Assignment 1

Yaswanth Naidu - EE18BTECH11024

Download all C codes from

https://github.com/Yaswanthkarri999/assignment1/blob/main/assignment1.c

and latex-tikz codes from

https://github.com/Yaswanthkarri999/assignment1/blob/main/assignment1.tex

1 Problem

Consider three registers R1, R2, R3 that store numbers in IEEE-754 single precision floating point format. Assume that R1, R2 contain the values (in hexadecimal notation). 0x42200000, 0xc1200000. If R3 = (R1/R2) what is the value stored in R3.

```
#include<stdio.h>
#include<stdlib.h>
#include<string.h>
#include<math.h>
double number(char str[]){
    char S[1];
    strncpy(S, str + 0, 0);
    char E[8];
    strncpy(E, str + 1, 8);
    char M[23];
    strncpy(M, str + 9, 22);
    int frac = 0;
    for(int i = 0;i<strlen(E);i++)
    {
        if(E[i] == '1')
        {
            frac = frac*2 + 1;
        }
        else
        {
            frac = frac*2 + 0;
        }
    }

    double expnt = 0;
    for(int i = 0;i<strlen(M);i++)
    {
        if(M[i] == '1')
        {
            expnt += 1*(pow(2,-i-1));
        }
    }
    int sign = 1;
    if(str[0] == '1'){
        sign = -1;
    }
    double exp = 1+expnt;
    double ten = pow(2,frac-127);
    double ans = sign*exp*ten;
        return ans;

}
int main(){
    int a = 0x42200000;
    char b[32];
    itoa(a,b,2);
    char op[] = "0";
    strcat(op,b);
    double first = number(op);

    int a_1 = 0xC1200000;
    char b_1[32];
    itoa(a_1,b_1,2);
    double second = number(b_1);
    printf("%lf",first/second);

}
```

2 Explanation

In IEEE-754 single precision format,a floating point number is represented in 32 bits.

1.Sign bit (MSB) - 1 bit
2.Biased exponent - 8 bit
3.Normalized mantissa - 23 bit

Sign bit value 0 means positive number and 1 means negative number.

The floating point number can be obtained by formula : $\pm 1.M * 2^{(E-127)}$
Given R1 as 0x 42200000
R2 as 0x c1200000

2.1 Calculation

Content of R1 in HEX (0x) is 42200000. After converting into binary it can be represented in IEEE-754 format as

0	100 0010 0	010 0000 0000 0000 0000 0000

Sign bit is 0, i.e, the number is positive

Biased exponent (E') = 100 0010 0 = 132

Normalized Mantissa (M) = 010 0000 0000 0000 0000 0000 = 0.25

Therefore the number in register R1 = +1.25 * 2^(132 – 127) = 1.25 * 32 = 40

Content of R2 in Hex (0x) is c1200000. After converting into binary it can be represented in IEEE-754 format as

1	100 0001 0	010 0000 0000 0000 0000 0000

Sign bit is 1. i.e, the number is negative

Biased exponent (E') = 100 0001 0 = 130

Normalized Mantissa (M) = 010 0000 0000 0000 0000 0000 = 0.25

Therefore the number in register R2 = -1.25 * 2^(130 – 127) = -1.25 * 8 = -10

R3 = (R1 / R2) = 40 / (-10) = -4

Since the number is negative, sign bit (MSB) = 1

Converting 4 into binary of a floating point gives : (100.0) 2

Representing it into normalized form gives (1.00000...) * 2**2
Therefore, mantissa is 23 bits of all 0s

Biased exponent (E') = E + 127 = 2 + 127 = 129 = (10000001) 2

It can be represented in IEEE-754 format as :

1	100 0000 1	000 0000 0000 0000 0000 0000

Converting it into Hexadecimal format gives : 0x C0800000