論 文 内 容 の 要 旨

氏 名 (SOMBATSIRI SALITA)	
論文題名	An Architecture Design Space Exploration Method of System-on-a-Chip for CNN-based Artificial Intelligence Platform (畳込みニューラルネットワークに基づく人工知能プラットフォーム向きSoC設計空間探索手法)

論文内容の要旨

Recent advancement of artificial intelligence (AI) algorithms and computing platforms enables various cutting-edge applications, such as image/video analytics and autonomous driving. In order to achieve a real-time response, the edge computing, which locates between the endpoint devices and the cloud, has become more compelling in the paradigm of an AI platform. Designing an AI-based edge computing device is very complicated and time-consuming since the edge computing devices have strict constraints in high-performance, yet compact and low-power. Therefore, the following three requirements are involved in designing optimal architectures for edge computing:(1) quickly evaluate the design quality of an architecture; (2) accelerate deep learning algorithms; (3) efficiently explore the design space to find optimal architectures. This thesis proposes an efficient method in designing edge system-on-a-chip (SoC) architecture for AI applications to fulfill the above-mentioned three requirements.

First, to quickly estimate the execution time of an application on each architecture, this thesis proposes an efficient performance estimation method for configurable multi-layer bus-based SoC. Performance estimation is time-consuming since it analyzes the behavior of the architecture. The speed of the estimation method is often more critical than the estimation accuracy in the early design stage because there are a massive amount of architectures to evaluate during the architecture exploration. The proposed performance estimation method provides a fast and accurate method to evaluate the execution time of each architecture. It analyzes system behavior based on system-level profiling, speculates dynamic bus contention, and predicts bus behavior with graph analysis. The experimental results show that the proposed method has achieved 25.6x speedup over the register-transfer level (RTL) simulation in evaluating the execution time of eight architectures. The error of the estimation results is within 8% compared to the conventional RTL simulation. Hence, the proposed method is efficient and suitable for architecture exploration process.

Second, a parallelism-flexible convolution core for sparse Convolutional Neural Network (CNN) is proposed in order to accelerate the deep learning algorithm, specifically a CNN, as a high-performance intellectual property (IP). A CNN becomes critical for real-time inference processing on edge devices for many applications. The proposed parallelism-flexible convolution core achieves high-performance by maximizing calculation-skip and parallel calculation in all convolutional layers of a CNN. It skips multiply-accumulates (MACCs) related to zero-valued weights efficiently with the use of the compressed CNN model together with the output-stationary scheme. It alternates dataflow and schedules MACCs flexibly according to the specification of each convolutional layer to improve multiplier utilization. The results have shown that the integration of both techniques improves performance by 4x speedup over the baseline architecture and 3x in effective GMACS over prior arts of CNN accelerator.

Third, an architecture exploration of SoCs for CNN-based AI platform is proposed to efficiently explore the design space with IP-based design and system-level design. The complexity of finding optimal architectures in the early design stage lies in IP selection and bus selection because there are a vast amount of IPs, bus architectures, and their parameters. In the proposed architecture exploration method, the IPs and bus architecture are parameterized and explored using a parameter set search tree. The proposed method consists of process mapping, channel mapping, bus protocol mapping, functional block's parameter mapping, functional block's and bus's execution frequency mapping, bus width mapping, and the number of buffer mapping. In the process mapping, a process is mapped to an IP. In the channel mapping, the data transfers are mapped onto either a hierarchical shared bus or configurable multi-layer bus by mapping the data transfers into clusters, each of which is connected to a bus matrix. Then, the functional block's parameters, i.e. the number of instances of each functional block and the number of processing elements (PEs) within each functional block, are selected considering data tiling in order to distribute and parallelize the workload of computation-intensive processes. This parameterization allows the MACCs to be parallelized on multiple instances of functional blocks. The results show that the proposed method discovers varieties of architecture having various functional blocks and multi-layer bus configurations.

This thesis contributes to designing SoCs for CNN-based AI platform at the edge, especially architecture design and optimization in the early design stage. It provides an efficient method to explore the architecture candidates, including a parameterized IPs and multi-layer bus, and evaluate their design qualities. The method can find architectures with superior design qualities within a short time. Hence, it is suitable for discovering good architecture candidates in the early stage of designing an SoC.

様式7

論文審査の結果の要旨及び担当者

氏　　名	(Salita Sombatsiri)		
	(職)		氏　　名
論文審査担当者	主 査	教授	橋本 昌宜
	副 査	教授	尾上 孝雄
	副 査	准教授	谷口 一徹
	副 査	教授	武内 良典 (近畿大学)

論文審査の結果の要旨

本論文は、高性能かつ小面積・低消費電力実装が求められるエッジプラットフォーム向けSystem-on-a-Chip (SoC)を短期間で設計する設計手法に関する研究の成果をまとめたものであり、以下の主要な結果を得ている。

(1)多様なバス構造に対応した高速性能見積もり手法の提案

System-on-a-Chipの開発においては、設計とその評価を繰り返し、最終的な設計を決定していく。そのため、設計評価に膨大な時間が必要であり、効率的な評価手法が望まれている。本研究では設計結果の性能を効率よく見積もる手法を提案している。提案手法は、SoC内のバス上で行われるデータのトランザクションから、性能見積もりを行う。提案手法は、従来手法と比較して8%以内の誤差で、25.6倍の見積もり高速化を達成した。

(2)並列構成が可変な高性能畳み込みニューラルネットワーク用コアの提案

畳込みニューラルネットワーク (Convolutional Neural Network: CNN) は、画像処理向けニューラルネットワークとして広く利用されるようになったが、エッジプラットフォームでの利用を考えるとその演算量は膨大である。本論文では、エッジプラットフォームで CNN を利用可能にするのためのアーキテクチャとして、データフローの変更や積和演算の柔軟なスケジューリングが可能で、疎なネットワーク荷重が計算量削減に活用でき、並列度が可変であるCNN用コアを提案した。提案したCNN用コアは基準アーキテクチャと比較して4倍の高速化を、また既存研究に対して一秒あたりの積和演算性能(GMACS)で3倍の性能を実現した。

(3) CNNベース人工知能プラットフォーム向けSoCのための設計探索手法の提案

設計資産 (Intellectual Property: IP) の活用とシステムレベルの設計を組合せた、設計初期段階向けのアーキテクチャ探索手法を提案した。設計初期段階での最適アーキテクチャ設計に重要なことは、使用するIPの選択、そのIP間を接続するバスアーキテクチャの選定、ならびにそれらに含まれるパラメータの決定である。本論文では、IP、バスアーキテクチャ、ならびにそれらのパラメータの決定問題を、パラメータ探索木を用いて定式化した。提案手法は、プロセスマッピング、チャネルマッピング、プロトコルマッピング、機能ブロックマッピング、バスアーキテクチャ選択、バスプロトコルマッピング等のパラメータを問題に含めることができ、(2)のCNNコアを含む様々な機能ブロックを有するマルチレイヤバス構成のアーキテクチャにも適用することが可能である。

以上のように、エッジプラットフォーム向けCNN SoCを効率良く設計する手法の研究により、既存手法では達成困難であった 高性能、小面積で低消費電力動作可能なAI用SoCが設計可能となる点で非常に有用である。これにより、エッジプラットフォームでのCNNベース AIの利用拡大が期待できる。したがって、本論文は博士 (情報科学) の学位論文として価値のあるものと認める。