様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ２０ 日現在

機関番号：１１３０１
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８００１２

研究課題名（和文）デバイス・アーキテクチャコデザインによるスマートユニバーサルメモリの創出

研究課題名（英文）A Universal Memory Architecture Based on Device-Architecture Co-Design

研究代表者
小林　広明（Kobayashi, Hiroaki）

東北大学・サイバーサイエンスセンター・教授

研究者番号：４０２０５４８０

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：本研究では、メモリサブシステムがアプリケーションプログラムの振る舞いに応じて知的にデータを管理し、それにより消費エネルギー最小でアプリケーションが求めるデータ供給能力を実現する新たなメモリアーキテクチャの基本技術の確立を研究の目的としている。本研究では、知的階層型メモリサブシステムを実現するために、高バンド幅のデータ供給を低消費電力で行うためのキャッシュアーキテクチャの設計に取り組み、その有効性と今後の課題を明らかにした。

研究成果の概要（英文）：The objective of this study is to establish a smart memory subsystem architecture that can consider memory access behaviors of applications and effectively manage data in the memory hierarchy in terms of performance and power efficiency. In particular, we have developed 1) a low-power/high-bandwidth cache architecture, 2) a cache management policy with an on-line evaluation of the memory request behavior of an application for reducing its working set in the memory hierarchy, 3) a cache partitioning mechanism to protect performance-sensitive shared data for chip multicore processors, 4)a memory address mapping mechanism with the performance/performance optimization by using an online-estimation of memory access behavior.

研究分野：計算機科学

キーワード：メモリシステム　キャッシュデータ管理ポリシー

１．研究開始当初の背景

近年のマイクロプロセッサは、多数の演算コアを集積し、アプリケーションに内在する並列性を活用して高いスループットコンピューティングを実現することを目的に設計・開発が行われている。これら大量のプロセッサコアが有効に機能し、その性能をフルに発揮するためには、データがメモリサブシステムからプロセッサコアによどみなく供給されることが求められる。一方、現在のメモリサブシステムの基本アーキテクチャは、メモリ容量とアクセス時間のトレードオフで階層化されたものであり、階層内ではアプリケーションのデータアクセスの振る舞いを十分に考慮できていないことから、階層内でのデータ管理に無駄が生じ、その結果、階層の深層化によるアクセス時間の増大やそれに伴う消費エネルギーの増加がプロセッサの性能を引き出す上で逆に足かせになっている。

２．研究の目的

本研究では、メモリサブシステムがアプリケーションプログラムの振る舞いに応じて知的にデータを管理し、それにより消費エネルギー最小でアプリケーションが求めるデータ供給能力を実現する新たなメモリアーキテクチャの基本技術の確立を研究の目的としている。

３．研究の方法

本研究では、知的階層型メモリサブシステムを実現するために、

1) 高バンド幅のデータ供給を低消費電力で行うためのキャッシュアーキテクチャの基本設計とその評価
2) アプリケーションの実行に必要なデータの再利用性のオンライン評価を行い、より再利用性の高いデータをキャッシュに高い優先度で配置すると共に、再利用性の低いデータはキャッシュから積極的に排除するキャッシュ管理ポリシーの検討を行い、それを実現するためのハードウェア機構と制御ソフトウェアの基本設計とその評価
3) マルチコア間でのデータ共有・データ非共有の状態をオンライン評価し、アプリケーションのマルチコア実行において性能向上に貢献するデータを積極的にキャッシュ階層に保持するキャッシュ管理機構の基本設計とその評価
4) アプリケーションが必要とするデータ供給量をオンラインで見積もり、それに応じてバンド幅とメモリでの消費エネルギーの適切なトレードオフを実現するメモリアドレスマッピング機構の基本設計とその評価

のサブテーマを設定し、3 年間でそれぞれに取り組んだ。

４．研究成果

(1) MVP キャッシュ

キャッシュメモリ中のデータアレイを複数のバンクに分割してデータアクセスを並列化することにより、高いバンド幅でのデータ供給を実現することが可能である。その一方で、データアレイを複数のバンクに分割するとタグアレイもそれぞれのデータアレイに付随させなければならず、タグ管理コストが増大し、消費電力が増加する問題がある。

本問題を解決するために MVP キャッシュを提案した。本提案では、対象とするプロセッサにおけるデータのロードがベクトル単位であるため、複数のデータアレイが格納するデータが隣接しやすい特徴に着目し、複数のデータアレイに格納されたデータを管理するために、隣接データの先頭アドレスのみを単一のタグアレイで管理する。図 1 は MVP キャッシュのブロック図である。図 1 から、本キャッシュが複数のデータアレイを単独のタグアレイで管理する構成をとっており、データアレイ毎にタグアドレスが付随する通常のキャッシュとは異なることがわかる。

図 2 に従来のキャッシュと MVP キャッシュの消費エネルギーを示す。図 2 の横軸はベンチマーク、縦軸は相対消費エネルギーを示す。図 2 から、低バンド幅な従来のキャッシュ(MBC-S)と比較して 15%、高バンド幅でタグ管理コストの大きい従来のキャッシュ(MBC-L)と比較して 3.7%の消費エネルギーを削減する事を可能とした。よって、提案手法により高バンド幅のデータ供給を低消費電力で行うためのキャッシュアーキテクチャを実現できることが明らかとなった。

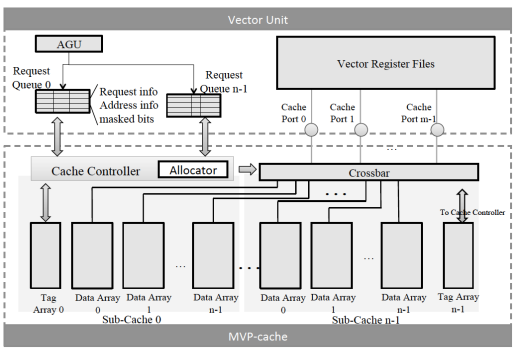

図１　MVP キャッシュのブロック図

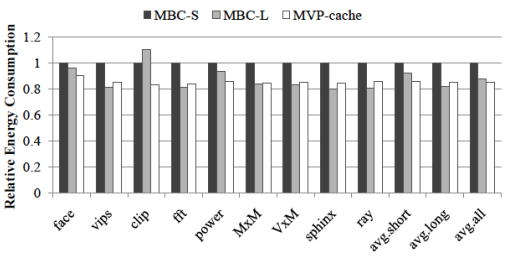

図２　MVP キャッシュの消費エネルギー

(2) FLEXII

キャッシュには再利用されないデータブロック（デッドオンフィルブロック）が保存

される場合が多い。このため、省電力化機能つきキャッシュメモリの一種であるウェイ適応型キャッシュにおいて、有効化された領域がデッドオンフィルブロックに占有され、無駄に電力を消費してしまう問題がある。

そこで、ウェイ適応型キャッシュのためのデータ管理ポリシーである FLEXII を提案した。本提案ポリシーでは、データの再利用性のオンライン評価を行いつつ、キャッシュに新しく保存されるデータブロックが再利用される可能性が低いと予測される場合に優先度を下げるデータ管理を行う。これにより、キャッシュに保存されたデータブロックのうちデッドオンフィルブロックのみを早期に追い出すことが可能である。本ポリシーの効果を他の比較手法(LRU、DIP)と比較したアクセスパターンの例を図 3 に示す。図 3 より、本提案ポリシーは比較手法と違い再利用可能な青いブロックが優先度（図 3 横軸）の高い位置に格納されていることがわかる。このことはデッドオンフィルブロックの優先度が相対的に低くなっていることを示す。ウェイ適応型キャッシュは優先度の低いブロックを格納するウェイから無効化するため、ウェイ適応型キャッシュは再利用されるブロックを維持しつつ無効化領域を増加させ、消費電力を削減することができる。

評価結果を図 4 に示す。図 4 は提案手法(FLEXII)と最新の比較手法(DAI-RRP)の消費エネルギーを表す。提案手法では、デッドオンフィルブロックを保存することによって発生していた無駄な消費電力を削減し、平均で 25%のエネルギーを削減することができた。一方で比較手法は微量な性能向上を優先して無効化ウェイを増加させてしまうため、エネルギーが大きく増加した。以上から、本提案手法は最新の比較手法と比べても有効であることが明らかとなった。

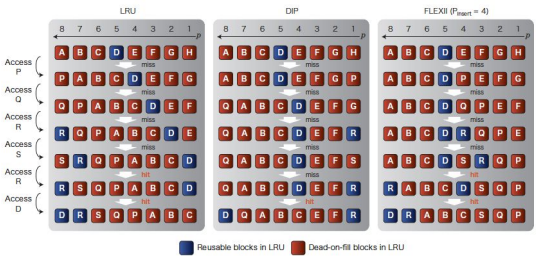

図 3　FLEXII のデッドオンフィルブロック早期追いだし効果の例

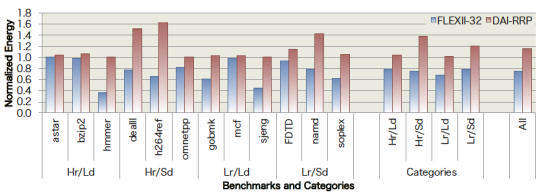

図 4　FLEXII のエネルギー削減効果

(3)　共有データ保護のためのキャッシュ分割機構

キャッシュには他のデータブロックと比較して頻繁にアクセスされやすいデータブロックがあり、そのようなデータブロックを保護することによって高効率なキャッシュメモリを実現可能である。そのようなデータとして、マルチコアプロセッサにおけるコア間共有データに着目した。コア間共有データは複数のコアからアクセスを受けることになるため、コア単独でのみしか使わない非共有データより多くのアクセスを受ける。このため、共有データをキャッシュ中で保護することにより高い実行時性能が得られると考えられる。

そこで、共有データを保護するためのキャッシュ分割機構を提案した。図 5 に本提案手法の概要を示す。本提案ではキャッシュパーティショニングを用いてキャッシュメモリを共有データのみ保存する領域(Shared space)と非共有データのみを保存する領域(Private space)に分割する。また、各領域の比率については、シャドウディレクトリ(Shadow directories)からのオンライン評価結果に基づき、パーティションを移動することによって調整可能である。これにより、様々な量の共有データをもつ並列アプリケーションで性能向上を実現することを可能とした。

本提案手法の評価結果として LRU 管理ポリシーからの速度向上を図 6 に示す。図 6 より提案手法(Proposal)の性能向上は、最大で 1.8 倍程度になっており、パーティションを最大性能となるように手動で割り当てた場合(StaticBest)と比較しても遜色ない。また、平均でも 10%程度の性能向上であり、最新の比較手法(SHP)を上回っている。以上より、本提案手法の有効性が明らかとなった。

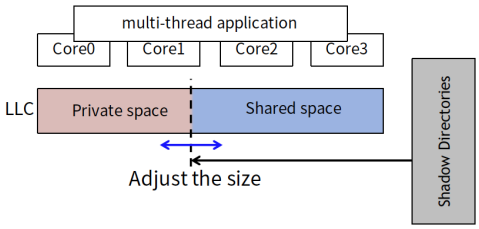

図 5　共有データを保護するキャッシュメモリ分割機構の概要

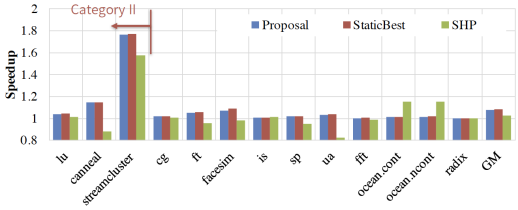

図 6　共有データを保護するキャッシュメモリ分割機構の性能評価結果

(4)　省エネルギー指向アドレスマッピング

計算機システム全体に対してメインメモリの消費エネルギーの増加が問題になりつつあり、計算機システムの構成方法によってはプロセッサよりメモリの消費エネルギー

が多くなる場合もある。一方で、プロセッサにはキャッシュが搭載されているため、メインメモリへの負荷はアプリケーションのメモリアクセス特性に基づき大きく変化する。

そこで、負荷に応じてメインメモリの消費エネルギーを削減する手法を提案した。本手法では、メモリアドレスマッピング手法によってメモリの消費電力とアプリケーション性能が変化することに着目し、消費電力削減を優先する手法と性能向上を優先する手法を切り替えてデータ供給性能と消費電力のトレードオフを実現可能である。図 7 にトレードオフをメモリアドレスマッピングで実現する方法の概要を示す。本提案手法ではメモリをランク(rank)単位で 2 つの領域に分割し、各領域でアドレスマッピングを変更することにより、上 2 ランクについては 1 ランクのみで連続アクセスを行う領域、下 2 ランクについては 2 ランクで連続アクセス行う領域とする。複数のランクで連続アクセスを行うと、性能は向上するが、消費電力も増加する。アプリケーションが必要とするデータ供給速度の見積もりは提案した指標に基づいてオンラインで見積もり、実行中にデータを分割領域間でマイグレーションすることでランク数の変更を可能とする。

提案手法の消費エネルギー評価結果を図 8 に示す。性能を優先するマッピング手法と比較して最大 16%、消費電力削減を優先するマッピング手法と比較して最大 22%の消費エネルギーを削減することが可能である。これにより、提案手法の有効性を明らかにした。

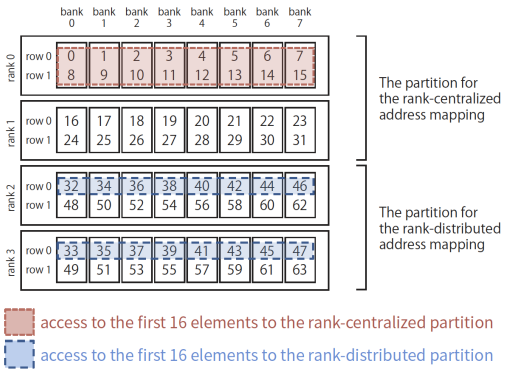

図 7　提案手法におけるアドレスマッピング

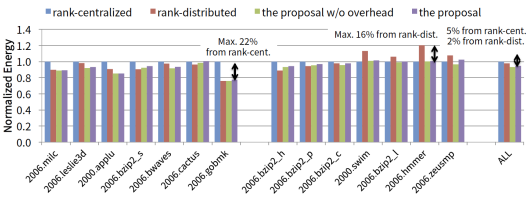

図 8　提案手法の消費エネルギー評価結果

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　6 件）

1) Masayuki Sato, Shin Nishimura, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “A Cache Partitioning Mechanism to Protect Shared Data for CMPs,” Proceedings of IEEE COOL Chips XIX, pp.1-22, April 2016(査読有).
2) 佐藤 雅之, 高井 拓実, 江川 隆輔, 滝沢 寛之, 小林 広明, “アプリケーション適応型キャッシュリサイズのためのバイパス機構”, 電子情報通信学会論文誌 , J99-D(3), pp.337-347, March 2016(査読有).
3) Masayuki Sato, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “FLEXII: A insertion policy for dynamic cache resizing mechanisms,” IEICE Transactions on Information and Systems, E98-C(7), pp.550-558, July 2015(査読有).
4) Masayuki Sato, Chengguang Han, Kazuhiko Komatsu, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “An energy-efficient dynamic memory address mapping mechanism,”pp.1-3, Proceedings of IEEE COOL Chips XVIII, 2015(査読有).
5) Ye Gao, Masayuki Sato, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “MVP-cache: A Multi-Banked Cache Memory for Energy-Efficient Vector Processing of Multimedia Applications,” IEICE Transaction on Information and Systems, Vol. E97-D, No. 11, pp.2835-2843, November 2014(査読有).
6) Ye Gao, Naoki Shoji, Ryusuke Egawa, Hiroyuki Takizawa, Hiroaki Kobayashi, “ Design and Evaluation of a Media-oriented Vector Processor with a Multi-banked Cache Memory, ” Processing of The 11th IEEE/ACM Symposium on Embedded Systems for Real-Time Multimedia, pp.78–87, 2013(査読有).

〔学会発表〕（計　4 件）

1) Masayuki Sato, Shin Nishimura, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “A Cache Partitioning Mechanism to Protect Shared Data for CMPs,” IEEE COOL Chips XIX,横浜情報文化センター, 神奈川県横浜市, 2016 年 4 月 20 日～4 月 22 日.
2) 西村秦, 佐藤雅之, 江川隆輔, 小林広明, “マルチコアプロセッサのためのスレッド間共有データを考慮したキャッシュ機構”, 並列/分散/協調処理に関するサマーワークショップ(SWoPP2015), 別府国際コンベンションセンター, 大分県別府市, 2015 年 8 月 4 日～ 8 月 6 日.
3) Masayuki Sato, Chengguang Han, Kazuhiko Komatsu, Ryusuke Egawa,

Hiroyuki Takizawa, and Hiroaki Kobayashi, “An energy-efficient dynamic memory address mapping mechanism,” COOL Chips XVIII,横浜情報文化センター，神奈川県横浜市，2015年4月13日～4月15日.
4) Ye Gao, Naoki Shoji, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “Design and Evaluation of a Media-oriented Vector Processor with a Multi-banked Cache Memory,” The 11th IEEE/ACM Symposium on Embedded Systems for Real-Time Multimedia, Montréal Marriott Château Champlain Hotel, Montreal, Canada, October 3-4, 2013.

〔図書〕（計　1件）

1) Masayuki Sato, Ryusuke Egawa, Hiroyuki Takizawa, and Hiroaki Kobayashi, “A data management policy for energy-efficient cache mechanisms.” Sustained Simulation Performance 2015, pages 61-75. Springer Berlin Heidelberg, 2015. ISBN 978-3-319-20340-9.

〔産業財産権〕

○出願状況（計　0件）

〔その他〕

特になし

６．研究組織

(1)研究代表者

小林　広明（KOBAYASHI　HIROAKI）
東北大学・サイバーサイエンスセンター・教授
研究者番号：40205480

(2)連携研究者

滝沢　寛之（TAKIZAWA　HIROYUKI）
東北大学・大学院情報科学研究科・准教授
研究者番号：70323996

江川　隆輔（EGAWA　RYUSUKE）
東北大学・サイバーサイエンスセンター・准教授
研究者番号：80374990