Design of a Floating-Point Fused Add-Subtract Unit Using Verilog

Article · June 2013

CITATIONS
0

READS
1,309

4 authors, including:

Dr. Ghanshyam Singh
Sharda University
45 PUBLICATIONS **118** CITATIONS
SEE PROFILE

R. M. Mehra
Sharda University
256 PUBLICATIONS **4,781** CITATIONS
SEE PROFILE

Available Online at www.ijecse.org **ISSN- 2277-1956**

Design of a Floating-Point Fused Add-Subtract Unit Using Verilog

Mayank Sharma, Prince Nagar, Ghanshyam Kumar Singh & Ram Mohan Mehra
Department of Electronics and Communication Engineering
School of Engineering & Technology
Sharda University, Knowledge Park-III, Greater Noida, (UP), India
Email- ghanshyam.singh@sharda.ac.in

Abstract- A floating-point (FP) fused add-subtract unit is presented that performs simultaneous floating-point operation of add-subtract on a common pair of single-precision data at the same time that it takes to perform in a single addition with a conventional floating-point adder. The system was placed and routed in 45nm process so that there will be less consumption of memory as well as power.

Keywords – Floating-point adder (FPA), Fused add-subtractor (FAS), Verilog

I. INTRODUCTION

The growing need of decimal arithmetic witnessed an increased attention in the recent years in many commercial applications and database systems, where the binary arithmetic is not sufficient. The floating-point fused multiply add (FMA) unit has been studied by so many researchers because it has several advantages in a floating-point unit design. The fused multiplier-add unit not only can reduce the latency of an application that executes a multiplication followed by an addition, but the unit may entirely replace a floating point co-processor's floating-point adder and floating-point multiplier [1-5]. Many DSP algorithms have been rewritten to take advantage of the presence of FMA units in a given system. For example, in a radix-16 FFT algorithm is presented that speeds up FFTs in systems with FMA units. High-throughput and digital filter implementations are possible with the use of FMA unit. FMA units were utilized in embedded signal processing and graphics applications, used to perform division, argument reduction , and this is why the FMA started to become an integral unit of many commercial processors such as IBM , HP and Intel.

Similar to operation performed by a FMA in many algorithms in DSP and other fields both of the sum and difference of a pair of operands are needed for subsequent processing. For example, this is required in computation of the FFT & DCT butterfly operations. 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 (FAS) unit accelerates the butterfly operation. Alternatively, the add and subtract may be performed in parallel with two independent floating point adders which is expensive [6-10]. This paper presents implementation of floating point add-subtract unit.

II. PROPOSED APPROACH

There are two design approaches that can be taken with discrete floating-point adders to realize add-subtract function. These are the parallel implementation shown in Figure 1 where two adders operate in parallel (one adding and one subtracting)and the serial implementation shown in Figure 2 where a single adder is used twice(once adding and once subtracting) with the same operands.

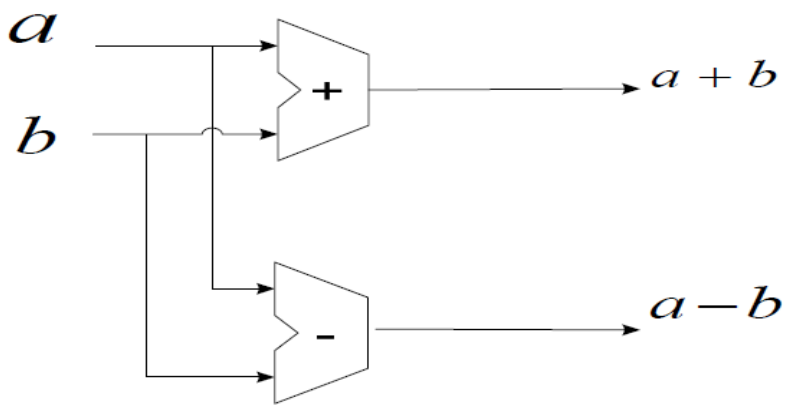

Figure 1 Conventional Parallel Realization of an Add-Subtract Unit

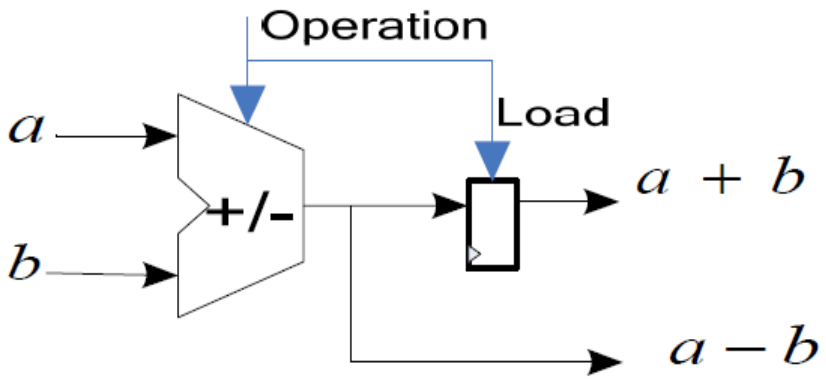

Figure 2 Conventional Serial Realization of an Add-Subtract Unit

In a parallel conventional implementation of the fused add-subtract (such as that shown in Figure 2) 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 serial conventional implementation of the fused add-subtract (such as that shown in Figure3) 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.

III. VERILOG MODELLING

The fused add-subtract unit, the following floating-point unit were implemented in synthesizable Verilog-RTL:

Fused floating add-subtract unit

The Verilog models were synthesized using 45 nm libraries. The area and the critical timing paths were evaluated. The floating-point adder and fused add subtract unit were designed to operate on single precision IEEE Std-754 operands [11-15].

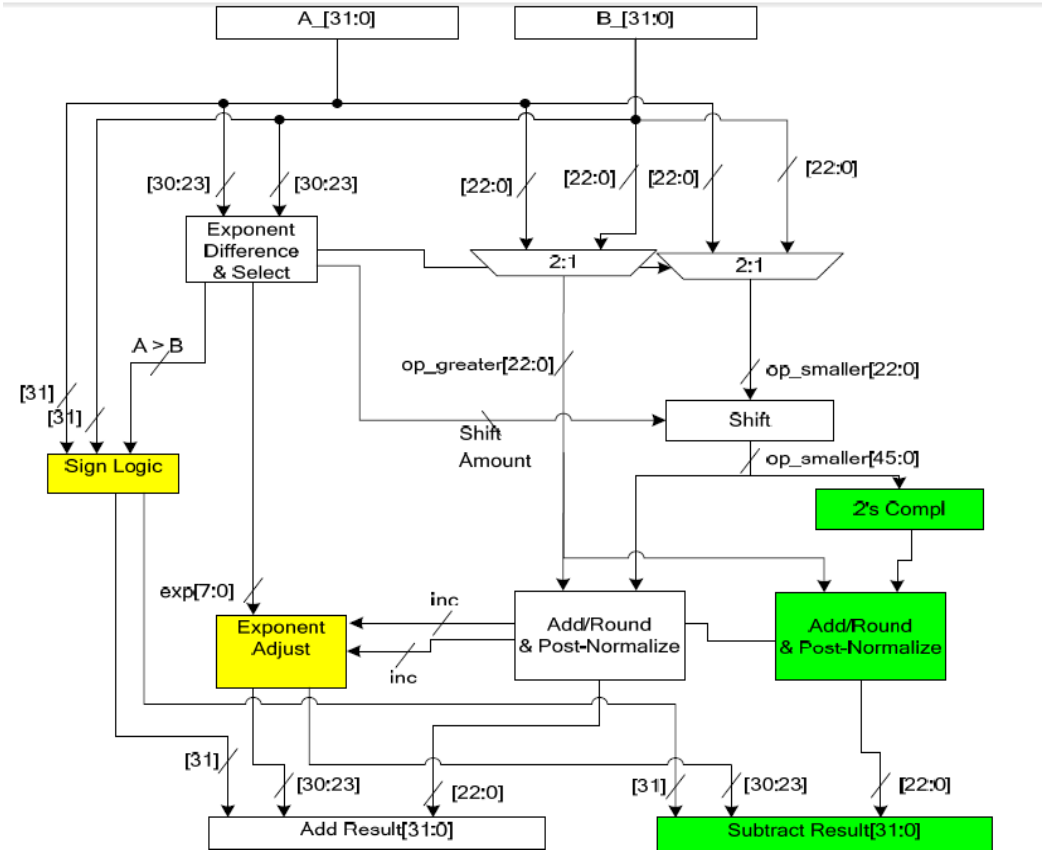

Figure 3 Floating-Point Fused Add-Subtract Unit

A. Simulation Results

The Floating point fused add-subtract unit architecture which has been proposed earlier has been implemented and designed through Verilog (ModelSim) and XILINX ISE ISIM Simulator. The output results have been shown below in Figure 4 and Figure 5. Here 2 floating point number has been performed through add and subtract operations in the following figure:

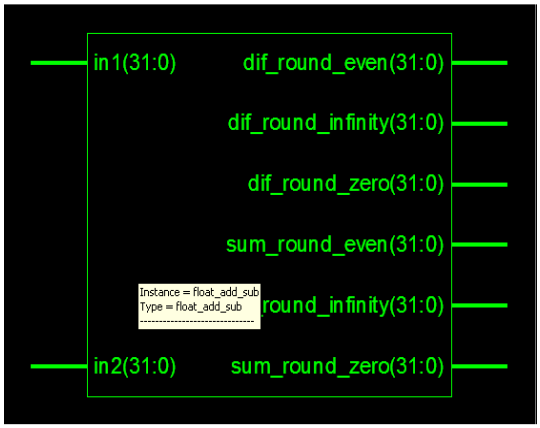

Figure 4: Input and Output block diagram

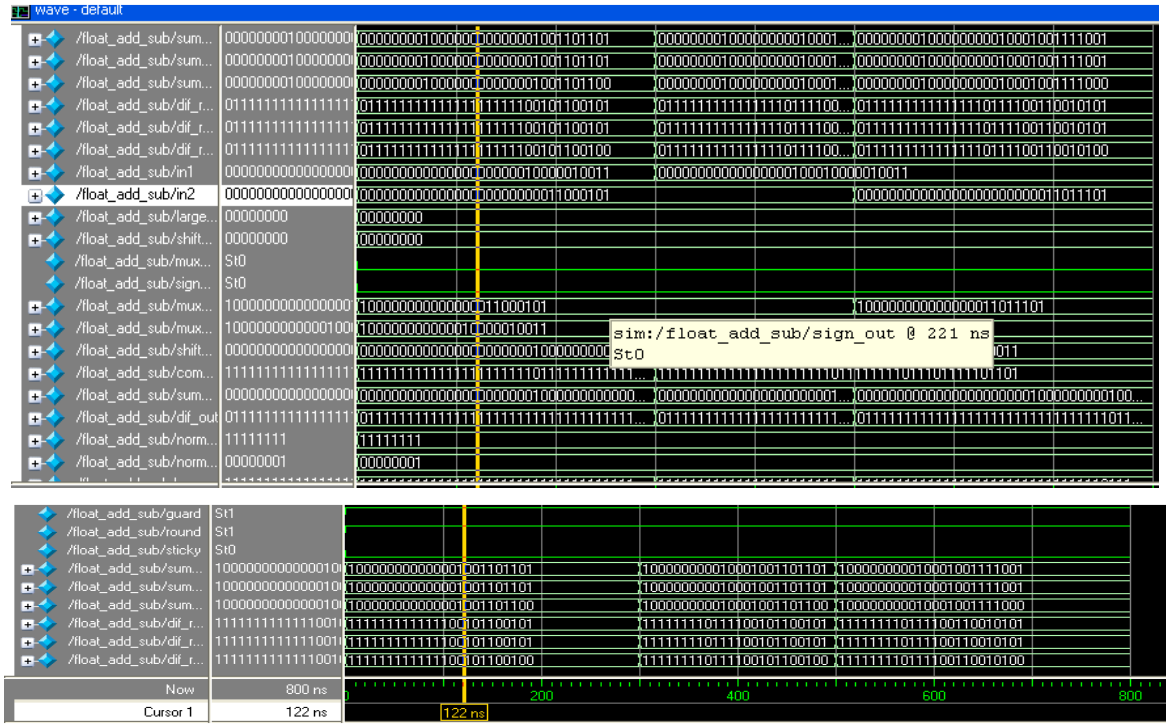

Figure 5: Simulation results Diagram

B. Synthesis Results

Here using the Xilinx XST tool for floating point fused add-subtract the following net-list has been generated which is shown in following Figure 6:

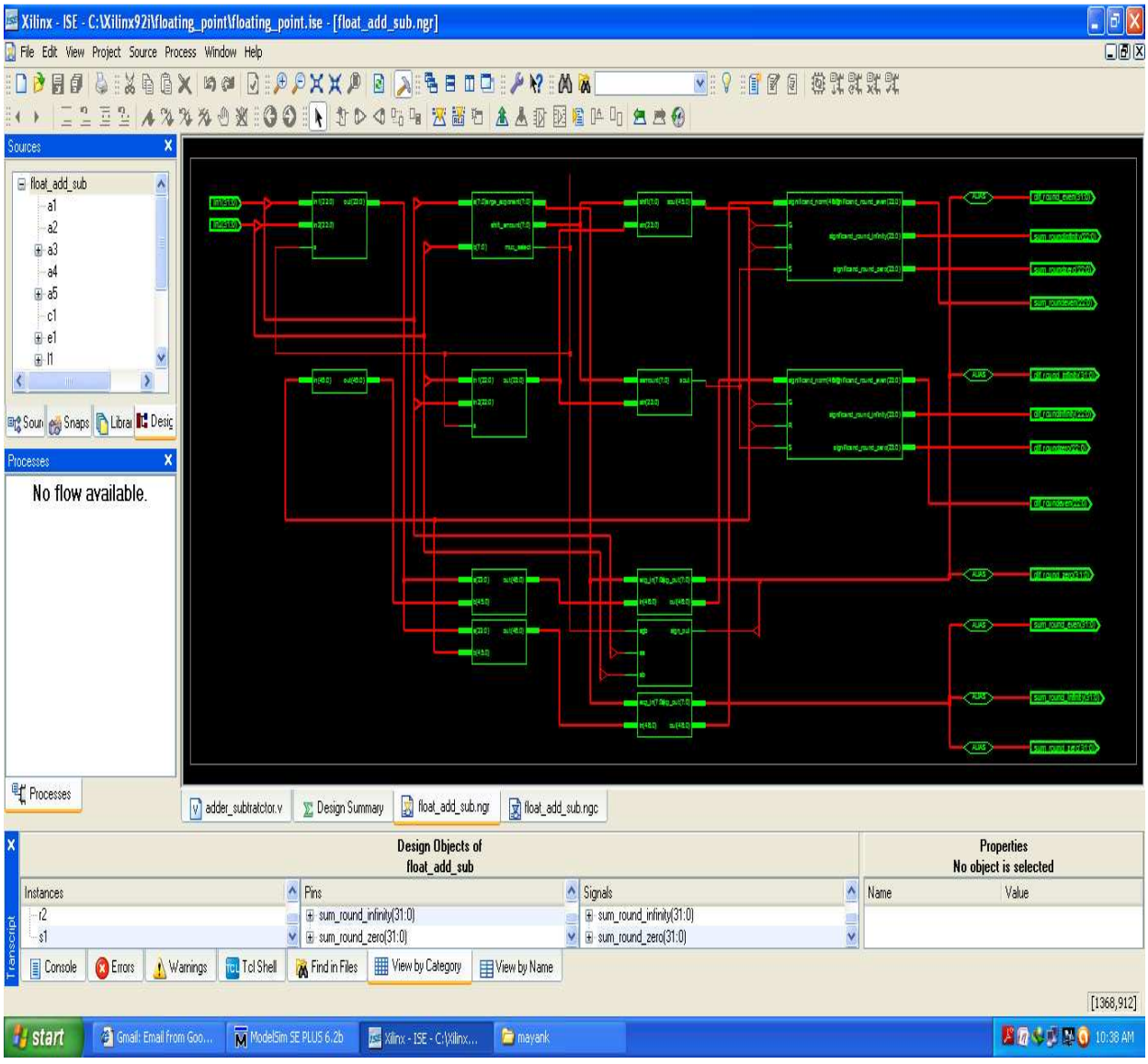

Figure 6: Net-list

IV. PLACE AND ROUTE VALUE

The floating-point adder (FPA) and fused add-subtract (FAS) unit were implemented using an automatic synthesize, place and route approach. A high performance 45 nm process was used for the implementation with a standard cell library designed for high speed applications [11-15].

Implementation value

	FPA	FAS
Format	**IEEE 754 Single-Precision**	
Cell area (unit inverter area unit)	**11342**	**15396**
Width	**48.9 µm**	**65.1 µm**
Total wire length	**14742 µm**	**25413 µm**
Critical Timing path	**2817ps**	**2813ps**
Dynamic power	**890 µw**	**1501 µw**
Leakage power	**142 µw**	**215 µw**
Height	**49.1µm**	**65.2 µm**
Total power	**1025 µw**	**1715 µw**

V.CONCLUSION

The new fused unit uses the IEEE-754 single-precision format and supports all rounding modes.The implementation results using a 45nm industry standard process and a simulations synthesis results shows that the fused primitives are faster, smaller, use less power and energy. It also provides slightly more accurate results. The fused add-subtract primitive unit can be used to realize many other DSP algorithms, including the basic butterfly computation of the discrete cosine transform and many forms of the wavelet transform.

The presented fused add-subtract design were implemented with no pipelines. The unit could be redesigned employing pipelining to achieve higher operation speeds. If proper pipeline gating were employed, then power consumption could be reduced as well. Future research can focus on optimizing the fused add-subtract unit using multipath approaches. These fusing concepts could be extended to other types of computation extensive applications and might result in delay, area and power consumption reduction

V. REFERENCE

[1] Akkas, A.; Schulte, M.J., "A decimal floating-point fused multiply-add unit with a novel decimal leading-zero anticipator," *Application-Specific Systems, Architectures and Processors (ASAP), 2011 IEEE International Conference on* , vol., no., pp.43,50, 11-14, Sept. 2011

[2] Samy, R.; Fahmy, H.A.H.; Raafat, R.; Mohamed, A.; ElDeeb, T.; Farouk, Y., "A decimal floating-point fused-multiply-add unit," *Circuits and Systems (MWSCAS), 2010 53rd IEEE International Midwest Symposium on* , vol., no., pp.529,532, 1-4 Aug. 2010

[3] Preiss, J.; Boersma, M.; Mueller, S.M., "Advanced Clockgating Schemes for Fused-Multiply-Add-Type Floating-Point Units," *Computer Arithmetic, 2009. ARITH 2009. 19th IEEE Symposium on* , vol., no., pp.48,56, 8-10 June 2009

[4] Swartzlander, E.E.; Saleh, H.H., "FFT Implementation with Fused Floating-Point Operations," *Computers, IEEE Transactions on* , vol.61, no.2, pp.284,288, Feb. 2012

[5] Takahashi, D., "A radix-16 FFT algorithm suitable for multiply-add instruction based on Goedecker method," *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on* , vol.2, no., pp.II,845-8 vol.2, 6-9 July 2003

[6] A.D. Robison, "N-Bit Unsigned Division Via N-Bit Multiply-Add," Proceedings of the 17th IEEE Symposium On Computer Arithmetic, pp. 131-139, 2005.

[7] R.-C. Li, S. Boldo and M. Daumas, "Theroems on Efficient Argument Reductions," Proceedings of the 16th IEEE Symposium on Computer Arithmetic, pp. 129-136, 2003.

[8] Kumar, "The HP PA-8000 RISC CPU," IEEE Micro Magazine, Vol. 17, Issue 2, pp. 27-32, April,1997

[9] Greer, J. Harrison, G. Henry, W. Li and P. Tang, "Scientific Computing on the Itanium Processor," Proceedings of the ACM/IEEE SC2001 Conference,pp. 1-8, 2004.

[10] M. P. Farmwald, On the Design of High Performance Digital Arithmetic Units, Ph.D. Thesis, Stanford University, 1981.

[11] IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Standard 754-1985.

[12] IEEE Standard for Floating-Point Arithmetic 754-2008, IEEE, 2008.

[13] IEEE Standard for Verilog Hardware Description Language, IEEE 1364-1995.

[14] IEEE Standard for Verilog Language, IEEE 1364-2001.

[15] IEEE Standard for System Verilog: Unified Hardware Design Specification and Verification, IEEE P1800-2005.

ISSN 2277-1956/V2N3-1007-1013

IJECSE
International Journal of Electronics and Computer Science Engineering
WWW.IJECSE.ORG