データインテンシブアプリケーション実行時のストレージ省電力に関する一検討

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

あらまし　近年デジタル情報量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台数も急増しており，ストレージの管理運用コストは見過ごせないものとなっている．さらに，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いる．これらのことからストレージにおけるデータの効率的管理に注目が集まっている．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た．また，データベースベンチマークである TPC-H 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い，ディスクの省電力状態を適用することによる TPC-H の省電力化が可能であることを示した．

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 Study on Runtime Disk Energy Saving of Data Intensive Applications

Naho IIMURA†, Norifumi NISHIKAWA††, Miyuki NAKANO††, and Masato OGUCHI†

† Ochanomizu University
2-1-1 Otsuka, Bunkyou-ku, Tokyo, 112-8610 JAPAN
†† Institute of Industrial Science, the University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 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

近年デジタル情報量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過ご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，ディスクの状態遷移における消費電力を調査し，ディスクの省電力状態を適用することによる TPC-H の省電力化が可能であることを見積と実測により示した．

2. 関連研究

関連研究として，アプリケーション協調型のストレージ省電力手法がある [3]- [5]．特に [4] 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, 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．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，クエリ遅延時間等，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．

3. 基礎性能測定

基本的なディスク性能を測定するために 2 種類の実験を行った．本節ではその結果と考察を述べる．3.2 では，簡単なディスク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測定する．3.3 では，データベースベンチマークである TPC-H 実行時のディスクの消費電力の測定と，ディスクの I/O トレースを取得し，分析する．

3.1 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 1TB × 6，OS が CentOS 5.6 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用 PC で操作する．またサーバ PC，電力計，電力計操作用 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．

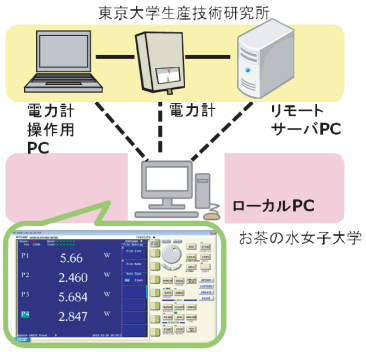

図 1 測定環境

3.2 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．表 1 より，シーケンシャルアクセスの場合，バッファサイズが大きいほど消費電力は大きく，スループットも向上した．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，Read 処理の方が Write 処理の場合よりもスループットが高いのは，ディスクの先読みによるものであると考えられる．

表 2 より，ランダムアクセスの場合は，消費電力・スループットはバッファサイズに関係なく同じであった．Read 処理より Write 処理の方がスループットが高いのは，ランダムアクセスを行っていることにより，先読みすることができないためである．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 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	9.4	33.65
	8	10.7	77.67
	16	10.7	77.77
Write	4	9.2	30.9
	8	9.7	52.94
	16	10.1	72.99

表 2 ランダムアクセス時

R/W	BufferSize(KB)	消費電力 (W)	throughput(MB/s)
Read	4	11.97	0.28
	8	11.92	0.55
	16	11.83	1.06
Write	4	9.69	0.45
	8	9.65	0.92
	16	9.62	1.79

3.3 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 I/O トレースの解析を行う．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．本計測では，blktrace と btrecord(I/O トレース取得ツール) [6] を用いて I/O トレースの取得・解析を行った．

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．上段は I/O トレースの取得結果を，下段は実行時の消費電力をクエリごとに色分けしたものを表している．結果より，前半よりも後半の方がアクセスが集中していることがわかる．これは後半のクエリはシーケンシャルアクセスを行っているため

である．また，I/O トレースと消費電力を比較すると，消費電力の上下とディスクアクセスの頻度が一致している．応答時間が長いクエリはランダムアクセスを行っていることがわかる．これはクエリプランが複雑であり，ランダムアクセスの方がシーケンシャルアクセスよりも時間がかかるためである．従って，この結果は妥当な振舞であると考えられる．

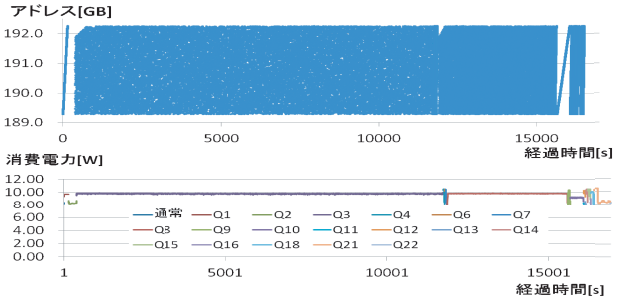

図 2　HDD1(SF=3)

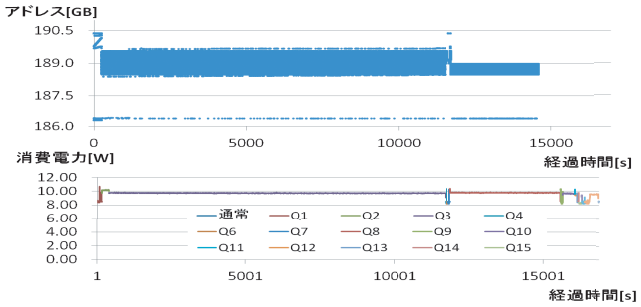

図 3　HDD2(SF=3)

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 [3]．

[3] 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 Stanby1，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，2.2 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．

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，

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

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 [7] 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 26 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 26 秒以上必要である．

図 4 は HDD1 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．

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

Disk	Standby1(W)	Standby2(W)	Idle(W)	Active(W)
HDD1	1.81	1.21	8.42	10.5
HDD2	1.92	1.24	8.43	10.8

Disk	Spindown(J)	Spinup1(J)	Spinup2(J)
HDD1	16.31	159.03	184.41
HDD2	13.77	181.39	180.05

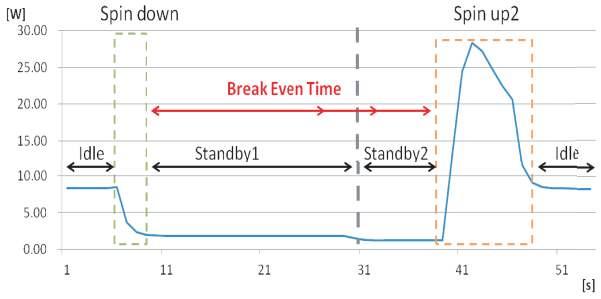

図 4　ディスクの状態遷移における消費電力と Break-Even Time(HDD1)

5.　実行時省電力可能性

本節では，前節をふまえて TPC-H 実行時の I/O 発行間隔を調査する．また，ディスクの省電力状態を適用することにより，TPC-H の実行時消費電力をどの程度削減することができるのか，という点について見積式をもとに実測を行い，省電力状態を適用した場合と，そうでない場合の実行時消費電力量を比較し，評価を行う．

5.1 I/O 発行間隔

ディスクに I/O が行われてから，次の I/O が発行されるまでの時間を I/O 発行間隔と呼ぶ．本研究では，TPC-H 実行時のディスク I/O の利用状況を取得・解析し，I/O 発行間隔を調査する．測定環境は 3 節同様で，調査対象のディスクは前節と同様のディスク 2 台とする．測定期間は TPC-H クエリ実行中で、時間は 16,903 秒 (4 時間 41 分 43 秒) とする．

測定の結果，図 5，6 に示すように，I/O 発行間隔が Break-Even Time 以上である回数が HDD1 では 4 回，HDD2 では 10 回であった．また，I/O 発行間隔が Break-Even Time 未満である回数は HDD1 では 8 回，HDD2 では 11 回であった．ここでは，I/O 発行間隔が Break-Even Time 未満のうち，最短 1 秒以上のものをカウントしている．Break-Even Time 以上の I/O 発行間隔のうち，最長は HDD1 では 322 秒，HDD2 では 157 秒であった．

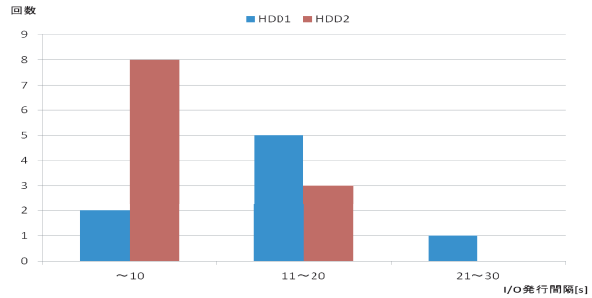

図 5　Break-Even Time 未満の I/O 発行間隔の回数

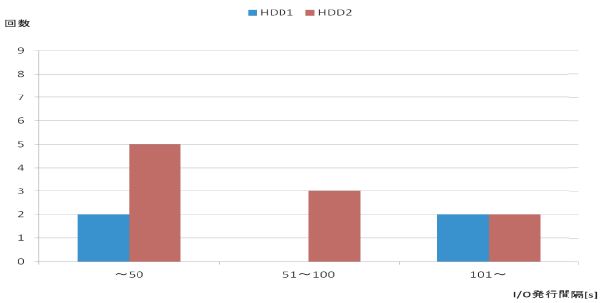

図 6　Break-Even Time 以上の I/O 発行間隔の回数

5.2 省電力状態適用時の実行時消費電力量

5.2.1 削減可能エネルギー見積

ディスクのスタンバイ状態を利用する場合を省電力状態適用あり，利用しない場合を省電力状態適用なしとする．さらに，省電力状態適用ありの場合は，ディスクをスタンバイ状態に移行するまでのタイムアウトを設定すると仮定して，実行時削減可能エネルギーの見積を行う．タイムアウトを設定することにより，タイムアウト時間より長くディスクへの I/O が発行されなかった場合に，ディスクを省電力状態へ移行することとする．すなわち，本研究では,【I/O 発行間隔 (s) －タイムアウト時間 (s)】を省電力状態適用期間とする．また，Spinup の契機は省電力状態 (Standby1，Standby2) 時にディスクに I/O が発行された時とする．

しかし，タイムアウトを設定する場合，Break-Even Time より短い I/O 発行間隔にも省電力状態を適用することになるため，省電力状態適用期間によって見積式を選択する必要がある．また，Break-Even Time より短い I/O 発行間隔に省電力状態を適用する場合，削減可能エネルギーがマイナスになるため，削減可能エネルギーの見積式ではタイムアウト時間も考慮する必要がある．提案する見積式では，Spindown に必要な時間，Standby1 状態の持続時間の合計時間と，省電力適用期間を比較することにより，見積式を選択する．3.2 節で使用した項目に加えて，I/O 発行間隔を T_i，設定するタイムアウトを T_t とするとき，それぞれの I/O 発行間隔における削減エネルギー E_s は，

- 省電力適用期間　$< T_d+T_{s1}$ の場合

$$E_s = (T_i - T_t) * P_i - E_d - E_{u1} + P_{s1} * (T_i - T_t - T_d)$$

- 省電力適用期間 $\geqq T_d+T_{s1}$ の場合

$$E_s = (T_i - T_t) * P_i - E_d - E_{u2} + P_{s1} * T_{s1} + P_{s2} * T_{s2}$$

により求めることができる．I/O 発行間隔ごとに算出した値の合計値を，削減可能エネルギーの見積値とする．

表 4 は，ディスクにタイムアウト時間を設定した場合の TPC-H 実行時の削減可能エネルギーを，見積式により算出した値を示している．タイムアウトが 20 秒以上の場合には，タイムアウトの増加に伴い，見積値が減少していたため，ここではタイムアウトが 20 秒までの見積値を載せている．表 4 より，削減可能エネルギーが最も大きいタイムアウト時間は，HDD1 では 15 秒 (3498.29J)，HDD2 では 10 秒 (2547.81J) であることがわかる．

表 4　実行時削減可能エネルギー (J)

Timeout(s)	5	10	15	20
HDD1	3224.64	3403.89	3498.29	3368.55
HDD2	2489.51	2547.81	2228.56	2028.73

クエリの遅延時間についても見積によって求めることができる．今回の見積では，Spinup の契機をディスクに I/O が発行された時としているため，ディスクの起動 1 回に必要な時間 (Spinup1，Spinup2 に必要な時間) の合計をクエリの遅延時間とすることができる．よって，各ディスクのクエリの遅延時間 T_{late} は，

$$T_{late} = T_{u1} * Spinup1 \text{ が行われた回数} + T_{u2} * Spinup2 \text{ が行われた回数}$$

により求めることができる．

タイムアウトを HDD1 では 15 秒，HDD2 では 10 秒に設定した際の，TPC-H 実行時のクエリ遅延時間の見積値は，HDD1 では 34.6 秒，HDD2 では 99 秒，全体の遅延時間は 133.6 秒であった．

5.2.2 省電力状態適用時の実行時消費電力量

見積式から得られた最適なタイムアウトを各ディスクに設定し，TPC-H 実行時の消費電力量を測定する．最適なタイムアウトとは，HDD1 では 15 秒，HDD2 では 10 秒を指す．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．測定値は 3 回の測定の平均値を使用する．

図 7 は省電力状態適用なし，省電力状態適用ありの実測値，省電力状態適用ありの見積値の TPC-H 実行時の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消費電力量は 164,794J，省電力状態を適用した場合の消費電力量の実測値は 162,135J，省電力状態を適用した場合の消費電力量の見積値は 161,296J であった．

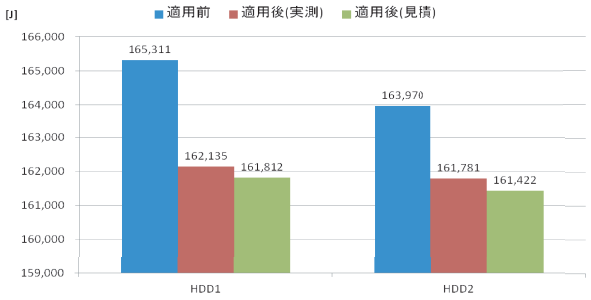

図 7　省電力状態適用による TPC-H 実行時消費エネルギー比較

クエリ遅延時間の実測値は 125 秒であった．見積値と比較して誤差が生じているが，これは 2 台のディスクに異なるタイムアウトを設定したことにより，TPC-H 実行中に，片方のディスクの処理を待つ等の動作が生じ，Spinup のタイミングや時間に誤差が生じたためであると考えられる．

消費エネルギー削減率の見積値は，HDD1 では 2.1%，HDD2 では 1.6% であるのに対し，実測値は HDD1 では 1.9%，HDD2 では 1.3% であり，誤差は 0.2〜0.3%に収まった．

TPC-H 実行時の I/O 発行間隔の長さや回数，ディスクの消費電力は毎回若干異なるため，見積値と実測値の誤差は許容範囲であると考えられる．これらのことから，ディスクの省電力状態と I/O 発行間隔を利用した TPC-H の省電力化は可能であるといえる．

5.2.3 削減可能エネルギー見積式の整合性

5.2.1 節で提案した，実行時削減可能エネルギーの見積式の整合性を示すために，5.2.2 節でディスクに設定したタイムアウト以外の値をタイムアウトとしてをディスクに設定し，同様の測定を行う．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し，測定値は 3 回の測定の平均値とする．ディスクを省電力状態に移行するまでのタイムアウトを，HDD1 には 5 秒，10 秒，HDD2 には 5 秒，15 秒のタイムアウトをそれぞれ設定し，TPC-H 実行時の消費電力量を測定する．

図 8 は，HDD1 に 5 秒，10 秒のタイムアウトを設定した時の，省電力状態適用なし，省電力状態適用ありの場合の実行時消費エネルギーの見積値，実測値を表している．図 9 は同様に HDD2 に 5 秒，15 秒のタイムアウトを設定した時の測定値を

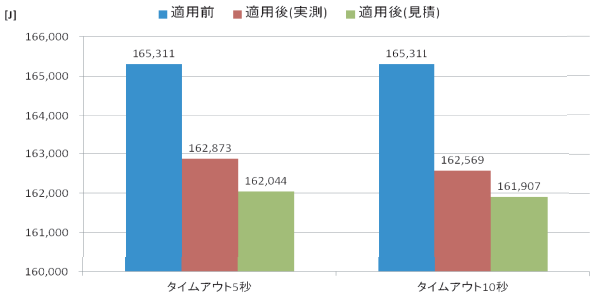

図 8　タイムアウト設定時消費エネルギー比較 (HDD1)

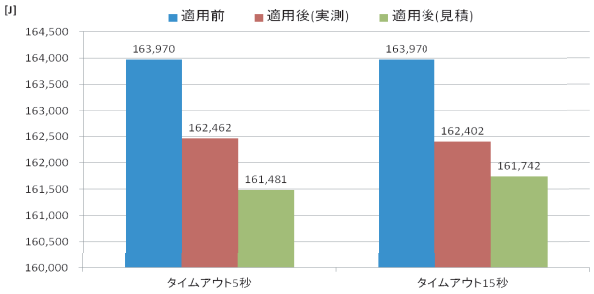

図 9　タイムアウト設定時消費エネルギー比較 (HDD2)

表 5　消費エネルギー削減率 (%)(HDD1)

Timeout(s)	5	10
見積値	1.9	2.0
実測値	1.5	1.7

表 6　消費エネルギー削減率 (%)(HDD2)

Timeout(s)	5	15
見積値	1.5	1.4
実測値	0.9	1.0

表している．2 つの図より，タイムアウトを 5.2.2 節で設定した値以外のものを設定した場合も，TPC-H 実行時の消費エネルギーを削減することが可能であることがわかる．

HDD1，HDD2 においてそれぞれタイムアウトを設定した場合の実行時消費エネルギー削減率の見積値と実測値の比較を表 5，6 に示す．見積値と実測値の間における誤差はそれぞれ約 0.3〜0.8%程度に収まっている．

この誤差は，ディスクにタイムアウトを設定したことによって起動オーバヘッド等の待ち時間が生じ，I/O 発行間隔の回数と長さにずれが生じたためであると考えられる．本研究では，TPC-H の省電力化を目標としており，見積式の整合性はある程度保たれていれば良いものとする．従って，本研究で提案した，TPC-H 実行時削減可能エネルギーの見積式は妥当であると言える．

6. まとめと今後の課題

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，データベースベンチマークである TPC-H の実行時省電力化に向けて，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．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．また，TPC-H 実行時の I/O 発行間隔を調査し，ディスクの省電力状態を利用することによって実行時の消費電力量

が削減可能であること，提案した見積式が妥当であることを実測によって示した．

ただし，今回の検討はディスクの基本性能や消費電力，TPC-H 実行時のディスク I/O の自然な振舞を考察しているだけであり，I/O 発行間隔とディスクの省電力状態を考慮して削減することができた消費電力も数%と微量である．そのため，研究方針である，アプリケーションの性能劣化を最小限に抑えたストレージの省電力化の提案としては十分ではない．今後は，見積式の整合性をさらに高めるために，TPC-H 実行時に使用する DB の規模をさらに大きくし，削減可能エネルギーが数十%になる場合の見積値と実測値の比較も行いたい．さらに，データ配置の調整などによる，TPC-H 実行時の I/O 発行間隔の制御など，アプリケーション実行時のストレージ省電力化に向けてさらに実践的な取り組みを行っていきたい．

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] 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

[4] 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

[5] 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

[6] Alan D. Brunelle: btrecord and btreplay User Guide, http://www.cse.unsw.edu.au/ aaronc/iosched/doc/btreplay.html, 2007

[7] Y.H. Lu, G.D.Micheli: Comparing System-Level Power Management Policies, IEEE Design & Test of Computers, 2010