	Signed bit	Exponent	Fraction	Decimal
i)	0	0100	1.0100	
ii)	1	1100	1101	
iii)	0	1111	1100	448
iv)	0	1110	0001	
v)	1	1011	1100	
vi)	0	0011	1111	

Table 7: Floating point representation

$M = 1,25$

$E = 4 - (8-1) = -3$

$1,25 \cdot 2^{-3}$

$M = 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^4} = \frac{8+8+16}{16}$

$E = 8+4-7 = 5 \qquad = \frac{29}{16}$

$- \frac{29}{16} \cdot 2^5$

$= -58$

$$M = 1{,}75$$
$$E = 15 - 7 = 8$$

$$1{,}75 \cdot 2^8$$

$$M = \frac{1}{16}$$

$$E = 14 - 7 = 7.$$

$$\frac{17}{16} \cdot 2^7 = 136$$

$$M = \frac{3}{2^2} + \frac{4}{4} = \frac{7}{2^2}$$

$$E = 11 - 7 = 4.$$

$$-\frac{7}{2^2} \cdot 2^4 = -7 \cdot 4 = -28$$

$\Omega =$

$E = 3 - 7 = -4$

$\frac{16}{16} + \frac{8}{16} + \frac{4}{16} + \frac{2}{16} + \frac{1}{16} = \frac{31}{16}$

$\frac{31}{16} \cdot 2^{-4} = \frac{31}{256}$

$n + 8 = 9n$

(b)

(i) $2^n - 1$

(ii) $\frac{1}{2^n}$

(iii) $-2^n + 1$

(iv) $-\frac{1}{2^n}$

(c) 2.5

$0.5 \cdot 2 = 1.0$

$M = (10.1)_2 = (2.5)_{10}$

$= (.01) \cdot 2^1$

$1 = E - B$

$\Rightarrow E = 1 + B = 128$

$x = ($ 0 10000000 010 0000 $)$

(ii) 3.75

$= (11.11)_2$

$(.111)_2 \cdot 2^1$

$E = 128$

(0 10000000 11100000000000...)

(iii) $(100.1)_2$

$= (.001)_2 \cdot 2^2$

$2 = E - B$

$\Rightarrow E = 2 + B$

$= 129$

(1 10000001 00100000...)

(iv) $(101)_2 = (.01) \cdot 2^1$

$\Rightarrow (1\ 10000000\ 010\ldots)$

(v) $6.25 = (110.01)_2$

$= (.1001)_2 \cdot 2^2$

$\Rightarrow (0\ 10000001\ 10010000\ldots)$

(a) $0.75 = (0.11)_2$

$(1.1)_2 \cdot 2^{-1}$

$E = 126$

$X = (1\ 01111110\ 10000\ldots)$

$$-1 = E - B$$

$$= E - 1023$$

$$\Leftrightarrow E = 1022$$

$$X = (0\ 01111111110\ldots100000\ldots$$

Exercice 7

(a) $8 + 2 = 10$

$\frac{3}{4} + \frac{1}{16} = \frac{13}{16}$

(b) $10 + \frac{13}{16} - 128$

$= -117.1875.$

(c) $41.0625 - 128 = -86.375$

(d) $-128 + 4 + 16 = -108$ 3125

$+ \frac{1}{2} + \frac{1}{8} + \frac{1}{16}$

$= -108 + 0,5 + 0,125 + 0,0625$

$= -107.6875$

e) $1 + 4 + 16 + 64$

$1/2 + 1/16$

f) $2 + 8 + 16 + 64$

$1/2 + 1/4 + 1/8 + 1/16$

(b) (i) $2^{n_1-1} - 1 + \left(1 - \frac{1}{2^{n_2}}\right)$

↓ rec

(ii) $\frac{1}{2^{n_2}}$

(iii) -2^{n_1-1}

(iv) $-\frac{1}{2^{n_2}}$

(c) (i) 1.5

0001.1000

(ii) 0101.0100

(iii) −0.125

875

500
250
125

1111.1110

(iv) - 3.0

0011.0000

→ 1101.1111

(v) -6.75

"rajoute 1 à la dernière pos"

0110.1100

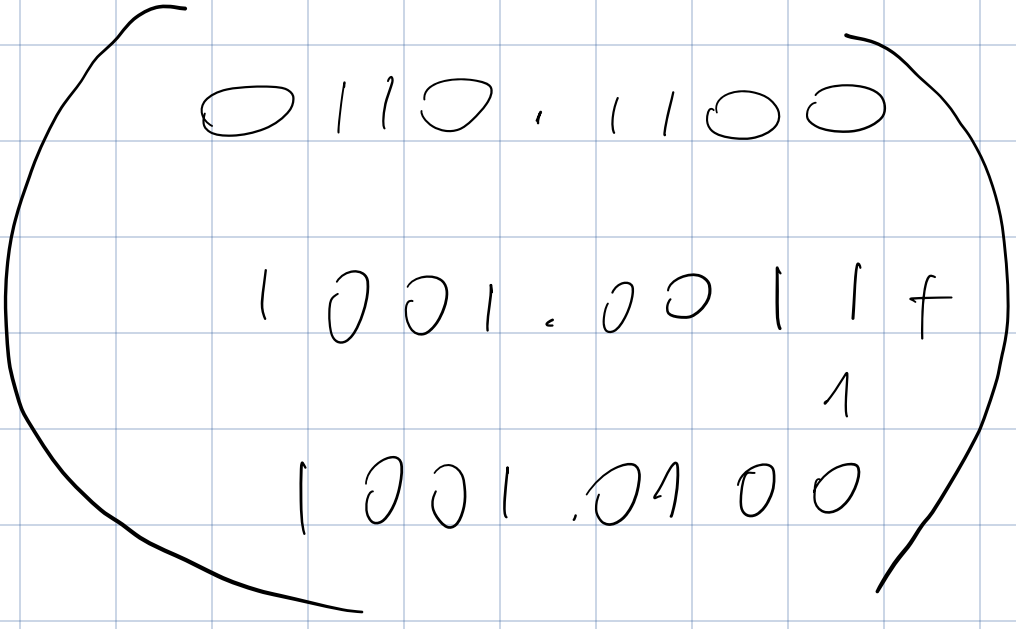
- valeur & &cb...

0011

+1

(1001).(0100)

-7



×

10001 0011

0.0011

1.110

1.111

0 1 2 3

-4 -3 -2 -1

110.000

110000

(d)

(i) $(.1010010)_2 \cdot 2^3$

$3 = E - B$

$= E - 127$

$\Leftrightarrow 130 = E$

(0 10000010 10100100...)

(ii) $(.0101101)_2 \cdot 2^3$

$E = 127 + 4 = 131$

(1 10000010 0101101...)

(iii) $(.1111111)_2 \cdot 2^3$

$E = 130$

$(0\ 10000010\ 11111100\ldots)$

(iv) $(.0011001)_2 \cdot 2^4$

$E = 131$

$(1\ 10000010\ 00110010\ldots)$

Exercise 8

(a) (I) from $(2^{24}-1)\cdot 2^{2^8-1-127}$
to $-(2^{24}-1)\cdot 2^{2^8-1-127}$

range = $2^{24}-1 \quad 2^8-1-127$

$M \cdot 2^{E} \cdot (-1)^{S}$ (E → $2^8-1-127$, S → 1)

$$\frac{1}{2^{24}-1}\cdot 2^{2^8-1-127}$$

Max value for magnitude:

1.111 1111 1111 1111 $\quad a_0\cdot\left(\frac{1-q^{n+1}}{1-q}\right)$

$1+\frac{1}{2}+\frac{1}{4}+\dots$

$$= 1\cdot\left(\frac{1-q^{24}}{1-q}\right)$$

$$= \frac{1 - \frac{1}{2^{24}}}{\frac{1}{2}}$$

$$= 2 - 2^{-23}$$

$$\left(2 - \frac{1}{2^{23}}\right) \cdot 2^{E}$$

$$= \left(2 - \frac{1}{2^{23}}\right) \cdot 2^{2^7}$$

$$= 2^{128} - \frac{2^{128}}{2^{23}}$$

$$= 2^{128} - 2^{105}$$

$$= 2^{105}\left(2^{24} - 1\right)$$

S and M is symmetric

$$\pm\,(0\,123\ldots)$$

(2)

$$\left(1 + \frac{1}{2}\,\frac{1-(-1/2)^{52}}{1-(-1/2)}\right)\cdot 2^{2^{11}-1-1023}$$

$$= \left(1 + 1 - \frac{1}{2^{52}}\right)\cdot 2^{2^{10}}$$

$$= \left(2 - \frac{1}{2^{52}}\right)\cdot 2^{1024}$$

$$= 2^{1024}\left(2 - \frac{1}{}\right.$$

$$= 2^{1025} - 2^{972}$$

$$= 2^{972}\left(2^{52} - 1\right)$$

(ii) **Resolution** : smallest diff.

(biased) min value

0 00000000 1.00000000000000000000000

$127 = 2^7 - 1$

$E = +1 - B$

↑ doc min -127 ; ↑ min is 0

doc $E_{min} = -2^7 + 1 = -2^7 + 1$

$1 \cdot 2^{(1-2^7)} = 2^{-127}$

π E

500
+250

500
+125
+0625

b) i) 2.70

excès

$\Rightarrow$ (10.1011 .0XXX

$= (.0110) \cdot 2^1$

$1 = E - 1$

$\Rightarrow E = 2$

(0 10 0110) = x_1

ii) 3.9

500
+250 } 875
+125
+0625 } 935...

$\Rightarrow (011.111)$

$1 = E - 1$

$\Rightarrow E = 2$

(0 10 111)

(iii) -0.2

on essaye 0.1875

0.0011 $\quad (.1)\cdot 2^{-3}$ $\quad -3 = E-1 \Rightarrow E = -2$ → précision trop petite

0.01 $\quad (.0)\cdot 2^{-2}$ $\quad -2 = E-1$ → pas possible

On essaye 0.25

$(0.1)_2 \quad .0 \cdot 2^{-1} \quad -1 = E-1 \Rightarrow E = 0$

$(1\ 00\ 0000)$

$0.125 + 0.0625 = 0.1875 \pm 0.1$

0.250 ± 0.50

(iv) 1.33

$250 \mid 375$

$+125$

$(1.0110)_2$

$(.0110) \cdot 2^0 \quad 0 = E-1 \Rightarrow E = 1$

$(0\ 01\ 0110)$

(v) -0.67

(0.1011)

0.500
$+ 0.125$
$+ 0.0625$

$(.011) \cdot 2^{-1}$ $\quad -1 = E - 1$

$\Rightarrow E = 0$

$(1 \;\; 00 \;\; 0110)$

(c) (i) $(7.2)_{10}$

0.125
0.0625
] 1875

$(0111.0011)_2$

(ii) 8 - 6,42

$-6.42 = -8 + 0.08 + 0.5 + 1$

(1001.1001) 3125 [+ 250 + 0625

375 [+ 250 + 125

(iii) 8 - 3.67

$-3.67 = -8 + 4 + 0.33$

$(1100.0101)_2$

(iv) 5.33

$(0101.0101)_2$

Ex 10

A	B	SP(A)	SP(B)	A + B	SP(A) + SP(B)
0.125	0.25				
-0.375	0.5				
0.625	0.75				
-0.875	-1.0				
1.125	-1.25				

Table 12: Addition in floating point representation

0.125

(0.001)

$(.01) \cdot 2^{-3}$

$-3 = E - B$

$\Leftrightarrow E = -3 + 127$

$= 124$

$\rightarrow$

$X = (0 \quad 01111100 \quad 0000\ldots)$

0.25

$(0.01)_2$

$(.0) \cdot 2^{-2}$

$-2 = E - B$

$\Leftrightarrow E = 2 + 127$

$= 125$

$Y = (0\ 01111101\ 000000\ldots)$

0.375

$0\ 01111101\ 00000000$

$+\ 0\ 01111101\ 00000000$

$= 0\ 01111101\ 10000000\ldots$

-0.375

250
+125

$(0.011)_2$
$= (-1) \cdot 2^{-2}$

$-2 = E - B$
$\Leftrightarrow E = -2 + 127$
$= 125$

$Z = (1\ 01111101\ 10000000\ldots)$ ✓

0.5

$(0.1)_2$
$(-0)_2 \cdot 2^{-1}$

$E = 126$

$A = (0\ 01111110\ 00000\ldots)$ ✗

↑ ⊕ grand

hidden bit

shift

$(0\ 01111110\ 1.0000000000\ldots$

$-(1\ 01111101\ 0.1000000000\ldots$

$0.0100000\ldots$ *

⚠ on shift de 2, l'exposant passe de -1 à -3

donc on passe de 126 à 124

$\Rightarrow 1\ 01111100\ 0000000\ldots$

* $\begin{array}{r} 1\ \overset{1}{1}00 \\ -011 \\ \hline 001 \end{array}$ ← on emprunte 1 $\Rightarrow 10 - 01 = 01$

0.625

$(0.101)_2$

$(.01)_2 \cdot 2^{-1}$

$-1 = E - B$

$\Leftrightarrow E = 126$

B = (0 01111110 01000000...)

0.75

$(0.11)_2$

$(.1)_2 \cdot 2^{-1}$

C = (0 01111110 10000000...)

(0 01111110 1.01000000000

0 01111110 1.10000000000

10.11

⇒ (0 01111111 01100000000)

on passe de 126 à 127

500
+ 250
+ 125

-0.875

$(0.111)_2$

$(.111)_2 \cdot 2^{-1}$

D = (1 01111110 11100000...)

-1.0

$(1.0)_2$

$(.0)_2 \cdot 2^0$

$0 = E - 127$

$\Rightarrow E = 127$

E = (1 01111111 00000000...)

(1 01111111 0.11100000...)
+ (1 01111111 1.00000000...)

$$= (1\ 01111111 \quad 1.1110000\ldots)$$

$$(1\ 01111111 \quad 1110000\ldots)$$

1.125

$(1.001)_2$

$(.001) \cdot 2^0$

$E = 127$

$F = (0\ 01111111\ 00100000\ldots)$

1

0

-1

-1.25

$(1.01)_2$

$(.01) \cdot 2^0$

$G = (1\ 01111111\ 01000000\ldots)$

$$(1\ 01111111\ \ 1,01000000\ldots) - (0\ 01111111\ \ 1,00010000\ldots)$$

$$(1\ 01111100\ \ 001.0000\ldots)$$

3 shift

note, si on avait fait l'inverse, on aurait eu une retenue infinie et on aurait dû shifter vers la gauche jusqu'à avoir le premier 1 devant

(b)

625

0
−0 1
1

−8+6

−8+2+1

A	B	FiP(A)	FiP(B)	A + B	FiP(A) + FiP(B)
1.125	3.25	0001.0010	0011.0100	4.375	
-4.375	6.5	1011.1010	0110.1000	2.125	
2.625	7.75	0010.1010	0111.1100	10.375	
-0.875	-1.125	1111.0010	1110.1110		
6.25	-3.875	0110.0100	1100.0010	2.375	

Table 13: Addition in fixed point representation

1 1
0001.0010
+ 0011.0100
0100.0110

2.5 − 0.375
− 2.125

1 1 1 1
1011.1010
+ 0110.1000
0010.0010

1 1 1 1 1 1
1111.0010
+ 1110.1110
1110,0000

1 1 1 1
0010.1010
+ 0111.1100
1010.0110

1
0110.0100
+ 1100.0010
0010.0110

Exercice 9

16 64
32

$$1110.10000001$$

$$(.1101000000 1) \cdot 2^3$$

$$B = 15$$

$$3 = E - 15$$

$$\Leftrightarrow E = 18$$

$$10010$$

$$X = (0\ 10010\ 1101000000 1)$$

$$-21.75 = \overset{-24}{-32} + \overset{-22}{8} + 2 + 0.25$$

$$(10101.1100)_{16}$$

$$(.010111) \cdot 2^{750}$$

$$(1 \quad 10011 \quad 0101\,1100\ldots)$$

$$4 = E - 15$$

7.45

$$2 + 1 + 0.5 + 0.25 + 0.0625 = 3.75$$

$$0.5 + 0.25 + 0.125 + 0.0625 = 0.96875$$

$$(111.01111)$$

$$=$$

$$(1.1101111001110) \cdot 2^{2}$$

$$2 = E - B$$

$$\Rightarrow 17 = E$$

$$(0\ \ 10001\ \ 11011110011)$$

$1 = E - 15$

-2.725

$10.10111) \cdot 2^1$

A.(1 10000 01011100011)

Max: 2^4 (19)

$2^4 \rightarrow 2^3$
on shift right de 3

14.51

(0 10011 1101000001
01110 10000)

$2^4 \rightarrow 2^2$
on shift right de
19 - 17 = 2

7.45

(0 10011 00111011110)

-2.725

(1 10011 00010101111)