Solutions - Homework 1

(Due date: September 20th)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (15 PTS)

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

UNSIGNED		SIGNED	
1.101010 + 0.0111101	1.10011 − 0.0100101	10.001 + 1.001101	0.1101 − 1.0100101
11.1101 + 1.0001	100.1 + 0.1101101	1000.0101 − 101.01011	111.0001 + 1.0111101

UNSIGNED:

$c_8=1, c_7=1, c_6=1, c_5=1, c_4=1, c_3=1, c_2=0, c_1=0, c_0=0$
```
  1.1 0 1 0 1 0 0 +
  0.0 1 1 1 1 0 1
-----------------
1 0.0 0 1 0 0 0 1
```

$b_7=0, b_6=1, b_5=0, b_4=0, b_3=1, b_2=1, b_1=1, b_0=0$
```
1.1 0 0 1 1 0 0 -
0.0 1 0 0 1 0 1
---------------
1.0 1 0 0 1 1 1
```

$c_6=1, c_5=1, c_4=0, c_3=0, c_2=0, c_1=1, c_0=0$
```
  1 1.1 1 0 1 +
    1.0 0 0 1
-------------
1 0 0.1 1 1 0
```

$c_{10}=0, c_9=0, c_8=0, c_7=1, c_6=0, c_5=0, c_4=0, c_3=0, c_2=0, c_1=0, c_0=0$
```
1 0 0.1 0 0 0 0 0 0 +
    0.1 1 0 1 1 0 1
---------------------
1 0 1.0 1 0 1 1 0 1
```

SIGNED:

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

```
0.1 1 0 1 0 0 0 -
1.0 1 0 0 1 0 1
```
➡

$c_9=0, c_8=0, c_7=1, c_6=1, c_5=1, c_4=1, c_3=0, c_2=0, c_1=0, c_0=0$
```
0 0.1 1 0 1 0 0 0 +
0 0.1 0 1 1 0 1 1
-----------------
0 1.1 0 0 0 0 1 1
```

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

$c_9=0, c_8=0, c_7=0, c_6=0, c_5=0, c_4=0, c_3=0, c_2=0, c_1=0, c_0=0$
```
1 0 0 0.0 1 0 1 0 +
0 0 1 0.1 0 1 0 1
-------------------
1 0 1 0.1 1 1 1 1
```

$c_{10}=1, c_9=1, c_8=1, c_7=0, c_6=1, c_5=1, c_4=1, c_3=0, c_2=0, c_1=0, c_0=0$
```
1 1 1.0 0 0 1 0 0 0 +
1 1 1.0 1 1 1 1 0 1
---------------------
1 1 0.1 0 0 0 1 0 1
```

PROBLEM 2 (25 PTS)

- Multiply the following signed fixed-point numbers.

01.101 × 1.001101	10.1001 × 01.00101	1000.000 × 10.010011	0.1111010 × 10.0011011

```
01.101   x  ➡  01.101   x  ➡  1 1 0 0 1 1 x
1.001101       0.110011             1 1 0 1
                              -------------
                                1 1 0 0 1 1
                            1 1 0 0 1 1
                          1 1 0 0 1 1
                        -------------------
                        1 0 1 0 0 1 0 1 1 1
                                ⬇
                        0 1.0 1 0 0 1 0 1 1 1
                                ⬇
                        1 0.1 0 1 1 0 1 0 0 1
```

```
10.1001 x  ➡  01.0111 x  ➡      1 0 0 1 0 1 x
01.00101      01.00101            1 0 1 1 1
                               -------------
                                1 0 0 1 0 1
                              1 0 0 1 0 1
                            1 0 0 1 0 1
                          1 0 0 1 0 1
                        -------------------
                          1 1 0 1 0 1 0 0 1 1
                                  ⬇
                          0 1.1 0 1 0 1 0 0 1 1
                                  ⬇
                          1 0.0 1 0 1 0 1 1 0 1
```

```
1000.000 x  ➡  01000.000 x  ➡  1 1 0 1 1 0 1 x
10.010011      01.101101       1 0 0 0 0 0 0
                               -------------
                               0 0 0 0 0 0 0
                   1 1 0 1 1 0 1
                   -------------------------
                   1 1 0 1 1 0 1 0 0 0 0 0 0
                               ⬇
                        0 1 1 0 1.1 0 1
```

```
0.111101 x   ➡  0.111101 x  ➡  1 1 1 0 0 1 0 1 x
10.0011011      01.1100101           1 1 1 1 0 1
                               ---------------
                               1 1 1 0 0 1 0 1
                           1 1 1 0 0 1 0 1
                         1 1 1 0 0 1 0 1
                       1 1 1 0 0 1 0 1
                     1 1 1 0 0 1 0 1
                   -------------------------------
                   1 1 0 1 1 0 1 0 0 1 0 0 0 1
                                 ⬇
                   0 1.1 0 1 1 0 1 0 0 1 0 0 0 1
                                 ⬇
                   1 0.0 1 0 0 1 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 ÷ 1.0001	11.0101 ÷ 1.0101	10.0100 ÷ 01.11	0.111010 ÷ 100.0111

✓ $\frac{101.1001}{1.0001}$: To unsigned and then alignment, $a = 4$: $\frac{101.1001}{1.0001} = \frac{010.0111}{0.1111} \equiv \frac{100111}{1111}$

```
          0000101001
1111 ) 1001110000
         1111
         010010
           1011
           0011000
              1111
              1001
```

Append $x = 4$ zeros: $\frac{100111\mathbf{0000}}{1111}$

Integer Division:

$Q = 101001, R = 1001$
$\rightarrow Qf = 10.1001\ (x = 4)$

Final result (2C): $\frac{101.1001}{1.0001} = 010.1001$

✓ $\frac{11.0101}{1.0101}$: To unsigned and then alignment, $a = 5$: $\frac{00.1011}{0.1011} = \frac{1011}{1011} = 1$

Integer Division:
$Q = 1, R = 0$
$\rightarrow Qf = 1$

Final result (2C): $\frac{11.0101}{1.0101} = 01$

✓ $\frac{10.01}{01.11}$: To unsigned (numerator) and then alignment, $a = 4$: $\frac{01.11}{01.11} = \frac{111}{111} = 1$

Integer Division:
$Q = 1, R = 0$
$\rightarrow Qf = 1$

Final result (2C): $\frac{10.01}{01.11} = 2C(01) = 1$

✓ $\frac{0.11101}{100.0111}$: To unsigned (denominator) and then alignment, $a = 5$: $\frac{0.11101}{011.10010} \equiv \frac{11101}{1110010}$

```
              000000100
1110010 ) 111010000
            1110010
              001000
```

Append $x = 4$ zeros: $\frac{11101\mathbf{0000}}{1110010}$

Integer Division:

$Q = 100, R = 1000$
$\rightarrow Qf = 0.0100 (x = 4)$

Final result (2C): $\frac{0.11101}{100.0111} = 2C(0.01) = 1.11$

PROBLEM 3 (20 PTS)

- We want to represent numbers between -511.97 and 256.25. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0025? (5 pts).

2C representation for integers: -2^{n-1} to $2^{n-1}-1$. For $-2^{n-1} \leq -511$, 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 0.0025:

$$2^{-p} \leq 0.0025 \rightarrow p \geq 8.6439 \rightarrow p = 9$$

Then the Fixed Point format required in $[19\ 9]$.

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

-127.125	-232.1875	-68.625	255.3125

✓ 127.125 = 01111111.001 → -127.125 = 10000000.111
✓ 232.1875 = 011101000.011 → -232.1875 = 100010111.1101
✓ 68.625 = 01000100.101 → -68.625 = 10111011.011
✓ 255.3125 = 011111111.0101

PROBLEM 4 (10 PTS)

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

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 the following floating point formats (which resemble the IEEE-754 standard) with 16, 24, 48 bits. 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 (30 PTS)

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

Single (32 bits)		Double (64 bits)	
✓ 50DAD800	✓ 800FACEA	✓ FA09D3784D039B70	✓ 800CABADE049AB80
✓ 80BEEF80	✓ 7FC0FEE0	✓ 80BEEFACE9700400	✓ FE800CD009AB00D8
✓ 3DE32860	✓ FACEB00C	✓ 7FFDECADEFEEBEE9	✓ DFC0FC0FFEE10800

✓ 50DAD800: 0101 0000 1101 1010 1101 1000 0000 0000
$e + bias = 10100001 = 161 \rightarrow e = 161 - 127 = 34$
$Mantissa\ ([24\ 23]) = 1.101101011011 = 1.709716796875$
$X = 1.709716796875 \times 2^{34} = 2.93727 \times 10^{10}$

✓ 80BEEF80: 1000 0000 1011 1110 1110 1111 1000 0000
$e + bias = 00000001 = 1 \rightarrow e = 1 - 127 = -126$
$Mantissa = 1.0111110111011111 = 1.49168395$
$X = -1.49168395 \times 2^{-126} = -1.753466 \times 10^{-38}$

✓ 3DE32860: 0011 1101 1110 0011 0010 1000 0110 0000
$e + bias = 01111011 = 123 \rightarrow e = 123 - 127 = -4$
$Mantissa = 1.110001100101000011 = 1.77466964$
$X = 1.77466964 \times 2^{-4} = 0.11091685295$

✓ 800FACEA: 1000 0000 0000 1111 1010 1100 1110 1010
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$
$Mantissa = 0.00011111010110011101010 = 0.1224644184$
$X = -0.1224644184 \times 2^{-126} = -1.4395623 \times 10^{-39}$

✓ 7FC0FEE0: 0111 1111 1100 0000 1111 1110 1110 0000
$e + bias = 11111111 = 255, f \neq 0$
$X = NaN$

✓ FACEB00C: 1111 1010 1100 1110 1011 0000 0000 1100
$e + bias = 11110101 = 245 \rightarrow e = 245 - 127 = 118$
$Mantissa = 1.10011101011000000001100 = 1.614747524261475$
$X = -1.614747524261475 \times 2^{118} = -5.36592 \times 10^{35}$

✓ FA09D3784D039B70: 1111 1010 0000 1001 1101 0011 0111 1000 0100 1101 0000 0011 1001 1011 0111 0000
$e + bias = 11110100000 = 1952 \rightarrow e = 1952 - 1023 = 929$
$Mantissa\ ([53\ 52]) = 1.1001110100110111100001001101000000111001101101111 = 1.614128399692813$
$X = -1.614128399692813 \times 2^{929} = -7.324939 \times 10^{279}$

✓ 80BEEFACE9700400: 1000 0000 1011 1110 1110 1111 1010 1100 1110 1001 0111 0000 0000 0100 0000 0000
$e + bias = 00000001011 = 11 \rightarrow e = 11 - 1023 = -1012$
$Mantissa = 1.1110111011111010110011101001011100000000001 = 1.93351451098$
$X = -1.93351451098 \times 2^{-1012} = -4.405465 \times 10^{-305}$

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

✓ 800CABADE049AB80: 1000 0000 0000 1100 1010 1011 1010 1101 1110 0000 0100 1001 1010 1011 1000 0000
$e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
$Mantissa = 0.1100101010111010110111100000001001001101010111 = 0.79191386$
$X = -0.79191386 \times 2^{-1022} = -1.7620668 \times 10^{-308}$

✓ FE800CD009AB00D8: 1111 1110 1000 0000 0000 1100 1101 0000 0000 1001 1010 1011 0000 0000 1101 1000
$e + bias = 11111101000 = 2024 \rightarrow e = 2024 - 1023 = 1001$
$Mantissa = 1.000000001100110100000000100110101011000000000011011 = 1.00312808777$
$X = -1.00312808777 \times 2^{1001} = -2.1497207 \times 10^{301}$

✓ DFC0FC0FFEE10800: 1101 1111 1100 0000 1111 1100 0000 1111 1111 1110 1110 0001 0000 1000 0000 0000
$e + bias = 10111111100 = 1532 \rightarrow e = 1532 - 1023 = 509$
$Mantissa = 1.0000111111000000111111111110111000010000100000000000 = 1.061538692113118$
$X = -1.061538692113118 \times 2^{509} = -1.77911336 \times 10^{153}$