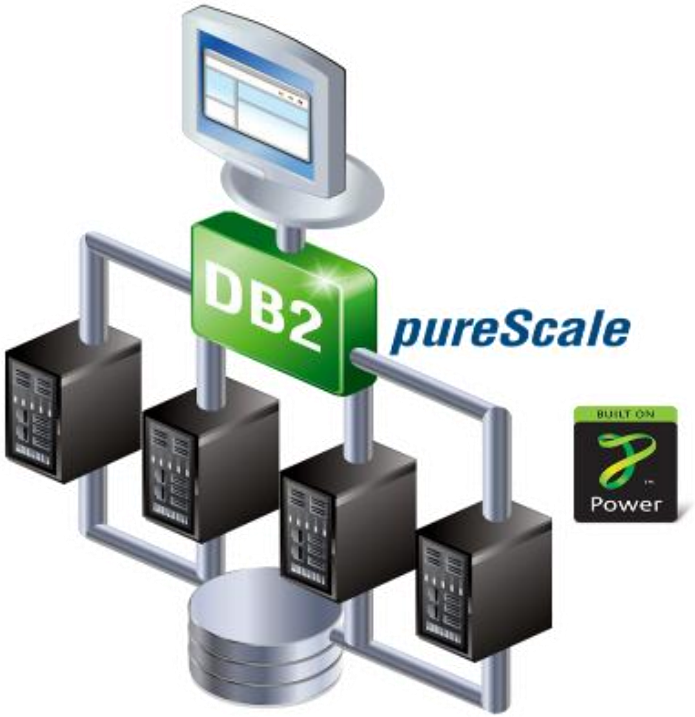
群集数据库系统的发展

- **多处理机系统主流时代**
  - SMP大规模出现之后，加上处理机性能的革命性突破
    - 数据库系统在CPU上的压力减小
    - IO压力成为主攻对象
  - 现有的单台服务器可扩展近百颗CPU，几乎能应对所有OLTP的应用场景
  - 高可靠性则可以通过承担不同任务的服务器相互监控和故障接替来实现
    - 在多分区环境下有更多的配置选择
- **群集数据库产品的发展**
  - 20年来只有Oracle数据库提供双机共享磁盘设备的解决方案
    - 近年来称为RAC
    - 应用效果“见仁见智”
  - 其它任何厂商均没有提供同类产品
    - 市场观点、技术路线选择
- **新时代、新要求**
  - 低成本、多服务器、共享文件系统
  - 网格、云计算



DB2 pureScale —— 满足新时代需求的群集数据库系统

- **无限产能**
  - 仅购买所需要的设备，随需提高产能
- **应用透明性**
  - 避免应用变更带来的风险和成本
- **持续可用性**
  - 交付不中断的数据访问，确保性能一致

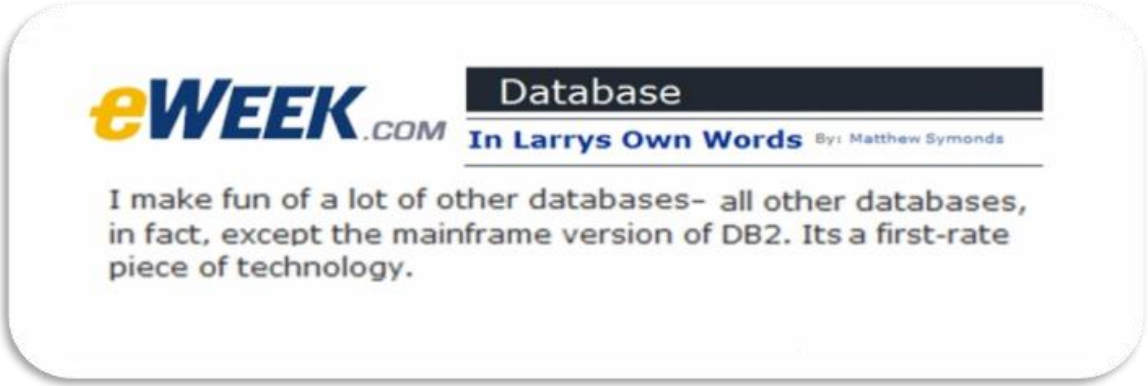


借鉴自无可争议的黄金标准……System z

DB2 pureScale 从何而来？

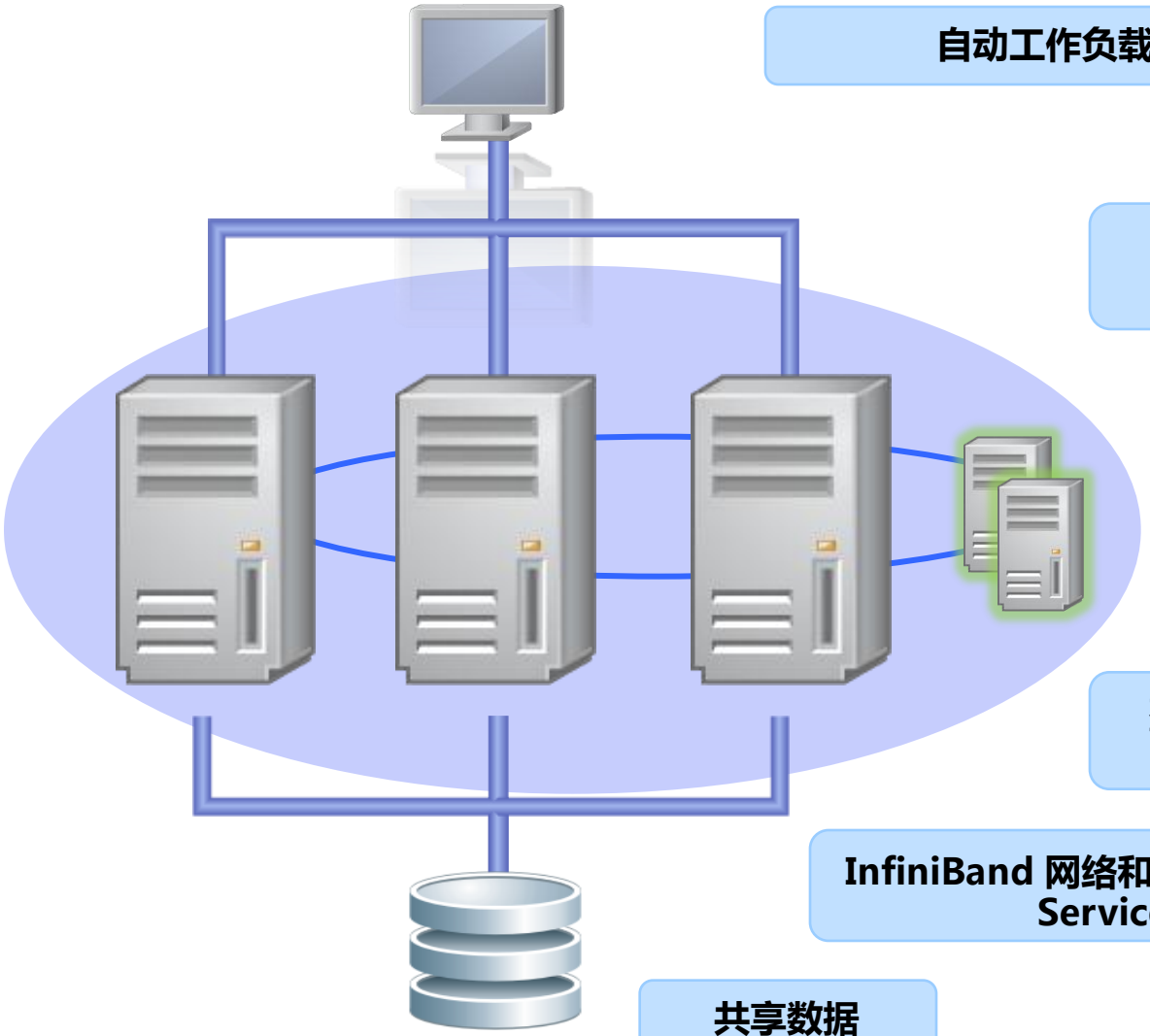
DB2 for z/OS 数据共享是“黄金标准”

- **每个人都认可 DB2 for z/OS 是可伸缩性和高可用性的**
- **甚至 Oracle 也同意：**



- **为什么？**
  - Coupling Facility！
    - 集中锁定、集中缓冲池交付了优异的可伸缩性和优异可用性
  - z/OS 上的整个环境都可用使用 Coupling Facility
    - CICS、MQ、IMS、Workload Management 等

DB2 pureScale 的架构



自动工作负载均衡

运行于 Power 服务器上的 DB2 成员集群

利用 z/OS 提供的全局锁定和内存管理器技术

整合的 Tivoli System Automation

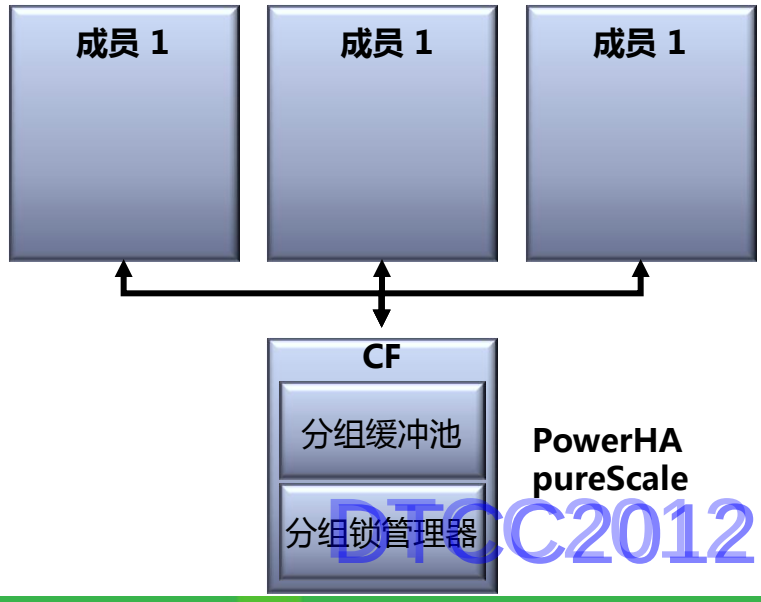
InfiniBand 网络和 DB2 Cluster Services

共享数据

DTCC2012

可伸缩性和高可用性的关键

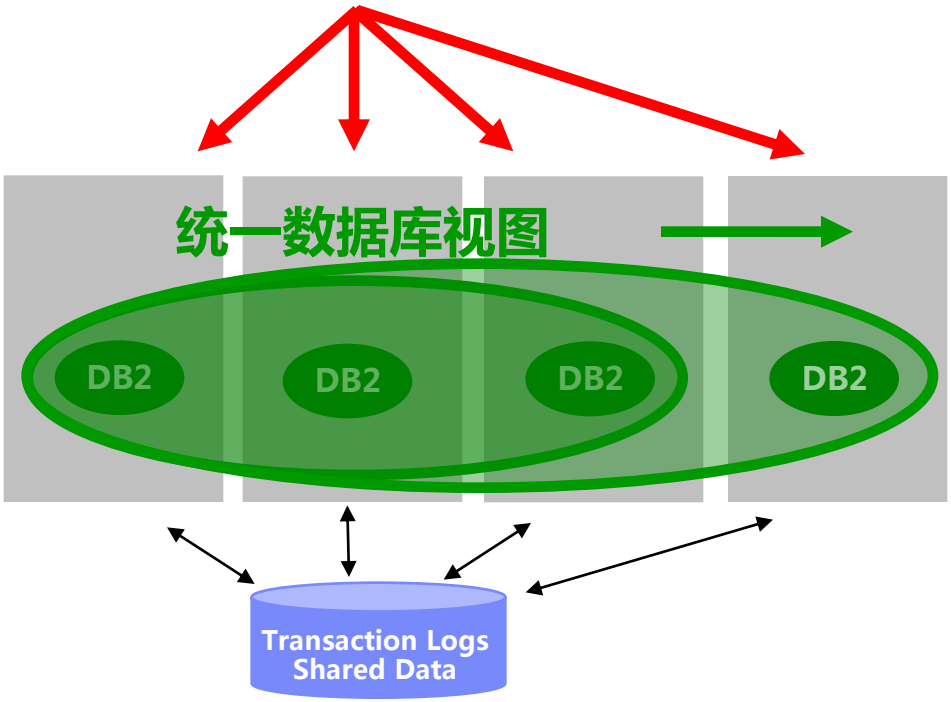
- **有效的集中锁定和缓存**
  - 随着集群的不断增长，DB2 会始终在 CF 维护锁定信息和共享页面
  - 针对超高速访问而优化
    - DB2 pureScale 使用 Remote Direct Memory Access (RDMA) 与 PowerHA pureScale 服务器通信
    - 没有 IP 套接字调用、没有中断、没有上下文切换
- **结果**
  - 为大量服务器提供接近线性的可伸缩性
  - 持续感知各成员当时的工作状态
    - 如果其中一个成员出现故障，不会造成
      - 其他成员 I/O 阻塞
      - 以内存速度恢复运行



易扩展

扩展

- ✓ 不需应用程序显著修改的完美扩展
- ✓ 对于数据所属节点没有限制
- ✓ 灵活适应工作负载路由



快速部署新成员

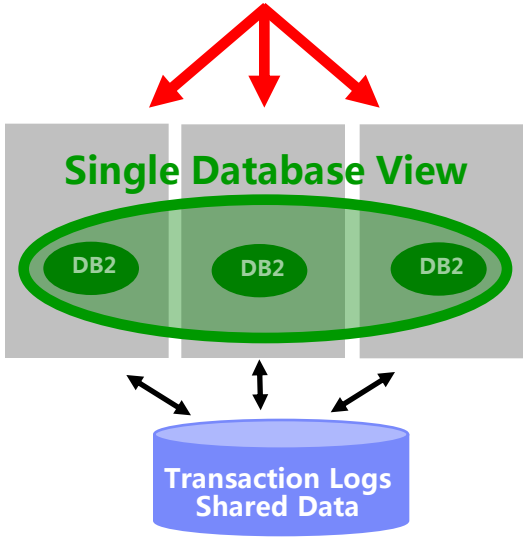
- ✓ 不需要数据重新分布

DTCC2012

易维护和升级

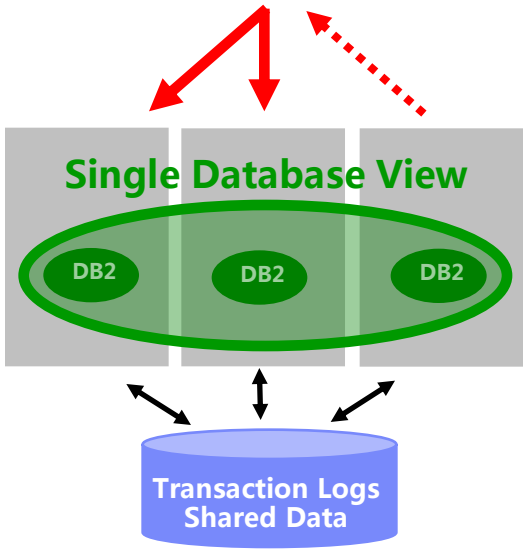
1)运行系统

- **设定目标节点**
- **（可选）增加一个新的节点以保证整个系统的整体能力**



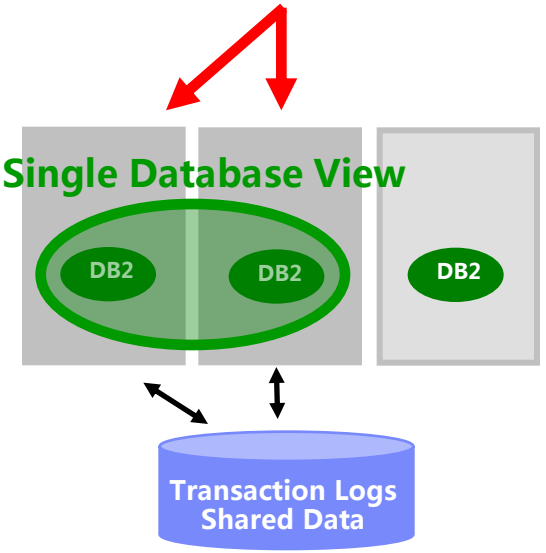
2) 排干（Drain）目标节点

- **停止新的路由**
- **允许已有交易完成**



3) 执行维护工作

- **排干完成后**

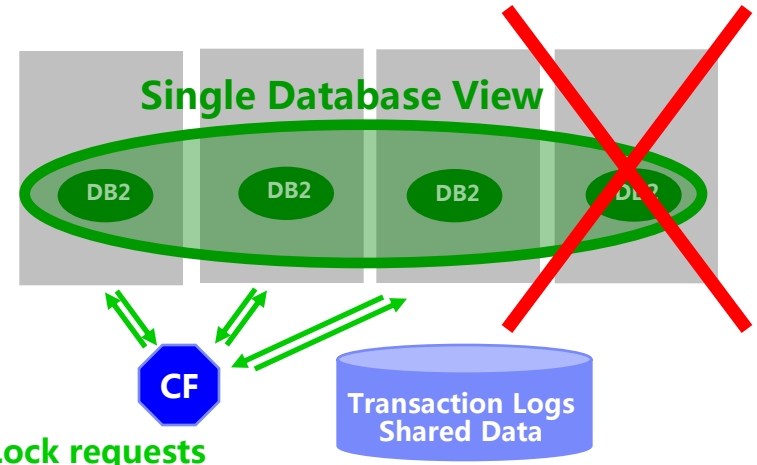


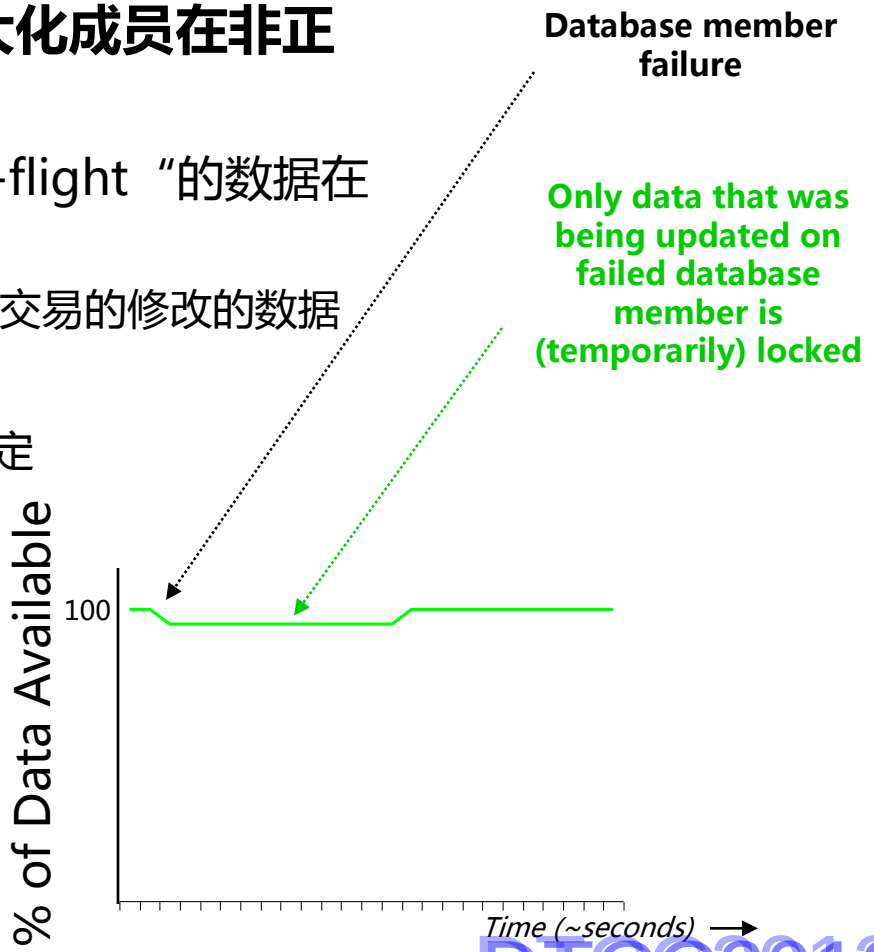
在系统可用性上无断点
无Quiesce时间; 不需要对已有工作强制回滚

DTCC2012

最小化非计划宕机时间

- **DB2 pureScale 的设计重点就是最大化成员在非正常宕机的情况下的可用性**
  - 当数据库成员失败的情况下，只有”in-flight“的数据在成员恢复完成前被锁定
    - In-flight = 在成员失败时在该成员上参与交易的修改的数据
  - 目标成员恢复时间： 10-15 秒
    - 失败成员上的只读数据在这段时间不被锁定





This example assumes about 5% of the database data was being updated on the database node that failed, at the time of the failure.

Information Management

IBM

Data Management

DB2 pureScale 技术概览

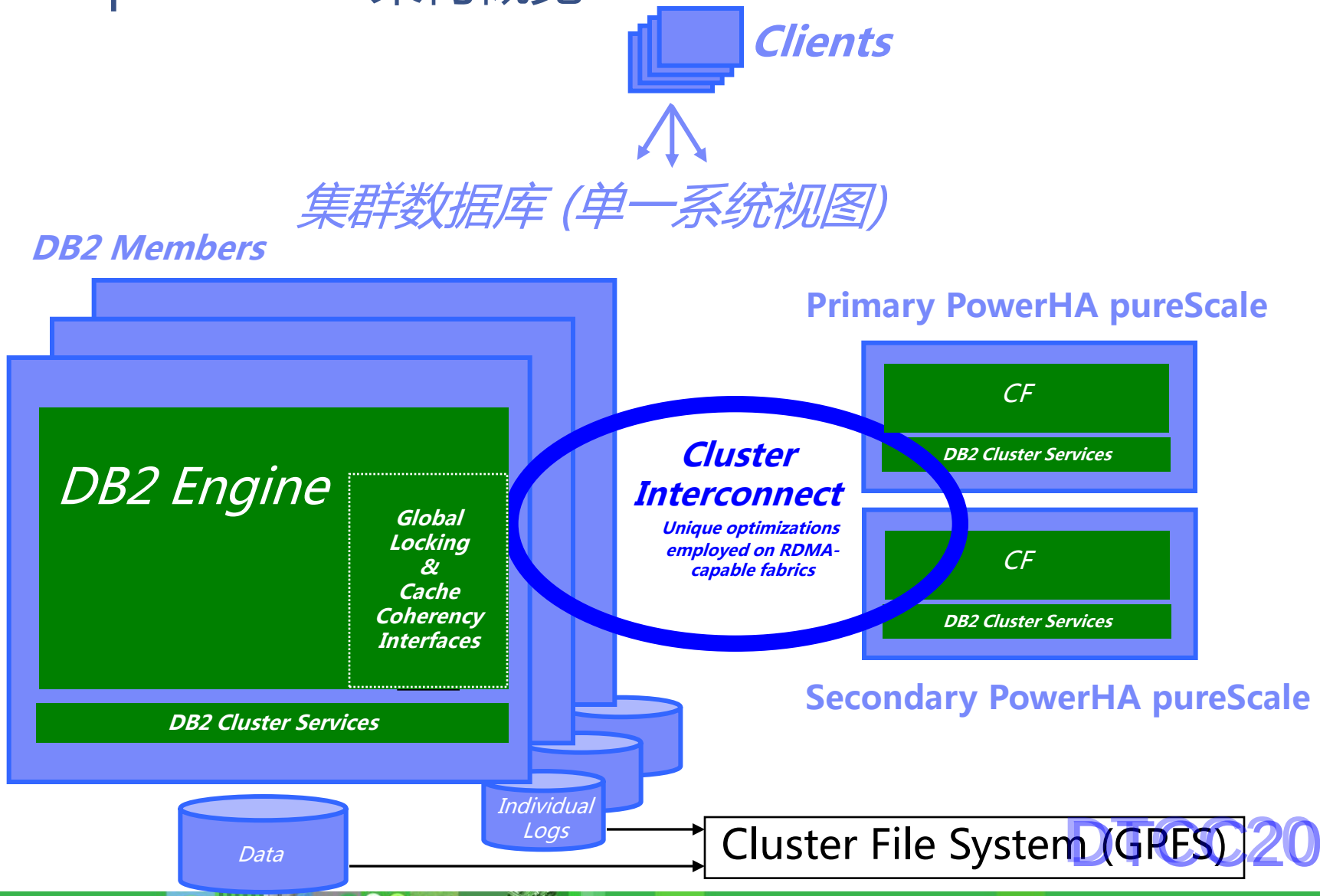
DTCC2012

DB2 pureScale 整体架构

DB2 pureScale 系统

- **集群数据库服务器**
  - 持续可用性
  - 可扩展
- **集群中有多台主机，每台主机可运行成员（members）或 PowerHA pureScale，或二者同时运行**
  - 成员和 PowerHA pureScale 由 cluster interconnect (network) 和 集群文件系统紧密联系
  - 成员和 PowerHA pureScale 共同组成 DB2 pureScale 实例
- **实例，主机，成员，PowerHA pureScale，网络和集群文件系统的状态有 DB2 Cluster Services 管理和监控**

DB2 pureScale 架构概览



什么是“成员”（member）？

- **一个通过集群文件系统访问共享数据的数据库引擎**
  - 一个 db2sysc 进程的实例
- **拥有自己的内存，缓冲池，交易日志和锁机制，且自行编译和执行**
- **可以是物理机器或逻辑节点**
  - 可以是:
    - 一个成员每台主机 – 生产推荐
    - 多台逻辑成员运行在一台主机上 – 开发和 QA 推荐
  - 和 DPF 中的“node”和数据库分区类似

什么是 PowerHA pureScale?

- **PowerHA pureScale 是 DB2 pureScale Feature 的一个集成组件**
  - 在 DB2 的文档和讲义中可能使用 ‘CF’ 来替代 PowerHA pureScale
- **协调多个成员对共享数据的访问**
  - 为所有成员提供锁定和数据缓存一致性服务
  - DB2 使用它来保证数据在所有的节点上都是一致的
- **包括3个主要部件**
  - Group Buffer Pool (GBP)
    - 确保所有成员都能读到最新提交的数据页
  - Global Lock Manager (GLM)
    - 提供给成员以能够序列访问对象
  - Shared Communications Area (SCA)
    - 提供 DB2 控制数据的一致性机制，包括 control blocks, log sequence numbers (LSN) 等
- **PowerHA pureScale 应该配置一对以避免单点故障**
  - 只配置一个的做法是支持但不推荐的，只适用于测试环境

注: GBP 和 GLM 并不能替代成员在本地维持本地缓冲池和锁管理的需求

双 PowerHA pureScale 保证连续可用性

- **如果主 PowerHA pureScale 失败，辅 PowerHA pureScale 可以接管避免造成整个系统宕机**
  - 无单点故障
- **辅 PowerHA pureScale 由 DB2 保持可用状态**
  - 更新会同时发给主辅 PowerHA pureScale，请求只送给主 PowerHA pureScale
  - 辅 PowerHA pureScale 也复制 GBP, GLM, SCA
- **当启动实例时**
  - 指定的 PowerHA pureScale 会被启动为主 PowerHA pureScale
    - 辅 PowerHA pureScale 保持 peer 状态
    - 且由复制保持 peer 状态

Cluster Interconnect

- **需求**
  - RDMA capable fabric
    - 直接修改内存不消耗 CPU 资源
    - 为 zSeries Sysplex 发明
      - implemented in their proprietary interconnect
  - 高速，大容量成员间交换，且使用主辅 PowerHA purescale
- **解决方案**
  - 使用 InfiniBand (IB) 和 uDAPL (User Direct Access Programming Library) 解决性能问题
    - InfiniBand 支持 RDMA 且支持高速，大容量网络交换
    - uDAPL 降低 AIX 的 CPU Kernel 时间

Cluster File system

- **需求**
  - 共享磁盘盒共享文件系统
  - 失败成员上文件系统的 fencing
- **解决方案**
  - General Parallel File System – GPFS
    - 由 DB2 提供许可证、安装和配置
    - 同时，客户自己预先配置的 GPFS 文件系统也可以接受
      - 允许客户在企业内部统一配置 GPFS
      - DB2 不再管理预先配置的文件系统 和负责对 GPFS 的升级
  - “SCSI-3 永久保留”(Persistent Reservations)实现快速 I/O Fencing

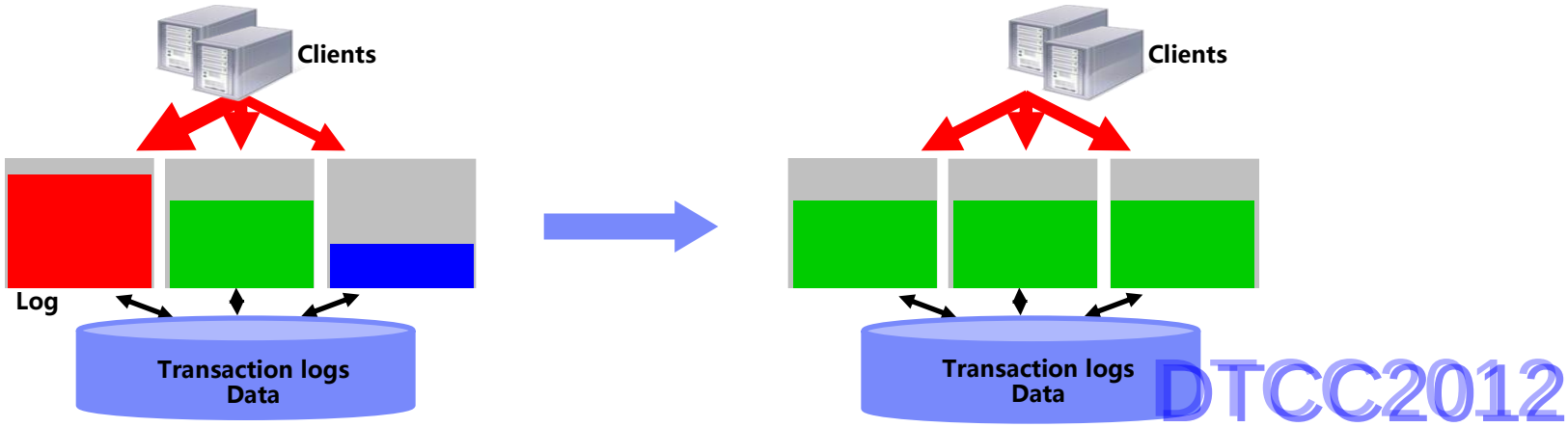
DB2 Cluster Services

- **协奏 (Orchestrate)**
  - 非计划事件提示保证无缝恢复和系统可用性
    - DB2 成员检测失败将自动地触发在本机或其它成员上的成员恢复
    - 主 PowerHA pureScale 的检测失败将会把辅 PowerHA pureScale 初始化为主 PowerHA pureScale
    - 主机的检测失败将初始化 I/O Fencing 并重启主机
  - 计划事件
    - ”秘密“（Stealth）维护
    - 增减成员和 PowerHA pureScale
- **整合**
  - TSA (Tivoli System Automation) 提供集群管理
    - 监控、失败检测，重启和恢复成员和 PowerHA pureScale
  - GPFS (General Parallel File System) 提供集群文件系统
  - 作为 Feature，TSA 和 GPFS 由 DB2 pureScale 提供介质，安装和配置

工作负载平衡和自动路由

工作负载平衡和自动路由

- **运行时负载信息用于负载平衡 (和 Z Sysplex 一样)**
  - 每个成员记载自己的工作负载
  - 回复给访问的客户端
  - 对下一个连接或者可选下一个交易进行路由
  - 路由对应用程序时透明的
- **Failover ： 失败成员上的工作负载平均地分布到其它存活的成员上**
  - 一旦失败的成员恢复，恢复的成员重新承担负载



客户端路由

- **工作负载平衡**
  - 服务器驱动
  - 基于成员的工作负载
  - 路由发生在：
    - Connection Level 或 Transaction Boundary
  - 平衡机制
    - 连接路由到工作负载最低的成员
- **也可以采用 Affinity 方式驱动**
  - 客户端驱动
  - 不考虑工作负载
  - 2种方式
    - 指定的
    - Round robin

Information Management

IBM

Data Management

DB2 pureScale 技术特色

DTCC2012

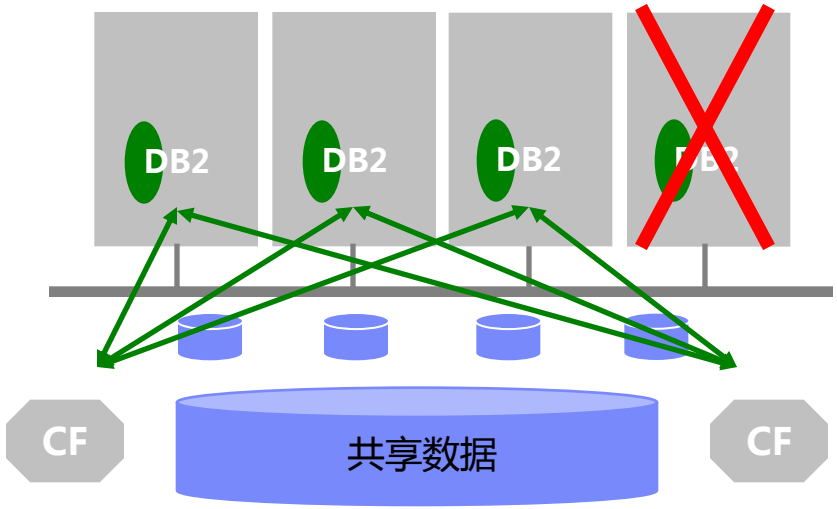
DB2 pureScale 技术特色和竞争分析

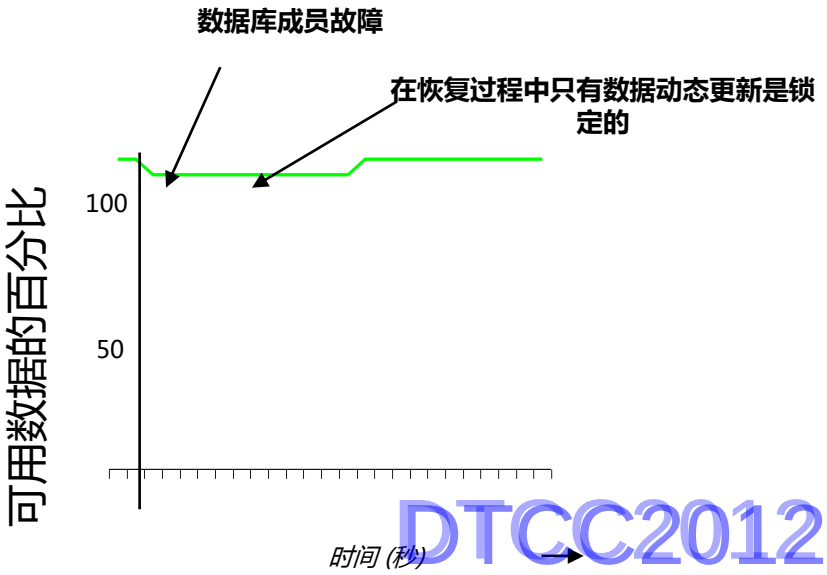
- **高可用性**
- **可用性竞争优势比较**
- **透明的应用可伸缩性**
- **可伸缩性竞争优势比较**

高可用性

在线恢复

- **DB2 pureScale 的设计初衷是最大限度地提高在故障恢复处理中的可用性**
- **当数据库成员出现故障时，只有*动态*数据仍然保持为锁定，直到成员恢复完成**
  - 动态 = 成员出现故障时正在更新的数据
- **行可用性的目标时间**
  - <20 秒





DB2 pureScale 成员故障中所涉及的步骤

- **故障检测**
- **恢复流程直接从 CF 读取：**
  - 需要修复的页面
  - 开始恢复的日志文件的位置
- **重新启动 Light Instance，重做和撤销恢复**

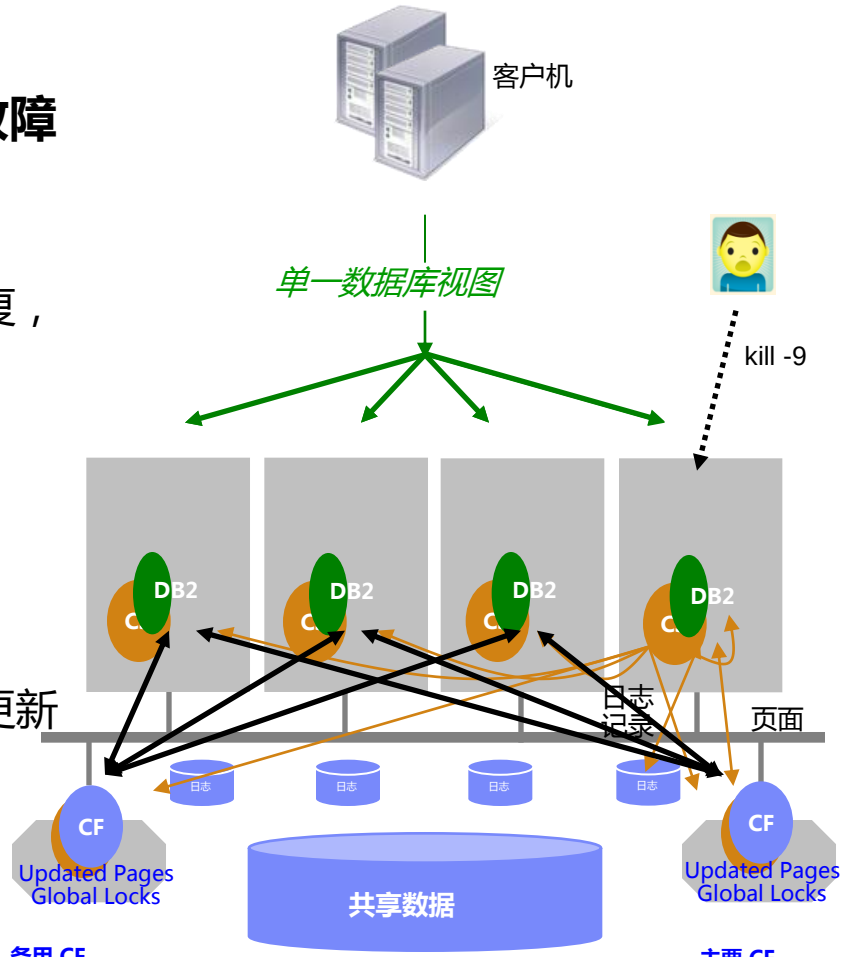
针对成员的故障检测

- **DB2 通过监视程序流程来监视自身的软件故障**
  - 当 DB2 成员出现故障时，监视程序会接收到消息
  - 监视器会中断集群管理器，通知它开始恢复
  - 软件故障检测时间不超过 1 秒
- **DB2 集群管理器在极低层次上以亚秒级速度运行（对资源利用的影响可以忽略）**
  - DB2 集群管理器执行其他检查来确定是否出现阻塞或故障
  - 其结果是，硬件故障检测可在 3 秒内完成，且没有错误的故障转移

成员故障概述

在幻灯片播放模式下运行

- **成员故障**
- **DB2 Cluster Services 自动检测成员的故障**
  - 通知其他成员和 CF
  - 在相同或远程主机上自动重新启动成员
  - 成员的重新启动类似于单一系统中的崩溃恢复，但速度要快很多
    - 重做操作仅限于处理中的事务
    - 利用 CF 中的页面缓存
- **客户机将透明地重新连接到健康的成员**
- **其他成员将随时可用 – “在线故障恢复”**
  - CF 存放针对故障成员的更新锁
  - 其他成员可以继续读取和更新故障成员未因更新而锁定的数据
- **成员重新启动完成**
  - 锁释放，所有数据完全可用



可用性竞争优势比较

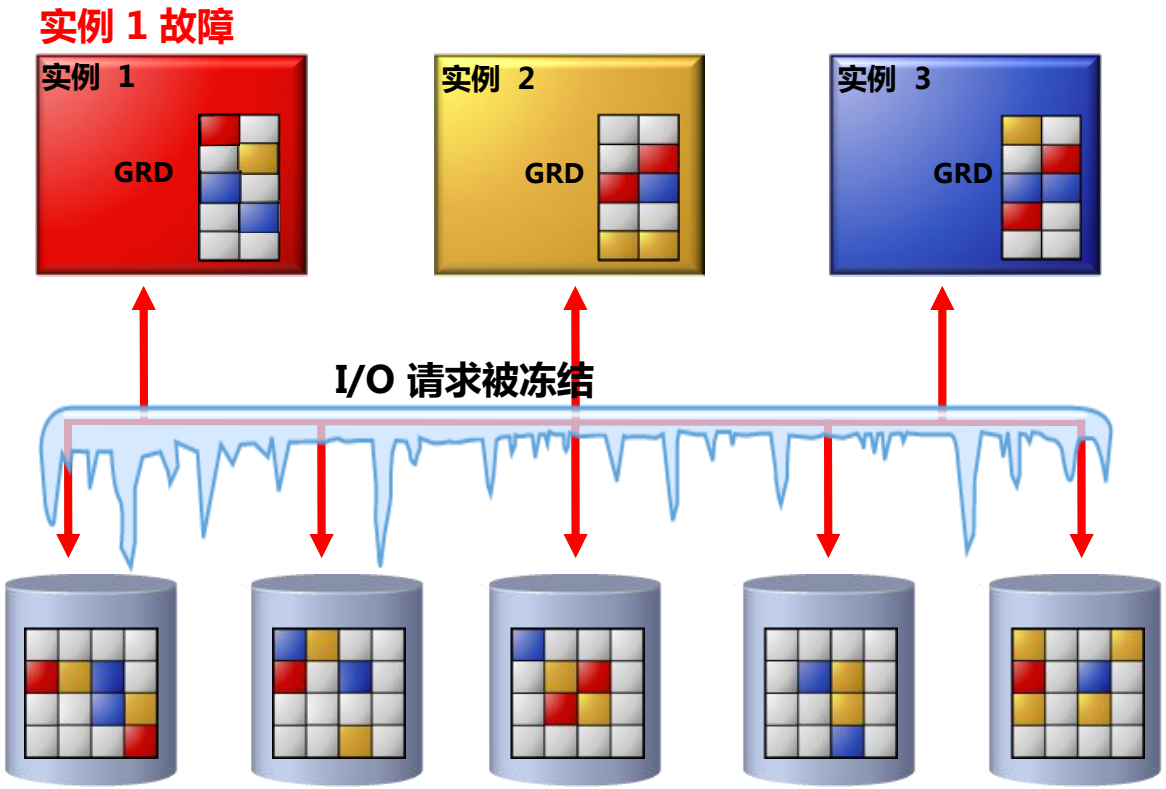
RAC 节点故障中所涉及的步骤

- **节点故障检测**
- **Remaster 数据块**
- **锁定需要恢复的页面**
- **重做和撤销恢复**

与 DB2 pureScale 不同，Oracle RAC 并不会集中化锁或数据缓存

通过 RAC – 访问 GRD 并冻结磁盘

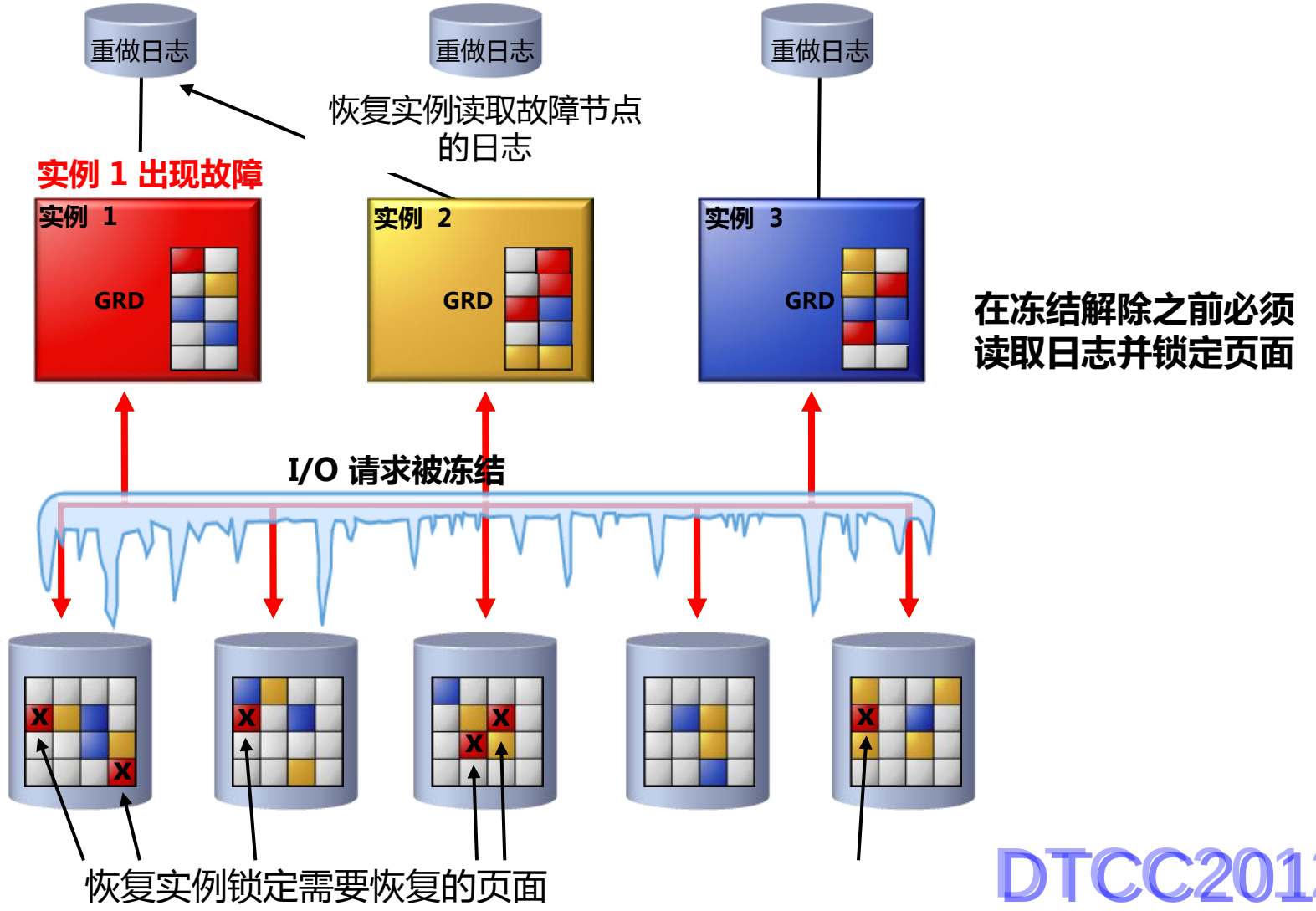
- **Global Resource Directory (GRD) 重新分配**



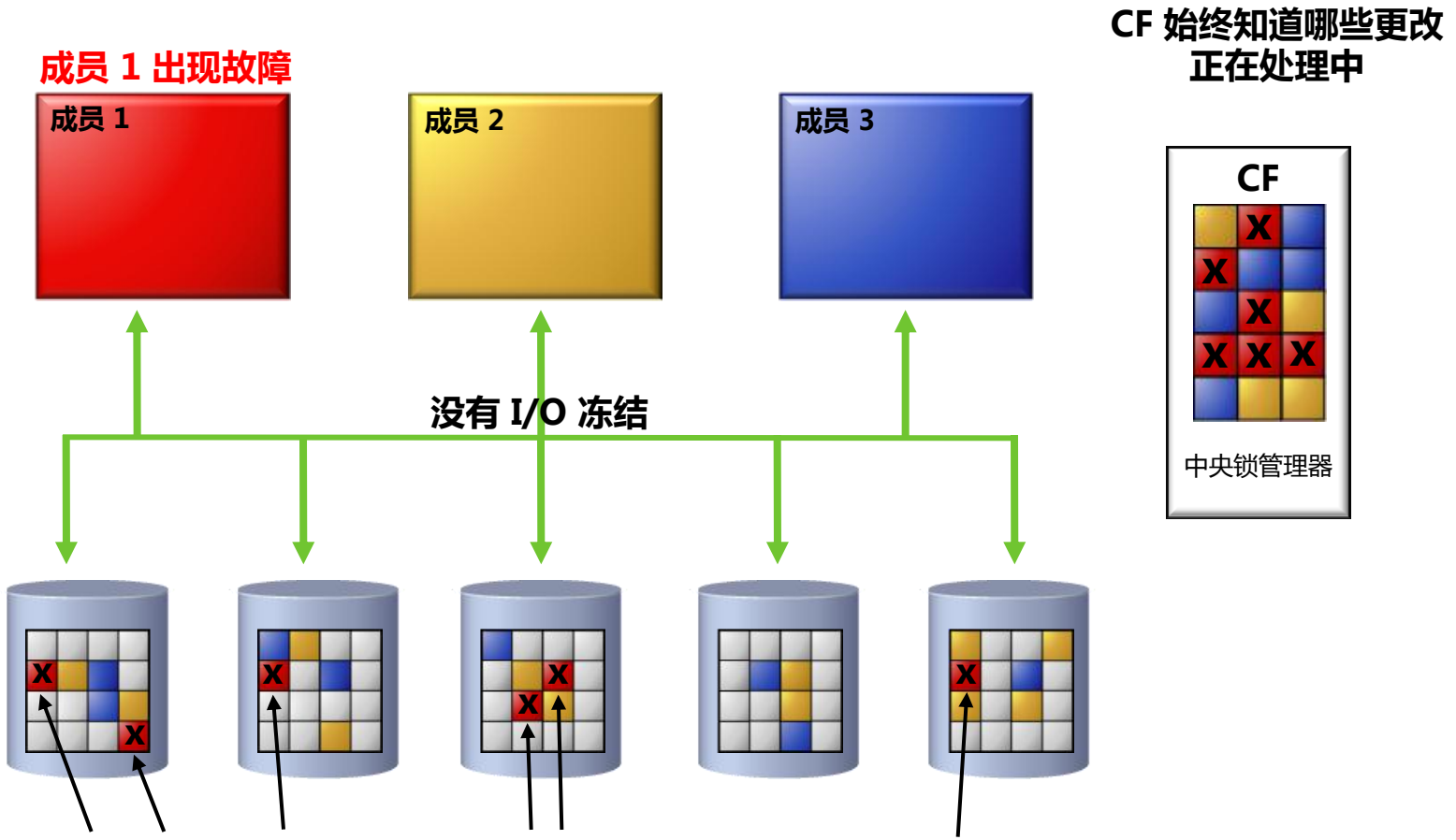
在需要恢复的页面被锁定之前没有更多 I/O

没有锁更新

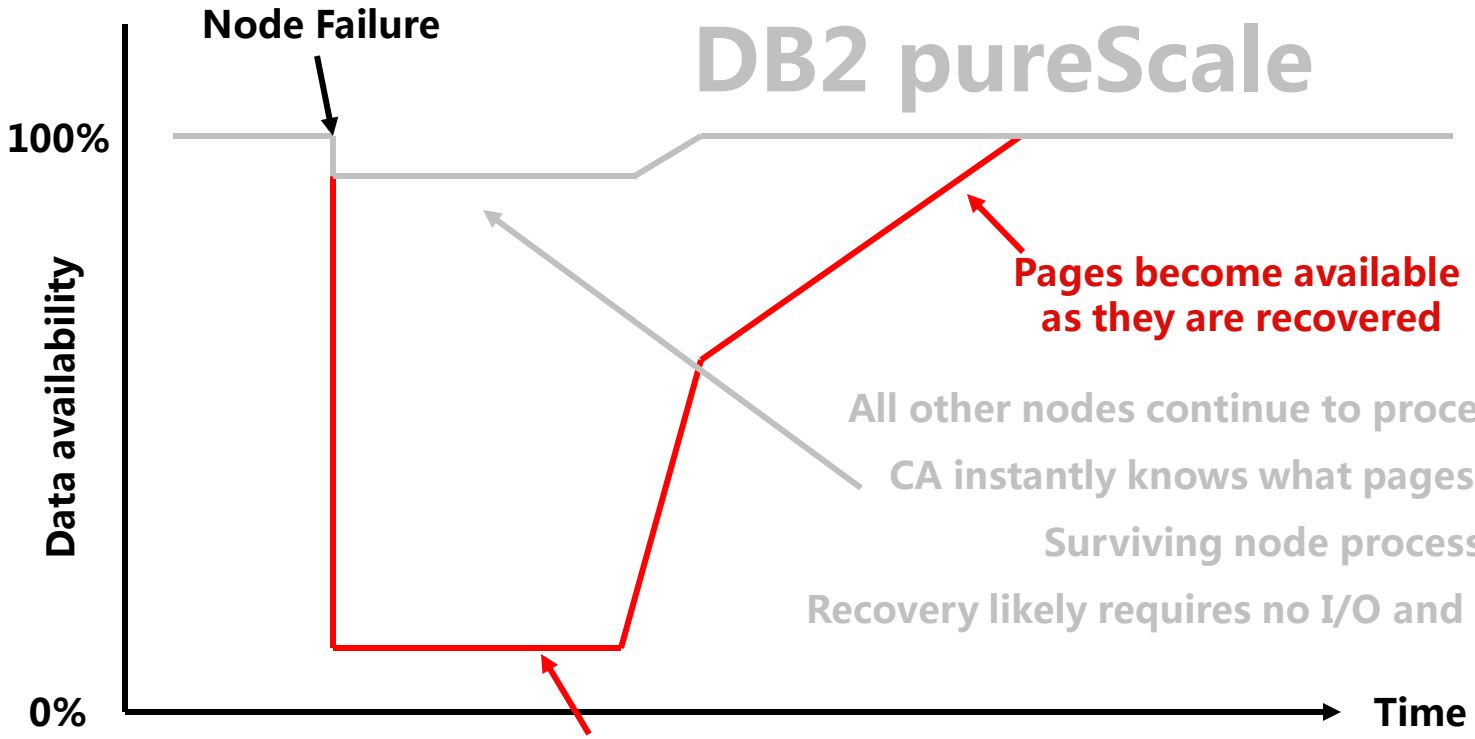
通过 RAC – 需要恢复的页面被锁定



DB2 pureScale – 完全没有冻结



崩溃恢复比较



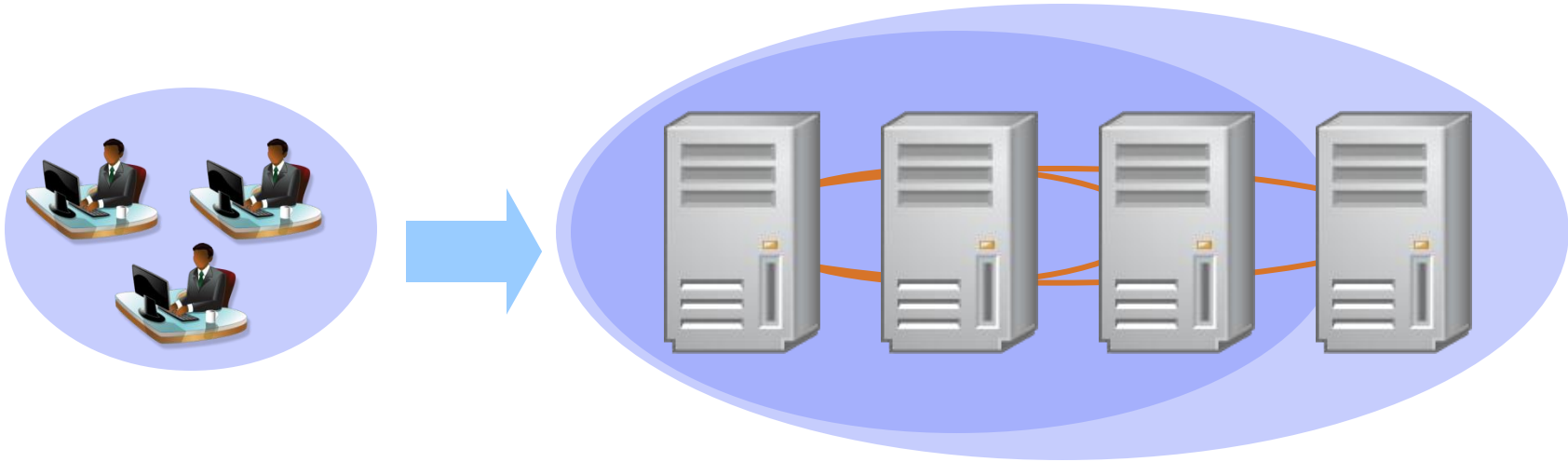
Oracle RAC

DTCC2012

透明的应用伸缩性

DB2 pureScale 的应用透明性

- **立即利用额外的产能**
  - 不需要修改您的应用代码
  - 不需要调优数据库基础设施



管理员可以增加产能，而不需要重新调优或重新测试

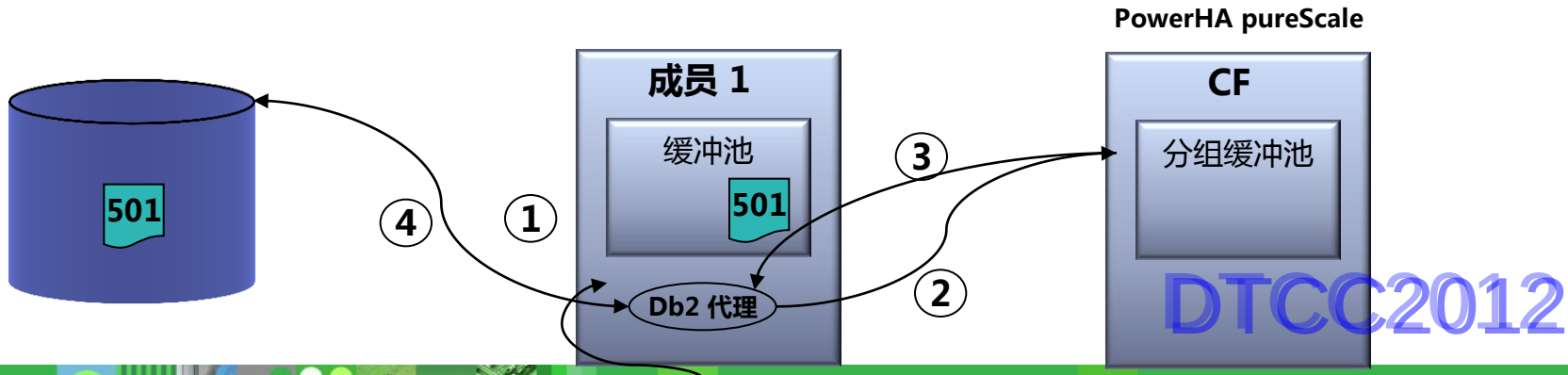
开发人员甚至不需要知道增加了更多节点

透明的应用可伸缩性

- **无需应用或数据库分区的可伸缩性**
  - 支持 RDMA 访问的集中锁定和全局缓冲池可以带来高可伸缩性，而不会让应用集群感知到
    - 数据页面的共享将在实际共享的缓存中通过 RDMA 来实现
      - 服务器之间的进程中断造成访问无法同步
    - 不需要通过应用或数据分区来实现可伸缩性
      - 降低了管理和应用开发成本
  - RAC 中的分布式锁定会增加开销并降低可伸缩性
    - Oracle RAC 最佳实践建议
      - 每个页面使用较少的行（避免热页面）
      - 通过数据库分区来避免热页面
      - 通过应用分区来获取一定水平的可伸缩性
      - 所有这些都会造成管理和开发成本增加

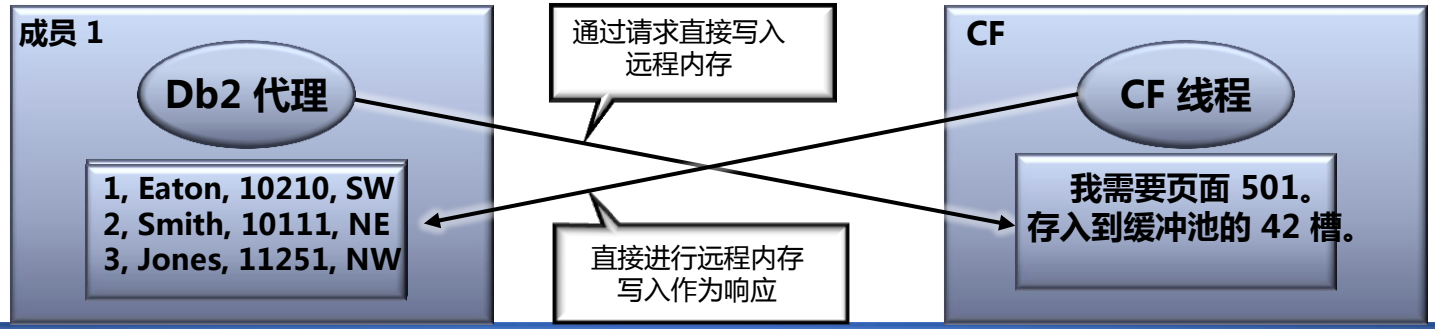
DB2 pureScale 在读取页面时发生了什么

- **成员 1 上的代理希望读取页面 501**
  - db2agent 检查本地缓冲池：未找到页面
  - db2agent 对 CF 内存执行 Read And Register (RaR) RDMA 调用
    - 没有上下文切换、没有内核调用。
    - 同步请求 CF
  - CF 回复它没有页面（再次通过 RDMA）
  - db2agent 从磁盘读取页面



DB2 通过 RDMA 执行读取和注册的优势

- **成员 1 上的 DB2 代理点程序直接向 CF 内存写入：**
  - 希望读取的页面号
  - 希望存入页面的缓冲池槽位
- **CF 的响应方式是向成员 1 上的内存直接写入：**
  - 它没有请求的页面 或者
  - 请求的数据页面
- **RAR 的总端到端时间将只有数微秒**
- **调度速度非常快，代理甚至可能会停留在 CPU 上以便进行响应**



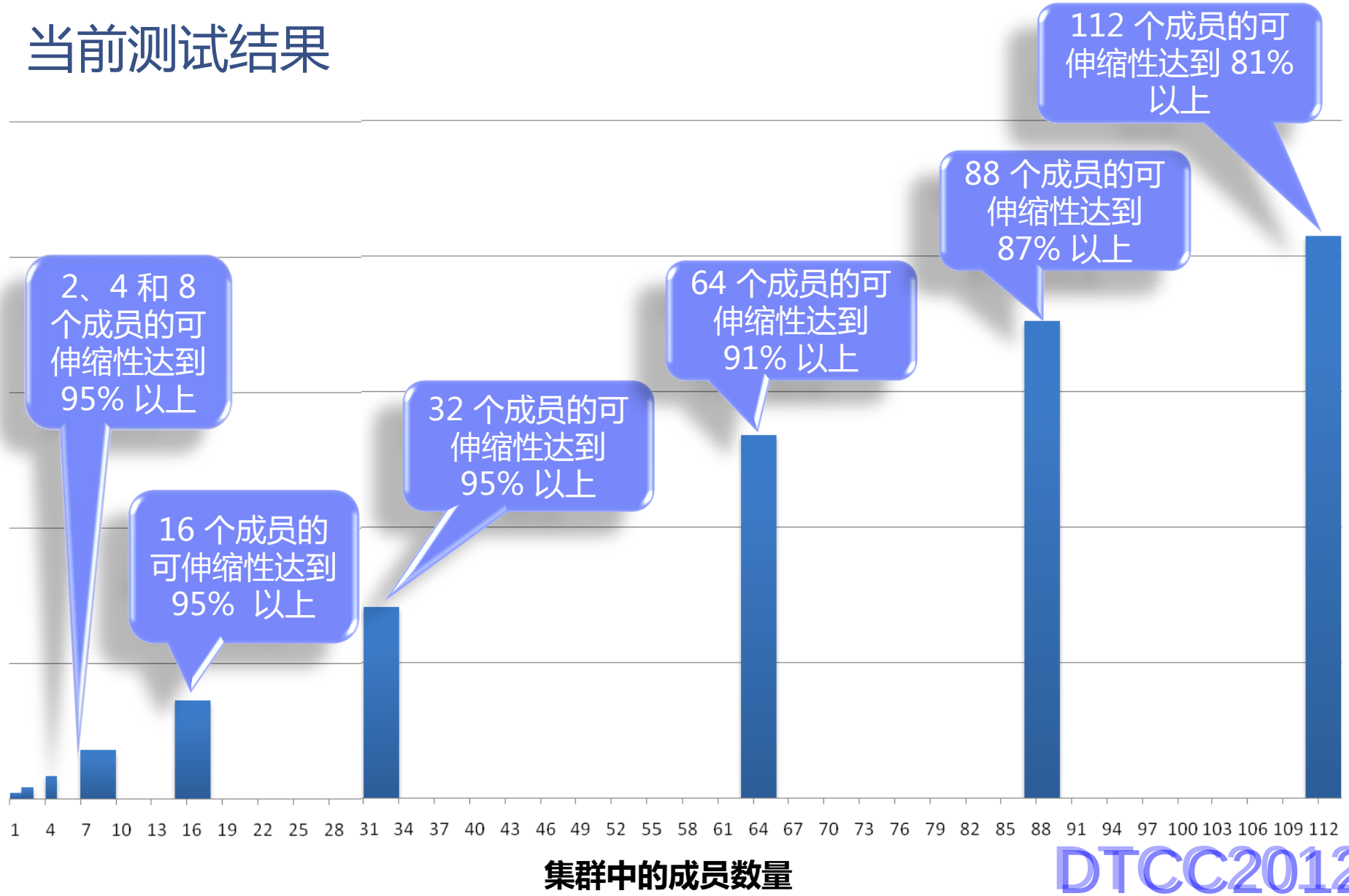
可伸缩性更高，数据不需要具备本地性

PS：我们的 InfiniBand 实现与 Oracle 有何差异?

- **DB2 利用 Remote Direct Memory Access (RDMA) 直接对远程服务器的内存执行写操作**

- **可以将 InfiniBand 应用于 Oracle RAC 集群**
  - 但是 Oracle 没有利用 RDMA
  - 这意味着 Oracle 通信将使用速度较慢的套接字协议，并且需要一些开销较大的 CPU 中断，从而影响集群效率

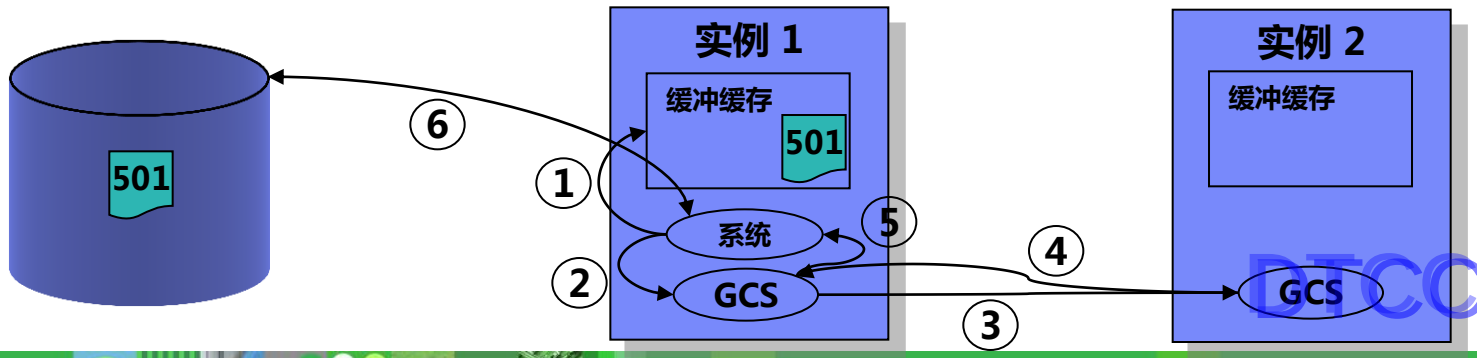
当前测试结果



可伸缩性竞争优势比较

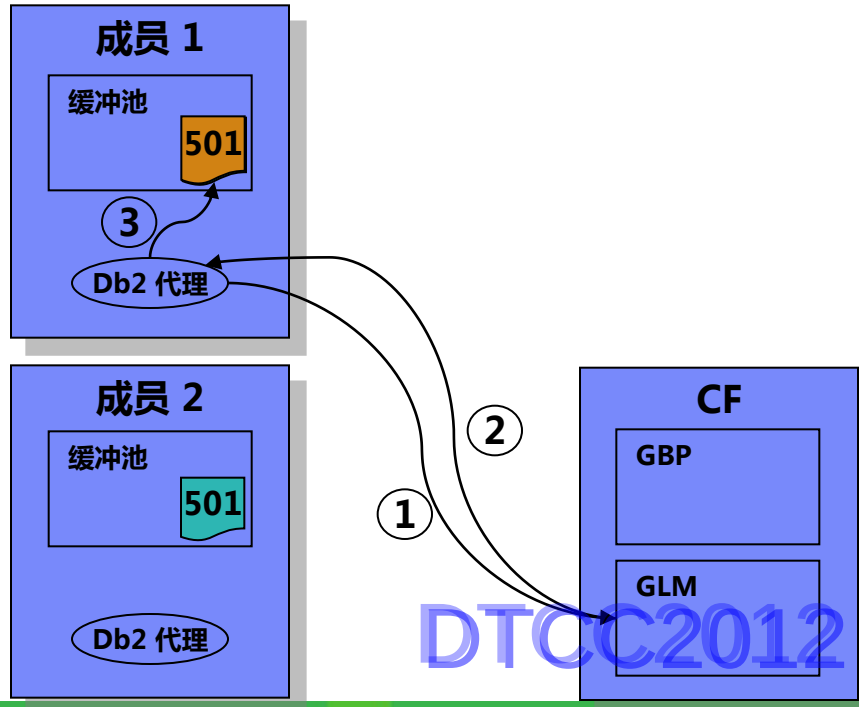
Oracle RAC – 单一实例希望读取页面

- **实例 1 上的进程希望读取实例 2 管理的页面 501**
  - 系统进程检查本地缓冲池：未找到页面
  - 系统进程向 Global Cache Service 进程发送 IPC，以获取页面 501
    - 通过上下文切换在 CPU 上调度 GCS
    - GCS 将请求复制到内核内存，以发起 TCP/IP 栈调用
  - GCS 向实例 2 发送请求
    - IP 接收调用需要中断远程节点上的处理操作
  - 远程节点通过 IP 回应，中断实例 1 上的 GCS
  - GCS 向系统进程发送 IPC（再一次上下文切换到进程请求）
  - 系统进程执行 I/O 以获取页面



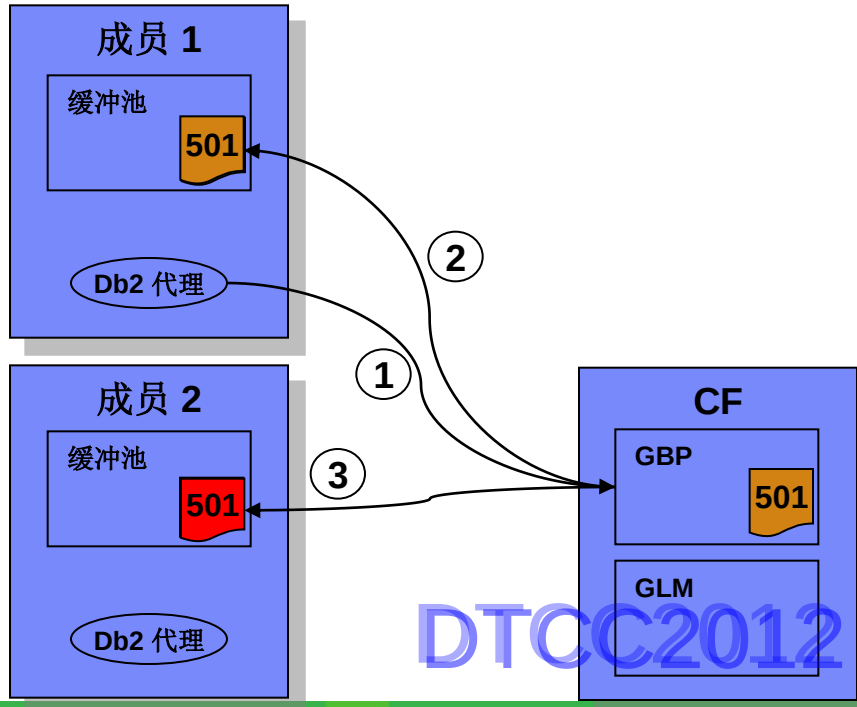
DB2 pureScale – 成员 1 更新一行

- **代理对 CF 发起 Set Lock State (SLS) RDMA 调用，对行设置 X-lock 和 P-lock，以指示要更新的页面**
  - 防止其他成员与此成员在相同时间对页面执行字节更改
  - SLS 调用完成所需的时间只有 15 微秒
- **CF 通过 RDMA 批准锁请求**
- **更新页面**
- **此时，成员 1 不需要执行任何其他操作**
  - P-lock 仅采用松散方式释放
  - 如果另一个成员需要它，则锁再提交
    - 之前都由成员 1 保存



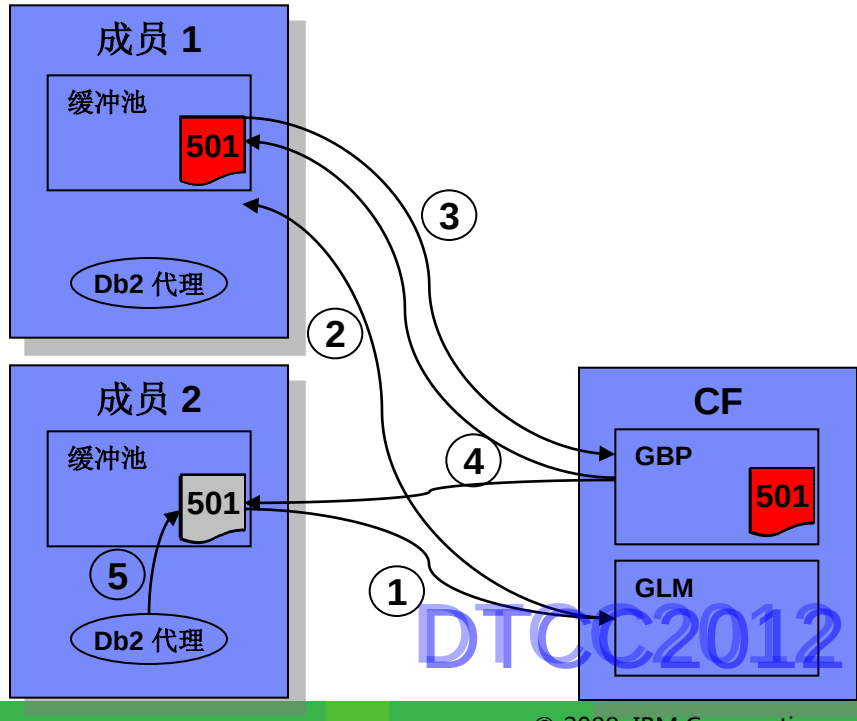
DB2 pureScale – 成员 1 提交更新

- **代理通过 Write And Register Multiple (WARM) RDMA 调用向 CF 请求已经更新的页面**
- **CF 直接从成员 1 的内存地址将所有更新的页面提取到它的全局缓冲池中**
  - 如果 P-Lock 未被释放，它将被释放（行的 X-lock 也是如此）
- **通过直接更新其他成员缓冲池目录中的某数据位，CF 将读取了此页面的所有其他成员中的页面都设置为无效**
  - 在成员再次访问这个经过更改的页面之前，它必须从 Global Buffer Pool 获取最新副本



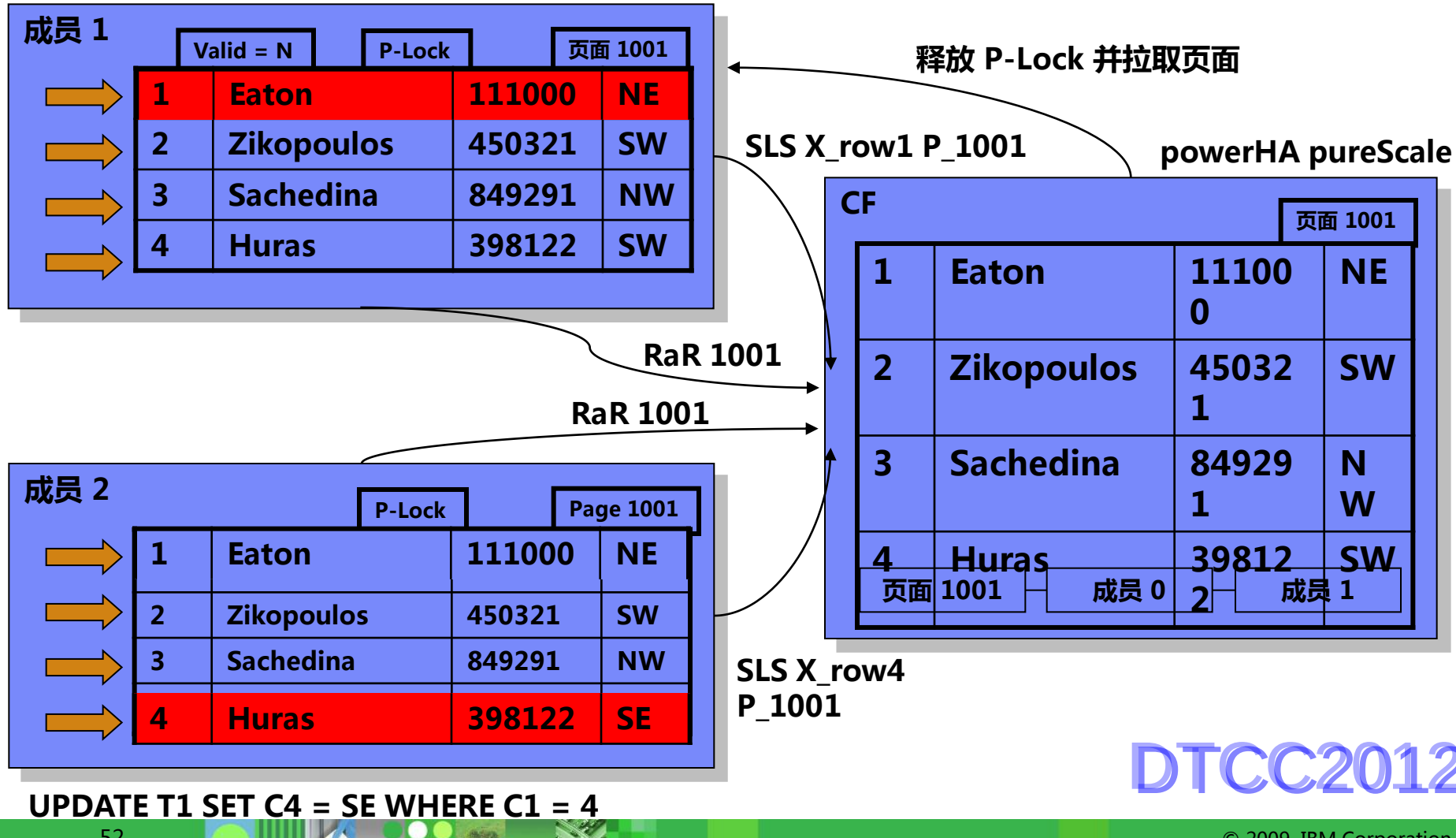
DB2 pureScale – 两个成员更新相同页面

- **成员 2 上的代理对 CF 发起 Set Lock State (SLS) RDMA 调用，对行设置 X-lock 和 P-lock，以指示要更新的页面**
  - P-Lock 争用由成员 1 持有的锁
- **GLM 通知成员 1 释放它的 P-lock**
- **成员 1 完成更新并通过 WARM 请求将页面从本地内存提取到 GBP 中**
  - 注意，事务边界不需要 P-lock
  - 仅在对页面执行字节更改的时间内有效
- **CF 通过 RDMA 批准锁请求，将更新的页面推入成员 2 的内存并将其他成员副本设置为无效**
- **更新页面**



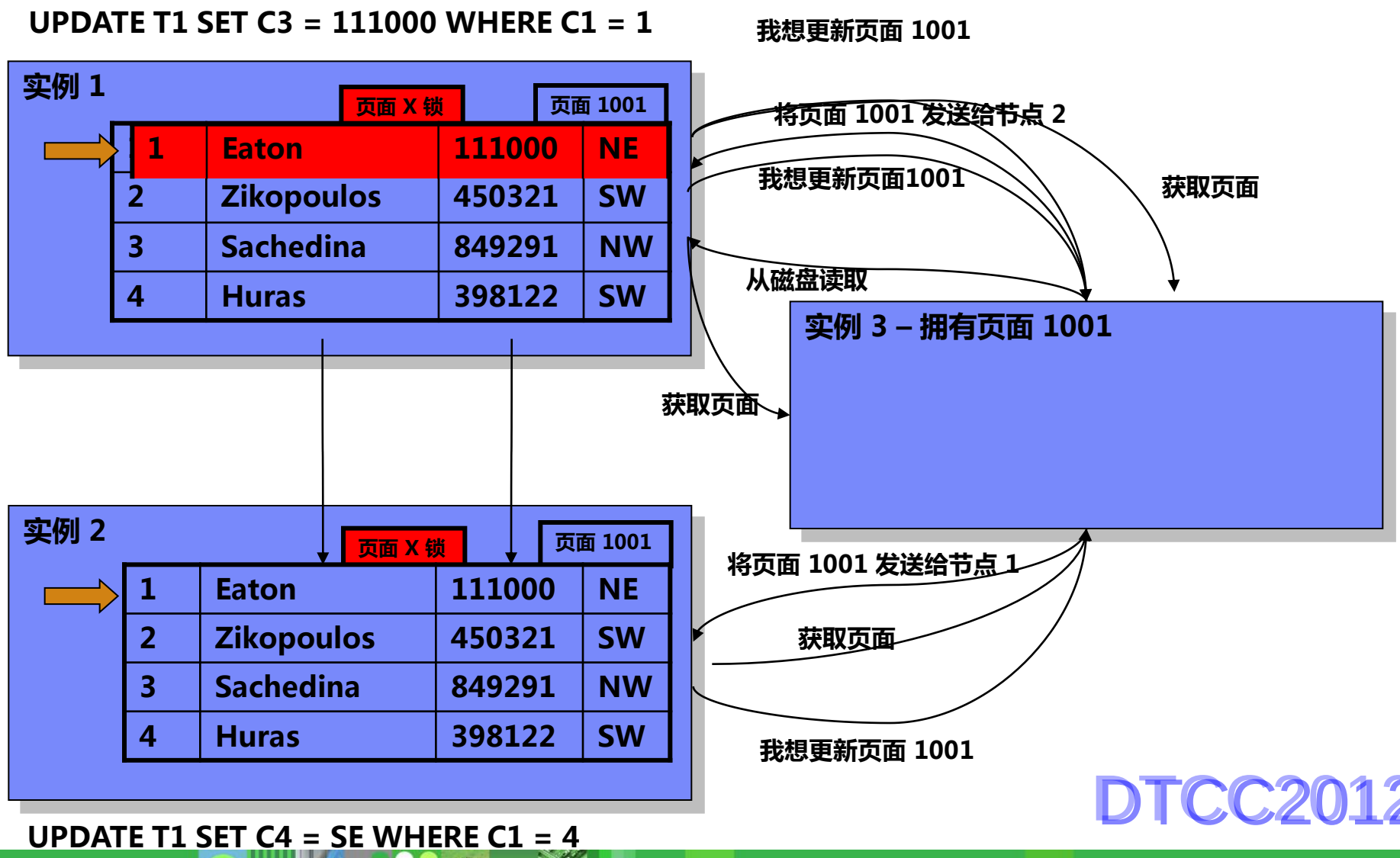
详细分析两个成员更新相同页面的情形（不同行）

UPDATE T1 SET C3 = 111000 WHERE C1 = 1

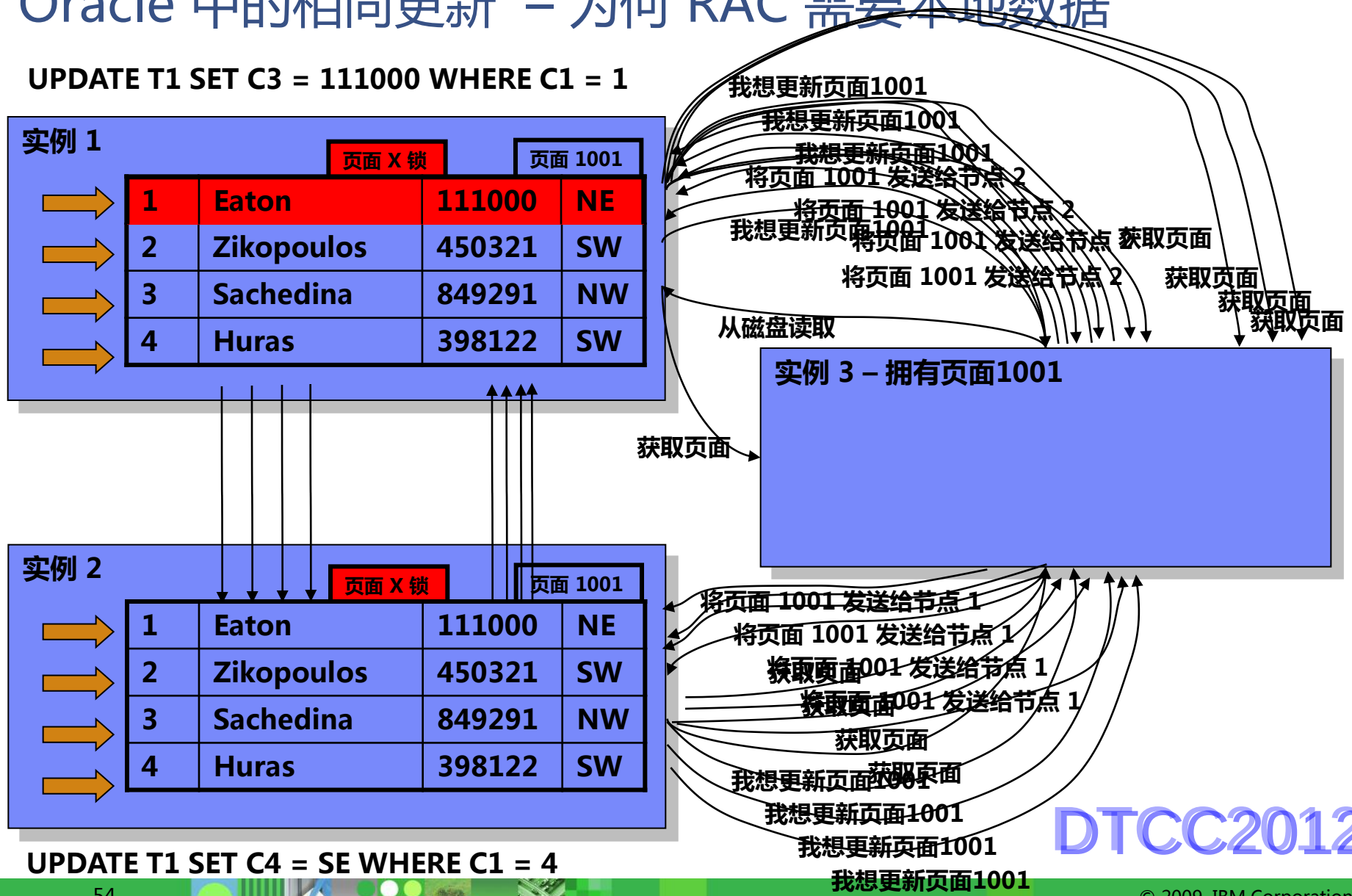


UPDATE T1 SET C4 = SE WHERE C1 = 4

Oracle 中的相同更新 – 为何 RAC 需要本地数据



Oracle 中的相同更新 – 为何 RAC 需要本地数据



IBM

Data Management

谢谢！

DTCC2012