Computer Basics
IEEE Floats (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

Model: Machine Precision

- Floating point ⇒ limited precision
  - singles (32b): 6-7 decimal places
  - doubles 15-16 places
  - affects calculations: $7 + 10^{-7} = ?$

$$7 = 0\ \ 1000\ 0010\ 1110\ 0000\ 0000\ 0000\ 0000\ 000,$$

$$10^{-7} = 0\ 0110\ 0000\ 1101\ 0110\ 1011\ 1111\ 1001\ 010$$

- Different exponents ⇒ can't add mantissas
- Shift bits (insert 0's) till same exponent

$$10^{-7} = 0\ 1000\ 0010\ 0000\ 0000\ 0000\ 0000\ 0000\ 000$$

$$\Rightarrow \quad 7 + 10^{-7} = 7$$

  - Last digits lost (truncation error)
  - Ignore 10^{-7} in single precision

``Machine Precision'' ε_m

- Measure of precision in computation
- Definition: *max* # add to 1 gives 1 as answer

$$1_c + \varepsilon_m = 1_c \qquad (1)$$

- $\Rightarrow$ *x* stored on computer:

$$\varepsilon_m \approx 10^{-7} \text{ (single)}, \quad \varepsilon_m \approx 10^{-16} \text{ (double)} \qquad (2)$$

$$x_c \approx x\,(1 \pm \varepsilon), \qquad |\varepsilon| \le \varepsilon_m \qquad (3)$$

- Scientists: Just say *No! to singles*

Time for Exercises in Lab

Landau's Rules of Education

1. Most of ed is learning names; ideas simple
2. Confusion is the first step towards understanding
3. Traumatic experiences are good teachers

Exercise: Determine Your ε_m

1. Determine the machine precision ε_m of your computer

Sample pseudocode:

```
eps = 1.
begin do N times
       eps = eps/2.                    // Make smaller
       one = 1. + eps
       write: loop number, one, eps
end do
```

2. Determine ε_m for single & double precision (floats)

3. * Determine precise values
   (decimal OK, octal, hexadecimal/octal better)

Exercise: Summing Series

(sum =) $$e^{-x} \simeq 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots + \frac{x^N}{N!}$$ + remainder

(term)

1. Calculate e^{-x} : $x = 0.1, \quad 1, \quad 10, \quad 100, \quad 1000$

2. Require "absolute" error less than 1 part in 10^8

$$|term/sum| < 10^{-8}$$

Presumptions: remainder ≈ term

no roundoff error

Series Exercise (cont)

$$(sum =)\ e^{-x} \simeq 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots + \frac{x^N}{N!}$$

(term)

```
term = 1, sum = 1, eps = 10**(-8)     // Initialize
do
    term = -term * x/i                // "Good", factorials "bad"
    sum = sum + term                  // Add in term
    while abs(term/sum) > eps         // Continue if error
end do
```

3. Output as table

```
x     N     sum      |sum  - exp(-x)| /sum
```

"**exp(-x)**" = built-in exponential `(Math.exp)`

4. Compare "good" to "bad" (factorials and powers)