ELEC8550: Computer Arithmetic, Fall 2020

Solution to Assignment 3

updated on Nov 24

Requirements:

- Organize your assignment solution into one single file in Microsoft Word format or PDF. Note that image format (incl. jpg etc) is not acceptable, however, you may convert jpg files to one single pdf file and then submit it.
- Submission must be conducted on-line through Blackboard website.

1. (12 marks)

Solution:

A		1	0	0	1				-7
X	$\times$	1	0	1	0				-6
$P^{(0)} = 0$		0	0	0	0				
$x_0 = 0 \Rightarrow$ Shift		0	0	0	0	0			
$x_1 = 1 \Rightarrow$ Add A	$+$	1	0	0	1	0			
		1	0	0	1	0			
Shift		1	1	0	0	1	0		
$x_2 = 0 \Rightarrow$ Shift		1	1	1	0	0	1	0	
$x_3 = 1 \Rightarrow$ Correction $($Add $-A)$	$+$	0	1	1	1				
		0	1	0	1	0	1	0	42

□

2. (12 marks)

Solution:

This problem is a little bit more advanced and needs a step of further research investigation.

Since $X > D$, shift the dividend X to the right by one bit to have $X' = X/2 = 0.011000$

such that $X' < D$. We perform the division operation $X' \div D$ as follows:

$r_0 = X'$			0	.0	1	1	0	0	0	
$2r_0$		0	0	.1	1	0	0	0		
Add $-D$	+	1	1	.0	1	1				
$2r_0 - D \geq 0$		0	0	.0	0	1	0	0		set $q_1 = 1$
$r_1 = 2r_0 - D$ and $2r_1$		0	0	.0	1	0	0			
Add $-D$	+	1	1	.0	1	1				
$2r_1 - D < 0$		1	1	.1	0	1	0			set $q_2 = 0$
$r_2 = 2r_1$		0	0	.0	1	0	0			
$2r_2$		0	0	.1	0	0	0			
Add $-D$	+	1	1	.0	1	1				
$2r_2 - D < 0$		1	1	.1	1	1				set $q_3 = 0$
$r_3 = 2r_2$		0	0	.1	0	0				

The final results are $Q = (0.100)_2 = 1/2$ and $r_3 = 1/2 \Rightarrow R = r_m 2^{-m} = r_3 2^{-3} = 1/16$. Note that $X' = 0.011_2 = 3/8$ and $D = 0.101_2 = 5/8$. we can verify the results as follows:

$$D \times Q + R = \frac{5}{8} \times \frac{1}{2} + \frac{1}{16} = \frac{6}{16} = \frac{3}{8} = X'$$

Note that the above results Q and R are obtained for $X' \div D = (X/2) \div D$. Certain adjustment step is needed for obtaining the quotient and remainder for $X \div D$.

Discussion on the adjustment step for the case $X > D$ (Without a discussion part similar to the following, there are maximal 6 marks):

- If the division operation is performed for floating-point numbers, then we can safely assume that X and D are the significand part of two floating-point numbers. Then applying the operation of shift-to-right by k-bit is equivalent to add k to exponent part (assuming the base is 2).
- If the division operation is performed for fixed-point numbers, then the scenario is more complicated and some further discussion can be given as follows:

  $$X = 2 \times X' = 2(Q \times D + R) = (2Q) \times D + 2R.$$

  It can be seen that the quotient needs to shift to left by one bit and further adjustment on the remainder is also needed.

□

3. (16 marks)

Solution:

$$100_{10} \times 2^6 = (2^6 + 2^5 + 2^2) \times 2^6 = 1100100_2 \times 2^6 = 1.1001 \times 2^{12}.$$

i).

The sign bit $s = 1$.
The exponent field $E = 12 + 127 = 2^7 + 2^3 + 2 + 1 = 1000\ 1011$.
The significand $f = \underbrace{1001000 \cdots 00}_{\text{23 bits}}$.

The format of IEEE single-precision is
$\boxed{s \mid E \mid f}$.

ii).

The sign bit $s = 1$.
The exponent field $E = 12 + 1023 = 2^{10} + 2^3 + 2 + 1 = 100\ 0000\ 1011$.
The significand $f = \underbrace{1001000 \cdots 00}_{\text{52 bits}}$.

The format of IEEE double-precision is $\boxed{s \mid E \mid f}$.

□

4. (16 marks)

Solution:

The sign bit $s = 1$.
The exponent field $E = 11000010 = 194_{10}$. So $E_{\text{true}} = E - 127 = 67_{10}$.
The significand $0.f = 0.0111 = 0.4375_{10}$.

$$F = (-1)^s \times 1.f \times 2^{E-127} = -1.4375 \times 2^{67}.$$

□

5. (16 marks)

Solution:

$x_1x_0.x_{-1}x_{-2}$	$Y = \text{round}(X)$
$\times.00$	$\times$
$\times.01$	$\times$
$\times.10$	$\times + 1$
$\times.11$	$\times + 1$

Table 1: Truth table

The output Y can be given as $Y = x_1x_0 + x_{-1}$. A block diagram for this rounding scheme can be easily drawn that uses one 2-bit adder with inputs of x_1x_0 and x_{-1}. □

6. (12 marks)

Solution:

- Truth table:

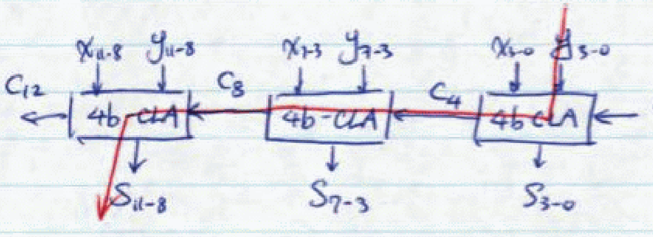
ROM scheme with $\ell = 3$ and $d = 2$		
Input:	Output:	Error:
x	ROM(x)	ROM$(x) - x$
$\times 00.00$	$\times 00.$	0
$\times 00.01$	$\times 00.$	$-1/4$
$\times 00.10$	$\times 01.$	$+1/2$
$\times 00.11$	$\times 01.$	$+1/4$
$\times 01.00$	$\times 01.$	0
$\times 01.01$	$\times 01.$	$-1/4$
$\times 01.10$	$\times 10.$	$+1/2$
$\times 01.11$	$\times 10.$	$+1/4$
$\times 10.00$	$\times 10.$	0
$\times 10.01$	$\times 10.$	$-1/4$
$\times 10.10$	$\times 11.$	$+1/2$
$\times 10.11$	$\times 11.$	$+1/4$
$\times 11.00$	$\times 11.$	0
$\times 11.01$	$\times 11.$	$-1/4$
$\times 11.10$	$\times 11.$	$-1/2$
$\times 11.11$	$\times 11.$	$-3/4$

- Error and bias: It can be seen from the above table that the maximal error is $e^{-}_{\text{max}} = -3/4$.

$$\text{Bias} = \frac{1}{16}\Big(3 \times \frac{1}{2} + 3 \times \frac{1}{4} - 4 \times \frac{1}{4} - \frac{1}{2} - \frac{3}{4}\Big) = 0.$$

7. (16 marks)

Solution:

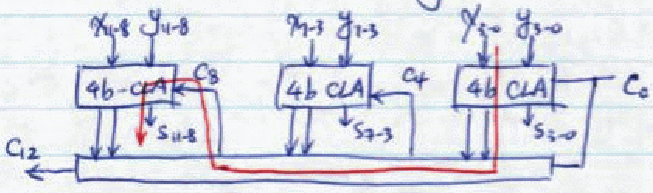


one-leve carry Lookahead.

One cp:

$x_0 \to C_4 \to C_8 \to S_{11}$

$$T_{cpp} = T_{x_0 \to C_4} + T_{C_4 \to C_8} + T_{C_8 \to S_{11}}$$
$$= 3\Delta_G + 2\Delta_G + 3\Delta_G$$
$$= 8\Delta_G$$



Two-Level Carry Lookahead.

One cp.

$x_0 \to G_{0,3} \to C_8 \to S_{11}$

$$T_{cpp} = T_{x_0 \to G_{0,3}} + T_{G_{0,3} \to C_8} + T_{C_8 \to S_{11}}$$
$$= 3\Delta_G + 2\Delta_G + 3\Delta_G$$
$$= 8\Delta_G$$