Vol.2015-DPS-162 No.1
Vol.2015-CSEC-68 No.1
2015/3/5

動作 HDD 数制御を用いた仮想化環境における データ再配置によるストレージ省電力化

若色匠[†1]　谷貝俊介[†2]　山口実靖[†3]

近年，情報技術が普及しデータセンター等において多数のサーバ計算機が稼動するようになった．これに伴い，サーバ計算機の消費電力の増加が問題となっている[1]．この問題に対する解決策の一つとして，アプリケーションの動作情報を用いてディスク上のデータレイアウトを変更し，HDD の消費電力を削減する手法がある．本研究では当該手法の仮想化環境下への適用について考察する．具体的には，仮想化環境にて TPC-E を実行する環境において，各テーブルデータへのアクセス間隔を調査し，HDD の I/O 処理の能力を超えない範囲でアクセス間隔の長いテーブルデータを特定の HDD にまとめ，その HDD 停止時間の拡大を図る．そして，HDD 使用率を考慮した配置手法を提案し，性能評価によりその有用性を示す．

1. はじめに

データセンターにて膨大な数の計算機が稼働しており，多くの電力が消費されている．ストレージ機器はその中でも消費電力が大きい装置の一つであり，この電力消費の削減は重要な課題の一つとなっている．この問題に対する解決策の一つとして，アプリケーションの動作情報を用いてディスク上のデータレイアウトを変更し，HDD の消費電力を削減する手法がある[2].

本研究では仮想化環境における HDD の消費電力の削減に着目し，HDD の使用率を考慮したデータ配置手法を提案する．そして，性能評価によりその有効性を示す．具体的には，仮想化環境下での TPC-E の各テーブルデータから HDD へのアクセス頻度を調査する．そして，長いアクセス間隔が多く，短いアクセス間隔の少ないテーブルデータを 1 つの HDD にまとめ，HDD 停止時間がどの程度確保できるか，スループットの低下はどの程度起こるか調査する．また HDD において停止時間を設定し，停止によるスループットの低下はどの程度か，停止によって生じる消費電力減少の程度を調査する．

2. 既存研究

ストレージ省電力化手法の一つに，応用(アプリケーション)情報を用いたデータレイアウト変更手法[1]がある．当該手法では，データ(テーブル)のアクセス頻度を考慮しディスクへのデータ配置を制御することによりディスクの省電力機能を適用できるだけの I/O 発行間隔を生成している．アクセス数が多いデータを Hot データ，アクセス数が少ないデータを Cold データと呼び，この Cold データを 1 つの HDD に集中させることでアクセス間隔の拡大させ省電力化を図っている [3]．図 1 に応用情報を用いたデータレイアウトの変更法について示す．

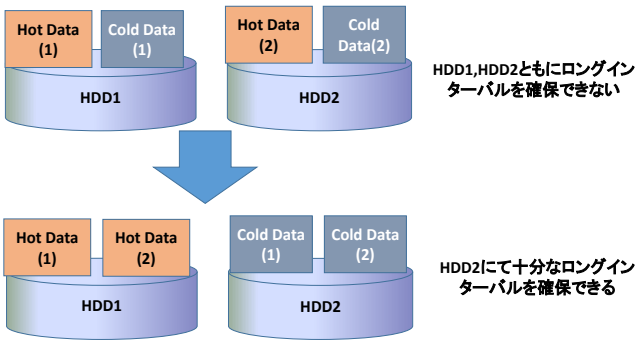


図 1　応用情報を用いたデータレイアウトの変更

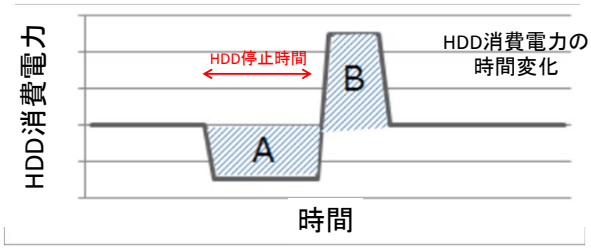


図 2　停止と再起動時の電力の変化

図 2 にてストレージの停止と再起動時の電力の変化について示す．ストレージ停止により削減できる電力量(A)とストレージ再稼働により失われる電力量(B)が等しくなる(A=B)ストレージ停止時間をブレークイーブンタイムと呼び，それより長くなる HDD アクセス間隔(A>B)をロングインターバルと呼び，上記手法ではロングインターバルを作り出すことで，省電力化を実現している．

ブレイクイーブンタイムは使用した HDD により異なり，文献[3，4]ではそれぞれブレークイーブンタイムが 25 秒，10 秒と定義されている．

†1 工学院大学大学院工学研究科電気・電子工学専攻
Electrical Engineering and Electronics, Kogakuin University Graduate School
†2 工学院大学大学院工学研究科電気・電子工学専攻
Electrical Engineering and Electronics, Kogakuin University Graduate School
†3 工学院大学工学部情報通信工学科
Department of information and Communications Engineering, Kogakuin University

Vol.2015-DPS-162 No.1
Vol.2015-CSEC-68 No.1
2015/3/5

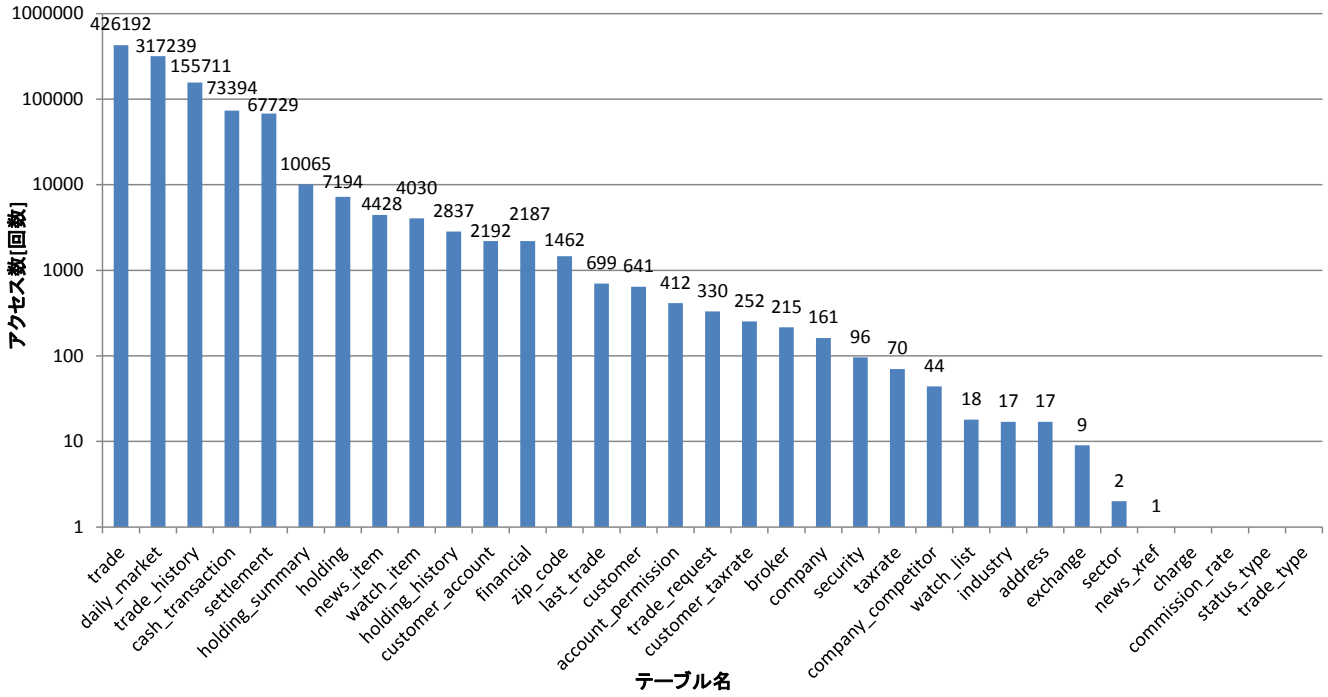


図 3　各テーブルデータへのアクセス数

3. 基本調査

3.1 アクセス頻度調査

基本調査として 1 台の物理計算機上に 1 台の VM を稼働させ, VM 上でベンチマークソフト tpcemysql を 2 時間実行させ，各テーブルデータへのアクセス数とアクセス頻度を調査した．調査は表 1 の環境にて行った．アクセス頻度は Linux カーネルの SCSI サブシステム内で観察した．よってページキャッシュなどによりストレージアクセス(およびそれに伴う電力消費)を発生させないアクセスは計測に含まれていない．ただし，データサイズの合計はゲスト OS メモリの約 16 倍であり, ページキャッシュはほぼヒットしないようになっている．

各テーブルデータへのアクセス数を図 3 に示す．図 3 よりテーブルデータへのアクセス数には大きな偏りがあり，

表 1　提案手法評価の測定環境

CPU	Intel Celeron CPU G1101 2.27[GHz]
MEMORY	4[GB]
HDD(OS用)	VB0160EAVEQ　160[GB]
HDD(データ用)	WD5000AZRX-0　500[GB]
OS(共通)	CentOS6.3　x86_64
カーネル(共通)	2.6.32.57
ホストOSメモリ	2[GB]
ゲストOSメモリ	512[MB]
仮想HDD	100[GB]
合計データサイズ	7.8[GB]

表 2　各テーブルデータのアクセス間隔

テーブル名＼アクセス間隔[秒]	0～10	10～30	30～100	100～
trade	426191	0	0	0
daily_market	317238	0	0	0
trade_history	155710	0	0	0
cash_transaction	73393	0	0	0
settlement	67728	0	0	0
holding_summary	9978	86	0	0
holding	6973	220	0	0
news_item	4372	55	0	0
watch_item	3974	55	0	0
holding_history	2570	233	33	0
customer_account	2031	93	67	0
financial	1980	203	3	0
zip_code	1202	246	13	0
last_trade	464	134	100	0
customer	397	185	58	0
account_permission	209	116	86	0
trade_request	87	152	90	0
customer_taxrate	180	17	26	28
broker	1	105	108	0
company	68	20	49	23
security	31	9	29	26
taxrate	16	4	16	33
company_competitor	1	1	12	29
watch_list	4	1	1	11
industry	0	0	0	16
address	2	2	3	9
exchange	0	0	0	8
sector	1	0	0	0

Vol.2015-DPS-162 No.1
Vol.2015-CSEC-68 No.1
2015/3/5

trade などの多いものでは 40 万以上，exchange などの少ないものでは 10 回以下であり 1 度もアクセスの無い charge などのテーブルデータもあった．

各テーブルデータへのアクセス頻度を表 2 に示す(表 2 にはアクセス数が 1 以下のテーブルはアクセス間隔が存在しないため省いてある)．表 2 より，30 秒以下のアクセス間隔が 10 回以下である company_competitor よりアクセス数の少ないテーブルは VM 数を増やしたとしても短い間隔でのアクセスが増えないと考えられるので，ロングインターバルを確保できると考えられる．また，100 秒以上の間隔が比較的大きく見られ，30 秒以下の回数が少ない company，security，taxrate もロングインターバルを得られるのではないかと考えられる．逆に，10 秒以下の間隔しかない watch_item より上のテーブルデータは短いアクセスが多くロングインターバルが存在しないため，これらのテーブルが 1 つでも存在すると HDD の停止時間を得ることが不可能であることが分かる．また，30 秒以上の間隔がある broker や trade_request より上のテーブルは 100 秒以上の間隔が無く 30 秒以下のアクセスも多く見られるため，ロングインターバルを得ることは難しいと考えられる

3.2 HDD の I/O 処理性能(Imax)の調査

3.1 の環境において HDD の 1 秒あたりの I/O 処理性能である Imax(SCSI 層への要求量/測定時間)を調査した．測定した結果，本実験で用いる HDD の Imax は 329[回数/sec]であることがわかった．

3.3 ロングインターバルの調査

HDD をスピンダウンすると一時的に消費電力を下げることができるが，スピンアップ時に一定の大きな電力消費が発生する．スピンダウン時に減少した電力がスピンアップ時に増加した電力より多くなる時間がロングインターバルである．表 3 の環境でロングインターバルの調査を行った．ただし，“WD5000AZRX-0”は， hdparm コマンドによるスピンダウンの設定をすることができなかったため．HDD に“VB0160EAVEQ”を用いて行った．

図 4 は今回使用した HDD のスピンダウンからスピンアップまでの電力をワットモニターを用いて調査した結果で

表 3　アクセス頻度調査の測定環境

CPU	Intel Celeron CPU G1101 2.27[GHz]
MEMORY	12[GB]
HDD(OS用)	VB0160EAVEQ　160[GB]
HDD(データ用)	WD5000AZRX-0　500[GB]
OS(共通)	CentOS6.3　x86_64
カーネル(共通)	2.6.32.57
ホストOSメモリ	2[GB]
ゲストOSメモリ	3[GB]
仮想HDD	30[GB]
合計データサイズ	6.5[GB]

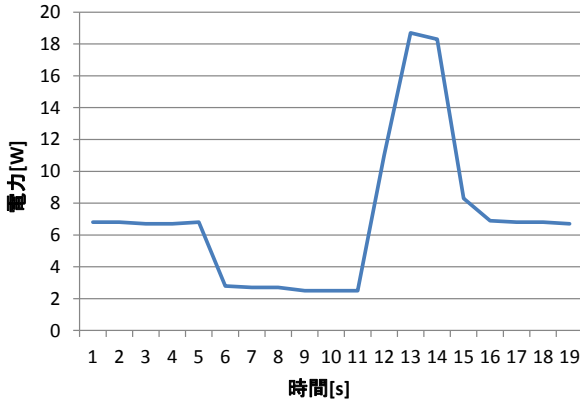
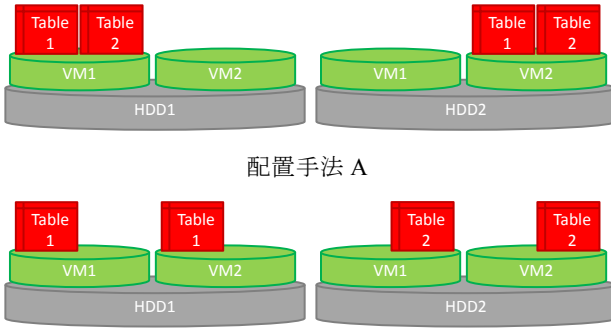


図 4　スピンダウンからスピンアップまでの推移



配置手法 A

配置手法 B

図 5　テーブル配置手法

ある．図 4 より，5 秒の位置でスピンダウンが行われたことで消費電力が半分以下になったが，11 秒の位置でのスピンアップにより一時的に倍以上の消費電力が発生していることがわかる．スピンアップの際に生じる消費電力は一定であり，この HDD では 8 秒以上のアクセス間隔がロングインターバルとなる．

4. VM イメージファイル配置と HDD 使用率

複数の HDD に複数の VM のファイルを配置する場合，特定の VM のファイル群を複数の HDD に分散配置すると，I/O バウンドの処理であっても I/O 使用率の低下をまねきアプリケーション性能が低下することがある．換言すると，ある VM のファイル群を特定の HDD に全て格納すると，(その VM の処理が I/O バウンドであれば)その HDD へのアクセス要求は常に発行され，その HDD の使用率は常に高くなる．

TPC-E の テーブルファイルの配置を図 5 に示す．図内の VM 上の四角の Table1 と Table2 は TPC-E のテーブル群 1 とテーブル群 2 である．配置手法 B では HDD へのアクセス数がなるべく均等になるようにテーブル群を分けている．表 1 の環境において図 5 の配置手法 A の様に(1 つの VM のテーブルを 1 つの HDD に集中)配置した場合と，配置手法 B のように(1 つの VM のテーブルを複数の HDD に

Vol.2015-DPS-162 No.1
Vol.2015-CSEC-68 No.1
2015/3/5

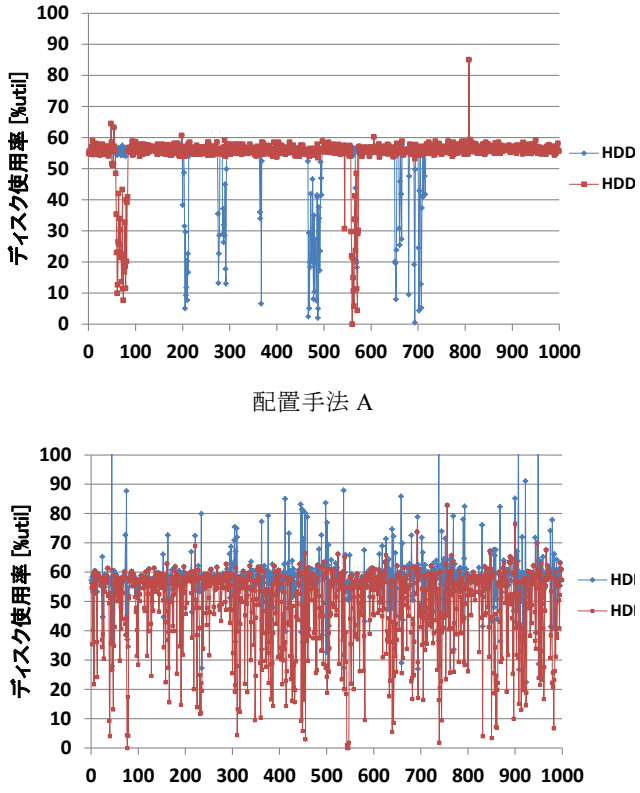


図 6　ディスク使用率

ファイルを分散)配置した場合の I/O 使用率と TPC-E 性能を図 6，図 7 に示す．

図 6 より，配置手法 A では常に I/O 使用率が高いが，配置手法 B では I/O 使用率が低下することがあることが分かる．配置手法 A では，I/O バウンドであるアプリケーションが常にそれぞれの HDD に I/O 要求を発行し続け，結果として HDD の使用率が低下することがない．これに対して配置手法 B では，両 VM が偶然同時に同一の HDD に対して I/O 要求を発行することでもう片方の HDD がアイドル状態(または使用率が低い状態)となる．図 7 より，アイドル状態(または使用率が低い状態)の少ない配置手法 A の方がスループットが優れていることがわかる．

5.　データベーステーブル再配置手法

5.1　再配置手法

前述のように TPC-E ではデータベーステーブルが複数作られ，またアクセス頻度もテーブルによって大きく異なり，ロングインターバルの存在するテーブルも複数存在する．本研究ではロングインターバルのあるテーブルを 1 つの HDD にまとめることによりその HDD のアクセス間隔を拡大し，HDD のロングインターバルを確保することにより停止による省電力化を実現する．

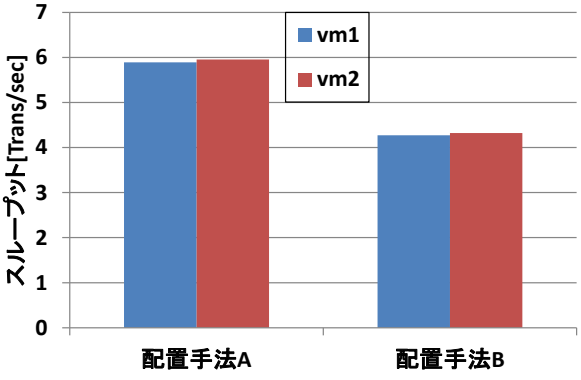


図 7　スループット

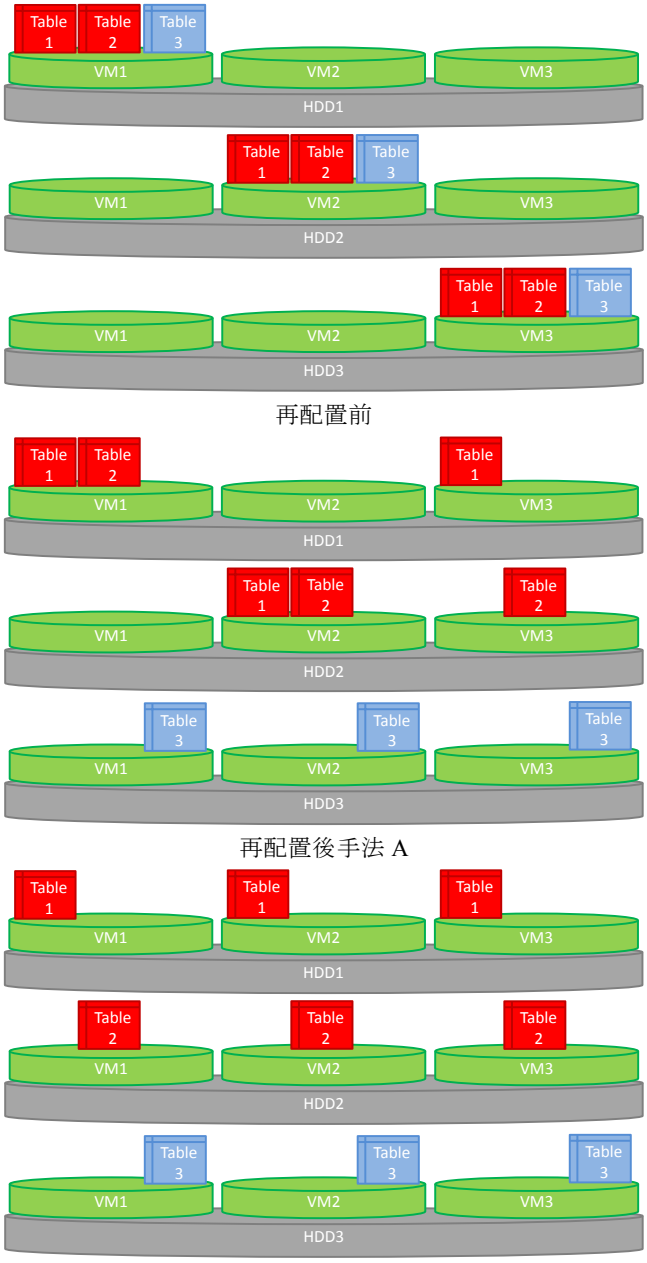


図 8　テーブル配置手法

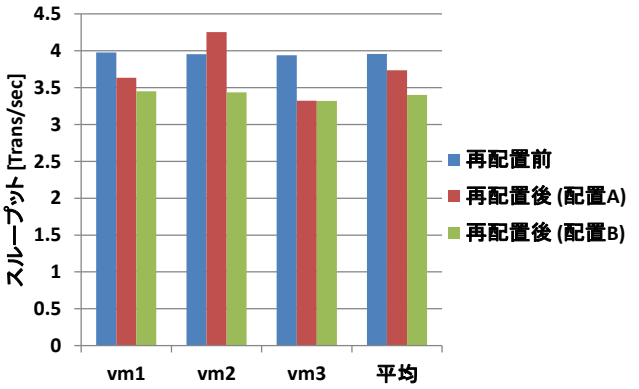


図 9　再配置前後のスループット

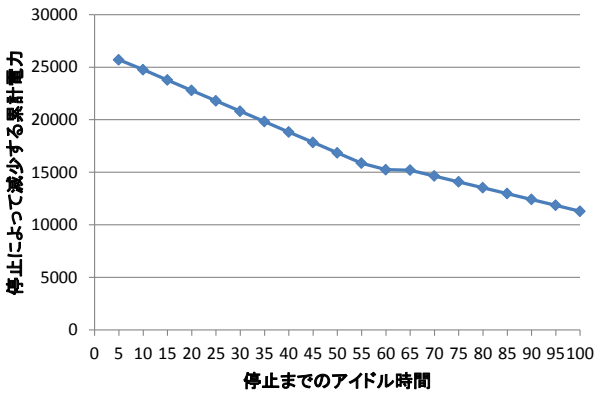


図 10　停止によって減少する累計電力と
アイドル時間の関係図

5.2 性能評価

提案手法の有効性を確認するための準備調査として，性能測定を表 3 の環境にて行った．表 1 と比べてメモリを 3[GB]に増やし，合計のデータサイズを 1[GB]ほど減らした．表 3 の設定で 1VM での秒間の I/O 処理性能(IOPS)を測定した．その結果，IOPS は 216[回数/sec]となった．

次に提案手法の評価を行うため，1 台の物理計算機上に 3 台の VM を稼働させ，それぞれの VM 上でベンチマークソフト tpcemysql を実行し，テーブルデータ移動前後の各 VM のスループットとロングインターバルのあるテーブルデータを集約した HDD のアクセス間隔を調査した．また，ロングインターバルのあるテーブルデータを集約した HDD にスピンダウン時間を設定し停止した場合のスループットと消費電力を調査した．

実験環境には物理 HDD が 4 台あり，1 台は OS のシステムファイルを格納し，残りの 3 台には TPC-E のテーブルデータを配置する．TPC-E の テーブルファイルの配置を図 8 に示す．図 8 上の赤い Table はロングインターバルの無いデータ，青い Table はロングインターバルのあるデータである．図より，再配置によりロングインターバルの無いテーブルを配置する HDD 数(2 個)が VM 数(3 個)を下回るので，配置手法 A では 1 つの VM のロングインターバルの無いテーブルだけをアクセスが均等になるように 2 つの HDD に配置する．準備調査で IOPS が Imax の約 2/3 であるため，3 台の VM を 2 台の HDD に配置しても性能劣化は起きにくいと考えられる．今回は確実にロングインターバルが得られると考えられる図 3 の company_competitor 以下のテーブルデータを移動し測定を行った．

表 4　アクセス間隔の測定結果

	アクセス回数		
アクセス間隔 [sec]	再配置前	再配置後（配置A）	再配置後（配置B）
～120	0	19	19
120～60	0	15	20
60～30	0	9	7
30～10	0	1	0
10～1	209	43	46
1～0	474766	357	328

ロングインターバルのあるテーブルデータを配置した HDD における移動前後のアクセス間隔の発生頻度を表 4 に示す．表 4 より，移動前はロングインターバルは 1 回も存在しなく，停止できる時間は無かった．しかし，移動後はロングインターバルのある表のみを集めたため HDD3 のアクセス頻度が千分の一ほどに減少し，120 秒以上の間隔も多く得ることができた．そのため，HDD のスピンダウン設定により省電力化ができると考えられる．

各 VM 上の tpcemysql のスループットの結果を図 9 に示す．図 9 より，性能劣化は配置手法 A では平均約 6%であり，配置手法 B では平均約 14%である．これより配置手法 A の方が少ない性能劣化で HDD 停止時間を確保できていることがわかる．

再稼働によって増加する消費電力は一定であるため，表 4 の結果を用いて停止によって減少する累計消費電力と停止までのアイドル時間の関係を調査した．“アクセス間隔”が“停止までのアイドル時間”より大きいときに“アクセス間隔”－“停止までのアイドル時間”だけ HDD が停止すると仮定し，停止により減少する累計電力は，“停止によって減少する電力の合計”－“再稼働によって増加する電力合計”の差によって求めた．Linux の hdparm は 5 秒間隔でアイドル時間を設定するため，”停止までのアイドル時間”は 5 秒間隔で計算している．計算結果を図 10 に示す．図 10 より，アイドル時間を短くすればするほど削減される消費電力が大きくなることがわかる．

ロングインターバルのあるテーブルデータを配置した HDD における配置手法 A のときのスループットとスピンダウン時間を設定したときの平均消費電力，10 分毎の平均消費電力推移の結果を図 11 と図 12，図 13 に示す．

図 11 より，HDD3 にはロングインターバルのあるテーブルのみを配置しているため，HDD 停止によるスループット

の低下はほどんど見られないことがわかる．

図 12 より，HDD 停止によって消費電力が低下しており，図 10 からもわかるように停止設定時間が短いほど平均の消費電力が短くなっていることがわかる．スピンダウン時間設定後は多くの停止時間が確保されているため，停止前と比べて約半分ほど消費電力を下げることができている．

図 13 より，スピンダウン設定時間が 10 秒のときや 30 秒のときでは，停止されてないときよりも 10 分の平均電力が大きくなっているときがあることがわかる．これは HDD が停止後すぐにアクセスが来たためにロングインターバルが稼げなかったため，消費電力が逆に増えてしまった現象が発生している．また，停止時間 60 秒のときでは 10 分間の平均消費電力が停止時間 5 秒よりも大きくなってるときが半数近く見らる．これは，スピンダウン時間前にアクセスが来てしまったために停止ができなくなってしまい，平均の消費電力が上がってしまっているからである．

6. まとめ

本稿では，アプリケーションの動作情報を用いてディスク上のデータレイアウトを変更し，HDD の消費電力を削減する手法を紹介した．そして，その手法を仮想化環境に適用し，HDD の使用率を考慮したデータ配置手法を提案し，評価結果を示した．結果から，小さい性能劣化でアクセス間隔の拡大により大幅な省電力化ができることがわかった．またスピンダウンの設定時間によってスループットに大きな差は見られなかったが，設定時間を適切に設定しないと消費電力が増えてしまうことがわかった．

今後は，メモリの使用について調査し，その効果の検証を行っていく予定である．

謝辞

本研究は JSPS 科研費 24300034, 25280022, 26730040 の助成を受けたものである．

参考文献

[1] GIPC Survey and Estimation Committee Report FY2009 (Summary)，http://www.greenit-pc.jp/activity/reporting /100707/index.html，2009

[2] 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),

[3] Norifumi Nisikawa , Miyuki Nakano and Masaru Kitsuregawa, ” Energy Efficient Storage Management Cooperated wuth Large Data Intensive Applications,” 28th IEEE International Conference on Data Engineering (IEEE ICDE 2012),

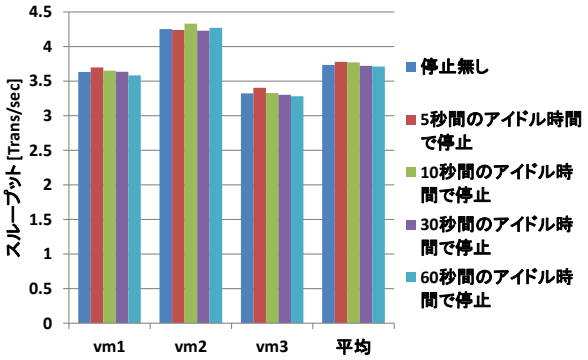


図 11　停止時間によるスループット

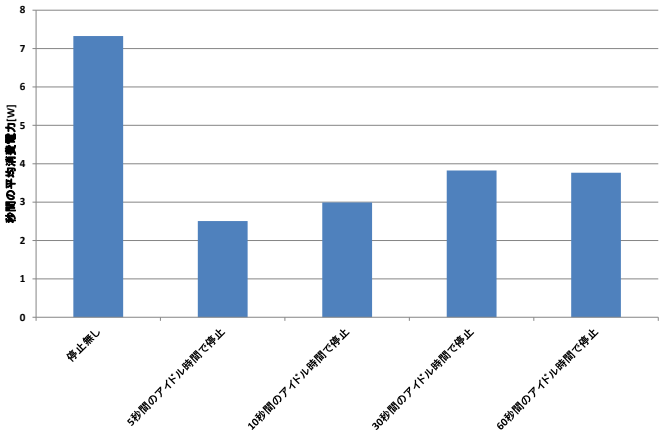


図 12　停止時間による平均消費電力

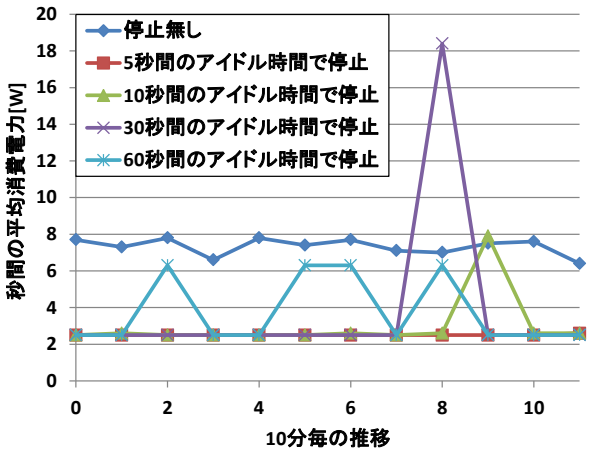


図 13　停止設定による電力の推移

[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, J95-D, 3, 1-13 (2012.03)