分布式存储取代集中式存储可行性分析

Feasibility Analysis of Distributed Storage Replacing Centralized Storage

金天骄（中国移动通信集团浙江有限公司，浙江 杭州 310030）
Jin Tianjiao（China Mobile Group Zhejiang Co.，Ltd.，Hangzhou 310030，China）

摘　要：
分布式存储在运营商多业务场景的使用过程中，在性能和软件成熟度两方面逐渐暴露出一系列问题。从性能、成熟度和成本3个维度，对分布式存储和集中式存储展开对比分析，深入探讨分布式存储和集中式存储的适用场景，并指出分布式存储的优化方向，为国内运营商云计算资源池中存储资源的规划设计提供有效建议。

关键词：
IO时延；吞吐量；数据重平衡；软件成熟度
doi：10.12045/j.issn.1007-3043.2019.10.019
中图分类号：TN919
文献标识码：A
文章编号：1007-3043（2019）10-0084-04

Abstract：
In the process of using distributed storage in multi-service scenarios of operators，a series of problems are gradually exposed in terms of performance and software maturity. From three dimensions of performance，maturity and cost，the distributed storage and centralized storage are compared and analyzed，the applicable scenarios of distributed storage and centralized storage are deeply explored，and the optimization direction of distributed storage is put forward，which provides effective suggestions for the planning and design of storage resources in the cloud computing resource pool of domestic operators.

Keywords：
IO delay；Throughput；Data rebalance；Software maturity

引用格式：金天骄. 分布式存储取代集中式存储可行性分析[J]. 邮电设计技术，2019（10）：84-87.

0 引言

近年来，国内运营商高举“去IOE”大旗，对传统的IT架构进行了大刀阔斧的改革。去“I”方面，通过应用改造将系统迁移至x86资源池，完成应用系统“去I”；通过将OLTP型核心系统数据库迁移至一体机，非核心迁移至x86资源池，将OLAP型数据库迁移至大数据平台，完成数据库系统“去I”。去“O”方面，通过区分在线交易、联机分析和实时查询等不同场景，引入PostgreSQL、MySQL、Hadoop和Gbase等技术组件，逐步替换Oracle数据库。去“E”方面，这些年通过大规模引入分布式存储，停止采购集中式存储，部分运营商省分公司集中式存储容量占比已降至20%以下。在分布式存储替代集中式存储的过程中，尤其是分布式存储经过多年的业务场景历练，准备向高性能、核心业务场景覆盖的关键时刻，本文对分布式存储技术进行了深入思考，分析现阶段分布式存储是否能够全面取代集中式存储，并且对未来存储的发展方向提出观点。

1 分布式存储的主要问题

软件定义存储（SDS——Software-defined storage）是一种数据存储方式，使用通用的x86设备组成分布式存储替换传统集中式存储，是一种新的高性价比的

收稿日期：2019-07-03

存储技术解决方案。某运营商省分公司自2015年开始对国内几个大厂商的分布式存储产品进行功能及可靠性测试，2016年初开始在虚拟化资源池进行小范围部署，2017年扩大部署规模并接入重要业务系统，根据2018年的年中统计，目前分布式存储承载的核心系统包含中心化、CBOSS、CRM、统一开通等；非核心系统包含日志分析、数据稽核、统一认证和渠道管理系统等。

目前现网由H公司和T公司2个厂家的分布式存储产品提供服务。在使用过程中，2家公司的产品集群产生了相同的故障，其中最主要的3类故障如表1所示。

表1　分布式存储产品故障表

故障现象	故障诱因	故障根因	问题归类
文件系统只读、IO慢	节点硬件HANG住	硬件HANG检测机制问题	软件成熟度问题
MQ业务报错	数据重平衡时间长、过程中对IO影响较大	数据重平衡算法、机制问题	软件成熟度+性能问题
营业厅、渠道业务量波动	集群IO性能达到瓶颈	服务器硬件配置不合理/集群节点数不足	性能问题

通过故障分析总结，目前现网分布式存储存在的问题可分为集群本身的性能和软件成熟度两大类问题，再结合现状深入剖析发现这两大类问题又包含以下5个具体问题，具体如表2所示。

表2　分布式存储问题分析表

性能	IO吞吐问题	单个分布式存储集群节点数有限且承载各类业务系统，经常出现主机iowait较高，前端响应缓慢，达到性能瓶颈
	IO时延问题	集群负载较高或者节点异常导致存储时延增高，进而导致时延敏感型业务频繁报错
	数据重平衡问题	节点异常后剔除或恢复加进集群时，数据三副本重构过程占用较大IO资源
软件成熟度	软件功能问题	集群软件对节点硬件异常时的hang检测机制不成熟，不能快速隔离异常节点导致整个集群受到影响；数据重平衡周期较长，长时间影响集群IO性能
	软硬件兼容问题	专用分布式存储软件和通用硬件的兼容问题，导致某些硬件关键指标无法采集和监控，诸如hang检测等部分功能无法实现

2 集中式存储和分布式存储的对比分析

2.1 全方位对比分析

根据上述分布式存储存在的问题，结合目前分布式存储的使用现状，从性能、软件成熟度和成本估算3个维度对集中式存储和分布式存储进行了全方位的对比分析，具体如表3所示。

表3　集中式存储和分布式存储总体对比分析表

分类	具体指标	集中式存储	分布式存储
性能	IO时延	稳定时延1~2 ms	平均时延3~5 ms
	IO吞吐（IOPS、带宽）	受控制器限制，双控~16控，无法线性扩展，性能存在瓶颈	采用分布式控制器，可线性扩展至4 096节点，性能容量线性增长
	数据重平衡	底层raid磁盘组数据同步，性能几乎无影响	多副本冗余、同步性能无影响
软件成熟度	软件功能	功能成熟稳定	版本灵活定制升级，支持多种新特性
	软硬件兼容性	软硬一体、完全兼容	兼容性较强
成本	每T成本	软硬一体，专有设备，成本高昂	软硬解耦；软件+廉价通用X86硬件；成本低廉

2.2 性能对比

2.2.1 时延

集中式存储由控制器与磁盘柜组成，主要部件包括各类控制、数据处理芯片以及高速缓存。分布式存储系统采用可扩展的系统结构，利用多台存储服务器资源构成一个虚拟的存储设备。由于架构不同，集中式存储和分布式存储的IO路径和时延也存在差异，通过对比发现前者的IO写时延有较大优势，具体如表4所示。

表4　集中式存储和分布式存储时延对比表

指标	集中式存储	分布式存储
读时延	①IO路径：集中式存储由于读缓存命中率低，需从HDD硬盘读数据，数据从硬盘读取后即返回读取成功，耗时较少 ②物理链路：SAN链路时延较低且比较稳定	①IO路径：分布式存储IO读流程较写流程路径更短，从任一副本磁盘获取数据即可，时延较写入明显降低 ②物理链路：IP链路时延相对较高且容易受干扰产生波动
写时延	①IO路径：集中式存储IO写流程路径较短，数据由控制器写入cache后即返回IO写入成功，耗时较少 ②物理链路：SAN链路时延较低且比较稳定	①IO路径：分布式存储IO写流程路径较长，数据写入主OSD后需等三副本OSD均完成写入方可返回IO写入成功，耗时较长 ②物理链路：分布式IP链路时延相对较高且容易受干扰产生波动

综上所述，集中式存储由于写入环节较少，在写时延方面明显优于分布式存储。读时延方面集中式存储和分布式存储均相对较低，但通过对比发现集中式存储通过专用设备优化，在读时延上依然有一定的优势。

2.2.2 IO吞吐

IO吞吐包括IOPS和带宽2项指标，集中式存储取决于硬盘的类型数量、控制器、缓存等硬件配制，而分布式存储性能取决于单个节点性能和节点数量。理论上分布式存储可通过增加节点来线性增加IO吞吐，但在实际使用中鉴于稳定性、管理成本、时延等因素，

各分布式存储厂商对于实际生产中集群规模的建议一般是20~40个节点。

a）稳定性考虑：一个资源池节点数过多，发生故障的概率也随之增加，发生故障后数据需要进行重平衡，平衡过程对系统感知影响极大。节点数增多会使集群复杂度增加，进而导致集群IO调度开销增加，加上时延等因素导致性能提升偏离线性曲线。

b）硬件及管理成本考虑：随着集群中节点数量不断增加，存储集群的硬件成本和管理成本会随之增加，内存资源、网络资源、故障域管理复杂度都会呈几何倍数增加。

c）时延考虑：业界分布式存储的最佳实践是20~40个节点，通过1台48端口标准交换机进行组网。集群规模越大，节点越多，整个集群需要横跨更多交换机，导致时延增加。

综上所述，由于集群节点数量并不能真正线性扩展，因此资源池的性能也就有了上限。比如常见的40个节点分布式存储集群性能可到50万IOPS、5 ms左右时延，高端配制可超过80万IOPS，2 ms左右时延。目前相同容量的主流高配集中式存储性能接近200万IOPS、1.3 ms左右时延。分布式存储相比集中式存储还是有一定差距的。此外，由于分布式存储的三副本特性，一份数据需要写入3份，额外占用IO资源，因此对于写密集型应用，分布式存储在性能上不如集中式存储。

2.2.3 数据重平衡

当存储介质磁盘出现故障损坏或进行修复时，存储软件均会进行数据重构以保证数据的多副本可靠性。数据重构周期与存储的硬件性能、重构数据量、前端业务压力有关。综合对比分析分布式存储和集中式存储在数据重构上的性能，得出如表5所示结论。

表5 集中式存储和分布式存储数据重平衡对比表

指标	集中式存储	分布式存储
数据平衡	①损坏及影响范围：磁盘规格性能统一，单个磁盘组内磁盘数量少，单个数据盘损坏，重构数据量较小，影响范围小 ②重构特点：数据重构在存储磁盘组内进行，路径短，时延小 总结：实际生产中数据重构业务几乎无感知	①损坏及影响范围：数据分散在集群中不同节点设备上，涉及节点和硬盘数量较大，在负载较大情况下影响可能较大，尤其时存储节点宕机或raid卡损坏等情形，会涉及大量数据盘 ②重构特点：数据重构在网络中各个节点之间进行，路径长，时延大，增加网络带宽消耗，存储性能降低 总结：实际生产中数据重构时间较长，占用IO较高，业务感知明显

2.3 成熟度对比

集中式存储在IT架构中长期扮演重要角色，久经市场和时间考验；而分布式存储作为近年来的新技术产品，本身发展起步较晚，成熟度还需要不断提高，加之硬件非定制化，通用软件结合通用x86硬件的稳定性还需要进一步验证。两者在成熟度方面的对比分析如表6所示。

表6 集中式存储和分布式存储成熟度对比表

指标	集中式存储	分布式存储
成熟度	①软件功能：商业产品起步发展较早，多年来历经产品迭代，各种数据功能丰富成熟，如容灾、重删、压缩、快照等，经过市场长期验证 ②软硬件兼容性：百分百兼容，软硬一体研发，深度定制优化，硬件性能及特性得到充分发挥利用，数据可靠性强，产品稳定性高	①软件功能：分布式存储起步发展较晚，虽有厂家定制研发，但x86硬件本身可靠性不高，目前企业级功能尚不完善且缺乏市场验证 ②软硬件兼容性：各厂商软件只针对自己硬件配置模型研发定制，对于运营商采购的通用x86服务器不能做到完全兼容，硬件性能不能完全发挥，且部分场景下基础监控管理功能也无法启用

2.4 成本对比

集中式存储成本取决于控制器数量、配置规格、硬盘数量和类型，以及各种附带软件功能特性，因此单位容量成本需考虑的因素较多。同样分布式存储也取决于x86服务器的CPU、内存、硬盘数量及类型、分布式软件特性。表7和表8分别是普通配置和全闪存配置的集中式存储和分布式存储价格对比。

表7 普通配置集中式存储和分布式存储价格对比

普通配制	集中式	分布式
配置	双控制器、512G缓存、16×16G FC、240×1.2T 10K SAS	4个服务器节点、2×4114 CPU、4×32G内存、2×600G 10K SAS，2×960G SSD，12×6T SATA
性能	18 947/8.6 ms	27 312/9.74 ms
裸容量/TB	288	288
可用容量/TB	216（按3/4的利用率）	96（三副本）
价格/万元	144	56（含软件）
每TB成本/万元	0.66	0.58

表8 全闪存配置集中式存储和分布式存储价格对比

全闪存	集中式	分布式
配置	双控制器，1 024G缓存，24×960G SSD	4节点X86服务器，2×5118 CPU，单节点4×32G内存，12×480G SSD
性能	181 132 /1 ms	201 600/2.64 ms
裸容量/TB	23	23
可用容量/TB	17.25（按3/4的利用率）	7.67（三副本）
价格/万元	65.3（经验值）	32.7（经验值，含软件）
每TB成本/万元	3.79	4.26

在性能相近的基础上，分布式存储采用三副本机制保障系统可靠性，相比集中式存储整体利用率较低

且需要较高的软件授权费用，仅在低性能场景中价格稍占优势，在高性能场景的全闪存配置上，分布式存储价格要略高于集中式存储。

2.5 对比总结

综上所述，目前分布式存储实测结论与部分业界流行观点偏差较大，分布式存储相比集中式存储在某些方面的优势并不明显。然而，上述结论的依据为产品现阶段的实测数据，后续随着分布式存储技术的不断成熟，分布式存储软件license的不断降低，结论也会随之发生变化。

传统集中式存储虽然有技术封闭、商务绑定等风险，但软硬一体化及成熟度高的特点使其在运维、性能、稳定性等方面有不容忽视的优势。分布式存储虽然在软硬件解耦、价格方面有优势，但目前其在性能和稳定性方面还存在着诸多不成熟之处。现阶段分布式存储需要通过升级软硬件技术和整合维护体系来解决目前遇到的问题。

3 结论

3.1 现阶段集中式存储与分布式存储适用场景建议

由于现阶段集中式存储在时延、稳定性和软件成熟度方面有明显优势，因此较适合应用于有数据强一致性要求、高IOPS、低时延的OLTP类交易业务场景；而分布式存储适用于海量数据、大数据块、顺序读写、时延不敏感的场景。这2类存储技术的使用建议如表9所示。

表9 集中式存储和分布式存储使用建议表

场景	计费账务	CRM订单	OA、邮件	资源池	备份归档
IO	高	高	高	中	高（备份） 中（归档）
时延要求/ms	低（<2）	低（<2）	中（<5）	中高（<10）	高（10~30）
容量要求	中高（<300TB）	中（<100 TB）	高（200~500 TB）	高（200 TB~1 PB）	高（200 TB~几PB）
所需功能	双活、复制、快照	双活、复制、快照	重删、压缩	重删、压缩、QoS	重删、压缩、分级存储
推荐架构	高端集中存储或全闪存分布式+IB	高端集中存储或全闪存分布式+IB	分布式块或文件SAS+IB	分布式块或文件SAS+10GE	分布式块或文件HDD+10GE

3.2 分布式存储优化方向

根据分布式存储在生产中遇到的问题，结合第2章中集中式存储和分布式存储的对比分析，分布式存储需要从性能和成熟度2方面不断进行优化改进。具体措施如表10所示。

表10 分布式存储优化思路表

分类	对应问题	优化思路
性能	IO时延	升级到高端产品系列，采用全SSD硬盘和IB交换机的方案，降低链路和数据存取时延
	IO吞吐	提高服务器SSD和HDD的缓存配比，用PCIE接口SSD替换SATA接口SSD，使用25G、40G以太网组网
	数据重平衡	对资源池接入的业务系统加以控制，针对不同业务类型接入不同存储池，避免存储集群在数据重平衡时的性能瓶颈影响到业务
软件成熟度	软件功能	升级集群软件版本并定期更新
	软硬件兼容性	针对部分IO敏感高、稳定性要求高的场景，可采用软硬一体化方案
成本	单位成本昂贵	使用ceph、MooseFS、GlusterFS等开源产品，加快定制化服务器研究

4 结束语

近年来分布式存储技术的发展及广泛应用，是因为网络技术的飞速进步以及闪存SSD盘的大量应用，解决了存储高时延和单节点IO性能瓶颈两大问题。将来随着硬件的不断发展和分布式软件功能的不断成熟完善，相信分布式存储定会逐步从中低端存储市场走向高端存储市场。

本文建议国内运营商在现阶段扩大分布式存储使用范围，但短期内不宜完全舍弃集中式存储，尤其对于技术研发力量较弱的单位，应当根据应用场景合理规划存储类型，确保生产系统稳定运行。同时在分布式存储的日常运维过程中要加强架构管控，确保分布式存储单节点以及集群整体的性能，加强兼容性评估测试，定期升级软件版本，加紧核心技术课题攻关，培养出新一代的能创新、善思考的技术团队，加强核心能力内化，为运营商后续业务快速迭代创新奠定扎实的IT基础。

参考文献：

[1] 金天骄，张式勤，王晓征. 运营商云化之路借鉴“BAT”的思考[J]. 移动通信，2017(19)：7-14.

[2] 金天骄. SDS技术在运营商资源池中的应用[J]. 电信技术，2016，8(10)：66-69.

作者简介：

金天骄，工程师，硕士，主要从事云计算、大数据相关行业咨询、方案规划以及技术研究工作。