Solutions - Homework 2

(Due date: February 13th @ 11:59 pm)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (20 PTS)

a) Compute the result of the additions and subtractions for the following fixed-point numbers (5 pts):

UNSIGNED		SIGNED		
1101.1 + 0.1100101	1.01011 − 0.0011011	1010.01 − 101.01001	10.1100 + 01.01100101	0.0011011 − 1.01011

UNSIGNED:

$c_{11}=0\ c_{10}=0\ c_9=0\ c_8=1\ c_7=1\ c_6=0\ c_5=0\ c_4=0\ c_3=0\ c_2=0\ c_1=0\ c_0=0$

```
  1 1 0 1.1 0 0 0 0 0 0 +
        0.1 1 0 0 1 0 1
  ---------------------
  1 1 1 0.0 1 0 0 1 0 1
```

$b_8=0\ b_7=0\ b_6=0\ b_5=1\ b_4=0\ b_3=0\ b_2=1\ b_1=1\ b_0=0$

```
  1.0 1 0 1 1 0 0 -
  0.0 0 1 1 0 1 1
  ---------------
  1.0 0 1 0 0 0 1
```

SIGNED:

```
1 0 1 0.0 1 0 0 0 -
1 1 0 1.0 1 0 0 1
```

$c_9=0\ c_8=0\ c_7=1\ c_6=0\ c_5=0\ c_4=0\ c_3=0\ c_2=0\ c_1=0\ c_0=0$

```
  1 0 1 0.0 1 0 0 0 +
  0 0 1 0.1 0 1 1 1
  -----------------
  1 1 0 0.1 1 1 1 1
```

$c_{11}=1\ c_{10}=1\ c_9=1\ c_8=1\ c_7=1\ c_6=0\ c_5=0\ c_4=0\ c_3=0\ c_2=0\ c_1=0\ c_0=0$

```
  1 1 0.1 1 0 0 0 0 0 0 +
  0 0 1.0 1 1 0 0 1 0 1
  ---------------------
  0 0 0.0 0 1 0 0 1 0 1
```

```
0.0 0 1 1 0 1 1 -
1.0 1 0 1 1 0 0
```

$c_8=0\ c_7=0\ c_6=0\ c_5=1\ c_4=0\ c_3=0\ c_2=0\ c_1=0\ c_0=0$

```
  0.0 0 1 1 0 1 1 +
  0.1 0 1 0 1 0 0
  ---------------
  0.1 1 0 1 1 1 1
```

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

01.1001 × 1.001011	10.0101 × 11.011	1010.001 × 01.100101

```
01.1001 x
1.001011

01.1001 x         1 1 0 1 0 1 x
0.110101              1 1 0 0 1
                -------------
                  1 1 0 1 0 1
            1 1 0 1 0 1
        1 1 0 1 0 1
        -------------------
        1 0 1 0 0 1 0 1 1 0 1

0 1.0 1 0 0 1 0 1 1 0 1

1 0.1 0 1 1 0 1 0 0 1 1
```

```
10.0101 x
11.011

01.1011 x
00.101
              1 1 0 1 1 x
                  1 0 1
          -----------
              1 1 0 1 1
          1 1 0 1 1
          -------------
        1 0 0 0 0 1 1 1

0 1.0 0 0 0 1 1 1
```

```
1010.001 x
01.100101

0101.111 x              1 0 1 1 1 1 x
01.100101             1 1 0 0 1 0 1
                      -------------
                        1 0 1 1 1 1
                    1 0 1 1 1 1
              1 0 1 1 1 1
          1 0 1 1 1 1
          -----------------------
        1 0 0 1 0 1 0 0 0 1 0 1 1

0 1 0 0 1.0 1 0 0 0 1 0 1 1

1 0 1 1 0.1 0 1 1 1 0 1 0 1
```

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

100.1001 ÷ 1.0101	1.011 ÷ 0.01001	1.01011 ÷ 010.1011

✓ $\frac{100.1001}{1.0101}$: To unsigned and then alignment, $a = 4$: $\frac{011.0111}{0.1011} = \frac{011.0111}{0.1011} \equiv \frac{110111}{1011}$

```
          0001010000
1011 ) 1101110000
          1011
            1011
            1011
             00000
```

Append $x = 4$ zeros: $\frac{110111\mathbf{0000}}{1011}$

Unsigned Integer Division:
$Q = 1010000, R = 0000 \quad \rightarrow Qf = 101.1000 (x = 4)$

Final result (2C): $\frac{100.1001}{1.0101} = 0101.1$

✓ $\frac{1.011}{0.01001}$: To unsigned (denominator) 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:
$Q = 100011, R = 101 \quad \rightarrow Qf = 10.0011\ (x = 4)$ ✶ Qf: unsigned number

Final result (2C): $\frac{0.101}{1.10111} = 2C(010.0011) = 101.1101$

✓ $\frac{1.01011}{010.1011}$: To unsigned (denominator), 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:
$Q = 11, R = 10100100 \quad \rightarrow Qf = 0.0011\ (x = 4)$ ✶ Qf: unsigned number

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

PROBLEM 2 (9 PTS)

a) We want to represent numbers between -256 and 255.97. What is the fixed-point format that requires the fewest number of bits for a resolution better or equal than 0.001? (3 pts).

Resolution: $2^{-p} \leq 0.001 \rightarrow p \geq 9.9657 \rightarrow p = 10$

Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
Here, we use the lower bound to determine n:
$\Rightarrow -2^{n-p-1} \leq -256 \rightarrow 2^{n-p-1} \geq 256 \rightarrow n - p - 1 \geq 8 \rightarrow n - p = 9 \rightarrow n = 19$

∴ Then the fixed-point format required in $[19\ 10]$.

b) We want to represent numbers between -63.42 and 65.69. What is the fixed-point format that requires the fewest number of bits for a resolution better or equal than 0.0015? (3 pts).

Resolution: $2^{-p} \leq 0.0015 \rightarrow p \geq 9.38 \rightarrow p = 10$

Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
Here, we use the upper bound to determine n:
$\Rightarrow 2^{n-p-1} - 2^{-p} \geq 65.69 \rightarrow 2^{n-p-1} \geq 65.69 + 2^{-10} \rightarrow n - p - 1 \geq 6.037623 \rightarrow n - p = 8 \rightarrow n = 18$

∴ Then the fixed-point format required in $[18\ 10]$.

c) Represent these numbers in Fixed Point Arithmetic (signed numbers). Use the FX format [16 4]. Truncate (the LSB) and perform Saturation when required.

2048.25	-117.53125	-129.375

✓ 2048.25 = 0100000000000.01 = 011111111111.11 (saturation) = 011111111111.1100
✓ 117.53125 = 01110101.10001 → -117.3125 = 10001010.01111 = 111110001010.0111 (truncation)
✓ 129.375 = 010000001.011 → -129.375 = 101111110.101 = 111101111110.1010

PROBLEM 3 (8 PTS)

a) Complete the table for the following fixed-point formats (signed numbers): (3 pts)

Fractional bits	Integer Bits	FX Format	Range	Dynamic Range (dB)	Resolution
11	5	[16 11]	$[-2^4, 2^4 - 2^{-11}] = [-16, 15.99951171]$	90.3089 dB	0.00048828125
15	9	[24 15]	$[-2^8, 2^8 - 2^{-15}] = [-256, 255.999969]$	138.4738 dB	0.000030517578125

b) 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)	Significand bits (p)	Significand FX format	Min	Max	Range of e	Range of significand
8	6	[7 6]	1.1754×10^{-38}	3.37623×10^{38}	$[-2^7 + 2, 2^7 - 1] = [-126, 127]$	[1,1.984375]
10	13	[14 13]	2.9833×10^{-154}	1.34069×10^{154}	$[-2^9 + 2, 2^9 - 1] = [-510, 511]$	[1,1.9998779296875]
12	32	[33 32]	2^{-2046}	$(2 - 2^{-32})2^{2047}$	$[-2^{11} + 2, 2^{11} - 1] = [-2046, 2047]$	[1,1.9999999997671]

PROBLEM 4 (19 PTS)

a) For the given IEEE-754 floating-point numbers (displayed as hexadecimals), complete: bits in the fields (sign, biased exponent, significand) and significand's FX format (4 pts)

✓ `90DECADE` **(single – 32 bits)**

sign: 1
e+bias: 001 0000 1
FX format of significand: [24 23]
significand: 1.101 1110 1100 1010 1101 1110

✓ `3DECAFC0FFEE5000` **(double – 64 bits)**

sign: 0
e+bias: 011 1101 1110
FX format of significand: [53 52]
significand: 1.1100 1010 1111 1100 0000 1111 1111 1110 1110 0101 0000 0000 0000

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

Single (32 bits)		Double (64 bits)	
✓ 803AD0BE ✓ BEA7BEEF	✓ 7FCE4710	✓ 7FFCABFEEBA5ED00 ✓ 000C0FFEEFAD0000	✓ ECE4710A96B80C60

✓ 803AD0BE: 1000 0000 0011 1010 1101 0000 1011 1110
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$
$Significand([24\ 23]) = 0.01110101101000010111110 = 0.4594953$
$X = -0.4594953 \times 2^{-126} = -5.40134 \times 10^{-39}$

✓ BEA7BEEF: 1011 1110 1010 0111 1011 1110 1110 1111
$e + bias = 01111101 = 125 \rightarrow e = 125 - 127 = -2$
$Significand = 1.01001111011111011101111 = 1.31051433$
$X = -1.31051433 \times 2^{-2} = -0.32762858$

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

- ✓ `7FFCABFEEBA5ED00`: `0111 1111 1111 1100 1010 1011 1111 1110 1110 1011 1010 0101 1110 1101 0000 0000`
  $e + bias = 11111111111 = 2047, f \neq 0$
  $X = NaN$

- ✓ `000C0FFEEFAD0000`: `0000 0000 0000 1100 0000 1111 1111 1110 1110 1111 1010 1101 0000 0000 0000 0000`
  $e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
  $Significand\ ([53\ 52]) = 0.1100000011111111111011101111110101101 = 0.75390523$
  $X = +0.75390523 \times 2^{-1022} = +1.6774948 \times 10^{-308}$

- ✓ `ECE4710A96B80C60`: `1110 1100 1110 0100 0111 0001 0000 1010 1001 0110 1011 1000 0000 1100 0110 0000`
  $e + bias = 11011001110 = 1742 \rightarrow e = 1742 - 1023 = 719$
  $Significand = 1.0100011100010000101010010110101110000000110001 1 = 1.277597988839965$
  $X = -1.277597988839965 \times 2^{719} = -3.52339 \times 10^{216}$

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: <u>how you got the significand and the biased exponent bits of the result</u>. Provide the 32-bit result.

✓ `7F800000 + C512290A` ✓ `C0D90000 + 42EAC000`	✓ `B3BEE000 - 8037C000` ✓ `80123000 - 004E8000`	✓ `5A09C000 × CD080000` ✓ `7CDA0000 × 80200000`	✓ `C9744000 ÷ 81C90000` ✓ `000C0000 ÷ BACA0000`

- ✓ $X =$ `7F800000 + C512290A`:
  `7F800000`: `0111 1111 1000 0000 0000 0000 0000 0000`
  $e + bias = 11111111 = 255, f = 0$
  `7F800000` $= +\infty$

  $X = (+\infty) + \# = +\infty$
  $X =$ `7F800000`

- ✓ $X =$ `C0D90000 + 42EAC000`:
  `C0D90000`: `1100 0000 1101 1001 0000 0000 0000 0000`
  $e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $Significand = 1.1011001$
  `C0D90000` $= -1.1011001 \times 2^2$

  `42EAC000`: `0100 0010 1110 1010 1100 0000 0000 0000`
  $e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $Significand = 1.110101011$
  `42EAC000` $= 1.110101011 \times 2^6$

  $X = -1.1011001 \times 2^2 + 1.110101011 \times 2^6 = -\frac{1.1011001}{2^4} \times 2^6 + 1.110101011 \times 2^6$
  $X = -0.00011011001 \times 2^6 + 1.110101011 \times 2^6$

  $X = (1.110101011 - 0.00011011001) \times 2^6$ (unsigned subtraction)

  ```
    0 0 0 1 1 1 0 1 0 0 1 1 0
     1.1 1 0 1 0 1 0 1 1 0 0  -
     0.0 0 0 1 1 0 1 1 0 0 1
    ------------------------
     1.1 0 1 1 1 0 1 0 0 1 1
  ```

  $X = 1.10111010011 \times 2^6$, $e + bias = 6 + 127 = 133 = 10000101$
  $X =$ `0100 0010 1101 1101 0011 0000 0000 0000 = 42DD3000`

- ✓ $X =$ `B3BEE000 - 8037C000`:
  `B3BEE000`: `1011 0011 1011 1110 1110 0000 0000 0000`
  $e + bias = 01100111 = 103 \rightarrow e = 103 - 127 = -24$ $Significand = 1.0111110111$
  `B3BEE000` $= -1.0111110111 \times 2^{-24}$

  `8037C000`: `1000 0000 0011 0111 1100 0000 0000 0000`
  $e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $Significand = 0.011011111$
  `8037C000` $= -0.011011111 \times 2^{-126}$

  $X = -1.100111011 \times 2^{-24} + 0.011011111 \times 2^{-126} = -1.0111110111 \times 2^{-24} + \frac{0.011011111}{2^{102}} \times 2^{-24}$
  Representing the number divided by 2^{102} requires more than $p + 1 = 24$ bits. Thus, we round down this operand to 0.
  $X = -1.100111011 \times 2^{-24}$
  $X =$ `1011 0011 1011 1110 1110 0000 0000 0000 = B3BEE000`

✓ X = 80123000 - 004E8000:

80123000: 1000 0000 0001 0010 0011 0000 0000 0000

$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $Significand = 0.00100100011$

80123000 = $-0.00100100011 \times 2^{-126}$

004E8000: 0000 0000 0100 1110 1000 0000 0000 0000

$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $Significand = 0.10011101$

004E8000 = $0.10011101 \times 2^{-126}$

$X = -0.00100100011 \times 2^{-126} - 0.10011101 \times 2^{-126}$ (unsigned addition)

$X = -0.11000001011 \times 2^{-126}, e + bias = 0 = 00000000$

X = 1000 0000 0110 0000 1011 0000 0000 0000= 8060B000

```
  0 0 0 1 1 1 1 0 0 0 0 0 0
    0.0 0 1 0 0 1 0 0 0 1 1 +
    0.1 0 0 1 1 1 0 1 0 0 0
   -------------------------
    0.1 1 0 0 0 0 0 1 0 1 1
```

✓ X = 5A09C000 × CD080000:

5A09C000: 0101 1010 0000 1001 1100 0000 0000 0000

$e + bias = 10110100 = 180 \rightarrow e = 180 - 127 = 53$ $Significand = 1.000100111$

5A09C000 = $1.000100111 \times 2^{53}$

CD080000: 1100 1101 0000 1000 0000 0000 0000 0000

$e + bias = 10011010 = 154 \rightarrow e = 154 - 127 = 27$ $Significand = 1.0001$

CD080000 = -1.0001×2^{27}

```
    1 0 0 0 1 0 0 1 1 1 x
              1 0 0 0 1
    -------------------
    1 0 0 0 1 0 0 1 1 1
1 0 0 0 1 0 0 1 1 1
-----------------------
1 0 0 1 0 0 1 0 0 1 0 1 1 1
```

$X = (1.000100111 \times 2^{53}) \times (-1.0001 \times 2^{27}) = -1.0010010010111 \times 2^{80}$

$e + bias = 8 + 127 = 207 = 11001111$

X = 1110 0111 1001 0010 0101 1100 0000 0000 = E7925C00

✓ X = 7CDA0000 × 80200000:

7CDA0000: 0111 1100 1101 1010 0000 0000 0000 0000

$e + bias = 11111001 = 249 \rightarrow e = 249 - 127 = 122$ $Significand = 1.101101$

7CDA0000 = 1.101101×2^{122}

80200000: 1000 0000 0010 0000 0000 0000 0000 0000

$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $Significand = 0.01$

80200000 = -0.01×2^{-126}

$X = (-0.01 \times 2^{-126}) \times (1.101101 \times 2^{122}) = -0.01101101 \times 2^{-4} = -1.101101 \times 2^{-6}$

$e + bias = -6 + 127 = 121 = 01111001$

X = 1011 1100 1101 1010 0000 0000 0000 0000 = BCDA0000

✓ X = C9744000 ÷ 81C90000:

C9744000: 1100 1001 0111 0100 0100 0000 0000 0000

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

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

81C90000: 1000 0001 1100 1001 0000 0000 0000 0000

$e + bias = 0000011 = 3 \rightarrow e = 3 - 127 = -124$ $Significand = 1.1001001$

81C90000 = $-1.1001001 \times 2^{-124}$

$$X = \frac{1.111010001 \times 2^{19}}{1.1001001 \times 2^{-124}}$$

Alignment:

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

Append x = 8 zeros: $\frac{111101000100000000\mathbf{0}}{11001001000}$

Integer division

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

```
             0000000000100110111
11001001000 ) 1111010001100000000
              11001001000
              -----------
                101011011000
                 11001001000
                ------------
                 100100100000
                  11001001000
                 ------------
                   101101100000
                    11001001000
                   ------------
                     101000110000
                      11001001000
                     ------------
                        11111010000
                        11001001000
                        -----------
                          110001000
```

Thus: $X = \frac{1.1110100001 \times 2^{19}}{1.1001001 \times 2^{-124}} = 1.00110111 \times 2^{143}$
$e + bias = 143 + 127 = 270 > 254$

Here, there is an overflow. The value X is assigned to $+\infty$.
X = 0111 1111 1000 0000 0000 0000 0000 0000 = 7F800000

✓ X = 000C0000 ÷ BACA0000:
000C0000: 0000 0000 0000 1100 0000 0000 0000 0000
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$ $Significand = 0.00011$
000C0000 = 0.00011×2^{-126}

BACA0000: 1011 1010 1100 1010 0000 0000 0000 0000
$e + bias = 01110101 = 117 \rightarrow e = 117 - 127 = -10$ $Significand = 1.100101$
BACA0000 = -1.100101×2^{-10}

$$X = -\frac{0.00011 \times 2^{-126}}{1.100101 \times 2^{-10}}$$

```
          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: $X = -\frac{0.00011 \times 2^{-126}}{1.100101 \times 2^{-10}} = -0.00001111 \times 2^{-116} = -(0.00001111) \times 2^{-116} = -1.1111 \times 2^{-121}$
$e + bias = -121 + 127 = 6 = 00000110$
X = 1000 0011 0111 1000 0000 0000 0000 0000 = 83780000