Floating point getallen

Voorstelling van positieve en negatieve kommagetallen

Vraag 1. Fictieve float

Bepaald de binaire waarde van volgende getallen in de fictieve 8-bits floating point notatie:

a) -1,375

$$\begin{array}{r} 0{,}375 \\ *16 \\ \hline 6{,}0 \end{array}$$

$$-1{,}375 \xrightarrow{\text{hexadecimaal}} -1{,}6H \xrightarrow{\text{binair}} -1{,}0110$$

normaliseren naar 1,.....

-1,0110 = -1,01100 * 2^0

Tekenbit = 1

Exponent = +0 +2 = 2 → 10

Mantisse = 01100

7	6	5	4	3	2	1	0	bitpositie
1	1	0	0	1	1	0	0	bitwaarde

b) -0,875

$$\begin{array}{r} 0{,}875 \\ *16 \\ \hline 14{,}0 \end{array}$$

$$-0{,}875 \xrightarrow{\text{hexadecimaal}} -0{,}EH \xrightarrow{\text{binair}} -0{,}1110$$

normaliseren naar 1,.....

-0,111 = -1,11000 * 2^{-1}

Tekenbit = 1

Exponent = -1 +2 = 1 → 01

Mantisse = 11000

7	6	5	4	3	2	1	0	bitpositie
1	0	1	1	1	0	0	0	bitwaarde

c) -0,5

$$\begin{array}{r} 0{,}5 \\ *16 \\ \hline 8{,}0 \end{array}$$

$$-0{,}5 \xrightarrow{\text{hexadecimaal}} -0{,}8H \xrightarrow{\text{binair}} -0{,}1000$$

normaliseren naar 1,.....

-0,1 = -1,0 * 2^{-1}

Tekenbit = 1

Exponent = -1 +2 = 1 → 01

Mantisse = 00000

7	6	5	4	3	2	1	0	bitpositie
1	0	1	0	0	0	0	0	bitwaarde

Vraag 2. Fictieve float

Wat is het grootste positieve getal dat met deze fictieve 8-bits floating point notatie en met normalisatie naar 1,... kan worden gemaakt?

7	6	5	4	3	2	1	0	bitpositie
0	1	1	1	1	1	1	1	bitwaarde

Tekenbit = 0
Exponent = 3 - 2 = +1
Mantisse = 11111

= + 1,11111 * 2^{+1}
= + 11,1111
= + 3 + 15/16
= + 3,9375

Hier is geen rekening gehouden met een eventuele oneindignotatie of een nulnotatie.

Vraag 3. Fictieve float

Geef de decimale waarde van de volgende fictieve 8-bit floating points.

a) 11010001

7	6	5	4	3	2	1	0	bitpositie
1	1	0	1	0	0	0	1	bitwaarde

Tekenbit = 1
Exponent = 2 - 2 = 0
Mantisse = 10001

= - 1,10001 * 2^{+0}
= - 1 + (½ + 1/32) = - 1 + (16/32 + 1/32) = -1 + 17/32
= - 1,53125

b) 00011010

7	6	5	4	3	2	1	0	bitpositie
0	0	0	1	1	0	1	0	bitwaarde

Tekenbit = 0
Exponent = 0 - 2 = -2
Mantisse = 11010

= + 1,11010 * 2^{-2}
= + 0,011101
= + 1/4 + 1/8 + 1/16 + 1/64
= 16/64 + 8/64 + 4/64 + 1/64 = 29/64
= + 0,453125

Vraag 4. Float (32bit)

Geef de float notatie van de volgende decimale getallen.

a) 256,28125

Decimale waarde	⇒ 256 + 9/32 = 256 + 0,28125
De binaire waarde met komma-notatie	⇒ 100000000,01001
Normalisatie	⇒ $1{,}0000000001001 \times 2^{8}$
Bit 31: het is een positief getal	⇒ 0
Exponent in plus 127-notatie = 127 + 8 = 135	⇒ 10000111
De mantisse	⇒ 0000000001001

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0

b) +56,428

Decimale waarde	⇒ +56 + 0,428
De binaire waarde	⇒ 111000,01101101100100010110 …
Normalisatie	⇒ $1{,}11000011011011001000101\ldots \times 2^{5}$
Bit 31: het is een positief getal	⇒ 0
Exponent in plus 127-notatie = 127 + 5 = 132	⇒ 10000100
De mantisse	⇒ 11000011011011001000101

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	0	1	1	0	1	1	0	0	1	0	0	0	1	0	1

Vraag 5. Float (32bit)

Wat is het grootste positieve getal in float notatie?

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

De tekenbit is 0, de exponent is maximaal en de mantisse is gelijk aan nul,
DUS $+\infty$

Op ∞ na is het grootste werkelijke getal dus:

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

De tekenbit is 0 = positief
De exponent is op 1 na maximaal = 254 –127 = 127
De mantisse is maximaal

$\Rightarrow 1{,}11111111111111111111111000\ldots \times 2^{254-127}$

$= 1{,}11111111\ldots\ldots \times 2^{127} = \pm 1 \times 2^{128} - 1 \times 2^{104} - \ldots.$

$= \pm 3.4 \cdot 10^{38}$

$= \pm 3.4\ E^{38}$

Vraag 6. Float (32bit)

Bereken de decimale waarde van volgende float getallen:

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0

Teken = positief

Exponent = 1000 0100 = 132, maar in plus-127, dus 132-127=+5

$\Rightarrow$ +1,10000011010001 x $2^{132-127}$

= 1,10000011010001 x 2^{5}

= 110000,011010001

= 48 + 209/512 = + 48,408203125

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	1	0	0	0	0	1	0	1	1	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	0	1	1	1	1	0	1

Teken = negatief

Exponent = 1000 0101 = 13

-1,11101110000101000111101 x $2^{133-127}$

= -1,11101110000101000111101 x 2^{6}

= -1111011,10000101000111101

= -123 + 68157/131072

= -123,5199966

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Teken = negatief

Exponent = 0000 0000 = 0, dus dit is een **gedenormaliseerd getal**

-**0,**11101 x 2^{0-127}

= -0,11101

= -29/32 x 2^{-127}

= -5,32645877716349027048333743428 83e^{-39}

Let er dus op om bij de minimale exponent 0 de impliciete 0 te plaatsen voor de mantisse!!!

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0

Dit is een NaN.