2D－6

メモリ管理を考慮したNUMAマルチプロセッサにおける2レベルスケジューリングの評価

小板 隆浩†, 片山 徹郎†, 最所 圭三†, 福田 晃†
†：奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

複数のプロセスが同時に実行されるマルチプログラミング環境において，プロセスの実行効率はスケジューリング方式に大きく影響される．一方，NUMAマルチプロセッサにおいては，メモリのアクセスコストが性能に与える影響は大きく，メモリ管理方式が重要となる．しかし，これまでメモリ管理を考慮したスケジューリング方式の研究は少なく，特に2レベルスケジューリングにおいてはほとんど行なわれていない [1, 2]．

本稿では，クラスタ型NUMAマルチプロセッサを対象に，2レベルスケジューリングのメモリ管理の方式とスケジューリング方式の基本的な選択肢を提案し，シミュレーションにより評価する．

2 2レベルスケジューリング

2レベルスケジューラは上位スケジューラ(グローバルスケジューラ)と下位スケジューラ(プライベートスケジューラ)から構成される．グローバルスケジューラはプロセッサグループの管理を行ない，各プロセッサグループにフリープロセッサを割り当てる．プライベートスケジューラは，グループ内のプロセッサを用いて，プロセス内のスレッドのスケジューリングを行ない，また，アイドルプロセッサをグループから解放する．

2レベルスケジューリングは，動的な空間分割スケジューリングの1つで，大規模マルチプロセッサに有効なスケジューリング方式である．我々はこれまで，2レベルスケジューリングの基本的な選択肢についてUMA/NUMAマルチプロセッサを対象に評価を行なってきた．

3 スケジューラとメモリ管理

2レベルスケジューリングにおいて，プロセッサグループはプロセスの実行中に動的に変化する．このとき，メモリ管理を考慮せずプロセッサの割り当て，解放を行なうと，リモートメモリへのアクセス(リモートアクセス)やページフォルトなどのオーバヘッドが発生する．プロセスを効率良く実行するためには，スケジューラの方式とメモリ管理方式の間に何らかの連携が必要であり，メモリ管理の情報を利用したスケジューリング方式が必要である．

3.1 メモリ管理

メモリ管理の方式としては，仮想ページのロード，移動，複製，置換などの方式が考えられるが，本稿では仮想ページの物理メモリへのロード方式をメモリ管理方式の対象とする．メモリ管理の方式として次の方式が考えられる．

集中ロード：1つのプロセスの仮想アドレス空間内の仮想ページを1つのクラスタ(ホームクラスタ)内の物理メモリにロードする．

分散ロード：仮想ページをページフォルトを起こしたプロセッサのあるクラスタの物理メモリにロードする．

集中ロード方式では，1つのプロセスは1つのホームクラスタを持つ．ホームクラスタ内の物理メモリに仮想ページを集中ロードするため，動的にプロセッサグループを変更する2レベルスケジューリングにおいて，ホームクラスタを中心としてプロセッサを割り当てるスケジューリング方式が考えられる．すなわち，スケジューラがメモリ管理の情報を利用することができる．これに対し，分散ロードは比較的広く用いられる方式だが，あるプロセスにとって特別なクラスタは存在せず，スケジューラはメモリ管理に関する情報は利用しない．

3.2 グローバルスケジューラ

プロセッサグループへのフリープロセッサ(どのプロセッサグループにも属していないプロセッサ)の割り当て方式として，以下の2つの方式を考える．

クラスタフリー：すべてのクラスタからフリープロセッサを割り当てる．

クラスタ限定：ただ1つのクラスタからフリープロセッサを割り当てる．

ロード方式が集中ロード方式の場合，あるプロセスのホームクラスタ内のプロセッサは，他プロセスには割り当てられない．それ以外のクラスタ内のプロセッサはプロセッサの割り当て属性により割り当てが決定される．これは，分散方式の場合も同様である．

Performance Evaluation of Two-level Scheduling in collaboration with Memory Management for Multiprogrammed NUMA Mulitprocessors
Takahiro Koita†, Tetsuro Katayama†, Keizo Saisho†, and Akira Fukuda†
†: Graduate School of Information Science, NAIST.

3.3 プライベートスケジューラ

あるプロセッサで，スレッド実行終了時，またはページフォルト時に，プロセス内に割り当て可能なスレッドが存在しない場合，そのプロセッサはアイドルプロセッサとなる．アイドルプロセッサ処理方式として次の方式を考える．

保持: すべてのアイドルプロセッサをプロセッサグループに保持する．

ホームクラスタ保持: ホームクラスタ内のプロセッサのみプロセッサグループに保持する．

解放: すべてのアイドルプロセッサをプロセッサグループから解放する．

3.4 各方式のまとめ

スケジューラとロード方式の組合せを表 1 に示す．A,C の集中ロード方式が，グローバルスケジューラが，メモリ管理を考慮したスケジューリング方式となり，B,D がプロセッサの割り当て情報のみを用いた方式となる．さらに，A-3 はプライベートスケジューラにおいてもメモリ管理を考慮している．

表 1: 各方式のまとめ

方式	ロード方式	グローバルスケジューラ	プライベートスケジューラ
A-1	集中	クラスタフリー	保持
A-2	集中	クラスタフリー	解放
A-3	集中	クラスタフリー	ホームクラスタ保持
B-1	分散	クラスタフリー	保持
B-2	分散	クラスタフリー	解放
C-1	集中	クラスタ限定	保持
C-2	集中	クラスタ限定	解放
D-1	分散	クラスタ限定	保持
D-2	分散	クラスタ限定	解放

4 評価

対象とするシステムは，クラスタ型 NUMA マルチプロセッサとする．各クラスタは 4 個のプロセッサとキャッシュ，1 つの物理メモリから構成され，システムは 64 プロセッサから構成される．プロセスは，fork-join タイプとする．

システム負荷を変化させた時のプロセスの平均応答時間を図 1 に示す．負荷値はプロセスの実行時間が最悪の場合を仮定して，算出した値であり，100% を越えても本図の範囲内では定常状態である．ここでは，表 1 の組合せの中で，A,B,C,D 各方式の中で応答時間の一番短いものだけを選んでいる．

図 1 より，すべての負荷値において，A-3 が優れていることがわかる．逆にメモリ管理を考慮せずに割り当てを行なった B の場合，負荷が重くなるにつれて，応答時間が長くなっている．また，クラスタ限定 C,D では，クラスタ内の少ないプロセッサしか割り当てられないため，各選択肢や負荷の影響は少ない．

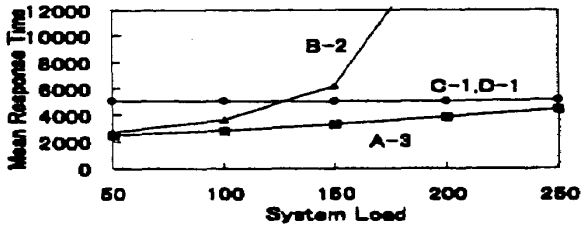

図 1: 平均応答時間

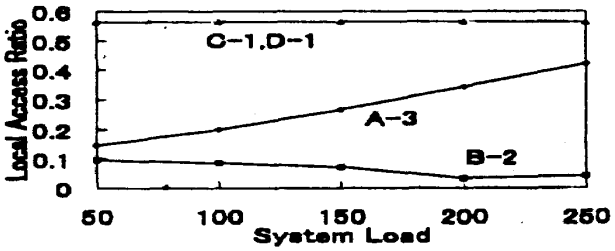

図 2: ローカルアクセス率

図 2 に，図 1 の場合のローカルメモリに対するアクセス率を示す．C,D は，高いローカルアクセス率を示している．メモリ管理を考慮しない B は負荷が重くなるにつれ，適当なプロセッサが割り当てられず，ローカルアクセス率が低下している．逆にメモリ管理を考慮した A は，負荷が重くなるにつれて，ローカルアクセス率が向上している．

5 まとめ

本稿では，2 レベルスケジューリングにおいて，メモリ管理の方式とそれを考慮したスケジューリングの方式を提案し，シミュレーションにより評価した．すべての負荷値においてメモリ管理を考慮した方式がそれ以外の方式に比べて平均応答時間の短縮がみられ，ローカルアクセス率も改善された．今後の課題は，他のスケジューリング方式との比較と，より詳細なシミュレーションによる評価である．

参考文献

[1] 大石 幸雄, 最所 圭三, 福田 晃, NUMA マルチプロセッサにおける 2 レベルスケジューリングアルゴリズムの評価, 信学会論文誌, Vol.J80-D-I, No.1, pp.31–41 (1997).

[2] 信国 陽二郎, 松本 尚, 平木 敬, 汎用超並列 OS SSS-CORE における資源管理機構, JSPP'97, pp.21–28 (1997).