Solutions - Midterm Exam

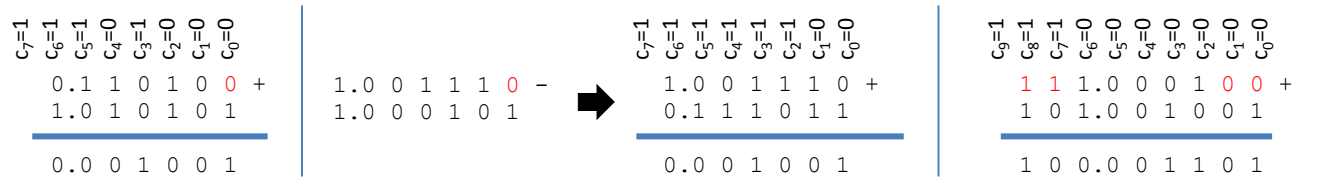
(October 19th @ 7:30 pm)

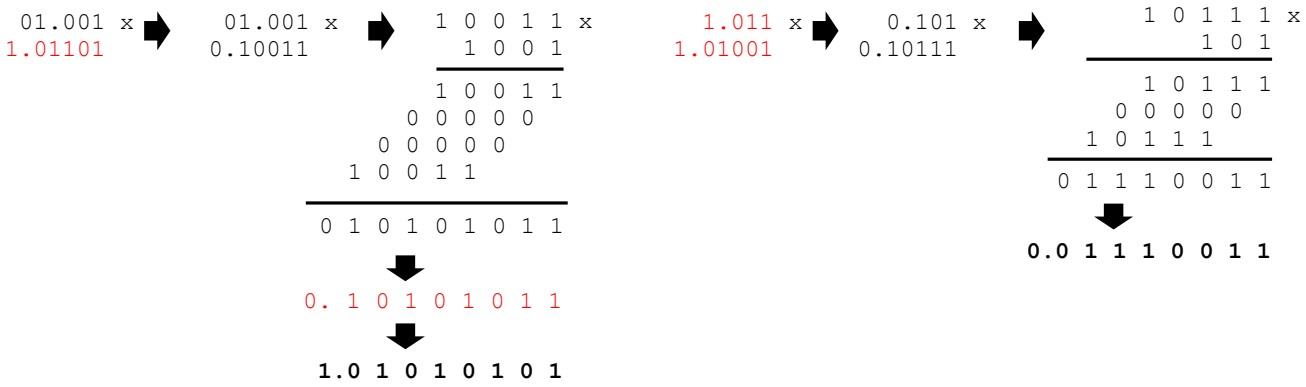
Presentation and clarity are very important! Show your procedure!

PROBLEM 1 (20 PTS)

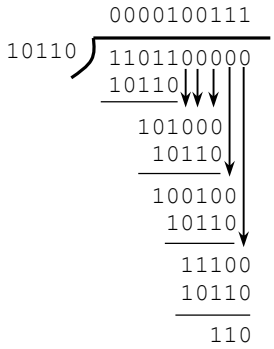
- Compute the result of the following operation with signed fixed-point numbers. For the division, use $x = 5$ fractional bits.

0.11010 + 1.010101	1.00111 - 1.000101	1.0001 + 101.001001
01.001 × 1.01101	1.011 × 1.01001	10.01010 ÷ 01.011





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



Append $x = 5$ zeros: $\frac{1101100000}{10110}$

Integer Division:

$Q = 100111, R = 110$

$\rightarrow Qf = 1.00111 (x = 5)$

Final result (2C): $\frac{01001.001}{10.101} = 2C(01.00111) = 10.11001$

PROBLEM 2 (30 PTS)

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

✓ 40B00000 + C2FA8000	✓ 10DAD000 – 90FAD000	✓ 7AB80000 × 81800000	✓ FA390000 ÷ 48400000

✓ X = 40B00000 + C2FA8000:

40B00000: 0100 0000 1011 0000 0000 0000 0000 0000

$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ $\quad Mantissa = 1.011$

40B00000 = 1.011×2^2

C2FA8000: 1100 0010 1111 1010 1000 0000 0000 0000

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

C2FA8000 = -1.11110101×2^6

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

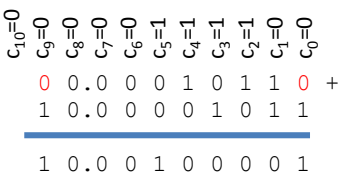
To subtract these unsigned numbers, we first convert to 2C:

$R = 0.0001011 - 01.11110101 = 0.0001011 + 10.00001011$

The result in 2C is: $R = 10.00100001$, $-R = 01.11011111$

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

$\Rightarrow R(SM) = -1.11011111$



$X = -1.11011111 \times 2^6$, $e + bias = 6 + 127 = 133 = 10000101$

X = 1100 0010 1110 1111 1000 0000 0000 0000 = C2EF8000

✓ X = 10DAD000 – 90FAD000:

10DAD000: 0001 0000 1101 1010 1101 0000 0000 0000

$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$ $\quad Mantissa = 1.10110101101$

10DAD000 = $1.10110101101 \times 2^{-94}$

90FAD000: 1001 0000 1111 1010 1101 0000 0000 0000

$e + bias = 00100001 = 33 \rightarrow e = 33 - 127 = -94$ $\quad Mantissa = 1.11110101101$

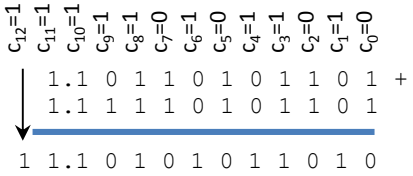
90FAD000 = $-1.11110101101 \times 2^{-94}$

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

$X = 11.1010101101 \times 2^{-94} = 1.11010101101 \times 2^{-93}$

$e + bias = -93 + 127 = 34 = 00100010$

X = 0001 0001 0110 1010 1101 0000 0000 0000 = 116AD000



✓ X = 7AB80000 × 81800000:

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}

81800000: 1000 0001 1000 0000 0000 0000 0000 0000

$e + bias = 00000011 = 3 \rightarrow e = 3 - 127 = -124$ $\quad Mantissa = 1.0$

90FAD000 = -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 = FA390000 ÷ 48400000:

FA390000: 1111 1010 0011 1001 0000 0000 0000 0000

$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117$ $\quad Mantissa = 1.0111$

FA380000 = $-1.0111001 \times 2^{117}$

48400000: 0100 1000 0100 0000 0000 0000 0000 0000

$e + bias = 10010000 = 144 \rightarrow e = 144 - 127 = 17$ $\quad Mantissa = 1.1$

48400000 = 1.1×2^{17}

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

```
          000000001111
11000000 )101110010000
           11000000
           101100100
            11000000
            101001000
             11000000
             100010000
              11000000
               1010000
```

Alignment:

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

Append $x = 4$ zeros: $\frac{10111001\mathbf{0000}}{11000000}$

Integer division

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

Thus: $X = -0.1111 \times 2^{100} = -1.111 \times 2^{99}$
$e + bias = 99 + 127 = 226 = 11100010$

X = 1111 0001 0111 0000 0000 0000 0000 0000 = F1700000

PROBLEM 3 (20PTS)

- 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
F2A + 0A9			F99-092		
C00 + F13			F33-6A9		

✓ FA2+0A9:

```
111100101010 +        ➡  1 1 1 0 0 1 0.1 0 1 0 +
000010101001             0 0 0 0 0 1 0.1 0 1 0 0 1 (struck out: 0 1)
                         ---------------------------
                         1 1 1 0 1 0 1.0 1 0 0
```

1110101.0100 ⇒ To DFX 12_6_4 ($num0$): 010101010000 = 550 Overflow = 0

✓ C00+F13:

```
110000000000 +        ➡  1 1 0 0 0 0 0 0.0 0 0 0 +
111100010011             1 1 1 1 0 0 0 1.0 0 1 1
                         -------------------------
                         1 0 1 1 0 0 0 1.0 0 1 1
```

10110001.0011 ⇒ To DFX 12_6_4 ($num0$): 010001001100 = not a $num0$!
⇒ To DFX 12_6_4 ($num1$): 101100010011 = not a $num1$! Overflow = 1

✓ F99-092:

```
111110011001 -   ➡  1111001.1001 -    ➡  1 1 1 1 0 0 1.1 0 0 1 +
000010010010        0000010.010010       1 1 1 1 1 0 1.1 0 1 1 1 0 (struck out: 1 0)
                                         ---------------------------
                                         1 1 1 0 1 1 1.0 1 0 0
```

1110111.0100 ⇒ To DFX 12_6_4 ($num0$): 010111010000 = 5D0 Overflow = 0

✓ F33-6A9:

```
111100110011 -   ➡  1110011.0011 -    ➡  1 1 1 0 0 1 1.0 0 1 1 +
011010101001        1111010.101001       0 0 0 0 1 0 1.0 1 0 1 1 1 (struck out: 1 1)
                                         ---------------------------
                                         1 1 1 1 0 0 0.1 0 0 0
```

1111000.1000 ⇒ To DFX 12_6_4 ($num0$): 011000100000 = 620 Overflow = 0

PROBLEM 4 (15 PTS)

- Calculate the square root (in binary) of the following unsigned number. Use $x = 2$ extra precision bits for your answer.

$$Df = 1011.011001$$

$Df = 1011.011001 = 11.390625, p = 3, n = 5$. Format $[10\ 6]$.
Qf format: $[n + x\ p + x] = [7\ 5]$. $x = 2$: extra precision bits.

Step 1: Get the integer D.
$\Rightarrow D = 1011011001$

Step 2: Add (optionally) $2x = 4$ zeros
$\Rightarrow Dp = 10110110010000 = 11664$

Step 3: Get $Qp = \sqrt{Dp}$

Then: $Dp = 10110110010000 = 11664$. Then $nq = n + x = 5 + 2 = 7$
$k = 6$: $q_6 = 1$ $(Q = 1000000)$. $11664 < 64^2$? No
$k = 5$: $q_5 = 1$ $(Q = 1100000)$. $11664 < 96^2$? No
$k = 4$: $q_4 = 1$ $(Q = 1110000)$. $11664 < 112^2$? $Yes \rightarrow q_4 = 0$ $(Q = 1100000)$
$k = 3$: $q_3 = 1$ $(Q = 1101000)$. $11664 < 104^2$? No
$k = 2$: $q_2 = 1$ $(Q = 1101100)$. $11664 < 108^2$? No . $Note\ that\ 11664 = 108^2, we\ could\ have\ stopped\ here$
$k = 1$: $q_1 = 1$ $(Q = 1101110)$. $11664 < 110^2$? $Yes \rightarrow q_1 = 0$ $(Q = 1101100)$
$k = 0$: $q_0 = 1$ $(Q = 1101101)$. $11664 < 109^2$? $Yes \rightarrow q_0 = 0$ $(Q = 1101100)$
Result: $Qp = 1101100, Rp = Dp - Qp^2 = 00000000$

Final Result $(p + x = 5)$: $Qf = 11.01100 = 3.375 = \sqrt{11.390625}$

PROBLEM 5 (15 PTS)

- Complete the following 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