DBSJ Journal, Vol.9, No.3
February 2011

データセンタの階層的データ管理と省電力化を支援するモニタリング機構の開発

Development of Monitoring Mechanisms for Hierarchical Data Management and Power-Saving Data Center

西川 記史♥　中野 美由紀♠
喜連川 優♠

Norifumi NISHIKAWA　Miyuki NAKANO
Masaru KITSUREGAWA

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と並行してストレージの省電力化を図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も考慮した階層的なデータ管理手法とそれを支援するモニタリング機構を提案する．本論文では，提案手法の特長を述べると共に，運用中のデータセンタを具体例とした階層的なデータ管理と省電力効果について検討する．

Power Consumption of data centers is increased rapidly. Especially, quick growth of data causes drastic increase of storage power consumption. The power saving of storage becomes one of the most significant problems on serving data center effectively. Today, a hierarchical data management (HDM) is focused from a point of data center management view. The purpose of HDM is to reduce the management cost of data centers. In the HDM, data are classified into some varieties of hierarchical level based on users' requirements. Then data of each levels are stored into appropriate storage hierarchies. Many power saving methods of data centers are proposed, these works, however, do not consider a power saving of the storages using HDM. In this paper, we propose a HDM method for storage power saving, and develop a monitoring system that supports the HDM. We describe our proposed method and discuss its effectiveness for storage power saving.

♥正会員　東京大学大学院情報理工学系研究科博士後期課程
norifumi@tkl.iis.u-tokyo.ac.jp および株式会社日立製作所
システム開発研究所 Norifumi.nishikawa.mn@hitachi.com
♠正会員　東京大学生産技術研究所
{miyuki, kitsure}@tkl.iis.u-tokyo.ac.jp

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の伸びは年率57%，その消費電力の増加率は年率19%[3]と他のIT機器を圧倒している．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．

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，これらの研究でデータセンタ運用前の静的な省電力化の試みであり，データセンタのサービス開始後の実行時省電力は対象となっていない．例えば，データの動的な追加等における省電力化手法は提案されていない．

近年，高いアクセス性能や短い復旧時間など高いサービスレベルが要求される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[20]．この目的は，データの性能や信頼性に合わせてストレージをアクセス性能や冗長度が異なる階層に分割し，データをサービスレベルに適した階層に配置することで，増大し続けるデータの管理コストを低減することにある．

階層的なデータ管理では，ユーザが，応用処理の特性に基づきデータの管理階層を定める．このデータの管理階層をストレージの省電力化に用いることにより，ユーザが設計したアプリケーションの要件をストレージの省電力に活用することが可能になる．

そこで我々は，データセンタにおける階層的なデータ管理と省電力運用を支援する，新たなモニタリング機構を提案する．本モニタリング機構は特に省電力という観点からストレージに着目し，その特長は，i) ストレージ階層の性能，消費電力，温度の収集・蓄積，ii) データ毎のアクセス性能及びアクセス頻度の収集・蓄積，iii) 階層的なデータ管理のための指標の提供，iv) データの性能要件に適した格納先の検索，である．これらの特長により管理者はストレージ階層の性能や消費電力や，要件に適したストレージ階層にデータが配置されているかを知ることができる．このモニタリング機構を利用することで階層的なデータ管理が容易に実現されデータセンタの省電力化に貢献できる．

本論文では，データセンタにおける階層的なデータ管理を支援するモニタリング機構について述べると共に，開発したモニタリング機構により得られた情報を用いたデータセンタのストレージシステムの省電力化の試みを報告する．

以下，2章で従来のデータセンタ省電力手法を紹介し，3章で省電力化の対象となるデータセンタの概要について述べる．4章で階層的データ管理の要件と開発中の階層的データ

管理を支援するモニタリング機構について述べる．また5章で運用中のデータセンタの具体例として階層的データ管理と省電力効果について検討し，6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，階層的なデータ管理と並行した省電力手法の提案は行われていない．利用者が求める様々なサービスレベルを満たしつつ，データセンタの大規模なストレージを省電力化するには，従来のモニタリング情報に加え階層的なデータ管理を支援する情報が必要となる．また，階層的なデータ管理においてストレージ階層への適切なデータの配置の評価についても検討しなくてはならない．さらに，新たなデータが追加された後も消費電力を低く抑えるための支援についても必要となる．

3. 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．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[21]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している．

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．

我々は，DIAS 上にサーバやストレージの性能，消費電力，温度を計測・蓄積するモニタリング機構を構築・稼動させており，2009 年度からデータを蓄積している．DIAS の写真を図 1 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．

図 1　データ統合解析システム DIAS

Fig.1 Data Integration & Analysis System DIAS

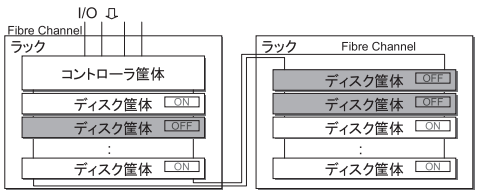

図 2 DIAS のストレージ装置の概要

Fig.2 Overview of the DIAS Storage System

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 2 はストレージの構成の概要である．ストレージは，13 台から 18 台のディスク筐体(1 ディスク筐体当り容量約 10TB)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．

4. 省電力モニタリング機構

本章では，階層的なデータ管理手法の概念，及び我々が提案するデータセンタの運用を支援するモニタリング機能について述べる．

4.1 階層的なデータ管理

階層的なデータ管理を行うには，まず稼動するアプリケーションやユーザの要件などを基にデータ管理階層を定める．次に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する．そして，データを適切なストレージ階層に配置する．

(a)データ管理階層の決定：ユーザの要件からデータが満たすサービスレベルを決定し，それに基づきデータ管理を階層化する．サービスレベルにはデータのアクセス性能やアクセス頻度，データのアクセス待ち時間，最大容量などがある．例えばユーザの要件がアクセス性能とアクセス頻度であり，ユーザがそれぞれを高・低の 2 つに分けた場合，データ管理階層は図 3 に示すようになる．この例では，アクセス性能やアクセス頻度が低いデータが電力コスト低減の対象と考えられる．アクセス性能とアクセス頻度は直交する軸であるため，低アクセス性能・高アクセス頻度データと高アクセス性能・低アクセス頻度データは同一階層にある．

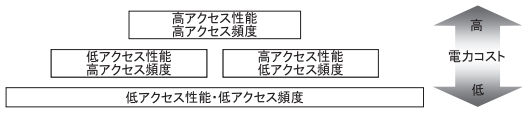

図 3 データ管理階層： 例

Fig.3 Data Management Hierarchy: An Example

(b)ストレージの階層化とデータの配置：データ管理の階層に合わせてストレージの階層を決定する．アクセス性能及びアクセス頻度をサービスレベルの指標とする場合，データのアクセス性能をストレージの性能に，データのアクセス頻度をストレージの電力制御方式にそれぞれ対応させる．図 3 のデータ管理階層を例にとると，ストレージの階層は，それぞれのサービスレベルを満たすように高性能高消費電力(H1)，低性能高消費電力(H2)，高性能低消費電力(H3)，及び低性能低消費電力(H4)の 4 つになる．ストレージ階層を構成するディスク筐体の数は，ユーザが求めるアクセス性能やデータ量を満たすように決定する．1 台のディスク筐体では十分なアクセス性能が出せない場合は，複数のディスク筐体間でデータをストライピングすることにより性能を確保する．また，ほぼ毎日アクセスされるデータを格納するストレージ階層は常時電源を ON にするなど，アクセス頻度に基づき省電力方式を決定する．図 4 は図 3 に示す 4 つのデータ管理階層がある場合のストレージ階層(H1～H4)とデータ配置の例である．図 4 では，ユーザが求める高いアクセス性能を単一のディスク筐体で満たす高性能ストレージ，高いアクセス性能を複数のディスク筐体間でストライピングを行うことにより満たすことができ，かつディスク筐体毎の電源 ON/OFF が可能な中性能ストレージ，ディスク筐体単位の電源 ON/OFF が可能であるがユーザが求める高いアクセス性能は出せない低性能ストレージ，の 3 種類のストレージがあると仮定している．管理者はこれら 3 種類のストレージを，高性能ストレージ内の階層(H1)，常時電源 ON である中性能ストレージ(H2)，アクセス時のみ電源 ON かつ高い性能を出すためにストライピングされた中性能ストレージ(H3)，及びアクセス時のみ電源 ON にする低性能ストレージ(H4)，の 4 階層に分割する．そして，アクセス性能とアクセス頻度の両方が高いデータを H1 に，アクセス性能は低いがアクセス頻度が高いデータを H2 に，アクセス性能が高くアクセス頻度が低いデータを H3 に，アクセス性能，頻度とも低いデータを H4 に配置する．この例では，高性能ストレージのディスク筐体間ではストライピングを行っていない．しかし高性能ストレージのディスク筐体 1 台ではユーザのアクセス性能要件が満たせない場合は，高性能ストレージのディスク筐体間でもストライピングを行う．

ここではアクセス性能とアクセス頻度を用いたストレージ階層の例を述べた．データ応答時間，データ最大容量等の多様なユーザの要求に基づく多様なデータ管理階層の特性に基づいて，ストレージ階層を決める必要がある．より汎用的な議論は今後の課題としたい．

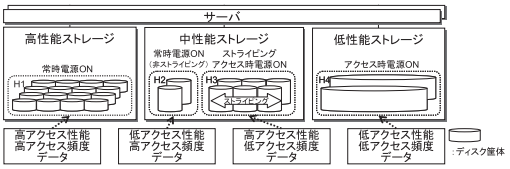

図 4 ストレージ階層とデータの配置

Fig.4 Storage Hierarchy and Data Placement

4.2 モニタリング機構の特長

前節で述べた階層的なデータ管理を支援するためのモニタリング機構の特長は次の通りである．

- ストレージ階層の性能，消費電力，温度の収集・蓄積
- データ毎のアクセス性能及びアクセス頻度の収集･蓄積
- 階層的なデータ管理のための指標の提供
- データの格納先に適したストレージの検索

階層的なデータ管理のために，我々が開発中のモニタリング機構は，従来のストレージ装置単位やディスク筐体単位の性能に加え[22]，ストレージの消費電力，ストレージ階層の性能及び消費電力値の提供，及びデータ毎のアクセス性能やアクセス頻度を提供する．ストレージ階層の定義とそれに基づくモニタ値の集計，階層的なデータ管理のためのモニタ項目の提示が，本モニタリング機構の特長である．これらの情報を用いて，データ毎のアクセス性能やストレージ階層のアクセス性能，消費電力の把握，データが配置されているストレージ階層の適切さの評価，及び適切なストレージへのデータの配置が可能になる．その結果，データセンタに求められる多様なアクセス性能を満たしつつ省電力化を達成できる．

4.3 モニタリング機構の機能

(a)サーバ及びストレージの性能情報，電力・温度情報の収集・蓄積：サーバ CPU のビジー率，ストレージのコントローラ筐体内プロセッサのビジー率，ディスク筐体のアクセス性能，ディスク筐体内の HDD のビジー率を一定時間間隔で収集し，性能情報管理 DB に格納する．またサーバやストレージの消費電力，及び温度を一定時間間隔で収集し，電力・温度情報管理 DB に格納する．ストレージの消費電力及び温度は，コントローラ筐体及び個々のディスク筐体単位で収集・蓄積する．

(b)サーバ・ストレージ性能情報及び電力・温度の可視化：サーバの CPU ビジー率，ストレージのディスク筐体毎のアクセス性能，及びコントローラ筐体のプロセッサのビジー率，ディスク筐体内の HDD のビジー率の推移及び平均値，現在

値を可視化する．またサーバ，及びストレージの筐体毎の消費電力及び温度の推移，平均値，及び現在値を可視化する．
(c)ストレージ階層の性能・消費電力情報の可視化：ストレージ階層のアクセス性能，ストレージ階層に含まれるディスク筐体内の HDD のビジー率の推移，平均値，及び現在値を可視化する．またストレージ階層に含まれるディスク筐体の合計消費電力の推移，平均値，現在値を可視化する．
(d)データのアクセス性能とアクセス頻度の可視化：データにアクセスが行われた時の単位時間当りのアクセス性能，及び単位時間当りのデータのアクセス回数(データ内のファイルのオープン回数の合計値)の推移，平均値，現在値を可視化する．
(e)ストレージ階層の検索：データの性能要件を満たすストレージ階層を検索し，管理者に提示する．次の 1，2 の条件を満たす階層を選択する: 1. 階層のアクセス性能+新規データのアクセス性能< 階層が提供できる最大アクセス性能，2. アクセス頻度が階層毎に決められた閾値以内．

図 5 モニタリングシステム画面

Fig.5 Data Viewer of Monitoring System

図 5 は，開発中のモニタリング機構の GUI を示している．左側はストレージの単位時間当たりアクセス性能，消費電力，及び温度の推移である．グラフ内の要素の色によってストレージ装置を区別している．右側は消費電力が高い時刻(左側図の点線枠内)のストレージのコントローラ筐体及びディスク筐体の消費電力を示している．1 ストレージを 2 列で表しており，小さな箱が筐体を示している．箱の色が赤に近い(色が濃い)ほど筐体の消費電力が高いことを示している．白い箱は筐体の電源が OFF であることを示している．

5. DIAS における階層的データ管理と省電力化

5.1 DIAS への階層的データ管理の適用

(a)データ管理の階層化：4 章で述べた考え方に基づき DIAS ユーザの求める性能要件を満たすサービスレベルを決定した．DIAS ユーザのヒアリングに基づき作成したデータ毎のサービスレベルとその管理方針を表 1 に示す．DIAS では，階層的データ管理を行うために必要なサービスレベルはアクセス性能とアクセス頻度，同時アクセス数であった．またアクセス性能は 50MB/s 以上か否か，アクセス頻度は一ヶ月で 24 日以上アクセスされるか否かで区分可能であることが分かった．アクセス性能とアクセス頻度，同時アクセス数は，他の科学技術計算向けのシステムでも見られる指標であるが，性能条件の具体的な数値は DIAS のアプリケーションによるものである．DIAS はストレージと主記憶の間のデータ移動を高いスループット(最大数 GB/s)で行う必要がある．最大数 GB/s にもなるアクセス性能を出すためには複数台のディスク筐体を並列に動作させる必要があり，電力が必要であるが，常時高スループットが必要な訳ではない．我々は，DIAS が用いるストレージを省電力化するためにはアクセス性能とアクセス頻度，同時アクセス数によるデータの分類が必要であると考え，これらを指標として選んだ．
(b)ストレージの階層化とデータ管理階層との対応付け：データ管理階層 D1～D4 のサービスレベルを満たすよう，DIAS のストレージをストレージ階層 H1 から H4 に分割した(図 6)．DIAS のディスク筐体のデータ転送性能は全て同一である[1]．そこで，1 台のディスク筐体でユーザの求めるアクセス性能を満たせない場合は，複数のディスク筐体間でデータをストライピングしユーザの求めるアクセス性能を出せるようにした．ストレージ階層 H1 は，D1 に対応するデータを格納する．D1 の性能要件はデータアクセス性能 50MB/s かつ同時アクセス数 10 以上である．このため 5 台のディスク筐体間でデータをストライピングし要求されたアクセス性能(500MB/s)を満たすようにした．また，D1 のデータはアクセス頻度が高いため，電源は常時 ON とした．ストレージ階層 H2 は D2 のデータを格納する．D2 のデータ量は約 21TB であるため 3 台のディスク筐体を用いた．データのアクセス性能要件は低いためストライピングは行わない．H2 も常時電源 ON である．H3 は D3 のデータを格納する．データアクセス性能は 50MB/s 以上でありかつ同時アクセス数が 3 以上であるため，2 台のディスク筐体間でデータのストライピングを行う．データのアクセス頻度は低いため，データにアクセスがある場合のみ電源を ON にする．残りのディスク筐体は D4 のデータを格納するストレージ階層 H4 である．D4 のデータ量は約 700TB である．データにアクセスがある場合のみディスク筐体の電源を ON にする．アクセス性能は低いためストライピングは行わない．

表 1 DIAS におけるデータ管理階層と管理方針

Table.1 Data Management Hierarchy and Policy of DIAS

データ管理階層のサービスレベル	データ管理方針
[D1] データのアクセス性能が 50MB/s，以上(同時アクセス数 10)，一月当り 24 回以上アクセス(高アクセス頻度)	高い転送性能と応答性とを維持，省電力化は積極的には行わない
[D2] データのアクセス性能が 50MB/s 未満であるが高アクセス頻度．	高い応答性能維持，省電力化は積極的には行わない
[D3] データのアクセス性能が 50MB/s 以上(同時アクセス数 3)，低アクセス頻度	高い転送性能を維持，省電力化を優先
[D4] データのアクセス性能が 50MB/s 未満かつ低アクセス頻度．	省電力化優先，転送性能，応答性能は低くてもよい

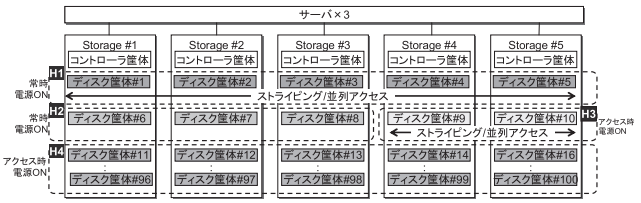

図 6 DIAS のストレージ階層

Fig.6 Storage Hierarchy of DIAS

5.2 階層的データ管理による DIAS の省電力化

DIAS の消費電力及び性能に関し，階層的なデータ管理の有効性を運用中のデータに基づくシミュレーションにより検討した．比較手法として i)全てのディスク筐体の電源を ON のままにする(省電力制御なし)，ii)データが入っていないディスク筐体の電源を OFF にする(現在の DIAS の運用形態; DIAS 現状)，iii)データに一日以上アクセスがない場合は

[1] ディスク筐体 1 台当りの最大データ転送性能を 120MB/s としてストレージ階層を構築した．

ディスク筐体にデータが格納されていても電源を OFF にする(電源 OFF)，iv) データのアクセス性能要件は考慮せずアクセス頻度が高いデータをできるだけ少数のディスク筐体にストライピングせずに配置しデータに一日以上アクセスがない場合にディスク筐体の電源を OFF にする(Non-SLA)，及び v)我々の提案方式(階層的データ管理)の 5 つの手法を用いた．階層的データ管理のストレージ階層の構築とデータの配置，及び方式 iv)のデータ配置を決定するために，DIAS から収集した 2010/4 月のデータ毎の性能情報を用いた．消費電力及び性能の計算には，2010/5 月及び 6 月分の性能情報及び消費電力情報を用いた．方式 ii)の性能及び消費電力は 2010 年 5, 6 月の実測値である．それ以外の方式は，各ディスク筐体に配置されたデータ毎の秒当り I/O 数よりディスク筐体毎の秒当り I/O 数を求め，式 1 を用いて消費電力を計算した．式 1 は DIAS のストレージにおいてファイルシステムのランダムアクセスの実測値から算出した消費電力である．i はディスク筐体に対する秒当り I/O 数である．

$$\left.\begin{aligned} P(i) &= -1.594\times10^{-5}i^2 + 0.036i + 287.5 \quad (i \le 2000) \\ P(i) &= -1.840\times10^{-5}i^2 + 0.094i + 285.4 \quad (i > 2000) \end{aligned}\right\} \text{(式1)}$$

また，ディスク筐体の電源 ON/OFF の切り替えは，午前 0 時より 6 時間当該ディスク筐体にアクセスが行われていなければ当該ディスク筐体の電源を OFF にすると仮定した．その後もしアクセスが行われれば，その時点で電源を ON にし，翌日の午前 6 時まで電源 ON 状態が継続すると仮定した．

図 7 に消費電力の比較結果を，DIAS 現状を 100%とした時の比で示した．図 8 にデータ管理階層 D1～D4 内のデータの 2010 年 5，6 月の二ヶ月間における一回のアクセスにおけるアクセス待ち時間(左)と平均アクセス性能(右)を示す．

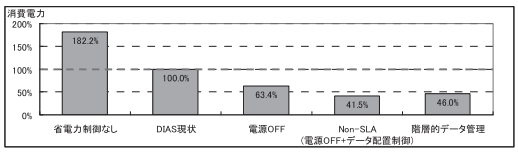

図 7 消費電力比較

Fig.7 Comparison of Power Consumption

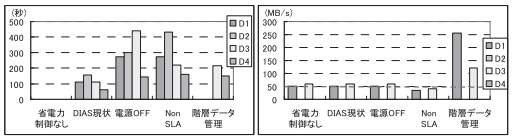

図 8 アクセス待ち時間(左)及びアクセス性能(右)

Fig.8 Wait Time (left) and Data Access Rate (right)

図 7 より，消費電力が最も少ない制御方式は Non-SLA であることが分かる．しかし図 8(右)より Non-SLA 方式の D1 及び D3 のアクセス性能はそれぞれ 38MB/s と 40MB/s であり，利用者が指定した 50MB/s を満たしていないことが分かる．これは，アクセス性能・アクセス頻度とも高いデータを単一のディスク筐体上に配置した結果，ディスク筐体でアクセス競合が発生したためである．さらに，図 8(左)より Non-SLA では高アクセス頻度のデータ(D1, 2)に対するアクセス待ち時間は D1 のデータではデータ当り平均約 280 秒，D2 では 400 秒以上ある．一方，図 7 より階層的データ管理のストレージの消費電力は Non-SLA 方式よりわずかに多いが他の方式と比較すると少ない．また図 8(右)より D1, 3 のデータアクセス性能はそれぞれ 257MB/s，119MB/s とサービスレベルの要件を満たしている．図 8(左)より D1, D2 の待ち時間がないことが分かる．また，図 8(左)において現在の DIAS の運用(DIAS 現状)でも待ち時間は少ないが，これはデータが格納されたディスク筐体は全て常時電源 ON のためである．つまり，現在の DIAS の運転ではデータが増加するにつれ消費電力が増大し，最終的には省電力制御無しの運用となる．以上より「階層的データ管理」方式に基づく省電力化のみが利用者が求める性能要件を満たしつつ，ストレージの消費電力を削減できる可能性があることが分かる．より汎用的な議論は今後の課題としたい．

5.3 低消費電力運用支援

本モニタリング機構を用いることによりストレージのさらなる省電力運用が可能になることを示す．

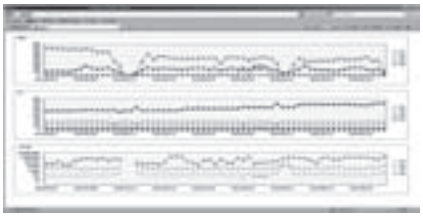

図 9 ストレージ階層のアクセス性能，電力，電力効率推移

Fig.9 Transition of Data Access Rate, Energy Consumption, and Energy Efficiency of Storage Hierarchy

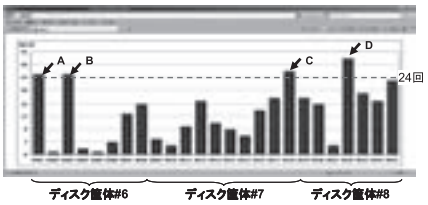

図 10 H2 階層内のアクセス頻度

Fig.10 Access Frequency of Storage Hierarchy H2

図 9 はストレージ階層(H1～H4)の一日毎のアクセス性能の平均値の推移(上段)，電力消費量の平均値の推移(中段)，及び電力消効率(1GB の転送に必要な消費電力量(MJ))の平均値の推移(下段)を示している．ストレージ側の視点であるディスク筐体毎ではなく，管理者側の視点であるストレージ階層に基づく性能及び消費電力，電力効率を表示する．これらの機能により，管理者は，階層毎の性能や消費電力の傾向の把握，問題点の発見を容易に行うことが可能となる．具体的には，電力効率の可視化により，アクセスを行っていない，あるいはアクセス性能が低いにも関わらず高い電力を消費している階層の発見が可能となる．図 9 の下段を参照することにより，階層 H1，4 と比較して階層 H2 の電力効率が悪いことが分かる．さらに図 10 に示す階層 H2 内のデータ毎のアクセス頻度と，ディスク筐体とデータとの対応関係表を参照することにより，階層 H2 を構成するディスク筐体それぞれに月当たりアクセス回数が 24 回以上のデータ(図 10 の A～D)が存在することが分かる．これから，ディスク筐体#7 のデータ C とディスク筐体#8 のデータ D をディスク筐体#6 に移動することで，ディスク筐体#7，8 の全てのデータのアクセス回数を月当たり 24 回未満にできることが分かる．もしディスク筐体#7, 8 のデータのアクセス待ちが可能であり，かつ D2 のデータが増加しないのであれば，ディスク筐体#7, 8 の階層を H2 から H4 に見直すことにより省電力化の可能

性を高めることができる．このように，階層的なデータ管理を支援するモニタリング機構を用いて継続的に監視を行うことにより，データのアクセス性能や頻度の変動に伴うデータとストレージ階層の不一致を発見することができる．

5.4 新規データ追加支援

DIASには，2010/8月から9月にかけて約24.2TBのデータが追加された．このデータを対象に，新規データの追加に対して階層的なデータ管理を用いた場合と用いない場合(DIAS現状)の消費電力とアクセス性能を比較した．ここで，新たに追加されたデータのユーザ要件を，アクセス性能140MB/s以上かつ計測期間の80%以上の期間でアクセスがある，とした．これらのデータはデータ管理階層D1に割当てられた．図11に比較結果を示す．図11より，階層的データ管理を用いた場合の消費電力は，DIAS現状の消費電力の約68.3%であるがアクセス性能は170.7MB/sとユーザ要件を満たしている．DIAS現状のアクセス性能は118.4MB/sでありユーザ要件をみたさない．これは，階層的データ管理がストレージ階層H1の5台のディスク筐体(常時電源ON)にストライピングしてデータを入れたのに対し，DIAS現状は新たなディスク筐体の電源をONにし，ストライピングを行わずにデータを追加したためである．

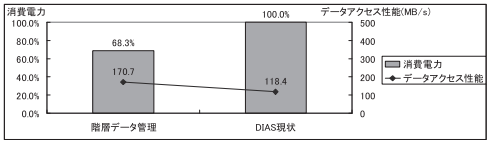

図11 新規データ追加後の消費電力とアクセス性能

Fig.11 Power Consumption and Access Performance after Loading New Data

5.5 実行時省電力支援のためのモニタ機構の課題

DIASに格納されているデータの特性や階層毎のデータ量は時間経過と共に変化する．我々は，データの変化に対応でき，かつストレージによるデータの識別を実現するモニタリング機構が必要と考える．

6. まとめと今後の予定

我々は，データセンタの省電力化も考慮した階層的なデータ管理とストレージ階層の導入，及び階層的なデータ管理を支援するモニタリング機構を提案した．階層的なデータ管理は，ユーザの要件に基づきデータを階層化し，それに合わせてストレージを性能や消費電力の異なる階層に分割しデータを配置する．モニタリング機構はストレージ階層の性能や電力，データの性能を可視化することにより階層的なデータ管理を支援する．我々は，東京大学で実際に運用されているデータ統合・解析システムに階層的データ管理とモニタリング機構を適用することにより，利用者の要求性能を損なうことなく消費電力を現状の運用と比較し54.0%削減できる可能性を得た．

今後は，現在開発中のモニタリング機構を拡張し自律的ストレージ省電力運転のための機構を開発する予定である．

[文献]

[1] Eastwood. M, 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] Patrick, B., et. al.: “GREEN STORAGE II: Metrics and Measurement” http://net.educause.edu/ir/library/pdf/churiedel.pdf. (2010).
[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 Intl. Conf. on Supercomputing (2004).
[10] Weddle, C., et. al: “PARAID: A Gear-Shifting Power-Aware RAID”, 5th USENIX Conf. 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 Conf. on USENIX Annual Technical Conference (ATEC ’05) (2005).
[13] Tang, Q, et. al: “Energy- Efficient, Thermal-Aware Task Scheduling for Homogeneous, High Performance Computing Data Centers: A Cyber-Physical Approach”, IEEE Trans. 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 Intl. Symposium on Autonomic Computing (ICAC ’10) (2010).
[15] Pakbaznia, E., Pedram. M.: “Minimizing Data Center Cooling and Server Power Costs”, Proc. of the 14th ACM/IEEE Intl. Symposium on Low Power Electronics and Design (ISLPED ’09) (2009).
[16] Ahmad. F, Vijaykumar, N. T.: “Joint Optimization of Idle and Cooling Power in Data Centers While Maintainging Response Time”, Proc. of the 15th Edition of ASPOLS on Architectural Support for Programming Languages and Operating Systems (ASPLOS ’10) (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 (COMPUTE ’09) (2009).
[19] “DIAS データ統合・解析システム”, http://www.editoria.u-tokyo.ac.jp/dias/ (2008).
[20] Tallon P.P.: “Undersitanding the Dynamics of Information Management Costs”, CACM Vol. 53, No.5 (2010).
[21] “My Atlas and Plot Service”, http://www.jamstec.go.jp/drc/maps/j/
[22]”日立ストレージソリューション ストレージシステム稼動管理”, http://www.hitachi.co.jp/products/it/storage-solutions/products/software/hsms/htm/

西川 記史 Norifumi NISHIKAWA
平元年神戸大学工学部計測工学科卒業．平3年同大学大学院工学研究科計測工学専攻修士課程修了．同年（株）日立製作所入社．システム開発研究所にてストレージ管理ソフトウェアの研究開発に従事．現在同研究所主任研究員及び東京大学大学院情報理工学系研究科電子情報学専攻．1998年度情報処理学会山下記念賞受賞．情報処理学会，日本データベース学会会員．

中野 美由紀 Miyuki NAKANO
東京大学理学部情報科学科卒業．博士（情報理工学）．富士通（株）勤務．1985年7月東京大学生産技術研究所助手(2004年助教)．2008年7月特任准教授．データベースシステム，ストレージシステム，データ工学の研究に従事．IEEE，電子情報通信学会，情報処理学会，ACM，日本データベース学会各会員．

喜連川 優 Masaru KITSUREGAWA
昭53東大・工・電子卒．昭58同大大学院工学研究科情報工学専攻博士課程修了．工学博士．同年同大生産技術研究所講師．現在，同教授．平15より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Webマイニングに関する研究に従事．情報処理学会フェロー，日本データベース学会理事．ACM SIGMOD Japan Chapter Chair，本会データ工学研究専門委員会委員長歴任．VLDB Trustee，IEEE ICDE，PAKDD，WAIMなどステアリング委員，SNIA日本支部顧問，文科省特定領域研究「情報爆発IT基盤」領域代表を務める．