

TD n° 1 : REPRESENTATION DES NOMBRES

1. Représentation des entiers

Soient des entiers en représentation en complément à 2 où un nombre N correspond à

$$N = -a_{n-1}2^{n-1} + \sum_{i=0}^{i=n-2} a_i 2^i$$

Rappel : dans cette représentation, on peut obtenir l'opposé -N d'un nombre N en prenant le complément à 1 de N (complémentation bit à bit), puis en ajoutant 1.

Additions et soustractions sur n bits

Faire les additions suivantes sur un octet et indiquer si le résultat est correct ou s'il y a débordement :

15 _H + 48 _H	72 _H + F9 _H
F5 _H + AF _H	47 _H + 3A _H
15 _H + A3 _H	81 _H + 95 _H

Additions n bits + p bits (avec p < n)

Faire les additions suivantes :

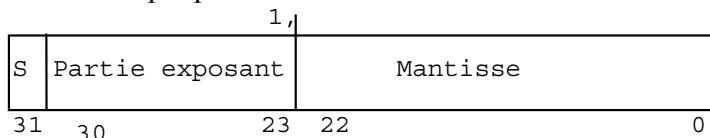
1560 _H + 48 _H	7200 _H + F9 _H
F500 _H + AF _H	47F0 _H + 3A _H
15FF _H + A3 _H	8100 _H + 95 _H

2. Nombres en représentation “virgule flottante” (Standard IEEE 754)

Le codage est donné dans la table ci-dessous où s est le bit de signe, e est la partie exposant et f représente la partie fractionnaire après le 1 implicite.

PE	f	représente
0 ; s ±1	0	±0
0	≠ 0	s x 0, f x 2 ⁻¹²⁶
0 < PE < 255	quelconque	s x 1, f x 2 ^(PE-127)
255	0	±∞
255	≠ 0	NaN

Format simple précision:



1) Quels nombres simple précision correspondent aux mots de 32 bits suivants :

- a) 41300000_H
- b) 41E00000_H
- c) BF800001_H

d) 00A00000H

2) Ecrire 1 et -1000 de façon normalisée.

3) Donner le plus grand positif et son prédécesseur, indiquer leur écart; le plus petit positif normalisé et dénormalisés; le plus grand et le plus petit négatif.

3. Conversions

Soient les déclarations C suivantes :

```
int x ;  
float f,  
double d
```

où x est un entier sur 32 bits, f est un flottant 32 bits (simple précision) et d un flottant 64 bits double précision.

On utilise les opérateurs de conversion de C.

Indiquer si les assertions suivantes sont vraies ou fausses, en justifiant

- a) $x == (\text{int})(\text{float}) x$
- b) $x == (\text{int})(\text{double}) x$
- c) $f == (\text{float})(\text{double}) f$
- d) $d == (\text{float}) d$
- e) $f == -(-f)$
- f) $2/3 == 2/3.0$
- g) $d < 0.0 \Rightarrow (2*d) < 0.0$
- h) $d > f \Rightarrow -f < d$
- i) $d*d \geq 0.0$
- j) $(d+f) - d == f$

4. Représentation virgule fixe

Pour les nombres en virgule fixe, on appellera $Q_{m,f}$ un format avec m bits de mantisse (avant la virgule) et f bits de fraction (après la virgule). Le nombre total de bits est m+f.

Q1) Avec les formats $Q_{12,4}$ et $Q_{1,15}$ en complément à 2

- quelle est la plus grande valeur positive représentable ?
- Quelle est la plus petite valeur positive représentable ?
- Quelle valeur (décimale) représente une configuration « plein 1 » ?
-

Q2) Soit le format $Q_{1,7}$ en complément à 2

Quels sont les nombres décimaux correspondants à

01001101 ₂ ,	01111001 ₂ ,
11100100 ₂ ,	10001011 ₂

Q3) Avec le format $Q_{1,7}$ en complément à 2, effectuer les opérations suivantes et indiquer les cas de débordement

01001101₂+11100100₂ =
01111001₂+10001011₂ =
01001101₂-11100100₂ =
10001011₂-00110111₂ =