Exploiting Error Resilience of Iterative and Accumulation based Algorithms for Hardware Efficiency[1]

G.A. Gillani, Faculty of EEMCS, University of Twente, Netherlands.

3 x 3 ≈ 7 Chip-area, Power, and Latency Improvements 3 x 3 ≈ 11

Introduction

- While the efficiency gains due to process technology improvements are reaching the fundamental limits of computing, emerging paradigms like approximate computing provide promising efficiency gains for error resilient applications [2].
- Keeping in view a wide range of iterative and accumulation based algorithms in digital signal processing, this thesis investigates systematic approximation methodologies to design high-efficiency accelerator architectures for such algorithms.
- As a case study of such algorithms, we have applied our proposed approximate computing methodologies to a radio astronomy calibration application [7].

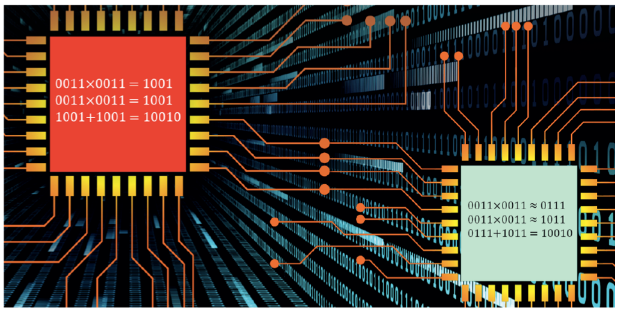

Figure 1. Accurate chips (ICs) are hot and bulky, one might have to wait or slow down; approximate chips are cool and smart, one might get the required results.

Energy-efficient Accelerator Design for Iterative Algorithms

- Our proposed error-resilience resilience analysis methodology, adaptive statistical approximation model [3], provides a way to quantify the number of iterations that can be processed using an approximate core while complying with the quality constraints.
- Targeting energy efficiency, we propose an accelerator design for iterative algorithms [4], see Fig. 2. Our design is based on a heterogeneous architecture, wherein many initial iterations are run on the approximate core and the rest on the accurate core to achieve a reduction in energy consumption.
- The proposed accelerator design does not increase the number of iterations (that are necessary for the conventional accurate counterpart) and provides sufficient precision to converge to an acceptable solution.

The proposed heterogeneous architecture shows 23.4% of the reduction in energy consumption for radio astronomy calibration processing.

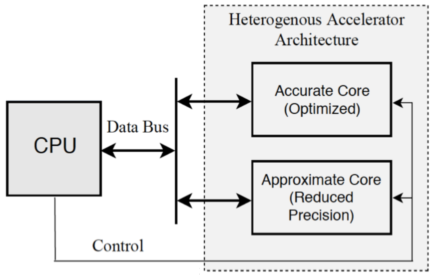

Figure 2. Our design methodology for iterative algorithms enables initial iterations to be processed on an approximate core (while the rest on an accurate core) to achieve an overall energy-efficiency.

Self-healing Methodology for Accumulation based Algorithms

Unlike the conventional approximate design methodology, the proposed self-healing [5] methodology provides the following benefits,

- Increased approximation space
- Error cancellation for the intermediate computing stage
- A more effective quality-efficiency trade-off

Internal self-healing

We further propose an internal self-healing methodology [6] that allows exploiting self-healing within a computing element, internally, without requiring a parallel module. This extends the applicability of self-healing methodology to irregular datapaths.

Self-healing methodology demonstrates up to 25% and 18.6% better area and power efficiency, respectively, as compared to the conventional methodology.

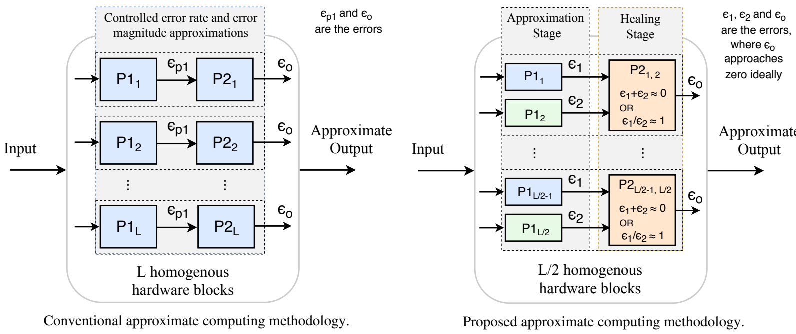

Figure 3. The proposed **self-healing** methodology [5] does not restrict the approximations based on an error profile but provides the opportunity for **error cancellation** to achieve an effective quality-efficiency trade-off.

References (Publications List)

[1] Thesis: https://doi.org/10.3990/1.9789036550116
[2] Survey: https://doi.org/10.1145/3394898
[3] CF'17: https://doi.org/10.1145/3075564.3078891
[4] CF'19: https://doi.org/10.1145/3310273.3323161
[5] SquASH: 10.1109/ACCESS.2018.2868036
[6] MACISH: 10.1109/ACCESS.2019.2920335
[7] Review: https://doi.org/10.1145/3310273.3323427

Conclusions

This research has contributed towards effective *error resilience analysis* and *energy-efficient accelerator design* for iterative algorithms like radio astronomy calibration algorithm. Moreover, for accumulation based algorithms like multiply-accumulate, our *self-healing* and *internal self-healing* methodologies provide a more effective quality-efficiency trade-off as compared to the state-of-the-art approximate computing methodology.

Acknowledgements: The author thanks Dr. ir. André Kokkeler for his invaluable supervision during this Ph.D. work. This research has been conducted in the context of the ASTRON and IBM joint project, DOME, funded by Netherlands Organization for Scientific Research (NWO), the Dutch Ministry of EL&I, and the Province of Drenthe.

UNIVERSITY OF TWENTE.