Proposal for C23
WG14 N2746

Title: overflow and underflow definitions
Author, affiliation: C FP group
Date: 2021-04-24
Proposal category: Editorial
Reference: N2596

7.12.1 in the current C23 draft (N2596) defines overflow and underflow:

> [5] A floating result overflows if the magnitude (absolute value) of the mathematical result is finite but so large that the mathematical result cannot be represented without extraordinary roundoff error in an object of the specified type. …
>
> [6] The result underflows if the magnitude (absolute value) of the mathematical result is nonzero and less than the minimum normal number in the type.249) …
>
> 249)The term underflow here is intended to encompass both "gradual underflow" as in IEC 60559 and also "flush-to-zero" underflow.

Problem 1: The use of “mathematical result” is not appropriate here. It might well be taken to mean the infinitely precise value of the mathematical function corresponding to the C function. But C doesn’t require correct rounding. The implementation might compute an estimate that overflows where the mathematical result would not. The following suggested changes eliminate these uses of “mathematical result”.

The C23 draft contains one other use of “mathematical result”, which the changes in another proposal eliminate.

Problem 2: The definition of underflow excludes the IEC 60559 underflows that are outside the normal range before but not after rounding. This is contrary to footnote 249. The following suggested changes broaden the definition of underflow to include all IEC 60559 underflows. Broadening the definition does not break implementations because reporting of underflow range errors is optional in C.

The changes also add a sentence to footnote 249 explaining why the definition is broader than might be expected.

Suggested changes:

Changes in 7.12.1:

> [5] A floating result overflows if ~~the~~ a finite result value with ordinary accuracy would have magnitude (absolute value) ~~of the mathematical result is finite but so large that the mathematical result cannot be represented without extraordinary roundoff error~~ too large for representation in an object of the specified type. …
>
> [6] The result underflows if ~~the~~ a nonzero result value with ordinary accuracy would have magnitude (absolute value) ~~of the mathematical result is nonzero and less~~ no larger than the minimum normalized number in the type.249) …
>
> 249)The term underflow here is intended to encompass both "gradual underflow" as in IEC 60559 and also "flush-to-zero" underflow. IEC 60559 underflow can occur in cases where the magnitude of the rounded result equals the minimum normalized number in the format.