Floating-point Numbers

Appendix B

Principles of Floating Point (1)

- Must separate range from precision
- Use scientific notation $n = f \times 10^e$
  - *f* is the fraction or mantissa
  - *e* is the exponent (a positive or negative integer)
- Examples

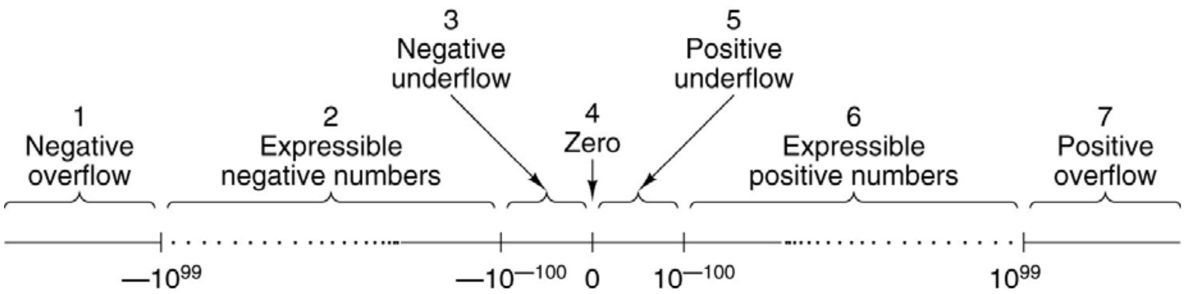
$3.14 = 0.314 \times 10^1 = 3.14 \times 10^0$
$0.000001 = 0.1 \times 10^{-5} = 1.0 \times 10^{-6}$
$1941 = 0.1941 \times 10^4 = 1.941 \times 10^3$

Principles of Floating Point (2)

Seven Regions of Real Number Line

- Large negative numbers less than -0.999×10^{99}.
- Negative numbers between -0.999×1099 and -0.100×10^{-99}.
- Small negative numbers, magnitudes less than $0.100 \times 10-^{99}$.
- Zero.
- Small positive numbers, magnitudes less than 0.100×10^{-99}.
- Positive numbers between 0.100×10^{-99} and 0.999×1099.
- Large positive numbers greater than 0.999×10^{99}.

Principles of Floating Point (3)



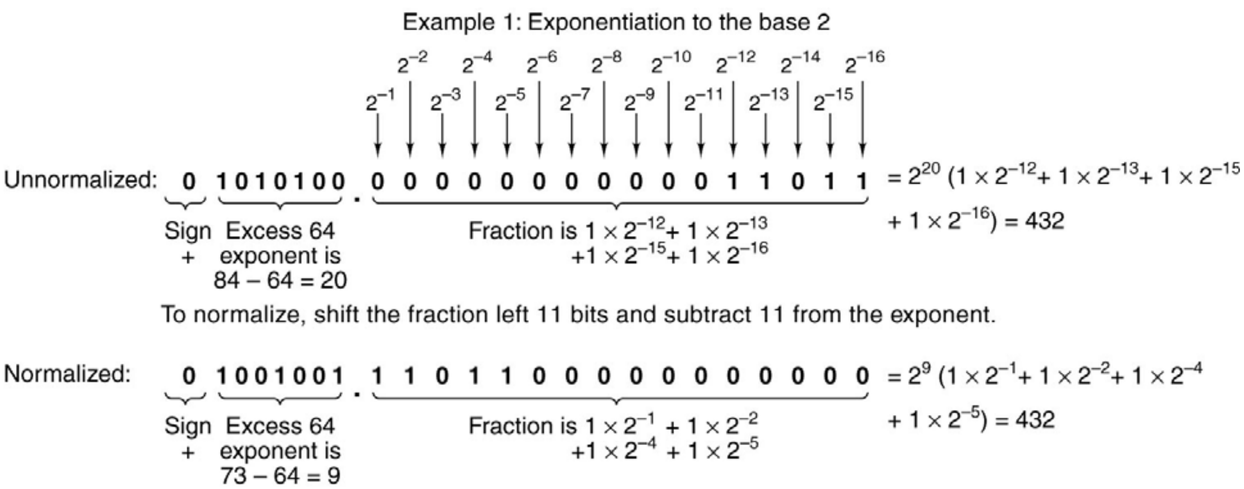
The real number line can be divided into seven regions.

Principles of Floating Point (4)

Digits in fraction	Digits in exponent	Lower bound	Upper bound
3	1	10^{-12}	10^{9}
3	2	10^{-102}	10^{99}
3	3	10^{-1002}	10^{999}
3	4	10^{-10002}	10^{9999}
4	1	10^{-13}	10^{9}
4	2	10^{-103}	10^{99}
4	3	10^{-1003}	10^{999}
4	4	10^{-10003}	10^{9999}
5	1	10^{-14}	10^{9}
5	2	10^{-104}	10^{99}
5	3	10^{-1004}	10^{999}
5	4	10^{-10004}	10^{9999}
10	3	10^{-1009}	10^{999}
20	3	10^{-1019}	10^{999}

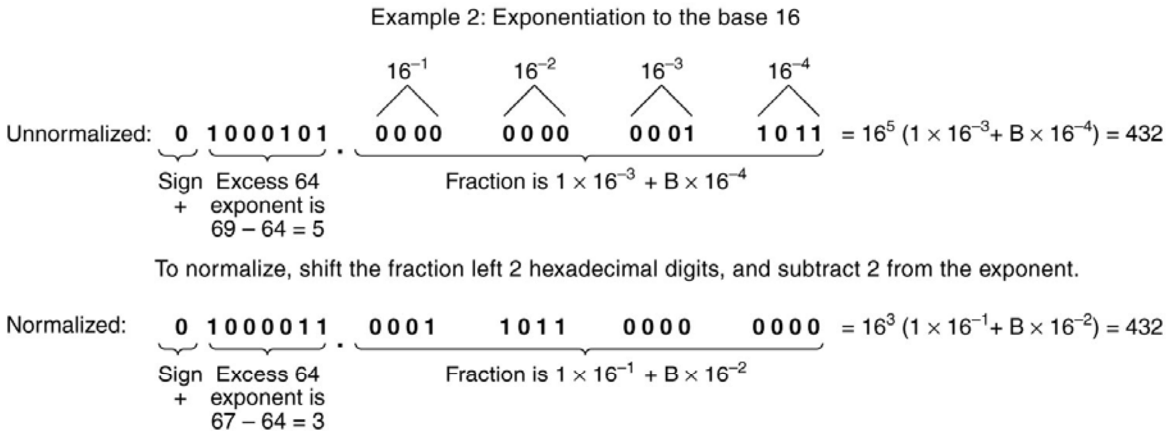
The approximate lower and upper bounds of expressible (unnormalized) floating-point decimal numbers.

IEEE Floating-point Standard 754 (1)



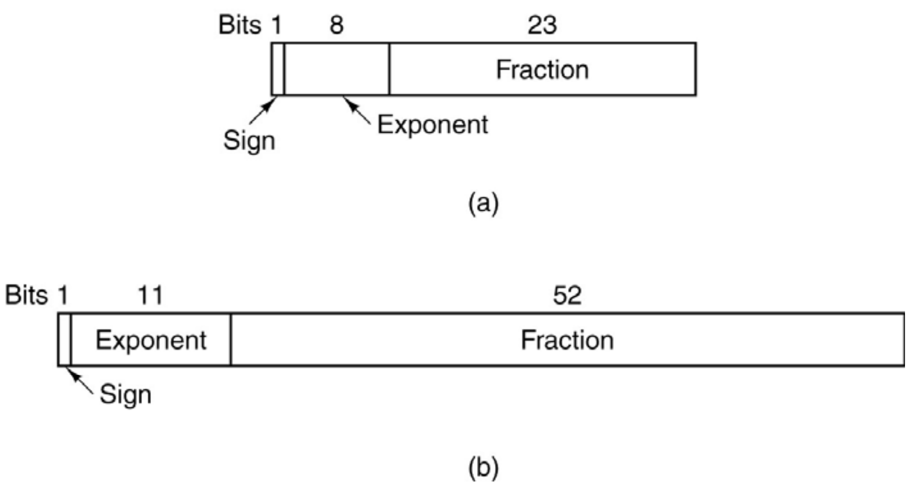
Examples of normalized floating-point numbers.

IEEE Floating-point Standard 754 (2)



Examples of normalized floating-point numbers.

IEEE Floating-point Standard 754 (3)



IEEE floating-point formats.
(a) Single precision. (b) Double precision.

IEEE Floating-point Standard 754 (4)

Item	Single precision	Double precision
Bits in sign	1	1
Bits in exponent	8	11
Bits in fraction	23	52
Bits, total	32	64
Exponent system	Excess 127	Excess 1023
Exponent range	−126 to +127	−1022 to +1023
Smallest normalized number	2^{-126}	2^{-1022}
Largest normalized number	approx. 2^{128}	approx. 2^{1024}
Decimal range	approx. 10^{-38} to 10^{38}	approx. 10^{-308} to 10^{308}
Smallest denormalized number	approx. 10^{-45}	approx. 10^{-324}

Characteristics of IEEE floating-point numbers.

IEEE Floating-point Standard 754 (5)

Normalized	±	0 < Exp < Max	Any bit pattern
Denormalized	±	0	Any nonzero bit pattern
Zero	±	0	0
Infinity	±	1 1 1…1	0
Not a number	±	1 1 1…1	Any nonzero bit pattern

Sign bit

IEEE numerical types.