Chapitre 5

Etude des nombres réels en machine

La représentation standard (norme *IEEE 754*) des nombres à virgule flottante dits flottants en base 2 sur machine (par exemple Linux), notés X est $(-1)^s \times (1+m) \times 2^{e-d}$ (flottants normalisés).
Avec :
- s le signe (1 pour négatif, 0 pour positif) sur un bit ;
- m la partie fractionnaire en base 2 sur 23 bits en simple précision (SP, 32 bits) et 52 en double précision (DP, 64 bits), $\frac{1}{2} \leq m < 1$;
- e l'exposant entier non signé sur 8 bits en SP (respectivement 11 en DP) ;
- d le décalage de l'exposant pour coder un nombre entier non signé. Si l'on note q le nombre de bits codant l'exposant, on a $d = 2^{q-1} - 1$ (en SP 127, en DP 1023).

Le premier bit de la mantisse $M = 1 + m$, ($1 \leq M < 2$) qui est implicitement 1 en normalisé n'est pas codé.
Pour densifier les nombres codés autour de 0 on prend des flottants dénormalisés de représentation standard $(-1)^s \times m \times 2^{1-d}$ (l'exposant vrai est en SP -126 et en DP -1022 avec le codage spécial : exposant codé à 0, mantisse codée $\neq 0$).
Le premier bit de la mantisse $0 + m$ qui est implicitement 0 en dénormalisé n'est pas codé.

Les flottants caractéristiques positifs en norme *IEEE 754* sur machine sont les suivants :
- l'$\epsilon_{machine}$, le plus petit flottant positif tel que sur machine $1 \oplus \epsilon_{machine} > 1$;
- X^n_{min} normalisé, $X^d_{min} = 0^+$ (voisin par valeur supérieure de 0) dénormalisé ;
- X^n_{max} normalisé, $X^d_{max} = X^{n-}_{min}$ (voisin par valeur inférieure de X^n_{min}) dénormalisé ;
- le zéro : 0 (codage spécial : exposant codé à 0, mantisse codée à 0) et l'unité : 1 normalisé ;
- l'∞ (codage spécial : exposant codé avec des 1, mantisse codée à 0) et NaN (Not a Number) (codage spécial : exposant codé avec des 1, mantisse codée $\neq 0$).
- l'unité d'arrondi, u égal à $\epsilon_{machine}/2$ (en arrondi au plus près) est tel que pour tout réel x, avec $X^n_{min} \leq |x| \leq X^n_{max}$ on ait $X = x(1+\delta)$ avec $|\delta| \leq u$.
- Soit X un flottant normalisé, $X^+ = X(1+\xi)$ avec $|\xi| \leq \epsilon_{machine}$.

Représentation mémoire des flottants :

$$\underline{|s|e_7 \cdots e_0|m_1 \cdots m_{23}|} \text{ sur 32 bits (SP)}$$

$$\underline{|s|e_{10} \cdots e_0|m_1 \cdots m_{52}|} \text{ sur 64 bits (DP)}$$

$$e = \begin{cases} \displaystyle\sum_{i=0}^{7} e_i 2^i & e_i \in \{0,1\} \text{ en SP} \\ \displaystyle\sum_{i=0}^{10} e_i 2^i & e_i \in \{0,1\} \text{ en DP} \end{cases}$$

$$m = \begin{cases} \displaystyle\sum_{i=1}^{23} m_i 2^{-i} \quad m_i \in \{0,1\} \text{ en SP} \\ \displaystyle\sum_{i=1}^{52} m_i 2^{-i} \quad m_i \in \{0,1\} \text{ en DP} \end{cases}$$

Étude de flottants caractéristiques

1 L'$\epsilon_{machine}$

L'$\epsilon_{machine}$ est le plus petit flottant positif tel que sur machine $1 \oplus \epsilon_{machine} > 1$.

1.1 Simple précision

1.1.1 Codage binaire

Sur 32 bits, on a :

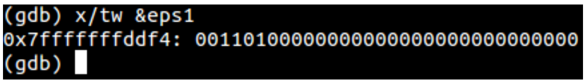


FIGURE 1 – Codage binaire de l'$\epsilon_{machine}$ en SP

Le nombre flottant, en SP sur 32 bits lus de gauche à droite :
— est positif (bit de signe à 0) ;
— est normalisé (l'exposant n'est pas codé par 00000000) ;
— a un exposant codé par 01101000 ;
— a une mantisse codée par 00000000000000000000000.

1.1.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de la mantisse codée $m = 0$.

A partir de l'expression binaire de l'exposant codé $(e)_2 = (01101000)_2 = (e_7 e_6 \ldots e_0)_2$, sa valeur est égale à :

$$\sum_{i=0}^{7} e_i 2^i = 2^3 + 2^5 + 2^6 = 8 + 32 + 64 = 104$$

La valeur de l'exposant vrai $e - d = 104 - 127 = -23$.
La valeur exacte de l'$\epsilon_{machine}$ en SP est égale à :

$$(-1)^0 \times (1 + 0) \times 2^{-23} = 2^{-23}$$

1.1.3 Valeur décimale approchée

La valeur décimale approchée de l'$\epsilon_{machine}$ en SP est égale à :

$$2^{-23} \approx 1.1920929 \times 10^{-7}$$

1.2 Double précision

1.2.1 Codage binaire

Sur 64 bits, on a :
Le nombre flottant, en DP sur 64 bits lus de gauche à droite :

FIGURE 2 – Codage binaire de l'$\epsilon_{machine}$ en DP

— est positif (bit de signe à 0) ;
— est normalisé (l'exposant n'est pas codé par 00000000) ;
— a un exposant codé par 01111001011 ;
— a une mantisse codée par 0000000000000000000000000000000000000000000000000000.

1.2.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de la mantisse codée $m = 0$.
A partir de l'expression binaire de l'exposant codé $(e)_2 = (01111001011)_2 = (e_{10}e_9 \ldots e_0)_2$, sa valeur est égale à :
.

$$\sum_{i=0}^{10} e_i 2^i = 2^0 + 2^1 + 2^3 + 2^6 + 2^7 + 2^8 + 2^9 = 971$$

La valeur de l'exposant vrai $e - d = 971 - 1023 = -52$.
La valeur exacte de l'$\epsilon_{machine}$ en DP est égale à :

$$(-1)^0 \times (1 + 0) \times 2^{-52} = 2^{-52}$$

1.2.3 Valeur décimale approchée

La valeur décimale approchée de l'$\epsilon_{machine}$ en DP est égale à :

$$2^{-52} \approx 2.2204460492503131 \times 10^{-16}$$

2 X_{min}^{n} et X_{min}^{d}

X_{min}^{n} est le plus petit flottant positif normalisé et X_{min}^{d} est le plus petit flottant strictement positif, 0^+.

2.1 X_{min}^{n} normalisé en simple précision

2.1.1 Codage binaire

Sur 32 bits, on a :

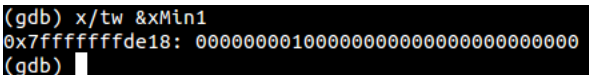


FIGURE 3 – Codage binaire de X_{min}^{n} en SP

Le nombre flottant, en SP sur 32 bits lus de gauche à droite :
— est positif (bit de signe à 0) ;

— est normalisé (l'exposant n'est pas codé par 00000000) ;
— a un exposant codé par 00000001 ;
— a une mantisse codée par 00000000000000000000000.

2.1.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de la mantisse codée $m = 0$.

A partir de l'expression binaire de l'exposant codé $(e)_2 = (00000001)_2 = (e_7 e_6 \ldots e_0)_2$, sa valeur est égale à :

$$\sum_{i=0}^{7} e_i 2^i = 2^0 = 1$$

La valeur de l'exposant vrai $e - d = 1 - 127 = -126$.
La valeur exacte de X^n_{min} en SP est égale à :

$$(-1)^0 \times (1 + 0) \times 2^{-126} = 2^{-126}$$

2.1.3 Valeur décimale approchée

La valeur décimale approchée de X^n_{min} en SP est égale à :

$$2^{-126} \approx 1.1754943 \times 10^{-38}$$

2.2 X^n_{min} normalisé en double précision

2.2.1 Codage binaire

Sur 64 bits, on a :
Le nombre flottant, en DP sur 64 bits lus de gauche à droite :

```
(gdb) x/tg &xMin2
0x7fffffffde20: 0000000000010000000000000000000000000000000000000000000000000000
(gdb)
```

FIGURE 4 – Codage binaire de X^n_{min} en DP

— est positif (bit de signe à 0) ;
— est normalisé (l'exposant n'est pas codé par 00000000) ;
— a un exposant codé par 00000000001 ;
— a une mantisse codée par 00000000000000000000000000000000000000000000000000.

2.2.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de la mantisse codée $m = 0$.

A partir de l'expression binaire de l'exposant codé $(e)_2 = (00000000001)_2 = (e_{10} e_9 \ldots e_0)_2$, sa valeur est égale à :

$$\sum_{i=0}^{10} e_i 2^i = 2^0 = 1$$

La valeur de l'exposant vrai $e - d = 1 - 1023 = -1022$.
La valeur exacte de X^n_{min} en DP est égale à :

$$(-1)^0 \times (1 + 0) \times 2^{-1022} = 2^{-1022}$$

2.2.3 Valeur décimale approchée

La valeur décimale approchée de X^n_{min} en DP est égale à :

$$2^{-1022} \approx 2.2250738585072014 \times 10^{-308}$$

2.3 X^d_{min} dénormalisé en simple précision

2.3.1 Codage binaire

Sur 32 bits, on a :
Le nombre flottant, en SP sur 32 bits lus de gauche à droite :

FIGURE 5 – Codage binaire de X^d_{min} en SP

— est positif (bit de signe à 0) ;
— est dénormalisé (l'exposant est codé par 00000000) ;
— a une mantisse codée par 00000000000000000000001.

2.3.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de l'exposant vrai $= -126$
A partir de l'expression binaire de la mantisse codée $(m)_2 = (0,00000000000000000000001)_2$, sa valeur est égale à :

$$2^{-23}$$

La valeur exacte de X^d_{min} en SP est égale à :

$$(-1)^0 \times 2^{-23} \times 2^{-126} = 2^{-149}$$

2.3.3 Valeur décimale approchée

La valeur décimale approchée de X^d_{min} en SP est égale à :

$$2^{-149} \approx 1.4012984 \times 10^{-45}$$

2.4 X^d_{min} dénormalisé en double précision

2.4.1 Codage binaire

Sur 64 bits, on a :
Le nombre flottant, en SP sur 32 bits lus de gauche à droite :
— est positif (bit de signe à 0) ;
— est dénormalisé (l'exposant est codé par 00000000) ;
— a une mantisse codée par 0000000000000000000000000000000000000000000000000001.

```
(gdb) x/tg &xMin4
0x7fffffffde28: 0000000000000000000000000000000000000000000000000000000000000001
(gdb)
```

FIGURE 6 – Codage binaire de X^d_{min} en DP

2.4.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de l'exposant vrai $= -1022$

A partir de l'expression binaire de la mantisse codée
$(m)_2 = (0,0000000000000000000000000000000000000000000000000001)_2$, sa valeur est égale à :

$$m = 2^{-52}$$

La valeur exacte de X^d_{min} en DP est égale à :

$$(-1)^0 \times 2^{-52} \times 2^{-1022} = 2^{-1074}$$

2.4.3 Valeur décimale approchée

La valeur décimale approchée de X^d_{min} en DP est égale à :

$$2^{-1074} \approx 4.9406564584124654 \times 10^{-324}$$

3 X^n_{max} et X^d_{max}

X^n_{max} est le plus grand flottant normalisé etX^d_{max} est le plus grand flottant dénormalisé c'est-à-dire le flottant voisin par valeur inférieure de X^n_{min}, X^{n-}_{min}.

3.1 X^n_{max} normalisé en simple précision

3.1.1 Codage binaire

Sur 32 bits, on a :

```
(gdb) x/tw &xMax1
0x7fffffffde18: 01111111011111111111111111111111
(gdb)
```

FIGURE 7 – Codage binaire de X^n_{max} en SP

Le nombre flottant, en SP sur 32 bits lus de gauche à droite :
— est positif (bit de signe à 0) ;
— est normalisé (l'exposant n'est pas codé par 00000000) ;
— a un exposant codé par 11111110 ;
— a une mantisse codée par 11111111111111111111111.

3.1.2 Valeur exacte

— le bit de signe $s = 0$;

A partir de l'expression binaire de l'exposant codé $(e)_2 = (11111110)_2 = (e_7e_6 \ldots e_0)_2$, sa valeur est égale à :

$$\sum_{i=0}^{7} e_i 2^i = 2^1 + 2^2 + 2^3 + 2^5 + 2^5 + 2^6 + 2^7 = 2 \times \frac{1-2^7}{1-2} = 254$$

La valeur de l'exposant vrai $e - d = 254 - 127 = 127$.
A partir de l'expression binaire de la mantisse codée $(m)_2 = (0,11111111111111111111111)_2 = (0, m_1m_2 \ldots m_{23})_2$, sa valeur est égale à :

$$\sum_{i=1}^{23} m_i 2^{-i} = 2^{-23}\frac{1-2^{23}}{1-2} = 1 - 2^{-23}$$

La valeur exacte de X^n_{max} en SP est égale à :

$$(-1)^0 \times (2 - 2^{-23}) \times 2^{127} = 2^{128} - 2^{104}$$

3.1.3 Valeur décimale approchée

La valeur décimale approchée de X^n_{max} en SP est égale à :

$$2^{128} - 2^{104} =\approx 3.4028234 \times 10^{38}$$

3.2 X^n_{max} normalisé en double précision

3.2.1 Codage binaire

Sur 64 bits, on a : Le nombre flottant, en DP sur 64 bits lus de gauche à droite :

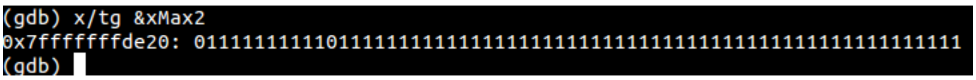


Figure 8 – Codage binaire de X^n_{max} en DP

- est positif (bit de signe à 0) ;
- est normalisé (l'exposant n'est pas codé par 00000000000) ;
- a un exposant codé par 11111111110 ;
- a une mantisse codée par 1111111111111111111111111111111111111111111111111111.

3.2.2 Valeur exacte

- le bit de signe $s = 0$;

A partir de l'expression binaire de l'exposant codé $(e)_2 = (11111111110)_2 = (e_{10}e_9 \ldots e_0)_2$, sa valeur est égale à :

$$= \sum_{i=0}^{10} e_i 2^i = 2^1 + 2^2 + 2^3 + 2^5 + 2^5 + 2^6 + 2^7 + 2^8 + 2^9 + 2^{10} = 2 \times \frac{1-2^{10}}{1-2} = 2046$$

La valeur de l'exposant vrai $e - d = 2046 - 1023 = 1023$.
A partir de l'expression binaire de la mantisse codée
$(m)_2 = (0,1111111111111111111111111111111111111111111111111111)_2 = (0, m_1m_2 \ldots m_{52})_2$,

sa valeur est égale à :

$$\sum_{i=1}^{52} m_i 2^{-i} = 2^{-52}\frac{1-2^{52}}{1-2} = 1-2^{-52}$$

La valeur exacte de X^n_{max} en DP est égale à :

$$(-1)^0 \times (2-2^{-52}) \times 2^{1023} = 2^{1024} - 2^{971}$$

3.2.3 Valeur décimale approchée

La valeur décimale approchée de X^n_{max} en DP est égale à :

$$2^{1024} - 2^{971} =\approx 1.7976931348623157 \times 10^{308}$$

3.3 X^d_{max} dénormalisé en simple précision

3.3.1 Codage binaire

Sur 32 bits, on a :
Le nombre flottant, en SP sur 32 bits lus de gauche à droite :

```
(gdb) x/tw &xMax3
0x7fffffffde1c: 00000000011111111111111111111111
(gdb)
```

FIGURE 9 – Codage binaire de X^d_{max} en SP

- est positif (bit de signe à 0) ;
- est dénormalisé (l'exposant est codé par 00000000) ;
- a une mantisse codée par 11111111111111111111111.

3.3.2 Valeur exacte

- le bit de signe $s = 0$;
- la valeur de l'exposant vrai $= -126$;

A partir de l'expression binaire de la mantisse codée $(m)_2 = (0,11111111111111111111111)_2 = (0, m_1 m_2 \ldots m_{23})_2$, sa valeur est égale à :

$$\sum_{i=1}^{23} m_i 2^{-i} = 2^{-23}\frac{1-2^{23}}{1-2} = 1-2^{-23}$$

La valeur exacte de X^d_{max} en SP est égale à :

$$(-1)^0 \times (1-2^{-23}) \times 2^{-126} = 2^{-126} - 2^{-149}$$

3.3.3 Valeur décimale approchée

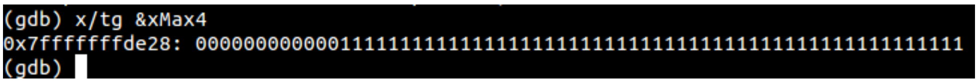
La valeur décimale approchée de X^d_{max} en SP est égale à :

$$2^{-126} - 2^{-149} \approx 1.1754942 \times 10^{-38}$$

3.4 X^d_{max} dénormalisé en double précision

3.4.1 Codage binaire

Sur 64 bits, on a :



```
(gdb) x/tg &xMax4
0x7fffffffde28: 0000000000001111111111111111111111111111111111111111111111111111
(gdb)
```

FIGURE 10 – Codage binaire de X^d_{max} en DP

Le nombre flottant, en DP sur 64 bits lus de gauche à droite :
— est positif (bit de signe à 0) ;
— est dénormalisé (l'exposant est codé par 00000000000) ;
— a une mantisse codée par 1111111111111111111111111111111111111111111111111111.

3.4.2 Valeur exacte

— le bit de signe $s = 0$;
— la valeur de l'exposant vrai $= -1022$
A partir de l'expression binaire de la mantisse codée $(m)_2 = (0,1111111111111111111111111111111111111111111111111111)_2 = (0, m_1 m_2 \dots m_{52})_2$, d'où :

$$\sum_{i=1}^{52} m_i 2^{-i} = 2^{-52} \frac{1-2^{52}}{1-2} = 1 - 2^{-52}$$

La valeur exacte de X^d_{max} en DP est égale à :

$$(-1)^0 \times (1 - 2^{-52}) \times 2^{-1022} = 2^{-1022} - 2^{-1074}$$

3.4.3 Valeur décimale approchée

La valeur décimale approchée de X^d_{max} en DP est égale à :

$$2^{-1022} - 2^{-1074} \approx 2.2250738585072009 \times 10^{-308}$$