Design and Optimization of Energy-Efficient Neural Network Accelerators Using Emerging Computing Paradigms

Name: Zhu Guangxian
Laboratory's name: Computing Architecture
Supervisor's name: Yasuhiko Nakashima

Abstract:

The demand for energy-efficient, high-performance computing in data-intensive applications like machine learning necessitates novel computing architectures. This paper contributes to power-efficient computing by leveraging new principles such as stochastic computing, innovative neural network topologies, and superconducting devices. Firstly, we introduce an ultra-compact calculation unit with temporal-spatial reconfigurability based on a Bisection Neural Network (BNN) topology. This design allows flexible partitioning of processing elements, achieving spatial reconfigurability by adjusting unit shapes and locations. Temporal reconfigurability is achieved through stochastic computing logic by adjusting bit-stream lengths, enhancing accuracy. Experimental results show our unit outperforms state-of-the-art approximate units in energy efficiency. Secondly, we present SuperSIM, a benchmarking framework for neural networks using superconducting Josephson devices, focusing on Adiabatic Quantum Flux Parametron based Processing-in-Memory architectures. SuperSIM offers detailed simulations and assessments, supporting various AQFP memory types, clocking methods, and both single and multi-bit designs. It provides precise measurements of energy consumption, delay, and area. Case studies examine the impacts of data precision, crossbar size, operating frequency, and clocking schemes, validating SuperSIM's effectiveness for advancing AQFP PIM chip development. These contributions advance energy-efficient computing architectures and neural network accelerators by introducing temporal and spatial reconfigurability and providing tools for optimizing superconducting technologies, paving the way for more efficient and scalable computing systems.