Nombres flottants

I : INTRODUCTION : DE LA VIRGULE FIXE À LA VIRGULE FLOTTANTE

Point de vue mathématique : rappels sur les nombres

- Les nombres entiers : peuvent s'écrire sans aucun chiffre après la virgule.
- Les nombres décimaux : peuvent s'écrire *exactement* avec un nombre fini de chiffres après la virgule.
- Les nombres réels : peuvent s'écrire par une partie entière et une liste finie ou infinie de décimales.

Point de vue physique : rappels sur la notation scientifique : mantisse et puissance de 10

La notation scientifique permet de concentrer l'écriture d'un nombre en se débarassant des éventuels nombreux zéros. Cela se fait en décalant la virgule : on découpe le nombre en deux informations. D'un côté la mantisse dans l'intervalle $[1; 10[$ et de l'autre une puissance de 10.

- $127\ 000\ 000\ 000{,}0$ donne $1{,}27 \times 10^{11}$
- $56\ 050\ 000\ 000\ 000\ 000{,}0$ donne $5{,}605 \times 10^{16}$
- $0{,}000\ 000\ 000\ 006\ 42$ donne $6{,}42 \times 10^{-12}$
- $0{,}000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 987$ donne $9{,}87 \times 10^{-37}$

Quels nombres peut-on représenter avec seulement 24 chiffres (et 1 ou 2 signes) ?

(1) Représentation virgule fixe	(2) Représentation virgule décalée
+ 000 000 000 123, 456 789 000 000	+ 123 456 789 000 000 000 000 |+003
- 876 000 000 000, 000 000 000 000	- 876 000 000 000 000 000 000 |+012
- 000 000 000 000, 000 654 000 321	- 654 000 321 000 000 000 000 |-003
- 000 000 000 000, 000 000 123 456	- 123 456 000 000 000 000 000 |-006
+ 123 000 000 000, 123 456 789 666	X XXX XXX XXX XXX XXX XXX XXX |XXXX
+ 123 000 000 000, 123 456 789 000	+ 123 000 000 000 123 456 789 |+012
X XXX XXX XXX XXX, XXX XXX XXX XXX	+ 123 000 000 000 000 000 000 |+145
X XXX XXX XXX XXX, XXX XXX XXX XXX	+ 123 000 000 000 000 000 000 |-145

Inconvénient de la virgule décalée :

Inconvénient de la virgule fixe :

Bilan :

Représentation binaire en virgule fixe (hors programme)

La représentation à virgule fixe est un type de données correspondant à un nombre qui possède un nombre fixe de chiffres après la virgule. Les bits à gauche de la virgule représentent la partie entière du nombre. Les bits à droite de la virgule correspondent aux inverses d'une puissance de 2 et donnent la partie décimale.

Rappels des puissances de 2 usuelles :

Exposant	5	4	3	2	1	0	-1	-2	-3	-4
Puissance de deux	32	16	8	4	2	1	0,5	0,25	0,125	0,0625
Puissance de deux : écriture alternative							$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$

Quelques exemples avec 6 bits à gauche et 4 bits à droite de la virgule *en complément à deux* :

0`10110 1010` correspond à :

- Nombre positif à cause du bit de poids fort (0)
- Partie entière égale à $16 + 0 + 4 + 2 + 0 = 21$
- Partie décimale égale à $1 \times \frac{1}{2} + 0 \times \frac{1}{4} + 1 \times \frac{1}{8} + 0 \times \frac{1}{16} = 0{,}5 + 0 + 0{,}125 + 0 = 0{,}625$
- Au final : $21{,}625$

1`10100 0010` correspond à :

- Nombre négatif à cause du bit de poids fort (1)
- *On inverse donc tous les bits strictement à gauche du premier 1 en partant de la droite :* `001011 1110`
- Partie entière égale en valeur absolue à $0 + 8 + 0 + 2 + 1 = 11$
- Partie décimale égale à $1 \times \frac{1}{2} + 1 \times \frac{1}{4} + 1 \times \frac{1}{8} + 0 \times \frac{1}{16} = 0{,}5 + 0{,}25 + 0{,}125 = 0{,}875$
- Au final : $-11{,}875$

II Représentation approximative des nombres réels en virgule flottante : nombre flottant

1) Première approche

On procède comme pour l'écriture binaire des entiers en utilisant la norme IEEE-754 (qui n'est pas à connaître)

Précision	Encodage	Signe	Exposant	Mantisse	Valeur d'un nombre	Chiffres significatifs
Simple précision	32 bits	1 bit	8 bits	23 bits	$(-1)^S \times (1 + M) \times 2^{E-127}$	environ 7
Double précision	64 bits	1 bit	11 bits	52 bits	$(-1)^S \times (1 + M) \times 2^{E-1023}$	environ 16
Quadruple précision	128 bits	1 bit	15 bits	112 bits	$(-1)^S \times (1 + M) \times 2^{E-16383}$	environ 38

Remarque 1: le fait qu'on utilise $(1 + M)$ pour le nombre est expliqué plus bas.
Remarque 2: on peut observer que l'on n'utilise pas le complément à deux pour gérer le signe.

Méthode (Simple précision) : 10,3125
étape 1 : décomposition en "écriture scientifique base deux"

```
10.3125
8 + 2.3125
8 + 2 + 0.3125
8 + 2 + 0.25 + 0.0625
```

On obtient `1010,0101`.
Donc : $10{,}3125_{10} = 1010{,}0101_2 = 1{,}0100101_2 \times 2^3$

$$10{,}3125 = 1{,}0100101_2 \times 2^3$$

étape 2 : obtention de la mantisse

Puisque le résultat obtenu à l'étape précédente commencera systématiquement par 1, la mantisse ne garde pas la premier 1 puisque cela gaspillerait un bit (qui serait toujours positionné à 1).

La mantisse M est donc : `0100101 000 000 000 000 000 0`

étape 3 : obtention de l'exposant

$E - 127 = 3$

$E = 130$ que l'on représente en binaire sur 8 bits

$E = 128 + 0 + 0 + 0 + 0 + 0 + 2 + 0$

L'exposant est donc : `10000010`

étape 4 : obtention du signe

Le nombre est positif, son bit de signe est donc : `0`

Conclusion

La représentation de 10,3125 selon la norme IEE-754 est : `0 10000010 01001010000000000000000`

Exercice

Faire la même chose (préciser chaque étape) en simple précision pour :

$74,625$

74,625 = 64 + 0 + 0 + 8 + 0 + 2 + 0 + 0,625

74,625 = 64 + 0 + 0 + 8 + 0 + 2 + 0 + 0.5 + 0 + 0.125

$74,625 = 1001010,101_2$

$74,625 = 1,001010101_2 \times 2^6$

La mantisse sur 23 bits est donc 001010101 0 0000000000 000

L'exposant E vérifie E-127=6 et donc $E=133=128+0+0+0+0+4+0+1=10000101_2$

Le bit de signe est 0

La représentation de 74,625 selon la norme IEE-754 est : 0 10000101 00101010100000000000000

$-0,4375$

- 0,4375 = - 0 - 0,25 - 0,125 - 0.625

$- 0,4375 = - 0,0111_2$

$- 0,4375 = -1,11 \times 2^{-2}$

La mantisse sur 23 bits est donc 110 0000000000 0000000000

L'exposant E vérifie E-127=-2 et donc $E=125=0+64+32+16+8+4+0+1=01111101_2$

Le bit de signe est 1

La représentation de-0.4375 selon la norme IEE-754 est : 1 01111101 11000000000000000000000

$-1024,03125$ $(0{,}03125 = 2^{-5})$

- 1024,03125 = 1024 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0,03125

- 1024,03125 = 1024 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0.03125

$- 1024,03125 = 10000000000,00001_2$

$- 1024,03125 = 1,000000000000001_2 \times 2^{10}$

La mantisse sur 23 bits est donc 000000000000001 00000000

L'exposant E vérifie E-127=10 et donc $E=137=128+0+0+0+8+0+0+1=10001001_2$

Le bit de signe est 1

La représentation de -1024,03125 selon la norme IEE-754 est : 1 10001001 00000000000000100000000

Exercice

Écrire les résultats précédénts sous forme hexadécimale.

```
0 10000101 00101010100000000000000 --> 01000010 10010101 0100000 000000000
                                   --> 42 95 40 00
1 01111101 11000000000000000000000 --> 10111110 11100000 00000000 00000000
                                   --> BE E0 00 00
1 10001001 00000000000000100000000 --> 11000100 10000000 00000001 00000000
                                   --> C4 80 01 00
```

2) Seconde approche (pour l'étape 1)

On peut appliquer l'algorithme des divisions successives pour la partie entière.
Et l'algorithme des multiplications successives pour la partie décimale.

10,3125	
Partie entière : *remontée* des divisions successives	Partie décimale : *descente* des multiplications successives
10//2 = 5 et 10%2 = 0 ^ 5//2 = 2 et 5%2 = 1 | 2//2 = 1 et 2%2 = 0 | 1//2 = 0 et 1%2 = 1 | L'écriture binaire de 10 est 1010	0,3125 * 2 = 0,625 = 0 + 0,625 | 0,625 * 2 = 1,25 = 1 + 0,25 | 0,25 * 2 = 0,5 = 0 + 0,5 | 0,5 * 2 = 1,0 = 1 + 0 v L'écriture binaire de ,3125 est ,0101
L'écriture binaire de 10,3125 est 1010,0101	

Exercice facultatif :

Appliquer cette méthode pour obtenir la représentation en virgule flottante :
1056,4140625

$1,00001000000110101_2 \times 2^{10}$
0 10001001 00001000000110101000000
01000100 10000100 00001101 01000000
44 84 0D 40

$-49{,}59375$

$1,10001100110000000000000 \times 2^5$
1 10000100 10001100110000000000000
11000010 01000110 01100000 00000000
C2 46 60 00

On donnera les résultats sous forme hexadécimale.

Exercice :

a) Obtenir la représentation en virgule flottante du nombre 14,25
On trouve 0 10000010 11001000000000000000000

a) En déduire la représentation en virgule flottante du nombre 28,5
Il suffit de multiplier par deux le nombre précédent.
Ce qui revient à augmenter de 1 l'exposant.
On obtient donc : 0 10000011 11001000000000000000000

Exercices : sur notebook

Conclusion : Ce qui est mentionné au programme

Représentation approximative des nombres réels : notion de nombre flottant	Calculer sur quelques exemples la représentation de nombres réels : 0.1, 0.25 ou 1/3.	0.2 + 0.1 n'est pas égal à 0.3. Il faut éviter de tester l'égalité de deux flottants. Aucune connaissance précise de la norme IEEE-754 n'est exigible.