Amath 483

HW1

by Cynthia Hong

Problem 1

The practical measure of machine's SP precision is 1.19209×10^{-7} and that of DP's is 2.22045×10^{-16} by taking the difference of 2 numbers and comparing the result to zero in each data type

Problem 2

Q: Find the largest and smallest SP & DP numbers that can be represented in IEEE arithmetic.

For SP:

32 bit

$S = 1 \quad k = 8 \quad n = 23$

the largest SP numbers (largest unnormalised):

$E = 2^{k-1} - 1 = 2^7 - 1 = 127$

$f = 1 - 2^{-n} = 1 - 2^{-23}$

$M = 1 + f = 2 - 2^{-n} = 2 - 2^{-23}$

$V = (2 - 2^{-n}) \cdot 2^{2^{k-1}-1}$

$= (2 - 2^{-23}) \cdot 2^{127}$

$= 3.4028 \times 10^{38}$

the smallest SP numbers:

Since the question ask for the smallest SP IEEE number,
the smallest SP number will be negative number for $-1 <$ smallest representable positive number

Therefore, we could directly get from:

$V = -1 \cdot V_{\text{largest normalised}}$

$= -3.4028 \times 10^{38}$

For DP:

64 bit

$S = 1 \quad k = 11 \quad n = 52$

the largest DP numbers (largest unnormalised):

$E = 2^{k-1} - 1 = 2^{10} - 1 = 1023$

$f = 1 - 2^{-n} = 1 - 2^{-52}$

$M = 1 + f = 2 - 2^{-n} = 2 - 2^{-52}$

$V = (2 - 2^{-n}) \cdot 2^{2^{k-1}-1}$

$= (2 - 2^{-52}) \cdot 2^{1023}$

$= 1.7977 \times 10^{308}$

the smallest DP numbers:

Since the question ask for the smallest DP IEEE number,
the smallest DP number will be negative number for $-1 <$ smallest representable positive number

Therefore, we could directly get from:

$V = -1 \cdot V_{\text{largest normalised}}$

$= -1.7977 \times 10^{308}$

Problem 3

1) The result of 200 * 300 * 400 * 500 I got is -884901888.

2) I find that the result is a negative number, which is not correct. The correct result should be 1.2×10^{10}.

The effect I observed is overflow.

3) The definition of underflow is that the result of an arithmetic operation is too small to be represented within the given number of bits. This effect occurs when the result is smaller the the smallest normalized floating point number which could be stored within the given precision. In another word, underflow happenes when any number smaller than the smallest positive normalized fp number cannot be represented.

The math formula for underflow is:

$$|result| < 2^{-2^{k-1}+2}$$

For SP,

$$2^{-2^{8-1}+2} \approx 1.1755 \times 10^{-38}$$

The smallest number SP could be represented by SP is 1.1755×10^{-38}.

The formula is $|result| < 1.1755 \times 10^{-38}$

For DP,

$$2^{-2^{11-1}+2} \approx 2.2251 \times 10^{-308}$$

The smallest number DP could be represented by DP is 2.2251×10^{-308}

The formula is $|result| < 2.2251 \times 10^{-308}$

4) The definition of overflow is that the result of an arithmetic operation is too large to be represented within the given number of bits. This effect occurs when the result is larger the the largest normalized floating point number which could be stored within the given precision. In another word, overflow happenes when any number larger than the largest positive normalized fp number cannot be represented.

The math formula for underflow is:

$$|result| > (2-2^{-n}) \times 2^{2^{k-1}-1}$$

For SP,

$$(2-2^{-23}) \cdot 2^{127} \approx 3.4028 \times 10^{38}$$

The largest number SP could be represented by SP is 3.4028×10^{38}.

The formula is $|result| > 3.4028 \times 10^{38}$.

For DP,

$$(2-2^{-52}) \cdot 2^{1023} \approx 1.7977 \times 10^{308}$$

The largest number DP could be represented by DP is 1.7977×10^{308}

The formula is $|result| > 1.7977 \times 10^{308}$.

Problem 4

The Question asks for the normalized floating point number of SP and DP.

For SP:

32 bit

$S=1 \quad K=8 \quad n=23$

For sign, there are 2 possible outcomes, where is 0 or 1.

For exponent, there are $2^8-2=254$ possible outcomes.

For fraction, there are $2^{23}=8388608$ possible outcomes.

$2 \times 254 \times 8388608 = 4261412864$

herefore, the normalized floating point number of SP is 4261412864.

For DP:

64bit

$S=1 \quad K=11 \quad n=52$

For sign, there are 2 possible outcomes, where is 0 or 1.

For exponent, there are $2^{11}-2=2046$ possible outcomes.

For fraction, there are $2^{52}=4.5036 \times 10^{15}$ possible outcomes.

$2 \times 2046 \times 4.5036 \times 10^{15} = 1.8429 \times 10^{19}$

herefore, the normalized floating point number of SP is 1.8429×10^{19}.

Problem 6

6 bit

$s=1 \quad k=3 \quad n=2$

Normalized $E = e - \text{bias}$, $e = e_2 e_1 e_0$, $\text{bias} = 2^2 - 1 = 3$

$E_{001} = -2$	$n_1 n_0$	f	$M = 1+f$
$E_{010} = -1$	0 0	0	1
$E_{011} = 0$	0 1	1/4	5/4
$E_{100} = 1$	1 0	1/2	3/2
$E_{101} = 2$	1 1	3/4	7/4
$E_{110} = 3$			

For M_{00}:

$s=1,\ v<0 \Rightarrow \{-\frac{1}{4}, -\frac{1}{2}, -1, -2, -4, -8\}$ from the support materials

$s=0,\ v>0 \Rightarrow \{\frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$

For M_{01}:

$V = (-)^s M 2^E \qquad M_{01} = \frac{5}{4}$

$M_{01} \cdot 2^{E_{001}} = \frac{5}{4} \cdot 2^{-2} = \frac{5}{16}$

$M_{01} \cdot 2^{E_{010}} = \frac{5}{4} \cdot 2^{-1} = \frac{5}{8}$

$M_{01} \cdot 2^{E_{011}} = \frac{5}{4} \cdot 2^{0} = \frac{5}{4}$

$M_{01} \cdot 2^{E_{100}} = \frac{5}{4} \cdot 2^{1} = \frac{5}{2}$

$M_{01} \cdot 2^{E_{101}} = \frac{5}{4} \cdot 2^{2} = 5$

$M_{01} \cdot 2^{E_{110}} = \frac{5}{4} \cdot 2^{3} = 10$

$\Rightarrow$ $s=1,\ v<0 \Rightarrow \{-\frac{5}{16}, -\frac{5}{8}, -\frac{5}{4}, -\frac{5}{2}, -5, -10\}$

$s=0,\ v>0 \Rightarrow \{\frac{5}{16}, \frac{5}{8}, \frac{5}{4}, \frac{5}{2}, 5, 10\}$

For M_{10}:

$M_{10} \cdot 2^{E_{001}} = \frac{3}{2} \cdot 2^{-2} = \frac{3}{8}$

$M_{10} \cdot 2^{E_{010}} = \frac{3}{2} \cdot 2^{-1} = \frac{3}{4}$

$M_{10} \cdot 2^{E_{011}} = \frac{3}{2} \cdot 2^{0} = \frac{3}{2}$

$M_{10} \cdot 2^{E_{100}} = \frac{3}{4} \cdot 2^{1} = \frac{3}{2}$

$M_{10} \cdot 2^{E_{101}} = \frac{3}{4} \cdot 2^{2} = 6$

$M_{11} \cdot 2^{E_{110}} = \frac{3}{4} \cdot 2^{3} = 12$

$\Rightarrow$ $s=1,\ v<0 \Rightarrow \{-\frac{3}{8}, -\frac{3}{4}, -\frac{3}{2}, -3, -6, -12\}$

$s=0,\ v>0 \Rightarrow \{\frac{3}{8}, \frac{3}{4}, \frac{3}{2}, 3, 6, 12\}$

For M_{11}:

$M_{11} \cdot 2^{E_{001}} = \frac{7}{4} \cdot 2^{-2} = \frac{7}{16}$

$M_{11} \cdot 2^{E_{010}} = \frac{7}{4} \cdot 2^{-1} = \frac{7}{8}$

$M_{11} \cdot 2^{E_{011}} = \frac{7}{4} \cdot 2^{0} = \frac{7}{4}$

$M_{11} \cdot 2^{E_{100}} = \frac{7}{4} \cdot 2 = \frac{7}{2}$

$M_{11} \cdot 2^{E_{101}} = \frac{7}{4} \cdot 2^{2} = 7$

$M_{11} \cdot 2^{E_{110}} = \frac{7}{4} \cdot 2^{3} = 14$

$\Rightarrow$ $s=1,\ v<0 \Rightarrow \{-\frac{7}{16}, -\frac{7}{8}, -\frac{7}{4}, -\frac{7}{2}, -7, -14\}$

$s=0,\ v>0 \Rightarrow \{\frac{7}{16}, \frac{7}{8}, \frac{7}{4}, \frac{7}{2}, 7, 14\}$

Denormalized

$e_2 e_1 e_0 = 000 \qquad E = -2$

$V = (-)^s M \cdot 2^E$

$M_{00} = 0 \qquad M_{00} \cdot 2^{-2} = 0$

$M_{01} = \frac{1}{4} \qquad M_{01} \cdot 2^{-2} = \frac{1}{16}$

$M_{10} = \frac{1}{2} \qquad M_{10} \cdot 2^{-2} = \frac{1}{8}$

$M_{11} = \frac{3}{4} \qquad M_{11} \cdot 2^{-2} = \frac{3}{16}$

$\Rightarrow$

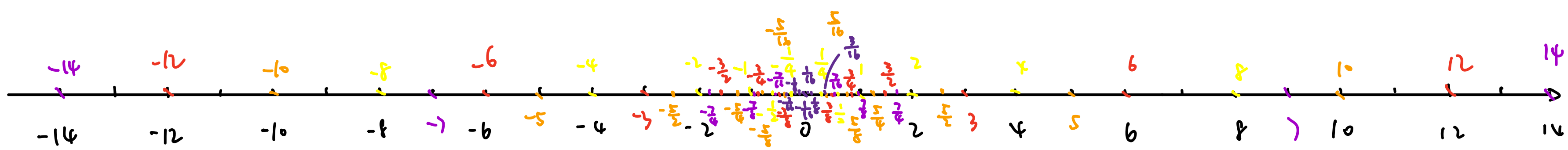
$s=0,\ v \Rightarrow \{0\} \qquad s=1,\ v \Rightarrow \{0\}$

$s=0,\ v \Rightarrow \{\frac{1}{16}\} \qquad s=1,\ v \Rightarrow \{-\frac{1}{16}\}$

$s=0,\ v \Rightarrow \{\frac{1}{8}\} \qquad s=1,\ v \Rightarrow \{-\frac{1}{8}\}$

$s=0,\ v \Rightarrow \{\frac{3}{16}\} \qquad s=1,\ v \Rightarrow \{-\frac{3}{16}\}$

plot



M_{00}

M_{01}

M_{10}

M_{11}