HomeWork_5

Name：卫焱滨（Wei Yanbin）

SID：11710823

chapter 3

Exercise 3.27

Reapresent -1.5625 × 10^-1 as binary :

-1.5625 ×10^1 = -0.15625 × 10^0 = -0.00101 × 2^0

Shaft 3 bits to normalization :

-0.00101 × 2^0 = 1.01 × 2^-3

Exponenet : add the bias -3+15 = 12. Fraction：-0.0100000000

By the format, we get the answer:

1011000100000000

Comparision of Range:

Because 1.11111.....(Binary) is near to 2, treat them as 2.0 here.

Range of the 16 bit format :

-2.0×2^15~ -1.0×2^-14 and 1.0×2^-14~2.0×2^15，+∞，-∞，NaN

Range of single presicion IEEE 754：

-2.0×2^127~ -1.0×2^-126 and 1.0×2^-126~2.0×2^127, +∞，-∞， NaN

Comparision of accuracy(Here discuss Relative precision and ulp):

Relative precision of the 16 bit format :

$\Delta A/|A|=2^{-10}\times 2^{exponent}/|1\times 2^{exponent}|=2^{-10}$

ulp of the 16 bit format： **one-half ulp**

Relative precision of single presicion IEEE 754 :

$\Delta A/|A|=2^{-23}\times 2^{exponent}/|1\times 2^{exponent}|=2^{-23}$

ulp of the single presicion IEEE 754 : **one-half ulp**

Exercise 3.29

Step 1 —To Binary normalization form :

```
2.6125 × 10^1 = 26.125 = 11010.001 = 1.1010001000 × 2^4

4.150390625 × 10^-1 = 0.4150390625 =  0.011010100111 = 1.1010100111 × 2^-2
```

1. Align binary points

   Shift number with smaller exponent 6 bit to align

```
1.1010100111 × 2^-2 = 0.0000011010100111 × 2^4
```

2. Add significands

```
1.1010001000 × 2^4 + 0.0000011010100111 × 2^4 = 1.1010100010100111× 2^4
```

3. Normalize result & check for over/underflow

```
1.1010100010(10100111)× 2^4    No overflow.
```

4. Round and renormalize if necessary

```
 using GRS round to the nearest even :
 guard = 1, round = 0, sticky = 1
 Because round to the nearest even, guard=1 and round=0, the last bit of significant bit is 0 :
 So The nearest even become  1.1010100010 × 2^4
 There is No need to renormalize.
```

The answer is **1.1010100010** × **2^4**

Exercise 3.30

Express The Result as :

```
-8.0546875 × -1.79931640625 × 10^-1
```

Express every operacand as normalized binary form :

```
-8.0546875 = -1.0000000111 × 2^3
-1.79931640625 × 10^-1 = -1.0111000010 × 2^-3
```

Do the multiplication by hand:

```
Exponent:  2^3 × 2^-3 = 2^0

Fraction:  1.0000000111
        × 1.0111000010
        ----------------
           00000000000
          10000000111
         00000000000
        00000000000
       00000000000
      00000000000
     10000000111
    10000000111
   10000000111
  00000000000
 10000000111
 1.01110011000001001110
```

Round to the nearest even(by G.R.S.) and renormalization:

```
1.0111001100(0001001110)
   guard:0 round: 0 sticky:1
By round to the nearest even: need Truncate
So The answer is 1.0111001100, and no need to renormalize.
```

Express as 16-bit form described in 3.27 and also as a decimal number.

```
1.0111001100 × 2^0 = 0100000111001100 (1.0111001100 = 1.44921875)
```

Above is the result computed by hand.

Following is the result by using calculator to compute the product is:

```
-8.0546875 × -1.79931640625 × 10^-1 = 1.449293137
```

Accuracy and compare to the number by a calculator is as follows:

```
Accuracy:
  one-half ulp.
Compare to the result ogf calculator:
  Some information was lost because the result did not fit into the available 10-bit field: Answer only off by 0.00007438659667.
```