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　５ 年　 ６ 月　 ５ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2019～2022
課題番号：１９Ｈ０４０７５

研究課題名（和文）性能最適化が容易なマルチパラダイム型高位合成フレームワークの創出

研究課題名（英文）Multi-Paradigm High-Level Synthesis Framework with Productive Performance Optimization Capability

研究代表者

高前田　伸也（Takamaeda, Shinya）

東京大学・大学院情報理工学系研究科・准教授

研究者番号：６０７３８８９７

交付決定額（研究期間全体）：（直接経費）　　12,600,000円

研究成果の概要（和文）：ドメイン特化計算の高効率化を目的に、性能最適化が容易なハードウェア高位設計フレームワークに関する研究を進めた。研究代表者が開発を進めている、演算データフローと制御を分離して記述するマルチパラダイム型ハードウェア設計フレームワークのVeriloggenを発展させ、メモリ容量と帯域の制約下での性能最適化を容易に行うためのハードウェア・プログラミングモデルと、対応する効率的な演算回路・メモリシステム合成技術を開発した。また、拡張されたVeriloggenを用いて、ニューラルネットワーク特化ハードウェアコンパイラNNgenの機能拡張を行い、実アプリケーションにおいて開発した技術の有効性を示した。

研究成果の学術的意義や社会的意義
アプリケーションやドメインに特化したハードウェア構成を用いることで高い計算性能と電力効率を達成する、ドメイン特化アーキテクチャ（Domain Specific Architecture）が、機械学習分野を筆頭に注目されており、多くの実用例が報告されている。高い計算効率を達成するドメイン特化ハードウェアを簡単に実現するためのハードウェア設計技術が求められている。本研究で開発を進めたVeriloggenは、オープンソースで提供されるハードウェア設計ソフトウェアであり、半導体開発の民主化技術として、利活用されている。

研究成果の概要（英文）：For the efficiency of domain-specific computations, we have conducted research on a high-level hardware design framework with high performance optimization capability. Based on the Veriloggen, a multi-paradigm hardware design framework developed by the Principal Investigator, which describes arithmetic dataflow and control-flow separately, we developed a novel hardware programming model for easy performance optimization under memory capacity and bandwidth constraints, and developed corresponding efficient arithmetic circuit and memory system synthesis techniques. Using the extended Veriloggen, we also extended the functionality of NNgen, a neural network specific hardware compiler, and demonstrated the effectiveness of the developed techniques in real applications.

研究分野：コンピュータアーキテクチャ

キーワード：ハードウェア設計技術　FPGA　Python

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

幅広い用途で高い電力性能を達成する計算機の実現方法として FPGA が活用されている。高い性能を達成するには FPGA 上に実現する多数の演算器をできるだけ絶え間なく稼働する必要がある。FPGA は高速なメモリブロックをチップ内に複数搭載しており、演算回路の構成に応じて、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、一般的なアプリケーションではオンチップメモリ容量を超える大きなデータを取り扱うことが多く、外部メモリ帯域はオンチップメモリ帯域と比較して狭い。そのため、演算回路構成だけではなく、メモリアクセスパターンと再利用性に応じて最適化したメモリシステムが重要となる。

従来のハードウェア記述言語（HDL）よりも効率的な FPGA 回路の開発方法として、C や C++ などのアプリケーションの動作記述から回路構成を生成する高位合成ツールが普及しつつある。しかし、アプリケーションが持つ並列性を十分に活用し高い性能を達成するには、高並列な演算回路とそれにデータを連続的に供給するメモリシステムの構造を設計し、高位合成ツールの挙動を理解した上でディレクティブ（指示子）により、その構成を指定しなければならない。そのため、高位合成ツールを用いたとしても、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。FPGA を今以上に幅広いアプリケーションドメインで活用するには、ソースコードから性能の見通しが良く、性能最適化が容易な、高い開発効率を持つ高位設計ツールが必要である。

２．研究の目的

本研究の目的は、アプリケーションやアルゴリズムに内在する並列性を、FPGA 上の並列演算回路とメモリシステムに対応付けることに適した高位設計方式を明らかにし、それに対応する、性能最適化が容易なハードウェア高位設計方式および高位合成フレームワークを創出することである。そして、具体的なアプリケーションの実装を通じて、その有用性を示すことである。

３．研究の方法

研究代表者が開発を進めている、演算データフローと制御を分離して記述するマルチパラダイム型ハードウェア設計フレームワークの Veriloggen (https://github.com/PyHDI/veriloggen) を発展させ、メモリ容量と帯域の制約下での性能最適化を容易に行うためのハードウェア・プログラミングモデルと、対応する効率的な演算回路・メモリシステム合成技術を開発する。また、拡張された Veriloggen を用いて、ニューラルネットワーク特化ハードウェアコンパイラ NNgen の機能拡張を行い、実アプリケーションにおいて開発した技術の有効性を示す。

４．研究成果

これらの以前よりオープンソースソフトウェアとして開発を進めているハードウェア設計コンパイラの Veriloggen の拡張する形で本研究を進め、いくつかの技術は既に本ソフトウェアの機能として取り込まれている。以下に本研究で実装された技術を示す。

2019 年度は、従来の Veriloggen では表現できない、もしくは、高性能な計算回路が生成することができない演算パターンをサポートするために、Veriloggen のストリーム計算型プログラミングモデルと、それを支えるメモリシステムの拡張を行った。具体的には、間接参照等の不規則なメモリアクセスパターンを持つ場合においてもストリーム計算により高速処理するための、ストリーム型プログラミングモデルにおけるオペレータの追加と、対応するメモリアクセス回路の開発を行った。従来の Veriloggen のストリーム計算モデルでは、ストリーム計算の中間結果を用いたメモリアクセスができなかったが、本技術の開発によりデータの流れを止めることなく、ポインタチェイシングといったランダムアクセスが可能になった。

2020 年度は、近年の多くの SoC FPGA が搭載するハードマクロの DMA コントローラを活用するために、AXI-Stream インタフェースへの対応を行った。従来の Veriloggen では、オフチップとのデータ転送は AXI-Master インターフェースを用いる方法のみが提供されていたが、本拡張により、データ転送機能を内包しない計算回路の実装が可能になり、ハードマクロの DMA コントローラを組み合わせて利用する場合の回路資源が効率化された。あわせて、ストリーム計算回路と AXI-Stream インターフェースを直接接続するために、ストリーム計算パイプラインの適応型ストール機構を実装した。また、従来の Veriloggen のストリーム計算モデルでは、オンチップメモリ（ブロック RAM）に対する入出力のみを対象としていたが、同メモリを FIFO として利用するモードにも対応させることで、必ずしもすべてのクロックサイクルにデータが入出力されないインターフェースや、データ入出力間隔が異なるストリーム計算パイプラインを結合することが可能になり、複数のマイクロタスクで構成されるアルゴリズムの表現が可能になった。さらに、本質的にパイプライン処理に適さない演算に対応するために、ストリーム計算の中にマルチサイクルの演算を埋め込むためのプログラミングモデルの拡張を行った。加えて、複雑なメモリアクセスパターンを表現するためのプログラミングモデルの拡張を行った。

2021 年度は、オフチップメモリアクセス時の性能を高めるための、DMA コントローラの拡張を行った。専用ロジックによる計算高速化では、オンチップメモリとオフチップメモリ間のデータ転送を計算とオーバーラップさせることが重要である。従来の DMA コントローラが複数のショートバーストのリクエストを連続して発行する際に、メモリ帯域を十分に活用できておらず、データ転送の遅延が大きいことを解消するために、DMA コントローラの多重リクエスト管理機構を導入し、複数の in-flight なリクエストをメモリコンローラに対して発行することを可能にした。Veriloggen コンパイラに当該機構を実装し、実装の正当性検証を複数のテストケースにより確認した。そして、DNN ハードウェア生成フレームワーク NNgen に本拡張を適用し、動作確認を行った。加えて、HBM などの高帯域なオフチップメモリの性能を引き出すための、カスタムコンピューティング向けのメモリシステムのアーキテクチャとプログラミングモデルのプロトタイプを開発した。データを読み出し消費するロード操作を、読み出し、消費、待ち合わせの 3 つのマイクロ命令に分離し、計算などの他の操作とのオーバーラップを可能にした。また、対応するメモリシステムのアーキテクチャを開発した。ソフトウェアシミュレーションによるハードウェア構成の検討と、グラフ処理を対象としたアプリケーションの性能評価を行い、キャッシュメモリに基づく方式よりも高い性能を達成することを確認した。

最終年度の 2022 年度は、FPGA による計算高速化では、データパスに効率的にデータを供給するメモリシステムが重要となる。一般的な高位合成ツールにおいては、オンチップメモリとオフチップメモリ間のデータ転送と計算本体をオーバーラップさせ、データ転送の遅延を隠蔽することは容易ではなく、遅延隠蔽をするハードウェア記述はプログラマの負担を大きく増やすことになる。本年度は、データ転送と計算のオーバーラップを意識しない容易なハードウェア記述から、データ転送と計算をオーバーラップさせて計算を行う効率的なハードウェアを自動的に合成する高位合成技術を開発した。計算とデータ転送を逐次的に行う記述を入力として、ソースコード静的解析によりデータパス稼働時にアクセスするオンチップメモリ領域を特定する。そして、未使用のオンチップメモリ領域に先行的にオフチップメモリからデータを転送し、データ転送完了後にデータパスからアクセスするアドレスを張り替えることでデータ転送遅延を隠蔽する。また、間接参照を含む場合でも計算とデータ転送をオーバーラップさせるための、計算とデータ転送の軽量な同期方式を開発した。密行列積や疎行列積等において性能向上を確認した。

また、Veriloggen の実アプリケーションへの応用として、単眼動画像を入力とする奥行き推定の FPGA ベースアクセラレータ FADEC (https://github.com/casys-utokyo/fadec) を開発した。図 1 に、FADEC で実装した奥行き推定アルゴリズム DeepVideoMVS の計算フローを示す。アプリケーション全体をハードウェア化することが容易ではないアルゴリズムであるため、ハードウェア化による恩恵と開発コストの見積に基づくハードウェア・ソフトウェア協調設計を行い、ハードウェアアクセラレータと CPU 上のソフトウェアが連携し並行処理するアーキテクチャを採用した。開発には Veriloggen をバックエンドに持つ DNN ハードウェア高位合成コンパイラ NNgen (https://github.com/NNgen/nngen) を用いることで、本研究で開発したハードウェア設計の要素技術を活かして、高性能なハードウェアシステムを短期間で開発できることを示した。

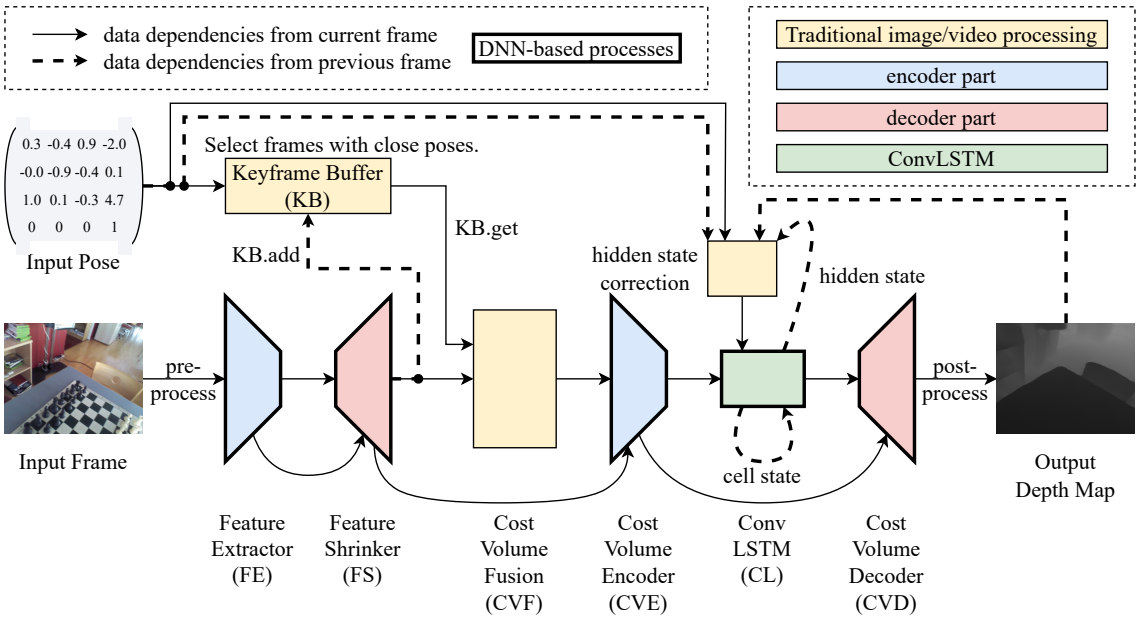

図 1 FADEC における DeepVideoMVS の計算フロー

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　9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 Nobuho Hashimoto, Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 An FPGA-Based Fully Pipelined Bilateral Grid for Real-Time Image Denoising	５．発行年 2021年
３．雑誌名 2021 31st International Conference on Field-Programmable Logic and Applications (FPL 2021)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/FPL53798.2021.00035	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 Yoshiki Fujiwara, Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 ASBNN: Acceleration of Bayesian Convolutional Neural Networks by Algorithm-hardware Co-design	５．発行年 2021年
３．雑誌名 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP 2021)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ASAP52443.2021.00041	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 Keisuke Kamahori and Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 Accelerating Decision Tree Ensemble with Guided Branch Approximation	５．発行年 2022年
３．雑誌名 International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies 2022 (HEART 2022)	６．最初と最後の頁 24--32
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3535044.3535048	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 Yuki Hirayama, Tetsuya Asai, Masato Motomura, and Shinya Takamaeda-Yamazaki	４．巻 Vol.10, No.2
２．論文標題 A Hardware-efficient Weight Sampling Circuit for Bayesian Neural Networks	５．発行年 2020年
３．雑誌名 International Journal of Networking and Computing	６．最初と最後の頁 84--93
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 本村 真人，高前田 伸也，植吉 晃大，安藤 洸太，廣瀬 一俊	４．巻 Vol.J103-C, No.5
２．論文標題 深層ニューラルネットワーク向けプロセッサ技術の実例と展望	５．発行年 2020年
３．雑誌名 電子情報通信学会論文誌C	６．最初と最後の頁 288-297
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 Taiga Ikeda, Kento Sakurada, Atsuyoshi Nakamura, Masato Motomura, and Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 Hardware/Algorithm Co-optimization for Fully-Parallelized Compact Decision Tree Ensembles on FPGAs	５．発行年 2020年
３．雑誌名 16th International Symposium on Applied Reconfigurable Computing (ARC 2020)	６．最初と最後の頁 345～357
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-44534-8_26	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 Peiqi Zhang and Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 Multi-Input Adaptive Activation Function for Binary Neural Networks	５．発行年 2022年
３．雑誌名 10th International Workshop on Computer Systems and Architectures (CSA 2022)	６．最初と最後の頁 90--96
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CANDARW57323.2022.00062	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 Nobuho Hashimoto and Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 FADEC: FPGA-based Acceleration of Video Depth Estimation by HW/SW Co-design	５．発行年 2022年
３．雑誌名 International Conference on Field Programmable Technology (FPT 2022)	６．最初と最後の頁 1--9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICFPT56656.2022.9974565	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 Tatsuya Kubo and Shinya Takamaeda-Yamazaki	４．巻 -
２．論文標題 Cachet: A High-Performance Joint-Subtree Integrity Verification for Secure Non-Volatile Memory	５．発行年 2023年
３．雑誌名 IEEE Symposium on Low-Power and High-Speed Chips and Systems (COOL Chips 26)	６．最初と最後の頁 1--6
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/COOLCHIPS57690.2023.10122117	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 －

〔学会発表〕　計28件（うち招待講演　9件／うち国際学会　3件）

１．発表者名 橋本 信歩，高前田 伸也
２．発表標題 FPGAを用いたフルパイプラインによるバイラテラルフィルタの高速化手法
３．学会等名 電子情報通信学会研究会報告RECONF2021-8
４．発表年 2021年

１．発表者名 藤原 良樹，高前田 伸也
２．発表標題 アルゴリズム・ハードウェア協調設計によるベイジアン畳み込みニューラルネットワークの高速化
３．学会等名 情報処理学会研究報告2021-ARC-245
４．発表年 2021年

１．発表者名 小池 亮，高前田 伸也
２．発表標題 セキュアな不揮発性メモリのクラッシュ一貫性支援の高速化
３．学会等名 情報処理学会研究報告2021-ARC-245
４．発表年 2021年

１．発表者名
高前田 伸也

２．発表標題
アーキテクチャとアルゴリズムの協調による高効率深層学習システムの創出

３．学会等名
第20回情報科学技術フォーラム（FIT 2021）イベント企画「Society5.0を支える革新的コンピューティング技術」（招待講演）

４．発表年
2021年

１．発表者名
Keisuke Kamahori, Shinya Takamaeda-Yamazaki

２．発表標題
GBA: Guided Branch Approximation

３．学会等名
The Fourth Young Architect Workshop (YArch 2022) (Co-located with ASPLOS 2022)（国際学会）

４．発表年
2022年

１．発表者名
高前田 伸也

２．発表標題
機械学習に適したハードウェア，ハードウェアに適した機械学習アルゴリズム

３．学会等名
情報処理学会第84回全国大会 イベント企画「知能と計算とアーキテクチャの新しい関係を目指して」（招待講演）

４．発表年
2022年

１．発表者名
高前田 伸也

２．発表標題
多様性と環境変化に寄り添う分散機械学習基盤の実現に向けて

３．学会等名
電子情報通信学会情報論的学習理論と機械学習研究会 2022-03-IBISML（招待講演）

４．発表年
2022年

１．発表者名
釜堀 恵輔，高前田 伸也

２．発表標題
分岐命令の選択的近似による決定木アンサンブルの高速化

３．学会等名
情報処理学会研究報告2021-ARC-248（招待講演）

４．発表年
2022年

１．発表者名
橋本 信歩，高前田 伸也

２．発表標題
機械学習ベースの動画像処理における近似計算手法の検討

３．学会等名
電子情報通信学会研究会報告CPSY2021-59

４．発表年
2022年

１．発表者名
菅 研吾，高前田 伸也

２．発表標題
高帯域幅メモリ搭載FPGAを用いたランダムアクセス指向メモリアーキテクチャとプログラミングモデルの検討

３．学会等名
情報処理学会研究報告2021-ARC-248

４．発表年
2022年

１．発表者名
久保 龍哉，小池 亮，高前田　伸也

２．発表標題
不揮発性メインメモリにおける効率的な整合性検証手法の検討

３．学会等名
情報処理学会研究報告2021-ARC-248

４．発表年
2022年

１．発表者名
山野 龍佑，高前田 伸也

２．発表標題
カスタマイズ可能！AIアクセラレータジェネレータNNgenを大解剖！

３．学会等名
Design Solution Forum 2020

４．発表年
2021年

１．発表者名
高前田 伸也

２．発表標題
オープンソースコンパイラNNgenでつくるエッジ・ディープラーニングシステム

３．学会等名
第3回ACRiウェビナー：Softwareエンジニアにも使って欲しいFPGAの実力（招待講演）

４．発表年
2021年

１．発表者名
Shinya Takamaeda-Yamazaki, Shinya Fujisawa, Shuichi Fujisaki

２．発表標題
NNgen: A Model-Specific Hardware Synthesis Compiler for Deep Neural Network (Demonstration)

３．学会等名
Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)（国際学会）

４．発表年
2019年

１．発表者名
高前田 伸也，藤澤 慎也，藤崎 修一

２．発表標題
ディープニューラルネットワークのモデル特化ハードウェア合成コンパイラ

３．学会等名
第2回機械学習工学研究会（MLSE夏合宿2019）

４．発表年
2019年

１．発表者名
空閑 康太，高前田 伸也

２．発表標題
CNNのクラスタリングによる圧縮と推論アクセラレータの検討

３．学会等名
電子情報通信学会研究会報告CPSY2021-51

４．発表年
2022年

１．発表者名
高前田 伸也

２．発表標題
多様性と環境変化に寄り添うエッジAI基盤の実現に向けて

３．学会等名
DAシンポジウム2022（招待講演）

４．発表年
2022年

１．発表者名
高前田 伸也

２．発表標題
多様性と環境変化に寄り添う信頼される分散機械学習基盤のための要素技術とその応用

３．学会等名
情報処理学会 連続セミナー 2022 「その先へ 情報技術が貢献できること」（招待講演）

４．発表年
2022年

１．発表者名
橋本 信歩，高前田 伸也

２．発表標題
動画像を入力とした深度推定のHW/SW協調設計によるFPGAベースの高速化手法

３．学会等名
情報処理学会研究会報告2022-ARC-250

４．発表年
2022年

１．発表者名
小池 亮，高前田 伸也

２．発表標題
セキュアNVMの高性能化のためのツリー事前更新

３．学会等名
情報処理学会研究会報告2022-ARC-250

４．発表年
2022年

１．発表者名
空閑 康太，高前田 伸也

２．発表標題
回帰木に基づく畳み込み演算の直接近似手法

３．学会等名
電子情報通信学会研究会報告CPSY2022-26

４．発表年
2022年

１．発表者名
Shinya Takamaeda-Yamazaki

２．発表標題
Algorithm/Hardware Co-design for Reliable AI

３．学会等名
The 2022 International Meeting for Future of Electron Devices, Kansai (IMFEDK 2022)（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
薄井 真之，高前田 伸也

２．発表標題
高位合成における分離型データオーケストレーションの自動合成

３．学会等名
電子情報通信学会研究会報告，Vol.122，No.402，VLD2022-90

４．発表年
2023年

１．発表者名
釜堀 恵輔，高前田 伸也

２．発表標題
オンチップの脅威に対処するためのセキュアなキャッシュシステム

３．学会等名
電子情報通信学会研究会報告，Vol.122，No.403，HWS2022-66

４．発表年
2023年

１．発表者名
空閑 康太，高前田 伸也

２．発表標題
定数係数畳み込み演算を対象とした論理圧縮アルゴリズムの検討

３．学会等名
電子情報通信学会研究会報告，Vol.122，No.402，VLD2022-97

４．発表年
2023年

１．発表者名
深見 匡，高前田 伸也

２．発表標題
複数のインデクスにより競合性ミスを低減した圧縮キャッシュ

３．学会等名
電子情報通信学会研究会報告，Vol.122，No.402，VLD2022-98

４．発表年
2023年

１．発表者名
小池 亮，高前田 伸也

２．発表標題
一時的メモリアクセスリダイレクションによる高性能かつプログラマ・フレンドリーなセキュアNVM

３．学会等名
電子情報通信学会研究会報告，Vol.122，No.402，VLD2022-106

４．発表年
2023年

１．発表者名
高前田 伸也

２．発表標題
魔法戦士のすすめ～科学技術分野の文部科学大臣表彰・若手科学者賞受賞に際して～

３．学会等名
電子情報通信学会研究会報告，Vol.123，No.62，CPSY2023-7（招待講演）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Veriloggen
https://github.com/PyHDI/veriloggen
NNgen
https://github.com/NNgen/nngen
FADEC
https://github.com/casys-utokyo/fadec

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関