Représentation des nombres flottants

Ensimag 1A

2022

Système de numérotation binaire

Dans le système décimal, un nombre s'écrit

$$9,90625 = 9 \times 10^0 + 9 \times 10^{-1} + 0 \times 10^{-2} + 6 \times 10^{-3} + 2 \times 10^{-4} + 5 \times 10^{-5}$$

Le même nombre s'écrit dans le système binaire (i.e. en base 2) sous la forme d'une suite de 0 et 1 (bits) :

$$(1001,11101)_2 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 0 \times 2^{-4} + 1 \times 2^{-5}$$

Nombre flottant

Définition

Une représentation en virgule flottante dans une base arithmétique donnée (généralement 2 sur ordinateur) comprend 3 éléments :

- un signe ± 1,
- l'exposant (entier relatif, variable)
- la mantisse, qui correspond aux chiffres après la virgule.

Dans ce qui suit, nous allons donner un bref aperçu du système des nombres flottants en "double précision", représentés avec la norme IEEE 754 ("Institute of Electrical and Electronics Engineers").

Système standard double précision

- Les nombres sont représentés sur des blocs de mémoire de taille 64 bits. Ce stockage comprend :

  1 bit de signe, 11 bits pour l'exposant, 52 bits pour la mantisse

- Nombres flottants normalisés :
  $(-1)^{\text{signe}} (1.b_{51}b_{50}\dots b_1 b_0)_2 \times 2^{e-1023}$ avec $0 \leq e \leq 2047$
- Plus petit flottant > 0 normalisé : $2^{-1022} \approx 2,225 \times 10^{-308}$
- Plus grand flottant (fini) : $\approx 1,797 \times 10^{308}$
- Si une opération donne un nombre non représentable dans la norme IEEE-754, il faut l'arrondir (souvent au plus proche nombre représentable).
- Le plus petit flottant > 1 est $1 + \%\text{eps}$ avec
  $\%\text{eps} = 2^{-52} \approx 2,22 \times 10^{-16}$

Exemples de calculs en double précision sous Scilab

```
-->  format(20)                  // résultats avec 20 caractères max
-->  1+%eps                      // ans=1.0000000000000022
-->  1+(%eps/2+%eps/2)           // ans=1.0000000000000022
-->  1+%eps/2                    // ans=1.
-->  (1+%eps/2)+%eps/2           // ans=1.
-->  1+%eps/2+%eps/2             // ans=1.
-->  (-1+1+%eps/2)/%eps          // ans=0.5
-->  (1+%eps/2-1)/%eps           // ans=0 ("annulation catastrophique").
```