早稲田大学大学院情報生産システム研究科

# 博士論文審査結果報告書

## 論　文　題　目

# A Study on Hierarchical Cache System Control based on Access Pattern Analysis for Chip Multiprocessor Systems

申　請　者

Huatao　ZHAO

情報生産システム工学専攻
ASIC 自動設計研究

2020 年　10 月

LSI の設計・製造技術とコンピュータアーキテクチャの発展により、数十から数百に及ぶ複数のプロセッサ・コア(以下、コアと略す)を 1 チップ上に搭載する Chip Multiprocessor System (以下、CMP)が可能となり、高度な演算処理機能が実現されている。しかし、コアとメモリの処理速度の違いやメモリアクセス時の信号伝搬遅延がさらなる性能向上の点で問題となってきた。この問題を解決するために、高速なキャッシュメモリを階層化したオンチップ階層型キャッシュシステムが利用されている。CMP の階層型キャッシュシステムは、コアに近い側から順にレベル 1 (L1)、レベル 2 (L2)などと称するキャッシュメモリ群から構成されている。L1 は通常、各コアが独占的に使用する専用キャッシュ(Private Cache)であり、L2 以降の下位層は複数コアが共用する共有キャッシュ(Shared Cache)となっている。共有キャッシュはバンクと称する一定容量のメモリ単位に分割され、各コアが演算実行時に要求するメモリ量に応じて複数バンクが割り当てられ、不要になればバンク単位で解放される。解放されたバンクは待機状態になる。しかし、このような階層型キャッシュシステムでは、CMP の規模が増大するにつれてキャッシュの消費エネルギーの問題および階層型キャッシュにおけるデータ一貫性問題(Cache Coherency)がより深刻になっている。本論文ではこの 2 つの問題の解法を提案し、実験により効果を確認したものである。

消費エネルギーに関しては、最近の CMP の全消費エネルギーの半分をキャッシュシステムが占めることさえあり、大きな課題となっている。その原因の一つは共有キャッシュのバンク数増加に伴う消費電力増である。一方で共有キャッシュのバンク数が過少になると、キャッシュメモリ内にコアが要求するデータが存在しないキャッシュミスが頻繁に発生し、さらに下層のメモリへのアクセスが必要となるため、余分なエネルギー消費と遅延が発生する。このため共有キャッシュ内の待機状態バンクへの不要な電力を遮断するパワーゲーティングなどの低消費電力技法を導入するとともに、各コアの動作状況に応じてバンク数を調整し、消費電力を最小化する方法が知られている。そこではバンク数の増減を決定するタイミングと増減を実際に処理する方法が重要である。Rawlins [IEEE T COMPUT、2013]や Chen [Microproc.Microsyst、2016]は、各コアが一定数の命令を実行する間隔毎(たとえば $10^7$ 命令実行毎)にその区間のキャッシュミス率をチェックし、次の間隔内でのバンク数を調整する方法を提案している。また、Adegbija [IEEE T VLSI SYST、2018]や Wei [ACM T ARCHIT CODE OP、2017]はキャッシュミス率にしきい値を設定してバンク数を調整する方法を提案している。しかし、いずれも予め決めた一定の命令数の間隔で必要バンク数を計算し、次の間隔で適用するという手法なので、コアが実際に演算処理時に必要とするバンク数とは乖離が生じるという問題があった。そのため、一律な命令実行間隔ではなく、コアが処理している演算内容に適応したバンク数の調整が必要である。

次に、キャッシュにおけるデータ一貫性の問題では、あるコアが L1 キャ

ッシュ内のデータを書き換えた場合、当該データは下層の共有キャッシュや別のコアの L1 キャッシュにも格納されている可能性がある。このため、L1 から最下位層まで辿り、次に逆方向に別のコアの L1 まで辿りながら、これらの経路上で当該データを探索して順次更新することが必要である。しかし、この処理には時間がかかり、キャッシュ性能および CMP 性能を低下させ、消費エネルギーの増大にもつながる。改善手法として、Lotfikamran [IEEE HPCA、2017]による先見的キャッシュ割当法や Gupta [ICPP、2015]によるデータの空間的局所性を利用したキャッシュ分割方法などが提案されている。しかし、本論文の 2 章で行ったキャッシュアクセスパターンの分析によれば、あるコアが要求するデータが別のコアの L1 キャッシュや共有キャッシュに同時に格納されている割合は 30%以上であり、既提案手法ではそのようなデータの探索機構が複雑で、依然として時間がかかるという問題が残されていた。そのため、データの探索および更新処理の一層の効率化・高速化が可能な機構が求められている。

以上のような階層型キャッシュシステムに関わる 2 つの課題を解決するために、本研究ではまずキャッシュアクセスパターンの実際の状況を解析し、その結果に基づいて新たな解決手法を提案し、実験で有効性を確認している。以下、各章ごとに本論文の概要を述べ、評価を加えることとする。

第 1 章 “Introduction”では、CMP の構成と研究開発動向を述べ、特に階層型キャッシュシステムの構成の研究背景を紹介している。その上で、本論文で対象とするキャッシュシステムの 2 つの問題を説明し、共有キャッシュのバンクをコアに最適に割り当てる手法の提案により消費エネルギーを削減すること、および、オンチップキャッシュネットワークの提案によりデータ一貫性を効率よく維持することが本論文の目的であることを述べている。

第 2 章 “Access Pattern Analysis” では、階層型キャッシュシステムへのデータアクセスを解析するための CMP 構成のモデルを説明した上で、SPEC ベンチマークのアプリケーションプログラムを用いて実験を行い、コアに割り当てるバンク数に対するキャッシュヒット率および消費エネルギーの変化と、一定数の命令実行間隔ごとの消費エネルギーの変化を解析し、さらにプログラム中のサブルーチンの呼び出し頻度の統計調査を行っている。またコア数 2、共有キャッシュのバンク数 32 の構成で、PARSEC ベンチマークを用いて、キャッシュアクセスのタイプ毎の発生割合を調査している。これらの調査・解析データは本論文の第 3 章、4 章の提案の根拠となるだけでなく、キャッシュ構成に関する今後の研究開発の参考データとして学術的にも有用である。

第 3 章 “Controllable Cache Resource Allocation”ではコアへの共有キャッシュのバンク数割当を一定の実行命令数の間隔で行う従来手法を説明した上で、それよりも優れた手法としてアプリケーション中で実行されるサブルーチンをベースにした間隔で行う手法を提案している。サブルーチン実行時のバンク数は、消費エネルギーの計算式をもとに PID 制御技法を適用し

て調整を行い、サブルーチンの要求に見合った適切な数のバンクを割り当てる機構を導入している。SPEC ベンチマークデータと Gem5 シミュレータによる実験の結果、提案手法は一律の命令数ごとの間隔による方法と比べて、共有キャッシュの消費エネルギーが 39.7%削減され、供給電力停止のスリープモードを追加した場合の比較でも 13.6%の削減が得られている。また、従来手法と比較しても 11.6%(対 Chen2016)および 18.2%(対 Adegbija 2018)の削減が得られている。提案手法は新たな観点から階層型キャッシュシステムの消費エネルギーの低減を可能にし、有効性も示されており、高く評価できる。

第 4 章“Stacked 3D On-chip Cache Network”では、キャッシュデータの一貫性を効率的に維持するためのルータネットワークを有する 3 次元実装のオンチップアーキテクチャを提案している。階層型キャッシュシステムの共有キャッシュのレベルでルータネットワークを構成するとともに、専用キャッシュのレベルでもネットワークを構成し、これらのネットワーク経由でデータ更新を含めキャッシュ内のすべてのデータの共有を効率よく処理する機構を導入した。PARSEC ベンチマークと Gem5 シミュレータによる実験の結果、ネットワークなしの基本システムと比較して IPC 値(Instructions/Cycle)で 31.9%の高速化となり、従来手法(Lotfikamran 2017)と比較しても 18.1%の高速化が得られ、提案手法による CMP の性能改善が示されている。また、この高速化により消費エネルギーも基本システムと比較して 17.6%削減できている。本章の後半では、提案機構のハードウェア化として 3 次元積層型のキャッシュネットワークの実装設計を行い、チップサイズや熱解析の評価を行っている。提案手法の有効性が示されたとともに、その機構を組み込んだ CMP が実現可能であることを明らかにしており、学術的ならびに実用的な価値が認められる。

最後に第 5 章 “Conclusion and Future Work”では、本論文の成果をまとめるとともに今後の課題を述べている。

以上、本論文では CMP における階層型キャッシュシステムについて、キャッシュアクセスパターンの詳細な解析に基づいて、共有キャッシュの消費エネルギー低減と性能向上に寄与する制御手法を提案し、またキャッシュのデータ一貫性を保つ効率的な機構を提案したもので、実験によりその有効性を確認している。これらは当該分野の研究に大きく寄与し、学術的かつ実用的な価値がある。よって本論文は博士(工学)の価値があると認める。

2020 年　9 月 14 日

審査員

主査　早稲田大学　教授　工学博士(東北大学)　渡邊　孝博

早稲田大学　教授　工学博士(京都大学)　木村　晋二

早稲田大学　教授　博士(工学)(筑波大学)　大澤　隆