CSCI 516 - Fundamental Concepts in Computing and Machine Organization
Homework Assignment 4 Solutions

Perform each of the following computation using IEEE-754 single precision and IEEE-754 double precision representation. Clearly show all the steps.

[Method: Convert each of the decimal values to IEEE-754 single precision representation. Perform IEEE-754 computation (addition or multiplication) and convert the IEEE-754 single precision result back to decimal value. Repeat the above method for each computation using IEEE-754 double precision representation].

a. -2.25 + 15 =
$-2.25_{ten} = -9/4_{ten} = -1001 * 2^{-2}_{two} = -10.01_{two} = -1.001_{two} * 2^1$
$15_{ten} = 1111_{two} = 1.111_{two} * 2^3$

IEEE-754 single precision representation:
1. Align binary points, shift the fraction of the number with smaller exponent.
$-2.25_{ten} = -1.001_{two} * 2^1$
$= 1100\ 0000\ 0001\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
$= 1100\ 0001\ 0010\ 0100\ 0000\ 0000\ 0000\ 0000_{two}$ (shifting right)
$15_{ten} = 0100\ 0001\ 0111\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
2. Add significands (Subtract 15 from 2.25 here)
$sum = 0100\ 0001\ 0100\ 1100\ 0000\ 0000\ 0000\ 0000_{two}$
3. Normalize result and check for over/underflow
$sum = 0100\ 0001\ 0100\ 1100\ 0000\ 0000\ 0000\ 0000_{two}$ result no changed
4. Round and renormalize if necessary
$sum = 0100\ 0001\ 0100\ 1100\ 0000\ 0000\ 0000\ 0000_{two}$ result no changed

$$\begin{aligned} sum &= 0100\ 0001\ 0100\ 1100\ 0000\ 0000\ 0000\ 0000_{two} \\ &= (-1)^0 * (1 + 2^{-1} + 2^{-4} + 2^{-5}) * 2^{(130-127)} \\ &= 12.75 \end{aligned}$$

IEEE-754 double precision representation:
1. Align binary points, shift the fraction of the number with smaller exponent.
$-2.25_{ten} = 1100\ 0000\ 0000\ 0010\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
$= 1100\ 0000\ 0010\ 0100\ 1000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$(shifting right)
$15_{ten} = 0100\ 0000\ 0010\ 1110\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
2. Add significands (Subtract 15 from 2.25 here)
$sum = 0100\ 0000\ 0010\ 1001\ 1000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
3. Normalize result and check for over/underflow
$sum = 0100\ 0000\ 0010\ 1001\ 1000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
4. Round and renormalize if necessary
$sum = 0100\ 0000\ 0010\ 1001\ 1000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$

$$\begin{aligned} sum &= 0100\ 0000\ 0010\ 1001\ 1000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two} \\ &= (-1)^0 * (1 + 2^{-1} + 2^{-4} + 2^{-5}) * 2^{(1026-1023)} \\ &= 12.75 \end{aligned}$$

b. -7.5 * 3.25 =
$-7.5 = -15/2 = -1111/2^1_{two} = -111.1_{two} = -1.111 * 2^2_{two}$
$3.25 = 13/4 = 1101/2^2_{two} = 11.01_{two} = 1.101 * 2^1_{two}$
The product of significands: $1.111 * 1.101 = 11.000011 = 1.1000011 * 2^1$
The sign of the product is: -1.

IEEE-754 single precision representation:
$-7.5 = 1100\ 0000\ 1111\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
$3.25 = 1100\ 0000\ 0101\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
1. add components
The sum of exponents is: 2 + 1 = 3
2. Multiply significands(The bit in red is the 1.0 in fraction)
$1|111000...000 * 1|101000...000 = 11|000011000...000 = 1|100011000...000 * 2^1$
3. Normalize result and check for over/underflow
$sum = 1100\ 0001\ 1100\ 0011\ 0000\ 0000\ 0000\ 0000_{two}$ new exponent = 130 + 1 = 131
4. Round and renormalize if necessary
$sum = 1100\ 0001\ 1100\ 0011\ 0000\ 0000\ 0000\ 0000_{two}$
5. Determine Sign: -ve * +ve = -ve

$$\begin{aligned} sum &= 1100\ 0001\ 1100\ 0011\ 0000\ 0000\ 0000\ 0000_{two} \\ &= (-1)^1 * (1 + 2^{-1} + 2^{-6} + 2^{-7}) * 2^{(131-127)} \\ &= -24.375 \end{aligned}$$

IEEE-754 double precision representation:
$-7.5_{ten} = 1100\ 0000\ 0001\ 1110\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
$15_{ten} = 0100\ 0000\ 0000\ 1010\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
1. add components
The sum of exponents is: 2 + 1 = 3
2. Multiply significands(The bit in red is the 1.0 in fraction)
$1|111000...000 * 1|101000...000 = 11|000011000...000 = 1|100011000...000 * 2^1$
3. Normalize result and check for over/underflow
$sum = 11100\ 0000\ 0011\ 1000\ 0110\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$ new exponent = 1026 + 1 = 1027
4. Round and renormalize if necessary
$sum = 11100\ 0000\ 0011\ 1000\ 0110\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two}$
5. Determine Sign: -ve * +ve = -ve

$$\begin{aligned} sum &= 1100\ 0000\ 0011\ 1000\ 0110\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000_{two} \\ &= (-1)^0 * (1 + 2^{-1} + 2^{-6} + 2^{-7}) * 2^{(1027-1023)} \\ &= -24.375 \end{aligned}$$