Alexandria University
Faculty of Engineering

Assignment NO. 1

By:

Ali Mohamed Taima Hummus

Q1

- The smallest
  - for Single Precision

    0 00000001 00000000000000000000000

    Exponent – bias = 1 – 127 = –126

    Significand = (1.0 … 0)2 = 1

    $1 * 2^{-126}$

    Value in decimal = $1.17549 * 10^{-38}$

  - For double precision

    0 00000000001 0000000000000000000000000000000000000000000000000000

    Exponent – bias = 1 – 1023

    Significand = (1.0 … 0)2 = 1

    $1 * 2^{-1022}$

    Value in decimal = $2.22507 * 10^{-308}$

- The largest
  - for Single Precision

    0 11111110 11111111111111111111111

    Exponent – bias = 254 – 127 = 127

    Significand = $(1.11…1)_2 \approx 2$

    $2 * 2^{127} = 2^{128}$

    Value in decimal = $3.4028 * 10^{38}$

  - For double precision

    0 11111111110 1111111111111111111111111111111111111111111111111111

    Exponent – bias = 2046 – 1023 = 1023

    Significand = $(1.11…1)_2 \approx 2$

    $2 * 2^{1023} = 2^{1024}$

    Value in decimal = $1.79769 * 10^{308}$

Q2

(32) –binary→ 0010 0000 ----2's complement → 1110 0000 = -32

(98) –binary→ 0110 0010 ----2's complement → 1001 1110 = -98

-32
-98

-130

1110 0000
1001 1110
----

1 0111 1110 → the result in decimal is 126 but the ture result is -130

The two numbers are negative but the result is positive, So there is an overflow occurs and the result is incorrect

To represent -130 we need to 16 bits

Q3

124.625

Sign = 0 → the number is positive

First convert 124 to binary

124 → 0111 1100

Second convert the fraction 0.625

Operation	Result	
0.625 * 2	1.25	1
0.25 * 2	0.5	0
0.5 * 2	1	1

The result is .101

The wole number = 0111 1100.101 = 1. 11 1100 101 * 2^6

Exponent = 6 + 127 = 133 -- binary -- → 1000 0101

The floating number in binary format is

0 1000 0101 1111 0010 1000 0000 0000 000

Q4

1- 0100 0011 0101 0100 0000 0000 0000 0000

Sign = 0 → The number is positive

1000 0110 = 134

Exponent – bias = 134 – 127 = 7

1. 101 0100 0000 0000 0000 0000 = 1. 65625

Value in decimal = +1. 65625 * 2^7 = 212

2- 1111 1111 1011 0100 0000 0000 0000 0000

Sign = 1 → The number is negative

1111 1111 = 255

The exponent is Maximum(255) and the fraction $\neq 0$, So the result is NaN(Not a Number)