Arithmétique en Virgule Fixe

Virgule fixe

- Facteur d'échelle implicite

$$\text{facteur} = 2^e \text{ avec } e \in \mathbb{Z}$$

Ceci revient à mettre une virgule virtuelle

$$\langle F, e \rangle \leftrightarrow (f_{n-1}, f_{n-2}, \cdots, f_0)$$

$$\text{val}(\langle F, e \rangle) = \begin{cases} \left(\sum_{k=0}^{k=n-1} f_k * 2^k \right) * 2^e & \text{si non signé} \\ \left(-f_{n-1} * 2^{n-1} + \sum_{k=0}^{k=n-2} f_k * 2^k \right) * 2^e & \text{si signé} \end{cases}$$

Exemples

$$0.625 = {}^{5}\!/_{8} = 5 \times 2^{-3} \leftrightarrow \langle 5, -3 \rangle$$

0000 0000 0000 0101

Partie entière Partie décimale

$$0.12109375 = {}^{31}\!/_{256} = 31 \times 2^{-8} \leftrightarrow \langle 31, -8 \rangle$$

$$-0.0018768311 = {}^{-123}\!/_{65536} = -123 \times 2^{-16} \leftrightarrow \langle -123, -16 \rangle$$

Codage optimal

v (valeur)	e (exposant)	m (mantisse)
0.0	0	0
$0.0 \leq v \leq 65535.0$	$-\left\lfloor \log_2 \frac{65535.0}{v} \right\rfloor$	$\left\lfloor 2^{-e} \times v + 0.5 \right\rfloor$
$v > 65535.0$	$\left\lfloor \log_2 \frac{v}{65535.0} \right\rfloor + 1$	$\left\lfloor 2^{-e} \times v \right\rfloor$
$-32767.0 \leq v \leq 32767.0$	$-\left\lfloor \log_2 \frac{32767.0}{\lvert v \rvert} \right\rfloor$	$\left\lfloor 2^{-e} \times v \right\rfloor$
$v > 32767.0$ ou $v < -32767.0$	$\left\lfloor \log_2 \frac{\lvert v \rvert}{32767.0} \right\rfloor + 1$	$\left\lfloor 2^{-e} \times v \right\rfloor$

Exemples de codages

	Non signé	Signé
1.2345	$40452\ 2^{-15}$	$20226\ 2^{-14}$
107573	$58787\ 2^{1}$	$26893\ 2^{2}$
-3.14158	-	$-25736\ 2^{-13}$
$3\ 10^{8}$	$46875\ 2^{6}$	$23438\ 2^{7}$

Exemples de codages

	Non signé	Signé
1.2345	0x9E04	0x4F02
107573	0xD21B	0x690D
-3.14158	-	0x9B78
$3\ 10^8$	0xB71B	0x5B8E

Addition

$$0.6250 + 0.1211 \leftrightarrow$$

$$\langle 20480, -15 \rangle + \langle 31745, -18 \rangle =$$

$$\langle 20480, -15 \rangle + \langle 03968, -15 \rangle =$$

$$\langle 20480 + 3968, -15 \rangle =$$

$$\langle 24448, -15 \rangle \leftrightarrow 0.7461$$

Mais problème de débordement :

$$0.99 \leftrightarrow \langle 32440, -15 \rangle$$

$$\langle 32440, -15 \rangle + \langle 32440, -15 \rangle =$$

$\langle 64880, -15 \rangle$ càd un nombre négatif!

Multiplication

Calculer 0.5^2 en format Q15

$$0.5 \leftrightarrow \langle 16384, -15 \rangle \leftrightarrow 0x4000$$

$$\times \quad 0.5 \leftrightarrow \langle 16384, -15 \rangle \leftrightarrow 0x4000$$

$$\langle 268435456, -30 \rangle$$

Format Q30 à convertir en Q15

$$\text{or } 1.0 \leftrightarrow \langle 32768, -15 \rangle$$

$$\langle 268435456, -30 \rangle = \frac{\langle 268435456, -30 \rangle}{\langle 32768, -15 \rangle} =$$

$$\left\langle \frac{268435456}{32768}, -30 - (-15) \right\rangle =$$

$$\langle 8192, -15 \rangle \leftrightarrow 0.25 \text{ en Q15}$$

Circonférence

1.22 cm ≤ D ≤ 20.8 cm. Calculer la circonférence.

D est non signé et D < 32 = 2^5 ⇒

5 bits pour partie entière

11 bits pour partie fractionnaire

Donc format $\langle *, -11 \rangle$

$C = \pi D$

Fonction monotone croissante

Maximum pour D=20.8 → C=65.345

Donc le résultat en format <*,-9>

Circonférence (calcul)

$$C = \pi \times D$$

$$\pi \leftrightarrow \langle 51472, -14 \rangle$$

$$\langle 51472, -14 \rangle \times \langle d, -11 \rangle =$$

$$\langle 51472 \times d, -25 \rangle =$$

$$\left\langle \frac{51472 \times d}{65536}, -25 - (-16) \right\rangle =$$

$$\langle 51472 \times d >> 16, -9 \rangle$$

Circonférence (programme)

```
typedef unsigned short UWord;
typedef unsigned long ULong;
/* D non signé, <*,-11>
   retourne non signé, <*,-9> */
UWord circonf (UWord D)
{
     UWord x;
     x=(UWord)(51472L*(ULong)D)>>16;
     return x;
}
```

Conversion

Degré Farenheit → degré Celcius

$$°C = \frac{(°F - 32) \times 5}{9}$$

$-40 \leq F \leq 250$ Signé, <*,-7> c-à-d Q7

Fonction monotone croissante

$-40 \leq C \leq 121$ Signé, <*,-8> c-à-d Q8

$32 = 2^5 = 2^{12} \times 2^{-7} \leftrightarrow \langle 4096, -7 \rangle$

$5/9 \leftrightarrow \langle 18204, -15 \rangle$

Conversion (calcul)

$$(F-32)\times 5/9 \leftrightarrow$$

$$(\langle f,-7\rangle-\langle 4096,-7\rangle)\times\langle 18204,-15\rangle=$$

$$\langle f-4096,-7\rangle\times\langle 18204,-15\rangle=$$

$$\langle (f-4096)\times 18204,-22\rangle$$

Résultat Q22 qu'il faut convertir en format Q8

$$\langle ((f-4096)\times 18204)>>14,-8\rangle$$

Conversion (programme)

```
typedef short Word;


/* F est représenté en format Q7
   retourne un résultat en Q8 */
Word FtoC (Word f)
{
  return (((long)(f-4096)*18204)>>14);
}
```