

# Fixed precision numbers

Classes **Fix16**, **Fix24**, **Fix32**, and **Fix48** support operations on 16, 24, 32, or 48 bit quantities that are considered as real numbers in the range [-1, +1). Such numbers are often encountered in digital signal processing applications. The classes may be used in isolation or together. Class **Fix32** operations are entirely self-contained. Class **Fix16** operations are self-contained except that the multiplication operation **Fix16 \* Fix16** returns a **Fix32**. **Fix24** and **Fix48** are similarly related.

The standard arithmetic and relational operations are supported (**=**, **+**, **-**, **\***, **/**, **<<**, **>>**, **+=**, **-=**, **\*=**, **/=**, **<=<**, **>=>**, **==**, **!=**, **<**, **<=**, **>**, **>=**). All operations include provisions for special handling in cases where the result exceeds +/- 1.0. There are two cases that may be handled separately: ```overflow``` where the results of addition and subtraction operations go out of range, and all other ```range errors``` in which resulting values go off-scale (as with division operations, and assignment or initialization with off-scale values). In signal processing applications, it is often useful to handle these two cases differently. Handlers take one argument, a reference to the integer mantissa of the offending value, which may then be manipulated. In cases of overflow, this value is the result of the (integer) arithmetic computation on the mantissa; in others it is a fully saturated (i.e., most positive or most negative) value. Handling may be reset to any of several provided functions or any other user-defined function via **set\_overflow\_handler** and **set\_range\_error\_handler**. The provided functions for **Fix16** are as follows (corresponding functions are also supported for the others).

## **Fix16\_overflow\_saturate**

The default overflow handler. Results are ```saturated```: positive results are set to the largest representable value (binary 0.111111...), and negative values to -1.0.

## **Fix16\_ignore**

Performs no action. For overflow, this will allow addition and subtraction operations to ```wrap around``` in the same manner as integer arithmetic, and for saturation, will leave values saturated.

## **Fix16\_overflow\_warning\_saturate**

Prints a warning message on standard error, then saturates the results.

**Fix16\_warning**

The default range\_error handler. Prints a warning message on standard error; otherwise leaving the argument unmodified.

**Fix16\_abort**

prints an error message on standard error, then aborts execution.

In addition to arithmetic operations, the following are provided:

**Fix16 a = 0.5;**

Constructs fixed precision objects from double precision values. Attempting to initialize to a value outside the range invokes the range\_error handler, except, as a convenience, initialization to 1.0 sets the variable to the most positive representable value (binary 0.1111111...) without invoking the handler.

**short& mantissa(a); long& mantissa(b);**

return  $a * \text{pow}(2, 15)$  or  $b * \text{pow}(2, 31)$  as an integer. These are returned by reference, to enable "manual" data manipulation.

**double value(a); double value(b);**

return a or b as floating point numbers.