Midterm Exam

(October 15th @ 7:30 pm)

Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (20 PTS)

- Compute the result of the following operations. The operands are signed fixed-point numbers. The result must be a signed fixed point number. For the division, use $x = 5$ fractional bits.

1.001001 + 1.001	1001.0101 − 1.010101	0.01001 + 01.11111
1.01101 × 01.011	1.011 × 1.0101	10.10010 ÷ 0.101

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

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

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

→

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

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

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

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

```
1.01101 x  →  0.10011 x  →      1 0 0 1 1 x
01.011        01.011              1 0 1 1
                              -----------
                                1 0 0 1 1
                              1 0 0 1 1
                            0 0 0 0 0
                          1 0 0 1 1
                      -------------------
                      0 1 1 0 1 0 0 0 1
                              ↓
                      0.1 1 0 1 0 0 0 1
                              ↓
                      1.0 0 1 0 1 1 1 1
```

```
1.011 x  →  0.101 x  →      1 0 1 1 x
1.0101      0.1011              1 0 1
                          -----------
                            1 0 1 1
                          0 0 0 0
                        1 0 1 1
                      ---------------
                      0 0 1 1 0 1 1 1
                            ↓
                      0.0 1 1 0 1 1 1
```

✓ $\frac{10.10010}{0.101}$: To unsigned (numerator) and then alignment, $a = 4$: $\frac{01.0111}{0.101} = \frac{01.0111}{00.1010} = \frac{010111}{001010} \equiv \frac{10111}{1010}$

```
        0001001001
1010 ) 1011100000
       1010
       ----
         1100
         1010
         ----
          10000
           1010
           ----
            110
```

Append $x = 5$ zeros: $\frac{10111\mathbf{00000}}{1010}$

Integer Division:

$Q = 1001001, R = 110$

$\rightarrow Qf = 10.01001 \; (x = 5)$

Final result (2C): $\frac{01.01110}{1.011} = 2C(010.01001) = 101.10111$

PROBLEM 2 (30 PTS)

- Calculate the result (provide the 32-bit result) of the following operations with single floating point numbers. Truncate the results when required. When doing fixed-point division, use $x = 4$ fractional bits.

✓ C2FA8000 + 40E00000	✓ D0DA8000 - 50FA8000	✓ 80400000 × FAB80000	✓ FB380000 ÷ 48C00000

✓ X = C2FA8000 + 40E00000:

C2FA8000: 1100 0010 1111 1010 1000 0000 0000 0000

$e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $Mantissa = 1.11110101$

C2FA8000 = -1.11110101×2^6

40E00000: 0100 0000 1110 0000 0000 0000 0000 0000

$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $Mantissa = 1.11$

40E00000 = 1.11×2^2

$$X = -1.11110101 \times 2^6 + 1.11 \times 2^2 = -1.11110101 \times 2^6 + \frac{1.11}{2^4} \times 2^6$$

$X = -(1.11110101 - 0.000111) \times 2^6$

We perform unsigned subtraction: $X = -1.11011001 \times 2^6$

```
b9=0 b8=0 b7=0 b6=0 b5=1 b4=1 b3=0 b2=0 b1=0 b0=0
 1.1 1 1 1 0 1 0 1 -
 0.0 0 0 1 1 1 0 0
 -----------------
 1.1 1 0 1 1 0 0 1
```

$X = -1.11011001 \times 2^6$, $e + bias = 6 + 127 = 133 = 10000101$

X = 1100 0010 1110 1100 1000 0000 0000 0000 = C2EC8000

✓ X = D0DA8000 - 50FA8000:

D0DAD000: 1101 0000 1101 1010 1000 0000 0000 0000

$e + bias = 10100001 = 161 \rightarrow e = 161 - 127 = 34$ $Mantissa = 1.10110101101$

D0DA8000 = $-1.10110101 \times 2^{34}$

50FA8000: 1101 0000 1111 1010 1000 0000 0000 0000

$e + bias = 10100001 = 161 \rightarrow e = 161 - 127 = 34$ $Mantissa = 1.11110101101$

50FA8000 = 1.11110101×2^{34}

$X = -1.10110101 \times 2^{34} - 1.11110101 \times 2^{34}$ (unsigned addition)

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

$X = -11.10101010 \times 2^{34} = -1.110101010 \times 2^{35}$

$e + bias = 35 + 127 = 162 = 10100010$

X = 1101 0001 0110 1010 1000 0000 0000 0000 = D16A8000

✓ X = 80400000 × FAB80000:

80400000: 1000 0000 0100 0000 0000 0000 0000 0000

$e + bias = 00000000 = 0 \rightarrow Denormal\ number \ \rightarrow e = -126$ $Mantissa = 0.1$

80400000 = $= -0.1 \times 2^{-126}$

FAB80000: 1111 1010 1011 1000 0000 0000 0000 0000

$e + bias = 11110101 = 245 \rightarrow e = 245 - 127 = 118$ $Mantissa = 1.0111$

FAB80000 = -1.0111×2^{118}

$X = (-0.1 \times 2^{-126}) \times (-1.0111 \times 2^{118}) = 0.10111 \times 2^{-8} = 1.0111 \times 2^{-9}$

$e + bias = -9 + 127 = 118 = 01110110$

X = 0011 1011 0011 1000 0000 0000 0000 0000 = 3B380000

✓ X = FB380000 ÷ 48C00000:

FB380000: 1111 1011 0011 1000 0000 0000 0000 0000

$e + bias = 11110110 = 246 \rightarrow e = 246 - 127 = 119$ $Mantissa = 1.0111$

FB380000 = -1.0111×2^{119}

48C00000: 1100 1000 1100 0000 0000 0000 0000 0000

$e + bias = 10010001 = 145 \rightarrow e = 145 - 127 = 18$ $Mantissa = 1.1$

48C00000 = 1.1×2^{18}

$$X = -\frac{1.0111 \times 2^{119}}{1.1 \times 2^{18}} = -\frac{1.0111}{1.1} \times 2^{101}$$

```
         000001111
11000 ) 101110000
         11000
          101100
           11000
           101000
            11000
            100000
             11000
              1000
```

Alignment:

$$\frac{1.0111}{1.1} = \frac{1.0111}{1.1000} = \frac{10111}{11000}$$

Append $x = 4$ zeros: $\frac{10111\mathbf{0000}}{11000}$

Integer division

$$Q = 1111 \rightarrow Qf = 0.1111$$

Thus: $X = -0.1111 \times 2^{101} = -1.111 \times 2^{100}$, $e + bias = 100 + 127 = 227 = 11100011$

X = 1111 0001 1111 0000 0000 0000 0000 0000 = F1F00000

PROBLEM 3 (13 PTS)

- Convert the following signed fixed point numbers in format [12 8] to the dual fixed point format 12_8_4.

FX	A.CD	F.EA	8.CA	1.CA
DFX				

✓ A.CD:
1010.1100 1101 ⇒ To DFX 12_8_4 ($num0$): 001011001101 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 111110101100 = FAC

✓ F.EA:
1111.1110 1010 ⇒ To DFX 12_8_4 ($num0$): 011111101010 = 7EA

✓ 8.CA:
1000.1100 1010 ⇒ To DFX 12_8_4 ($num0$): 000011001010 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 111110001100 = F8C

✓ 1.CA:
0001.1100 1010 ⇒ To DFX 12_8_4 ($num0$): 000111001010 = 1CA

PROBLEM 4 (22 PTS)

- Calculate the result of the following operations where the numbers are represented in dual fixed-point arithmetic. Note that the results must be in the same format. Include an overflow bit when necessary.

DFX Format 12_6_4	Result	Overflow		Result	overflow
C2A + C0B			FBA-073		
ACD + B98			F33-4BF		

✓ C2A + C0B:

```
                      1 1 0 0 0 0 0 0 1 0 1 0 0
110000101010 +    ➡    1|1 0 0 0 0 1 0.1 0 1 0 +
110000001011           1|1 0 0 0 0 0 0.1 0 1 1
                       ------------------------
                       1 0 0 0 0 0 1 1.0 1 0 1
```

10000011.0101 ⇒ To DFX 12_6_4 ($num0$): 000011.010100 = not a $num0$!
⇒ To DFX 12_6_4 ($num1$): 10000011.0101 = 835 → not a $num1$! Overflow = 1

✓ ACD + B98:

```
                      0 0 1 1 1 0 0 1 1 0 0 0 0
101011001101 +    ➡    0|0 1 0 1 1 0 0.1 1 0 1 +
101110011000           0|0 1 1 1 0 0 1.1 0 0 0
                       ------------------------
                       0 1 1 0 0 1 1 0.0 1 0 1
```

01100110.0101 ⇒ To DFX 12_6_4 ($num0$): 000110.010100 = not a $num0$!
⇒ To DFX 12_6_4 ($num1$): 11100110.0101 = E65 → not a $num1$! Overflow = 1

✓ FBA − 073:

111110111010 −
000001110011

➡ 1111011.1010 −
0000001.1100~~11~~
optional to keep

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

1111001.1110 ⇒ To DFX 12_6_4 ($num0$): 011001.111000 = 678   Overflow = 0

✓ F33 − 4BF:

111100110011 −
010010111111

➡ 1110011.0011 −
1110010.1111~~11~~
optional to keep

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

000000.0100 ⇒ To DFX 12_6_4 ($num0$): 000000.010000 = 010   Overflow = 0

PROBLEM 5 (15 PTS)

- Complete the timing diagram of the following iterative unsigned multiplier ($N = 4, M = 4$).
  Register: $sclr$: synchronous clear. Here, if $sclr = E = 1$, the register contents are initialized to 0.
  Parallel access shift registers: If $E = 1$: $s\_l = 1 \to$ Load, $s\_l = 0 \to$ Shift