
GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「数据与数据库」专场
本期议题：阿里云数据库架构演进及内核解读
于巍（花名 漠雪）2023年05月15日

《阿里云数据库架构演进及内核解读》目录

GOTC

1、阿里云瑶池数据库概述

阿里云全新“瑶池数据库”品牌； Gartner领导者象限、国内市场份额第一；All In Serverless战略

2、阿里云云原生数仓AnalyticDB架构演进

OLAP技术趋势；云原生数仓AnalyticDB；多 Master 架构； Laser 计算引擎； DADI缓存

3、阿里云云原生关系型数据库PolarDB架构演进

云原生数据库PolarDB，自研创新，步履不停；三层解耦；软硬协同创新； HTAP, 事务处理与计算分析一体化

4、阿里云开源PolarDB-X 架构介绍

自研创新 PolarDB：全面开源；PolarDB-X 兼容 MySQL 生态；PolarDB-X 技术架构

5、阿里云开源PolarDB-PG架构及代码解读

存储计算分离；存储层；缓存层；日志层；基于共享存储的MPP架构；PolarDB社区内核课程

01　阿里云数据库概述

阿里云瑶池数据库：云原生一站式数据管理与服务

Alibaba Cloud Apsara Database: One-stop, Cloud-native Data Management & Data Serving Platform

全球云数据库市场格局

阿里云数据库产品进入Gartner领导者象限

阿里云数据库产品蝉联
Gartner 2021 数据库魔力象限领导者

Gartner

Magic Quadrant

Figure 1: Magic Quadrant for Cloud Database Management Systems

Source: Gartner (December 2021)

不仅是阿里云数据库的重要突破，也是
中国数据库的重要里程碑

云原生数据仓库 AnalyticDB

TPC® 第一

TPC-DS 基准测试跑分较之第二名

速度领先 64%　成本减少 3/4

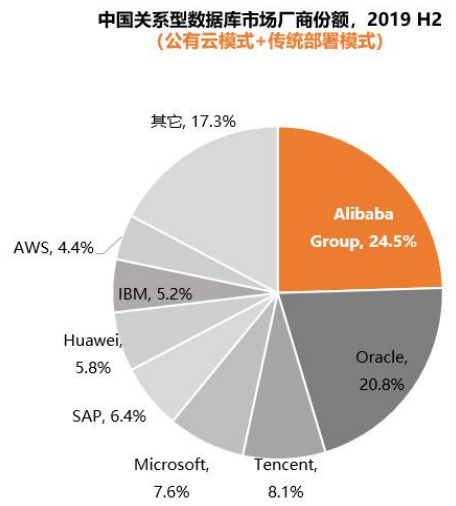
TPC-H 基准测试跑分较之第二名

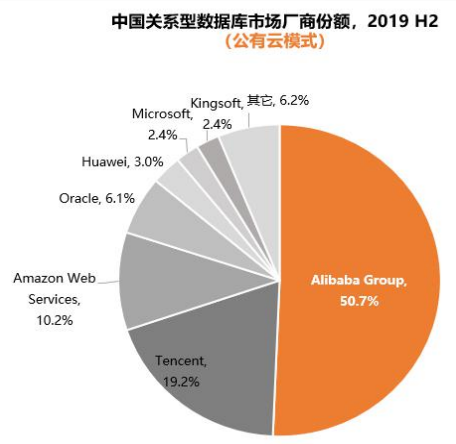
速度领先 3x

Forrester 全球云数据仓库卓越表现者象限

来源：Forrester 2021，TPC 2020

阿里云国内市场份额第一





阿里云数据库All In Serverless - 按需取用，按量付费的云原生数据库

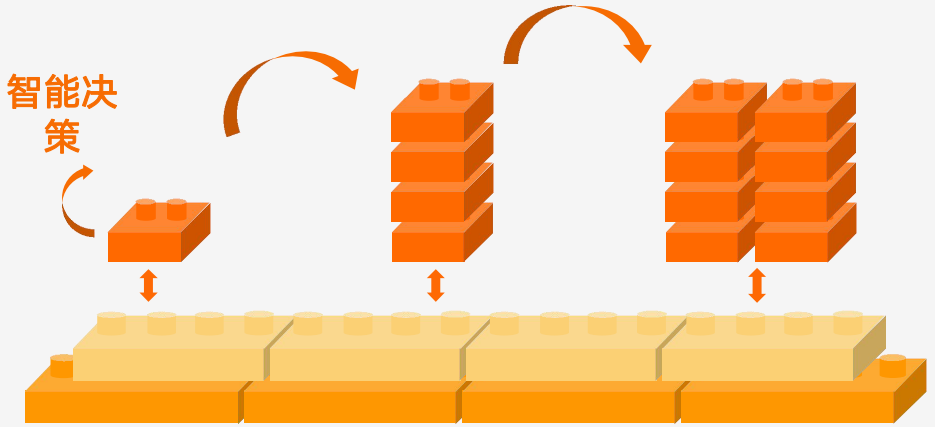
信通院Serverless认证，PolarDB和RDS获评**最高“先进级”**，AnalyticDB “增强级”

PolarDB业内第一个上限1000核的企业级Serverless数据库

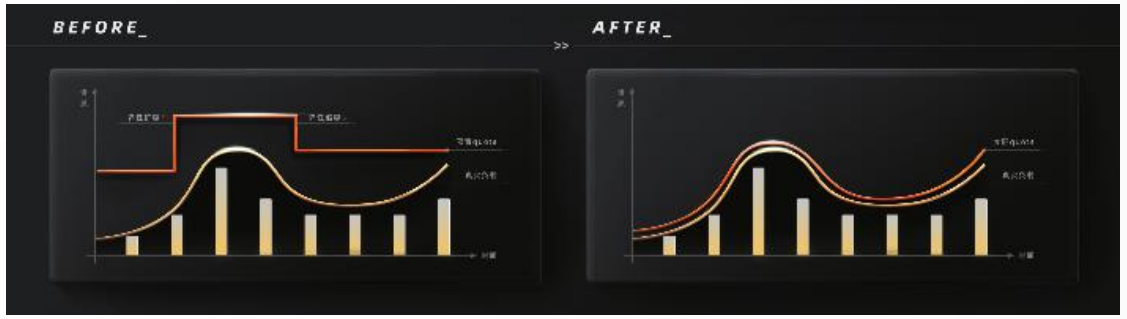
本地 ScaleUp
无感 BP Resize

跨机 ScaleUp
基于热备技术的秒级切换
连接和事务跨机续传

跨机 ScaleOut
集群维度高性能全局一致性
热资源池实现秒级横向弹性



智能无感秒级弹性



弹的更广
- 0~1000核，业内范围最广

弹的更稳
- 全场景（本机&跨机、HTAP、全局数据强一致）无感弹升，性能线性提升

弹的更快
- 秒级探测，秒级切换

弹的更细
- 三层解耦独立弹升，对比传统架构Serverless成本再降低60%

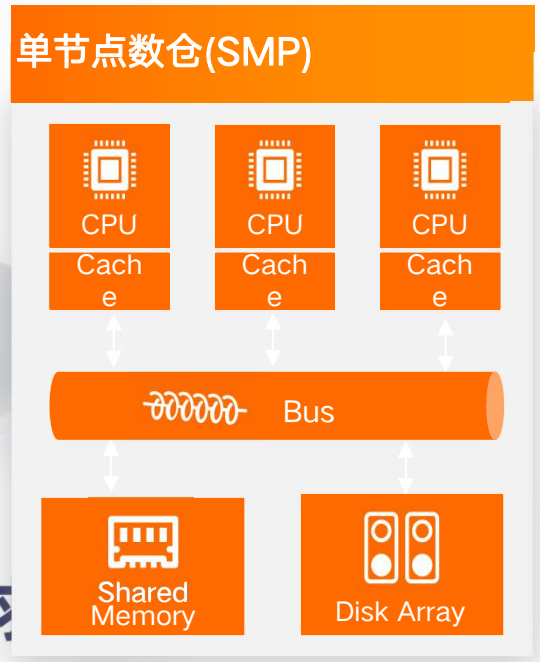
阿里云、中国信通院云大所联合发布《Serverless数据库技术研究报告》

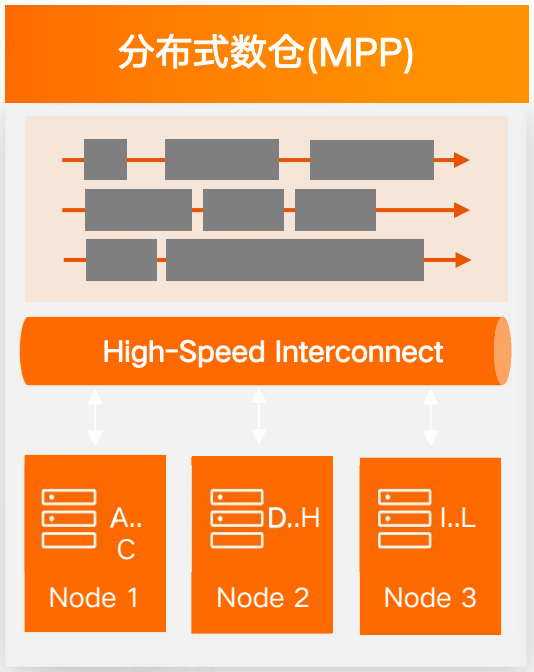
02 阿里云云原生数仓 AnalyticDB架构演进

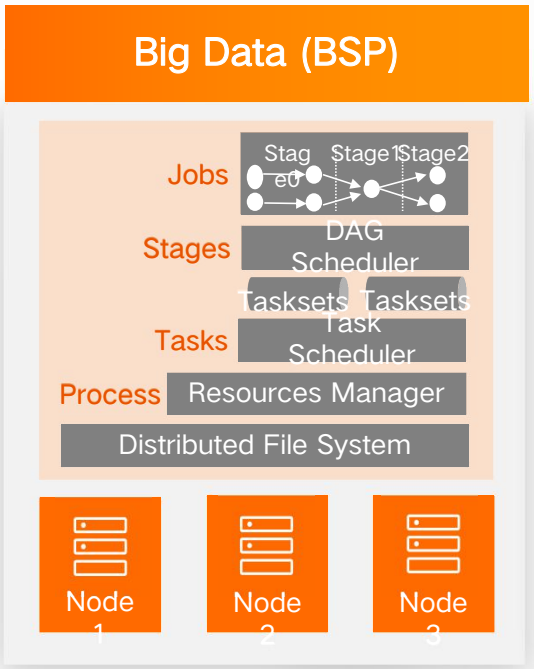
OLAP技术趋势：加速向云原生+离在线一体化演进

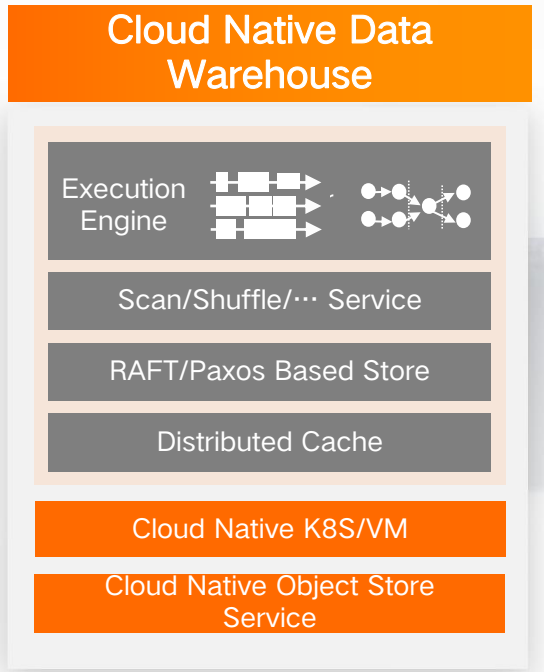
GOTC

1970	1993	Google 2005	Snowflake 2014
Relational [SQL+OLTP]	12 Rules [OLAP (R+M+H)]	GFS+MR+BigTable [Big Data]	Cloud Native+Serverless [Cloud DW]
✓ Oracle ✓ IBM DB2 ✓ SQL Server ✓ Informix	✓ Cognos ✓ ESSBase ✓ MicroStrategy ✓ Greenplum	✓ Impala ✓ Spark ✓ Dremel ✓ AnalyticDB V1-V2	✓ AWS Redshift ✓ Snowflake ✓ Azure Synapse ✓ AnalyticDB V3.0

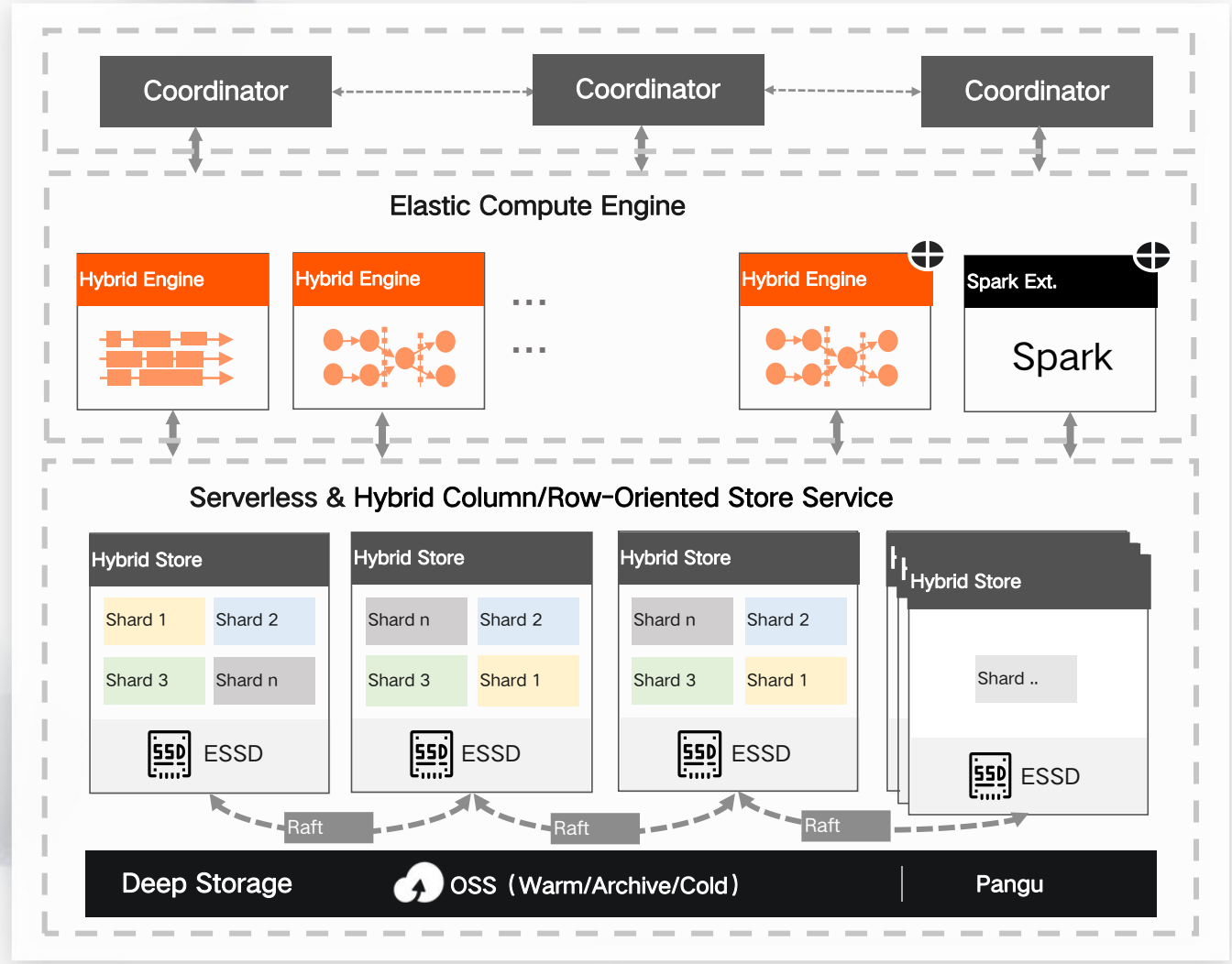








云原生数仓AnalyticDB：云原生+离在线一体化架构



生态兼容性

- 兼容MySQL，Oracle
- 完全兼容PG
- 完全兼容Spark

Teradata 迁移能力

在离线一体化

MPP + BSP 融合执行模型

- 一个引擎一份资源池同时支持 Ad hoc 和 ETL

混合负载

- 基于Page的实时计算调度
- 支持低延时Partition并行IO、高吞吐Block并行IO

高效执行引擎

- 支持Vectorized Execution、Code Gen
- 支持Operator Cache，支持统一内存池

实时CRUD

- 无需中间ETL，实时同步源库

融合存储

- 支持结构化、非结构化数据

云原生

自适应

- 分层存储：最低120元/TB/月
- 索引/分区支持自调整

算子级云原生

- Shuffle/Scan/Cache 服务化

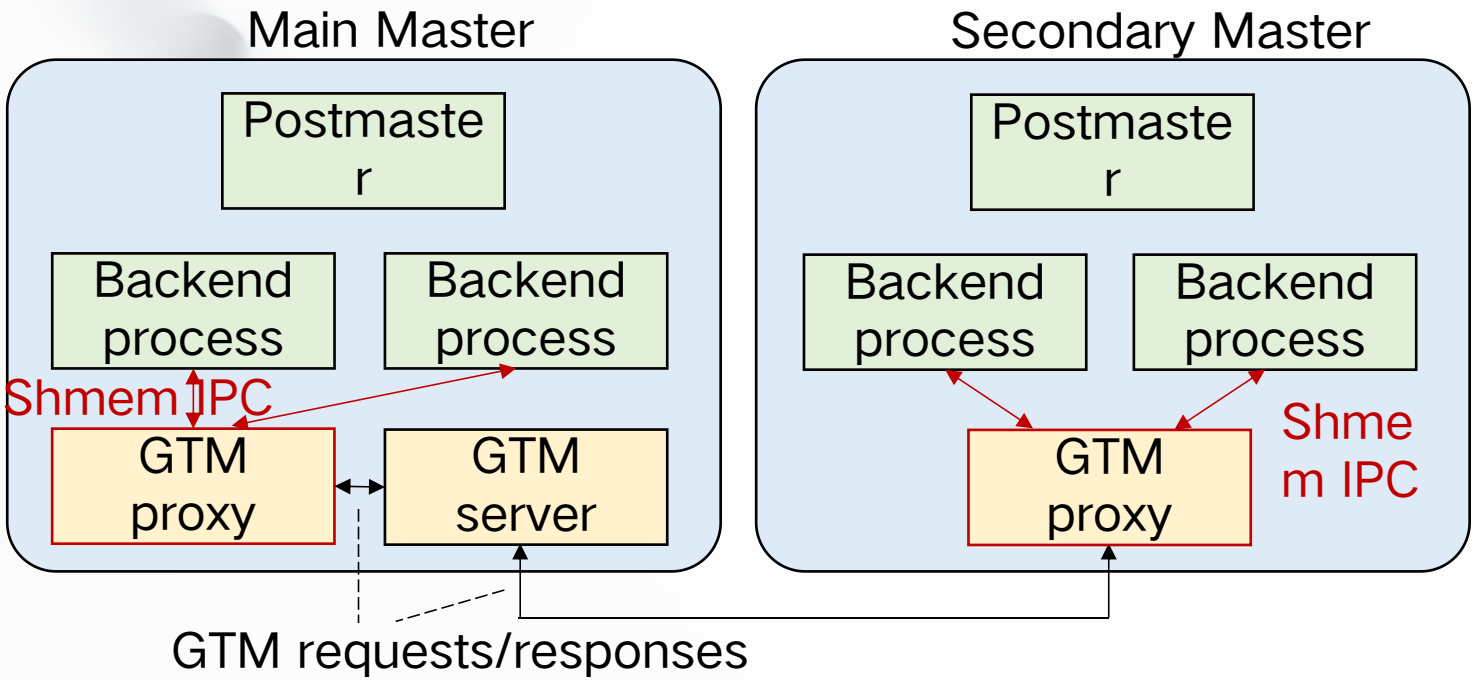
规模弹性

- 存储：支持扩展到100PB
- 计算：支持扩展到2000 节点

Serverless

- 存储：按实际使用付费
- 计算：基于负载Query级弹性

ADB-PG多 Master 分布式事务架构



多主：

- 多主GTM Server
- 多主GTM Proxy
- DDL同步
- 事务两阶段提交
- DTX恢复
- 死锁检查
- 三副本高可用

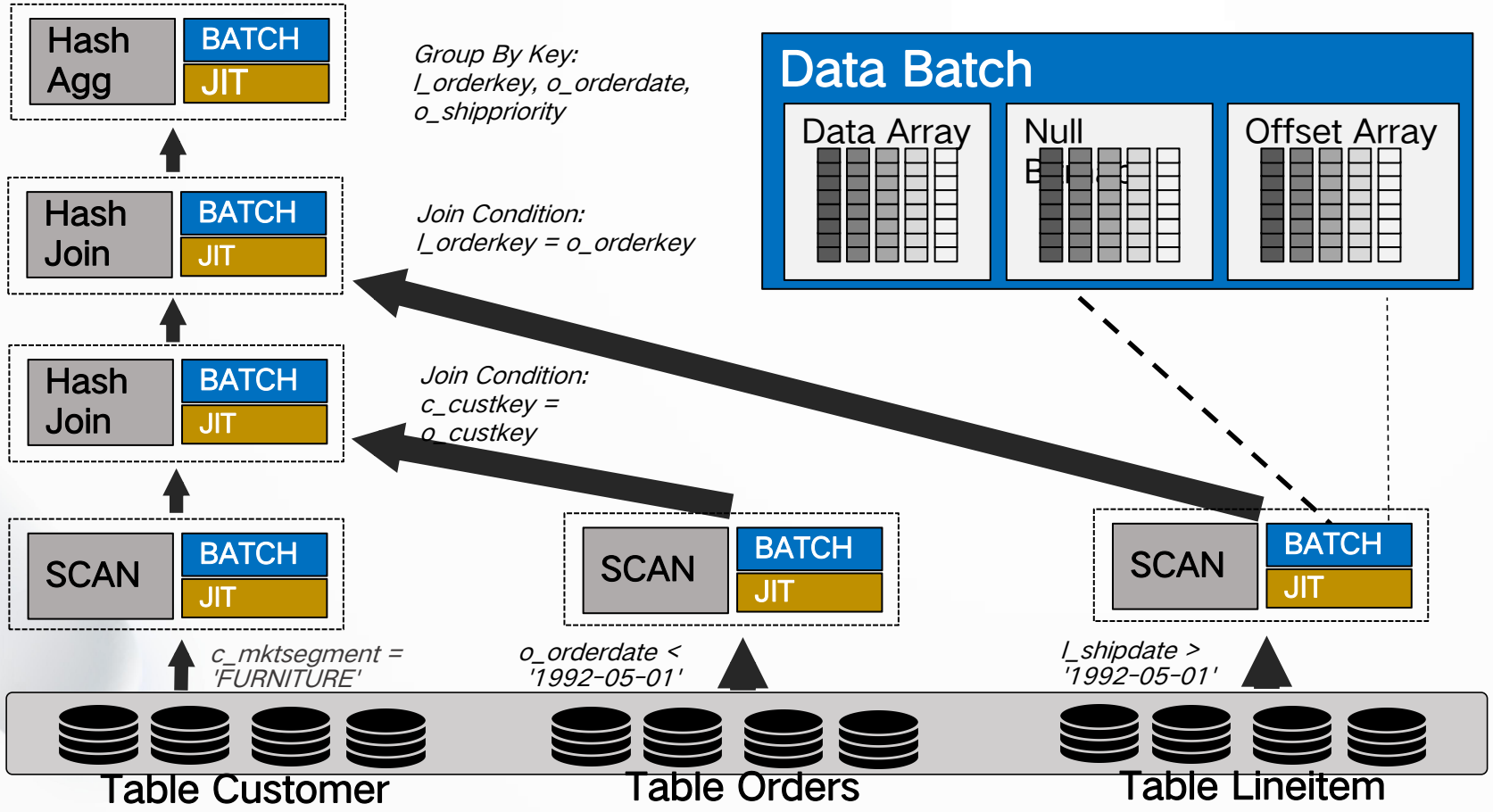
GTM管理(txn id counter，proc array, register node, ect.)
timestamp-gxid

- timestamp为main master的GTM server的启动时间戳，通过GTM连接握手以及FTS探活时同步到所有master
- gxid由GTM server统一分配

公有云支持：

- OSS
- 管控能力
- 安全能力

ADB-PG Laser 计算引擎



向量计算引擎（Q1，Q18）

行列式内存模型

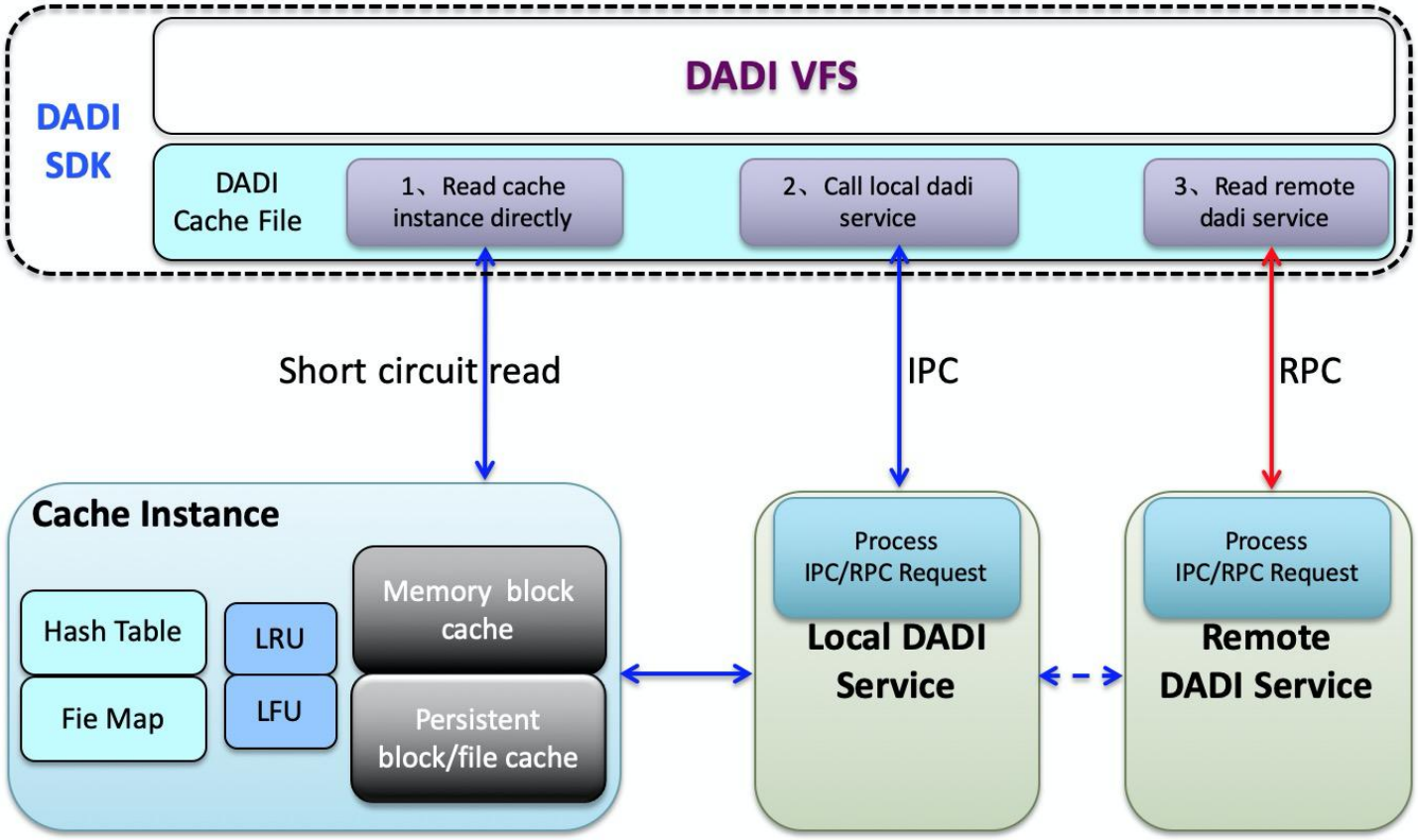
JIT 加速（LLVM，Spark）

SIMD 指令加速

FUSION SCAN

2020 年5 月20 号，TPC-H 30TB 场景测试，拿到了世界第一的成绩。相比于第二名微软SQL Server 2019，整体性能提升了290%，且成本只有SQLServer 的1/4。

ADB-PG DADI缓存



维度	RT		Throughput	
产品	DADI	Alluxio-Fuse	DADI	Alluxio-Fuse
命中内存	6~7 us	408 us	单线程： 4.0 GB/s 四线程： 16.2 GB/s	2.5 GB/s
命中磁盘	127 us	435 us	四线程： 541 MB/s	0.63 GB/s

03 阿里云云原生关系型数据库 PolarDB架构演进

云原生数据库PolarDB，自研创新，步履不停

PolarDB是阿里云自研的云原生数据库，100%兼容MySQL和PostgreSQL ，高度兼容Oracle语法，基于云原生架构、存储计算分离、软硬件一体化设计，支持云原生的集中式和分布式部署形态，为用户提供具备极致弹性、超高性能、海量数据、高可用、高性价比的数据库服务。

2个形态	3个生态	云原生数据库，覆盖传统关系型数据库全场景	
集中式	MySQL	PolarDB for MySQL PolarDB-M 云原生数据库PolarDB MySQL版	• 阿里巴巴自研的新一代云原生关系型数据库 • CPU、内存、存储三层解耦，软硬件结合 • 极致弹性、高性能、海量存储（100TB）、安全可靠 • 100%兼容MySQL 5.6/5.7/8.0
	PostgreSQL Oracle	PolarDB for PostgreSQL PolarDB-PG 云原生数据库PolarDB PostgreSQL版	• 100%兼容PostgreSQL数据库，高度兼容Oracle语法 • 基于分布式架构和普通PC服务器，提供与商用数据库相当的能力 • 具备高可用、高可靠、线性扩展、低成本、高性能等核心技术优势
分布式	MySQL	PolarDB for Xscale PolarDB-X 云原生数据库PolarDB分布式版	• PolarDB分布式版本 • 融合分布式SQL引擎与分布式自研存储 • 面向海量数据存储、超高并发吞吐、复杂计算与分析

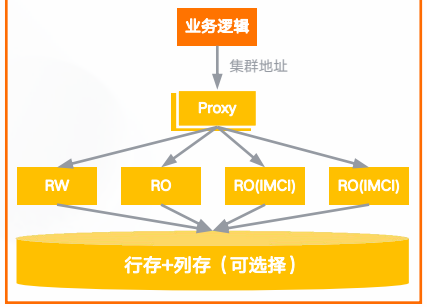
云原生数据库PolarDB，自研创新，步履不停

GOTC

1 云原生HTAP

- 一体化云原生HTAP产品，极简运维
- 内存列存索引 (In-Memory Column Index, IMCI）加持，MPP + 列存技术，实现分析性能百倍加速

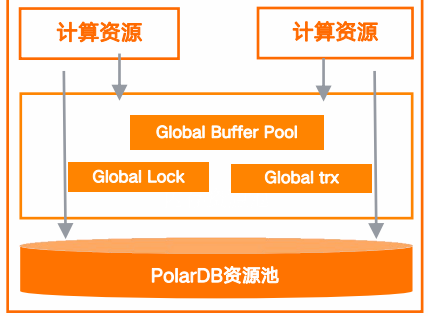
原生HTAP同时提供交易和分析能力



2 Serverless

- 资源秒级扩缩容，根据负载与资源动态匹配的按量付费，对于有间歇、潮汐等场景的可节省大量成本

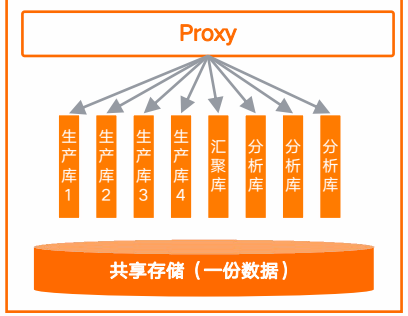
三层解耦，按需收费



3 多主多写

- 多主多写架构，最大支持32节点同时写入，更好的性能和扩展能力
- 跨节点动态调度，故障秒级完成切换。

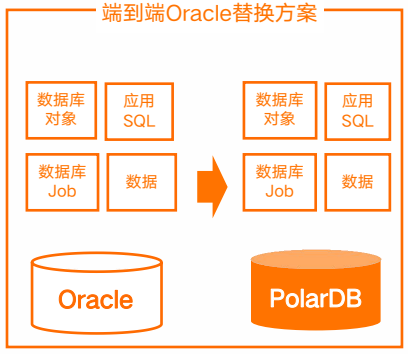
原生多生产库和汇聚库能力



4 企业级能力

- 100%兼容MySQL/PG，高度兼容Oracle语法、同城/异地容灾，完善的去O方案
- 支持多样化的芯片和操作系统，全栈国产化
- 开源开放真正意义上避免“Lock In”

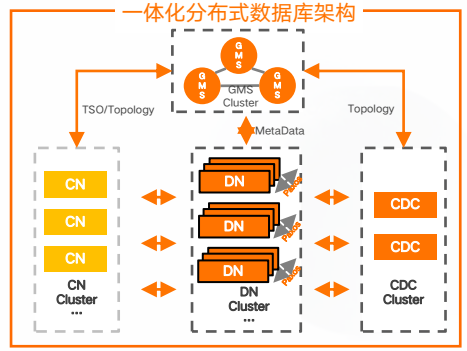
端到端Oracle替换方案



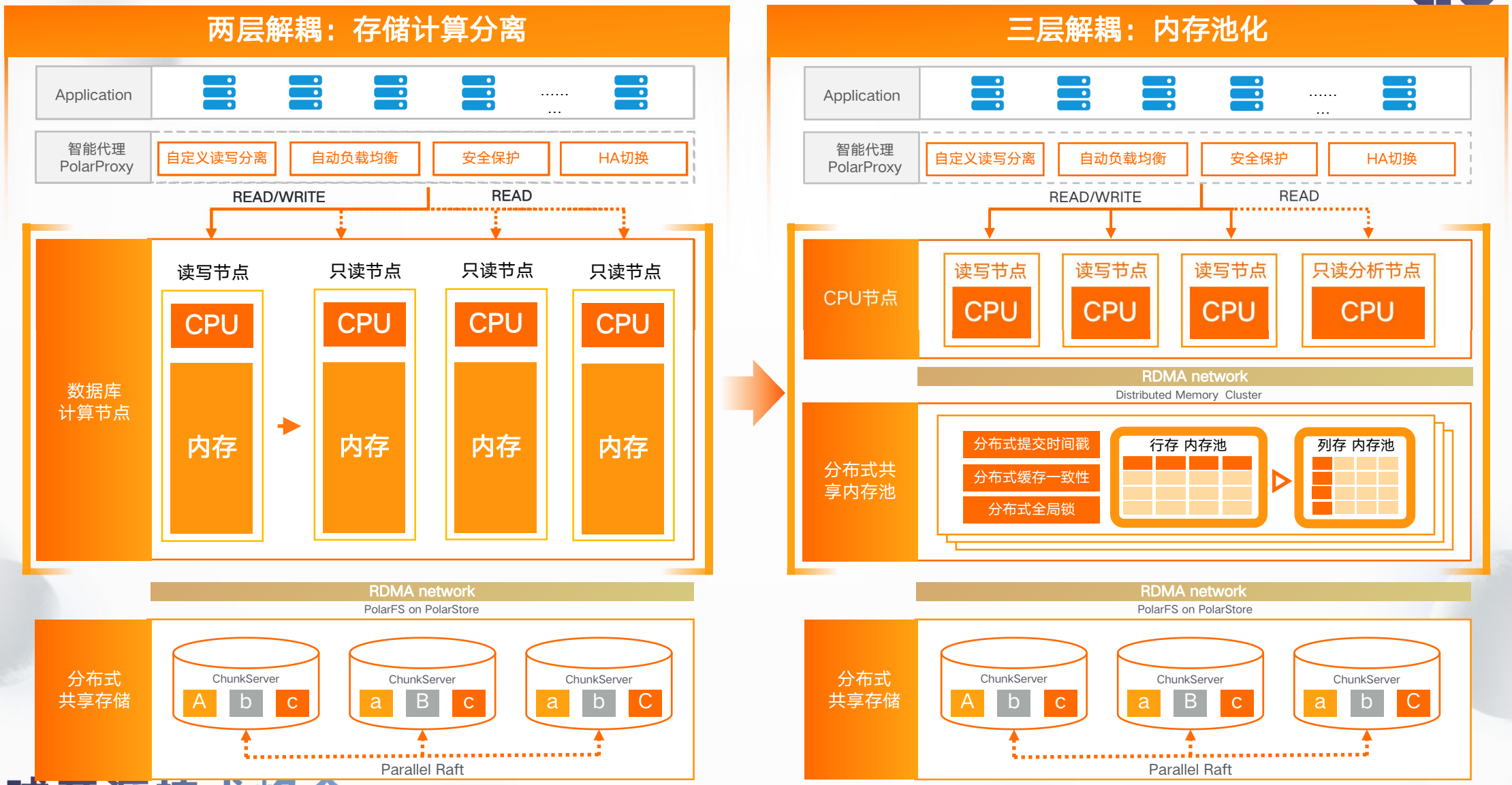
5 云原生分布式

- 云原生分布式数据库，原生MySQL生态
- 支持自动负载均衡、分布式高可用、分布式事务、全局二级索引等重要分布式特性
- 透明分布式提供单机MySQL的用户体验

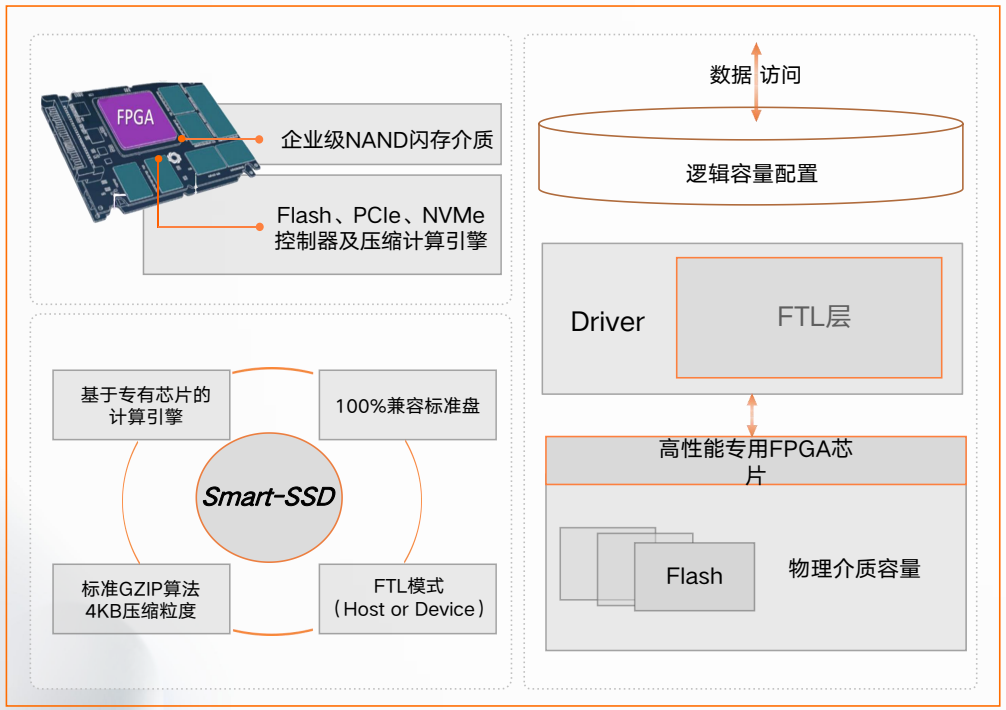
一体化分布式数据库架构



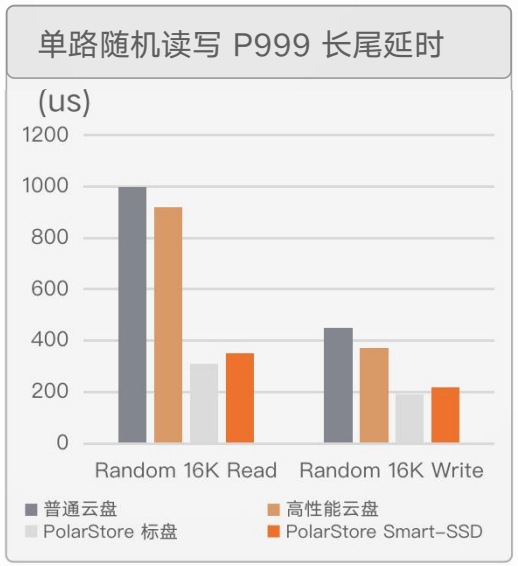
自研创新 PolarDB：三层解耦

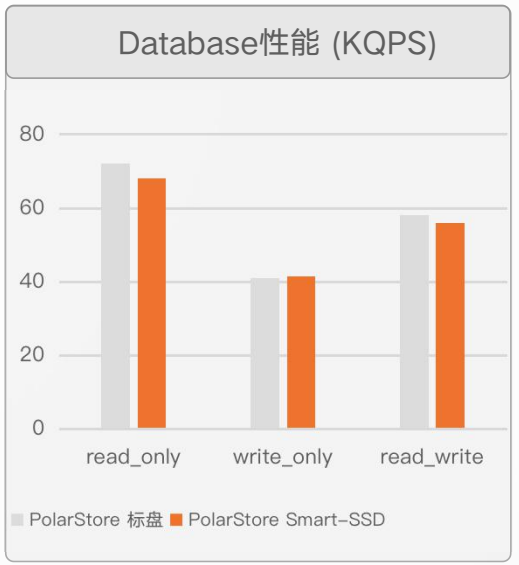


自研创新 PolarDB：软硬协同创新



PolarDB on Smart SSD 性能数据





3X
常规压缩比

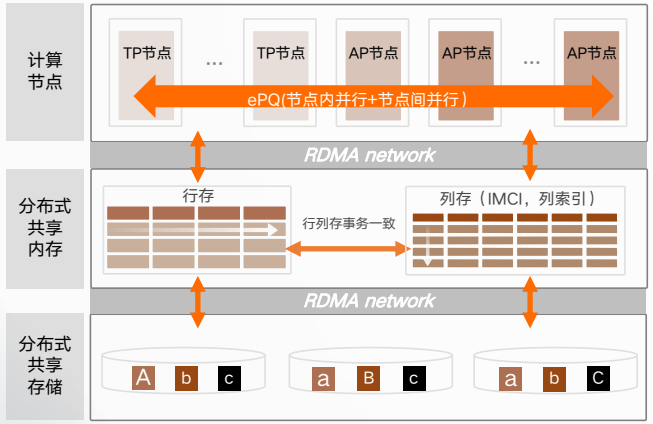
0
DB性能影响

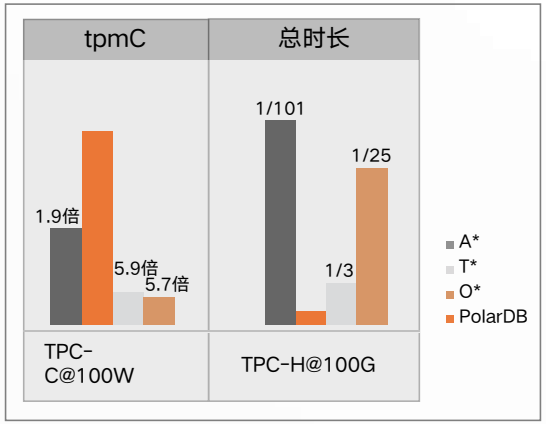
60%
成本节省

*依据标准FIO基准测试和实验环境数据，最终效果以实际产品和场景测试数据为准。

自研创新 PolarDB：HTAP，事务处理与计算分析一体化

云原生HTAP：PolarDB on IMCI



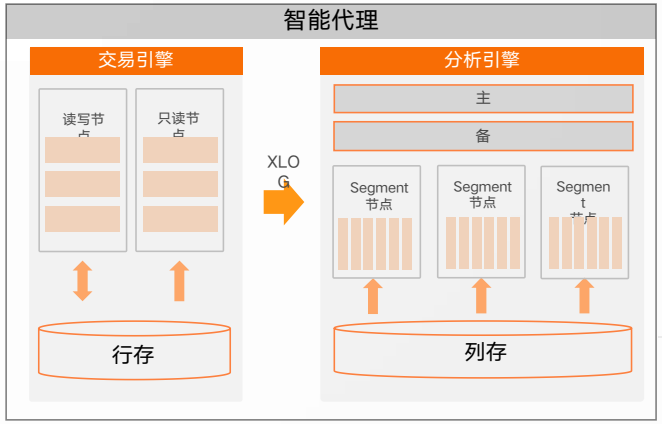


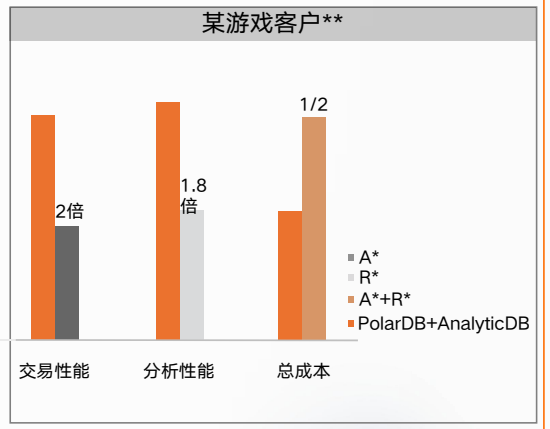
100%兼容MySQL

交易与分析0时延

TPC-C、TPC-H全面优于A*、T*、O*

一体化HTAP： PolarDB+AnalyticDB





一体化购买、使用体验

资源严格隔离

50%的成本，1.8倍的性能

*基于客户真实场景测算，最终效果以实际产品和场景测试数据为准。

04　开源PolarDB-X 架构介绍

自研创新 PolarDB：全面开源

携手生态伙伴为用户创造价值，推动开源协作与人才发展

全面兼容PostgreSQL与MySQL

开源PolarDB-X V2.2

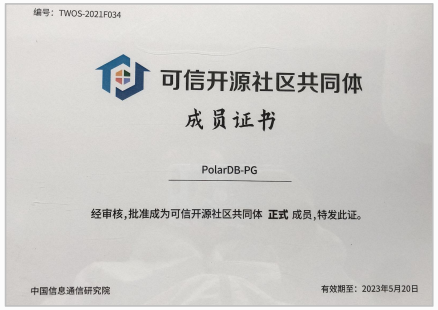
- 数据强一致性：RPO=0, TPCC提升23%
- 冷热数据分离：存储成本节省80%
- 全面兼容开源工具：canal/maxwell等
- 企业特性：MPP、闪回、存储过程、审计、容灾

开源PolarDB for PostgreSQL V11

- 三节点：2.5副本, 存储节省30%
- 存算分离集群：云原生HTAP
- 安全加密：支持TDE
- 性能优化：TPCC提升20% TPCH提升30倍

持续增长的用户与开发者社区

- 社区用户与贡献者: 21K+
- PolarDB开源大赛: 230名开发者参赛
- 荣获2022 OSCAR 尖峰开源社区奖项



全链路伙伴生态

开源PolarDB-X V2.2

- 50+生态伙伴: 韵达、莲子数据、网易数帆、龙蜥、武汉大学
- 11个SIG、7个合作项目、1个联合实验室 (韵达)

芯片　存储　操作系统　安全　人才培养　行业应用

全面兼容PostgreSQL与MySQL

10+内容栏目:

- 内核开发
- 应用开发
- 运维管理
- 架构咨询
- 教学内容: 100小时+
- 学习人次: 150万+
- 实验人次: 2万+
- 持证人才: 2000+

阿里云数据库PolarDB开源人才培养计划

PolarDB-X 兼容 MySQL 生态

工具

(MySQL生态兼容)

DTS
数据库迁移/容灾/多活

DMS
数据库流程/权限管理

DBS
数据库跨云备份

DAS
数据库自动化运维

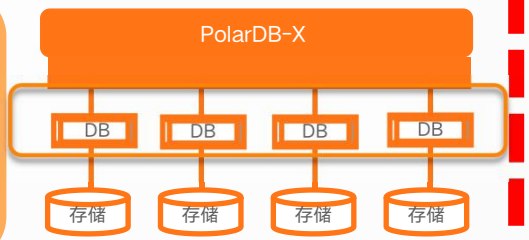
MySQL

（引擎按需转换）

PolarDB-X（原DRDS品牌升级）云原生分布式数据库

1.0版（计算层DRDS+存储RDS/PolarDB） / 2.0版（金融级一致性保障，透明分布式）

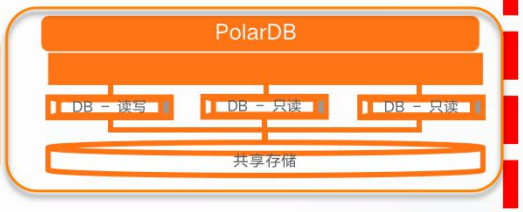
PB级数据+大表过亿+高并发大促+混合负载，上不封顶



PolarDB 云原生数据库

MySQL版（共享存储，一写多读）

超强弹性+百TB级容量+快照备份



RDS系列

MySQL/MariaDB

云上最强MySQL+高可靠+高安全

管控

（变速箱+仪表盘）

管控平台

高可用、读写分离、安全访问、实时变配、备份恢复…..

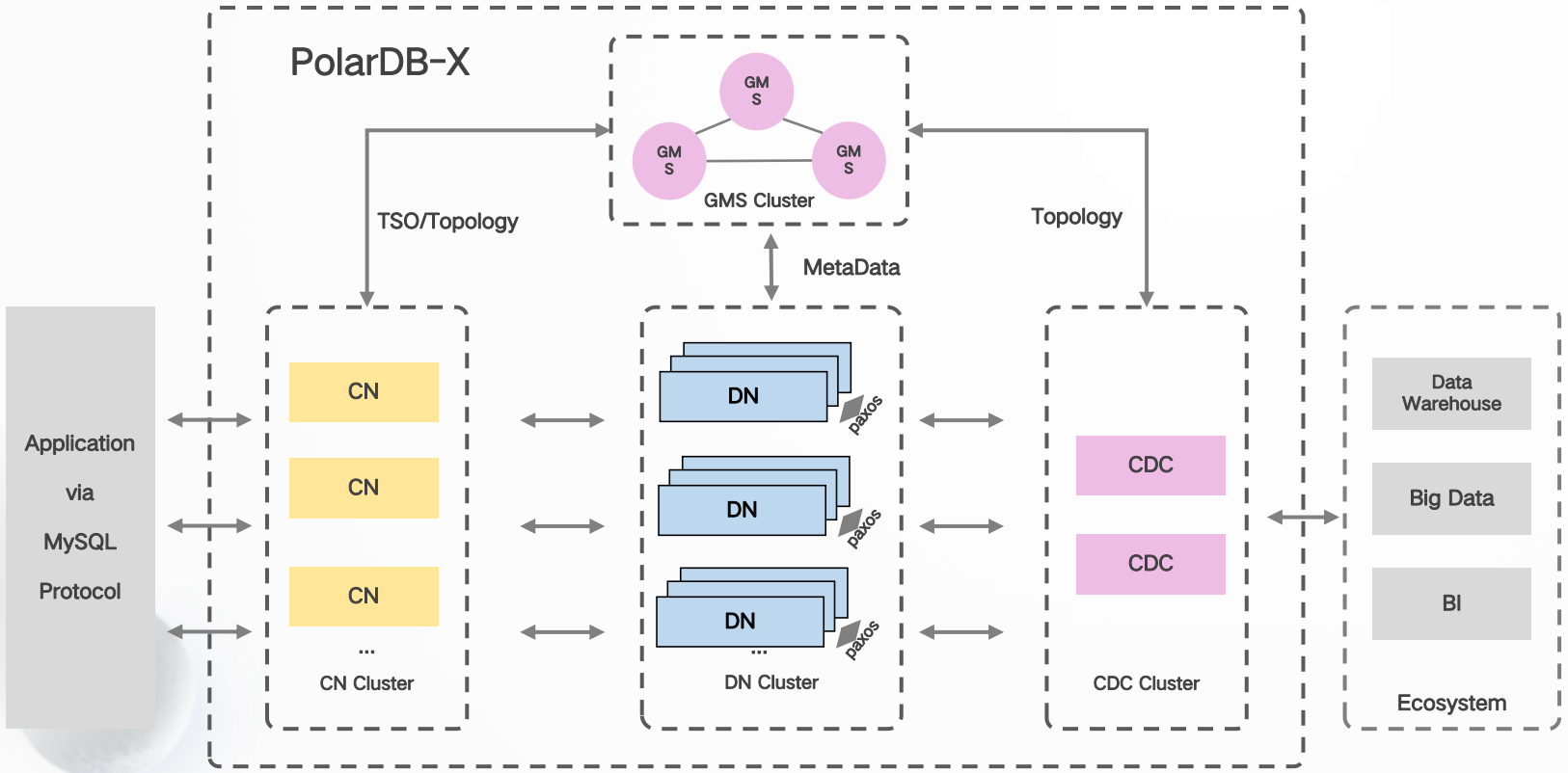
全球最大应用场景10年实战积淀

全链路监控和分析

秒级监控、告警、CloudDBA、SQL洞察、性能分析….

业界顶级DBA团队千锤百炼而成

PolarDB-X 技术架构



元数据服务（Global Meta Service，GMS）

- 提供全局授时服务(TSO)
- 维护Table/Schema、Statistic等Meta信息
- 维护账号、权限等安全信息

存储节点 （Data Node，DN）

- 基于多数派Paxos共识协议的高可靠存储
- 处理分布式MVCC事务的可见性判断

计算节点（Compute Node，CN）

- 基于无状态的SQL引擎提供分布式路由和计算
- 处理分布式事务的2PC协调、全局索引维护等

日志节点（Change Data Capture，CDC）

- 提供兼容MySQL生态的binlog协议和数据格式
- 提供兼容MySQL Replication主从复制的交互

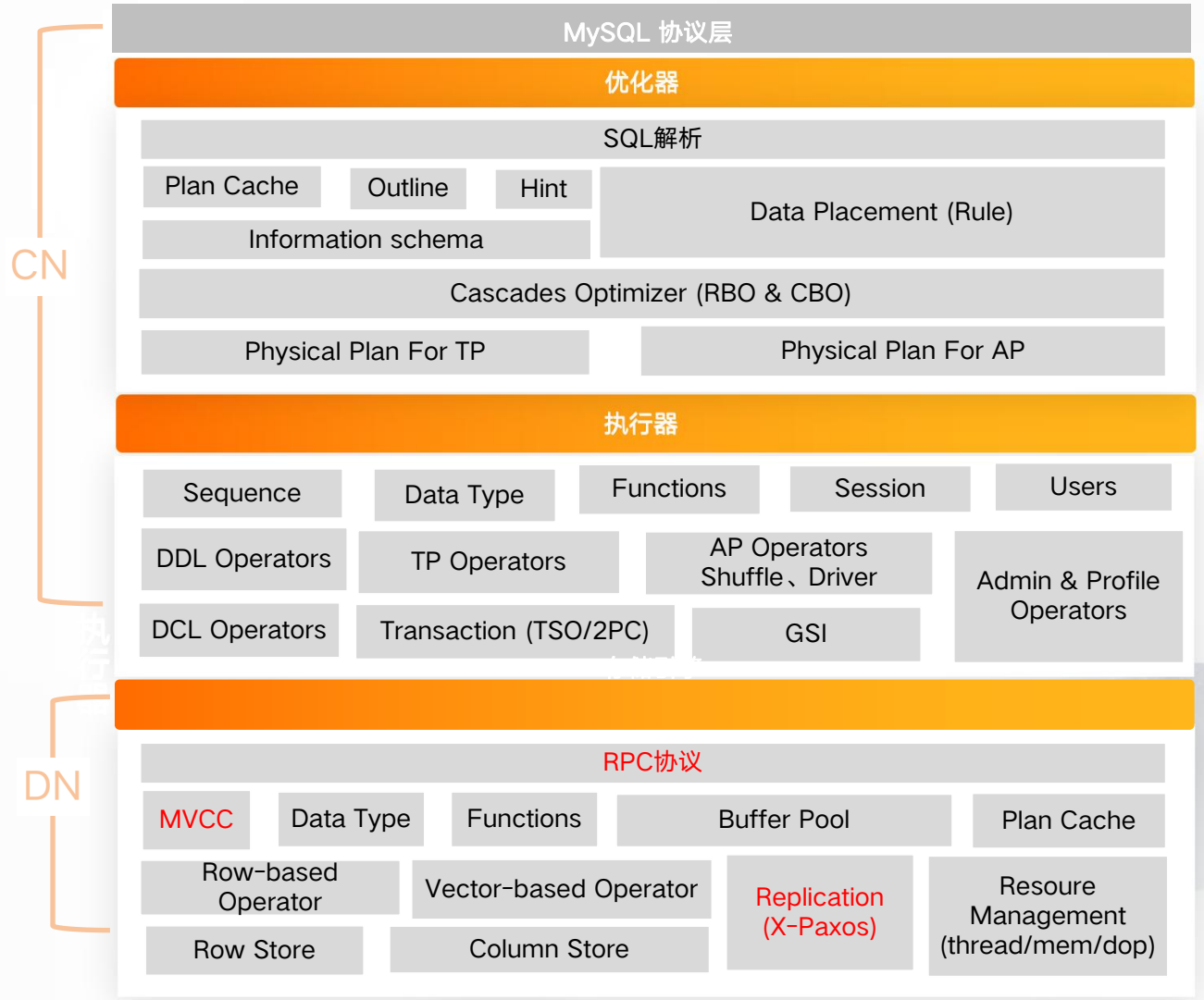
PolarDB-X ~ CN/DN组件

计算节点

- 经历多年实战磨练，MySQL语法高度兼容
- 完整的SQL解析层，实现精准算子下推
- Serverless无状态，弹性能力对业务透明
- 提供HTAP 并行计算能力，应对混合负载场景

数据节点

- 基于AliSQL内核，历经多年考验，稳定可靠
- 基于Paxos强一致协议，高可用能力进一步提升
- 全局MVCC改造，满足持金融级一致性要求
- RPC协议改造，提升节点间通讯性能



PolarDB-X ~ CDC组件

CDC节点

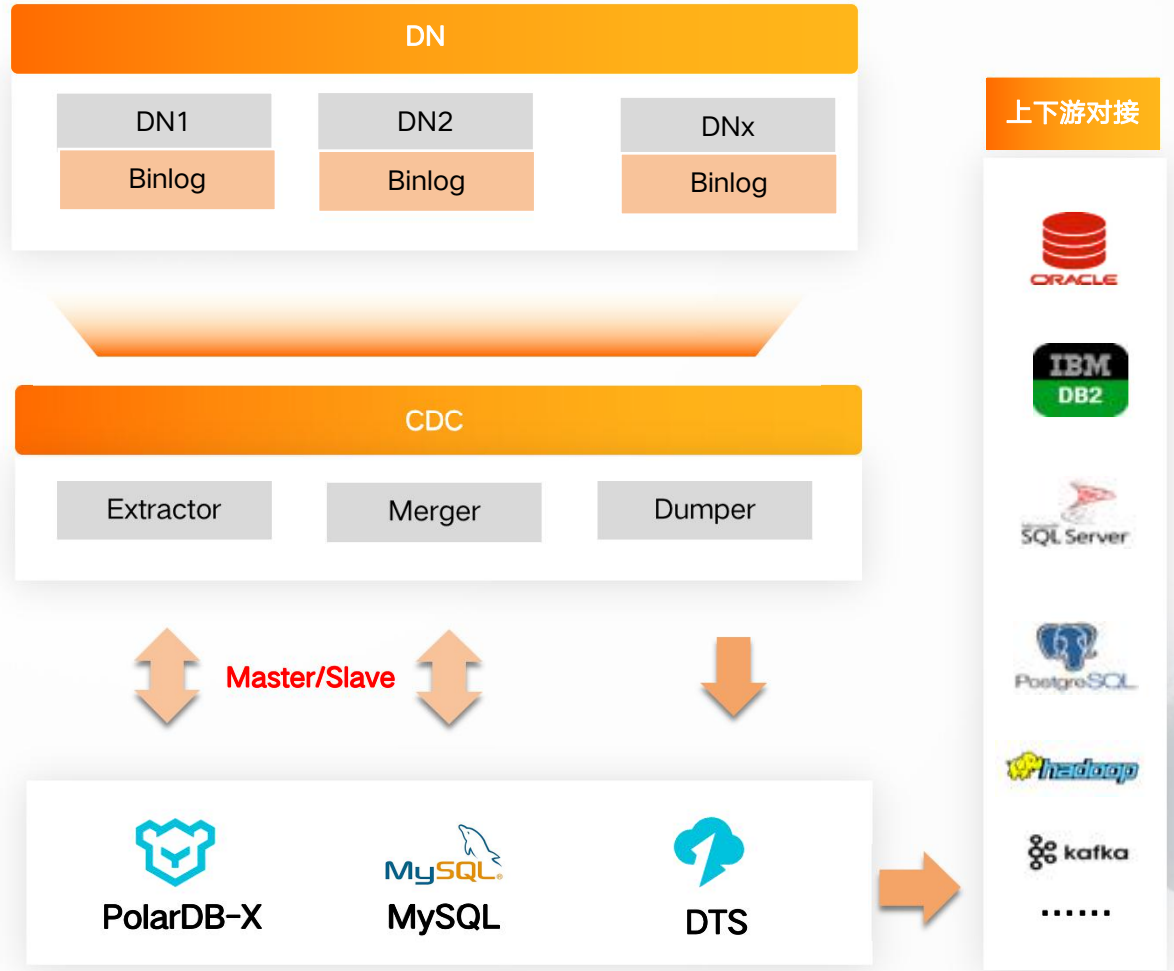
- EX：并行采集所有DN的变更日志
- MR：分布式事务日志/DDL排序重组
- DP：全局日志落盘并提供标准Binlog服务

全局Binlog

- 兼容事务 (分布式事务全局排序)
  例：基于TraceId、TSO信息对Binlog全局排序
- 兼容分布式DDL
  例：可支持DDL同步到下游，比如ADB
- 兼容分布式扩缩容
  例：屏蔽内部分片迁移、广播表、索引等数据干扰

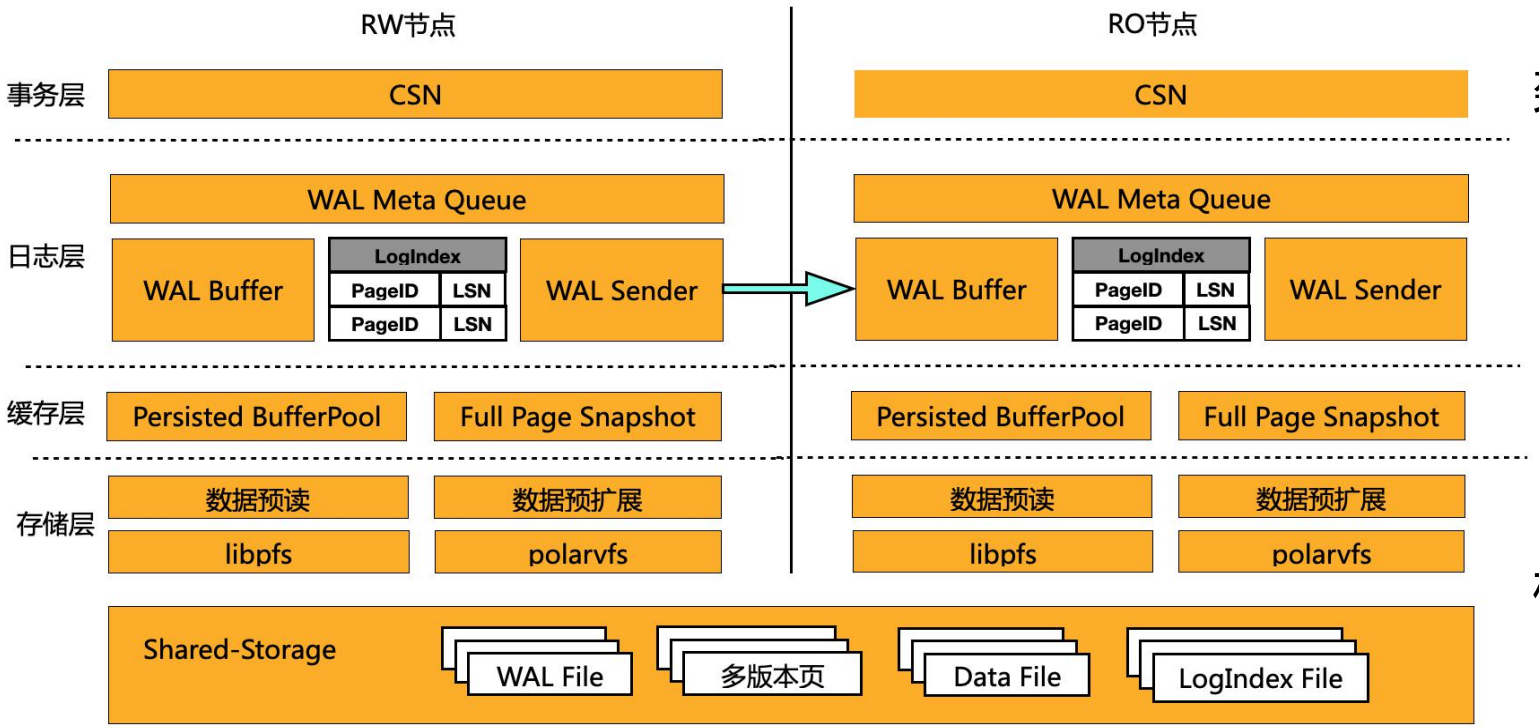
主备Replication

- 兼容MySQL生态的主备复制
- 兼容DTS的上下游生态



05 开源PolarDB-PG架构及代码解读

关键技术1 - 存储计算分离、存储层



文件存储：

1、共享存储（数据文件，WAL日志文件，pg_control等）

2、所有节点各自存储一份（pg_stat， postgresql.conf等）

3、master 读写共享存储，replica 本地维护一份（pg_xact等）

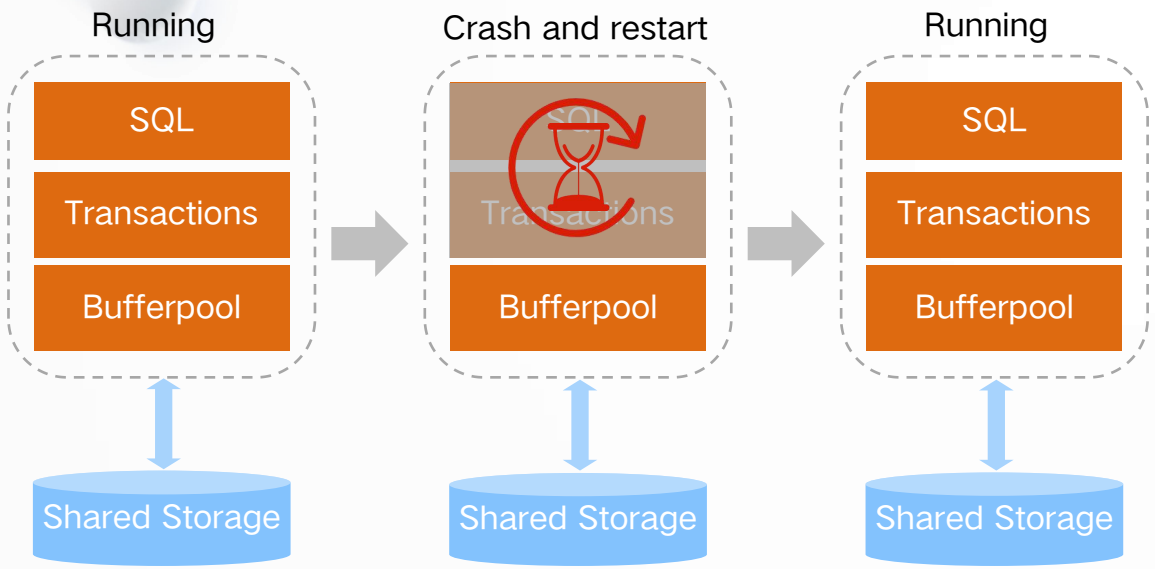
架构原理

- 事务层：CSN快照
- 日志层：复制WAL Meta/Lazy/并行回放
- 缓存层：常驻BufferPool/多版本页面
- 存储层：数据预读/扩展/PolarVFS

核心代码

- polar_vfs插件：Polar_vfs接口
- vfs_mgr结构：VFS存储访问接口
- polar_make_file_path_level3：
  文件位置定位
- POLAR_FILE_IN_SHARED_STORAGE：
  判断是否共享存储
- vfs_vfd_cache: 管理文件cache

关键技术2- 缓存层、Bufferpool



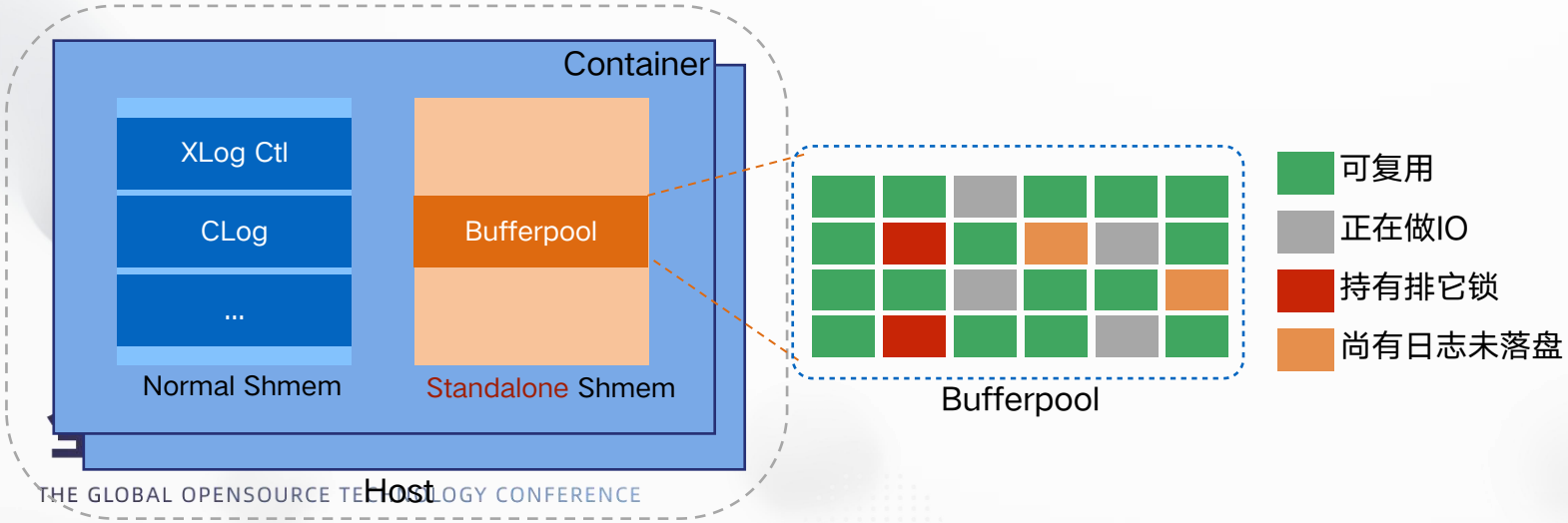
缓存层：

- 常驻BufferPool
- 多版本页面
- 刷脏控制（copybuffer，flushlist）

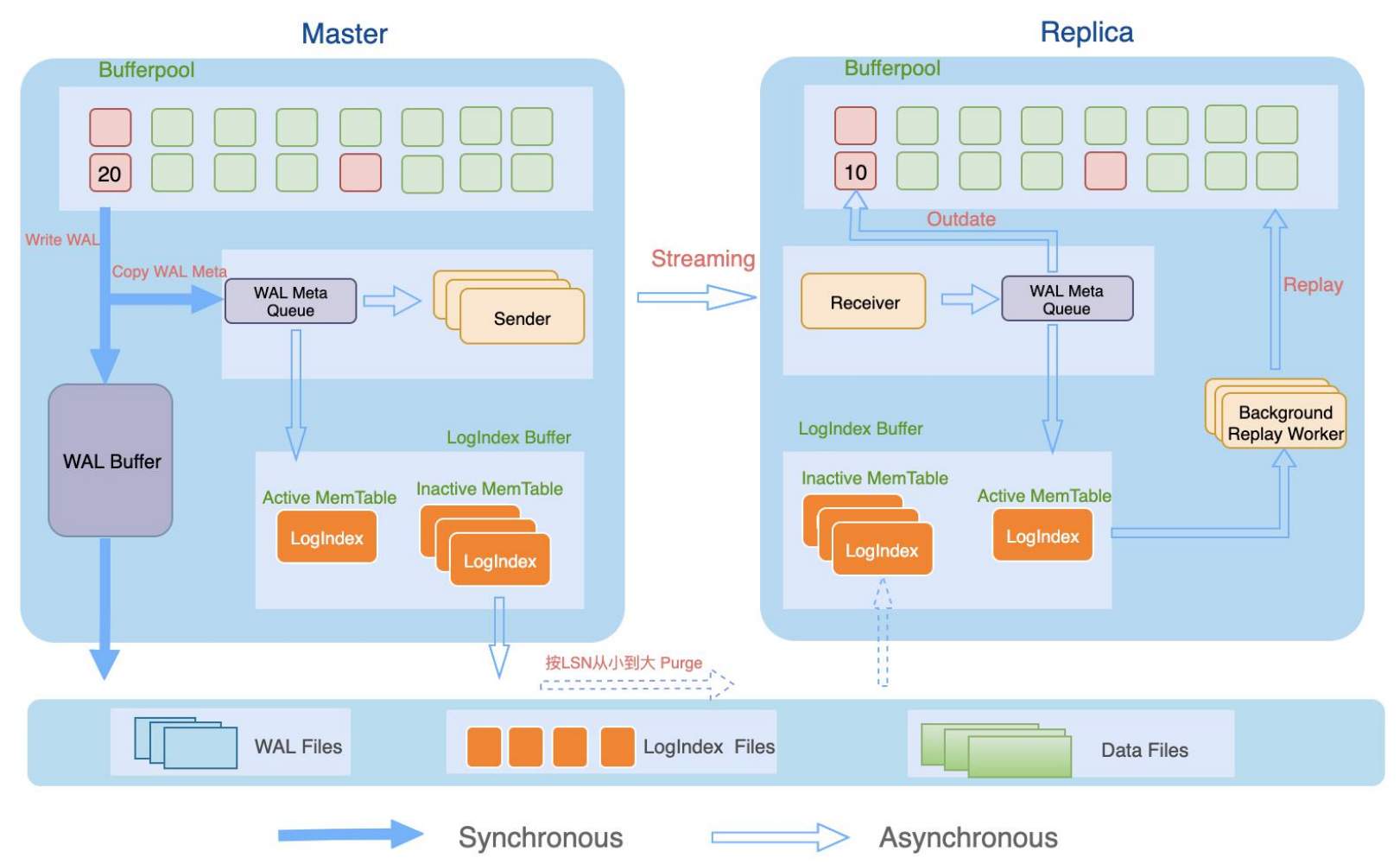
核心代码

- polar_buffer_can_be_flushed：刷脏判断
- BufferDesc：flush_next、copy_buffer
- polar_launch_parallel_bgwriter_workers：并行刷脏

POLAR_FILE_IN_SHARED_STORAGE：判断是否共享存储



关键技术3 – 日志层 、LogIndex



日志层：

- LogIndex
- 复制WAL Meta
- Lazy回放
- 并行回放
- 数据页Prefetch
- 异步DDL

核心代码：

1、PG_RMGR(symname,name,redo,polar_idx_save, polar_idx_parse, polar_idx_redo, polar_redo,desc,identify,startup,cleanup,mask)

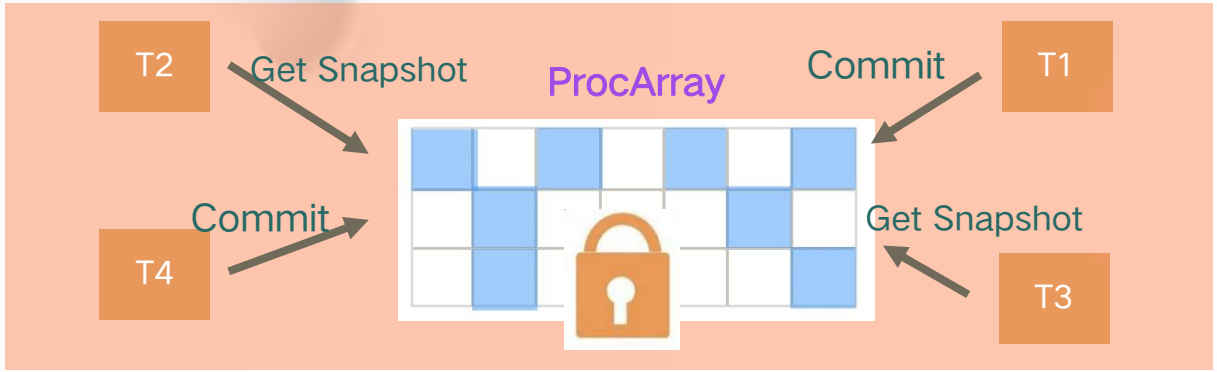
2、 XLogCtlData中consistent_lsn、 bg_replayed_lsn;

3、 log_mem_table_t、 log_idx_table_data_t

4、 polar_log_index_truncate_mem_table

5、 ReadBuffer_common中 polar_log_index_apply_page_from

关键技术4 – 事务、CSN快照

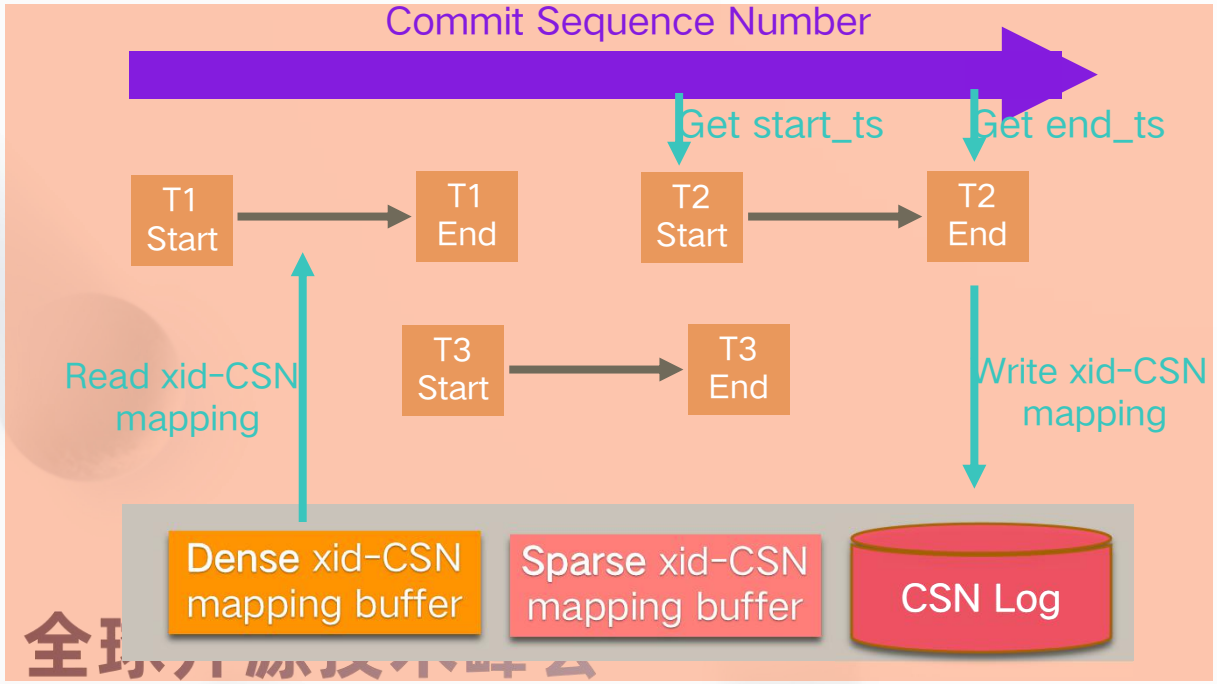


CSN MVCC原理：

- 用单调递增数值CSN替代Snapshot快照
- 维护事务ID与CSN的对应关系

效果：

- 消除Snapshot锁等待



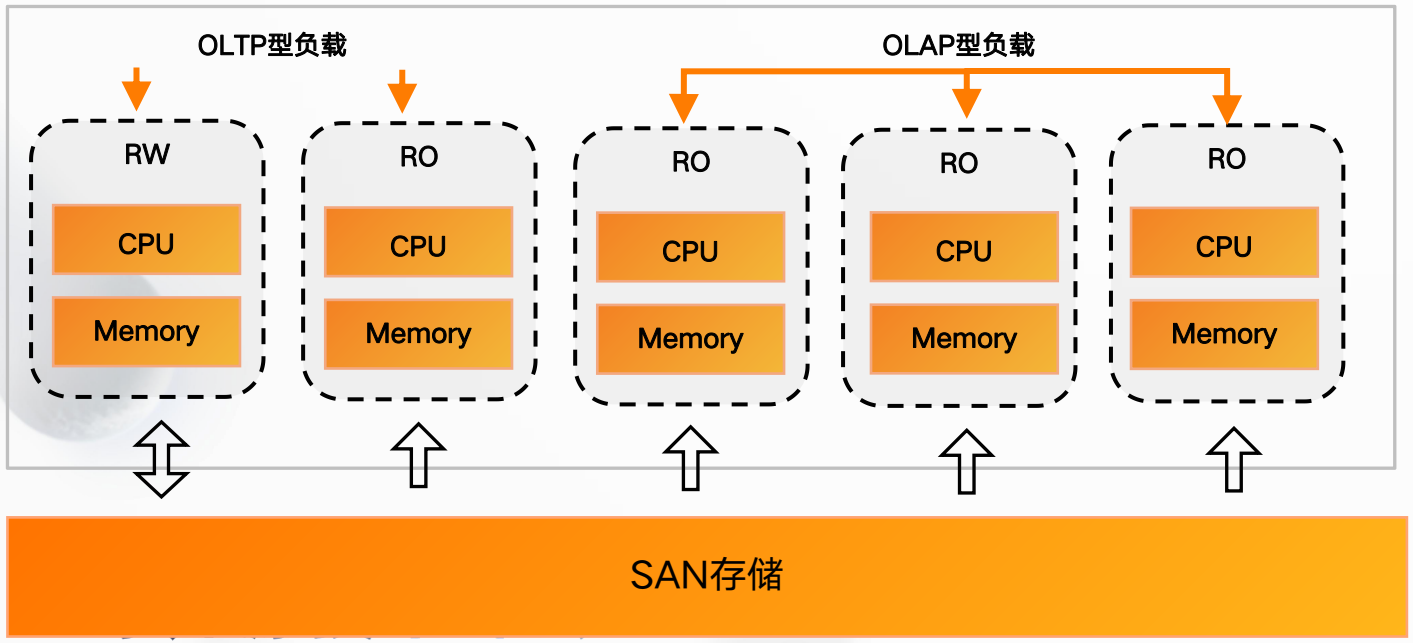
Percolator: Google在分布式Key-value store(BigTable)上利用TSO和行原子性操作实现分布式事务处理协议.
Large-scale Incremental Processing Using Distributed Transactions and Notifications（OSDI'2010）

HLC: 使用TSO(全局时间戳)对分布式事务进行定序，TSO会成为局部单点，影响扩展性，混合逻辑时钟根据事务相关性生成时间戳，解决扩展性问题。
Logical Physical Clocks and Consistent Snapshots in Globally Distributed Databases
CockroachDB: The Resilient Geo-Distributed SQL Database (SIGMOD'2020)

关键技术5- HTAP – 基于共享存储的MPP架构

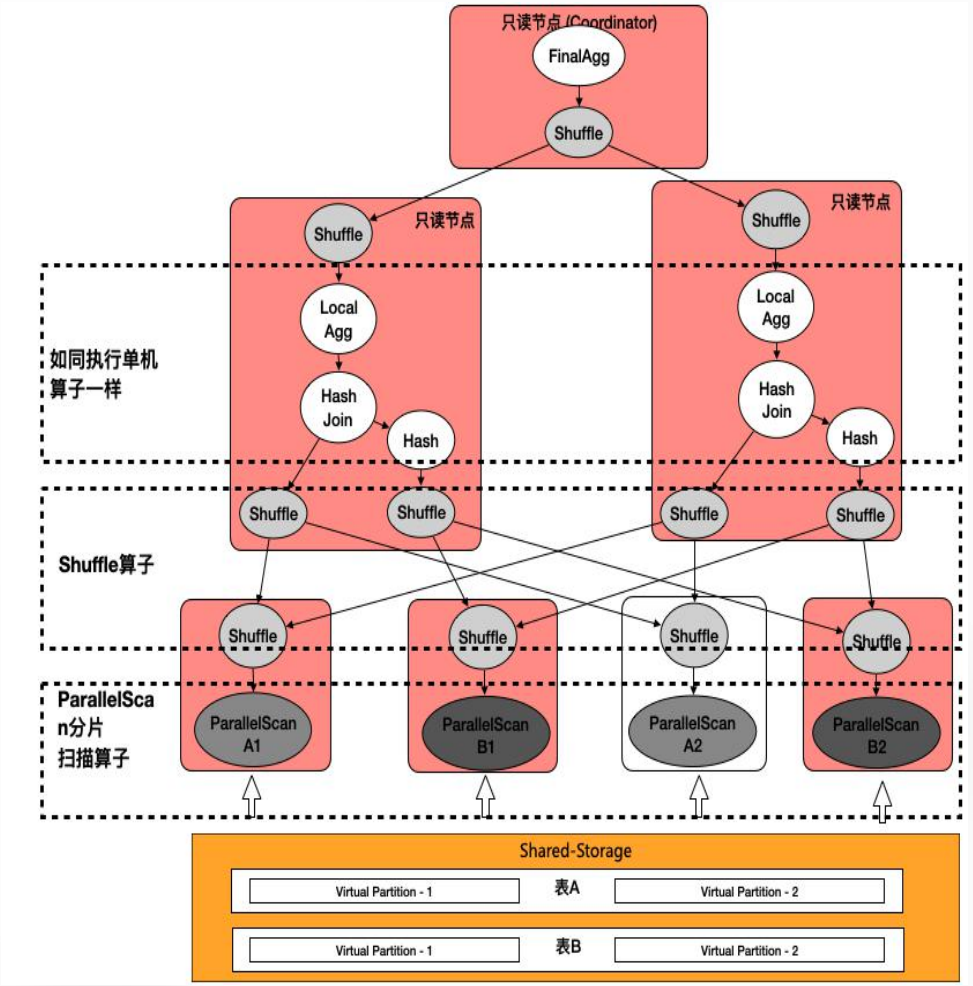
PolarDB HTAP

- 加速AP查询：发挥RO节点硬件资源
- 物理隔离：TP/AP隔离
- 一套数据，两套计算引擎



基于共享存储的MPP原理（首创）

- Virtual Partition
- Shuffle算子
- 单机算子并行化



开源PolarDB——打造世界级云原生数据库开源社区

GOTC

愿景：打造世界级云原生数据库开源社区

社区架构

由理事会统领，技术委员会（TOC）指导管理专项兴趣小组、用户组、其他组。

社区运营

生态伙伴建设

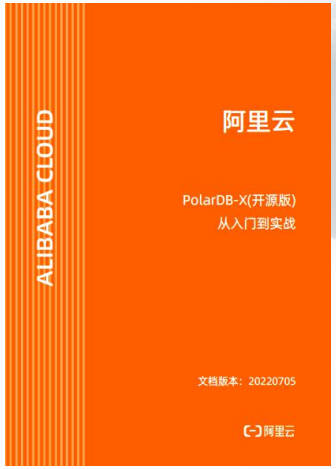
关注数据库全栈伙伴的建设；重点行业的头部伙伴合作，打造行业专属的云原生数据库。

面向开发者

鼓励开发者积极参与开源软件的开发维护，促进优秀开源软件社区的蓬勃发展

面向用户

通过理论与场景实验，让用户充分学习并使用开源产品。如动手实验室、训练营、课程等

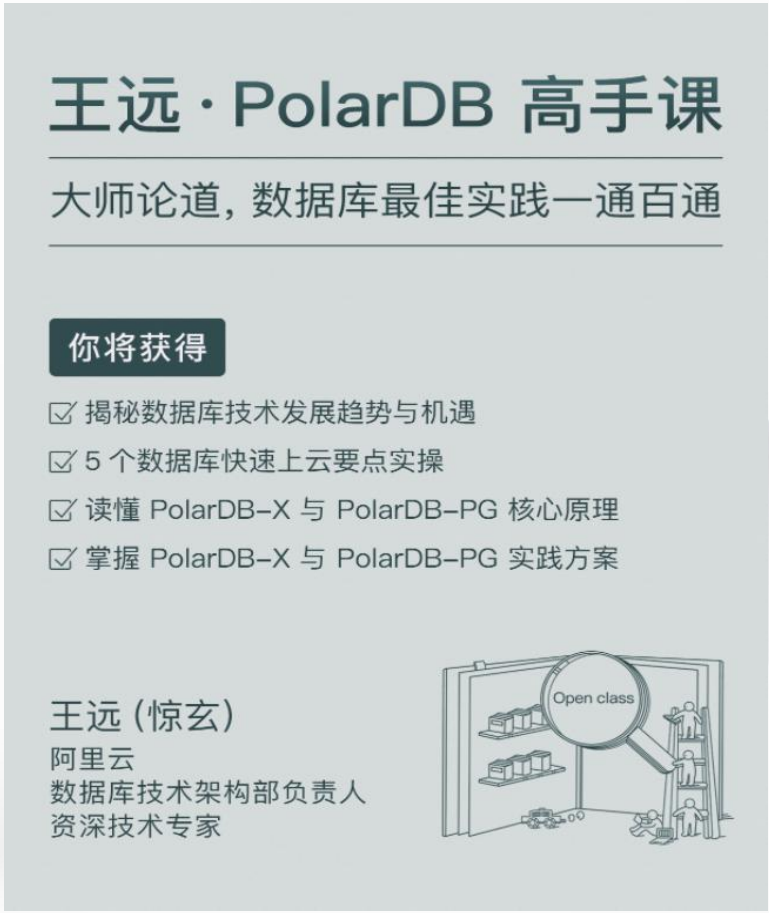


PolarDB开源社区正在开展的课程

本系列课程将面向DBA、高校学生、内核爱好者，介绍PG内核架构、各模块基本原理、用法、代码实现。

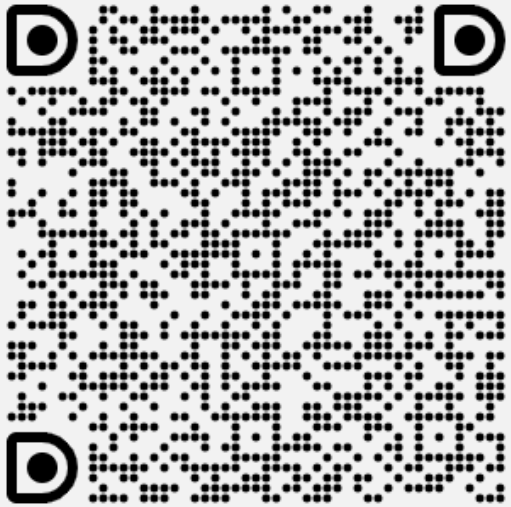
阿里云数据库专家携手高校教师带你从零开始系统化地学习数据库理论知识，由浅入深提升数据库实践能力。

本系列课程将对PolarDB开源技术和实践、上云方法论进行一个整体的解读。



加入PolarDB开源社区

PolarDB-X 开源交流钉钉群



PolarDB-PG 开源交流钉钉群

商务合作微信

内核交流微信

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE