ENERGY-ASSISTED RECORDING: SYSTEM ARCHITECTURE & APPLICATIONS

Bernd LAMBERTS, Classica JAIN and Remmelt PIT
Western Digital, San Jose, California, USA

I. Introduction

The introduction of energy-assisted recording technologies changes the way rotating magnetic storage systems have been perceived in a fundamental way. Traditionally recording devices have been treated as bimodal regarding reliability and durability. The designs consider only two states, the Head & Media components perform as measured during manufacturing or are defective. Energy-assisted recording introduces the need to manage components over the used lifetime [1].

For energy-assisted recording, the achievable performance is a function of the parameters which limit component lifetime. This introduces the need to understand workload distributions of modern storage architectures such as distributed file systems and hyper-scale applications. Another key aspect of the device use is related to the transition from direct static mapped disk layouts like PMR recording, to systems using indirection mapped layouts such as SMR recording. In the second part of the paper we discuss solutions to manage component lifetime while maximizing the recording system performance.

II. SYSTEM CHALLENGES

For systems with a finite lifetime of the head disk interface, a detailed understanding of the access statistics is crucial. The total volume of data and the temporal distribution as well as the spatial distribution are all important. We used device logs and detailed system traces to analyze the workload distributions. Our results show dramatic differences in data consumption as function of the application [see charts 1 & 2]. We find that hyper-scale workloads show significantly higher utilization than traditional enterprise applications [chart 2]. This result is consistent with other studies on access density requirements [2]. We investigated the time dependency of the workload via categorizing device logs into age classes. Our results indicate that almost no systems experience increased write workload over time. Typical use cases show an increased workload in the 1st 12 months followed by stabilization and a decline towards the end of the warranty life. This is consistent with the observation of data “aging” over time.

Details of the access statistics were investigated via analysis of device logs and trace analysis of setups in our System Integration Group. One of the key findings is that file system choices and device driver selection have similar impact to the overall results as the choice of applications. We find that poor choices in the Host software stack may lead to excessive write amplification for individual components. In some cases, 30% of all I/O’s over 24 hours are associated with less than 1% of the LBA’s [chart 3].

The reasons for this behavior are related to check-pointing by the application layer and/or the use of special meta-data index systems to facilitate synchronization across processes. A contributing effect is the fact that modern disk drives can store typical ‘iNodes’ of file systems such as Ext4 [3] on a few dozen tracks. Depending on the device format layout, this may lead to excessive use of individual heads.

III. DEVICE SOLUTIONS

There are three areas of opportunity where some of the impact of finite lifetime can be mitigated:

1) **Component lifetime management:** We present an integrated component lifetime management approach that aims to maximize the utilization of the given hardware. As shown in the previous paragraphs, only a few devices experience the maximal workload. Since the lifetime of energy-assisted recording devices is inversely proportional to their energy output, we use a reinforcement learning [4] approach to optimize component lifetime.

2) **Workload distribution opportunities:** As outlined in chapter II, challenging situations may occur for systems which feature excessive use of single components or localized areas in the device. This situation is challenging for statically mapped designs like traditional PMR recording arrangements. One solution

Bernd Lamberts
E-mail: Bernd.Lamberts@wdc.com
tel: +1-408-717-8761

of this problem is the use of an intermediate de-staging buffer or cache. This technique is well known from write cache designs in computer science.

3) **Workload management opportunities:** Even considering the options discussed in 1) & 2) ultimately true wear leveling similar to solid state devices [5] may be required. A major challenge to this objective is the fact that the true component lifetime is known. Hence most approaches are limited to workload balancing. Even for this objective we find a significant difference between systems based on full indirection like HA-SMR[6] and statically mapped systems such as PMR. For SMR systems it will be significantly easier to insure balanced use of all components.

REFERENCES

1) Budaev, Bair V., and David B. Bogy. "On the lifetime of plasmonic transducers in heat assisted magnetic recording." *Journal of Applied Physics* 112.3 (2012): 034512.
2) Cisco, “Cisco global Cloud Index 2015-2020”, http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf
3) Aneesh Kumar et al, “Ext4 block and inode allocator improvements”, Proceedings of the Linux Symposium, (2008)
4) Sutton, Richard S.; Barto, Andrew G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.
5) Yang, Ming-Chang, et al. "Garbage collection and wear leveling for flash memory: Past and future." *Smart Computing (SMARTCOMP), 2014 International Conference on*. IEEE, 2014.
6) Feldman, Tim, and Garth Gibson. "Shingled magnetic recording areal density increase requires new data management." *USENIX; login: Magazine* 38.3 (2013).

IV. ILLUSTRATIONS

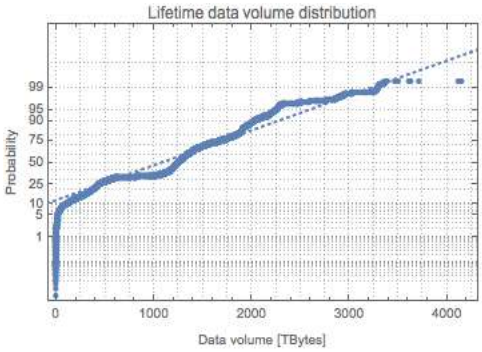

Fig. 1 Observed population data Volume distribution

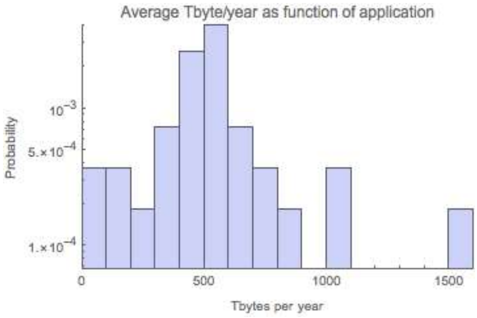

Fig 2. Data volume by application

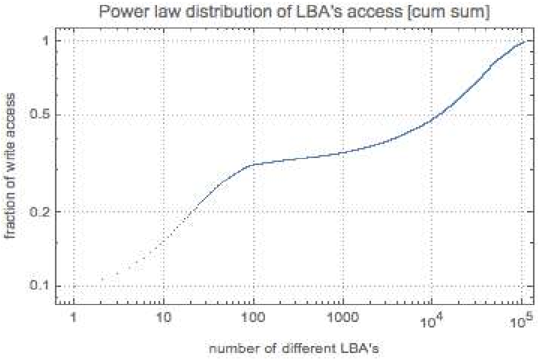

Fig. 3 File system 24h access distribution by LBA