Floating-point Exercises

Floating-point Exercise Questions:

1) Converting from decimal/binary to IEEE 754 32-bit Floating-point Standard.

Link to converter: https://www.h-schmidt.net/FloatConverter/IEEE754.html

The IEEE Standard for Floating-point Arithmetic (IEEE 754) divides the 32-bit sequence into three fields:
1) Sign field (1 bit)
2) Exponent filed (8 bits) — Biased notation representation with bias of 127
3) Significand (23 bits)

S	Exponent (Biased Notation)	Significand

To convert from decimal to the IEEE Floating-point Standard:
1) If number sign is positive, we place 0 in sign field, otherwise we place 1 in sign field
2) Convert decimal magnitude (ignore sign) to binary
3) Place number in normalized scientific notation
4) If exponent is not less than or equal to -127:
   A. Convert exponent to biased notation with bias 127 (add 127 then convert to unsigned binary), then fill the significand field with that binary value
   B. Fill significand field with significand
5) If exponent is less than or equal to -127, then we should use the denormalized format:
   A. Place number in this format: .xxxxx * 2^{-126}
   B. Fill significand field with significant (no implied 1 to the left of binary point)
   C. Fill exponent field with 00000000

Example 01: Convert the following decimals to the IEEE Floating-point Standard:

A) 20.75

We place 0 in sign field since number is positive, we then convert number magnitude to binary:

20.75 is equal to 10100.11

We place in normalized scientific notation:

1.010011 * 2^4

Since exponent is not less than or equal to -127, we convert exponent to biased notation with bias 127

4 + 127 = 131. The binary representation of 131 is 10000011 (1 + 2 + 128)

S	Exponent (Biased Notation)	Significand
0	1 0 0 0 0 0 1 1	0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Floating-point Exercises

B) -120

We place 1 in sign field since number is negative, we then convert number magnitude to binary:

120 is equal to

1111000 (8 + 16 + 32 + 64)

We place in normalized scientific notation:

$1.111000 * 2^6$

Since exponent is not less than or equal to -127, we convert exponent to biased notation with bias 127

6 + 127 = 133. The binary representation of 133 is 10000101 (1 + 4 + 128)

S	Exponent (Biased Notation)								Significand																					
1	1	0	0	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Example 02: Convert the following binaries to the IEEE Floating-point Standard:

A) 1110001010.111001001001001

*Unless there was a negative sign before the binary, you should assume that the binary is positive.

First step is to place binary in normalized scientific notation. To do that we need to move the binary point to the left (9 positions)

$1110001010.111001001001001 = 1.110001010111001001001001 * 2^9$

We notice that the significand in this case is more than 23 bits, therefore we cannot convert that binary to the IEEE floating-point representation

Floating-point Exercises

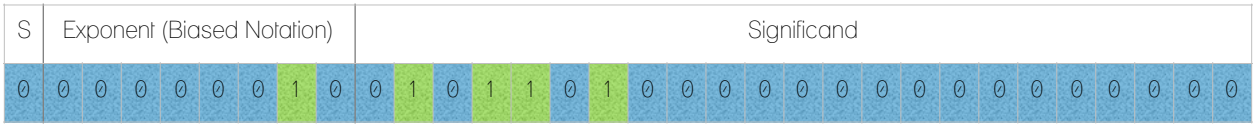
B) $10101.101 * 2^{-129}$

First step is to place binary in normalized scientific notation. To do that we need to move the binary point to the left (4 positions)

$10101.101 * 2^{-129} = 1.0101101 * 2^{4} * 2^{-129} = 1.0101101 * 2^{-125}$

Since exponent is not less than or equal to -127, we convert exponent to biased notation with bias 127

$-125 + 127 = 2$. The binary representation of 2 is 00000010



C) $.00010101101101 * 2^{-125}$

First step is to place binary in normalized scientific notation. To do that, we need to move the binary point to the right (4 positions)

$.00010101101101 * 2^{-125} = 1.010101101 * 2^{-4} * 2^{-125} = 1.010101101 * 2^{-129}$

Since exponent is less than or equal to -127 we need to place number in this format $.xxxxx * 2^{-126}$; to do that we need to move the binary point to the left by three positions:

$1.010101101 * 2^{-129} = .001010101101 * 2^{3} * 2^{-129} = .001010101101 * 2^{-126}$

Now we can place significand in significand field and place 00000000 in exponent field (when exponent is all 0s, that means that there is no implied 1 to the left of the binary point and that the exponent value should be -126)