A low-latency digital Flash-Attention co-processor in 18 nm FD-SOI CMOS using dataflow-optimized architecture

Joaquin Antonio Cornejo[(1)], Filipe Pouget[(1)], Sylvain Clerc[(1)],
Tifenn Hirtzlin[(2)], Benoit Larras[(3)], Andreia Cathelin[(1)], Antoine Frappé [(3)]

[(1)]STMicroelectronics, Crolles, France,
[(2)]CEA-LETI, Université Grenoble-Alpes, France
[(3)]Univ. Lille CNRS, Centrale Lille, JUNIA, IEMN

Attention mechanisms have emerged as a fundamental building block of modern machine learning systems [1], yet their efficient deployment on silicon remains constrained by data movement, memory footprint, and numerical precision requirements [3]. These limitations are particularly pronounced in edge environments, where strict energy and latency budgets preclude conventional high-throughput implementations.

Here, we investigate the on-chip realization of attention mechanisms in 18 nm FD-SOI CMOS technology, adopting a hardware-oriented perspective that departs from standard software-centric formulations. Recent works such as FlashAttention have highlighted the critical role of memory access optimization in improving performance [2], motivating alternative implementations that better align with hardware constraints. Instead of relying on memory-intensive computation patterns, we explore structured dataflows and adapted numerical representations that reduce intermediate storage and improve execution efficiency.

We present a system-level architecture combining a programmable RISC-based processing unit with a dedicated accelerator designed for attention workloads. The proposed approach emphasizes locality of computation, reduced off-chip memory access, and efficient mapping of accumulation operations, enabling low-latency processing of sequential data under tight resource constraints.

By co-designing algorithmic primitives with hardware structures, this work highlights a pathway toward practical deployment of attention-based models in embedded systems. More broadly, it illustrates how advanced learning mechanisms can be reinterpreted to match the physical realities of modern semiconductor technologies [4].

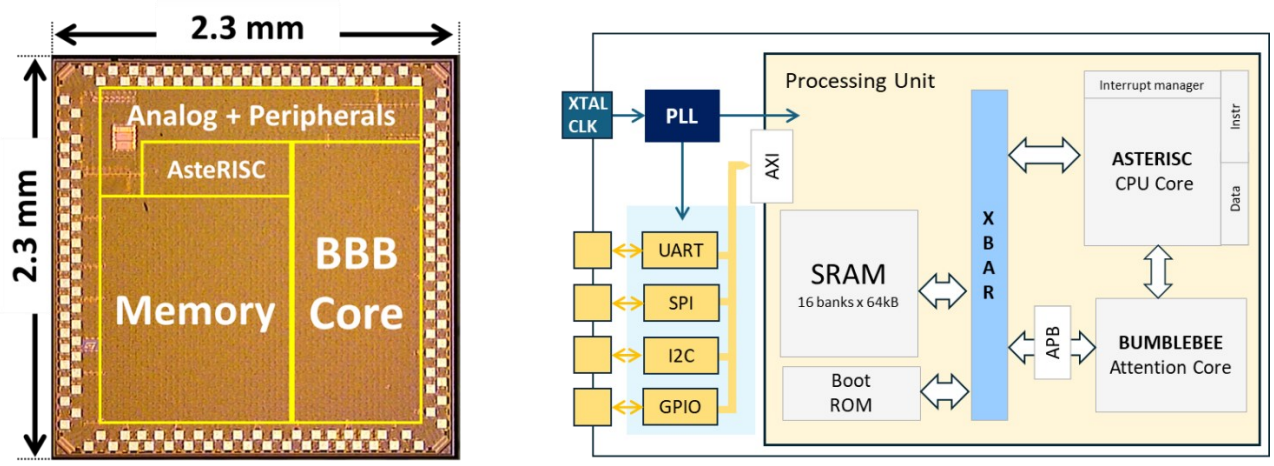

Figure 1 : System-level integration of attention acceleration in 18 nm FD-SOI CMOS.

[1] Vaswani, A. et al. *Attention is all you need.* Adv. Neural Inf. Process. Syst. ***30**, 5998–6008 (2017).*
[2] Dao, T. et al. *FlashAttention* NeurIPS *(2022).*
[3] Sze, V. et al. *Efficient processing of deep neural networks.* Proc. IEEE ***105**, 2295–2329 (2017).*
[4] Jouppi, N. P. et al. *In-datacenter performance analysis of a TPU.* ISCA *(2017).*