Computing Basics (Part II)

Rubin H Landau
With
Sally Haerer and Scott Clark

Computational Physics for Undergraduates
BS Degree Program: Oregon State University

"*Engaging People in Cyber Infrastructure*"
Support by EPICS/NSF & OSU

Facts of Life: Computers are Finite

- **Regardless of power, still finite**
  - internal memory
  - external storage
  - speed (CPU, storage)
- **Finite precision of (most) numbers**
- **Scientific representation: finite # digits**
- **Consequences? Not to be minimized**
  rockets explode

"Finite" Number Representations

- **Multiple meanings for "finite"**
- **Bits = *binary integer* = *(0, 1)* = (↑, ↓)**
- **All computer numbers ultimately binary**
- **Integers*:* *N* bits ⇒ 2^N max**
- **⇒ Limited range [0, 2^{N-1}] (1 sign, 32 b ⇒ 10^9)**
- **Binary: 110010010101: OK if computer**
- **People: *octal, decimal, hexadecimal* ()**
- **Decimal: nice but reduces precision, final output**

Finite Number Representations (2)

- *"Word length"* = # bits for stored number
- 1 byte = 1 B = 8 bits ()
- Careful: 1 K = 1 KB = 2^{10} B =1024 B
- 1 byte ≈ memory for 1 character ("a")
- 1 typed page ≈ 3 KB
- 1st PCs: 8-b words (2^7= 128)
- *"Overflow"*: too large a number; "*Underflow*"
- Now: 32, 64 bit PC, fastest processing
- $2^{31} \approx 2 \times 10^9$: OK for banks, signals, not science

Fixed-point Numbers (ints)

$$I_{fix} = \pm\left(\alpha_n 2^n + \alpha_{n-1}2^{n-1} + \cdots + \alpha_0 2^0 + \cdots + \alpha_{-m}2^{-m}\right) \qquad (1)$$

- *1* bit: sign, *N-1* bits: α_i
- *N*, *m*, *n* values: machine-dependent
- 32-bit machine $\rightarrow$ 4B for *int*
- *Good*: absolute error $\equiv$ $2^{-(m+1)}$ (left off)
- *Bad*: large relative error for small #
- *Bad*: small range:

$$-2147483648 \le \text{int} \le 2147483647$$

IEEE Number Types

Name	Type	Bits	Range
boolean	logical	1	*true* or *false*
char	string	16	'\u0000' ↔ '\uFFFF'
byte	integer	8	-128 ↔ +127
short	integer	16	-32,768 ↔ +32,767
int	integer	32	-2,147,483,648 ↔ +2,147,483,647
long	integer	64	-9,223,372,036,854,775,808 ↔ +9,223,372,036,854,775,807
float	floating point	32	$1.401298 \times 10^{-45} \leftrightarrow 3.402923 \times 10^{+38}$
double	floating point	64	$4.9406564584124654 4 \times 10^{-324} \leftrightarrow 1.797693134862315 7 \times 10^{+308}$

IEEE Floating Point

- **Binary version of scientific notation**
- **$c \approx +2.997924 \times 10^8$ m/sec**
- **2.997924 = mantissa, 7 significant figures**
- **+8 = exponent**
- **. = *decimal* point (base 10)**
- **Storage: (sign) (exponent) (mantissa)**