Floating Point Arithmetic

A = 40420000 = 0 100 0000 0 100 0010 0000 0000 0000 0000

Sign = 0 ∴ the number is positive; the exponent = 10000000 = 128 = 2^1
The fraction = 10000100000000000000000; the significand = 1.10000100000000000000000
Multiplying by 2^1 moves the binary point (bp) one position to the right giving 11.00001 = 3 + 1/32 = 3.03125

B = C0380000 = 1 100 0000 0 011 1000 0000 0000 0000 0000

Sign = 1 ∴ the number is negative; the exponent = 10000000 = 128 = 2^1
The fraction = 01110000000000000000000, the significand = -1.01110000000000000000000
Multiplying by 2^1 moves the binary point one position to the right giving -10.111 = 2 + 1/2 + 1/4 + 1/8 = 2.875

Addition.

Since the exponents are both 2^1 we can simply add 1.10000100000000000000000 and -1.01110000000000000000000 and multiply the result by 2^1

$$\begin{array}{r} 1.10000100000000000000000 \times 2^1 \\ -1.01110000000000000000000 \times 2^1 \\ \hline 0.00010100000000000000000 \times 2^1 \end{array}$$

we must now normalize the result so that it will have a leading 1. This requires moving the bp 4 places to the right. We must therefore subtract 4 from the exponent to give $1.01000000000000000 \times 2^{-3}$. The exponent of the floating point result is therefore 127 - 3 = 124. The sign is 0 and the fraction is 01000000000000000. The result is 0 01111100 01000000000000000 = 3E200000

Subtraction

Since the exponents are both 2^1 we can simply subtract -1.01110000000000000000000 from 1.10000100000000000000000 and multiply the result by 2^1

$$\begin{array}{r} 1.10000100000000000000000 \times 2^1 \\ --1.01110000000000000000000 \times 2^1 \\ \hline 10.11110100000000000000000 \times 2^1 \end{array}$$

we must now normalize the result so that it will have only one bit in front of the BP. This requires shifting the bp one position to the left and adding 1 to the exponent. The significand is therefore $1.01111010000000000000000 \times 2^2$. The exponent is therefore 127 + 2 = 129. The sign is 0 and the fraction is 01111010000000000000000 The result is 0 10000001 01111010000000000000000 = 40BD0000

Mu ltiplication

$$\begin{array}{r} 1.100001 \times 2^1 \\ 1.011100 \times 2^1 \\ \hline 110000100 \\ 1100001 \\ 1100001 \\ 11000010 \\ \hline 10001011011100 \end{array}$$

(Note there are 6 places after thebp in each number. There will be 12 places after the bp in the result)

$10001011011100 = 10.001011011100 \times 2^2$ (we add the exponents) = $1.0001011011100 \times 2^3$ Since the signs are opposite, the result is -., the exponent is 127 + 3 = 130 and the fraction is 0001011011100 The result is 1 10000010 0001011011100 = C10B7000

Division

In division we subtract the exponents. Since they are both 2^1 the resulting exponent will be 2^0. It is easiest to divide by a whole number rather than a fraction. Therefore move the bp of the divisor and dividend 4 places to the right. This gives

```
                1000011011110100110111101001

10111   11000010000000000000000000
        10111
        0000101000
         10111
        0100010
          10111
        000101100   ← Repeating fraction
           10111
          101010
           10111
          100110
           10111
          0011110
            10111
            0011100
              10111
              00101000
                 10111
                100010
                 10111
                001011   ← Repeating fraction
```

Since the signs are opposite, the sign of the result is -. The exponent is 2^0 which gives 127 as the biased exponent. The significand is 1.00001101111010011011110100 11. the fraction (to 23 places) is 00001101111010011011110 The truncated result is therefore 1 01111111 00001101111010011011110 = BF86F4DE, however the 24th bit is a 1 which would give a rounded result of BF86F4DF