2H－2

超流動OSのための大域的仮想仮想記憶(GVVM)とプロセッサ割り付けアルゴリズムの適合性の評価

平野 聡　田沼 均　須崎 有康　一杉 裕志
電子技術総合研究所

要旨

超並列システム用オペレーティングシステム「超流動OS」のための大域的仮想記憶GVVMをシミュレータを用いてマルチプロセス環境で評価した。プロセスがPEアレイ上でファーストフィットあるいはベストフィットアルゴリズムにより割り付けけられ生成消滅をする環境においてGVVMの性能を調べた結果について述べる。

1　大域的仮想仮想記憶(GVVM)の概要

仮想記憶(Virtual Memory)の目的は、プログラムに実記憶よりも大きな仮想記憶空間を提供し、実記憶の大きさを超える問題を解くことである。超並列システムにおいては、多数のPE(Processing Element)に対して見合うだけのバンド幅を備える二次記憶を用意することは困難であるため、仮想記憶による性能低下は逐次マシンの場合よりも著しくなる。また、プログラムの有するメモリ使用量の偏在性と局所性が原因で特定PEでページングが頻発し、実行性能が著しく低下する。

我々は汎用超並列システムにおけるオペレーティングシステム「超流動OS」[6]の仮想記憶管理方式として大域的仮想仮想記憶(GVVM)を提案した[4, 5]。GVVMは、多数のプログラムが混在して動作する分散メモリアーキテクチャの超並列システム上での実行時間の短縮を目的として、PE空間全体でのページ使用頻度に基づくデマンドページング、及び、他PE上の使用頻度の低いメモリページをスワップ領域として用いることにより、メモリの自動的な負荷分散を行なう。図1にGVVMの実行の様子を示す。小さな四角がPEを表しており、白枠付きのPEにはプロセスが割り付けられている。色の明るさはPEのメモリの使用頻度に対応している。GVVMは自律的に動作する各PE上のGVVM処理系が協調して動作するモデルをとる。メモリ不足でページアウトを行ないたいPEはシステム中のいくつかのPEを対象として「入札」を行い、自PEよりメモリ使用頻度の低いPEが発見されたらそのPEへ仮想ページアウトを行なう。発見されなかった場合は二次記憶へのページアウトを行なう。

2　スワップ領域探索法

あるPEでメモリが不足して仮想ページアウトを行なう場合、メモリの使用頻度が低いPE上のメモリをスワップ領域として用いる。全PEを対象として最もメモリ使用頻度が低いPEを発見する事は多大なコストがかかり現実的ではない。そこで、ほどほどに低いメモリ使用頻度のPEを低いコストで探索する「スワップ領域探索」の戦略が必要となる。論文[4]で7種類のスワップ領域探索法を比較したところ、システム中で最も使用頻度の低いPEを選ぶことより、自PEよりもほどほどにメモリ使用頻度が低いPEのうちからネットワークトポロジ上でより近くにあるPEを選ぶ方が全体としての性能は高くなることが明かとなった。本論文では、メモリ使用頻度よりネットワーク上での距離を優先する次の4種類のスワップ領域探索法を比較する。

ランダム探索法 乱数によって選んだ一定数のPEに対してメモリ使用頻度の比較を行なうための入札を行なう。より低いメモリ使用頻度のPEが複数存在した場合、最短距離のPEを選択し、スワップ先PEとする。

コストフリー探索法 全てのPEのメモリ使用頻度をネットワークを用いずに比較する(探索コストをかけない)。自PEよりメモリ使用頻度が低いPE群の中からネットワーク上で最短距離のPEを選ぶ。

メタシェア探索法 超流動OSの管理情報共有機構である「メタシェア[3]」の機能を使用する。メタシェアは一定の時間間隔でシステム中の「メモリ使用頻度の低いPE一覧」を各PE上のGVVMモジュールに提供する。GVVMモジュールは自PEよりメモリの使用頻度が低いPEをネットワーク上の距離が近い順に一定個数選び、入札を行なう。

DVM探索法 仮想ページアウトはネットワークで直接隣接するPEのうち最もメモリ使用頻度の低いPEへ行なう。1ホップずつの転送を繰り返すことにより、よりページ使用頻度の低いPE領域への負荷拡散が発生することを期待する[1]。

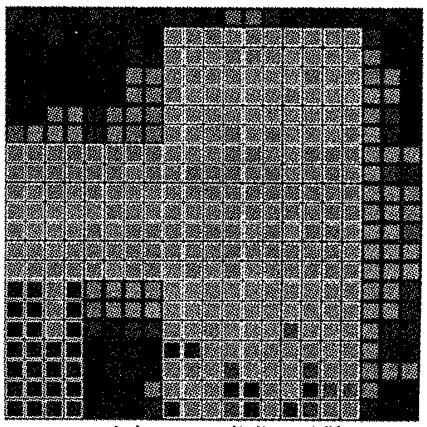

図1:GVVMによりメモリ負荷が分散される様子

GVVM under Multi Process Environment for "Fluid", An OS for Massively Parallel Systems
Hirano S.,Tanuma H.,Suzaki K.,Ichisugi Y., Electrotechnical Laboratory, Japan

3 シミュレーションによるマルチプロセス環境での性能評価

シミュレータを用い、マルチプロセス環境下でのスワップ領域探索法の性能を比較する評価を行なった。システムの形状は2次元正方形のPEアレイである。辺長をn、PE数を$N(=n\times n)$とする。ネットワークは2次元トーラスで、ブロッキングのあるバーチャルカットスルー転送を行なう。ディスクは9PE(3×3)単位で配置される。ネットワークと二次記憶装置の転送速度比は6.25:1とする。

投入するプロセスの形状は長方形とし、大きさは乱数で決定する。分布は縦、横のPE数ともに、$U(n\times 0.2, n\times 0.8)$の一様分布とする。プロセッサ割り付けアルゴリズムとしては、文献[2]によるファーストフィット・アルゴリズム(以下FF)とベストフィット・アルゴリズム(同BF)を用いた。これらは、二次元メッシュPEアレイ上で任意の長方形型をしたプロセスを到着順に割り付けるアルゴリズムである。FFはPEアレイを順に走査し、最初に見つかった配置可能な空き領域にプロセスを投入する。BFはプロセスを配置することによってできるフラグメンテーションが最小となる領域に投入する。

予め40個のプロセスを生成しておき、割り付け可能になった時点で直ちに投入する。評価の尺度として、40個のプロセスが終了するまでの実行時間(シミュレータの規定する仮想時間)を用いる。入札やページの転送等に要するネットワークの転送時間、及び、ディスクの転送時間は実行時間に含まれるが、CPUの使用時間は含まれていない。

FF及びBFを用いた評価の結果を図2、図3に示す。横軸はシステムのPE数Nである。縦軸は、GVVMを用いず各PEで独立に仮想記憶を行なった場合を1として正規化した実行時間である。結果はFF、BFともコストフリー、ランダム、メタシェア、DVMの順に良い性能を示した。

FFとBFの実行時間はほぼ同じであった(図4)。文献[2]はFFはBFと同等かやや上回るシステム使用率になると述べている。BFはFFよりもプロセスが詰めて配置されるため、プロセスの回りの空き領域が遠くなりGVVMの性能が低くなると予想していたが、影響は見られなかった。

謝辞 本研究の一部はRWC計画の一環として「超並列システムアーキテクチャに関する研究」で行なわれたものである。関係各位に感謝する。

参考文献

[1] M. Malkawi, D. Knox, and M. Abaza. Dynamic Page Distribution in Distributed Virtual Memory Systems. *Proc. of 4th ISMM Int. Conf. on Parallel and Distributed Computing and Systems*, pp. 87–91, 1991.

[2] Y. Zhu. Efficient Processor Allocation Strategies for Mesh-Connected Parallel Computers. *J. of Parallel and Distributed Computing*, pp. 328–337, 16 1992.

[3] 田沼均, 平野聡, 須崎有康. 超流動OSのための管理情報共有機構(MetaShare)の設計. 情報処理学会研究報告 *93-OS-61 (SWoPP'93)*, pp. 73–80, 1993.

[4] 平野聡, 一杉裕志, 田沼均, 須崎有康. 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. 情報処理学会研究報告 *93-OS-61 (SWoPP'93)*, pp. 65–72, 1993.

[5] 平野聡, 田沼均, 須崎有康. 超並列システム用OS「超流動OS」における大域的仮想仮想記憶. *JSPP'93*, pp. 237–244, 1993.

[6] 平野聡, 田沼均, 須崎有康, 濱崎陽一, 塚本享治. 超並列システム用オペレーティングシステム「超流動OS」の構想. 情報処理学会研究報告 *93-OS-58*, pp. 17–24, 1993.

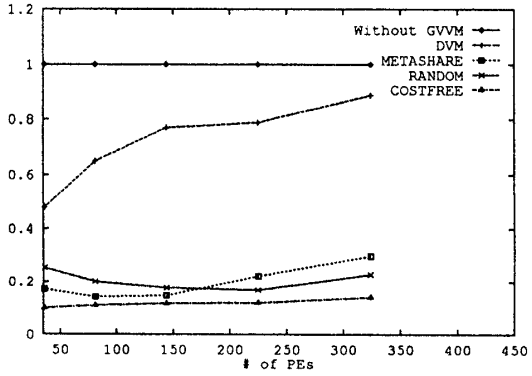

図2:GVVM不使用を1とした時の実行時間比(FF)

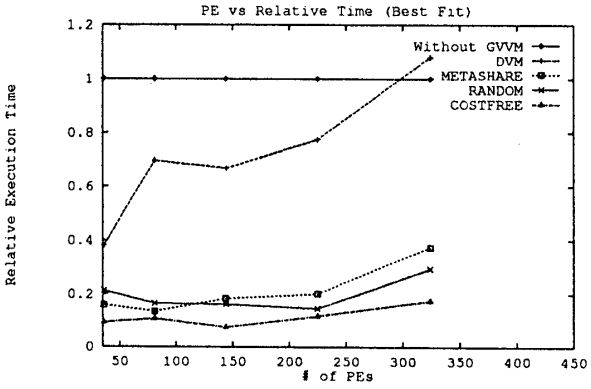

図3:GVVM不使用を1とした時の実行時間比(BF)

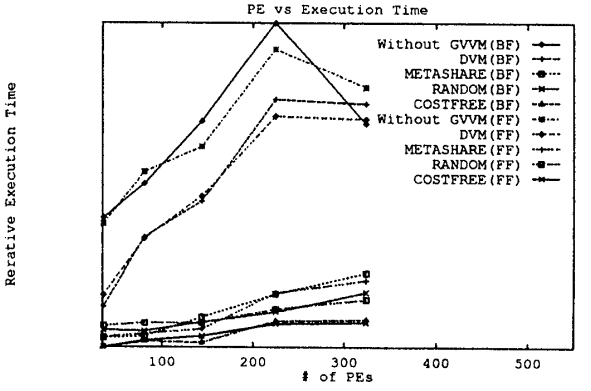

図4: 相対実行時間(FF, BF)