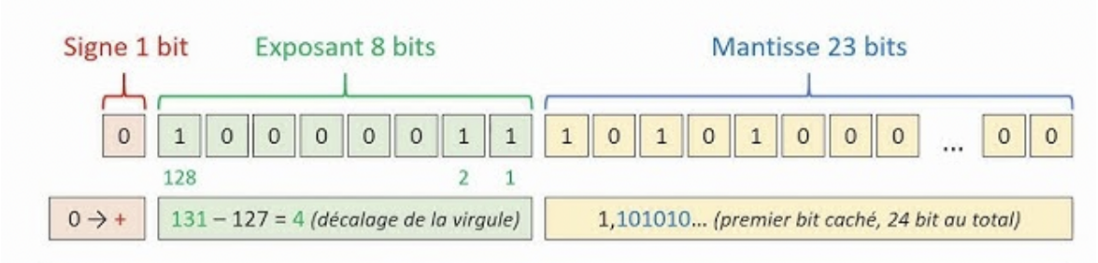
Norme IEEE 754

La norme IEEE 754 est un codage standardisé pour représenter les nombres binaires à virgules (flottantes) ainsi que pour effectuer de l'arithmétique sur ces nombres. La norme définit également les formats de représentation des valeurs spéciales (par exemple infinis et NaN). La version d'origine de la norme, datant de 1985, définissait quatre formats pour représenter des nombres à virgule flottante en base 2 : les formats simple précision ; simple précision étendue (obsolète) ; double précision et double précision étendue. [1]. Cette norme a été étendue par une révision majeure en 2008 à d'autres formats (128 bits) puis révisé en 2019.

La norme représente un nombre binaire *normalisé* [2] comme : $\pm\,(1,m)_2\,.2^{e-127}$ où le nombre flottant est formé de trois éléments : la (pseudo)mantisse m, l'exposant (biaisé) e et le signe codé sur un bit. Le bit de poids fort est le bit de signe : si ce bit est à 1, le nombre est négatif, et s'il est à 0, le nombre est positif. Les bits suivants représentent l'exposant [3] (sauf valeur spéciale), et les bits suivants (bits de poids faible) représentent la mantisse. En conséquence, un nombre décimal à virgule doit d'abord être codé en binaire (à virgule fixe) puis **obligatoirement** mis sous la forme ci-dessus pour être codé.

Un nombre flottant simple précision est stocké dans un mot de 32 bits : 1 bit de signe, 8 bits pour l'exposant et 23 bits pour la mantisse. Le format double [4] précision possède 52 bits de mantisse et 11 bits d'exposant.



Valeurs particulières

Les valeurs de l'exposant avec tous les bits à 0 ou 1 ($E = 00000000$ et $E = 11111111$ en 32-bits) sont réservées pour représenter des valeurs particulières, à savoir :

Exposant	Pseudomantisse	Valeur
Tous les bits à 0	Tous les bits à 0	Représentation de la valeur zéro.
Tous les bits à 0	Au moins un bit non nul	Représentation d'une valeur dénormalisée.
Tous les bits à 1	Tous les bits à 0	Représentation d'une valeur infinie
Tous les bits à 1	Au moins un bit non nul	Représentation d'une valeur *NaN*

1. Par exemple, dans le langage C, le compilateur gcc pour les architectures compatibles Intel 32 bits utilise le format simple précision pour les variables de type *float*, double précision pour les variables de type *double*, et la double précision ou la double précision étendue (suivant le système d'exploitation) pour les variables de type *long double*.
2. Puisque le premier chiffre est toujours 1, on ne le représente pas (on gagne 1 bit !!!)
3. L'exposant peut être positif ou négatif. Cependant, la représentation habituelle des nombres signés (complément à 2) rendrait la comparaison entre les nombres flottants un peu plus difficile. Pour régler ce problème, l'exposant est « biaisé », afin de le stocker sous forme d'un nombre non signé.
4. La mantisse est très élargie, alors que l'exposant est peu élargi. Ceci est dû au fait que, selon les créateurs de la norme, la précision est plus importante que l'amplitude.