4A-3

HPC 向け省電力階層ストレージにおける 性能スケーラビリティの検証

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 横浜研究所† 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，高性能と消費電力削減を両立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．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，ジョブ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．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]．

スーパーコンピュータとストレージの間には，一般的にファイルサーバが複数設置されており，高速なファイルサービスを提供している． 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全体を省電力階層ストレージと呼ぶ．試作機に設置のファイルサーバは 2 台である．この複数ファイルサーバ間の負荷分散手法として，我々はジョブスケジューラ連携負荷分散を提案している[2]．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 2 中(i))，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 2 中(ii))．移動完了までジョブ実行を遅延させ(図 2 中(iii))，移動完了後にジョブは実行開始する(図 2 中(iv))．結果としてジョブは負荷の小さいファイルサーバにアクセスする (図 2 中(v))．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．しかし，ジョブ実行の遅延操作により，スーパーコンピュータの CPU がアイドル状態となるため，利用率が低下してしまう問題があった．

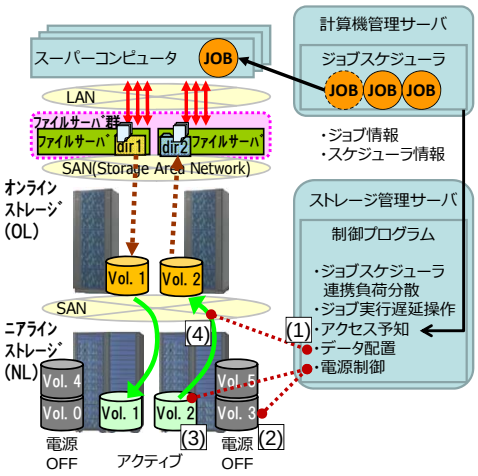

図 1 省電力階層ストレージの概要

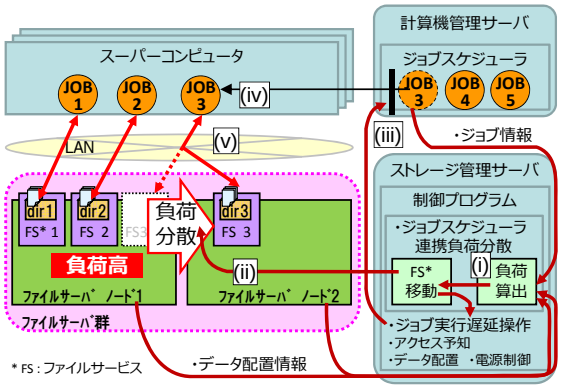

図 2 ジョブスケジューラ連携負荷分散の動作

3. 負荷分散アルゴリズムの改善

従来のアルゴリズムを図 3(A)に示す．従来のアルゴリズムでは，ジョブ実行開始の直前(時刻 t2)にファイルサーバにアクセスしている実行中ジョブの数(時刻 t2 負荷算出対象)を負荷として計算し，負荷の小さいファイルサーバにファイルサービスを移動する負荷分散処理を行っていた．処理が完了するまでは，ジョブ実行の遅延操作によりジョブはキュー内で待機状態となるため，負荷分散の開始(時刻 t2)から終了(時刻 t3)まで CPU がアイドル状態となり，利用率の低下が発生する．負荷分散処理の所要時間は 2 分程度である．表 1 の条件で実験を行ったところ，利用率の低下は 2%であった．スーパーコンピュータの CPU 数が 1024 個の場合，これは 20 個の CPU がアイドルすることに相当し，無視できない大きさである．

そこで，利用率低下を防止する新しい負荷分散アルゴ

Verification of Performance and Scalability of an Energy-efficient High Speed Tiered-Storage System with Proactive Migration for HPC Systems

† Hirotoshi Akaike, Yokohama Laboratory, Hitachi, Ltd.

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

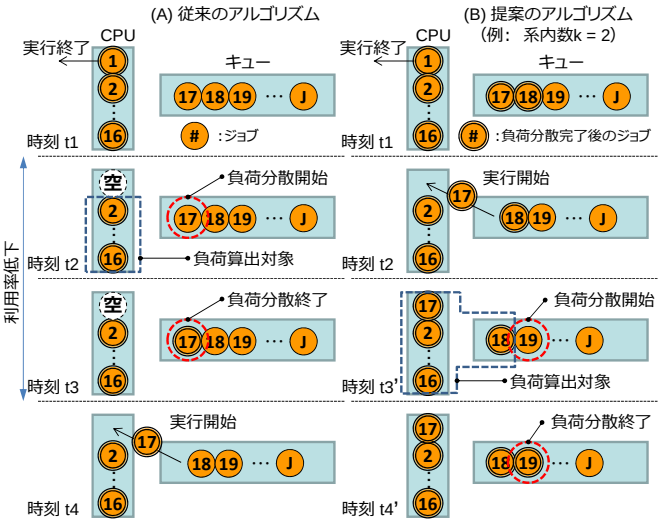

図 3 アルゴリズムの動作例

リズムを提案する．提案のアルゴリズムは，実行開始直前のジョブではなく，先頭から k 番目(系内数 k)のジョブに対して負荷分散処理を行う．提案のアルゴリズムの動作例を図 3(B)に示す．例では k=2 の場合を示している．時刻 t3’においてジョブ 19 が先頭から 2 番目の位置に到達した．この時，ファイルサーバにアクセスする実行中ジョブの数とキューの先頭から k-1 番目までのジョブの数の合計(時刻 t3’負荷算出対象)を負荷の推定値として計算し，負荷の小さいファイルサーバにファイルサービスを移動する負荷分散処理を行う．ジョブ 19 の実行開始までに負荷分散が終了(時刻 t4’)すれば，ジョブ実行の遅延操作の必要はなく，利用率の低下は発生しない．

従来アルゴリズムではジョブ実行開始の直前に負荷分散処理を行うので，実行中ジョブの最新の負荷を計算することで最適な負荷分散を実行することができる．しかし，提案アルゴリズムでは利用率の低下を防止するために事前に負荷分散処理を行うので，負荷を推定して計算する必要があり，誤差により適切に負荷が分散できない場合がある．特に，スーパーコンピュータの規模が大きく(すなわち最大のジョブ実行数が大きい)，ジョブの実行開始頻度が高い場合には，利用率の低下を防止するために系内数 k を大きく設定する必要があり，負荷の推定の誤差が大きくなってしまう．そこで，スーパーコンピュータの規模を拡大した場合の転送速度を評価することで，負荷分散の効果を検証する．

4. スケーラビリティの検証

提案のアルゴリズムを試作機に実装し，表 1 の条件で実験を行った．ここで，利用率の低下は 0.1%以内を目標とした．スーパーコンピュータの CPU 数が 1024 個の場合，これは約 1 個の CPU がアイドルすることに相当し，十分小さい値である．表 1 の条件下では利用率の条件を満たす系内数 k は 2 以上であったので，k=2 と設定した．ファイル書き込み時の転送速度を測定した結果を表 2 に示す．実験の結果，提案アルゴリズムは利用率の低下を防止しながら，従来アルゴリズムと同等の転送速度を示し，負荷分散なしの場合と比較して転送速度は 9%向上した．

次に，スーパーコンピュータの規模を拡大した場合の転送速度をシミュレーションにより評価した．ファイル

表 1 実験条件

実験条件	設定
実験時間	約100時間 (1000ジョブ投入)
ジョブ	スーパーコンピュータは同時に最大16ジョブを実行
ジョブ 投入間隔	平均投入間隔 = 1 job/4 (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)

表 2 実験結果

転送速度の評価結果			
負荷分散	従来アルゴリズム	提案アルゴリズム	なし
転送速度 (比率)	1.0	0.99	0.91

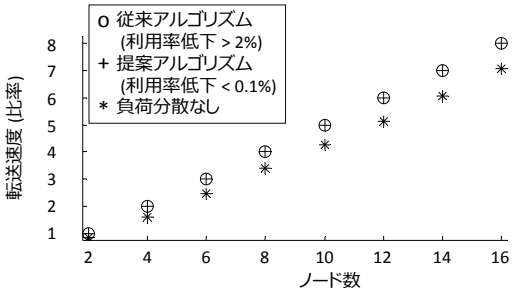

図 4 シミュレーションによる転送速度の評価結果

サーバのノード数はスーパーコンピュータの規模に比例して設定した．ただし，ノード数は試作機で用いているファイルサーバの最大構成を上限として，最大 16 ノードとした．また，シミュレーションでは，スーパーコンピュータとファイルサーバ間ネットワークのデータ転送の帯域不足など，転送速度のボトルネックが存在しない理想的な場合を仮定した．系内数 k は，各ノード数において利用率の条件を満たすように設定した．評価結果を，従来アルゴリズムでノード数=2 の転送速度を 1 とした時の比率で図 4 に示す．その結果，規模を拡大しても従来アルゴリズムと同等の転送速度を示し，ノード数とともに比例する転送性能を示した．

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について，スーパーコンピュータの利用率低下を防止する負荷分散アルゴリズムを提案した．実験の結果，提案のアルゴリズムは利用率の低下を防止しながら，転送速度を向上することを確認した．また，シミュレーションの結果，ノード数が 2～16 の範囲でノード数とともに比例する転送性能を示し，性能スケーラビリティを確認した．今後は，提案の負荷分散アルゴリズムを一般的なスケジューラ構成である複数キューに対応させることが課題となる．

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] 赤池洋俊，藤本和久，黒川大樹，三浦健司，村岡裕明，“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”， 第 73 回情報処理学会全国大会， 2011 年 3 月．