**An Energy Efficient Multiply Add Core (MAC) for Hardware Accelerators using 32nm Finfet Technology**

**ABSTRACT:**

Artificial intelligence (AI) is the process of human thought that can be mechanised, is just an assumption. AI is all about statistical learning of low dimensional structures from high dimensional data. Actual perspective on the second wave of AI which is short Deep Neural Networks (DNN) and deep learning. This paper go over custom chip implementations of DNNs from the more recent publications. In the charge domain, analog computation will have a role in future AI systems. The work has been configure the architecture, implementation and measurements from a mixed-signal. Using array capacitors and charge redistribution, the configuration performs the multiplication operation in charge domain at the thermal noise limit with minimum energy dissipation. The charge redistribution multiplier core is fabricated in a 32nm FinFET CMOS process, with measured 1.4 fJ for the analog multiplication operation. This design compares with digital implementation with same technology and achieves the performance at 37%less energy.

**Keywords**—Analog multiplier, Deep Neural Networks, Hardware AI, LT Spice

**1.Introduction**

Over the last few years engineers realize to build super computers that can match the parallel thinking of brain, this emerged as an alternative approach to AI. Today with the great success on Deep learning and Deep convolutional network supported by theoretical advances and partial understanding about the deep neural network In the early 90’s with foundry FinFET CMOS technologies at 32nm, 16nm, 10nm and 7nm,analog neural network have not obtain improvement of area density in the field of Digital CMOS transistor that match the human brain functions. As we know that analog transistor match the biophysics of neurons and digital transistor must perform arithmetic equivalent of the behaviour of neurons. In the energy efficiency perspective, digital circuits show low precision computation hence we use mixed signal circuits therefore we use charge- based multiplier from the architecture of the Successive Approximation Register (SAR) analog to digital converters. The successive approximation principle and charge re-distribution employed in the multiply-accumulate core is based on the widely used Successive Approximation Analog to Digital Converter (SA ADC)

SA ADC is used widely because of its high-speed low –power data conversion. This SA architecture consists of capacitor array comparator and digital decoding logic designed for high resolution and high speed implemented in small area. SAR logic is used convert data from by performing binary search which leads to digital output representing analog input.

Section 2: Overview of the Literature work for charged based multiply add core in FinFET CMOS, Section 3: Overall methodology of the system, Section 4: Expected result of the system. Section 5: Conclusion for the proposed methodology.

**2.Litrature Work**

There are several works developed for charge based multiply add core in FINFET CMOS. In this section a broad overview of the literature is presented, starting with paper

“A Charge-Based Architecture for Energy-Efficient Vector-Vector Multiplication in 65nm CMOS” by K. Sanni, T. Figliolia, G. Tognetti, P. O. Pouliquen, and A. G. Andreou,[1]. In this work, an energy-efficient architecture for computing inner products/vector-vector multiplications is proposed and fabricated in a 65nm CMOS process. Exploiting the advantages of charge-based computing, this architecture computes inner products in the analog domain as charge at the kTC thermal noise limit minimizing energy cost.

“8-bit,16input,3.2pJ/op switched capacitor dot product circuit in 28-nm FDSOI CMOS”by D. Bankman and B. Murmann.In this work it represented a switched-capacitor dot product circuit capable of performing arithmetic operations in an example neural network application. The proposed solution can be viewed as a more efficient drop-in replacement for a digital static CMOS combinational dot product block in a massively parallel neural network ASIC.[2]

“Eyeriss: An Energy Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,”by Y.-H. Chen, T. Krishna,J. S. Emer and V. Sze. This paper says that The overall low-frequency behavior of an MOS transistor, in all possible bias situations and from very low to high current, can be captured in a simple model based on just three independent parameters.[3]

“A low-power analog CMOS vector quantizer,”by T. Tuttle, S. Fallahi and A. A. Abidi. This paper says that An 8-bit, 16 input, switched-capacitor dot product circuit in 28-nm FDSOI CMOS is presented. The design uses sixteen 8- bit passive charge redistribution digital-to-analog multipliers followed by an 8-bit SAR ADC. Measured energy per dot product operation is 3.2 pJ.[4]

“Characterization of a pseudo-DRAM Crossbar Computational Memory Array in 55nm CMOS”by G. Tognetti, J. Sengupta, P. O. Pouliquen, and A. G. Andreou. This paper presents a power-efficient 80 MS/s, 11-bit ENOB ADC. It is realized in 28 nm CMOS and is based on two interleaved pipelined SAR ADCs. It includes an on-chip reference generator [5].

“A switched capacitor implementation of the generalized linear integrate-and-fire neuron,” by F. O. Folowosele, A. Harrison, A.S.Cassidy,A.G.Andreou,R.EtienneCummings,S.Mihalas,E.Niebur,and T. Hamilton. This paper presents the circuits and simulation results for a silicon neuron which is based on a modified version of the Mihalas-Niebur neural model. This silicon neuron produces 15 of the 20 known neural spiking and bursting behaviours. It has low complexity and reliable matching and can thus be easily integrated into more complex neuromorphic systems. Implemented in a 0.15 mum 1.5 V CMOS process, each neuron consumes about 7.5 nW of power at 1 kHz and occupies an area of 70 mum by 70 mum. [6]

“Charge-domain integrated circuits for signal processing” by T. L. Vogelsong, J. J. Tiemann, and A. J. Steckl. This paper presents all signal-processing operations are accomplished by splitting, routing and combining charge packets, thus overcoming many of the limitations of alternative devices such as charge-coupled device (CCD) split-electrode transversal filters and switched capacitor filters.[7]

“All-MOS charge redistribution analog to-digital conversion techniques - Part I” by J. L. McCreary and P. R. Gray. In this paper A new, all MOS A/D conversion technique has been demonstrated which, with the addition of an external reference, can be used to realize a standard-process A/D converter on a single chip. Experimental data were presented which indicated that 8-bit resolutions can be attained at very high yield and low cost using standard N-channel MOS technology, and that 10- bit resolution can be achieved at somewhat lower yield. It is believed by the authors that more careful control of photolithographic processing would result in very high yield at the 10-bit level and significant yield at even higher resolutions [8].

 We found out that there is a need for designing and simulating Charge-based Successive Approximation MAC operation in 32 nm FinFET CMOS technology using LT Spice software.

**3.Methodology**

Analog transistors where employed to emulate the biophysics of neurons and perform analog multiplication and other non-linear computations an energy efficiency perspective, analog computation and mixed signal circuits are a viable alternative to digital for low precision computations. The following figure (1) shows the flow diagram of the system



**Figure (1) Flow diagram of the system**

**3.1 Design of Successive Approximation Multiply-ADD core**

Charge redistribution SA ADC, which comprises of an analog switched capacitor array, a comparator, and some digital decoding logic, can be designed for both high resolution and high-speed while being implemented in relatively small area. Using a successive approximation register (SAR), the SA ADC achieves efficient data conversion by performing a binary search through all possible quantization levels to converge to the correct digital output representing the analog input.SA ADCs have been designed in silicon-on-sapphire (SOS) CMOS, and in 32nm FinFET CMOS.

The proposed mixed-signal Multiply-ADD (MADD) architecture, computes a fixed-point multiply-add operation as

 y = wx + c;

****

**3.2 Simulating the Design**

First, in the Compute phase, a charge equivalent to the sum of the weighted partial product bits is injected into the array of capacitors. The top plates of all capacitors are connected to a common ground node, while the bottom plates are isolated and connected to either a power supply with +V volts or the common ground node depending on the partial product bits. After this first phase, the charge QY is translated into a voltage VY in the Redistribution phase. In this phase, charge is redistributed evenly across all the capacitors in the array by disconnecting the top plates of the capacitors to ground, and connecting all the bottom plates of the capacitors to ground.

Finally, this voltage representing the output of computation is converted back into a digital value in the Decode phase. Using a digital control word si , the capacitor array is reconfigured from a parallel configuration to a series-parallel configuration in order to perform the binary search to decode the voltage VY into the digital word. In the series-parallel configuration, the VY is voltage divided according to each bit position in the digital word, and compared to the common ground node indicating the magnitude of that bit for the decoded result.

****

**3.3 Analysis and Comparing**

In order to compare the computational efficiency of this design to that of conventional architectures, circuit simulations along with an energy analysis is done. The mixed-signal SA architecture was constructed to compute 8-bit multiply-add operations with 5-bit signed weights and inputs, and an 8-bit offset. The capacitor array was designed with unit capacitors in the order of 4fF, with an input voltage bias of 50mV. We can see the simulation of each component of this system. (figure 2) shows the AND gate schematic and its logic, (figure 3) shows the comparator schematic and its logic, (figure 4 a and b) shows the Capacitor array schematic and its logic. These are designed for both high resolution and high-speed while being implemented in relatively small area. A successive approximation register (SAR), the SA ADC achieves efficient data conversion by performing a binary search through all possible quantization levels to converge to the correct digital output representing the analog input



Figure (2) AND gate schematic and its logic



Figure (3) Comparator schematic and its logic



Figure (4a) Capacitor array schematic



Figure (4b) capacitor array logic



Figure 5 SA Architecture for Multiply-Add Operations

**4. EXPECTED RESULT**

The SPICE models of both implemented designs (charge based and traditional) were simulated with different supply voltages and clock rates in order to measure performance and efficiency. It is done using predictive technology model (ptm) in 32nm low power FinFET technology. With a nominal process supply voltage of 0.8V, the power supply is swept from 0.4-0.8V for both designs, and the total energy from these simulations were measured.

We are expecting a power efficiency of around 37% compared to a conventional digital processor for the same technology. Since it’s a simulation the values may vary with the actual implementation as the parasitic capacitance values are neglected here.

**5.Conclusion**

Analog transistors where employed to emulate the biophysics of neurons and perform analog multiplication and other non-linear computations an energy efficiency perspective, analog computation and mixed signal circuits are a viable alternative to digital for low precision computations.

Thus, from the energy efficiency perspective analog processing seems a better alternative compared to its digital counterpart. In this project we are designing and simulating 8-bit Charge-based Successive Approximation MAC operation in 32 nm FinFET CMOS technology using LT Spice software.

We are expecting a power efficiency of around 37% compared to a conventional digital processor for the same technology.

**6.References**

[1] “ A Charge-Based Architecture for Energy-EfficientVector-Vector Multiplication in 65nm CMOS” K. Sanni, T. Figliolia, G. Tognetti, P. O. Pouliquen, and A.G.Andreou, 2018 IEEE International Symposium on Circuits and Systems.

[2] “An 8-bit, 16 input, 3.2 pJ/op switchedcapacitor dot product circuit in 28-nm FDSOI CMOS” D. Bankman and B. Murmann, in 2016 IEEE Asian Solid-State Circuits Conference (A-SSCC). IEEE, Nov. 2016, pp.21–24.

[3]“Eyeriss: An Energy Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” Y.-H. Chen, T. Krishna,J. S. Emer andV. Sze,2017

[4] “A low-power analog CMOS vector quantizer,” T. Tuttle,S. Fallahi andA. A. Abidi 1993.

[5] “Characterization of a pseudo-DRAM Crossbar Computational Memory Array in 55nm CMOS,” G. Tognetti, J. Sengupta,P. O. Pouliquen, and A. G. Andreou ,Mar 2019

[6] “A switched capacitor implementation of the generalized linear integrate-and-fire neuron” F. O. Folowosele, A. Harrison, A. S. Cassidy, A. G. Andreou, R. EtienneCummings, S. Mihalas, E. Niebur, and T. Hamilton,2009

[7] “Charge-domain integrated circuits for signal processing” T. L. Vogelsong, J. J. Tiemann, and A. J. Steckl

[8] “All-MOS charge redistribution analogto-digital conversion techniques - Part I” J. L. McCreary and P. R. Gray, IEEE Journal of Solid-State Circuits, vol. 10, no. 6, pp. 371–379, Dec. 1975.

[9] Tong, X. Y., Zhang, W. P., & Li, F. X. (2014). Low-energy and area-efficient switching scheme for SAR A/D converter. Analog Integrated Circuits and Signal Processing, 80(1), 153–157.

[10] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, “A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute,” IEEE Journal of Solid-State Circuits, vol. 54, no. 6, pp.1789–1799, Jun. 2019.

[11] C. Mayr, J. Partzsch, M. Noack, S. H¨anzsche, S. Scholze, S. H¨oppner,G. Ellguth, and R. Schueffny, “A Biological-Realtime Neuromorphic System in 28 nm CMOS using Low-Leakage Switched Capacitor Circuits,” arXiv.org, p. arXiv:1412.3233, Dec. 2014.

[12] E. A. Vittoz, “Analog VLSI signal processing: Why, where, and how?”The Journal of VLSI Signal Processing, vol. 8, no. 1, pp. 27–44, Feb.1994.

[13] U. Ramacher and U. Ruckert, VLSI Design of Neural Networks. KluwerAcademic Publishers, 1991.

[14] U. Ramacher, “SYNAPSE - a Neurocomputer That Synthesizes Neural

Algorithms on a Parallel Systolic Engine,” Journal of Parallel and Distributed Processing, vol. 14, no. 3, pp. 306–318, Mar. 1992.

[15] S. Nakamura and Y. Nagazumi, “A matched filter design by chargedomain

operations,” IEEE Transactions on Circuits and Systems I Fundamental Theory and Applications, vol. 52, no. 5, pp. 867–874,

May 2005.

[16] D. W. Hammerstrom, “A VLSI architecture for high-performance, lowcost,

on-chip learning,” in 1990 International Joint Conference on Neural Networks, Feb. 1990, pp. 537–544.