# Worksheet 1 Floating point form

**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

(b)

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

Mantissa Exponent

(c)

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

(d)

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

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

**Task 2**

3. Convert the following binary numbers into normalised form:

(a)

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

(b)

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

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

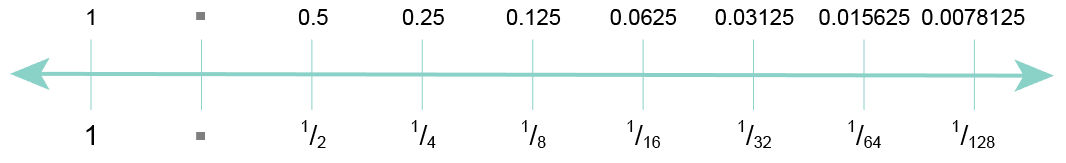
(b) -14.5

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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

**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?

(ii) the relative error, as a percentage?

(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?

(ii) the relative error, as a percentage?

(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?

(ii) the relative error, as a percentage?

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

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

(b) underflow

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