Introduction to Languages for Scientific Computing

Prof. **Paolo Bientinesi**

pauldj@aices.rwth-aachen.de

Floating Point Arithmetic – 1

- **[Q1]** Consider the IEEE settings for single precision arithmetic:

$$\beta = 2, \quad t = 24, \quad e_{\min} = -125, \quad e_{\max} = 128$$

1. What is the smallest floating point number larger than 2? $2 + \frac{1}{2}^{22}$
2. What is the largest floating point number smaller than 8? $8 - \frac{1}{2}^{21}$
3. How many floating points in [1/64, 1/32] ? $2^{23} + 1$
4. Distance between 65536 and the next floating point number? $\frac{1}{2}^{7}$
5. What is the first integer that cannot be represented exactly? $2^{24} + 1$

Floating Point Arithmetic – 2

- **[Q2]** Consider the following ternary arithmetic with normalization:

$$\beta = 3, \quad t = 3, \quad e_{\min} = -2, \quad e_{\max} = 3$$

1. How is π represented? $1.00 \times 3^1 = 3$

   What is the representation error? 0.04507 (relative)

   0.14159 (absolute)

2. What is the largest floating point number? 78

3. First 5 positive integers that cannot be represented? $28, 29, 31, 32, 34$

Floating Point Arithmetic – 3

- **[Q3]** Consider the following binary arithmetic with normalization:

$$\beta = 2, \quad t = 4, \quad e_{\min} = -2, \quad e_{\max} = 4$$

1. How is π represented? $1.101 \times 2^1 = 3.25$

   What is the representation error? 0.0333 (relative)

   0.1084 (absolute)

2. Smallest absolute distance between 2 floating point numbers? 2^{-5}

   (2^{-5})

3. Smallest relative distance between 2 floating point numbers? 2^{-4}

   (2^{-4})

Floating Point Arithmetic – 4

- $d_i = 2^{29} - 1$
- $D = 16\ d_i = 2^{33} - 16$
- Is D representable exactly in single precision? $\rightarrow$ **NO**
  If not, what are the absolute and relative representation errors?

$$\texttt{single}(d_i) = 2^{29}, \quad \texttt{single}(D) = 2^{33}$$

$$\text{Abs.Err.} = |2^{33} - 16 - 2^{33}| = 16$$

$$\text{Rel.Err.} = \frac{16}{D} = 1.8610^{-9} < \mathbf{u}$$

- Is D representable exactly in double precision? $\rightarrow$ **YES**