氏名（本籍地）	佐藤（さとう） 雅之（まさゆき）
学位の種類	博　士（情報科学）
学位記番号	情　博　第524号
学位授与年月日	平成24年3月27日
学位授与の要件	学位規則第4条第1項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）情報基礎科学専攻
学位論文題目	A Hardware-Software Co-designed Cache Memory System for Energy-efficient Microprocessors（高エネルギ効率マイクロプロセッサのためのハードウェア・ソフトウェア協調型キャッシュメモリシステムに関する研究）
論文審査委員	（主査）東北大学教授　小林　広明 東北大学教授　亀山　充隆　東北大学教授　青木　孝文 東北大学准教授　滝沢　寛之

論文内容の要旨

第1章　序論

マイクロプロセッサの性能向上は，半導体製造技術の微細化とアーキテクチャ設計の発展によって成立してきた．半導体製造技術の微細化が進むことでトランジスタサイズが縮小し，オンチップ上に搭載されるトランジスタ数と動作周波数の増加，トランジスタの動的電力の削減が実現されてきた．一方で，アーキテクチャ設計では増加したトランジスタを計算資源として効率よく利用することを目標とし，マイクロプロセッサのパイプライン化，アウトオブオーダー化，投機実行，キャッシュメモリの搭載などを実現してきた．しかし，近年の半導体製造技術の微細化は電力密度の著しい増加を招き，それに伴う発熱問題も深刻になりつつある．そこで，電力当たりの性能を向上させるために，アーキテクチャ設計においてオンチップハードウェア資源としてのトランジスタの効率的な利用が課題となっている．

そこで，本研究では，マイクロプロセッサの重要な構成要素であるキャッシュメモリに着目する．現代のマイクロプロセッサの著しい性能向上の一方で，主記憶として使われている DRAM メモリのアクセス速度の向上は緩やかである．このため，マイクロプロセッサと主記憶の性能ギャップが大きく開いている．このような状況下でマイクロプロセッサが主記憶にアクセスする場合，データ到着までの長い時間を待たなければならず，性能低下が発生する．そこで，近年のマイクロプロセッサでは，最近利用したデータをキャッシュメモリに保存し，次の機会に再利用される場合にそのデータをキャッシュメモリから読み出すことによって，高速なアクセスを実現している．

しかし，キャッシュメモリは2つの大きな問題を抱えている．第1の問題は電力消費である．近年の半導体加工技術の微細化は，消費電力の中でも特に静的電力の増加をもたらしたが，静的電力は半導体チップ上の面積に比例する．近年のキャッシュメモリは，大容量化によってその面積が増加しているため，キャッシュメモリの消費電力も増加している．第2の問題は資源競合による性能低下である．近年では，複数の実行コアによる複数スレッドの同時実行によって，単位時間あたりの性能向上を目指す CMP がマイクロプロセッサの主流となっている．しかし，共有キャッシュで発生する資源

競合によってキャッシュの有効的利用が阻害され，CMP によって期待される性能が得られない場合が発生する．資源競合の要因として inter-thread kickouts (ITKOs)とキャッシュ容量不足が上げられる．あるスレッドの性能向上に必要なデータが他のスレッドの必要ないデータによってキャッシュから追い出されるのが ITKOs であり，キャッシュ容量不足は複数のスレッドの性能向上にデータキャッシュ容量より大きく，キャッシュメモリに保存しきれない状態である．よって，資源競合によって CMP における性能低下が発生する．

過去の研究ではウェイ適応型キャッシュが提案されている．ウェイ適応型キャッシュは，マイクロプロセッサが性能を発揮するために必要なキャッシュ容量に相当する領域のみを各スレッドに割り当てり．そして，それ以外の必要ない領域は電源供給を停止する．これにより，高性能かつ低消費電力なキャッシュメモリを実現可能である．領域の割り当てはスレッドごとに排他的に行われるため，ITKO を防止することが可能である．また，必要ない領域の電源供給を停止による消費電力の削減が可能である．しかし，ウェイ適応型キャッシュでは，資源競合のもう一つの要因であるキャッシュ容量不足を回避する事はできない．このため，ウェイ適応型キャッシュにおいて，キャッシュ容量をより効率的に利用し，性能向上を実現する必要がある．そこで，本論文ではハードウェア・ソフトウェア協調型キャッシュメモリシステムを提案する．本提案は第 2 章から第 4 章までで記述される 3 つのアプローチにより構成される．

第 2 章　動的キャッシュリサイズのための多数決に基づくワーキングセット評価手法

第 1 のアプローチではワーキングセット評価の妥当性について議論する．ウェイ適応型キャッシュは，サンプリング区間ごとにキャッシュアクセスを統計情報として収集し，ワーキングセット評価に用いる．しかし，この情報には全体的なキャッシュアクセスの傾向と異なる例外的なキャッシュアクセスが含まれる場合もある．このような場合，ワーキングセット評価の精度が低下し，必要と見積もられるキャッシュ容量が増大する．しかし，例外的キャッシュアクセスを原因とした割り当て領域増加による性能向上は微少である．この結果，性能向上を得られないまま消費エネルギが増加する．

そこで，例外的キャッシュアクセスのワーキングセット評価に対する影響を防ぐために，多数決に基づくワーキングセット評価手法を提案する．本手法では例外的キャッシュアクセスの影響を，細粒度のワーキングセット評価と多数決によって無視することが可能である．本手法では従来手法より短い区間（投票区間）でキャッシュアクセスをサンプリングし，この情報を用いて投票区間ごとにワーキングセット評価を行う．これらの投票区間のうち，いくつかは例外的キャッシュアクセスを含んでいるが，そのような投票区間の数は全体の数と比較して少ない．そこで，複数の投票区間のワーキングセット評価結果の多数決をとることにより，例外的キャッシュアクセスの影響でワーキングセット評価結果が他の投票区間と異なる区間の結果を除外する．この結果，例外的キャッシュアクセスの影響を受けずに最終的なワーキングセット評価結果を得ることができる．シミュレーション結果から，提案手法を適用したウェイ適応型キャッシュにおいて，最大 24%，平均 10%のエネルギ削減を実現できることが明らかになった．

第 3 章　オンキャッシュデータ管理のための高効率挿入ポリシ

第 2 のアプローチではキャッシュ中で再利用されないデータについて議論する．一般的にキャッシュメモリの容量は主記憶の容量に対して非常に小さい．そのため，キャッシュメモリはどのデータブロックを保存し，その代償としてどのブロックを追い出すか決定しなければならない．LRU 置換ポリ

シは,追い出すブロックを決定するために最もよく使われるデータ管理ポリシである.このポリシは,最近アクセスされたデータが最も再利用される可能性が高いという参照局所性に基づき,キャッシュに保存されたブロックをなるべく長く保存しようとする.しかし,実際には全てのブロックが再利用されるわけではなく,保存されてから一度も再利用されないブロックも存在する.LRU 置換ポリシにおいては,性能向上に貢献しない再利用されないブロックもキャッシュに長時間保存されるため,ウェイ適応型キャッシュのエネルギ効率を低下させる恐れがある.

このようなブロックを早期に追い出すため,本章では高容量効率挿入ポリシを提案する.従来の LRU 置換ポリシでは,ブロック保存時の優先度を一律で最大とすることによって,新しく保存されたブロックを長時間保存していた.一方で,提案手法ではブロックを保存する場合にその優先度を自由に変更することを可能とする.しかし,最適な優先度の決定にはトレードオフが存在する.もし,新しく保存されたブロックが再利用されないブロックであれば,早期追い出しのためにより低い優先度で保存する必要がある.逆に再利用されるブロックであれば,低い優先度で保存すると再利用される前に追い出されてしまう可能性がある.このトレードオフを考慮し,提案手法では最低の優先度を持つブロックに対する初回再利用数を評価する.もし,最低の優先度を持つブロックに対する初回再利用がほとんど無ければ,再利用されるブロックは充分高い優先度のうちに初回再利用されていると判断される.そして,新しく保存されるブロックにより低い優先度を与えるようにする.一方で,最低の優先度における初回再利用が多い場合は,再利用されるはずのブロックが初回再利用される前に追い出されている可能性が高いとして,新しく保存されるブロックにより高い優先度を与えるようにする.これにより,再利用されるブロックをキャッシュ上に維持しつつ,再利用されないブロックのみを早期に追い出すことが可能である.シミュレーション結果から,提案手法をウェイ適応型キャッシュに適用することにより,最大 30%,平均 6%のエネルギ削減を実現可能であることが明らかになった.

第 4 章　CMP のためのワーキングセット評価に基づくスレッドスケジューリング

第 3 のアプローチではキャッシュ容量不足の解決について議論する.多くの CMP では複数のスレッドが一つのキャッシュを共有している.スレッドのワーキングセットサイズはキャッシュアクセスの特徴に依存するため,キャッシュを共有するスレッドのワーキングセットサイズの合計は,スレッドの組み合わせによって大きく異なる.もし,ワーキングセットサイズの大きいスレッド同士がキャッシュを共有すると,キャッシュ容量不足により性能低下が発生する.一方で,もしワーキングセットサイズの小さいスレッド同士がキャッシュを共有すると,キャッシュ容量不足は発生しない.予備実験結果では,ワーキングセットサイズが大きくなるほどキャッシュ容量不足による性能低下が大きくなることが明らかになった.特に,ワーキングセットサイズがキャッシュ容量を超えると性能低下が著しくなる.

これらの事実に基づき,本章ではワーキングセット評価に基づくスレッドスケジューリングを提案する.近年では CMP は複数のキャッシュメモリを搭載しており,それぞれが複数のコアに共有されている場合が多くなった.このような場合,キャッシュを共有するスレッドの組み合わせを自由に変更するスレッドスケジューリングが可能である.そこで,本提案手法ではキャッシュ容量不足を低減するようにスレッドの組み合わせを決定する.提案手法は,ワーキングセットサイズの評価と,スケジューリングアルゴリズムを使ったコアに対するスレッドの割り当ての決定の 2 つのステージによって構成される.スケジューリングアルゴリズムは,ワーキングセットサイズの大きいスレッド同士が同じキャッシュを共有することを避け,ワーキングセットサイズが最大と最小のスレッドを組み合わ

せる．これにより，キャッシュ容量不足による資源競合を低減することができる．シミュレーション結果より，平均1.9%，最大8.1%の性能向上を実現できることが明らかになった．

第5章　結論

半導体製造技術の微細化とアーキテクチャ設計の発展は，マイクロプロセッサの性能向上と消費エネルギの削減に貢献してきた．しかし，半導体製造技術の微細化による貢献は困難になりつつあるため，アーキテクチャ設計の発展がますます求められている．

このような状況のもと，本論文ではキャッシュメモリに着目した．キャッシュメモリはマイクロプロセッサの性能向上にとって重要な構成要素であるが，その消費電力とキャッシュ資源競合による性能低下は大きな問題である．そこで本論文では，資源競合の要因であるキャッシュ容量不足の低減を目的とし，ハードウェア・ソフトウェア協調型キャッシュメモリシステムを提案した．これにより，キャッシュ資源の有効利用を促進し，消費電力当たりの性能向上を実現可能であることを示した．

論文審査結果の要旨

半導体加工技術の進歩に依存したマイクロプロセッサの高性能化と低消費電力化は，トランジスタの微細化に伴うチップの消費電力や電力密度の上昇により困難になってきている．このため，マイクロプロセッサ設計において，チップに搭載された計算資源を効率的に利用し，消費電力あたりの性能を高めるアーキテクチャ設計が重要になりつつある．本論文は，マイクロプロセッサの主要な構成要素であり，かつ性能，および消費電力に多大な影響を及ぼすキャッシュメモリに着目し，ハードウェアとこれらを制御するソフトウェアを協調させることで，エネルギ効率を高めるキャッシュメモリシステムについて論じたものであり，全編5 章からなる．

第1 章は緒論である．

第2 章では，プログラムが高い性能を維持するための必要なキャッシュ容量（ワーキングセット）を評価する際に，性能向上に貢献せず，消費電力だけを増加させる例外的なキャッシュアクセスを除去可能なワーキングセット評価手法を提案している．まず，マイクロプロセッサのベンチマークプログラムを用いて，例外的なキャッシュアクセスの頻度とこれらのアクセスが性能に与える影響を詳細に解析している．この解析に基づき，多数決に基づくワーキングセット評価手法を提案し，ウェイ適応型キャッシュ機構に適用している．評価の結果，提案手法を用いることで，高い性能を維持したまま，消費電力を大幅に削減できることを明らかにしている．本章で得られた知見は，高エネルギ効率を実現する次世代のキャッシュメモリ設計において非常に有用である．

第3 章では，キャッシュの効率的な利用を目的に，キャッシュメモリに保存される再利用されないデータブロックの削減を目的としたデータ管理ポリシを提案している．本ポリシでは，新しくキャッシュに保存されるデータブロックを保持する優先度を，アプリケーションの特徴に応じて柔軟に設定することが可能になる．これにより，再利用されないブロックを早期にキャッシュから追い出し，キャッシュの高効率な利用が可能になる．性能評価により，ウェイ適応型キャッシュと提案手法を組み合わせることで性能向上とエネルギ削減を実現することを明らかにしている．これは有益な成果である．

第4 章では，キャッシュメモリを共有するスレッドのワーキングセットの増大に伴い，容量不足によるスレッド間キャッシュ競合が増加する問題を解決することを目的に，ワーキングセットに基づくスレッドスケジューリング手法を提案している．本手法により，ワーキングセットの大きいスレッドが同じキャッシュを共有しないようにスケジューリングすることが可能となり，その結果，キャッシュ競合が削減され，性能とエネルギ効率の向上を実現している．これらは次世代のハードウェア・ソフトウェア協調型のキャッシュ管理手法の要素技術として極めて重要な成果である．

第5 章は，本論文を総括し，結論としている．

以上要するに本論文は，ハードウェアとソフトウェアの協調により，次世代マイクロプロセッサのためのキャッシュメモリシステムの高効率化を実現するための重要な知見を与えたもので，情報基礎科学および計算機科学の発展に寄与するところが少なくない．

よって，本論文は博士（情報科学）の学位論文として合格と認める．