Université de Nantes
Région des Pays de La Loire

Feuille de travaux dirigés n° 2
Arithmétique entière et arithmétique flottante

Exercice 2.1

Effectuer les calculs suivants directement en binaire :

Additions	Soustractions	Multiplications	Divisions
$1011010 + 1110101$	$1010011 - 1111$	111×1111	$100110 \div 110$
$111010 + 110110$	$110101 - 1001$	1010×11001	$110101 \div 1010$
$1111111 + 1010$	$100010 - 101$	10001×10100	$110010 \div 111$

▽ **Correction**

Opération	Résultat	Opération	Résultat
$1011010 + 1110101$	11001111	111×1111	1101001
$111010 + 110110$	1110000	1010×11001	11111010
$1111111 + 1010$	10001001	10001×10100	101010100
$1010011 - 1111$	1000100	$100110 \div 110$	110
$110101 - 1001$	101100	$110101 \div 1010$	101
$100010 - 101$	11101	$110010 \div 111$	111

Exercice 2.2

Un ordinateur de type ix86 possède les quatre indicateurs suivants pouvant prendre les valeurs 0 ou 1 en fonction du résultat de la dernière opération entière :

SF (*Sign Flag*) : positionné si le résultat est négatif ;

CF (*Carry Flag*) : positionné en cas de présence d'une retenue finale (bit sur-numéraire) ;

ZF (*Zero Flag*) : positionné si le résultat est nul ;

OF (*Overflow Flag*) : positionné en cas de changement anormal de signe.

Donner la valeur des indicateurs après chacune des opérations suivantes :

$10001010 + 01101001$	$01110100 + 01011101$
$10001000 + 11100101$	$11101000 + 00111010$
$01001001 + 00100010$	$11111111 + 00100101$

Interpréter les résultats et indiquer les indicateurs pertinents dans les deux cas suivants :

1. Les opérandes sont des entiers non-signés ;
2. Les opérandes sont des entiers signés, codés en complément à 2.

▽ **Correction**

Opération	Résultats	Signé	Non signé	SF	CF	ZF	OF
$10001010 + 01101001$	11110011	$-118 + 105 = -13$	$138 + 105 = 243$	1	0	0	0
$01110100 + 01011101$	11010001	$116 + 93 = -47$	$116 + 93 = 209$	1	0	0	1
10001000+11100101	**1** 01101101	$-120 + -27 = 109$	$136 + 229 = 109$	0	1	0	1
11101000+00111010	**1** 00100010	$-24 + 58 = 34$	$232 + 58 = 34$	0	1	0	0
01001001+00100010	01101011	$73 + 34 = 107$	$73 + 34 = 107$	0	0	0	0
11111111+00100101	**1** 00100100	$-1 + 37 = 36$	$255 + 37 = 36$	0	1	0	0

Exercice 2.3

Dans la suite, on considère le format `tiny` (1,2,2) vu en cours.

1. Représenter 2.5 et 0.25 dans le format `tiny`. Faire la somme de ces deux nombres (en binaire). Que se passe t'il ?
2. Faire la somme de 1.5 et 1.75 dans le format `tiny`. Que se passe t'il ? Donner les différents résultats possibles.
3. Étant donnés deux nombres x et y au format `tiny`, que peut-on dire de la relation $x = y \iff x - y = 0$ en l'absence des nombres dénormalisés ? Donner les avantages et inconvénients de la représentation normalisée ;
4. Donner un exemple illustrant la non-associativité de l'addition et de la multiplication pour les calculs en nombres flottants exprimés dans le format `tiny` (hors dépassement de capacité). On considérera que tous les calculs sont arrondis par troncation.

▽ **Correction**

1. On a : $0.25_{10} = 0.01 \times 2^0{}_2 = \boxed{0\,|\,00\,|\,01}$ et $2.5_{10} = 1.01 \times 2^1{}_2 = \boxed{0\,|\,10\,|\,01}$, Le nombre 0.25 est dénormalisé. Pour faire la somme, il faut d'abord amener les deux nombres au même exposant (celui d'exposant le plus grand). Il vient :

$$0.25_{10} = 0.01 \times 2^0{}_2 = 0.001 \times 2^1{}_2$$

En reformatant la nouvelle mantisse sur 3 bits, on obtient :

$$2.5_{10} + 0.25_{10} = 1.01 \times 2^1{}_2 + 0.00 \times 2^1{}_2 = 1.01 \times 2^1{}_2 = 2.5_{10}$$

Il y a eu *absorption*.

2. On a : $1.5_{10} = 1.10 \times 2^0{}_2 = \boxed{0\,|\,01\,|\,10}$ et $1.75_{10} = 1.11 \times 2^0{}_2 = \boxed{0\,|\,01\,|\,11}$. La somme donne $1.101 \times 2^1{}_2$ après normalisation. Comme on a une partie fractionnaire sur deux bits, il faut arrondir le résultat. Suivant l'arrondi choisi, on obtiendra 3 ou 3.5.
3. La relation n'est plus vérifiée. Exemple : pour $x = 1.25$ et $y = 1$ on a $x - y = 0.25$ qui est non représentable dans le format `tiny` sans nombres dénormalisés. Le résultat de la soustraction sera donc 0 (arrondi au plus proche) sans que x et y soient égaux. Avantage de la représentation normalisée : on a le maximum de précision pour une taille de mantisse fixée. Inconvénient : la normalisation fait perdre de l'information sur la précision des calculs (ex. $9.8776456 \times 10^0 - 9.8776454 \times 10^0 = 0.0000002 \times 10^0 = 2.0000000 \times 10^{-7}$. On est donc porté à croire que l'on connaît la différence avec 8 chiffres de précision plutôt que 1 ;
4. On a :

$$r_1 = (3 + 0.25) + 0.25 = (3) + 0.25 = \mathbf{3}$$

et

$$r_2 = 3 + (0.25 + 0.25) = 3 + (0.5) = \mathbf{3.5}$$

De même :

$$r_3 = (2 * \frac{1}{2}) * \frac{1}{4} = \mathbf{1}$$

et

$$r_4 = 2 * (\frac{1}{2} * \frac{1}{4}) = \mathbf{0}$$

Exercice 2.4

Donner la représentation binaire en flottant simple précision (1,8,23) des nombres suivants :

$$30.5 \quad -0.5625 \quad \tfrac{1}{3} \quad 0.85$$

On utilisera un arrondi par troncation si nécessaire.

▽ Correction

On a un biais de 127 en format simple précision (1,8,23).

$$\begin{aligned} 30.5 &= \mathrm{41f40000} \\ -0.5625 &= \mathrm{bf100000} \\ 1/3 &= \mathrm{3eaaaaab} \\ 0.85 &= \mathrm{3f59999a} \end{aligned}$$

Exercice 2.5

Déterminer les nombres représentés en flottant simple précision (1,8,23) donnés par les chaînes suivantes :

$$1000\ 1111\ 1110\ 1111\ 1100\ 0000\ 0000\ 0000_{2}$$
$$\mathrm{FF800000}_{16}$$

▽ Correction

$1000\ 1111\ 1110\ 1111\ 1100\ 0000\ 0000\ 0000_{2}$ = $-2.36411752533422e-29_{10}$ (s=1, E=-96, f=.110111111)

$\mathrm{FF800000}_{16}$ = -inf (s=1, e=**255**, m=.0)

Exercice 2.6

- Représenter les nombres 12.5859375, 108.5, 9.1 en notation scientifique binaire normalisée avec une mantisse de 8 bits ;
- En utilisant les représentations binaires obtenues à la question précédente, effectuer les opérations ci-dessous :

$$12.5859375 + 108.5 \quad 108.5 - 9.1 \quad 12.5859375 \times 9.1$$

▽ Correction

1. Représentation normalisée :

$$\begin{array}{lll} 12.5859375 & = 1100.10010110 & = 1.1001001 \times 2^3 \\ 108.5 & = 1101100.1 & = 1.1011001 \times 2^6 \\ 9.1 & = 1001.0\overline{0011} & = 1.0010001 \times 2^3 \end{array}$$

2. On a :

$$\begin{aligned}
12.5859375 + 108.5 &= 0.0011001 \times 2^6 + 1.1011001 \times 2^6 &= 1.1110010 \times 2^6 \\
108.5 - 9.1 &= 1.1011001 \times 2^6 - 0.0010010 \times 2^6 &= 1.1000111 \times 2^6 \\
12.5859375 \times 9.1 & &= 1.1100011 \times 2^6
\end{aligned}$$