

2. Convert the following base 10 numbers to binary and express each as a floating point number $fl(x)$ by using the Rounding to Nearest Rule: (a) 9.5 (b) 9.6 (c) 100.2 (d) 44/7

a. $fl(9.5)$

$$x = (-1)^S \times 2^{e-1023} \times (1.f)_2$$

$$S = 0$$

$$9.5 = 1001_2 + 0.5_{10}$$

$$= 1001.1$$

$$\begin{array}{r|l} 0.5 & \times 2 \\ 1 & \\ \hline 0.1 & \end{array}$$

$$9.5_{10} = 1001.1_2$$

$$= 1.0011 \times 2^3$$

$$X = (-1)^0 \times 2^{1026-1023} \times (1.0011)$$

b. $fl(9.6)$

$$9.6 = 1001.\overline{1001}$$

$$\begin{array}{r|l} 0.6 & \times 2 \\ 1.2 & \\ 0.4 & \\ 0.8 & \\ 1.6 & \\ 1.2 & \\ 0 & \end{array}$$

$$fl(9.6) = (-1)^0 \times 1.0011001 \dots 0011 \times 2^{1026-1023}$$

c. $fl(100.2)$

$$100.2 = 1100100.00110011 \dots$$

$$fl(100.2) = (-1)^0 \times 1.100100 \dots 1101 \times 2^{1029-1023}$$

d $fl(44/7)$

$$6 \frac{2}{7} = 110.\overline{001}$$

$$\begin{array}{r|l} 0 & \frac{2}{7} \end{array} \times 2$$

$$\begin{array}{r|l} 0 & \frac{4}{7} \end{array}$$

$$\begin{array}{r|l} 1 & \frac{1}{7} \end{array}$$

$$\begin{array}{r|l} 0 & \frac{2}{7} \end{array}$$

$$\begin{array}{r|l} 0 & \end{array}$$

$$\begin{array}{r|l} 1 & \end{array}$$

$$fl(\frac{44}{7}) = (-1)^0 \times 1.1001\overline{001} \times 2^{1025-1023}$$

$$8_{10} = 1000_2$$

$$f(8) = (-1)^0 \times 2^{1026-1023} \times (1.00 \dots 0)$$

5 11 52

0	100000000010	$00000 \dots 0$
-----	----------------	-----------------

4
 0
 2

0
 0
 0

4020000000000000


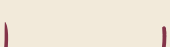
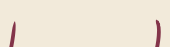
13 to

c. $\frac{1}{8}$

$$\begin{array}{c|c} 0 & \frac{1}{8} \\ 0 & \frac{1}{4} \\ 0 & \frac{1}{2} \\ 1 & \end{array} \quad \begin{array}{c} \text{XL} \\ \\ \\ \end{array}$$

0.001_2

$$f_l\left(\frac{1}{8}\right) = (-1)^0 \times 2^{1020-1023} \times (1.00\dots 0)$$

0	0111111100	000 - - - - 000
<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>3 F C</p> </div> <div style="text-align: center;">  <p>0</p> </div> <div style="text-align: center;">  <p>0 - - - 0</p> </div> </div>		

3FCD0000000000000000

13 ↑ 0

$$e) fl(\frac{2}{3})$$

$$\begin{array}{r|l} 0 & \frac{2}{3} \\ 1 & \frac{1}{3} \\ 0 & \frac{2}{3} \\ 1 & \frac{1}{3} \\ 0 & \frac{2}{3} \end{array} \quad \times 2$$

$$\frac{2}{3}_{10} = 0.\overline{10}_2$$

$$fl(\frac{2}{3}) = (-1)^0 \times 2^{1022-1023} \times (1.01010 \dots 10)$$

0	011111110	010	1010	10101
	3	FE	5	5

3FE 55555555555555

13 ↑ 5

$$g.) fl(-0.1)$$

$$\begin{array}{r|l} 0.1 & \times 2 \\ 0.2 & \\ 0.4 & \\ 0.8 & \\ 1.6 & \\ 1.2 & \\ 0.4 & \\ 0.8 & \end{array}$$

$$0.\overline{00011} \quad (-1)^1 \times 2^{1019-1023} \times (1.10011)$$

1	0111111011	10011001	10011010
	B	FB	

BFB999999999999a

12 ↑ 9 one a

12. Find the IEEE double precision representation $\text{fl}(x)$, and find the exact difference $\text{fl}(x) - x$ for the given real numbers. Check that the relative rounding error is no more than $\epsilon_{\text{mach}}/2$.

(a) $x = 1/3$ (b) $x = 3.3$ (c) $x = 9/7$

a. $x = \frac{1}{3}$

$$\text{fl}\left(\frac{1}{3}\right) = (-1)^0 \times 2^{1021-1023} \times (1.\overline{010101}\dots)$$

$$\begin{array}{r|l} 0 & \frac{1}{3} \\ 0 & \frac{2}{3} \\ 1 & \frac{1}{3} \\ 0 & \frac{2}{3} \end{array} \times 2$$

$$\begin{aligned} \text{fl}\left(\frac{1}{3}\right) - \frac{1}{3} &= -0.\overbrace{00\dots00}^{52 \text{ bit}}\overline{01} \\ &= -0.\overline{01} \times 2^{-54} \end{aligned}$$

$$\frac{|-0.\overline{01} \times 2^{-54}|}{|\frac{1}{3}|} = 2^{-54} < \frac{\epsilon_{\text{mach}}}{2}$$

$$0.\overline{010101}$$

b. $x = 3.3$

$$\text{fl}(3.3) = (-1)^0 \times 2^{1024-1023} \times (1.10\overline{10011001}\dots)$$

$$\begin{aligned} \text{fl}(3.3) - 3.3 &= -0.0\overbrace{\dots\dots\dots}^{52 \text{ bit}}\overline{0110} \\ &= -0.\overline{0110} \times 2^{-51} \end{aligned}$$

c. $x = \frac{9}{7}$

$$\frac{|-0.\overline{0110} \times 2^{-51}|}{|3.3|} < \frac{\epsilon_{\text{mach}}}{2}$$

$$\begin{array}{r|l} 1 & \frac{2}{7} \\ 0 & \frac{4}{7} \\ 1 & \frac{1}{7} \\ 0 & \frac{2}{7} \\ 0 & \frac{4}{7} \\ 1 & \frac{1}{7} \end{array} \times 2$$

$$\text{fl}\left(\frac{9}{7}\right) = (-1)^0 \times 2^{1023-1023} \times (1.0\overline{100100}\dots)$$

$$\text{fl}\left(\frac{9}{7}\right) - \frac{9}{7} = -0.0\overline{100} \times 2^{-52} < \frac{\epsilon_{\text{mach}}}{2}$$

16. Find the IEEE double precision representation $\text{fl}(x)$, and find the exact difference $\text{fl}(x) - x$ for the given real numbers. Check that the relative rounding error is no more than $\epsilon_{\text{mach}}/2$.

(a) $x = 2.75$ (b) $x = 2.7$ (c) $x = 10/3$

a. $x = 2.75$ 1.011×2^{-1}

10.11_2

error = $0 < \frac{\epsilon_{\text{mach}}}{2}$

b. $x = 2.7$

10.10110_2

$\text{fl}(2.7) = (-1)^0 \times 2^{1024-1023} \times (1.0101101001)$

$\text{fl}(2.7) - 2.7 = 0.\overbrace{0 \dots 0}^{52 \text{ bit}} 1001$

$= -0.1001 \times 2^{-5} < \frac{\epsilon_{\text{mach}}}{2}$

c. $x = \frac{10}{3}$

10.01_2

$\text{fl}(\frac{10}{3}) = (-1)^0 \times 2^{1024-1023} \times (1.001 \dots 010)$

$\text{fl}(\frac{10}{3}) - \frac{10}{3} = -0.\overbrace{0 \dots 0}^{52 \text{ bit}} 1010 \dots$

$= -0.10 \times 2^{-52} < \frac{\epsilon_{\text{mach}}}{2}$

$$\begin{array}{r} \frac{1}{3} \times 2 \\ 0 \frac{2}{3} \\ 1 \frac{1}{3} \end{array}$$

