FPGA に基づいた AI アクセラレータのアーキテクチャとソフトエラー率の関係評価

1250084 四戸 颯（量子･古典集積回路研究室）
（指導教員 廖 望 講師）

1. はじめに

機械学習の普及により組込み向けに小型で高性能な並列計算デバイスである FPGA が AI のモデルを高速計算するアクセラレータの実装が注目されている。ソフトエラーは放射線が半導体に作用する一時的な故障で FPGA は回路の構成情報を記録する CRAM がソフトエラーに弱く誤動作を起こしてしまう問題がある [1] 。本稿では、FPGA を用いて回路規模や計算処理時間の異なるアーキテクチャを設計し、エラー率評価を行う環境を構築してエラー率の変化を調査し、エラーに強いアーキテクチャ設計に貢献する。

2. アクセラレータのアーキテクチャ設計

アクセラレータのアーキテクチャ設計として、(1) 各ニューロンの演算を複数の演算部に分け、1 つの演算部で複数の積和演算をまとめて計算を行う性能の高い大規模な分散アクセラレータ、(2) 回路規模が小さく、すべてのニューロンの演算を 1 つの演算部のみで順次行い、複数の積和演算をまとめて計算を行う小規模な集中アクセラレータ、(3) エラーに強いと想定した回路規模が小さく、各ニューロンの演算を複数の演算部に分け、1 つの演算部で単一の積和演算を順次行う小規模な分散アクセラレータを提案し、この 3 種類のアーキテクチャに対して入出力が同様になるように設計を行いアクセラレータがエラーによる誤動作について評価する。

3. アクセラレータのソフトエラー評価手法

FPGA に構築したアクセラレータのソフトエラー評価方法は少ないコストや時間で実施することのできるエラー挿入手法を用いた。図 1 のようにエラー挿入を行う環境を構築してアクセラレータのアーキテクチャ部分のみを抽出して対象全てにエラー挿入を行った。

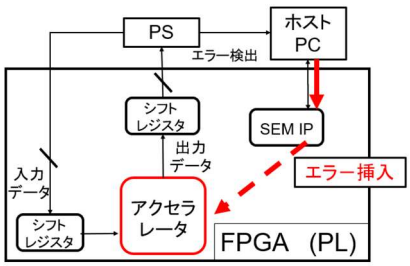

図 1 エラー挿入の構造

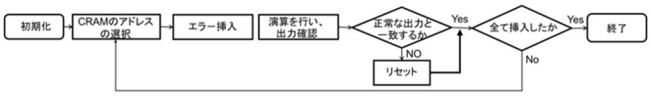

図 2 エラー挿入の流れ

誤動作の内訳として動作への影響が大きい順に (1) アクセラレータ動作停止、(2) 部分的ニューロンの演算結果が 0 に固まった 0-stuck、(3) 複数のニューロン出力誤り、(4) 1 つの出力誤りに分けて評価する。

挿入後エラーのアクセラレータの信頼性比較として $AVF_{DUT\_per\_cal}$ の値を用いる。値が小さいほど信頼性は高くなる。

$$AVF_{DUT} = \sigma \cdot P_{critical} \cdot factor_{cal} \quad (1)$$

$$AVF_{DUT\_per\_cal} = \frac{AVF_{DUT}}{factor_{cal}} \quad (2)$$

$$factor_{cal} = \frac{1}{N_{cycle}} \quad (3)$$

4. 実験結果

エラー挿入を行った結果を表 1､その誤動作の内訳を図 3 に示す。誤動作を起こした CRAM ビットをクリティカルビット（cb）として定義する。

表 1 エラー挿入の結果

アーキテクチャ	CRAM 数	cb 数	誤動作率
大規模な分散	398053	12601	3.1%
小規模な集中	62892	4679	7.0%
小規模な分散	75721	4409	5.6%

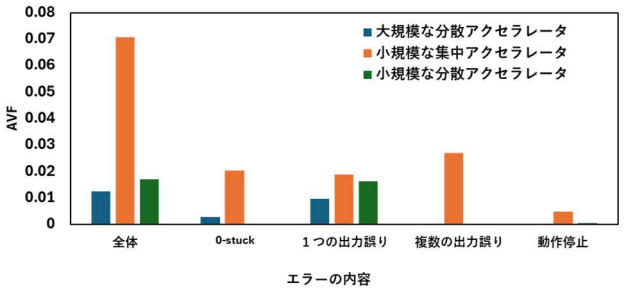

図 3 各出力エラーの内訳

以上の結果から全体で見ると大規模な分散アクセラレータが最も信頼性の高いアーキテクチャになった。しかし、影響度の大きい 0-stuck の数が小規模な分散アクセラレータよりも多いため、演算結果の信頼性に関しては小規模な分散アクセラレータが優れる。さらに小規模なアクセラレータ同士で比較した結果、回路面積が小さい場合に分散化設計を優先することで誤動作を低減できる傾向を示唆した。

5. まとめ

アクセラレータのアーキテクチャ設計部分のエラーに強い設計の提案とよるエラー率の関係について挿入実験結果についての分析を行った。

参考文献

[1] I. Souvatzoglou et al., “The impact of hardware folding on dependability in space-borne fpga-based neural networks,” in 2022 International Conference on Field-Programmable Technology (ICFPT), 2022, pp. 1–1.

[2] I. Souvatzoglou et al., “Assessing the reliability of fpga-based quantized neural net-works under neutron irradiation,” IEEE Transactions on Nuclear Science, vol. 71,no. 12, pp. 2565–2577, 2024.

[3] Z. Gao et al., “Modeling the effect of seus on the configuration memory of sram-fpga-based cnn accelerators,” IEEE Journal on Emerging and Selected Topics inCircuits and Systems, vol. 14, no. 4, pp. 799–810, 2024.

[4] Q. Cheng et al., “Reliability exploration of system-on-chip with multi-bit-width ac-celerator for multi-precision deep neural networks,” IEEE Transactions on Circuitsand Systems I: Regular Papers, vol. 70, no. 10, pp. 3978–3991, 2023.