1.8 Write the number 38.8125 in the following forms (in part (*c*), follow the IEEE-754 standard): (*a*) Binary form. (*b*) Base 2 floating point representation. (*c*) 64-bit double-precision string.

Solution

(*a*) The largest power of 2 that can be divided into 38 is $2^5 = 32$. Thus, the number 38 in binary form is 100110. The number 0.8125 in binary form is simply $1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-4}$ or 0.1101. Thus, the number 38.8125 in binary form is: 100110.1101.

(*b*) Using part (*a*), the binary floating point representation of 256.1875 is:

$$\frac{38.8125}{2^5}\times 2^5 = 1.212890625\times 2^5$$

(*c*) According to the IEEE-754 standard, 38.8125 in single precision form is as follows:

- Since the number is positive, the first bit is 0
- From part (*b*), the exponent is 5. Adding a bias of 1023, the value of the exponent that must be stored is 5+1023 = 1028. The number 1028 in binary form is:

  $$1028 = 1\times 2^{10} + 1\times 2^2 = 1024 + 4$$

  Thus the number 1028 in binary form is 10000000100. In double precision, 11 bits can be used to store the exponent so that 1028 is stored as 10000000100 without the need for rounding or chopping.
- Next, the mantissa 0.212890625 is converted to binary form: $1\times 2^{-3} + 1\times 2^{-4} + 1\times 2^{-6} + 1\times 2^{-7} + 1\times 2^{-9}$ or, 0.001101101.
- Since 52 bits are allocated for the mantissa, the binary number stored is 0011011010000000000000000000000000000000000000000000

Thus, the number 256.1875 in double precision is stored as: |0|10000000100|0011011010000000000000000000000000000000000000000000|.