INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue3)

Available online at: www.Ijariit.com

A REVIEW ON IEEE-754 STANDARD FLOATING POINT ARITHMETIC UNIT

Monika Maan
ECE Department
monikamaan0001@gmail.com

Abhay Bindal
ECE Department
abhaybindal@gmail.com

MAHARISHI MARKANDESHWAR UNIVERSITY, MULLANA (AMBALA)-133207, HARYANA

Abstract* - *Floating point operations in digital systems form an integral part in the design of many digital processors. Digital Signal Processor is the most important application of floating point operations. In the recent years many approaches for floating point operations have been proposed and their merits and demerits are compared. For floating point operations the operands are first converted into IEEE 754 format in either single precision or double precision format. The arithmetic operations are performed on the significant part of the IEEE format. In this paper various floating point operation unit architectures are reviewed. Few designers work on high speed architectures for reducing the delay of the overall circuit while others work on the area utilization parameters. Then the conclusion is drawn based on various architectural analyses.

Keywords— Floating Point Unit, FPGA, IEEE 754.

I. INTRODUCTION

In past decades, there were several of difficulties and issues occurred in the development of conventional computing technologies. The main difficulty of the conventional computing technologies is power dissipation and it a significant problem in today's computer chip [1]. The advancement in VLSI designs especially in portable device technologies lead to faster, smaller and more complex electronic system design. In VLSI design, the conventional logic circuits dissipate more power. Reversible computing is one of the methods used in low power dissipating circuit design which presents a method for constructing computers that produce no heat dissipation [2]. This design model is used in various current technologies such as low power CMOS design, Nanotechnology, quantum computing etc. it also helps to prevent the loss on data and reduces power dissipation. Reversible logic can eliminate the heat dissipation due to information loss [3]. This is due to the amount of energy dissipated in a system which makes a direct relationship to the number of bits erased during computation. The difference between the reversible circuits and conventional logic circuits is that the reversible circuits are built from reversible logic gates [4]. The basic principle of reversible computing is that device with an identical number of input and output lines will produce a computing environment where the electrodynamics of the system allow for prediction of all future states based on known past states, and the system reaches every possible state, resulting in no heat dissipation. Reversible logic is a promising field of research that finds applications in low power computing, quantum computing, optical computing, and other emerging computing technologies [5].

Further, floating point multiplication is one of the major operations in image and digital signal processing applications. The single precision floating-point multiplier requires the design of efficient 24x24 bit integer multiplier [6]. In modern days, computer use convention for representing non integer numbers. The IEEE754 Standard for Floating-Point Arithmetic gives binary representation for floating-point numbers of varying precision. Some examples of IEEE754 Standard are single precision format (binary 32), double precision format (binary64).There are four Operations in the floating point unit.The conceptual overview of floating point unit is shown in figure

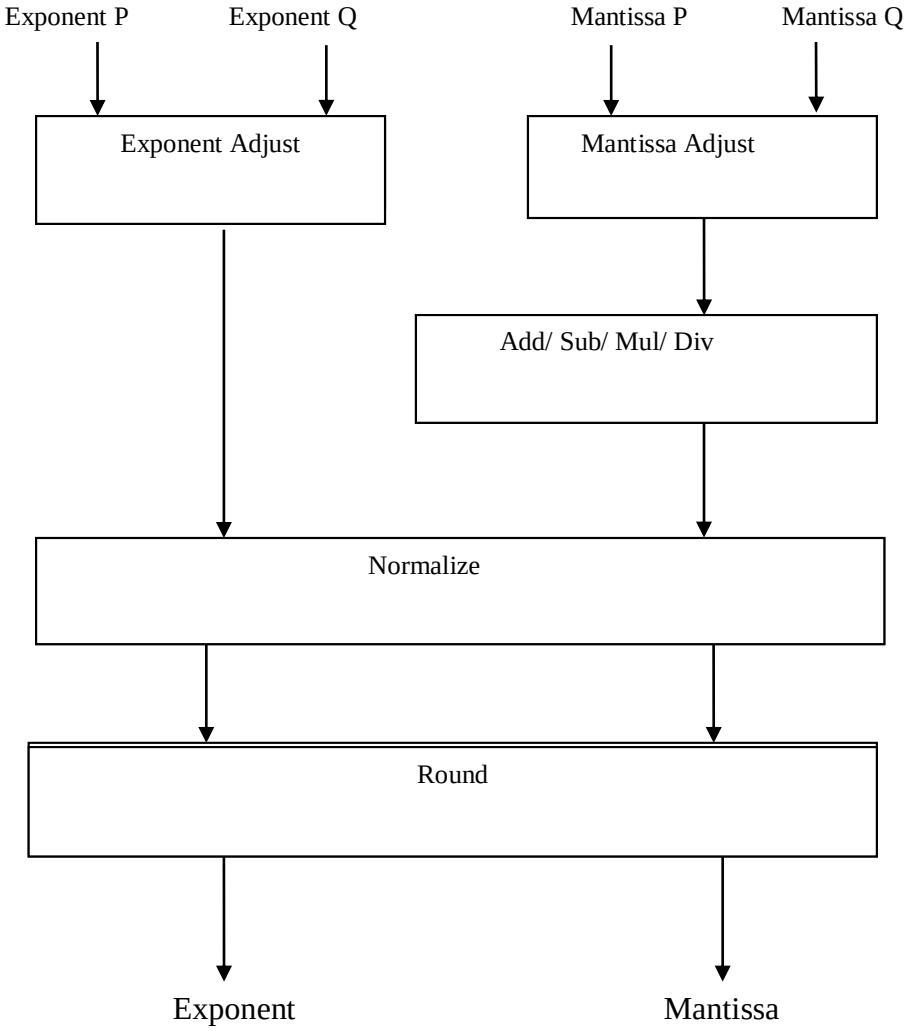

Fig1: Conceptual Overview of Floating Point Unit.

Floating-point addition is the most frequently used floating-point operation. The main components in IEEE 754 standard are sign, mantissa and exponent. Exceptions are arises during floating point operation which are gives as below:

1. Overflow exception: this exception is occurred when the result cannot be shown as definite number in precision format of the destination.
2. Underflow exception: when an intermediate result is small to be correctly calculated then underflow exception took place.
3. Division by zero exception: this exception took place when zero divides a finite nonzero number.
4. Invalid operation exception: this occurs when given inputs are not suitable for the operation to be performed.

II. LITERATURE REVIEW

This standard "IEEE Standard for Floating-Point Arithmetic, ANSI/IEEE Standard 754-2008, New York:" IEEE, Inc. [1], 2008 describes interchange and arithmetic methods and formats for binary and decimal floating-point arithmetic in computer programming environments. This standard specifies exception conditions and their default handling. An implementation of a floating-point system conforming to this standard may be recognized entirely in software, entirely in hardware, or in any combination of software and hardware. The normative part of this standard, numerical results and exceptions specified for different operations in are uniquely determined by the, sequence of operations, values of the input data and destination formats, all under user control.

Jain, Jenil, and Rahul Agrawal et al. [2] this paper presents design of high speed floating point unit using reversible logic. In recent nanotechnology, Programmable reversible logic design is trending as a prospective logic design style for implementation and quantum computing with low impact on circuit heat generation. There are various reversible implementations of logical and arithmetic units have been proposed in the existing research, but very few reversible floating-point designs has been designed. Floating-point operations are used very frequently in nearly all computing disciplines. The analysis of proposed reversible circuit can be done in terms of quantum cost, garbage outputs, constant inputs, power consumption, speed and area.

Gopal, Lenin, Mohd Mahayadin et al. [3] in the paper, eight arithmetic and four logical operations has been presented. In the proposed design 1, Peres Full Adder Gate (PFAG) is used in reversible ALU design and HNG gate is used as an adder logic circuit in the proposed ALU design 2. Both proposed designs are analyzed and compared in terms of number of gates count, garbage output, quantum cost and propagation delay. The simulation results show that the proposed reversible ALU design 2 outperforms the proposed reversible ALU design 1 and conventional ALU design.

Nachtigal, Michael, Himanshu Thapliyal et al. [4] In this work, a new reversible design of single precision floating point multiplier has been proposed based on operand decomposition approach. Furthermore, a new reversible design of the 8x8 bit Wallace tree multiplier has proposedthatis optimized in terms of quantum cost, delay, and number of garbage outputs. Wallace tree multiplication consists of three conceptual stages: Partial product generation, partial product compression using 4:2 compressors, full adders, and half adders, and then the final addition stage to generate the product. In this work we perform optimization at each of these three stages.

Dhanabal, R., Sarat Kumar Sahoo et al. [5] presents a design using reversible gates. Reversible gates namely TSG gate performs 1-bit addition with carry. This is the first reversible gate which alone can acts as full adder. Gate is used to perform logical operations like AND, OR. In this works, designing 1-bit alum has also been presented using pass transistor with virtuoso tool of cadence. Based on analysis of the result, this design using reversible gates is better than that using the irreversible gates.

Nachtigal, Michael, Himanshu Thapliyal, and Nagarajan Ranganathan [6] Floating-point operations are needed very frequently in nearly all computing disciplines, and studies have shown floating-point addition to be the most oft used floating-point operation. This paper presents for the first time a reversible floating-point adder that closely follows the IEEE754 specification for binary floating-point arithmetic. This design requires reversible designs of a controlled swap unit, a subtracter, an alignment unit, signed integer representation conversion units, an integer adder, a normalization unit, and a rounding unit.

Alaghemand, Fatemeh et al. [7] presented a reversible floating-point adder design, because the fixed-point adder is less precise in the representation of numbers. The proposed design is made up of several parts, including: Conditional swap, Alignment unit, Converter, Addition and Normalization. We attempted to improve the parameters of quantum cost, garbage outputs and constant inputs for these parts and finally compared this design with the existing designs. This proposed design has reduced 78% and 30% of the quantum cost, 78% and 26% of the garbage output and 79% and 30% of the constant input in compared with other approaches.

Kahan et al. [8] proposed a dozen commercially vital arithmetic's boasted various word sizes, precisions, misestimating procedures and over/underflow behaviours, and additional were within the works. "Portable" software system meant to reconcile that numerical diversity had become unbearably expensive to develop. 13 years past, once IEEE 754 became official, major microchip makers had already adopted it despite the challenge it exhibit to implementers. With new selflessness, hardware designers had up to its challenge within the belief that they might ease and encourage a huge burgeoning of numerical software system. They did succeed to a substantial extent. Anyway, misestimating anomalies that preoccupied all folks within the Seventies afflict solely CRAY X-MPs — J90s currently.

Ykuntam et al. [9] proposed Addition is that the heart of arithmetic unit and also the arithmetic unit is commonly the work horse of a machine circuit. Thus adders play a key role in planning Associate in Nursing arithmetic unit and additionally several digital integrated circuits. Carry choose Adder is one amongst the quickest adders employed in several information processors and in digital circuits to perform arithmetic operations. However CSLA is area-consuming as a result of it consists of twin ripple carry adder within the structure. To cut back the world of CSLA, a CSLA with Binary to Excess-1 converter is already designed that reduces the world of adder. However

there area unit different techniques to style a CSLA to cut back its space. One amongst such technique is victimization Associate in Nursing add one circuit technique. This paper proposes the planning of root CSLA victimization add one circuit with vital reduction in space.

Quinnell et al. [10] proposed several new architectures for floating-point amalgamate multiplier-adders employed in the x87 units of microprocessors. These new architectures are designed to produce solutions to the implementation issues found in modern amalgamate multiply-add units, at the same time increasing their performance and decreasing their power consumption. every new design, additionally as a group of contemporary floating-point arithmetic units used as reference styles for comparison, are styled and enforced victimization the Advanced small Devices sixty five micro-millimetre atomic number 14 on dielectric junction transistor technology logic gate design toolset. All styles use the AMD 'Barcelona' native quad-core standard-cell library as in study building block to make and distinction the new architectures in an exceedingly with-it and realistic industrial technology.

III. **CONCLUSIONS**

Floating point arithmetic is one of the most important units in modern day digital systems. Many researchers have proposed various approaches which includes the use of reversible logic in the design of adder or multiplier architecture. In one approach, Peres full adder architecture is implemented and the proposed architecture outperforms the basic conventional approach. In other approach Wallace Tree architecture is proposed and evaluated in terms of quantum cost, delay and latency. Most of the designs using the reversible logic architecture optimize the delay, latency and the area utilization. In the future approach high speed adder and multiplier architectures must be implemented and delay of the circuit must be reduced.

ACKNOWLEDGMENT

Every success stands as a testimony not only to the hardship but also to hearts behind it. Likewise, the present work has been undertaken and completed with direct and indirect help from many people and I would like to acknowledge all of them for the same.

REFERENCES

[1] "IEEE Standard for Floating-Point Arithmetic", in *IEEE Std 754-2008* , vol., no., pp.1-70, Aug. 29 2008.

[2] Jain, Jenil, and Rahul Agrawal. "Design And Development of Efficient Reversible Floating Point Arithmetic unit." In *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*, pp. 811-815. IEEE, 2015.

[3] Gopal, Lenin, Mohd Mahayadin, Nor Syahira, Adib Kabir Chowdhury, Alpha Agape Gopalai, and Ashutosh Kumar Singh. "Design and synthesis of reversible arithmetic and Logic Unit (ALU)." In *Computer, Communications, and Control Technology (I4CT), 2014 International Conference on*, pp. 289-293. IEEE, 2014.

[4] Nachtigal, Michael, Himanshu Thapliyal, and Nagarajan Ranganathan. "Design of a reversible single precision floating point multiplier based on operand decomposition." In *Nanotechnology (IEEE-NANO), 2010 10th IEEE Conference on*, pp. 233-237. IEEE, 2010.

[5] Dhanabal, R., Sarat Kumar Sahoo, V. Bharathi, V. Bhavya, Patil Ashwini Chandrakant, and K. Sarannya. "Design of Reversible Logic Based ALU." In*Proceedings of the International Conference on Soft Computing Systems*, pp. 303-313. Springer India, 2016.

[6] Nachtigal, Michael, Himanshu Thapliyal, and Nagarajan Ranganathan. "Design of a reversible floating-point adder architecture." In *Nanotechnology (IEEE-NANO), 2011 11th IEEE Conference on*, pp. 451-456. IEEE, 2011.

[7] Alaghemand, Fatemeh, and Majid Haghparast. "Designing and Improvement of a New Reversible Floating Point Adder." (2015).
[8] Kahan, William. "IEEE standard 754 for binary floating-point arithmetic."*Lecture Notes on the Status of IEEE* 754.94720-1776 (1996): 11.

[9] Ykuntam, Yamini Devi, MV Nageswara Rao, and G. R. Locharla. "Design of 32-bit Carry Select Adder with Reduced Area." *International Journal of Computer Applications* 75.2 (2013).

[10] Quinnell, Eric, Earl E. Swartzlander Jr, and Carl Lemonds. "Floating-point fused multiply-add architectures." *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on*. IEEE, 2007.

