3J-5

ディスクの自律制御による大規模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‡

筑波大学情報学群情報科学類† 筑波大学システム情報系情報工学域‡

1. はじめに

近年，クラウドコンピューティングの普及に伴い，データセンターでは，追加されるデータの増大とストレージシステムにおける消費電力の増加が問題となっている．そこで本研究では，大規模な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．特にここでは，写真や動画を対象としたストレージサービスのように大量のデータが追加される環境を想定し，高いスケーラビリティを有するシステムの構築を目的とする．本提案手法における基本的なアイデアは，各ディスクの自律制御により，アクセス頻度に着目したデータの交換を他のディスクと行い，一部のディスクにアクセス頻度の高いデータを集約するというものである．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，消費電力を削減する．また本研究では，稼働時間と応答時間をシミュレーションと実装を用いて評価し，提案手法の有用性を示す．

2. システム構成

本研究では，数千台の計算機がネットワークで相互に接続されている環境を想定しており，各計算機に接続されたディスクは，OS によりスピンダウンやスピンアップを行うものとする．

本研究で提案するシステムは，ディスク群 A とディスク群 B により構成される．追加されるデータはディスク群 A のディスクに書き込まれる．また，ディスク群 A におけるディスクの台数は，追加されるデータの容量によって決定され，複数台の場合はハッシュ値等を用いて負荷分散が行われる．一定容量まで達したディスクは，ディスク群 B に移動し，新たな空のディスクがディスク群 A に追加される．ディスク群 B では，一定時間ごとにファイルの交換がディスク間で行われ，一部のディスクにアクセスが集約される．そして，アクセスの少ないディスクをスピンダウンさせることにより，消費電力を削減する．ファイルの交換では，（1）交換相手の選定，（2）負荷の計算，（3）ファイルの交換，の一連の処理が行われる．また，ファイルの位置管理は，全ディスクにおける分散管理により行われる．

3. シミュレーション

本研究では，シミュレーションにおけるワークロードを決定するため，Flickr における写真データのアクセス傾向およびファイルサイズの分析を行った．アクセス傾向の分析では，20000 ファイルのアクセス回数を 2 週間に渡り計測した．その結果を基に，シミュレーションで各写真データに与える ID 番号i（$1 \leq i \leq 20000$），経過時間x（時間）およびアクセス回数yの間には，

$$y = 55.209 \cdot \left(641 \cdot \frac{i}{20000 \cdot 0.04}\right)^{-0.65} \cdot x^{-0.845}$$

の関係式が成立するものと仮定する．ただし，これは$1 \leq i \leq 20000 \cdot 0.04$のファイルにのみ成立し，他のファイルのアクセスは常に 0 とする．

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 は 1 台とし，ディスク群 B は，初期は 0 台，最大で 1000 台とする．各ディスクは，1TB まで記録可能であり，スピンアップに 10 秒を要し（[1], [3]），最後のアクセスから 60 秒後にスピンダウンを行う．転送速度は，ディスクの内部，ネットワークに共通して 100 MB/sec とした．使用する写真データは，3000 枚/min で追加され，サイズは 3 MB とする．

シミュレーションでは，720 時間におけるディスクの稼働時間を，交換相手の選定における制限の有無（有りの場合は近傍 10 台）と交換回数の制限の有無（有りの場合は各ファイル 5 回以内）により，4 つの場合で計測した．シミュレーションの結果は図 1 であり，交換相手の制限を行い交換回数には制限を設けない場合が最も稼働時間が短かった．また，720 時間における稼働時間の合計は，交換を行わない場合を 1 とすると，短い順から 0.11，0.14，0.17，0.34 であり，すべての場合で 60%以上の電力が削減できることがわかった．

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.

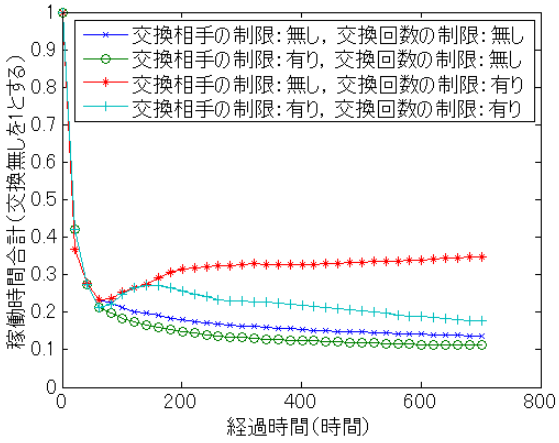

図 1 シミュレーション結果

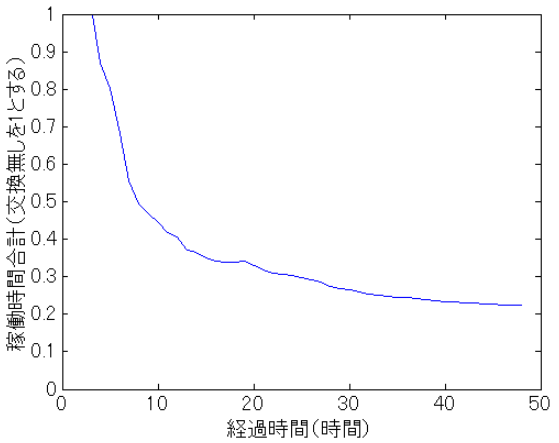

図 2 実装結果

4. 実装

本研究では，クラスターマシンを用いて実装を行い，稼働時間と応答時間を測定した．実験環境において，ファイルの追加はシミュレーション同様に行われるが，ディスクの容量が 500 GB である点と，ディスク群 B の最大が 50 台である点がシミュレーションとは異なる．また，ネットワーク帯域の制約により，応答時間は，要求が与えられてからデータがディスクからメモリ上に展開されるまでとする．加えて，実際のディスクは 36 GB であり，異なるファイルであっても同一のファイルへのアクセスをすることで，仮想的に 500 GB のディスクが存在するものとする．さらに，現在のハードウェア構成の制約から，実際にスピンアップやスピンダウンは行わず，最後にアクセスした時刻を記録し，その経過からディスクの状態を判断する．

実装結果は図 2 であり，48 時間における稼働時間の合計は，交換を行わない場合を 1 とすると，交換を行った場合は 0.23 であった．また，応答時間の平均値は，交換を行わない場合が 747 msec，交換を行った場合が 192 msec であった．従って，交換を行うことで応答性が向上し，77%の電力が削減された．

5. 関連研究

ストレージシステムの省電力化については，過去に多くの研究がなされており，データのアクセス頻度に着目した MAID[1]や PDC[2]，冗長化したデータに着目した DIV[3]，RAID 環境における EERAID[5]などがある．

近年では，大規模な分散ストレージにおける省電力化手法も提案されている[5]．本研究は，大規模な分散ストレージを想定することに加え，これまで十分に考慮されていなかった大量のデータの流入と各データのアクセス頻度の変化までを想定する点で他の研究と異なり，近年注目を集めているサービスにより近い環境を想定している．

6. 結論と今後の課題

本研究では，大規模な分散ストレージシステムにおける省電力化手法の提案を行った．シミュレーションと実装では，応答性を悪化されることなく，稼働時間が短縮されることを示した．ファイルの位置管理について具体的な言及は行わなかったが，DHT を用いて実現可能であると考えている．また，他の課題として耐障害性を考慮する必要があり，現在その実現方法について検討している．

参考文献

[1] D. Colarelli and D. Grunwald. Massive array of idle disks for storage archives. *SC'02,* pp.1-11, 2002.

[2] E. Pinheiro and R. Bianchini. Energy conservation techniques for disk array-based servers. *ICS'04,* pp.68-78, 2004.

[3] E. Pinheiro, R. Bianchini and C. Dubnicki. Exploiting redundancy to conserve energy in storage systems. *SIGMETRICS/Performance'06,* pp.15-26, 2006.

[4] D. Harnik, D. Naor and I. Segall. Low power mode in cloud storage systems. *IPDPS'09,* pp.1-8, 2009.

[5] D. Li and J. Wang. EERAID: energy efficient redundant and inexpensive disk array. *SIGOPSEW'04,* 6 pages, 2004.

* Autonomous Control of Disks for Power-Saving in Large-Scale Distributed Storage Systems
† College of Information Science, School of Informatics, University of Tsukuba
‡ Division of Information Engineering, Faculty of Engineering, Information and Systems, University of Tsukuba

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.