C-046

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

Study on a Power-aware Proactive Storage-tiering Management with Access Prediction for an Energy-efficient High-speed Tiered-Storage System (eHiTS) for HPC Systems.

岡田 尚也†　藤本和久†　赤池 洋俊‡　三浦 健司†　村岡 裕明†
Naoya Okada,　Kazuhisa Fujimoto,　Hirotoshi Akaike,　Kenji Miura,　Hiroaki Muraoka

1．はじめに

情報インフラが整備された現代社会において扱われる情報量が飛躍的に増加しており、それらを保存・蓄積するためのストレージの需要が高まっているとともに省電力性を配慮したシステムが強く求められている。システムが消費する電力の大部分は HDD によるものであり HDD の稼働制御によって消費電力を抑える手法が必要となっている。省電力アーキテクチャーとして提案された MAID(Massive of Arrays of Inactive Disks) [1]ではアクセスの来ない HDD をスピンダウンさせることによって消費電力低減を図るものの、停止状態の HDD へアクセスがあった場合の性能低下が著しく用途が限られていた。そこで我々は高速オンラインストレージ(OL)と大容量低電力ニアラインストレージ(NL)を階層化したストレージシステムにおいて、階層間で最適なデータ配置を行い性能と省電力の両立を図る。最適なデータ配置を行うためにアプリケーションからのヒントを用いてストレージへのアクセス予測を行う。本報告ではアクセス予測の予測確率を向上させる手法を検討し、その効果と影響について考察を行った。

2．従来の予測手法の問題点

Fig. 1 に示すように、ユーザーによって投入されたジョブはジョブスケジューラーによって管理されている。ジョブが投入された後、実行開始されるまでにジョブがアクセスするデータを NL から OL へコピーを完了することにより OL 上で高速な入出力を実現する[2]。ジョブがキューに投入され、実行が終了するまでの一連の過程は待ち行列理論の M/M/1 出生死滅過程モデルで説明される。ある時刻 t においてジョブがキューに投入されたとすると、そのジョブが実行開始されるまでの待ち時間 T_{wait} 内に、ジョブがアクセスするデータの NL から OL へのコピー開始を完了させなければならない。すなわちコピーに要する時間 T_{copy} が待ち時間 T_{wait} 以下であれば予測成功(OL 上での高速な入出力が可能)となる。よって予測成功条件は(1)式で表される。

$$T_{wait} \geq T_{copy} \qquad (1)$$

単位時間当たりのジョブの実行回数がランダム(ポアソン分布に従う)と仮定する。待ち行列が平衡状態にある時、ある時刻で投入されたジョブの待ち時間 T_{wait} がコピー時間 T_{copy} 以上である事象が起きる確率は(2)式で表される[3]。

$$P(T_{wait} \geq T_{copy}) = \rho \cdot e^{-(1-\rho)\mu T copy} \qquad (2)$$

ここで ρ はジョブの単位時間当たりの平均ジョブ実行回数と平均投入回数の比である利用率を表している。(2)式より成功確率は利用率 ρ とコピー時間 T_{copy} に依存していることがわかる。つまりこれはシステムの利用率 ρ が低下した場合、もしくはコピー時間 T_{copy} が増大した場合、予測成功確率が低下してしまうことを意味している。大容量高速転送を常に維持しなければならない HPC において、予測失敗による待ち時間発生のペナルティは致命的であるため、予測確率を向上させる手法について検討を行う。

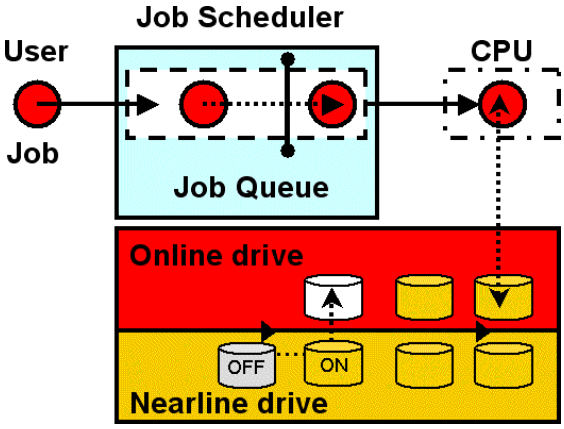

Fig. 1 提案システムの概要図

3. 予測確率向上手法

予測成功確率を大きく落としている要因の一つとして待ち行列の長さが挙げられる。ジョブの投入間隔と実行時間は分布を持っているため、場合によっては処理間隔に大きな偏りが出てしまい、その結果キューにジョブが溜まらなくなってしまうことがある。キューが空で実行中のジョブもない場合、もしくはキューの長さが極端に短くなってしまった場合、その時点で投入されたジョブは即座に実行に移る、または待ち時間がコピー時間より短くなるため、必要なデータの OL へのコピーを実行前に完了できない。そこで待ち行列がある一定の長さを下回った時においても、ジョブ実行開始前に確実にデータコピー時間を完了させるために、その場合にはジョブ実行をデータコピーに要する時間だけ遅延させる方法を考えた。ここでジョブ実行遅延を実施する条件として、系の待ち行列の長さの下限値 k_{length} を考える。Fig. 1 の Job Scheduler に予測確率改善手法を示す。キューの長さが下限値 k_{length} を下回った時点でキューに入ってきたジョブに対し一定の遅延時間すなわちデータのコピー時間 T_{stop} を与え、その時間だけジョブの実行を遅延させる。この機構をシミュレーションモデルに組み込み、待ち行列に様々な下限値 k_{length} を設定し遅延時間を与えた時に予測確率や待ち時間に与える影響ついて検討を行った。

3.1 シミュレーション手法

J. Jann らにより、並列計算機におけるジョブ発生間隔及び実行時間は超アーラン分布に従うことを示されている[3]。

実際に近い条件下での予測確率を求めるため我々はイベント駆動型シミュレータでジョブスケジューラーと計算機を模擬し、ジョブ発生間隔と実行時間に超アーラン分布の特別な場合である超指数分布を用いてシミュレーションを行い提案手法の予測確率を評価した。待ち行列の系を平衡状態にするため $1.6*10^7$ min のウォームアップ時間を与え、40 万個のジョブについてサンプリングを行った。ジョブの到着間隔と実行時間にそれぞれ Table 1, 待ち行列の下限値 k_{length} に 0, 1, 2 を与えた。遅延時間 T_{stop} は(3)式のように与えられる。

$$T_{stop} = \begin{cases} 0 & (k_{length} < k) \\ T_{copy} & (0 \le k \le k_{length}) \end{cases} \quad (3)$$

条件を満たしたときのみに与えるコピー時間 T_{copy} はディスク電源投入時間 T_{spinup} とデータ転送時間 $T_{Transfer}$ の和で表され、(4)式のように表される。

$$T_{copy} = T_{spinup} + \sum_{File} T_{Transfer} \quad (4)$$

ディスク電源投入時間 T_{spinup} は一定である。データ転送時間 $T_{Transfer}$ は分布を持ったファイル数とファイルサイズに比例した転送時間の積算で求められる。これは、ジョブ毎に試用するファイル数とファイルサイズが異なるためである。ファイル数と 1 つあたりのファイル転送時間の積を計算し $T_{Transfer}$ を与える。個々のジョブについて (1)式の予測成功条件に合致するジョブを数え成功確率を算出した。

Table 1. シミュレーションに用いたパラメータ

	ジョブ到着間隔	ジョブ実行時間
分布	超指数分布	超指数分布
平均	60 min	54 min
標準偏差 / 平均値	2.38035	2.07285
T_{spinup}	0.5, 1.0, 5.0 min	
File 数	1 ~ 20	
File size	2GB (平均)	
データ転送レート	100 MB / sec	

3.2 シミュレーション結果

シミュレーションを行った結果 Fig. 2 と Fig. 3 を得た。

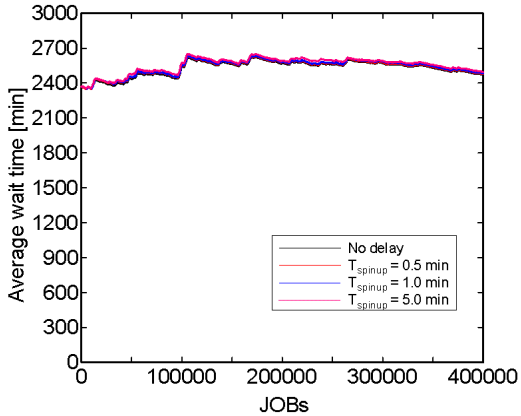

Fig.2　ジョブの平均待ち時間の推移

Fig.2 は T_{spinup} を変化させた時のジョブの平均待ち時間の推移を表している。T_{spinup} を変化させた場合、T_{spinup} に応じた時間だけ平均待ち時間が延びている。利用率が 0.9 と高く、平均待ち時間が 2400 分前後と非常に大きいため T_{spinup} を変化させても変化量が非常に小さいことがわかる。

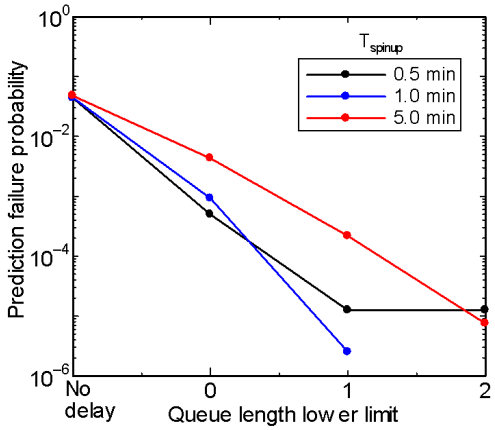

Fig. 3　キューの下限値 k_{length} と予測失敗確率の関係

Fig. 3 はキューの下限値を変化させた時の予測失敗確率を表している。T_{spinup} を下げることによって失敗確率を低下させることが出来る。つまりディスクのスピンアップに時間がかからないため、速やかにデータ転送時間が完了させることができるので結果予測が失敗する場合がすくなるためだと考えられる。以上から提案手法を用いることで利用率が大きい場合に限っては待ち時間を大幅に増大させることなく予測成功確率を大幅に向上させることができると示唆される。

4．まとめ

待ちの長さの下限値を 1 以上にすると予測失敗確率を大幅に低下させることが出来た。向上手法を適用した場合の平均待ち時間及び実際の待ち時間の、適用しない場合に対する伸びは非常に微小なのでユーザーの待ち時間に与える影響は小さいと考えられる。

5. 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・ストレージ基盤技術の開発」の援助を得て行った。ここに謝意を表する。

参考文献

[1] Dennis Colarelli, Dirk Grunwald, "Massive Arrays of Idle Disks For Storage Archives," sc,pp.47, ACM/IEEE SC 2002 Conference (SC 2002), 2002

[2] Kazuhisa Fujimoto, Hirotoshi Akaike, Naoya Okada, Kenji Miura, and Hiroaki Muraoka, “Power-Aware Storage-Tiering Management for High-Speed Tiered-Storage Systems”, http://www.usenix.org/events/fast09/wips_posters/ fujimoto_poster.pdf, USENIX Conference on File and Storage Technologies (FAST'09), Feb, 2009.

[3] 森村英典, 大前義次, “応用待ち行列理論”, 日科技連出版社, 1975.

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

†東北大学, 電気通信研究所, RIEC

‡株式会社日立製作所, システム開発研究所