Fiche TD8 : Représentation d'entiers et de flottants

Exercice 1 *Conversions sur les entiers*

Donner les représentations des entiers suivants dans la base indiquée :

1. $(123)_{10}$ dans les bases 2 puis 5 puis 16.
2. $(2187)_{10}$ en héxadécimal.
3. $(110110101)_2$ dans la base 10.
4. $(201021)_3$ dans la base 10.
5. $(2BA4)_{16}$ dans la base 10.
6. $(1F7A)_{16}$ dans la base 2 sans repasser par la base 10.
7. $(11010111010001010)_2$ dans la base 16 sans repasser par la base 10.

Exercice 2 *Opérations sur les entiers naturels en base p*

1. Soit p un entier supérieur ou égal à 2 et n un entier dont on connaît l'écriture en base p.
   a) Comment obtenir facilement l'écriture en base p de $p \times n$? Le prouver.
   b) Comment obtenir facilement l'écriture en base p de $\lfloor \frac{n}{p} \rfloor$ où $\lfloor \cdot \rfloor$ dénote la partie entière ? Le prouver.
2. Dans cette question, on cherche à concevoir un algorithme permettant de multiplier deux entiers naturels M et N représentés en base 2 à l'aide d'additions et de l'opérateur de décalage à gauche défini comme suit : si $a = (a_{n-1} \cdots a_0)_2$ et $k \in \mathbb{N}$, $a \ll k$ est l'entier représenté par $(a_{n-1} \cdots a_0 \underbrace{0 \cdots 0}_{k \text{ zéros}})_2$.
   a) Si les entiers sont représentés sur n bits, combien faut-il de bits pour être certain que le calcul du produit $M \times N$ ne provoque pas de dépassement de capacité ?
   b) Traduite sur les entiers, à quelle opération correspond un décalage à gauche de k bits ?
   c) Si $M = 2^p$, comment facilement obtenir le produit $M \times N$?
   d) En déduire un algorithme permettant de multiplier M et N en n'utilisant que des tests sur les bits de M, des additions et des décalages à gauche.

Exercice 3 *Entiers relatifs et complément à 2*

Dans cet exercice, on encode les entiers relatifs sur 8 bits en complément à 2.

1. Quels sont les entiers qu'il est possible de représenter ?
2. Donner les représentations des entiers suivants : $0, -1, 42, -42, -128$ et 128.
3. Donner la valeur en base 10 des entiers suivants : $(01101010)_2$, $(11000011)_2$ et $(10101011)_2$.

Exercice 4 *Additions, opposés, soustractions*

1. En supposant que les nombres a et b sont des entiers naturels codés sur 8 bits, effectuer l'addition $a + b$ dans chacun des cas suivants et dire quels cas provoquent un dépassement de capacité :
   a) $a = (11111111)_2$ et $b = (00000001)_2$.
   b) $a = (10010111)_2$ et $b = (01011001)_2$.
   c) $a = (11010100)_2$ et $b = (00111001)_2$.
   d) $a = (01101010)_2$ et $b = (01101001)_2$.
2. Reprendre la question précédente en supposant cette fois que a et b sont des entiers relatifs codés sur 8 bits en complément à 2.
3. Soit N un entier relatif codé sur n bits en complément à 2. On note $\widetilde{N}$ l'entier obtenu en transformant tous les bits de N égaux à 1 en zéros et tous les bits de n égaux à 0 en uns.

a) Que vaut $N + \widetilde{N}$?

b) En déduire comment obtenir l'opposé d'un nombre écrit en complément à 2. Que se passe-t-il lorsqu'on applique ces opérations à l'entier -2^{n-1} ? Expliquez.

c) Déduire de la question précédente un algorithme permettant de soustraire deux entiers relatifs s'il n'y a pas de dépassement de capacité et que le problème ci-dessus n'est pas rencontré.

Exercice 5 *Codage des flottants*

On rappelle le nombre de bits utilisés dans les formats classiques (simple précision sur 32 bits et double précision sur 64 bits) pour représenter des flottants et on introduit également ceux qu'on utilisera pour un format maison sur 8 bits :

format	signe	exposant décalé	mantisse
8 bits	1 bit	3 bits	4 bits
32 bits	1 bit	8 bits	23 bits
64 bits	1 bit	11 bits	52 bits

1. Quelle est la valeur du décalage dans chacun des formats ci-dessus ?
2. A quoi sont égaux les nombres suivants si les suites de bits suivantes représentent des flottants codés sur 8 bits avec le format décrit ci-dessus ? Lesquels sont des flottants dénormalisés ?
   a) 00010011.
   b) 10000000.
   c) 11101101.
   d) 01110101.
   e) 01101111.
3. On suppose que la chaîne héxadécimale $(C3AC0000)_{16}$ représente un nombre flottant au format simple précision. Quelle est la valeur décimale codée ?
4. Dans chacun des cas suivants, coder le réel x dans le format simple précision en arrondissant au flottant le plus proche si nécessaire et indiquer si x est dyadique.
   a) $x = 21.59375$. b) $x = -13.1$. c) $x = 18.13$.

Exercice 6 *Flottants normalisés, flottants dénormalisés*

On considère une représentation avec un bit de signe, e bits d'exposant (décalé) et m bits de mantisse.

1. a) Combien de flottants normalisés peut-on représenter ?
   b) Quel est le plus grand flottant normalisé que l'on peut représenter ? Et le plus petit ?
   c) Quel est le plus petit flottant normalisé strictement positif représentable ? Et le plus petit qui soit strictement supérieur à 1 ?
2. a) Combien de flottants dénormalisés peut-on représenter ?
   b) Quels est le plus grand nombre flottant dénormalisé représentable ? Calculer la différence entre ce flottant et le plus petit flottant normalisé représentable. Que se serait-il passé si on avait choisi de décaler l'exposant des flottants dénormalisés de la même façon que ceux des flottants normalisés ?
   c) Quel est le plus petit flottant dénormalisé strictement positif ?
3. En utilisant le tableau de l'exercice 5, donner des valeurs numériques à vos réponses précédentes dans le cas de flottants en simple précision.

Exercice 7 *Précision et arrondis*

Dans cet exercice on représente des flottants positifs normalisés à l'aide de 5 bits divisés en 2 bits d'exposant et 3 bits de mantisse.

1. a) Indiquer toutes les valeurs représentables dans ce format.
   b) Donner la représentation dans ce format des valeurs suivantes : 1.12 1.43 2.3 2.25.
2. a) On rappelle que l'erreur relative commise en remplaçant un réel x par sa représentation $\widetilde{x}$ est par définition $\left|\frac{x-\widetilde{x}}{x}\right|$. Déterminer l'erreur relative commise dans chacun des cas précédents et vérifiez que cette erreur est inférieure à la précision du format.
   b) Si 2.25 et 2.3 sont représentés en utilisant ce format, quel est le résultat de l'opération $2.3 - 2.25$? Quelle est l'erreur relative commise ? Pourquoi n'est-ce pas en contradiction avec la précision du format ?