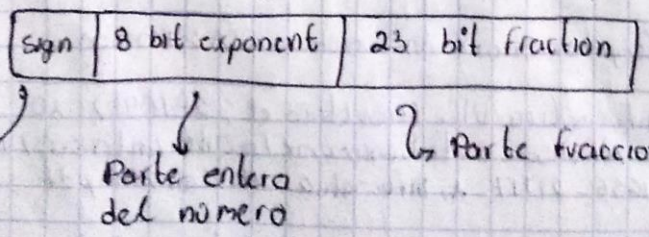
Díaz Hernández Marcos Bryan
317027253

Tarea 3

Estandar IEEE 754

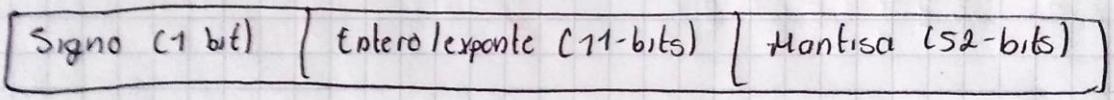
- Float (32 bits): el formato utiliza 1 bit de signo, 8 bits para el exponente y 23 bits para la representación la fracción del número

El valor es calculado: $(-1)^S \times 1.F \times 2^{(E-127)}$
el signo es positivo en (0) y negativo en (1)
y los 23 bits son parte de la mantisa.

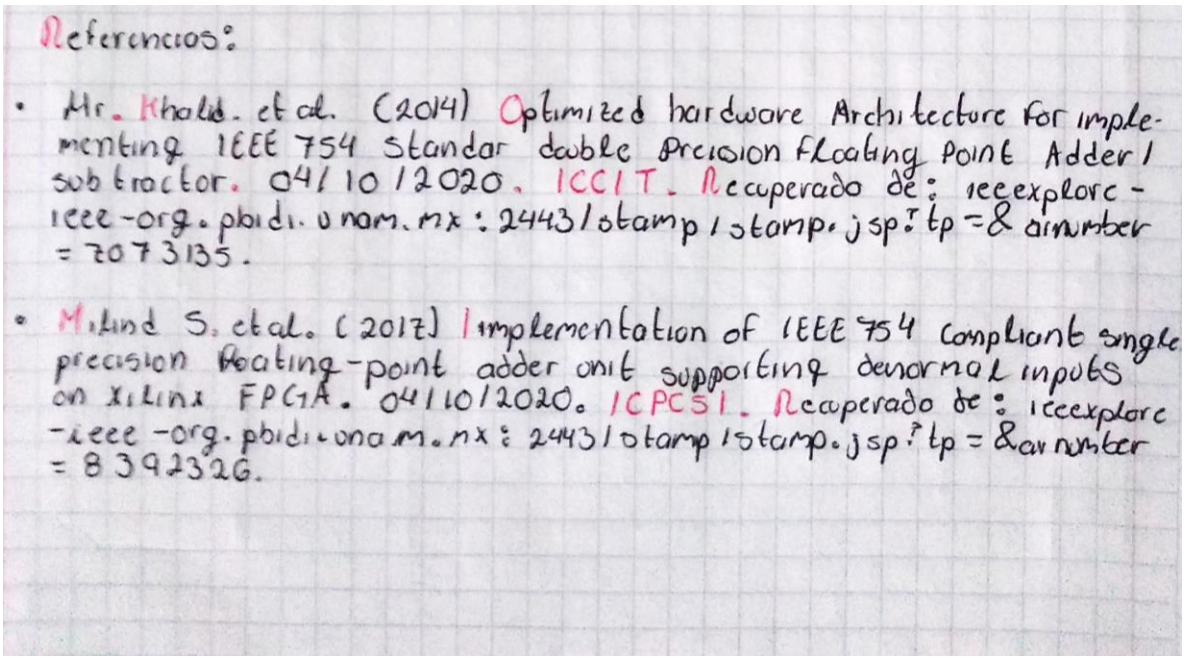


- Punto flotante de doble precisión: el Formato utiliza un bit de signo, 11 bits de exponente y 52 bits de mantisa. El mas 1 bit más significante de la mantisa debera ser 1 y el exponente mayor a 0 y menor que 1023

Calcularlo: $[(-1)^{signo} \times 2^{exponente - 1023} \times 1.\ mantisa]$



En general los dos formatos tienen dos formas de transformarse, de el numero real al formato binario y viceversar.



Referencias:

- Mr. Khalid. et al. (2014) Optimized hardware Architecture for implementing IEEE 754 Standar double Precision Floating Point Adder/subtractor. 04/10/2020. ICCIT. Recuperado de: ieeexplore-ieee-org.pbidi.unam.mx:2443/stamp/stamp.jsp?tp=&arnumber=7073135.

- Milind S. et al. (2017) Implementation of IEEE 754 Compliant single precision floating-point adder unit supporting denormal inputs on xilinx FPGA. 04/10/2020. ICPCSI. Recuperado de: ieeexplore-ieee-org.pbidi.unam.mx:2443/stamp/stamp.jsp?tp=&arnumber=8392326.