

Title	マルチスレッド型プロセッサ向けのキャッシュ機構の パイプライン化に関する研究
Author(s)	相原, 孝一
Citation	
Issue Date	1997-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1001
Rights	
Description	Supervisor:日比野 靖, 情報科学研究科, 修士

マルチスレッド型プロセッサ向きの キャッシュ機構のパイプライン化に関する研究

相原孝一

北陸先端科学技術大学院大学 情報科学研究科

1997年2月14日

キーワード: マルチスレッド, パイプライン, スループット, キャッシュメモリ.

1 はじめに

プロセッサの高性能化において、パイプラインアーキテクチャの持つ性能を有効に引き出すマルチスレッド型プロセッサが提案されている [5]。プロセッサの高度なパイプライン化が進むにつれ、プロセッサのクロックサイクルおよびスループットはキャッシュ機構に依存することになる。またマルチスレッド型プロセッサは、複数の命令ストリームを扱わなければならないので、キャッシュの容量もそれに見合った容量を用意しなければならない。そうするとキャッシュの大容量化にともなって配線遅延によるキャッシュメモリのアクセス速度も低下することになる。

そのため本稿ではキャッシュでのスループットを稼ぐべくキャッシュ機構のパイプライン化と、メモリアクセス速度の向上を図るためにキャッシュのメモリアレイの分割を提案する。

次にマルチスレッド型プロセッサ上でのキャッシュのヒット率、およびスループット問題について検討、評価を行なう。

2 キャッシュ機構のパイプライン化

マルチスレッド型プロセッサでは、キャッシュメモリのアクセス時間ではなくスループットが重要である [1]。プロセッサの高いスループットを実現するためのキャッシュメモリのパイプライン化を提案する。

キャッシュアクセスは、アドレスデコード、メモリセルアレイの読みだし、判定、データ排出の手順で行なわれる。これらの各操作を分割することによって、4 ステージから構成されるパイプライン化キャッシュを実現することができる。

マルチスレッド型プロセッサではキャッシュの大容量化が必要である。この大容量化が進むにつれメモリセルアレイの配線遅延が問題となってくる。そこでメモリセルアレイを分割し、各ブロックにデコーダを配置し階層的にデコードを行なう。メモリセルアレイを分割することによって、配線遅延時間を短縮することができる。またデコードの部分においてはデコードとメモリセルのブロック選択が並列に行えるため、両者により大容量化したキャッシュへの高速なメモリアクセスが期待できる。

さらに、クロックサイクルがキャッシュの操作に依存することになる。キャッシュ中の操作でも特に遅いデコーダを、数段の処理に分割しパイプライン化することによって階層的にデコードを行うことを検討する [2]。

3 キャッシュメモリの方式とメモリ構成

マルチスレッド型プロセッサでは各スレッド間でキャッシュブロックの競合が起こるのではないかという問題が懸念される。

スレッド毎にキャッシュを用意するスレッド専用キャッシュの場合、各スレッドのマッピングは自身に用意されたキャッシュ内でのみ許されるが、キャッシュを共用する場合そのような制約がないので、スレッド専用のキャッシュではブロック数が足りない場合でも、キャッシュを共用することでその部分がカバーできる可能性がある。

しかし、各スレッド間でのブロックの競合も起こり得る。この場合キャッシュのマッピング方式をセットアソシアティブ化することによって回避することが考えられる。

1つのスレッドがキャッシュミスを起こしても他のキャッシュはキャッシュへのアクセスを要求する。このため、メインメモリのアクセスが頻繁に起こり得る。しかも、プロセッサとメインメモリの速度差は非常に大きい。メインメモリのスループットよりもその時間あたりにキャッシュから発生するメインメモリへの参照数が大きければメインメモリへのアクセスに待ち列が発生することになる。

この、待ち列の長さの短縮には、キャッシュの書き込み方式を、メモリのアクセス要求を削減する効果のあるライトバック方式を採用し、メインメモリをマルチユニット化し、ユニット毎に要求待ち列を設けることが有効である。この方策のねらいは、メモリバンク毎のメモリサイクル利用率を低下させることである。

キャッシュミスの発生からブロック更新までの時間の短縮は、プロセッサのスループット向上にそのままつながる。このため、メインメモリからデータがあがってきた時のキャッシュへの書き込みに対してどのような方式をとるかによってもプロセッサのスループットに大きな影響を及ぼす。

4 シミュレーション実験

3章で述べた方策の有効性を示すために、キャッシュのヒット率、およびスループットについてシミュレーション実験を行なった。

実験で示されているスループットは、実行されたクロックサイクルに対して、どのくらいの比率でデータが処理されたかを表すものである。よって、このスループットがシステム本来の性能を計る尺度になる。

5 実験結果と考察

キャッシュヒット率の実験結果から、キャッシュをスレッド間で共用することによって、独立でキャッシュをもつ場合の総和に対して小さな容量で済む可能性を示している。

また、キャッシュの連想度を増やすことにより、キャッシュのヒット率を増加させる効果がある。ダイレクトマップでヒット率が低いのは、スレッド間での競合が起きているためである。連想度を持たせることにより、競合するブロックを同じラインにマッピングすることができる。しかし、連想度が増加するにつれヒット率の上昇は見られなくなる。

メモリアクセスペナルティが増大するにつれてスループットが急激に落ち込んでくる。これは、待ち列の発生条件がメモリアクセスペナルティが大きくなると成立するためである。

キャッシュの書き込み方式はライトスルー方式を採用した場合、メモリアクセスペナルティが大きくなった場合のスループットの低下が著しく見られた。これはキャッシュにある程度プログラムがマッピングされるとメモリアクセスはストア命令の生起が支配的となる。ライトスルーの場合、ストア命令が起きるたびにメモリへアクセスしているためメモリ要求待ち列が発生しやすい状況になる。このような状況を防ぐには、メモリへのアクセス要求頻度を少なくすることが必要である。この点から、メモリへのアクセス要求頻度の削減を期待することができるライトバック方式は、この状況に有効な方式であると言える。

また、メモリをマルチユニット構成とし、ユニット毎に待ち列をつくることにより、待ち列長の短縮を期待することができると考えられる。この手法も、ライトバック方式同様にスループットの低下を防ぐのに役立つと考えられる。

6 むすび

本稿ではマルチスレッド型プロセッサ向けのキャッシュ機構のパイプライン化を提案した。キャッシュ機構のパイプライン化はレイテンシを増加させるが、クロックサイクルの短縮によりスループットを向上させることができる。キャッシュメモリアレイを分割して物理寸法を小さくし配線遅延を削減することによって、パイプラインピッチをつめると同時に、キャッシュの大容量化も可能になると考えられる。大容量化はヒット率の向上も期

待できる。ヒット率の向上はキャッシュミスペナルティの低下にそのままつながり、システム全体のスループットの向上につながる。

メモリの書き込み方式をライトバック方式とすることでメモリトラフィックを低く抑えることができるが、問題点はキャッシュ機構が複雑化することでありレイテンシが増加する。しかし、キャッシュ機構のパイプライン化をさらにすすめることにより、ライトバック方式を採用してもキャッシュのスループットの低下にはつながらない。

このため、プロセッサ全体のスループットを向上させる上でキャッシュ機構のパイプライン化は大きな役割を果たしているといえる。

本稿でのキャッシュメモリの構成、実験を通じて、キャッシュのヒット率のみで正しい性能の評価はできないといえることができる。キャッシュのヒット率がダイレクトマップ方式でかなり高い結果を得ていたのにもかかわらず、スループットで高い性能を得ることができなかったことから明らかである。

マルチスレッド型プロセッサの高性能化のためにはメモリ要求待ち列の長さを抑えることが重要であり、それにはメインメモリの性能が重要であるということがいえる。しかし、メインメモリの速度向上の比はプロセッサの速度向上の比には追い付いていない。そのため、メインメモリのマルチユニット化などによりメモリのバンド幅を広げることが全体のシステムの性能を向上させる鍵となる。

参考文献

- [1] 伊藤英治, 相原孝一, 丹 康雄, 日比野 靖, “関数型プログラムの実行に適したマルチスレッド型プロセッサ・アーキテクチャの提案”, 情報処理学会研究会報告, 96-ARC-121, Vol.96, No.121, pp.81-88, 1997
- [2] Toshihiko Hirose, et al., “A 20ns 4Mb CMOS SRAM with Hierarchical Word Decoding Architecture”, *ISSCC DIGEST OF TECHNICAL PAPERS*, pp.132-133, 1990.
- [3] Koichiro Ishibashi, et al., “A 300MHz 4-Mb Wave-Pipeline CMOS SRAM using a Multi-Phase PLL”, *ISSCC DIGEST OF TECHNICAL PAPERS*, pp.308-309, 1995.
- [4] David A. Patterson and John L. Henney, “COMPUTER ORGANIZATION & DESIGN THE HARDWARE/SOFTWARE INTERFACE”, Morgan Kaufmann Publishers Inc, 1994.
- [5] B.Wilkinson, “Computer Architecture : Design and Performance”, Prentice Hall, 1991.