

# Lecture 10: Floating Point

---

- Today's topics:
  - IEEE 754 representations

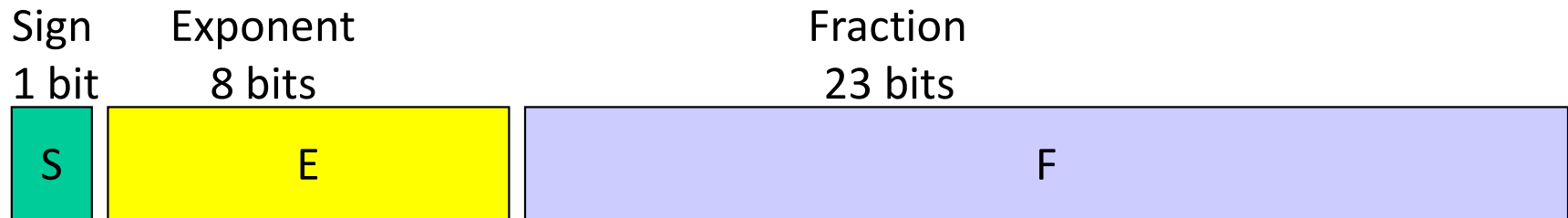
# Floating Point

---

- Normalized scientific notation: single non-zero digit to the left of the decimal (binary) point – example:  $3.5 \times 10^9$
- $1.010001 \times 2^{-5}_{\text{two}} = (1 + 0 \times 2^{-1} + 1 \times 2^{-2} + \dots + 1 \times 2^{-6}) \times 2^{-5}_{\text{ten}}$
- A standard notation enables easy exchange of data between machines and simplifies hardware algorithms – the IEEE 754 standard defines how floating point numbers are represented

# Sign and Magnitude Representation

---



- More exponent bits → wider range of numbers (not necessarily more numbers – recall there are infinite real numbers)
- More fraction bits → higher precision
- Register value =  $(-1)^S \times F \times 2^E$
- Since we are only representing normalized numbers, we are guaranteed that the number is of the form 1.xxxx..  
Hence, in IEEE 754 standard, the 1 is implicit  
Register value =  $(-1)^S \times (1 + F) \times 2^E$

# Exponent Representation

---

- To simplify sort, sign was placed as the first bit
- For a similar reason, the representation of the exponent is also modified: in order to use integer compares, it would be preferable to have the smallest exponent as 00...0 and the largest exponent as 11...1
- This is the biased notation, where a bias is subtracted from the exponent field to yield the true exponent
- IEEE 754 single-precision uses a bias of 127 (since the exponent must have values between -127 and 128)...double precision uses a bias of 1023

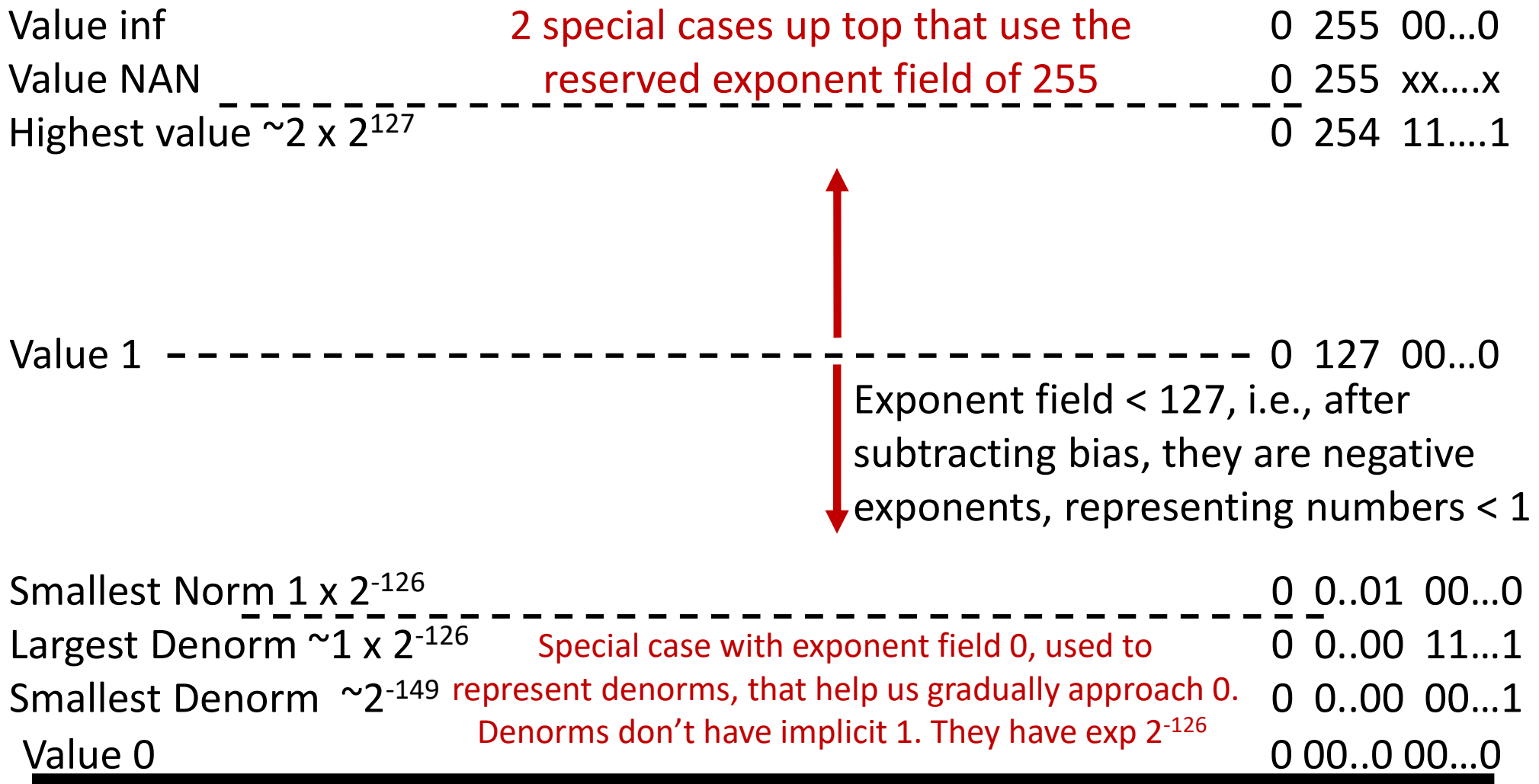
Final representation:  $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$



# Details

---

- The number “0” has a special code so that the implicit 1 does not get added: the code is all 0s  
(it may seem that this takes up the representation for 1.0, but given how the exponent is represented, that’s not the case)  
(see discussion of denorms in the textbook)
- The largest exponent value (with zero fraction) represents +/- infinity
- The largest exponent value (with non-zero fraction) represents NaN (not a number) – for the result of 0/0 or (infinity minus infinity)
- Note that these choices impact the smallest and largest numbers that can be represented



Same rules as above, but the sign bit is 1  
 Same magnitudes as above, but negative numbers



# Examples

---

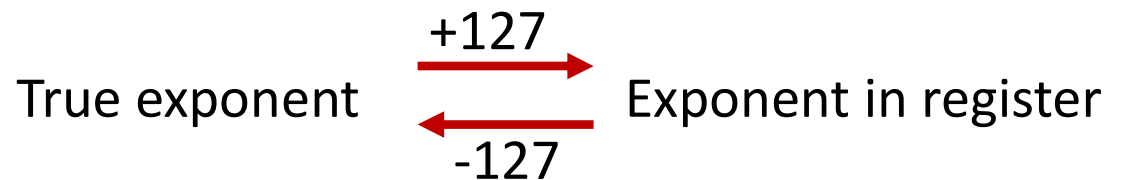
Final representation:  $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

- Represent  $-0.75_{\text{ten}}$  in single and double-precision formats

Single:  $(1 + 8 + 23)$

Double:  $(1 + 11 + 52)$

Remember:



- What decimal number is represented by the following single-precision number?

1 1000 0001 01000...0000



# Examples

---

Final representation:  $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

- Represent  $-0.75_{\text{ten}}$  in single and double-precision formats

Single:  $(1 + 8 + 23)$

1 0111 1110 1000...000

Double:  $(1 + 11 + 52)$

1 0111 1111 110 1000...000

- What decimal number is represented by the following single-precision number?

1 1000 0001 01000...0000

-5.0

## Example 2

---

Final representation:  $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

- Represent  $36.90625_{\text{ten}}$  in single-precision format

$$36 / 2 = 18 \text{ rem } 0$$

$$18 / 2 = 9 \text{ rem } 0$$

$$9 / 2 = 4 \text{ rem } 1$$

$$4 / 2 = 2 \text{ rem } 0$$

$$2 / 2 = 1 \text{ rem } 0$$

$$1 / 2 = 0 \text{ rem } 1$$



36 is 100100

$$0.90625 \times 2 = 1.81250$$

$$0.8125 \times 2 = 1.6250$$

$$0.625 \times 2 = 1.250$$

$$0.25 \times 2 = 0.50$$

$$0.5 \times 2 = 1.00$$

$$0.0 \times 2 = 0.0$$



0.90625 is 0.1110100...0

## Example 2

---

Final representation:  $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

We've calculated that  $36.90625_{\text{ten}} = 100100.1110100\dots 0$  in binary  
Normalized form =  $1.001001110100\dots 0 \times 2^5$   
(had to shift 5 places to get only one bit left of the point)

The sign bit is 0 (positive number)

The fraction field is 001001110100...0 (the 23 bits after the point)

The exponent field is  $5 + 127$  (have to add the bias) = 132,  
which in binary is 10000100

The IEEE 754 format is 0 10000100 001001110100.....0  
sign exponent 23 fraction bits

Remember:

True exponent

Exponent in register