DTCC 13 2010 2022

第十三届中国数据库技术大会

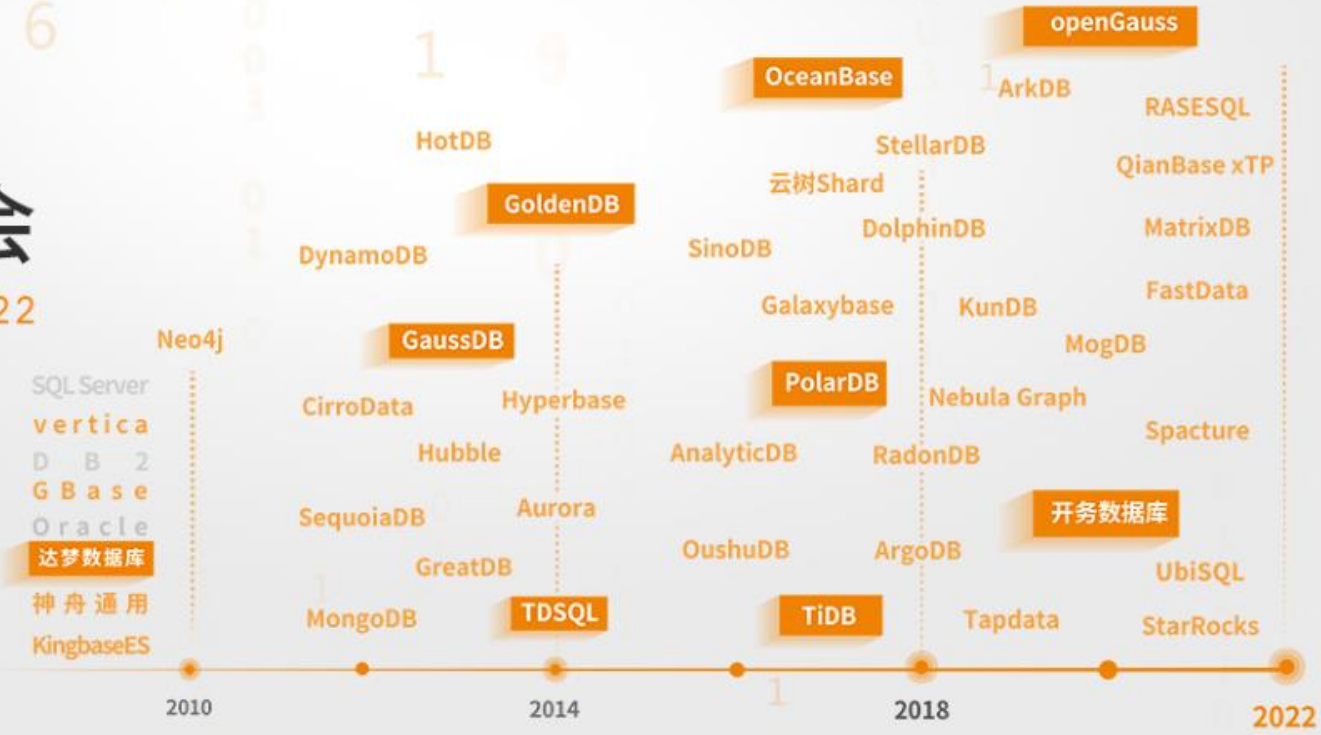
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



京东

京东云

DTCC 2022
第十三届中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

云海统一存储平台架构与实践

京东科技 – 京东云事业群
存储架构师 郑静

云海架构与实践

云海统一存储平台架构介绍 | 存算分离实践

云海统一存储平台是什么

阿里云盘古分布式存储

云海统一存储平台

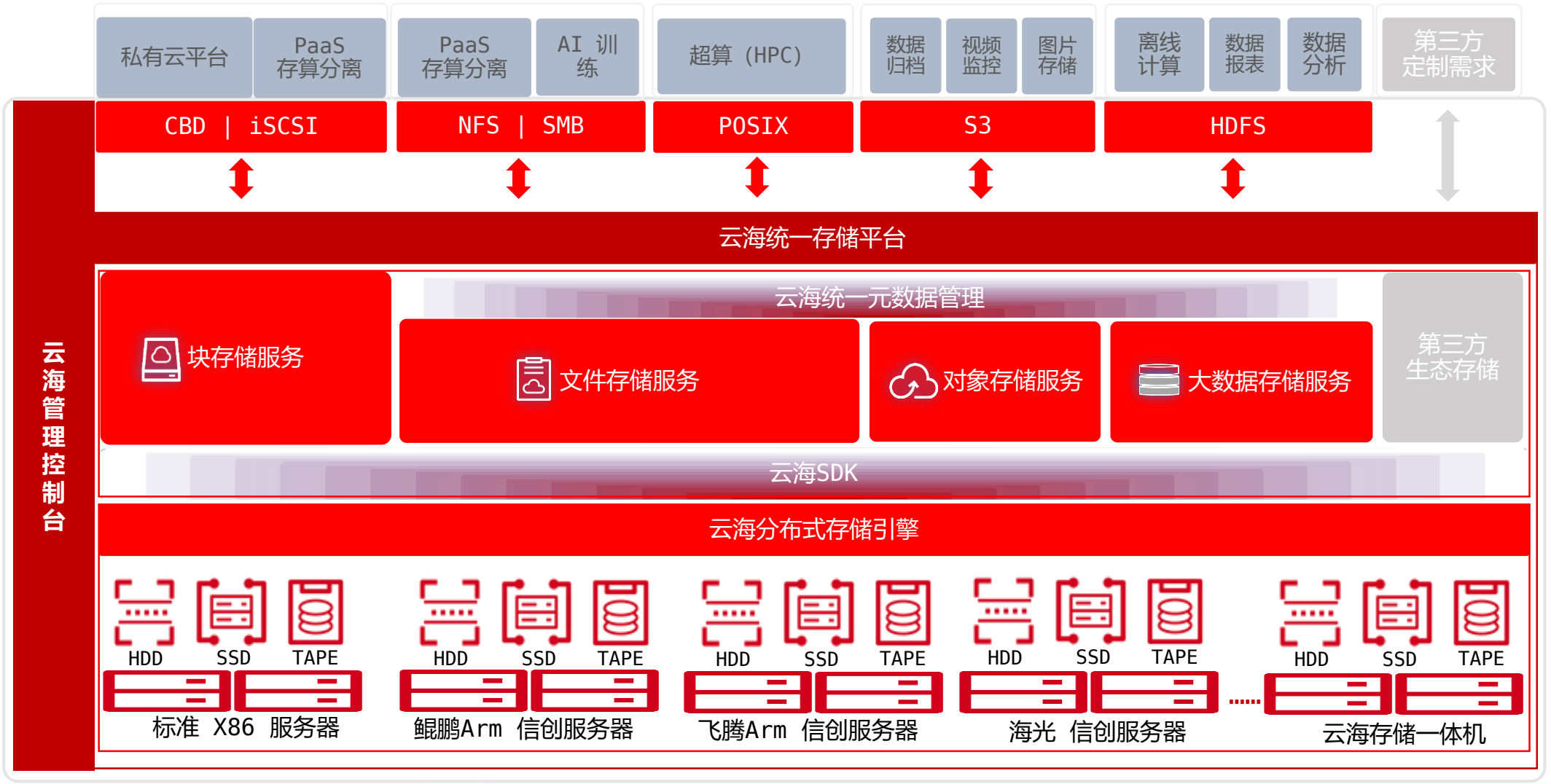
京东全自研分布式存储系统，高性能，高可靠，低成本，深度软硬件一体加速，协议互通

云海统一存储平台架构

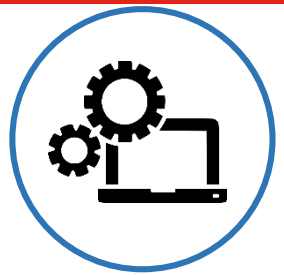
全场景

更敏捷

更高效

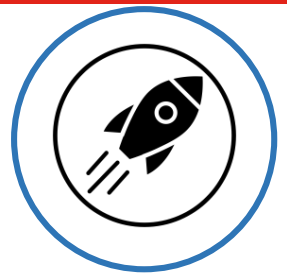


云海统一存储平台 领先性



全自研存储引擎

基于京东10+年在存储产品上的积累，针对新硬件的不断发展和传统存储架构不足，全新自研新一代存储引擎，充分满足存算分离、AI训练等新兴场景趋势的需求



高性能

基于全自研架构及京刚芯片加速，云海在公有云生产环境上做到单IO延迟 *100μs*；**云硬盘单实例百万IOPS纯写场景下延迟 *<200μs***；性能接近本地NVMe磁盘

低成本

低冗余支持从1.5-1.14 副本的存储方式

对QLC-SSD 友好的架构设计，使QLC-SSD能大规模应用于高性能场景的生产环境；官方认证为国内最大规模使用QLC的云厂商

强稳定

京东在经典的RS算法的基础上，推出RSDP数据修复模型，可以将修复网络开销降到最低，低冗余存储条件下的故障数据修复速度提升5~10倍

多机房高可用 / 精细化流控 / 副本故障下高可用性

国产化替代	支持全信创硬件	数据强一致	操作日志审计
安全可靠的存储	安全（国密加密）	多机房高可用	多地数据灾备

云海统一存储平台 高性能

软硬件一体技术

自研RDMA 网络栈

只用verbs接口

自流控

大小包拆分

读写分离

无Leader一致性协议

全链路异步模型

多流异步同时服务

全程无锁IO路径

不跨CPU，支持Polling

日志结构文件系统

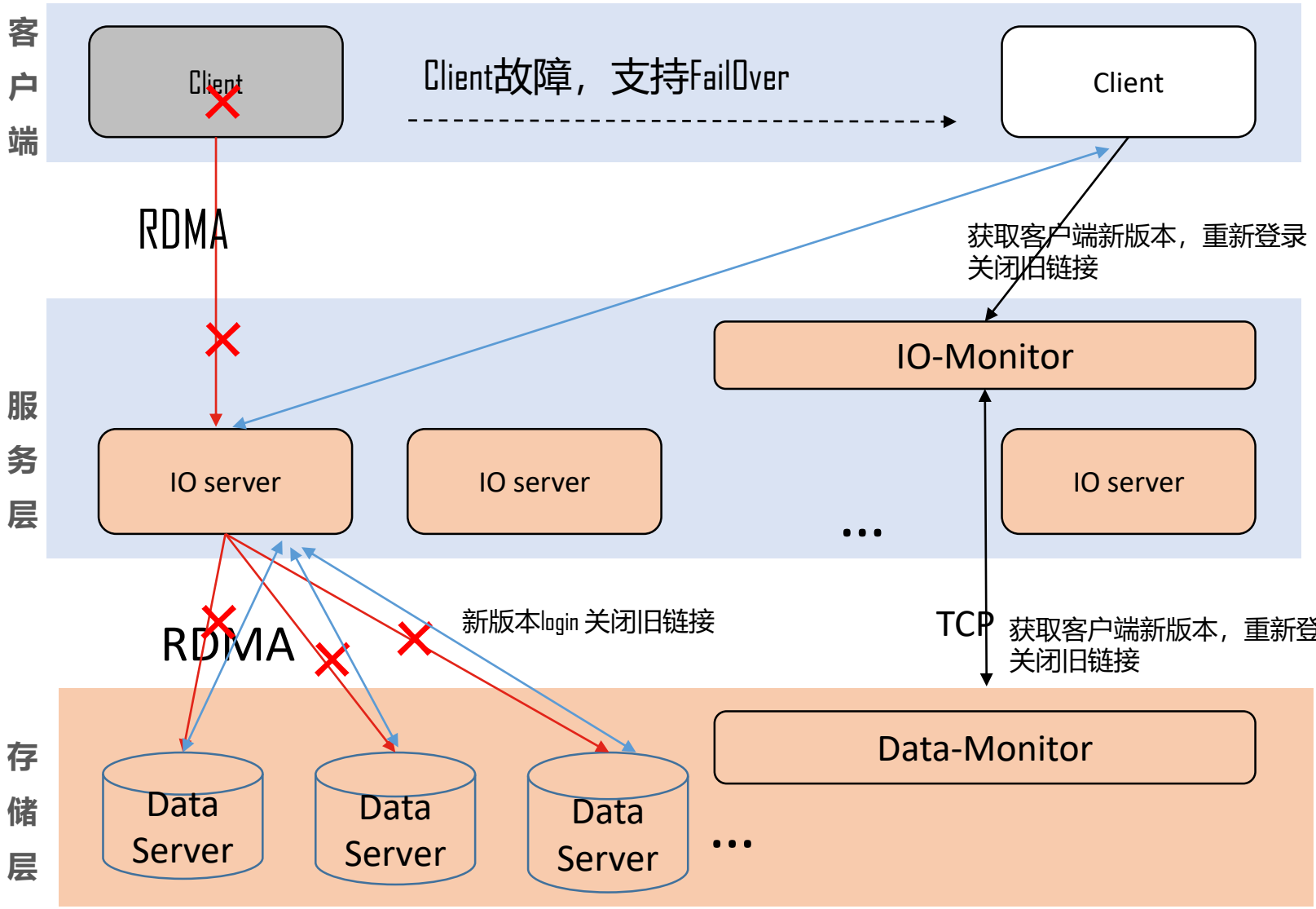
日志减少写放大，增加

预读，减少延迟性

```
Jobs: 8 (f=8): [w(8)][63.4%][r=0KiB/s,w=3906MiB/s][r=0,w=1000k IOPS][eta 00m:26s]
Jobs: 8 (f=8): [w(8)][78.9%][r=0KiB/s,w=3906MiB/s][r=0,w=1000k IOPS][eta 00m:15s]
Jobs: 8 (f=8): [w(8)][94.4%][r=0KiB/s,w=3914MiB/s][r=0,w=1002k IOPS][eta 00m:04s]
Jobs: 8 (f=8): [w(8)][100.0%][r=0KiB/s,w=3908MiB/s][r=0,w=1001k IOPS][eta 00m:00s]
test: (groupid=0, jobs=8): err= 0: pid=70617: Sat Jun 18 20:16:17 2022
  write: IOPS=999k, BW=3903MiB/s (4092MB/s)(229GiB/60001msec)
    slat (usec): min=2, max=116, avg= 5.67, stdev= 1.23
    clat (usec): min=59, max=17545, avg=186.05, stdev=123.67
     lat (usec): min=64, max=17551, avg=191.78, stdev=123.67
    clat percentiles (usec):
     |  1.00th=[   89],  5.00th=[  103], 10.00th=[  114], 20.00th=[  129],
     | 30.00th=[  141], 40.00th=[  153], 50.00th=[  167], 60.00th=[  184],
     | 70.00th=[  202], 80.00th=[  229], 90.00th=[  273], 95.00th=[  314],
```

云海存储，带负载压力实测结果，三节点单实例云硬盘纯4K写 IOPS 达到 *100*万时，平均延迟仍在 *200μs* 内

云海统一存储平台- 高性能 存储网络栈



链路聚合

四栈合一（用户态TCP/**内核态**TCP/RDMA/IB**）**

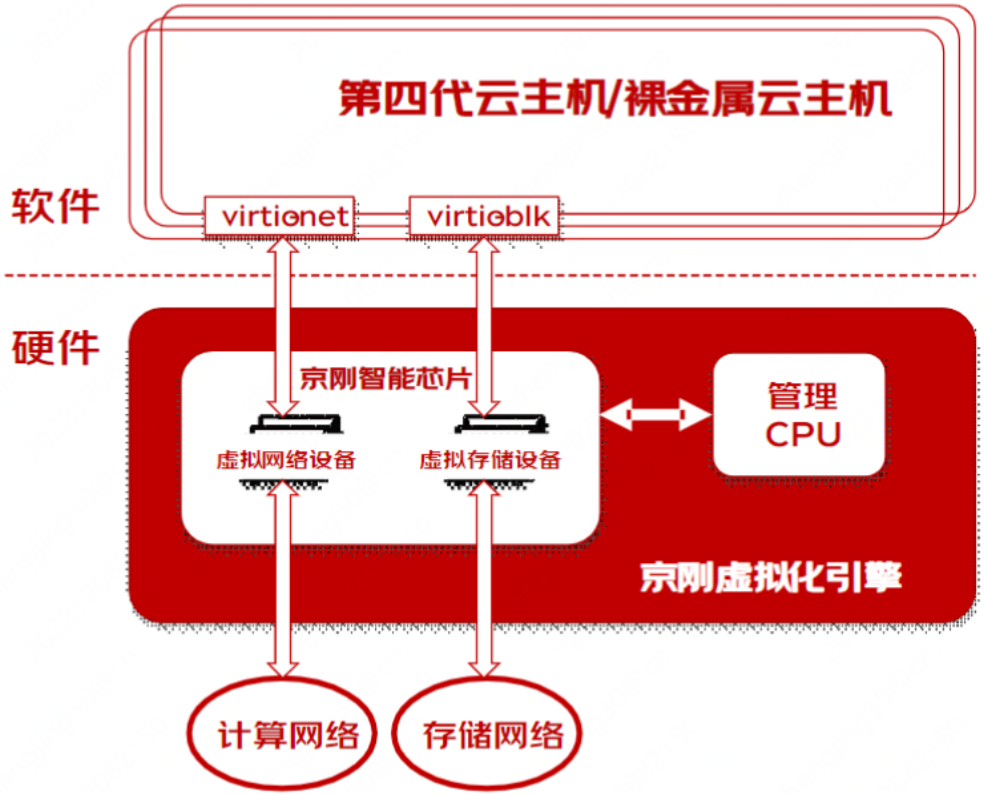
组件基于名字空间寻址

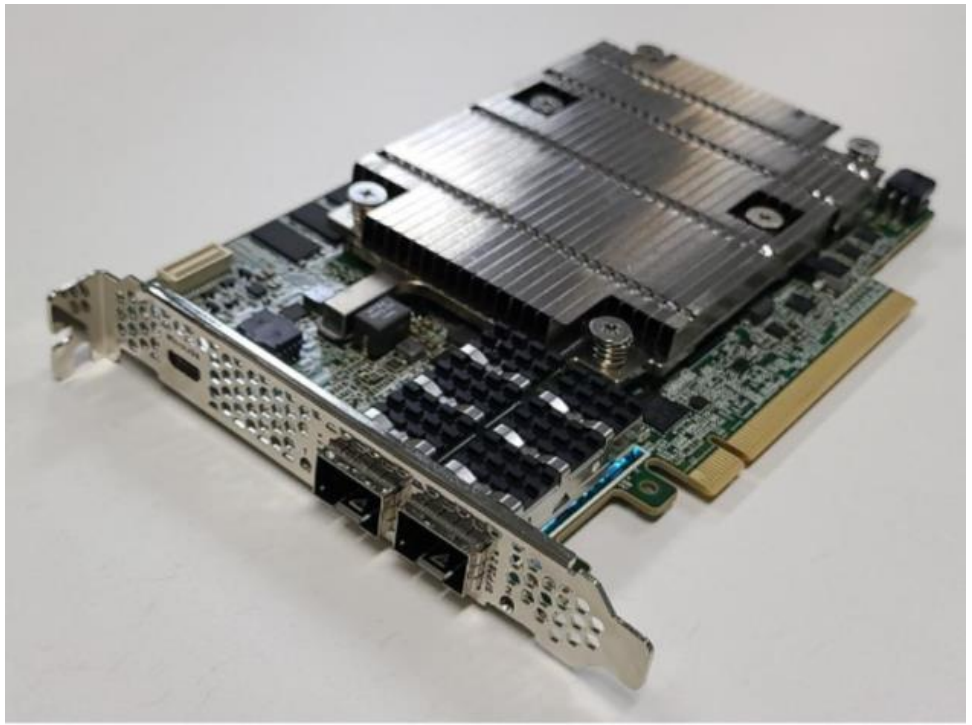
网络层实现Failover

支持IO fence

IO**全路径支持**RDMA，**零拷贝**

云海统一存储平台 – 高性能 软硬件一体技术





VM **侧硬件化**vrtio

存储网络硬件卸载RDMA

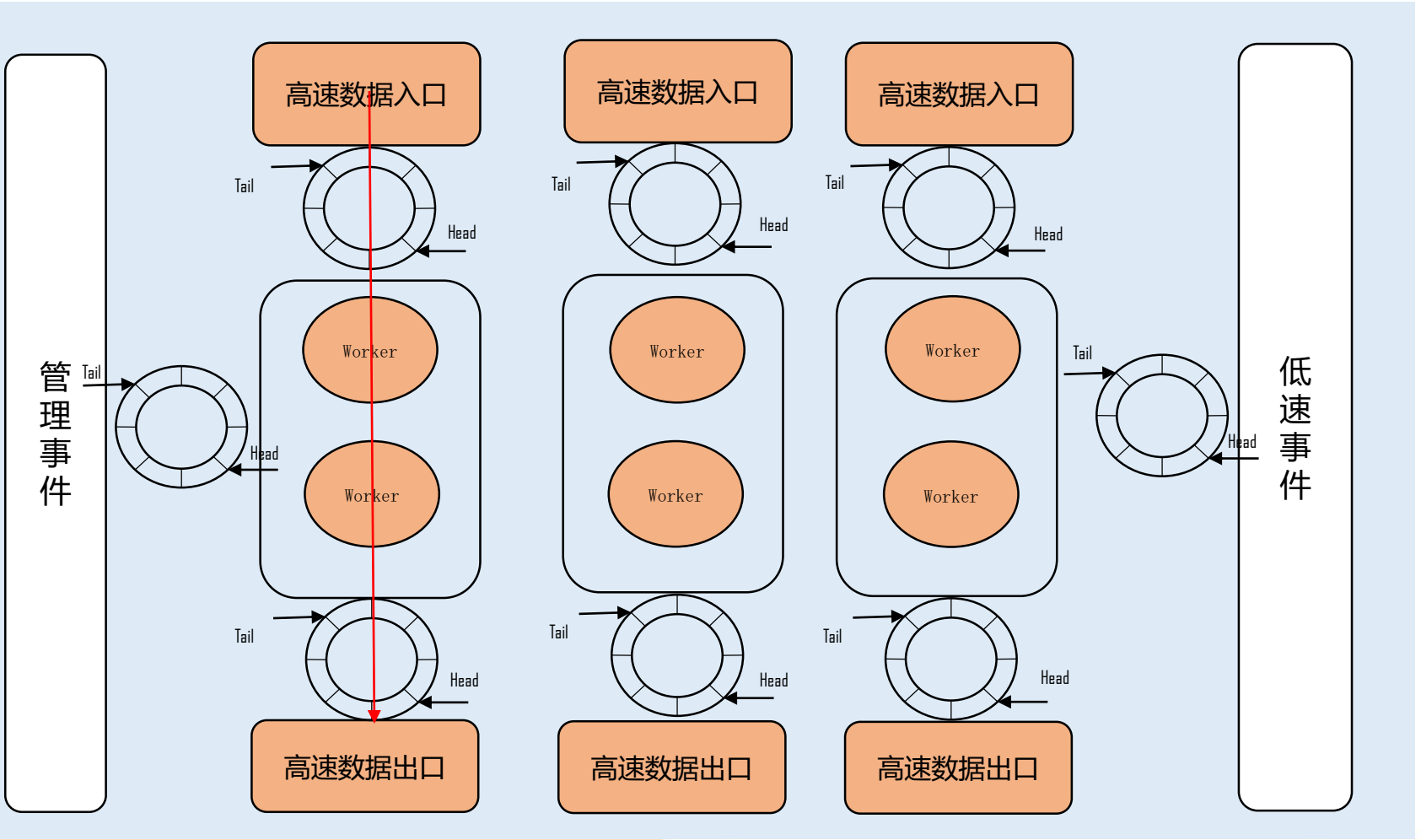
虚拟网络全卸载

使用京刚产品，存储全链路

延迟低于100us

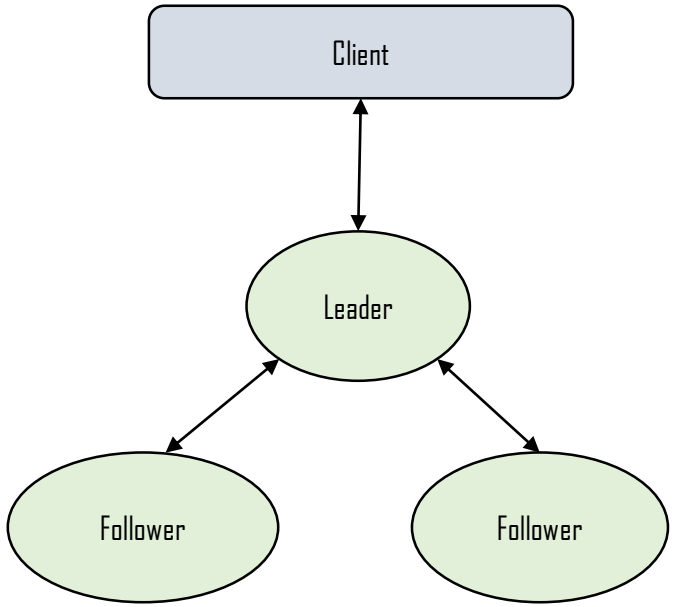
数据路径全程零拷贝

云海统一存储平台 – 高性能 异步模型

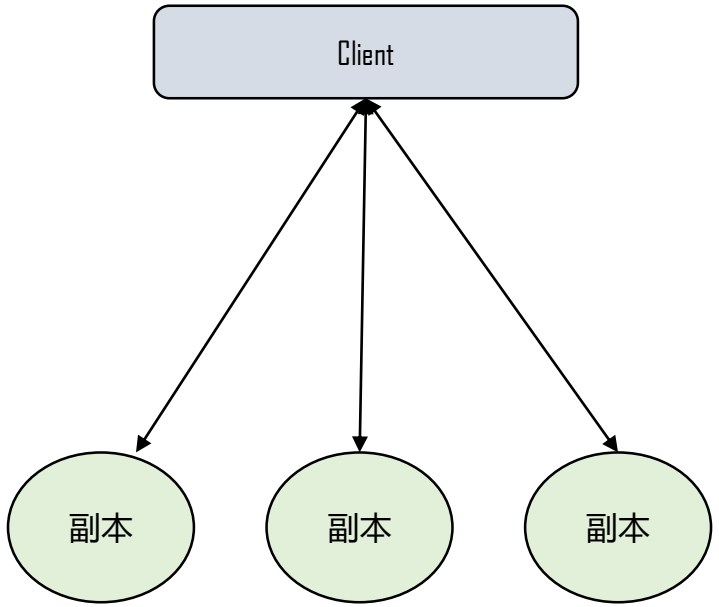


- 高速低速事件隔离
- 无锁队列交互
- IO不跨核
- 全无锁实现
- 全异步编程
- 全内存池
- 全用户态

云海统一存储平台 – 分布式一致性协议



采用客户端控制
副本一致性协议



Raft**选举期间无法提供服务**

Raft**串行同步**journal**性能受限**

Raft In-Place Update**架构，故障容忍度低**

无Leader**模式，减少故障**IO**卡顿**

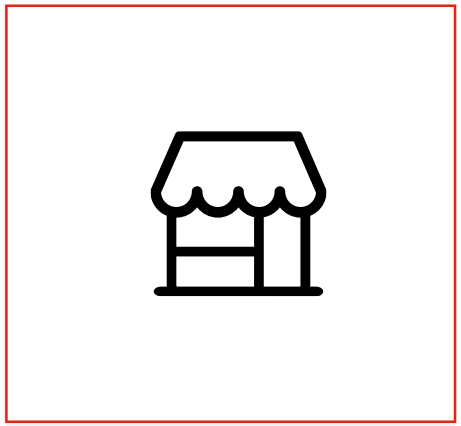
客户端主动探测异常触发IO**切换，保障**SLA

支持更高密度机型

云海统一存储平台 – 低成本

新硬件能力

自研引擎对QLC-SSD的特性做针对性设计，官方认证为国内最大规模使用QLC的云厂商



低冗余副本

冗余度最低到1.14副本; 并且数据修复速度极高
自动数据分层存储

软件栈优化

通过软件栈优化，降低软件消耗，并提升磁盘的利用率

云海统一存储平台 – 强稳定

故障快速恢复

无Leader副本模式，
毫秒级副本切换
快速 EC 数据恢复

异地容灾

支持三副本位于不同故障域
多副本可读

全链路数据管理

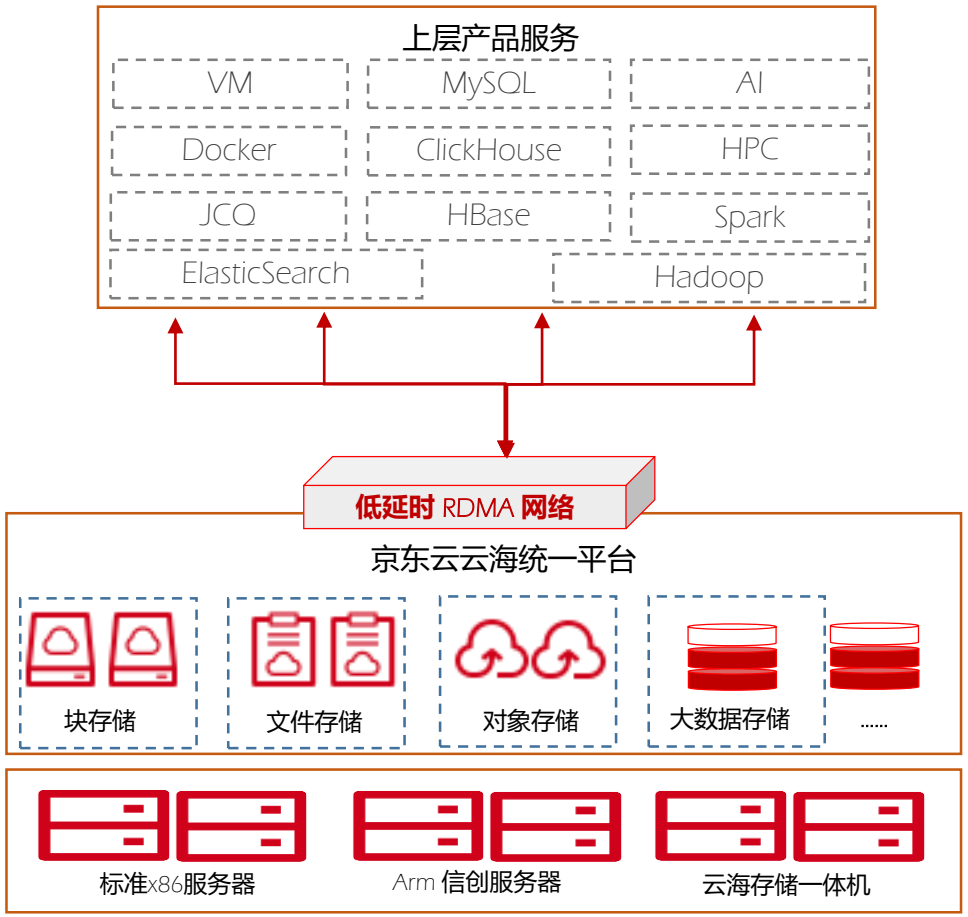
全链路数据端到端校验
从硬件、软件、服务层
实时监控预警
智能流控

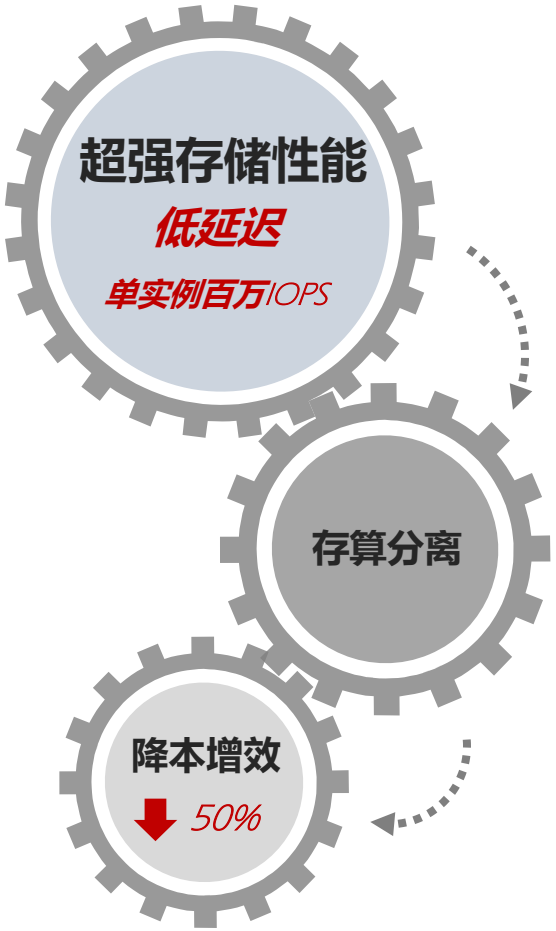
云海架构与实践

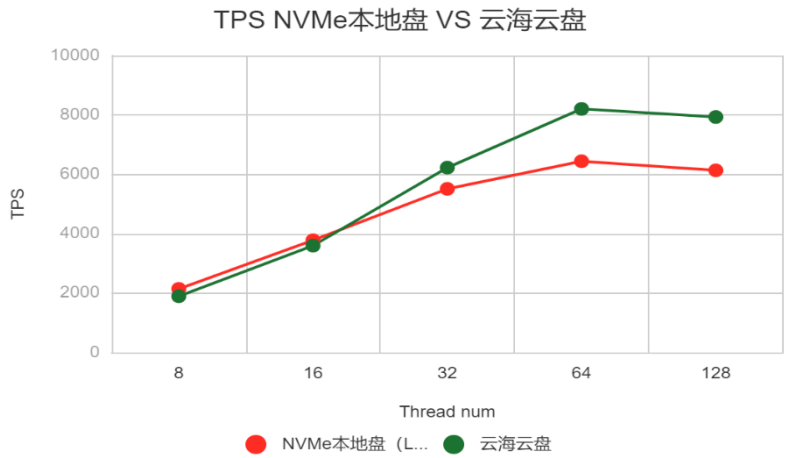
云海统一存储平台架构介绍 | 存算分离实践

云海存算分离实践：释放最大降本价值

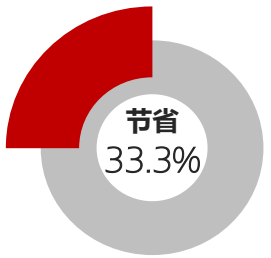
多场景支持，极致存储分离，降本增效



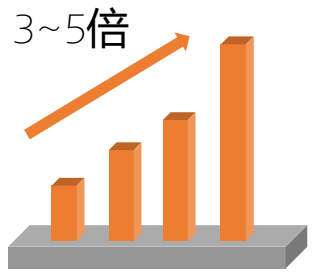




资源节省



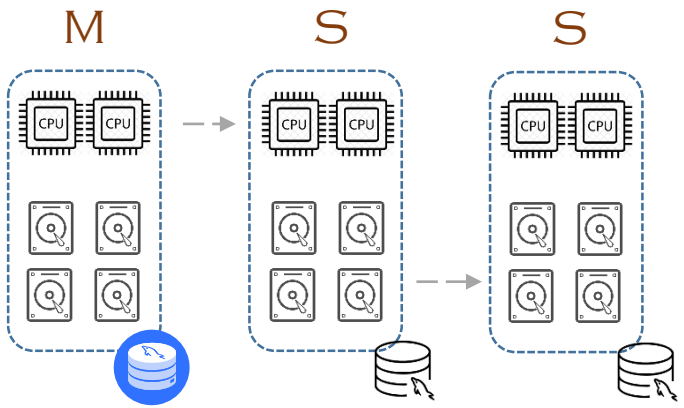
计算资源



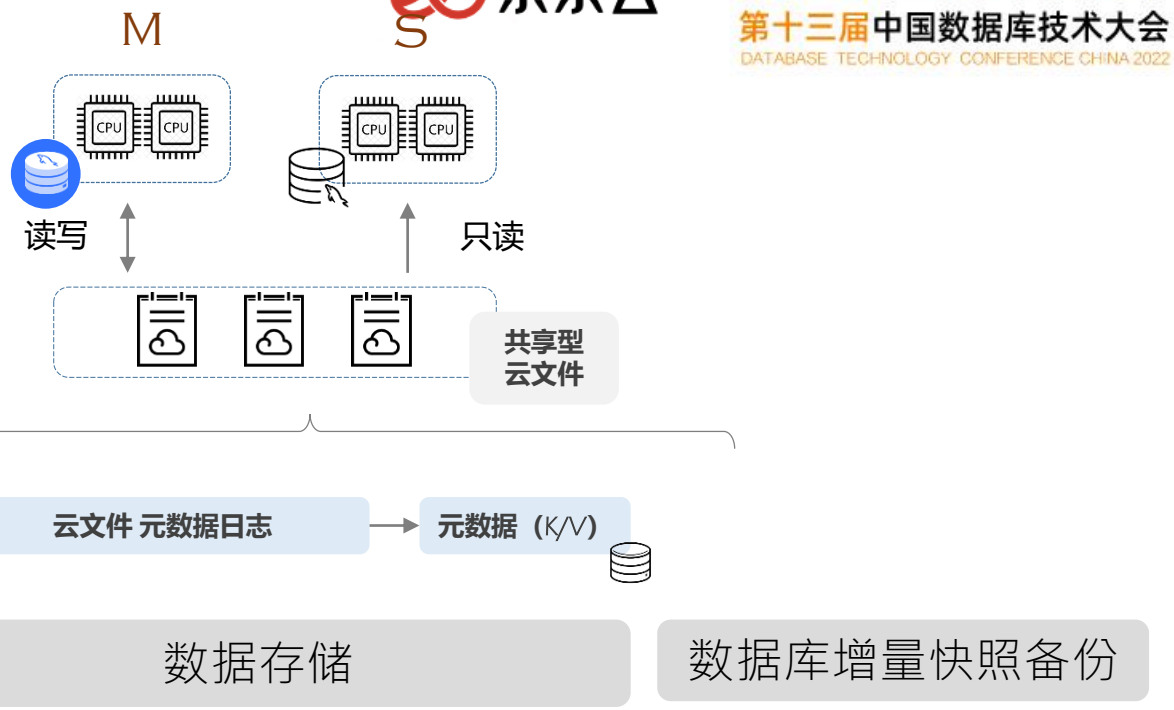
存储资源利用率

云海存算分离实践：共享存储 MySQL

存算分离架构



云海存储引擎



现有架构问题

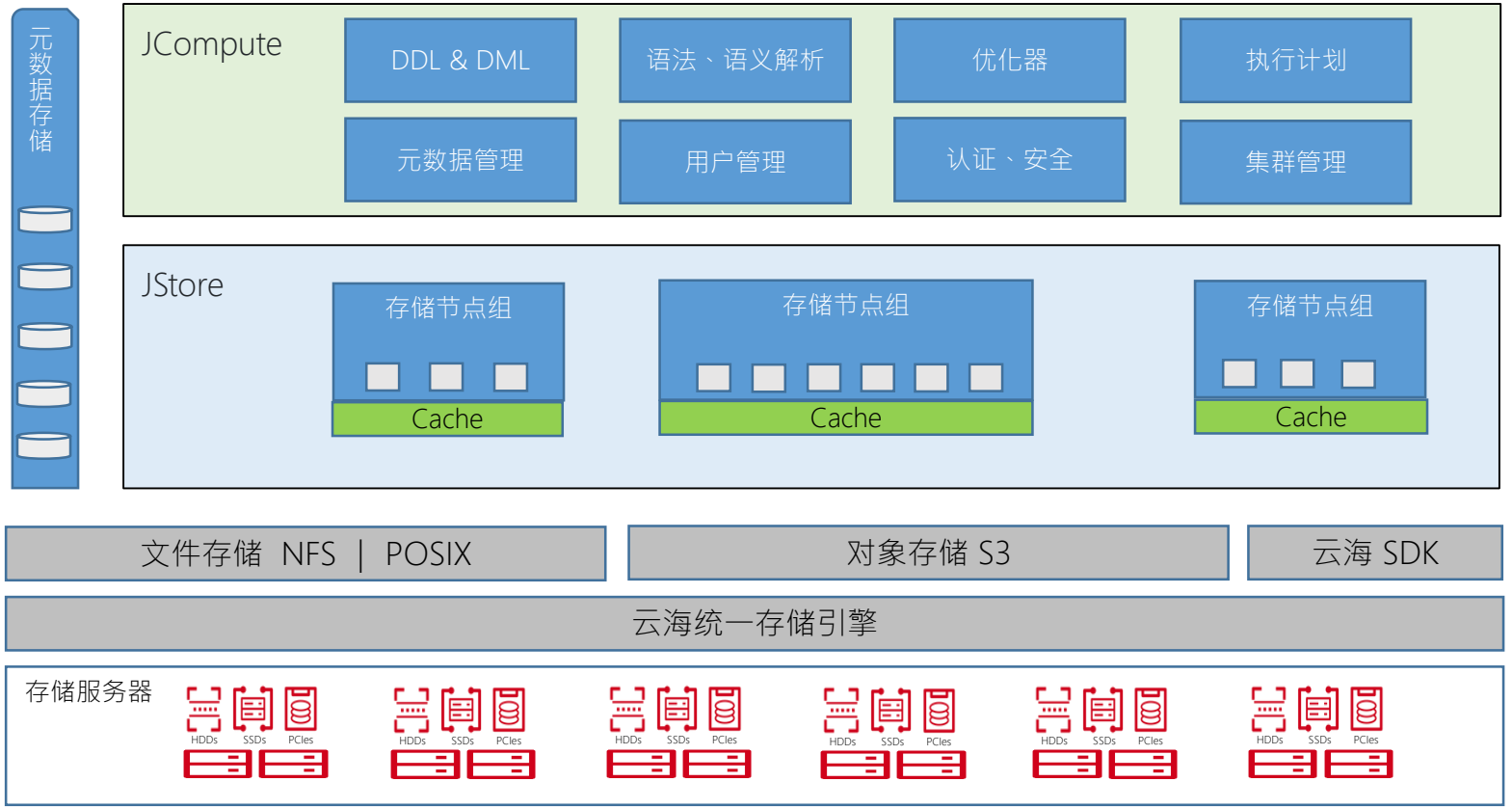
- **资源浪费** 不同时期对存储和计算的需求不同，绑定两者需要同时对两种资源进行扩展或收缩，会造成资源的明显浪费
- **使用效率低** 云计算技术上，数据库、中间件的 SERVERLESS 开始出现并在逐步主流化。背后的关键基础即是计算与存储资源解耦
- **应对突发难度大** 突发用户访问，超高并发下单等需求对计算资源的需求会爆炸增长，存算耦合扩容会使得扩容成本巨大

存算分离收益

- 数据库由 主-从-从 模式演进到 主-从 模式，计算资源需求降低 1/3
- 存储资源成本，降低 3~5 倍
- 计算、存储弹性秒级扩容；数据库快照；故障快速恢复

云海存算分离实践：京东云原生数仓架构

存算分离架构




openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云时Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
神舟通用
KingbaseES
TDSQL
TiDB
Tapdata
UbiSQL
StarRocks
2010
2014
2018

THANKS