Solutions - Homework 2

(Due date: February 15th @ 7:30 pm)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (15 PTS)

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

01.101 × 1.011001	100.101 × 01.10001	10.000 × 10.10101

```
  01.101 x
1.011001
    ↓
  01.101 x        1 0 0 1 1 1 x
0.100111              1 1 0 1
              -----------------
                  1 0 0 1 1 1
              1 0 0 1 1 1
          1 0 0 1 1 1
        -------------------
          1 1 1 1 1 1 0 1 1
                ↓
      0.1 1 1 1 1 1 0 1 1
                ↓
        1.0 0 0 0 0 0 1 0 1
```

```
100.101 x
01.10001
    ↓
011.011 x              1 1 0 1 1 x
01.10001             1 1 0 0 0 1
                   ---------------
                       1 1 0 1 1
               1 1 0 1 1
             1 1 0 1 1
          ------------------------
          1 0 1 0 0 1 0 1 0 1 1
                   ↓
        0 1 0 1.0 0 1 0 1 0 1 1
                   ↓
        1 0 1 0.1 1 0 1 0 1 0 1
```

```
10.000 x
10.10101
    ↓
010.000 x          1 0 1 0 1 1 x
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
```

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

101.0101 ÷ 1.101	10.0101 ÷ 01.11	1.1011 ÷ 1.01101

✓ $\frac{101.0101}{1.101}$: 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): $\frac{101.0101}{1.101} = 0111.001$

✓ $\frac{10.0101}{01.11}$: 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: $\frac{11100\mathbf{0000}}{11100}$

Integer Division:

$Q = 1111, R = 1100$

$\rightarrow Qf = 0.1111\ (x = 4)$

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

✓ $\frac{1.1011}{1.01101}$: 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): $\frac{1.1011}{1.01101} = 0.1$

PROBLEM 2 (11 PTS)

- We want to represent numbers between -512 and 511.9997. What is the fixed-point format that requires the fewest number of bits for a resolution better or equal than 0.0005? (4 pts).

  Resolution: $2^{-p} \leq 0.0005 \rightarrow p \geq 10.96578 \rightarrow p = 11$

  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.
  $\Rightarrow -2^{n-p-1} \leq -512 \rightarrow n - p - 1 \geq 9 \rightarrow n - p \geq 10 \rightarrow n - p = 10 \rightarrow n = 21$

  However, since -512 is so close (in absolute value) to 511.9997, we also use the upper bound to determine the value of n, since there is a chance it might be different.
  $\Rightarrow 2^{n-p-1} - 2^{-p} \geq 511.9997 \longrightarrow 2^{n-p-1} \geq 511.9997 + 2^{-11} \rightarrow n - p - 1 \geq 9.000000530531985 \rightarrow n - p = 11 \longrightarrow n = 22$

  From these different n values, we use the largest one ($n = 22$).
  $\therefore$ The required Fixed-Point format is $[22\ 11]$.

  Another valid solution:
  We can also use $p = 12$. This provides better resolution than $p = 11$, but at the risk of increased number of total bits.
  Then $2^{n-p-1} - 2^{-p} \geq 511.9997 \longrightarrow 2^{n-p-1} \geq 511.9997 + 2^{-12} \rightarrow n - p - 1 \geq 8.99999984260147 \rightarrow n - p = 10 \longrightarrow n = 22$.
  $\therefore$ The Fixed-Point format is $[22\ 12]$. Same number of total bits, but better resolution.

- We want to represent numbers between -127.05 and 116.25. 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: $2^{-p} \leq 0.0015 \rightarrow p \geq 9.38 \rightarrow p = 10$

  Range of signed fixed-point format $[n\ p]$: $[-2^{n-p-1}, 2^{n-p-1} - 2^{-p}]$
  Here, we can safely use the lower bound to determine n.
  $\Rightarrow -2^{n-p-1} \leq -127.05 \rightarrow n - p - 1 \geq 6.9893 \rightarrow n - p \geq 7.9893 \rightarrow n - p = 8 \rightarrow n = 18$
  $\therefore$ The required Fixed-Point format is $[18\ 10]$.

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

-129.625	-69.1875	113.3125

  ✓ `129.625 = 010000001.101 → -129.625 = 101111110.011`
  ✓ `69.1875 = 01000101.0011 → -69.1875 = 10111010.1101`
  ✓ `113.3125 = 01110001.0101`

PROBLEM 3 (10 PTS)

- Complete the table for the following fixed-point formats (signed numbers): (4 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 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
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 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)	
✓ 10DBD800	✓ 7F8CACA0	✓ DECAFC0FFEE80000	✓ ACCEDE90BEAD5000
✓ 800BEEF0	✓ 70DECADE	✓ C9A7DEAFBEE00000	✓ 800CBEBEFACE0000

✓ 10DBD800: 0001 0000 1101 1011 1101 1000 0000 0000
$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$
$Significand\ ([24\ 23]) = 1.101101111011 = 1.717529296875$
$X = 1.717529296875 \times 2^{-94} = 8.6713 \times 10^{-29}$

✓ 800BEEF0: 1000 0000 0000 1011 1110 1110 1111 0000
$e + bias = 00000000 = 0 \rightarrow Denormal\ number\ \rightarrow e = -126$
$Significand = 0.00010111111011101111 = 0.093229293$
$X = -0.093229293 \times 2^{-126} = -1.09590508 \times 10^{-39}$

✓ 7F8CACA0: 0111 1111 1000 1100 1010 1100 1010 0000
$e + bias = 11111111 = 255, f \neq 0$
$X = NaN$

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

✓ DECAFC0FFEE80000: 1101 1110 1100 1010 1111 1100 0000 1111 1111 1110 1110 1000 0000 0000 0000 0000
$e + bias = 10111101100 = 1516 \rightarrow e = 1516 - 1023 = 493$
$Significand\ ([53\ 52]) = 1.1010111111000000111111111110111101000 = 1.686538692214526$
$X = -1.686538692214526 \times 2^{493} = -4.3130468 \times 10^{148}$

✓ C9A7DEAFBEE00000: 1100 1001 1010 0111 1101 1110 1010 1111 1011 1110 1110 0000 0000 0000 0000 0000
$e + bias = 10010011010 = 1178 \rightarrow e = 1178 - 1023 = 155$
$Significand = 1.0111110111101010111110111110111 0 = 1.491866822820157$
$X = -1.491866822820157 \times 2^{155} = -6.8136431 \times 10^{46}$

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

✓ 800CBEBEFACE0000: 1000 0000 0000 1100 1011 1110 1011 1110 1111 1010 1100 1110 0000 0000 0000 0000
$e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
$Significand = 0.11001011111010111110111110101100111 0 = 0.796568851190386$
$X = -0.796568851190386 \times 2^{-1022} = -1.772424527 \times 10^{-308}$

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 results. Provide the 32-bit result.

✓ 3DE38C80 + 3A80D980	✓ 80A18000 - 83CEC000	✓ 7A09D300 × 4D080000	✓ 800C0000 ÷ 494C0000
✓ 80123000 + 804E8000	✓ 09DECAF0 - 7AD90000	✓ 90DECADE × FF800000	✓ 7F800000 ÷ 800ABBAA
✓ 7FEEFCA0 + FACADE90	✓ F0B1ABEE - 7F800000	✓ 0B09A000 × 8FACC000	✓ C9746000 ÷ 40490000

✓ $X =$ 3DE38C80 + 3A80D980:

3DE38C80: 0011 1101 1110 0011 1000 1100 1000 0000
$e + bias = 01111011 = 123\ \rightarrow\ e = 123 - 127 = -4$ $Significand = 1.1010011100011001$
3DE38C80 $= 1.1100011100011001 \times 2^{-4}$

3A80D980: 0011 1010 1000 0000 1101 1001 1000 0000
$e + bias = 01110101 = 117 \rightarrow\ e = 117 - 127 = -10$ $Significand = 1.0000000110110011$
3A80D980 $= 1.000000011011 \times 2^{-10}$

$X = 1.1100011100011001 \times 2^{-4} + 1.000000011011 \times 2^{-10}$

$X = 1.1100011100011001 \times 2^{-4} + \dfrac{1.0000000110110011}{2^6} \times 2^{-4}$

$X = 1.1100011100011001 \times 2^{-4} + 0.0000010000000110110011 \times 2^{-4}$

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

$X = 1.1100101100011111110011 \times 2^{-4}, e + bias = -4 + 127 = 123 = 01111011$

X = `0011 1101 1110 0101 1000 1111 1110 0110 = 3DE58FE6`

✓ X = `80123000 + 804E8000`:

`80123000`: `1000 0000 0001 0010 0011 0000 0000 0000`

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

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

`804E8000`: `1000 0000 0100 1110 1000 0000 0000 0000`

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

`804E8000` = $-0.10011101 \times 2^{-126}$

```
 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
```

$X = -0.00100100011 \times 2^{-126} - 0.10011101 \times 2^{-126}$

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

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

✓ X = `7FEEFCA0 + FACADE90`:

`7FEEFCA0`: `0111 1111 1110 1110 1111 1100 1010 0000`

$e + bias = 11111111 = 255, f \neq 0$

`7FBEADED` = NaN

$X = NaN + \# = NaN$

X = `7FEEFCA0`

✓ X = `80A18000 - 83CEC000`:

`80A18000`: `1000 0000 1010 0001 1000 0000 0000 0000`

$e + bias = 00000001 = 1 \rightarrow e = 1 - 127 = -126 \qquad Significand = 1.01000011$

`80A18000` = $-1.01000011 \times 2^{-126}$

`83CEC00`: `1000 0011 1100 1110 1100 0000 0000 0000`

$e + bias = 00000111 = 7 \rightarrow e = 7 - 127 = -120 \qquad Significand = 1.100111011$

`83CEC000` = $-1.100111011 \times 2^{-120}$

$X = -1.01000011 \times 2^{-126} + 1.100111011 \times 2^{-120} = -\dfrac{1.01000011}{2^6} \times 2^{-120} + 1.100111011 \times 2^{-120}$

$X = -0.00000101000011 \times 2^{-120} + 1.100111011 \times 2^{-120}$

To subtract these numbers, we first convert to 2C representation:

$R = 1.100111011 - 0.00000101000011 \equiv 01.100111011 + 1.11111010111101$ (2C addition)

The result (in 2C): $R = 01.10011000011101$

* *Note that you can also do unsigned subtraction.*

For floating point, we need to convert R to sign and magnitude:

$\Rightarrow R(SM) = +1.10011000011101$

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

$X = +1.10011000011101 \times 2^{-120}, e + bias = -120 + 127 = 7 = 00000111$

X = `0000 0011 1100 1100 0011 1010 0000 0000 = 03CC3A00`

✓ X = `09DECAF0 - 7AD90000`:

`09DECAF0`: `0000 1001 1101 1110 1100 1010 1111 0000`

$e + bias = 00010011 = 19 \rightarrow e = 19 - 127 = -108 \qquad Significand = 1.10111101100101011111$

`09DECAF0` = $1.1011110110010101111 \times 2^{-108}$

`7AD90000`: `0111 1010 1101 1001 0000 0000 0000 0000`

$e + bias = 11110101 = 245 \rightarrow e = 245 - 127 = 118 \qquad Significand = 1.1011001$

`7AD90000` = 1.1011001×2^{118}

$X = 1.1011110110010101111 \times 2^{-108} - 1.1011001 \times 2^{118}$

$X = \frac{1.1011110110010101111}{2^{226}} \times 2^{118} - 1.1011001 \times 2^{118}$

Representing the number divided by 2^{226} requires more than $p+1 = 24$ bits. Thus, we round down this operand to 0.

$X = -1.1011001 \times 2^{118}, e + bias = 118 + 127 = 245 = 11110101$

X = 1111 1010 1101 1001 0000 0000 0000 0000 = FAD90000

✓ X = F0B1ABEE - 7F800000:

7F800000: 0111 1111 1000 0000 0000 0000 0000 0000

$e + bias = 11111111 = 255, f = 0$

7F800000 = $+\infty$

$X = \# - (+\infty) = -\infty$

X = FF800000

✓ X = 7A09D300 × 4D080000:

7A09D300: 0111 1010 0000 1001 1101 0011 0000 0000

$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $Significand = 1.000100111010011$

7A09D300 = $1.000100111010011 \times 2^{117}$

4D080000: 0100 1101 0000 1000 0000 0000 0000 0000

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

4D080000 = 1.0001×2^{27}

$X = 1.000100111010011 \times 2^{117} \times 1.0001 \times 2^{27} = 1.0010010011100000011 \times 2^{144}$

$e + bias = 144 + 127 = 271 > 254$

Here, there is an overflow. The value X is assigned to $+\infty$.

X = 0111 1111 1000 0000 0000 0000 0000 0000 = 7F800000

✓ X = 90DECADE × FF800000:

FF800000: 1111 1111 1000 0000 0000 0000 0000 0000

$e + bias = 11111111 = 255, f = 0$

FF800000 = $-\infty$

$X = (-|\#|) \times -\infty = \infty$

X = 0111 1111 1000 0000 0000 0000 0000 0000 = 7F800000

✓ X = 0B09A000 × 8FACC000:

0B092000: 0000 1011 0000 1001 1010 0000 0000 0000

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

0B092000 = $1.0001001101 \times 2^{-105}$

8FACC000: 1000 1111 1010 1100 1100 0000 0000 0000

$e + bias = 00011111 = 31 \rightarrow e = 31 - 127 = -96$ $Significand = 1.010110011$

0FACE000 = $1.010110011 \times 2^{-96}$

$X = 1.0001001101 \times 2^{-105} \times -1.010110011 \times 2^{-96} = -1.0111001101111010111 \times 2^{-201} = -0 \times 2^{-126}$

$e + bias = -201 + 127 = -74 < 0$

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 = 800C0000 ÷ 494C0000:

800C0000: 1000 0000 0000 1100 0000 0000 0000 0000

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

800C0000 = -0.00011×2^{-126}

494C0000: 0100 1001 0100 1100 0000 0000 0000 0000

$e + bias = 10010010 = 146 \rightarrow e = 146 - 127 = 19$ $Significand = 1.10011$

494C0000 = 1.10011×2^{19}

$X = -\frac{0.00011 \times 2^{-126}}{1.10011 \times 2^{19}}$

```
        0000001111
110011 )1100000000
         110011
         1011010
          110011
          1001110
           110011
            110110
            110011
                11
```

Alignment:

$$\frac{0.00011}{1.10011} = \frac{000011}{110011}$$

Append $x = 8$ zeros: $\frac{1100000000}{110011}$

Integer division

$Q = 1111, R = 11 \rightarrow Qf = 0.00001111$

Thus: $X = -\frac{0.00011\times 2^{-126}}{1.10011\times 2^{19}} = -0.00001111 \times 2^{-145} = -(0.00001111 \times 2^{-19}) \times 2^{-126}$

$X = -0.000\ 0000\ 0000\ 0000\ 0000\ 0000\ 1111 \times 2^{-126}$. $Denormal \rightarrow e + bias = 00000000$

X = 1000 0000 0000 0000 0000 0000 0000 0000 = 80000000

✓ X = 7F800000 ÷ 800ABBAA:

7F800000: 0111 1111 1000 0000 0000 0000 0000 0000

$e + bias = 11111111 = 255, f = 0$

0F800000 = $+\infty$

$X = +\infty \div -|\#| = -\infty$

X = FF800000

✓ X = C9746000 ÷ 40490000:

C9742000: 1100 1001 0111 0100 0110 0000 0000 0000

$e + bias = 10010010 = 146 \rightarrow e = 146 - 127 = 19$ $\quad Significand = 1.1110100011$

C97460000 = $-1.1110100001 \times 2^{19}$

40490000: 0100 0000 0100 1001 0000 0000 0000 0000

$e + bias = 10000000 = 128 \rightarrow e = 128 - 127 = 1$ $\quad Significand = 1.1001001$

40490000 = 1.1001001×2^{1}

$$X = -\frac{1.1110100011 \times 2^{19}}{1.1001001 \times 2^{1}}$$

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

Alignment:

$$\frac{1.1110100011}{1.1001001} = \frac{1.1110100011}{1.1001001000} = \frac{11110100011}{11001001000}$$

Append $x = 8$ zeros: $\frac{1111010001100000000}{11001001000}$

Integer division

$Q = 100110111, R = 110001000 \rightarrow Qf = 1.00110111$

Thus: $X = -\frac{1.1110100011\times 2^{19}}{1.1001001\times 2^{1}} = -1.00110111 \times 2^{18}$

$e + bias = 18 + 127 = 145 = 10010001$

X = 1100 1000 1001 1011 1000 0000 0000 0000 = C89B8000