TD : Représentation des entiers/flottants

1 Exercices sur la numération

Dans toute cette section, on ne travaillera qu'avec des entiers positifs.

Exercice 1. *De la base 10 à une autre base.* Donner les représentations des nombres suivants dans la base indiquée.

1. 123 en bases 2, 3 et 5.
2. 2142 en base 16.
3. 666 en base 2.
4. 1027 en base 7.

Exercice 2. *D'une base à la base 10.* Donner l'interprétation des nombres suivants en base 10.

1. $\overline{2B4}^{16}$.
2. $\overline{11010010}^{2}$.
3. $\overline{412}^{5}$.
4. $\overline{20102}^{3}$.

Exercice 3. *Bases et puissances.* Convertir les nombres qui suivent dans la base indiquée, sans repasser pas par la base 10 (dix !).

1. $\overline{4FC3}^{16}$ en base 2.
2. $\overline{1231231}^{4}$ en base 2.
3. $\overline{10110011011010011}^{2}$ en base 16.
4. $\overline{10110011011010011}^{2}$ en base 8.
5. $\overline{184}^{9}$ en base 3.
6. $\overline{41023}^{5}$ en base 25.

Exercice 4. Combien faut-il de bits pour représenter tous les nombres à n chiffres (en base 10) ?

Exercice 5. Quelle est la représentation binaire des nombres 1,7,31,127 ? Expliquer.

Exercice 6. Soit p un entier supérieur ou égal à 2, et N un entier. Si l'on connaît l'écriture de N en base p, comment obtient-on (facilement) celles de $p \times N$ et $\left\lfloor \frac{N}{p} \right\rfloor$? (On rappelle que $\lfloor . \rfloor$ dénote la partie entière). Le démontrer.

Exercice 7. *Schéma de Hörner pour le changement de base.* Soit $N = \overline{a_{n-1} \cdots a_0}^{b}$ un entier encodé dans la base b. L'algorithme vu en cours pour obtenir N dans une base où on sait calculer (pour nous, la base 10), consiste à évaluer la somme $\sum_{k=0}^{n-1} a_k b^k$. L'algorithme de Hörner pour cette évaluation consiste à utiliser l'idée suivante :

$$N = a_0 + b \times (a_1 + b \times (a_2 + b \times (a_3 + \cdots + b \times (a_{n-2} + b \times a_{n-1}) \cdots)))$$

1. Écrire l'algorithme en pseudo-code.
2. L'implémenter en Python, en supposant les chiffres de N dans la base b stockés dans une liste.
3. Discuter de l'intérêt de cet algorithme par rapport à celui du cours, en nombre d'opérations.

2 Les entiers relatifs en base 2

Dans toute cette section, on ne travaillera qu'avec des nombres entiers. Pour les deux premiers exercices, on travaille avec des entiers naturels, pour les autres, on travaille sur des entiers relatifs représentés en complément à 2.

Exercice 8. *Additions d'entiers naturels sur 8 bits.* Pour chacun des couples de nombres a, b de 8 bits suivants, effectuer l'addition $a + b$. Quelles sont celles qui donnent lieu à un dépassement de capacité ?

1. $(a, b) = (\overline{11111111}^{2}, \overline{00000001}^{2})$
2. $(a, b) = (\overline{10010111}^{2}, \overline{01011001}^{2})$
3. $(a, b) = (\overline{11010100}^{2}, \overline{00111001}^{2})$
4. $(a, b) = (\overline{01101010}^{2}, \overline{01101001}^{2})$

Exercice 9. *Multiplieur sur les entiers naturels.* Dans cet exercice, on va voir comment réaliser une multiplication à l'aide d'additions et de l'*opérateur de décalage vers la gauche.*

1. Soit N et M deux entiers naturels de n bits. Combien faut-il de bits (dans le pire de cas) pour représenter le produit $N \times M$? (On discutera suivant si $n = 1$ ou $n > 1$).

2. On suppose dorénavant que $n > 1$, et que l'on dispose de suffisamment de bits pour écrire le résultat de la multiplication. Soit $a = \overline{a_{n-1}\cdots a_0}^2$. On dénote par $a \ll 1$ l'entier dont la représentation binaire est $\overline{a_{n-1}\cdots a_0 0}^2$, et plus généralement par $a \ll p$ l'entier représenté par $\overline{a_{n-1}\cdots a_0 \underbrace{0\cdots 0}_{p \text{ zéros}}}^2$. Traduite sur les entiers, à quelle opération correspond un décalage à gauche ?
3. Si $M = 2^p$, quel opération simple permet d'obtenir le produit $N \times M$?
4. En déduire un algorithme permettant de multiplier N et M, n'effectuant que des tests sur les bits de M, des décalages à gauche, et des additions.

Exercice 10. *Entiers relatifs sur 8 bits.* On rappelle qu'en représentation en complément à 2 sur n bits (complément à 2^n pour être exact !), on peut représenter tous les nombres de l'intervalle $[\![-2^{n-1}, 2^{n-1}-1]\!]$. Donner les représentations des entiers suivants, sur 8 bits.

1. $N = 0$
2. $N = -1$
3. $N = 42$
4. $N = -42$
5. $N = 127$
6. $N = -128$

Exercice 11. *Entiers relatifs sur 8 bits, l'inverse.* À l'inverse de l'exercice précédent, donnez maintenant la valeur des entiers relatifs suivants, codés sur 8 bits en complément à 2.

1. $N = \overline{01101010}^2$
2. $N = \overline{11000101}^2$
3. $N = \overline{10001110}^2$

Exercice 12. *Additions d'entiers relatifs.* Effectuer les additions des couples suivants représentants des entiers relatifs. Pour lesquelles y a-t-il dépassement de capacité ? (c'est à dire que la somme attendue n'est pas représentable sur 8 bits ?)

1. $(a, b) = (\overline{11111111}^2, \overline{00000001}^2)$
2. $(a, b) = (\overline{10010111}^2, \overline{01011001}^2)$
3. $(a, b) = (\overline{11010110}^2, \overline{10001010}^2)$
4. $(a, b) = (\overline{11010100}^2, \overline{10011001}^2)$
5. $(a, b) = (\overline{01010100}^2, \overline{00011001}^2)$
6. $(a, b) = (\overline{01010100}^2, \overline{01011001}^2)$

Exercice 13. *Opposé d'un entier.* Soit N un entier relatif codé en complément à 2 sur n bits. On note $\overline{N}$ l'entier obtenu en transformant les bits de N égaux à 1 en zéro et réciproquement.

1. Que vaut $N + \overline{N}$?
2. Comment obtenir la représentation de l'opposé d'un nombre en complément à 2, avec une telle transformation et une addition ?
3. Que se passe-t-il avec $N = -2^{n-1}$ sur n bits ? Expliquer.

Ainsi, toujours avec la même méthode, il est possible de réaliser facilement des soustractions, car $N - M = N + (-M)$. C'est ainsi qu'elles sont réalisées dans un processeur.

Exercice 14. *Caractérisation des dépassements de capacité sur entiers relatifs.* Soit $a = \overline{a_{n-1}a_{n-2}\cdots a_0}^2$ et $b = \overline{b_{n-1}b_{n-2}\cdots b_0}^2$ deux entiers relatifs de n bits en complément à 2. On note $r_0 = 0$, et pour tout $i \in [\![1, n]\!]$, on note r_i la retenue correspondant à l'addition des bits a_{i-1}, b_{i-1} et r_{i-1}. On cherche à montrer que l'addition se fait sans dépassement de capacité si et seulement si, $r_n = r_{n-1}$.

1. Vérifier que le résultat est vrai sur les exemples de l'exercice 12.
2. Supposons que a et b représentent deux entiers naturels. Que valent a_{n-1}, b_{n-1} et r_n ? Justifier que l'addition se fait sans dépassement si et seulement si $r_{n-1} = 0$, et conclure pour les entiers naturels.
3. Supposons que a et b soient deux entiers strictement négatifs. Que valent a_{n-1}, b_{n-1} et r_n ? Justifier que l'addition se fait sans dépassement si et seulement si $r_{n-1} = 1$, et conclure pour les entiers négatifs.
4. Enfin, supposons que a est un entier naturel, et b un entier strictement négatif. Justifier qu'il n'y a pas dépassement de capacité. Que vaut $a_{n-1} + b_{n-1}$? Conclure en discutant suivant les valeurs de r_{n-1}.

En pratique, c'est exactement comme ça que l'on teste le dépassement de capacité sur entiers relatifs dans un processeur !

3 Exercices sur les flottants

Dans toute cette section, on considérera des nombres flottants, donnés par leur représentation par signe, mantisse, exposant. On rappelle ici le nombre de bits utilisés dans les formats classiques (simple précision 32 bits et double précision 64 bits), et on donne également une représentation de flottants sur 9 bits « maison » que l'on utilisera pour certains exercices (ça fait moins de chiffres !) Dans la représentation ci-dessous, l'exposant décalé est un entier naturel, et on utilise un décalage correspondant à $2^{\text{taille de l'exposant décalé}-1} - 1$.

format	signe	exposant décalé	décalage	mantisse	signification (nombre normalisé)
32 bits	1 bit	8 bits	$2^{8-1}-1=127$	23 bits	$(-1)^{\text{signe}} \times 1, \underbrace{\cdots}_{\text{mantisse}} \times 2^{\text{exposant décalé}-127}$
64 bits	1 bit	11 bits	$2^{11-1}-1=1023$	52 bits	$(-1)^{\text{signe}} \times 1, \underbrace{\cdots}_{\text{mantisse}} \times 2^{\text{exposant décalé}-1023}$
9 bits	1 bit	4 bits	$2^{4-1}-1=7$	4 bits	$(-1)^{\text{signe}} \times 1, \underbrace{\cdots}_{\text{mantisse}} \times 2^{\text{exposant décalé}-7}$

On rappelle qu'un *nombre normalisé* a son exposant décalé qui n'est ni $0\cdots0$, ni $1\cdots1$.

Exercice 15. *Quelques représentations.* On considère la représentation sur 9 bits donnée plus haut. À quoi sont égaux les nombres suivants ?

1. 110010000
2. 000111010
3. 011101111
4. 101111010

Exercice 16. *Représentations de dyadiques.* Donnez la représentation des dyadiques suivants sur 9 bits. On garantit qu'on peut les représenter de manière exacte.

1. 16.0
2. 0.3125
3. -8.5

Exercice 17. *Approximation.* Donner la représentation sur 9 bits du flottant le plus proche de π. (On pourra s'aider d'une calculatrice...)

Exercice 18. *Nombres représentables normalisés.* On considère une représentation avec 1 bit de signe, e bits d'exposant et m bits de mantisse.

1. Combien de nombres normalisés peut-on représenter ?
2. Quel est le plus grand nombre que l'on peut représenter ? Le plus petit ?
3. Quel est le plus petit nombre représentable strictement supérieur à 1 ? Le plus petit strictement positif normalisé ?

On rappelle maintenant ce que sont les nombres *dénormalisés*. Ceux-ci ont leur exposant décalé égal à $0\cdots0$. Si la mantisse est nulle, le nombre vaut zéro (il y a alors un zéro positif et un zéro négatif), sinon l'interprétation est

$$(-1)^{\text{signe}} \times 0, \underbrace{\cdots}_{\text{mantisse}} \times 2^{-2^{\text{taille de l'exposant décalé}-1}+2}$$

En d'autres termes, l'interprétation est la même que pour les normalisés (car l'exposant décalé vaut 0), mais le décalage est réduit de 1.

Exercice 19. *Quelques nombres dénormalisés.* On considère les nombre dénormalisés suivants, sur 9 bits. À quoi sont ils égaux ?

1. 000001111
2. 100000101
3. 000000001

Exercice 20. *Nombres dénormalisés représentables.* On considère une représentation avec 1 bit de signe, e bits d'exposant et m bits de mantisse.

1. Combien de nombres dénormalisés peut-on représenter ?
2. Quel est le plus grand nombre dénormalisé que l'on peut représenter ? Quelle est sa différence avec le plus petit normalisé positif ? Justifier la valeur différente du décalage.
3. Quel est le plus petit nombre dénormalisé strictement positif ?

Exercice 21. *Comparaison de flottants.* Montrer que pour comparer deux flottants positifs, il suffit de les comparer bit à bit. (on excluera le cas où l'un au moins des exposants décalés est $1\cdots1$)

Exercice 22. *Nombres non représentables.* Les nombres non décimaux ne sont pas représentables exactement avec des flottants. Proposer des nombres entiers ou décimaux non entiers qui ne le sont pas non plus dans les cas suivants :

— parce qu'ils sont trop grands ou trop petits.

— pour des raisons de précision.

Lorsque l'exposant décalé d'un flottant est égal à $11\cdots1$, on parle de NAN (Not A Number). Les NAN sont utilisés pour signaler des opérations non valides (par exemple le calcul de $\sqrt{-1}$). Une exception : si la mantisse est nulle, le flottant représente $+\infty$ ou $-\infty$ suivant son signe. En C par exemple, le calcul de $1/0$ produit $+\infty$ (on obtient une erreur en Python). Les infinis sont utilisés pour représenter des résultats de calculs trop grands en valeur absolue.