遠隔メモリページングにおける各ページ固有のスワップ履歴を利用するページ置換アルゴリズム

斉藤 和広† 緑川 博子† 甲斐 宗徳†

1. はじめに

筆者らはローカル物理メモリサイズに制限されず，クラスタの各ノードの遠隔メモリを集めて仮想的な大容量メモリとして利用できるようにする分散大容量メモリシステム DLM(Distributed Large Memory)を構築，評価してきた[1]． DLM では，他研究で広く行われているOSスワップ処理を利用する手法ではなく，OS とは独立にユーザレベルソフトウェアによって実装することで，むしろ高速，かつ高安定な性能が得られることを示した．

このようなユーザソフトウエアにおいて，OS の仮想メモリ管理で行われているようなメモリアクセス記録を基に LRU ポリシーのページ置換アルゴリズムを用いることは高コストであり，現実的ではない．このため，従来の DLM では，遠隔メモリとのページ置換アルゴリズムには低コストのアドレス順にスワップアウトページを選択するクロックアルゴリズムを用いてきた．クロックアルゴリズムは単純であるが，対コスト性能がよく，応用においても一定レベルの性能を満たすことができた．

今回さらなる性能向上を可能にするために，アクセス記録などの高コストな処理を用いず，各ページ固有のスワップイン履歴を反映したページ置換アルゴリズムを考案し，評価したので，報告する．

2. DLM の遠隔ページスワップ機構

DLM は，大規模データ処理を行う計算ノードに対して，遠隔メモリサーバノードが適宜メモリを提供するシステムである．ユーザプログラムで利用するデータは，図 1 のように DLM で利用するノード全体のメモリにマップされる．計算プロセスはプログラム実行を行う計算スレッドとメモリサーバとやりとりする通信スレッドからなる．計算スレッドのユーザプログラムが，ローカルマップされていないデータ（ページ）へアクセスすると，ハンドラが起動し，遠隔メモリサーバへ必要ページを要求する．メモリサーバから要求ページが送られてくると，通信スレッドがページを受け取り計算スレッドに知らせ，計算スレッドは処理を再開する．通信スレッドはローカルメモリへのデータマッピング量を一定量に保つため，代わりのページを選択しメモリサーバに送る．（図 2 参照）このスワップアウトページ選択に従来，クロックアルゴリズムを用いてきた．

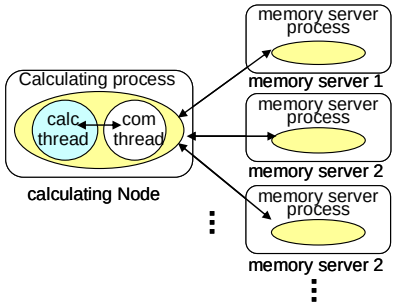

図 1 DLM のマシン構成

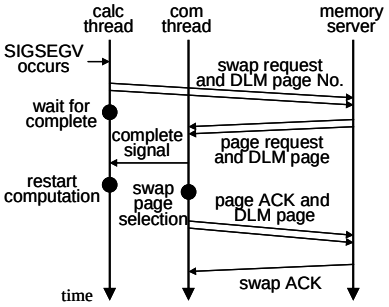

図 2 DLM の遠隔スワップの手順

3. ページ固有のスワップ履歴を利用するページ置換アルゴリズム

考案したページ置換アルゴリズムは，高コストなアクセス記録を用いず，スワップイン履歴をもとに，クロックアルゴリズムを高性能化する手法である．

†成蹊大学工学研究科情報処理専攻，
Graduate School of Engineering, Seikei University

計算ノードに新しくページをスワップインする時に，該当ページの前回スワップイン後に引き続き起きたスワップインページ番号を一定量（**履歴反映数**）覚えておき，これらのページは今回もこの後に引き続き利用する可能性が高いと仮定し，スワップアウトするページ候補から除外するというポリシーを加える．実装は図３のように，スワップイン履歴リストを環状に作成して，各ページから各履歴ポイントを指すようなデータ構造を用いる．

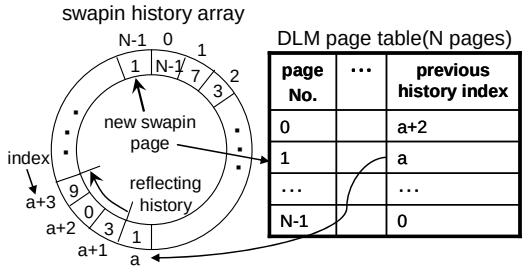

図３ スワップイン履歴と DLM ページ表(履歴反映数 3)

4. 性能評価

本報告では，幾つかの応用について行った評価のうち，Cluster3.0（遺伝子分析処理）における性能結果を示す．実験クラスタには表１の東大 T2K（ha8000）を用いた．DLM におけるページスワップ単位（DLM ページ）は 1MB に設定している．

まず従来方式で，**ローカル率**（プログラム利用データ量に対する計算ノードへマップされたデータ量の比）別で実行時間を測定した．次に新方式で**履歴反映率**（計算ノードローカルページ数に対する履歴反映数）別・**ローカル率**別に同様に測定した．

図 4 は従来方式に対する新方式の遠隔スワップ回数比で，図 5 は実行時間比である．遠隔スワップ回数，実行時間共に履歴反映率と共に減っている．ローカル率 58%のとき，遠隔スワップ回数は 50%，実行時間は 65%まで短縮されている．**履歴反映数**は大きいほど性能が良いが，ローカル率があまりに小さいと選択できるページ候補が絞られるため，かえって性能向上が小さくなる．遠隔スワップ回数の減少に比べ実行時間の減少が少ないが，この例では遠隔スワップの時間よりも計算時間の割合が比較的大きいためである．

5. おわりに

今回，低コストのスワップイン履歴を用いたページ置換アルゴリズムを考案し，遠隔メモリページングにおいて効果があることを示した．ここで述べたローカル率は，プログラムで利用する全体データ量に対する高速アクセスできるキャッシュサイズのような意味があり，このサイズに対してどのサイズまで履歴を反映するかについてなんらかの関連があると思われる．

今後は，様々な応用について，さらに詳細な分析と総合的な性能評価を行う予定である．

表 1 実験環境（T2K, ha8000）

	T2K HA8000
machine	HITACHI HA8000-tc/RS425
CPU	AMD Opteron 8356 2.3GHz QuadCore x 4CPU
memory	32GB
Cache	L2 : 2MB/CPU(512KB/Core) L3 : 2MB/CPU
OS	Linux kernel2.6.18- 53.1.19.el5 x86_64
Compiler	gcc version 4.1.2 20070626 mpicc for 1.2.7
Network	Myrinet-10G

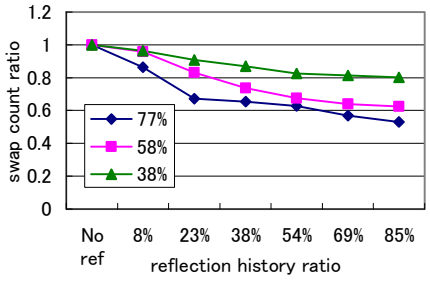

図 4 Cluster3.0 のローカル率別・スワップ回数の従来比

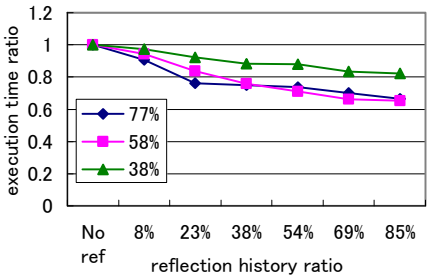

図 5 Cluster3.0 のローカル率別・実行時間の従来比

参 考 文 献

[1] 緑川，黒川，姫野，“遠隔メモリを利用する分散大容量メモリシステム DLM の設計と 10GbEthernet における初期性能評価”，情報処理学会論文誌，Vol. 49 , No.4, pp. 1-22, Apr. 2008