QS4D: Quantization-aware training for efficient state-space model acceleration

S. Siegel[(1)], M.-J. Yang[(1)], Y. Bouhadjar[(2)], M. Fabre[(2)], E. Neftci[(2, 3)], and J.P. Strachan[(1, 3)]

[(1)] Peter-Grünber-Institute (PGI-14), Forschungszentrum Jülich GmbH, Jülich, Germany, [(2)] Peter-Grünber-Institute (PGI-14), Forschungszentrum Jülich GmbH, Jülich, Germany, [(3)] Faculty of Electrical and Information Technology, RWTH Aachen University, Aachen, Germany

State-Space Models (SSM) [1] have emerged as a powerful neural network model for processing of sequential data for applications such as language analysis and generation, video understanding, and, in general, edge sensor data processing. In this capacity, they rival the performance of contemporary Transformer models, but demand only a fixed amount of memory. This feature makes them promising for resource-constrained edge-computing applications. However, there are few studies so far that lay out clear deployment pathways for SSMs on edge computing substrates that often operate at low fixed-point accuracies, have limited storage, or are subject to analog noise in the case of emerging concepts like analog in-memory computing (AIMC).

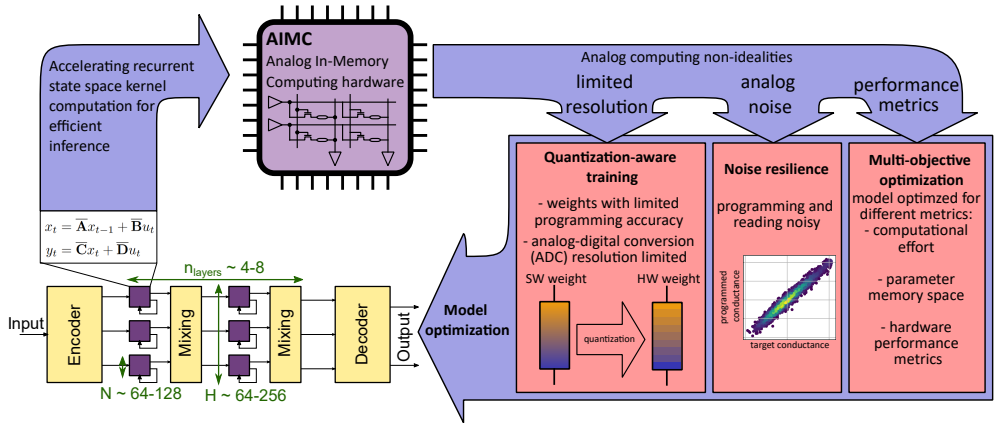

Figure 1: Quantization-aware training opens up deployment pathways for State-Space models on edge computing hardware by reducing the computing complexity and size, and increasing noise resilience.

We investigate how quantization-aware training (QAT) of SSMs allows for more aggressive quantization of weights, states, and activations in these models than traditional post-training quantization (PTQ) alone [2]. We quantify the benefits in relevant metrics, showcasing a reduction of computational complexity by up to 130x, a lowering of model size of 12x, and a reduction of the complexity of analog-to-digital conversion in AIMC by 10x.
Furthermore, we highlight hardware-software co-design opportunities like the trade-off of quantization and model size, increased noise resilience, and structural pruning potential by QAT. Finally, we present a kernel fusion method dedicated to AIMC, allowing it to perform the entire SSM kernel function in a single operation step in hardware [3].
This study is an important step toward the efficient deployment of SSMs in computing hardware in edge applications by paving the route to address this substrate's constrained computing capabilities.

[1] A. Gu et al. NEURIPS 2022, 35, 35971-35983, 2022.

[2] S. Siegel et al. AIS: e202501019, 2025.

[3] S. Siegel et a., ISCAS 2025, pp. 1-5. IEEE, 2025.