计算机研究与发展 DOI:10.7544/issn1000-1239.2020.20190549
Journal of Computer Research and Development 57(2):306-317,2020

基于“采集—预测—迁移—反馈”机制的主动容错技术

杨洪章[1] 杨雅辉[1] 屠要峰[2] 孙广宇[3] 吴中海[1]
[1](北京大学软件与微电子学院 北京 102600)
[2](中兴通讯股份有限公司 广东深圳 518057)
[3](北京大学信息科学技术学院 北京 100871)
(yanghongzhang@pku.edu.cn)

Proactive Fault Tolerance Based on “Collection—Prediction—Migration—Feedback” Mechanism

Yang Hongzhang[1], Yang Yahui[1], Tu Yaofeng[2], Sun Guangyu[3], and Wu Zhonghai[1]
[1](*School of Software & Microelectronics, Peking University, Beijing* 102600)
[2](*ZTE Corporation, Shenzhen, Guangdong* 518057)
[3](*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

Abstract Hard disk fault has become the main source of failure in data centers, which seriously affects the reliability of data. The traditional data fault tolerant technology is usually realized by increasing data redundancy, which has some shortcomings. Proactive fault tolerant technology has become a research hotspot, because it can predict hard disk failures and migrate data ahead of time. However, the existing technology mostly studies hard disk fault prediction, but lacks the research of collection, migration and feedback, which causes difficulty in commercialize. This paper proposes a whole process proactive fault tolerant on “Collection—Prediction—Migration—Feedback” mechanism, which includes time-sharing hard disk information collection method, sliding window record merging and sample building method, multi-type hard disk fault prediction method, multi-disk joint data migration method, and two-level validation of prediction results with fast feedback method. The test results show that the impact of collecting hard disk information on front-end thread is only 0.96%, the recall rate of hard disk fault prediction is 94.66%, and data repair time is 55.10% less than traditional methods. This work has been used stably in ZTE's data center, which meets the objectives of proactive fault tolerance technology, such as high-reliability, high-intelligence, low-interference, low-cost and wide-application.

Key words disk failure; storage reliability; fault tolerance; artificial intelligence; operation & maintenance

摘 要 硬盘故障是数据中心最主要的故障,严重影响了可靠性.传统的数据容错技术一般都是通过增加数据冗余来实现的,存在缺陷.主动容错技术通过预测硬盘故障提前将数据迁移,成为研究热点.现有

收稿日期:2019-08-15;**修回日期**:2019-11-06
基金项目:国家重点研发计划项目(2018YFB1003302);国家自然科学基金项目(61672062);江苏省工业和信息产业转型升级专项资金项目(2018GX02517)
This work was supported by the National Key Research and Development Program of China (2018YFB1003302), the National Natural Science Foundation of China (61672062), and the Jiangsu Provincial Program of Industrial & Information Industry Transformation (2018GX02517).
通信作者:杨雅辉(yhyang@ss.pku.edu.cn)

技术大多研究硬盘故障预测,缺乏采集、迁移、反馈的研究,难以商用.提出“采集—预测—迁移—反馈”全流程主动容错机制,包括:分时硬盘信息采集方法、滑动窗口记录合并及样本构建方法、多类型硬盘故障预测方法、多盘联合数据迁移方法、预测结果二级验证及快速反馈方法.测试表明:采集硬盘信息对业务影响仅0.96%,硬盘故障预测召回率达94.66%,数据修复时间较传统方法减少55.10%.该工作已在中兴通讯的数据中心稳定商用,满足了主动容错技术在高可靠、高智能、低干扰、低成本、广适用等核心目标.

关键词 硬盘故障;存储可靠性;容错;人工智能;运维

中图法分类号 TP309.3

1 概 述

1.1 硬盘故障频发并带来灾难性后果

据国际数据公司(International Data Corporation, IDC)发布的《数据时代2025》白皮书[1],到2025年全球数据量将会达到175 ZB,如果以12 TB容量的硬盘来计算,大约需要170亿块.而硬盘的年故障率在1%左右[2],因此全球每年将有数以亿计的故障硬盘出现.对于一个PB级规模的数据中心,硬盘故障每天都在发生[3].不可否认的是,硬盘故障已经成为数据中心最主要的故障来源[4].硬盘故障会直接导致灾难性后果,如数据丢失、业务中断等,这无疑严重影响了数据的可靠性.

1.2 重新审视硬盘故障

正如人类个体在一生中经历的“生老病死”,硬盘个体也在经历着“健康态、亚健康态、濒临故障态、故障态”的必然性周期,这是因为随着时间的流逝,硬件各部件的老化、磨损等原因导致的结果.并且,与人类因先天生理缺陷导致的新生儿夭折,以及诸如车祸、地震、灾害、动物袭击等意外造成的偶发性死亡类似,硬盘也存在因出厂缺陷和意外故障(如甲醛、震动、电压突变、空气湿度过大、运维人员操作不当等)而造成的偶发性故障.

显然硬盘的必然性故障具有一定时间的濒临故障窗口期,是具备预测的可能性的.如果在濒临故障窗口期内及时将数据迁移,则可避免系统降级服务.而硬盘的偶发性故障往往是突发的,几乎不存在濒临故障窗口期,因此不具备预测的可能性,或需要相当大的代价去预测,但故障不可避免地迅速发生,数据根本来不及迁移,这样的预测也没有意义.

1.3 传统被动容错存在缺陷

传统的数据容错技术一般都是通过增加数据冗余来实现的[5],主要有副本[6-7]、纠删码[8-9]、备份快照[10-11]等.这些技术都是在硬盘故障发生后,通过冗余满足读写访问和数据恢复,是被动式的处理手段.其缺陷体现在2方面:

1) 硬盘容量、恢复速度、读写服务之间存在矛盾.在数据恢复过程中,系统资源不可避免地被数据恢复占用,系统立即降级服务.因此,系统面临两难选择——如果数据恢复过快,系统正常读写受到严重干扰;如果数据恢复过慢,再出现新的故障盘将有可能引发数据永久丢失.并且,大容量硬盘的恢复需要更多的时间,进一步加深了上述矛盾.

2) 成本、性能、可靠性之间存在矛盾.副本技术虽然可以保证性能,但是存储空间利用率仅为$1/n$(其中n为副本个数),由此带来成本上涨问题.纠删码技术在每一次读写操作时额外增加了计算量,对性能产生负面影响,但是存储空间利用率较副本技术有了大幅提升.备份及快照技术可以保证性能,也不占用过多的额外存储空间,但是备份及快照时间节点之后的数据却无法保护.

1.4 新型主动容错成为研究热点

随着智能运维[12]的不断发展,通过采集硬盘SMART(self-monitoring analysis and reporting technology)指标,结合机器学习算法来预测硬盘故障,从而提前将数据迁移的主动容错技术[13]成为新的研究热点.一些国内外的研究成果[14-21]对单一品牌型号的硬盘预测其故障的准确率达到85%以上.

主动容错的好处不言而喻:1)在硬盘故障之前准确预测并将数据迁移到其他硬盘,避免了系统降级服务,提升了系统的可靠性.2)能够有效指导硬盘采购规划,克服了传统被动容错在故障发生后才去采购硬盘的缺陷.3)减少了运维人员的干预,被动容错需要运维人员的临场判断和快速响应,而主动容错在预测和处理过程中充分发挥了机器的优势.4)准确性高,传统被动容错依赖运维人员的经验来对硬盘故障进行手工检测,费时费力且误判率高.

虽然主动容错的现有技术在实验原型系统中取得了令人惊叹的故障预测准确率，但是在真实的业务场景中仍然问题百出、难以商用.其主要体现在：

1）大规模 SMART 采集引发的灾难问题亟待解决.以固定周期对全体硬盘采集 SMART 时不可避免地占用系统资源.经测试，在 1 万块硬盘的情况下，并行采集过程需经过 6 s 以上，期间系统几乎无法响应任何正常读写操作，且经常出现卡死的情况.因此，在大规模数据中心一次性采集全体硬盘 SMART 的方式迫切地需要改进.

2）针对 SAS(serial attached small computer system interface)硬盘和固态硬盘(solid state disk，SSD)的故障预测的空白亟待填补.现有文献均仅针对 SATA(serial advanced technology attachment)硬盘建立预测模型，而在实际的数据中心，SAS 硬盘和固态硬盘的数量十分庞大.直接将 SATA 硬盘的预测模型套用在 SAS 硬盘和固态硬盘是不可行的，这是因为它们的 SMART 差别巨大.此外，开源采集工具 smartmontools 直到版本 6.0 和 6.1 才陆续支持了在 Windows 操作系统和 Linux 操作系统中对 SAS 硬盘的 SMART 采集，直接导致对 SAS 硬盘的 SMART 采集年限短、积累的故障样本少.固态硬盘作为电子式的存储器件，相较于机械硬盘，其年故障率低[22-23]，且大规模应用的年限远短于 SATA 硬盘和 SAS 硬盘，固态硬盘同样存在故障数量少的问题.因此迫切的需要专门对 SAS 硬盘和固态硬盘的故障进行建模预测，如此才能完整地预测数据中心的全体硬盘故障.

3）正负样本严重不均的难题亟待解决.现阶段阻碍硬盘故障研究的最大问题是故障盘数量少、健康盘数量多、正负样本严重不均衡.现有的技术文献往往通过 SMOTE(synthetic minority over-sampling technique)算法[24]来人工合成若干正样本，以缓解该问题，但是缺点在于易产生模型过拟合的问题.因此迫切地需要提出新的上采样方法，在不引起过拟合的前提下增加数倍的正样本，从而彻底解决该难题.

4）难以快速数据修复的问题亟待解决.现有技术让濒临故障盘完全独立地承担数据修复，短时间内连续集中的数据访问会加速故障的发生，并且重构时间窗口过长，因此迫切地需要提出多盘联合修复，以加快数据修复.

5）预测结果的验证与反馈机制亟待建立.预测错误包括故障盘的误判、漏判、迟判.误判会浪费硬盘生命周期.漏判和迟判会使得系统进入降级状态，需要依赖传统被动容错保障可靠性.通过算法优化能在一定程度上减少预测错误，但无法完全做到 100% 避免误判，在出现预测错误时，现有技术往往通过更新预测模型的方式进行矫正，然而这种方式存在时间滞后性，因此迫切需要更加灵活的反馈机制.

1.5 本文的主要贡献

为了克服现有主动容错技术的上述缺陷，本文提出了一系列的关键技术，涵盖“采集—预测—迁移—反馈”的全周期主动容错，包括：1)提出分时硬盘信息采集方法，解决了大规模 SMART 采集引发的灾难问题；2)提出滑动窗口记录合并及样本构建方法，填补了针对 SAS 硬盘和固态硬盘故障预测的空白；3)提出多类型硬盘故障预测方法，攻克了正负样本严重不均的难题；4)提出多盘联合数据迁移方法，加快了数据的修复速度；5)提出预测结果二级验证及快速反馈方法，灵活并快速地处理了预测错误.

2 相关研究

基于硬盘故障预测和数据迁移的主动容错技术受到了学术界的持续关注，国内外的研究人员从不同角度相继提出了一些方案.

关于硬盘故障预测准确性提升：Pitakrat 等人[14]仅选取硬盘 SMART 作为状态特征数据，测试了 21 种不同的分类算法，其中 NNC，RF，C4.5，REPTree，RIPPER，PART，K-Star，SVM 等算法可将故障预测的准确性达到 90%以上；Zhu 等人[15]使用了 SMRAT 相关项在最近一段时间的变化值作为特征数据输入，采用 SVM 作为分类算法建立预测模型，故障预测的准确性达到最高 95%；Li 等人[16]提出了基于决策树的预测模型，能达到 95% 以上的准确率，并且能够提前一周预测出故障，决策规则清晰地解释了 SMART 属性值与故障之间的关系，为采取措施并减少故障提供依据；柳永康[17]提出二级预测方法，在预测磁盘是否即将要故障的基础上，进一步预测磁盘故障的发生时间范围，使用逻辑回归算法，在提前 5 天预测的情况下，预测准确率最高达 85.01%.

关于硬盘故障模型架构：Xiao 等人[18]提出一种基于在线随机森林的硬盘故障预测模型架构，解决了离线训练和模型老化的问题，故障预测准确率达到 93%～99%.Xie 等人[19]提出一种基于一对多建模的硬盘故障预测模型 OME(optimized modeling engine)，准确率总体比以前的工作高出 18.5%.

关于数据提前修复:Ji 等人[20]在预测到硬盘将要发生故障的情况下,主动将该硬盘上的数据迁移到健康盘,但限制数据迁移的速率,从而在硬盘故障发生时难以完成全部的数据迁移,仍旧需要依靠传统被动容错;Qin 等人[21]提出了 Fatman 系统,对冷、热数据分别使用 RS(Reed Solomon)码和副本机制,对于将要发生故障的硬盘提前进行数据迁移,在故障发生以后,热数据可以通过其他副本来提供服务,而冷数据需要重构丢失的数据,这样可减少 76.3%的重构开销.

综上所述,现有研究工作主要关注提升 SATA 硬盘的故障预测准确率,忽略了针对 SAS 硬盘和固态硬盘的预测,并且对采集、迁移、反馈的研究较少,难以形成一个完整的主动容错技术方案.

3 主动容错技术

本节围绕中兴通讯承建并运营的国内华南某数据中心的硬盘情况进行分析研究,提出完整的主动容错机制,涵盖采集、预测、迁移、反馈等各个环节.该数据中心共有 129 887 块硬盘,在 2018 年共出现 1995 块故障盘,其品牌、类型的情况如表 1 所示,H_{true}为实际健康盘的数量,F_{true}为实际故障盘的数量.

Table 1 Hard Disk Situation in Data Center

表 1 本文研究数据中心的硬盘情况

Brand	SATA		SAS		SSD	
	H_{true}	F_{true}	H_{true}	F_{true}	H_{true}	F_{true}
Western Digital	46 378	737	3 641	62	1 334	42
Seagate	8 776	112	14 187	164	0	0
Intel	0	0	0	0	4 843	73
HITACHI	5 536	84	42 783	716	0	0
Toshiba	0	0	349	4	65	1

Note: H_{true} stands for the truly healthy disk; F_{true} stands for the truly failed disk.

3.1 分时硬盘信息采集方法

硬盘在故障前必然存在一系列的内外部征兆,现有文献过多地关注以 SMART 为代表的内部征兆,而对硬盘 IO 情况、CPU 使用率、内存占用等外部征兆研究较少.处于亚健康和濒临故障的硬盘,其 SMART 情况不稳定、时好时坏,一旦在采集的瞬间其恰巧处于健康状态,这会直接导致误判.在这种情况下,研究硬盘故障的外部征兆则是非常必要的补充,例如硬盘的 IOPS 和吞吐量很低,并且 CPU 使用量也很低但时延很高,这种情况同样预示了硬盘即将故障.因此,本文采集硬盘的信息主要包括:硬盘 SMART、硬盘 IO 情况、CPU 使用率、内存占用等.

在采集频率上,现有的文献均采用固定周期的采集方式,例如每天采集 1 次、每 3 h 采集 1 次等.诚然,采集越频繁越有利于对故障的预测,能够更加敏锐地捕捉到“震荡”型的 SMART 表现.然而考虑到采集硬盘信息对数据中心业务的必然影响,采集的频率也不能过于频繁.结合中兴通讯数据中心视频业务的特点,本文的采集方法为:1)考虑视频业务高峰,原则上每个盘每小时采集 1 次,但关闭业务高峰时段(11:30—13:00,18:00—23:30)的采集.2)考虑视频业务特点,电视节目多数以整点和半点作为切换点,因此每次采集时段为 5～25 min,35～55 min.3)避免同时采集全体硬盘,应以 1 s 为间隔,每次采集 10 块盘.至此,以 2～4 TB 构成的百 PB 级的数据中心,在 1 h 之内可以采集完毕,且对业务影响降到最低.4)在业务低谷时段(2:35—5:25)更新模型、批量发送采集数据至计算节点.

由于采集间隔非固定,对硬盘的写 IO 通过当前写入速度描述,其计算公式为

$$WriteSpeed_k=\frac{Sector_k-Sector_{k-1}}{Time_k-Time_{k-1}}, \tag{1}$$

其中,k 表示第 k 次采集.读 IO 的计算同理,此处不再赘述.

由于硬盘 SMART 数据各数据项取值的规整方法不同,所以各数据项在数值上差异很大,如果按原值输入作模型训练时,数值较大的项会带来较大影响,因为每个数据项其原始数值的取值规则不同,为了防止个别数据项对模型带来较大偏差,因此进行归一化处理,其计算公式为

$$a_{normal}=\frac{a-a_{min}}{a_{max}-a_{min}}. \tag{2}$$

3.2 滑动窗口记录合并及样本构建

通过 3.1 节的采集方式,每天每块硬盘采集 16 条记录,经过 1 年的采集,129 887 块硬盘中共出现故障盘 1 995 块.健康盘与故障盘的比例约为 64:1,面对如此严重的不均衡,本文采用滑动窗口记录合并及样本构建的方法解决该问题.

如图 1 所示,对于故障盘,其在故障时刻前 30 天之内的记录作为故障记录,每个记录项按采样时间

先后进行排序,设定 3 天为时间窗口,截取时间窗口内的 48 条记录.时间窗口起始位置放在硬盘的故障时间上,然后时间窗向前滑动 0.5 天距离,即向前移动 8 条记录,共滑动 55 次,直至时间窗口涵盖故障前第 30 天的所有记录.如图 2 所示,对每次移动时间窗口所截取的记录,以记录项为单位对其计算平均值、方差、极差,从而将连续多个时间点的各记录项信息合并为 1 条正样本,将其计入训练样本集中.通过窗口的滑动,将构建 55 倍于原始故障盘个数的正样本数.对于健康盘,则随机选取连续 3 天的 48 条记录,同样针对每个记录项计算平均值、方差、极差,作为负样本计入训练样本集中.

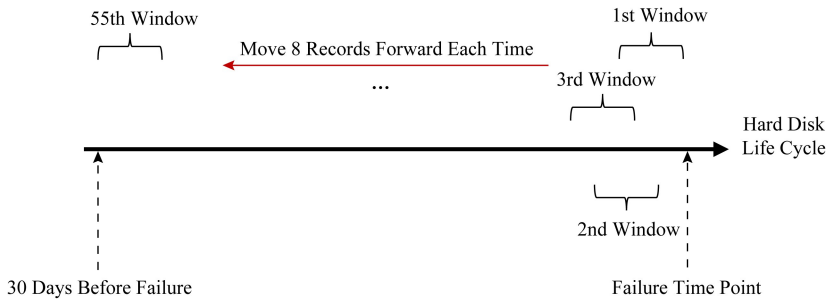

Fig. 1 Sliding window record merging

图 1 滑动窗口记录合并

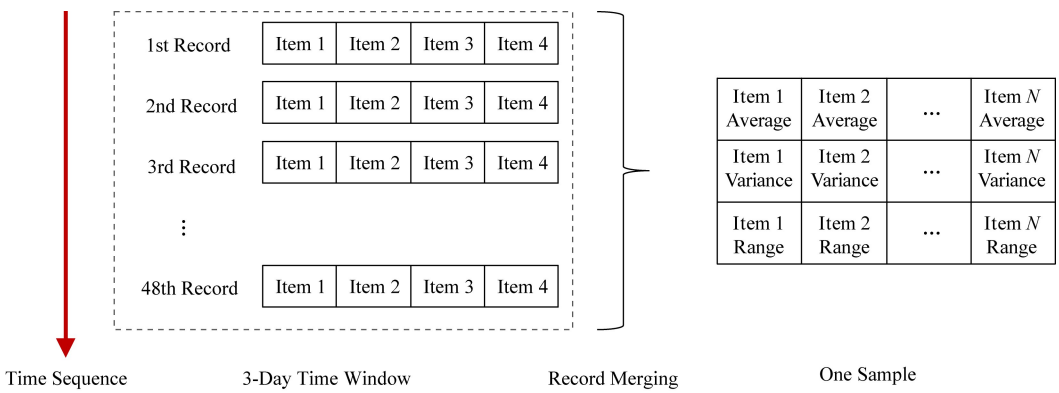

Fig. 2 Sample construction

图 2 样本构建

通过滑动窗口记录合并及样本构建方法,构建出 55 倍故障盘的正样本,相较于健康盘与故障盘之间 64 倍的比例差距,这几乎填补了二者之间的鸿沟,正负样本达到了相同的数量级.此外,在单条样本中综合体现多个时间点的硬盘状态,不再孤立的审视单一时间点的硬盘状态,能够更加敏锐地发现并刻画硬盘在故障前的各项指标的“陡增”、“陡减”、“震荡”等状态,从而为故障硬盘的判定提供了充分的依据.

3.3 多类型硬盘故障预测

硬盘 SMART 是对硬盘各组件,如磁头、马达、盘片等部件的状态进行分析监控的技术,并非所有信息均与硬盘故障相关,图 3 展现了 4 类典型的 SMART 信息在硬盘故障前的表现,分别是震荡型、陡增型、陡减型以及平静型,显然平静型的 SMART 是无助于区分硬盘是否将要故障的,因此需要特征筛选.

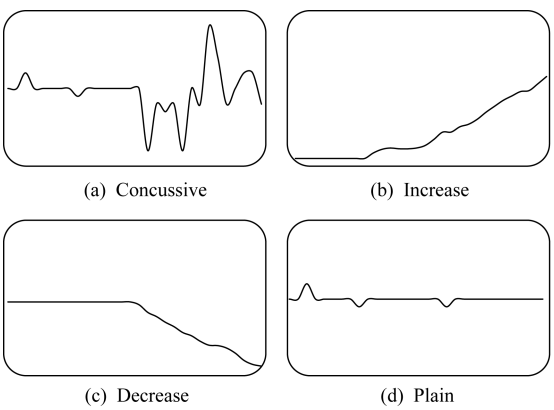

Fig. 3 Typical SMART information before disk failure

图 3 典型的 SMART 信息在硬盘故障前的表现

本文研究的数据中心含有 10 种不同类型及品牌的硬盘,因不同型号的硬盘其采集到的 SMART 存在差异,甚至同一个 ID 的含义也可能不同,因此需针对每一种类型及品牌的硬盘分别进行建模.本文在特征筛选时,采用专家经验判断、变化趋势观察、卡方检验、属性方差、树的特征选择这 5 种方法,在 3 种以上方法出现的 ID 最终被选取.受篇幅限制,表 2 仅展现了 3 个典型的硬盘类型品牌用于预测建模的 SMART 选取情况.这些 SMART 与 3.1 节所述的外部征兆一起,用于建立故障预测模型.

Table 2 Selected SMART for Predicting Different Hard Disks

表 2 本文在不同类型硬盘选取的 SMART 指标

Disk	Brand	Selected SMART Item
SATA	Western Digital	Raw_Read_Error_Rate, Spin_Up_Time, Reallocated_Sector_Ct, Seek_Error_Rate, Power_On_Hours, Temperature_Celsius, Current_Pending_Sector, Offline_Uncorrectable, UDMA_CRC_Error_Count
SAS	HITACHI	Accumulated start-stop cycles, Accumulated load-unload cycles, Number of background scans performed, Attached phy identifier, Invalid DWORD count, Running disparity error count, Loss of DWORD synchronization, Phy reset problem, Phy event descriptors, Loss of DWORD synchronization count
SSD	Intel	Erase_Fail_Count, End-to-End_Error, Temperature_Case, Current_Pending_Sector, CRC_Error_Count, Host_Writes_32MiB, Media_Wearout_Indicator, Thermal_Throttle, Host_Writes_32MiB, NAND_Writes_32MiB

在算法方面,本文使用人工神经网络算法,设置神经网络隐含层个数为 4 个,每层的神经元个数分别设为 1 000,500,200,100,输入层神经元个数根据输入数据特征维度确定,输出层为 3 个,激活函数选择 tanh.为避免神经网络过拟合,将交叉熵代价函数和 L2 正则之和作为网络的损失函数.模型网络的优化算法使用了批量梯度下降法.事实上,本文在尝试了 10 余种不同的算法及百余种不同的参数设置后,发现算法的优劣对于硬盘故障预测的准确率影响微乎其微,而数据质量高情形下对预测的准确性明显优于数据质量低的情形,这充分说明了在硬盘故障预测的问题中数据质量的重要性远高于算法.

如图 4 所示,在对 10 种硬盘接口、品牌分别进行建模后,多类型硬盘故障即具备上线运行的能力,在数据中心连续采集 3 天硬盘信息后,即可预测硬盘是否将要发生故障,首先根据硬盘类型选择相应的模型,随后将连续 48 条样本进行输入,在超过 36 条样本被判定为正样本的情况下,该硬盘会被预测为濒临故障盘.在商用场景中,先引入初始模型,在系统的运行过程中定期进行模型更新和参数调节.

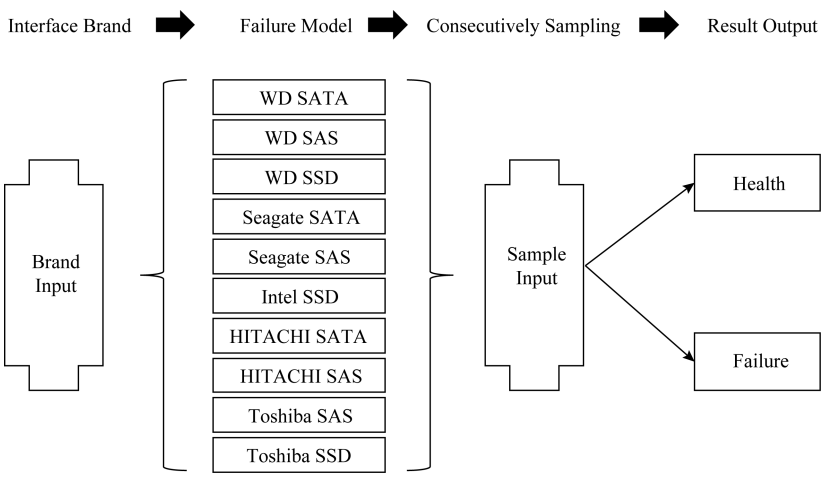

Fig. 4 Failure prediction of multi-type hard disk

图 4 多类型硬盘故障预测

3.4 多盘联合数据迁移

主动容错的最大价值是在预测到硬盘即将故障之时,利用该硬盘的剩余寿命把数据迁移到健康盘,从而避免进入降级模式,能够同时保证高性能和高可靠.然而现有技术让濒临故障盘独立承担数据修复,短时间内连续集中的数据访问会加速故障的发生,往往导致数据还未完全修复时故障就已发生,最终仍需依靠被动容错方法.因此,在主动容错

技术中,恢复数据的任务不应仅由濒临故障盘单独承担.

本节提出了多盘联合数据迁移技术,基于纠删码系统,通过多盘联合修复、拷贝与编解码计算相结合的修复手段,加快了修复速度,有效避免系统进入降级状态,同时均衡了网络传输压力.

多盘联合数据修复的主要步骤包括:

1) 在硬盘 A 被预测为濒临故障后,立即启动主动容错数据修复.

2) 确定参与共同修复的硬盘组,访问系统元数据,遍历硬盘 A 中的所有 p 个数据块(strip),读取其所属条带(stripe)的硬盘编号,这些硬盘参与共同修复,称之为硬盘组 B.

3) 选择空闲容量最大的健康硬盘作为修复目标盘,上述硬盘 A 和硬盘组 B 不得被选入,称之为硬盘 C.

4) 分别确定由硬盘 A 和硬盘组 B 负责修复的数据块集合.假设共有 p 个数据块待修复,为集合 Q.由硬盘 A 承担 $p\times y$ 个数据块的修复,为集合 T.由硬盘组 B 承担剩余的 $p-p\times y$ 个数据块的修复,为集合 R.

5) 随后同时进行硬盘 A 和硬盘组 B 的数据修复:通过拷贝的方法,将集合 T 中所有的数据块由硬盘 A 修复到硬盘 C;通过编解码计算的方法,将集合 R 所有数据块由硬盘组 B 修复到硬盘 C.

6) 如果硬盘 A 修复完毕,硬盘组 B 仍有 20% 以上的数据块尚未修复完成,则硬盘 A 接管部分硬盘组 B 的数据修复;反之亦然.

7) 如果硬盘 A 未修复完毕即发生故障,则系统进入降级模式,由硬盘组 B 承担全部剩余修复任务.

3.5 预测结果二级验证及快速反馈

现有技术文献缺乏对预测结果的进一步验证,并且在出现误判情况或主动修复策略不当时,无法及时改进,需要等待一定时间再将新采集的硬盘信息批量更新预测模型.因此本节提出预测结果二级验证及快速反馈方法,其原理如图 5 所示:

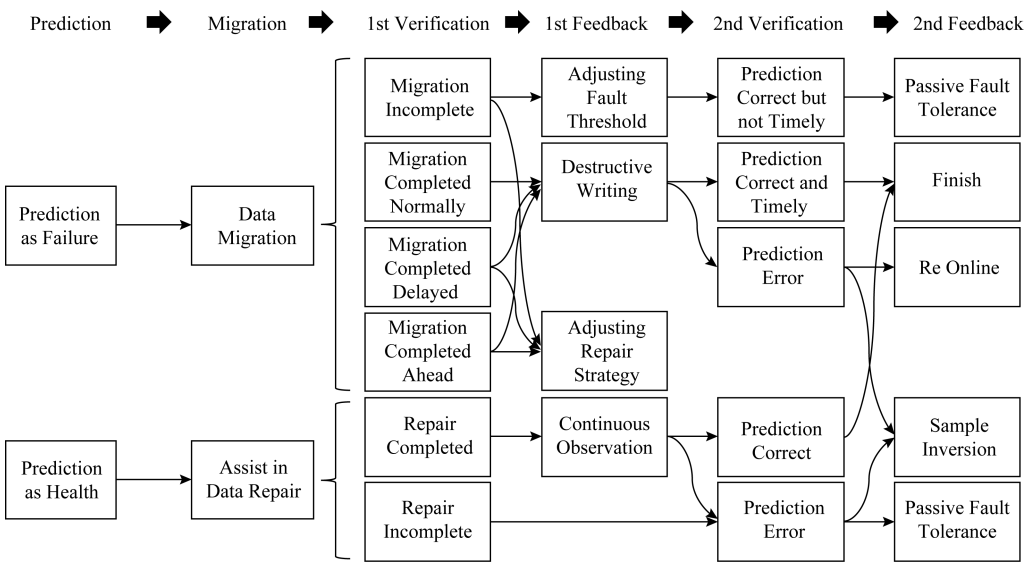

Fig. 5 Two-level verification and feedback

图 5 二级验证及反馈

对于被预测为故障的硬盘,立即进行主动修复:

1) 如果在修复过程中已出现故障,则系统降级服务,由健康盘完成剩余的所有修复工作,并且需调整故障阈值,后续应尽早将该盘预测为故障盘;

2) 如果该盘的修复顺利完成,但明显快于或慢于协助其修复的其他健康盘的修复工作,则需调整修复策略,增大或减小由濒临故障盘承担的修复数据比例;

3) 如果以上情况未出现,则对该盘进行破坏式写入,直至该盘发生故障为止,或写入时间达剩余生命周期阈值 z 时为止.记录写入时间 h.若 $h=z$,则将其认定为误判,反馈给训练模型;若 $h>z/4$,则适当上调故障阈值 x,并适当上调修复比例阈值 y;若 $h<z/4$,则参数保持不变.

对于被预测为健康的硬盘,应立即协助故障盘进行主动修复:

1) 如果在联合修复过程中出现故障,则该盘被误判为健康盘,需要样本反转为正样本.

2) 如果修复顺利完成,则持续观察,如果在 1 个月之内未出现故障,则预测正确,否则为预测错误.

在上述过程中,有 3 个重要阈值是反馈机制的重要组成部分,包括:

1) 故障阈值 x,若同一块硬盘的连续多条样本健康度低于 x,则认为该硬盘为濒临故障盘.

2) 修复比例阈值 y,由濒临故障盘承担的数据修复比例 y,由其他健康盘承担的数据修复比例为 $1-y$.该值的最理想情况是,由濒临故障盘的数据修复时间恰好等于其他健康盘承担的数据修复时间.一次主动数据修复由该 2 种修复同时进行,显然,主动数据修复的时间等于用时较大者的时间.

3) 剩余生命周期阈值 z,在主动数据修复之后,硬盘剩余生命周期越小越理想,最理想的情况是修复完毕时该盘恰好故障.若剩余生命周期大于该阈值 z,则适当调节 x 和 y.

通过二次验证及快速反馈方法,针对不合理的主动修复策略和不及时的故障预测,能够及时调整参数,在避免系统降级的前提下,能够精准利用濒临故障硬盘的剩余生命周期;针对错误的预测,能够快速甄别,及时反转正负样本标记,为模型更新提供准确的依据.

3.6 本节小结

主动容错技术框架如图 6 所示.通过分时硬盘信息采集方法、滑动窗口记录合并及样本构建方法、多类型硬盘故障预测方法、多盘联合数据迁移方法、预测结果二级验证及快速反馈方法五大技术,将主动容错技术形成完整闭环,从而具备了商业应用的条件.

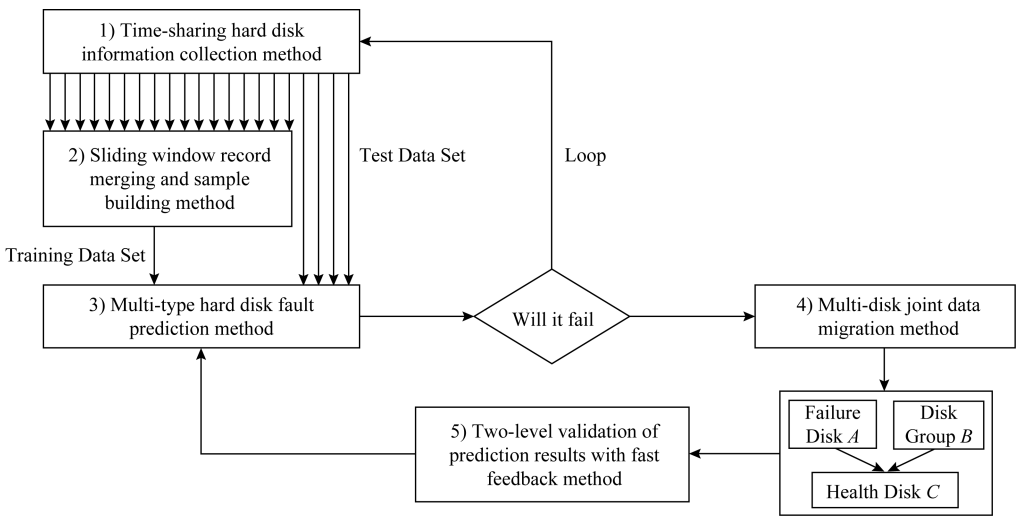

Fig. 6 Framework of proactive fault tollerance technology

图 6 主动容错技术框架

4 实验与结果

本节从 3 个方面对本文工作进行测试:

1) 测试采集硬盘信息对前台业务的干扰,其评价指标是相较于不采集硬盘信息的情况,前台业务的带宽下降比例以及用户视频播放的实际体验情况;

2) 测试硬盘故障预测的准确率,其评价指标是召回率和误检率;

3)测试数据修复速度,其评价指标是完成数据修复的时间.

4.1 测试采集硬盘信息对前台业务的干扰

前台业务是 20 个客户端分别播放 30 min 直播视频,其正常状态是该数据中心对所有的客户端提供等量、恒定的读数据带宽.图 7(a)为不采集硬盘信息的情况,图 7(b)为传统方法一次性采集全体硬盘的情况,图 7(c)为本文工作的情况.图 7 的纵坐标为客户端的播放带宽.

与图 7(a)相比,图 7(b)虽然平均带宽整体仅下降 0.88%,但在采集 SMART 时连续 9.18 s 系统不可服务,其前 11.21 s 及其后 4.23 s 出现明显性能抖动,用户在观看直播视频的过程中出现了超过 20 s 的严重卡顿;与图 7(a)相比,图 7(c)虽然平均带宽整体下降 0.96%,但全程未出现带宽性能抖动,用户对于视频的播放未察觉任何异常.由此可见,本文工作提升了用户体验,降低了对前台业务的干扰.

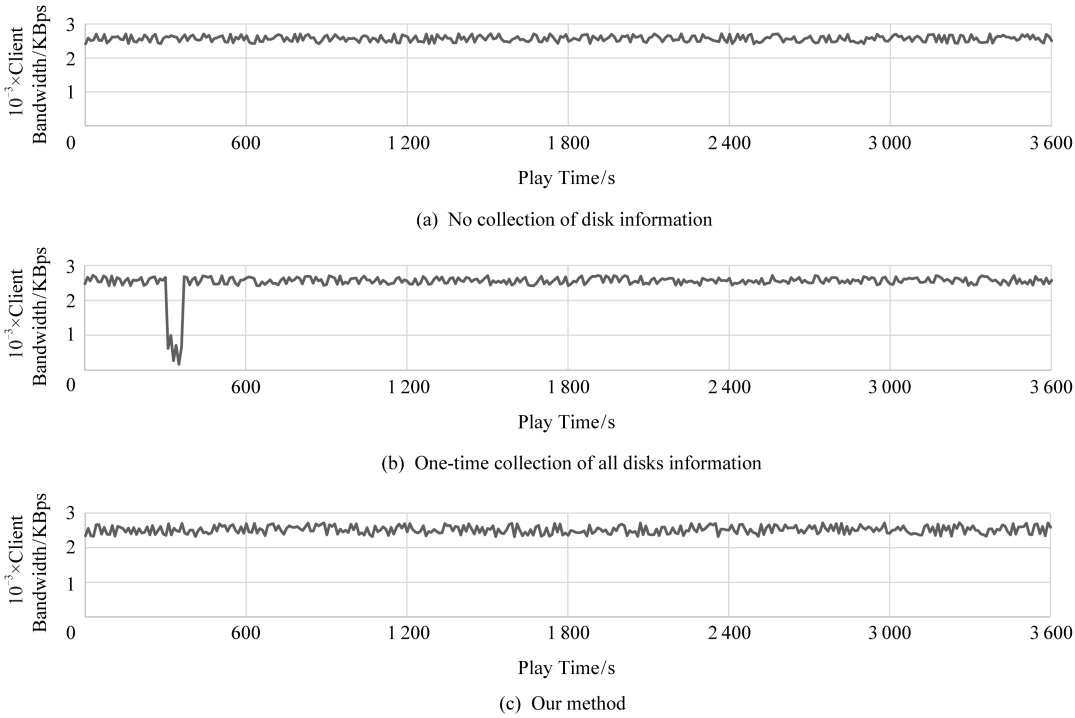

Fig. 7 Client bandwidth under different conditions

图 7 在不同情况下客户端播放视频带宽

4.2 测试硬盘故障预测准确率

在本节测试中,场景涵盖:1)基于中兴通讯已采集的数据进行建模和预测;2)基于中兴通讯已采集的数据进行建模,在中兴通讯真实场景中进行预测;3)基于 Backblaze 数据集进行建模和预测.并与其他文献公开的方法进行了比较.

4.2.1 面向已采集的硬盘数据进行预测

中兴通讯已采集的数据包括 127 892 块健康盘、1 995 块故障盘,其中 70%用于训练建模,30%用于预测测试.其总体测试结果如表 3 所示,$H_{\text{Predicted}}$ 为被预测为健康盘的数量,$F_{\text{Predicted}}$ 为被预测为故障盘的数量.本文工作总体的召回率为 94.66%,误检率为 0.34%.对于各类型的细分测试结果如表 4 所示,SATA,SAS,SSD 的召回率分别为 94.64%,94.37%,97.14%,误检率分别为 0.36%,0.35%,0.10%.可以看到,相较于机械硬盘,固态硬盘的预测准确性高、误检率低.

为了与本文工作进行比较,将文献[15]和文献[17]的方法分别作为对比系统 1 和对比系统 2,但受困于现有文献仅针对单一品牌类型的硬盘故障进行预测,本节仅选取了希捷 SATA 数据作为比较.如表 5 所示,本文工作、对比系统 1、对比系统 2 的召回率分别为 94.12%,88.24%,85.29%,误检率分别为 0.30%,2.10%,1.01%,本文工作显著降低了误检率.

Table 3 Prediction Results Based on Collected Data in Overall

表 3 基于已采集数据的总体预测结果

Type	$F_{\text{Predicted}}$	$H_{\text{Predicted}}$
F_{true}	567	32
H_{true}	102	38 266

Nete: $H_{\text{Predicted}}$ stands for the predicted healthy disk; $F_{\text{Predicted}}$ stands for the predicted failed disk.

Table 4 Prediction Results Based on Collected Data in Detail

表 4 基于已采集数据的细分预测结果

Disk	Type	$F_{\text{Predicted}}$	$H_{\text{Predicted}}$
SATA	F_{true}	265	15
	H_{true}	52	18 155
SAS	F_{true}	268	16
	H_{true}	49	18 239
SSD	F_{true}	34	1
	H_{true}	1	1 872

4.2.2 面向真实场景进行预测

为了进一步验证预测的准确性,本文系统在中兴通讯的真实数据中心进行了安装部署,对硬盘未来的故障进行预测,但关闭数据修复,以观察是否真正

故障.对于预测结果的判断,需等待 1 个月后得出结论.系统运行 3 个月,共出现故障盘 521 块.在模型训练时使用 100%已采集的硬盘数据.测试结果如表 6 所示,召回率 93.86%;误检率 0.33%,预测准确性得到了真实场景的验证.

Table 5 Comparison of Prediction Results Based on Seagate SATA

表 5 基于希捷 SATA 数据的预测结果比较

System	Type	$F_{Predicted}$	$H_{Predicted}$
Our work	F_{true}	32	2
	H_{true}	6	2 627
Comparison System 1[15]	F_{true}	30	4
	H_{true}	52	2 581
Comparison System 2[17]	F_{true}	29	5
	H_{true}	22	2 611

Table 6 Prediction Results Based on Real Scene

表 6 基于真实场景的预测结果

Type	$F_{Predicted}$	$H_{Predicted}$
F_{true}	489	32
H_{true}	392	127 500

4.2.3 在 Backblaze 数据集评测

为了充分验证本文工作的普遍适用性,本文选取了 Backblaze 网站免费公开的 2017 年的全年故障盘的数据集进行测试,其中希捷品牌的 SATA 硬盘共 65 003 块,故障盘 1 431 块,将其 70%用于建模、30%用于测试.然而遗憾的是,该数据集的数据质量较低,主要存在的问题包括:1)硬盘信息采集频率较低,每天 1 次,且个别数据丢失;2)采集硬盘信息单一化,是仅采集硬盘 SMART 信息,未采集 IO 信息,这对于硬盘剩余寿命的衡量是不利的;3)硬盘类型单一化,仅涉及 SATA 硬盘,不涉及 SAS 硬盘和固态硬盘.4)硬盘品牌单一化,该数据集中希捷品牌的硬盘占据绝大多数,其他品牌凤毛麟角.正因上述原因,测试时无法将 4.2.1 节中的模型直接与该数据集对接,在经过一定的技术处理,例如不考虑 IO、减少样本合并数量、剔除故障盘数量在 20 块以下的硬盘型号后,最终建立预测模型.

其结果如表 7 所示,本文工作、对比系统 1、对比系统 2 的召回率分别为 80.43%,75.19%,82.60%,误检率分别为 3.45%,3.84%,2.57%.可以看到,在数据质量较低的情况下,本文工作与 2 个对比系统的预测准确性均不理想,充分说明了数据质量对预测结果的重要性.

Table 7 Test Results from the Backblaze Dataset

表 7 在 Backblaze 数据集的测试结果

System	Type	$F_{Predicted}$	$H_{Predicted}$
Our work	F_{true}	1 151	280
	H_{true}	1 965	61 607
Comparison System 1[15]	F_{true}	1 076	355
	H_{true}	2 144	61 428
Comparison System 2[17]	F_{true}	1 182	249
	H_{true}	1 421	62 151

4.3 测试数据修复的速度

在本节测试时,所有涉及的硬盘均为希捷 ST8000DM002 型号,纠删码类型为 6+3 型.因硬盘在濒临故障期间的读写速度不稳定,为尽量排除干扰,每项测试都是 10 次,去掉 2 个最高值和 2 个最低值,剩余值取平均值.为了充分体现本文工作的优势,将濒临故障盘独立且全速修复的传统方法作为对比系统 3,将文献[20]的方法作为对比系统 4,如图 8 所示,在不同修复数据量的情况下,本文的工作均大幅减少了修复时间,相较于对比系统 3,4 的方法分别平均减少 55.10%和 84.56%的修复时间.

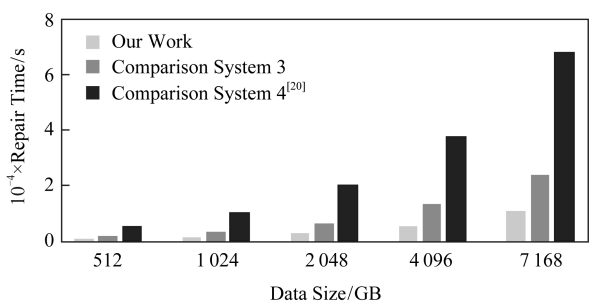

Fig. 8 Comparison of repair time under different data quantities

图 8 在不同数据量的情况下的修复时间对比

4.4 实验小结

本节从对前台业务的干扰程度、对故障预测的准确性以及故障后的修复速度这 3 个维度对本文的工作进行了测试,其中采集硬盘信息对前台业务影响仅为 0.96%,对硬盘故障预测的召回率达到 94.66%,数据修复较传统方法减少 55.10%的时间.该系统已在中兴通讯的数据中心正常运行 9 个月以上,期间超过 90%的硬盘故障均通过主动容错得以解决,系统降级服务的次数显著降低,且未出现数据丢失.

5 总 结

主动容错技术的核心目标包括：1)高可靠.故障预测准确率高，避免存储系统降级服务.2)高智能.解放运维人员劳动力，自动处理.3)低干扰.在采集、分析、迁移等阶段都尽可能降低对前端业务干扰，不影响数据中心的正常服务.4)低成本.避免浪费硬盘寿命，充分利用硬盘剩余寿命.5)广适用.能够适应真实的大规模数据中心业务场景，对不同品牌、型号、类型的硬盘都能支持，且安装部署灵活方便.

然而遗憾的是，现有的主动容错技术仅仅局限于对硬盘故障的高准确预测，但对采集、迁移、反馈等方面研究较少，这使得主动容错技术在难以真正商用.

本文首次提出了“采集—预测—迁移—反馈”的全流程主动容错技术方案，包括：1)提出了分时硬盘信息采集方法，解决了大规模 SMART 采集引发的灾难问题；2)提出了滑动窗口记录合并及样本构建方法，填补了针对 SAS 硬盘和固态硬盘故障预测的空白；3)提出了多类型硬盘故障预测方法，攻克了正负样本严重不均的难题；4)提出了多盘联合数据迁移方法，加快了数据的修复速度；5)提出了预测结果二级验证及快速反馈方法，灵活并快速地处理了误判、漏判、迟判的情形.测试表明，采集硬盘信息对前台业务影响仅为 0.96%，对硬盘故障预测的召回率达到 94.66%，数据修复较传统方法减少 55.10%的时间.本文的工作已在中兴通讯的大规模数据中心稳定商用，满足了主动容错技术在高可靠、高智能、低干扰、低成本、广适用等方面的核心目标.

参 考 文 献

[1] Reinsel D, Gantz J, Rydning J. An IDC White Paper: Data Age 2025 [OL]. [2019-07-12]. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] Schroeder B, Gibson G. Disk failures in the real world: What does an MTTF of 1 000 000 hours mean to you? [C] //Proc of the 5th USENIX Conf on File and Storage Technologies. Berkeley, CA: USENIX Association, 2007: 1-16

[3] Qin Xin, Miller E, Schwarz T, et al. Reliability mechanisms for very large storage systems [C] //Proc of the 20th IEEE Conf on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2003: 1-11

[4] Sankar S, Shaw M, Vaid K. Impact of temperature on hard disk drive reliability in large datacenters [C] //Proc of the 41st Int Conf on Dependable Systems & Networks. Piscataway, NJ: IEEE, 2011: 530-535

[5] Wang Yijie, Sun Weidong, Zhou Song, et al. Key technologies of distributed storage for cloud computing [J]. Journal of Software, 2012, 23(4): 962-986 (in Chinese)
(王意洁，孙伟东，周松，等. 云计算环境下的分布存储关键技术[J]. 软件学报，2012，23(4)：962-986)

[6] Weil S, Brandt S, Miller L, et al. Ceph: A scalable, high-performance distributed file system [C] //Proc of the 7th USENIX Symp on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2006: 307-320

[7] Shvachko K, Kuang H, Radia S, et al. The Hadoop distributed file system [C] //Proc of the 26th IEEE Conf on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2010: 1-10

[8] Fu Yingxun, Wen Shilin, Ma Li, et al. Survey on single disk failure recovery methods for erasure coded storage systems [J]. Journal of Computer Research and Development, 2018, 55(1): 1-13 (in Chinese)
(傅颖勋，文士林，马礼，等. 纠删码存储系统单磁盘错误重构优化方法综述[J]. 计算机研究与发展，2018，55(1)：1-13)

[9] Calder B, Wang Ju, Ogus A, et al. Windows Azure storage: A highly available cloud storage service with strong consistency [C] //Proc of the 23rd ACM Symp on Operating Systems Principles. New York: ACM, 2011: 143-157

[10] Yu Xiao, Tan Yuan, Zhang Changyou, et al. A High-performance hierarchical snapshot scheme for hybrid storage Systems [J]. Chinese Journal of Electronics, 2018, 27(1): 76-85

[11] Xiao Weijun, Yang Qing, Ren Jin, et al. Design and analysis of block-level snapshots for data protection and recovery [J]. IEEE Transactions on Computers. 2009, 58(12): 1615-1625

[12] Liu Dapeng, Zhao Youjian, Xu Haowen, et al. Opprentice: Towards practical and automatic anomaly detection through machine learning [C] //Proc of the 15th ACM Internet Measurement Conf. New York: ACM, 2015: 211-224

[13] Li Jing, Wang Gang, Liu Xiaoguang, et al. Review of reliability prediction for storage system [J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(3): 341-354 (in Chinese)
(李静，王刚，刘晓光，等. 存储系统可靠性预测综述[J]. 计算机科学与探索，2017，11(3)：341-354)

[14] Pitakrat T, Hoorn A, Grunske L, et al. A comparison of machine learning algorithms for proactive hard disk drive failure detection [C] //Proc of the 4th Int ACM SIGSOFT Symp on Architecting Critical Systems. New York: ACM, 2013: 1-10

[15] Zhu Bingpeng, Wang Gang, Liu Xiaoguang, et al. Proactive drive failure prediction for large scale storage systems [C/OL] //Proc of the 29th IEEE Conf on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2013 [2019-08-15]. https://www.storageconference.us/2013/Papers/2013.Paper.08.pdf

[16] Li Jing, Ji Xinpu, Jia Yuhan, et al. Hard drive failure prediction using classification and regression trees [C] //Proc of the 44th Annual IEEE/IFIP Int Conf on Dependable Systems and Networks. Piscataway, NJ: IEEE, 2014: 383-394

[17] Liu Yongkang. Predicting disk failures for large-scale datacenter by machine-learning method [D]. Wuhan: Huazhong University of Science & Technology, 2015 (in Chinese)
(柳永康. 基于机器学习的硬盘故障预测系统[D]. 武汉: 华中科技大学, 2015)

[18] Xiao Jiang, Xiong Zhuang, Wu Song, et al. Disk failure prediction in data centers via online learning [C/OL] //Proc of the 47th Int Conf on Parallel Processing. 2018[2018-12-20]. https://www.researchgate.net/publication/326910575_Disk_Failure_Prediction_in_Data_Centers_via_Online_Learning

[19] Xie Yanwen, Feng Dan, Wang Fang, et al. OME: An optimized modeling engine for disk failure prediction in heterogeneous data center [C] //Proc of the 36th IEEE Int Conf on Computer Design. Piscataway, NJ: IEEE, 2018: 561-564

[20] Ji Xinpu, Ma Yuxiang, Ma Rui, et al. A proactive fault tolerance scheme for large scale storage systems [C] //Proc of the 15th Int Conf on Algorithms and Architectures for Parallel Processing. Berlin: Springer, 2015: 337-350

[21] Qin An, Hu Dianming, Liu Jun, et al. Fatman: Cost-saving and reliable archival storage based on volunteer resources [C] //Proc of the 40th Int Conf on Very Large Data Bases. New York: ACM, 2014: 1748-1753

[22] Meza J, Wu Qiang, Kumar S, et al. A large-scale study of flash memory failures in the field [C] //Proc of the 20th ACM Int Conf on Measurement and Modeling of Computer Systems. New York: ACM, 2015: 177-190

[23] Bianca S, Raghav L, Arif M. Flash reliability in production: The expected and the unexpected [C] //Proc of the 14th USENIX Conf on File and Storage Technologies. Berkeley, CA: USENIX Association, 2016: 67-80

[24] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321-357

Yang Hongzhang, born in 1989. PhD candidate. Member of CCF. His main research interests include file system, fault tolerance, and distributed system.

Yang Yahui, born in 1966. PhD, professor, PhD supervisor. Member of CCF. Her main research interests include network security management and network performance optimization.

Tu Yaofeng, born in 1972. Master, professor. Member of CCF. His main research interests include big data and cloud computing.

Sun Guangyu, born in 1981. PhD, associate professor, PhD supervisor. Member of CCF. His main research interests include energy efficient memory architecture and system.

Wu Zhonghai, born in 1968. PhD, professor. Member of CCF. His main research interests include big data and cloud security.