

## TD2 : Représentation des flottants en machine, norme IEEE 754

Le type double du Cest défini dans la norme IEEE 754  $^{\rm 1}$  comme un *nombre à virgule flottante en double précision*. Il est codé sur 8 octets.

Cette norme repose sur un double constat :

- on manipule rarement de très grands nombres,
- plus les nombres manipulés sont grands, moins on se soucie de la précision.

Un nombre flottant est composé de trois éléments : son signe, son exposant et sa partie significative (liée à la mantisse), comme montré dans la figure ci-dessous.



Les valeurs de s, e et m sont lues en base 2. On appelle biais la valeur

$$b = 2^{10} - 1 = 1023$$
,

où l'exposant 10 dans le calcul est obtenu par 11-1, où 11 est le nombre de bits qui permettent de coder e. Ce biais sert à représenter des exposants négatifs sans passer par le complément à 2 (qui rendrait les comparaisons compliquées).

Le réel représenté par un nombre à virgule flottante en double précision est :

valeur de <i>e</i>	valeur de <i>m</i>	valeur du réel représenté	
0	0	±0	zéros
0	≠0	$(-1)^s \times m * 2^{-52} \times 2^{-(b-1)}$	flottant dénormalisé
$2^{11}-1$	0	±∞	
$2^{11}-1$	≠0	NaN (Not a Number)	
autre	quelconque	$(-1)^s \times \underbrace{(1+m*2^{-52})}_{\text{partie significative}} \times 2^{e-b}$	flottant <i>normalisé</i>

Le nombre  $1 + m * 2^{-52}$  doit être vu comme 1 suivi d'une virgule puis des chiffres de m (en base 2), tandis que le nombre  $m * 2^{-52}$  doit être vu comme 0 suivi d'une virgule puis des chiffres de m (en base 2).

**Exercice 1** Quels réels (en base 10) sont représentés de la façon suivante en nombres à virgule flottante double précision par la norme IEEE 754?

1	1	0	0	0	0	0	0	0	0	0	0	1	1	0.	…0
0 1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	00

**Exercice 2** Donner la représentation en nombre à virgule flottante en double précision (en suivant la norme IEEE 754) des nombres 11.28125 et 1035.28125.

Exercice 3 Expliquer pourquoi le test

s'évalue à faux.

MP2I 1 TD

<sup>1.</sup> IEEE: Institute of Electrical and Electronics Engineers