Analog Hardware Acceleration for AI Inference

Hardware accelleration of deep learning inference through analog matrix multiplication

Students

Gian-Luca Brazerol

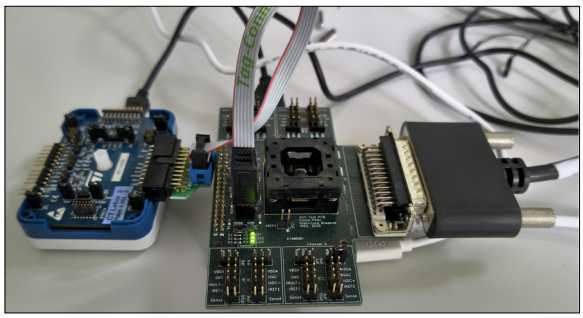
Flavio Peter

Initial Situation: The boom of DNNs has revolutionized the entire industry, driving unprecedented advancements. However, integrating DNNs into industrial applications is hindered by significant challenges relating to high energy consumption and processing delays caused by the load/store architecture of µCs. Up to 80% of clock-cycles in a µC are used to move data in and out of registers, while calculations are performed in only 20% of the time. Furthermore, digital multiplications are computationally expensive, the power consumption and latency increase with millions of such operations. Thus, a custom ASIC was designed in the previous work, implementing the concept of Near Memory Computing (NMC) in a single chip. This work builds on the previous investigation, in which the main functionalities of the ASIC were designed. This work focuses on layout and testing of the ASIC.

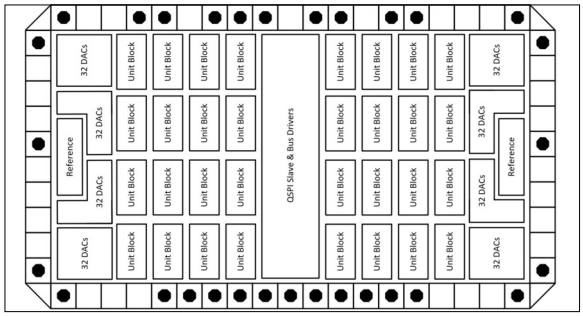
Approach: The chip design and layout were finalized and sent to the fabrication plant for production. During this time, a package for the finished ASIC was selected, along with a corresponding socket for a test PCB, which was also developed. Test cases and setups with suitable equipment were specified to characterize the chip once it was produced. Once the entire setup was running, the functionalities of the ASIC were tested, characterized and compared with the expected figures.

Result: Initial diode tests conducted on I/O pads were successful, but uncovered some minor design flaws. Despite the problems encountered, the write and read operation of the SRAM could be validated as well as the setting registers. The reference current of the bias circuit was measured and behaved as expected. Tests using the analog matrix multiplication were successfully conducted, indicating that all analog blocks worked as intended. This was validated by testing the dedicated test structures within the pad ring. The computing performance achieved was 1.07TOPS/A or 323GOPS/W with a core voltage of 3.3V. The maximum supported clocking speed achieved was 21.25MHz, which was more than twice as fast than the simulated 10MHz. Thus, a figure of merit of 50.2GOPS/A/MHz could be achieved.

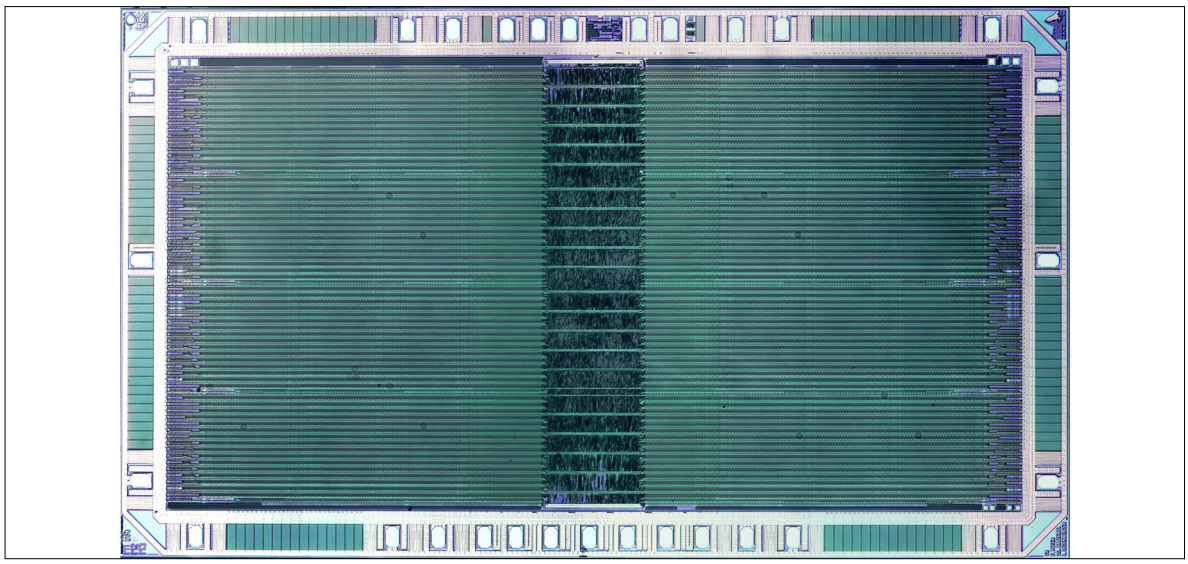
Test Setup: Left, the programmer for the µC; In the center, the test PCB with the ASIC; Right, the connector to the SMU
Own presentment



Chip Floor Plan
Own presentment



Chipfoto taken under a 10x magnifying microscope
Own presentment



Advisor
Lars Kamm

Subject Area
Electrical Engineering