5H－9

# プロセス管理と協調した管理情報共有機構 (MetaShare) による メモリ負荷分散方式の検討

田沼 均　平野 聡　須崎 有康　一杉 裕志　塚本 享治

電子技術総合研究所

## 1 はじめに

多数のプロセッサ・エレメント (PE) を有する超並列システム上で様々なプログラムを混在して実行させると、各プログラムの性質が異なり、各 PE におけるメモリの使用量が異なるため、PE 空間上のメモリ負荷の不均一性が発生する。不均一性は貴重なメモリ資源の有効利用を妨げ、ある PE ではメモリがほとんど使用されていないのに別の PE ではメモリが非常に逼迫してプログラムの処理能力を著しく低下させるような問題が発生する。このような状況を解決する手段の一つとして我々は大域的仮想仮想記憶 (GVVM)[1, 2] を提案した。GVVM は超並列システム全体の PE のメモリ使用頻度を把握して仮想記憶を行うこと、及び、他 PE 上のメモリをページ退避領域として用いることによりメモリの自動負荷分散を図る。

GVVM を使用する際、メモリ負荷の大きな PE からどの PE をページ退避先とすべきか決定するページ退避先 PE 探索法が重要な問題となる。これまでに、メモリ負荷などの管理情報を効率的に収集、管理、配布する管理情報共有機構 (MetaShare)[3, 4] を利用する探索法を開発してきた。ここでは MetaShare の収集、管理、配布する情報にプロセス管理情報を加え、より効率的なメモリ負荷分散を行なう方法について検討する。

## 2 同期型 MetaShare の問題点

GVVM のページ退避先 PE は以下の条件を満たす PE である必要がある。

1. 他の PE のページを受け入れる余裕があるほど十分にメモリ使用量の少ない PE であること。
2. PE 間距離が小さいこと。

   ページアウトを行なう際、さらに将来ページアウトされた内容が必要となった時にページインを行なう際、大量のメモリ転送が必要となる。ネットワークに対し負荷をかけないためにも相互の PE 間距離が小さいことが必要である。

A Memory Load Balancing Method by An Administrative Information Management System (MetaShare) Using Process Allocation Information

Hitoshi TANUMA,Satoshi HIRANO,Kuniyasu SUZAKI,
Yuji ICHISUGI,Michiharu TSUKAMOTO
Electrotechnical Laboratory

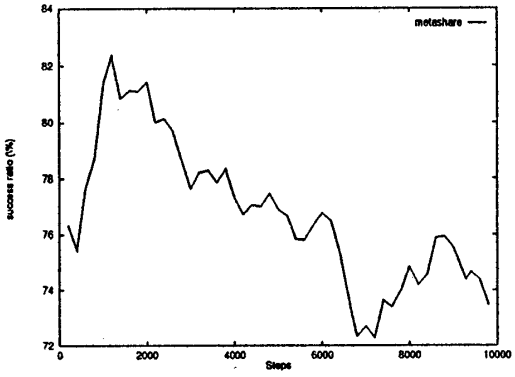


図 1: PE へのページアウト行なわれた割合の時間的変動

以上の条件を満たす PE を探索するために MetaShare はまず以下のような「負荷分散木 (LBT)」を構成する。負荷分散木の各ノードは PE の集合である「領域」を表す。木構造はこの「領域」間の包含関係を表現する。各「領域」には「領域直径」を定めておく。この「領域直径」は、その「領域」に属す二つの PE 間の最大通信距離を制限する。「領域直径」は負荷分散木のルート (最上位) においてはシステム全体を覆う大きさであり、リーフ (最下位) は各 PE 一つが一つの「領域」となるように 0 となる。負荷分散木はこの「領域直径」の制約の元に同じレベルにある「領域」の平均メモリ負荷が均衡するように木を構成する。ページアウト要求が発生した際、以上のように構成する負荷分散木をページアウト要求の発生した PE に相当するリーフよりたどりメモリ負荷の低い PE を探索する。通信距離が短くページアウトした際にシステム全体でメモリ負荷の均衡が実現するような PE を得ることが可能となる。MetaShare ではこの負荷分散木を PE 空間上に展開し、木の構成及び木のトラバースによる探索を並列処理し、高速に実行する。

以前開発した MetaShare(同期型 MetaShare)[3, 4] では、システム中の全 PE よりメモリ負荷を収集し負荷分散木を構成する操作を一定時間間隔毎に行なう。この同期型 MetaShare を利用してメモリ負荷分散を行なうシステムを並列マシンシミュレータ [5] 上に構成し評価した結果、期待したほどの性能が得られなかった。解析の結果、同期型 MetaShare ではページ退避先 PE として十分通信距離の短い PE を探す能力を有するが、探してきた PE へのページアウトの成功率が低いことが判明した。

図 1は、同期型 MetaShare を探索に利用した場合、全体のページアウトの内で PE へページアウトできた割合

の時間変化をグラフにしたものである。横軸のstepは並列マシンシミュレータの仮想時間であり、グラフでは収集時間間隔を10000にとり、情報収集の開始時点を0にとって集計した。図1より同期型MetaShareにおいては良い時で82%程度で時間経過と共に成功率は低下し、72%程度まで低下する。従ってこのページアウトの成功率を向上させればさらに高い性能を期待することができる。

## 3 プロセス管理情報の利用の検討

ページアウト成功率が低い理由としてはMetaShareの蓄えている情報の不正確さが考えられる。MetaShareは一定時間間隔毎に情報収集を行なっているが、収集からの時間経過とともに蓄積情報は古くなり、その正確さが低下する。その結果、図1の様に時間経過と共にページアウトの成功率が低下してゆく。蓄積情報の正確さの低下の原因としてプログラムの実行が進みメモリ負荷が変化したりPEへのページアウトが行なわれ負荷状況が変化しその変化が反映されていないなどが考えられるが、プロセスの生成や消滅といった大きな変化に一定間隔毎の情報収集では十分に追随できなかったことが大きな原因であったと考えられる。

そこでMetaShareの蓄積情報の精度を向上させる一つの方法としてプロセススケジューラと協調してプロセス割り当て情報を利用する方法が考えられる。プロセスの割り当てがなされ、またはプロセスが終了するとシステムの状況が大きく変化する。この二つのタイミングの情報とその際どのプロセスがどのPEに割り当てられたかまたは終了したかの情報を利用すればシステムの状況を適切にとらえることが可能となる。

1. プロセスが割り当てられた時点。

   新たなプロセスが割り当てられたPEでは、今まで0に近かったメモリ負荷が急に大きな値となる。一定時間間隔毎の情報収集ではこの変化への追随が遅れ、大きなメモリ負荷が発生しているPEをほとんどメモリ負荷の存在しないPEと誤る可能性がある。そこでプロセスの割当がなされた時点でプロセス管理モジュールから通知を受け、情報を更新する。プロセスが割り当てられた時点ではそのプロセスの正確なメモリ負荷は不明である。そこで新たなプロセスに割り当てられたPEは、メモリ負荷が予め定めておいた予測値をとるものとして、負荷分散木の再構成を行なう。具体的にはプロセス管理モジュールを新たなプロセスを割り付けるとそのプロセスの使用するPEをMetaShareに通知する。通知を受けたMetaShareは、予測値を割り付けられたPEのメモリ負荷として負荷分散木の再構成を行なう。

2. PE上でプロセスが終了した時点。

   PE上のプロセスが終了すると、PEのメモリ負荷はほとんどなくなりそのPEのメモリをページ退避領域として使用することができる。特に、ページ退避先PEは通信距離が短いPEであることが望ましいことを考慮すると、同一のプロセスに割り当てられるPEは相互に近接しているのでPEによってプロセスの終了にばらつきがある場合は、早期に終了したPEを他のPEのページの退避領域とすることは有効である。

   そこでMetaShareは、PE上のプロセスの終了をPE上のプロセス管理モジュールより通知を受けると、軽くなったメモリ負荷を収集する。収集した情報は負荷分散木に反映さる。

以上の動作を行なわせることにより、プロセス生成と消滅の際の大規模な負荷変動に迅速に対応することが可能となり、MetaShareの情報の精度の向上が期待でき、効率良いメモリ負荷分散が実現できるものと期待できる。

## 4 おわりに

ここでは管理情報共有機構(MetaShare)を利用したメモリ負荷分散の方式において、プロセス管理情報を利用する方法について検討した。プロセスの生成、消滅の時点を適切にとらえ、新たなプロセスに割り当てられるPEやプログラムの実行の終了するPEをいち早くとらえMetaShareの蓄積情報を更新することは、PEへのページアウトの成功率を向上させ、適切なメモリ負荷分散を実現し、システムのスループットの向上を図ることが期待できる。

現在、本方式の有効性を確認するために並列マシンシミュレータ[5]上でシステムを構築し、評価実験の準備を進めている。

## 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。関係各位に感謝いたします。

## 参考文献

[1] 平野, 田沼, 須崎. 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, pp237-244, 1993.

[2] 平野, 田沼, 須崎, 一杉, 塚本. 大域的仮想仮想記憶(GVVM)のマルチプロセス環境での評価.JSPP'94, pp365-362, 1994

[3] 田沼, 平野, 須崎, 一杉, 塚本. 適合型マップを用いた超並列システム用管理情報共有機構の提案. 情報処理学会研究会報告 94-OS-63,Vol.94, No.32,pp33-39, 1994.

[4] 田沼, 平野, 須崎, 浜崎, 塚本. 管理情報共有機構(MetaShare)を大域的仮想仮想記憶で利用した場合の基本性能評価. 情報処理学会研究会報告 94-OS-65(SWoPP'94), 1994.

[5] 平野, 一杉, 田沼, 須崎, 塚本. 超流動OS開発用並列マシンシミュレータ. 情報処理学会第49回全国大会,1994