

# Datasheet of FP divisor and square root for variable precision

Lei Li (*lile@iis.ee.ethz.ch*)

08/06/2017

## 1. Algorithms and functions

Divisor employs the non-restoring binary divisor algorithm (NRBD)(K. Jun, and E. E.Swartzlander, Modified non-restoring division algorithm with improved delay profile and error correction, IEEE). And square root uses the non-restoring square root calculation algorithm (NRSC)(Y. Li, and W. Chu, Implementation of single precision floating square root on FPGAs,IEEE). To reduce the area overhead, they are designed with one shared control logic and share the used iteration cells. For divisors, to improve the accuracy, an extra MSC and adder is added (K. Jun, and E. E.Swartzlander, Modified non-restoring division algorithm with improved delay profile and error correction, IEEE).

Both support IEEE 754 for single precision. The three basic components in IEEE 754 are sign(S), exponent(E) and mantissa(M).

Table 1 IEEE 754 for single precision

Precision	Sign	Exponent	Mantissa	Bias
Single	1[31]	8[30:23]	23[22:0]	127

$$= (-1)^S \times (1.M)^E$$

Table 2 Special bit patterns in IEEE 754

	M=0	M≠0
E=0	0	Denormalized with real exponent=1
E=255	$\pm \infty$	NaN

The IEEE 754 2008 standard supports all floating point operations. It handles all special inputs including signaling NaN, quiet NaN, +Infinity, -Infinity, positive zero and negative zero. Our design fully supports IEEE 754 and offers three exceptions namely overflow(OF), underflow(UF) and division by zero(DZ). Besides, our design supports four different rounding modes: RNE(Round to Nearest, ties to Even, 00), RTZ(Round towards Zero, encoding as 01), RDN(Round Down, encoding as 10), RUP(Round Up, encoding as 11).

This document is organized as follows: Chapter 2 will summarize all inputs/outputs. Chapter 3 will introduce the architecture. Chapter 4 will address normalization. Chapter 5 will show the rounding modes. Chapter 6 will present exceptions. Chapter 7 will provide some waveforms and latency for simulations. Chapter 8 will give the synthesized results.

## 2. Inputs and Outputs

*div\_sqrt\_top\_tp* is the name of our design, which can be used to divide two floating point operands: Operand\_a\_DI by Operand\_b\_DI to produce a floating-point quotient, Result\_DO, and compute the floating-point square root of a floating-point operand, Operand\_a\_DI. The input RM\_SI is a 2-bit rounding mode. A parameter Precision\_ctl\_Enable\_S is introduced for enabling precision control with active high. When Precision\_ctl\_Enable\_S=1, Precision\_ctl\_SI can be used to control the needed precision. Precision\_ctl\_SI not only is used to control the finite state machine in *control\_tp* module, but also is employed to select the corresponding outputs. The values of Precision\_ctl\_SI are listed in Section 7 with the corresponding latency.

Table 3 Inputs and outputs

or	width	direction	Function
Clk_CI	1	IN	Clock
Rst_RBI	1	IN	Reset, active low
Div_start_SI	1	IN	Start the operation of divisor. Active high for one cycle.
Sqrt_start_SI	1	IN	Start the operation of square root. Active high for one cycle.
Operand_a_DI	32bits	IN	Div: Numerator; Sqrt:Radicand
Operand_b_DI	32bits	IN	Div: Denominator
RM_SI	2 bits	IN	Rounding mode.
Precision_ctl_SI	5bits	IN	Precision control
Result_DO,	32bits	OUT	Div: Quotient with one cycle ; Sqrt: Square root of Operand_a_DI with one cycle
Done_SO	1	OUT	Active high for one cycle
Ready_SO	1	OUT	Active high. It will hold high state until the next Div_start_SI or Sqrt_start_SI arrives

### 3. Architecture

According to radix 2 ( $r=2$ ) NRBD and NRSC,  $n$  iterations are needed for  $n$ -bit operands. For IEEE single precision, 24 iterations are needed with 23-bits mantissa and 1 hidden bit. 24 iterations can be implemented using an iteration unit with 24 cycles, or using  $m$  iteration units with  $24/m$  cycles. Thus, the appropriate  $m$  should be chosen.

For division, each iteration can be seen to be same and the control logic is comparatively simple. On-the-fly conversion and an extra MSC and adder are used to produce the final quotient. The key point of the control logic is how to store the generated quotients each cycle and how to select the needed quotient to choose the appropriate operands at the first iteration unit. An efficient method is to shift the quotient registers by  $24/m$  each cycle. Thus we can use a fixed register to choose.

The control logic of square root is more complex than that of divisor. It is because each iteration of square root is different with different intermediate operands. We have to add some fine-grained control. The corresponding selectors are controlled by a finite state machine (FSM).

The employed architecture is shown in Fig.1. *div\_sqrt\_top\_tp* is the top module, which is consisted of three modules: *preprocess*, *nrbd\_nrsc\_tp* and *fpu\_norm*. *nrbd\_nrsc\_tp* contains a control logic and four iteration units. The design can be seen as three stages: the first stage, the middle stage and last stage. To reach the target clock period of 2.8ns, using UMC65nm process technology, the solution based on four iteration units at the middle stage is chosen.  $8(=1+24/4+1)$  cycles are needed for producing the final results for single precision. The first cycle is used to store operands and generate control signals at the first stage. The 2nd-7th cycles are used to finish 24 iterations at the middle stage. The 8th cycle is used to normalize and round the result. The output result is then ready at the last stage (without flip/flops). In other words, the output results can be captured at the rising clock edge of the 8th cycle. Precision\_ctl\_SI is introduced to control the needed width of mantissa for variable precision and also the needed latency, which is shown in Section 7. Big margins are kept for the inputs and outputs, 1ns for the input delay and 0.8ns for the output delay.

In the *preprocess* module, two operands are unpacked into two IEEE-754 encoded numbers into corresponding sign bits, biased binary exponents, and mantissa. To support denormal numbers, two leading zero detectors(LZD) are added to counter the number of leading zeros in mantissa part of both operands. With LZD1 and LZD2, two operands are normalized. The resultant exponent for division can be calculated by  $(\text{Exp\_a\_D} - \text{Exp\_b\_D} + \text{Bias} + \text{LZ2} - \text{LZ1})$ . For square root, the resultant exponent can be computed as

$$\begin{aligned} & \frac{\text{Exp\_a\_D} - \text{LZ1} - \text{C\_BIAS}}{2} + \text{C\_BIAS} \\ &= \frac{\text{Exp\_a\_D} - \text{LZ1}}{2} + \text{C\_HALF\_BIAS} + (\text{Exp\_a\_D} - \text{LZ1})\%2 \end{aligned}$$

The result exponent and normalized operands are stored into flip/flops (*Exp\_norm\_D* and *Mant\_norm\_D*) for next stage. The sign of final result is calculated by using sign

of both operands or one based on Div\_start\_SI and Sqrt\_start\_SI and stored into a flip/flop. Operand detection is added to generate Inf\_a\_S, Inf\_b\_S, Zero\_a\_S, Zero\_b\_S, NaN\_a\_S and NaN\_b\_S for normalization of the final result. For the special cases NaN\_a\_S=1 or NaN\_b\_S=1, the input operands are needed to store into flip/flops for normalization. Special case control for low power is based on these signals.

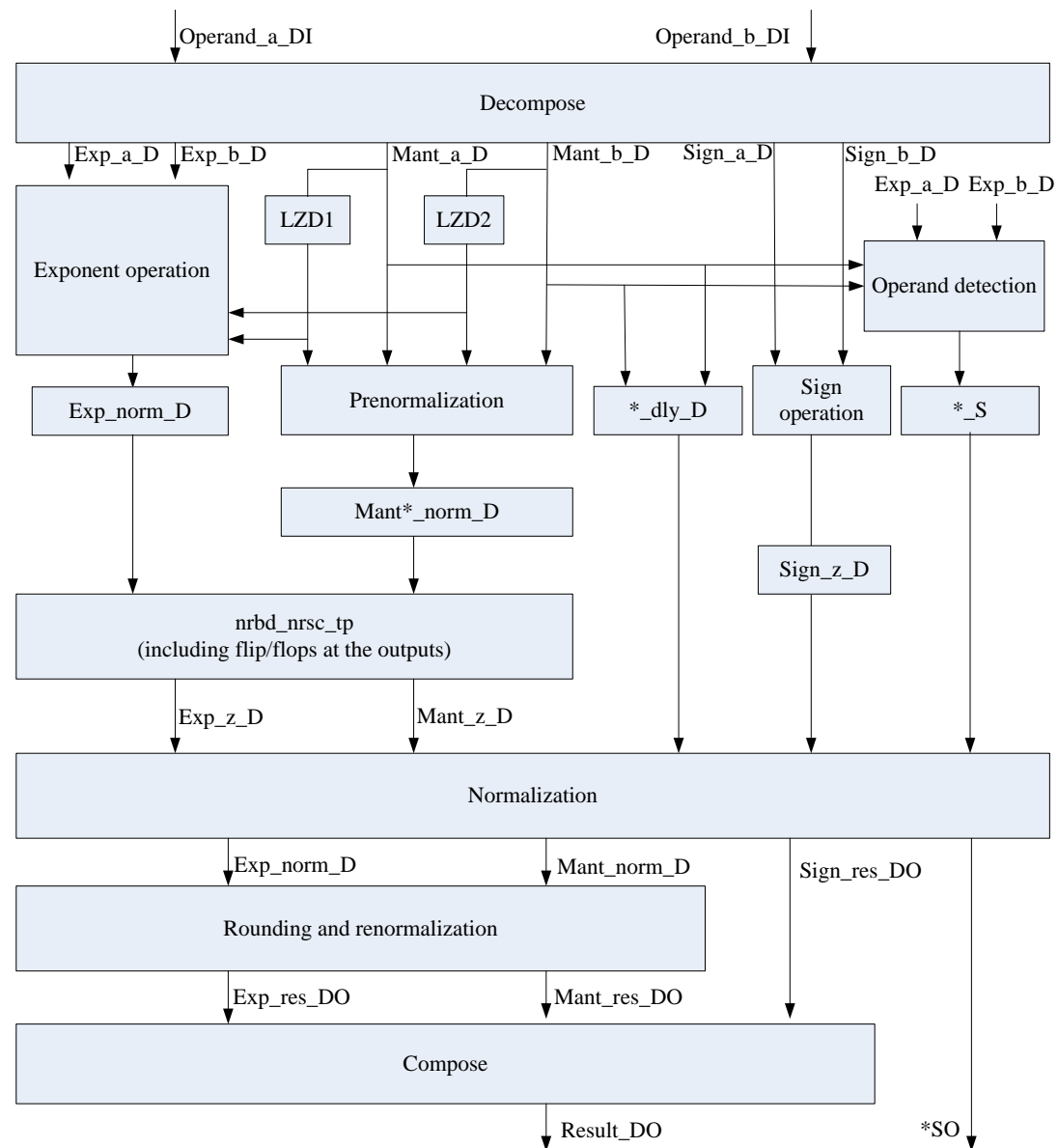


Fig.1 The architecture for the shared FP divisor and square root  
 \*\* \*\_D or \*\_S in a block are flip/flops.

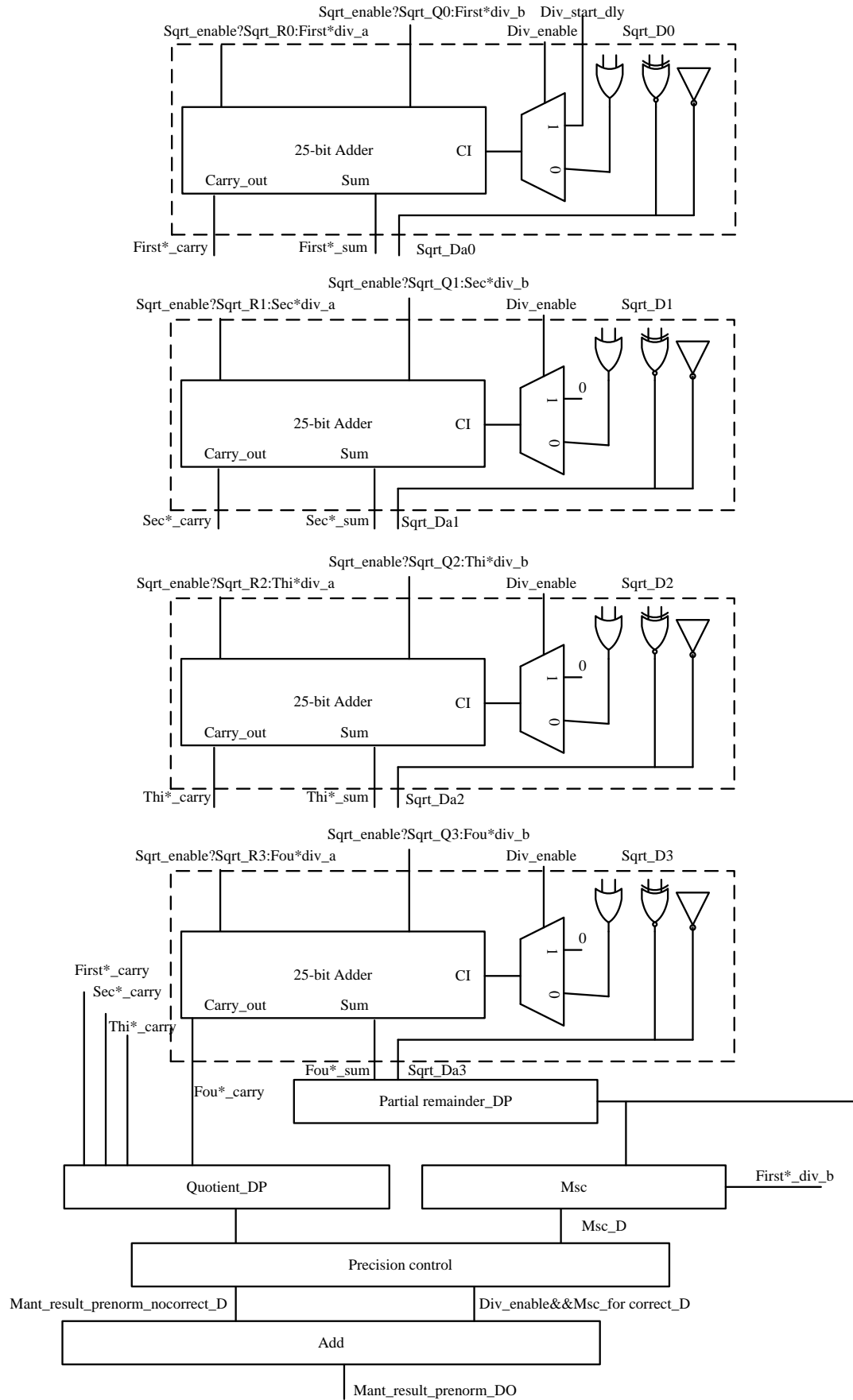


Fig.2 The data flow of the shared divisor and square root in *nrbd\_nsrc\_tp*

Fig.2 shows the data flow in *nrbd\_nsrc\_tp*. A finite state machine is used to control it. The final state of the used finite state machine is set base on the value of the Precision\_ctl\_SI. Sqrt\_enable\_S is used to choose the operands for square root and division.

For division, in the first iteration the left shift is not needed (Just like “the traditional pencil-and-paper method”, when we do a division, no left shift is needed before the first substraction. Besides, it breaks through the restriction  $Mant\_a\_D < Mant\_b\_D$  for some divisors) and the 2's complement introduced 1 should be added as carry-in in iteration\_cell\_for first by adding Div\_start\_dly\_SI. First\*div\_a is chosen from Mant\_a\_norm\_D from *preprocess* or Partical\_remiander\_DP based on Div\_start\_dly\_SI. The other input \*div\_a of next iteration cell are from Sum of the previous iteration cell directly, \*\_sum. For example, the input sec\*div\_a of the second iteration cell is from first\*\_sum, the Sum of the first iteration cell. \*div\_b is chosen from +denominator or – denominator according to the Carry\_out of the previous iteration cell \*\_carry. All the Carry\_outs of iteration cells are stored into Quotient\_DP. The final quotient Mant\_result\_prenorm\_D depends on Quotient\_DP[MANT-1:0] and Msc\_D based on Precision\_ctl\_SI, shown in Fig. 2. Herein Partical\_remiander\_DP and Quotient\_DP are flip/flops.

Square root is more complex than division. Sqrt\_D\* is chosen from Sqrt\_mant\_a\_norm\_D, 2bits for each iteration. Sqrt\_R0 is chosen from '0 or Partical\_remiander\_DP based on Sqrt\_start\_dly\_SI. The other input Sqrt\_R\* of next iteration cell are from Sum(\*\_sum) and D\_DO(Sqrt\_Da\*) of the previous iteration cell directly. Sqrt\_Q\* are different in each iteration with increase numbers, which are the Carry\_outs of the finished iterations. Mant\_result\_prenorm\_D is the result of square root before normalization.

The control signal from instruction decoder are Div\_start\_SI and Sqrt\_start\_SI. They will be stored in flip/flops as Div\_start\_dly\_S and Sqrt\_start\_dly\_S, and be used to generate Div\_enable\_S and Sqrt\_enable\_S in control module.

*Fpu\_norm* include normalization and rounding and renormalization. The employed schemes are shown in Section 4 and rounding in Section 5. The produced exception flags will be given in Section 6.

## 4. Normalization

### 4.1 division

(1)For normal IEEE754 operands, the result mantissa of division should start with 1 or 01. Therefore, we just need to care about the MSB of the quotient. When the quotient is 1.XXX, we check if the resultant exponent  $Exp\_a\_D - Exp\_b\_D + bias$  is out of the range of exponent. When the quotient is 0.1XX, we need to shift the mantissa one bit to the left and correct the exponent to:  $Exp\_a\_D - Exp\_b\_D + bias - 1$ .

$$\frac{1.M1}{1.M2} = 1.XXX \text{ or } 0.1XX$$

(2) If the numerator is a denormal number,

$$\frac{0.M1}{1.M2}$$

The resultant exponent is  $\text{Exp\_a\_D} - \text{Exp\_b\_D} + \text{bias}$ . May be a negative number. Index of LZD should be checked for normalization. If it is a negative number, we have to right shift the mantissa by Index of LZD to check if the resultant exponent ( $E + \text{Index of LZD}$ ) is negative. If it is negative, it is overflow.

(3) If the denominator is a denormal number,

$$\frac{1.M1}{0.M2} > 1$$

It is analyzed above.

(4) If these two operands are denormal numbers,

$$\frac{0.M1}{0.M2}$$

We should detect the first ones of these operands. It is the reason that we added two LZDs before operation.

Solution: If the hidden bit is 0, we should detect the first one of operands and left shift these two operands to normal mantissas. Thus we just care about exponent.  $\text{Exponent} = \text{Exp\_a\_D} - \text{Exp\_b\_D} + \text{Bias} + \text{LZ2} - \text{LZ1}$ , can be positive or negative. The leading one of the quotient should be detected for normalization.

(5) Other cases

If  $1 \leq E \leq 254$ , it is a normal result, return E and M;

If  $E=0$  and  $M \neq 0$ , it is a denormal number, return  $\gg (M)$  and the adjusted E;

If E is negative, the numbers cannot be represented. UF is signaled and return  $E=0, M=0$  (If so, it is all right for testbench. If not, cannot pass the check);

If  $E=255$  and  $M \neq 0$ , NaN is signaled and return qNaN;

If  $E > 255$ , OF is signaled and return  $E=255, M=0$ .

(6) Special cases

Table 4 IEEE 754-2008 specification for  $|x|/|y|$

$ x / y $		$ y $			
		+0	(sub)normal	$+\infty$	NaN
$ x $	+0	qNaN	+0	+0	qNaN
	(sub)normal	$+\infty$	$\circ( x / y )$	+0	qNaN
	$+\infty$	$+\infty$	$+\infty$	qNaN	qNaN
	NaN	qNaN	qNaN	qNaN	qNaN

“Except when otherwise stated, if the result of a floating-point operation is NaN, it is the canonical NaN. The canonical NaN has a positive sign and all significand bits

clear except the MSB, a.k.a. the quiet bit. For single-precision floating-point, this corresponds to the pattern 0x7fc00000.” Risc-spec V2.1 says.

Table 5 Special cases for division

Division	operation	return
1	a/NaN	qNaN*
2	a/Inf	0, the sign depending on the two operands
3	a/0	Inf, the sign depending on the two operands
4	0/b	0, the sign depending on the two operands
5	Inf/b	Inf, the sign depending on the two operands
6	NaN/b	qNaN
7	0/Inf	0, the sign depending on the two operands
8	Inf/0	Inf, the sign depending on the two operands
9	0/0	qNaN
10	Inf/Inf	qNaN

\*qNaN=0x7fc00000

## 4.2 square root

(1) The operand is a normal number

For normal IEEE754 operands, the result mantissa of square root should start with 1. Thus, we can check the final exponent. The 1.M will be left shifted one or zero-bit so that the new exponent  $e'$  makes  $e' - 127$  even. The shifted fraction will be 1X.XXX or 01.XXX. The result value will be 1.XXX. The resultant exponent can be computed as

$$\frac{e - 127}{2} + 127 = \frac{e}{2} + 63 + e\%2$$

(2) The operand is a denormal number

If the input operand is a denormal number, we can left shift like division.  $e$  can be a negative number. If  $e$  is even, it needs left shift 1-bit more.

(3) Other cases

Same to division.

(4) Special cases



Table 6 IEEE 754-2008 specification for square root

Operand x	+0	$+\infty$	-0	Less than zero	NaN
Result r	+0	$+\infty$	-0	qNaN	qNaN

Table 7 Special cases for square root

Square root	operands	
1	+0	+0
2	+NaN	qNaN
3	+Inf	+Inf
4	-0	-0
5*	-a	qNaN

\* can not be covered by division

## 5. Rounding

The design supports four different rounding modes: RNE(Round to Nearest, ties to Even, 00), RTZ(Round towards Zero, encoding as 01), RDN(Round Down, encoding as 10), RUP(Round Up, encoding as 11).

Table 8 Rounding modes

Mode	Code
RNE	00
RTZ	01
RDN	10
RUP	11

## 6. Exceptions

The design supports three exceptions namely overflow(OF), underflow(UF) and division by zero(DZ).

Div\_zero\_SO can be given by the LZD2 with resultant sign directly. Returns infinity (positive or negative ) as result

Exp\_OF\_SO is signaled if the exact result has an exponent that cannot be represented in the format. Returns infinity (positive or negative ) as result.

Exp\_UF\_SO is signaled when the result is denormal and rounded.

## 7. Waveforms and latency

The design was tested and it works well. Fig.3-9 present some waveforms. Div\_start\_SI or Sqrt\_start\_SI is coming with the operands when Ready\_SO=1. When Done\_SO=1, the Result\_DO is ready. The latency depends on the Precision\_ctl\_SI.

Table 9 Latency

Precision_ctl_SI	Latency	Max loss of mantissa
8-11	5	2**(23- Precision_ctl_SI) with MSC
12-15	6	
16-19	7	
20-23	8	
23(single precision)	8	1
Special cases*	2	-

\* No matter which value is Precision\_ctl\_SI set, the results will be ready at the third clock edge, which is shown in Fig. 5, Fig. 8 and Fig. 9. This feature has been removed in our transprecision design.

(1) Precision\_ctl\_SI=20, Latency=8 clock cycles

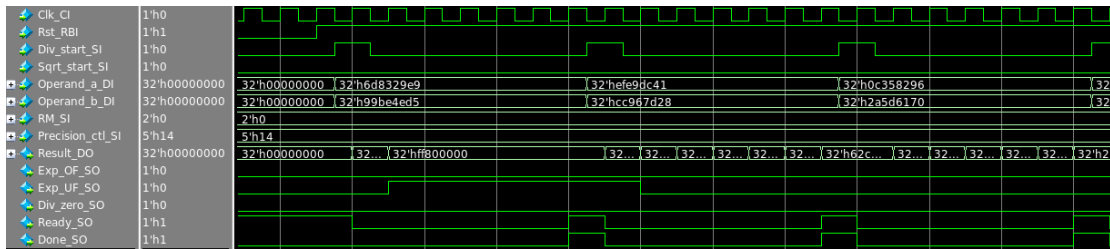


Fig.3 The waveform at the beginning with all inputs delayed by half of clock period

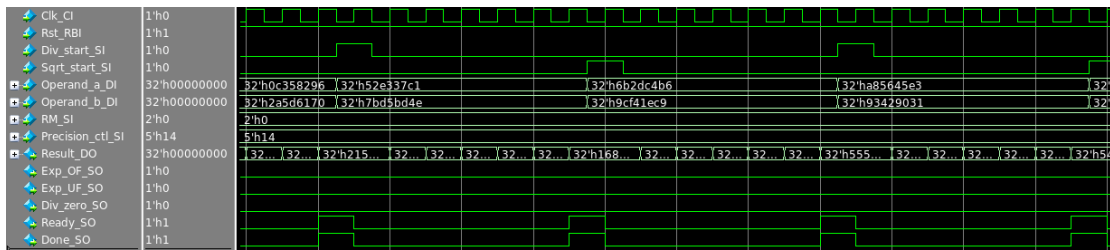


Fig.4 The waveform of division and square root

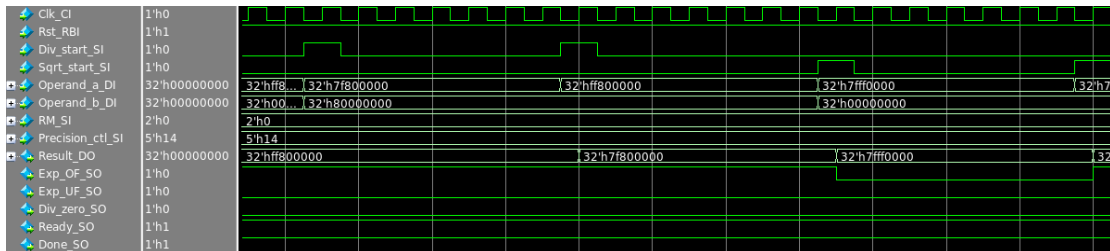


Fig.5 The waveform of special cases

(2) Precision\_ctl\_SI=10, Latency=5 clock cycles

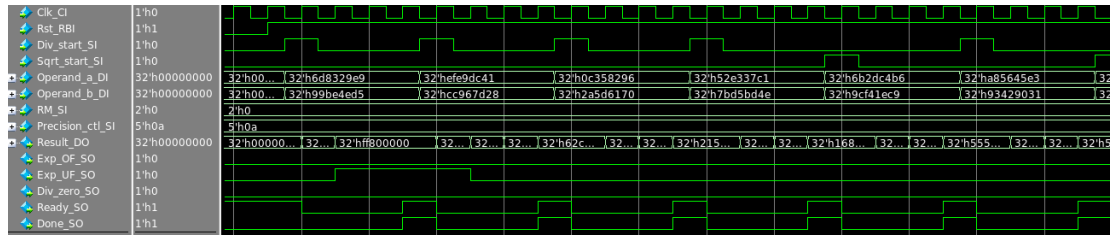


Fig.6 The waveform at the beginning with all inputs delayed by half of clock period

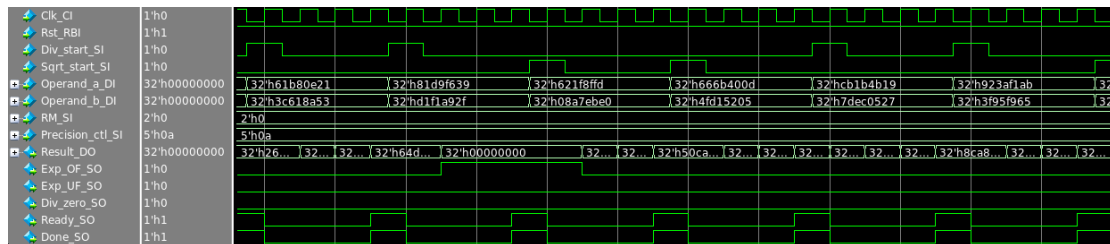


Fig.7 The waveform of division and square root

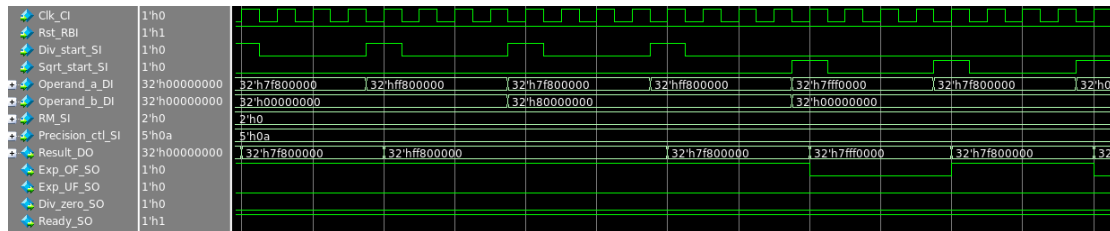


Fig.8 The waveform of special cases

(3) Precision\_ctl\_SI=10, Latency=2 clock cycles just for testing special cases

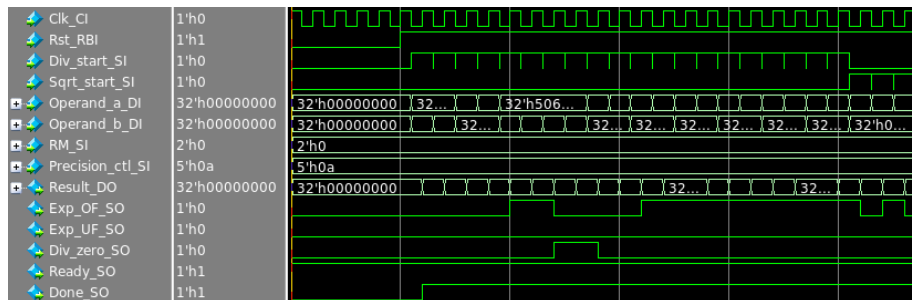


Fig.9 The waveform of special cases

## 8. Synthesized results

UMC65nm process technology was used for synthesis.

Tool: Design Compiler

Operating Conditions: uk65lscllmvbbl\_108c125\_wc

Library: uk65lscllmvbbl\_108c125\_wc

Input\_delay :1ns

Output\_delay:1ns for the original design, 0.8ns for others

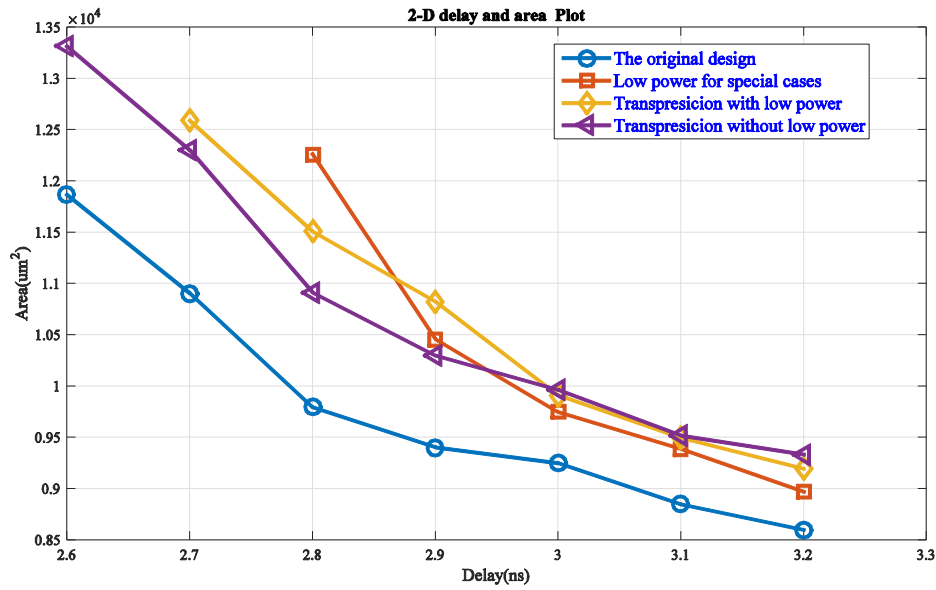


Fig.10 The sweep synthesized results

\*\*The area of a gate equivalent is  $1.44\mu\text{m}^2$

\*\* The direct synthesis will produce [10167um²@2.8ns](#) for the original design with the same constraint, about 7KGE.

## 9 Power estimation

UMC65nm process technology was used for power estimation.

Tool: Primetime

Operating Conditions: uk65lsc1lmvbb1\_108c125\_wc

Library: uk65lsc1lmvbb1\_108c125\_wc

Estimation Conditions: Postsyn without spef.

SDC: from Design Compiler with the clock period of 2.8ns.

Fig.11 plots the results for 23 special cases (23 vectors). 23 special cases cover all the list special cases in Section 4, which is defined in testbench. We just estimate the cases of latency=2 and 8 clock cycles. From Fig.1, it can be concluded that the used technique is very efficient in low power for special cases. When latency = 8 clock cycles, it can achieve more than 40% energy savings. The results of special cases will be ready after 2 clock cycles with low power. We can compare the needed energy @latency=2 clock cycles and @ latency=8 clock cycles. For transprecision with low power, the ratio of the needed energy @latency=2 clock cycles and @ latency=8 clock cycles is 27%. The power reduces with the increase of latency. This is because the needed energy for an operation can be seen to be constant, ignoring the leakage power and other logic. Why not reduce rapidly? It is because that the FSM used for control is wake up @ latency=8 clock cycles.

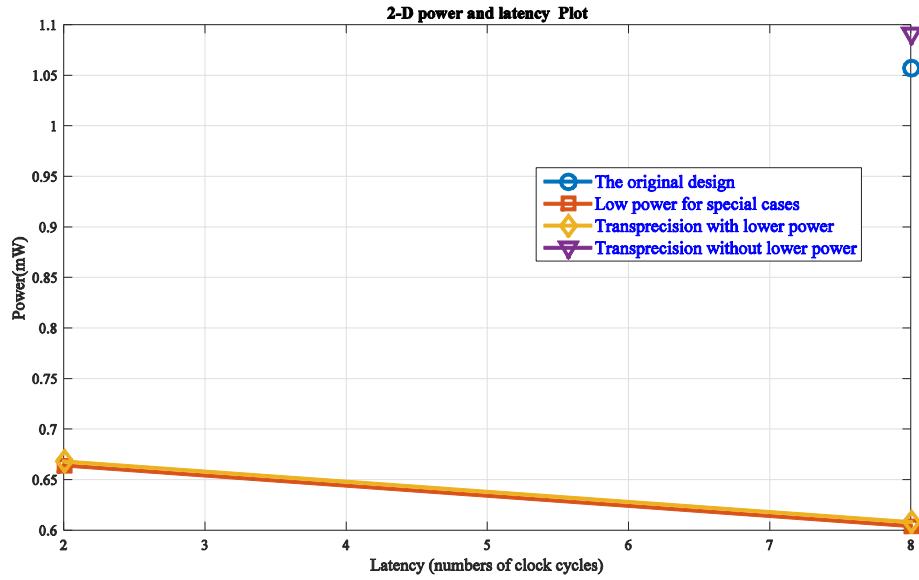


Fig.11 Special cases

Fig.12 plots the results for 1000 vectors including 23 special vectors. Low power control will introduce 14% more power for single precision. Lower power control will introduce about 4% more power for transprecision. Reducing the precision can reduce the needed energy. The needed energy of the precision with 5-cycle latency is about 61% of that of the precision with 8-cycle latency for an operation.

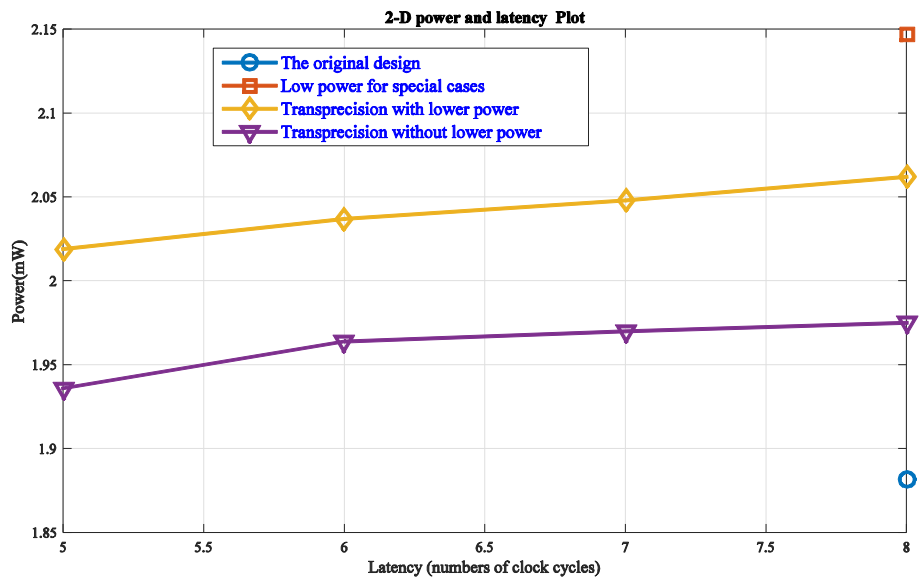


Fig.12 All cases