3A-2

# アクセス予測を利用したHPC向け省電力・高速階層ストレージの性能向上のための負荷分散アルゴリズムに関する検討

黒川大樹† 藤本和久† 赤池洋俊‡ 三浦健司† 村岡裕明†

東北大学 電気通信研究所† (株)日立製作所 システム開発研究所‡

## 1. はじめに

近年、ユーザが扱うデータ量が増大したことにより、データセンタ内でのストレージが占める電力消費が増加している。これまでにも、MAID[1]やStorage tieringといったストレージの省電力化技術に関する研究がなされてきたが、それらの手法は電力消費と応答性能がトレード・オフの関係になっていた。このことは現在のデータセンタへの高速性の要求の解決方法としては不十分であり、このことから従来の手法はアーカイブなどの限られた用途での使用にとどまっていた。そこで、我々はストレージシステムにおいて、データセンタで要求される応答時間や転送速度を維持したまま消費電力の削減を実現する新たな手法として、アクセス予測を利用した階層ストレージの階層管理方式を提案している[2]。

## 2. 提案手法と検証目的

提案手法のシステムの概略図をFig. 1に示す。高速にデータ転送を行えるオンラインストレージ(以下、OL)と低消費電力なニアラインストレージ(以下、NL)において階層構造を構成し、2つのストレージ間でファイル配置の制御やNLの電源制御を適宜行うことにより高速・大容量かつ低消費電力なシステムを実現する。全てのデータは予め、NLに保存しておき、ジョブが投入されキューに並び始めてから、実行されるまでの待ち時間$T_w$を利用し、予めジョブがアクセスするファイルをNLからOLへコピーすることにより、ジョブはデータファイルの高速な読み出しが可能になる。$T_w$はジョブの投入間隔と実行時間の分布を考えることにより、理論式を用いて見積もることができる。これまで、提案手法に関して待ち行列理論に基づく確率的検証がなされてきた[3]。この方式を用いると1PBクラスのシステムを想定した場合、従来の50%程度の電力で稼働できることが示されている[2]。

Investigation of load-balancing algorithm in high-speed and mass storage systems for HPC systems with access prediction

† Hiroki Kurokawa, Kazuhisa Fujimoto, Kenji Miura, Hiroaki Muraoka, RIEC, Tohoku University.

‡ Hirotoshi Akaike, Systems Development Laboratory, Hitachi, Ltd.

しかし、ジョブがアクセスするファイルはランダムであり、ファイルサービス専用サーバとして複数台用意されているNetwork Atached Storage(NAS)の中で特定のものだけに負荷が集中し、ファイルサービスの低下を招くことが懸念されていた。そのため、ジョブスケジューラと連携し、ジョブの情報を利用することにより、ジョブ実行前に予めNASノードに対するアクセス負荷を分散して割り当てるアルゴリズムを作成し、シミュレーションでその効果を検証した。

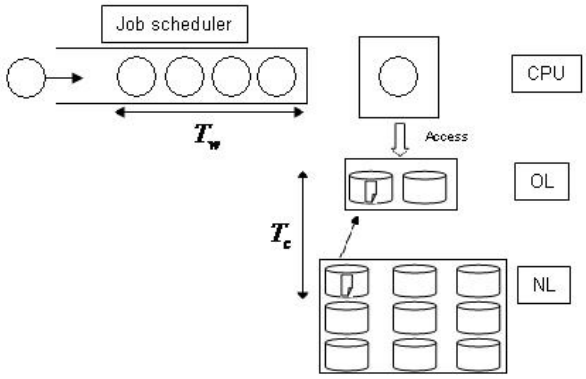


Fig. 1 提案手法のシステムの概略図

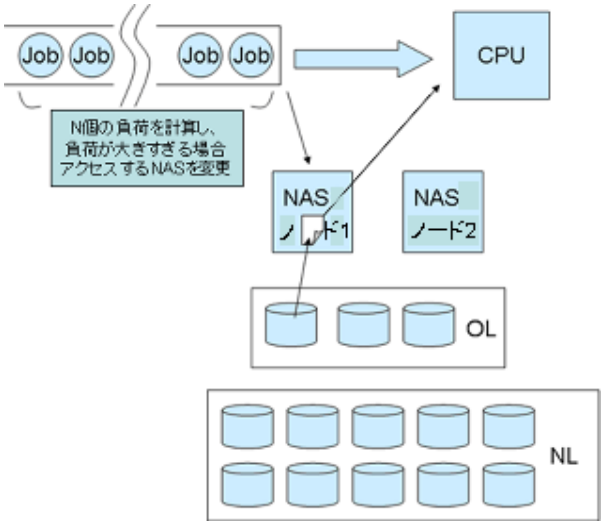


Fig. 2 提案アルゴリズムの模式図

## 3. 負荷分散手法

負荷を分散させるアルゴリズムとして、検討した方式の模式図を Fig. 2 に示す。この方式では、ジョブスケジューラ内で待機中のジョブが所定の順番になったときに、そのジョブより前に並んでいる任意の n 個のジョブの NAS アクセス負荷を計算し、当該ジョブのアクセス先の NAS ノード負荷が閾値を超えている場合、他の NAS ノードの中で最も負荷の低い NAS ノードにアクセスするように設定する。ジョブ投入時にジョブスケジューラ内のジョブが m 個(m<n)の場合は 1～m-1 番目の負荷を計算し、同様の操作を行う。

検証では、イベント駆動型シミュレータである Hyperformix 社、WorkBench5.3 を使用した。ジョブの投入間隔と実行時間は超指数分布に従うとした[4]。超指数分布とは確率密度関数が複数の独立なアーラン分布の重ねあわせで表される超アーラン分布の最も簡単な場合であり、複数の独立な指数分布の重ね合せで表される。このうち、ジョブの投入間隔は到着率の大きい指数分布に従う時間帯と到着率が小さい指数分布に従う時間帯を重ねあわせた超指数分布とした[5]。また、アクセス負荷を計算するジョブ数 n は 12、NAS の台数は 2 台、CPU 数は 12、1 つのジョブは実行のために 1 つの CPU を占有するという条件を想定した。この条件のもとでジョブの投入と実行をシミュレートし、2 台の NAS ノードの負荷を、アクセス負荷の分散を行わない場合と、アルゴリズムを実装し、負荷分散を行った場合で比較した。

## 4. 検証結果

アルゴリズム実装前と実装後の 2 台の NAS の負荷を経過時間で表した結果をそれぞれ Fig. 3,4 に示す。CPU 数が 12、NAS 台数が 2 台であることから、一方の NAS ノードに 7 以上の負荷がかかっている場合負荷が偏っていると言える。負荷分散アルゴリズムを実装する前では負荷が偏っている時間帯が存在していたが、アルゴリズム実装後は負荷が分散されていることが確認できる。

## 謝辞

本研究の一部は、文部科学省による次世代 IT 基盤構築のための研究開発「高機能・低消費電力スピンデバイス基盤技術の開発」の援助を得て行いました。ここに謝意を表します。

## 参考文献

[1] Dennis C, Dirk G, “Massive Arrays of Idle Disks for Storage Archives”, Univ. of Colorado, Boulder, July 26, 2002.

[2] Kazuhisa Fujimoto, Hirotoshi Akaike, Naoya Okada, Kenji Miura, Hiroaki Muraoka, “ Power-aware Proactive Storage-tiering Management for High-speed Tiered-storage Systems”, Sustain IT '10

[3] 岡田尚也， 藤本和久，赤池洋俊，三浦健司，村岡裕明“アクセス予測を利用した HPC 向け高速大容量階層ストレージの階層管理方式の予測精度向上手法に関する検討”, FIT2009, ‘09

[4] J. Jann, P. Pattnaik, H. Franke, et al, “Modeling of Workload in MPPs”, LNCS, Vol1291/1997, pp.95-116, 2006.

[5]黒川大樹, 藤本和久, 赤池洋俊, 三浦健司, 村岡裕明, “Analysis of probabilistic distribution about job arrival interval and execution interval in high performance computing”, 電気関係学会東北支部連合大会 講演論文集, pp.2, 2010

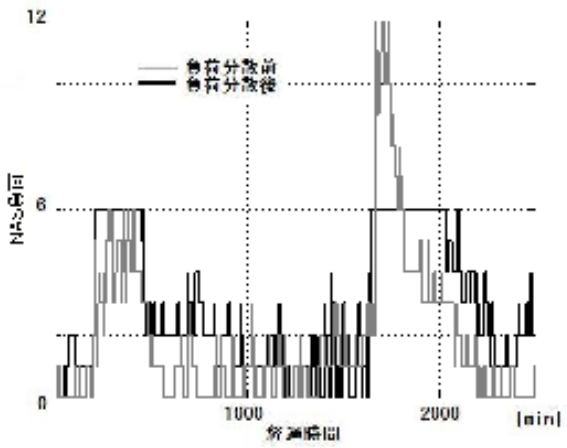


Fig. 3 NAS1 の負荷状態

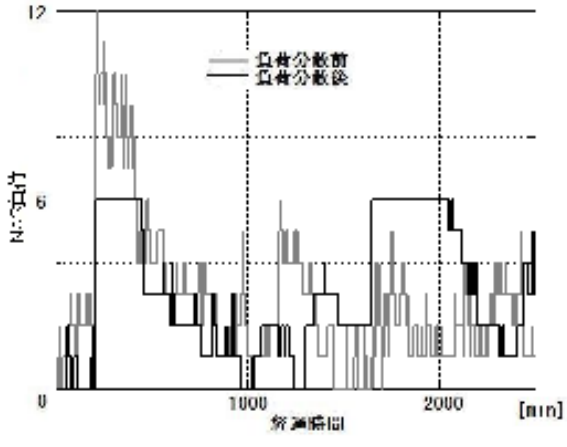


Fig. 4 NAS2 の負荷状態