Représentation des nombres entiers

Définition Écriture d'un nombre dans une base

Dans un système de numération en base B, un nombre noté N_B peut s'écrire sous la forme : $N_B = \sum_{k=0}^{n} a_k \cdot B^k$ s'écrit symboliquement sous la forme : $N_B = \underbrace{(a_n a_{n-1} \cdots a_2 a_1 a_0)_B}_{n+1 \text{ chiffres}}$ On note :

- B : la base ou nombre de chiffres différents qu'utilise le système de numération ;
- a_k : chiffre de rang k ;
- B^k la pondération associée à a_k.

Représentation des nombres entiers relatifs

Méthode Pour représenter l'opposé d'un nombre positif par son complément à deux, on inverse les bits 0 et 1 et on ajoute 1 au mot binaire obtenu.

Représentation des nombres réels

Méthode Conversion d'une partie fractionnaire en binaire

1. On multiplie la partie fractionnaire par 2.
2. La partie entière obtenue représente le poids binaire (limité aux seules valeurs 0 ou 1).
3. La partie fractionnaire restante est à nouveau multipliée par 2.
4. On procède ainsi de suite jusqu'à ce qu'il n'y ait plus de partie fractionnaire ou que le nombre de bits obtenus correspond à la taille du mot mémoire dans lequel on stocke cette partie.

Pour représenter des réels, nombres pouvant être positifs, nuls, négatifs et non entiers, on utilise la représentation en virgule flottante (*float* en anglais) qui fait correspondre au nombre 3 informations :

$$-243,25_{(10)} = \underbrace{-}_{1} \; \underbrace{0,24325}_{2} \cdot 10^{\overbrace{3}^{3}}$$

On appelle alors :

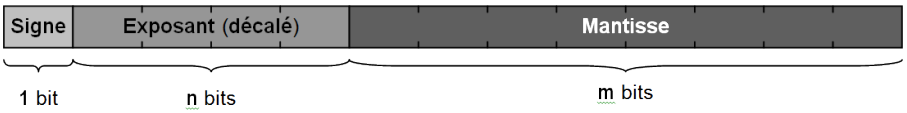
1. le signe (positif ou négatif) ;
2. la mantisse (nombre de chiffres significatifs) ;
3. l'exposant : puissance à laquelle la base est élevée.

Sous cette forme normalisée, il suffit de mémoriser le signe, l'exposant et la mantisse pour avoir une représentation du nombre en base 10. Il n'est pas utile de mémoriser le 0 avant la virgule puisque tous les nombres vont commencer par 0. En faisant varier l'exposant, on fait « flotter » la virgule décimale.

C'est cette méthode que l'on va adapter pour coder les réels en binaire naturel. Il faut au préalable les écrire sous la forme (norme IEEE 754 – Institute of Electrical and Electronics Engineers) :

$$\text{signe } 1, \text{mantisse x } 2^{\text{exposant}}$$

Le mot binaire obtenu sera la juxtaposition de 3 parties :



Le tableau décrit la répartition des bits selon le type de précision : la taille de la mantisse (m bits) donne la précision mais suivant la valeur de l'exposant, la précision sera totalement différente.
Ainsi :

	Signe	Exposant	Mantisse
Simple précision – 32 bits	1	8	23
Double précision – 64 bits	1	11	52
Précision étendue – 80 bits	1	15	64

- erreur relative : 2^{-m} (poids du dernier bit)
- erreur absolue : erreur relative * 2^{exposant}

Simple précision : $2^{-23} = 1,192... * 10^{-7}$ Double précision : $2^{-52} = 2,220... * 10^{-16}$

Procédure de conversion de réel en binaire (hexadécimale)

Méthode

1. Convertir en binaire les partie entière et fractionnaire du nombre sans tenir compte du signe.
2. Décaler la virgule vers la gauche pour le mettre sous la forme normalisée (IEEE 754).
3. Codage du nombre réel avec les conventions suivantes :
   - signe = 1 : Nombre négatif (Signe = 0 : Nombre positif) ;
   - le chiffre 1 avant la virgule étant invariant pour la forme normalisée, il n'est pas codé ;
   - on utilise un exposant décalé au lieu de l'exposant simple (complément sur octet). Ainsi, on ajoute à l'exposant simple la valeur 127 en simple précision et 1023 en double précision (c'est à dire $2^{n-1}-1$ où n est le nombre de bits de l'exposant) ;
   - la mantisse est complétée à droite avec des zéros.

Représentation chaînes de caractères