以記憶體為中心大型神經網路加速器之模型分析與軟硬體協同設計空間探索

Design Space Exploration for Scalable DNN Accelerators Using a Memory-Centric Analytical Model for HW/SW Co-Design

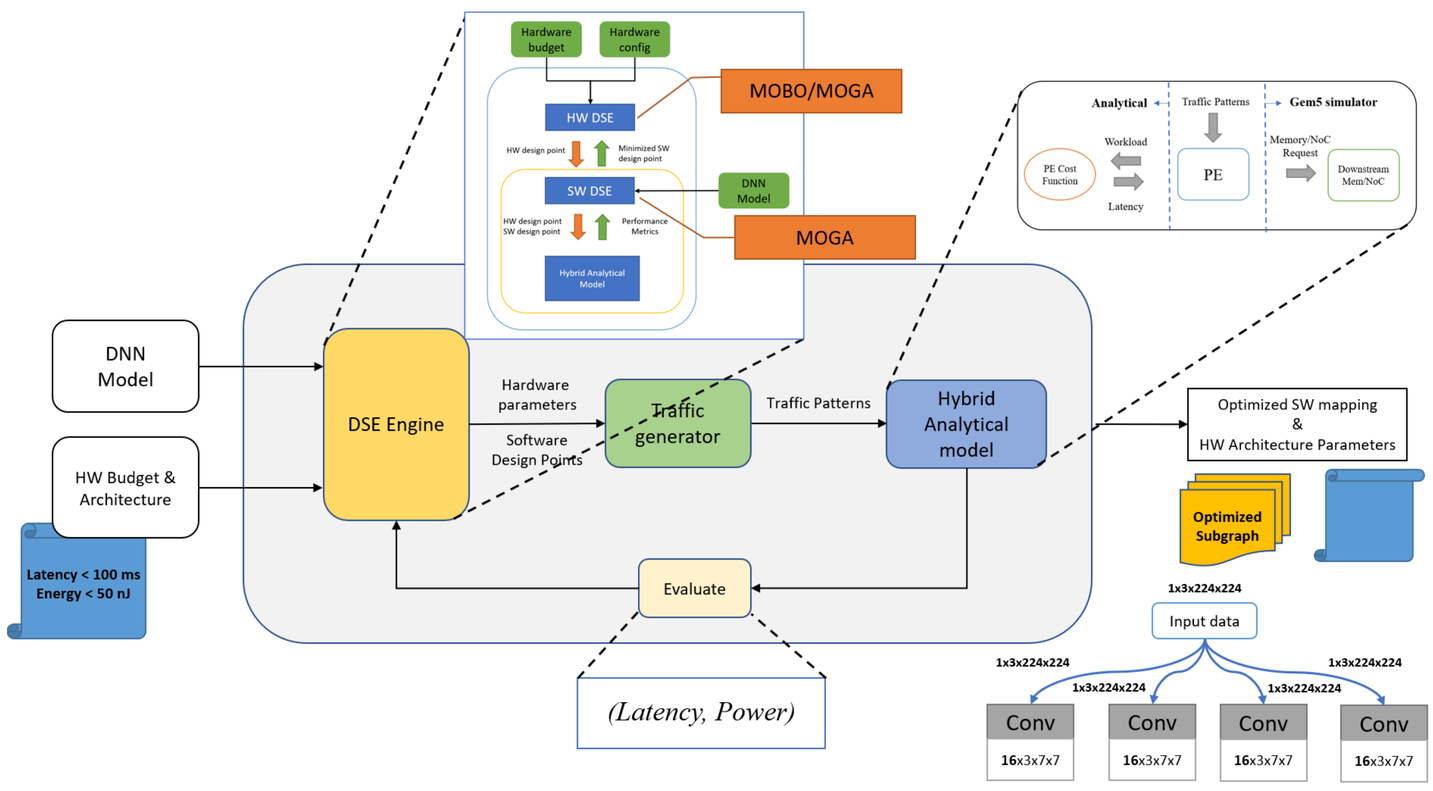
國立陽明交通大學 資通安全碩士學位學程

研究生:黃威淳 指導教授:陳添福

Introduction

The growing complexity of deep learning models has increased the need for efficient DNN accelerators. This has driven research into AI models and hardware architectures, but many designs aren't always efficient or cost effective. A key challenge is to create a flexible hardware simulator and ensure efficient software mapping. While there's been progress in optimizing hardware and software individually, their combined co-design remains complicated. Most design space exploration (DSE) approaches rely on data reuse metrics, which can be misleading. Our work presents a comprehensive DNN accelerator co-design framework that integrates a DSE engine, a traffic generator, and a hybrid analytical model to enable fast and accurate performance evaluation.

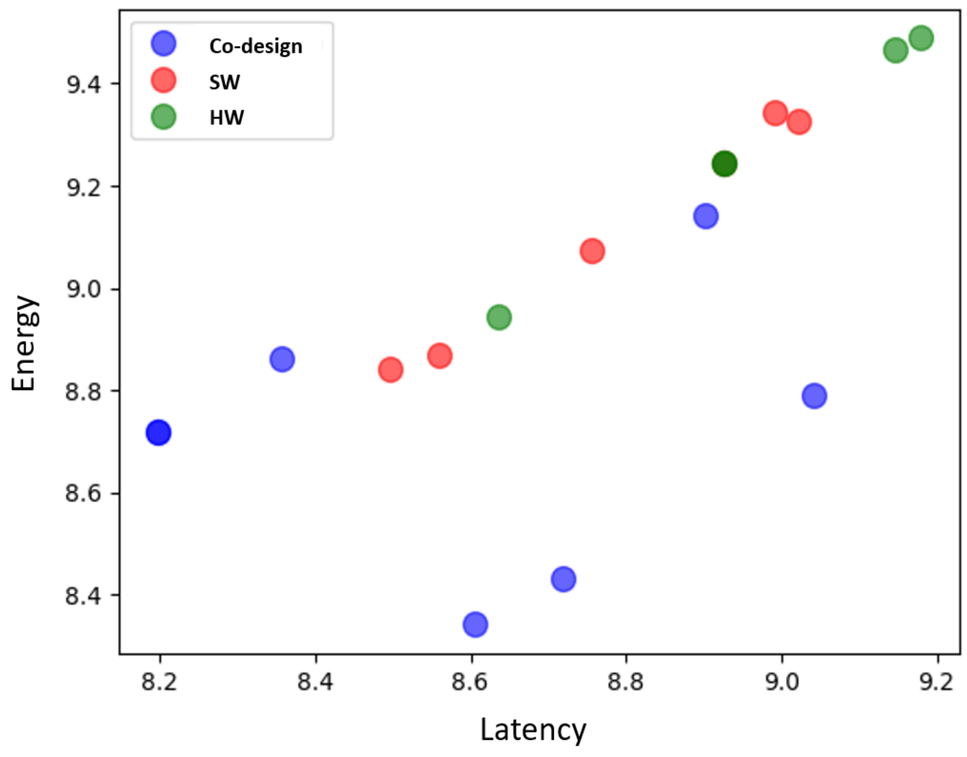
System Architecture



- **DSE Engine**
  - Generate HW/SW design space and design space pruning
  - We provide an interface for design space manipulation for different search algorithms
  - Search algorithms select a design point as the input of the traffic generator
- **Traffic Generator**
  - Analyze the execution behavior of the DNN workload
  - Generate the traffic patterns based on the selected design point
- **Hybrid Analytical Model**
  - PE computation (Analytical)
  - NoC/Off-chip traffic (Simulation)
  - Evaluate the given DNN model and user-defined hardware architecture according to traffic patterns

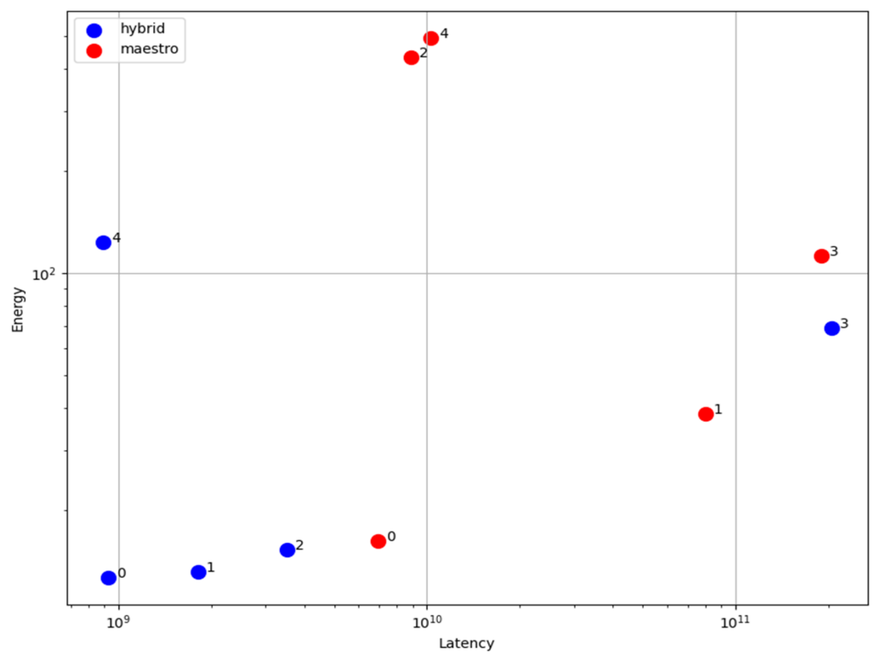
Evaluation

Evaluation of HW, SW, and Co-design DSE



The solution of Co-design are better than only SW of HW optimization

Comparison of design points evaluated by hybrid analytical model and evaluated by MAESTRO



Better performance is due to the **comprehensive** and **accurate memory system** simulated by the hybrid analytical model during the evaluation