CSE 306
Computer Architecture Sessional

Floating Point Adder

Group: 02
Section: A2

Group Members:
1705036
1705037
1705038
1705039
1705044

1 Introduction

A Floating Point Adder is a circuit specifically designed for floating point arithmetic addition operation. Floating point numbers are used to represent numbers with fractions in computer arithmetic. The binary coding of floating-point real numbers is composed of three fields : 1) Sign, 2) Exponent, 3) Fraction. The term Significand is (1 + fraction). The actual number can be represented in normalized form by:

$$(-1)^{Sign} \times (1 + Fraction) \times 2^{(Exponent-Bias)} \quad (1)$$

Thus, the addition of floating point numbers in comparison to integer values is a more complex function which requires a specifically designed hardware which is the floating point adder.

In order to add two floating point values, a traditional floating point adder has to follow a series of steps. It assumes that the inputs are given in normalized form. Firstly, it compares the exponents and aligns the input numbers. Secondly, it adds the Significands of the two numbers and it normalizes the sum by changing the exponent accordingly. Then it checks whether overflow or underflow occurs to report an exception. Finally the adder rounds the sum and then reports the final output in normalized form.

Our implementation of the floating point adder follows the same principles. It takes two normalized floating point numbers as inputs and outputs the result along with two flags which signal whether overflow or underflow has occured.

2 Problem Specification

In the assignment, we had to implement a floating point adder circuit which takes two floating point numbers as input and provides the sum as another floating point number. Each floating point numbers are 16 bits long.

Figure 1: Bit Representation.

We also had to implement the Overflow/Underflow flags. Truncating the final floating point number output was also necessary.

3 Flowchart of addition algorithm

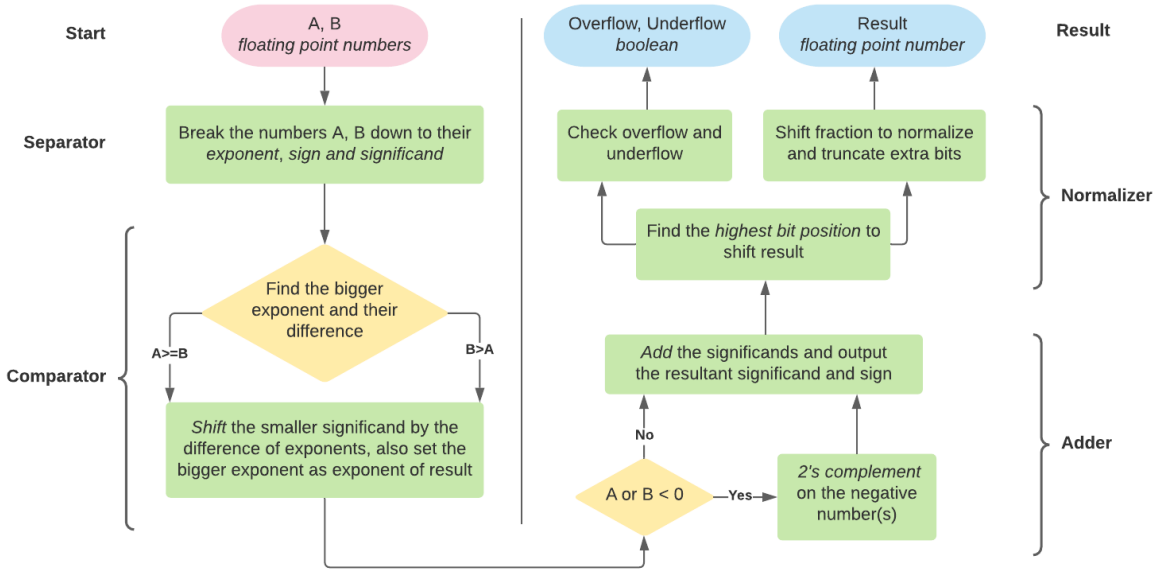

Figure 2: Addition/Subtraction Algorithm

4 Block Diagram of architecture

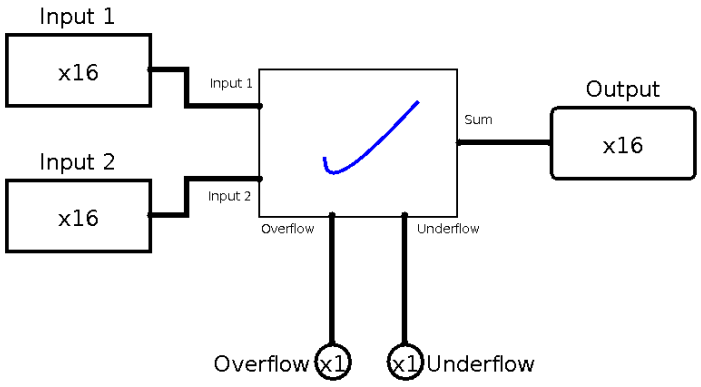

Figure 3: Floating Point Adder

5 Detailed Circuit Diagram

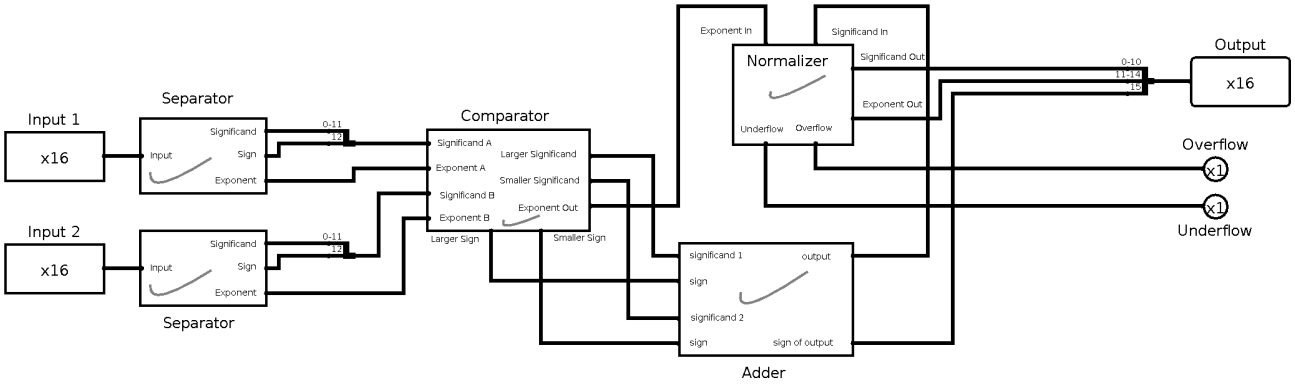

Figure 4: Floating Point Adder Detailed

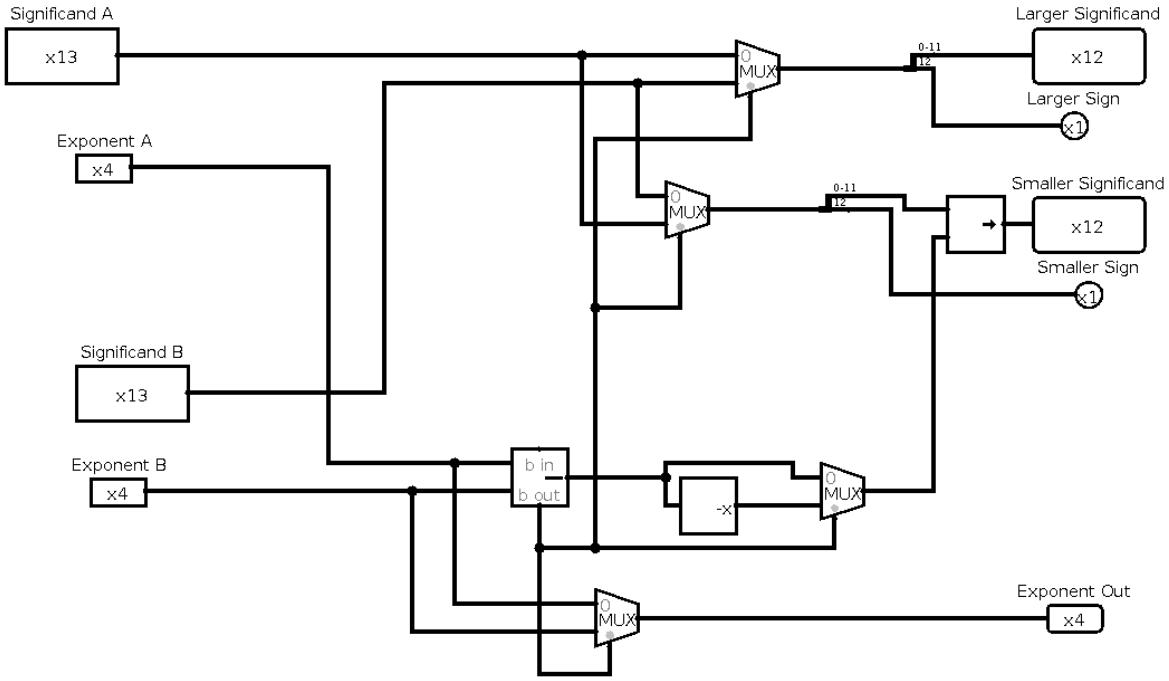

Figure 5: Comparator Block

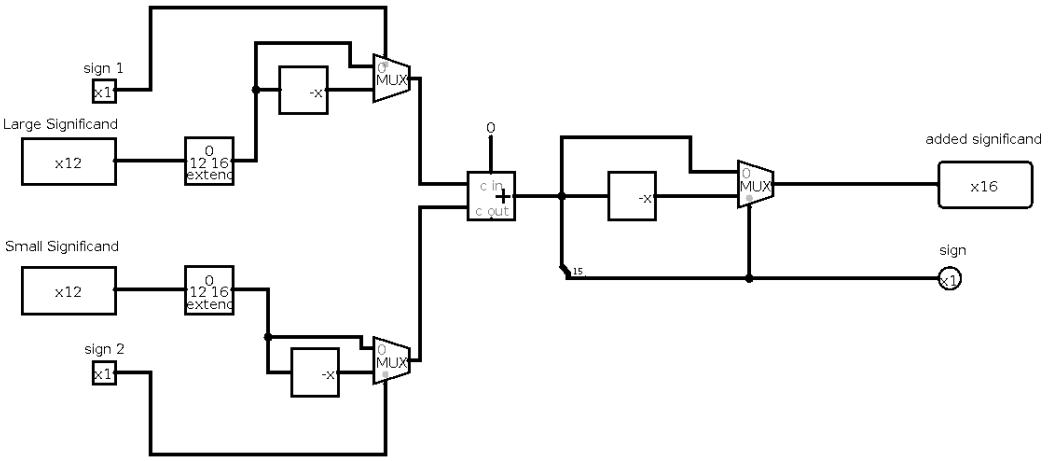

Figure 6: Adder Block

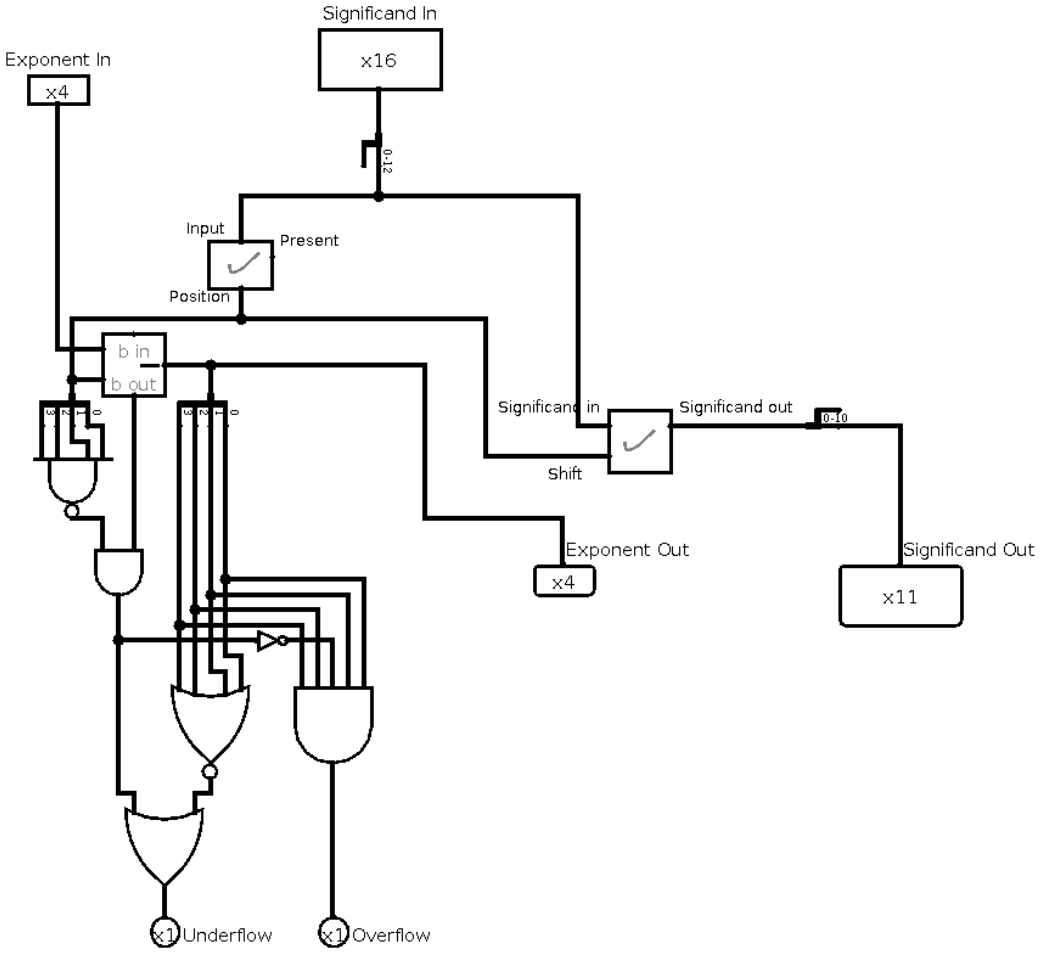

Figure 7: Normalizer Block

6 Required ICs and Components

IC Name	Count
IC 7404	1
IC 7408	3
IC 7432	1
IC 7402	1
IC 7483	7
IC 7486	4
IC 74157	2

Component Name	Count
Bit Shifter	2
Negator	4
Bit Finder	1
Bit Extender	1

7 Software Version

Logisim-Win-2.7.1

8 Discussion

In this assignment, we were tasked with implementing a floating point adder circuit that takes two floating point numbers and outputs the resultant floating point number in normalized and truncated form. In the Normalizer block, the Overflow/Underflow conditions were also checked and shown in the output. Due to the complicated nature of the calculations, running test cases and debugging was tougher than usual logical circuits. However, various test cases were checked and the circuit was made as efficient as possible.