ESE_2014_WEEK11

Submitted to- Takis Zourntos

Submitted by- Amonjot Singh Chhina(C0772326)

1) Assume a 16-bit word, with an 8-bit fraction, i.e., M=8. Provide the fixed-point representations for the following numbers. In each case, what is the error associated with the representation?

a) 3.14159

b) 0.2378

c) 5.125

d) 125.32

ANSWER==

a) x_norm= 3.14159

x= 3.14159*(1/ 2^2)(2^8)=201.0617

It would be represented as 201

Now, reversing

x_actual = 201 * 2^3 / 2^8 = 3.140625

error = x_norm - x_actual = 314159 - 3.140625 = 0.000965

b) x_norm= 0.2378

x= 0.2378*(1/ 2^0)(2^8)=60.8768

It would be represented as 60

Now, reversing

x_actual = 60* 2^0 / 2^8 = 0.234375

error = x_norm - x_actual = 0.2378 - 0.234375 = 0.003425

c) x_norm= 5.125

x= 5.125*(1/ 2^3)(2^8)=164

It would be represented as 164

Now, reversing

x_actual = 164 * 2^3 / 2^8 = 5.125

error = x_norm - x_actual = 5.125-5.125 = 0

d) x_norm= 125.32

x= 125.32*(1/ 2^7)(2^8)=250.64

It would be represented as 250

Now, reversing

x_actual = 250 * 2^7 / 2^8 = 125

error = x_norm - x_actual = 125.32 - 125 = 0.32

Ques2. repeat the above, but use a 10-bit fraction, i.e., M=10.

a) x_norm= 3.14159

x= 3.14159*(1/ 2^2)(2^10)=804.24704

It would be represented as 804

Now, reversing

x_actual = 804 * 2^2 / 2^10 = 3.140625

error = x_norm - x_actual = 314159 - 3.140625 = 0.000965

b) x_norm= 0.2378

x= 0.2378*(1/ 2^0)(2^10)= 243.5072

It would be represented as 243

Now, reversing

x_actual = 243 * 2^0 / 2^10 = 0.2373047

error = x_norm - x_actual = 0.2378 - 0.2343047 = 0.0004953

c) x_norm= 5.125

x= 5.125*(1/ 2^4)(2^10) = 328

It would be represented as 328

Now, reversing

x_actual = 328 * 2^4 / 2^10 = 5.125

error = x_norm - x_actual = 5.125-5.125 = 0

d) x_norm= 125.32

x= 125.32*(1/ 2^7)(2^10) = 1002.56

It would be represented as 1002

Now, reversing

x_actual = 1002 * 2^7 / 2^10 = 125.25

error = x_norm - x_actual = 125.32 - 125.25 = 0.07

3) repeat the above, but assume a 32-bit word and 16-bit fraction. How do the errors compare with the 16-bit, M=8, case?

a) x_norm= 3.14159

x= 3.14159*(1/ 2^2)(2^16)= 51471.810

It would be represented as 51471

Now, reversing

x_actual = 51471 * 2^2 / 2^16 = 3.1415405

error = x_norm - x_actual = 314159 - 3.1415405 = .0000495

error with m = 8 is 0.000965. Error with fraction m = 10 is (0.0009005) less than the error with fraction m = 8.

b) x_norm= 0.2378

x= 0.2378*(1/ 2^0)(2^16)= 15584.4608

It would be represented as 155854

Now, reversing

x_actual = 15584 * 2^0 / 2^16 = 0.23779

error = x_norm - x_actual = 0.2378 - 0.2779 = 0.00001

error with m = 8 is 0.003425. Error with fraction m = 10 is (0.000865) less than the error with fraction m = 8.

c) x_norm= 5.125

x= 5.125*(1/ 2^4)(2^16) = 41984

It would be represented as 41984

Now, reversing

x_actual = 328 * 2^4 / 2^16 = 5.125

error = x_norm - x_actual = 5.125-5.125 = 0

error with m = 8 is 0. Error with fraction m = 10 is (0) same as the error with fraction m = 8.

d) x_norm= 125.32

x= 125.32*(1/ 2^7)(2^16) = 64163.84

It would be represented as 64163

Now, reversing

x_actual = 64163 * 2^7 / 2^16 = 125.3183

error = x_norm - x_actual = 125.32 - 125.3183 = 0.0017

error with m = 8 is 0.32. Error with fraction m = 10 is (0.3183) less than the error with fraction m = 8.

It concludes that greater the number of bits lesser is the error