Research on General-Purpose Power-Efficiency Coarse Grained Linear Arrays

Name: Tomoya Akabe
Laboratory's name: Computing Architecture
Supervisor's name: Yasuhiko Nakashima

Abstract:

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.