Reserved Exponent Values

Exponent Value	Mantissa	Represents
11111111	All zeros	Infinity (∞)
11111111	Not all zeros	Not a number (NAN)
00000000	All zeros	Zero
00000000	Not all zeros	Subnormal (very small)

Practice problems

Convert the following decimal numbers to 32-bit IEEE 754 format by hand:

a. 1.0
b. −0.1
c. 2016.0
d. 0.00390625
e. −3125.3125
f. 0.33
g. −0.67
h. 3.14

▼ Answer

a. `3f800000`
b. `bdcccccd`
c. `44fc0000`
d. `3b800000`
e. `c5435500`
f. `3ea8f5c3`
g. `bf2b851f`
h. `4048f5c3`

Convert the following hexadecimal numbers to decimal by hand using the 32-bit IEEE 754 format:

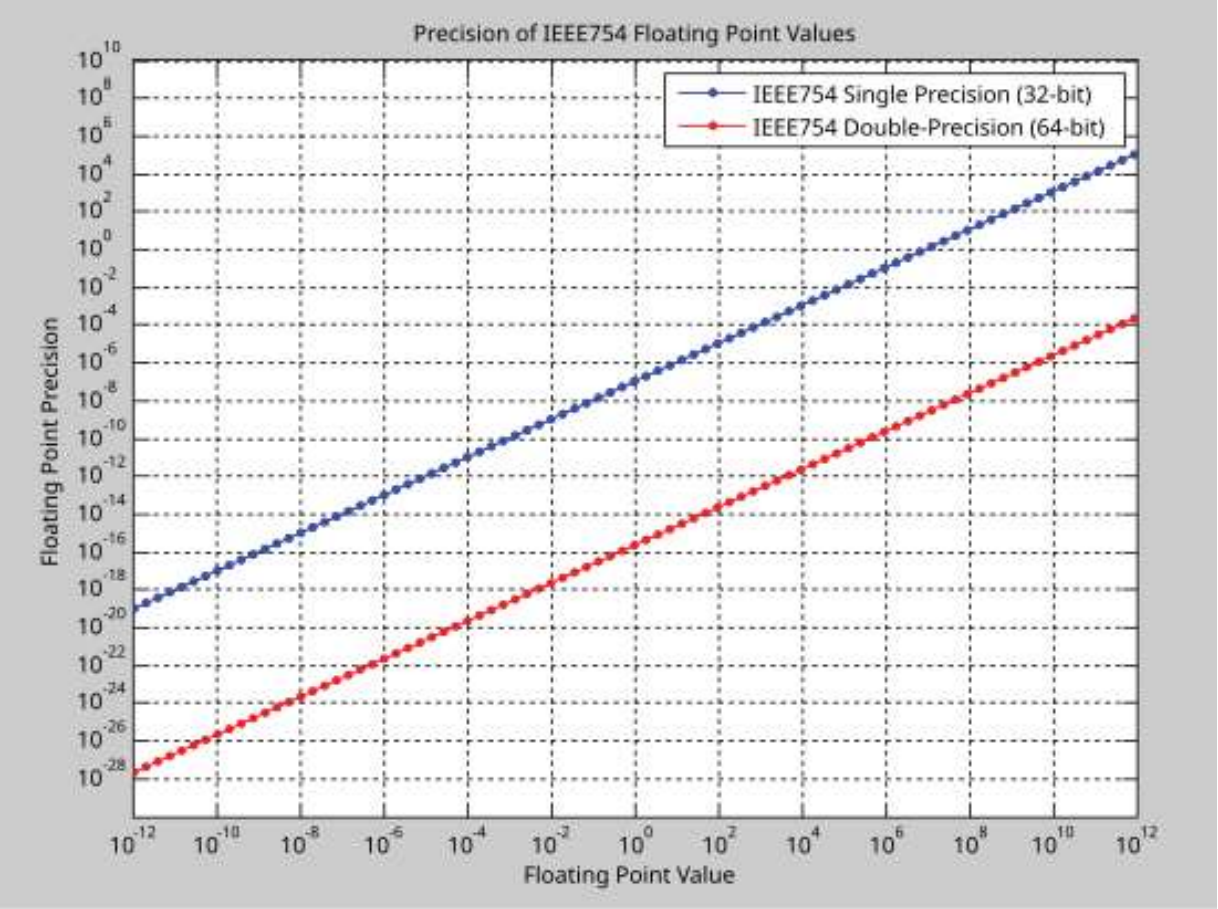
a. `40000000`
b. `bf800000`
c. `3d800000`
d. `c1804000`
e. `42c81000`
f. `3f99999a`
g. `42f6e666`
h. `c25948b4`

▼ Answer

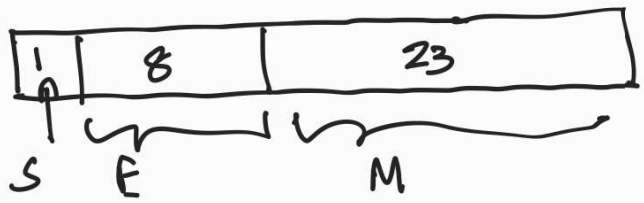
a. +2.0
b. −1.0
c. +0.0625
d. −16.03125
e. 100.03125
f. 1.2
g. 123.449997
h. −54.320999

Ack: https://bob.cs.sonoma.edu/IntroCompOrg-RPi/exercises-32.html

Precision and range in IEEE 754



https://en.wikipedia.org/wiki/IEEE_754



IEEE 754

Range: More is better
Resolution: Less is better.

Biased Exponent : 1 to 254
True exponent : -126 to 127

Smallest value : $1.0000\ldots0 \times 2^{-126} = 2^{-126}$
Largest value : $1.1111\ldots1 \times 2^{127} = (2-2^{-23}) \times 2^{127}$

$10.0000\ldots0 \rightarrow 2$
$-\ 0.000\ldots01 \rightarrow 2^{-23}$

2

Subnormal numbers

-2^{-126} 0 2^{-126}

-10^{+38}

$(2-2^{23}) \times 2^{127} \approx 10^{38}$

Range of signed numbers.

-2^{-31} to $2^{31}-1$
$\approx 10^{10}$

Signed no.

Resolution $= 1$ ✓

Resolution -

smallest $\begin{cases} -1.0000\ldots0 \times 2^{-126} \\ -1.0000\ldots1 \times 2^{-126} \end{cases}$

$.0000\ldots1 \times 2^{-126}$

$2^{-23} \times 2^{-126} = 2^{-149} \approx 10^{-45}$

$1.000\ldots0 \times 2^{0}$ ✓
$1.000\ldots1 \times 2^{0}$
$.000\ldots1 \times 2^{0}$ $\rightarrow 2^{-23} \times 2^{0} : 2^{-23}$

$1.0000\ldots0 \times 2^{25}$

$1.0000 \quad\quad 1 \times 2^{25}$

$2^{-23} \times 2^{25} = 2^2 = 4$

Worst case:

$1.000\ldots0 \times 2^{127}$

$1.000\ldots1 \times 2^{127}$

$2^{-23} \times 2^{127} = 2^{105} \approx 10^{31}$ ✓

Floating point addition.

$\rightarrow \quad 9.876 \times 10^4$

$\rightarrow + 5.678 \times 10^3 \rightarrow$

① Exponent Comparison.

$4 > 3$

② Mantissa alignment:

9.876×10^4

0.5678×10^4

③ Perform operation on mantissa.

9.876

0.5678

10.4438

④ Apply the common exponent to mantissa.

10.4438×10^4

⑤ Normalization $\rightarrow$ 1.04438×10^5.

Subtraction.

9.876×10^4

$- 5.678 \times 10^3$

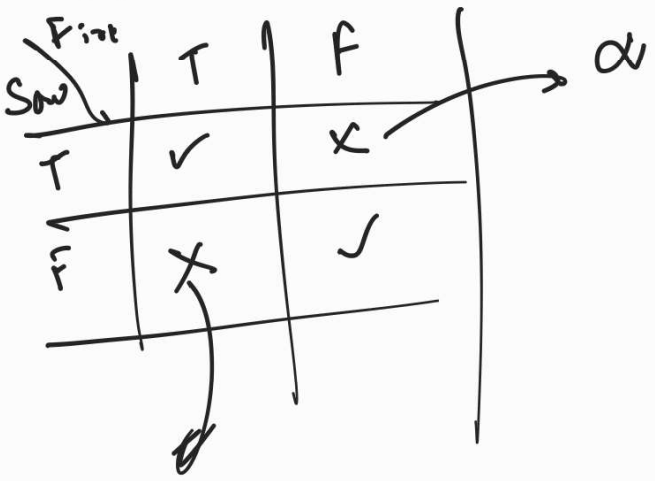
9.876

-0.5678

Conversion:
the large exponent value is chosen for alignment.

9.3082×10^4

Binary Addition



(more serious, fire but alarm does not sound).

0·8

·11001100.

100000

·11001 ~~100~~

+1

11010