Synthese - IEEE754

Jaunart Gilles

1 Paramètres et valeurs speciales:

Lettre majuscule représente la taille
Lettre minuscule représente la valeur

E = exposant
M = mantisse
B = biais
S = bit de signe

1.1 Paramètres IEEE754:

Précision	**Taille (bits)**	**E**	**M**	**B**
Simple	32	8	23	127
Double	64	11	52	1023

1.2 Valeurs speciales:

s	**e**	**m**	**valeur**
0	2^{E-1}	0	$+\infty$
1	2^{E-1}	0	$-\infty$
0 / 1	2^{E-1}	$\neq 0$	NaN

2 Conversion decimal vers IEEE754

2.1 Signe:

Si $x > 0$, le bit de signe vaut 0. Si $x < 0$, le bit de signe vaut 1.
Par la suite on prendra la valeur absolue de X.

2.2 Normalisation:

x est normalisé à l'aide de n divisions successives afin que $x'.2^n = x$ où $1 \leq x' < 2$.
Par exemple, $x = 6.2$ est divisé $n = 2$ fois par 2 pour obtenir $x' = 1,55$ tel que $1,55.2^2 = 6,2$

2.3 Déduction de l'exposant:

On pose $2^n = 2^{e-B}$, ce qui permet de déduire $e = n + B$. On vérifie alors que le nombre est bien représentable.
Si $0 < e < 2^{E-1}$, alors la représentation normalisée doit etre utilisée.
Si $e \leq 0$, alors on doit utiliser la dénormalisée.

2.4 Déduction de la mantisse en représentation normalisée:

On pose $x' = 1 + \frac{m}{2^M}$, ce qui permet de déduire $m = (x' - 1).2^M$

2.5 Déduction de la mantisse en représenation dénormalisée:

Ici l'exposant est fixe, $e = 1$. On pose $x = \frac{m}{2^M}.2^{1-B}$, ce qui permet de déduire $m = x.2^{M-(1-B)}$

3 Arrondi avec IEEE754

3.1 Round-to-nearest-even:

On arrondit au nombre le plus proche. Ex: $2,4 \rightarrow 2$; $1,7 \rightarrow 2$
Si on se retrouve pile entre 2 nombres pairs (ex: $1,35$; $2,65$; $3,15$) alors on arrondit au nombre pair le plus proche. Ex: $1,35 \rightarrow 1,4$ car 4 est pair et pas 3 ; $2,65 \rightarrow 2,6$; $3,15 \rightarrow 3,2$

3.2 Round-toward-zero:

On tronque la partie non-représentable. Ex: $1,4 \rightarrow 1$; $2,7 \rightarrow 2$; $-2,5 \rightarrow -2$

3.3 Round-down:

Il s'agit de l'arrondi par défaut. Ex: $1,4 \rightarrow 1$; $2,7 \rightarrow 2$; $-2,5 \rightarrow -3$

3.4 Round-up:

Il s'agit de l'arrondi par excés. Ex: $1,4 \rightarrow 2$; $2,7 \rightarrow 3$; $-2,5 \rightarrow -2$

4 Typologie des erreurs

4.1 Erreur vraie:

$\Delta x = X - \hat{X}$ où $\hat{X}$ est la valeur de nombre représenté.

4.2 Erreur absolue:

$|\Delta x| = |X - \hat{X}|$

4.3 Erreur relative:

$\epsilon_X = \frac{|X-\hat{X}|}{|X|}$

4.4 Epsilon machine:

La précision machine ou "epsilon machine" est une borne sur l'erreur relative qui dépend du format de représentation, en particulier de la taille de la mantisse M.

$$\frac{|X-\hat{X}|}{|X|} \leq \epsilon_M \text{ où } \epsilon_M = 2^{-(M+1)}$$