5B-5

ストレージ省電力手法RAPoSDAのプロトタイプシステムの試作および評価

引田 諭之† 小栗 寛生† 横田 治夫†

†東京工業大学大学院情報理工学研究科計算工学専攻

1 はじめに

日々増大するデジタルデータを保存するために，企業やデータセンター等におけるストレージシステムは益々大規模化しており，その省電力化は重要な課題である．ストレージシステムの省電力化に関してはこれまでに様々な研究がなされてきており，中でもアクセスに局所性のある大規模ストレージシステムの省電力化手法ではMAID[1]がよく知られている．MAIDはディスクをキャッシュディスクとデータディスクに分け，アクセス頻度の高いデータをキャッシュディスクに格納することでデータディスクへのアクセス頻度を低下させ，一定時間以上アイドル状態が続いたデータディスクをスピンダウンさせることでストレージシステム全体の省電力化を実現する．しかしMAIDは信頼性の確保に関しては特に考慮していないことや，固定数のキャッシュディスクが性能のボトルネックに陥る恐れがある等の問題も存在する．

一方，我々はこれまでにストレージ省電力化手法であるRAPoSDA(Replica Assisted Power Saving Disk Array)[2]を提案してきた．RAPoSDAでは信頼性および可用性を確保するためにデータをプライマリ・バックアップ構成で二重化し，個々のディスクドライブの回転状況を考慮したアクセス制御により省電力化を実現している．これまでにシミュレーションプログラムを用いた評価実験によりRAPoSDAの有効性を検証してきたが，実環境では未検証であった．

本報告では実際の計算機上にストレージシステムの消費電力および性能を評価するための実験基盤を構築し，その基盤システム上に我々の提案するRAPoSDAを実装し，実機上での省電力効果および性能に関してその有効性を検証する．

2 RAPoSDAの概要

2.1 構成

RAPoSDA[2]は主記憶上のバッファとディスクドライブから構成され，ディスクドライブは更に少数のキャッシュディスクと多数のデータディスクに分けられる．バッファは個別の電源系統に接続された複数の主記憶装置からなり，UPS等で断電対策されているものとする．RAPoSDAでは個々の電源系統に接続されているバッファを一つの単位として扱う．また，キャッシュディスクは常に回転しており，クライアントからの読み出し要求に対するキャッシュとして働く．データディスクは一定時間アクセスがなければ回転を停止し，それによりストレージシステム全体での省電力化を実現する．

また，バッファとデータディスクはプライマリ・バックアップ構成によりデータを二重化している．

2.2 readおよびwrite動作

2.2.1 read

readはバッファ，キャッシュディスク，データディスクの順番にアクセスし，対象データが見つかった時点でデータを読み出す．データディスクにアクセスする際は，対象データはプライマリディスクとバックアップディスクの両方に存在しているので，其々のディスクが回転中かどうかを確認し，回転中のディスクからデータを読み出す．両方とも回転中の場合，バッファ上のキューが長い方のディスクから読み出す．両方とも停止中だった場合は，回転停止期間の長い方のディスクからデータ読み出す．

2.2.2 write

writeデータがバッファに書き込まれる際は，書き込み対象バッファのプライマリ領域と，別バッファのバックアップ領域に書き込まれる．もしバッファに設定されている容量閾値を超えてしまう場合は，データディスクに書き込みを行う．該当データディスクが回転中の場合，そのままデータを書き込むが，停止中であればデータディスクをスピンアップさせた後で書き込む．また，データディスクに書き込む際は，そのデータディスクのプライマリ層データとバックアップ層データの両方を書き込む．

3 プロトタイプシステムの概要

3.1 プロトタイプシステムの構成

本節では実機上に試作したストレージシステムについて説明する．プロトタイプシステムを実装した計算機ではOSにLinux Kernel 2.6.32-5-amd64を用い，主にJava言語によってKey-Value Storeとして実装している．但しファイルシステムが提供するキャッシュ機構を回避するためディスクアクセス処理の部分はC++言語で書かれたネイティブコードのモジュールをJNI経由で使用している．

また，本来であればRAPoSDAのバッファは別々の電源系統に接続された個々のメモリから構成されるのであるが，市販の計算機上ではそのような構成を採用するのが困難であることと，バッファが個別の電源系

An Evaluation of Storage Power Reduction Method RAPoSDA based on a Prototype System

Satoshi HIKIDA† Hiroki OGURI† Haruo YOKOTA†

†Dept. of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

†hikida@de.cs.titech.ac.jp †oguri.h.aa@m.titech.ac.jp

†yokota@cs.titech.ac.jp

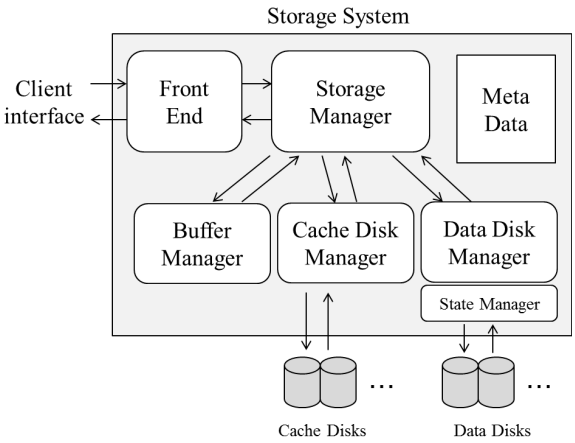

図 1: プロトタイプシステムの構成図

統であるかどうかは動作検証を行う上で影響がほとんど無いため，プロトタイプシステムでは物理的には一つの電源系統を共有する，論理的に分離されている複数のバッファを用いている．

図 1 は今回試作したプロトタイプシステムの構成図である．プロトタイプシステムは 6 つのモジュールから構成されている．以下では各モジュールの役割について概要を述べる．

Front End このモジュールは外部のクライアントとのインターフェースを担当する．今回試作したプロトタイプシステムは Key-Value Store のインターフェースを持ち，Key は任意の文字列で，Value はバイト配列を前提としている．

Storage Manager メタデータの管理および read や write 動作を統括するモジュールである．リクエストを受け付けたら，メタデータを参照し Key に対する値がどのバッファ（もしくはどのディスク）にマッピングされているかを調べ，適切なデバイスにリクエストを転送し，その結果を Front End に返す．write 時にはどのデータをどのバッファ（もしくはどのディスク）に割り当てるのかを決定するのもこのモジュールの役割である．

Buffer Manager 個々のバッファを統括管理するモジュールである．Storage Manager からの要求に応じて各バッファに対する read/write そのた必要な操作を行う．

Cache Disk Manager キャッシュディスクを管理するモジュールである．Storage Manager からの要求に応じて各キャッシュディスクに対する read/write を行う．

Data Disk Manager データディスクを管理するモジュールである．Storage Manager からの要求に応じて各データディスクに対する read/write を行う．また State Manager と連携して個々のディスクの回転状況も管理し，ディスクの Spin-up/Spin-down の制御も行う．

State Manager 個々のデータディスクそれぞれの回転状態を管理する．データディスクの状態の変化を監視しており，Spin-up/Spin-down の実行に必要なアイドル状態期間やその時点でのディスク状態を Data Disk Manager に提供する．

3.2 プロトタイプシステムの動作

現段階では外部との連携は，アプリケーションに直接組み込みプロトタイプシステムが提供する API を直接呼び出すか，ソケットベースで通信を行うかのいずれかを想定しており，今回試作したプロトタイプシステムでは，Key-Value Store として簡易な read/write の API を持つように設計している．

データを write するには Key として任意の文字列を与え，Value には書き込むデータのバイト配列を与える．read するときは Key を与え，もしストレージに該当データが存在すればそのバイト配列が返される．

```
// read
String key = “lookingfor”
byte[] value = null;
value = storage.read(key);

// write
String key = “tostore”
byte[] value = getOrGenerateValue();
boolean result = storage.write(key, value);
```

図 2: read および write の使用例

図 2 は Java 言語による read 時および write 時のコード例を示している．

4 まとめおよび今後の課題

本報告では我々が提案するストレージ省電力化手法 RAPoSDA を実機上に試作したプロトタイプシステムについて概要を述べた．

今後の予定としては，試作したプロトタイプシステムを用いて実機環境における RAPoSDA の省電力効果および性能に関して評価を行い，RAPoSDA の有効性を実証していく予定である．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金基盤研究 (A)(# 22240005, #25240014) の助成により行われた．

参考文献

[1] Dennis Colarelli and Dirk Grunwald. Massive arrays of idle disks for storage archives. In *Supercomputing '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, pp. 1–11, Los Alamitos, CA, USA, 2002. IEEE Computer Society Press.

[2] Satoshi Hikida, Hieu Hanh Le, and Haruo Yokota. A power saving storage method that considers individual disk rotation. In *The 17th International Conference on Database Systems for Advanced Applications (DASFAA)*, Vol. 7239/2010, pp. 138–149, April 2012.