論文内容の要旨

申請者氏名　　赤部　知也

In recent years, the rapid advancement of AI and HPC has led to a surge in computational demand and energy consumption. AI applications, particularly those involving deep learning and CNNs, require massive computational resources and high memory bandwidth, creating critical challenges in efficiency and power consumption. In response, there is an increasing need for novel architectures that combine flexibility and computational efficiency. While conventional hardware solutions such as GPUs, FPGAs, and ASICs offer specific advantages, they still face challenges in terms of flexibility, design cost, and energy efficiency. To address these issues, this study proposes a novel architecture called Coarse-Grained Linear Arrays (CGLA), aimed at delivering both generality and high efficiency as a computational platform. CGLA adopts a linear data flow structure, which differs from Coarse-Grained Reconfigurable Arrays (CGRAs), enabling reduced communication latency between computational units and more efficient data processing. This study develops a CGLA-based architecture tailored for AI and edge computing and achieves the following technical breakthroughs. First, a method for adjusting the bit width during CNN training was devised, successfully reducing the required bit width from 32 bits to 15 bits while maintaining training accuracy. This approach not only reduced computational load but also optimized memory utilization. Second, a stochastic computing-based fused multiply-add (FMA) unit was designed, achieving a 39% reduction in circuit area and a maximum 63% increase in operating frequency compared to conventional 32-bit floating-point FMA units. As a result, a 46-fold improvement in computational speed was achieved for tasks such as handwritten character recognition. Furthermore, the IMAX3 architecture was proposed, incorporating double buffering to mitigate communication latency and optimize processing for large-scale pipeline applications such as FFT and sparse matrix multiplication. These enhancements equipped IMAX3 with the ability to handle such tasks efficiently. Experimental results demonstrated that IMAX3 reduced computation time by up to 38% compared to its predecessor, IMAX2, and significantly outperformed conventional GPUs in terms of energy efficiency.

論文審査結果の要旨

申請者氏名　　赤部　知也

近年、AIやHPCの急速な進歩により、計算需要とエネルギー消費が急増している。特にディープラーニングやCNNを含むAIアプリケーションでは、膨大な計算リソースと高メモリ帯域幅が必要であり、効率と電力消費の面で重大な課題が生じている。この問題に対応できる、柔軟性と計算効率を兼ね備えた新しいアーキテクチャの必要性が高まっている。GPU、FPGA、ASIC などの従来型ハードウェアには各々の利点があるものの、柔軟性、設計コスト、エネルギー効率の面で依然として課題がある。本研究は、計算プラットフォームとして汎用性と高効率の両方を実現することを目的とした、粗粒度線形アレイ(CGLA)と呼ぶ新しいアーキテクチャを提案している。CGLAは、粗粒度再構成可能アレイ(CGRA)とは異なる線形データフロー構造を採用しており、計算ユニット間の通信遅延の削減とデータ処理の効率化を可能にしている。AIとエッジコンピューティング向けにカスタマイズされたCGLAベースのアーキテクチャを開発し、次の技術的ブレークスルーを達成している。第1に、CNNの学習時にビット幅を調整する方法を考案し、学習精度を維持しながら必要ビット幅を32ビットから15ビットに削減することに成功している。本アプローチは、計算負荷を軽減するだけでなく、メモリ使用率を最適化する。第2に、確率計算ベースの融合積和演算(FMA)ユニットを設計し、従来の32ビット浮動小数点FMAユニットと比較して回路面積を39% 削減し、動作周波数を最大63%向上させた。この結果、手書き文字認識タスクでは、計算速度が46倍向上した。また、FFTや疎行列乗算などの大規模パイプラインアプリケーションの処理を最適化するために、ダブルバッファリングを組み込んだIMAX3アーキテクチャを提案した。これらの機能強化により、IMAX3は、各タスクを効率的に処理できるようになった。実験結果は、IMAX3が、前身のIMAX2と比較して計算時間を最大38%削減し、エネルギー効率では従来型GPUを大幅に上回ることを実証している。
以上、本論文は学術上、実際上寄与するところが少なくない。よって、本論文は博士（工学）の学位論文として価値あるものと認める。