IUT de Montpellier, Dép. Informatique
R1.03 - Introduction à l'architecture des ordinateurs (1ère année) Responsable : V. Poupet

TD nº 3 - Nombres à virgule flottante

La norme IEEE 754

Un nombre décimal est représenté en simple précision (32 bits) ou en double précision (64 bits) de la façon suivante :

s : signe (1 bit)	e : exposant biaisé (8 ou 11 bits)	m : mantisse (23 ou 52 bits)

Le nombre n ainsi représenté est :
— si $e \neq 0$ et $e \neq 255(11111111)$ (ou $\neq 2047(11111111111)$ en double précision) :

$$n = (-1)^s \times 2^{e-\text{biais}} \times 1, \ldots \text{mantisse} \ldots$$

où $bin(m)$ est la représentation binaire de la mantisse, et le biais vaut 127 (01111111) en simple précision et 1023 (01111111111) en double précision ;
— si $e = 0$:

$$n = (-1)^s \times 2^{1-\text{biais}} \times 0, \ldots \text{mantisse} \ldots$$

— si $e = 255(11111111)$ (ou $= 2047(11111111111)$ en double précision), le nombre représenté est l'infini si la mantisse est nulle et NaN (*not a number*) sinon.

Exercice 1. *Représentation des nombres à virgule flottante*

1 Donner l'écriture des nombres 1, 2, 3, 4 et 5 au format IEEE 754 en simple précision.

2 Si l'exposant biaisé est différent de 0 et 255 (ou 2047 en double précision), quels sont les plus petits et les plus grands nombres strictement positifs représentables en simple et double précision ?

3 Même question avec un exposant biaisé valant 0. Quel est le plus petit nombre strictement positif représentable ?

4 Donnez les représentations en simple précision à virgule flottante de 2^5 et $2,125$.

5 Quelle est la représentation en simple précision à virgule flottante de $\frac{1}{10}$? Quelle est l'erreur obtenue ?

6 Même question pour $\frac{1}{5}$.

Exercice 2. *Perte d'information en arithmétique flottante*

(vous pouvez écrire de petits programmes sur machine pour vous aider)

Soient les nombres

$$A = 0\ 10001110\ 00000000000000000000000$$

$$B = 0\ 10011010\ 00000000000000000000000$$

au format IEEE 754 en simple précision.

1 Donnez les représentations des nombres
a. $C = A + 1$
b. $D = A + B$
c. $E = B + C$

Voici un algorithme :

```
A <- 1
tant que (((A + 1) - A) - 1) = 0 faire A <- 2 * A
B <- 1
tant que (((A + B) - A) - B) <> 0 faire B <- B + 1
```

2. Expliquez pourquoi la première boucle s'arrête.
3. Expliquez pourquoi la seconde boucle s'exécute au moins une fois.
4. Que contient la variable B en fin d'exécution ?

On veut maintenant calculer en simple précision la somme

$$S_n = \sum_{i=1}^{n} \frac{1}{i}$$

Nous proposons d'écrire deux variantes : une où les termes sont ajoutés du plus petit au plus grand et une autre où les termes sont ajoutés du plus grand au plus petit.

5. Expliquez pourquoi les résultats peuvent être différents entre les deux variantes de l'algorithme.
6. Laquelle des deux variantes donne le résultat le plus précis ?

Exercice 3. *Secret Robot Internet*