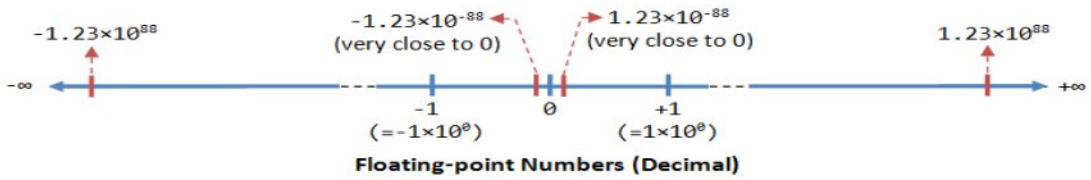
FLOATING POINTS:

- A floating point no is a real number which is used to represent a very large number as well as very small number value. It could also represent very small negative number and very small negative number as shown below.



Floating-point Numbers (Decimal)

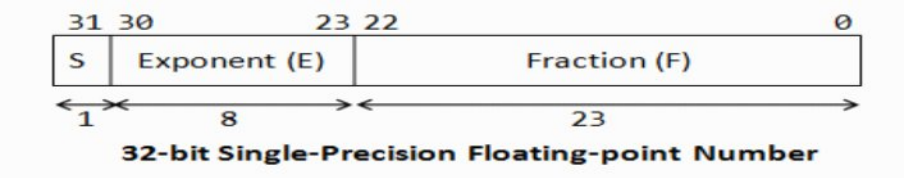
There are three parts in the floating-point representation:

- The sign bit (S) is self-explanatory (0 for positive numbers and 1 for negative numbers).
- For the exponent (E), a so-called *bias* (or *excess*) is applied so as to represent both positive and negative exponent. The bias is set at half of the range. For single precision with an 8-bit exponent, the bias is 127 (or excess-127). For double precision with a 11-bit exponent, the bias is 1023 (or excess-1023).
- The fraction (F) (also called the mantissa or significand) is composed of an implicit leading bit (before the radix point) and the fractional bits (after the radix point). The leading bit for normalized numbers is 1; while the leading bit for denormalized numbers is 0.
- Modern computers adopt IEEE 754 standard for representing floating-point numbers.
- There are two representation schemes:
    1. 32-bit single-precision and
    2. 64-bit double-precision.

32 BIT SINGLE PRECISION FLOATING POINT NUMBERS:

In 32-bit single-precision floating-point representation:

- The most significant bit is the sign bit (S), with 0 for positive numbers and 1 for negative numbers.
- The following 8 bits represent exponent (E).
- The remaining 23 bits represents fraction (F).

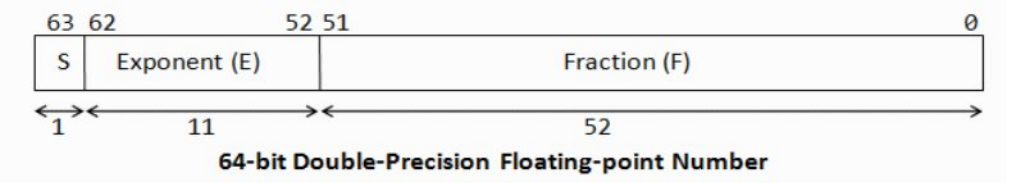


32-bit Single-Precision Floating-point Number

64 BIT DOUBLE PRECISION FLOATING POINT NUMBERS:

The representation scheme for 64-bit double-precision is similar to the 32-bit single-precision:

- The most significant bit is the sign bit (S), with 0 for positive numbers and 1 for negative numbers.
- The following 11 bits represent exponent (E).
- The remaining 52 bits represents fraction (F).



64-bit Double-Precision Floating-point Number