Principles of Operation for Shingled Disk Devices

Garth Gibson, Greg Ganger

CMU-PDL-11-107
April 2011

Parallel Data Laboratory
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

A leading strategy for driving the areal density of magnetic disk drives through 1 – 10 terabit/inch^2 (the coming decade) is to shingle (partially overlap) adjacent tracks, imposing significant restrictions on where data can be written without incurring multi-track read-modify-write penalties. These restrictions and penalties can be 1) fully hidden from system software using techniques familiar in NAND Flash disks; 2) minimally exposed as multi-track, shingled bands of predetermined size that can be read normally, but only appended to or trimmed (erased); or 3) maximally exposed as dynamically sized bands of shingles separated by guard regions of previously erased tracks, allowing maximal capacity to be obtained by the most sophisticated system software. While the latter options require significant changes in system software, there is a rich history of demonstrations of log-structured file systems that should be able to do this, and a profusion of write-once cloud storage systems that could provide the economic "killer application" [Kleiman11].

Now is a very good time for systems software experts to take interest and weigh in as magnetic disk technologists are experimenting and prototyping shingled disks. Experience shows that changes in the interface model for magnetic disks can take decades to change (for example, 512B to 4096B sectors) unless device vendors and systems software developers work together toward mutually desired principles of operation.

Acknowledgements: The work in this paper is based on research supported in part by the Department of Energy, under award number DE-FC02-06ER25767, by the Los Alamos National Laboratory, under contract number 54515-001-07. We also thank the member companies of the PDL Consortium (including APC, EMC, Facebook, Google, Hewlett-Packard, Hitachi, IBM, Intel, LSI, Microsoft, NEC, NetApp, Oracle, Panasas, Riverbed, Samsung, Seagate, STEC, Symantec, and VMware) for their interest, insights, feedback, and support.

Keywords: storage systems, shingled magnetic recording, magnetic disk interface

1 Introduction

Solid-state storage devices such as NAND flash are dramatically changing the playing field for durable storage systems, especially for smaller, randomly accessed, performance-sensitive storage systems. Still for at least the next decade the bulk of online stored information will magnetically recorded on hard disks because of the small size of magnetically recorded bits and the low cost for a device containing dozens of terabits. Consumer expectations for ever larger capacity magnetic disk drives, and the economics of the magnetic disk drive marketplace, call for an annual aggressive increase in areal density (more than 30% per year, recently about 40% per year). Maintaining this rate of increase in areal density in the face of the impending superparamagnetic limit is the core challenge of magnetic recording technologists today.

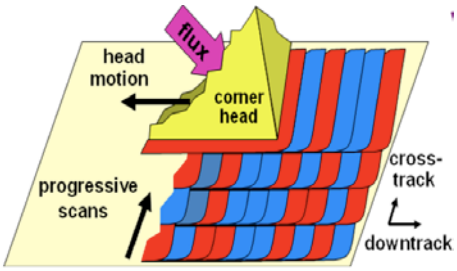

Figure 1: Shingled writing does not write data in non-overlapping tracks with non-magnetized "guard regions" between tracks, as is done in today's magnetic disks. Instead, shingled writing seeks only a fraction of the width of the written track before it writes the next track, largely overlapping adjacent tracks. This significantly increases tracks per inch, increasing areal density without significantly changing materials or write head design.

The manufacturing challenges inherent in next generation (higher areal density) magnetic recording techniques such as heat-assisted (integrating a laser into the write head, HAMR) [Rottmeyer06] or bit-patterned (fabricating discs with densely stippled protrusions, BPMR) [Richter06] magnetic recording encourage technologists to consider a shingled writing approach [Gibson09, Greaves09, Wood09, Amer10, Cassuto10]. As illustrated in Figure 1, shingled tracks overlap previously written tracks, exploiting the easier task of reading thin tracks than of writing thin tracks. Shingled magnetic recording (SMR) [Wood09] offers fewer technology and manufacturing challenges, but has a significant impact on the data organization and behavior of data access. Specifically, rewriting a sector on a track that has been shingled over cannot be done without overwriting subsequent ("down-band") tracks. As a result, data on a shingled disk will be organized into bands of shingled tracks with a non-shingled guard region between bands. The modification of a portion of a shingled band will require pre-reading and rewriting the modified sectors and all down-band sectors (which will be overwritten as a result of the modification).

Given the example of a "Flash Translation Layer" (FTL) in solid state storage devices (SSD) based on NAND flash devices [Gal05], one approach to integrating shingled writing into magnetic (hard) disk drives is to emulate the full range of non-shingled disks in an embedded controller, a "Shingle Translation Layer" (STL) [Gibson09]. As in the wide range of SSD product prices, speeds and sophistication, STL firmware could, at one extreme, inexpensively implement slow read-modify-write for essentially all writes to the shingled disk. Or, at the other extreme, it could expansively remap the physical location of written data to avoid read-modify-write, dynamically defining band boundaries, employing large write-back caches and overprovisioned disk capacity to hide a background garbage collection and defragmentation process.

We firmly expect many sophisticated implementations of a shingle translation layer to be developed in research, and perhaps in products. The compelling benefit of doing this is that user and system software can be used without change. However, experience with SSDs indicates that FTL performance can be significantly unexpected [Bjorling10], leading to system software reverse engineering SSDs to achieve higher level goals. And where there are large performance differences between hard disks and SSDs, systems software will naturally be modified as needed to optimize SSD performance advantages [Andresen09]. Moreover, the same market trends that drive geometric improvements in areal density and deliver geometric improvements in price per bit will inhibit the integration of an expensive STL onto each shingled disk.

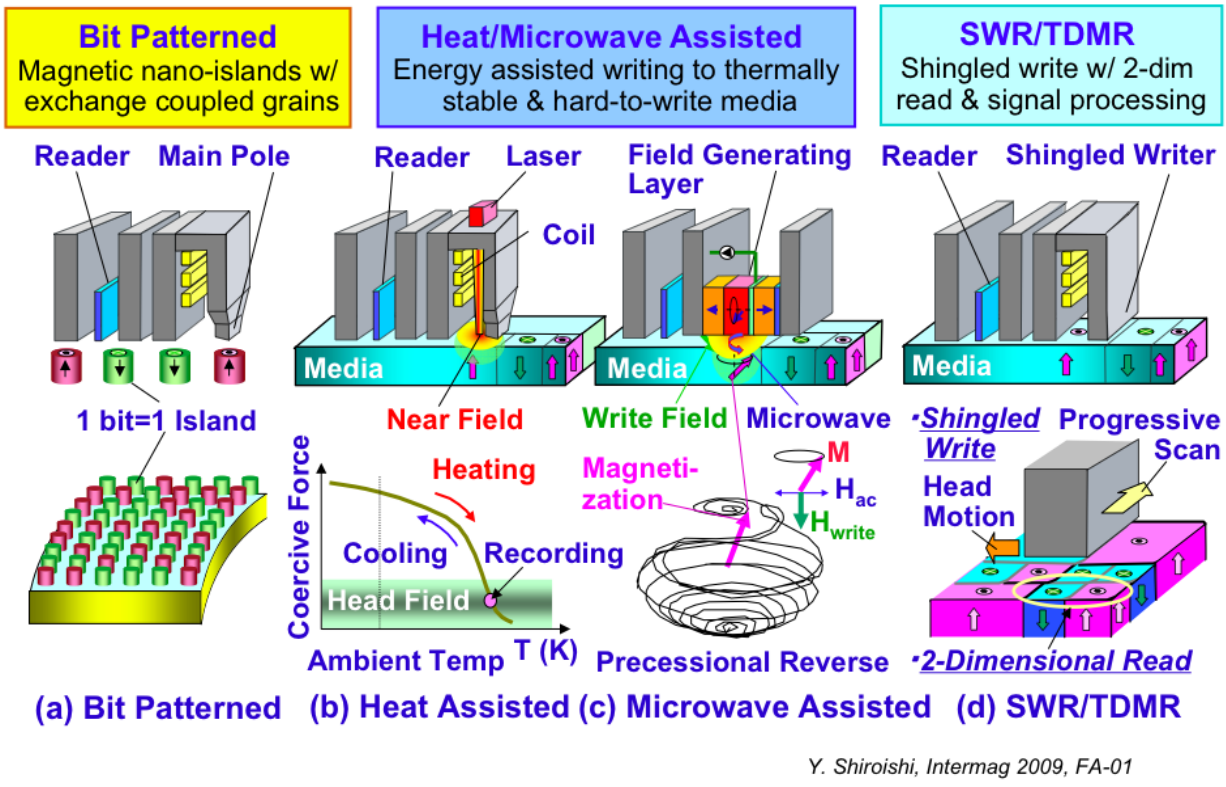

Figure 2: Cartoon characterization of the primary magnetic recording technology alternatives for achieving multiple terabits per square inch this decade, from the work of Shiroishi et. al. presented at the International Magnetics Conference, May 2009.

Our premise, then, is that a simple and inexpensive STL be embedded in each shingled disk, and that the system model of shingled disks be extended to facilitate system software to optimize for shingling without fighting or reverse engineering the STL. Specifically, system software should:

- make extensive use of the SSD-inspired TRIM command [T13-2009] to tell the STL when overwriting a down-band sector is free of consequences because no future read of that down-band sector is expected to return the currently stored data[1],
- be aware of bands of shingled tracks – possibly controlling the size and alignment of the bands – so it can avoid writing in the middle of a previously written band (without TRIMing down-band sectors),
- conceptualize writing to shingled bands as overwriting consecutively or appending to a log,
- perform necessary data remapping and garbage collection with read, TRIM and append operations.

With these principles of operation, the STL may never need to execute read-modify-write (which perhaps should still be implemented for the rare cases in which higher level software has been incompletely optimized) or garbage collection (which should probably not be implemented in the STL to simplify its logic and minimize its cost).

One concession an STL can provide for the inevitable remaining requirements randomly written sectors would be:

- a small fraction of the shingled disk, or perhaps one band, could be formatted unshingled, so random write does not require read-modify-write or garbage collection.

[1] With extensive use comes the need for negligible response time and efficient background processing, which should not be taken for granted.

A more impactful mode of operation for the STL and shingled disk mechanism could be:

- with a baseline of one continuous band of shingles across the entire surface, any erased (TRIMed) extent large enough to terminate down-band overwriting could be seen as a dynamic band boundary.

With some variation these model, we believe that shingled disks can achieve marketing targets for geometric improvements in cost per bit through 1-10 Tbit/inch^2 [Wood09], and system software can fully exploit the underlying shingled architecture without reverse engineering complex STL behaviors.

2 Magnetic Recording Technology Trends

Today's best areal density on magnetic disks exceeds 550 gigabits/inch^2 and historical trends (important to marketing magnetic hard disks) predict an annual increase in areal density of at least 40% per year. But at 1 terabit/inch^2 today's perpendicular recording hits the often quoted Superparamagnetic Limit [Wood2000], the density at which a bit written with current materials and writing mechanisms has too few magnetic grains to resist random grain orientation flips (for commercial thresholds on durability of ten years and one hundred writes to neighboring sectors or tracks). To continue the commercially expected 40% per year increase in areal density, new technologies are being explored, the leading of which are illustrated in Figure 2 [Shiroishi09].

The three leading approaches to achieving areal densities of 1-10 terabit/inch^2 are:

- Bit Patterned Magnetic Recording (BPMR): develop techniques for patterning the media into "islands" of grains in well-defined and well-isolated positions, to reduce inter-grain interference and allow new freedom in the materials to be employed,
- Heat (or Microwave) Assisted Magnetic Recording (HAMR): if only the bit being written can be heated then a material with smaller gains and more resistance to orientation flips at normal temperatures can be used (less well understood, but perhaps more manufacturable may be the use of a nano spin-torque microwave oscillator to magnetically write small grain media), and
- Shingled Magnetic Recording (SMR, sometimes SWR): using traditional materials and heads, abandon non-overlapping tracks so that adjacent tracks can be much closer together, because reading ever thinner tracks is easier than writing ever thinner tracks.

For BPMR and HAMR technologies to be commercially successfully significant and expensive changes in the disk manufacturing process are needed. For BPMR, the manufacturing process is stressed by the extreme nano-lithography and/or imprinting needed to pattern the media. For HAMR, in addition to the development of new media, recording heads with integrated, precise heat sources (lasers) must be developed, and will need new manufacturing processes. For both BPMR and HAMR the failure characteristics of these new materials and manufacturing processes are not well understood. The practical consequence these manufacturing process challenges is slower, more expensive product development, threatening to slow the rate of increase in areal density and damage the economic role of magnetic disks in the next decade of computing.

SMR technology, by defining a down-band adjacent track that will be overwritten, allows significantly more engineering freedom in the design of the write head. Conventional write heads need to write a narrow track with sharply defined edges on both sides, tolerant of the skew introduced by different angular positions (inner to outer diameters). Shingling means that the track written can be wider, allowing a stronger write field, only one edge needs to be sharp so it can be sharper. Of course, the read head has to read thinner tracks with much closer adjacent tracks causing interference (this will be discussed again later).

One recent study [Tagawa09] modeled the shingled writer track twice as large as the conventionally written head, predicting a write field strength 50% larger,

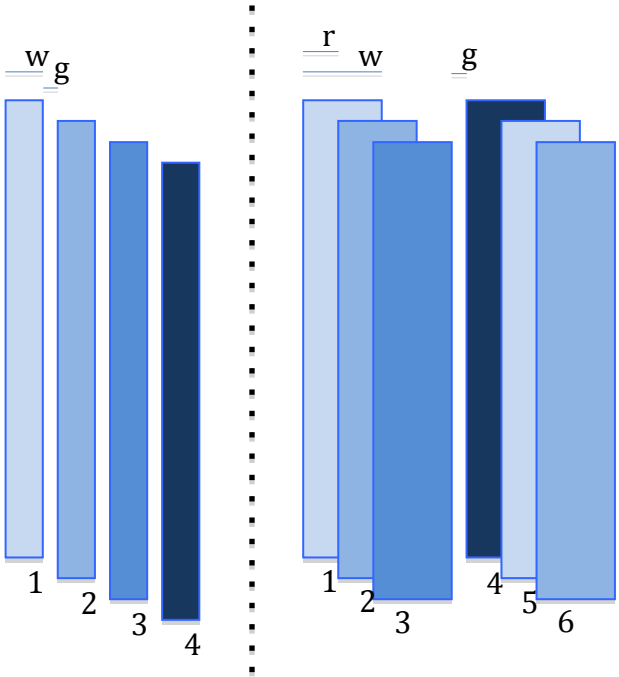

Figure 3: On the left, a conventionally written track has a non-overlapping width w separated by guard gaps of width g. Shingled writing, on the right, uses wider writing, width w', but overlaps with spacing width r, the residual track width, and the same guard gap.

inducing a magnetic grain size 33% smaller, leading to a predicted maximum areal density increase of a factor of 2.25 (with one huge band, an perhaps 10% less with bands of 100 adjacent tracks). Doubling the capacity of a disk drive with little change in the design or manufacturing of heads and media is highly interesting, if only to gain about 2.5 years on the 40% per year expected growth in which BPMR or HAMR manufacturing can be worked out.

3 Shingling Geometry Model

Inspired by Amer's model for disk density sacrificed by dedicating a fraction of each surface to unshingled tracks [Amer10], we develop the following "cartoon" model for how density is increased in shingled writing, as illustrated in Figure 3. Conventional recording writes track of, lets say, w=25 nanometers, with guard gaps between tracks of, say g=5 nanometers [Tagawa09]. Shingled writing, with its wider tracks, say w'=70 nanometers, achieves a sharper edge so that the adjacent, down-band track may be as close as r=10-20 nanometers, offering a maximum increase in areal density of 2.25 times. Adding a division of tracks into (1-f) shingled and (f) unshingled, following Amer10, gives an areal density increase factor (A), as a function of tracks per band, B (B=3 in Figure 3), illustrated in Figure 4.

$$A = S(f/S' + B(1-f)/(S'+B-1))$$
$$S = (w+g)/r, \quad S' = (w'+g)/r$$

First, we are not claiming that residual track width can be as small as 10 nanometers, but if the technologists can achieve shingled write heads with very sharp edges, then this is a model of the potential areal density increase. In fact most technologists talk about initial products that achieve less than a factor of two in areal density, but in the world of continually improving products, we expect improvement over time.

One interpretation of Figure 4 is that if residual track width is not much smaller than conventional track width, very large band sizes will be needed to achieve significant increases in areal density. Or conversely, if residual track width can be much smaller than conventional track width, then relatively small band sizes can be quite effective.

Another interpretation of Figure 4 is that 1% unshingled tracks has negligible impact on total disk capacity, and even 10% unshingled tracks is viable, if higher level software needs it.

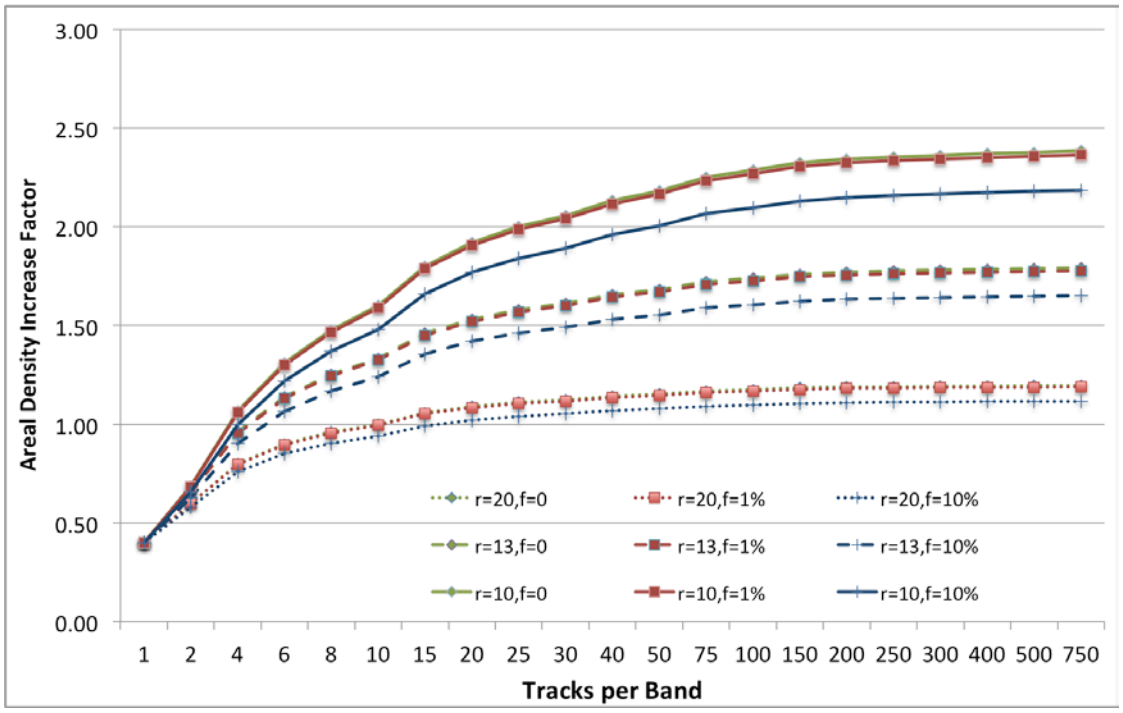

Figure 4: Multiplier on number of tracks per surface (and areal density) as a function of the number of tracks per shingled band. Three families of three curves are shown: three residual track widths (20, 13, 10 nanometers) and three fractions of unshingled tracks (0%, 1%, 10%). Conventional track width is 25 nanometers, shingled write size is 70 nanometers and guard bands are 5 nanometers.

4 Shingled Writing Systems Issues

When we were first introduced to shingled magnetic recording, three years ago, the principle question from magnetic technologists was "can system software cope with shingled writing?" Our first level answer at the time was to say that read-modify-write of entire bands on every small random write would yield such a reduction in write performance that such a product would be difficult to use with unmodified file and database systems. But the flash translation layer in modern SSDs has the same problem, because of large erase blocks and slow erase functions, and it has been overcome with remapping of logical block addresses to flash pages, overprovisioning of flash capacity and background cleaning/defragmenting of logically overwritten values [Gibson09]. Recently Amer et. al. and Cassuto et. al. have delved more deeply into the use of dynamically remapping and log-structuring of changing data on shingled written disks, as well as embedded non-volatile solid-state storage for delaying updates to exploit coalescing of changes [Amer10, Cassuto10].

If a shingled disk has an interface at all different from the current SSDs and HDDs, the determination of this new interface is a key challenge that systems software experts should engage in determining [Amer10, Cain10]. A partial list of such interfaces includes:

- No change – full emulation of traditional HDD operations embedded in the shingled disk,
- Large sector disks – if sectors are logically tens to hundreds of megabytes and fixed, then bands will always be logically erased and written consecutively – this is often cited as a backup or cloud storage specialized device,
- Append only fixed sized bands – sequential writing of bands is achieved by systems software logging changes and performing garbage collection at a higher level,
- Separate unshingled region – a small portion of the disk can be used for bands of one track only, eliminating down-band adjacent tracks, and allowing small random writes to be unpenalized – notice that if shingled tracks are written wider, this section of the disk has a bad areal density, and
- A data management and hints interface – like the TRIM command, band geometry discovery, creation/deletion of unshingled bands, etc., may enhance the performance of shingled disks,
- An object storage interface – SCSI has a command set for access by object ID with variable length data per object – this "network inode-like" model for storage is widely used for scalable storage systems in high-performance and data-intensive computing storage, but its use at the SCSI device level is still rare [T10-2004],
- A virtual tape model – if the entire device can be shingled into one band and higher level software creates band gaps (tape file marks and gaps) dynamically as needed, then maximal density is available to higher level software.

A key aspect in many of these will be the explicit exposure of the band size, alignment and placement to higher-level software and good support for "usable" band sizes—most higher level software is going to work with ranges that are reasonable numbers of power-of-two-sized blocks, not arbitrary counts of sectors. The choice of band size represents some trade-offs along several dimensions, relating to effective density (space wasted to boundaries between bands and to rounding off to usable band sizes) and consistency (is variation across zones or disks allowed?). An appropriate interface will expose or negotiate the band size(s) for a disk and the number of bands. Reads and writes may use logical block numbers that are explicitly tuples (band number and offset into band, very object like) or for which the mapping to those tuples is known.

These questions and considerations need research, discussion and, in general, airing among systems software experts now, before the magnetic disk technologists finalize the principles of operation for shingled magnetic recording.

5 Two Dimensional Magnetic Recording

The potential areal density increase of 2.25 predicted recently [Tagawa09] did not assume the residual track, after shingling the down-band track, was much smaller, so much of the benefit predicted comes from fewer gaps between tracks (only between bands). For much smaller residual tracks with overlapped tracks written down-band, technologists are exploring Two Dimensional Magnetic Recording (TDMR) [Wood09]. In TDMR the signal read by one head is compared to signals read at slight offsets on either side to integrate the interference from adjacent tracks. This change in the signal processing is a larger challenge for SMR, but even more importantly, if the offset signals are acquired by additional (2 or more) full rotations of the surface, the response time penalty for small

random reads will be dramatic and unacceptable. For large sequential reads this overhead may only be a couple of rotations in addition to many rotations needed to read the data, but for small random reads it is easily factors of 2 or 3 on the response time.

As we have said, significant benefits in areal density may be achieved with little reduction in the residual (readback) track width. However, technologists are pursuing read heads that actually have three read heads at slightly different offsets, closely aligned along the motion of the track rotating under the head. This has a larger impact on manufacturability and signal processing systems, so it is likely to be integrated into magnetic devices more slowly.

6 Summary

Shingled magnetic recording is a leading technological approach for scaling magnetic disk areal density over the next decade at the aggressive annual rates customers have come to expect. Shingled recording, however, overlaps adjacent tracks so that rewriting a prior track cannot be done without damaging a subsequent track. The systems issues in terms of the interface that such disks should offer is a topic that systems software experts should engage with magnetic disk technologists to decide.

7 References

[1] Amer, A., D.D.E. Long, E.L. Miller, J.-F. Paris, T. Swarz S.J., "Design Issues for a Shingled Write Disk System," 26th IEEE (MSST2010) Symposium on Massive Storage Systems and Technologies, Lake Tahoe, NV, May 2010.

[2] Andersen, D.G., J. Fanklin, M. Kaminsky, A. Phanishayee, L. Tan, V. Vasudevan, "FAWN: A Fast Array of Wimpy Nodes," 22nd ACM Symposium on Operating Systems Principles (SOSP), Big Sky, MT, Oct. 2009.

[3] Bjorling, M., L.L. Folgoc, A. Mseddi, P. Bonnet, L. Bouganim, B.P. Jonsson, "Performing Sound Flash Device Measurements: Some Lessons from uFLIP," 2010 ACM SIGMOD/PODS Conference, Indianapolis, IN, June 6-11, 2010.

[4] Cain, W., E. Champion, C. Stevens, "Future HDD Technologies and Prospects for Shingled Recording," 9th Perpendicular Magnetic Recording Conf., Sendai, Japan, May, 2010.

[5] Cassuto, Y., M.A.A. Sanvido, C. Guyot, D.R. Hall, Z.Z. Bandic, "Indirection Systems for Shingled-Recording Disk Drives," 26th IEEE (MSST2010) Symposium on Massive Storage Systems and Technologies, Lake Tahoe, NV, May 2010.

[6] Gal, E., S. Toledo, "Algorithms and data structures for flash memories," ACM Computing Surveys, v37, n2, June 2005.

[7] Gibson, G., M. Polte, "Directions for shingled-write and two-dimensional magnetic recording system architectures: Synergies with solid-state disks," Carnegie Mellon Univ. Parallel Data Lab Techn. Report, CMU-PDL-09-104, May 2009. Also FA-06 presentation in INTERMAG 2009, Sacramento, CA, May 2009.

[8] Greaves, S., Y. Kanai, H. Muraoka, "Shingled recording for 2-3 Tbit/in^2," IEEE Trans. on Magnetics, v45, n10, 2009. Also FA-03f presentation in INTERMAG 2009, Sacramento, CA, May 2009.

[9] Kleiman, S., NetApp University Day presentation, Sunnyvale, CA, Feb 14, 2011.

[10] T10 Object-Based Storage Device Commands (OSD), ANSI standard INCITS 400-2004, 2004.

[11] Richter, H., A. Dobin, O. Heinonen, K. Gao, R. Veerdonk, R. Lynch, J. Xue, D. Weller, P. Asselin, M. Erden, B. Brockie, "Recording on bit-patterned media at densities of 1 Tbit/in2 and beyond," IEEE Trans. on Magnetics, v42, n10, 2006.

[12] Rottmeyer, R.E., S. Batra, D. Buechel, W.A. Challener, J. Hohlfeld, Y. Kubota, L. Li, B. Lu, C. Michalcea, K. Mountfiled, K. Pelhos, R. Chubing, T. Rausch, M.A. Seigler, D. Weller, Y, Xiaomin, "Heat-assisted magnetic recording," IEEE Trans. on Magnetics, v42, n10, 2006

[13] Shiroishi, Y., K. Fukuda, I. Tagawa, H. Iwasaki, S. Takenoiri, H. Tanaka, H. Mutoh, N. Yoshikawa, "Future Options for HDD Storage," IEEE Trans. on Magnetics, v45, n10, 2009. Also FA-01 presentation in INTERMAG 2009, Sacramento, CA, May 2009.

[14] Tagawa, L., M. Williams, "Shingle-Write Technology and Gain Estimation," FA-02 presentation in INTERMAG 2009, Sacramento, CA, May 2009.

[15] T13 Technical Cmte, "Information technology–ATA/ATAPI Command Set 2 (ACS-2), T13/2015-D, rev 2," Aug 2009, pp. 99. www.t13.org/Documents/UploadedDocuments/docs2009/d2015r2-ATAATAPI_Command_set_-_2_ACS-2.pdf

[16] Wood, R., "Feasibility of Magnetic Recording at 1 Tbit/inch2," IEEE Trans. on Magnetics, v36, n1, 2000.

[17] Wood, R., M. Williams, A. Kavcic, J. Miles, “The Feasibility of Magnetic Recording at 10 Terabits per square inch on Conventional Media,” IEEE Trans. on Magnetics, v45, n10, 2009.
[18] Wood, R., “Shingled Magnetic Recording and Two-Dimensional Magnetic Recording,” IEEE Magnetics Society Meeting, Santa Clara, CA, Oct. 19 2010.