ALCA-Next
「グリーンコンピューティング・DX」領域
2024 年度　年次報告書

2023 年度採択

［研究開発代表者名：中島　康彦］

［奈良先端科学技術大学院大学先端科学技術研究科　教授］

［研究開発課題名：多連装マルチレベルパイプライン CGRA］

主たる共同研究者：なし

実施期間　：　2024 年 4 月 1 日～2025 年 3 月 31 日

§1. 研究開発成果の概要

目的は、AI+セキュリティを含む計算インテンシブアプリに、汎用高効率計算基盤を提供し、2050年に 274,000TWhとされる世界のサーバ電力消費量を 11.000TWh に削減し、太陽光発電量現実的上限 87,600TWh の 1/8 に抑えること(2018 年電力消費量=105TWh[LCS/JST]、2019 年太陽光発電量=679TWh[AIST])。独自性は、チップレット適合スケーラブル PiM アーキテクチャ、電力消費が大きいメモリインタフェース狭小化に有効な脱水平(SIMD)・マルチレベルパイプライン(MISD)アーキテクチャ、高速コンパイルが可能なロケーションフリー・多入力CISCベース・リングアレイ型アーキテクチャ、高効率確率的計算技術、および、統合フレームワーク。

【0】インメモリ・ニューロモーフィックデバイス

目標:新規購入のデバイスアナライザにより取得した特性と HSPICE シミュレーションに基づき、大規模積和演算器を実現可能なモデルを構築し、メムキャパシタの将来性を検証。成果:デジタルスイッチドキャパシタによる試作 LSI では、MNIST 認識精度 96%を達成し、180nm としては驚異的な 163GOP/W を達成[**論文1**]。また、メモリスキャパシタを用いる小型回路を試作。さらに、デバイスアナライザにより取得した特性と HSPICE シミュレーションに基づくパラメタを Python 記述のトランスフォーマ・フレームワーク(元は超電導素子向け環境)へ組み込み、予定通り、評価モデルを構築。

【1】中段・後段向け大規模デジタル汎用 CGRA

目標:アプリケーションを大規模言語モデル(LLM)に切り替え、FPGA プロトタイプを用いて、マクロパイプラインを記述できる C 言語フレームワークを実装。成果:AI に限定しない Polybench 等様々なアプリケーションを微細化レベル同等 GPGPU と本格的に比較。PDP で 0.5 倍～657 倍(0.5 倍は、主記憶バンド幅律速の大規模データ処理1件)を達成し、SG 目標 150 倍をクリア[**論文2**]。また、LLM の ggml と llama-v2 のマルチスレッドフレームワークをマルチレーン IMAX に対応付けるフレームワークの実装を完了。さらに、性能向上には HOST-CPU コア数増強の必要性を確認。

【2】初段向け超小型確率的デジタル CGRA

目標:新たに考案した Multi Radix Coding(MRC)の入力数を増やした大規模積和演算器に取り組み、積和演算器単体性能 100TOPS を達成。成果:7nm HSPICE-siml により、MRC ベース大規模積和演算器検証を完了。900.9TOPS/W を達成し、GPGPU(H100)の INT8 効率 8.4TOPS/W に対し 108 倍を達成。また、完全並列 SNN アクセラレータに積和演算器を使用し、FPGA で 40.1 GSOPS(synapse operation/s)/W を達成[**論文3**]。

【3】次世代セキュリティ向けデジタル専用 CGRA

目標:軽量暗号等、より最新のアルゴリズムに対応し、U2CA を IMAX3 フレームワークに適合。成果:昨年度実装した SHA2,SHA3,BLAKE,SM3,ASCON-Hash 等に続き、本年度は、ブロック暗号(AES、SM4 等)、ストリーム暗号(Salsa20、ChaCha20 等)、および軽量暗号(ASCON、ESCH256 等)に対応した。高いエネルギー効率性と再構成可能性が評価された[**論文4**]。なお、IMAX3 にキャッシュ間データ転送機構を追加する統合設計を実施中。

【代表的な原著論文情報】

1. Reon Oshio, Takumi Kuwahara, Takeru Aoki, Mutsumi Kimura and Yasuhiko Nakashima: “Neuromorphic System using Capacitor Synapses”, Scientific Reports, Vol. 15, 3954, DOI: 10.1038/s41598-025-87924-6, Jan. (2025)
2. Tomoya Akabe, Vu Trung Duong Le and Yasuhiko Nakashima: "IMAX: A Power-efficient Multilevel Pipelined CGLA and Applications", IEEE Access, 10.1109/ACCESS.2024.3524415, Jan. (2025)
3. Mingyang Li, Yirong Kan, Renyuan Zhang and Yasuhiko Nakashima: “A Fully-Parallel Reconfigurable Spiking Neural Network Accelerator with Structured Sparse Connections”, IEEE International Symposium on Circuits and Systems (ISCAS) , DOI: 10.1109/ISCAS58744.2024.10558156, Jul. (2024)
4. Pham Hoai Luan, Vu Trung Duong Le, Tuan Hai Vu, Van Duy Tran, Van Tinh Nguyen and Yasuhiko Nakashima, "MRCA2.0: Area-Optimized Multi-grained Reconfigurable Cryptographic Accelerator for Securing Blockchain-based IoT Systems," in IEEE Micro, Nov. 2024.