Computer Architecture Floating Point Accuracy Notes (Section 3.5)

Extra Bits Used During Calculation (three bits total):

Two additional least-significant bits of fraction beyond the 23 (single precision) or 52 (double precision in the final result.

Guard bit: one position less significant than the least-significant bit of result

Round bit: two position less significant than the least-significant bit of result

One additional bit that is set if any 1's have been shifted right past the Round bit (by any amount)

Sticky bit

Example with 4 bits of fraction and no extra bits:

$1.\underline{1111} \text{ x } 2^{3} + 1.\underline{1110} \text{ x } 2^{-1} = 1111.1 + 0.1111 = 15.5 + 0.96875$

$= 16.46875$ (exact value)

$1.\underline{1110} \text{ x } 2^{-1} = 0.\underline{0001} \text{ x } 2^{3}$

$\text{Sum} = (1.\underline{1111} + 0.\underline{0001}) \text{ x } 2^{3} = 10.0000 \text{ x } 2^{3} = 1.\underline{0000} \text{ x } 2^{4} = 16$

Example with 4 bits of fraction and extra bits:

$1.\underline{1111} \text{ x } 2^{3} + 1.\underline{1110} \text{ x } 2^{-1} = 1111.1 + 0.1111 = 15.5 + 0.96875$

$= 16.46875$ (exact value)

$1.\underline{1110}\mathbf{00} \text{ x } 2^{-1} = 0.\underline{0001}\mathbf{11} \text{ x } 2^{3}$ Sticky = **1**

$\text{Sum} = (1.\underline{1111}\mathbf{00} + 0.\underline{0001}\mathbf{11}) \text{ x } 2^{3} = 10.0000\mathbf{11} \text{ x } 2^{3}$

$= 1.\underline{0000}\mathbf{01} \text{ x } 2^{4} = 16$ Sticky = **1**

In this case extra bits not used since $1.0000 \text{ x } 2^{4} = 16$ and $1.0001 \text{ x } 2^{4} = 17$

16 is as close to 16.46875 as possible

Rounding rules when true result is halfway between allowed values:

If true result is -4.5 and allowed values are -4 and -5

Round up (toward +∞):	use -4
Round down (toward -∞)	use -5
Truncate	use -4
Round to nearest even	use -4

If true results is -4.500001, always use -5