RESEARCH ARTICLE OPEN ACCESS

# A Fast Floating Point Double Precision Implementation on Fpga

Monika Maan*, Abhay Bindal**
**(PG Scholar, Department of Electronics and Communication, Maharishi Markandeshwar University, Mullana (Ambala)-133207,Haryana.India)*
*** (Assistant Professor, Department Of Electronics And Communication, Maharishi Markandeshwar University, Mullana (Ambala)-133207,Haryana.India)*

**ABSTRACT**
In the modern day digital systems, floating point units are an important component in many signal and image processing applications. Many approaches of the floating point units have been proposed and compared with their counterparts in recent years. IEEE 754 floating point standard allows two types of precision units for floating point operations, single and double. In the proposed architecture double precision floating point unit is used and basic arithmetic operations are performed. A parallel architecture is proposed along with the high speed adder, which is shared among other operations and can perform operations independently as a separate unit. To improve the area efficiency of the unit, carry select adder is designed with the novel resource sharing technique which allows performing the operations with the minimum usage of the resources while computing the carry and sum for '0' and '1'. The design is implemented using the Xilinx Spartan 6 FPGA and the results show the 23% improvement in the speed of the designed circuit.
***Keywords***: Carry Select Adder, Floating point unit, FPGA, IEEE-754 Standard, Reversible logic gates.

## I. INTRODUCTION

An Arithmetic circuit has multiple applications in digital coprocessors ,microprocessors and in application specific circuits because they performs the digital arithmetic operations Since, there is a lot of advancements in VLSI technology, many complex algorithms that appears inoperable to put into practice, have become facilely realizable today with desired performance parameters so that new designs can be incorporated [4]. However, since the 1990s,the most popular code for representing the real numbers is called the IEEE -754 Floating-Point Standard through which the floating point operations carried out efficiently with modest storage requirements [1].

### 1.1 The IEEE-754 Floating-Point Standard

According to the IEEE-754 standard there are many important operations which must be performed in order to get the accurate results [1].An overall introduction to the floating-point standard is presented: 1) Floating-point number system, 2) Rounding modes, 3) Exceptions.

### 1.2 Floating-Point Number System

The floating-point number has the basic three components:1) Sign, 2) Exponent, and 3) Significand. Floating point numbers are presented according to the sign magnitude representation which means "0" indicates a positive number and "1" indicates a negative number.

The exponent field according to the 32 bit floating point representation is 8 bits wide. This field has an exponential form with base of 2 for binary and 10 for decimal [3]. Since it is the most commonly used format, only the binary format is covered in this paper. The exponent is biased by the half the maximum exponent so that it can represent both positive and negative exponents. The significand bits represent a fraction that is multiplied by the exponent term. The significand needs to be normalized in order to attain the form of 1.xxx, so that MSB is always "1".


**Figure 1:** IEEE-754 Floating-Point Single and Double Precision Formats.

In IEEE-754 floating point standard, the single precision format consists of 1 sign bit, 8 exponent bits and 23 significand bits. The double precision format consists of 64 bits which is divided into 1 sign bit which can be 0 or 1 as described above, 11 exponent bits and 52 significand bits.

## II. LITERATURE STUDY

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.

Kahan et al. [2] proposed a dozen commercially vital arithmetic's boasted various word sizes, precisions, misestimating procedures and over/underflow behaviors, 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.

Ykuntam et al. [3] explained 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.

Quinnell et al. [4] described 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. 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.

Bruintjes et al. [5] proposed a solution within the type of a brand new design that mixes number and floating- purpose arithmetic in an exceedingly single data path. each varieties of arithmetic area unit tightly integrated by mapping practicality to identical basic hardware elements. The advantage of such Associate in approach is two-fold. as a result of the floating-point unit are often regular for number instruction, we tend to area unit ready to cut-down on number dedicated resources creating floating-point units justify ready in an exceedingly inexpensive surroundings.

Hafiz Md. Hasan Babu et. al. [6] explained that a reversible gate has the equal number of inputs and outputs and one-to-one mappings between input vectors and output vectors; so that, the input vector states can be always uniquely reconstructed from the output vector states. This correspondence introduces a reversible full adder circuit that requires only three reversible gates and produces least number of "garbage outputs", that is two.

K.Saranya et.al. [7] proposed that Carry Select Adder (CSLA) is one of the fastest adders used in many data-processing processors to perform fast arithmetic functions. The proposed design has reduced area and power as compared with the regular SQRT CSLA with only a slight increase in the delay. This work evaluates the performance of the proposed designs in terms of delay, area, power, and their products by hand with logical effort and through custom design and layout in 0.18-m CMOS process technology. The results analysis shows that the proposed CSLA structure is better than the regular SQRT CSLA**.**

## III. PROPOSED TECHNIQUE

In the proposed methodology, the operations of the arithmetic unit is divided into various operations utilizes the IEEE 754 double precision format. The steps used for the implementation are:

### 3.1 Conditional Swap

In conditional swap unit, firstly all the different parts of the IEEE 754 double precision unit are extracted i.e. significand, exponent and sign. In this unit the exponents are compared and their difference is calculated. According to the difference calculated between the exponents the significands are arranged. The significand with the higher exponent value is assigned to x and the exponent with lower value is assigned to the variable y. Now in order to perform arithmetic operations on floating point unit, they first be converted to the $1.xxxx$ format. For the conversion into this format 1 is concatenated on the MSB side of the significand. After adding the MSB, now the significand y must be shifted to the right by the value of difference of the exponents. The output of the conditional swap unit is shifted significands x and y and the exponent and sign value which is further supplied to the arithmetic unit.

### 3.2 Arithmetic Operations

In the second unit the fundamental arithmetic operations are performed i.e. addition, subtraction, multiplication and division. For performing the basic operations the basic ripple carry adder is replaced by the fast and efficient carry select adder. Also the resource sharing among the operations is performed. It means that the addition unit used in the multiplication operation is also used to perform the addition and subtraction

operations and also the subtraction performed in the division operation.

The carry select adder is implemented using the basic ripple carry adders which computes the value of the sum and carry on the basis of carry input taken as '1' and '0'. The basic ripple carry adder is implemented using the reversible Peres gates. The unit computes the addition of the significand values along with the multiplication and division values. Figure 2 shows the flowchart of the basic arithmetic operations performed by the floating point unit using the reversible gates. The reversible gates present here are used to preserve the input values and can be used for further calculation.

### 3.3 Post Normalization

Post Normalization operation is performed after the addition operation. In this operation if there is a carry then the exponent in the third unit is also increased or decreased according to the value obtained after the operation. Normalization shift quantity is deducted just in case large cancellation happens throughout thesubtraction.

## IV. RESULTS AND DISCUSSIONS

The proposed methodology is implemented using the Xilinx Spartan 6 FPGA. The language used for the implementation is VHDL and the environment is Xilinx ISE. Figure 3 shows the top level module of the proposed methodology.

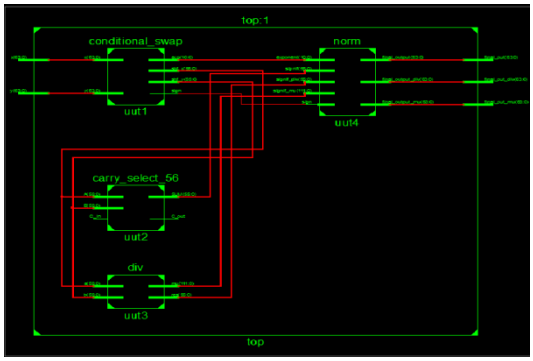


**Figure 3:** Top Module of the Proposed Approach

**Table 1:** Comparison Table

| Parameters | Basic Approach | Proposed Approach |
|---|---|---|
| Delay (ns) | 47.328 | 35.672 |
| Number of Slice LUTs | 673 | 690 |
| Number of DSP48A1s | 10 | 10 |
| Max Frequency (MHz) | 21.12 | 28.03 |

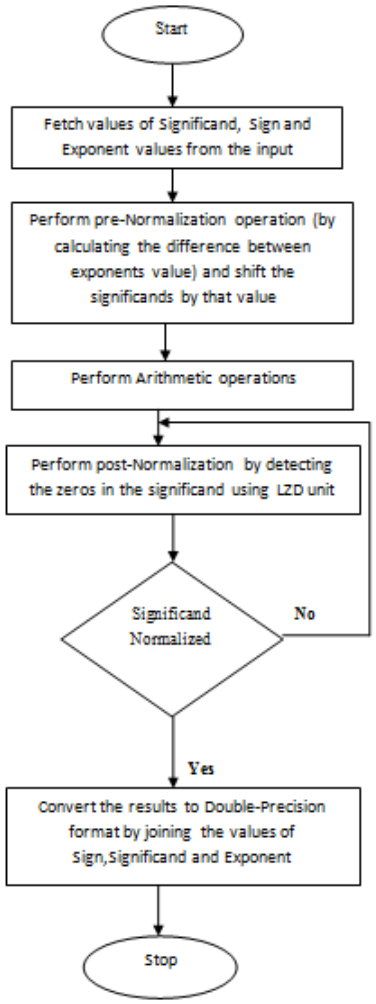


**Figure 2:** Flow Diagram of the Proposed Methodology

The Delay is defined as the critical path delay calculated for the implementation. The critical path Delay is the maximum delay which a circuit must have from input to output. The proposed approach shows the decrease in delay in the proposed approach with an increase in number of devices used. Figure 4 shows the simulation waveforms of the proposed methodology.

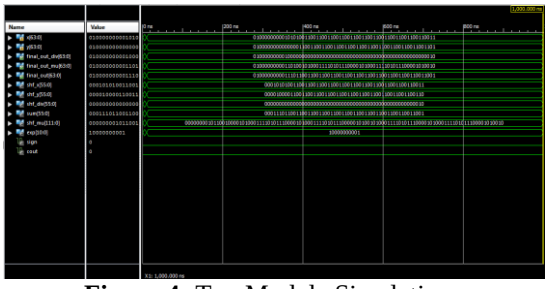

**Figure 4:** Top Module Simulation

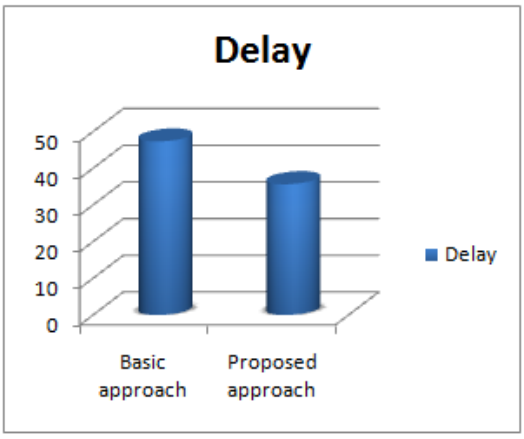


**Figure 5:** Comparison of delay between proposed design and Basic approach

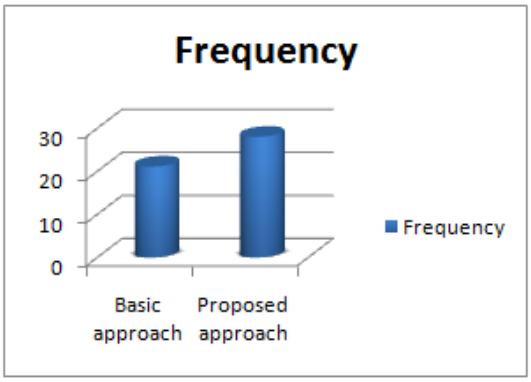


**Figure 6:** Comparison of frequency between proposed design and Basic approach

The proposed approach shows the decrease in delay i.e. 35.672ns as compared to the basic approach, haddelay of 47.328 ns. Similarly the maximum operating frequency of proposed approach is 28.03 MHz which is better than the basic approach's operating frequency i.e. 21.12MHz.

## V. CONCLUSION

In various Digital and Signal processing units, floating point arithmetic operations forms an important application on FPGA. IEEE 754 standard is used for the implementation of the unit with double precision standard. The unit is implemented using the reversible gates which preserve the input for further utilization. The maximum combinational path delay is improved by around 25% with a slight increase in area. The implementation of the proposed methodology uses the Carry Select Adder which is fastest adder. In future other units must also be optimized along with the adder to decrease the area utilization.

## REFERENCES

[1]. "IEEE Standard for Floating-Point Arithmetic", *ANSI/ IEEE Standard 754-2008, New York:IEEE Inc., Aug. 29 2008.*

[2]. Kahan, William. "IEEE standard 754 for binary floating-point arithmetic."*Lecture Notes on the Status of IEEE 754.94720-1776 (1996): 11.*

[3]. 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, ISSN:0975 – 8887, Volume 75, Issue No.2,PP: 47-51, August 2013.*

[4]. 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.*

[5]. Bruintjes, Tom M. "Design of a fused multiply-add floating-point and integer datapath." (*2011).*

[6]. Hafiz Md. Hasan Babu Md. Rafiqul Islam Syed Mostahed Ali Chowdhury Ahsan Raja Chowdhury, "Synthesis of Full-Adder Circuit Using Reversible Logic", *Proceedings of the 17th International Conference on VLSI Design , PP 757 – 760, DOI 10.1109/ICVD.2004.1261020, 2004.*

[7]. K,Saranya, "Low Power and Area-Efficient Carry Selet Adder", *International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-6,PP 114-117, Jauary2013*