Homework 2

Madiba Hudson-Quansah

Question 1

What are the unsigned and signed decimal values of the following binary and hexadecimal numbers?

1. 10110110
2. C1B3

Solution:

1. Unsigned

$$\begin{aligned}10110110 &= 1 \times 2^7 + 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^2 + 1 \times 2^1 \\ &= 182\end{aligned}$$

Signed

$$\begin{aligned}10110110 &= 01001001 + 00000001 \\ &= 01001010 \\ &= 1 \times 2^6 + 1 \times 2^3 + 1 \times 2^1 \\ &= -74\end{aligned}$$

2. Unsigned

$$\begin{aligned}C1B3 &= 12 \times 16^3 + 1 \times 16^2 + 11 \times 16^1 + 3 \times 16^0 \\ &= 49587\end{aligned}$$

Signed

$$\begin{aligned}C1B3 &= 1100000110110011 \\ &= 0011111001001100 + 1 \\ &= 0011111001001111 \\ &= 3D4F\end{aligned}$$

Question 2

Carry out the following additions. Indicate whether there is a carry or overflow.

1. 11010010 (binary) + 10111101 (binary)
2. A1CF (hexadecimal) + B2D3 (hexadecimal)

Solution:

1.

$$\begin{array}{r} 11010010 \\ \underline{+10111101} \\ 110001111 \end{array}$$

Overflow: yes as the result is 9 bits
Carry: yes as there is a carry out of the MSB

(a)

$$\begin{array}{r} 1010000111001111 \\ \underline{+1011001011010011} \\ 10111010100100010 \end{array}$$

Overflow: yes as the result is 17 bits
Carry: yes as there is a carry out of the MSB

Question 3

Carry out the following subtractions. Indicate whether there is a borrow or overflow.

1. 11010010 (binary) - 10111101 (binary)
2. 71CF (hexadecimal) - B2D3 (hexadecimal)

Solution:

1.

$$\begin{array}{r} 11010010 \\ \underline{-10111101} \\ 11010010 \\ \underline{+01000011} \\ 00010101 \end{array}$$

Borrow: yes as there is a borrow out of the MSB
Overflow: No as the result fits in 8 bits

2.

$$\begin{array}{r} 71CF \\ \underline{-B2D3} \\ 29135 \\ \underline{-45795} \\ -16660 \\ -4104 \end{array}$$

Question 4

What is the decimal value of the following single-precision floating-point numbers?

1. 1 01011010 001 0100 0000 0000 0000 0000 (binary)
2. 0100 0110 1100 1000 0000 0000 0000 0000 (binary)

Solution:

1.

$$\begin{aligned} S &= 1 \\ F &= 1 + 1 \times 2^{-3} + 1 \times 2^{-5} \\ E &= 90 - 127 \\ &= -37 \\ D &= (-1)^S \times F \times 2^E \\ &= -1 \times 1.125 \times 2^{-37} \\ &= -8.412825991399586e^{-12} \end{aligned}$$

2.

$$
\begin{aligned}
S &= 0 \\
F &= 1 + 1 \times 2^{-1} + 1 \times 2^{-4} \\
E &= 141 - 127 \\
&= 14 \\
D &= (-1)^S \times F \times 2^E \\
&= 1 \times 1.5625 \times 2^{14} \\
&= 25600.0
\end{aligned}
$$

Question 5

Show the IEEE 754 binary representation for: -95.4 in:

1. Single Precision
2. Double precision

Solution:

$$
\begin{aligned}
S &= 1 \\
F &= 0.4 \times 2 = 0.8 \\
&= 0.8 \times 2 = 1.6 \\
&= 0.6 \times 2 = 1.2 \\
&= 0.2 \times 2 = 0.4 \\
&= 0.4 \times 2 = 0.8 \\
&= 0.8 \times 2 = 1.6 \\
&= 0.6 \times 2 = 1.2 \\
&= 0.2 \times 2 = 0.4 \\
&= 0.4 \times 2 = 0.8 \\
&= 0.8 \times 2 = 1.6 \\
&= 0.6 \times 2 = 1.2 \\
&= 0.2 \times 2 = 0.4 \\
&= 0.4 \times 2 = 0.8 \\
&= 0.8 \times 2 = 1.6 \\
&= 0.6 \times 2 = 1.2 \\
&= 0.2 \times 2 = 0.4 \\
&= 0.4 \times 2 = 0.8 \\
&= 0.8 \times 2 = 1.6 \\
&= 0.6 \times 2 = 1.2 \\
&= 0.2 \times 2 = 0.4 \\
&= 0.4 \times 2 = 0.8 \\
&= 0.8 \times 2 = 1.6 \\
&= 0.6 \times 2 = 1.2 \\
E &= 95 \div 2 = 47 \text{ rem } 1 \\
&= 47 \div 2 = 23 \text{ rem } 1 \\
&= 23 \div 2 = 11 \text{ rem } 1 \\
&= 11 \div 2 = 5 \text{ rem } 1 \\
&= 5 \div 2 = 2 \text{ rem } 1 \\
&= 2 \div 2 = 1 \text{ rem } 0 \\
&= 1 \div 2 = 0 \text{ rem } 1 \\
\text{Un-normalized} &= 1011111.01100110011001100110011 \\
\text{Normalized} &= 1.01111101100110011001101 && \text{Rounded}
\end{aligned}
$$

1.

$$
\begin{aligned}
E &= 6 + 127 \\
&= 133
\end{aligned}
$$

$$
\underbrace{1}_{\text{Sign}} \quad \underbrace{10000101}_{\text{Exponent}} \quad \underbrace{01111101100110011001101}_{\text{Fraction}}
$$

2.

$$
\begin{aligned}
E &= 6 + 1023 \\
&= 1029
\end{aligned}
$$

$$\underbrace{1}_{\text{Sign}} \quad \underbrace{10000000101}_{\text{Exponent}} \quad \underbrace{0111110110011001100110011001100110011001100110011001}_{\text{Fraction}}$$

Question 6

Given the following numbers:
x = 1100 0110 1101 1000 0000 0000 0000 0000 (binary) and
y = 0011 1110 1110 0000 0000 0000 0000 0000 (binary) are single-precision floating-point numbers.
Perform the following operations showing all work:

1. x + y
2. x * y

Solution:

$$
\begin{aligned}
S &= 1 \\
F &= 1 + 1 \times 2^{-1} + 1 \times 2^{-3} + 1 \times 2^{-4} \\
E &= 141 - 127 \\
&= 14 \\
D &= (-1)^S \times F \times 2^E \\
&= -1 \times 1.6875 \times 2^{14} \\
&= -27648.0
\end{aligned}
$$

$$
\begin{aligned}
S &= 0 \\
F &= 1 + 1 \times 2^{-1} + 1 \times 2^{-2} \\
E &= 125 - 127 \\
&= -2 \\
D &= (-1)^S \times F \times 2^E \\
&= 1 \times 1.75 \times 2^{-2} \\
&= 0.4375
\end{aligned}
$$

1.

$$
\begin{array}{r}
11000110110110000000000000000000 \\
\underline{+00111110111000000000000000000000} \\
-27648.0 \\
\underline{+0.4375} \\
-27647.5625
\end{array}
$$

$$
\begin{aligned}
S &= 1 \\
F &= 0.5625 \times 2 = 1.125 \\
&= 0.125 \times 2 = 0.25 \\
&= 0.25 \times 2 = 0.5 \\
&= 0.5 \times 2 = 1.0 \\
E &= 27647 \div 2 = 13823 \text{ rem } 1 \\
&= 13823 \div 2 = 6911 \text{ rem } 1 \\
&= 6911 \div 2 = 3455 \text{ rem } 1 \\
&= 3455 \div 2 = 1727 \text{ rem } 1 \\
&= 1727 \div 2 = 863 \text{ rem } 1 \\
&= 863 \div 2 = 431 \text{ rem } 1 \\
&= 431 \div 2 = 215 \text{ rem } 1 \\
&= 215 \div 2 = 107 \text{ rem } 1 \\
&= 107 \div 2 = 53 \text{ rem } 1 \\
&= 53 \div 2 = 26 \text{ rem } 1 \\
&= 26 \div 2 = 13 \text{ rem } 0 \\
&= 13 \div 2 = 6 \text{ rem } 1 \\
&= 6 \div 2 = 3 \text{ rem } 0 \\
&= 3 \div 2 = 1 \text{ rem } 1 \\
&= 1 \div 2 = 0 \text{ rem } 1 \\
\text{Un-normalized} &= 110101111111111.1001 \\
\text{Normalized} &= 1.101011111111111001 \\
E &= 14 + 127 \\
&= 141
\end{aligned}
$$

$$
\underbrace{1}_{\text{Sign}} \quad \underbrace{10001101}_{\text{Exponent}} \quad \underbrace{101011111111111001}_{\text{Fraction}}
$$

2.

$$
\begin{array}{r}
11000110110110000000000000000000 \\
\underline{\times 00111110111000000000000000000000} \\
-27648.0 \\
\underline{\times 0.4375} \\
-12096
\end{array}
$$

$$
\begin{aligned}
S &= 1 \\
E &= 12096 \div 2 = 6048 \text{ rem } 0 \\
&= 6048 \div 2 = 3024 \text{ rem } 0 \\
&= 3024 \div 2 = 1512 \text{ rem } 0 \\
&= 1512 \div 2 = 756 \text{ rem } 0 \\
&= 756 \div 2 = 378 \text{ rem } 0 \\
&= 378 \div 2 = 189 \text{ rem } 0 \\
&= 189 \div 2 = 94 \text{ rem } 1 \\
&= 94 \div 2 = 47 \text{ rem } 0 \\
&= 47 \div 2 = 23 \text{ rem } 1 \\
&= 23 \div 2 = 11 \text{ rem } 1 \\
&= 11 \div 2 = 5 \text{ rem } 1 \\
&= 5 \div 2 = 2 \text{ rem } 1 \\
&= 2 \div 2 = 1 \text{ rem } 0 \\
&= 1 \div 2 = 0 \text{ rem } 1 \\
\text{Un-normalized} &= 10111101000000.0 \\
\text{Normalized} &= 1.01111010000000 \\
E &= 13 + 127 \\
&= 140
\end{aligned}
$$

$$
\underbrace{1}_{\text{Sign}} \quad \underbrace{10001100}_{\text{Exponent}} \quad \underbrace{01111010000000000000000}_{\text{Fraction}}
$$

Question 7

IA-32 offers an 80-bit extended precision option with a 1-bit sign, 16-bit exponent, and 63-bit fraction (64-bit significand including the implied 1 before the binary point). Assume that extended precision is similar to single and double precision.

1. What is the bias in the exponent?
2. What is the range (in absolute value) of normalized numbers that can be represented by the extended precision option?

Solution:

1. The bias for an exponent field of k bits is given by

$$2^{k-1} - 1.$$

I.e.:

$$2^{15} - 1 = 32768 - 1 = 32767.$$

2. For normalized numbers the exponent field e runs from 1 to $2^{16} - 2 = 65534$ (since 0 and all 1's are reserved). Therefore, the true exponent $E = e - 32767$ varies from:

$$E_{\text{min}} = 1 - 32767 = -32766 \quad \text{to} \quad E_{\text{max}} = 65534 - 32767 = 32767.$$

Hence, the range of normalized numbers is from:

$$1.0 \times 2^{-32766} \quad \text{up to} \quad (2 - 2^{-63}) \times 2^{32767}.$$

Question 8

Using the refined division hardware, show the unsigned division of:

$$\text{Dividend} = 11011001 \quad \text{by} \quad \text{Divisor} = 00001010$$

The result of the division should be stored in the Remainder and Quotient registers.
(Eight iterations are required. Show your steps.)

Solution:

1. **Initialize:**
$$\text{Remainder } R = 0, \quad \text{Dividend bits: } 1\,1\,0\,1\,1\,0\,0\,1.$$

2. **Iteration 1:**
$$R \leftarrow (0 \ll 1) \,|\, 1 = 1.$$
$1 < 10, \Rightarrow q_7 = 0.$

3. **Iteration 2:**
$$R \leftarrow (1 \ll 1) \,|\, 1 = 3.$$
$3 < 10 \Rightarrow q_6 = 0.$

4. **Iteration 3:**
$$R \leftarrow (3 \ll 1) \,|\, 0 = 6.$$
$6 < 10 \Rightarrow q_5 = 0.$

5. **Iteration 4:**
$$R \leftarrow (6 \ll 1) \,|\, 1 = 13.$$
$13 \geq 10 \Rightarrow 13 - 10 = 3, \Rightarrow q_4 = 1.$

6. **Iteration 5:**
$$R \leftarrow (3 \ll 1) \,|\, 1 = 7.$$
$7 < 10 \Rightarrow q_3 = 0.$

7. **Iteration 6:**
$$R \leftarrow (7 \ll 1) \,|\, 0 = 14.$$
$14 \geq 10 \Rightarrow 14 - 10 = 4, \Rightarrow q_2 = 1.$

8. **Iteration 7:**
$$R \leftarrow (4 \ll 1) \,|\, 0 = 8.$$
$8 < 10 \Rightarrow q_1 = 0.$

9. **Iteration 8:**
$$R \leftarrow (8 \ll 1) \,|\, 1 = 17.$$
$17 \geq 10 \Rightarrow 17 - 10 = 7, \Rightarrow q_0 = 1.$

   Final Registers:

- Quotient bits (from q_7 to q_0): $0\,0\,0\,1\,0\,1\,0\,1 = 00010101_2$ (which is 21_{10}).
- Remainder: 7 (or 00000111_2).

Question 9

Using the refined signed multiplication algorithm, show the multiplication of:

$$\text{Multiplicand} = 00101101 \quad \text{by} \quad \text{Multiplier} = 11010110 \quad \text{(signed)}$$

The multiplication result should be a 16-bit signed number stored in the HI and LO registers.
(Eight iterations are required because there are 8 bits in the multiplier. Show your steps.)

Solution: Define registers:

$$A \text{ (Accumulator, 8 bits)}, \quad Q \text{ (Multiplier, 8 bits)}, \quad Q_{-1} \text{ (1 bit)}, \quad \text{and } M \text{ (Multiplicand, 8 bits)}.$$

Compute $-M$:

$$M = 00101101_2, \quad -M = 11010011_2.$$

Initialize:

$$A = 00000000, \quad Q = 11010110, \quad Q_{-1} = 0.$$

1. **Iteration 1:** Look at $(Q_0, Q_{-1}) = (0, 0)$. $\longrightarrow$ No addition/subtraction.
   Perform arithmetic right shift on $[A, Q, Q_{-1}]$:

   $$A\,Q\,Q_{-1}: \quad 00000000\;11010110\;0 \quad \longrightarrow \quad 00000000\;01101011\;0.$$

2. **Iteration 2:** Now, $(Q_0, Q_{-1}) = (1, 0)$. $\longrightarrow$ Subtract M:

   $$A \leftarrow A - M = 00000000 - 00101101 = 11010011.$$

   Then, arithmetic right shift:

   $$11010011\;01101011\;0 \quad \longrightarrow \quad A = 11101001, \quad Q = 10110101, \quad Q_{-1} = 1.$$

3. **Iteration 3:** Now, $(Q_0, Q_{-1}) = (1, 1)$. $\longrightarrow$ No operation.
   Shift:

   $$11101001\;10110101\;1 \quad \longrightarrow \quad A = 11110100, \quad Q = 11011010, \quad Q_{-1} = 1.$$

4. **Iteration 4:** $(Q_0, Q_{-1}) = (0, 1)$. $\longrightarrow$ Add M:

   $$A \leftarrow A + M = 11110100 + 00101101 = 00100001 \quad \text{(with overflow discarded)}.$$

   Shift:

   $$00100001\;11011010\;1 \quad \longrightarrow \quad A = 00010000, \quad Q = 11101101, \quad Q_{-1} = 0.$$

5. **Iteration 5:** $(Q_0, Q_{-1}) = (1, 0)$. $\longrightarrow$ Subtract M:

   $$A \leftarrow 00010000 - 00101101 = 11100011.$$

   Shift:

   $$11100011\;11101101\;0 \quad \longrightarrow \quad A = 11110001, \quad Q = 11110110, \quad Q_{-1} = 1.$$

6. **Iteration 6:** $(Q_0, Q_{-1}) = (0, 1)$. $\longrightarrow$ Add M:

   $$A \leftarrow 11110001 + 00101101 = 00011110.$$

   Shift:

   $$00011110\;11110110\;1 \quad \longrightarrow \quad A = 00001111, \quad Q = 01111011, \quad Q_{-1} = 0.$$

7. **Iteration 7:** $(Q_0, Q_{-1}) = (1, 0)$. $\longrightarrow$ Subtract M:

   $$A \leftarrow 00001111 - 00101101 = 11100010.$$

   Shift:

   $$11100010\;01111011\;0 \quad \longrightarrow \quad A = 11110001, \quad Q = 00111101, \quad Q_{-1} = 1.$$

8. **Iteration 8:** $(Q_0, Q_{-1}) = (1, 1)$. $\longrightarrow$ No operation.
   Final shift:

   $$11110001\;00111101\;1 \quad \longrightarrow \quad A = 11111000, \quad Q = 10011110, \quad Q_{-1} = 1.$$

Final Product: The 16-bit product is the concatenation of A (HI) and Q (LO):

$$\text{HI:LO} = 11111000\;10011110.$$