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 $= 24 - 127 = -103$

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

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

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

Q2 Solution
0 0C000000 = 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 = 0 0 0 0 1 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

sign is positive
exponent = 0 x18 = 24-127= -103
there is a hidden 1
mantissa = 0
answer 1.0 x 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 × 100 = 111111.01 × 20
normalize, move binary point five to the left
1.1111101 × 25
sign = positive, 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 x 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.

-1.5625×10^{-1}

exponent	Fraction
12 bits	24 bits
Two's complement	Two's complement

No Hidden 1 is used

$$
\begin{array}{r|c}
0.15625 \times 2 & 0 \\
0.3125 \times 2 & 0 \\
0.625 \times 2 & 1 \\
0.25 \times 2 & 0 \\
0.5 \times 2 & 1
\end{array}
$$

$-0.15625 \longrightarrow -0.00101$

No leading 1
No hidden bit

-0.101×2^{-2}

Fraction = 101

exponent = -2

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

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

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

$= 0.625$

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

In Two's complement

$+0.625 \longrightarrow 0.10100\cdots\cdots 0$

Invert bits $\longrightarrow 1.010111\cdots\cdots 1$

add 1 $\longrightarrow 1.011000\cdots\cdots 0$

Fraction number $\longrightarrow 101100\cdots\cdots 0$

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

-2 in 12-bit two's complement

$+2$: $0\cdots 00010$

Invert: $1\cdots\cdots 1101$

add 1: $1\cdots\cdots 110$

Exponent $= 1\cdots\cdots 1110$

Full 36-bit Pattern

$\underbrace{1111\cdots 1110}_{12\text{ bits}}\ \underbrace{101100\cdots\cdots 0}_{24\text{ bits}}$

$\boxed{\text{Range}}$: The PDP-8 format has a much larger exponent range

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

vs

± 127 for single ± 1023 for double

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

$2^{-23} \to 23 \log_{10} 2 = 23 \times 0.3 \simeq 6$ 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 × 10-1 = −0.15625 × 100
= −0.00101 × 20
move the binary point two to the right
= −0.101 × 2^{-2}
exponent = −2,
fraction = −0.1010000000000000000000000
answer: 111111111110 1011000000000000000000000