BRADLEY University

CS 220 Computer Architecture

HW 04 - Floating-Point Numbers

Fall 2023

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION SYSTEMS

PART 0: READING

- **Reading:** **Chapter 11** - Computer Arithmetic | **11.4 - 11.5**

PART 1: FLOATING-POINT NUMBER REPRESENTATION [24 PTS]

QUESTION 1

List four alternative methods of rounding the result of a floating-point operation and major benefits and drawbacks. [4 pts]

- Round to nearest:
  - The representable value nearest to the infinitely precise result shall be delivered
  - Benefit: simple method of rounding
  - Drawback: introduces small cumulative bias
- Round towards pos infinity:
- Round toward neg infinity:
  - Benefits: Interval arithmetic using both results at once provides a more accurate average using bounds
  - Drawbacks: Carries sign of divisor
- Round toward 0:
  - Benefits: Ignores truncation, but produces a term less or equal to the more precise original value
  - Drawbacks: Carries sign of dividend

QUESTION 2

Describe the following terms in the IEEE standard for binary floating-point arithmetic. [4 pts]

a. Infinity - The infinity arithmetic is treated as the limiting case of real arithmetic

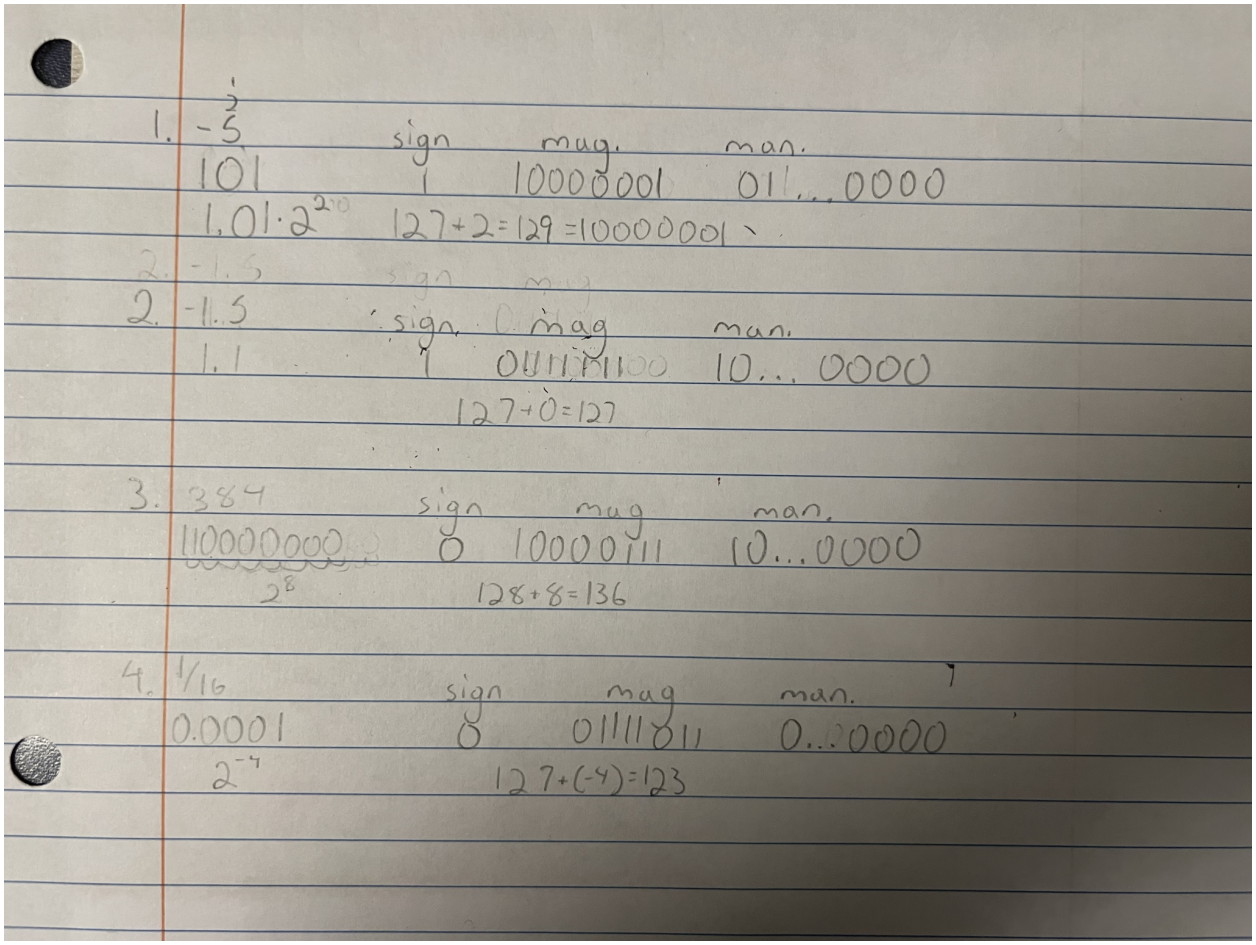
b. NaN - a symbolic floating-point representation which is neither a signed infinity nor a finite number

QUESTION 3

Represent the following numbers in IEEE 32-bit floating-point format. [16 pts]

a. -5
    a. 11000000101000000000000000000000
b. -1.5
    a. 10111111110000000000000000000000
c. 384
    a. 01000011110000000000000000000000
d. 1/16
    a. 00111101100000000000000000000000

Show ALL steps.



PART 2: FLOATING-POINT NUMBER ARITHMETIC [76 PTS]

QUESTION 1

Convert the IEEE 32-bit floating-point format to decimal and hexadecimal values. [6 pts]

Binary **# of bolded bits: 23**	Decimal	Hexadecimal
1 – 1000 0011 – **1100 … 0000**	-28	0xc1e00000
0 – 0111 1110 – **1010 … 0000**	0.8125	0x3f500000
0 – 1000 0000 – **1010 … 0000**	3.25	0x40500000

QUESTION 2

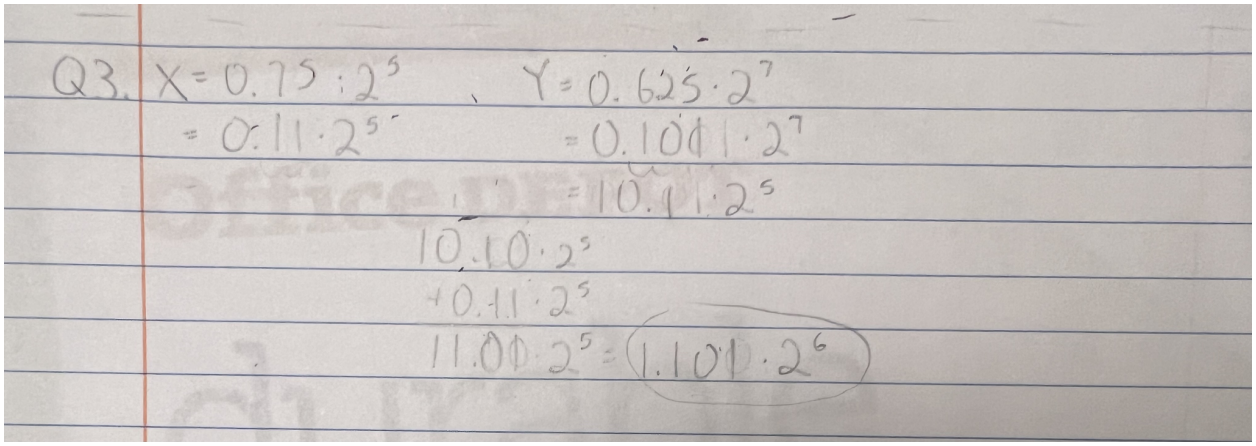
Compare and contrast IEEE 754 formats for Binary32, Binary64, and Binary128 in the following table. [15 pts]

	# of the sign bit	# of bits Biased Exponent	# of bits Trailing significant	Min value	MAX value
Binary32	1	8	23	10^-38	10^38
Binary64	1	11	52	10^-308	10^308
Binary128	1	15	112	10^-4932	10^4932

QUESTION 3

Given **X = 0.75 * 2^5** and **Y = 0.625 * 2^7**, perform Z = X + Y using Normalized form. [10 pts]

Show ALL steps.



1.101*2^6

QUESTION 4

Given **X = 0.75 * 2^5** and **Y = 0.625 * 2^7**, find the **IEEE 754 32-bit format** of X, and Y and specify its corresponding components in the tables. [5 pts]

Show ALL steps.

Normalized X:	1.1 * 2^(4)	
Exponent	Decimal form: 4 Biased + 127 ==> 131	Binary form: 100 1000-0011
Significand	Decimal form: 1.5 Due to shifting	Binary form: 1. 1 1 followed by 22 0s
Normalized Y:	1.01*2^6	
Exponent	Decimal form: 6 Biased + 127 -> 133	Binary form: 10000101
Significand	Decimal form: 1.25 Due to shifting	Binary form: 1. 01 01 Followed by 21 zeros

IEEE 754 32-bit format

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31

X	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: X - Blank slots from indexed 10 - 31 hold a digit of 0

QUESTION 5

Following the above question (4), perform Z = X + Y using **Floating-Point** Addition and Subtraction as discussed in class. [10 pts]

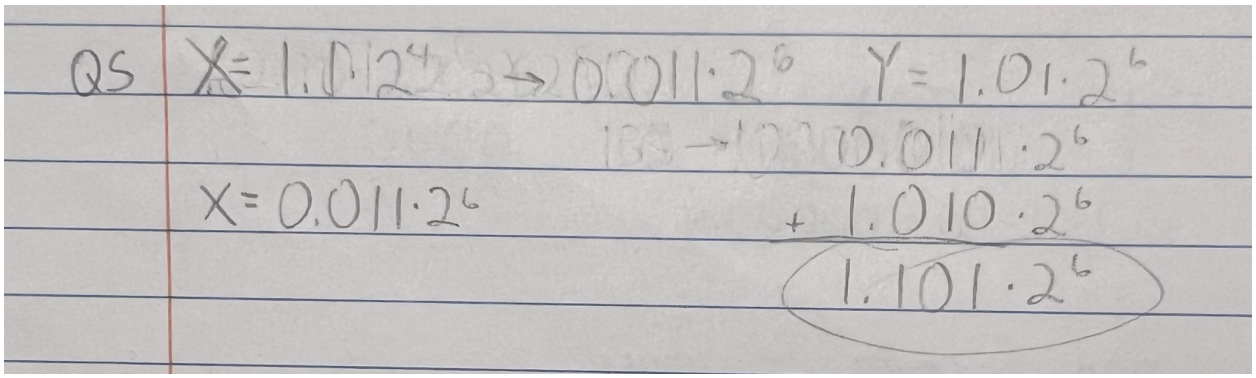
#1 shift X or Y for equal exponents

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31	32
X	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
After shift																																
X	0	1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

#2 Add signed significands and normalize results.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31	32
X	0	1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	1	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



QUESTION 6

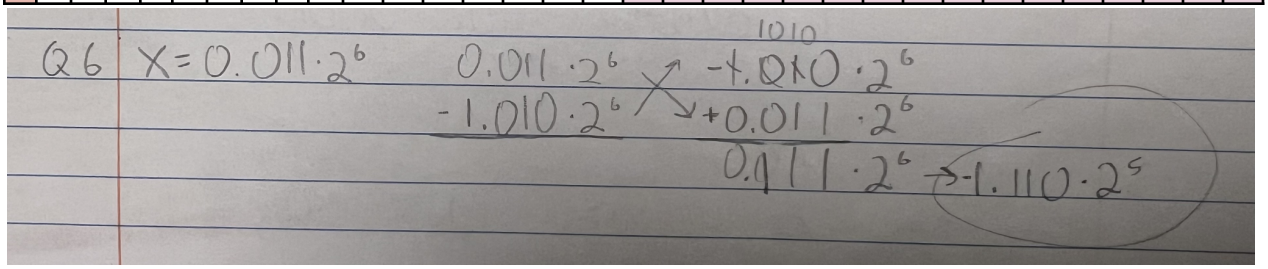
Following the above question (4), perform Z = X - Y using Floating-Point Addition and Subtraction as discussed in class. [10 pts]

Use a table similar to the above questions and show all steps.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31	32
X	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
After shift																																
X	0	1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31	32
X	0	1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Y	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

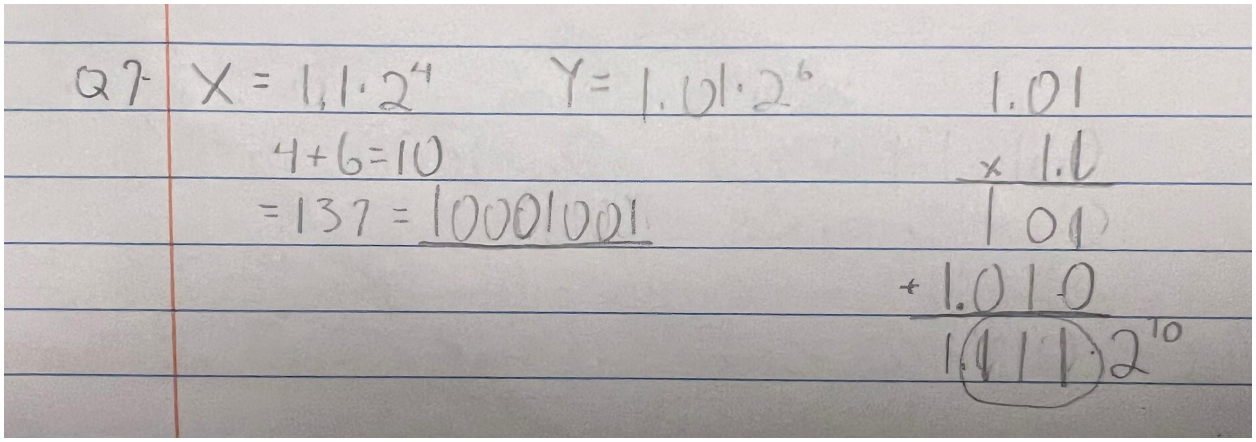


QUESTION 7

Following the above question (4), perform Z = X * Y using Floating-Point Multiplication as discussed in class. [10 pts]

Use a table similar to the above questions and show all steps.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31	32
X	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

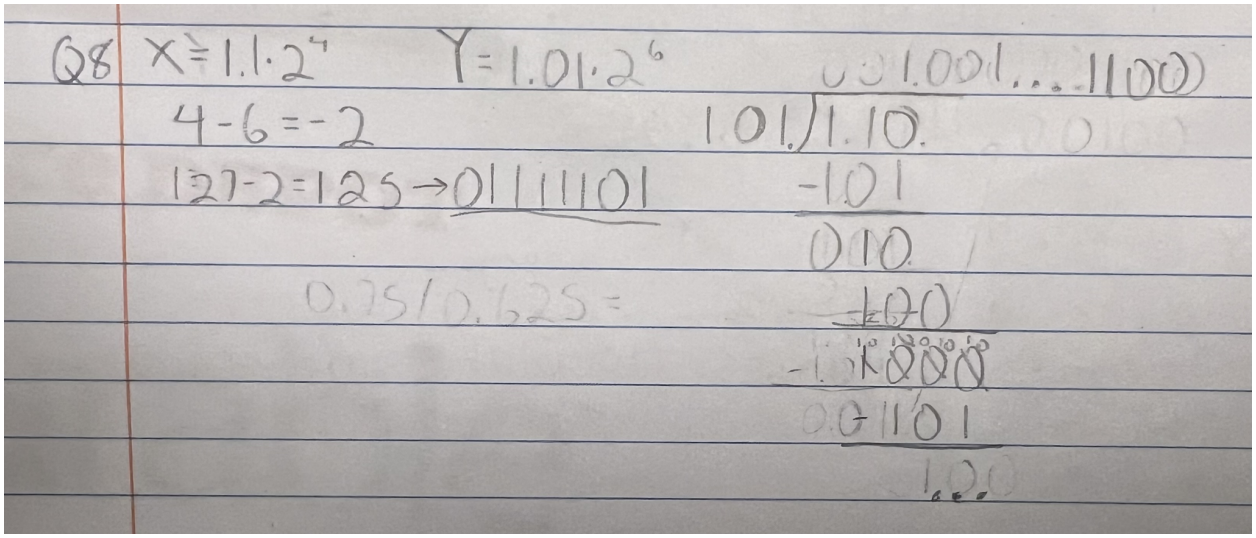


QUESTION 8

Following the above question (4), perform Z = X / Y using Floating-Point Division as discussed in class. [10 pts]

Use a table similar to the above questions and show all steps.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28	30	31	32
X	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Y	0	1	0	0	0	0	1	0	**1**	**0**	**1**	**0**	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**0**


Q8
X=1.1·2^4
Y=1.01·2^6
4-6=-2
127-2=125→01111101
0.75/0.625=
01.001...1100
101/1.10.
-101
010
-000
-10000
00110
1.0