High Performance Computing
Homework 2

Francisco Jose Castillo Carrasco

September 20, 2018

Problem 1

(a) What is the normalized IEEE single-precision representation of the number 5.5?

Solution: We can find the binary representation of the number as follows. First, the integer part

$$5/2 = 2 + 1;$$
$$2/2 = 1 + 0;$$
$$1/2 = 0 + 1;$$

which implies

$$5_{10} = 101_2.$$

Second, the decimal part

$$0.5 \times 2 = 1 + 0;$$

which gives

$$0.5_{10} = 0.1_2.$$

Finally, we have that $5.5_{10} = 101.1_2$. We express that in the normalized format and obtain

$$5.5_{10} = 1.011000\ldots000 \times 2^2.$$

This means that $a_2 = a_3 = 1$ and the rest of a_j are zero.

(b) If we change the significand of 5.5 by one ulp, by how much does the value of the floating point representation change? Express your answer as a power of 2.

Solution: We change the *unit of last place*, a_{23}, from 0 to 1. This produces a change in the value of $\Delta x = 2^{-23+2} = 2^{-21}$

Problem 2

Let $x = 1/3$. (a) Find the binary representation of x.

Solution: Similarly as in the previous problem, we obtain

$$x = \frac{1}{3}_{10} = 0.0101\overline{01}\ldots$$

(b) Find the IEEE-754 single precision representation $\hat{x}$ of x when rounding to nearest.

Solution: To find the normalized representation we must make $a_0 = 1$ such as

$$1.01\overline{01}\cdots \times 2^2,$$

and cut at the $23rd$ digit,

$$\hat{x} = 1.01010101010101010101011 \times 2^2.$$

Note that we have rounded to nearest in a_{23}.

(c) What is the absolute error due to rounding? In other words, what is $|\hat{x} - x|$?

Solution: We begin using geometric series to represent both quantities as

$$x = S = \frac{a}{1-r},$$

and

$$\hat{x} = S_N + 2^{-23} = \frac{a(1-r^{N+1})}{1-r} + 2^{-23}$$

with $a = r = 1/4$ and $N = 11$. Therefore,

$$|\hat{x} - x| = |S_N + 2^{-23} - S| = \left|2^{-23} - \frac{r^{N+1}}{1-r}\right| = \left|2^{-23} - \frac{(1/4)^{N+1}}{3/4}\right| = 9.934 \cdot 10^{-9}.$$

Problem 3

How does the spacing depend on e?

Solution: We obtain the spacing by focusing in the last digit a_{23} that is multiplied by 2^e. Hence,

$$1\ ulp = 2^{-23}2^e = 2^{e-23}.$$

Problem 4

How many possible nonnegative normalized IEE single precitions floating point numbers are there?

Solution: Since we have 24 bits of mantisa and one of them is fixed, we have 2^{23} possibilities due to the mantissa. The exponent multiplies those possibilities by 127×2. Hence, the number of possible nonnegative normalized IEE single precitions floating point numbers is $N = 254 \times 2^{23}$

Problem 5

Consider IEEE single-precision representations. (a) Is $1,000,000.0$ exactly representable in IEEE single precision?

Solution: Yes, we find the representation doing the same as in Problem 1, although this is a much longer case. We obtained

$$1,000,000.0 = 1.11101000010010000000000 \times 2^{19}$$

(b) What is the smallest positive integer M that does not have an exact IEEE single-precision representation?

Solution: Since we have 24 bits of mantisa, the largest value that those bits can represent is $2^{24} = 16777216$. Therfore the next integer will not be representable. Therefore the smallest positive integer M that does not have an exact IEEE single-precision representation is

$$M = 2^{24} + 1 = 16,777,217.$$

Problem 6

True or False: If x has a terminating base-2 expansion, then x has a terminatiing base-10 expansion.

Solution: Let $x = p/q$. Since x has base-2 expansion, q divides some power of 2,

$$\frac{2^e}{q} = K \in \mathbb{Z}.$$

Now we have some power e of 10 divided by q,

$$\frac{10^e}{q} = \frac{2^e 5^e}{q} = K5^e = C \in \mathbb{Z}.$$

Hence, if q divides some power of 2, it also divides some that same power of 10 (since 10 is multiple of 2). Therefore, since q divides some power of 10, $x = p/q$ has a terminating base-10 expansion.

The assertion is TRUE.

Problem 7

The machine epsilon for IEEE single-precision numbers is $2^{-23} \approx 1.2 \times 10^{-7}$. In this respect, single-precision IEEE floating-point is roughly equivalent to 7 decimal digits. On the other hand, 7 decimal digits do not suffice to represent an IEEE single-precision floating-point number uniquely. This exercise outlines a proof. Consider real numbers x such that $10 \leq x < 16$, (a) The numbers in this interval that are exactly representable in 7 decimal digits are 10.00000, 10.00001, 10.00002,..., 15.99999. How many numbers are in the set?

Solution: There are 10^5 numbers for the decimal values multiplied by 6 for the integer part from 0 to 5. There are a total of

$$N_a = 6 \cdot 10^5.$$

The numbers in [10, 16) that are exactly representable in IEEE format. How many numbers are in this set?

Solution: There are 2^{21} for the bits after the second decimal bit, multiplied by 3i for the combinations of the first 2, making a total of

$$N_b = 3 \cdot 2^2 1 = 6291456 > N_a.$$

Explain how the pigeonhole principle implies that at least two different IEEE single-precision numbers have the same 7-digit decimal representation (and why that proves the result).

Solution: It is inmediate that since $N_b > N_a$, the seven digit decimali representation cannot represent all the numbers represented by the IEE single precision format. In addition, least two different IEEE single-precision numbers have the same 7-digit decimal representation.

Explain why 8 decimal digits also dont suffice to represent IEEE single-precision numbers uniquely.

Solution: Same as in part (a), we have that the total number of elements in this set is

$$N_c = 6 \cdot 10^6 < N_b.$$

Since $N_c < N_b$, 8 decimal digits also don't suffice to represent IEE single-precision numbers uniquely. The discussion is the same as above.

Problem 8

(a) What real number is represented by $(+, 1.5, 0)$ $(-, 1.5, 1)$ $(+, 1.5, 2)$ $(-, 1.5, 1)$?

Solution:

$$\begin{aligned}
(+, 1.5, 0) &= 1.5 \\
(-, 1.5, 1) &= -2^{1.5} = -2\sqrt{2} \\
(+, 1.5, 2) &= 2^{2^{1.5}} = 2^{2\sqrt{2}} = 4^{\sqrt{2}} \approx 7.103 \\
(-, 1.5, 1) &= -2^{-1.5} = -\frac{1}{2^{1.5}} = -\frac{1}{2\sqrt{2}}
\end{aligned}$$

(b) What is the set of representable values for $l = 0, \pm 1, \pm 2, \pm 3, \pm 4$?

Solution:

- $l = 0$,

$$S = \{x; x = \pm s, s \in [1, 2)\}.$$

- $l = \pm 1$,

$$S = \{x; |x| = 2^{\pm s}, s \in [1, 2)\}.$$

- $l = \pm 2$,

$$S = \{x; |x| = 2^{\pm 2^{s}}, s \in [1, 2)\}.$$

- $l = \pm 3$,

$$S = \{x; |x| = 2^{\pm 2^{2^{s}}}, s \in [1, 2)\}.$$

- $l = \pm 3$,

$$S = \{x; |x| = 2^{\pm 2^{2^{2^{s}}}}, s \in [1, 2)\}.$$

(c) The number $G = 10^1 00$ is called a *googol*; 10^G is a *googolplex*. Express the largest representable value of (+, s, 4) as an approximate power of G.

Solution: First we have that the largest value of $(+, s, 4)$ is

$$x \approx 2^{2^{2^{2^{2}}}} = 2^{2^{2^{4}}} = 2^{2^{16}}.$$

To express it as an approximate power of G we have

$$x = G^y = \left(10^{100}\right)^y = 10^{100y}.$$

Using logarithms we get

$$\log_{10} x = 100y \Rightarrow y = \frac{1}{100} \log_{10}\left(2^{2^{16}}\right) = \frac{2^{16}}{100} \log_{10}(2) \approx 197.$$

(d) Is the largest representable value of (+, s, 5) greater or less than a googolplex? Explain.

Solution: Let $z = 10^G$ and $w = (+, s, 5)$. Using logarithms as before

$$\log_{10} z = 10^{100} = G,$$

and

$$\log_{10} w = 2^{2^{16}} \log_{10} 2 = x \log_{10} 2 = G^y \log_{10} 2.$$

It is obvious that $\log_{10} 2 << G$ and therefore

$$\log_{10} w = G^y \log_{10} 2 < \frac{G^y}{G} = G^{y-1} > G = \log_{10} z.$$

Thus,

$$\log_{10} w = \log_{10} z \Rightarrow (+, s, 5) > 10^G.$$

(e) Invent your own terminology as necessary to describe the largest representable values of $(+, s, 6)$ and $(+, s, 7)$.

Solution: Let $b^{\#_n^e}$ denote the number obtained by calculating the consecutive power of b n times and ending with the power e. For example,

$$5^{\#_3^2} = 5^{5^{5^2}},$$

and

$$8^{\#_4^6} = 8^{8^{8^{8^6}}}.$$

With that notation we have,

$$(+, s, 6) = +2^{\#_6^s},$$

and

$$(+, s, 7) = +2^{\#_7^s}.$$