Energy-Efficient and Low-Latency SSM-Transformer Accelerator for LLMs

Presented by:
Ming Zhang, SDU Microelectronics, University of Southern Denmark

This PhD project centers on an energy-efficient and low-latency SSM-Transformer accelerator for large language models through hardware-algorithm co-design. Aiming to overcome the high latency and energy consumption of conventional Transformers, as well as the insufficient hardware support for hybrid SSM-Transformer workloads, the project will propose, implement, and prototype a unified and reconfigurable architecture with adaptive dataflow and hierarchical memory, enabling efficient deployment of LLMs on edge and low-power platforms with validated performance and publishable outcomes.

Background

Large language models based on the Transformer architecture demonstrate outstanding natural language processing performance, but their self-attention mechanism causes high latency, memory pressure, and energy consumption, restricting deployment on edge and low-power devices.

State Space Models (SSMs) offer linear complexity and efficient long-sequence modeling, and recent hybrid SSM-Transformer models combine the advantages of both structures. However, hardware acceleration for such hybrid models remains under-explored. Traditional accelerators cannot efficiently support SSM's unique computation patterns, and heterogeneous workloads lead to dataflow and scheduling inefficiencies. Lacking unified reconfigurable hardware creates a gap between algorithm advances and practical deployment.

This research will focus on a dedicated hardware-algorithm co-designed accelerator to achieve high energy efficiency, ultra-low latency, and efficient deployment of hybrid SSM-Transformer LLMs.

Primary Objective

Develop an energy-efficient and low-latency accelerator for hybrid SSM-Transformer LLMs via hardware-algorithm co-design, outperforming conventional accelerators while maintaining performance.

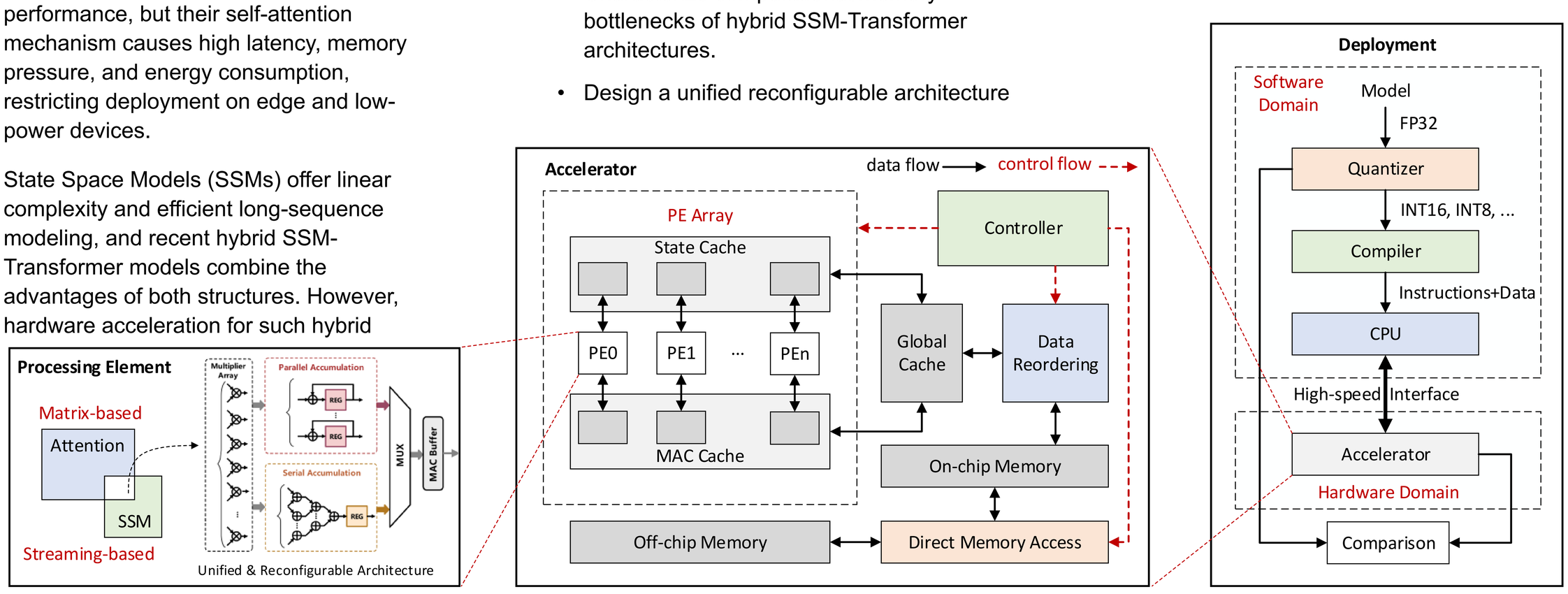
Secondary Objectives

- Characterize computational/memory bottlenecks of hybrid SSM-Transformer architectures.
- Design a unified reconfigurable architecture natively supporting SSM and Attention workloads.
- Implement and simulate the accelerator in RTL, then prototype and evaluate on FPGAs against state-of-the-art baselines.
- Develop a full software toolchain (quantization + compilation) for model deployment.
- Publish findings and support potential chip tape-out/productization.

Methodology

This research adopts a four-stage methods:

- Analyze computation intensity, memory access, and data utilization of hybrid models to identify bottlenecks in SSM-Transformer architectures. Explore hardware-friendly optimizations such as block fusion and simplified gating to reduce computational overhead.
- Design a reconfigurable computing core with shared computing logic to support SSM and Transformer operations. Propose adaptive dataflow, hierarchical on-chip memory, and dynamic scheduling to improve resource utilization and reduce latency.
- Implement the accelerator in RTL, verify functionality on FPGA, and evaluate latency, throughput, power, and efficiency against mainstream platforms. Complete ASIC implementation flow to validate chip feasibility.
- Develop a quantization and compilation toolchain for model deployment. Validate the entire system on real-world hybrid SSM-Transformer models to demonstrate performance and efficiency improvements.



A Software-Hardware Co-Design Framework of Unified and Reconfigurable SSM-Transformer Accelerator for LLMs

Ming Zhang, SDU Microelectronics, University of Southern Denmark

Ming Zhang, SDU Microelectronics, University of Southern Denmark