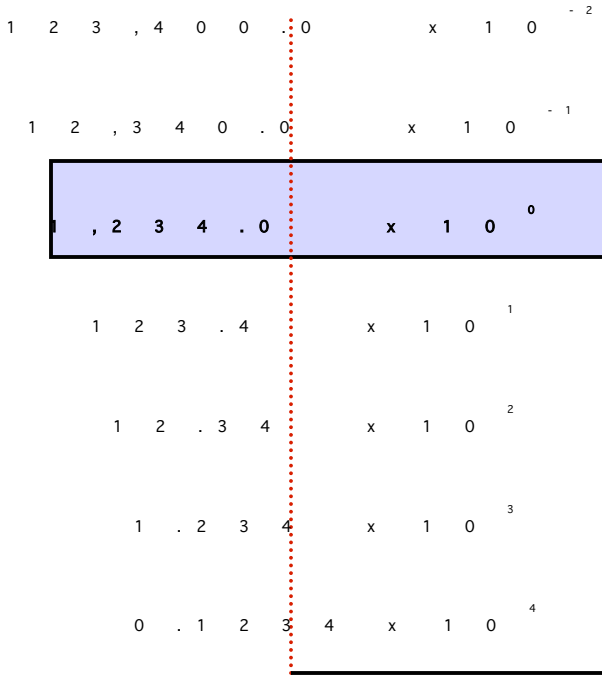
Représentation des nombres flottants

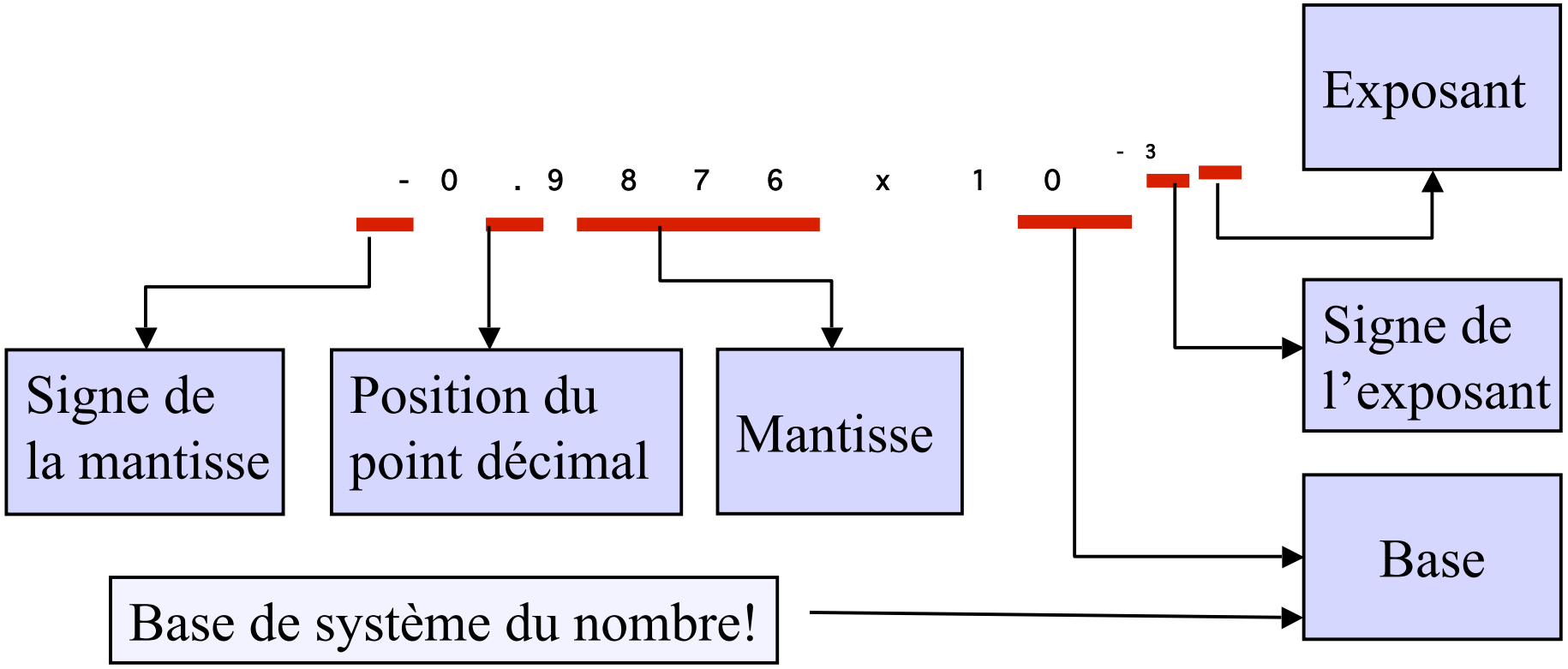
Notation exponentielle

- Représentations équivalentes dans la base 10 de 1,234



Le point décimal “flotte” (ajustement approprié de l’exposant).

Éléments de la notation exponentielle



Représentation normalisée

- Un nombre représenté en virgule flottante est normalisé s'il est sous la forme:
  - $\pm$ 0,M * $X^{\pm c}$
  - M – un nombre dont le premier chiffre est non nul
- Exemple:
  - + 59,4151 * 10^{-5} =>

    Normalisé: +0,594151 * 10^{-3}

Représentation de l'exposant et de son signe

- L'exposant est translatée de manière à toujours coder en interne une valeur positive
- Avec 2 digits réservés au codage de l'exposant
  - Les valeurs positives: [+0, +99]
  - En appliquant une translation k=50:
    - Les exposants représentables => [-50,49]
- La constante k est appelée constante d'excentrement

Représentation en virgule flottante

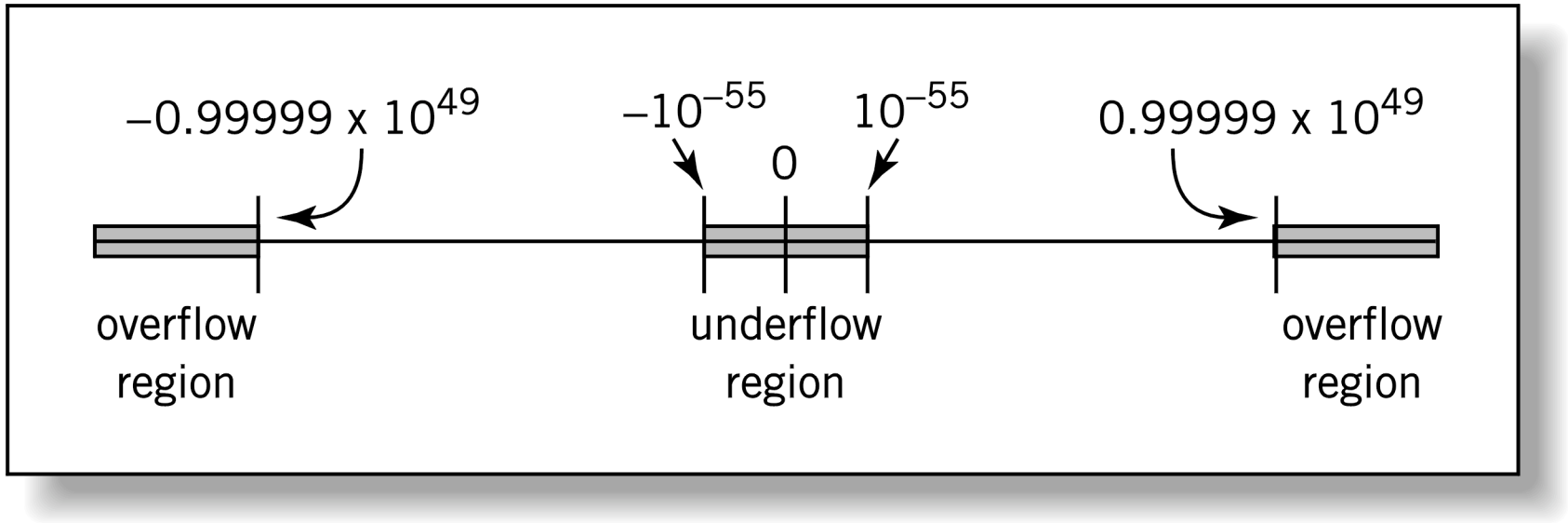
- Avec 2 digits réservés au codage de l'exposant avec un excentrement égal à 50_{10} et 5 digits pour la mantisse on peut représenter
  - de .00001 x 10^{-50} à .99999 x 10^{49}

Englander: The Architecture of Computer Hardware and Systems Software, 2nd edition
Chapter 5, Figure 05-01

Overflows / Underflows

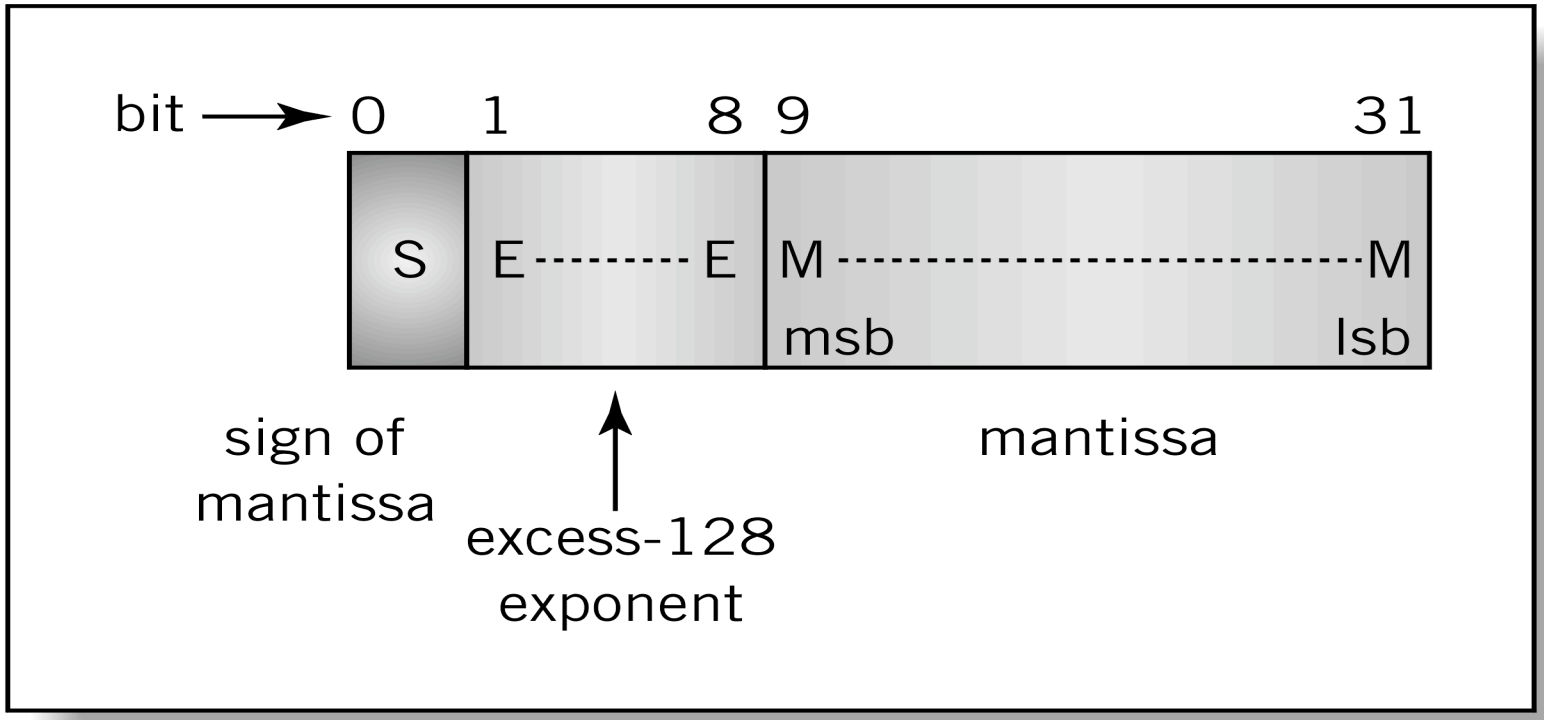
- De $.00001 \times 10^{-50}$ à $.99999 \times 10^{49}$

  1×10^{-55} à $.99999 \times 10^{49}$



Englander: The Architecture of Computer Hardware and Systems Software, 2nd edition
Chapter 5, Figure 05-02

Format typique

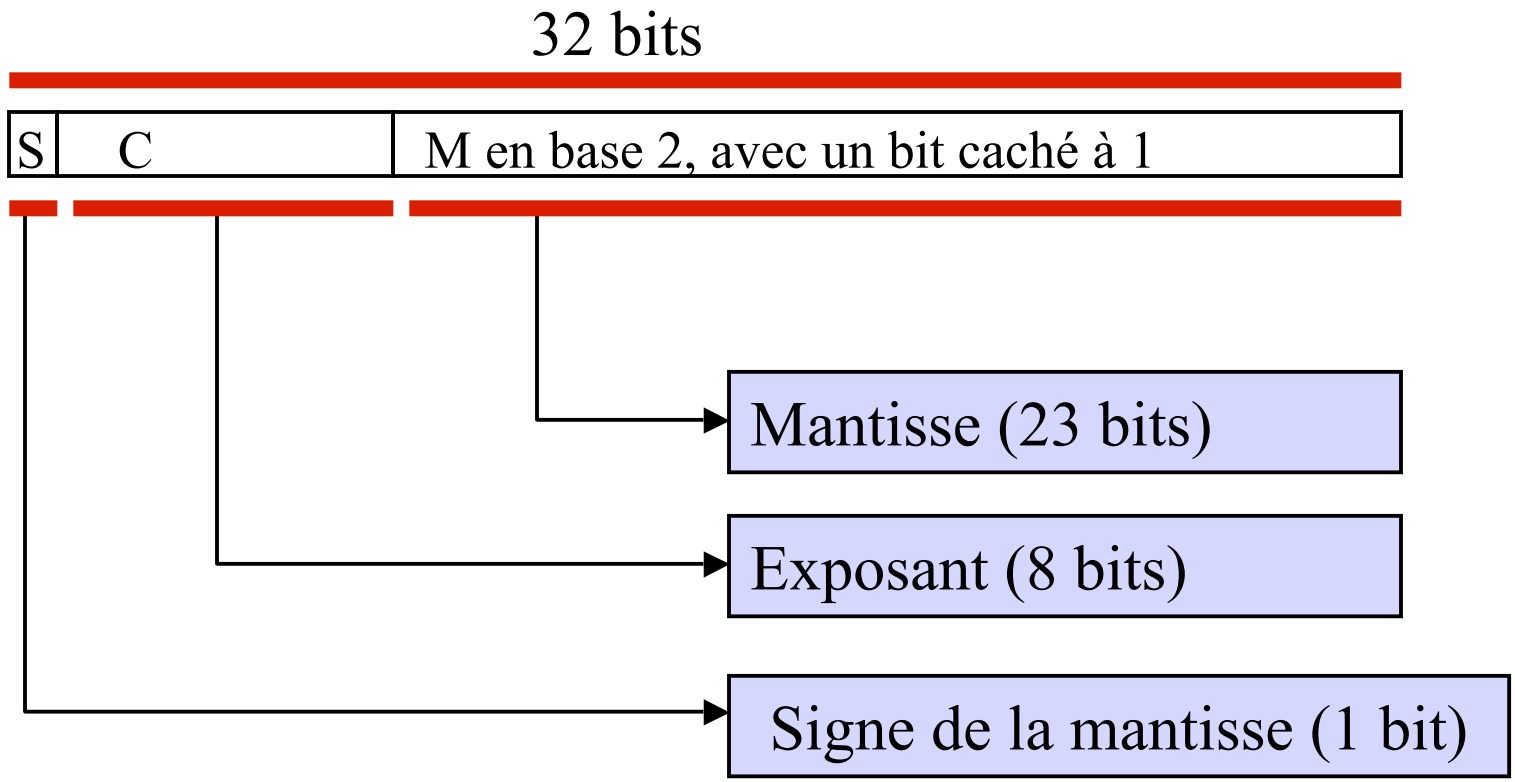


Englander: The Architecture of Computer Hardware and Systems Software, 2nd edition
Chapter 5, Figure 05-04

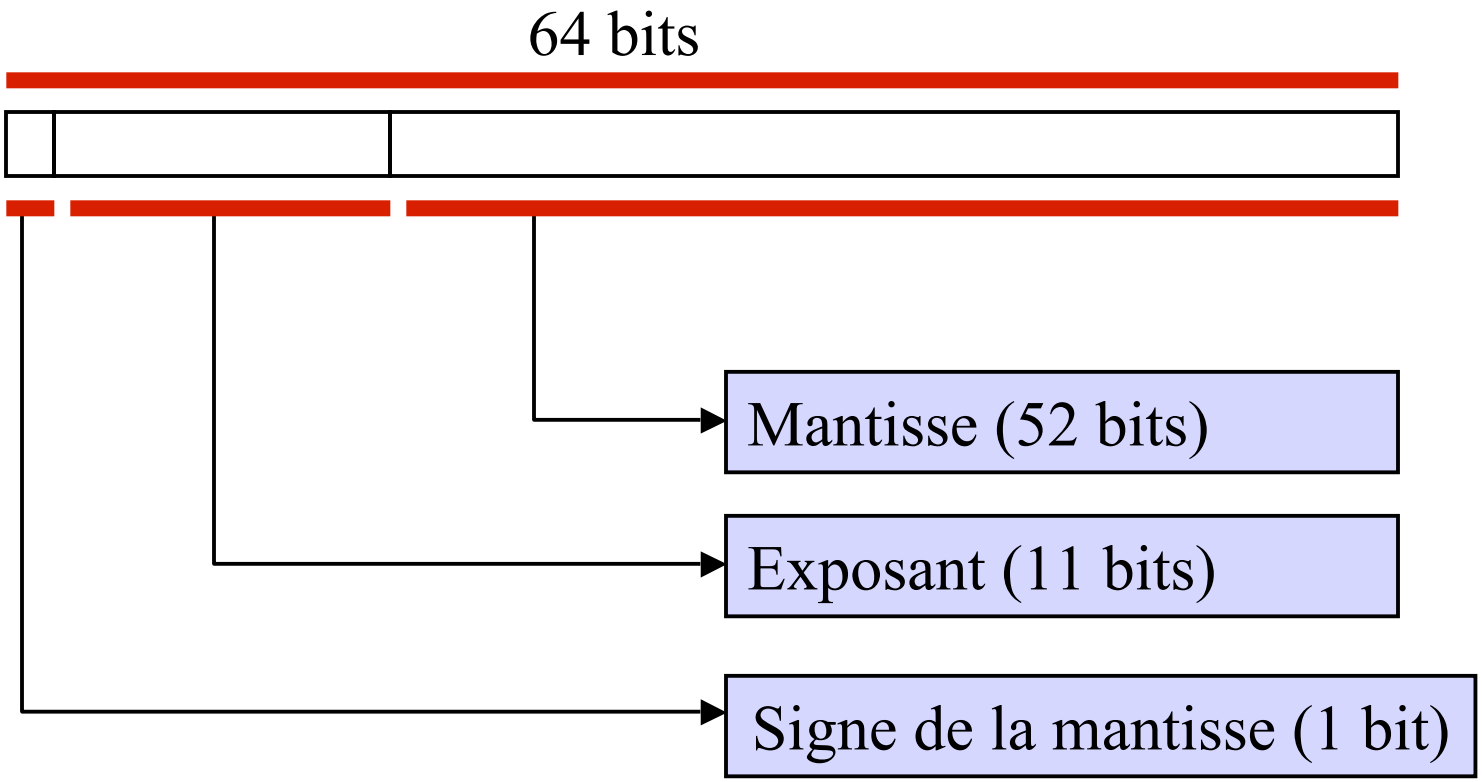
La norme IEEE 754

- Un format standardisé
- **Format simple précision**: 32 bits
  - Bit du signe (1 bit)
  - Exposant (8 bits)
  - Mantisse (23 bits)
- **Format double précision**: 64 bits
  - Bit du signe (1 bit)
  - Exposant (11 bits)
  - Mantisse (52 bits)

Format simple précision



Format Double Précision



Normalisation dans le format IEEE 754

- La mantisse est ***normalisé*** sous la forme
  - $\pm 1,M*2^{\pm c}$
  - Pseudo mantisse
  - Le 1 précédant la virgule n'est pas codé en machine et est appelé **bit caché**
- Exemple:
  - Mantisse:
  - Représentation: 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

$1.101_2 = 1.625_{10}$

IEEE 754, Représentation de l'exposent

- Constante k d'excentrement appliquée à l'exposant
  - Simple précision: $+127_{10}$
  - Double précision: $+1023_{10}$
- L'exposant c codé en interne
  - $\pm c + 127_{10}$
  - $\pm c + 1023_{10}$
- Ex., $-k = 127_{10}$,
  - Exposant: 10000111_2
  - Représentation: $135_{10} - 127_{10} = 8_{10}$ (valeur)

Représentation de l'exposant et de son signe

- Exemple -

Représentez l'exposant 14_{10} avec un excentrement 127:

$\mathbf{127_{10}}$	=	+ $\mathbf{01111111_2}$
$\mathbf{14_{10}}$	=	+ $\mathbf{\underline{00001110}_2}$
Représentation	=	$\mathbf{10001101_2}$

Représentation de l'exposant et de son signe

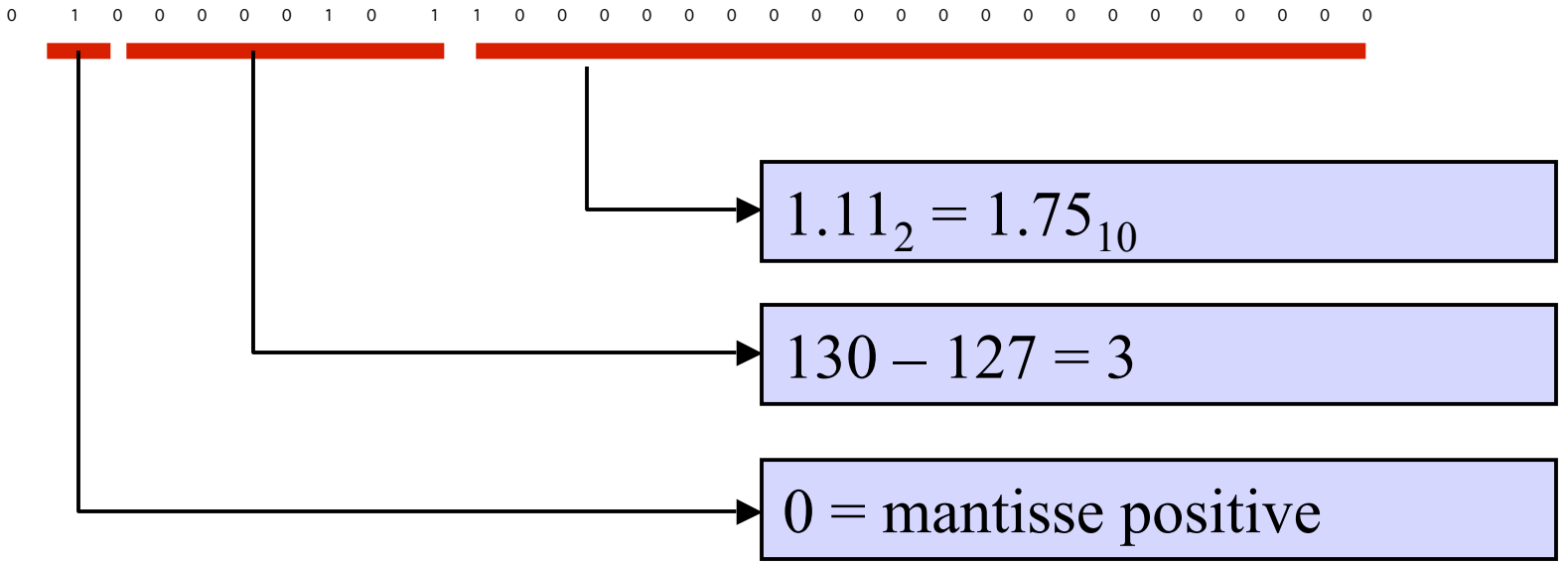
- Exemple -

Représentez l'exposant -8_{10} avec un excentrement 127:

$\mathbf{127_{10}}$	=	+ $\mathbf{01111111_2}$
$\mathbf{-\ 8_{10}}$	=	- $\mathbf{\underline{00001000}_2}$
Représentation	=	$\mathbf{01110111_2}$

Exemple

- Simple précision



$+1.75 \times 2^3 = 14.0$

Exercice – Conversion en virgule flottante IEEE 754

- Quelle est la valeur décimale des représentations internes suivantes?

1 1 0 0 0 0 0 1 0

1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

- Réponse: -

Exercice – Conversion en virgule flottante IEEE 754

Réponse

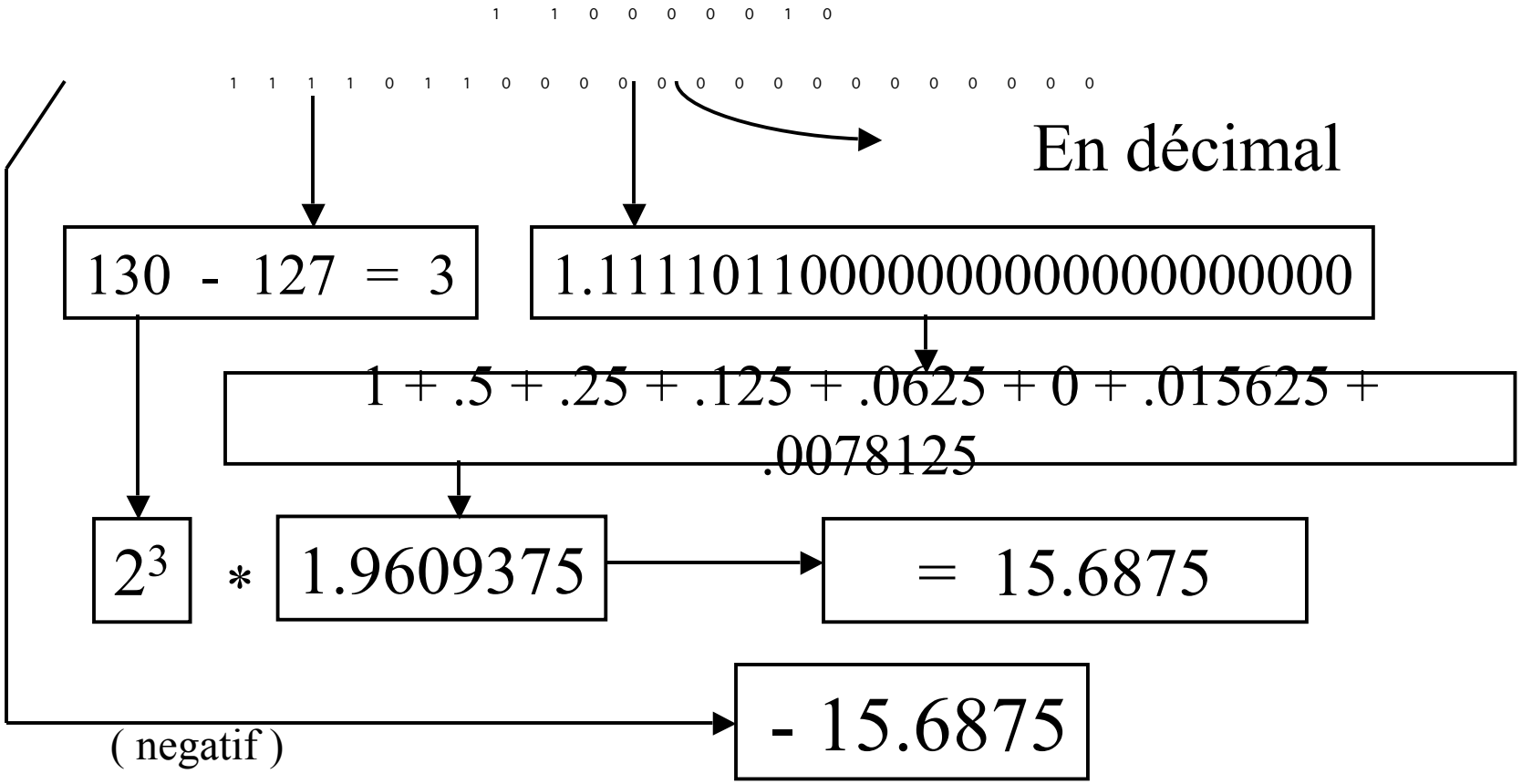
- Quelle est la valeur décimale des représentations internes suivantes?

1 1 0 0 0 0 0 1 0

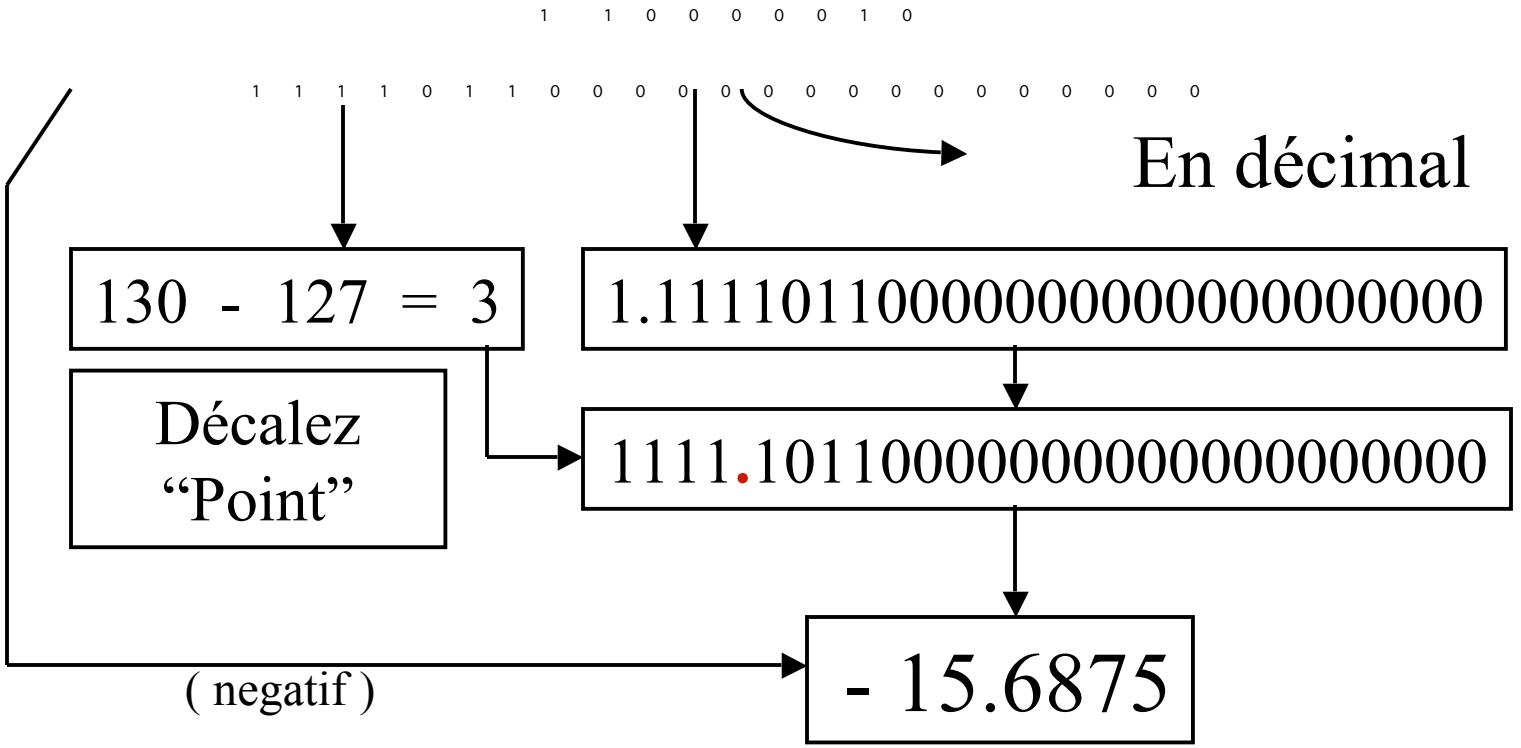
1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

- Réponse: -15.6875

Solution


1 1 0 0 0 0 0 1 0
1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
En décimal
130 - 127 = 3
1.11110110000000000000000000
1 + .5 + .25 + .125 + .0625 + 0 + .015625 + .0078125
2^3 * 1.9609375
= 15.6875
- 15.6875
(negatif)

Solution : Méthode Alternative



Exercice – Conversion en virgule flottante IEEE 754

- Quelle est la représentation interne du nombre 3.14_{10}?
- Remarque: utiliser seulement les 10 chiffres significatifs pour la mantisse

- Réponse: _

Exercice – Conversion en virgule flottante IEEE 754

Réponse

- Quelle est la représentation interne du nombre 3.14_{10}?
- Remarque: utiliser seulement les 10 chiffres significatifs pour la mantisse

- Réponse: 0 1 0 0 0 0 0 0 0 0

1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0

Solution : 3.14 en IEEE Simple Précision

3.14 En Binaire (approx):

11.001000111101

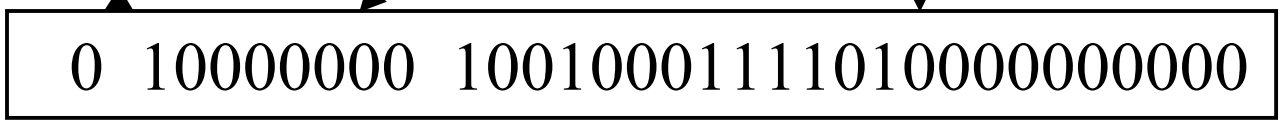
•Normalisez (2^1)
•Enlevez le bit caché

1001000111101

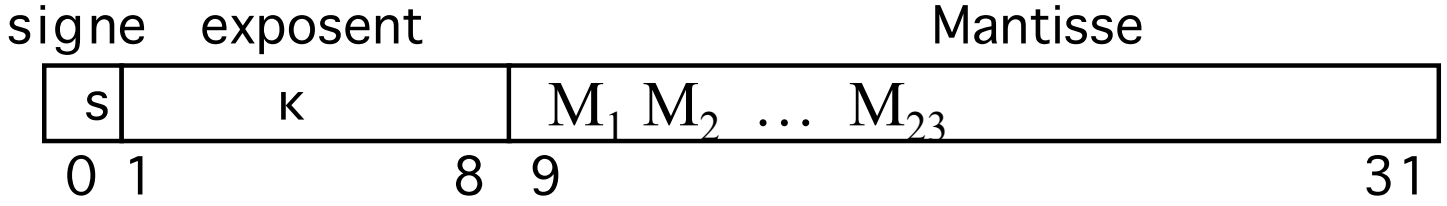
Exposant = 127 + 1

10000000

Valeur est positive: Bit de signe = 0



IEEE 754 Simple Précision Format (Résumé)



- Signe – 1 bit (0 – “+”; 1 – “-”)
- Exposant – 8 bits (excentrement-127)
- Mantisse – 23 bits
- Format binaire
- Normalisation : 1.MMMM…
- Bit caché

Représentation du zéro, des infinis, représentations dénormalisées

- Le norme IEEE admet des codages spéciaux pour la représentation
  - 0
  - $+\infty$
  - $-\infty$
  - Représentations dénormalisées

Représentation du zéro, des infinis, représentations dénormalisées

Exposant	Mantisse	Valeur
0	±0	0
0	Non 0	$\pm 2^{-126} * 0.M$
-126 - +127	Tout	$\pm 2^{E+127} * 1.M$
±128	±0	±∞
±128	Non 0	Conditions spéciales

Addition et soustraction de deux nombres décimales en virgule flottante

Opérandes	Alignement	Normaliser et arrondir
6.144×10^{2}	0.06144×10^{4}	1.003644×10^{5}
$+9.975 \times 10^{4}$	$+9.975 \times 10^{4}$	$+ .0005 \times 10^{5}$
	10.03644×10^{4}	1.004×10^{5}

Opérandes	Alignement	Normaliser et arrondir
1.076×10^{-7}	1.076×10^{-7}	7.7300×10^{-9}
-9.987×10^{-8}	-0.9987×10^{-7}	$+ .0005 \times 10^{-9}$
	0.0773×10^{-7}	7.730×10^{-9}

Calcul en virgule flottante: Addition

- Nombres doivent être alignés : avoir les mêmes exposants (le plus élevé pour protéger la précision)
- Additionner mantisses. Si overflow, ajuster l'exposant
- Ex. 0 51 99718 (e = 1) et 0 49 67000 (e = -1)
  - Aligner les nombres: 0 51 99718
0 51 00670
  - Additionner: 99718
+ 00670
~~1 00388~~ ← Overflow
  - Arrondir le nombre et ajuster l'exposant: 0 52 10039

Calcul en virgule flottante: Multiplication

- $(a * 10^e) * (b * 10^f) = a * b * 10^{e+f}$
- Règle: multiplier les mantisses; additionner les exposants

  But: Codage en excédent, $(n + e) + (n + f) = 2 * n + e + f$

  ➔ Besoin soustraire constante d'excentrement n a partir du résultat
- Ex. 0 51 99718 (e = 1) and 0 49 67000 (e = -1)

  Mantisses: .99718 * .67000 = 0.6681106

  Exposants: 51 + 49 = 100 and 100 – 50 = 50

  Normaliser: .6681106 → .66811

  Résultat: .66811 * 10^0 (50 signifie e = 0)