JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IMPLEMENTATION OF FAULT TOLERANT SYSTOLIC ARRAY MATRIX MULTIPLIER AS PROCESS ELEMENT BLOCKs FOR HIGH-SPEED COMPUTATIONS

Nagamalli. Arasavalli[1], Balaji.Abburi[2], Manideep.P[3], Teja. T[4], Nasiha Farheen.SK[5], Kavya.S[6]

[1]*Associate Professor,* [2,3,4,5,6] *Student,* Department of Electronics and Communication Engineering,

PBR Visvodaya Institute of Technology & Science, SPSR Nellore, Kavali, Andhra Pradesh, India 524201

ABSTRACT

In the era of Big data, Internet of Things (IoT), Industry 4.0 and 5G communication based applications where high speed computations with minimal or reduced faults and errors are intended. Digital signal processing (DSP) multiplier and accumulator (MAC) structures are the core building blocks in which integer or float point multiplications from low to high complexity computations at higher speeds are expected in this Digital era. Systolic Array Matrix Multiplier (SAMM) is implemented in which Fault Tolerance can be expected such that it can detect logic level faults and timing errors due to intermittent clocks at transistor level. The SAMM process element (PE) architectures using Verilog HDL in Xilinx Vivado are implemented results with reduced error less MAC units at higher complex operations with reduced voltages, detecting errors on-the-fly to avoid energy demanding round-trips over the complex computations are achieved.

Keywords:, fault detection, systolic array, matrix multiplication, hardware and Network architectures.

I. INTRODUCTION

With the MOS technology revolution, supply voltages are scaled down for CPUs, GPUs, and FPGAs by about 12%, 20%, and even 30% respectively[1]. Practically, an increment of up to 8x using voltage regimes close(V_{th}) is observed [2],[8],[9]. The lack of confidence prevents manufacturers from using aggressive voltage reduction strategies to make use of near-threshold and sub-threshold areas[3]. To deploy and coordinate "intelligence" in the edge computing resources, data transmission speed and latency limits must disappear. Among the 5G technology enablers that are computationally taxing and difficult to implement in an energy-efficient huge mu-MIMO (multi-user Multiple-Input Multiple-Output) radios and neural inference, due to its low latency from the sensor and actuator nodes with the ultra-densification of wireless infrastructures[4]. In the hardware FPGA implementations without occurrence of a single defect or performance loss, an energy saving of up to 60% is observed. In this paper, a technique for the algorithm-dependent identification of mistakes from forceful voltage decreases is proposed. Similar methods, like Error-Correction-Code (ECC) algorithms, can be useful in the identification and correction of memory faults. The identification of

errors in data and control routes, however, is not possible with them [5].

Although they offer error-resilience, traditional fault tolerance techniques like Triple-Module-Redundancy (TMR) or Double-Module-Redundancy (DMR) and their various variations significantly increase gate activity and gate count. In contrast to this, our strategy relies on fault-tolerance to achieve energy efficiency. This results in fault-resiliency at the expense of energy efficiency [6]. T-FLOPS on 484 processors returns a correct result while one process failure has happened. This represents 65% of the machine peak efficiency and less than 12% overhead with respect to the fastest failure-free implementation. We predict (and have observed) that, as we increase the processor count, the overhead of the fault tolerance drops significantly [7]. The region of operation below(V_{th})zones have the potential to increase energy efficiency by 10x to 20 times[8] , It might be difficult due to decreasing the supply voltage without optimizing the clock frequency can lead to a reduced in reliability, clock performance and larger time and cost during the development process.

In the hardware FPGA implementations without occurrence of a single defect or performance loss, an energy saving of up to 60% is observed[9].

In the digital logic designs the circuits are operated close to and below transistor threshold voltage(V_{th})was proposed[10].

In the earlier architectures the focus was on reduction of lengths of carry propagation in the final addition and the accumulation. by integrating a part of additions into the partial product reduction process. MAC unit is composed of two individual blocks: a multiplier and an accumulator. An N-bit MAC unit includes an N-bit multiplier and a (2N+α-1)-bit accumulator (adder), where α is the number of guard bits used to avoid overflow. A Multiplier has the following three sequence of operations as shown

- The first step is the partial product generation (PPG) process.
- The second step is the partial product reduction (PPR) process.
- The third step is the final addition.

II. Implementation of PE-MAC

By the variations in the voltage and frequency logical faults occur at the computing logic's data output. Matrix operations, which constitute the most energy-intensive calculations such as DSP high complex computations and convolutional and deep neural networks in which large data sets are to be handled.

Systolic array is one of the most energy-efficient and high-performance designs for matrix computations. Systolic designs are being used in Google Tensor-Processor-Units (TPU) to speed up neural network computations. In this work, presenting and testing the performance of multiplier solution for matrix multiplication focused systolic array error detection.

By co-simulating transistor and system level components and implementing a systolic matrix multiplier in an FPGA, With minimal development overheads, the solution makes it possible to use decreased voltage working areas. In the earlier architectures, Lack of performance verification, error detection Less energy performance efficiency are the key issues to overcome in the proposed SAMM PE architectures.

This makes it possible to identify computational problems by looking at the finished product, although rectification is only possible in a small number of situations if hardware support has been included into the design. The corrective method involves recalculating the change in voltage or clock frequency is suggested.

Assuming, [A] is an $_{N\times N}$ matrix, then a row checksum matrix [A^r] is defined as N × (N + 1) matrix, as below

$$A^r = [A \quad Ae^T] \qquad (1)$$

Where, e is column vector e_N = [1, 1,..., 1], the n^{th} element of the column vector A_e contains the sum of elements of the corresponding row of matrix A, i.e.,

$$a^r_{n,(N+1)} = \Sigma a_{n,I} \qquad (2)$$

Similarly, column checksum, [A^c] and full checksum, [A^f] matrices are as follows in equation (3).

$$A^C = \begin{bmatrix} A \\ eA \end{bmatrix} \text{ and } A^f = \begin{bmatrix} A & Ae^T \\ eA & eAe^T \end{bmatrix} \qquad (3)$$

The row vector e^TA denotes the sums of the elements in the columns of matrix A, and e^TAe is the sum of all elements in matrix A, a scalar value. The checksums can be used to detect errors instead by multiplying matrices $A_{N\times N}$ and $B_{N\times N}$

$$C = A \times B \qquad (4)$$

we multiply the column checksum matrix Ac and row check-sum matrix B^c to obtain full checksum matrix C

$$C^f = A^c \times B^r = \begin{bmatrix} A \\ eA \end{bmatrix} [B \quad Be^T] = \begin{bmatrix} AB & ABe^T \\ eAB & eABe^T \end{bmatrix} \quad (5)$$

The matrix multiplication of the 2-D structured systolic array SAMM-PE logic depicted. The array is made up of a grid of identical Processing Elements (PE), each of which performs multiply-accumulate (MAC) operations on data received from the next-door PEs on the top and left before passing the output or input data to the next-door PEs on the right and below. Only matrix B has to be increased since the output's row checksum is already being checked by the circuitry. When the matrix B elements enter the array as seen in Figure.1 done instantly.

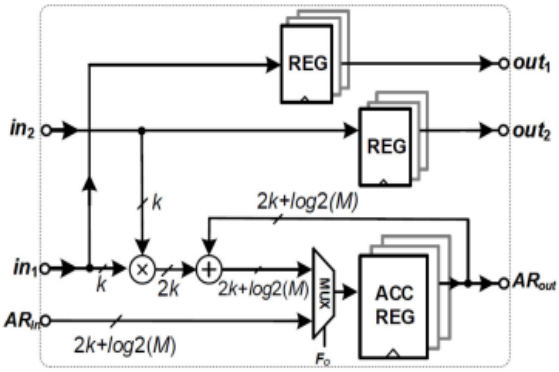

Fig. 1: A systolic array of processing element (k and M are data lengths)

The design is unaffected by augmentation, except for a two-clock cycle delay between checksum computation and inspection. As the result matrix is clock-out from the array, faults are immediately discovered, which is a benefit of the approach. Following examination, the checksums may be discarded. When targeting transient mistakes in low voltage situations rather than addressing persistent defects, the additional cost from columns for checksum and error checking is simply O(1/N), Feasibility of detecting errors, with efficiency in performance verification and High efficiency in energy performance is achieved.

III. Simulation Results

RTL Schematic of designed MAC processing element is extracted as shown in the figure (2), by implementing the design in Verilog simulated and synthesis reports are extracted and analysed . the synthesized results are represented as shown in table [1] with respect to time, frequency as performance parameters. RTL simulation results of PE are shown in figure (3) and the technology schematic of a PE as shown in figure (4) respectively.

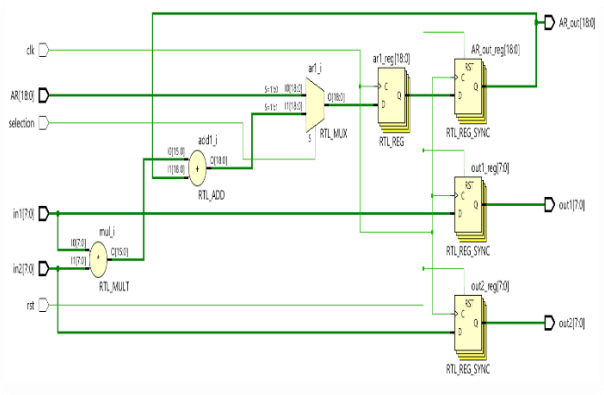

Figure. [2] RTL Schematic PE block

Figure [3] simulation results of Single PE

Table [1] Timing analysis of a PE Block.

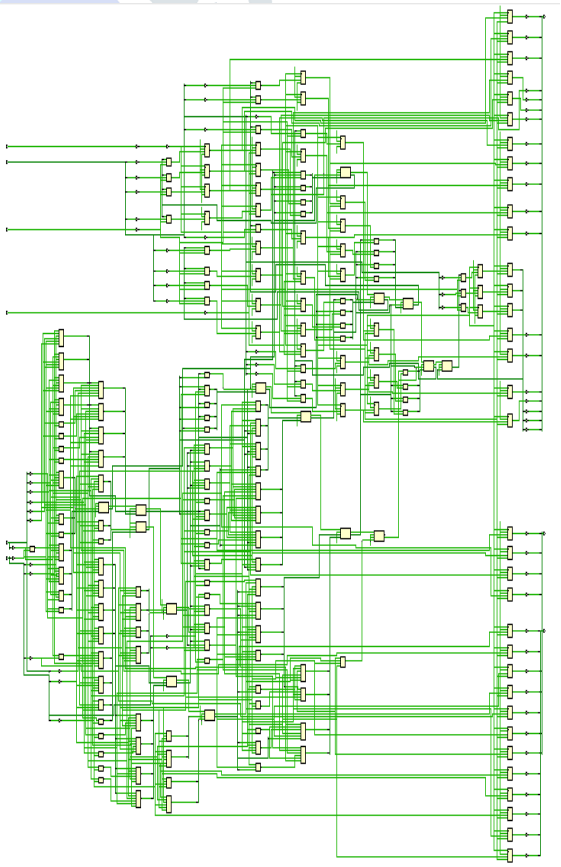

Figure. [4] Technology schematic of a PE block.

Cell: in->out	fanout	Gate delay	Net delay	LOGICAL/ NET name
RAM4X1D: DPRA5 -> DPO	1	0.043	0.279	Mram_B31 (count[2]_read_port_28_OUT<0>)
SRLC16E: D		0.005		n87/Mshreg_out2_0
FD: C -> Q	19	0.232	0.363	count_FSM_FFd1 (count_FSM_FFd1)
Total		1.050ns (0.408ns logic, 0.642ns route) (38.9% logic, 61.1% route)		

Timing analysis
All value displayed in nanoseconds (ns)
Timing constraint: Default period analysis for Clock 'clk'
Clock period: 1.050 ns (frequency: 952.472MHz)
Total number of paths / destination ports: 129 / 67
Delay: 1.050 ns (Levels of Logic: 1)
Source: count FSM_FFd1 (FF)
Destination: n87/Mshreg_out2_0 (FF)
Source Clock: clk rising
Destination Clock: clk rising
Data Path: count_FSM_ FFd1 to n87/Mshreg_out2_0

CONCLUSION

The proposed PE system works under the clock frequency of 952.38 MHz whereas the clock frequency of existing system is 400MHz. On comparing the clock frequency of the existing system with the proposed system, the proposed PE gets improved by 55% highest performance in terms of frequency. Hence, the speed of the system gets improved, and the delay gets reduced from 2.5nsec to 1.05nS is observed as shown in the table.(1). Hence the manufacturers, without performance compromises, by using efficient near-threshold operation points can be reached when the clock rate is adjusted as well with reduced
error rate, the scheme fits a wide field of applications from 5G wireless communications to artificial intelligence. As a result it is concluded that in the best future work, utilization of the proposed solution for approximate high speed computing architectures with best performance is preferred.

REFERENCES

[1] G. Papadimitriou et al., "Exceeding conservative limits: A consolidated analysis on modern hardware margins," IEEE Trans. Device Mater. Rel., vol. 20, no. 2, pp. 341–350, Jun. 2020.

[2] M. Hiienkari et al., "A 0.4-0.9V, 2.87pJ/cycle near-threshold ARM Cortex-M3 CPU with in-situ
monitoring and adaptive-logic scan," in Proc. IEEE Symp. Low-Power High-Speed Chips (COOL CHIPS), 2020, pp. 1–3., April 2020.

[3] "Dynamic margining: The minima approach to near-threshold design," Minima Processor Oy,Oulu, Finland, Rep., Nov. 2017. [Online]. Available: https://minimaprocessor.com/wp-content/uploads/2017/11/minima-margining-white-paper.pdf.

[4] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," IEEE Trans. Inf. Theory, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.

[5] H. David, C. Fallin, E. Gorbatov, U. R. Hanebutte, and O. Mutlu, "Memory power management via dynamic voltage/frequency scaling," in Proc. 8th ACM Int. Conf. Auton. Comput., 2011, pp. 31–40.

[6] G. Bosilca, R. Delmas, J. Dongarra, and J. Langou, "Algorithm-based fault tolerance applied to high performance computing," J. Parallel Distrib. Comput., vol. 69, no. 4, pp. 410–416, 2009.

[7] "Algorithmic Based Fault Tolerance Applied to High Performance Computing" George Bosilca, Jack Dongarra, Julien Langou
Available:https://www.researchgate.net/publication/1737443_Algorithmic_Based_Fault_Tolerance_Applied_to_High_Performance_Computing, July 2008.

[8] A. Wang and A. Chandrakasan, "A 180-mV subthreshold fft processor using a minimum energy design methodology," IEEE J. Solid-State Circuits, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[9] D. Ernst et al., "RAZOR: Circuit-level correction of timing errors for low-power operation," IEEE Micro, vol. 24, no. 6, pp. 10–20, Nov./Dec. 2004.

[10] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," IEEE J. Solid-State Circuits, vol. 7, no. 2, pp. 146–153, Apr. 1972.