ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Floating Point Fused Add-Subtract and Fused Dot-Product Units

S. Kishor[1] , S. P. Prakash[2]

PG Scholar (VLSI DESIGN), Department of ECE Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India

Assistant Professor (Sr.G), Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India

ABSTRACT*:* A single precision floating-point fused add-subtract unit and fused dot-product unit is presented that performs simultaneous floating-point add and multiplication operations. It takes to perform a major part of single addition, subtraction and dot-product using parallel implementation. This unit uses the IEEE-754 single-precision format and supports all rounding modes. The fused add-subtract unit is only about 56% larger than a conventional floating-point multiplier, and consumes 50% more power than the conventional floating-point adder. The speed of the fused dot-product is about 27% faster than the conventional parallel approach. This will combine to use for FFT algorithms mainly. The simulation results are obtained using Xilinx 14.3 EDA tool. The results show that the RTL view and synthesis reports.

KEYWORDS*:* Fused Add-Subtract unit (FAS), Single precision floating point Dot-Product unit (FDP), Rounding modes, Number Of LUT'S, Delay, Verilog, Xilinx.

I. INTRODUCTION

Fixed-point arithmetic has been used for the longest time in computer arithmetic calculations due to its ease of implementation compared to floating-point arithmetic and the limited integration capabilities of available chip design technologies in the past. The design of binary fixed-point adders, multipliers, subtractions, and dividers is covered in numerous textbooks and conference papers. However, advanced technology applications require a data space that ranges from the infinitesimally small to the infinitely large. Such applications require the design of floating-point hardware. A floating point number representation can simultaneously provide a large range of numbers and a high degree of precision. As a result, a portion of most microprocessors is often dedicated to hardware for floating point computation. Floating-point arithmetic is attractive for the implementation for a variety of Digital Signal Processing (DSP) applications because it allows the designer and user to concentrate on the algorithms and architecture without worrying about numerical issues such as scaling, overflow, and underflow. In the past, many DSP applications used fixed point arithmetic due to the high cost (in time, silicon area and power consumption) of floating-point arithmetic units. In IEEE-754, the 32-bit with base 2 format is officially referred to as single precision or binary32. It was called single in IEEE 754.

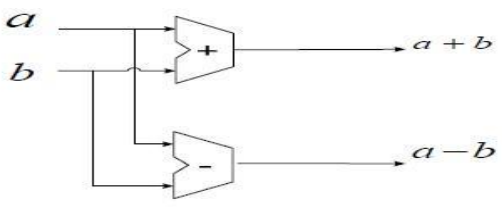

Fig.1 Parallel Implementation of FAS.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

This is required, for example, in computation of the FFT butterfly operation. In traditional floating-point hardware these operations may be performed in a serial fashion which limits the throughput. The use of a fused add-subtract (fused AS) unit in fig.1 and fused dot-product unit (FDP) in fig.2 accelerates the butterfly operation. Alternatively, the addition and subtraction may be performed in parallel with two floating-point adders which is expensive (in silicon area and in power consumption).

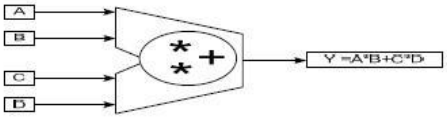

Fig.2 Parallel Implementation of FDP

This paper is organized as following. Section I describes the introduction about FAS and FDP for FFT applications. Section II Floating Point FAS. Section III Floating-Point FDP. Section IV Simulation and performance analysis in table forms are shown in Table (2-3), followed by the conclusion in Section V.

II. FLOATING POINT FUSED ADD-SUBTRACT UNIT

The architecture of the fused add-subtract unit is derived from the floating-point add unit. The exponent difference, significant shift and exponent adjustment functions can be performed once with a single set In Fig.3 shows the architecture of the fused add-subtract unit, the blocks with white background are the same blocks used for a single floating-point add operation. The blocks with green background are additional blocks used to perform the subtract operation, and the blocks with yellow background are similar to the floating point add blocks, but with extended functionality to calculate the sign and exponent for the new subtract operation

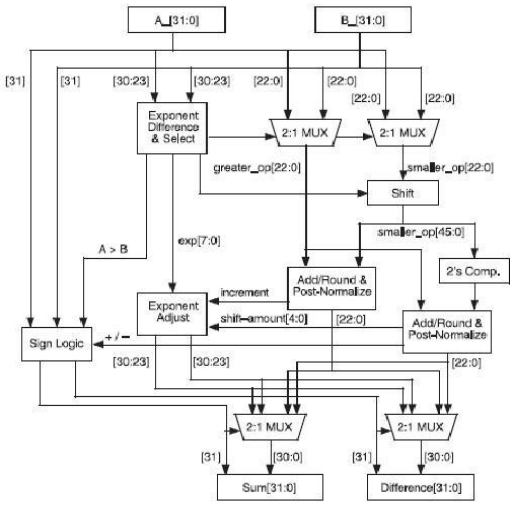

Fig.3 Floating-Point Fused Add-Subtract Unit

It detects the effective operation based on the signs of the two operands and the intended operation. It also generates guard and pre-sticky bits that aid in the proper rounding of the final results. In a parallel conventional

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

implementation of the fused add-subtract such as that two floating-point adders are used to perform the operation. This approach is fast, however, the area and power overhead is large because two floating point add/subtract units are used.

In a conventional implementation of the fused add-subtract one floating-point adder/Subtractor is used to perform the operation in addition to a storage element to store the addition or subtraction result. This approach is very efficient in terms of area. However, due to the serial execution of both operations, the time needed to get both results is twice the time needed by the parallel approach. Also since a storage element is used, it adds slightly to the area and power overhead, through two floating-point adders operating.

III. FLOATING POINT FUSED DOT-PRODUCT UNIT

The architecture of the fused dot-product unit is derived from the floating-point add unit. The exponent difference, significand shift and exponent adjustment functions can be performed once with a single set of hardware, with the results shared by both the add and the subtract operations in fig.3. New add and normalize blocks are needed for the new subtract operation. It shows the architecture of the fused add-subtract unit, the blocks with white background are the same blocks used for a single floating-point add operation. The blocks with green background are additional blocks used to perform the subtract operation, are similar to the floating point add blocks, but with extended functionality to calculate the sign and exponent for the new subtract operation. Since two operations are explicitly performed for sum and difference results (e.g., if the addition is used for the sum, the subtraction is used for the difference), the addition and subtraction are separately placed and only one LZA and normalization (for the subtraction) is required.

Assuming both sign bits are positive, the addition and subtraction are performed separately. Then, two multiplexers select the sum and difference with the operation decision bit, which is the XOR of the two sign bits. This will realize their Dot-product format of multiplication and sum them again to make as FDP for better than serial implementation. This FDP will increase the efficiency of FFT implementation.

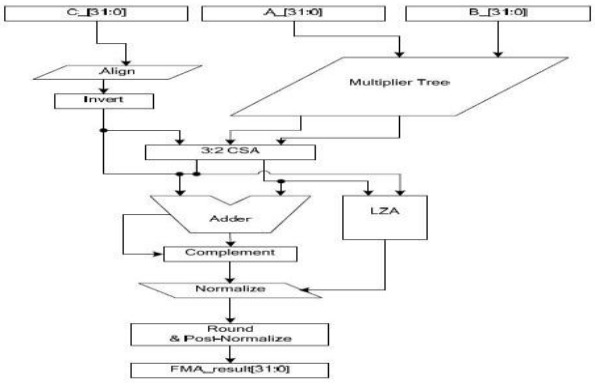

Fig.4 Floating point Fused Dot-Product Unit

IV. SIMULATION RESULTS AND PERFORMANCE ANALYSIS

The various arithmetic modules, Conventional floating point multiplier, Fused Add-Subtract unit and Fused dot-product unit are developed using Verilog. (fig.5 - fig.6 show that FAS & FDP symbol and fig.7 – fig.9 show that RTL view of FAS & FDP).

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

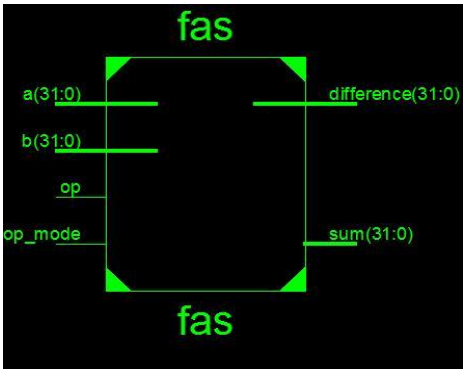

Fig.5 FAS symbol

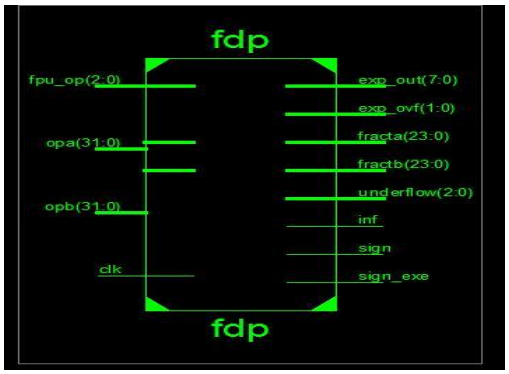

Fig.6 FDP symbol

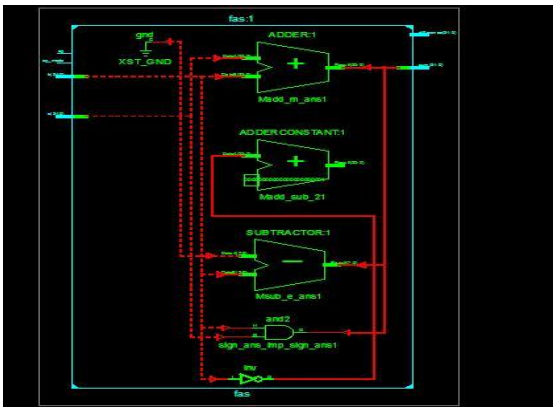

Fig.7 RTL view of FAS

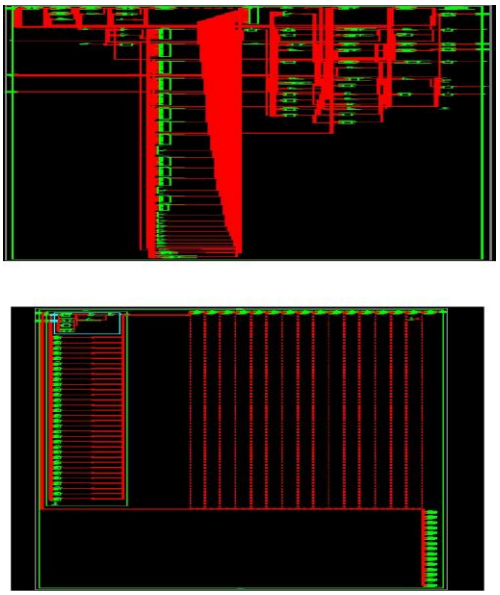

Fig.8 RTL view of FAS

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Fig.10 FAS Outputs

TABLE 1 Performance Analysis of FAS

Logic Utilization	Used	Available	Utilization
Number of Slice LUT's	360	27288	1%
Number of Fully Used LUT-FF Pairs	4	363	1%
Number of Bonded IOBs	136	296	45%
Number of BUFG/BUFGCTRLS	1	16	6%

V. CONCLUSION

This proposed fused architecture has been specifically designed for faster than conventional floating point add-subtract unit and multiplier and provide a slightly more accurate result. Both Fused FDP and FAS unit is more efficient than the older designs. FAS are more sense only rounding is performed over 3 rounding in parallel approaches. Mainly these architecture is useful for FFT implementation where we use of 2 FAS and 2 FDP unit for Radix- 2 FFT will designed for future. The proposed system reduces the shift amount and normalization is applied to reduce the size of significand addition and LZA reduces the reduction tree. Thus it improves more in Future work of FFT implementations.

REFERENCES

[1] Saleh and E.E. Swartzlander, Jr., "A Floating-Point Fused Add-Subtract Unit," Proc. IEEE Midwest Symp. Circuits and Systems

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

(MWSCAS), pp. 519-522, 2008.

[2] H.H. Saleh, "H.Fused Floating-Point Arithmetic for DSP," PhD dissertation,Univ. of Texas, 2008.

[3] Jorge Tonfat, Ricardo Reis, "Improved Fused Floating Point Add-Subtract and Multiply-Add Unit for FFT Implementation", in 2014 2nd International Conference on Devices, Circuits and Systems (ICDCS).

[4] Jongwook Sohn, Earl E. Swartzlander, "Improved Architectures for a Fused Floating-Point Add Subtract Unit," Ieee Transactions On Circuits And Systems—I: Regular Papers, Vol. 59, No. 10, October 2012

[5] V. Oklobdzija, "High-Speed VLSI Arithmetic Units: Adders and Multipliers", in "Design of High-Performance Microprocessor Circuits", Book Chapter, Book edited by A. Chandrakasan, IEEE Press, 2000.

[6] Earl E. Swartzlander, Hani H.M. Saleh, "FFT Implementation with Fused Floating-Point Operations," in IEEE Transactions On Computers, Vol. 61, No. 2, February 2012.

[7] Jongwook Sohn, Earl E. Swartzlander "Improved Architectures for a Floating-Point Fused Dot Product Unit" in 2013 IEEE 21st Symposium on Computer Arithmetic.

[8] Zhang Zhang, Dongge Wang "FFT Implementation with Multi-Operand Floating Point Units" in 978-1-61284-193-9/11/©2011 IEEE Transactions.