Boris Topalov

bnt4yb

2/9/2020

Filename: floatingpoint.pdf

Base 10 to hex: 10 = a, 11 = b, 12 = c, 13 = d, 14 = e, 15 = f

My floating point number was 17.625

17.625 floating point to hexadecimal:

1. 17 = 10001 in binary
2. 0.625 in binary =
    a. 0.625 * 2 = 1.25 so 1 bit
    b. 0.25 * 2 = 0.5 so 0 bits
    c. 0.5 * 2 = 1.0 so 1 bit
        i. So 0.625 in binary = 0.101
3. 17.625 in binary = 10001.101 * 2^0 = 1.0001101 * 2^4
4. Exponent: 4 + 127 = 131
    a. 131 in binary = 10000011
5. So we have 1.0001101 and the exponent is 10000011
    a. Sign is 0 since number is positive
6. So 17.625 in binary = 0 11000001 0001101 0000000000000000
    a. 0100 0001 1000 1101 0000 0000 0000 0000
        i. 0x418d in big-endian
        ii. 0x8d41 in little-endian

0x00809ec2 in hex (0x809ec2) to 32 bit floating point number:

1. Convert to big-endian: 0xc29e80
2. Convert to binary: 1100 0010 1001 1110 1000 0000 0000 0000
3. Sign is 1 so number will be negative, exponent is 1000 0101 which is 128 + 4 + 1 = 133
    a. 133 – 127 = 6 so exponent is 6
4. Mantissa is 001 1110 1000 0000 0000 0000
    a. So we have an 8^{th}, 16^{th}, 32^{nd}, 64^{th}, and 256^{th} = 0.23828125 + 1 = 1.23828125
5. Multiply mantissa by 2^exponent = 1.23828125 * 2^6 = 79.25
6. Sign is negative so number is **-79.25**