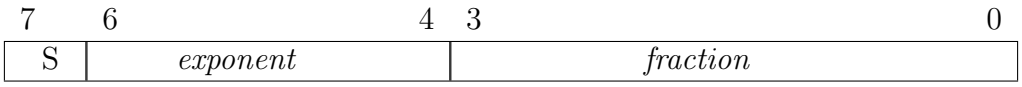
►Solution◄

Question 1: (20 points)

NVIDIA has defined a "tiny-precision" floating pointing format for use in its Graphics Processing Units (GPUs). A floating-point number is represented in this format in 8 bits as follows: the most significant bit is the sign bit, next there are 3 bits used for the exponent, and 4 bits for the fraction. This format is illustrated below:



The exponent is expressed in excess-8 format (also known as a bias representation). Given the binary representation above, the decimal value of the number represented can be computed by the following expression:

$$N = \begin{cases} (-1)^S \times 0.0 & \text{if } exponent = 0 \text{ and } fraction = 0 \\ (-1)^S \times 0.fraction \times 2^{-2} & \text{if } exponent = 0 \text{ and } fraction \neq 0 \\ (-1)^S \times 1.fraction \times 2^{exponent-3} & \text{if } 0 < exponent < 7 \\ (-1)^S \times \infty & \text{if } exponent = 7 \text{ and } fraction = 0 \\ NaN & \text{if } exponent = 7 \text{ and } fraction \neq 0 \end{cases}$$

a. (8 points) Give the bit pattern for the representation of the number -1.25_{10} in this notation.

Solution: Pattern = 1 011 0100
$-1.25_{10} = -(2^0+2^{-2}) = (-1)^1 \times 1.01 \times 2^0 \Rightarrow exponent-3 = 0 \Rightarrow exponent = 3$

b. (8 points) What is the largest positive number that can be represented in this format? Give both the binary pattern in the representation and the decimal value.

Solution: Pattern = 0 110 1111

$$\begin{aligned} N = (-1)^0 \times 1.1111_2 \times 2^{6-3} &= 1 \times (2^1 - 2^{-4}) \times 2^3 = 2^4 - 2^{-1} \\ &= 16 - 0.5 \\ &= 15.5 \end{aligned}$$

c. (4 points) What is the smallest (closest to zero) positive number that can be represented in this format? Give both the bit pattern and the decimal value.

Solution: Pattern = 0 000 0001

$$\begin{aligned} N = (-1)^0 \times 0.0001_2 \times 2^{-2} &= 1 \times 2^{-4} \times 2^{-2} \\ &= 2^{-6} = \frac{1}{2^6} = \frac{1}{2^6} \\ &= \frac{1}{64} = 0.015625 = 1.56 \times 10^{-2} \end{aligned}$$