

openGauss 技术特性和数据库基础



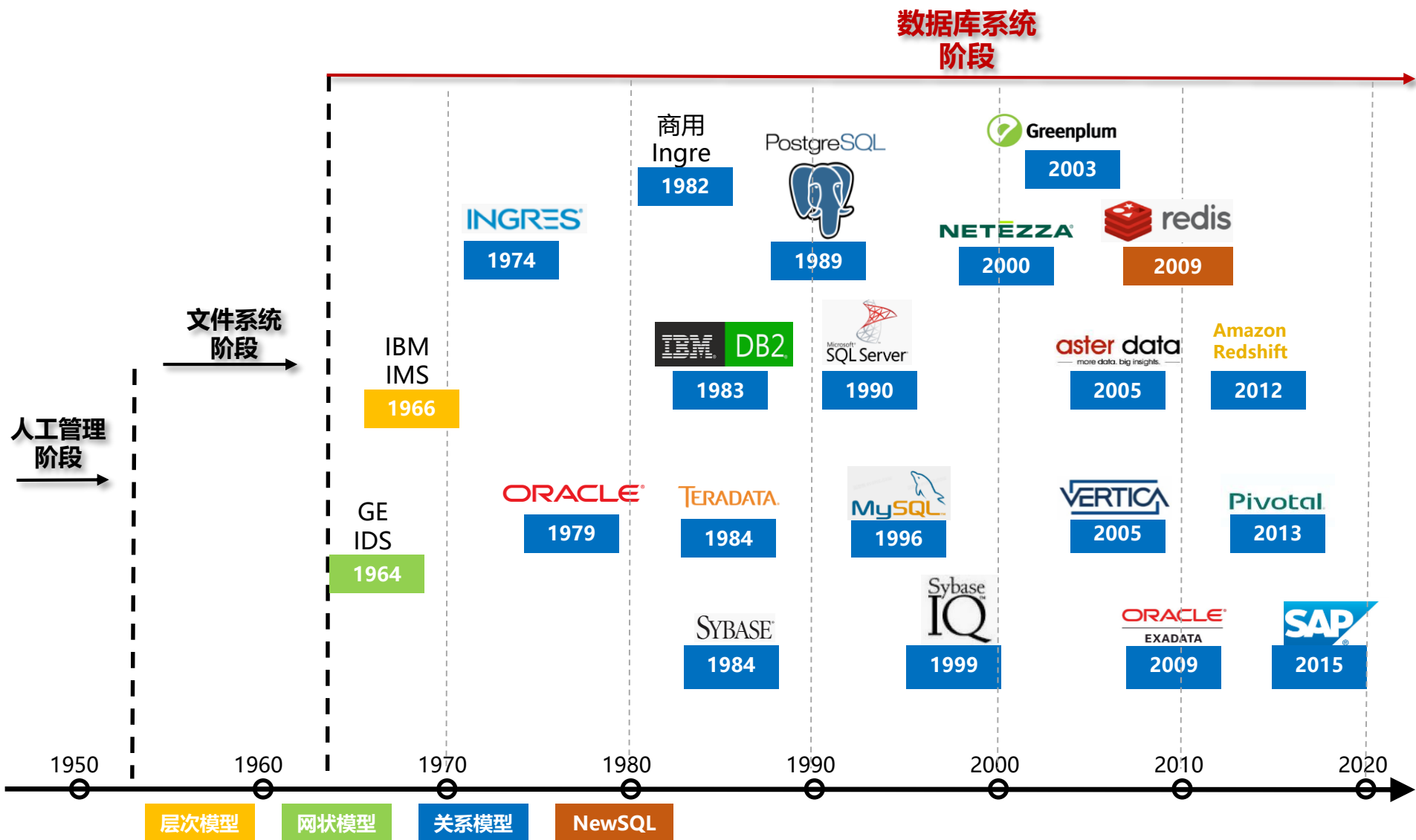
目录

第一部分：数据库概述

第二部分：openGauss技术特性介绍

第三部分：openGauss社区介绍

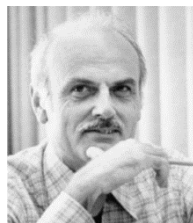
数据库技术发展史



Charles W. Bachman

1973年图灵奖

- 设计开发最早的网状数据库管理系统IDS并推动标准的制定



Edgar F. Codd

1981年图灵奖

- 首创关系模型理论



Jim Gray

1998年图灵奖

- 事务处理技术上的创造性思维和开拓性工作



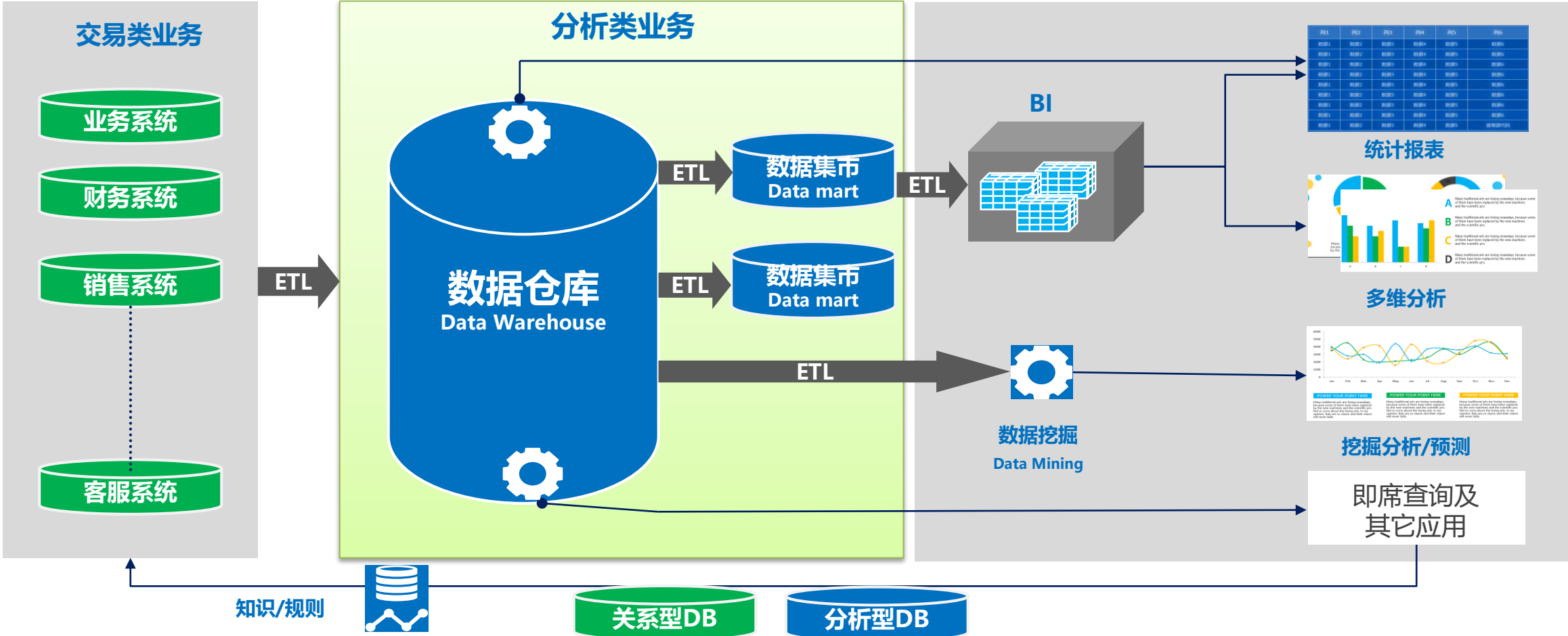
Michael Stonebraker

2014年图灵奖

- 对现代数据库的概念和实践作出的根本性贡献
- 主导参与Ingres、Postgres、Vertica、Aurora等项目

典型场景介绍：数据库在应用过程中主要有OLTP和OLAP场景

- 联机事务处理(OLTP)： 存储/查询业务应用中活动的数据以支撑日常的业务活动；
- 联机分析处理(OLAP)： 存储历史数据以支撑复杂的分析操作， 侧重决策支持；



主流关系型数据库系统架构对比：技术匹配应用诉求

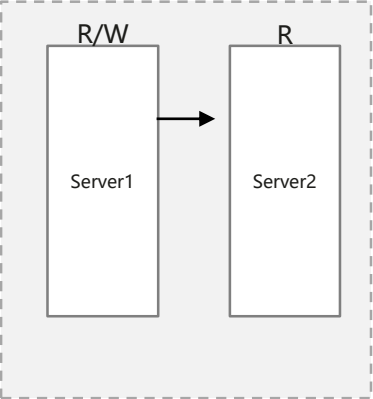
类型

技术架构

单机

代表：PG、MySQL
特点：
1、单机主备架构
2、扩展能力不足

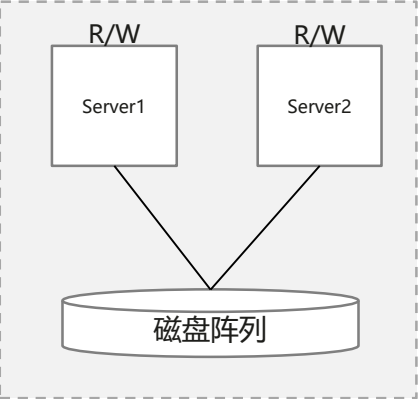
经典架构



Shared-Disk

代表：RAC/pureScale
特点：
1、多写、多读
2、Scale-up性能较高；
Scale-out扩展性不足

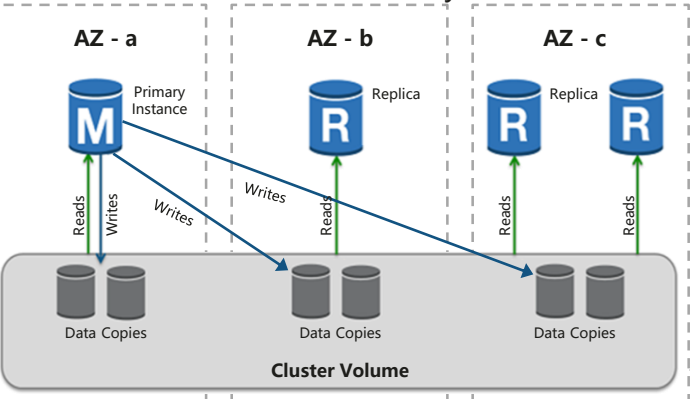
Oracle RAC



Shared-Virtual-Disk (分布式存储架构)

代表：Aurora
特点：
1、一写、多读；计算、存储分离；多副本
读提升性能，写性能受单Primary节点限制
2、支持paxos/raft多AZ高可用

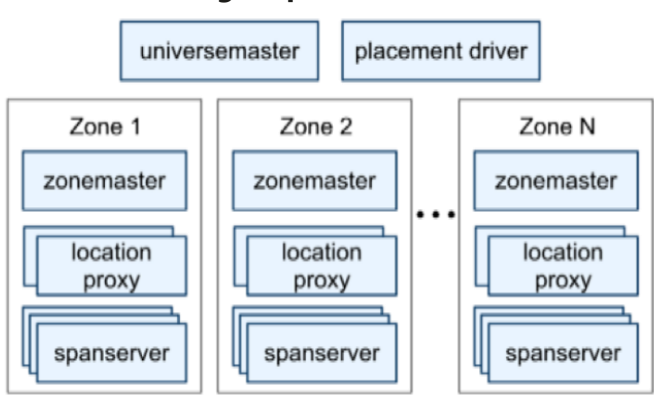
AWS Aurora (基于MySQL和PG)



Shared-Nothing (分布式架构)

代表：Spanner
特点：
1、sharding多写、多读；读写性能均可
Scale-out，不受限于单节点处理能力
2、支持paxos/raft多AZ高可用

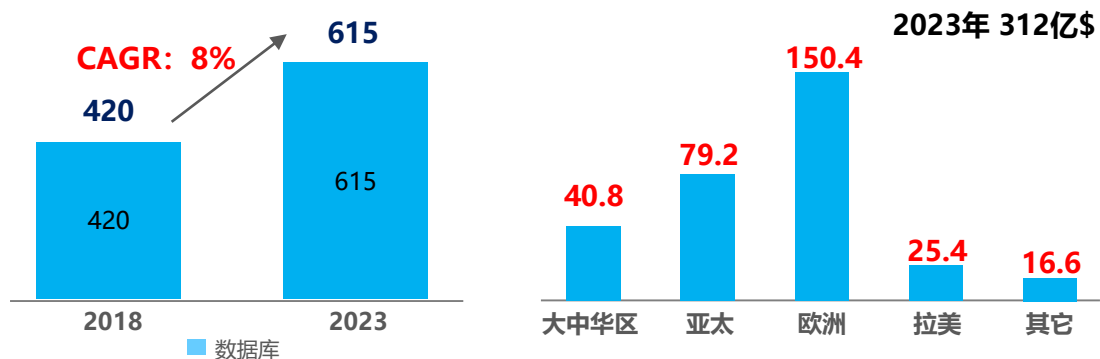
Google Spanner



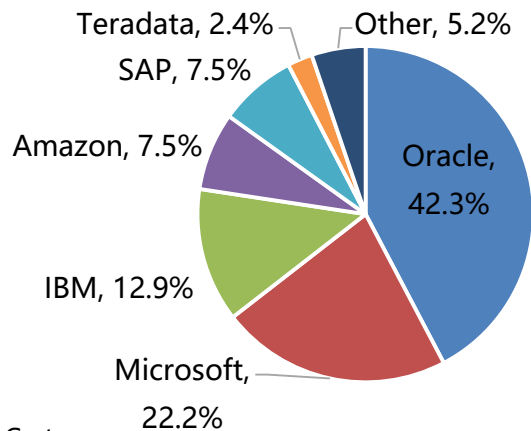
数据库市场空间：2018年数据库全球市场460亿\$；中国市场23.3亿\$，增长率59%

全球数据库市场空间分析

- 2023年全球数据库市场空间615亿美金



- 全球数据库市场Oracle、微软、IBM、Amazon、SAP、Teradata六大厂商瓜分90%以上的市场；大机、小机及一体机占据了20%高端市场



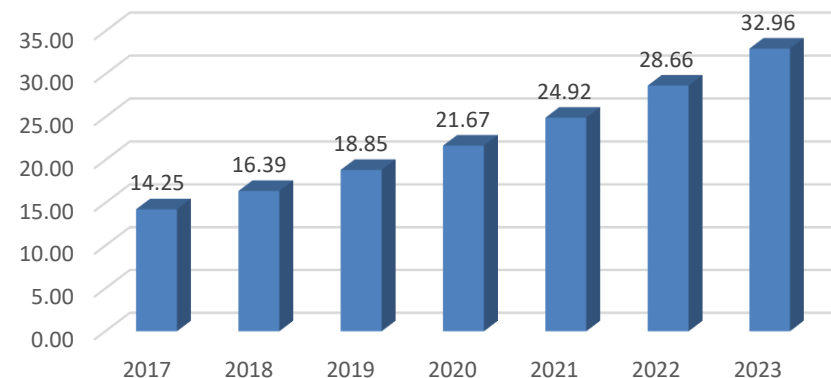
Notes:

1. IBM大机+小机共\$64亿
2. Oracle一体机\$36亿，共100亿美金
3. 一体机模式软硬件比约 1:5

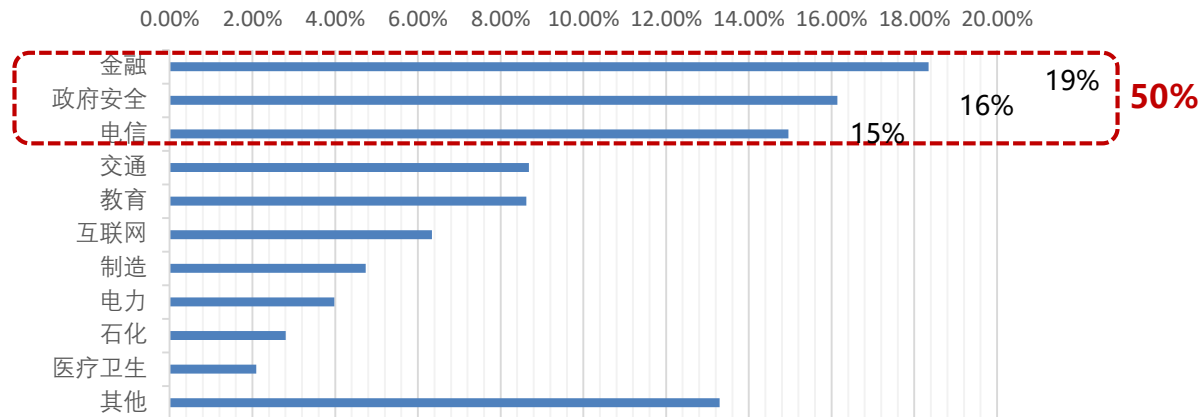
数据来源：Gartner

中国数据库市场空间分析

- 预计未来5年，中国区数据库市场空间累计约\$130亿，复合增长率超15%；



- 金融、运营商、政府、安全占据50%的市场空间

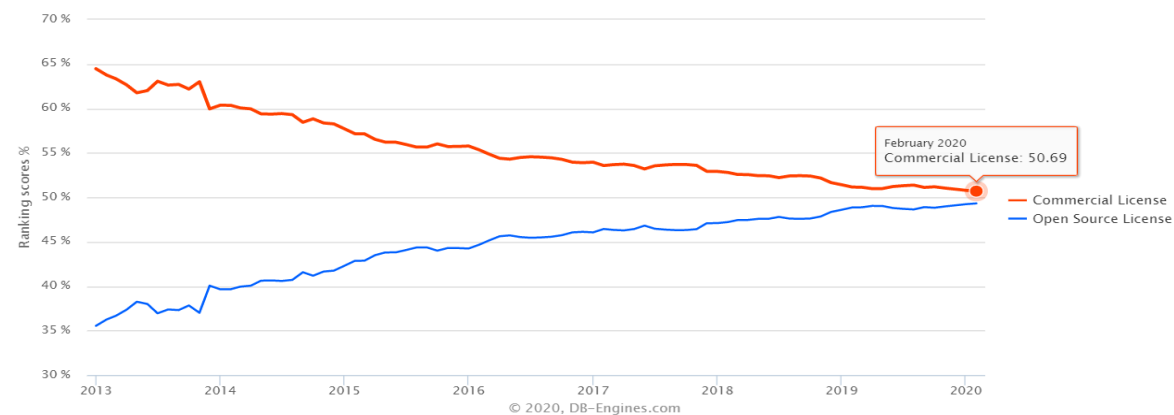


数据来源：赛迪顾问

开源数据库趋势：流行程度逐渐赶超商业数据库，关系数据库在市场中占主流

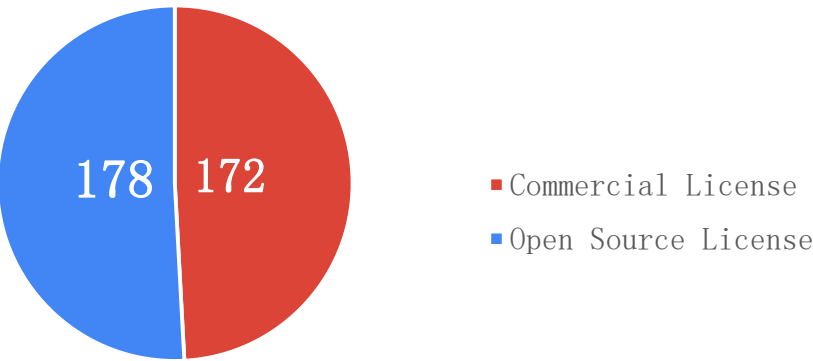
开源数据库的流行程度逐渐赶超商业数据库

- 2020年2月，DB-Engine上对比了的开源和商业数据库管理系统普及历史趋势显示，开源数据库的**流行度为49%**与商业数据库基本持平，并有超越趋势



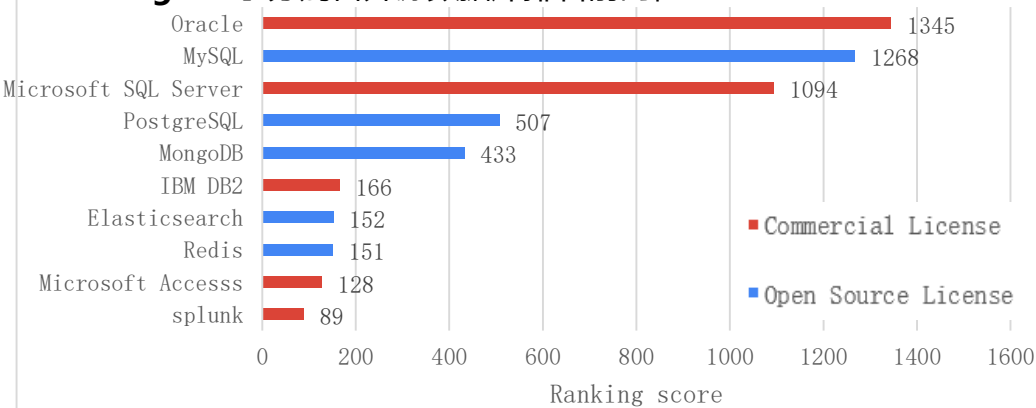
开源数据库的数量超越商业数据库

- DB-Engine社区调研对比了350种数据库，结果显示开源数据库的数量超越商业数据库，开源数据库在业界已逐渐成流行趋势



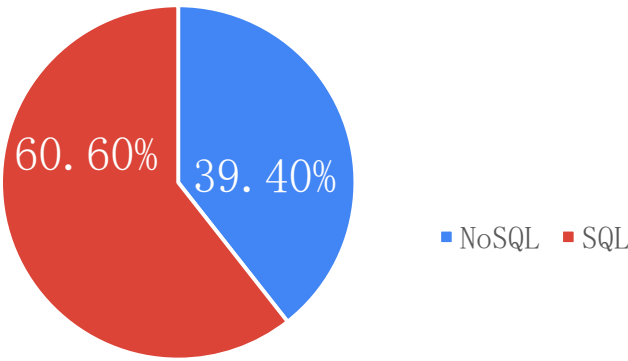
关系型数据库在开源社区中仍占主流

- 全球数据库产品流行度排名中，关系型开源数据库**MySQL和PostgreSQL**分别占开源数据库排名前两位



开源SQL数据库的数量占主流

- 在对开源数据库流行程度调研中，关系数据库类型占比超过60%，基于SQL的关系数据库在主流开源数据库产品中仍为主流



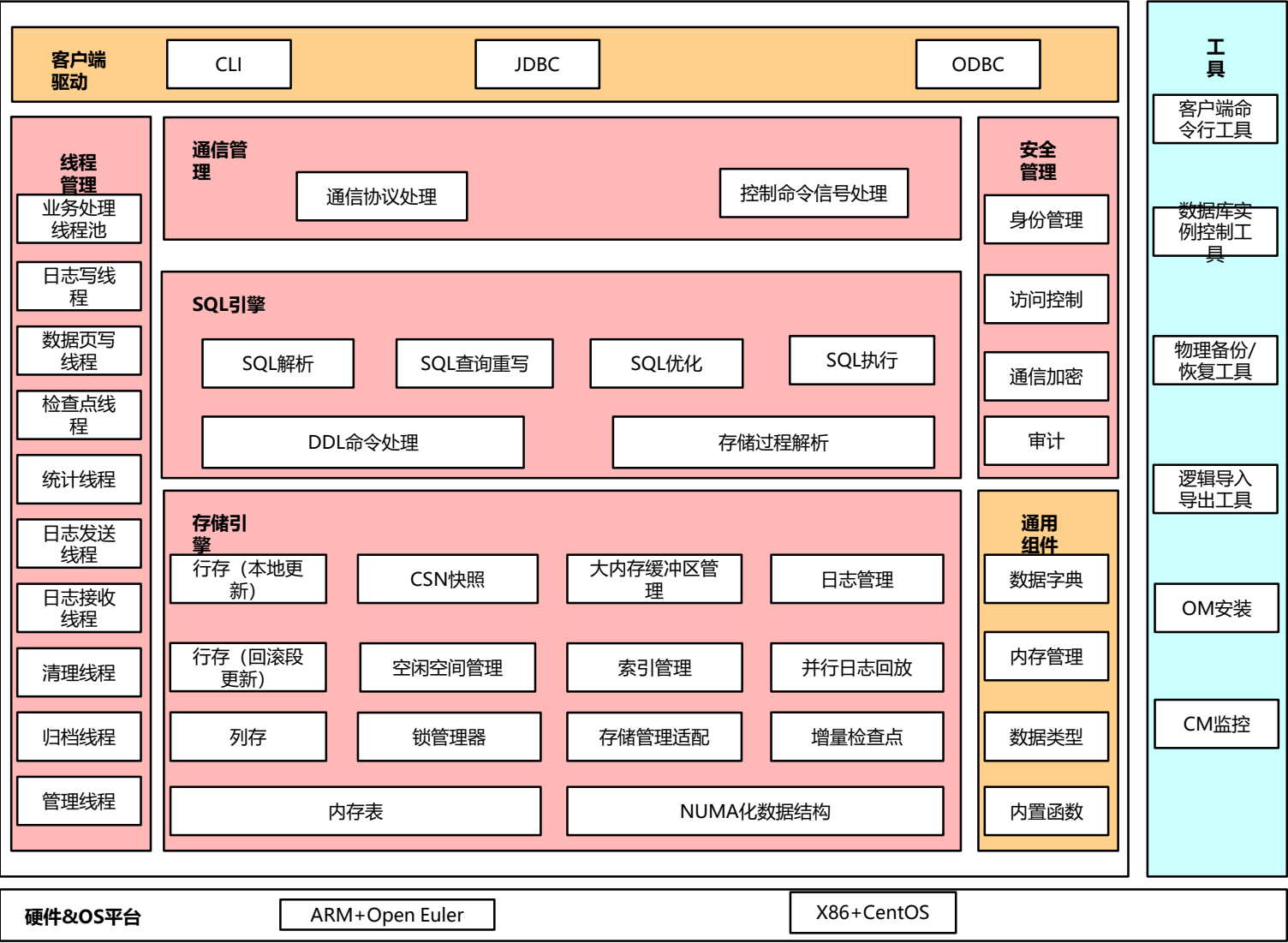
目录

第一部分：数据库概述

第二部分：openGauss技术特性介绍

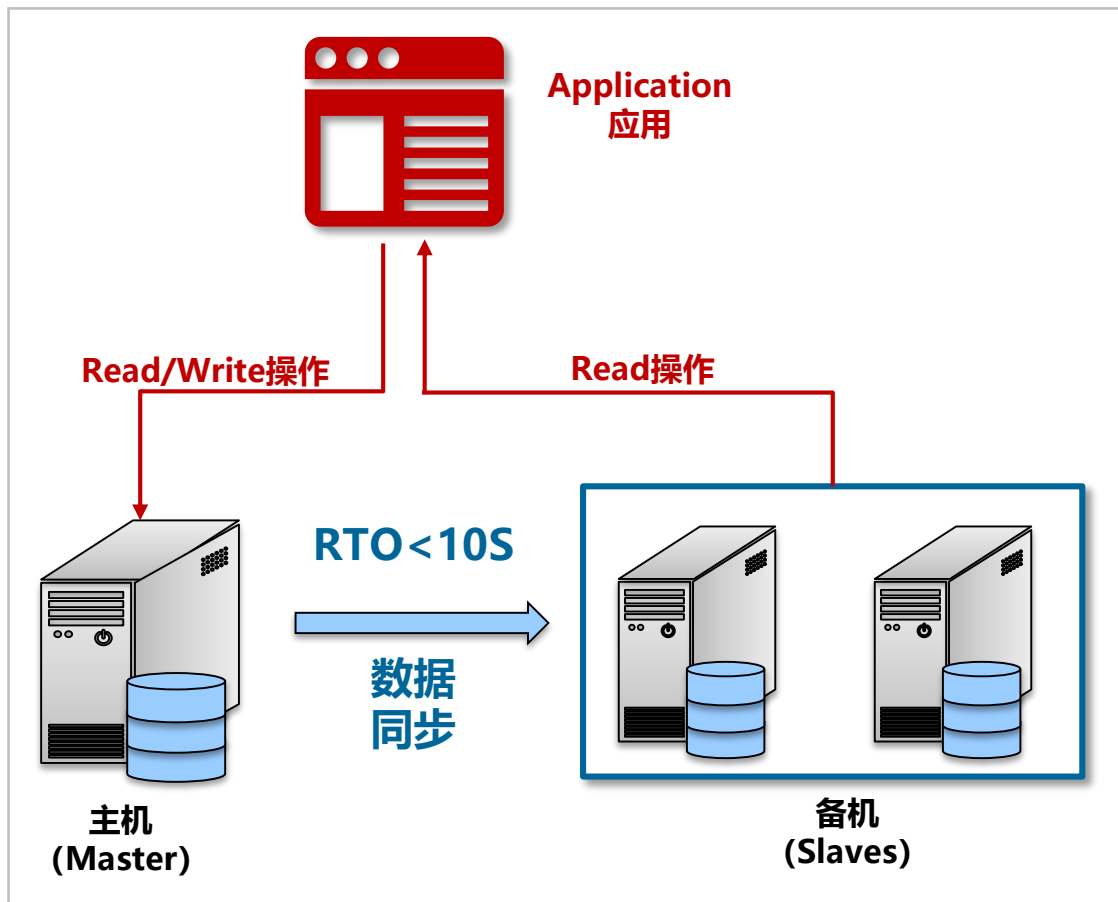
第三部分：openGauss社区介绍

openGauss介绍：单机主备架构，源于PG9.2，深度修改，内核自研占比74%



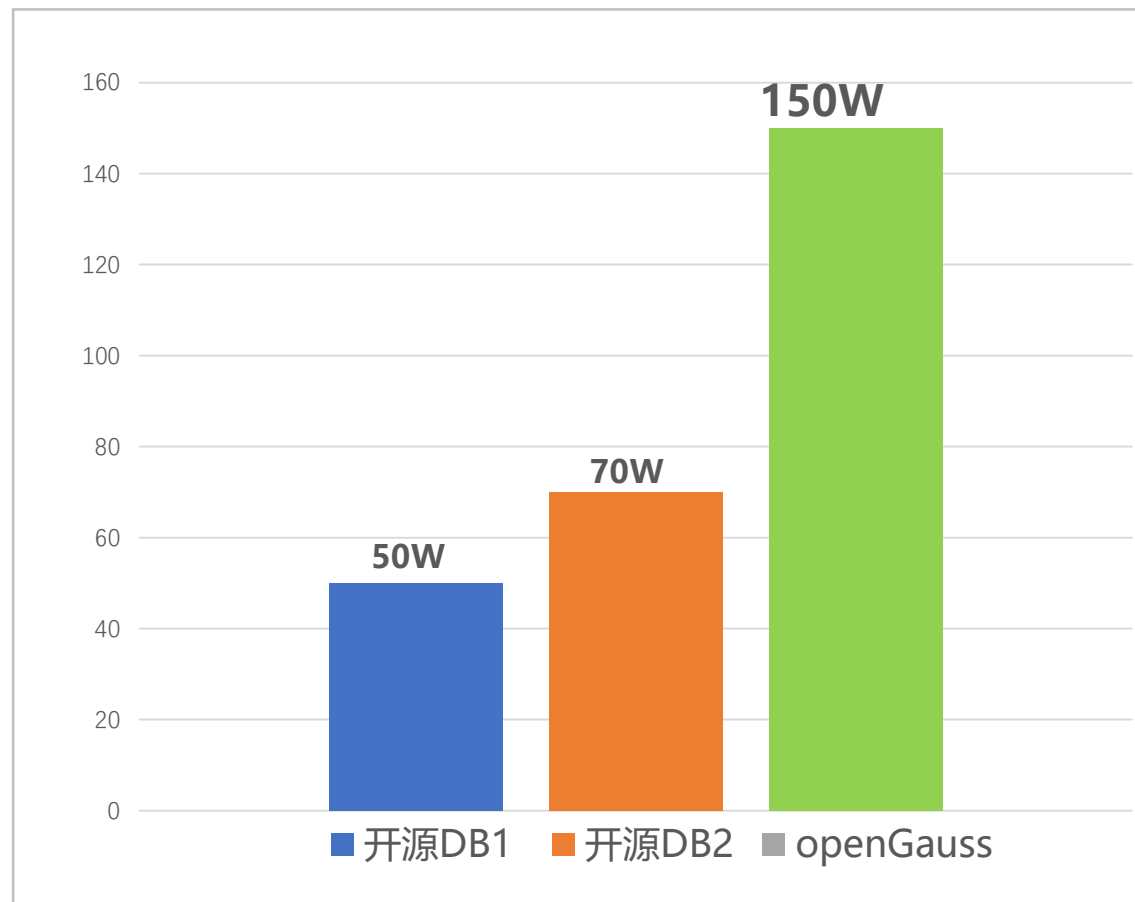
- 源自于PostgreSQL9.2和PG XC，深度修改
- openGauss总代码量120w行，
 - 其中内核代码95w万行代码；内核中修改和新增了70w行核心代码；
 - 保留了PostgreSQL的接口和公共函数25w行；
- openGauss并不等于PostgreSQL的简单增强版，openGauss着重在架构、事务、存储引擎、优化器、和鲲鹏芯片优化上进行深度修改，经过时间积累，市场项目打磨，当前已经成为一款企业级开源数据库产品。

openGauss介绍：单机主备架构，具有高性能、高可用、高可靠等技术优势



一主多备示例

高可用：业内最快故障恢复 (RTO < 10秒)



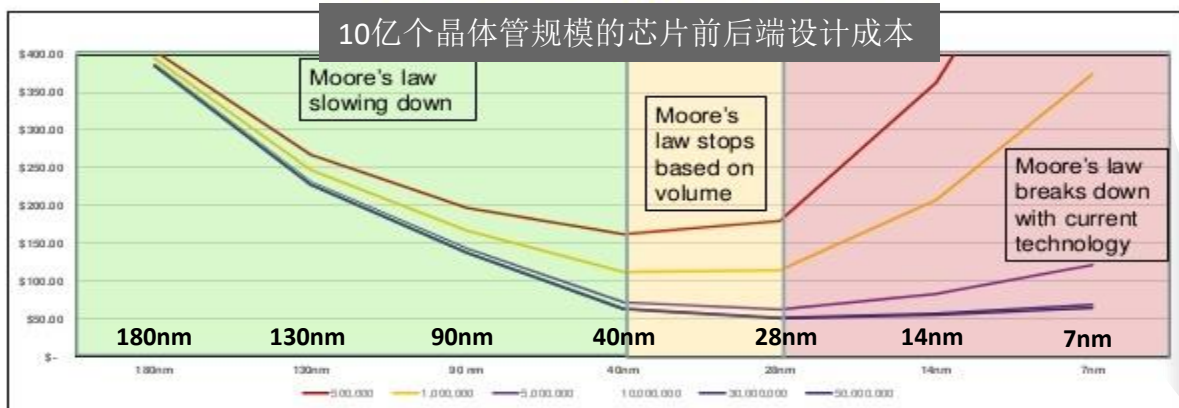
OLTP-TPCC标准Benchmark

高性能：两路鲲鹏下性能达到150W tpmC

硬件发展趋势：单核架构向多核架构发展，单位芯片面积提供更强算力

芯片工艺进入28nm以下空间时，靠先进工艺提升性能成本急剧上升

Total Cost of Moore's Law



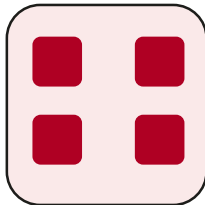
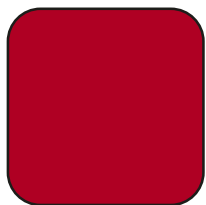
Aurora Semiconductor Proprietary



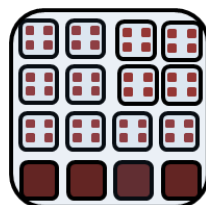
2

工艺、主频遇到瓶颈后，开始转向增加核数的横向扩展来提升性能

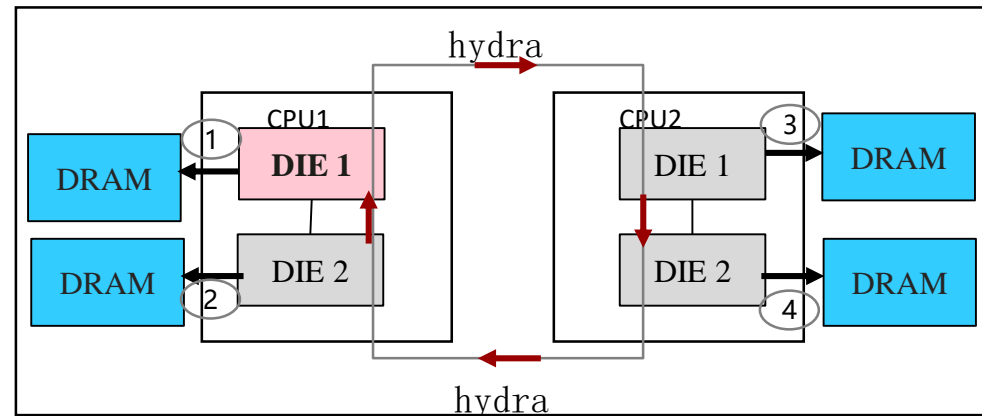
单核



多核



鲲鹏服务器系统架构



Multi-core, high-concurrency
多核高并发

- 芯片的物理尺寸有限制，不能无限制的增加
- ARM的多核横向扩展空间优势明显

多核下数据库优化研究策略：提高数据缓存局部性、降低多核数据同步开销

算法优化示例

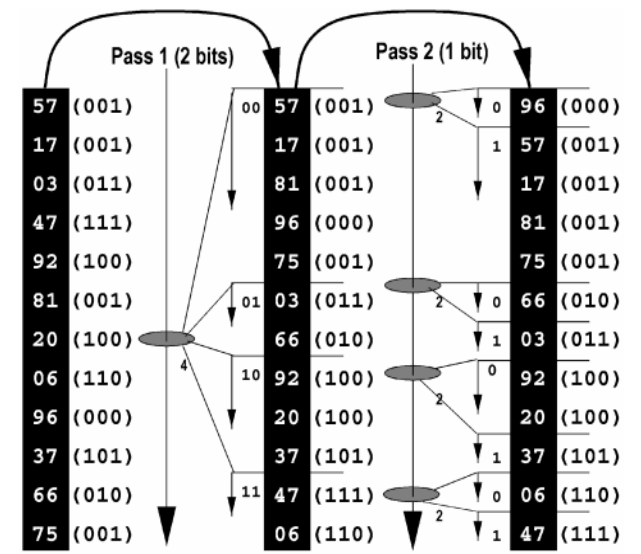
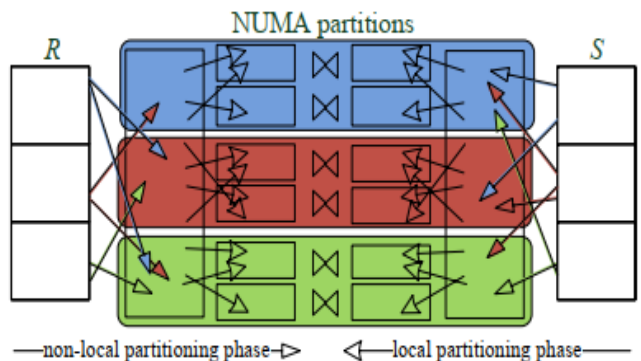


Fig. 9. Two-pass/3-bit radix cluster (lower bits indicated between parentheses).



Radix join processing

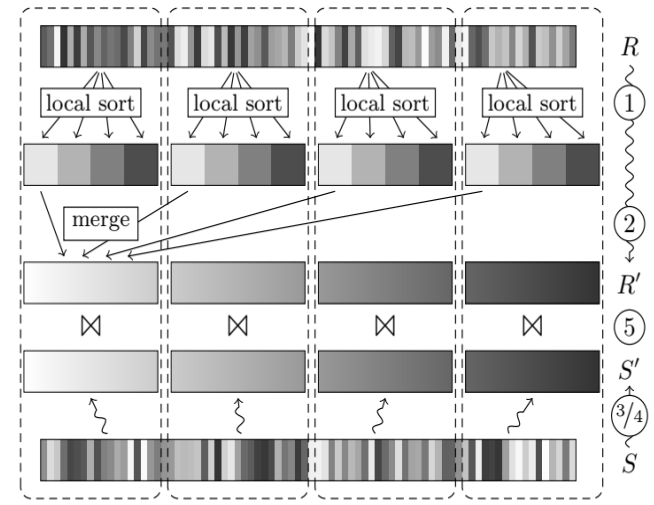
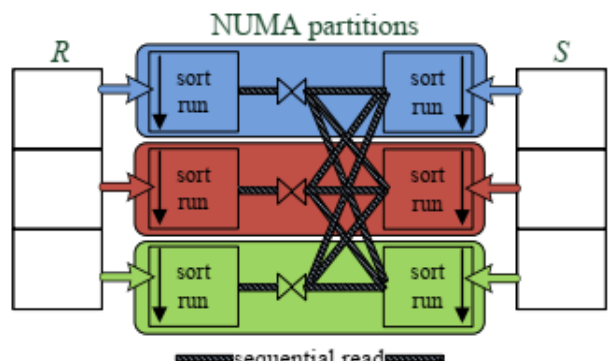
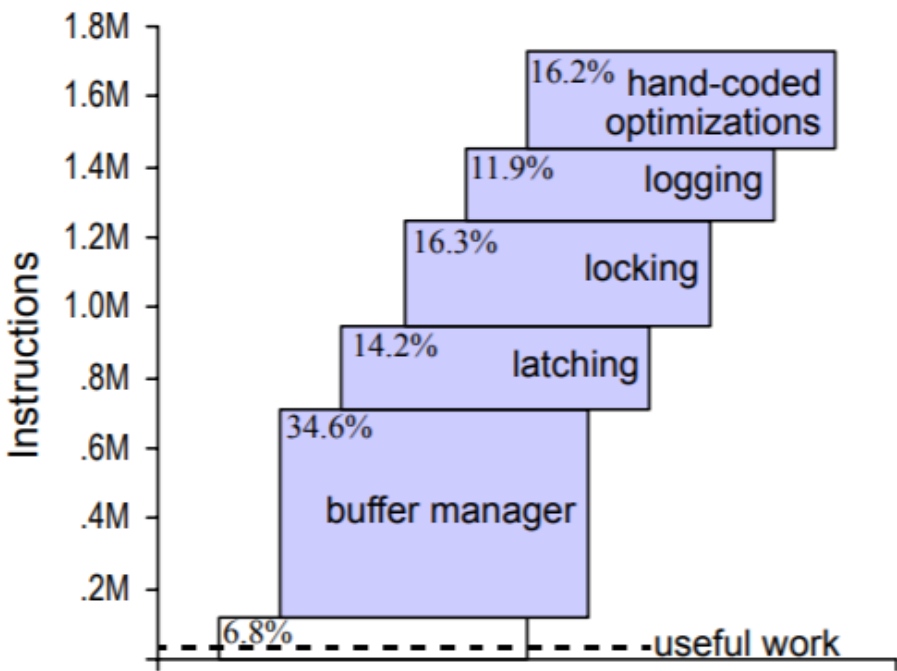


Figure 4: *m*-way: NUMA-aware sort-merge join with multi-way merge and SIMD.



Sort Merge join processing



锁、日志、协议优化：

- 轻量级意图锁 (LIL) (Lightweight Intent Locks)
- 投机性锁继承技术 (SLI) (SpeculativeLock Inheritance)
- 可伸缩日志
- 基于乐观锁的多版本控制协议

➢

高性能：通过NUMA化改造，构筑多核扩展性、实现极致性能

场景和背景：

OLTP场景下DML语句（Insert, Update, Delete）大量并发操作trx_sys全局结构体中的关键数据结构，造成临界区的竞争和同步瓶颈。

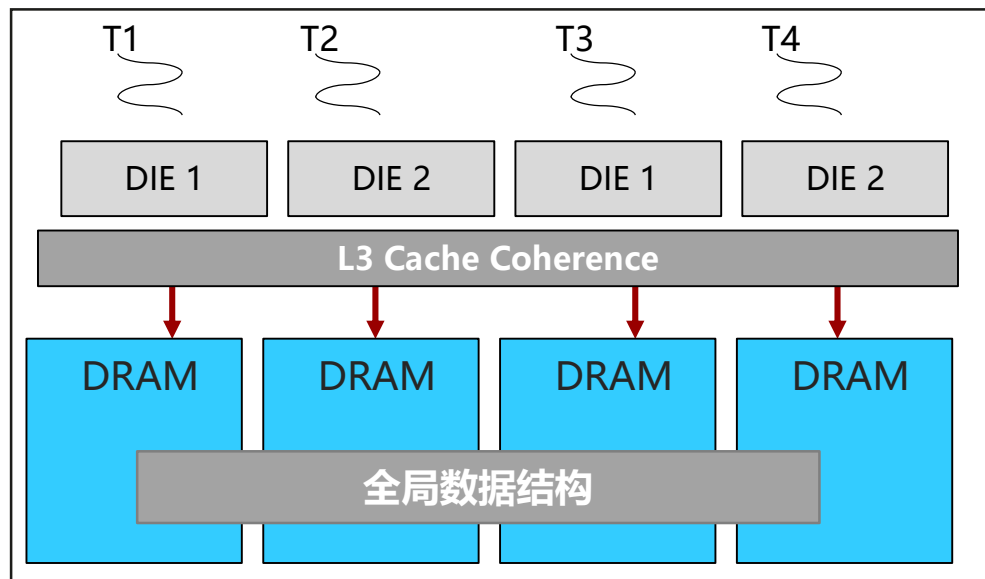
```
[minqi@localhost dTrace]$ ./lwlock_wait.d
```

Lock Id	Mode	Count
CLogControlLock	Exclusive	2085
CLogControlLock	Shared	2336
WALInsertLock	Exclusive	3993
ProcArrayLock	Shared	7400
ProcArrayLock	Exclusive	34405
WALWriteLock	Exclusive	208354
XidGenLock	Exclusive	275126

Lock Id	Mode	Combined Time (ns)
CLogControlLock	Exclusive	28956662
CLogControlLock	Shared	29554120
WALInsertLock	Exclusive	58710394
ProcArrayLock	Shared	125250951
ProcArrayLock	Exclusive	446470964
WALWriteLock	Exclusive	19448626395
XidGenLock	Exclusive	450481068249

TPCC result profiling

- 对传统数据库（PG等）事务执行Profiling，存在五个关键性能瓶颈点：Clog、WALInsert、WALWrite、ProcArray、XidGen。
- 只有消除串行化点，尽可能多核并行化才能释放算力优势。



- 基于操作系统能力，对工作进程进行NUMA绑核，减少跨核访问
- 全局数据结构（ProcArray/Buffer/B-Tree等）**NUMA分区化**改造，减少跨核、跨处理器竞争冲突；
- 并发控制原语改造，高并发Spin Lock原子锁效率和临界区代价高2-3倍
- Cache line对齐，减少cache miss，提升整体性能

高性能：软硬结合，指令级优化，复用硬件能力，提升系统整体性能

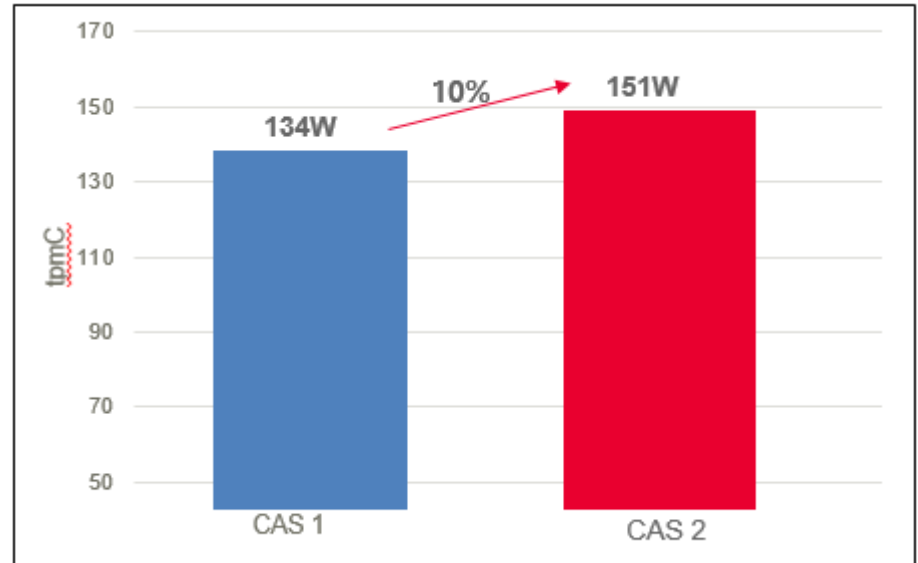
CAS (Compare and Swap) 有3个操作数，内存值V，旧的预期值A，要修改的新值B。当且仅当预期值A和内存值V相同时，将内存值V修改为B，否则什么都不做。

CAS实现方式一：由ldxr和stlxr指令来实现

```
20      os_compare_and_swap_uint(&old,old,new);
0x000000000004005bc <+24>:   ldr     w0, [x29,#24]
0x000000000004005c0 <+28>:   mov     w2, w0
0x000000000004005c4 <+32>:   ldr     w1, [x29,#28]
0x000000000004005c8 <+36>:   add     x0, x29, #0x18
0x000000000004005cc <+40>:   ldxr    w3, [x0]
0x000000000004005d0 <+44>:   cmp     w3, w2
0x000000000004005d4 <+48>:   b.ne    0x4005e0 <main+60>
0x000000000004005d8 <+52>:   stlxr   w4, w1, [x0]
0x000000000004005dc <+56>:   cbnz    w4, 0x4005cc <main+40>
0x000000000004005e0 <+60>:   dmb     ish
```

CAS实现方式二：由casal一条指令来实现

```
4005b8:      b9001ba0      stl     w0, [x29,#24]
// int *ptr=100;
os_compare_and_swap_uint(&old,old,new);
4005bc:      b9401ba0      ldr     w0, [x29,#24]
4005c0:      2a0003e2      mov     w2, w0
4005c4:      b9401fa1      ldr     w1, [x29,#28]
4005c8:      910063a0      add     x0, x29, #0x18
4005cc:      2a0203e3      mov     w3, w2
4005d0:      88e3fc01      casal   w3, w1, [x0]
4005d4:      6b02007f      cmp     w3, w2
```

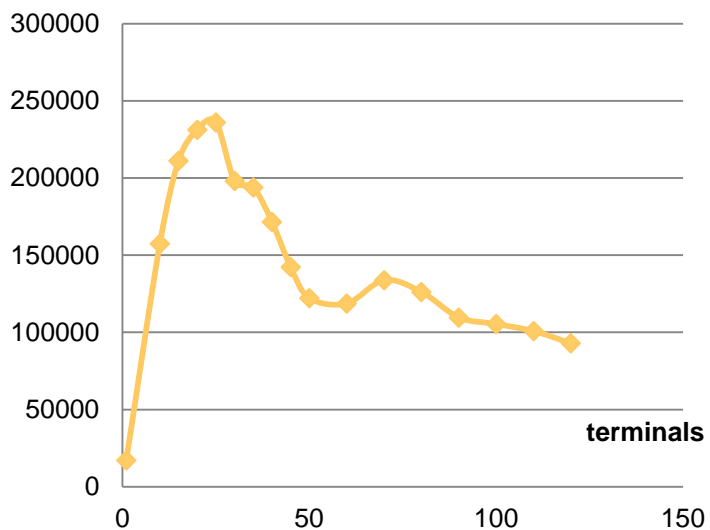


四个指令→1个指令，提升执行效率；

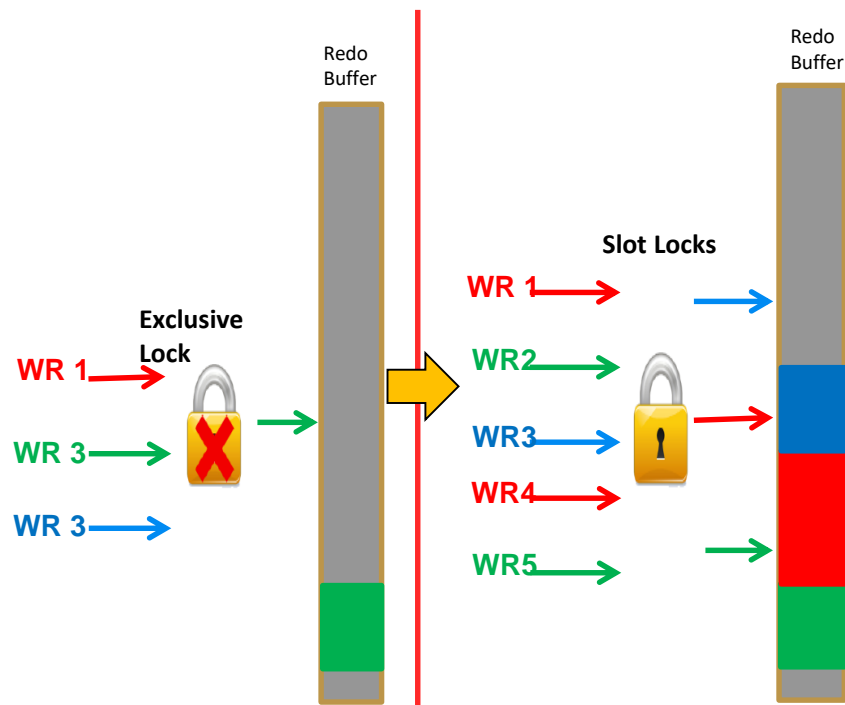
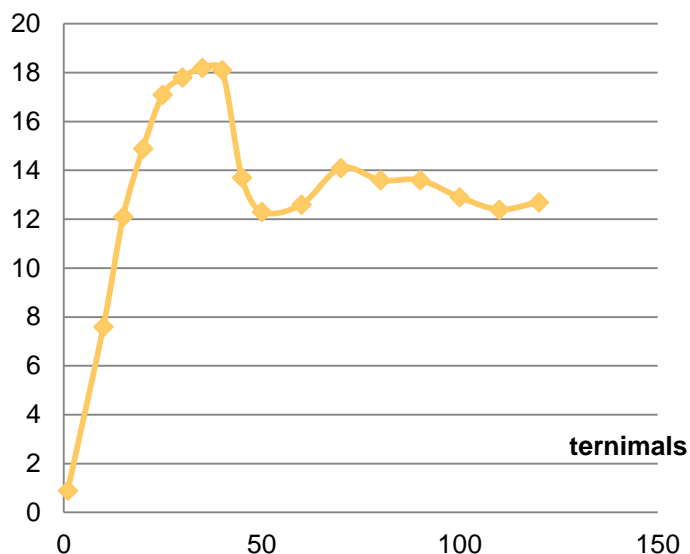
高性能：面向多核的并发控制和日志提交算法，提升高并发事务处理性能

- 关键算法面向多核优化，提升高并发事务处理性能

tpmC随并发负载的变化情况示意图



不同并发负载cpu利用率变化情况示意图



在高并发事务处理负载下，核心事务处理算法通常成为系统瓶颈，限制系统事务处理性能

多路WAL日志算法，提升高并发下的事务提交性能

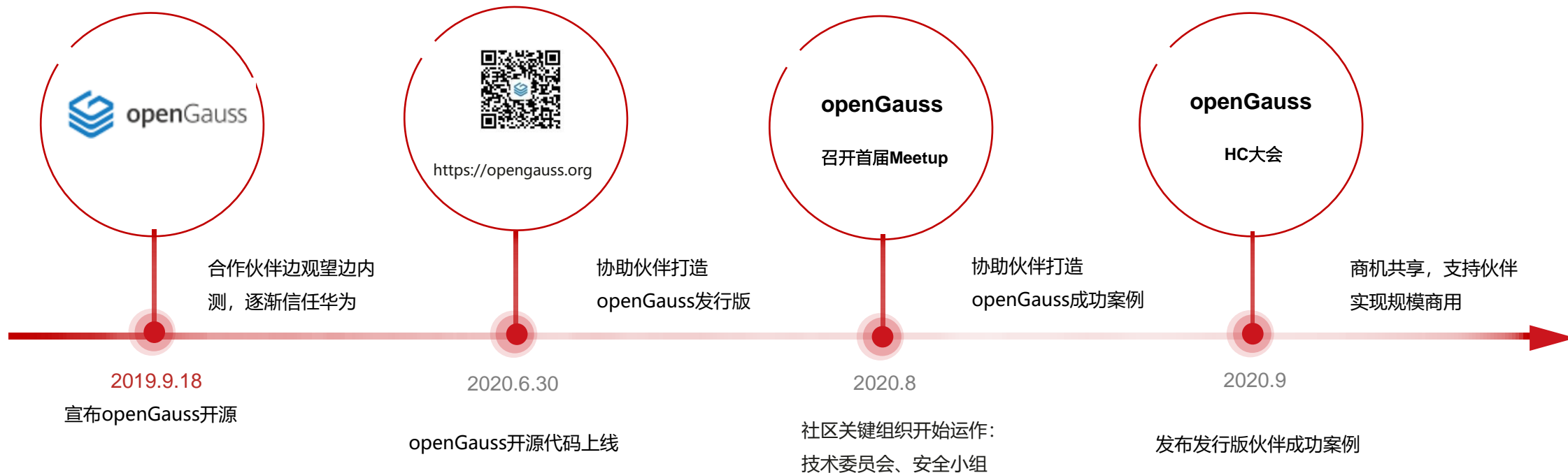
目录

第一部分：数据库概述

第二部分：openGauss技术特性介绍

第三部分：openGauss社区介绍

openGauss开源社区介绍:



官方网站: <https://opengauss.org>



openGauss组织仓库: <https://gitee.com/opengauss>

openGauss镜像仓库: <https://github.com/opengauss-mirror>

openGauss社区简介：社区活跃，内容持续构筑完善

运营数据

从6月30日开源到8月1日：

- 官方网站访问量：210734
- 官方网站访客人数：16770
- 官方网站视频播放量：52455
- 官方网站安装包下载量：10785



数据库介绍



<https://opengauss.org/zh/events.html>


openGauss性能调优

<https://opengauss.org/zh/video.html>


协议友好，欢迎基于openGauss，开展教学、科研活动

开源数据库	License	类型
openGauss	MulanPS L	BSD类,允许随意商业集成
MySQL	GPL v2	GPL 类,传染性,商业不友好
MariaDB	GPL v2	GPL 类
PostgreSQL	PostgreSQL	BSD类
TBase	BSD 3-Clause	BSD类
TiDB	Apache 2.0	BSD类


PostgreSQL is released under the **PostgreSQL License**, a liberal Open Source license, similar to the BSD or MIT licenses.




Very Large Data Bases




Very Large Data Bases Endowment Inc.




ICDE




IEEE International Conference on Data Engineering



ACM SIGMOD






TKDE



Transactions on Knowledge and Data Engineering




.....



ARM、X86 openGauss

软硬结合，释放算力潜能

多核优化



Blockchain openGauss

Blockchain in DB

.....

openGauss、MySQL支持相同的SQL标准，通用接口，欢迎基于openGauss参赛



标准SQL支持:

- 支持标准的SQL92/SQL2003规范
- 支持GBK和UTF-8字符集等

应用程序接口:

- 支持标准JDBC 特性
- 支持标准ODBC 特性

数据库开发环境

opengauss.org/zh/video/20200503.html

openGauss 主页 下载 文档 社区 安全 新闻 活动 博客 视频

JDBC运行和测试 (2)

- 编译运行JDBC应用程序代码如下:

```
public class jdbc_test{
    public void test() {
        // 驱动类
        String driver = "org.postgresql.Driver";
        // 数据库连接描述符
        String sourceURL = "jdbc:postgresql:@10.255.255.1:16666";
        Connection conn = null;
        try {
            // 加载数据库驱动
            Class.forName(driver).newInstance();
        } catch (Exception e) {
            // 抛出异常
            e.printStackTrace();
        }
    }
}
```

8:12 / 33:14

数据库开发环境

数据库基础 2020-05-03

<https://opengauss.org/zh/video/20200503.html>



Thank you.



把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

**Copyright©2018 Huawei Technologies Co., Ltd.
All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

