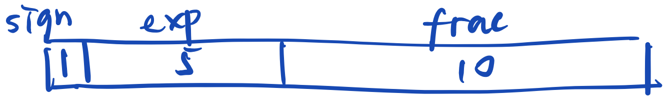
ICS Homework 13



Floating Point

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**.

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.

10101 (15) -12

Description	Hex	M	E
-21/2	0xC940	21/16	3
5/8	0x3900	5/4	-1
-85/64	0xBD50	85/64	0
$-3*2^{-18}$	0x80C0	3/16	-14
32	0x5000	1	5
-0	0x8000	0	-14
Largest negative normalized value	0x8400	1	-14
+∞	0x7C00	-	-
Largest denormalized value	0x03FF	$\frac{1023}{1024}$	-14

1|011 11|01 0101 0000

15-15=0

$\frac{5+16}{2^6} = \frac{21}{64}$ +1

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.

0x470C

(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.

2^15: 0|111 10|00 0000 0000

0.5: 0|011 10|00 0000 0000

2^15+0.5: 1.0000 000000

+ 0.0000 0000 0000 0001 0000 0000 00

= 1.0000 0000 0000 0001 0000 0000 00

M =1.0000 0000 00 E=30

2^15+0.5 = 0111 1000 0000 0000 = 2^15

∴ 2^15+0.5-2^15 =0

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

00 1010 1001

a: 0|100 01|10 0110 0011 $*2^2$

b: 0|011 10|01 0100 1100 $*2^{-1}$

11 0000 1100

0|100 01|11 0000 1100

0x470c