Parametric, Floating Point Arithmetic Logic Unit

by Nathan Sketch, Saher Al-Khrayyef, Raymon Benjamin

Objectives

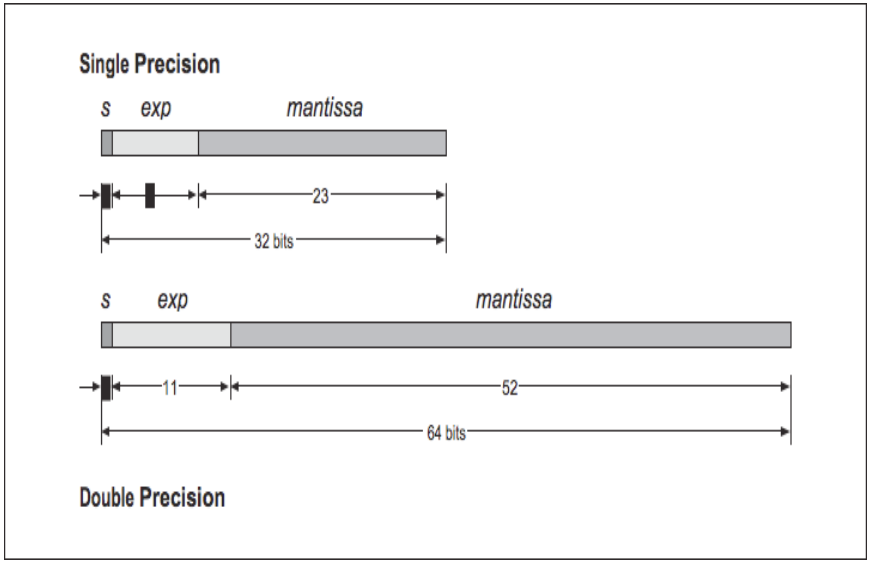
- Design Parametric Floating Point Arithmetic Logic Unit
  - ❖ Addition
  - ❖ Subtraction
  - ❖ Multiplication
- Interface the FPU to a ARM processor via an AXI4 lite bus.

IEEE single and double precision floating point representation

This figure shows the IEEE single and double precision

floating point format.

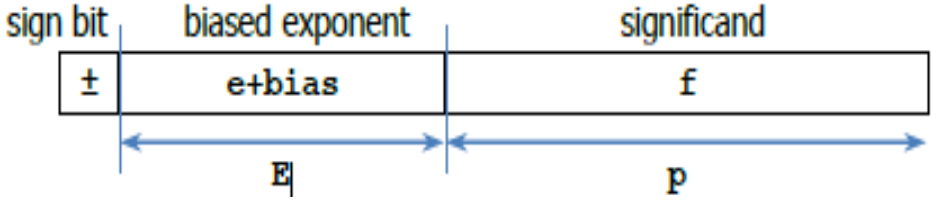
- Where s is the sign bit.
- Exp is the exponent, it depends on the binary number size.
- Mantissa, also it depends on the binary number size.
- Bias single is 127 and double is 1023



IEEE-754 Standard Representation

- The representation is as follows:

$X = \pm 1.f \times 2^{e}$



Significand: f is the mantissa. We should add 1 to the beginning of the mantissa before start addition/subtraction.

Significand range should be $[1,2 - 2^{-p}] = [1,2)$

Biased Exponent:

➢ E is the number of the bits.

➢ Bias = 127 or 1023

➢ e= exp - bias

Floating Point Addition/Subtraction Equations

$b1=\pm s1*2\hat{}e1, \qquad s1=1.f1$

$b2=\pm s2*2\hat{}e2, \qquad s2=1.f2$

- $b1+b2=\pm s1*2\hat{}e1 \pm s2*2\hat{}e2$

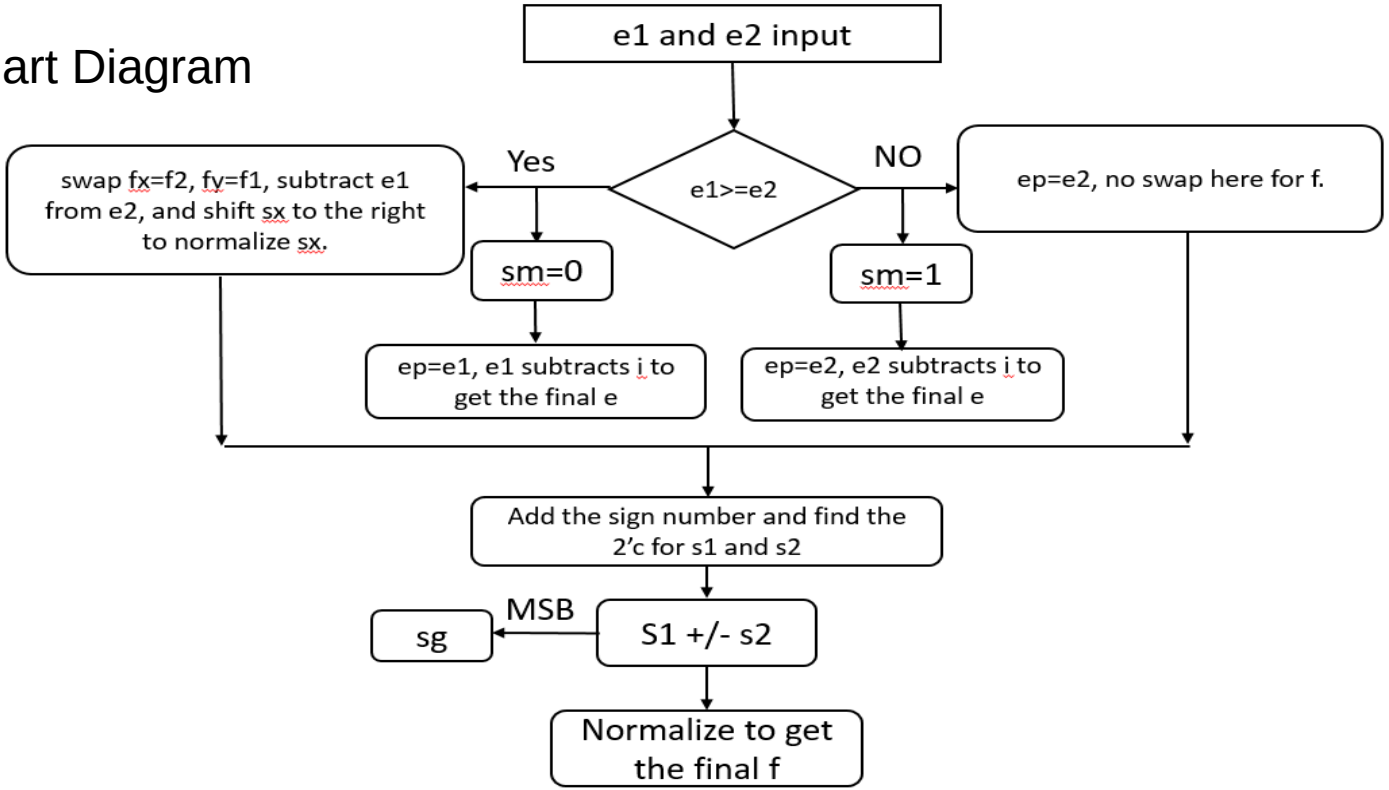
Floating Point Adder/Subtractor Sign

- An add-subtract operation has three sign inputs, a_sign, b_sign, and op.
- We can transform the normal equation into one that better serves the output floating-point format.

-a+(-b)	a-(+b)
-a-b	a-b
-(a+b)	a-b

DESIGN OF FLOATING POINT ADDITION/SUBTRACTION

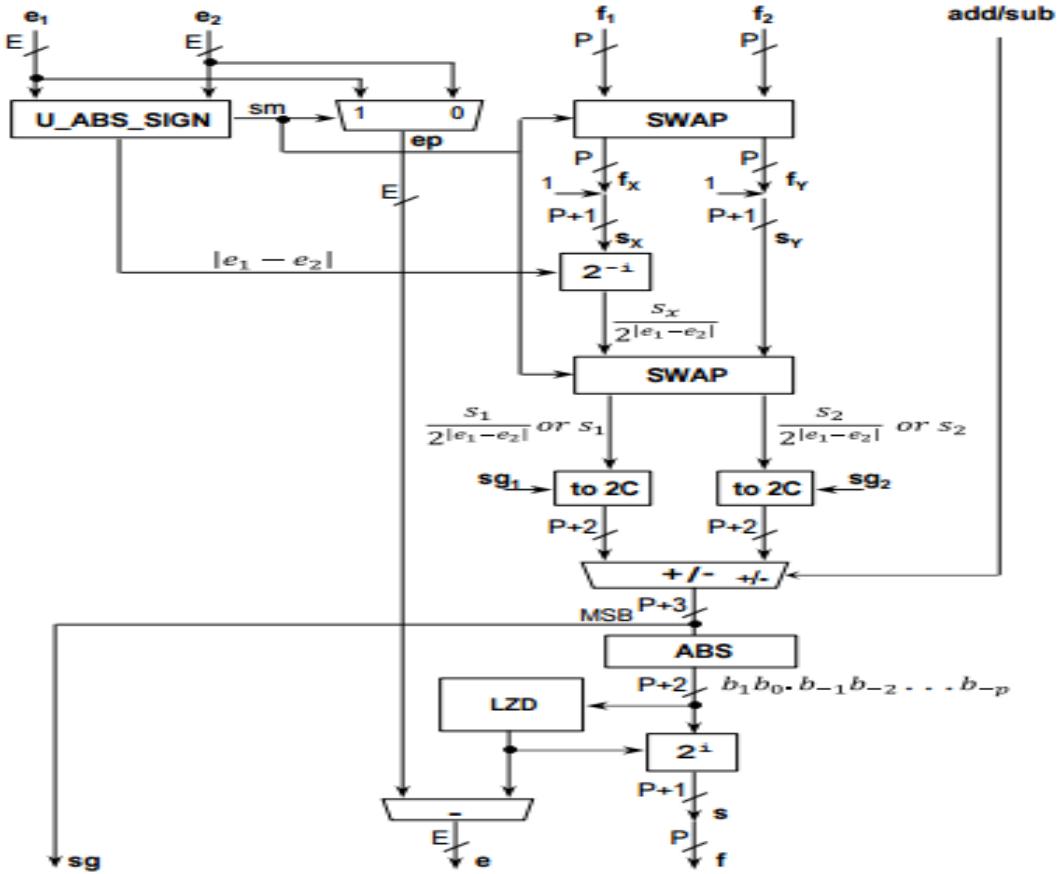
- Flow Chart Diagram



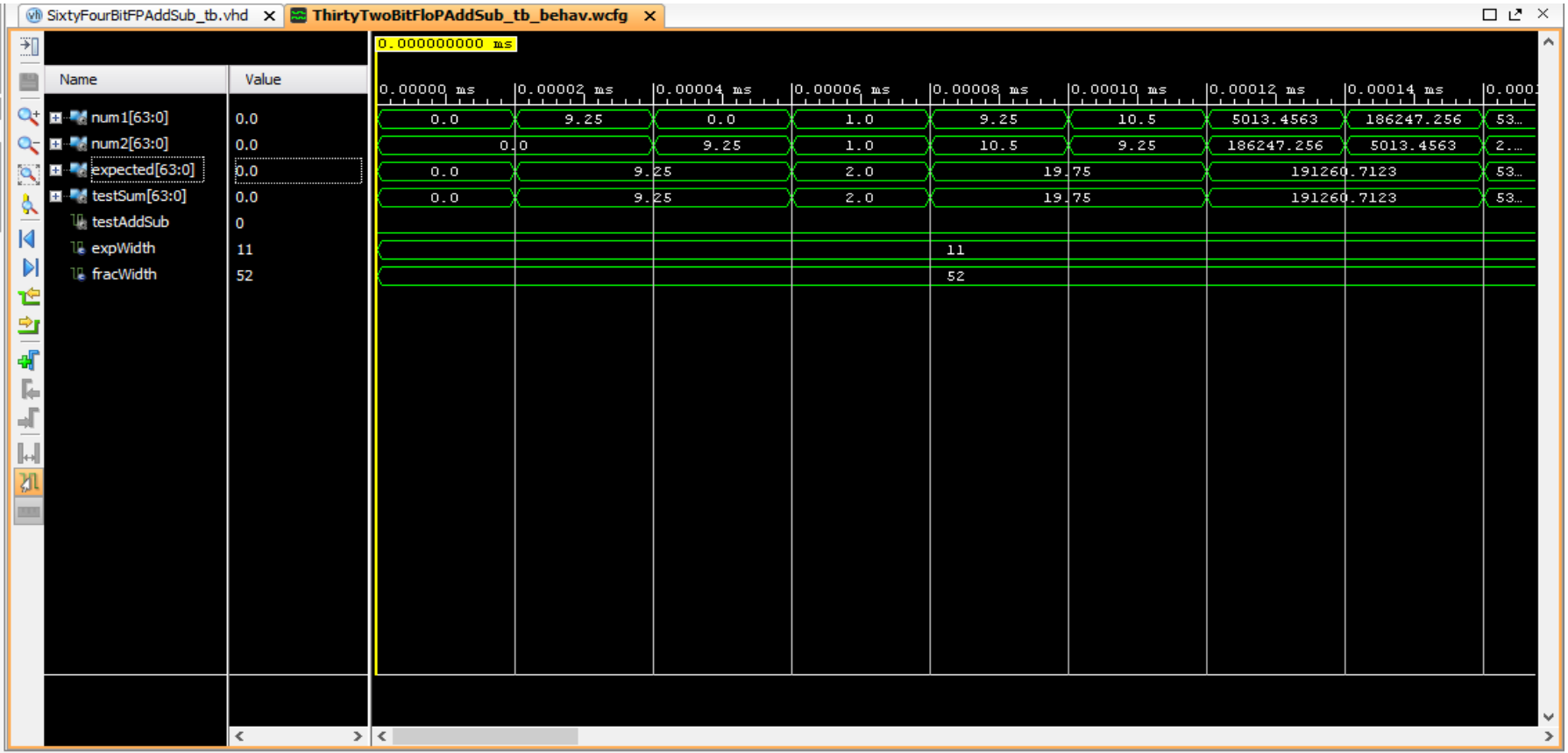
Normalized left shift

- The normalized left shift in the post-normalization step removes leading zeros.

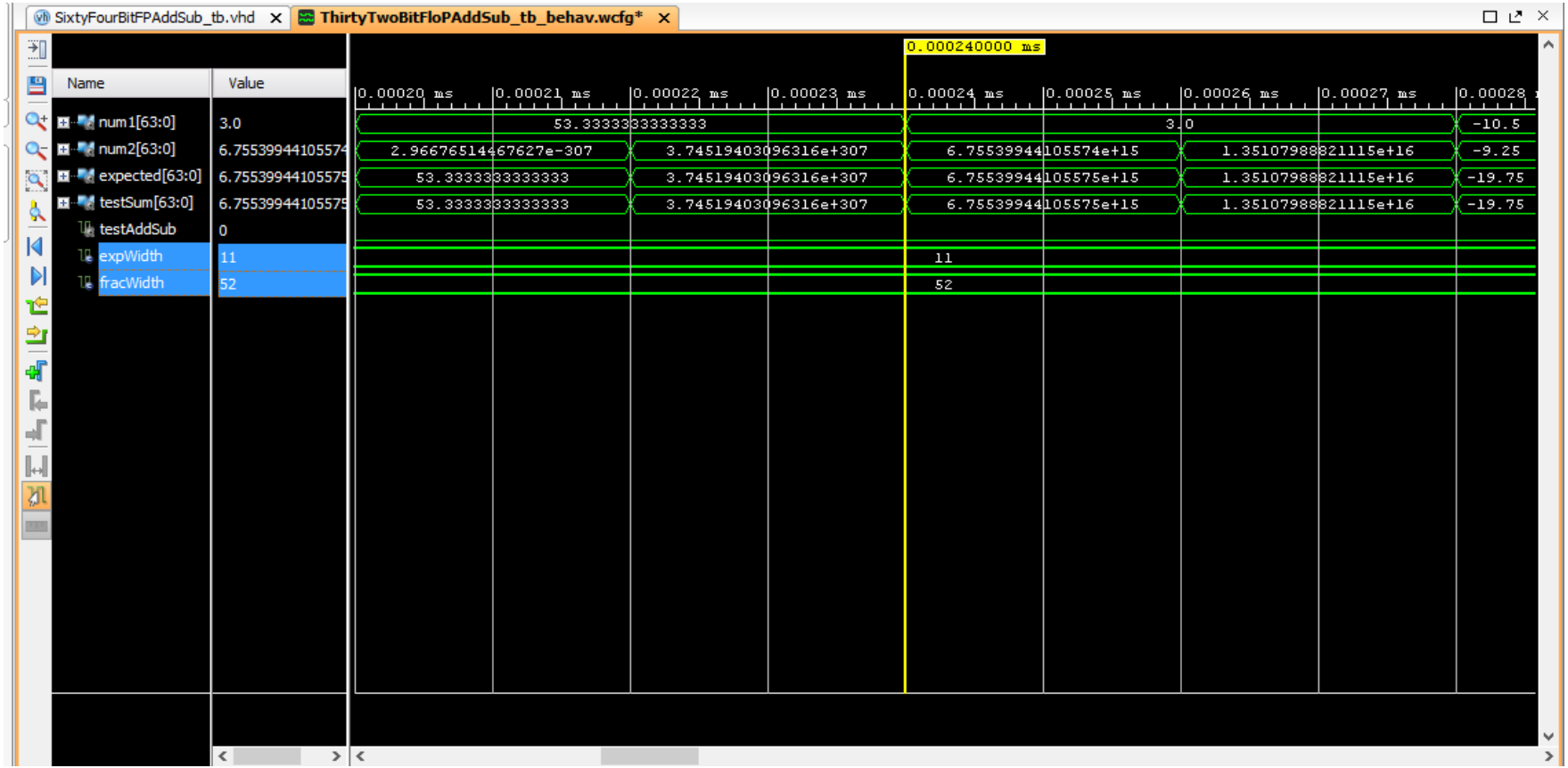
DESIGN OF FLOATING POINT ADDITION/SUBTRACTION …. Cont.



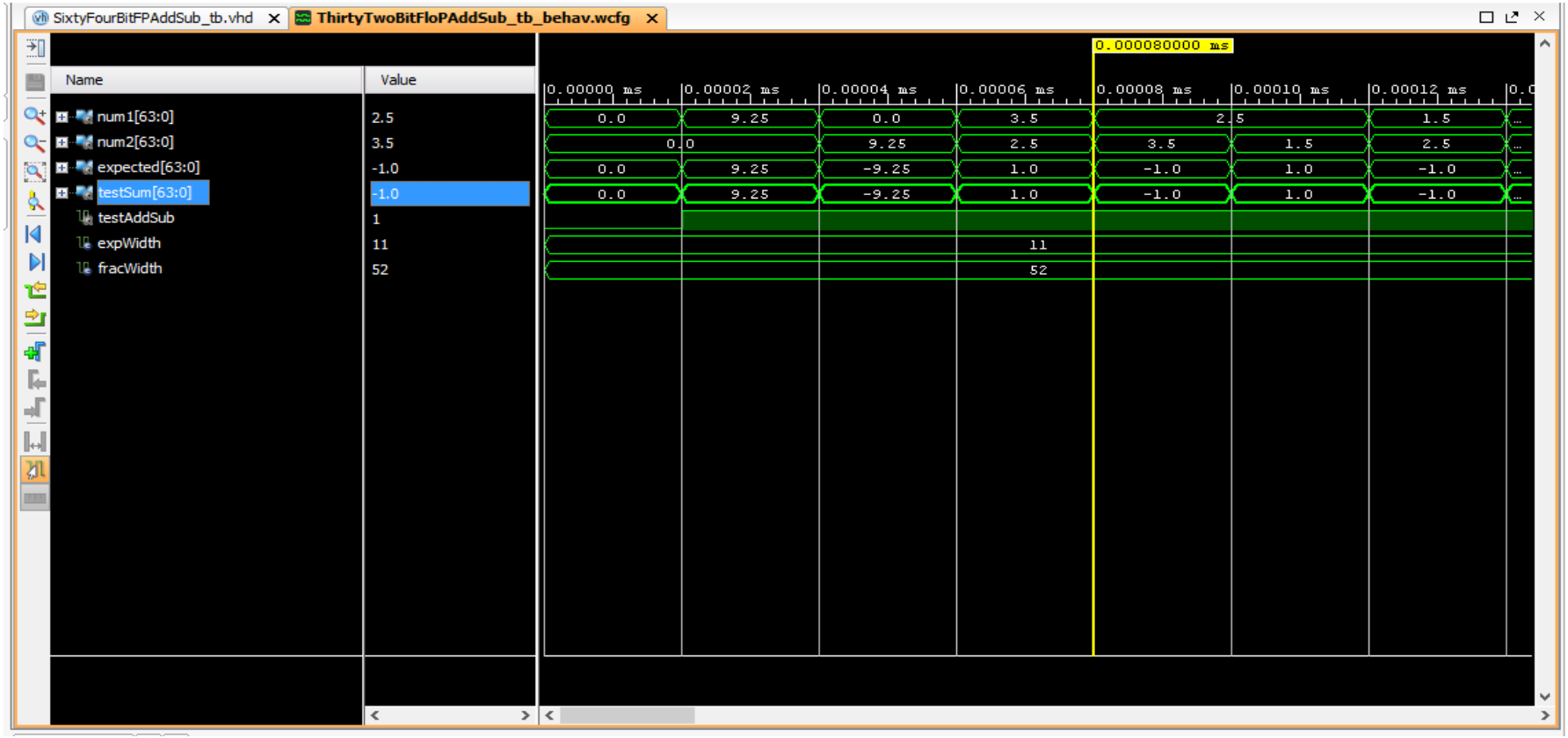
Addition Simulation Part 1



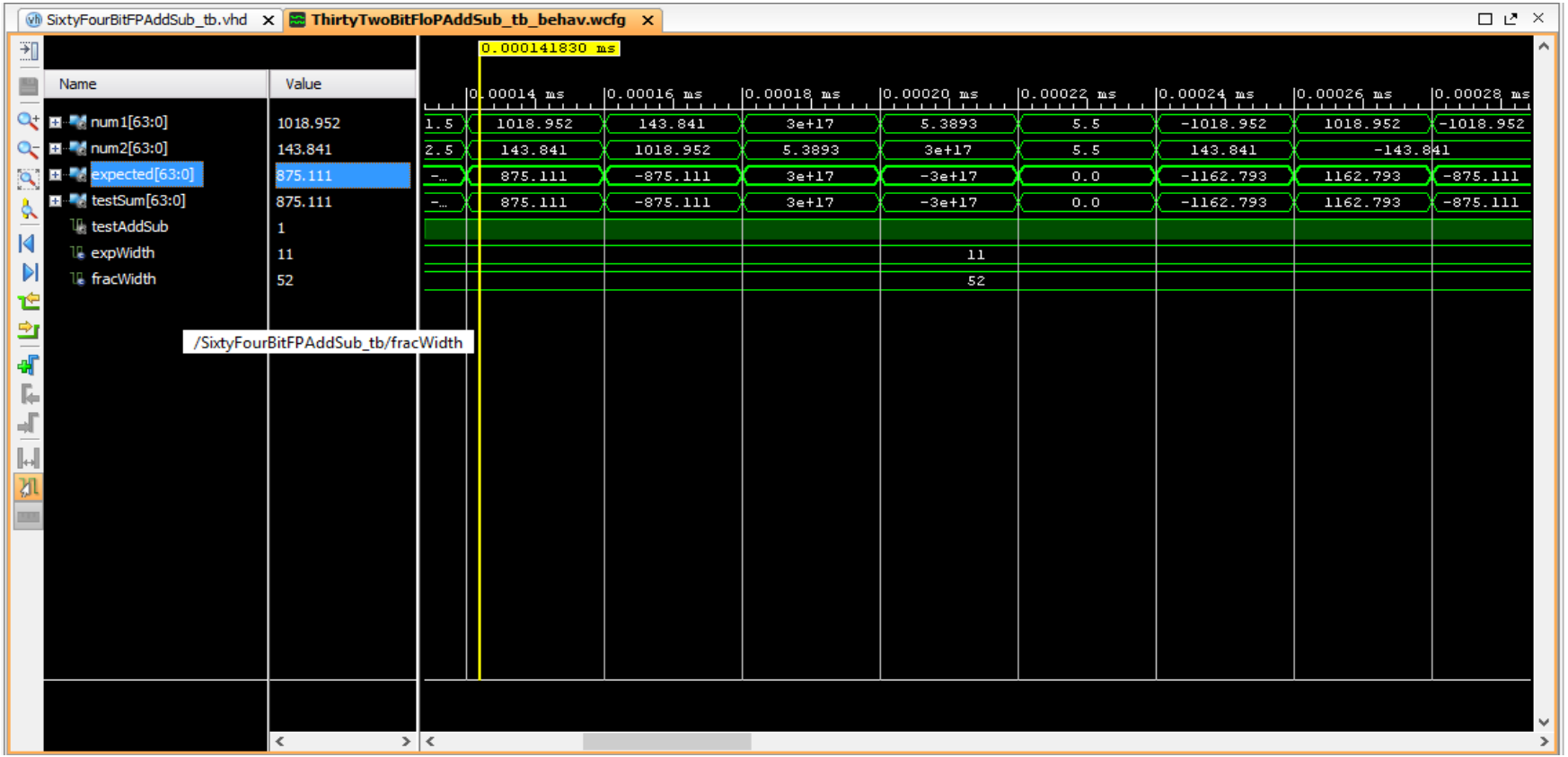
Addition Simulation Part 2



Subtraction Simulation Part 1



Subtraction Simulation Part 2

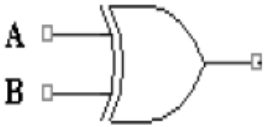


DESIGN OF FLOATING POINT MULTIPLICATION EQUATIONS

- $b1$ = $\pm s1$*2 ^$e1$
  $b2$ = $\pm s2$*2^$e2$
- $b1 \times b2$ = ($\pm s1$*2^$e1$) × ($\pm s2$*2^$e2$) = $\pm(s1 \times s2\ )$2^($e1$+$e2$)
- s = ($s1 \times s2\ ) \in [1,4)$.

DESIGN OF FLOATING POINT MULTIPLICATION

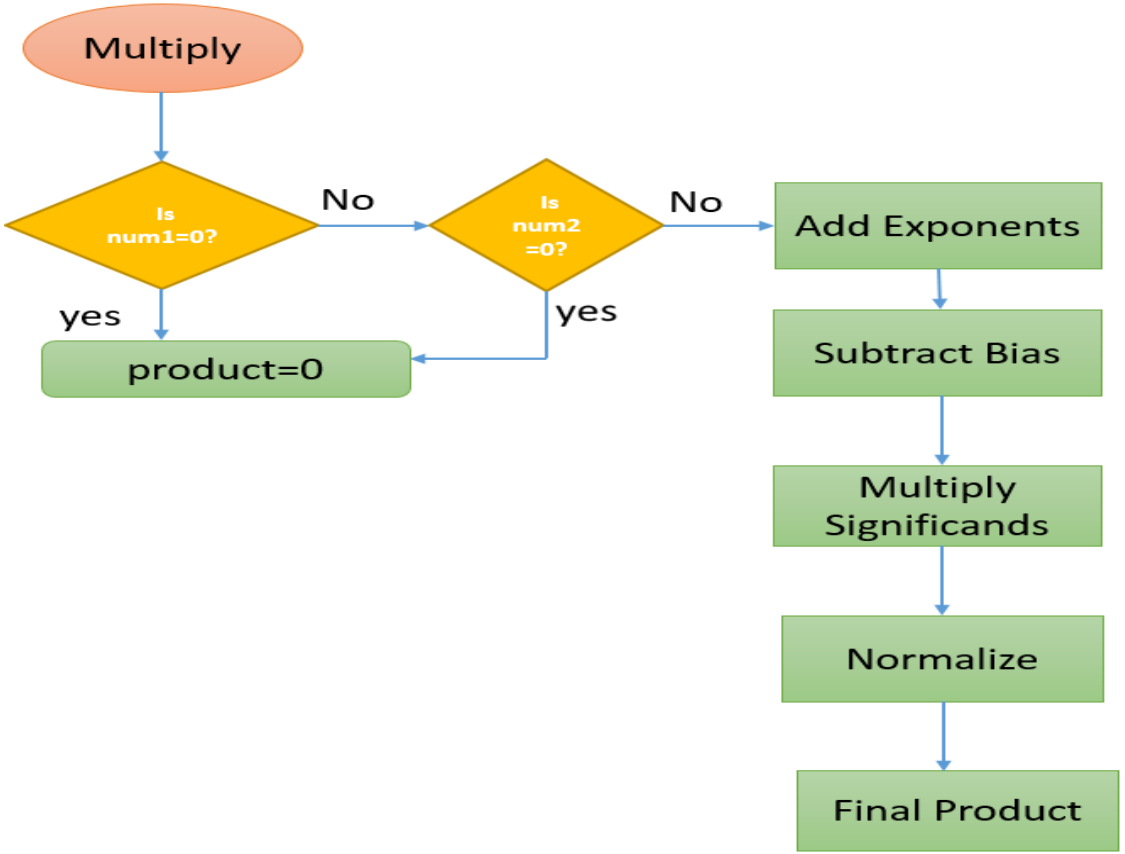
- Sign Bit Calculation
- Multiplying two number’s result is a negative sign if one of the multiplied numbers is of a negative value. By the aid of a truth table we find that this can be obtained by XORing the sign of two inputs.



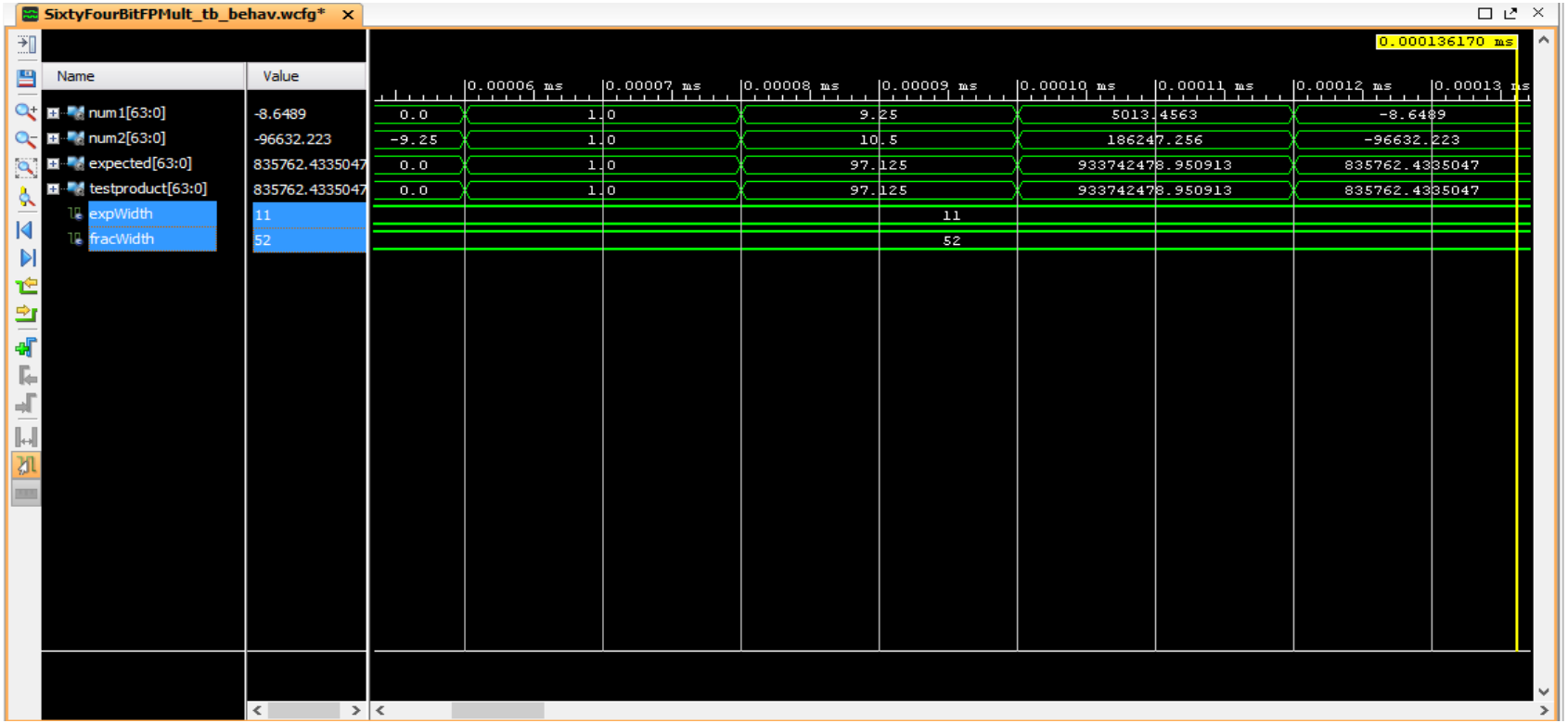
A	B	Y=A⊕B
0	0	0
0	1	1
1	0	1
1	1	0

DESIGN OF FLOATING POINT MULTIPLICATION …. Cont.

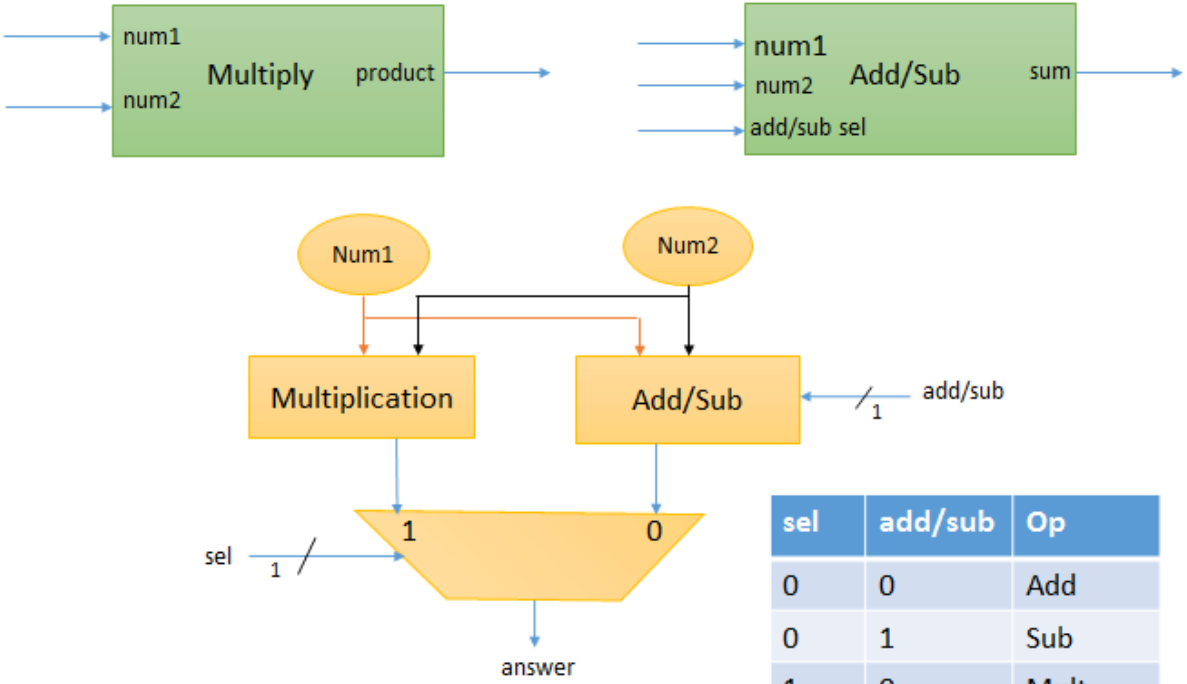
Flow chart Diagram



Multiplication Simulation

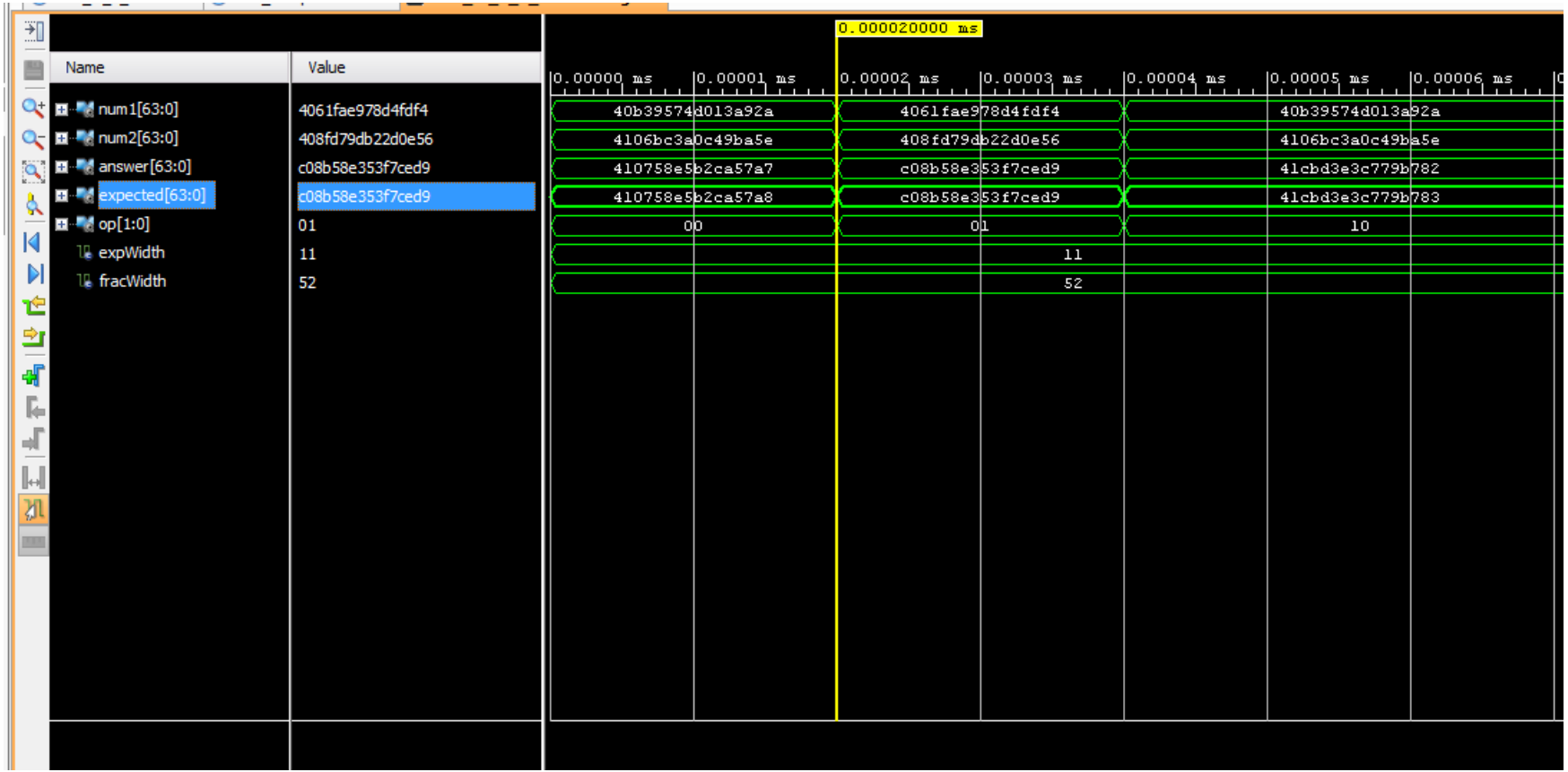


Floating Point Top Circuit

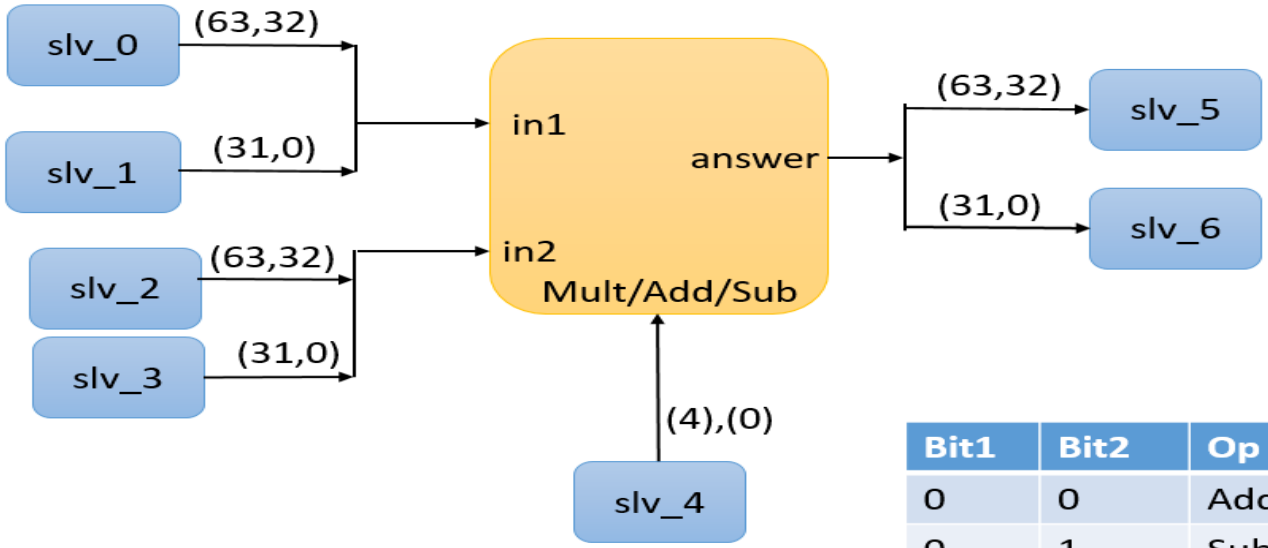


sel	add/sub	Op
0	0	Add
0	1	Sub
1	0	Mult
1	1	Mult

Top Floating Point Arithmetic Circuit Simulation

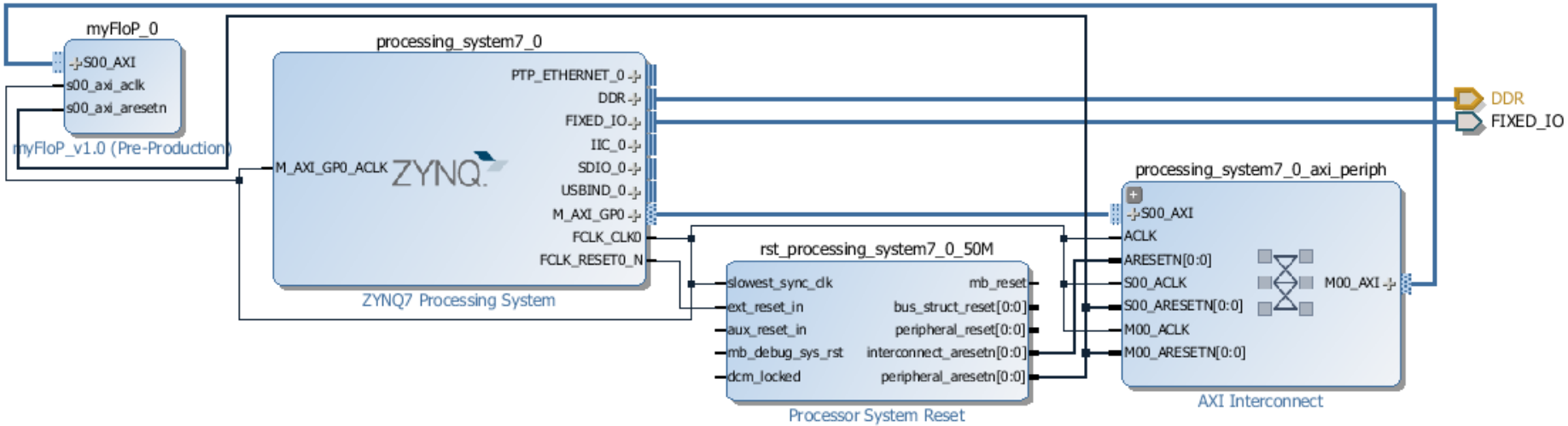


Addition/Subtraction/Multiplication AXI4 lite Interface



Bit1	Bit2	Op
0	0	Add
0	1	Sub
1	0	Mult
1	1	Mult

IP Block Diagram



Challenges/Improvements

- Challenges VHDL
- Improvements Testbench and Divider

Any Questions?

Thank You