Review Questions

$2^{-1} = 0.5$
$2^{-2} = 0.25$
$2^{-3} = 0.125$
$2^{-4} = 0.0625$

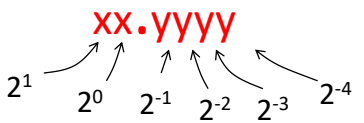
- Convert 11.375_{10} to normalized binary scientific notation
- What is the value encoded by the following floating-point number?

  0b 0 | 1000 0000 | 110 0000 0000 0000 0000 0000
  - bias = $2^{w-1}-1$
  - exponent = E – bias
  - mantissa = 1.M

Representation of Fractions

- "Binary Point," like decimal point, signifies boundary between integer and fractional parts:

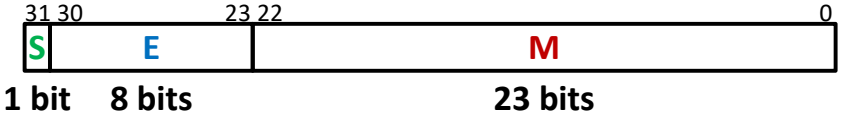
  Example 6-bit representation:



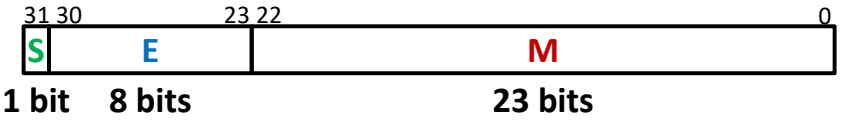
- Example: $10.1010_2 = 1\times2^1 + 1\times2^{-1} + 1\times2^{-3} = 2.625_{10}$

Floating Point Encoding (Review)

- Use normalized, base 2 scientific notation:
  - Value: ±1 × Mantissa × $2^{Exponent}$
  - Bit Fields: $(-1)^S \times 1.M \times 2^{(E-bias)}$
- Representation Scheme:
  - Sign bit (0 is positive, 1 is negative)
  - Mantissa (a.k.a. significand) is the fractional part of the number in normalized form and encoded in bit vector **M**
  - Exponent weights the value by a (possibly negative) power of 2 and encoded in the bit vector **E**



The Mantissa (Fraction) Field (Review)



$(-1)^S \times (1 . M) \times 2^{(E-bias)}$

- Note the implicit leading 1 in front of the M bit vector
  - Example: 0b 0011 1111 1100 0000 0000 0000 0000 0000 is read as $1.1_2 = 1.5_{10}$, *not* $0.1_2 = 0.5_{10}$
  - Gives us an extra bit of *precision*
- Mantissa "limits"
  - Low values near M = 0b0…0 are close to 2^{Exp}
  - High values near M = 0b1…1 are close to 2^{Exp+1}

Normalized Floating Point Conversions

- FP → Decimal
  1. Append the bits of M to implicit leading 1 to form the mantissa.
  2. Multiply the mantissa by $2^{E - bias}$.
  3. Multiply the sign $(-1)^S$.
  4. Multiply out the exponent by shifting the binary point.
  5. Convert from binary to decimal.
- Decimal → FP
  1. Convert decimal to binary.
  2. Convert binary to normalized scientific notation.
  3. Encode sign as S (0/1).
  4. Add the bias to exponent and encode E as unsigned.
  5. The first bits after the leading 1 that fit are encoded into M.

Practice Question

$2^{-1} = 0.5$
$2^{-2} = 0.25$
$2^{-3} = 0.125$
$2^{-4} = 0.0625$

- Convert the decimal number **-7.375** into floating point representation

New Representation Limits

- New largest value (besides ∞)?
  - E = 0xFF has now been taken!
  - E = 0xFE has largest: $1.1\ldots1_2 \times 2^{127} = 2^{128} - 2^{104}$
- New numbers closest to 0:
  - E = 0x00 taken; next smallest is E = 0x01
  - a = $1.0\ldots00_2 \times 2^{-126} = 2^{-126}$
  - b = $1.0\ldots01_2 \times 2^{-126} = 2^{-126} + 2^{-149}$
  - Normalization and implicit 1 are to blame
  - *Special case:* E = 0, M ≠ 0 are denormalized numbers
    - Mantissa has implicit 0 instead of implicit 1
    - Store much smaller numbers