# Worksheet 1 Floating point form Answers

**Task 1**

1. Convert the following floating point numbers from binary to decimal. Show your working.

(a)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |  | 0 | 1 | 1 | 0 |

Mantissa Exponent

Mantissa: 0.1110101

Exponent = 6,

so move binary point 6 places right 0111010.1 = 58.5

(b)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |  | 0 | 1 | 0 | 0 |

Mantissa Exponent

Mantissa: 1.0001111

1s complement 0.1110000

Add 1 to get 2s complement 0.1110001

Exponent = 4,

so move binary point 4 places right 01110.001 = -14.125

(c)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |  | 1 | 1 | 1 | 1 |

Mantissa: 0.1110000

Exponent = -1 (the negative exponent can be calculated using 2s complement or as -8 + 7, as sign bit has the value -8)

so move binary point left 1 place 0.0111000 =.25 + .125 + .0625

= 0.4375

(d)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |  | 1 | 1 | 1 | 0 |

Mantissa: 1.0010000

1s complement 0.1101111

Add 1 to get 2s complement 0.1110000 (-)

Exponent = -2,

so move binary point 2 places left -0.0011100 =-( .125 + .0625 + .03125)

= - 0.21875

2. What is the largest number, in decimal that can be represented using this floating point system?

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | 0 | 1 | 1 | 1 |

Exponent = 7 so move binary point 7 places right

01111111 =127

Largest number = 127

**Task 2**

3. Convert the following binary numbers into normalised form:

(a)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |  | 0 | 0 | 1 | 1 |

Move the binary point 2 places right so that there is a 1 directly after the sign bit

This makes the number larger, so subtract 2 from the exponent

The normalised number is 0.1011000 0001

(b)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |  | 0 | 1 | 1 | 0 |

Move the binary point three places to the right so that there is a zero directly after the sign bit

Subtract 3 from the exponent

The normalised number is 1.0110000 0011

4. Convert the following from decimal to normalised binary floating point, using an 8-bit mantissa and a 4-bit exponent. show your working.

(a) 45.5

convert to binary 0101101.1

move point 6 places left 0.1011011 exponent 6

Answer 0.1011011 0110

(b) -14.5

convert 14.5 to binary 01110.100

one’s complement 10001.011

two’complement 10001.100

move point 4 places left 1.0001100 exponent 4

Answer: 1.0001100 0100

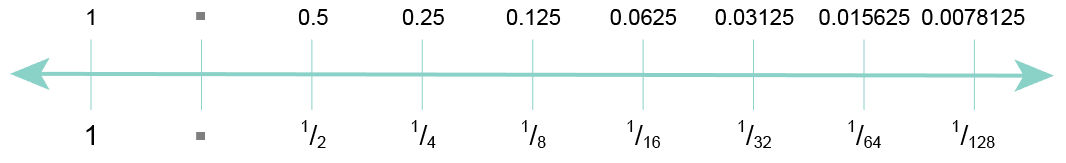
5. What is the most negative number that can be held in a 8-bit mantissa and a 4-bit exponent? Express the answer as a normalised floating point binary number.

Answer:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |  | 0 | 1 | 1 | 1 |

**Task 3**

6. Suppose a binary number is held in fixed point binary with 9 places before the binary point and 7 places after the binary point.



(a) A measurement of 0.007 is made in a scientific experiment. In fixed point binary, this is represented as 0.0078125. What is

(i) the absolute error? 0.0008125

(ii) the relative error, as a percentage? 11.607% (approx.)

(b) A different measurement is made. The new measurement is 1.007. In fixed point binary, this is represented as 1.0078125. What is

(i) the absolute error? 0.0008125

(ii) the relative error, as a percentage? 0.08%

(c) A third measurement is made. The new measurement is 100.007. In fixed point binary, this is represented as 100.0078125. What is

(i) the absolute error? 0.0008125

(ii) the relative error, as a percentage? 0.0008125%

Comment on the difference in relative errors for large and small magnitude numbers.

A small inaccuracy in a large number is a small relative error. The smaller the number, the larger the relative error for the same absolute error.

7. Using the fixed point representation of binary numbers shown in question 6, and assuming 9 bits before the binary point, with the first being the sign bit, give examples of calculations that would cause

(a) overflow

Max positive number that can be held is 127. So any calculation that results in a number greater than 127 will cause overflow.

(b) underflow

Minimum number that can be held is 0.0078125. Any calculation , e.g. subtraction or division, or multiplication by a number less than 1, will cause underflow.

8. Why do rounding errors sometimes occur when a computer performs calculations? Give an example using fixed point binary

Because some numbers, e.g. denary 0.1, cannot be represented exactly in binary, however many bits are allocated.

Rounding may also occur if not enough bits are allocated to the mantissa to hold a number precisely. e.g. 1.0078126 cannot be held precisely using the representation held in question 6.