Solutions - Homework 2

(Due date: February 16th @ 11:59 pm)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (15 PTS)

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

1000.001 × 01.100101	01.10001 × 10.1101	01.101 × 1.001011

```
1000.001 x
01.100101
    ↓
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
```

```
01.10001 x
10.1101
    ↓
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
```

```
01.101 x
1.001011
    ↓
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
```

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

1.010110 ÷ 010.1011	101.1001 ÷ 1.0101	11.011 ÷ 1.10111

✓ $\frac{1.010110}{010.1011}$: To unsigned (numerator), and then alignment, $a = 5$: $\frac{0.10101}{010.1011} = \frac{000.10101}{010.10110} \equiv \frac{10101}{1010110}$

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

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

Unsigned Integer Division:
$Q = 11, R = 10100100$ $\rightarrow Qf = 0.0011\ (x = 4)$ ✶ Qf: unsigned number

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

✓ $\frac{101.1001}{1.0101}$: 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: $\frac{100111\mathbf{0000}}{1011}$

Unsigned 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 alignment: $a = 5$: $\frac{00.101}{0.01001} = \frac{0.10100}{0.01001} \equiv \frac{10100}{1001}$

```
      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): $\frac{11.011}{1.10111} = 010.0011$

PROBLEM 2 (11 PTS)

- We want to represent numbers between -251.5 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? (4 pts).

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

  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  Here, we use the upper bound to determine n:
  $\Rightarrow 2^{n-p-1} - 2^{-p} \geq 256 \rightarrow 2^{n-p-1} \geq 256 + 2^{-10} \rightarrow n - p - 1 \geq 8.0000055 \rightarrow n - p = 10 \rightarrow n = 20$

  $\therefore$ Then the fixed-point format required in $[20\ 10]$.

- We want to represent numbers between -127.69 and 120.69. 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: $2^{-p} \leq 0.0025 \rightarrow p \geq 8.6438 \rightarrow p = 9$

  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  Here, we use the lower bound to determine n:
  $\Rightarrow -2^{n-p-1} \leq -127.69 \rightarrow 2^{n-p-1} \geq 127.69 \rightarrow n - p - 1 \geq 6.996501 \rightarrow n - p = 8 \rightarrow n = 17$

  $\therefore$ Then the fixed-point format required in $[17\ 9]$.

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

-128.1875	-147.3125	79.125

✓ 128.1875 = 010000000.0011 → -128.1875 = 101111111.1101
✓ 147.3125 = 010010011.0101 → -147.3125 = 101101100.1011
✓ 79.125 = 01001111.001

PROBLEM 3 (10 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 4 (20 PTS)

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

Single (32 bits)		Double (64 bits)	
✓ 7FCE4710	✓ 803ACBAC	✓ FEAAFC0FEE000000	✓ 000ABBAF25C00000
✓ BDE32856	✓ 7BEAD360	✓ 7A09D3784D039800	✓ FFFECE4710ABCDEF

✓ 7FCE4710: 0111 1111 1100 1110 0100 0111 0001 0000
$e + bias = 11111111 = 255, f \neq 0$
$X = NaN$

- ✓ `803ACBAC`: `1000 0000 0011 1010 1100 1011 1010 1100`
  $e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$
  $Significand([24\ 23]) = 0.01110101100101110101100 = 0.459340572$
  $X = -0.459340572 \times 2^{-126} = -5.3995224 \times 10^{-39}$

- ✓ `BDE32856`: `1011 1101 1110 0011 0010 1000 0101 0110`
  $e + bias = 01111011 = 123 \rightarrow e = 123 - 127 = -4$
  $Significand = 1.11000110010100001010110 = 1.774668455123901$
  $X = -1.774668455123901 \times 2^{-4} = -0.110916778445244$

- ✓ `7BEAD360`: `0111 1011 1110 1010 1101 0011 0110 0000`
  $e + bias = 11110111 = 247 \rightarrow e = 247 - 127 = 120$
  $Significand\ ([24\ 23]) = 1.11010101101001101100000 = 1.834575$
  $X = +1.834575 \times 2^{120} = +2.4385 \times 10^{36}$

- ✓ `FEAAFC0FEE000000`: `1111 1110 1010 1010 1111 1100 0000 1111 1110 1110 0000 0000 0000 0000 0000 0000`
  $e + bias = 11111101010 = 2026 \rightarrow e = 2026 - 1023 = 1003$
  $Significand = 1.1010111111000000111111110111 = 1.686538$
  $X = -1.686538 \times 2^{1003} = -1.44571 \times 10^{302}$

- ✓ `000ABBAF25C00000`: `0000 0000 0000 1010 1011 1011 1010 1111 0010 0101 1100 0000 0000 0000 0000 0000`
  $e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
  $Significand\ ([53\ 52]) = 0.1010101110111010111100100101111 = 0.6708213$
  $X = +0.6708213 \times 2^{-1022} = +1.492627 \times 10^{-308}$

- ✓ `7A09D3784D039800`: `0111 1010 0000 1001 1101 0011 0111 1000 0100 1101 0000 0011 1001 1000 0000 0000`
  $e + bias = 11110100000 = 1952 \rightarrow e = 1952 - 1023 = 929$
  $Significand\ ([53\ 52]) = 1.100111010011011110000100110100000001110011 = 1.61412$
  $X = +1.6142 \times 2^{929} = +7.3249 \times 10^{279}$

- ✓ `FFFECE4710ABCDEF`: `1111 1111 1111 1110 1100 1110 0110 0111 0001 0000 1010 1011 1100 1101 1110 1111`
  $e + bias = 11111111111 = 2047, f \neq 0$
  $X = NaN$

PROBLEM 5 (44 PTS)

- Perform the following 32-bit floating point operations. For fixed-point division, use 8 fractional bits. Truncate the result when required. Show your work: how you got the significand and the biased exponent bits of the result. Provide the 32-bit result.

✓ `801A8000 + 33CEC000` ✓ `ECE4710A + FF800000`	✓ `40D90000 - 42EAC000` ✓ `CF4A8000 - 30A90000`	✓ `0E2CE000 × 8B092000` ✓ `AD0BEBED × 7F800000`	✓ `C9744000 ÷ 40C90000` ✓ `000C0000 ÷ C94A0000`

- ✓ X = `801A8000 + 33CEC000`:
  `801A8000`: `1000 0000 0001 1010 1000 0000 0000 0000`
  $e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126 \quad Significand = 0.00110101$
  `801A8000` = $-0.00110101 \times 2^{-126}$

  `33CEC000`: `0011 0011 1100 1110 1100 0000 0000 0000`
  $e + bias = 01100111 = 103 \rightarrow e = 103 - 127 = -24 \qquad Significand = 1.100111011$
  `33CEC000` = $+1.100111011 \times 2^{-24}$

  $X = -0.00110101 \times 2^{-126} + 1.100111011 \times 2^{-24} = -\frac{0.00110101}{2^{102}} \times 2^{-24} + 1.10011101 \times 2^{-24}$
  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 + ECE4710A`:
  `FF800000`: `1111 1111 1000 0000 0000 0000 0000 0000`
  $e + bias = 11111111 = 255, f = 0$
  `FF800000` = $-\infty$
  $X = (-\infty) + \# = -\infty$
  X = `FF800000`

✓ X = 40D90000 - 42EAC000:

40D90000: 0100 0000 1101 1001 0000 0000 0000 0000

$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $\quad$ $Significand = 1.1011001$

40C00000 = 1.101101×2^2

42EAC000: 0100 0010 1110 1010 1100 0000 0000 0000

$e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $\quad$ $Significand = 1.110101011$

42EA9000 = 1.110101011×2^6

$X = 1.1011001 \times 2^2 - 1.110101011 \times 2^6 = \dfrac{1.1011001}{2^4} \times 2^6 - 1.110101011 \times 2^6$

$X = 0.00011011001 \times 2^6 - 1.110101011 \times 2^6$

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 - 30A90000:

CF4A8000: 1100 1111 0100 1010 1000 0000 0000 0000

$e + bias = 10011110 = 158 \rightarrow e = 158 - 127 = 31$ $\quad$ $Significand = 1.10010101$

CF4A8000 = $-1.10010101 \times 2^{31}$

30A90000: 0011 0000 1010 1001 0000 0000 0000 0000

$e + bias = 01100001 = 97 \rightarrow e = 97 - 127 = -30$ $\quad$ $Significand = 1.0101001$

30A90000 = 1.0101001×2^{-30}

$X = -1.10010101 \times 2^{31} - 1.0101001 \times 2^{-30} = -1.10010101 \times 2^{31} - \dfrac{1.0101001}{2^{61}} \times 2^{31}$

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 = AD0BEBED × 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 × 0E2CE000:

8B092000: 1000 1011 0000 1001 1010 0000 0000 0000

$e + bias = 00010110 = 22 \rightarrow e = 22 - 127 = -105$ $\quad$ $Significand = 1.0001001101$

8B092000 = $-1.0001001101 \times 2^{-105}$

0E2CE000: 0000 1110 0010 1100 1110 0000 0000 0000

$e + bias = 00011100 = 29 \rightarrow e = 29 - 127 = -98$ $\quad$ $Significand = 1.0101100111$

0E2CE000 = $1.0101100111 \times 2^{-98}$

$X = -1.0001001101 \times 2^{-105} \times 1.0101100111 \times 2^{-98} = -1.0111001110111111011 \times 2^{-203} = -0 \times 2^{-126}$

$e + bias = -203 + 127 = -76 < 0$

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 = 000C0000 ÷ C94A0000:

000C0000: 0000 0000 0000 1100 0000 0000 0000 0000

$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126 \quad Significand = 0.00011$

000C0000 = 0.00011×2^{-126}

C94A0000: 1100 1001 0100 1010 0000 0000 0000 0000

$e + bias = 10010010 = 146 \rightarrow e = 146 - 127 = 19 \qquad Significand = 1.100101$

C94A0000 = -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: $\frac{11\mathbf{00000000}}{1100101}$

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 = C9744000 ÷ 40C90000:

C9744000: 1100 1001 0111 0100 0100 0000 0000 0000

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

C97440000 = $-1.111010001 \times 2^{19}$

40C90000: 0100 0000 1100 1001 0000 0000 0000 0000

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

40C90000 = 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