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における予測確率に関する検討

岡田　尚也†, 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

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。蓄積される情報量の増大によりシステムに内蔵されるHDD(ハードディスクドライブ)の数が増えたため消費電力が急激に増加した。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)を階層化したシステムにおいて、階層ストレージの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、対象アプリケーションにHPCを選択しスーパーコンピューターのジョブスケジューラーの情報をヒントとしてアクセスを予測する手法を提案し、提案システムをモデル化した解析式及びシミュレーションにより予測確率について検討した結果について報告する。

2.従来システムの課題と提案手法

データセンターではOLとNLに階層化した階層ストレージシステムが用いられている。OLには高性能なHDDが用いられており、高性能である反面容量当たりの消費電力が大きいという問題がある。NLは安価で大容量なHDDから構成されており容量当たりの消費電力が小さいという特徴がある。データセンターではILM(Information Lifecycle Management)と呼ばれるデータ管理により、アクセス履歴から使用頻度が低いと判定されたデータをOLからNLへ移動することにより消費電力が大きいOLのHDD数を抑え、システム全体として省電力化を図っている。OLのHDD数を抑えすぎるとOLに無いデータにアクセスが発生する確率が高まりその場合、低速なNLから対象データを読む、またはOLへ対象データをコピーしなければならないので性能ペナルティが生じる。

この問題を解決するためFig.1に示すストレージシステムを提案した。データはNLに格納しHDDは停止状態とする。ジョブ実行開始までに対象データを格納したNLのHDDをスピンアップさせ、OLへのデータのコピーを完了させることによって、電力消費を抑えつつデータの高速転送を維持する。NL上の対象データのコピー開始タイミング予測は、ジョブスケジューラーのキュー情報から待ち行列理論により計算した、ジョブ実行開始までの待ち時間をもとに行う。待ち時間内にNLのHDDのスピンアップとNLからOLへのデータコピーを完了できるようコピーを開始する。ジョブ実行前にOLにデータのコピーを完了させることが出来れば予測成功となる。予測の成功確率を上げることにより、性能を犠牲にすることなく省電力性を向上させることができる。

3.予測確率の見積もり

3.1解析的手法による見積もり[2]

ジョブをキューに投入、実行したのちジョブが消滅するまでの一連の過程は待ち行列理論における出生死滅過程モデルで説明される。解析的に予測確率を求めるため、待ち行列の系はFCFS(First Come First Serve)のM/M/1モデルであるとする。ある時刻tにおいて、あるジョブがアクセスするデータのOLからNLへのコピー開始を指示したとすると、そのジョブが実行開始されるまでの待ち時間γの間までにコピー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_{th}が待ち時間γ以下であれば予測成功となる。単位時間当たりのジョブの実行回

Storage-tiering management for high-speed mass storage systems by predicting job execution start times in HPC systems
† Naoya Okada, Kazuhisa Fujimoto, Kenji Miura, Hiroaki Muraoka, RIEC, Tohoku University.
‡ Hirotoshi Akaike, Systems Development Laboratory, Hitachi Ltd.

数がランダム(ポアソン分布に従う)と仮定する。時間間隔 t_{th} の間に k 個のジョブが完了する確率は以下の式で表される。

$$v_k(t_{th}) = e^{-\mu} \cdot \frac{(\mu t_{th})^k}{k!} \quad (1)$$

ここで μ は単位時間当たりの平均ジョブ実行回数である。また、ある時刻で予測対象のジョブの前に、計算実行中のジョブを含め k 個のジョブがある確率を P_k とすると、予知が成功する確率は以下の式で表される。

$$P(\gamma \geq t_{th}) = \sum_{k=1}^{\infty} p_k \sum_{r=0}^{k-1} e^{-\mu t_{th}} \frac{(\mu t_{th})^r}{r!} \quad (2)$$

待ち行列が平衡状態にあるとき P_k は以下の式で表される。

$$p_k = (1-\rho) \cdot \rho^k \quad (3)$$

以上より、予知成功確率は以下の式で表される[2]。

$$P(\gamma \geq t_{th}) = \rho \cdot e^{-(1-\rho)\mu t_{th}} \quad (4)$$

3.2 シミュレーションによる検討

J. Jann らは、並列計算機におけるジョブ発生間隔及び実行時間が超アーラン分布に従うことを示している[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、超アーラン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。

4. 検討結果

ジョブ発生間隔とジョブ時間が指数分布に従う場合と超アーラン分布の特別な場合である超指数分布に従う場合について、前者は解析式(4)、後者はシミュレーションから予測が失敗する確率を求めた。それらの結果を Fig.2 に示す。横軸は、平均実行時間(1/μ)に対するコピーに要する時間(t_{th})の比率(μt_{th})である。失敗確率は μt_{th} が小さく、利用率が高いほど低くなることがわかった。また、指数分布に比べて超指数分布の失敗確率が低く出ているのは超指数分布の標準偏差が平均値の 2 倍以上と大きいためジョブ発生間隔及び実行時間に大きなばらつきが生じ、指数分布の場合に比べ平均の待ち時間が長くなるためであると考えられる。しかしながら、利用率 ρ が 0.967 と 1 に近い場合でも失敗確率は 1.7%と高く、実用に耐えると考えられる 10^{-3} %以下の確率に抑えるためには方式の改善が必要である。

5. まとめ

階層間でデータの最適配置を行う階層ストレージにおいてデータ最適配置のためのアクセス予測手法を提案し、提案手法の予測確率について評価した。予測確率を実用レベルまで向上する方式改善のための検討を進めていく予定である。

6. 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。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]森村英典, 大前義次, "応用待ち行列理論", 日科技連出版社, 1975.

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

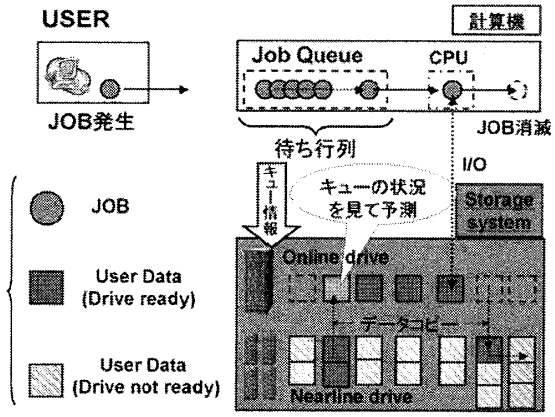

Fig.1 提案システム概要図

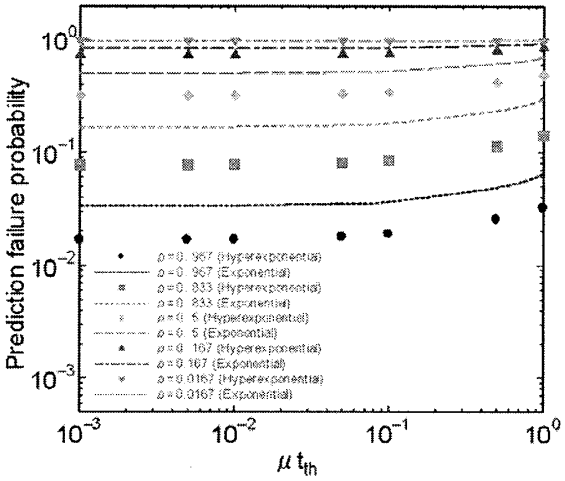

Fig.2 指数分布と超指数分布の失敗確率