Advanced Methods in Computational Physics

Dr. Movahed

Exercise set 1

Pooria Dabbaghi 98416029

1)

We assumed that list_arrange and sunspot.txt files are connected and therefore we wrote a bash script that splits the sunspot.txt file to directories that have a file with specified country name. For running a program we simply wrote a c++ program that prints out the argument that is passed to it.

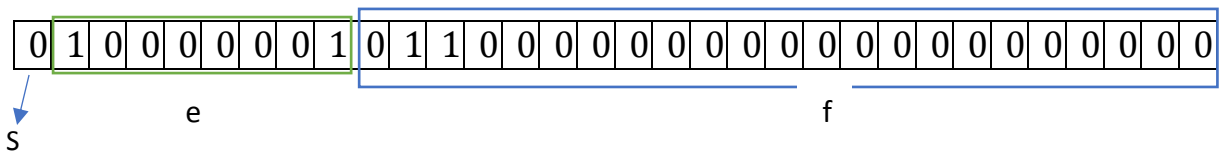
2)

First we represent 5.5 and 10^{-8} in floating point representation

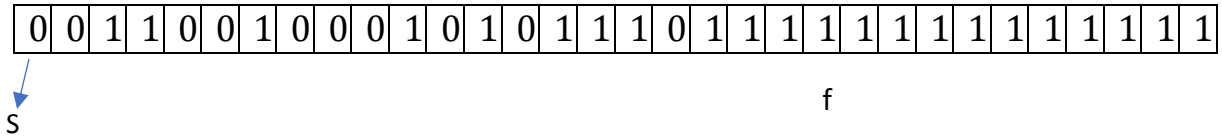
$$\begin{matrix}(5)_{10} = 1*2^2 + 0*2^1 + 1*2^0 = (101)_2 \\ (0.5)_{10} = 1*2^{-1} = (0.1)_2\end{matrix} \Rightarrow (5.5)_{10} = (101.1)_2 = \left(1.\underbrace{011}_{f} * 2^2\right)_{fpr}$$

$$\begin{aligned}(10^{-8})_{10} = {} & 1*2^{-27} + 0*2^{-28} + 1*2^{-29} + 0*2^{-30} + 1*2^{-31} + 0*2^{-32} + 1*2^{-33} + 1*2^{-34} + 1*2^{-35} \\ & + 0*2^{-36} + 1*2^{-37} + 1*2^{-38} + 1*2^{-39} + 1*2^{-40} + 1*2^{-41} + 1*2^{-42} + 1*2^{-43} \\ & + 1*2^{-44} + 1*2^{-45} + 1*2^{-46} + 1*2^{-47} + 1*2^{-48} + 1*2^{-49} + 1*2^{-50} + 1*2^{-51} \\ & + \cdots = \left(1.\underbrace{01010111\overline{01111111111111}}_{f \to 23\ bits} * 2^{-27}\right)_{fpr}\end{aligned}$$

So for 5.5 representation is like this($e = 129 = (10000001)_2$):



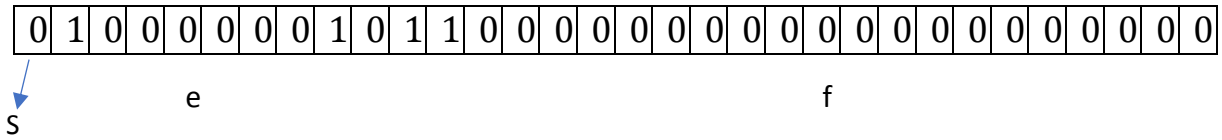
and for 10^{-8} ($e = 100 = (1100100)_2$):



So by adding these two numbers we get

$$(5.5 + 10^{-8})_{10} = \left(101.1\underbrace{0\ldots0}_{25}10101011101\overline{1111}\right)_2 = \left(1.011\underbrace{0\ldots0}_{20} * 2^2\right)_{fpr}$$

Floating point representation for $5.5 + 10^{-8}$



Therefore we won't get the right answer by this method because we can't show more than 23 decimals in single precision and as shown we get $5.5 + 10^{-8} = 5.5$! Thus we should try another methods like using double precision.

3)

a) Single Precision

Max $(s = 0,\ e = 254\ ,\ f = 1\underbrace{1\ldots1}_{21}1\ )$

$$\rightarrow max = \left(1.1\underbrace{1\ldots1}_{21}1 * 2^{127}\right)_{fpr} = 1 * 2^{127} * 1.\sum_{m=0}^{22} 2^{-(m+1)} = 2^{128}$$
$$\cong 3.4028 * 10^{38}$$

Min $(s = 0, e = 1, f = 0\underbrace{0\ldots0}_{21}0)$

$$\rightarrow min = \left(1.0\underbrace{0\ldots0}_{21}0 * 2^{-126}\right)_{fpr} = 1 * 2^{-126} * \left(\underbrace{0 + 2^{-23}}_{IEEE}\right) = 2^{-149}$$
$$\cong 1.4013 * 10^{-45}$$

b) Double Precision

Max $(s = 0,\ e = 2046\ ,\ f = 1\underbrace{1\ldots1}_{50}1\ )$

$$\rightarrow max = \left(1.1\underbrace{1\ldots1}_{50}1 * 2^{1023}\right)_{fpr} = 1 * 2^{1023} * 1.\sum_{m=0}^{51} 2^{-(m+1)} = 2^{1024}$$

$$\cong 1.7977 * 10^{308}$$

$$\text{Min}\ (s = 0, e = 1, f = 0\,\underbrace{0\ldots0}_{50}\,0)$$

$$\rightarrow min = \left(1.0\,\underbrace{0\ldots0}_{50}\,0 * 2^{-126}\right)_{fpr} = 1 * 2^{-1022} * \left(\underbrace{0 + 2^{-52}}_{IEEE}\right) = 2^{-1074} =$$

$$\cong 4.9407 * 10^{-324}$$