CompPhysics2

Vandan Pokal

September 2020

1

A) The formula for single precision float is as follows:

$$float(x) = (-1)^{(s)} \times \left[1 + \sum_{i=0}^{22} b_i \times 2^{-(23-i)}\right] \times 2^{e-127}$$

Then,
1) the value of the sign would be: $S = 1$
2) the value of the exponent would be $e = 2^6 + 2^4 + 2^2 + 2 = 86 \Rightarrow 2^{-41}$
3) the value of the mantissa is:$1 + 2^{-1} + 2^{-3} + 2^{-5} + 2^{-6} = 1.671875$

B)The number composed would be:

$$-1.671875 \times 2^{-41} = -1.671875 \times 10^{-41 log_{10} 2} = -7.6028 \times 10^{-13}$$

2

A) Since there are 12 bits for the exponent:
1) The exponent part of the single-precision float would change as:

$$2^{e-(2^{10}+2^9+2^8+2^7+2^6+2^5+2^4+2^3+2^2+2^1+2^0)} = 2^{e-2047}$$

2) The mantissa would change as:

$$\left[1 + \sum_{i=0}^{19} b_i \times 2^{-(20-i)}\right]$$

Then the largest normal number would be:

$$\left[(2 - 2^{-20})\right] \times 2^{4094-2047} = 2^{2048} = 3.2317 \times 10^{616}$$

The smallest positive number that can be stored is:

$$\left[2^{-20}\right] \times 2^{-2047} = 2^{-2067} = 5.9019 \times 10^{-623}$$

B) The machine precision for the system in A is $2^{-20} = 10^{-6}$

C) Lets x bits for the exponent , then the mantissa will have $31 - x$. Then the general formula would be:

$$float(x) = (-1)^{(s)} \times \left[1 + \sum_{i=0}^{30-x} b_i \times 2^{-((31-x)-i)}\right] \times 2^{e - \left(\sum_{i=2}^{x} 2^{i-2}\right)}$$

The largest normal number and smallest positive number would be, respectively:

$$(2 - 2^{31-x}) \times 2^{\sum_{i=2}^{x} 2^{i-2}} \tag{1}$$

$$2^{-(31-x)} \times 2^{(-\sum_{i=2}^{x} 2^{i-2})+1} = 2^{-(30-x+\sum_{i=2}^{x} 2^{i-2})} \tag{2}$$

Taking the log of the numbers:

$$\log_{10} 2^{1+\sum_{i=2}^{x} 2^{i-2}} = \left[1 + \sum_{i=0}^{x-2} 2^i\right] \log_{10} 2 = \log_{10}(N_{max})$$

$$-(30 - x + \sum_{i=2}^{x} 2^{i-2} \log_{10} 2 = \log_{10} N_{min}$$

Considering the fact that the largest number for single bit in the exponent is 1. The plots $\log_{10} N_{max}$ and $\log_{10} N_{min}$ are as follows:

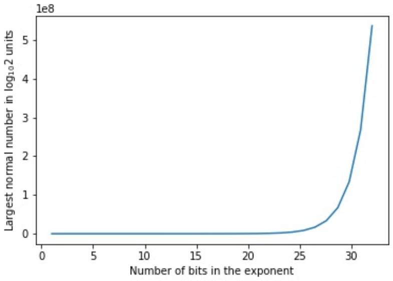

Figure 1: $\log_{10} N_{max}(x)$

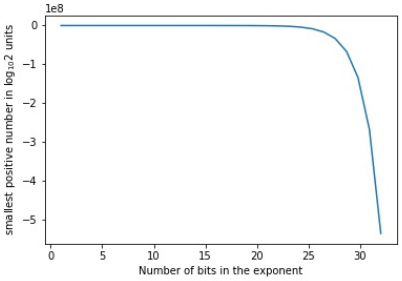

Figure 2: $\log_{10} N_{min}(x)$