ICS Homework 13

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			
	0xBEA8		
$-3*2^{-34}$			
	0x4800		
-0			
Largest negative normalized value			
$+\infty$		--	--
Largest denormalized value			

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.

(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.