Representation des nombres: Norme IEEE 754

1 byte = 1 octet = 8 bits

entier represente par 4 byte =32 bits

$$-2^{31} \leq N \leq 2^{31}$$
$$-10^{9} \leq N \leq 10^{9}$$

1 Representation des decimaux

Un decimal s'ecrit sous la forme $m10^e$ m mantisse , e exposant

exemple Nb Avogadro = $6.0221407610^{23} mol^{-1}$

En machine le decimal s'ecrit $M2^E$

2 decimaux representes sur 4 byte

31	30 ... 23	22 ... 0
signe	exposant biaise	mantisse
1 bit	8 bits	23 bits

Table 1: Representation d'un flottant sur 4 octets (IEEE 754)

$$(-1) * *\text{signe} * (1 + \text{mantisse}) * 2 * *(\text{exposantbiaise} - 127)$$

maximum = $2^{2^7} = 2^{128} \approx 2^{3^{40}} \approx 10^{40}$
minimum = $2^{-2^7} \approx 10^{-40}$

3 Zero machine et precision machine

Zero machine

Definition 3.1 *Le plus grand x tel que x soit represente par 0 en machine.*

depend de l'exposant

```
-Inf |--------------|  zero machine        |------------------|Inf

   -10^{40}      -10^{-40}               10^{-40}           10^{40}
```

Precision machine

Definition 3.2 *Le plus grand x tel que $1+x$ soit represente par 1 en machine.*

depend de la mantisse
$2^{23} \approx 10^7$ 7 chiffres significatifs
precision machine 10^{-7}

exemples
1. + 0.000 000 1 = 1.000 000 1
1. + 0.000 000 01 = 1.000 000 0 = 1

equation du second degre

$$x^2 - x + 10^{-20} = 0$$

$$x = \frac{1 \pm \sqrt{1 - 4 \cdot 10^{-20}}}{2}$$

Une racine pose probleme, laquelle?

4 decimaux representes sur 8 byte

63	62 ... 52	51 ... 0
signe	exposant biaise	mantisse
1 bit	11 bits	52 bits

Table 2: Representation d'un flottant sur 8 octets =64 bits (IEEE 754)

$$(-1) * *\text{signe} * (1 + \text{mantisse}) * 2 * *(\text{exposantbiaise} - 1023)$$

zero machine 10^{-308}
precision machine 10^{-14}

5 Complexes

En **Fortran** les complexes sont par defaut representes sur 8 bytes 4 byte partie reelle, 4 byte partie imaginaire

Les complexes sur 16 bytes sont declares
double precision complex

En Matlab et Python les flottants sont par defaut sur 8 octets

6 Arithmetique etendue Norme IEEE 754

Probleme d'Ariane 5 !
Le depassement de capacite doit etre bien gere

La norme introduit
Inf $+\infty$ Overflow
-Inf $-\infty$ Underflow
NaN not a number, Invalid operation, operation illegale (ex $\sqrt{-1}$)

Ces quantites obeissent a des regles d'operation bien precises
Inf $\pm$ x = Inf
Inf $\times$ x = Inf
Inf - Inf = NaN
Inf * 0 = NaN

N'importe quelle operation sur un NaN donne un NaN

```
      program tst4
      parameter(n=39)
      real x , y , z

      y=0.
      x = 1. / y
      write(*,*)'1/0  =', x

      x = 10.**n
      write(*,*)'10**n =', x

      y = -10.**n
      write(*,*)'-10**n =', y

      z= x + y
      write(*,*)'10**n-10**n =',z

      x = 10.**(-2*n)
      write(*,*)'10**(-2*n) =', x

      stop
      end
---------------execution------------------------
tst4
 1/0  =  INF
 10**n =  INF
 -10**n = -INF
 10**n-10**n =  NAN
 10**(-2*n) =  0.
```