Solutions - Homework 1

(Due date: September 15th)

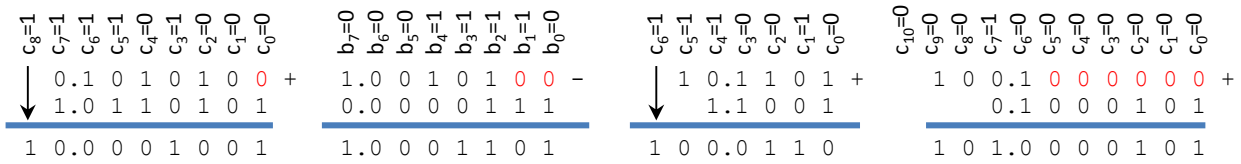
Presentation and clarity are very important! Show your procedure!

Problem 1 (15 pts)

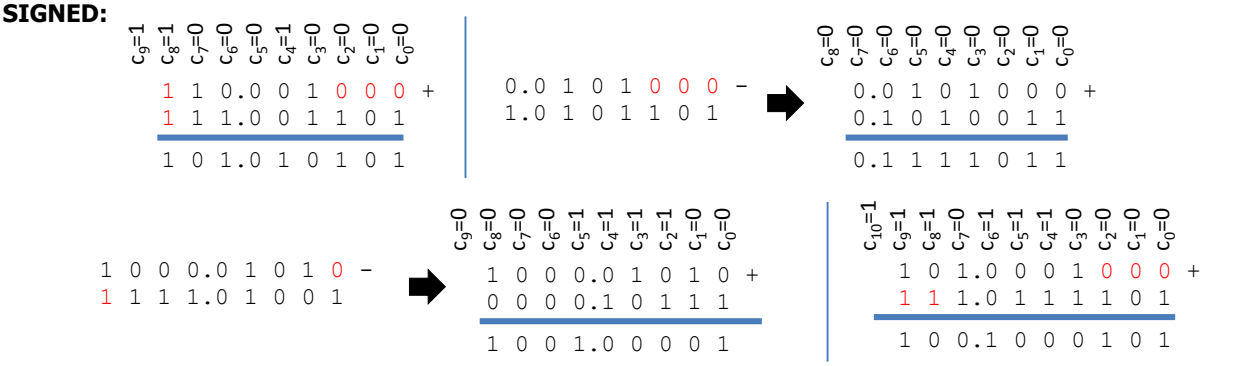
- Calculate the result of the additions and subtractions for the following fixed-point numbers.

UNSIGNED		SIGNED	
0.101010 + 1.0110101	1.00101 – 0.0000111	10.001 + 1.001101	0.0101 – 1.0101101
10.1101 + 1.1001	100.1 + 0.1000101	1000.0101 – 111.01001	101.0001 + 1.0111101

UNSIGNED:



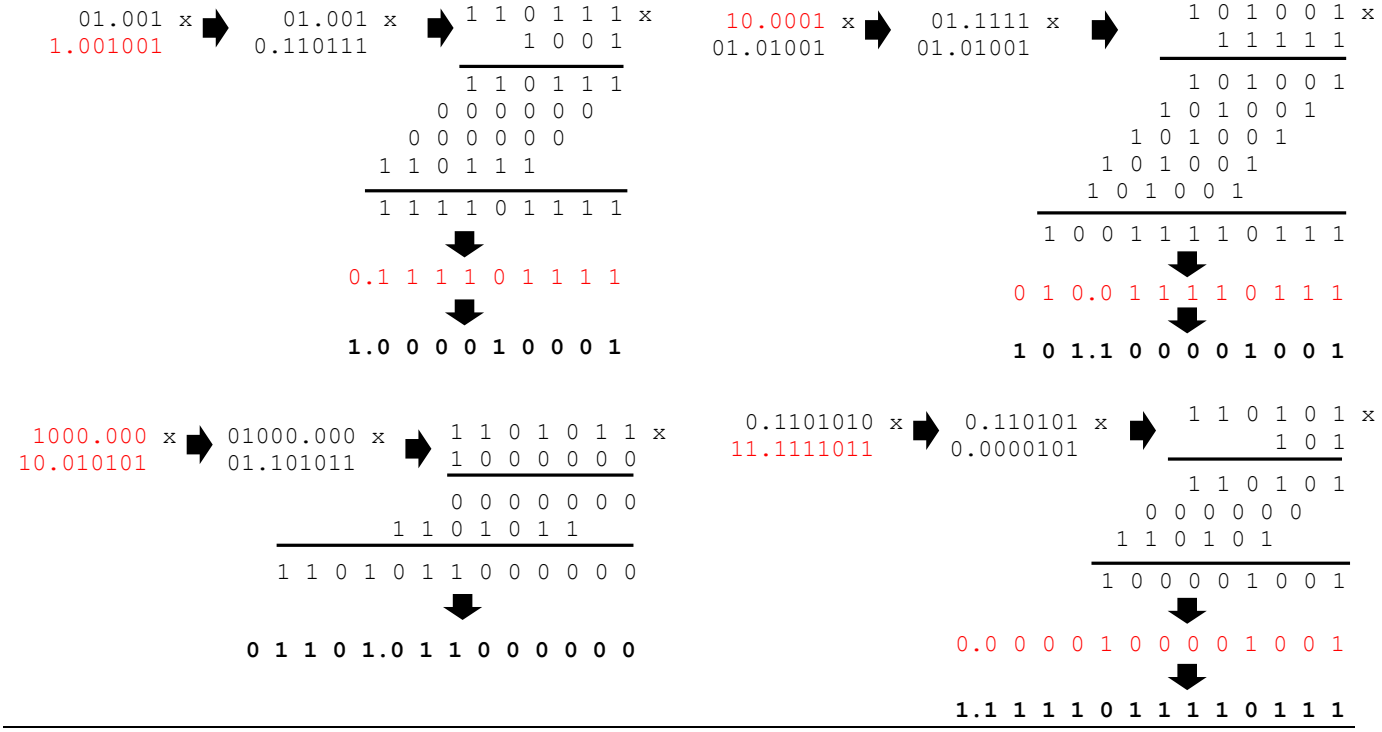
SIGNED:



Problem 2 (25 pts)

- Multiply the following signed fixed-point numbers:

01.001 × 1.001001	10.0001 × 01.01001	1000.000 × 10.010101	0.1101010 × 11.1111011



- 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	0.101010 ÷ 110.1001

✓ $\frac{101.1001}{1.011}$: 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): $\frac{10.0110}{01.01} = 2C(01.01) = 10.11$

✓ $\frac{0.101010}{110.1001}$: To unsigned (denominator) and then alignment, $a = 5$: $\frac{0.10101}{001.0111} = \frac{0.10101}{001.01110} \equiv \frac{10101}{101110}$

```
        000000111
    ____________
101110 | 101010000
         101110
         ------
          1001100
           101110
           ------
            111100
            101110
            ------
              1110
```

Append $x = 4$ zeros: $\frac{10101\mathbf{0000}}{101110}$
Integer Division:

$$Q = 111, R = 1110$$
$$\rightarrow Qf = 0.0111 (x = 4)$$

Final result (2C): $\frac{0.101010}{110.1001} = 2C(0.0111) = 1.1001$

PROBLEM 3 (20 PTS)

- We want to represent numbers between -255.9 and 234.5. 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 $-2^{n-1} \leq -255$, we have that $n \geq 9$, so we pick $n = 9$.

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 $[18\ 9]$.

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

-125.125	-232.625	-78.1875	212.3125

✓ 125.125 = 01111101.001 → -125.125 = 10000010.111
✓ 232.625 = 011101000.101 → -232.625 = 100010111.011
✓ 78.1875 = 01001110.0011 → -78.1875 = 10110001.1101
✓ 212.3125 = 011010100.0101

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
7	5	[12 7]	$[-2^4, 2^4 - 2^{-7}] = [-16, 15.9921875]$	66.2266 dB	0.0078125
12	4	[16 12]	$[-2^3, 2^3 - 2^{-12}] = [-8, 7.999755859375]$	90.3089 dB	0.000244140625
17	7	[24 17]	$[-2^6, 2^6 - 2^{-17}] = [-64, 63.999992370605469]$	138.4738 dB	0.00000762939453125

- 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
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 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)	
✓ E8000978	✓ 800BCCAA	✓ FA09D3784D039B7D	✓ FACADE50404900DB
✓ 80DE0FEE	✓ 7FFCAFEA	✓ 80DEADBEE9700400	✓ FE80000009AB00DE
✓ 3DE32856	✓ FACEB00C	✓ DECAFAAA0BEEF0A0	✓ DFC0FC0FFEE10800

✓ E8000978: 1110 1000 0000 0000 0000 1001 0111 1000
$e + bias = 11010000 = 208 \rightarrow e = 208 - 127 = 81$
$Mantissa$ ([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
  $e + bias = 00000001 = 1 \rightarrow e = 1 - 127 = -126$
  $Mantissa = 1.10111100000111111101110 = 1.734861135482788$
  $X = -1.734861135482788 \times 2^{-126} = -2.0393193 \times 10^{-38}$

- ✓ 3DE32856: 0011 1101 1110 0011 0010 1000 0101 0110
  $e + bias = 01111011 = 123 \rightarrow e = 123 - 127 = -4$
  $Mantissa = 1.11000110010100001010110 = 1.774668455123901$
  $X = 1.774668455123901 \times 2^{-4} = 0.110916778445244$

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

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

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

- ✓ FA09D3784D039B7D: 1111 1010 0000 1001 1101 0011 0111 1000 0100 1101 0000 0011 1001 1011 0111 1101
  $e + bias = 11110100000 = 1952 \rightarrow e = 1952 - 1023 = 929$
  $Mantissa\ ([53\ 52]) = 1.10011101001101111000010011010000001110011011011111101 = 1.614128399692816$
  $X = -1.614128399692816 \times 2^{929} = -7.3249 \times 10^{279}$

- ✓ 80DEADBEE9700400: 1000 0000 1101 1110 1010 1101 1011 1110 1110 1001 0111 0000 0000 0100 0000 0000
  $e + bias = 00000001101 = 13 \rightarrow e = 13 - 1023 = -1010$
  $Mantissa = 1.11101010110110111110111010010111000000000010000000000 = 1.917418395867799$
  $X = -1.917418395867799 \times 2^{-1010} = -1.7475 \times 10^{-304}$

- ✓ DECAFAAA0BEEF0A0: 1101 1110 1100 1010 1111 1010 1010 1010 0000 1011 1110 1110 1111 0000 1010 0000
  $e + bias = 10111101100 = 1516 \rightarrow e = 1516 - 1023 = 493$
  $Mantissa = 1.10101111101010101010000010111110111011110000010100000 = 1.686197325339016$
  $X = -1.686197325339016 \times 2^{493} = -4.31217 \times 10^{148}$

- ✓ FACADE50404900DB: 1111 1010 1100 1010 1101 1110 0101 0000 0100 0000 0100 1001 0000 0000 1101 1011
  $e + bias = 11110101100 = 1964 \rightarrow e = 1964 - 1023 = 941$
  $Mantissa = 1.10101101111001010000010000000100100100000000011011011 = 1.67927575217623$
  $X = -1.67927575217623 \times 2^{941} = -3.121439 \times 10^{283}$

- ✓ FE80000009AB00DE: 1111 1110 1000 0000 0000 0000 0000 0000 0000 1001 1010 1011 0000 0000 1101 1110
  $e + bias = 11111101000 = 2024 \rightarrow e = 2024 - 1023 = 1001$
  $Mantissa = 1.00000000000000000000000010011010101100000000011011110 = 1.00000003601604$
  $X = -1.00000003601604 \times 2^{1001} = -2.1430173 \times 10^{301}$

- ✓ DFC0FC0FFEE10800: 1101 1111 1100 0000 1111 1100 0000 1111 1111 1110 1110 0001 0000 1000 0000 0000
  $e + bias = 10111111100 = 1532 \rightarrow e = 1532 - 1023 = 509$
  $Mantissa = 1.00001111110000001111111111101110000100001000000000000 = 1.061538692113118$
  $X = -1.061538692113118 \times 2^{509} = -1.77911336 \times 10^{153}$