32 bit Floating Point Adder

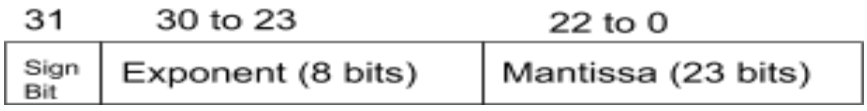
Representation of Floating Point Numbers in Binary

A floating point number can be represented in the form of 32 bit binary number (Single Precision) as below.

$N=Fx2^E$, Where F is fractional part (mantissa) and E is the exponent.

It is necessary to normalise mantissa i.e. it should be of the form 1.xxxxx

Most significant bit of normalised mantissa is always 1 so it can be ignored while storing the number.



Floating point adder is implemented as 3 stage pipeline circuit, when given two positive number as input, gives their sum as output.

Stage 1 : Compare and Shift

This stage compares the exponents of two operands and shift one of the mantissas in order to make exponents same.

Stage 2 : Addition

This stage adds mantissa of operands.

Stage 3 : Normalisation

This stage normalises the output of addition stage.

Future Work : Including Negative Numbers

Logic : If XOR of sign bits is 1 then change mantissa of smaller number by its 2's complement.

Reference

http://cstl-csm.semo.edu/xzhang/Class%20Folder/CS280/Workbook_HTML/FLOATING_tut.htm

Saumya Shah