Computer Architecture HW # 3
Understanding Integers and Floating Point

Name: Xu Haoran
Student ID: 72344187

1. What decimal values do the following 32-bit twos-complement integer numbers represent?

a. 0000 0000 0000 0000 0000 0000 0000 1111
b. 0000 0000 0000 0000 0000 0000 0001 0000
c. 0000 0000 0000 0000 0000 0000 0010 0000
d. 1111 1111 1111 1111 1111 1111 1111 1111
e. 1111 1111 1111 1111 1111 1111 1111 0000

a. 2^0 + 2^1 + 2^2 + 2^3 = 15
b. 2^4 = 16
c. 2^5 = 32
d. - 2^32 + 2^31 + 2^30 + … + 2^1 + 2^0 = -1
e. -2^32 + 2^31 + 2^30 + … + 2^5 + 2^4 = -1 - (2^0 + 2^1 + 2^2 + 2^3) = -16

2. Tell me what decimal values the following 32-bit floating point numbers represent:
(First, identify the sign bit, exponent, and significand, then work from there. Also, f and g are only slightly different, but I want to know what that difference is!)

a. 0000 0000 0000 0000 0000 0000 0000 0000
b. 0011 1111 1000 0000 0000 0000 0000 0000
c. 1011 1111 1000 0000 0000 0000 0000 0000
d. 0011 1111 0000 0000 0000 0000 0000 0000
e. 0100 0000 0000 0000 0000 0000 0000 0000
f. 0100 0000 0100 0000 0000 0000 0000 0000
g. 0100 0000 0100 0000 0000 0000 0000 0001
h. 0100 0000 0100 1001 0000 1111 1101 1010

(This answer follows IEEE-754 standard)
32-bit floating point number:

0(1 bit)	0000000(8 bits)	00000000000000000000000(23 bits)
sign(s)	exponent (E)	fraction (M)

decimal value:
for normal number: V = (-1)^s * (1+M) * 2^(E-127)
for subnormal number (E=0): V = (-1)^s * M * 2^(-126)

a. 0 00000000 00000000000000000000000
S = 0, M = 0, E = 0 (subnormal)
V = (-1)^0 * 0 * 2^(-126) = +0

b. 0 01111111 00000000000000000000000
S = 0, M = 0, E = 127
V = (-1)^0 * (1+0) * 2^(127-127) = 1 * 1 * 1 = 1

c. 1 01111111 00000000000000000000000
S = 1, M = 0, E = 127
V = (-1)^1 * (1+0) *2^(127 – 127) = -1 * 1 * 1 = -1

d. 0 01111110 00000000000000000000000
S = 0, M = 0, E = 126
V = (-1)^0 * (1+0) * 2^(126-127) = 1 * 1 * 2^(-1) = 0.5

e. 0 10000000 00000000000000000000000
S = 0, M = 0, E = 128
V = (-1)^0 * (1+0) * 2^(128-127) = 1 * 1 * 2^1 = 2

f. 0 10000000 10000000000000000000000
S = 0, M = 2^(-1) = 0.5, E = 128
V = (-1)^0 * (1+0.5) * 2^(128-127) = 1 * 1.5 * 2^1 = 3

g. 0 10000000 10000000000000000000001
S = 0,
M = 2^(-1) + 2^(-23) = 0.500000119209289550781 25
E = 128
V = (-1)^0 * (1+0.500000119209289550781 25) * 2^(128-127)
= 1 * 1.500000119209289550781 25 * 2^1
= 3.000000238418579101562 5

h. 0 10000000 1001 0010 0001 11111011010
S = 0,
M = 2(-1) + 2(-4) + 2(-7) + 2(-12) + 2(-13) + 2(-14) +
2^(-15) + 2^(-16) + 2(-17) + 2(-19) + 2(-20) + 2(-22)
= 0.5 + 0.0625 + 0.0078125 + 0.000244140625 + 0.0001220703125 + 0.00006103515625 +
0.000030517578125+ 0.0000152587890625 + 0.00000762939453125 +
0.0000019073486328125 + 0.00000095367431640625 + 0.000000238418579101562 5
= 0.5707962512969970703125
E = 128
V = (-1)^0 * (1+0.5707046985626220703125) * 2^(128-127)
= 1 * 1.5707962512969970703125 * 2
= 3.141592502593994

The difference between f and g is 0.000000238418579101562 5, which is equal to 2^(-22). In binary, this difference corresponds to 0.00000000000000000000001, which is twice the smallest positive difference representable (ULP) in a 32-bit floating point number.