CS 367 - Recitation #3 Worksheet **Floating Point Representation**

NameAdilet Kuroda _____ **G#**3 _____

Group Member Name: Prashant Shrestha _____

Group Member Name: _____

Today's Goals: We want to get comfortable with floating point representation and binary fractional rounding.

Work in groups of 2-3 students. Every group will turn in what they've got at the end of recitation on paper, with all the members' names written down.

Grading is based on participation. Get as much done as you can. You will also be given feedback in the form of a 'score' (1-3) and possibly some comments. This doesn't affect your grade – it is solely for feedback. A score of 3 means everything looks great. A score of two indicates some minor problems. And a score of 1 indicates that there were some major issues. If you get a 1, don't panic - go see your prof or a GTA to get more extensive feedback.

For this recitation, assume **12 bit floating point representation** which uses the first bit as a sign bit, the next 6 for the exponent and the last 5 for the fraction:

s (1 bit)	exp (6 bit)	frac (5 bit)

Floating Point Range

Floating point can be classified into **normalized**, **denormalized**, or **special** depending on the exp bit sequence. Fill in **normalized**, **denormalized**, or **special** in the table. If special, write the special value.

0 000000 00000	D	1 111100 10111	N
1 011011 11111	N	0 111111 00000	S (infinity)
1 111111 10011	S (NaN)	1 000000 01100	D

How are each of these (12bit) values represented? Show their components by filling in the chart below.

	s	exp (in bits)	frac (in bits)	Decimal Value
Largest possible normalized	0	111110	11111	$1\ 31/32 \times 2^{31}$
Smallest possible (non-zero) positive normalized	0	000001	00000	1×2^{-30}
Largest possible de-normalized	0	000000	11111	$31/32 \times 2^{-30}$
Smallest possible (non-zero) positive de-normalized	0	000000	00001	$1/32 \times 2^{-30}$

s (1 bit)	exp (6 bit)	frac (5 bit)

Floating Point Conversion

Here we will learn the conversion between float and binary. Note that we are still following the 12bit floating point representation as in the top of the page.

What is the bias value?	31

Fill in the chart below. `s`, `E`, and `M` are from the actual represented quantity of `Value = (-1)s*M*2E`. `exp` and `frac` are from the bit-encodings of `E` and `M` respectively. Be careful determining the classification and then using the appropriate encoding approach to navigate between `E` and `exp` and between `M` and `frac`.

Fill in the chart using the types listed. (dec) is Base-10 decimal. X/32 is a fraction. Y/32 is an improper fraction (eg. 43/32). (+/- Y/32 * 2^E) is an improper fraction. (eg. -45/32 * 2^-3) for $-45/32 * 2^{-3}$

Bit representation	s	exp (dec)	E (dec)	frac (X / 32)	M (Y/32)	Value (+/- Y/32 * 2^E)
0 011000 01011	0	24	-7	11/32	43/32	+43/32*2^-7
1 100111 10010	1	39	8	18/32	50/32	-50/32*2^8
0 000000 11110	0	0	-30	30/32	30/32	+30/32*2^-30
0 100011 10000	0	35	4	16/32	48/32	+48/32*2^4
0 100100 10100	0	36	5	20/32	52/32	$52/32 * 2^5$
0 011100 01100	0	28	-3	12/32	44/32	$44/32 * 2^{-3}$
0 011000 10000	0	24	-7	16/32	48/32	$48/32 * 2^{-7}$

Floating Point Rounding

Fill in the table below. Rounding of the binary fractional numbers should be done to nearest 1/8 (3bits right of binary point) following the round-to-even rule.

Value	Binary	Rounded Binary	Rounded Value
3 3/32	11.00011	11.001	3 1/8
5/16	0.010	0.010	1/4
2 29/32	**10.11101**	10.111	2 7/8
1 11/32	**1.01011**	1.011	1 3/8