
Humboldt University of Berlin

Einführung in das wissenschaftliche Rechnen

# Floating Point Arithmetic

*Christian Parpart & Kei Thoma*

May 29, 2019

# Contents

**Example 0.1.** Let $z_1 = 67.0$. We want to find the normalized binary form of this integer and and ten decimal places accurate. According to lemma **??**, we have

$$\begin{aligned} 67.0 \div 2 &= 33.0 + 1 \\ 33.0 \div 2 &= 16.0 + 1 \\ 16.0 \div 2 &= 8.0 + 0 \\ 8.0 \div 2 &= 4.0 + 0 \\ 4.0 \div 2 &= 2.0 + 0 \\ 2.0 \div 2 &= 1.0 + 0 \\ 1.0 \div 2 &= 0.0 + 1, \end{aligned}$$

therefore, we have $z_1 = 67.0 = (1000011)_2$. To normalize this number, we just have to move the decimal point six digits to the left. Since $z_1$ only has seven digits, we do not need to round. We have

$$z_1 = 67.0 = (1.000011 \times 2^6)_2$$

**Example 0.2.** Let $z_2 = 287.0$. To find the normalized binary form with respect to ten decimal places, we have

$$\begin{aligned} 287.0 \div 2 &= 143.0 + 1 \\ 143.0 \div 2 &= 71.0 + 1 \\ 71.0 \div 2 &= 35.0 + 1 \\ 35.0 \div 2 &= 17.0 + 1 \\ 17.0 \div 2 &= 8.0 + 1 \\ 8.0 \div 2 &= 4.0 + 0 \\ 4.0 \div 2 &= 2.0 + 0 \\ 2.0 \div 2 &= 1.0 + 0 \\ 1.0 \div 2 &= 0.0 + 1, \end{aligned}$$

therefore, $z_2 = 287.0 = (100011111)_2$. Again, there is no need to round any digits. Its normalized binary form is

$$z_2 = 287.0 = (1.00011111 \times 2^8)_2$$

**Example 0.3.** For a non-integer example, let $z_3 = 10.625$. To find the binary form of this number, we first separate $z_3 = 10.0 + 0.625$ and apply the algorithm of **??** on each summand. For 10.0 we have

$$\begin{aligned} 10.0 \div 2 &= 5.0 + 0 \\ 5.0 \div 2 &= 2.0 + 1 \\ 2.0 \div 2 &= 1.0 + 0 \\ 1.0 \div 2 &= 0.0 + 1 \end{aligned}$$

and for 0.625 we will multiply it with 2 until we get 0

$$0.625 \times 2 = 0.25 + 1$$
$$0.25 \times 2 = 0.5 + 0$$
$$0.5 \times 2 = 0.0 + 1$$

Combining both results together, we get $z_3 = (1010.101)_2$. To normalize, we move the decimal place three digits to the left and we have

$$z_3 = 10.625 = (1.010101 \times 2^3)_2.$$

**Example 0.4.** Perhaps a more interesting example is needed. Let $z_4 = 1.01$. As we did in **??**, we will separate $z_4$ in two parts; however, we immediately see that 1 is 1 in both decimal and binary system. We will therefore consider 0.01.

$$0.01 \times 2 = 0.02 + 0$$
$$0.02 \times 2 = 0.04 + 0$$
$$0.04 \times 2 = 0.08 + 0$$
$$0.08 \times 2 = 0.16 + 0$$
$$0.16 \times 2 = 0.32 + 0$$
$$0.32 \times 2 = 0.64 + 0$$
$$1.28 \times 2 = 0.28 + 1$$
$$0.28 \times 2 = 0.56 + 0$$
$$0.56 \times 2 = 0.12 + 1$$
$$0.12 \times 2 = 0.24 + 0$$

We could go on, but since we only need to find the normalized binary form with respect to ten decimal places. We have

$$z_4 = 1.01 = (1.0000001010 \times 2^0)_2$$

which is already normalized.