CS 4400 - Problem Set 3

Rob Johansen

u0531837

1. Problem 2.86:

Description	Hex	M	E	V
-0	0x8000	0	0	——
Smallest value > 2	0x4001	257/256	1	257 x 2^{-7}
512	0x4800	1	9	——
Largest denormalized	0x00FF	255/256	-62	255 x 2^{-53}
−∞	0xFF00	——	——	——
Hex number 3BB0	——	477/256	13	477 x 2^{5}

2. Problem 2.87:

Format A		Format B	
Bits	**Value**	**Bits**	**Value**
1 01111 001	-9/8	1 0111 0010	-9/8
0 10110 011	176	0 1110 0110	176
1 00111 010	-5/1024	1 1111 0100	NaN
0 00000 111	7/(2^{17})	0 1000 1100	7/(2^{17})
1 11100 000	-0	1 0100 0000	-0
0 10111 100	384	0 1111 1000	NaN

3. Problem 2.88:

A. Yes. Although casting dx from double to float is technically a loss of precision, this will always yield 1 because the original type from which dx came was int.

B. No. This will not always yield 1 because the (x-y) operation could result in overflow before being cast to double, whereas the operation dx - dy cannot overflow. One example is when x is TMax and y is -1.

C. Yes. Although floating-point addition is not associative, it must again be noted that int was the original type of dx, dy, and dz. Thus, they cannot possibly contain large/small enough values to create associativity problems.

D. No. Floating-point multiplication is not associative, even if the operands were originally cast from int. One example is when dx, dy, and dz are all TMax.

E. No. If either side of the expression results in NaN, the expression will yield 0.