Design of High Speed IEEE-754 Double Precision Floating Point Multiplier Using Dadda Algorithm

Vivek .K(M.Tech), Gundam Narahari, and S.Jagadeesh

ABSTRACT: Floating Point (FP) multiplication is widely used in large set of scientific and signal processing computation. Multiplication is one of the common arithmetic operations in these computations. A high speed floating point double precision multiplier is implemented in HDL. This paper presents a high speed binary double precession floating point multiplier based on Dadda Algorithm. To improve speed multiplication of mantissa is done using Dadda multiplier replacing Carry Save Multiplier. In addition, the proposed design is compliant with IEEE-754 format and handles over flow, under flow, rounding and various exception conditions. The design achieved the operating frequency of 414.714 MHz with an area of 648 slices.
KEY WORDS: Dadda Algorithm, Double precision, Floating point, Multiplier, IEEE-754, Verilog HDL.

INTRODUCTION

The real numbers represented in binary format are known as floating point numbers. Based on IEEE-754 standard, floating point formats are classified into binary and decimal interchange formats. Floating point multipliers are very important in DSP applications. This paper focuses on double precision normalized binary interchange format. Figure I shows the IEEE- 754 double precision binary format representation. Sign (S) is represented with one bit, exponent (E) and fraction (M or Mantissa) are represented with eleven and fifty two bits respectively. For a number is said to be a normalized number, it must consist of 'one' in the MSB of the significand and exponent is greater than zero and smaller than 1023. The real number is represented by equations (I) & (2).

$$Z = (-1^{S}) * 2^{(E - Bias)} * (1.M) \quad (1)$$

$$Value = (-1^{Sign\ bit}) * 2^{(Exponent - 1023)} * (1.Mantissa) \quad (2)$$

Floating point implementation has been the interest of many researchers. In an IEEE-754 single precision pipelined floating point multiplier is implemented with custom 16/18 bit three stage pipelined floating point multiplier, that doesn't support rounding modes [1]. L.Louca, T.A.Cook, W.H. Johnson [2] implemented a single precision floating point multiplier by using a digit-serial multiplier. The design achieved 2.3 MFlops and doesn't support rounding modes. The multiplier handles the overflow and underflow cases but rounding is not implemented. The design achieves 30 I MFLOPs with latency of three clock cycles. The multiplier was verified against Xilinx floating point multiplier core.

1.Vivek .K(M.Tech),2.Gundam Narahari,3.S.Jagadeesh, 1.M.TechStudent in SSJ Engineering College,Hyderabad, anudeep405@gmail.com. 2.Ass.prof In ECE Dept,SSJ Engg,Hyderabad,3. HOD and Ass.prof In ECE Dept,SSJ Engg,Hyderabad, jaaga.ssjec@gmail.com.

Figure1. IEEE 754 Double Precision Floating Point Format.

The double precision floating point multiplier presented here is based on IEEE-754 binary floating standard. We have designed a high speed double precision floating point multiplier using Verilog language. It operates at a very high frequency of 414.714 MFlops and occupies 648 slices. It handles the overflow, underflow cases and rounding mode.

I.FLOATING POINT MULTIPLICATION ALGORITHM

Multiplying two numbers in floating point format is done by

1. Adding the exponent of the two numbers then subtracting the bias from their result.
2. Multiplying the significand of the two numbers
3. Calculating the sign by XORing the sign of the two numbers.

In order to represent the multiplication result as a normalized number there should be I in the MSB of the result (leading one).

The following steps are necessary to multiply two floating point numbers.

1. Multiplying the significand i.e. (I.MI * I.M2)
2. Placing the decimal point in the result
3. Adding the exponents i.e. (E I + E2 - Bias)
4. Obtaining the sign i.e. sl xor s2
5. Normalizing the result i.e. obtaining I at the MSB of the results "significand"
6. Rounding the result to fit in the available bits
7. Checking for underflow/overflow occurrence

II.IMPLEMENTATION OF DOUBLE PRECISION FLOATING POINT MUL TIPLTER

In this paper we implemented a double precision floating point multiplier with exceptions and rounding. Figure 2 shows the multiplier structure that includes exponents addition, significand multiplication, and sign calculation. Figure 3 shows the multiplier, exceptions and rounding that are independent and are done in parallel.

A_exponent B_exponent A_mantissa B_mantissa

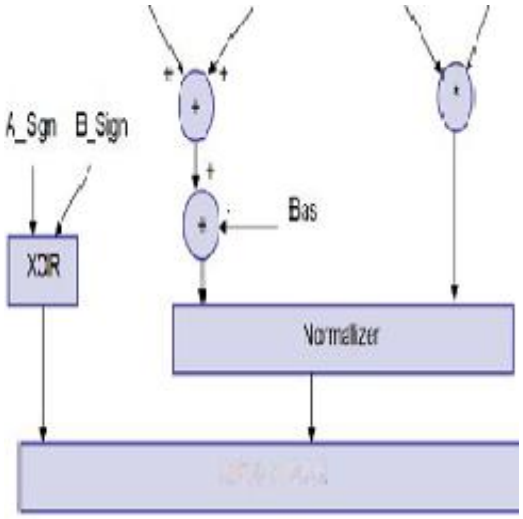

Figure 2. Multiplier structure

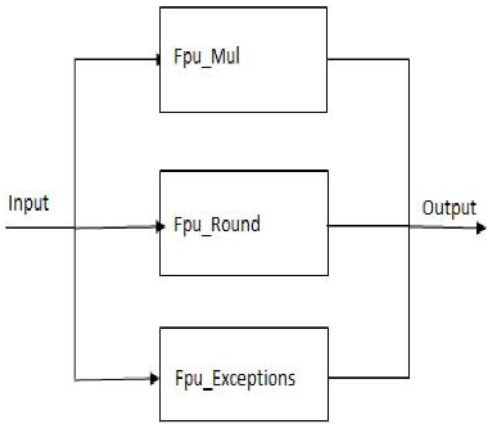

Figure 3. Multiplier structure with rounding and exceptions

III.MULTIPLIER

Existing Multiplier:

Carry Save Multiplier:

This unit is used to multiply the two unsigned significand numbers and it places the decimal point in the multiplied product. The unsigned significand multiplication is done on 24 bit. The result of this significand multiplication will be called the IR. Multiplication is to be carried out so as not to affect the whole multiplier's performance. In this carry save multiplier architecture is used for 24X24 bit as it has a moderate speed with a simple architecture. In the carry save multiplier, the carry bits are passed diagonally downwards (i.e. the carry bit is propagated to the next stage). Partial products are generated by ANDing the inputs of two numbers and passing them to the appropriate adder. Carry save multiplier has three main stages:

1. The first stage is an array of half adders.
2. The middle stages are arrays of full adders. The number of middle stages is equal to the significand size minus two.
3. The last stage is an array of ripple carry adders.

This stage is called the vector merging stage.

The count of adders (Half adders and Full adders) in each stage is equal to the significand size minus one. For example, a 4x4 carry save multiplier is shown in Figure 8 and it has the following stages:

1. The first stage consists of three half adders.
2. Two middle stages; each consists of three full adders.
3. The vector merging stage consists of one half adder and two full adders.

The decimal point is placed between bits 45 and 46 in the significand multiplier result. The multiplication time taken by the carry save multiplier is determined by its critical path. The critical path starts at the AND gate of the first partial products (i.e. a1b0 and a0b1), passes through the carry logic of the first half adder and the carry logic of the first full adder of the middle stages, then passes through all the vector merging adders. The critical path is marked in bold in Figure 4.

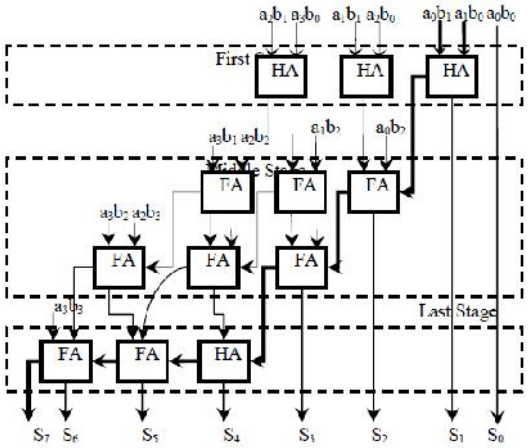

Fig. 4. 4x4 bit Carry Save multiplier

In Figure 4

1. Partial product: aibj ai and bj
2. HA: half adder.
3. FA: full adder.

Proposed multiplier

Dadda Multiplier:

Dadda proposed a sequence of matrix heights that are predetermined to give the minimum number of reduction stages. To reduce the N by N partial product matrix, dada multiplier develops a sequence of matrix heights that are found by working back from the final two-row matrix. In order to realize the minimum number of reduction stages, the height of each intermediate matrix is limited to the least integer that is no more than 1.5 times the height of its successor.

The process of reduction for a dadda multiplier [3] is developed using the following recursive algorithm

1. Let d1=2 and dj+1 = [1.5*dj], where dj is the matrix height for the jth stage from the end. Find the smallest j such that at least one column of the original partial product matrix has more than dj bits.
2. In the jth stage from the end, employ (3, 2) and (2, 2) counter to obtain a reduced matrix with no more than dj bits in any column.
3. Let j = j-1 and repeat step 2 until a matrix with only two rows is generated.

This method of reduction, because it attempts to compress each column, is called a column compression technique.

Another advantage of utilizing Dadda multipliers is that it utilizes the minimum number of (3, 2) counters. {Therefore, the number of intermediate stages is set in terms of lower bounds: 2, 3, 4, 6, 9 . . .

For Dadda multipliers there are N2 bits in the original partial product matrix and 4.N-3 bits in the final two row matrix. Since each (3, 2) counter takes three inputs and produces two outputs, the number of bits in the matrix is reduced by one with each applied (3, 2) counter therefore}, the total number of (3,2) counters is #(3, 2) = N2 – 4.N+3 the length of the carry propagation adder is CPA length = 2.N–2.

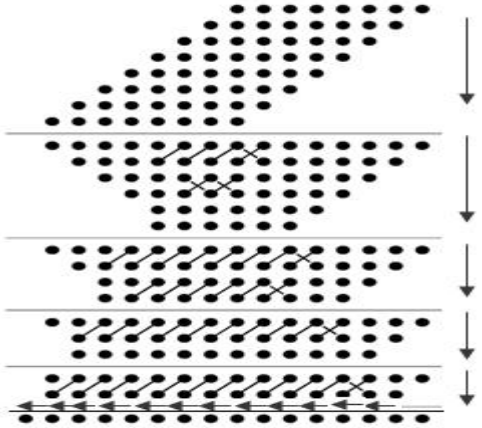

Fig. 5. Dot diagram for 8 by 8 Dadda Multiplier

The number of (2, 2) counters used in Dadda's reduction method equals N-1.The calculation diagram for an 8X8 Dadda multiplier is shown in figure 5. Dot diagrams are useful tool for predicting the placement of (3, 2) and (2, 2) counter in parallel multipliers. Each IR bit is represented by a dot.

The output of each (3, 2) and (2, 2) counter are represented as two dots connected by a plain diagonal line. The outputs of each (2, 2) counter are represented as two dots connected by a crossed diagonal line.

The 8 by 8 multiplier takes 4 reduction stages, with matrix height 6, 4, 3 and 2. The reduction uses 35 (3, 2) counters, 7 (2, 2) counters, reduction uses 35 (3, 2) counters, 7 (2, 2) counters, and a 14-bit carry propagate adder. The total delay for the generation of the final product is the sum of one AND gate delay, one (3, 2) counter delay for each of the four reduction stages, and the delay through the final 14-bit carry propagate adder arrive later, which effectively reduces the worst case delay of carry propagate adder. The decimal point is between bits 45 and 46 in the significand IR. Critical path is used to determine the time taken by the Dadda multiplier. The critical path starts at the AND gate of the first partial products passes through the full adder of the each stage, then passes through all the vector merging adders. The stages are less in this multiplier compared to the carry save multiplier and therefore it has high speed than that.

IV.ROUNDING AND EXCEPTIONS

The IEEE standard specifies four rounding modes round to nearest, round to zero, round to positive infinity, and round to negative infmity. Table 1 shows the rounding modes selected for various bit combinations of rmode. Based on the rounding changes to the mantissa corresponding changes has to be made in the exponent part also.

Bit combination	Rounding Mode
00	round_nearest_even
01	round_to_zero
10	round_up
11	round_down

Table1: Rounding modes selected for various bit combinations of rmode In the exceptions module, all of the special cases are checked for, and if they are found, the appropriate output is created, and the individual output signals of underflow, overflow, inexact, exception, and invalid will be asserted if the conditions for each case exist.

RESULTS

The double precision floating point multiplier design was simulated in Modelsim 6.6c and synthesized using Xilinx ISE 12.1.

Table 2 shows the area and operating frequency of double precision floating point multiplier, Single precision floating point multiplier [4] and Xilinx core respectively.

Device parameters	Present Work	M.Al-Ashrafy, A.Salem and W.Anis [6]	Xilinx Core
	Double Precision	Single precision	Single Precision
No. of slices	648	604	266
Maximum Frequency (MHz)	414.714	301.114	221.484

Table 2: Area and operating frequency of double precision floating point multiplier, single precision floating point multiplier [4] and Xilinx core

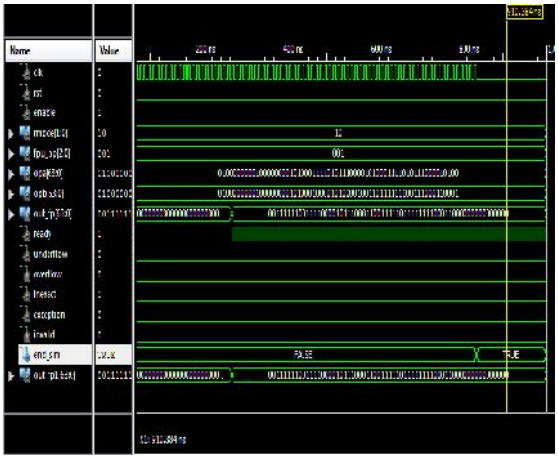

Fig. 6. Simulation Result of Double Precision Floating Point Multiplier

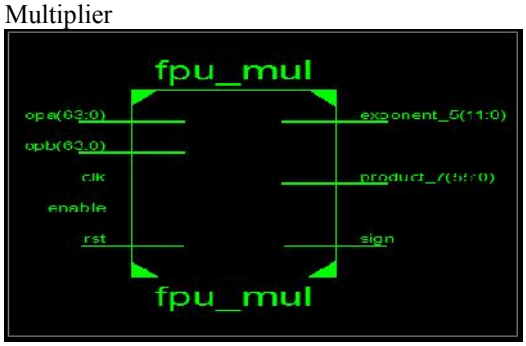

Fig.7. RTL Schematic for top level module

CONCLUSION

The double precision floating point multiplier supports the IEEE-754 binary interchange format. The design achieved the operating frequency of 414.714 MFLOOPS with area of 648 slices. The implemented design is verified with single precision floating point multiplier [4] and Xilinx core, it provides high speed and supports double precision, which gives more accuracy compared to single precession. This design handles the overflow, underflow, and truncation rounding mode.

REFERENCES

[1] N. Shirazi, A. Walters, and P. Athanas, "Quantitative Analysis of Floating Point Arithmetic on FPGA Based Custom Computing Machines," Proceedings of the IEEE Symposium on FPGAs for Custom Computing Machines (FCCM"95), pp.155-162, 1995.

[2] L. Louca, T. A. Cook, and W. H. Johnson, "Implementation of IEEE Single Precision Floating Point Addition and Multiplication on FPGAs," Proceedings of 83rd IEEE Symposium on FPGAs for Custom Computing Machines (FCCM"96), pp. 107-116,1996.

[3] Whytney J. Townsend, Earl E. Swartz, "A Comparison of Dadda and Wallace multiplier delays". Computer Engineering Research Center, The University of Texas.

[4] Mohamed AI-Ashraf)', Ashraf Salem, Wagdy Anis., "An Efficient Implementation of Floating Point Multiplier ", Saudi International Electronics, Communications and Photonics Conference (SIECPC), pp. 1-5,24-26 April 2011.

[5] B. Lee and N. Burgess, "Parameterisable Floating-point Operations on FPG A," Conference Record of the ThirtySixth Asilomar Conference on Signals, Systems, and Computers, 2002.

[6] Xilinx13.4, Synthesis and Simulation Design Guide", UG626 (v13.4) January 19, 2012.

[7] N. Shirazi, A. Walters, and P. Athanas, "Quantitative Analysis of Floating Point Arithmetic on FPGA Based Custom Computing Machines," Proceedings of the IEEE Symposium on FPGAs for Custom Computing Machines (FCCM'95), pp.155-162, 1995.

[8] L. Louca, T. A. Cook, and W. H. Johnson, "Implementation of IEEE Single Precision Floating Point Addition and Multiplication on FPGAs," Proceedings of 83 the IEEE Symposium on FPGAs for Custom Computing Machines (FCCM'96), pp. 107-116, 1996.