Floating point multiplication

$10\ 000\ 000$

$= 10 \times 10^6$

$= 1010_2 \times 10^6$

$\approx 1.19 \times 2^{23}$

$1.19_{10} = 01.0111111$

$= 0.10111111 e 0001$

$\approx 0.10111111 e (0001 + 010111)$

$\approx 0.10111111 e 011000$

$\approx 0.1 e 011000$ Q 0.1 e 6

$= 8388608$

111110

$-\ 000010$

111100

$y_1 \times y_2 = (M_1 \times M_2) e (E_1 + E_2)$

$0.10111 e 011000 \times 0.00101 e 111110$

Normalise:

$0.10111 e 011000 \times 0.10100 e 111100$

$= (0.10111 \times 0.10100) e (011000 + 111100)$

Multiply mantissas.

010111 Q 0.5

$\times\ 010100$ Q 0.5

000000

0000000

Σ $010111\ 00$

000000000

01011100000

0000000000000

001110011100 Q 0.10

$0.111001100 e 111111$

Add exponents:

011000 24_{10}

$+\ 111100$ -4_{10}

$[1]\ 010100$ $+20$

$+\ 111111$

$[1]\ 010011$ $+19_{10}$

Answer: $0.111001100 e 010011$

$= 0.11100 e 010011$ Q 0.5 e 6

Floating point division

$y_1 \div y_2 = (M_1 \div M_2)\, e\, (E_1 - E_2)$

1. Normalise
2. Unsign (if necessary)
3. Divide mantissas
4. Subtract exponents
5. Normalise
6. Re-sign (if necessary).

1.01101 e 0001 ÷ 0.001011 e 0101
= 1.01101 e 0001 ÷ 0.101100 e 0011 (normalised)
= −1 × (0.10011 e 0001 ÷ 0.101100 e 0011)
→ unsigned

101101
010010
+1
010011

Divide mantissas:

```
           000000 | 011011
0101100 ) 010011  | 000000
          010011  | 001
          -1011   | 00
          1000    | 1000
          -101    | 100
          0110    | 10000
          -1      | 01100
          1       | 001000
```

= 0.011011
= 0.11011 e 1111

Subtract exponents:

```
      0001
    - 0011
[1]   1110   ✓
+     1111   (from divided mantissas)
[1]   1101   ✓
```

$0.110110e1101$

Re-sign!

$z = 1.0010010e1101$

$$\begin{array}{r} 0110110 \\ 1001001 \\ + \quad\quad 1 \\ \hline 1001010 \end{array}$$

Calculate area of a circle with radius 0.3333 in R0.7e8

$\pi = 3.141592654$

$= 011.001001$

$= 0.11001001e00000010$

0.3333

$= 0.01010101 = 0.328125$ (1.6% error).

$= 0.10101001e11111111$

$$0.\dot{3}^2 = 01010100 \times 01010100$$

$$\begin{array}{r} 00000000 \\ 00000000 \\ 01010100 \\ 00000000 \\ 01010100 \\ 00000000 \\ 01010100 \\ 00000000 \\ \hline 0011011100010000 \end{array}$$

$0.0110111e00000000$

$= 0.1101110e11111111$

Add Exponents!

$11111111 + 11111111 + 11111101 = 11111101$

$0.1101110e11111101$ (0.3^2)

$\times\ 0.1100100e00000010$ (π)

$$
\begin{array}{r}
01101110 \\
\times\ 01100100 \\
\hline
00000000 \\
00000000 \\
01101110 \\
00000000 \\
00000000 \\
01101110 \\
01101110 \\
\hline
01010110\,11111000 \\
= 0.10101011e00000000
\end{array}
$$

Add exponents:

$$
\begin{array}{r}
11111101 \\
+\ 00000010 \\
\hline
11111111
\end{array}
$$

$\therefore$ Area $= 0.10101011e11111111$

What every programmer should know about floating point numbers

Guard bits

Method for preserving precision

Add another column after least significant bit.

a. Introduce at normalisation

b. Preserve until last formatting.