eXtremeDB内存数据库性能提升方案分享

美国麦科捷有限公司 高级工程师

黄东旭

Agenda

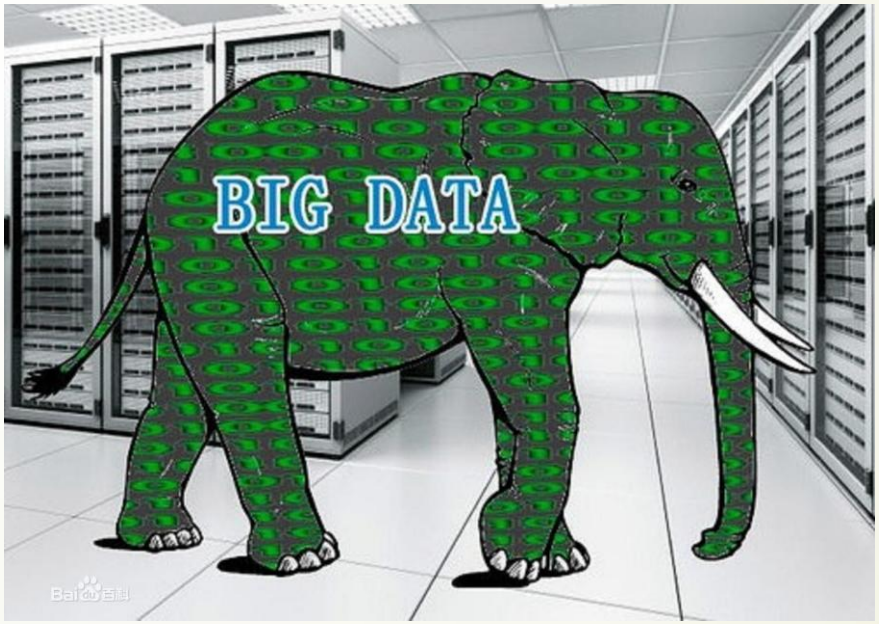
- 内存数据库的革命
- 为什么要使用内存数据库
- 什么是内存数据库
- 内存数据库特点
- 内存数据库性能提升方案
- 内存数据库扩展功能
- 内存数据库应用

内存数据库的革命？

- 内存数据库的历史
  - 文件系统
  - 嵌入式
  - 缓存
- 什么革命？流行？
  - 纯内存数据库
  - 商业化 – 市场需求
  - 从嵌入式到服务器， 交易处理到缓存数据处理
  - 行业特性的内存数据库
  - 磁盘数据库的功能实现在内存数据库 – 日志，事务，可靠性，集群等
- 行业趋势？
  - Oracle, IBM 数据库老大们争相推出内存数据库产品
  - 成熟商业内存数据库产品

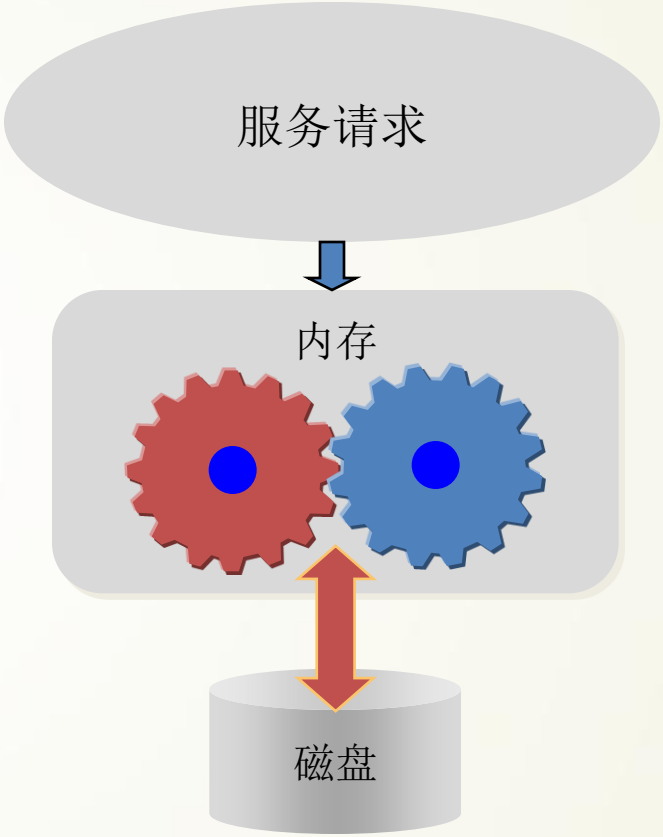
为什么要使用内存数据库?

- 系统变得越来越大，使用的软件越来越多
  - 系统复杂度
  - 数据结构，类型
- 大数据时代来临，数据量爆发增长
  - 海量数据
  - 高性能
- 磁盘数据库性能瓶颈
  - 性能
- 硬件成本降低
- 内存数据的可靠性技术发展
  - 日志
  - 热备
  - 集群



什么是内存数据库-1

- 内存为主要存储介质
  - 所有的表及索引总是在内存中
- 提供最快的访问速度
  - 为访问内存而设计的最佳访问方法和索引模式，并在数据缓存、快速算法、并行操作也进行了相应的改进
  - 消除了 I/O 瓶颈



什么是内存数据库-2

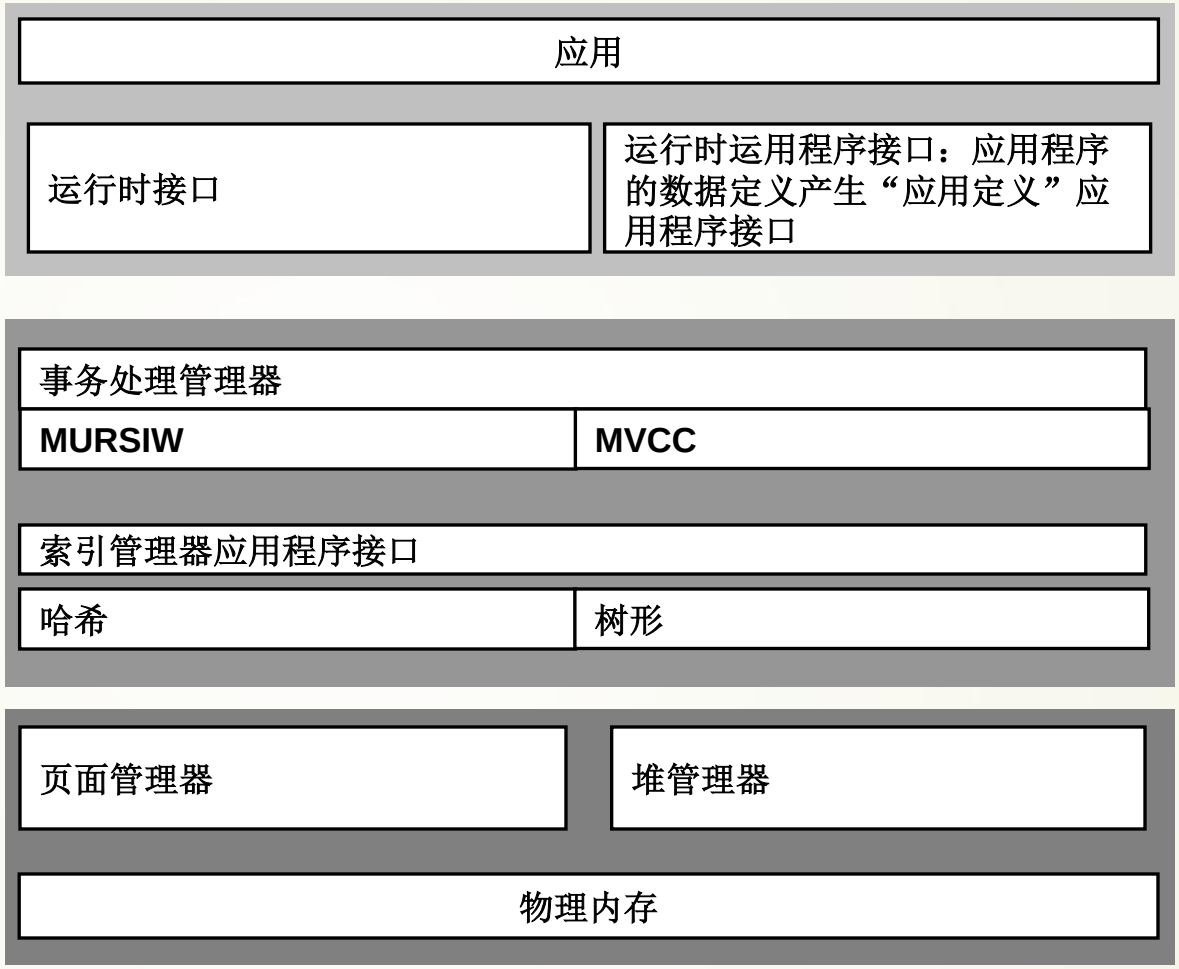
- 比较传统磁盘数据库
  - 访问成本
  - 容量小
  - 数据持久性
  - 索引算法和内存管理算法
  - 系统架构中的角色
- 比较自己开发模块
  - 可移植性
  - 完整解决方案
  - 接口统一性

内存数据库-特点

- 快：追求绝对的性能
  - 速度是内存数据库最大的优势
  - 单笔绝对消耗时间或吞吐率是评测的指标
- 小：占用系统资源
  - 内存稀缺资源，高效利用内存
  - 部署方式？与磁盘数据库的关系
- 易失性：
  - 内存本身掉电丢失的特性
  - 需要提前对内存上的数据做保护

内存数据库-性能提升方案

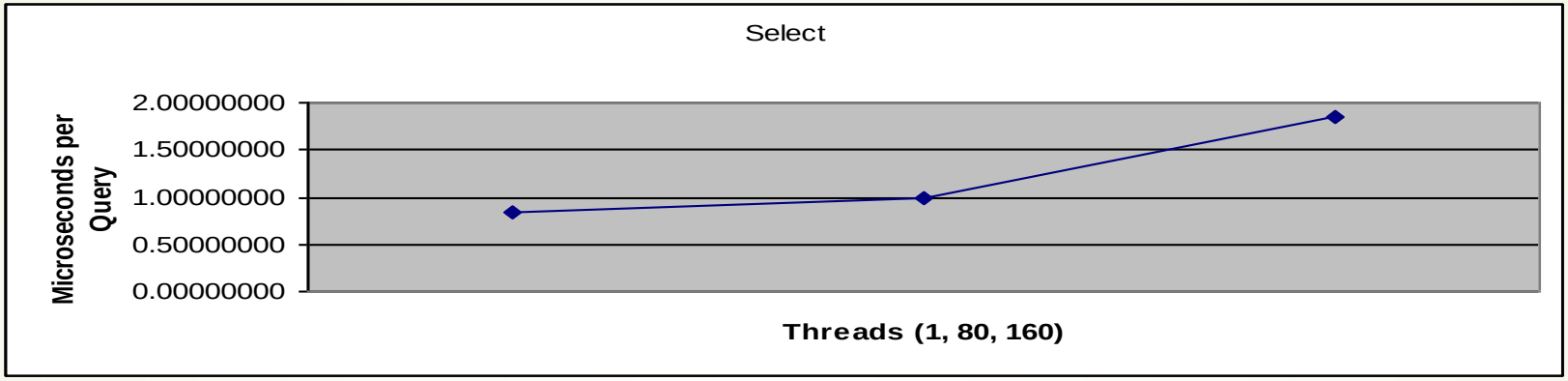
- 如何提升内存数据管理性能？
- 运行时接口
- 事务处理管理器
- 索引管理器
- 内存管理



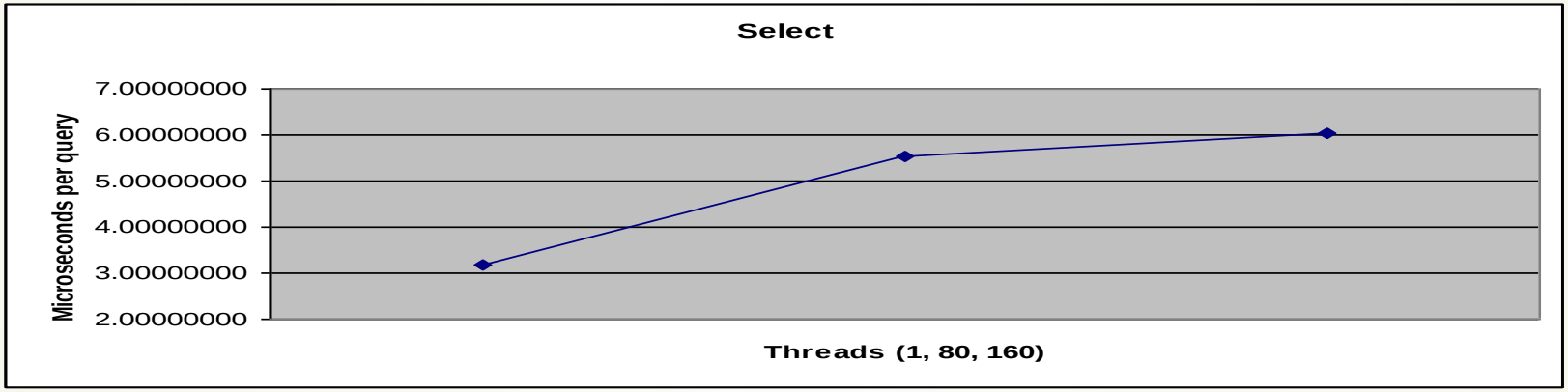
内存数据库-运行时接口1

- SQL/JDBC/ODBC
  - 标准化
  - 简单易用
  - 高度非过程化
- 本地API
  - 执行路径短
  - 专有接口
  - 最大的灵活性
- 其他接口
  - JNI，C#
- 混合接口

内存数据库-运行时接口2



eXtremeDB 本地API接口 查询性能

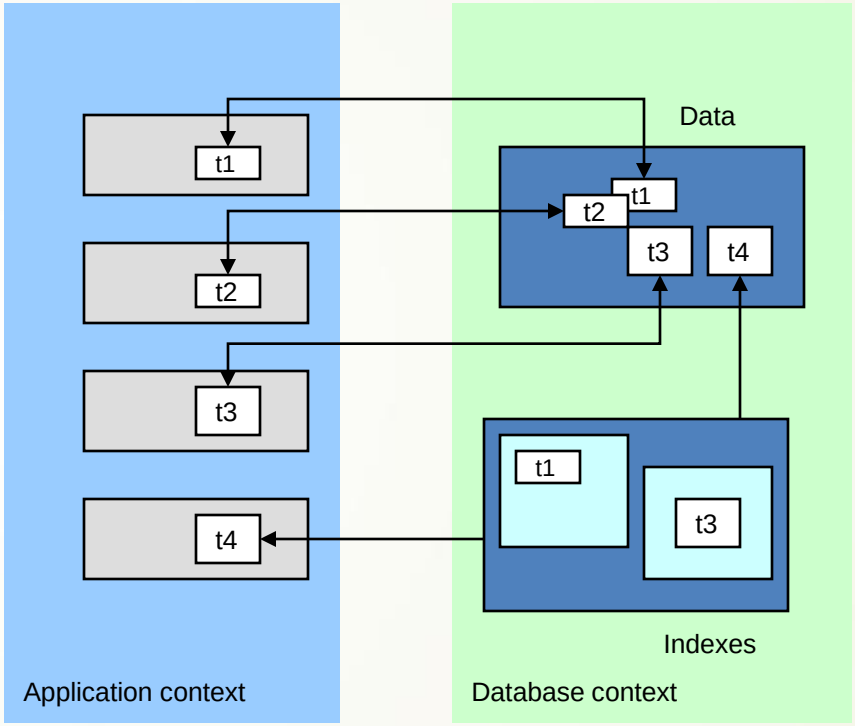


eXtremeDB SQL接口 查询性能

测试报告下载地址：http://www.mcobject.com/terabytebenchmark.pdf

内存数据库-多任务处理

- MURSIW(多读单写）
- MVCC（多版本并发控制）
  - 增强数据库的并发管理
  - 允许同时的读写事务
  - 提高多核CPU利用效率
  - 事务隔离级别(RC，RR，SE)
- 轻量且高效的事务处理
  - 排它锁
  - 共享锁



内存数据库-索引优化

- Hash
  - 动态哈希表
- B-Tree
  - 基于磁盘特点优化
- T-tree
  - 基于内存特点优化
- R-tree
  - 二维索引（X/Y坐标系，经纬度）
- KD-tree
  - 多维索引
- Patricia trie
  - 前缀匹配算法
- 用户自定义索引
  - 用户自定义检索匹配算法

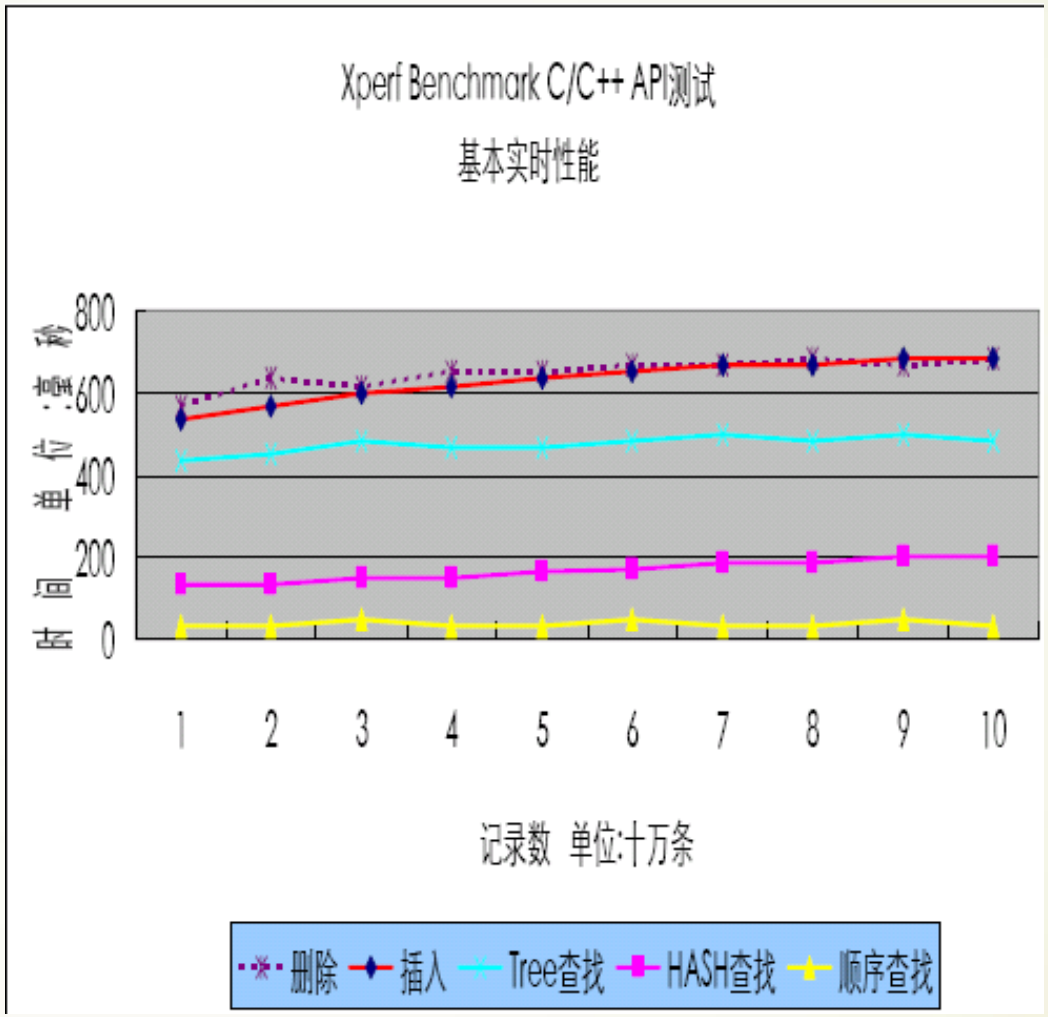
内存数据库-内存管理

- 静态分配OR动态分配？
  - 静态分配
    - 不依赖操作系统内存管理机制
    - 避免给操作系统造成大量内存碎片，影响系统稳定
  - 动态分配
    - 内存管理算法简单

- 页管理器
  - 逻辑设备管理机制
  - 设备统一分页管理
  - 数据对齐
- 降低内存占用

eXtremeDB-性能

- 数据库建立在主内存中，程序可以直接使用，数据库操作的速度以微秒计
- 静态内存分配及定制的API缩短了代码执行路径，既提高了性能，更提高了系统健壮性
- 嵌入进程的运行方式,剔除了进程间通信的开销
- 右图显示了eXtremeDB操作一条记录能达到惊人的个位微秒级性能



内存数据库-功能扩展

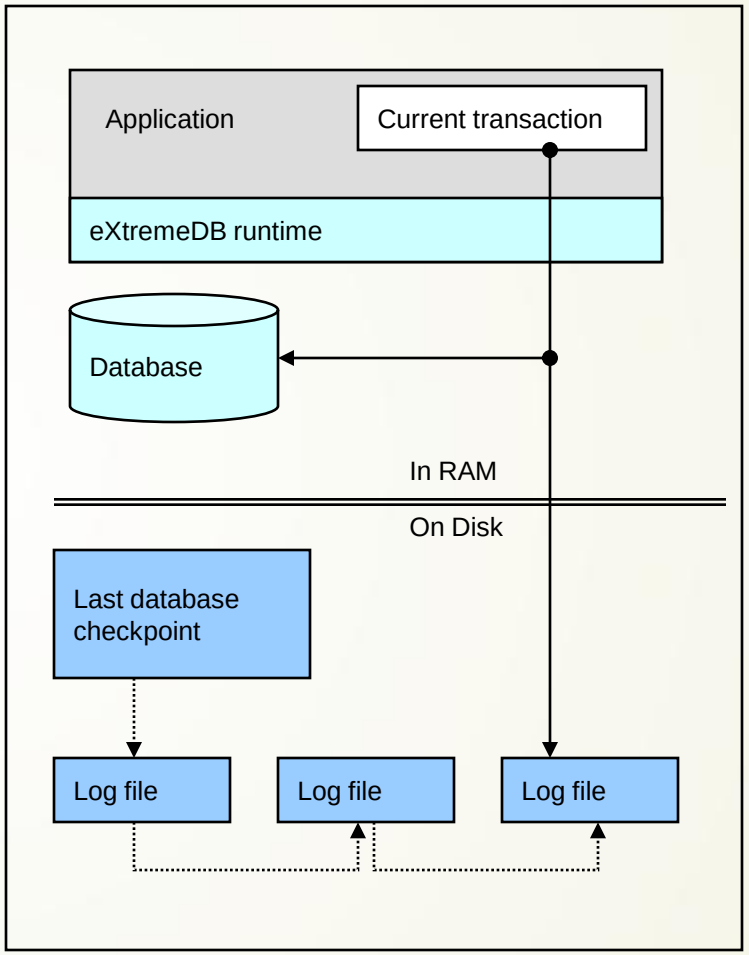
- 内存数据保护
- 事务日志
- 高可用版
- 集群
- 特殊版本

内存数据库-内存数据保护

- 备份/修复
  - 自定义备份介质（本地文件系统，网络设备等）
- 事务日志
  - 定期检查点
  - 前滚/回滚恢复
- 非易失性存储
  - NVRAM，有后备电池的静态存储器
- 高可用性组件和集群组件
  - 在不同设备中同步多个数据库实例
- 与磁盘数据库进行数据同步

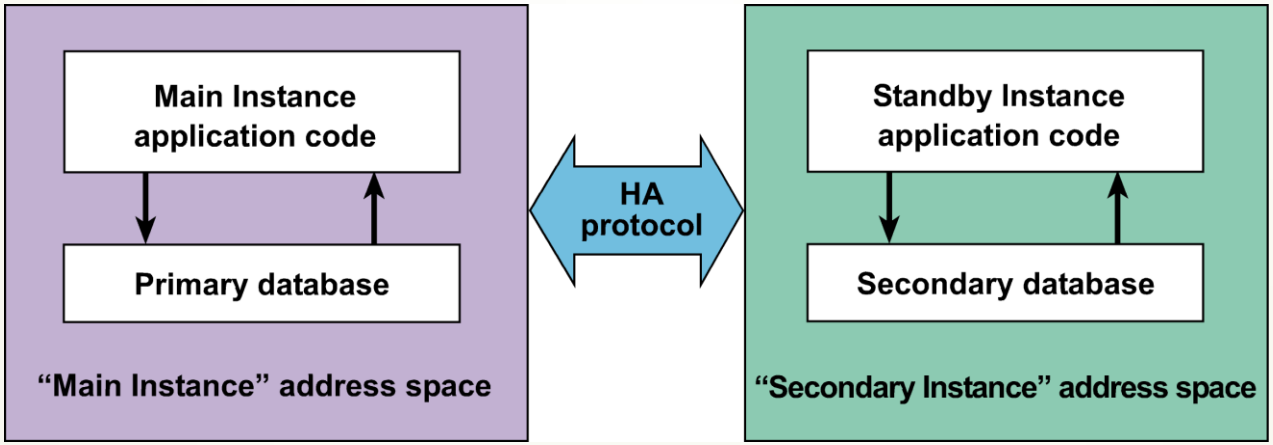
内存数据库-事务日志

- 以事务位单位记录事务日志
- 设备故障或系统出错的情况下，对于RAM中内存数据库实例提供恢复能力
- 前滚或回滚
- 缓存策略
- 事务日志存储介质：
  - 本地或网络文件系统
  - Socket 设备
  - Pipe
- 性能影响：
  - 查询操作不受事务日志的影响
  - 写入操作会比直接写入传统磁盘数据数据库要快得多
- 提供”Data Relay” 特性，与后端数据库进行无缝通讯



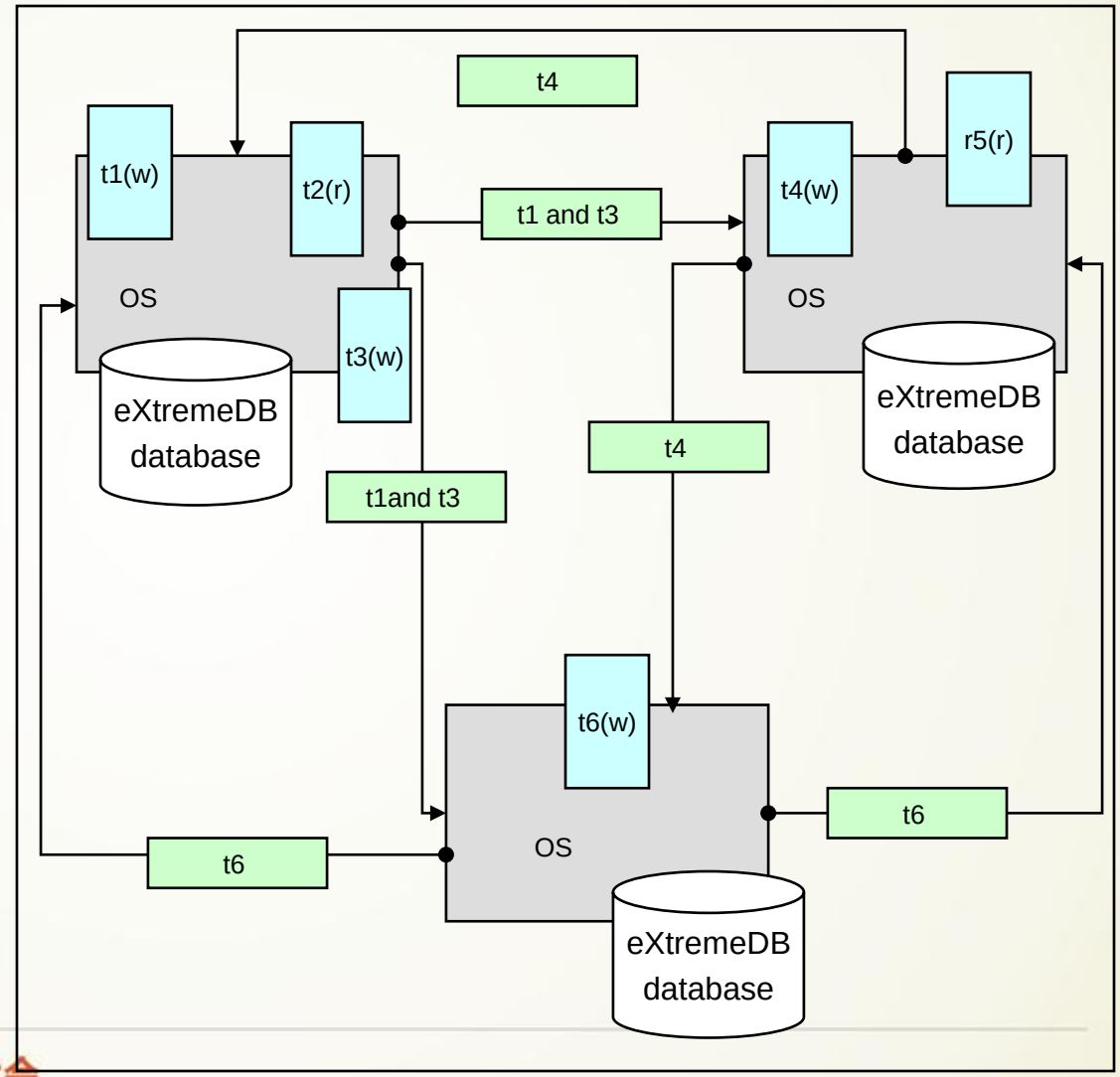
内存数据库-高可用特性

- 在不同节点，高可用功能同步多个独立的数据库实例，保证它们的一致性。较常见的有：
  - 在一个局域网内两个或两个以上的计算机
  - 在同一硬件环境个的多个进程或线程
- 数据传输方式
  - 同步
  - 异步



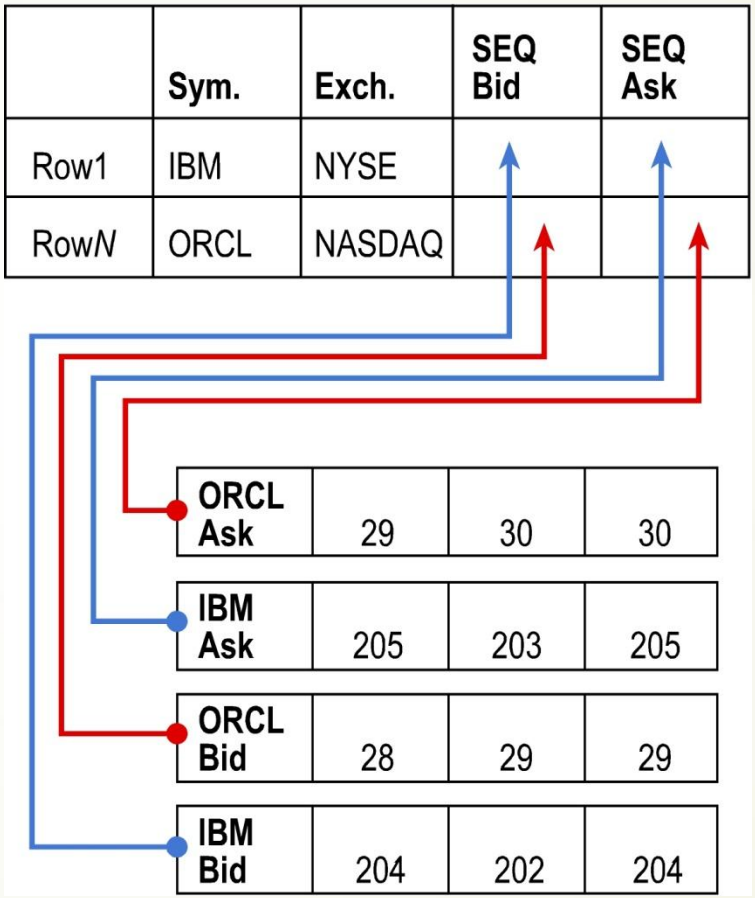
内存数据库-集群

- 数据库的事务分发到网络上的多个数据库结点。
- 节点间共享数据，数据一致性
- 两阶段提交策略
- 所有的结点都支持只读和读写的事务。
- 所有结点都拥有全库数据。
- 负载均衡算法

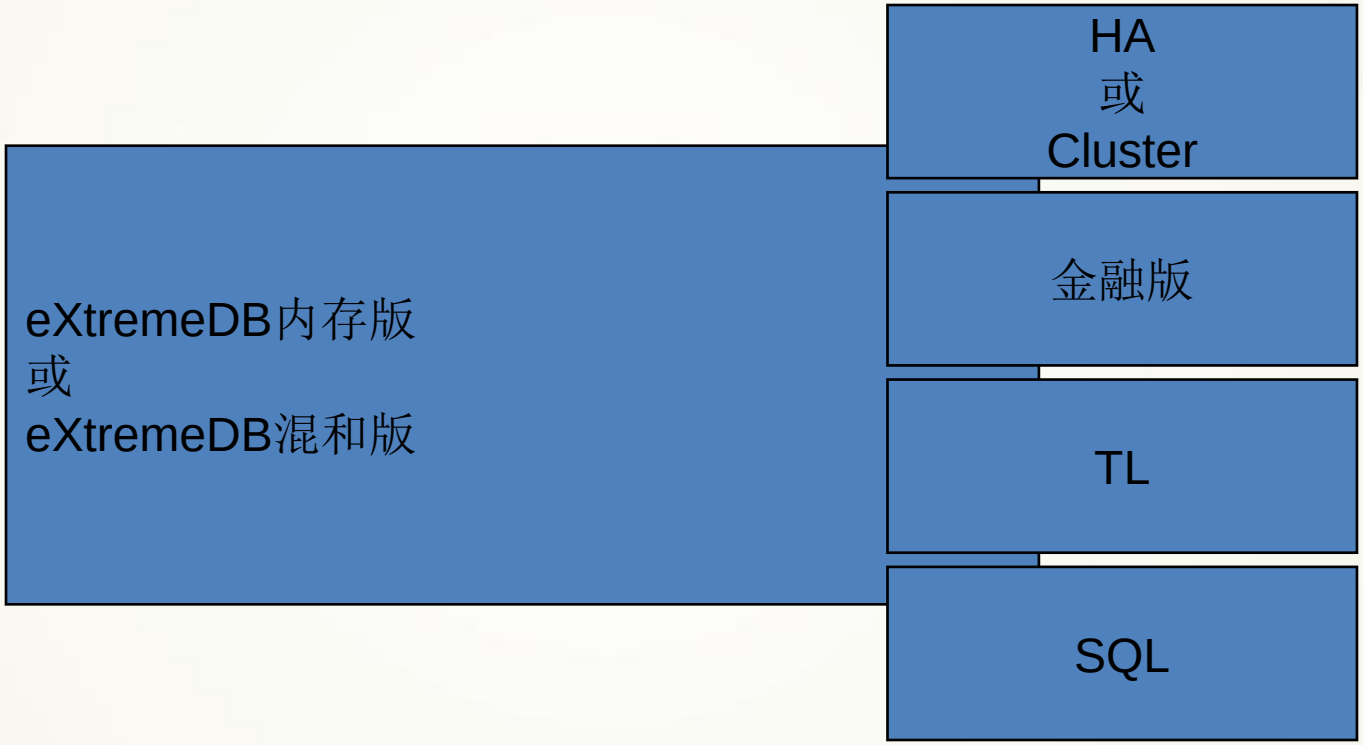


eXtremeDB-金融版

- 纵列数据布局
- 分笔数据流、历史交易数据和其他顺序数据
- 丰富的基于矢量的数学函数，通过最大限度地提高一级/二级缓存的使用率来加快对时间序列数据的管理。通过组织函数形成顺序的运算流水线来支持统计/定量分析。
- 集群技术“本地表”扩展
- 全新数据库监控界面：交易吞吐量、内存占用量和其他关键因素



eXtremeDB-产品结构



McObject公司

- McObjec成立于 2001年，总部位于美国华盛顿. 2008年正式成立中国办事处.
- 所有管理层和技术人员都来自于数据库和实时数据处理领域，公司创始者在实时数据库行业有20多年从业经验.
- McObject从一开始就创建了eXtremeDB内存数据库系统，现已拥有大量的成熟解决方案案例，拥有超过300个全球各行业领军客户.
  - 大连商品交易所，郑州期货交易所，烽火科技，京信，和黄电信，国电南瑞，中国华东电网，中船重工集团，中国电子科技集团，中国航天科工集团，广州思维奇，京信通信，三一重机，黑龙江福彩，香港赛马会，南车
  - NSE.IT, Boeing, EADS, Lockheed Martin, Northrop Grumman, Tyco, Motorola, Nokia Siemens Networks, SOMA Networks, NextPoint Networks, F5 Networks, Siemens, Lifetree, DadeBehring, Chrysler, JVC, SAIC, CA, Maximizer, Philips, Pioneer, Peiker Acousti

McObject客户

郑州商品交易所
Zhengzhou Commodity Exchange

大连商品交易所
DALIAN COMMODITY EXCHANGE

Comba 京信通信

ptSwitch

龙工风采

FiberHome
烽火通信@光通信专家

中国农业银行
AGRICULTURAL BANK OF CHINA

CSR
中国南车

国家电网
STATE GRID
中国电力科学研究院
CHINA ELECTRIC POWER RESEARCH INSTITUTE

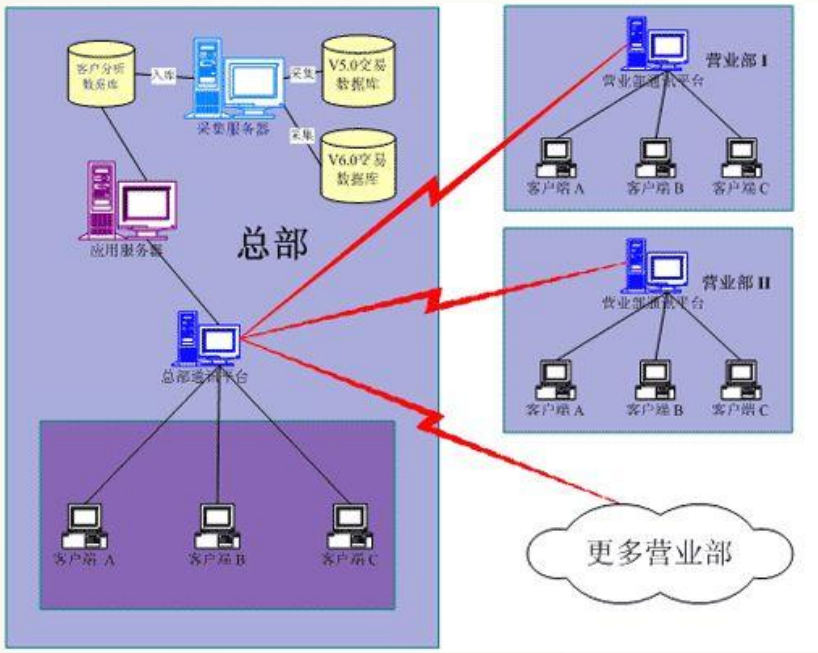
NARI
国电南瑞科技股份有限公司
NARI Technology Development Co.,Ltd.

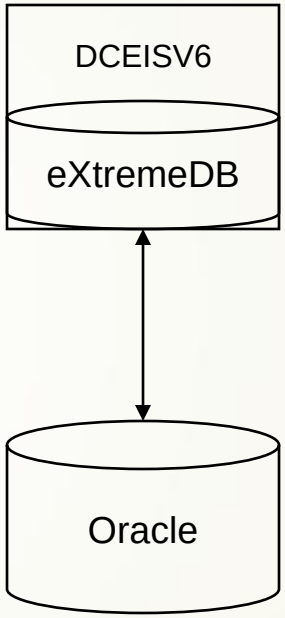
内存数据库-应用篇

- 金融领域
  - 实时交易结算系统，期货交易系统，交易中心软件
- 电信领域
  - 实时计费，实时会话
- 安全控制系统
  - 安全通讯设备，安防系统，智能交通
- 工业控制
  - 数据采集，工业监控设备
- 消费电子
  - 智能手机，机顶盒，路由器，游戏机
- 军事/航空航天
  - 武器装备，导航，火控，战地指挥
- 其他
  - 彩票，赛马
- **任何需要实时数据处理和高吞吐率的场景**

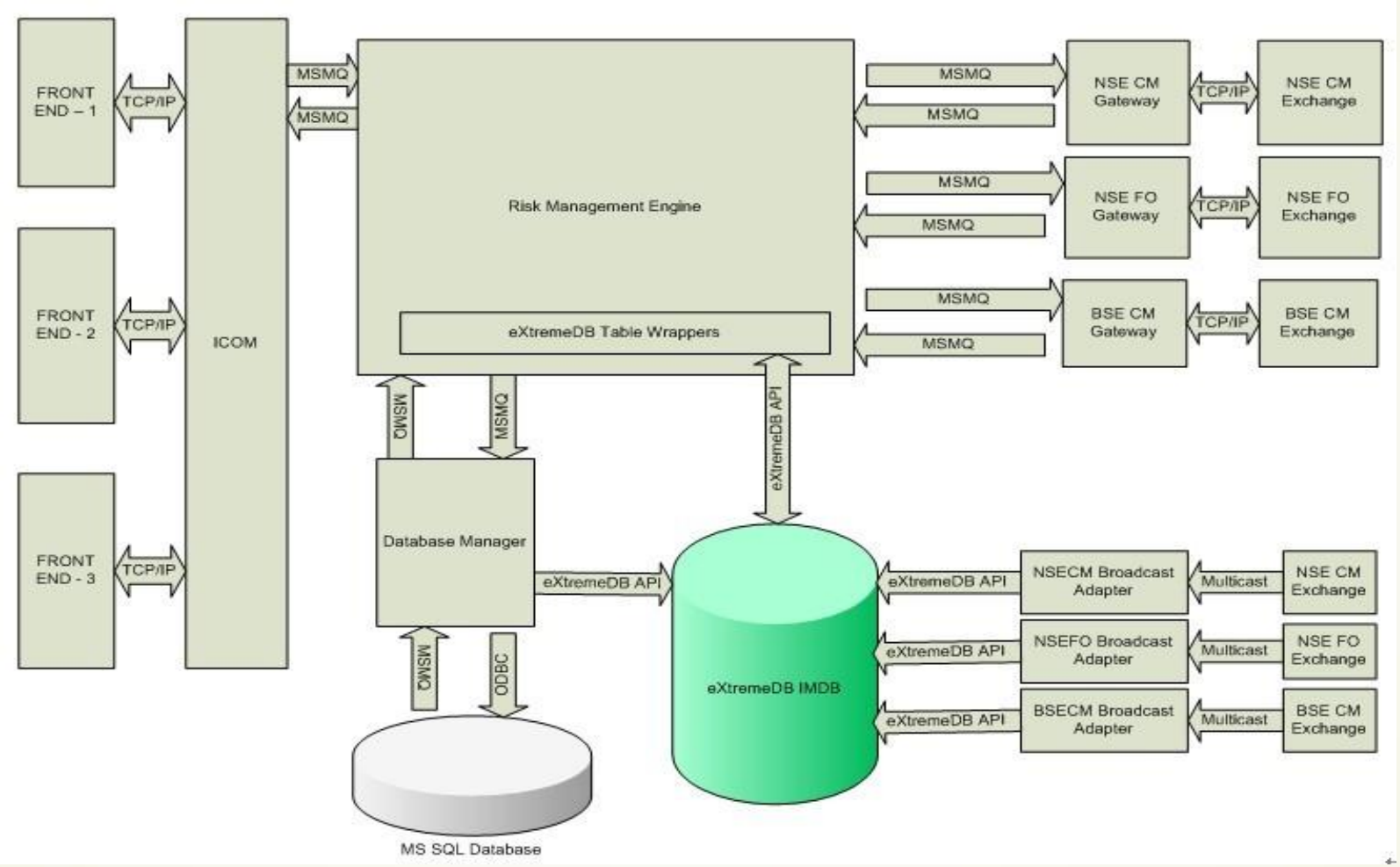
金融应用-DCE

- 大连商品交易所 DALIAN COMMODITY EXCHANGE



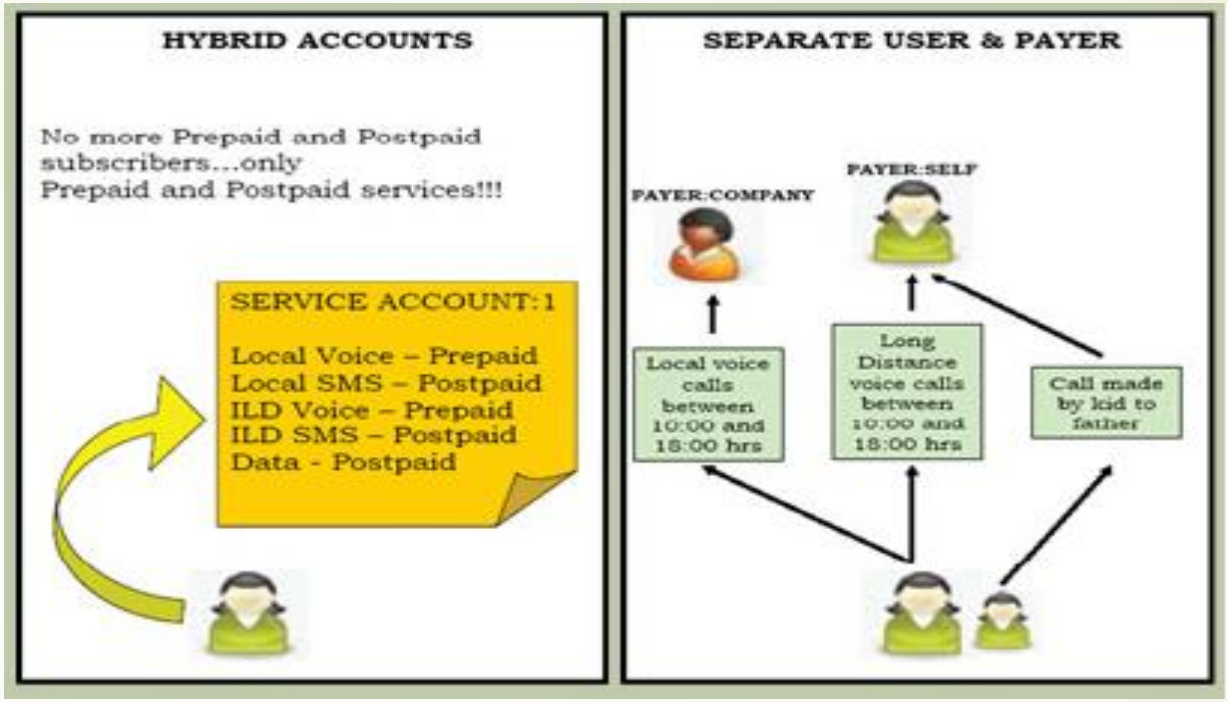


金融应用-印度国家证券交易所（NSE）



电信应用-Lifetree 公司

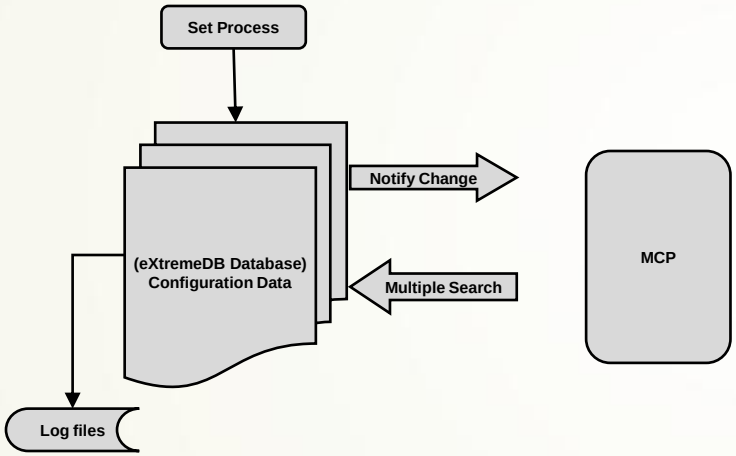
- 实时计费
- “Lifetree的开发团队评估过多个现有的商用数据库系统作为J@nus内存缓存的潜在供应商，并最终选择了eXtremeDB。那是由于eXtremeDB最快速的响应速度，高可用性和对64位的支持。” Lifetree CTO Naim Kazi说



操作系统:	Linux X64 Bit
CPU架构:	Intel Pentium
通讯方式:	Share Memory
使用数据库:	eXtremeDB HA

通讯应用-F5

- “为了达到如BIG-IP网络应用的性能要求，快速地设置查询结构参数是非常重要的。eXtremeDB的微秒级查询性能符合了这个要求。”F5 Networks的高级架构师 Dave Schmitt说。

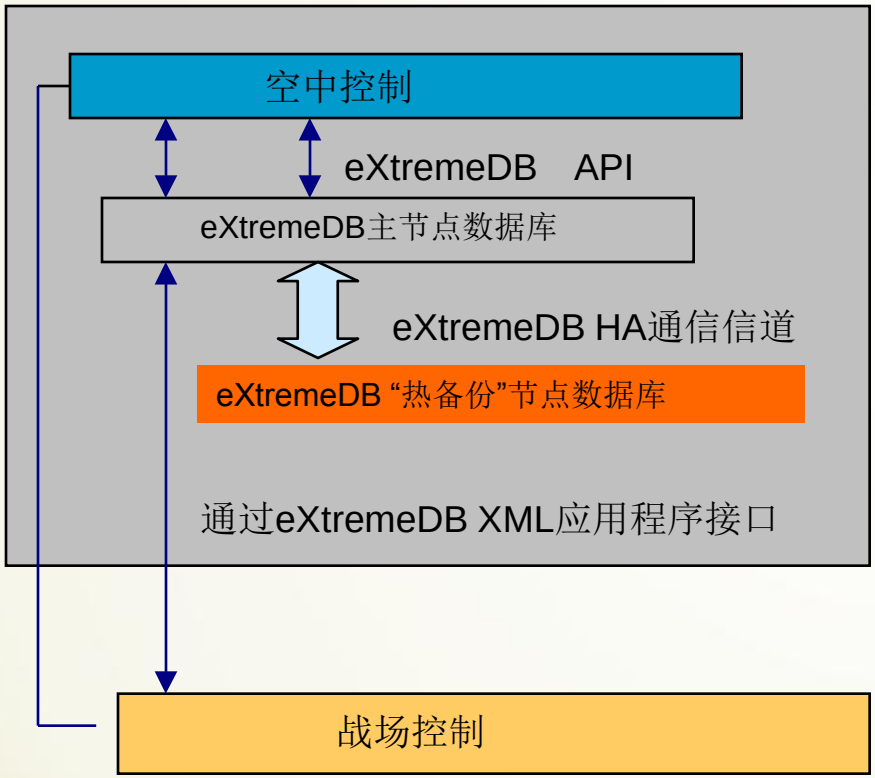


- 烽火通信

操作系统:	Embedded Linux
CPU架构:	x86
通讯方式:	Share Memory
使用数据库:	eXtremeDB Log

军工应用-波音公司

- 由波音公司生产的世界领先的**AH-64D Apache Longbow** 多功能的战斗直升机，利用eXtremeDB作为飞机关键的实时板载系统研发的一部分。



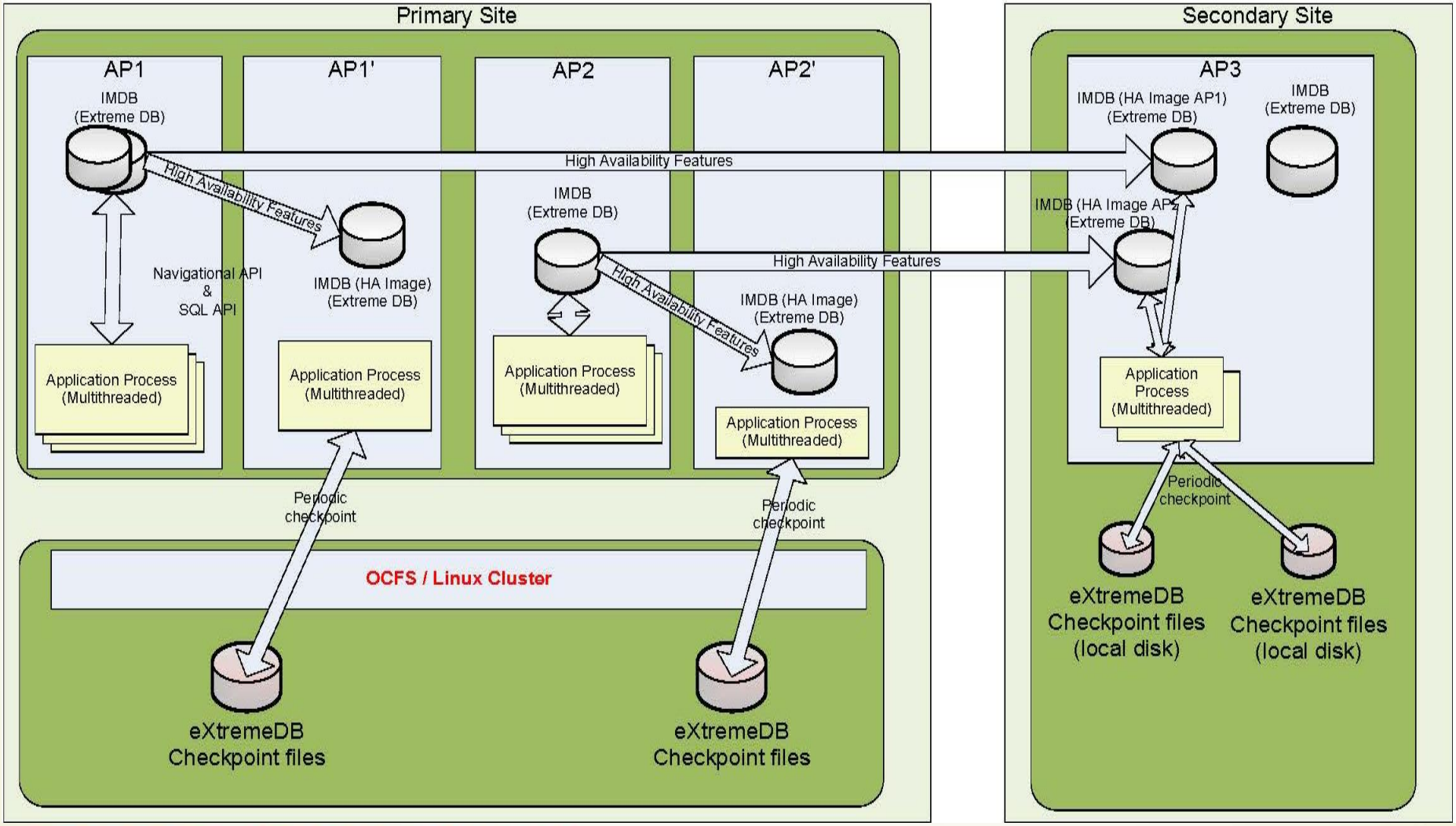
操作系统:	Commercial RTOS
CPU架构:	PowerPC
通信:	XML-based
On-device 数据库:	eXtremeDB HA

其他行业-福利彩票

- 黑龙江福彩研发团队，评估了多个内存数据库，以应对当期彩票销售状况实时监控和分析。eXtremeDB最终以极高的响应速度和高可靠的特性，成为黑龙江福彩的最终选择。尤其是HA特性，即可以保证数据的安全性，同时保证故障发生时的系统可靠性。

操作系统:	Redhat 6.1 64Bit
CPU架构:	X86
通讯:	TCP
使用数据库:	eXtremeDB HA

其他行业-香港赛马会


Primary Site
AP1
AP1'
AP2
AP2'
IMDB
(Extreme DB)
High Availability Features
Navigational API
&
SQL API
IMDB (HA Image)
(Extreme DB)
Application Process
(Multithreaded)
Periodic
checkpoint
OCFS / Linux Cluster
eXtremeDB
Checkpoint files
Secondary Site
AP3
IMDB (HA Image AP1)
(Extreme DB)
IMDB (HA Image AP2)
(Extreme DB)
eXtremeDB
Checkpoint files
(local disk)

谢谢!

Q&A

- Mac.Huang(黄东旭)
- McObject LLC
- +86-(0)13810448048
- @eXtremeDB
- mac.huang@mcobject.cn