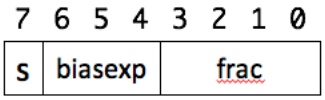
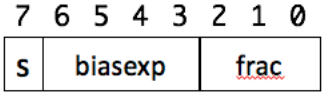
Question 2 (10 points):

(A)



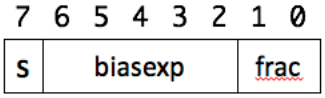
$$N = \begin{cases} (-1)^S \times 0.\text{fraction} \times 2^{-2} & \text{if biasexp} = 0 \\ (-1)^S \times 1.\text{fraction} \times 2^{\text{biasexp}-3} & \text{if } 0 < \text{biasexp} \leq 7 \end{cases}$$

(B)



$$N = \begin{cases} (-1)^S \times 0.\text{fraction} \times 2^{-6} & \text{if biasexp} = 0 \\ (-1)^S \times 1.\text{fraction} \times 2^{\text{biasexp}-7} & \text{if } 0 < \text{biasexp} \leq 15 \end{cases}$$

(C)



$$N = \begin{cases} (-1)^S \times 0.\text{fraction} \times 2^{-14} & \text{if biasexp} = 0 \\ (-1)^S \times 1.\text{fraction} \times 2^{\text{biasexp}-15} & \text{if } 0 < \text{biasexp} \leq 31 \end{cases}$$

Figure 1: Three alternative formats for 8-bit floating-point representation.

Recently Microsoft announced that they are now using an 8-bit floating-point representation for their Field-Programmable Gate Array (FGPA) hardware to support deep neural networks. However, they have not disclosed the specifics of this format. Figure 1 shows three possibilities for the definition of and FP8 format. All these formats assume that there is no need for special value representations such as NaN and $\pm\infty$.

1. (**4 points**) Which format(s) is(are) suitable if the upper limit of the range of values that must be represented is 448.0_{10}. Explain your answer.

   Both formats B and C can be used. The largest number that can be represented in format B is $1.111 \times 2^8 = 2^9 - 2^5 = 512 - 32 = 480$.
   Obviously the maximum value that can be represented with format C is much larger.
   The largest value that can be represented by format A, $1.11 \times 2^3 = 2^4 - 2^1 = 16 - 2 = 14$, is too small.

2. (**6 points**) Let $X = 2.25$. What is the binary representation of X in each of the three formats? If the value cannot be represented in the format, represent the closest value, rounding to the nearest even, and indicate that it is an approximation.

   $2.25 = 10.01 = 1.001 \times 2^1$

   (a) format A: biasexp-3 $= 1 \Rightarrow$ biasexp $= 4 = 100_2$
   0100 0010

   (b) format B: biasexp-7 $= 1 \Rightarrow$ biasexp $= 8 = 1000_2$
   0100 0001

   (c) format C: biasexp-15 $= 1 \Rightarrow$ biasexp $= 16 = 10000_2$
   0100 0000 $= 2.0$