September 7, 2016

Floating point mathematics

In this class:

- *How arithmetic operations involving floating point numbers work.*
- *IEEE rounding modes and the guarantees of an IEEE system.*
- *(An example of why even simple computations oven discard many significant digits.*

Next class

QUIZ and floating point math
Floating Point
G&C – Chapter 5

Next next class

Monte Carlo algorithms
G&C – Chapter 3

The most important person you’ve never heard of (yet)!

William Kahan

Fought to get a standard to floating point arithmetic that provided useful mathematical properties.

Won a Turing award (the “Nobel prize” of CS) for this!

Quick review

A floating point number

Quick review

A floating point number

- a sign
- an exponent
- a mantissa

Toy system

1 bit for sign

2 bits of mantissa

2 bits for exponent (-1,0,1,∅)

$1\ 10\ 0 = (-1)^{1} \times (1.10)_2 \times 2^{0}$

Real system

$(-1)^{\text{sign}}$
$\times\ (1.\text{mantissa})_2$
$\times\ 2^{\text{exponent}-\text{bias}}$

Inf and NaN values too!

IEEE Single

- 1 bit for sign
- 8 bits for exponent
- 23 bits for mantissa
- Bias=127, ∅ = 0, Inf=255

IEEE Double

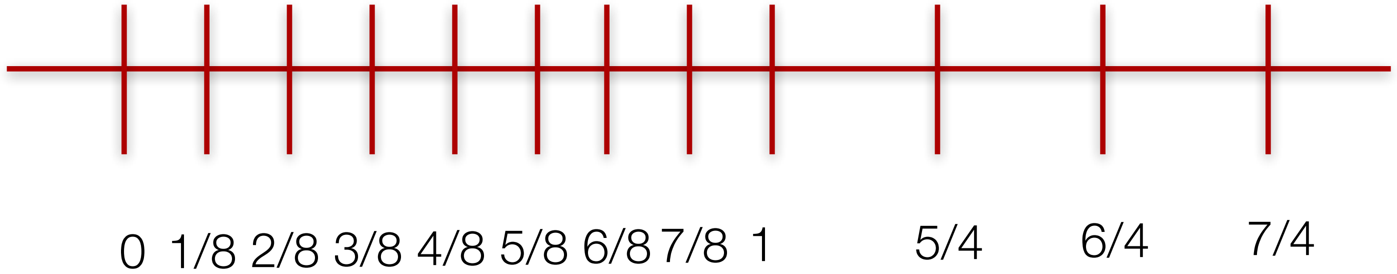
- 1 bit for sign
- 11 bits for exponent
- 52 bits for mantissa
- Bias=1023, ∅=0, Inf=2048

IEEE Quad

- 1 bit for sign
- 15 bits for exponent
- 112 bits for mantissa
- Bias = , ∅ = 0

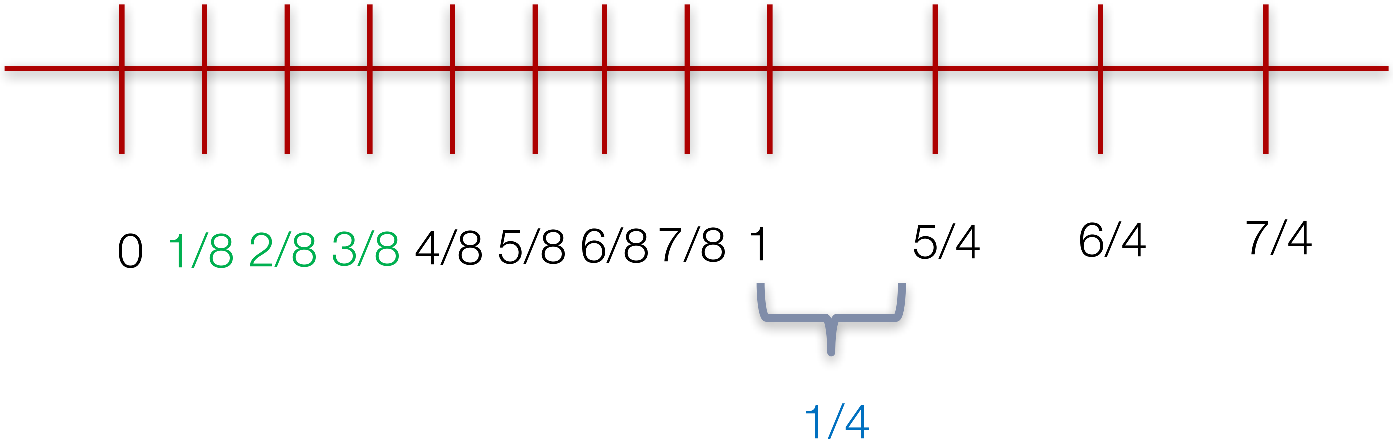
An important property of floats

- Subnormal numbers
- Machine epsilon (the difference between 1 and the next largest floating point number)



An important property of floats

- Subnormal numbers
- Machine epsilon (the difference between 1 and the next largest floating point number)



… demo ...

… back to board ...