システムソフトウェアと
オペレーティング・システム 76－13
（1997．8．21）

大規模並列システムにおける大域的ページ配置方針の性能評価

宮田忠明　正木宏和　芦原評　清水謙多郎
電気通信大学 情報工学科

概要

本論文では，分散メモリ型大規模並列システムにおける大域的ページ配置方針について論じる．各ノードの主記憶を仮想記憶の一時的な退避先として割り当てる大域的ページ配置方式は，従来の仮想記憶におけるディスクアクセスをより高速なノード間のメモリアクセスに置き換えることができる．限定されたノードにしかディスクを持たない大規模並列システムでは，大域的ページ配置は特に有効と考えられる．本論文では，状態情報の収集，置き換えページの選択，ページを転送するノードの選択などについて，大規模システムを想定した大域的ページ配置の様々な方針を提案し，これらの性能をシミュレーションを用いて比較・評価する．

Performance Evaluation of Global Page Placement Policies in Massively Parallel Systems

Tadaaki Miyata, Hirokazu Masaki, Kentaro Shimizu, and Hyo Ashihara

Department of Computer Science
The University of Electro-Communications

Abstract

This paper describes global page placement policies in distributed memory massively parallel systems. Global page placement provides an alternative to the expensive disk swapping in conventional virtual memory systems by taking advantage of faster internode memory access. In support of global page placement, we have developed various policies for state information exchange, selection of a page for replacement, selection of a node to which a page should be sent, and so on. We have evaluated them by using a simulation method.

1　はじめに

分散メモリ型並列システムにおいて，分散するノードの主記憶(ローカルメモリ)にページを割り当てる大域的ページ配置は，主記憶を他ノードの一時的なスワップ領域(ページ退避領域)として利用することを許すもので，これにより，従来の仮想記憶ではボトルネックとなるようなディスクアクセスを，より高速なノード間のメモリアクセスに置き換えることができる．この大域的ページ配置により，ページ退避領域をノード間で分散して保持する分散仮想記憶(DVM)が実現される．大域的ページ配置は，いくつかの特定のノードしかディスクを持たない大規模並列システムにおいて，ページの転送やディスクアクセスにかかるオーバヘッドを減らすことができるため，特に効力を発揮する

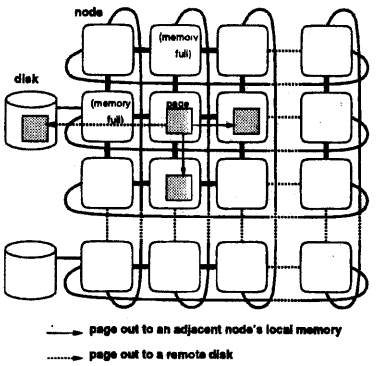

図 1: 大域ページ配置

ものと考えられる．図 1に大規模並列システムにおける大域的ページ配置の機構を示す．

プロセッサ割当てがプロセッサ資源の共有，すなわち負荷分散を直接の目的とするのに対し，大域的ページ配置はメモリ資源の共有を通して負荷分散を実現する．一般に，ページ転送のコストはプロセス移送のコストに比べてはるかに小さいため，大域的ページ配置はプロセッサ割当てよりもきめの細かい資源共有を実現することが期待される．

大域的ページ配置に関する従来の研究としては，次のようなものがある．Clancey ら [1] は，プログラムのページフォールト頻度を用いてページの配置先を決定する静的ページ割当て方針を提案している．この研究は，特定のプログラムの平均応答時間を短縮することが目的である点，提案された方針が静的割当てに基づくものであるため，適用範囲に限界があるという点で，我々の研究と異なっている．Dahlin ら [2] は，Nチャンス転送を用いたリモートメモリ割当てのいくつかのアルゴリズムを評価している．これらのアルゴリズムは，選択された置き換えページがクラスタ内で最後の複製であったとき，ランダムに選ばれたノードにそれを転送し，そうでなければページを破棄するというものである．全てのページは，N回転送されたとき破棄される．Malkawi ら [4] はシミュレーションを用いて大域的ページ配置におけるページ配置方針の評価を行った．その結果，隣接ノードの中で未使用のメモリ領域が最も大きいノードを選択する least loaded neighbor(LLN) アルゴリズムが良い性能を示した．Feeley ら [3] は，ワークステーションクラスタ上の global age information を用いたページ配置アルゴリズムを提案しているが，後述するように，大規模なシステムへの適用は困難である．

本論文では，大規模並列システムを想定した大域的ページ配置に関し，状態情報の収集，置き換えページの選択，ページを転送するノードの選択など様々な方針について提案・分類を行うとともに，これらの性能を共通の条件の下で評価した．性能評価においては，様々なページ配置方針が柔軟に実装できるシミュレータを新たに開発した．

2 大域ページ配置方針

2.1 情報収集方針

効果的なページ配置には他のノードの状態情報 (例えばメモリ使用量) が必要であり，情報収集方針はシステム内の他ノードの状態についての情報をいつどのように収集すべきかを決定する方針である．以下のような方針が考えられる:

- 周期的 – 定期的に状態情報を伝える．この方針はシステム状態の動的な変化に対応しにくい．
- 状態変化時 – 状態が変化したときに状態情報を伝える．この方針は，ノードの情報が変わるたび (例えば，局所メモリの空きページ枠の数が変わったとき) に情報を伝達していてはコストが高すぎる．そこで，我々は，各ノードの空きページ枠の数が高水準値を越えたときと低水準値を下回ったときにのみ情報を伝えるしきい値方針を用いる．
- 要求時 – この方針は，ページアウトをするときに他ノードの情報を収集するものである．我々は，転送先ノードを決めるのに入札アルゴリズムを用いる．

ノードがLANによって結合されている分散システムでは，システムはブロードキャストを用いて状態情報を大域的に収集することができるが，ポイント・ツー・ポイントの直接結合に基づく大規模並列システムの場合ブロードキャストの使用はネットワーク混雑を引き起こす可能性がある．そこで，

我々は，情報交換を制限するために次の2つのモデルを考える．

- ドメインモデル - 各ノードは状態情報の交換やページ転送の範囲を定義したドメインを持つ．各ノードがドメインの中央ノードとして，状態情報を収集し，ドメイン内でページ転送を行う．ドメイン間のオーバラップを許すことにより，メモリ資源を広範囲に共有することを目指す．中央ノードからドメイン内の他ノードへの最大距離(ホップ数)で定義されるドメインサイズは，性能を決定する重要なパラメータである．

- 傾斜モデル - ページを転送する指針として，それぞれのノードが隣接ノードの勾配マップを作る．ノードの proximity は，空きページ枠を持っているもっとも近いノードへの最短距離によって定義される．ノードのローカルメモリに空きページ枠があるときは proximity を 0 にする．空きページ枠がないときは，隣接ノード n_i を持つノード a の proximity は次のように計算する:

$$\text{proximity}(a) = \min_i(\text{proximity}(n_i)) + 1.$$

ノードの proximity が変化したとき，その値を近隣ノードに伝達する．この方式はもともと負荷分散を実現するためのものとして提案された [5].

2.2 ページ置き換え方針

プロセスのスレッドがページフォールトを起こしたとき，要求されたページはディスクまたはリモートメモリからローカルメモリへ転送される．ローカルメモリに空きページ枠がない場合は，置き換えのためのページを選択しなければならない．置き換えアルゴリズムには FIFO(first in first out), LRU(least recently used), LFU(least frequently used) など従来の方針の適用が考えられる．

DVM では，ページ置き換えに関連し，各ノードがローカルメモリに保持するページを以下の2つに分類して扱うことができる:

- 局所ページ - 局所ページはそのノードのスレッドからのみアクセスされる．

- 退避ページ - 退避ページは他ノードのためにローカルメモリに置かれ，そのノードのスレッドには参照されない．

本論文では，退避ページと局所ページの両方に対し対等に LRU を適用する方針を用いる．なお，異なったノード間でページを共有する分散共有メモリについては考えていない．

2.3 ページ配置方針

置き換えにより選択されたページが転送されるべきノードは，ページ配置方針により決定される．ページがメモリに存在している時間を大域的に管理し (global age information)[3]，それに基づいて全てのリモートメモリに対し LRU を適用する方式も考えられるが，大規模並列システムでは現実的ではない．本論文では，ページアウト先として空きページ枠を持ったノードのみを選択することにする．空きページ枠を持つノードが見つからない場合はディスクへページアウトされる．ノードの決定の基準として考えらるものを以下に挙げる:

- 空きページ枠数 - ドメイン内のノードの中で最も空きページ枠数を多く持つノードを選択する．

- 自ノードからの距離 - 空きページ枠を持つノードの中から最も近いノードを選択する．

- ランダム - 空きページ枠を持ったノードの中からランダムに選択する．

空きページ枠数に基づく方針はページアウトの成功率を上げることができる．自ノードからの距離に基づく方針はページ転送のコストを小さくすることができるという点で有利である．それぞれの基準を組み合わせた方針も考えられる．

ページアウトは，選択されたノード (転送先ノード) で空きページ枠がなかったときに失敗する．これは，ノード選択の際に用いた情報が古かったこと，ひいては通信遅延による情報伝達の遅れが原因である．ページアウトが失敗したとき，システ

ムはページを転送すべき他のノードを選択しなければならない (再ページアウトを必要とする). そのようなノード選択のためにいくつかの方針を考えることができる.

- 転送先ノードでページを置き換える. 置き換えられたページがページアウトされる.
- 転送元ノードに戻り他ノードを選択する.
- 転送先ノードで他ノードを選択する (カスケードページアウト).
- ページをディスクに退避する. この方針は, 通常ディスクアクセスの回数を増大させるが, 再ページアウトの発生を防ぐことができる.

2.4 ページ予約方針

再ページアウトは, ページングのオーバヘッドを増大させネットワークを混雑させる原因となる. そのような現象は, 従来の仮想記憶システムにおけるページスラッシングと類似している. 再ページアウトの数を減らすために, 我々はページアウト先としてリモートノードのメモリのいくつかのページ枠を予約する方針を提案する. 以下に, ページ枠を割り当てるタイミングとして考えられる方針を挙げる:

- 予約なし - ページ枠の予約をしないで直接ページを転送する. 情報が古いときにページアウトが失敗する可能性がある.
- 要求時割り当て - ページアウトの要求があったときにページ枠をリモートノードのメモリに割り当てる.
- 先行割り当て - リモートノードのメモリの空きページ枠をページアウトの要求が出される前に予め割り当てる.

3 シミュレーションモデル

シミュレーションでは, 相互結合網が2次元トーラスメッシュ結合である $N \times N$ ノードからなる分散メモリ型並列システムを仮定する. 各ノードは, プロセッサ, 容量 M のローカルメモリ, 4つの隣接ノードへのポイント・ツー・ポイント結合を持つ. ディスクは $n \times n$ のサブメッシュごとに置かれるものとする.

システムへのプロセス到着率はポアソン分布に従う. プロセスは1つ, または複数のスレッドからなる. 各プロセスのスレッド数は正規分布に従う. 各ノードでは同時に1つのスレッドのみが実行される. スレッドが移送されることはない.

各スレッドから出されるメモリアクセスは1つまたは重なり合う複数のフェイズからなっていると仮定する. 各フェイズにおいて, スレッドは局所参照集合 (locality set) と呼ばれるページの集合を参照する. 各フェイズには, そのフェイズのメモリ参照位置の中心を与える placement とその広がりを与える locality が定義される. プログラムの大きさは平均 M の一様分布に従い, 異なるフェイズの placement は仮想アドレス空間内の一様分布に従う. フェイズ内のメモリ参照のアドレスは, フェイズの placement を平均とし標準偏差が locality である正規分布によって決定される. locality を増やすことにより, フェイズ内でより大きな局所参照性がシミュレートされる. メモリアクセスには読み出しと書き込みの2種類があり, 書き込まれたページ (dirty ページ) のみ, 実際のページアウト操作が必要となる.

各ノードはページテーブルと拡張ページテーブルを持つ. ページテーブルについては従来の仮想記憶システムと同じである. 拡張ページテーブルはリモートノードに移動したページの位置を記録しておく.

表1に主なシステムパラメータとシミュレータで用いたデフォルトの値を示す. 特定のパラメータの値を変化させることによりシミュレーションを行った. 本論文ではプロセスの平均応答時間を性能指標とする. ローカルメモリサイズ, 仮想記憶サイズなどのいくつかのパラメータの値は一般的とは言えない. しかし我々は, 基本的なシステムの振舞いはスケールダウンした値でも解析することができると考えている. これらのパラメータは過去のいくつかの論文 [4] でも同様の値が用いられている.

4 結果と考察

表 1: シミュレーションパラメータ

パラメータ	デフォルト値
ノードの数	256 (16 × 16 メッシュ)
ディスクの数	16
1プロセスあたりの平均スレッド数	9
1スレッドあたりのページアクセス数	3600
ページサイズ	8 Kbyte
ローカルメモリサイズ	64 ページ
仮想記憶サイズ	1 ～ 128 ページ (正規分布)
フェイズ長	60 アクセス
locality	仮想記憶サイズの 10.0%
read/write 比	10 : 1
ネットワーク速度	100 MBytes/sec
ドメインサイズ	4
情報収集方針	状態変化時
ページ配置方針	ランダム, カスケード
ページ予約方針	予約なし

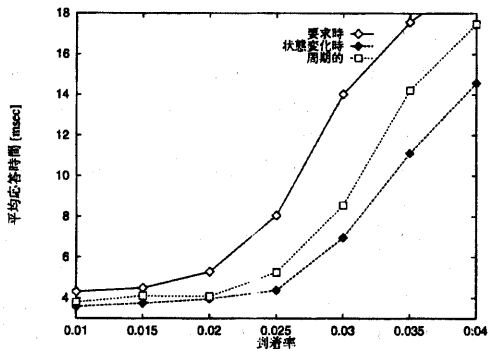

図 2: 情報収集方針

図 2にそれぞれの情報収集方針の性能を示す.「状態変化時」方針が「要求時」方針, 及び「周期的」方針よりも良い性能を示している. これは, 後者2つの方針は情報交換の通信オーバヘッドが大きいためである.「状態変化時」方針では, 2.1節で述べたしきい値として, ローカルメモリの 87.5%と 100%を高水準値, 75%と 0%を低水準値としている. これらのしきい値を 12.5%ずつ値を変化させて性能評価を行った結果, 上記の値が最も良い性能を示した.「周期的」方針では, 1msec の周期を選んだ. この値もまた, 他のいくつかの周期よりもより良い性能を示した. 大域ページ配置を適用しない (ページが常にディスクへ退避される) ときの平均応答時間は, 全到着率を通して 33msec であった. これらの値は, 大域ページ配置を適用したときよりも1桁大きい.

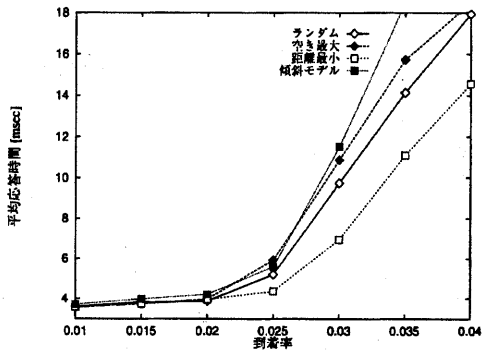

図 3: ページ配置方針

図 3にページ配置方針の性能を示す. 到着率が高い時に「距離最小」方針が良い性能を示した. 到着率が高い状態はシステム全体でネットワーク負荷が高いため, ページを転送する距離の差が処理時間に大きく影響してくる. この状態でのリモートメモリからのページの平均転送時間は,「距離最小」方針では 0.59msec(平均転送距離 1.9) であったのに対し,「空き最大」方針では 0.82msec(平均転送距離 3.0) であった. 到着率が高い時に「空き最大」方針は「ランダム」に選択する方針よりも性能が悪い. 到着率が高い状態の「空き最大」方針は, ドメイン内の複数のノードから同時に同じノードがページアウト先として選択されページアウト要求の衝突を起こしやすい. この状態の「空き最大」方針の再ページアウトの数は「ランダム」方針の 1.2 倍であった.「ランダム」方針はページアウト要求が平均的に分散されるためページアウト要求の衝突が少なく,「空き最大」方針よりは有利である.

「傾斜モデル」は, 基本的には「距離最小」方針であるが, 他の方針よりも性能が悪化した. これは, この方針が 1 ホップ間隔で直接ページを転送して空きページ枠の存在を調べているので, ページ転送のオーバヘッドが非常に大きいためである.

図 4にページ予約方針の性能を示す. 到着率の低いところでは空きページ枠が十分にあり,「予約なし」でもページアウトの失敗が少ないので, 予約

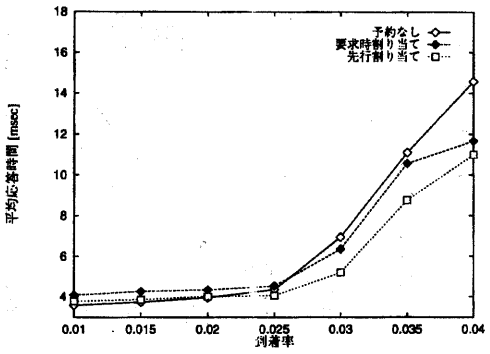

図 4: ページ予約方針

にかかる時間だけ予約方針の性能が悪くなっている．それに対して，到着率が高くなると空きページ枠が少なくなるので，「予約なし」はページアウトの失敗が増え，再ページアウトによるネットワーク混雑のために性能が悪化する．予約方針は小さいコストでページの予約ができ，予約することで再ページアウトが起こらないので性能が良い．シミュレーションでは，先行割り当てを行うページの数は 1 ノードあたり 1 ページとした．それ以上のページ数を割り当てても，性能の向上は見られなかったためである．

ドメインサイズは 3 と 4 のときが最も良い性能を示した．ドメインサイズ 5 の性能は状態情報交換のコストの増大により性能が悪い．一方，ドメインサイズが小さいと，到着率が高くなったときにドメイン内で空きページ枠を見つけるのが困難になり，性能が低下する．

5 結論

様々な大域ページ配置方針の実験で得られた結論を以下に示す:

- 情報収集方針については，「状態変化時」方針が，しきい値が適切であれば，最も良い性能を示すことがわかった．
- ページアウトする距離が短い「距離優先」方針がページ配置方針の中で良い性能を示した．
- ページ予約は再ページアウトを減らすことができ，到着率の高い状態で性能の改善が見られた．
- ドメイン分割は大規模並列システムにおいて効果的であり，通信コスト，ハードウェア構成などに依存するが，我々のシステムモデルではドメインサイズが 3〜4 が最適であった．

今後は，再ページアウトに対して異なったページ置き換え方針やページ配置方針を使用した場合のページ割り当ての振る舞いを調べたり，通信コスト，ディスク配置，ローカルメモリサイズや平均仮想記憶サイズなどのシステムパラメータを変化させた場合の性能の変化を調べたいと考えている．

参考文献

[1] P. M. Clancey and J. M. Fransioni. Distribution of Pages in a Distributed Virtual Memory. *Proceedings of the 1990 International Conference on Parallel Processing II*, 258-265, 1990.

[2] M. D. Dahlin, R. Y. Wang, T. E. Anderson and D. A. Patterson. Cooperative Caching: Using Remote Client Memory to Improve File System Performance. *Proceedings of the USENIX Conference on Operating Systems Design and Implementation*, 267-280, 1994.

[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. *Proceedings of the 15th ACM Symposium on Operating Systems Principles*, 201-212, 1995.

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

[5] M.H.Willebeek-LeMair and A.P.Reeves. Strategies for Dynamic Load Balancing on Highly Parallel Computers. *IEEE Trans. on Parallel and Distributed Systems, Vol. 4*, No. 9, 979-993, 1993.