Representing Numbers

Floating point numbers :

1. IEEE standard :Stores like scientific notation :

625,9 becomes 0,6259 \* 10^3

* 0,6259 is the significand – or mantissa, or coefficient
* 10^3 is the exponent

1. 32-bits Floating Point Number

Fixed point numbers :

When we write a number such as 1.234 as a decimal fraction (that has a decimal point), we mean one unit + 2 tenths + 3 hundredths + 4 thousandths:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 100 | 10 | Units |  | 1/10 | 1/100 | 1/1000 |
| 0 | 0 | 0 | 1 | . | 2 | 3 | 4 |

We can use this same principle for binary numbers except with the use of a binary point. Each decimal place is divided by 2, moving from left to right.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 16 | 8 | 4 | 2 | Units | . | ½ | ¼ | 1/8 | 1/16 | 1/32 |
| 0 | 0 | 0 | 0 | 1 | . | 1 | 1 | 0 | 1 | 1 |

With fixed point representation, we use a fixed number of bits for the fractional part of the number (that is, a fixed number of bits after the binary point).

It doesn’t work well with very big or very small numbers though – that’s where we use floating point instead.

For example, the binary 1.1101 is:

1 + ½ + ¼ + 0/8 + 1/16 + 1/32 = 1 + 0.5 + 0.25 + 0.0625 + 0.03125 = 1.84375

Convert a decimal number to a fixed-point binary:

We will convert 7.75 to the fixed-point binary using 4 bits after the binary point:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8 | 4 | 2 | Units | . | ½ | ¼ | 1/8 | 1/16 |
| 0 | 1 | 1 | 1 | . | 1 | 1 | 0 | 0 |

7.75 is 111.110 in binary fixed-point representation.