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

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

which is already normalized.

Example 0.5. As we already fell into the rabit hole of numbers which have endlessly long binary forms, let's continue with $z_5 = 0.0002$. For this example, we must stay

diligent and iterate many times over the algorithm.

$$\begin{aligned}
0.0002 \times 2 &= 0.0004 + 0 \\
0.0004 \times 2 &= 0.0008 + 0 \\
0.0008 \times 2 &= 0.0016 + 0 \\
0.0016 \times 2 &= 0.0032 + 0 \\
0.0032 \times 2 &= 0.0064 + 0 \\
0.0064 \times 2 &= 0.0128 + 0 \\
0.0128 \times 2 &= 0.0256 + 0 \\
0.0256 \times 2 &= 0.0512 + 0 \\
0.0512 \times 2 &= 0.1024 + 0 \\
0.1024 \times 2 &= 0.2048 + 0 \\
0.2048 \times 2 &= 0.4096 + 0 \\
0.4096 \times 2 &= 0.8192 + 0 \\
0.8192 \times 2 &= 0.6384 + 1
\end{aligned}$$

We got our first 1! Now we only have to find a maximum of 10 more digits.

$$\begin{aligned}
0.6384 \times 2 &= 0.2768 + 1 \\
0.2768 \times 2 &= 0.5536 + 0 \\
0.5536 \times 2 &= 0.1072 + 1 \\
0.1072 \times 2 &= 0.2144 + 0 \\
0.2144 \times 2 &= 0.4288 + 0 \\
0.4288 \times 2 &= 0.8576 + 0 \\
0.8576 \times 2 &= 0.7152 + 1 \\
0.7152 \times 2 &= 0.4304 + 1 \\
0.4304 \times 2 &= 0.8608 + 0 \\
0.8608 \times 2 &= 0.7216 + 1
\end{aligned}$$

Therefore, we have $z_5 = 0.0002 \approx (0.00000000000011010001101)_2$ and normalized we have

$$z_5 = 0.0002 \approx (1.1010001101 \times 2^{-13})_2$$

Example 0.6. For the more mathematically minded, we have last but not least $z_6 = \frac{1}{3}$.

$$\begin{aligned}
\frac{1}{3} \times 2 &= \frac{2}{3} + 0 \\
\frac{2}{3} \times 2 &= \frac{1}{3} + 1
\end{aligned}$$

We already see a patern here; further calculation is not needed. We simply have

$$z_6 = \frac{1}{3} \approx (1.0101010101 \times 2^{-2})_2$$

For posterity and stripped from tedious calculation, in the following is a table summerizing the results of **??**.

decimal representation	normalized binary representation
67.0	1.000011×2^{6}
287.0	1.00011111×2^{8}
10.625	1.010101×2^{3}
1.01	$1.0000001010 \times 2^{0}$
0.0002	$1.1010001101 \times 2^{-13}$
$\frac{1}{3}$	$1.0101010101 \times 2^{-2}$