Solutions - Homework 2

(Due date: January 31st @ 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	
0.11010 + 1.0101101	1.00111 − 0.0000111	1.0001 + 1.001001	0.0101 − 1.0101101
10.10101 + 1.1001	100.1 + 0.10101	1000.0101 − 11.010101	101.0101 + 1.0111101

UNSIGNED:

```
c8=1 c7=1 c6=1 c5=0 c4=1 c3=0 c2=0 c1=0 c0=0
      0.1 1 0 1 0 0 0 +
      1.0 1 0 1 1 0 1
    1 0.0 0 1 0 1 0 1
```

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

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

```
c7=0 c6=0 c5=1 c4=0 c3=0 c2=0 c1=0 c0=0
  1 0 0.1 0 0 0 0 +
      0.1 0 1 0 1
  1 0 1.0 0 1 0 1
```

SIGNED:

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

```
0.0 1 0 1 0 0 0 −
1.0 1 0 1 1 0 1

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

```
1 0 0 0.0 1 0 1 0 0 −
1 1 1 1.0 1 0 1 0 1

c10=0 c9=0 c8=0 c7=0 c6=0 c5=0 c4=0 c3=0 c2=0 c1=0 c0=0
1 0 0 0.0 1 0 1 0 0 +
0 0 0 0.1 0 1 0 1 1
1 0 0 0.1 1 1 1 1 1
```

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

PROBLEM 2 (15 PTS)

- Multiply the following signed fixed-point numbers (6 pts):

01.001 × 1.001001	10.0001 × 01.01001	1.11010 × 110.11011

```
 01.001 x
1.001001

 01.001 x
0.110111

      1 1 0 1 1 1 x
            1 0 0 1
      1 1 0 1 1 1
    0 0 0 0 0 0
  0 0 0 0 0 0
1 1 0 1 1 1
1 1 1 0 1 1 1 1

0.1 1 1 1 0 1 1 1 1

1.0 0 0 0 1 0 0 0 1
```

```
 10.0001 x
01.01001

 01.1111 x
01.01001

          1 0 1 0 0 1 x
            1 1 1 1 1
          1 0 1 0 0 1
        1 0 1 0 0 1
      1 0 1 0 0 1
    1 0 1 0 0 1
  1 0 1 0 0 1
1 0 0 1 1 1 1 0 1 1 1

0 1 0.0 1 1 1 1 0 1 1 1

1 0 1.1 0 0 0 0 1 0 0 1
```

```
  1.1101 x
10.11011

  0.0011 x
01.00101

  1 0 0 1 0 1 x
            1 1
  1 0 0 1 0 1
1 0 0 1 0 1
1 1 0 1 1 1 1

0.0 0 1 1 0 1 1 1 1
```

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

101.1001 ÷ 1.011	11.011 ÷ 1.01011	10.0110 ÷ 01.01

✓ $\frac{101.1001}{1.011}$: To unsigned and then alignment, $a = 4$: $\frac{101.1001}{1.011} = \frac{010.0111}{0.1010} \equiv \frac{100111}{1010}$

```
           0000111110
1010  )    1001110000
            1010
            10011
             1010
            -----
             10010
              1010
             -----
              10000
               1010
              -----
                1100
                1010
                ----
                 100
```

Append $x = 4$ zeros: $\frac{100111\mathbf{0000}}{1010}$
Integer Division:

$Q = 111110, R = 100$
$\rightarrow Qf = 11.1110 \ (x = 4)$

Final result (2C): $\frac{101.1001}{1.011} = 011.111$

✓ $\frac{11.011}{1.01011}$: To unsigned and then alignment, $a = 5$: $\frac{00.101}{0.10101} = \frac{0.10100}{0.10101} \equiv \frac{10100}{10101}$

```
            000001111
10101  )    101000000
             10101
             100110
              10101
             ------
              100010
               10101
              ------
                11010
                10101
                -----
                  101
```

Append $x = 4$ zeros: $\frac{10100\mathbf{0000}}{10101}$
Integer Division:

$Q = 1111, R = 101$
$\rightarrow Qf = 0.1111(x = 4)$

Final result (2C): $\frac{11.011}{1.01011} = 0.1111$

✓ $\frac{10.0110}{01.01}$: To unsigned (numerator) and then alignment, $a = 4$: $\frac{01.1010}{01.01} = \frac{01.1010}{01.0100} \equiv \frac{11010}{10100}$

```
           000010100
10100  )   110100000
           10100
           ------
            11000
            10100
            -----
             10000
```

Append $x = 4$ zeros: $\frac{11010\mathbf{0000}}{10100}$
Integer Division:

$Q = 10100, R = 10000$
$\rightarrow Qf = 1.0100(x = 4)$

Final result (2C): $\frac{10.0110}{01.01} = 2C(01.01) = 10.11$

PROBLEM 3 (11 PTS)

- We want to represent numbers between -128.87 and 127.12. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0015? (4 pts).

  Resolution: $2^{-p} \leq 0.0015 \rightarrow p \geq 9.38 \rightarrow p = 10$
  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  $\Rightarrow -2^{n-p-1} \leq -128.87 \rightarrow n - p - 1 \geq 7.009 \rightarrow n - p \geq 8.009 \rightarrow n - p = 9 \rightarrow n = 19$
  $\therefore$ The required Fixed-Point format is $[19\ 10]$.

- We want to represent numbers between -255.12 and 256.91. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0025? (4 pts).

  Resolution: $2^{-p} \leq 0.0025 \rightarrow p \geq 8.643 \rightarrow p = 9$
  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  $\Rightarrow 2^{n-p-1} - 2^{-p} \geq 256.91 \longrightarrow 2^{n-p-1} \geq 256.91 + 2^{-9} \rightarrow n - p - 1 \geq 8.0051 \rightarrow n - p = 10 \longrightarrow n = 19$
  $\therefore$ The required Fixed-Point format is $[19\ 9]$.

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

-128.1875	-78.125	107.3125

✓ 128.1875 = 010000000.0011 → -128.1875 = 101111111.1101
✓ 78.125 = 01001110.001 → -78.1255 = 10110001.111
✓ 107.3125 = 01101011.0101

PROBLEM 4 (10 PTS)

- Complete the table for the following fixed point formats (signed numbers): (4 pts)

Fractional bits	Integer Bits	FX Format	Range	Dynamic Range (dB)	Resolution
8	4	[12 8]	$[-2^3, 2^3 - 2^{-8}] = [-8, 7.99609375]$	66.2266 dB	0.00390625
10	6	[16 10]	$[-2^5, 2^5 - 2^{-10}] = [-32, 31.9990234375]$	90.3089 dB	0.0009765625
16	8	[24 16]	$[-2^7, 2^7 - 2^{-16}] = [-128, 127.9999847412109]$	138.4738 dB	0.000015258789063

- 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
8	7	1.1755×10^{-38}	3.3895×10^{38}	$[-2^7 + 2, 2^7 - 1] = [-126, 127]$	$[1, 1.9921875]$
10	13	2.9833×10^{-154}	1.3406×10^{154}	$[-2^9 + 2, 2^9 - 1] = [-510, 511]$	$[1, 1.9998779296875]$
12	35	2^{-2046}	1.9999×2^{2047}	$[-2^{11} + 2, 2^{11} - 1] = [-2046, 2047]$	$[1, 1.999969482421875]$

PROBLEM 5 (20 PTS)

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

Single (32 bits)		Double (64 bits)	
✓ E8000978	✓ 800BCCAA	✓ 7FFDECADEFEEBEE9	✓ 8009BEBEFACE8000
✓ 80DE0FEE	✓ 7FFCAFEA	✓ 49A5DEAF8FAD8000	✓ 70800FEDCAB09000

✓ E8000978: 1110 1000 0000 0000 0000 1001 0111 1000
$e + bias = 11010000 = 208 \rightarrow e = 208 - 127 = 81$
$Significand\ ([24\ 23]) = 1.00000000000100101111000 = 1.000288963317871$
$X = -1.000288963317871 \times 2^{81} = -2.41855 \times 10^{24}$

✓ 80DE0FEE: 1000 0000 1101 1110 0000 1111 1110 1110
$e + bias = 00000001 = 1 \rightarrow e = 1 - 127 = -126$
$Significand = 1.10111100000111111101110 = 1.734861135482788$
$X = -1.734861135482788 \times 2^{-126} = -2.0393193 \times 10^{-38}$

✓ 800BCCAA: 1000 0000 0000 1011 1100 1100 1010 1010
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$
$Significand = 0.00010111100110010101010 = 0.092183351516724$
$X = -0.092183351516724 \times 2^{-126} = -1.08361 \times 10^{-39}$

✓ 7FFCAFEA: 0111 1111 1111 1100 1010 1111 1110 1010
$e + bias = 11111111 = 255, f \neq 0$
$X = NaN$

✓ 7FFDECADEFEEBEE9: 0111 1111 1111 1101 1110 1100 1010 1101 1110 1111 1110 1110 1011 1110 1110 1001
$e + bias = 11111111111 = 2047, f \neq 0$
$X = NaN$

✓ 49A5DEAF8FAD8000: 0100 1001 1010 0101 1101 1110 1010 1111 1000 1111 1010 1101 1000 0000 0000 0000
$e + bias = 10010011010 = 1178 \rightarrow e = 1178 - 1023 = 155$
$Significand = 1.0101110111101010111110001111101011011000 = 1.366866646996641$
$X = 1.366866646996641 \times 2^{155} = 6.24274325 \times 10^{46}$

- ✓ `8009BEBEFACE8000`: 1000 0000 0000 1001 1011 1110 1011 1110 1111 1010 1100 1110 1000 0000 0000 0000
  $e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
  $Significand = 0.1001101111101011111011111010110011101000 = 0.609068851197662$
  $X = -0.609068851197662 \times 2^{-1022} = -1.35522317 \times 10^{-308}$

- ✓ `70800FEDCAB09000`: 0111 0000 1000 0000 0000 1111 1110 1101 1100 1010 1011 0000 1001 0000 0000 0000
  $e + bias = 11100001000 = 1800 \rightarrow e = 1800 - 1023 = 777$
  $Significand = 1.0000000011111110110111001010101100001001 = 1.003888885265951$
  $X = 1.003888885265951 \times 2^{777} = 7.97980496 \times 10^{233}$

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.

✓ `40C00000 + C2EA9000` ✓ `801A8000 + 92CE8000`	✓ `5A09D378 – 4C490FD8` ✓ `10DAD000 – 90FAD000`	✓ `7A09C000 × 8BEE0000` ✓ `FA19D800 × CD100000`	✓ `C9680000 ÷ 80700000` ✓ `7A390000 ÷ C8400000`

- ✓ X = `40C00000 + C2EA9000`:

  `40C00000`: 0100 0000 1100 0000 0000 0000 0000 0000
  $e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $\quad Significand = 1.1$
  `40C00000` = 1.1×2^2

  `C2EA9000`: 1100 0010 1110 1010 1001 0000 0000 0000
  $e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $\quad Significand = 1.11010101001$
  `C2EA9000` = $-1.11010101001 \times 2^6$

  $X = 1.1 \times 2^2 - 1.11010101001 \times 2^6 = \frac{1.1}{2^4} \times 2^6 - 1.11010101001 \times 2^6$
  $X = 0.00011 \times 2^6 - 1.11010101001 \times 2^6$
  To subtract these numbers, we first convert to 2C:
  $R = 0.00011 - 01.11010101001 = 0.00011 + 10.00101010111$ (2C addition)
  The result (in 2C) is: $R = 10.01000010111$, $-R = 01.10111101001$
  ** Note that you can also do unsigned subtraction.*
  For floating point, we need to convert to sign-and-magnitude:
  $\Rightarrow R(SM) = -1.10111101001$

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

  $X = -1.10000101 \times 2^6, e + bias = 6 + 127 = 133 = 10000101$
  X = 1100 0010 1101 1110 1001 0000 0000 0000 = `C2DE9000`

- ✓ X = `801A8000 + 92CE8000`:

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

  `92CE8000`: 1001 0010 1100 1110 1000 0000 0000 0000
  $e + bias = 00100101 = 37 \rightarrow e = 37 - 127 = -90$ $\quad Significand = 1.10011101$
  `92CE8000` = $-1.10011101 \times 2^{-90}$

  $X = -0.00110101 \times 2^{-126} - 1.10011101 \times 2^{-90} = -\frac{0.00110101}{2^{36}} \times 2^{-90} - 1.10011101 \times 2^{-90}$
  Representing the division by 2^{36} requires more than $p + 1 = 24$ bits. Thus, we round down the 1st operand to 0.
  $X = -1.10011101 \times 2^{-90}$
  X = 1001 0010 1100 1110 1000 0000 0000 0000= `92CE8000`

- ✓ X = `50A9D378 – 4C490FD8`:

  `50A9D378`: 0101 0000 1010 1001 1101 0011 0111 1000
  $e + bias = 10100001 = 161 \rightarrow e = 161 - 127 = 34$ $\quad Significand = 1.01010011101001101111$
  `50A9D378` = $1.01010011101001101111 \times 2^{34}$

  `4C490FD8`: 0100 1100 0100 1001 0000 1111 1101 1000
  $e + bias = 10011000 = 152 \rightarrow e = 152 - 127 = 25$ $\quad Significand = 1.10010010000111111011$
  `4C490FD8` = $1.10010010000111111011 \times 2^{25}$

$X = 1.01010011101001101111 \times 2^{34} - 1.10010010000111111011 \times 2^{25}$
$X = 1.01010011101001101111 \times 2^{34} - \dfrac{1.10010010000111111011}{2^9} \times 2^{34}$
$X = 1.01010011101001101111 \times 2^{34} - 0.00000000110010010000111111011 \times 2^{34}$ (unsigned subtraction)

```
0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 1 0 0 0 0 1 1 1 1|1 1 1 1 1 0 0 0
  1.1 0 1 1 0 0 1 1 1 0 1 0 0 1 1 0 1 1 1 1 0 0 0|0 0 0 0 0 0 0 0 -
  0.0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 1 1|1 1 1 0 1 1 0 0
-------------------------------------------------------------------
  1.0 1 0 1 0 0 1 0 1 1 0 1 1 1 0 1 1 1 1 0 0 0 0|0 0 0 1 0 1 0 0
```

$X = 1.0101001011011101111 0000 \times 2^{34}, e + bias = 34 + 127 = 161 = 10100001$
X = 0101 0000 1010 1001 0110 1110 1111 0000 = 50A96EF0

✓ X = 10DAD000 - 90FAD000:
10DAD000: 0001 0000 1101 1010 1101 0000 0000 0000
$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$ $\quad$ $Significand = 1.10110101101$
10DAD000 = $1.10110101101 \times 2^{-94}$

90FAD000: 1001 0000 1111 1010 1101 0000 0000 0000
$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$ $\quad$ $Significand = 1.11110101101$
90FAD000 = $-1.11110101101 \times 2^{-94}$

$X = 1.10110101101 \times 2^{-94} - (-1.11110101101 \times 2^{-94})$ (unsigned addition)
$X = 11.1010101101 \times 2^{-94} = 1.11010101101 \times 2^{-93}$
$X = 1.11010101101 \times 2^{-93}, e + bias = -93 + 127 = 34 = 00100010$
X = 0001 0001 0110 1010 1101 0000 0000 0000 = 116AD000

```
1   1 1 1 1 0 1 0 1 1 0 1 0
    1.1 0 1 1 0 1 0 1 1 0 1 +
    1.1 1 1 1 0 1 0 1 1 0 1
----------------------------
1   1.1 0 1 0 1 0 1 1 0 1 0
```

✓ X = 7A09C000 × 8BEE0000:
7A09C000: 0111 1010 0000 1001 1100 0000 0000 0000
$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $\quad$ $Significand = 1.000100111$
7A09C000 = $1.000100111 \times 2^{117}$

8BEE0000: 1000 1011 1110 1110 0000 0000 0000 0000
$e + bias = 00010111 = 23 \rightarrow e = 23 - 127 = -104$ $\quad$ $Significand = 1.110111$
8BEE0000 = $-1.110111 \times 2^{-104}$

$X = 1.000100111 \times 2^{117} \times -1.110111 \times 2^{-104} = -10.000000000100001 \times 2^{13}$
$X = -1.0000000000100001 \times 2^{14}, \quad e + bias = 14 + 127 = 141 = 10001101$
X = 1100 0110 1000 0000 0001 0000 1000 0000 = C6801080

✓ X = FA19D800 × CD100000:
FA19D800: 1111 1010 0001 1001 1101 1000 0000 0000
$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $\quad$ $Significand = 1.001100111011$
FF19D800 = $-1.001100111011 \times 2^{117}$

CD100000: 1100 1101 0001 0000 0000 0000 0000 0000
$e + bias = 10011010 = 154 \rightarrow e = 154 - 127 = 27$ $\quad$ $Significand = 1.001$
CD100000 = -1.001×2^{27}

$X = -1.001100111011 \times 2^{117} \times -1.001 \times 2^{27} = 1.010110100010011 \times 2^{144}, \quad e + bias = 144 + 127 = 271 > 254$
Here, there is an overflow. The value X is assigned to $+\infty$
X = 0111 1111 1000 0000 0000 0000 0000 0000 = 7F800000

✓ X = C9680000 ÷ 80700000:
C9680000: 1100 1001 0110 1000 0000 0000 0000 0000
$e + bias = 10010010 = 146 \rightarrow e = 146 - 127 = 19$ $\quad$ $Significand = 1.1101$
C9680000 = -1.1101×2^{19}

80700000: 1000 0000 0111 0000 0000 0000 0000 0000
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $\quad$ $Significand = 0.111$
80700000 = -0.111×2^{-126}

$$X = \frac{1.1101 \times 2^{19}}{0.111 \times 2^{-126}}$$

```
        0001000010010
1110 ) 1110100000000
       1110
         010000
           1110
            10000
             1110
               100
```

Alignment:

$$\frac{1.1101}{0.111} = \frac{1.1101}{0.1110} = \frac{11101}{01110}$$

Append $x = 8$ zeros: $\frac{1110\mathbf{100000000}}{1110}$

Integer division

$Q = 1000010010, R = 100 \rightarrow Qf = 10.00010010$

Thus: $X = \frac{1.1101\times 2^{19}}{0.111\times 2^{-126}} = 10.0001001 \times 2^{145} = 1.00001001 \times 2^{146}$

$e + bias = 146 + 127 = 273 > 254$

Here, there is an overflow. The value X is assigned to $+\infty$

X = 0111 1111 1000 0000 0000 0000 0000 0000 = 7F800000

✓ X = 7A390000 ÷ C8400000:

7A390000: 0111 1010 0011 1001 0000 0000 0000 0000

$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $Significand = 1.0111001$

7A390000 = 1.0111001×2^{117}

C8400000: 1100 1000 0100 0000 0000 0000 0000 0000

$e + bias = 10010000 = 144 \rightarrow e = 144 - 127 = 17$ $Significand = 1.1$

C8400000 = -1.1×2^{17}

$$X = -\frac{1.0111001 \times 2^{117}}{1.1 \times 2^{17}}$$

```
             0000000011110110
11000000 ) 1011100100000000
           11000000
            101100100
             11000000
              101001000
               11000000
                100010000
                 11000000
                  101000000
                   11000000
                    100000000
                     11000000
                      10000000
```

Alignment:

$$\frac{1.0111001}{1.1000000} = \frac{10111001}{11000000}$$

Append $x = 8$ zeros: $\frac{10111001\mathbf{00000000}}{11000000}$

Integer division

$Q = 11110110, R = 10000000 \rightarrow Qf = 0.11110110$

Thus: $X = -\frac{1.0111001\times 2^{117}}{1.1\times 2^{17}} = -0.1111011 \times 2^{100} = -1.111011 \times 2^{99}$, $e + bias = 99 + 127 = 226 = 11100010$

X = 1111 0001 0111 0110 0000 0000 0000 0000 = F1760000