

Chapitre 4 : Représentation approximative des nombres réels

Nous avons vu comment sont représentés les entiers relatifs au sein d'un ordinateur. Nous allons maintenant voir comment sont représentés les nombres réels, appelés ici nombres flottants.

1. Écriture de la partie décimale d'un nombre du point de vue mathématique

1.1. Généralités sur les bases 10 et 2

Un **nombre décimal** est un nombre s'écrivant sous la forme :

$$\frac{x}{10^n}$$

Exemple :

$$\frac{13}{4} = \frac{13 \times 25}{4 \times 25} = \frac{325}{100} = \frac{325}{10^2} = 3,25$$

Si l'on cherche à faire la **décomposition décimale** d'un nombre en base 10, les chiffres à gauche de la virgule correspondent à des puissances de 10 positives ou nulles, tandis que ceux à droite correspondent à des puissances de 10 négatives.

Exemples :

$$3,25 = 3 \times 10^0 + 2 \times 10^{-1} + 5 \times 10^{-2}$$

$$18,053 = 1 \times 10^1 + 8 \times 10^0 + 5 \times 10^{-2} + 3 \times 10^{-3}$$

Par analogie avec les nombres décimaux, on appelle **nombres dyadiques** les nombres s'écrivant sous la forme : $\frac{x}{2^n}$

Exemple : $\frac{13}{4} = \frac{13}{2^2}$

est un nombre dyadique. On peut en faire la **décomposition dyadique**

$$(13)_2 = (1101)_2$$

$$\left(\frac{13}{4}\right)_2 = \left(\frac{1101}{100}\right)_2 = (11,01)_2$$

1.2. Conversions

1.2.1. De la base 2 vers la base 10

Il est possible de retrouver une représentation décimale en base 10 à partir d'une représentation en binaire.

Partons de : $100,0101_2$

$$= 1 \times 2^2 + 1 \times 2^{-2} + 1 \times 2^{-4}$$

$$= 4 + \frac{1}{4} + \frac{1}{16}$$

$$= (4,3125)_{10}$$

Exercice 1 :

Trouver la représentation décimale de $100,001_2 = 1 \times 2^2 + 1 \times 2^{-3}$

1.2.2. De la base 10 vers la base 2

$$= 4 + \frac{1}{8}$$

$$= 4,125$$

Partons tout de suite d'un exemple : comment représenter 5,1875 en binaire ?

Il nous faut déjà représenter 5. Pas de souci, nous savons faire : $5 = 4 + 1 = 2^2 + 2^0 = 101$

Mais comment représenter le ",1875" ?

- on multiplie 0,1875 par 2 : $0,1875 \times 2 = 0,375$
- on multiplie 0,375 par 2 : $0,375 \times 2 = 0,75$
- on multiplie 0,75 par 2 : $0,75 \times 2 = 1,5 \rightarrow$ on regarde que la partie décimale
- on multiplie 0,5 par 2 : $0,5 \times 2 = 1,0$

Lorsque la partie décimale est à 0, on arrête le processus.

On obtient une succession de "a + 0,b" (" $0 + 0,375$ ", " $0 + 0,75$ ", " $1 + 0,5$ " et " $1 + 0,0$ ").

Il suffit maintenant de "prendre" tous les "a" dans l'ordre de leur obtention afin d'obtenir la partie décimale de notre nombre : **0011**.

On a finalement : $5,1875_{10} = 101,0011_2$

Exercice 2 :

Trouver la représentation binaire de $4,125_{10}$

$$4_{10} = 100_2$$

$$0,125 \times 2 = 0,25$$

$$0,25 \times 2 = 0,5$$

$$0,5 \times 2 = 1,0$$

$$,125_{10} = ,001$$

$$4,125_{10} = 100,001$$

* <http://www.science-phynum.fr/>

Exemple :

$$\frac{7}{11} = 0,636363...$$

Et certains ont une écriture binaire infinie périodique alors même que leur écriture décimale est finie.

Prenons par exemple 0,2 :

$$0,2 \times 2 = 0,4$$

$$0,4 \times 2 = 0,8$$

$$0,8 \times 2 = 1,6$$

$$0,6 \times 2 = 1,2$$

$$0,2 \times 2 = \dots$$

$$(0,2)_{10} = (0,001100110011001100\dots)_2$$

Exercice 3 :

Trouver la représentation binaire de : $0,1_{10} = 0,1 \times 2 = 0,2$ $0,1_{10} = (0,00011001100110\dots)_2$

Exercice 4 :

1) Donner la représentation binaire de la fraction : $\frac{11}{16} = 0,6875 \times 2 = 1,375$

2) Donner la représentation binaire de la fraction : $\frac{11}{15} = 0,7333 \times 2 = 1,4666$ $0,7333 \times 2 = 1,4666$ $0,4666 \times 2 = 0,9332$ $0,9332 \times 2 = 1,8664$ $0,8664 \times 2 = 1,7328$ $0,7328 \times 2 = 1,4656$ $0,4656 \times 2 = 0,9312$ $0,9312 \times 2 = 1,8624$ $0,8624 \times 2 = 1,7248$ $0,7248 \times 2 = 1,4496$ $0,4496 \times 2 = 0,8992$ $0,8992 \times 2 = 1,7984$ $0,7984 \times 2 = 1,5968$ $0,5968 \times 2 = 1,1936$ $0,1936 \times 2 = 0,3872$ $0,3872 \times 2 = 0,7744$ $0,7744 \times 2 = 1,5488$ $0,5488 \times 2 = 1,0976$ $0,0976 \times 2 = 0,1952$ $0,1952 \times 2 = 0,3904$ $0,3904 \times 2 = 0,7808$ $0,7808 \times 2 = 1,5616$ $0,5616 \times 2 = 1,1232$ $0,1232 \times 2 = 0,2464$ $0,2464 \times 2 = 0,4928$ $0,4928 \times 2 = 0,9856$ $0,9856 \times 2 = 1,9712$ $0,9712 \times 2 = 1,9424$ $0,9424 \times 2 = 1,8848$ $0,8848 \times 2 = 1,7696$ $0,7696 \times 2 = 1,5392$ $0,5392 \times 2 = 1,0784$ $0,0784 \times 2 = 0,1568$ $0,1568 \times 2 = 0,3136$ $0,3136 \times 2 = 0,6272$ $0,6272 \times 2 = 1,2544$ $0,2544 \times 2 = 0,5088$ $0,5088 \times 2 = 1,0176$ $0,0176 \times 2 = 0,0352$ $0,0352 \times 2 = 0,0704$ $0,0704 \times 2 = 0,1408$ $0,1408 \times 2 = 0,2816$ $0,2816 \times 2 = 0,5632$ $0,5632 \times 2 = 1,1264$ $0,1264 \times 2 = 0,2528$ $0,2528 \times 2 = 0,5056$ $0,5056 \times 2 = 1,0112$ $0,0112 \times 2 = 0,0224$ $0,0224 \times 2 = 0,0448$ $0,0448 \times 2 = 0,0896$ $0,0896 \times 2 = 0,1792$ $0,1792 \times 2 = 0,3584$ $0,3584 \times 2 = 0,7168$ $0,7168 \times 2 = 1,4336$ $0,4336 \times 2 = 0,8672$ $0,8672 \times 2 = 1,7344$ $0,7344 \times 2 = 1,4688$ $0,4688 \times 2 = 0,9376$ $0,9376 \times 2 = 1,8752$ $0,8752 \times 2 = 1,7504$ $0,7504 \times 2 = 1,5008$ $0,5008 \times 2 = 1,0016$ $0,0016 \times 2 = 0,0032$ $0,0032 \times 2 = 0,0064$ $0,0064 \times 2 = 0,0128$ $0,0128 \times 2 = 0,0256$ $0,0256 \times 2 = 0,0512$ $0,0512 \times 2 = 0,1024$ $0,1024 \times 2 = 0,2048$ $0,2048 \times 2 = 0,4096$ $0,4096 \times 2 = 0,8192$ $0,8192 \times 2 = 1,6384$ $0,6384 \times 2 = 1,2768$ $0,2768 \times 2 = 0,5536$ $0,5536 \times 2 = 1,1072$ $0,1072 \times 2 = 0,2144$ $0,2144 \times 2 = 0,4288$ $0,4288 \times 2 = 0,8576$ $0,8576 \times 2 = 1,7152$ $0,7152 \times 2 = 1,4304$ $0,4304 \times 2 = 0,8608$ $0,8608 \times 2 = 1,7216$ $0,7216 \times 2 = 1,4432$ $0,4432 \times 2 = 0,8864$ $0,8864 \times 2 = 1,7728$ $0,7728 \times 2 = 1,5456$ $0,5456 \times 2 = 1,0912$ $0,0912 \times 2 = 0,1824$ $0,1824 \times 2 = 0,3648$ $0,3648 \times 2 = 0,7296$ $0,7296 \times 2 = 1,4592$ $0,4592 \times 2 = 0,9184$ $0,9184 \times 2 = 1,8368$ $0,8368 \times 2 = 1,6736$ $0,6736 \times 2 = 1,3472$ $0,3472 \times 2 = 0,6944$ $0,6944 \times 2 = 1,3888$ $0,3888 \times 2 = 0,7776$ $0,7776 \times 2 = 1,5552$ $0,5552 \times 2 = 1,1104$ $0,1104 \times 2 = 0,2208$ $0,2208 \times 2 = 0,4416$ $0,4416 \times 2 = 0,8832$ $0,8832 \times 2 = 1,7664$ $0,7664 \times 2 = 1,5328$ $0,5328 \times 2 = 1,0656$ $0,0656 \times 2 = 0,1312$ $0,1312 \times 2 = 0,2624$ $0,2624 \times 2 = 0,5248$ $0,5248 \times 2 = 1,0496$ $0,0496 \times 2 = 0,0992$ $0,0992 \times 2 = 0,1984$ $0,1984 \times 2 = 0,3968$ $0,3968 \times 2 = 0,7936$ $0,7936 \times 2 = 1,5872$ $0,5872 \times 2 = 1,1744$ $0,1744 \times 2 = 0,3488$ $0,3488 \times 2 = 0,6976$ $0,6976 \times 2 = 1,3952$ $0,3952 \times 2 = 0,7904$ $0,7904 \times 2 = 1,5808$ $0,5808 \times 2 = 1,1616$ $0,1616 \times 2 = 0,3232$ $0,3232 \times 2 = 0,6464$ $0,6464 \times 2 = 1,2928$ $0,2928 \times 2 = 0,5856$ $0,5856 \times 2 = 1,1712$ $0,1712 \times 2 = 0,3424$ $0,3424 \times 2 = 0,6848$ $0,6848 \times 2 = 1,3696$ $0,3696 \times 2 = 0,7392$ $0,7392 \times 2 = 1,4784$ $0,4784 \times 2 = 0,9568$ $0,9568 \times 2 = 1,9136$ $0,9136 \times 2 = 1,8272$ $0,8272 \times 2 = 1,6544$ $0,6544 \times 2 = 1,3088$ $0,3088 \times 2 = 0,6176$ $0,6176 \times 2 = 1,2352$ $0,2352 \times 2 = 0,4704$ $0,4704 \times 2 = 0,9408$ $0,9408 \times 2 = 1,8816$ $0,8816 \times 2 = 1,7632$ $0,7632 \times 2 = 1,5264$ $0,5264 \times 2 = 1,0528$ $0,0528 \times 2 = 0,1056$ $0,1056 \times 2 = 0,2112$ $0,2112 \times 2 = 0,4224$ $0,4224 \times 2 = 0,8448$ $0,8448 \times 2 = 1,6896$ $0,6896 \times 2 = 1,3792$ $0,3792 \times 2 = 0,7584$ $0,7584 \times 2 = 1,5168$ $0,5168 \times 2 = 1,0336$ $0,0336 \times 2 = 0,0672$ $0,0672 \times 2 = 0,1344$ $0,1344 \times 2 = 0,2688$ $0,2688 \times 2 = 0,5376$ $0,5376 \times 2 = 1,0752$ $0,0752 \times 2 = 0,1504$ $0,1504 \times 2 = 0,3008$ $0,3008 \times 2 = 0,6016$ $0,6016 \times 2 = 1,2032$ $0,2032 \times 2 = 0,4064$ $0,4064 \times 2 = 0,8128$ $0,8128 \times 2 = 1,6256$ $0,6256 \times 2 = 1,2512$ $0,2512 \times 2 = 0,5024$ $0,5024 \times 2 = 1,0048$ $0,0048 \times 2 = 0,0096$ $0,0096 \times 2 = 0,0192$ $0,0192 \times 2 = 0,0384$ $0,0384 \times 2 = 0,0768$ $0,0768 \times 2 = 0,1536$ $0,1536 \times 2 = 0,3072$ $0,3072 \times 2 = 0,6144$ $0,6144 \times 2 = 1,2288$ $0,2288 \times 2 = 0,4576$ $0,4576 \times 2 = 0,9152$ $0,9152 \times 2 = 1,8304$ $0,8304 \times 2 = 1,6608$ $0,6608 \times 2 = 1,3216$ $0,3216 \times 2 = 0,6432$ $0,6432 \times 2 = 1,2864$ $0,2864 \times 2 = 0,5728$ $0,5728 \times 2 = 1,1456$ $0,1456 \times 2 = 0,2912$ $0,2912 \times 2 = 0,5824$ $0,5824 \times 2 = 1,1648$ $0,1648 \times 2 = 0,3296$ $0,3296 \times 2 = 0,6592$ $0,6592 \times 2 = 1,3184$ $0,3184 \times 2 = 0,6368$ $0,6368 \times 2 = 1,2736$ $0,2736 \times 2 = 0,5472$ $0,5472 \times 2 = 1,0944$ $0,0944 \times 2 = 0,1888$ $0,1888 \times 2 = 0,3776$ $0,3776 \times 2 = 0,7552$ $0,7552 \times 2 = 1,5104$ $0,5104 \times 2 = 1,0208$ $0,0208 \times 2 = 0,0416$ $0,0416 \times 2 = 0,0832$ $0,0832 \times 2 = 0,1664$ $0,1664 \times 2 = 0,3328$ $0,3328 \times 2 = 0,6656$ $0,6656 \times 2 = 1,3312$ $0,3312 \times 2 = 0,6624$ $0,6624 \times 2 = 1,3248$ $0,3248 \times 2 = 0,6496$ $0,6496 \times 2 = 1,2992$ $0,2992 \times 2 = 0,5984$ $0,5984 \times 2 = 1,1968$ $0,1968 \times 2 = 0,3936$ $0,3936 \times 2 = 0,7872$ $0,7872 \times 2 = 1,5744$ $0,5744 \times 2 = 1,1488$ $0,1488 \times 2 = 0,2976$ $0,2976 \times 2 = 0,5952$ $0,5952 \times 2 = 1,1904$ $0,1904 \times 2 = 0,3808$ $0,3808 \times 2 = 0,7616$ $0,7616 \times 2 = 1,5232$ $0,5232 \times 2 = 1,0464$ $0,0464 \times 2 = 0,0928$ $0,0928 \times 2 = 0,1856$ $0,1856 \times 2 = 0,3712$ $0,3712 \times 2 = 0,7424$ $0,7424 \times 2 = 1,4848$ $0,4848 \times 2 = 0,9696$ $0,9696 \times 2 = 1,9392$ $0,9392 \times 2 = 1,8784$ $0,8784 \times 2 = 1,7568$ $0,7568 \times 2 = 1,5136$ $0,5136 \times 2 = 1,0272$ $0,0272 \times 2 = 0,0544$ $0,0544 \times 2 = 0,1088$ $0,1088 \times 2 = 0,2176$ $0,2176 \times 2 = 0,4352$ $0,4352 \times 2 = 0,8704$ $0,8704 \times 2 = 1,7408$ $0,7408 \times 2 = 1,4816$ $0,4816 \times 2 = 0,9632$ $0,9632 \times 2 = 1,9264$ $0,9264 \times 2 = 1,8528$ $0,8528 \times 2 = 1,7056$ $0,7056 \times 2 = 1,4112$ $0,4112 \times 2 = 0,8224$ $0,8224 \times 2 = 1,6448$ $0,6448 \times 2 = 1,2896$ $0,2896 \times 2 = 0,5792$ $0,5792 \times 2 = 1,1584$ $0,1584 \times 2 = 0,3168$ $0,3168 \times 2 = 0,6336$ $0,6336 \times 2 = 1,2672$ $0,2672 \times 2 = 0,5344$ $0,5344 \times 2 = 1,0688$ $0,0688 \times 2 = 0,1376$ $0,1376 \times 2 = 0,2752$ $0,2752 \times 2 = 0,5504$ $0,5504 \times 2 = 1,1008$ $0,1008 \times 2 = 0,2016$ $0,2016 \times 2 = 0,4032$ $0,4032 \times 2 = 0,8064$ $0,8064 \times 2 = 1,6128$ $0,6128 \times 2 = 1,2256$ $0,2256 \times 2 = 0,4512$ $0,4512 \times 2 = 0,9024$ $0,9024 \times 2 = 1,8048$ $0,8048 \times 2 = 1,6096$ $0,6096 \times 2 = 1,2192$ $0,2192 \times 2 = 0,4384$ $0,4384 \times 2 = 0,8768$ $0,8768 \times 2 = 1,7536$ $0,7536 \times 2 = 1,5072$ $0,5072 \times 2 = 1,0144$ $0,0144 \times 2 = 0,0288$ $0,0288 \times 2 = 0,0576$ $0,0576 \times 2 = 0,1152$ $0,1152 \times 2 = 0,2304$ $0,2304 \times 2 = 0,4608$ $0,4608 \times 2 = 0,9216$ $0,9216 \times 2 = 1,8432$ $0,8432 \times 2 = 1,6864$ $0,6864 \times 2 = 1,3728$ $0,3728 \times 2 = 0,7456$ $0,7456 \times 2 = 1,4912$ $0,4912 \times 2 = 0,9824$ $0,9824 \times 2 = 1,9648$ $0,9648 \times 2 = 1,9296$ $0,9296 \times 2 = 1,8592$ $0,8592 \times 2 = 1,7184$ $0,7184 \times 2 = 1,4368$ $0,4368 \times 2 = 0,8736$ $0,8736 \times 2 = 1,7472$ $0,7472 \times 2 = 1,4944$ $0,4944 \times 2 = 0,9888$ $0,9888 \times 2 = 1,9776$ $0,9776 \times 2 = 1,9552$ $0,9552 \times 2 = 1,9104$ $0,9104 \times 2 = 1,8208$ $0,8208 \times 2 = 1,6416$ $0,6416 \times 2 = 1,2832$ $0,2832 \times 2 = 0,5664$ $0,5664 \times 2 = 1,1328$ $0,1328 \times 2 = 0,2656$ $0,2656 \times 2 = 0,5312$ $0,5312 \times 2 = 1,0624$ $0,0624 \times 2 = 0,1248$ $0,1248 \times 2 = 0,2496$ $0,2496 \times 2 = 0,4992$ $0,4992 \times 2 = 0,9984$ $0,9984 \times 2 = 1,9968$ $0,9968 \times 2 = 1,9936$ $0,9936 \times 2 = 1,9872$ $0,9872 \times 2 = 1,9744$ $0,9744 \times 2 = 1,9488$ $0,9488 \times 2 = 1,8976$ $0,8976 \times 2 = 1,7952$ $0,7952 \times 2 = 1,5904$ $0,5904 \times 2 = 1,1808$ $0,1808 \times 2 = 0,3616$ $0,3616 \times 2 = 0,7232$ $0,7232 \times 2 = 1,4464$ $0,4464 \times 2 = 0,8928$ $0,8928 \times 2 = 1,7856$ $0,7856 \times 2 = 1,5712$ $0,5712 \times 2 = 1,1424$ $0,1424 \times 2 = 0,2848$ $0,2848 \times 2 = 0,5696$ $0,5696 \times 2 = 1,1392$ $0,1392 \times 2 = 0,2784$ $0,2784 \times 2 = 0,5568$ $0,5568 \times 2 = 1,1136$ $0,1136 \times 2 = 0,2272$ $0,2272 \times 2 = 0,4544$ $0,4544 \times 2 = 0,9088$ $0,9088 \times 2 = 1,8176$ $0,8176 \times 2 = 1,6352$ $0,6352 \times 2 = 1,2704$ $0,2704 \times 2 = 0,5408$ $0,5408 \times 2 = 1,0816$ $0,0816 \times 2 = 0,1632$ $0,1632 \times 2 = 0,3264$ $0,3264 \times 2 = 0,6528$ $0,6528 \times 2 = 1,3056$ $0,3056 \times 2 = 0,6112$ $0,6112 \times 2 = 1,2224$ $0,2224 \times 2 = 0,4448$ $0,4448 \times 2 = 0,8896$ $0,8896 \times 2 = 1,7792$ $0,7792 \times 2 = 1,5584$ $0,5584 \times 2 = 1,1168$ $0,1168 \times 2 = 0,2336$ $0,2336 \times 2 = 0,4672$ $0,4672 \times 2 = 0,9344$ $0,9344 \times 2 = 1,8688$ $0,8688 \times 2 = 1,7376$ $0,7376 \times 2 = 1,4752$ $0,4752 \times 2 = 0,9504$ $0,9504 \times 2 = 1,9008$ $0,9008 \times 2 = 1,8016$ $0,8016 \times 2 = 1,6032$ $0,6032 \times 2 = 1,2064$ $0,2064 \times 2 = 0,4128$ $0,4128 \times 2 = 0,8256$ $0,8256 \times 2 = 1,6512$ $0,6512 \times 2 = 1,3024$ $0,3024 \times 2 = 0,6048$ $0,6048 \times 2 = 1,2096$ $0,2096 \times 2 = 0,4192$ $0,4192 \times 2 = 0,8384$ $0,8384 \times 2 = 1,6768$ $0,6768 \times 2 = 1,3536$ $0,3536 \times 2 = 0,7072$ $0,7072 \times 2 = 1,4144$ $0,4144 \times 2 = 0,8288$ $0,8288 \times 2 = 1,6576$ $0,6576 \times 2 = 1,3152$ $0,3152 \times 2 = 0,6304$ $0,6304 \times 2 = 1,2608$ $0,2608 \times 2 = 0,5216$ $0,5216 \times 2 = 1,0432$ $0,0432 \times 2 = 0,0864$ $0,0864 \times 2 = 0,1728$ $0,1728 \times 2 = 0,3456$ $0,3456 \times 2 = 0,6912$ $0,6912 \times 2 = 1,3824$ $0,3824 \times 2 = 0,7648$ $0,7648 \times 2 = 1,5296$ $0,5296 \times 2 = 1,0592$ $0,0592 \times 2 = 0,1184$ $0,1184 \times 2 = 0,2368$ $0,2368 \times 2 = 0,4736$ $0,4736 \times 2 = 0,9472$ $0,9472 \times 2 = 1,8944$ $0,8944 \times 2 = 1,7888$ $0,7888 \times 2 = 1,5776$ $0,5776 \times 2 = 1,1552$ $0,1552 \times 2 = 0,3104$ $0,3104 \times 2 = 0,6208$ $0,6208 \times 2 = 1,2416$ $0,2416 \times 2 = 0,4832$ $0,4832 \times 2 = 0,9664$ $0,9664 \times 2 = 1,9328$ $0,9328 \times 2 = 1,8656$ $0,8656 \times 2 = 1,7312$ $0,7312 \times 2 = 1,4624$ $0,4624 \times 2 = 0,9248$ $0,9248 \times 2 = 1,8496$ $0,8496 \times 2 = 1,6992$ $0,6992 \times 2 = 1,3984$ $0,3984 \times 2 = 0,7968$ $0,7968 \times 2 = 1,5936$ $0,5936 \times 2 = 1,1872$ $0,1872 \times 2 = 0,3744$ $0,3744 \times 2 = 0,7488$ $0,7488 \times 2 = 1,4976$ $0,4976 \times 2 = 0,9952$ $0,9952 \times 2 = 1,9904$ $0,9904 \times 2 = 1,9808$ $0,9808 \times 2 = 1,9616$ $0,9616 \times 2 = 1,9232$ $0,9232 \times 2 = 1,8464$ $0,8464 \times 2 = 1,6928$ $0,6928 \times 2 = 1,3856$ $0,3856 \times 2 = 0,7712$ $0,7712 \times 2 = 1,5424$ $0,5424 \times 2 = 1,0848$ $0,0848 \times 2 = 0,1696$ $0,1696 \times 2 = 0,3392$ $0,3392 \times 2 = 0,6784$ $0,6784 \times 2 = 1,3568$ $0,3568 \times 2 = 0,7136$ $0,7136 \times 2 = 1,4272$ $0,4272 \times 2 = 0,8544$ $0,8544 \times 2 = 1,7088$ $0,7088 \times 2 = 1,4176$ $0,4176 \times 2 = 0,8352$ $0,8352 \times 2 = 1,6704$ $0,6704 \times 2 = 1,3408$ $0,3408 \times 2 = 0,6816$ $0,6816 \times 2 = 1,3632$ $0,3632 \times 2 = 0,7264$ $0,7264 \times 2 = 1,4528$ $0,4528 \times 2 = 0,9056$ $0,9056 \times 2 = 1,8112$ $0,8112 \times 2 = 1,6224$ $0,6224 \times 2 = 1,2448$ $0,2448 \times 2 = 0,4896$ $0,4896 \times 2 = 0,9792$ $0,9792 \times 2 = 1,9584$ $0,9584 \times 2 = 1,9168$ $0,9168 \times 2 = 1,8336$ $0,8336 \times 2 = 1,6672$ $0,6672 \times 2 = 1,3344$ $0,3344 \times 2 = 0,6688$ $0,6688 \times 2 = 1,3376$ $0,3376 \times 2 = 0,6752$ $0,6752 \times 2 = 1,3504$ $0,3504 \times 2 = 0,7008$ $0,7008 \times 2 = 1,4016$ $0,4016 \times 2 = 0,8032$ $0,8032 \times 2 = 1,6064$ $0,6064 \times 2 = 1,2128$ $0,2128 \times 2 = 0,4256$ $0,4256 \times 2 = 0,8512$ $0,8512 \times 2 = 1,7024$ $0,7024 \times 2 = 1,4048$ $0,4048 \times 2 = 0,8096$ $0,8096 \times 2 = 1,6192$ $0,6192 \times 2 = 1,2384$ $0,2384 \times 2 = 0,4768$ $0,4768 \times 2 = 0,9536$ $0,9536 \times 2 = 1,9072$ $0,9$

Pour passer de $1101,1001_2$ à $1,1011001_2$, nous avons décalé la virgule de 3 rangs vers la gauche, d'où le 2^{11_2} (attention à ne pas oublier que nous travaillons en base 2 : le 11_2 correspond bien à un décalage de 3 rangs de la virgule).

Si l'on désire décaler la virgule vers la droite, il va être nécessaire d'utiliser des puissances de deux négatives :

$0,01110_2 = 1,10 \times 2^{-2} = 1,10 \times 2^{-10_2}$. Nous décalons la virgule de 2 rangs vers la droite, d'où le -10_2 .

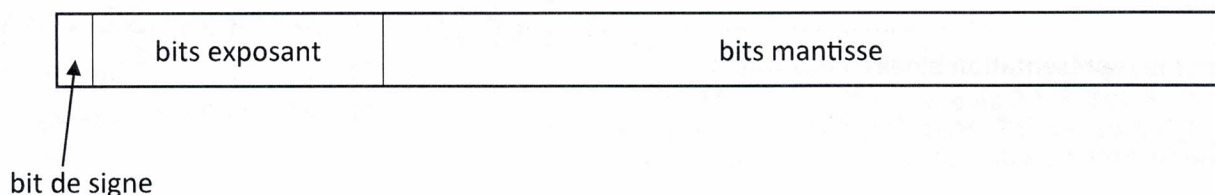
La norme **IEEE 754** est la norme (sa première version de cette norme date de 1985) la plus employée pour la **représentation des nombres à virgule flottante** dans le domaine informatique. Elle est basée sur cette **représentation en écriture scientifique**.

Nous allons étudier deux formats associés à cette norme : le format dit « simple précision » et le format dit « double précision ».

Le format « **simple précision** » utilise **32 bits** pour écrire un nombre flottant alors que le format « **double précision** » utilise **64 bits**. Dans la suite nous travaillerons principalement sur le format 32 bits.

Que cela soit en simple ou en double précision, la norme IEEE754 utilise :

- 1 bit de **signe** (1 si le nombre est négatif et 0 si le nombre est positif) ;
- des bits consacrés à l'**exposant** (8 bits pour la simple précision et 11 bits pour la double précision) ;
- des bits consacrés à la mantisse (23 bits pour la simple précision et 52 bits pour la double précision).



Nous pouvons vérifier que l'on a bien $1 + 8 + 23 = 32$ bits pour la simple précision et $1 + 11 + 52 = 64$ bits pour la double précision.

Pour écrire un nombre flottant en respectant la norme IEEE754, il est nécessaire de commencer par écrire le nombre sous la forme :

$$1, fraction_2 \times 2^{exposant_2}$$

Il faut obligatoirement qu'il y ait un seul chiffre à gauche de la virgule et il faut que ce chiffre soit un « 1 ».

Exemples :

Le nombre $1010,11001_2$ devra être écrit $1,01011001 \times 2^{11_2}$

De même, le nombre $0,00001001_2$ devra être écrit $1,001 \times 2^{-10_2}$

2.1.1. Pour la suite, nous analyserons la représentation machine de notre exemple $1010,11001_2$

On rappelle tout d'abord sa notation scientifique :

Partie bit de signe :

On met le premier bit à 0 s'il s'agit d'un nombre positif et à 1 s'il s'agit d'un nombre négatif.

Partie bits mantisse :

La partie **fraction** de 1, $fraction_2 \times 2^{exposant_2}$ constitue la mantisse.

Dans notre exemple $1010,11001_2$, la mantisse est 01011001.

Comme la mantisse comporte 23 bits en simple précision, il faudra compléter avec le nombre de zéros à droite nécessaire afin d'atteindre les 23 bits.

Toujours dans notre exemple avec pour partie fraction 01011001, il faudra ajouter $23 - 8 = 15$ zéros à droite, ce qui donnera en fin de compte pour mantisse 01011001000000000000000.

Partie bits exposant :

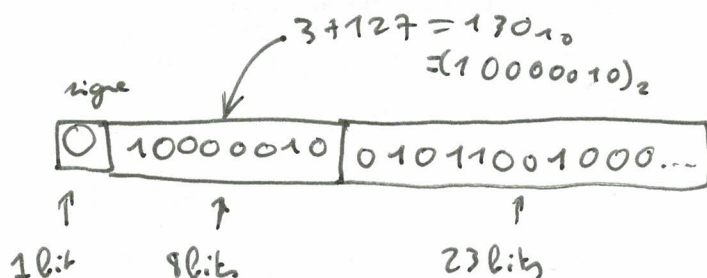
Notre première intuition serait de dire que la partie « bits exposant » correspond rigoureusement à la partie exposant de l'écriture scientifique.

Dans notre exemple $1010,11001_2 = 1,01011001_2 \times 2^{11_2}$, nous aurions 11.

En fait, c'est un peu plus compliqué que cela. En effet, comment représenter les exposants négatifs ?

Aucun bit pour le signe de l'exposant n'a été prévu dans la norme IEEE754, une autre solution a été choisie. Pour le format simple précision, 8 bits sont consacrés à l'exposant, il est donc possible de représenter 256 valeurs ($2^8 = 256$). Nous allons pouvoir représenter des exposants compris entre -126 et +127 (les valeurs -127 et +128 sont en fait des valeurs réservées, nous n'aborderons pas ce sujet ici). Pour avoir des valeurs uniquement positives, il **va falloir procéder à un décalage : ajouter systématiquement 127 à la valeur de l'exposant.**

Dans la suite, afin de simplifier les choses nous commencerons par écrire les exposants en base 10 avant de les passer en base 2 une fois le décalage effectué.



À noter que pour le format double précision le décalage est de 1023 (il faut systématiquement ajouter 1023 à l'exposant afin d'obtenir uniquement des valeurs positives).

2.1.2. Écrivons notre premier nombre au format simple précision.

Soit le nombre -10,125 en base 10. Représentons-le au format simple précision :

- Conversion en base 2 :

$$10 = 5 \times 2 + 0$$

$$5 = 2 \times 2 + 1$$

$$2 = 1 \times 2 + 0$$

$$1 = 0 \times 2 + 1$$

$$10 = 1010$$

$$0,125 \times 2 = 0,25$$

$$0,25 \times 2 = 0,5$$

$$0,5 \times 2 = 1,0$$

$$0,125 = 001$$

$$10,125 = 1010,001$$

- Écriture scientifique pour la mantisse : $1,010001_2 \times 2^{11_2}$

- Ajout de 127 à l'exposant : $127 + 3 = (130)_{10} = (1000010)_2$

- Concaténation (signe | exposant | mantisse) :

1 10000010 010001000000000000000000

32 bits

Remarque :

Nous n'avons étudié ici que les nombres dits **normalisés** de la norme IEEE 754.

La norme est en réalité un peu plus compliquée : d'une part il existe des **valeurs spéciales réservées** telles que +0, -0, +∞, -∞, NaN (« Not a Number » pour $\frac{1}{0}$ ou $\sqrt{-1}$ par exemple) ; d'autre part il existe des nombres dénormalisés permettant de représenter des nombres **très proches de zéros**.

Exercice 5 :

Déterminez la représentation au format simple précision de $0,25_{10}$ en binaire.

$$0,25 \times 2 = 0,5$$

$$0,5 \times 2 = 1,0$$

$$0,25 = 001$$

$$10 \times 2^{-10_2}$$

$$-2 + 127 = 125_{10} = (0111101)_2$$

00111101 000000000000000000000000

Remarque :

Il est aussi possible de passer d'une représentation au format IEEE 754 à une représentation classique en base 10.

Soit le nombre flottant au format simple précision : 00111110100000000000000000000000. Nous pouvons déjà dire que ce nombre est positif (bit de signe à 0). Les 8 bits suivants 01111101 nous donnent l'exposant décalé :

01111101 = 125, soit une fois le décalage supprimé, $125 - 127 = -2$. Les 23 bits suivants (la mantisse) sont uniquement des zéros, ce qui nous donne en fin de compte :

$$1,0 \times 2^{-2} = 1 \times \frac{1}{2^2} = \frac{1}{4} = 0,25$$

On retrouve bien le résultat de l'exercice précédent.

2.2. Les problèmes engendrés

Des problèmes de nombres flottants ont par le passé causé des dégâts économiques et humains. Commençons par en analyser les origines...

Exercice 6 :

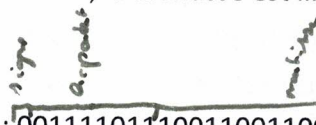
Déterminez la représentation au format simple précision de $0,1_{10}$ en binaire.

Nous avons ici **un problème** :

Comme déjà évoqué, nous nous retrouvons avec une "conversion" qui ne s'arrête jamais (le schéma "0011" se répète à "l'infini"). En effet, en simple précision, la mantisse est limitée à 23 bits.

Exercice 7 :

Soit le nombre flottant au format simple précision : 00111101110011001100110011001100. Trouvez la représentation en base 10 de ce nombre.



En toute logique, nous devrions trouver 0,1.

A cause de la limitation de la mantisse à 23 bits, nous avons dû "tronquer" notre résultat. Cette **représentation avec un nombre limité de bits des nombres flottants** est un problème classique en informatique. Cela peut entraîner des erreurs d'arrondis dans les calculs qui peuvent être très gênants si l'on n'y prend pas garde.

Exercice 8 :

Tapez dans la console dans Repl.it : $0.1 + 0.2$

Que constatez-vous ? Comment l'expliquez-vous en quelques mots ?

0,30000000000000004

ils bouclent

Exercice 9 :

Tapez dans la console dans Repl.it : $9007199254740992.0 + 1.0 + 1.0$

Puis : $1.0 + 1.0 + 9007199254740992.0$

Qu'obtient-on ? L'addition sur les flottants est-elle commutative ?

9007199254740994,0

* <http://www.science-phynum.fr/>

Exercice 10 :

Tapez dans la console dans Repl.it :

```
9007199254740992.0 + 1.0 == 9007199254740992.0
```

Que constatez-vous ?

Exercice 11 :

Certaines expressions peuvent générer des valeurs spéciales, comme « inf » ou « nan » (Not a Number).

1. Tapez les commandes suivantes : $x = 1e200$ et $x*x$ (remarque : $1e200$ signifie 1×10^{200}).

2. Puis : $(x * x) * 0$

Exercice 12 :

Tapez dans la console les commandes : $y = 10**400$ et $y + 0.5$

Que constatez-vous ? Analysez l'erreur engendrée.

Des solutions cependant...

De manière générale, il faudra faire attention lorsqu'on effectue des tests comme :

```
x = 0.1 + 0.2
```

```
y = 0.3
```

```
if x == y : car celui-ci renverrait 'False'.
```

On préférera utiliser quelque chose du type :

```
if abs(x-y) < 1e-12
```

On pourra également utiliser des bibliothèques spécialement conçues pour travailler avec des calculs exacts ou des calculs dont on pourra choisir le niveau de précision voulue. Attention cependant, le fait de travailler sur des flottants permet d'optimiser la vitesse de calcul du processeur, et leur relativement grande précision nous permet de les utiliser dans la plupart des situations de la vie courante.