DTCC

数，造，未，来

第十二届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2021

2021 年 10 月 18 日 - 20 日 | 北京国际会议中心

Spark
Hadoop
OLAP
HANA
MPP
Oracle
SQL

IT168.com ChinaUnix.net ITPUB

金山云分布式数据库DragonBase架构详解和实践

王天宇

内容提纲

- 分布式数据库发展背景
- 金山云DragonBase产品架构
- 金山云DragonBase关键能力
- 金山云DragonBase应用实践

1 分布式数据库发展背景

数据库系统演进

Oracle IBM DB2 SQL Server Sysbase Informix	MySQL Postgresql Teradata Sybase IQ Greenplum	Redis Hbase MongoDB Hadoop	Aurora Google Spanner CockroachDB/TiDB SequoiaDB	NEWDBs
商业进阶 (1980~1990)	开源+分析 (1990~2000)	NoSQL数据库 (2000~2010)	云原生分布式、 NewSQL、多模 (2010~2020)	软硬一体化 云化+分布式 +HTAP

在线关系计算

关系型数据库
RDBMS

在线/离线
数据抽取&分析

数据仓库
ETL+OLAP

多样化异构
数据计算

文档	KV
空间	向量
时序	结构化
图	半结构化

多模&混合计算
数据库统一服务

DBaaS

云原生分布式

HTAP

Muti Mode

软硬结合

数据库市场空间大，云化趋势明显

数据库市场规模保持年复合18%增长率，云数据库部署模式将成为主流，预计2024年占比75%

智研咨询，东兴证券研究所

来源：IDC中国，2020

2 金山云DragonBase产品架构

DragonBase发展里程碑与应用

DragonBase 单体版本 — 2016

- 上线公有云，
- 提供主备模式的高可用数据库服务

DragonBase 分布式版本1.0 — 2018

- 提供私有化部署
- 支持弹性扩展、分布式事务等基础能力

DragonBase 分布式版本2.0 — 2020

- 数据分布式强一致存储
- 内核性能优化
- 数据安全、运维监控、兼容性等

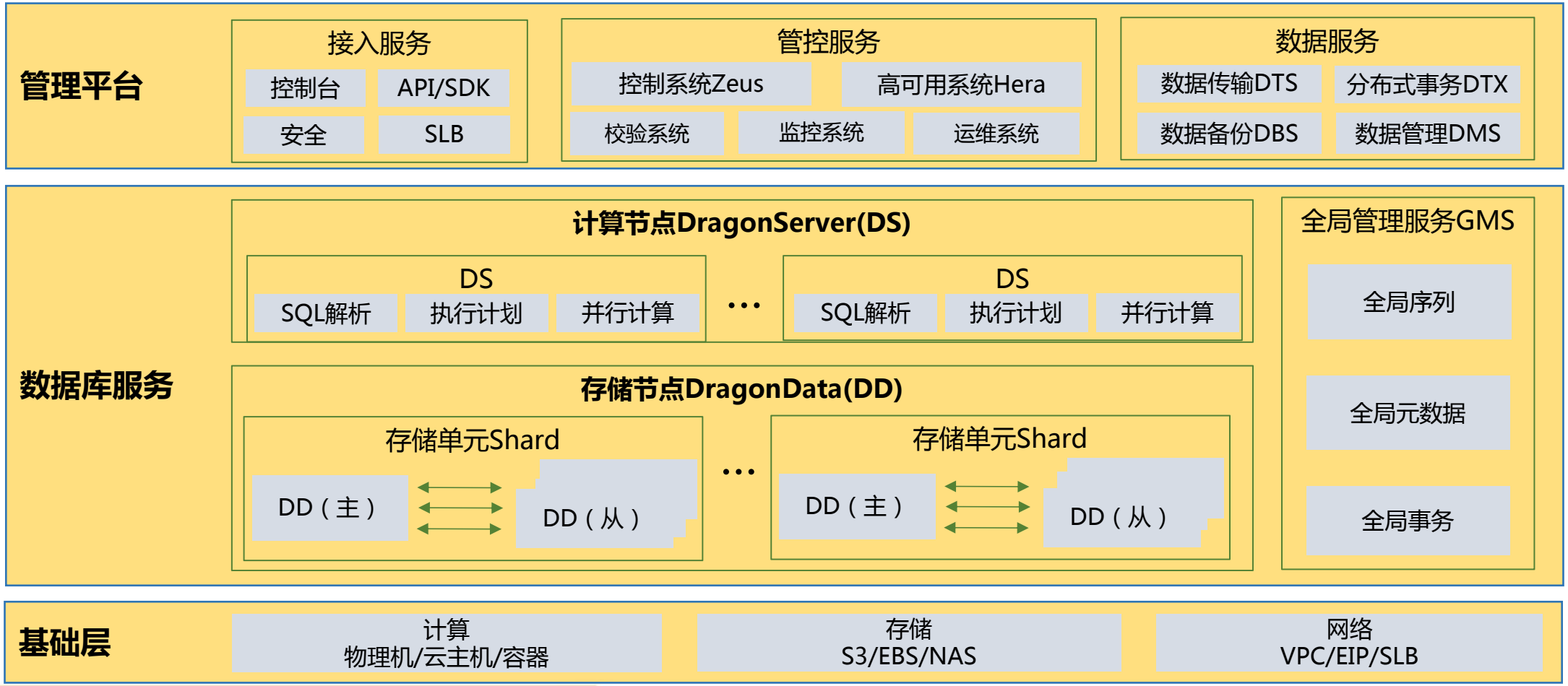
DragonBase 分布式版本3.0 — 2021

- 分布式一致性读
- 查询优化
- 一致性备份恢复
- 全局索引

……

DragonBase产品架构

DragonBase是能满足金融级业务需求，支持高可用、高可靠、高性能、高安全、可扩展、高兼容的分布式数据库。



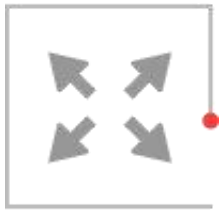
DragonBase主要功能特性

强一致

- 数据多副本实现强同步复制
- 缺省情况下读写操作在主副本进行
- 多重数据校验，保证数据不丢不错

高可用

- 数据采用多副本存储
- RPO -> 0；RTO<30s
- 实现同城双活、异地灾备



可扩展

- 水平扩展，在线扩容缩容，服务不停
- 单集群规模大，数据量百TB级
- 计算节点和存储节点均可扩展

兼容性

- 兼容MySQL /PG 功能及协议
- 业务零修改或少量修改即可迁移过来
- 兼容SQL标准和Oracle常用功能

高性能

- 采用计算存储分离，全链路优化
- 使用分布式并行计算技术
- 集群最大吞吐百万QPS

安全性

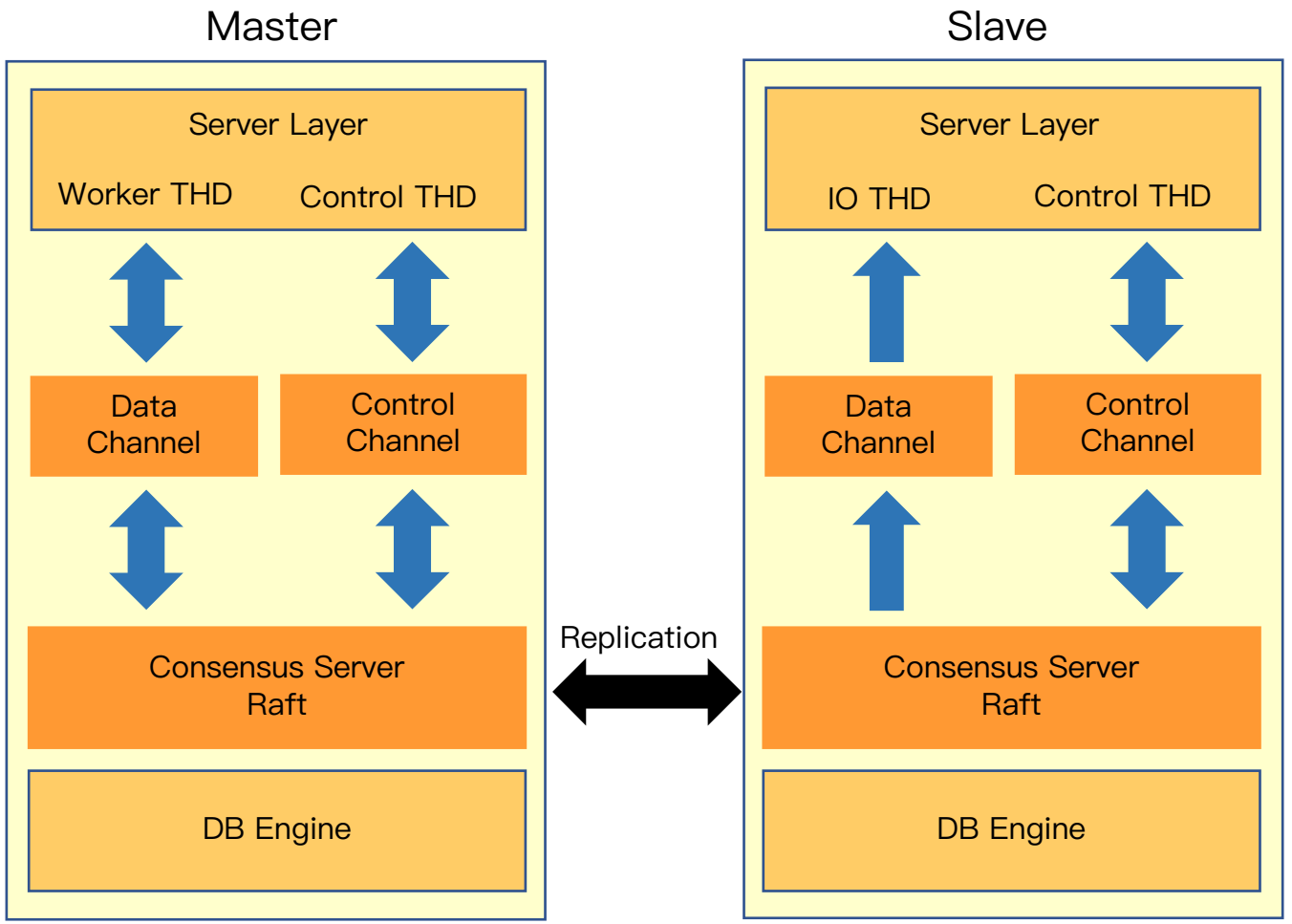
- 用户权限管理
- 传输加密、存储加密
- 提供安全审计功能

3 金山云DragonBase关键能力

DragonBase副本间数据一致性

- 主从副本间支持多种模式的数据复制方式：

- **异步复制：**
  - 性能好
  - 无法保证副本间的数据一致性
- **半同步复制：**
  - 正常场景下保证副本数据一致性；
  - 异常情况（如网络or节点问题等）下退化为异步，无法保证
- **强同步复制（On Raft）：**
  - RPO = 0
  - 高性能：多线程模型 + BATCH传输日志
  - 高可用：权重化选主 + 磁盘探活
  - 低成本：Logger节点



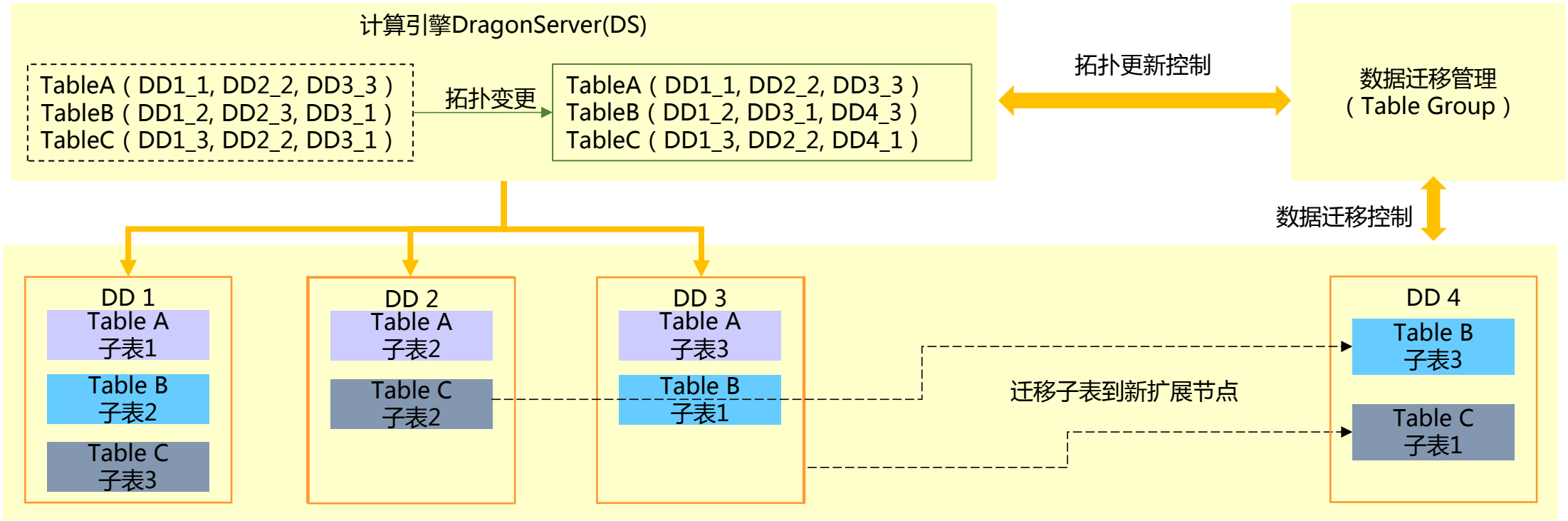
DragonBase支持灵活水平扩展

- **分库分表：**

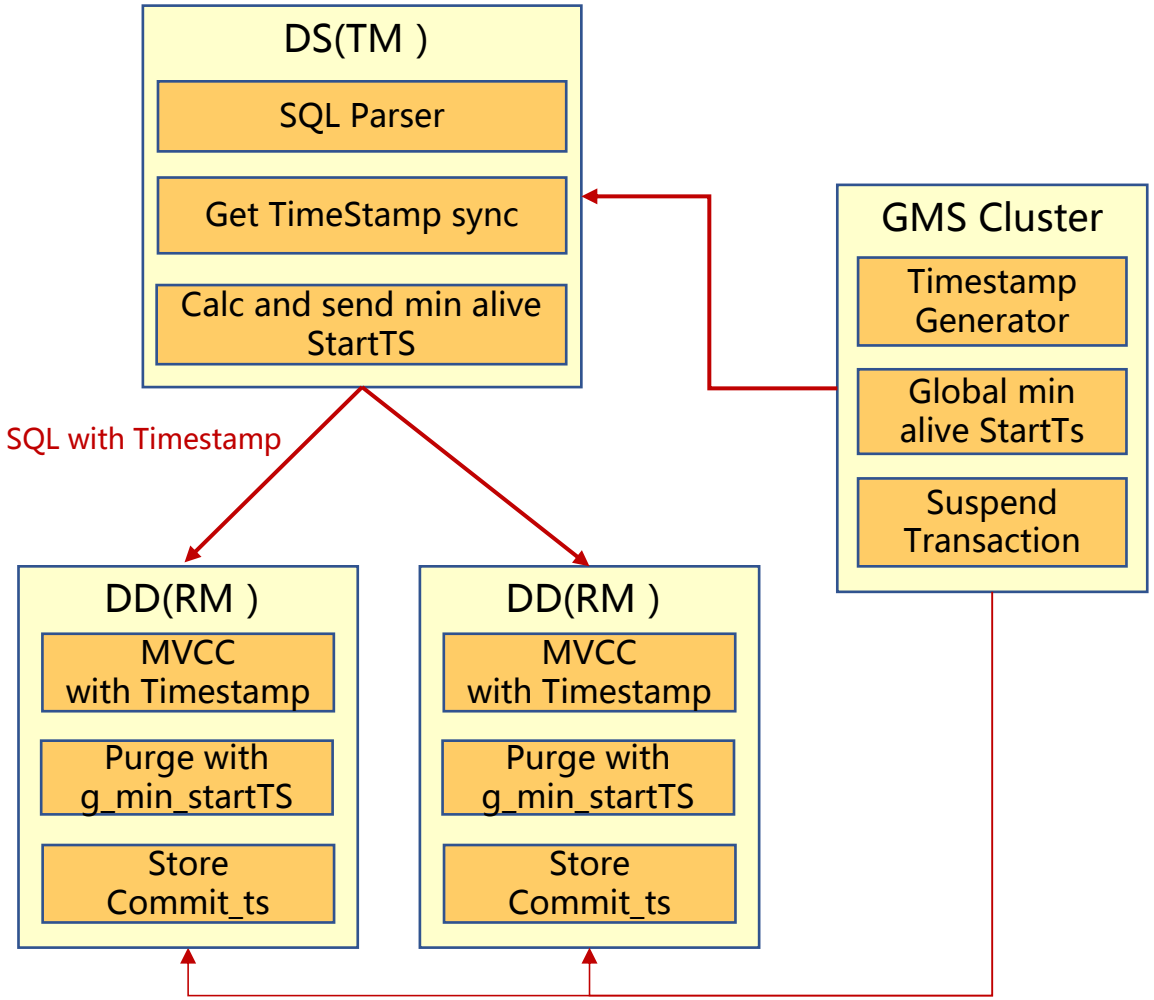
  支持Hash\range\datetime拆分方式，支持单策略和多策略

- **弹性扩展：**

  子表粒度，按需迁移，支持按容量或访问热度迁移，支持Table Group方式扩容

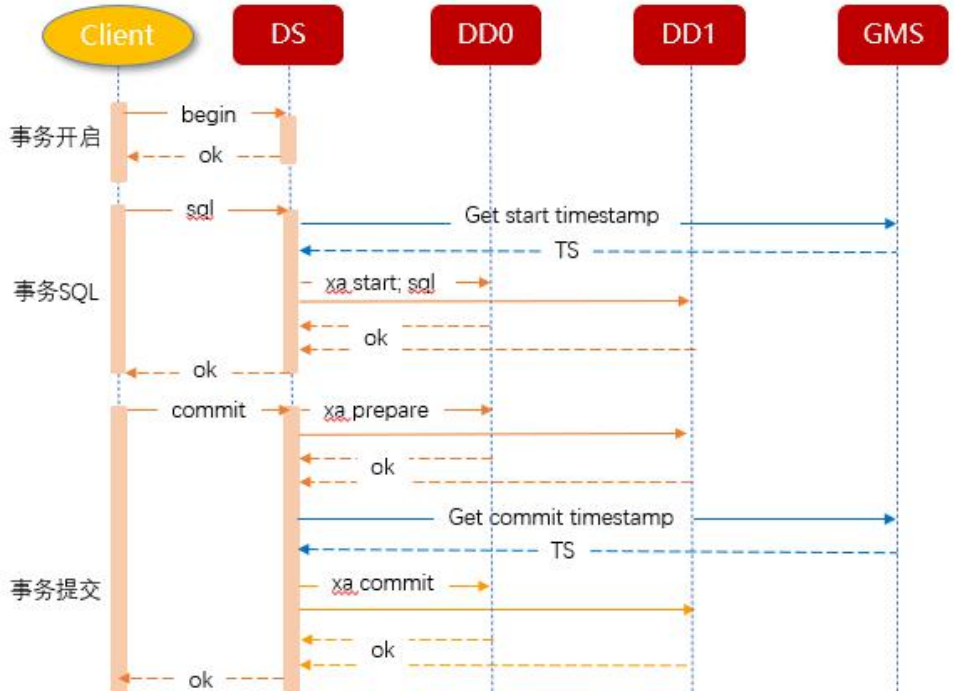


DragonBase全局事务读一致性



基于全局时钟，实现RR\RC隔离级别的全局读一致性

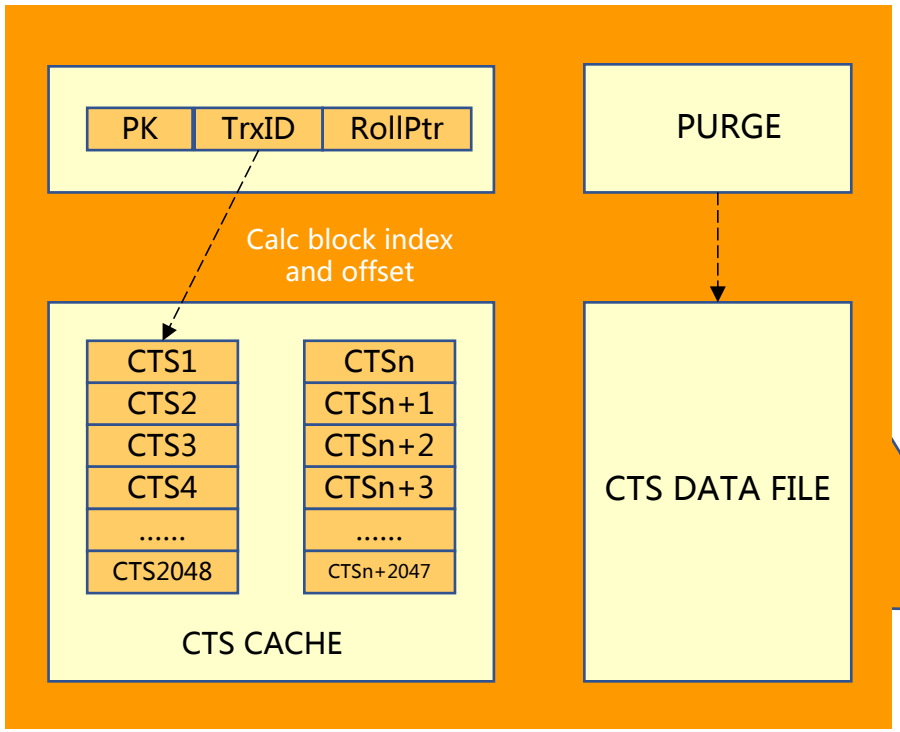
- 基于CTS（Commit TimeStamp）的MVCC机制
- 采用Batch和Pipeline的时间戳交互方式，降低延时、抖动的影响
- 单分片写入场景，进行一阶段事务提交优化，减少获取时间戳开销

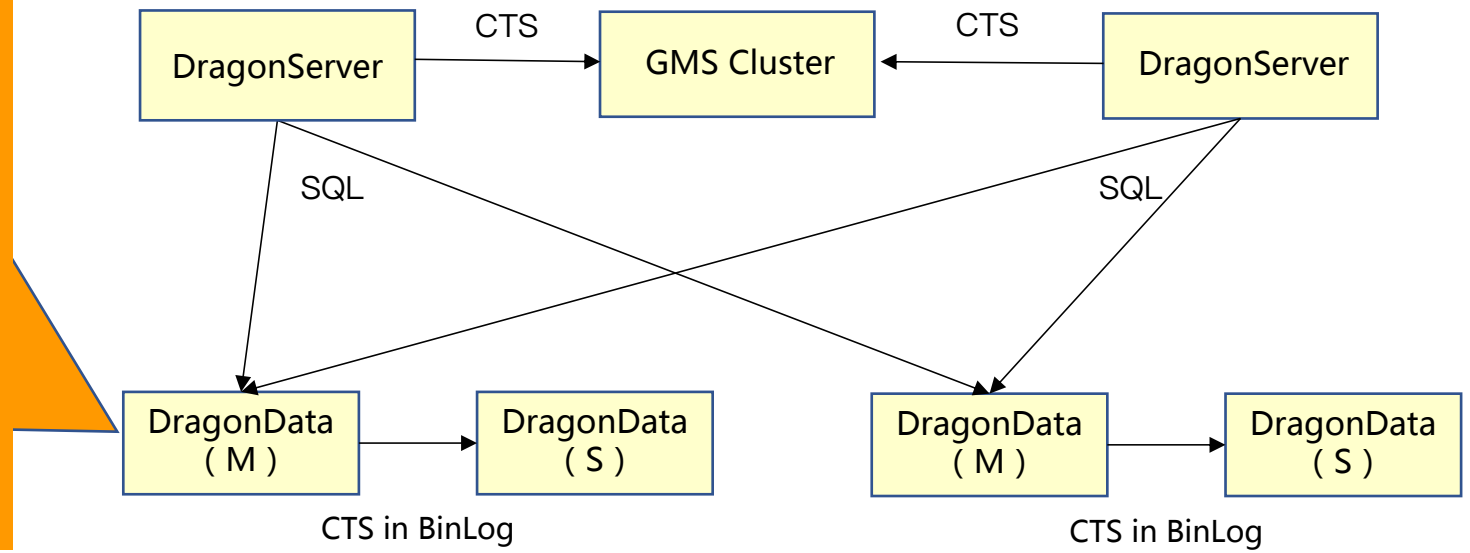


DragonBase全局事务读一致性（DMVCC设计）

基于CTS（Commit TimeStamp）的DMVCC

- 时间戳的生命周期：事务开启**申请**StartTS，事务提交**申请**CommitTS，并**持久化**CTS，并最终由CTS_worker线程**删除**
- 数据可见性判断：比对StartTS和数据行所属的CommitTS

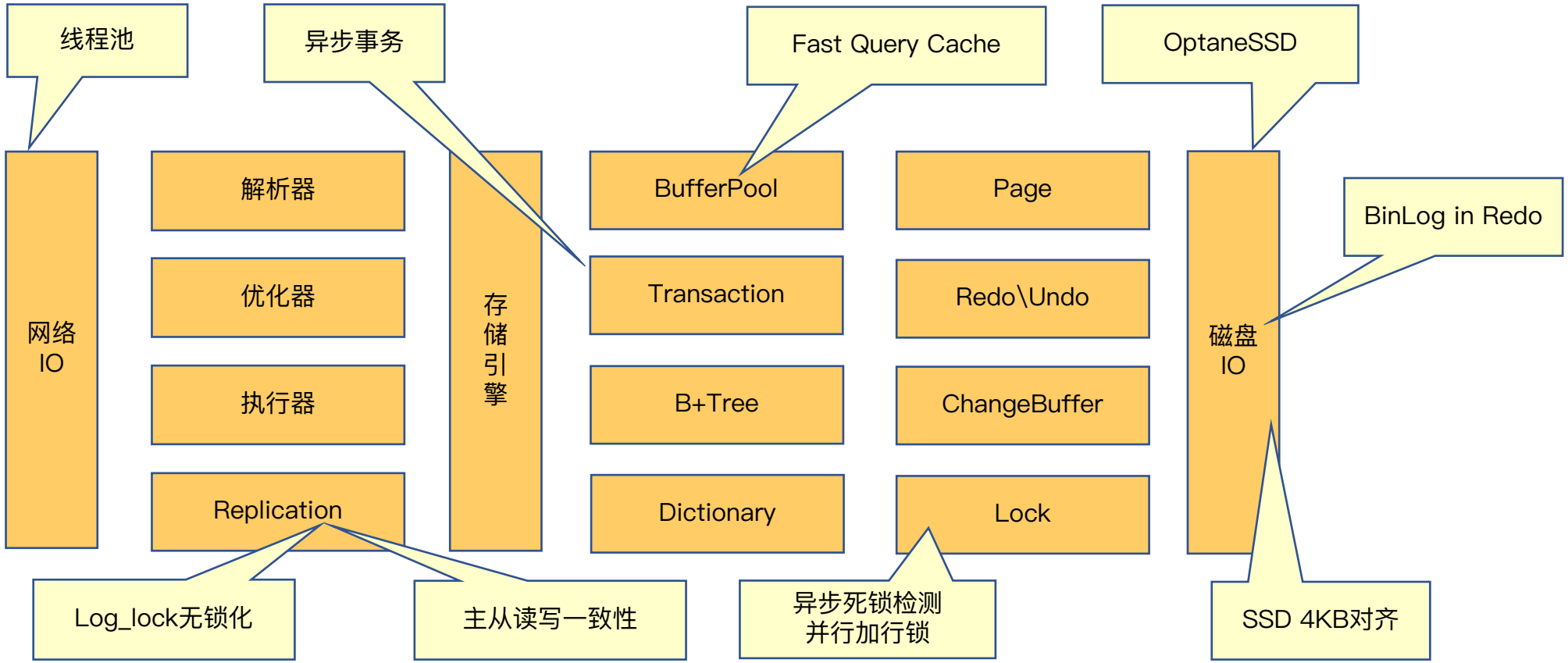




DragonBase存储层内核优化

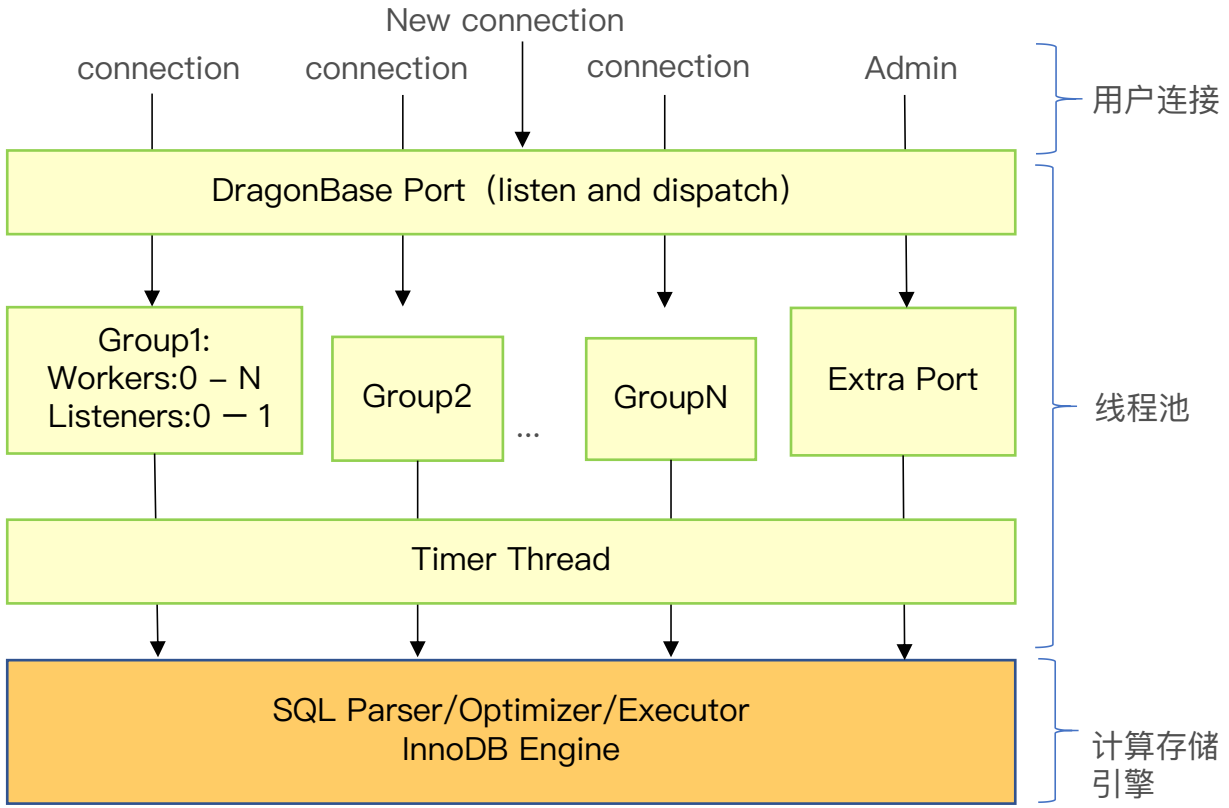
□ KingSQL性能优化：

异步 + 并行 + Batch + 少互斥（多入口\LockFree）

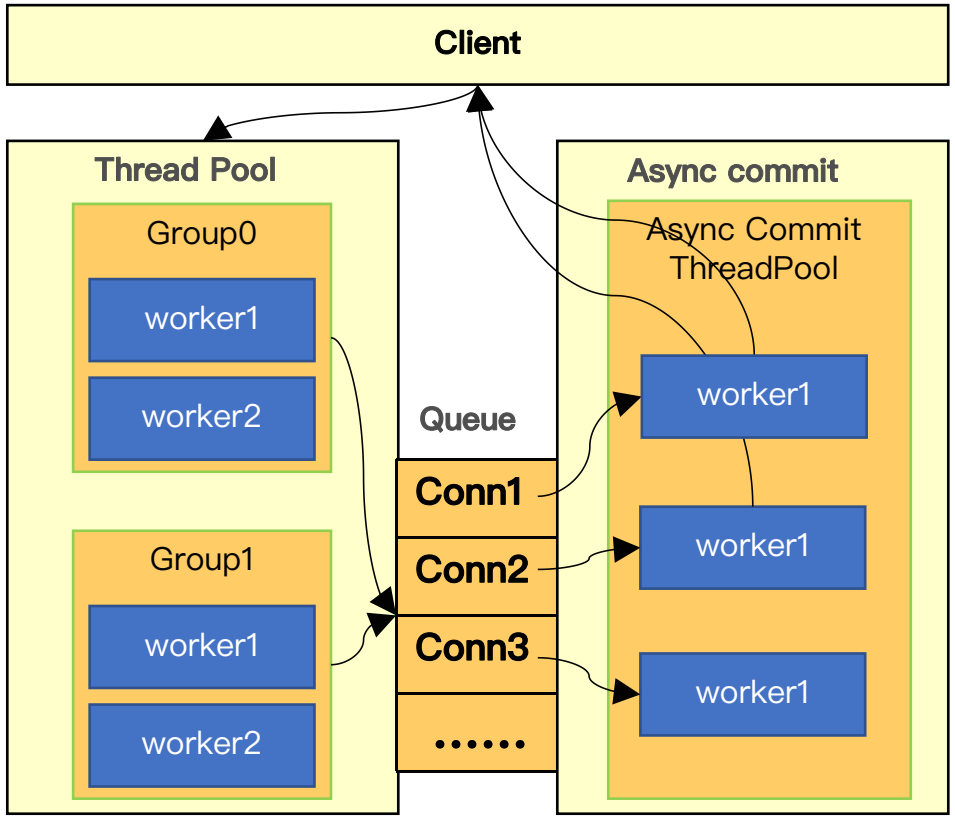


DragonBase存储层内核优化

- 高并发关键问题：线程切换开销过高
- 问题抽象：排队论（Queuing Theory）：
- 解决方案：线程池，每个CPU核处理适当的活跃线程，快启动、慢增长：

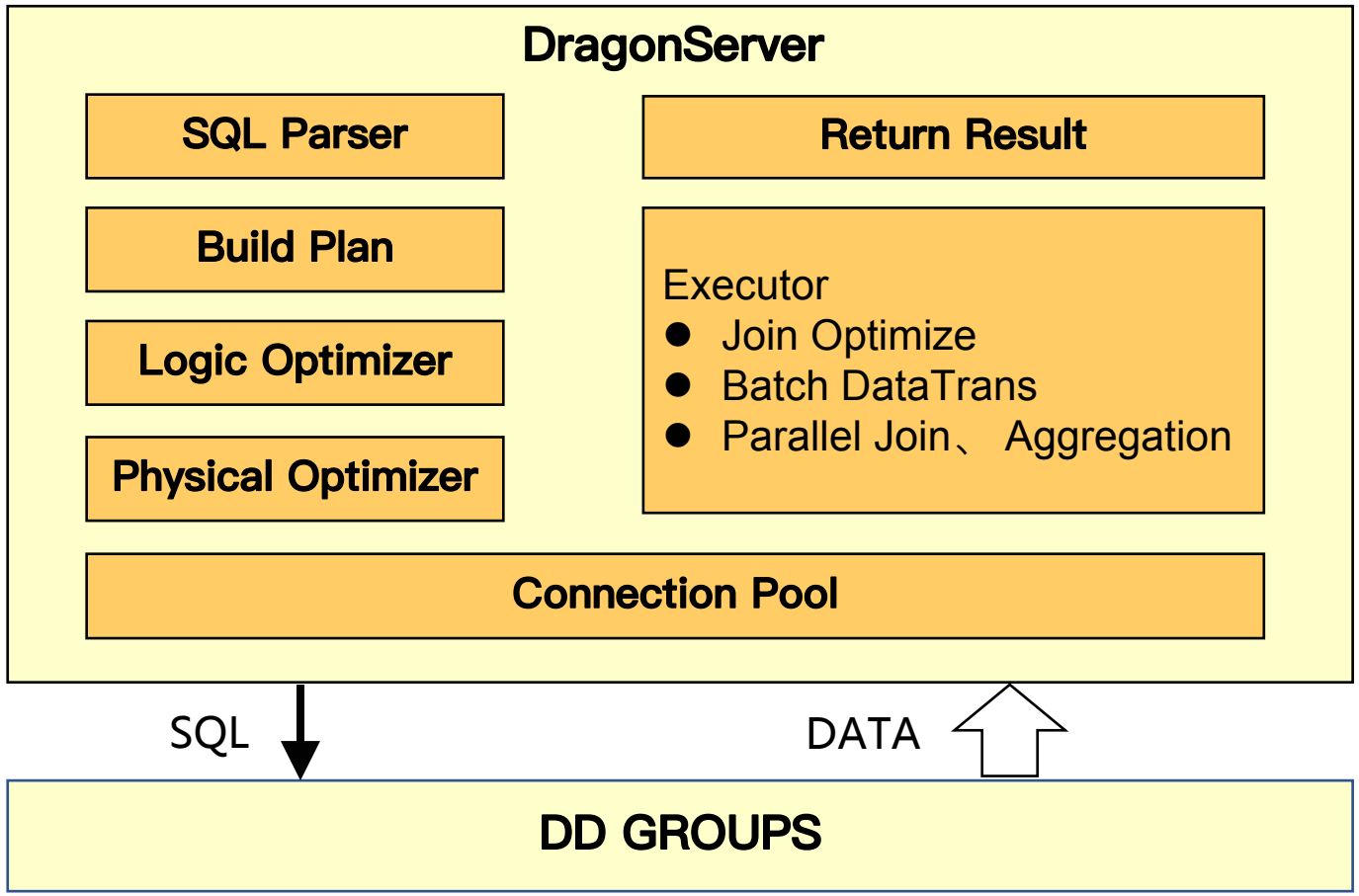


- 强同步关键问题：事务提交时间过长导致线程数大幅增加
- 问题拆解：内存计算和网络交互耦合性较高
- 解决方案：异步事务，解耦SQL处理和事务提交



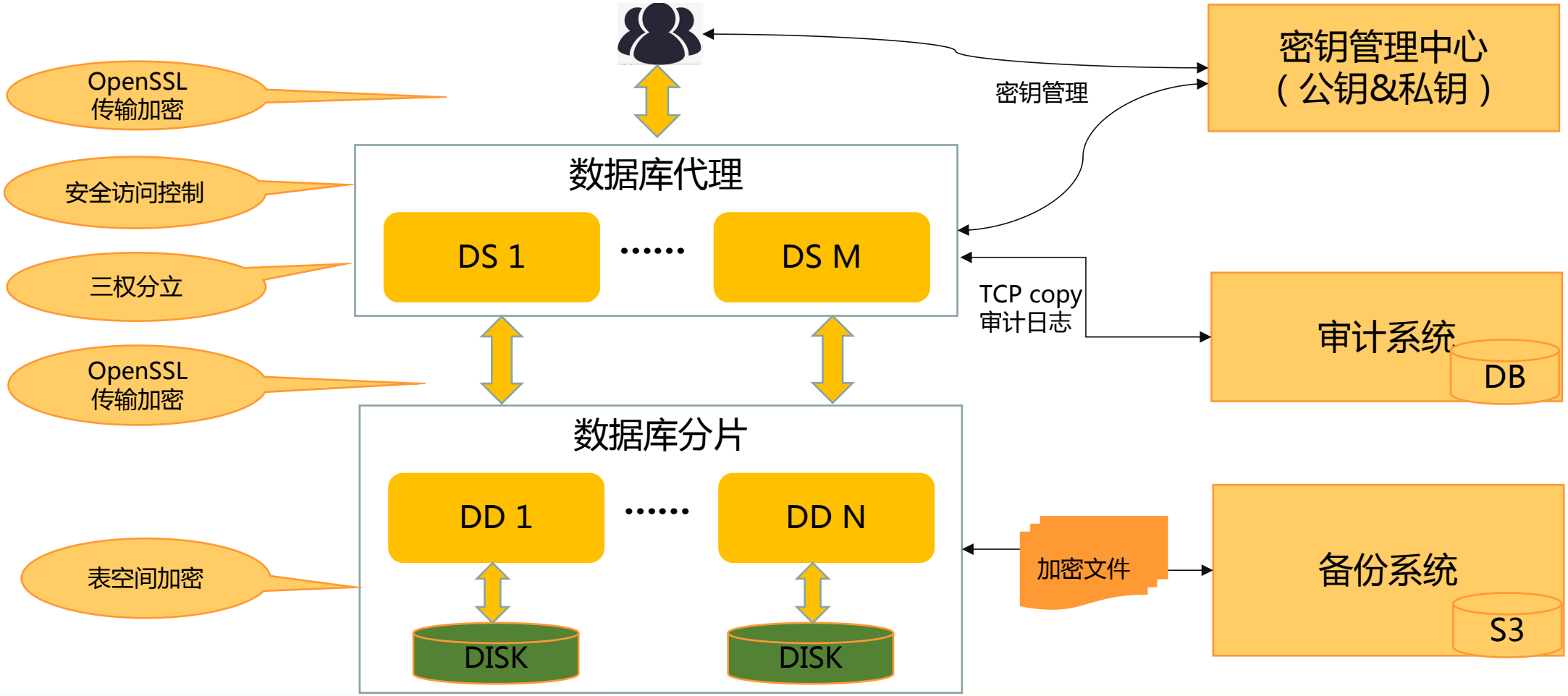
DragonBase计算层查询优化

- 普通点查询：高效连接复用的连接池（比较会话属性 + multi_query + async reset）
- 复杂查询：RBO\CBO\执行器优化。减少传输、计算开销，优化并行计算



DragonBase安全机制

三权分立 + 访问控制 + 安全审计 + 加密技术



DragonBase运维管控平台

- 资源管理
- 集群部署
- 参数管理
- 备份恢复管理
- 权限管理

- 监控指标管理
- 监控展示
- 告警策略管理
- 告警内容管理

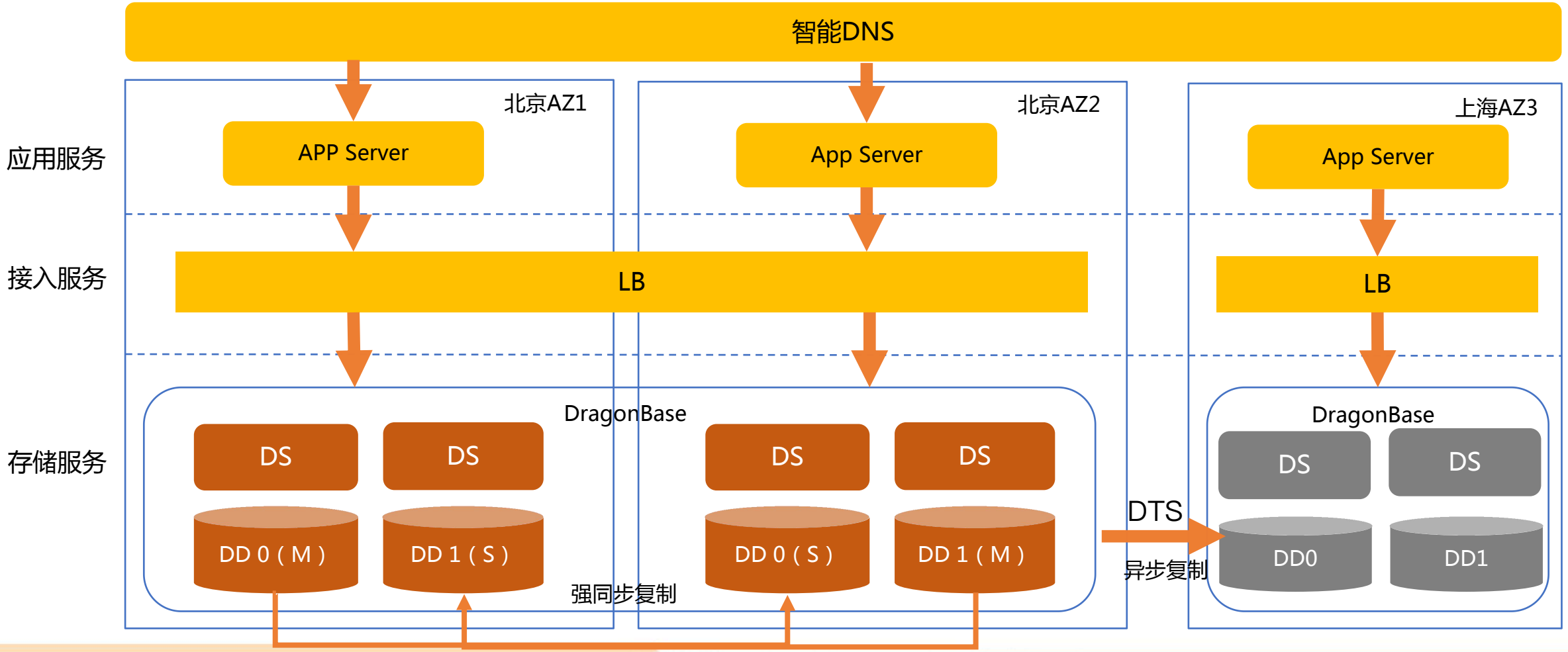
- 日志分析
- 慢日志检索
- 审计日志检索
- 故障巡检
- 故障诊断

- 服务日志
- 审计日志
- 慢日志
- 运维日志
- 全链路日志追踪
- 日志可视化/下载

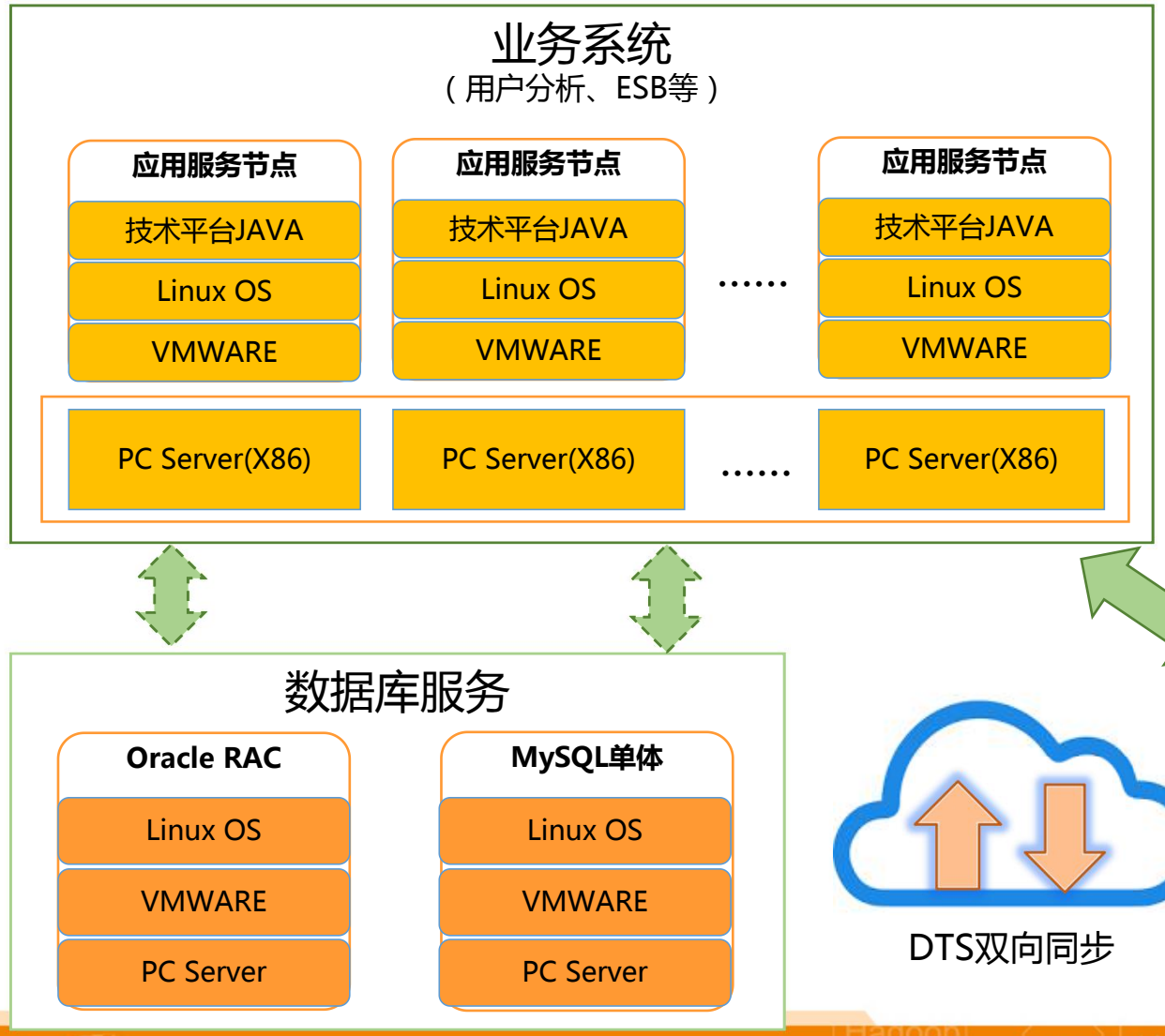
4 金山云DragonBase应用实践

DragonBase实践：某证券公司（两地三中心）

- 部署原则：**同城双活、异地灾备、单元化**（每个AZ都包含完整的应用、服务和数据）
- 同城使用强同步复制，保证RPO为0；跨城使用DTS异步复制，RPO毫秒级，RTO时间DNS路由生效时间，通常是秒级



DragonBase实践：某银行业务系统替换Oracle

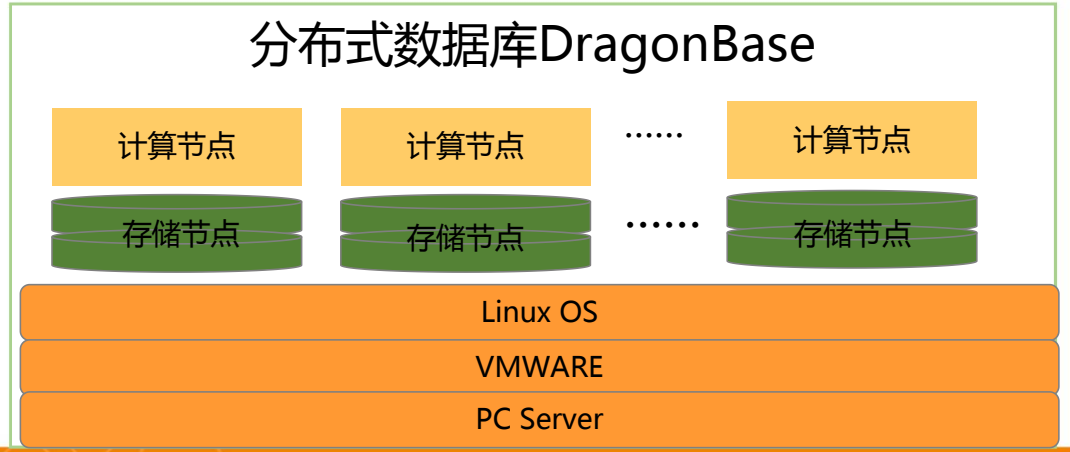


【产品优势】

- 分布式数据库替代商业数据库和传统数据库： 金融级强一致分布式数据库，可扩展。
- 降低成本： 降低硬件和商业软件成本

【客户价值】

- 分布式开放架构，弹性扩展，满足未来发展需要
- 提高新业务上线效率



DragonBase实践：OLTP&OLAP结合

在线业务

交易　CRM　风控　报表

分布式缓存

关系型TP数据库

DS　DS　……　DS

DD　DD　……　DD

数据同步软件

分析类数据库

APACHE HBASE　elastic　Greenplum　hadoop HDFS


THANKS
时序
开源
自研
分布式
文档
Spark
HANA
SQL
OLAP
Oracle
Hadoop
MPP
DB2