Solutions - Homework 1

(Due date: September 19th)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (14 PTS)

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

UNSIGNED (6 pts)		**SIGNED** (8 pts)	
1.10010 + 0.101001	1.10101 − 0.0100111	10.101 + 1.10101	0.0011 − 1.101101
10.1011 + 1.1101	100.1 + 0.0100101	1001.0111 − 111.01001	101.1101 + 1.1010011

UNSIGNED:

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

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

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

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

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

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

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

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

SIGNED:

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

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

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

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

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

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

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

```
 1 0 0 1.0 1 1 1 0 +
 0 0 0 0.1 0 1 1 1
 -----------------
 1 0 1 0.0 0 1 0 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 0 1.1 1 0 1 0 0 0 +
 1 1 1.1 0 1 0 0 1 1
 -------------------
 1 0 1.0 1 1 1 0 1 1
```

PROBLEM 2 (22 PTS)

- Multiply the following signed fixed-point numbers: (10 pts)

01.101 × 1.101001	100.001 × 01.10001	110.000 × 10.10101	0.11001 × 11.11011

```
 01.101   x  →   01.101   x  →      1 0 1 1 1 x
 1.101001        0.010111             1 1 0 1
                                  ---------
                                    1 0 1 1 1
                                1 0 1 1 1
                              1 0 1 1 1
                          -------------------
                          1 0 0 1 0 1 0 1 1
                                   ↓
                          0.1 0 0 1 0 1 0 1 1
                                   ↓
                          1.0 1 1 0 1 0 1 0 1
```

```
 100.001  x  →  011.111  x  →      1 1 1 1 1 x
 01.10001       01.10001         1 1 0 0 0 1
                                 -----------
                                   1 1 1 1 1
                           1 1 1 1 1
                         1 1 1 1 1
                         ---------------------
                         1 0 1 1 1 1 0 1 1 1 1
                                   ↓
                         0 1 0 1.1 1 1 0 1 1 1 1
                                   ↓
                         1 0 1 0.0 0 0 1 0 0 0 1
```

```
 110.000  x  →  010.000  x  →    1 0 1 0 1 1 x
 10.10101       01.01011           1 0 0 0 0
                                 -----------
                                 0 0 0 0 0 0 0
                           1 0 1 0 1 1
                           -------------------
                           1 0 1 0 1 1 0 0 0 0
                                   ↓
                         0 1 0.1 0 1 1 0 0 0 0
```

```
 0.11001  x  →  0.11001  x  →    1 1 0 0 1 x
 11.11011       00.00101               1 0 1
                                 ---------
                                 1 1 0 0 1
                               1 1 0 0 1
                               -------------
                               1 1 1 1 1 0 1
                                     ↓
                         0.0 0 0 1 1 1 1 1 0 1
                                     ↓
                         1.1 1 1 0 0 0 0 0 1 1
```

- Get the division result (with $x = 4$ fractional bits) for the following signed fixed-point numbers: (12 pts).

101.0101 ÷ 1.101	1.1011 ÷ 1.01101	10.0101 ÷ 01.11	0.10100 ÷ 110.101

✓ $\frac{101.0101}{1.101}$: To unsigned and then alignment, $a = 4$: $\frac{010.1011}{0.0110} \equiv \frac{101011}{110}$

```
         0001110010
110 ) 1010110000
       110
        1001
         110
          111
          110
           1000
            110
             100
```

Append $x = 4$ zeros: $\frac{101011\mathbf{0000}}{110}$

Integer Division:

$Q = 1110010, R = 100$
$\rightarrow Qf = 111.0010\ (x = 4)$

Final result (2C): $\frac{101.0101}{1.101} = 0111.001$

✓ $\frac{1.1011}{1.01101}$: To unsigned and then alignment, $a = 4$: $\frac{0.01010}{0.10011} = \frac{1010}{10111}$

```
          00001000
10011 ) 10100000
        10011
            1000
```

Append $x = 4$ zeros: $\frac{1010\mathbf{0000}}{10011}$

Integer Division:

$Q = 1000, R = 1000$
$\rightarrow Qf = 0.1000\ (x = 4)$

Final result (2C): $\frac{1.1011}{1.01101} = 0.1$

✓ $\frac{10.0101}{01.11}$: To unsigned (numerator) and then alignment, $a = 4$: $\frac{01.1011}{01.1100} = \frac{11011}{11100}$

```
          000001111
11100 ) 110110000
         11100
         110100
          11100
          110000
           11100
           101000
            11100
             1100
```

Append $x = 4$ zeros: $\frac{11110\mathbf{0000}}{11100}$

Integer Division:

$Q = 1111, R = 1100$
$\rightarrow Qf = 0.1111\ (X = 4)$

Final result (2C): $\frac{10.0101}{01.11} = 2C(0.1111) = 1.0001$

✓ $\frac{0.10100}{110.101}$: To unsigned (denominator) and then alignment, $a = 3$: $\frac{0.101}{001.011} \equiv \frac{101}{1011}$

```
         0000111
1011 ) 1010000
        1011
        10010
         1011
          1110
          1011
            11
```

Append $x = 4$ zeros: $\frac{101\mathbf{0000}}{1011}$

Integer Division:

$Q = 111, R = 11$
$\rightarrow Qf = 0.0111 (x = 4)$

Final result (2C): $\frac{0.10100}{110.101} = 2C(0.0111) = 1.1001$

PROBLEM 3 (16 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? (5 pts).

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

  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  $\Rightarrow -2^{n-p-1} \leq -128.87 \rightarrow n-p-1 \leq 7.009 \rightarrow n-p \geq 8.009 \rightarrow n-p = 9 \rightarrow n = 19$

  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? (5 pts).

  Resolution: $2^{-p} \leq 0.0025 \rightarrow p \geq 8.643 \rightarrow p = 9$

  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  $\Rightarrow 2^{n-p-1} - 2^{-p} \leq 256.91 \longrightarrow 2^{n-p-1} \leq 256.91 + 2^{-9} \rightarrow n-p-1 \geq 8.0051 \rightarrow n-p = 10 \longrightarrow n = 19$

  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	-78.1255	-34.625	207.3125

  ✓ 128.1875 = 010000000.0011 → -128.1875 = 101111111.1101
  ✓ 78.1255 = 01001110.001 → -78.1255 = 10110001.111
  ✓ 34.625 = 0100010.101 → -34.625 = 1011101.011
  ✓ 207.3125 = 011001111.0101

PROBLEM 4 (12 PTS)

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

Fractional bits	Integer Bits	FX Format	Range	Dynamic Range (dB)	Resolution
9	3	$[12\ 9]$	$[-2^2, 2^2 - 2^{-9}] = [-4, 3.998046875]$	66.2266 dB	0.001953125
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

- Complete the table for the following floating point formats (which resemble the IEEE-754 standard) with 16, 24, 48 bits. Only consider ordinary numbers. (9 pts)

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

Exponent bits (E)	Significant bits (p)	Min	Max	Range of e	Range of significand
8	6	1.1754×10^{-38}	3.37623×10^{38}	$[-2^7 + 2, 2^7 - 1] = [-126, 127]$	$[1, 1.984375]$
10	13	2.9833×10^{-154}	1.34069×10^{154}	$[-2^9 + 2, 2^9 - 1] = [-510, 511]$	$[1, 1.9998779296875]$
15	32	2^{-16382}	$(2 - 2^{-32})2^{16383}$	$[-2^{14} + 2, 2^{14} - 1] = [-16382, 16383]$	$[1, 1.9999999997671]$

PROBLEM 5 (36 PTS)

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

Single (32 bits)		Double (64 bits)	
✓ 90DBD800	✓ 7F85B0AC	✓ DECAFC0FFEE80000	✓ ACCEDE90BEAD5000
✓ 800BEEF0	✓ 70DECADE	✓ 49A5DEAF8FAD8000	✓ 8009BEBEFACE8000
✓ 3CF398C0	✓ 90BEE950	✓ FFFACADE19801990	✓ 70800FEDCAB09000

- ✓ `90DBD800`: `1001 0000 1101 1011 1101 1000 0000 0000`
  $e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$
  $Mantissa\ ([24\ 23]) = 1.101101111011 = 1.717529296875$
  $X = -1.717529296875 \times 2^{-94} = -8.6713 \times 10^{-29}$

- ✓ `800BEEF0`: `1000 0000 0000 1011 1110 1110 1111 0000`
  $e + bias = 00000000 = 0 \rightarrow Denormal\ number\ \rightarrow e = -126$
  $Mantissa = 0.00010111110111011111 = 0.093229293$
  $X = -0.093229293 \times 2^{-126} = -1.09590508 \times 10^{-39}$

- ✓ `3CF398C0`: `0011 1100 1111 0011 1001 1000 1100 0000`
  $e + bias = 01111001 = 121 \rightarrow e = 121 - 127 = -6$
  $Mantissa = 1.11100111001100011 = 1.90309906$
  $X = 1.90309906 \times 2^{-6} = 0.029735922$

- ✓ `7F85B0AC`: `0111 1111 1000 0101 1011 0000 1010 1100`
  $e + bias = 11111111 = 255, f \neq 0$
  $X = NaN$

- ✓ `70DECADE`: `0111 0000 1101 1110 1100 1010 1101 1110`
  $e + bias = 11100001 = 225 \rightarrow e = 225 - 127 = 98$
  $Mantissa = 1.10111101100101011011110 = 1.740566015$
  $X = 1.740566015 \times 2^{98} = 5.5160738 \times 10^{29}$

- ✓ `90BEE950`: `1001 0000 1011 1110 1110 1001 0101 0000`
  $e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$
  $Mantissa = 1.0111110111010010101 = 1.491486191$
  $X = -1.491486191 \times 2^{-94} = -7.53008094 \times 10^{-29}$

- ✓ `DECAFC0FFEE80000`: `1101 1110 1100 1001 1111 1100 0000 1111 1111 1110 1110 1000 0000 0000 0000 0000`
  $e + bias = 10111101100 = 1516 \rightarrow e = 1516 - 1023 = 493$
  $Mantissa\ ([53\ 52]) = 1.100111111100000011111111111011101000 = 1.686538692214526$
  $X = -1.686538692214526 \times 2^{493} = -4.3130468 \times 10^{148}$

- ✓ `49A5DEAF8FAD8000`: `0100 1001 1010 0101 1101 1110 1010 1111 1000 1111 1010 1101 1000 0000 0000 0000`
  $e + bias = 10010011010 = 1178 \rightarrow e = 1178 - 1023 = 155$
  $Mantissa = 1.010111011110101011111000111110101101 1000 = 1.366866646996641$
  $X = 1.366866646996641 \times 2^{155} = 6.24274325 \times 10^{46}$

- ✓ `FFFACADE19801990`: `1111 1111 1111 1010 1100 1010 1101 1110 0001 1001 1000 0000 0001 1001 1001 0000`
  $e + bias = 11111111111 = 2047, f \neq 0$
  $X = NaN$

- ✓ `ACCEDE90BEAD5000`: `1010 1100 1100 1110 1101 1110 1001 0000 1011 1110 1010 1101 0101 0000 0000 0000`
  $e + bias = 01011001100 = 716 \rightarrow e = 716 - 1023 = -307$
  $Mantissa = 1.1110110111101001000010111110101011010101 = 1.929337258178748$
  $X = -1.929337258178748 \times 2^{-307} = -7.3994507 \times 10^{-93}$

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

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