大規模並列システムにおける共有仮想記憶のページ配置方式

1B－5

正木 宏和　宮田 忠明　工藤 智子　芦原 評　清水 謙多郎

電気通信大学 情報工学科

1　はじめに

プロセッサ(PE)ごとにディスクを持たない大規模並列システムでは、従来のようにページングにおける二次記憶としてディスク装置だけを用いていたのでは主記憶とディスクとの距離やディスクのアクセス時間の遅さがあって効率が悪い。そこで、遠くのディスクではなく、より近くの他のPEの主記憶を二次記憶として利用することが考えられる。

本稿では、大規模並列システムにおいて共有仮想記憶を実現するためのページ配置方式として考えられるものを分類・整理し、それらから有効と考えられるものをシミュレーションを用いて比較・評価する。

2　ページ配置方式

ページ配置方針について検討すべき項目およびそれぞれについて考えられる方式をまとめると以下のようになる。

1. ページの分類
   - 局所ページ：自PEでのみ参照されるページ
   - 退避ページ：他のPEから退避され、自PEでは参照されることのないページ
   - 共有ページ：複数のPEにより共有されるページ
2. ページアウトのタイミング
   - ページフォールトが起こりページ枠が不足した時にページを退避する。
   - 常に空きページ枠を確保するために前もってページを退避する。
3. 置き換えるページの選択

(a) 適用するページ置き換えアルゴリズム
LRU、FIFO、ワーキングセット法 等

(b) アルゴリズムの適用方針
   - 局所ページ、共有ページ、退避ページに対し対等に適用する。
   - 退避ページを局所ページより優先する。退避ページを更に退避する場合には、退避元PEからの距離を考慮する。
   - 共有ページのコピーが最も近くに存在するものを優先する。コピーが存在する場合は、単にそのページを破棄する。

4. 退避先PE選択のための情報収集
   - 周期方式：各PEがページ受け入れのために定期的にそのPEに関する情報を周辺のPEに知らせる方式
   - 事象駆動方式：各PEがページ受け入れのために状態変化時にそのPEに関する情報を周辺のPEに知らせる方式
   - 要求時方式：退避側PEが必要とする時に他のPEにそのPEに関する情報を入札によって要求・獲得する方式
5. 獲得する情報
   上記PEに関する情報として次のようなものがある。
   - 空きページ数：そのPEの空きページ数
   - 空きページを持つPEまでの距離：そのPEから空きページを持つPEまでの距離
6. 退避ページ枠確保のタイミング
   退避先PE(の候補)における退避ページ枠の確保の方式として次のようなものが考えられる。
   - 確保なし：退避ページ枠を確保せずに直接ページを転送する。
   - 要求時確保：ページを転送する際に退避ページ枠を確保する。
   - 先行確保：ページの転送に先立って退避ページ枠を確保する。
7. 退避先PEの選択
   実際の退避先を選ぶ際の基準として次のようなものがある。
   - 空きページ数が大きい順
   - 距離が小さい順
   - ランダム
8. 退避失敗時の処理
   退避が失敗した時、以下の方針が考えられる。
   - 選択されたPEのページを置き換える。
   - 退避元からもう一度退避先PEを選択する。
   - 選択先PEから次の退避先PEを選択する。
   - すぐにディスクに格納する。

Page Placement Algorithms for Shared Virtual Memory in Massively Parallel Systems
Hirokazu Masaki, Tadaaki Miyata, Tomoko Kudo,
Hyo Ashihara, Kentaro Shimizu
Department of Computer Science, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo, Japan

3 性能評価

3.1 シミュレーションモデル

2次元トーラスにより結合された1024個のPEからなる分散メモリ型マルチプロセッサシステムを想定する。ディスク装置は、トーラスの縦横の一辺のPE(32個)に1台ずつ、計63台備えられているものとする。ディスクへの1ページの読み書きと通信リンクを介した隣接PEへのページ転送との速度比は30:1とする。また、PEおよび通信リンクの故障はないものとする。性能の指標として平均ページアクセス時間を用いる。ここでいうシステム負荷は、全PE中のタスクが割り当てられたPEの割合で与えるものである。図1,2において、「ディスクのみ」は二次記憶としてディスクだけを用いる方式を、「ランダム」は退避先PEを全くランダムに選択する方式を表しており、これらは今回、比較の対象として載せた。

3.2 退避先PEのための情報収集

退避先PEのための情報収集(2節4. 参照)の方式に関するシミュレーションの実行結果を以下に示す。周期方式については、明らかに効率が悪いと考えられるので省略した。図1は、要求時方式および事象駆動方式について、システム負荷に対する平均ページアクセス時間の変化を示したものである。事象駆動方式については、獲得する情報として、空きページ数(空き)と、空きページをもつPEまでの距離(距離)の双方についての結果を示している。退避ページ枠確保のタイミング(2節6. 参照)は、先行確保方式を用いた。

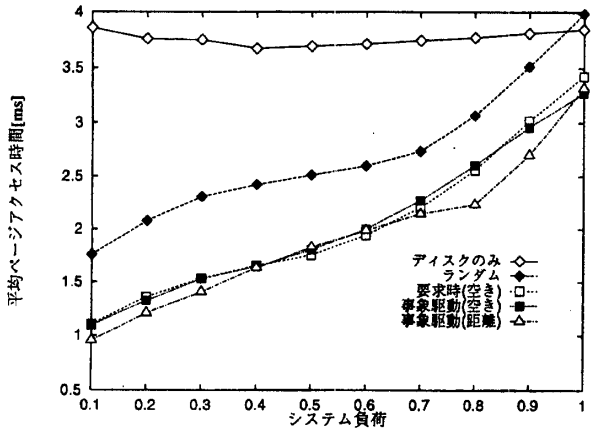

図 1: 退避先PEのための情報収集

図1からわかるように、事象駆動(距離)方式は、他の2つの方式に比べて良い性能を示している。これは、情報の性質上、事象駆動(距離)方式はシステム内全体の空きページ枠を見渡せることのよるものと考えられる。

3.3 退避ページ枠確保のタイミング

退避ページ枠確保(2節6. 参照)のタイミングについて、システム負荷に対する平均ページアクセス時間の変化を示したものが図2である。評価する3つの方式について、退避先PEは距離が小さい順で選択するものとした。

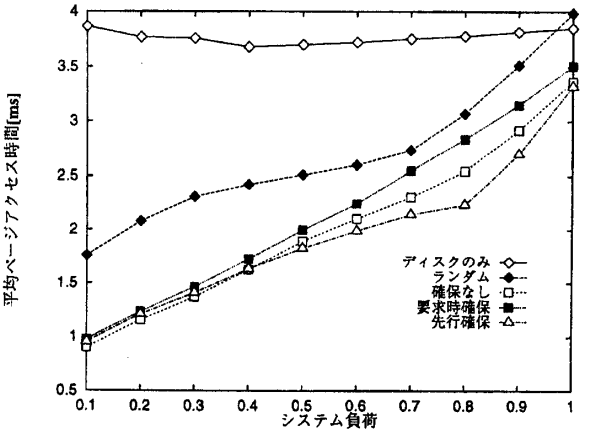

図 2: 退避ページ枠確保のタイミング

図2からわかるように、負荷が低い時には、システム内の空きページも多いため、シンプルな方式である確保なし方式が良い性能を示す。負荷が高くなると、空きページ枠も減少し、直接ページを転送しても退避に失敗する可能性が出てくる。そのため、先行確保方式が良い性能を示す。要求時確保方式は、退避の失敗はないが、確保のためのオーバヘッドがかかり良い性能が得られない。

4 まとめ と 今後の課題

分散メモリ型マルチプロセッサ上でのページ配置方針を扱った論文として[1, 2, 3]があるが、本研究のようにページ配置方針について広範な分類・評価を行なったものはない。

今後、さらに、今回性能評価できなかった共有ページや、他の相互結合網トポロジ、負荷分散への応用を加味した方式なども含めて検討していく予定である。

参考文献

[1] 平野, 一杉, 田沼, 須崎. 超流動 OS の大域的仮想仮想記憶におけるページ探索法の比較. 情報処理学会研究会報告 93-OS-61(SWoPP'93),pp.65-72,1993.

[2] M.Malkawi, D.Knox, and M.Abaza. "Dynamic Page Distribution in Distributed Virtual Memory systems". *Proc. of 4th ISMM International Conference on Parallel and Distributed Computing and Systems*, pp.87-91, 1991.

[3] M.J.Feeley, W.E.Morgan, F.H.Pighin, A.R.Karlin and H.M.Levy."Implementing Global Memory Management in a Workstation Cluster". *Proc. of the Fifteenth ACM Symposium on Operating Systems Principles:SIGOPS'95*, pp.201-212,1995.