Nombres Rationnels

Représentation à virgule flottante

Un nombre est rationnel s'il peut s'écrire sous la forme d'un quotient de deux entiers. Inversement, un nombre est irrationnel lorsqu'il n'est pas rationnel, c'est à dire qu'il ne peut s'écrire sous forme de fraction. Exemple Le nombre (pi) est irrationnel (Π = 3,14159265…).

Les nombres réels ne peuvent pas tous être représentés, Ils sont approximés par les nombres rationnels les plus proches. On peut dire que les nombres rationnels sont un sous-ensemble des réels.

Pour représenter un nombre rationnel dans une machine informatique, il a été nécessaire de trouver une écriture des nombres compatible avec la taille mémoire qu'on lui accorde. On va privilégié la notation scientifique et l'écriture en virgule flottante.

1. Principe de la représentation

Le nombre 9,750 se trouvera mémorisé sous la forme suivante :

9,750 = 9 + 0,75

conversion en binaire de la partie entière :
9 → 1001

conversion en binaire de la partie fractionnaire: 0,75 = 3/4 = 1/2 + 1/4
0,75 * 2 = 1,5 → **1**$*2^{-1}$
0,5 * 2 = 1,0 → **1**$*2^{-2}$
0 * 2 = 0 → **0**$*2^{-3}$
0 * 2 = 0 → **0**$*2^{-4}$
etc

$9{,}75_{(10)} = 1001{,}1100_{(2)}$

En utilisant cette notion de virgule, notre nombre peut s'écrire de la manière ci-après :
$N=1001{,}1100\times2^{0}$
$N=100{,}11100\times2^{1}$
$N=10{,}011100\times2^{2}$
$N=1{,}0011100\times2^{3}$

La dernière expression présente l'avantage de représenter la grandeur par un nombre supérieur à 1 et inférieur à 2 multiplié par une puissance de 2.
L'exposant 3 est bien entendu représentatif de la position de la virgule.
Donc pour définir totalement notre information (9,750) il faudra dans le système de représentation **deux termes** :

Le terme 10011100 appelé **Mantisse** (M),
le terme 11 appelé **Exposant** (E).

Dans une machine les informations sont représentées en virgule flottante, elles se présenteront de la manière suivante :
1001110011.
10011100 est la Mantisse et correspond à notre nombre N de départ (1001,1100) mais sans "écrire ou indiquer" la virgule,
11 est l'Exposant (11 en binaire vaut 3 en décimal) et donne la position de la virgule.
On retrouve ainsi notre nombre :
$N=1,00111\times2^{3}$
N=1001,11

Les nombres réels ne peuvent pas tous être représentés, Ils sont approximés par les nombres à virgule flottante les plus proches. On peut dire que les nombres à virgule flottante sont un sous-ensemble des réels.

Il vient que, lors de la conception du programme de traitement des nombres, il faudra déterminer la plage de représentation des nombres et la précision désirée. On conviendra alors du nombre de bits pour représenter la Mantisse qui donnera la précision sur les nombres, et du nombre de bits pour l'Exposant qui procurera l'intervalle des nombres représentables.

2. Les Normes de représentation IEEE 754

Il y a nécessité de trouver une norme pour la représentation en mémoire :

L'exposant peut être positif ou négatif. Cependant, la représentation habituelle des nombres signés (complément à 2) rendrait la comparaison entre les nombres flottants un peu plus difficile. Pour régler ce problème, l'exposant est «biaisé», afin de le stocker sous forme d'un nombre non signé.

3 formats sont normalisés:

Sur 32 bits → simple précision mot clé **float**
- 1 bit pour le signe
- 8 bits pour l'exposant Biais est l'excès 127 compris entre (−126 à 127)
- 23 bits pour la mantisse (un bit 1 implicite)

Sur 64 bits → double précision mot clé **double**
- 1 bit pour le signe
- 11 bits pour l'exposant Biais est l'excès 1023 compris entre (−1022 à 1023)
- 52 bits pour la mantisse (un bit 1 implicite)

Sur 80 bits → précision étendue mots clés **long double**
- 1 bit pour le signe
- 15 bits pour l'exposant Biais est l'excès 16 383
- 64 bits pour la mantisse (pas de bis implicite)

Exemple avec **13,8125**

- S = 0
- M = 1011101 – sur 23 bits : M = 101 1101 0000 0000 0000 0000
- E = 127+3 = 130 – sur 8 bits : E = 1000 0010
- 13,8125 en IEEE754 simple précision :
- 0100 0001 0101 1101 0000 0000 0000 0000
- En hexa : 41 5D 00 00

Vérification :

- signe S = 0
- mantisse M = 101 1101 0000 0000 0000 0000 → 6 094 848
- format de la mantisse n = 23
- exposant biaisé E = 130
- biais = 127

$$(-1)^{S} \times (1+ M / 2^{n}) \times 2^{(E-biais)}$$

- $(-1)^{0} = 1$
- $1+ 6094848/ 2^{23} = 1,7265625$
- $2^{(130-127)} = 2^{3} = 8$

- $1 \times 1,7265625 \times 8 = 13,8125$

Vous pouvez vérifier vos résultats avec le site suivant.

https://www.h-schmidt.net/FloatConverter/IEEE754.html