贾晨刚,王玮. 基于气象大数据云平台的文件存储优化设计[J]. 陕西气象,2025(5):82-86.
文章编号:1006-4354(2025)05-0082-05

基于气象大数据云平台的文件存储优化设计

贾晨刚[1,2],王 玮[1]
(1. 陕西省气象信息中心,西安 710014;
2. 中国气象局秦岭和黄土高原生态环境气象重点开放实验室,西安 710014)

摘 要:在气象大数据云平台运行阶段,其分布式文件存储系统出现异常现象:服务器端存储空间使用量与设备管理界面显示数据存在显著偏差,实际可用容量低于预设设计阈值。经开展系统性排查与机理分析,明确该问题的核心诱因在于存储系统数据管理颗粒度粗放,以及资源动态分配机制存在设计缺陷。为解决上述问题,研究团队实施了存储管理系统升级方案,同步完成设备参数的优化配置。具体措施包括:(1)对数据存储单元进行精细化管控,建立存储块的动态映射机制;(2)构建负载均衡模型,实现数据在存储节点间的最优分布;(3)设计多重容错架构,通过数据同步机制降低单点故障风险。实践验证表明,此次技术优化使平台文件存储服务的可用性、可靠性显著提升,为气象监测预报业务的多源数据整合、实时分析及历史归档提供了规范化的存储支撑,有效保障了业务流程的连续性与稳定性,为后续数据存储方案的优化设计与实践应用提供了可借鉴的技术参考与理论依据。

关键词:气象大数据云平台;分布式存储;存储引擎;优化设计
中图分类号:TP302 **文献标识码**:A

气象大数据云平台构建了数据驱动与算法集成深度融合的集约化架构,形成统一开放的服务型技术平台。该平台通过“数算一体”创新范式实现异构数据资源与计算资源的协同治理,其中“数”指多模态气象数据集,“算”指智能分析算法库,“数算一体”则强调二者在存储层、计算层及应用层的三维耦合机制。该平台支撑全维度气象数据资产的汇聚治理与服务化封装,覆盖观测数据流、业务产品库、政务数据、行业关联数据、设备状态日志及系统监控流等多源异构数据体系。平台构建弹性计算引擎,可系统化提供数值模式预报之外的数据产品加工流水线、深度挖掘分析服务及可视化建模支持,适配气象业务全域应用场景[1-4]。当前平台已集成 12 大类 460 项标准化数据服务接口,构建 PB 级混合存储体系,其中文件存储系统承载占比 40%。值得注意的是,文件存储服务依赖的分布式存储集群若发生读写异常,将导致基础数据服务链断裂,对强数据驱动型的气象业务体系将产生级联式影响,凸显了该存储架构在平台运行体系中的关键支撑作用。

1 分布式存储

气象大数据云平台基于分布式存储架构构建文件服务子系统,采用对象存储与块存储协同机制,重点承载雷达回波图像、卫星遥感影像、数值模式输出等海量非结构化数据集。系统通过标准化 API 接口实现数据服务化封装,支持按需检索、格式转换及流式传输等功能[5-7]。同时,承担气象观测结构化数据的二级备份使命,采用文件级冗余策略保障地面观测报文、高空探测数据等关键业务资料的持久化存储,具体配置见表 1。

分布式存储采用 EC 8+2:1 冗余策略。EC(erasure code),全称为纠删码。“:1”含义为将存

收稿日期:2023-12-15
作者简介:贾晨刚(1982—),男,汉族,陕西户县人,硕士,高工,主要从事气象信息系统运维与技术研发工作。
基金项目:秦岭和黄土高原生态环境气象重点实验室开放基金课题(2024G-10)

表 1　分布式存储配置

型号	节点数	具体配置	可用空间	空间利用率
H3C UniStor X10000 G3	6	每节点配置:2 颗 Intel 4110(2.1 GHz/8 核/85 W) CPU;256GBi 内存;1 块 960GBi SSD 硬盘;2 块 600GBi SAS 硬盘;35 块 6TBi SATA 硬盘	900 TBi	88.89%

储所包含 6 个存储节点中其中的一个存储节点作为备份节点,“8+2”含义为将存储节点上的数据划分条带,每个条带由 8 个数据块(Data)和 2 个校验块(Parity)组成,同一个条带的 8+2 个块均匀地存储在所有节点上,以实现节点和磁盘级别容灾。因此,使用 EC 8+2:1 策略,任意 1 个节点或者不同磁盘组中的 2 个磁盘故障,仍可保证整套存储中数据可用性。

2　研究背景与问题界定

2022 年 8 月,运维团队监测到气象大数据云平台分布式存储子系统出现异常 I/O 阻塞,文件服务层出现持续读写超时。通过深度巡检发现,该存储集群挂载的两个 LUN 卷物理容量合计 393TBi,仅占理论可用空间的 47.8%,但存储管理系统界面显示已用容量达 91.28%,且集群健康度指标跌落至 66%。进一步分析揭示存储数据层存在显著偏差:服务端通过 df 命令获取的挂载点使用率为 48.3%,而存储管理界面统计的数据使用率已达阈值(90%)。这种“逻辑满-物理空”的矛盾现象,初步判定为由数据碎片化引发的空间错配。当存储池数据占用超过预设阈值时,触发集群自我保护机制,导致存储服务进程僵死,最终引发存储节点宕机。值得注意的是,本案例中出现了服务端与存储端的空间计量差异。这种异构计量体系的冲突,暴露出传统阈值监控机制在分布式存储场景下的局限性,需在数据存储模型、空间计量算法及健康度评估体系等维度开展深入研究。

3　实验室环境下的验证试验

3.1　故障定位与归因

首先从存储系统架构进行分析。该存储是以 Ceph 分布式存储系统作为底层技术架构而搭建起来的。Ceph 使用 CRUSH 算法(controlled replication under scalable hashing),提供块设备存储、文件系统存储和对象存储三种存储功能。经对存储系统参数核查发现,BlueStore HDD 存储引擎的最小分配单元为 64 KBi,而存储系统中存在大量如地面资料、高空资料等小文件;基于此,进一步对存储文件的类型划分及各类文件的空间占用情况展开梳理分析[8]。梳理情况具体见表 2。

表 2　单个文件存储量级和各类数据空间占用情况汇总表

文件量级	数据种类	空间占用量	空间占用率
[1 KBi, 50 KBi]	地面资料、高空资料、海洋资料、辐射数据、农气数据、大气成分、气象灾害、服务产品、其他数据	约 169.0 TBi	约 43%
(50 KBi, 5 MBi]	雷达数据	约 27.5 TBi	约 7%
(5 MBi, 400 MBi]	数值预报、卫星数据(部分)	约 59.0 TBi	约 15%
(400 MBi, 500 MBi]	卫星数据(部分)	约 137.5 TBi	约 35%

汇总分析结果显示,64 KBi 以下的小文件占总存储用量的 45%。鉴于 BlueStore HDD 存储引擎的最小分配单元为 64 KBi,即使单个文件大小不足 64 KBi,系统仍按 64 KBi 进行空间分配。据此推测,存储容量统计不一致的具体成因可能在于:大量小文件在当前分配机制下产生了显著的空间冗余损耗[9]。

3.2　故障复现实验

为探明其具体成因,在 H3C 实验室搭建一台同型号存储设备进行试验。试验存储采用与实际

生产环境相同的版本、软件配置，采用 EC 4+2:1 的冗余策略(试验存储节点数有限)，存储引擎 BlueStore HDD 的最小分配单元同样设置为 64 KBi。

通过存储性能测试工具 Vdbench 写入 100 万个 33 KBi 大小(按文件存储量级的小文件平均值得出)的文件，数据量为 31.5 GBi，占用容量为 47.2 GBi，然而存储集群全局变量 GLOBAL 中的已用存储空间总量参数 RAW USED 为 456 GBi，说明试验存储存在空间浪费情况。

用试验结果分析空间浪费原因。在采用 EC 4+2:1 冗余策略下，默认条带大小为 16 KBi，trunk(对于小文件，采用合并存储的方式，合并存储后的文件称为 trunk 文件)大小为 4 KBi(根据数据块数量进行计算)，同时 BlueStore HDD 的最小分配单元为 64 KBi，则一个 33 KBi(考虑冗余为 49.5 KBi)的文件在 BlueStore 层占用的空间如下图所示，需要占用 64 KBi×6=384 KBi 的空间，大约浪费 334.5 KBi 的容量，且该空间无法回收再利用。因此，计算 1 个 33 KBi 文件造成的空间损失是实际文件大小的 10.14 倍(334.5/33≈10.14)。具体数据块分布情况见图 1。

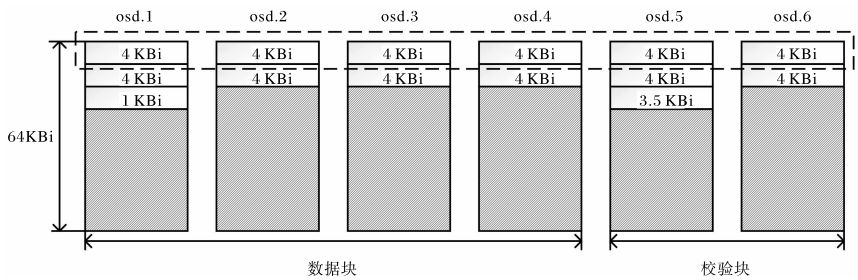

图 1 数据块分布图

图 2 中 64 KBi 为 BlueStore HDD 的最小分配单元；EC 4+2:1 策略下一份数据分为 4 个数据块+2 校验块分布在 6 个 osd 上面，4 KBi 为 trunk 大小，Stripe 为条带。

3.3 归纳与梳理

根据试验结果推算可得，造成空间容量浪费的主要原因是在现有 EC 模式下，由于存储引擎 BlueStore HDD 的最小分配单元为 64 KBi，实际写入的文件约有 50%为大小不足 64 KBi 的小文件。根据现有业务分析，假设小文件大小均取中值为 33 KBi，理论上 1 个 33 KBi 的文件需要存储空间是(考虑冗余)为 49.5 KBi，但实际在存储设备上落盘按照现有 EC 8+2:1 冗余策略，则需要占用 64 KBi×10=640 KBi 空间，大约浪费 590.5 KBi(640−49.5=590.5)的存储空间，且这部分空间被占用无法回收。由此得出，1 个 33 KBi 文件造成的空间损失是实际文件大小的 17.89 倍(590.5/33≈17.89)。

因此，当文件存储的数据量达到 393 TBi，同时小文件占比超过 45%，而存储集群全局变量 GLOBAL 中的已用存储空间总量参数 RAW USED 为 1 008 TBi，导致文件落盘容量超过阈值，从而引起存储节点宕机。

4 优化路径与实施步骤

4.1 优化路径

根据试验结果进行问题总结，在存储引擎 BlueStore HDD 最小分配单元设置为 64 KBi 的情况下，一个 33 KBi 小文件造成约 17.89 倍的空间浪费，如果将 BlueStore HDD 参数调小，可以降低空间浪费；但试验时同时发现，在存储系统的计算硬件即 CPU 和内存不变的情况下，将存储引擎 BlueStore HDD 调整到更小，是以增加存储落盘计算量为代价的，会引起 CPU 和内存一定程度的增加，造成落盘效率降低[10]。由试验得到，将 BlueStore HDD 的最小分配单元设置为 8 KBi，既能保证存储系统的 CPU 和内存运行在稳定合理范围内，又能降低≤16 KBi 的小文件的空间浪费率。

故障处置思路如下：鉴于气象大数据云平台作为核心数据支撑平台，对实时业务连续性存在

极高要求,为避免对实时数据服务产生干扰,需优先实施历史数据迁移操作。具体方案为:新增一台华为存储设备,将原故障存储中文件大小处于1 KBi～50 KBi区间的小文件全部迁移至该新存储节点;通过释放故障存储的部分空间,使此前超限的单块磁盘容量回落至阈值范围内,从而恢复实时数据服务的正常运行。

故障修复后,升级存储卷管理系统,引入对象存储服务(Object Storage Service)概念。同时优化相关存储设备参数,将备用存储BlueStore HDD的最小分配单元为8 KBi,设计其为针对≤16 KBi的小文件存储设备,实现梳理的数据种类按单个文件大小情况进行不同存储分类存放。具体文件存放分类见表3。

表3　优化改造后分布式存储配置和文件存放情况

型号	节点数	具体配置	可用空间	空间利用率/%	存放数据种类
H3C UniStor X10000 G3	6	每节点配置:2颗Intel 4110(2.1 GHz/8核/85 W)CPU;256GBi内存;1块960GBi SSD硬盘;2块600GBi SAS硬盘;35块6TBi SATA硬盘	900 TBi	25.77	雷达数据 数值预报 卫星数据
华为OceanStor 9000 V5	6	每节点配置:2颗Intel4214R(2.4GHz/12核/100W)CPU;128GBi内存;2块1.6TBi NVMe SSD硬盘;2块600GBi SAS硬盘;36块8TBi SATA硬盘。	1 152 TBi	14.67	地面资料 高空资料 海洋资料 辐射数据 农气数据 大气成分 气象灾害 服务产品 其他数据

4.2　实施步骤

(1)针对新增小文件存储设备,采用如下存储策略[11-12]。

升级华为存储卷管理系统OBS,替代本地文件系统Ext3/Ext4。同时,通过存储引擎BlueStore HDD调整条带块大小为8 KBi,使数据分布更加均衡,存储实际可用容量最大程度接近预期可用容量。

(2)针对原有大文件存储设备,进行存储I/O性能优化策略[13]。

升级H3C存储卷管理系统OBS,使条带块大小自适应,增加条带宽度,依据数据控制器节点和存储设备的容量进行数据均衡,提高数据访问性能。

(3)优化数据修复性能[14]

存储系统目录开启小文件聚合,将若干个小文件"聚合"成大文件进行修复,极大降低了修复的对象数量。

优化前后系统架构和数据部分对比见图2。

5　结语

经系统性优化后,气象大数据云平台文件存储系统完成了存储架构的精细化升级。具体而言,引入动态分级存储机制,并结合一致性哈希算法对数据分布模型进行了重构,实现了存储结构的优化迭代。实测结果表明,优化后实际可用容量与理论可用容量的偏差系数控制在±1.5%范围内,从根本上消除了"逻辑满-物理空"空间计算方法不一致的矛盾。优化后的存储模型有效解决了存储碎片化问题,为海量气象数据的全生命周期管理提供了规范化的技术保障。

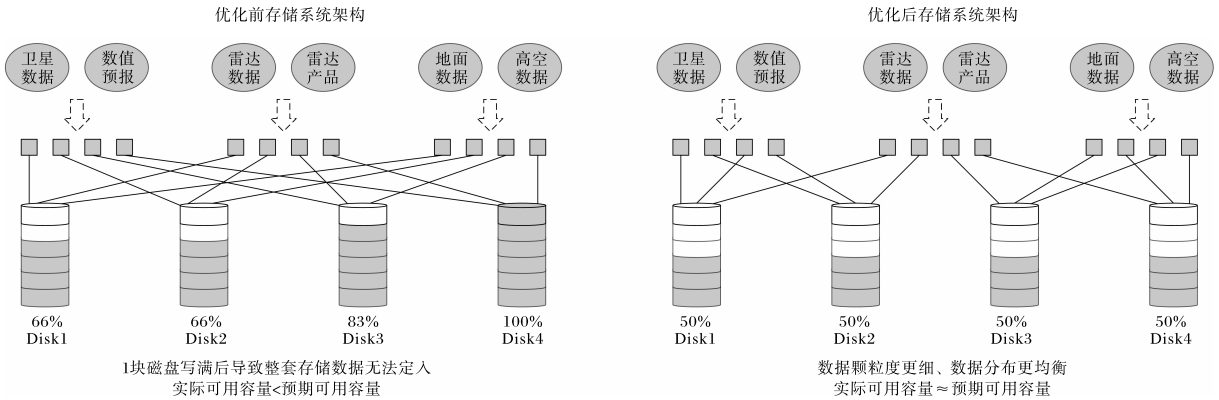

图 2 优化前后系统架构和数据分布对比

参考文献：

[1] 中国气象局.“十四五”气象信息网络业务发展规划[Z]. 北京：中国气象局，2021.

[2] 国家气象信息中心. 气象大数据云平台“天擎”1.0应用开发手册(V1.2)[Z]. 北京：国家气象信息中心，2020.

[3] 国家气象信息中心. 气象大数据云平台设计方案[R]. 北京：国家气象信息中心，2018.

[4] 刘媛媛，何文春，王妍，等. 气象大数据云平台归档系统设计及实现[J]. 气象科技，2021，49(5)：697-706.

[5] 许竹霞，张春燕，徐娟. 甘肃省气象大数据云平台的存储与服务系统设计[J]. 信息技术与信息化，2022(2)：53-57.

[6] 周笑天，冯勇，陈益玲，等. 基于 Hadoop 的气象数据分布式存储技术研究[J]. 信息技术，2022(1)：68-74.

[7] 雷鸣. 气象大数据分布式存储设计与实现[J]. 计算机技术与发展，2021，31(5)：193-197.

[8] 穆彦良. Ceph 存储技术中 CRUSH 算法的研究与改进[D]. 成都：成都信息工程大学，2016.

[9] 徐敏，胡聪，王萍，等. 基于强化学习的 Ceph 文件系统的性能优化[J]. 微型电脑应用，2022，38(3)：83-86.

[10] 夏亚楠，王勇. Ceph 存储系统中节点的容错选择算法[J]. 桂林电子科技大学学报，2022，42(5)：384-390.

[11] 陈晓丹，庞双龙，曾德生，等. Ceph 存储系统在云计算环境中的应用[J]. 电子技术，2020，49(8)：40-42.

[12] 夏畅. Ceph 分布式存储技术研究与设计[J]. 网络安全和信息化，2021(6)：93-96.

[13] 席磊. Ceph 分布式存储系统在 OpenStack 云平台的设计与实现[J]. 数字传媒研究，2021，38(8)：37-43.

[14] 张晓，张思蒙，石佳，等. Ceph 分布式存储系统性能优化技术研究综述[J]. 计算机科学，2021，48(2)：1-12.