

# Floating point representation

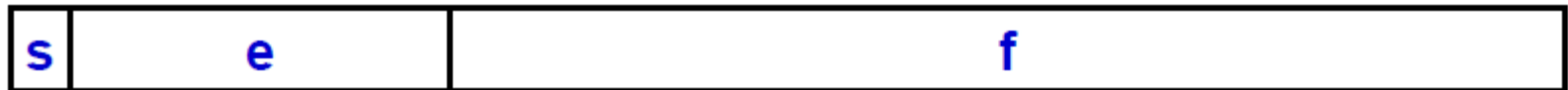


# Floating-Point Representation

IEEE numbers are stored using a kind of scientific notation.

$$\pm \text{mantissa} * 2^{\text{exponent}}$$

We can represent floating-point numbers with three binary fields: a sign bit **s**, an exponent field **e**, and a fraction field **f**.



- ❑ The field **f** contains a binary fraction.
- ❑ The actual mantissa of the floating-point value is  $(1 + f)$ .
  - In other words, there is an implicit 1 to the left of the binary point.
  - For example, if **f** is **01101...**, the mantissa would be **1.01101...**
- ❑ The **e** field represents the exponent as a **biased** number.
  - It contains the actual exponent **plus 127** for single precision, or the actual exponent **plus 1023** in double precision.
  - This converts all single-precision exponents from **-126** to **+127** into unsigned numbers from 1 to 254, and all double-precision exponents from **-1022** to **+1023** into unsigned numbers from 1 to 2046.

# Mapping Between e and Actual Exponent

e		Actual Exponent
0000 0000		Reserved
0000 0001	$1-127 = -126$	$-126_{10}$
0000 0010	$2-127 = -125$	$-125_{10}$
...		...
0111 1111		$0_{10}$
...		...
1111 1110	$254-127=127$	$127_{10}$
1111 1111		Reserved

## Special Values (single-precision)

E	F	meaning	Notes
00000000	0...0	0	+0.0 and -0.0
00000000	X...X	Valid number	Unnormalized $=(-1)^S \times 2^{-126} \times (0.F)$
11111111	0...0	Infinity	
11111111	X...X	Not a Number	

# Converting an IEEE 754 number to decimal



- The decimal value of an IEEE number is given by the formula:

$$(1 - 2s) * (1 + f) * 2^{e-\text{bias}}$$

- Here, the s, f and e fields are assumed to be in decimal.
  - (1 - 2s) is 1 or -1, depending on whether the sign bit is 0 or 1.
  - We add an implicit 1 to the fraction field f, as mentioned earlier.
  - Again, the bias is either 127 or 1023, for single or double precision.

# Example IEEE-decimal conversion

- Let's find the decimal value of the following IEEE number.

1      01111100      110000000000000000000000

- First convert each individual field to decimal.

- The sign bit  $s$  is 1.
- The  $e$  field contains  $01111100 = 124_{10}$ .
- The mantissa is  $0.11000... = 0.75_{10}$ .

- Then just plug these decimal values of  $s$ ,  $e$  and  $f$  into our formula.

$$(1 - 2s) * (1 + f) * 2^{e-\text{bias}}$$

- This gives us  $(1 - 2) * (1 + 0.75) * 2^{124-127} = (-1.75 * 2^{-3}) = -0.21875$ .

## Exercise

---

- What is the single-precision representation of 639.6875

$$\begin{aligned}639.6875 &= 1001111111.1011_2 \\ &= 1.0011111111011 \times 2^9\end{aligned}$$

$$s = 0$$

$$e = 9 + 127 = 136 = 10001000$$

$$f = 0011111111011$$

The single-precision representation is:

0 10001000 0011111111011000000000



# Decimal value of the IEEE number

- 1 10000001 11000000000000000000000000000000
- 0 10001000 10110000000000000000000000000000





# Single precision representation of

- 534,625
- -0,00345
- -430,5625
- 0,09375

