Solutions - Homework 2

(Due date: February 1st @ 7:30 pm)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (12 PTS)

- Calculate the result of the additions and subtractions for the following fixed-point numbers.

UNSIGNED		SIGNED	
1.1011010 + 0.010101	1.00101 - 0.0000111	10.001 + 1.001101	0.011 - 1.1011101
10.1101 + 1.1001	1100.1 + 0.100101	1001.101 - 111.10001	101.0001 + 1.1001001

UNSIGNED:

```
c8=1 c7=1 c6=1 c5=1 c4=1 c3=0 c2=1 c1=0 c0=0
    1.1 0 1 1 0 1 0 +
    0.0 1 0 1 0 1 0
  ---------------------
  1 0.0 0 0 0 1 0 0
```

```
c6=1 c5=1 c4=1 c3=0 c2=0 c1=1 c0=0
    1 0.1 1 0 1 +
    0 1.1 0 0 1
  ---------------
  1 0 0.0 1 1 0
```

```
b8=0 b7=0 b6=0 b5=0 b4=1 b3=1 b2=1 b1=1 b0=0
  1.0 0 1 0 1 0 0 -
  0.0 0 0 0 1 1 1
  -----------------
  1.0 0 0 1 1 0 1
```

```
c10=0 c9=0 c8=0 c7=0 c6=1 c5=0 c4=0 c3=0 c2=0 c1=0 c0=0
  1 1 0 0.1 0 0 0 0 0 +
  0 0 0 0.1 0 0 1 0 1
  ---------------------
  1 1 0 1.0 0 0 1 0 1
```

SIGNED:

```
c9=1 c8=1 c7=0 c6=0 c5=0 c4=1 c3=0 c2=0 c1=0 c0=0
  1 1 0.0 0 1 0 0 0 +
  1 1 1.0 0 1 1 0 1
  -------------------
  1 0 1.0 1 0 1 0 1
```

```
0.0 1 1 0 0 0 0 -
1.1 0 1 1 1 0 1
        ➡
c8=0 c7=0 c6=1 c5=0 c4=0 c3=0 c2=0 c1=0 c0=0
  0.0 1 1 0 0 0 0 +
  0.0 1 0 0 0 1 1
  -----------------
  0.1 0 1 0 0 1 1
```

```
1 0 0 1.1 0 1 0 0 -
1 1 1 1.1 0 0 0 1
        ➡
c9=0 c8=0 c7=0 c6=1 c5=1 c4=1 c3=1 c2=0 c1=0 c0=0
  1 0 0 1.1 0 1 0 0 +
  0 0 0 0.0 1 1 1 1
  -------------------
  1 0 1 0.0 0 0 1 1
```

```
c10=1 c9=1 c8=1 c7=0 c6=0 c5=0 c4=1 c3=0 c2=0 c1=0 c0=0
  1 0 1.0 0 0 1 0 0 0 +
  1 1 1.1 0 0 1 0 0 1
  ---------------------
  1 0 0.1 0 1 0 0 0 1
```

PROBLEM 2 (18 PTS)

- Multiply the following signed fixed-point numbers:

10.011 × 0.110101	10.1101 × 01.10001	0111.111 × 10.011011

```
10.011 x
0.110101
   ⬇
 01.101 x   ➡          1 1 0 1 0 1 x
0.110101                   1 1 0 1
                     -------------
                       1 1 0 1 0 1
                   1 1 0 1 0 1
               1 1 0 1 0 1
             ---------------------
             1 0 1 0 1 1 0 0 0 1
                     ⬇
           0.1 0 1 0 1 1 0 0 0 1
                     ⬇
         1 0.1 0 1 0 0 1 1 1 1
```

```
10.1101 x
01.10001
   ⬇
01.0011 x   ➡          1 1 0 0 0 1 x
01.10001                   1 0 0 1 1
                     -------------
                       1 1 0 0 0 1
                     1 1 0 0 0 1
               1 1 0 0 0 1
             ---------------------
             1 1 1 0 1 0 0 0 1 1
                     ⬇
           0 1.1 1 0 1 0 0 0 1 1
                     ⬇
         1 0.0 0 1 0 1 1 1 0 1
```

```
0111.111 x
10.011011
   ⬇
0111.111 x   ➡             1 1 1 1 1 1 x
01.100101                1 1 0 0 1 0 1
                       ---------------
                           1 1 1 1 1 1
                       1 1 1 1 1 1
                 1 1 1 1 1 1
               1 1 1 1 1 1
             -------------------------
               1 1 0 0 0 1 1 0 1 1 0 1 1
                           ⬇
             0 1 1 0 0.0 1 1 0 1 1 0 1 1
                           ⬇
           1 0 0 1 1.1 0 0 1 0 0 1 0 1
```

- Get the division result (with $x = 4$ fractional bits) for the following signed fixed-point numbers:

101.1001 ÷ 1.0101	11.011 ÷ 1.10111	0.101010 ÷ 101.0101

✓ $\frac{101.1001}{1.0101}$: To unsigned and then alignment, $a = 4$: $\frac{010.0111}{0.1011} = \frac{010.0111}{0.1011} \equiv \frac{100111}{1011}$

```
         0000111000
1011 ) 1001110000
         1011
         -----
          10001
           1011
          -----
            1100
            1011
           -----
               1000
```

Append $x = 4$ zeros: $\frac{100111\mathbf{0000}}{1011}$

Integer Division:

$$Q = 111000, R = 1000$$
$$\rightarrow Qf = 11.1000 (x = 4)$$

Final result (2C): $\frac{101.1001}{1.0101} = 011.1$

✓ $\frac{11.011}{1.10111}$: To unsigned and then to unsigned: $a = 5$: $\frac{00.101}{0.01001} = \frac{0.10100}{0.01001} \equiv \frac{10100}{1001}$

```
         000100011
1001 ) 101000000
         1001
         ----
            10000
             1001
            -----
              1110
              1001
             -----
                101
```

Append $x = 4$ zeros: $\frac{10100\mathbf{0000}}{1001}$

Unsigned Integer Division:

$$Q = 100011, R = 101$$
$$\rightarrow Qf = 10.0011\ (x = 4)$$

Final result (2C): $\frac{11.011}{1.10111} = 010.0011$

✓ $\frac{0.101010}{101.0101}$: To positive (denominator), alignment, and then to unsigned, $a = 5$: $\frac{0.10101}{010.1011} = \frac{000.10101}{010.10110} \equiv \frac{10101}{1010110}$

```
            000000011
1010110 ) 101010000
            1010110
            -------
            10100100
             1010110
            --------
             1001110
```

Append $x = 4$ zeros: $\frac{10101\mathbf{0000}}{1010110}$

Integer Division:

$$Q = 11, R = 10100100$$

$\rightarrow Qf = 0.0011\ (x = 4)$ ✶ Qf here is represented as an unsigned number

Final result (2C): $\frac{0.101010}{101.0101} = 2C(0.0011) = 1.1101$

PROBLEM 3 (10 PTS)

- We want to represent numbers between -214.9 and 256.7. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0015? (5 pts).

  2C representation for integers: -2^{n-1} to $2^{n-1} - 1$. For $2^{n-1} - 1 \geq 256$, we have that $n \geq 10$, so we pick $n = 10$.

  For the fractional part, we select the number of fractional bits p that make the resolution better or equal than 0.0005:

$$2^{-p} \leq 0.0015 \rightarrow p \geq 9.3808 \rightarrow p = 10$$

  Then the Fixed Point format required in $[20\ 10]$.

- Represent these numbers in Fixed Point Arithmetic (signed numbers). Select the minimum number of bits in each case.

-128.625	-231.3125	112.125

✓ 128.625 = 010000000.101 → -128.625 = 101111111.011
✓ 231.3125 = 011100111.0101 → -231.3125 = 100011000.1011
✓ 112.125 = 01110000.001

PROBLEM 4 (12 PTS)

- Complete the table for the following fixed point formats (signed numbers):

Fractional bits	Integer Bits	FX Format	Range	Dynamic Range (dB)	Resolution
7	5	[12 7]	[-16, 15.9922]	66.23	0.0078125
12	4	[16 12]	[-8,7.9998]	90.31	0.0002441
17	7	[24 17]	[-64, 63.99999]	138.47	0.00000763

- Complete the table for these floating point formats (which resemble the IEEE-754 standard). Only consider ordinary numbers.

$min = 2^{-2^{E-1}+2}$, $max = (2-2^{-p})2^{2^{E-1}-1}$, $e \in [-2^{E-1}+2, 2^{E-1}-1]$, $significand \in [1, 2-2^{-p}]$

Exponent bits (E)	Significant bits (p)	Min	Max	Range of e	Range of significand
7	8	2.1684×10^{-19}	1.841×10^{19}	$[-2^6+2, 2^6-1] = [-62,63]$	$[1,1.99609375]$
8	15	1.1755×10^{-38}	3.4027×10^{38}	$[-2^7+2, 2^7-1] = [-126,127]$	$[1,1.999969482421875]$
11	36	2.2251×10^{-308}	1.7977×10^{308}	$[-2^{10}+2, 2^{10}-1] = [-1022,1023]$	$[1,1.999999999985448]$

PROBLEM 5 (16 PTS)

- Calculate the decimal values of the following floating point numbers represented as hexadecimals. Show your procedure.

Single (32 bits)		Double (64 bits)	
✓ FDEAD360	✓ 803ACBAC	✓ FA09D3784D039800	✓ 7FFBEEFC0FFEEBEE
✓ 3DE32856	✓ 7FCBEEFE	✓ DECAFC0FEE000000	✓ 800ABBAF25C00000

✓ FDEAD360: 1111 1101 1110 1010 1101 0011 0110 0000
$e + bias = 11111011 = 251 \rightarrow e = 251 - 127 = 124$
$Mantissa\ ([24\ 23]) = 1.11010101101001101100000 = 1.834575$
$X = -1.834575 \times 2^{124} = -3.9017 \times 10^{37}$

✓ 3DE32856: 0011 1101 1110 0011 0010 1000 0101 0110
$e + bias = 01111011 = 123 \rightarrow e = 123 - 127 = -4$
$Mantissa = 1.11000110010100001010110 = 1.774668455123901$
$X = 1.774668455123901 \times 2^{-4} = 0.110916778445244$

✓ 803ACBAC: 1000 0000 0011 1010 1100 1011 1010 1100
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$
$Mantissa = 0.01110101100101110101100 = 0.459340572$
$X = -0.459340572 \times 2^{-126} = -5.3995224 \times 10^{-39}$

✓ 7FCBEEFE: 0111 1111 1100 1011 1110 1110 1111 1110
$e + bias = 11111111 = 255, f \neq 0$
$X = NaN$

✓ FA09D3784D039800: 1111 1010 0000 1001 1101 0011 0111 1000 0100 1101 0000 0011 1001 1000 0000 0000
$e + bias = 11110100000 = 1952 \rightarrow e = 1952 - 1023 = 929$
$Mantissa\ ([53\ 52]) = 1.1001110100110111100001001101000000111001 1 = 1.61412$
$X = -1.6142 \times 2^{929} = -7.3249 \times 10^{279}$

✓ 7FFBEEFC0FFEEBEE: 0111 1111 1111 1011 1110 1110 1111 1100 0000 1111 1111 1110 1110 1011 1110 1110
$e + bias = 11111111111 = 2047, f \neq 0$
$X = NaN$

✓ DECAFC0FEE000000: 1101 1110 1100 1010 1111 1100 0000 1111 1110 1110 0000 0000 0000 0000 0000 0000
$e + bias = 10111101100 = 1516 \rightarrow e = 1516 - 1023 = 493$
$Mantissa = 1.1010111111000000111111110111 = 1.686538$
$X = -1.686538 \times 2^{493} = -4.313 \times 10^{148}$

✓ 800ABBAF25C00000: 1000 0000 0000 1010 1011 1011 1010 1111 0010 0101 1100 0000 0000 0000 0000 0000
$e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
$Mantissa\ ([53\ 52]) = 0.101010111011101011110010010111 = 0.6708213$
$X = -0.6708213 \times 2^{-1022} = -1.492627 \times 10^{-308}$

PROBLEM 6 (32 PTS)

- Calculate the result (provide the 32-bit result) of the following operations with 32-bit floating point numbers. Truncate the results when required. When doing fixed-point division, use 8 fractional bits. Show your procedure.

✓ 40D90000 + C2EAC000	✓ 801A8000 – B3CEC000	✓ FACADE80 × 7F800000	✓ 800C0000 ÷ 494A0000
✓ CF4A8000 + B0A90000	✓ FF800000 – DECAFF00	✓ 8B092000 × 0FACE000	✓ 49744000 ÷ C0C90000

✓ $X =$ 40D90000 + C2EAC000:
40D90000: 0100 0000 1101 1001 0000 0000 0000 0000
$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $Significand = 1.1011001$
40C00000 $= 1.101101 \times 2^2$

C2EAC000: 1100 0010 1110 1010 1100 0000 0000 0000
$e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $Significand = 1.110101011$
C2EA9000 $= -1.110101011 \times 2^6$

$$X = 1.1011001 \times 2^2 - 1.110101011 \times 2^6 = \frac{1.1011001}{2^4} \times 2^6 - 1.110101011 \times 2^6$$
$X = 0.00011011001 \times 2^6 - 1.110101011 \times 2^6$
To subtract these numbers, we first convert to 2C:
$R = 0.00011011001 - 01.110101011$
$R = 0.00011011001 + 10.001010101$ (2C addition)
The result (in 2C) is: $R = 10.01000101101$, $|R| = 01.10111010011$

```
  0 0.0 1 1 1 0 1 0 0 0 0 0
  0 0.0 0 0 1 1 0 1 1 0 0 1 +
  1 0.0 0 1 0 1 0 1 0 1 0 0
  -------------------------
  1 0.0 1 0 0 0 1 0 1 1 0 1
```

For floating point, we need to convert to sign-and-magnitude:
$\Rightarrow R(SM) = -1.10111010011$

$X = -1.1011101001 \times 2^6, e + bias = 6 + 127 = 133 = 10000101$
X = 1100 0010 1101 1101 0011 0000 0000 0000 = C2DD3000

✓ $X =$ CF4A8000 + B0A90000:
CF4A8000: 1100 1111 0100 1010 1000 0000 0000 0000
$e + bias = 10011110 = 158 \rightarrow e = 158 - 127 = 31$ $Significand = 1.10010101$
CF4A8000 $= -1.10010101 \times 2^{31}$

B0A90000: 1011 0000 1010 1001 0000 0000 0000 0000
$e + bias = 01100001 = 97 \rightarrow e = 97 - 127 = -30$ $Significand = 1.0101001$
B0A90000 $= -1.0101001 \times 2^{-30}$

$$X = -1.10010101 \times 2^{31} - 1.0101001 \times 2^{-30} = -1.10010101 \times 2^{31} - \frac{1.0101001}{2^{61}} \times 2^{31}$$
Representing the number divided by 2^{61} requires more than $p + 1 = 24$ bits. Thus, we round down this operand to 0.
$X = -1.10010101 \times 2^{31}, e + bias = 31 + 127 = 158 = 10011110$
X = 1100 1111 0100 1010 1000 0000 0000 0000 = CF4A8000

✓ $X =$ 801A8000 – B3CEC000:
801A8000: 1000 0000 0001 1010 1000 0000 0000 0000
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $Significand = 0.00110101$
801A8000 $= -0.00110101 \times 2^{-126}$

B3CEC000: 1011 0011 1100 1110 1100 0000 0000 0000
$e + bias = 01100111 = 103 \rightarrow e = 103 - 127 = -24$ $Significand = 1.100111011$
B3CEC000 $= -1.100111011 \times 2^{-24}$

$$X = -0.00110101 \times 2^{-126} + 1.100111011 \times 2^{-24} = -\frac{0.00110101}{2^{102}} \times 2^{-24} + 1.10011101 \times 2^{-24}$$

Representing the number divided by 2^{102} requires more than $p+1=24$ bits. Thus, we round down this operand to 0.
$X=+1.100111011\times 2^{-24}$
X = 0011 0011 1100 1110 1100 0000 0000 0000 = 33CEC000

✓ X = FF800000 – DECAFF00:
FF800000: 1111 1111 1000 0000 0000 0000 0000 0000
$e+bias=11111111=255, f=0$
FF800000 = $-\infty$
$X=(-\infty)-\#=-\infty$
X = FF800000

✓ X = FACADE80 × 7F800000:
7F800000: 0111 1111 1000 0000 0000 0000 0000 0000
$e+bias=11111111=255, f=0$
7F800000 = $+\infty$
$X=(-|\#|)\times+\infty=-\infty$
X = 1111 1111 1000 0000 0000 0000 0000 0000 = FF800000

✓ X = 8B092000 × 0FACE000:
8B092000: 0000 1011 0000 1001 1010 0000 0000 0000
$e+bias=00010110=22\rightarrow e=22-127=-105$ $Significand=1.0001001101$
8B092000 = $-1.0001001101\times 2^{-105}$

0FACE000: 0000 1111 1010 1100 1110 0000 0000 0000
$e+bias=00011111=31\rightarrow e=31-127=-96$ $Significand=1.0101100111$
0FACE000 = $1.0101100111\times 2^{-96}$

$X=-1.0001001101\times 2^{-105}\times 1.0101100111\times 2^{-96}=-1.01110011101111111011\times 2^{-201}=-0\times 2^{-126}$
$e+bias=-201+127=-74<0$

Here, there is underflow (not even denormalized numbers different than zero can represent it). Then $X\leftarrow -0$.
X = 1000 0000 0000 0000 0000 0000 0000 0000 = 80000000

✓ X = 800C0000 ÷ 494A0000:
800C0000: 1000 0000 0000 1100 0000 0000 0000 0000
$e+bias=00000000=0\rightarrow Denormal\ number\rightarrow e=-126$ $Significand=0.00011$
800C0000 = -0.00011×2^{-126}

494A0000: 0100 1001 0100 1010 0000 0000 0000 0000
$e+bias=10010010=146\rightarrow e=146-127=19$ $Significand=1.100101$
494A0000 = 1.100101×2^{19}

$$X=-\frac{0.00011\times 2^{-126}}{1.100101\times 2^{19}}$$

```
          00000001111
1100101 ) 11000000000
           1100101
           --------
           10110110
            1100101
            --------
            10100010
             1100101
             -------
              1111010
              1100101
              -------
                10101
```

Alignment:
$$\frac{0.000110}{1.100101}=\frac{0000110}{1100101}$$

Append x = 8 zeros: $\frac{11000000000}{1100101}$

Integer division
$Q=1111, R=111111\rightarrow Qf=0.00001111$

Thus: $X=-\frac{0.00011\times 2^{-126}}{1.100101\times 2^{19}}=-0.00001111\times 2^{-145}=-(0.00001111\times 2^{-19})\times 2^{-126}$
$X=-0.000\ 0000\ 0000\ 0000\ 0000\ 0000\ 1111\times 2^{-126}$. $Denormal\rightarrow e+bias=00000000$
X = 1000 0000 0000 0000 0000 0000 0000 0000 = 80000000

✓ X = 49744000 ÷ C0C90000:

49744000: 0100 1001 0111 0100 0100 0000 0000 0000

$e + bias = 10010010 = 146 \rightarrow e = 146 - 127 = 19$ $\quad Significand = 1.111010001$

497440000 = $1.111010001 \times 2^{19}$

C0C90000: 1100 0000 1100 1001 0000 0000 0000 0000

$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $\quad Significand = 1.1001001$

C0C90000 = -1.1001001×2^{2}

$$X = -\frac{1.111010001 \times 2^{19}}{1.1001001 \times 2^{2}}$$

```
              0000000000100110111
11001001000 ) 1111010001000000000
              11001001000
              -----------
                101011010000
                 11001001000
                 -----------
                 100100010000
                  11001001000
                  -----------
                   101100100000
                    11001001000
                    -----------
                    100110110000
                     11001001000
                     -----------
                      11011010000
                      11001001000
                      -----------
                         10001000
```

Alignment:

$$\frac{1.111010001}{1.1001001} = \frac{1.1110100010}{1.1001001000} = \frac{11110100010}{11001001000}$$

Append x = 8 zeros: $\frac{11110100010\mathbf{00000000}}{11001001000}$

Integer division

$Q = 100110111, R = 10001000 \rightarrow Qf = 1.00110111$

Thus: $X = -\frac{1.1110100001 \times 2^{19}}{1.1001001 \times 2^{2}} = -1.00110111 \times 2^{17}$

$e + bias = 17 + 127 = 144 = 10010000$

X = 1100 1000 0001 1011 1000 0000 0000 0000 = C81B8000