Math 1310 – Technical Math for IT

ASSIGNMENT 3 Name:_____Markus Afonso_____

Due: October 10th at 11:59 PM (all three sets)

Online submission, ONE pdf file ID :____ A01333486______

After completing all the questions on separate paper, place all answers on THIS sheet. Be sure to attach your work <u>showing all intermediate steps in a clear and well organized fashion</u> for full credit.

1. [2]

Convert the following single precision IEEE floating point number to decimal.

1	1011 1011	01011010000000000000000

$\underline{1}$ 1011 1011 01011010000 0000 0000 0000

$1011\ 1011 = 2^7 + 2^5 + 2^4 + 2^3 + 2^1 + 2^0$

$= 187 - 127$

$= 60$

-1.0101101×2^{60}

$\boxed{= -1.558245471 \times 10^{18}}$

2. [3] The following sequence of 32 bits is stored in memory:

1010 1110 1111 1011 0100 0000 0000 0000

What is the decimal value of the number stored if the binary string given represents a number in:

a) Unsigned binary form? 2935701504

b) Twos-complement binary form? -1359265792

c) IEEE-754 single precision floating point form?
$-1.142552719 \times 10^{-10}$

2a)

1010 1110 1111 1011 0100 0000 0000 0000

$2^{31} + 2^{29} + 2^{27} + 2^{26} + 2^{25} + 2^{23} + 2^{22} + 2^{21} + 2^{20} + 2^{19} + 2^{17} + 2^{16} + 2^{14}$

$\boxed{= 2935701504}$

b)

31	27	23	19	15	11	7	3
1010	1110	1111	1011	0100	0000	0000	0000
0101	0001	0000	0100	1011	1111	1111	1111
							+1
0101	0001	0000	0100	1100	0000	0000	0000

$= 2^{30} + 2^{28} + 2^{24} + 2^{18} + 2^{15} + 2^{14}$

$= 1359265792$

$\boxed{= -1359265792}$

c) 1 0101 1101 11110110100000000000000

$0101\ 1101 = 64 + 16 + 8 + 4 + 1$

$= 93 - 127$

$= -34$

$-1.111101101 \times 2^{-34}$

$\boxed{= -1.142552719 \times 10^{-10}}$

3. [7] Find the decimal number corresponding to each of the mini-standard floating point representations in the table to the right.
Be sure to check special cases!

Math 1310 – Technical Math for IT

Mini-Standard Floating Point Rep	Decimal Number
(a) 1 0000 00000	−0 SC #1
(b) 1 1111 00000	−∞ SC # 2
(c) 0 0101 01110	0.359375
(d) 0 0010 01110	0.044921875
(e) 0 0000 00110	0.001464844
(f) 1 1111 11101	NaN SC # 3
(g) 1 1110 01010	−168

3a) 1 0000 0000 = $(-0)_{10}$ (special case 1)

b) 1 1111 0000 = $-\infty$ (special case 2)

c) 0 0101 01110

0101 = 4+1
= 5 −7
= −2

1.01110×2^{-2}

$2^{-2} + 2^{-4} + 2^{-5} + 2^{-6}$

$\boxed{= 0.359375}$

e) 0 0000 00110 unnormilized

= −7

0.0011×2^{-7}

1.1×2^{-10}

$= 2^{-10} + 2^{-11}$

$\boxed{= 0.001464844}$

d) 0 0010 01110

0010 = 2 − 7
= −5

1.01110×2^{-5}

$= 2^{-5} + 2^{-7} + 2^{-8} + 2^{-9}$

$\boxed{= 0.044921875}$

f) 1 1111 11101 (special case 3)

$\boxed{= NaN}$

g) 1 1110 01010

1110 = 14 − 7
= 7

-1.0101×2^{7}

$\boxed{= -168}$

4. [12 marks] Use the **mini-standard** floating point representation (1 sign bit, 4 exponent bits, and 5 mantissa bits assuming the hidden bit, with the exponent recorded in bias 7) to perform the arithmetic operations below. Each of the following steps should be performed for each problem (use the template to record the results, but attach sheets with detailed work to support these results):

i. Both of the given numbers should first be coded in the mini-standard.

ii. the numbers in the mini-standard should converted back to decimal form and the precise loss of precision recorded.

iii. The addition or subtraction should be performed in standardized form with **only five bits to the right of the radix point**.

iv. The result should then be recorded in the normalized mini form, **if possible**.

v. Finally, the result should be interpreted as a **decimal** floating point number**.**

vi. Any **further** loss of precision (resulting from the standardization process or from renormalization) should be noted (yes/no).

a. 38 + 53

	Decimal	Mini-Standard	Convert back to decimal	Actual loss of Precision or "none"
	38	0 1100 00110	38	none
	53	0 1100 10101	53	none
sum	91	0 1101 01101	90	1

b. 70 - 173

	Decimal	Mini-Standard	Convert back to decimal	Actual loss of Precision or "none"
	70	0 1101 00011	68	2
	-173	1 1110 01011	-172	-1
sum	-103	1 1101 10100	-104	1

c. 101.4 + 5.525

	Decimal	Mini-Standard	Convert back to decimal	Actual loss of Precision or "none"
	101.4	0 1 1 0 1 1 0 0 1 0	100	1.4
	5.525	0 1 0 0 1 0 1 1 0 0	4	1.525
sum	106.925	0 1 1 0 1 1 0 1 0 0	104	2.925

4)

i) $38 = 32 + 4 + 2$

00100110

1.00110×2^5

$5 + 7 = 12 = 1100$

$= 0\ 1100\ 00110$

$53 = 32 + 16 + 4 + 1$

00110101

1.10101×2^5

$5 + 7 = 12 = 1100$

$= 0\ 1100\ 10101$

ii) 0 1100 00110

$1100 = 12 - 7 = 5$

1.0011×2^5

$= 2^5 + 2^2 + 2^1$

$= 38 = 38$ ← no precision lost

0 1100 10101

$1100 = 12 - 7 = 5$

1.10101×2^5

$2^5 + 2^4 + 2^2 + 2^0$

$= 53 = 53$ ← no preision lost

iii)

$$
\begin{array}{r}
0\ 1100\ \ 1.00110 \\
+\ 0\ 1100\ \ 1.10101 \\
\hline
0\ 1100\ 10.11011
\end{array}
$$

lost

$10.1101\underline{1} \times 2^5$

$= 1.01101 \times 2^6$

$= 2^6 + 2^4 + 2^3 + 2^1$

$= 90 \neq 91$ ← precision lost After renormilization by 1

iv) $1.01101 \times 2^6 = 01101\ 01101$

$6 + 7 = 13 = 1101$

v) $90 = 1.01101 \times 2^6$

vi) precision lost after renormilizing the sum by 1.

b)

i) $70 = 64 + 4 + 2$

01000110

1.00011×2^{6}

$6 + 7 = 13 = 1101$

$= 0\ 1101\ 00011$

$-173 = 128 + 32 + 8 + 4 + 1$

10101101

1.0101101×2^{7} ← lost

$7 + 7 = 14 = 1110$

$1\ 1110\ 01011$

ii) $0\ 1101\ 00011$

$1101 = 13 - 7 = 6$

$1.00011 \times 2^{6} \xrightarrow[\text{greater exponent}]{\text{normilize to}} 0.10001 \times 2^{7}$ ← lost

$= 2^{6} + 2^{2} = 68 \neq 70$ ∴ precision lost after renormilizing of 2

$\underline{1}\ 1110\ 01011$

$1110 = 14 - 7 = 7$

-1.01011×2^{7}

$2^{7} + 2^{5} + 2^{3} + 2^{2}$

$= -172 \neq -173$ ∴ precision lost of -1

iii)
$$\begin{array}{r} \cancel{1}.\cancel{0}1011 \times 2^{1110} \\ -\ 0.10001 \times 2^{1110} \\ \hline 0.11010 \times 2^{1110} \end{array}$$

$$\begin{array}{r} 1.\cancel{1}\cancel{1}\cancel{0}\cancel{0}1 \\ -1.01011 \\ \hline 0.01110 \end{array}$$

$= -0.11010 \times 2^{1110}$

Normilize

iv) $= -1.10100 \times 2^{1101}$

$= 1\ 1101\ 10100$

v) -1.10100×2^{6}

$2^{6} + 2^{5} + 2^{3}$

$= -104$

vi) no further loss of precision

c) i) 101.4

$101 = 64 + 32 + 4 + 1$

$0110\ 0101.\overline{0110}$

$0.4 \times 2 = 0.8$

$0.8 \times 2 = 1.6$

$0.6 \times 2 = 1.2$

$0.2 \times 2 = 0.4$

$0.4 \times 2 = 0.8$

$0110|0110|\ldots$

$01100101.\overline{0110}$ lost

1.10010×2^6

$6 + 7 = 13 = 1101$

$= 0\ 1101\ 10010$

5.525

$5 = 4 + 1$

$= 0000\ 0101.1000\overline{0110}$

$0.525 \times 2 = 1.05$

$0.05 \times 2 = 0.1$

$0.1 \times 2 = 0.2$

$0.2 \times 2 = 0.4$

$0.4 \times 2 = 0.8$

$0.8 \times 2 = 1.6$

$0.6 \times 2 = 1.2$

$0.2 \times 2 = 0.4$

$0.4 \times 2 = 0.8$

$0.8 \times 2 = 1.6$

$0.6 \times 2 = 1.2$

$.1000|0110|0110|0110$

$= .1000\overline{0110}$

$= 0000\ 0101.1000\overline{0110}$ lost

1.01100×2^2

$2 + 7 = 9 = 1001$

$= 0\ 1001\ 01100$

ii) 0 1101 10010

$1101 = 13 - 7 = 6$

1.10010×2^6

$2^6 + 2^5 + 2^2$

$= 100 \neq 101.4$ ∴ precision lost of 1.4

0 1001 01100

$1001 = 9 - 7 = 2$

1.01100×2^2

lost

Normilize to match greater exponent

0.00010×2^6

$= 2^2$

$= 4 \neq 5.525$ ∴ precision loss of 1.525.

iii)

$$\begin{array}{r} 1.10010 \times 2^{1101} \\ 0.00010 \times 2^{1101} \\ \hline 1.10100 \times 2^{1101} \end{array}$$

iv) 1.10100×2^{1101}

$= 0\ 1101\ 10100$

v) 1.10100×2^6

$2^6 + 2^5 + 2^3$

$= 104$

vi) the sum did not need to be renormilized ∴ no further precision lost.