Aide-mémoire pour les systèmes de numération et le codage des nombres

Yann Thorimbert

Systèmes de numération positionnels

Expression d'un nombre entier dans une base b :

$$n_b = \sum_{i=0}^{k-1} a_i \cdot b^i, \tag{1}$$

Pour la base 2 on a :

$$n_2 = \sum_{i=0}^{k-1} a_i \cdot 2^i, \tag{2}$$

En base 2, si l'on dispose de k chiffres, l'entier maximum que l'on peut représenter est $2^k - 1$.

En base 2, si l'on dispose de k chiffres, le nombre de valeurs différentes que l'on peut représenter est 2^k.

Codage des nombres

Une fonction de codage $cod(n)$ prend en entrée une quantité n en représentation externe et donne en sortie un nombre en représentation interne, donc un état binaire.

Une fonction de décodage $dec(b)$ prend en entrée un état binaire b (représentation interne d'une quantité) et retourne la représentation externe correspondante.

Codage des entiers naturels

La fonction $cod_{\mathbb{N}_{(k)}}(n)$ retourne la séquence de bits correspondant aux chiffres de n en base 2.

Ensemble des entiers naturels exprimables avec k :

$$\mathbb{N}_{(k)} = \{x \in \mathbb{N} | 0 \leq x < 2^k\}. \tag{3}$$

Codage des entiers relatifs

Entiers biaisés

$$cod_{\mathbb{Z}B^+_{(k)}}(n) = cod_{\mathbb{N}_{(k)}}(n + 2^{k-1} - 1). \tag{4}$$

$$dec_{\mathbb{Z}B^+_{(k)}}(b) = dec_{\mathbb{N}_{(k)}}(b) - 2^{k-1} + 1. \tag{5}$$

Complément à 2

Ensemble des entiers relatifs exprimables avec k bits en complément à 2 :

$$\mathbb{Z}_{(k)} = \{x \in \mathbb{Z} | -2^{k-1} \leq x < 2^{k-1}\} \tag{6}$$

$$cod_{\mathbb{Z}_{(k)}}(n) = \begin{cases} cod_{\mathbb{N}_{(k)}}(2^k - |n|) & \text{si } n < 0, \\ cod_{\mathbb{N}_{(k)}}(n) & \text{sinon.} \end{cases} \tag{7}$$

$$dec_{\mathbb{Z}_{(k)}}(b) = \begin{cases} dec_{\mathbb{N}_{(k)}}(b) - 2^k & \text{si } b_{k-1} = 1, \\ dec_{\mathbb{N}_{(k)}}(b) & \text{sinon.} \end{cases} \tag{8}$$

Codage des réels en virgule flottante (norme IEEE 754)

Dans tous les cas :
— Le signe est calculé à partir de son codage b comme : $s = (-1)^b$.
— L'exposant est calculé à partir la séquence binaire b_E comme :

$$E = dec_{\mathbb{Z}_{(8)}}(b_E) - 2^{k_E - 1} + 1.$$

— La mantisse est calculée à partir la séquence binaire b_m comme :

$$m = 1 + \sum_{i=1}^{k_m} 2^{-i} \cdot b_{m_i}.$$

— Le nombre représenté vaut $s \cdot m \cdot 2^E$.

Simple précision

Dans ce cas, le nombre est codé sur 32 bits, avec 1 bit pour le signe, 8 bits pour l'exposant et 23 bits pour la mantisse. En particulier, $E = dec_{\mathbb{Z}_{(8)}}(b_E) - 127$.

Double précision

Dans ce cas, le nombre est codé sur 64 bits, avec 1 bit pour le signe, 11 bits pour l'exposant et 52 bits pour la mantisse. En particulier, $E = dec_{\mathbb{Z}_{(11)}}(b_m) - 1023$.