IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024

Superconducting Array of Arrays for Acceleration of Transformers

Manu Perumkunnil, Kartik Lakshminarasimhan, Udara De Silva, Debjyoti Bhattacharjee, Trent Josephson, Quentin Herr, Anna Herr

imec, Leuven, Belgium

E-mail: manu.Perumkunnil@imec.be

***Abstract*–Superconducting Digital (SCD) computing systems operate at extremely low temperature ranges (4K -77K) having ultra-fast, low power switching characteristics and zero resistance wires. However, complex logic scaling and dense memories have been particularly challenging for SCD technology even with recent innovations [1]. Given these limitations, simple yet dense throughput architectures like Systolic Arrays are more suited for SCD as compared to complex Out-of-Order cores. Even from an algorithmic perspective, for large Machine Learning (ML) models like Transformers, Systolic Arrays have been considered as an optimal architectural choice [2],[3]. However, the latency overhead due to the RC parasitics of wires in room temperature implementation of systolic architectures results in skewed systolic fill and drain cycles and thus under-utilized MAC units. In recent transformer models, this becomes more prominent since the matrix sizes are bigger with different aspect ratios.**

In this work, we explore customized superconducting 'Regular Arrays' (based on Josephson Junctions) for the acceleration of Transformers. These SCD arrays eliminate the fill and drain cycles in the systolic array due by means of additional wires across the different Processing Elements (PEs). Our custom implementation keeps the array active and saves buffer space by avoiding the skewed systolic cycles to optimize for SCD memory constraints. A scaled-out implementation of multiple such SCD array chips (with 3D stacked local superconducting Josephson-SRAM) can be integrated on a superconducting interposer via superconducting 3D interconnects to function as a standalone superconducting accelerator blade for a wide variety of ML workloads. A GPT-3 transformer is mapped on such an array of arrays, parallelizing the self-attention layer of the transformer block. The superconducting blade implementation achieves a speedup of >25x over a room temperature scaled-out systolic array.

Keywords (Index Terms)–Superconducting digital, Josephson SRAM, systolic arrays, regular arrays, transformers

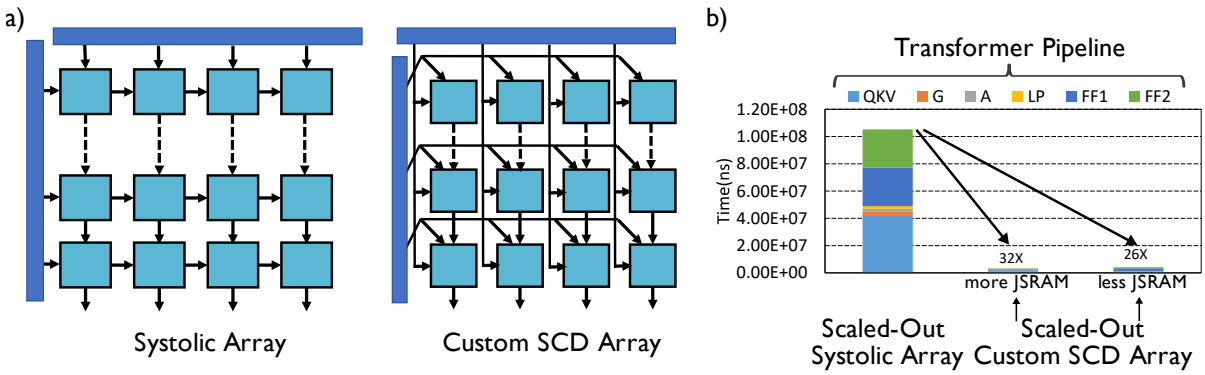

Fig. 1: a) *Systolic Array vs Custom SCD array (Regular Array architecture).* b) *Execution time for the custom SCD array implementation with varying amounts of local JSRAM memory.*

References

1. Quentin Herr, Trent Josephson, Anna Herr, Appl. Phys. Lett. 122, 182604 (2023).
2. Kung, IEEE computer, Vol. 15, Issue 1, DOI: 10.1109/MC.1982.1653825, (1982).
3. Jouppi, Norman P et. al., ISCA'17, DOI: 10.1145/3079856.3080246, (2017).

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024.
Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.