TD 2 : Architecture des Ordinateurs, Les Nombres Flottants

Première année d'IUT Informatique

1 Les nombres en 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(1)	exp biaisé (8 ou 11)	... mantisse (23 ou 52)

Le nombre ainsi représenté est :

si, exp biaisé $\neq 0$ et $\neq 255(11111111)$ (ou $\neq 2047(11111111111)$ en double précision) :

$$(-1)^s \times 2^{(\text{exp biaisé})-\text{biais}} \times 1, ...\text{mantisse}...$$

Le biais vaut 127 (01111111) en simple précision, et 1023 (01111111111) en double précision.

si, exp biaisé $= 0$:

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

si, exp biaisé $= 255(11111111)$ (ou $= 2047(11111111111)$ en double précision) :
alors le nombre représente l'infini si la mantisse est nulle, ou NaN (not a number) sinon.

2 Exercices

1. Donner l'écriture des nombres $1, 2, 3, 4$ et 5 au format IEEE 754 en simple précision.
2. Si exp biaisé $\neq 0$ et $\neq 255(11111111)$ (ou $\neq 2047(11111111111)$ en double précision), quels sont les plus petits et les plus grands nombres positifs représentables en simple et double précision ?
3. Même question si exp biaisé $= 0$. Et quel est le plus petit nombre positif non nul représentable ?
4. Donner la représentation simple précision virgule flottante de 2^5 puis de $2,125$.
5. Quelle est la représentation simple précision virgule flottante de $\frac{1}{10}$? Quelle est l'erreur obtenue ?
6. Même question pour $\frac{1}{5}$.

3 Perte d'information en arithmétique flottante

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

1. Soient les nombres $A = 0\ 10001110\ 00000000000000000000000$ et $B = 0\ 10011010\ 00000000000000000000000$ au format IEEE 754 en simple précision. Donner les représentations des nombres :
   (a) $C = A + 1$
   (b) $D = A + B$
   (c) $E = B + C$
2. 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
```

   (a) Pour quelle raison la première boucle s´arrête ?
   (b) Pour quelle raison la seconde boucle s´exécute-t-elle au moins une fois ?
   (c) Que contient la variable B en fin d´exécution ?

3. Nous désirons 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 cumulés du plus petit au plus grand, et l'autre où l'on cumule dans l'ordre inverse.
   (a) Si les résultats obtenus diffèrent d'une variante à l´autre, quelle peut en être la raison ?
   (b) Quelle est la plus précise des deux variantes ?
4. Nous voulons évaluer $I_n = \int_0^1 x^n e^{-x} dx$. Pour ceci, nous effectuons une intégration par parties, nous obtenons $I_n = n * I_{n-1} - e^{-1}$ avec $I_0 = 1 - e^{-1}$.
   Nous proposons deux implantations pour évaluer I_n avec $n < 100$. La première implantation reprend la récurrence, la seconde inverse le sens de la récurrence :

```
VERSION I                                VERSION II
InvE <- .36787944117144232159      InvE <- .36787944117144232159
I <- 1- InvE                       I <- 22
Pour k = 1 à n faire               Pour k = 150 à n faire
     I <- k * I - InvE                  I <- ( I +  InvE ) / k
```

   Pensez-vous obtenir dans les deux cas une bonne approximation du résultat ?