# Homework 2

(Due date: January 31st @ 7:30 pm)
Presentation and clarity are very important! Show your procedure!

## PROBLEM 1 (12 PTS)

- Calculate the result of the additions and subtractions for the following fixed-point numbers.

| **UNSIGNED** | | **SIGNED** | |
|---|---|---|---|
| 0.11010 + 1.0101101 | 1.00111 - 0.0000111 | 1.0001 + 1.001001 | 0.0101 - 1.0101101 |
| 10.10101 + 1.1001 | 100.1 + 0.10101 | 1000.0101 - 11.010101 | 101.0101 + 1.0111101 |

## PROBLEM 2 (15 PTS)

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

| | | |
|---|---|---|
| 01.001 × 1.001001 | 10.0001 × 01.01001 | 1.11010 × 110.11011 |

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

| | | |
|---|---|---|
| 101.1001 ÷ 1.011 | 11.011 ÷ 1.01011 | 10.0110 ÷ 01.01 |

## PROBLEM 3 (11 PTS)

- We want to represent numbers between $-128.87$ and $127.12$. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than $0.0015$? (4 pts).
- We want to represent numbers between $-255.12$ and $256.91$. What is the fixed point format that requires the fewest number of bits for a resolution better or equal than $0.0025$? (4 pts).
- Represent these numbers in Fixed Point Arithmetic (signed numbers). Select the minimum number of bits in each case.

| | | |
|---|---|---|
| -128.1875 | -78.125 | 107.3125 |

## PROBLEM 4 (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** |
|---|---|---|---|---|---|
| 8 | 4 | | | | |
| 10 | 6 | | | | |
| 16 | 8 | | | | |

- 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 | 7 | | | | |
| 10 | 13 | | | | |
| 12 | 35 | | | | |

## PROBLEM 5 (20 PTS)

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

| **Single (32 bits)** | | **Double (64 bits)** | |
|---|---|---|---|
| ✓ E8000978 | ✓ 800BCCAA | ✓ 7FFDECADEFEEBEE9 | ✓ 8009BEBEFACE8000 |
| ✓ 80DE0FEE | ✓ 7FFCAFEA | ✓ 49A5DEAF8FAD8000 | ✓ 70800FEDCAB09000 |

## PROBLEM 6 (32 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.

| | | | |
|---|---|---|---|
| ✓ 40C00000 + C2EA9000 | ✓ 5A09D378 - 4C490FD8 | ✓ 7A09C000 × 8BEE0000 | ✓ C9680000 ÷ 80700000 |
| ✓ 801A8000 + 92CE8000 | ✓ 10DAD000 - 90FAD000 | ✓ FA19D800 × CD100000 | ✓ 7A390000 ÷ C8400000 |