TECHNICAL REPORT

ISO/IEC TR 24733

First edition
2011-11-01

Information technology — Programming languages, their environments and system software interfaces — Extensions for the programming language C++ to support decimal floating-point arithmetic

Technologies de l'information — Langages de programmation, leurs environnements et interfaces de logiciel système — Extensions pour le langage de programmation C++ pour supporter l'arithmétique flottante décimale

Reference number
ISO/IEC TR 24733:2011(E)

Contents

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, when the joint technical committee has collected data of a different kind from that which is normally published as an International Standard (“state of the art”, for example), it may decide to publish a Technical Report. A Technical Report is entirely informative in nature and shall be subject to review every five years in the same manner as an International Standard.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC TR 24733 was prepared jointly by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 22, *Programming languages, their environments and system software interfaces.*

Introduction

0.1 General

Most of today's general purpose computing architectures provide binary floating-point arithmetic in hardware. Binary float-point is an efficient representation that minimizes memory use, and is simpler to implement than floating-point arithmetic using other bases. It has therefore become the norm for scientific computations, with almost all implementations following the IEEE-754 standard for binary floating-point arithmetic.

However, human computation and communication of numeric values almost always uses decimal arithmetic, and decimal notations. Laboratory notes, scientific papers, legal documents, business reports and financial statements all record numeric values in decimal form. When numeric data are given to a program or are displayed to a user, binary to-and-from decimal conversion is required. There are inherent rounding errors involved in such conversions; decimal fractions cannot, in general, be represented exactly by floating-point values. These errors often cause usability and efficiency problems, depending on the application.

These problems are minor when the application domain accepts, or requires results to have, associated error estimates (as is the case with scientific applications). However, in business and financial applications, computations are required either to be exact (with no rounding errors) unless explicitly rounded, or to be supported by detailed analyses that are auditable to be correct. Such applications therefore have to take special care in handling any rounding errors introduced by the computations.

The most efficient way to avoid conversion error is to use decimal arithmetic. Recognizing this, the IEEE 754-2008 Standard for Floating-Point Arithmetic specifies decimal floating-point encodings and arithmetic. This technical report specifies extensions to the International Standard for the C++ programming language to permit the use of decimal arithmetic in a manner consistent with the IEEE 754-2008 standard.

0.2 Arithmetic model

This Technical Report is based on a model of decimal arithmetic which is a formalization of the decimal system of numeration (algorism) as further defined and constrained by the relevant standards, IEEE 854, ANSI X3-274, and IEEE 754-2008.

There are three components to the model:

- *data*- numbers and NaNs, which can be manipulated by, or be the results of, the core operations defined in the model
- *operations*- (such as addition, multiplication, etc.) which can be carried out on data
- *context* - the status of operations (namely, exceptions flags), and controls to govern the results of operations (for example, rounding modes).

The model defines these components in the abstract. It defines neither the way in which operations are expressed (which might vary depending on the computer language or other interface being used), nor the concrete representation (specific layout in storage, or in a processor's register, for example) of data or context.

From the perspective of the C++ language, *data* are represented by data types, *operations* are defined within expressions, and *context* is the floating environment specified in `<cfenv>` and `<fenv.h>`. This Technical Report specifies how the C++ language implements these components.

0.3 The formats

In the C++ International Standard, the representation of a floating-point number is specified in an abstract form where the constituent components of the representation are defined (sign, exponent, significand) but the internals of these components are not. In particular, the exponent range, significand size and the base (or radix), are implementation defined. This allows flexibility for an implementation to take advantage of its underlying hardware architecture. Furthermore, certain behaviors of operations are also implementation defined, for example in the area of handling of special numbers and in exceptions.

This approach was inherited from the C programming language. At the time that C was first standardized, there were already various hardware implementations of floating-point arithmetic in common use. Specifying the exact details of a representation would make most of the existing C implementations at the time not conforming.

This Technical Report specifies decimal floating-point arithmetic according to the IEEE 754-2008 standard, with the constituent components of the representation defined. This is more stringent than the approach taken for the floating-point types in the C++ standard. Since it is expected that all decimal floating-point hardware implementations will conform to the IEEE 754-2008 standard, binding to this standard directly benefits both implementers and programmers.

Information technology — Programming languages, their environments and system software interfaces — Extensions for the programming language C++ to support decimal floating-point arithmetic

1 Scope

This Technical Report specifies an extension to the programming language C++, specified by ISO/IEC 14882:2003. The extension provides support for decimal floating-point arithmetic that is consistent with the specification in IEEE 754-2008. Any conflict between the requirements described here and that specification is unintentional.

The binary floating-point arithmetic specified in IEEE 754-2008 is not considered in this Technical Report.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 14882:2003, *Programming languages — C++*

ISO/IEC TR 19768:2007, *Information technology — Programming languages — Technical Report on C++ Library Extensions*

ISO/IEC TR 24732:2009, *Information technology — Programming languages, their environments and system software interfaces — Extension for the programming language C to support decimal floating-point arithmetic*

IEEE 754-2008, *IEEE Standard for Floating-Point Arithmetic*