ICS Homework 13

Floating Point Operations

Consider a 16-bit floating point representation based on the IEEE floating-point format, with 1 sign bit, 5 exp bits, 10 frac bits, called **Float16**.

(1) 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.

(2) 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.