Floating Point

Madiba Hudson-Quansah

Contents

Chapter 1

IEEE 754

1.1 Introduction

Definition 1.1.1: Floating Point

A way to represent real numbers in a computer system.

Floating point numbers are represented in scientific notation bitwise under the IEEE 754 specification, i.e.:

$$\pm d.\text{fraction} \times 2^{\text{exponent}}$$

Floating point numbers should be normalized, i.e. exactly one non-zero digit should appear before the decimal point. This number can be any digit in the number base except zero. For example in decimal the possible numbers are 1 to 9, and in binary the possible numbers are 1.

A floating point number is represented in the following way: Where:

Sign Bit	**Exponent**	**Fraction / Mantissa**

Sign Bit - 0 is positive and 1 is negative.

Exponent - The exponent is stored in a biased form. The bias is the value that is added to the exponent to get the actual exponent. Very large numbers have large exponents and very small close to zero numbers have negative exponents. The more bits in the exponent field increases the range of numbers that can be represented.

Fraction / Mantissa - The fraction is the part of the number that is multiplied by the base raised to the exponent. The more bits in the fraction field increases the precision of the number. The fraction is also called the mantissa.

Single precision floating point numbers are represented in 32 bits, with:

Sign Bit - 1 bit

Exponent - 8 bits

Fraction / Mantissa - 23 bits

Double precision floating point numbers are represented in 64 bits, with:

Sign Bit - 1 bit

Exponent - 11 bits

Fraction / Mantissa - 52 bits

For a normalized floating point number in binary this is the representation:

Sign Bit	Exponent	Fraction / Mantissa
S	$E = e_1, \ldots, e_n$	$F = f_1, \ldots, f_m$

Definition 1.1.2: Significand

The significand is the part of a number that contains its significant digits. The significand is also called the mantissa. IEEE 754 assumes that the significand is normalized, i.e. it is in the form of $1.f_1 f_2 \ldots f_m$. The significand is 1 bit longer than the fraction field, i.e. it is 1 bit longer than the mantissa. The significand is equal to:

$$(1.F)_2 = \left(1.f_1 f_2 f_3 f_4 \ldots f_m\right)$$

Therefore the value of the significand is:

$$1 + f_1 \times 2^{-1} + f_2 \times 2^{-2} + f_3 \times 2^{-3} + \ldots + f_m \times 2^{-m}$$

IEEE 754 uses biased exponent representation, i.e.:

$$E = \text{Exponent} - \text{Bias}$$

The bias term depends on the number of bits in the exponent field. Defined by the formula:

$$\text{Bias} = 2^{(k-1)} - 1$$

Where k is the number of bits in the exponent field. Therefore the bias for the exponent field is:

Single Precision - 127, i.e. $2^{(8-1)} - 1$

Double Precision - 1023, i.e. $2^{(11-1)} - 1$

Finally the value of a normalized floating point number is:

$$(-1)^S \times \left(1 + f_1 \times 2^{-1} + f_2 \times 2^{-2} + f_3 \times 2^{-3} + \ldots + f_m \times 2^{-m}\right) \times 2^E$$

Example 1.1.1

Question 1

What is the decimal value of the following single precision floating point number?

$$\underbrace{1}_{\text{Sign}} \quad \underbrace{01111100}_{\text{Exponent}} \quad \underbrace{01000000000000000000000}_{\text{Fraction}}$$

Solution:

$$\begin{aligned}
S &= 1 \\
E &= (01111100)_2 - 127 \\
&= 124 - 127 \\
&= -3 \\
F &= 1 + 1 \times 2^{-2} \\
\\
D &= (-1)^1 \times (1.25) \times 2^{-3} \\
&= -0.15625
\end{aligned}$$

Question 2

What is the decimal value of the following single precision floating point number?

$$\underbrace{0}_{\text{Sign}} \quad \underbrace{10000010}_{\text{Exponent}} \quad \underbrace{01001100000000000000000}_{\text{Fraction}}$$

Solution:

$$\begin{aligned}
S &= 0 \\
E &= (10000010) - 127 \\
&= 130 - 127 \\
&= 3 \\
F &= 1 + 1 \times 2^{-2} + 1 \times 2^{-5} + 1 \times 2^{-6} \\
&= 1.296875 \\
D &= (-1)^{0} \times (1.296875) \times 2^{3} \\
&= 10.375
\end{aligned}$$

Example 1.1.2

Question 3

Convert -0.8125 to single and double precision floating point IEEE 754 format.

Solution:

$$\begin{aligned}
S &= 1 \\
F &= 0.8125 \times 2 = 1.625 \\
&= 0.625 \times 2 = 1.25 \\
&= 0.25 \times 2 = 0.5 \\
&= 0.5 \times 2 = 1 \\
&= (0.1101)_2 \\
&= \underbrace{(1.101)_2 \times 2^{-1}}_{\text{The exponent is the number of times we moved the decimal point to the right or left. Negative for right and positive for left.}}
\end{aligned}$$

$$\begin{aligned}
E &= -1 + 127 \\
&= 126 && \text{Add bias when converting to IEEE 754}
\end{aligned}$$

$$\underbrace{1}_{\text{Sign}} \quad \underbrace{01111110}_{\text{Exponent}} \quad \underbrace{10100000000000000000000}_{\text{Fraction}}$$

$$\begin{aligned} E &= -1 + 1023 \\ &= 1022 \end{aligned}$$

$$\underbrace{1}_{\text{Sign}} \quad \underbrace{01111111110}_{\text{Exponent}} \quad \underbrace{1010000000000000000000000000000000000000000000000000}_{\text{Fraction}}$$

Question 4

Convert -125.343 into a single and double precision floating point integer.

Solution:

$$\begin{aligned} S &= 1 \\ F &= 0.343 \times 2 = 0.686 \\ &= 0.686 \times 2 = 1.372 \\ &= 0.372 \times 2 = 0.744 \\ &= 0.744 \times 2 = 0.488 \\ &= 0.488 \times 2 = 0.976 \\ &= 0.976 \times 2 = 1.952 \\ &= 0.952 \times 2 = 1.904 \\ &= 0.904 \times 2 = 1.808 \\ &= 0.808 \times 2 = 1.616 \\ &= 0.616 \times 2 = 1.232 \\ &= 0.232 \times 2 = 0.464 \\ &= 0.464 \times 2 = 0.928 \\ &= 0.928 \times 2 = 1.856 \\ &= 0.856 \times 2 = 1.712 \\ &= 0.712 \times 2 = 1.424 \\ &= 0.424 \times 2 = 0.848 \\ F &= 1111101.0100011111001110 \\ F &= 1111101.0100011111001110 \\ F &= 1.1111010100011111001110 \end{aligned}$$

$$\begin{aligned} E &= 6 + 127 \\ &= 133 \end{aligned}$$

$$\underbrace{1}_{\text{Sign}} \quad \underbrace{10000101}_{\text{Exponent}} \quad \underbrace{11110101000111110011100}_{\text{Fraction}}$$

$$\begin{aligned} E &= 6 + 1023 \\ &= 1029 \end{aligned}$$

$$\underbrace{1}_{\text{Sign}} \quad \underbrace{10000000101}_{\text{Exponent}} \quad \underbrace{1111010100011111001110000000000000000000000000000000}_{\text{Fraction}}$$

The largest normalized single precision IEEE 754 float is:

$$\underbrace{0}_{\text{Sign}} \quad \underbrace{11111110}_{\text{Exponent}} \quad \underbrace{11111111111111111111111}_{\text{Fraction}}$$

$\therefore (-1)^0 \times 1 + \sum_{i=1}^{23} 1 \times 2^{-i} \times 2^{127} = 3.4028\ldots \times 10^{38}$ And largest normalized double precision IEEE 754 float is:

$$\underbrace{0}_{\text{Sign}} \quad \underbrace{11111111110}_{\text{Exponent}} \quad \underbrace{1111111111111111111111111111111111111111111111111111}_{\text{Fraction}}$$

$\therefore (-1)^0 \times 1 + \sum_{i=1}^{52} 1 \times 2^{-i} \times 2^{1023} = 1.79769...\ldots \times 10^{308}$

Chapter 2

MIPS Floating-Point Instructions

Definition 2.0.1: MIPS

Microprocessor without Interlocked Pipeline Stages

Definition 2.0.2: Computer Architecture vs Computer Organization

- Computer Architecture: The attributes of a system that are visible to a programmer.
- Computer Organization: The operational units and their interconnections that realize the architectural specifications.