IEEE754

November 27, 2019

1 Représentation approximative des nombres réels

1.1 Conversion d'un nombre décimal en binaire - virgule fixe

1.2 Conversion d'un nombre décimal en binaire - virgule flottante

1.2.1 Analogie avec la notation scientifique en base 10

1.2.2 Forme normalisée - standard IEEE754

1.2.3 Exemple

1.3 Cas particuliers

1.4 Conversion d'un nombre décimal en binaire - virgule fixe

La méthode consiste à décomposer la partie entière et la partie fractionnaire suivant les puissances de deux (puissances positives pour la partie entière et puissances négatives pour la partie fractionnaire). Ces deux parties étant séparées par la virgule (qui est fixe dans ce cas). La conversion binaire $\rightarrow$ décimale est alors évidente.
Par exemple la conversion de 101,1101 donne:

2^2	2^1	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}
1	0	1	1	1	0	1

soit 5,8125
L'Algorithme décimal $\rightarrow$ binaire consiste en:

1) Convertir la partie entière (voir cours précédents)

2) Convertir la partie fractionnaire en adoptant l'algorithme suivant:

- multiplier la partie fractionnaire par deux;
- extraire la partie entière qui donne un des bits de la partie fractionnaire;

- répéter tant que la partie fractionnaire restante est différente de zéro.

1.5 Conversion d'un nombre décimal en binaire - virgule flottante

La deuxième façon d'encoder un nombre décimal est inspirée de la notation scientifique: $\pm m \times 10^n$ mais en base deux, c'est-à-dire $(-1)^s \times 1, f \times 2^{e-biais}$. Il s'agit des nombres à virgule flottante.
Ce format est constitué de trois parties essentielles:
* 1 bit de signe s; * un exposant; pour éviter d'avoir que des grandes valeur, on décale cet exposant d'une

signe (s)	exposant (e)	partie fractionnaire ou mantisse (f)

certaine valeur (biais);
* une partie fractionnaire f appelé encore mantisse.

La représentation des nombres à virgule flottante est entièrement définie dans la norme IEEE 754. Celle-ci prévoit une représentation simple précision sur 32 bits ou double précision sur 64 bits.

	exposant (e)	fraction (f)	valeur
32 bits	8 bits	23 bits	$(-1)^s \times 1, f \times 2^{e-127}$
64 bits	11 bits	52 bits	$(-1)^s \times 1, f \times 2^{e-1023}$