階層的データ管理手法を用いた 大規模省電力ストレージ構築方式の提案

西川 記史† 中野 美由紀† 喜連川 優†

†東京大学生産技術研究所

E-mail: †{norifumi, miyuki, kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし データセンタの消費電力は増加の一途を辿っている．特に，データ量の急増に伴うストレージの消費電力の増加は著しく，ストレージの消費電力の削減はデータセンタの省電力化を行う上で最重要の課題である．近年，データセンタでは増大し続けるデータの管理コストの低減を目的に，ストレージを階層化しデータをその要件に適した階層に配置する階層データ管理が着目されている．これまでデータセンタの省電力化のために様々なアプローチが提案されているが，階層的なデータ管理をストレージの省電力化に適用する手法は提案されていない．そこで我々は，ストレージの省電力化のための階層的なデータ管理と階層的なデータ管理手法を用いた省電力ストレージの構築手法を提案する．我々の提案手法は，まずデータに求められるアクセス性能に基づきデータの管理階層を定める．そして各階層に求められる性能や消費電力を提供するストレージ階層を構築し，データを適切なストレージ階層に配置する．本論文では，提案手法の処理方式について述べるとともに運用中のデータセンタを具体例とした省電力効果について検討する．

キーワード データセンタ，ストレージ，省電力，階層的データ管理

1. はじめに

データセンタの総運用コストは年々増加している．特に電力や冷却のためのコストの増加は著しく，1997 年にはハードウェアコストの 10%程度でしかなかった電力及び冷却のコストは，2011 年には 75%程度にまで上昇するとの報告もある[1].

データセンタにおけるデータ量の増加率は年率 30%から 60%と非常に高く[2]，この結果ストレージの出荷容量は年率 50%，その消費電力の増加率は年率 20%と他の IT 機器を圧倒している[3]．データセンタの主要なアプリケーションの一つである Online Transaction Processing (OLTP)が稼動するシステムでは，ストレージの消費電力が IT 機器全体の消費電力の 70%以上を占めるとの報告もある[4]．このように，データセンタの電力コストの低減にはストレージの省電力化が最重要の課題の一つであると言える．

近年，高いアクセス性能や短いダウンタイム(データにアクセスできない時間)など高いサービスレベルが要求される重要なデータの管理にコストを掛け，そうではないデータの管理コストを低く抑える「階層的なデータ管理」が注目されている[19]．階層的なデータ管理は，データに求められるサービスのレベルに基づきデータの管理を階層化する．そしてデータ管理階層毎にアクセス性能や冗長度が異なるストレージを割当てる．さらに，データに求められるサービスのレベルが時間の経過とともに変化することに着目し，データに求められるサービスのレベルに見合ったデータ管理階層にデータを移動することで，増大し続けるデータの管理コストを低減する．

これまで，データセンタの電力コストを下げるための様々なアプローチが提案されている．これらには，IT 機器及び空調設備を合わせたデータセンタ全体の消費電力を削減する研究[15-18]，空調機器の省電力化を図る研究[12-14]，サーバの消費電力の低減を図る研究[5-7]，ストレージの消費電力の低減を図る研究[8-11]などがある．しかし，これらの研究では何れも，データに求められるサービスのレベルがデータごとに異なり，かつそれらが時間の経過とともに変化することについては考慮されていない．

そこで我々は，サービスレベルが異なり，かつそれが時間の経過とともに変化するデータを蓄積するストレージの省電力化を目的とした，階層的なデータ管理手法を用いたストレージの電力管理手法を提案する．本手法は，まずデータに求められるサービスのレベルを定義し，それに基づきデータ管理の階層を定める．次にデータ管理の階層に合わせて入出力性能や電力制御方式(アクセスがない場合は電源 OFF 等)が異なるストレージ階層を作成し，データを配置する．さらに，データに求められるサービスレベルの変化に応じてデータを適切なデータ管理階層に再配置する．これにより，サービスレベルが異なり，かつそれが時間の経過とともに変化するデータを蓄積するストレージの省電力化を図る．

本論文では，提案手法の処理方式について述べるとともに運用中のデータセンタを具体例とした省電力効果について検討する．以下，2 章で関連研究について

述べ，3 章で階層的データ管理手法について述べる．4 章で省電力化を対象とした階層的データ管理手法について述べ，5 章で我々が具体例として用いたデータセンタについて述べる．6 章で評価について述べ，最後に 7 章でまとめる．

2. 関連研究

2.1. ファシリティ制御

ファシリティ制御とは，空調機器と IT 機器を併せた電力の削減を試みる研究である．例えば，文献[18]は，データ解析，可視化，知識発見技術の使い方の調査結果，およびこれらを電力，冷却，計算の 3 サブシステムに適用する際のユースケース，効果的な使い方を提案している．文献[17]は，ラックに収められた blade server を対象に，blade server の消費電力と blade server に冷気を送るファンの消費電力の合計を最小化する手法を提案している．文献[16]はサーバのアイドル時消費電力と空調の消費電力のトレードオフを図ると同時にサーバにジョブを過剰に配置することによりサーバの稼働台数を減らす手法を，文献[15]は空調電力とサーバ電力を最小化するジョブの配置を Linear Programming により求める手法を提案している．

2.2. 空調制御

文献[12]は，データセンタ内の冷却能力が場所により異なるため，ジョブをサーバに均一に配備したのではホットスポットが生じ冷却により多くの電力が必要になることを指摘し，温度が低いサーバにより多くのジョブを配備する手法を提案している．また，文献[13]は機器が取り込む空気の温度をできるだけ高くするようジョブのスケジューリングを行う手法を提案している．文献[14]では，温度に加えエア・フローを考慮したジョブのスケジューリングを提案している．

2.3. サーバ省電力

サーバ省電力化の研究には，エージェントを用いてサーバの消費電力を目標電力以下に制御する手法[5]，仮想化環境において仮想サーバの物理サーバへの配置を制御し物理サーバの電力削減を行う手法[6, 7]などがある．

2.4. ストレージ省電力

文献[8]にて提案された Massive Arrays of Idle Disks (MAID)は，高アクセス頻度のデータをキャッシュディスクに保存し，他のディスクの電源を OFF することにより省エネルギー化を図る手法である．文献[9]は，主にファイルサーバ向けにアクセス頻度の高いデータを少数のディスクに集中させ，他のディスクをスタンバイ状態とする手法を提案している．また，RAID を構成するストレージを対象とした省電力手法も提案されている．PARAID[10]は，データセンタ等での負荷の変動に対応すべく，RAID のパリティ配置を偏らせることによりアクティブ状態のディスク数を動的に変える．文献[11]は大規模データセンタで用いられるストレージを対象に RAID グループを高頻度でアクセスされるブロックを格納する Hot RAID グループとそれ以外のブロックを格納する Cool RAID グループに分け，Cool RAID グループの省電力化を図る手法を提案している．

2.5. 考　察

上記に述べたように，関連研究では，個々の IT 機器に対する省電力制御手法の提案が主であり，ユーザがデータに求められるサービスレベルはデータ毎に異なっている点については考慮されていない．また時間の経過とともにデータに求められるサービスレベルが変化することについても考慮されていない．

3. 階層的データ管理

データには，例えば金融機関における口座情報など高いアクセス性能や短いダウンタイム，データ復旧時間，堅牢なセキュリティなどの高いサービスレベルが求められるデータ，ファイルサーバのように口座情報ほどの性能や信頼性は求められないがユーザが不満に感じない程度のサービスレベルを必要とするデータ，あるいはメールの履歴などほとんどアクセスされることがない大量データの長期保管のようにコストをできる限り抑えつつ最低限のサービスレベルを満たせばよいデータなど，様々なサービスレベルのデータが存在する．

階層的データ管理では，データに求められるサービスレベルがデータ毎に異なることに着目し，サービスのレベルに基づきデータの管理階層を構築する．そして各データ管理階層に求められるサービスを提供するための最低限の装備や運用方法を備えたストレージをそれぞれの管理階層に割当てる．これにより，ストレージ及びその運用に関わるコストの低減とデータに求められるサービスレベルの維持の両立を図る．さらにデータに求められるサービスレベルが変化した場合には，適切なデータ管理階層にデータを移動することにより，サービスレベルの維持とコストの低減の持続を図る[19].

4. ストレージ省電力化を対象とした階層的データ管理

前章で述べた階層的ストレージ管理におけるコストには，ストレージインフラ(ハードウェア及びソフトウェア)，保守，バックアップとリカバリ，人的コスト，災害対策やデータロス対策のコストが含まれる[19]．しかし近年急増している消費電力については考慮されていない．我々は，階層的データ管理に省電力に関す

る観点を追加することで，データセンタの構築時のみではなく運用時においてもコスト(ストレージの運転コスト)の低減が可能になると考えた．本節では，我々が提案するストレージ省電力化を対象とした階層的データ管理方式について述べる．

4.1. データ管理指標

ストレージの省電力化を目的としたデータの管理階層を構築するために，我々は，サービスのレベルを示す指標として，(1) データのサーバ主記憶への最大転送性能(秒当り転送データ量や秒当り I/O 数)，(2) データのアクセス契機とアクセス持続時間，及び(3)データ量，を選んだ．

高いデータ転送性能を実現するためにはディスク筐体を同時に稼動させる必要がありその分単位時間当たりの電力を要する．逆に高いデータ転送性能が求められない場合は同時に稼動させるディスク筐体は 1 台でよく，単位時間当たりの電力は少ない．ストレージの消費電力をデータ転送性能に見合ったものにするために，我々は最大データ転送性能をデータ管理指標として選んだ．

また，ストレージを省電力化するためには，ストレージの省電力機能を最大限活用する必要があるが，このためにはストレージが Idle 状態である時間を長く取る必要がある．そこで，データのアクセス契機やアクセス持続時間が近いデータを同じストレージに配置することにより Idle 時間を長く取るようにする．これを可能とするために，データのアクセス契機とアクセス持続時間を第二のデータ管理指標として選んだ．

さらに，データ量もストレージの台数を決める上で必要になる．そこで我々は第三の指標としてデータ量を選択した．

4.2. データ管理指標の取得

データの管理指標である最大データ転送性能，データのアクセス契機とアクセス持続時間，及びデータ量はデータのユーザからのヒアリングなどにより取得する．しかし，データの数は膨大であり全てのデータについて管理指標を設定することは現実的ではない．このため，ユーザから指示があったデータについてはその管理指標を用い，それ以外のデータについてはデータに対するアクセス履歴から，データに求められる最大データ転送性能，データアクセス回数及びアクセス時間を推定する．

4.3. データ管理階層の構築

ストレージの消費電力削減を目的とした階層的データ管理では，データに求められるデータ転送性能を満たしつつストレージの消費電力を削減するためのデータの管理階層を構築する．

データの管理階層を構築するに当たっては，まずデータの最大転送性能に着目する．高いデータ転送性能は，ストレージのディスク筐体を並列稼動させることにより実現する．ストレージの消費電力はディスク筐体の並列稼働台数にほぼ比例するため，データの最大転送性能を達成するために必要なディスク筐体台数に基づきデータの管理階層を定める．その後，データのアクセス契機やアクセス持続時間が近いデータを配置するストレージ階層を構築する．これらの方式により，単位時間当たりのデータ転送量が高いデータに対しては単位時間当たりより多くの電力を，そうではないデータに対してはより少ない消費電力を割当てられるようにする．

4.4. ストレージ省電力方式

データのアクセス契機やアクセス持続時間はデータに求められる最大転送性能とは独立である．データのアクセス契機やアクセス持続時間の長さから求まるディスク筐体のアイドル時間の長さによりストレージの省電力機能の適用可否が決まる．そこでストレージの各階層をさらにストレージ省電力機能が適用可能な部分と不可能な部分に分け，ストレージの電力管理を行う．

図 1 は並列に動作させる最大ディスク筐体数が 3 の場合の省電力向けデータ管理階層を示している．

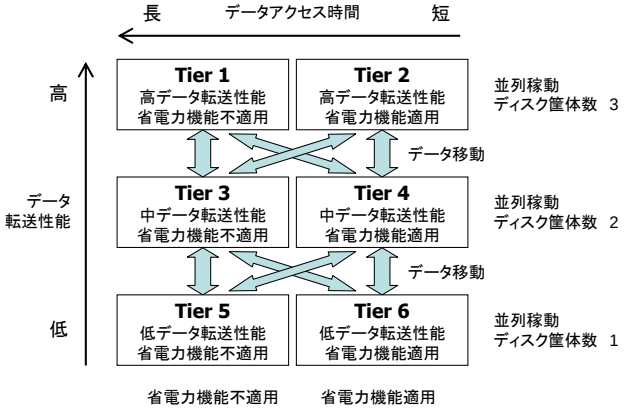

図 1. ストレージ省電力向けデータ管理階層

4.5. ストレージ階層の構築

ストレージ階層の構築では，データ管理階層に合わせてストレージ階層を決定する．まず，データ管理階層に必要なディスク筐体数であるが，これはデータ管理階層内のデータの同時最大データ転送性能とデータの合計容量を満たすために必要となるディスク筐体数の多い方とする．ディスク筐体当りのデータ転送性能を T，データ管理階層内のデータの同時最大データ転送性能を t，RAID グループ当りの容量を S，データ管理階層内のデータの合計容量を s とすると，データ転送性能を満たすために必要な RAID グループ数 m_t，容量を満たすために必要な RAID グループ数 m_s 及び RAID グループ数 m はそれぞれ以下の式により求める．

$$\left.\begin{aligned} m_t &= \lceil t/T \rceil \\ m_s &= \lceil s/S \rceil \\ m &= \max(m_t, m_s) \end{aligned}\right\} \quad \text{（式 1）}$$

次に，式 1 により求めたディスク筐体数に基づき，ストレージ階層を構築する．まずデータ転送性能を満たすために並列にアクセスしなければならないディスク筐体数 m_t が最大の階層を構築する．この時，ストレージのパス及びプロセッサがディスク筐体間で共有されないようにディスク筐体の配置を選ぶ．m が m_t よりより大きい場合(m_s>m_t)は，必要な容量が提供できるまでディスク筐体を追加する．この場合，ストレージのパス及びプロセッサはディスク筐体間で共有されてよい．次に，並列にアクセスしなければならないディスク筐体が 2 番目に多い階層を構築する．その後，順次並列にアクセスしなければならないディスク筐体数が少ない階層を構築する．さらに，各階層の省電力制御方式を決定する．これは，当該階層に格納するデータにより決まるディスク筐体のアイドル時間長に基づき決める．

4.6. ストレージ階層へのデータの配置

次に，構築したストレージ階層にデータを配置する．データに求められる最大データ転送性能及びデータアクセス時間に基づき，データを各階層に配置する．ストレージ階層内にデータ配置先となるディスク筐体が複数存在する場合は，組み合わせ最適化問題である Bin Packing 問題の解法の一つである Best Fit Decreasing 法を用いて，データ量と最大データ転送性能がディスク筐体のそれを越えず，かつ最もデータのアクセス契機やアクセス持続時間が近いストレージにデータを配置する．Best Fit Decreasing を用いるのは，より少数のディスク筐体にデータを集約し稼動させるディスク筐体数を削減することにより，ストレージの消費電力の削減を図るためである．

また，データに求められるサービスレベルは時間の経過に伴い変化する．これに伴いデータを配置する階層も変える必要がある．このため提案手法は，サービスレベルの変化に伴いデータが属する管理階層を見直す．階層間でのデータ移動が必要な場合は，データの初期配置時と同様 Best Fit Decreasing 法を用いてデータを配置する．階層内に複数個のストレージがあり，階層内でデータの移動が必要と予測される場合は，ストレージに配置されているデータの転送量がストレージの最大データ転送量を超えると予想されるストレージのデータを移動対象として選択し，その中から最も少ない移動数でストレージの最大データ転送量を下回ることが可能なデータを選択し，Best Fit Decreasing 法を用いて移動先を決定する．もし移動先が見つからない場合は，データの移動回数は多くなるが最大データ転送量がより小さいデータを移動対象として選ぶ．それでも移動先ストレージが割当てられない場合は，新たな(自ストレージ階層内あるいは他ストレージ階層内の未使用の)ストレージ領域を割り当て，そこにデータを移動する．

なお，実際の運用においては将来データに求められるデータ転送性能やアクセス回数，時間は必ずしもユーザより得られるものではない．このため将来の挙動の予測が必要となるが，これは今後の課題としたい．

4.7. 従来の階層データ管理との差異

提案手法は，従来の階層データ管理をストレージの省電力化にも適用できるように拡張している．従来の階層データ管理は，データに求められる性能や容量効率に基づき階層を決定しており，データは階層内のディスク筐体に均一に配置される傾向にある．一方，提案手法は，消費電力削減の観点から稼働させなければならないディスク筐体数を減少させるようデータを片寄らせて配置する．これにより従来の階層データ管理と比較して省電力状態のディスク筐体数や省電力状態の持続時間を増やすことが可能となる．

5. データ統合・解析システム DIAS とストレージ省電力機構

次に，我々が提案手法を適用し評価を行ったデータセンタについて述べる．

5.1. データ統合･解析システム DIAS

DIAS とは，地球規模の観測や各地域で得られたデータを収集，蓄積，統合，解析し，地球規模の環境問題や自然災害の脅威に対する危機管理に有益な情報を提供するデータ統合・解析システム[19]である．

図 2 データ統合解析システム DIAS

その主なアプリケーションは，海洋の気候変動の分析，ユーラシア寒冷圏の氷河の長期的な変動の明確化，地球上の天候変動と植生変動の関連の分析などである．これらのアプリケーションは，ストレージより数十 GB～1TB のデータを読み出してサーバの主記憶に展開し，解析やシミュレーションを行う．そして結果をストレージに書き戻す．DIAS は 3 台のサーバと約 1.6PB の容量を持つストレージ(全部で 5 台)を有する地球環境デジタルライブラリシステムであり，日々，計測デー

タやシミュレーション結果などのデータが追加されている．運用開始は 2008 年度である．

DIAS の写真を図 2 に示す．サーバは(株)日立製作所の SR16000/VL1，ストレージは同じく(株)日立製作所の Adaptive Modular Storage 2500 である．

5.2. ストレージ省電力機構と電力特性

DIAS が使用しているストレージの概要について説明する．図 3 はストレージの構成の概要である．ストレージは, 13 台のディスク筐体(1 ディスク筐体当り容量約 10TB，アクセス性能約 150MB/s)と，1 台のコントローラ筐体を有している．ディスク筐体は RAID 6 (13D+2P)構成を取る 15 台の HDD を格納している．ディスク筐体の電源状態を ON あるいは OFF にすることによりストレージの消費電力を制御する．コントローラ筐体とサーバは，Fibre Channel ケーブルで接続されており，サーバとの入出力，及びディスク筐体の電源の ON/OFF の切り替えを行う．

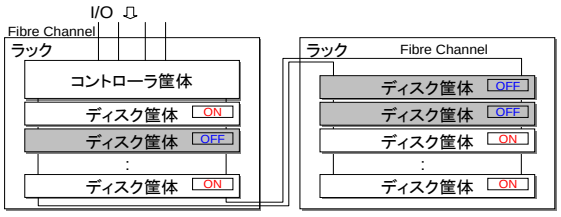

図 3 DIAS のストレージ装置の概要

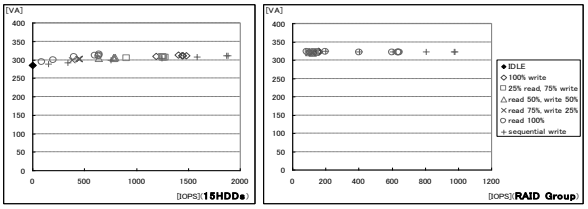

図 4. I/O 時のディスク筐体消費電力(左)とコントローラ筐体消費電力(右)

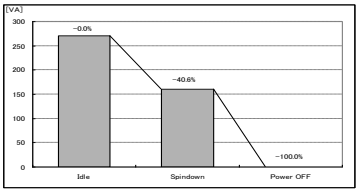

図 5. ストレージ省電力機能使用時の消費電力

次に我々は，ストレージのコントローラ筐体及ディスク筐体の電源ケーブルにクランプオンセンサを装着し，電力特性を計測した．ディスク筐体及びコントローラ筐体の消費電力の計測結果を図 4 に，省電力機能使用時のディスク筐体の消費電力を図 5 にそれぞれ示す．図 3(左)からわかるように，ディスク筐体の消費電力は I/O 数が増加するに従いわずかに増加している(アイドル時+10.6%). 一方コントローラ筐体の消費電力は I/O 数が増加してもほとんど変化していない．また，図 5 より Spin down 機能を用いた場合のディスク筐体の消費電力は-40.6%, Power Off 機能を用いた場合の消費電力は 0 であることが分かる．このことから，高い省電力効果を得るためにはディスク筐体の省電力機能を最大限活用しなければならないことが分かる．

また，ディスク筐体の起動には数万ジュールの電力損が発生する．このため，ディスク筐体の省電力機能を使用するためには約 100 秒の Idle 時間が必要である．

6. 階層的データ管理の省電力効果

我々は，5 章及び 6 章で述べた階層的データ管理手法を DIAS に適用した場合の省電力効果を確認すべく，特にアクセス頻度の高いストレージ装置を対象に省電力効果の検証を実施した．

6.1. DIAS の I/O 挙動特性と性能要件

(a) I/O 挙動特性

まず我々は，データの I/O 挙動特性と性能要件の調査を行った．I/O 挙動特性の調査を行った理由は，全てのデータの性能要件をユーザより取得することは現実的ではなく，データに求められる性能要件を自動的に決定しなければならないためである．DIAS が保有するファイル数も膨大であり個々のファイル毎に I/O 挙動を調査し性能要件を決定することは非現実的であると考えられるため, 我々はデータを 52 個のデータセットに分割し，データセット単位で I/O 挙動の調査を行った．分析対象期間は 2010 年 4 月及び 5 月である．図 6 上段に並列に動作させるディスク筐体数を決定する最大データ転送量とその推移を，図 6 下段にストレージ階層の省電力手法適用可否を決定する平均 Idle 時間とその推移をそれぞれ示す．

図 6 上段から分かるように，データ転送性能の最大値は約 150MB/s であった．また，いくつかのデータについては, 2010 年 4 月には 120MB/s 以上であった最大データ転送性能が翌月では 50MB/s 以下となった場合や，あるいは逆にデータ転送性能の最大値が数 MB/s しかなかったデータの最大が翌月には 140MB/s を超えるなど，いくつかのデータについてデータが属すべき管理階層が変化していることが分かる．また，図 6 下段より，ストレージ省電力機能の適用可否の閾値となる平均 Idle 時間が 100 秒未満のデータが 2010 年 4 月にはいくつか見られたが 2010 年 5 月には見られなくなるなど，データのアクセス契機やアクセス持続時間の観点からもデータの属すべき管理階層が変化していることが分かる．

(b) ユーザ要件

本評価において我々は，ユーザ要件の取り込み時の効果を確認するために，転送データ量が特に高かった 6 種類のデータ(30, 33, 40, 43, 44, 53)及びアクセス時間が長かった(平均 Idle 時間が短かった)5 種類のデータ(25 , 29, 37, 41, 45)のアクセス時間がユーザより指定

されたものとして評価を行った．データ転送量は観測値の 2 倍，アクセス時間は観測値が指定されたものと仮定して評価を行った．

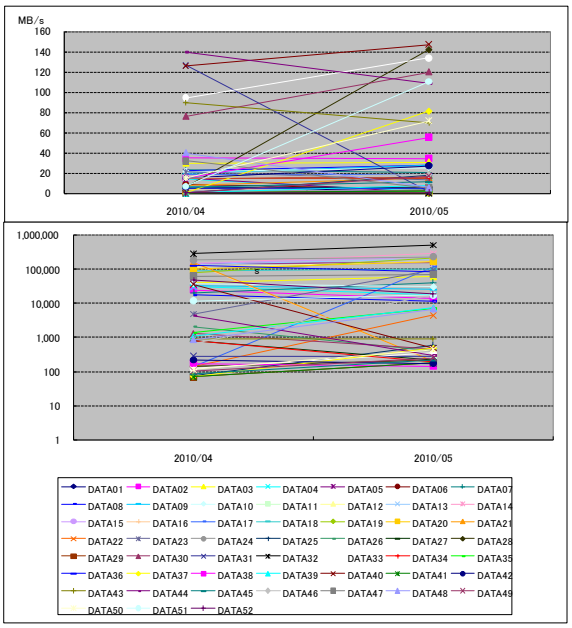

図 6 データ転送性能(上段)及び平均 Idle 時間(下段)とそれらの推移

6.2. データ管理階層

表 1 は，データの観測結果及び我々が仮定したユーザ要件に基づくデータ管理階層(階層 1～4)の構築及び各階層へのデータの配置結果を示している．最大データ転送量はディスク筐体 1 台の最大データ転送性能 150MB/s を，平均 Idle 時間(データのアクセス契機やアクセス持続時間より求まる)はストレージの省電力機能適用可否の閾値となる 100 秒を基準とした．

表 1 I/O 挙動観測結果に基づくデータ管理階層の定義とデータの配置

データ管理階層			データの配置	
階層 #	最大 MB/s	平均 Idle 時間	2010/4 月	2010/5 月
1	150MB/s 以上 300MB/s 未満	100 秒未満	-	-
2		100 秒以上	30, 33, 43	30, 33, 40, 44, 53
3	150MB/s 未満	100 秒未満	25, 29, 37, 41, 45	-
4		100 秒以上	残り 44 データ	残り 47 データ

次にデータの配置であるが，階層 1 に属するデータは 2010 年 4 月，5 月を通して存在せず，階層 2 に属するデータは 2010 年 4 月が 2, 2010 年 5 月が 5 データあった．また，階層 3 には 2010 年 4 月では 5 データが該当するが 2010 年 5 月では 0 であった．そして，階層 4 には 40 以上のデータが属する結果となった．

各階層にデータを配置した場合の階層ごとの最大データ転送量は階層 2 が 294.2MB/s，階層 3 が 113.8MB/s，階層 4 は 294.2MB/s であった．各階層のデータ量は，階層 2 及び 3 が約 8.5TB，階層 4 は 80.0TB であった．

また，2010 年 4 月から 5 月にかけてデータ管理階層を変更しなければならないデータの数は全体の約 15%程度あることが分かる．但し，データ管理階層におけるストレージ省電力機能の使用可否の変更にはデータ移動は必ずしも必要ではない．

6.3. DIAS への階層データ管理の適用

6.3.1. DIAS ストレージの構成

DIAS，は図 3 に示したディスク筐体を持つ 5 台のストレージを持つ．各ストレージはそれぞれ 13 台のディスク筐体を持っており，ディスク筐体の数は合計 65 台である．

6.3.2. ストレージ階層の構築

6 章で示した方式に基づき，表 1 に示したデータ管理階層に対応するストレージ階層を構築した結果を図 7 に示す．

表 1 より，階層 2 に対応するデータを格納するためにはディスク筐体 2 台が必要でありそれらを並列で動作させなければならない．そこでディスク筐体#1 及び#2 を一つのストレージ階層(H2)とし，階層 2 に割当てる．次に階層 3 であるが，データ転送性能，容量ともディスク筐体 1 台でよいため，ディスク筐体#3 を一つのストレージ階層(H3)とし，階層 3 に割当てた．階層 4 については，ディスク筐体が 8 台あればよいため，残りのディスク筐体(62 台)を全て階層 4 に割当てた．

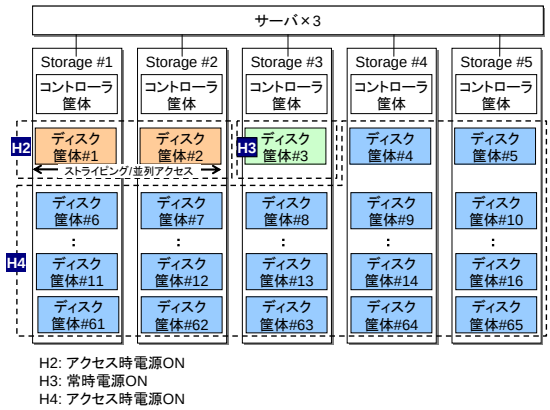

図 7 ストレージ階層

ストレージ階層毎の省電力方式であるが，ストレージ階層 H2 及び H4 は平均アイドル時間がストレージ省電力機能の適用可否の閾値となる平均 Idle 時間が 100 秒より長いため，データにアクセスがある場合のみ電源を ON にする(省電力機能適用)．ディスク筐体#3 については，2010 年 4 月の平均 Idle 時間は 100 秒未満であるため常時電源 ON(省電力機能未適用)とし，データのサービスレベルの変更時に省電力機能の適

用可否を見直し，省電力機能の適用が可能であれば適用を行う方針とした．

6.3.3. ストレージ階層へのデータの配置

次に，6 章で示した方式に基づき，各ストレージ階層にデータを配置する．階層 2，3 については，データの配置先となるストレージ階層内のディスク筐体数は 1 つ(階層 2 はストライピングを行うため一つとして扱う)のため，ストレージ階層 H2，H3 にそれぞれデータを配置した．

階層 4 については 4 章で述べたように Best Fit Decreasing 法に基づきデータを配置した．

6.4. 省電力効果の確認

次に，我々は階層的データ管理を用いたストレージ省電力化の効果を確認するために，シミュレーションによる評価を行った．

6.4.1. シミュレーション条件

シミュレーションには，7.1 節で示したデータ及びデータ管理階層の定義を用い，7.2 節で示したストレージ階層とその初期データ配置を用いた．ストレージの消費電力は図 4 及び図 5 により計測した値を用いた．階層 2 及び 3 に属するデータの最大データ転送性能及びアクセス回数，アクセス時間はユーザが月の変わり目にそれらを指定したものとし，階層 4 については一ヶ月前の観測値より得られた結果に基づきデータを再配置すると仮定した．ストレージの消費電力とデータ転送性能については 2010 年 4 月の観測結果に基づきデータを配置した上で，2010 年 5 月のデータを用いた場合の効果を計測した．データ移動については 2010 年 4 月の観測結果に基づくデータ配置から，2010 年 5 月の観測結果に基づくデータ配置にデータを移動する場合のデータ移動量を評価した．

さらに，性能のみを考慮する従来の改装データ管理手法を用いた場合の消費電力および性能との比較も行った．従来手法では省電力化を考慮しないため，階層 2 および 4 のみが存在する．また，階層 4 については負荷分散を目的に Worst Fit Decreasing 法に基づきデータを配置するとした．

6.4.2. 評価結果

(a) ストレージ消費電力とデータ転送性能

省電力制御を行わない場合，現状，アクセスが行われていないディスク筐体の電源を OFF にした場合，提案手法を用いた場合，および従来の階層データ管理を用いた場合の省電力効果の比較結果を図 8 に示す．図 8 より，提案手法を用いることでストレージの消費電力を最大 77.9%削減できる可能性があることが分かる．また，従来の階層データ管理は負荷分散が主たる目的であり階層 4 の全ディスク筐体の負荷ができるだけ均等になるようデータを配置する．このため，ストレージ消費電力の削減率は 44.0%に留まっている．

図 9 は，ユーザよりデータ転送性能が指定されたデータの，要求性能及び転送性能を示している．図から分かるように提案手法および従来の階層データ管理の双方においてユーザの求めるデータ転送性能を満たしていることが分かる．提案手法を用いない場合はデータが単一ディスク筐体上にあるため単一ディスク筐体の最大データ転送性能以上のデータ転送性能を出すことができず，ユーザ要件を満たしていない．

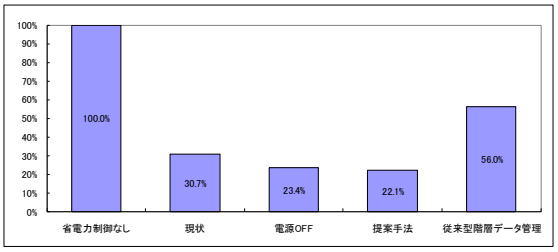

図 8 ストレージ消費電力

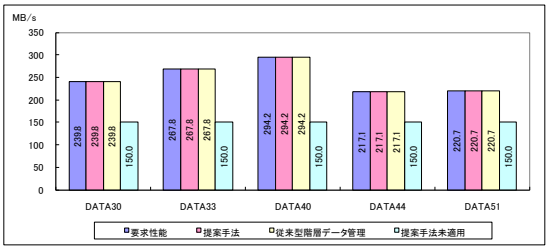

図 9 データ転送性能(階層 2)

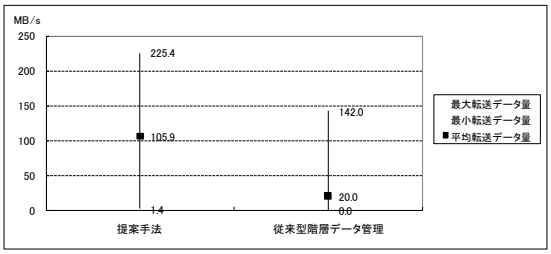

図 10 データ転送性能(階層 4)

図 10 は階層 4 におけるディスク筐体のデータ転送性能の計算値を示している．図から分かるように，提案手法では最大データ転送性能がディスク筐体 1 台分の最大データ転送性能 150MB/s を超えておりデータの転送性能が不十分であった可能性がある．一方，従来型の階層データ管理手法では，最大データ転送性能は 150MB/s 以下となっている．以上の結果より，提案手法はユーザの要件を満たしつつストレージの消費電力を低く抑える可能性があることが分かった．その一方で将来のデータ転送量に関する情報が不十分な場合は性能ボトルネックとなる可能性があることが分かった．

(b) データ移動量

2010 年 4 月から 5 月にかけて移動しなければならないデータ数は 5 であったまた，階層 4 内でも最大データ転送量がディスク筐体の最大値を超えないようにす

るためのデータの配置の見直しが必要である．2010 年 4 月から 5 月にかけては，合計 2 データの移動が必要であった．このため，合計のデータ移動量は約 11.9TB，移動時間はデータを一つずつ移動させた場合で約 24 時間である．

(c) アクセスパターンが既知の場合のデータ転送性能

データ統合解析システムのような，科学技術計算向けのシステムでは，ジョブのスケジュールなどの情報を活用することによりアクセスパターンを事前に知ることが可能になる場合がある．このような場合における提案手法の効果を確認するために，我々は 2010 年 5 月のアクセスパターンが既知であるとした場合のストレージの消費電力および階層 4 のデータ転送性能をシミュレーションにより求めた．結果を図 11 に示す．

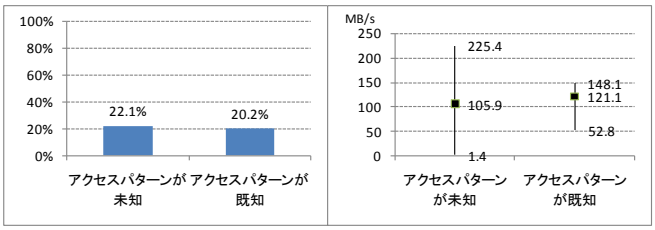

図 11 ストレージ消費電力(左)およびデータ転送性能(右) (アクセスパターンが既知の場合)

図 11 から分かるように，アクセスパターンが既知の場合はデータ転送性能を満たしつつ消費電力をさらに削減できることが分かる．これは，アクセスパターンの情報を用いることによりディスク筐体の最大性能を超えるようなデータ配置を避けることができ，かつ電力あたりのデータ転送量を増加させることができたためである．図 11 右においてアクセスパターンが既知の場合は平均および最小のデータ転送性能が向上していることもこれを裏付けていると考えられる．

7. まとめ

ストレージの省電力化のための階層的なデータ管理と階層的なデータ管理手法を用いた省電力ストレージの構築手法を提案した．また，提案手法の処理方式について述べるとともに運用中のデータセンタを具体例とした省電力効果について検討した．この結果，提案手法を用いることにより，データに求められる性能要件を満たしつつストレージの消費電力を最大約 78%低減できる見込みを得た．

今後は提案手法の実装及び DIAS ストレージ上での実機を用いた評価を行う予定である．

参 考 文 献

[1] Eastwood, et. al, “The Business Value of Consolidating on Energy-Efficient Servers: Customer Findings”, IDC White Paper #218185, 2009.

[2] Fellows, R, “Data Center Transformation”, https://www.eiseverywhere.com/file_uploads/bde68f8d6aa42fe8abbf315aa10e29ed_Fellows_Monday_0920_SNWF10.pdf, 2010

[3] Rajecki, K., “Tiered Storage Architectures”, http://www.sun.com/solutions/landing/industry/education/pdf/webinar_050708.pdf, 2008.

[4] Poess, M., et. al, “Energy cost, the key challenge of today’s data centers: a power consumption analysis of TPC–C results”, Intl. Conf. on Very Large Data Base, 1229-1240, 2008.

[5] Das, R., et. al, “Autonomic Multi-Agent Management of Power and Performance in Data Centers”, Proc. of 7th Intl. Conf. on AAMAS 2008, pp.107-114, 2008.

[6] R. Nathuji, et. al, “VirtualPower: Coordinated Power Management in Virtualized Enterprise Systems”, 21st ACM SOSP ’07 , 2007

[7] R. Nathuji, et. al, “VPM Tokens: Virtual Machine-Aware Poweer Budgeting in Datacenters”, ACM HPDC ’08, 2008.

[8] Colarelli, D., et. al, “Massive Arrays of Idle Disks For Storage Archives”, Supercomputing ACM/IEEE Conference, 2002.

[9] Pinheiro,E., et. al, “Energy Conservation Techniques for Disk Array Based Servers”, 18th Annual International Conference on Supercomputing, 2004.

[10] Weddle, C., et. al “PARAID: A Gear-Shifting Power-Aware RAID”, 5th USENIX Conference on File and Storage, 2007.

[11] Otoo, E., “Dynamic Data Reorganization for Energy Saving in Disk Storage Systems”, Scientific and Statistical Database Management Conference, 2010.

[12] Moore, J., et. al, “Making Scheduling “Cool”: Temperature-Aware Workload Placement in Data Centers”, Proc. of the Annual Conference on USENIX Annual Technical Conference,2005.

[13] Tang, Q, et. al, “Energy- Efficient, Thermal-Aware Task Scheduling for Homogeneous, High Performance Computing Data Centers: A Cyber-Physical Approach”, IEEE Transactions on Parallel and Distributed Systems, Vol.19, Issue 11 , 2008.

[14] Vasic, N., Scherer. T., Schott, W., “Thermal-Aware Workload Scheduling for Energy Efficient Data Centers”, Proc. of the 7th International Symposium on Autonomic Computing, 2010.

[15] Pakbaznia, E., Pedram. M., “Minimizing Data Center Cooling and Server Power Costs”, Proc. of the 14th ACM/IEEE International Symposium on Low Power Electronics and Design, 2009.

[16] Ahmad. F, et. al., “Joint Optimization of Idle and Cooling Power in Data Centers While Maintaining Response Time”, Proc. of the 15th Edition of ASPOLS on Architectural Support for Programming Languages and Operating Systems, 2010.

[17] Wang, Z., Tolia, N., Bash. C., “Opportunities and Challenges to Unify Workload, Power, and Cooling Management in Data Centers”, ACM SIGOPS Operating System Review, Vol.44, Issue 3, 2010.

[18] Marwah, M., Sharama, R., Shih, R., Patel, C., “Data analysis, Visualization and Knowledge Discovery in Sustainable Data Centers”, Proc. of the 2nd Bangalore Annual Compute Conference, 2009.

[19] Tallon P.P., “Undersitanding the Dynamics of Information Management Costs”, CACM Vol. 53, No.5, 2010.

[20] “DIAS データ統合・解析システム”, http://www.editoria.u-tokyo.ac.jp /dias/, 2008.