



## 2/ Limites du codage des flottants

### 2.1/ Plus grand, plus petit nombre

$$x = 2^{1023} \left( 1 + \sum_{i=1}^{52} 2^{-i} \right) = 2^{1023} + 2^{971} \cdot \sum_{i=0}^{51} 2^i = 2^{1023} + 2^{971} (2^{52} - 1) = 2^{1023} + 2^{1023} - 2^{971} \\ = (2^{1024} - 2^{971}) \approx \pm 1,7976931348623157 \times 10^{308}$$

### 2.2/ Nombres proches de 0

Selon la règle générale :

Le plus petit nombre positif peut être obtenu, selon la règle, pour l'exposant minimum (-1022) et la mantisse minimale (tous les  $m_i$  à 0). Alors :

$$x = 2^{-1022} \approx 2,2250738585072020 \times 10^{-308}$$

Si  $e = -1023$ , la règle ne s'applique plus. On considère que le nombre est nul.

L'exception est basée sur le fait que  $x$  est suffisamment proche de 0 pour être codé par :

$$x = (-1)^s \cdot 2^{-1022} \sum m_i \cdot 2^{-i}$$

### 2.3/ Arrondi intrinsèque

Il est rare que le résultat d'un calcul faisant intervenir deux nombres à virgule flottante donne un résultat représentable exactement sur 64 bits. Même sans effectuer de calcul, la plupart des nombres décimaux ne sont pas représentables exactement dans ce format.

Ainsi, par exemple, le nombre 0,4 admet pour développement en base 2 :

$$\frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^6} + \frac{1}{2^7} + \frac{1}{2^{10}} + \frac{1}{2^{11}} + \frac{1}{2^{14}} + \frac{1}{2^{15}} + \dots = \underline{0,011001100110011\dots}$$

Il faudrait donc un nombre de bits infini pour le représenter.

Si nous utilisons 32 bits, une valeur approchée serait : **0-01111101-10011001100110011001101** ou  $2^{-2} \left( 1 + \frac{1}{2} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{512} + \frac{1}{4096} + \frac{1}{8192} + \frac{1}{65536} + \frac{1}{131072} + \frac{1}{2097152} + \frac{1}{4194304} + \frac{1}{16777216} \right) \approx 0,39970703423023224$

Il n'existe pas de représentation flottante plus proche de 0,4.

Le codage implique donc une imprécision sur le nombre codé.

Il ne pourra être représenté qu'avec :

- la précision des 54 chiffres binaires significatifs de la mantisse ;
- soit une erreur d'arrondi relative de  $2^{-55} = 2,8 \cdot 10^{-17} \approx 10^{-16}$  ;
- soit 16 chiffres décimaux significatifs.

2.4/ Figures explicatives

