CSE115 / CSE503
Introduction to Computer Science I

Dr. Carl Alphonce

343 Davis Hall

alphonce@buffalo.edu

Office hours:

Tuesday 10:00 AM – 12:00 PM*

Wednesday 4:00 PM – 5:00 PM

Friday 11:00 AM – 12:00 PM

OR request appointment via e-mail

**Tuesday adjustments: 11:00 AM – 1:00 PM on 12/6*

ROADMAP

Last time (before exam)

- int representation in detail

Today

- floating point representation

Coming up

- search

© Dr. Carl Alphonce

ANNOUNCEMENTS

FINAL EXAM NOTICE

CHECK YOUR ENTIRE FINAL EXAM SCHEDULE!

http://blogs.advising.buffalo.edu/beadvised/posts/have-you-checked-your-final-exam-schedule-4/

Room assignments will be announced at a later date.

© Dr. Carl Alphonce

double / float

(IEEE 754)

Primitives in Java

- Java has eight primitive types
  - boolean
  - integral types:
    - signed: long, int, short, byte
    - unsigned: char
  - floating point types: double, float
- Values of the primitive types are not objects
  - no properties
  - no capabilities

https://docs.oracle.com/javase/specs/jls/se8/html/jls-4.html#jls-4.2.3

Decimal (base 10) – with a *decimal* point

—	—	—	—	•	—	—	—	—
10^3	10^2	10^1	10^0		10^{-1}	10^{-2}	10^{-3}	10^{-4}
1000	100	10	1		1/10	1/100	1/1000	1/1000

Binary (base 2) – with a *binary* point

—	—	—	—	•	—	—	—	—
2^3	2^2	2^1	2^0		2^{-1}	2^{-2}	2^{-3}	2^{-4}
8	4	2	1		1/2	1/4	1/8	1/16

Examples

$101.1_2 = 1*2^2 + 0*2^1 + 1*2^0 + 1*2^{-1} = 4 + 1 + ½ = 5½$

$0.111 = 1*2^{-1} + 1*2^{-2} + 1*2^{-3} = ½ + ¼ + ⅛ = ⅞$

double

values: 0.0, 1.0, -3.5, 3141.5926535e-3

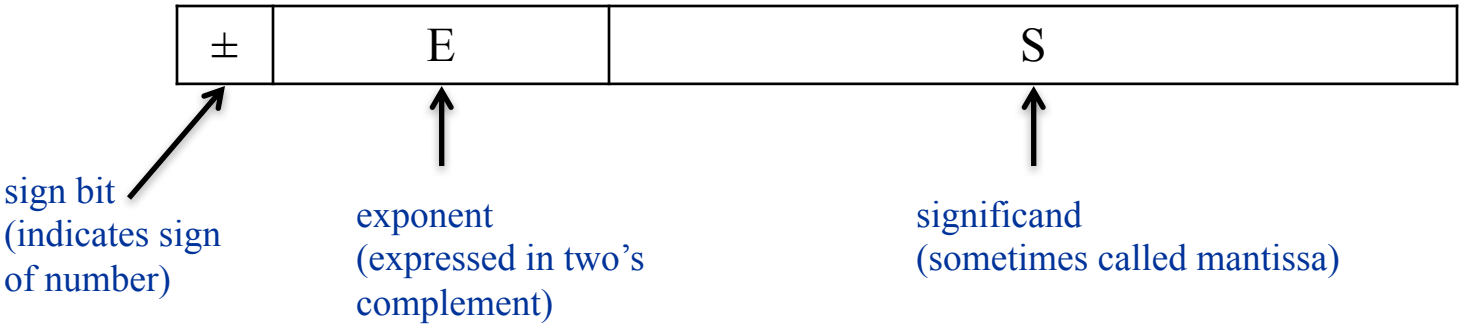
inexact representation (!)

operations: + - * /

5.0 + 2.0 = 7.0	+	double X double	→ double
5.0 – 2.0 = 3.0	-	double X double	→ double
5.0 * 2.0 = 10.0	*	double X double	→ double
5.0 / 2.0 = 2.5	/	double X double	→ double

floating point types' representation

both double and float use the IEEE754 representation scheme:



represents $\pm 1S * 2^E$

- S is represented in normalized form (no leading 0)
- since first bit is always 1, it is not stored

size of representation differs according to type:

- the representation of a float is 4 bytes wide
- the representation of a double is 8 bytes wide

NB: I gloss over several important details of the IEEE754 standard here. The intent is not to give a full presentation of the standard, but to give you a rough idea of how floating point numbers are represented, and especially that integral types of floating point types use very different representation schemes!

Things to watch for!

Representation is inexact

- in base 10, 1/3 does not have an exact finite representation
  - 0.33333333…
- likewise in base 2: 0.010101…
- in base 2, 1/10 does not have an exact finite representation
  - 0.000110011…

Mixing magnitudes

- it is possible that x+y is the same as x
  - 1.0e-15 + 1.0e-15 → 2.0 e-15
  - 1.0e+15 + 1.0e-15 → 1.0 e+15 (whoops!)

Round-off errors – relevant for making comparisons

```
double d = 0.1;
double pointNine = d+d+d+d+d+d+d+d+d;
```

pointNine is not the same as 0.9 (whoops!)

relational operators

We can form expressions like:

x < y

which have a value of true or false (i.e. the type of this expression is boolean)

relational operators: < <= > >= ==

BUT write:

Math.abs(x-y) < eps

for some small value eps, not

x == y