CEE432/CEE532/MAE 541
Developing Software for Engineering Applications

Lecture 5: Introduction to Numerical Analysis
(Chapter 5)

Numerical Representation

Decimal

$$(4096)_{10} = 6\times10^{0} + 9\times10^{1} + 0\times10^{2} + 4\times10^{3}$$

Binary

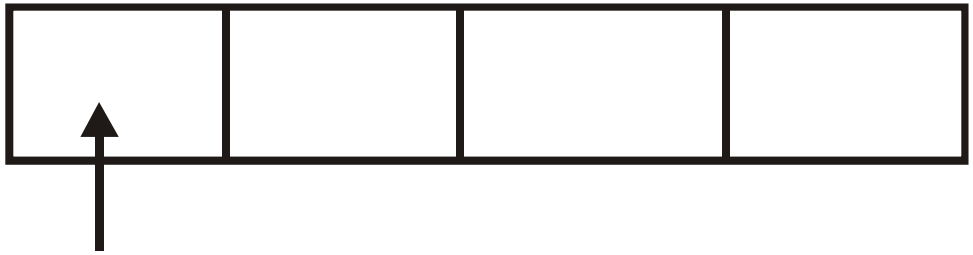
$$(10011)_{2} = 1\times2^{0} + 1\times2^{1} + 0\times2^{2} + 0\times2^{3} + 1\times2^{4} = (19)_{10}$$

Bit: Smallest storage unit (stores a 0 or a 1)

Byte: 8 bits make a byte

Word: 4 bytes (32-bit word) or 8 bytes (64-bit word)

An Example Storage



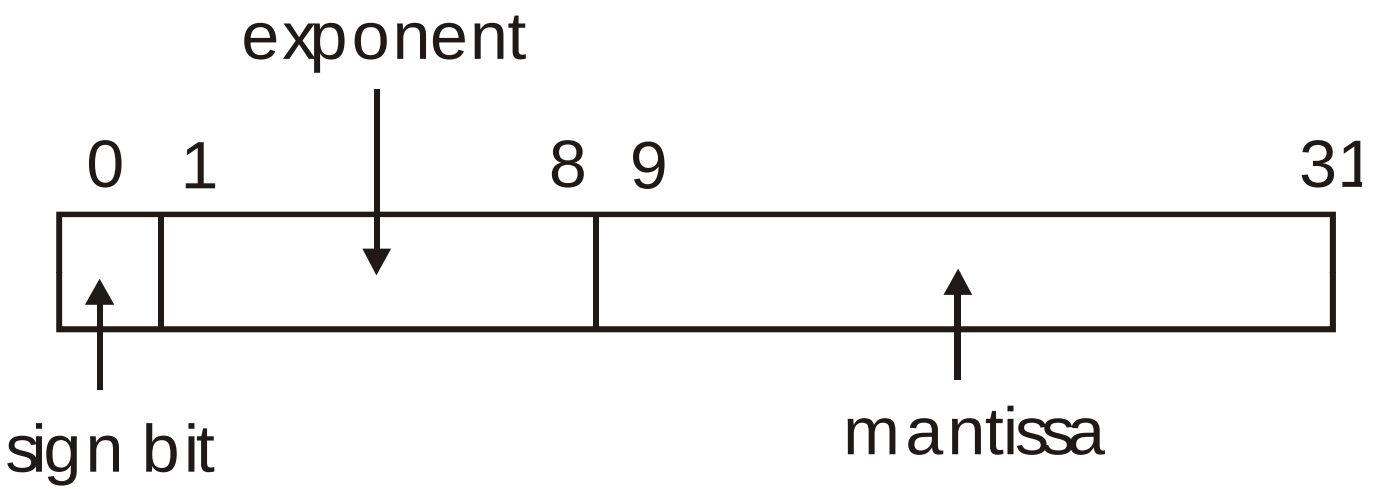
Largest Value

$$(111)_2 = 1\times 2^0 + 1\times 2^1 + 1\times 2^2 = (7)_{10}$$

Smallest Value

$$-(111)_2 = 1\times 2^0 + 1\times 2^1 + 1\times 2^2 = (-7)_{10}$$

Floating Point: IEEE Representation




```
S EEEEEEEE FFFFFFFFFFFFFFFFFFFFFFF
0 1       8 9                    31
```

Types of Errors

Absolute Error

$$E_{abs} = \left| x_t - x_a \right|$$

Relative Error

$$E_{rel} = \frac{\left| x_t - x_a \right|}{\left| x_t \right|}$$

Types of Errors

- Round-Off: Fixed number of bits are used to represent numbers

$$x = x_a \times 10^n + x_e \times 10^{n-d} = \text{approx } x + \text{error}$$

- Truncation: An approximation is used in the place of an exact representation

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - .....$$

Machine Epsilon

- Upper bound on the relative error in representing a floating point number

$$\left|\frac{x - x_a}{x}\right| \leq \varepsilon \qquad \varepsilon = x_a - 1$$