社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

磁気ディスクドライブの性能モデルの自動調整に向けた初期検討

別所祐太朗† 　合田　和生†† 　早水　悠登†† 　喜連川　優††,†††

† 東京大学大学院 情報理工学系研究科
†† 東京大学 生産技術研究所
††† 国立情報学研究所

あらまし　高密度化のすすむ近年の磁気ディスクドライブでは，不良セクタによるレイテンシ増加が発生しやすい傾向にある．こうした高いレイテンシは全体としてごく僅かしか発生していなかったとしても，システム全体の性能に無視できない影響を与えることが知られており，不良セクタによるレイテンシ増加の多い個体は機械的に検出できることが望ましい．本稿では異なる性能モデルを持つ磁気ディスクドライブ製品に対し, ランダムアクセスの性能からレイテンシモデルを自動調整する手法を提案する．これにより, 異なる製品に対する不良セクタへのアクセスを外れ値として検出できる．初期評価の結果, 異なるレイテンシ特性を持つ実際の異なる 3 製品に対して正常なレイテンシの範囲を正確に予測可能であること，また 1,000 秒程度で十分に安定した調整結果が得られることを確認した．
キーワード　磁気ディスクドライブ, レイテンシ, 性能モデル, 不良セクタ, 自動調整

A Preliminary Study Towards Autotuning of Performance Models for Magnetic Disk Drives

Yutaro BESSHO†, Kazuo GODA††, Yuto HAYAMIZU††, and Masaru KITSUREGAWA††,†††

† Graduate School of Information Science and Technology, The University of Tokyo
†† Institute of Industrial Science, The University of Tokyo
††† National Institute of Informatics

Abstract　With the trend of increasing sector density, hard disk drives are becoming more prone to abnormal latencies caused by bad sectors. It is known that such latencies could degrade overall system performance to a non-negligible extent, even if their frequency is quite low. In such a context, it is profitable for users to be able to identify drives in an automated manner that have more bad sectors, which produces more latenies. This work proposes a method to automatically tune the model parameters of different disk drive products by running a simple random access benchmark. Accesses to bad sectors are detected as latency outliers of trained models. Our initial implementation and evaluation shows that for 3 real-world drives with different latency characteristics, our tuning method are capable of accurately predicting their latency range of good sector accesses. It is also shown that sufficiently stable model outputs can be obtained by running the benchmark for around 1,000 seconds.
Key words　magnetic disk drive, latency, performance model, bad sector, auto tuning

1. 序　　論

磁気ディスクドライブは最も主要なストレージメディアであり [1]，2017 年の全世界のメディア出荷容量のうち 70% 以上を磁気ディスクドライブが占める [2]．データセンタにおけるストレージの大容量化等に牽引され，磁気ディスクドライブは高密度化を続けている [2]．情報を保持する磁性体が微細化するにつれて，製造段階の僅かな揺らぎや経年劣化の影響によってデータを正しく記録することができない，いわゆる不良セクタが増加する傾向にあることが知られている [3]．一般的な磁気ディスクドライブでは，少数の不良セクタが発生してもメディアとして利用可能であるように，ECC によるエラー訂正や代替セクタ割当て等の機能が備えられているが，こうした不良セクタへのアクセスには，正常なセクタにアクセスする場合よりも長いアクセスレイテンシが発生する．

通常, 不良セクタの数は正常なセクタと比較して極めて少数のものであるため，磁気ディスクストレージへの入出力命令のうち不良セクタによるレイテンシ増加の影響を受けるものは限

定的である．しかし，ごく一部の命令のレイテンシ増加によって，処理全体の性能に無視できない影響が生じる現象が所謂 tail latency として知られており，例えば RAID アレイを構成する特定のドライブのレイテンシ増加によって RAID アレイの性能低下が生じるといった現象 [4] や，複数ティア構成をとるサービスの下位層で発生したレイテンシが上位層へ伝播する現象 [5] などが報告されている．このことから，磁気ディスクドライブの不良セクタによるレイテンシ増加は，システム全体の性能への影響を無視することができない．

近年の磁気ディスクドライブは，個品レベルで不良セクタに起因すると見られる性能のばらつきが大きい機種も一定数みられ，磁気ディスクドライブを用いたシステムを構築する観点からは，こうした個品を機械的に検出可能であることが望ましい．正常なセクタへのアクセスレイテンシは, 磁気ディスクドライブを構成する物理機構からモデル化を行い，一定の精度で予測可能能であることが知られている [6], [7]．こうしたモデルを用いることで，不良セクタへのアクセスによるレイテンシ増加を検出することができる．しかしながら，シーク時間やディスクの回転速度などは個別の製品に固有のものであるため，幅広い磁気ディスクストレージの製品を対象として不良セクタアクセスによるレイテンシ増加を検出するためには，こうした各製品固有のパラメータの導出が必要となる．

本論文では，磁気ディスクドライブにランダムアクセスを行うワークロードを対象として，入出力ベンチマークによってレイテンシを計測することで，レイテンシモデルを自動調整する手法を提案する．そして，複数の磁気ディスクドライブに提案手法を適用することで，正常なレイテンシの範囲を正確に予測できることを実測によって確認することで，自動調整したモデルによって不良セクタアクセスによるレイテンシ増加を検出可能であることを示す．

2. 磁気ディスクドライブの機構とレイテンシモデル

本章では, 自動チューニングの対象とする磁気ディスクドライブの物理的機構による性能モデルについて述べる．

2.1 磁気ディスクドライブの物理的機構

磁気ディスクドライブにおいて, データの読み書きを行う物理的機構は, 主にディスクとアームの 2 つによって構成される．プラッターは表面に塗布された磁性体にデータを記録する薄い円盤状の部品であり, 作動時は一定速度で回転している．アームはデータの読み書きを行う磁気ヘッドを先端に持つ棒状の部品であり, アクセスしたい情報が記録されているプラッターの部位に, ヘッドを移動 (シーク) させる．

情報が記録されているディスクの表面は同心円状に細かく分割され, それぞれトラックと呼ばれる．トラックはさらに円弧状の記録単位に分割され, それぞれセクタと呼ばれる．セクタは磁気ディスクドライブがデータの読み出しと編集を行う最小の単位であり, 通常 4KB の記憶容量をもつ．

磁気ディスクドライブの記憶空間におけるブロックアドレスは，一般に外周のトラックから内周に向かって, セクタを単位として昇順に割り当てられる．セクタの中には，磁性体の製造時のばらつきや経年劣化により，正しくデータの読み書きを行うことができない不良セクタが一定数存在することが知られている．不良セクタにアクセスする場合には，ECC によるエラー訂正や，プラッタ最内周等に確保された予備領域への代替セクタ割当て，参照といった処理が通常のセクタアクセスに加えて発生する．即ち，磁気ディスクドライブにおいて，不良セクタは確率的に通常より長いアクセスレイテンシを生じされる主要な要因となっている．

2.2 磁気ディスクドライブの物理的機構に基づくレイテンシモデル

磁気ディスクドライブのアクセスレイテンシ, すなわち, データの読み出しあるいは書き込みの要求を受けてから操作が完了するまでの時間を, 該当セクタへアクセスする場合について, 以下の 3 つに分解してモデル化する．

- **シーク時間** t_{seek}: アームが読み出しヘッドを目的のデータを記録するトラックに移動させる際に経過する時間．シーク時間 t_{seek} は, ヘッドの移動距離 d の関数 $t_{seek}(d)$ とみなすことができる．ヘッドの位置 p をディスクの最外周トラックからの物理的な距離として表すことにすると (図 1), d は, ヘッドの移動前後の位置 p_{prev}, p_{access} を用いて $d = |p_{access} - p_{prev}|$ として表せる．

  ここでは，文献 [7] に示される前提に倣い，ヘッドが移動し始めてから停止するまでの運動は，ヘッドの始動時（停止から最高速度に至るまで）と，制動時（最高速度から停止に至るまで）には等加速度運動をするものとし，また最高速度での移動時は等速運動するものと仮定する．即ち，$t_{seek}(d)$ の関数形については, d が閾値 d_{th} より小の時は平方関数で, d_{th} より大のときは線形関数 (ただし, $d = d_{th}$ において s は連続および微分可能) としてモデル化することができる．

- **回転待ち時間** t_{rot}: ヘッドが目的のトラックへ移動を完了してから, 目的のセクタがディスクの回転によってヘッドの下に到達するまでの時間．t_{rot} はシーク終了時に目的のセクタがヘッド下にある場合に最小値 0 をとる．目的のセクタがヘッド下を直前に通過していた場合はディスクが更に 1 回転するのを待機する必要があり, t_{rot} は最大値 T_R (ディスクの回転周期) をとる．

- **データ転送時間** t_{trans}: ヘッドが目的のセクタの上に到達してから, データを転送が終了するまでに経過する時間．ディスクの回転速度, アクセスするセクタ数, セクタの密度，SATA 等接続インターフェースの通信速度などに依存する．

以上の議論より，磁気ディスクドライブの正常なセクタに対するアクセスレイテンシ t_{total} の上限値 t_{upper} および下限値 t_{lower} は, $0 \leq t_{rot} \leq T_R$ より, 以下のように表される．

$$t_{lower} = t_{seek} + t_{trans}$$

$$t_{upper} = t_{seek} + t_{trans} + T_R$$

2.3 シークプロファイルと定式化

磁気ディスクドライブに対して, ランダムに選択されたオフセットに対して同じ長さの読み出し要求を発行するワークロー

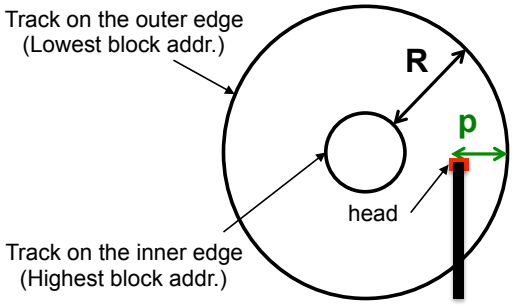


図 1: ディスク上のヘッド位置 p の表し方

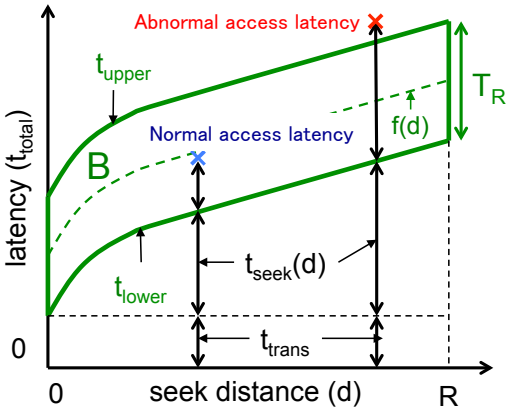


図 2: シークプロファイルによる不良セクタへのアクセスの検出

ドを実行した時の, 各アクセスのレイテンシを考える.

データ転送時間 t_{trans} は, 読み出しのサイズが一定のため, d に依存しない定数とみなすことができる.

各アクセスに対し横軸にシーク距離 d, 縦軸にレイテンシ t_{total} をプロットする (以降, このプロットをシークプロファイルと呼ぶ), 先述の議論を踏まえると, 正常なアクセスの点群は, 図 2 のように, t_{lower} および t_{upper} に囲われた帯状の領域 B に収まる. t_{upper} より上方に出ている点は, 不良セクタへのアクセスであるとみなせる. ただし, 不良セクタへのアクセスであっても t_{upper} を超えないことがあるため, 領域 B 内にあれば必ずしも正常なセクタへのアクセスであるとは限らない.

領域 B の範囲は, t_{lower} および t_{upper} の平均を $f(d)$ とおいて, 次のように定式化できる. ディスクの回転周期を T_R, 最外周トラックから最内周トラックへのヘッド移動距離を R とする. また, 係数 c_0, c_1 を導入して

$$0 \leq d \leq R,\ f(d) - \frac{T_R}{2} \leq t_{total} \leq f(d) + \frac{T_R}{2}$$

ただし

$$f(d) = \begin{cases} c_1\sqrt{d} + c_0 & \text{if } d \leq d_{th}. \\ l_1 d + l_0 & \text{if } d > d_{th}. \end{cases}$$

$$l_1 = \frac{c_1}{2\sqrt{d_{th}}},\ \ l_0 = c_1\sqrt{d_{th}} + c_0 - l_1 d_{th}$$

2.4 ブロックアドレスからヘッド位置への変換

磁気ディスクドライブ内のデータは，プログラムからはブロックアドレス a を用いてアクセスされるため，ヘッド位置 p を直接観測することができない．そのため，シークプロファイルを作成するにはアクセスしたブロックアドレス a をヘッド位置 p へ変換する手段が必要である．

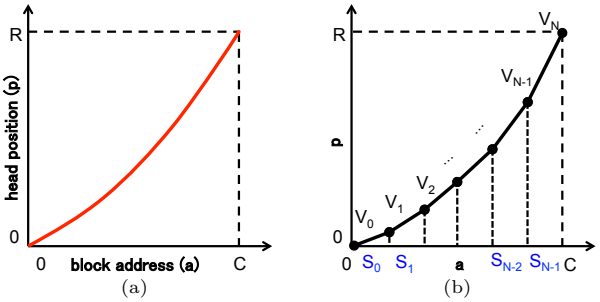


図 3: (a)ZBR 製品のブロックアドレスとヘッド位置の対応 (b) スループット計測による折れ線近似.

市場におけるほとんどの磁気ディスクドライブドライブ製品は，外周の半径が大きいトラックにより多いセクタを配置する Zone Bit Recording (以下, ZBR) と呼ばれる技術を採用しており， ブロックアドレスとヘッド位置の関係は線形ではない.

ZBR 方式の特性を言い換えると, 同じブロックアドレス間隔のシークであっても, 低位ブロックアドレス, すなわち外周トラックにおけるシークの方がヘッドの移動距離が小さい. 図 3(a) にグラフを用いた図解を示す. 赤の線に着目すると, 同じブロックアドレス a の差分に対してヘッド位置 p の変化が小さい. 高位ブロックアドレス (内周トラック) については，その逆の関係が成り立つ.

この曲線は, 図 3(b) のように, ディスクのアドレス空間全体を N 個の領域 $S_n (n = 0, 1, ..., N-1)$ に等分し, 各領域の中では, トラック当たりのセクタ数が一定, すなわちブロックアドレスとヘッド位置の関係が線形であると近似することができる. このような折れ線状のモデルは, ベンチマークを実行することで自動的に較正が可能である.

3. 磁気ディスクドライブ性能モデルの自動調整手法

この章では, 前章で説明した磁気ディスクドライブ性能モデルを異なる製品に対して自動的に較正する手法を説明する. 較正すべきパラメータは, ブロックアドレスとヘッド位置の対応関係, ディスクの回転周期 T_R と,$f(d)$ のパラメータ c_1, c_0, d_{th} である.

3.1 ブロックアドレスとヘッド位置の変換関数の較正

2.4 節で, ブロックアドレスとヘッドの位置の対応は線形でなく曲線状であることを述べ, 折れ線で近似する手法に触れた. 本節では, ベンチマークを実行してこの近似的なモデルを自動的に生成する手法を説明する.

S_n の 先頭アドレスから 20 秒間シーケンシャルリードを発行し, 平均スループット TP_n を計測する. 全てのトラックの幅が等しいという仮定の下で, トラック当たりのセクタ数と, 一定のブロックアドレス幅がまたぐトラックの距離は反比例する.

トラック当たりのセクタ数とスループットは比例し, また, S_n 内でトラック当たりのセクタ数が一定であるという近似的仮定から, TP_n と S_n がディスク内で占める幅は反比例の関係にある. S_n がディスク内で占める幅は, 全トラック幅 R を TP_n の逆数で比例配分することで決定すればよい.

定式化すると以下の通りである. 領域 S_n の先頭のブロックアドレスとヘッド位置は図 3(b) の点 $V_n = (v_{ia}, v_{ip})$ に対応する. ディスクの終端に対応する $V_n = (C, R)$ も含めると, $n = 0, 1, ..., N$ で

$$v_{na} = \frac{n}{N}C, \;\; v_{np} = \frac{\sum_{i=0}^{n} TP_i^{-1}}{\sum_{i=0}^{N} TP_i^{-1}}R$$

と表される. 与えられたブロックアドレス a に対し, ヘッド位置 p は, $v_{(n-1)a} \leq a < v_{na}$ を満たす n を選ぶことで,

$$p = (1-\alpha)v_{(n-1)p} + \alpha v_{np} \qquad (1)$$

と求められる. ただし

$$\alpha = \frac{a - v_{(n-1)a}}{C}N$$

3.2 回転周期 T_R の較正

ディスクの回転周期 T_R は通常, 各製品のデータシートに分間回転数として記載されているが, この値は必ずしも正確ではない. 実験中においても, 5400rpm モデルの異なる 2 製品が, 300rpm 程度異なる回転数を示していることを発見した.

モデルパラメータである T_R の自動較正ツールとして, たとえば以下のような簡易ベンチマークを考えることができる.

今回実行した測定方法は以下の通りで, T_R を 0.1ms の精度で測定できる.

- 1GB 離れた 2 つの 20MB の連続した領域に対して, 5000 往復シークするような 10000 回のセクタリードを実行する. (注1)
- 各リードに対してレイテンシを測定し, ヒストグラムを 0.1 秒単位で作成する. シーク距離とアクセスサイズを固定しているので T_R の長さの連続した階級にレイテンシが集中しているはずである.
- 50 以上の頻度をもつ最小の階級値と最大の階級値を求め, その差を T_R とする. (注2)

3.3 ランダムアクセスベンチマークによるモデルパラメータの調整

この節では, ランダムアクセスのベンチマーク実行結果を利用して製品固有の $f(d)$ のパラメータ c_1, c_0, d_{th} をフィッティングする手法について説明する.

まず, ディスクアクセス全体の範囲から乱数を用いて選ばれた N_{access} 個の 16KB 長の領域を順番にアクセスし, 各アクセスのレイテンシを記録することでシークプロファイルを作成する.

(注1) : オンディスクのキャッシュヒットが発生しないように, 同じブロックアドレスには 1 度までしかアクセスしない. また, ディスクのリードアヘッドを無効にする.

(注2) : 50 という閾値は一般的な回転数において適切に機能する. 外れ値がなければ, 5000rpm なら 12.0ms の連続した階級に平均 83.3 個, 10000rpm なら 6.0ms の連続した階級に平均 167 個のレイテンシが集中する.

表 1: 実験用サーバ諸元

Dell Precision T3620	
Processor	Intel(R) Xeon(R) CPU E3-1240
Memory	16GB DDR4
HDD conectivity	SATA 3.0
OS	Linux 4.14.101.el7.elrepo.x86_64

不良セクタの割合は正常セクタと比較して少数であるため, 実際に得られるシークプロファイルでは, 大多数の点は帯状のクラスタに集中する. このクラスタの形状に領域 B を自動的にフィットする, すなわち実際の分布における c_1, c_0, d_{th} を自動的に推定する簡易的なアルゴリズムを次のように考案した.

まず, 領域 B の関数形の閾値である d_{th} を固定する. シークプロファイルにある N_{access} 個の点のうち, $d \leq d_{th}$ の点群を S_{sqrt}, $d > d_{th}$ の点群を S_{linear} とする. S_{sqrt} は平方関数でフィットされ, S_{linear} は線形関数でフィットされる.

ここで, 次のような試行を n_{tr} 回繰り返し, 得られたスコアの最大値を記録する.

試行

- S_{sqrt} から $\alpha|S_{sqrt}|$ 個 (α は定数で $0 < \alpha < 1$) の点をランダムに選択し, これを S_{sample} とする.
- S_{sample} を用いて, $d \leq d_{th}$ における $f(d) = c_1\sqrt{d} + c_0$ を最小二乗法でフィットする. $x = \sqrt{d}$ と変数変換することで, 通常の線形回帰に帰着できる. この時, 領域 B の形が定まる.
- 全プロット (S_{sqrt} と S_{linear}) のうち, 領域 B に収まっているものの数をスコアとする.

固定した d_{th} に対する最大スコアが得られたら, d_{th} を 0 から R まで徐々に移動させながら同様にスコアを算出し, 最大のスコアが得られた d_{th} および, c_1, c_0 を解とする. その時, 領域 B に含まれていない点群は外れ値, すなわち不良セクタへのアクセスの結果と判定する.

このアルゴリズムの根拠は以下の通りである. 試行回数 n_{tr} が十分多ければ, S_{sample} として外れ値のほぼない組合せが選択される確率は十分に高い. その S_{sample} の元で c_1, c_0 のフィットを行うと, 観測データにおける非外れ値のクラスタの形に近い領域 B が生成される. 非外れ値クラスタの点密度は高いため, 高いスコアが得られ, 解に適用される.

試行の結果を次の S_{sample} の選び方に適用するような高度な方法も考えられるが, 今回の実験では以上の単純な手法のみを用いた.

4. 実　　験

4.1 実験環境

実験は表 1 に示すサーバ 1 台を用いて行った.

4.2 シークプロファイルの作成

3 種類の磁気ディスクドライブ製品 1 台ずつに対して, 3.3

表 2: 実験に利用した磁気ディスクドライブ製品

ベンダ	モデル	容量	回転数 (仕様)
Seagate	ST6000DM003	6TB	5400 rpm
Seagate	ST6000VN0033	6TB	7200 rpm
Western Digital	WD60EZRZ	6TB	5400 rpm

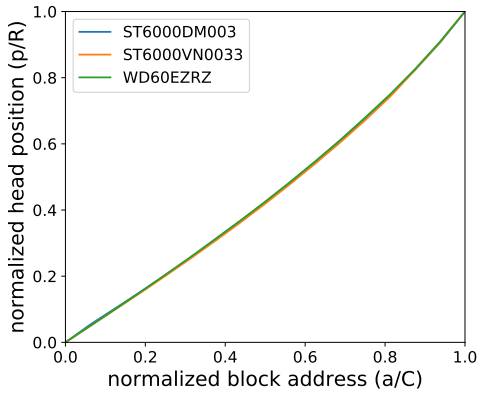


図 4: 測定で得られたブロックアドレスとヘッド位置の対応. ブロックアドレスおよびヘッド位置は 0〜1 の範囲にスケーリングされている. (すなわち容量を C, 最外周〜最内周トラックの距離を R として, 横軸は a/C, 縦軸は p/R を表す.)

節で説明したランダムアクセスのワークロードの実行, およびシークプロファイルの作成を行った. 利用した製品のモデル名, 容量等は表 2 に示す通りである.

ワークロードは, ディスク全体のアドレス空間から乱数を用いて選択された N_{access} 個の各 16KB 長のアドレス領域に対して, 順番にシングルプロセスから read() システムコールを発行するものである. なお, 実験中はシステムキャッシュを無効化した.

レイテンシは, システムコールの組 lseek(), read() の実行にかかる時間として測定した.

試行は, 各製品モデルに対し, $N_{access} = 2^7, 2^8, ..., 2^{16}$ とサンプル数を変化させながら, それぞれ 5 回ずつ行った. 5 回の試行のアクセスパターンは, 乱数のシードを変化させることで全て異なるものを生成した. 試行は互いに干渉しないよう, 異なるドライブに対する試行に対しても同時には実行しなかった.

4.3 ブロックアドレスとヘッド位置の変換

3 台のディスクにおけるブロックアドレスとヘッド位置の関係を推定するため, 2.4 節で示した手法を $N = 16$ で実行した. すなわち, アドレス空間を 16 等分したのち, 各領域の先頭から 20 秒間シーケンシャルリードを実行し, 平均スループットを測定した. 式 1 を適用し, ブロックアドレスとヘッド位置の近似関係を得た. ただし, 物理的な距離 R は推定不可のため, 全製品に対して $R = 1$ と設定した. グラフ化したものを図 4 に示す. いずれも, 低位ブロックアドレスでは傾きが小さく, 高位ブロックアドレスでは傾きが大きいことから, ZBR の特徴が見て取れる. 3 製品に対するプロットはほぼ互いに重なっている.

表 3: ベンチマークにより測定された T_R および回転数

モデル	回転周期 T_R	回転数 (実測)
ST6000DM003	10.9 ms	5.50×10^3 rpm
ST6000VN0033	8.3 ms	7.23×10^3 rpm
WD60EZRZ	10.3 ms	5.83×10^3 rpm

表 4: フィッティングにより得られた c_1, c_0, d_{th} の値

モデル	c_1	c_0	d_{th}	# of outliers
ST6000DM003	16.8	8.47	0.94	1762
ST6000VN0033	22.1	4.98	0.64	4458
WD60EZRZ	15.9	7.86	0.30	1174

4.4 フィッティング

3.3 節で示したフィッティング手法を実装し,[注3] 4.2 節で得たシークプロファイル群に対して適用した.

まず, 3.2 節に示した手法でディスクの回転数を測定し, 表 3 の結果を得た. これを用いて, フィッティングを行った.

$\alpha = 0.125, N_{tr} = 100$ として, d_{th} は $0 \leq d_{th} \leq 1$ の範囲で 0.02 ずつ変化させて最高スコアを計算した. 今回試した最大の $N_{access} = 2^{16}$ のプロファイルに対しては, シングルプロセスで 5 秒程度で実行が終了した.

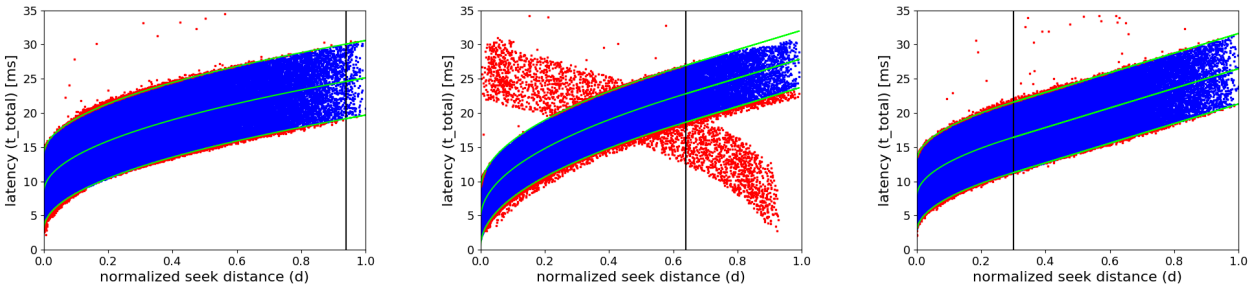
$N_{access} = 2^{16}$ における 3 製品の, 1 度目の試行で得られたシークプロファイルとフィット結果を, 図 5 に示す (読み方はキャプションを参照のこと). また, 得られた c_1, c_0, d_{th} の値を表 4 に示す. # of outliers は, 領域 B 外のプロットの数を示す.

3 つのプロットで, ほとんどの点群が領域 B の形状をした帯状の領域に密なクラスタを形成しており, 周囲には疎に分布していることが読みとれる. いずれにおいてもフィッティングした領域 B が, クラスタの形とおよそ一致していることが読み取れ, 3. 章で考察したモデルの正当性を示している. ただし, T6000VN0033 においては, 最も密なクラスタの他に, それを左右反転したようなやや薄いクラスタが観測され, 3. 章で示していない機構が背後にあると考えられる.

次に, N_{access} と, 出力されるモデルの安定性の関係を検証した. 3 製品で作成したシークプロファイルに対して全てフィッティングを行い, $d = 0.1, 0.2, ..., 1.0$ に対する $f(d)$ の値を記録した. ST6000DM003 に対して, N_{access} を変化させた時のこれらの値の推移を図 6 に示す. N_{access} の増加に従い, 各値の平均値の上下は徐々に縮まり, $N_{access} = 2^{16}$ までにはほぼ変化が無くなっている. 標準偏差も, N_{access} の増加に従い減少している. これは, サンプルの増加に従い, 試行ごとのクラスタ内での点が均一に近づくことで, 生成されるモデルが安定しているためだと考えられる. 他の 2 製品に関しても, $N_{access} = 2^{16}$ までに値が安定する傾向が見られた.

$N_{access} = 2^{16}$ の試行 1 回にかかる時間は, およそ 1000 秒から 1150 秒である. 高々この時間をかければ, 3 モデルに対しては, ディスクの性能モデルとしておよそ安定した値を推定することができる.

(注3) : 線形回帰には Python ライブラリ scikit-learn の LinearRegression クラスを利用した.



(a) ST6000DM003　(b) ST6000VN0033　(c) WD60EZRZ

図 5: $N_{access} = 2^{16} = 65536$ における 3 製品のシークプロファイルとフィット結果.

3 本の緑色の曲線のうち, 中央にあるものが $f(d)$, 上下にあるものがそれぞれ $f(d) + \frac{T_R}{2}, f(d) - \frac{T_R}{2}$ で, この 2 つがモデルが生成した領域 B の上限値, 下限値を示す. 領域 B の内側にある点を青色, 外側にある点を赤色に着色している. 関数形の境界 $d = d_{th}$ を黒い縦線で示す.

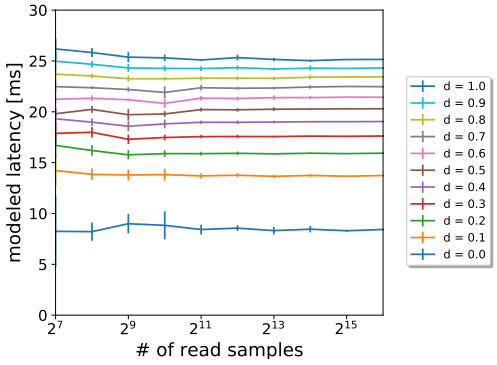


図 6: N_{access} と, $d = 0.1, 0.2, ..., 1.0$ のときのモデル出力 $f(d)$ の関係. 各プロットは 5 回の試行に対する平均および標準偏差を示している.

5. 関 連 研 究

磁気ディスクドライブの性能モデルを考察し, 性能シミュレータを開発する研究は 90 年代より数多く試みられている. 特定製品に対するモデルの構築の例として [6], [7], パラメータ可変のシミュレータの例として [8], またパラメータを自動推定する研究として [9], [10] などが挙げられる. トラック及びセクタの配置パターンを推定の対象とする研究例 [11] もあり, 近年では, トラックの高密度化のために普及している瓦記録 (SMR) 技術と呼ばれる記録方式を対象とした詳細な性能モデルの構築が研究されている [12]〜[14].

6. 結　　論

本論文では，磁気ディスクドライブに対するランダムアクセスを対象として，入出力ベンチマークを用いてレイテンシモデルを自動調整する手法を提案した．提案した手法を 3 つの異なる磁気ディスクドライブ製品に適用した結果，いずれの製品においても，自動調整したレイテンシモデルによって，正常なアクセスレイテンシの範囲を正確に予測可能であることを確認した．また，1 つのモデルに対して 1,000 秒から 1,150 秒の入出力ベンチマークを実効することで，安定したモデルの調整結果を得られることを確認した．

文　　献

[1] K. Goda and M. Kitsuregawa, “The history of storage systems,” Proceedings of the IEEE, vol.100, no.Special Centennial Issue, pp.1433–1440, May 2012.
[2] R.E. Fontana, “Ten year storage technology landscape for hdd, nand, and tape,” MSST 2018 (invited talk), 2018.
[3] M. Hao, G. Soundararajan, D. Kenchammana-Hosekote, A.A. Chien, and H.S. Gunawi, “The tail at store: A revelation from millions of hours of disk and SSD deployments,” 14th USENIX Conference on File and Storage Technologies (FAST 16), pp.263–276, USENIX Association, Santa Clara, CA, 2016.
[4] M. Hao, G. Soundararajan, D.R. Kenchammana-Hosekote, A.A. Chien, and H.S. Gunawi, “The tail at store: A revelation from millions of hours of disk and ssd deployments.,” FAST, pp.263–276, 2016.
[5] Q. Wang, C.-A. Lai, Y. Kanemasa, S. Zhang, and C. Pu, “A study of long-tail latency in n-tier systems: Rpc vs. asynchronous invocations,” 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), pp.207–217, 2017.
[6] D. Kotz, S.B. Toh, and S. Radhakrishnan, “A detailed simulation model of the hp 97560 disk drive,” 1994.
[7] C. Ruemmler and J. Wilkes, “An introduction to disk drive modeling,” Computer, vol.27, pp.17–28, 1994.
[8] G.R. Ganger, B.L. Worthington, and Y.N. Patt, “The disksim simulation environment version 2.0 reference manual,” 1999.
[9] J. Schindler and G.R. Ganger, Automated disk drive characterization (poster session), vol.28, ACM, 2000.
[10] N. Talagala, R.H. Arpaci-Dusseau, and D.A. Patterson, Micro-benchmark based extraction of local and global disk characteristics, Computer Science Division, University of California, 1999.
[11] J. Gim and Y. Won, “Extract and infer quickly: Obtaining sector geometry of modern hard disk drives,” TOS, vol.6, pp.6:1–6:26, 2010.
[12] A. Aghayev, M. Shafaei, and P. Desnoyers, “Skylight—a window on shingled disk operation,” ACM Trans. Storage, vol.11, no.4, pp.16:1–16:28, Oct. 2015. http://doi.acm.org/10.1145/2821511
[13] R. Pitchumani, A. Hospodor, A. Amer, Y. Kang, E.L. Miller, and D.D. Long, “Emulating a shingled write disk,” Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2012 IEEE 20th International Symposium onIEEE, pp.339–346 2012.
[14] M. Shafaei, M.H. Hajkazemi, P. Desnoyers, and A. Aghayev, “Modeling drive-managed smr performance,” ACM Trans. Storage, vol.13, no.4, pp.38:1–38:22, Dec. 2017.