Binary Numbers

Binary Numbers

Math Fact:
- any number to the zero power is 1
- any number to the one power is itself
- after that the power is the number of times a number is multiplied times itself.

Binary numbers are in base 2, so 2^x is

$2^0 = 1$

$2^1 = 2$

$2^2 = 2 \times 2 = 4$

$2^3 = 2 \times 2 \times 2 = 8$

etc

decimal number is base 10 so 10^x

$10^0 = 1$

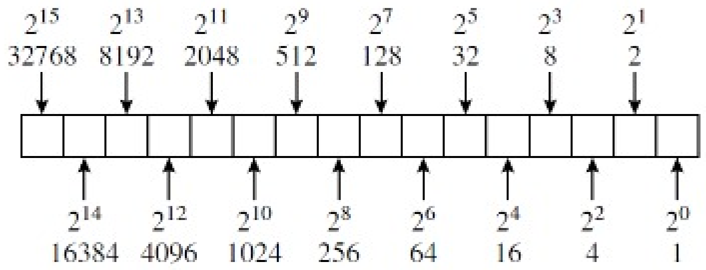
$10^1 = 10$

$10^2 = 10 \times 10 = 100$

$10^3 = 10 \times 10 \times 10 = 1000$

$10^4 = 10 \times 10 \times 10 \times 10 = 10000$

Like decimal numbers, a binary numbers' place value uses base number to the x power.



Start with 1 and progress to the left for each new place.

Continues Indefinitely

10^5	10^4	10^3	10^2	10^1	10^0
100,000 Hundred-thousands place	10,000 Ten-thousands place	1,000 Thousands place	100 Hundreds place	10 Tens place	1 Ones place

Normalization and Precision

convert number to 10 cubed power by moving the decimal point one place to the left for every power I'm adding to the exponent.

100 is $1.00 \times 10^2 = 0.100 \times 10^3$

1000 is 1.000×10^3

sum is 1.100×10^3

In the problem below, a variable of type double variable CANNOT hold all the decimal places of the sum. It can only hold 15 decimal places. So everything to the right of .250 is LOST because it cannot be stored in the variable.

329,169,278,153.25
\+ 0.0000000000000000001

329,169,278,153.2500000000000000001

double has 15 decimal places of precision.

this data is lost

To add two numbers in scientific notation together, the smaller number has to be converted the same exponent of the larger first.

\- This makes sense when you think about the fact that 1dollar plus one million dollars is one million and one dollars, not 2 million dollars.

1,000,000 + 1 = 1,000,001

1 x 10^6 + 1 x 10^1 = ~~2 x 10^6~~

miro

Double and float representation

The number -10 is shown in binary IEEE format for type `float`.

	sign	exponent		mantissa
$-10.0 = -1.25 \times 2^3 =$	1	1000001 0	1.	0100000 00000000 00000000
	31	30 ... 23		22 ... 0

- The sign bit is 1, indicating a negative number
- The exponent 10000010 is represented in excess 127 notation, which means that we must subtract 127 from the binary number shown to get the true exponent: $130 - 127 = 3$
- The mantissa is $1.01000\ldots$, which means $1 + 1/4 = 1.25$

Figure 7.6. Binary representation of reals.

These are the minimum value ranges for the IEEE floating-point types. The names given in this table are the ones defined by the C standard.

Type Name	Digits of Precision	Name of C Constant	Minimum Value Range Required by IEEE Standard
float	6	±FLT_MIN...±FLT_MAX	±1.175E−38 ... ±3.402E+38
double	15	±DBL_MIN...±DBL_MAX	±2.225E−308 ... ±1.797E+308

Figure 7.7. IEEE floating-point types.

Number of decimal digits that the mantissa can represent

Largest/smallest power the exponent can represent

You need to know the parts that make up a float/double. Sign bit, exponent and mantissa.

The IEEE defined these values based on the kinds of numbers Engineers would need to be able to use and the ISO standard adopted their requirements.

miro