Student's Name: WANG XIANGBO **Course Name: CSPP58001**

1) Answer:

The Double Precision Floating point number is 64 bits. Its sign bit, exponent bits and mantissa bits are 1, 11 and 52 respectively. the minimum value of it is

$$2^{-1022} \approx 2.23 \times 10^{-308}$$

And the maximum value is

$$(1 + (1 - 2^{-52})) \times 2^{1023} \approx 1.798 \times 10^{308}$$

2)Answer:

First, converting 2^24 into Single Precision Floating Format, we can get that

$$2^{24} = 0\ 00011000\ 0000000\ldots\text{(all the rest are zeros)} = (1.000\ldots\text{(total 23 zeros)})_2 \times 2^{24}$$

If we converting 1 into Single Precision Floating Format we can get that

$$1 = (1.000\ldots\text{(total 23 zeros)})_2 \times 2^0 = (0.00\ldots01\text{(total 23 zeros after decimal point)})_2 \times 2^{24}$$

Then if we add these two numbers, it will be

$$(1.00\ldots01\text{(total 23 zeros after decimal point)})_2 \times 2^{24}$$

which has a 24 bits Mantissa. It exceeds the precision limit of the Single Precision Floating Format since there are only 23 bits of Mantissa.

Therefore 2^24 + 1 is still 2^24 in Single Precision Floating Format.

3)Answer:

$111 = (1101111)_2$

$0.875 = (0.111)_2$

$111.875 = (1101111.111)_2 = (1.101111111)_2 \times 2^6$

Since the offset is 129 and exponent is 10 bits, the exponent bits should be

$$129+6 = 135 = (0010000111)_2$$

Then the result should be

0 0010000111 101111111000...(all the rest are zeros)

format sign+exponent+mantissa

4)Answer:

Since for single precision, the fraction is 23 bits, then the epsilon should be 2^-24 = 5.96e-08.

Since for double precision, the fraction is 52 bits, then the epsilon should be 2^-53 = 1.11e-16.

9)Answer:

Based on the code for Gaussian Elimination in Q10, there is a 'while loop' outside the codes which cost O(n^2). In the wile loop, there are several O(n) 'for loops' to do swapping and subtracting/dividing. Therefore the approximate is O(n^3).

From another view, as the Gaussian Elimination just picking one pivot and doing elimination on the other rows, and there is n pivots to pick, therefore the approximate is n(n-1)+(n-1)(n-2)+...2*1 = O(n^3).