Solutions - Homework 2

(Due date: February 9th @ 5:30 pm)
Presentation and clarity are very important!

PROBLEM 1 (10 PTS)

- Complete the following table. Use the fewest number of bits in each case.
  You MUST show your conversion procedure. **No procedure ≡ zero points**.

REPRESENTATION			
Decimal	**Sign-and-magnitude**	**1's complement**	**2's complement**
-129	11000 0001	101111110	101111111
-46	1101110	1010001	1010010
154	010011010	010011010	010011010
-64	11000000	10111111	1000000
-10	1001010	10101	10110

PROBLEM 2 (16 PTS)

a) Perform the following additions and subtractions of the following unsigned integers. Use the fewest number of bits n to represent both operators. Indicate every carry (or borrow) from c_0 to c_n (or b_0 to b_n). For the addition, determine whether there is an overflow. For the subtraction, determine whether we need to keep borrowing from a higher bit. (6 pts)

Example (n=8):

✓ 54 + 210

```
                 c8=1 c7=1 c6=1 c5=1 c4=0 c3=1 c2=1 c1=0 c0=0

 54 = 0x36 =        0    0    1    1    0    1    1    0  +
210 = 0xD2 =        1    1    0    1    0    0    1    0
-------------------------------------------------------
Overflow! →    1    0    0    0    0    1    0    0    0
```

✓ 77 - 194

```
Borrow out! → b8=1 b7=0 b6=0 b5=0 b4=0 b3=0 b2=1 b1=0 b0=0

 77 = 0x4D =        0    1    0    0    1    1    0    1  -
194 = 0xC2 =        1    1    0    0    0    0    1    0
-------------------------------------------------------
                    0    0    0    0    1    0    1    1
```

✓ 244 + 267
✓ 39 + 218

```
No Overflow  c9=0 c8=0 c7=0 c6=0 c5=0 c4=0 c3=0 c2=0 c1=0 c0=0

244 = 0x0F4 =   0    1    1    1    1    0    1    0    0  +
267 = 0x10B =   1    0    0    0    0    1    0    1    1
-------------------------------------------------------
511 = 0x1FF =   1    1    1    1    1    1    1    1    1
```

```
               c8=1 c7=1 c6=1 c5=1 c4=1 c3=1 c2=1 c1=0 c0=0

 39 = 0x27 =        0    0    1    0    0    1    1    1  +
218 = 0xDA =        1    1    0    1    1    0    1    0
-------------------------------------------------------
Overflow! →    1    0    0    0    0    0    0    0    1
```

✓ 251 - 126
✓ 169 - 201

```
No Borrow Out  b8=0 b7=1 b6=1 b5=1 b4=1 b3=1 b2=0 b1=0 b0=0

251 = 0xFB =        1    1    1    1    1    0    1    1  -
126 = 0x7E =        0    1    1    1    1    1    1    0
-------------------------------------------------------
125 = 0x7D =        0    1    1    1    1    1    0    1
```

```
Borrow out! →  b8=1 b7=1 b6=0 b5=0 b4=0 b3=0 b2=0 b1=0 b0=0

169 = 0xA9 =        1    0    1    0    1    0    0    1  -
201 = 0xC9 =        1    1    0    0    1    0    0    1
-------------------------------------------------------
       0xE0 =       1    1    1    0    0    0    0    0
```

b) Perform the following operations, where numbers are represented in 2's complement arithmetic: (10 pts)

✓ -70 + 63
✓ -257 + 256
✓ 490 + 47
✓ -127 - 183

- For each case:
  - ✓ Determine the minimum number of bits required to represent both summands. You might need to sign-extend one of the summands, since for proper summation, both summands must have the same number of bits.
  - ✓ Perform the binary addition in 2's complement arithmetic. The result must have the same number of bits as the summands.
  - ✓ Determine whether there is overflow by:
    - i. Using c_n, c_{n-1} (carries).
    - ii. Performing the operation in the decimal system and checking whether the result is within the allowed range for n bits, where n is the minimum number of bits for the summands.
  - ✓ If there is overflow and we want to avoid it, what is the minimum number of bits required to represent both the summands and the result?

n = 8 bits

$c_8 \oplus c_7=0$
No Overflow

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

```
-70 = 1 0 1 1 1 0 1 0 +
 63 = 0 0 1 1 1 1 1 1
---------------------
 -7 = 1 1 1 1 1 0 0 1
```

$-70 + 63 = -1 \in [-2^7, 2^7-1] \rightarrow$ no overflow

n = 10 bits

$c_{10} \oplus c_9=0$
No Overflow

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

```
-257 = 1 0 1 1 1 1 1 1 1 1 +
 256 = 0 1 0 0 0 0 0 0 0 0
--------------------------
  -1 = 1 1 1 1 1 1 1 1 1 1
```

$-257+256 = -1 \in [-2^9, 2^9-1] \rightarrow$ no overflow

n = 10 bits

$c_{10} \oplus c_9=1$
Overflow!

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

```
490 = 0 1 1 1 1 0 1 0 1 0 +
 47 = 0 0 0 0 1 0 1 1 1 1
-------------------------
      1 0 0 0 0 1 1 0 0 1
```

$490+47 = 537 \notin [-2^9, 2^9-1] \rightarrow$ overflow!

To avoid overflow:
n = 11 bits (sign-extension)

$c_{11} \oplus c_{10}=0$
Overflow!

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

```
490 = 0 0 1 1 1 1 0 1 0 1 0 +
 47 = 0 0 0 0 0 1 0 1 1 1 1
---------------------------
      0 1 0 0 0 0 1 1 0 0 1
```

$490+47 = 537 \in [-2^{10}, 2^{10}-1] \rightarrow$ no overflow

n = 9 bits

$c_9 \oplus c_8=1$
Overflow!

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

```
-127 = 1 1 0 0 0 0 0 0 1 +
-183 = 1 0 1 0 0 1 0 0 1
------------------------
       0 1 1 0 0 1 0 1 0
```

$-127-183 = -310 \notin [-2^8, 2^8-1] \rightarrow$ overflow!

To avoid overflow:
n = 10 bits (sign-extension)

$c_{10} \oplus c_9=0$
No Overflow

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

```
-255 = 1 1 1 0 0 0 0 0 0 1 +
-230 = 1 1 0 1 0 0 1 0 0 1
--------------------------
-310 = 1 0 1 1 0 0 1 0 1 0
```

$-127-183 = -310 \in [-2^9, 2^9-1] \rightarrow$ no overflow

PROBLEM 3 (27 PTS)

a) Calculate the result of the additions and subtractions for the following signed fixed-point numbers. Use the minimum number of bits for both operands and result so that overflow is avoided. (6 pts.)

0.1110 + 1.010111	11.11101 − 0.001101	1.0001 + 1.001001	0.00101 − 101.01101

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

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

```
1 1.1 1 1 0 1 0 −
0 0.0 0 1 1 0 1
```

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

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

```
0 0 0.0 0 1 0 1 −
1 0 1.0 1 1 0 0
```

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

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

b) Calculate the result of the multiplication of the following signed fixed-point numbers: (9 pts.)

10.0101 × 0.10111	101.1101 × 110.011	01.101 × 100.01011

```
10.0101 x
0.10111

01.1011 x        1 1 0 1 1 x
0.10111          1 0 1 1 1
                 ---------
                 1 1 0 1 1
               1 1 0 1 1
             1 1 0 1 1
           1 1 0 1 1
     -------------------
     1 0 0 1 1 0 1 1 0 1

     0 1.0 0 1 1 0 1 1 0 1

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

```
101.1101 x
110.011

010.0011 x       1 0 0 0 1 1 x
001.101                1 1 0 1
                 -----------
                 1 0 0 0 1 1
               0 0 0 0 0 0
             1 0 0 0 1 1
           1 0 0 0 1 1
     ---------------------
     0 1 1 1 0 0 0 1 1 1

     0 1 1.1 0 0 0 1 1 1
```

```
01.101 x
100.01011

01.101 x         1 1 1 0 1 0 1 x
011.10101                1 1 0 1
                 -------------
                 1 1 1 0 1 0 1
               0 0 0 0 0 0
             1 1 1 0 1 0 1
           1 1 1 0 1 0 1
     -----------------------
     1 0 1 1 1 1 1 0 0 0 1

     0 1 0 1.1 1 1 1 0 0 0 1

     1 0 1 0.0 0 0 0 1 1 1 1
```

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

10.0101 ÷ 01.011	01.0111 ÷ 01.11	01.01110 ÷ 1.011	1.1101 ÷ 10.001

✓ $\frac{10.0101}{01.011}$: To positive (numerator), alignment, and then to unsigned: $a = 4$: $\frac{01.1011}{01.011} = \frac{01.1011}{01.0110} \equiv \frac{11011}{10110}$

```
          000010011
10110 )  110110000
         10110
           101000
            10110
           100100
            10110
             1110
```

Append $x = 4$ zeros: $\frac{11011\mathbf{0000}}{10110}$
Unsigned Integer Division:

$Q = 10011, R = 1110$
$\rightarrow Qf = 1.0011\ (x = 4)$ ✶ Qf here is represented as an unsigned number

Final result (2C): $\frac{10.0101}{01.011} = 2C(01.0011) = 10.1101$

✓ $\frac{01.0111}{01.11}$: Alignment, and then to unsigned: $a = 4$: $\frac{01.0111}{01.1100} \equiv \frac{10111}{11100}$

```
          000001101
11100 )  101110000
          11100
          100100
           11100
           100000
            11100
              100
```

Append $x = 4$ zeros: $\frac{10111\mathbf{0000}}{11100}$
Unsigned Integer Division:

$Q = 1101, R = 100$
$\rightarrow Qf = 0.1101 (x = 4)$

Final result (2C): $\frac{01.0111}{01.11} = 0.1101$ (this is represented as a signed number)

✓ $\frac{01.0111}{1.011}$: To positive (denominator), alignment, and then to unsigned, $a = 4$: $\frac{01.0111}{0.101} = \frac{01.0111}{0.1010} \equiv \frac{10111}{1010}$

```
         000100100
1010 )  101110000
        1010
          1100
          1010
           1000
```

Append $x = 4$ zeros: $\frac{10111\mathbf{0000}}{1010}$
Integer Division:

$Q = 100100, R = 1000$
$\rightarrow Qf = 10.01 (x = 4)$ ✶ Qf here is represented as an unsigned number

Final result (2C): $\frac{01.0111}{1.011} = 2C(010.01) = 101.1100$

✓ $\frac{1.1101}{10.001}$: To unsigned and then alignment, $a = 4$: $\frac{0.0011}{01.111} = \frac{0.0011}{01.1110} \equiv \frac{11}{11110}$

Append $x = 4$ zeros: $\frac{11\mathbf{0000}}{11110}$
Integer Division:

```
          000001
11110 )  110000
          11110
          10010
```

$Q = 1, R = 10010$
$\rightarrow Qf = 0.0001 (x = 4)$

Final result (2C): $\frac{1.1101}{10.001} = 0.0001$

PROBLEM 4 (10 PTS)

a) We want to represent numbers between -128.7 and 179. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0005? (3 pts.)

2C representation for integers: -2^{n-1} to $2^{n-1} - 1$. For $2^{n-1} - 1 \geq 179$, 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 0.0005:

$$2^{-p} \leq 0.0005 \rightarrow p \geq 10.0658 \rightarrow p = 11$$

Then the Fixed Point format required in $[20\ 11]$.

b) 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? (3 pts.)

2C representation for integers: -2^{n-1} to $2^{n-1}-1$. For $-2^{n-1} \leq -255.9$, 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 $[18\ 9]$.

c) Represent these numbers in Fixed Point Arithmetic (signed numbers). For each case, select the minimum number of bits

-127.3125	232.21875

✓ 127.3125 = 01111111.0101 → -125.125 = 10000000.1011
✓ 232.21875 = 011101000.00111

PROBLEM 5 (9 PTS)

a) 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]	[-16, 15.9922]	66.23	0.0078125
12	4	[16 12]	[-8, 7.9998]	90.31	0.0002441
17	7	[24 17]	[-64,63.99999]	138.47	0.00000763

b) 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
6	9	9.31×10^{-10}	4.29×10^{9}	$[-2^5+2, 2^5-1] = [-30,31]$	$[1, 2-2^{-9}] = [1,1.998046875]$
7	16	2.17×10^{-19}	1.84×10^{19}	$[-2^6+2, 2^6-1] = [-62,63]$	$[1, 2-2^{-16}] = [1,1.9999847412109 38]$
10	37	2.98×10^{-154}	1.34×10^{154}	$[-2^9+2, 2^9-1] = [-510,511]$	$[1, 2-2^{-37}] = [1,1.9999999999992724]$

PROBLEM 6 (28 PTS)

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

Single (32 bits)		Double (64 bits)	
✓ F8000378	✓ 7FFCDEAC	✓ 8009DECADE080000	✓ 7FF0000000000000
✓ 801DECAF	✓ B300D959	✓ FFFDECAFC0FFEE90	✓ FACADEDECADE1990

✓ F8000378: 1111 1000 0000 0000 0000 0011 0111 1000
$e + bias = 11110000 = 240 \rightarrow e = 240 - 127 = 113$
$Mantissa\ ([24\ 23]) = 1.00000000000001101111000 = 1.000105857849121$
$X = -1.000105857849121 \times 2^{113} = -1.0385693 \times 10^{34}$

✓ 7FFCDEAC: 0111 1111 1111 1100 1101 1110 1010 1100
$e + bias = 11111111 = 255, f \neq 0$
$X = NaN$

✓ 801DECAF: 1000 0000 0001 1101 1110 1100 1010 1111
$e + bias = 00000000 = 0 \rightarrow Denormal\ number \rightarrow e = -126$
$Mantissa = 0.00111011110110010101111 = 0.233785510063171$
$X = -0.233785510063171 \times 2^{-126} = -2.748135 \times 10^{-39}$

✓ `B300D959`: 1011 0011 0000 0000 1101 1001 0101 1001
$e + bias = 01100110 = 102 \rightarrow e = 102 - 127 = -25$
$Mantissa = 1.00000001101100101011001 = 1.006632924079895$
$X = -1.006632924079895 \times 2^{-25} = -2.99999 \times 10^{-8}$

✓ `8009DECADE080000`: 1000 0000 0000 1001 1101 1110 1100 1010 1101 1110 0000 1000 0000 0000 0000 0000
$e + bias = 00000000000 = 0 \rightarrow Denormal\ number \rightarrow e = -1022$
$Mantissa\ ([53\ 52]) = 0.1001110111101100101011011110000010000000000000000000 = 0.616892688092776$
$X = -0.616892688092776 \times 2^{-1022} = -1.372631 \times 10^{-308}$

✓ `7FF0000000000000`: 0111 1111 1111 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
$e + bias = 11111111111 = 2047, f = 0$
$X = +\infty$

✓ `FFFDECAFC0FFEE90`: 1111 1111 1111 1101 1110 1100 1010 1111 1100 0000 1111 1111 1110 1110 1001 0000
$e + bias = 11111111111 = 2047, f \neq 0$
$X = NaN$

✓ `FACADEDECADE1990`: 1111 1010 1100 1010 1101 1110 1101 1110 1100 1010 1101 1110 0001 1001 1001 0000
$e + bias = 11110101100 = 1964 \rightarrow e = 1964 - 1023 = 941$
$Mantissa = 1.1010110111101101111011001010110111100001100100010000 = 1.679411690188832$
$X = -1.679411690188832 \times 2^{941} = -3.1216419 \times 10^{283}$

b) Calculate the result of the following operations with 32-bit floating point numbers. Truncate the results when required. When doing fixed-point division, use 8 fractional bits. Show your procedure. (20 pts.)

✓ `40B00000 + C2FA8000` ✓ `42FA8000 + C0E00000`	✓ `10DAD000 - 90FAD000` ✓ `3DE38866 - B300D959`	✓ `7AB80000 × 81800000` ✓ `FA09D300 × 4D080000`	✓ `FA390000 ÷ 48400000` ✓ `FF800000 ÷ 09FE0090`

✓ $X =$ `40B00000 + C2FA8000`:
`40B00000`: 0100 0000 1011 0000 0000 0000 0000 0000
$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $Mantissa = 1.011$
`40B00000` $= 1.011 \times 2^2$

`C2FA8000`: 1100 0010 1111 1010 1000 0000 0000 0000
$e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $Mantissa = 1.11110101$
`C2FA8000` $= -1.11110101 \times 2^6$

$$X = 1.011 \times 2^2 - 1.11110101 \times 2^6 = \frac{1.011}{2^4} \times 2^6 - 1.11110101 \times 2^6 = (0.0001011 - 1.11110101) \times 2^6$$

To subtract these unsigned numbers, we first convert to 2C:
$R = 0.0001011 - 01.11110101 = 0.0001011 + 10.00001011$
The result in 2C is: $R = 10.00100001$, $-R = 01.11011111$
For floating point, we need the result in sign-and-magnitude:
$\Rightarrow R(SM) = -1.11011111$

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

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

$X = -1.11011111 \times 2^6$, $e + bias = 6 + 127 = 133 = 10000101$
X = 1100 0010 1110 1111 1000 0000 0000 0000 = `C2EF8000`

✓ $X =$ `42FA8000 + C0E00000`:
`42FA8000`: 0100 0010 1111 1010 1000 0000 0000 0000
$e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $Mantissa = 1.11110101$
`42FA8000` $= 1.11110101 \times 2^6$

`C0E00000`: 1100 0000 1110 0000 0000 0000 0000 0000
$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $Mantissa = 1.11$
`C0E00000` $= -1.11 \times 2^2$

$$X = 1.11110101 \times 2^6 - 1.11 \times 2^2 = 1.11110101 \times 2^6 - \frac{1.11}{2^4} \times 2^6 = (1.11110101 - 0.000111) \times 2^6$$

To subtract these unsigned numbers, we first convert to 2C:
$R = 01.11110101 - 0.000111 = 01.11110101 + 1.111001$
The result in 2C is: $R = 01.11011001$
For floating point, we need to convert to sign-and-magnitude:
$\Rightarrow R(SM) = +1.11011001$

```
c10=1 c9=1 c8=1 c7=1 c6=1 c5=0 c4=0 c3=1 c2=0 c1=0 c0=0
 0 1.1 1 1 1 0 1 0 1 +
 1 1.1 1 1 0 0 1 0 0
 -------------------
 0 1.1 1 0 1 1 0 0 1
```

$X = 1.11011001 \times 2^6$, $e + bias = 6 + 127 = 133 = 10000101$
X = 0100 0010 1110 1100 1000 0000 0000 0000 = 42EC8000

✓ X = 10DAD000 - 90FAD000:
10DAD000: 0001 0000 1101 1010 1101 0000 0000 0000
$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$ $Mantissa = 1.10110101101$
10DAD000 = $1.10110101101 \times 2^{-94}$

90FAD000: 1001 0000 1111 1010 1101 0000 0000 0000
$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$ $Mantissa = 1.11110101101$
90FAD000 = $-1.11110101101 \times 2^{-94}$

$X = 1.10110101101 \times 2^{-94} + 1.11110101101 \times 2^{-94}$

```
c12=1 c11=1 c10=1 c9=1 c8=1 c7=0 c6=1 c5=0 c4=1 c3=1 c2=0 c1=1 c0=0
   1.1 0 1 1 0 1 0 1 1 0 1 +
   1.1 1 1 1 0 1 0 1 1 0 1
   -----------------------
 1 1.1 0 1 0 1 0 1 1 0 1 0
```

$X = 11.1010101101 \times 2^{-94} = 1.11010101101 \times 2^{-93}$
$e + bias = -93 + 127 = 34 = 00100010$
X = 0001 0001 0110 1010 1101 0000 0000 0000 = 116AD000

✓ X = 3DE38866 - B300D959:
3DE38866: 0011 1101 1110 0011 1000 1000 0110 0110
$e + bias = 01111011 = 123 \rightarrow e = 123 - 127 = -4$ $Mantissa = 1.11000111000100001100110$
3DE38866 = $1.11000111000100001100110 \times 2^{-4}$

B300D959: 1011 0011 0000 0000 1101 1001 0101 1001
$e + bias = 01100110 = 102 \rightarrow e = 102 - 127 = -25$ $Mantissa = 1.00000001101100101011001$
B300D959 = $-1.00000001101100101011001 \times 2^{-25}$

$$X = 1.11000111000100001100110 \times 2^{-4} + 1.00000001101100101011001 \times 2^{-25}$$
$$X = 1.11000111000100001100110 \times 2^{-4} + \frac{1.00000001101100101011001}{2^{21}} \times 2^{-4}$$
$$X = 1.11000111000100001100110 \times 2^{-4} + 0.00000000000000000000001 \times 2^{-4}$$

$X = 1.11000111000100001101010 \times 2^{-4} = 1.7776 \times 2^{-4} = 0.111100003123282$
X = 0011 1101 1110 0011 1000 1000 0110 1010 = 3DE3886A

✓ X = 7AB80000 × 81800000:
7AB80000: 0111 1010 1011 1000 0000 0000 0000 0000
$e + bias = 11110101 = 245 \rightarrow e = 245 - 127 = 118$ $Mantissa = 1.0111$
7AB80000 = 1.0111×2^{118}

81800000: 1000 0001 1000 0000 0000 0000 0000 0000
$e + bias = 00000011 = 3 \rightarrow e = 3 - 127 = -124$ $Mantissa = 1.0$
81800000 = -1.0×2^{-124}

$X = 1.0111 \times 2^{118} \times (-1.0 \times 2^{-124}) = -1.0111 \times 2^{-6}$
$e + bias = -6 + 127 = 121 = 01111001$
X = 1011 1100 1011 1000 0000 0000 0000 0000 = BCB800000

✓ X = FA09D300 × 4D080000:
FA09D300: 1111 1010 0000 1001 1101 0011 0000 0000
$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $Mantissa = 1.00010011101001100000000$
FA09D300 = $-1.000100111010011 \times 2^{117}$

4D080000: 0100 1101 0000 1000 0000 0000 0000 0000
$e + bias = 10011010 = 154 \rightarrow e = 154 - 127 = 27$ $Mantissa = 1.00010000000000000000000$
4D080000 = 1.0001×2^{27}

$X = -1.000100111010011 \times 2^{117} \times 1.0001 \times 2^{27}$
$X = -1.0010010011100000011 \times 2^{144}$
$e + bias = 144 + 127 = 271 > 254$

In this case, there is an overflow. The value X is assigned to $-\infty$.
X = 1111 1111 1000 0000 0000 0000 0000 0000 = FF800000

✓ X = FA390000 ÷ 48400000:

FA390000: 1111 1010 0011 1001 0000 0000 0000 0000
$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $Mantissa = 1.0111$
FA390000 = $-1.0111001 \times 2^{117}$

48400000: 0100 1000 0100 0000 0000 0000 0000 0000
$e + bias = 10010000 = 144 \rightarrow e = 144 - 127 = 17$ $Mantissa = 1.1$
48400000 = 1.1×2^{17}

$$X = -\frac{1.0111001 \times 2^{117}}{1.1 \times 2^{17}} = -\frac{1.0111001}{1.1} \times 2^{100}$$

```
          0000000011110110
11000000 )1011100100000000
           11000000
            101100100
             11000000
              101001000
               11000000
                100010000
                 11000000
                  101000000
                   11000000
                    100000000
                     11000000
                      10000000
```

Alignment:
$$\frac{1.0111001}{1.1} = \frac{1.0111001}{1.1000000} = \frac{10111001}{11000000}$$

Append $x = 8$ zeros: $\frac{10111001\mathbf{00000000}}{11000000}$

Integer division
$Q = 11110110 \rightarrow Qf = 0.1111011$

Thus: $X = -0.1111011 \times 2^{100} = -1.111011 \times 2^{99}$
$e + bias = 99 + 127 = 226 = 11100010$

X = 1111 0001 0111 0110 0000 0000 0000 0000 = F1760000

✓ X = FF800000 ÷ 09FE0090:

FF800000: 1111 1111 1000 0000 0000 0000 0000 0000
$e + bias = 11111111 = 255, f = 0$
FF800000 = $-\infty$

$X = -\infty \div \# = -\infty$
X = FF800000