様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ４ 日現在

機関番号：１３９０４
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２６５８

研究課題名（和文）生産的かつ高効率なFPGAアクセラレータ開発のためのメモリ参照局所性の高位最適化

研究課題名（英文）High-level optimization of memory references for productive and efficient development of FPGA accelerators

研究代表者

佐藤　幸紀（Sato, Yukinori）

豊橋技術科学大学・工学（系）研究科（研究院）・准教授

研究者番号：３０４５２１１３

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：近年、消費エネルギ効率の優位性により注目を集めているFPGAを用いたアプリケーション特化型のアクセラレーションにおいては、アプリケーションに内在するソフト的特徴とアクセラレータ自体のハード的特徴の双方を踏まえてメモリ参照局所性を活用することが処理の高効率化の鍵となる。本研究では、個々のアプリケーションについて高度にカスタマイズされたアクセラレーション技術を生産的かつ効率的に適応することを目指した高位最適化技術の開発を行った。本研究を通して、メモリ参照局所性の高位最適化において多階層メモリを活用することが有効であることを示し、生産的な高位最適化のための変換方式についての技術的課題を整理した。

研究成果の学術的意義や社会的意義
FPGAによるアプリケーション特化型のカスタムアクセラレーションは、金融工学、ビッグデータ解析、人工知能分野など多岐にわたる分野でCPUやGPUによる汎用的なアプローチと比べて消費エネルギ効率の面で大きな優位性があることが報告されており、産業界の実利用や社会実装も広がっている。一方で、アプリケーション開発における開発コスト、設計の抽象化に伴う性能効率の低下が課題となっていた。本研究では、これらの課題に対して新規の高位最適化手法を提案し学術の発展に寄与したほか、アプリケーション特化型処理技術の普及に向けた基盤技術の1つとして展開していくことを目指すことを通して社会への還元を試みている。

研究成果の概要（英文）：Recently, application-specific custom acceleration using FPGA is becoming popular and attracting attention. In the development process of FPGA acceleration systems, utilizing locality of memory references based on software features within an application and hardware features inherent in the accelerator itself is a key for highly efficient processing. In this project, we developed high-level optimization technique that aims at productive and efficient development of FPGA accelerators customized for each application. Throughout this project, we showed that utilizing multiple levels of hierarchical memory can improve the gain of such high-level optimization for locality. Finally, we summed up the challenges in the process of code transformation for productive high-level optimization.

研究分野：計算機アーキテクチャ

キーワード：FPGAアクセラレータ　カスタムコンピューティング　高位合成　ハード・ソフト協調設計　Polyhedral最適化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、FPGA を用いて実装されたハードウェアアクセラレータによる処理の高効率化に大きな注目が集まっている。FPGA によるアプリケーション特化型のカスタムアクセラレーションは、CPU や GPU による汎用的なアプローチと比べて特に消費エネルギ効率の面で大きな優位性があることが様々な分野のアプリケーションで実証されて、特に、金融工学、ビッグデータ解析、機械学習などの分野では産業界の実利用も報告されている。このような利用の広がりの背景には高水準言語からハードウェアを生成する高位合成技術が著しい発展を遂げたことも大きく貢献しており、実用的なアプリを個々にハードウェア化する際の生産性は大きく向上した。

しかしながら、現状の高位合成技術により自動化されるのは RTL レベルの低水準の最適化、最内ループのパイプライン化やデータフローに基づく演算スケジューリングに限定されてきた。一方で、システムレベルで高い処理効率を達成するにはメモリサブシステムを構成するハードウェアに対する低水準のノウハウ、更に、時間空間並列性やメモリ参照局所性の利用といった高水準の最適化技術も必要不可欠であることが知られている。現状の開発の実際においては、このような高位最適化はシステム開発者が高位合成ツールの生成する回路を想定しながらソースコード上の記述を手作業で書き換えていくことにより経験的に実現されてきた。例えば、回路の並列性制御は明示的なループアンローリングによる多重化にて、階層的なパイプラインタイミング制御は FIFO の通信チャネルにて記述することが可能であった。一方で、メモリサブシステムに関してはオンチップ BRAM、オンボード DRAM、ホストマシン、ネットワークサーバに分散する階層化されたメモリ領域に対するメモリ参照局所性をアプリケーションプログラムの持つメモリアクセス特性を踏まえて制御する必要があり、記述の複雑さや適切な設計方針の導出の困難さが大きな障壁となりメモリ局所性に関する高位最適化が多くの場合なされていないのが実情であった。

メモリ参照局所性を最適化するためのアプローチとしては、ループタイリングに代表されるループ変換やその発展形である Polyhedral モデルに基づく最適化を行うことが効果的であるとスパコンや HPC の分野で実証されてきた。これらの最適化においてはループ反復レベルで計算順序のアルゴリズム的な変更を行い、メモリアクセス順序を再編することにより、メモリ参照局所性を改善させる。当時においては、UCLA の研究チームが Polyhedral モデルに基づくデータ再利用性最適化を FPGA におけるハードウェア設計に適応した事例[FPGA'13]や、Stanford 大学の研究チームがループタイリングを HW 向け並列パターンテンプレートに応用した事例[ASPLOS'16]が発表されていることから、FPGA アクセラレータ設計に本アプローチを適応する有効性が徐々に認識されつつあり、その体系化や生産性の向上が課題であった。

このように、研究開始当時においては、社会から求められている課題の解決に寄与するためには、論理回路のレイヤにとどまらない真のハードウェアとソフトウェアの融合を志すシステム指向の研究開発が必要不可欠であり、メモリデバイスやネットワークからのデータの移動を含めたシステム全体の最適化技術の開発と最適化プロセスの生産性の向上が強く求められていた。

２．研究の目的

本研究では、FPGA によるカスタムアクセラレータを主要なターゲットとして、個々のアプリケーションについて高度にカスタマイズされたアクセラレーション技術を適応するために Polyhedral モデルに基づきループ反復レベルで計算順序をアルゴリズム的に変更するメモリ参照局所性の高位最適化を実施する。その過程でメモリ参照局所性の本質をソフト・ハードの両面からモデリングするという新規のアプローチによりアクセラレーション効率を向上させることが研究の目標である。現状の Polyhedral モデルに基づくデータ再利用性最適化では考慮されていないアプリケーションに内在するソフト的特徴と FPGA アクセラレータのハード的特徴の双方を踏まえたメモリ参照局所性の高位最適化手法を提案することにより、FPGA アクセラレータを用いたシステムにおけるハード・ソフト協調設計技術の更なる進化を目指す。

３．研究の方法

第一に、メモリ局所性の高位最適化技術を FPGA アクセラレータ向けに展開する取り組みとして、Polyhedral モデルによるループ最適化を FPGA アクセラレータの高位合成フローにおけるカスタム回路設計に適応することを行った。Polyhedral 最適化を行うツールチェインとして LLVM IR のレベルでループの依存関係やスケジュール制約を表現する Polly を利用し評価システムを構築した。高位最適化による利得の検証については、Maxeler 社のデータフロー駆動型カスタムアクセラレータの開発環境を利用し、実際に個々のアプリケーションについてカスタムアアクセラレータハードウェアを合成した。本プロトタイプにおいてはアプリケーションの対象カーネル部の高位最適化を C 言語レベルで LLVM+Polly にて実施した後、FPGA アクセラレータ向けの高位合成を行うためにカーネル部分について手作業で MaxJ 言語による記述に変換するというフローを適応した。FPGA アクセラレータのハードウェアとしては、Altera(Intel) Stratix V GX

(5SGXMA5N3F45C4)と 12GB の DRAM を搭載した Maxeler 社の GALAVA ボードを x86 の汎用的なワークステーションに PCIe 経由で接続した構成を利用した。

第二に、高位最適化のアルゴリズムと自動化についての評価を行うために、LLVM+Polly の環境で開発しているメモリ局所性の高位最適化技術を FPGA アクセラレータに加えて、メニーコア型アクセラレータにも適応することを行った。利用したメニーコア型アクセラレータは、64 コアを持つ Intel Xeon Phi processor 7210 であり 1.3GHz で動作するものであった。

４．研究成果

メモリ局所性の高位最適化技術を FPGA アクセラレータ向けに展開する取り組みの評価として、ピアソンの相関係数を求めるプログラム(Correlation)を題材として用いた。Correlation 処理のカーネルを構成するループのタイリングに加えて、オンボード DRAM や FPGA チップ内部の BRAM を最大限活用する多階層メモリに対応する高位最適化技術の開発を進めた。BRAM の局所性を活用する評価については、Polyhedral モデルに基づくループタイリングを適応することにより FPGA に集積された BRAM メモリの容量に適するようにデータを分割することが可能となるため、任意の問題サイズを入力とした場合でも FPGA アクセラレータに展開可能であることを実証した。オンボード DRAM については、DRAM アクセスのバースト長や局所性を踏まえ、最適なデータアクセスパターンが出力されるようなメモリコマンド生成器の設計を行い、その有用性を評価した。FPGA 上で 192 並列の計算カーネルを実装し評価した結果、1.4GB 程度のサイズの配列に対して CPU の参照実装に対して 8 倍程度の速度向上が得られることが明らかになった。これらの成果より、メモリ参照局所性の高位最適化の基本原理の一つである多階層メモリの活用による実行速度改善の効果が FPGA アクセラレータの環境で実証された。

加えて、ループタイリングのパラメータについての考察や、最適なオンボード DRAM へのデータアクセスパターンが生成されるようなデータ格納形式の変換についての詳細な評価を行った。ループタイリングについては、空間方向の並列性を活用するベクトル化と時間方向の並列性を活用するパイプライン化の双方と融合させつつデータ参照局所性を向上させるようにタイリングを行う手法の実装を行い、有用性を確認した。データ格納形式の変換については、DRAM へのアクセスが連続となるように行列を転置させるという変換や、転送のバーストサイズに合わせた処理を毎サイクル行うようにループの形状を調整することを行い、これらの変換の有用性が確認できた。

一方で、FPGA で動くコードについてはハードウェアでのデータ処理の制御フローをたどることにより性能モデルを構築することは比較的容易であるが、ハードウェア実装がされる前に設計段階で性能モデルを作ることは容易な作業ではないことが明らかになった。すなわち、CPU で動作するリファレンスコードの実行プロファイルからアクセラレータにおける性能を推測するなどが必要であり、技術的な課題であった。そこで、CPU での多階層メモリを意識したタイリングに際する性能モデリングを起点に研究を行う方針とし、多階層メモリを持つアクセラレータの一種であるメニーコア型アクセラレータを対象に高位最適化のアルゴリズムと自動化評価を行った。実装においては、高位最適化の中でもループタイリングが性能に与える影響が最も大きいであろうという着想の下、LLVM+Polly の Polyhedral 最適化におけるタイルサイズ選択機構を独自のロードバランス見積もりと Hill-Climbing アルゴリズムによる探索の併用による数理最適化により実装した。評価実験の結果、メニーコア環境においては既存のコスト関数に基づく手法では達成できなかった性能向上が得られることが分かった。

評価を行った高位最適化方式のコード変換機構においては、現状では手作業でのソースコードの書き換えが必要となった。ハードウェアのメモリ階層構造を的確にマッピングするソフトウェアの自動での高位最適化が可能となるツールチェインの開発は将来的な課題である。この課題に対して考察を行うことを通して、本研究で対象としている CPU/FPGA が混載されるヘテロジーニアス環境を考えた際、それらの最適化においてはプログラミング環境が対象の抽象度を決定付けているという着想に至った。そこで、FPGA/GPU/CPU を搭載した SoC の開発ボードを事例にインタラクティブに入力に追従するシステムのフルスタックでの実装を行い、高位合成と RTL 実装を比べることによりプログラミングの抽象度が与える性能への影響を観測した。結果から、完全自動による高位最適化への道筋における抽象度の面での技術的課題を考察した。

これらの研究成果について、国際会議 ICAICTA2018 における Keynote 講演、Journal first publication モデルに基づく ACM の論文誌 TACO での論文出版および国際会議 HiPEAC2019 での口頭発表、IEEE COOL Chips 22 におけるポスター発表など国際的に発表を行ってきた。また、ACM SIGPLAN が主催する並列システム向けのソフトウェア工学に関する国際ワークショップ SEPS2017 のパネルセッションのポジショントークにおいても本研究の目的や成果の一部を発表した。国内においても、電子情報通信学会 北陸支部 2019 年度支部講演会での招待講演や、IPSJ プログラミングシンポジウム、IEICE CPSY+IPSJ ARC 研究会の場において発表し該当分野の研究者らと深く議論することを行った。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 佐藤健太，佐藤幸紀	４．巻 61
２．論文標題 FPGA/GPU/CPUが集積されたヘテロSoC環境におけるプログラミング	５．発行年 2020年
３．雑誌名 情報処理学会第61回プログラミング・シンポジウム予稿集	６．最初と最後の頁 pp.75-88
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Sato Yukinori、Yuki Tomoya、Endo Toshio	４．巻 15
２．論文標題 An Autotuning Framework for Scalable Execution of Tiled Code via Iterative Polyhedral Compilation	５．発行年 2019年
３．雑誌名 ACM Transactions on Architecture and Code Optimization	６．最初と最後の頁 1～23
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3293449	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計10件（うち招待講演　3件／うち国際学会　4件）

１．発表者名 Kenta Sato, Yukinori Sato
２．発表標題 Designing an FPGA Accelerator with Optimization and Specialization for On-Board DRAM
３．学会等名 Poster Session, IEEE Symposium on Low-Power and High-Speed Chips and Systems (COOL Chips 23)（国際学会）
４．発表年 2019年

１．発表者名 佐藤幸紀
２．発表標題 ソフトウェア性能最適化技術の概要とドメイン特化型カスタムコンピュータへの展開
３．学会等名 電子情報通信学会 北陸支部 2019 年度支部講演会，能美市，2020年3月6日（招待講演）
４．発表年 2020年

１．発表者名
佐藤健太，佐藤幸紀

２．発表標題
FPGA/GPU/CPUが集積されたヘテロSoC環境におけるプログラミング

３．学会等名
情報処理学会第61回プログラミング・シンポジウム

４．発表年
2020年

１．発表者名
Yukinori Sato.

２．発表標題
Computer systems and performance engineering for upcoming AI applications.

３．学会等名
5th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA 2018). Krabi, Thailand. August 14 - 17, 2018.（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Yukinori Sato, Tomoya Yuki, and,Toshio Endo.

２．発表標題
An Autotuning Framework for Scalable Execution of Tiled Code via Iterative Polyhedral Compilation.

３．学会等名
The HiPEAC 2019 conference, Spain, Valencia, January 21-23, 2019（国際学会）

４．発表年
2019年

１．発表者名
Yukinori Sato

２．発表標題
Engineering software performance of hardware accelerators using open source compilers and tools

３．学会等名
Position talk at panel discussion, The 4th ACM SIGPLAN International Workshop on Software Engineering for Parallel Systems (SEPS 2017)（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考