CS 2110 — Quiz 1, Group A
January 2018

Answer Key
This quiz is worth a total of **100 points**.

Binary Addition, Subtraction and Base Conversion

For questions 1, 2 and 3, let A and B be 8-bit binary integers such that $A = 01011000_2$ and $B = 10011000_2$. **Represent all binary results in 8 bits**, disregarding overflows and truncating down to 8 bits when necessary. **When asked to convert to decimal, convert the 8-bit value.**

If you see that your results, when converted to decimal, don't match the results you would expect - don't worry, as the operations might produce overflow and the 8-bit truncation is expected to cause otherwise unreasonable results.

1. For the following calculations, interpret $\boldsymbol{A}$ and $\boldsymbol{B}$ as **unsigned integers**.

   (a) Calculate $\boldsymbol{A}+\boldsymbol{B}$ in binary, representing the result as an unsigned binary integer. 8

   Solution: $$\begin{array}{r} 0\,1\,0\,1\,1\,0\,0\,0 \\ +1\,0\,0\,1\,1\,0\,0\,0 \\ \hline 1\,1\,1\,1\,0\,0\,0\,0 \end{array}$$

   (b) Express the result of $\boldsymbol{A}+\boldsymbol{B}$ as a decimal integer: $\mathbf{128+64+32+16=240}$ 4

   (c) Calculate $\boldsymbol{B}-\boldsymbol{A}$ in binary, representing the result as an unsigned binary integer. 8

   Solution: $$\begin{array}{r} 1\,0\,0\,1\,1\,0\,0\,0 \\ -0\,1\,0\,1\,1\,0\,0\,0 \\ \hline 0\,1\,0\,0\,0\,0\,0\,0 \end{array}$$

   (d) Express the result of $\boldsymbol{B}-\boldsymbol{A}$ as a decimal integer: **64** 4

2. For the following calculations, interpret $\boldsymbol{A}$ and $\boldsymbol{B}$ as **signed magnitudes**.
   Hint: For signed magnitudes, the most significant bit is interpreted as the sign, and the rest of the binary value is interpreted as the magnitude much like an unsigned integer. Remember that the 7 least significant bits represent the absolute value of the number. Do the calculations on these 7 bits, changing the operation according to the sign bits if necessary. When done, truncate to 7 bits and add the correct sign bit.

   (a) Calculate $\boldsymbol{B}+\boldsymbol{A}$ in binary, representing the result as a signed magnitude. 5

   Solution: Since $\boldsymbol{B}$ is negative, and $\boldsymbol{A}$ is positive, $\boldsymbol{B}+\boldsymbol{A} = -|\boldsymbol{B}| + |\boldsymbol{A}| = |\boldsymbol{A}| - |\boldsymbol{B}| = \mathbf{1011000_2 - 0011000_2}$

   $$\begin{array}{r} 1\,0\,1\,1\,0\,0\,0 \\ -0\,0\,1\,1\,0\,0\,0 \\ \hline 1\,0\,0\,0\,0\,0\,0 \end{array}$$
   Then we add the sign to the magnitude: **01000000**.

   (b) Calculate $\boldsymbol{B}-\boldsymbol{A}$ in binary, representing the result as a signed magnitude. 5

   Solution: Since $\boldsymbol{B}$ is negative, and $\boldsymbol{A}$ is positive, $\boldsymbol{B}-\boldsymbol{A} = -|\boldsymbol{B}| - |\boldsymbol{A}| = -(|\boldsymbol{A}| + |\boldsymbol{B}|) = \mathbf{-(1011000_2 + 0011000_2)}$

   $$\begin{array}{r} 1\,0\,1\,1\,0\,0\,0 \\ +0\,0\,1\,1\,0\,0\,0 \\ \hline 1\,1\,1\,0\,0\,0\,0 \end{array}$$
   Then we add the minus sign to the magnitude: **11110000**.

   (c) Express the result of $\boldsymbol{B}-\boldsymbol{A}$ as a decimal integer: $\mathbf{(-1)*(64+32+16)=-112}$ 3

3. For the following calculations, interpret $\boldsymbol{A}$ and $\boldsymbol{B}$ as **2's complement integers**.

   (a) Calculate $\boldsymbol{B}+\boldsymbol{A}$ in binary, representing the result in 2's complement. $\boxed{8}$

   Solution: This is the same as *1a*:

   $$\begin{array}{r} 01011000 \\ +10011000 \\ \hline 11110000 \end{array}$$

   We know the number is negative, we flip it and add one to find its magnitude:
   $\boldsymbol{\sim 11110000_2 + 1 = 00001111_2 + 1 = 00010000_2 = 16}$

   (b) Express the result of $\boldsymbol{B}+\boldsymbol{A}$ as a decimal integer: **−16** $\boxed{4}$

   (c) Calculate $\boldsymbol{A}-\boldsymbol{B}$ in binary, representing the result in 2's complement. $\boxed{8}$
   Hint: Convert the operation to addition first and then do the calculation.

   Solution: $\boldsymbol{A - B = A + (-B) = A + (\sim B + 1) = 01011000_2 + (01100111_2 + 1) = 01011000_2 + 01101000_2}$

   $$\begin{array}{r} 01011000 \\ +01101000 \\ \hline 11000000 \end{array}$$

   We know the number is negative, we flip it and add one to find its magnitude:
   $\boldsymbol{\sim 11000000_2 + 1 = 00111111_2 + 1 = 01000000_2 = 64}$

   (d) Express the result of $\boldsymbol{A}-\boldsymbol{B}$ as a decimal integer: **−64** $\boxed{4}$

Bitwise Operations

4. Fill in the blanks with valid integers such that the equations are correct. Assume unsigned values.
   Hint: The 0x prefix is used to denote an hexadecimal number. E.g. 0x8 is equal to 8_{16}

   (a) 10110110_2 & $\sim$ (0x3 << 4) $= 10000110_2$ — *Express answer in decimal* $\boxed{5}$

   (b) $10110110_2 \wedge$ 11100011_2 $= 01010101_2$ — *Express answer in 8-bit binary* $\boxed{5}$

5. For this question, let C and D be 4-bit **2's complement** integers such that $C = 0110_2$ and $D = 1001_2$.

   (a) Extend C to be an 8-bit signed integer, writing the result in binary: $\boxed{2}$

   Solution: 00000110_2

   (b) Extend D to be an 8-bit signed integer, writing the result in binary: $\boxed{2}$

   Solution: 11111001_2

   (c) What is different between extending positive and negative signed integers? What would have happened to the values if you had extended both positive and negative integers in the same manner? $\boxed{10}$

   Solution: The difference is in what leading bits are inserted during extension - positive numbers are extended with 0s and negative numbers with 1s - this process is called **sign extension**. If negative numbers were also extended with 0s, the result would no longer be a negative number, and it would have a larger magnitude as well.

IEEE-754 Floating-Point Numbers

6. Interpret X and Y as IEEE-754 floating point numbers and answer the following questions.

Function	Sign	Exponent								Mantissa																						
Bit Index	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
X	1	1	1	0	1	1	1	1	0	1	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0
Y	1	1	1	0	0	0	1	1	0	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	0	0	0	1	0

(a) Which of the following is true for numbers X and Y? ○ $X > Y$ ○ $X = Y$ $\surd$ $X < Y$ $\boxed{5}$

(b) Justify your answer: What do the parts (denoted as "function" in the above table) of an IEEE-754 float represent? Which one(s) did you have to compare to arrive at your answer for part a? $\boxed{10}$

> **Solution:** The IEEE-754 float represents a number in base-2 scientific notation. The sign bit represents the sign (where 0 is positive and 1 is negative), the exponent group modifies the power to which 2 is raised, and the mantissa group modifies the ranged fraction the power of two is multiplied with. For this question, the sign bit is compared first, since the numbers are both equal, the one with a larger exponent is smaller. The exponents are then compared: since these are unsigned, X has a larger exponent, and thus must be the smaller of the two numbers. The mantissa does not need to be compared in this case.