IEEE Rounding

Ø, 1, 2, 3, 4 round ↓

6, 7, 8, 9 round ↑

5 round to the nearest even digit

1.5 ⇒ 2.Ø

2.5 ⇒ 2.Ø

Remember, there are an infinite number of values between Ø.Ø → 1.Ø. We can only represent 2^{53} with double precision FP

Wednesday, February 28, 2024 2:10 PM

example

$2.56_{10} \times 10^{0} + 2.34_{10} \times 10^{2}$

We perfor

$$\begin{array}{r} 2.3400 \\ +\ 0.0256 \\ \hline 2.3456 \end{array}$$

So, 2.3656×10^{2}

We must now round to 2 significant digits to the right of the decimal.

00 ⇒ 49		round ↓
51 ⇒ 99		round ↑
50		to nearest even

In this case 2.37×10^{2}

Misc Trivia ?

IEEE 754 defines 32, 64, & 128 bit FP formats

C# has a decimal type that is 128 bits

C/C++ float is 32 bits
double is 64 bits

Check on 128 bits?

1/10 in binary is

0.0001100110011 0011 . . .

1/10 is rational in decimal, but irrational in binary.

Most monetary calculations are performed using binary coded

Most [illegible]
performed using binary coded decimals.

Wednesday, February 28, 2024 2:24 PM

When performing binary floating point operations, the hardware will (typically) provide two additional bits that are named guard and round

With 5 significant digits

$$\begin{array}{rl} & 0.1011 \times 2^{0} \\ + & 0.1111 \times 2^{-2} \end{array}$$

5 sig digits, guard, round

$$\begin{array}{rl} & \overbrace{0.1011}^{5\ \text{sig digits}} \\ + & 0.001111 \\ \hline & 0.1110\underbrace{11} \end{array}$$

round up since this is 3_{10}

0.1111×2^{0}

patterns
00 & 01 round ↓
& 10 & 11 round ↑