Tutorial 6 - Week of Oct. 25th

Question 1) (3.22)

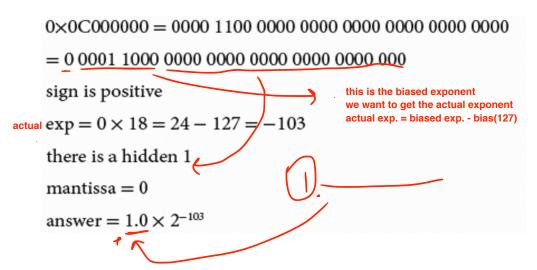
What decimal number does the bit pattern

0×0C000000

4 * 8 = 32

represent if it is a floating-point number? Use the IEEE 754 standard.

Solution:



Question 2) (3.23)

Write down the binary representation of the decimal number 63.25 assuming the IEEE 754 single precision format.

Solution:

$$63.25 \times 10^{0} = 1111111.01 \times 2^{0}$$

normalize, move binary point five to the left

$$1.11111101 \times 2^{5}$$

sign = positive,
$$\exp = \underbrace{127}_{0} + \underbrace{5}_{0} = \underbrace{132}_{0}$$

 $= 0100\ 0010\ 0111\ 1101\ 0000\ 0000\ 0000\ 0000 = 0x427D0000$

Question 3) (3.24)

Write down the binary representation of the decimal number 63.25 assuming the IEEE 754 double precision format.

Solution:

$$63.25 \times 10^{0} = 1111111.01 \times 2^{0}$$

normalize, move binary point five to the left

fraction part $3 \qquad 1.1111101 \times 2^{5}$

normalized (scientific notation) binary representation

a hidden 1 sign = positive, $\exp = 1023 + 5 = 1028$

biased exponent = actual exponent + bias

Final bit pattern:

 $0\ 100\ 0000\ 0100\ 1111\ 1010\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000$

= 0x404FA000000000000

Question 4) (3.27)

IEEE 754-2008 contains a half precision that is only 16 bits wide. The leftmost bit is still the sign bit, the exponent is 5 bits wide and has a bias of 15, and the mantissa is 10 bits long. A hidden 1 is assumed. Write down the bit pattern to represent -1.5625×10^{-1} assuming a version of this format, which uses an excess-16 format to store the exponent.

Solution:

$$-1.5625 \times 10^{-1} = -0.15625 \times 10^{0}$$

$$= -0.00101 \times 2^{\circ}$$

move the binary point three to the right, $= -1.01 \times 2^{-3}$

caponent

exponent =
$$-3 = -3 + 15 = 12$$
, fraction = -0.01000000000

answer: 10110001000000000

sign = negative = 1