NYU Tandon School of Engineering

Fall 2025, ECE 6913

Homework Assignment 4

---

1. How would you test for overflow, the result of an addition of two 8-bit operands if the operands were (i) unsigned (ii) signed with 2s complement representation.

Add the following 8-bit strings assuming they are (i) *unsigned* (ii) *signed and represented using 2's complement*. Indicate *which of these additions overflow*.

A. 0110 1110 + 1001 1111

B. 1111 1111 + 0000 0001

C. 1000 0000 + 0111 1111

D. 0111 0001 + 0000 1111

2. One possible performance enhancement is to do a shift and add instead of an actual multiplication. Since 9×6, for example, can be written $(2\times2\times2+1)\times6$, we can calculate 9×6 by shifting 6 to the left three times and then adding 6 to that result. Show the best way to calculate $0xAB_{hex} \times 0xEF_{hex}$ using shifts and adds/subtracts. Assume both inputs are 8-bit unsigned integers.

3. What decimal number does the 32-bit pattern 0×DEADBEEF represent if it is a floating-point number? Use the IEEE 754 standard

4. Write down the binary representation of the decimal number 78.75 assuming the IEEE 754 *single precision* format. Write down the binary representation of the decimal number 78.75 assuming the IEEE 754 *double precision* format

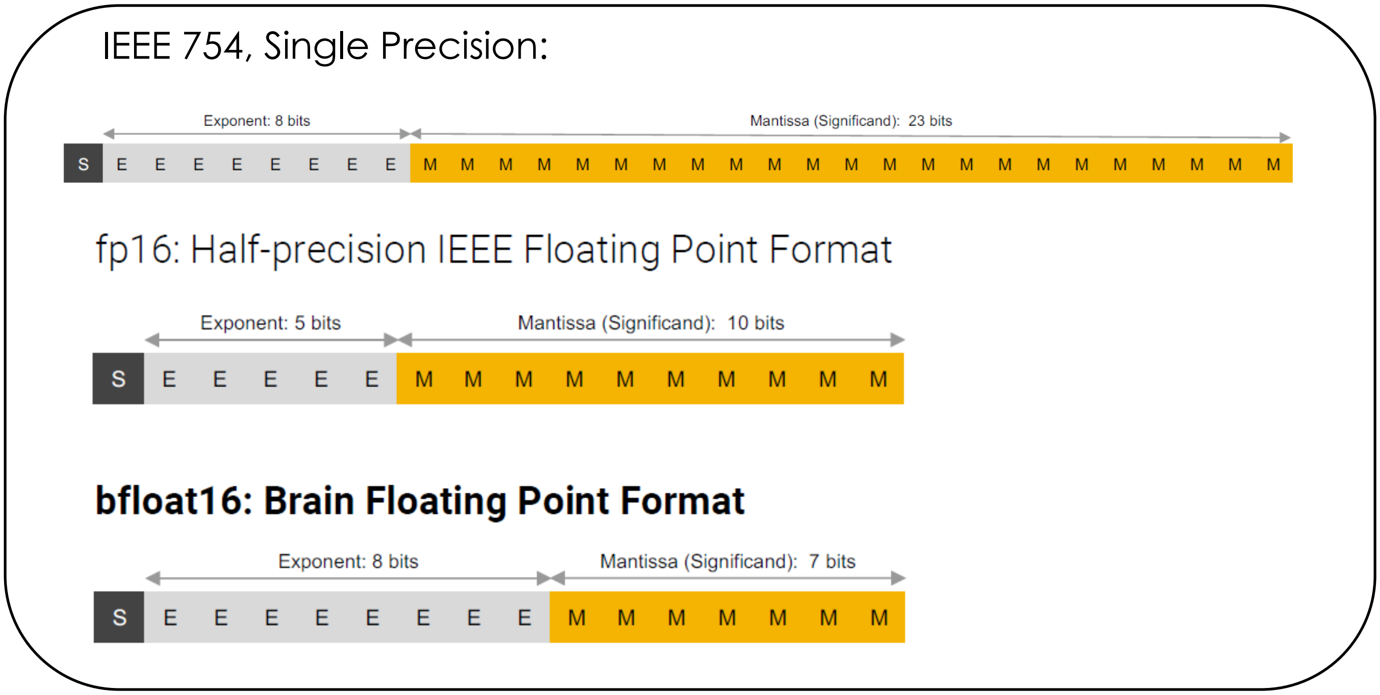
5. Write down the binary representation of the decimal number 78.75 assuming it was stored using the single precision **IBM format** (base 16, instead of base 2, with 7 bits of exponent).

6. IEEE 754-2008 contains a half precision that is only 16 bits wide. The leftmost bit is still the sign bit, the exponent is 5 bits wide and has a bias of 15, and the mantissa (fractional field) is 10 bits long. A hidden 1 is assumed.
(a) Write down the bit pattern to represent $\mathbf{-1.3625 \times 10^{-1}}$ Comment on how the range and accuracy of this 16-bit floating point format compares to the single precision IEEE 754 standard.

(b) Calculate the sum of $\mathbf{1.6125 \times 10^{1}}$ (A) and $\mathbf{3.150390625} \times 10^{-1}$ (B) by hand, assuming operands A and B are stored in the 16- bit half precision described in problem a. above Assume 1 guard, 1 round bit, and 1 sticky bit, and round to the nearest even. Show all the steps.

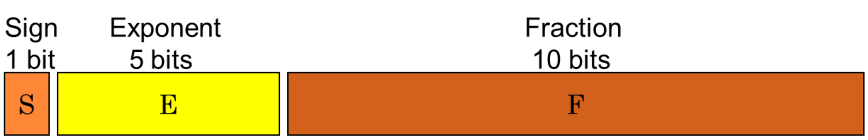
7. What is the range of representation and relative accuracy of positive numbers for the following 3 formats:

(i) IEEE 754 Single Precision (ii) IEEE 754 – 2008 (described in Problem 6 above) and (iii) 'bfloat16' shown in the figure below



8. NVIDIA has a "half" format, which is similar to IEEE 754 except that it is only 16 bits wide. The leftmost bit is still the sign bit, the exponent is 5 bits wide and the Fraction field is 10 bits long. A hidden 1 is assumed.

REPRESENTATION RANGE OF THE NVIDIA 'HALF FORMAT'



For each of the following, write the *binary value* and *the* ***corresponding*** *decimal value* <u>of the 16-bit floating point number that is the closest available representation of the requested number</u>. If rounding is necess2ary use round-to-nearest. Give the decimal values either as whole numbers or fractions. Show your work.

Number	Binary	Decimal
0	00000 0000000000	0
Mass of a neutron: 1.674×10^{-27} (Kg)		
Smallest positive normalized number		
Smallest positive **denormalized** number > 0		
Largest positive **denormalized** number > 0		
Largest positive number < infinity		
Average distance b/w proton and neutron in Hydrogen atom = 0.8751×10^{-15} m		
Number of Years since ancient 'supercontinent Gondwana' broke up 180,000,000		