We will start with a definition of floating point and double using IEEE format definitions. They are

- 32-bit: 1 sign bit, 8 exp bits, 23 frac bits
- 64-bit: 1 sign bit, 11 exp bits, 52 frac bits
- 8-bit: 1 sign bit, 4 exp bits, 3 frac bits

1. Provide the representation of the value 3.14159265 in 32-bit IEEE format.
2. Provide the representation of the value 3.14159265 in 64-bit IEEE.
3. Provide the representation of the value 3.14159265 in the 8 bit format

3.14159265 → 11.001001000011111101101010100010001
000110101001111000 01 ← Because Rounding

$(3)_{10} = (11)_2$

32-bit bias: $2^{(8-1)}-1 = 127$

64-bit bias: $2^{(11-1)}-1 = 1023$

8-bit bias: $2^{(4-1)}-1 = 7$

0.14159265

× 2	bit
0.28318530	0
0.56637060	0
1.13274120	1
0.26548240	0
0.53096480	0
1.06192960	1
0.12385920	0
0.24771840	0
0.49543680	0
0.99087360	0
1.98174720	1
1.96349440	1
1.92698880	1
1.85397760	1
1.70795520	1
1.41591040	1
0.83182080	0
1.66364160	1
1.32728320	1
0.65456640	0
1.30913280	1
0.61826560	0
1.23653120	1
0.47306240	0
0.94612480	0
1.89224960	1

× 2	bit
1.89224960	
1.78449920	1
1.56899840	1
1.13799680	1
0.27599360	0
0.55198720	0
1.10397440	1
0.20794880	0
0.41589760	0
0.83179520	0
1.66359040	1
1.32718080	1
0.65436160	0
1.30872320	1
0.61744640	0
1.23489280	1
0.46978560	0
0.93957120	0
1.87914240	1
1.75828480	1
1.51656960	1
1.03313920	1
0.06627840	0
0.13255680	0
0.26511360	0
0.53022720	0
1.06045440	1

$1.100\ldots \times 10^1$

e: 1

64: E = 1 + 1023 = 1024

$(1024)_{10} = (10000000000)_2 = E$

32: E = 1 + 127 = 128

$(128)_{10} = (10000000)_2 = E$

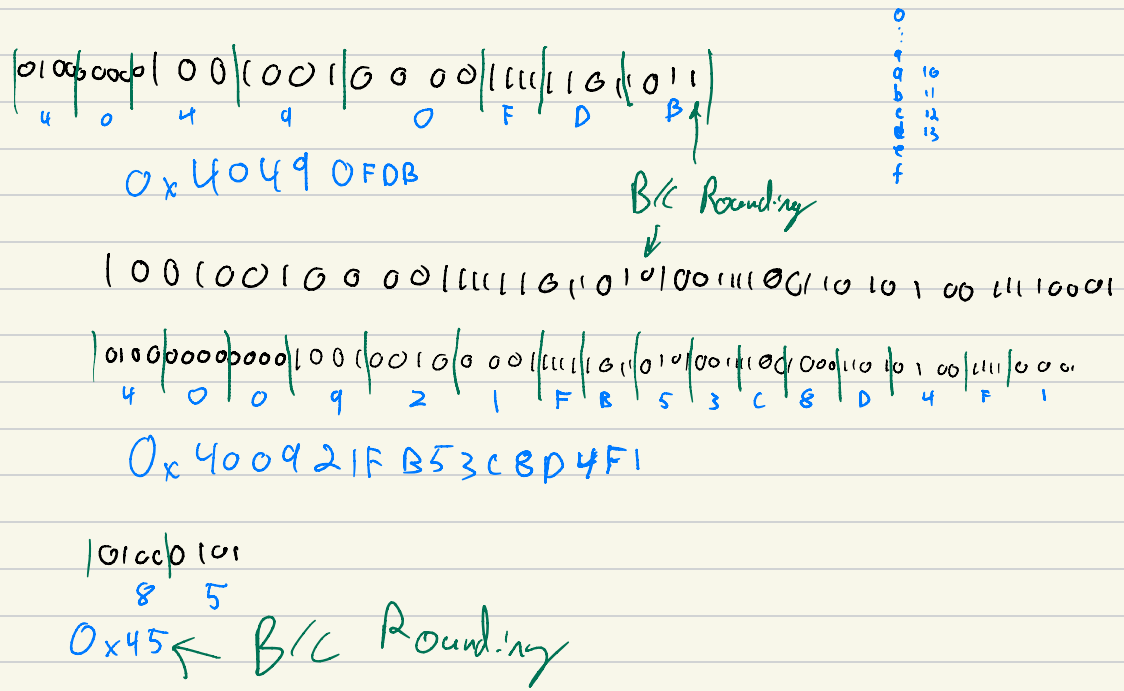
8: E = 1 + 7 = 8

$(8)_{10} = (1000)_2 = E$

We will start with a definition of floating point and double using IEEE format definitions. They
are

- 32-bit: 1 sign bit, 8 exp bits, 23 frac bits
- 64-bit: 1 sign bit, 11 exp bits, 52 frac bits
- 8-bit: 1 sign bit, 4 exp bits, 3 frac bits

1. Provide the representation of the value 3.14159265 in 32-bit IEEE format.
2. Provide the representation of the value 3.14159265 in 64-bit IEEE.
3. Provide the representation of the value 3.14159265 in the 8 bit format



Converting Back:

64 Bit value

S: 0

E: 10000000000 → 1024 ∴ 1024 − 1023 = 1 = Exp

M: 1001001000011111101101010100010001000010110100011000010001101001100010011000110011

$M\_normalized$ = 1.1001001000011111101101010100010001000010110100011000010001101001100010011000110011

$1^S \times M\_normalized \times 2^{Exp}$

1 × 1.1001001000011111101101010100010001000010110100011000010001101001100010011000110011 × 2

2^n 1 0 -1 -2 -3 ...

11.001001000011111101101010100010001000010110100011000010001101001100010011000110011

↓

$2^1 + 2^0 + 2^{-3} + 2^{-6} + 2^{-11} + 2^{-12} + 2^{-13} + 2^{-14} + 2^{-15} + 2^{-16} + 2^{-18} + 2^{-19} + 2^{-21} + 2^{-23} + 2^{-26} + 2^{-27} + 2^{-28} + 2^{-29} + 2^{-32} + 2^{-33} + 2^{-35} + 2^{-37} + 2^{-40} + 2^{-41} + 2^{-42} + 2^{-43} + 2^{-47}$

$= 3.14159265$

32 bit value

S: 0

E: 10000000 → 128 : 128 - 127 = 1 = Exp

M: 10010000111111011011

M_Normalized: 1.10010000111111011011

$1^S \times M\_normalized \times 2^{Exp}$

$1 \times 1.10010000111111011011 \times 2$

2^1 1 0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -10 -11 -12 -13 -14 -15 -16 -17 -18 -19 -20 -21 -22

1 1.10010000111111011011

$2^1 - 2^0 + 2^{-3} + 2^{-6} + 2^{-11} + 2^{-12} + 2^{-13} + 2^{-14} + 2^{-15} + 2^{-16} + 2^{-19} + 2^{-19} + 2^{-21} + 2^{-22}$

$= 3.14159274101$

8 bit value

S: 0

E: 1000 $\Rightarrow 8 \quad \therefore 8-7 \Rightarrow 1 = Exp$

M: 101

M_Normalized: 1.101

$1^S \times M\_normalized \times 2^{Exp}$

$1 \times 1.101 \times 2$

$2^1 \quad 1 \; 0 \; -1 \; -2$

11.01

$2^1 + 2^0 + 2^{-2}$

$= 3.125$

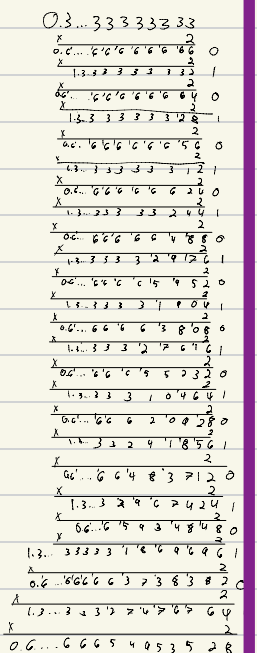
Precision Loss!

64: 3.14159265
32: 3.1415924
16: 3.125

$3.14159265 - 3.1415924 = -9.1 \times 10^{-8}$

$3.14159265 - 3.125 = 0.01659265$

$0.\overline{3}$



32bit

S: 0

F: 2+127 = 125 : 0 1111101

M: 01010101010101010101010

M_norm : 1.01010101010101010101010 $\times 2^{-2}$

assume hidden

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

0.01 01010101010101010101010

$2^{-2}+2^{-4}+2^{-6}+2^{-8}+2^{-10}+2^{-12}+2^{-14}+2^{-16}+2^{-18}+2^{-20}+2^{-22}+2^{-24}$

32bit: 0.33333335135

6 bit

S: 0

E: $-2+7 = (01)_2$

M: 010

M_Norm 1.01×2^{-2}

0.0101

$2^{-2}+2^{-4} = 0.3125$

$0.33333335135 - 0.3125 = 0.02083335135$