

# Homework 2

(Due date: February 6<sup>th</sup> @ 7:30 pm)

Presentation and clarity are very important! Show your procedure!

## PROBLEM 1 (15 PTS)

- Multiply the following signed fixed-point numbers (6 pts):

01.101 × 1.101001	100.001 × 01.10001	110.000 × 10.10101
----------------------	-----------------------	-----------------------

- Get the division result (with  $x = 4$  fractional bits) for the following signed fixed-point numbers:

101.0101 ÷ 1.101	10.0101 ÷ 01.11	1.1011 ÷ 1.01101
---------------------	--------------------	---------------------

## PROBLEM 2 (11 PTS)

- We want to represent numbers between  $-512$  and  $511.9997$ . What is the fixed point format that requires the fewest number of bits for a resolution better or equal than  $0.0005$ ? (4 pts).
- We want to represent numbers between  $-127.05$  and  $116.25$ . What is the fixed point format that requires the fewest number of bits for a resolution better or equal than  $0.0015$ ? (4 pts).
- Represent these numbers in Fixed Point Arithmetic (signed numbers). Select the minimum number of bits in each case.

-129.625	-69.1875	113.3125
----------	----------	----------

## PROBLEM 3 (10 PTS)

- Complete the table for the following fixed point formats (signed numbers): (4 pts)

Fractional bits	Integer Bits	FX Format	Range	Dynamic Range (dB)	Resolution
9	3				
11	5				
15	9				

- Complete the table for these floating point formats (which resemble the IEEE-754 standard). Only consider ordinary numbers.

Exponent bits (E)	Significant bits (p)	Min	Max	Range of e	Range of significand
8	6				
10	13				
15	32				

## PROBLEM 4 (20 PTS)

- Calculate the decimal values of the following floating point numbers represented as hexadecimals. Show your procedure.

Single (32 bits)		Double (64 bits)	
✓ 90DBD800	✓ 7F85B0AC	✓ DECAF0FFEE80000	✓ ACCEDE90BEAD5000
✓ 800BEEF0	✓ 70DECADE	✓ 49A5DEAF8FAD8000	✓ 8009BEBEFACE8000

## PROBLEM 5 (44 PTS)

- Calculate the result (provide the 32-bit result) of the following operations with 32-bit floating point numbers. Truncate the results when required. When doing fixed-point division, use 8 fractional bits. Show your procedure.

✓ 3DE38C80 + 3A80D980	✓ 80A18000 - 83CEC000	✓ 7A09D300 × 4D080000	✓ 800C0000 ÷ 494C0000
✓ 80123000 + 804E8000	✓ 09DECAF0 - 7AD90000	✓ 90DECADE × FF800000	✓ 7F800000 ÷ 800ABBAA
✓ 7FEEFCA0 + FACADE90	✓ F0B1ABEE - 7F800000	✓ 0B09A000 × 8FACC000	✓ C9746000 ÷ 40490000