Floating-Point Representation

- The IEEE Standard for Floating-Point Arithmetic (IEEE 754) is a technical standard for floating-point computation which was established in 1985 by the **Institute of Electrical and Electronics Engineers (IEEE)**

IEEE 754 numbers are divided into two representation based on the three components (Sign, Exponent and Mantissa):

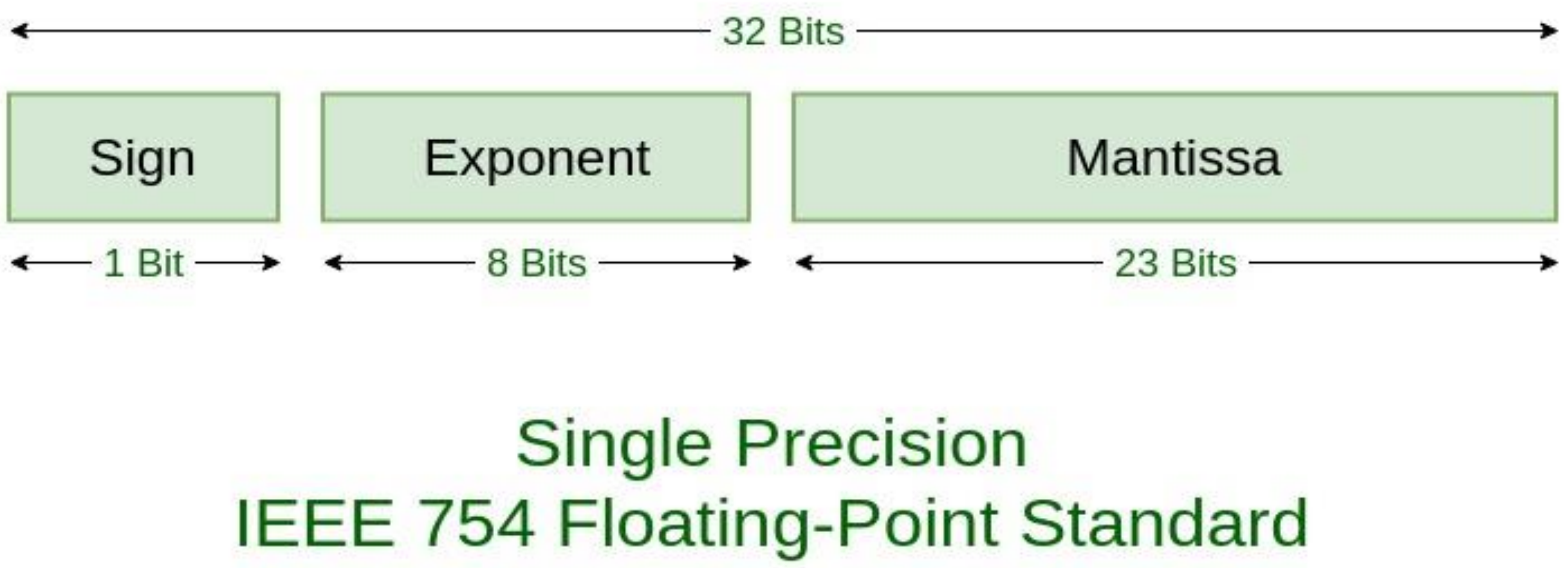
- ❑ Single precision (32-bit)
- ❑ Double precision (64-bit)

IEEE 754 has 3 basic components:

1. **The Sign of Mantissa :**
   This is as simple as the name. 0 represents a positive number while 1 represents a negative number.
2. **The Biased exponent:**
   The exponent field needs to represent both positive and negative exponents. A bias is added to the actual exponent in order to get the stored exponent.
3. **The Normalised Mantissa:**
   The mantissa is part of a number in scientific notation or a floating-point number, consisting of its significant digits. In binary, we have only 2 digits, i.e. O and 1. So a normalised mantissa is one with only one 1 to the left of the decimal.

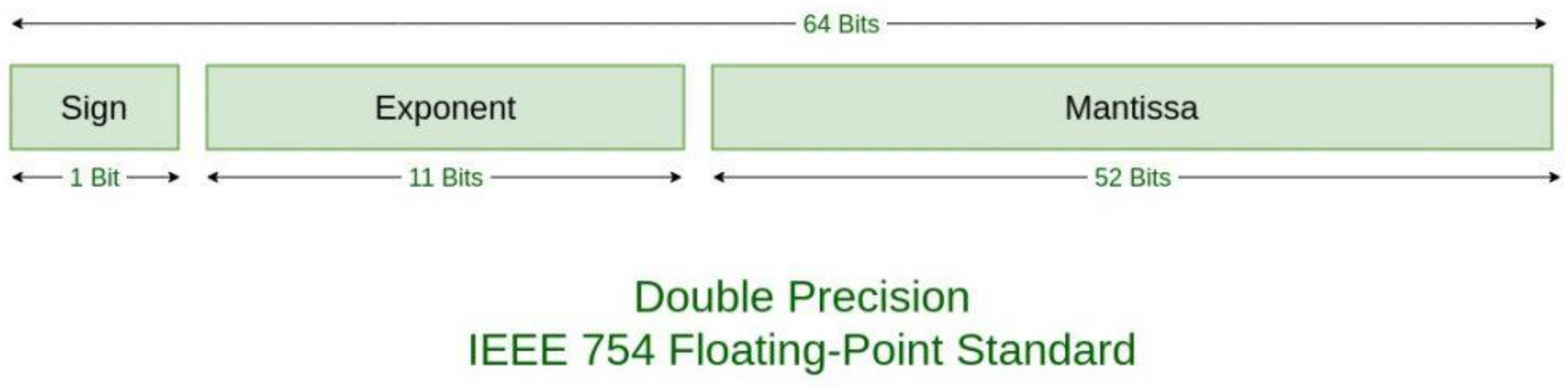
IEEE-754 single precision (32-bit)

The IEEE-754 single precision floating point standard uses 1-bit for sign an 8-bit exponent (with a bias of 127) and a 23-bit significand



IEEE-754 double precision (64-bit)

The IEEE-754 double precision standard uses 1-bit for sign and 11-bit exponent (with a bias of 1023) and a 52-bit significand



IEEE-754 Representation

- In both the IEEE single-precision and double precision floating-point standard, the significant has an implied 1 to the LEFT of the radix point.
- The format for a significand using the IEEE format is: 1.xxx…
- For example, 4.5 = .1001 x 23 in IEEE format is 4.5 = 1.001 x 22. The 1 is implied, which means is does not need to be listed in the significand (the significand would include only 001).

SINGLE-PRECISION RANGE

- Exponents 00000000 and 11111111 are reserved
- Smallest value
  - Exponent: 00000001
    $\Rightarrow$ actual exponent = 1 – 127 = –126
  - Fraction: 000…00 $\Rightarrow$ significand = 1.0
  - $\pm 1.0 \times 2^{-126} \approx \pm 1.2 \times 10^{-38}$
- Largest value
  - exponent: 11111110
    $\Rightarrow$ actual exponent = 254 – 127 = +127
  - Fraction: 111…11 $\Rightarrow$ significand $\approx$ 2.0
  - $\pm 2.0 \times 2^{+127} \approx \pm 3.4 \times 10^{+38}$

DOUBLE-PRECISION RANGE

- Exponents 0000…00 and 1111…11 are reserved
- Smallest value
  - Exponent: 00000000001
    $\Rightarrow$ actual exponent = 1 – 1023 = –1022
  - Fraction: 000…00 $\Rightarrow$ significand = 1.0
  - $\pm 1.0 \times 2^{-1022} \approx \pm 2.2 \times 10^{-308}$
- Largest value
  - Exponent: 11111111110
    $\Rightarrow$ actual exponent = 2046 – 1023 = +1023
  - Fraction: 111…11 $\Rightarrow$ significand $\approx$ 2.0
  - $\pm 2.0 \times 2^{+1023} \approx \pm 1.8 \times 10^{+308}$

IEEE 754 Special Number Representation

Single Precision		Double Precision		Number Represented
Exponent	Significand	Exponent	Significand	
0	0	0	0	0
0	nonzero	0	nonzero	Denormalized number[1]
1 to 254	anything	1 to 2046	anything	Floating Point Number
255	0	2047	0	Infinity[2]
255	nonzero	2047	nonzero	NaN (Not A Number)[3]

[1] May be returned as a result of underflow in multiplication

[2] Positive divided by zero yields “infinity”

[3] Zero divide by zero yields NaN “not a number”

Example for Single Precision

- Example: Express -3.75 as a floating point number using IEEE single precision.
- First, let's normalize according to IEEE rules:
  - $-3.75 = -11.11_2 = -1.111 \times 2^1$
  - The bias is 127, so we add 127 + 1 = 128 (this is our exponent)

1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(implied)

  - Since we have an implied 1 in the significand, this equates to

    $-(1).111_2 \times 2^{(128 - 127)} = -1.111_2 \times 2^1 = -11.11_2 = -3.75.$

Floating-Point Example

- What number is represented by the single-precision float

  11000000101000…00

  - S = 1
  - Fraction = $01000\ldots00_2$
  - Exponent = $10000001_2 = 129$

- $x = (-1)^1 \times (1 + 01_2) \times 2^{(129 - 127)}$

  $= (-1) \times 1.25 \times 2^2$

  $= -5.0$

- Represent –0.75
  - $-0.75 = (-1)^1 \times 1.1_2 \times 2^{-1}$

    $= -1 \times 1.\,½ \times ½$

    $= -1.5 * .5 = -0.75$
  - $S = 1$
  - Fraction = $1000\ldots00_2$
  - Exponent = –1 + Bias
    - Single: $-1 + 127 = 126 = 01111110_2$
    - Double: $-1 + 1023 = 1022 = 01111111110_2$

- Single: 1011111101000…00
- Double: 1011111111101000…00

Summary

IEEE Floating-Point Format

	single: 8 bits double: 11 bits	single: 23 bits double: 52 bits
S	Exponent	Fraction

$$\boxed{x = (-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}}$$

- Exponent: excess representation: actual exponent +
  Bias
  - Ensures exponent is unsigned
  - Single precision: Bias = 127;
  - Double precision: Bias = 1203