*Proceedings of 4th RIT Post Graduates Conference (RIT PG Con-18)*
**NOVATEUR PUBLICATIONS**
*JournalNX- A Multidisciplinary Peer Reviewed Journal (ISSN No:2581-4230)*
**April, 13th, 2018**

# IMPLEMENTATION OF HALF PRECISION FLOATING POINT ARITHMETIC OPERATIONS FOR DSP APPLICATIONS

Miss. Supriya Sunil Phalle
Department of E&TC
Rajarambapu Institute of Technology, Sakharale
Islampur, India
supriyaphalle@gmail.com

Prof. M.R.Jadhav
Department of E&TC
Rajarambapu Institute of Technology, Sakharale
Islampur, India
maruti.jadhav@ritindia.edu

***Abstract*—For dealing with digital signals in real time, parameters like, speed of operation, hardware requirement, power and area, must take into consideration. Implementation of FFT, with less number of logic gates which helps to reduce area and power required for the design. With this motto multipliers are replaced with pass logic. To represent twiddle factors, standard IEEE floating point format is used. By considering The end user application, twiddle factors are represented in half precision format. So that it helps to increase the speed of application. FFT is completed with complex floating point multiplier, complex floating point adder/subtractor. All design is implemented in Verilog HDL in Quartus II web edition for Cyclone 4E FPGA family. The Synthesized RTL description is tested /simulated in ModelSim simulator**

***Keywords — Half precision IEEE754 floating point, Multiplier, adder, subtractor, QuartusII, Verilog, hardware descriptive language (HDL).***

## I. INTRODUCTION (*HEADING 1*)

In Digital signal processing of any signal, DFT is getting used in commonly. While the implementation of DFT takes complexity of $O(N^2)$, So Cooley and Tukey have given the approach of FFT with reduced in complexity upto *O(Nlog2N)* where N denotes the FFT size. Because of the Twiddle factor multiplication in the Cooley and Tukey algorithm, FFT processing unit takes large power and area. While designing any VLSI architecture, power requirements, area utilization must get considered.

The Purpose of this project is to implement the VLSI architecture for N-point FFT with reduced area and power consumption. N- point FFT DIF structure has N/2 no of multipliers in the $n^{th}$ stage of the butterfly which is huge count. It can be reduced by using Pass logic in the $n^{th}$ stage of the DIF structure. So that it will help to reduce the no of complex multiplication in the architecture. In the remaining stage we will use simple, complex adder and complex subtractor followed by complex multiplier, all modules follows the standard floating point format as the twiddle factor follows the float data type.

In this work, to represent twiddle factor, The IEEE floating point format is used. Precision of floating point number depends upon the end user application. Standard IEEE floating point format is represented in three formats, i.e. 1) Double precision (64 bit), 2) single Precision (32 bit), 3) Half Precision (16 bit).

For the implementation of the multipliers, one of the inputs is of the twiddle factor (floating point number) and second one is output of the subtractor. To do multiplication, shift and add multiplication algorithm is used. In N-point FFT, there are $n^{th}$ stages each stage has N/2 no of addition and N/2 no of subtractions with N/2 no of twiddle multiplications (i.e. Complex multiplication). While designing the adder and subtractor unit, operation must support the floating point input, which is given as input to the unit. So here focus is to reduce the number of multiplications and no of gates required in the architecture of FFT unit. This all work has to design in the VLSI design tool Of Quartus II 10.0 Web Edition, i.e. Designing RTL behavior of the design. After successfully completion of synthesis, RTL description goes through simulation of design in ModelSim simulator. Design process follows the VLSI design flow stages.

## II. RELATED WORK

**V. Arunachalam , Alex Noel Joseph Raj,** has proposed logic to minimize the no of multipliers at the input stage of the FFT in DIF structure. In OFDM application, as data is present in the form of the digital (binary data), so it is possible to replace the multipliers with the proposed pass logic. Twiddle factors are stored in the extra RAM registers so as to achieve the faster memory access. Because of the pass logic used in the design, the no of gates are getting reduced and that's why the area optimized and low power architecture has made for FFT structure. [1]

**Xiaolin Cao, Ciara Moore, Student, M´ aire O'Neill, Elizabeth O'Sullivan, Neil Hanley,** proposes the novel large multipliers hardware architecture using FFT and low hamming weights designs are proposed. FFT has designed with the help of serial integer multiplier architecture with feature of low hardware cost and reduced latency. [2]

**Trong-Yen Lee, Chi-Han Huang, Wei-Cheng Chen, Min-Jea Liu,** has proposed FFT architecture with reconfigurable modules. Shift and add method of multiplier with fixed point format is used to reduce the complexity of design. The design is flexible for the 64 to 512 points FFT. Design flexibility is achieved using partial dynamic reconfigurable FPGA, so that the hardware resource used is get minimized. It gives low area dynamic reconfigurable processor for wireless networks.[3]

**Mario Garrido , Miguel Ángel Sánchez, María Luisa López-Vallejo, and Jesús Grajal 's**, system Radix -4 memory based architecture of point 4096. Here memory based design has proposed. Design includes four memories of N/4 samples in the parallel instead of the series.The

design has Conflict free memory strategy that only requires only N –size of memory and few no of multipliers. [4]

**Mingyu Wang, Fang Wang, Shaojun Wei, ZhaolinLi,** Focuses on the pipeline architecture of the FFT implementation. Work uses the reconfigurable processor to implement the variable length single precision Floating point FFT/IFFT.In this work has proposed of reconfigurable butterfly to reduces up to the 75% adders as comparison with the conventional radix 4 algorithm and intermediated data cashing is achieved efficiently.[5]

**Zeke Wang,Xue Liu, Bingsheng He, and Feng Yu** has introduced the combined single path delay commutator- feedback radix-2 pipeline architecture of Fast Fourier Transform architecture. It has $log_2N-1$ SDC stage and *1* SDF stage. Design helps to reduce the 50% of complex multipliers as compare to the other radix 2 FFT design. Design has tested in the Virtex-5 FPGA , XC5VSX240T-2 FF1738. [6]

**Hao Zhang , Dongdong Chen , Seok-Bum Ko,** focused on the area and power efficient iterative pipelined architecture of floating point multiplier. This architecture support both double precision and single precision operations. Because of the iterative nature of the design, architecture requires less power, area get reduced as compared to the fully pipelined design. [7]

**L.P. Thakare, Dr. A.Y. Deshmukh** , Has focused on the complex multiplication in FFT implementation. Proposed FFT architecture very less resources in terms of the flip-flops, and multipliers, so that design is less power consuming. As the design uses the 43% of the total hardware so architecture is area efficient as well.[8]

## III. IEEE FLOATING POINT FORMAT

IEEE floating point format is used to represent the twiddle factor. Twiddle factor for the FFT algorithm comes in floating form. So to do operations on the floating number, it must represent the standard format. So here IEEE floating point format is used. IEEE 754 standardized representation for binary floating point numbers. Depends upon the application, precision of no is gets decided. As per required the precision, no of bits get to decide on the IEEE 754 format, i.e. 1) half precision(16 bit),2)single precision(32 bit), 3)Double precision (64 bit) etc. Precision in floating point format denotes the accuracy required in the floating point number. Accuracy of the floating point number depends upon the user application.

In many applications such as embedded, graphics, signal processing half precision format get used. The standard representation of half precision floating point format is as below:

### *A) Half Precision Floating Point Format*

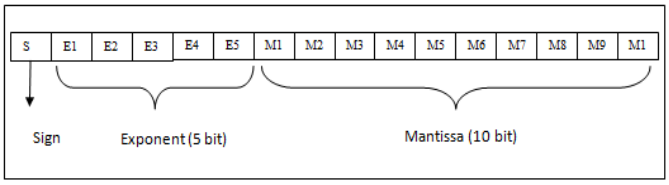


Fig.1. IEEE Floating point Format

- Sign bit: It represents sign of number. '1'for negative sign and '0' for positive sign
- Exponent: It represent 5 bit with biased form of true exponent. Bias is depends on the number of bit. Present in field, i.e. ($2^n$). True exponent is given as:- True Exponent = Exponent- Bias
- Mantissa: - It represents the Fraction part of data in 10 bit format with one hidden bit.

## IV. MULTIPLICATION OF COMPLEX NUMBER.

For performing floating-point complex multiplication we referred the basic mathematical step i.e.(A+jB)*(C+jD)= (AC-BD)+j(AD+BC).So for performing one complex multiplication we required four floating point multiplication and two floating point adder/subtractor unit.Seperate units of complex multiplier are as follows:

### *A. Floating point Multiplier*

Multiplication of two floating point no's are as per considering floating point format. First the sign of final answer is decided as Xoring sign bit of two numbers. Then adjust the exponent field of final answer with considering the upper and lower range of exponent (i.e.±14).If addition of two exponents by subtracting one bias from that exceeds exponent range the one have to shift the multiplied value which comes from the 11 bit multiplier. Finally just normalizing all final answer gets.

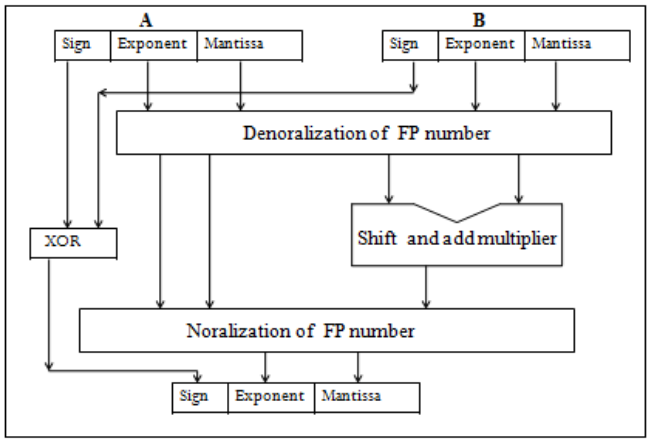


Fig.2. Floating point Multiplier

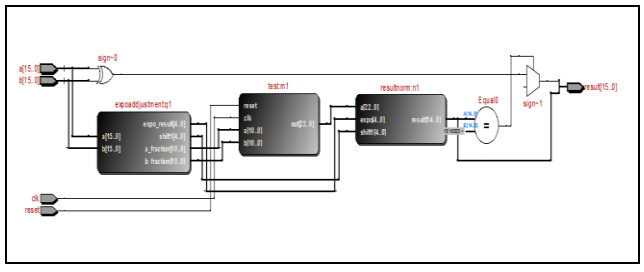

Fig.3. RTL Design of FP Multiplier

Register transfer level(RTL) is abstraction Module of digital circuits which connects/ transfer logic into hardware interface with the help of logic gates.

### *B. Floating point ADDER/Subtractor:*

For performing floating point addition/subtraction of 16 bit number , number is divided into three parts as per

format:1)sign bit,2)exponent bit,3)Mantissa bit. Which operation is to be performing is decided from xoring the sign bits given inputs. If it is zero, then addition operation takes place, and if it is one, then subtraction is taking place. By selecting proper exponent, respective mantissa is aligned. Alignment of mantissa is achieved by shifting proper mantissa from one of. Then subtraction/Addition of aligned mantissa takes place. Finally result is also arranged in IEEE 754 format as inputs are. The basic architecture of design is as follows:

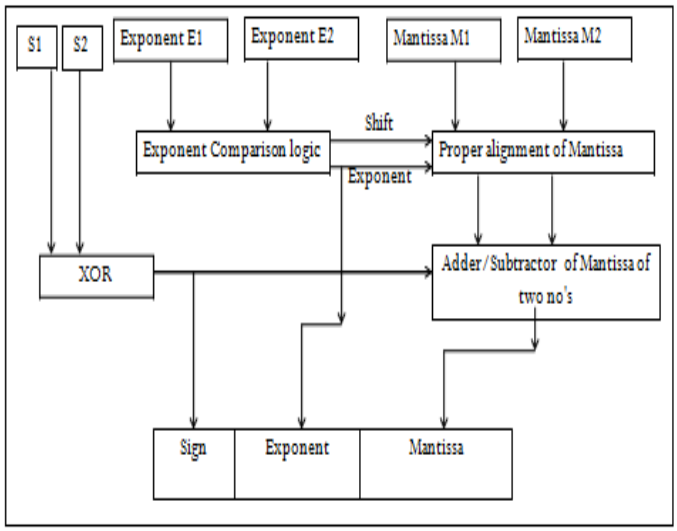


Fig.4. Floating point Adder/Subtractor

## V. Results and Discussion

All Design is implemented in RTL Description using Verilog HDL. After successful completion of synthesis, RTL Design get created. So that logical design get converted into resister transistor logic , and hardware for the design gets created. Functionality of Design gets tested using Modelsim Software.

### 1) *Floating point Multiplier:*

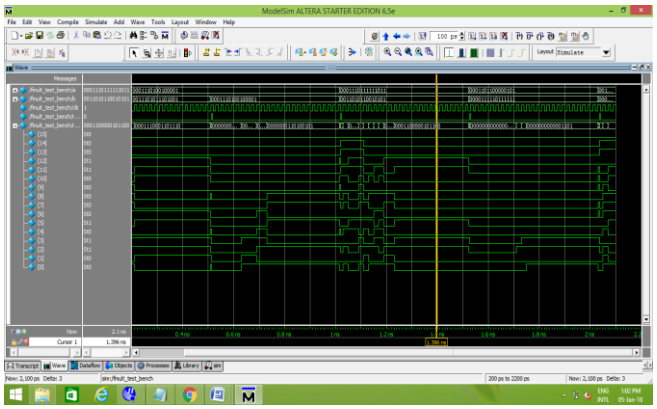

Fig.5. Simulation Result of FP Multiplier

### 2) *Floating point Adder:*

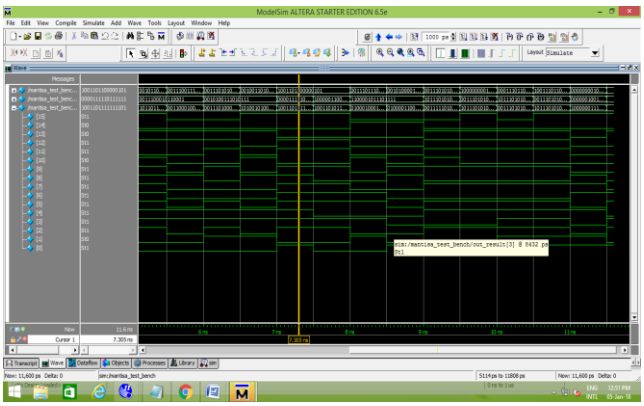

Fig.6. Simulation Result of FP Adder

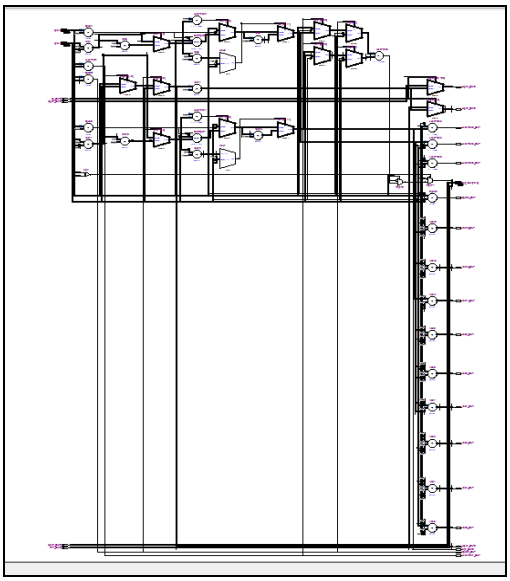

Fig.7. RTL Design of FP Adder/SubtractorI

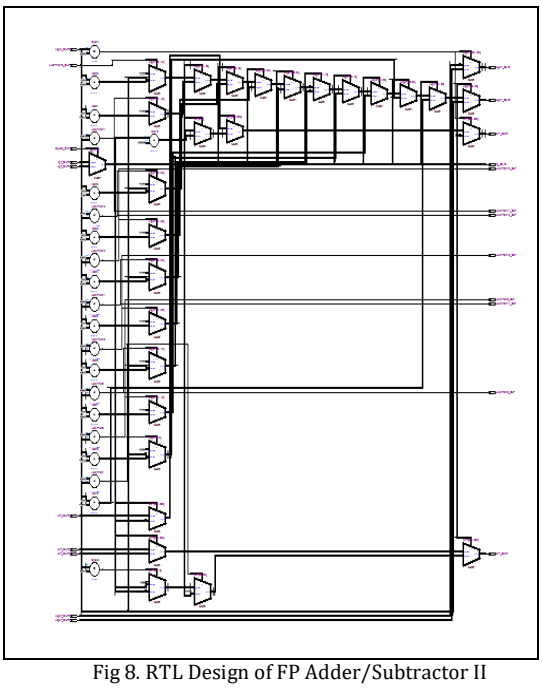

Fig 8. RTL Design of FP Adder/Subtractor II

*3) Floating point Subtractor:*

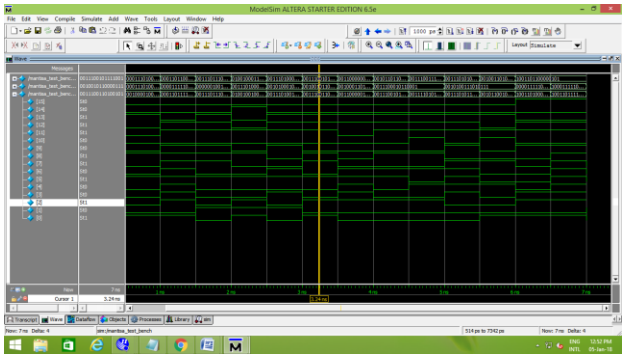

Fig.9. Simulation Result of FP Subtractor

*4) 8 Point FFT*

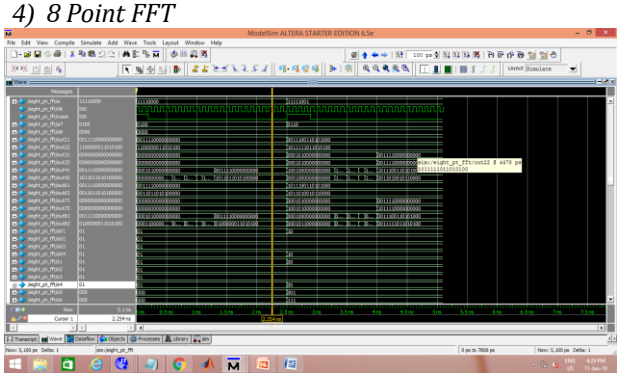

Fig.10. Simulation Result of 8 Pt FFT

Fig. 5,6 and 9,10 shows the results of Functionally verification of modules. From this functionality of design gets verified by giving proper inputs to the design. So that in out functionality gets tested.Fig.7&8 shows the Resistor transistor logic design of the respected modules. This is obtained by after successful synthesizing the Verilog modules.

## VI. Concusion

The designed Floating point multiplier using shift and multiplier, and standard architecture for adder/subtractor structure work for any point FFT with giving correct results. Hardware part utilization of designed fp multiplier uses <1% of logic elements whereas for 8 pt-FFT 7% of logic elements so design becomes area efficient. The device used for proposed design is EP4CE115F29C7FPGA of Cyclone 4E family.

## References

[1] V. Arunachalam, Alex Noel Joseph Raj, "Efficient VLSI implementation of FFT for orthogonal frequency division multiplexing applications", IET Circuits Devices Syst., 2014, Vol. 8, Iss. 6, pp. 526–531.

[2] Xiaolin Cao, Ciara Moore, Student Member, IEEE, M´aire O'Neill, Senior Member, IEEE, Elizabeth O'Sullivan, Neil Hanley, "Optimized Multiplication Architectures for Accelerating Fully Homomorphic Encryption", IEEE Transactions on Computers,2015.

[3] Trong-Yen Lee,Chi-Han Huang ,Wei-Cheng Chen , Min-Jea Liu, "A low-area dynamic reconfigurable MDC FFT processor design" , Microprocessors and Microsystems Elsevier Journal 2016.

[4] Mario Garrido, Member, IEEE, Miguel Ángel Sánchez,María Luisa López-Vallejo, and Jesús Grajal, "A 4096-Point Radix-4 Memory-Based FFT Using DSP Slices", IEEE Transactions On Very Large Scale Integration (VLSI) Systems 2016.

[5] Mingyu Wang , Fang Wang , Shaojun Wei , Zhaolin Li, "A pipelined area-efficient and high-speed reconfigurable processor for floating-point FFT/IFFT and DCT/IDCT computations", Microelectronics Journal of ELSEVIER , 47(2016)19–30.

[6] Zeke Wang, Xue Liu, Bingsheng He, and Feng Yu , "A Combined SDC-SDF Architecture for Normal I/O Pipelined Radix-2 FFT " , IEEE Transactions On Very Large Scale Integration (VLSI) Systems, April 16, 2014.

[7] Hao Zhang, Dongdong Chen, Seok-Bum Ko, "Area- and power-efficient iterative single/ double-precision merged floating-point multiplier on FPGA ", IET Computer Digit. Tech., 2017, Vol. 11 Iss. 4, pp. 149-158.

[8] L. P.Thakare, Dr. A.Y.Deshmukh ,"Area Efficient Complex Floating Point Multiplier for Reconfigurable FFT/IFFT Processor Based on Vedic Algorithm ",Elsevier Publication,7th International Conference on Communication, Computing and Virtualization 2016.

[9] K.Harikrishna, T. Rama Rao, Vladimir A. Labay, "FPGA Implementation of FFT Algorithm for IEEE 802.16e (Mobile WiMAX)", International Journal of Computer Theory and Engineering, Vol. 3, No. 2, April 2011.