最近の磁気ディスクドライブに於ける高遅延特性の観測とデータベース処理性能への影響の考察

佐藤　佑紀† 　早水　悠登†† 　合田　和生†† 　喜連川　優†††

† 東京大学情報理工学系研究科　〒 113–8656 東京都文京区本郷 7–3–1
†† 東京大学生産技術研究所　〒 153–8505 東京都目黒区駒場 4–6–1
††† 国立情報学研究所　〒 100–1003 東京都千代田区一ツ橋 2–1–2
E-mail: †{satoyuki,haya,kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　磁気ディスクドライブはそれが開発されて以降ストレージ技術に於いて中心的役割を担ってきたが, 時代の変遷により扱うデータ量が増えると同時に, 磁気ディスクドライブにも常に大容量化が望まれてきた. 現在, 大容量化を牽引し注目を集めている技術が, 瓦書き磁気記録技術であるが, その性能特性については詳細な調査がなされていない. 本論文では, 最近の磁気ディスクドライブの中でも瓦書き磁気ディスクドライブに着目し, マイクロベンチマークを用いて高遅延特性の観測を行うとともに, 書き込み負荷が性能特性に与える影響を調査した.

キーワード　磁気ディスク, データベース, 高遅延

1. はじめに

磁気ディスクドライブの歴史は大容量化の歴史でもある. 世界で初めての磁気ディスクドライブは 1956 年に誕生した. 磁気ディスクドライブは誕生した当時から大容量化が熱望されており, IBM によって開発された IBM RAMAC 305 に搭載された世界最初の磁気ディスクドライブである IBM 350 は 24 インチのプラッタ 50 枚で構成されていたが, その記憶容量は 5MB に満たなかった. 1961 年に導入された IBM 1301 はヘッドを読み取り用と書き込み用の 2 つ用いる技術, ヘッドを空気抵抗でディスクから自力で浮かせることで書き込み, 読み取りを行う技術が初めて用いられ, IBM 350 と比べて同面積で 13 倍の記憶容量を実現した. その後も技術革新は進み, 1980 年に現在の Seagate が記憶容量 5MB でパソコン用の 5.25 インチ HDD を開発して以来, パソコン用磁気ディスクの開発が進み, 1980 年代後半頃には 3.5 インチ HDD の開発が主流となっていった. 大きさが 3.5 インチで統一されたことでますます大容量化に求められる技術は高度化していく. まずは, 読み込み, 書き込みに用いられ, トラック幅に直接関係してくるヘッド技術が 1990 年代に MR ヘッド (Magneto Resistance Head) へと変わり,2000 年に入ると GMR ヘッド (Giant Magnetic Resistance Head) そして TMR ヘッド (Tunnel Magneto Resistance Head) へと移行していった. またヘッド技術のみならず垂直磁気記録方式といった記録方式においても技術が向上し現在に至る.

最近では, エネルギーアシスト磁気記録, ビットパターンメディア, 瓦書き磁気記録などの技術が注目を集めている. エネルギーアシスト磁気記録技術は大きく分けて熱アシスト磁気記録 (Thermal Assisted Magnetic Recording) とマイクロ波アシスト磁気記録 (MAMR : Microwave Assisted Magnetic Recording) の 2 種類が存在し, これらは記録メディアの磁性粒を小さくして記録密度を高めることにより失われてしまう熱安定性を高保磁力材料で補った際に, 現在の磁気ヘッドの磁気では書き込めなくなってしまう問題が生じるので熱やマイクロ波を照射することで, 高保磁力材料に磁気を通りやすくする方式である. ビットパターンメディア (BPM : Bit Pattened Media) は磁気メディアの表面に加工をすることでノイズの発生を抑制する方式である. これらの技術は HDD を構成する部品の性能を向上させることで高密度化, 大容量化を図る方式であるが, 瓦書き磁気記録 (SMR : Shingled Magnetic Recording) は部品レベルではなく記録方式の変更によって高密度化, 大容量化を図る方式であるため他の方式に比べて実現しやすく, すでに製品化もなされており, 現在の磁気ディスクドライブの高密度化を牽引している.

本論文では, この SMR 技術を用いた磁気ディスクドライブに対して書き込み負荷を与えた後のアクセスパターンについて観測及び考察をするとともに, データベースシステムに於いてそれらのアクセスパターンが及ぼす影響について考察を行う.

本論文は以下のように構成される. 第 2 章で瓦書き磁気ディスクドライブ技術を説明する. 第 3 章では著者らが行った性能評価試験の内容について述べる. 第 4 章では評価試験で得られた結果を示し考察する. 最後に第 5 章に於いてまとめと今後の展望について述べる.

2. 瓦書き磁気ディスクドライブ

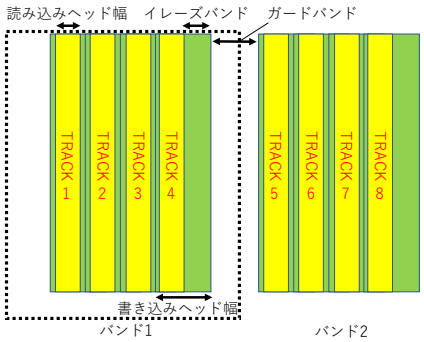


図 1: SMR の記録方式

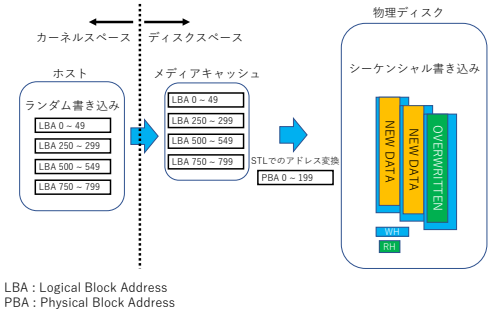


図 2: Shingled Translation Layer での動作

瓦書き磁気ディスクドライブ (SMR ディスク) の技術は 2009 年に R.Wood らによって提唱され, 2014 年に一般向け HDD として初めて製品として販売された [3]. 図 1 に SMR の記録方式を示す. SMR 技術とは, 書き込み用の記録磁気ヘッドの技術的限界によりトラック幅を狭めることができないことにより記録磁気ヘッドより相対的に幅が狭い読み込み用の磁気ヘッドにトラックの幅を合わせるために, 以前に書いたトラックを少しずつずらしながら重ねて書いていく方式である. この記録方式により高密度化が可能な理由は大きく分けて 2 つあり, 強い記録磁界が得られることと, 従来ディスクの記録面上に必要だった余分なスペースが減らすことができたことである. これは, 従来の方式でトラック密度を高めるためには記録磁気ヘッドの幅を狭めなくてはならなかったが, SMR 方式ではトラック幅よりも記録磁気ヘッドの幅を大きく設定できるため記録磁界を強く保つことができることにある. 後者はより複雑で, 磁気ディスクの記録密度を高めるためにはトラック間の幅 (トラックピッチ) を狭めなくてはならないが, その際には記録素子から生じる記録磁界が隣接する記録トラックに干渉して記録された情報を消去してしまう書き滲みと呼ばれる問題を考慮しなくてはならないことによる. 書き滲みによってデータが消されてしまう領域をイレーズバンドと呼ぶが, トラックピッチは通常このイレーズバンドを考慮して広めに取られている. 記録磁気ヘッドはディスクの内周部にアクセスする際と, 外周部にアクセスする際ではディスクに対しての傾きが異なり, これにより磁気ディスクの内周部では外周側の, 外周部では内周側のイレーズバンドが大きくなる. そのため従来は外周部や内周部はトラックピッチを大きく取っていた. しかし, SMR 方式ではイレーズバンドの狭い側のみを使用することが可能である. つまり, ディスクの外周では内周側に, 内周では外周側に向かって重ねて記録していけば今まで必要だったイレーズバンド間に必要な余分なスペースを減らすことができ, トラックピッチを狭めることが可能になるのだが, 余分なスペースが全く必要なくなるわけではなく, その代わりに 1 度に書き込む最小単位であるバンドと呼ばれるトラックのまとまった領域の間に存在するバンド間干渉を防ぐための仕組みであるガードバンドが必要になる.

次に SMR のファームウェアについて説明する. SMR ディスクを既存のディスクと同様に互換的に扱うのか, それとも新たなコマンドセットを用意し SMR ディスクに対してのアクセスを shingled なものに限定するのかといった調整を行うのがファームウェアである. ファームウェアは 3 種類が提案されており, SMR を従来の磁気ディスクとの互換性を保ちながら扱う方式を Drive Managed 方式, SMR ディスクへのワークロードをホストで shingled なアクセスに最適化して扱う方式を Host Managed 方式, 従来のディスクと互換的に扱うか SMR ディスクへのワークロードを調整するかホスト側で選択することができる方式を Host Aware 方式と呼ぶ. Host Managed 方式を採用すれば, SMR へのワークロードを適切に管理することができるので, SMR の shingled 構造に対して精緻な制御が可能になり, 安定したパフォーマンスが実現できる. しかし, Drive Managed 方式は既存のホストに対して一切の変更を加えずに SMR ディスクを扱う方式であるのに対して, Host Managed 方式ではまったく新しいソフトウェアが必要になり, さらにはファイルシステム, OS, ハードウェア等にも変更が必要となるため, 導入コストが非常に高い方式である. Host Managed 方式の基本は, 物理的な単位であるバンドに対して, 論理的な単位であるゾーンを割り振り, ゾーン単位でのアクセスを行うことなのだが, SMR への最適化のため Host Managed 方式ではゾーンアクセスに対して様々な制約が存在する. そこでその制約を緩和することで, Drive Managed 方式のように互換性を保ちながら, Host Managed のようなゾーン単位でのアクセスを可能にした方式が Host Aware 方式である. Host Aware 方式はその特徴から Drive Managed 方式と Host Managed 方式の長所と短所を併せ持っている. 現在は Drive Managed 方式が主流であるので以降は Drive Managed 方式を前提として説明する.

Drive Managed 方式の SMR ではデータの記録方式も従来と異なっている. 従来のデータ記録方式は, OS から受け取ったデータはディスクに内蔵されたメモリ上に蓄えられ, ディスクコントローラーがアクセス時間を最小にするように, 記録するデータのスケジューリングを行った後に実際に記録が行われていた. 一方, SMR 方式のデータ記録方式は, OS から受け取ったデータをメモリ上に蓄えるところまでは変わらないが, 新たにメディアキャッシュと呼ばれるキャッシュをディスクに内蔵し, バッファメモリ上のデータをメディアキャッシュへと書き移す. SMR はその特性上ランダム書き込みはできないので, メディ

アキャッシュ上のデータはシーケンシャル書き込み用に再構成され, バンド単位で書き込みを行う. 記録済みのデータを書き換える場合は, そのデータが存在するバンドのデータを 1 度メディアキャッシュ上に読み出し, メディアキャッシュ上で書き換える部分を書き換えた後, バンドに書き込みを行うという方式になっている. この際に, 元のバンドに書き込む場合に比べ, 新しいバンドに書き込む場合の方が高速であるが, 不要になったデータを削除するためのガベージコレクションの仕組みが必要になってくるため, ディスクコントローラに多少複雑な機構が必要になってくる. ガベージコレクションなどのこれらの SMR 独自の操作が行われるのが Shingled Translation Layer (STL) と呼ばれるレイヤで行われ, 図 2 のように最終的に物理ブロックアドレスへと変換がなされてディスクへ書き込まれる.

3. 性能試験手法

表 1: 性能試験環境

CPU	Intel(R) Xeon(R) CPU E3-1240 v5 @ 3.50GHz
Memory	DDR4 8192MB × 2
OS	CentOS release 6.8 (Final)
Kernel	2.6.32-642.4.2.el6
HDD	Seagate Archive Disk 8TB × 2

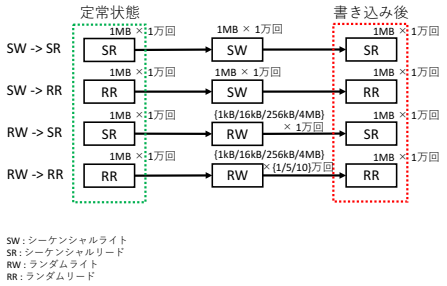


図 3: マイクロベンチマーク負荷

本節で述べる性能試験は SMR に書き込みを行った時, STL がどのような影響を及ぼすのかを確認することを目的とするものである. 性能試験は表 1 のような環境で行った. SMR ディスクは同じ型番のものを 2 種類用いて計測を行い, まず初期状態においての 2 つのディスクのアクセスパターンをシーケンシャル読み込みとランダム読み込みによって計測した. 計測の際は O_DIRECT によって OS のキャッシュは介さないようにし, 磁気ディスクのバッファキャッシュについてはある場合とない場合の計 4 通り計測した. 計測に於いては, バッファキャッシュがある場合とない場合の 2 通りについて計測を行ったが, 同様の傾向が見られたので以下ではバッファキャッシュがない場合の結果についてのみ触れる.

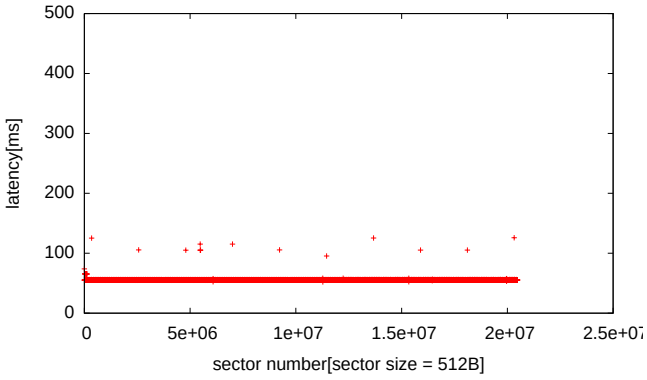
図 4 は初期状態の瓦書き磁気ディスクに対して, 先頭セクタ 10GB 分を 1MB 単位でシーケンシャル読み込みを行った際の結果及び, ランダム読み込みをディスク全体に対して 1MB 単位で 10GB 分行った結果であり, 縦軸は 1MB を読むのにかかった時間 (ms) を, 横軸は読み込んだセクタ番号をそれぞれ表している. 結果については両方のディスクで同様の結果が得られたので, 片方のディスクでの結果のみを示している.

初期状態において両方のディスクで同様の結果が得られたので双方のディスクにそれぞれ違う書き込み負荷を与えて, 読み込み性能の変化を観測した. 図 3 に実験に用いた負荷を示す. 読み込み性能の変化は, 書き込みを行う直前の読み込みによるレイテンシと書き込み直後の読み込みによるレイテンシを比較することで評価し, 書き込みと読み込みはそれぞれランダムとシーケンシャルの 2 通りずつ行い, 計 4 通り行った. シーケンシャル書き込みは先頭セクタから 1MB 単位で 10GB, ランダム書き込みはディスク全体に対して 1kB, 16kB, 256kB, 4MB の 4 種類の単位でそれぞれがほぼ同じ回数書き込まれるようにそれぞれ 10GB ずつ書き込み, シーケンシャル読み込みは先頭セクタから 1MB 単位で 10GB, ランダム読み込みはディスク全体に対して 1MB 単位で 10GB 読み込んだ. さらにランダム書き込み後のランダム読み込みについては定常状態に落ち着くまでの時間を見るために, 書き込み量を 1GB, 5GB, 10GB の 3 通りに変化させて読み込みレイテンシの時間変化の観測を行った. ここでいう定常状態とはメディアキャッシュに書き込みデータが残っていない状態のことを指す.

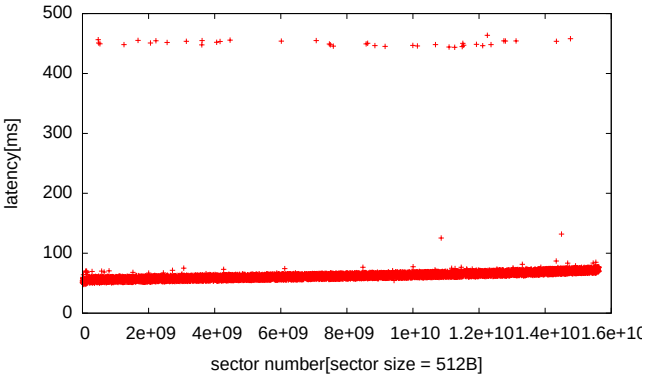
4. 実験結果と考察

図 5 は書き込み負荷を与える前後の読み込み性能を示している. 図中の緑の系列が初期状態を赤の系列が書き込み直後のレイテンシをそれぞれ表しており, 横軸はセクタ番号を縦軸はレイテンシを表している. また, 図のキャプションの SW はシーケンシャル書き込み, RW はランダム書き込み, SR はシーケンシャル読み込み, RR はランダム読み込みをそれぞれ表しており, SW → SR はシーケンシャル書き込みの前後にシーケンシャル読み込みを行っていることを表している. 図 6 は図 5 の実験におけるそれぞれのレイテンシの割合を cumulative curve で示しており, 横軸はレイテンシを縦軸はパーセンテージを表している. 図 7 は RW → RR において書き込み量を変化させた場合の時間経過に伴う読み込み性能の変化を示しており, 横軸は書き込みが終了した時刻を 0 秒とした経過時間を縦軸はレイテンシを表している.

図 6 と図 5 によると SW → SR, RW → SR では有意な性能低下は見られないのに対して, SW → RR では若干のレイテンシの増大が見られ, RW → RR では著しいレイテンシの増大が見られる. またランダム読み込みにおいてはいずれも高遅延の増大が見られ, SW → RR では 100ms 以上の高遅延が 3%に, RW → RR では 15%に増大していた. 図 7 によると書き込み終了後は読み込みレイテンシのバースト的な増大が観測された. このバーストが開始する時間は書き込み量によって再現性があるのである程度予測が可能であり, バーストの継続時間は書き込み量が増えるに従って増大することが分かった. SMR は従来型の磁気ディスクドライブとは異なるレイテンシの増大現象が見られるため, データベース処理において大量の書込みの後

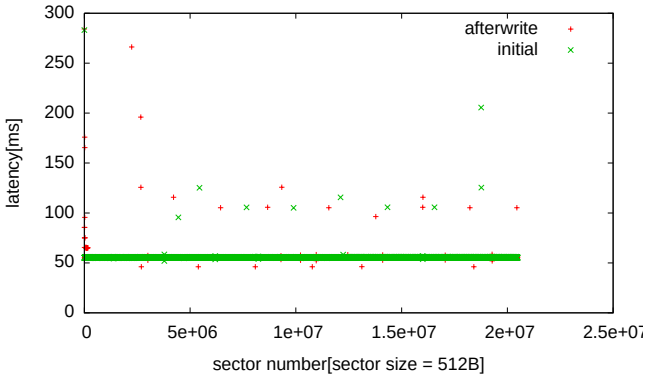


(a) シーケンシャル読み込み

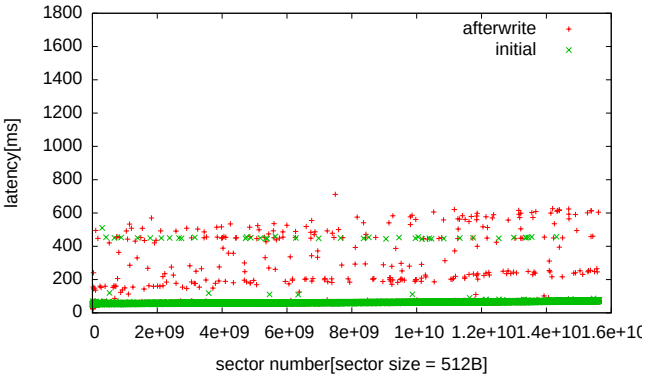


(b) ランダム読み込み

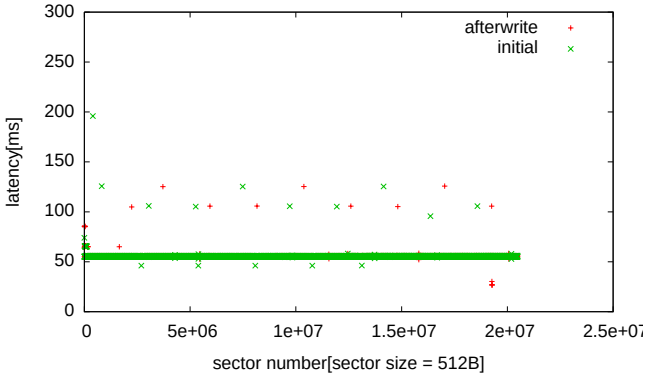
図 4: 初期状態



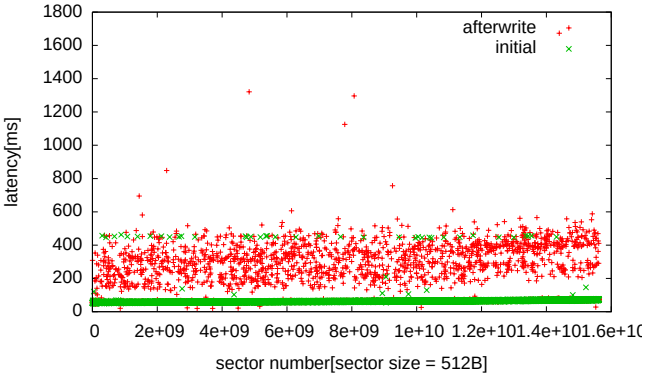
(a) SW → SR



(b) SW → RR

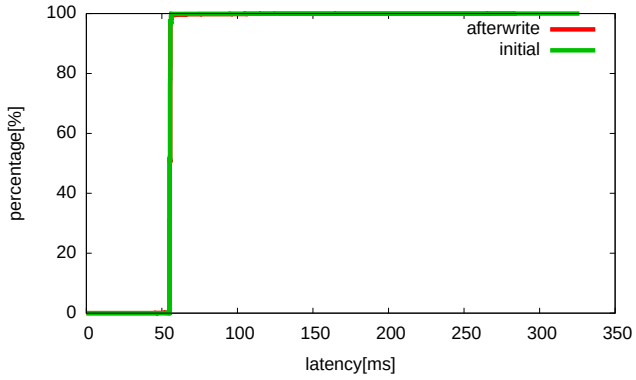


(c) RW → SR

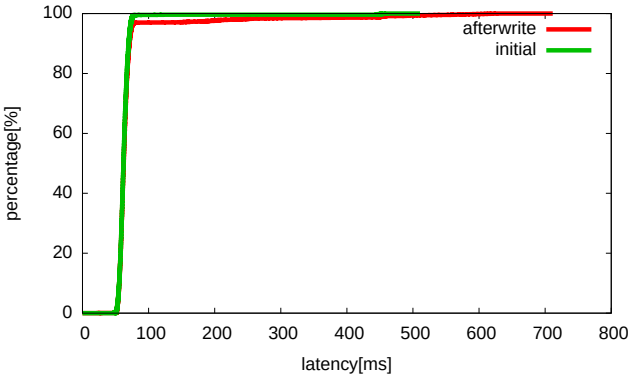


(d) RW → RR

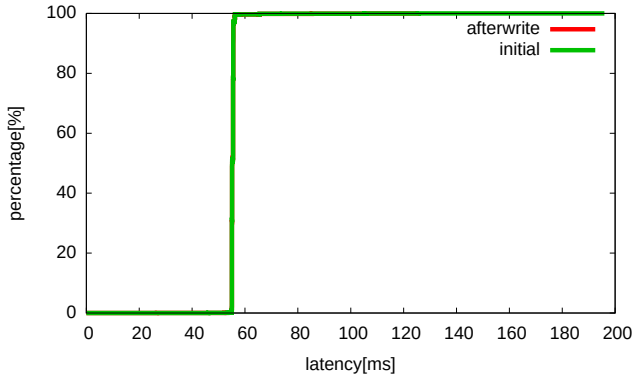
図 5: 読み込み性能の変化



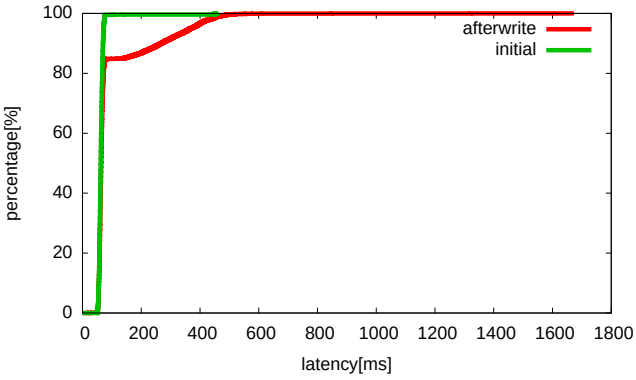
(a) SW → SR



(b) SW → RR

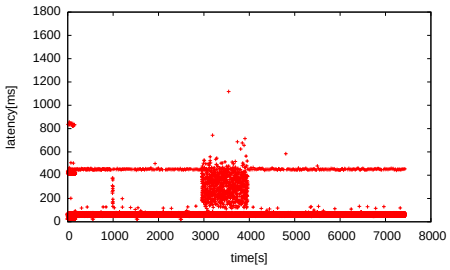


(c) RW → SR

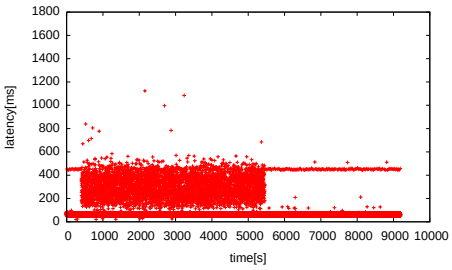


(d) RW → RR

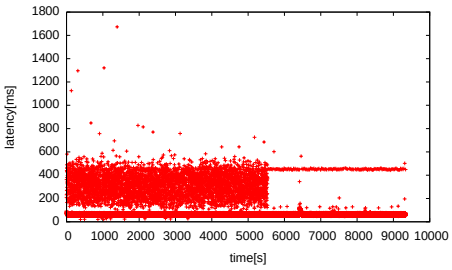
図 6: 読み込み性能の変化 (cumulative curve)



(a) 1GB 書き込み



(b) 5GB 書き込み



(c) 10GB 書き込み

図 7: RW → RR の書き込み量に対する読み込み性能の時間変化

に読込みレイテンシが増大する可能性があり, 性能管理上考慮が必要である.

5. おわりに

本稿では, マイクロベンチマークを用いて書き込み負荷に対する読み込み性能の観測を行った. その結果, 書き込み負荷によらずランダム読み込みでは性能が悪化すること, および書き込み後は書き込み量の増大に伴い長期化する読み込みレイテンシのバースト的な増大が生じること明らかになった.

今後の課題としては, 本稿で明らかになった高遅延特性がデータベースシステムへ与える影響の詳細な調査を行い, SMR のファームウェア方式によらず高遅延をホスト上で減らす機構の考案を行っていきたい.

謝　　辞

本研究の一部は, 国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）委託業務「エネルギー・環境新技術先導プログラム/革新的な省エネルギー型データベース問合せコンパイラの研究開発」及び「IoT 推進のための横断技術開発プロジェクト／先進 IoT サービスを実現する革新的超省エネルギー型ビッグデータ基盤の研究開発」に拠る.

文　　献

[1] 喜連川優, ”ストレージ技術　クラウドとビッグデータの時代”, オーム社, 2015.
[2] K. Goda, M. Kitsuregawa, ”The History of Storage Systems.”, Proceedings of the IEEE, 2012, 100.Centennial-Issue: 1433-1440.
[3] R. Wood, M. Williams, A. Kavcic, J. Miles, ”The feasibility of magnetic recording at 10 Terabits per square inch on conventional media”, IEEE Trans. Magn., vol. 45, no. 2, pp. 917-923, Feb. 2009.
[4] A. Amer, D. D. E. Long, E. L. Miller, J.-F. Paris, T. Schwarz, ”Design issues for a shingled write disk system”, 26th IEEE Symposium on Mass Storage Systems and Technology, pp. 1-12, 2010.
[5] M. Dunn, T. Feldman, ”Shingled Magnetic Recording Models, Standardization - SNIA”, 2014.
[6] 田河育也, ”次世代高密度ハードディスクドライブ開発における発想を転換したアプローチ”, 精密工学会誌, vol. 76, no. 7, pp. 755-758, 2010.
[7] 下村和人, ”HDD の大容量化をけん引する瓦記録技術”, 東芝レビュー, vol. 70, no. 8, pp. 29-32, 2015.
[8] Pitchumani, Rekha, et al. ”Emulating a shingled write disk.” 2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. IEEE, 2012.
[9] Aghayev, Abutalib, Mansour Shafaei, and Peter Desnoyers. ”Skylight—a window on shingled disk operation.” ACM Transactions on Storage (TOS) 11.4 (2015): 16.
[10] Hall, David, John H. Marcos, and Jonathan D. Coker. ”Data handling algorithms for autonomous shingled magnetic recording hdds.” IEEE Transactions on Magnetics 48.5 (2012): 1777-1781.
[11] Le Moal, Damien, Zvonimir Bandic, and Cyril Guyot. ”Shingled file system host-side management of shingled magnetic recording disks.” 2012 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2012.
[12] Jin, Chao, et al. ”HiSMRfs: A high performance file system for shingled storage array.” 2014 30th Symposium on Mass Storage Systems and Technologies (MSST). IEEE, 2014.