CSE 351 Section 3 – Integers and Floating Point

Welcome back to section, we're happy that you're here ☺

Signed Integers with Two's Complement

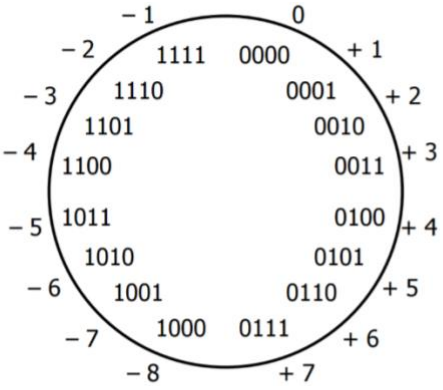
Two's complement is the standard for representing signed integers:

- The most significant bit (MSB) has a negative value; all others have positive values (same as unsigned)
- Binary addition is performed the same way for signed and unsigned
- The bit representation for the negative (additive inverse) of a two's complement number can be found by:

  flipping all the bits and adding 1 (i.e. $-x = \sim x + 1$).

The "number wheel" showing the relationship between 4-bit numerals and their Two's Complement interpretations is shown on the right:

- The largest number is 7 whereas the smallest number is -8
- There is a nice symmetry between numbers and their negative counterparts except for -8



Exercises: (assume 8-bit integers)

1) What is the **largest integer**? The **largest integer + 1**?

Unsigned: 1111 1111 -> 0000 0000	Two's Complement: 0111 1111 -> 1000 0000

2) How do you represent (if possible) the following numbers: **39**, **-39**, **127**?

Unsigned: 39: 0010 0111 -39: Impossible 127: 0111 1111	Two's Complement: 39: 0010 0111 -39: 1101 1001 127: 0111 1111

3) Compute the following sums in binary using your Two's Complement answers from above. *Answer in hex.*

a. 39 -> 0b 0 0 1 0 0 1 1 1 +(-39) -> 0b 1 1 0 1 1 0 0 1 0x 0 0 <- 0b 0 0 0 0 0 0 0 0	**b.** 127 -> 0b 0 1 1 1 1 1 1 1 + (-39)-> 0b 1 1 0 1 1 0 0 1 0x 5 8 <- 0b 0 1 0 1 1 0 0 0
c. 39 -> 0b 0 0 1 0 0 1 1 1 +(-127)-> 0b 1 0 0 0 0 0 0 1 0x A 8 <- 0b 1 0 1 0 1 0 0 0	**d.** 127 -> 0b 0 1 1 1 1 1 1 1 + 39 -> 0b 0 0 1 0 0 1 1 1 0x A 6 <- 0b 1 0 1 0 0 1 1 0

4) Interpret each of your answers above and indicate whether or not overflow has occurred.

a. 39 + (-39) Unsigned: 0 overflow Two's Complement: 0 no overflow	**b.** 127 + (-39) Unsigned: 88 overflow Two's Complement: 88 no overflow
c. 39 + (-127) Unsigned: 168 overflow Two's Complement: -88 no overflow	**d.** 127 + 39 Unsigned: 166 no overflow Two's Complement: -90 overflow

Goals of Floating Point

Representation should include: [1] a large range of values (both very small and very large numbers), [2] a high amount of precision, and [3] real arithmetic results (*e.g.* ∞ and NaN).

IEEE 754 Floating Point Standard

The value of a real number can be represented in scientific binary notation as:

$$\text{Value} = (-1)^{\text{sign}} \times \text{Mantissa}_2 \times 2^{\text{Exponent}} = (-1)^{S} \times 1.M_2 \times 2^{E\text{-bias}}$$

The binary representation for floating point values uses three fields:

- **S**: encodes the *sign* of the number (0 for positive, 1 for negative)
- **E**: encodes the *exponent* in **biased notation** with a bias of 2^{w-1}-1
- **M**: encodes the *mantissa* (or *significand*, or *fraction*) – stores the fractional portion, but does not include the implicit leading 1.

	S	E	M
`float`	1 bit	8 bits	23 bits
`double`	1 bit	11 bits	52 bits

How a `float` is interpreted depends on the values in the exponent and mantissa fields:

E	M	Meaning
0	anything	denormalized number (denorm)
1-254	anything	normalized number
255	zero	infinity (∞)
255	nonzero	not-a-number (NaN)

Exercises:

5) What is the largest, finite, positive value that can be stored using a `float`?

$S = 0, E = 254, M = 0b1…1.$ Value is then $1.1 \ldots 1 \times 2^{127} = (2 - 2^{-23}) \times 2^{127} = 2^{128} - 2^{104}$.

6) What is the smallest, positive, normalized value that can be stored using `float`?

7) $S = 0, E = 1, M = 0b0…0.$ Value is then $1.0 \times 2^{-126} = 2^{-126}$.

8) Convert the decimal number 1.25 into single precision floating point representation:

0	0	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

9) What are the decimal values of the following `float`s?

0x80000000	0xFF94BEEF	0x41180000
–0	NaN	+9.5

0x41180000 = 0b 0|100 0001 0|001 1000 0…0.
S = 0, E = 128+2 = 130 → Exponent = E – bias = 3, Mantissa = 1.0011_2
$1.0011_2 \times 2^3 = 1001.1_2 = 8 + 1 + 0.5 = 9.5$

Floating Point Mathematical Properties

- Not associative: $(2 + 2^{50}) - 2^{50}$!= $2 + (2^{50} - 2^{50})$
- Not distributive: $100 \times (0.1 + 0.2)$!= $100 \times 0.1 + 100 \times 0.2$
- Not cumulative: $2^{25} + 1 + 1 + 1 + 1$!= $2^{25} + 4$

Exercises:

10) Based on floating point representation, explain why each of the three statements above occurs.

Associative: Only 23 bits of mantissa, so $2 + 2^{50} = 2^{50}$ (2 gets rounded off). So LHS = 0, RHS = 2.

Distributive: 0.1 and 0.2 have infinite representations in binary point ($0.2 = 0.\overline{0011}_2$), so the LHS and RHS suffer from different amounts of rounding (try it!).

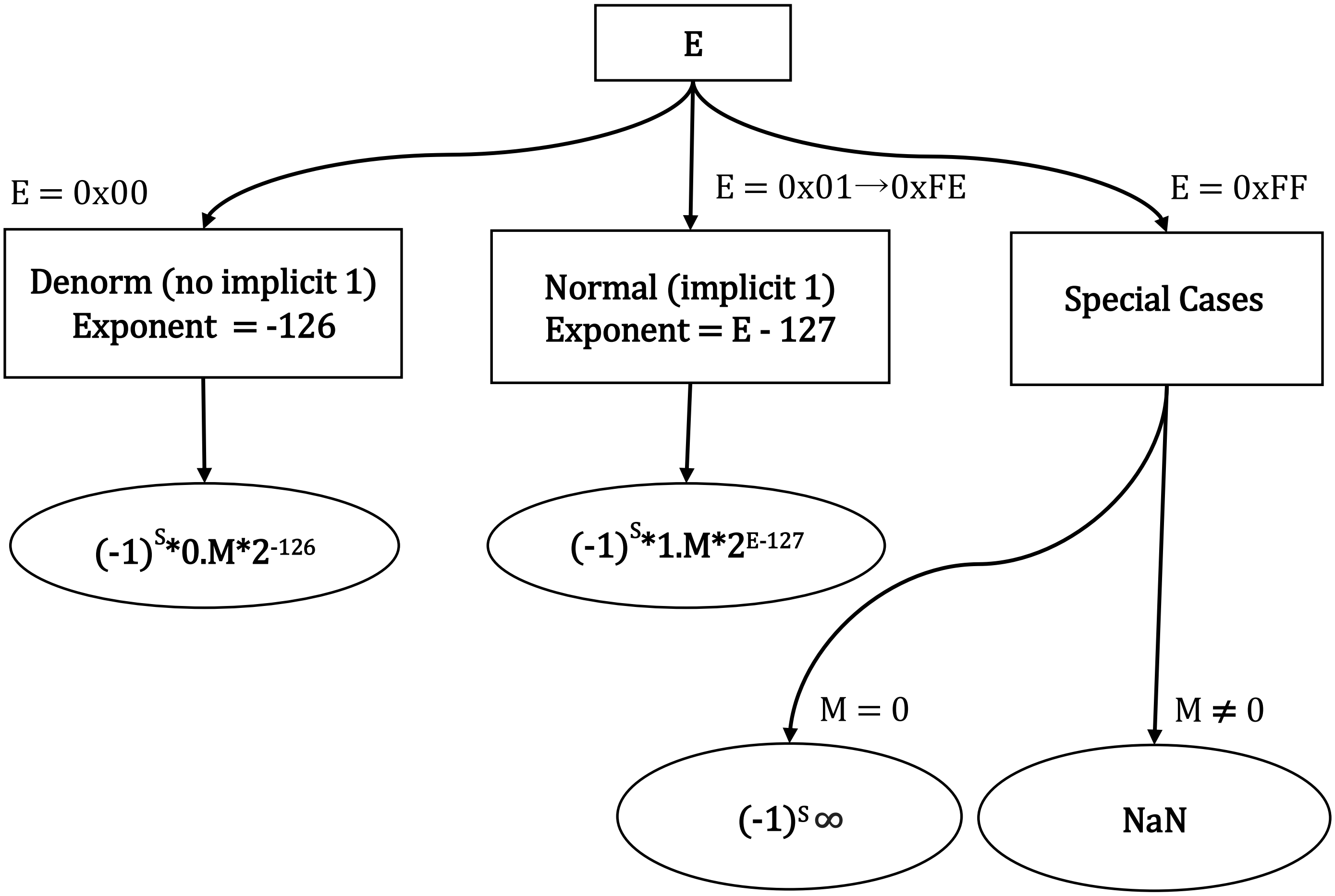
Cumulative: 1 is 25 powers of 2 away from 2^{25}, so $2^{25} + 1 = 2^{25}$, but 4 is 23 powers of 2 away from 2^{25}, so it doesn't get rounded off.

11) If `x` and `y` are variable type `float`, give two *different* reasons why `(x+2*y)-y==x+y` might evaluate to false.

(1) Rounding error: like what is seen in the examples above.

(2) Overflow: if `x` and `y` are large enough, then `x+2*y` may result in infinity when `x+y` does not.

1EEE 754 Float (32 bit) Flowchart


E
E = 0x00
E = 0x01→0xFE
E = 0xFF
Denorm (no implicit 1)
Exponent = -126
Normal (implicit 1)
Exponent = E - 127
Special Cases
(-1)^S*0.M*2^-126
(-1)^S*1.M*2^E-127
M = 0
M ≠ 0
(-1)^S ∞
NaN