ISO/IEC 10967-1:2012-07 (E)

Information technology - Language independent arithmetic - Part 1: Integer and floating point arithmetic

Contents

Page

Foreword vii

Introduction viii

1 Scope 1
- 1.1 Inclusions . . . 1
- 1.2 Exclusions . . . 2

2 Conformity 3

3 Normative references 4

4 Symbols and definitions 4
- 4.1 Symbols . . . 4
  - 4.1.1 Operators and relations . . . 4
  - 4.1.2 Sets and intervals . . . 5
  - 4.1.3 Exceptional values . . . 5
  - 4.1.4 Special values . . . 6
  - 4.1.5 The Boolean datatype . . . 6
  - 4.1.6 Operation specification framework . . . 6
- 4.2 Definitions of terms . . . 7

5 Specifications for integer and floating point datatypes and operations 12
- 5.1 Integer datatypes and operations . . . 13
  - 5.1.1 Integer result function . . . 14
  - 5.1.2 Integer operations . . . 14
    - 5.1.2.1 Comparisons . . . 14
    - 5.1.2.2 Basic arithmetic . . . 15
- 5.2 Floating point datatypes and operations . . . 17
  - 5.2.1 Conformity to IEC 60559 . . . 19
  - 5.2.2 Range and granularity constants . . . 19
  - 5.2.3 Approximate operations . . . 19
  - 5.2.4 Rounding and rounding constants . . . 20
  - 5.2.5 Floating point result function . . . 21
  - 5.2.6 Floating point operations . . . 22
    - 5.2.6.1 Comparisons . . . 22
    - 5.2.6.2 Basic arithmetic . . . 24
    - 5.2.6.3 Value dissection . . . 27
    - 5.2.6.4 Value splitting . . . 29
- 5.3 Operations for conversion between numeric datatypes . . . 29
  - 5.3.1 Integer to integer conversions . . . 30
  - 5.3.2 Floating point to integer conversions . . . 31
  - 5.3.3 Integer to floating point conversions . . . 31
  - 5.3.4 Floating point to floating point conversions . . . 32
  - 5.3.5 Floating point to fixed point conversions . . . 32
  - 5.3.6 Fixed point to floating point conversions . . . 34
- 5.4 Numerals as operations in a programming language . . . 34

5.4.1 Numerals for integer datatypes . . . 34
5.4.2 Numerals for floating point datatypes . . . 35

6 Notification **35**
6.1 Model for handling of notifications . . . 35
6.2 Notification alternatives . . . 36
6.2.1 Notification by recording in indicators . . . 36
6.2.2 Notification by alteration of control flow . . . 38
6.2.3 Notification by termination with message . . . 38
6.3 Delays in notification . . . 39
6.4 User selection of alternative for notification . . . 39

7 Relationship with language standards **39**

8 Documentation requirements **41**

Annex A (informative) **Partial conformity** **43**
A.1 Integer overflow notification relaxation . . . 44
A.2 Infinitary notification relaxation . . . 44
A.3 Inexact notification relaxation . . . 44
A.4 Underflow notification relaxation . . . 45
A.5 Subnormal values relaxation . . . 45
A.6 Accuracy relaxation for add, subtract, multiply, and divide . . . 45
A.7 Accuracy relaxation for floating point conversion . . . 47

Annex B (informative) **IEC 60559 bindings** **51**
B.1 Summary . . . 51
B.2 Notification . . . 55

Annex C (informative) **Rationale** **57**
C.1 Scope . . . 57
C.1.1 Inclusions . . . 57
C.1.2 Exclusions . . . 57
C.1.3 Companion parts to this part . . . 58
C.2 Conformity . . . 58
C.2.1 Validation . . . 59
C.3 Normative references . . . 59
C.4 Symbols and definitions . . . 59
C.4.1 Symbols . . . 60
C.4.2 Definitions of terms . . . 60
C.5 Specifications for integer and floating point datatypes and operations . . . 61
C.5.1 Integer datatypes and operations . . . 62
C.5.1.0.1 Unbounded integers . . . 62
C.5.1.0.2 Bounded non-modulo integers . . . 63
C.5.1.0.3 Modulo integers . . . 64
C.5.1.1 Integer result function . . . 64
C.5.1.2 Integer operations . . . 64
C.5.1.2.1 Comparisons . . . 64
C.5.1.2.2 Basic arithmetic . . . 65
C.5.2 Floating point datatypes and operations . . . 65
C.5.2.0.1 Constraints on the floating point parameters . . . 66
C.5.2.0.2 Radix complement floating point . . . 68
C.5.2.1 Conformity to IEC 60559 . . . 68
C.5.2.1.1 Subnormal numbers . . . 69
C.5.2.1.2 Signed zero . . . 69
C.5.2.1.3 Infinities and NaNs . . . 69
C.5.2.2 Range and granularity constants . . . 70
C.5.2.2.1 Relations among floating point datatypes . . . 70

C.5.2.3 Approximate operations . . . 71
C.5.2.4 Rounding and rounding constants . . . 71
C.5.2.5 Floating point result function . . . 73
C.5.2.6 Floating point operations . . . 73
C.5.2.6.1 Comparisons . . . 73
C.5.2.6.2 Basic arithmetic . . . 73
C.5.2.6.3 Value dissection . . . 74
C.5.2.6.4 Value splitting . . . 74
C.5.2.7 Levels of predictability . . . 75
C.5.2.8 Identities . . . 76
C.5.2.9 Precision, accuracy, and error . . . 78
C.5.2.9.1 LIA-1 and error . . . 79
C.5.2.9.2 Empirical and modelling errors . . . 80
C.5.2.9.3 Propagation of errors . . . 80
C.5.2.10 Extra precision . . . 81
C.5.3 Operations for conversion between numeric datatypes . . . 82
C.5.4 Numerals as operations in a programming language . . . 83
C.5.4.1 Numerals for integer datatypes . . . 83
C.5.4.2 Numerals for floating point datatypes . . . 83
C.6 Notification . . . 83
C.6.1 Model handling of notifications . . . 84
C.6.2 Notification alternatives . . . 84
C.6.2.1 Notification by recording in indicators . . . 84
C.6.2.2 Notification by alteration of control flow . . . 85
C.6.2.3 Notification by termination with message . . . 86
C.6.3 Delays in notification . . . 86
C.6.4 User selection of alternative for notification . . . 86
C.7 Relationship with language standards . . . 87
C.8 Documentation requirements . . . 88

Annex D (informative) **Example bindings for specific languages** **89**
D.1 Ada . . . 90
D.2 C . . . 96
D.3 C++ . . . 104
D.4 Fortran . . . 111
D.5 Common Lisp . . . 115

Annex E (informative) **Example of a conformity statement** **121**
E.1 Types . . . 121
E.2 Integer parameters . . . 121
E.3 Floating point parameters . . . 122
E.4 Expressions . . . 122
E.5 Notification . . . 122

Annex F (informative) **Example programs** **125**
F.1 Verifying platform acceptability . . . 125
F.2 Selecting alternate code . . . 125
F.3 Terminating a loop . . . 126
F.4 Estimating error . . . 126
F.5 Saving exception state . . . 126
F.6 Fast versus accurate . . . 127
F.7 High-precision multiply . . . 127

Bibliography **129**