Fixed Point Numbers

32-bit integers have an implied number point on the right

Setting the number point in the middle allows fractional values

- bits to the left of the point represent non-negative powers of 2
- bits to the right of the point represent negative powers of 2

To illustrate assume 8-bit patterns with point in middle:

Implied number point

0 1 1 0 1 0 1 0 = 4 + 2 + 0.5 + 0.125 = 6.625

2^{+2} 2^{+1} 2^{-1} 2^{-3}

In hex: 6.A = $6 * 16^{0} + 10*16^{-1} = 6 + 10*0.0625 = 6.625$

The position of the number point is set, thus the name “fixed-point”

We would like to be able to represent ranges of values

Very large integer parts

Very small fractional parts

This requires that the number point *slide* or *float*

Representation must specify number point location

Floating Point Representation must include:

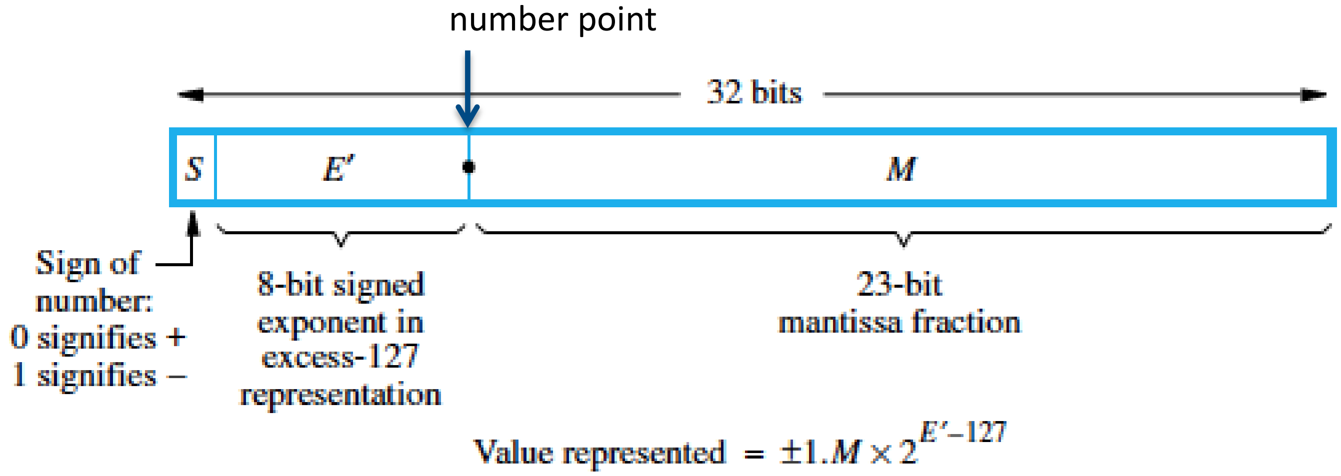
- Sign of number
- Significant bits
- Signed exponent for an implied base of 2

Width of exponent field determines range

Number of significant bits determines precision

IEEE-754 Standard specifies a common floating format

- 32-bit single precision
- 64-bit double precision



E’ = exponent + 127 (excess-127 form) also called the “characteristic”

“Normalized” numbers have an implied 1 to the left of the number point

Single Precision Floating Point

With single precision numbers $0 \leq E' \leq 255$

- But endpoints (0 and 255) are used for special cases
- 0 denotes exponent of -126 for denormalized
  - Denormalized numbers have 0 to left of number point
  - $E'=0$ and $M \neq 0$ for denormalized numbers
  - $E'=0$ and $M=0$ for true zero
  - Denormalized numbers allow gradual underflow

- $E'=255$ denotes ±infinity or NaN (not a number)
  - $E'=255$ and $M \neq 0$ for NaN (e.g. 0/0 or $\sqrt{-1}$)
  - $E'=255$ and $M=0$ for $\pm\infty$

Single Precision Floating Point

Provides 7 decimal places of precision

With the hidden 1, the significand is essentially 24 bits

X, the number of decimal places would be such that:

$10^x \cong 2^{24}$

$\mathrm{Log}(10^x) \cong \log(2^{24})$

$X = 24 * \log(2) = 24 * 0.301 = 7.22$

Approximate range for normalized values is $\pm(10^{\pm 38})$

$\pm 2^{-126}$ to $\pm 2^{+128}$

0x14140000

0	0 0 1 0 1 0 0 0	• 0 0 1 0 1 0 . . .	0

$$\text{Value represented} = 1.001010\ldots0 \times 2^{-87}$$

Exponent = characteristic 00101000 – 127 = 40 – 127 = -87