TÉCNICO LISBOA

# Instituto Superior Técnico Sistemas de Processamento Digital de Sinais (SPDSina)

## Fixed-point numerical representation

Consider the real numbers $x = 17.35$ and $y = 0.15$ and its processing using fixed point arithmetic. Assume a symmetric two's complement representation: the number $w$ is represented in the interval $-(2^i - 2^{-m}) \leq w \leq 2^i - 2^{-m}$ where $i$ is the number of integer bits and $m$ the number of fractional bits (total number of bits is $n$=1+$i$+$m$).

a) Determine the arithmetic formats which allow the most accurate representation of $x$ and $y$ with $n$=16 bit and $z = x \cdot y$ with 32 bit.

b) Determine the values of $x$ and $y$ in these formats and the resulting value of $z$, $z_{\text{real}}$. Compute the relative absolute error of $z$, $\varepsilon_{\text{rel}}(z) = \left|\frac{z_{ideal} - z_{real}}{z_{ideal}}\right|$. How could this error be made smaller? Compute the same error if $z$ is represented with 16 bit instead of 32.

c) Assume you know *apriori* the values of the operands $x$ and therefore of $z$. Determine the arithmetic formats which allow the most accurate representation of $z$ and determine the relative error.

Exercise: $x = 17.35$, $y = 0.15$, fixed point 16 bit

$z = x \cdot y = 2.6025 = z_{true}$

$y_{Q15}$, $|x| > 16$, $< 32$ → need 5 integer bit → $Q_{10}$

a) Assume $y_{Q15}$, $x_{Q10}$ but actual values are unknown.

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

$x = \text{round}(2^{10} \times 17.35) = \text{round}(17766.4) = 17766$

$y = \text{round}(2^{15} \times 0.15) = \text{round}(4915.2) = 4915$

b)

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

$z = 2 \times 87319890 = 174639780$

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

$\varepsilon_r = \left| 1 - \frac{z_{real}}{2.6025} \right| = 6.32 \times 10^{-5} \approx 10^{-4.2}$ (4.2 decimal places)

(com 32 bit)

Code:

```
Int16  x = 1766, y = 4915; z16;
Int32  z
z = (x * y) << 1;        (Q26)
z16 = (z >> 16);   or   z16 = (((Int32)(x * y)) << 1) >> 16;
                                                  (Q10)
```

$\log_{10} 2^{-15} \approx -15 \times 0.3 = -4.5$

c) 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 $|z_{true}| < 4$! ⇒

```
z = ((x * y) << 1) << 3;
```

($Q_{26}$)

$Q_{29}$!

→ $z_{real} = \frac{174639778 \gg 13}{2^{13}} = \frac{21318}{2^{13}} = 2.602294922$

$\varepsilon = 7.88 \times 10^{-5}$ → almost equal to 32 bit

$= 10^{-4.1}$

Note: symmetric interval: $-(2^i - 2^{-m}) \leq \omega \leq 2^i - 2^{-m}$

With 16 bit ($z_{16}$)

$z_{real} = 174639780 \gg 16 = 2664$ ($Q_{11}$)

$z_{real} = \frac{2664}{2^{10}} = 2.601562^5$

$\varepsilon = 3.6 \times 10^{-4} \approx 10^{-3.44}$