ディスクアレイ装置の電源制御による省エネルギー機構の解析的評価

根本　利弘†　　喜連川　優‡

†東京大学地球観測データ統融合連携研究機構　〒153-8505 東京都目黒区駒場 4-6-1
‡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:　†nemoto@tkl.iis.u-tokyo.ac.jp,　‡kitsure@tkl.iis.u-tokyo.ac.jp

あらまし　ディスクアレイ装置全体の電源制御を行う省エネルギー機構について解析的に評価を行う．ディスクアレイ装置自体がアクセス要求に応じてディスクドライブの回転停止などを行う省エネルギー機構では，アクセス要求がない場合においてもコントローラには通電させる必要がある．これに対し，ホストからの指令によりディスクアレイ装置の電源制御を可能とすることで，アクセス要求がない場合にはディスクアレイ装置は電源制御のための部位のみを通電するだけでよく，大幅な消費電力の削減が可能である．本稿では，アクセス要求がポアソン分布であると仮定した場合において，一定時間アクセス要求がない場合にディスクアレイ装置全体の電源をオフにするとともに，アクセス要求が生じた場合にはディスクアレイ装置の電源をオンにする制御を行う際の平均消費電力，応答遅延時間を導出し，解析的に評価を行う．

キーワード　ディスクアレイ，電源制御，省エネルギー，解析的評価

Analytic Evaluation of Energy Reduction by Power Management of Disk Array

Toshihiro NEMOTO†　　and　Masaru KITSUREGAWA‡

†EDITORIA, The University of Tokyo　4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
‡Institute of Industrial Science, The University of Tokyo　4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
E-mail:　†nemoto@tkl.iis.u-tokyo.ac.jp,　‡kitsure@tkl.iis.u-tokyo.ac.jp

1. はじめに

近年，地球温暖化が極めて重要な問題となり，温室効果ガスの削減が重要課題となっている．IT 機器の消費電力量が急増しており，その中においてもストレージの占める割合は増加傾向にあり[1]，ストレージ機器の省電力化は急務である．このような背景の下，アクセスがないときにディスクドライブの回転数を下げたり，停止させたりすることにより省エネルギー化する機能を有するディスクアレイ装置が商用化されている．このようなディスクアレイ装置においては，ホストからのアクセス要求を受けると，ディスクドライブを回転させて要求に応じる．すなわち，低消費電力状態においても，ホストからのアクセス要求を受け，対応する必要があるためにコントローラやファンなどは稼働状態にある．このため，ディスクアレイ装置の消費電力は，まったくアクセスがない低消費電力状態においても，ディスクドライブが通常回転時にあるときの 40～80%程度となっている[2][3]．そこで，我々はより大幅な消費電力の削減を実現するため，ホスト側でディスクアレイ装置に対するアクセスの制限を行い，ディスクアレイ装置全体の電源を制御するシステムを構築している[4]．低消費電力状態においては，ディスクアレイ装置はホストからの要求を受けることがないため，ホストからの指令による電源のオン・オフ制御を可能とする部分のみを稼働させるだけでよく，大幅な消費電力の削減が実現できる．本稿では，本機構について述べ，消費電力，応答性能を解析的に評価する．

2. 電源制御機構

ディスクアレイ全体の電源制御を行う試作システムは，市販のディスクアレイ装置[5]にリモートからのオン・オフ制御を可能とする電源コンセント[6]を利用し，構築している．この電源コンセントはネットワークポートを有し，HTTP サーバとして動作する．当該 HTTP サーバに対して，所定のアクセス要求を発行することによりコンセントの各ポートのオン・オフの制御が可能である．

ホストにおけるアクセス要求に応じた自動的なディスクアレイ装置の停止，起動には，UNIX，Linux などにおいて利用されている automount（自動マウント）デーモンを使用する．automount に対し，アクセス要求によりディスクのマウントが行われる直前，および，アクセス終了後，ディスクがアンマウントされる直後にストレージ電源管理コマンドを起動するよう変更をする．ストレージ電源管理コマンドは，マウントポイントと電源制御ポートの対応が記述されているストレ

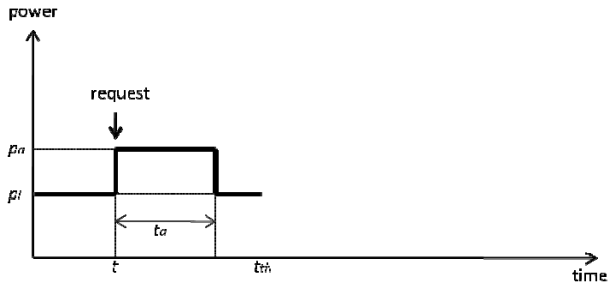

図 1 消費電力モデル（$t \le t_{th}$）

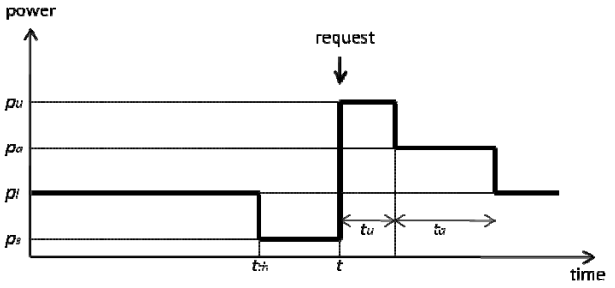

図 2 消費電力モデル（$t > t_{th}$）

ージ電源管理ファイルを参照し，要求されたマウントポイントに対応する電源制御ポートに対して，電源オンコマンドを発行し，必要とされるディスクアレイ装置を起動する．その後，ストレージ電源管理コマンドは，指定されたデバイスが使用可能となるまで待ち，終了する．アンマウント時には，ストレージ電源管理コマンドは，1 つのコンセントに対して複数のディスクアレイ装置が接続されている場合を考慮し，マウントポイントに対応する電源制御ポートに接続された全てのディスクアレイ装置がマウントされていない場合に，電源制御ポートに対して電源オフコマンドを発行し，終了する．このように，アクセス要求に対してディスクアレイ装置の電源制御を実現する．

3. 解析式の導出

本節では，ディスクアレイ装置に一定時間アクセス要求が発行されない場合にディスクアレイ装置の電源をオフにし，その後，アクセス要求が発行された場合には，ディスクアレイの電源をオンにしてアクセス要求に応じるという制御を行う場合の平均消費電力，平均応答遅延時間に関する式の導出を行う．アクセス要求はポアソン分布に従うとする．

ディスクアレイ装置は，アクセス処理時，アクセスがない通常運転時，低消費電力（電源オフ）状態，低消費電力状態から通常運転への移行時における消費電力をそれぞれ，p_a，p_i，p_s，p_uとする．アクセス終了からタイムアウト時間（t_{th}），アクセス要求がない場合に直ちに低消費電力状態へ移行し，また，低消費電力状態においてアクセス要求を受けた際には通常運転状態へと移行し，このためにt_uを要するとする．

3.1. 消費電力

あるアクセス終了時からt経過後に次のアクセス要求が発行された場合について考える．$t \le t_{th}$の場合，ディスクアレイ装置は通常運転状態を保つ（図 1）．計算を容易にするため，p_iを基準としてその差分のみを考えた場合の，直前のアクセス終了時から次のアクセス終了までに消費される電力量E_1は，

$$E_1 = (p_a - p_i)t_a$$

となる．一方，$t > t_{th}$の場合，一度低消費電力状態へ移行し，その後，アクセス要求を受けた時に通常状態へ移行し，アクセスが行われる（図 2）．したがって，このときの電力量E_2は，

$$E_2 = (p_s - p_i)(t - t_{th}) + (p_u - p_i)t_u + (p_a - p_i)t_a$$

となる．平均リクエスト間隔がθである場合の確率密度関数は，

$$f(t) = \frac{1}{\theta}e^{-\frac{t}{\theta}}$$

であることより，アクセス要求あたりの平均消費電力量$\overline{P_{req}}$は

$$\overline{P_{req}} = \int_0^{t_{th}} E_1 f(t)dt + \int_{t_{th}}^{\infty} E_2 f(t)dt$$

$$= \{(p_s - p_i)\theta + (p_u - p_i)t_u\}e^{-\frac{t_{th}}{\theta}} + (p_a - p_i)t_a$$

となる．アクセス処理をしていない時間における消費電力を明確化するため，アクセスデータ量が小さく，処理時間がきわめて短い場合を想定する．

$$t_a \to 0$$

基準をp_iから 0 とした場合の平均消費電力$\overline{P}$は，単位時間当たり $1/\theta$ のアクセス要求を受けることより

$$\overline{P} = p_i + \frac{1}{\theta}\overline{P_{req}}$$

$$= \boldsymbol{p_i + \left\{p_s - p_i + \frac{(p_u - p_i)t_u}{\theta}\right\}e^{-\frac{t_{th}}{\theta}}} \qquad \textbf{数式 1}$$

となる．

数式 1 に示されるように，アクセス要求がポアソン分布に従うと仮定した場合の消費電力に関しては，平均リクエスト到着時間（θ）とブレークイーブン時間（$t_{be} = \frac{(p_u - p_i)t_u}{p_i - p_s}$）の大小関係により最適なタイムアウト時間が決まり，$\theta > t_{be}$の場合 t_{th}=0，$\theta < t_{be}$の場合 t_{th}=∞である．すなわち平均リクエスト到着時間がブレークイーブン時間よりも長い場合はすぐにディスクアレイ装置を低消費電力状態に移行させ，平均リクエスト到着時間がブレークイーブン時間よりも短い場合には，ディスクアレイ装置を低消費電力状態に移行させずに通常運低状態を保つことが，消費電力の点では最適となる．

3.2. 遅延時間

遅延は，ディスクアレイ装置が低消費電力状態にアクセス要求を受けた場合にのみ，t_u生じる．したがって，アクセス要求あたりの平均遅延時間$\overline{D}$は

$$\overline{D} = \int_{t_{th}}^{\infty} t_u f(t) dt$$

$$= t_u e^{-\frac{t_{th}}{\theta}} \qquad \text{数式 2}$$

となる．

4. 解析式による評価

本節では，3 節で導出した数式により，ディスクアレイ全体の電源制御を行う機構の評価を行う．

4.1. 試作システム

試作システム[4]に利用したディスクアレイ装置自体がディスクドライブの回転を停止する省エネルギー機能を有しており，この機能による省エネルギー効果とディスクアレイ全体の電源制御を行う場合の比較を行う．試作システムの各パラメータは表 1 の通りであり，これらは実測値に基づいている．従来方式のディスクの回転停止による省エネルギー機構を用いた場合のブレークイーブン時間は 43.1 秒，ディスクアレイ装置全体の電源停止による省エネルギー機構でのブレークイーブン時間は約 9.6 秒である．

	ディスク回転停止	電源オフ
t_u	43.1 秒	26.1 秒
p_u	260W	260W
p_i	190W	190W
p_s	120W	0W

表 1　試作システムのパラメータ

図 3 は，試作システムにおける平均リクエスト到着間隔と消費電力の関係を表したものである．一般に低消費電力状態へ移行するまでのタイムアウト時間としてブレークイーブン時間が用いられることより，タイムアウト時間として 0 秒，9.6 秒，43.1 秒の場合についての結果を示している．この図によると，同じタイムアウト時間の場合を比較すると，いずれのタイムアウト時間においても，大幅に消費電力を削減することが可能であることが分かる．一方，それぞれのブレークイーブン時間をタイムアウト時間として設定した場合を比較すると，平均リクエスト到着間隔が約 9 秒以下の場合には，ディスクアレイ全体を電源オフする場合の方が，消費電力は大きくなる．これは，平均リクエスト到着率が短い場合には，タイムアウト後にさらにブレークイーブン時間が経過する前にアクセス要求が到着する確率が高く，このため，タイムアウト時間

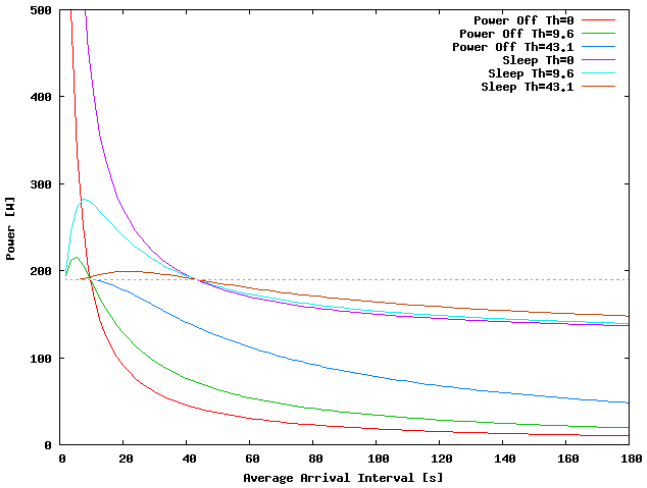

図 3　試作システムの消費電力

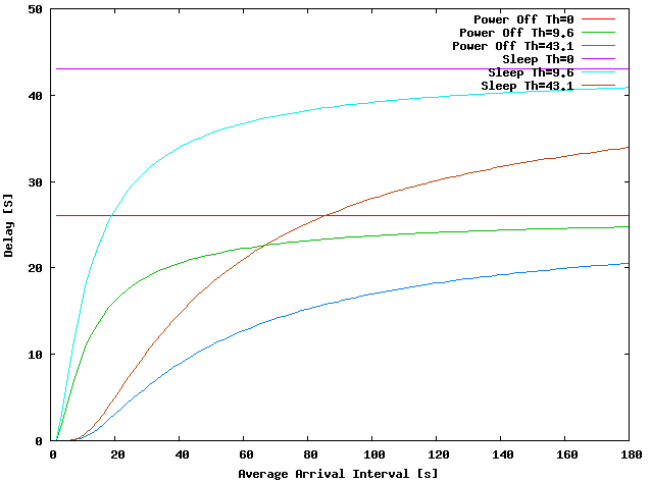

図 4　試作システムの応答遅延時間

が短い場合にはより平均消費電力が上がってしまうためである．

図 4 は，試作システムにおける平均リクエスト到着間隔と応答遅延時間の関係である．応答遅延時間においても，同じタイムアウト時間の場合を比較すると，いずれのタイムアウト時間においても，応答遅延時間が短縮されることが示されている．同様に，それぞれのブレークイーブン時間をタイムアウト時間として設定した場合を比較すると，平均リクエスト到着時間が約 60 秒以下の場合には，平均応答遅延時間は悪化している．これは，タイムアウト時間が短い場合にはより高頻度で低消費電力状態へ移行するためである．一方，リクエスト到着間隔が長い場合には，タイムアウト時間が長くても多くのアクセス要求はタイムアウト後に受けることとなるため，低消費電力状態から通常状態への移行に要する時間が短い，ディスクアレイ全体の電源をオフにする方が，平均応答遅延時間は短くなる．

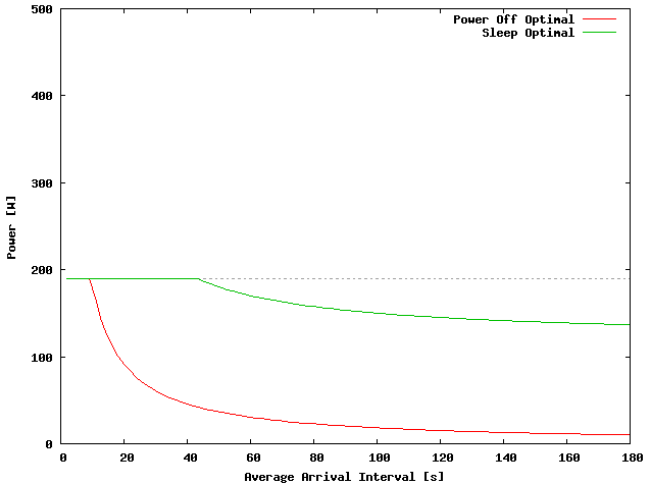

図 5　消費電力に最適化した場合の消費電力

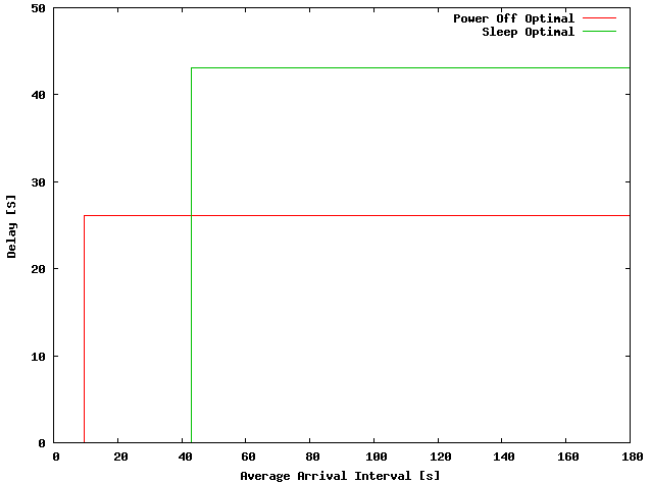

図 6　消費電力に最適化した場合の応答遅延時間

図 5，図 6 はそれぞれ，消費電力が最小となるように，平均リクエスト到着時間がブレークイーブン時間よりも短い場合には，低消費電力状態へ移行せずに通常状態を保ち，平均リクエスト到着時間がブレークイーブン時間よりも長い場合には，タイムアウト時間を 0 とするときの，平均リクエスト到着時間と消費電力，応答遅延時間の関係である．この場合ブレークイーブン時間より平均リクエスト到着間隔が短い場合には低消費電力状態へ移行しないので，消費電力が増加することはなくなる．ディスクアレイ装置全体の電源をオフにする場合にはブレークイーブン時間が短縮されるため，平均リクエスト到着間隔が両者のブレークイーブン時間の間にある場合には，応答遅延時間はディスクアレイ装置全体の電源をオフにする場合の方が悪化することとなる．

4.2. 立ち上げに要する時間が増加する場合

試作システムでは，低消費電力状態から通常状態への移行に要する時間が，ディスクアレイ装置全体の電源をオフにすることによる省エネルギー機構によって短縮されるが，それに対して，移行に要する時間が増加する場合を仮定する．

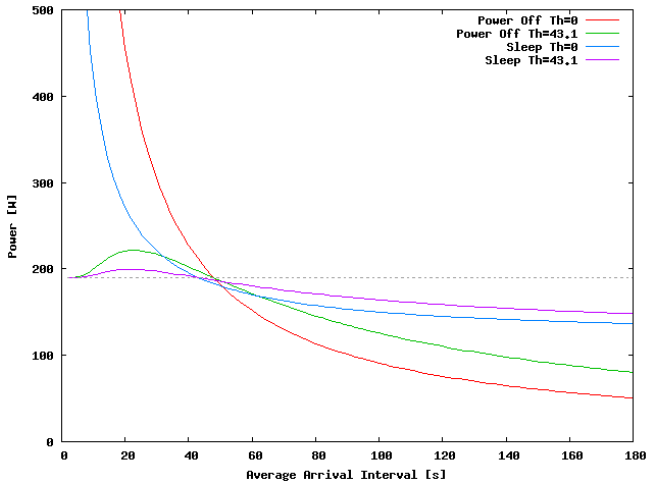

図 7　立ち上げに要する時間が増加するとした場合の消費電力

図 7 は，ディスクアレイ装置全体の電源をオフにした状態からアクセス可能となるまでに 130 秒を要するとした場合の，平均リクエスト到着間隔と消費電力の関係である．このとき，ブレークイーブン時間は約 48 秒である．タイムアウト時間が 0 秒，および 43.1 秒の結果を示しているが，いずれの場合においても，ブレークイーブン時間よりも平均リクエスト到着間隔が短い場合には消費電力は悪化するが，ブレークイーブン時間よりも平均リクエスト到着間隔が長い場合には消費電力は低下する．立ち上げに要する時間が増加したため，その間の消費電力量が増加し，平均リクエスト到着間隔が短い場合には消費電力が増加し，一方，平均リクエスト間隔が長い場合には，低消費電力状態の時間が長くなるため，低消費電力状態での消費電力が小さい電源オフによる機構の場合の方が，消費電力は小さくなる．消費電力に最適化した場合には，平均リクエスト到着間隔がブレークイーブン時間よりも短い場合の消費電力の増加を抑えることが可能である．

図 8 は，ディスクアレイ装置全体の電源をオフにした状態からアクセス可能となるまでに 130 秒を要するとした場合の，平均リクエスト到着間隔と応答遅延時間の関係である．応答遅延時間については，平均リクエスト到着間隔によらず，常に悪化してしまう．

4.3. 立ち上げ時の消費電力が増加する場合

試作システムでは，低消費電力状態から通常状態への移行時の消費電力はディスクアレイ装置全体の電源をオフにすることによる省エネルギー機構によって変化しないが，それに対して，移行に要する消費電力が増加する場合を仮定する．

図 9 は，ディスクアレイ装置全体の電源をオフにした状態からアクセス可能とする際に 500W を要すると仮定した場合の平均リクエスト到着間隔と消費電力の

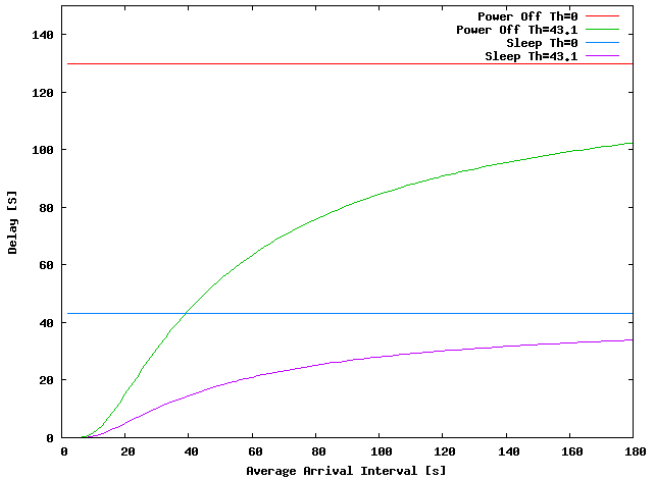

図 8　立ち上げに要する時間が増加するとした場合の応答遅延時間

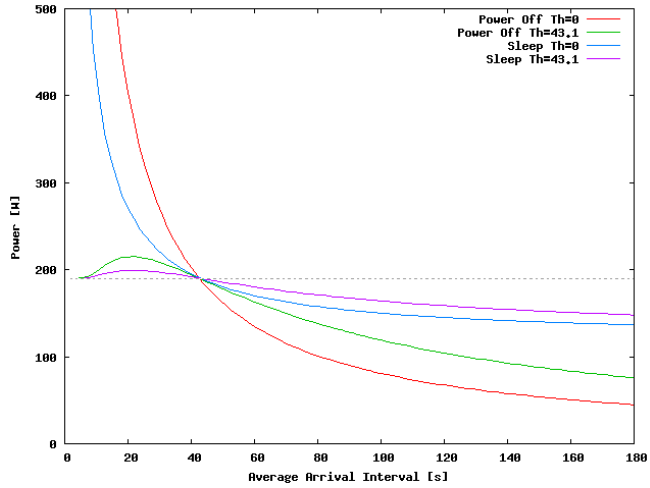

図 9　立ち上げに要する消費電力が増加する場合の消費電力

関係である．このときのブレークイーブン時間は約 42.6 秒である．数式 1 における，$e^{-\frac{t_{th}}{\theta}}$の係数がほぼ同じ値となることより，図 7 の場合とほぼ同様の結果となる．一方，応答遅延時間については，数式 2 においてt_uは変化しないため，図 4 から変化しない．

5. 関連研究

ストレージシステムの省エネルギー機構に関する研究は，まずパーソナルコンピュータにおける単一ディスクドライブを対象として始まった．初期の研究においては，主にディスクドライブを停止させるまでのタイムアウト時間の決定方式について検討が行われている．[7][8]では，ディスクドライブを停止させるまでの固定されたタイムアウト時間をアクセストレースを用いて評価し，数秒程度が最適であることを示した．また，[10][11][12]では，アクセス履歴やディスクの停止状況等を用いてタイムアウト時間を適応的に変更する手法を提案した．[13]では，複数の手法について比較を行っている．近年では，主たる対象をサーバにおける複数ディスク群とし，また，積極的にディスクドライブの停止時間を長くし，省エネルギー効果を高めることに注力がなされている．これらは，キャッシュを用いてディスクへのアクセス頻度を下げる，データへのアクセスのタイミングを変更して集中させ，アクセス時間の短縮を図る，オリジナルデータ，さらにはレプリカを含めたデータの配置により，特定のディスクへのアクセス頻度を下げる，といったことにより，ディスクの停止時間を延ばし，より高い省エネルギー効果を実現するものに大きく分類される．[16]では，応答時間と消費電力により RAID-4，RAID-5，RAID-10 の評価を行った．[15][22][24]はアプリケーションのプロファイリング，改修によりデータへのアクセスタイミングを変更して集中させることにより，ディスク装置のアイドル時間を延ばし，[14]では少数のディスクドライブをキャッシュとして用い，他のディスクへのアクセス頻度を下げてアイドル時間を延ばす方式を提案した．[20]では，より省エネルギー効果の高いキャッシュアルゴリズムを提案している．[21]では，RAID-1，RAID-5 において，ミラーディスク，パリティディスクを効果的に利用して，新たに起動されるディスクが少なくなるような起動ディスクの選択法を提案し，[26]ではさらにキャッシュを含めたより高い省エネルギー手法を提案した．[19]では，アクセス頻度に応じたデータ配置手法を提案し，低アクセス頻度データを特定のディスクに集中させることにより，当該ディスクのアイドル時間を延ばしている．[28][30]では，レプリカを作成して，負荷に応じて段階的にドライブを起動するシステムを提案している．これらの手法はいずれもディスクドライブを停止し，省エネルギー化を図るものである．本稿の手法は，アクセス制御を上位側で行うことによりディスク装置停止時の消費電力を低減させるものであり，これらの手法における停止するディスクドライブに対して適用することにで，より一層の省エネルギー効果が期待できる．

[17][18]では，回転数が可変で，低回転状態でもデータへアクセスが可能なディスクドライブ（DRPM）を仮定し，その効果を示している．また，[23]では，DRPM 利用時におけるキャッシュアルゴリズムを提案し，[25]では，DRPM ドライブを階層的に利用し，省エネルギー化を図る手法を提案している．DRPM は比較的負荷が高い場合の省エネルギー化を目的としているのに対し，本稿での手法は低負荷時の省エネルギー化に貢献するものであり，適用範囲が異なる．しかしながら，DRPM に対しても同時に相補的に適用可能である

と考えられる．

[9]は，本稿と同様に，アクセス要求がポアソン分布に従うと仮定して，消費電力，応答遅延時間の式を導出している．しかしながら，本稿と異なり，低消費電力状態での消費電力を 0 としており，低消費電力状態においても電力が消費される場合が考慮されていない．

6. おわりに

本稿では，ディスクアレイ装置全体の電源をオフにすることによる省エネルギー機構の平均消費電力，応答遅延時間を解析的に導出し，評価を行った．本機構により，大幅な消費電力の削減が可能であることを示した．

参 考 文 献

[1] U.S. Environmental Protection Agency ENEGY STAR Program, “Report to Congress on Server and Data Center Energy Efficiency, Public Law 109-431”, Aug. 2007.

[2] 富士通株式会社, “富士通 ETERNUS ディスクアレイの MAID 技術による省エネルギー”, http://storage-system.fujitsu.com/jp/products/diskarray/download/pdf/MAID_whitepaper.pdf, Nov. 2007.

[3] G. Schulz, “MAID 2.0: Energy Saving without Performance Compromises, Energy Saving for Secondary and Near-line Storage Systems”, http://www.nexsan.com/solutions/energysavings/whitepaper.php, Jan. 2008.

[4] 根本利弘，喜連川優, “自動マウントを利用した外部ストレージの電源制御”, DEIM Forum 2010, E6-2, Mar. 2010.

[5] ネクサン・テクノロジーズ・インク, “ハイパフォーマンス SATA ストレージ”, http://www.nexsan.jp/sataboy.html.

[6] オムロン株式会社, “マルチコントロールコンセント RC1504 取扱説明書 詳細版”, http://www.omron.co.jp/ped-j/dengen/download/rc1504/rc1504_manual_j04.pdf.

[7] K. Li, et al., “A Quantitative Analysis of Disk Power Management in Portable Computers”, In Proceedings of the USENIX Winter 1994 Technical Conference, pp.279-291, Jan. 1994.

[8] F. Douglis, et al., “Thwarting the Power-Hungry Disk”, In Proceedings of the USENIX Winter 1994 Technical Conference, pp.292-306, Jan. 1994.

[9] P. Greenwalt, “Modeling Power Management for Hard Disks”, In Proceedings of MASCOTS’94, pp.62-66, Jan. 1994.

[10] F. Douglis, et al., “Adaptive Disk Spin-down Policies for Mobile Computers”, In Proceedings of MLICS’95, pp.121-137, Apr. 1995.

[11] D. P. Helmbold, et al., “A Dynamic Disk Spin-down Technique for Mobile Computing”, In Proceedings of MobiCom’96, pp.130-142, Nov. 1996.

[12] Y. Lu, et al., “Adaptive Hard Disk Power Management on Personal Computers”, In Proceedings of the Ninth Great Lakes Symposium on VLSI 1999, pp.50-53, Mar. 1999.

[13] Y. Lu, et al., “Quantitative Comparison of Power Management Algorithms”, In Proceedings of DATE 2000, pp.27-30, Mar. 2000.

[14] D. Colarelli, et al. “Massive Arrays of Idle Disks for Storage Archives”, In Proceedings of SC2002, pp.1-11, Nov. 2002.

[15] A. Weissel, et al., “Cooperative I/O – A Novel I/O Semantics for Energy-Aware Applications”, In Proceedings of OSDI’02, pp.117-129, Dec. 2002.

[16] S. Gurumurthi, et al., “Interplay of Energy and Performance for Disk Arrays Running Transaction Processing Workloads”, In Proceedings of ISPASS’03, pp.123-132, Mar. 2003.

[17] S. Gurumurthi, et al., “DRPM: Dynamic Speed Control for Power Management in Server Class Disks”, In Proceedings of ISCA 2003, pp.169-179, Jun. 2003.

[18] E.V.Carrera, et al., “Conserving Disk Energy in Network Servers”, In Proceedings of ICS’03, pp.86-97, Jun. 2003.

[19] E. Pinheiro, et al., “Energy Conservation Techniques for Disk Array-Based Servers”, In Proceedings of ICS’04, pp.68-78, Jun. 2004.

[20] E. Papathanasiou, et al., “Energy Efficient Prefetching and Caching”, In Proceedings of the USENIX Annual Technical Conference, pp.255-268, Jun. 2004.

[21] D. Li, et al., “EERAID: Energy Efficient Redundant and Inexpensive Disk Array”, In Proceedings of SIGOPSEW’04, Sep. 2004.

[22] S. W. Son, et al., “Software-Directed Disk Power Management for Scientific Applications”, In Proceedings of IPDPS’05, Apr. 2005.

[23] Q. Zhu, et al., “Power Aware Storage Cache Management”, IEEE Transactions on Computers, Vol.54, pp.587-602, May 2005.

[24] S. W. Son, et al., “Disk Layout Optimization for Reducing Energy Consumption”, In Proceedings of ICS’05, pp.274-283, Jun. 2005.

[25] Q. Zhu, et al., “Hibernator: Helping Disk Arrays Sleeping through the Winter”, In Proceedings of SOSP’05, pp.177-190, Oct. 2005.

[26] X. Yao, et al., “RIMAC: A Novel Redundancy-based Hierarchical Cache Architecture for Energy Efficient, High Performance Storage Systems”, In EuroSys’06, pp.249-262, Apr. 2006.

[27] E. Pinheiro, et al., “Exploiting Redundancy to Conserve Energy in Storage Systems”, In Proceedings of SIGMETRICS/Performance 2006, pp.15-26, Jun. 2006.

[28] C. Weddle, et al., “PARAID: A Gear-Shifting Power-Aware RAID”, In Proceedings of FAST’07, pp.245-260, Feb. 2007.

[29] T. Xie, “SEA: A Striping-Based Energy-Aware Strategy for Data Placement in RAID-Structured Storage Systems”, IEEE Transactions on Computers, Vol.57, No.6, pp.748-761, Jun. 2006.

[30] J. Kim, “Using Replication for Energy Conservation in RAID Systems”, In Proceedings of PDPTA’10, Jul. 2010.