Bangladesh University of Engineering and Technology

Course No: CSE306 - Computer Architecture Sessional
Assignment-2: Floating Point Adder
Group: 2

Date of Submission: January 8, 2023

Group Members:

1. 1905003
2. 1905017
3. 1905023
4. 1905028
5. 1805066

1 Introduction

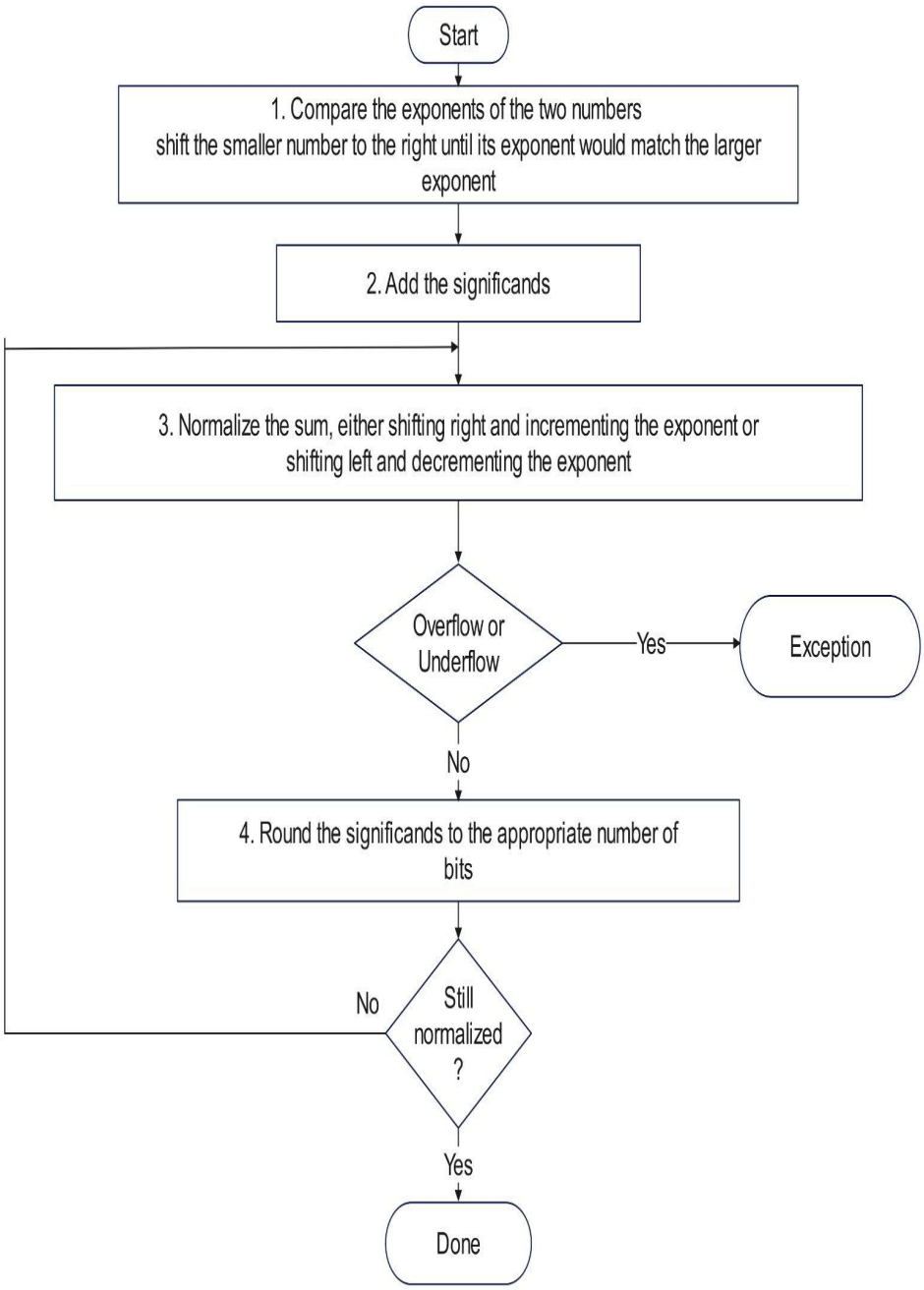
Floating-point addition is the most frequent floating-point operation and accounts for almost half of the scientific operation. Therefore, it is a fundamental component of math coprocessors, DSP processors, embedded arithmetic processors, and data processing units. These components demand high numerical stability and accuracy and hence are floating- point based. Floating-point addition is a costly operation in terms of hardware and timing as it needs different types of building blocks with variable latency. In floating-point addition implementations, latency is the overall performance bottleneck. Standard IEEE754 format is used for floating point representation. In this experiment, we simulate a floating point adder which takes two floating point numbers and produce output in the same format.

2 Problem Specification

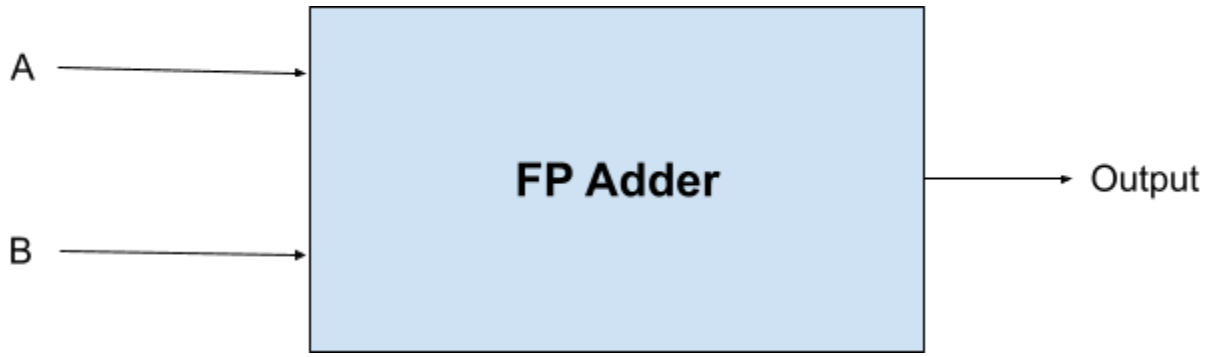
Here, we are required to design a floating point adder that takes two floating point numbers as input and produces their sum, another floating point number as output. Each floating point number will have to be 32 bits long with following representation,

Sign	Exponent	Fraction
1 bit	12 bits	19 bits (Lowest bits)

3 Flowchart of the Algorithm



4 High Level Block Diagram



5 Circuit Diagram

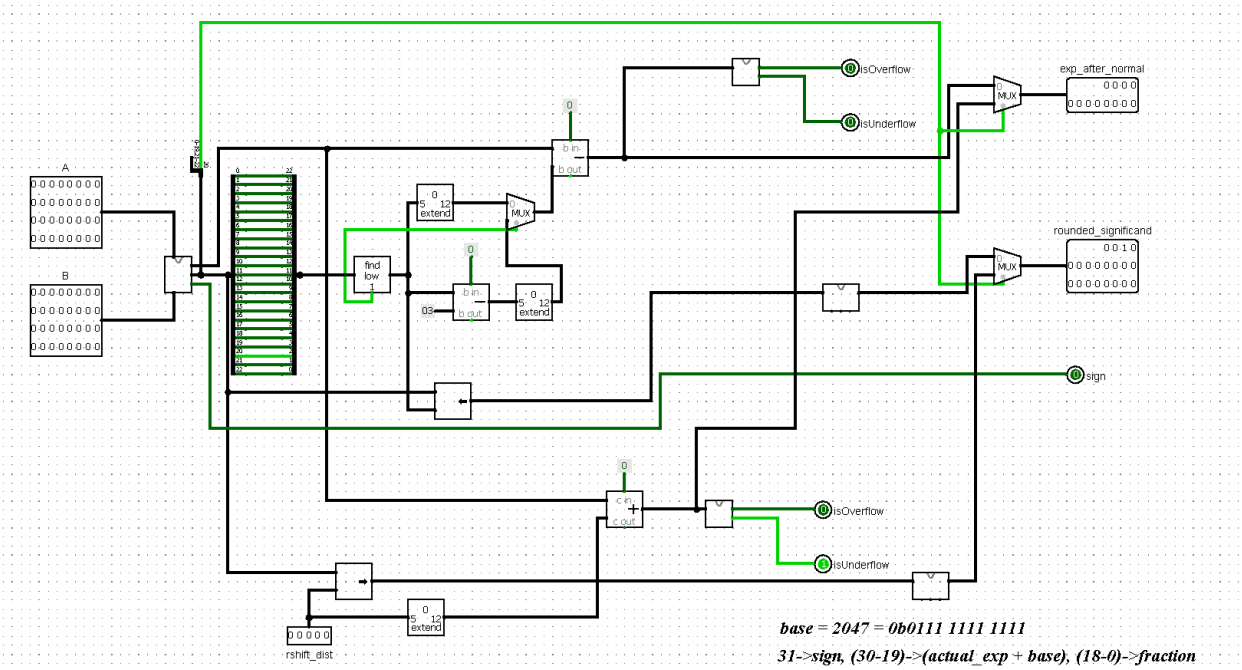


Figure 1: Floating Point Adder

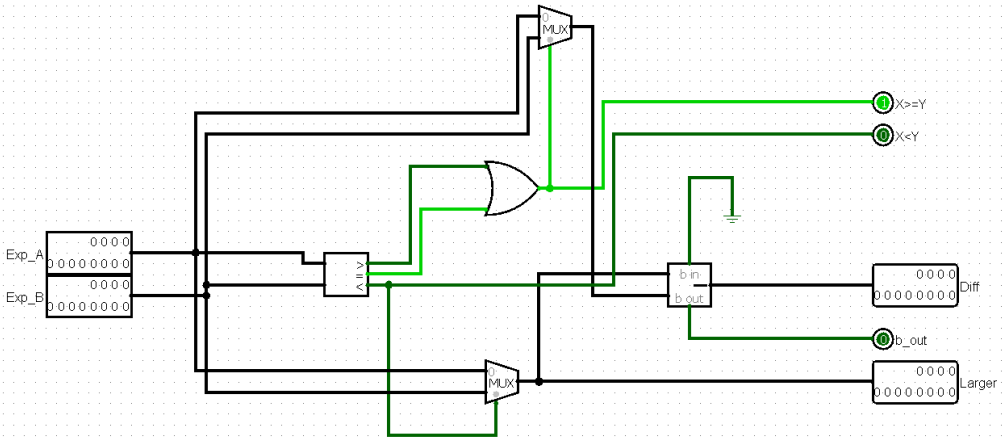


Figure 2: Exponent Comparator

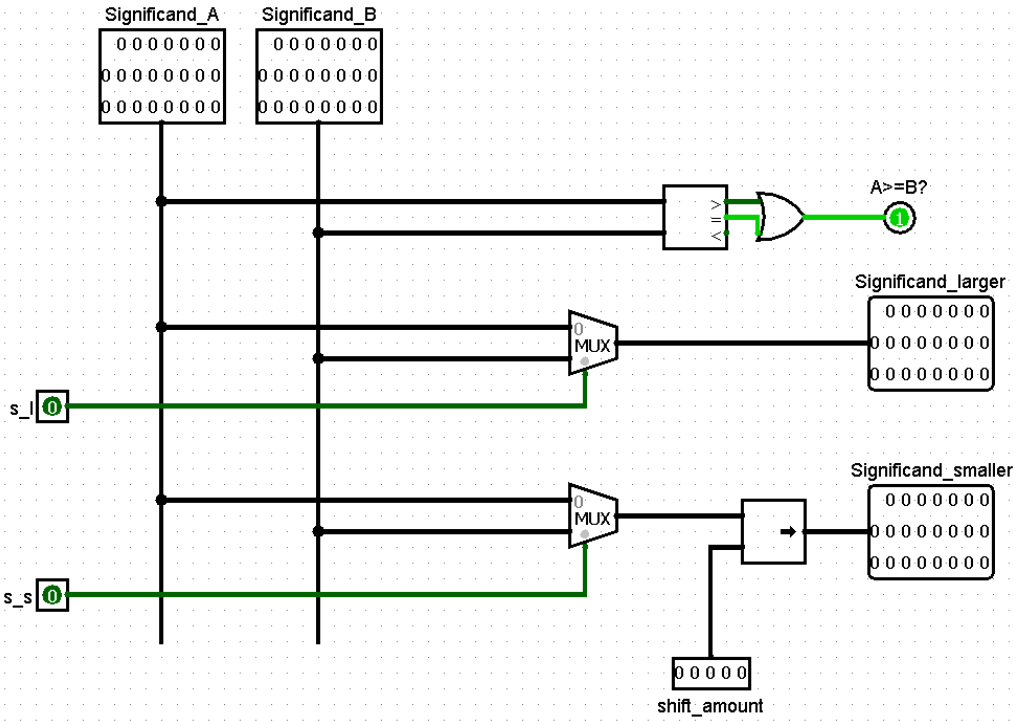


Figure 3: Fraction Shifter

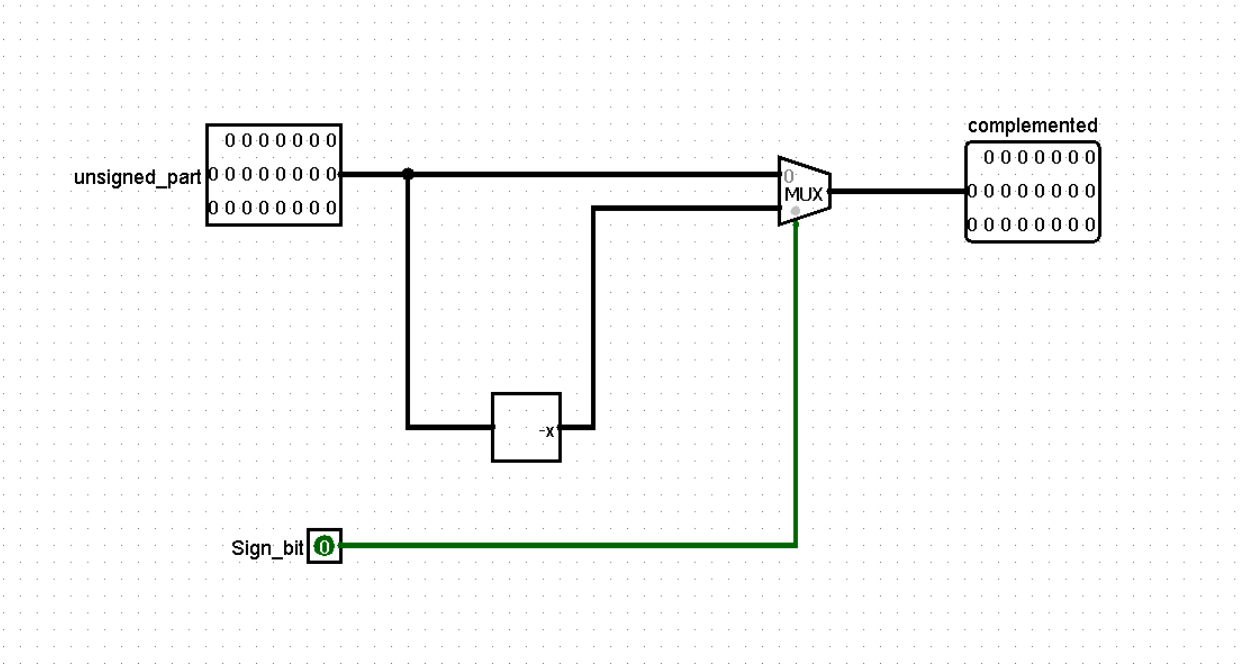


Figure 4: Adjust Input Sign

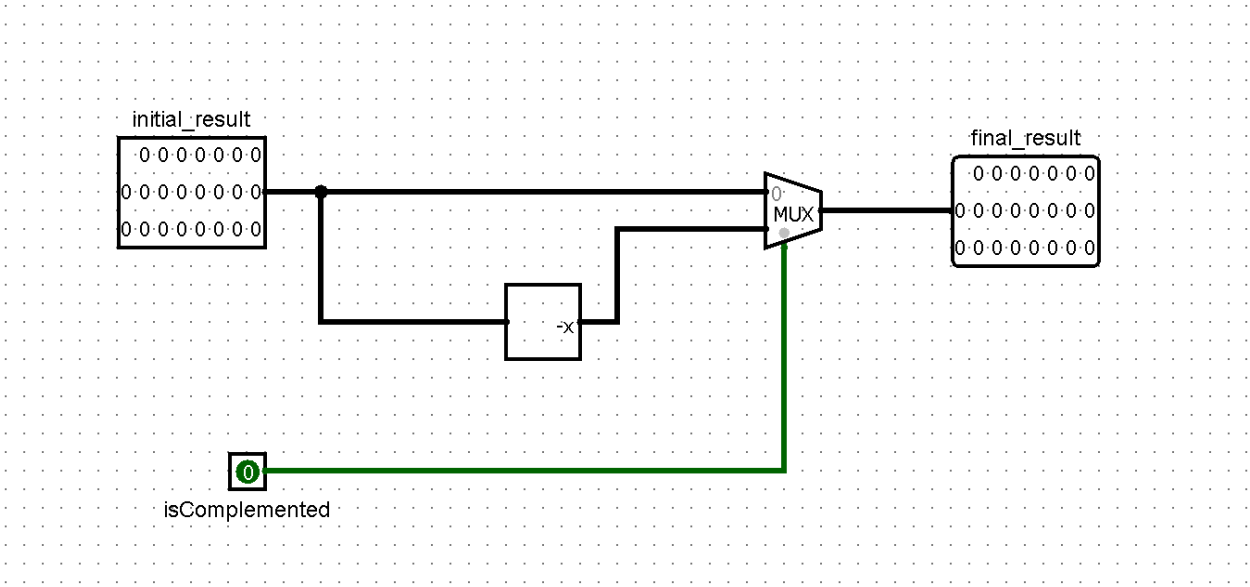


Figure 5: Negation of Result

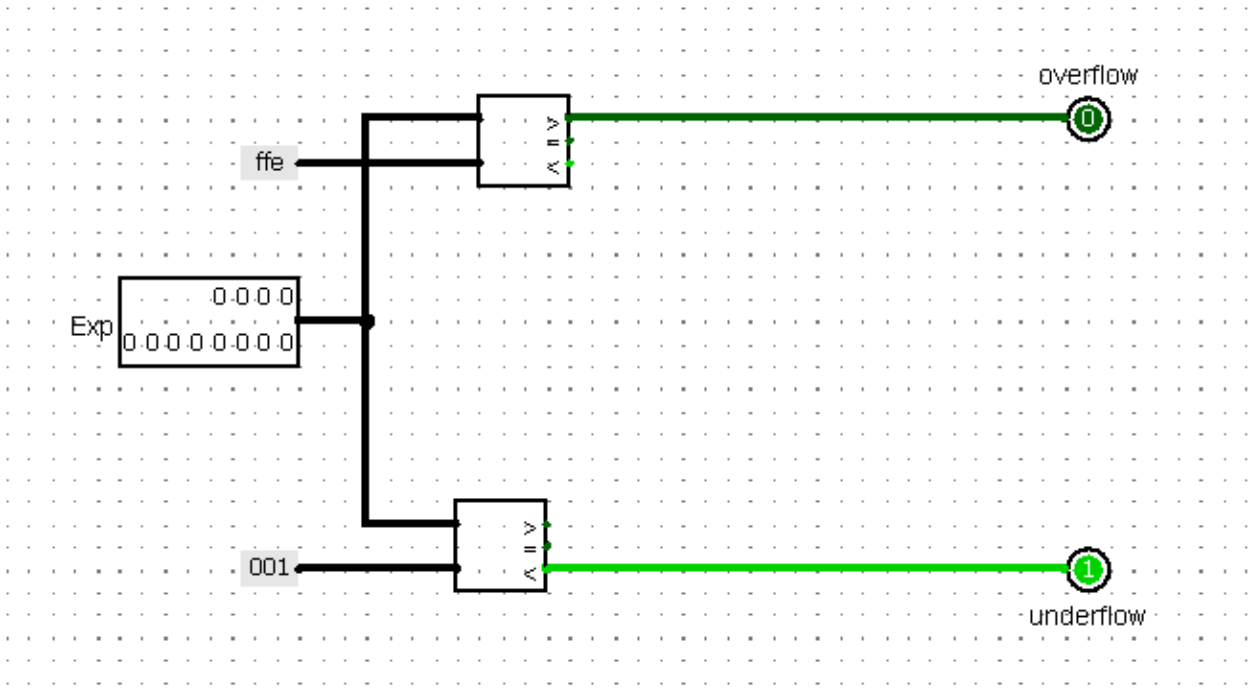


Figure 6: Check Overflow Underflow

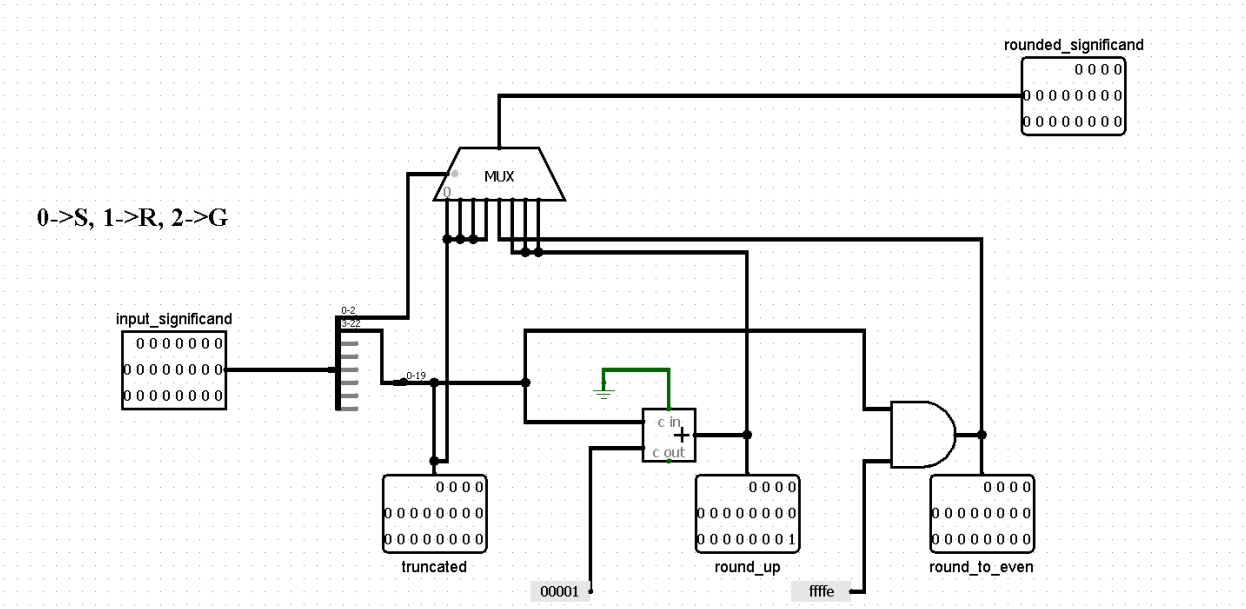


Figure 7: Rounding

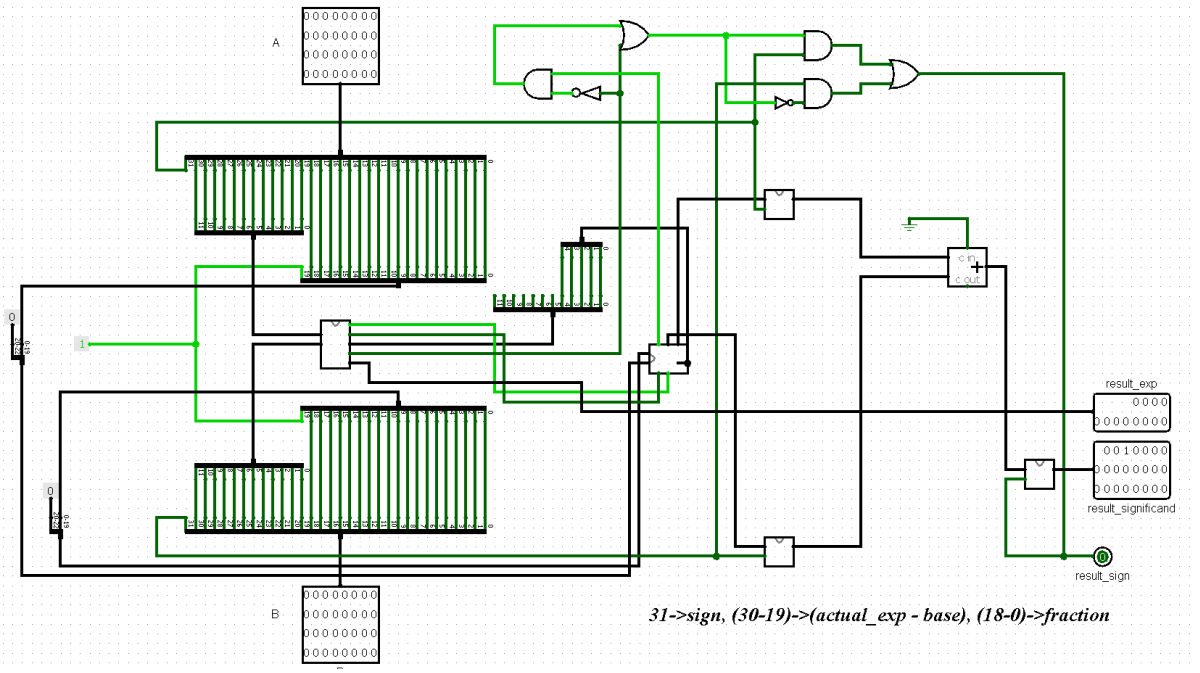


Figure 8: Find Exponent and Add Significant

6 Components

Gate	Gate Count	IC	IC Count
2 bit OR	4	7432	1
2 bit AND	4	7408	1
NOT	2	7404	1

Component	Component Count
23 bit Adder	1
20 bit Adder	1
12 bit Adder	1
12 bit Subtractor	2
5 bit Subtractor	1
20 bit 3x8 MUX	1
12 bit 2x1 MUX	4
23 bit 2x1 MUX	4
23 bit logical left shifter	1
23 bit logical right shifter	2
23 bit Bit Finder	1
5 to 12 bit extender	3

8 Simulator Used

In this experiment Logisim version 2.7.1 was used.

7 Discussion

The floating-point adder has been designed using AND, OR and NOT gates as well as some Logisim components as described in the “Components” section. Optimization of the total number of gates used has also been our concern while designing and subsequently modifying the circuit through many trial and error. For this purpose, different types of multiplexers have been used instead of just the basic logic gates. Overflow and underflow has been indicated with seperate output pins in the circuit. In order to simulate rounding of the normalized sum, 000 is appended at the start of the input significants to make total number of bits 23. The final result is expressed by 12-bit output pin for exponent, 20-bit output pin for the significant which includes the '1' in MSB and 1-bit output pin for displaying sign. Computations for indicating overflow/underflow and rounding had to be performed twice even though only one of them is finally selected through multiplexer. With all these things in consideration, the final design has been as optimized as possible.