A Review on Single Precision Floating Point Arithmetic Unit of 32 bit Number

Mr. Ankit Trivedi[1], Mr. Apoorv Verma[2]

[1]Assistant Professor, Axis Institute of Technology & Management Kanpur, up, India
[2]Student, Axis Institute of Technology & Management Kanpur, up, India

Abstract - Roughly computing has turn out to be one of the most popular computing paradigms in the era of the Internet of things and big data. It takes advantages of the error- tolerable feature of many applications, such as machine learning and image/signal processing, to reduce the resource required to deliver certain level of computation quality. Floating point arithmetic unit is widely used in many areas, especially scientific computation and signal processing. For many application signal processing and graphic it is acceptable to trade off some accuracy for faster and better implementations. Multiplication is the second basic operation of arithmetic Floating point unit.

FPU may be a part of ADP system specially designed to hold out operation on floating purpose variety.

Floating point unit is widely used in many areas. This paper shows review of IEEE 754 standard floating point arithmetic unit which will perform multiplication, addition and subtraction function on 32bit operand. A system's performance is usually determined by the performance of the multiplier, because the multiplier is usually the delay element in the system.

Single-precision floating-point IEEE-754 customary Adder/Subtractor and number modules with high speed and space economical square measure best owed.

The arithmetic operations square measure performed on the many a part of the IEEE format.

Key words: Floating Point Unit, IEEE 754

I. INTRODUCTION

A. *Floating-Point Units(FPU)*

It is a math coprocessor which is designed specially to carry out operations on floating point numbers [1]. The FPUs will perform operations like addition, subtraction, multiplication and division. Main function of FPUs can execute different functions such like as exponential or trigonometric calculations, although these are done with software library routine in nearly all recent processors. Our FPU is basically a 32 bit (single precision) IEEE754.

B. *Floating-Point Units(FPU)*

CPU executes a program then it is calling for a floating- point (FP) operation, which have three ways it can carry out the operation also be called by floating-point unit operation emulator which is floating-point library, used a series of simple fixed-point arithmetic operations which can run on the integer ALU. These emulators can keep the additional hardware price of a FPU but are significantly slow. Secondly it may us can add-on FPUs that are entirely separate from the CPU, and are typically sold as an optional add-ons which are purchased only when they are needed to speed up math- intensive operations and integrated FPU present in the system[2].

The FPU design as format of single precision IEEE754 compliant integrated unit also can't handle only basic floating point operations but also handle operations like shifting, square root determination and operation of transcendental functions like sine, cosine and tangential function.

C. *IEEE 754 Standard*

IEEE 754 used as floating-point computation, followed by many hardware (CPU and FPU) and software implementations [3]. Main standard for IEEE 754 Floating-Point calculations defines binary representation for floating-point numbers is in two types

1) Binary 32 bit (or single precision) format

2) Binary 64 bit (or double precision) format.

Bias	Fraction	Exponent	Sign	
127	23(22-00)	8(30-23)	1(31)	Single Precision
1023	52(51-00)	11(62-52)	1(63)	Double Precision

Table 1: Bit Range for Floating-Point Values

There is given IEEE 754 single-precision floating-point format that occupies 32 bits in a computer memory and given range of values in floating point unit. In IEEE 754, 32-bit with base 2 value format is referred to as single precision or binary32. It was called by single in IEEE 754-1985. The IEEE 754 standard given a format of single precision number which have sign bit maximum 1 bit length and exponent of width 8 bits , significant precision bits have 24 bits out of 23 bits are stored and where 1 bit is implicit 1.

There is given sign bit which is determining the value of sign number where 0 denotes a positive number and 1 denotes a negative number. It is known as sign bits (mantissa). Where exponent value is an 8 bit signed integer from −128 to 127 (2's Complement) or can be an 8 bit unsigned integer from 0 to 255 and given form of IEEE 754 single precision definition. Value of exponent value 127 represents actual zero, mantissa includes 23 fraction bits to the right of the binary point and an implicit leading bit and given left of the binary point with value 1 if the exponent is stored with all zeros. The value of 23 fraction bits of the mantissa appears in the memory format but there total precision is 24 bits.

For example:

M 23 Bit	E 8 Bit	S 1 Bit

Table 2-Floating Point Number Representation

An IEEE754 standard defines format which have a position of illustration of numerical values and symbols and also comprise the sets of bits are encoded.

II. LITERATURE REVIEW

"IEEE Standard given for Floating-Point Arithmetic Standard 754-2008, New York:" IEEE, It describes the interchange and an arithmetic method formats for binary and decimal floating-point arithmetic in computation. These standards have such conditions to default handling. The work of a floating-point system has conform to standard may be in software, entirely in hardware, or in any combination of software and hardware. In this standard, numerical results and exclusions specified for different operations in are exceptionally determined by the, sequence of operations, values of the input data and destination formats, all below user control.

Jain, Jenil, and Rahul Agrawal et al. [2] this paper grants design of high speed floating point unit using reversible logic. There are various alterable 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 processes are used very often in nearly all computing disciplines. The analysis of projected 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 newspaper, eight arithmetic and four logical operations has been presented. In the intended design 1, Peres Full Adder Gate (PFAG) is use in reversible ALU plan and HNG gate is used as an adder logic circuit in the planned ALU design 2. Both planned design are analyze and compare in terms of number of gates calculate, garbage output, quantum price and propagation interruption. The model results show that the proposed reversible ALU design 2 outperforms the proposed reversible ALU design 1 and conventional ALU design.

Nachtigal, Michael ,Himanshu Thapliyaletal.[4] In this work, a innovative design of single precision floating point multiplier has been proposed based on operand decomposition approach. Moreover, a new mutable design of the 8x8 bit Wallace tree multiplier has proposed that is accustomed in terms of quantum cost, delay, and number of garbage outputs. Wallace tree multiplication involves of three intangible steps: Partial product generation, partial product compression spending 4:2 compressors, full adders, and half adders, and then the ultimate accumulation stage to produce the product. In this slog we perform optimization at each of these three stages.

Dhanabal, R., Sarat Kumar Sahoo et al. [5] present a design using reversible gates. Alterable gates namely TSG gate performs 1-bit addition with carry. This is the first alterable gate which alone can acts as full adder. Gate is used to perform logical actions like AND, OR. In this works, designing 1-bit alum has also been presented using pass transistor with virtuoso tool of cadence. Based on examination 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 actions are needed very frequently in nearly all computing disciplines, and studies have shown floating-point addition to be the most often used floating-point operation. These paper offerings 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 subtractor, 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 planned design is made up of several parts, including: Conditional swap, Alignment unit, Converter, Addition and Normalization. We tried 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 planned 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.

Kahanetal.[8] proposed a dozen commercially vital arithmetic's boasted various word sizes, precisions, misestimating procedures and over/underflow behaviors', and additional were within the works. Suitable software system meant to reconcile that numerical diversity had become unbearably expensive to develop. 13years previous, 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 a normalizes that preoccupied all folks within the Seventies afflict solely CRAY X-MPs — J90scurrently.

Ykuntam et al. [9] Planned Addition is that the heart of arithmetic unit and also the arithmetic unit is commonly the work horse of a machine circuit. Therefore adders play a key role in planning Associate in Nursing arithmetic unit and additionally several digital integrated circuits. Carry pick 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 are 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 offers the planning of root CSLA victimization add one circuit with vital reduction in space.

Quinnell et al. [10] planned several new architectures for floating-point amalgamate multiplier-adders employed in the x87 units of microprocessors. These fresh 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 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-millimeter atomic number 14 on dielectric junction transistor technology logic gate design tool set. All style use the AMD'Barcelona'native quad-core standard-cell library as in study basics of making and differentiating the new architectures in an exceedingly with-it and more reliable and economic industrial technology.

This paper show evaluation of IEEE floating point unit (FPU) which will carry out multiplication, addition, subtraction and division reason on 32bit operand that uses the IEEE-754 regulation. Floating point numbers representation can support a much wider range of values than fixed point representation. The work is to tool and analyses floating point unit operation and hardware module were implemented using VHDL and synthesized using Xilinx ISE suite.

III. CONCLUSION

This paper presents the floating point unit according to IEEE 754 Standard. We resolve floating point representation concept which have large range of values as well as accuracy. Hence hardware necessity is reduced, thereby reducing power consumption and delay. So, designing of power efficient 32-bit single precision Floating point unit (FPU) based on IEEE-754 standard based on FPGA. In future, one can adopt Vedic mathematics also.

REFERENCES

[1] "IEEE Standard for Floating-Point Arithmetic", in IEEE STD 754-2008, vol., no., pp.1-70, Aug. 292008.

[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, MohdMahayadin, Syahira, AdibKabir 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 Nano technology (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."IEEEstandard754forbinaryfloating-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

[11] F. Conti, D. Rossi, A. Pullini, I. Loi, and L. Benini, “Energy-efficient vision on the PULP platform forultra-low power parallel computing,” in Signal Processing Systems (SiPS), 2014 IEEE Workshop on, Oct2014.

[12] V. Camus, J. Schlachter, and C. Enz, “Energy-efficient in exact speculative adder with high performance and accuracy control,” in Circuits and Systems (ISCAS), 2015 IEEE International Symposium on, May2015.

[13] Naresh Grover, M.K.Soni Design of FPGA based 32-bit Floating Point Arithmetic Unit and verification of its VHDL code using MATLAB I.J. Information Engineering and Electronic Business, 2014, 1, 1-14 Published Online February 2014 in MECS

[14] Syed M. Qasim, Ahmed A. Telba and Abdulhameed Y. AlMazroo FPGA Design and Implementation of Matrix Multiplier Architectures for Image and Signal Processing Applications IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.2, February 2010.