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

平成　３０　年　　６　月　１３　日現在

機関番号：　１０１０１
研究種目：　若手研究(B)
研究期間：　2016～2017
課題番号：　１６Ｋ１６０２６

研究課題名（和文）メモリ性能を最大限活用するFPGAアクセラレータ最適設計フレームワーク

研究課題名（英文）A Framework for FPGA-based Accelerators with Maximum Memory Performance

研究代表者

高前田　伸也（Takamaeda, Shinya）

北海道大学・情報科学研究科・准教授

研究者番号：６０７３８８９７

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：FPGAがもつオンチップメモリや再構成可能ロジックなどのリソースを最大限活用し最大性能を達成する，マルチパラダイム型高位設計フレームワークの実現に向けて研究を行った．研究代表者が以前より開発を進めている，プログラミング言語Python上のドメイン固有言語として実装したハードウェア設計ライブラリVeriloggenをベースとして，逐次処理，ストリーム処理，レジスタ転送レベルの3つの異なるパラダイムを持つ高位合成コンパイラを実現した．また，本フレームワークをバックエンドとして用いて，ディープニューラルネットワークを主な対象とした，データフロー型ハードウェア・コンパイラの開発に取り組んだ．

研究成果の概要（英文）：We developed a multi-paradigm high-level hardware design framework that easily exploits on-chip memory blocks and memory bandwidth of an FPGA. The framework is based on Veriloggen, a Python-based domain-specific language for hardware design. The newly developed framework supports 3 different programming paradigms; The compiler supports Sequential, Stream, and RTL. In addition to the framework, we developed a highly-abstracted dataflow-based hardware compiler for deep neural networks.

研究分野：　コンピュータアーキテクチャ

キーワード：　FPGA　高位合成　Python

１．研究開始当初の背景

性能と電力効率の向上を目的に、アプリケーションに特化した回路を搭載できる FPGA をアクセラレータとして用いるヘテロジニアスな計算機が注目されている。FPGA は利用者がカスタム可能なディジタル回路であるため、アプリケーションに特化した回路を構成することで、CPU や GPU と比較して高い電力効率を達成可能である。そのため、電力制約の厳しい組み込みシステムへの適用は活発に研究されてきた。近年では、マイクロソフトが機械学習アクセラレータとして FPGA をデータセンターに採用するなど、大規模システムへも広がりつつある。

FPGA の動作周波数は 100M～400MHz と低く、高い性能を達成するには、多数の演算を並列に実行する必要がある。そして、演算器に効率的にデータを供給する機構が重要となる。

FPGA は高速な内部メモリブロックを複数搭載しており、演算回路の構成に応じて、複数のメモリブロックを連結して大きな容量のメモリを実現するほかに、それぞれを独立動作させ多ポート・広帯域のメモリを実現することが可能である。一方で、大きなデータを取り扱う際には、外部メモリ帯域と演算器スループットの差が問題となる。つまり、FPGA 上の演算器の稼働率を高めて高い性能を達成するには、演算に適した内部メモリ構成を採用し限られた容量と広い帯域を活用するのと同時に、外部メモリの帯域の最大限引き出すことが重要である。

従来のハードウェア記述言語(HDL)による回路設計は抽象度が低く、開発コストが大きいという問題点がある。そこで、C 等のソフトウェア記述から専用コンパイラにより回路記述を生成する高位合成が普及しつつある。例えば、Vivado HLS は、C/C++で回路動作を記述し、ディレクティブ（指示文）により回路構成（パイプライン化等）を選択する方式を採用する Xilinx 社の商用ツールであり、HDL と比べて開発が簡単である。一方で、高い性能の達成には、内部メモリによるデータ再利用と外部メモリ帯域の活用が求められる。設計者は、高並列な演算器とそのためのメモリシステムを自ら設計し、コンパイラの挙動を把握した上で、ディレクティブによりその構成を指定しなければならない。そのため開発コストは依然として高い。また、逐次ソフトウェア記述を用いているため、データ転送と演算のオーバラップが難しく、最大性能の達成がそもそも困難である。このように、現状の高位合成環境は、ソースコードからの性能の見通しが悪く、FPGA の性能を最大限に引き出すチューニング、特にメモリシステムの最適化を行うのは容易ではない。今後も巨大化が進む FPGA の更なる活用には、性能の見通しが良く、最大性能の達成のためのチューニングが簡単な、高い開発効率を持つ高位設計方式が求められている。

２．研究の目的

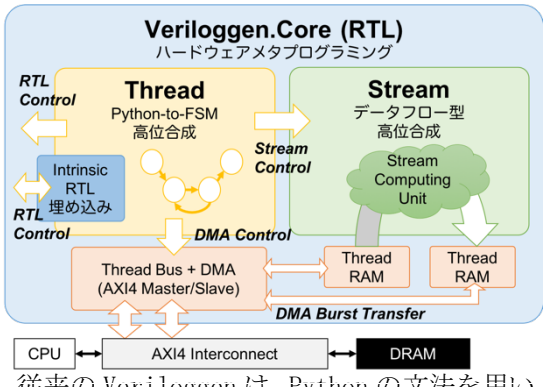
本研究の目的は、FPGA が持つメモリの総容量や帯域の制約下で、演算器および内部メモリシステムの構成を最適化し、少ない開発コストで最大性能を達成する高位設計フレームワークの実現である。本研究の本質は、これまで多く研究されてきた、高位合成により生成される論理回路そのものの品質（回路面積・周波数）の向上を目指すものではなく、メモリを主体とした設計手法・性能解析技術・最適化手法を開発することにより、高生産・高性能な FPGA コンピューティング環境を実現することである。

３．研究の方法

本研究では、記述から性能の予測が容易な高位設計方式の開発に取り組んだ。具体的には、研究代表者が以前より開発に取り組んでいる、プログラミング言語 Python を用いたハードウェア設計技術である Veriloggen (https://github.com/PyHDI/veriloggen) をベースに、単一言語上で計算とデータ移動を明示的に分割して定義可能なマルチパラダイム型高位合成ツールの開発に取り組んだ。

４．研究成果

図１に本研究で開発したマルチパラダイム型高位合成ツールの全体構成を示す。なお、高位合成ツールは既存の Veriloggen を発展させたものであるため、ツールの名前に変更はない。



従来の Veriloggen は、Python の文法を用いて明示的に RTL（レジスタ転送レベル）の回路記述を生成するメタプログラミングを行うためのツールであった。つまり、HDL（ハードウェア記述言語）のソースコードをどのように生成するかを Python で表現していると言える。本機能は図中の Veriloggen.Core に相当する。本機能により、ディジタル回路の最も細かいレベルで回路生成パターンを表現することが可能になる一方で、表現の抽象度は高くなくアプリケーションに特化したハードウェアを短時間で設計するには十分ではない。

本研究では、より高い抽象度でハードウェアの振る舞いを記述できる、2 つの高位合成モデルの開発に取り組んだ。ひとつは、Python の一般的な逐次処理記述をステートマシンに基づくハードウェアに変換する

Thread である。もうひとつは、データフロー形式で計算パターンを記述し、高いスループットのストリーム計算ハードウェアを合成する Stream である。アプリケーションを実現する上で必要となる回路的振る舞いを、それぞれ異なるパラダイムに分解して記述・実装することで、従来の高位合成ツールよりも性能のチューニングを容易にする。

Thread は、外部メモリとオンチップメモリとの間のデータ転送、およびオンチップメモリから Stream で定義されるストリーム計算ユニットへのデータ入出力パターンの制御に関する振る舞いを記述するものである。外部メモリおよびオンチップメモリ間はバースト転送を主として行うことで、高いメモリバンド幅利用効率を達成する。また、Thread では、ストリーム計算ユニットへのデータ入出力パターンのみを設定し、実際のデータの読み書きは各オンチップメモリに付随する制御回路が毎クロックサイクル連続的に行うことで、高い演算性能を達成する。

一方、Stream は計算パターンのみを記述するものである。プログラマは、DAG（有向非巡回グラフ）の形で演算間の関係を記述し、それに特化した演算器パイプラインを自動的に合成する。どのデータをいつ供給するかのデータ入出力パターンは Thread 記述により定義され、実行時に動的に変更することができる。このように、演算パターンとデータ供給パターンを分離して記述することで、回路して実装しなければならない演算器群の構成を静的に決定しつつ、複数の異なる演算で演算器群を効率的に再利用することが可能になる。

従来から存在する RTL レベルの回路を記述・生成する機能と併用することで、FPGA チップの入出力から直接データを得て、そのままチップ外へ出力するような、低遅延コンピューティングに対応することも可能である。

初期評価として、行列積や畳み込みなどの計算パターンについて性能を測定したところ、オンチップバスが提供可能なメモリ帯域を活用し、同時にストリーム計算ユニットが高い効率で稼働して演算を行っていることを確認した。

開発したマルチパラダイム型高位合成ツールは、プログラマが直接ソースコードを記述しハードウェアを開発するための環境としてだけではなく、ツールが提供する各種関数を API として用いることで、さらに抽象度の高い高位設計ツールのバックエンドとして用いることが可能である。そこで、本研究では、上記のマルチパラダイム型高位合成ツールをバックエンドとして用いて、深層学習計算のためのハードウェアを、高い抽象度で効率的に設計するための新たな高位設計ツールの開発に取り組み、現在も開発中である。

また、深層学習 LSI のプロトタイプ環境として FPGA を用いる際に，メモリシステムの抽象化および IP コアとしてシステムへのインテグレーションを行うための設計環境として、本研究で開発した高位設計フレームワークを用いることで、効率的に開発を行うことができた。

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 2 件）

(1) Hoang Gia Vu, Shinya Takamaeda-Yamazaki, Takashi Nakada, and Yasuhiko Nakashima: A Tree-based Checkpointing Architecture for the Dependability of FPGA Computing, IEICE Transactions on Information and Systems, Vol.E101-D, No.2, pp.288-302, February 2018. DOI: 10.1587/transinf.2017RCP0010 （査読有り）

(2) 渡邊 実, 佐野 健太郎, 高前田 伸也, 三好 健文, 中條 拓伯:FPGA ハードウエア・アクセラレーション向け日の丸高位合成ツール, 電子情報通信学会論文誌 B 招待論文, Vol.J100-B, No.1, pp.1-10, January 2017. DOI:10.14923/transcomj.2016JBI0002 （査読有り）

〔学会発表〕（計 14 件）

(1) Shinya Takamaeda-Yamazaki: Invited Talk (Keynote), Accelerating Deep Learning by Hardware/Algorithm Co-Design, International Workshop on Advances in Networking and Computing (WANC 2017), Aomori, Japan, November 2017.

(2) 高前田 伸也: 招待講演, アルゴリズムとハードウェアの協調設計によるディープラーニングアクセラレーション, Design Solution Forum 2017, 於 新横浜国際ホテル, 2017 年 10 月 13 日, October 2017.

(3) 高前田 伸也: 招待講演, アルゴリズムとハードウェアの協調設計による新時代コンピューティング, 電子情報通信学会集積回路研究会(IEICE-ICD)・シリコン材料・デバイス研究会(IEICE-SDM)・映像情報メディア学会メディア工学研究会(ITE-IST), 於 北海道大学情報教育館, 2017 年 7 月 31 日, July 2017.

(4) Shinya Takamaeda-Yamazaki: Invited Talk (Mini Keynote), Energy-Efficient In-Memory Neural Network Processor, 17th International Forum on MPSoC for Software-defined Hardware (MPSoC 2017), Les Tresoms Hotel, Annecy, France, July, 2017.

(5) Kasho Yamamoto, Huang Weiqiang, Shinya Takamaeda-Yamazaki, Masayuki Ikebe, Tetsuya Asai, and Masato Motomura: A Time-Division Multiplexing Ising Machine on FPGAs, International Symposium on Highly-Efficient Accelerators and

Reconfigurable Technologies (HEART 2017), Ruhr University, Bochum, Germany, June 2017.
(6) 熊澤 輝顕, 高前田 伸也, 池辺 将之, 浅井 哲也, 本村 真人: メモリアクセスパターンを考慮した遅延評価による ZDD 構築の高速化, 第 30 回回路とシステムワークショップ, 於 北九州国際会議場, 2017 年 5 月 11 日発表, May 2017.
(7) Keisuke Fujimoto, Takashi Nakada, Shinya Takamaeda-Yamazaki, and Yasuhiko Nakashima: A Multi-Level Power-Capping Mechanism for FPGAs (Outstanding M2 Student Award (OM2)), The 1st. cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming (xSIG 2017), April 2017.
(8) Hoang Gia Vu, Shinya Takamaeda-Yamazaki, Takashi Nakada, and Yasuhiko Nakashima: CPRring: A Structure-aware Ring-based Checkpointing Architecture for FPGA Computing, The 25th IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM2017) (Poster), April 2017.
(9) 高前田 伸也: 招待講演, ゆるふわコンピュータ, 情報処理学会第 79 回全国大会 IPSJ-ONE, 於 名古屋大学, 2017年3月18日, March 2017.
(10) 高前田 伸也: 招待講演, Python によるカスタム可能な高位設計技術, Design Solution Forum 2016, 於 新横浜国際ホテル, 2016 年 10 月 14 日, October 2016.
(11) 高前田 伸也: 招待講演, ハードウェアはやわらかい, 第 15 回情報科学技術フォーラム (FIT 2016) 助教が吼える！各界の若手研究者大集合, 於 富山大学, 2016 年 9 月 9 日, September 2016.
(12) Shinya Takamaeda-Yamazaki: Invited Talk (Mini Keynote), Customizable Hardware Abstraction, 16th International Forum on MPSoC for Software-defined Hardware (MPSoC 2016), Nara, Japan, July 2016.
(13) Hoang Gia Vu, Supasit Kajkamhaeng, Shinya Takamaeda-Yamazaki, and Yasuhiko Nakashima: CPRtree: A Tree-based Checkpointing Architecture for Heterogeneous FPGA Computing, 4th International Symposium on Computing and Networking (CANDAR 2016), November 2016.
(14) Keisuke Fujimoto, Shinya Takamaeda-Yamazaki, and Yasuhiko Nakashima: Stop the World: A Lightweight Runtime Power-Capping Mechanism for FPGAs, 4th International Workshop on Computer Systems and Architectures (CSA 2016), November 2016.

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況 (計 2 件)
(1) 名称：ニューラル電子回路
発明者：高前田伸也、植吉晃大、本村真人
権利者：同上
種類：特許
番号：特許願 2018-19251
出願年月日：2018 年 2 月 6 日
国内外の別：国内
(2) 名称：ニューラル電子回路
発明者：高前田伸也、植吉晃大、本村真人
権利者：同上
種類：特許
番号：特許願 2018-19252
出願年月日：2018 年 2 月 6 日
国内外の別：国内

○取得状況 (計 0 件)

〔その他〕
ホームページ等
https://sites.google.com/site/shinyaty/
https://github.com/PyHDI/veriloggen

６．研究組織
(1)研究代表者
高前田 伸也 (TAKAMAEDA, Shinya)
北海道大学・大学院情報科学研究科・准教授
研究者番号：60738897

(2)研究分担者

(3)連携研究者

(4)研究協力者