Travaux dirigés n° 3

Représentation des réels

Rappels et compléments

Considérons le codage des réels respectant la norme IEEE754 (vue en cours) :
- 1 bit de signe S
- k bits d'exposant E (obtenu par la méthode du biais, représenté comme non signé [1])
- n bits de mantisse M

Taille des champs :
- flottants simple précision (4 octets) : 1 bit de signe, 8 bits d'exposant, 23 bits de mantisse
- flottants double précision (8 octets) : 1 bit de signe, 11 bits d'exposant, 52 bits de mantisse

La valeur du biais (pour l'exposant) est égale à $2^{k-1} - 1$:
- flottants simple précision : biais de $2^{8-1} - 1 = 127$
- flottants double précision : biais de $2^{11-1} - 1 = 1023$

Représentations normalisées :
La plage des représentations "normalisées" est fixée par la valeur de l'exposant : e' doit être compris entre 1 et $2^k - 2$ ($2^8 - 2 = 254$ en simple précision). La valeur du flottant alors est égale à $(-1)^s.(1,m)_2.2^{e'-biais}$, soit $(-1)^s.(1,m)_2.2^{e'-127}$ pour les flottants simple précision.

Cas particuliers :
Les cas particuliers sont les cas pour lesquels $e' = 0$ ou $e' = 2^k - 1$ ($e' = 2^8 - 1 = 255$ en simple précision) :
- $e' = 2^k - 1$, $M = [0...0]$: la valeur représentée est l'infini (+ ou - l'infini selon le bit de signe)
- $e' = 2^k - 1$, $M \neq [0...0]$: la valeur représentée est NaN (*Not a Number*). La valeur de M sert à coder le type d'erreur.
- $e' = 0$, $M = [0...0]$: la valeur représentée est 0
- $e' = 0$, $M \neq [0...0]$: la valeur représentée est $\pm(0,m)_2.2^{-126}$ ($\pm$: selon le bit de signe)

Pour ce TD, nous considérons le codage des réels respectant la norme IEEE754 sur les flottants simple précision (4 octets).
Par convention, nous noterons $[...]_{X1,2}$ et $[...]_{X2,2}$ les représentations de flottants simple précision (respectivement double précision) selon la norme IEEE754, ici en binaire (même chose pour les écritures hexadecimales : $[...]_{X1,16}$ et $[...]_{X2,16}$).

1. E est la représentation signée par la méthode du biais de l'exposant e
c'est-à-dire la représentation non signée de l'exposant biaisé e'

Exercice 1 (préliminaires)
On considère des valeurs entières non signées représentées sur un octet :
- effectuez les calculs à partir des représentations : addition, soustraction, multiplication, division
- vérifiez en base 10

1. — $A = [00111110]_2$
   — $B = [00101011]_2$
2. — $C = [ef]_{16}$
   — $D = [a9]_{16}$

Exercice 2 (décodage)
Donner la valeur exacte et une valeur approchée des représentations suivantes :
- $R_1 = [00000000\ 10000000\ 00000000\ 00000000]_{X1,2}$
- $R_2 = [10111111\ 10000000\ 00000000\ 00000000]_{X1,2}$
- $R_3 = [98\ 58\ 00\ 00]_{X1,16}$
- $R_4 = [61\ 18\ 00\ 00]_{X1,16}$
- $R_5 = [80\ 00\ 08\ 00]_{X1,16}$
- $R_6 = [10000000\ 00000000\ 00000000\ 00000000]_{X1,2}$

Exercice 3 (codage)
Donner la représentation IEEE754 simple précision des valeurs suivantes :
- $v_1 = 256$
- $v_2 = -118,375$
- $v_3 = -12,7734375.2^{-70}$
- $v_4 = 9.2^{-143}$
- $v_5 = -84,1.2^{67}$
- $v_6 = 0,15.2^{-125}$
- $v_7 = 262144,03$ (indication : $262144/1024 = 256$)
- $v_8 = -27.2^{-151}$
- $v_9 = 2^{80} + 2^{79} + 2^{78} + 2^{77} + \ldots + 2^{42} + 2^{41} + 2^{40}$

Exercice 4 (opérations)
Dans cet exercice, on donne les représentations (IEEE754 simple précision) :
- effectuez les calculs à partir des représentations (éventuellement après recallage)
- vérifiez en base 10

1. — $A = [00111110\ 10100000\ 00000000\ 00000000]_{X1,2}$
   — $B = [00111110\ 01000000\ 00000000\ 00000000]_{X1,2}$
   — opérations : addition, soustraction, multiplication, division
2. — $C = [53\ 88\ 00\ 00]_{X1,16}$
   — $D = [E0\ 28\ 00\ 00]_{X1,16}$
   — opérations : addition, soustraction, multiplication
3. — $E = [96\ 60\ 00\ 00]_{X1,16}$
   — $F = [28\ 20\ 00\ 00]_{X1,16}$
   — opération : multiplication