A Flexible Accelerator Architecture for State Space and Linear Recurrent Models

Yuankang Zhao[(1)], Osman Burak Cevik[(2)], Younes Bouhadjar[(3)] and Emre Neftci[(1)]

[(1)] PGI-15, Forschungszentrum Juelich, RWTH Aachen [(2)] Sabanci University

[(3)] Fraunhofer Institute for Integrated Circuits, IIS

State-space and linear recurrent models have recently emerged as an attractive alternative to attention-based transformer architectures, offering favorable scaling, especially in terms of memory footprint, for long-context processing[1, 2]. However, their efficient hardware realization remains challenging because these recurrent models such as Gated Linear Attention (GLA)[3] combine matrix operations, element-wise gating, recurrent states and updates with diverse data dependencies that do not map cleanly to conventional dense tensor-based accelerators[4].

In this work, we present a programmable accelerator architecture for such workloads, built around a multi-functional systolic array and a flexible processor pipeline. The proposed design supports reconfigurable operating modes, adaptable dataflows, and embedded recurrence within the array, enabling both feed-forward and recurrent computation patterns to be executed efficiently with this architecture. In particular, output-stationary execution is exploited to retain intermediate states locally and realize recurrent/state-update behavior with reduced external data movement. This allows the same compute fabric to support the heterogeneous arithmetic operations appearing in modern linear-attention and state-space-inspired models, while preserving architectural flexibility for other recurrent algorithms. By combining reconfigurable systolic processing with workload-aware pipeline control, the proposed system provides a practical hardware template for accelerating emerging recurrent and linear-time sequence models.

[1] A Gu et al., Mamba: Linear-Time Sequence Modeling with Selective State Spaces, arXiv, 2024.

[2] T Dao et al.,Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality, arXiv, 2024.

[3] S Yang et al., Gated Linear Attention Transformers with Hardware-Efficient Training, arXiv, 2024.

[4] J Li et al., MARCA: Mamba Accelerator with Reconfigurable Architecture, ICCAD'2025, 234, 9, 2025.