

Imperial College London
Department of Computing

Computer Systems (113) / Architecture (110)
Exercises – *Floating Point Numbers*

- 1) Convert the following decimal numbers to binary: a) 5.5 b) 8.25 c) 9.3
- 2) Convert the binary number 1001.1010101 to decimal.
- 3) Normalise the following binary numbers: a) 101.1 b) 1000.01 c) 0.00010101
- 4) Convert –31.3 to IEEE Single Precision format.
- 5) Interpret the 32-bit hexadecimal value C154 0000 as an IEEE Single Precision number.
- 6) Carry out the operation $31.3 + 13.25$ in IEEE Single Precision arithmetic
- 7) Fill in the missing entries

Fraction	Binary	Decimal
1/4	0.01	0.25
3/8		
23/16		
	10.1101	
	1.011	
		5.625
		3.0625

- 8) Consider a five-bit floating representation based on the IEEE floating point format with 1 sign bit, two exponent bits and 2 significand bits. For this format fill in the missing entries:

Bits	Binary Value or Special Value	Decimal value or Special Value
0 00 00		
0 00 01		
0 00 10		
0 00 11		
0 01 00		
0 01 01		
0 01 10		
0 01 11		
0 10 00		
0 10 01		
0 10 10		
0 10 11		
0 11 00		
0 11 01		
0 11 10		
0 11 11		

Imperial College of Science, Technology and Medicine
Department of Computing

Computer Systems (113) / Architecture (110)
Solutions – *Floating Point Numbers*

1) Binary fractions are:

a) 5.5 is **101.1**

b) 8.25 is **1000.01**

c) 9 is 1001

0.3 \Rightarrow **0.6, 1.2, 0.4, 0.8, 1.6, 1.2,**
= 01001 1001 1001 etc.

9.3 is **1001. 01001 1001 1001** repeating etc

2) Convert the binary number 1001.1010101 to decimal.

1001 binary is 9 decimal

.	1	0	1	0	1	0	1
128	64	32	16	8	4 .	2	1

Sum=85

Fraction = $85 / 128 = 0.6640625$

Number = **9.6640625**

3) a) $101.1 = \mathbf{1.011} \times 2^2$

b) $1000.01 = \mathbf{1.00001} \times 2^3$

c) $0.00010101 = \mathbf{1.0101} \times 2^{-4}$

4) Convert -31.3 to IEEE Single Precision format.

First convert to a binary number -31.3 = -11111.01001 1001 1001

Next Normalise

$1.11110\ 1001\ 1001\ 1001\ 1001\ 1001 \times 2^4$

Significand field is **1111 0100 1100 1100 1100 110** (23 bits with 1. omitted)

Exponent field is $4+127 = 131 = \mathbf{1000\ 0011}$

Number is -ve therefore Sign field is **1**

Sign	Exponent	Significand
1	1000 0011	1111 0100 1100 1100 1100 110

5) Convert the IEEE Single Precision format hex value C154 0000 to decimal.

C154 0000 = 1100 0001 0101 0100 0000 0000 0000 0000

Sign	Exponent	Significand
1	1000 0010	1010 1000 0000 0000 0000 000

Exponent field = 1000 0010 = 130 => Exponent = 130 - 127 = 3

Significand field = 10101 Adding Hidden Bit => 1.10101

Therefore number is $1.10101 \times 2^3 = 1101.01 = \text{Decimal } 13.25$

Sign is 1 therefore number is **-13.25**

6) Carry out the operation 31.3 + 13.25 in IEEE single precision arithmetic

Number	Sign	Exponent	Significand
31.3	0	1000 0011	1111 0100 1100 1100 1100 110
13.25	0	1000 0010	1010 1000 0000 0000 0000 000

Significand of Larger Number = 1.1111 0100 1100 1100 1100 110

Significand of Smaller Number = 1.1010 1000 0000 0000 0000 000

Exponents differ by 1. Therefore shift binary point of Smaller Number 1 place.

Significand of Larger Number = 1.1111 0100 1100 1100 1100 1100

Significand of Smaller Number = 0.1101 0100 0000 0000 0000 0000

Significand of Sum = 10.1100 1000 1100 1100 1100 1100

Sum = $10.1100 1000 1100 1100 1100 1100 \times 2^4$

Normalise $1.01100 1000 1100 1100 1100 1100 \times 2^5$

Sign	Exponent	Significand
0	1000 0100	0110 0100 0110 0110 0110 011

7)

Fraction	Binary	Decimal
1/4	0.01	0.25
3/8	0.011	0.375
23/16	1.0111	1.4375
45/16	10.1101	2.8125
11/8	1.011	1.375
45/8	101.101	5.625
49/16	11.0001	3.0625

8)

Bits	Binary value or special value	Decimal value or special value
0 00 00	0	0
0 00 01	0.01	0.25
0 00 10	0.10	0.50
0 00 11	0.11	0.75
0 01 00	1.00	1
0 01 01	1.01	1.25
0 01 10	1.10	1.5
0 01 11	1.11	1.75
0 10 00	10.0	2
0 10 01	10.1	2.5
0 10 10	11.0	3
0 10 11	11.1	3.5
0 11 00	∞	∞
0 11 01	NaN	NaN
0 11 10	NaN	NaN
0 11 11	NaN	NaN