3A-1

HPC 向け省電力階層ストレージの性能向上のための負荷分散手法と効果の検証

赤池　洋俊† 藤本　和久‡ 黒川　大樹‡ 三浦　健司‡ 村岡　裕明‡
(株) 日立製作所 システム開発研究所† 東北大学電気通信研究所‡

1. はじめに

近年、IT 機器の消費電力は無視できないほど増加しており、大きな問題となっている。ストレージシステムはその中でも多くの電力を消費するシステムの一つである。特にスーパーコンピュータと接続するストレージシステムには大量のデータを高速に入出力することを目的として高い性能が要求される。そのため、性能を維持しながら消費電力を削減するストレージアーキテクチャと、その管理方式が求められている。

2. 省電力階層ストレージ

この背景の下で、図 1 に示す様に高性能なオンラインストレージ(以下、OL)と大容量のニアラインストレージ(以下、NL)の階層構成においてアクセス予知(図 1 中(1))に基づくデータ配置(図 1 中(4))とディスク電源の ON／OFF 制御(図 1 中(3)(2))を行う低消費電力化方式を提案した。さらに、提案方式を試作機に実装し、実際に消費電力を測定することで省電力効果を検証した。その結果、階層ストレージにおいて使用頻度に基づくデータ管理とディスクのスピンダウン制御を行う従来方式と比較して、提案方式はシステム容量 1024TB の場合の試算で性能を維持しながら消費電力を 50%以上削減する見込みを得た[1]。

アクセス予知には、キューアクセス予知方式を用いている。これは、計算機管理サーバのスケジューラ情報やジョブ情報をヒントにして、ジョブのアクセス先データの特定とジョブ実行開始までの時間的余裕の予測を行う。本方式では、ジョブがキュー内で待機している間に、ジョブのアクセス先データを NL から高速な OL ディスクにコピー(データ配置)することで CPU はジョブ実行時に高速な OL 上のデータにアクセスできる。逆に、データ配置がジョブ実行に間に合わない時は低速な NL ディスクにアクセスするために性能低下が発生してしまうが、本方式ではジョブ実行遅延操作(図 1 中(5))で一時的にジョブを待機状態にすることで性能低下を防止している。

このように提案方式は性能維持と低消費電力を両立する。ところが、実際の環境では、スーパーコンピュータは直接ストレージシステムのディスクにブロックアクセスするわけではなく、ストレージシステムに接続したファイルサーバにファイルアクセスを行う。そのため、ストレージ側だけでなく、ファイルサーバ側においても同様に性能向上が重要となる。この OL と NL からなる階層ストレージにファイルサーバを含めた試作機全体を省電力階層ストレージと呼ぶ。

The Verification of the Load Distribution Technique for an Energy-efficient High Speed Tiered-Storage System (eHiTS) with Proactive Migration for HPC Systems.
† Hirotoshi Akaike, Systems Development Laboratory, Hitachi, Ltd.
‡ Kazuhisa Fujimoto, Hiroki Kurokawa, Kenji Miura, Hiroaki Muraoka, RIEC, Tohoku University.

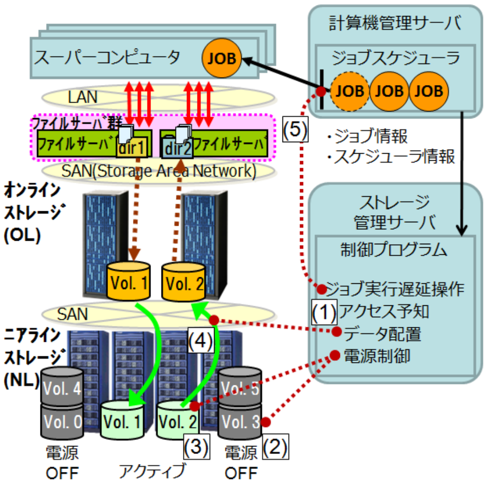

図 1. 省電力階層ストレージの概要

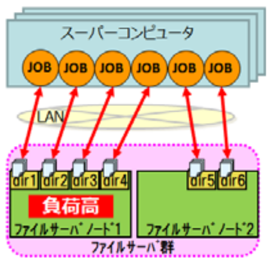

図 2. 負荷集中の例

3. 負荷分散手法

スーパーコンピュータ環境ではファイルサーバを複数設置することで、高速なファイルサービスを提供している。省電力階層ストレージにおいてはファイルサーバを 2 台設置している。この省電力階層ストレージでジョブをランダムに投入する実験を行ったところ、一時的にジョブのアクセスに偏りが発生し、一部のファイルサーバに負荷が集中することが明らかになった(図 2)。アクセスが偏らないようにジョブのアクセス先データを 2 つのファイルサーバに振り分けても、ランダムなジョブ投入によりアクセスに偏りが発生してしまう。負荷が集中したファイルサーバではファイルサービス性能が低下してしまうことから、ファイルサーバの負荷を分散することでファイルサービス性能を向上することが求められる。

そこで、負荷分散手法として、ジョブスケジューラ連携負荷分散を提案する。ジョブスケジューラ連携負荷分散はジョブ情報・スケジューラ情報とデータ配置情報に基づきファイルサーバの負荷を算出し(図 3(i))、次のジョブ実行前に予めアクセス先データのファイルサービスを負荷の小さいファイルサーバに移動する(図 3(ii))。移動完了がジョブ実行に間に合わなかった時に限り、ジョブ実行を遅延させ(図 3(iii))、移動完了後にジョブは実行開始する(図 3(iv))。結果としてジョブは負荷の小

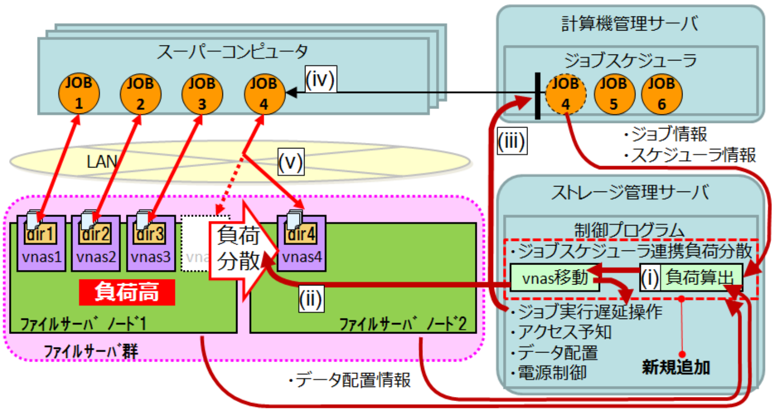

図 3. ジョブスケジューラ連携負荷分散の動作

さいファイルサーバにアクセスするため(図 3(v))、負荷分散したことになる。本手法はジョブスケジューラとの連携でジョブ実行前に予め負荷を分散でき、ジョブの単位でファイルサービス性能を向上できる特徴がある。

本手法の実装として試作機の制御プログラムにジョブスケジューラ連携負荷分散を新規追加した(図 3)。ファイルサーバ間のファイルサービス移動(図 3(ii))は、ファイルサーバを仮想化した Virtual NAS(以下、vnas)を用いて実装した。vnas はファイルサービスを継続したままファイルサーバ間を移動できる。ただし、移動中はファイルサービスの性能低下が発生する。省電力効果向上のために今回は vnas を使用し、ファイルシステムが NL の電源境界(図 1 中(2)(3))を跨がない様に vnas がボリュームをマウントして管理する実装とした。なお、vnas 以外にもファイルサーバにファイルサービスを移動する機能があれば、本手法を適用可能である。

4. 負荷分散手法の効果の検証

負荷分散の動作確認として、表 1 の条件で実験を行った。今回は簡単のため、ファイルサーバの負荷を vnas 数と定めた。比較のため、負荷分散がある場合とない場合で、同じ条件の実験を行った。J.Jann らにより、スーパーコンピュータで実行されるジョブの投入間隔と実行時間は超アーラン分布に従うことが示されており[2]、投入間隔と実行時間に超アーラン分布に従うランダムな値を指定した。図 4 は実験結果の一部で、横軸に実験開始後の経過時間、縦軸にファイルサーバの各ノードについての vnas 配置と負荷を時系列で示している。vnas は全部で 16 個あり、vnas 配置では起動中を実線、停止中を破線で示した。負荷はジョブがアクセスしている vnas の数を示している。vnas の初期配置としては、両ノードに 8 個ずつ均等に vnas を配置した。なお、今回はスーパーコンピュータの 12CPU を用いて実験を行った。vnas 数が両方のノードで 6 以下であれば負荷分散の成功を意味する。

負荷分散なしの場合、時刻 A に負荷が 7 となり、負荷オーバーが発生した。一方で、負荷分散ありの場合、時刻 A の 1 分前に制御プログラムが vnas 移動を行い、負荷分散に成功した。約 12 時間の実験中に vnas 移動は 2 回発生した。結果として負荷分散なしの場合、全投入ジョブの中で高速ファイルアクセスを行ったジョブは 87%で、残り 13%が負荷オーバーにより性能低下したファイルサーバにアクセスした。一方で、負荷分散ありの場合、高速ファイルアクセスを行ったジョブは 100%に改善した。

表 1. 実験条件

実験条件	設定
実験時間	約 12 時間 (100 ジョブ投入)
vnas	vnas 数=16 (実験開始時は 2 台のファイルサーバに各 8 個割当)
ジョブ	各ジョブは 1 つの vnas にアクセス。スーパーコンピュータは同時に最大 12 ジョブを実行。
ジョブ投入間隔	平均投入間隔 = 1 job/6 (min) (超アーラン分布からランダムにサンプリング)
ジョブ実行時間	平均実行時間 = 60 (min)/1 job (超アーラン分布からランダムにサンプリング)

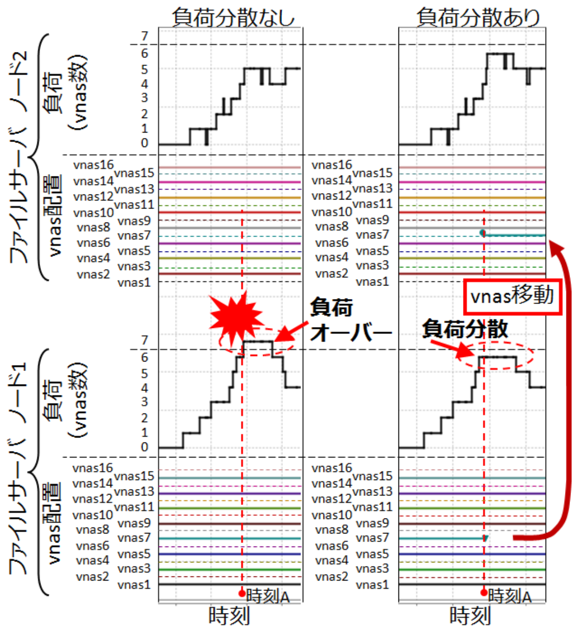

図 4. 実験結果

5. まとめ

省電力階層ストレージの性能向上のための負荷分散手法であるジョブスケジューラ連携負荷分散を実装し、動作確認を行った。実験の結果、ジョブスケジューラ連携負荷分散は正常動作し、ファイルサーバ間で負荷が分散することを確認した。今後は、実ジョブを用いた実験で実際に性能を測定し、性能向上を詳しく評価していく。

謝辞 本研究は、文部科学省の委託研究「高機能・超低消費電力スピンデバイス・ストレージ基盤技術の開発」の成果の一部である。

参考文献

[1] 赤池洋俊,藤本和久,岡田尚也,三浦健司,村岡裕明, “HPC 向けストレージの省電力化を図るアクセス予知階層ストレージの予知成功確率改善手法と効果の検証”, 第 72 回情報処理学会全国大会, 2010 年 3 月

[2] J.Jann, P.Pattnaik, H.Franke, et al, “Modeling of Workload in MPPs.”, LNCS, Vol 1291/1997, pp. 95-116, 2006.