



**TÉCNICO**  
LISBOA

**Instituto Superior Técnico**

**Sistemas de Processamento Digital de Sinais**

**Digital Signal Processing Systems**

## Problem: Numerical representation and fixed-point operations

Consider the real numbers  $x = 17.35$ ,  $y = 0.15$  and its representation and processing using fixed point arithmetic.

1. Determine the arithmetic formats which allow the most accurate representation of  $x$  and  $y$  with 16 bit words and  $z = x \cdot y$  with 32 bit words. Determine the values of  $x$  and  $y$  in these formats and the resulting value of

$$z, z_{\text{real}}. \text{ Compute the relative absolute error of } z, \varepsilon_{\text{rel}}(z) = \left| \frac{z - z_{\text{real}}}{z} \right|.$$

How could this error be made smaller?

Write the C code that implements this computation including variable declarations and initialization.

2. Since in this case the true value of  $z$  is known beforehand, what is the most precise format that could be used to represent it?

Exercise:  $x = 17.35$ ,  $y = 0.15$ , fixed point 16 bit  
 $z = x \cdot y = 2.6025 = z_{true}$   
 $y_{Q_{15}}$ ,  $|x| > 16$ ,  $< 32 \rightarrow$  need 5 integer bit  $\rightarrow Q_{10}$

① Assume  $y_{Q_{15}}$ ,  $x_{Q_{10}}$  but actual values are unknown.

$z = x \cdot y \rightarrow$  needs  $5+0 = 5$  integer bits  $\Rightarrow Q_{10}$  or  $Q_{26}$  (32 bit)

$$x = \text{round}(2^{10} \times 17.35) = 17766$$

$$y = \text{round}(2^{15} \times 0.15) = 4915$$

$$z = 4915 \times 17766 = 87319890 \quad \text{but is in } Q_{25} \text{ because of extra sign bit}$$

$$z = 2 \times 87319890 = 174639780$$

$$z_{\text{real}} = \frac{174639780}{2^{26}} = 2.602335512638 \dots$$

$$\epsilon_r = \left| 1 - \frac{z_{\text{real}}}{2.6025} \right| = 6.32 \times 10^{-5}$$

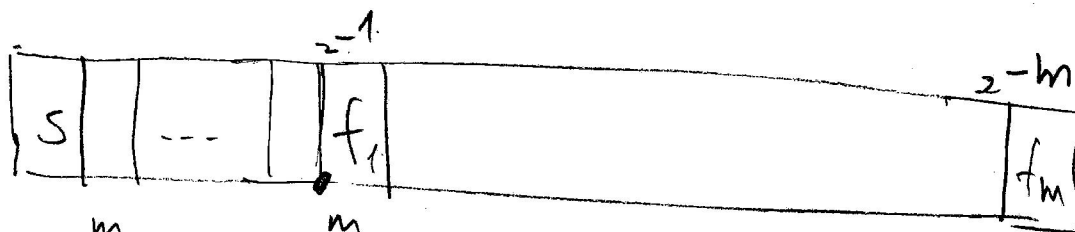
Code:  $\begin{matrix} \text{Int}+16 & x = 17766, y = 4915, z_{16}; \\ \text{Int}+32 & z \end{matrix}$

$$z = (x * y) \ll 1; \quad (Q_{26})$$

$$z_{16} = (z \gg 16); \quad \text{or} \quad z_{16} = ((x * y) \ll 1) \gg 16; \quad (.Q_{10})$$

② In this case we know the values of the operands and the result, which is  $2.6025 \Rightarrow$  in fact need only 2 integer bits because  $|z_{true}| < 4! \Rightarrow$  can store in  $Q_{29}$  (or  $Q_{13}$ )

$$z = \underbrace{((x * y) \ll 1)}_{Q_{26}} \ll 3; \quad Q_{29}!$$



$$\sum_{h=1}^m 2^{-h} = \sum_{h=0}^m 2^{-h} - 1 = \frac{1 - 2^{-(m+1)}}{1 - 2^{-1}} - 1 = 2 - 2^{-(m+1)+1} - 1 = \underbrace{1 - 2^{-m}}_{\text{mantissa maximum value}}$$

$$\epsilon_a = |x - \hat{x}|, \quad \hat{x} = x \pm 2^{-(m+1)}$$

$$\epsilon_a = 2^{-(m+1)} \equiv \text{absolute error}$$

$$\epsilon_r = \frac{2^{-(m+1)}}{x} \equiv \text{absolute relative error (depends on } x)$$