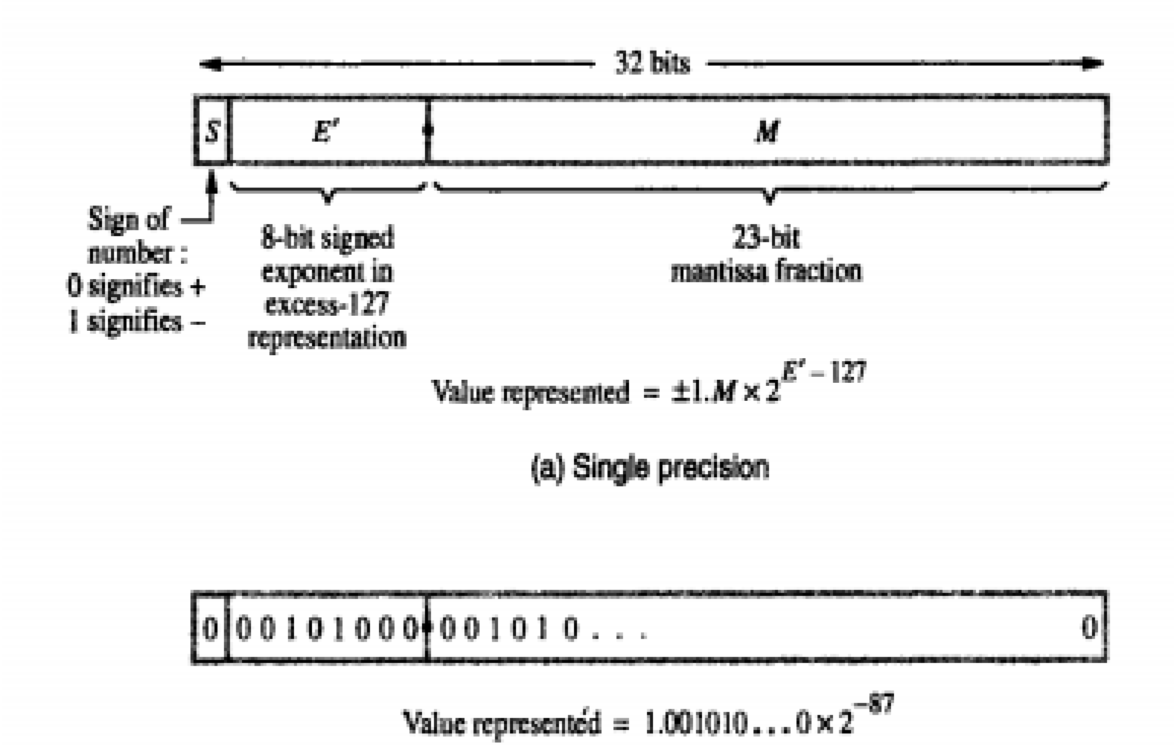
FLOATING-POINT NUMBERS AND OPERATIONS

IEEE STANDARD FOR FLOATING-POINT NUMBERS

SINGLE PRECISION


32 bits
S
E′
M
Sign of number : 0 signifies + 1 signifies –
8-bit signed exponent in excess-127 representation
23-bit mantissa fraction
0
0 0 1 0 1 0 0 0
0 0 1 0 1 0 . . .
0

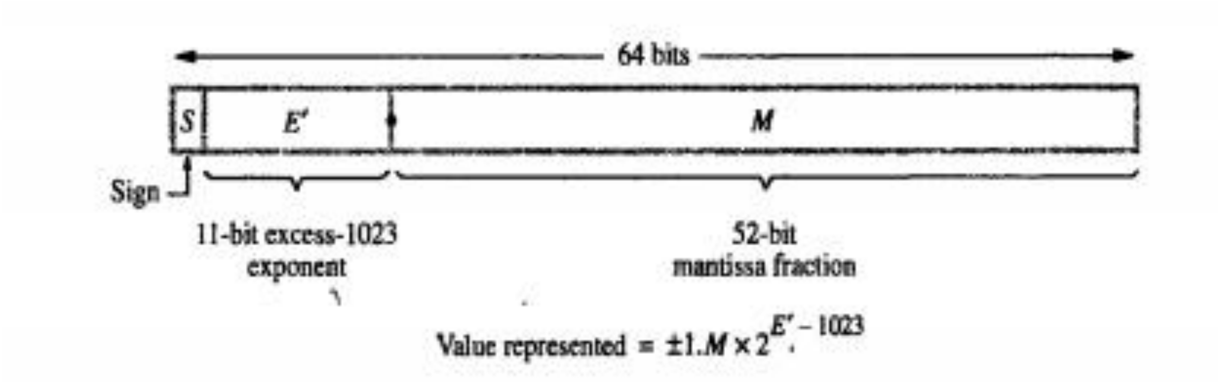
Value represented $= \pm 1.M \times 2^{E' - 127}$

(a) Single precision

Value represented $= 1.001010 \ldots 0 \times 2^{-87}$

(b) Example of a single-precision number

DOUBLE PRECISION

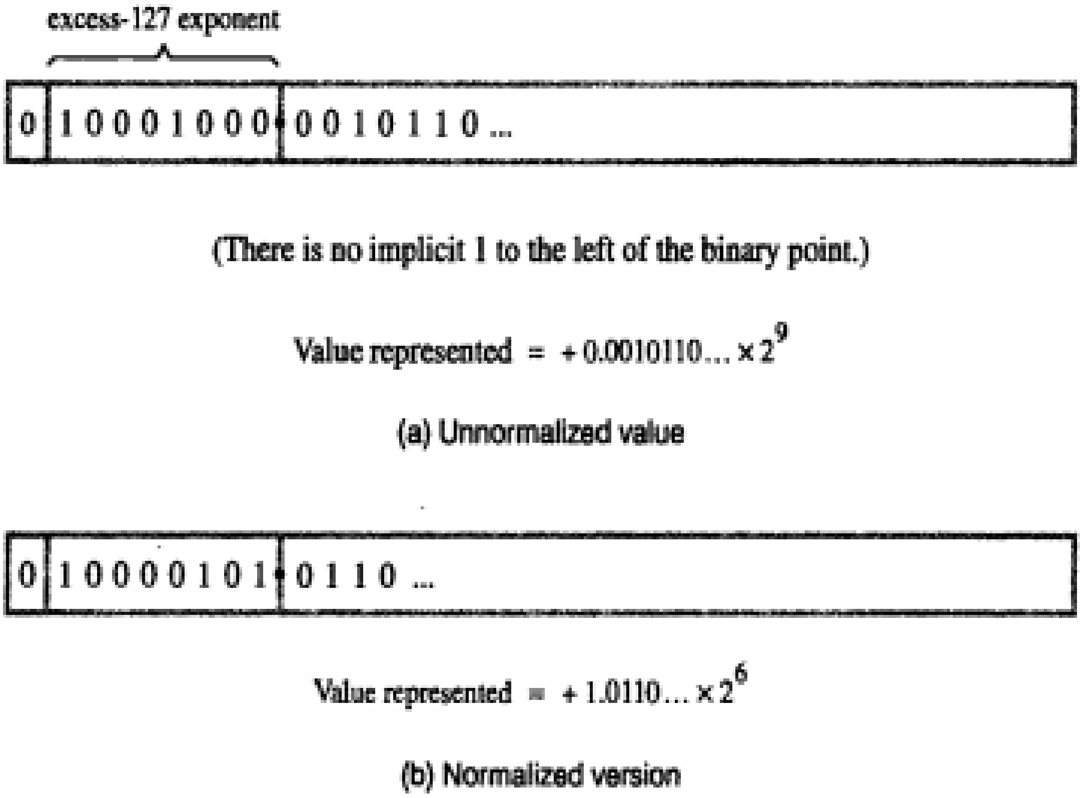


Value represented $= \pm 1.M \times 2^{E' - 1023}$

(c) Double precision

Figure 6.24 IEEE standard floating-point formats.

FLOATING-POINT NORMALIZATION



(There is no implicit 1 to the left of the binary point.)

Value represented = $+0.0010110\ldots \times 2^{9}$

(a) Unnormalized value

Value represented = $+1.0110\ldots \times 2^{6}$

(b) Normalized version

Figure 6.25 Floating-point normalization in IEEE single-precision format.

ARITHMETIC OPERATIONS ON FLOATING-POINT NUMBERS

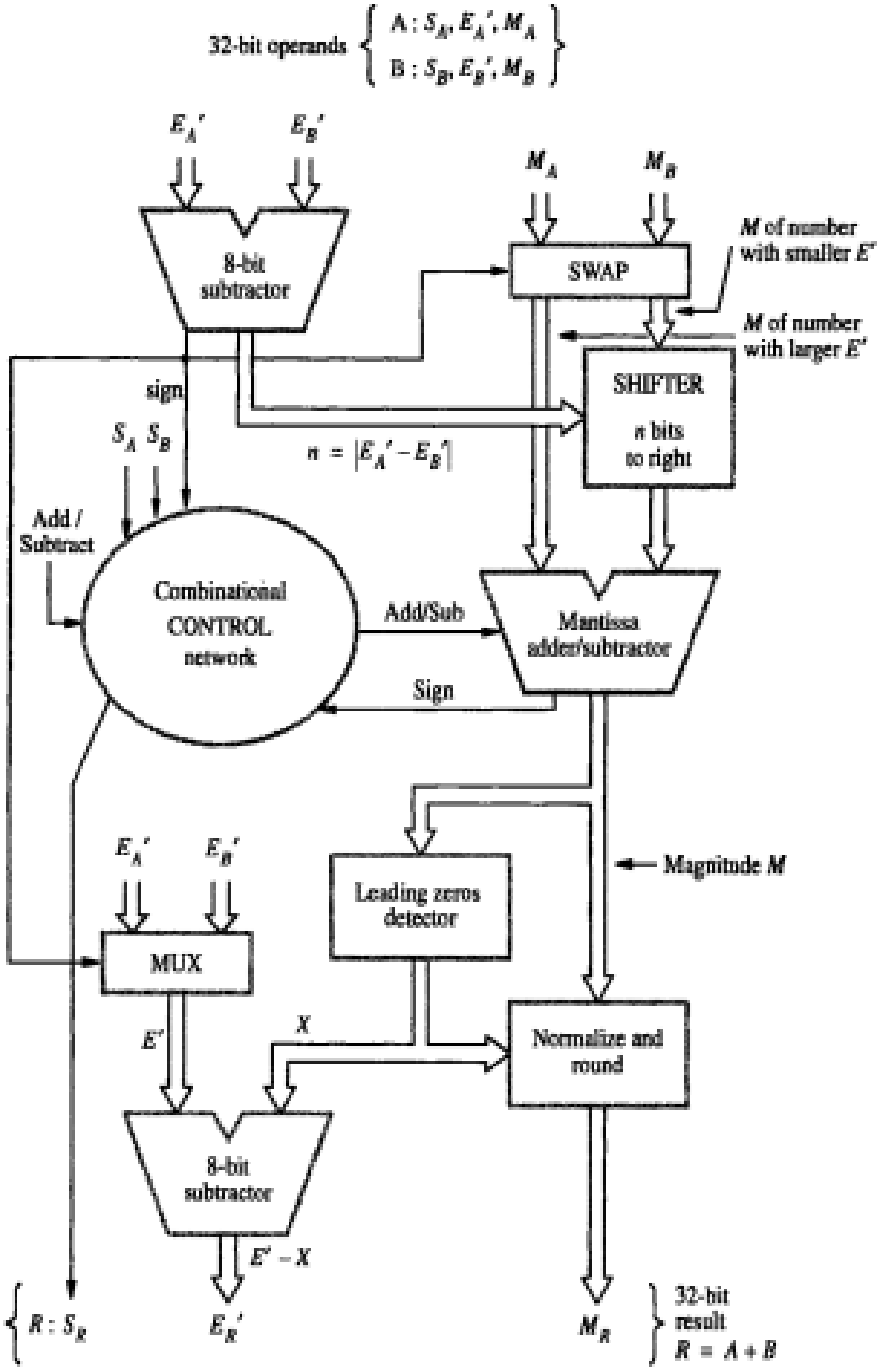


Figure 6.26 Floating-point addition-subtraction unit.

Add/Subtract Rule

1. Choose the number with the smaller exponent and shift its mantissa right a number of steps equal to the difference in exponents.
2. Set the exponent of the result equal to the larger exponent.
3. Perform addition/subtraction on the mantissas and determine the sign of the result.
4. Normalize the resulting value, if necessary.

Multiplication and division are somewhat easier than addition and subtraction, in that no alignment of mantissas is needed.

Multiply Rule

1. Add the exponents and subtract 127.
2. Multiply the mantissas and determine the sign of the result.
3. Normalize the resulting value, if necessary.

Divide Rule

1. Subtract the exponents and add 127.
2. Divide the mantissas and determine the sign of the result.
3. Normalize the resulting value, if necessary.

The addition or subtraction of 127 in the multiply and divide rules results from using the excess-127 notation for exponents.