TÉCNICO LISBOA

Sistemas de Processamento Digital de Sinais (SPDSina)

Digital Representation – Fixed Point

II — Consider the fractional numbers x and y both in Q_{14} format and their processing in a 16 bit fixed-point processor.

a) Determine the format that should be used to represent $z = 2 \cdot (x^2 - y^2)$ with 16 bit such that z is always correct for all possible values of x and y. Justify.

b) Consider now that $x = 1.25$ and $y = -0.75$. Write the most efficient and accurate fixed-point C code (32 bit processing when possible) which computes z in the format determined in a), including variable declarations and initializations, using only 16 bit words.

c) Compute the true value of z and the value obtained with the computation in b). Explain the result.

$z = 2(x^2 - y^2)$, x, y em Q_{14} → regras!

a) formato de z?

$x^2 \to Q_{13}$, $y^2 \nearrow$ $\quad x^2 - y^2 \to Q_{12}$ $\quad 2(x^2 - y^2) \to Q_{11}$

$-(2 - 2^{-14}) \leq x, y \leq 2 - 2^{-14}$ $\quad x^2 - y^2 \to$ máximo quando $y = 0$, x^2 Q_{13}
→ mínimo quando $x = 0$, $-y^2$ Q_{13}

⇒ z pode ser representado em Q_{12}! e não Q_{11} ⇒ 1 bit a mais sem custo

b)
$x = 1.25$, $x_{Q_{14}} = \text{round}(1.25 \times 2^{14}) = 20480$

$y = -0.75$, $y_{Q_{14}} = \text{round}(-0.75 \times 2^{14}) = -12288$

Int1 $x = 20480$, $y = -12288$, z; ficam MSBs

$z = (((((Int32)x * x - (Int32)y * y) << 1) << 1) >> 1) >> 16$; $Q_{28} \to Q_{12}$

$x^2 - y^2 \equiv Q_{13}$; sign ; $\times 2$ $Q_{13} - Q_{12}$

ou deixar logo z em Q_{12}

$z = (((Int32)x * x - (Int32)y * y) << 1) >> 16$;

Q_{12} ; $\times 2$; MSB's

$z = 2(1.25^2 + (-0.75)^2) = 2$

$z_{real} = (20480^2 - (-12288)^2) \times 2 \times \underbrace{\frac{1}{2^{16}}}_{\text{MSB's}} \times \underbrace{\frac{1}{2^{12}}}_{Q_{12}} = 2$

$z = z_{real} \Rightarrow$ erro = 0 porque o resultado é representável por uma soma de potências (positivas e negativas) de 2 com 16 bit