Floating-Point Number System

- Two kinds of computers:
  - ◇ Analog Computer: Numbers are represented by some physical quantities, such as the length of a bar or the intensity of a voltage.
  - ◇ Digital Computer: Numbers are represented by a sequence of digits where each digit is represented by a specific physical quantity.
- Most modern computers are digital computers using the floating-point arithmetic.
- A floating-point number is of the form

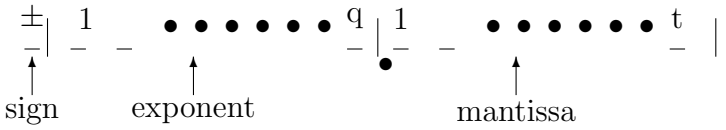
$$\pm .d_1 \dots d_t \times \beta^e$$

where

$$\begin{aligned} \beta &= \text{integer} = \text{fixed base;} \\ e &= \text{integer} = \text{exponent;} \\ t &= \text{number of digits;} \\ & 1 \le d_1 \le \beta - 1; \\ & 0 \le d_i \le \beta - 1; \quad \text{for } 2 \le i \le t. \end{aligned}$$

  - ◇ The fractional part $.d_1 \dots d_t = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t}$ is called the mantissa of the floating-point number.

- In a digital computer with word length n, a floating-point number is sto red as follows:



  where each horizontal bar represents either a bit (a binary digit), a hexadecimal (base 16) or so on.

  - ◇ Note that the exponent portion should be able to represent integers from 0 to $\beta^q - 1$ (why?), but these are only positive integers. To represent negative integers, we assume some kind of shifting strategy by sacrificing some larger exponents.

- Consider the normalized number system with $\beta = 2$, $t = 3$ and $q = 2$. With exponent shifted by 1. The positive half (Recall that to represent the negative half will need another bit to denote the "-" sign) of the system contains the following numbers:

	-1	0	1	2
.100	4/16	4/8	4/4	4/2
.101	5/16	5/8	5/4	5/2
.110	6/16	6/8	6/4	6/2
.111	7/16	7/8	7/4	7/2

  - ◇ Note that with each fixed e, in the interval $[\beta^{e-1}, \beta^e)$, each floating-point is equally spaced by β^{e-t}.
  - ◇ Note the first bit is always "1". In practice, there is no need to store this universal number. We may therefore use the mantissa to represent four instead of three bits, increasing the accuracy.

- The distribution of the above system of numbers on the real line is particularly informative:

  - ◇ Any real number that is not within the floating-point number system of a particular computer has to be rounded or truncated, and hence the round-off error is introduced.

- You really need to get yourself familiarized with the floating-point number system. **A project is given today!**