A Case for Redundant Arrays of Hybrid Disks (RAHD)

Frank Wang[1], Na Helian[2], Sining Wu[1], Yike Guo[3], Derek Deng[1], Vineet Khare[1], C. Liao[1], M. Rashidi[1], and Andy Parker[4]

[1]Centre for Grid Computing, Cambridge-Cranfield HPCF, Cranfield, MK43 0AL, U.K.
[2]Department of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, U.K.
[3]Department of Computing, Imperial College London, London SW7 2AZ, U.K.
[4]Cavendish Laboratory, Cambridge University, Cambridge CB3 0HE, U.K.

Hybrid Hard Disk Drive was originally concepted by Samsung, which incorporates a Flash memory in a magnetic disk. The combined ultra-high-density benefits of magnetic storage and the low-power and fast read access of NAND technology inspires us to construct Redundant Arrays of Hybrid Disks (RAHD) to offer a possible alternative to today's Redundant Arrays of Independent Disks (RAIDs) and/or Massive Arrays of Idle Disks (MAIDs). We first design an internal management system (including Energy-Efficient Control) for hybrid disks. Three traces collected from real systems as well as a synthetic trace are then used to evaluate the RAHD arrays. The trace-driven experimental results show: in the high speed mode, a RAHD outplays the purely-magnetic-disk-based RAIDs by a factor of 2.4–4; in the energy-efficient mode, a RAHD4/5 can save up to 89% of energy at little performance degradation.

Index Terms—**Flash memory, hybrid disk, Massive Arrays of Idle Disks (MAID), Redundant Arrays of Independent Disks (RAID).**

I. Introduction

HYBRID hard disk drive, proposed by Samsung in 2005, is the first fully functional disk drive to combine NAND-based Flash with rotating storage media. The ultra-high-density benefits of magnetic storage technology are preserved, while the ultra-low-power, high-reliability and fast read/write access of advanced NAND technology enhances the overall value of the hybrid drive at little additional cost [1]. It is reported that Samsung has teamed up with Microsoft to make this hybrid drive work with the next version of the Windows operating system [1].

Norman Ken Ouchi at IBM was awarded a 1978 U.S. patent 4 092 732 [2] titled "System for recovering data stored in failed memory unit." The claims for this patent describe what would later be termed RAID5 with full stripe writes. This 1978 patent also introduced disk mirroring or duplexing (what would later be termed RAID1) and protection with dedicated parity (that would later be termed RAID4).

The term RAID was first defined by David A. Patterson, Garth A. Gibson and Randy Katz at the University of California, Berkeley, in 1987 [3]. They studied the possibility of using two or more drives to appear as a single device to the host system and published a paper: "A Case for Redundant Arrays of Inexpensive Disks (RAID)" in June 1988 at the SIGMOD conference [3]. This specification suggested a number of prototype "RAID levels", or combinations of drives. Each had theoretical advantages and disadvantages. Over the years, different implementations of the RAID concept have appeared.

A massive array of idle disks (MAID) is a system using hundreds to thousands of hard drives for near-line data storage [4]. MAID is designed for Write Once, Read Occasionally (WORO) applications. In a MAID each drive is only spun up on demand as needed to access the data stored on that drive. Large scale disk storage systems based on MAID architecture allow dense packaging of drives and are designed to have only 25% of disks spinning at any one time. This allows for high throughput to get data from this platform quickly. Since persistent data is accessed very little, any data can be accessed at any time, and stay within the power budget of 25% of drives spinning [4].

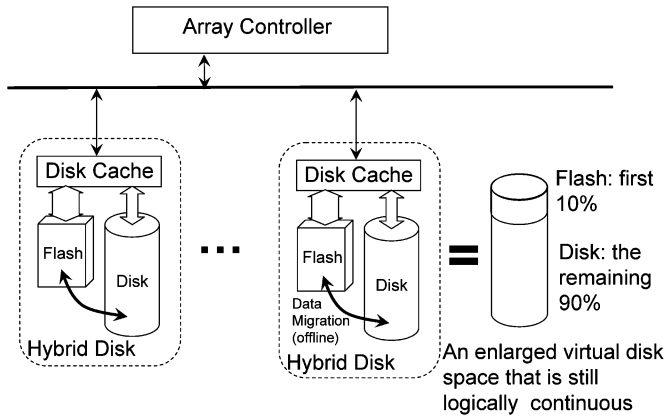

Fig. 1. Redundant Arrays of Hybrid Disks (RAHD). Based on the observation that the majority of the disk I/Os are probably going to less than 10% of the total disk space, a flash memory occupies the first 10% of an enlarged disk space that is virtually continuous. Accordingly, the most frequently-accessed data blocks are migrated into the flash memory periodically.

In this work, we propose to construct Redundant Arrays of Hybrid Disks (RAHD), as shown in Fig. 1. RAHD is intended to imitate RAID Level 0 (Striped disks lacking any particular fault tolerance), Level 1 (Simple disk mirroring), Level 2 (Mirroring augmented with Hamming Error Correction Code), Level 3 (Byte striping with dedicated parity), Level 4 (Block striping with dedicated parity) and Level 5 (Block striping with distributed parity). It is believed that a RAHD will provide improvements in performance, power consumption, noise level and scalability, resulting from the combined properties of hybrid disks in ultra-high-density and the low-power consumption and fast read access.

II. Related Work and Our Innovation

The Hewlett-Packard scientists proposed to incorporate MEMS-based storage into disk arrays [5]. Micro-electro-

Digital Object Identifier 10.1109/TMAG.2008.2002531

mechanical Systems (MEMS) is a disruptive new storage technology. The typical access times for MEMS are in the order of 1–2 ms, which is ideal to bridge the gap between NVRAM and disk drives. They examined several possible placements for the MEMS storage in the disk array by: (1) replacing all the disks with MEMS storage; 2) replacing the NVRAM cache with MEMS storage; and 3) replacing half the disks with MEMS storage. Unfortunately, MEMS-based storage chips are still not commercially available.

There are efforts in reducing power consumption in RAIDs. In [6], an energy-efficient RAID system architecture called EERAID is designed by taking advantage of RAID redundant information. Trace-driven simulation experiments showed that 1) For single-speed disks, EERAID 1 can achieve up to 30% energy savings by EERAID 5, and 2) For multi-speed disks, compared with DRPM, EERAID 1 can achieve 22% extra energy savings and 11% more for EERAID 5. In all experiments, there is either better performance gain or little performance degradation. In [7], a family of energy-efficient disk layouts has been found in RAID1. The scheme called DiskGroup distributes the workload between the primary disks and secondary disks based on the characteristics of the workload.

There is not a reported work to date in the literature on building hybrid disk arrays due to its recent introduction. On the other hand, even the internal management for hybrid disks remains unclear since the hybrid disk proposers have never published the details. In this work, we first design an internal management system (including Energy-Efficient Control) for hybrid disks. We then establish an analytical model for the RAHD arrays. The trace-driven experiments will be carried out to evaluate the performance.

III. Design and Analysis

The combined ultra-high-density benefits of magnetic storage and the low-power and fast read access of NAND technology inspires us to construct Redundant Arrays of Hybrid Disks (RAHD, Fig. 1) to offer a possible alternative to today's RAID [2], [3] and/or MAID [4]. We have designed two separable working modes for RAHD arrays: (1) High-Speed (HS) mode targeting at displaying the full potential speed of hybrid disks; and (2) Energy-Efficient (EE) mode targeting at reducing the energy consumption while maintaining acceptable performance.

High Speed (HS) Mode and an Internal Management Strategy of Hybrid Disks: In the HS mode, the Flash memory always contains frequent data objects. We have designed two schemes to move the frequent data blocks into the Flash memory: (1) Using Decision Tree from Data Mining to predict the frequencies of data blocks (file attributes of a newly-generated file are used to predict its frequency). Blocks with high (predicted) frequency are then written onto the Flash memory [8]; (2) Tracking the frequencies of data blocks (when the disk is being used) and then migrating the most frequently-accessed data blocks into the Flash memory during the system idle periods [9], as illustrated in Fig. 2. Because of the high skew in common application loads, most of the requests are likely to be satisfied in the Flash memory without bothering the slower disks.

Three real-system traces (Cello-96, TPC-D, and Cello-99) [10] are used to investigate the data access patters. These traces

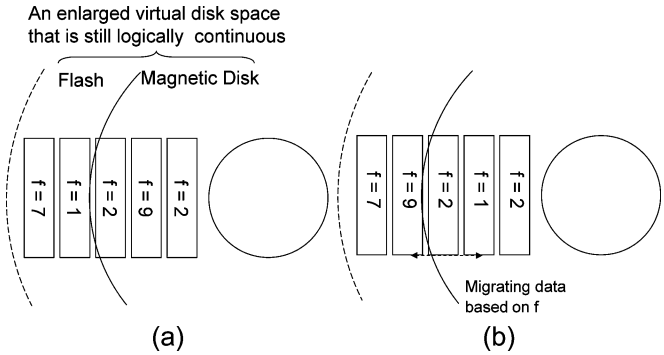

Fig. 2. In a hybrid disk for RAHD, the Flash memory (the first 10% of the enlarged virtual disk space) absorbs the frequent data whereas the magnetic disk contains those relatively infrequent data. Data are migrated in objects periodically according to their frequencies (f). An object may be placed anywhere in Flash or the disk, which means we only perform a partial ordering of the objects rather than a full ordering. (a) Before data migration. (b) after data migration.

capture all low-level disk I/O performed by the system at Hewlett-Packard Laboratories. Only TCP-D is read-intensive (read occupies 98%) whereas Cello-96 (66%) and Cello-99 (46%) are not. Our simulator combines the modified DiskSim simulator [11] with an integrated Object Mover to simulate an array of hybrid disks. We use the above modified Disksim to configure an 8-hybrid-disk RAHD. Reads and writes in Flash are processed in terms of pages. The page size for the Samsung Flash memory used in the work is (2K + 64) Bytes. Note that the page size is intended to fit a disk sector in size, hence a disk (Seagate Cheetah Ultra SCSI) in a RAHD is formatted to have a sector size of 2 KB. Based on the observation that the majority of the disk I/O's (including both writes and reads) are probably going to less than 10% of the total disk space [9], a Flash memory is conservatively designed to occupy the first 10% of an enlarged disk space that is virtually continuous (inset of Fig. 1).

Table I summarizes all level RAHD characteristics in the High Speed mode. It is expected that RAHD works with supercomputer applications that need a huge data rate or transaction-processing applications that need a high I/O rate.

Energy Efficient (EE) Mode: In the EE mode, some "active drives" remain constantly spinning whereas the remaining "passive drives" are allowed to spin-down following a period of inactivity. A request is directed to the Flash in the first instance. If the request is fulfilled by hitting the Flash memory, there is no need to awaken the sleeping disk. Otherwise, the request will be forwarded to the disk that is already active or needs to be awakened from sleep.

Note that RAHD3 uses disk stripping to achieve a high data transfer rate and a separate data protection disk drive to store error checking data. As shown in Fig. 3, an individual write to a single sector still involves all the disks in a group because (1) the check disk must be rewritten with the new parity data, and (2) the rest of the data disks must be read to be able to calculate the new parity data. Recall that each parity bit is just a simple exclusive OR of all the corresponding data bits in a group. Therefore the EEC for small write/large write in RAID3 is zero.

Table II summarizes all level RAID/RAHD characteristics. Note that all small reads help save energy and all large writes do not contribute to power-saving. It is expected that RAHD

TABLE I
HIGH SPEED (HS) MODE FOR ALL LEVEL RAHDS. G = number of data disks IN A GROUP. X IS RAID LEVEL

Access time (ms) & Speedup	Small transfer		Large transfer	
	A request of 2KB to a single disk, a RAID of 8 magnetic disks, or a RAHD of 8 hybrid disks.		A request of 16KBx8=128KB to a single disk, a RAID of 8 magnetic disks, or a RAHD of 8 hybrid disks.	
	Small Write (SW)	Small Read (SR)	Large Write (LW)	Large Read (LR)
A single disk	4.7 ms	4.7 ms	7.85 ms	7.85 ms
RAID (G=8)	4.7 ms	4.7 ms	5.05 ms	5.05 ms
RAHD (G=8)	0.67 ms	0.96 ms	2.11 ms	1.17 ms
$S_1 = \frac{T_{SimgleDisk}}{T_{RAID}}$	1.0	1.0	1.6	1.6
$S_2 = \frac{T_{SimgleDisk}}{T_{RAHD}}$	7.0	4.9	3.7	6.7
$S_3 = \frac{T_{RAID\_x}}{T_{RAHD\_x}}$	7.0	4.9	2.4	4.3

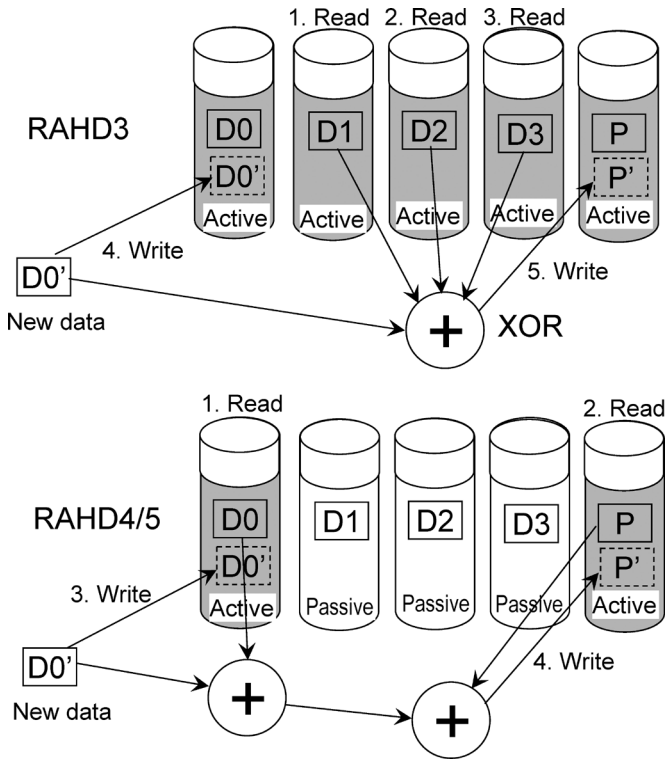

Fig. 3. Small write update on RAHD3 versus RAHD4/5. the optimization for small writes reduces the number of disk accesses as well as the number of disks occupied, which is good to save power. A small write in RAHD4/5 involves 1 Data Disk & 1 Check Disk (if the write is missed in Flash in the worst case) and the remaining disks are kept idle to save energy.

works energy-efficiently in read-intensive applications that have little need for write operation performance, such as historic or archival databases.

IV. REAL-SYSTEM-TRACE-DRIVEN RESULTS

The real-system-trace-driven results show: in the high speed mode, a RAHD outplays the purely-magnetic-disk-based RAIDs by a factor of 2.4–4 (Table III); in the energy-efficient mode, a RAHD4/5 can save up to 89% of energy at little performance degradation (Fig. 4). The highest performance comes from either Level 4 or Level 5. Considering that RAHD4 has an inherent bottleneck on the parity drive, Level 5 looks best in terms of the trade between the performance and the power-saving.

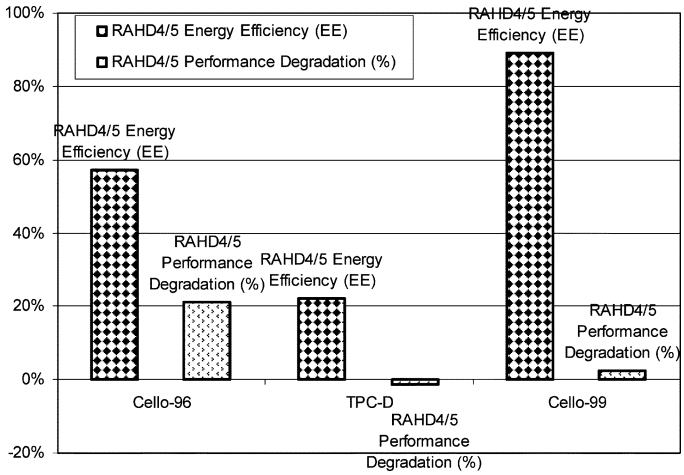

Fig. 4. In the energy-efficient (EE) mode, a RAHD4/5 is found to have 89% of energy saving at 2.6% of performance degradation under Cello-99, 22% of energy saving and even 1.3% of performance improvement under TPC-D.

V. CONCLUSION

Hybrid disks are originally designed for personal and mobile applications whereas the proposed RAHD can be used more broadly [12]. It is found that a RAHD will provide improvements in performance, power consumption, and scalability. It is a conceptually simple technique for dramatically improving disk array performance, which is desirable for supercomputers, transaction processing and data streaming (e.g., media, video, audio, etc.).

Our technique (including an intelligent Energy-Efficient Control) can be implemented in an embedded device driver in a hybrid disk controller, which means each hybrid disk independently manages its own space without considering the overall array management. Our internal management strategy is to move data to their optimal locations during idle periods, so as not to conflict with system use.

Although Flash memory costs far less than byte-programmable EEPROM [13], the cost of RAHDs may slightly increase with the addition of a flash memory compared with RAIDs/MAIDs. This increase will be mitigated by fourfold deduction of access/boot time in the High-Speed Mode or up to 89 percent power savings in the Energy-Efficient Mode. All of these changes are crucial to the ever demanding needs of today's high performance computing applications.

When combining technologies of flash and rotating media, there are both "life expectancies" and "wear" issues. Flash drives have limited endurance (typically, 300 000–1 000 000 write cycles). However, all Flash drives employ a technique known as wear leveling, where writes are smoothly distributed over all blocks [13], [14].

In principle, the proposed RAHD architecture and its management principles are extensible by replacing Flash memory with other nonvolatile memories, such as battery-powered SRAMs, MRAMs, MEMS [15] and Phase-Change Memories (PCM or PRAM). In particular, Toggle MRAM (Magneto-resistive RAM) is a hopeful candidate in terms of effectively

TABLE II
EEC FOR ALL LEVEL RAHDS IN THE EE MODE. G = number of data disks IN A GROUP. D = total number of disks WITH DATA (NOT INCLUDING EXTRA CHECK DISKS); C = number of check disks IN A GROUP; $n_G = D/G$ = number of groups. HIT RATIO H = 0.8 FOR HYBRID DISKS. NOTE THAT THE RAID MAY SPIN DOWN ITS MAGNETIC DISKS TO SAVE ENERGY

Disk I/Os	Energy-Efficient Coefficient (EEC)				
	RAID0	RAID1	RAID3	RAID4	RAID5
	RAHD0	RAHD1	RAHD3	RAHD4	RAHD5
Small Write (SW)	(G-1)/G	(G-1)/G	0	(G-1)/(G+1)	(G-1)/(G+1)
	(G-1+H)/G	(G-1+H)/G	H	(G-1+2H)/(G+1)	(G-1+2H)/(G+1)
Small Read (SR)	(G-1)/G	(2G-1)/2G	G/(G+1)	G/(G+1)	G/(G+1)
	(G-1+H)/G	(2G-1+H)/(2G)	(G+H)/(G+1)	(G+H)/(G+1)	(G+H)/(G+1)
Large Write (LW)	0	0	0	0	0
	H	H	H	H	H
Large Read (LR)	0	0.5	1/(G+1)	1/(G+1)	1/(G+1)
	H	(1+H)/2	(1+G·H)/(G+1)	(1+G·H)/(G+1)	(1+G·H)/(G+1)
Capacity Utilization (CU)	1	0.5	G/(G+1)	G/(G+1)	G/(G+1)

TABLE III
COMPARISON OF RAID, RAIDFASTBAND [9], AND THE RAHD. G (= 8) IS THE NUMBER OF DATA DISKS IN THE GROUP. THE AVERAGE PER-REQUEST RESPONSE TIME FOR A SINGLE DISK IS 23.0 ms UNDER THE SAME WORKLOAD

	RAID based on purely magnetic disks	RAIDfastband with data migration	RAHD based on hybrid disks
	Average per-request response time	Average per-request response time	Average per-request response time
Level 0	10.1 ms	4.3 ms	2.6 ms
Level 1	14.9 ms	6.3 ms	3.5 ms
Level 4	20.5 ms	8.7 ms	5.2 ms
Level 5	20.3 ms	8.4 ms	5.3 ms
Level 10	12.3 ms	5.2 ms	3.3 ms

eliminating the single-line disturb phenomenon present in previous approaches to MRAM switching [16], [17].

ACKNOWLEDGMENT

This work is supported by the U.K. government and European Commission (EC) through an EPSRC/DTI grant (Pound sterling 1 million) "Grid-oriented Storage (GOS)", an EPSRC grant (£470 k) "GOS2", an EC grant (euro 1 million) "QuickLinux" and an EC grant (euro400 k) "EuroAsiaGrid". The authors also thank Prof. G. Gibson of Carnegie Mellon University, Prof. Kai Li of Princeton University, Prof. R. Parker of Rolls Royce, Dr. R. Wright of BBC, Dr. F. Donno of CERN, Dr. S. Vandebroek of Xerox, Dr. P. Francis of EADS, and Prof. L. Ni of HKUST for viewing their demonstrations and providing them with comments that much improved the work.

REFERENCES

[1] Samsung teams with Microsoft to develop first hybrid hard drive with NAND flash memory (2007) [Online]. Available: www.physorg.com/news3862.html
[2] N. K. Ouchi, "System for recovering data stored in failed memory unit," U.S. Patent 4 092 732, May 30, 1978.
[3] D. Patterson, G. A. Gibson, and R. Katz, "A case for redundant arrays of inexpensive disks (Raid)," in *SIGMOD*, 1988.
[4] D. Colarelli and D. Grunwald, "Massive arrays of idle disks for storage archives," *2002 IEEE*.
[5] M. Uysal, A. Merchant, and G. Alvarez, "Using MEMS-based storage in disk arrays," in *Proc. e 2nd Usenix Conf.on File and Storage Technologies*, San Francisco, CA, Mar. 2003.
[6] D. Li, P. Gu, H. Cai, and J. Wang, "Eeraid: Energy efficient redundant and inexpensive disk array," in *Proc. 11th Workshop on ACM SIGOPS Eur. Workshop: Beyond The Pc, P.29-Es*, Leuven, Belgium, Sep. 19–22, 2004.
[7] L. Lu, P. Varman, and J. Wang, "Diskgroup: Energy efficient disk layout for raid1 systems," in *Int. Conf. on Networking, Architecture, And Storage*, Guilin, Jul. 29–31, 2007.
[8] C. Liao, "Accelerating file systems by predicting access frequency," in *UK e-Science All-Hand Meeting*, Nottingham, U.K., Nov. 2007.
[9] F. Wang, Y. Deng, N. Helian, S. Wu, V. Khare, and C. Liao, "Evolutionary storage: Speeding up a magnetic disk by clustering frequent data," *IEEE Trans. Magn.*, vol. 43, no. 6, Dec. 2007.
[10] "HP Storage Systems: Tools And Traces," [Online]. Available: www.hpl.hp.com 2008
[11] *"Disksim 3.0 Manual,"* 2007.
[12] F. Wang, S. Wu, N. Helian, Y. Deng, A. Parker, Y. Guo, and V. Khare, "Grid-oriented Storage," *IEEE Trans. Computers*, vol. 56, no. 4, 2007.
[13] [Online]. Available: En.Wikipedia.Org/Wiki/Solid_State_Disk2008
[14] [Online]. Available: www.esacademy.com/faq/docs/flash2008
[15] M. Uysal, A. Merchant, and G. Alvarez, "Using MEMS-based storage in disk arrays," in *FAST*, Mar. 2003.
[16] F. Z. Wang, "Diode-free magnetic random access memory," *Applied Physics Letters*, vol. 77, no. 13, 2000.
[17] MRAM, Freescale Semiconductor, Inc., 2008.

Manuscript received March 03, 2008. Current version published December 17, 2008. Corresponding author: F. Wang (e-mail: frankwang@ieee.org).