データベースベンチマーク実行時のストレージ省電力手法に関する評価

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　　‡ 東京大学生産技術研究所

1. はじめに

近年，情報爆発に伴いデータセンタの大規模化が進み，データセンタの管理運用コストは見過ごせないものとなっている．特に電力コストの削減は急務であり，一定の消費電力割合を占めるストレージの省電力化に注目が集まっている [1]. 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，大規模システム環境において TPC-H 実行時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．また，1 つの指標となる Break-Even Time の詳細な解析と，提案手法であるデータ配置制御の評価を行った．

2. 研究方針

関連研究である [2] では，データベースベンチマークである TPC-H 実行時のストレージ省電力化手法としてデータ配置制御を提案し，評価している．しかし，実験環境として一つの小規模なユニットを用いているため，大規模システム環境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た場合の実行時ストレージ省電力化に向けて，本研究で提案するデータ配置制御手法の評価を行うことを研究方針とする．

3. 基礎性能測定

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 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．

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．測定対象のディスクは今後使用する予定の 3 台とし，アクセス単位 (バッファサイズ) を 4，8，16KB と変化させ，シーケンシャルの場合はアクセス単位*1M 回，ランダムアクセスの場合はアクセス単位*1K 回の I/O を行う．I/O の方法としては Direct I/O を用いる．測定は 3 回行い，平均値を取得する．

測定結果として，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，ランダムアクセス時の方が消費電力が大きかったが，これはディスクアドレスが毎回異なるため，ヘッドを動かすのに電力を必要とするためである．これらのことから，測定結果は妥当であると考えられる．

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 IOPS 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．

各クエリの 1 秒あたりの消費電力と IOPS を照らし合わせると，IOPS が小さいクエリは実行時間が長いことがわかった．これはクエリプランが複雑で処理が重いためである．また，測定した 3 種類の SF すべてにおいて同じ傾向であった．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．

4. ディスクの消費電力特性

4.1 ディスクの遷移状態と消費電力特性

ディスクの遷移状態には Idle，Active，Standby，Sleep の 4 種類がある．本研究では使用するディスクの遷移状態を Idle，Active，Standby1，Standby2 の 4 種類とする．Idle/Active から Standby1 に移行することを Spindown，Standby1/Standby2 から Idle に移行することをそれぞれ Spinup1/Spinup2 と呼ぶ．Standby 状態に関しては，本研究で使用したディスクの遷移状態を詳細に調査したところ，消費電力の異なる 2 種類の状態が見られたため，区別している．今後使用する予定のあるディスク 2 台について各遷移状態における消費電力の平均値を測定したところ、結果はどれもほぼ同様であった．各状態における消費電力の平均は Standby1 では 1.05W，Standby2 では 0.88W，Idle では 5.22W，Active では 7.25W であった．Spindown に必要なエネルギーは 6.79J，Spinup1 では 108.5J，Spinup2 では 105.5J であった．

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

A Evaluation of Storage Power Saving Method on Runtime Database Benchmark
† Naho Iimura, ‡ Norifumi Nishikawa, ‡ Miyuki Nakano, † Masato Oguchi
Ochanomizu University (†) , Institute of Industrial Science, The University of Tokyo (‡)

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 [4].

本研究で使用している HDD 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 発行間隔がそれぞれのディスクにおいて約 24 秒以上必要である．図 1 は今回の測定で使用したディスクが Idle 状態から Standby1 状態，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．

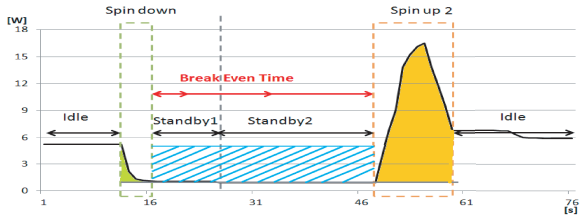

図 1: ディスクの状態遷移における消費電力と Break-Even Time

5. 提案手法の評価

5.1 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．使用する HDD およびデータ配置は 3.3 節と同様とする．SF は 1, 2, 3 の 3 種類で，1 秒毎に pdbufls(DBMS 付属のバッファ情報表示ツール) を用いて入出力状況を取得する．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．

5.2 データ配置の変更

5.1 節で調査した結果に基づいて，データ配置を変更する．HDD1 には TPC-H 実行中に入出力が見られなかったデータを配置し，HDD2 には入出力が見られたデータを配置する．また，HDD1 には省電力状態を適用するものとする．配置前後での TPC-H 実行時における消費電力と実行時間の比較を行った．図 2 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，今回使用したデータ配置制御が，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．

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)[5] 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 が見れられず，

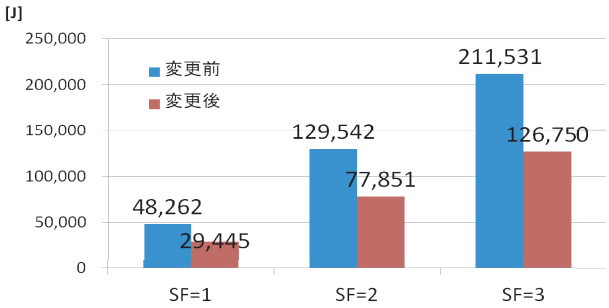

図 2: データ配置変更前後での消費電力比較

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．

以上のことからも今回使用したデータ配置は適切であると考えられる．

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，大規模システム環境での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，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，I/O 発行間隔を利用することによる，TPC-H の省電力化が可能であることを示した．今後はディスクの台数を増やし，さらに大規模システム環境を想定した評価を行いたい．

参考文献

[1] GIPC Survey and Estimation Committee Report FY2009 (Summary): http://www.greenit-pc.jp/activity/reporting/100707/index.html

[2] 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，DICOMO シンポジウム 2013, p.1646-1652，7C-1， 2013 年 7 月

[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] 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

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