Mock Quiz #1

A normal number can be written in floating point representation as,

$$(-1)^{sign} \quad 1.M \;\; 2^{e-bias}$$

where M is mantissa and e is the unbiased exponent.

1. For double-precision (IEEE 64-bit) floating point number representation, write the number of bits used to store the following:
    a. Sign
    b. Exponent
    c. Mantissa

2. What is the value of bias for 64-bit floating point representation? Why do we need a bias?

3. What is the largest floating-point number that can be represented in double precision?
   Hint: largest number will have the largest mantissa and the largest exponent.

4. What is the smallest normalized floating-point number that can be represented in 64-bit precision?
5. What are subnormal numbers? What is the smallest subnormal number that can be written in double precision (IEEE 64-bit) format?

6. Show the results of the addition of the following two numbers in double-precision format.

   a. 2^{53} + 1
   b. $2^{53} + 2$