Analog Hardware Acceleration for AI Inference

Hardware Accelleration of Deep Learning Inference through Analog Matrix Multiplication

Students

Gian-Luca Brazerol

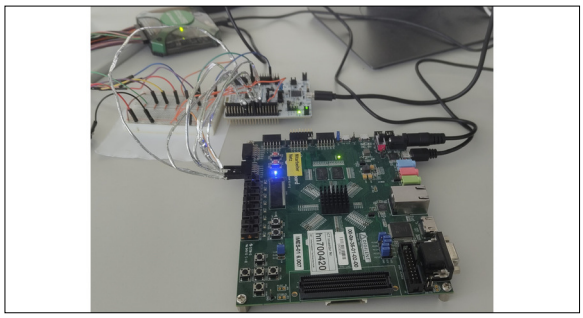
Flavio Peter

Introduction: The boom of Deep Neural Networks (DNNs) has revolutionized the entire industry, driving unprecedented advancements and innovations. However, the integration of DNNs into industrial applications is hindered by significant challenges related to high energy consumption and processing delays created by the load/store architecture of microcontrollers. Up to 80% of clock-cycles in a microcontroller are used to move data in and out of registers and a calculation is only computed in 20% of the time. This research addresses these limitations through Near Memory Computing (NMC). In contrast to In Memory Computing (IMC), where RAM cells are altered to perform calculations, this NMC implementation leverages analog computation which offers the promise of substantial reductions in both energy usage and inference time, thus improving the efficiency and scalability of DNNs. In contrast to traditional IMC which uses 1-bit operations, this NMC solution uses multibit analog operations. A fundamental operation in deep learning inference and other fields such as computer vision or digital signal processing, is the multiplication of a vector and a matrix. This operation involves numerous Multiply-Accumulate (MAC) operations to compute the resulting vector.

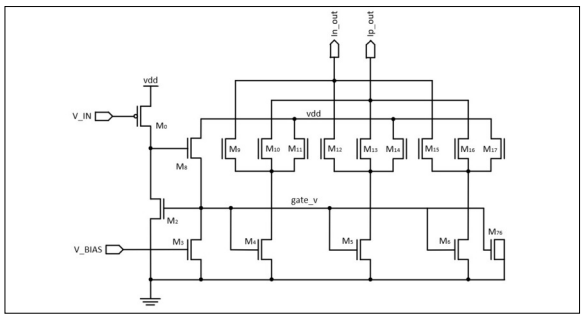
Approach: To reduce costly data transfers between memory and the processing unit, this approach stores layer weights in SRAM blocks adjacent to an analog computation block, eliminating the need to move data from memory to the processing unit. The analog block, connected directly to the SRAM, performs the MAC operation in the analog domain by multiplying an input current using weighted current mirrors and implementing addition via Kirchhoff's first law. Weights and input data are constrained to four bits, with the input vectors converted to currents using DACs. These currents are fed into a 32x32 matrix of analog multipliers, which multiply these currents with the digital weights. The weights are set once before inference, with output currents converted back to the digital domain via six bit SAR ADCs. Four bit precision for input, weight and output values suffices for most DNNs, offering a balance between accuracy and efficiency. Digital simulations with the MNIST dataset achieved 95% accuracy using four bit quantized weights and inputs. High-resolution calculations consume excessive space, power, and cost. The four bit quantization effectively preserves necessary DNN resolution while minimizing chip area.

Result: To solve this issue, we designed an Application Specific Integrated Circuit (ASIC) using the standard 350nm kit from X-FAB. Due to die size limitations and the large technology node, we could only realize a matrix containing 32x32 analog multipliers. For larger matrices, iterative calculations are necessary. However, this approach is feasible due to the linear properties of matrix multiplication. Intermediate results are stored digitally and accumulated using multibit adders. This hampers the performance massively, as with every iteration, the DAC and ADC are necessary instead of pure analog calculation. Without these limitations, the analog MAC configuration achieved a performance of 5.13 TOPS/A or 1.56 TOPS/W with a core voltage of 3.3V. When normalized with the maximum clock speed of 16MHz supported by this technology node, the Figure of Merit is 321GOPS/A/MHz, which can compete with modern IMC implementations. The entire chip area was estimated at 12.5mm^2.

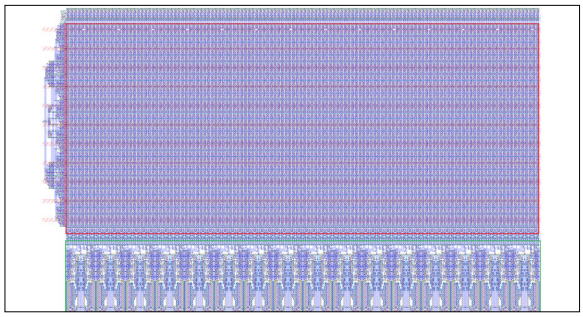
Test Setup with the QSPI Slave and Address Decoder programmed on the FPGA. The Microcontroler acts as QSPI Master.
Own presentment



Schematic of a single four bit analog multiplier
Own presentment



Chip Layout of an SRAM unit Block with 32 analog multipliers at the bottom
Own presentment



Advisor
Lars Kamm

Subject Area
Data Science, Electrical Engineering