2L-7

ホスティングサービスにおけるアクセス傾向予測を用いた分散ストレージシステムの省電力化手法

大越 淳平 † 長谷部 浩二 ‡ 加藤 和彦 ‡

筑波大学大学院システム情報工学研究科 † 筑波大学システム情報系情報工学域 ‡

1 序論

近年，Youtube[*1]やFlickr[*2]等のホスティングサービスに代表される大規模なインターネットサービスにおいて，データセンターの運用コストの削減が大きな課題となっている．そこで本研究では，ホスティングサービスを提供するための分散ストレージシステムにおける省電力化手法を提案する．ホスティングサービスの特徴として，大量のデータの逐次的な追加やアクセス傾向の急激な変化が挙げられる．本提案手法では，データの経過時間とアクセス数の二つから予測される将来のアクセス傾向に基づきデータをディスク上に配置し，アクセス頻度の高いデータを一部のディスクに集約する．これにより，アクセスの少ないディスクをスピンダウンさせ，システム全体の消費電力を削減する．本研究では，消費電力と応答時間をFlickr上の公開写真ファイルのアクセスパターンを用いたシミュレーションと実装により評価し，提案手法の有用性を示す．

2 システム構成

提案するシステムの構成を図1に示す．提案するシステムは数千台のディスクで構成され，それらの各ディスクはディスク群A，ディスク群Bおよびディスク群Cの3つのディスク群のいずれかに分類される．初期の状態では，ディス群Aに数台のディスク（ディスク台数はデータの流入量により決定される），ディスク群Bに0台のディスク，ディスク群Cに残りのすべてのディスクが格納されている．

クライアントからのデータの書き込みはディスク群Aのいずれかのディスクに対して行われ，空き容量のなくなったディスクはディスク群Bに移動する（図中における（1））．また，ディスクの移動に応じて新たな空のディスクがディスク群Cより追加され，ディスク台数は一定に保たれる．

ディスク群Aより移動したディスクはディスク群Bを構成し，このディスク群においてディスクは論理的に矩形状に配置される．ここで，i行j列のディスクを$D(i,j)$と表記する．新たなディスクがディスク群Aより追加された場合，そのディスクは$D(1,1)$となり，1行目に存在したディスクは列番号が1増加する．また，jの最大値は（例えば，ディスク群Bのディスク台数をNとした場合，$j=\lfloor N \rfloor$などと）あらかじめ定められており，その値を超えた場合，その行において一番大きな列番号を有するディスクは，それより小さい列番号を有するディスクとアクセス頻度（アクセス頻度の詳細については，3章で述べる）に着目したデータの交換を行い，アクセス頻度の小さいデータを集約する（図中における（2））．その後，このディスクは次の行の先頭に移動する（図中における（3））．以上の一連の手続きにより，各ディスクにデータがアクセス頻度順に格納される（この場合，iやjが小さいほどアクセス頻度の高いデータが集約される）．また，ディスク群Bに所属するディスクは，一定時間（例えば，60秒間など）アクセスがなければスピンダウンし，消費電力の削減を行う．しかしながら，iの小さい行に関しては，応答性の悪化を防ぐため常時稼動させる．

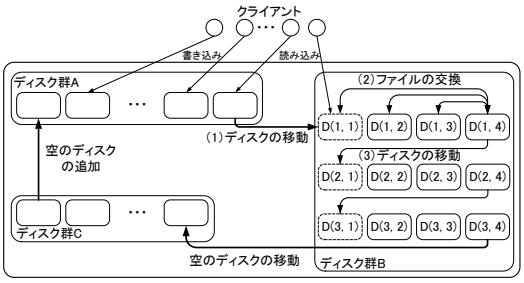

図1 システム構成

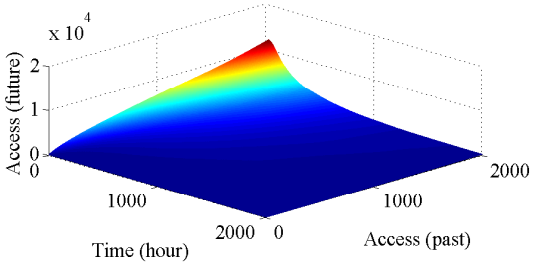

図2 アクセス傾向

3 Flickrにおける写真データの分析

図2は，Flickrにアップロードされた写真データのアップロードされてからの経過時間（Time（hour）），それまでのアクセス回数（Access（past））およびそれ以降の1,000時間のアクセス回数（Access（future））の関係を示している．この図は，生のデータを平滑化し，さらに各時間ごとに累乗関数で近似することにより得ている．この図より，アップロードされてからの経過時間とそれまでのアクセス回数より，それ以降のアクセス傾向を予測することで，ディスク群Bにおける効率的なデータの移動を実現する．

この分析には，Flickrにアップロードされた46,106枚の写真データのアクセスパターンを用いている．アクセスパターンは，Flickrの Web APIを用いて1時間ごとのアクセス回数を計測することにより収集した．本研究におけるシミュレーションと実装のワークロードにおいても，このアクセスパター

Power-Saving using Access Prediction for Distributed Storage Systems in Internet Hosting Services

†Graduate School of Systems and Information Engineering, University of Tsukuba

‡Division of Information Engineering, Faculty of Engineering, Information and Systems, University of Tsukuba

*1 http:/www.youtube.com/

*2 http:/www.flickr.com/

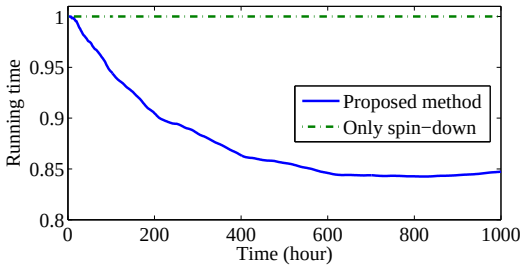

図 3　稼働時間

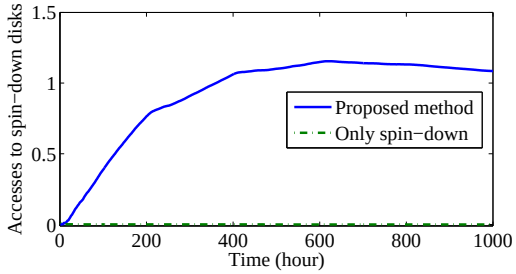

図 4　スピンダウンしているディスクへのアクセス

ンを用いている．

4　シミュレーション

本研究におけるシミュレーションでは，稼働時間とスピンダウンしているディスクへのアクセス数を計測することで提案手法の省電力性と応答性を評価した．なお，本研究では稼働時間と消費電力がほぼ比例すると仮定している．

シミュレーションで想定する環境を次に述べる．ディスク群 A のディスクは数台とし（なお，本シミュレーションではディスク群 A のディスクの稼働時間は考慮しない），ディスク群 B は初期状態で 0 台，ディスク群 C は最大で約 1,000 台とした．各ディスクの容量は 500 GB であり，スピンアップに 5 秒を要するものとする．転送速度は，ディスク内部，ネットワークに共通して 100 MB/sec とした．また，ディスク群 B の各ディスクはアクセスが終了してから 30 秒でスピンダウンし，全ディスクの 10% は応答性の悪化を防ぐため常時稼動させる．

図 3 にディスクの稼働時間の推移を示す．これは常時稼動させた場合を 1 とした相対的な値を示しており，1,000 時間経過後には提案手法により 15.3% の稼働時間が削減されていることを示している（比較のためにデータの交換を行わず，スピンダウンのみを行う設定を記載した．この設定では，稼働時間の削減率は 0.0% である）．

図 4 に全アクセスに占めるスピンダウンしているディスクへのアクセス数の比率の推移を示す．この図より，1,000 時間経過後には，提案手法において 1.0% のアクセスがスピンダウンしているディスクへのアクセスであることが観察される．

5　実装

本研究では，クラスターマシン 10 台を用いて実装を行い，応答時間を計測することで提案手法の応答性を評価した．実験環境において，各サーバのディスク容量は 36 GB であるが，同一のファイルに対して読み書きを行うことで仮想的に 500 GB の容量があるものとした．また，実験環境におけるネットワーク帯域の制約により，応答時間はディスクからメモリ中にファイルが展開されるまでの時間を計測した．加えて，現在のシステム構成では実際にディスクをスピンダウンさせることが困難であるため（各サーバのディスクは 1 台であり，OS がインストールされている），スピンダウンはアクセスがなされた際にプログラム上で遅延を生じさせることにより実現した．

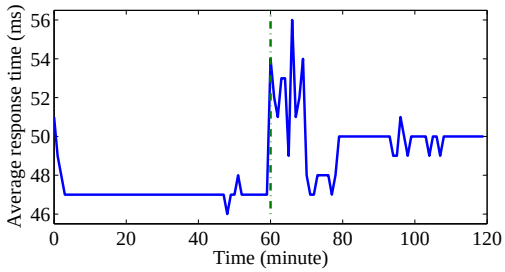

図 5　応答時間

図 5 は，200 時間経過後のディスク群 B において，最もアクセス頻度の低い 10 台の平均応答時間の推移を示している．図中の点線はデータの移動を開始した時点を示している．この図より，全アクセスにおける平均応答時間は 48.6 ms であったことが観察される．また，データの移動開始前の平均応答時間は 47.1 ms，データの移動開始後の平均応答時間は 50.0 ms であり，データの移動により応答性のわずかな悪化が観察された．

6　関連研究

ストレージシステムの省電力化は，過去の研究により多くの手法が提案されていおり，データのアクセス頻度に着目した MAID [1] や PDC [5]，RAID 環境における PARAID [6] などが挙げられる．

近年では，大規模な分散ストレージにおける手法も提案されている [2, 3]．本研究は [4] を拡張したものであり，ホスティングサービスに代表されるデータが逐次的に追加され，アクセス頻度が大きく変化するストレージシステムを対象とした省電力化手法を提案している．

7　結論と今後の課題

本研究では，Flickr や Youtube に代表されるホスティングサービスのストーレジシステムを対象とした省電力化手法を提案した．評価により，ある程度の応答性の悪化は観察されたものの，15.3% の稼働時間の削減が観察された．今後の課題として，分散ハッシュテーブルなどを組み込んだ，より現実に近い環境での評価が挙げられる．

参考文献

[1] Dennis Colarelli and Dirk Grunwald. Massive arrays of idle disks for storage archives. Supercomputing'02, pp. 1–11, 2002.
[2] Danny Harnik, Dalit Naor, and Itai Segall. Low power mode in cloud storage systems. IPDPS'09, pp. 1–8, 2009.
[3] Rini T. Kaushik and Milind Bhandarkar. GreenHDFS: towards an energy-conserving, storage-efficient, hybrid Hadoop compute cluster. HotPower'10, pp. 1–9, 2010.
[4] Jumpei Okoshi, Koji Hasebe, and Kazuhiko Kato. Power-Aware Autonomous Distributed Storage Systems for Internet Hosting Service Platforms. CloudComp'12, 10 pages, 2012.
[5] Eduardo Pinheiro and Ricardo Bianchini. Energy conservation techniques for disk array-based servers. ICS'04, pp. 68–78, 2004.
[6] Charles Weddle, Mathew Oldham, Jin Qian, An-I Andy Wang, Peter Reiher, and Geoff Kuenning. PARAID: A gear-shifting power-aware RAID. *TOS*, Vol. 3, No. 3, 2007.