Homework3-solution

1 . Suppose all the numbers are 8-bit long. Give the result of the following expressions of C language in full 8-bit 2's complement form and signed decimal form.

Expression	2's-complement	Signed 8-bit Decimal
3+2	0000 0101	5
99/2	0011 0001	49
(-23)/2	1111 0101	-11
-127-1	1000 0000	-128
125+3	1000 0000	-128
125>>3	0000 1111	15

2. The following C code pieces are executed on a typical 32-bit machine with 2's complement encoding. Please give the output and show how you can get the result in detail.

```
int main()
{
int x = 257;
char y = -10;
int z = 128;
char a = (char)x;
short b =(short)y;
unsigned short d = (unsigned short)b;
char c = (char)z;
unsigned int e = (c > 0) ? 0 : 1;
int f = ((unsigned) z<<24)>>24;
Int g =(z<<24)>>24;
```

```
printf("a=%d,b=%d,d=%x,c=%d,e=%d,f=%d,g=%d\n",a,b,d,c,e,f,g);
}
```

a=1, b=-10, d=fff6, c=-128, e=1, f=128, g=-128

3.

Consider a **16-bit** floating-point representation based on the IEEE floating-point format, which is illustrated below.

S	E	E	E	E	E	E	F	F	F	F	F	F	F	F	F

1. Filling the blanks with proper values.

   1) The denormalized values can be represented in a form $V = (-1)^{\textbf{sign}} * (0.\textbf{fraction}) * 2^{E}$, where **E** = ____**[1]**____;

   2) The normalized values can be represented in a form $V = (-1)^{\textbf{sign}} * (1.\textbf{fraction}) * 2^{(e - \textbf{bias})}$, where **bias** = ____**[2]**____, and the value of e ranges from ____**[3]**____ to ____**[4]**____.

2. Give the equivalent value of the following numeric numbers or FP representation.

Numeric value	FP representation (in hex)
$(12.625)_{10}$	$(0x$ **[1]** $)_{16}$
$(-0.09375)_{10}$	$(0x$ **[2]** $)_{16}$
[3]	$(0x4C18)_{16}$
[4]	$(0x7EB0)_{16}$

3. Calculate both the **sum** and **multiplication** of $(12.625)_{10}$ and $(0x4C18)_{16}$, and then round the results to **5** bits to the right of the binary point with **Round-to-Even** rounding modes. Give your steps detailed.

1. (1) $E = -30$

   $bias = 2^k - 1$ $k = 6$ $E = 1 - bias$

   (2) $bias = 31$ (3) $e_{min} = 1$ (4) $e_{max} = 62$

   $bias = 2^k - 1$ $k = 6$ $E = e - bias$

eisneitherallzeros(000000B)norallones(111111B),soerangesfrom 000001B(1D) to 111110B(62D)

2. (1) 4528

   12.625D 1100.101B

   F = 100101000B

   E= 3D + 31D = 34D = 100010B

   0100010100101000B

   (2) $B700$

   -0.09375D -0.00011B

   F= 100000000

   E = -4D+31D = 27D = 011011B

   1011011100000000

(3) 134

0x4C18 $0100110000011000B$

E = $38e$ = $38 - 31 = 7$

$1000011000B1.046875B$

1000011000B1.046875B

1.046875B $*$ 27 = 134

(4) NaN

E = 111111 and F is not all zeros

3. Sum:

$12.625D$+0x4C18 = $146.625D$

$10010010.101000000B$

$1.0010010101000000B * 2^6$ after rounding: $1.00101B * 2^7$

$E = 7D + 31D = 38D100101B$

$F = 00101B$

$0100101001010000B$, Sum : $0x4C50$

Mul:

$12.625D$*0x4C18 = $1691.75D$

$11010011011.110000000B$

$1.101001101111B * 2^{10}$, after rounding: $1.10101B * 2^{10}$

$E = 10D + 31D = 41D101001B$

$F = 10101B$

$0101001101010000B$, Mul : $0x5350$