Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

仮想化環境における DB データ配置に関する一考察

谷貝　俊輔†　　山口　実靖†‡

近年，情報技術が普及に伴い，データセンターなどで多数のサーバ計算機が稼働するようになり，これらの消費電力が大きな問題の 1 個になっている．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することで HDD の消費電力を削減する方法がある．この手法では HDD へのアクセス間隔を調査し，アクセス量が少ないデータを特定の HDD デバイスにまとめ，HDD のアクセス間隔の拡大と停止時間をはかり省電力化をはかる．本研究では上記手法を仮想化環境に適用し，そして本調査結果をもとに仮想化環境に適したデータ配置方法の提案を行った．評価の結果，本データ配置手法を用いることで少ない性能劣化で大幅な HDD アクセス間隔の拡大が可能であることが確認された

1. はじめに

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働するようになり今後 10 年でデジタル情報量は約 44 倍になるといわれている[1]. これに伴い，サーバの消費電力の増加が問題となり，データセンター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 [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レイアウトを変更することで HDD の消費電力削減する方法がある[2，3].

本研究では上記手法を仮想化環境に適用し，その有効性の調査をおこなう．具体的には代表的な仮想計算機システムである Xen を用いて, 仮想計算機上に mysqlDB を立ち上げ TPC-C 実行時の各テーブルのアクセス量を調査し，アクセス量を考慮したテーブルの再配置を行い HDD アクセス間隔の拡大の程度とトランザクション性能の評価を行なった．そして上記調査結果に基づくデータ配置手法の提案を行った.

2. 応用情報を用いたストレージ省電力

前節で説明した手法の一つに応用(アプリケーション)の動作情報を用いたストレージ省電力手法としてデータ(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 発行間隔を生成する手法が提案されている．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 データをひとつの 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はかる手法が存在する[2]. 図 1 にて応用情報を用いたデータレイアウトの変更について示す.

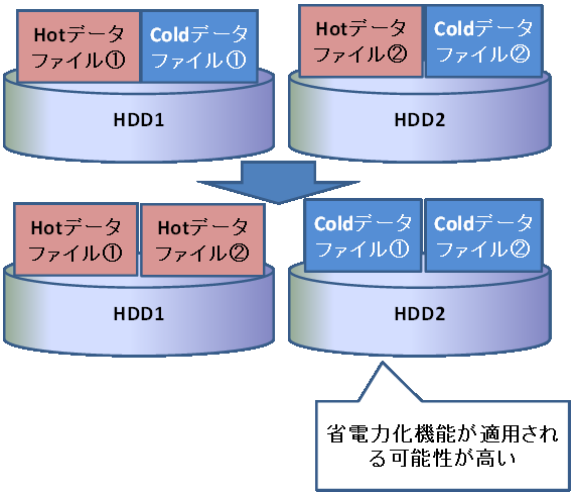

図 1　応用情報を用いたデータレイアウトの変更

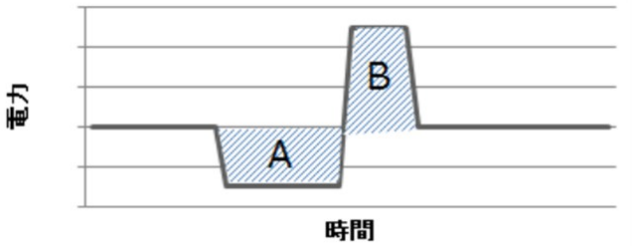

図 2 ストレージの停止と再起動時の電力の変化

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とストレージ再稼働により失われる電力量が等しくなるストレージ停止時間(A=B)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，上記手法により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.

文献[2, 3]ではそれぞれブレークイーブンタイムが 25 秒，10 秒と定義されており，使用した HDD ごとにこの時間は多少の変化があると考えられる.

†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 工学院大学工学部情報通信工学科
Department of information and Communications Engineering, Kogakuin University.

Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

3. 仮想環境における応用情報を用いたストレージ省電力

本章で，仮想化環境において HDD 上のファイルを移動することにより，特定の HDD におけるアクセス間隔を大きくする手法を提案する．

カーネル内でベンチマークソフト tpcc-mysql 実行時の DB テーブルファイルへのアクセス要求数を監視し，各ファイルのアクセス頻度を調査する．そしてアクセス要求の少ないテーブルファイルを，特定の HDD 上に集中して配置する．これにより，要求の少ない HDD のアクセス間隔が大きくなると予想される．

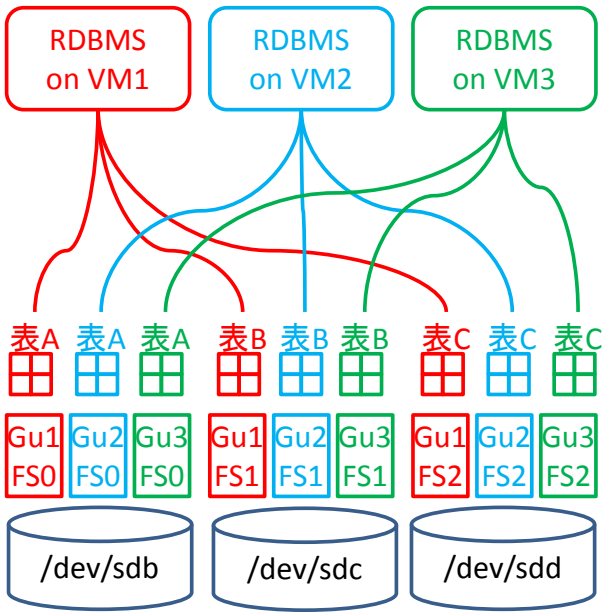

図 3　負荷を考慮した配置方法

4. 性能評価

4.1 実験方法

以下の実験にて提案手法の性能評価を行った．仮想計算機システム Xen を用いて 1 台の物理計算機上に 3 台の VM(VM1，VM2，VM3)を起動させ，全 VM 上に MySQL を立ち上げた．MySQL 上に TPC-C の表を作成し，前章の手法に従い表の配置を変更し，配置状況ごとの性能を測定した．測定はベンチマークソフト tpcc-mysql を使用して測定を行った．配置方法等は複数用意し，DB 性能とアクセス間隔について調査する．

HDD は 4 台使用し，1 つの HDD(HDD1)にはゲスト OS のシステムファイルを格納し，ほかの 3 台の HDD(HDD2, HDD3, HDD4)には MySQL のテーブルファイルを格納した．HDD のアクセスログはカーネル内で取得した．また実験は VM を 1 台から 3 台使用した実験をおこなう．

4.2 実験環境

表 1，表 2 にて実験で使用した機器の仕様について示す．本実験ではアクセス間隔の拡大のために，linux サーバの設定である dirty_expire_centisecs を変更している．この値はキャッシュ上に存在しているページの存在できる時間を指しておりこの値をすぎてもライトバックされないデータがある場合，自動的にキャッシュされる．Dirty_expire_centisecs この値を初期設定の 30 秒から 300 秒に拡大している．

表 1　物理計算機仕様

HostOA	CentOS release 6.4 (Final)
Host Kernel	Linux 2.6.32.57
仮想化システム	Xen version 4.1.2
HDD	2TB×4
File System	ext2
dirty_expire_centisecs	300秒

表 2　仮想計算機仕様

Guest OS	CentOS release 6.3 (Final)
Guest kernel	Linux vm02 2.6.32.57
Virtual CPU Core	1
Virtual Memory	2048[MB]
Vitual HDD	50GB+100GB×3
File System	ext2

4.3 応用情報について

実験で用いる tpcc-mysql は 9 種類のテーブルファイルを用いて計測を行う．表 3 に各テーブルの要領について，図 3 に各テーブルのアクセス頻度についての図を示す．テーブルの配置方法についてはこれらのデータを用いて配置を行う．図 3 から Stock，Order_line，Customer などのファイルはアクセス間隔が小さく省電力化は期待できないファイルだと分かる．対して，Warehouse，District，item などのファイルはアクセス間隔が比較的長く省電力化が期待できるファイルだと考えられる．

表 3　DB 仕様

テーブル名	データ [KB]	インデックス [KB]	合計 [KB]
stock	1,739,776	83,584	1,823,360
order_line	1,075,200	509,952	1,585,152
customer	925,696	100,288	1,025,984
history	104,080	52,368	156,448
orders	66,144	46,704	112,848
new_orders	12,816	0	12,816
item	9,744	0	9,744
district	80	0	80
warehouse	16	0	16
合計	3,933,552	792,896	4,726,448

Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

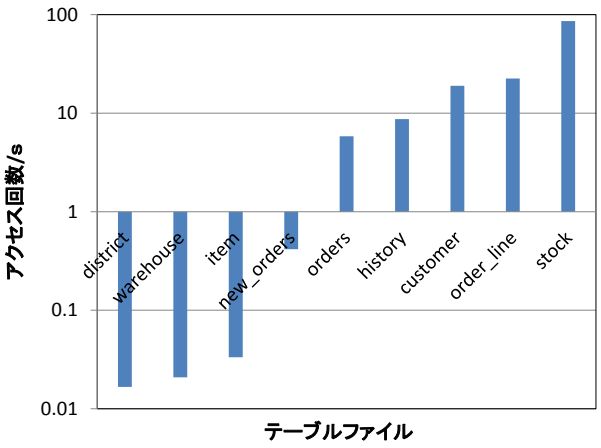

図 4　各テーブルの平均アクセス間隔

4.4 評価方法(仮想マシン 1 台仕様)

仮想計算機 1 台を使用した場合のテーブル配置の詳細以下に示す．配置“9-0-0”では全テーブルファイルをひとつの HDD にまとめる配置方法である．

配置“3-3-3”では，テーブルサイズ均等になるようテーブルを 3 等分し，3 つの HDD に配置した．この配置では，各 HDD のアクセス量はほぼ同等となると予想される．本配置はアクセス頻度を考慮しておらず，本研究ではこれを標準的な配置方法として考える．

配置“1-7-1”では，VM1 にて HDD2, HDD3, HDD4 に表を 1 個, 7 個, 1 個配置した．HDD2 には stock を配置し，HDD4 にはもっともアクセスの少ない表（district）を配置した．

配置“1-6-2”では VM1 にて HDD2, HDD3, HDD4 に表を 1 個, 6 個, 2 個配置した．HDD2 には同様に stock を配置し，HDD4 にはもっともアクセスの少ない表(warehouse, district)を配置した．これは，前述の“1-7-1”の状態から warehouse を移動した状態である．配置“1-5-3&1-5-3”では，両 VM とも HDD2, HDD3, HDD4 に表を 1 個, 5 個, 3 個配置し，HDD4 には次にアクセスの少ない item を追加で配置した．

以降配置“1-4-4”では new_orders，配置“1-3-5”では orders“1-2-6”では history，配置“1-1-7”では customer とアクセス数が少ないファイルを順に HDD4 に追加で配置していく．

この前述した複数の配置は 1-7-1 が特に HDD4 の負荷が少なくなる傾向が強く，1-6-2，1-5-3 と HDD4 に配置されるファイル数が多くなるたびにその傾向は弱くなっていくと考えられる．

4.4.1 測定結果(仮想マシン 1 台使用時)

前節で説明した配置により TPC-C を実行し，トランザクション性能と最大アクセス間隔とアクセス間隔ごとの発生頻度を測定した．VM を 1 台を使用した測定結果を図 5，図 6，図 7，図 8 に示す．図 5，6，7，8 より，アクセス間隔の拡大が確認された配置ではトランザクション性能が約 10%ほどの劣化が見られたが大きな劣化ではないことが確認できた．

図 6，図 7 より，標準的な配置方法“3-3-3”では最大アクセス間隔が 5 秒以下であるが，提案手法では 30 秒以上のアクセス間隔が多数得られていることが分かる．ブレークイーブンタイムは HDD の実装に依存するが，文献[2]や文献[3]に示される例を超えるアクセス間隔が多数回得られており，ストレージ省電力が可能になったと考えられる．

また図 9 に 10 秒以上のアクセス間隔の後に発生した I/O の read，write の比率を示す．

また参考のために read 要求のみに着目したアクセス間隔の最大値と頻度分布を付録に示す．これは遅延書き込みが無限の長さで許された場合のアクセス間隔となり dirty_expire_centisecs の拡大により達成できるアクセス間隔の最大値といえる．

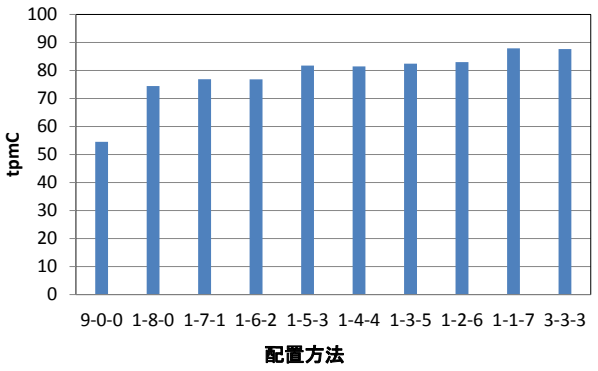

図 5　各配置方法のトランザクション性能

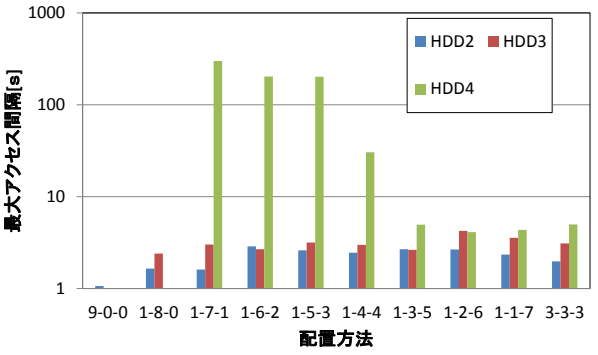

図 6　HDD2,3,4 の最大アクセス間隔

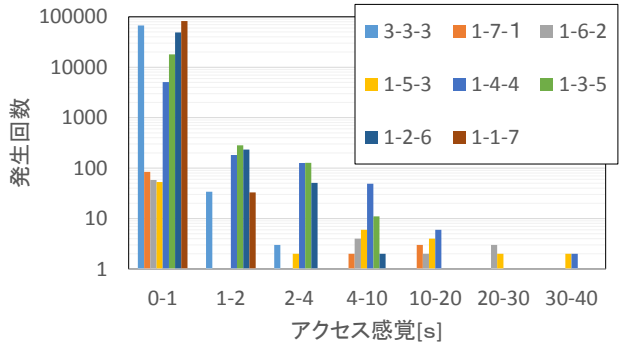

図 7　HDD4 のアクセス間隔頻度分布(1/2)

Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

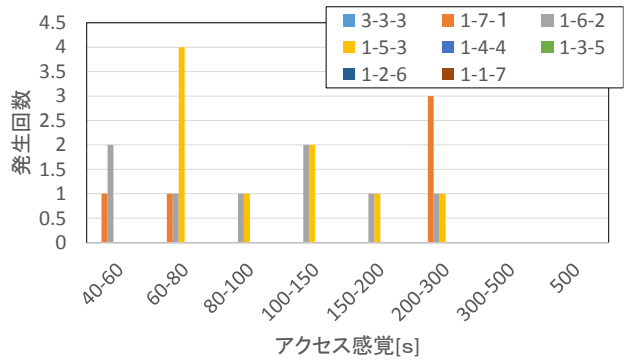

図 8　HDD4 のアクセス間隔頻度分布(2/2)

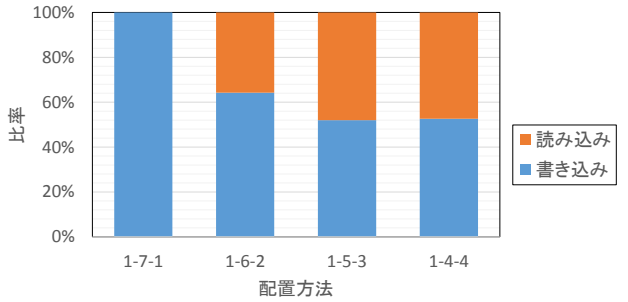

図 9　read，write 比率

4.5 評価方法(仮想マシン 2 台使用時)

配置“3-3-3&3-3-3”では VM1，VM2 ともにテーブルサイズ均等になるようテーブルを 3 等分し，3 つの HDD に配置した．この配置方法はアクセス頻度を考慮していない配置方法により標準的な配置方法と考える．

以下“3-3-3&3-3-3”を含め“1-7-1&1-7-1，1-6-2&1-6-2，1-5-3&1-5-3～1-2-6&1-2-6，1-7-1&1-7-1”の配置方法は前節で説明した配置方法を，両 VM ともに同条件で配置する方法である．

この配置(1-7-1&1-7-1，1-6-2&1-6-2，1-5-3&1-5-3～1-7-1&1-7-1)はすべて前節と同様に HDD4 の負荷が少なくなる様になっているが，“1-7-1&1-7-1”が特にその傾向が強く，1-6-2&1-6-2，1-5-3&1-5-3 と HDD4 に配置されるファイル数が多くなるたびに弱くなっていく．また，アクセスの少ない表が特定の HDD に集中すると同時に，アクセスの多い表も特定の HDD に集中するようになっていることからこの配置方法を総じて少アクセス集中多アクセス集中(少集中多集中)と呼ぶ．

配置“1-7-1&7-1-1”では，HDD4 には“1-7-1&1-7-1”と同様に district 表を配置し，HDD2 と HDD3 の配置は VM1 と VM2 で逆にしてある．これはアクセスの少ない表が特定の HDD に集中するが，アクセスの多い表は残りの HDD に分散するようになっている．

配置“1-6-2&6-1-2～1-1-7&1-1-7”も上記配置と同様にアクセスの少ない表を順に HDD4 に集中配置し，残りの表の配置は VM1 と VM2 で逆としてアクセスの多い表は分散配置している．これらの配置方法では VM 1 と VM2 で、HDD2 と HDD3 の配置を逆にすることでアクセスが少ないファイルを集中させると同時にアクセスが多いファイルを分散させるような配置方法から総じて少アクセス集中多アクセス分散(少集中多分散)と呼ぶ．

また参考のために read 要求のみに着目したアクセス間隔の最大値と頻度分布を付録に示す．

4.5.1 測定結果(仮想マシン 2 台使用時)

VM を 2 台使用した測定結果，DB 性能とアクセス間隔について図 10，11，12，13，14，15 に示す．

図 10 より少集中多分散の配置方法が少集中多集中の結果よりもトランザクション性能が優れていることがわかる．この結果よりアクセス頻度が高いデータを分散させる配置方法は性能が良くなると考察出来る．また図 11，12，13，14，15 から 100 秒以上を超えるアクセス間隔を確認することができ，その間隔の拡大は複数回確認できることから省電力化を実現できたといえる．またアクセス間隔の拡大を確認できた配置は標準配置“3-3-3&3-3-3”の配置の性能の 4.4%ほどの劣化であったことから，少ない性能劣化で省電力化を実現できたといえる．

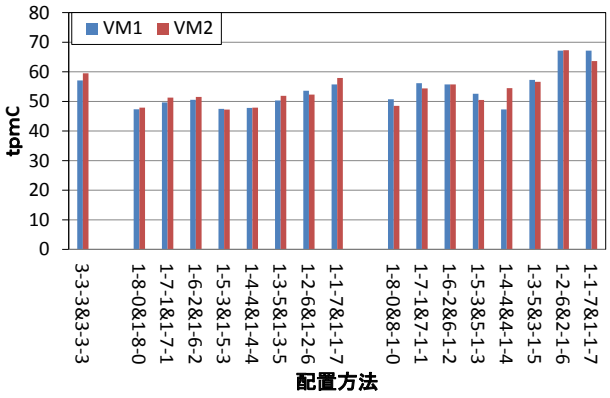

図 10　各配置方法のトランザクション性能

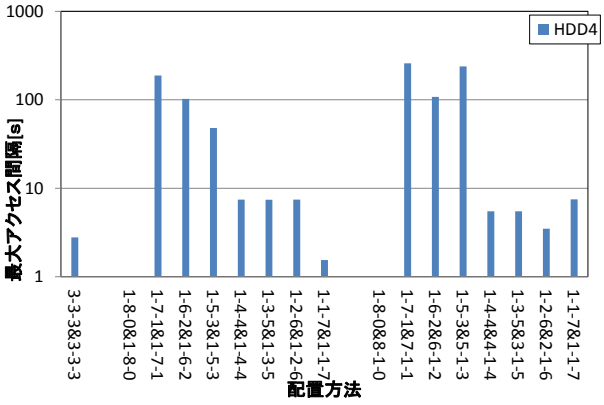

図 11　HDD4 の最大アクセス間隔

Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

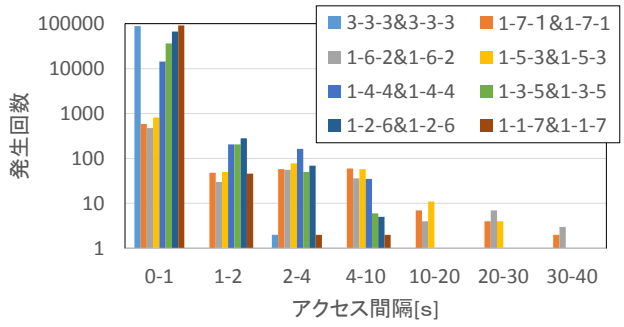

図 12　配置“3-3-3”と少集中多集中の
アクセス間隔頻度分布(1/2)

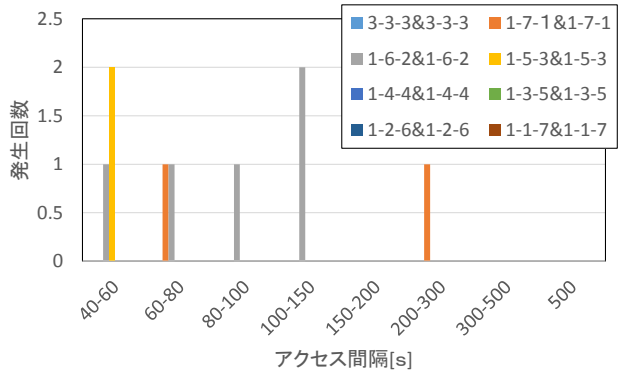

図 13　配置“3-3-3”と少集中多集中の
アクセス間隔頻度分布(2/2)

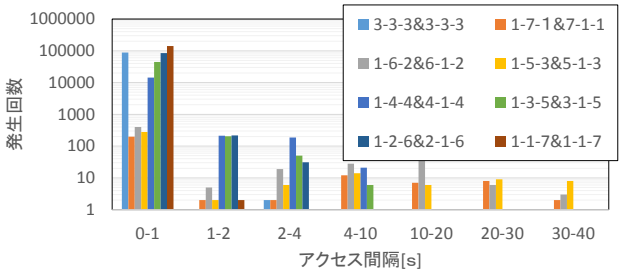

図 14　配置“3-3-3”と少集中多分散の
アクセス間隔頻度分布(1/2)

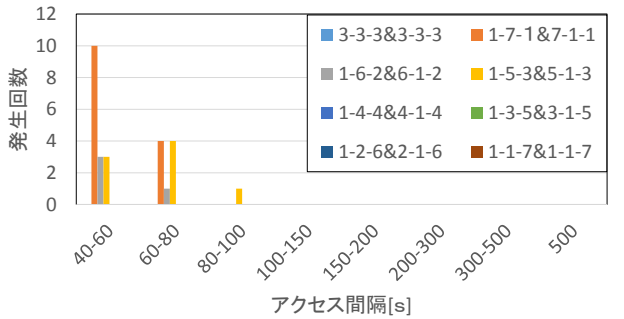

図 15　配置“3-3-3”と少集中多分散の
アクセス間隔頻度分布(2/2)

4.6　評価方法(仮想マシン３台使用時)

この節では VM3 台を使用した場合の配置方法を以下に示す．配置“3-3-3”では VM1，VM2，VM3 ともにテーブルサイズ均等になるようテーブルを 3 等分し，3 つの HDD に配置した．この配置方法はアクセス頻度を考慮していない配置方法により標準的な配置方法と考える．

配置“1-8-0”～“1-1-7”の配置では前節と同様に HDD4 の負荷が少なくなるようになっているが HDD に配置されるファイルが増えるにつれてその傾向は弱くなっていく．また VM1 と VM3 は HDD2 と HDD3 を同配置するが，この時 HDD2 にアクセス頻度の高いファイルが集中している．VM2 は HDD2 と HDD3 の配置を逆にすることでアクセス頻度が高いファイルを HDD2 と HDD3 で分散させている．

また参考のために read 要求のみに着目したアクセス間隔の最大値と頻度分布を付録に示す．

4.6.1 測定結果（仮想マシン３台使用時）

VM を 3 台使用した測定結果を示す. 図 16 が DB 性能を，図 17，18 がアクセス間隔について示している図 17 からアクセス間隔の拡大には成功したが最大 34 秒と比較敵低く，図 18 で頻度分布からも拡大に成功した回数が少ない. また図 16 にて性能はアクセス間隔の拡大に成功した配置と比べ標準配置“3-3-3”から 11%下がっていた．このことから 11%ほどの性能劣化でアクセス間隔の拡大に成功したが、大きなアクセス間隔の拡大を得られなかったので，さらなる改善方法が必要であると考えられる．

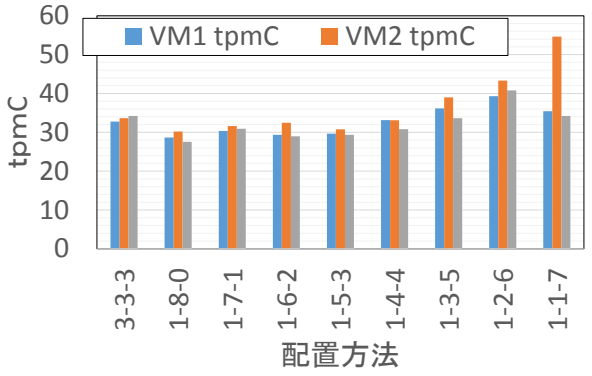

図 16　各配置方法のトランザクション性能

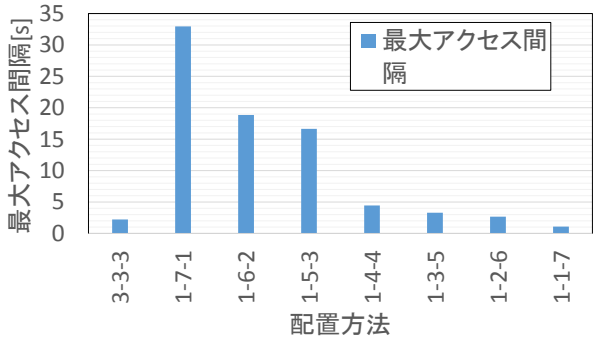

図 17　HDD4 の最大アクセス間隔

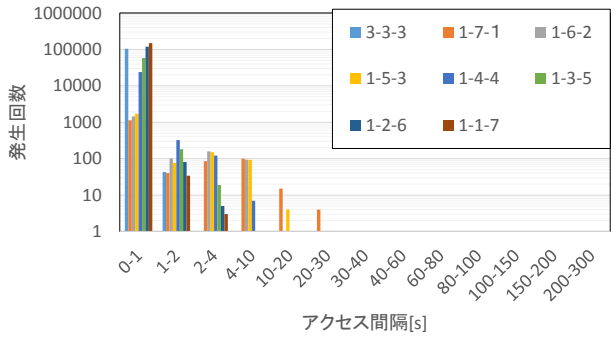

図 18　HDD04 のアクセス間隔の頻度分布

5. まとめ

本研究では，応用情報を利用した HDD のアクセス間隔の拡大手法に着目し，それを仮想化環境に適用し性能評価を行った．評価の結果，本手法による大幅なアクセス間隔の拡大が確認され，仮想化環境においても本手法が有効であることが確認された．

今後は，先読みや遅延書き込みの拡大などによるアクセス間隔のさらなる拡大について考察していく予定である．

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] 飯村　菜穂, 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”(DICOMO2013 7/12　7C-1)

[2] 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),

[3]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)

[4]若色　匠，山口　実靖　“仮想化環境における応答性能を考慮したストレージ稼働時間の低減” 情報処理学会 2013 年全国大会　1L-9

[5] Transaction Processing Performance Council (TPC) TPC BENCHMARKC
Standard Specification　Revision 5.11　February 2010

付録

多くの場合書き込み要求は処理を遅延することが可能であり，書き込み要求発生後も HDD 停止を継続することが可能であると考えることが出来る．逆に読み込み要求では覆うの場合即時処理が求められ，処理を遅延することができず，停止中の HDD は再稼動させることが要求される．

本付録にて読み込み要求のみに着目して測定した HDD アクセス間隔を示す．

図 19 から図 21 が仮想マシン 1 台におけるアクセス間隔を示しており，図 21 から図 24 と図 25 から図 27 が少集多集と少集多分の仮想マシン 2 台におけるアクセス間隔を図 28 から図 30 が仮想マシン 3 台におけるアクセス間隔を示している．これから測定値は遅延書き込みを最大限に活用した場合における停止時間の上限と考えることができる．

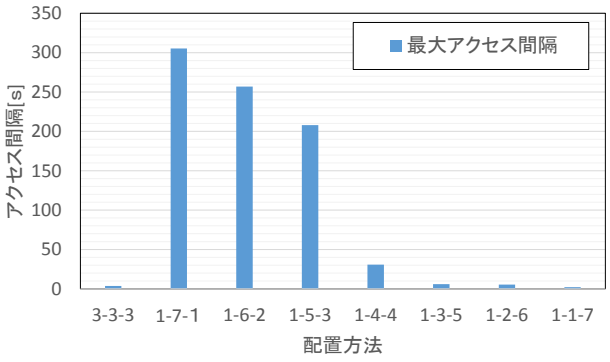

図 19　HDD4 の最大アクセス間隔(read のみ)

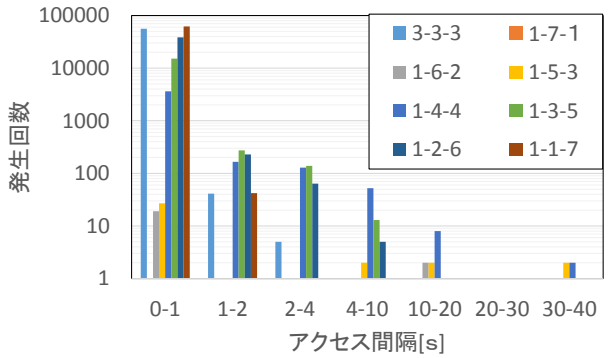

図 20　HDD4 のアクセス頻度分布(read のみ，1/2)

Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

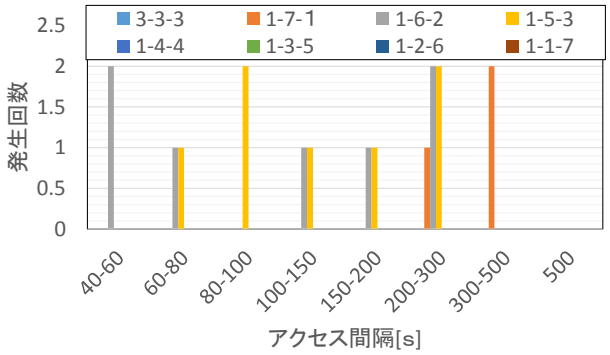

図 21　HDD4 のアクセス頻度分布(read のみ，2/2)

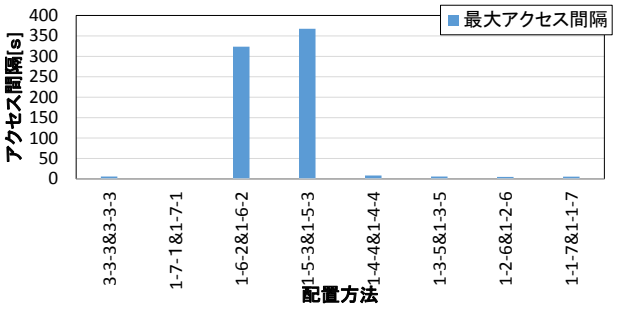

図 22　各配置の HDD4 の最大アクセス間隔
(read のみ，1-7-1&1-7-1 は無限)

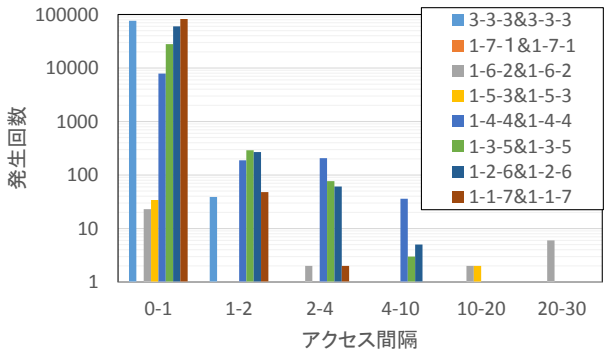

図 23　HDD4 のアクセス間隔頻度分布(read のみ，1/2)

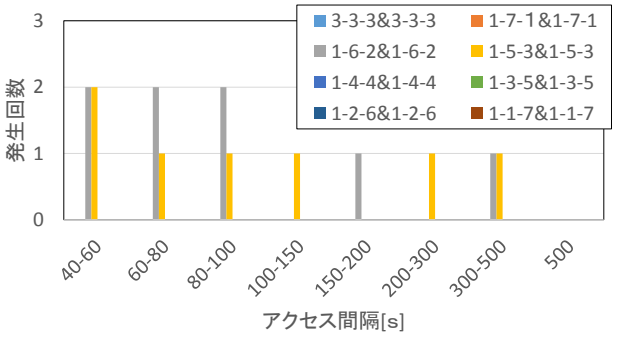

図 24 HDD4 のアクセス間隔頻度分布(read のみ，2/2)

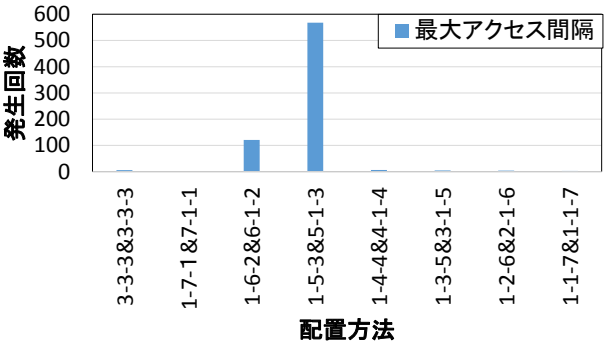

図 25 各配置の HDD4 の最大アクセス間隔
(read のみ，1-7-1&7-1-1 は無限)

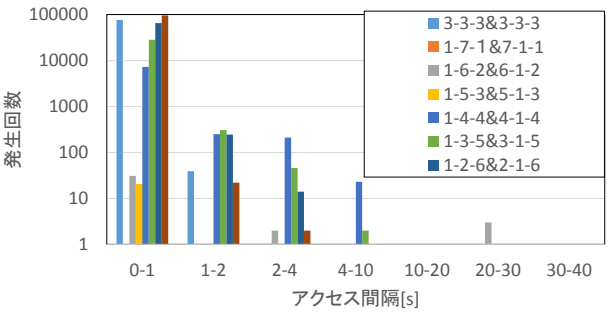

図 26　HDD4 のアクセス間隔頻度分布(read のみ，1/2)

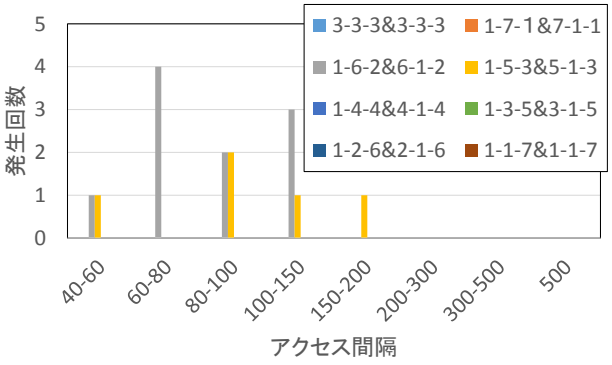

図 27　HDD4 のアクセス間隔頻度分布(read のみ，2/2)

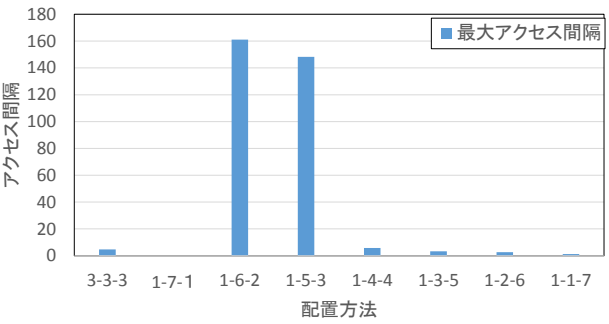

図 28

Vol.2014-DPS-159 No.33
Vol.2014-MBL-71 No.33
2014/5/16

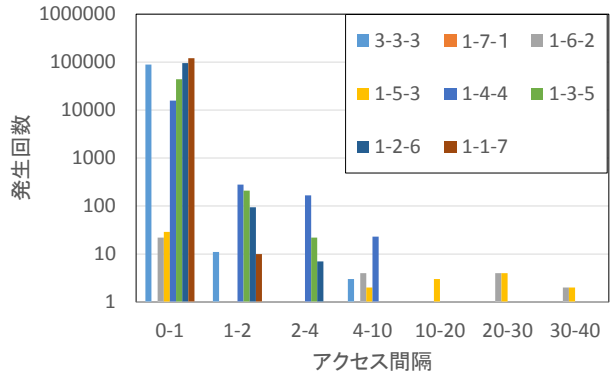

図 29　HDD4 のアクセス間隔頻度分布(read のみ，1/2)

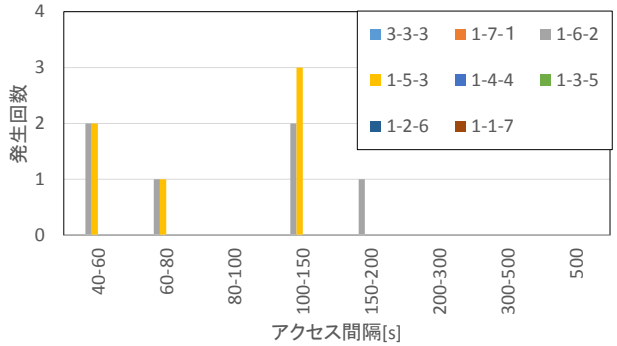

図 30　HDD4 のアクセス間隔頻度分布(read のみ，2/2)