Floating point numbers

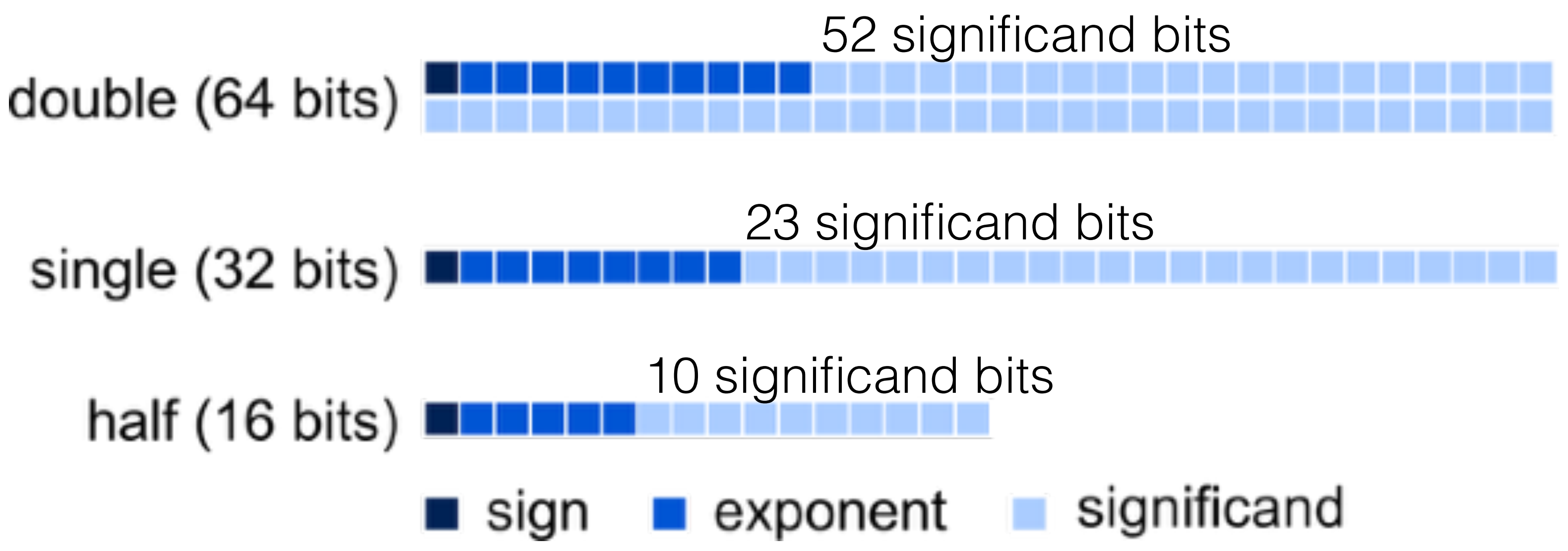
Method to encode numbers in binary

$$x = (-1)^{\text{sign}} \times \overset{\text{Significand}}{\underset{\text{Precision}}{\left(1 + \sum_{n=1}^{N} s_n 2^{-n}\right)}} \times \overset{\text{Exponent}}{\underset{\text{Magnitude}}{2^e}}$$

Think of

$$65504 = 6.5504 \times 10^4$$

Computers have standards layouts for these numbers



This talk: focus on the significand (precision).

“New” types of computers

Lower precision, parallel computations

GPU - Graphs processing unit

- Massively parallel.
- Used for machine learning, where high precision is often unnecessary.
- Support half-precision floats.

FPGA - Field programmable gate arrays

- Programming at a logic gate level (very hard).
- Configure a chip to solve only your equations (very power efficient).
- Can use arbitrary numerical precisions (not just double, single, half).
- Now available on cloud computing, e.g. Amazon, Microsoft.

Can we take advantage of these developments?