（ 続紙 1 ）

京都大学	博士（情報学）	氏名	程　全（Quan CHENG）
論文題目	Design and Reliability Analysis of System-on-Chips for Artificial Intelligence Applications at the Edge （エッジ人工知能アプリケーション向けのシステムオンチップの設計と信頼性解析）		

（論文内容の要旨）

This thesis targets energy-efficient edge artificial intelligence (AI) system-on-chip (SoC) designs with necessary and sufficient reliability consideration and studies (1) design of edge AI SoCs, and (2) reliability analysis of SoCs running AI applications. The former aims at energy-efficient SoC designs, especially the parallel computation to improve power efficiency, while the latter intends to explore the reliability of the proposed SoC platforms, providing valuable insights and serving as crucial references for future research.

Chapter 1 lays the foundation by presenting the background and objectives of the dissertation. This chapter provides a brief overview of the current research landscape concerning neural networks (NNs) for image processing and then outlines the progress of hardware platforms in edge AI. Subsequently, this chapter delves into the research status of single-event effects (SEEs). Finally, the chapter outlines the objectives and provides an overview of the overall organization of the dissertation.

Chapter 2 introduces three proposed AI accelerators. In the multi-bit-width (MBW) Booth accelerator, a multi-precision (MP) multiplier based on the Booth algorithm is proposed to realize INT2/4/8 multiplication for energy-efficient computation. Additionally, a vector-based systolic array is proposed to accelerate matrix multiplications effectively. A pre-encoding radix-4 Booth algorithm, highly compatible with data-reuse and weight offline pre-processing, is proposed to simplify multiplication for area and power consumption reduction. Furthermore, for low-power design, the non-zero weights above 50% sparsity are steered into arithmetic units, such that each computation logic supporting the computation of sparse data can reduce the number of MACs by half and realize the same amount of computation as a regular accelerator without sparsity mechanism. Regarding the Booth-value-confined (BVC) MP accelerator, a radix-8 Booth-based MP multiplier is proposed with lower overhead than traditional multipliers. Prohibiting the ±3 cases in both the model training and inference stages and eliminating the encoder based on the feature of offline weights can simplify multiplication in NN calculations, substantially optimizing the latency, area, and power in the PE array. Moreover, we propose an accelerator with 16 near-memory engines (NMEs) enabling the data-reuse of weights and activations for NN inference from inter-/intra-NME-order views. The weight memory is based on the near-memory computing (NMC) framework and is directly connected to the PE array to reduce the FF consumption and the path for data movement. The developed AI accelerators are experimentally evaluated with simulation and silicon measurement.

Chapter 3 describes the design and construction of SoC platforms. Three SoCs are realized on field programmable gate array (FPGA) and application-specific integrated circuit (ASIC) platforms. Regarding the flash-based FPGA platform, using an open-source lightweight RISC-V core as a foundation, we construct an SoC with an MBW accelerator to create a practical AI hardware platform. On the ASIC platform, we design two SoCs at the edge. Leaning on a lightweight RISC-V core, we propose a single-core AI processor incorporating a BVC MP accelerator with NMC-friendly data flow. Furthermore, we propose a multi-core multi-mode SoC for versatile safety-critical AI applications. Targeting different application scenarios, the

four supported modes are: 1) accelerator mode, 2) multi-core mode, 3) fault-tolerant mode supporting dual-core-lock-step (DCLS), and 4) fully-access mode. The performance of the developed SoCs is evaluated with simulation and silicon measurement.

Chapter 4 performs fault injection experiments and radiation experiments for the proposed SoCs running AI applications to evaluate the reliability of the proposed SoCs in Chapter 2. A pre-silicon fault injection simulation framework is proposed to simulate the ASIC design from the netlist view. Post-silicon fault injection frameworks are constructed to inject the errors into hardware platforms from the physical view. Also, neutron radiation experiments and alpha particle radiation experiments are performed to explore the reliability of the proposed SoCs thoroughly.

Chapter 5 summarizes the thesis and outlines the possible directions for future work.

（論文審査の結果の要旨）

本論文は、必要十分な信頼性の考慮を伴うエネルギー効率の良いエッジ人工知能（AI）システムオンチップ（SoC）設計を目的として、エッジAI SoCの設計、ならびにAIアプリケーションを実行するSoCの信頼性解析に取り組んだものである。得られた主要な成果は以下の通りである。

1. 畳み込みおよび行列計算を加速するために、3つのアクセラレータアーキテクチャを提案した。乗算器の開発では、マルチ精度計算および近似計算に基づく手法を考案した。提案したマルチ精度ブース積和演算器は、INT2/4/8の計算をサポートし、従来の基数4ブース積和演算器と比較して回路規模削減を実現した。提案した基数8ブースに基づくブース値制限(BVC)マルチ精度演算素子(PE)は、マルチ精度ブースPEと比較して電力と面積の削減を達成した。

2. ベクター型シストリックアレイ、構造的スパース性、およびニアメモリ計算（NMC）を活用したデータフローを提案した。NMCデータフローとベクター型シストリックデータフローを組み合わせることで、高いピークエネルギー効率を達成した。

3. 上述のアクセラレータアーキテクチャを活用して、BVCマルチ精度アクセラレータを搭載したシングルコアSoC、ならびに複雑なAIワークロードに対応するマルチコアSoCを開発した。

4. SoCの信頼性評価のために、製造前後に利用する2つの故障注入フレームワークを提案した。開発したSoCに対して、中性子放射実験およびアルファ粒子放射実験を実施し、プロセッサの命令メモリ、データメモリが最も脆弱なコンポーネントであることを明らかにした。アクセラレータ内のほとんどのソフトエラーは影響が小さく、SoCの信頼性向上のためにはプロセッサコアの強化に取り組むべきであることを示した。

以上のように、本論文は高い信頼性が求められるAIエッジアプリケーションの高エネルギー効率実装について、アーキテクチャ提案とシリコンチップによる評価を行っており、学術上・実用上の寄与が認められる。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和６年８月１日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。
なお、本論文は、京都大学学位規程第１４条第２項に該当するものと判断し、公表に際しては、（令和８年９月３０日までの間）当該論文の全文に代えてその内容を要約したものとすることを認める。