移动云百万IOPS块存储系统架构分享

移动云块存储架构演进和技术展望

邓瑾
中国移动云能力中心高级系统架构师

目录

一、移动云及其存储现状

二、极速型云盘自研转型

三、极速型云盘架构详解

四、未来展望

移动云发展历程及现状

人员规模

突破4000人

算力资源多元布局

全国建成40+个直管资源池，100%省份覆盖

产品体系丰富

230+款云产品

合作引入SaaS产品1300+

2007 2010 2014 2016 2020 2023

大云计划启动
展开Hadoop研究

发布大云1.0
云计算&大数据商用

云能中心成立
拥抱Openstack

大云规模化商用
Openstack super user

产业赋能
承载海量行业应用

跃居第一阵营
全网收入突破500亿
公有云、政务云份额TOP3
边缘云份额TOP1

移动云存储团队现状

产品
云硬盘、云硬盘备份
对象存储、归档存储、蓝光存储、云存储网关
文件存储、并行文件存储

社区
向Ceph/Gluster社区提交补丁300+个，Ceph社区排名国内第3
国内首位Gluster社区Peer

标准化
参与制定全国信委云计算标准《信息技术 云数据存储和管理》等
软件著作权7件、申请专利70+

2015
落地首个资源池
Sheepdog
ceph

2018
产品体系形成
容量型、性能优化型云盘商用

2020
自研转型
下一代超高性能云盘产品（ESSD）研发
销量规模累计超过300PB
2019年国内分布式块存储收入进入前列（第4）

2021
ESSD商用
极速型L2云盘：100K IOPS/1GBps
极速型L3云盘：1M IOPS/4GBps

一、移动云及其存储现状

二、极速型云盘自研转型

三、极速型云盘架构详解

四、未来展望

极速型云盘自主研发的必要性

Ceph局限性

面向HDD时代、IO链路性能差
社区历史包袱重、新硬件引入周期长
性能需求对系统架构产生了本质的变化

公有云架构趋势

趋向于中心化设计
软硬件融合对技术栈、系统架构的全新需求

自主可控

支撑移动云差异化技术能力的升级
统一存储平台以整合技术资源、优化效率

新架构的设计思路

系统分层

天权存储底座

1. 解耦业务层和存储层，实现统一存储能力
2. 业务层基于底层存储语义进行业务语义支持

简单

易开发、易运维

1. 基于主从强一致性，抛弃Quorum复杂性
2. 基于中心化的路由和集群管理
3. 追加写语义

高性能

极致性能

1. 面向NVMe全闪介质
2. 支持RDMA高速网络能力
3. 全链路极致性能（rpc升级、用户态技术、零拷贝等）
4. IO链路至多一次落盘能力

系统功能分层

计算侧接入层
高性能虚机IO接入、路由存储节点、SLA流控、热迁移

块设备业务层
存储协议转换、LBA索引管理、负载打散、数据回收

分布式存储语义层
多副本强一致性、数据分层、数据修复

存储引擎层
多介质管理、空间管理、磁盘管理

一、移动云及其存储现状

二、极速型云盘自研转型

三、极速型云盘架构详解

四、未来展望

云盘数据模型

如何将云盘映射到业务层数据模型

伸缩性

容量、性能易拓展

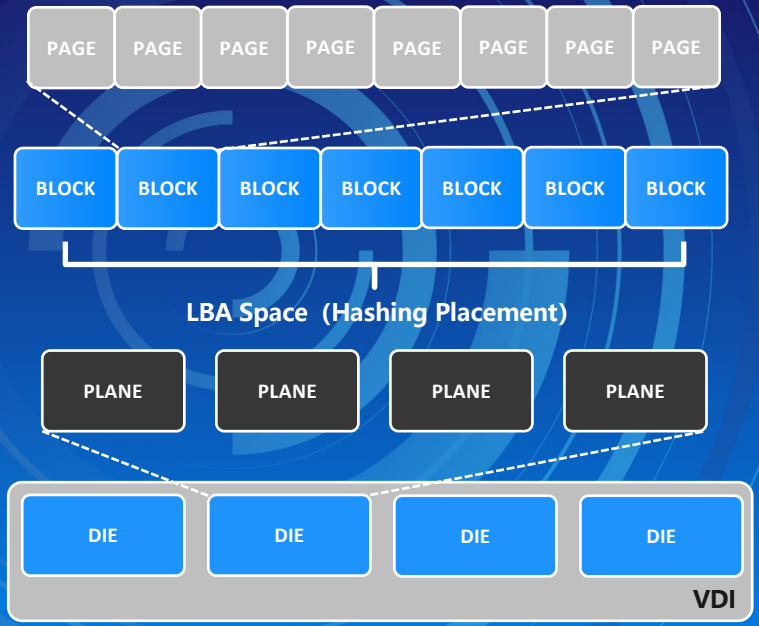
性能打散

降低性能热点

简化路由

消除元数据负担

1. 每块云盘可通过动态添加DIE来实现容量和性能的 scale up
2. 通过调整PLANE大小实现性能和容量的分配关系，实现细粒度性能打散
3. 通过动态哈希映射BLOCK到PLANE，通过基于负载的静态路由放置PLANE，整体路由规模可控



业务层数据模型

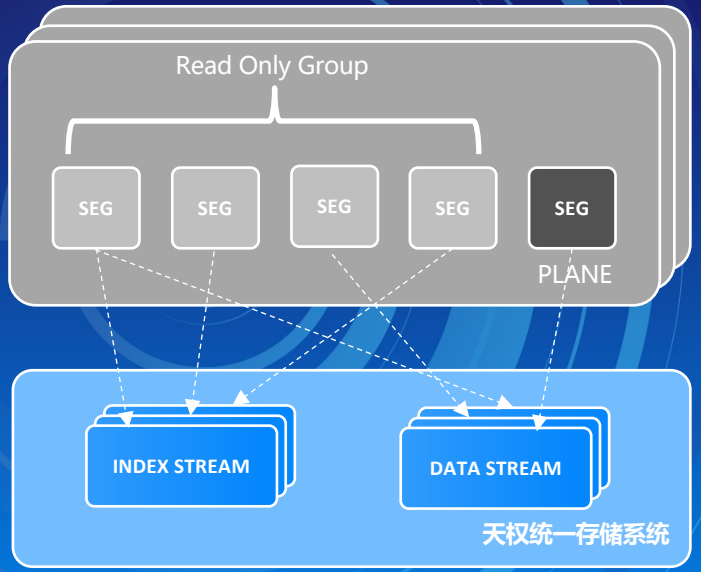
如何将业务层数据模型映射到存储系统数据模型

无状态
灵活调度能力

极致性能
最短IO路径

单点写入
协调写冲突

1. 通过将所有状态持久化在底层存储实现无状态（share everything），调度灵活、容灾迅速、disaggregated storage
2. 索引查询全内存化（索引压缩技术）、索引异步下刷（保证索引完整性）、全链路零拷贝
3. 随机转顺序能力
4. 单写保证，协调多点挂载的写冲突，保证索引局部化

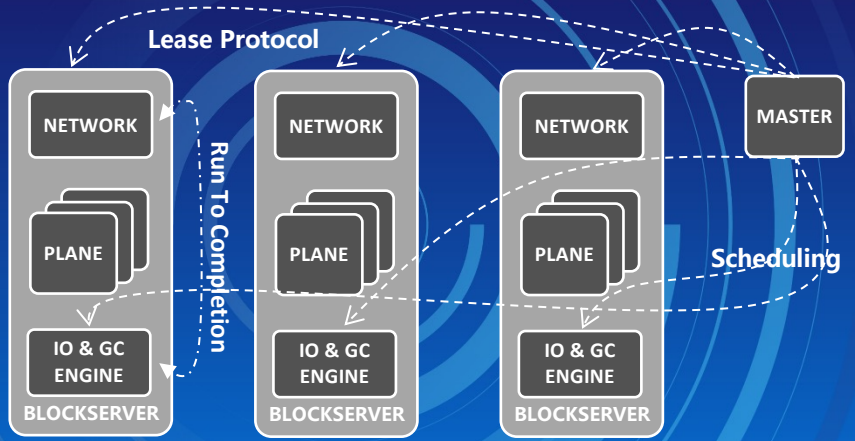


业务层系统模型

业务特性　集群管理　SLA管理

块服务特性　弹性伸缩、容灾、空间回收　负载调度、流控、波动处理

1. 实现了完整的快服务功能：秒级无损崩溃一致性快照、备份回滚能力
2. 中心化管控实现对处理能力的灵活伸缩
3. 通过分布式租约进行节点状态和路由管理
4. 通过数据回收操作进行 SLA 可控的回收操作
5. 对后端存储进行观测，对 IO 波动/慢操作进行主动规避和响应



存储层数据模型

追加写

简洁、介质友好

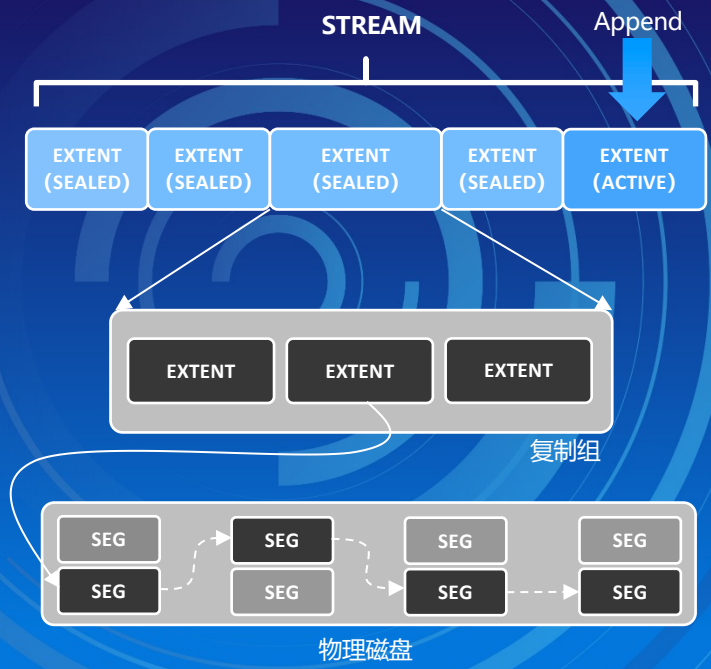
大分片

索引规模小

主从同步

实现简单、强一致性

1. STREAM是由非定长EXTENT组成的有序列表，只有尾部EXTENT可写
2. 追加写天然对介质友好，在Flash场景尤为重要，且系统简单（快照、ransomware方案、数据修复）
3. EXTENT通常为GB级别，大分片导致索引规模小易管理（降低数据库依赖）
4. EXTENT调度灵活，可应对写波动快速故障转移
5. 多副本采用主从强一致，无复杂一致性协议的负担，支持 N – 1 节点故障时的可用性



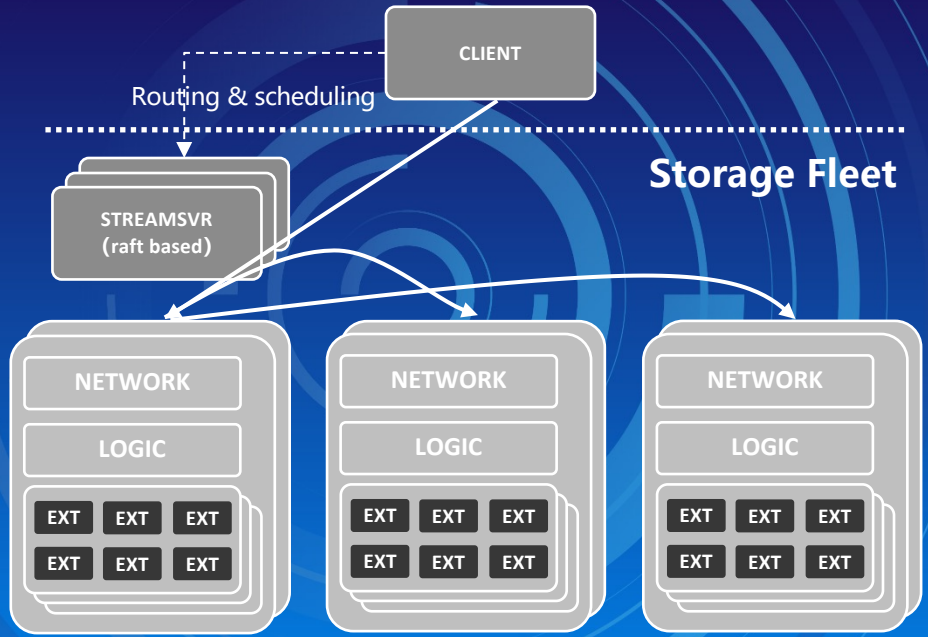
存储层系统模型

全RDMA
双栈网络支持

用户态IO
极低开销

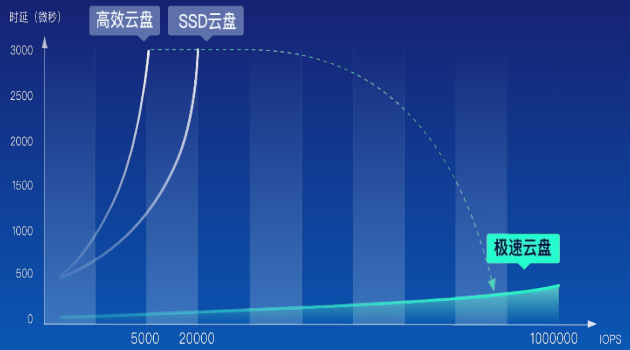
超大规模
弹性伸缩

1. 数据面全RDMA部署，支持TCP降级服务
2. IO链路全用户态实现，RTC机制
3. 轻量级元数据面中心管控、伸缩灵活



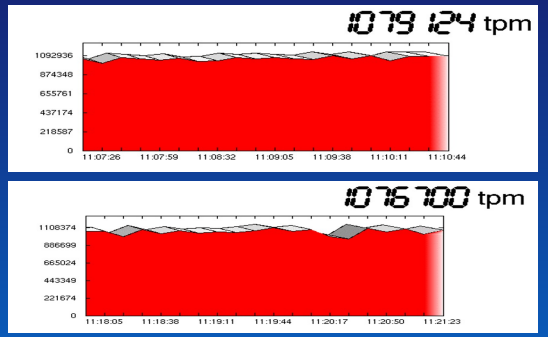
产品进展

极速型PL2（100K IOPS）、PL3（1M IOPS）已在移动云官网可订购



较前代产品，IOPS性能提升超10倍；时延下降超80%

可有效支撑大型数据库、实时日志分析等IO密集型以及AI训练、基因测序等高吞吐型业务场景。



OLTP性能测试：AWS io2 Block Express vs 极速型 PL2 云盘

8C32G云主机、2TB PL2云盘可达 107 万tmp能力

一、移动云及其存储现状

二、极速型云盘自研转型

三、极速型云盘架构详解

四、未来展望

技术布局

网络

软硬件融合

用户态协议栈建设（资源利旧）
全协议栈硬件卸载
RDMA异构设备标准化

存储

软件定义闪存

Zoned Namespace / FDP
提升产品 QoS

产品

效能提升

QLC & 纠删码应用
Disaggregate Storage

Thanks_