Les nombres à virgules en C
en IEEE 754

Les **float**: 32 bits = 4 octets

- 23 bits pour la mantisse
- 8 bits pour l'exposant
- 1 bit pour le signe

Les **double**: 64 bits = 8 octets

- 52 bits pour la mantisse
- 11 bits pour l'exposant
- 1 bit pour le signe

Les **long double**: 80 bits = 10 octets

- 64 bits pour la mantisse
- 15 bits pour l'exposant
- 1 bit pour le signe

La précision des réels est approchée: $(-1)^{signe}$ x mantisse x $base^{exposant}$ (base = 2, ou 16)

- Généralement $(-1)^{signe}$ x [partie fractionnaire en base 2] x $2^{exposant}$
  - On calcule la valeur d'un flottant IEEE 754 en utilisant la formule suivante:
    - $(-1)^{signe}$ x bit implicite.[mantisse] x $2^{(exposant - décalage)}$
    - Le décalage est là pour pouvoir coder les nombres très petits
      - décalage de 127 pour les float (et plus généralement le décalage est de $2^{nb\_bit\_exposant - 1} - 1$)
    - Exemple pour un **float** normalisé: bit implicite à 1
      - valeur = $(-1)^{signe}$ x $(1+(\text{mantisse en base 2})/2^{\text{nb bits mantisse}})$ **x** $\mathbf{2^{exposant-decalage}}$
      - -3141.5 est codé en float par 1 10001010 $\mathbf{10001000101}00000000000_{base\ 2}$
      - signe = 1 => négatif
      - exposant = 10001010 => = 138 =127+11
      - valeur = -1 x $(1+(\mathbf{10001000101}00000000000_{base\ 2}\mathbf{)/2^{23}})$ x 2^{11} = - 1 x (1 + 4478976/8388608) x 2048 = -3141,5
    - Exemple pour un **float** dénormalisé (nombre très petit): bit implicite à 0
      - valeur = $(-1)^{signe}$ x $(0+(\text{mantisse en base 2})/2^{\text{nb bits mantisse}})$ **x** $\mathbf{2^{exposant-decalage+1}}$
- **La plupart du temps une erreur de calcul est introduite**
  - 6 chiffres de précision pour les float en décimal
    - 43.1337 a 6 chiffres significatifs.
    - On compte les chiffres significatifs à partir du premier chiffre de gauche différent de 0.
  - 15 chiffres de précision pour les double en décimal
  - 17 chiffres de précision pour les long double en décimal
  - **La mantisse représente la partie significative du nombre**, un nombre flottant a autant de chiffres significatifs **en base 2** que sa mantisse occupe de bits + le bit implicite.
    - Soit 23 + 1 = 24 chiffres significatifs en binaire pour les float (nombre normalisé)
    - 53 chiffres significatifs en binaire pour les double (nombre normalisé)

    - 65 chiffres significatifs en binaire pour les long double (nombre normalisé)
    - Pour les nombres dénormalisés, le nombre de chiffres significatifs dépendra du nombre de zéros au début de la mantisse, soit entre 23 et 0.

Sources:

- http://fr.wikipedia.org/wiki/IEEE_754
- http://fr.openclassrooms.com/informatique/cours/nombres-flottants-et-processeurs
- http://progdupeu.pl/tutoriels/85/les-nombres-a-virgule-flottante/#2-les-nombres-a-virgule-en-c
- http://perso.limsi.fr/pruvost/res/teach-doc/ieee754.pdf
- Page 4 du document: http://hal.inria.fr/docs/00/07/14/77/PDF/RR-5105.pdf