ICS Homework 13 Solution

Floating Point

Consider a 16-bit floating point representation based on the IEEE floating-point format, with 1 sign bit, 6 exp bits, 9 frac bits, called **Float16**.

(1) Fill in the table below. Please represent M in the form `x` or `x/y` where `x` is an integer and `y` is an integral power of 2.

Description	Hex	M	E
-21/2	0xC4A0	21/16	3
5/8	0x3C80	5/4	-1
-85/64	0xBEA8	85/64	0
$-3*2^{-34}$	0x8060	3/16	-30
32	0x4800	1	5
-0	0x8000	0	-30
Largest negative normalized value	0x8200	1	-30
+∞	0x7E00	--	--
Largest denormalized value	0x01FF	511/512	-30

Suppose the Float16 is formatted with 1 sign bit, 5 exp bits, 10 frac bits.

(2) Assume we use IEEE round-to-even mode to do the approximation. Now a, b are both Float16, with a = 0x4663 and b = 0x394c represented in hex. Compute a+b and represent the answer in hex.

0x470c

(3) Using Float16, what's the difference between $2^{15} + 0.5 - 2^{15}$ and $2^{15} - 2^{15}+0.5$? Calculate them to explain why.

2^15 : 0|111 10|00 0000 0000
0.5: 0|011 10|00 0000 0000
2^15+0.5:
1.0000 0000 00
+ 0.0000 0000 0000 0001 0000 0000 00
= 1.0000 0000 0000 0001 0000 0000 00
M=(1.0000 0000 00|00 0001 0000 0000 00)2 =1.0000 0000 00
E=30
2^15+0.5 = 0|111 10|00 0000 0000 = 2^15
So 2^15+0.5-2^15 = 0

But 2^15-2^15=0
2^15-2^15+0.5=0.5

0.5 is rounded during the calculation of 2^15 + 0.5.