Solutions - Midterm Exam

(October 16th @ 5:30 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.

1.0111 + 1.101001	1.010101 − 1000.0101	01.11111 + 0.10001
10.101 × 1.01101	1.001 × 0.1011	10.1010 ÷ 0.101

```
c8=1 c7=1 c6=1 c5=1 c4=0 c3=0 c2=0 c1=0 c0=0
  1.0 1 1 1 0 0 +
  1.1 0 1 0 0 1
  -------------
  1.0 0 0 1 0 1
```

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

=>

c10=1 c9=1 c8=1 c7=1 c6=1 c5=1 c4=1 c3=1 c2=0 c1=0 c0=0
  1 1 1 1.0 1 0 1 0 1 +
  0 1 1 1.1 0 1 1 0 0
  -------------------
  0 1 1 1.0 0 0 0 0 1
```

```
c8=0 c7=0 c6=1 c5=1 c4=1 c3=1 c2=1 c1=1 c0=0
  0 0 1.1 1 1 1 1 +
  0 0 0.1 0 0 0 1
  ---------------
  0 1 0.1 0 0 0 0
```

```
10.101  x  =>  01.011  x  =>        1 0 0 1 1 x
1.01101        0.10011                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
                                    |
                                    v
                            0.1 1 0 1 0 0 0 1
```

```
1.001  x  =>  0.111  x  =>        1 0 1 1 x
0.1011        0.1011                1 1 1
                                ---------
                                  1 0 1 1
                                1 0 1 1
                              1 0 1 1
                            ---------------
                            0 1 0 0 1 1 0 1
                                  |
                                  v
                            0.1 0 0 1 1 0 1
                                  |
                                  v
                            1.0 1 1 0 0 1 1
```

✓ $\frac{10.1010}{0.101}$: To unsigned (numerator) and then alignment, $a = 3$: $\frac{01.0110}{0.101} = \frac{01.011}{00.101} \equiv \frac{1011}{101}$

```
        001000110
101 ) 101100000
      101
      ---
        01000
          101
          ---
           110
           101
           ---
            10
```

Append $x = 5$ zeros: $\frac{101\mathbf{100000}}{101}$

Integer Division:

$$Q = 1000110, R = 10$$
$$\rightarrow Qf = 10.00110\ (x = 5)$$

Final result (2C): $\frac{10.1010}{0.101} = 2C(010.0011) = 101.1101$

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.

✓ C0D00000 + 42EA0000	✓ 50A90000 - 4F480000	✓ 80400000 × 7AB80000	✓ FB380000 ÷ 48C00000

✓ X = C0D00000 + 42EA0000:

C0D00000: 1100 0000 1101 0000 1000 0000 0000 0000

$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $\quad Mantissa = 1.101$

C0D00000 = -1.101×2^2

42EA0000: 0100 0010 1110 1010 0000 0000 0000 0000

$e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ $\quad Mantissa = 1.110101$

42EA0000 = 1.110101×2^6

$$X = -1.101 \times 2^2 + 1.110101 \times 2^6 = -\frac{1.101}{2^4} \times 2^6 + 1.110101 \times 2^6 = (-0.0001101 + 1.110101) \times 2^6$$

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

$R = 01.110101 - 0.0001101 = 01.110101 + 1.1110011$

The result in 2C is: $R = 01.1011101$

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

$\Rightarrow R(SM) = +1.1011101$

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

$X = 1.1011101 \times 2^6$, $e + bias = 6 + 127 = 133 = 10000101$

X = 0100 0010 1101 1101 0000 0000 0000 0000 = 42DD0000

✓ X = 50A90000 – 4F480000:

50A90000: 0101 0000 1010 1001 0000 0000 0000 0000

$e + bias = 10100001 = 161 \rightarrow e = 161 - 127 = 34$ $\quad Mantissa = 1.0101001$

50A90000 = 1.0101001×2^{34}

4F480000: 0100 1111 0100 1000 0000 0000 0000 0000

$e + bias = 10011110 = 1581 \rightarrow e = 158 - 127 = 31$ $\quad Mantissa = 1.1001$

4F480000 = 1.1001×2^{31}

$$X = 1.0101001 \times 2^{34} - 1.1001 \times 2^{31} = 1.0101001 \times 2^{34} - \frac{1.1001}{2^3} \times 2^{34}$$

$X = (1.0101001 - 0.0011001) \times 2^{34}$ (unsigned subtraction)

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

$X = 1.001 \times 2^{34}$, $e + bias = 34 + 127 = 161 = 10100001$

X = 0101 0000 1001 0000 0000 0000 0000 0000 = 50900000

✓ X = 80400000 × 7AB80000:

80400000: 1000 0000 0100 0000 0000 0000 0000 0000

$e + bias = 00000000 = 3 \rightarrow Denormal\ number \rightarrow e = -126$ $\quad Mantissa = 0.1$

80400000 = -0.1×2^{-126}

7AB80000: 0111 1010 1011 1000 0000 0000 0000 0000

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

7AB80000 = 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 = 1011 1011 0011 1000 0000 0000 0000 0000 = BB380000

✓ X = FB380000 ÷ 48C00000:

FB380000: 1111 1011 0011 1000 0000 0000 0000 0000

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

FB380000 = -1.0111×2^{119}

48C00000: 0100 1000 1100 0000 0000 0000 0000 0000

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

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: $\frac{101110000}{11000}$

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.D7	F.AE	7.1F	8.C4
DFX				

✓ A.D7:
1010.1101 0111 ⇒ To DFX 12_8_4 ($num0$): 001011010111 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 111110101101 = FAD

✓ F.AE:
1111.1010 1110 ⇒ To DFX 12_8_4 ($num0$): 011110101110 = 7AE

✓ 7.1F:
0111.0001 1111 ⇒ To DFX 12_8_4 ($num0$): 011100011111 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 100001110001 = 871

✓ 8.C4:
1000.1100 0100 ⇒ To DFX 12_8_4 ($num0$): 000011000100 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 111110001100 = F8C

PROBLEM 4 (22 PTS)

- Calculate the result of the following operations where the numbers are represented in dual fixed-point arithmetic (12_8_4). Note that the results must be in the same format. Include an overflow bit when necessary.

DFX Format 12_8_4	Result	Overflow		Result	overflow
FAC + 7EE			10A-C0A		
40B + 78B			999-674		

✓ FAC + 7EE:

```
                      1 1 1 1 1 1 1 1 1 0 0 0
111110101100 +          1 1 1 1 0 1 0.1 1 0 0 +
011111101110   ➡        1 1 1 1 1 1 1.1 1 1 0 1 1 1 0
                                                       optional
                        1 1 1 1 0 1 0.1 0 1 0 1 1 1 0
```

1010.10101110 ⇒ To DFX 12_8_4 ($num0$): 0010.10101110 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 11111010.1010 = FAA Overflow = 0

✓ 40B + 78B:

```
                            1 1 0 0 0 0 0 0 1 0 1 1 0
010000001011 +              1|1 0 0.0 0 0 0 1 0 1 1 +
011110001011          ➡     1|1 1 1.1 0 0 0 1 0 1 1
                            -------------------------
                            1 0 1 1.1 0 0 1 0 1 1 0
```

1011.10010110 ⇒ To DFX 12_8_4 ($num0$): 0011.10010110 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 11111011.1001 = FB9 Overflow = 0

✓ 10A – C0A:

```
                          optional to keep      0 0 1 1 1 1 1 1 0 0 0 0 0
000100001010 -            001.00001010 -        0|0 0 0 0 0 0 1.0 0 0 0 +
110000001010       ➡    1000000.1010      ➡     0|0 1 1 1 1 1 1.0 1 1 0
                                                -------------------------
                                                0 1 0 0 0 0 0 0.0 1 1 0
```

01000000.0110 ⇒ To DFX 12_8_4 ($num0$): 0000.01100000 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 11000000.0110 = 806 → not a $num1$! Overflow = 1

✓ 999 – 674:

```
                                          0 0 0 0 0 0 1 1 0 0 0 0
100110011001 +       0011001.1001 -       0 0 1 1 0 0 1.1 0 0 1 +
011001110100    ➡    110.01110100    ➡    0 0 0 0 0 0 1.1 0 0 0 1 1 0 0
                                          -------------------------      optional
                                          0 0 1 1 0 1 1.0 0 0 1 1 1 0 0
```

011011.0001 ⇒ To DFX 12_8_4 ($num0$): 0011.00010000 = not a $num0$!
⇒ To DFX 12_8_4 ($num1$): 10011011.0001 = 9B1 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 register: If $E = 1$: $s\_l = 1 \rightarrow$ Load, $s\_l = 0 \rightarrow$ Shift