Exercise 2. Convert the following base 10 numbers to binary and express each as a floating point number $\mathrm{fl}(x)$ by using the Rounding to Nearest Rule. Use a 52 bit mantissa.

(a) 9.5

(b) 9.6

(c) 100.2

Solution.

Exercise 4. Do the following sums by hand in IEEE double precision computer arithmetic, using the Rounding to Nearest Rule: (Check your answers with MATLAB.)

(a) $(1 + (2^{-51} + 2^{-52} + 2^{-54})) - 1$

(b) $(1 + (2^{-51} + 2^{-52} + 2^{-60})) - 1$

Solution.

Exercise 5. Write each of the given numbers using `format hex`. Show your work. Check your answers with MATLAB.

(b) 21

(c) 1/8

(d) $\mathrm{fl}(1/3)$

Solution.

Exercise 10. Find the IEEE double precision representation $\mathrm{fl}(x)$, and find the exact difference $\mathrm{fl}(x) - x$ for the given real numbers. Show that the relative rounding error is no more than $1/2\epsilon_{\mathrm{mach}}$.

(a) $x = 2.75$

(b) $x = 2.7$

Solution.