

Q1- What decimal number does the bit pattern 0×0C000000 represent if it is a floating point number? Use the IEEE 754 standard.

$(0x0C000000)_{16}$ Mantissa

$= 0000\ 1100\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000$

- 1-bit sign (S) $= 0 \Rightarrow$ positive
- 8-bit exponent (E) $= 00011000 = 24$
- 23-bit fraction/mantissa (M)
 $= 0$

actual exponent $= 24 - 127 = -103$

$$= (-1)^0 \times (1 + \overset{\text{Mantissa}}{F}) \times 2^{(E-127)}$$

$$= (-1)^0 (1.0) \times 2^{-103}$$

$$= 1.0 \times 2^{-103}$$

Q2 Solution

$$\begin{aligned}00C000000 &= 00001100000000000000000000000000 = 00001100000 \\ 0000000000000000000000\end{aligned}$$

sign is positive

$$\text{exponent} = 0 \times 18 = 24 - 127 = -103$$

there is a hidden 1

mantissa = 0

answer 1.0×2^{-103}

Q2- Write down the binary representation of the decimal number 63.25 assuming the IEEE 754 double precision format.

Q2 Solution

$$63.25 \times 100 = 111111.01 \times 20$$

normalize, move binary point five to the left

$$1.1111101 \times 25$$

$$\text{sign} = \text{positive}, \text{exp} = 1023 + 5 = 1028$$

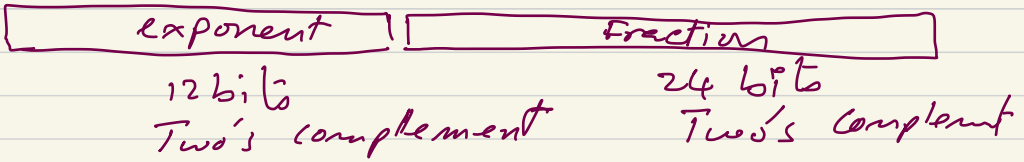
Final bit pattern:

0 100 0000 0100 1111 1010 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000

$$= 0x404FA00000000000$$

Q3- Write down the binary bit pattern to represent -1.5625×10^{-1} assuming a format similar to that employed by the DEC PDP-8 (the leftmost 12 bits are the exponent stored as a two's complement number, and the rightmost 24 bits are the fraction stored as a two's complement number). No hidden 1 is used. Comment on how the range and accuracy of this 36-bit pattern compares to the single and double precision IEEE 754 standards.

$$-1.5625 \times 10^{-1}$$



No Hidden 1 is used

0.15625×2	0
0.3125×2	0
0.625×2	1
0.25×2	0
0.5×2	1

$$-0.15625 \rightarrow -0.00101$$

No leading 1
No hidden bit

$$-0.101 \times 2^{-2}$$

$$\text{Fraction} = 101$$

$$\text{exponent} = -2$$

Encode the fraction (24-bit two's complement)

$$0.101_2 = 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3}$$

$$= \frac{1}{2} + 0 + \frac{1}{8} = 0.5 + 0.125 = 0.625$$

$$-0.101)_2 \rightarrow -0.625)_{10}$$

In Two's complement

$$+0.625 \rightarrow 0.10100\text{-----}0$$

$$\text{Invert bits} \rightarrow 1.01011\text{-----}1$$

$$\text{add } 1 \rightarrow 1.011000\text{-----}0$$

$$\text{Fraction number} \rightarrow 101100\text{-----}0$$

Encode the Exponent (12-bit two's complement)

-2 is 12-bit two's complement

$$+2: 0\text{-----}00010$$

$$\text{Invert: } 1\text{-----}1101$$

$$\text{add } 1: 1\text{-----}110$$

$$\text{Exponent} = 1\text{-----}110$$

Full 36-bit Pattern

$$\underline{1111\text{-----}1110} \quad \underline{101100\text{-----}0}$$

12 bits

24 bits

Range: The PDP-8 format has a much larger exponent range

far beyond IEEE standard $(-2048 \rightarrow 2047)$

± 127 for single ^{Vs} ± 1023 for double

Precision: PDP-8 matches single-precision's decimal digits

$$2^{-23} \rightarrow 23 \log_{10} 2 = 23 \times 0.3 \approx 6 \text{ decimal digits}$$

but lack the hidden bit

Double precision offers 16 digits with 52 bits

$$52 \log_{10} 2 = 52 \times 0.3 = 16$$

Q3 Solution

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

$$= -0.00101 \times 2^0$$

move the binary point two to the right

$$= -0.101 \times 2^{-2}$$

exponent = -2,

fraction = -0.101000000000000000000000

answer: 11111111110 101100000000000000000000