Selective scrubbing based on algorithmic randomness

Rahul Vishwakarma
California State University Long Beach
California, United States
rahuldeo.vishwakarma01@student.csulb.edu

Bing Liu
Dell Technologies
Beijing, China
bing.liu@dell.com

Peter Gatsby
California State University Long Beach
California, United States
peter.gatsby01@student.csulb.edu

Jinha Hwang
California State University Long Beach
California, United States
jinha.hwang01@student.csulb.edu

ABSTRACT

Disk scrubbing is a background process to fix read errors by reading the disks. However, scrubbing the entire storage array can significantly increase the system load and degrade system performance when there is high incoming IO. Deciding "which disk to scrub" complemented with "when to scrub" can significantly improve the data centre's overall reliability and power saving. We present a solution on an open-source SMART dataset that performs selective scrubbing and designs a scrub frequency based on the scrub cycle. The method leverages an algorithmic randomness framework to quantify the health of the concerned drives and ranks them for selective scrubbing.

CCS CONCEPTS

• **General and reference** → Reliability; • **Hardware** → Failure recovery, maintenance and self-repair;

KEYWORDS

selective scrubbing, conformal prediction

1 SOLUTION

Unlike more conventional studies that focused on the use of machine learning for disk scrubbing schemes, we implement conformal prediction to first quantify the degree of health and then categorize the disks as concern and no-concern drives. The concerned drives are further ranked for selective scrubbing and mapped with the scrub frequency.

Here, we focus on identifying the disks in the system which are of 'concern' or may become 'concern' drives soon; and only select those disks for scrubbing [1] as it will reduce the number of disks meant for scrubbing. We create a set of the above categories and quantify "how much" concern/no-concern is it across the entire storage pool based on prediction's confidence and use this metric to prioritize selective scrubbing of the drives.

2 RESULTS

We use an open-source SMART dataset [2] from Baidu with 23395 disks with twelvel features, and implement conformal prediction [3]. We then identify the concern disk, select a set of disks for scrubbing, and assign scrubbing frequency based on system load. The system load identification is based on multivariate fuzzy time series. The proposed method, when implemented in a production environment, can help resource and power savings [4] by spinning disks down. We achieved this by prioritizing the scrubbing of disks likely to have failures over the disks with a low likelihood of failure.

REFERENCES

[1] Rahul Deo Vishwakarma. Bing Liu. 2022. Managing storage device scrubbing. https://patentimages.storage.googleapis.com/d4/da/3f/e9c833ae55ad04/US20220019564A1.pdf. [Online; accessed 11-March-2022].

[2] Open Source SMART dataset. 2013. Smart Dataset from Nankai University and Baidu, Inc. https://pan.baidu.com/share/link?shareid=189977&uk=4278294944/. [Online; accessed 04-March-2022].

[3] Peter Gatsby Jinha Hwang. 2022. Source Code for Selective scrubbing based on algorithmic randomness. https://github.com/jinzzup/systor2022-disk-scrubbing/. [Online; accessed 11-March-2022].

[4] Guanying Wang, Ali Raza Butt, Chris Gniady, and Virginia Tech. 2008. On the Impact of Disk Scrubbing on Energy Savings.. In *HotPower*.

SYSTOR'22, June 2022, Haifa, Israel

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4