32-bit Single Precision Floating Point Adder

Implementation Details:

The entire 32-bit Floating Point Adder will be designed in VHDL. The design follows a bottom-up approach i.e all the building blocks (Adders/Muxes, etc) will be designed first and then instantiated in the top module.

The project uses the IEEE-754 Single Precision floating point format. The following table lists how the 32-bit floating number is divided into 3 parts viz. Mantissa, Exponent and the Sign bit -

Representative Name	Number-of-bits
Mantissa	23
Exponent	8
Sign	1

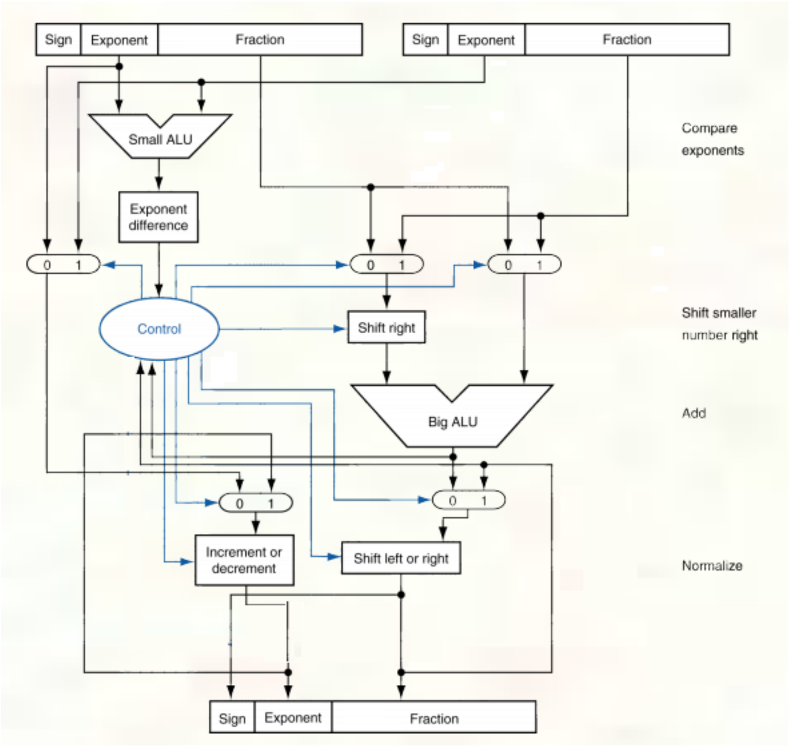
Using the above format, the floating point number can be expressed as follows -

$$\mathbf{(-1)^{s}\ 1.f \times 2^{exp}}$$

In the IEEE-754 format the first bit of mantissa is assumed to be 1 and the remaining part is the fractional part.

Block Diagram of the adder -

The diagram given below gives a simple architecture of the Floating-Point adder -



The above hardware implementation is taken from the following book -
"*Computer Organization and Architecture 3rd Edition – Hennessy and Patterson*"

To simplify the implementation, we have ignored the use of rounding hardware at the end of the FP-Adder.

Building Blocks:

I. **ALU** – The adder uses to ALU (small and big) one for the mantissa bits and the other for the exponent bits. The ALU is responsible for finding the bigger number such that the normalization can take place easily.

II. **Shifters** – The shifters are used to perform the normalization. Depending on the shift value generated by the Control unit, the Mantissa/Exponents are shifted accordingly.

III. **Control Unit –** The Control Unit uses inputs at various stages and generates the expected control values for shifters/ALUs.