

TPC-H 実行時省電力化のためのストレージ制御手法の提案

飯村 奈穂[†] 西川 記史^{††} 中野美由紀^{†††} 小口 正人[†]

[†] お茶の水女子大学

〒 112-8610 東京都文京区大塚 2-1-1

^{††} 東京大学生産技術研究所

〒 153-8505 東京都目黒区駒場 4-6-1

^{†††} 芝浦工業大学

〒 135-8548 東京都江東区豊洲 3-7-5

E-mail: [†]naho@ogl.is.ocha.ac.jp, ^{††}{norifumi,miyuki}@tkl.iis.u-tokyo.ac.jp, ^{†††}oguchi@computer.org

あらまし 近年デジタル情報は爆発的に増加しており、データセンタの大規模化が進んでいる。これにより、データセンタ内の管理運用コストも増加しており、その中でも電力コストの増加は見逃げせないものとなっている。そこで本研究ではデータセンタの省電力化に向け、データセンタの電力消費の中で一定の割合を占めるストレージに注目し、アプリケーションの性能劣化を最小限に抑えつつ、消費電力を抑えることを研究目的とする。本稿では、データベースベンチマークである TPC-H に注目し、大規模システム環境での TPC-H 実行時におけるストレージの省電力化に向けて、実行時の消費電力と性能評価およびデータ配置による省電力化手法の検討を行った。

キーワード 省電力、ストレージ、データインテンシブアプリケーション、性能評価、TPC-H

A proposal of Storage Control Method for Energy Saving on Runtime Database Benchmark

Naho IIMURA[†], Norifumi NISHIKAWA^{††}, Miyuki NAKANO^{†††}, and Masato OGUCHI[†]

[†] Ochanomizu University

2-1-1 Otsuka, Bunkyo-ku, Tokyo, 112-8610 JAPAN

^{††} Institute of Industrial Science, the University of Tokyo

4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 JAPAN

^{†††} Shibaura Institute of Technology

3-7-5 Toyosu, Koto-ku, Tokyo, 135-8548 JAPAN

E-mail: [†]naho@ogl.is.ocha.ac.jp, ^{††}{norifumi,miyuki}@tkl.iis.u-tokyo.ac.jp, ^{†††}oguchi@computer.org

1. はじめに

近年デジタル情報は爆発的に急増しており、今後 10 年で約 44 倍になるとも言われている。これに伴いストレージの出荷容量も急増していることからストレージの管理運用コストは見逃げせないものとなっており、データの効率的管理に注目が集まっている。

データセンタのエネルギー消費量は 2050 年には 2010 年度の日本の発電電力量の約 3 倍になると予測されており、社会全体での節電が求められる中でデータセンタの消費電力を削減することは急務になっている [1]。また、データセンタの消費電力割合の中でストレージの消費電力比率は約 13%を占めていること

から、ストレージの消費電力を削減することでデータセンタ全体の省電力化が可能であると言える。

本研究ではデータの効率的管理という点からクラウド上のデータベースの省電力化を考え、アプリケーションの性能劣化を抑えつつ、ストレージの消費電力を削減することを研究目的とする。本稿ではデータベースベンチマークである TPC-H [2] の省電力化に向けて、実行時のシステム性能と消費電力量の解析を行った。さらに、大規模システム環境を想定し、提案手法であるデータ配置制御のための実装と評価に向けた TPC-H 実行時のシステム性能評価、および提案手法の評価を行った。

これまでも多くのストレージ省電力手法が提案されてきた [3]- [7]。これらの研究では、ストレージレベルの入出力頻度

に従ってディスクを停止する等の手法が提案されている．しかし、実際にストレージレベルでの入出力を予測することは必ずしも容易ではない．

省電力アプリケーション協調型のストレージ省電力手法もある [8]-[10]．特に [9] では、アプリケーション協調型のストレージ省電力システムの構築を目指し、データインテンシブアプリケーションの I/O 挙動特性を解析、評価して、ストレージ電力制御モデルの提案を行っている．これらの研究では、提案手法におけるストレージの消費電力削減を考慮した上で、アプリケーションの性能評価を行っている．

そこで、本研究では、データインテンシブアプリケーションの SLA(Service Level Agreement) に注目し、アプリケーションの性能劣化を最小限に抑えつつ、ストレージの消費電力を削減し、性能評価を行っていくことを研究方針とする．

これまでの研究で、データ配置制御によるストレージ省電力化が有効であるということが示されている [11]．しかし、[11] の環境では評価に使用している HDD の台数や容量は少なく、大規模なデータを想定しているとは言い難い．そこで本稿ではより大規模な環境を想定し、提案手法の評価を行う．

2. 測定環境

本研究では、サーバ PC として、CPU が AMD Athlon 64 FX-74 3GHz (4 cores) × 2、主記憶が 8GB、HDD が Seagate Barracuda 3TB × 11(うち OS 用として 1 台)、OS が CentOS 5.10 64 ビット版、DBMS は Hitachi HiRDB Single Server Version 9 を使用する．また、電力計は YOKOGAWA WT1600 Digital Power Meter を使用する．

これらの測定環境は全て遠隔アクセスによる実験が可能である．電力計はサーバ PC の HDD に繋がれており、電力計は電力計操作 PC で操作する．またサーバ PC、電力計、電力計操作 PC は、ローカル PC と全てリモート接続されている．測定環境の簡単な模式図と、電力計の操作画面を図 1 に示す．

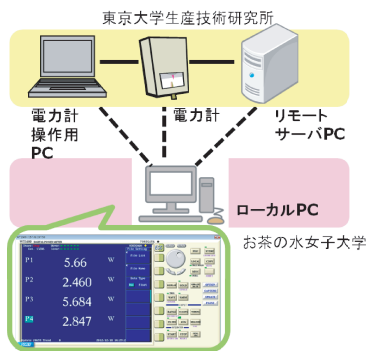


図 1 測定環境

3. 基本性能測定

3.1 ディスクアクセス時の性能測定

ディスクに対してシーケンシャルおよびランダムアクセスを行った際のディスクの消費電力とスループットを、1 秒毎に測定する．アクセスの方法は Read 処理、Write 処理の 2 種類で

測定を行う．測定対象のディスクは 1 台とし、アクセス単位 (バッファサイズ) を 4KB, 8KB, 16KB と変化させる．データの読み書き量はシーケンシャルの場合は 4GB, 8GB, 16GB(アクセス単位*1M 回)、ランダムアクセスの場合は 4MB, 8MB, 16MB(アクセス単位*1K 回) とする．I/O の方法としては Direct I/O を用いる．スループットは /proc/diskstats の内容を 1 秒毎に読み込んで測定する．

表 1, 2 は測定結果を示しており、消費電力とスループットはそれぞれ最大値を表している．シーケンシャルアクセス、ランダムアクセス共にバッファサイズが大きいくほど消費電力は大きく、スループットは向上した．これはバッファサイズが大きくなったことで一度に I/O を行うデータ量が増加したためである．また、シーケンシャルアクセス時と比較して、ランダムアクセス時の方が消費電力が大きいのは I/O を行う際のディスクアドレスが毎回異なるため、任意の位置に針を移動する際にエネルギーが必要だからである．これらのことから測定結果は妥当であると考えられる．

表 1 ディスクアクセス性能 (シーケンシャルアクセス時)

R/W	BufferSize(KB)	消費電力 (W)	throughput(MB/s)
Read	4	7.7	63.2
	8	7.9	102.9
	16	8.0	144.3
Write	4	7.5	59.3
	8	7.9	92.7
	16	8.3	131.1

表 2 ディスクアクセス性能 (ランダムアクセス時)

R/W	BufferSize(KB)	消費電力 (W)	throughput(MB/s)
Read	4	8.7	0.38
	8	8.8	0.76
	16	8.9	1.54
Write	4	7.6	2.5
	8	7.8	5.3
	16	9.0	7.8

3.2 TPC-H 実行時のディスクの性能測定

データベースベンチマークである TPC-H を動作させた際のディスクの消費電力と IOPS(Input Output Per Second) の測定を行う．測定対象のディスクは 2 台、HDD1 には TPC-H の LINEITEM テーブルを、HDD2 にはその他のテーブルを配置してある．DB の規模を決めるスケールファクタ (SF) を 1, 2, 3 と変化させ、それぞれの SF ごとに DB とクエリを用意して測定を行う．また、各項目共に 1 秒毎に取得する．

図 2 には SF=3 の時の HDD1、図 3 には SF=3 の時の HDD2 の測定結果を示す．SF=1, 2 の時もそれぞれ同様の傾向であった．上段が消費電力、下段が IOPS を表している．上段では、1 秒毎の消費電力の値がしばらく一定の部分が 2 箇所あり、こ

のとき下段の IOPS は一定して低い値を示している．この 2 箇所は、それぞれクエリプランが複雑で処理が重いクエリを処理している期間である．また、他の部分に関しても消費電力の増減と IOPS の増減のタイミングが一致していることがわかる．これらのことから、この測定結果は妥当であると考えられる．

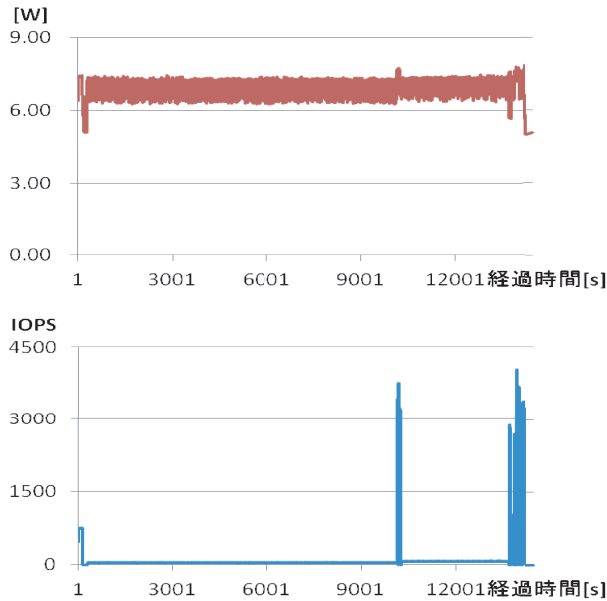


図 2 TPC-H 実行時のディスク性能 (HDD1, SF=3)

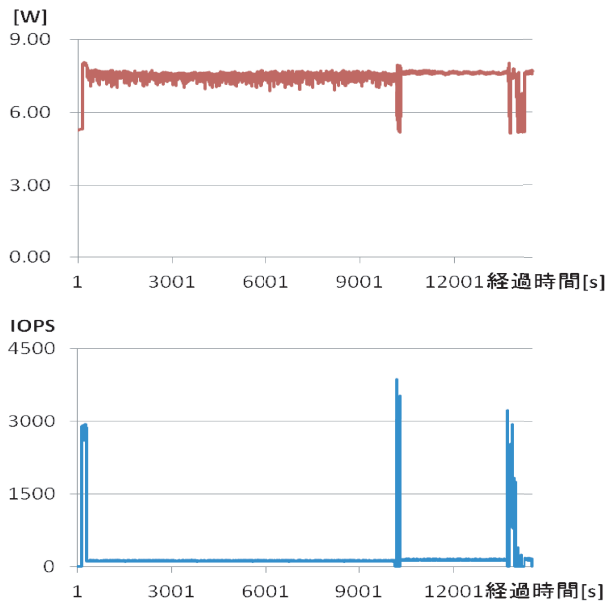


図 3 TPC-H 実行時のディスク性能 (HDD2, SF=3)

4. ディスクの消費電力特性

ディスクの遷移状態の種類と、各状態における消費電力を調査し、それに基づいて、省電力可能性の 1 つの指標となる、Break-Even Time を算出する．

4.1 ディスクの遷移状態と消費電力

本研究で使用したディスクの遷移状態は Standby1, Standby2, Idle, Active の 4 種類である．Idle/Active 状態から Standby1 状態に移行することを Spindown, Standby1 状

態から Idle/Active 状態に移行することを Spinup1, Standby2 状態から Idle/Active 状態に移行することを Spinup2 と呼ぶ[8]．

[8] では使用するディスクの状態を Standby, Idle, Active の 3 種類としているが、本研究で使用するディスクの遷移状態を詳細に調査したところ、Standby 状態時に消費電力が異なる 2 種類の期間がみられたため、本研究では 2 種類の状態を Standby1, Standby2 と区別している．

各状態におけるディスクの消費電力の測定を行った．測定対象のディスクは、4.2 節の測定に使用したものと同様のディスク 2 台である．Standby1 時、Standby2 時、Idle 時、Active 時の最大消費電力と、Spindown, Spinup1, Spinup2 に必要なエネルギーを表 3 に示す．

表 3 ディスクの遷移状態における消費電力とエネルギー量

Standby1(W)	Standby2(W)	Idle(W)	Active(W)
1.05	0.88	5.22	7.25
Spindown(J)	Spinup1(J)	Spinup2(J)	
6.79	108.5	105.5	

4.2 Break-Even Time

ディスクの Spindown 及び Spinup により消費されるエネルギーと、ディスクを Standby 状態に移行し、その状態を維持することにより削減できるエネルギーが等しくなる Standby 状態の持続時間を Break-Even Time と呼ぶ．これは Spindown に必要なエネルギーを E_d , Spinup2 に必要なエネルギーを E_{u2} , Standby1 状態の消費電力を P_{s1} , Standby2 状態の消費電力を P_{s2} , Idle 状態の消費電力を P_i , Spindown と Spinup2 に必要な時間をそれぞれ T_d, T_{u2} , Standby1, Standby2 状態の持続時間をそれぞれ T_{s1}, T_{s2} とすると、Break-Even Time T_{be} は、

$$T_{be} = (E_d + E_{u2} - P_{s2} * T_d - P_{s2} * T_{u2} + T_{s1} * (P_{s1} - P_{s2})) / (P_i - P_{s2})$$

により求めることができる．

本研究では Standby を 2 種類の状態に区別するため、Break-Even Time の算出式は[13]を参考に作成した．

本研究で用いた HDD1, HDD2 では Break-Even Time はそれぞれ約 24 秒であった．これより Standby 状態を利用して省電力化を実行するためには、ディスクへの I/O 発行間隔が HDD1, HDD2 それぞれのディスクにおいて約 24 秒以上必要である．

図 4 は今回使用したディスクにおいて Idle 状態から Standby1 状態を経て、Standby2 状態に移行した後、再び Idle 状態に移行した時の消費電力の推移を示している．

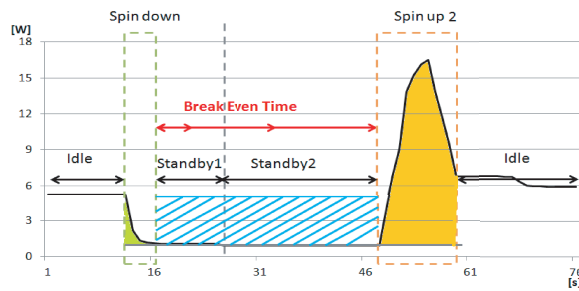


図 4 ディスクの状態遷移における消費電力と Break-Even Time

5. 省電力状態適用時の実行時消費電力量

本節では省電力状態適用時の TPC-H 実行時消費電力量を測定し、省電力状態適用前後での消費電力量を比較する。また、ディスクを省電力状態 (Standby 状態) に移行することをここでは省電力状態適用と呼ぶ。測定に使用した TPC-H の SF は 3、測定期間は TPC-H の実行開始から終了までとする。使用する HDD は 2 台、データ配置も 4.2 節と同じである。測定は 3 回行い、各測定の平均値を使用する。SF は 3 とする。

図 5 は省電力状態適用なし、省電力状態適用ありの場合の TPC-H 実行時における消費電力量の比較を示している。省電力状態を適用しなかった場合のディスク 2 台の消費電力量の合計値は 212,208J、省電力状態を適用した場合の消費電力量の実測値は 209,579J であり、削減率は約 1.2%であった。また、クエリの遅延は 125 秒であり、遅延率は約 0.8%であった。

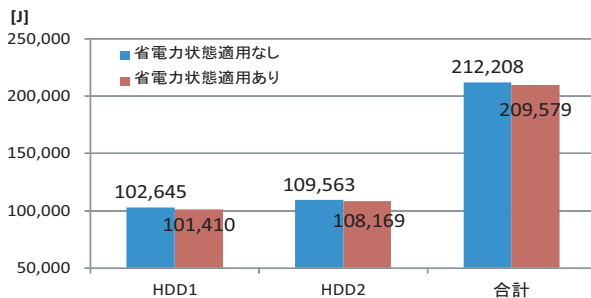


図 5 省電力状態適用による TPC-H 実行時消費エネルギー比較

クエリの遅延が発生しているのはディスクに省電力状態を適用したことによるオーバーヘッドが生じたためであると考えられる。この測定により、ディスクに対する I/O 発行間隔とディスクの省電力状態を適用することによって TPC-H 実行時の省電力化が可能であることが示された。

6. データ配置制御

これまで、我々は TPC-H 実行時のディスクに対する I/O 発行間隔と Break-Even Time を利用することによる TPC-H の実行時省電力化が可能であることを示した。しかし、これはアプリケーションに手を加えずにディスクの単純な振舞いを利用しているナイーブな手法である。

本節では、TPC-H 実行時の I/O 発行間隔を制御するという

目的のもと、TPC-H の各表と索引のバッファに対する I/O の状況を調査し、データの配置に手を加えることでさらなる実行時省電力化を目指す。また、ディスク 2 台と 3 台の 2 種類の環境で提案手法の評価を行った。

6.1 入出力状況の調査

本研究におけるストレージ省電力化のための提案手法である、データ配置制御の評価を行うために、TPC-H 実行時の各データに対する入出力状況を調査する。SF は 1, 2, 3 の 3 種類で、1 秒毎に pdbufstats (DMBS に付属する DB バッファ統計情報取得ツール) [14] を用いて入出力状況を取得する。DB は raw デバイス上に直接配置しており、調査対象のバッファは 23 個、調査期間は TPC-H クエリ実行開始から終了までとする。また、SF=1, 2, 3 の 3 種類において測定を行い、データは 1 秒毎に取得する。今回は I/O 状況を把握するために、得られるデータの項目の中でも実 READ 回数に注目し、解析を行う。通常の DBMS はバッファに DB の一部が載った状態 (Hot 状態) で使用されるため、本測定は Hot 状態で行うものとする。DB バッファのサイズは、すべての SF において表のデータが格納されたバッファの合計が約 0.58GB、索引のデータが格納されたバッファの合計が約 0.21GB である。DB のサイズは SF によって異なるため、表 4 に示す。また、本測定は 3. 節と同一の条件で行う。調査の結果、実行期間中に入出力が見られたデータが格納されたバッファは 13 個、入出力が見られなかったデータが格納されたバッファは 10 個であった。

表 4 DB サイズ (GB)

SF	1	2	3
表	1.38	2.75	4.13
索引	0.29	0.57	0.86

6.2 データ配置の変更

6.2.1 ディスク 2 台での評価

7.1 節で調査した結果に基づいて、データ配置を変更する。SF=1 ~ SF=3 のすべての場合において「実 READ 回数が 0 であったデータ」、「実 READ 回数に変化が見られたデータ」の 2 種類に分類し、前者を HDD1 に、後者を HDD2 に配置する。ここで使用する HDD1, HDD2 はこれまでの測定で使ったものと同様である。また、HDD1 には省電力状態を適用するものとする。データ配置のみを変更し、ディスクの省電力状態を適用した際の TPC-H 実行時における各 HDD の消費電力量とクエリの実行時間の測定を行い、データ配置変更前と変更後の各項目を比較する。測定はこれまでと同様、SF=1, 2, 3 の 3 種類について行う。また、ディスクがスタンバイ状態に移行するまでのタイムアウトを HDD1, HDD2 の 2 台とも 5 秒に設定する。測定期間は TPC-H 実行開始から終了までとする。消費電力量に関してはディスク 2 台の合計を比較対象とする。また、タイムアウトを 5 秒に設定したのはデータ配置を変更したことによって I/O 発行間隔が伸びたため、できるだけ早く HDD をスタンバイ状態に移行した方がより多く消費電力を削減することができると考えたからである。

図 6 は各 SF におけるデータ配置変更前後での消費電力の比較，図 7 はクエリ実行時間の比較を示している．消費電力の削減率は各 SF とも約 40%，クエリの遅延率は約 3～5%程度に収まった．これより，データ配置制御がアプリケーション実行時のストレージ省電力化に有効であることを示せた．

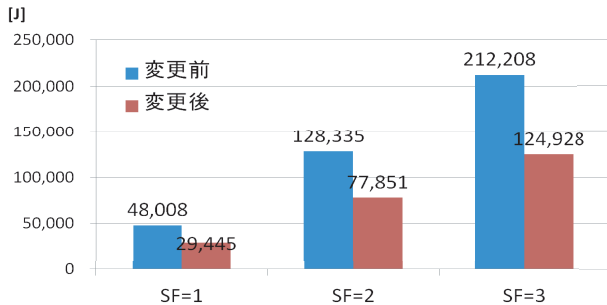


図 6 データ配置変更前後での消費電力比較 (ディスク 2 台)

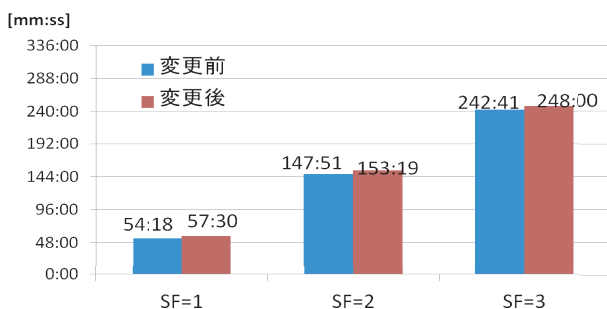


図 7 データ配置変更前後でのクエリ実行時間比較 (ディスク 2 台)

また，データ配置変更後におけるディスクに対する I/O トレースを，blktrace と btrecord(I/O トレース取得ツール) [12] を用いて取得・解析を行った．入出力が無いデータを配置した HDD1 にはクエリ実行中の I/O が見られず，入出力のあるデータを配置した HDD2 に I/O が集中していた．さらに，変更前後での iostat コマンドで HDD2 における I/O 期間中のディスク使用率を比較したところ，変更前は最高が約 95%，平均が約 66.7%であったのに対し，変更後は最高が約 100%，平均が約 95.3%であった．データ配置を変更したことにより，入出力のあるデータを配置したディスクの負荷が増加し，クエリの遅延が発生したと考えられる．以上のことから今回使用したデータ配置は適切であると考えられる．

6.2.2 ディスク 3 台での評価

提案手法をより大規模な環境で評価するため，ディスク 3 台での提案手法の評価を行う．7.2.1 節で使ったディスク 2 台に加えて，同じ種類のディスクを 1 台追加する．また，7.2.1 節においてディスク 2 台で提案手法の評価を行った際に，ディスク 2 台にデータ量を均等に配置した場合を提案手法の比較対象としていた．そのため，ディスク 3 台の測定でもできるだけ 7.2.1 と同じ条件に合わせたい．

そこで，まずデータをディスク 3 台に入出力頻度に関わらず，データ量ができるだけ均等になるように配置する．これを「データ配置変更前」とする．今回，データ配置変更前において HDD3

台に配置したデータ量の比は，HDD1:HDD2:HDD3=1:1:2 であった．HDD3 のデータ量が HDD1 および HDD2 の 2 倍になっているのは，各バッファに格納されたデータのうち，ある 1 つのバッファに格納されたデータが全体のデータ量の半分を占めており，これを HDD3 に配置したためである．続いて，提案手法の評価として，TPC-H 実行時において入出力が見られなかったデータを HDD1 に配置し，入出力が見られたデータについては，データ量を約半分にして HDD2 と HDD3 に配置する．これを「データ配置変更後」とする．

データ配置変更前と変更後において，これまでと同様に SF=1，2，3 の 3 種類で TPC-H 実行時のディスクの消費電力とクエリの実行時間を測定した．また，配置変更後の測定では入出力が見られなかったデータを配置している HDD1 は Standby 状態であるものとする．

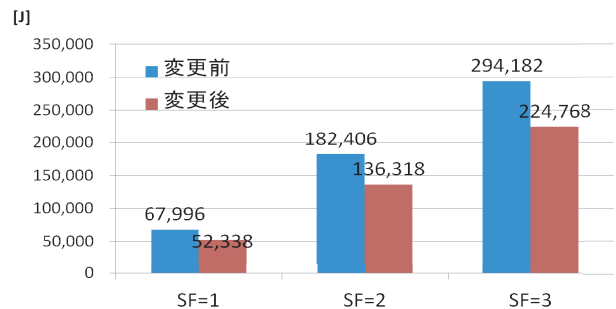


図 8 データ配置変更前後での消費電力比較 (ディスク 3 台)

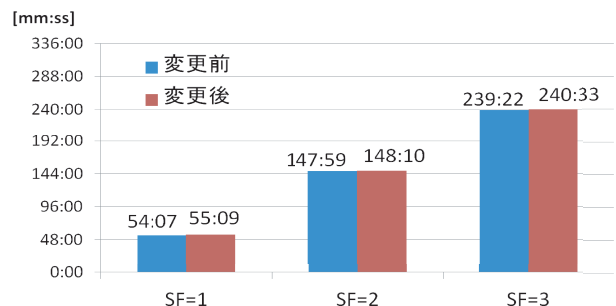


図 9 データ配置変更前後でのクエリ実行時間比較 (ディスク 3 台)

図 8 は各 SF におけるデータ配置変更前後での消費電力の比較，図 9 はクエリ実行時間の比較を示している．消費電力の削減率は各 SF とも約 22～25%，クエリの遅延率は約 0.5～1.5%程度に収まった．ディスク 2 台の場合と比較して削減率が少ないのはディスクの台数を増やし，入出力の見られるデータをディスク 2 台に配置したことにより，Idle/Active 状態のディスクが増えたためである．また，遅延率が減少した理由も同様であると考えられる．

さらに，ディスク 2 台の時と同様に iostat コマンドで I/O 期間中のディスク使用率の調査した．結果を表 5 に示す．

表 5 より，データ配置を変更した結果，入出力の見られないデータを配置した HDD1 のディスク使用率は 0%で，入出力の見られるデータを配置したディスクの使用率は上昇，またはほぼ変化がないことがわかる．また，データ配置変更後の消

表 5 ディスク使用率 (%)

		HDD1	HDD2	HDD3
変更前	平均	64.3	1.43	29.4
	最高	80.8	55.1	50.8
変更後	平均	0	65.8	29.3
	最高	0	95.8	51.1

費電力に関しても HDD1 は常に Standby 状態を保っており，HDD2 と HDD3 は常に Idle または Active 状態であった．以上のことから，今回設定したデータ配置は妥当であると考えられる．

7. まとめと今後の課題

本研究ではデータの効率的な管理という点からストレージの消費電力を削減することで，データセンタの省電力化が可能であると考えた．本稿では，既存研究をふまえて大規模システム環境での提案手法であるデータ配置制御の評価に向け，簡単なディスクアクセス時と TPC-H 実行時のシステム性能と消費電力量の解析を行った．また，ディスクの省電力状態が 2 種類あることを考慮し，その時のディスクの消費電力から，Break-Even Time を詳細に算出した．さらに，提案手法の評価として，TPC-H 実行時の各データに対する入出力状況に基づいてデータ配置を変更することにより実行時の消費電力が削減できることを示した．また，より大規模な環境を想定した場合でも本研究の提案手法が有効であることを示した．

今後の課題としては，現在行っているディスク 3 台でのデータ配置の検討について，I/O の間隔をより詳細に解析し，それに基づいてデータ配置を変更することによってさらなる省電力化が期待できると考えている．また，データを格納しているバッファをさらに分割することによる省電力化も期待できると考えている．これら検討も今後取り組みたい．

文 献

- [1] GIPC Survey and Estimation Committee Report FY2009 (Summary), <http://www.greenit-pc.jp/activity/reporting/100707/index.html>, 2009
- [2] TPC-H: <http://www.tpc.org/tpch/default.asp>
- [3] Jorge Guerra, Himabindu Pucha, Joseph Glider, Wendy Belluomini, Raju Rangaswami. Cost Effective Storage using Extent Based Dynamic Tiering. In 9th USENIX Conference on File and Storage Technologies, pp.1-14, 2011.
- [4] Dushyanth Narayanan, Austin Donnelly, Antony Rowstron. Write Off-Loading: Practical Power Management for Enterprise Storage. In 6th USENIX Conference on File and Storage Technologies, pp.253-267, 2008.
- [5] Athanasios E Papathanasiou and Michael L Scott. Energy Efficient Prefetching and Caching. In Proceedings of the annual conference on USENIX Annual Technical Conference, 2004.
- [6] Akshat Verma, Ricardo Koller, Luis Useche, and Raju Rangaswami. SRCMap: Energy Proportional Storage using Dynamic Consolidation. In Proceedings of the 8th USENIX Conference on File and Storage Technologies, 2010.
- [7] Charles Weddle, Mathew Oldham, Jin Qian, An-I Andy Wang, Peter Reiher, and Geo Kuenning. PARAD: A Gear-Shifting Power-Aware RAID. In 5th USENIX Conference on File and Storage Technologies, Vol. 3, pp. 245—260, October 2007.

- [8] 西川記史，中野美由紀，喜連川優:アプリケーション協調型大規模ストレージ省電力システムの開発と評価，DEIM Forum 2012，D6-1，2012 年 3 月
- [9] 西川記史，中野美由紀，喜連川優:アプリケーション処理の I/O 挙動特性を利用したディスクの実行時省電力手法とその評価: オンライントランザクション処理における省電力効果，電子情報通信学会論文誌，Vol.J95-D，No.3，pp.447-459，2012 年 3 月
- [10] Norifumi Nishikawa, Miyuki Nakano, and Masaru Kitsuregawa: Energy Efficient Storage Management Cooperated with Large Data Intensive Applications, 28th IEEE International Conference on Data Engineering (IEEE ICDE 2012), pp.126-137, 2012.04
- [11] 飯村奈穂，西川記史，中野美由紀，小口正人:データベース処理実行時における省電力化のためのストレージ制御手法の提案，DICOM シンポジウム 2013，p.1646-1652，7C-1，2013 年 7 月
- [12] Alan D. Brunelle: btrecord and bt replay User Guide, <http://www.cse.unsw.edu.au/aaronc/iosched/doc/btreplay.html>, 2007
- [13] Y.H. Lu, G.D.Micheli: Comparing System-Level Power Management Policies, IEEE Design & Test of Computers, Vol.18, No.2, pp.10-19, March 2001
- [14] pdbuffs: <http://www.hitachi.co.jp/Prod/comp/soft1/manual/pc/d635540/W3550027.HTM>