Instituto Superior Técnico

Sistemas de Processamento Digital de Sinais

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

   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)

$x = \text{round}(2^{10} \times 17.35) = 17766$
$y = \text{round}(2^{15} \times 0.15) = 4915$

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

$z = 2 \times 87319890 = 174639780$

$$z_{real} = \frac{174639780}{2^{26}} = 2.60233551 2638\ldots$$

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

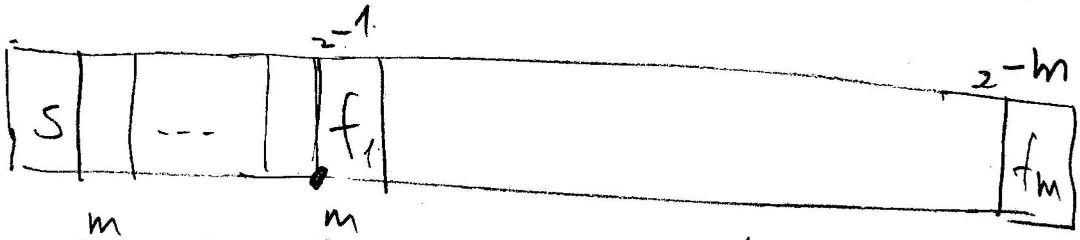
code:

```
Int16   x = 17766, y = 4915, z16;
Int32   z

        z = (x * y) << 1;     (Q26)
        z16 = (z >> 16);   or   z16 = ((x * y) << 1) >> 16;
                                                         (Q10)
```

(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})

$$z = \underbrace{\underbrace{((x * y) \ll 1)}_{Q26} \ll 3}_{Q29!};$$



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

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

$\varepsilon_a = 2^{-(m+1)}$ = absolute error

$\varepsilon_r = \dfrac{2^{-(m+1)}}{x}$ = absolute relative error (depends on x)